DAGSTUHL
REPORTS

Volume 8, Issue 10, October 2018

 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
http://www.dagstuhl.de/dagpub/2192-5283

Publication date
April, 2019

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 http://dnb.d-nb.de.

Digital Object Identifier: 10.4230/DagRep.8.10.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
http://www.dagstuhl.de/dagrep

Report from Dagstuhl Seminar 18411

Progressive Data Analysis and Visualization

Edited by

Jean-Daniel Fekete[1], Danyel Fisher[2], Arnab Nandi[3], and Michael Sedlmair[4]

1 **INRIA Saclay – Orsay, FR, jean-daniel.fekete@inria.fr**
2 **Honeycomb – San Francisco, US, danyel@gmail.com**
3 **Ohio State University – Columbus, US, arnab@arnab.org**
4 **Universität Stuttgart, DE, michael.sedlmair@visus.uni-stuttgart.de**

Abstract

We live in an era where data is abundant and growing rapidly; databases storing big data sprawl past memory and computation limits, and across distributed systems. New hardware and software systems have been built to sustain this growth in terms of storage management and predictive computation. However, these infrastructures, while good for data at scale, do not well support exploratory data analysis (EDA) as, for instance, commonly used in Visual Analytics. EDA allows human users to make sense of data with little or no known model on this data and is essential in many application domains, from network security and fraud detection to epidemiology and preventive medicine. Data exploration is done through an iterative loop where analysts interact with data through computations that return results, usually shown with visualizations, which in turn are interacted with by the analyst again. Due to human cognitive constraints, exploration needs highly responsive system response times: at 500 ms, users change their querying behavior; past five or ten seconds, users abandon tasks or lose attention. As datasets grow and computations become more complex, response time suffers. To address this problem, a new computation paradigm has emerged in the last decade under several names: online aggregation in the database community; progressive, incremental, or iterative visualization in other communities. It consists of splitting long computations into a series of approximate results improving with time; in this process, partial or approximate results are then rapidly returned to the user and can be interacted with in a fluent and iterative fashion. With the increasing growth in data, such progressive data analysis approaches will become one of the leading paradigms for data exploration systems, but it also will require major changes in the algorithms, data structures, and visualization tools.

This Dagstuhl Seminar was set out to discuss and address these challenges, by bringing together researchers from the different involved research communities: database, visualization, and machine learning. Thus far, these communities have often been divided by a gap hindering joint efforts in dealing with forthcoming challenges in progressive data analysis and visualization. The seminar gave a platform for these researchers and practitioners to exchange their ideas, experience, and visions, jointly develop strategies to deal with challenges, and create a deeper awareness of the implications of this paradigm shift. The implications are technical, but also human–both perceptual and cognitive–and the seminar provided a holistic view of the problem by gathering specialists from all the communities.

Seminar October 7–12, 2018 – http://www.dagstuhl.de/18411

2012 ACM Subject Classification Human-centered computing → Visualization, Theory of computation → Models of computation, Computing methodologies → Machine learning, Software and its engineering → Software organization and properties

Keywords and phrases Approximate Query Processing, Online Aggregation, Exploratory Data Analysis, Visual Analytics, Progressive Data Analysis, Scalability

Digital Object Identifier 10.4230/DagRep.8.10.1

Edited in cooperation with Sriram Karthik Badam

Progressive Data Analysis and Visualization, *Dagstuhl Reports*, Vol. 8, Issue 10, pp. 1–40
Editors: Jean-Daniel Fekete, Danyel Fisher, Arnab Nandi, and Michael Sedlmair
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Jean-Daniel Fekete (INRIA Saclay – Orsay, FR, jean-daniel.fekete@inria.fr)
Danyel Fisher (Honeycomb – San Francisco, US, danyel@gmail.com)
Michael Sedlmair (Universität Stuttgart, DE, michael.sedlmair@visus.uni-stuttgart.de)

We live in an era where data is abundant and growing rapidly; databases to handle big data are sprawling out past memory and computation limits, and across distributed systems. New hardware and software systems have been built to sustain this growth in terms of storage management and predictive computation. However, these infrastructures, while good for data at scale, do not support data exploration well.

The concept of exploratory data analysis (EDA) was introduced by John Tukey in the 1970's and is now a commonplace in visual analytics. EDA allows users to make sense of data with little or no known model; it is essential in many application domains, from network security and fraud detection to epidemiology and preventive medicine. For most datasets, it is considered a best practice to explore data before beginning to construct formal hypotheses. Data exploration is done through an iterative loop where analysts interact with data through computations that return results, usually shown with visualizations. Analysts reacting to these results and visualizations issue new commands triggering new computations until they answer their questions.

However, due to human cognitive constraints, exploration needs highly responsive system response times (see https://www.nngroup.com/articles/powers-of-10-time-scales-in-ux/): at 500 ms, users change their querying behavior; past five or ten seconds, users abandon tasks or lose attention. As datasets grow and computations become more complex, response time suffers. To address this problem, a new computation paradigm has emerged in the last decade under several names: online aggregation in the database community; progressive, incremental, or iterative visualization in other communities. It consists of splitting long computations into a series of approximate results improving with time; the results are then returned at a controlled pace.

This paradigm addresses scalability problems, as analysts can keep their attention on the results of long analyses as they arrive progressively. Initial research has shown promising results in progressive analysis for both big database queries and for machine learning.

The widespread use of progressive data analysis has been hampered by a chicken-and-egg problem: data visualization researchers do not have online database systems to work against, and database researchers do not have tools that will display the results of their work. As a result, progressive visualization systems are based on simulated systems or early prototypes. In many cases, neither side has currently incentive, skills, or resources to build the components needed.

Recently, data analysis researchers and practitioners have started conversations with their colleagues involved in the data analysis pipeline to combine their efforts. This standard pipeline includes the following core communities: data management, statistics and machine learning and interactive visualization. These initial conversations have led to fruitful evolutions of systems, combining two or three of these communities to complete a pipeline. Database and visualization have collaborated to create systems allowing progressive, approximate query results. Machine-learning and visualization have collaborated to create systems combining progressive multidimensional projections with appropriate scalable visualizations, such as Progressive t-SNE. Most current machine learning algorithms are designed to examine the entirety of a dataset. A major contribution of work like Progressive t-SNE is to have a decomposable algorithm that can compute a meaningful partial result, which then can be passed on to a visual interface for fluent exploration. In these few existing collaborations, the researchers are able to work together and find concrete mechanisms by adapting existing systems for these without re-building them from the ground up. A systematic and widespread linkage between the involved communities, however, is still largely absent.

This Dagstuhl seminar brought the researchers and practitioners who have started this software evolutionary process to exchange their ideas, experience, and visions. We are convinced that in the forthcoming years, progressive data analysis will become a leading paradigm for data exploration systems, but will require major changes in the algorithms and data structures in use today. The scientific communities involved need to understand the constraints and possibilities from their colleagues to converge faster, with a deeper awareness of the implications of this paradigm shift. The implications are technical, but also human, both perceptual and cognitive, and the seminar will provide a holistic view of the problem by gathering specialists from all the communities.

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

- defining and formalizing the concept of progressive data analysis,
- addressing fundamental issues for progressive data analysis, such as software architecture, management of uncertainty, and human aspects,
- identifying evaluation methods to assess the quality of progressive systems, and threats to research on the topic,
- examining applications in data science, machine learning, and time-series analysis.

As a major result from the seminar, the following problems have been identified:

1. Implementing fully functional progressive systems will be difficult, since the progressive model is incompatible with most of the existing data analysis stack,
2. The human side of progressive data analysis requires further research to investigate how visualization systems and user interfaces should be adapted to help humans cope with progressiveness,
3. The potentials of progressive data analysis are huge, in particular it would reconcile exploratory data analysis with big data and modern machine learning methods,
4. Yet, there are many threats that should be addressed to bring progressive data analysis and visualization mainstream in research and application domains.

2 Table of Contents

3 Overview of Talks

3.1 Interactive Data Exploration: A Database Perspective

Carsten Binnig (TU Darmstadt, DE)

In recent years interactive data exploration (IDE) has gained much attention in both the database and visualisation community. Many specialised database execution engines have been developed to accommodate for and leverage the inherent workload characteristics of IDE. In this talk, I presented an overview of the existing techniques (e.g., for Approximate Query Processing) in databases to better support IDE workloads and discussed recent research results and presented challenges to be addressed by the database community.

3.2 Progressive Data Analysis and HCI

Adam Perer (Carnegie Mellon University – Pittsburgh, US)

In this lightning talk, I focus on two aspects of PDA and HCI: the design of PDA systems and user studies of PDA systems. Without PDA, data analysis workflows tend to force users to wait for analytics to complete. However, the design of PDA allows users to analyze and interpret partial results. I present early designs which give users two options: 1) trust the incremental result or 2) wait for everything to complete. However, further research has provided a set of design goals and requirements for PDA. Sampling the space of user studies of PDA systems suggest the need for more studies that take longer amounts of time and that utilize real analysts working on real problems, something missing from literature to date. In summery, there are many opportunities to further advance the design and evaluation of PDA systems.

3.3 Progressive Algorithms and Systems

Florin Rusu (University of California – Merced, US)

Joint work of Chengjie Qin, Yu Cheng, and Weijie Zhao
Main reference Yu Cheng, Weijie Zhao, Florin Rusu: "Bi-Level Online Aggregation on Raw Data", in Proc. of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, June 27-29, 2017, pp. 10:1–10:12, ACM, 2017.
URL https://doi.org/10.1145/3085504.3085514

In this lightning talk, we present two systems for parallel online aggregation and the sampling algorithms they implement in order to generate progressively more accurate results. The first system–PF-OLA (Parallel Framework for Online Aggregation)–defines a unified API for online estimation and provides a parallel/distributed implementation. The estimation is supported by several parallel sampling operators that make use of node stratification. We show how the PF-OLA API is used to detect better hyper-parameters for gradient descent optimization applied to machine learning model training. The second system–OLA-RAW (Online Aggregation over Raw Data)–supports progressive estimation over raw files–not loaded inside the database. It does not require any data pre-processing, i.e., random shuffling. Estimation is supported by a novel bi-level sampling algorithm that is resource-aware, i.e., the number of rows processed (sampled) from a block is determined based on the system load. The "inspection paradox" inherent to parallel sampling is avoided by carefully inspecting blocks at certain time intervals. The common theme of these systems is that they allow for progressive data exploration. They have as a goal to find those queries that do not produce relevant results for the user. This can be done quite fast. We think this is a more realistic goal for progressive computation and visualization–steer the user away from uninteresting data.

3.4 Progressive Analysis–A VA+ML Perspective

Cagatay Turkay (City – University of London, GB)

Main reference Cagatay Turkay, Erdem Kaya, Selim Balcisoy, Helwig Hauser: "Designing Progressive and Interactive Analytics Processes for High-Dimensional Data Analysis", IEEE Trans. Vis. Comput. Graph., Vol. 23(1), pp. 131–140, 2017.
URL https://doi.org/10.1109/TVCG.2016.2598470

Progressiveness offer a fresh look at how visual analytics integrate machine learning algorithms within analysis processes and how machine learning algorithms can be designed and used in real-world settings. There are already several promising efforts both in visual analytics and machine learning literature that provide a basis for further research in progressive data analysis. In visual analytics, there are already frameworks, guidelines and also early examples of iterative machine learning algorithms embedded in visual analysis pipelines. In machine learning literature, the body of work on iterative learning models provide a strong basis to build upon. As expected from any emergent discipline, there are several challenges that lie at this intersection of incorporating machine learning methods in progressive settings, such as, the issue of concept drift, model adaptivity, e.g., in terms of complexity or flexibility, providing specific progressiveness qualities such as interactive steerability and convergent behaviour in terms of quality. On top of these algorithmic challenges, practical issues such as reproducibility and modelling provenance are critical for the future adoption of progressive methods. Considering the increasingly growing complexities of models and volume of data, and with all the various challenges that it brings with it, progressive analysis offers exciting opportunities for impactful, timely research.

3.5 A Visualization Perspective on Progressive Data Analysis

Marco Angelini (Sapienza University of Rome, IT)

Main reference Marco Angelini, Giuseppe Santucci, Heidrun Schumann, Hans-Jörg Schulz: "A Review and Characterization of Progressive Visual Analytics", Informatics, Vol. 5(3), p. 31, 2018.
URL https://doi.org/10.3390/informatics5030031

Progressive Data Analysis (PDA) is a novel analysis paradigm aimed at providing a result for long computational processes in the form of a sequence of approximating intermediate results, where the cause of the long time needed can be the complexity of the process and/or the amount of the data to process. Visualization plays a big role in allowing a human to evaluate, monitor and understand the results of a data analysis process, eventually steering it toward the desired goal. This lighting talk introduces the state of the art of Visualization and Visual Analytics for supporting PDA scenarios (PDAV or PVA) in terms of which characteristics govern the production of intermediate visualizations to evaluate partial results, workflow of analysis that a user follows interacting with a PVA system and differences with respect to a more traditional Visual Analytics system, management of additional like the control of the progression parametrization or the steering of the analysis, management of visually encoded uncertainty, stability, progress, and user's interaction. Important open problems, challenges and research directions conclude the talk.

3.6 Need for Progressive Analytics on Very Large Data Series Collections

Themis Palpanas

Main reference Kostas Zoumpatianos, Themis Palpanas: "Data Series Management: Fulfilling the Need for Big Sequence Analytics", in Proc. of the 34th IEEE International Conference on Data Engineering, ICDE 2018, Paris, France, April 16-19, 2018, pp. 1677–1678, IEEE Computer Society, 2018.
URL https://doi.org/10.1109/ICDE.2018.00211

Massive data series collections are becoming a reality for virtually every scientific and social domain, and have to be processed and analyzed, in order to extract useful knowledge. This leads to an increasingly pressing need, by several applications, for developing techniques able to index and mine such large collections of data series. It is not unusual for these applications to involve numbers of data series in the order of hundreds of millions to several billions, which are often times not analyzed in their full detail due to their sheer size. Bringing into the picture interactions with users and visualizations in support of explorative analytics introduces novel challenges related to the different kinds of queries (e.g., precise vs imprecise) and answers (e.g., exact vs approximate) that should be available to users, the amount of information to visualize, and the time performance of all these operations, which calls for progressive analytics in order to allow interactivity.

3.7 User Interface Elements and Visualizations for Supporting Progressive Visual Analytics

Sriram Karthik Badam (University of Maryland – College Park, US)

Joint work of Niklas Elmqvist, Jean-Daniel Fekete
Main reference Sriram Karthik Badam, Niklas Elmqvist, Jean-Daniel Fekete: "Steering the Craft: UI Elements and Visualizations for Supporting Progressive Visual Analytics", Comput. Graph. Forum, Vol. 36(3), pp. 491–502, 2017.
URL https://doi.org/10.1111/cgf.13205

Progressive visual analytics (PVA) has emerged in recent years to manage the latency of data analysis systems. When analysis is performed progressively, rough estimates of the results are generated quickly and are then improved over time. Analysts can therefore monitor the progression of the results, steer the analysis algorithms, and make early decisions if the estimates provide a convincing picture. In this talk, I will describe interface design guidelines for helping users understand progressively updating results and make early decisions based on progressive estimates. To illustrate the ideas, I will also present a prototype PVA tool called InsightsFeed for exploring Twitter data at scale. Our user evaluation showcased better usage patterns in making early decisions, guiding computational methods, and exploring different subsets of the dataset using the progressive user interface, compared to a sequential analysis without progression.

3.8 Progressive Visual Analytics: User-Driven Visual Exploration of In-Progress Analytics

Charles D. Stolper (Google, US)

Joint work of Adam Perer, David Gotz
Main reference Charles D. Stolper, Adam Perer, David Gotz: "Progressive Visual Analytics: User-Driven Visual Exploration of In-Progress Analytics", IEEE Trans. Vis. Comput. Graph., Vol. 20(12), pp. 1653–1662, 2014.
URL https://doi.org/10.1109/TVCG.2014.2346574

As datasets grow and analytic algorithms become more complex, the typical workflow of analysts launching an analytic, waiting for it to complete, inspecting the results, and then relaunching the computation with adjusted parameters is not realistic for many real-world tasks. This paper presents an alternative workflow, progressive visual analytics, which enables an analyst to inspect partial results of an algorithm as they become available and interact with the algorithm to prioritize subspaces of interest. Progressive visual analytics depends on adapting analytical algorithms to produce meaningful partial results and enable analyst intervention without sacrificing computational speed. The paradigm also depends on adapting information visualization techniques to incorporate the constantly refining results without overwhelming analysts and provide interactions to support an analyst directing the analytic. The contributions of this paper include: a description of the progressive visual analytics paradigm; design goals for both the algorithms and visualizations in progressive visual analytics systems; an example progressive visual analytics system (Progressive Insights) for analyzing common patterns in a collection of event sequences; and an evaluation of Progressive Insights and the progressive visual analytics paradigm by clinical researchers analyzing electronic medical records.

3.9 Optimistic Visualization and Progressive Interactions

Dominik Moritz (University of Washington – Seattle, US)

Joint work of Danyel Fisher, Dominik Moritz
Main reference Dominik Moritz, Danyel Fisher, Bolin Ding, Chi Wang: "Trust, but Verify: Optimistic Visualizations of Approximate Queries for Exploring Big Data", in Proc. of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017., pp. 2904–2915, ACM, 2017.
URL https://doi.org/10.1145/3025453.3025456

Analysts need interactive speed for exploratory analysis, but big data systems are often slow. With sampling, data systems can produce approximate answers fast enough for exploratory visualization at the cost of accuracy and trust. We propose optimistic visualization, which lets users analyze approximate results interactively and detect and recover from errors later. It gives users interactivity and trust. In addition, progressive interactions let users start with low-resolution interactions and later interact at the pixel resolution.

3.10 Visplore–a highly responsive Visual Analytics Application

Thomas Mühlbacher (VRVis – Wien, AT) and Harald Piringer

Joint work of Visual Analytics Group at VRVis
Main reference Harald Piringer, Christian Tominski, Philipp Muigg, Wolfgang Berger: "A Multi-Threading Architecture to Support Interactive Visual Exploration", IEEE Trans. Vis. Comput. Graph., Vol. 15(6), pp. 1113–1120, 2009.
URL https://doi.org/10.1109/TVCG.2009.110

In my brief talk, I presented a highly responsive Visual Analytics software called "Visplore", which has been developed at VRVis since 2005. The software enables a flexible in-depth exploration of process- and timeseries data for process engineers, R&D staff and data scientists. Solutions based on Visplore are used by more than 10 companies in industries like automotive design, manufacturing, and energy transmission. The goal is to allow more in-depth analyses than business intelligence tools, while being simpler to use than standard statistics software. As one focus of the talk, I described the software's multi-threading architecture that enables highly responsive feedback for interaction, visualization, and computations based on millions of data records [1]. User interactions like brushing, for example, will immediately halt any ongoing computations, trigger invalidations of all dependencies, and continue computations based on the new input. As a second focus, I showcased a success story of Visplore from the manufacturing sector, where the high performance of the software has enabled process engineers to perform daily tasks of quality control and process operation based on millions of data records, leading to siginifcant cost savings and efficiency gains.

References

1 Harald Piringer, Christian Tominski, Philipp Muigg and Wolfgang Berger. *A Multi-Threading Architecture to Support Interactive Visual Exploration*. IEEE Transactions on Visualization and Computer Graphics, 15(6):1113–1120, Nov. 2009.

4 Working groups

4.1 Progressive Data Analysis and Visualization: Definition and Formalism

Charles D. Stolper (Google, US), Michael Aupetit (QCRI – Doha, QA), Jean-Daniel Fekete (INRIA Saclay – Orsay, FR), Barbara Hammer (Universität Bielefeld, DE), Christopher M. Jermaine (Rice University – Houston, US), Jaemin Jo (Seoul National University, KR), Gaëlle Richer (University of Bordeaux, FR), Giuseppe Santucci (Sapienza University of Rome, IT), and Hans-Jörg Schulz (Aarhus University, DK)

Big data, complex computations, and the need for fluent interaction are the source of challenges for contemporary visual analytics systems. Visual analytics systems must support a fluent, interactive flow between the human analyst, and visualizations and computational analysis; these challenges can impair this flow. To address them, the idea of Progressive Visual Analytics (PVA) [11, 4, 2] becomes increasingly appealing. A PVA approach can either subdivide the data under analysis in chunks, processing each data chunk individually, or it can subdivide the analytic process into computational steps that iteratively refine analytic results [10]. By doing so, PVA yields partial results of increasing completeness or approximate results of increasing correctness, respectively.

The expected benefits of introducing a progression within a Visual Analytics application include reducing latencies; allowing users to monitor the computational process; explaining the computational process through algorithm animation; reducing visual overload; and allowing the user to steer the running computation.

The Definition and Formalism group aimed to frame and formally characterize this emerging approach in the context of other well defined and related approaches.

4.1.1 Computation methods relevant in the progressive context.

Several approaches exist in the context of producing results from different data sources according to different constraints about quality, time and resources.These approaches can be described by their characteristics: by the data they deal with (i.e., bounded, unbounded), the way they digest such data (i.e., as a whole or in chunks), the number of outputs they produce, the flexibility they offer on response time, the measures they provide about uncertainty, and the possibility of being steered during their execution.

As an example, streaming computations deal with unbounded data and produce an unbounded number of results, while any-time computations deal with bounded data and are ready to produce a single output at user demand, output characterized by a quality depending on the computation time. Progressive computation, instead, deals with bounded data providing a converging sequence of output characterized by increasing quality, allow to be steered and are able to produce the first result in a previously specified time frame, characteristic that they share with bounded queries.

In order to better characterize a progressive computation we have investigated similarities and dissimilarities among various time and quality related computations (e.g., streaming data, online, approximation, anytime, iterative, etc.), characterizing them along input data, output, process, and accuracy. In the following section we analyze the most relevant proposals in the field, showing their properties and the relationships hey have with progressive computation.

4.1.2 Definition

A progressive process produces a sequence of useful/meaningful *intermediate results* in contrast with pure iterative processes that inherently generate intermediate results whose utility is mainly related to monitoring, tuning, steering, and explaining the process itself. The first result is constrained by a time value and must be be generated within that value. The sequence should be of increasing utility to the user, where utility can be defined according to the following aspects:

- actual result measures, like uncertainty, progression (data, time, iterations);
- comparison with previous result: delta in the data, (perceptual) delta in the visualization;
- analysis of fluctuating results that may be skipped to avoid confusing the user
- results that are produced at a too fast rate that may be delayed to not confuse the user with a too fast update generating a manageable rate for user understanding;

To sum up, a *progressive* computation is:

- a computation
  - **bounded on time and data**
- which **reports intermediate outputs**, namely
  - a result,
  - a measure of quality, and
  - a measure of progress,
- within **bounded latency**,
- **converging** towards the true result, and
- **controllable by a user during execution** either by
  - aborting, pausing, etc., or
  - by tuning parameters (steering).

4.1.3 Formalism & Comparison

As per the previous section, here we give more formal definitions with unified notations, of the computation approaches related to *Progressive* computation, namely: *Approximate*, *Optimistic*, *Online*, *Streaming* and *Iterative* computations.

Let F be a standard data analysis function with parameters $P = \{p_1, \ldots, p_n\}$. The computation of F takes a time t, starts with a machine state S_z, and modifies this state that becomes S_{z+1}. Also, for data analysis, we distinguish the parameters P from a set of input *data* D that the function takes as input. The function also yields a result vector R, and possibly a quality vector Q and a progression vector π:

$$F(S_z, P, D) \mapsto (S_{z+1}, R_z, t_z, Q_z, \pi_z) \tag{1}$$

The *eager* function F_e is the baseline that uses a fixed set of parameters P and can work on the full dataset D in one run in time t_D, yielding ideally the final exact result R_D with baseline quality Q_D.

To compare the different computation approaches, we focus on 3 properties of F: *accuracy*, *performance*, and *real-time* capabilities. Unlike its mathematical counterpart, the result R can suffer from *inaccuracies*[1]. Some methods are *real-time*, meaning that there is a mechanism to guarantee that the computation will complete within a specified constant time.

[1] "Accuracy" is defined by ISO 5725-1 [1] as "the closeness of a measurement to the true value"

Approximate Query Processing (AQP) returns either an approximate result unbounded quality in a user specified time θ (F_t) or an approximate result with bounded error ϵ in unbounded time (F_q).

Optimistic computation is a combination of AQP with bounded in time and unbounded quality in a first step and eager computation that yields the exact result R_D and baseline quality in the second step, in unbounded time, i.e., $t_D + \theta$ for AQP$_t$. It returns explicitly the progress π it made at the first and final steps.

Online computation is related to the way the data D are made available to the function F [3]. Informally, when an online function F_o is called iteratively with growing chunks of data D_z ($D_{z-1} \subset D_z \subset D$), it returns its results more accurately and more efficiently than if it were called directly from scratch with D_z. Online algorithms are popular in machine-learning and analytics in general, in particular to handle dynamic data.

Formally, the growth of the data tables can be modeled with a partition of D into n_C non-intersecting subsets or chunks C_i: $\bigcup C_i = D$, $C_i \neq \emptyset$ and $C_i \cap C_j = \emptyset$. At step z, F_o is called with $\bigcup_{j=1\ldots z} C_j$. Online functions guarantee that:

- **Accuracy** each result R_z is accurate and after running on all data chunks R_{n_C} is equal to R_D.
- **Time** the total time t_z to compute $F_o(S_z, P, \bigcup C_{1\ldots z})$ after R_{z-1} is much shorter than the total time t_D to compute $F_e(S_1, P, \bigcup C_{1\ldots z})$,
- **Unbounded time** however, online functions offer no guarantee that t_z is bounded.

4.1.4 Streaming methods

They are related to memory usage and time constraints [6]. According to Wikipedia: "*streaming algorithms are algorithms in which the input is presented as a sequence of items and can be examined in only a few passes (typically just one). These algorithms have limited memory available to them (much less than the input size) and also limited processing time per item.*"

Compared to an online function, a streaming function F_s only takes the chunks C_z and not the whole dataset D. The maximal size of a chunk can be limited by the algorithm.

Streaming functions properties are:

- **Accuracy** results R_i can be inaccurate so $R_z \approx R_D$,
- **Time** the total time t_z to compute $F_s(S_z, P, C_z)$ after R_{z-1} is much shorter than the total time to compute $F_e(S_1, P, \bigcup C_{1\ldots z})$, which would be $\sum_{j=1,z} t_j$,
- **Real-time** streaming functions are real-time so t_z is bounded,

Due to the real-time constraints, streaming algorithms can sometimes compute rough approximations of their results. Furthermore, streaming algorithms are usually difficult to design and to implement; relying on streaming algorithms only to implement an entire data analysis infrastructure severely limits the range of possible applications that could be built. For example, at the time of the writing on this document, the Spark streaming machine-learning library [8] provides 86 classes compared to the 245 provided by the scikit-learn machine-learning library [7].

Iterative computation is defined in Wikipedia as: "a mathematical procedure that generates a sequence of improving approximate solutions for a class of problems." An iterative function performs its computation through some internal iteration and can usually provide meaningful partial results R_j after each iteration, although some extra work might be required to obtain the partial result in an externally usable form [9].

Compared to an online function, an iterative function F_i always takes the whole dataset D and returns improving results R_z that eventually becomes the exact R_D. Iterative functions guarantee that:

- **Accuracy** the results R_i will converge to the real value R_D: $\lim_{j\mapsto\infty} R_j = R_D$, but the convergence may not improve at each step,
- **Time** the total time to compute R_D is usually controlled indirectly either through a tolerance parameter or a maximal number of iterations; none of these parameters allow analysts to predict the completion time,
- **Bounded time** iterative functions offer no guarantee that t_z is bounded.

Progressive computation is designed to deliver results for analysts at a controlled pace and with controlled accuracy. A progressive function F_p has a form very similar to the AQP_{time} function except it works on a subset $D_z \subset D$ of the full data with no constraint on D_z regarding any other $D_{j<z}$. In addition to the result and time, progressive function output explicitly quality Q and progress π.

Progressive functions guarantee that:

- **Accuracy** just like online and iterative functions, the results R_i will converge to the real value R_D. in a finite number Z of steps: $\lim_{j\mapsto Z} R_j = R_D$. Additionally, like streaming functions, each call will provide an estimate of the result R_i that can be inaccurate, but like iterative functions, the accuracy can be improved by calling F_p several times with the same input ($D_z = D_{z-1}$),
- **Time** progressive functions are similar to online functions when D_z is growing, and similar to iterative functions when D_z is constant,
- **Real-time** progressive functions offer the guarantee that t_z is bounded ($t_z \leq \theta$).
- **Steerable** Progressive functions parameter P_z can be modified between each call. This is called *computational steering* [5], and has emerged from analysts who could monitor the evolution of their computations by visualizing the progression, and then wanted to influence the computation without restarting from scratch when they realized that some parameters would have to be tuned.

The table 1 summarizes all these computations.

4.1.5 Open questions and Future work

Good quality metrics. In the formal definition we have defined quality as a simple scalar. The exact quality function depends very much on the progressive algorithm under consideration, but we can make a few statements on properties of ideal quality functions:

- They are a direct function of the distance of R_z to the end result. In practice, defining this function exactly is impossible since the end result is not (yet) known. Here, we will have to settle on an approximate quality function that is close enough to the actual quality. Depending on the definition of this approximate function Q', there will be additional uncertainty introduced by the difference between Q and Q' which may be variable. It is unknown how this influences user perception of quality.
- Consistent across a single computation: in order to accurately gauge the state of the computation, there should be a total order on the values given by the quality function. In other words, the quality should be comparable across the entire computation.
- They are strictly increasing: ideally a quality function strictly increases over the duration of the progressive algorithm. Depending on the progressive algorithm however, there may be situations where quality decreases over time (for example the algorithm is moving out of a local minimum solution). In these cases a user will have to decide on the suitability

Table 1 Comparison of computation methods properties.
R is the desirable result
t is the computation time
P are the algorithm parameters
Q is the quality
π is the progression
t_D is time taken on full dataset D.
C_i are chunks of the full data: $\bigcup_{z\in\{1\ldots n_C\}} C_z = D$, $C_i \neq \emptyset$ and $C_i \cap C_j = \emptyset$.
$D_i \subseteq D$.
θ is a user-defined upper-bound on latency
Output is always: $(S_{z+1}, R_z, t_z, Q_z, \pi_z)$

	Input	R	t	P	Q	π
Eager	$F_e(S, P, D)$	Exact R_D	t_D	Fixed	Baseline	Irrelevant
$\mathrm{AQP_t}$	$F_t(S, P, D, \theta)$	Apx $d(R, R_D) \geq 0$	$\leq \theta$	Fixed	Unbd	Irrelevant
$\mathrm{AQP_q}$	$F_q(S, P, D, \epsilon)$	Apx $d(R, R_D) \leq \epsilon$	Unbd	Fixed	Bounded	Irrelevant
Optimistic	AQP→eager	$R_2 = R_D$	$\leq t_D + \theta$	Fixed	Baseline	Explicit
Online	$F_o(S_z, P, \bigcup C_{1,\ldots,z})$	Exact for $z = n_C$	$\ll t_D$	Fixed	Baseline	Implicit
Streaming	$F_s(S_z, P, C_z)$	Apx $d(R_z, R_D) \geq 0$	$\ll t_D$	Fixed	Unbd	Implicit
Iterative	$F_i(S_z, P, D)$	$R_z \xrightarrow[z\to\infty]{} R_D$	Unbd	Fixed	Unbd	Implicit
Progressive	$F_p(S_z, P_z, D_z, \theta)$	$R_z \xrightarrow[z\to Z]{} R_D$	$\leq \theta$	Ctrled	Unbd	Explicit

of an intermediate result R_z based on the time taken, a visual assessment of R_z or the current quality relative to the initial quality.

Lastly, in some cases, a suitable quality function may simply not be available. In that case, the only measure of quality may be the algorithm's progression. Whether the algorithm can still be deemed progressive (as opposed to iterative) is an open question.

Good progress metrics. The progress π_z was defined abstractly as a variable between 0 and N. Since the computation is bounded, an upper bound N is known at the start of the computation, at least as a parameter of the algorithm itself. This implies we can always output an abstract progress indicator. While this abstract progress of the algorithm is strictly increasing, it is better to present the user with more actionable progress measures, such as time remaining until finish (can be determined by sampling) or percentage of records analyzed.

Semantics of algorithm parameters. In the formal definition we have assumed a parameter set P, containing parameters driving the algorithm. In practice, this set P will consist of parameters that can influence the convergence rate of the chosen algorithm (e.g., parameters that control a speed/quality trade-off internal to the algorithm) and parameters that can influence the end result (e.g., the number of clusters in a cluster operation). Changing one of the former should ideally allow the progressive algorithm to continue from its current state. Changing one of the latter typically necessitates a full recompute. Depending on the progressive algorithm in question, there may be cases where the last intermediate result R_z can be reused in the new progressive computation to try to maintain some continuity in the progression.

Latency bound. Choosing the latency bound between intermediate results (θ) is a trade-off between faster-coming results and more precise results. At the same time, it is closely related to both the user's visual processing capabilities and the rate of data processing. Indeed, ideally, θ is expected to lay under several seconds for the system to appear reactive to the user, but still be large enough for the changes to be digestible to the user. Since producing intermediate results typically introduces an overhead, θ should fit the rate to which the underlying computation takes place to avoid producing successive results without changes as a result of overload instead of stabilization of the algorithm. Consequently, choosing the optimal value of θ relative to a particular algorithm and user task is an open question as well as the extents to which a user could interactively change it during progression.

Mapping to use-cases. While we have presented a semi-formal definition of what progressive computation is, and what its parameters are, we still need to validate this definition against real-world use cases. It would be a good exercise to take some of the use-cases presented in other teams and see how these fit the definitions presented above. Any mismatches can then be used to sharpen or correct the definition.

References

1 ISO 5721-1. Accuracy (trueness and precision) of measurement methods and results - Part 1: General principles and definitions, 1994.

2 Marco Angelini, Giuseppe Santucci, Heidrun Schumann, and Hans-Jörg Schulz. A review and characterization of progressive visual analytics. *Informatics*, 5(3):31, 2018.

3 Allan Borodin and Ran El-Yaniv. *Online Computation and Competitive Analysis.* Cambridge University Press, 1998.

4 Jean-Daniel Fekete and Romain Primet. Progressive analytics: A computation paradigm for exploratory data analysis. *arXiv.org e-print*, 1607.05162, 2016.

5 Jurriaan D. Mulder, Jarke J. van Wijk, and Robert van Liere. A survey of computational steering environments. *Future Gener. Comput. Syst.*, 15(1):119–129, February 1999.

6 S. Muthukrishnan. Data streams: Algorithms and applications. *Found. Trends Theor. Comput. Sci.*, 1(2):117–236, August 2005.

7 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

8 N. Pentreath. *Machine Learning with Spark.* Community experience distilled. Packt Publishing, 2015.

9 Yousef Saad. *Iterative methods for sparse linear systems.* Siam, 2003.

10 Hans-Jörg Schulz, Marco Angelini, Giuseppe Santucci, and Heidrun Schumann. An enhanced visualization process model for incremental visualization. *IEEE Trans. Vis. Comput. Graphics*, 22(7):1830–1842, 2016.

11 C.D. Stolper, A. Perer, and D. Gotz. Progressive visual analytics: User-driven visual exploration of in-progress analytics. *IEEE Trans. Vis. Comput. Graphics*, 20(12):1653–1662, 2014.

4.2 Tasks, Users, and Human Biases in PVA Scenarios

Hans-Jörg Schulz (Aarhus University, DK), Remco Chang (Tufts University – Medford, US), Luana Micallef (University of Copenhagen, DK), Thomas Mühlbacher (VRVis – Wien, AT), and Adam Perer (Carnegie Mellon University – Pittsburgh, US)

4.2.1 Introduction

This report describes the result of two breakout groups that focused on the role of human users in PVA scenarios. The first group, "Tasks and User" compiled a list of concrete **example usage scenarios** of PVA, intending to (1) raise awareness of PVA benefits for all communities based on tangible stories, and (2) to investigate recurring patterns in a bottom-up fashion. Based on these scenarios, the group formulated five **high level tasks** [T0-T4] that only become possible through the progressiveness of the described applications.

The second breakout group, "Humans in the progressive Loop", characterized different **types of users** of PVA applications. The discussion focused on how their backgrounds, skills, expectations and benefits from progressiveness differ, and which implications for the design of PVA applications arise. Moreover, the group built on the usage scenarios of the first breakout group to establish a more **structured characterization of PVA scenarios** regarding the dimensions concurrency, task dependency and interactivity. Finally, the effect of **human biases** that may occur in each scenario is discussed, to outline potential challenges and pitfalls for future developers of PVA systems.

4.2.2 Example Usage Scenarios / "Stories" (Adam Perer)

During our discussions, we established a set of usage scenarios to demonstrate the breadth and variety of progressive data analysis. In each usage scenario, we highlight the progressive aspects in **bold**. For brevity, we only describe a representative subset of our stories here:

4.2.2.1 Rapid Queries on Very Large Data

User wants to write query computing an average from a large amount of data, the **execution of which could take hours**.

- User runs the query against a progressive visualization system.
- At some point, first **intermediate result** is visualized
- From this, the user notices implausible / unexpected scaling of Y-Axis
- The user's initial thought is: "maybe the dataset seen yet is not representative"
- The user modifies the sampling to **prioritize data chunks** that she thought would be more representative.
- Still, the intermediate results are of the same strange scaling
- The user has the insight: "Oh, the query uses sum instead of average"
- Based on these preliminary results, user **cancels and restarts** the corrected query

However, not all initial queries tell the full story. Consider the following scenario where an inconsistency emerges after the initial approximation:

- The user wants to see what the origin and destinations of mid-range flights for Hawaiian airlines are.
- She **queries** for the histogram and **sees** that most flights are from California to Hawaii. This observation is based on the approximate result.

- Later she **runs the query over the complete data** – which takes a few minutes to complete – and she sees that she **missed a flight** from California to Pennsylvania. This flight was missing in the approximation and only included in the later result.

4.2.2.2 Model Search

Doctors wants to find the best model for predicting body mass index (BMI) from a list of clinical data to find external risk factors. **The optimization of hyperparameters using AutoML is very costly and often takes hours to days to execute**.

- The user initiates AutoML to build a regression model based on automatically selected features from the set of all available physical and diagnostic features that are available
- The system **progressively returns potential models** and **shows models' accuracies**
- User realizes that accuracy is "too good"; almost 100% accurate for all models
- She explores the correlations between features. By examining a scatterplot matrix she notices that the factors Weight and Height perfectly correlate with BMI.
- She realizes that she had included these two factors mistakenly as features (as BMI is a direct function of Weight / Height, but she wanted "external" risk factors).
- She changes the parameterization by removing these two variables, **retriggering the execution** of AutoML, which results in more informative/realistic models.

4.2.2.3 Exploratory Search

A user wants to find frequent event sequences in a large database of time-stamped events. She launches a frequent sequence mining algorithm, specifying a support threshold that defines common patterns that occur in at least 5% of the population. This particular algorithm returns short patterns first, as they are cheaper to compute, and then longer ones as the algorithm continues, in a breadth-first manner. This particular algorithm can run concurrently and independently, spread across multiple threads and servers.

