DAGSTUHL
REPORTS

Volume 13, Issue 1, January 2023

Dagstuhl Reports, Vol. 13, Issue 1 ISSN 2192-5283

ISSN 2192-5283

Published online and open access by
Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at
https://www.dagstuhl.de/dagpub/2192-5283

Publication date
September, 2023

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at `https://dnb.d-nb.de`.

Digital Object Identifier: 10.4230/DagRep.13.1.i

Aims and Scope
The periodical *Dagstuhl Reports* documents the program and the results of Dagstuhl Seminars and Dagstuhl Perspectives Workshops.
In principal, for each Dagstuhl Seminar or Dagstuhl Perspectives Workshop a report is published that contains the following:

- an executive summary of the seminar program and the fundamental results,
- an overview of the talks given during the seminar (summarized as talk abstracts), and
- summaries from working groups (if applicable).

This basic framework can be extended by suitable contributions that are related to the program of the seminar, e. g. summaries from panel discussions or open problem sessions.

Contact
Schloss Dagstuhl – Leibniz-Zentrum für Informatik
Dagstuhl Reports, Editorial Office
Oktavie-Allee, 66687 Wadern, Germany
reports@dagstuhl.de
https://www.dagstuhl.de/dagrep

Report from Dagstuhl Seminar 23021

Media Forensics and the Challenge of Big Data

Irene Amerini[*1], **Anderson Rocha**[*2], **Paul L. Rosin**[*3], **and Xianfang Sun**[*4]

1 **Sapienza University of Rome, IT.** amerini@diag.uniroma1.it
2 **State University – Campinas, BR.** arrocha@unicamp.br
3 **Cardiff University, GB.** paul.rosin@cs.cf.ac.uk
4 **Cardiff University, GB.** sunx2@cardiff.ac.uk

Abstract

With demanding and sophisticated crimes and terrorist threats becoming more pervasive, allied with the advent and widespread of fake news, it becomes paramount to design and develop objective and scientific-based criteria to identify the characteristics of investigated materials associated with potential criminal activities. We need effective approaches to help us answer the four most important questions in forensics regarding an event: "who," "in what circumstances," "why," and "how." In recent years, the rise of social media has resulted in a flood of media content. As well as providing a challenge due to the increase in data that needs fact-checking, it also allows leveraging big-data techniques for forensic analysis.

The seminar included sessions on traditional, deep learning-based methods, big data, benchmark and performance evaluation, applications, and future directions. It aimed to orchestrate the research community's efforts in such a way that we harness different tools to fight misinformation and the spread of fake content.

Seminar January 8–13, 2023 – https://www.dagstuhl.de/23021
2012 ACM Subject Classification Applied computing → Computer forensics; Computing methodologies → Image manipulation
Keywords and phrases Digital forensics, Image and video authentication, Image and video forensics, Image and video forgery detection, Tampering detection
Digital Object Identifier 10.4230/DagRep.13.1.1

1 Executive Summary

Anderson Rocha (State University – Campinas, BR)

This summary summarizes the outcomes of our Dagstuhl Seminar. The seminar focused on

- important issues,
- relevant problems, and
- adequate solutions.

In the end, we provide a panorama of the last 20 years of the area, its main advances, and its challenges ahead. We go through several key aspects regarding research and development, the translational gap between academia and industry, and what we need to fill this gap. We also highlight key areas and decisions we must focus on in the years ahead. Digital Forensics is part of our lives, and we need to bring together the best minds to tackle its open problems and challenges.

[*] Editor / Organizer

Media Forensics and the Challenge of Big Data, *Dagstuhl Reports*, Vol. 13, Issue 1, pp. 1–35
Editors: Irene Amerini, Anderson Rocha, Paul L. Rosin, and Xianfang Sun
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

In our discussions, we confront traditional techniques with a range of new data-driven solutions, clearly pointing out the advantages and disadvantages of each kind of formulation. We also discuss their needs regarding scaling up to deal with ever-growing data sets.

We bring to bear aspects related to the development of fair, accountable, unbiased, and explainable solutions respecting directives such as the General Data Protection Regulation.

Finally, we point out that one of the biggest challenges nowadays in the presence of big data is the emergence of artificial intelligence generative techniques that easily allow the creation of never-seen-before content at unprecedented scale and speed, giving rise to what we have been referring to as synthetic realities. Only an orchestrated effort taking advantage of all different techniques from various formulations will allow us to fight back against such synthesized realities.

2 Table of Contents

Executive Summary

Overview of Talks

Round Table Discussions

3 Overview of Talks

3.1 Traditional Methods in Forensics

Mauro Barni (University of Siena, IT)

The dawn of multimedia forensics traces back to some seminal works published in the early 2000s by researchers previously working on steganalysis. Such works focused mostly on camera identification and detection of double JPEG compression. Since then, a large number of techniques have been developed dealing with a wide variety of forensic problems, including[1] detection of image resizing, color correction, detection of copy-move editing, detection of geometric and illumination inconsistencies etc…The methods developed in the first decade of multimedia research were based on the intuition that every step in the life of a multimedia document leaves within it a specific trace, often referred to as fingerprint or footprint, whose presence (or absence) can be used to derive some useful information about the past history of the document. Most methods developed in that period were adopting a model-based approach, according to which the process leading to the generation of the footprint was carefully modeled (by means of geometric or statistical tools), and the model used to develop sound footprint detection and/or localization techniques. In some cases, the forensic models were quite accurate allowing the development of extremely powerful tools. This was the case, for instance, of source camera identification based on PRNU (Photo-Response-Non-Uniformity) and detection of copy-move forgeries. This approach contrasts with more recent data-driven techniques based on deep neural network architectures, which base their success on the availability of massive amounts of training data. It is the goal of this talk to review the early history of multimedia forensics techniques and compare them with the most recent developments in the field, by paying particular attention to discuss the pros and cons of model-based and data-driven solutions, eventually advocating a synergistic use of both approached so to leverage on their complementary strengths.

3.2 Deep Learning in Multimedia Forensics

Christian Riess (Universität Erlangen-Nürnberg, DE)

Deep learning drives the development of new methods in Multimedia Forensics. Since deep learning derives decision rules from examples, it not only improves traditional model-based forensic tasks, but it also enables entirely new forensic tasks where analytic models cannot be constructed. However, after harvesting the immediate benefits of deep learning in forensics, we are now entering a period where its challenges become more visible.

In this talk, we discuss the most pressing challenges, and we raise the question for future directions of research. We hypothesize that a combination of the virtues of traditional methods with the power of deep learning can move the field significantly forward. The talk reviews four recent examples for such combinations, namely GAN fingerprints, image self-consistency, NoisePrint, and Bayesian learning.

[1] Here and afterwards we focus mainly on image forensics.

3.3 Compliance Challenges in Forensic Image Analysis Under the Artificial Intelligence Act

Benedikt Lorch (Universität Innsbruck, AT)

Joint work of Benedikt Lorch, Nicole Scheler, Christian Riess

In many applications of forensic image analysis, state-of-the-art results are nowadays achieved with AI methods. However, concerns about their reliability and opacity raise the question whether such methods can be used in criminal investigations from a legal perspective. In April 2021, the European Commission proposed the Artificial Intelligence Act, a regulatory framework for the trustworthy use of AI. Under the draft AI Act, high-risk AI systems for use in law enforcement are permitted but subject to compliance with mandatory requirements. In this paper, we summarize the mandatory requirements for high-risk AI systems and discuss these requirements in light of two forensic applications, license plate recognition and deep fake detection. The goal of this talk is to raise awareness of the upcoming legal requirements and to point out avenues for future research. For full details, see: [1].

References

1 Benedikt Lorch, Nicole Scheler, and Christian Riess. Compliance Challenges in Forensic Image Analysis Under the Artificial Intelligence Act. In *30th European Signal Processing Conference (EUSIPCO)*, pages 613–617. IEEE, 2022.

4 Round Table Discussions

4.1 Day 1 – Initial Introductory Discussions

Christian Riess (Universität Erlangen-Nürnberg, DE) – recorder of the session

Thorsten Beck introduces his work and background. He works on scientific integrity education. He reports about a database of images that was compiled by researchers at the Humboldt-Elsevier Advanced Data + Text Centre (HEADT Centre), supported by publishers such as Elsevier, PLOS, Frontiers and others. The images stem from retracted papers. From the point of view of a journal reviewer, he is interested in solutions to detecting the (very diverse) types of manipulations to support the reviewing process with an automated screening for image-based scientific fraud.

A discussion emerges on the challenges of analyzing such images. Concerns are raised that even though the database consists of about 500 papers (which may seem to be a lot from some point of view), the individual cases are too diverse to think about a "universal" forensic tool. HEADT Centre also came to this conclusion, which is why they work with major publishers to collect enough data for creating a training set for machine learning approaches to specific types of tampering, and to develop specific tools for scientific reviewers. Such a tool might inform a reviewer for example whether an image has been previously used in a publication (image repurposing), or whether there are indications for a copy-move forgery in an image. It is clear that such tools cannot cover all cases of fraud and cannot replace humans in the decision-making process. On the other hand, the Dagstuhl participants agree that such well-defined computational tasks are feasible goals to achieve, and may help to catch some cases of scientific fraud.

The discussion shifts towards the different roles of images in different scientific fields. In biomedical imaging, an image sometimes constitutes the actual contribution of a paper, as a proof for some type of (expected or unexpected) behavior. Similarly, for imaging or image generation tasks, the image is the "proof of work", and hence an integral part of the contribution. In contrast, in other fields of computer science, images oftentimes only serve as illustrations, and are hence less of a priority for forensic verification. It is also noted that in various fields (biology, computer graphics, computer vision) images often serve the purpose of advertising a work. It is also pointed out that a single image of a successful experiment may in most cases not be sufficient scientific proof per se, since it does not indicate anything about error probabilities. An analogy to COVID tests is made, which may be positive, but to get a satisfying statement, one should actually present a number of different tests and a confidence value associated with their accuracy.

The discussion then shifts towards the difficulty of realistically, conservatively assessing the performance of tools. Scientific results are often too optimistic. One notorious issue are evaluation setups that are too simple and do not cover the diversity of real-world data. Another prevalent issue are side channels in the evaluation dataset that greatly simplify the classification task. Several participants report first-hand experiences with such side channels across various application fields.

The discussion further shifts towards comparisons in scientific works. It is raised that one issue in the community are unfair comparisons due to a lack of care in fully tuning the competing algorithms for a comparative evaluation. Martin Steinebach mentions the "Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection" (PAN) challenges at the workshop of the CLEF initiative (Conference and Labs of the Evaluation Forum). Here, instead of performing a self-evaluation, the workshop is centered around challenges where participants submit a docker image and all code is evaluated at a central site, to ensure a fairer comparison of scientific results. Another example is SHREC, a shape retrieval contest. Here, a list of results is computed by each participant, and sent to the organizers for comparison to the ground truth. A criticism is that the participants can look at the test set, which is not possible in the docker approach.

The discussion returns to the challenges that Thorsten Beck initially raised. In particular, participants address the question what forensic algorithms can be considered to work robustly. The participants agree that copy-move is quite mature, and up until a couple of years ago Photo-Response Non-Uniformity (PRNU) was also a go-to forensic cue. A conversation around copy-move emerges. Two possible use cases for copy-move forgeries are to either cover something (e.g. an airplane in the sky, or a car by wood) or to emphasize something (e.g., a crowd of people). For low-texture content, block-based detectors work better, but they are quite expensive to compute. For high-texture content, keypoint-based descriptors work quite well. Aerial images are a good use case for copy-move forgeries, since there are fewer perspective constraints. It was pointed out that creating a large-scale copy-move dataset is a challenge: if done manually, it takes a large amount of time. If done automatically, it exhibits typically hard edges at the cuts which put the usefulness of the data into question. An experience is reported that one can splice semi-automatically foreground and background objects, and thereby create a larger dataset.

One challenge in the transition from academic research to practice is that in practice the priors are greatly skewed. In academia, classification tasks are often set up such that there is a 50-50 chance to be correct when guessing. In practice, for example in CSAM detection or steganalysis, the odds are skewed to a prior probability of 1 out of 10^6. Hence, even a low false positive rate overwhelms an analyst if she/he has to skim through all of these cases.

A short detour to video analysis. A case is reported where a Ph.D. student achieves better results on real data than on minimal, clean, academic data. It is acknowledged that this may be possible depending on the exploited cue. However, it may also be the case that the data preparation is just flawed, and a side channel is opened.

The discussion returns to the question on a characterization of forensic tools and their practical use. It is agreed among the group that forensic tools for proof in court are different from forensic tools to fight disinformation. Image reuse detection can be a good tool to fight disinformation.

The discussion then turns towards the broad family of detection or localization of synthetically generated visual content. First, how big is the actual thread from so-called DeepFakes? Maybe the actual threat vector is relatively narrow. A counter-example could be the Zelenskiy video ("drop your guns, surrender"), even though this was debunked after publication. However, variations of this task could bear realistic threat vectors in the future, for example to generate a synthetic image from a line of text. Hence detecting such synthetically generated data can be quite relevant. There are now also advanced possibilities for image retouching, e.g., by asking a model to replace a logo from a truck. With respect to the practical applicability, the networks are currently not good at creating interaction, e.g.: "draw a picture where Biden chokes Trump". From the perspective of image creation, it is better to take an image of someone choking another person, and to replace the two involved persons by Biden and Trump.

Finally, the discussion turns towards the role of deep learning in multimedia forensics research (this is an anticipation of the following seminar days). Deep learning papers are highly cited and are taking over many communities that would in principle also be interested in other approaches. For example, "traditional SIGGRAPH" people might also be happy about other methods, but deep learning dominates the conference. Deep learning also highly affects the funding landscape, and it is difficult to get a grant through without deep learning. Also, it impacts the culture of evaluation, in the sense that much more empirical comparisons are required and it is difficult to get something published without a demonstrated improvement over related work.

4.2 Day 1 – Traditional Methods in Forensics

Christian Riess (Universität Erlangen-Nürnberg, DE) – recorder of the session

Opening. Two statements are made to enter the conversation on traditional methods in forensics. First, a thought on traditional copy-move forgery detection (CMFD) algorithms is raised. They are tedious to parameterize, and a good strategy for overcoming that is unclear. However, it is appreciated that this is a classic, explainable image processing pipeline. Second, a thought on traditional methods versus deep learning methods is raised: it would be interesting to see hybrid approaches that make the best use of both paradigms.

Remarks on Explainability. An extended block of the discussion then focuses on explainability, which is oftentimes attributed as a key advantage of traditional forensic methods. The conversation is very lively, every seminar member contributes his or her perspective.

Why is explainability important? One could also do a controlled experiment to modify something and then check how good such a modification is detected, in order to convince for example a court of law of the workings of a method. It is a problem to base a court decision on an empirical evaluation without even any chance of understanding what is going on in the ML model. To illustrate this statement with an example from the US: parole decisions based on machine learning achieve the same performance as decisions that are reduced to three simple features: age, sex, and prior convictions [1]. However, the three features are much better understandable, and based on that understanding one can then discuss whether these features are a agreeable basis for the decision.

Hence, explainability is a critical asset in forensic investigations. It is noted that some classes of model-based methods are indeed well explainable, at least the main intuition behind them. Examples are physics-based geometric cues, like shadows and lighting conditions. One practical example from Brasil are video recordings that allegedly document a case of bribery. A forensic expert showed that there is a 1 in a million chance that the video is a forgery, otherwise it is real. This straw was used by the defendant, and only explainable systems can add further trust in the analysis.

There are several remarks that question the advantage of explainability in traditional forensic methods. It is noted that the claim that traditional forensic methods are inherently explainable comes with limitations, in particular when interacting with representatives from law enforcement without technical background. In this case, forensic cues that would otherwise be considered to be quite elementary from an information theoretic point of view, for example JPEG artifacts. This is even exacerbated in court, where the lawyer from the other party acts as an opponent. It happened in the past that expert witnesses failed to even explain linear interpolation in a satisfying way. On the other hand, law enforcement officers arguably also do not need to understand every detail of a method (who understands DNA analysis? Raise your hand!). From that point of view, input modifications and tracking of the change of output or an associated heatmap is the closest to the needs of the police. Hence, what you can explain to a non-technical audience is the ability of the tool, and the false positives, but you cannot explain the method itself. As a side note, judges then treat traditional methods and deep learning methods the same, since both are not explainable from their point of view. That doesn't negate the difference between traditional methods, whose functioning can be explained to suitably trained professionals, and deep learning tools, whose decision process is often obscure. However, the "level of obscurity" for AI-based methods differs with the type of task that AI fulfills. A binary classifier might indeed be unpredictable in its results. However, one could think of hybrid methods that use traditional elements and AI elements (e.g., AI for denoising the image, traditional methods for extracting hand-crafted features) which can be expected to satisfy these requirements very well. To conclude, it is important that our community develops more awareness to the other stakeholders (lawyers, judges) that are supposed to use our methods.

Regarding explainability in the context of the combination of traditional and deep learning techniques: a relatively easy scenario is when an image region is locally manipulated. In this case, deep learning and traditional methods can be cascaded. The deep learning approach can be used to find the relevant region, and traditional tools can be used in a manual analysis to verify this finding. The explainability comes in this case from the manual analysis. Such an approach is pursued in the analysis of fraud in scientific papers. Another option for combining traditional and AI methods is to use (AI-)learned filters and to re-inject them into a traditional method, e.g., by training a random forest.

It is noted that deep networks are also not entirely opaque. Instead, one can aim to get an impression about their behavior and some confidence that the correct functions are learned 1) by modifying the input and observing how the output behaves (e.g., noisier input should lead to less crisp results), 2) by backpropagating the decision scores to understand which parts of the input were most relevant for that decision, as it is done in gradCAM, and 3) by manually checking the learned filters. However, while this three-element list is not questioned per se, several participants note that these tools do not fit well to some multimedia forensics tasks. For example, heatmaps are usually not quantitative, oftentimes hard to interpret ("messy"), they are only useful for artifacts that coincide with certain locations in the image. For example, a sensor fingerprint (PRNU) can not really be visualized, so how can it be explained? Another example for a lack of possibilities for explainability is authorship attribution of a post at a social media platform. To make this example even more difficult, how can such an attribution be distinguished from a spoofing attempt?

Compression of the next Generation – an opportunity for traditional forensic methods? One remark is that HEIF images have not been forensically investigated. One problem here could be the lack of data. Researchers at Florence studied HEIF images and collected a small HEIF dataset. The analysis showed that, although the sensor pattern noise is still present on HEIF images, it is much more attenuated than in JPEG images, posing serious limitations to its effectiveness in realistic scenarios [2]. It is also noted that creating forgeries in HEIF data requires particularly high effort. On the other hand, it is not clear whether there is a forensic use case for such manipulations.

Standardization of Forensic Methods

Regarding generalizability, existing forensic methods are doing quite good on known attacks and known processing chains, but we fail on generalization of social network laundered data and unknown new generators. So, generalization, explainability, certification/standardization are central issues. If a method is standardized, then it does not need to be explainable anymore. For example, DNA testing is standardized because at some point in the past scientist have proven that it works. However, how could possibly a deep neural network be standardized? And what if someone then demonstrates an adversarial example attack on the network? Will this not immediately destroy the validity of the proof and destroy the standardization, because a judge sees two images that look the same, but they create different predictions? Against this concerns one can argue that in Western countries, certification is usually done for the operator of a method, not the method itself.

Evaluation of forensic methods: too far away from practice?

However, we are not quite in the situation to standardize methods, and one issue towards this lies in the evaluation.

One critique of traditional physics-based methods (e.g., methods that assess inconsistencies in shadows, perspective, or lighting) is that they only work in very controlled scenarios, whereas they are typically too constrained to be used in real world examples.

However, to be fair, this limitation to methods that work in lab environments is not only limited to physics-based methods. In practice, strong laundering of forensic traces happens when sharing images over social media. The sharing introduces recompression artifacts and geometrical modification on the uploaded visual content that degrades or erases the traces previously left by a manipulation, thus hindering the analysis. One specific example is that there is to our knowledge no paper on deep fake detection in the web.

In research evaluation we often make the simplifying assumption that we only need to decide whether one specific attack did occur or not. For example, we check for copy/paste, scaling or double compression. In real-world scenarios the challenge is more open. One often is tasked with stating whether *something* happened to the image, resulting in a manipulation of its perceived content. In practice, one strategy could be to run several detectors on the image, like double compression detection, inverse image search, stitching detection and more, and then to aggregate the different results in a graphical interface with an alert function that is fed back to the analyst who needs to decide about the evidence. The potential usefulness of such an approach is also reported in a TIP paper by Anderson Rocha's group. There, the output of many detectors is combined in a Bayesian way into a probability map. The success of this method may indicate that one needs multiple complementary methods.

Public Code

It is acknowledged that today more code is made publicly available than "back in the days" for traditional forensic methods. However, it would be good to have more efforts to collect code and to benchmark existing approaches. Meanwhile, code is available from various groups. There are also some benchmarks.

However, there is no grander community work to publish code. Biometrics has good practices by conducting challenges. In forensics, there was the 1st IEEE IFS Challenge, and there were some other minor events (ICASSP 2017, NIST/DARPA MFC, deepfake challenges). Maybe it would be good to do a challenge with a) synthetic generators e.g. based on stable diffusion, b) photoshop, and c) synthetic generators and photoshop. The evaluation should then be done in a way that the generators are not known.

Acceptance of Various Types of Evidence in Court

A generated piece of data, like a synthetic license plate, can not be accepted as proof at a court. However, this situation changes if an algorithm enhances an existing license plate, and an expert witness reads what he can decipher from this enhanced license plate. Besides the scenario of a court case, the second scenario is to read a license plate as an investigative cue. In this case, also machine learning classification is admissible (which would be impossible in court, due to the unknown error probability). As always, there are exceptions to this rule: in a case from the US, a person was sent to jail because his/her face was matched with a database, even though the person was innocent [4].

Then, it is discussed what national regulations exist for using a photograph or social media images in court. In a court case in the US, a social media picture was used to establish a link between the person and a gang. The photographer was asked whether the image/scene is real, and the photographer confirmed it. Hence, it does not necessarily need a technical method to authenticate images, there are also other ways. In Italy, it depends on a case-by-case basis. If the opposing lawyer does not challenge an image, then it should be admissible. Amped had a case where they challenged an image that was allegedly transported through WhatsApp, but in general, the judge can decide what is accepted as evidence.

From a technical point of view, it can be interesting to look into confidences for decisions. For example, a neural network can provide a confidence, and this can be a real benefit over traditional forensic methods, for example in super-resolution. This can also be a reason for revisiting AI methods in court cases: If I trust a network better than some flawed assumptions about a Gaussian distribution in a traditional forensic method, then it is probably better to use that network, argue why the method is better trusted, and provide its empirical accuracy to the court.

The threat of adversarial attacks should probably not be too much overstated for multimedia forensics. Adversarial attacks also exist for example for face detection. However, face recognition is a widely accepted technique, maybe because it is a visible cue. In our case, we are dealing with invisible cues, which could be the reason why it is more difficult for us to argue against adversarial attacks. However, in principle the threat assessment from adversarial attacks should in both cases be equal.

GoF versus DL Forensics. Traditional methods are maybe better suited for looking at one individual object, e.g., whether the shadow is fitting. However, in order to establish context between objects, then maybe machine learning methods can learn correlations that are otherwise inaccessible.

It is important to note that even our strongest traditional methods are limited in their generalizability in the field. For example, traditional CMFD detectors have a recall of about 20% on sufficiently difficult data (like scientific papers that are screened for fraud). This leaves a lot of work open.

From the perspective of a researcher: did we stop to do research in traditional methods because everything was done, or did we move to AI because we had no other choice due to the overall "AI wave"?

The rise of AI methods has also brought more datasets. Is there a way that we can benefit from these datasets with traditional methods? Arguably, the low-hanging fruits of traditional methods are taken, and the deep learning fruits were much easier to reach. For traditional methods, it could also be a selling point that the method only needs 10 images to calibrate, or that the method can generalize better than deep learning methods. But in any case, it is necessary to compare novel methods that follow the traditional paradigm also to AI methods. Such a comparison is difficult to do in a reasonable way, since traditional cues only pick up isolated aspects oftentimes, but nevertheless it has to be done.

References

1 Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning certifiably optimal rule lists. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 35–44, 2017.

2 Daniele Baracchi, Massimo Iuliani, Andrea G. Nencini, and Alessandro Piva. Facing image source attribution on iphone x. In Xianfeng Zhao, Yun-Qing Shi, Alessandro Piva, and Hyoung Joong Kim, editors, *Digital Forensics and Watermarking*, pages 196–207, Cham, 2021. Springer International Publishing.

3 Jonathan W. Hak. The admissibility of video and photographs posted to social media: Inconsistent court rulings. `https://tinyurl.com/yc3eymcc`, 2020.

4 K. Hill. Another arrest, and jail time, due to a bad facial recognition match. The New York Times, Dec. 29, 2020, available: https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html.

5 ICCV. 1st workshop on traditional computer vision in the age of deep learning (TradiCV). `https://sites.google.com/view/tradicv`, 2021.

6 Wikipedia. Phantom of Heilbronn: Example issues of DNA forensics analysis. `https://en.wikipedia.org/wiki/Phantom_of_Heilbronn`, 2022.

4.3 Day 2 – Discussion on the AI Act

Christian Riess (Universität Erlangen-Nürnberg, DE) – recorder of the session

Benedikt Lorch presents challenges for the use of AI in criminal investigations that arise from the draft Artificial Intelligence Act. The presentation is followed by a discussion. It is clarified that the AI Act aims at companies/providers of AI solutions, not on AI methods per se. For example, it is not a DeepFake detector that is 'high risk' per se, but instead it is the application of a DeepFake detector in a court case, where the fundamental rights of the defendant are at stake.

The following discussion touches a number of concerns. One concern is that GDPR is preventing research on faces, because all data/models that is created in a non-GDPR conforming way is tainted, and (strictly speaking) it can not be used for research. Another concern is that the AI act will create obstacles not only to companies using AI for commercial use, but also to researchers. All the more that it is not clear if the restrictions and obligations also extend to the models used as initial point for fine tuning and transfer learning. Another question that is raised is whether the transparency requirements for companies in the AI Act should also be extended to research? Stating the limitations of the system is a good practice in papers, but not everyone does it, and some people write pseudo justifications.

4.4 Day 2 – Deep Learning Based Methods

Irene Amerini (Sapienza University of Rome, IT) – recorder of the session

4.4.1 Christian Riess: Deep Learning in Multimedia Forensics

Christian Riess opens the session with a stimulating presentation on the advantages and challenges of deep learning methods in multimedia forensics. As a sidenote, Teddy Furon's WIFS 2021 keynote is mentioned which highlights the analogies between ML security and the typical goals in information forensics and security [2].

He cites three works:

- GAN fingerprint (Marra et al) depends on upsampling in GAN. However, this trace is easily removed by compression [3]
- Self-consistency (Efron et al), they didn't do any assumption on the kind of the attacks
- NoisePrint (Verdoliva et al) [1]
- Detection of out-of-distribution samples (cases in multimedia forensics of out-of-distribution samples are an huge amount)
  - Supervised approach calibration: needs another dataset
  - Bayesian methods that model weights as probability distributions
  - Bayesian approach

The talk ends with some final questions:

- Tangible benefits of DL?
- Are we just replacing models by dataset?
- Other interesting DL methods?

References

1 Davide Cozzolino and Luisa Verdoliva. NoisePrint: A CNN-Based Camera Model Fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2019.

2 Teddy Furon. WIFS 2021 keynote. https://www.youtube.com/watch?v=Gh2_tR-hgyU, 2021.

3 Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do GANs Leave Artificial Fingerprints? In *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 506–511. IEEE, 2019.

4.4.2 Deep learning – Discussion

Paul Rosin: There are also limitations in deep learning, and it is unsettling that the choice of architecture feels somewhat random: e.g. is tanh activation better than sigmoid? You don't know unless you try it out empirically.

Luca Cuccovillo: Neural networks should learn forensically useful properties. For example, features to describe the reverberation properties of the room in which the recording took place.

Martin Steinebach: Adding transformed input to a neural network, e.g., additional frequency information, really helps.

Christian Riess: Agnostic about the attack. I would like a method that generalizes.

Luca Cuccovillo: Algorithms for audio synthesis are meant to create speech which sounds plausible, and produced by the voices the network was trained upon – not to create audio meant to overcome a forensics analysis. Rather than looking *only* for synthesis traces, we should also look if the distributions of (meaningful) features inside the evidence about, e.g., speaker identity, recording device, room acoustics fit the allegation or not. If not, then something is off with the recording.

Paul Rosin: My experience with deep learning is that although the results are often good in general, if we look in detail there can be a lot of flaws. I found this when we had to compare our non-DL approach to colorization with competing DL approaches; the latter were not as good as I expected from an initial superficial view. Sometimes, in an attempt to achieve good results it seems that you rather then spend time on hand crafting features, instead you had craft loss functions. But it can be difficult to control the output of these deep learning models. In comparison, with the traditional methods, to do what you want is trivial.

Luca Cuccovillo: When you want to deal with a lot of complexity you should use deep learning to cover this complexity. This can be done directly – e.g., to perform end-to-end single/double encoding detection – or indirectly, – e.g. to perform microphone identification in presence of strong background noise, using a network to remove the noise while preserving the colour of the microphone.

Martin Steinbach: In detection CSAM or fake news deep learning methods are working. Manipulation detection is not working well with deep learning methods of the box. We added spectral transformation as a second channel to the input data and the performance improved a lot.

Paul Rosin: An interesting topic is neurosymbolic AI, which combines neural and symbolic AI in order to better capture prior information than purely using machine learning.

Benedikt Lorch: In the past few years, deep learning has been applied to almost any application in multimedia forensics. In light of all the success stories, little attention has been given to the limitations of deep learning. Only now are the failure cases of deep learning receiving increasing attention.

Isao Echizen: Benefit of DL, data. For a Deepfake detector for a company you should vary the dataset. Provide simple models to companies and companies improve the model, continuing to train the model. For rolling out a deep neural network in a company, then the data is often quite limited.

Thorsten Beck: What are the implications of the lack of sufficiently large datasets for the development of DeepLearning models and resulting tools? Are artificially generated datasets able to contribute to the development of effective tools?

Tiziano Bianchi: Deep Learning for analyzing robustness of deep learning, but not used a lot. Maybe one of the tools that we need is on the explanation of out-of-distribution samples.

Mauro Barni: My impression is that with DL we are just replacing models with datasets. The limits of model-based methods is that they cannot be used in the absence of good models and they cannot be used outside the precise limits used to build the models. The limit of data-driven methods, conversely, is that they cannot be used in the absence of representative and vast datasets, and they cannot be used with data which is not coherent with the datasets used for training. One may argue, though, that datasets are easier to build, while good models describing the complexity of real life may simply not exist. On the other hand, model-based methods seems to generalize a bit better to situations deviating from the models.

Mauro Barni: Maybe it is the right thing to replace models with datasets: if you want to describe real life then data are better than models. So maybe models are more robust. Confidence is the key. If I want to describe images why shouldn't I use as many images as possible as examples. Dataset mismatch: is the same as model mismatch.

Christian Riess: Unsupervised confidence measure. Bayesian neural network – I like the paradigm.

Marco Fontani: Dempster Shafer theory to measure the confidence measure [2]. Law enforcement 5% authentication, 95% enhancement (AI dangerous for enhancement). Paper by Boato and Pasquini [9]: more real than real (AI-generated images are considered more real than real images by humans). AMPED also published a paper where celebrities faces were upsampled with bicubic interpolation and with deep learning, and the recognition rate was not really affected [3]. With AI super-resolution, you create an average face, but real faces may contain strange artifacts (e.g., scars, moles) that the network tends to neglect; these artifacts are the most valuable for law enforcement when doing face recognition.

Irene Amerini: The work proposed by Mayer at al [12] is an interesting DL-based method. The authors introduce a digital image forensics approach called forensic similarity, which determines whether two image patches contain the same forensic trace or different forensic traces. The system is evaluated determining whether two image patches were captured by the same or different camera model and manipulated by the same or a different editing operation and the same or a different manipulation parameter, given a particular editing operation. Regarding Deepfake detection many different DL-methods exist in the literature. Those methods suffer from a number of shortcomings some of which are particularly relevant for their applications, so to say, in the wild, where strictly controlled laboratory conditions do not hold. Another point that should be addressed is the detection of Deepfakes in real-time such as recognizing the fake contents in a video-call on a device like a smartphone. For this purpose it is necessary to design models with low inference time and a small number of parameters, able to run on hardware with limited memory but able to recognize the fake with an high accuracy.

Alessandro Piva: Farid had another paper with similar results to Boato and Pasquini [13]. What are your experiences of Continual Learning?

Christian Riess: You add training data on the fly without going to catastrophic forgetting. Good paper but I don't know if I want to use it in forensics. It is autonomous in general you have a plan when you decide to retrain. So I think it is difficult to apply in real forensics scenario. It is used in network intrusion detection and in biometrics.

Lakshmanan Nataraj: We have a couple of papers on seam carving, most recently in the CVPR media forensics workshop. Our experience in deep learning methods in video forensics: training and test data should be the same; changing model changes a little bit the accuracy

Roberto Caldelli: In our experiments, deep learning works super great for specific tasks, but generalization and vulnerability to adversarial attacks are a problem. Ablations are important to understand the impact of certain design decisions. Confidences are also important. Input perturbations are fundamental to understand what is happening inside of the network.

Xianfang Sun: image segmentation, super resolution. Data hungry not only forensic application but also other areas. The results should be scalable. Weak supervised learning is not so popular in forensics.

Christian Riess: the community is sometimes a bit slow to absorb insights from the ML and vision communities. For example, we used shallow networks for a while. The vision community extensively explored self-supervised learning to mitigate the data bottleneck. This is probably something that we should be paying more attention to.

Law Ngai Fong: extract noise pattern Siamese network, forensic similarity, metrics

Roberto Caldelli: 1. When you do a good training, with a sufficient number of data, the performances that deep neural networks can achieve are amazing but what about generalization, black box scenario, adversarial point of view? All these kinds of issue should be put on the table and be analyzed. 2. Methods that look inside the box (looks for activations and so on). A paper, we gave at ICIP 2019 analyzes the confidence, the internal layers and tries to understand what it is inside the black box (explainability); it considers the point of view of adversarial.

Anderson Rocha: Bot classification can be done either by content with a language model (this is our community) or based on connectivity (which is done in the field of network analysis)

Anderson Rocha: Smartphone authentication with multimodal: image, video, and audio reflection(!) Fusion is an important topic in forensics, because sometimes one signal is not strong enough. Example: we record biosignals with smart watches, then do anxiety classification, because the person e.g. is sweating, heart rate is increasing, but the person is standing. This needs to be validated with medical insights.

Anderson Rocha: What do you think are the biggest challenge in cross-modal algorithm design?

Paul Rosin: Are there datasets available?

Anderson Rocha: Yes, for various tasks.

References

1 Thierry Denoeux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics*, 25(5):804–813, 1995.

2 Velan Erik, Marco Fontani, Sergio Carrato, Jerian Martino, et al. Does deep learning-based super-resolution help humans with face recognition? *Frontiers in Signal Processing*, 2, 2022.

3 Federica Lago, Cecilia Pasquini, Rainer Böhme, Hélène Dumont, Valérie Goffaux, and Giulia Boato. More real than real: A study on human visual perception of synthetic faces [applications corner]. *IEEE Signal Processing Magazine*, 39(1):109–116, 2021.

4 Sophie Nightingale and Hany Farid. Synthetic faces are more trustworthy than real faces. *Journal of Vision*, 22(14):3068–3068, 2022.

4.5 Day 2 – Cross-Media Approaches for Multimedia Forensics

Irene Amerini (Sapienza University of Rome, IT) – recorder of the session

Luca Cuccovillo: It is a problem to analyse image and audio over time.

Martin Steinbach: Lipsynch movement of lips compare to voice. Synthetic tools that do that are good. Can be applicable

Isao Echizen: detection of fake news, inconsistency

Martin Steinbach: disinformation detection, take image out of the original content and reuse it, image search and take the text

Luca Cuccovillo: Finding duplicates is a problem, figure out how to do cross in social media, Next media to consider: the metaverse

Paul Rosin: Could GPS tracking be considered a new media?

Thorsten Beck: What about using video codecs?

Christian Riess: You can use image forensic tools that analyze key frames.

Mauro Barni: It is pretty obvious that a video sequence provides more information than its single frames taken in isolation, yet the current state of the art in video forensics shows that in most cases Working at the frame level is enough to get very good accuracy. Problems like lack of generalization are not easily solved by passing from frames to video sequences. Of course, I am not saying that working at the sequence level does not provide any advantage, this looks more like a limitation of currently available techniques.

Possibly, temporal based analysis of deepfake based on LSTM is a little bit less more prone to adversarial attacks in terms of transferability [1], still I do not know if this small advantage makes temporal analysis worth.

Martin Steinbach: Fraunhofer study on a tool for fake news detection

Roberto Caldelli: We have studied how to improve source identification by using different sensors on-board of a smartphone (e.g. accelerometer, gyroscope). Not necessarily adding different media improve the identification.

Tiziano Bianchi: useful for disinformation detection: text

Irene Amerini: Multi-modal approach is useful in the context of fact-checking. The general idea is to do topic mining on tweets to identify facts, e.g., the first tweets about covid at the time when it was not yet well known what it was. So the goal is to work on a system that knows how to map tweets and the images associated with it into a multi-modal embedding in which images and text pertaining to the same facts are close to each other. Why is this useful? Imagine that we find a tweet about a new fact, but we do not have enough elements in the tweet to tell whether it is true or false. With this system we can do retrieval of all tweets similar to the one I am considering, and through these I can get more information about that fact. Another example on the use of different media is related to social media provenance where images and videos data are considered together. The main reason behind such choice is that collecting datasets large enough to train neural networks for the task has become difficult because of the privacy regulations that have been enacted in recent years. To mitigate this limitation, in [10] authors propose two different solutions based on transfer learning and multitask learning to determine whether a video has been uploaded from or downloaded to a specific social platform through the use of shared features with images trained on the same task. Moreover they introduce a model based on multitask learning, which learns from both tasks simultaneously.According to our knowledge, this is the first work that addresses the problem of social media platform identification of videos through the use of shared features.

Anderson Rocha: authorship attribution. Connectivity graph (Facebook), authorship (emoji are important)

Mauro Barni: Multi-modal approaches surely make sense yet it is important that the various modalities are fused properly to avoid inheriting the weaknesses of the various modes rather than their strengths.

Anderson Rocha: Cross-modality parental control in real time. Images, videos, caption, audio. Process them in real time to block the video. It is a classification problem but you don't have time coherence or series of classifier that we combine over time. How to combine different modality over time → fusion. Doing this real time with no deep learning in our case. Sometimes audio says one thing, but the image says something different.

Paul Rosin: Data fusion is a common topic in computer vision, and there are many different approaches. Perhaps we can use some of these in forensics.

Anderson Rocha: For recent papers this is true. Jointly optimizing different modalities. Early fusion or decision fusion if you don't have a network. Which one is better depends.

Irene Amerini: Most of the methods for Deepfake detection rely on extracting salient features from RGB images to detect through a binary classifier if the image is fake or real. In [11] is proposed DepthFake, a study on how to improve classical RGB-based approaches with depth-maps. The depth information is extracted from RGB images with monocular depth estimation techniques. Using multi-modal information can help increase the performance of the detectors and in generalization capacity of these features with respect to deepfake generation techniques that have not been seen in training.

Anderson Rocha: How to combine temporal information. One of the challenges when dealing with multiple detectors across time is how to combine the different responses overtime so that temporal information is incorporated. This was, for instance, discussed in the paper "Multimodal data fusion for sensitive scene localization"[10] in which the authors propose a novel multimodal fusion approach to sensitive scene localization. The solution can be applied to diverse types of sensitive content, without the need for step modifications. Such solutions are key to deal with the ever-changing scenario of forensics in which actors keep proposing new ways of defeating detectors.

Alessandro Piva: Our experience on data fusion concerns the exploitation of both content-based features and file structure-based features for the identification of the source of video content (e.g. which brand of the source device, or in which social network was the content uploaded). The idea is to extend the work to exploit both audio and video features.

Anderson Rocha: Fusion is important and promising path in forensics. We are using smartwatches to capture biosignals. With different data we are able to understand what it is doing.

Anderson Rocha: What prevents you using a cross modality?

Paul Rosin: Lack of datasets.

Mauro Barni: Video and audio lip synchronization is quite popular, still frame by frame analysis seems to work better.

Anderson Rocha: This is a dataset bias

Benedetta Tondi: Maybe we need a bigger dataset.

Anderson Rocha: Generalization could help solve working with more modalities.

Mauro Barni: Maybe the current networks do not exploit well the availability of more than one single modality. For sure we need larger datasets, which are not easy to build in the multimodal case.

Anderson Rocha: Example of a work by Christian Riess on his PhD on reflectance for forgery detection. So do not exploit well the availability of more than one single modality. For sure it is important to transform the input.

Mauro Barni: I really think the way to go is to fuse results from GOF and AI-based methods.

Alessandro Piva: Continual learning: we investigate the potential of continual learning techniques to build an extensible social network identification neural network where multiple new tasks, each one comprising multiple new social platforms, are considered, in order to simulate the possibility that new social media can appear.

Marco Fontani: Reproducibility of the methods found in the literature is often impossible. The results are quite different. A problem can be a different dataset.

Benedetta Tondi: We should do all make efforts to release the code including the trained models, and also all the instructions for methods' training. Without that, reproducing results turns out to be a hard task in deep learning. Also, research advances so fast that we need to be able to run comparisons in a fast way.

Anderson Rocha: and if you can publish because of that?

Anderson Rocha: In video you cannot do cross-validation and this often happens.

Mauro Barni: Often the problem is the way you test your algorithm.

Benedetta Tondi: It helps in reproducibility (for company and for us to compare our results).

Anderson Rocha: In a Nature paper [14] they analyze 62 algorithm COVID Xray image detection. None working! When you submit a paper reviewer ask to compare with arXiv

Christian Riess: What do you do?

Mauro Barni: If the AE is not responding or insist that you should consider arXiv papers as state of the art, then you should talk to the EIC. IEEE, for instance, has a clear policy stating that arXiv papers CANNOT be considered state of the art and asking a comparison against arXiv papers is not allowed.

Anderson Rocha: And you have to compare with published papers not arxiv!

References

1 Dongdong Lin, Benedetta Tondi, Bin Li, Mauro Barni, Exploiting temporal information to prevent the transferability of adversarial examples against deep fake detectors. *2022 IEEE International Joint Conference on Biometrics (IJCB).*

2 Thierry Denoeux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics*, 25(5):804–813, 1995.

3 Velan Erik, Fontani Marco, Sergio Carrato, Jerian Martino, et al. Does deep learning-based super-resolution help humans with face recognition? *Frontiers in Signal Processing*, 2, 2022.

4 Marco Fontani, Enrique Argones-Rúa, Carmela Troncoso, and Mauro Barni. The watchful forensic analyst: Multi-clue information fusion with background knowledge. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 120–125. IEEE, 2013.

5 Information Processing Fraunhofer Institute for Communication and Ergonomics. Software that can automatically detect fake news. `https://www.fraunhofer.de/en/press/research-news/2019/february/software-that-can-automatically-detect-fake-news.html`, 2019.

6 Chandrakanth Gudavalli, Erik Rosten, Lakshmanan Nataraj, Shivkumar Chandrasekaran, and BS Manjunath. SeeTheSeams: Localized detection of seam carving based image forgery in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–11, 2022.

7 Oren Halvani and Philipp Marquardt. An unsophisticated neural bots and gender profiling system. In *CLEF (Working Notes)*, 2019.

8 Muhammad Rifki Kurniawan. Catastrophic forgetting in neural networks explained. `https://mrifkikurniawan.github.io/blog-posts/Catastrophic_Forgetting/`, 2021.

9 Federica Lago, Cecilia Pasquini, Rainer Böhme, Hélène Dumont, Valérie Goffaux, and Giulia Boato. More real than real: A study on human visual perception of synthetic faces [applications corner]. *IEEE Signal Processing Magazine*, 39(1):109–116, 2021.

10 Luca Maiano, Irene Amerini, Lorenzo Ricciardi Celsi, and Aris Anagnostopoulos. Identification of social-media platform of videos through the use of shared features. *Journal of Imaging*, 7(8), 2021.

11 Luca Maiano, Lorenzo Papa, Ketbjano Vocaj, and Irene Amerini. Depthfake: a depth-based strategy for detecting deepfake videos, 2022.

12 Owen Mayer and Matthew C Stamm. Forensic similarity for digital images. *IEEE Transactions on Information Forensics and Security*, 15:1331–1346, 2019.

13 Sophie Nightingale and Hany Farid. Synthetic faces are more trustworthy than real faces. *Journal of Vision*, 22(14):3068–3068, 2022.

14 Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3(3):199–217, 2021.

15 Inna Vogel and Peter Jiang. Bot and gender identification in twitter using word and character n-grams. In *CLEF (Working Notes)*, 2019.

16 Kyra Wittorf, Martin Steinebach, and Huajian Liu. Automated image metadata verification. *Electronic Imaging*, 33:1–6, 2021.

4.6 Day 3 – Big Data Challenges

Tiziano Bianchi (Polytechnic University of Turin, IT) – recorder of the session

The participants were asked to report their experience with big data.

Tiziano Bianchi: Large scale PRNU search, dataset of about 25 million images. The problem is that for this kind of data search does not scale sub-linearly (e.g, log) with the size of dataset. Data is noise-like, standard indexing techniques (e.g LSH) are unstable.

Isao Echizen: Problems with bias on datasets. Construction of large dataset by starting with reference dataset and doing preprocessing and augmentation. Promote construction of large datasets involving more communities.

Luca Cuccovillo: Experience with speech matching (a sort of specialized Shazam). Problem with scalability, e.g., the need to replicate pairwise correlations for aggregation of similar speech. Problem mainly related to engineering, e.g., how to design short fingerprint with enough quality. One challenge is doing audio phylogeny, i.e, finding relation graphs of audio signals. Including synthetic audio. Need of collaboration, common understanding. Need of many different tools for dataset generation, different community should provide them to have scalability, single institutions cannot do this. Some datasets in challenges may have biases (e.g ASVspoof) [3]

Martin Steinebach: Working with real datasets has many issues not found in scientific research (transcoding, etc.). Research does not often consider efficiency on large scale.

Luca Cuccovillo: Agrees on additional engineering for managing speed required for big data.

Irene Amerini: Dealing with big data is a huge challenges due to many issues in order to have access to it. One of them is the time needed to collect dataset since big datasets are not always already available and, secondly, the data storage if you are a small institution. Furthermore all of the collected data need to be filtered and pre-processed in order to be used. Multimodal is even a bigger problem if should scale to big data.

Paul Rosin: A challenge with 3D datasets can be the large amount of data required to be stored. The feasibility of large scale digitization has been demonstrated for museum artifacts where companies have captured millions of images. As part of a project we were working on automatic segmentation, and needed to manually segment a large amount of data ($>$ 1000 images) for ground truth.

Mauro Barni: There is a lack of real big data due to problems in gathering them.

I have two experiences in this sense.

In a first case we were trying to develop a print and scan attack against a detector of synthetic images. The original detector does not work after print and scan, so we had to retrain it on printed and scanned images. To do that we had to build a dataset of printed and scanned images. The required effort was huge, and were able to get *only* 20,000 images, obtained with one single printer and one single scanner. Generalization to other devices was out of reach due to the lack of equipment. As a result, at test time the detector does not work well with different printers and scanners. [4]

As a second example, we collected 4 million outdoor images to classify geographic provenance (country recognition) by relying on the cultural features of urban architecture, social habits, etc ... The country was determined from the GPS position of the image. We got a huge improvement when the dataset grew from 0.5M to 4M images. Yet, gathering the images was not easy. We had to filter the images based on their content, to retain only urban scenes, remove persons etc ... We also had to ensure enough diversity gathering images from more source, including, street view, Flickr and Mapillary). Diversity and representativeness are big problems in large-scale image collection. For example, in our dataset there are many more images from the US & Europe than from some small countries (e.g. in Africa). We tried to solve this bias by balancing the dataset, i.e., building macro classes with the same size, but in this way our classifier was less discriminative. We also tried by weighting underrepresented countries, however the overall accuracy decreased. [1]

Similar problems are surely present in other application domains. For instance, how can we gather face images from small ethnic groups?

Benedikt Lorch: We gained practical experience with big data on an image retrieval task where the image database was growing every day. The concern was that the search would slow down with size of the database. We were able to address this concern with an approximate search method. Looking ahead, larger machine learning models create a demand for larger datasets. However, it becomes increasingly difficult to screen larger datasets and assert data quality properties, which is also required by the Artificial Intelligence Act. To this end, an interesting direction for future research would be quality metrics for datasets and automatic methods to assess data quality.

Luca Cuccovillo: An example of dataset quality assessment is to classify degraded training data, to get what most representative data are. [9, 13]

Christian Riess: experience in building dataset for image superresolution. One problem is the exponential number of combinations of parameters in dataset, to be done manually. [10]

License plate recognition project with police, mix of real and synthetic to ease annotation. Use of augmentation, and post-processing. Built a rack of different cameras to automate acquisition of renderings on screen. Real acquisition of cars can be done but is very time consuming (700 labeled so far). Difficult to add realistic features like weather effects, lighting, etc. [16]

Thorsten Beck: Mentions dataset of images from retracted articles from Elsevier. Annotation is manually done from article retraction notices. The dataset is not really large. Compiling such a dataset comes with significant legal challenges, e.g. when results are published (e.g., only for research use). The dataset does not cover all forms of manipulation, consequently representativity is another issue, larger collection of images exist only for few categories (duplicates). The dataset comprises of multiple kinds of manipulations (since it is build of real-world data). Automatically generated manipulation datasets (e.g., copy-move) are not very realistic. Getting enough publishers to the table is a demanding task, since they are not necessarily ready to invest resources and man power. Still, the problem of inacceptable image manipulation in scholarly works will hardly be resolved without a contribution from the side of the publishers. [17]

Roberto Caldelli: not much experience on big data for forensics. We have been gathering data for testing image provenance from social media (Facebook, Twitter) and we developed automatic tools for crawling and downloading. Problems are the interaction with social network API, which can be time consuming and complying with their policies for gathering data. We also experimented with a copy-move detection tool on print and scan images by testing with different devices. A comment is that limited availability of very big datasets for everybody sometimes makes research less democratic.

Lakshmanan Nataraj: Detection of GAN generated images. Collection of datasets from different GAN tools (6 types of GAN). Millions of images. Classification of GAN types. [5, 11]

Alessandro Piva: In PRNU estimation for video there is the need to process multiple frames, but this process is hindered by the presence of video stabilization, requiring the synchronization of ech frame. No efficient methods found in the state of the art when research was done. Managing crops, resize, rotation. Analyzing large datasets of videos requires huge computational effort. Experience in building dataset (VISION, multimedia forensic challenge), one of the problems is organizing the dataset before starting the collection of data. For recent datasets this is complicated by multiple acquisition settings and resolutions available. Usually only few settings/resolution are considered. For some published datasets there is not enough information on video settings used during acquisition, or inconsistent setting were adopted. Care must be taken on these aspects when building new datasets, such that in our opinion a single research group is not enough for the task.

Marco Fontani: Our products are for case works (mainly police), not many big data cases. We've been testing automatic analysis of images for insurance companies, there's a problem with the complexity of real cases (acquisition pipeline, etc.), and unclear definitions of authenticity in some scenarios. In video surveillance, a large amount of data is collected and must be stored for possible later use as evidence. Some storage and evidence management systems do not preserve the integrity of data (e.g., they systematically use transcoding of the original footage); this is a problem for forensics. Also, it is expensive to use commercial storage systems. Some police forces try to revert to local storage lately.

Benedetta Tondi: country recognition task, joint work with Mauro. Satellite images, and the detection of manipulated satellite images. Problem of datasets of satellite images, especially large scale datasets. Different sources are different domains. Tools trained on one sensor do not generalize to other sensor (e.g., Google Maps, other satellites). Need to include images from multiple sensors in the dataset.

Anderson Rocha: scientific retraction papers (DARPA 2017). 5000 papers with retraction notes. System receives pdf and extract images from pdf. Analyze images for forgery. Compare all images from papers of same authors. Analysis of images in suspicious scientific papers. Compare all papers from Scholar profile of authors, build a graph with similarities. The system produces a report to help human expert. No automatic decision should be allowed according to rules. Library to create copy move and forgery with different tools, to generate data for training. Completely annotated since synthetically produced. Freely available. (DARPA semaphore project). Only can detect about 20-25% of forgery right now. (Papers in biology and medicine). Detector should be improved. Right now only images are used, no content or text from papers. [12]

Detection of pornographic images/videos. 200 hours of pornography in dataset, problems with authorization from University for storing them. For illegal material (child pornography) training should be done on virtual machine by police. Multiple levels of training: Imagenet, fine tuning on generic pornography, fine tuning on police virtual machine for child pornography. Should have very low false positives. 40000 child pornography cases in police dataset. 40000 normal (including non pornography and regular pornography). 35% detection, less than 5% FP. No decision, only filtering. Usually run on suspect's harddrive, the tool gives the most likely files, manual inspection is required. You should reduce the number of hours used by manual expert inspection, so low FP is required in this application. One video is enough for prosecution, so even if few videos are detected over the total is perfectly fine. Right now we are collecting everything form social media (whatsapp, telegram, tiktok, facebook, twitter) on attacks in Brasil. Billions of data, most is garbage, should be filtered.

Mauro Barni: What are you looking for?

Anderson Rocha: To localize faces and identify spreaders, i.e.,most frequent faces seen in videos. Collecting related text.

Roberto Caldelli: what kind of real images did you use during training for pornography detection? Common images such as objects, landscapes, cars and so on, or did you select specific cases of presence of normal nudity?

Anderson Rocha: We selected difficult cases, for example we use images selected by skin detector (beach, swimming pools, etc.).

Then, the discussion turned on discussing challenges and opportunities offered by big data. The following challenges are identified:

- Copyright
- How to manage storage requirements
- How to distinguish what is useful in collected data
- How to generate synthetic data
- How to guarantee diversity and representativity.
- Computational power to collect all required data.
- Versioning.
- For university is difficult to have storage and computation capabilities.
- Problems of privacy when collecting some kind of data (e.g. faces).

Then discussion follows:

Mauro Barni: You get outstanding performance if and only if you have enough data. You cannot use AI without enough data. Someone claims that with big data and enough computational power you can explain everything? I do not quite agree with this view, understanding is more than just finding patterns in data.

Anderson Rocha: Most of the correlations are spurious, but how to separate useful from spurious? There are three levels for acquiring knowledge:

1. Find correlations, machines are very good at this
2. Find possibilities, like cause – effect relations, AI is usually bad at this
3. Analyse past decisions, project alternative future based on different choices, machines cannot do this.

Marco Fontani: Are machines accountable? Who is accountable? Producers will say this is just a help for human decision, the expert operating the system should be accountable.

Benedetta Tondi: A theoretical challenge is represented by the security of networks in an adversarial environment. Data should be representative of possible attacks. This turns out to be a very big challenge for forensic tools.

Mauro Barni: With big data it is easier to hide poisoned samples, and more difficult to spot them.

Anderson Rocha: Attacks exploiting triggers. How can we inspect a network to see whether we have a backdoor. This can be a forensic problem.

Mauro Barni: A possible solution is to inspect the datasets used during training, not only the trained network. Attacks can be carried out at different levels. Backdoor can be used also to watermark a network. We have a good experience in checking datasets. In [7] we used cluster analysis in the latent space to detect poisoned samples.

Martin Steinebach: This is important for autonomous driving. Training robust classifier. More machine learning security. But also forensic if you analyze the dataset for anomalies.

Luca Cuccovillo: Opportunities of federated learning in big data (privacy, complexity, but also vulnerabilities to attacks).

4.7 Day 3 – Benchmark and Performance Evaluation

Tiziano Bianchi (Polytechnic University of Turin, IT) – recorder of the session

Anderson Rocha: There is a need of a validation protocol, for comparisons among different tools. Problems to be solved are how to access to data, how to choose test and training data, how to choose proper metrics.

Martin Steinebach: Huge datasets sometimes are prone to overfitting, if not diverse enough. A black-box evaluation protocol could be more fair. Give a blind test set to prevent overfitting over it.

Anderson Rocha: : We need to be responsible for this black box.

Martin Steinebach: A public body could be the standardization body. Blind virtual machine for evaluation of security of tools, including AI tools.

Paul Rosin: What about the feedback, will this be useful for improving tools?

Marco Fontani: The main issue for the practitioner is explainability. Heat maps are often not enough.

Paul Rosin: When benchmarking for image standardization, often a benchmark dataset is used both for training and testing, with a random split. It is better to use a separate benchmark dataset for only testing purposes. Collecting different data for training can be left up to the developer. I advocate a structured benchmark where different levels of difficulty are provided in the testing set (depends on application) A small testing dataset can be more curated. [15]

Anderson Rocha: : We have good dataset for deepfake detection, however the performance when performing intra-dataset evaluation is saturated. This is observed also for spoofing detection and copy-move detection. Cross-dataset evaluation is needed as next step. Difficult to have different levels for testing in forensics.

Marco Fontani: confirm the experience of cross-dataset evaluation, performance drops in this case.

Paul Rosin: One issue is the diversity of types of images in forensic datasets.

Anderson Rocha: most of datasets include natural images. Experience with separation of specific images (biomedical) from natural during dataset preparation.

Mauro Barni: Most work is done fine tuning network trained on natural images. In some fields, e.g. GAN generated images, few architectures are available. Better cross-validation by training on images produced by one architecture and testing on another one is needed.

Anderson Rocha: : there is a shift from the real-fake detection problem to fingerprinting of GAN generation algorithm. Maybe in the future we will shift to fingerprinting, which is a more challenging problem and requires training on all available tools.

Paul Rosin: I recently came across ForgeryNet dataset for benchmarking [8], which contains a lot of data: 3 millions images, 200000 videos. This dataset should be considered a useful resource.

References

1 Omran Alamayreh, Giovanna Maria Dimitri, Jun Wang, Benedetta Tondi, and Mauro Barni. Which country is this picture from? New data and methods for DNN-based country recognition. *arXiv preprint arXiv:2209.02429*, 2022.

2 João P Cardenuto and Anderson Rocha. Benchmarking scientific image forgery detectors. *Science and Engineering Ethics*, 28(4):35, 2022.

3 Luca Cuccovillo, Christoforos Papastergiopoulos, Anastasios Vafeiadis, Artem Yaroshchuk, Patrick Aichroth, Konstantinos Votis, and Dimitrios Tzovaras. Open challenges in synthetic speech detection. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2022.

4 Anselmo Ferreira, Ehsan Nowroozi, and Mauro Barni. VIPPrint: A large scale dataset of printed and scanned images for synthetic face images detection and source linking. *arXiv preprint arXiv:2102.06792*, 2021.

5 Michael Goebel, Lakshmanan Nataraj, Tejaswi Nanjundaswamy, Tajuddin Manhar Mohammed, Shivkumar Chandrasekaran, and BS Manjunath. Detection, attribution and localization of GAN generated images. *arXiv preprint arXiv:2007.10466*, 2020.

6 Chandrakanth Gudavalli, Erik Rosten, Lakshmanan Nataraj, Shivkumar Chandrasekaran, and BS Manjunath. SeeTheSeams: Localized detection of seam carving based image forgery in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–11, 2022.

7 Wei Guo, Benedetta Tondi, and Mauro Barni. A master key backdoor for universal impersonation attack against DNN-based face verification. *Pattern Recognition Letters*, 144:61–67, 2021.

8 Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. ForgeryNet: A versatile benchmark for comprehensive forgery analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4360–4369, 2021.

9 Zohaib Amjad Khan, Giuseppe Valenzise, Aladine Chetouani, and Frédéric Dufaux. Towards an image utility assessment framework for machine perception. In *30th European Signal Processing Conference (EUSIPCO)*, pages 568–572. IEEE, 2022.

10 Thomas Köhler, Michel Bätz, Farzad Naderi, André Kaup, Andreas Maier, and Christian Riess. Toward bridging the simulated-to-real gap: Benchmarking super-resolution on real data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2944–2959, 2019.

11 Tajuddin Manhar Mohammed, Jason Bunk, Lakshmanan Nataraj, Jawadul H Bappy, Arjuna Flenner, BS Manjunath, Shivkumar Chandrasekaran, Amit K Roy-Chowdhury, and Lawrence Peterson. Boosting image forgery detection using resampling features and copy-move analysis. *arXiv preprint arXiv:1802.03154*, 2018.

12 Daniel Moreira, João Phillipe Cardenuto, Ruiting Shao, Sriram Baireddy, Davide Cozzolino, Diego Gragnaniello, Wael Abd-Almageed, Paolo Bestagini, Stefano Tubaro, Anderson Rocha, et al. SILA: A system for scientific image analysis. *Scientific Reports*, 12(1):18306, 2022.

13 Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.

14 Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Hachette UK, 2018.

15 Paul L Rosin, Yu-Kun Lai, David Mould, Ran Yi, Itamar Berger, Lars Doyle, Seungyong Lee, Chuan Li, Yong-Jin Liu, Amir Semmo, et al. NPRportrait 1.0: A three-level benchmark for non-photorealistic rendering of portraits. *Computational Visual Media*, 8(3):445–465, 2022.

16 Andreas Spruck, Maximilane Gruber, Anatol Maier, Denise Moussa, Jürgen Seiler, Christian Riess, and André Kaup. Synthesizing annotated image and video data using a rendering-based pipeline for improved license plate recognition. *arXiv preprint arXiv:2209.14448*, 2022.

17 Humboldt-Universität zu Berlin. Image integrity database. `https://rs.cms.hu-berlin.de/iidb/pages/home.php`, 2023.

4.8 Day 4 – Morning Discussion

Benedikt Lorch (Universität Innsbruck, AT) – recorder of the session

Luca Cuccovillo presents open challenges in synthetic speech detection based on his talk from IEEE WIFS 2022 [1]. The goal of the WIFS paper was to review limitations of current datasets and discuss requirements for good synthetic speech datasets.

Neural speech synthesis: Ground-breaking applications vs. unprecedented forms of misuse

Synthetic speech detection:

- Potential: Plenty of room for research and development
- Danger: Lack of common planning/directions
  - Unclear technical requirements for datasets
  - No interpretability of model outputs
  - Lack of robustness/generalization
  - Lack of exchange between research and potential end users

Datasets: Large number of datasets available, but all of them have problems: Undisclosed synthesis algorithms, synthesized voices do not have real counterparts (speaker recognition would solve the task), single female speaker, not redistributable in original/derivative form, single text-to-speech pipeline

Detection algorithms: Many excellent proposals with hand-crafted features and deep neural networks. But they also have some issues: Unseen synthesis methods are problematic, unseen speaker/recording conditions, methods based on flawed dataset, lack of interpretability and explainability, unclear functional/non-functional experiments

How to do data collection right?

- Curating the data: Balance the speakers, gender, age, languages, accents
- Has to have transcriptions, enough data for training/fine-tuning
- Adhere to legal constraints.

Requirements for the creation of synthetic data:

- High linguistic and voice variability
- Diverse vocoding qualities
- Diverse feature extraction qualities
- Maximum expressiveness

Efforts and costs should be shared:

- Data collection and storage requirements
- What about federated learning (FL), leaving data on-premise?
- Is federated learning feasible for non-IID audio data?

Explainability is more than nice-to-have:

- Right of explanation prompted by the EU
- Current AI Act proposal considers forensic algorithms "high-risk"
- Journalists and forensic analysts have strong demand for explainability

- Question: Should we rely on XAI methods from image domain, or go further?
- Question: Are saliency maps on spectrograms understandable to end users, or only to researchers? Useful as debugging tools but not really explainable

Discussion: How many of these challenges are related to synthetic image detection?

- Image datasets can also contain biases
- Difficult not to inject any side channels in speech
- Possession asymmetry: A few companies possess the most amount of speech data, which gives them an advantage. In speech, this asymmetry arose earlier than in vision.
- The general problems are the same across application domains: dataset diversity, dataset size, explainability. The way these problems manifest themselves are different, calling for different mitigation strategies. Visualization maps can be more difficult to interpret for audio. In other words, audio and image forensics share the same general problems, but solutions can be very different.

Mauro Barni: Research in AI (and AI-based forensics) proceeds in a chaotic way. Everyone is somehow steered by their own goals. But we can do small things to advance our field: serving on the editorial board of a journal allows you to some extent direct the community. Similarly, competition steers the communities for the next years.

Martin Steinebach: There are many parallel, duplicate efforts, just using other taxonomies and not knowing about each other.

Mauro Barni: The newly proposed AI-based watermarking methods are rarely compared to the traditional watermarking techniques. Yet, classical watermarking provides satisfactory, sometimes excellent, solutions to many problem, so a comparison would be really needed.

4.9 Day 4 – Current and Future Applications

Benedikt Lorch (Universität Innsbruck, AT) – recorder of the session

There were eight short talks on current and future applications.

4.9.1 Marco Fontani about Amped Software

About Amped software:

- Mission: Provide customers with reliable algorithms based on scientific papers
- More than 100 users around the world
- Quest to provide good support, provide a complete product, forensically sound, widely adopted and accepted worldwide, deeply involved in the scientific community

Amped ecosystem:

- Amped Five, the top tool with all filters (Swiss knife)
- Amped Replay (simplified Swiss knife): an advanced player with streamlined processing and basic enhancement
- DvrConv: CCTV systems use proprietary video formats, and this software allows batch conversion of such formats in a forensically sound manner
- Amped Authenticate: Authenticate images; since recently Authenticate includes a DL detector for deepfake detection, but with all the necessary warning messages.

Survey on video forensics state of the art based on user survey [2]

- Main issues: Low image quality, proprietary CCTV/DVR video files, amount of cases/data, interpretation of video evidence; budget is not an issue
- Increase of video casework in recent years
- Increase of crime, change on image and video quality, pandemic made an increase of casework
- Evidence used to solve a crime: CCTV, mobile device data, images and videos from other sources
- Training: Vendor training, self-learning, job training,...

Should AI be used in forensics:

- Only 2% say "avoid AI"
- Majority said the use of AI should be limited to cases when proven reliable
- A good percentage said "to be used for investigative leads only"

Question: What tool would Amped like to develop?

- Users want Amped FIVE to be faster
- Functionality: Image enhancement, e.g., improve denoising
- Authenticate: Need for video authentication tools; users request tools for deepfake detection, although they are to date not very relevant in practice yet

Amped also contributed to the ENFSI best practice guidelines for audio authentication [9] and image authentication [8].

4.9.2 Isao Echizen: Fake media detection and its practical application

- Examples of where fake faces have been seen recently
- Five types of face synthesis methods
- Detection approaches: MesoNet, Capsule network, joint facial video detection and segmentation
- AIaaS for automatic detection of fake facial videos: AI-based web service accessible via web API
- License status of SYNTHETIQ VISION: Will be used by several companies, including CyberAgent Inc (advertising company in Japan).

Common issues for fake media detection:

- Performance degrades when images are redistributed via social networks (item Detection of unseen types of fakes. Periodic updating of training data and model training are necessary
- Users do not necessarily need a generic detector. Accurate detection of a specific kind of fake media is acceptable, e.g. digital twin: faceswap / eKYC: facial reenactment

Question about compliance with AI regulations: Are there similar restrictions and laws about privacy in Japan as in Europe? Data comes from companies.

Question: What is the most important face synthesis technique to detect for companies? Facial reenactment used in KYC.

4.9.3 Martin Steinebach: KIKU: Utilizing AI for the protection of cultural property

KIKU = Künstliche Intelligenz für den Kulturgutschutz (Artificial intelligence for the protection of cultural assets)

The project KIKu[2] (for a video demo please see [11]) is a follow-up project of the BMBF project Illicid[3]. In Illicid, various technical methods for the protection of cultural assets were developed, including a machine-learning based app that classifies robbery excavations and thus helps, for example, customs officers to detect illegal imports. This part was considered so relevant by users that KIKu was designed, with several stages to further develop the application. The core continues to be the detection of robbery excavations through machine learning [6] [4]. Deep learning will be used both to classify and to recognize similar objects. However, the project will also address issues such as the detection of forgeries of cultural goods.

The project is relevant to security because illegal excavations and lootings serve, among other things, to finance terrorist groups. The excavation sites are occupied, cultural goods are looted and then smuggled to third countries, where they are sold. The proceeds then flow back to the terrorists. To identify artifacts that come from looted excavations, the knowledge of experts is necessary. But these are not available where the objects are brought across the border or offered for sale. It is not possible for the customs officers to verify the information on the objects, for example regarding origin or age. This is where the KIKu tool comes in: Items assessed by experts become training data with images and metadata. The trained

[2] https://www.sit.fraunhofer.de/de/kiku/

[3] https://www.sifo.de/sifo/de/projekte/schutz-vor-kriminalitaet-und-terrorismus/schutz-vor-organisierter-kriminalitaet/illicid/illicid-verfahren-zur-erhellun-beispiel-antiker-kulturgueter.html

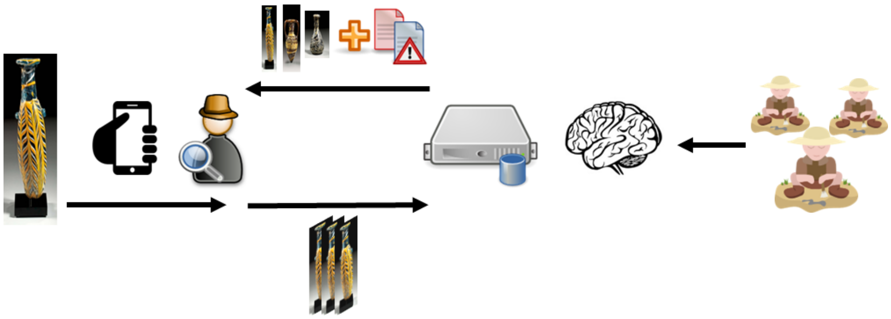


Figure 1 KIKu workflow.

network can then be accessed with an app from a smartphone with a photo of an item to be examined. The customs officer can now compare the information on the object with KIKu's assessment and take further steps in the event of discrepancies (see figure 1).

As common in machine learing, training data is an important issue. In the first project Illicid, 2-3 archaeologists provided 3,000 hand-labelled datasets, which did show promising results. The strategy here was not to aim for a generic recognition of cultural goods, but learning only items of a narrow area and epoch. In KIKu, crawling of museum data was utilized, with currently 140,000 training sets available, increasing the potential of generalization.

For cultural good recognition there are already applications in Poland for recognizing stolen paintings. However, the KIKU project also includes similar paintings and other objects. The core goal is classification and not re-identification.

Discussion about collecting more images from museums, whether users can prefer texture or shape features, whether the network prefers any particular features

Discussion about maturity of technology: Retrieval tasks seem to work quite well, checking constraints for pixel-level differences is error prone to do at scale in practice

4.9.4 Lakshmanan Nataraj: Current and future applications in media forensics

- Seam carving and seam insertion
- Seam carving detection with a CNN
- Object removal examples with and without heatmaps
- Satellite image object removal with heatmaps [12]
- Seam carving for object displacements
- Potential future applications:
  - Satellite image forensics
  - Different domains: image, video, audio, metaverse, NERFs, diffusion, etc.

Discussion how to do object displacement using seam carving.

4.10 Day 4 – Challenges Ahead

Benedikt Lorch (Universität Innsbruck, AT) – recorder of the session

4.10.1 Jane Wang: Convergence of signal processing to machine learning

- Signal processing (SP) and image processing (IP) plays a key role in the *preprocessing and transformation and feature extraction*, before the DL design. SP-based processing is critical in digital media security and forensics research.
- Relationship between SP/IP operations and DL components
- The SP/ML boundary is getting blurred

Future: signal/model-driven DL

- Challenges: data-driven (lack of generalizibility in out-of-sample scenarios); limited/noisy training samples; interpretablity/explainability; DL security/trustfulness; robustness to noise/attacks; uncertainty in deep learning
  - Potential direction: combine domain knowledge and the DL's learning capabilities to mitigate deficiencies of traditional SP/IP and black-box DNN approaches
  - Bring DSL(?) in statistical SP into DL, e.g. statistical DL, Bayesian DL
  - Bring DL into SP, e.g. deep unfolding
- Combining physics-based modeling and DL
- Perspective: Seeing will no longer be believing

Adversarial ML:

- scrutinize potential security vulnerabilities of DL models by (virtually) attack them
- requires proper threat model

Analogies to forensics, anti-forensics, and counter anti-forensics: Both digital images and DL models are vulnerable to manipulations and attacks, intentionally or unintentionally, posing critical challenges in trusting digital images

Potential directions:

- combine both SP and IP with DL
- leverage domain knowledge in signal/image processing
  - investigate interpretation for DL-based digital image forensics problems
  - focusing on the vulnerability of digital images themselves
  - focusing on vulnerability of current DL models

Both attack and defense side will improve. It is harder to fool the traditional image processing features

Paul Rosin: There is no guarantee that combining learning- and model-based techniques can gain the benefits of the two. In fact, how can we be sure that the combination does not inherit the weaknesses of the two?!

Discussion about interpretability: Use domain knowledge where possible.

4.10.2 Sebastien Marcel: Biometrics Security and Privacy

Biometrics security and privacy (BSP) research group: signal processing and ML applied to BSP, e.g. biometric recognition, security, privacy, multi-modal fusion AI and responsible datasets: fairness, Trojan/backdoors, ethics and synthetic datasets

- BATL: Create face anti-spoofing technology with a multi-spectral sensor. Created a multi-spectral PAD dataset
- FairFace: Metric to measure fairness in biometric systems (fairness discrepancy rate), now working on fairness mitigation strategy [3, 7]
- SAFER: Generate synthetic datasets for training and testing
- Media forensics challenges ahead:
  - Hyper-realistic and real-time audio-visual fakes
    - Detection: generalization to unseen attacks
    - Attribution: identification of the source of the attacks
  - Fairness and transparency compliance (e.g. EU Artificial Intelligence Act)
    - bias assessment and mitigation (biometrics and forensics)
    - synthetic datasets (e.g. face datasets) for training/testing classifiers to circumvent data protection issues
    - certification labs: push for AI certification scheme

What is bias? When you consider the error rate for the general population, you have a low error rate. As soon as the population is broken down into groups, the performance of the subdistribution is diverse. Same errors for everybody.

Bias mitigation strategies:

1. post-processing of the scores
2. if you have access to the model: regularization in order to balance the errors
3. fix the dataset

4.10.3 Anderson Rocha: Key challenges ahead

1. Synthetic realities: People with their own view of the world, fabrication of views and images, fake news, deep fakes: How to deal with synthetic generators for faces/images/videos?
2. How to generalize, dealing with the openness and unseen scenarios, e.g., in spoofing or deepfake detection? Try to devise methods that adaptively train themselves, i.e. self-supervised learning.
3. Fusion: Combining different sensors and modalities for solving a particular problem.
4. Solutions to compliance problems with privacy laws: Federated learning, self-supervised learning with access to some information only

4.10.4 Irene Amerini: Multimedia forensics: Challenges ahead

Research objectives:

- Design multimedia forensics techniques able to detect manipulated contents
- Scale forensic investigations to real-world applications: deepfake detection, social network provenance

Future trends:

- Forgery detection and source identification on internet-style data, not only on lab datasets. Semantic forensics on multimedia/multimodal assets

- Defense solutions against disinformation attacks, e.g. images generated with text-to-image techniques
- Adversarial deep learning: understanding the robustness and security of developed techniques. Build platforms and procedures to test robustness of models.

Future trends in deep fake detection:

- Continual deepfake detection (continual learning)
- Multimodal approach for deepfake detection (or generative models)
- Generalization issues
- Real-time deepfake detection
- Certifying authorship (even of deepfakes) via blockchain. Back to watermarking?

Future trends:

- Datasets are huge but not huge enough. Potential solutions are self-supervised learning, generating synthetic training images, augmenting datasets with generative models
- Problems: Biased datasets
- Computational cost for training and hardware resources. Potential solution: Creating lightweight models that require less hardware resources but without sacrificing much of performance

Common themes in all 8 talks: Self-supervised learning, combining different ways (e.g. model-based and learning-based techniques, different modalities), and synthetic generators pose a pressing problem.

References

1 Luca Cuccovillo, Christoforos Papastergiopoulos, Anastasios Vafeiadis, Artem Yaroshchuk, Patrick Aichroth, Konstantinos Votis, and Dimitrios Tzovaras. Open challenges in synthetic speech detection. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2022.

2 Amped Software. Survey results: The state of video forensics 2022. `https://blog.ampedsoftware.com/2022/12/20/survey-results-the-state-of-video-forensics-2022/`, 2022.

3 Kimmo Karkkainen and Jungseock Joo. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.

4 Waldemar Berchtold, Huajian Liu, Simon Bugert, York Yannikos, Jingcun Wang, Julian Heeger, Martin Steinebach, and Marco Frühwein. Recognition of objects from looted excavations by smartphone app and deep learning. *Electronic Imaging*, 34:1–4, 2022.

5 Gabriel Bertocco, Antônio Theófilo, Fernanda Andaló, and Anderson Rocha. Reasoning for complex data through ensemble-based self-supervised learning. *arXiv preprint arXiv:2202.03126*, 2202.

6 Simon Bugert, Huajian Liu, Waldemar Berchtold, and Martin Steinebach. Cultural assets identification using transfer learning. *Electronic Imaging*, 34:1–4, 2022.

7 Tiago de Freitas Pereira and Sébastien Marcel. Fairness in biometrics: A figure of merit to assess biometric verification systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(1):19–29, 2022.

8 European Network of Forensics Science Institutes – Digital Imaging Working Group. Best practice manual for digital image authentication. `https://enfsi.eu/wp-content/uploads/2021/10/BPM_Image-Authentication_ENFSI-BPM-DI-003-1.pdf`, 2021. Issue No. 001.

9 European Network of Forensics Science Institutes – Forensic Speech and Audio Analysis Working Group. Best practice manual for digital audio authenticity analysis. https://enfsi.eu/wp-content/uploads/2022/12/FSA-BPM-002_BPM-for-Digital-Audio-Authenticity-Analysis.pdf, 2022. Issue No. 001.

10 Moreira, Daniel and Avila, Sandra and Perez, Mauricio and Moraes, Daniel and Testoni, Vanessa and Valle, Eduardo and Goldenstein, Siome and Rocha, Anderson. Multimodal data fusion for sensitive scene localization. Elsevier Information Fusion, v45, pp 307–323, 2019

11 Fraunhofer SIT. The KiKu-App: Using artificial intelligence to automatically recognize cultural assets. https://youtu.be/un4ED05Ag_I.

12 Chandrakanth Gudavalli, Erik Rosten, Lakshmanan Nataraj, Shivkumar Chandrasekaran, and BS Manjunath. SeeTheSeams: Localized detection of seam carving based image forgery in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–11, 2022.

4.11 Day 5 – Forensics Questions and the Future of the Field

Thorsten Beck (HU Berlin, DE) – recorder of the session

4.11.1 Discussion questions for the report

- How can we advance the field?
- How has the field changed in the past 5 years?
- What do you predict will happen in the next to 5-10 years?
- What is the biggest challenge in the field at the moment?
- What are the most critical changes that we must make to face the future effectively?
- What effect has deep learning made on the field?
- Who is making the greatest advancements in the field, and what are they doing?

4.11.2 The Future

10 year perspective:

- improved quality of synthetic media
- pervasiveness of synthetic media
- link between physical world and digital content will be broken – then crypto?
  and forensics may not help to reinforce trustworthiness/authenticity
- few generators will emerge possibly watermarked
- integration of AI and GoF (AI comes first)
- self-supervised DL

4.11.3 Research Challenges in the Field of Media Forensics

Core Challenges

- Generalization (if I know how to identify one deep fake, how do I know to detect different ones?)

- Distribution mismatches/distributional shift (how can we handle out of distribution samples?)
  (above items boil down to lack of realistic models in GoF MMF)
- Modeling (various kinds of) uncertainty/dealing with uncertainty

see also: limitations due to amount of training data

Data-related problems (different twist for GoF)

- representativity (number of variables considered)
- privacy / copyright and legal restrictions (* see security)
- bias (exists as variable, but does not necessarily consider real-world distribution of age/gender etc.)
- for A.I. forensic approaches, big data is required (sometimes one might be confronted with one-shot problems, that require GoF approaches, see §What speaks for A.I.?)

Marco Fontani: from the point of view of COURTS and JUDGES, it is generally not plausible to make decision about an individual by data derived from other sources.

Explainability (resp. Interpretability?) – for "AI eyes" only?

- Check for correct behavior
- for forensic use (how do machines "see the invisible")

Marco Fontani: Research papers ought to distinguish between explainability and interpretability.

Martin Steinebach: Interpretation of forensic results in court and trials must generally represent not only the perspective of the prosecution, but also the perspective of the defense (neutrality).

Luca Cuccovillo: It should be considered that explainability in the literature is discussed as subset of trustworthiness.

Security

- enlarged attack surface wrt GoF (also because of training) – see also adversarial examples
- develop suitable threat models
- cat and mouse loop

What speaks for the application of A.I.?

- lack of good models for GoF fitting the complexity of real life
- coping with dynamic changes (e.g. software updates for cameras)
- benefits from pre-training/immediate benefit from available standard computer vision models (?)
- less domain knowledge needed (?)

Christian Riess: greybox/blackbox examples cannot always be sufficiently addressed via GoF

Martin Steinebach: problem with A.I. – in real world cases: one needs maybe 10 photos to identify a camera, but with A.I. you need thousands of images to train models. Real-life scenarios may require GoF approaches. It may be hard to explain criminal investigators or the police that large amounts of data is required to make A.I. work.

Participants

- Irene Amerini
Sapienza University of Rome, IT
- Mauro Barni
University of Siena, IT
- Thorsten Beck
Humboldt Universität zu Berlin, DE
- Tiziano Bianchi
Polytechnic University of Turin, IT
- Luca Cuccovillo
Fraunhofer IDMT – Ilmenau, DE
- Isao Echizen
National Institute of Informatics – Tokyo, JP
- Benedikt Lorch
Universität Innsbruck, AT
- Christian Riess
Friedrich-Alexander-Universität Erlangen-Nürnberg, DE
- Paul Rosin
Cardiff University, UK
- Martin Steinebach
Fraunhofer SIT – Darmstadt, DE

Remote Participants

- Roberto Caldelli
CNIT – Florence and Mercatorum University, IT
- Marco Fontani
Amped Software – Trieste, IT
- Haiying Guan
NIST – Gaithersburg, US
- Zulfiqar Habib
Comsats University – Lahore, PK
- Lakshmanan Nataraj
Trimble Inc. – Chennai, IN
- Ngai Fong Law
The Hong Kong Polytechnic University, HK
- Sebastian Marcel
Idiap Research Institute – Martigny, CH
- Alessandro Piva
University of Florence, IT
- Anderson Rocha
State University – Campinas, BR
- Xianfang Sun
Cardiff University, UK
- Benedetta Tondi
University of Siena, IT
- Z. Jane Wang
University of British Columbia – Vancouver, CA

Report from Dagstuhl Seminar 23022

Inverse Biophysical Modeling and Machine Learning in Personalized Oncology

George Biros*[1], Andreas Mang*[2], Björn H. Menze*[3], and Miriam Schulte*[4]

1 **University of Texas at Austin, US.** biros@oden.utexas.edu
2 **University of Houston, US.** andreas@math.uh.edu
3 **Universität Zürich, CH.** bjoern.menze@uzh.ch
4 **Universität Stuttgart, DE.** miriam.schulte@ipvs.uni-stuttgart.de

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 23022 "Inverse Biophysical Modeling and Machine Learning in Personalized Oncology".

This seminar brought together leading experts in mathematical, computational, and medical imaging sciences with research interests in data science, scientific machine learning, modeling and numerical simulation, optimization, and statistical and deterministic inversion, and image analysis with applications in medical imaging, and, in particular, oncology. A central theme of the seminar was the integration of data-driven methods with model-driven approaches for predictive modeling.

The seminar had several main thrusts including design and analysis of novel mathematical models, recent developments in medical imaging, machine learning in the context data analytics and data-driven model prediction, predictive computational modeling through (statistical) inversion, integration of machine learning with model-based priors and use of these methods to aid decision-making. We discussed these topics through the lens of foundational algorithmic complications and mathematical and computational challenges. The participants explored how advances in the applied sciences (e.g., data analytics, medical imaging, or radiomics) can aid us to tackle challenges in the application domain. We also discussed the significant challenges associated with the validation of the proposed methodology, and a lack of reproducibility due to the absence of standard protocols for validation of data- and model-driven methods by translational research groups.

Seminar January 8–13, 2023 – https://www.dagstuhl.de/23022

2012 ACM Subject Classification Applied computing → Imaging; Applied computing → Life and medical sciences; Computing methodologies → Machine learning algorithms; Computing methodologies → Machine learning approaches; Mathematics of computing → Mathematical optimization; Mathematics of computing → Mathematical software performance; Mathematics of computing → Numerical analysis; Mathematics of computing → Solvers

Keywords and phrases Bayesian inverse problems, image segmentation, inverse problems, machine learning, medical image analysis, parallel computing, tumor growth simulation and modeling

Digital Object Identifier 10.4230/DagRep.13.1.36

* Editor / Organizer

Inverse Biophysical Modeling and Machine Learning in Personalized Oncology, *Dagstuhl Reports*, Vol. 13, Issue 1, pp. 36–67
Editors: George Biros, Andreas Mang, Björn H. Menze, and Miriam Schulte
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Andreas Mang (University of Houston, US)
George Biros (University of Texas at Austin, US)
Björn H. Menze (Universität Zürich, CH)
Miriam Schulte (Universität Stuttgart, DE)

Our Dagstuhl Seminar brought together leading experts in computational and applied mathematics, computer science, biomedical imaging, and medical imaging sciences with research interests in data science, machine learning, modeling, optimization, and statistical and deterministic inversion with applications in medical imaging, and – in particular – oncology. Overall, 22 participants (and 5 remote participants) from various scientific disciplines contributed with scientific presentations about their current and future research efforts in these areas.

The seminar had four main thrusts: (i) machine learning in the context of data analytics and data-driven model prediction, (ii) predictive computational modeling through statistical and deterministic inversion, (iii) integration of machine learning with model-based priors, and (iv) use of these methods to aid decision making. We discussed these topics through the lens of foundational algorithmic complications and mathematical and computational challenges. We also explored how advances in the applied sciences (e.g., data analytics, medical imaging, radiomics, genomics, or experimental design) can aid us to tackle challenges associated with the design of computational and mathematical methods.

In the context of predictive computational modeling and deterministic and statistical inversion, we addressed topics ranging from uncertainty quantification, model choices (multiscale versus macroscopic; model-complexity; multispecies versus single-species), regularization strategies, sensitivity analysis, strategies to address the massive computational costs, challenges in the design of hardware-accelerated computational methods with optimal energy efficiency, and strategies to yield the throughput, robustness, and reliability required in practical applications under given hardware constraints. In the context of machine learning and its integration with predictive modeling and priors, we discussed issues associated with limited reproducibility beyond the training data, robustness against outliers, issues with small-sample size problems, uncertainty quantification for learning from data, and generic strategies to enrich the available data. Lastly, we also explored the availability and use of advanced imaging technologies that can help to (i) provide a better data basis for predictive modeling, (ii) trigger community efforts to enrich available data, and (iii) enable validation and standardize population-based studies. We also discussed reproducibility issues, given that in many cases (medical imaging) data is proprietary, challenges associated with the validation of the proposed methodology, and a lack of reproducibility due to the absence of standard protocols for validation of data- and model-driven methods by translational research groups.

The seminar started with opening remarks by two of the main organizers (Biros and Menze). They reviewed their contributions to the field and gave an overview of the state-of-the-art from their perspective. This opened up the floor for a first discussion on where we are and where we would like to go with our future research. During the first two and a half days different scientists contributed to our seminar with presentations about their recent activities and their view on the state-of-the-art. We did not keep a tightly fixed schedule. This allowed participants to engage and discuss the presented material, shed light on potential future research avenues, identify common areas of interest between participants and research groups, as well as exchange ideas on how to address potential shortcomings of the state-of-the-art methods. Overall, this led to an active exchange about open issues, potential solutions, and current activities among participants of our seminar. The topics discussed during the research presentations include computational and mathematical approaches targeted at aiding clinical treatment (with contributions form, e.g., Brüningk, Fuster Garcia, Hormuth, Menze), the design of new mathematical models of cancer/tumor progression (with contributions from, e.g., Biros, Deutsch, Gomez, Menze, Schulte, Wohlmuth), the design and analysis of methodology for machine learning (with contributions from, e.g., Erhardt, Konukoglu, Pati), inverse problems and optimization (with contributions from, e.g., Biros, Erhardt, Latz, Mang, Schulte), scientific machine learning with applications in medical imaging (with contributions from e.g., Brüningk, Erhardt, Fuster Garcia, Konukoglu, Li, Merhof, Van Leemput), hardware-accelerated computational methods, high performance computing, and computational complexity (with contributions from Biros, Mang, Schulte), the integration of modeling integration of data-driven methods with model-driven approaches for predictive modeling (with contributions from, e.g., Biros, Brüningk, Hormuth, Lorenzo, Menze, Schulte, Wiestler), and advances in medical imaging and medical image analysis (with contributions from, e.g., Li, Lundervold, Merhof, Paech, Pati, Van Leemput, Weidner, Wiestler). Several of these contributions are briefly described in the abstracts included in this report.

As mentioned above, during the discussions after each scientific presentation, we identified several *open problems and challenges* that we believe should be addressed by the community at large. We briefly list some of the main points raised during these discussions here:

Regarding the integration of computational models with medical imaging, a key challenge is to establish if a model is of use in the clinical context. Many of the available mathematical models are oversimplifications, particularly in the context of modeling cancer progression at a tissue scale. As such, one generic use of these models is to utilize them as “priors” for more classical image analysis tasks such as image segmentation or image registration.

As for generating model-based predictions, a key remaining challenge is how simple or complicated mathematical models need to be, to be of clinical value. While some tasks (e.g., patient classification or tissue characterization) can potentially be helped by simple models, an open question is how complicated models can or have to be to aid clinical decision-making or enable model-based predictions (e.g., if one envisions forecasting the benefit of certain types of clinical intervention in individual patients).

Another key challenge in this context is the scarcity of the available data. Moreover, how do we validate and compare the performance of these approaches and how can we establish good benchmarks to test methods developed by individual research teams? A related open question is, which scale is most suitable to simulate certain aspects of cancer growth/disease progression and/or treatment? Are microscopic rule-based approaches required, or can we utilize coarser, macroscopic models that typically formulate tumor/cancer progression in terms of partial differential equations? Another question is to what extent and if physics-informed methods (i.e., methods based on the simulation of biophysical phenomena) add significant value to clinical diagnosis and treatment planning versus more standard, machine learning-based predictions generated from features derived from imaging data. One key question that was also discussed in this context was how these methods are plagued by model and data uncertainties.

Moreover, we discussed how to integrate modeling with machine learning in the most efficient way. Can we, e.g., use machine learning as a tool to initialize more classical (e.g., variational) methods for inference of model parameters and/or integration of simulation with data? Conversely, can machine learning benefit from an integration of physics-based principles prescribed by biophysical models? Likewise, can machine learning be used as a tool to improve model selection, i.e., can we use it to decide how complex a mathematical model needs to be?

From an imaging perspective, one question that arose was how to combine different types of data (e.g., structural imaging, biomedical markers, radiomics, functional imaging, patient questionnaires) most effectively. In many studies, one typically does not integrate information from multiple sources but relies on specific types of medical data. Would such a more complete integration aid model-based predictions? How does the designed methodology generalize for data acquired at different imaging sites and/or imaging modalities? Another key issue is the scarcity of publicly available (good quality) data and how to address it as a community. One solution presented at the seminar was the use of federated learning.

Lastly, if we envision pushing these methods toward clinical applications, how can we deal with low-performance computing infrastructure at clinical sites? We also discussed clinical scenarios for applying the designed methods and how they could be of use in clinical practice (for example, to plan a therapeutic intervention or post-therapy assessment).

On Wednesday, we engaged participants in scientific discussions during an excursion to an art show at the “Völklinger Hütte”. We concluded this social event with a joint dinner in one of the local restaurants.

The scientific presentations were followed by a brief discussion about selected topics in two working groups to identify immediate goals and further discuss existing challenges. The first group included researchers with a key interest in designing methods to analyze medical (imaging) data and integrate mathematical and computational methods with imaging and medical data. The second group discussed topics associated with the design of mathematical and computational methods for inference, simulation, and optimization. We list the key findings in these two groups and some of the questions that remain to be addressed by the community at large in this report. We concluded our seminar with a plenary discussion about the findings of our working group discussions. This enabled us to identify commonalities toward a more concrete outline of follow-up work after the conclusion of our seminar. As a first concrete goal for the entire group, we agreed that we should start our work with a joint (public) dataset that compiles available medical imaging data for model development and testing. Spearheaded by Gomez and Hormuth, a first list of publicly available data was curated on the Mathematical Oncology webpage: `https://mathematical-oncology.org/resources/datasets`. Moreover, they have started to collect information for relevant conferences and workshops of interest for the community at large (`https://mathematical-oncology.org/conferences`).

2 Table of Contents

Executive Summary

Overview of Talks

Panel discussions

3 Overview of Talks

3.1 Harnessing machine learning and mechanistic modelling for personalized radiotherapy of pediatric diffuse midline glioma

Sarah Brüningk (ETH Zürich – Basel, CH)

Joint work of Sarah Brüningk, Catherine Jutzeler, Javad Nazarian, Karsten Borgwardt, Sabine Mueller
Main reference Sarah C. Brüningk, Jeffrey Peacock, Christopher J. Whelan, Renee Brady-Nicholls, Hsiang-Hsuan M. Yu, Solmaz Sahebjam, Heiko Enderling: "Intermittent radiotherapy as alternative treatment for recurrent high grade glioma: a modeling study based on longitudinal tumor measurements". Scientific Reports, 11(1), 2021.
URL https://doi.org/10.1038/s41598-021-99507-2

Pediatric diffuse midline glioma is a rare, yet fatal disease, with currently no curative treatment. Owing to the delicate location of these tumors, treatment options and surgical interventions are greatly limited. Radiotherapy (RT) is one of the few life-prolonging treatments, but its therapeutic efficacy varies between individuals. Currently, it is impossible to predict RT benefit a priori and there is a great unmet clinical need to improve patient stratification and survival.

The overarching aim of this project is to build a treatment decision support platform facilitating personalized RT optimization based on non-invasive magnetic resonance imaging. To this end, we develop an analytical pipeline bridging mechanistic modelling and data-driven machine learning to refine patient stratification, discover imaging biomarkers, and inform RT scheduling and dosing by an individualized radiosensitivity score (RSS).

Imaging and clinical data from ∼250 patients centralized from different international institutions are at the centre of this analysis. Image classification will be based on a scalable combination of local and global image features reflecting the biological hallmarks of DMGs. The challenge of limited, multi-domain data is addressed via the model architecture together with transfer learning from adult glioblastoma and data augmentation. We employ interpretability analysis to identify imaging biomarkers driving classification, and use regression analysis to infer a RSS. An ordinary differential equation model of longitudinal tumor growth under RT is fitted to follow-up patient data. Based on the fitted model parameters and the RSS, alternative RT strategies can then be simulated and the gain in time to progression of an *in silico* trial comparing conventional and personalized RT will be quantified. At this point we are in the early phase of the study and have centalized patient data from the University of Californa, San Francisco, from The DMG Centre Zurich, and from patients treated as part of clinical trials within the Pacific Pediatric Neuro-Oncology Consortium (PNOC).

This study investigates personalized RT for a group of pediatric patients for which treatment individualization is inevitable. The treatment decision support tool and the identified imaging biomarkers should be translatable to clinical practice, while our *in silico* trial may motivate clinical evaluation to provide validation of our predictions. By focussing on imaging data and available, cost effective RT, our approach is feasible in treatment facilities worldwide with clear application of digital pediatric health. Relevant references are [1, 2].

Acknowledgements

This project is funded by the Botnar Research Centre for Child Health as part of their postdoctoral excellence programme under #PEP-2021-1008. The ETH Career Seed grant 2021-HS-329 supported the centralization of the data sets.

References

1 Kuijs, M., Jutzeler, C. R., Rieck, B., and Brüningk, S. C. (2021). Interpretability aware model training to improve robustness against out-of-distribution Magnetic Resonance Images in Alzheimer's disease classification. arXiv. https://doi.org/10.48550/arXiv.2111.08701

2 Brüningk, S. C., Peacock, J., Whelan, C. J., Brady-Nicholls, R., Yu, H.-H. M., Sahebjam, S., and Enderling, H. (2021). Intermittent radiotherapy as alternative treatment for recurrent high grade glioma: a modeling study based on longitudinal tumor measurements. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-021-99507-2

3.2 Mechanisms of cancer invasion and progression: insights from cellular automaton models

Andreas Deutsch (TU Dresden, DE)

Tumour invasion and progression may be viewed as collective phenomena emerging from the interplay of biological cells with their environment. Cell-based mathematical models in which cells are regarded as separate discrete entities can be used to decipher the rules of interaction. Here, we focus on the dynamics of glioma and breast cancer. We introduce lattice-gas cellular automaton models [1, 5] to analyse the role of phenotypic plasticity in cancer invasion, define spatial and non-spatial Moran processes to shed light on the size of the tumour originating niche, and adopt Markov chain models to investigate the origin of genetic heterogeneity in glioblastoma [2, 3, 4].

References

1 A. Deutsch and S. Dormann, Cellular automaton modeling of biological pattern formation: characterization, applications, and analysis, Birkhauser, Boston, 2018.

2 T. Buder, A. Deutsch, B. Klink and A. Voss-Böhme, Patterns of tumor progression predict small and tissue-specific tumor-originating niches, Front. Oncol., 8, 668, 2019.

3 Anne Dirkse, Anna Golebiewska, Thomas Buder, Petr V. Nazarov, Arnaud Muller, Suresh Poovathingal, Nicolaas H. C. Brons, Sonia Leite, Nicolas Sauvageot, Dzjemma Sarkisjan, Mathieu Seyfrid, Sabrina Fritah, Daniel Stieber, Alessandro Michelucci, Frank Hertel, Christel Herold-Mende, Francisco Azuaje, Alexander Skupin, Rolf Bjerkvig, Andreas Deutsch, Anja Voss-Böhme and Simone P. Niclou, Stem cell-associated heterogeneity in Glioblastoma results from intrinsic tumor plasticity shaped by the microenvironment, Nature Commun., 10, 1, 1787, 2019.

4 Olga Ilina, Pavlo G. Gritsenko, Simon Syga, Jürgen Lippoldt, Caterina A. M. La Porta, Oleksandr Chepizhko, Steffen Grosser, Manon Vullings, Gert-Jan Bakker, Jörn Starruß, Peter Bult, Stefano Zapperi, Josef A. Käs, Andreas Deutsch and Peter Friedl, Cell–cell adhesion and 3D matrix confinement determine jamming transitions in breast cancer invasion, Nature Cell Biol., 1103–1115, 2020.

5 A. Deutsch, J. M. Nava-Sedeño, S. Syga, H. Hatzikirou, BIO-LGCA: a cellular automaton modelling class for analysing collective cell migration, PLOS Comp. Biol., 17, 6, e1009066, 2021.

3.3 Machine Learning meets Inverse Problems: Bilevel Learning

Matthias J. Ehrhardt (University of Bath, GB)

Joint work of Matthias J. Ehrhardt, Lindon Roberts, Ferdia Sherry, Martin Benning, Juan Carlos de los Reyes, Martin J.Graves, Georg Maierhofer, GuyWilliams, Carola-Bibiane Schönlieb
Main reference Matthias J. Ehrhardt, Lindon Roberts: "Inexact Derivative-Free Optimization for Bilevel Learning", J. Math. Imaging Vis., Vol. 63(5), pp. 580–600, 2021.
URL http://dx.doi.org/10.1007/s10851-021-01020-8

Inverse problems are omnipresent in any imaging related field and is as such a backbone in oncology, too. Here we focussed on the connections of machine learning to the particular inverse problem of image reconstruction but many concepts generalise to other inverse problems such as estimating parameters in PDEs. Solving inverse problems can be approached via variational regularization techniques which are dominant in the field of inverse problems in general. A drawback of these techniques is that they are dependent on a number of parameters which have to be set by the user. This issue can be approached by machine learning where we estimate these parameters from data. This is known as "Bilevel Learning" and has been successfully applied to many tasks, some as small-dimensional as learning a regularization parameter, others as high-dimensional as learning a sampling pattern in MRI. While mathematically appealing this strategy leads to a nested optimization problem which is computationally difficult to handle. We discussed several applications of bilevel learning for imaging [2, 1] as well as new computational approaches [1, 3].

References

1 Ehrhardt, M. J., and Roberts, L. (2021).Inexact Derivative-Free Optimization for Bilevel Learning. Journal of Mathematical Imaging and Vision, 63(5), 580–600. `https://doi.org/10.1007/s10851-021-01020-8`

2 Sherry, F., Benning, M., de los Reyes, J. C., Graves, M. J., Maierhofer, G., Williams, G., Schönlieb, C.-B., and Ehrhardt, M. J. (2020). Learning the Sampling Pattern for MRI. IEEE Transactions on Medical Imaging, 39(12), 4310–4321.

3 Ehrhardt, M. J., and Roberts, L. (2023). Analyzing Inexact Hypergradients for Bilevel Learning. `http://arxiv.org/abs/2301.04764`

3.4 Computational Radiology & Artificial Intelligence in Cancer

Elies Fuster Garcia (Technical University of Valencia, ES)

Joint work of Atle Bjornerud, Kyrre E. Emblem, Siri Fløgstad Svensson, Elies Fuster Garcia, Juan M. García-Gómez, F. Javier Gil-Terrón, Qinghui Liu, Carles Lopez-Mateu, Donatas Sederevičius
Main reference Elies Fuster-Garcia, Ivar Thokle Hovden, Siri Fløgstad Svensson, Christopher Larsson, Jonas Vardal, Atle Bjørnerud, Kyrre E. Emblem: "Quantification of Tissue Compression Identifies High-Grade Glioma Patients with Reduced Survival", Cancers, Vol. 14(7), 2022.
URL http://dx.doi.org/10.3390/cancers14071725

Recent advances in medical imaging, coupled with the analysis capabilities offered by artificial intelligence, have led to significant progress in personalized oncology. Advanced MRI sequences in neuroimaging are now able to provide critical biophysical parameters for the study of tumor growth, response to therapies, and clinical decision-making. Furthermore, the integration of multi-parametric information, which would be otherwise infeasible, is now made possible through artificial intelligence. This presentation will introduce the collaborative efforts between the Biomedical Data Science Lab (Universitat Politècnica de València, UPV) and the MRI research and technology (Oslo University Hospital, OUH) to combine these two disciplines and make a real impact on clinical practice, particularly on high-grade glial tumors.

The OUH is improving its MRI protocol for neuro-oncology studies by incorporating advanced MRI sequences, such as Vessel Caliber MRI, Vessel Architectural Imaging, and MR Elastography. These sequences offer valuable information at the voxel level, such as vessel caliber size and density [1], vessel type dominance and microvascular efficiency [2], and tissue biomechanics by stiffness and viscosity [3, 8]. This enables researchers to gather a wider range of information on the brain's blood vessels and tissue, providing a more comprehensive understanding of neuro-oncology.

To integrate all of the information gathered through advanced MRI sequences, processing pipelines and multi-parametric artificial intelligence models are being developed. The collaboration between the Oslo University Hospital (OUH) and the Universitat Politècnica de València (UPV) has led to the creation of AI systems that can accurately segment regions of interest [4], identify functional habitats [5], and analyze longitudinal series and growth dynamics [6], among others. An example of such a system is the publicly available ONCOhabitats platform developed by the UPV, which studies vascular heterogeneity in patients with high-grade glial tumors [7].

The success of these AI technologies in clinical practice depends on their integration into a relevant environment at the moment of decision-making. To achieve this, OUH's models and associated pipelines are being integrated into a computation framework connected with the hospital PACS through the TrackGrowth, Chronos, and Progress research projects (see Acknowledgements). This setup allows for the direct evaluation of AI-based solutions in PACS by deploying hospital-approved software in the hospital interface.

Acknowledgements

We gratefully acknowledge support from the European Union's Horizon 2020 Programme: ERC Grant [758657-ImPRESS], Research and Innovation Grant [668039-FORCE], Marie Skłodowska-Curie grant [844646-GLIOHAB]; South-Eastern Norway Regional Health Authority [2017073, 2013069, 2021057]; The Research Council of Norway Grant [261984, 325971, 325971]; the National Institutes of Health R21 grant [EB030757]; the German Research Foundation [SCHR 1542/1-1]; Grant PID2021-127110OA-I00 (PROGRESS) funded by MCIN/AEI/ 10.13039/501100011033 and by ERDF A way of making Europe; Grant PID2019-104978RB-I00/AEI/10.13039/501100011033 (ALBATROSS) funded by MCIN/AEI/ 10.13039/501100011033.

References

1 Emblem, K. E. et al. Vessel caliber–a potential MRI biomarker of tumour response in clinical trials. Nat. Rev. Clin. Oncol. 11, (2014).

2 Emblem, K. E. et al. Vessel architectural imaging identifies cancer patient responders to anti-angiogenic therapy. Nat. Med. 19, 1178–1183 (2013).

3 Fløgstad Svensson, S. et al. Decreased tissue stiffness in glioblastoma by MR elastography is associated with increased cerebral blood flow. Eur. J. Radiol. 147, 110136 (2022).

4 Juan-Albarracín, J., Fuster-Garcia, E., Manjon, J. V., Robles, M., Aparici, F., Martí-Bonmatí, L., and Garcia-Gomez, J. M. Automated glioblastoma segmentation based on a multiparametric structured unsupervised classification. PLoS ONE 10, (2015).

5 Álvarez-Torres, M. D. M. et al. Robust association between vascular habitats and patient prognosis in glioblastoma: An international multicenter study. J. Magn. Reson. Imaging JMRI (2019) https://doi.org/10.1002/jmri.26958.

6 Fuster-Garcia, E. et al. Quantification of Tissue Compression Identifies High-Grade Glioma Patients with Reduced Survival. Cancers 14, 1725 (2022).

7 Juan-Albarracín, J., Fuster-Garcia, E., García-Ferrando, G. A. and García-Gómez, J. M. ONCOhabitats: A system for glioblastoma heterogeneity assessment through MRI. Int. J. Med. Inf. 128, 53–61 (2019).

8 Siri Fløgstad Svensson, Kyrre Eeg Emblem, and Elies Fuster-Garcia. (2021). MR Elastography, perfusion and diffusion data in 9 patients with glioblastoma and 17 healthy subjects [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4926005

3.5 An image-driven computational modeling approach to forecast radiotherapy response in gliomas

David Hormuth (University of Texas at Austin, US)

Joint work of David A. Hormuth II, Karine A. Al Feghali, Andrew M. Elliott, Thomas E. Yankeelov, Caroline Chung

Main reference David A. Hormuth II, Karine A. Al Feghali, Andrew M. Elliott, Thomas E. Yankeelov, Caroline Chung: "Image-based personalization of computational models for predicting response of high-grade glioma to chemoradiation". Scientific Reports. 2021; 11:1-14.

URL https://doi.org/10.1038/s41598-021-87887-4

Radiotherapy (RT) is a foundational component of clinical management for high-grade glioma (HGG) used to target residual and infiltrative disease following surgical resection. Variability in patient response to radiotherapy can depend on the tumor's underlying sensitivity to treatment as well as the ability to accurately target the biologically relevant malignant tissue. To improve patient outcomes, RT treatment plans could be adapted for individual patients to target tumor sub-regions demonstrating treatment resilience and higher aggressive potential. Towards this goal, we developed a family of biologically-based mathematical models of HGG growth and response, which are initialized and calibrated using patient-specific multi-parametric magnetic resonance imaging (mpMRI) data [1, 2]. Our family of models is built upon a 3D, two-species model of enhancing and non-enhancing tumor that describes tumor cell proliferation, diffusion, and treatment response. Unique to our approach is the use of mpMRI collected weekly during RT which reports on both tumor extent and cellularity dynamics. Using patient imaging data collected before treatment onset and weekly up to mid-treatment, we identified patient-specific tumor growth and response parameters via a non-linear least squares optimization. These patient-specific model parameters were then used to forecast tumor growth and response dynamics at the remaining weekly imaging visits during RT. In an initial cohort of 13 patients, we observed that our computational framework was able to predict total tumor cell count with a Pearson correlation coefficient of 0.95 and concordance correlation coefficient of 0.91 at 1-month post-RT. Likewise, the forecasted total tumor volume agreed spatially with the observed tumor volume with Dice similarity coefficients greater than 0.73. At the individual voxel-level, the forecasted distribution of tumor growth was able to predict areas of significant increases or decreases in tumor cell with an accuracy, specificity, and sensitivity greater than 0.76. The results of this initial study demonstrates the ability for image-driven modeling to predict HGG response to RT that with further development may enable anticipatory adaption of RT.

Acknowledgements

NCI R01CA235800, U24CA226110, U01CA174706, CPRIT RR160005 and CPRIT RP220225. Caroline Chung is supported by the Andrew Sabin Family Foundation Fellowship, MD Anderson Cancer Center Institutional Research Grant and CCSG Radiation Oncology and Cancer Imaging Program, and the Marnie Rose Foundation. T.E.Y. is a CPRIT Scholar in Cancer Research.

References

1 Hormuth II DA, Feghali KA Al, Elliott AM, Yankeelov T, Chung C. Image-based personalization of computational models for predicting response of high-grade glioma to chemoradiation. Scientific Reports. 2021;11:1-14. `https://doi.org/10.1038/s41598-021-87887-4`

2 Hormuth DA, Farhat M, Christenson C, et al. Opportunities for improving brain cancer treatment outcomes through imaging-based mathematical modeling of the delivery of radiotherapy and immunotherapy. Advanced Drug Delivery Reviews. 2022;187:114367. `https://doi.org/10.1016/j.addr.2022.114367`

3.6 On the well-posedness of Bayesian inverse problems

Jonas Latz (Heriot-Watt University – Edinburgh, GB)

Main reference Jonas Latz: "On the Well-posedness of Bayesian Inverse Problems", SIAM/ASA J. Uncertain. Quantification, Vol. 8(1), pp. 451–482, 2020.
URL http://dx.doi.org/10.1137/19M1247176

Mathematical models that are used in science and engineering often need to be calibrated with respect to observational data. In the context of tumour modelling, for instance, image data can be used to estimate chemotaxis, consumption, and proliferation of a tumour [1]. Such parameter estimation problems are often referred to as "inverse problems". Due to observational noise and complexity of models, inverse problems are usually difficult to solve and also *ill-posed*: a well-posed problem on the opposite is one, that has a solution, the solution is unique, and the solution depends continuously on the data. Well-posedness is important. Without existence, the problem has no solution and is, thus, not solvable. Uniqueness is required to prevent ambiguity between different solutions. The continuity assumption is a stability condition: the data is noisy, thus, we should hope that the influence of the noise on the parameter estimate is restricted in a certain sense.

The Bayesian approach to inverse problems gives a way to turn an ill-posed inverse problem into a well-posed problem. Here, we consider the calibration problem to be a statistical problem and model noise and unknown parameter as random variables. Through conditioning we are then able to incorporate the information from the data into the parameter. The conditioning can be achieved through Bayes' formula.

As shown in [2], the resulting "Bayesian inverse problem" will be well-posed under very, very mild assumptions, allowing for parameter estimation in blackbox models and, e.g., with respect to data-driven prior models.

References

1 Christian Kahle, Kei Fong Lam, Jonas Latz and Elisabeth Ullmann. Bayesian parameter identification in Cahn-Hilliard models for biological growth. SIAM/ASA Journal on Uncertainty Quantification 7(2): 526-552, 2019.

2 Jonas Latz. On the Well-posedness of Bayesian inverse problems. SIAM/ASA Journal on Uncertainty Quantification 8(1): 451–482, 2020.

3.7 Intelligent Neuroimaging for Precision Neuro-oncology

Chao Li (University of Cambridge, GB)

Brain tumour comprises a spectrum of malignant and benign entities. The complex pathophysiology of brain tumours poses challenges to effective clinical decision-making and treatment for patients. Multi-modal neuroimaging provides a non-invasive technique for probing brain tumours [5, 3, 13]. Based on neuroimaging, artificial intelligence (AI) offers an automated solution to optimise patient management, promising to accelerate precision neuro-oncology. Typically, the clinical applications of AI include tools for automatic diagnostics and guiding precise treatment. Together, these AI models promise to improve the overall efficiency of healthcare. Through engaging clinical domain knowledge, AI models can be tailored to the critical challenges in neuro-oncology, which could further advance our understanding of brain tumours and accelerate individualised and precise therapeutics.

Glioma is the most common malignant brain tumour in adults, characterised by remarkable heterogeneity and extensive invasion. To characterize tumour heterogeneity based on imaging, we designed novel radiological features to characterize tumour morphology and spatial heterogeneity [12]. Combined with machine learning methods, these features show robust performance in subtyping patients across diverse tissues and imaging modalities. The identified patient sub-groups show distinct molecular characteristics and prognostics. Advanced MRI techniques, e.g., perfusion and diffusion MRIs [4, 6], provide sensitive information for characterising tumour invasion over contrast-enhanced MRI. However, advanced MRI are typically in low resolution, which hinders full training labels for developing supervised models. To mitigate this challenge, we develop weakly supervised deep learning models that can identify the tumour invasion outside of contrast enhancement [2]. Further, glioma is considered a systematic disease, as it frequently spreads along white matter tracts into the whole brain. To characterize the tumour invasion globally, we developed an iterative tract-based spatial statistics method to quantify the structural connectivity of the brain and measure tract integrity in brain tumour patients [11]. Through comparing patients to healthy controls, we identified regional disrupted connectome in glioblastoma patients, which shows significance in predicting patient survival and indicating treatment targets [10]. Following this study, we introduced brain connectome into the AI model to better characterise glioma. Specifcially, we developed a multi-modal learning model, which leverages three encoders to extract features of focal tumour image, tumour geometrics and global brain network in predicting the isocitrate dehydrogenase (IDH) mutation, achieving higher performance over other state-of-the-art models [9].

In translating AI models into real-world practice, we need to tackle the challenges from heterogeneous clinical datasets, e.g., missing scans, and low image resolution. Therefore, we develop AI approaches to enhance image quality and standardisation [7, 1, 8]. For a trustworthy AI solution, we develop biophysics-informed deep learning models to enhance model explainability and generalisability. With these AI prototypes developed, we test the models in the real-world clinical setting, by connecting model development with the clinical system to obtain clinical and biological validations. We develop multi-centre imaging trials to validate the efficacy of imaging tools, where MR images are processed using reproducible and transparent pipelines. In the next step, we will test the imaging tools at scale through connections to large population data. Our vision is to transform the healthcare of brain tumour patients using image-based AI models.

Acknowledgements

Supported by Guarantors of Brain.

References

1 Jiang, L., Mao, Y., Chen, X., Wang, X., Li, C.: Cola-diff: Conditional latent diffusion model for multi-modal mri synthesis. arXiv preprint arXiv:2303.14081 (2023).

2 Li, C., Huang, W., Chen, X., Wei, Y., Price, S.J., Schönlieb, C.B.: Expectation-maximization regularized deep learning for weakly supervised tumor segmentation for glioblastoma. arXiv preprint arXiv:2101.08757 (2021).

3 Li, C., Wang, S., Serra, A., Torheim, T., Yan, J.L., Boonzaier, N.R., Huang, Y., Matys, T., McLean, M.A., Markowetz, F., et al.: Multi-parametric and multi-regional histogram analysis of MRI: Modality integration reveals imaging phenotypes of glioblastoma. European Radiology, 29, 4718–4729 (2019).

4 Li, C., Wang, S., Yan, J.L., Piper, R.J., Liu, H., Torheim, T., Kim, H., Zou, J., Boonzaier, N.R., Sinha, R., et al.: Intratumoral heterogeneity of glioblastoma infiltration revealed by joint histogram analysis of diffusion tensor imaging. Neurosurgery 85(4), 524–534 (2019).

5 Li, C., Wang, S., Yan, J.L., Torheim, T., Boonzaier, N.R., Sinha, R., Matys, T., Markowetz, F., Price, S.J.: Characterizing tumor invasiveness of glioblastoma using multiparametric magnetic resonance imaging. Journal of Neurosurgery 132(5), 1465–1472 (2019)

6 Li, C., Yan, J.L., Torheim, T., McLean, M.A., Boonzaier, N.R., Zou, J., Huang, Y., Yuan, J., van Dijken, B.R., Matys, T., et al.: Low perfusion compartments in glioblastoma quantified by advanced magnetic resonance imaging and correlated with patient survival. Radiotherapy and Oncology 134, 17–24 (2019).

7 Liu, P., Li, C., Schönlieb, C.B.: Ganredl: Medical image enhancement using a generative adversarial network with real-order derivative induced loss functions. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. pp. 110–117. Springer (2019).

8 Mao, Y., Jiang, L., Chen, X., Li, C.: Disc-diff: Disentangled conditional diffusion model for multi-contrast mri super-resolution. arXiv preprint arXiv:2303.13933 (2023).

9 Wei, Y., Chen, X., Zhu, L., Zhang, L., Schönlieb, C.B., Price, S., Li, C.: Multi-modal learning for predicting the genotype of glioma. IEEE Transactions on Medical Imaging (2023).

10 Wei, Y., Li, C., Cui, Z., Mayrand, R.C., Zou, J., Wong, A. L. K. C., Sinha, R., Matys, T., Schönlieb, C.B., Price, S.J.: Structural connectome quantifies tumour invasion and predicts survival in glioblastoma patients. Brain 146(4), 1714–1727 (2023).

11 Wei, Y., Li, C., Price, S.J.: Quantifying structural connectivity in brain tumor patients. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24. pp. 519–529. Springer (2021)

12 Wu, J., Li, C., Gensheimer, M., Padda, S., Kato, F., Shirato, H., Wei, Y., Schönlieb, C.B., Price, S.J., Jaffray, D., et al.: Radiological tumour classification across imaging modality and histology. Nature Machine Intelligence 3(9), 787–798 (2021).

13 Yan, J.L., Li, C., Boonzaier, N.R., Fountain, D.M., Larkin, T.J., Matys, T., van der Hoorn, A., Price, S.J.: Multimodal MRI characteristics of the glioblastoma infiltration beyond contrast enhancement. Therapeutic advances in neurological disorders 12, 1756286419844664 (2019)

3.8 Personalized computational forecasting of prostate cancer growth during active surveillance

Guillermo Lorenzo (University of Pavia, IT & UT Austin, US)

Joint work of Guillermo Lorenzo, Jon S. Heiselman, Michael A. Liss, Michael I. Miga, Hector Gomez, Thomas E. Yankeelov, Thomas J. Hughes, Alessandro Reali
Main reference Guillermo Lorenzo, Jon S. Heiselman, Michael A. Liss, Michael I. Miga, Hector Gomez, Thomas E. Yankeelov, Thomas J. Hughes, Alessandro Reali: "Abstract 5064: Patient-specific forecasting of prostate cancer growth during active surveillance using an imaging-informed mechanistic model", Cancer Research, Vol. 82(12_Supplement), pp. 5064–5064, 2022.
URL http://dx.doi.org/10.1158/1538-7445.AM2022-5064

Active surveillance (AS) is a feasible management option for low to intermediate-risk prostate cancer (PCa), which represents almost 70% of newly-diagnosed cases. During AS, patients have their tumor monitored via multiparametric magnetic resonance imaging (mpMRI), serum prostate-specific antigen (PSA), and biopsies [1]. If any of these data reveal tumor progression towards an increased clinical risk, the patient is prescribed a curative treatment. However, clinical decision-making in AS is usually guided by population-based protocols that do not account for the unique, heterogenous nature of each patient's tumor. This limitation complicates the personalization of monitoring plans and the early detection of tumor progression, which constitute two unresolved problems in AS. To address these issues, we propose to forecast PCa growth using personalized simulations of an mpMRI-informed mechanistic model solved over the 3D anatomy of the patient's prostate [1, 2, 3]. We describe PCa growth *via* the dynamics of tumor cell density with a diffusion operator, representing tumor cell mobility, and a logistic reaction term, which accounts for tumor cell net proliferation [1, 2]. Model calibration and validation rely on assessing the mismatch between model predictions of the tumor cell density map with respect to corresponding mpMRI-based estimates [2]. Our preliminary results on a cohort of seven patients show a median concordance correlation coefficient (CCC) and Dice score (DSC) of 0.55 and 0.82, respectively, for the spatial fit of tumor cell density during model calibration using two mpMRI datasets. Then, model validation at the date of a third mpMRI scan resulted in median CCC and DSC of 0.33 and 0.76, respectively. Thus, while further improvement and testing in larger cohorts are required, we believe that our results are promising for the potential use of our methods to personalize AS protocols and predict tumor progression.

References

1 G. Lorenzo, J.S. Heiselman, M.A. Liss, M.I. Miga, H. Gomez, T.E. Yankeelov, T.J.R. Hughes, and A. Reali. Patient-specific forecasting of prostate cancer growth during active surveillance using an imaging-informed mechanistic model, *Cancer Research*, **82**(12 Supp): 5064, 2022.

2 G. Lorenzo, D.A. Hormuth II, A.M. Jarrett, E.A.B.F. Lima, S. Subramanian, G. Biros, J.T. Oden, T.J.R. Hughes, and T.E. Yankeelov. *Quantitative in vivo imaging to enable tumor forecasting and treatment optimization.* In: *Cancer, Complexity, Computation.* Eds.: Igor Balaz and Andrew Adamatzky, Springer, pp. 55-97, 2023.

3 G. Lorenzo, M.A. Scott, K. Tew, T.J.R. Hughes, Y.J. Zhang, L. Liu, G. Vilanova, and H. Gomez. Tissue-scale, personalized modeling and simulation of prostate cancer growth, *Proceedings of the National Academy of Sciences of the United States of America*, **113**(48): E7663-E7671, 2016.

Table 1 Performance evaluation for the multi-GPU implementation of CLAIRE for the registration of images of size 256^3 of different individuals. We report (from left to right) the considered hardware architecture, the used memory, the relative mismatch after registration, the runtime (in seconds) as well as the speedup compared to the CPU implementation.

version	hardware	mem	mis	runtime	speedup
CLAIRE	24 core x86		2.9e-2	1.5e2	1×
	P100	4.6GB	2.6e-2	5.2e0	28×
	V100	4.6GB	2.6e-2	4.2e0	36×
	RTX3080	5.0GB	2.6e-2	3.2e0	47×
CLAIRE*	P100	8.1GB	3.6e-2	2.9e0	52×
	4×V100	2.6GB	3.6e-2	2.1e0	71×
	RTX3080	8.5GB	3.6e-2	2.3e0	65×

3.9 Scalable Algorithms for Diffeomorphic Image Registration

Andreas Mang (University of Houston, US)

Joint work of Andreas Mang, George Biros, Malte Brunn, Naveen Himthani, Miriam Schulte
Main reference Andreas Mang, Amir Gholami, Christos Davatzikos, George Biros: "CLAIRE: A Distributed-Memory Solver for Constrained Large Deformation Diffeomorphic Image Registration", SIAM Journal on Scientific Computing, Vol. 41(5), pp. C548–C584, 2019.
URL http://dx.doi.org/10.1137/18M1207818

We present a framework for diffeomorphic image registration termed CLAIRE [1, 6]. This algorithm is an integral part of some of our efforts to develop algorithms for the analysis of brain tumor imaging data [4, 5, 7, 8, 9]. Diffeomorphic image registration is a non-linear, ill-posed inverse problem that poses significant mathematical and computational challenges. Generally speaking, we seek a $\mathbb{R}^d$-diffeomorphism $y \in \mathrm{diff}(\mathbb{R}^d)$, $d \in \{2, 3\}$ that establishes a point-wise spatial correspondence between two views (images) of the same scene. In our work, we consider a variational formulation governed by hyperbolic transport equations.

Our contributions are new algorithms and dedicated computational kernels to reduce the runtime. We study the performance of our solver in terms of rate of convergence, registration accuracy, and time-to-solution. We demonstrate that we can solve problems for clinically relevant data of sizes (256^3 voxels; $\sim$50 million unknowns) in under 5 seconds (see table below). Our formulation and numerical algorithms are described in [6, 10, 11, 13]. Our parallel CPU implementation is discussed in [6, 12]. Our parallel GPU implementation is presented in [2, 3]. The integration of our registration algorithm with models of tumor progression is presented in [4, 5, 7, 8, 9]. We overview the computational performance of our framework for diffeomorphic image registration for an image of size 256^3 in the table below. Compared to our CPU implementation we observe a speedup between 28× and 71× depending on the GPU and implementation (CLAIRE: standard implementation; CLAIRE*: additional intermediate variables kept in memory). We report from left to right the version of CLAIRE, the hardware CLAIRE is executed on, the memory use, the mismatch between the data after registration, the runtime in seconds and the speedup. We can see that the GPU implementation is significantly faster than our GPU implementation without sacrificing accuracy.

Acknowledgments

This work was partly supported by the National Science Foundation under the awards DMS-1854853, DMS-2012825, and DMS-2145845. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily reflect the views of NSF.

References

1 M. Brunn, N. Himthani, G. Biros, M. Mehl and A. Mang. CLAIRE: Constrained Large Deformation Diffeomorphic Image Registration on Parallel Computing Architectures. The Journal of Open Source Software, 6(61), 3038, 2021.

2 M. Brunn, N. Himthani, G. Biros, M. Mehl and A. Mang. Fast GPU 3D diffeomorphic image registration. Journal of Parallel and Distributed Computing, 149:149–162, 2021.

3 M. Brunn, N. Himthani, G. Biros, M. Mehl and A. Mang. Multi-node multi-GPU diffeomorphic image registration for large-scale imaging problems. Proc ACM/IEEE Conference on Supercomputing, pp. 523–539, 2020.

4 A. Mang, S. Bakas, S. Subramanian, G. Biros and C. Davatzikos. Integrated biophysical modeling and image analysis: Application to neuro-oncology. Annual Review of Biomedical Engineering, 22:309–341, 2020.

5 K. Scheufele, S. Subramanian, A. Mang, G. Biros and M. Mehl. Image-driven biophysical tumor growth model calibration. SIAM Journal on Scientific Computing, 42(3):B549–B580, 2020.

6 A. Mang, A. Gholami, C. Davatzikos and G. Biros. CLAIRE: A distributed-memory solver for constrained large deformation diffeomorphic image registration. SIAM Journal on Scientific Computing, 41(5):C548–C584, 2019.

7 K. Scheufele, A. Mang, A. Gholami, C. Davatzikos, G. Biros and M. Mehl. Coupling brain-tumor biophysical models and diffeomorphic image registration. Computer Methods in Applied Mechanics and Engineering, 237:533–567, 2019.

8 A. Mang, A. Gholami, C. Davatzikos and G. Biros. PDE-constrained optimization in medical image analysis. Optimization and Engineering, 19(3):765–812, 2018.

9 A. Gholami, A. Mang, K. Scheufele, C. Davatzikos, M. Mehl and G. Biros. A framework for scalable biophysics-based image analysis. Proc ACM/IEEE Conference on Supercomputing, 19:1–19:13, 2017.

10 A. Mang and G. Biros. A semi-Lagrangian two-level preconditioned Newton–Krylov solver for constrained diffeomorphic image registration. SIAM Journal on Scientific Computing, 39(6):B1064–B1101, 2017.

11 A. Mang and L. Ruthotto. A Lagrangian Gauss–Newton–Krylov solver for intensity- and mass-preserving diffeomorphic image registration. SIAM Journal on Scientific Computing, 39(5):B860–B885, 2017.

12 A. Mang, A. Gholami and G. Biros. Distributed-memory large-deformation diffeomorphic 3D image registration. Proc ACM/IEEE Conference on Supercomputing, pp. 842–853, 2016.

13 A. Mang and G. Biros. Constrained H^1-regularization schemes for diffeomorphic image registration. SIAM Journal on Imaging Sciences, 9(3):1154–1194, 2016.

14 A. Mang and G. Biros. An inexact Newton–Krylov algorithm for constrained diffeomorphic image registration. SIAM Journal on Imaging Sciences, 8(2):1030–1069, 2015.

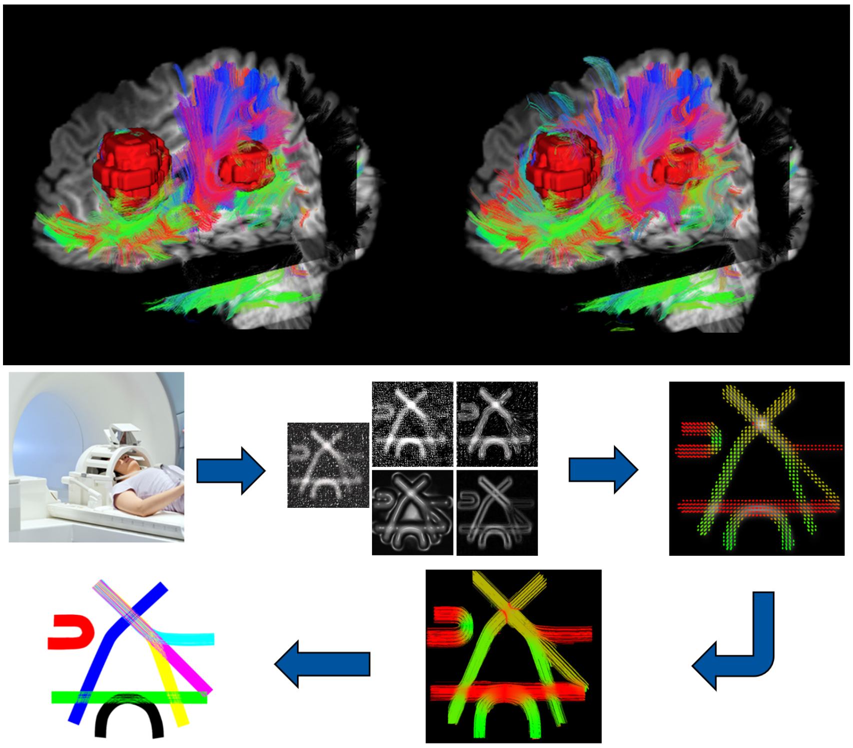


Figure 1 Visualizations associated with our work on deep-learning based analysis of diffusion MRI data.

3.10 Deep-Learning based Analysis of Diffusion MRI Data

Dorit Merhof (Universität Regensburg, DE)

Joint work of Dorit Merhof, Leon Weninger, Simon Koppers, Chuh-Hyoun Na, Kerstin Jütten, Hans Clusmann, Veren Mainz, Seigfried Gauggel, Ferdinand Binkofski, Martin Wiesmann, Christoph Haarburger, Sandro Romanzetti, Julia Ebert, Kathrin Reetz, Mushawar Ahmad, Lu Bloy

Main reference Leon Weninger, Chuh-Hyoun Na, Kerstin Jütten, Dorit Merhof: "Analyzing the effects of free water modeling by deep learning on diffusion MRI structural connectivity estimates in glioma patients", PLOS ONE, Vol. 15(9), pp. 1–17, Public Library of Science, 2020.

URL http://dx.doi.org/10.1371/journal.pone.0239475

Artificial Intelligence approaches, and especially recent Deep Learning techniques, have shown to outperform conventional image processing algorithms in many medical image analysis scenarios.

This presentation will present Deep Learning approaches for Diffusion MRI Data for (1) diffusion signal augmentation [1], (2) free water correction [6, 2, 4] and (3) signal harmonization [5, 6, 7].

Finally, limitations of neuronal networks as well as current challenges and trends in Deep Learning will be discussed.

References

1 Simon Koppers, Christoph Haarburger and Dorit Merhof: Diffusion MRI Signal Augmentation – From Single Shell to Multi Shell with Deep Learning. In: MICCAI Workshop on Computational Diffusion MRI (CDMRI), 2016.

2 Leon Weninger, Simon Koppers, Chuh-Hyoun Na, Kerstin Juetten and Dorit Merhof: Free-Water Correction in Diffusion MRI: A Reliable and Robust Learning Approach. In: MICCAI Workshop on Computational Diffusion MRI (CDMRI), 2019.

3 Leon Weninger, Chuh-Hyoun Na, Kerstin Jütten and Dorit Merhof: Analyzing the effects of free water modeling by deep learning on diffusion MRI structural connectivity estimates in glioma patients. In: PLOS ONE 15 (9), 2020.

4 Kerstin Jütten, Leon Weninger, Verena Mainz, Siegfried Gauggel, Ferdinand Binkofski, Martin Wiesmann, Dorit Merhof, Hans Clusmann and Chuh-Hyoun Na: Dissociation of structural and functional connectomic coherence in glioma patients. In: Scientific Reports 11 (16790), 2021.

5 Simon Koppers, Luke Bloy, Jeffrey I. Berman, Chantal M.W. Tax, J. Christopher Edgar and Dorit Merhof: Spherical Harmonic Residual Network for Diffusion Signal Harmonization. In: MICCAI Workshop on Computational Diffusion MRI (CDMRI), 2018.

6 Leon Weninger, Sandro Romanzetti, Julia Ebert, Kathrin Reetz and Dorit Merhof: Harmonization of diffusion MRI data obtained with multiple head coils using hybrid CNNs. In: ECCV AIMIA Workshop, 2022.

7 Leon Weninger, Mushawar Ahmad and Dorit Merhof: From supervised to unsupervised harmonization of diffusion MRI acquisitions. In: IEEE International Symposium on Biomedical Imaging (ISBI), 2022.

3.11 Federated Learning and Reproducibility in Healthcare

Sarthak Pati (University of Pennsylvania, US)

Joint work of Sarthak Pati, Spyridon Bakas

Real-world applicability of artificial intelligence (AI) in the clinical setting [39, 40, 41] is hampered by the *i)* lack of available diverse (training and validation) data affecting the robustness and generalizability of AI models towards unseen/unknown population groups, and *ii)* limitations on defining reproducible computational pipelines for local hardware resources at clinical sites.

The current paradigm towards sufficiently large and diverse data for training and validating AI models is via centralization of data from multiple institutions [29, 30, 32, 33, 31, 17, 6, 49, 50, 45, 46, 47, 48, 51]. However, this paradigm faces limitations when it comes to scale due to various legal, regulatory, cultural, and ownership concerns [8, 9]. Federated Learning (FL) offers an alternative paradigm to train robust AI models and a potential solution to the data sharing hurdles, as demonstrated in multiple simulated [8, 9, 2, 43] and real-world studies [1]. Furthermore, beyond training robust AI models, the evaluation of their effectiveness and durability over time on real-world patient data from large and diverse population demographics poses another challenge towards their clinical translation. Federated evaluation (FE) studies through persistent data registries and streamlined workflows may provide a solution on such performance evaluations, obviating the need of data sharing. Together, federated ***learning*** and ***evaluation*** form complementary mechanisms to generate meaningful clinical impact by enabling access to data silos in a way that is compatible with regulations and cultural concerns.

There have been numerous community-driven efforts to provide either common definitions towards results' reproducibility [13, 14, 16, 29, 30, 32, 33, 31, 17, 49, 50, 45, 46, 47, 48, 51], or common benchmarking environments (i.e., challenges) for fair AI model evaluation [15].

Although a substantial number of closed-source and commercial solutions achieve clinical reproducibility [42], having widely available, community-driven, and well-documented open-source projects [18, 19, 20, 21, 22, 23, 34, 7] that focus on the *reproducibility* of research, while being driven by the clinical stakeholders would be critical towards ensuring that cutting edge scientific breakthroughs make it for clinical validation sooner. This further allows computational scientists to explore their methodological interests while allowing clinical partners to deploy these methods in an easy manner in their *existing* hardware infrastructure.

Our collaborative group has collectively produced open-source publicly-available software solutions to address this space. Starting with the largest real-world FL study to-date (the Federated Tumor Segmentation (FeTS) initiative)[1], which also describes the largest reported study on the rare cancer of glioblastoma, involving data of 6,314 patients from 71 institutions across 6 continents [1]. The tool used by the FeTS Initiative has been open-sourced as "The FeTS Tool" [4], which provides an end-to-end point solution for studies related to brain tumor boundary detection/segmentation. This solution includes all the required computational steps, starting from data curation, anonymization, brain extraction (also known as skull-stripping [35, 34]), to pre-processing, generation of baseline automated annotations leveraging methods considered state of the art [53, 54, 55], an interface to manually refine these automated annotations and sign off ground truth labels [18, 19, 20], as well as to allow a user to either train their AI model or join an existing FL study. Moreover, the FL component of the FeTS tool is enabled by the Open Federated Learning (OpenFL) library [11, 10], which is designed for general-purpose FL and being agnostic to use-case and framework. Further to the FeTS initiative, OpenFL has facilitated studies on the i) effect of cosmic radiation on astronauts by the Frontier Development Lab (FDL) of the National Aeronautics and Space Administration (NASA), and ii) prediction of respiratory distress syndrome and death for COVID-19 patients by the 11 sites of the Montefiore Health System in New York.

Building upon the collaborative network of the FeTS initiative, we further conducted the first-ever computational challenge in FL, which happened in conjunction with the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2021 and 2022 [5], and followed the principle of clinical trials [52]. The focus of the FeTS challenge was two-fold: *i)* the development of aggregation methods for FL, and *ii)* the federated evaluation of brain tumor segmentation algorithms in-the-wild, by circulating AI models on unseen data from multiple sites of the FeTS initiative collaborative network. The FeTS challenge was orchestrated by MedPerf [24], in which the challenge organizers initiated the design of the study, the collaborating sites registered information about their datasets, and the AI models of the challenge participants were described as independent experiments evaluated against these datasets. Finally, towards broader clinical workflows, we developed the Generally Nuanced Deep Learning Framework (GaNDLF) [12], which enables users to design and manage AI algorithms for multiple tasks and various data/organ/modality types, such as segmentation on brain tumor MRI [2, 1], breast mammograms [37, 36] & dynamic contrast enhanced MRI [6], as well as classification on histology whole slide images [3], RGB images [38], & breast mammograms [43]. The wide applicability and obtained results showcase the generalizability of GaNDLF. Additionally, GaNDLF offers automated post-training optimization of AI models [56, 44], allowing their execution/inference on consumer-grade computers without requiring specialized hardware, such as deep learning acceleration cards.

[1] www.fets.ai

In conclusion, there is a need to i) assess the generalizability of AI models by capturing ample patient demographics, ii) address bias and inequities in AI, especially those related to underserved/underrepresented patient populations, and iii) on the continuous monitoring of AI models requiring further developments in automated quality control, monitoring of drift & bias, and model calibration. Towards fulfilling these goals, the open federated ecosystem consisting of GaNDLF [12], OpenFL [11], and MedPerf [24] provide a holistic end-to-end open-sourced federated learning and evaluation solution that supports multiple data types, and that be easily used by both experienced and novice researchers.

Acknowledgements

Research reported in this publication was partly supported by the National Cancer Institute (NCI) of the National Institutes of Health (NIH) under award number U01CA242871. The content of this publication is solely the responsibility of the authors and does not represent the official views of the NIH or any other funding body.

References

1 Pati, S., Baid, U., Edwards, B., Sheller, M., Wang, S., Reina, G., Foley, P., Gruzdev, A., Karkada, D., Davatzikos, C. et al. Federated Learning Enables Big Data for Rare Cancer Boundary Detection.*Nature Communications*. 13 (2022)

2 Baid, U., Pati, S., Thakur, S., Edwards, B., Sheller, M., Martin, J. & Bakas, S. The Federated Tumor Segmentation (FeTS) Initiative: The First Real-World Large-Scale Data-Private Collaboration Focusing On Neuro-Oncology. Neuro-Oncology. 23 pp. 135-135 (2021)

3 Baid, U., Pati, S., Kurc, T., Gupta, R., Bremer, E., Abousamra, S., Thakur, S., Saltz, J. and Bakas, S. Federated learning for the classification of tumor infiltrating lymphocytes. ArXiv Preprint ArXiv:2203.16622. (2022)

4 Pati, S., Baid, U., Edwards, B., Sheller, M., Foley, P., Reina, G., Thakur, S., Sako, C., Bilello, M., Davatzikos, C. et al. The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research. Physics In Medicine & Biology. 67, 204002 (2022)

5 Pati, S., Baid, U., Zenk, M., Edwards, B., Sheller, M., Reina, G., Foley, P., Gruzdev, A., Martin, J., Albarqouni, S. et al. The Federated Tumor Segmentation (FeTS) Challenge. ArXiv Preprint ArXiv:2105.05874. (2021)

6 Chitalia, R., Pati, S., Bhalerao, M., Thakur, S., Jahani, N., Belenky, V., McDonald, E., Gibbs, J., Newitt, D., Hylton, N. et al. Expert tumor annotations and radiomics for locally advanced breast cancer in DCE-MRI for ACRIN 6657/I-SPY1.*Scientific Data*. 9, 440 (2022).

7 Bounias, D., Singh, A., Bakas, S., Pati, S., Rathore, S., Akbari, H., Bilello, M., Greenberger, B., Lombardo, J., Chitalia, R. et al. Interactive machine learning-based multi-label segmentation of solid tumors and organs. Applied Sciences. 11, 7488 (2021)

8 Sheller, M., Reina, G., Edwards, B., Martin, J. and Bakas, S. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. International MICCAI Brainlesion Workshop. pp. 92-104 (2018)

9 Sheller, M., Edwards, B., Reina, G., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. et al. Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. Scientific Reports. 10, 1-12 (2020)

10 Reina, G., Gruzdev, A., Foley, P., Perepelkina, O., Sharma, M., Davidyuk, I., Trushkin, I., Radionov, M., Mokrov, A., Agapov, D. et al. OpenFL: An open-source framework for Federated Learning. ArXiv Preprint ArXiv:2105.06413. (2021)

11 Foley, P., Sheller, M., Edwards, B., Pati, S., Riviera, W., Sharma, M., Moorthy, P., Wang, S., Martin, J., Mirhaji, P. et al. OpenFL: the open federated learning library. Physics In Medicine & Biology. 67, 214001 (2022)

12 Pati, S., Thakur, S., Bhalerao, M., Baid, U., Grenko, C., Edwards, B., Sheller, M., Agraz, J., Baheti, B., Bashyam, V. et al. Gandlf: A generally nuanced deep learning framework for scalable end-to-end clinical workflows in medical imaging. ArXiv Preprint ArXiv:2103.01006. (2021)

13 Zwanenburg, A., Vallières, M., Abdalah, M., Aerts, H., Andrearczyk, V., Apte, A., Ashrafinia, S., Bakas, S., Beukinga, R., Boellaard, R. et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology. 295, 328-338 (2020)

14 McNitt-Gray, M., Napel, S., Jaggi, A., Mattonen, S., Hadjiiski, L., Muzi, M., Goldgof, D., Balagurunathan, Y., Pierce, L., Kinahan, P. et al. Standardization in quantitative imaging: A multicenter comparison of radiomic features from different software packages on digital reference objects and patient data sets. Tomography. 6, 118-128 (2020)

15 Eisenmann, M., Reinke, A., Weru, V., Tizabi, M., Isensee, F., Adler, T., Godau, P., Cheplygina, V., Kozubek, M., Ali, S. et al. Biomedical image analysis competitions: The state of current participation practice. ArXiv Preprint ArXiv:2212.08568. (2022)

16 Pati, S., Verma, R., Akbari, H., Bilello, M., Hill, V., Sako, C., Correa, R., Beig, N., Venet, L., Thakur, S. et al. Reproducibility analysis of multi-institutional paired expert annotations and radiomic features of the Ivy Glioblastoma Atlas Project (Ivy GAP) dataset. Medical Physics. 47, 6039-6052 (2020)

17 Borovec, J., Kybic, J., Arganda-Carreras, I., Sorokin, D., Bueno, G., Khvostikov, A., Bakas, S., Eric, I., Chang, C., Heldmann, S. et al. ANHIR: automatic non-rigid histological image registration challenge. IEEE Transactions On Medical Imaging. 39, 3042-3052 (2020)

18 Davatzikos, C., Rathore, S., Bakas, S., Pati, S., Bergman, M., Kalarot, R., Sridharan, P., Gastounioti, A., Jahani, N., Cohen, E. et al. Cancer imaging phenomics toolkit: Quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. Journal Of Medical Imaging. 5, 011018 (2018)

19 Pati, S., Singh, A., Rathore, S., Gastounioti, A., Bergman, M., Ngo, P., Ha, S., Bounias, D., Minock, J., Murphy, G. et al. The Cancer Imaging Phenomics Toolkit (CaPTk): Technical Overview. International MICCAI Brainlesion Workshop. pp. 380-394 (2019)

20 Rathore, S., Bakas, S., Pati, S., Akbari, H., Kalarot, R., Sridharan, P., Rozycki, M., Bergman, M., Tunc, B., Verma, R. et al. Brain cancer imaging phenomics toolkit (brain-CaPTk): an interactive platform for quantitative analysis of glioblastoma. International MICCAI Brainlesion Workshop. pp. 133-145 (2017)

21 Fathi Kazerooni, A., Akbari, H., Shukla, G., Badve, C., Rudie, J., Sako, C., Rathore, S., Bakas, S., Pati, S., Singh, A. et al. Cancer imaging phenomics via CaPTk: Multi-institutional prediction of progression-free survival and pattern of recurrence in glioblastoma. JCO Clinical Cancer Informatics. 4 pp. 234-244 (2020)

22 Wolf, I., Vetter, M., Wegner, I., Nolden, M., Bottger, T., Hastenteufel, M., Schobinger, M., Kunert, T. and Meinzer, H. The medical imaging interaction toolkit (MITK): a toolkit facilitating the creation of interactive software by extending VTK and ITK. Medical Imaging 2004: Visualization, Image-Guided Procedures, And Display. 5367 pp. 16-27 (2004)

23 Kikinis, R., Pieper, S. and Vosburgh, K. 3D Slicer: a platform for subject-specific image analysis, visualization, and clinical support. Intraoperative Imaging And Image-guided Therapy. pp. 277-289 (2014)

24 Karargyris, A., Umeton, R., Sheller, M., Aristizabal, A., George, J., Bala, S., Beutel, D., Bittorf, V., Chaudhari, A., Chowdhury, A. et al. MedPerf: Open Benchmarking Platform for Medical Artificial Intelligence using Federated Evaluation. ArXiv Preprint ArXiv:2110.01406. (2021)

25 Beam, A., Manrai, A. and Ghassemi, M. Challenges to the reproducibility of machine learning models in health care. Jama. 323, 305-306 (2020)

26 Belbasis, L. and Panagiotou, O. Reproducibility of prediction models in health services research. BMC Research Notes. 15, 1-5 (2022)

27 Benjamin Haibe-Kains, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, Massive Analysis Quality Control Society Board of Directors, Levi Waldron, Bo Wang, Chris McIntosh, Anna Goldenberg, Anshul Kundaje, Casey S. Greene, Tamara Broderick, Michael M. Hoffman, Jeffrey T. Leek, Keegan Korthauer, Wolfgang Huber, Alvis Brazma, Joelle Pineau, Robert Tibshirani, Trevor Hastie, John P. A. Ioannidis, John Quackenbush and Hugo J. W. L. Aerts. Transparency and reproducibility in artificial intelligence. Nature. 586, E14-E16 (2020)

28 Carter, R., Attia, Z., Lopez-Jimenez, F. and Friedman, P. Pragmatic considerations for fostering reproducible research in artificial intelligence. NPJ Digital Medicine. 2, 1-3 (2019)

29 Menze, B., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R. et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions On Medical Imaging. 34, 1993-2024 (2014)

30 Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R., Berger, C., Ha, S., Rozycki, M. et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. ArXiv Preprint ArXiv:1811.02629. (2018)

31 Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F., Pati, S. et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. ArXiv Preprint ArXiv:2107.02314. (2021)

32 Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K. and Davatzikos, C. Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. The Cancer Imaging Archive. 286 (2017)

33 Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K. and Davatzikos, C. Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. The Cancer Imaging Archive. 286 (2017)

34 Thakur, S., Doshi, J., Pati, S., Rathore, S., Sako, C., Bilello, M., Ha, S., Shukla, G., Flanders, A., Kotrotsou, A. et al. Brain extraction on MRI scans in presence of diffuse glioma: Multi-institutional performance evaluation of deep learning methods and robust modality-agnostic training. NeuroImage. 220 pp. 117081 (2020)

35 Thakur, S., Doshi, J., Pati, S., Ha, S., Sako, C., Talbar, S., Kulkarni, U., Davatzikos, C., Erus, G. and Bakas, S. Skull-stripping of glioblastoma MRI scans using 3D deep learning. Brainlesion: Glioma, Multiple Sclerosis, Stroke And Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held In Conjunction With MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I. pp. 57-68 (2020)

36 Ahluwalia, V., Mankowski, W., Pati, S., Bakas, S., Brooks, A., Vachon, C., Conant, E., Gastounioti, A. and Kontos, D. Artificial-intelligence-driven volumetric breast density estimation with digital breast tomosynthesis in a racially diverse screening cohort.. (American Society of Clinical Oncology,2022)

37 Ahluwalia, V., Mankowski, W., Pati, S., Bakas, S., Brooks, A., Vachon, C., Conant, E., Gastounioti, A. and Kontos, D. Deep-learning-enabled volumetric breast density estimation with digital breast tomosynthesis. Cancer Research. 82, 1929-1929 (2022)

38 Güley, O., Pati, S. and Bakas, S. Classification of Infection and Ischemia in Diabetic Foot Ulcers Using VGG Architectures. Diabetic Foot Ulcers Grand Challenge. pp. 76-89 (2021)

39 Giger, M. Machine learning in medical imaging. Journal Of The American College Of Radiology. 15, 512-520 (2018)

40 Kelly, C., Karthikesalingam, A., Suleyman, M., Corrado, G. and King, D. Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine. 17 pp. 1-9 (2019)

41 Ashton, J., Young, A., Johnson, M. and Beattie, R. Using machine learning to impact on long-term clinical care: principles, challenges, and practicalities. Pediatric Research. pp. 1-10 (2022)

42 Rexilius, J., Spindler, W., Jomier, J., König, M., Hahn, H., Link, F. and Peitgen, H. A framework for algorithm evaluation and clinical application prototyping using ITK. Insight Journal—MICCAI Open-Source Workshop. (2005)

43 Katti, K., Muthukrishnan, R., Heyler, A., Pati, S., Alahari, A., Sanborn, M., Conant, E., Scott, C., Winham, S., Vachon, C. et al. MammoDL: Mammographic Breast Density Estimation using Federated Learning. ArXiv Preprint ArXiv:2206.05575. (2022)

44 Thakur, S., Pati, S., Panchumarthy, R., Karkada, D., Wu, J., Kurtaev, D., Sako, C., Shah, P. and Bakas, S. Optimization of Deep Learning Based Brain Extraction in MRI for Low Resource Environments. International MICCAI Brainlesion Workshop. pp. 151-167 (2022)

45 Baheti, B., Waldmannstetter, D., Chakrabarty, S., Akbari, H., Bilello, M., Wiestler, B., Schwarting, J., Calabrese, E., Rudie, J., Abidi, S. et al. The brain tumor sequence registration challenge: establishing correspondence between pre-operative and follow-up MRI scans of diffuse glioma patients. ArXiv Preprint ArXiv:2112.06979. (2021)

46 Simpson, A., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B., Litjens, G., Menze, B. et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. ArXiv Preprint ArXiv:1902.09063. (2019)

47 Bilic, P., Christ, P., Li, H., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G., Chartrand, G. et al. The liver tumor segmentation benchmark (lits). Medical Image Analysis. 84 pp. 102680 (2023)

48 Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B., Litjens, G., Menze, B., Ronneberger, O., Summers, R. et al. The medical segmentation decathlon. Nature Communications. 13, 4128 (2022)

49 Davatzikos, C., Barnholtz-Sloan, J., Bakas, S., Colen, R., Mahajan, A., Quintero, C., Capellades Font, J., Puig, J., Jain, R., Sloan, A. et al. AI-based prognostic imaging biomarkers for precision neuro-oncology: the ReSPOND consortium. Neuro-oncology. 22, 886-888 (2020)

50 Bakas, S., Ormond, D., Alfaro-Munoz, K., Smits, M., Cooper, L., Verhaak, R. and Poisson, L. iGLASS: imaging integration into the Glioma Longitudinal Analysis Consortium. Neuro-oncology. 22, 1545-1546 (2020)

51 Dorent, R., Kujawa, A., Ivory, M., Bakas, S., Rieke, N., Joutard, S., Glocker, B., Cardoso, J., Modat, M., Batmanghelich, K. et al. CrossMoDA 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. Medical Image Analysis. 83 pp. 102628 (2023)

52 Larson, D., Harvey, H., Rubin, D., Irani, N., Justin, R. and Langlotz, C. Regulatory frameworks for development and evaluation of artificial intelligence–based diagnostic imaging algorithms: summary and recommendations. Journal Of The American College Of Radiology. 18, 413-424 (2021)

53 Kamnitsas, K., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Rueckert, D. and Glocker, B. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical Image Analysis. 36 pp. 61-78 (2017)

54 McKinley, R., Meier, R. and Wiest, R. Ensembles of densely-connected CNNs with label-uncertainty for brain tumor segmentation. International MICCAI Brainlesion Workshop. pp. 456-465 (2018)

55 Isensee, F., Jaeger, P., Kohl, S., Petersen, J. and Maier-Hein, K. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods. 18, 203-211 (2021)

56 Baheti, B., Thakur, S., Pati, S., Karkada, D., Panchumarthy, R., Wu, J., Mohan, S., Nasrallah, M., Shah, P. and Bakas, S. Optimization of artificial intelligence algorithms for low-resource/clinical environments: Focus on clinically-relevant glioma region delineation. Neuro-Oncology. 24 pp. 167-167 (2022)

3.12 Coupling Brain Tumor Models and Image Registration

Miriam Schulte (Universität Stuttgart, DE)

Joint work of Miriam Schulte, George Biros, Malte Brunn, Amir Gholami, Andreas Mang, Klaudius Scheufele, Shashank Subramanian
Main reference Klaudius Scheufele, Andreas Mang, Amir Gholami, Christos Davatzikos, George Biros and Miriam Mehl: "Coupling brain-tumor biophysical models and diffeomorphic image registration". Computer Methods in Applied Mechanics and Engineering, Volume 347, pages 533-567, 2019.
URL https://doi.org/10.18419/opus-10396

We present computational coupling of inverse tumor simulation and diffeomorphic image registration that allows to achieve two tasks that can be relevant for clinicians: (i) registration of a healthy statistical atlas brain to a patient brain with tumor in order to transfer labels and brain region boundaries; (ii) identification of tumor growth parameters such as diffusion and reaction rates or initial tumor. For both tasks, we have to solve a combined inverse problem involving image registration and the tumor model to 'move' from an atlas image to a patient images with a tumor. We present various ways to achieve this by combining separate registration and tumor solvers in [1, 2]. More details on the single components are presented in [3] for the tumor growth inversion and in [4] for image registration. Both software components show very good scalability on high performance computing hardware such that we can solve problems at 256^3 resolution in a couple of minutes.