After a few seconds, patterns start returning from the algorithm and she gets a ranked list of the patterns found so far. This helps her **get an overview** of the common patterns that the user was previously unaware of, as the data was too big to query directly. At this point, she is able to **start exploring** the frequent patterns already and learn what is common in its dataset. After looking at this list, she notices that one of the patterns is irrelevant to its goal, so **directs the algorithm** to not continue looking for longer patterns that begin with this sequence. Furthermore, she notices one short pattern that seems quite promising and is curious if there are longer patterns that begin with this sequence. She tells the system to **prioritize this search space**, and look for longer patterns that begin with this sequence. This progressive process gave her the opportunity to help guide the search algorithm, which gave her a better understanding of how the search algorithm worked, as well as being able to take advantage of her domain knowledge. The progressive process also gave the user to speed up the algorithm by focusing on relevant spaces.

4.2.2.4 Domain Example: Disaster Response

A crisis manager has to analyze a **vast amount of data to trigger emergency response** under time, human and equipment constraints. She collects millions of tweets relative to a crisis from diverse sources and other sensor data like UAV photographs, satellite images or seismic sensors, wind speed sensors. She analyzes the data to detect events that require emergency response. The data can all be stored, but their amount requires automatic processing to extract the data relevant to the crisis manager. The data are not easy to analyze, some events on tweets can be rumors, some images can be difficult to interpret. Automatic processing is necessary but the more data it processes the more confident it is relative to the certainty of occurrence and the severity of an event. The user can select the type of event to focus on depending on their certainty and severity as evaluated **as per the current status of the progressive processing**, correcting badly classified events (named entities like location, person and organization in tweets) to guide the classifier in doing its task, triggering relevant emergency response as soon as she decides an event has reached her subjective threshold of certainty and severity.

4.2.3 High-Level Tasks / Benefits of PVA (Thomas Mühlbacher)

Based on the usage scenarios from the previous section, we extracted a set of recurring high-level tasks that were only made possible by the early involvement of users in the ongoing computations. The ability of performing each task can be seen as a concrete benefit added by the progressiveness. Fulfillment of the task list may also serve as a qualitative benchmark for the evaluation of future PVA architectures. The tasks are:

T0: Making decisions under time- and resource constraints
T1: Discovering wrong assumptions early to save time and not lose train of thought
T2: Rapid exploration of very large data / information spaces
T3: Incorporating human knowledge to steer algorithms and speed them up
T4: Building trust and deepening the understanding of algorithms

4.2.4 Characterization of Users (Hans-Jörg Schulz)

Users utilize PVA in different ways and for different objectives. Accordingly, we distinguish between three main types of users:

U0: "The Consumer" who makes decisions based on the output of a PVA model (usually without knowledge of that model or even PVA as a concept)
U1: "The Domain Expert" who makes decisions with the PVA model (domain-specific, but no CS/Stats knowledge of underlying model or architecture)
U2: "The Data Scientists" who makes decisions about the PVA model (does have CS/Stats knowledge, but not necessarily domain knowledge of underlying model or architecture)

These prototypical users perform different subsets of the established progressive tasks [T0-T4], use different aspects of PVA to perform these tasks, have different expectations on PVA, which in turn incur different properties of the UI in which they work with PVA.

4.2.4.1 U0: The Consumers

The consumers want to make decisions under time- and resource constraints [T0], for which they need to rapidly explore very large data / information spaces [T2], and discover wrong assumptions early to save time and not lose their train of thought [T1]. An everyday example is the use of a progressively loading online map to explore different routes from the current position to a desired goal. To do so, they make use of intermediate PVA results:

- to identify those assumptions that were wrong (e.g., map showing the wrong place),
- to assess the progress of the loading map (e.g., am I already looking at the finest resolution?) or the "aliveness" of the process (e.g., does the server still respond?), and
- to realize a simple "steering" in the sense of quickly correcting a wrong assumption by adjusting the focus of the PVA output (e.g., panning the map to a region of interest).

Their expectation of the underlying PVA process is that of a well-behaved, calmly converging output that arrives at a predictable end. The mental analogy of a "loading process" probably describes this expectation best. To have the consumers being able to use PVA right away – without hesitation or an extra learning curve for the progressiveness – it should ensure this expectation is met by means like:

- using/building on commonplace metaphors and familiar abstractions that are already part of the interface, e.g., the "online map application",
- reducing visual complexity, fluctuations, jumps, flickering or any other indication of the underlying process not calmly and predictably producing progressive outputs, and
- maintain the users' mental map, e.g., by introducing anchors or waypoints.

4.2.4.2 U1: The Domain Experts

The domain experts want to do all the things the consumers do, but on top they want to have more control over the internal model execution – not necessarily about what these models are, but how they are parametrized and utilized. In most cases, this means to steer the algorithm based on their knowledge and speed them up while already running [T3]. An example is an online user study run via Mechanical Turk that shows after the first 20+ responses that certain questions are formulated ambiguously. So, the domain expert can step in and adjust the question on the fly to get answers in the future that are more to the point. For doing this, intermediate PVA results are used:

- to identify assumptions that are wrong (e.g., about the choice of questions in the study, or the quality of the study results)
- to decide, whether to wait or step in (e.g., wait for 20+ more results or make the change now?)
- to steer the result towards the best match with the domain knowledge of the experts (e.g., for A/B testing alternative formulations for the identified problematic questions)

Their expectation of the underlying PVA is that of a predefined, but malleable process that can be tuned while running to overcome corner cases or to deal with unexpected complications. To do so, the UI needs to give access this adaptability of the process by means of:

- not glossing over, but reflecting – maybe even emphasizing – the fluctuations, jumps, and other indications of the underlying model not behaving as expected,
- exposing degrees of freedom and steering possibilities provided by the underlying model in a suitable language (e.g., building only on domain concepts and semantic interaction, rather than exposing statistical/algorithmic parameters directly), and
- providing a feeling of uncertainty that allows to judge the results from interactively performed parameter changes – also in a domain-specific manner.

4.2.4.3 U2: The Data Scientists

The data scientists need to be able to do everything the domain experts do, but as it is usually them who are building and optimizing the underlying models, they may also need to gain a deep understanding of the algorithms to increase their trust in the model [T4]. An example is the debugging of complex models and algorithms in multiple steps and with data of increasing sizes. To facilitate this, intermediate PVA results are used:

- to identify wrong assumptions early (e.g., by testing different models,)
- to recognize processes getting stuck in local optima, and

- to efficiently allocate time for different tasks involved in model building, testing, and revision.

Their expectation of PVA is that of a completely open, transparent process that can not only be reparametrized, but in which modules and methods can be exchanged at intermittent stages. For this, the UI must:
- expose all details, parameters and process metrics to the user on demand AND making them adjustable on the fly,
- incorporate as much as possible of statistics over the outputs and the process on demand, to judge stability and convergence of the method as a whole, and
- allow possible changes in a What-If scenario that branches off alternatives and compares them later.

These three user types rather describe a continuum with other possible types in between – e.g., a U1.5 "The Apprentice / Learner of Algorithms", either for curious/suspicious domain experts (U1) or for aspiring data scientists (U2), that monitor PVA processes without changing them. The task characterizing a U1.5 is [T4], i.e., to deepen the understanding and trust in algorithmic methods by looking at their progression. In addition, we note that any person can take on multiple user types of PVA during an analysis, as it may become necessary. It may thus also make sense in many application scenarios to refer to U0-U2 as "user roles" instead of "user types".

4.2.5 Structured Characterization of PVA Scenarios (Remco Chang)

We consider three dimensions to the use of progressive visualization: concurrency, task dependency, and interactivity. Concurrency refers to whether the user interacts with each task in a serial (blocking) or concurrent manner. Task dependency refers to whether the multiple tasks that the user is trying to perform have outcomes that depend on each other. Lastly, interactivity refers to whether the user is required to interact with the progressive visualization; and if so, if the user would interact with the final outcome of the computation or interact during the process of progression. Table 2 shows these dimensions, the types of usage scenarios that they most associate with (refer to Section 4.2.2 for more detail), as well as the most threatening biases (described in Section 4.2.6).

For example, when progressive visualization is used for **debugging** queries (see Section 4.2.2.1), the user would *not interact* with the progressive visualization (other than to terminate the query), the task of executing a query is *serial*. Task dependency in this case does not apply as debugging can be used in both contexts.

For **exploring** data, the user would *interact* with the outcomes of the progressive computation after the computation is complete or the user is satisfied with the partial result (as such, the process is *serial*). In contrast, for **model steering**, the user would interact during the progressive computation to guide the execution of a machine learning process.

Progressive visualization can also be used for managing or executing multiple tasks *concurrently*. In the most simple case, the user can **monitor** the progression of multiple tasks without interacting with the visualization. Further, the user could *interact* with the tasks. For example, the user could reduce the search space in an **exploratory search** task when multiple *independent* search processes are *concurrently* executed.

Lastly, we consider the case of a *concurrent*, *dependent*, and *interactive* progressive visualization usage scenario. This scenario has not been defined in Section 4.2.2 because few modern systems can support this without the use of a progressive database engine. For example, imagine that an analyst is looking for the "Top 10" customers in their database.

Table 2 Characterization of Usage Scenarios.

Usage Scenarios	Concurrency	Task Dependency	Interactivity	Biases (negative effect)
explain	serial	indep/dep	no	illusion of transparency
debugging	serial	indep/dep	no	anchoring, recency
explore	serial	indep/dep	with results	sunk cost fallacy
steering	serial	indep/dep	with results/ progression	illusion of control, confirmation bias, superiority bias
monitoring	concurrent	indep/dep	concurrent	conservatism, change blindness
exploratory search	concurrent	indep	with results/ progression	cognitive overload, inattentional blindness
continuous querying	concurrent	dep	with results/ progression	base rate fallacy

When this search query is being executed and the partial results are progressively streamed in, the analyst *concurrently* executes a join query to see the purchases of these customers before the first query completes. This capability has the potential to significantly speed up the user's performance since the user can begin to formulate new questions and queries without being *blocked* by the original query or queries. However, the execution of this query is different from what traditional databases support, in particular because the database engine will need to join two tables that are dynamically changing.

In summary, the use of progressive visualization affords the user in performing wide ranging usage scenarios, tasks, and strategies. However, with the benefits of progressive visualization also comes potential costs to the user. In the section below, we describe the cognitive biases that can most threaten the use of progressive visualization in each of the usage scenarios.

4.2.6 Biases in Human Reasoning (Luana Micallef)

A large number of factors can affect human reasoning. One widely studied phenomena is cognitive biases in human reasoning in general [3] and in visualization in particular [1, 2]: a systematic and involuntary way of how humans deviate from rational judgement, regardless of intelligence and domain expertise. The majority of the cognitive biases lead to inaccurate judgment and distorted perception. However, a small number of biases lead to a positive effect in certain analytic tasks, such as progressive analytics. It is important to be aware of these biases and their effect, so visualizations are carefully designed to reduce negative effects of biases, while attempting to fully exploit positive ones.

For instance, reasoning about probabilities is hard for humans, and under uncertainty, probabilities are either disregarded or hugely overrated (neglect of probability bias). Moreover, humans are insensitive to sample size and often disregard it when reasoning about probabilities. Both of these biases could have a large negative effect in progressive analytics due to the partial and uncertain results generated by a progressive system.

On the contrary, a few biases and effects support the benefits of progressive analytics. The generation effect indicates that humans can better recall information that they generate than information that they read about. Thus, users that are active in progressive analytics are more likely to understand the problem and the results of the system than passive users. This is also supported by the spacing effect, which indicates that studying a problem over time leads to improved learning than studying the same content in one instance.

Other biases and effects (e.g., confirmation bias) that are specific to the user's analytic task during the progression (e.g., when steering the progression) should also be considered. Examples of such biases that could have a negative effect are listed in Table 2.

List of participants (alphabetical order)

Breakout Group "Task and Users": Michael Aupetit, Remco Chang, Luana Micallef, Dominik Moritz, Thomas Mühlbacher, Themis Palpanas, Adam Perer, Hendrik Strobelt, Emanuel Zgraggen

Breakout Group "Humans in the Progressive Loop": Marco Angelini, Remco Chang, Jörn Kohlhammer, Luana Micallef, Thomas Mühlbacher, Adam Perer, Giuseppe Santucci, Hans-Jörg Schulz

Both groups received input from the organizers Jean-Daniel Fekete, Danyel Fisher, and Michael Sedlmair.

References

1 Dimara, E; Franconeri, S; Plaisant, C; Bezerianos, A; Dragicevic, P. "*A Task-based Taxonomy of Cognitive Biases for Information Visualization*". IEEE Transactions on Visualization and Computer Graphics, to appear.
2 Ellis, G. (Ed.) "*Cognitive Biases in Visualizations*". Springer, 2018.
3 Evans, J S B. "*Bias in human reasoning: Causes and consequences*". Lawrence Erlbaum Associates, 1989.

4.3 ProgressiveDB: Architecture Report

Florin Rusu (University of California – Merced, US), Carsten Binnig (TU Darmstadt, DE), Jörn Kohlhammer (Fraunhofer IGD – Darmstadt, DE), Stefan Manegold (CWI – Amsterdam, NL), Themis Palpanas (Paris Descartes University, FR), Nicola Pezzotti (TU Delft, NL), Charles D. Stolper (Google, US), and Chris Weaver (University of Oklahoma – Norman, US)

4.3.1 Introduction

The goal of this session was to determine the specifications of a computing architecture for progressive computation in order to support interactive visual analytics. In a nutshell, design the architecture for "Progressive Tableau". Since the participants came from two distinct groups–visualization and databases–the discussion followed a pattern in which the visualization experts defined the requirements for progressive visualization and the database experts proposed a data computing architecture that implements the specification. Essentially, visualization is the user of a database engine capable to provide progressive results that enhance the display experience. Thus, the interaction between the visual dashboard, e.g., Tableau, and the computing engine is done through standard SQL queries operating over relational data.

The visualization requirements asked for progressiveness to be included in all the elements of the dashboard interface. Results to long-running queries should be returned progressively for continuous display. For these partial, i.e., approximate, results to make sense, they have to be statistically meaningful. Progressive results can be materialized in an evolving time series to provide continuity. The query itself should be allowed to evolve based on the progressive results. This evolution can take the form of minor changes to the original query or the triggering of (related) different queries. The refinement process generates a hierarchy of queries connected through their lineage. The end-user analyst can interact with the queries in two ways. It can monitor the execution by inspecting the progressive results. Or it can steer the query hierarchy by halting uninteresting queries and adding derived queries. In addition to the functional requirements, visualization imposes strict response time constraints that are highly-dependent on the graphical widget. For example, the movement of a slider elicits immediate response times lower than 100 ms, while the pressing of a button increases the feedback time to 1 second. If a custom text entry is required, the feedback time can be as large as 10 seconds. This "time engineering" process interacts with progressive execution by making some graphical widgets more likely to be included in progressive visualization and disqualifying others.

Given the functional specification and the performance constraints, the database experts proposed an architecture for progressive computation. Rather than designing a novel architecture from scratch, the approach taken was to embed progressiveness in a modern relational database architecture accessed through SQL. There are two solutions to realize this approach. First, a middleware that sits between the visual interface and the database engine can implement progressiveness by intercepting the queries and split them into smaller partial subqueries. This is a completely non-intrusive solution. The second approach requires changes to the database operators and tackles the situations that cannot be handled at the middleware layer. Independent of the approach, the progressive architecture has to provide support for progressive results and progressive queries. These are discussed in the following sections.

4.3.2 Progressive Results

The main obstacle to generating progressive results in a standard database are the blocking operators, e.g., join or group-by, which require processing all of their inputs before producing any output tuple. This has led to the development of non-blocking operators, such as hybrid-hash join, which generate result tuples much earlier and continuously. Their functioning principle is to operate on blocks (or chunks) of tuples that are processed by all the operators in the query execution tree at once. This is identical to performing the entire query on a subset of the data, i.e., the selected blocks or chunks. While a partial result is generated, the important question is "What is the significance of this result?" In the case of queries that produce tuples for which the output order does not matter, the partial tuples are as good as it gets. For queries that compute (grouping) aggregates, the partial tuples have to be statistically significant in order to produce an estimate of the result. This is realized by taking random samples from the base tables and performing the subquery on them. The process can be repeated multiple times, with a partial result generated each time. Are these results progressive? Only to the extent that the estimate is refined with each additional executed subquery. This is not the case if the subqueries are independent, only if each subsequent query "builds" on top of the previous ones. This requires the operator state to be preserved across subqueries. Since this is not done by any of the existing database engines, architectural changes may be required.

The XDB system (https://github.com/initialDLab/XDB) that implements the Wander-Join algorithm in PostgreSQL 9.4 is a possible approach. XDB provides online aggregation-style estimates which improve progressively. The progressiveness integration is at all query processing levels, including optimization and execution. A progressive query is specified as follows:

SELECT ONLINE SUM ($l_{extendedprice}$ * (1 − $l_{discount}$)), COUNT(*)
FROM customer, orders, lineitem
WHERE $c_{mktsegment}$=‘BUILDING’ AND $c_{custkey} = o_{custkey}$ AND $l_{orderkey} = o_{orderkey}$
WITHINTIME 20000 CONFIDENCE 95 REPORTINTERVAL 1000

This tells the database engine that it is an online aggregation query which reports the estimations and their associated confidence intervals, calculated with respect to 95% confidence level, every 1,000 milliseconds, and for up to 20,000 milliseconds. In this case, the partial results–consisting of the estimate and its confidence intervals–are managed by the database engine and they are pushed progressively to the visualization client. Another alternative is to allow the client to pull new results by itself. Rather than always transferring the complete result, it is also possible to send only the difference, i.e., delta, from the previous result. In order to execute the query efficiently, B+-tree indexes are built on all the attributes having selection predicates and all the join attributes. This allows for all the blocking operators to be transformed into non-blocking operators, thus, *GetNext*() from the query execution tree root can be pushed to a single table. At this table, *GetNext*() retrieves samples that satisfy the selection predicate efficiently since there is an index. Joins become index lookups because of the indexes built for all the join attributes. As more samples are extracted from the source table, the accuracy of the progressive results increases. This requires maintaining the state of the estimate instead of just the state of the aggregate.

Another strategy for progressive result generation is to extract samples from each of the source tables and execute the complete query over the samples. A middleware architecture is likely more appropriate in this case and the database has to provide quite minor support. State has to be maintained by the middleware between queries for improved results. This requires efficient sampling from the source tables. In PostgreSQL, this can be done by selecting all the tuples stored in the same block (information is available in the catalog). For sampling without replacement, a block has to be accessed only once and this can be controlled by the middleware. In order to provide accuracy guarantees, the middleware can implement several statistical techniques, e.g., CLT-bounds or bootstrap.

4.3.3 Progressive Queries

Producing and visualizing the results of a single query in a progressive manner is not the only aspect of progressive visual data analysis. Rather, the fact that results of a long running query are produced and presented in a progressive fashion during the entire course of the query execution enables the user to intervene and steer their analysis based on the evolving results at any point during query execution, once they realize that the analysis does not yield the desired insight in the data. The user performs their steering actions via the respective interaction mechanisms in the visualization interface, i.e., adjusting sliders, selecting regions, panning, zooming, brushing, adding/removing attributes, etc. These actions result in modifications of the currently running query to produce the required data. The modifications to the query include changing predicate boundaries, adding or removing filters, adding or removing result attributes, adding or removing group-bys (drill-down and roll-up), and adding joins for pivoting. In order to properly expose query changes to the database engine, additions to the SQL language are likely necessary. For instance, queries have to be given names. While this can be done automatically by the system, it is also necessary to expose the names to the user in order to allow non-sequential jumps. Moreover, the query changes have to be recorded. A possible solution is the addition of an ALTER QUERY statement to SQL. For example, if we want to remove the predicate on the *customer* table in the query introduced in the previous section–named Q1–we write:

ALTER QUERY Q1 WHERE: DELETE $c_{mktsegment}$=‘BUILDING’

A naive approach to implement progressive queries is to stop the running query–preferably whenever a result chunk has been produced–and start the execution of the revised query. While semantically correct, this approach does not explicitly convey the information to the database engine that–let alone how–the original query and its modified version are related. Consequently, the queries are treated independently, even though they share the common results. A more effective approach is to refine the previous results and devise a method to combine the new results with the old ones. Moreover, the execution of the derived query should also be optimized based on the previous queries. For instance, the modification to query Q1 given in the example ALTER QUERY statement could be handled by performing Q1 with the predicate $c_{mktsegment}$ <> ‘BUILDING’ and then summing-up this estimate with the progressive result of Q1 computed up to the query change. This choice may be optimal if the number of ‘BUILDING’ tuples is large, thus, they can be pruned away before the costly join. Another option is to perform both Q1 and its modification–essentially, a group-by on $c_{mktsegment}$ with two groups–and then summing the results. Which of the three alternatives–the third being the execution of the modified query from start–depends on many aspects, such as the time when the transition happens, the quality of the Q1 estimate, the data distribution, etc. A progressive architecture should at least be aware of all these options and consider them in a meaningful way when a query is modified.

The progressive architecture we envision allows users to not only execute a single query in a progressive manner, but also to expand their data exploration by issuing additional related queries, which can also be executed in a progressive manner. In this case, the system is making sure that all the progressive queries are collectively optimized and executed. We illustrate this functionality with the following example. Consider a user that explores the census dataset and starts by issuing a progressive SQL query to get overall statistics on the gender distribution. Initial results show that the distribution is heavily skewed towards females, so the user decides to issue a second progressive SQL query to find out the distribution of females over different geographic areas. This distribution being uniform prompts the user to issue a third progressive query that computes the distribution of females for different age groups. At this point the system is running in parallel three progressive queries, where the last two are dependent on the first one. Notice, though, that they do not replace the original query – they run alongside with the original query. This difference is captured in SQL with the FORK QUERY statement which permits more extensive changes compared to ALTER QUERY. Essentially, FORK QUERY creates a graph of query dependency relationships that have to be all satisfied. The system is responsible for creating and running a query execution plan that includes all the queries in the graph, thus optimizing the overall system resource usage. This is a complicated problem because several decisions have to be made. First, the queries have to be prioritized. A simple approach is to give preference to the newer queries over the older ones. While this makes sense if we assume that the attention span of the user is skewed towards the immediate past, we may miss or delay significant trends for the older queries. The second decision is to extract the queries that can processed concurrently and assign them the required resources. The end goal is for the user to make fast progress on the data exploration and analysis tasks, which is possible since they are able to quickly formulate and check different hypothesis, without having to wait for each query to complete execution.

The idea of progressive multi-queries can be pushed one step further by trying to speculate on how the next query may look like. For example, consider a progressive histogram over an attribute of a table in the database. An initial progressive histogram result may be coarse-grained, with a fixed number–say 2 or 10–of bins. Progressive results may consist of finer-grained bins, such as increasing from 2 to 4 or from 10 to 100 bins. We can speculate that such a finer-grained histogram will be requested and begin processing through a speculative forked query of the initial query. Since the overlap between the two queries is significant, most of the computation can be reused across them. If the database knew that the client would change the query in this or a similar way, the database could have served more accurate results faster by scheduling and beginning executing the forked query before the client even made the query. In the course of executing progressive queries, there is a clear expectation that subsequent queries will be made to the database. Moreover, these queries are likely to be similar to the queries before and after. This similarity may come in many forms, for example, a changed WHERE-clause parameter, group-by attribute, or addition or removal of a WHERE-clause. As such, the progressive architecture can "learn" the query sequence from past user sessions and predict the next queries in the sequence with high probability. The next step is to incorporate them into the query graph and schedule them for processing. This can be easily achieved within the progressive multi-query architecture by setting the priorities accordingly.

4.3.4 Conclusions

This report identifies the main components of a progressive computing architecture starting from the functional requirements of a visual analytics system and their performance constraints. This architecture is designed within the confines of modern databases, rather than proposing something from scratch. Nonetheless, all the elements are added as novel SQL constructs. In order to generate progressive results, non-blocking operators have to avoided. For these results to be statistically significant, the processing order matters. Progressive queries can take the form of delta modifications to the query statement, concurrent multi-queries spawned by progressive results, and speculative queries "learned" from prior investigations. They can all be modeled as a query relationship graph and optimized accordingly. We believe that the proposed architecture represents a first step for a fully progressive visual analytics system and opens many research directions in this area.

4.4 Evaluation and Benchmarks for Progressive Analytics

Sriram Karthik Badam (University of Maryland – College Park, US), Jaemin Jo (Seoul National University, KR), Stefan Manegold (CWI – Amsterdam, NL), and Hannes Mühleisen (CWI – Amsterdam, NL)

Progressive data analysis system design contains several degrees of freedom. System proposals require evidence to demonstrate their feasibility. The dynamic nature of progressive systems poses new challenges for the creation of this evidence, since two previously disconnected classes of systems are involved: visualization and data processing.

The approaches to evaluate visualization and data processing systems are fundamentally different and incompatible. Evaluating visualization systems usually borrows methodologies from Human-Computer Interaction (HCI), while data processing utilizes standardized performance benchmarks. Investigating their performance individually is near-meaningless, and an integrated evaluation is necessary to demonstrate the value of a progressive data analysis system.

In the context of progressive visual analytics, success criteria for users can for example be (1) the cognitive ability of participants to interpret uncertainty, (2) behavioral and psychological factors (e.g., keeping the attention and limiting fatigue), (3) the visibility of changes during computation, (4) how informative early intermediate results are, and 5) the speed and accuracy of insights that human can derive from progressive visualization. For data processing systems benchmarking, visual analytics performance can include (1) metrics of the indication of uncertainty in intermediate results, (2) bounds on time until the first approximation is available, and (3) the stability of those intermediate results.

However, how informative early intermediate results are and the time until those are first available are very much at odds with each other: If more time is spent on computation, these results are likely to improve in quality, but will make the user wait longer, which might overstress their patience and thus their ability to derive knowledge. What is needed is a connection of user and task goals with specific system performance metrics. This can for example be compared with the 100ms time bound commonly cited as the threshold for interactivity which is an HCI-derived metric that has direct impact on system performance requirements.

We propose an integrated evaluation and benchmarking framework that reconciles the different methods and metrics of visualizations and data processing for visual analytics. Given a specific task, the framework adopts an iterative approach for reconciliation. Each iteration consists of two phases: (1) visualization-to-algorithm (V2A) reflection and (2) algorithm-to-visualization (A2V) reflection. In the V2A reflection, metrics and limits relevant for usefulness of a progressive visualization to humans are derived. From these metrics, data processing algorithms are selected and configured. Then, the A2V reflection follows where the specific properties of those algorithms in turn are expressed in terms of a set of different metrics and limits. In turn, they are fed back as clarifications into the progressive visualization experiments, again possibly influencing metrics and limits. In addition, local iteration is possible using the set of metrics and limits from the previous global iteration, allowing for example to optimize data processing for a specific response time or – after performance bounds are known – to experimentally determine the best result presentation.

4.4.1 Scenario

Greg wants to dispatch emergency services in a flooding disaster response coordination center by monitoring the locations and times of emergency calls. The amount of calls is overwhelming the capabilities of aid resources such as search teams. Hence, prioritization is required to send aid to where it can be most effective. A progressive visual analysis system supports Greg in this critical decision. Calls and other real-time information are aggregated and visualized in an interactive map. This map also shows detailed topographical information and population density data since lower and crowded areas are more likely to need attention by search teams. Due the large volume of static and incoming data the visualization and the underlying fluid dynamics models that predict water movement suffer from a large computational delay of over fifteen minutes.

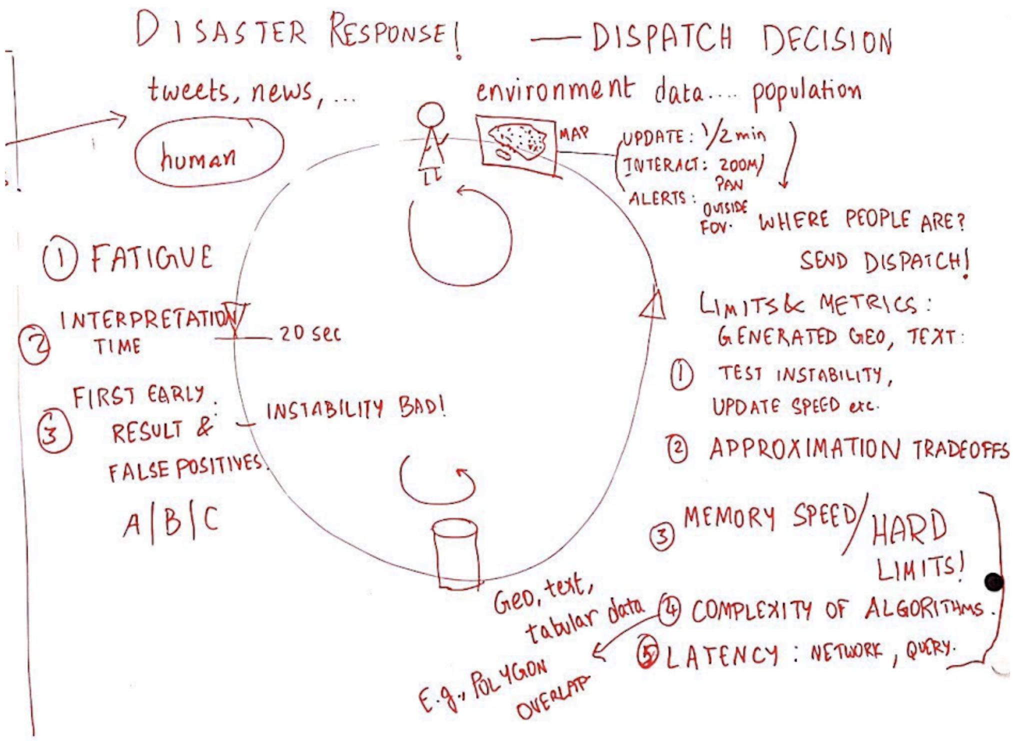

Figure 1 An example evaluation scenario.

The measures that can be used to evaluate the effectiveness of this visualization system can be (1) human fatigue resulted from monitoring and (2) time before decisions can be made. A first round of user study experimentation reveals that the computational delay of 15 minutes is not acceptable in this disaster response scenario, requiring an approximate approach in computation. At the same time, avoiding user fatigue and confusion dictates a large degree of visual stability in the displayed information.

Based on these requirements, a progressive algorithm is chosen that iteratively improves result while providing early intermediate results for early and continuously improving visualization (i.e., V2A reflection). This is made possible by limiting iterations through the fluid dynamics model and using an approximate polygon overlap algorithm to integrate topographic information. Unfortunately, this introduces a fair bit of uncertainty and volatility in the early responses from the system. It is found that sometimes false positives are created.

The possible false positives threaten the utility of the entire system. If operators cannot trust the displayed incidents, they may be reluctant to dispatch aid. This limitation from false-positive cases can be handled by visualizing the uncertainty and allowing a weighing of displayed incidents w.r.t. their certainty and their severity. For example, an incident involving one person with 100% certainty might be weighed against an incident involving 100 people with 80% certainty. As another V2A reflection cycle, the algorithm is extended to provide the uncertainty of the intermediate results.

The design of visualization is modified to show the uncertainty computed by the algorithm (i.e., the A2V reflection). This raises another aspect for evaluation such as how well the user can interpret the uncertainty and how correct the decision made during analysis is.

We have described a framework to reconcile the competing and sometimes incompatible goals between progressive visualization and the required data processing systems. We have shown that through iterating over their respective metrics and limits we can derive both a compromise in their requirements as well as usable bounds for their evaluation and benchmarking. Regardless, open questions remain concerning the more general questions of how various metrics can be quantified and qualified (e.g., through experimental protocols concerning the question of trust). Similarly, the implications of "steering" a computation through interaction are not clear yet, while this could be simulated through restarts, already computed information might be re-used to speed up computation. Immediate next steps would be to investigate how visualization and data processing metrics are connected to each other concretely, what choices exist to achieve trade-offs between them, and how perform various different types of studies in the evaluation loop, e.g., formative or summative studies in HCI.

4.5 Uncertainty in Progressive Data Analysis

Sriram Karthik Badam (University of Maryland – College Park, US), Marco Angelini (Sapienza University of Rome, IT), Jean-Daniel Fekete (INRIA Saclay – Orsay, FR), Barbara Hammer (Universität Bielefeld, DE), Christopher M. Jermaine (Rice University – Houston, US), Hannes Mühleisen (CWI – Amsterdam, NL), Cagatay Turkay (City – University of London, GB), Frank van Ham (IBM Netherlands – Weert, NL), Anna Vilanova (TU Delft, NL), and Yunhai Wang (Shandong University – Jinan, CN)

4.5.1 Sources of uncertainty

In the progressive visual analytics pipeline, we can find different sources of uncertainty on the different steps of the pipeline. Initially, there is uncertainty in the data itself (e.g., measuring errors, sampled population). The algorithm can produce uncertain results independently of being progressive (e.g., local minima tSNE embedding). There is also a source of uncertainty introduced when an algorithm is made progressive. Furthermore, the visual encoding and visualization introduce uncertainty given, for example, the representation chosen or the resolution of the screen. There is also uncertainty associated to the perception and interpretation bias of the user on the visualization. The task to develop might also be uncertain. Finally, if uncertainty is estimated, there is also uncertainty in the uncertainty estimate, e.g., assumptions needed for the estimation.

All sources of uncertainty are relevant for the final decision-making process of the user. For this report, we focus on the uncertainty specific to progressive VA algorithms, and not present to other pipelines. We focus, therefore, on the uncertainty introduced due to making progressive the underlying algorithm. There is substantial amount of research studying each one of the sources of uncertainty named above. All uncertainties should be considered in combination with the progressive uncertainty, it is important for the final decision making process of the user. For example, it makes no sense to very accurately calculate a result for a data set that is very uncertain.

We can consider that there are two main ways to make an algorithm or method progressive:

- **Data based** uses the full data but rather a subsets or an aggregated version.
  - Sampling approximates the data fed to the method, for example by a sampling strategy.
  - Aggregation approximates by building aggregation data structures to facilitate the calculations, e.g., hierarchies.

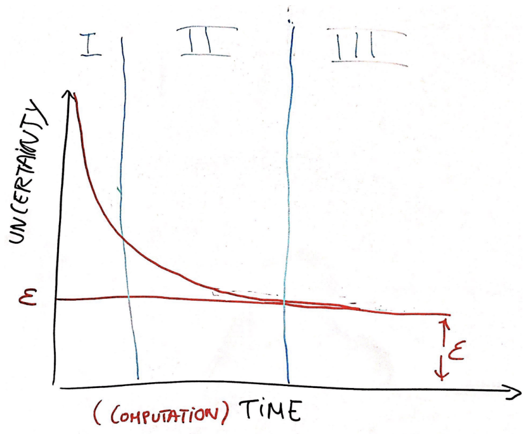

Figure 2 The three phases in progression.

- **Computationally based** approximates the computations themselves, for example, using heuristics.

4.5.2 Model of uncertainty in progression (Phases)

The stage of progression has a say over the uncertainty and how the user responds to it. To model uncertainty, we considered three phases during the progressive analysis process. Our current characterization is based upon certain assumptions about the data as well as the analytical process.

Assumptions: Our first assumption is that the progressive analysis system works with fixed data. This implies the world of the system is closed and there is no new data added during the analytical process. This assumption implies that in an ideal situation, the uncertainty decreases with the computational progress and converges to a minimum.

Beyond this, we focus on hypothesis generation for modeling the uncertainty. During hypothesis generation, the analyst is still open and undirected–exploring patterns and outliers that pop up from the data. In contrast, hypothesis confirmation creates more challenges as the analysts need to collate all the steps of the progression to actually confirm or reject their hypothesis with a certain confidence.

Based on these assumptions, we characterize three phases (Figure 2) in progressive analysis that showcase different effects of uncertainty on the user.

1. Phase I: The first phase represents an early part of the progressive analysis, where the estimates are improving rapidly. This phase is very uncertain as the system is just beginning to process the data either in batches or through iterations of an algorithm. This phase should help the analyst evaluate the methods to find major mistakes and fix parameters to adapt the analysis or restart it. Therefore this is the phase of **estimating suitability**. This is not the time to make concrete insights or even decisions from the data.
2. Phase II: During this phase, the analyst sees clear patterns in the progression capturing some artifacts within the data. The certainty or confidence increases and the analysts can develop early insights. Therefore this is the phase of **early response**. However at this stage, there is a tradeoff between time and value: the longer an analyst waits the more certain they can be. It is the responsibility of the system to balance the time and value to fit the constraints of the analyst.

3. Phase III: This is the phase of **convergence**. The uncertainty has stabilized to its minimum, in some cases it can approach zero at the end. The added value of watching the progression in this phase is little as the changes are minimal and the analyst already has enough understanding to make decisions.

This uncertainty evolution model represents the ideal world–what we hope to achieve. However, quantifying uncertainty from the system to represent such evolution is a complex problem. In fact, this problem needs to be resolved both from the data end–by creating measures that are driven by the user tasks–and from the visualization end–by creating visual representations that can capture the uncertainty of the intermediate estimates in these phases. In reality, there can be a synthesis of these two methods to convey a concrete picture of uncertainty to the user.

4.5.3 Measures of uncertainty

Given the considerations made on the model of uncertainty in progression, in this section we explain the different ways in which uncertainty can be computed. The definition of progressive uncertainty requires to consider the characteristics of the dataset on which the progression is applied. Starting from this characteristics a hierarchy of ways to compute uncertainty can be defined, reported in the following list of measures that are relevant to the decision making in progressive visual analytics:

1. Quality: For obtaining a measure of uncertainty it is needed that, for the original data some statistical properties are known, e.g., distribution of the data (provided by the DB in which they are stored), or assumptions on the statistical properties of the data (e.g., independent identically distributed, i.i.d., that allows to estimate a distribution for the data. In this case the quality at the various stages of the progression can be measured comparing the processed data with the distribution, obtaining confidence intervals.
2. Stability: if the distribution of the data is not known, it is not possible to measure quantitatively uncertainty, so a different indicator must be used based on the data computed up to the actual stage of progression. This indicator can be stability, that effectively estimate when a process is globally in a stable state; Stability measures are based on the computation and aggregation of difference between the values assigned to visual elements in two consecutive stages of the progression. If this measure is minimized under a certain threshold defined by the user/system for a certain sequence of stages of length N, the process can be considered stable.
3. Progress: Stability can be always computed independently from the task or dataset, but in the case in which the progressive computation never converges toward a stable state it can be used only as an evidence that the result is still uncertain. In this case the definition of uncertainty is again relaxed toward the concept of progress of the progressive computation. The computation of the progress is based on the assumption that the end of the progressive process can be computed or estimated.
   In order to allow this estimation we defined two methods:
   - When the problem is data-bounded (e.g., computation of average value) the progress is estimated with respect to the percentage of the total data that has been processed up to the actual stage of the computation.
   - When the problem is process-bounded (e.g., clustering) we make the assumption that the complexity bound of the particular algorithm used is used to estimate its termination time.

Given the availability of this information it is possible to estimate at every stage of the progressive computation how much it lasts to its natural end, obtaining the actual progress.

As an additional consideration it is important to note that each measure of this hierarchy can be evaluated in a Global fashion (on the whole progressive process) or in a local fashion based on the selection of areas of interest for the user. So while in order to use a level of the hierarchy as uncertainty measure all the relative assumptions must be valid for both global/local fashions, the decision-making process of the user could be based on different granularity. As an example, if it is possible to define the stability concept, it would be possible to have a progressive computation that at a generic stage m is globally not stable, while it is locally stable in the area of interest of the user.