For a glance at more general concepts for coupling of two or more computational components, refer to [5] and [1].

References

1 Klaudius Scheufele. Coupling Schemes and Inexact Newton for Multi-Physics and Coupled Optimization Problems, Doctoral Thesis, University of Stuttgart, Germany, 2019, `https://doi.org/10.18419/opus-10396`.

2 Klaudius Scheufele, Andreas Mang, Amir Gholami, Christos Davatzikos, George Biros and Miriam Mehl. Coupling brain-tumor biophysical models and diffeomorphic image registration. Computer Methods in Applied Mechanics and Engineering, Volume 347, pages 533-567, 2019, `https://doi.org/10.1016/j.cma.2018.12.008`.

3 Shashank Subramanian, Klaudius Scheufele, Miriam Mehl and George Biros. Where did the tumor start? An inverse solver with sparse localization for tumor growth models. IOP Publishing Ltd, Inverse Problems, Volume 36, Number 4, 2020, `https://doi.org/10.1088/1361-6420/ab649c`.

4 Malte Brunn, Naveen Himthani, George Biros, Miriam Mehl and Andreas Mang. Multi-Node Multi-GPU Diffeomorphic Image Registration for Large-Scale Imaging Problems. SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, Atlanta, GA, USA, 2020, pp. 1-17, `https://doi.org/10.1109/SC41405.2020.00042`.

5 Gerasimos Chourdakis, Kyle Davis, Benjamin Rodenberg, Miriam Schulte, Frédéric Simonis, Benjamin Uekermann, Georg Abrams, Hans-Joachim Bungartz, Lucia Cheung Yau, Ishaan Desai, Konrad Eder, Richard Hertrich, Florian Lindner, Alexander Rusch, Dmytro Sashko, David Schneider, Amin Totounferoush, Dominik Volland, Peter Vollmer and Oguz Ziya Koseomur. preCICE v2: A sustainable and user-friendly coupling library. Open Res Europe 2022, `https://doi.org/10.12688/openreseurope.14445.2`.

3.13 Generative Models for Generalizable and Interpretable Analysis of Brain Tumor Images

Koen Van Leemput (Martinos Center – Charlestown, US)

Main reference Mikael Agn, Per Munck af Rosenschöld, Oula Puonti, Michael J. Lundemann, Laura Mancini, Anastasia Papadaki, Steffi Thust, John Ashburner, Ian Law, Koen Van Leemput: "A modality-adaptive method for segmenting brain tumors and organs-at-risk in radiation therapy planning", Medical Image Analysis, Vol. 54, pp. 220–237, 2019.
URL http://dx.doi.org/10.1016/j.media.2019.03.005

In my talk I will discuss the use of *generative models* for generalizable and interpretable analysis of brain tumor images. Specifically, I will highlight the fundamental differences that exist between analyzing tightly-standardized images in well-controlled group studies, vs. analyzing images acquired "in the wild", i.e., as part of the clinical treatment of brain diseases. I will present our work on generative models that can naturally extrapolate beyond the narrow characteristics of manually labeled training data, and how these techniques are implemented within the well-known open-source software suite FreeSurfer. Specific attention will be paid to modeling lesions (such as white matter lesions or brain tumors) within whole-brain segmentation settings, and to leveraging the temporal consistency between follow-up scans in longitudinal data. Time permitting, I will also touch on the need for *interpretable* image prediction models, where the generative aspect encodes the *causal* effect of disease on brain shape. Such models are much easier to interpret and explain to clinicians than the "black box" discriminative methods that are often used for predicting diagnoses or disease scores from images.

3.14 A Clinical and Biological Validation Study of a Tumor Growth Model

Benedikt Wiestler (TU München – Klinikum rechts der Isar, DE)

Joint work of Marie-Christin Metz, Ivan Ezhov, Lucas Zimmer, Jan C. Peeken, Josef A. Buchner, Jana Lipkova, Florian Kofler, Diana Waldmannstetter, Claire Delbridge, Christian Diehl, Denise Bernhardt, Friederike Schmidt-Graf, Jens Gempt, Stephanie E. Combs, Claus Zimmer, Bjoern Menze, Benedikt Wiestler
Main reference Marie-Christin Metz, Ivan Ezhov, Lucas Zimmer, Jan C. Peeken, Josef A. Buchner, Jana Lipkova, Florian Kofler, Diana Waldmannstetter, Claire Delbridge, Christian Diehl, Denise Bernhardt, Friederike Schmidt-Graf, Jens Gempt, Stephanie E. Combs, Claus Zimmer, Bjoern Menze, Benedikt Wiestler:" Towards Image – Based Personalization of Glioblastoma Therapy A Clinical and Biological Validation Study of a Novel, Deep Learning – Driven Tumor Growth Model". Preprint (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-2262631/v1]
URL https://doi.org/10.21203/rs.3.rs-2262631/v1

The diffuse growth pattern of glioblastoma is one of the main challenges for improving patient survival. Computational tumor growth modeling has emerged as a promising tool to infer tumor cell distribution and thereby guide personalized therapy.

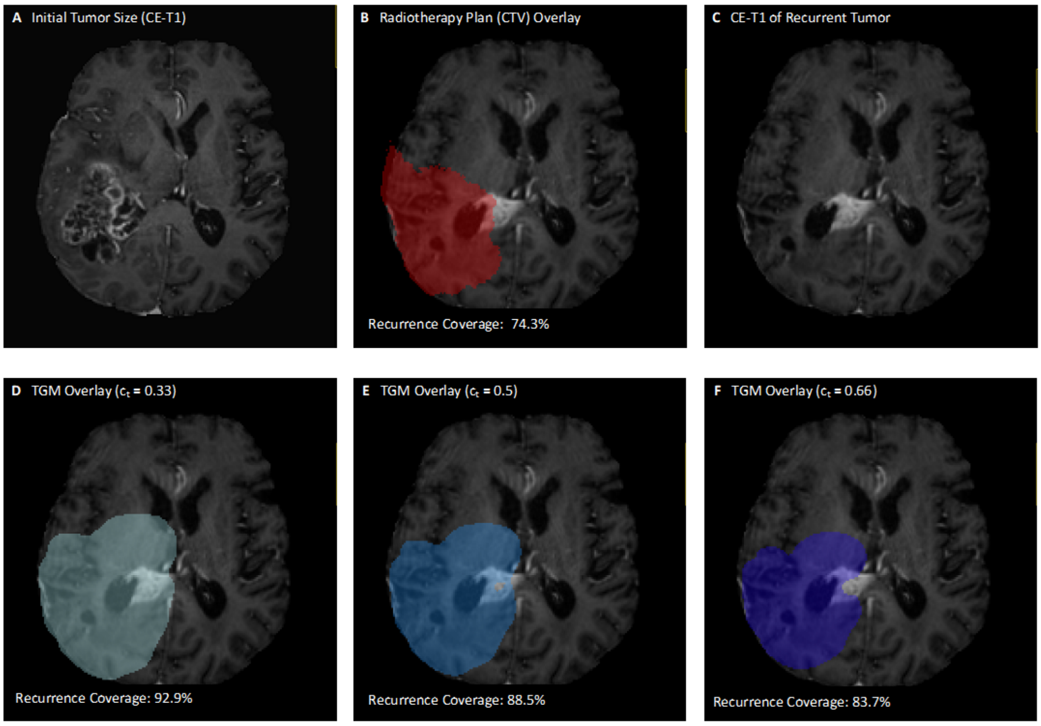

A Initial Tumor Size (CE-T1)
B Radiotherapy Plan (CTV) Overlay
Recurrence Coverage: 74.3%
C CE-T1 of Recurrent Tumor
D TGM Overlay (c_t = 0.33)
Recurrence Coverage: 92.9%
E TGM Overlay (c_t = 0.5)
Recurrence Coverage: 88.5%
F TGM Overlay (c_t = 0.66)
Recurrence Coverage: 83.7%

Figure 2 Comparison of standard clinical target volume (CTV) and computed target delineations derived from isolines of different estimated tumor cell densities by the tumor growth model (TGM). Underlying images are contrast-enhanced T1 (CE-T1).

In [1], we performed clinical and biological validation of a novel, deep learning – based growth model [2], aiming to close the gap between the experimental state and clinical implementation. In more detail, we wanted to investigate how well this Fisher-Kolmogorov model correlates with (i) tumor biology, (ii) survival and – perhaps most importantly – (iii) location of tumor recurrence.

To answer these questions, we included a total of three data sets into our study. For (i) and (ii), we analysed 124 patients from The Cancer Genome Archive network and 397 patients from the UCSF glioma MRI data set for correlations between clinical data, genetic pathway activation maps (generated with PARADIGM; TCGA only), and infiltration (Dw) as well as proliferation (ρ) parameters stemming from a Fisher-Kolmogorov growth model adjusted to the patients' preoperative images [2]. To address (iii), we correlated later tumor recurrence in an in-house data set with 30 glioblastoma patients with radiotherapy plans and growth model-derived tumor cell distribution.

Interestingly, we observed a significant correlation between 11 signaling pathways that are associated with proliferation, and the estimated proliferation parameter ρ. The parameter ratio Dw/ρ ($p<0.05$ in TCGA) as well as the simulated tumor volume ($p<0.05$ in both TCGA and UCSF) were significantly inversely correlated with overall survival in Cox survival modeling. Depending on the cutoff value for tumor cell density, we observed a significant improvement of recurrence coverage without significantly increased radiation volume utilizing model-derived target volumes instead of standard radiation plans (example shown in figure 2).

Identifying a significant correlation between computed growth parameters, and clinical and biological data, we highlight the potential of tumor growth modeling for individualized therapy of glioblastoma. This holds promise to improve accuracy of personalized radiation planning in the near future. Future research directions include more complex growth models (e.g., including necrosis or mass effect), including imaging information for model calibration, and ultimately also going from global to local modeling, explicitly incorporating tumor heterogeneity.

References

1 Metz M, Ezhov I, Zimmer L, Peeken JC, Buchner JA, Lipkova J, Kofler F, Waldmannstetter D, Delbridge C, Diehl C, Bernhardt D, Schmidt-Graf F, Gempt J, Combs SE, Zimmer C, Menze B and Wiestler B. Towards Image – Based Personalization of Glioblastoma Therapy A Clinical and Biological Validation Study of a Novel, Deep Learning – Driven Tumor Growth Model. Preprint (Version 1) available at Research Square https://doi.org/10.21203/rs.3.rs-2262631/v1

2 Ezhov I, Scibilia K, Franitza K, Steinbauer F, Shit S, Zimmer L, Lipkova J, Kofler F, Paetzold JC, Canalini L, Waldmannstetter D, Menten MJ, Metz M, Wiestler B and Menze B. Learn-Morph-Infer: A new way of solving the inverse problem for brain tumor modeling. Med Image Anal. 2023 Jan;83:102672. https://doi.org/10.1016/j.media.2022.102672.

3.15 Conceptual mathematical tumor models on different scales

Barbara Wohlmuth (TU München – Garching, DE)

Joint work of Barbara Wohlmuth, Tobias Duswald, Marvin Fritz, Tobias Köppl, T. Oden, Andreas Wagner

Tumor simulations require complex multiscale models ranging from discrete agent-based models to continuum models, with various hybrid type models in between [1]. Extremely small scales require an agent-based formulation for the tumor and the capillaries, where only signaling molecules, drugs, and nutrients are described by continuous fields [2]. Larger tumors inside rat brains might be resolved with a continuous phase field approach, where still the capillary flow is described by 1D-models, and their growth is modeled by a rule-based algorithm [5]. On the macro scale, the capillaries might be further simplified to a porous medium, requiring only the resolution of larger vessels by 1D-models for breast tumors [7] or 2D surface sources in the lung [8]. Often a problem-dependent coupling is required to achieve a biologically meaningful value range, particularly for the pressure of the 1D blood flow where 0D models have to damp down oscillations.

Besides the choice between discrete and continuum approaches, there remains the question of which biophysical mechanisms are considered relevant for the application at hand and thus have to be modeled, often leading to increasingly complex models of various species. The tumor typically consists of necrotic, hypoxic, and proliferative cell species. For agent-based models, the latter might be further divided into the Q, G1, SG2 subspecies [2]. Further, matrix degenerative enzymes acting on the extra-cellular matrix might be added [3]. The nutrient field might be split up into various porous media resulting in double continuum models [7, 8]. For angiogenesis, vascular endothelial growth factors have to be included [2, 5]. Often mechanical deformations have to be considered [6], and, depending on the clinical therapy, one or more drug species have to be included [6, 8].

All these modeling choices lead to complex, heavily-coupled, nonlinear models, which pose mathematical challenges to the analysis of their well-posedness [3, 4], to the creation of stable numerical schemes and efficient decoupled solvers [5].

In a second, more applied and challenging step, these models have to be calibrated against real-world data [9] and verified against clinical measurements [8]. Here, the amount of data is often the bottleneck and requires a strong multidisciplinary effort to acquire, evaluate and interpret. Especially the derivation of generally accepted benchmark problems for model validation would be extremely valuable for future work.

References

1 Lowengrub, John S., et al. "Nonlinear modelling of cancer: bridging the gap between cells and tumours." Nonlinearity 23.1 (2009): R1.

2 Duswald, Tobias, et al. "A Hybrid PDE and Agent Based Model for Cancer Simulations" (preliminary title). in preparation.

3 Fritz, Marvin, et al. "Local and nonlocal phase-field models of tumor growth and invasion due to ECM degradation." Mathematical Models and Methods in Applied Sciences 29.13 (2019): 2433-2468.

4 Fritz, Marvin, et al. "Analysis of a new multispecies tumor growth model coupling 3D phase-fields with a 1D vascular network." Nonlinear Analysis: Real World Applications 61 (2021): 103331.

5 Fritz, Marvin, et al. "Modeling and simulation of vascular tumors embedded in evolving capillary networks." CMAME (2021).

6 Fritz, Marvin, et al. "On a subdiffusive tumour growth model with fractional time derivative." IMA Journal of Applied Mathematics 86.4 (2021): 688-729.

7 Fritz, Marvin, et al. "A 1D–0D–3D coupled model for simulating blood flow and transport processes in breast tissue." IJNMBE (2022).

8 Fritz, Marvin, et al. "A phase-field model for non-small cell lung cancer under the effects of immunotherapy." to be submitted.

9 Lima, E. A. B. F., et al. "Calibration of multi-parameter models of avascular tumor growth using time resolved microscopy data." Scientific reports 8.1 (2018): 14558.

4 Panel discussions

4.1 Working Groups and Panel Discussions

Andreas Mang (University of Houston, US), George Biros (Univ. of Texas at Austin, US), Björn H. Menze (Universität Zürich, CH), and Miriam Schulte (Universität Stuttgart, DE)

As mentioned in the executive summary, the scientific presentations were followed by a brief discussion about selected topics in two working groups to identify immediate goals and further discuss existing challenges. The first group included researchers with a key interest in designing methods to analyze medical (imaging) data and integrate mathematical and computational methods with imaging and medical data. The second group discussed topics associated with the design of mathematical and computational methods for inference, simulation, and optimization. Below we list the key findings in these two groups and some of the questions that remain open.

Working Group 1

We summarize the main topics discussed in the *first working group* below:

One of the questions discussed during our meeting was if it is possible to curate a database of (publicly) available data for model validation on unseen data in both machine learning and classical modeling. Several questions arose in this context. For example: What are the quality requirements for the data and what datasets are already available? (What are the resolution requirements? How do we deal with medical imaging artifacts?) What are the most pertinent/viable applications that this database is generated from? Do we only include/want longitudinal data included in this study? What types of imaging modalities are most pertinent/relevant and available? Do we require multi-modal/multi-parametric data? What is the best entry-level for these data, i.e., what preprocessing should be applied? How does one coordinate IRB approval across institutions? Another key aspect discussed during this session was the inclusion of meta-data in such a database. Such inclusion is decisive for clinicians and the reproducibility of (modeling and simulation) results. From purely a technological point of view, one needs to decide how to store/curate this metadata in the most efficient way. Moreover, one needs to define a precise protocol to avoid confusion and have documentation. In addition, standards need to be established for data pre-processing. For example, one could improve data sets by offering data correction algorithms to generate a harmonized reference data set and correct for most common imaging artifacts. This would aid reproducibility. Additionally, one could provide data with respect to different processing levels using already available tools deployed by the medical imaging and image computing community.

Another key question that was discussed during this session is how to establish a benchmark and demonstrators for mathematical modeling and data processing. Some of the main questions that arose during this discussion include: What are the representations that models have to return to be useful for clinical evaluation? Can we provide a benchmark that is useful for model development and/or model validation? If so, what are the best metrics for such an effort? How can we quantify tumor or patient status and what are the key metrics most clinicians trust in this context? One possibility is to establish a benchmark similar to the BraTS dataset at the organ level for tumor models. A first step towards establishing such a benchmark could be to develop internal demonstrators to showcase what we can accomplish with available modeling tools to the community at large as well as to clinicians.

Working Group 2

We summarize the main topics discussed in the *second working group* below:

One of the main challenges for many research groups working in medical imaging sciences is access to clinical data (of high quality). Data is rarely shared amongst groups. One major outcome we hope to accomplish with this seminar is to establish and curate a list of available datasets.

We also discussed aspects surrounding model selection. We discussed the option to drive an initiative for model selection and provide guidance to people with the following aspects in mind: (*i*) How do we identify required model complexity with a specific application in mind (i.e., what do the models need to capture in the context of a particular application), (*ii*) When is a model useful and for what purpose? Can we provide guidance on the usefulness of particular models for specific applications/clinical questions? (*iii*) What aspects can and should be captured by mathematical models to make them clinically useful? For example, can we include models of radio-necrosis? Are we able to design mathematical modes that can predict pseudo-progression? In this context, we concluded that an attainable concrete goal for participants in this seminar is to curate a list that identifies classes of models and their potential applications. We intend to curate this list in an online platform.

Another key aspect we discussed was model validation and the design of benchmarks for computational models. One challenge in developing benchmarks for mathematical models of disease progression is the definition of a clinical goal and/or a biological phenomenon one wants to capture and how to measure a model's performance in capturing it. Moreover, we discussed that it will also be instrumental for developing predictive capabilities to rigorously equip our simulations and model-based predictions with certificates about our belief in their accuracy (uncertainty quantification).

Panel Discussion

After the two breakout sessions described above, we came together for a *panel discussion*. We focussed on the following main items in an attempt to curate some of the information that may help us to push forward community efforts towards developing computational methods to aid clinical decision-making:

As a first attainable goal, we agreed that we would curate a list that identifies individual researchers one reaches out to for computational tools and medical imaging data. Moreover, we discussed how we could support such an endeavor of establishing a clinical benchmark financially, i.e., we identified potential funding agencies to support such an effort. We also identified several long-term goals of key clinical relevance such as differentiation of progression and pseudoprogression (i.e., radio necrosis). Moreover, we established that such a database should provide information about publicly available data sets as well as different classes of models and computational tools for data pre- and post-processing developed by individual groups. We agreed to use GitHub as a starting point to curate a platform to share our research results, methods, algorithms, and data as well as provide a platform for young researchers to showcase their academic profile.

Participants

- Michal Balcerak
Universität Zürich, CH
- George Biros
Univ. of Texas at Austin, US
- Sarah Brüningk
ETH Zürich – Basel, CH
- Malte Brunn
Universität Stuttgart, DE
- Andreas Deutsch
TU Dresden, DE
- Matthias J. Ehrhardt
University of Bath, GB
- Elies Fuster Garcia
Technical University of Valencia, ES
- David Hormuth
University of Texas at Austin, US
- Ender Konukoglu
ETH Zürich, CH
- Jonas Latz
Heriot-Watt University – Edinburgh, GB
- Guillermo Lorenzo
University of Pavia, IT & UT Austin, US
- Arvid Lundervold
University of Bergen, NO
- Andreas Mang
University of Houston, US
- Björn H. Menze
Universität Zürich, CH
- Dorit Merhof
Universität Regensburg, DE
- Daniel Paech
Universitätsklinikum Bonn, DE
- Sarthak Pati
University of Pennsylvania, US
- Miriam Schulte
Universität Stuttgart, DE
- Koen Van Leemput
Martinos Center – Charlestown, US
- Jonas Weidner
TU München – Klinikum rechts der Isar, DE
- Benedikt Wiestler
TU München – Klinikum rechts der Isar, DE
- Barbara Wohlmuth
TU München – Garching, DE

Remote Participants

- Atle Bjornerud
University of Oslo, NO
- Chao Li
University of Cambridge, GB
- Jana Lipkova
Harvard University – Boston, US
- Assad Oberai
USC – Los Angeles, US
- Russell Rockne
City of Hope – Duarte, US

Report from Dagstuhl Seminar 23031

Frontiers of Information Access Experimentation for Research and Education

Christine Bauer*[1], Ben Carterette*[2], Nicola Ferro*[3], Norbert Fuhr*[4], and Guglielmo Faggioli†[5]

1 **Utrecht University, NL.** c.bauer@uu.nl
2 **University of Delaware and Spotify, US.** carteret@acm.org
3 **University of Padua, IT.** nicola.ferro@unipd.it
4 **University of Duisburg-Essen, DE.** norbert.fuhr@uni-due.de
5 **University of Padua, IT.** faggioli@dei.unipd.it

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 23031 "Frontiers of Information Access Experimentation for Research and Education", which brought together 37 participants from 12 countries.

The seminar addressed technology-enhanced information access (information retrieval, recommender systems, natural language processing) and specifically focused on developing more responsible experimental practices leading to more valid results, both for research as well as for scientific education.

The seminar brought together experts from various sub-fields of information access, namely Information Retrieval (IR), Recommender Systems (RS), Natural Language Processing (NLP), information science, and human-computer interaction to create a joint understanding of the problems and challenges presented by next generation information access systems, from both the research and the experimentation point of views, to discuss existing solutions and impediments, and to propose next steps to be pursued in the area in order to improve not also our research methods and findings but also the education of the new generation of researchers and developers.

The seminar featured a series of long and short talks delivered by participants, who helped in setting a common ground and in letting emerge topics of interest to be explored as the main output of the seminar. This led to the definition of five groups which investigated challenges, opportunities, and next steps in the following areas: *reality check, i.e. conducting real-world studies*, *human–machine-collaborative relevance judgment frameworks*, *overcoming methodological challenges in information retrieval and recommender systems through awareness and education*, *results-blind reviewing*, and *guidance for authors.*

Seminar January 15–20, 2023 – https://www.dagstuhl.de/23031

2012 ACM Subject Classification Information systems → Information retrieval; Information systems → Recommender systems; Computing methodologies → Natural language processing; Information systems → Users and interactive retrieval; Information systems → Evaluation of retrieval results

Keywords and phrases evaluation, experimentation, information access systems, simulation, user interaction

Digital Object Identifier 10.4230/DagRep.13.1.68

* Editor / Organizer

† Editorial Assistant / Collector

Frontiers of Information Access Experimentation for Research and Education, *Dagstuhl Reports*, Vol. 13, Issue 1, pp. 68–154
Editors: Christine Bauer, Ben Carterette, Nicola Ferro, and Norbert Fuhr
DAGSTUHL REPORTS
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Christine Bauer (Utrecht University, NL, c.bauer@uu.nl)
Ben Carterette (University of Delaware and Spotify, US, carteret@acm.org)
Nicola Ferro (University of Padua, IT, nicola.ferro@unipd.it)
Norbert Fuhr (University of Duisburg-Essen, DE, norbert.fuhr@uni-due.de)

Information access – which includes Information Retrieval (IR), Recommender Systems (RS), and Natural Language Processing (NLP) – has a long tradition of relying heavily on experimental evaluation, dating back to the mid-1950s, a tradition that has driven the research and evolution of the field. However, nowadays, research and development of information access systems are confronted with new challenges: information access systems are called to support a much wider set of user tasks (informational, educational, and entertainment, just to name a few) which are increasingly challenging, and as a result, research settings and available opportunities have evolved substantially (e.g., better platforms, richer data, but also developments within the scientific culture) and shape the way in which we do research and experimentation. Consequently, it is critical that the next generation of scientists is equipped with a portfolio of evaluation methods that reflect the field's challenges and opportunities, and help ensure internal validity (e.g., measures, statistical analyses, effect sizes, etc., to support establishing a trustworthy cause-effect relationship between treatments and outcomes), construct validity (e.g., measuring the right thing rather than a partial proxy), and external validity (e.g., critically assessing to which extent findings hold in other situations, domains, and user groups). A robust portfolio of such methods will contribute to developing more *responsible experimental practices*.

Therefore, we face two problems: Can we re-innovate how we do research and experimentation in the field by addressing emerging challenges in experimental processes to develop the next generation of information access systems? How can a new paradigm of experimentation be leveraged to improve education to give an adequate basis to the new generation of researchers and developers?

This Dagstuhl Seminar brought together experts from various sub-fields of information access, namely IR, RS, NLP, information science, and human-computer interaction to create a joint understanding of the problems and challenges presented above, to discuss existing solutions and impediments, and to propose next steps to be pursued in the area.

To stimulate thinking around these themes, prior to the seminar, we challenged participants with the following questions:

- Which experimentation methodologies are most promising to further develop and create a culture around?
- In which ways can we consider the concerns related to Fairness, Accountability, and Transparency (FAccT) in the experimentation practices? How can we establish FaccT-E, i.e. FaccT in Experimentation?
- How can industry and academia better work together on experimentation?
- How can critical experimentation methods and skills be taught and developed in academic teaching?
- How can we foster collaboration and run shared infrastructures enabling collaborative and joint experimentation? How to organize shared evaluation activities taking advantage of new hybrid forms of participation?

We started the seminar week with a series of long and short talks delivered by participants, also in response to the above questions. This helped in setting a common ground and understanding and in letting emerge the topics and themes that participants wished to explore as the main output of the seminar.

This led to the definition of five groups which explored challenges, opportunities, and next steps in the following areas

- **Reality check**: The working group identified the main challenges in doing real-world studies in RS and IR research – and points to best practices and remaining challenges in both how to do domain-specific or longitudinal studies, how to recruit the right participants, using existing or creating new infrastructure including appropriate data representation, as well as how, why and what to measure.
- **Human-machine-collaborative relevance judgment frameworks**: The working group studied the motivation for using Large Language Models (LLMs) to automatically generate relevance assessments in information retrieval evaluation, and raises research questions about how LLMs can help human assessors with the assessment task, whether machines can replace humans in assessing and annotating, and what are the conditions under which human assessors cannot be replaced by machines.
- **Overcoming methodological challenges in IR and RS through awareness and education**: Given the potential limitations of today's predominant experimentation practices, we find that we need to better equip the various actors in the scientific ecosystem in terms of scientific methods, and we identify a corresponding set of helpful resources and initiatives, which will allow them to adopt a more holistic perspective when evaluating such systems.
- **Results-blind reviewing**: The current review processes lead to undue emphasis on performance, rejecting papers focusing on insights in case they show no performance improvements. We propose to introduce a results-blind reviewing process forcing reviewers to put more emphasis on the theoretical background, the hypotheses, the methodological plan and the analysis plan of an experiment, thus improving the overall quality of the papers being accepted.
- **Guidance for authors**: The Information Retrieval community has over time developed expectations regarding papers, but these expectations are largely implicit. In contrast to adjacent disciplines, efforts in the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR) community have been rather sparse and are mostly due to individuals expressing their own views. Drawing on materials from other disciplines, we have built a draft set of guidelines with the aim of them being understandable, broad, and highly concise. We believe that our proposal is general and uncontroversial, can be used by the main venues, and can be maintained with an open and continuous effort driven by, and for, the community.

2 Table of Contents

Executive Summary

Overview of Talks

Working Groups

3 Overview of Talks

3.1 Kickoff on Frontiers of Information Access Experimentation for Research and Education

Ian Soboroff (National Institute of Standards and Technology, US, ian.soboroff@nist.gov)

The goal of this talk is to set out a common starting point for the seminar, and I approach this from the perspective of test collections and information retrieval. I start from the structure of a test collection and describe the pooling and relevance assessment process, highlighting known issues in those processes, including incompleteness, assessor disagreement, shallow pooling, and integrating results from multiple test collections. I close the talk with a list of hard problems in evaluation such as handling low run coverage and the absence of external ground truth.

3.2 Goodhart's Law and the Lucas Critique

Justin Zobel (University of Melbourne, AU, jzobel@unimelb.edu.au)

The discipline of IR has a deep literature examining how best to measure performance, in particular the practice of assessing retrieval systems using batch experiments based on collections and relevance judgements. However, this literature has only rarely considered an underlying principle: that measured scores are inherently incomplete as a representation of human behaviour. In other disciplines, the significance of the principle has been examined through the perspectives of Goodhart's law and the Lucas critique. Here I argue that these apply to IR and show that neglect of this principle has consequences in practice, separate from issues that can arise from poor experimental designs or the use of effectiveness measures in ways that are known to be questionable. Specifically, blind pursuit of performance gains based on the optimisation of scores, and analysis based solely on aggregated measurements, can lead to misleading or meaningless outcomes.

This talk was based on SIGIR Forum paper "When Measures Mislead: The Limits of Batch Assessment of Retrieval Systems" [1], available at `https://www.sigir.org/wp-content/uploads/2022/07/p12.pdf`.

References

1 J. Zobel. When measurement misleads: The limits of batch assessment of retrieval systems. *SIGIR Forum*, 56(1), June 2022.

3.3 User-centric Evaluation

Bart P. Knijnenburg (Clemson University, US, bartk@clemson.edu)

I presented an evaluation framework to study the user experience of interactive systems. It involves measuring users' perception and experiences with questionnaires and then triangulating these with behaviour. The subjective constructs explain why users' behaviour is different for different systems – this explanation is the main value of our framework.

I also addressed the filter bubble, and proposed to evaluate and build information systems in a way that supports rather than replaces decision-making; covers users' tastes, plural; and focuses on exploration and preference development rather than consumption.

Finally, I addressed the challenge of designing human subjects studies that preserve research participants' privacy and security while still generating robust results.

3.4 Offline Evaluation Based on Preferences

Charles L. A. Clarke (University of Waterloo, CA, claclark@uwaterloo.ca)

Traditional offline evaluation of search, recommender, and other systems involves gathering item relevance labels from human editors. These labels can then be used to assess system performance using offline evaluation metrics. Unfortunately, this approach does not work when evaluating highly-effective ranking systems, such as those emerging from the advances in machine learning. Recent work demonstrates that moving away from pointwise item and metric evaluation can be a more effective approach to the offline evaluation of systems.

3.5 The Impact of Human Assessors on Judgements, Labels, Supervised Models, and Evaluation Results

Gianluca Demartini (The University of Queensland, AU, demartini@acm.org)

When we evaluate systems or train supervised models we make use of human annotations (e.g., judgements or labels). In this talk, I have presented examples of how different people may provide different annotations for the same data items. First, I have shown how misinformation judgements are prone to political background bias [1, 2]. Then, I have shown how human annotators discriminate based on the socio-economic status of the persons depicted in the annotated content [3]. The way human annotators are biased also depends on how the annotation task is framed and on what extra information we provide them with [4]. Finally, I have shown what it means to train supervised models with such biased labels and how these models behave very differently when they are trained with labels provided by different human annotators [5]. It is thus important for us to start considering tracking information about who the human assessors and annotators are and to include this as meta-data of our test collections [6].

References

1 Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. Can the crowd identify misinformation objectively?: The effects of judgment scale and assessor's background. In Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 439–448. ACM, 2020.

2 David La Barbera, Kevin Roitero, Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. Crowdsourcing truthfulness: The impact of judgment scale and assessor bias. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval – 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II*, volume 12036 of *Lecture Notes in Computer Science*, pages 207–214. Springer, 2020.

3 Shaoyang Fan, Pinar Barlas, Evgenia Christoforou, Jahna Otterbacher, Shazia W. Sadiq, and Gianluca Demartini. Socio-economic diversity in human annotations. In *WebSci '22: 14th ACM Web Science Conference 2022, Barcelona, Spain, June 26 – 29, 2022*, pages 98–109. ACM, 2022.

4 Jiechen Xu, Lei Han, Shazia Sadiq, and Gianluca Demartini. On the role of human and machine metadata in relevance judgment tasks. *Information Processing & Management*, 60(2):103177, 2023.

5 Periklis Perikleous, Andreas Kafkalias, Zenonas Theodosiou, Pinar Barlas, Evgenia Christoforou, Jahna Otterbacher, Gianluca Demartini, and Andreas Lanitis. How does the crowd impact the model? A tool for raising awareness of social bias in crowdsourced training data. In Mohammad Al Hasan and Li Xiong, editors, *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pages 4951–4954. ACM, 2022.

6 Gianluca Demartini, Kevin Roitero, and Stefano Mizzaro. Managing bias in human-annotated data: Moving beyond bias removal. *CoRR*, abs/2110.13504, 2021.

3.6 A Plea for Result-Less Reviewing

Norbert Fuhr (University of Duisburg-Essen, DE, norbert.fuhr@uni-due.de)

Scientific experiments aim at testing hypotheses and gaining insights into cause-and-effect for the setting studied. Unfortunately, most IR publications focus on the first aspect, while papers addressing the second aspect get rejected if they fail to show improvements in terms of performance. However, many published papers suffer from severe flaws in their experimental analysis part, which makes their results almost useless. Focusing on performance numbers, top IR conferences and journals accept only papers showing improvements, which also leads to publication bias. As PhD students must publish to get a degree, they might be tempted to cheat if their proposed method does not yield the desired results.

As a way out, we propose to switch to result-less reviewing, which is standard e.g. in some psychological journals. Here reviewers cannot see the actual experimental results and have to base their decision on the theoretical background, the hypotheses, the methodological plan and the analysis plan. In case of acceptance, the experimental results are included in the paper published.

This approach could help to achieve higher scientific quality and better reproducibility of experimental studies in IR.

3.7 Understanding your User, Process Tracing as a User-centric Method

Martijn C. Willemsen (Eindhoven University of Technology & JADS – 's-Hertogenbosch, NL, m.c.willemsen@tue.nl)

In evaluating our information access systems, we get more insights if we combine subjective measures (e.g. satisfaction) with interaction data [1]. However, most interaction data used nowadays, like simple clickstreams, do not provide sufficient insights into the underlying cognitive processes of the user. In this talk, I show how richer process measures (like hovers and eye-tracking) can provide deeper insights into the underlying decision processes of a user. For example, they help to understand when and why users search more superficially or more deeply into a list of results from the algorithm.

3.7.1 Process tracing in decision making

In decision-making, process tracing methods are commonly used to better understand human decision processes [2]. In the talk, I demonstrated one technique that I developed myself, called mouselabWEB[1]. This information board tool allows users to acquire information by hovering over boxes. It can be regarded as a cheap and simple eye-tracker-like tool that can be used in online studies. The tool allows users to easily design a mouselabWEB table and page and takes care of data storage and handling [3].

3.7.2 Process tracing used in Recommender Systems

We already used process tracing-like measures in earlier RS work to better understand the decision processes. In our work on latent feature diversification [4], we presented diversified lists of movie recommendations by their titles. Only when hovering the titles, additional movie information and poster were shown. This measured how much effort people spend and how many recommendations were inspected. We found that a top-20 list of recommendations resulted in more effort than a top-5 list, which subsequently increased choice difficulty and reduced satisfaction. In work on user inaction [5], we investigated why users do not interact with some recommended items, questioning if we should keep showing these recommendations. We found diverse reasons for inaction and showed that some reasons provide good reasons for not recommending the item again, whereas others indicate that it would actually be very beneficial to show the item again in the next round of recommendations.

[1] `https://github.com/MCWillemsen/mouselabWEB20`

References

1 Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User Model. User Adapt. Interact.*, 22(4-5):441–504, 2012.

2 M. Schulte-Mecklenbeck, J.G. Johnson, U. Böckenholt, D.G. Goldstein, J.E. Russo, N.J. Sullivan, and M.C. Willemsen. Process-tracing methods in decision making: on growing up in the 70s. *Current Directions in Psychological Science*, 26(5):442–450, October 2017.

3 Martijn C. Willemsen and Eric J. Johnson. *(Re)Visiting the Decision Factory: Observing Cognition with MouselabWEB*, pages 76–95. Taylor and Francis Ltd., United Kingdom, 2nd edition, 2019. Publisher Copyright: © 2019 selection and editorial matter, Michael Schulte- Mecklenbeck, Anton Kühberger, and Joseph G. Johnson; individual chapters, the contributors.

4 Martijn C. Willemsen, Mark P. Graus, and Bart P. Knijnenburg. Understanding the role of latent feature diversification on choice difficulty and satisfaction. *User Model. User Adapt. Interact.*, 26(4):347–389, 2016.

5 Qian Zhao, Martijn C. Willemsen, Gediminas Adomavicius, F. Maxwell Harper, and Joseph A. Konstan. Interpreting user inaction in recommender systems. In Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan, editors, *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pages 40–48. ACM, 2018.

3.8 From Living Lab Studies to Continuous Evaluation

Philipp Schaer (Technische Hochschule Köln, DE, philipp.schaer@th-koeln.de)

In this short talk, I briefly introduced the basic idea behind using living labs for information retrieval or recommender system evaluation. I also outlined a framework to extend living labs to enable a continuous evaluation environment.

3.8.1 Living labs

Livings labs were introduced in CLEF and TREC by initiatives like NewsREEL [1], Open-Search [2] or, more recently, LiLAS [3], with a particular focus on academic search evaluation. The general motivation behind living labs is to enable in-vivo evaluation in real-world settings and to extend the Cranfield-style in-vitro evaluations. Limitations of Cranfield studies like being static and not incorporating real-world users should be avoided. Instead of using (domain-specific) experts to evaluate retrieval results, the behaviour of real-world users is logged to measure their usage of different system implementations. Approaches like A/B testing or interleaving allow comparing the amount and type of interactions with these different systems to infer the underlying system performance. By integrating real-world systems and users into the evaluation process, organizers of living lab evaluations can hope to bring more diversity and heterogeneity in the set of evaluators and, therefore, a higher level of realism. In industry, these types of online evaluations in real-world applications are common but not in academia, as access to these systems is usually not possible for external researchers and their systems. Although in principle, systems like STELLA [4] would make this possible, it is rarely used.

Most living lab CLEF and TREC initiatives suffered from a common set of issues, like, the small number of click events gathered in the experiments, therefore long-running experiments, missing user information or anonymous profiles, no differentiating in click events and no possibility to include expert feedback and generally the problem of being confronted with constant change in the systems and their data sets.

3.8.2 Continuous evaluation

A framework for continuous evaluation was outlined to overcome some of the previously outlined issues. The framework is based on a living lab installation within a real-world system but extends it with the following components:

- Different user profiles – (regular) platform users whose user interaction data is logged and expert users that can directly annotate relevance labels on results in the systems.
- Relevance assessments – The expert assessments will be added to a constantly growing test collection that has to support versioning.
- Simulation module – As both expert and regular user feedback is expected to be insufficiently small at the beginning, different user types or interaction patterns can be simulated based on the interaction and relevance data gathered so far.

These components within the framework can run over a long time and create a constantly growing set useful for evaluating systems – running in the living lab as an online study or using the distilled/simulated evaluation data available for offline evaluation.

A first version of this framework will be implemented in the DFG-funded STELLA II project[2].

References

1 Frank Hopfgartner, Krisztian Balog, Andreas Lommatzsch, Liadh Kelly, Benjamin Kille, Anne Schuth, and Martha A. Larson. Continuous evaluation of large-scale information access systems: A case for living labs. In Nicola Ferro and Carol Peters, editors, *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*, pages 511–543. Springer, 2019.

2 Rolf Jagerman, Krisztian Balog, Philipp Schaer, Johann Schaible, Narges Tavakolpoursaleh, and Maarten de Rijke. Overview of TREC opensearch 2017. In Ellen M. Voorhees and Angela Ellis, editors, *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017*, volume 500-324 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2017.

3 Philipp Schaer, Timo Breuer, Leyla Jael Castro, Benjamin Wolff, Johann Schaible, and Narges Tavakolpoursaleh. Overview of lilas 2021 – living labs for academic search. In K. Selçuk Candan, Bogdan Ionescu, Lorraine Goeuriot, Birger Larsen, Henning Müller, Alexis Joly, Maria Maistro, Florina Piroi, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction – 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21-24, 2021, Proceedings*, volume 12880 of *Lecture Notes in Computer Science*, pages 394–418. Springer, 2021.

4 Timo Breuer and Philipp Schaer. A living lab architecture for reproducible shared task experimentation. In Christian Wolff and Thomas Schmidt, editors, *Information between Data and Knowledge: Information Science and its Neighbors from Data Science to Digital Humanities – Proceedings of the 16th International Symposium of Information Science, ISI 2021, Regensburg, Germany, March 8-10, 2021*, pages 348–362. Werner Hülsbusch, 2021.

[2] https://stella-project.org/

3.9 An Idea for Evaluating Retrieve & Generate Systems

Laura Dietz (University of New Hampshire, US, dietz@cs.unh.edu)

Natural language generation models (like GPT*) are here to stay, and they are a huge opportunity to build systems that combine retrieval and language generation in a combined system.

But: how can we evaluate the quality of such systems?

We discuss an idea for a new paradigm, the EXAM Answerability Metric [1], which uses a Question Answering (QA) system along with some human-written exam questions to evaluate whether the systems retrieve good *information* (instead of the right terms).

The paradigm has other advantages such as no need for highly trained assessors, no fixed corpus for retrieval (open web is possible), and comparison of retrieval-only systems and fully-generated systems on equal footing. Moreover, additional systems can be added for evaluation later without bias against non-participating systems. There is the possibility to add additional exam questions at a later point, to increase resolution between systems.

We compare the EXAM evaluation metric to the official TREC quality metrics on the TREC Complex Answer Retrieval Y3 track. We observe a Spearman Rank Correlation coefficient of 0.73. In contrast, ROUGE yields a correlation of 0.01.

There are also many open questions about the evaluation paradigm, I would like to discuss with participants in this Dagstuhl Seminar.

References

1 David P. Sander and Laura Dietz. EXAM: how to evaluate retrieve-and-generate systems for users who do not (yet) know what they want. In Omar Alonso, Stefano Marchesin, Marc Najork, and Gianmaria Silvello, editors, *Proceedings of the Second International Conference on Design of Experimental Search & Information REtrieval Systems, Padova, Italy, September 15-18, 2021*, volume 2950 of *CEUR Workshop Proceedings*, pages 136–146. CEUR-WS.org, 2021.

3.10 Metadata Annotations of Experimental Data with the `ir_metadata` Schema

Timo Breuer (Technische Hochschule Köln, DE, timo.breuer@th-koeln.de)

In this talk, we present the current status of `ir_metadata` [1] – a metadata schema for annotating run files of information retrieval experiments. We briefly outline the logical plan of the schema that is based on the PRIMAD model (first introduced as part of the Dagstuhl Seminar 16041 [2]). The acronym stems from the six components that can possibly affect the reproducibility of an experiment including the Platform, Research Goal, Implementation, Method, Actor, and Data. In addition, we extended the taxonomy with related subcomponents, for which details can be found on the project's website[3].

[3] https://www.ir-metadata.org/

Furthermore, we demonstrate how run files can be annotated in practice, describe the current software support and include example experiments in the form of reproducibility studies. Open points of discussion include what kinds of additional software features could be implemented to reduce the annotation effort or how the schema can be made a community standard in general. By introducing this resource to the community, we hope to stimulate a more reproducible, transparent, and sustainable use of experimental artefacts.

References

1 Timo Breuer, Jüri Keller, and Philipp Schaer. ir_metadata: An extensible metadata schema for IR experiments. In *SIGIR*, pages 3078–3089. ACM, 2022.

2 Juliana Freire, Norbert Fuhr, and Andreas Rauber. Reproducibility of data-oriented experiments in e-science (Dagstuhl Seminar 16041). *Dagstuhl Reports*, 6(1):108–159, 2016.

3.11 Measuring Fairness

Maria Maistro (University of Copenhagen, DK, mm@di.ku.dk)

In recent years, the discussion on the fairness of Machine Learning (ML) models has gained increasing attention and involved different research communities, including Information Retrieval (IR) and Recommender Systems (RS). In the ML community, well-defined fairness criteria have been proposed and applied to the risk assignment score returned by classifiers. Assume that there are two (or more) groups, denoted by $\mathcal{A}$ and $\mathcal{B}$, defined on attributes that should not be used to discriminate people, e.g., gender, ethnicity, or age. Kleinberg et al. [1] propose 3 fairness criteria: (1) calibration within groups; (2) balance for the positive class; and (3) balance for the negative class. Calibration within groups means that the probability score estimated by a classifier is well-calibrated, i.e., if a classifier returns a probability x for people in group $\mathcal{A}$ to belong to the positive class, then an x percentage of people in $\mathcal{A}$ should truly belong to the positive class. Balance for the positive class states that the average estimated probability for people truly belonging to the positive class should be the same in groups $\mathcal{A}$ and $\mathcal{B}$. Balance for the negative class is the counterpart defined for the negative class. Kleinberg et al. [1] proves that these criteria are incompatible, except for two non-realistic cases.

The above criteria are not directly applicable when the output of a system is a ranking. Ekstrand et al. [2] identify several reasons, some of which are mentioned in the following. First, items are organized in a ranking, where they receive different levels of attention due to the position bias [3]. Therefore decisions based on model scores, i.e., how to generate the ranking, are not independent and can not be evaluated independently. Second, users can access IR and recommendation systems multiple times over a period of time and decisions based on model predictions are repeated over time. Thus, fairness should be evaluated for the whole process, not at a single point in time. Third, multiple stakeholders are involved with IR and RS systems and they have different fairness constraints. For example, users of the system might be concerned about receiving results that are not biased towards some of their attributes, e.g., gender, while providers might be concerned about their items not being underrepresented in the ranking.

Due to the above reasons, there has been a proliferation of fairness definitions and measures, targeting different nuances of the same problem and trying to adapt more general fairness definitions to the ranking problem. Recent surveys identify more than 6060 different variants of fairness definitions resulting in more than 4040 different fairness measures [4, 5].

In this talk, I argue that there is a need for a better understanding of different fairness definitions and measures. I present some open questions and future research directions which include: an exploration of the relationship between bias, data distribution, and fairness [6]; an analysis of formal properties and pitfalls of fairness measures as done for IR measures [7]; evaluation approaches able to accommodate multiple aspects, e.g., relevance, fairness and credibility [8]; guidelines, benchmarks, and tools to advise researchers and practitioners in designing the most appropriate evaluation protocol for fairness.

References

1 Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, volume 67 of *LIPIcs*, pages 43:1–43:23. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2017.

2 Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. Fairness in information access systems. *Found. Trends Inf. Retr.*, 16(1-2):1–177, 2022.

3 Nick Craswell, Onno Zoeter, Michael J. Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In Marc Najork, Andrei Z. Broder, and Soumen Chakrabarti, editors, *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California, USA, February 11-12, 2008*, pages 87–94. ACM, 2008.

4 Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Trans. Inf. Syst.*, 2022.

5 Enrique Amigó, Yashar Deldjoo, Stefano Mizzaro, and Alejandro Bellogín. A unifying and general account of fairness measurement in recommender systems. *Inf. Process. Manag.*, 60(1):103115, 2023.

6 Yashar Deldjoo, Alejandro Bellogín, and Tommaso Di Noia. Explaining recommender systems fairness and accuracy through the lens of data characteristics. *Inf. Process. Manag.*, 58(5):102662, 2021.

7 Marco Ferrante, Nicola Ferro, and Maria Maistro. Towards a formal framework for utility-oriented measurements of retrieval effectiveness. In James Allan, W. Bruce Croft, Arjen P. de Vries, and Chengxiang Zhai, editors, *Proceedings of the 2015 International Conference on The Theory of Information Retrieval, ICTIR 2015, Northampton, Massachusetts, USA, September 27-30, 2015*, pages 21–30. ACM, 2015.

8 Maria Maistro, Lucas Chaves Lima, Jakob Grue Simonsen, and Christina Lioma. Principled multi-aspect evaluation measures of rankings. In Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong, editors, *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 – 5, 2021*, pages 1232–1242. ACM, 2021.

3.12 (Aspects of) Enterprise Search

Udo Kruschwitz (University of Regensburg, DE, udo.kruschwitz@ur.de)

Search and IR is commonly associated with Web search but there are plenty of other areas that fall outside the scope of Web search and which are nevertheless interesting and challenging. One example is enterprise search which describes search within companies or other organisations [1]. This is an area that has attracted little attention in academia (as well as in shared tasks and competitions) yet it affects millions of users who try to locate relevant information as part of their everyday work. Key challenges include the silo structure of data sources, privacy issues, the lack of link structure and the fact that there may only be a single relevant document (or none at all) for a given information need. All this has implications, and in the context of this seminar, some of the main challenges include the absence of test collections, problems with data sharing and reproducibility as well as the domain-specific nature of each use case.

References

1 Udo Kruschwitz and Charlie Hull. Searching the enterprise. *Found. Trends Inf. Retr.*, 11(1):1–142, 2017.

3.13 Identification of Stereotypes: Retrieval and Monitoring

Paolo Rosso (Technical University of Valencia, ES, prosso@dsic.upv.es)

In the short talk, I addressed the problem of the retrieval of text fragments containing implicit and subjective information such as stereotypes, framing them, and annotating them. Part of the work was done in collaboration with OBERAXE, the Spanish observatory of racism and xenophobia. Transcribed speeches of the Spanish Congress of Deputies with immigrants as the target were framed as a threat or victims using a taxonomy where the negative/neutral/positive attitudes of the speaker were taken into account. Moreover, social media memes with women as a target were retrieved and annotated. The low inter-annotator agreement shows the necessity to go beyond the aggregated ground truth and consider the pre-aggregated information of each individual annotator in order to give voice also to minorities in disagreement with the opinion of the majority. Using, for instance, the learning with disagreements paradigm should allow the development of more equitable systems in the name of fairness.

3.14 Coordinate Research, Evaluation, and Education in Information Access: Towards a More Sustainable Environment for the Community

Nicola Ferro (University of Padua, IT, nicola.ferro@unipd.it)

The information access research field is characterized by several areas, such as IR, RS, and NLP. These areas, in turn, offer various venues where the community can meet, discuss, and grow; typically, a mix of *scientific conferences*, *evaluation fora*, and *summer/winter schools*. For example, in the IR area, there are several such venues around the world. In Europe, there is European Conference on Information Retrieval (ECIR)[4] as scientific conference; Conference and Labs of the Evaluation Forum (CLEF)[5] [1] as evaluation forum; and, European Summer School on Information Retrieval (ESSIR)[6] as summer school. In America, there is SIGIR[7] as scientific conference, which is also the premier international venue for the area; Text REtrieval Conference (TREC)[8] [2] as evaluation forum; however, they lack a summer/winter school. In Asia, there is the newly born Information Retrieval in the Asia Pacific (SIGIR-AP)[9] as scientific conference; NII Testbeds and Community for Information access Research (NTCIR)[10] [3] and Forum for Information Retrieval Evaluation (FIRE)[11] as evaluation fora; and, Asian Summer School in Information Access (ASSIA)[12].

All these venues are independent events, coordinated by their own steering committees (or equivalent bodies), with their own vision and strategic goals. Obviously, being the members of the community shared across the different committees and part of most of them, there is some informal level of coordination among these venues, which are cooperating for the overall growth of the community rather than competing for acquiring "shares" of it.

However, the main question of this talk is whether we can make better use of the venues we have in the field in order to fully unveil the potential of (research, evaluation, and education) in a more coordinated way and deliver further benefits to our community in terms of quality and volume of the research produced, robustness of the experimental results achieved, effective and smooth training and education to make our junior members the new leaders.

And, if this were possible in an area, such as IR, what would it mean for the information access field at large? How would we cross the boundaries of the different areas?

3.14.1 Examples of Coordination between Research, Evaluation, and Education

In the following, we provide some possible examples of coordination between research, evaluation, and education, considering the case of ECIR, CLEF, and ESSIR.

[4] https://www.bcs.org/membership-and-registrations/member-communities/information-retrieval-specialist-group/conferences-and-events/european-conference-on-information-retrieval/
[5] https://www.clef-initiative.eu/
[6] https://www.essir.eu/
[7] https://sigir.org/general-information/history/
[8] https://trec.nist.gov/
[9] http://www.sigir-ap.org/
[10] https://research.nii.ac.jp/ntcir/index-en.html
[11] http://fire.irsi.res.in/
[12] https://goassia.github.io/

As a preliminary note, all of them happen in Europe, all of them follow an annual cycle, and their schedules match well enough[13]:

- ECIR: submission deadline in October, conference in March/April;
- CLEF: evaluation activities in January-May/June, submission deadline in June/July, conference in September;
- ESSIR: school in July-August.

ECIR ↔ CLEF: Research ↔ Evaluation

There are already some coordination activities in place between ECIR and CLEF:

- ECIR hosts a section dedicated to CLEF labs, in order to stimulate participation in the CLEF evaluation activities;
- CLEF solicits its participants to follow-up their work in the labs with a submission to ECIR.

This link is possible because the new labs for CLEF are selected around July and this matches with the submission deadline to ECIR the next October; moreover, the ECIR session happens in March/April, which is still in due time for allowing participation in a CLEF lab up to May/July. On the other side, CLEF activities end in July (labs, papers), even if the actual event is later on in September; therefore, CLEF participants have time for planning a follow-up submission to ECIR in October.

Why is this link needed? Even if both ECIR and CLEF are part of the same IR area, being it a large community, the audience of ECIR and CLEF is only partially overlapping. On the other hand, this audience may benefit from participation in both venues, not only because of more opportunities of conducting research but also because of the organized progress of such activities throughout the year, with intermediate delivery points, which help in making it smoother and break-down the overall work.

In his talk, Fuhr, see Section 3.6, argued for the need for a *result-less reviewing* approach, where papers are assessed on the basis of their methodology, innovation, research questions, soundness of the planned experiments and, if accepted, the actual experiments will be conducted later on, possibly in a follow-up publication.

This could represent another area of coordination between ECIR and CLEF: result-less papers are submitted at ECIR and, if accepted, their experimental part is then submitted to CLEF as a follow-up publication. Also in this case the schedule of the two venues aligns well enough to make this possible. And, again, this would allow the community to have more regular and intermediate steps at which to deliver their research, with the additional benefit of focusing each step on a specific aspect of the research and, possibly, improving the overall quality of the output, both the methodology and the experiments.

ECIR ↔ ESSIR: Research ↔ Education

There is currently no specific joint activity between ECIR and ESSIR.

A first example of activity could be for ESSIR to offer a mentorship program for the students attending it, in order to help them in preparing their submission to ECIR and getting feedback about it. Conferences sometimes offer mentorship programs to students but these are often asynchronous exchanges of emails or, at best, remote calls. In this case, students and senior researchers would be back-to-back in the same place for a week and this

[13] The alignment of the schedule is a partially intentional decision by the committees behind these venues.

would allow for a much more smother and productive interaction. This link between ECIR and ESSIR would be possible because ESSIR happens in July/August and the submission deadline for ECIR is in October.

During the discussion that followed-up after the presentation, it was correctly asked how this link would compare/relate to a Doctoral Consortium activity. It is true that the two activities would share some commonalities, both being a form of mentorship to students. However, in the case of a Doctoral Consortium, the purpose is to provide students with overall feedback about the PhD theme or thesis; in this case, we would focus on a much more specific goal, which is the submission of a paper to a conference. As a side note, ESSIR already hosts a form of Doctoral Consortium which is Future Directions in Information Access (FDIA).

Another form of activities could be to present at ESSIR "digested" research breakthrough highlights from the latest ECIR edition. In organizing a summer/winter school there is always a trade-off between offering foundational and advanced lectures; in both cases, the lectures are expected to cover in a reasonably complete way the topic they are about. This forces school organizers to select some topics and makes it impossible to cover all the frontier of the research in the field. These "digested highlights" could be a partial solution: they could provide a taste of other areas of the research frontier, still not being fully-fledged lectures.

ESSIR ↔ CLEF: Education ↔ Evaluation

There is currently no specific joint activity between ESSIR and CLEF.

A possible activity could be to organize a permanent educational lab at CLEF, focusing on some basic tasks such as ad-hoc retrieval. This would allow us to address another trade-off typical of summer/winter schools: lectures versus hands-on sessions. Indeed, it is often difficult to find the right balance between the two and, due to limited time available or even hardware/software setup, the hands-on sessions are often at risk to be an oversimplification. On the other hand, a permanent lab at CLEF could be seen as a very extensive hands-on session of ESSIR, giving the possibility of exploring further details, also of practical nature. Moreover, this would allow for addressing some foundational concepts (and ensuring they are well understood) before the school, giving them additional freedom when school organizers have to balance between foundational and advanced topics.

3.14.2 Towards a More Sustainable Environment for Our Community

The examples discussed in the previous section provide a very basic idea of what better coordination among our venues could be. At the same time, they should help in making clear that a change in our perspective is required.

Indeed, we currently adopt a sort of *point-wise* vision, where we target and optimize for each venue separately, and the venues themselves are somewhat organized and managed in isolation. In a sense, this incurs in a *waste of resources*, since we (both organizers and participants) may need to redo some part of the same work when passing from one venue to another and, definitely, we do not exploit any synergy and interaction among venues.

On the other hand, the approach presented in the previous section would require us to adopt a more *flow-wise* vision, consisting of progressive stamps of quality, where the different steps of our research and education activities are part of an organized process, whose ultimate goal is to make them proceed in a smoother way along the pipeline, possibly also improving the quality of the outputs. Moreover, this could be also of further help for junior researchers who often are under the "publish or perish" pressure, forcing them to spread submissions to whatever venue, often repeating or slicing their work. In this case, for example, submitting a result-less paper to ECIR and the follow-up experiments to CLEF would preserve the publication volume but in a more controlled way, aimed at ensuring a better quality of each output, methodology first, and experiment after.

Obviously, this new vision will require training of both authors and reviewers, who should understand the model and how to properly apply it. For example, if a result-less paper is accepted at ECIR, when reviewing the experimental part at CLEF, its methodology should not be questioned again, especially if the reviewers happen to be different, but the review should focus just on the experimentation and the insights gathered from it.

Overall, this new coordinated vision aims at creating a *more sustainable environment* for our community, reducing the waste of resources for intermediate steps and optimizing the overall effort for delivering an improved quality.

References

1 Nicola Ferro and Carol Peters, editors. *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*. Springer, 2019.

2 D. K. Harman and E. M. Voorhees, editors. *TREC. Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge (MA), USA, 2005.

3 T. Sakai, D. W. Oard, and N. Kando, editors. *Evaluating Information Retrieval and Access Tasks – NTCIR's Legacy of Research Impact*, volume 43 of *The Information Retrieval Series*. Springer International Publishing, Germany, 2021.

3.15 Recommender Systems Evaluation 2017–2022

Alan Said (University of Gothenburg, SE, alansaid@acm.org)

Recommender systems research and practice is a fast-developing topic with growing adoption in a wide variety of information access scenarios. In this talk, I presented a snapshot of the evaluation landscape in RS research between 2017 and 2022. The talk is based on a systematic literature review analyzing 64 papers, focusing particularly on the evaluation methods applied, the datasets utilized, and the metrics used. The study shows that the predominant experiment method is offline experimentation and that online evaluations are primarily used in combination with other experimentation methods, e.g., an offline experiment. The analysis of the snapshot of the last six years of recommender systems research shows that the research community in recommender systems has consolidated the majority of experiments on a few metrics, datasets, and methods.

4 Working Groups

4.1 Reality Check – Conducting Real World Studies

Bruce Ferwerda (Jönköping University, SE, bruce.ferwerda@ju.se)
Allan Hanbury (TU Wien, AT, allan.hanbury@tuwien.ac.at)
Bart P. Knijnenburg (Clemson University, US, bartk@clemson.edu)
Birger Larsen (Aalborg University Copenhagen, DK, birger@ikp.aau.dk)
Lien Michiels (University of Antwerp, BE, lien.michiels@uantwerpen.be)
Andrea Papenmeier (Universität Duisburg-Essen, DE, andrea.papenmeier@uni-due.de)
Alan Said (University of Gothenburg, SE, alansaid@acm.org)
Philipp Schaer (Technische Hochschule Köln, DE, philipp.schaer@th-koeln.de)
Martijn Willemsen (Eindhoven University of Technology & JADS, NL, m.c.willemsen@tue.nl)

Information retrieval and recommender systems are deployed in real world environments. Therefore, to get a real feeling for the system, we should study their characteristics in "real world studies". This raises the question: What does it mean for a study to be realistic? Does it mean the user has to be a real user of the system or can anyone participate in a study of the system? Does it mean the system needs to be perceived as realistic by the user? Does it mean the manipulations need to be perceived as realistic by the user?

4.1.1 Background & Motivation

Arguably, the most realistic users can be found on existing systems, which will typically have a sufficiently large user base. However, this raises some additional questions. Firstly, there is the question of how to sample from this user base to obtain a representative sample. Secondly, these users may have some expectations of the system, which may make them somewhat resistant to (drastic) changes. On the other hand, recruiting new users comes with its own set of challenges, discussed further in Section 4.1.2.

In a similar vein, the largest degree of "system realism" would be achieved by studying real users of an existing system. For example, log-based studies have been considered the best examples of real world studies [26] since they capture behavior in a real-life setting, with little chance of contamination or bias. However, this limits the amount of control we, as researchers, can exert, and thus the research questions we can pose and answer. On the other hand, highly controlled experiments might lack realism in terms of the system, the user experience (users knowing they are being studied) and the generalizability of the study. Realism in a study is a continuum, as illustrated in Figure 1, ranging from highly controlled experiments towards real systems with real users, and researchers need to identify the appropriate experiment type for their purpose [59].

One central question in running real world studies is the influence of measurements on the behavior and experience of users. Following the Heisenberg principle [18], it is impossible to measure without influencing. If we study existing users in an existing system, and only use behavioral measures and logs from the system we will not affect users much but it will be hard to answer our question, as the evaluation of our manipulation will be difficult. On the other hand, when we start collecting additional measures, like intermediate

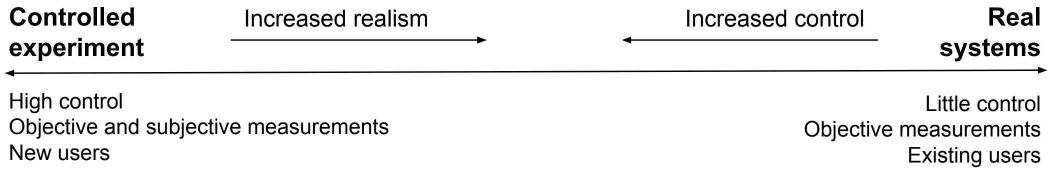


Figure 1 Control versus realism continuum.

surveys, users will know they are part of a study and modify their behavior because of that (Hawthorne effect [50]). Also, longer surveys might break the actual flow of system usage and demotivate people. Survey questions might provide the users with insights into the underlying research questions, resulting in unwanted demand characteristics or socially desirable answer patterns.

However, triangulating objective (behavioral) data with subjective measures will be crucial to understand how users experience the system [30], so a careful development and usage of a combination of subjective and objective measures is going to be central to balancing realism with adequate measurement. The challenge of 'How to measure' is further discussed in Section 4.1.3.

Then, we have the realism of the research question and experiment design. In any experiment, we manipulate the system, thus breaking some existing habits or patterns. Especially when studying users of an existing system, the realism of this manipulation is crucial. If users do not experience the manipulation as a realistic feature or implementation, the results may not be representative. Similarly, the degree of information given to the user may also influence the realism of the study. If we provide users with too much information, e.g., a very specific task and scenario to work from, users may perform actions they would not have in a realistic situation. On the other hand, if we provide too little information, e.g., when we introduce a new feature on an existing platform without any instruction, we require users to invest the time and effort to find out how the feature works before they can use it in the way we intended.

Another important consideration regarding experiment design is the assignment of users to different versions of a system. Should the experience of a single user be kept consistent throughout the entire study? Such between-subjects designs have the advantage of preventing any spill-over effects but users working side by side or communicating about the system might discover there are different versions of the system, accidentally revealing the experimental conditions and goals. Within-subject designs allow users to experience all experimental conditions, which increases statistical power (as we can control for participant variance) but ordering and spill-over effects have to be considered. Moreover, to make a real world study sufficiently realistic and also understand how behavior changes over time and how habits are formed, we will need to consider longitudinal studies which come with their own set of challenges discussed in Section 4.1.4.

Even when we carefully design our experiments and research questions and select the appropriate participants, we may arrive at conclusions that do not necessarily generalize beyond the domain. The tension between domain-specific experiments and generalizable findings is further discussed in Section 4.1.5.

Finally, the cost of running a real world study is typically many times higher than performing offline evaluation [59]. Therefore it is important to also consider the available research infrastructure, and promote the development of reusable research infrastructure, as elaborated in Section 4.1.6, and provide datasets in sufficiently general formats to promote reuse, as discussed in Section 4.1.7.

4.1.2 Recruiting Participants

Real-world user studies require recruiting efforts to find the "right" participants for the research. As a prerequisite, researchers need to have a clear understanding of the target user group and be able to **formalize the target user characteristics**. While some research can be conducted on a user sample with few limitations, others pose fine-grained requirements for user characteristics. In both cases, the user group needs to be carefully designed and adapted to the research problem at hand so that the user study is conducted on a sample representative for the user base [41].

Although some research communities have a broad consensus of what characteristics of participants should be reported, the RS and IR communities do not yet have a clear checklist of **reporting sample characteristics** and their information needs. Similarly, very few test collections, like the iSearch collection [37], actually report on the context and task users are in. Standardized reporting and metadata would also enable reproducibility [8] and data re-use (see Section 4.1.7). Inviting users that fit the recruitment criteria can be challenging. To invite users that fit the user group characteristics, information about the potential participants must be available in a structured format for filtering. Especially in IR and RS, systems often rely on user profiles [31]. Such profiles would therefore not only facilitate recruitment but also the usage of the system and avoid the "cold-start" problem [35]. With detailed user profiles, adhering to the GDPR and CCPA and formulating appropriate consent forms become additional points on a researcher's checklist.

Moreover, participants must be **recruited at the right moment**: People must be in the right mindset to start with the study. For some user groups, e.g., professionals, finding a good timing to ask for participation is crucial. Participants also must stay motivated throughout the session (or possibly even beyond) to deliver complete data. To gather high-quality data from users in real-life, ensuring that users participate for the right reasons is important too, e.g., participants should have an internal motive (that is, an actual information need) rather than generating data for financial compensation. That said, offering appropriate incentives also works towards data quality and participant motivation [14]. For that, a thorough understanding of user needs and motivations is needed. If the task/system provides users with a real benefit and actual value, the payment might not be needed and could even reduce realism and intrinsic motivation. Without such benefits, user behavior might be mostly driven by monetary incentives and divert from user behavior in the wild. However, these aspects are not necessarily in contradiction. Carefully designed, payment combined with benefits might reinforce each other. For example, in a recent longitudinal study on a music genre exploration tool, Liang and Willemsen [34] recruited new users online and paid them per session, with the system providing the additional benefit of exploring genre exploration and providing them with a personalized playlist. User drop-out was lower than common and engagement remained high across 6 weeks and 4 sessions, despite users having to respond to a medium-sized survey after every session.

Recruiting at the right time can also concern the time of day, week, or season. For example, recruiting during working hours might lead to a lack of users with full-time jobs. Defining filter criteria does not ensure that the diversity of the target user group is covered. Consequently, researchers must monitor the participant group to cover the full bandwidth of the user group under investigation. Neglecting the monitoring of incoming participants could lead to under- or over-representation of certain age, gender, or profession groups [5].

The **recruitment channel** is equally important for IR and RS studies in the wild. Several online recruiting platforms exist and can be used for studies in this field [1], e.g., MTurk or Prolific, each with their own participant characteristics [13, 43]. Other online recruiting channels include social media [41]. Offline recruiting for online experiments can pose additional challenges for participants. In some cases, IR and RS systems are already used in the wild and provide an established user base to invite for studies.

4.1.3 How to Measure

The abundance of various types of data is both a benefit and a curse of real world studies. Whereas the subsection on data representation (see Section 4.1.7 covers the proper management of this data, the current subsection addresses the measurement of data from the perspective of motivation (why do we measure?), best practices (what should we measure, and how can we make measurement easier?), and issues (what makes measurement difficult in realistic studies?). As real world studies often revolve around specific tasks and use contexts (Section 4.1.5), we also address the (lack of) generalizability of measurement.

4.1.3.1 Why to measure

Conduct theory-driven research. Real-world studies allow us to go beyond optimization of offline algorithmic performance in terms of performance metrics such as Mean Reciprocal Rank (MRR), normalized Discounted Cumulative Gain (nDCG) and recall, to a fine-grained analysis of how different system parameters can influence the system's performance at a given task.

Running a real world study requires researchers to think carefully about this "task", the right way of measuring how well the system performs at this task, and how the performance is impacted by the different system parameters. Tasks may range from highly domain-specific to more general, as discussed in Section 4.1.5. This domain-specificity means that if such studies aim to make generalizable contributions to an existing body of scientific knowledge, they should aim to explain why certain system parameters lead to higher performance.

Conducting theory-driven research requires additional measurement of intermediate (or mediating) variables that provide an explanation for the variance in performance indicators caused by system manipulations. Such mediating variables are often inherently user-centred; they can be characterized as subjective system aspects (users' perceptions of the manipulations) and user experience variables (users' self-relevant evaluation of the user experience) [30]. These can be measured with questionnaires, but there may exist behavioral proxies.

Define an evaluation target. In realistic studies, the evaluation target must shift from system performance to a multi-faceted consideration of stakeholder satisfaction [59].

As the main goal – and hence the standard metrics – of traditional IR and RS research is to optimize system performance, it avoids the question of who these metrics are optimized for. In realistic studies, metrics must be optimized to satisfy the stakeholders of the system, and the goals of these stakeholders – and hence the metrics to measure these goals – may not always align. Most prominently, measuring the satisfaction of the end-users of a system has traditionally involved user experience metrics like satisfaction, decision confidence, and self-actualization [30], while system owners tend to be interested in metrics related to conversions, such as click-through rate, session length and basket value [22, 21].

4.1.3.2 What to measure

Carefully determine what to measure. Realistic studies must capture a variety of measures that are closely related to the evaluation target and/or can explain how/why certain system aspects influence the evaluation target.

Realistic studies tend to support a variety of user behaviors, and researchers are encouraged to instrument their research systems to capture these behaviors, such as page visits, ratings, and purchases. At the same time, though, considerations of end-user privacy may prescribe that measurement be limited to the metrics that are essential to answer the research questions. It is important to acknowledge here that a user's behavior is not always an accurate representation of their own longer-term goals (let alone the goals of the system owner). As the "true" evaluation target may be difficult to measure (i.e., "user satisfaction" is an inherently latent variable, and "company profit" is an aggregate measure that depends on many other variables), researchers must decide which of the measurable behaviors are most closely related to the evaluation target (see also Section 4.1.3.3).

An important consideration here is that certain implicit behaviors may also provide valuable insights – especially when taking the importance of explanation into consideration. Users who are ignoring a recommendation, quickly navigating away from a page, or abandoning a shopping cart are providing important insights into their experience.

Users' subjective evaluations may also be important to measure: such measures may be a more accurate representation of their goals than behaviors, and even in cases where the value of behavioral metrics is clear, subjective evaluations can be used to explain the occurrence of certain behaviors. Subjective evaluations are inherently latent and must be measured using "indicator variables" [11]. The best practice to measure such evaluations is to use multi-item measurement scales, but administering such scales may be considered an intrusive practice (more suggestions on how to best do this are provided below).

Process data can also be used to explain how an evaluation target is or is not met. Process data consists of particularly granular navigational data – usually at the level of mouse-overs, intermediate clicks, or mouse movements – that can be used as evidence of a user's decision processes (e.g., which search result to visit, which product to buy, which movie to watch) [58, 49].

Make things more measurable. Realistic studies must trade off depth of measurement with user burden: more insightful measures are often more obtrusive, thereby reducing realism and participation. Below we provide suggestions on how to reduce the obtrusiveness of measurement.

While process measures are very useful to explain users' decision processes, precise process measures tend to require a certain system structure. For example, users' attention is easier to measure if certain information is hidden behind a click or a mouse-over if the user must perform a measurable action to acquire said information. More generally, behavioral data tend to be noisy due to the influence of external factors and system factors. The latter can be attenuated by reducing the number of available features and/or the amount of system personalization. Conversely, one can boost the "signal" to be measured by making the manipulated system aspect (e.g., a list of recommendations from a variety of different algorithms) more prominent in the system. Importantly, though, all of these practices may reduce the realism of the study.

Moreover, while subjective measures and process measures are invaluable in realistic studies – especially when it comes to explanation – subjective measurement is also more intrusive. Interrupting the user to fill out a questionnaire makes the interaction less realistic, and may cause asymmetric drop-outs from the study. An important consideration in this regard is when to measure users' subjective experience. The ideal but most intrusive timing is during the interaction; if the measurement occurs after the experience, it will be a retrospective and aggregate account of their experience. Aggregate retrospective evaluations of experiences have been shown to be unduly influenced by strong negative events (peaks), and events that occurred at the end of the experience [24]. Finally, if the measurement occurs too long after the experience, it may no longer accurately reflect the experience, as the user may simply no longer remember the experience. Similarly, in certain contexts users' subjective evaluations and even their interaction patterns may be inaccurate representations of their true interests – people's responses may fall prey to desirability bias, framing and default effects, or other heuristic influences that must be accounted for in measurement.

As a final consideration, one could suggest that rather than minimizing (the obtrusiveness of) measurement, one could attempt to promote measurement, e.g., by providing easily accessible and/or gamified feedback elements. Evidently, this may reduce the realism of the study.

Provide qualitative insights. Realistic studies benefit from qualitative evaluations that can be triangulated with quantitative metrics.

The metrics discussed above are well-suited for statistical evaluation – either in a correlational study, an intervention study, or a controlled experiment. When studies are sufficiently large, statistical significance may not be a suitable guideline to decide on the relevance of a finding, as even very small effects become significant when the sample size is large. In this case, researchers should focus on whether the size of the effect constitutes a meaningful contribution. Conversely, some real world studies may not attain the precision or sample size needed for statistical significance. Such studies may still provide valuable insights by treating them as pilot studies for more concerted (but perhaps less realistic) evaluation efforts.

If large sample sizes cannot be attained, a better approach may be to conduct a qualitative study. Regardless, there is immense value in deep, qualitative insights that such studies can provide. For example, one can conduct Grounded Theory studies to establish theories of users' psychology [9], or Contextual Design studies to gain a thorough understanding of users' experiences and their system needs [20]. Such studies are particularly useful when investigating evaluation targets that are highly context-dependent and/or not yet very well understood, such as fairness [23], serendipity [6] or surprise [25]. And while statistical methods are often not suitable for qualitative data, established methods exist that allow for systematic comparisons between users and/or systems (cf. "constant comparison" [9]).

Qualitative studies vary from purely observational studies to in-depth user interviews, and from single sessions to long-running studies where the researcher is "embedded" in a team or organization. As realism is often a prime consideration in such studies, other scholars have covered this aspect in much detail [20]. Note, though, that the collection and analysis of qualitative data are particularly labour-intensive, especially when they must integrate into a larger real world research infrastructure. It is also important to carefully report on qualitative procedures (e.g., procedures for "coding" qualitative data) and findings (e.g., by considering the researchers' positionality in conducting the study [9] and by providing ample evidence in the form of user quotes).

4.1.3.3 Towards best practices in measurement

Standardize measurement practices. To expedite generalizable research with real world systems, the field must adopt a set of theoretically-grounded measurement principles.

While most system-centric evaluation metrics in RS and IR have relatively standardized definitions that enjoy mostly universal adoption, this is not true for user behavior and experience metrics. While this is partially due to the highly contextual nature of relevant metrics in such studies, it may still be beneficial to identify a set of standardized metrics – or, at the very least, measurement principles that can improve the robustness of our evaluations and expedite comparisons between studies.

On the subjective side, the field could create a repository of validated measurement scales that have been proven useful in past studies. Care must be taken, though, that such a repository does not become an exclusive source of measurement instruments – there are usually limits to the applicability of existing scales. Researchers could be encouraged to particularly study the measurement principles of existing scales, such as how well they generalize to new tasks, contexts, and user groups (this can be done through the statistical process of "measurement invariance testing" [56]). Another way to address the context-specificity of measurement is to provide guidelines for researchers to adapt existing scales to their particular context, as well as guidelines for the development of completely new scales [11].

Finally, it is best if the selection, adaptation and development of scales are rooted in a theoretical framework, such as the Knijnenburg et al. [29] framework for the user-centric evaluation of recommender systems. This framework should be extended beyond recommender systems and augmented with theoretical considerations regarding users' long-term behaviors and goals.

Triangulate measures across multiple studies. To develop a set of robust and relevant metrics, IR and RS researchers should conduct a variety of studies – offline evaluations, controlled experiments, and A/B tests and observational studies with real world systems – and triangulate the data collected across these evaluation efforts.

Replication is a fundamental principle of robust scientific progress. Researchers who conduct realistic studies have an opportunity to conduct "conceptual replications" [10], where they try to replicate the findings from one domain (or one type of study) in their specific real world context. Such conceptual replications can particularly benefit from a theoretical framework like the Knijnenburg et al. framework [29], which can provide a high-level understanding of how the user experience of systems comes about (supporting the goal of explanation), provide guidance for the generation of measurement instruments and hypotheses for in-depth empirical research, and serve as a common frame of reference to compare and integrate findings across studies in different real world contexts. Furthermore, the Knijnenburg et al. framework specifically encourages the triangulation of user behaviors with their subjective evaluations – this grounds the subjective evaluations in observable actions, and in turn, explains the observable actions with subjective evaluations.

Relatedly, an important goal of conducting multi-faceted measurements in realistic studies is to test the validity and universality of the system-centric metrics that are commonly used in IR and RS research. Do these metrics correlate with positive, long-term, real world outcomes? In what contexts do they fail, and are there better system-centric metrics to optimize in these settings? As offline studies are likely not going away anytime soon, realistic studies can provide the all-important "reality check" that such studies need to validate their approach. Conversely, real world studies could provide a platform for researchers to test whether the offline performance of their solutions generalizes to a real world context. One could even create leaderboard-style challenges for each real world system to standardize this approach.

Measure unobtrusively, where possible. To maintain realism, researchers should aim to measure things unobtrusively wherever possible.

As mentioned in our introductory subsection (Section 4.1.1), it is impossible to measure users without influencing them. So while subjective evaluations are invaluable to better understand users' experiences, it would be better for the realism of our studies if such obtrusive measures could eventually be avoided. This could be supported by a concerted effort to establish behavioral proxies for subjective measures: which user behaviors best correlate with, e.g., user satisfaction? For example, Ekstrand et al. [12] showed that objective measures of diversity, novelty and accuracy correlated strongly with subjective measures based on items from a survey. In commercial systems, item ratings may – or may not – be a good proxy for user interests [38]. In dialogue-based systems, users' phrasing or tone of voice may be an indicator of their satisfaction or frustration. The answer to this question is likely highly context-dependent, so each real world study should identify its own best behavioral proxy metrics.

Similarly, researchers could benefit from easily measurable proxy metrics for longer-term (behavioral) outcomes. As outlined in Section 4.1.4, conducting longitudinal studies is a complicated affair, so the establishment of good proxy metrics could help set realistic long-term evaluation goals in studies that run over a shorter time span. Again, the best proxies for longer-term outcomes are likely context-dependent, so each real world study should aim to identify its own best proxies before reverting to shorter studies.

Conduct appropriate statistical evaluations. As real world data is messy and complex, researchers must take care to conduct the appropriate statistical evaluations of their study data.

Using the guidelines for measurement outlined above, researchers conducting realistic studies will likely collect datasets that are complex (i.e., users may have multiple sessions, or may interact in groups) and longitudinal: users are tracked over time, may interact in groups, and can drop out of and into studies at any given moment. Conducting statistical evaluations on such data is not straightforward – aggregating data to a point where simple statistics apply likely wastes much of the benefit of conducting realistic studies, so complex statistical methods are likely required to carefully analyze the data. Calculating the required sample size (both in terms of the number of users and the number of measures per user) is also not straightforward [7].

A potential benefit of longitudinal data is that such data can be used to analyze "cross-lagged panel models" [51], where metric A at timestep n is regressed on metric B at timestep n-1 and vice versa. This allows researchers to establish the causal order between metrics.

If studies are conducted on a real world system, then it is important to establish a baseline measurement of user behavior and subjective evaluation. Moreover, if this system is continuously updated, this baseline metric must be continuously updated as well.

Subsequently, researchers must aim to detect trends in the data that are caused by their interventions. Such trends may be difficult to detect, as external factors (e.g., seasonal patterns) and the effects of multiple overlapping studies influence the study data simultaneously. This means that the data must be "de-biased" to isolate the effect of the intended study. Another consideration is that study samples may not be representative (see Section 4.1.2), which may introduce bias in the statistical results. Stratified sampling and weighting may be used to avoid such biases.

A final statistical consideration in real world studies is that most study participants will have an established interaction history with the system before the study starts. Their past experiences may "spill over" into subsequent evaluations. It is thus possible that they may be biased against (or in favour of) changes made to the system as part of the experimental study. Ideally, such systems would have a steady stream of new users that can be used to avoid such effects.

4.1.4 Longitudinal Studies

Longitudinal studies conduct continuous measurements on their test subjects over a prolonged period of time. This temporal aspect provides opportunities to increase our understanding of the evolution of user experiences and behaviors over time in a way that does not only capture factors related to users' initial acceptance of a system or technology but also what influences their prolonged usage. Although longitudinal studies provide extended insights on experiences and behaviors and therefore contribute to a more realistic understanding of users, they are often considered too time-consuming and cumbersome to conduct [32]. We have defined several challenges and opportunities for longitudinal studies.

Types of longitudinal studies. The strength of longitudinal aspects lies within revealing behavioral and attitudinal changes of users over time. In the most traditional way, longitudinal studies use the same participants over the course of the study (so-called, panel studies). However, the measurement of temporal changes within panel studies comes with its own challenges. For example, researchers must keep participants motivated to continue their participation in the study. These types of longitudinal studies are particularly susceptible to attrition (e.g., missing data due to non-returning dropouts) [42]. Attrition becomes a problem when complete data is systematically different from missing data, as the impact of missing data can accumulate over time.

Time is an important factor when addressing attrition. Dropouts during a longitudinal study typically occur when the study is too long, or the sampling rate is too high (in particular for non-behavioral studies). Hence, careful consideration of temporal aspects within longitudinal studies is crucial to keep participants motivated. Besides time aspects, there are several alternative types of longitudinal studies [39] that can help to circumvent the negative effects of panel studies:

1. A cohort study: participants are drawn from a sample consisting of people sharing the same characteristics and events of interest
2. A retrospective study: analyzing historical data (e.g., offline data)

A cohort study allows for flexibility in the participants that one wants to use at a certain point in time as long as the participants show overlap in the characteristics of interest. This would allow for a lightened load on participants that would otherwise continuously be participating in the study. Alternatively, a retrospective study would make inferences based on historical data instead of collecting new data. Existing datasets such as datasets of LastFM[14] and MovieLens[15] could be used to analyze longitudinal behaviors in retrospect.

Confounding factors. Considering the reliability and the robustness of the collected data, not only the study design but also user and platform aspects play a role. Particularly in paid studies, participants could start multiple sessions to participate by creating multiple accounts or could influence one another when they are acquainted with each other and discuss the study. These activities by participants are difficult to detect and create potential confounds in the collected data. There are also several challenges with platform aspects. For example, adapting and changing the experimental platform based on interactions that were done during the longitudinal study. Adaptation of platform aspects based on participant interactions may contribute to the realism of the study (compared to a static platform) but can also collude how the data should be interpreted.

[14] E.g., `http://www.cp.jku.at/datasets/LFM-2b/`

[15] `https://grouplens.org/datasets/movielens/`

Analysis. A challenge with longitudinal studies is how to analyze the data meaningfully. Although behavioral data collection might be continuous (unobtrusive), attitudinal data is collected less frequently as this often involves questionnaires (obtrusive). Hence, the challenge in the analysis is how to distinguish correlation from causation within the collected data. A potential way to address the aforementioned issue is to triangulate the analysis between unobtrusively collected data and obtrusively collected data.

4.1.5 Domain-specific vs. General

In both RS and IR, real world experiments are often done in specific domains, for example, IR in the patent [44] and medical [40, 53] domains and recommender systems in the fashion [33] and travel [28] domains. The domains are specified by the data used, users, tasks, etc. These domains can be defined at varying levels of granularity, e.g., scientific paper search or recommendation as a domain, vs. a more specific sub-domain such as physics paper search or recommendation. Another example would be medical search as a domain, with medical search for dentists and for radiologists as sub-domains. While classification systems for research areas like DFG Subject Areas[16] or the Common European Research Classification Scheme (CERIF)[17] exist and might be a starting point, they do not catch all definitions of domains.

There is much value in small, in-depth studies, but the results from such studies are hard to generalise. With respect to research infrastructures (see Section 4.1.6) evaluation platforms should be customizable for different applications and domains but are most likely only one-shot implementations that cannot be used in different contexts. The challenge is therefore that domains tend to be treated as silos and there are few attempts to learn general principles that apply across multiple domains. Since the results of domain-specific studies cannot be compared at a numerical level, they must be compared at a conceptual level to allow for generalization. This can be seen as a continuum from general widely-applicable knowledge at one end to domain-specific knowledge at the other end, and the aim would be to shift knowledge from domain-specific to general. The widely applicable knowledge should then also allow theory to be developed – this theory would then allow researchers to make predictions about new domains, which aids the process of building tailored solutions and platforms for specific needs. This is illustrated in Figure 2.

An approach adopted in the DoSSIER project in the area of Professional Search[18] is to classify domains by knowledge task types [55], as shown in Figure 3. This would allow similarities between different domains to be more easily identified, which would assist in the generalization of results. Evaluations of approaches could then be done over similar tasks in different domains, rather than within specific domains, referred to as (semi-)replication[19], conceptual replication, or transitivity. Given the specifications of a new domain, the generalized knowledge and theory could be used to make predictions about how various approaches would work in the domains before any implementation or experiments are done. The ability to make predictions is also important for domains and tasks for which ethics and privacy concerns prevent large-scale experiments from being carried out.

[16] https://www.dfg.de/en/dfg_profile/statutory_bodies/review_boards/subject_areas/index.jsp

[17] https://www.arrs.si/en/gradivo/sifranti/sif-cerif-cercs.asp

[18] https://dossier-project.eu/

[19] In the sense of the ACM's definition on reproducibility: "Different team, different experimental setup", see https://www.acm.org/publications/policies/artifact-review-and-badging-current

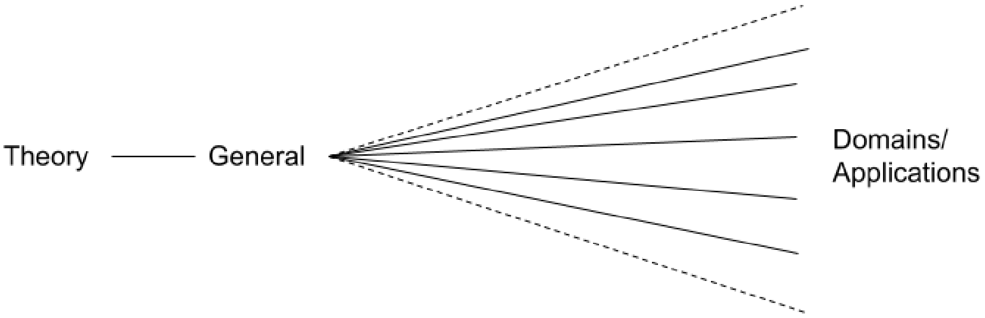


Figure 2 Theory development on a continuum from domain-specific to more general knowledge.

Such a classification would also assist in systematic reviews and meta-analyses across domains. Meta-analysis is a powerful tool to accumulate and summarize the knowledge in a research field [15]. While meta-analyses are very common in the medical area, they are more challenging in IR and RS as experiments tend to be less comparable and hence amenable to a statistical meta-analysis. A challenge here would be the different types of studies done, e.g., a controlled randomized trial is likely more easily generalizable than a large search log study. The classification should also facilitate a move toward more task-specific workshops (e.g., ALTARS 2022[20]) as a complement to domain-specific workshops (e.g., academic search in medicine or the social sciences [48] and legal IR workshops). The classification could also assist in identifying domains or task types for which too little experimental work has been done, especially to include domains that are most relevant for communities that are outside the commonly considered WEIRD (white, educated, industrialized, rich, democratic) communities [19]. It could also assist in identifying important theoretical questions and planning experiments that should be conducted to answer them (divide and conquer).

Challenges foreseen for this approach are:

- How should domains be differentiated? Medical search for dentists might be different from medical search for radiologists, or they may be considered as part of the broader domain of medical search. Where are the lines between different domains?
- What are the incentives for researchers to work on generalized insights? Solutions to domain-specific problems are likely more publishable.
- It is unlikely that we can find generalizable knowledge or theory for every aspect under evaluation. How can such limits be recognized?
- It makes sense to start this approach at a smaller scale as a proof-of-concept. How do we identify which domains and tasks to start from?
- Generalizable theory is also about people/users, not only about the systems. What does it mean for users to behave differently in some domains, and how can we generalize knowledge about user behaviors across domains?

4.1.6 Research Infrastructure

A well-functioning research infrastructure can significantly speed up and improve research in several ways, e.g., by lowering entry requirements, reducing the cost of conducting research, and making it possible to work on common goals from common standards while

[20] https://altars2022.dei.unipd.it/

task name: *the unique name assigned to the task, e.g., Prefiling Patentability Search*

definition: *a brief definition of the task*

rationale: *why is the task carried out? what should carrying out the task achieve? e.g., the task should lead to the identification of one or more patents that invalidate the query patent.*

initial information available: *what information is available at the beginning to start the search? e.g., a patent application document*

information source: *what information must be searched? e.g., all patent and non-patent information published prior to today.*

searcher: *who usually performs the search? e.g., subject expert or librarian*

query formulation methodology: *how are the queries formulated? e.g., extraction of keywords from the query document and formulation of a Boolean query using synonym expansion lists*

types of tools used: *what tools are commonly used in this type of search? e.g,. clustering results, merging results, Boolean search, ...*

search stopping criteria: *what criteria are used to decide when the search process must be stopped? e.g., a reasonable number of documents returned by a Boolean query*

output of the search: *what does the result list look like? e.g., a list of patents matching the Boolean query in reverse chronological order.*

how/if the search is documented: *is the search documented in some standard way? e.g., queries are placed into a search report along with the number of documents retrieved per query.*

post-processing, interpretation, and analysis of search results: *what is done with the result list once it is obtained? e.g., every patent is checked for relevance by an expert, if relevant it is marked as X or Y...*

any caveats to consider in the analysis or its interpretation: *e.g., the searcher needs to have a good understanding of what the requester is looking for to enable a quick review of the answers for relevance.*

Figure 3 Task definition template for professional search developed in the DoSSIER Project [55].

also increasing comparability between results [57]. Here, we consider challenges when using existing infrastructures and give overall recommendations for creating new research infrastructures that can facilitate real world studies.

4.1.6.1 Challenges of using existing infrastructure

We distinguish three types of research infrastructure used for real world studies. First, we have frameworks that can be (re)used to conduct small-scale user studies. Examples are the 3bij3 framework by [36], the Experiment Support System (ESS) and the Python Interactive Information Retrieval Evaluation (PyIRE) [16]. We will refer to these as “frameworks”. Secondly, there is a research infrastructure that is kept continuously running for longer periods of time. Examples are the MovieLens movie recommendation platform [17] and the Plista Open Recommendations Platform [54, 27], which has since been discontinued. We will refer to these as “live platforms”. Finally, the CLEF includes several labs that address challenges in both the IR and RS fields with offline datasets collected from real world systems for a specific task [3], or the ACM Conference on Recommender Systems (RecSys) challenges, which have run since 2010 [33, 2, 45]. We will refer to these as “real-world task datasets”. Below we discuss the key aspects to consider when deciding to reuse existing research infrastructure.

4.1.6.2 Recruiting participants

A clear advantage to reusing existing live platforms is that there is often no need to recruit new participants, which comes with its own set of challenges, as discussed in Section 4.1.2. The platform provides either access to real users on a real product, e.g., Plista, or may have obtained sufficient traction because of its value to the community, e.g., MovieLens. Similarly, real world task datasets are usually collected from live platforms, and therefore do not require the recruitment of participants. Frameworks, then, do not share this advantage.

4.1.6.3 Customizability/Flexibility

Frameworks allow for the most flexibility out of all the available options. Provided sufficient knowledge of the tool or some programming experience, frameworks can be customized such that a task of choice can be evaluated, as well as different experimental conditions created at will. At the other end of the spectrum, we find real world task datasets, where the task is set up front and there is no flexibility to change the data collection protocol or decide experimental conditions. In between, we find the live platforms that may have different degrees of flexibility. Flexibility is often at tension with the openness of the platform to the broader research community. On live platforms, users may have some expectations of the system. Therefore, they may be somewhat resistant to change, and therefore offer a limited degree of flexibility. This could be overcome provided a steady stream of new users who do not yet have these expectations of the system, however, on all platforms, only a few users will be converted to loyal users who will use the platform over longer periods of time.

Examples of this tension between flexibility and openness can be found in the RS community. While the NewsReel challenge allowed researchers to directly test algorithms with real users on their platforms, the task was set up front, i.e., obtain the best possible click-through rate, and the data collection protocol was fixed. Here, flexibility was limited in favor of broad community access. On the other hand, the MovieLens movie recommendation platform regularly releases new offline datasets but has thus far restricted access to its live platform to researchers within the GroupLens organization. However, research coming out of GroupLens is much more varied: it includes a larger variety of tasks, changes experimental conditions and uses a variety of data collection protocols. Here, flexibility is preferred over broad access.

4.1.6.4 Rich data

When an infrastructure draws on data from running systems with many active users realistic behavioral data can be collected. Collecting additional rich data, which can be of pivotal importance for research, can be a challenge though as system owners may be reluctant to, e.g., allow pop-up questionnaires that might annoy or drive users away. Even when these are allowed, the risk of self-selection bias is high. User behavior in a running system can also appear messy, non-targeted and display many confounding properties not related to the overall research goals. System updates can change the system properties and affect user behavior – especially in longitudinal studies [48].

4.1.6.5 Recommendations for creating new infrastructure

When existing research infrastructure is unable to support the researcher's needs, new research infrastructure has to be built.

Here we put forward some recommendations for building new research infrastructure so that it can benefit the entire research community, as building new infrastructure can be a lengthy and costly process.

The first challenge lies in obtaining sufficiently large content corpora, e.g., movies, articles or texts. An important consideration here is that after some amount of time, data will inevitably become stale. Therefore, whenever possible, we propose to integrate with APIs that give access to live content corpora that can be kept up-to-date over longer periods of time. The MovieLens platform, for example, integrates with TMDb, and as such has remained relevant for over a decade [17].

Another challenge lies in developing the system, getting the infrastructure up and running, maintaining it and providing support for both users of the system and researchers who wish to use it. Here, we recommend sufficient "realism": Funding applications should allocate sufficient funds towards software and infrastructure development, as well as the costs of running and supporting research infrastructure over prolonged periods of time. Conversely, funding institutions that wish to support reusable research infrastructure should allow for larger budget applications for the cost of development and running of research infrastructure. An interesting paradox is revealed here: The more successful the platform is with users, the more interesting it becomes for researchers, but also the higher the costs to keep it up and running.

Finally, researchers who wish to create reusable research infrastructure should dedicate significant time and effort towards documenting the system.

4.1.7 Data Representation

Information retrieval and recommender systems are critical components of modern information technology, as they allow for the efficient retrieval and recommendation of relevant information. However, for these systems to function effectively, they require underlying data to be present. This is true both in the real world, where these systems are used to process vast amounts of information, as well as in research, where the systems are being developed and tested. Without access to data sets, the research communities would not be able to perform the necessary studies and experiments to further our understanding of these systems.

Given the importance of data in information retrieval and recommender systems research, data representation is one of the cornerstones of this field. In order for datasets to be usable by the research communities, we should strive for a common understanding of what we mean by data, how we represent data, and what we communicate by (and in) data. This includes not only the format of the data but also the semantics and meaning behind the data, as well as the methods used to collect and pre-process the data [47].

Furthermore, data representation also includes the way data is organized, indexed, and stored, as well as how it can be queried and analyzed. By focusing on data representation, we can ensure that the datasets used in information retrieval and recommender systems research are of high quality and that they are accessible and usable by the entire research community. This in turn will facilitate the progress of research in our fields, and ultimately lead to the development of more effective information retrieval and recommender systems.

When sharing data, it is important to communicate the necessary details for understanding the context, use cases, and utility of the data. This includes providing detailed data descriptions, as well as data insights, which can be used by potential data users to understand the utility of the data for the intended research purposes. This information can help users to determine whether the data is appropriate for their research needs, and can also help to facilitate collaboration and sharing of data within the research community.

To ensure the reproducibility of research and to promote a deeper understanding of the data used, it is essential that researchers provide detailed information about the origin, version, and processing of the data. This includes information about the source of the data, any pre-processing or cleaning that was done, and any specific versions or updates of the data that were used in the research [4].

One way to achieve this is by adopting the practice of versioning data sets, similar to how software is versioned. This would facilitate easy identification of the specifics of the data set used in a particular study, making it simpler for others to replicate or build upon previous work. Furthermore, it would also allow researchers to clearly communicate which version of the data was used, in turn making it easier for others to access the same data set.

It is also important to remember that data processing is a crucial step in adapting certain datasets to specific use cases. Therefore, introducing the possibility of easily creating and keeping track of unique identifiers for the specific processed data sets used in research studies would facilitate reproducibility of studies. By doing so, researchers can clearly identify the specific processed data set that was used in a particular study, allowing others to easily access and use the same data set for replication or follow-up studies [46].

While keeping track of specific data versions we also need to adopt practices compatible with regulations such as General Data Protection Regulation (GDPR), making sure that users represented in data sets are sufficiently anonymized, and given the opportunity to retrospectively have their data deleted. This may create problematic scenarios if the original data is not sufficiently anonymized. However, this can in turn be used as a motivation for clear and concise privacy policies as to how to generalize, perturb, or as a last resort, censor data in order for it to be released to a wider community [52].

We should remember that data representation within systems may differ immensely between systems. However, when sharing data externally, it is important to ensure that the data representation is realistic in terms of what the data actually express and how. This includes aligning the data types used with the reality, for example, using integers for positive whole numbers and float for non-integer decimal numbers. Additionally, it is important to convey the quality of data realistically and to clearly communicate the purposes for which the shared data is created. This can help potential users to understand the limitations and potential biases of the data and can help to ensure that the data is used appropriately.

We generalize data into two specific data types commonly used in information retrieval and recommender systems, namely, **living** data, and **archival** data.

Living data refers to continuously updated data. Living data can be made available in various different formats, including continuous and uniquely identifiable downloadable snapshots, or through a so-called firehose where data is continuously delivered through an API endpoint or similar. While snapshots can provide a unique identifier making it easy to trace back to the exact version of the data, a firehose instead provides an easier way to maintain local data repositories containing up-to-date versions of the source data.

Furthermore, keeping in mind the data representation, it is important to keep the data in a format which is easily understandable, processable and accessible. This includes but is not limited to the type of format (text, image, audio, video etc.), the language of the data, the structure of the data, the size of the data, etc.

Overall, paying attention to data representation and sharing it in a clear and informative manner is crucial for the advancement of research in information retrieval and recommender systems. It can help to ensure that data is used appropriately, and can help to facilitate collaboration and sharing of data among members of the research community.

4.1.8 Next Steps

The following steps should be taken to carefully determine the **goals** of conducting real world studies:

- Classify domains by knowledge task types
- Establish context-specific evaluation targets
- Carefully consider users' information needs when conducting studies
- Develop a checklist of sample characteristics and user task details that should be collected and reported for each study

The following **resources** would expedite the design, execution and evaluation of real world studies:

- Provide researchers with access to flexible real world research infrastructure
- Obtain sufficiently large and rich content corpora that can be used in real world studies
- Create a repository of validated measurement scales
- Standardize practices for scale development
- Establish effective recruitment methods to find the "right" participants for a study
- Develop metrics that are as unobtrusive as possible to measure
- Design standardized but flexible ways to represent the data and meta-data collected in real world studies
- Study effective ways to limit attrition in longitudinal studies
- Produce best-practices guidelines for developing real world systems, getting infrastructures up and running, maintaining them and providing support for both users and researchers
- Establish guidelines to protect the privacy of research participants

The following steps must be taken to allow researchers to **integrate the findings of real world studies into generalizable knowledge**:

- Use theory to integrate domain-specific knowledge into a generalized knowledge
- Define a theoretical framework for measurement
- Develop an infrastructure for researchers to contribute analyses of and insights about real world datasets in a centralized manner
- Integrate research within specific domains as well as at the generalized knowledge level using systematic reviews, meta-analyses, task-specific workshops and domain-specific workshops
- Conduct studies to triangulate qualitative and quantitative insights, behavioral and subjective metrics, and short-term and long-term metrics

References

1 Omar Alonso and Stefano Mizzaro. Can we get rid of trec assessors? using mechanical turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, volume 15, page 16, 2009.

2 Vito Walter Anelli, Saikishore Kalloori, Bruce Ferwerda, Luca Belli, Alykhan Tejani, Frank Portman, Alexandre Lung-Yut-Fong, Ben Chamberlain, Yuanpu Xie, Jonathan Hunt, Michael M. Bronstein, and Wenzhe Shi. Recsys 2021 challenge workshop: Fairness-aware engagement prediction at scale on twitter's home timeline. In Humberto Jesús Corona Pampín, Martha A. Larson, Martijn C. Willemsen, Joseph A. Konstan, Julian J. McAuley, Jean Garcia-Gathright, Bouke Huurnink, and Even Oldridge, editors, *RecSys '21: Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September 2021 – 1 October 2021*, pages 819–824. ACM, 2021.

3 Alberto Barrón-Cedeño, Giovanni Da San Martino, Mirko Degli Esposti, Fabrizio Sebastiani, Craig Macdonald, Gabriella Pasi, Allan Hanbury, Martin Potthast, Guglielmo Faggioli, and Nicola Ferro, editors. *Experimental IR Meets Multilinguality, Multimodality, and Interaction – 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5-8, 2022, Proceedings*, volume 13390 of *Lecture Notes in Computer Science*. Springer, 2022.

4 Alejandro Bellogín and Alan Said. Improving accountability in recommender systems research through reproducibility. *User Model. User Adapt. Interact.*, 31(5):941–977, 2021.

5 Mindy E Bergman and Vanessa A Jean. Where have all the "workers" gone? a critical analysis of the unrepresentativeness of our samples relative to the labor market in the industrial–organizational psychology literature. *Industrial and Organizational Psychology*, 9(1):84–113, 2016.

6 Lennart Björneborn. Three key affordances for serendipity: Toward a framework connecting environmental and personal factors in serendipitous encounters. *J. Documentation*, 73(5):1053–1081, 2017.

7 Niall Bolger, Gertraud Stadler, and Jean-Philippe Laurenceau. *Power analysis for intensive longitudinal studies.*, pages 285–301. Handbook of research methods for studying daily life. The Guilford Press, New York, NY, US, 2012.

8 Timo Breuer, Jüri Keller, and Philipp Schaer. ir_metadata: An extensible metadata schema for IR experiments. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 – 15, 2022*, pages 3078–3089. ACM, 2022.

9 Kathy Charmaz. *Constructing grounded theory : a practical guide through qualitative analysis.* Sage Publications, London; Thousand Oaks, Calif., 2006.

10 Maarten Derksen and Jill Morawski. Kinds of replication: Examining the meanings of "conceptual replication" and "direct replication". *Perspectives on Psychological Science*, 17(5):1490–1505, 2022. PMID: 35245130.

11 Robert F. DeVellis. *Scale development: theory and applications.* Number v. 26 in Applied social research methods series. Sage, Newbury Park, Calif, 1991.

12 Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. User perception of differences in recommender algorithms. In Alfred Kobsa, Michelle X. Zhou, Martin Ester, and Yehuda Koren, editors, *Eighth ACM Conference on Recommender Systems, RecSys '14, Foster City, Silicon Valley, CA, USA – October 06 – 10, 2014*, pages 161–168. ACM, 2014.

13 D Jake Follmer, Rayne A Sperling, and Hoi K Suen. The role of mturk in education research: Advantages, issues, and future directions. *Educational Researcher*, 46(6):329–334, 2017.

14 Anja S Göritz. Incentives in web studies: Methodological issues and a review. *International Journal of Internet Science*, 1(1):58–70, 2006.

15 T. Greco, A. Zangrillo, G. Biondi-Zoccai, and G. Landoni. Meta-analysis: pitfalls and hints. *Heart, lung and vessels*, 5:219–225, 2013.

16 Mark M. Hall. To re-use is to re-write: Experiences with re-using IIR experiment software. In Toine Bogers, Samuel Dodson, Maria Gäde, Luanne Freund, Mark M. Hall, Marijn Koolen, Vivien Petras, Nils Pharo, and Mette Skov, editors, *Proceedings of the CHIIR 2019 Workshop on Barriers to Interactive IR Resources Re-use co-located with the ACM SIGIR Conference on Human Information Interaction and Retrieval, BIIRRR@CHIIR 2019, Glasgow, UK, March 14, 2019*, volume 2337 of *CEUR Workshop Proceedings*, pages 19–23. CEUR-WS.org, 2019.

17 F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4):19:1–19:19, 2016.

18 W. Heisenberg. über den anschaulichen inhalt der quantentheoretischen kinematik und mechanik. *Zeitschrift für Physik*, 43(3-4):172–198, 1927.

19 J. Henrich, S. Heine, and A. Norenzayan. Most people are not WEIRD. *Nature*, 466, 2010.

20 Karen Holtzblatt and Hugh R. Beyer. Contextual design. In Mads Soegaard and Rikke Friis Dam, editors, *Encyclopedia of Human-Computer Interaction.* The Interaction Design Foundation., Aarhus, Denmark, 2011.

21 Dietmar Jannach and Gediminas Adomavicius. Recommendations with a purpose. In Shilad Sen, Werner Geyer, Jill Freyne, and Pablo Castells, editors, *Proceedings of the 10th ACM Conference on Recommender Systems, Boston, MA, USA, September 15-19, 2016*, pages 7–10. ACM, 2016.

22 Dietmar Jannach and Michael Jugovac. Measuring the business value of recommender systems. *ACM Trans. Manag. Inf. Syst.*, 10(4):16:1–16:23, 2019.

23 Jean-Marie John-Mathews, Dominique Cardon, and Christine Balagué. From reality to world. a critical perspective on AI fairness. *Journal of Business Ethics*, 178(4):945–959, 2022.

24 Daniel Kahneman, Barbara L. Fredrickson, Charles A. Schreiber, and Donald A. Redelmeier. When more pain is preferred to less: Adding a better end. *Psychological Science*, 4(6):401–405, 1993.

25 Marius Kaminskas. Measuring surprise in recommender systems. 2014.

26 Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Found. Trends Inf. Retr.*, 3(1-2):1–224, 2009.

27 Benjamin Kille, Frank Hopfgartner, Torben Brodt, and Tobias Heintz. The plista dataset. In *NRS'13: Proceedings of the International Workshop and Challenge on News Recommender Systems*, ICPS, page 14–22. ACM, 2013.

28 Peter Knees, Yashar Deldjoo, Farshad Bakhshandegan Moghaddam, Jens Adamczak, Gerard Paul Leyson, and Philipp Monreal. Recsys challenge 2019: session-based hotel recommendations. In Toine Bogers, Alan Said, Peter Brusilovsky, and Domonkos Tikk, editors, *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019*, pages 570–571. ACM, 2019.

29 Bart P. Knijnenburg, Lydia Meesters, Paul Marrow, and Don Bouwhuis. User-centric evaluation framework for multimedia recommender systems. In Petros Daras and Oscar Mayora-Ibarra, editors, *User Centric Media – First International Conference, UCMedia 2009, Venice, Italy, December 9-11, 2009, Revised Selected Papers*, volume 40 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 366–369. Springer, 2009.

30 Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User Model. User Adapt. Interact.*, 22(4-5):441–504, 2012.

31 Alfred Kobsa. User modeling: Recent work, prospects and hazards. *Human Factors in Information Technology*, 10:111–111, 1993.

32 Sari Kujala, Talya Miron-Shatz, and Jussi J Jokinen. The cross-sequential approach: A short-term method for studying long-term user experience. *Journal of Usability Studies*, 14(2), 2019.

33 Nick Landia, Frederick Cheung, Donna North, Saikishore Kalloori, Abhishek Srivastava, and Bruce Ferwerda. Recsys challenge 2022: Fashion purchase prediction. In Jennifer Golbeck, F. Maxwell Harper, Vanessa Murdock, Michael D. Ekstrand, Bracha Shapira, Justin Basilico, Keld T. Lundgaard, and Even Oldridge, editors, *RecSys '22: Sixteenth ACM Conference on Recommender Systems, Seattle, WA, USA, September 18 – 23, 2022*, pages 694–697. ACM, 2022.

34 Yu Liang and Martijn C. Willemsen. Exploring the longitudinal effects of nudging on users' music genre exploration behavior and listening preferences. In Jennifer Golbeck, F. Maxwell Harper, Vanessa Murdock, Michael D. Ekstrand, Bracha Shapira, Justin Basilico, Keld T. Lundgaard, and Even Oldridge, editors, *RecSys '22: Sixteenth ACM Conference on Recommender Systems, Seattle, WA, USA, September 18 – 23, 2022*, pages 3–13. ACM, 2022.

35 Blerina Lika, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. Facing the cold start problem in recommender systems. *Expert Syst. Appl.*, 41(4):2065–2073, 2014.

36 Felicia Loecherbach and Damian Trilling. 3bij3–developing a framework for researching recommender systems and their effects. *Computational Communication Research*, 2(1):53–79, 2020.

37 Marianne Lykke, Birger Larsen, Haakon Lund, and Peter Ingwersen. Developing a test collection for the evaluation of integrated search. In Cathal Gurrin, Yulan He, Gabriella Kazai, Udo Kruschwitz, Suzanne Little, Thomas Roelleke, Stefan M. Rüger, and Keith van Rijsbergen, editors, *Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings*, volume 5993 of *Lecture Notes in Computer Science*, pages 627–630. Springer, 2010.

38 Sean M. McNee, István Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K. Lam, Al Mamunur Rashid, Joseph A. Konstan, and John Riedl. On the recommending of citations for research papers. In Elizabeth F. Churchill, Joseph F. McCarthy, Christine Neuwirth, and Tom Rodden, editors, *CSCW 2002, Proceeding on the ACM 2002 Conference on Computer Supported Cooperative Work, New Orleans, Louisiana, USA, November 16-20, 2002*, pages 116–125. ACM, 2002.

39 Bianca Melo, Rossana M. de Castro Andrade, and Ticianne Darin. Longitudinal user experience studies in the iot domain: a brief panorama and challenges to overcome. In Caroline Queiroz Santos, Maria Lúcia Bento Villela, Kamila Rios da Hora Rodrigues, and Ticianne de Gois Ribeiro Darin, editors, *Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems, IHC 2022, Diamantina, Brazil, October 17-21, 2022*, pages 23:1–23:13. ACM, 2022.

40 Henning Müller, Jayashree Kalpathy-Cramer, and Alba García Seco de Herrera. *Experiences from the ImageCLEF Medical Retrieval and Annotation Tasks*, volume 41 of *The Information Retrieval Series*, pages 231–250. Springer, 2019.

41 Alexander Newman, Yuen Lam Bavik, Matthew Mount, and Bo Shao. Data collection via online platforms: Challenges and recommendations for future research. *Applied Psychology*, 70(3):1380–1402, 2021.

42 Yanfang Pan and Peida Zhan. The impact of sample attrition on longitudinal learning diagnosis: A prolog. *Frontiers in psychology*, 11:1051, 2020.

43 Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163, 2017.

44 Florina Piroi and Allan Hanbury. Multilingual patent text retrieval evaluation: CLEF-IP. In Nicola Ferro and Carol Peters, editors, *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*, pages 365–387. Springer, 2019.

45 Alan Said. A short history of the recsys challenge. *AI Mag.*, 37(4):102–104, 2016.

46 Alan Said and Alejandro Bellogín. Replicable evaluation of recommender systems. In Hannes Werthner, Markus Zanker, Jennifer Golbeck, and Giovanni Semeraro, editors, *Proceedings of the 9th ACM Conference on Recommender Systems, RecSys 2015, Vienna, Austria, September 16-20, 2015*, pages 363–364. ACM, 2015.

47 Alan Said, Babak Loni, Roberto Turrin, and Andreas Lommatzsch. An extended data model format for composite recommendation. In Li Chen and Jalal Mahmud, editors, *Poster Proceedings of the 8th ACM Conference on Recommender Systems, RecSys 2014, Foster City, Silicon Valley, CA, USA, October 6-10, 2014*, volume 1247 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.

48 Philipp Schaer, Timo Breuer, Leyla Jael Castro, Benjamin Wolff, Johann Schaible, and Narges Tavakolpoursaleh. Overview of lilas 2021 – living labs for academic search (extended overview). In Guglielmo Faggioli, Nicola Ferro, Alexis Joly, Maria Maistro, and Florina Piroi, editors, *Proceedings of the Working Notes of CLEF 2021 – Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st – to – 24th, 2021*, volume 2936 of *CEUR Workshop Proceedings*, pages 1668–1699. CEUR-WS.org, 2021.

49 M. Schulte-Mecklenbeck, J.G. Johnson, U. Böckenholt, D.G. Goldstein, J.E. Russo, N.J. Sullivan, and M.C. Willemsen. Process-tracing methods in decision making: on growing up in the 70s. *Current Directions in Psychological Science*, 26(5):442–450, 2017.

50 D. Schwartz, B. Fischhoff, T. Krishnamurti, and F. Sowell. The hawthorne effect and energy awareness. *PNAS Online – Proceedings of the National Academy of Sciences*, pages 15242–5246, 2013.

51 James P. Selig and Todd D. Little. *Autoregressive and cross-lagged panel analysis for longitudinal data.*, pages 265–278. Handbook of developmental research methods. The Guilford Press, New York, NY, US, 2012.

52 Divesh Srivastava, Monica Scannapieco, and Thomas C. Redman. Ensuring high-quality private data for responsible data science: Vision and challenges. *ACM J. Data Inf. Qual.*, 11(1):1:1–1:9, 2019.

53 Hanna Suominen, Lorraine Goeuriot, Liadh Kelly, Laura Alonso Alemany, Elias Bassani, Nicola Brew-Sam, Viviana Cotik, Darío Filippo, Gabriela González Sáez, Franco Luque, Philippe Mulhem, Gabriella Pasi, Roland Roller, Sandaru Seneviratne, Rishabh Upadhyay, Jorge Vivaldi, Marco Viviani, and Chenchen Xu. Overview of the CLEF ehealth evaluation lab 2021. In K. Selçuk Candan, Bogdan Ionescu, Lorraine Goeuriot, Birger Larsen, Henning Müller, Alexis Joly, Maria Maistro, Florina Piroi, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction – 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21-24, 2021, Proceedings*, volume 12880 of *Lecture Notes in Computer Science*, pages 308–323. Springer, 2021.

54 Mozhgan Tavakolifard, Jon Atle Gulla, Kevin C. Almeroth, Frank Hopfgartner, Benjamin Kille, Till Plumbaum, Andreas Lommatzsch, Torben Brodt, Arthur Bucko, and Tobias Heintz. Workshop and challenge on news recommender systems. In Qiang Yang, Irwin King, Qing Li, Pearl Pu, and George Karypis, editors, *Seventh ACM Conference on Recommender Systems, RecSys '13, Hong Kong, China, October 12-16, 2013*, pages 481–482. ACM, 2013.

55 Elaine Toms, Sophia Althammer, Allan Hanbury, Wojciech Kusa, Ginar Santika Niwanputri, Ian Ruthven, Ayah Soufan, and Vasileios Stamatis. Knowledge task survey. Technical Report D1.1, DoSSIER EU Project, 2021.

56 Rens van de Schoot, Peter Lugtig, and Joop Hox. A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, 9(4):486–492, 2012.

57 E. Voorhees, D.K. Harman, National Institute of Standards, and Technology (US). *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridgeˆ eMA MA, 2005.

58 Martijn C. Willemsen and Eric J. Johnson. *(Re)Visiting the Decision Factory: Observing Cognition with MouselabWEB*, pages 76–95. Taylor and Francis Ltd., United Kingdom, 2nd edition, 2019. Publisher Copyright: © 2019 selection and editorial matter, Michael

Schulte- Mecklenbeck, Anton Kühberger, and Joseph G. Johnson; individual chapters, the contributors.

59 Eva Zangerle and Christine Bauer. Evaluating recommender systems: Survey and framework. *ACM Comput. Surv.*, 55(8):170:1–170:38, 2023.

4.2 HMC: A Spectrum of Human–Machine-Collaborative Relevance Judgment Frameworks

Charles L. A. Clarke (University of Waterloo, CA, claclark@plg.uwaterloo.ca)
Gianluca Demartini (University of Queensland, AU, demartini@acm.org)
Laura Dietz (University of New Hampshire, US, dietz@cs.unh.edu)
Guglielmo Faggioli (University of Padua, IT, guglielmo.faggioli@unipd.it)
Matthias Hagen (Friedrich-Schiller-Universität Jena, DE, matthias.hagen@uni-jena.de)
Claudia Hauff (Spotify, NL, claudia.hauff@gmail.com)
Noriko Kando (National Institute of Informatics (NII), JP, Noriko.Kando@nii.ac.jp)
Evangelos Kanoulas (University of Amsterdam, NL, e.kanoulas@uva.nl)
Martin Potthast (Leipzig University and ScaDS.AI, DE, martin.potthast@uni-leipzig.de)
Ian Soboroff (National Institute of Standards and Technology (NIST), US, ian.soboroff@nist.gov)
Benno Stein (Bauhaus-Universität Weimar, DE, benno.stein@uni-weimar.de)
Henning Wachsmuth (Leibniz Universität Hannover, DE, h.wachsmuth@ai.uni-hannover.de)

4.2.1 Motivation

IR evaluation traditionally needs human assessors to generate relevance judgements. Traditionally, human assessors are asked to judge the relevance of a document with respect to a topic [3]. Recently, work looking at preference judgements [2, 4] has looked at research questions related to how to best evaluate IR systems by asking human assessors which of two results is the better given an information need. The recent availability of LLMs has opened the possibility to use them to automatically generate relevance assessments in the form of preference judgements. While the idea of automatically generated judgements has been looked at before [1], new-generation LLMs drive us to re-ask the question of whether human assessors are still necessary.

New models tend to fail in a different and more diverse way compared to traditional approaches. Failure points for old models were more uniform and clear, with new systems it is harder to predict in which ways the model will fail. In most cases, LLMs (especially for what concerns generative aspects) focus on entertainment tasks. Models tend to report false facts in such a convincing way that they need to be carefully read by some expert to identify lacking factuality (e.g., Michel Foucault simulation[21]).

Our motivation to investigate the possibility of using LLMs in order to provide automatic annotations stems from some fundamental research questions that can be summarized as follows.

[21] `https://www.youtube.com/watch?v=L6c0xeAqEz4E`

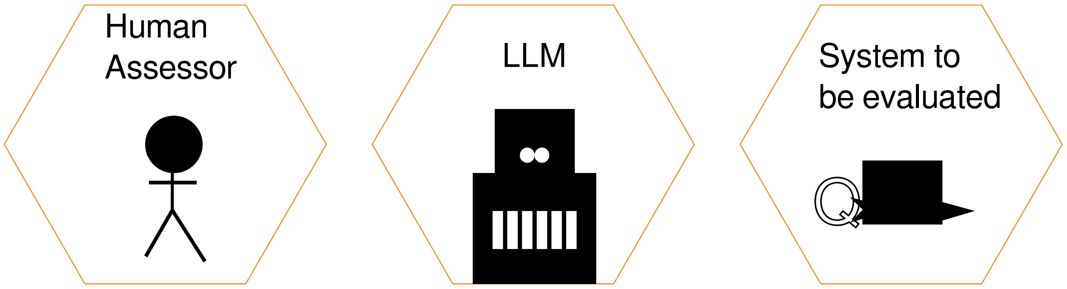


Figure 4 The three most relevant components in our system: the human assessor, the Large Language Model (LLM) that can help humans or replace them in annotating documents for relevance, and the system that we want to evaluate using the newly produced relevance judgements.

- **RQ1**: In which way automatic approaches, and in particular LLMs, can help assessors with the assessment task to yield the most reliable annotations while improving the efficiency of the annotation process? This question raises other interesting related inquiries. For example, if we were to build such a mixed human-machine annotation paradigm, which held out (not provided to the IR system) supporting information about the topic would yield the best and fastest annotations? What weighting between human and LLMs and AI-assisted annotations is ideal?
- **RQ2**: Can machines (either in the form of LLMs or in general as Artificial Intelligence (AI) models) replace humans in assessing and annotating? This question raises also concerns about what annotation target (e.g., relevance labeling, summarization, paragraph highlighting, exam questions [5]) would yield the best and fastest annotations.
- **RQ3**: What are the conditions under which human assessors cannot be replaced by machines? Alternatively, in which role can the Human assessor most productively provide relevance assessments?

Answering the questions mentioned would also require finding viable solutions for a set of additional questions and open issues that touch a number of IR evaluation process steps.

- Assessors And Collections:
  - How to use LLM to help assessors: some examples of possible usages include, summarising text, associating keywords and identifying the content of long podcasts to help assessors annotate the documents, for example by highlighting relevant fragments of text/podcast or segments with correct answers.
  - What is the effective role of the human assessor in annotating material for generative models? Should the annotator provide input at the beginning of the pipeline, by annotating the original documents, or are they more useful downstream, after the task has been carried out?
  - Generative models can be used to create new collections: corpora, conversations, queries, abstracts and so on.
- LLM and generative models to retrieve information in a broader sense:
  - IR tasks that employ LLMs have the means to provide more details: often a single answer is not satisfactory for the user. How to support the user in exploring the results further (for example via links and connected pages). Generative models can help, but is this helpful when the model simply generates the response without knowing where it comes from? In many cases, the user is not interested in receiving only the direct/short answer, but rather in seeing which documents contain it and related pieces of information to expand their knowledge.
- LLMs as an evaluation tool:

  - The model is biased: how can we use it to evaluate itself? If a model has been trained on biased data, then also the evaluation is prone to the same biases. How to detect and account for such biases?
- Evaluating LLMs and their trustworthiness:
  - Can we find a way to understand and measure to what level we can trust the results of a generative model?
  - How to carry out fact-checking, for example by identifying the source of information of a generative model and verifying that it is presented accurately.
  - Distinguish between human and machine-generated data: Important for many tasks, such as journalism, where it is of uttermost importance to verify the information. Human-generated data is more trusted.

We argue that the collaboration between humans and ML, especially under the form of LLMs, could be abstracted in the form of a spectrum. On the two extremes of this spectrum, we have either the human or the machine entirely tasked to annotate documents for relevance with respect to a query. Within the spectrum, humans and LLMs interact to a different extent. Theoretically, such a spectrum corresponds also to moving from highly expansive annotations in terms of human effort, cost and time, but with high-quality annotations, to a much less expensive annotation procedure with also a decreased annotation quality. We also argue that something exists beyond the spectrum; it corresponds to the scenario in which the machine overcomes the human, by producing relevance judgments without any form of bias. We observed this phenomenon happening already in several tasks and scenarios, and therefore we can aspect this to happen also with respect to the construction of the relevance judgments.

The remainder of this chapter is organized as follows: Subsection 4.2.2 reports details on the current state of the art and limitations associated with the current usage of LLMs ad AI in annotating documents. Subsection 4.2.3 illustrates our proposal of a spectrum of possible interactions between the human and the machine, to provide more efficient and effective annotations and relevance judgments. Subsection 4.2.4 outlines a possible experimental protocol that would allow us to verify at what point modern LLMs and whether they can be used to produce automatically relevance judgements.

4.2.2 State of the Art, Idea, and Gaps

4.2.2.1 Using LLMs to Generate Annotations and Label Automatically

Potential uses of LLMs to annotate documents, extract snippets, summarize and, in the end, annotate documents for relevance. If this can be made to work reliably, it opens up many opportunities for evaluation. For example, the LLM can be used directly to evaluate the output of other large language models (for example in summarization).

Assessments can arise from different sources, with different levels of quality and collection costs as follows.

- Human assessors or, in the enterprise scenario, final users. This scenario, at the current time, is the most expansive, but also likely to produce high-quality annotations.
- Human assessors aided by mild automatic support systems (e.g. remove redundancy, encourage consistency)
- Half of the judgments are produced by human assessors and half of the judgments are produced automatically.

- Automatic annotation of a collection, which is verified and corrected by human intervention.
- At some point even a fully automatic assessment.

An additional axis describes the type of annotations. Typically an annotation is a graded relevance judgment, but for example in EXAM [5], humans are used for generating questions instead. This can be generalized by asking human assessors for something different than traditional annotation while some Machine Learning (ML) converts the human responses into relevance assessments. This follows the paradigm of Competence Partitioning of Human-Machine-Collaboration where humans and machines are performing tasks they are best at (not vice versa).

One concern is that fully automatic assessment with LLMs can be very expensive, which is also the reason why we consider the application of LLMs as part of the retrieval process. In such a case, we could reduce the cost by considering a teacher-student training paradigm (knowledge distillation) in which a large and expensive LLM is used to train a smaller model that is less expensive to run.

Not all IR tasks focus on topics. For example, one may want to search for podcasts where two or more people interact or with a particular style. Another issue is regarding truth. For example, finding a podcast for the query "does lemon cure cancer?" that talks about healing cancer with lemon might be on topic. Nevertheless, it is unlikely to be factually correct, and therefore not relevant to correctly answering the information needs. To overcome this issue, assessors have to access external information to determine the trustworthiness of a source, or the truthfulness of a document. In a similar way, we can assume our LLM is used as an oracle that accesses external facts, verified by humans. To properly support different tasks, human intervention can be plugged into the collection and annotation of additional facts, to define relevance.

There are open questions for the special case of 100%-machine/0%human. How is this ranking evaluation different from being an approach that produces a ranking? (circularity problem). We can use multiple LLMs, possibly based on different rationales, such that it is possible to define an inter-annotation systems agreement, in which different systems are used to verify if there is an agreement between each other. An alternative approach is to endow the evaluation with additional information about relevant facts/questions/nuggets that the system (under evaluation) does not have access to.

It is yet to be understood what the risks associated with such technology are: it is likely that in the next few years, we will assist in a substantial increase in the usage of LLMs to replace human annotators. Nevertheless, a similar change in terms of data collection paradigm was observed with the increased use of crowd assessor. Up to that moment, annotations were typically made by in-house experts. Then, such annotation tasks were delegated to crowd workers, with a substantial decrease in terms of quality of the annotation, compensated by a huge increase in annotated data. It is a concern that machine-annotated assessments might degrade the quality, while dramatically increasing the number of annotations available.

The Cranfield paradigm [6] is based on simplifying assumptions that make manual evaluation feasible: *1)* independence of queries; *2)* independence of relevance of documents; *3)* Relevance is static (and not changing in time). Recently, the field is diverging from this paradigm, for example with TREC CAR and TREC CAsT/iKAT where the information needs are developing as the user learns more about the domain. The TREC Evaluation of CAST describes a tree of connected information needs, where one conversation takes a path through the tree. The Human-Machine evaluation paradigm might make it feasible to assess more connected (and hence, realistic) definitions of relevance.

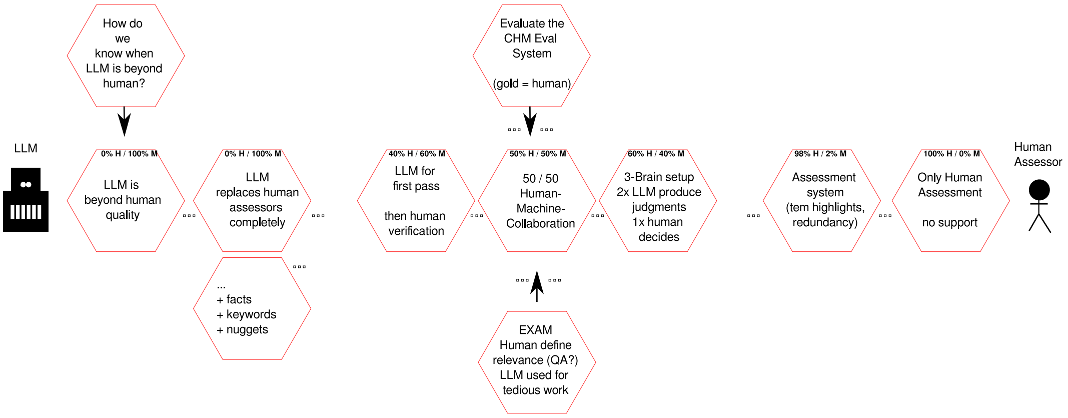


Figure 5 A spectrum of Collaborative-Human-Machine paradigms to create relevance judgments.

4.2.3 Collaborative Human-Machine Relevance Judgments

We can describe a spectrum of Collaborative-Human-Machine paradigms to create relevance judgments, where the weighting of human contributions vs machine contributions changes along the spectrum.

- **Only Human (100%H / 0%M)**: On one extreme, the human will do all assessments manually without any kind of support.
- **Human with assessment system (99%H / 1%M)**: This is a more realistic case for how TREC assessment is conducted, where humans have full control of what is relevant but are supported in the following ways: Humans can define "scan terms" that will be highlighted in the text, can limit view the pool that is already judged, ordering documents so that similar documents are near one another, produce readable presentations of retrieve content.
- **Human with document summaries (80%H/ 20%M)**: A text summarization model produces a generative summary representation of the document to be judged. The human assessor judges the representation, which is more efficient to do.
- **EXAM (60%H / 30% M)**: For each query, the human defines information nuggets that are relevant (e.g. exam questions). The machine is trained to automatically determine how many test nuggets are contained in the retrieved results (e.g. via a Q/A system).
- **Equal contribution (50%H / 50%M)**: A theoretic midpoint in the collaborative spectrum. Humans perform tasks that humans are good at. Machines perform the tasks that machines are good at. It is yet to be concretely defined what this might be.
- **3-Brain Setup (32%H / 58%M)**: Two machines each generate an assessment, and a human will select the best of the two assessments (+verification). Human decision trumps machines'.
- **LLM for first pass + human verification (30%H / 60%M)**: A first-pass assessment of the LLM is automatically produced as a suggestion. This can also be an assessment-supporting surrogate prediction like a rationale. The human assessment is based on this suggestion, but the human will have the final say.
- **LLM replaces humans completely (0%H / 100%M)**: We explore the possibility that a fully automatic assessment system might be as good as a human in producing high-quality relevance judgments.

- **LLM is beyond human (0%H / 100%M)**: Given known biases in human assessments, we contemplate the possibility that the automatic assessments might even surpass the human in terms of quality. While not feasible at the current time, this is an important case to consider when we evaluate the HMC evaluation.

4.2.3.1 Use LLMs to Help Humans in Annotating Documents

LLMs could be successfully applied in helping human assessors with annotating data. For example, LLMs might be particularly useful in recognizing near duplicates and using them to verify if the two near duplicates share the same relevance annotation – with the human entering the loop only in those cases where the system has a high degree of uncertainty.

Related to the case of (100%H / 0%M), we have the *human-in-the-loop*, helping the system in realizing its annotation goal. Such help might include providing annotated facts or verifying the annotation after a first pass from the system. In the 50%/50% case, equal contributions, we have a substantial equilibrium between both the human and the machine. We refer to this scenario as *competence partitioning*: the task is assigned to either the human or the machine, depending on who is currently better at the current moment. On the other side of the spectrum (%M > %H), the scenario is called *model-in-the-loop*: the model offers its contribution in organizing the data, where the human is used as a verification step. The concern is that any bias in the LLM might be affecting the relevance assessments, as the human will not be able to correct for information it will not see.

An alternative approach to the collaborative one is a complementary one, where the human and the machine both produce judgments, but different ones. This then becomes a task allocation problem where the aim is to predict who among the human and the machine assessor is best suited for any given judgment.

4.2.3.2 Beyond Human Performance

We could expect that, at a certain point in the future, the LLMs will overcome humans in a number of tasks that can be reconducted to annotate the documents. Humans are likely to make mistakes when annotating documents and are limited in the time dedicated to the annotation. In contrast, LLMs are likely to be more self-consistent and potentially capable of annotating all the documents perfectly. Machines can also annotate a much larger number of data points.

Furthermore, we have a series of assumptions, such as the fact that relevance does not change through time, that are enforced to make evaluation tractable. These assumptions can be relaxed if the machine annotates automatically.

It is an open issue in recognizing when the human is failing. All the above strategies assume human annotations are the gold standard without errors. This assumption is strong: the LLM, having access to more information, might be able to correct human mistakes.

We are likely to reach the limit of measurement: we will not be able to use differences between the current evaluation paradigms to evaluate such models. A problem is that if we surpass the quality of only human-annotated data, we will not be able to detect this if we use only human-annotated data as a gold standard. will not suffice and will fail in providing a gold standard.

Another research question is to identify optimal competence partitioning. One idea is to use the LLM to generate rationales for explaining the relevance. While humans are often considered experts for rational generation, recent advancements, including chatGPT, suggest that we are on the verge of a shift of paradigm, with LLMs constantly improving in identifying why a document is (non)-relevant, either considering information with the document, or other relevant external pieces of information.

4.2.3.3 Trust, Correctness, and Inter-annotator Agreement

One important difference between humans and automatic assessors concerns the assessment sample size. While it is possible to hire multiple assessors to annotate the documents and, possibly, resolve disagreements between annotators, this is not that trivial in the automatic assessor case. We can expect that LLMs which are trained on similar corpora will likely produce correlated answers – but we don't know whether these are correct. A possible solution to this would include the usage of different subcorpora based on different sets of documents. This, in turn, could lead to personalized LLMs, fine-tuned on data from different types of users, which would allow to auto-annotate documents directly according to the user's subjective point of view, while also helping with increasing the pool of annotations collected. While this technology is not available yet, mostly due to computational reasons, we expect it to be available in a few years.

A related idea that can be implemented today is to allow LLMs to learn by observing human annotators performing the task or following an active learning paradigm. The LLM starts with mild suggestions to the user on how to annotate the documents, then it continues to learn by considering actual decisions made by the annotator and finally improving the quality of the suggestions provided.

4.2.4 Next Steps

Tables 1 and 2 report two examples of document annotation done with two well-known LLMs: YouChat[22] and ChatGPT[23]. It is interesting to notice that, in both cases, both models provided the correct answer, correctly identifying the passage which was annotated as more relevant. It is possible to observe that, while ChatGPT simply repeats the relevant passage, YouChat is capable of correctly identifying the reason why a passage is more relevant than the other.

To assess the feasibility of the proposed approaches, next steps would include an experimental comparison of the different Collaborative-Human-Machine paradigms. This should include multiple test collections (e.g., TREC-8 and TREC Deep Learning), multiple types of judgments (e.g., binary, graded, preference), and multiple models (e.g., GPT-2, GPT-3, chatGPT, etc.). Comparison between human-generated judgments and machine-generated judgments may be performed both using inter-assessor agreement metrics as well as IR system ranking correlation methods.

[22] https://you.com/

[23] https://chat.openai.com/

Table 1 Response of two well-known LLMs in comparing two passages as answer for the query 54_8 from CAsT '19: "What are anemia's possible causes?"

Prompt	Select the passage that answers the question better. – If both answers are similar, select the one with the least extraneous information. – If both answers are still similar, select the one with the best formatting. Question: What are anemia's possible causes? Passage 1: If you have anemia, your blood does not carry enough oxygen to the rest of your body. The most common cause of anemia is not having enough iron. Your body needs iron to make hemoglobin. Hemoglobin is an iron-rich protein that gives the red color to blood. It carries oxygen from the lungs to the rest of the body. Anemia has three main causes: blood loss, lack of red blood cell production, and high rates of red blood cell destruction passage 2: Broadly, causes of anemia may be classified as impaired red blood cell (RBC) production, increased RBC destruction (hemolytic anemias), blood loss and fluid overload (hypervolemia
ChatGPT	Passage 1: If you have anemia, your blood does not carry enough oxygen to the rest of your body. The most common cause of anemia is not having enough iron. Your body needs iron to make hemoglobin. Hemoglobin is an iron-rich protein that gives the red color to blood. It carries oxygen from the lungs to the rest of the body. Anemia has three main causes: blood loss, lack of red blood cell production, and high rates of red blood cell destruction.
YouChat	Passage 1 is the better answer, as it provides an explanation of what anemia is, and then lists the three main causes in detail, which is the most relevant information for the question. Passage 2, while providing some good background information, does not provide the same level of specificity as Passage 1, and includes extraneous information that is not directly relevant to the question.

Table 2 Response of two well-known LLMs in comparing two passages as answer for the query 67_8 from CAsT '19: "What is a DC half smoke?"

Prompt	Select the passage that answers the question better. – If both answers are similar, select the one with the least extraneous information. – If both answers are still similar, select the one with the best formatting. Question: What is a DC half smoke? Passage 1: However, celebrities at Ben's are not limited to American citizens. In 1998, former DC mayor Marion Barry described having traveled to Ghana and meeting the Mayor of Accra, an alumnus of Howard University, whose greeting was, "Glad to have you in Accra. Is Ben's Chili Bowl still there?" And when French President Nicolas Sarkozy and his wife Carla Bruni-Sarkozy visited Washington, D.C. in March 2010, they reportedly each had two of Ben's half smokes during their visit to the restaurant. Passage 2: A half-smoke is a "local sausage delicacy" found in Washington, D.C. and the surrounding region. Similar to a hot dog, but usually larger, spicier, and with more coarsely-ground meat, the sausage is often half-pork and half-beef, smoked, and served with herbs, onion, and chili sauce.
ChatGPT	Passage 2: A half-smoke is a local sausage delicacy found in Washington, D.C. and the surrounding region. Similar to a hot dog, but usually larger, spicier, and with more coarsely-ground meat, the sausage is often half-pork and half-beef, smoked, and served with herbs, onion, and chili sauce.
YouChat	Passage 2 is the better answer as it more directly answers the question and does not include any extraneous information.

References

1 Stefan Büttcher, Charles L. A. Clarke, Peter C. K. Yeung, and Ian Soboroff. Reliable information retrieval evaluation with incomplete and biased judgements. In Wessel Kraaij, Arjen P. de Vries, Charles L. A. Clarke, Norbert Fuhr, and Noriko Kando, editors, *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 63–70. ACM, 2007.

2 Charles L. A. Clarke, Alexandra Vtyurina, and Mark D. Smucker. Assessing top-k preferences. *ACM Trans. Inf. Syst.*, 39(3):33:1–33:21, 2021.

3 Donna Harman. Information retrieval evaluation. 2011.

4 Martin Potthast, Lukas Gienapp, Florian Euchner, Nick Heilenkötter, Nico Weidmann, Henning Wachsmuth, Benno Stein, and Matthias Hagen. Argument search: Assessing argument relevance. In Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer, editors, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1117–1120. ACM, 2019.

5 David P. Sander and Laura Dietz. EXAM: how to evaluate retrieve-and-generate systems for users who do not (yet) know what they want. In Omar Alonso, Stefano Marchesin, Marc Najork, and Gianmaria Silvello, editors, *Proceedings of the Second International Conference on Design of Experimental Search & Information REtrieval Systems, Padova, Italy, September 15-18, 2021*, volume 2950 of *CEUR Workshop Proceedings*, pages 136–146. CEUR-WS.org, 2021.

6 Ellen M. Voorhees. The philosophy of information retrieval evaluation. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001, Revised Papers*, volume 2406 of *Lecture Notes in Computer Science*, pages 355–370. Springer, 2001.

4.3 Overcoming Methodological Challenges in Information Retrieval and Recommender Systems through Awareness and Education

Christine Bauer (Utrecht University, NL, c.bauer@uu.nl)
Maik Fröbe (Friedrich-Schiller-Universität Jena, DE, maik.froebe@uni-jena.de)
Dietmar Jannach (University of Klagenfurt, AT, dietmar.jannach@aau.at)
Udo Kruschwitz (University of Regensburg, DE, udo.kruschwitz@ur.de)
Paolo Rosso (Technical University of Valencia, ES, prosso@dsic.upv.es)
Damiano Spina (RMIT University, AU, damiano.spina@rmit.edu.au)
Nava Tintarev (Maastricht University, NL, n.tintarev@maastrichtuniversity.nl)

4.3.1 Background & Motivation

In recent years, we have observed a substantial increase in research in IR and RS. To a large extent, this increase is fueled by progress in ML (deep learning) technology. As a result, countless papers are nowadays published each year which report that they improved the state-of-the-art when adopting common experimental procedures to evaluate ML based systems. However, a number of issues were identified in the past few years regarding these reported findings and their interpretation. For example, both in IR and RS, studies point to methodological issues in *offline* experiments, where researchers for example compare their models against weak or non-optimized baselines or where researchers optimize their models on test data rather than on held-out validation data [4, 13, 48, 53].

Besides these issues in offline experiments, questions concerning the *ecological validity* of the reported findings are raised increasingly. Ecological validity measures how generalizable experimental findings are to the real world. An example of this problem in information retrieval is the known problem of mismatch between offline effectiveness measurement and user satisfaction measured with online experimentation [10, 5, 40, 46, 56] or when the definition of relevance does not consider the effect on a searcher and their decision-making. For example, the order of search results, and the viewpoints represented therein, can shift undecided voters toward any particular candidate if high-ranking search results support that candidate [19]. This phenomenon – often referred to as the *Search Engine Manipulation Effect (SEME)* – has been demonstrated for both politics [19, 20] and health [2, 43]. By being aware of the phenomena, methods have been adapted to measure its presence [14, 15], and studies to evaluate when and how it affects human decision-makers [16]. Similar questions of ecological validity were also raised in the RS field regarding the suitability of commonly used computational accuracy metrics as predictors of the impact and value such systems have on users in the real world. Several studies indeed indicate that the outcomes of offline experiments are often *not* good proxies of real-world performance indicators such as user satisfaction, engagement, or revenue [7, 25, 30].

Overall, these observations point to a number of open challenges in how experimentation is predominantly done in the field of information access systems. Ultimately, this leads to the questions of *(i)* how much progress we really make despite the large number of research works that are published every year [4, 35, 57] and *(ii)* how effective we are in sharing and translating the knowledge we currently have for doing IR and RS experimentation [23, 45]. One major cause for the mentioned issues, for example, seems to lie in the somewhat narrow way we tend to evaluate information retrieval and recommender systems: primarily based on various computational effectiveness measures. In reality, information access systems are interactive systems used over longer periods of time, i.e., they may only be assessed holistically if the user's perspective (task and context) is taken into account, cf. [36, 51, 55]. Studies on long-term impact furthermore need to consider the wider scope of stakeholders [6, 30]. Moreover, for several types of information access systems, the specific and potentially competing interests of multiple stakeholders have to be taken into account [6]. Typical stakeholders in a recommendation scenario include not only the consumers who receive recommendations but also recommendation service providers who for example want to maximize their revenue through the recommendations [29, 30].