4.5.4 Visual Encoding Characteristics

Uncertainty visualization is an active field of research within the visualization community. Visualization of uncertainty has as main challenge to identify effective visual encodings in a usually already cluttered display. Several visual encodings exist (e.g., textures, blurring, box plots) and their uncertainty perception effectiveness are being studied. An effective visual encoding is also data and task dependent. We will not deal with this general issues in uncertainty visualization here, we will focus on the specific characteristics that a visual encoding of progressive uncertainty should address. The main identified characteristics are as follows:

- The encoding should consider the notion of change which is an essential part of progressive methods. For example, enhance deviation in consecutive results, and provide visual indications of stability.
- In a general setting it is interesting to be able to identify differences of uncertainty in different areas of the data (i.e., locally), as well as having an overview of the overall uncertainty (i.e., global).
- The decision-making process needs to be able to combine progress, stability and quality. Therefore, the visual encoding should include this three aspects of the uncertainty.
- Bounded estimations will not always be possible, visual encodings that are flexible and allow to include soft-bounded representations for progression,quality and stability are relevant.
- The visual encoding based on animation/evolution should be taken with care to not mislead the user. It can also distract the user towards fast small changing areas which might blind from slow large changes.

4.5.5 Use Cases

In this section different scenarios are reported with respect to the characteristics of the uncertainty modelled in the previous sections.

4.5.5.1 (1) Information about distribution of data are available.

(progress – stability – quality) Dataset is composed of sales data of a company, and the task is to compute the Top 3 products that sold the most. Given the knowledge of the distribution, the progressive system computes data in chunks, and at each stage it produces a top 3 for which it is measured the relative quality/uncertainty with a confidence interval. At phase 1 the confidence interval will still be too large, so the user wait for the computation to unfold

more. The confidence interval will shrinks at every stage to the point in which at some stage in phase 2 it becomes acceptable for the user, Eventually the user will decide if it even needs it to stabilize or not (still in phase 2) for her final decision.

4.5.5.2 (2) Information about distribution of data are not available.

(progress – stability) The dataset and task are the same of Scenario 1, but this time the distribution of the data is not available. The progressive computation is still started on chunks of data and enter in phase 1. At this point no measure of quality/uncertainty can be computed, and the Stability and Progress are the only two measures available. The user, depending on their trends, choose to base her decision on the stability of the expected Top 3 (if it does not vary after several stages) and progress (proportion of data left to be computed). If the process does not become stable, user uses the Progress to make the decision.

4.5.5.3 (3) Stability is not definable.

(progress – quality) The dataset is based on a timely data processing coming from taxi data in New York City. The task is to identifying the taxi stops that earns more money during a month. In these conditions it is probable that even the identification of stable elements through the various stages of the progression could originate from natural skewness in the data (timestamps) and the stability measure would not result in a reliable indicator for decision-making. At that point the progress (quantity of processed data) is the only indicator that can be used to support the decision-making process.

4.5.5.4 (4) Identify groups in a dataset without knowledge of the groups or the data.

(progress, but contains visual indicators of stability) This scenario represents an exploratory analysis application in which the progression contributes to finding groups within a dataset. This is driven by a computation process that progressively outputs clusters from data (e.g., similarity of tweets) or classifies the data in presence of ground truth (e.g., identifying digits in MNIST). These groups are represented in a progressive visualization that highlights them. In terms of the three metrics described before, the progress can be concretely captured by following the iterations of the computations or the number of the batches processed within the data. This holds by assuming the developer bounded the number of iterations, as common for many ML toolkits. The stability and quality of the computation can be hard to quantify or even computationally expensive in some cases within this scenario. For example, computing the inter-cluster distance for k-means as the quality measure is computationally expensive. Therefore, stability and quality can only be visually assessed in this situation. It becomes the responsibility of the progressive visualization to indicate the changes and convey a sense of stability to the user through the global and local highlights.

4.5.5.5 (5) Pattern search in a high dimensional space

(progress only) Pattern search tasks are particularly challenging for progressive analysis. An example scenario for this task is identifying particular patterns of event sequences from a large event data. When dealing with pattern search, measuring progress is still possible when the event sequences are processed in batches. The progress can be conveyed to the user. However, the stability and quality of the search is hard to quantify. In simple terms, it can be hard to know how many more sequences may match the expected pattern without iterating through the rest of the data. Quality is also not possible to access as search represents a concrete outcome – hit or miss.

4.5.6 Open Questions

1. How to define the scheme in a general setting of progressiveness?
2. How best to convey all the uncertainty heuristics to the user to help decision making?
3. User and task dependency on the influence of uncertainty in the decision making process?

4.6 Threats to Research in Progressive Analytics and Visualization

Dominik Moritz (University of Washington – Seattle, US), Christopher M. Jermaine (Rice University – Houston, US), and Emanuel Zgraggen (MIT – Cambridge, US)

4.6.1 Introduction

Progressive analytics provides users with increasingly useful results with the results eventually converging to the precise answer. The claimed benefit is that users can catch errors early and make decisions based on the approximate results while not having to a priori decide on a necessary accuracy. Users can watch the results change until they are confident enough in the results. In this document, we describe threats to this ideologically pure vision of progressive analytics and visualization. These threats are meant as warnings against blindly working on this vision, and are designed to encourage critical reflection on what is needed to benefit users and not accidentally lead to confusion or false discoveries.

There are three main groups of threats that we identified. First, designing and implementing progressive systems has costs, and using them has additional overhead in terms of both computation and user effort. These costs may not outweigh the benefits, and there is a threat of producing over-engineered systems. Second, communicating uncertainty of approximations is challenging. Continuously-changing visualizations (aka. "dancing bars") can add an additional mental burden to users, and can lead to incorrect interpretations of uncertainty. Third, many variations on the idea of progressiveness have been well-studied by prior research, and methods exist that can address some of the concerns that lead to the idea of designing progressive systems in the first place. Hence, there is a danger that the development of progressive systems will be dismissed by academia or industry as unnecessary or not novel.

The next sections describe the three main threats to research on progressive analytics and visualization.

4.6.2 Cost vs. Benefit

We argue that a potential threat to the progressive analytics research agenda presented in this report is that cost to achieve it could outweigh the benefits.

Creating "pure" progressive systems, where users are continually presented with updated and refined results, are expensive to research, build, and run. The architecture of data management systems and query languages has to be fundamentally changed and machine learning, data mining and other algorithms need to be redesigned to be usable in a progressive setting. Furthermore, adding support for progressive results can increase the overall runtime, and therefore computation cost, of algorithms. For example, a blocking join implementation can use only a single hash table and thus use less memory and computer resources. Implementing

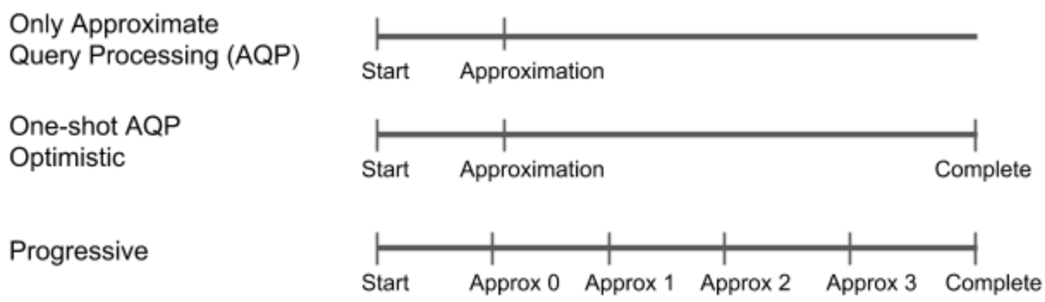

Figure 3 *Only Approximate Query Processing and One-Shot AQP Optimistic* are both approaches that can be implemented on top of existing databases, hence their development cost is low. A pure Progressive system requires more substantial changes, and therefore higher cost. We need to carefully analyze in which cases paying this cost is justified.

queries in a way so that they progressively converge to the true result adds overhead over approximation that does not have to make this guarantee.

Additionally, the progressive paradigm inflicts a cost on users: they need to learn how to use, interpret and interact with these systems. Progressive systems require users to understand and correctly interpret uncertainty, and progressively updating results are potentially distracting and increase the cognitive workload of users.

We urge caution before addressing these complex topics and creating engineering and research agendas around them. We believe that we must first gain a better understanding when progressiveness is really needed, critical, and clearly provides added value over perhaps simpler or existing solutions.

For example, one-shot sampling (i.e., approximation over a single sample) is already heavily used by data scientists to debug, test, and tune their machine learning pipelines before running them on the entire dataset. Such sampling techniques are much simpler to implement on top of existing systems without fundamental redesigns and therefore impose almost no additional cost. For a data scientist described in this example: does progressive analytics provide enough of an additional benefit to justify the redesign of widely used ML packages? As a research community it is important for us to find use cases and scenarios where progressiveness provably justifies its cost and compare and contrast it to other approaches.

4.6.3 Uncertainty

The vast majority of proposed methods for facilitating progressive computations utilize some sort of randomness or statistical inference, and hence they may result in uncertainty as to the final answer to the computation. Various statistical measures of uncertainty such as the standard error, bias, and variance, of the error distribution are computed to describe how the estimate relates to the correct answer. Since these are abstract measures of uncertainty that are often ill-understood by users, the standard approach to date as been to translate these metrics into confidence bounds on the answer to the computation, and present those bounds to the user as a measure of the uncertainty of the current guess as to the result of the computation. A confidence bound is a statement of the form, “With a probability of p%, the real answer is in the range x +- ϵ.” Confidence bounds are appealing because they have a natural, graphical representation and can easily be animated, showing shrinking or converging bounds when visualizing a progressive computation.

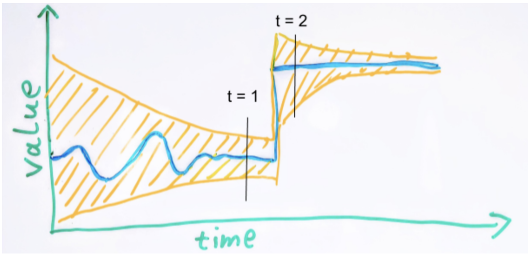

Figure 4 Estimation of a value (blue) over time with confidence bounds (yellow).

However, there are significant concerns regarding the use of confidence bounds in a progressive computation, relating to the fact that in an animation, a sequence of confidence bounds are presented to the user, and not just a single bound.

The first concern is that displaying an animation results in an instance of the so-called *multiple hypothesis testing* (MHT) problem. When ignored, MHT results in so-called "p-hacking." Imagine that a user wants to answer the question: "Were the profits in Europe in 2017 less than \$2.8 million?" She watches a visualization with 95% confidence bounds that progressively shrink over time, until the entire confidence region is above \$2.8 million, and she stops the computation. The problem is that the user may have been shown 100 different 95% confidence bounds in sequence, only the last of which excluded the \$2.8 million cutoff, and concluded that there was only a 5% chance that the profits were under \$2.8 million. In fact, rejecting the hypothesis with a p-value of 0.05 is not implied by the observed bounds, as the user has tested the "less than \$2.8 million" hypotheses 100 separate times, and found it to be rejected only one time. Depending upon the correlation between the various estimates, the "real" p-value may be much, much higher than 0.05.

A second problem regarding showing a series of estimates and bounds to a user is that the user may view the sequence of estimates and bounds as data to be interpreted, when there is no statistical basis for doing so. Consider the picture in Figure 4, showing an estimate and 95% confidence bounds as a function of time. A user may watch a visualization of the bounds and conclude at time $t = 1$ that they have converged. After all, at that time, the bounds have focused in on a particular value for some time, and now appear stable. However, there is no statistical basis for this conclusion. No matter how stable the bounds appear, the bounds still allow for a 5% chance that there will be a jump as shown at time $t = 2$. In effect, the animation itself has made it easy for a user to incorrectly draw conclusions about the accuracy of the computation.

A final problem with uncertainty in the context of progressiveness is that there is "uncertainty in the uncertainty." That is, statistical measures of uncertainty such as confidence bounds almost always depend upon estimating the variance of an error distribution, or else using a simulation-based method such as the bootstrap (typically, estimating the variance of an error distribution is as difficult as estimating the answer to the computation, and so computing it exactly is not feasible). Thus, the bounds themselves are actually an estimate, and may be incorrect. In practice, displayed bounds can shrink for some time, only to get wider (showing more uncertainty) over time as more data are processed. This may happen even though the variance or standard error of the underlying computation can be formally shown to monotonically decrease over time (this is commonly observed when using the standard, ripple-join style estimator for aggregate queries over relational joins). The concern is that a user may make a decision based upon confidence bounds that are tight only by virtue of themselves being a high-variance estimate for the "real" bounds.

4.6.4 Novelty

The three main use cases for progressive analytics and visualizations are massive datasets where full computation takes too long, monitoring and incremental computation of algorithms, and illustrating algorithms step by step. Some potential consumers of progressive technology may argue that these usage scenarios can all be addressed with existing ideas and research efforts (such as approximate query processing and anytime algorithms) and do not require a new research topic called "progressiveness."

For example, in exploratory analysis, users create visualizations, interpret the results, and then decide to refine their query or ask a new question. Speed is crucial to maintain attention and to ensure that the analyst will not lose their train of thought. However, while speed is crucial, perfect accuracy is often not necessary. For example, an analyst may have queried for the sum of values and only upon seeing the result notices that they actually wanted an average. With quick results, they can easily correct this mistake. However, instead of progressiveness, users can just use an approximation based on a sample of the dataset. Approximate query processing (AQP) is an established research area in databases. Data scientists already typically develop their analysis pipelines on a sample before running it on the full dataset, and may question the need for a fully progressive system.

For another example, monitoring machine learning algorithms allows model creators to see when their model converges, and to determine whether there are potential problems with the data or the model parameters. The machine learning community has long investigated anytime algorithms where users can "pull" results at any time, and monitoring is built into most machine learning frameworks. Incremental computation to update a model when new data arrives has also been an area of extensive research in the algorithms and machine learning communities. There is concern that these problems have already been studied, and gathering these ideas under a new umbrella called "progressive computation" does not add additional value.

Finally, the developing visualizations of computations that evolve over time, with the goal of giving people insight into the inner workings of a computation or to potentially steer the computation, is not new. In fact, it is a well-known idea, generally referred to as "algorithm visualization." Again, this may lead to questions regarding the necessity of branding a class of related ideas as "progressive computations."

4.6.5 Conclusion

In this document, we describe three threats to the area of progressive analytics and visualization. We do not see these threats as existential threats but as questions that will come up in grant reviews and challenges that should be considered. We hope that considering these threats encourages the community to investigate the design space between static and progressive. We should investigate how exposure to incorrect approximate results influences decision making even when users see the precise results later (e.g., anchoring effect). Working on these challenges may help us to overcome the threats outlines in this document.

Since we are lacking practical and usable progressive systems today, we should focus carefully on what is already possible with existing systems, and what cannot be done with them. Along the way, we may find "killer use cases" that convincingly show that alternatives like just approximation are not sufficient and proof the need for progressiveness. Before we strive for the "pure" progressiveness of continuously updating results, we should investigate better tool support for users today. Analysts already use samples and there is a need for better tools to communicate uncertainty, help users make decisions in the face of uncertainty, select appropriate samples for selective queries, and better support for incremental update to models when only one parameter changes.

4.7 Progressive Data Science: Potential and Challenges

Cagatay Turkay (City – University of London, GB), Carsten Binnig (TU Darmstadt, DE), Jean-Daniel Fekete (INRIA Saclay – Orsay, FR), Barbara Hammer (Universität Bielefeld, DE), Daniel A. Keim (Universität Konstanz, DE), Themis Palpanas (Paris Descartes University, FR), Nicola Pezzotti (TU Delft, NL), Florin Rusu (University of California – Merced, US), Hendrik Strobelt (IBM TJ Watson Research Center – Cambridge, US), and Yunhai Wang (Shandong University – Jinan, CN)

Data science often requires time-consuming iterative manual activities. In particular, the early activities including data selection, preprocessing, transformation and mining, highly depend on iterative trial-and-error processes that could be sped up significantly by quick feedback on the impact of changes. The idea of progressive data science is to compute the results of changes in a progressive manner, returning a first approximation of the results quickly with iterative refinements until converging to the final result. Enabling the user to interact with the intermediate results allows an early detection of wrong or suboptimal choices, the guided definition of modifications to the pipeline and their quick assessment. In this paper, we discuss the progressiveness challenges arising in the different steps of the data science pipeline. We describe how changes in each step of the pipeline affect the following steps and outline why a progressive data science process will help to make the process more effective. Computing progressive approximations of the results resulting from changes creates numerous research challenges, especially if the changes are made in the early steps of the pipeline. We discuss these challenges and then outline first steps towards more progressiveness in the data science process, which will ultimately help to significantly speed-up the data science process. An extended version of this report is available as [1].

References

1 Cagatay Turkay, Nicola Pezzotti, Carsten Binnig, Hendrik Strobelt, Barbara Hammer, Daniel A. Keim, Jean-Daniel Fekete, Themis Palpanas, Yunhai Wang, and Florin Rusu. Progressive Data Science: Potential and Challenges. *arXiv.org e-print*, 812.08032, 2019.

Participants

- Marco Angelini
Sapienza University of Rome, IT
- Michael Aupetit
QCRI – Doha, QA
- Sriram Karthik Badam
University of Maryland – College Park, US
- Carsten Binnig
TU Darmstadt, DE
- Remco Chang
Tufts University – Medford, US
- Jean-Daniel Fekete
INRIA Saclay – Orsay, FR
- Danyel Fisher
Honeycomb – San Francisco, US
- Hans Hagen
TU Kaiserslautern, DE
- Barbara Hammer
Universität Bielefeld, DE
- Christopher M. Jermaine
Rice University – Houston, US
- Jaemin Jo
Seoul National University, KR
- Daniel A. Keim
Universität Konstanz, DE
- Jörn Kohlhammer
Fraunhofer IGD – Darmstadt, DE
- Stefan Manegold
CWI – Amsterdam, NL
- Luana Micallef
University of Copenhagen, DK
- Dominik Moritz
University of Washington – Seattle, US
- Thomas Mühlbacher
VRVis – Wien, AT
- Hannes Mühleisen
CWI – Amsterdam, NL
- Themis Palpanas
Paris Descartes University, FR
- Adam Perer
Carnegie Mellon University – Pittsburgh, US
- Nicola Pezzotti
TU Delft, NL
- Gaëlle Richer
University of Bordeaux, FR
- Florin Rusu
University of California – Merced, US
- Giuseppe Santucci
Sapienza University of Rome, IT
- Hans-Jörg Schulz
Aarhus University, DK
- Michael Sedlmair
Universität Stuttgart, DE
- Charles D. Stolper
Google, US
- Hendrik Strobelt
IBM TJ Watson Research Center – Cambridge, US
- Cagatay Turkay
City – University of London, GB
- Frank van Ham
IBM Netherlands – Weert, NL
- Anna Vilanova
TU Delft, NL
- Yunhai Wang
Shandong University – Jinan, CN
- Chris Weaver
University of Oklahoma – Norman, US
- Emanuel Zgraggen
MIT – Cambridge, US

Report from Dagstuhl Seminar 18412

Encouraging Reproducibility in Scientific Research of the Internet

Edited by

Vaibhav Bajpai[1], Olivier Bonaventure[2], Kimberly Claffy[3], and Daniel Karrenberg[4]

1 **TU München, DE,** bajpaiv@in.tum.de
2 **UC Louvain, BE,** olivier.bonaventure@uclouvain.be
3 **San Diego Supercomputer Center, US,** kc@caida.org
4 **RIPE NCC - Amsterdam, NL,** dfk@ripe.net

Abstract

Reproducibility of research in Computer Science (CS) and in the field of networking in particular is a well-recognized problem. For several reasons, including the sensitive and/or proprietary nature of some Internet measurements, the networking research community pays limited attention to the of reproducibility of results, instead tending to accept papers that appear plausible.

This article summarises a 2.5 day long Dagstuhl seminar on Encouraging Reproducibility in Scientific Research of the Internet held in October 2018. The seminar discussed challenges to improving reproducibility of scientific Internet research, and developed a set of recommendations that we as a community can undertake to initiate a cultural change toward reproducibility of our work. It brought together people both from academia and industry to set expectations and formulate concrete recommendations for reproducible research. This iteration of the seminar was scoped to computer networking research, although the outcomes are likely relevant for a broader audience from multiple interdisciplinary fields.

Seminar October 7–10, 2018 – http://www.dagstuhl.de/18412
2012 ACM Subject Classification Networks
Keywords and phrases Computer Networks, Reproducibility
Digital Object Identifier 10.4230/DagRep.8.10.41

1 Executive Summary

Vaibhav Bajpai
Olivier Bonaventure
Kimberly Claffy
Daniel Karrenberg

Reproducibility in scientific research is a means to not only achieve trustworthiness of results, but it also lowers barriers to technology transition [40] and accelerates science by promoting incentives to data sharing. The networking research community however pays limited attention to the importance of reproducibility of results, instead tending to accept papers that appear plausible. Previous studies [29, 41, 18] have shown that a fraction of published papers release artifacts (such as code and datasets) that are needed to reproduce results. In order to encourage reproducibility of research, practitioners continue [33, 28, 37, 11, 20] to do service to educate the community on the need for this change. To provide incentives to authors, vehicles for publication of software and datasets are also emerging. For instance, Elsevier SoftwareX [3] is a new journal designed to specifically publish software contributions. DataCite [36, 27] provides mechanisms for supporting methods to locate and cite datasets. Community Resource for Archiving Wireless Data (CRAWDAD) [43] and Information Marketplace for Policy and Analysis of Cyber-risk & Trust (IMPACT) Cyber Trust [4] provide an index of existing measurement data to not only enable new research but also advance network science by promoting reproducible research. Traditional conferences bestow best dataset awards and actively solicit submissions that reproduce results. SIGCOMM Computer Communication Review (CCR) allows authors to upload artifacts during paper submission to allow reviewers to check for reproducibility, and relaxes page limits for reproducible papers. Association for Computing Machinery (ACM) has recently introduced a new policy [1] on result and artifact review and badging. The policy identifies a terminology to use to assess results and artifacts. ACM has also initiated a new task force on data, software and reproducibility in publication [8] to understand how ACM can effectively promote reproducibility within the computing research community. National Academies of Sciences, Engineering, and Medicine with a goal to move towards the open science ecosystem has recently (2018) released a report [31] with guidance and concrete recommendations on how to build strategies for achieving open science. The target is to ensure the free availability (and usability) of publications and associated artifacts. The National Science Foundation (NSF) is taking substantial steps [2] in this area whereby submitted proposals are required to provide a results dissemination plan to describe how produced research results are made available to the extent necessary to independently validate the findings. Towards this end, the proposal budget [5] may request funds for the costs of documenting, preparing, publishing or otherwise making available to others the findings and products of the work conducted under the NSF grant. Despite these continued efforts, reproducibility of research exist as an ongoing problem and few papers that reproduce existing research get published [17, 26, 34] in practise.

Goals

In this seminar, we discussed challenges to improving reproducibility of scientific Internet research, developed a set of recommendations that we as a community can undertake to initiate a cultural change toward increased reproducibility of our work. The goal of the seminar was to discuss the questions below and to propose recommendations that would improve the state of reproducibility in computer networking research.

- **What are the challenges with reproducibility?**

  How can researchers (and data providers) navigate concerns with openly sharing datasets? How should we cope with datasets that lack stable ground truth?

The first category of questions tried to identify the challenges with reproducibility [14]. For instance, concerns with openly sharing datasets led to discussions around legal restrictions and the advantages of researchers keeping data private for their own exclusive future use. Another consideration is double-blind review practices, which require that authors expend effort to obfuscate the source of their data. Would this time be better spent documenting the datasets for sharing to enable reproducibility? A "gap analysis" discussion to understand whether the problem is a lack of appropriate venues or lack of stable ground truth, or more broadly a lack of incentive to reproduce research since publishing (and funding) agents tend to prefer novelty was held. There is also the inherent risk of confirmation bias of existing results; discussion of ideas on how to train young researchers to recognize and counter this tendency was sought.

- **What incentives are needed to encourage reproducibility?**

  What can publishers do? What can conference organisation committees do? How can we ensure that reviewers consider reproducibility when reviewing papers? How can we manage and scale the evaluation of artifacts during peer review? Do we need new venues that specifically require reproducibility of the submitted research?

The second category of questions is about incentives. Questions about how publishers can promote reproducibility framed discussions on whether publishers can provide storage for authors to upload data artifacts with the associated paper in digital libraries, or whether mechanisms can be developed to highlight reproducible (and reproduced) papers. Questions on how conference organisation committees can inspire ideas for additional incentives (such as best dataset awards or relaxing page limits) for authors to make research reproducible. We identified questions to add to review forms to ensure reviewers pay attention to reproducibility aspects. This further lead to discussions on whether committees (in parallel to the regular technical program committee) should evaluate artifacts during the conference review process. Should such a committee be composed of purely young researchers or a blend of young and senior researchers? Questions on the need for specific venues triggered discussions on whether high-impact journals need to establish feature topics on reproducibility or devote a dedicated column for papers that reproduce existing research.

- **What tools and systems are available to facilitate reproducibility?**

  How effective are emerging interactive lab notebook tools (e.g., Jupyter) at facilitating reproducibility? Should CS course curricula integrate use of these tools for student projects to help develop skills and habits that enable reproducibility?

The third category of questions attempt to identify and review tools and systems that are available to facilitate reproducibility. Enormous interest has developed recently in tools for recording experimental observations and computational analytics on large data sets. Some researchers now document the entire process for a paper in a Jupyter lab notebook, greatly facilitating reproducibility and extension of the research. The learning curve for these tools may be daunting; we discussed how faculty can evolve CS course curricula to integrate use of these tools for student projects to help develop skills and habits that enable reproducibility.

- **What guidelines or (best practises) are needed to help reproducibility?**

  How can we ensure authors think about reproducibility? What guidelines would assist reviewers in evaluating artifacts?

The fourth category of questions attempts to develop guidelines (or best practises) to promote reproducibility of research. For instance, we discussed what language could be added to Call for Papers (CFP) to encourage authors to describe reproducibility aspects (of both measurements and results) in their paper submissions.

Structure

The seminar lasted 2.5 days. The seminar began with an introductory round where each participant presented one slide to give an overview of their experience that is relevant for the seminar and a set of open questions that the participant wished to discuss during the event. These slides were collected from each participant before the seminar. We had one invited talk (§3.1) that we used as a basis for triggering discussions and identifying areas for group work, while a major portion of the seminar time was dedicated to breakout sessions, whereby participants were split into small groups to discuss specific themes and develop ideas with consensus to propose to larger groups. The morning sessions the following day were dedicated to continuing parallel group work with presentations that reported the outcomes of each breakout session from the previous day. In the afternoons, we dedicated some time for seven minute lightning talks to invite ideas for subsequent breakout sessions. One evening, we had a social dinner activity. The afternoon of the third day was spent reviewing and collecting feedback from the participants and to initiating follow up actions identified during the seminar.

2 Table of Contents

3 Presentations

Participants were encouraged to volunteer for a lightning talk to provide their perspective on the topic and the presentations were intended as a basis for triggering discussions and identifying areas for breakout sessions.

3.1 Hyper papers and Open Co-Authoring

Alberto Dainotti (CAIDA) kicked off the discussion by presenting some history of the scientific publication process. Scientific papers were born as a mean to share novel scientific knowledge. However, over time publications have also become the main metric for career advancement. This shift has influenced the whole publishing process, from the generation of ideas, data and results to how they are shared. He proposed that perhaps there is a need to step back and look at the currently established process for scientific paper authoring and publishing, including conventions and formats, and wondered if there were room for optimization for the good of science and education. For example, have we struck the right balance between "secrecy" and openness? Are there opportunities from recent technologies and collaborative practices that we can leverage to address the following relevant issues: (i) ideas are often kept secret until a paper is published; (ii) studies are often not reproducible; (iii) incremental work is discouraged by lack of incentives and practical barriers; (iv) fixed-layout flat documents have limitations which are not addressed by simply attaching supplemental material. He proposed to explore the concept of "open collaborative hyperpapers": a paper writing paradigm where co-authorship is potentially open to any researcher, using formats and tools that by design incorporate reproducibility and accountability of contributions, enable incremental progress, and allow for experimenting with different models of code/data/paper reviewing. The talk led to a parallel group breakout (§4.1) where the idea was further developed.

3.2 SIGCOMM Reproducibility Workshop and Artifacts Survey

Recent years have shown an increasing awareness of issues of reproducibility of results as an essential part of research. To address this important issue, ACM has introduced a new policy [1] on result and artifacts review and badging. The policy defines the terminology to be used to assess results and artifacts but does not specify the review process or how to make research reproducible. Furthermore, there appears to be an inconsistency with the terminology defined by other scientific communities [7]. This concern led to a parallel group breakout (§4.5) where specifically the ontology of reproducibility was formalised. At SIGCOMM 2017, a workshop was organized with the specific purpose to tackle the issue. The objective was to trigger discussion and activity in order to craft recommendations on how to introduce incentives for authors to share their artifacts, and the details on how to use them, as well as defining the process to be used to evaluate reproducibility. Luigi Iannone (Télécom ParisTech) presented an overview of this workshop [38] and summarized the main discussions and outcomes.

To improve the current state and strengthen ongoing community efforts towards reproducibility, as a followup to the SIGCOMM reproducibility workshop, a survey was conducted among the authors of papers published at leading ACM computer networking conferences in 2017: CoNEXT, ICN, IMC, and SIGCOMM. Damien Saucez (INRIA Sophia Antipolis) presented the current state of artifacts availability and reproducibility based on this survey [23]. The goal of the survey was to serve as a starting point for further discussions to encourage researchers to ease the reproduction of scientific work published within the SIGCOMM community and to inspire program chairs of future conferences to emphasize the importance of reproducible research.

3.3 Artifacts Evaluation Committee & CoNEXT'18 Badges

Damien Saucez (INRIA Sophia Antipolis) introduced the new ACM SIGCOMM Artifacts Evaluation Committees (AEC) (similar to the AEC created in several SIGs or conferences). The objective of the AEC is to evaluate the artifacts of papers accepted at the SIGCOMM sponsored conferences and assign badges (using the badging system [1] recently established by the ACM) to these papers. Every paper of six pages or more that has been published or accepted by SIGCOMM CCR or any of the conferences sponsored by ACM SIGCOMM in 2018 was eligible for artifacts evaluation by the ACM SIGCOMM AEC. Authors submitted a revised version of their accepted paper that also included pointers to their publicly-available artifacts and in the appendix additional information to help the reviewers with evaluation of the artifact. The ACM Digital Library (DL) was updated to attach assigned badges and public reviews to all the badged papers. As of December 2018, 37 volunteers evaluated 32 submitted papers as part of the AEC[1]. Meanwhile, papers recently accepted at CoNEXT 2018 underwent such a badging process. 14 papers out of 32 accepted papers at CoNEXT were evaluated and 12 were awarded with badges with the help of 20 volunteers who evaluated the artifacts of submitted papers. The lessons learned from the evaluation of the these submitted artifacts will be appear as an editorial note that will be published in the beginning of 2019, with all public reviews and the list of badged papers.

3.4 Challenges with Reproducibility

Mirja Kühlewind (ETH Zürich) raised concerns on how the CS culture is receptive to accepting papers that are non-reproducible so long as they appear plausible. In her talk [14], she discussed some of the challenges that hinders authors and reviewers to embrase reproducibility. For instance, lack of dedicated publishing venues reduce the incentives for authors to reproduce existing research. When submitting papers to double-blind venues, authors have to spend time obfuscating the manuscript, which could instead be used to make artifacts available. On the other hand, paper submission systems do not generally allow authors to upload artifacts, which compels reviewers to fetch artifacts from provided locations discounting their anonymity during the review process. There is also lack of incentive to commit significant time to artifact review to ensure reproducibility of research.

Towards this end, she proposed a set of recommendations. For instance, authors can be encouraged to discuss reproducibility considerations in papers. Paper submission systems can provide authors mechanisms to upload artifacts for reviewing purposes and the review forms themselves can be augmented to guide reviewers to think about reproducibility of the submitted manuscript. After acceptance, reproducible papers can be also be highlighted in digital libraries for recognition purposes.

[1] https://sigcomm18ae.hotcrp.com

3.5 Experiences on Reproducing a Routing Security Paper

In this talk, Matthias Wählisch (FU Berlin) reported about his experiences with reproducing a paper about routing security [35]. He identified that data sources that are used to analyze secure inter-domain routing, usually lack sufficient description. He illustrated that the selection of vantage points is crucial within this context. Based on his experiences, he concluded with three observations. First, many authors are afraid to share their tools because they are afraid to reveal mistakes. Second, asking for reproducibility is especially important in inter-disciplinary research as this allows to self-check the level of competence of the reviewers. Finally, the community needs a change in culture: Making mistakes is not preferred but denying mistakes is worse. Resolving these pieces might help to advance reproducible research.

3.6 Reproducibility: A Problem of Economics, Not Science

Henning Schulzrinne (Columbia University) identified that lack of replicability (see [1] for a definition) tends to pollute the knowledge pool. He argued that reproducibility is a matter of aligning incentives, and the incentives all argue against reproducibility. Replicability has a higher opportunity cost, and is associated with a principal-agent problem where the funding bodies may want to encourage replication, but the researchers may not due to lack of incentives. This talk led to a parallel group breakout (§4.4) where the issue of aligning incentives for reproducibility was further discussed.

3.7 (Strict) Reproducibility Considered both Hard and Harmful

Much current discussion of reproducibility focuses on reproduction of the specific, concrete result obtained by the original researcher. John Wroclawski (USC) proposed to focus instead on reproduction of results at the semantic level, with the aim of validating the larger research conclusion of the original work rather than obtaining a precisely identical narrow result. To support this goal, John suggested the use of explicitly specified invariants to bound the region of applicability of a result, and then briefly touch on the range of discovery, specification, and enforcement tools that, if they existed, would facilitate the use of invariants in support of reproducible research. This talk led to a parallel group breakout (§4.5) where the ontology of reproducibility was formalised.

3.8 Interactive Data as a New Publication Model for Journals

Ralph Holz (The University of Sydney) proposed to make changes to the publication model of journals: instead of producing, and being the gatekeeper to the equivalent of a printout, publications should be 'containerised' - like websites, they should be runnable applications, with the real dataset in the background, where users can choose the appropriate form of presentation and even apply filters and make changes to code. The new form of publication combines write-up, dataset, and software in an instance that is playable and reproducible and makes reviews much easier and tractable. He discussed both advantages and possible problems that went into further discussion into the parallel group (§4.1) that converged into the concept of hyperpapers and new publication strategies.

3.9 Reproducibility vs. Measurement Infrastructures

Robert Kisteleki (RIPE NCC) highlighted a few examples of reproducibility issues related to measurement frameworks [15] such as (geographical and topological) biases, (stability) of vantage point allocation [13, 25], timing of the research, and unique properties of measurement infrastructures in relation to capabilities, data access formats, and data anonymisation. This talk led to the development of a parallel group that discussed data access formats (§4.8) in more detail.

3.10 Taming the Complexity of Artifact Reproducibility

Reproducing research results, as it is required for peer review, can be a time-consuming and difficult task. Thomas Zinner (TU Berlin), proposed three approaches to improve how research results can be substantiated and discussed their applicability. The proposals are based on a brief study [22] on evaluation methods (for Software-defined Network (SDN) research) and insights from a comprehensive discussion on reproducibility. The first approach proposes the use of 'meta-artifacts', which he defined as a structured piece of metadata that describes the tools and parameters that are used during the evaluation. He envisioned a community driven database holding a well-functional set of such meta-artifact templates that could assist in the documentation of the evaluation process. The second approach proposes to either share domain-specific 'evaluation environments' or at least establish well-known 'evaluation scenarios'. For instance, a description of traffic patterns and topology to realise a representative campus network could be one such evaluation scenario. The third approach proposes to adopt 'self-provisioning evaluation setups' using Vagrant (or Docker) to help reproduce the circumstances as close as possible to the original experiment. Some of these ideas became input for the parallel group breakout (§4.1) on new publication strategies.

3.11 Towards an Ecosystem of Reproducibility

Quirin Scheitle (TU München) argued that changing the culture in the network measurement field towards more reproducible results requires changes at many elements of the ecosystem, including authors and independent reproducers. The talk explored what these incentives [39] might be and what might be key factors to their success. Examples to get more reproductions may be making them a soft requirement in PhD programs, co-locating replication hackathons at major venues, or turning labs at PhD schools [6] into replication efforts. Some of these ideas were discussed further in the parallel group (§4.4) on creating incentives of reproducibility.

3.12 High-Quality Measurements and Modelling of Packet Processing

There is a trend towards increasing complexity of networked systems, which leads to the challenge to produce high-quality data for reproducible research. Georg Carle (TU München) presented methods to measure in a reproducible manner and to assess the quality of results. He presented four dimensions to assess the quality of measurement data based on precision, accuracy, coverage and scope. Precision and accuracy requires high-quality tools with hardware support for time-stamping and rate control, thereby limiting random and systematic errors of the traffic generation process. With respect to hard real-time guarantees [24], measurement data with high coverage requires various types of artificial load for rare and worst-case system states. Measurement data of higher "scope" characterises widely used hardware/software configurations. The tool MoonGen [21], a dedicated packet generator in broad use by the community, with its hardware-supported measurement capabilities supports these dimensions of high-quality data, thereby allowing network experiments to reliably reproduce measurement results.

3.13 Measuring Mobile Broadband Networks with MONROE

MONROE [9] is a platform for mobile broadband measurements explicitly designed with openness and reproducibility in mind. In this talk, Anna Brunstrom (Karlstad University) introduced MONROE and discussed its design from a reproducibility and repeatability perspective. MONROE measurement nodes are deployed both in fixed locations and on board trains and buses, offering the possibility to measure from the same vantage points or along the same routes over time. The platform is open to external users, allowing researchers to repeat or extend previous measurements. It is built on open source software and open hardware specifications, allowing others to extend the platform or reuse its design. Experiments in MONROE are designed as Docker containers, making them easy to reuse by others in the same or other environments. Rich metadata is available on the platform and saved in the MONROE database to provide context for the measurements. Open data is made available for all results published by the MONROE alliance.

3.14 Reproducible Research: Implications of Roaming in Europe

"Roam like Home" is the initiative of the European Commission (EC) to end the levy of extra charges when roaming within the European region. As a result, people are able to use data services more freely across Europe. This brings the need for operators to provide seamless service for their customers, similar to what they experience in their home country. However, the implications roaming solutions have on performance have not been carefully examined. In this talk, Andra Lutu (Telefónica Research) presented how they leveraged MONROE open source components to build a roaming measurement infrastructure [30] for 16 different mobile networks in six countries across Europe. With this infrastructure, they then measured different aspects of international roaming in 3G and 4G networks, including mobile network configuration, performance characteristics, and content discrimination. To facilitate reproducibility, Andra plans to extend this research into a hyperpaper which was further developed in the parallel group (§4.1) on new publication strategies.

3.15 Observatories for Internet Measurement

Brian Trammell (ETH Zürich) presented a general model for "measurement observatories", which supports comparability, repeatability, and protection of raw measurement data. This model is based around a metadata-first workflow, and normalizers that translate heterogeneous raw data into a common observation schema defined for a given measurement purpose. Metadata attached to raw data, normalizers, analyzers, observations, and queries allow tracking the provenance of each object. This model is validated through the implementation of the observatory for Internet path transparency measurements. The ideas presented in this talk were further developed in the parallel group (§4.8) on data formats.

4 Parallel Group Work

The afternoon sessions were used to discuss certain topics in more depth in smaller groups. This section summarises the discussions of each group.