Various factors contribute to our somewhat limited view of such systems, e.g., the difficulties of getting access to real systems and real-world data for evaluation purposes. Unfortunately, the IR and RS research communities to a certain extent seem to have accepted to live with the limitations of the predominant evaluation practices of today. Even more worryingly, the described narrow evaluation approach has become more or less a standard in the scientific literature, and there is not much debate and – as we believe – sometimes even limited awareness of the various limitations of our evaluation practices.

There seems to be no easy and quick way out of this situation, even though some of the problems are known for many years now [17, 5, 32, 46]. However, we argue that improved *education* of the various actors in the research ecosystem (including students, educators, and scholars) is one key approach to improve our experimentation practices and ensure real-world impact in the future. As will be discussed in the next sections, better training in experimentation practices is not only important for students, but also for academic teachers, research scholars, practitioners and different types of decision-makers in academia, business, and other organizations. This will, in fact, help address the much broader problem of reproducibility[24] and replicability [25] we face in Computer Science [12, 1] in general and in AI in particular [26].

This chapter is organized as follows: Next, in Section 4.3.2 we briefly review which kinds of actors may benefit from better education in information access system experimentation. Afterwards, in Section 4.3.3, we provide concrete examples of what we can do in terms of concrete resources and initiatives to increase the awareness and knowledge level of the different actors. Finally, in Section 4.3.4, we sketch the main challenges that we may need to be aware of when implementing some of the described educational initiatives.

4.3.2 Actors

As in any process related to the advancement, communication, and sharing of knowledge, knowing how to properly design and carry out correct and robust experimentation concerns people with various different roles.

This covers a broad spectrum including academia, industry, and public organizations, e.g., from a lecturer in IR and RS introducing evaluation paradigms to undergrad students and data scientists – not necessarily experienced in IR and RS – choosing metrics aligned to business Key Performance Indicators (KPIs) by looking at textbooks and Wikipedia pages. We have identified a number of actors that are involved in the education to experimentation in information access, who are listed below. Note that this categorization is not exhaustive nor exclusive, as actors may have multiple roles.

Students

This category embraces the different stages of academic training. Starting from students enrolled in IR & RS courses [41], including, for instance, undergraduate students in Computer Science degrees and Master's students in Data Science, AI, and Human-Computer Interaction. It also includes students enrolled in a doctoral degree, i.e., PhD students, including those jointly co-supervised with industry.

Educators

Academic roles related to education, such as course coordinators, lecturers, teaching assistants, as well as research student supervisors.

Scholars

Researchers and academics involved in academic services, including reviewers, journal editors, program chairs, grant writers, etc.

[24] https://www.wired.com/story/machine-learning-reproducibility-crisis/

[25] https://cacm.acm.org/magazines/2020/8/246369-threats-of-a-replication-crisis-in-empirical-computer-science/abstract

Practitioners

Data scientists, developers, User Experience (UX) designers, and other practitioners outside academia, that may need support in their lifelong learning.

Decision-makers

People that make strategic decisions in processes, policies, products and/or human resources (e.g., managers in industry or policy-makers) that may benefit from having a better understanding of IR and RS core concepts in evaluation and experimentation.

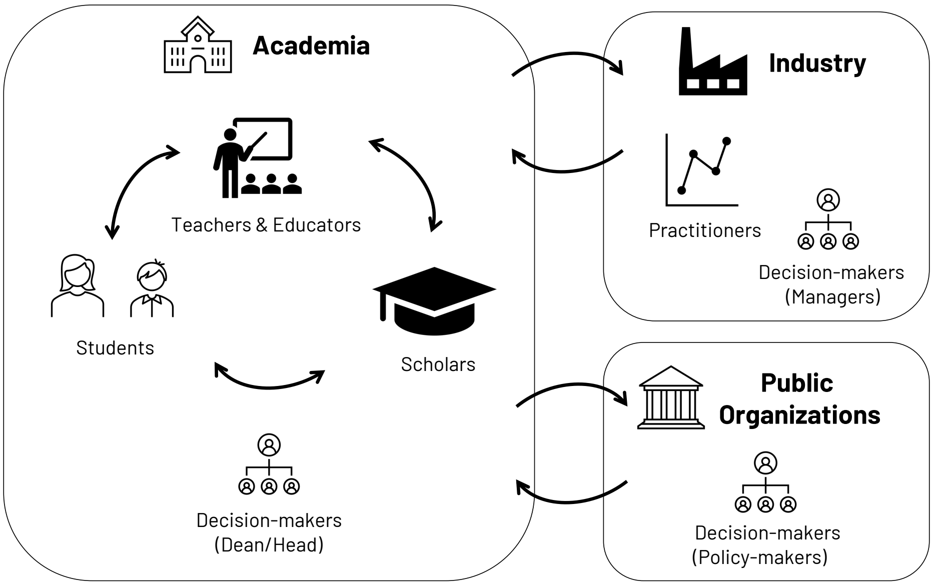


Figure 6 Interaction among actors involved in IR and RS experimental education.

Figure 6 shows the interaction among the identified actors. In academia, students, educators, and scholars are in continuous interaction through learning, teaching, and supervision processes, which are overseen and/or led by decision-makers such as deans, heads of departments, etc. In industry, decision-makers such as product and team managers, as well as practitioners, make use of training and education resources and initiatives to support experimentation in real-world domains. The cyclic arrows represent the active participation in the creation and development of those resources and initiatives. Decision-makers in public organizations, such as policy-makers, are also key actors in the definition of curricula, which has a direct impact on how and to which extent experimentation in IR and RS is included in Data Science, Computer Science, Computer-Human-Interaction (CHI), and AI programs.

4.3.3 What can we do?

In this section, we first provide examples of helpful *resources* to improve education in IR and RS evaluation. Then, we outline several possible *initiatives* that contribute to increasing awareness about current methodological issues and to disseminate knowledge about experimentation approaches.

4.3.3.1 Resources

The resources with which the actors interact are a way to share, maintain, and promote best practices while ensuring a low barrier of entry to the field. Given that those resources might be widely used in education, research (experimentation, etc.), and even production systems, resources have great potential to continuously grow the knowledge of future generations of scholars, practitioners, and decision-makers.

General Teaching Material. Textbooks quickly may become outdated,[26] but have the advantage that these typically reach a wide audience, whereas slides and tutorials that cover evaluation methodology in more depth might only reach smaller audiences. Often, today's online lectures primarily report on "mainstream" information retrieval (e.g., offline studies, common metrics), but foster reflection and discussion only to a very limited extent. More comprehensive resources should be made publicly available and shared across universities, summer schools, and meetups.[27] Finally, having the IR and RS community actively contribute to the curation of material in sources that are widely used by the general public – and, thus, also by students – as a starting point to get a basic understanding of a topic (e.g., Wikipedia) is advisable. Further, contributing to the documentation of software such as Apache Solr,[28] Elasticsearch,[29] Surprise,[30] Implicit,[31] etc. (see the report by Ferro et al. [22] for more that are widely used in practice), can help to make non-experts more aware of the best practices in IR and RS experimentation.

Apart from introducing modern information retrieval systems, **teaching material** should give more attention to a wider set of application fields of IR, including recommender systems and topics related to query and interaction mining and understanding, and online learning to rank [41]. To date, also online evaluation falls short in such resources although it is essential in the spectrum of evaluation types [41]. Students need to be introduced to concepts such as reproducibility and replicability, and it is essential that students understand what makes a research work impactful in practice. To lower the entry barrier to the field, students should be taught how to use available tools and environments that enable quick prototyping, and that have real-world relevance. Teaching fairness, privacy, and ethical aspects, both in designing experiments and also in how to evaluate them, is also important.[32]

Moreover, the participation in **shared tasks (challenges or competitions)** of evaluation campaigns in IR (e.g., TREC,[33] CLEF,[34] NTCIR,[35] or FIRE[36]) and RecSys (e.g., the yearly ACM RecSys challenges[37]) should be fostered. To facilitate the participation of

[26] In contrast to that, the main textbook in the area of natural language processing has for years only been available as an online draft and is continuously being updated: `https://web.stanford.edu/~jurafsky/slp3/`

[27] For instance, Sebastian Hofstätter released Open-Source Information Retrieval Courses: `https://github.com/sebastian-hofstaetter/teaching`.

[28] `https://solr.apache.org/`

[29] `https://www.elastic.co/es/elasticsearch/`

[30] `https://surpriselib.com/`

[31] `https://implicit.readthedocs.io`

[32] Cyprus Center for Algorithmic Transparency (CyCAT) project: `https://sites.google.com/view/biasvisualizationactivity/home`

[33] `https://trec.nist.gov/`

[34] `https://www.clef-initiative.eu/`

[35] `https://research.nii.ac.jp/ntcir/`

[36] `https://fire.irsi.res.in/fire/`

[37] `https://recsys.acm.org/challenges/`

students, it is worthwhile to make the timelines of such challenges and competitions compatible with the academic (teaching) schedules (e.g., in terms of semesters). Students will be provided with the datasets used in the benchmarks and will be able to learn more on evaluation methodologies (for instance, students from Padua, Leipzig, and Halle participated in Touché [8, 9] hosted at CLEF). At the same time, it is important to critically reflect with students on the limitations and dangers of competitions [11] and encourage them to go beyond leaderboard State Of The Art (SOTA) chasing culture – e.g., only optimizing on one metric or a limited set of metrics without reflection of the suitability of these metrics in a given application context [50, 30]. Hence, it is important that a student's (or student group's) grade does not depend on their rank in the leaderboard but to a large degree on their approach, reasoning, and reflection to counteract SOTA chasing and help students to focus on insights. Inspired by result-blind reviewing in Section 4.4, we might refer to this as "result-blind grading".

Test collections[38] and **runs/submissions** – typically combined with novel evaluation methodologies – are the main resources resulting from shared tasks or evaluation campaigns. Integrating the resulting test collections into tools such as `Hugging Face datasets` [34], `ir_datasets` [38] or `EvALL` [3] allows for unified access to a wide range of datasets. Furthermore, some **software components** such as `Anserini` [52], `Capreolus` [54], `PyTerrier` [39], `OpenNIR` [37], etc., can directly load test collections integrated into `ir_datasets` which substantially simplifies data wrangling for scholars of all levels. For instance, PyTerrier allows for defining end-to-end experiments, including significance tests and multiple-test correction, using a declarative pipeline and is already used in research and teaching alike (e.g., in a master course with 240 students [39]). Other resources for performance modeling and prediction in RS, IR, and NLP can also be found in the manifesto of a previous Dagstuhl Perspectives Workshop [22]. The broad availability of such resources makes it tremendously easier to replicate and reproduce approaches that were submitted to a shared task (challenge) before. Further, it lowers the entry barrier to experiment with a wider set of datasets and approaches across domains as switching between collections will be easy. New test collections can be added with limited effort. Still, further promoting the practice of sharing code and documentation,[39] or using software submissions with tools such as TIRA [24, 44] in shared tasks is important.

Combining and integrating the resources listed above in novel ways has the potential to reduce or even remove barriers between research and education, ultimately enabling Humboldt's ideal to combine teaching and research. Students who participate in shared tasks as part of their curriculum already go in this direction [18]. Continuously maintaining and promoting the integration of test collections and up-to-date best practices for shared tasks into a shared resource might further foster student participants because it becomes easier to "stand on the shoulders of giants" yielding to the cycle of education, research, and evaluation that is streamlined by ECIR, CLEF, and ESSIR (see Section 3.14).

4.3.3.2 Initiatives

We have identified a range of actors, and we argue that addressing the problems around education requires a number of different initiatives some of which target one particular type of actor but more commonly offer benefits for different groups. These initiatives should not

[38] In IR, an offline test collection is typically composed of a set of topics, a document collection, and a set of relevance judgments.

[39] `https://www.go-fair.org/fair-principles/`

be seen in isolation as our vision is in line with what has been proposed in Section 3.14 which calls for coordinated action around education, evaluation, and research. Here we will discuss instruments we consider to be essential on that path. There is no particular order in this discussion other than starting with well-established popular concepts.

Summer schools are a key instrument primarily aimed at graduate students. ESSIR[40] is a prime example of a summer school focusing on delivering up-to-date educational content in the field of IR; the Recommender Systems Summer School is organized in a similar manner focusing on RS. Beyond the technical content, summer schools do also serve the purpose of community-building involving different actors, namely students and scholars. Annually organized summer schools appear most effective as they make planning easier by integrating them into the annual timeline of IR- and RS-related events. This is in line with the *flow-wise* vision discussed earlier in Section 3.14.

Summer schools also provide a good setting to embed (research-focused) **Mentoring** programs and **Doctoral Consortia**. This allows PhD students as well as early-career researchers to learn from experts in the field outside their own institutions. Both instruments are well-established in the field. However, even though the established summer schools are repeatedly organized, these often happen on an irregular basis (sometimes yearly, sometimes with longer breaks) and using different formats. This irregular setting makes it difficult to integrate it into a PhD student's journey from the outset. Currently, Mentoring is often merely a by-product of other initiatives such as Summer Schools and Doctoral Consortia. It may be a fruitful path to see mentoring programs as an independent (yet, not isolated) initiative. For instance, the "Women in Music Information Retrieval (WiMIR) Mentoring program"[41] sets an example of a sustainable initiative that is organized independently of other initiatives and on yearly basis. A similar format seems a fruitful path to follow in the IR and RS communities, where it is advisable to facilitate exchange across (sub-)disciplines and open up the initiative to the entire community. We note that – similar to the WiMIR – mentoring may not only address PhD students but is well suited also for later-career stages.

While the IR and RS communities have a tradition of research-topic-driven **Tutorials** as part of the main conferences, **Courses** that address skills and practices beyond research topics (similar to courses hosted by the CHI conference[42]) would be an additional fruitful path to follow. Such courses may, for instance, address specific research and evaluation methods on an operational level[43] or how to write better research papers for a specific outlet or community[44]. With regard to support in writing better papers, see also Section 4.5.

In Bachelor and Master education, more resources in the form of Formal Educational Materials could be developed. For example, students could benefit from The Black Mirror Writers' Room exercise[45] which helps convey ethical thinking around the use of technology. Participants choose current technologies that they find ethically troubling and speculate about what the next stage of that technology might be. They work collaboratively as if they were science fiction writers, and use a combination of creative writing and ethical speculation to consider what protagonist and plot would be best suited to showcase the potential negative

[40] https://www.essir.eu

[41] https://wimir.wordpress.com/mentoring-program/

[42] https://chi2023.acm.org/for-authors/courses/accepted-courses/

[43] See, e.g., CHI 2023's C12: Empirical Research Methods for Human-Computer Interaction https://chi2023.acm.org/for-authors/courses/accepted-courses/#C12, C18: Statistics for CHI https://chi2023.acm.org/for-authors/courses/accepted-courses/#C18

[44] See, e.g., CHI 2021's C02: How to Write CHI Papers [42]

[45] https://discourse.mozilla.org/t/the-black-mirror-writers-room/46666

consequences of this technology. They plot episodes, but then also consider what steps they might take now (in regulation, technology design, social change) that might result in *not* getting to this negative future. More experienced Bachelor students and Master students could have assessments similar to paper reviews as part of their curriculum to practice critical thinking.

Topically relevant **Meetups** ranging from informal one-off meetings to more regular thematically structured events offer a much more flexible and informal way to learn about the field. Unlike summer schools they bring together the community for an evening and cater for a much more diverse audience involving *all* actors with speakers as well as attendees from industry, academia and beyond. Talks range from specific use cases of IR in the industry (e.g., search at Bloomberg), to the latest developments in well-established tools (such as Elasticsearch) to user studies in realistic settings. There is a growing number of information-retrieval-related and recommender-systems-related Meetups[46] and many of which have become more accessible recently as they offer virtual or hybrid events. Meetups offer a low entry barrier in particular for students at all levels of education and they help participants obtain a more holistic view of the challenges of building and evaluating IR and RS applications. Loosely incorporating Meetups in the curriculum, in particular when there is alignment with teaching content (e.g., **joint seminars**), has been demonstrated to be effective in our own experience. These joint initiatives may go beyond the dissemination of content, but also involve practitioners as well as decision-makers in terms of facilitating (or hindering) strategic alliances or setting strategic themes.

Knowledge Transfer through **collaboration between industry and academia** is another instrument offering a mutually beneficial collaboration between three key actors: PhD students, academic scholars, and practitioners in the industry. By tackling real-world problems (as defined by the industrial partner) using state-of-the-art research approaches in the fields of IR and RS (as provided by the academic partner) knowledge does not just flow in one direction but both ways. In the context of our discussion, this is an opportunity to gain insights into evaluation methods and concerns in the industry. There are well-established frameworks to foster knowledge transfer such as Knowledge Transfer Partnerships[47] in the UK with demonstrated impact in IR[48] and beyond.

Knowledge transfer should also be facilitated and supported at a higher level at conferences and workshops. This is where the RS community is particularly successful in attracting industry contributions to the RecSys conference series. In IR, there is still an observable gap between key academic conferences such as SIGIR and practitioners' events like Haystack (*"the conference for improving search relevance"*[49]). The annual Search Solutions conference is an example of a successful forum to exchange ideas between all different actors.[50]

With a view to improving evaluation practices in the long-term, the reviewing process and practices play an important role. Hence, **addressing reviewers and editors** is essential. Reviewers are important actors in shaping what papers will be published and which not. And it is essential that good evaluation is acknowledged and understood while poorly evaluated

[46] See, e.g., https://opensourceconnections.com/search-meetups-map/, https://recommender-systems.com/community/meetups/

[47] http://ktp.innovateuk.org

[48] https://www.gov.uk/government/news/media-tracking-firm-wins-knowledge-transfer-partnership-2015

[49] https://haystackconf.com

[50] https://www.bcs.org/membership-and-registrations/member-communities/information-retrieval-specialist-group/conferences-and-events/search-solutions/

papers are not let through. Similarly, it is crucial to have reviewers who acknowledge and understand information retrieval and recommendation problems in their broader context (e.g., tasks, users, organizational value, user interface, societal impact) and review papers accordingly. Hence, it is essential to develop educational initiatives concerning evaluation that address current and future reviewers (and editors) accordingly. Promising initiatives include the following:

- Clear reviewer guidelines acknowledging the wide spectrum of evaluation methodology and the holistic view on information retrieval and recommendation problems. For example, CHI[51] and Association for Computational Linguistics (ACL)[52] provide detailed descriptions of what needs to be addressed and considered in a review and what steps to take.[53] Care has to be taken, though, that such guidelines are kept concise to not overwhelm people before even starting to read. Further suggestions on results-blind reviewing and guidance for authors can be found in Sections 4.4 and Section 4.5 respectively.
- Next to reviewers, meta-reviewers and editors is another entity to address, which can be done in a similar manner as addressing reviewers. These senior roles can have strong momentum in inducing change – but have a strong power position in preventing it. Stronger resistance might be expected on that (hierarchical) level. Seemingly, only a few conferences and journals – for instance, ACL[54] – seem to offer clear guidelines for the meta-reviewing activity.
- Similar to courses on research methods or addressing paper-writing skills, it is advisable to provide courses that specifically address how to peer review.[55]
- Mentored reviewing is another promising initiative to have better reviews that, on the one hand, better assess submitted papers and, on the other hand, are more constructive to induce better evaluation practices for future research. Mentored reviewing programs are, for instance, established in Psychology[56]. The MIR community[57] has a New-to-ISMIR mentoring program[58] that mainly addresses paper-writing for people who are new to the community but will likely also have an impact on reviewing practices. Similar programs could be established in the IR and RS communities with a particular focus on evaluation aspects. It is worthwhile to note that a recent study (in ML and AI) indicates that novice reviewers provide valuable contributions in the reviewing process [47].
- Summer schools mainly address (advanced) students and are also a good opportunity to include initiatives addressing reviewing.

General Public Dissemination is another important aspect that needs to be addressed. Communication in the lay language of our field is very important. Editing and curating better relevant Wikipedia pages on evaluation measures for information retrieval[59] and recommender systems[60] will increase the potential of reaching a wider audience, including potential future students. Other actions can concern publishing papers in magazines

[51] ACM CHI Conference on Human Factors in Computing Systems

[52] Association for Computational Linguistics

[53] CHI 2023 Guide to reviewing papers `https://chi2023.acm.org/submission-guides/guide-to-reviewing-papers/`; ACL's How to Review for ACL Rolling Review `https://aclrollingreview.org/reviewertutorial`; Ken Hinckley's comment on what excellent reviewing is [28].

[54] ACL's Action Editor Guide to Meta-Reviewing `https://aclrollingreview.org/aetutorial`

[55] `https://chi2023.acm.org/for-authors/courses/accepted-courses/#C16`

[56] `https://www.apa.org/pubs/journals/cpp/reviewer-mentoring-program`

[57] `https://www.ismir.net`

[58] `https://ismir2022.ismir.net/diversity/mentoring`

[59] `https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval)` [Accessed: 20-Jan-2023]

[60] `https://en.wikipedia.org/wiki/Recommender_system#Evaluation` [Accessed: 20-Jan-2023]

Table 3 Actors generating or consuming resources and initiatives related to education in evaluation for IR and RS. ✓and (✓) indicate primary and secondary actors, respectively.

Actors:	Students	Educators	Scholars	Practitioners	Decision-makers
			Resources		
Teaching Materials	✓	✓			(✓)
Shared tasks/challenges/competitions	✓	✓	✓	✓	
Test collections & runs/submissions	✓	✓	✓	✓	
Software (components)	✓	✓	✓	✓	
			Initiatives		
Mentoring: Summer schools and Doctoral Consortia	✓		✓	(✓)	
Tutorials and courses	✓		✓	✓	
Meetups	(✓)	(✓)	✓	✓	✓
Joint seminars	✓	✓		✓	(✓)
Collaboration between industry and academia	✓		✓	✓	
Reviewing	(✓)		✓		
General public dissemination	(✓)	(✓)	✓	✓	✓

with a wider and differentiated audience, such as *Communications of the ACM*[61], *ACM Inroads*[62], *ACM XRDS: Crossroads*[63], *IEEE Spectrum*[64]. One of the final goals is to make IR and RS more popular to both attract students to the field and grow a healthy ecosystem of professionals at various levels.

We have described actors, resources, and initiatives that we think are worth considering in moving forward as a community towards creating more awareness, as well as sharing and transferring knowledge on experimental evaluation for IR and RS. We summarize the participation (either primary or secondary actors) in generating and consuming these resources and initiatives in Table 3. This is not intended as a definitive list but aimed to represent the primary and secondary actors which are involved.

4.3.4 Challenges & Outlook

Given the importance of reliable and ecologically valid results, one may ask oneself which obstacles occur in the path of developing better education for experimentation and evaluation of information access systems. We see different potential barriers (and possibilities) for the different actors: students, educators, scholars, practitioners, and decision-makers. We will investigate each actor in turn.

Scholars. As has also been identified in a previous Dagstuhl Seminar [22], it is significantly harder to test the importance of assumptions in user-facing aspects of the system, such as the presentation of results or the task model, as it is prohibitively expensive to simulate arbitrarily many versions of a system and put them before users. User studies are therefore

[61] https://cacm.acm.org/
[62] https://inroads.acm.org/
[63] https://xrds.acm.org/
[64] https://spectrum.ieee.org/

also at higher risk of resulting in hypotheses that cannot be clearly rejected (non-significant results), leading to fear of criticism and rejection from paper reviewers. There are some proponents of Equivalence Testing [33][65] and Bayesian Analysis [49] in Psychology which may also be useful in Computer Science.

As LLMs are becoming a commodity, policies to educate and guide authors and reviewers in how different AI tools can (or cannot) be used for writing assistance should be discussed and defined.[66] These guidelines may inspire educators on how to characterize the role of these tools in learning & teaching environments, including assessment design and plagiarism policies[67].

In addition, a current culture of 'publish or perish' incentivizes short-term and incremental findings[68], over more holistic thinking and thoughtful comparative analysis. The problem of 'SOTA-chasing' has also been discussed in other research areas, e.g., in NLP [11]. Change in academic incentive systems both within institutions and for conferences and journals change slowly but they do evolve.

Students and Educators. Thankfully, institutions are increasingly recognizing the need for reviewing studies before they are performed, such as Ethics and Data Management plan[69]. In Bachelor and Master education, in particular, this means that instructors may require training in writing such documents, and institutions appreciate and are equipped for timely review. Therefore, planning of education would benefit from allowing sufficient time for submission, review, and revision.

In that context, teaching evaluation methodologies may require some colleagues to re-train, in which case some resistance can be expected. Improving access to training initiatives and materials at post-graduate level can support colleagues who are willing but need additional support. Various forms of informal or even organized exchange between teachers may be a helpful instrument to grow the competency of educators.

Furthermore, certain evaluation concepts and methodologies cannot be taught before certain topics are covered in the curriculum. A student in recommender systems may need to understand the difference between a classification and regression problem; or the difference between precision and recall (for a given task and user it may be more important to retrieve accurate results, or to retrieve a wider range of results) before they can start thinking about the social implications.

Moreover, some students are prone to satisfice, thinking that "good enough is good enough": there are many methodologies available for evaluation, and the options are difficult to digest in a cost-effective way at entry-level – highlighting the need for availability of tutorials and low-entry level materials as indicated earlier in Section 4.3.3. Embedding participation to shared tasks and competitions (e.g., CLEF labs or TREC tracks) which provide a common framework for robust experimentation may help overcome this challenge – although the synchronization between the semester and participation timelines may not be straightforward.

[65] See also https://cran.r-project.org/web/packages/TOSTER/TOSTER.pdf

[66] For instance, see the ACL 2023 Policy on AI Writing Assistance: https://2023.aclweb.org/blog/ACL-2023-policy/.

[67] https://www.theatlantic.com/technology/archive/2022/12/chatgpt-ai-writing-college-student-essays/672371/

[68] https://harzing.com/resources/publish-or-perish

[69] Further proposals for methodological review are also under discussion in Psychology, but will likely take longer to reach Computer Science: https://www.nature.com/articles/d41586-022-04504-8

Finally, there is a growing number of experiments in developing multi-disciplinary curricula – with the appreciation that different disciplines bring to such a program. Successful initiatives include group projects consisting of students in both Social Sciences and Humanities (SSH) and Computer Science. In fact, one of the underlying principles of the continuously growing *iSchools consortium*[70] is to foster such interdisciplinarity. The challenge here is not only the design of the content but also accreditation and support from the strategic level of institutions.

Practitioners. Maintenance of resources used to translate knowledge about models and methodologies for evaluation is challenging given the fast pace of the field. This can make it hard to compare results across studies and to keep up with the SOTA of best practices in experimentation. In this regard lowering the entry barrier to participating in initiatives such as shared tasks/challenges [21, 27] and maintaining documentation of resources commonly used by non-experts are increasingly helpful.

Another issue is the homogeneity of actors. Often there is no active involvement of actors outside a narrow academic Computer Science sphere, who otherwise might have indicated assumptions or limitations early on. It can be challenging to set up productive collaborations between industry and academia, as well as across disciplines. Typical issues include, for instance, common terminology used in a different way, or different levels of knowledge of key performance indicators. Co-design in labs has set a good precedent in this regard. Examples are ICAI in the Netherlands[71], its extension in the new 10-year ROBUST initiative[72], and the Australian Centre of Excellence for Automated Decision-Making and Society (ADM+S)[73], where PhDs in multiple disciplines (Social Sciences & Humanities, Computer Science, Law, etc.) are jointly being trained in shared projects.

Research Advisory Boards are another effective instrument to draw in practitioners but here the challenge is to make the most of the little time that is usually available for the exchange of ideas between practitioners and academics.

Decision-makers. The output of evaluation and experimentation in IR and RS may be used to inform decision-making on the societal level. Consequently, if the evaluation is poorly done, or the results incorrectly generalized, the implications may also be poor decision-making with far-reaching impacts on society, e.g. [31, Ch. 10].

The ability of the other actors to support education on evaluation is constrained and shaped by decision-makers. Policy-makers in public organizations and program managers or deans in academia play a crucial role in curriculum design. Scholars and educators will have to communicate effectively the importance of experimental evaluation in information access in order to inform the decision-making process. The challenge here is to initiate change in the first place and to drive such changes. Any new initiative will necessarily involve not just a single decision-maker but more stakeholders and committees making this a more effortful but possibly also more impactful process than many of the other initiatives we have identified.

Additionally, decision-makers within academic institutions, namely libraries and career development centres, can play an important role towards developing the competency of students and educators. Making best practices in evaluation available as a commodity through these channels will require making resources more accessible for non-experts in IR and RS.

[70] https://www.ischools.org
[71] https://icai.ai/
[72] https://icai.ai/ltp-robust/
[73] https://www.admscentre.org.au/

4.3.5 Concluding Remarks

Education and dissemination represent key pillars to overcoming methodological challenges in Information Retrieval and Recommender Systems. What we have sketched here can be interpreted as a general roadmap to create more awareness among and beyond the IR and RS communities. We hope the recommendations – and the identified challenges to consider – on what we can do will help to support education for better evaluation in the different stages of the lifelong learning journey. We acknowledge that facets such as incentive mechanisms and processes in institutions are often slow-moving. The vision proposed in this section is therefore also aimed at a longer-term (5–10 years) perspective.

References

1 *Reproducibility of Data-Oriented Experiments in e-Science (Dagstuhl Seminar 16041)*, volume 6, 2016.

2 Ahmed Allam, Peter Johannes Schulz, and Kent Nakamoto. The impact of search engine selection and sorting criteria on vaccination beliefs and attitudes: Two experiments manipulating google output. *Journal of Medical Internet Research*, 16(4):e100, 2014.

3 Enrique Amigó, Jorge Carrillo de Albornoz, Mario Almagro-Cádiz, Julio Gonzalo, Javier Rodríguez-Vidal, and Felisa Verdejo. Evall: Open access evaluation for information access systems. In Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White, editors, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 1301–1304. ACM, 2017.

4 Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In David Wai-Lok Cheung, Il-Yeol Song, Wesley W. Chu, Xiaohua Hu, and Jimmy Lin, editors, *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*, pages 601–610. ACM, 2009.

5 Ahmed Hassan Awadallah, Rosie Jones, and Kristina Lisa Klinkner. Beyond DCG: user behavior as a predictor of a successful search. In Brian D. Davison, Torsten Suel, Nick Craswell, and Bing Liu, editors, *Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6, 2010*, pages 221–230. ACM, 2010.

6 Christine Bauer and Eva Zangerle. Leveraging multi-method evaluation for multi-stakeholder settings. In Oren Sar Shalom, Dietmar Jannach, and Ido Guy, editors, *Proceedings of the 1st Workshop on the Impact of Recommender Systems co-located with 13th ACM Conference on Recommender Systems, ImpactRS@RecSys 2019), Copenhagen, Denmark, September 19, 2019*, volume 2462 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019.

7 Jöran Beel and Stefan Langer. A comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems. In Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla, editors, *Research and Advanced Technology for Digital Libraries – 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015. Proceedings*, volume 9316 of *Lecture Notes in Computer Science*, pages 153–168. Springer, 2015.

8 Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Shahbaz Syed, Timon Gurcke, Meriem Beloucif, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. Overview of touché 2022: Argument retrieval. In Alberto Barrón-Cedeño, Giovanni Da San Martino, Mirko Degli Esposti, Fabrizio Sebastiani, Craig Macdonald, Gabriella Pasi, Allan Hanbury, Martin Potthast, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction – 13th International Conference of the CLEF Association, CLEF 2022, Bo-*

logna, Italy, September 5-8, 2022, Proceedings, volume 13390 of *Lecture Notes in Computer Science*, pages 311–336. Springer, 2022.

9 Alexander Bondarenko, Lukas Gienapp, Maik Fröbe, Meriem Beloucif, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. Overview of touché 2021: Argument retrieval. In K. Selçuk Candan, Bogdan Ionescu, Lorraine Goeuriot, Birger Larsen, Henning Müller, Alexis Joly, Maria Maistro, Florina Piroi, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction – 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21-24, 2021, Proceedings*, volume 12880 of *Lecture Notes in Computer Science*, pages 450–467. Springer, 2021.

10 Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. Meta-evaluation of online and offline web search evaluation metrics. In Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White, editors, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 15–24. ACM, 2017.

11 Kenneth Ward Church and Valia Kordoni. Emerging trends: Sota-chasing. *Nat. Lang. Eng.*, 28(2):249–269, 2022.

12 Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. Threats of a replication crisis in empirical computer science. *Commun. ACM*, 63(8):70–79, 2020.

13 Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In Toine Bogers, Alan Said, Peter Brusilovsky, and Domonkos Tikk, editors, *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019*, pages 101–109. ACM, 2019.

14 Tim Draws, Nirmal Roy, Oana Inel, Alisa Rieger, Rishav Hada, Mehmet Orcun Yalcin, Benjamin Timmermans, and Nava Tintarev. Viewpoint diversity in search results. In *ECIR*, 2023.

15 Tim Draws, Nava Tintarev, and Ujwal Gadiraju. Assessing viewpoint diversity in search results using ranking fairness metrics. *SIGKDD Explor.*, 23(1):50–58, 2021.

16 Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. This is not what we ordered: Exploring why biased search result rankings affect user attitudes on debated topics. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 295–305. ACM, 2021.

17 Michael D. Ekstrand, Michael Ludwig, Joseph A. Konstan, and John Riedl. Rethinking the recommender research ecosystem: reproducibility, openness, and lenskit. In Bamshad Mobasher, Robin D. Burke, Dietmar Jannach, and Gediminas Adomavicius, editors, *Proceedings of the 2011 ACM Conference on Recommender Systems, RecSys 2011, Chicago, IL, USA, October 23-27, 2011*, pages 133–140. ACM, 2011.

18 Theresa Elstner, Frank Loebe, Yamen Ajjour, Christopher Akiki, Alexander Bondarenko, Maik Fröbe, Lukas Gienapp, Nikolay Kolyada, Janis Mohr, Stephan Sandfuchs, Matti Wiegmann, Jörg Frochte, Nicola Ferro, Sven Hofmann, Benno Stein, Matthias Hagen, and Martin Potthast. Shared Tasks as Tutorials: A Methodical Approach. In *37th AAAI Conference on Artificial Intelligence (AAAI 2023)*. AAAI, 2023.

19 Robert Epstein and Ronald E. Robertson. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521, 2015.

20 Robert Epstein, Ronald E. Robertson, David Lazer, and Christo Wilson. Suppressing the search engine manipulation effect (SEME). *Proc. ACM Hum. Comput. Interact.*, 1(CSCW):42:1–42:22, 2017.

21 Nicola Ferro. What happened in CLEF \ldots for a while? In Fabio Crestani, Martin Braschler, Jacques Savoy, Andreas Rauber, Henning Müller, David E. Losada, Gundula Heinatz Bürki, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction – 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9-12, 2019, Proceedings*, volume 11696 of *Lecture Notes in Computer Science*, pages 3–45. Springer, 2019.

22 Nicola Ferro, Norbert Fuhr, Gregory Grefenstette, Joseph A. Konstan, Pablo Castells, Elizabeth M. Daly, Thierry Declerck, Michael D. Ekstrand, Werner Geyer, Julio Gonzalo, Tsvi Kuflik, Krister Lindén, Bernardo Magnini, Jian-Yun Nie, Raffaele Perego, Bracha Shapira, Ian Soboroff, Nava Tintarev, Karin Verspoor, Martijn C. Willemsen, and Justin Zobel. From evaluating to forecasting performance: How to turn information retrieval, natural language processing and recommender systems into predictive sciences (Dagstuhl Perspectives Workshop 17442). *Dagstuhl Manifestos*, 7(1):96–139, 2018.

23 Nicola Ferro and Mark Sanderson. How do you test a test?: A multifaceted examination of significance tests. In K. Selcuk Candan, Huan Liu, Leman Akoglu, Xin Luna Dong, and Jiliang Tang, editors, *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 – 25, 2022*, pages 280–288. ACM, 2022.

24 Maik Fröbe, Matti Wiegmann, Nikolay Kolyada, Bastian Grahm, Theresa Elstner, Frank Loebe, Matthias Hagen, Benno Stein, and Martin Potthast. Continuous Integration for Reproducible Shared Tasks with TIRA.io. In *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023)*, Lecture Notes in Computer Science, Berlin Heidelberg New York, 2023. Springer.

25 Carlos Alberto Gomez-Uribe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Trans. Manag. Inf. Syst.*, 6(4):13:1–13:19, 2016.

26 Odd Erik Gundersen and Sigbjørn Kjensmo. State of the art: Reproducibility in artificial intelligence. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1644–1651. AAAI Press, 2018.

27 D. K. Harman and E. M. Voorhees, editors. *TREC. Experiment and Evaluation in Information Retrieval.* MIT Press, Cambridge (MA), USA, 2005.

28 Ken Hinckley. So you're a program committee member now: On excellence in reviews and meta-reviews and championing submitted work that has merit, 2016.

29 Dietmar Jannach and Gediminas Adomavicius. Price and profit awareness in recommender systems. In *Proceedings of the ACM RecSys 2017 Workshop on Value-Aware and Multi-Stakeholder Recommendation*, Como, Italy, 2017.

30 Dietmar Jannach and Christine Bauer. Escaping the mcnamara fallacy: Towards more impactful recommender systems research. *AI Mag.*, 41(4):79–95, 2020.

31 Daniel Kahneman. *Thinking, fast and slow.* Penguin, 2011.

32 Joseph A. Konstan and Gediminas Adomavicius. Toward identification and adoption of best practices in algorithmic recommender systems research. In Alejandro Bellogín, Pablo Castells, Alan Said, and Domonkos Tikk, editors, *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation, RepSys 2013, Hong Kong, China, October 12, 2013*, pages 23–28. ACM, 2013.

33 Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, 8(4):355–362, 2017.

34 Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Sasko, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf. Datasets: A community library for natural language processing. In Heike Adel and Shuming Shi, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2021, Online and Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 175–184. Association for Computational Linguistics, 2021.

35 Jimmy Lin, Daniel Campos, Nick Craswell, Bhaskar Mitra, and Emine Yilmaz. Significant improvements over the state of the art? A case study of the MS MARCO document ranking leaderboard. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2283–2287. ACM, 2021.

36 Marianne Lykke, Ann Bygholm, Louise Bak Søndergaard, and Katriina Byström. The role of historical and contextual knowledge in enterprise search. *J. Documentation*, 78(5):1053–1074, 2022.

37 Sean MacAvaney. Opennir: A complete neural ad-hoc ranking pipeline. In James Caverlee, Xia (Ben) Hu, Mounia Lalmas, and Wei Wang, editors, *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 845–848. ACM, 2020.

38 Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan, and Nazli Goharian. Simplified data wrangling with ir_datasets. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2429–2436. ACM, 2021.

39 Craig Macdonald, Nicola Tonellotto, Sean MacAvaney, and Iadh Ounis. Pyterrier: Declarative experimentation in python from BM25 to dense retrieval. In Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong, editors, *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 – 5, 2021*, pages 4526–4533. ACM, 2021.

40 Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jingtao Song, Min Zhang, Shaoping Ma, Jiashen Sun, and Hengliang Luo. When does relevance mean usefulness and user satisfaction in web search? In Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel, editors, *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 463–472. ACM, 2016.

41 Ilya Markov and Maarten de Rijke. What should we teach in information retrieval? *SIGIR Forum*, 52(2):19–39, 2018.

42 Lennart E. Nacke. How to write CHI papers, online edition. In Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, and Takeo Igarashi, editors, *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama Japan, May 8-13, 2021, Extended Abstracts*, pages 126:1–126:3. ACM, 2021.

43 Frances A. Pogacar, Amira Ghenai, Mark D. Smucker, and Charles L. A. Clarke. The positive and negative influence of search results on people's decisions about the efficacy of

medical treatments. In Jaap Kamps, Evangelos Kanoulas, Maarten de Rijke, Hui Fang, and Emine Yilmaz, editors, *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2017, Amsterdam, The Netherlands, October 1-4, 2017*, pages 209–216. ACM, 2017.

44 Martin Potthast, Tim Gollub, Matti Wiegmann, and Benno Stein. TIRA integrated research architecture. In Nicola Ferro and Carol Peters, editors, *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*, pages 123–160. Springer, 2019.

45 Tetsuya Sakai. Laboratory experiments in information retrieval – sample sizes, effect sizes, and statistical power. 40, 2018.

46 Mark Sanderson, Monica Lestari Paramita, Paul D. Clough, and Evangelos Kanoulas. Do user preferences and evaluation measures line up? In Fabio Crestani, Stéphane Marchand-Maillet, Hsin-Hsi Chen, Efthimis N. Efthimiadis, and Jacques Savoy, editors, *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19-23, 2010*, pages 555–562. ACM, 2010.

47 Ivan Stelmakh, Nihar B. Shah, Aarti Singh, and Hal Daumé III. Prior and prejudice: The novice reviewers' bias against resubmissions in conference peer review. *CoRR*, abs/2011.14646, 2020.

48 Zhu Sun, Di Yu, Hui Fang, Jie Yang, Xinghua Qu, Jie Zhang, and Cong Geng. Are we evaluating rigorously? benchmarking recommendation for reproducible evaluation and fair comparison. In Rodrygo L. T. Santos, Leandro Balby Marinho, Elizabeth M. Daly, Li Chen, Kim Falk, Noam Koenigstein, and Edleno Silva de Moura, editors, *RecSys 2020: Fourteenth ACM Conference on Recommender Systems, Virtual Event, Brazil, September 22-26, 2020*, pages 23–32. ACM, 2020.

49 Johnny van Doorn, Don van den Bergh, Udo Böhm, Fabian Dablander, Koen Derks, Tim Draws, Alexander Etz, Nathan J Evans, Quentin F Gronau, Julia M Haaf, et al. The jasp guidelines for conducting and reporting a bayesian analysis. *Psychonomic Bulletin & Review*, 28(3):813–826, 2021.

50 Ellen M. Voorhees. Coopetition in IR research. In Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, page 3. ACM, 2020.

51 Ryen W. White. *Interactions with Search Systems.* Cambridge University Press, 2016.

52 Peilin Yang, Hui Fang, and Jimmy Lin. Anserini: Enabling the use of lucene for information retrieval research. In Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White, editors, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 1253–1256. ACM, 2017.

53 Wei Yang, Kuang Lu, Peilin Yang, and Jimmy Lin. Critically examining the "neural hype": Weak baselines and the additivity of effectiveness gains from neural ranking models. In Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer, editors, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1129–1132. ACM, 2019.

54 Andrew Yates, Siddhant Arora, Xinyu Zhang, Wei Yang, Kevin Martin Jose, and Jimmy Lin. Capreolus: A toolkit for end-to-end neural ad hoc retrieval. In James Caverlee, Xia (Ben) Hu, Mounia Lalmas, and Wei Wang, editors, *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 861–864. ACM, 2020.

55 Eva Zangerle and Christine Bauer. Evaluating recommender systems: Survey and framework. *ACM Comput. Surv.*, 55(8):170:1–170:38, 2023.

56 Fan Zhang, Jiaxin Mao, Yiqun Liu, Xiaohui Xie, Weizhi Ma, Min Zhang, and Shaoping Ma. Models versus satisfaction: Towards a better understanding of evaluation metrics. In Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 379–388. ACM, 2020.

57 J. Zobel. When measurement misleads: The limits of batch assessment of retrieval systems. *SIGIR Forum*, 56(1), 2022.

4.4 Results-blind Reviewing

Joeran Beel (University of Siegen, DE, joeran.beel@uni-siegen.de)
Timo Breuer (Technische Hochschule Köln, DE, timo.breuer@th-koeln.de)
Anita Crescenzi (University of North Carolina at Chapel Hill, US, amcc@unc.edu)
Norbert Fuhr (University of Duisburg-Essen, DE, norbert.fuhr@uni-due.de)
Meijie Li (University of Duisburg-Essen, DE, meijie.li@uk-essen.de)

4.4.1 Motivation

Campbell and Stanley defined experiments as “that portion of research in which variables are manipulated and their effects upon other variables observed” (p. 1 in [1]).” Scientific experiments are used in confirmatory research to test a priori hypotheses as well as in exploratory research to gain new insights and help to generate hypotheses for future research [7]. In information access research, the ultimate goal is to gain insights into cause and effect. Unfortunately, many reviewers of information access experiments place undue emphasis on performance, rejecting papers that contain insights if they fail to show improvements in performance. The focus on performance numbers not only leads to publication bias. It also puts additional pressure on early-career researchers who must publish or perish, thus being tempted to cheat if their proposed method does not yield the desired results. Moreover, reviewers pay little attention to the experimental methodology and analysis [4] in case the results are impressive. Focusing primarily on performance (and in particular aggregated performance) can lead to a neglect of insights; gaining insights is critical to move the information access field forward and essential to be able to make performance predictions [2].

We think that one important step to change the situation is if we alter the review process such that there is more emphasis on the theoretical background, the hypotheses, the methodological plan and the analysis plan of an experiment, while improvement or decline of performance should play less of a role when deciding about the quality of a paper. It is hoped that this will lead to a higher scientific quality of publications, more insights, and improved reproducibility (as there is less incentive for beautifying results). As Woznyj et al. [8] note in their survey of editorial board members, overall there are positive attitudes towards results-blind reviewing and advantages for the scientific community outweigh concerns.

In order to move the review focus away from performance improvement, appealing to reviewers alone will not be sufficient. A more drastic measure is the change of the review process such that reviewers decide about acceptance vs. rejection of a paper without knowing the outcome of the experiments described.

Table 4 Comparison of traditional and emerging approaches to peer review: results-blind, preregistered reports, and registered reports.

	Traditional	Results-Blind	Preregistered	Registered Report
protocol preregistration	optional	optional	yes (in journal repository)	no
protocol publication (separate from research article)	no	no	no	yes
peer review of research protocol before data collection	no	no	yes	yes
peer review of paper with blinded results	no	yes	no	no
peer review of full paper	yes	yes (if in-principle acceptance)	yes with focus on results (if in-principle acceptance)	yes (if in-principle acceptance)
Example publication(s)	ACM SIGIR, ACM CHIIR	BMC Psychology	PLOS Biology	PLOS ONE

4.4.2 Current Situation and Gaps

As part of IR or RS conferences, the peer-reviewing process usually involves the review of the full paper using double-blinded reviewing, i.e., both authors and reviewers remain anonymous to each other. Before submission, authors are informed about possible reviewing criteria and areas of interest in the Call for Papers (CfP) that can be found on the conference website. Upon submission, the paper should contain all of the relevant information regarding the motivation, the research methodology or study design, the experimental results, and finally, a discussion that puts the results into context.

For each submission, usually, a group of three reviewers is assigned. All of them should align their reviews to those criteria mentioned in the CfP and, depending on the submission system, express their opinion in written text or by pre-defined answers regarding particular aspects. In addition, they can assign (overall) scores. The final decision is based on a discussion among reviewers, which is governed by an additional meta-reviewer, and consolidation with the program chairs.

Even though this traditional review model has been established for several years, it can imply negative impacts on the stakeholders or the scientific community as a whole. Under the assumption that reviewers overemphasize positive outcomes, the authors might be inclined to “search for” performance gains in system-oriented experiments at the cost of scientific rigor and reasoning. Even more, there is the danger of fraud or selecting positive outcomes, considering the need to publish in order to proceed in an academic career.

Alternatives to the traditional review process have emerged with an initial round of peer review of a manuscript with the results blinded or a study protocol and a subsequent round of peer review of the full paper including results. Table 4 shows the traditional peer review model with our recommended results-blind reviewing and two other variants, each of which we describe below. The Center for Open Science notes that, as of January

2023, over 300 journals have adopted one or more variants of this approach.[74] In addition, several preliminary analyses of their implementation have been conducted and published (e.g., [3, 5, 8]).

A results-blind review involves an in-principle acceptance or rejection decision based on peer review of the paper *with the results blinded* from the reviewers (see the third column of Table 4). The reviewers can put more emphasis on judging the merits of the general motivation, the study design, and what kinds of scientific insights could be gained from the experiments. If the paper is accepted in-principle, it proceeds to a second stage of peer review of the *paper with the results* included for reviewers. The final decision about the acceptance is based on the second stage of the review in which the reviewers have access to the experimental outcomes.

Other peer-reviewing models have emerged in recent years as part of the growing awareness of preregistration[75,76] and its adoption [6]. One such approach to peer review involves the review and in-principle acceptance of the study protocol including the methods and analysis plan before data is collected or analysis begins. Variants of this approach include preregistered research articles and registered reports for confirmatory research [77]. Although preregistered reports and registered reports are typically used for confirmatory research, there are variants for exploratory research and some journals also use a separate approach for exploratory research projects which do not have a confirmatory component (e.g., an Exploratory Report article type in journal *Cortex*).

Preregistered research articles involve researchers submitting a research study protocol including the rationale and hypotheses, methodology including analysis plan, and materials to a journal for review and simultaneous depositing into a repository often associated with the journal (see the fourth column of Table 4). The preregistered protocol is peer-reviewed with a focus on methods and the analytic approach, and a provisional in-principle acceptance conditional upon the execution of the study as designed. The researchers execute the study, analyze the results, and submit a full manuscript. After peer review of the new sections, the completed manuscript is published.

Registered Reports also involve submission and peer review of a study protocol (see the third column of Table 4). A key difference from preregistered articles is that accepted protocols are published immediately and a future article with the results of the study is given an in-principle acceptance. After the study execution, the full manuscript is submitted and reviewed.

4.4.3 Next Steps

We propose several changes to the reviewing processes for information access papers to reduce publication biases. Our recommendations are that information access scholarly community:

1. adopts a pilot test of results-blind reviewing for a conference or journal,
2. considers starting from our initial process recommendation for results-blind reviewing,
3. ask authors, conference organizers, and reviewers to place more emphasis within papers on the insights that can be gained from their research,

[74] https://www.cos.io/initiatives/registered-reports
[75] https://www.cos.io/initiatives/prereg
[76] https://plos.org/open-science/preregistration/
[77] For examples of how preregistered research articles and research reports have been implemented, see the summary provided by PLOS. https://plos.org/open-science/preregistration/

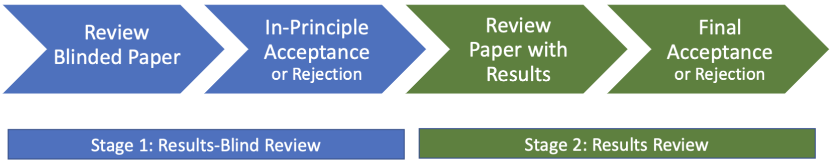

Review Blinded Paper
In-Principle Acceptance or Rejection
Review Paper with Results
Final Acceptance or Rejection
Stage 1: Results-Blind Review
Stage 2: Results Review

Figure 7 Proposed two-stage process for results-blind reviewing (figure adapted from BMC[78]).

4. considers allowing additional space for additional details about study methodology, and
5. considers whether to implement a two-stage review process in which research proposals and/or preregistered research reports are reviewed with a tentative acceptance decision before data collection and analysis are conducted.

Each of these is described in more detail below.

Recommendation 1: Pilot test of results-blind reviewing in conference(s) or journal(s)

Our first and most important recommendation is that the information access research communities (i.e., IR and RS communities) adopt a results-blind approach to peer reviewing for conference(s) and/or journal(s). We recommend that the community start with a pilot test of results-blind reviewing in an established conference track, perhaps with a new paper track with an earlier deadline to allow for a two-stage review process. In results-blind reviewing, the authors submit two versions of their manuscript: one version of the paper with the full results, and one version with the results blinded. The two submitted versions are the basis of a results-blind reviewing process with two major stages (see Figure 7).

Stage 1 consists of the Results-Blind Review. The results-blind version of the manuscript is reviewed and an in-principle acceptance (or rejection) is made. During Stage 1, as in the traditional reviewing process, the paper is reviewed by multiple reviewers who also make acceptance recommendations. In the case of conferences, the in-principle acceptance (or rejection) decision is made after discussion with the Senior Program Committee (SPC)/meta reviewer and in the Program Committee (PC) meeting. Papers that receive an in-principle acceptance proceed to Stage 2.

Stage 2 consists of the Results Review. The paper containing the results is reviewed by the same set of reviewers with a focus on the results. In the case of a conference, the final acceptance (or rejection) decision is made after a discussion period with the SPC and in the PC meeting.

Recommendation 2: Initial process recommendation for a results-blind reviewing pilot

Below, we recommend a high-level process for how a results-blind reviewing process pilot might be implemented and important considerations for conference organizers and reviewers as well as authors.

Conference organizers. Once the decision for results-blind reviewing has been made, conference organizers would have to take the following steps:

- First, the CfP for the new track should be written. As the proposed results-blind reviewing process with two stages of review will take longer to complete, an earlier deadline for this track should be set.

[78] https://www.biomedcentral.com/collections/RFPR

- Criteria for both stages of the review (blinded and with results) should be defined. Special attention should be given to the criteria for changing an initial acceptance recommendation into a rejection.
- Author instructions for the results-blind reviewing track have to be formulated, describing not only the new reviewing criteria and process but also specific instructions on how to prepare the blinded version of an article. For the results-blind version of the paper, the authors will need to blind all mentions of the results (e.g., in the abstract, introduction, discussion, and conclusion in addition to in a results section) in a way that it is not technically possible to recover the blinded text. There should be a way for reviewers to easily determine the differences between the results-blind version of the paper and the one with the results.
- Reviewers for the results-blind reviewing track have to be recruited. In the beginning, additional or different expertise will be required for this track. A special introduction of training for the reviewers might be necessary in order to make them familiar with the new process and criteria.
- The reviewing software will need to be configured for multiple stages of review for the results-blind reviewing. In the first stage of reviewing, only the blinded version of the papers should be distributed to reviewers (see below for the process for reviewers).
- After the final decision by the PC, the authors will be provided with the review and informed about the final accept or reject decision. In the case of a rejection decision, authors should also be notified at which stage the paper was rejected.
- The organizers should give special recognition to the PC member of the track (on the conference Web site and in the proceedings)
- The success of the new track and the process should be evaluated.

Reviewers. Once the reviewers are provided with instructions about the general process and received additional training, we recommend the following process:

- In the first stage, the reviewers are provided with the results-blind version of the submission and complete their review including a recommendation about the in-principle acceptance.
- Once the reviews are complete, a discussion phase with the SPC follows, leading to a recommendation for each paper.
- The PC for the track meets and makes an initial decision (in-principle acceptance or rejection) for each paper.
- For the second reviewing stage, only in-principle accepted papers are considered. Reviewers get the full versions of the papers they reviewed before. They add an additional part to their review focusing on the results which were previously blinded. Also, they make a second recommendation about acceptance.
- As for the first phase, a discussion phase with the SPC follows leading to a recommendation for each paper.
- The track PC meets for the second time and makes the final decision for each paper.

Authors. Authors will have to understand the new reviewing scheme, and possibly be trained/educated for preparing manuscripts that satisfy the new reviewing criteria. They will have to prepare and submit two versions of a paper, a version with the results as in the traditional model as well as one in which the results are blinded.

Recommendation 3: Emphasize insights in papers

We recommend that authors, conference organizers, and reviewers place additional emphasis on communicating expected insights to be gained from experiments. Guidelines (and review forms) should ask the reviewers to comment on the theoretical background, the hypotheses, the methodological plan and the analysis plan of the experiment(s) described. Special attention should be given to the expected insights to be gained from experiments, i.e. regarding cause and effect.

Recommendation 4: Extra space for methods information

Another recommendation is for the community to consider explicitly allowing methodological appendices for authors to provide additional methodological details outside of page and/or word limits and to include these appendices with the text of the paper and not as supplementary materials. While not needed for all publications, this would be very beneficial for some types of studies so that the authors can include all study materials. For example, in user studies, researchers may administer multiple questionnaires, conduct a semi-structured interview, and read from a script. It is not uncommon for researchers to administer multiple questionnaires and conduct a semi-structured interview.

This would be especially important if adopting a results-blind reviewing process as careful scrutiny of the study design and all study materials is needed to ascertain whether the authors will be able to answer the research questions. For example, due to page limits, it is common for authors to describe the topics of an interview but uncommon to include the full text of an interview guide due to page limits.

In addition, this would have an additional benefit for other researchers who wish to replicate the study. While, for example, authors can currently make supplementary materials available in ACM Digital Library (ACM DL), these materials are not included in the downloadable version of the article or when reading online in the ACM DL in the eReader or HTML formats.

Recommendation 5: Consider a two-stage review process adapted from preregistered or registered reports

Although our primary recommendation is for conference organizers or journal editors to embrace a results-blind reviewing approach, we also recommend that they consider piloting a conference track or article type in which the study protocol undergoes peer review and is accepted in-principle before data collection or analysis begins. This may be more appropriate for certain types of research (e.g., user studies).

4.4.4 Conclusion

At first glance, the new result-blind reviewing scheme might seem to be only attractive for papers describing failed experiments, while authors with successful results would go to the established tracks. In order to avoid this impression, it is essential that the new scheme is piloted as a highly visible and prestigious track in an established conference. Furthermore, it should be clearly communicated that the results-blind reviewing scheme aims at establishing high standards for the design, execution and analysis of experiments while shielding the reviewers from being blinded by shiny experimental results. Thus, it is our hope that papers published in this track will be regarded as high-quality publications which thoroughly address research questions and clearly demonstrate the insights that may be gained from the research.

References

1 Donald T. Campbell and Julian C. Stanley. *Experimental and quasi-experimental designs for research.* Houghton Mifflin Company, Boston, 1963.

2 Nicola Ferro, Norbert Fuhr, Gregory Grefenstette, Joseph A. Konstan, Pablo Castells, Elizabeth M. Daly, Thierry Declerck, Michael D. Ekstrand, Werner Geyer, Julio Gonzalo, Tsvi Kuflik, Krister Lindén, Bernardo Magnini, Jian-Yun Nie, Raffaele Perego, Bracha Shapira, Ian Soboroff, Nava Tintarev, Karin Verspoor, Martijn C. Willemsen, and Justin Zobel. From evaluating to forecasting performance: How to turn information retrieval, natural language processing and recommender systems into predictive sciences (Dagstuhl Perspectives Workshop 17442). *Dagstuhl Manifestos*, 7(1):96–139, 2018.

3 Michael G. Findley, Nathan M. Jensen, Edmund J. Malesky, and Thomas B. Pepinsky. Can Results-Free Review Reduce Publication Bias? The Results and Implications of a Pilot Study. *Comparative Political Studies*, 49(13):1667–1703, 2016. Publisher: SAGE Publications Inc.

4 Norbert Fuhr. Some common mistakes in IR evaluation, and how they can be avoided. *SIGIR Forum*, 51(3):32–41, 2017.

5 Daniel M. Maggin, Rachel E. Robertson, and Bryan G. Cook. Introduction to the special series on results-blind peer review: An experimental analysis on editorial recommendations and manuscript evaluations. *Behavioral Disorders*, 45(4):195–206, 2020.

6 Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.

7 William R Shadish, Thomas D Cook, and Donald T. Campbell. *Experimental and quasi-experimental designs for generalized causal inference.* Houghton, Mifflin and Company, New York, 2002.

8 Haley M. Woznyj, Kelcie Grenier, Roxanne Ross, George C. Banks, and Steven G. Rogelberg. Results-blind review: a masked crusader for science. *European Journal of Work and Organizational Psychology*, 27(5):561–576, 2018.

4.5 Guidance for Authors

Giorgio Maria Di Nunzio (University of Padova, IT, giorgiomaria.dinunzio@unipd.it)
Maria Maistro (University of Copenhagen, DK, mm@di.ku.dk)
Christin Seifert (University of Duisburg-Essen, DE, christin.seifert@uni-due.de)
Julián Urbano (Delft University of Technology, NL, j.urbano@tudelft.nl)
Justin Zobel (University of Melbourne, AU, jzobel@unimelb.edu.au)

4.5.1 Motivation

The IR community has over time developed a strong shared culture of expectations of published papers, particularly in our leading venues. However, these expectations are not explicit and the evidence of submitted papers is that many authors are not aware of what elements, or omissions, are likely to be of concern to reviewers. While accepted papers do provide an indication of what an author should do, they are, of course, uneven, and the small set of papers that an author is consulting in their new work could easily be unrepresentative of the best IR work as a whole.

In this section, our aim is to provide a basis for general guidance for authors and reviewers, with a focus on people who are new to the community. It should communicate to authors and reviewers a range of factors that the community regards as significant. Such guidance, if well designed, should help authors to lift the standard of their work and provide context should it not be accepted; for reviewers, especially those new to the task, it can provide checklists and (at a high level) advice about the field from beyond their immediate research environment.

Some elements in papers have attracted specific criticism in publications; this is particularly true of effectiveness measurement, where a long history of research on method has argued for and against a range of measures, forms of evidence for statistical validity, treatment of test collections, and so on. Such literature is critical to improving the quality of our research but does not necessarily represent a settled, shared view of best practice.

In our view, it is essential that general advice be constructive, readily understandable by new IR authors and reviewers, and – to the extent that is possible – not the subject of active debate. In the following, we have sought to follow this principle. We first explain the basis of the draft guidance for authors that we have developed and then present that guidance. How this work might develop over time is considered under "next steps".

4.5.2 Flaws in Submitted IR Papers

For our goal of developing draft guidelines for authors for the community, we have multiple sources of inspiration. As a first step, it is valuable to understand and list the kinds of issues that lead experienced reviewers to criticize papers, that is, to collect the opinions from the community based on their experience from different roles as scientists: authors, readers, reviewers and meta-reviewers. Another valuable source of information consists in existing guidelines in adjacent research fields, as they reflect a common agreement of what constitutes a good scientific paper in that community and point out commonly agreed issues that may lead to rejection.

By collecting, consolidating, and harmonising the collected information, we aim to establish a strong foundation for the synthesis of a new set of draft guidelines that comprehensively capture the community-agreed strengths aspects of good scientific papers as well as issues that commonly lead to rejection; and separately to identify significant emerging aspects that are not yet captured by existing guidelines.[78] To obtain concise, comprehensive, understandable, and actionable guidelines for early-career researchers, we translated the identified issues, points of criticism, and guideline items, which have been described at varying levels of detail, into observations on elements that papers should include and on elements that can lead to rejection.

We designed the following approach to create our guidelines: (1) search of existing guidelines; (2) brainstorming to identify common pitfalls; (3) categorization of the outcomes from the brainstorming exercise and comparison of these with existing guidelines; and (4) consolidation and integration with existing SIGIR guidelines.[79] Throughout each step of the process, we adhere to the principle of keeping only issues that we believe to be widely agreed upon within the community.

We now describe our approach.

[78] As an example, ACL 2023 includes a "Policy on AI Writing Assistance" in their call for papers `https://2023.aclweb.org/blog/ACL-2023-policy/`.

[79] `https://sigir.org/sigir2023/submit/call-for-full-papers/checklist-to-strengthen-an-ir-paper/`

Identifying existing guidelines

We started by searching for existing guidelines for authors and reviewers that have been proposed in adjacent research communities. In our search for existing guidelines, we considered the following sources.

- The ACM Special Interest Group on Information Retrieval (SIGIR) developed recommendations to strengthen IR papers. These are rather general suggestions concerning presentation and experimentation. We used them as the initial stage and extend them to design our list of recommendations for authors (see Section 4.5.3).
- Empirical Evaluation Guidelines from the ACM Special Interest Group on Programming Languages (SIGPLAN).[80] This is a checklist that presents best practices meant to support both authors and reviewers within the community. The checklist includes some broad categories (e.g., appropriate presentation of results) and examples of violations for each subcategory (e.g., a misleading summary of results). These are reported in Appendix 6.1.
- The Special Interest Group on CHI SIGCHI[81] published a guide for reviewing papers submitted to the CHI conference. This is a general overview of both quality considerations (e.g., whether the paper contribution is sufficiently original), and more practical considerations related to the paper length and the review process. SIGCHI also suggested the Equitable Reviewing Guide,[82] which is a list of recommendations to help reviewers write fair reviews. Some of their points include reflecting on personal bias or considering that many authors are not native English speakers, thus being lenient on writing style and typos.
- The ACL presented an online tutorial to instruct reviewers on the ACL Rolling Review process.[83] This tutorial presents some practical suggestions (e.g., planning the reading and reviewing time to avoid rushed reviews), as well as suggestions to evaluate the quality of the paper and a list of common reasons for rejection, which often lead to author complaints because such reasons are not actual weaknesses but rather easy, unreasonable grounds for rejection.
- Ulmer et al. [1] present a list of best practices and guidelines for experimental standards within NLP. These guidelines contain some broad categories, (e.g., data), and minimal requirements and recommendations for each category (e.g., publish the dataset accessibly and indicate changes). These are reported in Appendix 6.2.

Brainstorming to identify common issues

After our search for guidelines, we ran a brainstorming exercise among contributors of the working group. The goal of this exercise was to identify concerns and flaws that we, as reviewers, would not want to find in IR papers and can very likely lead to rejection. This list of reflections is included in Appendix 6.3.

We extended the brainstorming exercise to all participants in the Dagstuhl Seminar through an online survey. We asked participants to list “things we don’t like to see in papers”, and provided some examples for guidance and the full list of SIGPLAN categories for inspiration. We received 35 items. Comments concerning strategic issues, such as “I prefer

[80] https://www.sigplan.org/Resources/EmpiricalEvaluation/
[81] https://chi2022.acm.org/for-authors/presenting/papers/guide-to-reviewing-papers/
[82] https://chi2022.acm.org/for-authors/presenting/papers/equitable-reviewing-guide/
[83] https://aclrollingreview.org/reviewertutorial

to have a new paper category" were omitted from further analysis; others were integrated into our findings. As mentioned above, we adhere to the principle of keeping only issues that we did not regard as controversial issues or the subject of debate, with the aim of omitting points that might lead to disagreement in the community.

Integration and categorization

Inspired by the SIGPLAN and NLP guidelines, we developed an initial set of broad categories to organize the issues we identified above. We then mapped each item in our list of reflections to the corresponding category. We did the same for the suggestions collected from the participant survey, as well as for the pertinent points identified in the SIGPLAN and NLP guidelines and the SIGIR guidance. In this process, we focused on issues that specifically relate to IR papers and set aside more general issues such as "captions of tables should be clear".

There were several rounds of review to clarify and consolidate similar items, with minor re-categorizations when needed. The final result of this process is a list of what we believe are recognised as common flaws in IR papers. The final list consists of 57 items organized in the following 9 categories (see Appendix 6.4): (1) Design, motivation and hypothesis; (2) Literature; (3) Model and method; (4) Data, data gathering and datasets; (5) Metrics; (6) Experiments; (7) Analysis of results and presentation; (8) Repeatability, reproducibility, and replicability; and (9) Conclusions and claims.

Finally, we used this list of concerns to propose an update to the existing SIGIR guidelines. This is described in the next section.

4.5.3 Draft Guidance for Authors

Some years ago, SIGIR introduced brief guidance for authors as "Things that strengthen an IR paper".[84] One of us (Zobel) recently updated this guidance for SIGIR-AP'23, in consultation with the other Program Chairs, but we note that it represented the views of just a couple of individuals. The SIGIR guidance proposed, at a high level, aspects to consider in presentation and experiments. The SIGIR-AP revision primarily addressed some aspects – omissions, oversights, and shortcomings – that are offered as grounds for rejection.

Here, we took the SIGIR-AP draft guidance as a starting point and reviewed it against the list of concerns that we set out in Section 4.5.1. We also took note of generic writing advice that is widely available and decided to omit elements that we regarded as pertinent to computer science research in general. This led to the following, which we propose as a basis for the advice provided by venues that publish IR work.

We have sought to make the advice broad, understandable, and constructive; but it is of necessity brief and some readers may seek more detail. For that reason, when the advice (or a revision of it) is used, it might also be helpful to link to a version of the lists of concerns in Appendix 6.4.

Our proposed draft guidance is as follows.

[84] The earliest mention we are aware of is from SIGIR 2021, https://sigir.org/sigir2021/checklist-to-strengthen-an-ir-paper/.

Motivation and claims

- The problem is well characterised and motivated, and the potential impact is discussed.
- The proposed application of the work is contextualised by pertinent knowledge from that domain, including potential ethical, social, or environmental impacts.
- The research goals and original contributions (that is, the elements that are a contrast to the prior art) are stated and are clearly distinguished from prior work.
- The claims are properly scoped and supported.
- There are explicit statements of what was done and what was not.

Presentation

- The literature review considers competitive previous solutions for the problem, that is, it is not limited to consideration of other work on the same technology as that explored in the submission.
- There is a reasoned justification for each of the choices made in each step of the research and each element of the method.
- Results are presented in keeping with the norms in the field as exemplified in strong prior work.
- A substantive, focused, and insightful discussion accompanies the results taking into account limitations and scope of the work.

Experiments

- The experimental design and its scale are appropriate to the problem.
- In comparative studies, appropriate baselines are used; they are deployed and optimized in ways comparable to those used for the proposed method.
- The experimental results are reliable and generalizable, and preferably show illustrative individual cases as well as aggregated results.
- Where appropriate, a diversity of data sets are used, including public-domain data sets used in prior work.
- Sufficient details (with data and code where appropriate) are provided to enable other researchers to assess and reproduce the experiments; this includes the nature, source, and collection process for the data, and the data preparation steps.

Results and analysis

- The evaluation methods and measures address the research questions; the use of redundant or highly correlated measures should be avoided.
- Statistical analysis is used and reported appropriately.
- Development data, training data, and test data are distinguished from each other.
- User studies are based on adequately sized, representative cohorts; data is gathered in ways that meet ethical norms, or where appropriate in keeping with prescribed ethics practices.
- Final results were obtained after all development was complete, that is, not selected because they are the best outcomes amongst a larger set of experiments or hand-fitted to the data.

Common problems that lead to rejection

Issues with papers in relation to the recommendations above can lead to rejection. Other problems that can lead to rejection are as follows.

- Literature reviews that lack critical analysis of prior work or that largely consist of lists of papers, that is, do not have an insightful discussion.
- Contributions that consist of small modifications to established techniques, particularly where the contribution is a straightforward variation of the established technique or where there are numerous prior papers exploring similar variations.
- Methods that appear to be developed and hand-tuned on a specific data set without discussion or demonstration of their lessons for future work or of how the methods would be more generally applicable.
- Justification of a method solely by its score in experiments, lacking an a priori rationale for why the method is worth exploring.
- Experiments where the data volumes are too small to support the conclusions.
- Any form of academic fraud, misrepresentation, or dishonesty.

4.5.4 Next Steps

Guidance and lists of issues should be living documents that reflect a current and uncontroversial agreement in the community. Therefore, they should be open to change because there can always be some disagreements and expectations of authors can change over time, in some cases quite quickly, especially as the subjects of research shift to focus on new topics. For that reason, no set of advice should be regarded as fixed, but revision should be undertaken consultatively and with a spectrum of colleagues.

We suggest that the detailed list of issues of concern in Appendix 6.4 be made available in some form as educative for reviewers. We stress here that it is not our intention that reviewers simply reject papers because of these issues. It could also provide a resource at forums such as doctoral consortia.

We thus believe that it would be valuable for the community to:

- Ensure that the guidelines are prominent in the calls-for-papers at our major conferences and journals, or otherwise disseminated.
- Encourage the SIGIR executive committee to take ownership of the guidelines and to occasionally convene a panel to produce an update.
- Use these resources educatively for new members of the community and for new reviewers.

In this exercise, we have not produced guidance for reviewers, which in other disciplines tends to consist of two parts: general advice on how to approach the task and specifics for the field. An example that we found was produced by the ACL, as discussed above; a particular strength of these guidelines in our view is the enumeration of unfair grounds for rejection. We believe that such guidance would be of value to our community, and could make use of the materials we have presented here.

References

1 Dennis Ulmer, Elisa Bassignana, Max Müller-Eberstein, Daniel Varab, Mike Zhang, Rob van der Goot, Christian Hardmeier, and Barbara Plank. Experimental standards for deep learning in natural language processing research, 2022.

5 List of Acronyms

ACL Association for Computational Linguistics
ACM DL ACM Digital Library
ADM+S Automated Decision-Making and Society
AI Artificial Intelligence
ASSIA Asian Summer School in Information Access
CfP Call for Papers
CHI Computer-Human-Interaction
CLEF Conference and Labs of the Evaluation Forum
CyCAT Cyprus Center for Algorithmic Transparency
ECIR European Conference on Information Retrieval
ESS Experiment Support System
ESSIR European Summer School on Information Retrieval
FAccT Fairness, Accountability, and Transparency
FDIA Future Directions in Information Access
FIRE Forum for Information Retrieval Evaluation
GDPR General Data Protection Regulation
IR Information Retrieval
KPI Key Performance Indicator
LLM Large Language Model
ML Machine Learning
MRR Mean Reciprocal Rank
NLP Natural Language Processing
nDCG normalized Discounted Cumulative Gain
NTCIR NII Testbeds and Community for Information access Research
PC Program Committee
PyIRE Python Interactive Information Retrieval Evaluation
QA Question Answering
RS Recommender Systems
RecSys ACM Conference on Recommender Systems
SSH Social Sciences and Humanities
SPC Senior Program Committee
SEME Search Engine Manipulation Effect
SIGIR ACM SIGIR Conference on Research and Development in Information Retrieval
SIGIR-AP Information Retrieval in the Asia Pacific

SIGPLAN ACM Special Interest Group on Programming Languages
SOTA State Of The Art
TREC Text REtrieval Conference
UX User Experience
WiMIR Women in Music Information Retrieval

6 Author Guidance Appendix

This Appendix consists of annotated materials that helped to inform the list of concerns described in Section 4.5, and the list of concerns itself. As explained in the main text, our aim was to gather suggestions of guidance and issues from a range of sources and consolidate them into a resource for IR.

6.1 SIGPLAN Empirical Evaluation Guidelines with Annotations

The following is our annotation of the SIGPLAN Guidelines.[85] In this annotation, we underlined aspects deemed particularly worth reflecting in guidelines for IR. ~~Strikethrough~~ was used for aspects that we felt did not translate to our community well, and greyed text for aspects, we felt to be valuable but needing adaptation for an IR context.

Clearly stated claims

S1: Claims not explicit

- Claims must be explicit in order for the reader to assess whether the empirical evaluation supports them. ~~Missing claims cannot possibly be assessed. Claims should also aim to state not just what is achieved but how.~~

S2: Claims not appropriately scoped

- The truth of a claim should clearly follow from the evidence provided. ~~Claims that are not fully supported mislead readers.~~ “Works for all Java” is over-broad when based on a subset of Java. Other examples are “works on real hardware” when evaluating only with (unrealistic) simulation, and “automatic process” when requiring human intervention.