4.1 New Publication Strategies

Alberto Dainotti, Ralph Holz, Mirja Kühlewind, Andra Lutu, Joel Sommers, Brian Trammell

The group envisioned a publishing ecosystem for Internet science, supporting publications that are self-contained, interactive, multi-level, open, and collaborative. The idea leverages on recent developments in platforms and tools for data science and scientific collaboration to build an experimental publishing ecosystem for Internet measurements based on hyperpapers [10], similar to the WholeTale project [16] that also envisions to unite data products with research articles to create "living publications" (or tales).

Hyperpapers are self-contained and interactive. Ideally, a full hyperpaper contains all the data from which results, plots, and conclusions in the paper are drawn, as well as source code implementing the analytic tasks distilling those results from the raw source data. The paper is interactive, allowing both changes to the raw source data and to the analysis code to be reflected in the analytic products in the paper.

Hyperpapers are multi-level. The initial view of a full hyperpaper includes the typical prose of a paper. Analysis products, such as charts and tables, can be expanded to show how they were derived. However, the paper can also be expanded in other ways. A section of prose may be linked to an alternate view, information for an alternate audience, related content, or a drill down on some interesting set of a result.

The perennial problem of setting up environments for data analysis without needing to replicate a full tool chain with dependencies from scratch is largely solved today by virtualisation and containerization tools such as Vagrant and Docker. Problems of scale are addressed by the easy (if sometimes costly) widespread availability of cloud infrastructure from multiple providers. Integration of data analytics with authoring environment interleaving text and interactive visualizations is supported by data analysis notebooks such as JupyterLab and Apache Zeppelin. GitHub has emerged as the de-facto standard for integrating version control of digital artifacts with a collaboration environment, and its model of working is suited to open collaborative papers, which have a fair amount in common with the long-running open source projects GitHub was originally built to support.

The group identified two main gaps in technical infrastructure necessary for a full initial realization of this vision. First, while some research studies can be done with data or models that can easily be stored in an ad-hoc format within the hyperpaper itself, large-scale Internet measurement studies need access to large data sets mediated through some interface. This exists for certain data sources (such as the RIPE Atlas API), but a full realization would require the creation and standardization of interfaces for retrieval of data and metadata for each broad type of measurement activity. Second, the distribution of rendered versions of papers is currently possible for scientific notebook environments, but these render to a webpage that is not necessarily optimized for accessibility. Tooling to render a view of hyperpaper as a PDF according to the required format for a given venue, is necessary to support the full multi-rendering functionality of the vision above.

An editorial note [19] describing this concept with a call for action to publish demonstrations of hyperpapers and make preparations for an experimental hyperpaper platform recently appeared in SIGCOMM CCR.

4.2 Guidelines for Students

Vaibhav Bajpai, Anna Brunstrom, Anja Feldmann, Wolfgang Kellerer, Aiko Pras, Henning Schulzrinne, Georgios Smaragdakis, Matthias Wählisch, Klaus Wehrle

The group developed guidelines meant for researchers and for students working in experimental networking research early in their career, and as a reminder to others. General best practises on problem formulation and design, documentation, experimentation and data collection and data handling were outlined. For instance, it is essential to formulate the hypothesis, design the experiments to validate (or not) the hypothesis, conduct the necessary experiment, and eventually check the validity of the hypothesis. Planning and soliciting early feedback is crucial in such a workflow, whereby visualisations help to convey early results and help identify anomalies that may need further analysis. One-time experiments are prone to bias by transient effects and dynamism of the operational system in itself, which requires to reiterate the experiment to gain confidence in the results. Documenting all steps and observations (similar to the lab notebook approach common in natural sciences) during experimentation is key for repeatability. The gathered artifacts need to be accompanied with metadata to help understand how the data was created, what it contains and how to recreate it. Embracing version control helps identify regressions in code and analysis to help identify the root cause of the anomaly. It's crucial to keep regular backups to ensure data is safely stored. A good strategy is to run a series of small experiments to verify the tools and validate analysis and then scale up. Identify how not to reinvent the wheel, and which tools are readily available for use in the experiment. During the data collection phase, monitoring should be applied to ensure the smooth running of the experiment to avoid network/disk failures, host reboots, overwritten logs that may distort the data gathering process. It's important to respect the privacy constraints of external datasets that are used in the research. Similarly, before making your datasets available, consult others for any privacy concerns that perhaps could be alleviated by data anonymisation. Furthermore, ensure the integrity of the data to account of observation biases and document them together with the released datasets. Consider how the developed code will be licensed and made available, discuss and form agreement with the team, and perhaps also reach out within your organisation to make yourself aware of the guidelines that may be available. The group went further and also developed specific guidelines by research area, particularly for simulation studies, systems prototyping and evaluations, real-world measurements and subjective experiments along with a recommendations on tools that are generally used in these areas. Pointers to research papers that follow similar guidelines were also identified. An editorial note [12] describing the set of guidelines recently appeared in SIGCOMM CCR.

4.3 Guidelines for Reviewers

Olivier Bonaventure, Luigi Iannone, Daniel Karrenberg, Damien Saucez

This group discussed guidelines for reviewing of artifacts. The aim of the review process is to decide whether to award papers with the relevant ACM badges. The primary document is a form asking a number of questions to the reviewers. The purpose of the form is to structure the discussion among the reviewers. Awarding badges will be a decision based on that discussion. The group realised that this is a work in progress and review methods and standards will converge and best practises be established by the community as we go along. The intent is to provide a starting point. The group started from experience with a review in progress for CONEXT 2018 conference, which already used a form. The group re-worked this form moving some yes/no questions to more differentiated scores that enable reviewers to respond in a more differentiated way. The group also discussed examples across the whole range to provide as guidance to the reviewers. In practice, it is expected that such a review happens between submission of camera ready copy and conference presentation. The rationales are that this allows sufficient time available for review, the paper also becomes immutable after the camera-ready, and badging leads to recognition at the conference. The review form developed by the group is made publicly available[2]. The proposal is to utilise this form and process to review artifacts for upcoming conferences.

4.4 Incentives for Reproducibility

Olivier Bonaventure, Ken Calvert, Luigi Iannone, Damien Saucez, Quirin Scheitle, Jürgen Schönwälder

The group attempted to identify incentives for reproducible papers and workflows to evaluate them. The incentives for independent reproductions of published work and identifying venues for publishing papers reproducing research were studied. Perhaps a carrots (reproducers) and sticks (funding agencies) approach is needed in the long-run to establish a feedback loop to initiate a cultural change. In order to foster a positive view of the process, explicit incentives can be established. The incentives need to be viewed for all players involved. For instance, funding agencies require[3] that results created with their funding are open access with artifacts available (barring cases where data cannot be released due to privacy constraints). Conferences can be made to meet certain reproducibility standards for published papers to get support. Students who try to reproduce results can be handed travel grants to attend conferences and meet authors of papers they reproduce. Publishing venues (such as conferences and journals) can facilitate mechanisms to submit artifacts with published papers and can also integrate them early with the submission process. Reproduction of published results can be made a soft requirement of doctoral studies. A special track in

[2] https://goo.gl/JjXgjw
[3] https://goo.gl/P3L33S

a conference can be established where reproduction reports can be published. Dedicated "repathons" can be organised where authors and reproducers can sign up for attendance backed up by travel support by funding agencies. The repathons can be co-located with regular conferences whereby a list of possible papers available for reproduction are announced in advance. Encouraging reproduction allows authors to develop new ideas with the reproducers and create collaboration possibilities. It also allows the community to build upon the author's work and increases their impact. The badging system helps increase the visibility of work further. Artifact evaluators on the other hand get visibility by being part of the conference committee. The process helps them develop new skills and understanding of the scientific review process.

4.5 Ontology for Reproducibility

Steve Bauer, Georg Carle, Robert Kisteleki, Matt Mathis, Jörg Ott, Karen Sollins, John Wroclawski, Thomas Zinner

The group met to formulate an ontology to be used with reproducibility and their applicability dimensions. The group identified a taxonomy axis composed of observational reproducibility and model (or prediction validation) reproducibility. The former category applies to situations where one is attempting to reproduce a "data collection and analysis" activity. Within this category, the reproducer has no control over the system, but the goal is to collect data with sufficient accuracy to validate the conclusion of an analysis of the data. An example that falls within this category includes real-world measurements on the Internet to understand reality as is, using well-known test-beds. The key reproducibility criteria include vantage point selection, traffic dynamics, methodological description, handling of outliers, and hidden assumptions that need to be documented for increased reproducibility within this category.

The second category applies to situations where one is attempting to reproduce the results of "modeling (or prediction)" activity. The goal is to provide controlled inputs to a system, and then observe how the system responds, so as to observe modeled behavior or validate a predicted result. An example that falls within this category includes simulation experiments that include specific input conditions that do not necessarily arise in the existing system. The key reproducibility criteria include clarity and completeness of description of invariants and dynamic inputs and guidance about the space to explore for predictive study.

Another way to formulate the taxonomy is by identifying the objectives of reproducibility. One objective could be to determine whether a specific result is reproducible, by conceptually keeping all conditions (such as inputs, environment, etc) identical to the original. A different objective could be to reproduce the validity of a result over some range of invariants and inputs. Yet another objective could be on reproducing a methodology to understand whether it is applicable in a certain different circumstance. The key requirement here is to have a precise description of the methodology, along with invariants and some validating test cases.

Yet another way to formulate the taxonomy is by the quality (or strength) of the reproduction activity. For instance, one axes could be the resilience of reproduced result to variations in inputs (or experimental conditions). The precision or degree to which the reproduced result matches the original result could be another axes. The ability to explain (or defend) the proposition of why the reproduced and original results match could be yet another axes. The quality of reproduction is influenced by dynamics of the underlying system. Ideally, reason about appropriate failure to reproduce (vs inappropriate failure to reproduce) in different circumstances should be identified. On the contrary, results that are no longer reproducible for good reason continue to have value for educational use, for re-evaluation with new insights, and for evaluating predictive models.

A completely different view is to ask how effective is a reproducible result as one would go outside from the technical into the non-technical presentation of the result, to the rest of the world. This aspect is particularly important for consumers and policy makers. The clarity of presentation, pre-requisite knowledge and ease of understanding are important factors for considerations to cover this aspect.

4.6 Reproducibility Track

Anna Brunstrom, Georg Carle, Alberto Dainotti, Mirja Kühlewind, Andra Lutu, Damien Saucez, Quirin Scheitle, Georgios Smaragdakis, Matthias Wählisch

The group met with a focus on creating a Call for Papers for a reproducibility track for a conference (or a workshop). The objective of such a reproducibility track is many-fold. For one, the goal is to increase the number of incidents of reproducing a published work. A formal track allows such incidents to be documented in a 2-page reproduction report that can go together with the conference proceedings. The reproduction report shall provide a summary of the original paper, description of the reproduction process and findings including describing challenges with reproducing a certain piece of published result. The authors may be contacted for assistance, but may not become authors of the reproduction report. Existing templates [42] used by universities in their reproducibility seminars can be used as starting point for reproduction reports. The target for a conference (instead of workshop) track allows higher attendance and increases focus on in-depth quality. This also has the side-benefit of giving more visibility to authors that publish reproducible research. Furthermore, the track explicitly allows early-stage researchers to learn not only how to reproduce a published research, but also document and report about it. The track also gives new researchers an opportunity to get in touch or work directly with the potentially more senior authors of the original paper, which can be an additional incentive to participate. Unlike the AEC, where the focus is on replication, the focus of the reproducibility track is on independent reproduction of results by interested students. The process shall also uncover sanity of the employed methodology and eventually recognise the reproducers by recognising their contribution as a citable publication. The AEC badges can be used to identify papers that can be used as input for the reproducibility track. The authors of the original publication shall be allowed to provide a commentary on the reproduction report.

The proposal for a reproducibility track was made at the Internet Measurement Conference (IMC) 2018 conference and is currently under discussion.

4.7 Reproducibility in Post-publication Phase

Olivier Bonaventure, Kimberly Claffy, Luigi Iannone and Henning Schulzrinne

The group met with a goal to understand how reproducibility can be maintained in the post-publication phase. The timeline could be either after the camera-ready phase, or after the conference, or after the publication of the paper in the digital libraries. For instance, Social Science Research Network (SSRN) with minimal vetting allows updates to the paper in the post-publication phase. Certain fields (such as economics) have a notion of working papers instead. Examples of post-publication updates could include either replication concerns or corrections (or extensions) by authors themselves. Journals allow such mechanisms where the editor-in-chief mediates the communication with the authors and may eventually make the decision to publish the letter with (or without) authors' response. The model for conferences is unclear and a mechanism for vetting and mediation needs to be defined. The Association for Computing Machinery (ACM) has a mechanism to post comments on the DL. The Internet Engineering Task Force (IETF) has a mechanism to associate errata with immutable Request for Comments (RFC) since 2002 which are mediated by the IETF Area Director (AD). ACM SIGCOMM CCR also has a similar errata mechanism. A strawman proposal is for the conference steering committee to designate a Point of Contact (PoC) to handle post-publication concerns and also deal with possible misuse scenarios that are mediated with the authors with responses tagged in DL. While, a counter viewpoint is to avoid hierarchical control structures, but instead crowd source the problem by utilising systems that already exist. For instance, StackExchange, which uses reputation metrics to rank most useful answers and ResearchGate, a social network for scientists that other scientific communities have meanwhile adopted.

4.8 Data and Metadata Formats

Ken Calvert, Kimberly Claffy, Alberto Dainotti, Ralph Holz, Daniel Karrenberg, Robert Kisteleki, Joel Sommers, Brian Trammell and John Wroclawski

The group met with the goal to understand how data (and metadata) can be represented in a common understandable way to lower the barriers of collaboration and facilitate reproducibility of research. One aspect that is of particular significance is metadata which allows for search, categorisation and retrieval of the raw data. The metadata describes the context within which measurements were performed, the experimental parameters and overall interpretation of the data and how it was created, and its ownership and access rights. As such, it is essential to treat metadata as a first-class citizen because allowing it to be an afterthought often leads to its neglect. With this in mind, the goal is to drive creation of common tooling of interchange formats and APIs to facilitate the consumer of the data. BGPstream [32] is one such case study (of an API and a library) that provides common access to BGP data from multiple providers. The solution is specifically tailored to BGP datasets only and has seen increasing uptake in the research community. The group wondered whether it would be possible to learn from this effort and build infrastructures for other kinds of datasets with similar success? `traceroute` datasets are one such target type of datasets that are produced by multiple providers: CAIDA Archipelago, RIPE Atlas and Measurement Lab measurement infrastructures. The group plans to take this discussion forward by attempting to build a BGPstream-like architecture for `traceroute` data.

5 Conclusions and Next Steps

Participants with a mix of senior and junior researchers hailing from both academia and industry encouraged fruitful dialogue. A number of future research agendas were recognized. The group working on hyperpapers (§3.1 and §4.1) submitted the idea and call for actions as an editorial note [19] for the SIGCOMM CCR. The AEC plans to review the artifacts of 32 submitted papers with the help of 37 volunteers who evalute the artifacts and prepare public reviews. A report of the AEC activity with public badges will appear in 2019. Damien Saucez lead the activity of badging accepted papers that released artifacts for the CoNEXT 2018 conference. 14 (out of 32 accepted) papers were submitted for evaluation, and 12 papers were badged with the help of 20 volunteers. The badges are marked on the conference webpage[4] and also in the ACM DL. The group working on preparing students to think about reproducibility (§4.2) prepared an editorial note [12] on a beginners guide to reproducibility for experimental networking research which recently appeared in SIGCOMM CCR. The group working on preparing a reproducibility guidelines for reviewers (§4.3) produced a review form which informed the reviewers of the AEC. The group working on a reproducibility track (§4.6) worked with Anja Feldmann who made a proposal for such a reproducibility track at IMC 2018[5].

The organizing team also received valuable feedback. The participants felt productive in the groups and appreciated the continuation of the group activity in the mornings. Participants appreciated that the talks were limited to seven minutes to allow more time for group discussions, but also suggested to encourage presenters to present on topics that increased interactivity.

References

1 ACM Artifact Review and Badging. https://www.acm.org/publications/policies/artifact-review-badging [Online; last accessed 29-December-2018].
2 Computer and Network Systems (CNS): Core Programs. https://www.nsf.gov/pubs/2018/nsf18569/nsf18569.htm [Online; last accessed 29-December-2018].
3 Elsevier SoftwareX. https://www.journals.elsevier.com/softwarex [Online; last accessed 29-December-2018].
4 IMPACT Cyber Trust. https://www.impactcybertrust.org [Online; last accessed 31-December-2018].
5 NSF Proposal & Award Policies & Procedures Guide (PAPPG). https://www.nsf.gov/pubs/policydocs/pappg18 [Online; last accessed 29-December-2018].
6 TMA PhD School on Traffic Monitoring and Analysis. http://tma.ifip.org/2017/phd-school. [Online; accessed 25-December-2018].

[4] https://conferences2.sigcomm.org/co-next/2018
[5] https://conferences.sigcomm.org/imc/2018

7 Toward a Compatible Reproducibility Taxonomy for Computational and Computing Sciences. https://cfwebprod.sandia.gov/cfdocs/CompResearch/docs/SAND2018-11186.pdf [Online; last accessed 31-December-2018].

8 ACM Task Force on Data, Software and Reproducibility in Publication, 2015. https://www.acm.org/publications/task-force-on-data-software-and-reproducibility [Online; last accessed 29-December-2018].

9 Özgü Alay, Andra Lutu, Miguel Peón Quirós, Vincenzo Mancuso, Thomas Hirsch, Kristian Evensen, Audun Fosselie Hansen, Stefan Alfredsson, Jonas Karlsson, Anna Brunström, Ali Safari Khatouni, Marco Mellia, and Marco Ajmone Marsan. Experience: An Open Platform for Experimentation with Commercial Mobile Broadband Networks. In *MOBICOM*, pages 70–78, 2017. doi:10.1145/3117811.3117812.

10 Alberto Dainotti. A Conversation on Hyperpapers and Open Co-authoring. http://www.caida.org/workshops/aims/1803 [Online; last accessed 31-December-2018].

11 Vaibhav Bajpai, Arthur W. Berger, Philip Eardley, Jörg Ott, and Jürgen Schönwälder. Global Measurements: Practice and Experience (Report on Dagstuhl Seminar #16012). *Computer Communication Review*, 46(2):32–39, 2016. doi:10.1145/2935634.2935641.

12 Vaibhav Bajpai, Anna Brunström, Anja Feldmann, Wolfgang Kellerer, Aiko Pras, Henning Schulzrinne, Georgios Smaragdakis, Matthias Wählisch, and Klaus Wehrle. The dagstuhl beginners guide to reproducibility for experimental networking research. *SIGCOMM Computer Communication Review (CCR)*, 49(1):24–30, 2019. doi:10.1145/3314212.3314217.

13 Vaibhav Bajpai, Steffie Jacob Eravuchira, Jürgen Schönwälder, Robert Kisteleki, and Emile Aben. Vantage Point Selection for IPv6 Measurements: Benefits and Limitations of RIPE Atlas Tags. In *IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, pages 37–44, 2017. doi:10.23919/INM.2017.7987262.

14 Vaibhav Bajpai, Mirja Kühlewind, Jörg Ott, Jürgen Schönwälder, Anna Sperotto, and Brian Trammell. Challenges with Reproducibility. In *SIGCOMM 2017 Reproducibility Workshop*, pages 1–4, 2017. doi:10.1145/3097766.3097767.

15 Vaibhav Bajpai and Jürgen Schönwälder. A Survey on Internet Performance Measurement Platforms and Related Standardization Efforts. In *IEEE Communications Surveys and Tutorials*, pages 1313–1341, 2015. doi:10.1109/COMST.2015.2418435.

16 Adam Brinckman, Kyle Chard, Niall Gaffney, Mihael Hategan, Matthew B. Jones, Kacper Kowalik, Sivakumar Kulasekaran, Bertram Ludäscher, Bryce D. Mecum, Jarek Nabrzyski, Victoria Stodden, Ian J. Taylor, Matthew J. Turk, and Kandace Turner. Computing Environments for Reproducibility: Capturing the "Whole Tale". In *Future Generation Computer Systems*, 2018. doi:10.1016/j.future.2017.12.029.

17 Bryan Clark, Todd Deshane, Eli M. Dow, Stephen Evanchik, Matthew Finlayson, Jason Herne, and Jeanna Neefe Matthews. Xen and the Art of Repeated Research. In *USENIX Annual Technical Conference*, pages 135–144, 2004. URL: http://www.usenix.org/publications/library/proceedings/usenix04/tech/freenix/clark.html.

18 Christian S. Collberg and Todd A. Proebsting. Repeatability in Computer Systems Research. In *http://doi.acm.org/10.1145/2812803*, Communications of the ACM, 2016. doi:10.1145/2812803.

19 Alberto Dainotti, Ralph Holz, Mirja Kühlewind, Andra Lutu, Joel Sommers, and Brian Trammell. Open collaborative hyperpapers: a call to action. *SIGCOMM Computer Communication Review (CCR)*, 49(1):31–33, 2019. doi:10.1145/3314212.3314218.

20 Philip Eardley, Marco Mellia, Jörg Ott, Jürgen Schönwälder, and Henning Schulzrinne. Global Measurement Framework (Dagstuhl Seminar 13472). *Dagstuhl Reports*, 3(11):144–153, 2013. doi:10.4230/DagRep.3.11.144.

21 Paul Emmerich, Sebastian Gallenmüller, Daniel Raumer, Florian Wohlfart, and Georg Carle. MoonGen: A Scriptable High-Speed Packet Generator. In *Internet Measurement Conference (IMC)*, pages 275–287, 2015. doi:10.1145/2815675.2815692.

22 Matthias Flittner, Robert Bauer, Amr Rizk, Stefan Geissler, Thomas Zinner, and Martina Zitterbart. Taming the Complexity of Artifact Reproducibility. In *SIGCOMM 2017 Reproducibility Workshop*, 2017. doi:10.1145/3097766.3097770.

23 Matthias Flittner, Mohamed Naoufal Mahfoudi, Damien Saucez, Matthias Wählisch, Luigi Iannone, Vaibhav Bajpai, and Alex Afanasyev. A Survey on Artifacts from CoNEXT, ICN, IMC, and SIGCOMM Conferences in 2017. In *SIGCOMM Computer Communication Review*. ACM, April 2018. doi:10.1145/3211852.3211864.

24 Fabien Geyer and Georg Carle. Network Engineering for Real-time Networks: Comparison of Automotive and Aeronautic Industries Approaches. *IEEE Communications Magazine*, pages 106–112, 2016. doi:10.1109/MCOM.2016.7402269.

25 Thomas Holterbach, Emile Aben, Cristel Pelsser, Randy Bush, and Laurent Vanbever. Measurement Vantage Point Selection Using A Similarity Metric. In *Applied Networking Research Workshop*, pages 1–3, 2017. doi:10.1145/3106328.3106334.

26 Heidi Howard, Malte Schwarzkopf, Anil Madhavapeddy, and Jon Crowcroft. Raft Refloated: Do We Have Consensus? In *Operating Systems Review*, pages 12–21, 2015. doi:10.1145/2723872.2723876.

27 Catherine Jones, Brian Matthews, and Ian Gent. Software reuse, repurposing and reproducibility. In *Proceedings of the 12th International Conference on Digital Preservation, iPRES 2015, Chapel Hill, North Carolina, USA, November 2-6, 2015*, 2015. URL: http://hdl.handle.net/11353/10.429590.

28 Balachander Krishnamurthy, Walter Willinger, Phillipa Gill, and Martin F. Arlitt. A Socratic Method for Validation of Measurement-based Networking Research. In *Computer Communications*, pages 43–53, 2011. doi:10.1016/j.comcom.2010.09.014.

29 Stuart Kurkowski, Tracy Camp, and Michael Colagrosso. MANET Simulation Studies: The Incredibles. In *Mobile Computing and Communications Review*, pages 50–61, 2005. doi:10.1145/1096166.1096174.

30 Anna Maria Mandalari, Andra Lutu, Ana Custura, Ali Safari Khatouni, Özgü Alay, Marcelo Bagnulo, Vaibhav Bajpai, Anna Brunström, Jörg Ott, Marco Mellia, and Gorry Fairhurst. Experience: Implications of Roaming in Europe. In *MOBICOM*, pages 179–189, 2018. doi:10.1145/3241539.3241577.

31 National Academies of Sciences, Engineering, and Medicine. *Open Science by Design: Realizing a Vision for 21st Century Research*. The National Academies Press, 2018. doi:10.17226/25116.

32 Chiara Orsini, Alistair King, Danilo Giordano, Vasileios Giotsas, and Alberto Dainotti. BGPStream: A Software Framework for Live and Historical BGP Data Analysis. In *Internet Measurement Conference (IMC)*, 2016. URL: http://dl.acm.org/citation.cfm?id=2987482.

33 Vern Paxson. Strategies for Sound Internet Measurement. In *Internet Measurement Conference, (IMC)*, pages 263–271, 2004. doi:10.1145/1028788.1028824.

34 Diana Andreea Popescu and Andrew W. Moore. Reproducing Network Experiments in a Time-controlled Emulation Environment. In *Traffic Monitoring and Analysis (TMA)*, 2016. URL: http://tma.ifip.org/2016/papers/tma2016-final10.pdf.

35 Andreas Reuter, Randy Bush, Ítalo S. Cunha, Ethan Katz-Bassett, Thomas C. Schmidt, and Matthias Wählisch. Towards a Rigorous Methodology for Measuring Adoption of RPKI Route Validation and Filtering. In *SIGCOMM Computer Communication Review*, 2018. doi:10.1145/3211852.3211856.

36 Laura Rueda, Martin Fenner, and Patricia Cruse. Datacite: Lessons learned on persistent identifiers for research data. *IJDC*, 11(2):39–47, 2016. doi:10.2218/ijdc.v11i2.421.

37 Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. Ten Simple Rules for Reproducible Computational Research. *PLOS Computational Biology Journal*, 2013. doi:10.1371/journal.pcbi.1003285.

38 Damien Saucez and Luigi Iannone. Thoughts and Recommendations from the ACM SIGCOMM 2017 Reproducibility Workshop. *SIGCOMM Computer Communication Review*, 2018. doi:10.1145/3211852.3211863.

39 Quirin Scheitle, Matthias Wählisch, Oliver Gasser, Thomas C. Schmidt, and Georg Carle. Towards an Ecosystem for Reproducible Research in Computer Networking. In *SIGCOMM Reproducibility Workshop*, pages 5–8, 2017. doi:10.1145/3097766.3097768.

40 Henning Schulzrinne. Networking Research - A Reflection in the Middle Years. In *Computer Communications*, 2018. doi:10.1016/j.comcom.2018.07.001.

41 Patrick Vandewalle, Jelena Kovacevic, and Martin Vetterli. Reproducible Research in Signal Processing. *IEEE Signal Processing Magazine*, 2009. doi:10.1109/MSP.2009.932122.

42 Lisa Yan and Nick McKeown. Learning Networking by Reproducing Research Results. In *Computer Communication Review*, 2017. doi:10.1145/3089262.3089266.

43 Jihwang Yeo, David Kotz, and Tristan Henderson. CRAWDAD: A Community Resource for Archiving Wireless Data at Dartmouth. *SIGCOMM Computer Communication Review*, pages 21–22, 2006. doi:10.1145/1129582.1129588.

Participants

- Vaibhav Bajpai
TU München, DE
- Steve Bauer
MIT – Cambridge, US
- Olivier Bonaventure
UC Louvain, BE
- Anna Brunström
Karlstad University, SE
- Kenneth L. Calvert
University of Kentucky – Lexington, US
- Georg Carle
TU München, DE
- Kimberly Claffy
San Diego Supercomputer Center, US
- Alberto Dainotti
San Diego Supercomputer Center, US
- Anja Feldmann
MPI für Informatik – Saarbrücken, DE
- Ralph Holz
The University of Sydney, AU
- Luigi Iannone
Télécom Paris Tech, FR
- Daniel Karrenberg
RIPE NCC – Amsterdam, NL
- Wolfgang Kellerer
TU München, DE
- Robert Kisteleki
RIPE NCC – Amsterdam, NL
- Mirja Kühlewind
ETH Zürich, CH
- Andra Lutu
Telefónica Research – Barcelona, ES
- Matt Mathis
Google Inc. – Mountain View, US
- Jörg Ott
TU München, DE
- Aiko Pras
University of Twente, NL
- Damien Saucez
INRIA Sophia Antipolis, FR
- Quirin Scheitle
TU München, DE
- Jürgen Schönwälder
Jacobs University Bremen, DE
- Henning Schulzrinne
Columbia University – New York, US
- Georgios Smaragdakis
TU Berlin, DE
- Karen Sollins
MIT – Cambridge, US
- Joel Sommers
Colgate University – Hamilton, US
- Brian Trammell
ETH Zürich, CH
- Matthias Wählisch
FU Berlin, DE
- Klaus Wehrle
RWTH Aachen University, DE
- John Wroclawski
USC – Marina del Rey, US
- Thomas Zinner
TU Berlin, DE

Report from Dagstuhl Seminar 18421

Algorithmic Enumeration: Output-sensitive, Input-Sensitive, Parameterized, Approximative

Edited by
Henning Fernau[1], Petr A. Golovach[2], and Marie-France Sagot[3]

1 **Universität Trier, DE,** `fernau@uni-trier.de`
2 **University of Bergen, NO,** `petr.golovach@ii.uib.no`
3 **University Claude Bernard – Lyon, FR,** `marie-france.sagot@inria.fr`

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 18421 "Algorithmic Enumeration: Output-sensitive, Input-Sensitive, Parameterized, Approximative".

Enumeration problems require to list all wanted objects of the input as, e.g., particular subsets of the vertex or edge set of a given graph or particular satisfying assignments of logical expressions. Enumeration problems arise in a natural way in various fields of Computer Science, as, e.g., Artificial Intelligence and Data Mining, in Natural Sciences Engineering, Social Sciences, and Biology. The main challenge of the area of enumeration problems is that, contrary to decision, optimization and even counting problems, the output length of an enumeration problem is often exponential in the size of the input and cannot be neglected. This makes enumeration problems more challenging than many other types of algorithmic problems and demands development of specific techniques.

The principal goals of our Dagstuhl seminar were to increase the visibility of algorithmic enumeration within (Theoretical) Computer Science and to contribute to establishing it as an area of "Algorithms and Complexity". The seminar brought together researchers within the algorithms community, other fields of Computer Science and Computer Engineering, as well as researchers working on enumeration problems in other application areas, in particular Biology. The aim was to accelerate developments and discus new directions including algorithmic tools and hardness proofs.

Seminar October 14–19, 2018 – http://www.dagstuhl.de/18421
2012 ACM Subject Classification Mathematics of computing → Enumeration, Mathematics of computing → Combinatorial algorithms, Theory of computation → Design and analysis of algorithms, Theory of computation → Complexity classes
Keywords and phrases algorithms, input-sensitive enumeration, output-sensitive enumeration
Digital Object Identifier 10.4230/DagRep.8.10.63
Edited in cooperation with Benjamin Gras (University of Orleans, FR)

1 Executive Summary

Marie-France Sagot (University Claude Bernard – Lyon, FR)
Dieter Kratsch (Université de Lorraine, FR)
Henning Fernau (Universität Trier, DE)
Petr A. Golovach (University of Bergen, NO)

About fifty years ago, NP-completeness became the lens through which Computer Science views computationally hard (decision and optimization) problems. In the last decades various new approaches to solve NP-hard problems exactly have attracted a lot of attention, among them parameterized and exact exponential-time algorithms, typically dealing with decision and optimization problems.

While optimization is ubiquitious in computer science and many application areas, relatively little is known about enumeration within the "Algorithms and Complexity" community. Fortunately there has been important algorithmic research dedicated to enumeration problems in various fields of Computer Science, as, e.g., Artificial Intelligence and Data Mining, in Natural Sciences Engineering and Social Sciences.

Enumeration problems require to list all wanted objects of the input as, e.g., particular subsets of the vertex or edge set of a given graph or particular satisfying assignments of logical expressions. Contrary to decision, optimization and even counting problems, the output length of enumeration problems is often exponential in the size of the input and cannot be neglected. This motivates the classical approach in enumeration, now called output-sensitive, which measures running time in (input and) output length, and asks for output-polynomial algorithms and algorithms of polynomial delay. This approach has been studied since a long time and has produced its own important open questions, among them the question whether the minimal transversals of a hypergraph can be enumerated in output-polynomial time. This longstanding and challenging question has triggered a lot of research. It is open for more than fifty years and most likely the best known open problem in algorithmic enumeration.

Recently as a natural extension on research in exact exponential-time algorithms, a new approach, called input-sensitive, which measures the running time in the input length, has found growing interest. Due to the number of objects to enumerate (in the worst case), the corresponding algorithms have exponential running time. So far branching algorithms are a major tool. Input-sensitive enumeration is strongly related to lower and upper combinatorial bounds on the maximum number of objects to be enumerated for an input of given size. Such bounds can be achieved via input-sensitive enumeration algorithms but also by the use of combinatorial (non-algorithmic) means.

The area of algorithmic enumeration is in a nascent state, though it has a huge potential due to theoretical challenges and practical applications. While output-sensitive enumeration has a long history, input-sensitive enumeration has been initiated only recently. Natural and promising approaches like using parameterized or approximative approaches have not been explored yet in their full capacities.

The principal goals of our Dagstuhl seminar were to increase the visibility of algorithmic enumeration within (Theoretical) Computer Science and to contribute to establishing it as an area of "Algorithms and Complexity". The seminar brought together researchers within the algorithms community, other fields of Computer Science and Computer Engineering, as well as researchers working on enumeration problems in other application areas, in particular, in Biology. Besides the people already working with enumeration, researchers from other fields of Computer Science were invited. In particular, researchers who are interested in Parameterized Complexity and different aspects of counting problems were participating in the seminar. The aim was to accelerate developments and discuss new directions including algorithmic tools and hardness proofs.

The seminar collected 44 participants from 13 countries. The participants presented their recent results in 18 invited and contributed talks. Open problems were discussed in several open problem and discussion sessions.

2 Table of Contents

3 Overview of Talks

3.1 Modular counting of directed Hamiltonian cycles by enumerating solutions to quadratic equations

Andreas Björklund (Lund University, SE)

We show that the number of Hamiltonian cycles in an n-vertex directed graph can be counted modulo 2^k in expected time $O(1.619^n \binom{n}{k})$ for constant k. Our approach is based on an inclusion–exclusion formula over 2^n $n \times n$ matrix determinants. The speedup is obtained by observing that many of the determinants will be zero for the trivial reason of having k or more columns with every element even. We note that the determinants that are not trivially zero in the above sense can be seen as solutions to a special kind of quadratic equation system modulo 2. We show how to list the solutions in expected time $O(1.619^n \binom{n}{k})$.

3.2 Ideal-preferred Enumeration of Minimal Dominating Sets in Graphs

Oscar Defrain (Université Clermont Auvergne – Aubiere, FR)

Joint work of Oscar Defrain, Lhouari Nourine, Takeaki Uno

Due to its equivalence with the two problems of enumerating minimal transversals of a hypergraph, and minimal dominating sets of a graph, the dualization of a Boolean lattice is, in disguise, one of the most studied problem in algorithmic enumeration [1, 2, 3]. Its generalization to distributive lattices, however, is little-understood. In this work, we investigate ideal-preferred enumeration of minimal dominating sets in graphs, toward such a generalization, based on the recent framework of Staworko et al. on preferred consistent query answering in databases. We show that this problem is equivalent to the dualization of a distributive lattice, even when considering various combined restrictions on graphs classes and posets, including bipartite, split and co-bipartite graphs, and variants of neighborhood inclusion posets.

References

1 Fredman, M. L., Khachiyan, L. (1996). On the complexity of dualization of monotone disjunctive normal forms. Journal of Algorithms, 21(3).
2 Staworko, S., Chomicki, J., Marcinkowski, J. (2012). Prioritized repairing and consistent query answering in relational databases. Annals of Mathematics and Artificial Intelligence, 64(2-3), 209–246.
3 Babin, M. A., Kuznetsov, S. O. (2017). Dualization in lattices given by ordered sets of irreducibles. Theoretical Computer Science, 658, 316–326.

3.3 Enumerating Vertices of Covering Polyhedra with Totally Unimodular Constraint Matrices

Khaled M. Elbassioni (Masdar Institute – Abu Dhabi, AE) and Kazuhisa Makino (University of Tokyo, JP)

Main reference Khaled M. Elbassioni, Kazuhisa Makino: "Enumerating Vertices of 0/1-Polyhedra associated with 0/1-Totally Unimodular Matrices", in Proc. of the 16th Scandinavian Symposium and Workshops on Algorithm Theory, SWAT 2018, June 18-20, 2018, Malmö, Sweden, LIPIcs, Vol. 101, pp. 18:1–18:14, Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, 2018.
URL http://dx.doi.org/10.4230/LIPIcs.SWAT.2018.18

We give an incremental polynomial time algorithm for enumerating the vertices of any polyhedron $P = P(A, 1) = \{x \in \mathbb{R}^n \mid Ax \geq 1,\ x \geq 0\}$, when A is a totally unimodular matrix. Our algorithm is based on decomposing the hypergraph transversal problem for unimodular hypergraphs using Seymour's decomposition of totally unimodular matrices, and may be of independent interest.

3.4 The Minimal Extension of a Partial Solution

Henning Fernau (Universität Trier, DE)

Joint work of Katrin Casel, Henning Fernau, Mehdi Khosravian Ghadikolaei, Jérôme Monnot, Florian Sikora
Main reference Katrin Casel, Henning Fernau, Mehdi Khosravian Ghadikolaei, Jérôme Monnot, Florian Sikora: "On the Complexity of Solution Extension of Optimization Problems", CoRR, Vol. abs/1810.04553, 2018.
URL http://arxiv.org/abs/1810.04553

The very general problem of determining the quality of a given partial solution occurs basically in every algorithmic approach which computes solutions in some sense gradually. Pruning search-trees, proving approximation guarantees or the efficiency of enumeration strategies usually requires a suitable way to decide if a partial solution is a reasonable candidate to pursue. Consider for example the classical concept of minimal dominating sets for graphs. The task of finding a maximum cardinality minimal dominating set (or an approximation of it) as well as enumerating all minimal dominating sets naturally leads to solving the following extension problem: Given a graph $G = (V, E)$ and a vertex set $P \subseteq V$, does there exists a *minimal* dominating set S with $P \subseteq S$.

In an attempt to study the nature of such extension tasks, we propose a general, partial-order based framework to express a broad class of what we refer to as *extension problems*. In essence, we consider optimisation problems in NPO with an additionally specified set of presolutions (including the solutions) and a partial order on those. This partial order $\preccurlyeq$ reflects not only the notion of *extension* but also of *minimality* as follows. For a presolution P and a solution S, S *extends* P if $P \preccurlyeq S$. A solution S is *minimal*, if there exists no solution $S' \neq S$ with $S' \preccurlyeq S$. The resulting extension problem is then formally the task to decide for a given presolution P, if there exits a minimal solution S which extends P.

We consider a number of specific problems which can be expressed in this framework. Possibly contradicting intuition, these problems tend to be NP-hard, even for problems where the underlying optimisation problem can be solved in polynomial time. This raises the question of how fixing a presolution causes this increase in difficulty. In this regard, we study the parameterised complexity of extension problems with respect to parameters related to the presolution. We further discuss relaxation of the extension constraint asking only for a solution S which extends some presolution $P' \preccurlyeq P$. Here we do not want just any such presolution P' but we want P' to be as close to P as possible, in the sense that there exits no presolution $P'' \neq P'$ with $P' \preccurlyeq P'' \preccurlyeq P$ which can also be extended. These considerations yield some insight into the difficult aspects of extension problems.

3.5 Enumeration of Preferred Extensions in Almost Oriented Digraphs

Serge Gaspers (UNSW – Sydney, AU)

Joint work of Serge Gaspers, Rayi Li

In this talk, we present enumeration algorithms to list all preferred extensions of an argumentation framework. This task is equivalent to enumerating all semikernels of a directed graph. For directed graphs on n vertices, all preferred extensions can be enumerated in $O^*(3^{n/3})$ time and there are directed graphs with $\Omega(3^{n/3})$ preferred extensions. We give faster enumeration algorithms for directed graphs with at most $0.8004n$ vertices occurring in 2-cycles. In particular, for oriented graphs one of our algorithms runs in time $O(1.2321^n)$, and we show that there are oriented graphs with $\Omega(3^{n/6}) > \Omega(1.2009^n)$ preferred extensions.

A combination of three algorithms leads to the fastest enumeration times for various proportions of the number of vertices in 2-cycles. The most innovative one is a new 2-stage sampling algorithm, combined with a new parameterized enumeration algorithm, analyzed with a combination of the recent monotone local search technique (STOC 2016) and an extension thereof (ICALP 2017).

3.6 Listing All Maximal k-Plexes in Temporal Graphs

Anne-Sophie Himmel (TU Berlin, DE)

Joint work of Matthias Bentert, Anne-Sophie Himmel, Hendrik Molter, Marco Morik, Rolf Niedermeier, René Saitenmacher
Main reference Matthias Bentert, Anne-Sophie Himmel, Hendrik Molter, Marco Morik, Rolf Niedermeier, René Saitenmacher: "Listing All Maximal k-Plexes in Temporal Graphs" in Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, (ASONAM '18), p.41–46. IEEE, 2018.
URL https://doi.org/10.1109/ASONAM.2018.8508847

Many real-world networks evolve over time, that is, new contacts appear and old contacts may disappear. They can be modeled as temporal graphs where interactions between vertices (in case of social networks these would represent people) are represented by time-stamped edges. One of the most fundamental problems in (social) network analysis is community detection, and one of the most basic primitives to model a community is a clique. Addressing the problem of finding communities in temporal networks, Viard et al. [TCS 2016] introduced Δ-cliques as a natural temporal version of cliques. Himmel et al. [SNAM 2017] showed how to adapt the well-known Bron-Kerbosch algorithm to enumerate Δ-cliques. We continue this work and improve and extend this algorithm to enumerate temporal k-plexes (notably, cliques are the special case $k = 1$).

We define a Δ-k-plex as a set of vertices with a lifetime, where during the lifetime each vertex has an edge to all but at most $k-1$ vertices at least once within any consecutive $\Delta+1$ time steps. We develop a recursive algorithm for enumerating all maximal Δ-k-plexes and perform experiments on real-world social networks that demonstrate the feasibility of our approach. In particular, for the special case of Δ-1-plexes (that is, Δ-cliques), we observe that our algorithm is significantly faster than the previous algorithm by Himmel et al. [SNAM 2017] for enumerating Δ-cliques.

3.7 Constructive and nonconstructive enumeration of designs

Petteri Kaski (Aalto University, FI)

This talk looks at enumeration problems in the study of combinatorial designs where the enumeration is to be carried out up to isomorphism given by a group action. We look at two examples, namely one-factorizations of the complete graph and Latin squares, enumerated up to isomorphism by the sequences A000474 and A003090, respectively, in the Online Encyclopedia of Integer Sequences. A key technique used in both cases is McKay's method of canonical extensions [1].

References

1 Brendan D. McKay, Isomorph-free exhaustive generation, J. Algorithms 26 (1998), no. 2, 306–324.

3.8 Node Similarity with q-Grams for Real-World Labeled Networks

Andrea Marino (University of Pisa, IT)

Joint work of Conte, Alessio; Ferraro, Gaspare; Grossi, Roberto; Sadakane, Kunihiko; Uno, Takeaki
Main reference Alessio Conte, Gaspare Ferraro, Roberto Grossi, Andrea Marino, Kunihiko Sadakane, Takeaki Uno: "Node Similarity with q -Grams for Real-World Labeled Networks", in Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018, pp. 1282–1291, ACM, 2018.
URL http://dx.doi.org/10.1145/3219819.3220085

We study node similarity in labeled networks, using the label sequences found in paths of bounded length q leading to the nodes. (This recalls the q-grams employed in document resemblance, based on the Jaccard distance.) When applied to networks, the challenge is two-fold: the number of q-grams generated from labeled paths grows exponentially with q, and their frequency should be taken into account: this leads to a variation of the Jaccard index known as Bray-Curtis index for multisets. We describe NSIMGRAM, a suite of fast algorithms for node similarity with q-grams, based on a novel blend of color coding, probabilistic counting, sketches, and string algorithms, where the universe of elements to sample is exponential. In particular, after a preprocessing node coloring phase, our method estimates the similarity between two nodes x and y, sampling colorful q-paths, i.e. q-paths containing different colors ending in x and y, and using the Bray-Curtis index for the corresponding multisets of q-grams. We provide experimental evidence that our measure is effective and our running times scale to deal with large real-world networks. One lesson learned is that, when dealing with network analytic, in order to avoid to do statistics on an exponential number of solutions, sometimes we can successfully focus just on special solutions, e.g. colorful solutions, or consider just part of them by applying sampling.

3.9 On Solving Enumeration Problems with SAT Oracles

João Marques-Silva (University of Lisbon, PT)

Joint work of Mark H. Liffiton, Alessandro Previti, Ammar Malik, João Marques-Silva
Main reference Mark H. Liffiton, Alessandro Previti, Ammar Malik, João Marques-Silva: "Fast, flexible MUS enumeration", Constraints, Vol. 21(2), pp. 223–250, 2016.
URL http://dx.doi.org/10.1007/s10601-015-9183-0

Enumeration is often necessary in the context of reasoning about computationally hard problems, including NP-complete, PSPACE-complete, or harder. One example is enumeration of solutions. However, there are many other enumeration problems, that find important practical applications. One example is the enumeration of diagnoses or explanations of overconstrained systems. This talk overviews some of these enumeration problems, and reports on the progress achieved in recent years.

3.10 Minimal Connected Dominating Sets below 2^n

Michal Pilipczuk (University of Warsaw, PL)

Joint work of Daniel Lokshtanov, Michal Pilipczuk, Saket Saurabh
Main reference Daniel Lokshtanov, Michal Pilipczuk, Saket Saurabh: "Below all subsets for Minimal Connected Dominating Set", CoRR, Vol. abs/1611.00840, 2016.
URL http://arxiv.org/abs/1611.00840

We prove that the number of minimal connected dominating sets in an n-vertex graph is always bounded by $(2-\varepsilon)^n$, for some $\varepsilon > 0$, and moreover they can be enumerated in time $O((2-\varepsilon)^n)$. The proof relies on a fine understanding of combinatorial properties of minimal connected dominating sets and a probabilistic argument, which shows that if a graph is "robustly dense", then the probability that a random subset of its vertices is a minimal connected dominating set is exponentially small.

3.11 Enumerating minimal dominating sets in triangle-free graphs

Jean-Florent Raymond (TU Berlin, DE)

Joint work of Marthe Bonamy, Oscar Defrain, Marc Heinrich, Jean-Florent Raymond
Main reference Marthe Bonamy, Oscar Defrain, Marc Heinrich, Jean-Florent Raymond: "Enumerating minimal dominating sets in triangle-free graphs", CoRR, Vol. abs/1810.00789, 2018.
URL http://arxiv.org/abs/1810.00789

It is a long-standing open problem whether the minimal dominating sets of a graph can be enumerated in output-polynomial time. In this talk I will present the following results obtained jointly with Oscar Defrain, Marthe Bonamy, and Marc Heinrich:

- the enumeration of minimal dominating sets in triangle-free graphs can be performed in output-polynomial time;
- deciding if a set of vertices of a bipartite graph can be completed into a minimal dominating set is a NP-complete problem.

3.12 The Maximum Number of Minimal Dominating Sets in a Tree

Günter Rote (FU Berlin, DE)

Main reference Günter Rote: "The Maximum Number of Minimal Dominating Sets in a Tree", in Proc. of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019, pp. 1201–1214, SIAM, 2019.
URL https://doi.org/10.1137/1.9781611975482.73

A tree with n vertices has at most $95^{n/13}$ minimal dominating sets. The corresponding growth constant $95^{1/13} \approx 1.4194908$ is best possible.

I show how these results are obtained in a semi-automatic way with computer help, starting from the dynamic-programming recursion for computing the number of minimal dominating sets of a given tree. This recursion defines a bilinear operation on sixtuples, and the growth constant arises as a kind of "dominant eigenvalue" of this operation.

We also derive an output-sensitive algorithm for listing all minimal dominating sets with linear set-up time and linear delay between reporting successive solutions. It is open whether the delay can be reduced to a constant delay, for an appropriate modification of the problem statement.

3.13 Efficiently Enumerating Hitting Sets of Hypergraphs Arising in Data Profiling

Martin Schirneck (Hasso-Plattner-Institut – Potsdam, DE)

Joint work of Thomas Bläsius, Tobias Friedrich, Julius Lischeid, Kitty Meeks, Martin Schirneck

A reoccurring task in the design and profiling of relational data is the discovery of hidden dependencies between attributes, e.g., unique column combinations. Enumerating them is equivalent to the classical transversal hypergraph problem. We present a backtracking algorithm for the enumeration of inclusion-wise minimal hitting sets that achieves polynomial delay on hypergraphs for which the size of the largest minimal transversal is bounded.

This algorithm solves the extension problem for minimal hitting sets as a subroutine. The extension problem is known to be NP-complete. We show that it also remains hard in the parameterised sense. In fact, it is one of only a few natural W[3]-complete problems when parameterised by the size of the set to be extended. Despite the hardness results, we show that a careful implementation of the extension oracle can help avoiding the worst case on hypergraphs arising in the profiling of real-world databases, leading to significant speed-ups in practice.

This work is to appear at the 2019 Meeting on Algorithm Engineering and Experiments (ALENEX).

3.14 A panorama of enumeration complexity

Yann Strozecki (University of Versailles, FR)

Joint work of Yann Strozecki, Arnaud Mary, Florent Capelli
Main reference Florent Capelli, Yann Strozecki: "Incremental delay enumeration: Space and time", Discrete Applied Mathematics, 2018.
URL https://doi.org/10.1016/j.dam.2018.06.038

We review different complexity measures for enumeration algorithms: total time, incremental time and delay. To each of these measures, we associate one ore several complexity classes. We show how those classes relate to classical decision or search problems and how they relate together. In particular, we prove strict inclusions modulo complexity hypotheses such as ETH or TFNP≠NP. We then present a framework of saturation problems designed to investigate the limit between incremental polynomial time and polynomial delay. It allows to prove that a large number of natural saturation problems are solvable in polynomial delay and capture interesting problems such as the enumerations of maximal cliques in a graph or of the circuits in a matroid representable over a finite field. We also present low complexity classes with polynomial time precomputation and related algorithms: constant delay (Gray codes, amortization or queries over simple structures) and polynomial delay in the size of a single solution.

3.15 How to use the Solutions of Enumeration in Practice

Takeaki Uno (National Institute of Informatics – Tokyo, JP)

There are many applications of optimizations and enumerations in practice. However, users are often, always, not satisfied with the solution given by those algorithms. For the optimization, there is no alternatives even though they have several implicit unwritten rules, and for enumeration that is the given huge number of solutions. Actually, the users usually want to choose something well, thus want to understand the problems and data easily, what is important, and what solutions are possible. In this sense, users want to be given good explanations of problems and solutions, in an understandable and abstracted ways. For this sake, we propose three ways; one is abstract the solutions by clustering the solutions, one is decomposing the problem/solutions into parts composing the solutions, and one is to extract important vertices and edges of the solutions according to the distribution of the solutions to the enumeration.

3.16 A Complexity Theory for Hard Enumeration Problems

Heribert Vollmer (Leibniz Universität Hannover, DE)

Joint work of Nadia Creignou, Markus Kröll, Reinhard Pichler, Sebastian Skritek, Heribert Vollmer
Main reference Nadia Creignou, Markus Kröll, Reinhard Pichler, Sebastian Skritek, Heribert Vollmer: "On the Complexity of Hard Enumeration Problems", CoRR, Vol. abs/1610.05493, 2016.
URL http://arxiv.org/abs/1610.05493

We introduce a hierarchy complexity classes and reductions for enumeration problems. The hierarchy resembles the polynomial-time hierarchy for decision problems. We provide tools to prove membership in the hierarchy, e.g., based on self-reducibility. Furthermore, we introduce a new reduction among enumeration problems, based on oracle computations, i.e., a Turing-type reduction. We prove a basic "Completeness Theorem", providing an easy to obtain hardness for classes in our hierarchy. We demonstrate the applicability of our approach by providing a number of completeness results for well-studied enumeration problems from the areas of propositional logic, artificial intelligence, and database theory.

4 Open problems

4.1 The DIVERSE X Paradigm

Michael R. Fellows (University of Bergen, NO) and Frances Rosamond (University of Bergen, NO)

4.1.1 Introduction

Diverse X refers to a parameterized problem such as the following (being somewhat informal, to begin with):

DIVERSE VERTEX COVERS

Input: A graph G on n vertices, and positive integers k, r and T.

Parameter: (k, r)

Output: A set S of r different vertex covers V_i of G, each of size at most k, such that a diversity measurement for the set of solutions S is at least T.

To summarize the role of these different numbers: k measures the quality of the solutions (the vertex covers in S). A vertex cover of a graph is good when it is small; r measures the size of the sample S of these high quality solutions; T measures the diversity of the sample. For example, we might take the diversity measurement to be the sum of the Hamming distances between the pairs of n-length 0/1 indicator vectors, where the sum is taken over all possible pairs V_i and V_j of the solutions in S.

This is clearly a very general paradigm for problem parameterization. The problem defined above is a variation on the classic NP-hard decision problem X = VERTEX COVER.

Theorem. DIVERSE VERTEX COVERS is FPT.

4.1.2 Motivation for the Paradigm

A motivation to look at this parameterization has been developing for a long time, since some conversations in the early 1990s with biologists, in the collaborative setting of computational biology.[1]

Biologists typically do not want just *one* high quality solution to a combinatorial optimization problem relevant to their investigations. For example, they don't want just one high quality multiple sequence alignment. Because they have all of this other information. Call this *side information* which is not included in the simple combinatorial model. Given two sequence alignments of equally good score, they might prefer one of them, on the basis of side information, for further investigation.[2]

Biologists typically also do not want *all* of the solutions of quality k. There are usually too many to deal with.

A moment's reflection makes clear that this is the natural situation now in all areas of science and information systems. There is more and more data coming from all different directions. Relevant side information is *the rule* in practical computing, not the exception.

This direction of investigation is truly fundamental and profound!

The traditional approach in computational complexity, focusing on the existence (or construction) of a single high-quality solution, lending to optimality issues (in the worst-case asymptotic framework), is therefore almost always nonsense!

This generality and significance deserves a brave and colorful name: *The Second Main Heresy.*

In applied settings, they generally don't care about optimality. They want a diverse sample of high quality solutions, to which they can apply side information in picking one to work with.

In order to get an r-sized sample of quality k, and high diversity, you might have to move away from optimality. So this is sort of like approximation, but different. There will likely be interesting tradeoffs motivated by practical considerations and the properties of real datasets, and the playground for metatheorems should be rich.

4.1.3 What is a Diverse Collection of High Quality Solutions Good For?

Beyond the basic vanilla issue raised in the above section, one can point to several established areas, and speculative possibilities.

Concorde-style memetic heuristics. The Concorde heuristic for TSP patented by Paul Seymour and Bill Cook, when Paul was at Bell Labs, maintains a (smallish) population of high quality solutions, and alternates between two phases. In one phase, each solution in the population is improved (hopefully) by a local search heuristic, applied separately to each solution in the population. In the second phase (termed *recombination*) a small number of

[1] The conversations were with Ben Koop and/or Chris Upton. Why did it take so long for the DIVERSE X research direction to gel? The conversations were in the early days of parameterized complexity, and my recollection is that we were already taking so much heat for the very idea of a relevant secondary measurement apart from the input size — that while they had an undeniably good point, our plate was already full of trouble!

[2] It can even happen that the crucial side information is *secret*, as in an amusing anecdote about a scheduling problem that I heard about from Karsten Weihe at a Dagstuhl Workshop around 1998.

tours (like ten or so) from the population (randomly chosen), are combined by taking the union of the edges involved in these ten tours.

It turns out, empirically, for many applications of the TSP problem, the union of ten high quality tours tends to be a graph of treewidth less than 15, and on such a graph, TSP can be solved exactly and optimally (essentially by an FPT dynamic programming algorithm where the treewidth parameter is 15). The solution found will be at least as good as any of the tours in the union. This is then added to the population (which is maintained at a constant size by some pruning of less efficient solutions). These two phases are repeated some number of times.

Prominence in data – training sets for machine learning, and uses in data visualization. The work of Gunnar Carlsson and Harlan Sexton.

Maximizing relative prominence in the setting of abstract simplicial complexes is the same as sum of pairwise Hamming distances.

Speculative Uses: Heuristic Kernelization. If you have a highly diverse r-sized sample S of good solutions (k-small) for VERTEX COVER, then the vertices v that:

(i) Occur in many of the solutions in S, should perhaps be inducted into the solution: $G' = G - v$ and $k' = k - 1$. (ii) Occur in none of solutions in S, should perhaps be banned from a solution. $G' = G - -N[v]$ and $k' = k - deg(v)$.

Either of these simplifies the instance and is a heuristic kernelization.

Speculative Uses: Evidentiary Interpretation of Universal Quantification. Suppose we are interested in the graph property involving alternating quantification:

For every independent set J in G of size k, there exists a set J' of size k such tha $J \cup J'$ is a dominating set in G.

Considered classically, this is likely hard for Π_2^P, but if we could find a *very diverse* collection of k-independent sets, for each J_i of which, there is a J_i' such that $J_i \cup J_i'$ is a dominating set, then we might regard this as evidence that the graph has the property. Thus using diversity to interpret Π_2^P into NP in a heuristic, evidentiary manner.

This could potentially be quite interesting, as most of mathematical thought involves at most two, sometimes three, very rarely four alternating quantifications. It is probably the same with cognition in general.

This could also be interesting in the context of the Szeider-deHaan program.

4.1.4 Some Concrete Open Problems

Classical P-time Self-reducibility. DIVERSE X, generally speaking, is a search problem. In the setting of classical complexity (P *vs* NP and all that), search problems are sort of swept under the rug (see the discussion in the Garey and Johnson book). It is generally easy to work out a search algorithm given an oracle for the decision problem.[3]

Given an oracle for the decision problem ("Does G have a dominating set of size k?") is there (or, when is there) a P-time oracle algorithm that constructively produces a set S of r different solutions, of diversity measure at least T?

[3] It should be more widely understood that this is a *skeleton in the closet* of classical complexity. The problem is in defining what one means by *P-time self-reducibility*, the usual name for this passage from decision to construction. Formally, decision problems are defined by a formal language, but the notion of a solution requires some sort of relational framework. There is no generally agreed-upon definition of P-time self-reducibility. An attempt to frame things in relational terms was made in [1]. This might be relevant to the DIVERSE X program.

What about FPT self-reducibility for DIVERSE X? For various X and various diversity measures.

Dependence of Positive Results on Properties of the Diversity Measure. Can we give meta-theorems where if the diversity measure has certain properties, then positive results hold?

References

1 K. Abrahamson, M. Fellows, M. Langston and B. Moret. "Constructive Complexity." *Discrete Appl. Math.* 34 (1991), 3–16.

4.2 Various enumeration perspectives

Henning Fernau (University of Trier, Germany)

4.2.1 Planar graphs

- Planarity often gives opportunities for improved algorithms in decision or optimization problems: square root phenomenon, PTAS, ...
- Many worst-case examples in enumeration are planar. No hope for improvements here?
- Enumeration problems involving connectivity are notoriously hard.
  What about considering such problems on planar graphs?
  Intuition: There are not too many ways to interconnect planar graphs ...

4.2.2 Enumerating representative solutions

- One of the motivations for enumeration:
  User do not only want to see one optimal solution, because there might be ill-formalized aspects of the problem that require to also look at seemingly non-optimal solutions.
- Yet, users do not want to look at zillions of solutions.
- What about presenting a selection of k solutions that are far apart from each other, kind of being a representative selection of the whole space of solutions?

4.2.3 Enumeration beyond graphs

- Typically, the enumeration community focuses on problems concerning graphs or (sometimes) logic. What about other areas?
- To give some concrete example and also to indicate the different character of such problems, we suggest looking at *synchronizing words*, a notion related to the possibly most famous open combinatorial question in Formal Languages, the so-called Černý Conjecture.
- A word $x \in \Sigma^*$ is called *synchronizing* for a DFA A, $A = (S, \Sigma, \delta, s_0, F)$, if there is a state s_f, such that for all states s, $\delta^*(s, x) = s_f$.
- Deciding, given (A, k), if DFA A has synchronizing word of length at most k is NP-complete [1].
- For the purpose of enumeration, in order to define a good notion of minimality, several natural options exist, depending on the chosen partial ordering on the set of words. The most common are: the prefix, suffix, subword (infix), subsequence (scattered subword), lexicographic, or length-lexicographic orderings.

- In the meantime, we (together with S. Hoffmann) could at least determine the status of most of the extension problems that can be associated to these enumeration problem. More precisely, given A and a word u, the question if there is some synchronizing word w that is larger than u (in the considered ordering) and minimal among the synchronizing words can be solved in polynomial time for the prefix, suffix and lexicographic orderings but is NP-hard for the subsequence or length-lexicographic orderings. The status with respect to the subword ordering is unkonwn.

References

1 D. Eppstein. Reset sequences for monotonic automata. *SIAM Journal on Computing*, 19(3):500–510, 1990.

4.3 Enumeration of tree width-t-modulators

Fedor V. Fomin (University of Bergen, NO) and Saket Saurabh (Institute of Mathematical Sciences – Chennai, IN)

Let $t \geq 1$ be a fixed integer. Given a graph G, a set $S \subseteq V(G)$ is called *treewidth-t-modulator* if $G - S$ has treewidth at most t. Furthermore if $G[S]$ is connected it is called *connected treewidth-t-modulator* Observe that for $t = 1$, treewidth-1-modulator and connected treewidth-1-modulator are known as *feedback vertex set* and *connected feedback vertex set*, respectively. It is known that the number of minimal feedback vertex sets on a graph on n vertices is at most $\mathcal{O}(1.8527^n)$ [1].

1. Show that there exist a constant c such that any graph on n-vertices has at most
$$\left(2 - \frac{1}{c}\right)^n n^{\mathcal{O}(1)}$$
minimal connected feedback vertex sets.
Potential approach seems to be via methods given in [2] and [3].
2. Show that for evert $t \geq 2$, there exist a constant c_t such that any graph on n-vertices has at most
$$\left(2 - \frac{1}{c_t}\right)^n n^{\mathcal{O}(1)}$$
minimal treewidth-2-modulator sets. Even for $t = 2$, this is open. One could also study this problem when the objective is to upper bound the number of minimal vertex sets, whose deletion results in an outer-planar graph.
Potential approach seems to be via methods given in [1].

References

1 Serge Gaspers, Edward J. Lee: Exact Algorithms via Multivariate Subroutines. ICALP 2017: 69:1–69:13
2 Daniel Lokshtanov, Michal Pilipczuk, Saket Saurabh: Below all subsets for Minimal Connected Dominating Set. CoRR abs/1611.00840 (2016)
3 Fedor V. Fomin, Serge Gaspers, Artem V. Pyatkin, Igor Razgon: On the Minimum Feedback Vertex Set Problem: Exact and Enumeration Algorithms. Algorithmica 52(2): 293–307 (2008)

4.4 Minimal separators in graphs

Serge Gaspers (UNSW – Sydney, AU)

It is known that every graph on n vertices has $O(1.6181^n)$ minimal separators [2]. However, it is unknown whether this bound is tight. In 2008, a family of graphs with $\Omega(3^{n/3}) = \Omega(1.4422^n)$ minimal separators was shown [1], and this lower bound has recently been improved to $\Omega(1.4457^n)$ [3]]. The question is to improve either the upper bound or the lower bound.

References

1 Fedor V. Fomin, Dieter Kratsch, Ioan Todinca, Yngve Villanger: Exact Algorithms for Treewidth and Minimum Fill-In. SIAM J. Comput. 38(3): 1058–1079, 2008.
2 F. V. Fomin and Y. Villanger, Treewidth computation and extremal combinatorics, Combinatorica 32(3), 289–308, 2012.
3 Serge Gaspers and Simon Mackenzie. On the Number of Minimal Separators in Graphs. Journal of Graph Theory 87(4): 653–659, 2018.

4.5 Enumeration of vertex subsets with non-local properties

Petr A. Golovach (University of Bergen, NO)

The most standard approach for the input-sensitive enumeration of vertex subsets of a graph satisfying a given property Π is using the recursive branching algorithms (see the book [1]). Still, this approach has limitations. In particular, the technique works well for properties that are local in some sense. For example, a set of vertices X is an inclusion-maximal independent set of a graph G if and only if $|X \cap N[v]| = 1$ for every $v \in V(G)$ and this property is crucial for for the recursive branching algorithm that enumerates maximal independent sets of an n-vertex graph in time $\mathcal{O}^*(3^{n/3})$. It looks that for non-local properties, we have to develop new techniques. The following specific problems could be interesting as starting points.

1. A set $D \subseteq V(G)$ is a *connected dominating set* of G if D is a dominating set and $G[D]$ is connected. Lokshtanov, Pilipczuk and Saurabh [2] proved that all inclusion-minimal connected dominating sets of an n-vertex graph can be enumerated in time $\mathcal{O}(2^{(1-\varepsilon)n})$ where $\epsilon < 10^{-50}$ slightly improving upon the trivial $\mathcal{O}^*(2^n)$ algorithm. Is it possible to improve this results?
2. A set $D \subseteq V(G)$ is *irredundant* if for every $u \in D$, there is $v \in N[u]$ (called a *private vertex*) such that $v \notin N[w]$ for every $w \in D \setminus \{u\}$. Is it possible to enumerate all inclusion-maximal irredundant sets of an n-vertex graph faster that $\mathcal{O}^*(2^n)$?

References

1 Fedor V. Fomin, Dieter Kratsch: Exact Exponential Algorithms. Springer 2010
2 Daniel Lokshtanov, Michal Pilipczuk, Saket Saurabh: Below all subsets for Minimal Connected Dominating Set. CoRR abs/1611.00840 (2016)

4.6 Enumeration of minimal dominating sets for graph classes

Dieter Kratsch (Université de Lorraine, FR)

The number of minimal dominating sets of an n-vertex graphs has generated plenty of research in the last 15 years. In fact even in 1985 all known was that it could not be more than order of $2^n/\log n$.

Due to groundbreaking work in input-sensitive algorithms by Fomin, Grandoni, Pyatkin and Stepanov (2008) we know that there is an upper bound of 1.7159^n. This one is achieved by an intricate measure and conquer analysis of the branching algorithm enumerating all minimal dominating sets of the input graph. The best known lower bound for general n-vertex graphs is $15^{n/6}$, approximately 1.5704^n, is folklore (see e.g. the master thesis of Serge Gaspers in 2005, see also the above mentioned paper by Fomin et al). Hence there is a huge gap between upper and lower bound.

- Any even small improvement upon the lower or upper bound for graphs in general would be highly appreciated and may lead to improvements on various similar questions (minimal feedback vertex sets, minimal separators, minimal connected dominating sets, minimal subset feedback vertex sets, etc.).

A natural approach to achieve matching upper and lower bounds is to restrict lower and upper bounds to some particular class of graphs. Starting with a paper by Couturier Heggernes, Kratsch, van't Hoff (2013) matching upper and lower bounds for the number of minimal dominating sets in n-vertex graphs of a certain graph class have been obtained. It is worth mentioning that in almost all of these cases the upper bound has been obtained by the use of structural properties of the graph class (and without complicated measure-and-conquer analysis).

- An interesting class that despite various efforts still has non matching upper and lower bounds are the chordal graphs.
- Nothing has been published on bipartite graphs.
- Possibly the most challenging case are permutation graphs. For at least a year we tried to prove that $15^{n/6}$, approximately 1.5705^n, is a matching upper bound. However we never succeeded to fix all the holes in our proofs. Thus we give up on this one some years ago. Now with some new ideas we conjecture that 1.5705^n is an upper bound for permutation graphs and it might even be an upper bound for cocomparability graphs.

4.7 Enumeration of k-colorable induced subgraphs

Daniel Lokshtanov (University of California – Santa Barbara, US)

Question 1: Is the following statement true? For every integer $k > 0$ there exists reals $c_k < 2$ and $a_k > 0$ such that for every graph G on n vertices the number of inclusion maximal k-colorable induced subgraphs of G is at most $a_k \cdot c_k^n$.

Question 2: For a graph G and integer k, define the family $F_k(G)$ to contain all vertex sets S that are the union of k (not necessarily distinct) maximal independent sets of G. Is the following statement true? For every integer $k > 0$ there exists reals $c_k < 2$ and $a_k > 0$ such that for every graph G on n vertices $|F_k(G)| \leq a_k \cdot c_k^n$?

Remarks: A positive answer to question 2 implies a positive answer to question 1, but not necessarily vice versa. Question 2 was posed by Boris Bukh at ICGT 2018.

4.8 Primal and Dual Representations

Kazuhisa Makino (University of Tokyo, JP)

Enumeration is one of the fundamental topics of discrete mathematics. Recently, enumerar-ation (or generation) algorithms have recently been acheived much attention in computer science. In this note, we describe three famous problems, which are *not* known to be solved in output-polynomial (or polynomial total) time. They can be viewed as the decision problems on *primal and dual representations* for polytopes, monotone Boolean functions, and Horn Boolean functions.

Problem 1
Input: A set of points $S \in \mathbb{R}^n$ and a set of inequalities H in $\mathbb{R}^n$.
Question: Is the convex hull of S represented by H ?

Problem 2
Input: A monotone DNF φ and a monotone CNF ψ.
Question: Is φ logically equivalent to ψ ?

Clearly they belong to co-NP. As for Problem 1, one of the natural representations of polytope P is the set of extreme points in P, and the other one is the set of facets of P. It is not known whether Problem 1 can be solved in polynomial time, or it is co-NP-compelte. It is known that Problem 1 can be solved in polynomial time then the corresponding generation problems such as (1-1) given a set of point S, generating all facets of the polytope defined by S, or (1-2) given a set of inequalites H (that defines a polytope), generate all the extreme points of the polytope defined by H, can be solved in polynomial total time, On the other hand, if it is co-NP-complete, the correponding generation problems have no polynomial total time algorithm, unless P=NP.

As for Problem 2, monotone Boolean functions have two natural representations: monotone DNFs φ and monotone CNFs ψ. It is known that Problem 2 can be solved in quasi-polynomial time. However, it is still open whether Problem 2 can be solved in polynomial time, or it is co-NP-compelte, where most experts in complexity theory believe that Problem 2 is not co-NP-complete. Similarly to Problem 1, Problem 2 can be solved in polynomial time if and only if the corresponding generation problems can be solved in polynomial total time.

Let us consider primal and dual representations of Horn Boolean function $f : \{0,1\}^n \to \{0,1\}$. A CNF is called Horn if each clause contains at most one positive literal, and a Boolean function is called *Horn* if it can be represented by a Horn CNF. It is well-known that a Boolean function f is Horn if and only if its false set is closed under intersection, i.e., for any pair of vectors v and w in $\{0,1\}^n$ such that $f(v) = f(w) = 1$, we have $f(v \wedge w) = 1$. Here for vectors v and w in $\{0,1\}^n$, $v \wedge w$ denotes the vector in $\{0,1\}^n$ defined by $(v \wedge w)_i = v_i \wedge w_i$

for all $i = 1, \ldots, n$. Thus Horn functions have alternative representation. For a set S of vectors in $\{0,1\}^n$, the intersection closure $Cl_\wedge(S)$ is the set defined by

$$Cl_\wedge(S) = \{ \bigwedge_{v \in W} v \mid W \subseteq S\}.$$

It is not difficult to see that any set T of vectors in $\{0,1\}^n$ that is closed under intersection has a unique (inclusion-)minimal set S such that $T = Cl_\wedge(S)$. Such a set is called *characteristic set* of T.

Problem 3
Input: A set of vectors $S \in \{0,1\}^n$ and a Horn CNF ψ
Question: Is $Cl_\wedge(S)$ a set of true vectors of ψ ?

Problem 3 is polynomially equivalent to the problem that decides whether given characteristifc set and Horn CNF represent the same Horn Boolean function. Again the complexity status of Problem 3 is still open, although Problem 3 is polynomially equivalent to Problem 2 if ψ in Problem 3 contains all prime implicates, which implies that Problem 3 is at least as difficult as Problem 2. Moreover, Problem 3 can be solved in polynomial time if and only if the corresponding generation problems can be solved in polynomial total time.

References

1 Thomas Eiter, Kazuhisa Makino, Georg Gottlob: Computational aspects of monotone dualization: A brief survey. Discrete Applied Mathematics 156: 2035–2049 (2008)
2 L. Lovasz. Combinatorial optimization: some problems and trends. DIMACS Technical Report 92-53, Rutgers University, 1992.
3 Roni Khardon: Translating between Horn Representations and their Characteristic Models. J. Artif. Intell. Res. 3: 349–372 (1995)

4.9 Minimal dominating sets enumeration and hypergraph colorings

Michal Pilipczuk (University of Warsaw, PL)

It is known that the problem of finding the next solution for enumeration of minimal dominating sets in a graph reduces to the following problem. We are given two hypergraphs H_1 and H_2 over the same universe U with the property that every hyperedge of H_1 intersects every hyperedge of H_2. The question is whether one can color every element of U red or blue so that red vertices form a transversal of H_1 while blue vertices form a transversal of H_2. It is known that this problem can be solved in quasi-polynomial time (in the size of input), which immediately leads to incremental quasipolynomial-time enumeration algorithm.

There are two question about the problem stated above:

1. Can one exclude a polynomial-time algorithm for the problem under ETH? To do so, one would need to devise a reduction from 3SAT that takes an instance φ of total size N, and in time $2^{\mathcal{O}(N^{1-\varepsilon})}$ outputs an equivalent instance of the problem in question of size $2^{\mathcal{O}(N^{1-\varepsilon})}$, for some $\varepsilon > 0$.
2. Is the problem in coNP? If this was true, then this would exclude the existence of a reduction as above, unless 3SAT is solvable in co-nondeterministic subexponential time.

4.10 Generation of representative solutions

Yann Strozecki (University of Versailles, FR)

The aim of many enumeration algorithms is to build explicitly a set of elements so that an end user (a chemist, a biologist, a network engineer ...) can inspect them and select the best for him. When the user can express its preferences as a value to optimize, it is a classical optimization problem. If there are several critieria, we have multi-criteria problem and the set of pareto optimal solutions can be enumerated (or approximated). However, most of the time, the user is unable to give its preferences explicitly and formally. Hence we must generate the whole set of solutions, which can be slow and is often too large to be used by the user or requires complex data analysis methods.

It would be much better to generate only a small subsets of the solutions but diverse enough to represent well the whole set of solutions. There are several way of formalizing this problem which can or have been proposed during the workshop: generating k solutions which are different enough, generating a sufficiently large subset of solutions, representing succinctly the set of solutions using cartesian product, unions, wildcards ... Here we assume that we have a distance over the set of solutions. We say that a subset of solutions C is a d-cover of the whole set of solutions S if all solutions in S are at distance at most d of a solution in C. We say that an enumeration problem is efficiently approximable if for every d, there is an algorithm which produces a d-cover in a time polynomial in the size of the smallest d-cover. We can relax this definition by allowing non solutions to be in the cover and also by allowing randomized algorithms.

When the set of solutions S is given explicitly, the task of finding a small d-cover is a clustering problem, which is often hard. Hence, this notion is relevant only if the set of solutions is very simple and structured. As open problems and important practical examples we ask whether the two following problems are efficiently approximable for the Hamming distance between solutions:

- the set of (minimal) spanning trees of a graph
- the set of (minimal) $s - t$ paths

4.11 Strong polynomial delay conjectures

Yann Strozecki (University of Versailles, FR)

Joint work of Yann Strozecki, Florent Capelli
Main reference Florent Capelli, Yann Strozecki: "Enumerating models of DNF faster: breaking the dependency on the formula size", CoRR, Vol. abs/1810.04006, 2018.
URL http://arxiv.org/abs/1810.04006

A good enumeration algorithm is in polynomial delay, that is the delay between the production of two solutions is polynomial in the "input". However, for many problems, the input can be much larger than the size of the generated solutions. Hence, we should rather design strong polynomial delay algorithms, where after a precomputation phase polynomial in the input, the solutions are generated with a delay polynomial in the size of the previous solution.

The problem of enumerating the models of a DNF formula is denoted by EnumDNF. This problem seems exceedingly simple: we just need to compute the union of the models of the terms and the models of one term can be obtained in constant delay by Gray code enumeration. There are linear delay algorithm solving EnumDNF, for instance using a classical branching method with the relevant data structure. Let n be the number of variables and m the number of terms of a DNF. Note that a model is of size n while m can be exponentially larger than n. Hence the known algorithms are not in strong polynomial delay. While this problem seems simple, we conjecture that dealing with non disjoint union is hard and that there are no better algorithms than the classical one.

Weak conjecture: EnumDNF cannot be solved in strong polynomial delay.

Strong conjecture: EnumDNF cannot be solved in delay $O(m^{1-\epsilon}n^k)$ for any k and $\epsilon > 0$.

Remark: the weak conjecture is false for monotone DNF formulas. The strong conjecture is false if one considers amortized delay. We feel that the weak conjecture could be proved (assuming complexity hypothesis) by looking at the total time. Indeed, it is of the form $m^a + sn^b$ where s is the number of solutions, a and b are constants. When s is large this number is almost independent from m, which seems to strong.