S3: Fails to acknowledge limitations

- A paper should acknowledge its limitations to place the scope of its results in context. ~~Stating no limitations at all, or only tangential ones, while omitting the more relevant ones may mislead the reader into drawing overly-strong conclusions. This could hold back efforts to publish future improvements and may lead researchers down to wrong paths.~~

S4: Suitable comparison

[85] https://www.sigplan.org/Resources/EmpiricalEvaluation/

- Fails to compare against the appropriate baseline. ~~Empirical evidence for a claim that a technique/system improves upon the state-of-the-art should include a comparison against an appropriate baseline. The lack of a baseline means empirical evidence lacks context. A 'straw man' baseline that is misrepresented as state-of-the-art is also problematic, as it would inflate apparent benefit.~~

S5: Comparison is unfair

- Comparisons to a competing system should not unfairly disadvantage that system. Doing so would inflate the apparent advantage of the proposed system. For example, it would be unfair to compile the state-of-the-art baseline at -O0 optimization level, while using -O3 for the proposed system.

Principled Benchmark Choice

S6: Inappropriate suite

- Evaluations should be conducted using appropriate established benchmarks where they exist ~~so that claimed results are more likely to generalize. Not doing so may yield results that are not sufficiently general.~~ Established suites should be used in context; e.g., it would be wrong to use a single-threaded suite for studying parallel performance.

S7: Unjustified use of non-standard suite(s)

- The use of standard benchmark suites improves the comparability of results. However, sometimes a non-standard suite, such as one that is subsetted or homegrown, is the better choice. In that case, a rationale, and possible limitations, must be provided to demonstrate why using a standard suite would have been worse.

~~S8: Kernels instead of full applications~~

- ~~Kernels can be useful and appropriate in a broader evaluation. However, a claim that a system benefits applications should be tested on such applications directly, and not only on micro-kernels, which may lack important characteristics of full applications.~~

Adequate Data Analysis

S9: Insufficient number of trials

- Modern systems with non-deterministic performance properties may require many trials (e.g., of a single time measurement) to characterize their behavior adequately. Failure to do so risks treating noise as a signal. Similarly, more trials may be needed to get the system into an intended state (e.g., into a steady state that avoids warm-up effects).

S10: Inappropriate summary statistics

- Summary statistics such as mean and median can usefully characterize many results. But they should be selected carefully because each statistic presents an accurate view only under appropriate circumstances. An inappropriate summary may amplify noise or hide an important trend.

S11: No data distribution reported

- A measure of variability (e.g., variance, std. Deviation, quantiles) and/or confidence intervals, is needed to understand the distribution of the data. Reporting just a measure of central tendency (e.g., a mean or median) can mislead the reader, especially when the distribution is bimodal or has a fsignificant variance.

Relevant metrics

S12: Indirect or inappropriate proxy metric

- Proxy metrics can substitute for direct ones only when the substitution is clearly, explicitly justified. For example, it would be misleading and incorrect to report a reduction in cache misses to claim actual end-to-end performance or energy consumption improvement.

S13: Fails to measure all important effects

- All important effects should be measured to show the true cost of a system. For example, compiler optimizations may speed up programs at the cost of drastically increasing compile times of large systems, so the compile time should be measured as well as the program speedup. Failure to do so distorts the cost/benefit of the system.

Appropriate and Clear Experimental Design

S14: Insufficient information to repeat

- Experiments evaluating an idea need to be described in sufficient detail to be repeatable. All parameters (including default values) should be included, as well as all version numbers of software, and full details of ~~hardware~~ platforms. ~~Insufficient information impedes repeatability and comparison of future ideas and can hinder scientific progress.~~

~~S15: Unreasonable platform~~

- ~~The evaluation should be on a platform that can reasonably be said to match the claims; otherwise, the results of the evaluation will not fully support the claims. For example, a claim that relates to performance on mobile platforms should not have an evaluation performed exclusively on servers.~~

S16: Ignores key design parameters

- Key parameters should be explored over a range to evaluate sensitivity to their settings. Examples include the size of the heap when evaluating garbage collection and the size of caches when evaluating a locality optimization. All expected system configurations ~~(e.g., from warmup to steady state)~~ should be considered.

~~S17: Gated workload generator~~

- ~~Load generators for typical transaction-oriented systems should be 'open loop', to generate work independent of the performance of the system under test. Otherwise, results are likely to mislead because real-world transaction servers are usually open-loop.~~

S18: Tested on training set

- When a system aims to be general but was developed with close consideration of specific examples, it is essential that the evaluation explicitly perform cross-validation, so that the system is evaluated on data distinct from the training set. For example, static analysis should not be exclusively evaluated on programs used to inform its development.

6.2 Experimental Standards for Deep Learning Guidelines with Annotations

The following is our annotation of the highlighted material from the Experimental Standard for Deep Learning[86] [1]. A question mark (!) indicates the "must" category from the original paper and a plus (+) indicates recommendations from the original paper. We use ~~striketrough~~ for items we deemed not specifically relevant for the IR community, and gray text for relevant items with a valuable issue that needs to be adapted to be made pertinent to the IR community.

Data

D01 ! Consider dataset and experiment limitations when drawing conclusions (Schlangen, 2021);

D02 ! Document task adequacy, representativeness and pre-processing (Bender and Friedman, 2018);

D03 ! Split the data such as to avoid spurious correlations;,

D04 + Publish the dataset accessibly & indicate changes;

D05 + Perform exploratory data analyses to ensure task adequacy (Caswell et al., 2021);

D06 + Publish the data set with individual-coder annotations, besides aggregation;

D07 + Claim significance considering the dataset's statistical power (Card et al., 2020).

Codebase & Models

D08 ~~!~~ Publish a code repository with documentation and licensing to distribute for replicability;

D09 ! Report all details about hyperparameter search and model training;

D10 ! Specify the hyperparameters for replicability

D11 + Publish model predictions and evaluation scripts.;

D12 + ~~Use model cards~~;

D13 + Publish models;

Experiments and Analysis

D14 ! Report mean & standard deviation over multiple runs;

D15 ! Perform significance testing ~~or Bayesian analysis~~ and motivate your choice of method;

D16 ! Carefully reflect on the amount of evidence regarding your initial hypotheses.

Publications

D17 ! Avoid citing pre-prints (if applicable);

D18 ~~!~~ Describe the computational requirements;

D19 ~~!~~ Consider the potential ethical & social impact;

D20 + ~~Consider the environmental impact and prioritize computational efficiency;~~

D21 + Include an Ethics and/or Bias Statement.

[86] https://arxiv.org/pdf/2204.06251.pdf

6.3 Quick reflections

Our next resource was an unstructured collection of material we gathered by discussing of our individual experience as reviewers. We also gathered similar kinds of comments from other attendees, which we omit here (they are much less structured) but incorporated into the list of concerns below.

- No analysis of outliers or inspection of spread and diversity of results (aka just report the mean score).
- Lit reviews that are lists of papers without reflection, analysis or connections to the current work (gaps, bridges, etc); addition of max number of citations to each statement.
- Unreflective use of "the rubric" as a way of writing the paper; no insights, no meaningful analysis, no meaningful identification of contribution.
- Justification by score.
- Don't show examples of the method or only show the positive ones.
- Unjustified experimental settings such as hyperparameter choices, or a long sequence of unjustified design choices/decisions., and how they may be perpetuated thanks to citing work.
- Graph overload – thousands of results without explanation, choice of illustrative cases – lost in visualisation. Also, graphs that make no sense.
- Confident, bold statements of goals that are impossible to interpret in concrete terms.
- Model and problem are not related to each other.
- Problem and measures are not related.
- Scale of data absurdly out of keeping with the problem that the paper sets out to solve.
- Claims are overstated by comparison to the data.
- Naïve, outdated baselines – a single strong competitive baseline is better than a family of simplistic baselines.
- No consideration of the possibility or scale or presence of random error.
- Assumption that training data is perfect; use of cross-fold validation (the dataset defines the task) to draw general conclusions.
- Doing of user studies just to get a check-mark for making it real.
- Failure to get ethics clearances when required.
- Use of crowd-sourcing for experiments that require a laboratory setting.
- Use of students enrolled in a subject as experimental subjects when representativeness is required.
- Inadequate description of the data, lack of clarity on source and availability, and likewise for the code.
- Basic issues with clarity and obscurity; obfuscation.
- Badly implemented baseline or implementation is not comparable.
- Failure to consider Goodhart's law.
- Inference from aggregate data.
- Comparison between systems with different scales of hyperparameters (time-constrained tuning vs. grid search)
- Papers that just show summary statistics and don't show any examples.
- Lack of understanding of what is needed for repeatability, reproducibility, and replicability.
- Lack of distinction between development data and test data; selective presentation of results that are favourable.

6.4 Common Flaws in Submitted IR Papers

Our analysis of the materials above, and reading of other resources for authors in cognate fields, provided the basis for the categorisation of areas of concern. These areas of concern were subsequently analysed to inform the Guidance for Authors included in the main text. In this analysis, we identified bullet points that we regarded as essential; these are marked with a star ($*$).

Design, motivation, and hypothesis

- Basic issues with clarity and obscurity; make the design, motivation, and hypotheses difficult to understand and unintentionally obfuscate the main content of the research.
- Confident, bold statements of goals that are impossible to interpret in concrete terms.
- Unreflective use of "the rubric" as a way of writing the paper (i.e., a specific set of details about what is needed to structure a paper) with no insights, no meaningful analysis, and without any meaningful identification of contribution.
- Inclusion of elements just to follow a template, such as unhelpful user studies, use of ablation when it doesn't relate to the conclusions, and graphs showing irrelevant data.
- Lack of a clearly stated problem or research goal.
- Lack of appreciation that method design relies on domain knowledge; lack of inclusion of extra-disciplinary knowledge where relevant. $*$
- No acknowledgement of the social or ethical impact of the work. $*$

Literature

- Literature reviews that are mere lists of papers without reflection, analysis or connections to the current work; unreasonably large numbers of citations to each statement. $*$
- Citing of papers which would clearly fail the above guidelines.
- Obvious gaps in the bibliography due to poor literature search, such as missing foundational or key papers that are relevant to the work, recent citations or older citations that are still current.

Model and method

- Model (or the method or solution) and the problem are not related to each other.
- A long sequence of unjustified design choices and decisions, or justification from prior work that does not apply. $*$
- Lack of examples of how the model is going to work.
- Not clear how the method is distinct from and connected to, prior work. $*$

Data, data gathering, and datasets

- Inadequate description of the data, lack of clarity on creation, source or availability. $*$
- Inappropriate choice of human subjects (e.g., the researchers themselves, or students in cases where they do not represent the target populations).
- Use of crowd-sourcing for experiments that require a laboratory or controlled settings.
- Use of survey instruments that are not a good match to the problem, or that haven't been validated for it.
- Failure to get ethics clearances when required, lack of consideration of ethics, bias, confidentiality or privacy. $*$

- Scale of data clearly out of keeping with the problem.
- Lack of multiple datasets when readily available and appropriate to the problem. ∗
- Use of the wrong dataset, or no exploration of its suitability for the problem.

Metrics

- Problems and chosen measures are not related (for example, a classification problem and the use of inappropriate measures for this kind of problem). ∗
- Selective, post hoc use of metrics to find positive results.
- Reporting of multiple, correlated metrics as if they represented independent sources of evidence.
- Invented metrics, especially when they are not explained or difficult to interpret.

Experiments

- Lack of distinction between data partitions, such as training, validation, and test set. ∗
- Results that come from overfitting to the wrong data partition, especially hand-tuned models for that data, or results that are hand-picked from a large volume of trials.
- No exploration of the sensitivity of the method to the values of key (hyper-)parameters.
- Unjustified decisions in the experimental setting, such as hyperparameter settings.
- Use of default parameters for baselines while tuning the same parameters for the proposed model. ∗
- Lack of consideration about testing systems with very different numbers of hyperparameters (e.g., time-constrained tuning vs. grid search).
- Poor or naive choice of baselines (e.g., a single strong competitive baseline is better than a family of simplistic baselines).
- Badly implemented baselines, or implementation is not comparable.

Analysis of results and presentation

- Reporting only summary statistics without specific examples, positive examples and negative examples. ∗
- No consideration for variability and diversity of results and outliers (i.e., reporting only mean scores). ∗
- Only quantitative results, without studying whether modeling assumptions are reasonably held up and a qualitative discussion of error sources.
- Selective presentation of results that are favorable. ∗
- Selective, post hoc use of statistical analysis to find positive results; reporting of results as “nearly significant”.
- No consideration of the presence and scale of random error.
- Overstatement of the statistical precision of results.
- Data overload: unnecessarily large numbers of graphs and tables, or insufficient explanation as to how to interpret them.
- Poor statistical analysis, such as wrong choice of significance test, lack of consideration of power or effect size, statistical testing when sample size is unsuitable, or missing to mention what hypotheses are being tested and how.
- Superficial analysis or without interpretation.

Repeatability, reproducibility and replicability

- Lack of communication of what is needed for repeatability, reproducibility, and replicability; e.g., missing parameter settings, missing explanation of data preparation and pre-processing. ∗
- Failure to use an appropriate standard dataset.
- Failure to use a standard implementation (e.g., baselines, evaluation software).
- Lack of recognition of the value of publishing data and code.
- Inadequate description of code, lack of clarity on source and availability, documentation, licensing, key metadata, or not versioned.

Conclusions and claims

- Inference of general conclusions from aggregated data without individual analysis.
- Assumption that training data are perfect (e.g., that they are an ideal setting and representative of all possible data). ∗
- Claims of performance on unseen data based on cross-validation results.
- Claims that do not follow from the results. ∗
- Justification of innovation entirely by numerical results. ∗
- Use the current results to reformulate the initial hypotheses.
- No consideration of limitations of the proposed solution or experimentation.
- No noting of excessive or large-scale computational requirements.

Participants

- Christine Bauer
Utrecht University, NL
- Joeran Beel
Universität Siegen, DE
- Timo Breuer
TH Köln, DE
- Charles Clarke
University of Waterloo, CA
- Anita Crescenzi
University of North Carolina – Chapel Hill, US
- Gianluca Demartini
The University of Queensland – Brisbane, AU
- Giorgio Maria Di Nunzio
University of Padova, IT
- Laura Dietz
University of New Hampshire – Durham, US
- Guglielmo Faggioli
University of Padova, IT
- Nicola Ferro
University of Padova, IT
- Bruce Ferwerda
Jönköping University, SE
- Maik Fröbe
Friedrich-Schiller-Universität Jena, DE
- Norbert Fuhr
Universität Duisburg-Essen, DE
- Matthias Hagen
Friedrich-Schiller-Universität Jena, DE
- Allan Hanbury
TU Wien, AT
- Claudia Hauff
Spotify – Amsterdam, NL
- Dietmar Jannach
Alpen-Adria-Universität Klagenfurt, AT
- Noriko Kando
National Institute of Informatics – Tokyo, JP
- Evangelos Kanoulas
University of Amsterdam, NL
- Bart Knijnenburg
Clemson University, US
- Udo Kruschwitz
Universität Regensburg, DE
- Birger Larsen
Aalborg University Copenhagen, DK
- Meijie Li
Universität Duisburg-Essen, DE
- Maria Maistro
University of Copenhagen, DK
- Lien Michiels
University of Antwerp, BE
- Andrea Papenmeier
Universität Duisburg-Essen, DE
- Martin Potthast
Universität Leipzig, DE
- Paolo Rosso
Technical University of Valencia, ES
- Alan Said
University of Gothenburg, SE
- Philipp Schaer
TH Köln, DE
- Christin Seifert
Universität Duisburg-Essen, DE
- Ian Soboroff
NIST – Gaithersburg, US
- Damiano Spina
RMIT University – Melbourne, AU
- Benno Stein
Bauhaus-Universität Weimar, DE
- Nava Tintarev
Maastricht University, NL
- Julián Urbano
TU Delft, NL
- Henning Wachsmuth
Universität Paderborn, DE
- Martijn Willemsen
Eindhoven University of Technology & JADS – 's-Hertogenbosch- Eindhoven
- Justin Zobel
The University of Melbourne, AU

Report from Dagstuhl Seminar 23041

Integrated Rigorous Analysis in Cyber-Physical Systems Engineering

Erika Abraham*[1], Stefan Hallerstede*[2], John Hatcliff*[3], Danielle Stewart*[4], and Noah Abou El Wafa†[5]

1 **RWTH Aachen University, DE.** abraham@informatik.rwth-aachen.de
2 **Aarhus University, DK.** stefan.hallerstede@wanadoo.fr
3 **Kansas State University – Manhattan, US.** hatcliff@ksu.edu
4 **Galois – Minneapolis, US.** danielle@galois.com
5 **KIT – Karlsruher Institut für Technologie, DE.** noah.abouelwafa@kit.edu

Abstract

This report documents the program and the outcomes of the Dagstuhl Seminar 23041 “Integrated Rigorous Analysis in Cyber-Physical Systems (CPS) Engineering”.

This seminar brought together academic and industry representations from a variety of domains with backgrounds in different techniques to develop a roadmap for addressing the current challenges in the area of CPS engineering. An overarching theme was the potential use of integrated models and associated methodologies that support cross-technique information/results sharing and smooth workflow hand-offs between individual tools and methods.

Seminar January 22–27, 2023 – https://www.dagstuhl.de/23041

2012 ACM Subject Classification Computer systems organization → Embedded and cyber-physical systems; Security and privacy → Logic and verification

Keywords and phrases cyber-physical systems, formal methods, rigorous modelling and analysis, systems engineering

Digital Object Identifier 10.4230/DagRep.13.1.155

1 Executive Summary

Erika Abraham (RWTH Aachen University, DE)
Stefan Hallerstede (Aarhus University, DK)
John Hatcliff (Kansas State University – Manhattan, US)
Danielle Stewart (Galois – Minneapolis, US)

Overview

The design of cyber-physical systems (CPSs) typically balances requirements that concern function, performance and interaction between discrete and continuous subsystems. In the big picture CPS design must be considered in the context of systems engineering. When engineering a CPS, modelling plays a central role during early stages of the development. Depending on objectives and purpose different models are produced, say, for a concept of operation, a trade study, a preliminary design, and a detailed design. In recent years modelling methods and tools have been developed that can contribute to the development of CPSs.

* Editor / Organizer
† Editorial Assistant / Collector

Integrated Rigorous Analysis in Cyber-Physical Systems Engineering, *Dagstuhl Reports*, Vol. 13, Issue 1, pp. 155–183
Editors: Erika Abraham, Stefan Hallerstede, John Hatcliff, and Danielle Stewart
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Each method has a limited view of CPS development, say, focusing on correctness verification, scenario validation or evaluation of design alternatives. Each method is specialised on specific kinds of analyses depending on its purpose and objectives. Of course, this is necessary for reasons of effectiveness and efficiency. Unfortunately, then the outcomes of different analyses carried out on the various models of a CPS are not systematically exploited in the other models. The arguments connecting the different outcomes of independent methods and tools can be intricate and complex, potentially causing erroneous reasoning but missed opportunities when relevant outcomes remain unused.

This Dagstuhl Seminar explored systems engineering processes and methodology as a framework for rigorous reasoning to alleviate the problem of bridging different modelling methods, opening up a possibility to reason across method and stage barriers. The seminar brought together academic and industry representations from a variety of domains with backgrounds in different techniques. We developed a roadmap for addressing CPS challenges both in industry and academia, and identified ways that we can help each other overcome these challenges.

Outcomes of the Seminar

- Identified new techniques, tool capabilities and methodology improvements that will improve the ability to develop, assure, deploy, and evolve modern CPS.
- Identified gaps and needs that enumerates desired tool capabilities and methodology improvements that if successfully addressed, would improve the ability to develop, assure, deploy, and evolve modern CPS.
- Identified criteria and resources for community-based example systems that enable the interplay of multiple techniques to be evaluated across the life-cycle of system development.
- Created an activity plan for future meetings and smaller collaborative groups to build on the outcomes of the seminar.

The organizers thank all participants for their interesting ideas and viewpoints presented in talks, discussions, and informal meetings. Moreover, we would like to express our gratitude towards Schloss Dagstuhl and its staff for all the support before and during the seminar, which contributed to making this seminar a successful one.

2 Table of Contents

Executive Summary

Overview of Talks

Working groups

3 Overview of Talks

3.1 How to Prove That We Do Not Prove a Faulty Controller Safe

Wolfgang Ahrendt (Chalmers University of Technology – Göteborg, SE)

Joint work of Wolfgang Ahrendt, Yuvaraj Selvaraj, Jonas Krook, Martin Fabian

Cyber-physical systems are often safety-critical and their correctness is crucial, as in the case of automated driving. Using formal mathematical methods is one way to guarantee correctness. Though these methods have shown their usefulness, care must be taken as modeling errors might result in proving a faulty controller safe, which is potentially catastrophic in practice. This talk deals with two such modeling errors in differential dynamic logic. Differential dynamic logic is a formal specification and verification language for hybrid systems, which are mathematical models of cyber-physical systems. The main contribution is to express conditions that, when fulfilled, show the absence of certain modeling errors that would cause a faulty controller to be proven safe. The problems are illustrated with an example of a safety controller for automated driving, and it is shown that the formulated conditions have the intended effect both for a faulty and a correct controller. It is also shown how the formulated conditions aid in finding a loop invariant candidate to prove properties of hybrid systems with feedback loops. The results are proven using the interactive theorem prover KeYmaera-X.

3.2 Surrogate Verification – Neural Networks and Koopman Operator Approximations

Stanley Bak (Stony Brook University, US)

Main reference Stanley Bak, Hoang-Dung Tran: "Neural Network Compression of ACAS Xu Early Prototype Is Unsafe: Closed-Loop Verification Through Quantized State Backreachability", in Proc. of the NASA Formal Methods – 14th International Symposium, NFM 2022, Pasadena, CA, USA, May 24-27, 2022, Proceedings, Lecture Notes in Computer Science, Vol. 13260, pp. 280–298, Springer, 2022.
URL https://doi.org//10.1007/978-3-031-06773-0_15

Many systems are black-box in nature or too complex to directly verify. To work with such systems, surrogate model approaches can be used to create models that approximate system behaviors. We discussed two approaches for this problem, one for neural network approximations and one for approximations of nonlinear dynamical systems based on Koopman Operator approximations.

In Koopman Operator approximations a nonlinear system is approximated using a higher-dimension linear system. While reachability and verification of linear systems is usually much easier, the problem involves complex nonlinear initial sets of states. We overcome this using polynomial zonotopes, data structures originally designed for nonlinear reachability analysis. Further, to accommodate for the error in the model approximation, we explore conformant synthesis approaches. We are working toward developing scalable formal analysis methods that can still be applied towards complex and black-box systems.

3.3 Validation and verification approaches for safe and secure cyber-physical systems

Stylianos Basagiannis (Raytheon Technologies – Collins Aerospace – Cork, IE)

Ensuring the security and safety of cyber-physical systems (CPS) while reducing systems' environmental impact, fuel consumption, and operational cost is forcing a rethinking of future cyber-physical systems design cycles. In a continuously growing global market, next-generation CPS development requires methods and tools to promote early cross-discipline collaboration, allowing a system-wide accurate analysis, validation, and verification. Collaborative model-based design is a promising approach, in which diverse digital model representations of system elements are combined and analyzed in a virtual setting, but its full benefits have not been fully realized in the sector. At the same time, the multi-diverse engineering background of CPS teams forces requirements to be easily corrupted or misinformed from the (abstract) design till the (granular) prototype generation phase.

In this presentation, we will introduce some of our recent validation and verification approaches being applied in aerospace cyber-physical systems. The first (top-down) approach will involve the usage of simulation-based verification techniques through interval analysis approaches [1] for the safety verification of advanced engine control solutions. The second (bottom-up) approach will involve the usage of theorem-proving techniques at the instruction set level for embedded (RISC-V) micro-architectures for verifying security requirements. Activities described in this presentation are part of two European-funded projects in which Collins Aerospace Ireland is participating; namely the ECSEL VALU3S (2020-2023) [2] and Horizon Europe REWIRE (2022-2025) [3].

References

1 Vassilios A. Tsachouridis and Georgios Giantamidis and Stylianos Basagiannis and Kostas Kouramas, Formal analysis of the Schulz matrix inversion algorithm: A paradigm towards computer aided verification of general matrix flow solvers, In Journal of Numerical Algebra, Control and Optimization, v10 (2), pp. 177-206, 2020.

2 ECSEL VALU3S: Verification and Validation of Automated Systems' Safety and Security, 2020-2023, https://valu3s.eu/

3 Horizon Europe REWIRE: REWiring the ComposItional Security VeRification and AssurancE of Systems of Systems Lifecycle, 2022-2025, https://www.rewire-he.eu/

3.4 Developing a prototype of a mechanical ventilator controller from requirements to code with ASMETA

Andrea Bombarda (University of Bergamo – Dalmine, IT)

Joint work of Andrea Bombarda, Silvia Bonfanti, Angelo Gargantini, Elvinia Riccobene
Main reference Andrea Bombarda, Silvia Bonfanti, Angelo Gargantini, Elvinia Riccobene: "Developing a Prototype of a Mechanical Ventilator Controller from Requirements to Code with ASMETA", Electronic Proceedings in Theoretical Computer Science, Vol. 349, pp. 13–29, Open Publishing Association, 2021.
URL https://doi.org//10.4204/eptcs.349.2

Rigorous development processes aim to be effective in developing critical systems, especially if failures can have catastrophic consequences for humans and the environment. Such processes generally rely on formal methods, which can guarantee, thanks to their mathematical foundation, model preciseness, and properties assurance. However, they are rarely adopted in practice.

In this talk, I report the experience of my research group in using the Abstract State Machine formal method and the ASMETA framework in developing a prototype of the control software of the MVM (Mechanical Ventilator Milano), a mechanical lung ventilator that has been designed, successfully certified, and deployed during the COVID-19 pandemic.

Although due to time constraints and lack of skills, no formal method was applied for the MVM project, later we wanted to assess the feasibility of developing (part of) the ventilator by using a formal method-based approach. Our development process starts from a high-level formal specification of the system to describe the MVM main operation modes. Then, through a sequence of refined models, all the other requirements are captured, up to a level in which a C++ implementation of a prototype of the MVM controller is automatically generated from the model, and tested.

Along the process, at each refinement level, different model validation and verification activities are performed, and each refined model is proved to be a correct refinement of the previous level. By means of the MVM case study, we evaluate the effectiveness and usability of our formal approach.

3.5 Monitoring distributed cyber-physical systems: opportunities and challenges

Borzoo Bonakdarpour (Michigan State University – East Lansing, US)

CPS is becoming increasingly distributed, where a set of asynchronous agents deal with continuous signals that do not share a global clock. We advocate for runtime verification (RV) of distributed CPS as a complementary method, but a roadmap for enhancing its effectiveness and efficiency is much needed. This brief talk will go over the challenges, recent advances, open problem problems and opportunities in RV for distributed CPS. We will first explain the challenges of verification of a set of continuous signals subject to clock drifts against specifications expressed in the signal temporal logic (STL). We then explain how a practical assumption, namely, an off- the-shelf clock synchronization algorithm such as NTP, can drastically contribute to efficiency and effectiveness of RV. Finally, we show how exploiting special characteristics of CPS such as the knowledge of dynamics of physical processes can reduce the runtime overhead and discuss a roadmap of open problems, applications, and opportunities.

3.6 Optimizing different flavours of nondeterminism in hybrid automata with random clocks

Joanna Delicaris (Universität Münster, DE) and Anne Remke (Universität Münster, DE)

Joint work of Joanna Delicaris, Anne Remke, Carina da Silva, Stefan Schupp
Main reference Carina Pilch, Stefan Schupp, Anne Remke: "Optimizing Reachability Probabilities for a Restricted Class of Stochastic Hybrid Automata via Flowpipe-Construction", in Proc. of the Quantitative Evaluation of Systems – 18th International Conference, QEST 2021, Paris, France, August 23-27, 2021, Proceedings, Lecture Notes in Computer Science, Vol. 12846, pp. 435–456, Springer, 2021.
URL https://doi.org//10.1007/978-3-030-85172-9_23

Stochastic hybrid automata (SHA) are a powerful tool to evaluate the dependability and safety of critical infrastructures. However, the resolution of nondeterminism, which is present in many purely hybrid models, is often only implicitly considered in SHA. This paper instead proposes algorithms for computing maximum and minimum reachability probabilities for singular automata with *urgent* transitions and random clocks which follow arbitrary continuous probability distributions. We borrow a well-known approach from hybrid systems reachability analysis, namely flowpipe construction, which is then extended to optimize nondeterminism in the presence of random variables. Firstly, valuations of random clocks which ensure reachability of specific goal states are extracted from the computed flowpipes and secondly, reachability probabilities are computed by integrating over these valuations. We compute maximum and minimum probabilities for history-dependent prophetic and non-prophetic schedulers using set-based methods. The implementation featuring the library Hypo and the complexity of the approach are discussed in detail. Two case studies featuring nondeterministic choices show the feasibility of the approach.

3.7 Application of Reachability Analysis to MAPE-K Loops

Cláudio Gomes (Aarhus University, DK)

Joint work of Cláudio Gomes, Hao Feng, Casper Thule, Kenneth Lausdahl, Peter Gorm Larsen, Thomas Wright, Jim Woodcock
Main reference Thomas Wright, Cláudio Gomes, Jim Woodcock: "Formally Verified Self-adaptation of an Incubator Digital Twin", in Proc. of the Leveraging Applications of Formal Methods, Verification and Validation. Practice – 11th International Symposium, ISoLA 2022, Rhodes, Greece, October 22-30, 2022, Proceedings, Part IV, Lecture Notes in Computer Science, Vol. 13704, pp. 89–109, Springer, 2022.
URL https://doi.org//10.1007/978-3-031-19762-8_7

The performance and reliability of Cyber-Physical Systems are increasingly aided through the use of digital twins, which mirror the static and dynamic behaviour of a Cyber-Physical System (CPS) in software. Digital twins enable the development of self-adaptive CPSs which reconfigure their behaviour in response to novel environments. It is crucial that these self-adaptations are formally verified at runtime, to avoid expensive re-certification of the reconfigured CPS.

In this talk, I discuss formally verified self-adaptation in a digital twinning system, by constructing a non-deterministic model which captures the uncertainties in the system behaviour after a self-adaptation. We use Signal Temporal Logic to specify the safety requirements the system must satisfy after reconfiguration and employ formal methods based on verified monitoring over Flow* flowpipes to check these properties at runtime. This gives us a framework to predictively detect and mitigate unsafe self-adaptations before they can lead to unsafe states in the physical system.

3.8 Systems engineering with formal methods: darpa case successes, challenges, and gaps

David Hardin (Collins Aerospace – Cedar Rapids, US)

This talk provides a summary of experiences in the development of a Systems Engineering Environment using Formal Methods-based tools on the DARPA CASE program, highlighting notable successes, research and development challenges, as well as technology gaps.

3.9 Heterogeneous Approaches to Safety of Automated Driving Systems: Search-based Testing and Refinement-based Verification

Fuyuki Ishikawa (National Institute of Informatics – Tokyo, JP)

In this talk, I will introduce our research for safety of automated driving systems (ADS).

We had our intensive work on automated testing and debugging for the path planning function in ADS via simulation. Multiple requirements need to be satisfied such as safety, comfort, and compliance with traffic rules. Violations may occur in very specific traffic scenarios or simulator configurations such as positions of other cars. Our technical approach is to make use of automated testing and debugging techniques, originally for program code, by adapting them to the continuous and uncertain ADS problems. We employed search-based testing techniques to explore simulation configurations that lead to violations, e.g., [1]. We also applied fault localization techniques to analyze possible causes of detected violations [2]. Our techniques were evaluated with a simulator provided by our partner Mazda.

To complement these heuristics or search-based approaches, we are also working on a formal approach called Responsibility-Sensitive Safety (RSS). Intuitively, RSS is an approach to define rules such that no crash occurs if all the traffic participants obey them. We formulated RSS with Hoare-like pre-post decomposition and made a case study of refinement-based safety verification with the Event-B formalism for ADS that switches between a black-box AI-based controller and a conservative safe controller [4].

These studies have considered the control aspect of ADS while the emerging difficulties lie in the perception aspect, especially using deep neural networks (DNN). After interviews with industrial partners, our “Engineerable AI” project tackles the problem of safety-aware DNN update. We may want to “fix” our DNN component to mitigate the risk by specific errors, e.g., misclassifying pedestrian to something else. However, re-training with additional dataset can shuffle the millions of DNN parameters. We may not have intended improvement or even have unintended degradation for other error types. We defined benchmarks with our industry partners that evaluate many (10+) of fine-grained safety metrics for the prediction performance. We are tackling them by unique techniques to apply fault localization techniques to identify "suspicious neurons" in DNN for safety-aware, regression controlled update, e.g., [3].

We believe integrating these heterogeneous approaches is essential to deal with complexity and uncertainty of safety-critical CPS such as ADS.

References

1 Yixing Luo, Xiao-Yi Zhang, Paolo Arcaini, Zhi Jin, Haiyan Zhao, Linjuan Zhang, Fuyuki Ishikawa, Targeting Requirements Violations of Autonomous Driving Systems by Dynamic Evolutionary Search, The 36th IEEE/ACM International Conference on Automated Software Engineering (ASE 2021), pp.295-305, November 2021

2 Xiao-yi Zhang, Paolo Archani, Fuyuki Ishikawa, An Incremental Approach for Understanding Collision Avoidance of an Industrial Path Planner, IEEE Transactions on Dependable and Secure Computing, March 2023

3 Davide Li Calsi, Matias Duran, Xiao-Yi Zhang, Paolo Arcaini, Fuyuki Ishikawa, Distributed Repair of Deep Neural Networks,The 16th IEEE International Conference on Software Testing, Verification and Validation (ICST 2023), April 2023

4 Tsutomu Kobayashi, Martin Bondu, Fuyuki Ishikawa, Formal Modelling of Safety Architecture for Responsibility-Aware Autonomous Vehicle via Event-B Refinement, The 25th International Symposium on Formal Methods (FM 2023), March 2023

3.10 Data-Driven Verification for Dynamical Systems Under Uncertainty

Nils Jansen (Radboud University Nijmegen, NL)

Joint work of Thom S. Badings, Licio Romao, Alessandro Abate, Nils Jansen
Main reference Thom S. Badings, Licio Romao, Alessandro Abate, Nils Jansen: "Probabilities Are Not Enough: Formal Controller Synthesis for Stochastic Dynamical Models with Epistemic Uncertainty", CoRR, Vol. abs/2210.05989, 2022.
URL https://doi.org//10.48550/arXiv.2210.05989

Capturing both aleatoric and epistemic uncertainty in models of robotic systems is crucial to designing safe controllers. Most existing approaches for synthesizing certifiably safe controllers exclusively consider aleatoric but not epistemic uncertainty, thus requiring that model parameters and disturbances are known precisely. We present a novel abstraction-based controller synthesis method for continuous-state models with stochastic noise, uncertain parameters, and external disturbances. By sampling techniques and robust analysis, we capture both aleatoric and epistemic uncertainty, with a user-specified confidence level, in the transition probability intervals of a so-called interval Markov decision process (iMDP). We then synthesize an optimal policy on this abstract iMDP, which translates (with the specified confidence level) to a feedback controller for the continuous model, with the same performance guarantees. Our experimental benchmarks confirm that accounting for epistemic uncertainty leads to controllers that are more robust against variations in parameter values.

References

1 Badings, T., Abate, A., Nils Jansen, Parker, D., Poonawala, H. & Stoelinga, M. Sampling-Based Robust Control of Autonomous Systems with Non-Gaussian Noise. *AAAI.* pp. 9669-9678 (2022)

2 Badings, T., Romao, L., Abate, A., Parker, D., Poonawala, H., Stoelinga, M. & Jansen, N. Robust Control for Dynamical Systems with Non-Gaussian Noise via Formal Abstractions. *Journal Of Artificial Intelligence Resesarch.* **76** pp. 341-391 (2023)

3 Badings, T., Romano, L., Abate, A. & Jansen, N. Probabilities Are Not Enough: Formal Controller Synthesis for Stochastic Dynamical Models with Epistemic Uncertainty . *AAAI.* (2023)

3.11 Dynamic Model Composition in Digital Twins

Einar Broch Johnsen (University of Oslo, NO)

Digital twins are currently revolutionizing industry [9] and are entering into the world of medicine (e.g., [7]) and natural science (e.g., [1]). A digital twin is an information system that analyzes the behavior of a physical or cyber-physical system by connecting streams of observations of this twinned system to dynamic (e.g., simulation) and static (e.g., asset management) models. In complex settings, the digital twin will often need to manage several models that reflect different subsystems or different aspects of the twinned system. To analyze such complex systems, digital twins must ensure the correct composition of these models. However, the composition problem for models in digital twins remains unresolved [8]; e.g., models may be at different levels of abstraction, at different granularities or scales, and use different modeling concepts.

In this talk, we discuss this problem for digital twins, with a focus on the composition of heterogeneous dynamic models. For the integration and transfer of information between subsystems, digital twins may profit from a formalization of domain knowledge using ontologies, which has proven effective to unify data models. We have started to explore this approach to correctness and compositionality in digital twins by combining formalized asset models [2] with dynamic behavioral models [4, 6]. This has been done in the context of SMOL [5], a small object-oriented orchestration language[1] which can (a) dynamically create models and integrate them into a program and (b) lift the runtime configuration of a program into a static asset model which can be queried from inside the programs using semantic technologies [3].

Climate barometer for the Oslo Fjord. In a recently started project in collaboration with natural sciences, we tap into on-going efforts to equip the Oslo Fjord (see Fig. 1) with sensors. Our purpose is to combine these sensor streams with digital twin technology to analyze the effects of climate change on ecosystems in the fjord in "real time". During intense precipitation periods (extreme weather), the circulation in the fjord system will change, but it is not known how extreme weather changes the circulation in the fjord. We study this problem by combining two kinds of models. First, a low-resolution circulation model of the fjord. Second, a high-resolution hydro-dynamical models of turbulence in riverine floods

[1] https://smolang.org/

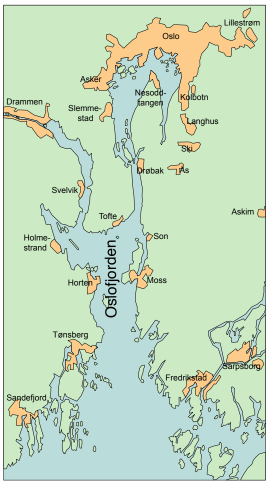


Figure 1 The Oslo Fjord System.

Figure 2 Extreme weather floods.

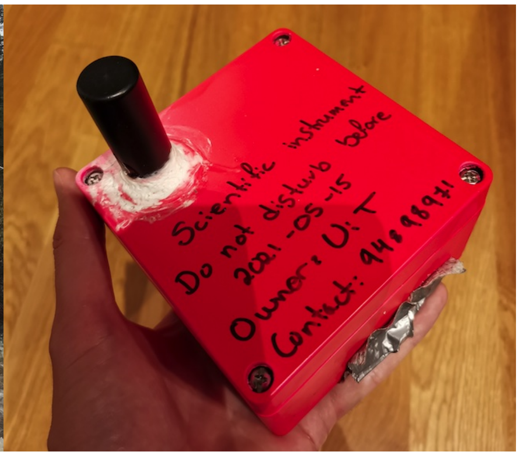


Figure 3 Custom drifter sensor.

(see Fig. 2). The composition of these models will be decided by sensor data from mobile sensors (using an "openSensor" solution, see Fig. 3), tracking the water from the river into the fjord. This composition poses several challenges: (a) the difference in scale between the models needs to be addressed and (b) the exact positioning of the models with respect to each other needs to be decided by the sensor data. Our aim is to formalize this notion of correct composition in a "fjord asset model" of the digital twin, such that the twin can use it together with the sensor data to dynamically compose and adjust the models.

Acknowledgment. This work is a collaboration with Atle Jensen (Dept. of Mathematics, Univ. of Oslo), Kai H. Christensen (Norwegian Meteorological Institute), and Eduard Kamburjan, S. Lizeth Tapia Tarifa, Rudolf Schlatte, Vidar Klungre, David Cameron, Martin Giese and Ingrid Chieh Yu (SIRIUS, Dept. of Informatics, University of Oslo).

References

1 P. Bauer, P. D. Dueben, T. Hoefler, T. Quintino, T. C. Schulthess, and N. P. Wedi. The digital revolution of earth-system science. *Nature Computational Science volume*, 1:104–113, 2021.

2 J. Heaton and A. K. Parlikad. *Asset information model to support the adoption of a digital twin: West Cambridge case study.* IFAC-PapersOnLine 53(3): 366–371, 2020.

3 P. Hitzler, M. Krötzsch, and S. Rudolph. *Foundations of Semantic Web Technologies.* Chapman and Hall/CRC Press, 2010.

4 E. Kamburjan and E. B. Johnsen. Knowledge structures over simulation units. In *Proc. Annual Modeling and Simulation Conference (ANNSIM 2022)*, pages 78–89. IEEE, 2022.

5 E. Kamburjan, V. N. Klungre, R. Schlatte, E. B. Johnsen, and M. Giese. Programming and debugging with semantically lifted states. In *Proc. 18th Intl. Conf. on the Semantic Web (ESWC 2021)*, *LNCS* 12731, pages 126–142. Springer, 2021.

6 E. Kamburjan, V. N. Klungre, R. Schlatte, S. L. Tapia Tarifa, D. Cameron, and E. B. Johnsen. Digital twin reconfiguration using asset models. In *Proc. 11th Intl. Symp. on Leveraging Applications of Formal Methods, Verification and Validation. Practice (ISoLA 2022)*, *LNCS* 13704, pages 71–88. Springer, 2022.

7 J. Masison, J. Beezley, Y. Mei, H. Ribeiro, A. C. Knapp, L. Sordo Vieira, B. Adhikari, Y. Scindia, M. Grauer, B. Helba, W. Schroeder, B. Mehrad, and R. Laubenbacher. A modular computational framework for medical digital twins. *Proceedings of the National Academy of Sciences*, 118(20), 2021.

8 J. Michael, J. Pfeiffer, B. Rumpe, and A. Wortmann. Integration challenges for digital twin systems-of-systems. In *Proc. 10th Intl. Workshop on Software Engineering for Systems-of-Systems and Software Ecosystems (SESoS 2022)*, pages 9–12. ACM/IEEE, 2022.

9 F. Tao and Q. Qi. Make more digital twins. *Nature*, 573: 490–491, 2019.

3.12 Assurance-based Learning-enabled Cyber-Physical Systems: A project summary

Gabor Karsai (Vanderbilt University – Nashville, US)

Joint work of Gabor Karsai, Ted Bapty, Abhishek Dubey, Taylor Johnson, Xenofon Koutsoukos, Janos Sztipanovits and many others
URL https://assured-autonomy.isis.vanderbilt.edu/

Cyber-Physical Systems (CPS) are increasingly incorporating what one can call Learning-Enabled Components (LEC) to implement complex functions. By LEC we mean a component (typically, but not exclusively, implemented in software) that is realized with the help of data-driven techniques, like machine learning. For example, an LEC in an autonomous car can implement a lane follower function such that one trains an appropriate convolutional neural network with a stream of images of the road as input and the observed actions of a human driver as output. The claim is that such LEC built via supervised learning is easier to implement than building a very complex, image processing driven control system that steers the car such that it stays on the road. In other words, if the straightforward design and engineering is too difficult, a neural network can do the job – after sufficient amount of training. For high-consequence systems the challenge is to prove that the resulting system is safe: it does no harm, and it is live: it accomplishes its goals. Safety is perhaps the foremost problem in autonomous vehicles, especially for ones that operate in a less-regulated environment, like the road network. The traditional technology for proving the safety of systems is based on extensively documented but often informal arguments – that are very hard to apply to CPS with LEC. The talk will focus on a recent project that aims at changing this paradigm by introducing (1) verification techniques whenever possible (including proving properties of the "learned" component), (2) monitoring technology for assurance to indicate when the LEC is not peforming well, and (3) formalizing the safety case argumentation process so that it can be dynamically evaluated. The application target is autonomous vehicles, with significant, but not exclusively used LECs. The goal is to construct an engineering process and a supporting toolchain that can be used for the systematic assurance of CPS with LECs.

3.13 Revisiting the challenges in combining requirements engineering and formal methods for CPS

Régine Laleau (IUT Sénart-Fontainebleau, FR)

Joint work of Marc Frappier, Amel Mammar, Tueno Fotso, Steve Jeffrey
Main reference Steve Jeffrey Tueno Fotso, Régine Laleau, Marc Frappier, Amel Mammar, Francois Thibodeau, Mama Nsangou Mouchili: "Assessment of a Formal Requirements Modeling Approach on a Transportation System", in Proc. of the Formal Methods and Software Engineering – 21st International Conference on Formal Engineering Methods, ICFEM 2019, Shenzhen, China, November 5-9, 2019, Proceedings, Lecture Notes in Computer Science, Vol. 11852, pp. 470–486, Springer, 2019.
URL https://doi.org//10.1007/978-3-030-32409-4_29
Main reference Steve Jeffrey Tueno Fotso, Amel Mammar, Régine Laleau, Marc Frappier: "Event-B Expression and Verification of Translation Rules Between SysML/KAOS Domain Models and B System Specifications", in Proc. of the Abstract State Machines, Alloy, B, TLA, VDM, and Z – 6th International Conference, ABZ 2018, Southampton, UK, June 5-8, 2018, Proceedings, Lecture Notes in Computer Science, Vol. 10817, pp. 55–70, Springer, 2018.
URL https://doi.org//10.1007/978-3-319-91271-4_5

A well-known rule says that the sooner a problem is identified in the development process, the better it is for the success of a project, its costs, time delivery and residual default rate. That is why requirements engineering (RE) is getting higher responsibility in the development of systems. RE, always, has to manage some tradeoffs between methods, languages, models and tools to capture well the initial goals defined in a natural language and the need to produce a clear, complete, unambiguous model of the specification for design and implementation phases. On the other hand, when developing critical systems, formal methods are used to strengthen the development process and to increase the level of confidence of the final product. In the last decade, several research works have been developed to combine requirements engineering and formal methods, mainly for software or embedded systems. Cyber-Physical Systems (CPS) combine interconnected computational and physical elements, possibly including human interactions. They are most often critical systems, especially in industrial domains like automotive, aeronautics, space, energy, medical, etc. Clearly, RE for CPS is more complex than RE for traditional embedded or software systems. Indeed, CPS design necessarily involves different engineering disciplines, such as mechanical, electrical, software engineering, relying on different sets of modeling languages. Similarly, different kinds of formal methods (e.g. logic for computational components, differential calculus for physical components) are essential to verify critical requirements such as consistency, safety, security, reliability, performance, while taking into account requirements involved by human interactions. In this talk I will introduce some of the challenges surrounding the modeling and verification of requirements for CPS through an illustrative example of a road transportation system.

3.14 Increasing Dependability of Cyber-Physical Systems by using Digital Twins

Peter Gorm Larsen (Aarhus University, DK)

Joint work of John S. Fitzgerald, Peter Gorm Larsen, Ken G. Pierce

Main reference John S. Fitzgerald, Peter Gorm Larsen, Ken G. Pierce: "Multi-modelling and Co-simulation in the Engineering of Cyber-Physical Systems: Towards the Digital Twin", in Proc. of the From Software Engineering to Formal Methods and Tools, and Back – Essays Dedicated to Stefania Gnesi on the Occasion of Her 65th Birthday, Lecture Notes in Computer Science, Vol. 11865, pp. 40–55, Springer, 2019.

URL https://doi.org//10.1007/978-3-030-30985-5_4

This presentation demonstrated the personal journey moving from formal modelling to using such models to realise digital twins for Cyber-Physical Systems (CPSs). There are many considerations that needs to be considered in order to make such a journey successful and many of these involve interdisciplinary engineering and research. Coupling models together with different mathematical backgrounds we are conducting using co-simulation. Here it is important to note that many mathematically-based models of the physical phenomena does not just have an anlytic solution and thus when simulating such models we get approximations, and these are not necessarily refinements of each other (and many of them need to be calibrated to be close to what happens in reality). Another challenge that needs to be overcome is the fact that receiving data from the physical system can be a complicated process and this will also result in discretations with approximations of the real values (e.g. due to noise). In case data is received wirelessly there will also be a time delay and this matters in a digital twin setting, in particular if one wish to control the physical twin from the digital side. Being able to estimate the state (and state transitions) of a physical twin can also be challenging when it needs to be done purely on the data that is accessible from the outside. Finally, since the models of a CPS will never be having a behaviour which will be identical to the physical system, so it is likely that there will be drifting and thus one will need to calibrate the models once in a while and determining when and how to do this is also not obvious.

3.15 Functional, Safe, Secure CPS In contact with human beings

Thierry Lecomte (CLEARSY – Aix-en-Provence, FR)

Main reference Thierry Lecomte, David Déharbe, Paulin Fournier, Marcel Oliveira: "The CLEARSY safety platform: 5 years of research, development and deployment", Sci. Comput. Program., Vol. 199, p. 102524, 2020.

URL https://doi.org//10.1016/j.scico.2020.102524

This presentation reports on the return of experience collected during the last two decades, while applying formal methods for software-based safety critical systems, from design to exploitation. These systems, legacy or brand new, are in close contact with people. Forthcoming systems have to be analysed through a huge number of dimensions (safety, security, cybersecurity, AI, autonomy, etc.). Who is going to specify, design, V&V, certify, qualify them ? We need tools, standards, and people to achieve this – people from the 30% of the population, able to use abstraction, are required. Target customers are those who do not sleep well at night – the financial argument (FM are going to save money) is quite usually ignored.

3.16 ProB after 20 Years

Michael Leuschel (Heinrich-Heine-Universität Düsseldorf, DE)

ProB has been developed over around 20 years and was initially developed in SICStus Prolog. In this talk I will discuss various lessons learned over this period, touching development, maintenance, certification and reaching industrial users.

3.17 Integrated Rigorous Analysis of CPS: Examples from the Airspace Domain

Paolo Masci (NASA Langley – Hampton, US)

This talk discusses a range of verification and validation approaches employed by the research team at NASA Langley for the analysis of new automated navigation systems for general aviation. Concrete examples will be given based on Detect-And-Avoid (DAA) systems. DAA is the capability of an aircraft to remain well clear of other aircraft and avoid collisions. The idea behind DAA is to define a safe region around the aircraft and use the current position and velocity vector of the aircraft to compute possible route conflicts with other aircraft flying nearby. DAA was originally created for Unmanned Aerial Vehicles (UAVs). The research team at NASA Langley has created a reference implementation of a DAA system [1], and is now adapting the DAA concept to manned aircraft. The ultimate goal is to create a technology that can be used by pilots in the cockpit to enhance traffic awareness and support maneuver guidance when required to comply with see and avoid regulations [2]. This research is carried out within NASA's Air Traffic Management Exploration (ATM-X) project [3], which is looking into the future of airspace operations and services.

References

1 NASA Detect and AvoID Alerting Logic for Unmanned Systems (DAIDALUS) https://shemesh.larc.nasa.gov/fm/DAIDALUS/
2 NASA Detect and Avoid in the cockpit (DANTi) https://shemesh.larc.nasa.gov/fm/DANTi/
3 NASA Air Traffic Management Exploration (ATM-X) Project https://www.nasa.gov/aeroresearch/programs/aosp/atm-x

3.18 Generative Engineering: A Paradigm for the Development of Cyber-physical Systems

Andrei Munteanu (Siemens PLM Software, BE)

Generative engineering is a new paradigm for developing cyber-physical systems. Rather than developing, increasingly more detailed model of a system, multiple architectural system variants are computationally generated and evaluated, which would be prohibitively expensive to do by hand. The components and parameters that make up this system model optionally maps to library components in various simulations and analytics tools, with architectural models for those tools then automatically generated. This methodology was successfully applied to different use cases, from vehicle transmission design and hybrid vehicles to safety in avionics.

3.19 Logic of Autonomous Dynamical Systems

André Platzer (KIT – Karlsruher Institut für Technologie, DE)

Main reference André Platzer: "Logical Foundations of Cyber-Physical Systems", Springer, 2018.
URL https://doi.org//10.1007/978-3-319-63588-0

This talk highlights some of the most fascinating aspects of the logic of dynamical systems which constitute the foundation for developing cyber-physical systems (CPS) such as robots, cars and aircraft with the mathematical rigor that their safety-critical nature demands. Differential dynamic logic (dL) provides an integrated specification and verification language for dynamical systems, such as hybrid systems that combine discrete transitions and continuous evolution along differential equations. In dL, properties of the global behavior of a dynamical system can be analyzed solely from the logic of their local change without having to solve the dynamics.

In addition to providing a strong theoretical foundation for CPS, differential dynamic logics as implemented in the KeYmaera X prover have been instrumental in verifying many applications, including the Airborne Collision Avoidance System ACAS X, the European Train Control System ETCS, automotive systems, mobile robot navigation, and a surgical robotic system for skull-base surgery. dL is the foundation to provable safety transfer from models to CPS implementations and is also the key ingredient behind autonomous dynamical systems for Safe AI in CPS.

References

1 Platzer, A.: Logical Foundations of Cyber-Physical Systems. Springer, Cham (2018). 10.1007/978-3-319-63588-0

2 Platzer, A.: Logics of dynamical systems. In: LICS. pp. 13–24. IEEE, Los Alamitos (2012). 10.1109/LICS.2012.13

3.20 Inspiration from NASA Formal Methods Success Stories

Kristin Yvonne Rozier (Iowa State University – Ames, US)

This invited talk offers inspiration from the significant history of successful integration of formal methods into NASA projects. We highlight the differences between software and flight software, overview lessons learned from practical experience, and identify the limits of, and future challenges for, formal verification of aerospace systems. After drawing on examples from design-time verification of automated Air Traffic Control and runtime verification on-board Robonaut2, we visit the current full-system-lifecycle verification plans published for the NASA Lunar Gateway. We conclude with a collection of real-life, full-scale, open-source resources for the formal methods research community.

References

1 H. Erzberger, K. Heere, Algorithm and operational concept for resolving short-range conflicts, Proc. IMechE G J. Aerosp. Eng. 224 (2) (2010) 225–243

2 Zhao, Yang, and Rozier, Kristin Yvonne. "Formal Specification and Verification of a Coordination Protocol for an Automated Air Traffic Control System." In AVoCS 2012

3 Y. Zhao and K.Y. Rozier. Formal specification and verification of a coordination protocol for an automated air traffic control system. *Science of Computer Programming Journal*, volume 96, number 3, pages 337-353, Elsevier, December, 2014

4 Zhao, Yang, and Rozier, Kristin Yvonne. "Probabilistic Model Checking for Comparative Analysis of Automated Air Traffic Control Systems." In IEEE/ACM 2014 International Conference on Computer-Aided Design (ICCAD), 2014

5 C. von Essen & D. Giannakopoulou"Analyzing the Next Generation Airborne Collision Avoidance System" *TACAS* 2014

6 Marco Gario, Alessandro Cimatti, Cristian Mattarei, Stefano Tonetta and Kristin Y. Rozier."Model Checking at Scale: Automated Air Traffic Control Design Space Exploration." In *Computer Aided Verification (CAV)*, 2016

7 Rohit Dureja and Kristin Yvonne Rozier. "FuseIC3: An Algorithm for Checking Large Design Spaces." In Formal Methods in Computer-Aided Design (FMCAD), IEEE/ACM, Vienna, Austria, October 2-6, 2017

8 Rohit Dureja and Kristin Yvonne Rozier. "More Scalable LTL Model Checking via Discovering Design-Space Dependencies (D^3)." In *Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, part I, volume 10805 of Springer LNCS, pages 309-327, Springer-Verlag, Thessaloniki, Greece, 14-21 April 2018

9 B.Kempa, P.Zhang, P.H.Jones, J.Zambreno, K.Y.Rozier. "Embedding Online Runtime Verification for Fault Disambiguation on Robonaut2." FORMATS, LNCS vol 12288, 2020

10 Dabney, James B., Julia M. Badger, and Pavan Rajagopal. "Adding a Verification View for an Autonomous Real-Time System Architecture." In AIAA Scitech 2021 Forum, p. 0566. 2021

11 James Bruster Dabney, "FSW 2021: Using Assume-Guarantee Contracts In Autonomous Spacecraft." Online: `https://www.youtube.com/watch?v=zrtyiyNf674`

12 James Bruster Dabney, "FSW 2022: Using Assume-Guarantee Contracts for Developmental Verification of Autonomous Spacecraft." Online: `https://www.youtube.com/watch?v=HFnn6TzblPg`

13 B. Kempa, C. Johannsen, K.Y.Rozier. "Improving Usability and Trust in Real-Time Verification of a Large-Scale Complex Safety-Critical System." *Ada User Journal*, 2022

14 Alexis Aurandt, Phillip Jones, and Kristin Yvonne Rozier. "Runtime Verification Triggers Real-time, Autonomous Fault Recovery on the CySat-I." In *Proceedings of the 14th NASA Formal Methods Symposium (NFM 2022)*, Caltech, California, USA, May 24-27, 2022

15 Zachary Luppen, Michael Jacks, Nathan Baughman, Benjamin Hertz, James Cutler, Dae Young Lee, and Kristin Yvonne Rozier. "Elucidation and Analysis of Specification Patterns in Aerospace System Telemetry." In *Proceedings of the 14th NASA Formal Methods Symposium (NFM 2022)*, Springer, Caltech, California, USA, May 24-27, 2022

16 Benjamin Hertz, Zachary Luppen, and Kristin Yvonne Rozier. "Integrating Runtime Verification into a Sounding Rocket Control System." In *Proceedings of the 13th NASA Formal Methods Symposium (NFM 2021)*, Springer, Virtual, May 24-28, 2021

17 Abigail Hammer, Matthew Cauwels, Benjamin Hertz, Phillip Jones, and Kristin Yvonne Rozier. "Integrating Runtime Verification into an Automated UAS Traffic Management System." In *Innovations in Systems and Software Engineering: A NASA Journal*, Springer, July, 2021

3.21 All Eyes on Extra-functional System Properties: On the Formalisation and Analysis of Explainability and Morality for Autonomous Traffic Agents

Maike Schwammberger (KIT – Karlsruher Institut für Technologie, DE)

Joint work of Maike Schwammberger, Verena Klös
Main reference Maike Schwammberger, Verena Klös: "From Specification Models to Explanation Models: An Extraction and Refinement Process for Timed Automata", in Proc. of the Proceedings Fourth International Workshop on Formal Methods for Autonomous Systems (FMAS) and Fourth International Workshop on Automated and verifiable Software sYstem DEvelopment (ASYDE), FMAS/ASYDE@SEFM 2022, and Fourth International Workshop on Automated and verifiable Software sYstem DEvelopment (ASYDE)Berlin, Germany, 26th and 27th of September 2022, EPTCS, Vol. 371, pp. 20–37, 2022.
URL https://doi.org//10.4204/EPTCS.371.2

Motivation. During the last years, autonomous cars are increasingly capturing the markets worldwide. As such autonomous cars involve both software and hardware aspects, these systems can be summarised as Cyber-Physical Systems. Often, these systems also involve cooperation or interaction with human operators or end-users, thus leading to *Human Cyber-Physical Systems (HCPS)*. Ensuring functional properties of these HCPS is of the utmost importance to allow for a desirable future with them. Examples for functional properties are safety (e.g. collision freedom for moving HCPS) or liveness (a desired goal is finally reached). Fortunately, different research directions for analysing and proving functional system properties exist.

Apart from functional system properties, a variety of important *extra-functional* system properties must be ensured, which is the focus of this talk. In our case, we consider self-explainability and morality to be such extra-functional properties. Both fields have gained more and more attention within the last years of success for autonomous systems. With *self-explainability*, we describe the capability of a system to self-explain its actions and decisions to an addressee. Such an addressee may, e.g., be an engineer, an end-user or another HCPS. When we say that an HCPS *acts morally*, we mean that it can follow a given set of moral rules, as is, e.g., presented through the societal, cultural or legal context of the HCPS.

Approach. We introduce the modular MAB-EX framework for self-explainability[1]. The framework comprises four phases: First, the system is **M**onitored, e.g. through an observer mechanism. In the second phase, **A**nalyse, the monitored data is examined w.r.t. unusual behaviour that needs explaining. If the need for an explanation is identified, the formal core of an explanation is extracted from an *explanation model* in the **B**uild phase. An explanation model is a structure that we extract from formal system models and that contains formalised versions of explanations[3]. In the last phase, **EX**plain, the extracted, formal, explanation is translated into an explanation that fits for the intended addressee.

For morality, we envision a step-wise procedure to include morality into *autonomous traffic agents (ATAs)*, thus gaining moral ATAs[2]. For this, we will analyse a formalised set of traffic rules for conflicts and solve them by introducing moral rules to the ATAs. Conflicts could exist between different traffic rule in different contexts, or between an agent's goals and traffic rules. For instance, if a traffic sign demands that cars drive only at 50km/h, while an agents goal is to drive faster, a moral rule might be used to implement that the agent (temporarily) adapts their own goal.

Challenges. We perceive and discuss a variety of challenges in the field of formal analysis of extra-functional system properties of ATAs:

- How far can we go with formal methods in the area of extra-functional system properties?
- The endeavour of proving extra-functional system properties like self-explainability will be an interdisciplinary operation. What disciplines need to be involved and how can we bridge potential gaps between different disciplines?
- What types of extra-functional properties must be analysed and ensured?
- What are the further challenges that exist?

References

1 Mathias Blumreiter, Joel Greenyer, Francisco Javier Chiyah Garcia, Verena Klös, Maike Schwammberger, Christoph Sommer, Andreas Vogelsang, Andreas Wortmann. *Towards Self-Explainable Cyber-Physical Systems.* In: MODELS Companion, IEEE, pp. 543–548, 2019.

2 Astrid Rakow, Maike Schwammberger. *Brake or Drive: On the Relation Between Morality and Traffic Rules when Driving Autonomously.* In: 20th Workshop on Automotive Software Engineering (ASE), 2023.

3 Maike Schwammberger, Verena Klös. *From Specification Models to Explanation Models: An Extraction and Refinement Process for Timed Automata.* In: FMAS@SEFM, volume 371 of EPTCS, pp. 20–37, 2022.

3.22 Modeling and Analysis of Cyber-Physical Systems Using Actors

Marjan Sirjani (Mälardalen University – Västerås, SE)

Main reference Marjan Sirjani, Edward A. Lee, Ehsan Khamespanah: "Verification of Cyberphysical Systems", Mathematics, Vol. 8(7), 2020.
URL https://doi.org//10.3390/math8071068
Main reference Marjan Sirjani, Luciana Provenzano, Sara Abbaspour Asadollah, Mahshid Helali Moghadam, Mehrdad Saadatmand: "Towards a Verification-Driven Iterative Development of Software for Safety-Critical Cyber-Physical Systems", J. Internet Serv. Appl., Vol. 12(1), p. 2, 2021.
URL https://doi.org//10.1186/s13174-021-00132-z
Main reference Marjan Sirjani: "Power is Overrated, Go for Friendliness! Expressiveness, Faithfulness, and Usability in Modeling: The Actor Experience", in Proc. of the Principles of Modeling – Essays Dedicated to Edward A. Lee on the Occasion of His 60th Birthday, Lecture Notes in Computer Science, Vol. 10760, pp. 423–448, Springer, 2018.
URL https://doi.org//10.1007/978-3-319-95246-8_25

Our world has become a network of connected software systems, communicating with each other, and controlling physical systems. We have autonomous cars driving around, interoperable medical devices monitoring and controlling the health of patients and collaborating robots interacting with humans without separating fences. These systems are generally concurrent, distributed, and dynamic, with critical timing properties.

I will present our approach for analysis of timing properties of interoperable systems, using actor models and formal verification. Rebeca was designed more than 20 years ago as an imperative actor-based language with the goal of providing an easy-to-use language for modelling concurrent and distributed systems, with formal verification support. It was extended a few years later to support modelling real-time network and computational delays, periodic events, and required deadlines; and then extended to Hybrid Rebeca to support hybrid systems.

At the dagstuhl, I will briefly present our work that may be of interest for the audience. I will reflect on how we used Rebeca, its extensions, and its toolset for timing analysis and safety assurance of different systems, for example sensor network applications, and medical interoperable systems. I will present Hybrid Rebeca and our design decisions in extending Rebeca to support hybrid systems. I will present our work on model checking CPS by connecting Timed Rebeca and Lingua Franca (of Edward Lee from UC Berkeley). I will also explain our work on anomaly detection of CPS using formal verification at design time, and runtime monitoring during operation using an abstract digital twin that we call Tiny Twin.

3.23 Rigorous Development & Certification of Complex, Software Intensive Systems – My Wish List

Alan Wassyng (McMaster University – Hamilton, CA)

The talk tackled the question "Can we achieve the safety & dependability we need for extremely complex systems of systems that combine hardware & software, and may even include machine learning components?" I presented my personal wish list for techniques and approaches that I believe can help us answer the question in the affirmative.

Top of my wish list is "Incremental Product Family Assurance" to complement "Incremental Product Family Development", which I think is already well established. To support this we need effective and practical "Change Impact Analysis & Bi-directional traceability". I presented our work on the Workflow+ modeling framework as one approach that can help in this regard.

I also presented 8 Support Wishes ranging from "Systematic Methods To Explore Emergent Behaviour" to "Integrated Methods", with an emphasis on Model Driven Engineering. I ended the presentation with 5 Foundational Wishes ranging from "Separation Of Concerns" to "[building the] Assurance Case Before Start [of development]".

References

1 Nicholas Annable, Thomas Chiang, Mark Lawford, Richard F. Paige and Alan Wassyng. *Generating Assurance Cases using Workflow+ Models.* In Computer Safety, Reliability, and Security, Munich, Germany, September 6–9, pp. 97–110. Springer, 2022

3.24 Towards a Unifying Framework for Uncertainty in Cyber-Physical Systems

James C. P. Woodcock (University of York, GB)

Main reference Kangfeng Ye, Ana Cavalcanti, Simon Foster, Alvaro Miyazawa, Jim Woodcock: "Probabilistic modelling and verification using RoboChart and PRISM", Softw. Syst. Model., Vol. 21(2), pp. 667–716, 2022.
URL https://doi.org//10.1007/s10270-021-00916-8

There are many challenges to the satisfactory operation of cyber-physical systems (CPSs). They include architectural issues, real-time properties, human interaction, autonomy, privacy, safety, security, and uncertainty. Researchers who have analysed CPSs cite problems linked to security and uncertainty as the most common causes of failure [1]. We focus on uncertainty, a lack of knowledge about a system's state.

Computer scientists have proposed several formalisms for dealing with uncertainty. Probabilistic and statistical model checkers, such as Prism [5] and Storm [4], analyse a range of semantic models for these formalisms. These include discrete and continuous-time Markov chains and their nondeterministic extensions. These tools are good at interoperability. Verification-oriented formalisms include the following: Hehner's probabilistic predicative programming [3], the conditional probabilistic guarded command language [7], probabilistic Hoare logic [2], and partially observable Markov decision processes [6].

Research on describing and analysing uncertainty raises many questions. What does a unifying theory for uncertainty look like? What are the connections between semantics and tools that support the different approaches? Can we establish more connections? Can we support probabilistic and statistical model checking with theorem proving? Contrariwise, can we support theorem proving with probabilistic and statistical model checking? Can we establish uncertainty properties using correctness by construction? What about probabilistic refinement-based model checking? Can we qualify one analysis tool (as in DO-178C) and then map soundly into that tool for high assurance? What is the formal testing theory for a CPS with (say) unknown MDP semantics? What are the testability hypotheses (in Gaudel's sense)? How do we exploit the interplay between testing, proof, and model checking? What about uncertainty modelling and runtime verification? What role can unifying uncertainty formalisms and tools play in the development, application, and evaluation of CPSs?

We describe some preliminary work towards answering these questions.

References

1 Mah Asmat, Saif Ur Rehman Khan, and Shahid Hussain. Uncertainty handling in cyber–physical systems: State-of-the-art approaches, tools, causes, and future directions. Journal of Software: Evolution and Process, 2022.

2 Jerry den Hartog and Erik P. de Vink. Verifying probabilistic programs using a Hoare-like logic. Int. J. Found. Comput. Sci., 13(3):315–340, 2002.

3 Eric C. R. Hehner. Probabilistic predicative programming. In Dexter Kozen and Carron Shankland, editors, Mathematics of Program Construction, 7th International Conference, MPC 2004, Stirling, Scotland, UK, July 12-14, 2004, Proceedings, volume 3125 of Lecture Notes in Computer Science, pages 169–185. Springer, 2004.

4 Christian Hensel, Sebastian Junges, Joost-Pieter Katoen, Tim Quatmann, and Matthias Volk. The probabilistic model checker Storm. Int. J. Softw. Tools Technol. Transf., 24(4):589–610, 2022.

5 Marta Z. Kwiatkowska, Gethin Norman, and David Parker. PRISM 4.0: Verification of probabilistic real-time systems. In Ganesh Gopalakrishnan and Shaz Qadeer, editors, Computer Aided Verification – 23rd International Conference, CAV 2011, Snowbird, UT, USA, July 14-20, 2011. Proceedings, volume 6806 of Lecture Notes in Computer Science, pages 585–591. Springer, 2011.

6 George E. Monahan. A survey of partially observable Markov decision pro- cesses: Theory, models, and algorithms. Management Science, 28(1):1–16, 1982.

7 Federico Olmedo, Friedrich Gretz, Nils Jansen, Benjamin Lucien Kaminski, Joost-Pieter Katoen, and Annabelle McIver. Conditioning in probabilistic programming. ACM Trans. Program. Lang. Syst., 40(1):4:1–4:50, 2018.

4 Working Groups

4.1 Formal methods for hybrid systems: Challenges and research directions

Erika Abraham (RWTH Aachen University, DE), Wolfgang Ahrendt (Chalmers University of Technology – Göteborg, SE), André Platzer (KIT – Karlsruher Institut für Technologie, DE), Marjan Sirjani (Mälardalen University – Västerås, SE), Frank Zeyda (Zapopan, MX)

Hybrid and *cyber-physical systems* started to attract the interest of the *formal methods* community in the 90s, followed by a diversity of great ideas, elegant methods and impactful tools. However, till today, these methods and tools did not find their way into regular industrial usage. What are the major problems delaying a wider adoption?

Model building: As a crucial enabling factor for verification, we first need formal models for these systems.

- A hybrid system is typically composed of a controller and a continuous dynamical system controlled by it. Whereas models for the controller are developed at relatively early design phases, the modeling process considers the dynamics often too late. There is a communication problem between engineers constructing the system and modeling people making it difficult to get the right starting point for realistic dynamical models on the suitable level of abstraction.
- Often, different sources offer different information (e.g. on the dynamics, control, uncertainty, environment, requirements, etc.) that needs to flow together for model building. However, there is no clear methodology for synthesizing models from partial information from different sources.
- A related problem is the scarcity of notion of compositionality/composability/modularity for hybrid systems. Compositionality of hybrid systems modeling and reasoning works in logic per operator, but it needs good design to succeed for larger system components. Furthermore, there is no established way to jointly represent models together with their specifications and verification results, which would ease their adoption and maintenance during the system's life cycle.
- In general, for modeling we might not yet have the right interface between the hybrid and the discrete world. Here, research may benefit from further novel principles.
- For the modeling, no standard language exists. Different languages differ in their semantics (if a semantics is defined at all) and expressivity, which makes model transformation challenging. Consequently, applying different tools on the same problem requires a lot of effort, or may even be impossible. For instance the system time-horizon has a huge impact on the performance of some tools but not on others.
- The model is often not sufficiently maintained during system construction, leading to major differences between the system implementation and the current model. Consequently, previous verification results are not applicable any more, and the whole modeling and verification process needs to be carried out anew.

Specification: What is still missing is a specification formalism that is easier to use for engineers than the formal languages but still captures assumptions as well as guarantees of (hybrid) components, composition operators, and composability constraints (relative to the desired properties of the composed system). Education and training enables engineers to use the required logic, but more gentle specification languages may make that specification process easier for engineers who are novice in formal techniques.

Verification: Intensive research efforts in the last decades have led to a number of formal verification techniques and tools, but only a few of them are used by a larger community.

- The verification problem for hybrid systems is inherently hard. People might change research direction such that available tools are not maintained any more.
- The development of usable tools requires both strong science and significant engineering effort that is often impossible to find funding for.
- Industrial applications are doable, but often need a PhD student's help. Bachelor's students can also do impressive verification studies but of more medium complexity.
- The combined analysis of the discrete-continuous behavior of hybrid systems is hard. Separating discrete behavior and continuous dynamics for the verification process is only partially possible, because the hybrid system's safety conditions impact the needs of the discrete controllers.
- Abstractions (e.g. discrete abstractions combined with a counterexample-guided abstraction refinement approach) are possible but they do not always solve the problem, unless clever problem-specific insight is given to the tools.
- Rigorous verification needs a stack of rigorous tools that is hard to sustain.
- Most techniques are developed to compute (or approximate) the set of all states that a system can reach from a given set of initial states during its execution. However, there is nearly no support for more complex (e.g. temporal or spatial) properties except in deductive logic approaches.
- The controller is discrete, coupled with the physical world in the time dimension; it would be great to find a way to exploit this fact to simplify the analysis.

4.2 Human models for human cyber-physical systems

Maike Schwammberger (KIT - Karlsruher Institut für Technologie, DE), Borzoo Bonakdarpour (Michigan State University - East Lansing, US), Simon Thrane Hansen (Aarhus University, DK), Joseph Roland Kiniry (Galois - Portland, US), Régine Laleau (IUT Sénart-Fontainebleau, FR), Ken Pierce (Newcastle University, GB)

Cyber-Physical Systems (CPSs), often acting autonomously, are used in more and more application domains of our daily lives: Driving assistance systems, smart factories and smart homes are just some examples. While the level of autonomy of these systems increases, so also does the need for these systems to interact with human end-users, operators or engineers. A new type of system is born: Human Cyber-Physical Systems (HCPSs). With that, one question comes to the fore: How we can capture, analyze and formally verify human behavior in CPS models?

Challenges and Opportunities: We discuss a selection of topics and challenges that need to be addressed for ensuring a satisfying and safe interaction of human and CPS.

- Human models for self-explainability: The capability of a Cyber-Physical System to self-explain its actions is a crucial feature for HCPS, especially if shared and safety-critical tasks of the CPS and the human exist. However, such explanations must be targeted towards a variety of addresses: E.g., end-users, engineers, operators or lawyers. For different addressees of explanations, we need different human models. What techniques do we have to model humans?

- To answer the previous question, cognitive models as they are used by psychologists come to our minds. However, how do we translate psychologists' cognitive models into formal models? A widely used approach to specify knowledge is, e.g., to use an (auto) epistemic logic.
- Should a human's behavior be modeled as a continuous-time or hybrid process?
- What assumptions do we need about human behavior? For this, literature often uses a notion of *rational agents*. A rational agent is an entity that generally tries to use optimal actions based on some given knowledge, rules and goals. Nonetheless, also notions of irrational or even evil agents should be taken into account for worst-case analyses.

4.3 Formal Methods and Certification

Danielle Stewart (Galois - Minneapolis, US), Kristin Yvonne Rozier (Iowa State University – Ames, US), Stefan Hallerstede (Aarhus University, DK), Stanley Bak (Stony Brook University, US), John Hatcliff (Kansas State University - Manhattan, US), David Hardin (Collins Aerospace – Cedar Rapids, US), Andrea Bombarda (University of Bergamo – Dalmine, IT), Fuyuki Ishikawa (National Institute of Informatics – Tokyo, JP), Gabor Karsai (Vanderbilt University, US), Thierry Lecomte (CLEARSY – Aix-en-Provence, FR), Michael Leuschel (Heinrich-Heine-Universität Düsseldorf, DE), Alan Wassyng (McMaster University – Hamilton, CA)

Notetaker: Danielle

Summary

Different regulatory agencies have different expectations and requirements for certification processes. Certain agencies are more comfortable with formal methods and verification approaches, such as the NSA. Dave described a Type I crypto system called Janus that required accreditation through the NSA. They were able to provide formal evidence of various code properties and the NSA gave approval for the system. Other agencies are not familiar enough with formal method approaches to understand the artifacts, let alone the benefit of the approach. Safety and security are totally different. The security process (Common Criteria) is fully defined, and formal methods – along with penetration testing – is part of the process. The certification authorities are well equipped to evaluate the formal methods artifacts. The nuclear regulation in Canada is more serious about formal methods artifacts. They decide where they think problems exist and focus on those parts of the system. But any discussion about formal methods in certification needs to also look at the problem of tool qualification. If you want to certify a system to a certain point, any formal methods tools must be qualified to that standard as well. This is a difficult and expensive process. Formal methods can, however, provide insight into the system during certification, even if to the developer alone. It can aid in understanding and documentation, even if those artifacts are not used directly within the certification process.

4.4 Stochastic cyberphysical systems

Cláudio Gomes (Aarhus University, DK), James C. P. Woodcock (University of York, GB), Joanna Delicaris (Universität Münster, DE), Noah Abou El Wafa (KIT - Karlsruher Institut für Technologie, DE)

Our group focused on the discussion of what barriers are there for the integration of stochastic behavior into formal methods. One barrier is that, currently, formalisms traditionally used to model stochastic behavior, like Discrete Time Markov Chains (DTMC), have no formal semantics, that enables them to be used in, e.g., theorem provers. Just like their deterministic counterparts, stochastic formalism have relationships between them. For example, each step of a DTMC is a simple Markov chain, and each transition in a Markov chain is a statistical distribution. Another barrier is therefore to represent the links between these formalisms by, e.g., defining Galois connections between these formalisms. Another barrier: What are the methodologies to build stochastic models from physical prototypes. This might be well known to statisticians, but not to computer scientists. Finally, we need methodologies on how to identify the sources of uncertainty. Barrier: how are uncertainties propagate through a coupled system, and do they affect the software elements' formal models? Here we can draw from a huge body of literature with formalisms to quantify and propagate uncertainty: sensitivity analysis, monte-carlo simulations, stochastic differential equations, etc.

4.5 Technology Needs – Self-Explainability of Cyber-Physical Systems

Maike Schwammberger (KIT – Karlsruher Institut für Technologie, DE), Thierry Lecomte (CLEARSY – Aix-en-Provence, FR), Alan Wassyng (McMaster University – Hamilton, CA)

As autonomous systems get more and more complex, we need to ensure that they remain or get understandable. For instance, if an AV needs to change a driving strategy unexpectedly, this should be explained to a passenger to retain usability and trustworthiness into the AV. Using formal methods, we can automatically generate formal explanations from specification models (e.g. UML diagrams, timed automata,...). Such *formal cores of explanations* allow for formal verification. From these explanation cores that have been extracted at design time, explanations may be translated at run-time whenever an explanation is needed.

Challenges and Solutions:

- **There is a need to identifying different addressees:** Different addressees mean that differently grained and detailed explanations are needed. For instance, an engineer needs a different explanation than an end-user. For this, expertise from requirements engineering could be taken into account (e.g. "Personas" or "User Classes") or different user models might be learned using AI techniques.
- **Verification and validation of formalized explanations:** A validation of explanations cannot be done isolated from the addressees, as , e.g., it is necessary to know what is relevant for different addressees. A formalization and verification of different addressee's mental models is needed for a joint verification of formalized explanations and user models.

- **Structure of formalized explanations:** Safety explanations could be produced from the discovery of why a system was designed in a certain way. Explanations start from feared events. These events are refined into assumptions agreed on by all experts and measures taken to avoid these events, to produce a tree. All leaves are either assumptions or functions that are sufficient when combined to avoid these events.
- **Automatically extractable explanations can help in system debugging:** If a formalized explanation has been automatically extracted from a specification model in a provably correct manner, and the explanation is still wrong or does not make sense, this means that something is wrong with the system specification. This can be especially helpful for very complex systems that are hard to understand or verify even by experts. This approach could be a fast way to identify some system faults before we start time and resource costly verification mechanisms.

4.6 Digital Twins

Peter Gorm Larsen (Aarhus University, DK), Einar Broch Johnsen (University of Oslo, NO), Leo Freitas (Newcastle University, GB), William Earl Scott (ScubaTx – Newcatle upon Tyne, GB), Andrei Munteanu (Siemens PLM Software, BE), Klaus Kristensen (Bang & Olufsen – Struer, DK)

This group worked to identify the main research challenges for Digital Twins (DT). This also involved assessing how Formal Methods (FM) may be incorporated to enhance the engineering and assurance of DT.

Requirements for the overall twin system:

- Who is responsible for what at different stages of the lifecycle of a digital twin system?
- The requirements of a DT system need to include the main purpose of the DT.
- Would it make sense to create new special DSLs for configurations, monitors and/or what-if scenarios?
- Declare properties of interest to be true of the system/module/unit.
- How to formally specify and evaluate hypothetical (what-if) scenarios?

Applications of FM in Digital Twins:

- Some engineering challenges are related to getting data (in a filtered form) into formal models in a satisfactory manner.
- When we need humans in the DT loop we also need human models. How do we get those models?

Challenges for applying FM in Digital Twins:

- What are the pros and cons of using FM inside DTs?
- How to determine the collection of different models to be included inside the DT (and consider how to select between them)?

Challenges in Digital Twins that would benefit from applying FM:

- Providing evidence of the “goodness” of the digital twin.

- The composition of DTs will benefit from an analysis from FM stakeholders.
- The placement of simulation models in a distributed setting will require different analysis.

Correctness criteria for Digital Twins:

- How to define the assumptions required before being able to verify properties?

Experiences with rigorous engineering of Digital Twins:

- How to discover calibration options/needs?
- Different case studies are important here (some of these will be reported about in a new book on engineering digital twins).

Models for safety and security of Digital Twins:

- Most likely these shall be indicated in some of the monitors.
- How can we trust the results from services considering what-if scenarios?

Fidelity of Digital Twins:

- How accurate does the DT models need to be in relation to the performance of the physical twin?

Validation of Digital Twins:

- Start with historical data and use this as arguments to FM models (potentially in a co-simulation context).
- Determine how it is possible to get data transferred from a physical twin to its digital twin (there can be significant challenges with respect to handling of data because of noise and the fact that the input to models may need to be derived from data that can be extracted).

Achievements applying FM to Digital Twins:

- FM has numerous opportunities for having impact on the DT domain in semantics for different notations, clarification of different concepts.
- Run-time verification is essentially the core of the monitors here.

Participants

- Noah Abou El Wafa
KIT – Karlsruher Institut für Technologie, DE
- Erika Abraham
RWTH Aachen University, DE
- Wolfgang Ahrendt
Chalmers University of Technology – Göteborg, SE
- Stanley Bak
Stony Brook University, US
- Ezio Bartocci
TU Wien, AT
- Stylianos Basagiannis
Raytheon Technologies – Collins Aerospace – Cork, IE
- Andrea Bombarda
University of Bergamo – Dalmine, IT
- Borzoo Bonakdarpour
Michigan State University – East Lansing, US
- Joanna Delicaris
Universität Münster, DE
- Leo Freitas
Newcastle University, GB
- Cláudio Gomes
Aarhus University, DK
- Stefan Hallerstede
Aarhus University, DK
- Simon Thrane Hansen
Aarhus University, DK
- David Hardin
Collins Aerospace – Cedar Rapids, US
- John Hatcliff
Kansas State University – Manhattan, US
- Fuyuki Ishikawa
National Institute of Informatics – Tokyo, JP
- Nils Jansen
Radboud University Nijmegen, NL
- Einar Broch Johnsen
University of Oslo, NO
- Gabor Karsai
Vanderbilt University – Nashville, US
- Joseph Roland Kiniry
Galois – Portland, US
- Klaus Kristensen
Bang & Olufsen – Struer, DK
- Régine Laleau
IUT Sénart-Fontainebleau, FR
- Peter Gorm Larsen
Aarhus University, DK
- Thierry Lecomte
CLEARSY – Aix-en-Provence, FR
- Michael Leuschel
Heinrich-Heine-Universität Düsseldorf, DE
- Paolo Masci
NASA Langley – Hampton, US
- Monica Moniz
Cambridge University Press, GB
- Andrei Munteanu
Siemens PLM Software, BE
- Ken Pierce
Newcastle University, GB
- André Platzer
KIT – Karlsruher Institut für Technologie, DE
- Anne Remke
Universität Münster, DE
- Kristin Yvonne Rozier
Iowa State University – Ames, US
- Maike Schwammberger
KIT – Karlsruher Institut für Technologie, DE
- William Earl Scott III
ScubaTx – Newcatle upon Tyne, GB & Newcastle University, GB
- Marjan Sirjani
Mälardalen University – Västerås, SE
- Danielle Stewart
Galois – Minneapolis, US
- Alan Wassyng
McMaster University – Hamilton, CA
- James C. P. Woodcock
University of York, GB
- Frank Zeyda
Zapopan, MX

Report from Dagstuhl Seminar 23042

Quality of Sustainable Experience (QoSE)

Katrien De Moor[*1], **Markus Fiedler**[*2], **Ashok Jhunjhunwala**[*3], **and Alexander Raake**[*4]

1 **NTNU – Trondheim, NO.** katrien.demoor@ntnu.no
2 **Blekinge Institute of Technology – Karlshamn, SE.** markus.fiedler@bth.se
3 **IITM Research Park – Madras, IN.** ashok@tenet.res.in
4 **TU Ilmenau, DE.** alexander.raake@tu-ilmenau.de

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 23042 "Quality of Sustainable Experience (QoSE)". The seminar aimed to bring together people from different fields, perspectives and backgrounds. The participants discussed how experiences – as the main selling point of products and services – in various ICT-related domains can be made more sustainable, how they can contribute to relevant sustainable development goals, and how the quality and degree of sustainability of such experiences may be evaluated and be better understood. The main objectives of the seminar were to foster new alliances, to inspire, to trigger scientific renewal, as well as to identify future opportunities and research challenges through a hands-on approach.

Seminar January 22–25, 2023 – https://www.dagstuhl.de/23042
2012 ACM Subject Classification Human-centered computing → Human computer interaction (HCI); Social and professional topics
Keywords and phrases Design for Sustainability, Digitalisation, Human-Computer Interaction, Information and Communication Technology, Wellbeing
Digital Object Identifier 10.4230/DagRep.13.1.184

1 Executive Summary

Katrien De Moor (NTNU – Trondheim, NO)
Markus Fiedler (Blekinge Institute of Technology – Karlshamn, SE)
Ashok Jhunjhunwala (IITM Research Park – Madras, IN)
Alexander Raake (TU Ilmenau, DE)

In line with the shift towards a more experience-centered paradigm in product and service design, Information and Communication Technology (ICT) is seen as an important enabler of immersive, and potentially transformative digital experiences. As such, ICT has a huge potential to address fundamental human needs (e.g., experiencing pleasure, relatedness); to tackle "slow-change problems" (e.g., adopting a sustainable lifestyle) and to keep up important social functions also in times of crisis (e.g., distance education, communication, entertainment) through experiences. However, two non-negligible downsides of ICT are its potential negative impact on wellbeing (e.g., addiction, blurring online/offline identities), and its growing ecological footprint, with ever-increasing demands to satisfy the Quality of Experience (QoE) of increasingly spoiled users.

[*] Editor / Organizer

Quality of Sustainable Experience (QoSE), *Dagstuhl Reports*, Vol. 13, Issue 1, pp. 184–215
Editors: Katrien De Moor, Markus Fiedler, Ashok Jhunjhunwala, and Alexander Raake
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

On this background, this Dagstuhl Seminar set out to discuss the topic of "Quality of Sustainable Experience", and hence the challenge of how to transform existing physical and digital experiences into more sustainable (ideally fossil-free), yet human-centered and well-appreciated ones. The aim was to bring together experts from different fields addressing the multi-faceted topic from their own perspective, using their distinct tools and methods. The main objectives with the seminar were to foster new alliances, inspire, trigger scientific renewal, as well as to identify future opportunities and research challenges through a hands-on approach. The participants discussed how experiences – as the main selling point of products and services – in various ICT-related domains can be made more sustainable, how they can contribute to relevant sustainable development goals, and how the quality and degree of sustainability of such experiences may be evaluated and be better understood. The seminar adopted a bottom-up approach to identify key areas for future work within the outlined scope and converged into four topics that were further discussed in smaller groups, with the aim of better understanding current knowledge gaps and challenges and to identify topics and areas where the represented disciplines could – in the short to longer term future – make a genuine impact towards more sustainable ICT-based experiences.

The group discussions during the seminar centered around four main topics, namely (1) collaborative XR and remote attendance, (2) quantification / measures of QoSE, (3) ICT as a means to drive sustainability and (4) Needs versus greeds. During the discussions, the groups identified a set of challenges and generated "NOW", "WOW" and "HOW" ideas [1], which are described further in Section 5.

The seminar has already resulted in a number of spin-off activities, for example at the 15th International Conference on Quality of Multimedia Experiences (QoMEX 2023), having a particular focus on the transition towards more inclusive and sustainable mutimedia experiences. More concretely, the conference is hosting a special session involving several seminar participants and organizers entitled "Towards the design and evaluation of sustainable multimedia experiences", and one of the seminar participants was invited to give a keynote at the conference. Another concrete outcome is the initiative to apply for funding of a COST Action in order to build a community on the topic of QoSE. Finally, a video trailer has also been compiled to put focus on and raise awareness of the topics discussed at the seminar [2].

References

1 COCD-box school of creative thinking. https://schoolofcreativethinking.nl/articles/cocd-box/. Accessed: 2023-05-18.

2 Dagstuhl Seminar 23041: Quality of Sustainable Experience (QoSE): short video trailer. https://youtu.be/D2vswi_8O7A. Accessed: 2023-05-29.

2 Table of Contents

Executive Summary

Overview of Talks

Group work: introduction

Reports from the working groups

3 Overview of Talks

3.1 Let's talk about designing Sustainable Interactions through Accessibility

Stepanie Arevalo Arboleda (TU Ilmenau, DE)

My current research focuses on experiences in immersive environments (augmented and virtual reality) together with the use of robotic systems to enhance communication for the aging population. Designing for sustainable interactions could be approached from an inclusive perspective, where technology is conceived and designed to allow for adaptable experiences. Sustainability through accessibility can be approached methodologically by understanding the current experiences of people with disabilities and the aging population using participatory design and experience-driven design. I consider that Sustainable HCI and QoSE could also include Disability Interaction and accessibility when conceptualizing sustainable experiences that go beyond designing for sustainable technology but invite reflection on technologies' uses and evoke self-evaluation of intentions and behavior. I would like to encourage discussions on how to include Disability Interaction and Accessibility in the QoSE agenda.

3.2 The user experience of assessing ethical issues of AI systems

Emma Beauxis-Aussalet (VU University Amsterdam, NL)

The users are tasked with assessing the ethical issues of AI systems, at different phases of a system life cycle. The users have very diverse backgrounds, e.g., with technical expertise or domain expertise(s) – but are generally Dutch. A most characteristic element of the user experience is the knowledge gap(s) between users, between users and the technology, or between users and the domain (e.g., poverty prevention, fraud detection, resource allocation). It makes collaboration between diverse stakeholders essential to succeed with the task. Misunderstandings and miscommunication are key issues in such collaboration. Fear and stress are also inherent to the user experience, due to the many impacts of AI on society – some of which already had devastating consequences. Conflicts of interests also arise, e.g., between technology suppliers (especially contractors) and policy makers. Our work relates to many societal aspects of sustainability, especially considering the many impacts of AI on sustainability. But energy consumption is outside of our scope.

The specificity of our approach is not to design new user interfaces, visualisations, or tutorials. Instead, we first focus on designing the assessment techniques (e.g., the appropriate metrics, statistics, sampling method), and designing the human organisation that is needed for assessing AI (e.g., gathering people with the right set of skills and responsibilities). But to do so, user-centered design may prove harmful (sometimes) due to the many knowledge gaps between stakeholders. Conflicts of interest are particularly important and challenging, and may occur in many endeavours towards sustainability – which is often considered an overhead with unwelcome costs.

3.3 Sustainable and inclusive innovation

Michael Best (Georgia Institute of Technology – Atlanta, US)

My overall research focuses on computing and global development. I use the UN SDGs to frame a lot of my work and so pain sustainability with a broad brush. One of my current projects is focused on inclusive innovation and sustainable entrepreneurship with a focus on the East Asia region. We are aiming to collaboratively develop up some new programs/facilities in Taiwan and perhaps Thailand. I would love to learn from this community inspired ways that we can act as valuable, ethical, and humble global collaborators as we partner on this endeavor.

3.4 Connecting people

Pablo Cesar (CWI – Amsterdam, NL)

My research combines human-computer interaction and multimedia systems, focusing on modelling and controlling complex collections of media objects (including real-time media and sensor data) that are distributed in time and space. My aim is to better integrate core human-computer interaction methodologies and computer science research. In particular, I am interested on "connecting people": how we can make remote togetherness possible. Since 2005, I have been involved in a number of research projects on Social TV, multi-party videoconferencing, and more recently social XR as a collaboration and communication medium. We are moving towards a connected intelligent world, in which always-on sensing and monitoring enable rich immersive media experiences (remote working, medical consultation, online cultural heritage experience, entertainment). These systems help towards a more resilient society, providing the means to communicate across distance in meaningful and natural manners, thus reducing the travel needs. Still, apart from the usage of resources, there are many sustainability goals, as identified by the UN: quality of education, good health, gender equality, decent work and economic growth, resilient infrastructure, sustainable cities and communities. My hope in this seminar is to discover the work of others and better understand how we as scientists can address the overall picture.

3.5 Towards more humane and sustainable experiences supported by digital technology

Katrien De Moor (NTNU – Trondheim, NO)

Recent forecasts show an alarmingly high carbon footprint of ICT in the middle-term future, due to, among others, the increasing energy demand of data centres, as well as increasing use and consumption, including unsustainable use and viewing practices (e.g., binge-watching, media-multitasking), which have become more common over the last years and have partly been associated with negative health and wellbeing effects. Moreover, the wide range of experiences enabled by digital technology (e.g., XR, AI and IoT-supported smart environments) come with a growing number of ethical concerns (e.g., safeguarding meaningful human agency, designing for genuine empowerment, privacy under threat, inclusivity and equity), which should be even more prominently on the agenda. Through my research, I aim to subscribe the growing plea for a shift towards a more sustainable and humanity-centered paradigm, which considers to a much larger extent how digital consumption, increased user expectations and data demand may impact individuals, society at large and our environment and which wants to better "align technology with humanity's best interests" (see e.g., humanetech.com). My interest and activities in this area are grounded in human-centered approaches and focus on:

1. Aspects related to the design, evaluation and use of audiovisual media (e.g., video conferencing, video streaming, immersive applications) and deal with aspects related to improving these more sustainable experiences, supporting inclusion and triggering more sustainable use practices.
2. The need to better understand users' awareness (and lack of it) of their own "invisible" digital carbon footprint; and explore strategies and concrete mechanisms that may help to trigger more conscious and responsible consumption (both from the well-being- and environmental point of view).
3. Human- and humanity-centric design principles and the need for meaningful ways to evaluate whether desired outcomes such as empowerment, meaningful human agency, inclusivity, equity are reached.

3.6 Sustainable Software Engineering for Sustainable Development

Yvonne Dittrich (IT University of Copenhagen, DK)

I would like to share 2 research points: 1) In a project on "Sustainable Irrigation Advice for Mid-Himalayan Farmers using Smart satellite Image Analysis" we address sustainability in 3 different ways:

1. Water management is part of climate change mitigation
2. We apply co-design to embed the irrigation advice in the farmers' irrigation practices
3. The project aims at not only addressing the technical feasibility, but also the economically viable deployment and evolution by taking a software ecosystem approach.

The other project explores the development of domain specific standards of reporting of environmental and societal impacts and corporate governance. The European Commission is developing legislation for reporting and investors increasingly ask for this data. In future we will be accountable for the energy consumption of our services. In both cases, technical solutions need to take the needs of different actors and stakeholders into account. They need to support cooperation of heterogeneous stakeholders and support decentralised governance structures.

References

1 Dittrich, Y. *Software engineering beyond the project–Sustaining software ecosystems.* In Information and Software Technology, 56(11), 1436-1456, 2014.

2 Wang, C., Østerlund, C., Jiang, Q., & Dittrich, Y. *Becoming Sustainable Together: ESG Data Commons for Fintech Startups*. In Proceedings/International Conference on Information Systems (ICIS), 2022.

3.7 Designing Sustainable Experiences

Markus Fiedler (Blekinge Institute of Technology – Karlshamn, SE)

Sky-rocketing energy prices have increased our awareness of resource limitations. Having worked with quality-versus-energy tradeoffs since 2010, the emerging multi-reality digiphysical experiences make me curious of their potential to reduce environmental footprints without sacrificing the essentials of the experiences. Bringing together the "Research through Design" and "Quality of Experience (QoE) by Design" [1] principles, I see a great potential to create beyond-expectation immersive experiences with sustainability in mind, for instance Extended Reality (XR) telemeetings. Thereby, creative design of experimental artefacts based on fundamental relationships between QoE and provisioning, measurements and modelling efforts will pave the way towards optimised quality-versus-energy performance, expressed for instance through measures such as "QoE per Watt" (QoEW) or "QoE per Joule" (QoEJ) [2] – or as "QoE per kWh" (QoEkWh) that relates directly to the energy bill.

References

1 Fiedler, M., Möller, S., Reichl, P., and Xie, M., *QoE vadis? (Dagstuhl Perspectives Workshop 16472)*. Dagstuhl Manifesto, 7(1):30-51, 2018.

2 Fiedler, M., Popescu, A., and Yao, Y., *QoE-aware sustainable throughput for energy-efficient video streaming*. In Proc. of 2016 IEEE BDCloud, SocialCom and SustainCom, Atlanta, GA, Oct. 2016.

3.8 Energy-Efficient Video Communications

Christian Herglotz (Universität Erlangen-Nürnberg, DE)

Nowadays, research targeting the energy efficient use of video communication technology is an important research topic. In this respect, our team focuses on the energy consumption of two important aspects in a video communication pipeline: First, the generation, compression, and provisioning of videos, second, the consumption of videos on end-user devices. Methodologically, we usually start by measuring the energy consumption of a video system, then analyze the behavior with respect to parameters such as hardware, software, and video properties, and come up with numerical models that are further exploited to reduce the energy consumption. We noticed that next to the energy efficiency of distinct devices, the overall energy consumption of video services draws more and more attention in academia and industry. Hence, in this seminar, interesting challenges are to jointly optimize the energy consumption of distinct devices and a complete video service while keeping a high QoE for the end user.

3.9 Sustainable Remote Work: How to make virtual / hybrid conferences enjoyable?

Oliver Hohlfeld (Universität Kassel, DE)

Many processes in work environments (including the prominent publication mode in Computer Science with in-person gatherings to present research output) have relied on in-person meetings, which often require travel. For many researchers, traveling to conferences may well be a significant, or even largest, contributor to their annual carbon footprint. To be sustainable, alternative processes – such as virtual and hybrid attendance modes – need to be established. To be successful, these must meet the goals of the gathering and provide a high experience (QoE). How to make virtual and hybrid meetings enjoyable and therefore sustainable is a question directly related to QoE research. To address this, I have studied the QoE of virtual conference attendance via a survey approach that identified areas in which this mode works and also exposes its limits. As a future trend, hybrid conferences are having their moment, primarily due to the prolonged and open-ended transition period from the COVID-19 pandemic. While hybrid conference also address rising concerns relating to the carbon footprint of air travel. Further, they promote inclusiveness of members of the community, e.g., those that are not able to attend due to family obligations, budget restrictions, difficulties obtaining a visa or disability. Yet, it remains unclear of how to design hybrid conferences well to achieve a high participant QoE, which will be an upcoming challenge to the QoE community. This imposes a direct question to this seminar on how to make work processes – such as hybrid attendance – enjoyable and thus sustainable by means of QoE research.

3.10 A Greener Experience: Trade-offs between QoE and CO_2 Emissions in Today's and 6G Networks

Tobias Hoßfeld (Universität Würzburg, DE)

Main reference Tobias Hossfeld, Martin Varela, Lea Skorin-Kapov, Poul E. Heegaard: "A Greener Experience: Trade-offs between QoE and CO_2 Emissions in Today's and 6G Networks", IEEE Communications Magazine, pp. 1–7, 2023.
URL https://doi.org/10.1109/MCOM.006.2200490

Quality of Sustainable Experience raises several research questions which are addressing the different pillars of sustainability: human, social, economic, environmental. In particular, environmental sustainability calls for the following: What is the trade-off between QoE and CO_2 emission? How can an optimal operational point be derived in practice? Is the ratio of goodness, e.g. QoE, and badness, e.g. CO_2 emissions, e.g. energy consumption, a meaningful key value indicator (KVI) for today's and 6G networks? This ratio goodness to badness is Kleinrock's power metric from queueing theory. How much reduction in CO_2 emission can be achieved by a green user as compared to a high-quality user? How much reduction in CO_2 emission can be achieved by moving towards a green network? What is the relative impact on the reduction of CO_2 emissions of green user behavior as compared to green networking? Is it more relevant to focus (i) on green user behavior and empowering green user behavior or (ii) on green networking technology today and in the future in year 2030? What are the implications of solution approaches on the networking and communications technology?

3.11 Ecologically Valid Experiments

Lucjan Janowski (AGH – Univ. of Science and Technology – Krakow, PL)

I work mainly with classical video quality. Right now, I am developing a Virtual Reality Laboratory. So I expect to work more with VR/AR in the context of 5G. My main focus right now is the development of ecologically valid experiments. Ecologically valid experiments are closer to the real-life scenario. We expect that such experiments can reveal situations where quality is less important than concluded from a classical lab study. It gives an option for further optimization of network resources, limiting energy consumption. Further, quality should not be the only goal and the trade between quality and the resources used should be better understood. An important component is not only the network, but also the habits of the users, like playing music from video, not even watching. A different essential aspect of the quality system is the development of algorithms for recompression from clear energy. We have to understand quality and user behavior outside the laboratory. Only then can specific solutions be proposed.

3.12 Sustainable India and World

Ashok Jhunjhunwala (IITM Research Park – Madras, IN) and Reema Saha (IITM Research Park – Madras, IN)

It was good to see young academicians from so many countries, concerned about the society and the world. On the one hand, they worry that climate change could do a irreparable damage to our earth in coming years. On the other hand, most of them felt somewhat powerless, as big Governments and big industry seem to drive every aspect of life on planet. The youngsters work very hard to just have a decent life. They seem to be powerless in the current situation. Recognising this, the seminar attempted to do two things. The first was the little actions that they could carry out individually and in groups to start making some difference. The second was to dream of a future society – may be 100 years from now. What would be the norms and ways such that the people would really be empowered, free from the control of big governments and big industry. The participants knew that it was a mere beginning, but felt that even imaging a future society would be the first step to move towards such society in future.

3.13 Multisensory User Experience in eXtended Reality

Effie Lai-Chong Law (Durham University, GB)

One of my current research foci is multisensory user experience (MUX) in extended reality (XR). Given the immersive and presence experience enabled by XR, the number of XR-based applications is ever-increasing, especially for social interaction such as in games, therapy, and training. Avatars representing interactants are typically used. Non-verbal sensory signals (i.e. facial expression, gesture, gait) are essential for emotion portrayal. The MUX of social XR is determined by the extent to which intended emotions can be conveyed and recognized by the interactants. As XR technologies are highly energy-demanding, they can have a very negative impact on sustainability. The higher the avatar fidelity is, the higher the MUX quality can be, but the higher the energy consumed and costs. To address how to improve the greenability of XR, I am investigating how the avatar fidelity can be minimised but without compromising the perception and recognition of the intended emotions required for successful social interaction and quality MUX. Extensive user-based studies are designed to identify the minimum fidelity level for each type of sensory signal per emotion in a range of contexts. Advanced rendering techniques and machine learning models for adjusting avatar fidelity will be deployed. How MUX varies with different avatar fidelity levels will be evaluated. Overall, the main challenges are to scope the large problem space, considering the nature of emotion and the rapid growth of XR tech and techniques.

3.14 Imagination, Climate Futures, and the Qualities of Sustainable Experiences

Dan Lockton (TU Eindhoven, NL)

My work explores designing tools for participatory (re-)imagining and futuring in an age of transitions (and crises) in climate, energy, health, and social inequalities. How we experience the world (interacting with technology, but also how we encounter societal and infrastructural systems) affects how we imagine, understand, live, and what we see as possible in our collective futures, with consequences for sustainability. Design has an important role to play in engaging with imagination and futures, and the urgency of climate crisis makes this acute: enabling people to share their experiences with others, giving voice to underrepresented experiences, and turning ideas into prototypes (including interfaces) which can be experienced, used, lived with, and reflected upon. Designers can bring plural possible futures to life, in the present. I see the qualities of how we experience the systems around us as important in building more sustainable ways of thinking and acting – better connections to impacts, consequences, and each other.

3.15 sustainability storytelling: mobilizing transformation

Colin Maclay (USC – Los Angeles, US)

While I integrate research, teaching and engagement like any good faculty member, I am a hacker of universities and find that a lot of my attention goes to less traditional work like community creation, demonstrating different ways to do things and building new institutions. I lead fellowships, research groups and programs for troublemaking practitioners and scholars, host a podcast on popular culture and social change, create welcoming and non-hierarchical environments and try to engage respectfully and generatively with the community and the world. After decades spent on the interaction of information and communications technology with organizational and institutional change, i have spent recent years more focused on sustainability, environmental justice and climate change, finding significant overlap and complementarities. A large part of my current attention is on reorienting the functions of my university around sustainability, where I focus primarily on research and engagement (and leadership, of course). I've learned that seeking a sustainability orientation echoes the challenges of the digital, diversity and other transformations before it, requiring not just modest changes to what we do or who we hire, but fundamental shifts for both individuals and organizations in how we see ourselves, our practices and our mission. It confronts identity, asks that we engage emotional complexities, requires us to engage our imagination, create different systems, communicate differently and address other seemingly distant considerations. I'm excited to hear what others are thinking about, experimenting with and learning as we navigate the unseen and deeper barriers that will begin to allow the sort of transformational developments that facilitate not just human survival, but thriving.

3.16 How to characterize QoSE [kō-zē] experiences?

Alexander Raake (TU Ilmenau, DE)

We investigate perception and experience for traditional and immersive audiovisual media, including video (e.g., high-resolution, high dynamic range), spatial audio, and technology for Augmented, Virtual and Mixed Reality (AR/VR/MR). Two types of "resources" may be considered in terms of sustainability: (1) Human mental and physical resources, for example measuring fatigue for telemeetings versus face-to-face, or the positive impact on wellbeing with mediated social presence. (2) Energy and natural resources consumed along the end-to-end chain (e.g., by one media system implementation versus another), or resources saved (e.g., meeting via videoconferencing or MR rather than travelling). In this context the question arises, how "sustainable experiences" can best be characterized, and the result be applied towards a more sustainable way of life. A holistic and collaborative approach is needed to achieve this. Here, I see the QoSE seminar as a possible crystallization point for sharpening the participants' views and future collaborative work.

3.17 From QoE to Digital Humanism and Digital Ecology

Peter Reichl (Universität Wien, AT)

Over the years, QoE turned out to be very successful in redirecting the attention of the communication networks community towards the user. However, the current multiple crises indicate that we have to extend our perspective on both sides, leading to two key questions: (1) Do we have the technology we need, and do we need the technology we have? (2) Which world are we currently building? Recently, several new movements formed to address these issues, especially in the context of the "Vienna Manifesto on Digital Humanism" or the "Rat für Digitale ökologie Berlin". That leads to question number (3): What can QoE research learn from, and how can QoE research contribute to this broader perspective on the Digital Change?

3.18 Moodlebox: A Broadband Connectivity with Sustainable Quality of Experience for e-Learning in Rural and Remote Areas?

Fatuma Simba (University of Dar es Salaam, TZ)

Rural areas are characterized by scattered settlements, lack or limited Information and Communication Technologies (ICT), hence they are disadvantaged in accessing e-learning resources. Different technologies have been proposed to address broadband connectivity for e-learning in rural and remote areas, such as the 3G UMTS operating in the 900MHz frequency band, and the Television White Spaces (TVWS), due to their wider coverage and capability to offer broadband connectivity. However, further research revealed that broadband networks configured in the best-effort approach cannot deliver video streaming with the required QoS for e-learning, which implies that users will end up unsatisfied, hence poor quality of experience. Trends in e-learning shows development of MoodleBox, which is a standalone mobile device that can provide both local broadband connectivity and e-learning resources. Potential research area here is to evaluate performance of MoodleBox in delivering multimedia e-learning contents in rural settings towards sustainable quality of experience in e-learning.

3.19 How to assess the value of services more holistically

Sascha Spors (Universität Rostock, DE)

We work in the field of digital signal processing with a focus on the processing of audio and medical signals. Many applications and services use signal processing to extract information, for signal enhancement, or for transformation into other representations. While traditionally, model-based techniques played a much more prominent role, the employment of data-driven methods (machine learning, artificial intelligence) has increased significantly in recent years. This enabled significant breakthroughs, for instance, in speech recognition. However, in many cases, at the cost of increased resource consumption and corrupted privacy. While some of the current and upcoming technical possibilities are of great use to society, their employment is often discussed on an economic level only, and sustainability plays a minor role. I want to discuss how we can assess the benefit of new applications and services more holistically, including society, sustainability, and economics.

3.20 Innovation experience management

Fee Steinhoff (Hochschule Koblenz – Remagen, DE)

Innovation experiences result from the (often unconscious) comparison of needs and offer and lead to emotion, cognitions and actions. For a convincing innovation experience, the following areas need to be actively "managed":

- Utility: Is the innovation addressing “real” human problems and needs which are relevant to the target customers? (exemplary tool: jobs-to-be-done approach).
- User Experience: Is the innovation providing a convincing experience in the product and usage context? Is the user able to use the innovation easily and does the user like the way the innovation looks and feels? (exemplary tool: iterative UX prototyping & testing).
- Customer Experience: Is the innovation providing a convincing experience in the broader market context? For example, does the innovation create positive moments of truth and emotional binding along the whole customer journey? (exemplary tool: customer experience blueprinting).
- Transforming innovation experiences into more sustainable ones is obviously a very challenging task in our days. From a management perspective, exemplary questions are: How can innovators create convincing sustainable innovation experiences? How should innovators deal with the current “more, better, higher” consumption mantra? Which tools and methods are helpful to create sustainable innovation experiences (e.g. sustainable business model design patterns)? etc.

3.21 Digitalization supporting the integration of sustainability in product development tools

Denny Carolina Villamil Velasquez (Blekinge Institute of Technology – Karlskrona, SE)

Researching in the field of sustainable product development and supporting manufacturing companies, we have identified that companies struggle to integrate sustainability in their processes. In Blekinge Institute of Technology, we have developed and tested tools and methods to guide companies to adopt a strategic sustainability perspective based on the Framework for Strategic Sustainable Development, by considering a holistic view, the environmental, economic and social dimensions of sustainability, the assessment of the complete product lifecycle, stakeholders’ collaboration and a long-term perspective. Finding that the sustainable society transformation requires the participation and support of many fields, where digitalization can be used to support this transformation. Moreover, digitalization might facilitate manufacturing processes and the usability of decision-support tools to develop solutions with a higher sustainability performance. Therefore, it is essential to discuss how digitalization can support the implementation of sustainability, considering trade-offs and additional requirements e.g., knowledge, infrastructure, management, social interaction and circularity.

3.22 QoSE for immersive communication

Irene Viola (CWI – Amsterdam, NL)

Remote telepresence is essential to enable connection among users at a distance, facilitating communication and collaboration, while decreasing the amount of travelling and commuting required; as such, it has become a key point in research agendas both in the European and national level to create sustainable travel habits and more liveable cities. Current telepresence solutions for telepresence have been shown to create exhaustion and fatigue, due to the unnatural way in which communication takes place, such as limited mobility and close-distance eye gaze. Extended Reality (XR) telecommunication systems promise to overcome the limitations of current real-time teleconferencing systems, enabling a better sense of immersion, enhancing the sense of presence and fostering more natural interpersonal interactions. To achieve their goals, they need to be designed keeping the user as the central perspective. How can we optimize the quality of such systems, such that they can maximise the Quality of Experience for the user, while ensuring the sustainability of their operating principles? How can we incorporate Quality of Sustainable Experience in the design, implementation and evaluation of such systems?

3.23 Beyond Human-Centeredness in Experience Design for Sustainability

Kaisa Väänänen (University of Tampere, FI)

I work with experiences that drive ways in which people's activities in the world can be more sustainable, both socially and environmentally. Recently, together with my team we have worked a lot with AI-driven systems such as social robots and the ways in which they could motivate/persuade people act more sustainably. (At the same time realising that hardware robots may not be very sustainable in themselves.) Methodologically, we employ human-centered design thinking, and especially co-design and co-creation approaches, both in-situ and (when needed, e.g. due to pandemic) online. We also work with industry to help them adopt Human-Centered AI (HCAI) design approaches. It is timely and relevant for sustainable experience design to move beyond human-centeredness, towards what has been labeled as "post-human", "more-than-human" or "planetary" design by various authors. While these concepts are attractive, they are currently still quite abstract and philosophical in terms of how to apply them in practical product and service design. Furthermore, qualities of AI – proactivity, dynamism and autonomy – introduce new possibilities to the system design process. Hence, we need to define practices for integrating the needs of humans, the ecosystem and AI to advance sustainability through experience design. These practices have to take into account the needs of the planet, not just of humans.

3.24 QoE for mobile immersive media

Hans-Jürgen Zepernick (Blekinge Institute of Technology – Karlskrona, SE)

Sustainability has been a crucial demand for all generations of wireless communications systems in terms of optimal resource allocation subject to given key performance indicators. Due to the required high data rates, low latency, signal processing complexity and other constraints, maintaining QoE and sustainable QoE is a challenging task in ultra-reliable low latency communication applications such as mobile immersive media. 6G technology shall support immersive mobile media experiences that extend over the entire continuum of digital computer-generated virtual worlds. A key emphasis in the growth of digital value platforms will be the convergence of multimodal engagement with media and the physicality of lived experience. In this context, architectures and technologies for green 6G networks shall be envisaged that offer sustainable QoE. In my current work, I conduct subjective experiments for mobile immersive media, subjective and objective quality assessment, mobile multimedia signal processing, analytical approaches on QoE-assured VR video streaming, energy harvesting in wireless networks. I am interested in discussing experimental designs for sustainable QoE, subjective and objective metrics for sustainable QoE assessment, technologies to enhance energy efficiency and low power consumption for 6G and beyond mobile telecommunication systems with application to mobile immersive media.

3.25 ICT and Sustainability – More than energy?

Thomas Zinner (NTNU – Trondheim, NO)

I am working in the broad area of networked systems and applications. Technical systems have become more and more complex, and many problems are solved adding additional resources, bringing resources closer to the users, or using programmable hardware. My work aims at designing mechanisms and algorithms to improve the operation of technical systems by enabling customization considering user-centric metrics or utility functions. While this can for instance improve system utilization, improve revenue and reduce the carbon footprint per user, it also puts additional burden on the control planes, and increases costs and computational complexity.

Hence, my work strives sustainability of ICT, but my interest also covers how new ICT systems can be used to enable applications improving sustainability, e.g., immersive haptic / XR applications further reducing traveling. For that I am trying to understand relationships between social, environmental and economic factors.

4 Group work: introduction

4.1 Seminar structure and used methods

Katrien De Moor (NTNU – Trondheim, NO), Markus Fiedler (Blekinge Institute of Technology – Karlshamn, SE), Ashok Jhunjhunwala (IITM Research Park – Madras, IN), and Alexander Raake (TU Ilmenau, DE)

The seminar adopted a genuinely bottom-up approach and started with the phase of *inventory and exploration.* After the short introductory talks of the participants which took place on the morning and part of the afternoon of day 1, a first clustering session took place. Participants were asked to write down their core expertise and knowledge areas related to the topic. These inputs were clustered into the following overall categories during the *analysis and condensation phase*:

- ICT / measurement: e.g., sustainability of ICT and potential of ICT to contribute to more sustainable experiences in other sectors; methodologies, metrics, best practices e.g., to increase energy efficiency, to measure environmental impact
- Experience measurement: e.g., knowledge and methods to evaluate users' experiences, expectations, quality perceptions
- Changing perceptions and behavior: e.g., insights on design for behavioral change, triggering motivation and engagement.
- Policies and broader implications: e.g., regulatory landscape, policy perspectives, role of activism, implications for digital ecosystems and business models.

This first initial clustering allowed us to situate the participants in different areas based on the perspectives, knowledge, methods, etc. they brought to the seminar. As a next step, we conducted a brainstorming session on concrete topics that participants would like to address during the seminar and that they considered important and potentially impactful towards driving sustainable experiences. First, the participants were asked to write down ideas on post-its (individual phase), after which all ideas were placed on the blackboard, and everyone could build upon the listed ideas. After a saturation of ideas was reached, all ideas were briefly explained and further elaborated upon in a plenary session, so that all participants would have a good understanding of what was meant with the different ideas / topics. The last step of this session was a prioritization of topics to work on. All participants were given three vote stickers (1st, 2nd and 3rd choice) and could indicate which topics they would be interested to discuss during the subsequent group work. This prioritization resulted in the following topics that were discussed in smaller groups during the *analysis and condensation phase*:

- Group 1: Collaborative XR and remote attendance
- Group 2: Quantification / measures of QoSE
- Group 3: ICT as a means to drive sustainability
- Group 4: Needs versus greeds

By matching the topical prioritizations and expertise clusters, the discussion groups were selected such that each had a representative from all four expertise areas / perspectives listed above. The *synthesis phase* consisted of a number of activities. The first task of the group work was to discuss the group's topic more in-depth, to explore the views represented within the group and to discuss where there is a potential for impact. For this discussion, we used the COCD method [1], which distinguishes between NOW, WOW and HOW-ideas.

- NOW-ideas have a more short-term focus, are relatively easy to implement, are low-risk and generally not controversial.
- WOW-ideas can also be implemented, but in a slightly longer time-frame. Such ideas are exciting, innovative, potentially breakthrough ideas.
- HOW-ideas are more longer term, are, from the current perspective, considered more as longer-term dreams and challenges, ideas for the future, "cathedral" ideas.

All groups identified NOW-, WOW- and HOW-ideas and discussed what would be needed to realize these ideas. The groups documented their ideas via the online tool Taskcards [2]. Finally, for the last phase of the group work the groups switched topics and provided peer feedback on another group's ideas by means of De Bonos' six thinking hats [3]. Each hat represents another type of perspective:

- *White hat*: Information. Facts and information, neutrality, objective point of view. What is needed in terms of facts and data? What is missing? Where can more information be found?

- *Red hat*: Feeling and intuition. What does your gut feeling say? (no justification needed), spontaneous reactions? Does it feel right? Both positive and negative feelings are welcome and do not need to be justified.
- *Yellow hat*: Possibilities. Identify positive sides and possibilities, visionary thinking. Why is this worth trying out? How can it lead to improvement / value? Visions and dreams are allowed, speculative thinking as well.
- *Green hat*: Creativity. Thinking creative, opportunities for growth, how to extend? Which ideas have been presented? How can they be further explored and further developed? Alternatives and suggestions for solutions? New ideas, build on each other ideas. Criticism is not allowed with this hat on.
- *Black hat*: Critical perspective. This hat represents the devil's advocate. Focus is on identification of negative aspects, risky elements, weaknesses. Focus on vulnerabilities. Objective, rational evaluation.
- *Blue hat*: Process perspective. Where in the process is the group with their idea? What is the intended goal / outcome? What should be done now? Any decisions that need to be made? How to continue the work with the presented idea? Think in terms of process-orientation.

References

1 COCD-box school of creative thinking. `https://schoolofcreativethinking.nl/articles/cocd-box/`. Accessed: 2023-05-18.

2 TaskCards. `https://www.taskcards.de/#/home/start`. Accessed: 2023-05-18.

3 Edward De Bono. *Six Thinking Hats: The multi-million bestselling guide to running better meetings and making faster decisions.* Penguin uk, 2017.

5 Reports from the working groups

In the following section, we provide a brief overview of the main outcomes and ideas discussed in the different groups. The rapporteur for the group is always denoted with an (*) in the list of group members.

5.1 Group 1: Collaborative XR and remote attendance

Markus Fiedler (Blekinge Institute of Technology – Karlshamn, SE), Stepanie Arevalo Arboleda (TU Ilmenau, DE), Pablo Cesar (CWI – Amsterdam, NL), Effie Lai-Chong Law (Durham University, GB), Fatuma Simba (University of Dar es Salaam, TZ), and Hans-Jürgen Zepernick (Blekinge Institute of Technology – Karlskrona, SE)

The group discussion on collaborative extended reality (XR) and remote attendance started with collecting items and ideas. In a second step, these were matched to the COCD questions. In the sequel, we present the emerging sets of items and ideas merged with feedback from the review group 3, amongst others wondering which organization/individuals would be most feasible to take care of the various challenges.

5.1.1 NOW-topics and cases

Collaborative XR and remote attendance approaches and solution should be aligned with the UN Sustainable Development Goals (SDGs) [1], in particular SDG4 (quality of education), SDG5 (gender equality), SDG10 (inclusion) and SDG11 (sustainable cities and communities). Ethical, Legal and Social Implications (ELSI) must be taken care of, ethics by design should be the preferred approach. Accessibility and inclusion are essential in XR, which necessitates inclusive and participative design. As current VR learning, training and medical experiences do not come close enough to reality, hybrid XR and digiphysical settings should be considered. Physicality (e.g. feedback) and visual representation (e.g. facial expressions) need to decoupled and customized in order to include users (SDG10), convey the intended content and allow for tradeoffs of experiences. The latter are frequently targeted in sustainability-inspired comparative studies, which in turn require reliable data to be telling. For example, energy consumption should be estimated with reliable precision. Last but not least, limitations such as delays and cost incurred by trendy technology need to be overcome.

5.1.2 WOW-ideas

1. Inclusive and accessible XR experiences will be designed, and it is expected that they will be constantly evolving and impact ELSI in a positive manner. Also, people will be able to express themselves in XR in various versions, which will help to address and overcome prejudices and expectations.
2. Fidelity and altered physicality have to be chosen and controlled carefully depending on task and content in order to enable acceptable holoportation experiences.
3. XR experiences powered by alternative energy (through various harvesting approaches) will reduce the energy footprint and allow usage in remote areas.
4. Virtual coffee breaks and other happenstances, allowing for true digiphysical meeting experiences supported by multisensory interfaces (incl. smell and taste) and 3D audiovisuals.

5.1.3 HOW-ideas

1. How to address technology-related and -induced inequalities, e.g. w.r.t. SDG4 (education)?
2. How to handle delays and latencies in XR multiparty communication systems?
3. CoVid in mind triggered the controversial idea of an XR Dagstuhl Experience, moving 2D meetings to XR meetings with improved interactivity and well (re-)presented behavioral cues. While XR is failing on the very motto of Saarland people "Hauptsach' gudd gess" (the main thing is to eat well), it might at least help to solve the enigma of Dagstuhl's "White Lady".

References

1 The 17 Goals. https://sdgs.un.org. Accessed: 2023-05-18.

5.2 Group 2: How to characterize QoSE [kō-zē] experiences? Towards a measurement framework for Quality of Sustainable Experiences

Yvonne Dittrich (IT University of Copenhagen, DK), Emma Beauxis-Aussalet (VU University Amsterdam, NL), Tobias Hossfeld (Universität Würzburg, DE), Lucjan Janowski (AGH – Univ. of Science and Technology – Krakow, PL), Alexander Raake (TU Ilmenau, DE), Daniel Schien (University of Bristol, GB), and Thomas Zinner (NTNU – Trondheim, NO)

The following text represents a summary of the work by Group 2 and the subsequent review by other groups, especially Group 4, and respective further dicsussions.

For both aspects individually, experiences on the one hand, and environmental impact on the other, different measures and measurement approaches have been developed in the past. In this group 2, we have discussed a framework for characterizing "Quality of Sustainable Experiences" (QoSE), with the aim to jointly characterize the user experience and the associated sustainability of the used system or service. The overarching perspective was referred to as the "cathedral view" within the group. In the group work, the framework was primarily instantiated for the case of media technology, for the example of videostreaming, for which some quantitative measures for both the "experience" and the associated consumption have been investigated in the past. A joint footprint and QoE measurement view had been addressed before by some of the group members, too, e.g. in [1] in terms of power consumption vs. video streaming QoE, and [2] with regard to CO_2 emission and streaming QoE. In the group's discussions, a novel component was the possible impact of the users' awareness of the degree of sustainability of their product-related behavior, which has been integrated into the experiencing process, updating an existing model view on QoE formation [3, 4, 5, 6].

To this aim, the group has sketched a first graphical representation of the framework, see Figure 1. Here, W2I refers to the willingness to invest, that is, to consume more moderately and at lower perceptual quality, if this reduces the environmental impact. Both during system operation (left) and system / service production and operation (right), environmental resources are being consumed. It is noted that the figure acknowledges the fact that the exact consumption of resources may be difficult to attribute to individual systems and/or manufacturers / service providers ("Attribution Problem"). It is general consensus in the QoE community now, that QoE happens in the users' minds and results from the appraisal of the experience with regard to expectations. Here, awareness and W2I were thought by the group to influence expectations, hence increasing QoE in spite of possibly lower sensory / perceptual quality. In case of the manufacturer / service provider, awareness for the sustainability impact of the "experiences they sell" may lead to a more careful handling of resources and acceptance, that users may not strive for better and better perception. On the very left hand side, examples for aspects that need to be measured or characterized are indicated. Here, at the border between system / service and "measures", the associated cost is given as a measure, which currently still strongly impacts user expectations and decisions in terms of acceptance. Also for the providers or manufacturers, cost is a key measure, determining many decisions. Here, too, awareness may be a driver for updated, more environmentally sustainable decisions. To raise awareness, sustainability- and/or experience-related measures can be used to display and inform (red in Figure 1) about the environmental impact and the role of experience therein, positively influencing production and consumption patterns.

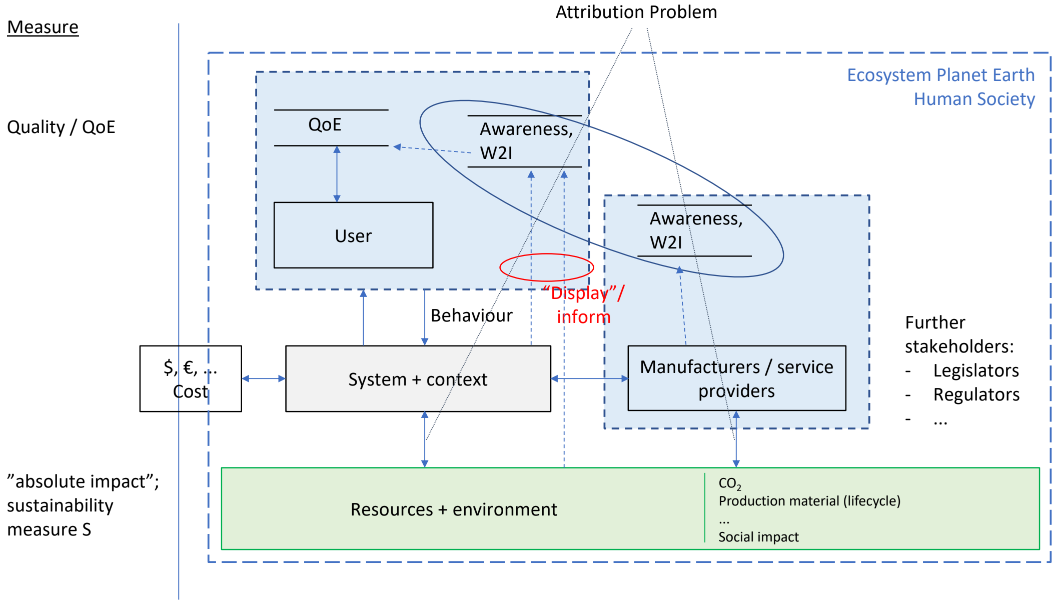


Figure 1 Conceptual model for QoSE measurement framework.

The ideas can be assigned to the categories of NOW, WOW and HOW as follows;

5.2.1 NOW-ideas

1. The framework concept and figure shall be incorporated into a conceptual paper
2. QoE and CO_2 or energy consumption may be considered together in research, also in conjunction with some aspect of "awareness"
3. The ongoing legislation on reports of companies of a certain size on environmental, societal and corporate governance need to be inspected for specific QoSE measures to be derived
4. The provision of usage reports for each user with regard to sustainability of each application / service could be a feasible goal

5.2.2 WOW-ideas

1. Making environmental and societal impact of the consumption subject of the quality of experience
2. Quantify "awareness"
3. Quantify QoE/sustainability with somewhat more evolved measure (beyond CO_2)
4. Changing the hidden optimisation criteria for technology design from economics only to also include sustainability
5. Conceive information approaches for sustainability, e.g., in terms of an intermediate, ICT-related, consumption-related "nutriscore"
6. Usage of service and substitution, e.g., in terms of "Drink less tea when watching TV (or other resource consuming activities like boiling water)?"

5.2.3 HOW-ideas

1. Enable the quantification and ultimately reduction of the "planetary resource usage" / sustainability and the associated and underlying "QoE", including a running measurement framework

2. Understand the relation between QoE and wellness
3. Have “planetary consumption” established as a sort of currency
4. Enable counter-weighing societal impact vs. the fun and its sustainability imprint

5.2.4 Feedback and conclusions Group 2

In the group and based on the feedback provided from other groups, especially Group 4, it was agreed that in future joint research, the started work will be complemented by a literature review, and by collecting existing as well as specifying new measures for QoE and user experience on the one hand, and sustainability on the other. Here, the validity and relevance of the measures was considered an important aspect. Besides the academic literature and recommendations from standardization bodies, considerations by policy-making agencies were identified as key resources, and such agencies also ultimately as the target group for the measurement framework. It was further agreed that the continued work will need to result into a clearer roadmap with reachable goals. Further, the notion of “what is needed”, “what is enough” and how these are being perceived by individual users have been discussed, motivated by the respective considerations in this regard by other groups.

References

1 Christian Herglotz, Werner Robitza, Matthias Kränzler, Andre Kaup, and Alexander Raake. Modeling of energy consumption and streaming video qoe using a crowdsourcing dataset. In *2022 14th International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2022.

2 Tobias Hossfeld, Martin Varela, Lea Skorin-Kapov, and Poul E. Heegaard. A greener experience: Trade-offs between qoe and CO_2 emissions in today’s and 6g networks. *IEEE Communications Magazine*, pages 1–7, 2023.

3 Ute Jekosch. *Voice and Speech Quality Perception.* Springer, Berlin, 2005.

4 Patrick Le Callet, Sebastian Möller, and Andrews Perkis, eds. Qualinet White Paper on Definitions of Quality of Experience (2012). European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), 2013.

5 Alexander Raake. *Speech quality of VoIP: assessment and prediction.* John Wiley & Sons, 2006.

6 Alexander Raake and Sebastian Egger. *Quality and Quality of Experience*, pages 11–33. Springer International Publishing, Cham, 2014.

5.3 Group 3: ICT as a means to drive sustainability

Ashok Jhunjhunwala (IITM Research Park – Madras, IN), Dan Lockton (TU Eindhoven, NL), Colin Maclay (USC – Los Angeles, US), Peter Reichl (Universität Wien, AT), Reema Saha (IITM Research Park – Madras, IN), Kaisa Väänänen (University of Tampere, FI), Irene Viola (CWI – Amsterdam, NL), Markus Fiedler (Blekinge Institute of Technology – Karlshamn, SE)

We present the emerging sets of questions and ideas related to time horizons and based on the discusssions within the group around the role of ICT and how it can be a means to drive sustainability.

5.3.1 5 years: NOW-ideas

There is a need to learn from recent CoViD experiences, and some of the new practices might be continued while observing longer-term implications. We observe relevant and promising trends such as moves to remote work and education, buying local, "right to repair" legislation as well as changes in consumption patterns and increased awareness of cooperatives, exploitation scenarios, outsourcing of externalities and more ethical supply chains, which is also visible in the younger "Generation Zero's" attitudes, approaches and trends [1]. Indeed, there are many "ICT for Good" examples such as "Mastodon" [2] and "Fediverse" [3] (decentralized and community-owned social media), "Do Not Pay" [4] (providing AI-based legal support), and messaging apps used for organizing communities. The question arises what (affirmative) values in terms of plurality, diversity, individuality and care of each other will matter beyond the 5-year time horizon?

5.3.2 25 years: WOW-ideas

The big transitions' effects on people will become obvious, implying tensions between survival, convenience, and bigger questions of existence. Fixation on (quality of) experience might become controversial, and the question how Quality of Life can be maintained or even improved with less consumption (and movements) will gain importance. ICT services should be incentivized to keep their users healthy, including the right to opt out. Hopefully, democratisation of access will provide more equal opportunities on a global scale. The question emerges which layers of the technology stack should be publicly owned, implementing a Fediverse [3] on which people can build their ICT solutions. How does governance of tech companies, the technology and the legislation need to evolve to serve public interest?

5.3.3 100 years: HOW-ideas

A vision for 100 years may include the Locavore idea "70% of what you consume is from 100-150 km of where you live" [5]. It becomes increasingly important where and how value is created (e.g. by sharing, repairing, re- and upcycling). Local governance, good relationships between communities, open access to networks and technologies will allow communities to create goods and services, yielding improved conditions for mankind. Accompanied by corresponding incentives for research, innovation and development (RID), ICT is a key tool to enable this utopia, keeping in mind the "authenticity of the human experience".

References

1 Millennials are shattering the oldest rule in politics. https://12ft.io/proxy?q=https%3A%2F%2Fwww.ft.com%2Fcontent%2Fc361e372-769e-45cd-a063-f5c0a7767cf4, 2023. Accessed: 2023-06-01.

2 Social networking that is not for sale. https://joinmastodon.org/en. Accessed: 2023-06-01.

3 Fediverse. https://www.fediverse.to. Accessed: 2023-06-01.

4 The world's first RobotLawyer. https://donotpay.com/, 2023. Accessed: 2023-06-01.

5 What is a Locavore? https://www.treehugger.com/what-is-a-locavore-1204001, February 2022. Accessed: 2023-06-01.

5.4 Group 4: Needs vs. Greeds

Michael Best (Georgia Institute of Technology – Atlanta, US), Katrien De Moor (NTNU – Trondheim, NO), Christian Herglotz (Universität Erlangen-Nürnberg, DE), Oliver Hohlfeld (Universität Kassel, DE), Sascha Spors (Universität Rostock, DE), Fee Steinhoff (Hochschule Koblenz – Remagen, DE), and Denny Carolina Villamil Velasquez (Blekinge Institute of Technology – Karlskrona, SE)

The considerations tackled in this topic targeted the question whether a high QoE is really needed or not and aimed to challenge the paradigm of always more and always better. For most applications, it would be possible to define a certain minimum, basic requirement, which is sufficient to satisfy the needs of the user while avoiding that the striving for fulfillment of fundamental psychological needs becomes an act of greediness. Such "sufficiency" thresholds could be investigated for different use cases and implemented as the default configuration of applications. In this regard, the group discussed the need for a definition and better understanding of "sufficiency": what does greedy or ungreedy behavior mean, what does it imply and what would be meaningful measures? Further, it was discussed that we should investigate which needs that drive human behavior are related to sustainability and how these could be used for pushing QoSE? A challenge here is however that the traditional test designs to evaluate quality are not suitable for this goal and would need to be redesigned. As a result, there is a need for new methodological approaches and metrics representing "sufficiency" as opposed to metrics targeting high quality. We briefly discuss the main NOW, WOW and HOW ideas, also incorporating the feedback from group 2 (De Bonos hats feedback round).

5.4.1 NOW-ideas

The main topic that was discussed is the need for a better understanding and definition of the common constructs and aspects (e.g., QoSE, sufficiency, striving for fulfillment of fundamental needs and balancing: when does need become greed?, theoretical models for consumption behavior). In particular, it was discussed that what greedy behavior entails is likely case-specific and that the line between greed and need-driven behavior is not clear-cut and represents a tension that should be investigated. Further, the group discussed the need to better understand what the outcomes and consequences of greedy behavior are (e.g., increased carbon footprint, impact on well-being or one's mental health), and whether and how they can be measured and visualised. The proposal in this respect was to write a white paper based on a thorough literature review, incorporating also literature from relevant related fields, to better define the relevant concepts and quality needs, greeds, and their multiple facets before deriving any metrics. As a part of this exercise, concrete use cases should be defined, since the means to address need, greed, sufficiency may be more actionable if specific use cases are targeted.

A second NOW idea is to (re-) run studies with an additional set of measures, e.g., including questions and measures related to needs (actual vs. perceived), tasks and purpose of using a specific service, acceptability and behavioral measures. However, a starting point here should be a thorough check of existing databases to get an overview of what is already available.

5.4.2 WOW-ideas

Concrete suggestions and goals that were discussed include:

1. Studies on sufficiency to understand where the thresholds are situated. However, here it was pointed out that this may be culturally different and that other variables may play in here. In addition, there may not be clear-cut thresholds, but rather grey zones, which should be better understood as they can help to map users' willingness to "sacrifice" or to contribute (when formulated positively). The concept of fairness was also coined in the discussion of this idea as potentially relevant. Finally, in "sufficiency modeling" and need-related research, it should be considered that needs may change over time (reduce or become stronger).
2. Definition of meaningful metrics and useful subjective measures of sufficiency. Efforts aiming to define relevant metrics and measures in this respect should consider what has been done in relevant related fields.
3. Triggering user empowerment, better consumer awareness, more informed decisions. The group identified a potential to allow users / consumers to take more informed decisions, based on the assumption that users are today not well informed and therefore lack the necessary insights into how their choices and use of various services may impact sustainability. However, this requires good measures and reliable indicators (e.g., to visualise carbon footprint associated to a service usage session). Overall, the importance of striking a good balance between paternalizing / dictating what is "good", non-greedy behavior vs. empowering users and letting them decide for themselves, was underlined.
4. Gamification: challenge, compare, compete to trigger behavioral change. The idea to use nudging and gamification mechanisms to help users to adopt more pro-environmental behavior when it comes to use of digital technology was generally considered positive, but it was also pointed out that such an approach also has important limitations and may not reach all segments of the population. A broader understanding of different measures that can be used to incentivize users, e.g., depending on contextual factors, is therefore needed. A goal should also be that it's hip, attractive, "in" to be an environmental-friendly, low-energy consumer, so that people feel motivated and inclined to adopt a more sustainable lifestyle.

5.4.3 HOW-ideas

They include

1. finding meaningful "punishments" for non-sustainable habits. The group discussed what might be an equivalent to solutions to prevent waste in a food context. Some of the ideas discussed include paying or apps that are not used, or rather paying only if you really use an app (after a testing period). Further exploration of such ideas is needed, but it should be ensured that there is room for individuality and differentiation and that such mechanisms do not have the opposite effect (e.g., that usage gets enforced in order to avoid having to pay).
2. Product and cost should cover the whole cost of a service (externalities). While this would potentially also lead to more conscious consumption, there are various challenges to consider (e.g., how to globally enforce this). Yet, approaches that address both the user perspective, economic implications and environmental impact together would be useful and overall, such an approach could trigger more transparency.
3. Move towards a post-growth economy which is not prevailed by capitalism and economic incentives.

6 Pictures

This section contains a set of visual impressions from the seminar – joint work (Figures 2 to 7) and social activities (Figures 8 to 11) – as well as pictures of the participants (Figure ??) and co-organisers (Figure 12).

Figure 2 Presentation by delegates (Daniel Schien).

Figure 3 Discussion in plenum.

Figure 4 Group work (Group 1).

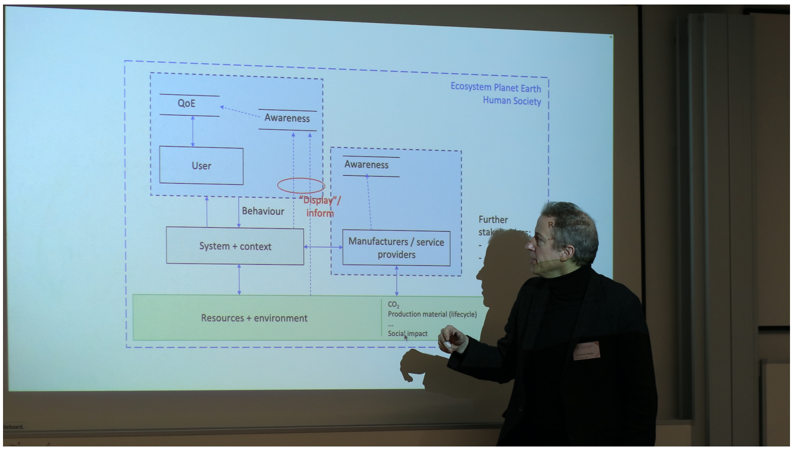


Figure 5 Presentation of group work (Group 2).

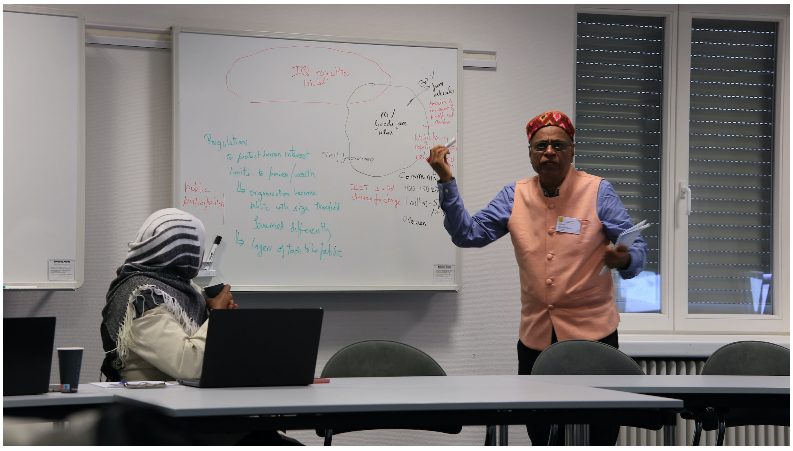


Figure 6 Presentation of group work (Group 3).

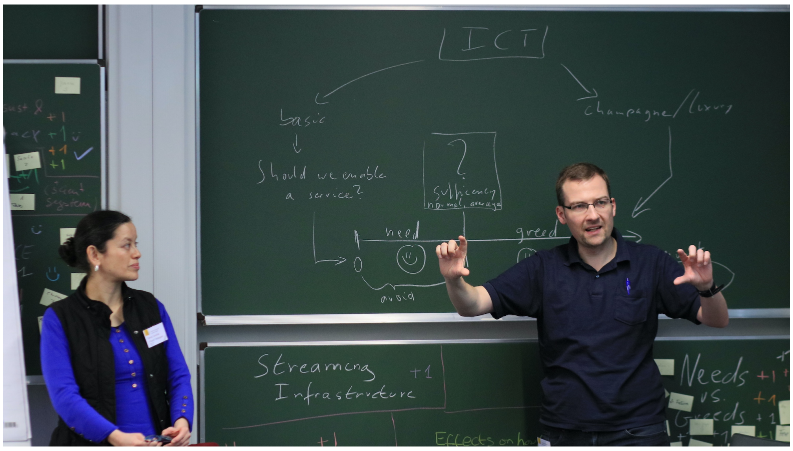


Figure 7 Presentation of group work (Group 4).

Figure 8 Sustainable social outing – walk in the surroundings.

Figure 9 Refreshing Kneipp experience.

Figure 10 Preparation of a music session.

Figure 11 In the wine cellar.

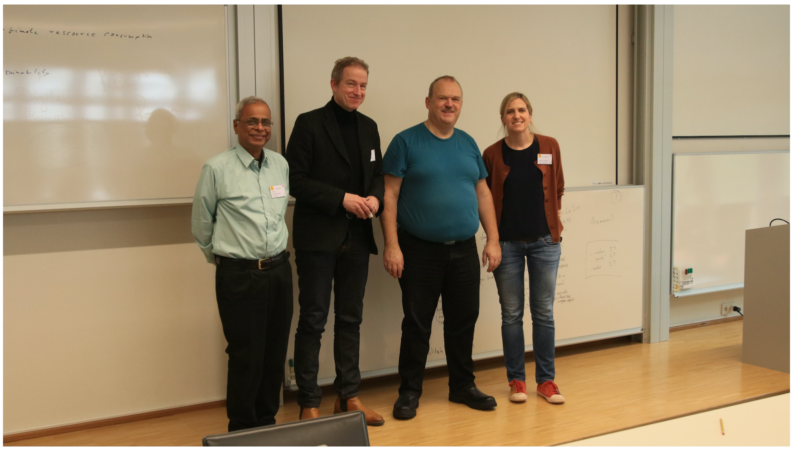


Figure 12 Co-organisers (from left to right: Ashok, Alex, Markus and Katrien).

7 Final reflections

7.1 Main outcomes

Katrien De Moor (NTNU – Trondheim, NO), Markus Fiedler (Blekinge Institute of Technology – Karlshamn, SE), Ashok Jhunjhunwala (IITM Research Park – Madras, IN), and Alexander Raake (TU Ilmenau, DE)

This seminar approached the topic of sustainable experiences and the quality of sustainable experience from different angles and disciplinary perspectives, both in terms of current understanding and knowledge, tools and methods and challenges for future research. The main aim was to bring together a diverse set of participants in order to foster new alliances, inspire, trigger scientific renewal and to explore and map future opportunities and research challenges. The bottom-up methodology that was followed resulted in many ideas, as described in Section 5. We may summarize the main outcomes from the various group works as follows:

1. Openings for innovative sustainability-relevant services;
2. Conceptual model for QoSE measurement framework (cf. Figure 1);
3. An up-to 100-years sustainability perspective on ICT and related circumstances;
4. Concept and modeling of sufficiency.

Beyond “Quality of”, there is a need for a wider take and a longer time horizon on “Sustainable Experience”, reflected in alternative notions such as SDE (Sustainable Digital Experiences) or SUE (Sustainable User Experience).

7.2 Next steps

Katrien De Moor (NTNU – Trondheim, NO), Markus Fiedler (Blekinge Institute of Technology – Karlshamn, SE), Ashok Jhunjhunwala (IITM Research Park – Madras, IN), and Alexander Raake (TU Ilmenau, DE)

To take these ideas further, a number of next steps have already been taken, including a special session on "Towards the design and evaluation of Sustainable Multimedia Experiences" at the International Conference on Quality of Multimedia Experience (QoMEX 2023). QoMEX 2023 has "Towards Sustainable and Inclusive Multimedia Experiences" as special focus and several of the Dagstuhl seminar organizers are also involved in the organization of the conference. In addition, one of the Dagstuhl participants, Dr. Daniel Schien, has been invited to hold a keynote speech at the conference. In addition, Dagstuhl co-organizer Prof. Markus Fiedler is also co-chair of the "Workshop on sustainability and QoE Management", co-located with QoMEX 2023. Further, a video trailer has been compiled, based on recordings made during the Dagstuhl seminar. This video is available on YouTube [1] and will be used for disseminating around the focus of the seminar. In addition, joint journal and conference publications ideas were discussed and are under work. Further, there are plans for a COST Action, which even bridges to the Dagstuhl Perspectives Workshop 23092 and which could offer an excellent vehicle to continue the discussions, do community-building and to join forces on a global stage.

References

1 Dagstuhl Seminar 23041: Quality of Sustainable Experience (QoSE): short video trailer. `https://youtu.be/D2vswi_807A`. Accessed: 2023-05-29.

Participants

- Stepanie Arevalo Arboleda
TU Ilmenau, DE
- Emma Beauxis-Aussalet
VU University Amsterdam, NL
- Michael Best
Georgia Institute of Technology – Atlanta, US
- Pablo Cesar
CWI – Amsterdam, NL
- Katrien De Moor
NTNU – Trondheim, NO
- Yvonne Dittrich
IT University of Copenhagen, DK
- Markus Fiedler
Blekinge Institute of Technology – Karlshamn, SE
- Christian Herglotz
Universität Erlangen-Nürnberg, DE
- Oliver Hohlfeld
Universität Kassel, DE
- Tobias Hoßfeld
Universität Würzburg, DE
- Lucjan Janowski
AGH – Univ. of Science and Technology – Krakow, PL
- Ashok Jhunjhunwala
IITM Research Park – Madras, IN
- Effie Lai-Chong Law
Durham University, GB
- Dan Lockton
TU Eindhoven, NL
- Colin Maclay
USC – Los Angeles, US
- Alexander Raake
TU Ilmenau, DE
- Peter Reichl
Universität Wien, AT
- Reema Saha
IITM Research Park – Madras, IN
- Daniel Schien
University of Bristol, GB
- Fatuma Simba
University of Dar es Salaam, TZ
- Sascha Spors
Universität Rostock, DE
- Fee Steinhoff
Hochschule Koblenz – Remagen, DE
- Kaisa Väänänen
University of Tampere, FI
- Denny Carolina Villamil Velasquez
Blekinge Institute of Technology – Karlskrona, SE
- Irene Viola
CWI – Amsterdam, NL
- Hans-Jürgen Zepernick
Blekinge Institute of Technology – Karlskrona, SE
- Thomas Zinner
NTNU – Trondheim, NO

Report from Dagstuhl Seminar 23051

Perception in Network Visualization

Karsten Klein*[1], Stephen Kobourov*[2], Bernice E. Rogowitz*[3], Danielle Szafir*[4], and Jacob Miller†[5]

1 **Universität Konstanz, DE.** karsten.klein@uni-konstanz.de
2 **University of Arizona – Tucson, US.** kobourov@cs.arizona.edu
3 **Visual Perspectives – New York, US.** bernice.e.rogowitz@gmail.com
4 **University of North Carolina at Chapel Hill, US.** danielle.szafir@cs.unc.edu
5 **University of Arizona – Tucson, US.** jacobmiller1@arizona.edu

Abstract

Networks are used to model and represent data in many application areas from life sciences to social sciences. Visual network analysis is a crucial tool to improve the understanding of data sets and processes over many levels of complexity, such as different semantic, spatial and temporal granularities. While there is a great deal of work on the algorithmic aspects of network visualization and the computational complexity of the underlying problems, the role and limits of human perception are rarely explicitly investigated and taken into account when designing network visualizations. To address this issue, this Dagstuhl Seminar raised awareness in the network visualization community of the need for more extensive theoretical and empirical understanding of how people perceive and make sense of network visualizations and the significant potential for improving current solutions when perception-based strategies are employed. Likewise, the seminar increased awareness in the perception community that challenges in network research can drive new questions for perception research, for example, in identifying features and patterns in large, often time-varying networks. We brought together researchers from several different communities to initiate a dialogue, foster exchange, discuss the state of the art at this intersection and within the respective fields, identify promising research questions and directions, and start working on selected problems.