Participants

- Faisal N. Abu-Khzam
Lebanese American University – Beirut, LB
- Ricardo Andrade
University Claude Bernard – Lyon, FR
- Cristina Bazgan
University Paris-Dauphine, FR
- Daniel Binkele-Raible
BEONTRA – Karlsruhe, DE
- Andreas Björklund
Lund University, SE
- Oscar Defrain
Université Clermont Auvergne – Aubiere, FR
- Arnaud Durand
Paris Diderot University, FR
- Martin Dyer
University of Leeds, GB
- Khaled M. Elbassioni
Masdar Institute – Abu Dhabi, AE
- Michael R. Fellows
University of Bergen, NO
- Henning Fernau
Universität Trier, DE
- Fedor V. Fomin
University of Bergen, NO
- Serge Gaspers
UNSW – Sydney, AU
- Petr A. Golovach
University of Bergen, NO
- Benjamin Gras
University of Orleans, FR
- Anne-Sophie Himmel
TU Berlin, DE
- Leandro Ishi Soares de Lima
University Claude Bernard – Lyon, FR
- Giuseppe F. Italiano
LUISS University, IT
- Mamadou Moustapha Kanté
Université Clermont Auvergne – Aubiere, FR
- Petteri Kaski
Aalto University, FI
- Mathieu Liedloff
University of Orleans, FR
- Daniel Lokshtanov
University of California – Santa Barbara, US
- Kazuhisa Makino
University of Tokyo, JP
- Alberto Marchetti-Spaccamela
Sapienza University of Rome, IT
- Andrea Marino
University of Pisa, IT
- João Marques-Silva
University of Lisbon, PT
- Arnaud Mary
University Claude Bernard – Lyon, FR
- Lhouari Nourine
University Clermont Auvergne, FR
- Michal Pilipczuk
University of Warsaw, PL
- Jean-Florent Raymond
TU Berlin, DE
- Frances A. Rosamond
University of Bergen, NO
- Günter Rote
FU Berlin, DE
- Marie-France Sagot
University Claude Bernard – Lyon, FR
- Saket Saurabh
Institute of Mathematical Sciences – Chennai, IN
- Mohamed Yosri Sayadi
Université de Lorraine – Metz, FR
- Martin Schirneck
Hasso-Plattner-Institut – Potsdam, DE
- Blerina Sinaimeri
Université Lyon I, FR
- Leen Stougie
CWI – Amsterdam, NL
- Yann Strozecki
University of Versailles, FR
- Takeaki Uno
National Institute of Informatics – Tokyo, JP
- Heribert Vollmer
Leibniz Universität Hannover, DE
- Kunihiro Wasa
National Institute of Informatics – Tokyo, JP
- Gerhard J. Woeginger
RWTH Aachen, DE
- Petra Wolf
Universität Tübingen, DE

Report from Dagstuhl Seminar 18422

Shape Analysis: Euclidean, Discrete and Algebraic Geometric Methods

Edited by

Michael Breuß[1], Alfred M. Bruckstein[2], Christer Oscar Kiselman[3], and Petros Maragos[4]

1 **BTU Cottbus, DE,** breuss@b-tu.de
2 **Technion – Haifa, IL,** freddy@cs.technion.ac.il
3 **Uppsala University, SE,** kiselman@it.uu.se
4 **National Technical University of Athens, GR,** maragos@cs.ntua.gr

Abstract

In computer vision, geometric processing and image analysis, the notation of a shape of a 3-D object has been studied either by an embedded manifold for the continuous stetting or as a collection of a discrete set of marker positions on the manifold. Within the last years, there have been many rapid developments in the field of shape representation, shape correspondence and shape manipulation with technical applications ranging from laser-range scanners to 3-D printing. Classic shape analysis tools from differential geometry have a fresh influence in the field, often powered by modern methods from optimization and numerical computing. At the same time, discrete geometric methods and related techniques such as from mathematical morphology have evolved significantly. Moreover, techniques like deep learning gained a significant influence in the development of corresponding methods and tools. New developments from tropical geometry have a high potential for use in shape analysis.

The topics in our seminar addressed the sketched challenges and developments that will be useful for shape analysis. Especially we aimed to discuss the possibilities of combining fields like tropical geometry with more classical techniques as for instance from mathematical morphology. We discussed possibilities of applying machine learning methods in this context and considered recent advances from more classical fields like differential geometry and partial differential equations that can be useful for setting up and understanding shape analysis methods in all of these approaches.

This seminar brought together 26 researchers from North America and Europe engaged in recent and upcoming developments in shape analysis who view these challenges from different perspectives and who together discussed the pressing open problems and novel solutions to them.

Seminar October 14–19, 2018 – http://www.dagstuhl.de/18422

2012 ACM Subject Classification Theory of computation → Computational geometry, Mathematics of computing → Geometric topology, Computing methodologies → Image processing, Computing methodologies → Machine learning algorithms, Mathematics of computing → Mathematical optimization, Mathematics of computing → Partial differential equations, Computing methodologies → Shape analysis

Keywords and phrases algebraic geometry, mathematical morphology, optimization, shape analysis, shape reconstruction

Digital Object Identifier 10.4230/DagRep.8.10.87

Edited in cooperation with Robert Dachsel

Shape Analysis: Euclidean, Discrete and Algebraic Geometric Methods, *Dagstuhl Reports*, Vol. 8, Issue 10, pp. 87–103
Editors: Michael Breuß, Alfred M. Bruckstein, Christer Oscar Kiselman, and Petros Maragos
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Michael Breuß (BTU Cottbus, DE)
Alfred M. Bruckstein (Technion – Haifa, IL)
Christer Oscar Kiselman (Uppsala University, SE)
Petros Maragos (National Technical University of Athens, GR)

Dagstuhl seminar 18422 Shape Analysis: *Euclidean, Discrete and Algebraic Geometric Methods* took place October 14–19, 2018. 26 researchers from North America and Europe discussed state-of-the-art, current challenges, and promising future research directions in the areas of 2-D and 3-D shape analysis from a cross-disciplinary point of view. Participants included international experts from the fields of continuous-scale shape analysis, discrete shape analysis, tropical geometry and numerical computing. The seminar consisted of an opening and getting to know session and 26 scientific presentation sessions. Furthermore, there was time for extensive discussions both between the talks and in the evenings.

The topics in our seminar addressed the sketched challenges and developments that will be useful for shape analysis. Especially we aimed to discuss the possibilities of combining fields like tropical geometry with more classical techniques as for instance from mathematical morphology. We discussed possibilities of applying machine learning methods in this context and considered recent advances from more classical fields like differential geometry and partial differential equations that can be useful for setting up and understanding shape analysis methods in all of these approaches.

The purpose of this seminar was to address these challenges with the latest tools related to geometric, algorithmic and numerical concepts. To do so, we brought together researchers working on shape analysis topics from different perspectives. The purpose in bringing together researchers from those different communities sharing substantial interest in shape analysis was to explore the benefits of a cross-disciplinary point of view.

Promising new ways to combine the latest techniques from these different fields were identified during in-depth discussions in small groups. Some especially promising research directions in the areas of deep learning, mathematical morphology, shape from shading, modelling deformable shapes, and tropical geometry were discussed in small groups between the talks and in the evenings.

2 Table of Contents

3 Overview of Talks

3.1 Tropical geometry, Optimal Control and Mean-payoff Games

Marianne Akian (Ecole Polytechnique – Palaiseau, FR)

Tropical algebra can be seen as the limit of a deformation of usual agebra or linked with usual algebra by using the notion of valuation. This allows one to define tropical geometry and in particular tropical convex sets. Tropical polyhedra can be seen as the set of solutions of a finite number of two sided tropical inequalities or equalities. In a work with Gaubert and Guterman (2012), we proved that the above external representation of tropical polyhedra is equivalent to the notion of mean-payoff zero-sum games. Then, the polyhedra is nonempty if and only if there exists a winning position for the maximization player. We then show how this result can be combined with more recent results of Grigoriev and Podolskii (2014) and Allamigeon, Gaubert and Skomra (2018) to obtain such characterizations for polynomial equations, and for stochastic games respectively.

3.2 Heat kernel semigroups and morphological semigroups on ultrametric spaces for hierarchical data processing

Jesús Angulo (Ecole des Mines de Paris, FR)

Joint work of Jesús Angulo, Santiago Velasco-Forero

Ultrametric spaces are the natural mathematical structure to deal with data embedded into a hierarchical representation. This kind of representations is ubiquitous in morphological image processing, from pyramids of nested partitions to more abstract dendograms from minimum spanning trees. This talk presents a formal study of morphological operators for functions defined on ultrametric spaces. First, the notion of ultrametric structuring function is introduced. Then, using as basic ingredient the convolution in (max,min)-algebra, the multi-scale ultrametric dilation and erosion are defined and their semigroup properties are stated. It is proved in particular that they are idempotent operators and consequently they are algebraically ultrametric closing and opening too. Some preliminary examples illustrate the behavior and practical interest of ultrametric dilations/erosions.

3.3 Descriptors of Q-convexity: Theory and Applications (Part 2)

Péter Balázs (University of Szeged, HU)

Motivated by a special reconstruction problem in binary tomography, we start out from a one-dimensional convexity measure and generalize it to two dimensions using the concept of Q-convexity. We show a way how to normalize this measure and study the performance of this descriptor in several image classification problems. Our results are comparable and in some cases even better than those achieved by other methods based on convexity descriptors. While these latter ones take 180 directions into consideration, our measure uses just two directions in the 2D plane.

3.4 Extending Classic Photometric Stereo

Michael Breuß (BTU Cottbus, DE)

In this talk the results on recent work in the area of photometric stereo for non-Lambertian surfaces is presented. The photometric stereo (PS) problem is a classic problem in computer vision. Given several input images, taken from the same point of view but under different lighting conditions, the task is to compute the 3D shape of objects depicted in the photographed scene. The crucial modeling steps to solve this inverse problems are the camera model that yields how the 3D world is mapped to a 2D image, and the light reflectance model that accounts for the light reflectance of objects in a scene. The classic simple setting features Lambertian i.e. completely diffuse reflectance and an orthographic camera for easy mathematical modeling, respectively. The talk shows how a perspective geometry changes the mathematical setting. Moreover complex non-Lambertian reflectance models like the Blinn-Phong model known from computer graphics are addressed for the PS problem.

3.5 On Geometric Centers of Point Constellations

Alfred M. Bruckstein (Technion – Haifa, IL)

Joint work of Alfred M. Bruckstein, Yael Yankelevsky

We present a new result on the characterization of Steiner Centers for planar point constellations. The Steiner Center is the location that minimizes the excess area of coverage of the convex hull of the planar point constellation by the superposition of the disks with diameters defined by the variable point and all points in the given constellation. Proving the higher dimensional equivalent of this characterization remains an open problem.

3.6 Descriptors of Q-convexity: Theory and Applications (Part 1)

Sara Brunetti (University of Siena, IT)

In this talk, we propose a new idea to design a measure for shape descriptors based on the geometrical concept of Quadrant-convexity. We may derive a “quantitative” representation which provides a measure of concavity, and permits to define complex relations like enlacement and interlacement between objects. In addition, we discuss a different approach based on the notion of salient points and Quadrant convex hull giving rise to a “qualitative” representation of the image. The derived shape descriptors have the following features: 1)their values range from 0 to 1; 2)they are invariant by reflection and point symmetry; 3) they are scale-tolerant 4) their computation can be easily and efficiently implemented.

3.7 Geometry of Neural Networks with Piecewise Linear Activations

Vasileios Charisopoulos (Cornell University, US)

Deep neural networks are considered to be the state of the art in many image recognition challenges, such as object detection, segmentation, and other tasks, often surpassing human performance. However, despite their accessibility, their theoretical properties are still being investigated.

We will present an approach based on tropical mathematics to investigate one of these properties for neural networks with piecewise linear activations. Our approach relies on the duality between the vertices of Newton Polytopes and maximizers of corresponding polynomials in the (max, +) or tropical algebra. It enables us to recover bounds on the number of linear regions of neural network layers, which serve as a measure of their expressive power. We will attempt to demonstrate the aforementioned duality and the geometric objects involved starting from a much simpler learning model, mimicking the linear perceptron in the (max, +) setting.

3.8 A Feature Descriptor based on Time Evolution Equations

Robert Dachsel (BTU Cottbus, DE)

A major task in non-rigid shape analysis is to retrieve correspondences between two almost isometric 3D objects. For the descriptor based approach, a simplified shape representation, called feature descriptor which is invariant under almost isometric transformations is required. A successful class of feature descriptors employ the spectral decomposition of the Laplace-Beltrami operator. Important examples are the heat kernel signature and the wave kernel signature, being expressed as a series expression of the Laplace-Beltrami's low frequency eigenfunctions and eigenvalues. This choice makes the descriptors robust to high frequent noise but particularly vulnerable to global distortions of a shape. In this talk we tackle this problem and explore an alternative solution strategy. We model physically phenomena on shapes as carried by the heat, Schrödinger and wave equation. Therefore we chose a direct discretization using a combined finite volume and time integration approach. This approach leads to a local acting feature descriptor class being able to handle global distortions of a shape. By a detailed evaluation at hand of a standard shape data set we demonstrate that our approach may yield significant improvements over state of the art methods for finding correct shape correspondences.

3.9 Data representation and topology-based analysis for data visualization

Leila De Floriani (University of Maryland – College Park, US)

Joint work of Leila De Floriani, Riccardo Fellegara, Federico Iuricich

This talk provides an overview on our research on data representation and topology-based approaches to scientific data visualization. We discuss first a new compact and modular data structure for simplicial complexes in arbitrary dimensions, the Stellar tree, and then topology-based techniques for data transformation based on topological tools. Specifically, we discuss discrete Morse theory, a combinatorial counterpart of smooth Morse theory, and its application in visualization as a segmentation tool as well as in computing persistent homology for shape analysis.

3.10 Geometry of facial shapes and machine learning methods for synthesis of expressive multimodal avatars

Panagiotis Filntisis (National Technical University of Athens, GR)

We explore the geometry of faces via statistical modeling and deformable shape models such as Active Appearance Models (AAMs), for modeling facial emotional expressions and movements. We also investigate how machine/deep learning methods can be coupled with AAMs in order to synthesize smooth trajectories of different facial shapes that occur during emotional speech.

To do this, we use both established machine learning methods such as hidden Markov models as well as "deeper" models (i.e., Deep Neural Networks) and show how deep models can better capture the complexity of facial shapes and the large variations of emotional speech and expressions.

The resulting machine learning models can be used to synthesize expressive multimodal faces (avatars-talking heads) that can also adapt their shape to target emotions using a small amount of data, as well as interpolate different facial shapes to create more complex shapes.

3.11 Towards the Analysis of Multivariate Data based on discrete Morse Theory

Federico Iuricich (Clemson University, US)

Due to the advances in sensor technology, we are in need of new tools for analyzing shapes of constantly increasing complexity. Topological Data Analysis is an active research area providing robust and data-agnostic descriptors of data. The most widely used tool in TDA is persistent homology which provides a way to compute topological features in a multi-scale manner and summarizes them in a persistence diagram. Persistence diagrams can be converted in descriptors suitable for learning tasks like shape classification or retrieval.

Multi-parameter Persistent Homology (MPH), is a generalization of persistent homology devoted to the investigation of multivariate data, i.e. data provided with multiple functions. Currently, computing MPH is very challenging even with data of modest size and scalable algorithms are needed for extracting this information efficiently. In this talk, we will present a new approach to Multidimensional Persistent Homology computation, inspired by Discrete Morse theory, based on reducing the input complex through the definition of a discrete gradient field. Experimental results will be presented showing the efficiency of our algorithm. Moreover, we will discuss future applications to scientific data analysis and visualization.

3.12 Computing and regularizing medial axes in 3D

Tao Ju (Washington University – St. Louis, US)

Medial axes is a classical concept in computational geometry and has been the basis of most of today's skeletal shape descriptors. In this talk, I will present our recent work in addressing two major roadblocks in using medial axes for 3D shape analysis: the difficulty in computing the medial axes of general 3D shapes, and the sensitivity of the medial axes to noise. First, I will describe a novel sampling-based algorithm for computing 3D medial axes that is numerically robust, simple to implement, and theoretically sound. Second, I will present a medial axes regularization method that is guided by a novel significance measure in 3D and capable of producing a family of skeletons that are descriptive and robust to noise. Finally, I will briefly discuss some applications of skeletons in biomedicine.

3.13 Geometric problems in isogeometric analysis

Bert Jüttler (Johannes Kepler Universität Linz, AT)

Part I: Isogeometric Segmentation of Free-form Solids
Part II: Arc Fibrations of Planar Domains

The two parts of this talk reported on the authors' recent work on selected geometric problems in isogeometric analysis (IGA). The framework of IGA, which was introduced in 2005 by T.J.R. Hughes et al., combines the mathematical technologies of numerical simulation and geometric design. The interaction of these two fields leads to the formulation of many new and challenging problems, two of which have been addressed in this talk.

The first part was devoted to the segmentation problem. Given a three-dimensional solid object in boundary representation, find a segmentation into topological cuboids, which admit a parameterization by tensor-product B-splines. We presented the isogeometric segmentation pipeline, which is a semi-automatic process that creates such a segmentation. The process, which is based on a recursive splitting algorithm, relies on a cost-function, which measures the quality of a splitting step. This part of the talk was based on joint work with M. Kapl, D. M. Nguyen, Q. Pan and M. Pauley.

The second part discussed a special instance of a parameterization. More precisely, we investigated the parameterization of planar domains by families of circular arcs. It is well known that star-shaped domains possess particularly simple polar parameterizations, which are formed by the line segments that connect a suitably chosen center with the points on the domain's boundary. The polar parameterization is valid (i.e., regular everywhere except for the center) if the center is located in the kernel of the domain. In the case of a domain with a smooth free-form boundary curve, the kernel is a convex polygon which is formed by (some of) the boundary's inflection tangents. These parameterizations possess numerous applications, most recently also including domain parameterization in isogeometric analysis.

Since the class of star-shaped domains is quite limited, we propose to increase the flexibility of the underlying polar parameterizations by considering circular arcs that connect the center with the points on the domain's boundary. Parameterizations that are regular everywhere except at the center are said to form an arc fibration of a planar domain. We analyze the existence of an arc fibration with a given center and present an algorithm that computes it in the affirmative case. In addition, we explore the arc fibration kernel of a domain, which contains the suitable center points. This part of the talk is joint work with M.-S. Kim and S. Maroscheck.

3.14 Statistical learning under group actions, with applications to cryo-electron microscopy

Joe Kileel (Princeton University, US)

In many problems in computer vision, robotics and image/signal processing, we wish to recover latent variables from observations suffering unknown shifts or rotations. One example is cryo-electron microscopy (cryo-EM), recognized by the Nobel Prize 2017 in Chemistry. Here the challenge is to estimate the 3D structure of a protein from many, very noisy 2D images taken at unknown viewing directions.

In this talk, I will place cryo-EM reconstruction inside a framework for statistical learning under noisy group actions. I will prove a tight relation between the sample complexity for statistical learning under noisy group actions and the invariant theory of the underlying symmetry group. On the algebra side, this motivates apparently new questions in invariant theory – to which we offer partial algorithmic answers in general. As for the cryo-EM case, we will give a novel ab initio 3D reconstruction algorithm, which is both sample and computationally efficient – at least under model assumptions.

I will review needed background information along the way. Overall, I hope to convey the flavor of some algebraic-geometric methods (still under development) in image analysis.

3.15 Invariant Representations of Shapes and Forms: Self Functional Maps

Ron Kimmel (Technion – Haifa, IL)

A classical approach for surface classification is to find a compact algebraic representation for each surface that would be similar for objects within the same class and preserve dissimilarities between classes. We introduce self functional maps as a novel surface representation that satisfies these properties, translating the geometric problem of surface classification into an algebraic form of classifying matrices. The proposed map transforms a given surface into a universal isometry invariant form defined by a unique matrix. The suggested representation is realized by applying the functional maps framework to map the surface into itself. The key idea is to use two different metric spaces of the same surface for which the functional map serves as a signature. As an example we suggest the regular and the scale invariant surface laplacian operators to construct two families of eigenfunctions. The result is a matrix that encodes the interaction between the eigenfunctions resulted from two different Riemannian manifolds of the same surface. Using this representation, geometric shape similarity is converted into algebraic distances between matrices. In contrast to geometry understanding there is the emerging field of deep learning. Learning systems are rapidly dominating the areas of audio, textual, and visual analysis. Recent efforts to convert these successes over to geometry processing indicate that encoding geometric intuition into modeling, training, and testing is a non-trivial task. It appears as if approaches based on geometric understanding are orthogonal to those of data-heavy computational learning. We propose to unify these two methodologies by computationally learning geometric representations and invariants and thereby take a small step towards a new perspective on geometry processing.

3.16 Duality of Convolution Operators: A Tool for Shape Analysis?

Christer Oscar Kiselman (Uppsala University, SE)

Duality is a term which represents a collection of ideas where two sets of mathematical objects confront each other. An important example is the dual of a normed space, where the linear forms on the space operate on its points. A most successful duality is that between the space $\mathcal{D}(\Omega)$ of test functions (smooth functions of compact support) and its dual space $\mathcal{D}'(\Omega)$ of distributions. The distributions that are not given by a locally integrable function live like ghosts in the dark, perceptible only through their actions on test functions.

The dual of a family of functions or sets reflects some but typically not all properties of the family. As an example, the supporting function of a set gives faithful information on the closed convex hull of the set, but forgets about any holes or cavities.

Mathematical morphology can be quite helpful in providing guiding concepts and ideas in the study of discrete convolution operators and many other topics, like discrete optimization. Here we apply these ideas to the duality of convolution operators.

It all goes back to Galois, two centuries ago. I put it under the common roof of lower and upper inverses of mappings between preordered sets.

3.17 Divergence-Free Shape Interpolation and Correspondence

Zorah Lähner (TU München, DE)

Joint work of Zorah Lähner, Marvin Eisenberger, Daniel Cremers

We present a novel method to model and calculate deformation fields between shapes embedded in $\mathbb{R}^D$. Our framework combines naturally interpolating the two input shapes and calculating correspondences at the same time. The key idea is to compute a divergence-free deformation field represented in a coarse-to-fine basis using the Karhunen-Loève expansion. The advantages are that there is no need to discretize the embedding space and the deformation is volume-preserving. Furthermore, the optimization is done on downsampled versions of the shapes but the morphing can be applied to any resolution without a heavy increase in complexity. We show results for shape correspondence, registration, inter- and extrapolation on the TOSCA and FAUST data sets.

3.18 Mathematical Morphology and Tropical Geometry

Petros Maragos (National Technical University of Athens, GR)

Mathematical Morphology and Tropical Geometry share the same max/min-plus scalar arithmetic and max/min-plus matrix algebra. In this Dagstuhl Seminar on shape analysis, we summarize their common ideas and algebraic structure, extend both of them using weighted lattices, and outline applications on geometry, dynamical systems, control and optimization, multimodal signal processing and machine learning. We begin our presentation by summarizing elementary max/min-plus morphological operators such as Minkowski shape and image operators and providing elementary concepts from tropical geometry. Further, we show how vision scale-space PDEs such as the Gaussian scale-space through Maslov Dequantization become morphological dilation/erosion scale-space PDEs. Then, we extend the underlying max-plus algebra to a max-* algebra where matrix operations and signal convolutions are generalized using a (max,*) arithmetic with an arbitrary binary operation "*" that distributes over max. This theory is based on complete weighted lattices and allows for both finite- and infinite-dimensional spaces. We outline applications of dynamical systems and geometry on weighted lattices including min-plus systems for distance transforms, max-product systems for tracking salient events in multimodal videos, max-fuzzy-norms systems, and max/min-plus perceptrons for machine learning. Finally, we outline the optimal solution of systems of max-* equations using weighted lattice adjunctions and projections, possibly with sparsity constraints, and show how it applies to optimal regression for fitting max-* tropical curves on data.

3.19 Towards polarisation in the wild: when the illumination makes the difference

Roberto Mecca (Italian Institute of Technology – Genova, IT)

In this talk I present recent results regarding the shape from polarisation problem where a new differential formulation has been derived. It consists of a homogeneous linear partial differential equation that provides the level-set of the surface for the diffuse polarisation phenomena. This new formulation allows to derive geometrical hints of the 3D shape for specular polarisation as well since the shift in the phase angle brings the level-set orthogonally to be shifted with respect to the diffuse case. The interchangeability of the proposed differential model between diffuse and specular reflection allows a step forward for studying the mixed polarisation problem.

The ability to extract level-set from dielectric objects using polarimetric images has been tested using real world data with different type of illuminations. The capability of this method to discern diffuse from specular reflections shows the importance of the illumination.

3.20 Geo'metric' Learning: Deep Isometric Manifold Learning Using Sparse Geodesic Sampling

Gautam Pai (Technion – Haifa, IL)

We explore a fully unsupervised deep learning approach for computing distance-preserving maps that generate low-dimensional embeddings for a certain class of manifolds. We use the Siamese configuration to train a neural network to solve the problem of least squares multidimensional scaling for generating maps that approximately preserve geodesic distances. By training with only a few landmarks, we show a significantly improved local and non-local generalization of the isometric mapping as compared to analogous non-parametric counterparts. Importantly, the combination of a deep-learning framework with a multidimensional scaling objective enables a numerical analysis of network architectures to aid in understanding their representation power. This provides a geometric perspective to the generalizability of deep learning.

3.21 Morphological Adjustments for Provably Robust Machine Learning

Frank R. Schmidt (Robert Bosch GmbH – Stuttgart, DE)

In recent years, deep learning has become the state of the art in machine learning for image classification. Trained on a specific dataset, it can usually generalize to images from the same dataset that were not observed during training. Nonetheless, it suffers severely from an artefact that is known as "adversarial perturbation", i.e., small perturbation of correctly classified images will be classified incorrectly. As long as adversarial examples are possible, deep networks cannot be used for products that are expected to make autonomous decisions in critical situations.

We will show how one can certify those elements of a dataset that do not suffer from these adversarial examples. To this end, each data point is augmented via a morphological operation. The resulting set is tracked through a network. In order to certify that no element of this set is misclassified, one has to solve an ILP. We show that a specific relaxation leads to reliable certificates and that this certification approach can be used in order to make deep networks more robust.

3.22 Principal geodesic analysis in shell space

William Smith (University of York, GB)

Joint work of William Smith, B. Heeren, C. Zhang, M. Rumpf

Important sources of shape variability, such as articulated motion of body models or soft tissue dynamics, are highly nonlinear and are usually superposed on top of rigid body motion which must be factored out. We propose a novel, nonlinear, rigid body motion invariant Principal Geodesic Analysis (PGA) that allows us to analyse this variability, compress large variations based on statistical shape analysis and fit a model to measurements. For given input shape data sets we show how to compute a low dimensional approximating submanifold on the space of discrete shells, making our approach a hybrid between a physical and statistical model. General discrete shells can be projected onto the submanifold and sparsely represented by a small set of coefficients. We demonstrate two specific applications: model-constrained mesh editing and reconstruction of a dense animated mesh from sparse motion capture markers using the statistical knowledge as a prior

3.23 Analysis of Planar Ornaments Within and Beyond Symmetry Groups

Sibel Tari (Middle East Technical University – Ankara, TR)

The scientific study of the ornaments constructed by repeating a base motif is the study of the repetition structure, i.e. symmetry. Planar case where only four geometric operations (translation, rotation, mirror reflection and glide reflection) are permitted has been thoroughly studied; the repetition structure in planar case could take one of the seventeen forms. In this talk, using examples from Islamic art and Escher, we argue that the artistic side of the problem is as interesting as the mechanical repetition structure, hence requires scientific inquiry. Specifically, intriguing color permutations, clever choices of asymmetric interlocking forms, hyper-symmetries, symmetry breaking ideas, all that come with the artistic freedom, make patterns appear more symmetric than they really are. After going through examples, we suggest a change of perspective when analyzing ornaments and present custom designed ornament dataset.

3.24 Differential approaches to Shape-from-X Problems

Silvia Tozza (Sapienza University of Rome, IT)

In this talk I will give a small overview of differential approaches used to solve problems belonging to the Shape-from-X class, limiting the presentation to the orthographic projection case. More in details, I will start from the classical orthographic Shape-from-Shading (SfS) problem with a single input image, explaining why it is an ill-posed problem and which are the possible directions to arrive to a well-posed problem. Hence, I will move to the Photometric stereo SfS problem firstly with two images plus boundary conditions and secondly with three input images in order to avoid the need of boundary conditions. Finally, I will talk about the Shape-from-Polarisation problem by showing how combine shading and polarisation information in order to directly reconstruct the surface height. We are able to linearise the constraints involved, arriving to a new unified PDE formulation of several proposed methods with different nice properties (e.g. albedo invariant or phase angle invariant).

3.25 Medians and related measures for multidimensional data

Martin Welk (UMIT – Hall in Tirol, AT)

Median filtering is a simple and robust procedure for e.g. denoising and aggregating 1D real data. Various generalisations to higher-dimensional data space have been proposed which differ in their theoretical properties such as degree of equivariance w.r.t. different transformation groups, behaviour when the data aggregate near hyperplanes etc., but differ also in their computational expense. In this talk important multidimensional median concepts are juxtaposed: the L^1 median, the transformation–retransformation L^1 median, Oja's simplex median, the half-space median and the convex-hull-stripping median. They are compared regarding their use in filtering images, in filtering geometric data in $\mathbb{R}^d$ and on Riemannian manifolds. Relations to PDE filters for images are discussed. Based on experimental evidence an interesting relation between the convex-hull-stripping median and a PDE well-known in shape analysis is shown.

3.26 Estimation of Laminar Coordinate Systems between Two Surfaces

Laurent Younes (Johns Hopkins University – Baltimore, US)

Given two disjoint open surfaces, we discuss the problem of estimating a parametrization $(x, y, z) \to F(x, y, z)$ of an open subset of the three-dimensional space such that the set $[z = 0]$ corresponds to one of the surfaces, $[z = 1]$ to the other one, and for each t, $(x, y) \to F(x, y, t)$ is an embedding onto the surface $[z = t]$, with the additional constraint that the derivative of F with respect to t is always normal to this surface.

The resulting construction is a natural way to build a foliation between the two surfaces for which the length of the transverse lines provides a good definition of thickness of the considered volume. It is designed with the objective of analyzing cortical volumes, for which a robust definition of thickness is essential for the characterization of degenerative diseases. Our estimation algorithm uses a version of the large deformation diffeomorphic metric mapping between surfaces in which normality constraints of the velocity field are enforced.

The estimated coordinate system can also be used to estimate "cortical layers" as laminar segmentations of the inter-surface region that satisfy a local equi-volumetric condition called Bok's hypothesis. Experimental results on human and feline brains will be presented as illustrations.

Participants

- Marianne Akian
Ecole Polytechnique – Palaiseau, FR
- Jesús Angulo
Ecole des Mines de Paris, FR
- Péter Balázs
University of Szeged, HU
- Michael Breuß
BTU Cottbus, DE
- Alfred M. Bruckstein
Technion – Haifa, IL
- Sara Brunetti
University of Siena, IT
- Vasileios Charisopoulos
Cornell University, US
- Robert Dachsel
BTU Cottbus, DE
- Leila De Floriani
University of Maryland – College Park, US
- Panagiotis Filntisis
National Technical University of Athens, GR
- Federico Iuricich
Clemson University, US
- Tao Ju
Washington University – St. Louis, US
- Bert Jüttler
Johannes Kepler Universität Linz, AT
- Joe Kileel
Princeton University, US
- Ron Kimmel
Technion – Haifa, IL
- Christer Oscar Kiselman
Uppsala University, SE
- Zorah Lähner
TU München, DE
- Petros Maragos
National Technical University of Athens, GR
- Roberto Mecca
Italian Institute of Technology – Genova, IT
- Gautam Pai
Technion – Haifa, IL
- Frank R. Schmidt
Robert Bosch GmbH – Stuttgart, DE
- William Smith
University of York, GB
- Sibel Tari
Middle East Technical University – Ankara, TR
- Silvia Tozza
Sapienza University of Rome, IT
- Martin Welk
UMIT – Hall in Tirol, AT
- Laurent Younes
Johns Hopkins University – Baltimore, US

Report from Dagstuhl Seminar 18431

Computational Aspects of Fabrication

Edited by

Bernd Bickel[1], Marc Alexa[2], Jessica K. Hodgins[3], and Kristina Shea[4]

1 **IST Austria, AT,** bernd.bickel@ist.ac.at
2 **TU Berlin, DE,** marc.alexa@tu-berlin.de
3 **Carnegie Mellon University – Pittsburgh, US,** jkh@cs.cmu.edu
4 **ETH Zürich, CH,** kshea@ethz.ch

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 18431 "Computational Aspects of Fabrication".

Seminar October 21–26, 2018 – http://www.dagstuhl.de/18431

2012 ACM Subject Classification Computing methodologies → Shape representations, Computing methodologies → Appearance and texture representations, Computing methodologies → Physical simulation, Computing methodologies → Shape modelingApplied computing → Computer-aided manufacturing, Hardware → Design for manufacturability

Keywords and phrases Computational Fabrication, Computational Design, Engineering Design, 3D Printing, 4D Printing

Digital Object Identifier 10.4230/DagRep.8.10.104

1 Executive Summary

Marc Alexa
Jessica K. Hodgins
Kristina Shea

As manufacturing goes digital, we are facing a fundamental change in the workflow of fabrication. While access to advanced digital fabrication and 3D-printing technology becomes ubiquitous and provides new possibilities for fabricating complex, functional, multi-material objects with unconventional properties, its potential impact is currently limited by the lack of efficient and intuitive methods for content creation. Existing tools are usually restricted to expert users, have been developed based on the capabilities of traditional manufacturing processes, and do not sufficiently take fabrication constraints into account. Scientifically, we are facing the fundamental challenge that existing simulation techniques and design approaches for predicting the physical properties of materials and objects at the resolution of modern 3D printers fail to scale well with possible object complexity.

To achieve significant progress, we need a deep understanding of interdisciplinary fundamentals: Shape, Appearance of Shape and Materials, Validated Simulation, and Engineering Design. The purpose of this Dagstuhl Seminar is to bring together leading experts from academia and industry in the area of computer graphics, geometry processing, mechanical engineering, human-computer interaction, material science, and robotics. The goal is to address fundamental questions and issues related to computational aspects of fabrication, build bridges between related fields, and further pioneer this area.

There has been a considerable growth in the number of articles treating aspects of computational fabrication, scattered across multiple disciplines and journals. In this seminar we gathered together these various threads and described the computational accomplishments and outstanding challenges. Researchers from different communities analyzed which existing fabrication workflows could benefit most from computation and identify novel application domains, with the aim of cross-fertilizing ideas between disciplines. The main goal of this seminar was identifying and reporting common grand challenges and developing a roadmap for addressing them. Additionally, the seminar seeked to discuss and establish standards and best practices for sharing research results, code, and hardware prototypes, facilitating reproducibility and reusability of results among disciplines. An important aspect of this was to analyze teaching and learning needs for new students in the field, and coordinating the development of teaching material.

2 Table of Contents

3 Overview of Talks

3.1 Computational Nanofabrication

Thomas Auzinger (IST Austria – Klosterneuburg, AT)

Joint work of Thomas Auzinger, Wolfgang Heidrich, Bernd Bickel
Main reference Thomas Auzinger, Wolfgang Heidrich, Bernd Bickel: "Computational design of nanostructural color for additive manufacturing", ACM Trans. Graph., Vol. 37(4), pp. 159:1–159:16, 2018.
URL https://doi.org/10.1145/3197517.3201376

Multiphoton lithography – also known as Direct Laser Writing – provides an accessible nanofabrication method that resembles the workflow of 3D printing. As it solidifies the photoresist only in the focus region of the femtosecond-pulsed laser beam, it allows nearly unrestricted freeform writing of nanostructures. We utilized this fabrication method to create structural colorization of glass surfaces by transparent nanostructures. The structures themselves were discovered by a fully automatic inverse design method based on electromagnetic simulation. Efforts by other groups used Direct Laser Writing to create nanolattices, whose specific strength surpases that of steel, as well as microrobots, microneedles, and cell cages.

3.2 Fine Art Appearance Fabrication

Vahid Babaei (MPI für Informatik – Saarbrücken, DE)

Reproduction of fine art objects has been a topic of interest for many decades, pursued by many technologies. In general, the results have not been satisfactory as the quality bar is very high for this particular application. Multi-material 3D printing is the latest technology and a new hope for a physical reproduction with archival quality. Among fine art artifacts, we find paintings as an excellent case-study due to the rich appearance, unique challenges and rather convenient fabrication. Multi-material 3D printing is not only able to reproduce the fine 3D geometry present in many forms of paintings, but also other appearance attributes, such as spectral color and gloss. In this talk, I discuss the opportunities enabled by 3D printing for fine art reproduction. In contoning, for example, one can eliminate the traditional halftoning artifacts using the inherent ability of 3D printers in layering inks on top of each other. Using the same property of 3D printers, I show that the spectral gamut of a 3D printer can exceed the one of a 2D printer significantly thereby enabling truly spectral reproduction of fine art. I also speak about challenges where aside from computational problems, such as accurate volumetric prediction of appearance, an open hardware platform that gives the possibility of tuning both machines and materials is indispensable.

3.3 Fabrication-aware design: Where we are? Where are we going?

Amit Haim Bermano (Princeton University, US)

We have recently published a state-of-the-art report and a book about the advances made in computational fabrication by the computer graphics community. By drawing conclusions from this work, I would like to discuss one of the pressing issues I believe should be investigated in the near future in the context of fabrication-aware design: Design through objectives, instead of geometry, using a hierarchical, modular, representation.

3.4 Computational Design of Physical Characters and Structures

Moritz Bächer (Disney Research – Zürich, CH)

Stimulated by advances in manufacturing, fabrication-oriented design has gained an increasing level of interest from the graphics community. With modern manufacturing technologies, we can build physical characters and large-scale structures of nearly unbounded complexity by shifting the design burden to computational approaches.

In a first part, I will talk about a technique that aids with the design and fabrication of elastically-deforming kinetic wire characters. Our technique takes as input a network of curves or a skeletal animation, then estimates a cable-driven, compliant wire structure which matches user-selected targets or keyframes as closely as possible. To enable large local deformations, we shape wire into functional spring-like entities, optimizing their stiffness. We use consumer-grade hardware to fabricate our optimized designs.

In a second part, I will talk about a worst-case optimization of structures that are weak in tension. I will introduce a technique to derive distance metrics from failure criteria, formulating a stress objective that accounts for asymmetries in the tensile and compressive strength of common build materials. Parameterizing uncertainties in load cases, we introduce a formulation, optimizing structures under worst-case loads. I will show several optimized structures, tailored for manufacturing on large-scale binder jetting technologies.

3.5 Meaningful Applications of 3D Printing. Key computational components for success.

Sabine Demey (Materialise HQ – Leuven, BE)

In Industry 4.0, the reasons why people turn to 3D Printing are still largely the same as before. The reasons can be summarized as Design, Cost and Time: Freedom of design, function before form, affordable small series, no tool-making required, faster design iterations, faster time-to-market, etc... People increase expectations though. Products should be highly personalized, be of high quality, with many details, in multiple materials, etc. and the higher complexity is expected to be supported at no added cost. In support of this complex journey from idea to 3D printed product, software has enormous computationally powerful algorithms under the hood and the need for these algorithms is increasing more than ever. The AI hype is also contributing to this. I will bring you in touch with a variety applications using 3D printing to create a better and healthier world. These applications will reveal one by one the need for powerful computations.

3.6 The Design and Fabrication of Smart Textiles

Laura Devendorf (University of Colorado – Boulder, US)

The term "smart textiles" describes fabrics with sensing and actuation capabilities integrated into their structure at the yarn-level. While a consumer-level infrastructure for the rapid fabrication of textiles is emerging, we have a lack of design tools to effectively leverage this infrastructure for new application domains. This talk will describe the pipeline of textiles fabrication, the structural properties of woven fabrics that open up new spaces for computational design, and how textiles might require new forms of human machine collaboration.

3.7 AM representation that enables design

Georges Fadel (Clemson University – Clemson, US)

Much of the focus on representation for additive manufacturing has been on the format that enables 3D printers to efficiently and reliably translate the designers' creation into a solid object. The STL format has been a significant component of the success of the AM, but has also been recognized to have shortcomings, and researchers have proposed new formats such as AMF. These formats however, are not adequate for the designer who seeks to take advantage of the full potential of AM, specifically the ability to modify the shape and topology, to include material variability and anisotropy. Once the design is done, existing formats would allow the 3D printer to print it, but since CAD tools are surface based representations that do not provide this flexibility, researchers keep trying to find alternate approaches. The voxel representation may be adequate again for the interface to the printers, but how can it be used in the design? Topology optimization and a two level approach may be appropriate, the lower level elements being of the scale of voxels. However, traditional homogenization approaches are limiting, and do not provide the designer with the full flexibility needed to design and manufacture novel artifacts since they depend on full Y-Periodicity and on a small size cell as compared to the design space in all directions. We need a representation, which provides the designer with a volumetric design tool that allows design for additive manufacturing.