Seminar January 29 – February 5, 2023 – https://www.dagstuhl.de/23051
2012 ACM Subject Classification Theory of computation → Data structures and algorithms for data management; Human-centered computing → Human computer interaction (HCI)
Keywords and phrases Network Visualization, Graph Drawing, Perception, Cognition
Digital Object Identifier 10.4230/DagRep.13.1.216

1 Executive Summary

Karsten Klein
Stephen Kobourov
Bernice E. Rogowitz
Danielle Szafir

The Dagstuhl Seminar "Perception in Network Visualization" addressed the issue that both established knowledge and current research on human perception are not represented well in network visualization research, and in particular not explicitly taken into account in the development of methods and measures for effective network representation.

[*] Editor / Organizer
[†] Editorial Assistant / Collector

Perception in Network Visualization, *Dagstuhl Reports*, Vol. 13, Issue 1, pp. 216–244
Editors: Karsten Klein, Stephen Kobourov, Bernice E. Rogowitz, Danielle Szafir, and Jacob Miller
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

A main goal of the seminar thus was to investigate the foundations of network visualization in the context of human perception and cognition. This included raising awareness in the network visualization community about potential gaps in the current state of the art, identifying specific research questions to fill these gaps, investigating the selected questions, and creating an agenda for future research. Similarly, we wanted to increase awareness in the perception community that challenges in network visualization research can drive new questions for perception research. An important purpose of the seminar was to engage network visualization researchers to increase the efforts to take into account knowledge on perception – its limits but also opportunities – when designing and evaluating network visualization approaches. The mechanisms and impact of specific perceptual phenomena are currently underexplored in network visualization research, and we wanted to put the investigation of these topics more prominently on the research agenda. To this end, the seminar initiated exchange between researchers in the network visualization community and researchers studying perception.

Perception can play an important role in nearly all aspects of network visualization. We aimed to cover diverse aspects in the topics investigated during the seminar, with the following short list serving as a starting point for further discussions:

- *Fundamentals of perception in relation to network visualization:* Basic questions about how humans read network visualizations in the context of specific network characteristics and tasks are not yet well understood. We would like to investigate some of these questions, including: What are main features that humans recognize and memorize from different network representations? How well can these features be distinguished? How sensitive are people to changes in these features? What are the main features that support orientation and navigation in large networks? What are the relationships between insight generation, perception and interaction in interactive exploration scenarios?
- *Quality metrics and layout styles:* Many quality metrics and optimization goals for different layout styles have been proposed (e.g., number of crossings, stress, number of bends). We want to investigate whether these metrics and goals are motivated or justified by modern theories of perception and align these metrics with relevant empirical evidence. Can the current knowledge on perception explain why certain approaches work better than others?
- *Experimental design:* Investigating the above questions requires new experimental paradigms that consider the complex relationship between elements in network visualizations (e.g., nodes and edges) and the insights that people develop with such visualizations. Experimental methods must both investigate perceptual aspects of network visualization and provide meaningful evaluations of new metrics and approaches.
- *Guidelines:* Network visualization covers more than algorithmic aspects, such as choosing different channels to represent data visually. Is it possible to develop guidelines that help steer the complex design process using perceptual principles?

Acknowledgments

We would like to thank all participants of the seminar for their contributions and lively discussions, and the scientific directorate of Dagstuhl for providing us with the opportunity to organize this seminar. Finally, the seminar would not have been possible without the amazing support by the staff of Dagstuhl.

2 Table of Contents

Executive Summary

Overview of Talks

Working Groups

3 Overview of Talks

While most of the week at Dagstuhl was spent in smaller working groups, on Monday we had two overview talks about perception and network visualization, and two overview talks about graphs and graph drawing. The rest of the week included several lightening talks on topics requested by the participants. Abstracts of all these talks follow.

3.1 Visual Perception, Visualization, and Network Visualization

Cindy Xiong (University of Massachusetts Amherst, US, cindy.xiong@cs.umass.edu), Danielle Szafir (University of North Carolina at Chapel Hill, US, danielle.szafir@cs.unc.edu)

Visualization has a long tradition of drawing on insights from human perception to inform effective design. In this talk, we review basic perceptual phenomena, including visual attention, visual search, and grouping, to highlight their basic mechanics and how they have been applied in past visualization design guidelines and experiments. We connect each phenomenon to past studies in network visualization to highlight key potential crossover between the fields and scaffold workshop discussion.

3.2 Introduction to Perception: the "Food Chain"

Bernice Rogowitz (Visual Perspectives – New York, US, bernice.e.rogowitz@gmail.com)

The term "perception" is used broadly in computer science, and can refer to a broad range of human behaviors. This talk painted an overview of the many, often parallel, processes involved when we interact with network representations, and with the world in general. Low level visual perception focuses on retinal processes such as sensing luminance and color variations, and the impact of differences in foveal and peripheral resolution. Early cortical processes provide binocular vision and enable low-level feature perception. The organization of these features into a unified whole is managed by processes of perceptual organization and attention. And cognitive processes imbue them with semantic meaning, enable memory, and support decision making. Emotion and aesthetic perception, often shaped by culture and experience, also contribute to our visual experience. Moreover, vision does not operate in a vacuum. We are simultaneously hearing, smelling, touching and moving through our world, guided by our intentions, tasks and desires. This is a "food chain," in the sense that projections feed from sensors to cortex to higher centers, often called "bottom -up," but it is really more of a network, with important "top down" feedback and modulation.

In this presentation, we reviewed key topics in early vision. The human visual system is designed to register variations in sensory stimuli. The absolute luminance is less important perceptually than the contrast, the difference between the highest and lowest luminance values, and our sensitivity to contrast depends on the way those luminance variations play out over space. In network visualization, there needs to be sufficient luminance contrast to make out nodes, edges, arrows and annotations, and the finer the spatial detail, the higher the required luminance contrast. Even for colored visualization features, like yellow or red edges on a gray background, legibility depends on luminance contrast.

Moving up the food chain, we can ask how visual information is organized perceptually. In the 1920, researchers from the Berlin School of Experimental Psychology developed a paradigm-shifting approach to studying perception. They focused not on bottom-up constructionist ideas of perception, but instead introduced the idea that top-down processes organized individual elements into Gestalts. The visual system actively constructs visual impression, causing us to perceive sets of elements as wholes. Principles such as proximity, continuity, symmetry and closure can be seen working when we extract perceive structures embedded in graphs. For small graphs, it may be easy to identify clusters or pick out embedded structures. But, how robust are these organizational forces when graphs become very large?

Humans are not passive recipients of visual information. We active move our bodies and our eyes to make sure that we can register important features. Some of these processes are bottom up. An object that has a different color from its surroundings or a different movement or orientation will attract our attention. Some types of low-level patterns are perceived instantly, no matter how many objects there are in the background. Others take time to suss out, and the more objects, the longer it takes to scrutinize the field to find them. In visualization, we can use these bottom-up cues to attract attention to features of interest. For example, marking a critical edge red will draw our attention to it automatically, marking all the nodes belonging to a class red will automatically group them together perceptually, even if they are not near each other spatially. There are so many visual cues bombarding our senses all the time that our perceptual system need mechanisms to segregate them into categories, and we can make use of these capabilities in network visualization to, for example, segregate sets of nodes by assigning them to a common color. We can even use hue to interactively "paint" a set of nodes in one network, and if that color is "brushed" onto corresponding nodes in another network, we see the correspondence immediately.

As we move up the food chain, we are struck by the powerful forces of top-down perception. As we said, we direct not only to low-level features that attract our attention, but to those objects that will give us information about the world. What we are trying to learn about the visual scene guides our gaze and our attention. This is especially important for high-level tasks like pattern recognition and decision-making. The layout of the graph will afford different types of visual observations, and different tasks will drive how we explore a visualization visually. Creating visualizations, thus, requires thinking about the intended audience, the task, and the multitude ways of representing data and relationships.

As we move up the food chain, individual differences play an increasing part in perception. The ability to detect and identify luminance, color, spatial, and movement is similar for everyone. However, where and how you look at a visualization, and what meaning you extract, depends on your training and experience. Some judgments, like naming colors, can depend on your cultural and linguistic background. The ability to perceive hidden shapes in a complex environment may not only reflect your spatial intelligence, but may even be tied to your personality. And at the top of the food chain, aesthetic judgments vary wildly from person to person, encompassing emotional and societal factors.

In network visualization, thus, we are not simply mapping data and relationships onto visual marks. These renderings are processed by the same mechanisms that have evolved to help us perceive and act in all the environments we encounter. Understanding how these mechanisms work, independently and together, can help guide the design of visualizations, and studying how human observers perceive and explore different visual metaphors can, likewise, help advance our understanding of visual perception.

3.3 Graphs and Their Visualizations

Giuseppe Liotta (University of Perugia, Italy, giuseppe.liotta@unipg.it)

Most datasets are relational in nature and can be conveniently modeled as graphs. Graph visualizations are a useful tool to extract knowledge from relational datasets. The talk briefly introduces some of the most common visualization metaphors and interaction paradigms. It also considers different approaches to identify the readability requirements for an effective graph visualization. Finally the talk proposes some research directions at the intersection of network visualization and perception, including multisensorial interaction user studies for non-planar networks, and experimental comparisons of different interaction paradigms.

3.4 Graph Drawing 101

Peter Eades (University of Sydney, Australia, peter.eades@sydney.edu.au)

The quality of a graph drawing can be measured in terms of its (1) faithfulness and (2) readability. There have been three kinds of algorithms proposed and deployed for graph drawing – planarity-based methods, force-directed methods, and layered drawing. The performance of these methods against faithfulness and readability requirements varies, especially with respect to scale – performance on large graphs differs from performance on small graphs. For large graphs, finding visual proofs of assertions seems to be a promising research direction.

3.5 Don't Trust the Object

Claus-Christian Cardon (Universität Bamberg, Germany, ccc@uni-bamberg.de)

In our daily life, to perceive and correctly recognize objects is key. We have to quickly process food, have to decide whether it's good or bad, healthy or lethal. We have to tell partners, friends and enemies apart. But looking behind the scenes of everyday perception we have to realize that the percept of an object is a mental construction. We cannot perceive the object, our cognitive apparatus makes believe we perceive a certain very determined object. Consequently we have to understand perception to understand how we perceive, asses and understand the object. Perception-based analysis will also help to understand emotional and personal reactions triggered by objects.

3.6 Embedding Neighborhoods Simultaneously by t-Distributed Stochastic Neighborhood Embedding (ENS-t-SNE)

Jacob Miller (University of Arizona – Tucson, US, jacobmiller1@arizona.edu)

When visualizing a high-dimensional dataset, dimension reduction techniques are commonly employed which provide a single 2 dimensional view of the data. We describe ENS-t-SNE: an algorithm for Embedding Neighborhoods Simultaneously that generalizes the t-Stochastic Neighborhood Embedding approach. By using different viewpoints in ENS-t-SNE's 3D embedding, one can visualize different types of clusters within the same high-dimensional dataset. This enables the viewer to see and keep track of the different types of clusters, which is harder to do when providing multiple 2D embeddings, where corresponding points cannot be easily identified. We illustrate the utility of ENS-t-SNE with real-world applications and provide an extensive quantitative evaluation with datasets of different types and sizes.

3.7 Perception as an Educated Guess

Alexander Pastukhov (Universität Bamberg, Germany, Alexander.Pastukhov@uni-bamberg.de)

Our perception feels effortless and accurate, yet it relies on inputs that are intrinsically ambiguous and not faithful. The only way to solve an unsolvable problem is by using prior knowledge about statistical regularities of the world. This reverses the way inference is processed, which becomes an educated guess, guided by sensory evidence. Rules of reverse perception apply particularly for graph visualization, where patterns and context effects strongly override local features and optimalities.

3.8 Perception of Graph Sampling

Daniel Archambault (Swansea University, UK, D.W.Archambault@swansea.ac.uk)

In this talk I present perceptual factors that lead to a graph sample being representative of the original graph. Factors such as coverage, cluster quality, and high degree nodes were determined to be of importance here.

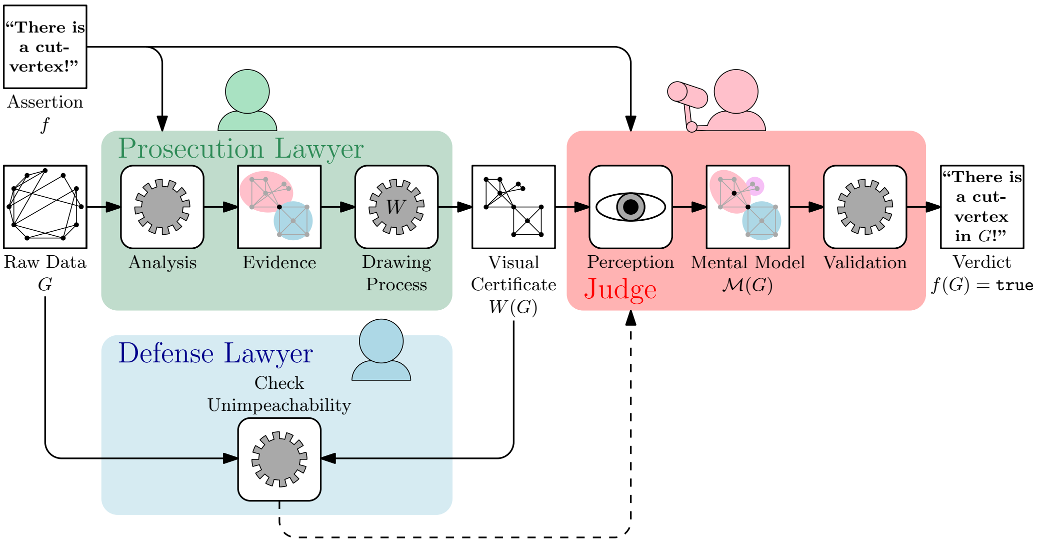


Figure 1 Our concept for visually proving an assertion about a given graph. The prover (prosecution lawyer) shows evidence and has to convince a third party (judge). Another party (defense lawyer) may interfere if the proof is not entirely trustworthy.

4 Working Groups

4.1 Visual Proofs of Network Properties

Tim Dwyer, Peter Eades, Henry Förster, Seok-Hee Hong, Felix Klesen, Stephen Kobourov, Giuseppe Liotta, Kazuo Misue, Fabrizio Montecchiani, Alexander Pastukhov, Falk Schreiber

Working group members: Tim Dwyer, Peter Eades, Henry Förster, Seok-Hee Hong, Felix Klesen, Stephen Kobourov, Giuseppe Liotta, Kazuo Misue, Fabrizio Montecchiani, Alexander Pastukhov, and Falk Schreiber

The working group discussed scenarios where it is necessary to convince an audience that a particular graph has some structural property. We assume that a prover already knows that the assertion is true. A visual proof is a visual representation used to convince a third party who does not know the proof. Visual proofs are commonly used in different areas of science to show the truth of a statement or to support an argument. Figure 1 provides on overview of our concept for visually proving an assertion about a given graph. The prosecution lawyer (prover) shows evidence for the assertion in a visual certificate drawing of the graph. In order to convince the judge (third party who does not know the proof), the visual proof must guide the judge's perception to form a mental model that makes the assertion easy to understand and validate. In addition, the visual proof needs to be entirely trustworthy as otherwise a defense lawyer may raise doubts about the validity of the prosecution lawyer's claims.

Visual proofs have some key differences to the standard motivations for network visualization. While typical network visualization approaches often seek a representation which shows as many graph properties as possible simultaneously, a visual proof may focus on showing optimally just one specific property. Also, common aesthetic criteria for network layout may not hold for specific visual proofs. A simple examples of a visual proof is proving that there

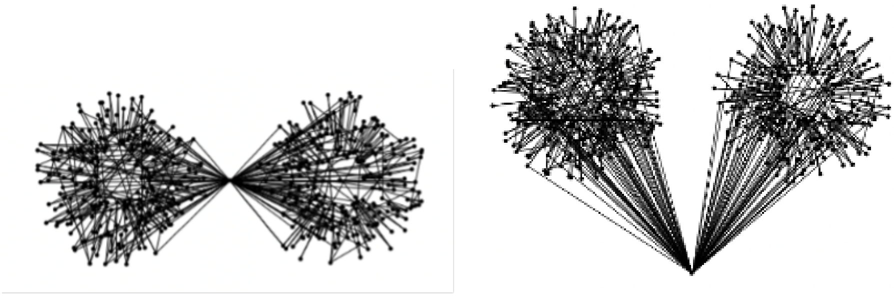


Figure 2 A simple visual proof that the graph is not biconnected. The right drawing is better as it is clear that no edge bypasses the cut vertex (straight line drawing).

is a cut vertex (i. e. that the graph is not biconnected). This is trivial: put the cut vertex in the middle and show that after removing there are two or more components. Figure 2 shows visual proofs for this property, the right drawing is better than the left drawing as it is clear there is no edge bypassing the cut vertex. We found a number of other easy examples initially, however, the problem turned out much more challenging quite quickly.

After initial discussions about the idea of visual proofs for network properties, the working group focused on two major areas: developing and formalizing the general concept of visual proofs for graph properties (see also Figure 1), and discussing and structuring examples for properties and related proofs. Research questions discussed included

- How can we visually prove properties of graphs?
- What does it mean to visually prove a property?
- What makes a good visual proof?
- What is a good formalization?
- What layout (or other graphical representation) is optimal with respect to proving a particular property?
- What is the relationship between visual complexity and complexity theory?
- What classes of properties are visually provable?
- When is the opposite of a property easy, when difficult to prove visually?

The working group continued after the Dagstuhl Seminar to further develop ideas, concept and examples, and recently submitted a paper describing the concept and applications of visual proofs of network properties.

4.2 Mapping Perception Mechanisms to Analytical Tasks in Network Visualization

Carsten Görg, Cindy Xiong, Danielle Szafir, Paul Rosen

Working group members: Carsten Görg, Cindy Xiong, Danielle Szafir, and Paul Rosen

4.2.1 Abstract

Network visualization is utilized in a variety of domains to analyze data, e.g., social network analysis, biological pathways, computer network analysis, etc. Researchers have designed a range of network representations to support data exploration and analysis, focusing on concepts such as faithfulness, a measurement of how well the visualization matches the data, and a number of heuristics, such as reducing edge crossings. However, visualization design can impact what people perceive and how, and therefore it is equally important to increase the readability of a network visualization by leveraging the way people see the world to help people optimally make sense of their data. Many other visualization types, e.g., scatterplots and bar charts, have been studied so people can design and optimize for underlying perceptual mechanisms to combat limitations in human perception and cognition.

While there exists some perceptually-oriented work that looked at specific aspects of design aesthetics, such as edge crossing and symmetry, and identified perceptual features that can harm or enhance the readability or memorability of network visualizations, most network visualizations lack the formal analysis regarding their perceptual features required to achieve this goal. As a result, most designs remain to largely rely on intuition, heuristics, and the outcomes of algorithmic processes optimized over a range of mathematical and visual parameters. We posit that network visualization design can and should be optimized based on perceptual principles.

There exist trade-offs between various perceptual mechanisms when people interpret network visualizations. For example, our visual system is optimized for spatial reasoning, while spatial information in a network visualization is not necessarily useful. The tension between these two paradigms can cause subtle biases or misleading features that may lead to poor task performance and undetected biases.

In this project, we focus on perceptual features in node-link diagrams and adjacency matrix visualizations of networks. The goal is to link the relevant literature to identify open questions about how people perform fundamental tasks with network visualizations and how our natural approaches to processing information may subtly mislead.

To do this, we derived an expanded taxonomy from Lee et al., characterized by visual workflows for completing analytic tasks. Example tasks include finding connected nodes, estimating the size of the graph, and follow a given path. We align these workflows with key perceptual tasks that may inform how people achieve these analytic tasks, such as scene perception object recognition, internal representation, perceptual organization, ensemble coding, low-level feature perception, visual search. We offer a connection between on-going research in visualization and perception to inspire future study. We plan to publish this taxonomy with connections to example systems and studies exemplifying core elements as a preliminary journal paper and apply for an NSF CISE Medium from the US National Science Foundation to fund a series of experiments quantifying the relations between analytical tasks, perceptual tasks, and visualization design.

By reconciling these fields, we hope to spur collaboration and innovation in both perception and network science to ultimately improve our abilities to make readily sense of network data.

4.2.2 Key Research Questions

- What are the tasks that people want to do with networks?
- Revisit past task taxonomies to figure out which connect to vision
- How do we map known visualization theory to these tasks?
- What are the gaps/new tasks that need to be understood?
- What do we not know how about people complete these tasks?
- Tasks for visualization (typically for smaller network exploration rather than bigger graphs) vs. Tasks for mathematical exploration

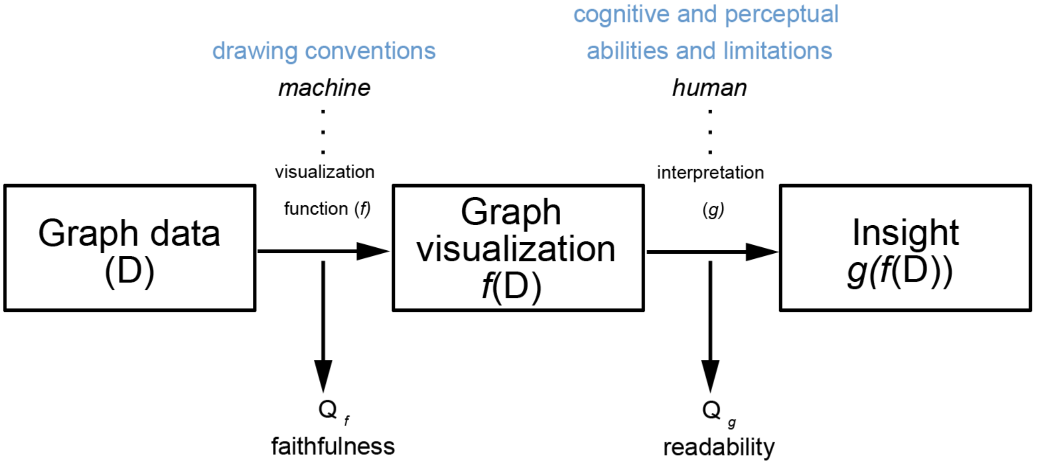


Figure 3 A typical graph visualization pipeline. A drawing function f takes a grapf from a graph dataset D. This produces a graph visualization (f(D)). An interpretation function g takes the graph visualization and produces an insight (g(f(D)).

- How do we walk the balance between control & ecological validity to provide actionable guidance for designers?
- Can we identify methods/experimental design considerations/visual design considerations for designing for network tasks?
- Where do we transition between precise/exact strategies to approximation strategies to "I give up"?
- Where does the spatial optimality of vision help us? Hurt us?
- The tasks may be broken down into more fine-grained components of tasks based on perception, can we find these low-level building blocks? (cognition/perceptual sensemaking → achieve a bigger goal)
- Does feature congestion make these tasks more difficult? E.g., metrics that measure visual clutters (take a look at Rosenholtz).

4.3 Perception-Based Framework for Measuring Quality of Graph Visualizations

Tamara Mchedlidze, Alexandru C. Telea, Marius H. Raab, Christophe Hurter, Natalia Melnik, Martin Nöllenburg, Bernice E. Rogowitz

Working group members: Tamara Mchedlidze, Alexandru C. Telea, Marius H. Raab, Christophe Hurter, Natalia Melnik, Martin Nöllenburg, and Bernice E. Rogowitz

The quality of data visualizations is typically assessed through their conformance to drawing styles and conventions and by measuring quality metrics. Drawing styles and conventions aim to provide a certain quality standard of a graph drawing, but do not explicitly measure it, while quality metrics measure the quality aspect of a graph drawing without an algorithmic way to optimize it. Drawing styles and conventions and quality metrics originating from a variety of perspectives, are complementary tools, but are rarely systematic. Our working group discussed approaches to systematize and unify currently existing metrics and conventions in a single framework for assessing graph visualization quality. During this Dagstuhl Seminar, we discussed the possibilities of integrating the abilities and peculiarities of the human perceptual system (what we here refer to as "perceptual principles") into a unified visualization pipeline (Figure 3).

Perception plays an important role in how humans judge and perceive visual information. One phenomenon that strongly affects one's perception is grouping. Perceptual grouping refers to processes in which discrete elements ("parts") are parsed into groups ("wholes") by the visual system, following so-called Gestalt principles [7]. In a typical example, rows of dots positioned closer together are perceived as grouped together more than the rows of dots that are sparely positioned (i.e., proximity principle). Such grouping is, for instance, one of the targeted perception optimization for graph visualization [22]. However, some similarities between the graph drawing and perceptual grouping can be observed. For example, in graph drawing, in stress model [12], edges are shortened, so that the related nodes are placed closer together and thus appear as grouped together. While there are clear differences between the mechanisms behind Gestalt perception and graph drawings, the outcome product can be seen as analogous: the items appear as grouped.

From this starting point we made progress by analyzing specific drawing styles and conventions and quality metrics, relating them to a list of perception principles. Our initial focus included such Gestalt principles as proximity, symmetry, common fate, closure, similarity, common region, and good continuation, as well as such additional concepts as curvature, visual complexity and global aesthetics, and clutter. Our analysis revealed that drawing styles and conventions and quality metrics often rely on one or another perception principle, at least to some extent. For instance, global shape, bundling, and number of crossings reflect the Gestalt principle of similarity. We acknowledge that some of the metrics will reflect a complex combination of perception principles. However, we believe that they can nevertheless be formally described. After the description, the sample space of all possible variations of the metrics can be measured and new graph layouts can be computed. How to effectively measure the space and design quality metrics from the measured spaces, still remains an open question though. Our connecting of metrics with perceptual principles can help, e.g., people choosing sets of metrics to optimize simultaneously (because they relate to the same principle) or, alternatively, see which tasks will be helped when optimizing certain metrics.

We believe that quality metrics can be also ordered along a hierarchical spectrum (i.e., low, mid, and high), potentially adhering to the hierarchical spectrum of visual perception (low-, mid-, high-level vision). Generic attributes of the drawings, which are typically less data- and less task-specific (to give a few examples: number of bends, or number of crossings) and are easier to quantify and measure automatically, could correspond to low-level perception. Other quality measurements are more data- and task-specific, and can thus be seen as high(er) level. They are typically the ones quantified by user experiments. Interestingly, there is a problem: good low-level metrics values do not imply good high-level metric values, and conversely. For example, a drawing may have few crossings, but a path-following task might still be hard without monotonicity of the paths [15]. The other way round: A drawing might not be symmetric in a strictly mathematical sense, but might appear symmetric due to local regularities [9] or the global shape [8]. While the metrics near the endpoints of the spectrum are relatively easy to identify, how exactly to order the metrics at the mid-level (potentially symmetry, bundling quality, clutter, complexity) is an open question. Additionally, there is a need for ways to measure some higher-level quality properties, e.g., global aesthetics. The challenge is that it is unclear how to reduce these to easily measurable properties, e.g., curvature or shape. The end product could be to be able to derive an aesthetics function a(D, f(D)) from a good set of samples (D, f(D)) of graph drawings D and their corresponding lower-level metrics f(D).

To sum up, our working group proposed the first steps towards the creation of a framework that unifies quality metrics and drawing conventions based on human perception. We suggest that new metrics and drawing conventions can be developed for perceptual mechanisms which are not reflected by the current metrics or conventions. We hope that the connection between the perceptual principles and the metrics can aid in optimization of certain aspects that relate to the same perceptual principle or development of tasks in order to see which ones will be improved when optimizing certain metrics. We ended up the Dagstuhl week with the intention to further develop the framework and publish a paper related to it.

4.4 Spatio-temporal Networks – Visualizing Time-dependent Touristic Route Planning

Annika Bonerath, Claus-Christian Carbon, Silvia Miksch, Maurizio Patrignani, Alessandra Tappini

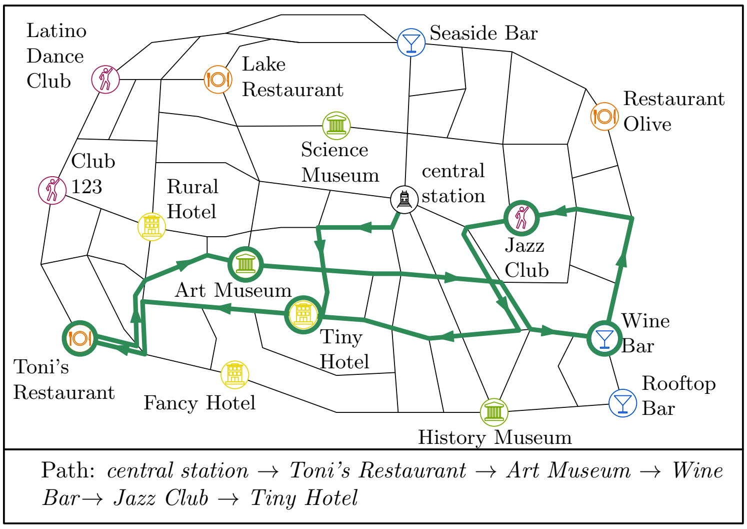
Working group members: Annika Bonerath, Claus-Christian Carbon, Silvia Miksch, Maurizio Patrignani, and Alessandra Tappini

Several user studies were performed to assess the ability of people to use maps, graphs, or other abstract representations of the relationship between objects in a spatial environment. In this report, we describe how a user study can be designed to address the complex scenario when the user has to cope with a network that is dynamic both with respect to nodes (that are “available” or “unavailable” in specific time windows) and with respect to edges (that change their congestion through time).

4.4.1 Introduction

Finding paths in networks is one of the tasks that are more challenging using a matrix-based representation rather than using a node-link representation of the network [13, 20]. The choice of the right path may be complicated by the fact that the network conditions may change over time, because the targets to reach may move from one place to another or because network congestion may discourage the choice of some paths during the day. Still, deciding the sequence of places to visit and the paths that allow us to reach them is an ordinary task in everyday life, and a good or a poor choice reflects on the time spent commuting, on pollution, on personal satisfaction, and on money.

Several user studies are available that address the ability of people of using node-link or matrix-based representations [13, 20]. Also, the domain of dynamic networks has been deeply studied (refer to [2] for a taxonomy of models and solutions). The most complex scenario is when a network is dynamic both with respect to space and with respect to time. This scenario is very rarely explored.



Path: *central station* → *Toni's Restaurant* → *Art Museum* → *Wine Bar*→ *Jazz Club* → *Tiny Hotel*

Figure 4 The users are provided with alternative options from which they should choose the optimal. Here, we display an exemplary solution.

The purpose of our work is to design an experimental study to assess what is best to help the user to cope with a network that is dynamic both with respect to nodes (that are "available" or "unavailable" in specific time windows) and with respect to edges (that change their congestion through time).

This report is organized as follows. Section 4.4.2 describes the application scenario we address. Section 4.4.3 explores related work about network visualization and dynamic network exploration. Section 4.4.4 describes the visualization we would like to adopt in our experiments. Section 4.4.5 describes the experimental design. Finally, Section 4.4.6 is devoted to the next steps and a timeline for our work.

4.4.2 Application Scenario

The application scenario we address is the following. The user starts from a starting point in space and time, i.e., she is located at a specific node of the network at a specific time. Her purpose is to reach a certain number of targets, using the shortest time possible to commute among the targets. The scenario is dynamic: targets have a specific time window when they are available; edges have a traversal time that changes through the day. Our scenario is inspired by the real-world scenario of tourists planning a path through a city that they want to visit; see Figure 4.

4.4.3 Related Work

Network visualization provides meaningful representations of networks/graphs, which are abstract data structures to define as a set of data points and relationships between them. In a recent paper [11], Filipov et al. conducted a survey of surveys to provide researchers and practitioners a "roadmap" elaborating the current research trends in the field of network visualization. They categorize recent surveys and task taxonomies published in the context of network visualization.

Interacting with Networks

In [13] a taxonomy of generic graph-related tasks is described and an evaluation is performed aiming at assessing the readability of two representations of graphs: matrix-based representations and node-link diagrams. The study shows that matrix-based visualizations perform better than node-link diagrams on most tasks when graphs are bigger than twenty vertices. Only path-finding is consistently in favor of node-link diagrams throughout the evaluation.

In [10] it was investigated the usability of Overloaded Orthogonal Drawings against classical Orthogonal Drawings, Hierarchical Drawings, and Matrix-based Representations for performing a collection of basic user tasks on directed graphs. Directed graphs are also the subject of the crowd-sourced user study in [1], where node-link diagrams, adjacency matrices, and bipartite layouts are compared mainly focusing on overview tasks for large instances.

Dynamic Network Visualizations

Surveys about the representations of dynamic networks can be found in [16, 5]. The two main strategies for representing dynamic phenomena on a network are the *time-to-space mapping* and the *time-to-time mapping* [5]. The first strategy encodes the time dimension into some geometric object, i.e., into a space dimension. This kind of representation may be very challenging for some application domains. A common technique, which will be used in our experiments, is that of relying on small multiples, i.e., replicating the representation for different discrete times. The second strategy, maps the time dimension of the dataset into the time of the user, actually showing a dynamic view, where the changes of the dataset through time are animated in a simulated time. This kind of representation is also planned in our experiments. Similar to our setting is the work described by Saraiya et al. [23], where a node-link diagram with static positions is used and only node attributes are time-varying (in our case edge attributes are). Comparing an animated slider solution to an approach with small time-series visualizations inside each node, they observe better performance of participants for the animated approach when only one or two points in time are involved, while the reverse happens when tasks involve more time steps. Archambault and Purchase contrast animation and small multiples techniques for the visualization of dynamically evolving graphs and show that when the stability of the drawing is low and important nodes in the task cannot be highlighted throughout the time series, animation can improve task performance when compared to the use of small multiples [3]. As in the present paper, the tasks in [3] also involve paths, but the purpose there is that of recognizing a given path while the positions of the nodes changes, rather than finding paths with specific properties. Boyandin et al. focus on the qualitative differences between the types of findings users make with animations and small multiples [6]. They show that animation tends to reveal more findings on adjacent time steps while small multiples foster the discovery of patterns lasting over longer periods. Based on the above results [5] concludes that small-multiples approaches seem to be preferable for tasks involving more than two time steps.

Geospatial Network Visualizations

Geospatial network visualizations associate nodes and links with geographic locations either on Earth or other planets [24]. These visualizations are used to show, for example, trade and financial connections between countries and regions [4] or to display flight connections [21].

Schöttler et al. [24] present a systematic review of geospatial network visualization approaches by establishing a design space, which supports designers in building appropriate and effective visualization for this type of networked data. The proposed design space consists of the following dimensions: (i) geographical facet representation, (ii) network representation (for both nodes and edges), (iii) composition (how the topology and geography are combined in the visualization), and (iv) use of interaction. The geographical representation tackles how to encode geospatial information, which ranges from explicit (representations that use a cartographic map), to distorted (representations that use displacement of spatial positions according to some property of the network), to abstract (representations that use encodings not based on map projections). However, geospatial network visualization captures several open challenges, like handling co-located nodes, link density, and uncertainty in geospatial networks.

4.4.4 Visualization Concept/Approach

From previous research, it is well known, that for the task of finding paths in networks, the node-link visualization outperforms other representation techniques such as matrix-based representations [13, 20]. Since we consider a geospatial network, we visualize it on a map where each node is at its spatial location. The traversal time of an edge is encoded by its thickness in the drawing.

In the experiment, we explore the influence of interaction. Especially, for exploration tasks, interactive user interfaces perform better with respect to non-interactive visualizations [18].

We distinguish two levels: the non-interactive case, where we display the traversal times for all commuting time ranges next to each other; and the interactive case, where the user can interactively choose the time range for which we visualize the traversal times. We expect that for simple networks, the static visualization performs better, while for more complex networks, the interactive visualization is more convenient.

4.4.5 Experimental Design

4.4.5.1 Experimental Factors

Our experimental design consists of four fully crossed experimental factors; see Figure 5:

1. *Interactivity* (non-interactive vs. interactive),
2. *NumAlt*: Number of route alternatives (2 vs. 3),
3. *NumSites*: Number of touristic sites to be visited (3 vs. 5), and
4. *NetSize*: Size of the overall network (small vs. large).

We consider a scenario to be less complex than another scenario if NumAlt, NumSites, or NetSize is smaller.

4.4.5.2 Research Hypotheses

For our experiment, we formulate four research hypotheses.

(H1) *For less complex scenarios, the non-interactive visualization leads to (a) shorter response time and (b) higher accuracy.*

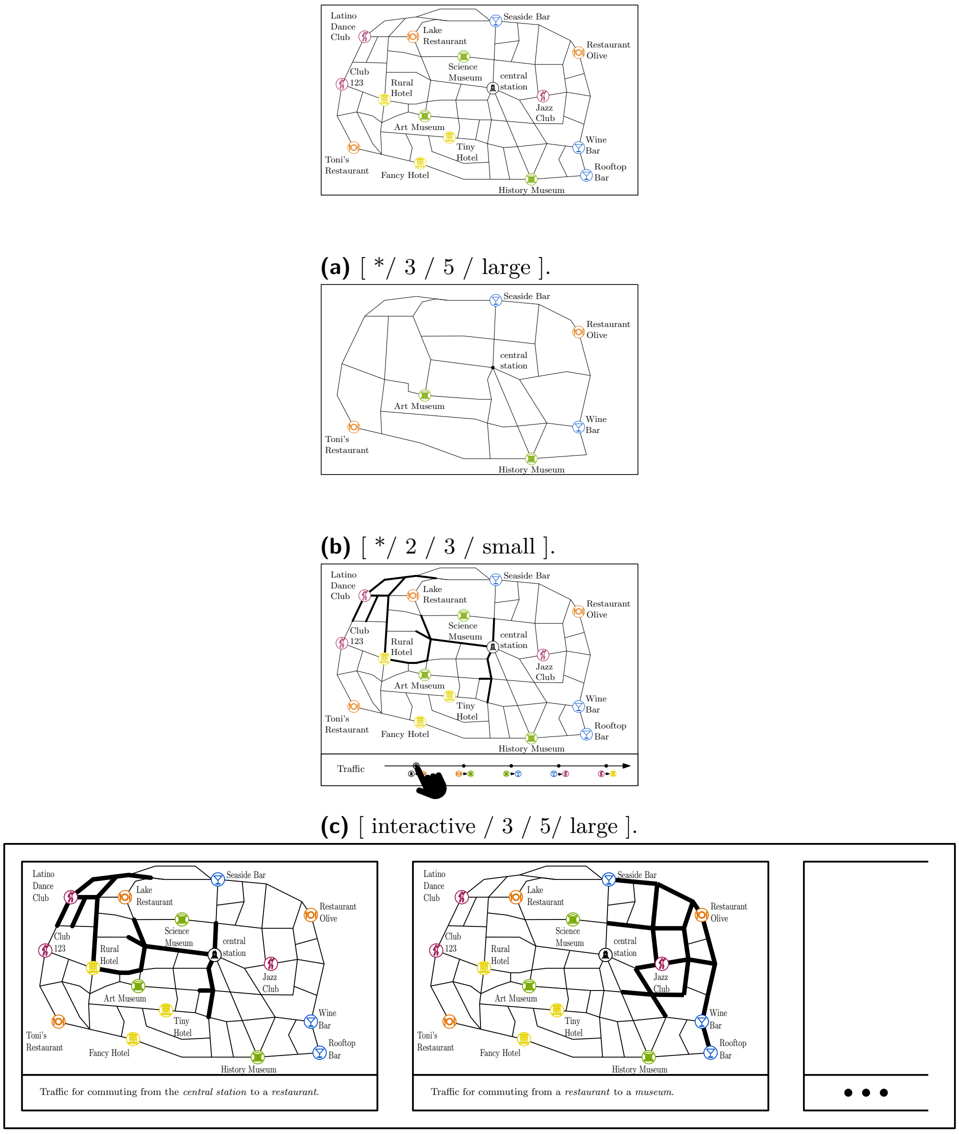
(H2) *For more complex scenarios, the interactive visualization leads to (a) shorter response time and (b) higher accuracy.*

Our intuition is that for a complex scenario, the non-interactive visualization is difficult for the participants since they need to find the sight locations over the different views.

(H3) *For more complex scenarios, the memorability will be higher.*

We base this hypothesis on the fact that users engage more with the data in complex scenarios.

(H4) *The visualization that is perceived as more aesthetically appealing leads to better memorability.*



(a) [*/ 3 / 5 / large].

(b) [*/ 2 / 3 / small].

(c) [interactive / 3 / 5/ large].

(d) [non-interactive/ 3 / 5/ large].

Figure 5 Visualization dimensions [*Interactivity*/ *NumAlt*/ *NumSites*/ *NetSize*]. Edges that are depicted with higher widths indicate more traffic for the displayed time window.

4.4.5.3 Participants

The relevant hypothesis for the power analysis, which determined the sample size, was hypothesis (H1a/b). As we base our analyses on multilevel data analysis using linear mixed models, we will employ R package simR [14] for power calculation. The main statistical model we use is based on a repeated measures design. The effect in question is about the additional fixed effects of NumAlt, NumSites and NetSize on accuracy and was set to $b1 = -5$, $b1 = -2$, and $b1 = -1$, respectively, which represent a mixture of small up to medium effect sizes [19]. To observe that this effect explains a significant amount of variance compared to the null model with $\alpha = 0.05$ and a satisfactory test power 1-β of 0.80, we aim to collect data from $N = 53$ participants. As we expect a drop off of about 20% of the participants including persons who do not respond to the tasks adequately, we will recruit $N = 64$ participants. We aim for recruiting participants without specific knowledge about routing in touristic scenarios but ordinary people that we can recruit online, e.g., volunteer workers recruited by specialist online recruiting companies such as Clickworker or Amazon Mechanical Turk (MTurk). Additionally, we will recruit student participants from the University of Bamberg.

4.4.5.4 Timeline of Experiment

The course of the experiment is as follows:

1. Demographic data of the participants and instruction,
2. Test phase 1 (T1): Exposure to the experimental stimuli one by another in a randomized order; participants are asked to find the most adequate route given certain time constraints,
3. Intermediate phase to distract participants,
4. Test phase 2 (T2): Exposure to the same stimuli from T1 plus the same number of stimuli matched for experimental design properties but being unfamiliar to the participants; the participants have to fulfill the same routing task as in T1 but preceded by a familiarity decision task and followed by ratings about a) aesthetic appeal and impression, b) perceived complexity, c) assessment of the experienced difficulty of solving the respective routing task, and (d) usefulness.

As we estimate the processing time of T1 to be approx. 1 hr and T2 to be 2 hr, we will split T1 and T2 and will operationalize the intermediate phase as a break of about one week – this assists the major aim to distract participants from the main task and to get back full attention at T2. The stimuli are node-link graphs with certain points of interest (POI) which are labeled as certain targets to get routed to. Additional nodes are added to the graph to increase complexity and to emulate typical complexities achieved by real-world touristic city maps (see Figure 4).

For each trial, participants are requested to route to a certain category of POI within a certain time window. The thickness of the links indicates the current traffic situation on the path between two nodes, i.e., two locations, which has an impact on travel times. The major task of the participants is to visit one instance of each category while maximizing the visiting time of all instances in sum.

4.4.5.5 Measures

We evaluate our experiments with quantitative measures: a) the response time, b) the correctness of the answer and c) the memorability in the second phase; and with qualitative measures from the ratings: a) aesthetic appeal and impression, b) perceived complexity, c) assessment of the experienced difficulty of solving the respective routing task, and (d) usefulness.

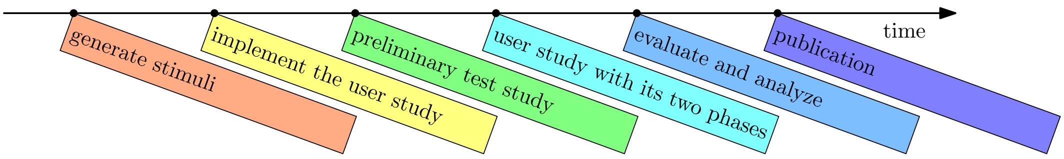


Figure 6 Project timeline.

4.4.6 Outline

We aim with this project at a human-subject study. We want to publish our results at a conference on information visualization such as IEEE VIS or Graph Drawing, or in a journal in the field of spatial cognition or information visualization such as IEEE TVCG. Figure 6 illustrates the timeline of this project.

4.5 Matrix Path Exploration

Carolina Nobre, Daniel Archambault, Rita Borgo, Andreas Kerren

Working Group members: Carolina Nobre, Daniel Archambault, Rita Borgo, Andreas Kerren

4.5.1 Introduction

Data visualizations have long been used to amplify human cognition and help make sense of the vast amount of data in the world. Research has shown that the visual analysis process itself is not universal. User-adaptive visualizations can adapt to the characteristics and preferences of the user. A recent state-of-the-art survey by Yanez et al. [26] proposes a workflow for user adaptive visualization, including user input, adaptation logic, and visual interventions. However, deeper thoughts about the spectrum of visual adaptations for different visualization types have not been explored. Here we report on the group's work exploring user adaptive visualization in the context of interactive network visualization as shown in Figure 7.

4.5.2 Discussions Over the Week

For the first part of the week, we provided further details around the conceptual framework (see Figure 7). In particular, we looked at visual interventions which applied more generally across populations and others that were more individual differences. We instantiated inputs and user representations in the context of network visualisation. The work also converged on three main research questions:

- What types of user-adaptive approaches can support improved visual analytics for networks?
- How can we assess if the adaptive approaches "work", i.e improve the analysis process? (Cognitive load, graph readability tasks - accuracy and time)
- How can we instantiate the user-adaptive visualizations model for network perception?

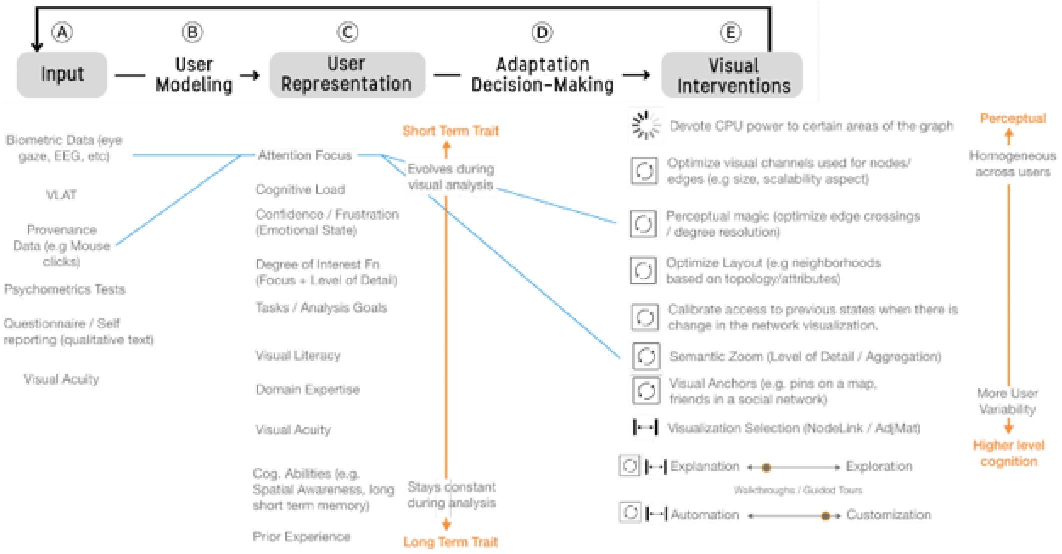

A
B
C
D
E
Input
User Modeling
User Representation
Adaptation Decision-Making
Visual Interventions
Biometric Data (eye gaze, EEG, etc)
VLAT
Provenance Data (e.g Mouse clicks)
Psychometrics Tests
Questionnaire / Self reporting (qualitative text)
Visual Acuity
Attention Focus
Cognitive Load
Confidence / Frustration (Emotional State)
Degree of Interest Fn (Focus + Level of Detail)
Tasks / Analysis Goals
Visual Literacy
Domain Expertise
Visual Acuity
Cog. Abilities (e.g. Spatial Awareness, long short term memory)
Prior Experience
Short Term Trait
Evolves during visual analysis
Stays constant during analysis
Long Term Trait
Devote CPU power to certain areas of the graph
Optimize visual channels used for nodes/ edges (e.g size, scalability aspect)
Perceptual magic (optimize edge crossings / degree resolution)
Optimize Layout (e.g neighborhoods based on topology/attributes)
Calibrate access to previous states when there is change in the network visualization.
Semantic Zoom (Level of Detail / Aggregation)
Visual Anchors (e.g. pins on a map, friends in a social network)
Visualization Selection (NodeLink / AdjMat)
Explanation
Exploration
Walkthroughs / Guided Tours
Automation
Customization
Perceptual
Homogeneous across users
More User Variability
Higher level cognition

Figure 7 Extension of Yanez et al. [26] pipeline (under submission). The two blue polylines represent two concrete realizations of an user-adaptive process starting with a measurement of a user input (e.g., a mouse click), a representation derived from the input (e.g., attention focus) and ending with a visual intervention (e.g., semantic zooming or edge crossing optimization close to the user focus).

During Thursday, all of a sudden an interactive technique for matrix path exploration broke out (Figure 8). The approach is similar to bring & go [17] interactive technique. Instead, rows and columns of the matrix are pulled and placed around an overlay of a focus node. The history of repeated interactions is visualized through a thread that touches all visited edges and nodes through the diagonal.

The group was able to create a draft of the logic behind the algorithm, we however agreed more work was needed to implement the technique and perform stress testing on a wider rage of cases. This matches fully with Dagstuhl seminars vision of seeding new collaborative research.

4.5.3 Conclusions and Next Steps

Based on the Dagstuhl discussions, we plan to write an article on the proposed workflow for user adaptive network visualization that will be submitted to an established journal/magazine in the visualization community, such as Computer Graphics Applications (CG&A). Moreover, we plan to implement the proposed algorithm for matrix path exploration with the help of PhD students and use the implemented approach as the basis for instantiating a user-adaptive visualization process. The final approach should be evaluated and the results published as a conference publication and/or journal article.

As always, Schloss Dagstuhl proved to be an excellent venue to nurture new and exciting research ideas and collaborations. We would like to thank the organizers and the staff for making this a very successful event but also the Dagstuhl staff for providing a friendly and stimulating working environment.

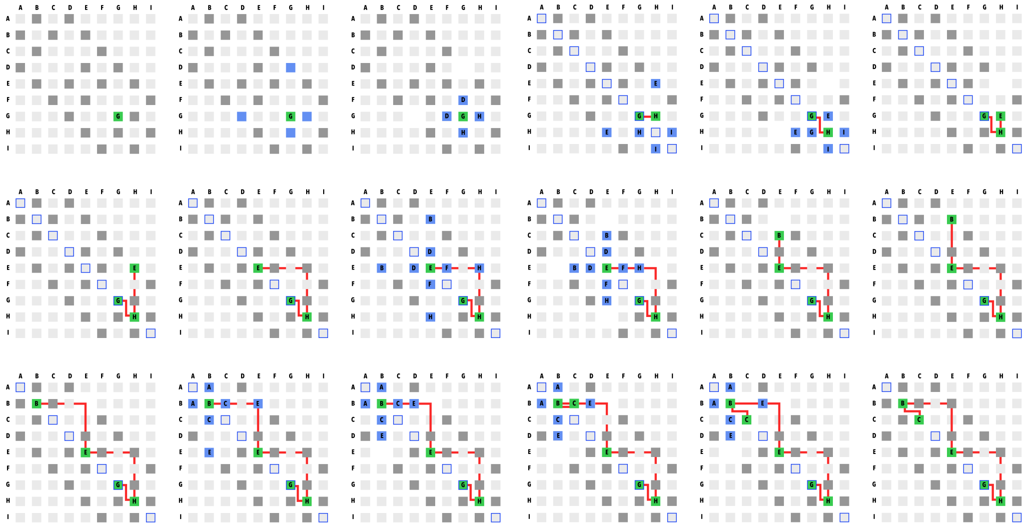


Figure 8 Snap & Go brings the rows and columns around a node in a matrix to interactively explore paths. The overlay shows where a user can go next and the history of the path persists as a line which is drawn through the nodes and edges of the path on the matrix.

4.6 Unintended Perceptual Inferences in Graph Drawing

Michael Aichem, Mohammad Ghoniem, Christophe Hurter, Karsten Klein, Oliver Kohlbacher, Mauro Martino, Jacob Miller, Helen C. Purchase, Bernice Rogowitz, Markus Wallinger, Hsiang-Yun Wu

Working group members: Michael Aichem, Mohammad Ghoniem, Christophe Hurter, Karsten Klein, Oliver Kohlbacher, Mauro Martino, Jacob Miller, Helen C. Purchase, Bernice Rogowitz, Markus Wallinger, and Hsiang-Yun Wu

Discussion. Layout algorithms are often designed to support different analysis functions (e.g., to enhance the salience of clusters or to reveal specific features in the data). For example, lin-log was created in order to create better spatial spread in the 2-D plane, to help reveal clusters that might be difficult to observe when a layout is locally dense. Selecting an algorithm or the parameterization of a layout algorithm to achieve particular goals, however, is often a dark art. Algorithm designers and users rely on intuitions and experience to make these design decisions. Despite their best efforts, many layouts induce incorrect inferences about the true structure of the data. The goal of this research is to identify and characterize classes of incorrect inference, and to provide experimental research with human observers, which can guide more perceptually-faithful renderings.

In particular, we are focusing on layouts that produce misleading results because they interfere with Gestalt Principles of Organization. In a graph layout, human observers tend to see spatially-proximal points as belonging together ("Proximity"), which helps them perceive individual nodes as belonging to the same cluster. If this principle is disrupted when an

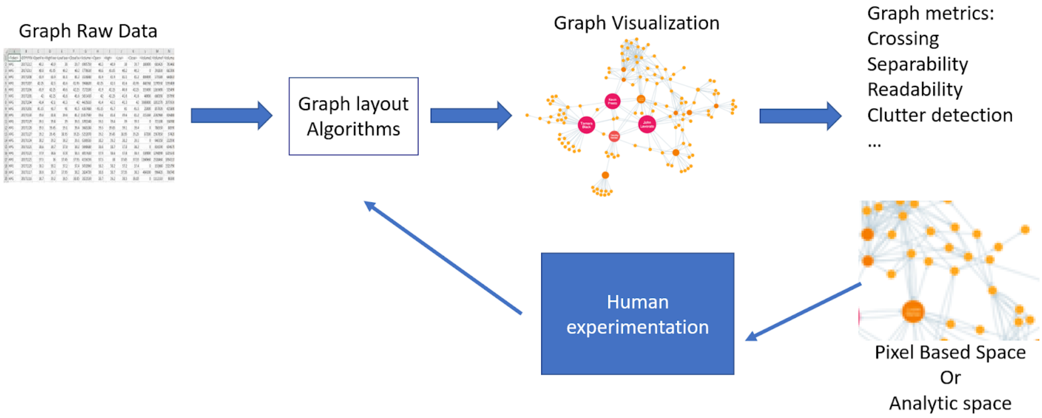


Figure 9 Overview of the framework. Graph drawing metrics and pixel-based metrics are computed, and parameters of the graph layout algorithms are shaped and tuned by experiments with human observers.

algorithm draws unconnected nodes (e.g., no common edges) in the same proximity, this may lead to an incorrect inference about cluster membership. These Gestalt principles exert a very strong impact on how we organize the spatial world, and operate automatically, without conscious control. Some principles we plan to explore are (1) Proximity, (2) Closure, (3) Grouping, (4) Symmetry, and (5) Good Continuation.

Some research questions:

- What types of incorrect inferences can occur in layouts?
- Do these occur because they interfere with Gestalt Principles?
- How important are these misinterpretations to understanding the structure in the data?
- How prevalent are they?
- How does the degree to which they mislead depend on the layout algorithm (e.g., vanilla force directed vs. lin-log), and their parameters?
- What experiments with human observers can measure how the correct interpretation of the data depends on the layout? And on the degree to which Gestalt principles of organization are abrogated?

4.6.1 Discussion through a General Framework

One avenue for exploration was a general approach to categorizing and exploring unintended inferences in graph layouts. For each layout type (e.g.,node-link force directed or circular), there would be three stages: (1) identifying illustrative exemplars, (2a) creating a database of layout drawings for different algorithms and their parameters, (2b) computing the graph-drawing and image-processing (pixel-based) metrics for cases where unintended interpretations are produced, and (3) conducting experiments with human observers to measure the saliency and strength of different misleading (but data-faithful) renderings. This perceptual data would, in turn, feedback to inform layout algorithms and metrics.

4.6.2 Subgroup 1: Unintended Perceptual Inferences in Node-Link Diagrams

Michael Aichem, Christophe Hurter, Karsten Klein, Oliver Kohlbacher, Mauro Martino, Bernice Rogowitz, Markus Wallinger.

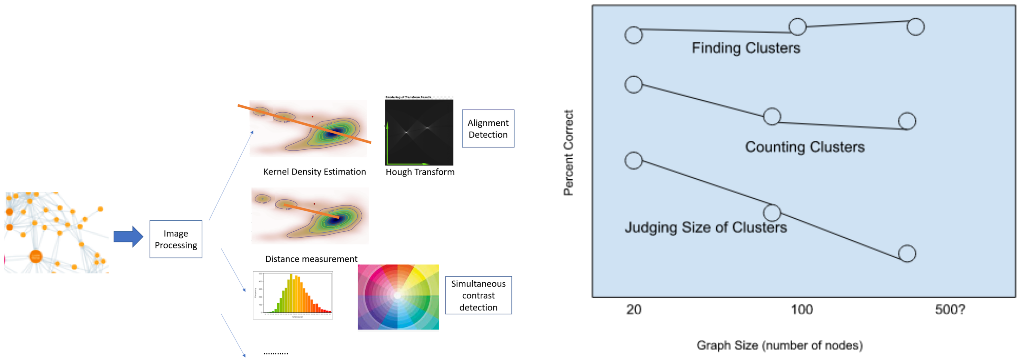


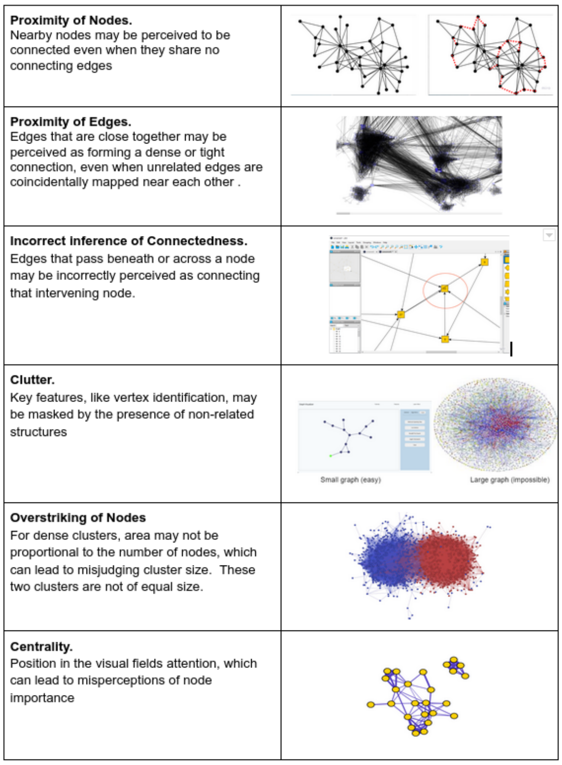
Figure 10 (left): Examples of pixel-based pipeline for generating examples of cases producing misleading perceptual inferences. (right): Hypothetical data from perceptual experiments aim at identify graph-drawing parameters that can create unintended perceptual inferences in graph understanding.

Discussion. Subgroup 1 worked to designing the general framework, with special focus on unintended perceptual inferences arising from node-link diagrams, such as misjudging the size or density of a cluster. This group was motivated by two related issues. First, graph-drawing algorithms are not aware of human perception, and can create layouts that can lead the user to misinterpret the distribution, relationship, and density of nodes and edges. For example, the size of a cluster is related to the number of closely-connected nodes. If non-connected nodes are in the same spatial neighborhood, however, they may be mis-perceived as belonging to that cluster. This unintended inference would have been produced by a Gestalt Principle of Proximity. A layout algorithm could apply a constraint that more aggressively separates non-connected nodes in the spatial layout.

Second, there are many graph-drawing algorithms, with many parameters. Some are designed for a specific purpose, such as lin-log, whose objective is to spread the nodes apart so that internal structures can be more easily appreciated. However, there are no perceptual guidelines for selecting the right algorithm and the right parameterization. One goal of this research is to more closely couple the algorithms with their perceptual effects, so they can be more easily and intuitively selected and tuned. To this end, we have begun designing simple experiments that explore unintended inferences in node-link diagrams produced by different algorithms, and to measure their impact on spatial judgments in graphs of different sizes (small, medium and large). Insight into the size and scope of these mis-perceptions will deepen our understanding of the relationship between algorithm parameters, spatial rendering, and perception, which can provide an objective guide to future algorithm evaluation and development.

Experiments. To provide perceptual feedback to layout algorithms, we would like to examine several tasks that may be prone to unintended lies, as a function of graph size. Hypothetical data are shown below for three tasks, finding clusters, counting clusters, and judging the size of clusters. In this mock-up of experimental results, judging cluster size is the most prone to error, and this error increases with graph size, since there is more opportunity for overstriking to obscure the estimation of the number of nodes involved.

Future Work. Our goal is to continue conceptualizing this approach, which, if successful, will have deep implications for the design of graph layouts and will extend the powerful concepts of perceptual psychology in the graph visualization domain.



Proximity of Nodes. Nearby nodes may be perceived to be connected even when they share no connecting edges	
Proximity of Edges. Edges that are close together may be perceived as forming a dense or tight connection, even when unrelated edges are coincidentally mapped near each other .	
Incorrect Inference of Connectedness. Edges that pass beneath or across a node may be incorrectly perceived as connecting that intervening node.	
Clutter. Key features, like vertex identification, may be masked by the presence of non-related structures	Small graph (easy) Large graph (impossible)
Overstriking of Nodes For dense clusters, area may not be proportional to the number of nodes, which can lead to misjudging cluster size. These two clusters are not of equal size.	
Centrality. Position in the visual fields attention, which can lead to misperceptions of node importance	

Figure 11 Examples of Unintended Perceptual Interferences.

4.6.3 Subgroup 2: Pattern Matching

Jacob Miller, Mohammad Ghoniem, Helen C. Purchase, Hsiang-Yun Wu

Discussion. We are interested in the phenomenon whereby small sub-graphs within a graph may be perceived to be identical when they are structurally dissimilar, or may be perceived to be dissimilar when they are identical. This primarily relates to the Gestalt principle of similarity, but may also include an element of symmetrical pattern-matching.

We believe that if sub-graphs are identical, they should be depicted identically; if they are nearly-identical, they should be depicted nearly-identically. While we are investigating this matter with abstract graphs, the principle is particularly important in domains where sub-structures hold meaning, and where their identification is important. For example, it may be important to identify all the five-node cliques in a social network, or the six-node cycles in a biological network.

We identified five sub-structures (which we call "motifs"): cliques, stars, double-cliques, bi-cliques, cycles. For each motif, we have defined variations on two dimensions – same or different structure, same or different shape. An example for the star motif is shown in Fig. 12.

We have developed an automatic means of creating larger random graphs which include two variants of a motif; the nodes of these motifs have fixed positions when the graph is laid out (see Fig. 13 for the cycle motif).

We have also created matrix and arc diagrams demonstrating the same same/different shape/structure phenomena for the same motifs. We ended our Dagstuhl week with the intention to conduct an experiment (or experiments) using these stimuli – testing the extent

Table 1 A matrix of possibilities for comparing two motifs.

	Same Structure	Different Structure
Same Shape	This is simple rotation. It will be easy for symmetric motifs like cycles, cliques, but less easy for non-symmetric motifs like bi-cliques or double-cliques.	Pattern matching of the node positions in the same shape will make it easy to see where edges are missing or have been added.
Different Shape	The change in shape will make it hard to recognize that the motifs are the same structure.	The change in shape will make it hard to recognize that the motifs are different in structure.

to which participants view the sub-structures as identical. Our preliminary expectations are expressed in Table 1.

We would like to demonstrate that where the identification of sub-graphs is important, the algorithm used for the layout of the whole graph should ensure that the associated motifs are clearly depicted.

References

1 Moataz Abdelaal, Nathan D. Schiele, Katrin Angerbauer, Kuno Kurzhals, Michael Sedlmair, and Daniel Weiskopf. Comparative evaluation of bipartite, node-link, and matrix-based network representations. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):896–906, 2023.

2 Jae-wook Ahn, Catherine Plaisant, and Ben Shneiderman. A task taxonomy for network evolution analysis. *IEEE transactions on visualization and computer graphics*, 20(3):365–376, 2013.

3 Daniel Archambault and Helen C. Purchase. Can animation support the visualisation of dynamic graphs? *Information Sciences*, 330:495–509, 2016. SI Visual Info Communication.

4 Alessio Arleo, Christos Tsigkanos, Roger A. Leite, Schahram Dustdar, Silvia Miksch, and Johannes Sorger. Visual Exploration of Financial Data with Incremental Domain Knowledge. *Computer Graphics Forum*, 41(2):101–116, 2023.

5 Fabian Beck, Michael Burch, Stephan Diehl, and Daniel Weiskopf. A taxonomy and survey of dynamic graph visualization. *Computer Graphics Forum*, 36(1):133–159, 2017.

6 Ilya Boyandin, Enrico Bertini, and Denis Lalanne. A qualitative study on the exploration of temporal changes in flow maps with animation and small-multiples. *Computer Graphics Forum*, 31(3pt2):1005–1014, 2012.

7 Joseph L. Brooks. Traditional and new principles of perceptual grouping. In *The Oxford Handbook of Perceptual Organization*. Oxford University Press, 08 2015.

8 Claus-Christian Carbon, Tamara Mchedlidze, Marius Hans Raab, and Hannes Wächter. The power of shape: How shape of node-link diagrams impacts aesthetic appreciation and triggers interest. *i-Perception*, 9(5):2041669518796851, 2018. PMID: 30210777.

9 Felice De Luca, Md. Iqbal Hossain, and Stephen Kobourov. Symmetry detection and classification in drawings of graphs. In Daniel Archambault and Csaba D. Tóth, editors, *Graph Drawing and Network Visualization*, pages 499–513, Cham, 2019. Springer International Publishing.

10 Walter Didimo, Fabrizio Montecchiani, Evangelos Pallas, and Ioannis G. Tollis. How to visualize directed graphs: A user study. In *IISA 2014, The 5th International Conference on Information, Intelligence, Systems and Applications*, pages 152–157, 2014.

11 Velitchko Filipov, Alessio Arleo, and Silvia Miksch. Are We There Yet? A Roadmap of Network Visualization from Surveys to Task Taxonomies. *Computer Graphics Forum*, page fothcomming, 2023.

12 Emden R. Gansner, Yehuda Koren, and Stephen North. Graph drawing by stress majorization. In János Pach, editor, *Graph Drawing*, pages 239–250, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

13 M. Ghoniem, J.-D. Fekete, and P. Castagliola. A comparison of the readability of graphs using node-link and matrix-based representations. In *IEEE Symposium on Information Visualization*, pages 17–24, 2004.

14 Peter Green and Catriona J. MacLeod. SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4):493–498, 2016.

15 Weidong Huang, Peter Eades, and Seok-Hee Hong. A graph reading behavior: Geodesic-path tendency. In *2009 IEEE Pacific Visualization Symposium*, pages 137–144, 2009.

16 Othon Michail. An introduction to temporal graphs: An algorithmic perspective. *Internet Mathematics*, 12(4):239–280, 2016.

17 Tomer Moscovich, Fanny Chevalier, Nathalie Henry, Emmanuel Pietriga, and Jean-Daniel Fekete. Topology-aware navigation in large networks. In *Proceedings of the 2009 CHI Conference on Human Factors in Computing Systems*, CHI '09, page 2319–2328, 2009.

18 Tamara Munzner. *Visualization Analysis and Design.* A K Peters/CRC Press Visualization Series. CRC Press, 2015.

19 P. Nieminen. Application of standardized regression coefficient in meta-analysis. *BioMedInformatics*, 2(3):434–458, 2022.

20 Mershack Okoe, Radu Jianu, and Stephen G. Kobourov. Node-link or adjacency matrices: Old question, new insights. *IEEE Trans. Vis. Comput. Graph.*, 25(10):2940–2952, 2019.

21 Peter Rodgers. Chapter 7 - Graph Drawing Techniques for Geographic Visualization. In *Exploring Geovisualization*, pages 143–158. Elsevier, 2005.

22 Amalia Rusu, Andrew J. Fabian, Radu Jianu, and Adrian Rusu. Using the gestalt principle of closure to alleviate the edge crossing problem in graph drawings. In *2011 15th International Conference on Information Visualisation*, pages 488–493, 2011.

23 Purvi Saraiya, P. Lee, and C. North. Visualization of graphs with associated timeseries data. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 225–232, 2005.

24 Sarah Schöttler, Yalong Yang, Hanspeter Pfister, and Benjamin Bach. Visualizing and Interacting with Geospatial Networks: A Survey and Design Space. *Computer Graphics Forum*, 40(6):5–33, 2021.

25 Frank Van Ham and Bernice Rogowitz. Perceptual organization in user-generated graph layouts. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1333–1339, 2008.

26 Fernando Yanez, Alvitta Ottley, Cristina Conati, and Carolina Nobre. The state of the art in user adaptive visualizations. *Computer Graphics Forum (EuroVis 2023)*, Under Submission.

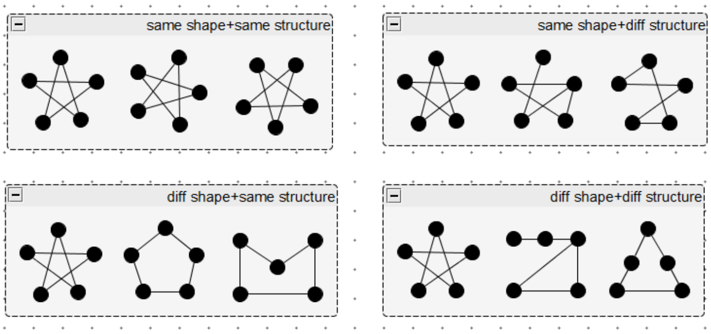

same shape+same structure
same shape+diff structure
diff shape+same structure
diff shape+diff structure

Figure 12 A visual example of how shape and structure effect the perception of a motif.

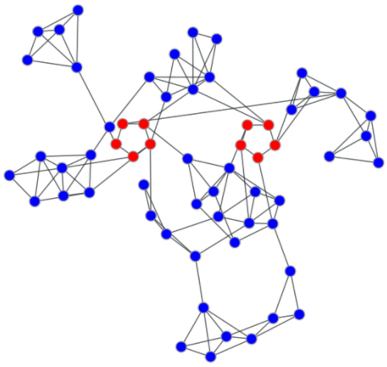


Figure 13 An initial prototype for the graphs we will generate for our experiment.

Participants

- Michael Aichem
Universität Konstanz, DE
- Daniel Archambault
Swansea University, GB
- Annika Bonerath
Universität Bonn, DE
- Rita Borgo
King's College London, GB
- Claus-Christian Carbon
Universität Bamberg, DE
- Tim Dwyer
Monash University – Clayton, AU
- Peter Eades
The University of Sydney, AU
- Henry Förster
Universität Tübingen, DE
- Mohammad Ghoniem
Luxembourg Inst. of Science & Technology, LU
- Carsten Görg
University of Colorado – Aurora, US
- Seok-Hee Hong
The University of Sydney, AU
- Christophe Hurter
ENAC – Toulouse, FR
- Andreas Kerren
Linköping University, SE
- Karsten Klein
Universität Konstanz, DE
- Felix Klesen
Universität Würzburg, DE
- Stephen G. Kobourov
University of Arizona – Tucson, US
- Oliver Kohlbacher
Universität Tübingen, DE
- Giuseppe Liotta
University of Perugia, IT
- Mauro Martino
MIT-IBM Watson AI Lab – Cambridge, US
- Tamara Mchedlidze
Utrecht University, NL
- Natalia Melnik
Otto-von-Guericke-Universität Magdeburg, DE
- Silvia Miksch
TU Wien, AT
- Jacob Miller
University of Arizona – Tucson, US
- Kazuo Misue
University of Tsukuba, JP
- Fabrizio Montecchiani
University of Perugia, IT
- Carolina Nobre
University of Toronto, CA
- Martin Nöllenburg
TU Wien, AT
- Alexander Pastukhov
Universität Bamberg, DE
- Maurizio Patrignani
University of Rome III, IT
- Helen C. Purchase
Monash University – Clayton, AU
- Marius Raab
Universität Bamberg, DE
- Bernice E. Rogowitz
Visual Perspectives – New York, US
- Paul Rosen
University of Utah – Salt Lake City, US
- Falk Schreiber
Universität Konstanz, DE
- Alessandra Tappini
University of Perugia, IT
- Alexandru C. Telea
Utrecht University, NL
- Markus Wallinger
TU Wien, AT
- Hsiang-Yun Wu
FH – St. Pölten , AT
- Cindy Xiong
University of Massachusetts – Amherst, US

Remote Participants

- Danielle Szafir
University of North Carolina at Chapel Hill, US