3.8 Human-Centered Interfaces for Autonomous Fabrication

Madeline Gannon (Atonaton – Pittsburgh, US)

This talk discusses the potentials and merits of fabrication machines becoming collaborative companions in computational fabrication processes. I highlight several technical challenges of bypassing computers to directly engage with fabrication machines. I then share my recent, ongoing research into fluid, intuitive interfaces for industrial robots. As one of our most versatile and adaptable fabrication machines, industrial robots are a reliable and agnostic hardware platform for several additive and subtractive processes. The goal of this research is to demonstrate novel relationships between people and machines that can augment our existing computational design and fabrication processes in fruitful ways.

3.9 New computational tools to support Design for Additive Manufacturing (DfAM) in the early stages of the Product Development Process

Serena Graziosi (Polytechnic University of Milan, IT)

The rapid development of additive manufacturing technologies is continuously providing new stimuli to creative people and industries. From modelling and printing complex geometries, their interest has now moved towards the possibility to create complex and smart systems by mimicking, replicating and eventually extending the complexity of natural systems. However, to reach such a challenging target, designers need support in understanding and mastering the complexity of the phenomena and thus of the aspects determining the system behaviour and architecture. Most of the current design tools were conceived to support detailed design activities and are not effective in supporting designers in such an exploration which should take place during the concept design phase of the product development process. Indeed, it is during this phase that new ideas are generated, and designers have enough time to investigate how to exploit the potentialities of additive manufacturing technologies in their products, i.e., how to Design for Additive Manufacturing (DfAM). This talk aims to stimulate the development of new computational tools for the concept design phase. These tools should help designers to understand the system behaviour in real-time through quick and easy to set-up simulations, even simplified ones, for example by combining 3D modelling with multiphysics analyses, and by letting designers make informed decisions.

3.10 Taking out the Hard Edges: New Printers and Processes for Fabricating Soft Materials

Scott Hudson (Carnegie Mellon University – Pittsburgh, US)

Much of what I do involves building unique new 3D printers and other fabrication machines, and I have been particularly interested in expanding our ability to create soft objects. Soft objects have unique advantages, particularly in objects used by humans, and yet we have comparatively few ways of digitally fabricating them. In this talk I will consider several machines which work with fibers as a material, such as a 3D printer which prints in needle felted yarn, and printing using electrospun fibers, as well as a new inexpensive and accessible silicone rubber printer.

3.11 Robust Geometry Processing: the Life Cycle of a Messy Shape

Alec Jacobson (University of Toronto, CA)

I will discuss my vision for a robust geometry processing pipeline. I propose that we reject the traditional "garbage in, garbage out" policy. Instead, we should return to first principles and hunt for solutions that adapt and gracefully degrade in the presence of messy inputs. I will highlight a few very recent successes (winding numbers for inside/outside classification, Boolean operations on triangle meshes, and tetrahedral volumetric meshing). I will emphasize their importance to computational fabrication via applications to 3D printed movies, stop motion and generalized Matryoshka dolls.

3.12 A Software Platform for Algorithmic Design

Lin Sebastian Kayser (Hyperganic Technologies AG – München, DE)

Five years ago I started a company called Hyperganic with the goal to enable algorithmic design of highly complex objects. I'd like to share the progress that we've made and show how using our voxel-based approach, we can break down the entry barriers of generative design applications, by making it very easy to create even complex algorithms and solutions.

3.13 Human Perception of 3D Shapes

Manfred Lau (City University – Hong Kong, HK)

In this talk, I will describe two projects that look at the human perceptual properties of 3D shapes. The first is tactile saliency. While there has been previous research in the area of visual saliency, both for images and for 3D meshes, we introduce the concept of tactile mesh saliency. For example, a point on a 3D mesh is more tactile salient than another if it is more likely to be grasped. I describe the data collection process and the learning method for computing tactile saliency. The second perceptual property is the softness of 3D meshes, where we look at how humans perceive the softness of the surface of virtual meshes. We take a similar approach as in the first project to compute softness. I will also discuss applications of these perceptual properties including for fabrication and describe potential challenges for the future.

3.14 Fused filament fabrication of parts with gradients of properties

Sylvain Lefebvre (LORIA & INRIA – Nancy, FR)

In this presentation I describe some of our ongoing research to grade properties within parts, such as elasticity and color. I will first discuss how to represent complex infill structures in a compact, efficient manner using pixel shaders written in the OpenGL shading language. These shaders can be provided directly to our slicing software (IceSL) which uses them to produce toolpaths within parts. I will then describe ongoing work on micro-layering to produce color gradients using FDM printers. Finally, I will discuss some still open challenges regarding support structures and surface finish.

3.15 Designing Volumetric Truss Structures

David I. W. Levin (University of Toronto, CA)

Voxels, voxels everywhere has been the long standing mantra of computational design algorithms. This seems at odds with output of such procedures which are typically well defined, low dimension primitives such as curves. In this talk I will discuss the implications of this representational conflict as well as detail a new voxel free method for optimal truss generation for 3D printing. I'll conclude with thoughts on future problems in computational design.

3.16 Appearance and Interiors Optimization for Extrusion-based Fabrication

Lin Lu (Shandong University – Qingdao, CN)

My talk basically includes two parts, appearance and interiors optimization. First, I share some attempts on printing complex shapes. Different shape classes such as hair/fibers and architectural models have driven research toward class-specific solutions. 3D trees are an especially challenging case for 3D manufacturing. They consist of non-volumetric patch leaves, an extreme amount of small detail often below printable resolution and are often physically weak to be self-sustainable for single material. I describe the knowledge based optimizations in terms of both geometric and physical constraints and show 3D printed trees. Then I discuss the defects of extrusion based fabrication, e.g, FDM, and take an image carving example to show the details adaption results based on the physical constraints. In the second part, I talk the interiors in closed-cell and open-cell structures and discuss the advantages and disadvantages of these structures in terms of manufacturing feasibilities and applications. Challenges remaining in the mentioned problems are discussed in the end.

3.17 Procedural and stochastic microstructures for AM

Jonas Martinez-Bayona (INRIA Nancy – Grand Est, FR)

Additive Manufacturing (AM) technologies are now capable of fabricating microstructures at the scale of microns, therefore enabling to precise control of the macroscopic physical behavior. This control empowers a wide range of industrial applications by bringing high-performance customized materials. Microstructures for AM will play a decisive role in the factory of the future, but several challenges remain aside. In this talk we consider procedural, stochastic, and fabricable microstructures, with a controlled macroscopic physical behavior. As a result of their stochastic nature such microstructures afford for free grading and embedding of microstructures into objects, hence avoiding the limitations imposed by periodic structures.

3.18 Machine Learning for AM Monitoring

Sara McMains (University of California – Berkeley, US)

I describe a machine learning based approach to in-situ quality monitoring for Selective Laser Melting (SLM). Our collaborators at Lawrence Livermore National Laboratory modified the SLM apparatus to include a high-speed camera that follows the mirror/galvanometer-controlled path of the scanning laser to gather in-situ video data of the melt pool. After fusing separated experimental tracks and removing unfused powder, a height field is obtained (ex-situ) by scanning with a structured light microscope. Our two-stage height field segmentation algorithm classifies track, etch, & background pixels in order to automatically label 10-frame videos with track width (for regression training) and presence or absence of breaks (for classification training).

A CNN architecture whose hyperparameters are tuned for one modality successfully predicts both width and continuity.

3.19 Fabrication and 3D Modeling at Adobe

Radomir Mech (Adobe Inc. – San José, US)

In this talk I overviewed technologies that we have most recently developed and research direction that we are pursuing at Adobe Research. I showed a project on mapping 2D dielines to 3D folded geometry that can be used to place decals so that they show correctly on the folded object. The second project was on easy manipulation of 3D objects using handles based on wires created on salient features. In addition, I presented future directions in the area of 3D modeling: assembly based modeling, parametric modeling and ways to explore parametric space of 3D models.

3.20 Creation for everyone: Broadening participation, increasing accessibility, democratizing engineering, and other warm fuzzy goals

Ankur Mehta (University of California at Los Angeles, US)

Computational fabrication broadly aims to automate challenging design tasks in the process of creation. We can therefore use this to bring the act of creation to those with minimal access to resources, expertise, or background; I posit that this can provide the most significant benefit to society by incentivizing education in the demographics that most need it. We therefore need to consider extremely low cost manufacturing processes, accessible and intuitive design interfaces, and the needs and expectations of the target users. I show some initial work building on this motivation towards design automation for inexpensive paper-based robots.

3.21 From Material to Autonomy via Programmability

Shuhei Miyashita (University of York, GB)

Living systems in general feature a functionality to heal their structures when damaged while artificial systems do not. This is mainly because living systems are made in bottom-up, by protein molecules. Such self-assembly processes happen in a way that structures increase their dimensions; parts (e.g. amino acids) form one dimensional strings that are further reconfigured to a higher dimensional structures (e.g. 3D proteins). This talk presents heat-driven self-folding origami robots: the mechanism, capabilities, potential, and the limitation.

3.22 Materializing performance-driven form for architecture

Caitlin Mueller (MIT – Cambridge, US)

My research (digitalstructures.mit.edu) focuses on performance-driven design in architecture, often from the perspective of structural engineering, and how freedom of design expression and authorship can be reconciled with efficiency and related performance goals. A subset of this work relates to materialization, and specifically to tackling the geometric complexities that emerge from performance-driven design processes. While building-scale construction currently still favors standardization and regularity to an overwhelming degree, there is potential in the future to achieve performance-driven design complexity through new computationally driven methods for fabrication and assembly. In my talk, I'll discuss examples in 3D spatial extrusion, engineered timber, and reinforced concrete, all at architectural scale.

3.23 Rapidly Deployable Elastic Gridshells

Julian Panetta (EPFL – Lausanne, CH)

I will present a new approach for designing elastic gridshells where the deployed shape is encoded directly into the layout and cross section geometry of the beams. Traditional gridshells employ a regular grid layout that by itself has no knowledge of the target shape; the deployed shape is determined by what planar boundary curve is cut from this grid and how the beam endpoints on this boundary are moved to their target locations. This deformation causes each beam to buckle into a 3D curve, but the final shape that arises can depend on the order in which the endpoints are moved. Our work seeks to simplify the deployment process by designing spatially varying grids with a single easily actuated deployment path from the flat assembly configuration to a uniquely specified curved shape. I will present some examples that we have created and discuss some of the challenges of designing and robustly simulating these structures.

3.24 Design of Complex Assemblies

Mark Pauly (EPFL – Lausanne, CH)

Interlocking assemblies have a long history in the design of puzzles, furniture, architecture, and other complex geometric structures. The key defining property of interlocking assemblies is that all component parts are immobilized by their geometric arrangement, preventing the assembly from falling apart. Computer graphics research has recently contributed design tools that allow creating new interlocking assemblies. However, these tools focus on specific kinds of assemblies and explore only a limited space of interlocking configurations, which restricts their applicability for design.

In this talk, I will describe a new general framework for designing interlocking assemblies. The core idea is to represent part relationships with a family of base Directional Blocking Graphs and leverage efficient graph analysis tools to compute an interlocking arrangement of parts. This avoids the exponential complexity of brute-force search. The algorithm iteratively constructs the geometry of assembly components, taking advantage of all existing blocking relations when constructing successive parts. As a result, our approach supports a wider range of assembly forms compared to previous methods and provides significantly more design flexibility. We show that our framework facilitates efficient design of complex interlocking assemblies, including new solutions that cannot be achieved by state of the art approaches.

3.25 Rapid prototyping of rapid prototyping machines

Nadya Peek (University of Washington – Seattle, US)

Digital fabrication machines are becoming less expensive and therefore more accessible. However, they largely follow the same workflow: g-code moves a 3-axis gantry with a 3d print head or spindle. Custom digital fabrication machines enable diverse fabrication practices, including different kinematic models, different end effectors, and different user interactions. But building custom digital fabrication machines requires time and expertise. In this talk, I will present a variety of modular parts for machine building, including modular networked controllers, modular mechanical machine axes, and workflow composition software. Finally, I'll show how non-expert machine builders are able to construct lots of different kinds of machines using this modular machine infrastructure.

3.26 Molding is the new Black

Nico Pietroni (University of Technology – Sydney, AU)

While 3D printing technologies are becoming faster and more precise, classical manufacturing techniques remain the first choice for most industrial application scenarios. Industrial production is still largely dominated by casting techniques: casting scales well with the number of copies, supports a wide spectrum of materials, and ensures high geometric accuracy.

In this talk I will show recent advancement on geometry processing and shape analysis for the automatic design and fabrication of 3D printed molds. I will show the technical details and the effectiveness of new technologies that use 3D printing to automatise industrial production processes.

3.27 Mobile Fabrication

Thijs Roumen (Hasso-Plattner-Institut – Potsdam, DE)

We have gotten accustomed to mobile computing, whenever we encounter information problems, we solve them on the go. But when it comes to mechanical problems, we either just accept this or panic. Would it be possible to use the power of digital fabrication in that mobile context to solve our mechanical problems as we encounter them? I think so! In this talk, I will present the vision and challenges for making mobile fabrication a reality, I outline some of the current and future projects I am involved in to make mobile fabrication a reality and highlight my goal of forming a community of people to overcome these hurdles.

3.28 Toolpathing for 3D Printing

Ryan Schmidt (Gradientspace – Toronto, CA)

3D Printing is widely used in fabrication research and in practice, but the focus of most research is either at the level of geometry or machine. Comparatively little research targets the toolpathing algorithms that provide the interface between designs and robots. I will discuss how toolpathing is a major determinant of manufacturability, and that current toolpathing techniques leave much to be desired. Specific topics will include examples of novel design spaces exposed by small changes to the toolpathing pipeline, attempts to resolve assembly tolerances at the toolpath level, and potential directions for shape-aware toolpathing strategies.

3.29 Predictability and Robustness in Design for Additive Manufacturing (AM)

Carolyn C. Seepersad (University of Texas at Austin, US)

Additive manufacturing (AM) is making a profound impact on the way engineers realize customized parts, but fully realizing the manufacturing freedom afforded by AM requires some significant advances in engineering design methods and tools. For some additive manufacturing applications, simulation-based design tools may be required to explore a hierarchy of features, ranging in size from microns to meters. When these features are fabricated, however, AM systems typically induce significant deviations from intended geometry and mechanical performance. Designers need comprehensive statistical models that characterize this variability. Furthermore, they need design tools that use these models to provide real-time feedback on the constraints and process-structure-property relationships relevant to specific AM technologies, and this Design-for-AM feedback is needed during the design process, rather than at the end. To address these challenges, a design exploration approach has been established for creating inverse maps of promising regions of a hierarchical structural/material design space. The approach utilizes Bayesian network classifiers for identifying sets of promising solutions to a materials design problem by efficiently utilizing information gained from simulations, experiments, and/or expert knowledge. It also makes use of statistical characterization of geometric features and material properties to identify robust designs. The capabilities of the approach are demonstrated by applying it to the hierarchical design of negative stiffness metamaterials for energy absorption applications.

3.30 4D Printing: The new frontier

Kristina Shea (ETH Zürich, CH)

4D printing considers the 3D printing of designs with active materials so that they can function as machines, e.g. providing locomotion or reconfiguring their shape, without the need for drop-in components or 3D printing of conductive elements. The fourth dimension in 4D printing is time. This talk highlights our research on novel designs for 3D printed, tunable, multi-stable structures and an untethered swimming robot both of which activate through changes in temperature using a combination of shape memory polymers and bi-stability. Finally, computational design problems and results are shown for computing and optimizing 3D printed, shape morphing 2.5D and 3D structures activated through bi-stability, shape memory polymers and pneumatics.

3.31 Design of Meta-materials for Digital Fabrication

Melina Skouras (INRIA – Grenoble, FR)

Meta-materials are materials that owe their bulk properties primarily from their geometric structure. Designing meta-materials with extremal properties is a challenging task and is usually done by hand. In this talk, I will present a pipeline to (1) automatically characterize the range – or gamut – of mechanical properties that can be achieved by assemblies of 3D-printed voxels of base materials, (2) automatically identify microstructures sharing common geometric traits and cluster them into distinct families, and (3) to generate parametric templates for each family allowing to represent the microstructures at arbitrary resolutions.

3.32 Experiments in Extrusion-based Clay Printing

Bernhard Thomaszewski (Université de Montréal, CA)

Extrusion-based clay printing is an accessible technology for additive manufacturing of customized ceramics. There are, however, several aspects which make this process challenging.

Unlike conventional thermoplastics, clay is a viscoplastic material that remains comparatively soft throughout printing. Its limited load bearing capacities can lead to deformations or even collapse during printing. A second challenge is that discontinuities in the print paths lead to artifacts that quickly amplify and lead to failure. This is particularly problematic when generating support structures, which typically rely on disconnected paths. Finally, uneven drying of the model after printing can lead to large deformations and even fracture. In this talk, I described these challenges in more detail and indicated some avenues for possible solutions.

3.33 Making 3D Prints more Functional using Electronics and Machine Learning

Nobuyuki Umetani (AUTODESK Research – Toronto, CA)

In this talk, I introduce our recent attempts to enhance the functionality of the 3D prints. First, I talk about our technique to firmly mount electronic circuits on the top of the 3D prints to add various modalities such as light, sound and movements. Then, I talk about the use of the machine learning in the context of the 3D shape generation and aerodynamically efficient shape design. Finally, I discuss the challenges in the data-driven 3D design.

3.34 Appearance Fabrication

Philipp Urban (Fraunhofer IGD – Darmstadt, DE)

The Bright Future of Metameric Blacks.

3.35 Geometric Computing for Multi-Axis Additive Manufacturing

Charlie Wang (The Chinese University of Hong Kong, HK)

I will present our recent development of using multi-axis motion to conduct material accumulation along dynamically varied directions. Our development results in two approaches that mainly focus on how to avoid the additional supporting structures in a framework of volume-to-surface and then surface-to-curve decomposition. I will discuss a few future extensions of this framework so that to strengthen the function of 3D printed parts.

3.36 Appearance Fabrication: Challenges of Production Deployment

Tim Weyrich (University College London, GB)

Appearance fabrication aims at creating custom reflectance properties on real-world surfaces. As the inverse of appearance acquisition, it starts from a digital description of spatio-angular reflectance properties and seeks to alter physical surface to match that description. My talk provides a brief overview over the state of the art and then discusses a variety of key challenges when carrying academic proofs of concept into production environments, including the combination of multiple working principles for appearance fabrication, the challenge of countering visual artefacts of fabrication methods, the needs of mass production, and last not least of finding appearance specifications that have value in the applied context.

3.37 Multi- Material/Modality/Scale/Axis: Realizing Multi-Functional Products with Next-Generation AM Processes

Christopher Bryant Williams (VPI – Blackburg, US)

The maturation of Additive Manufacturing processes has provided access to unparalleled design freedom in product realization. For perhaps the first time, manufacturing capability outpaced that of the available design processes. While the recent acceleration in research of design automation tools and methodologies has closed this perceived gap, emerging AM processes capable of working at multiple length scales using multiple AM modalities along multiple axes to fabricate with multiple materials are expanding the gap between design and manufacturing yet again. The goal of this talk is to present these emerging AM capabilities in order to reflect on the corresponding needs of the next-generation design and computation tools.

3.38 Multi-Species Robot Ecologies for Space Making

Maria Yablonina (Universität Stuttgart, DE)

Over the past few decades, digital fabrication processes have been gaining momentum in the field of architecture and design. While the construction industry is racing to increase the efficiency of existing processes through automation of work in a conventional construction environment; the field of architectural research is implementing robotic technology towards discovering new materials, fabrication methods and ultimately a new design space. An industrial robot arm has become a somewhat iconic symbol of this undertaking. Research labs and institutions across the world push the boundaries of what is possible in architecture by augmenting robots with custom end-effectors and software, appropriating them for architectural tasks in all possible materials from brick and wood to 3D printed concrete and carbon fiber. However, could it be that today, when we are arriving at the point where processes no longer need to be designed specific to their human agent, the metaphor of the arm extension that the industrial robot suggests is not enough? This research is focusing on mobile robotic fabrication strategies specific to filament materials. Deploying smaller robots for manipulation of lightweight thread-like materials allows building significantly larger structures. Multiple task-specific machines developed in this research are designed to carry, manipulate, anchor and pass filament materials in an on-site architectural environment of interior space. This paper presents the current state of the catalogue of robot species developed in this research as well as the experiments and demonstrators performed to evaluate them. Ultimately this research aims to create a larger toolbox of hardware and software tools and methods for custom single-task fabrication and construction robots.

4 Working Groups

4.1 Algorithmic Reproducibility

Thomas Auzinger (IST Austria – Klosterneuburg, AT)

Joint work of Workshop participants

There has been a discussion on how to best publish research results that depend heavily on algorithmic components; this is often the case in computational research. Academic participants strongly favor the release of open-source program code that implements a proposed method. Thus, a benchmark implementation is immediately available which covers all the details that are omitted from the method's description in the published article. However, such behavior needs to be incentivized by the community and needs to be part of recruitment requirements. Providing open-source implementation should be expected from both students and faculty when applying for positions in academia. Also, peer review would need to establish the submission of source code as a necessary criterion for reproducibility. Otherwise, the additional effort of supplying usable code disadvantages the person that does it in comparison with peers that only focus on publications. It was also emphasized that the existence of published source code should be mentioned when presenting the associated project.

In contrast, several companies prefer detailed description of the method in supplementals. This preference originates from legal issues: incorporating source code that is similar to published variants could create of vulnerability during IP-related litigation. For the same reason, some copyright- and trademark-focused companies do not open-source their code at all.

5 Open Problems

5.1 Fabrication Reproducibility

Thomas Auzinger (IST Austria – Klosterneuburg, AT)

Joint work of Workshop participants

An issue that was voiced by large number of participants is the reproducibility of fabrication result on different 3D printing devices. Even 3D printers of the *same* model can behave very differently at different sites. Several concerns and proposals were voiced:

Calibration models: It might be useful to develop a repository of 3D models that can serve as a 3D printer benchmark. This should be a collaborative effort of the community and should highlight various aspects of a 3D printers capabilities – both from geometric and material aspects. This would allow for a standardized way of cross-device and/or cross-location comparisons.

Metrology: Having these models fabricated on one's 3D printer, it is still necessary to reliably measure the result in order to judge the printer capabilities and quality. For this, both an open source hardware solution was advocated for and a centralized measurement service. Any of these two allows a quantitative analysis of the fabrication device and would also constitute a community consensus on the topic. However, it was acknowledges that open source hardware would require much more effort and it is unclear how many research groups would invest the time to build such a device. Sharing such measurements with the community will provide an overview of fabrication tolerances and device reliability. Moreover, it would pressure printer manufacturer to address fabrication biases of their devices. In any case, the simple collection of 3D models will not be useful and a full-feedback pipeline as it exists in 2D printing is strongly desired. This would permit the recalibration of a device based on a set of standardized models.

Manufacturers: However, several participants reported that 3D printer manufacturers are not overly cooperative in opening their APIs to allow fine-grained calibration methods. They often rely on business models that rely on lock-in and the necessity of service contracts. Thus, it is unclear if (and to what extent) such community efforts would influence their corporate strategy.

Legal issues: Several participants strongly favored a GPL-based license for such a community project to ensure that all users of this data are forced to contribute to it. However, enforceability is unclear and it might prevent companies as well as printer manufacturers from using it at all.

References: NIST is currently preparing a benchmark test for 3D printers but focus on the low-level material properties in a first step. The talk by Carolyn C. Seepersad presented an implementation of various parts of the aforementioned issues. On a public webpage, example prints on many machines are available for comparison; this highlights the device uncertainties across different printers. At the same time, various open-source 3D printers are packaged with calibration patterns. The soft robotics community can also serve as an example for such efforts, as it routinely shares fabrication recipes.

It was also mentioned that such a rich collection of data might be useful for machine learning efforts (e.g., for automatizes device calibration or design).

5.2 Geometry Representation Guidelines

Thomas Auzinger (IST Austria – Klosterneuburg, AT)

Joint work of Workshop participants

Several participants pointed out the missing consensus on what the best geometry representations for different tasks are. For various parts of the manufacturing pipeline, different representations (e.g., meshes or voxels) are used, and they exact trade-offs between different approaches are not always known. As examples from the workshop, the talks by Alec Jacobson and Lin Sebastian Kayser highlighted advantages and disadvantages of both representations.

Especially from industry, it would be highly appreciated if standardized benchmarks on algorithms could be provided to make a more informed decision. This is especially relevant for fully-fledged design and production pipelines, where many computational and manual tasks are usually chained together. In such a context, it cannot be expected that a certain geometry representation proves superior for every subtask. At the same time, conversions between representations – in order to adapt them to the various subtasks – often causes information loss. Thus, an overall recommendation is usually hard to identify and would require future research.

Participants

- Marc Alexa
TU Berlin, DE
- Thomas Auzinger
IST Austria – Klosterneuburg, AT
- Vahid Babaei
MPI für Informatik – Saarbrücken, DE
- Moritz Bächer
Disney Research – Zürich, CH
- Amit Haim Bermano
Princeton University, US
- Sabine Demey
Materialise HQ – Leuven, BE
- Laura Devendorf
University of Colorado – Boulder, US
- Georges Fadel
Clemson University – Clemson, US
- Madeline Gannon
Atonaton – Pittsburgh, US
- Serena Graziosi
Polytechnic University of Milan, IT
- Jessica K. Hodgins
Carnegie Mellon University – Pittsburgh, US
- Scott Hudson
Carnegie Mellon University – Pittsburgh, US
- Alec Jacobson
University of Toronto, CA
- Lin Sebastian Kayser
Hyperganic Technologies AG – München, DE
- Leif Kobbelt
RWTH Aachen, DE
- Manfred Lau
City University – Hong Kong, HK
- Sylvain Lefebvre
LORIA & INRIA – Nancy, FR
- David I. W. Levin
University of Toronto, CA
- Lin Lu
Shandong University – Qingdao, CN
- Jonas Martinez-Bayona
INRIA Nancy – Grand Est, FR
- Sara McMains
University of California – Berkeley, US
- Radomir Mech
Adobe Inc. – San José, US
- Ankur Mehta
University of California at Los Angeles, US
- Shuhei Miyashita
University of York, GB
- Caitlin Mueller
MIT – Cambridge, US
- Julian Panetta
EPFL – Lausanne, CH
- Mark Pauly
EPFL – Lausanne, CH
- Nadya Peek
University of Washington – Seattle, US
- Nico Pietroni
University of Technology – Sydney, AU
- Thijs Roumen
Hasso-Plattner-Institut – Potsdam, DE
- Ryan Schmidt
Gradientspace – Toronto, CA
- Carolyn C. Seepersad
University of Texas at Austin, US
- Kristina Shea
ETH Zürich, CH
- Melina Skouras
INRIA – Grenoble, FR
- Bernhard Thomaszewski
Université de Montréal, CA
- Nobuyuki Umetani
AUTODESK Research – Toronto, CA
- Philipp Urban
Fraunhofer IGD – Darmstadt, DE
- Charlie Wang
The Chinese University of Hong Kong, HK
- Tim Weyrich
University College London, GB
- Christopher Bryant Williams
VPI – Blackburg, US
- Maria Yablonina
Universität Stuttgart, DE

Report from Dagstuhl Seminar 18441

Data Physicalization

Edited by
Jason Alexander[1], Petra Isenberg[2], Yvonne Jansen[3], Bernice E. Rogowitz[4], and Andrew Vande Moere[5]

1 **Lancaster University, GB** j.alexander@lancaster.ac.uk
2 **Inria, FR** petra.isenberg@inria.fr
3 **CNRS & Sorbonne University, FR** jansen@isir.upmc.fr
4 **Visual Perspectives & Columbia University, US** bernice.e.rogowitz@gmail.com
5 **KU Leuven, BE** andrew.vandemoere@kuleuven.be

Abstract

Data physicalization involves representing numbers and relationships using physical, tangible displays. These displays provide tactile, as well as visual metaphors for expressing and experiencing data, and can unlock new analytical insights and emotional responses. This Dagstuhl seminar brought together a diverse group of researchers and practitioners to explore the benefits and challenges of physicalization – computer scientists trained in visualization, virtual reality and human-computer interaction; architects of virtual and augmented systems; perceptual and cognitive scientists; and artists and designers. Through interactive discussions and demonstrations, we explored physicalization, as a set of methodologies for representing data, for engaging audiences, and for artistic expression.

Seminar October 28–November 2, 2018 – http://www.dagstuhl.de/18441
2012 ACM Subject Classification Human-centered computing → Collaborative and social computing, Human-centered computing → Interaction design, Human-centered computing → Ubiquitous and mobile computing, Human-centered computing → Visualization
Keywords and phrases Art, Data Physicalization, Design, Human-Computer Interaction, Information Design, Visualization
Digital Object Identifier 10.4230/DagRep.8.10.127

1 Executive Summary

Jason Alexander (Lancaster University, GB)
Petra Isenberg (INRIA, FR)
Yvonne Jansen (CNRS-Sorbonne University, FR)
Bernice E. Rogowitz (Visual Perspectives & Columbia University, US)
Andrew Vande Moere (KU Leuven, BE)

This executive summary gives a brief overview of the goals and agenda for our seminar.

Introduction

Among the earliest man-made artifacts are physical representations of semantic concepts. These "physicalizations" externalize numerical and abstract concepts, providing physical metaphors that allow us to reason, remember, and communicate. Physicalizations are created

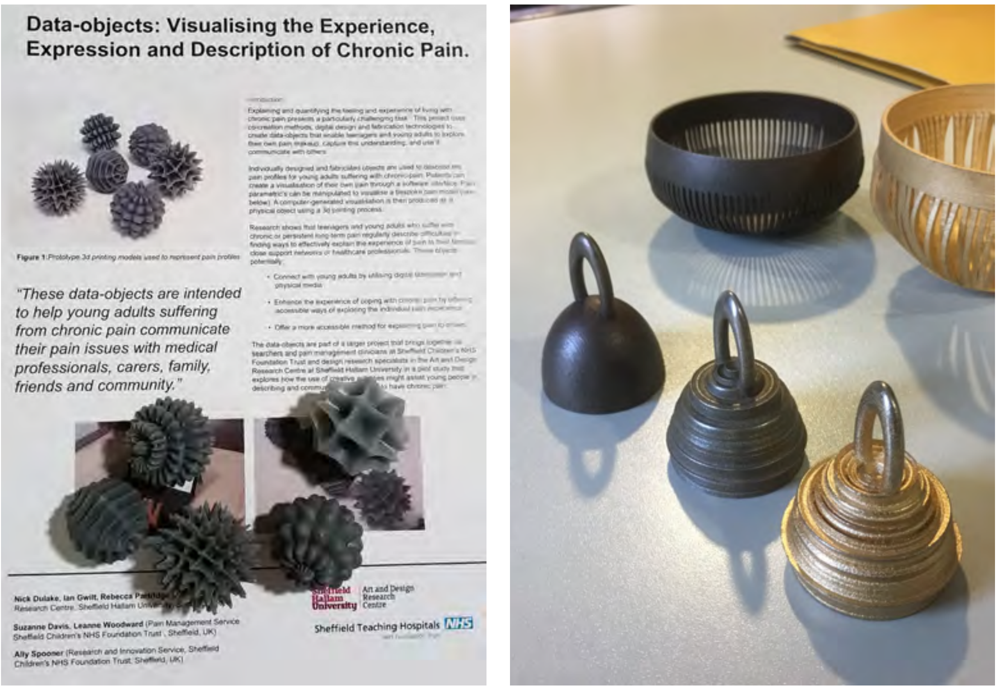

Figure 1 Examples of data physicalizations exhibited by our seminar participants. Left: Data objects for visualizing the experience, expression, and description of chronic pain. Right: Acoustic Sonifications whose sounds convey useful information about a dataset.

for many different purposes, or "intents." The sundial, for example, transforms shadows into a readable representation of time of day, the mercury thermometer transforms temperature into a displacement along a number line, and a scatterplot transforms the values of two variables into a form that allows the reader to interpret correlation. Our first data representations were based on natural objects, such as charcoal scraped onto walls or built from clay, or later, ink on paper. With the advent of computers, we've substituted physical representations with pixels on a computer screen. The resurgence of physicalization asks what we have lost in this transformation. Certainly, a computer-based visualization allows us to zoom an image, transform variables in real time, and to zoom through virtual computer-based world. However, these representations can sever the relationship to the natural world, depriving us of the touch, feel, and emotion that comes from interacting with real objects.

This Dagstuhl seminar brought together a diverse group of researchers and practitioners to explore the benefits and challenges of physicalization – computer scientists trained in visualization, virtual reality and human-computer interaction; architects of virtual and augmented systems; perceptual and cognitive scientists; and artists and designers. Through interactive discussions and demonstrations, we explored physicalization, as a set of methodologies for representing data, for engaging audiences, and for artistic expression. There were no formal paper presentations. Instead, the work was done through interactive discussions, hands-on workshops, and interactive demonstrations outside the lecture hall. Figure 1 shows two examples of data physicalizations exhibited by our participants.

Figure 2 A collaborative affinity diagramming session on Tuesday, to organize the large amount of thematic research questions into cross-disciplinary themes of research interest.

The Week at a Glance

Monday. After explaining various organization matters and formalities, the first day began with the physical construction of custom name badges. The hands-on session offered the first occasion to get to know each other, also by comparing the name badges themselves, then and during the whole duration of the seminar. After this session, each of the 40 participants gave a two-minute presentation of their achievements and interests in the field of physicalization. Some described their artistic creations based on data, some showed how mapping data onto physical dimensions enhanced data analysis. Others showed how scientists and artists collaborated to explain biological and physical principles through physicalization, or how principles of perception and cognition could be used to guide how data are effectively mapped onto visual dimensions. The electricity in the air was palpable; everyone realized that we were on the threshold of a new discipline. This energy drove the first deliverable for this seminar: Dagstuhl monograph, and perhaps a book, highlighting extended versions of these 40 contributions.

To allow a more focused approach towards yet unexplored research topics, the first activity focused on defining "pillars" of physicalization, crystallizing the learning and background of the different intellectual communities identified in our proposal. In break-out sessions, each group ideated around and then synthesized the most fundamental papers, examples, principles and challenges for their pillar, relative to physicalization. The Perception and Cognition Group focused on sensory processing, especially touch perception and embodied cognition, which deals with the way we learn about the world through our motor interactions with it. The Evaluation group shared their experiences in measuring human responses, and how methods might be extended to physicalization. The Design group explored artistic and design approaches to data physicalization. The Applications group identified existing and future application areas. The technology group surveyed the range of materials and devices for physicalization. In the reports-back, the group was encouraged to explore challenges and limitations for each of these core areas, to set the stage for the cross-disciplinary discussions beginning on Tuesday.

Tuesday. At the start the second day, each participant generated three questions for each of the five pillar areas. A massive exercise was undertaken wherein all these questions were organized and grouped into emergent categories. Spontaneous discussions started around the meaning and validity of the themes that emerged, as these would align the next round of thematic discussions (Figure 2). The "Design Patterns" group focused on identifying general templates for characterizing data physicalization. The "Emotion" group explored unique affordances that touching data enable. The Vis vs. Phys Group delved into the ways vision and touch were different, and what unique advantages that might enable. The "Critical" group explored the range of ethical and critical matters that could be related to physicalization practice in particular.

Wednesday. Wednesday morning was devoted to hands-on workshops where practitioners engaged small groups in interactive activities. Samuel Huron's group explored a set of physicalization examples, to identify common principles; Daniel Keefe demonstrated a virtual reality exploration of the human heart; Till Nagel, Laura Perovich, and Dietmar Offenhuber led a group into the forest to collect natural objects which they used to create physicalizations; Robert Friska provided surprising problem solving examples that drove a discussion about physical reasoning; Barbara Tversky showed examples of how physical gestures we make contribute to problem solving; Daniel Schneider led a workshop on computational embroidery, where the data are represented in yarn color and texture. By the time we piled into the bus for Trier, the large and diverse group of 40 had transformed into a dynamic community of scholars. Listening into the conversations, you could hear artists explaining how sculpture is taught in Art School, perceptual psychologists describing touch perception; and engineers revealing expertise on control systems for autonomous micro-robots.

Thursday. Thursday was the major work day. Building on the explorations of Tuesday, we divided into groups and worked on a variety of topics. One group, for example, focused on how to teach visualization and how to use data physicalization to teach other areas. Another group worked on a white paper on emotion and physicalization. Another worked on ideas for using physicalization for environmentally-situated projects. Another group focused on categorizing the critical considerations that emerged on Tuesday.

Friday. The final day aimed into transforming the progress of the past days into concrete contributions. For example, plans for a physicalization contest were discussed, ideas for the book of 40 contributions were pushed further, as well as on a special issue on physicalization for the Journal of Perceptual Imaging. Another group discussed the topic of scientific funding, among which the experience of past H2020 research proposals, and the potential of future research proposal initiatives that relate to physicalization.

Figure 3 Exhibition tables outside our seminar room.

Hands-on Demonstrations

The area outside the main conference room was populated with demonstrations provided by the participants. For example, Andrew Vande Moere brought a wireless, networked display that allowed participants to vote about particular topics that were shown; Volker Schweinsfurth, Daniel Schneider, and many others brought along and exhibited 3-D printed, hand-crafted, or even embroidered physicalizations. Bernice Rogowitz ran participants through several experiments around the topics of touch perception and touch/vision interactions. Section 4 details several of the hands-on experiences seminar participants and organizers could engage in; while Figure 3 shows the diversity of the physical artifacts that participants exhibited.

2 Table of Contents

Figure 4 The seminar included two very interesting lectures for participants. Left: Using Gesture to Communicate Semantics (Barbara Tversky); Right: Strong Spatial Cognition (Christian Freksa)

3 Overview of Talks

The seminar featured two talks on topics related to perception and cognition. Figure 4 shows Barbara Tversky and Christian Freksa lecturing. Details of the talks follow.

3.1 Visual Communication: Gesture and Diagrams

Barbara Tversky (Stanford University & Columbia Teachers College)

Like diagrams, sketches, and other graphics, gestures communicate more directly than words [1, 2, 3, 4]. Gestures can convey an overall spatial structure for thought, for examples, a relationships between two ideas or juxtaposing two sets of ideas or indicating a continuum of ideas, a dimension, such as events ordered in time or sports teams or movies in order of value or the workings of a complex system. Our work has shown that people learning static or dynamic spatial or conceptual relations spontaneously gesture when they study, and that doing so improves memory and inference. Similarly, people learning from explanations learn better when teachers gesture, especially dynamic information. The same holds true for diagrams and other graphics, but gestures are especially effective for dynamic information.

References

1 Barbara Tversky. Visualizing thought. *Topics in Cognitive Science*, 3(3):499–535, July 2011.
2 Barbara Tversky. The cognitive design of tools of thought. *Review of Philosophy and Psychology*, 6(1):99–116, March 2015.
3 Barbara Tversky. *Mind in Motion: How Action shapes thought.* Basic Books, 2019.
4 Barbara Tversky and Angela Kessell. Thinking in action. *Pragmatics & Cognition*, 22(2):206–223, 2014.

3.2 Strong Spatial Cognition

Christian Freksa (Universität Bremen, DE)

This workshop introduced the Strong Spatial Cognition (SSC) paradigm for spatial problem solving. Strong Spatial Cognition (SSC) – Definition 'SSC' is a cognitive agent's capability of processing spatial information through perception, representation, mental processing, and action. In comparison with spatial reasoning SSC includes the participation of spatial structures and processes in physical environments, the perception of configurations and spatial change in these environments, as well as the agent's locomotion and manipulation of spatial configurations.

In other words, the spatio-temporal substrate of an agent's environment is an integral part of his/her/its cognitive engine [2, 1, 3]. It allows to make direct use of factual, structural, and procedural knowledge in the world (Norman). The strong spatial cognition paradigm employs affordance-based object-level problem solving to complement knowledge-level computation.

According to SSC, cognitive agents equipped with suitable perception, mental capabilities, and effectors constitute not merely a tool in the study of the mind (as cognitive simulation programs on a computer do); rather, appropriately configured agents really are cognitive agents, in the sense that they can interact with spatial environments in a meaningful way and have cognitive states.

References

1 Christian Freksa. Strong spatial cognition. In Sara Irina Fabrikant, Martin Raubal, Michela Bertolotto, Clare Davies, Scott Freundschuh, and Scott Bell, editors, *Spatial Information Theory*, pages 65–86, Cham, 2015. Springer International Publishing.
2 Christian Freksa, Thomas Barkowsky, Frank Dylla, Zoe Falomir, Ana-Maria Olteteanu, and Jasper van de Ven. Spatial problem solving and cognition. In Jeffrey M. Zacks and Holly A. Taylor, editors, *Representations in Mind and World*, pages 156–183. Routledge, New York, USA, 2018.
3 Christian Freksa, Thomas Barkowsky, Zoe Falomir, and Jasper van de Ven. Geometric problem solving with strings and pins. *Spatial Cognition & Computation*, 2018. To appear.

4 Hands-On Workshops and Demos

Many of the workshop participants had previously explored physicalizations and haptic methods, and offered to lead workshops, demonstrations, and discussions that allowed others to share their experiences. These workshops provided hands-on experience with physicalization from many different perspectives.

Figure 5 Participants fabricating their own bespoke data badge.

4.1 Data Badges

Andrew Vande Moere (KU Leuven)
Sinem Görücü (KU Leuven) (via videoconference)
Georgia Panagiotidou (KU Leuven) (via videoconference)

All participants were invited to construct a bespoke, individualized "data badge" for themselves to be worn during the seminar (Figure 5). Specifically for this seminar, two graduate students of Andrew Vande Moere designed and then constructed a modular physicalization system via rapid prototyping methods. The system was meant to represent the research domain-related aspects of a person as a lightweight and wearable physicalization. The two students presented the core underlying ideas via a short videoconference session, and explained the instructions for building them. Four different colorful shapes each conveyed a particular domain (e. g. interactive technologies and hardware platform; evaluation and methodologies; arts, design and applications; perception and cognition), whereas their relative size correlated to the weight or importance of that domain. On the top left, these shapes conveyed the academic background of the attendee, whereas the bottom right corner related to future research directions. The length of white and blue wire corresponded with the duration of travel to Dagstuhl and the number of physicalisation-related publications respectively.

All attendees actively participated in the data badge fabrication session. The session formed the ideal informal platform to get to know one another by helping out or commenting on each other's badge. All resulting name badges were characterized by a sense of individual creativity and personal craftsmanship, while conveying shared types of information with which attendees could commence or refocus joint discussions and conversations (Figure 6). As a result, many data badges were worn for the whole length of the seminar.

Figure 6 Different examples of the hands-on fabricated data badges, reflecting the research interests of each attendee, as well as a certain sense of originality and craftsmanship.

4.2 Environmental Art as Physicalization Technique

Till Nagel (Hochschule Mannheim)
Laura Perovich (MIT Media Lab)
Dietmar Offenhuber (Northeastern University)

In this hands-on workshop (Figure 7), we explored environmental art as a visualization technique and investigated how natural materials can be employed in meaningful ways. After a brief introductory outline of this workshop's scope by the organizers, participants were asked to create physicalizations outdoors. We walked up the forest path to Burg Dagstuhl, observed the environment, collected materials, identified opportunities for intervention, and finally, created physicalizations. The three participating groups presented their physicalizations in-situ and reflected on the creation process.

One group focused on exploiting the biological shapes and textures to combine and connect them in ways that reflected quantitative data. Another group made a parcours of different zones, with each zone generating a different kind of sound when one stepped on it, and each sound representing a different numerical value. The last group constructed constellations of situated materials such as leaves and drops of water to resemble graphical data depictions. One more out-of-the-box example included the differences in length-of-flight of plant seeds that resemble a wing. This exploratory workshop was successful in bringing together different perspectives towards material and context from different fields around a new and unfamiliar task. This informed a conversation about the preliminary design space of a potential new research area. As a next step, we are going to fully document the process, results, and discussion and refine the workshop setup.

Figure 7 Participants engaging in the creation of data physicalizations using natural materials.

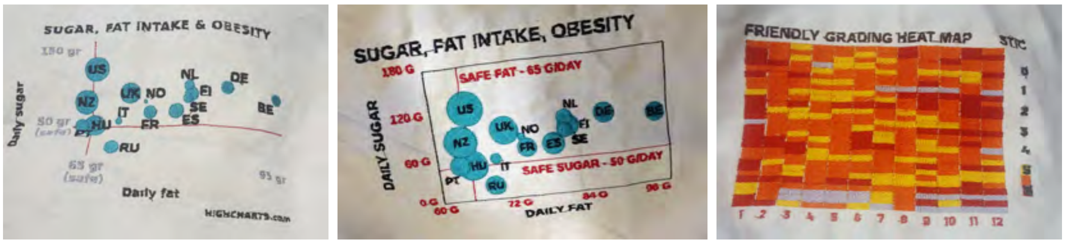

Figure 8 Examples of digital embroidery shared by Daniel Schneider.

4.3 Data Visualization with Machine Embroidery

Daniel K. Schneider (University of Geneva, CH)

This technical workshop introduced participants to principles of computerized embroidery (Figure 8), Participants discussed its potential for data physicalization. At the end of the workshop, each participant created a (very simple) data visualization that could be stitched with any low or high-end embroidery machine. The workshop agenda was as follows:

- Workflow(s) of computerized embroidery (10 min)
- Typology of embroidery stitch types (5min)
- Physical constraints (resolution, layers, tissue, size, etc.) (5 min)
- Hands-on: Using InkStitch (a free extension to the free Inscape program) (40 min)
- Hands-on: Generating, importing and adopting appropriate SVG graphics (30 min)

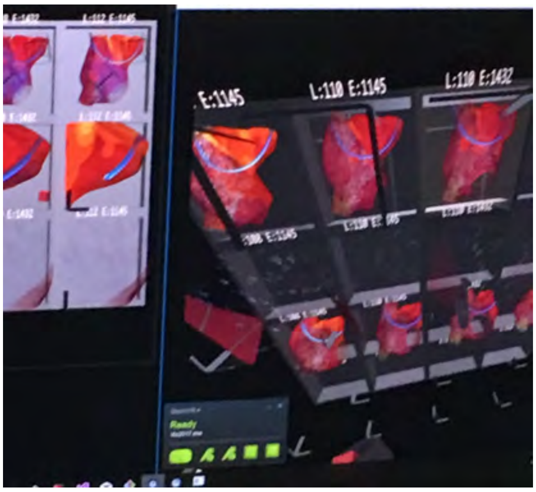

Figure 9 Multiple views of heart model parameters in Virtual Reality.

4.4 Physicalization for Scientific Visualization and Art+Science Collaborations

Daniel Keefe (University of Minnesota, US)

I presented a virtual reality demo of "Bento Box", a visualization method for arranging virtual spaces to make visual comparisons of multiple related datasets. In this case, we looked at a parameter study of 10 blood flows in the heart simulated on supercomputers using 10 different initial conditions. I also handed out a physical "pick up and touch" demo of 3D printouts of the same hearts. In the group we discussed how these two techniques might be combined to create more informative visualizations and intuitive user interfaces.

4.5 Screening "Visualizing Agriculture"

Leanne Elias (University of Lethbridge, CA)
Denton Frederickson (University of Lethbridge, CA)

The growing field of Data Visualization is situated at the fertile intersection of art and science. It explores the symbiotic coupling of these seemingly disparate disciplines, by finding connection among the distinct goals, methodologies, and contexts of artistic and scientific pursuit. Data, in its simplest sense, can be described as facts or statistics collected together for analysis. But things are rarely so simple, and in a moment of human development when data continues to escalate in quantity, complexity, and value, it may be timely to ask: What are the cultural, environmental, and political implications of data? How is data collected and used in contemporary society? Can art make sense of data? And how will data fundamentally change the character of art?

In September 2016, the Southern Alberta Art Gallery and the Data Physicalization Lab at the University of Lethbridge invited six artists to participate in a residency, documentary film, and exhibition that asked them to respond to agricultural data developed by Dr. Jamie Larson and Dr. Andre Laroche from the Lethbridge Research and Development Centre: Agriculture and Agri-food Canada. Two datasets were provided: The first is the result of an experiment on breeding cereal wheats and wheat grasses with the intention of developing a perennial wheat cultivar; the second relates to experimentation in genetic modification aimed at reducing wheat's susceptibility to the devastating pathogen known as stripe rust. The artists were invited to engage with the instruments, test subjects, contexts, methods, and people associated with the development of this data, which they would later consider while creating new work for this exhibition. The aim of this endeavour is to investigate the effect of intensive collaboration – how artists can use scientific process to guide their art, and how scientists can use artistic ways of knowing to approach their data in new ways.

The artwork that emerges from this investigation explores the potential that lies within scientific inquiry when strict standards for fact and method are allowed to be considered and probed through an expanded perspective. Through this inquiry, the work is allowed to affectively engage viewers by evoking feelings of wonder, curiosity, and consciousness about data and agricultural research, while creating a place for contemplation about the land and our inherent connection to it.

We are currently working on a documentary movie about our project. The Dagstuhl participant group received an early screening.

Summary Details on the project:
- Visualizing Agriculture: Southern Alberta Art Gallery
- February 17 to April 22, 2018
- Artists: Jackson 2Bears, Tori Foster, Mary-Anne McTrowe, Robyn Moody, Adrien Segal, Michelle Sylvestre
- Exhibition Design and Data Analysis: Christine Clark
- Curators / Organizers: Christina Cuthbertson, Leanne Elias and Denton Fredrickson (Data Physicalization Lab, University of Lethbridge)

4.6 The data physicalization card game

Samuel Huron (Telecom Paris Tech, FR), Anaelle Beignon (DSAA Villefontaine, FR)

This workshop was a hands-on activity to identify, describe and discover design patterns among the history of data physicalization. A design pattern is the description of the core of a solution to a recurrent problem. Design patterns are useful to describe, explain, understand, and generate novel solutions of a kind. In this activity we provide a card deck and a process to identify, describe, discover and structure data physicalization design patterns.

The data physicalization card deck (Figure 10) is composed of 280 cards, each one representing an artifact from the dataphyslist [1]. In a way this card deck is a physicalization of the data phys list. The data physicalization card game process (Figure 11) allows 10 to 100 researchers to engage in a hands-on open coding activity of these cards with several templates forms. The result of the activity is the identification of various design patterns and

Figure 10 From left to right: the data physicalization card deck; the design pattern templates; a summary of the activity process.

Figure 11 From right to left: picture of the activity; close up on a pattern; Example of the use of the cards by an others groups, here Sarah Hayes presenting.

meaningful categories that describe the data physicalization design space from the perspective of the participants.

The data physicalization card game is free, open source, and accessible online [2]. The data physicalization card game has already been used during the IEEE VIS 2018 workshop "Toward a Design Language for Data Physicalization" [3]. During the seminar participant were using it for all sort of activities. This workshop is part of a workshop series about data physicalization including let's get physical [5], vizkit [4], and other [6]. As a next step, I am going to document the results of the workshop and the use of the cards.

References

1. Pierre Dragicevic and Yvonne Jansen. List of physical visualizations. www.dataphys.org/list, 2012. Last accessed Jan 2019.
2. Samuel Huron and Anaelle Beignon. Data physical visualizations card game. data-physicalisation.github.io/data-phys-cards.html, 2018. Last accessed Nov 2018.
3. Jason Alexander Samuel Huron Sheelagh Carpendale Eva Hornecker Trevor Hogan, Uta Hinrichs, editor. *Toward a Design Language for Data Physicalization, Workshop at IEEE Visualization (VIS)*, 2018.
4. Samuel Huron, Sheelagh Carpendale, Jeremy Boy, and Jean-Daniel Fekete. Using viskit: A manual for running a constructive visualization workshop. In *Pedagogy of Data Visualization Workshop at IEEE Visualization (VIS)*, 2016.
5. Samuel Huron, Pauline Gourlet, Uta Hinrichs, Trevor Hogan, and Yvonne Jansen. Let's get physical: Promoting data physicalization in workshop formats. In *Proceedings of the Conference on Designing Interactive Systems (DIS)*, pages 1409–1422, New York, USA, 2017. ACM.
6. Trevor Hogan, Uta Hinrichs, Yvonne Jansen, Samuel Huron, Pauline Gourlet, Eva Hornecker, and Bettina Nissen. Pedagogy & physicalization: Designing learning activities around physical data representations. In *Proceedings of the Conference Companion on Designing Interactive Systems (DIS)*, pages 345–347, New York, NY, USA, 2017. ACM.

Figure 12 Volker Schweinsfurth presenting one of his data physicalizations.

4.7 Haptic Perception Experiments

Bernice Rogowitz (Visual Perspectives & Columbia University – New York, US)

One application of data physicalization is to communicate meaning in data. In visualization, Jacques Bertin explored how different visual dimensions could be used to effectively convey insight into category membership and magnitude. I presented a series of hands-on experiments that explored touch as a vehicle for communicating qualitative and quantitative mapping. The two-touch threshold experiment showed the participants that we can discriminate very fine differences with the fingertips, but may need up to 10x the separation on our forearms. Another experiment highlighted differences in texture discrimination using vision (very good), tactile discrimination without moving the finger (not so good) and tactile discrimination with movement (excellent). A third experiment demonstrated how tactile information can conflict with visual information. Two bottles of the exact same weight were presented. If the two bottles were placed on the palm, and judged by haptics alone, they seemed equal. If they were grasped without looking, the larger bottle was judged to be lighter. That is, extra weight had to be added to the large bottle for them to feel equal. When the bottles were grasped with visual as well as haptic cues, the larger bottle needed 50% more weight in order to be judged equal. This striking effect demonstrates that even a simple judgment of magnitude can be processed differently by the two senses, which has deep implications for data physicalizations that are both seen and felt.

4.8 Data Sculpture Exhibition

Volker Schweisfurth (Melies Art – Düsseldorf, DE)

During the breaks I laid out a collection of my (hand- to A4- sized) 3D-printed color datasculptures (based mostly on World Bank country data and Forbes, see Figure 12); they had been photographed by different participants and mostly covered topics from global and economic contexts; such as the size of 130 economies (GIP/ppp; $) vs. country risk for investments, the evolution of age pyramids of various countries, country risk profiles over the years, business figures of multinational companies, etc. As participants in the discussion of these physical objects, we emphasized the fact that, according to recent research (including my reference to Hutmacher/Kuhbandner; University of Regensburg), the brain stores information about the tactile sense longer than expected ("Detailed, durable, long-term memory representations are stored as a natural product of haptic perception"). In addition, I have pointed out that in my experience, the CAD files created for 3D printing can be used elegantly in two additional ways without media discontinuities: The creation of animations (.mov) and as input for mixed media (AR/VR) environments. In this manner, the virtualization of economic and global phenomena can effectively complement the physicalization of data- for teaching and industrial use, especially also for planning purposes. Regarding a patent, which I hold in the field of active data physicalization, I am unfortunately currently behind schedule in prototyping, mainly because the component market (small OLED-/AMOLED) displays is characterized by the strong requirements and demand on the part of VR glasses manufacturers.

4.9 Dagstuhl Data Physicalization

Pauline Gourlet (University of Paris VIII, FR)
Till Nagel (Hochschule Mannheim, DE)
Aurélien Tabard (Université de Lyon, FR)

As one of our workshop activities we got participants engaged in creating a data physicalization on their personal relationship to this Dagstuhl seminar. In Figure 13 you can see participants organized by how many Dagstuhl seminars they have attended (including this one) as well as their travel time to the seminar.

Figure 13 Participants joining in a data physicalization on their relationship to the seminar.

5 Working Groups

5.1 Data Physicalization Toward Sustainability and Peace

Yuri Engelhardt (University of Twente – Enschede, NL)
Brygg Ullmer (Clemson University, US)
Christian Freksa (Universität Bremen, DE)
Pauline Gourlet (University of Paris VIII, FR)
Samuel Huron (Telecom ParisTech, FR)

In this working group we explored the potentials of using data physicalizations to promote sustainability and peace. We discussed related initiatives that the members of our working group were affiliated or familiar with such as the UN sustainable development goals, the PRAM knowledge sharing network, the SDG DataViz Camp on Visualizing Inequalities, the Geospatial Information Section of the United Nations' project on "Making the World a Better Place with Maps", or the UNGlobalPulse's initiative on how data science and analytics can contribute to sustainable development. We also discussed the possibility to create a data physicalization challenge on the topic of sustainability and peace.

5.2 Physicalization vs. Visualization

David Kirsh (University of California – San Diego, US)

Our goal was to appreciate the relative strengths of the two approaches – in part by understanding how they differ in a principled way – and in part by a collection of looks at physicalization examples. There were four topics we opened.

1. **Embodiment:** how does having a body and using it in space, or using our hands, eyes in head, etc – how do these things affect how we frame problems, shape cognition and bias or facilitate how we think about things. Sometimes the body helps, sometimes it hinders or makes difficult certain transformations. Gestures help many things. We have volumetric understanding, knowing where we are etc
2. **Expressivity:** every representational system can be assessed by its expressivity – its capacity to encode / represent a certain dimensionality and form of data. We talked about what this means and in particular how there are so many dimensions in which data can be encoded in physical things. There are haptic dimensions – squeeze, texture,, there is shape – how expressive is that!!!, there is size, color, position, time.
3. **Readability:** the body plays a role in how we read off information. Essential that we note the interactive strategies available with a phyz and human body. There is no haptic reading without a haptic strategy to get it. These interactive strategies are often more complex than in viz. Embodiment also means that our bodies play a role in what we detect and read off.
4. **Concept formation:** phyz are often advertised as great for engendering new concepts or enabling insight into new sorts of relationships. We want to explore this much more.

5.3 The Role of Emotion and Engagement in Data Physicalization

Yun Wang (Microsoft Research – Beijing, CN)
Tim Dwyer (Monash University – Caulfield, AU)
Daniel Keefe (University of Minnesota – Minneapolis, US)
Petra Isenberg (Inria, Saclay, FR)
Roberta Klatzky (Carnegie Mellon University, US)
Eva Hornecker (Bauhaus-Universität Weimar, DE)
Jörn Hurtienne (Universität Würzburg, DE)
Leanne Elias (University of Lethbridge, CA)
Adrien Segal (California College of the Arts, US)
Steven Barras (Sonification – Ainslie, AU)

Visualisation, as a field of academic research, has for a long time sought rigour in terms of metrics and methodologies for the evaluation of data visualisation techniques. Typically, effectiveness or value of a visualisation is measured in terms of the efficiency of the technique in terms of user speed and comprehension in completing (relatively) basic data comprehension tasks. However, this pragmatic focus on readability and usability is missing consideration of some qualities of visualisation that are more difficult to define but, arguably, also very important. For example, a visualisation may not necessarily be the most efficient representation of a given data set, for a given task, but it may achieve greater user engagement through evoking an emotion or simply enjoyment in the user. By contrast, recent discussions about the possibilities of and proposals for data physicalisations have focussed on the importance of emotional engagement with the embodiment of the data. But what do we mean by emotional engagement, and why is it valuable? In this group, we exploreed these questions with consideration from findings in cognitive and perceptual psychology, art, design, marketing, activity theory, and user experience.

5.4 Sound in Data Physicalisation

Stephen Barrass (Sonification – Ainslie, AU)

In this working group we discussed about the potential application of sound and sonification in data physicalisation. Sound is a natural consequence of the manual manipulation of physical objects that conveys material properties, forces, modes of interaction and events over time. Film sound is used to convey emotions, cultural references, and to direct attention. Product sound is used for branding, guidance, engagement and other functional and aesthetic purposes. This leads to the idea that sound could be a way to augment physicalisations in similar ways. As an example an acoustic sonification is a dataset that is intentionally 3d printed in the form of a sounding object, such as a bell or a singing bowl. Since acoustic vibrations are a consequence of the 3D shape, the sound is effected by the dataset. Sounds can also be produced by embedding a digital synthesiser, such as the Mozzi sonification synth, inside an object. Sensors can be used to synthesis sounds in response to interactions such as shaking, squeezing or striking the object. For example, Zizi the Affectionate Couch is a piece of furniture that produces a range of different characteristic and emotional sounds such as whining, yipping and purring in response to human proximity, sitting and stroking its fur. Through further discussions in the workshop we heard about other examples, and thought about other ways that sounds could be applied in physicalisation. Many thanks to those who took part in an enjoyable and enlightening workshop.

6 Summary

The week went by very quickly, with impromptu evening discussions complementing the work during the day. New collaborations and new friendships emerged from the mixing of colleagues from so many different academic and artistic fields. Data physicalization is a new discipline; its practitioners are still defining its scope and limits. The Dagstuhl seminar will be seen as a critical moment in its crystallization, a formative time in the evolution of its research and aesthetic agenda. The seminar had several possible outcomes that were discussed in various working groups: developing a data physicalization challenge for the community, grant collaborations, the idea of a follow-on seminar, journal articles, the

Figure 14 After-hours collaborations.

possibility of a dedicated conference or symposium, and a special issue in a journal with a dedicated call-for-participation. These outcomes continue to be worked on by seminar participants and organizers. We thanks Dagstuhl for providing the necessary context and organization to make our seminar a very fruitful, engaging, and exciting event.

Participants

- Jason Alexander
Lancaster University, GB
- Stephen Barrass
Sonification – Ainslie, AU
- Stephen Brewster
University of Glasgow, GB
- Sheelagh Carpendale
University of Calgary, CA
- Tim Dwyer
Monash University – Caulfield, AU
- Leanne Elias
University of Lethbridge, CA
- Yuri Engelhardt
University of Twente – Enschede, NL
- Sean Follmer
Stanford University, US
- Denton Fredrickson
University of Lethbridge, CA
- Christian Freksa
Universität Bremen, DE
- Pauline Gourlet
University of Paris VIII, FR
- Ian Gwilt
University of South Australia, AU
- Sarah Hayes
Cork Institute of Technology, IE
- Uta Hinrichs
University of St Andrews, GB
- Trevor Hogan
Cork Institute of Technology, IE
- Eva Hornecker
Bauhaus-Universität Weimar, DE
- Samuel Huron
Telecom ParisTech, FR
- Jörn Hurtienne
Universität Würzburg, DE
- Petra Isenberg
Inria – Saclay, FR
- Daniel Keefe
University of Minnesota – Minneapolis, US
- David Kirsh
University of California – San Diego, US
- Roberta Klatzky
Carnegie Mellon University, US
- Till Nagel
Hochschule Mannheim, DE
- Bettina Nissen
University of Edinburgh, GB
- Lora Oehlberg
University of Calgary, CA
- Dietmar Offenhuber
Northeastern University – Boston, US
- Laura Perovich
MIT – Cambridge, US
- Bernice E. Rogowitz
Visual Perspectives & Columbia University – New York, US
- Daniel K. Schneider
University of Geneva, CH
- Volker Schweisfurth
Melies Art – Düsseldorf, DE
- Adrien Segal
California College of the Arts, US
- Aurélien Tabard
Université de Lyon, FR
- Barbara Tversky
Columbia Teachers College – New York, US & Stanford University, US
- Brygg Ullmer
Clemson University, US
- Andrew Vande Moere
KU Leuven, BE
- Karin von Ompteda
The Ontario College of Art and Design University, CA
- Yun Wang
Microsoft Research – Beijing, CN
- Wesley J. Willett
University of Calgary, CA

Report from Dagstuhl Seminar 18442

Visualization and Processing of Anisotropy in Imaging, Geometry, and Astronomy

Edited by

Andrea Fuster[1], Evren Özarslan[2], Thomas Schultz[3], and Eugene Zhang[4]

1 **TU Eindhoven, NL,** a.fuster@tue.nl
2 **Linköping University, SE,** evren.ozarslan@liu.se
3 **Universität Bonn, DE,** schultz@cs.uni-bonn.de
4 **Oregon State University – Corvallis, US,** zhange@engr.orst.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 18442, "Visualization and Processing of Anisotropy in Imaging, Geometry, and Astronomy", which was attended by 30 international researchers, both junior and senior. Directional preferences or anisotropies are encountered across many different disciplines and spatial scales. These disciplines share a need for modeling, processing, and visualizing anisotropic quantities, which poses interesting challenges to applied computer science. With the goal of identifying open problems, making practitioners aware of existing solutions, and discovering synergies between different applications in which anisotropy arises, this seminar brought together researchers working on different aspects of computer science with experts from neuroimaging and astronomy. This report gathers abstracts of the talks held by the participants, as well as an account of topics raised within the breakout sessions.

Seminar October 28–November 2, 2018 – http://www.dagstuhl.de/18442

2012 ACM Subject Classification Human-centered computing → Visualization, Applied computing → Life and medical sciencesApplied computing → Physical sciences and engineering, Computing methodologies → Modeling and simulation, Computing methodologies → Computer graphics, Mathematics of computing → Probability and statistics, Mathematics of computing → Continuous mathematics

Keywords and phrases Anisotropy, astronomy, diffusion-weighted imaging (DWI), geometry processing, tensor fields, topology, visualization, uncertainty, shape modeling, microstructure imaging, statistical analysis

Digital Object Identifier 10.4230/DagRep.8.10.148

Edited in cooperation with Marco Pizzolato

1 Executive Summary

Andrea Fuster (TU Eindhoven, NL)
Evren Özarslan (Linköpings Universitet, SE)
Thomas Schultz (University of Bonn, DE)
Eugene Zhang (Oregon State University, USA)

Topics and Motivation

Directional preferences or anisotropies are encountered across many different disciplines and spatial scales. For example, local anisotropies are imprinted in the cosmic microwave background radiation, the human brain contains elongated nerve fibers, etc. Such anisotropies lead to (physical) orientation-dependent quantities, i.e., quantities that take on different values when considered along different directions. Compared to scalar or vector-valued data, it is much more challenging to model, process, and visualize anisotropic quantities. Suitable mathematical models often involve tensors and other higher-order descriptors, and pose specific research challenges in several areas of computer science, such as visualization, image analysis, and geometry processing.

In order to explore synergies between different fields, to inform computer scientists about open application challenges, and domain experts about existing solutions, this seminar brought together researchers from three different disciplines:

- Medical **imaging,** where several modalities are now available to probe anisotropic behavior. In particular, Diffusion Weighted Magnetic Resonance Imaging (DW-MRI) is based on measuring anisotropic diffusion. It makes it possible to visualize and quantify microstructural information in fibrous tissues such as white-matter and muscles, and to infer larger-scale structures, such as fiber tracts in the human brain.
- Computer graphics and **geometry processing,** where tensor fields have a wide range of applications, such as quadrangular and hexahedral geometry remeshing, street network modeling, geometry synthesis, computational architecture, and path planning for environment scans.
- Cosmology and **astronomy,** where anisotropy plays a crucial role. For example, anisotropies in the cosmic microwave background (CMB) consist of small temperature fluctuations in the blackbody radiation left over from the Big Bang. Anisotropies are also found in the CMB in the form of a polarization tensor field, and they arise in the field of "cosmography", where efforts are united to map (parts of) the cosmos, e.g. the large-scale distribution of matter in the Universe or cosmic web.

Organization of the Seminar

This seminar was the seventh in a series of Dagstuhl seminars that was started in 2004, and has been devoted to the visualization and processing of tensor fields and higher-order descriptors. This particular instance of the seminar series focused on anisotropy in the fields of imaging, geometry, and astronomy.

To ensure a steady inflow of new ideas and challenges, we put an emphasis on inviting researchers who previously did not have the opportunity to attend one of the meetings in this series. This was true for almost half the attendees in the final list of participants.

The seminar itself started with a round of introductions, in which all participants presented their area of work within 100 seconds with help of a single slide. This helped to create a basis for discussion early on during the week, and was particularly useful since participants came from different scientific communities, backgrounds, and countries.

A substantial part of the week was devoted to presentations by 29 participants, who spent 20 minutes each on presenting recent advances, ongoing work, or open challenges, followed by ten minutes of discussion in the plenary, as well as in-depth discussions in the breaks and over lunch. Abstracts of the presentations are collected in this report. On Wednesday we held the traditional social event which was joined by almost all participants, and offered additional welcome opportunities for interaction.

A total of six breakout sessions were organized in the afternoons of Monday and Tuesday. Moderators summarized the respective discussion in the plenary on Thursday afternoon. The organizers came up with initial suggestions for session topics, which were refined further after discussion with the seminar participants. The session topics were as follows:

- Astronomy
- Time-varying anisotropy
- Theoretical tools
- Visualization
- Diffusion MRI
- Geometry

Notes were taken during all sessions, and the main points are summarized later in this report.

Outcomes

The participants all agreed that the meeting was inspiring and successful. It also stimulated new scientific collaborations and joint grant proposals. In addition we plan to publish another Springer book documenting the results of the meeting. Participants have pre-registered seventeen chapters already during the seminar, and we are in the process of collecting additional contributions both from participants and from researchers working on closely related topics who could not attend the meeting. We expect that the book will be ready for publication in 2020.

Acknowledgment

The organizers thank all the attendees for their contributions and extend special thanks to the moderators of the breakout sessions and the team of Schloss Dagstuhl for helping to make this seminar a success. As always, we enjoyed the warm atmosphere, which supports both formal presentations as well as informal exchanges of ideas.

2 Table of Contents

3 Overview of Talks

3.1 Quadrilateral and Hexahedral Mesh Generation

David Bommes (Universität Bern, CH)

Joint work of Heng Liu, Paul Zhang, Edward Chien, Justin Solomon, David Bommes
Main reference Heng Liu, Paul Zhang, Edward Chien, Justin Solomon, David Bommes: "Singularity-constrained octahedral fields for hexahedral meshing", ACM Trans. Graph., Vol. 37(4), pp. 93:1–93:17, 2018.
URL http://dx.doi.org/10.1145/3197517.3201344

Automatically generating quadrilateral and hexahedral meshes that smoothly align to freeform surfaces and offer a high amount of regularity and low distorted elements is a notoriously challenging task. Novel algorithms based on global optimization rely on the construction of integer-grid maps, which pull back a Cartesian grid of integer isolines from a 2D or 3D domain onto a structure aligned quadrilateral or hexahedral mesh. Such global optimization algorithms do not suffer from limitations known from local advancing front methods, as for instance a high rate of irregularity, and enable meshes comparable to manually designed ones by finding a good compromise between regularity and element distortion. The key for finding good solutions are 3D cross-fields that are employed to globally optimize the orientation and sizing of mesh elements. In my talk, I will give an overview of the state of the art and discuss the strengths and weaknesses of available algorithms, including open challenges for hexahedral meshing.

3.2 A Unifying Approach to the Processing of Polyspectral Images

Bernhard Burgeth (Universität des Saarlandes, DE) and Andreas Kleefeld (Jülich Supercomputing Centre, DE)

The processing of colour images is still an active field of research, especially if there are more than 3 or 4 channels. For instance, modern astronomical satellites provide polyspectral images with more than 100 channels. In this talk we propose generalizations of mathematical operations, such as linear combinations, multiplication, maximum, and minimum, that is, the fundamental building blocks in any image processing algorithm, to polyspectral data. The proposed technique takes advantage of the methodologies already available for tensor fields, and appears to be extendible to general vector data. This is joint work in progress with Andreas Kleefeld

3.3 Enforcing necessary constraints for common diffusion MRI models using sum-of-squares programming

Tom Dela Haije (University of Copenhagen, DK)

Diffusion-weighted magnetic resonance imaging (MRI) captures local micro-structural information by observing diffusing (water) molecules probing their surroundings at a microscopic scale. In order to analyze this type of data one can either estimate parameters that describe the diffusion itself, which provides a somewhat abstract but accurate description of the observed stochastics, or one can use a model of the ambient structure that re-expresses the observed diffusion in terms of more intuitive structural parameters. Both cases generally rely on optimization to reconstruct the descriptive parameters from diffusion-weighted images, and in this presentation I will introduce a specific set of basic constraints to improve such model reconstructions. These constraints are based on the non-negativity of the so-called ensemble average propagator or associated functions, but reformulated as the (relaxed) condition that these functions can be written as a sum of squared polynomials. For many commonly used models and basis expansions these constraints take the form of a positive-definiteness condition on a matrix that is linear in the model parameters, which can thus be enforced through the use of semidefinite programming or nonlinear optimization alternatives. In preliminary results I will show that the application of these constraints can be considered essential in many situations despite the associated computational costs.

3.4 Multiscale Visualization of 3D-Polarized Light Imaging Fields

Ali Can Demiralp (RWTH Aachen, DE)

3D-Polarized Light Imaging (3D-PLI) is a recent neuroimaging technique which is able to record the spatial orientation of nerve fibers within the micrometer range. This method utilizes the optical birefringence properties of myelin sheaths surrounding the axons, yielding a vector field corresponding to mean orientations of the nerve fibers. In this talk, I will focus on interactive visualization of 3D-PLI outputs including extraction of spherical harmonic representations through downscaling the vector field, which in turn may serve as guidance to the registration of the slices (a central problem in 3D-PLI data acquisition) as well as providing opportunities for voxel-wise comparison with diffusion Magnetic Resonance Imaging.

3.5 Geometry in uncertainty quantification

Aasa Feragen (University of Copenhagen, DK), Anton Mallasto, and Søren Hauberg

Main reference Anton Mallasto, Aasa Feragen: "Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes", in Proc. of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pp. 5665–5674, 2017.
URL http://papers.nips.cc/paper/7149-learning-from-uncertain-curves-the-2-wasserstein-metric-for-gaussian-processes

In this talk, discussed recent work with Anton Mallasto and Søren Hauberg on the role of geometry in quantification of uncertainty.

First, we discussed the situation where data is estimated from data, and therefore should be considered stochastic. This is a common scenario in medical imaging, where raw data is heavily preprocessed to create an image, which is next further processed in order to extract the data points of interest, for instance organ segmentation boundaries. While such boundaries are routinely considered deterministic data points, they should, in principle, be considered stochastic variables following a distribution. This is particularly relevant when data quality is low, or model fit is poor. We discuss the role of geometry in population analysis [2] when such data points (in our examples, white matter tracts [3]) are stochastic, and represented as Gaussian Processes.

Next, we moved to the situation where data is known to reside on a Riemannian manifold, where we discuss uncertainty quantification in submanifold learning [1]. Here, we introduced the Wrapped Gaussian Process Latent Variable Model, which learns a stochastic embedding of data into the Riemannian ambient manifold. We discussed the relation between sample size and prior knowledge, and how the manifold constraint becomes particularly important for quantifying uncertainty.

References

1 A. Mallasto, S. Hauberg, A. Feragen, Probabilistic Riemannian submanifold learning with wrapped Gaussian process latent variable models, AISTATS 2019 (to appear)
2 A. Mallasto, A. Feragen, Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes, NIPS 2017
3 S. Hauberg, M. Schober, M. Liptrot, P. Hennig, A. Feragen, A Random Riemannian Metric for Probabilistic Shortest-Path Tractography, MICCAI 2015

3.6 Control Triads for Geodesic Tractography in Diffusion Weighted Magnetic Resonance Imaging

Luc Florack (TU Eindhoven, NL)

Joint work of Rick Sengers, Stephan Meesters, Lars Smolders, Andrea Fuster, Luc Florack

We propose a novel (preliminary) Riemann-geometric method for geodesic tractography in diffusion tensor imaging (DTI) that lacks the rigidity of most existing ones. It does not presume a well-posed relation between DTI data evidence and tracts. Instead, it is endowed with control parameters for optimal adaptation to fiducial 'ground truth' tracts in (real or synthetic) data, in which data-extrinsic knowledge might be incorporated (e.g. provided by expert annotations or edits, statistically inferred via machine learning, etc.).

The method is not limited to DTI/Riemannian geometry, but can be (nontrivially) extended to generic diffusion MRI models in the context of Finsler geometry.

3.7 Local anisotropies in spacetime

Andrea Fuster (TU Eindhoven, NL)

Joint work of Andrea Fuster, Cornelia Pabst, Christian Pfeifer
Main reference Andrea Fuster, Cornelia Pabst, Christian Pfeifer: "Berwald spacetimes and very special relativity", Physical Review Vol. 98(8), pp. 084062, 2018.
URL https://doi.org/10.1103/PhysRevD.98.084062

In this talk I consider an scenario where local anisotropies may be present in spacetime, as suggested by Cohen and Glashow in the context of very special relativity (VSR). It turns out that the geometry underlying VSR is of a very particular type. But where is gravity?

3.8 Crease Enhancement Using MAFOD Filter

Shekoufeh Gorgi Zadeh (Universität Bonn, DE)

Main reference Shekoufeh Gorgi Zadeh, Stephan Didas, Maximilian W. M. Wintergerst, Thomas Schultz: "Multi-scale Anisotropic Fourth-Order Diffusion Improves Ridge and Valley Localization", Journal of Mathematical Imaging and Vision, Vol. 59(2), pp. 257–269, 2017.
URL http://dx.doi.org/10.1007/s10851-017-0729-1

Before extraction of the centerline (crease line) or the core-surface (crease surface) of anisotropic structures, such as vessels or brain's white matter skeleton in fractional anisotropy images, applying crease enhancing filters can improve the localization. For this purpose, we propose a multi-scale anisotropic fourth-order diffusion (MAFOD) filter that performs better than the other existing isotropic and anisotropic fourth-order filters. Plus we show that the 3D MAFOD filter can be steered to either enhance crease lines, or crease surfaces, or both at the same time.

References

1 Gorgi Zadeh, S., Didas, S., Wintergerst, M. W. M., Schultz, T.: Multi-scale Anisotropic Fourth-Order Diffusion Improves Ridge and Valley Localization. Journal of Mathematical Imaging and Vision 59(2):257–269 (2017)

3.9 Variance measures of diffusion tensors

Magnus Herberthson (Linköping University, SE)

Joint work of Magnus Herberthson, Evren Özarslan, Carl-Fredrik Westin

Calculating the variance of a family of diffusion tensors involves the formation of a fourth order tensor with the same symmetry properties as the elasticity tensor. This tensor has been studied w.r.t many properties: degrees of freedom, representations, invariants, decomposition, the equivalence problem et cetera. In this talk we discuss some of these properties, both in two and three dimensions.

3.10 Robust Extraction and Simplification of 2D Symmetric Tensor Field Topology

Ingrid Hotz (Linköping University, SE)

In this work, we propose a controlled simplification and smoothing strategy for symmetric 2D tensor fields that is based on the topological notion of robustness. Robustness measures the structural stability of the degenerate points with respect to variation of the underlying field. We consider an entire pipeline for the topological simplification of the tensor field by generating a hierarchical set of simplified fields based on varying the robustness values. Such a pipeline comprises of four steps: the stable exaction and classification of degenerate points, the computation and assignment of robustness values to the degenerate points, the construction of a simplification hierarchy, and finally the actual smoothing of the fields across multiple scales. We also discuss the challenges that arise from the discretization and interpolation of real world data.

3.11 Maximizing the information content of diffusion-relaxometry MRI data

Jana Hutter (King's College London, GB)

Novel acquisition techniques beyond traditional diffusion MRI include both variation in the shape of the diffusion encoding and the combination with relaxometry techniques (T1/T2). A versatile sequence, allowing variation in all dimensions are presented and possible alternatives in the sampling discussed.

3.12 Deep Learning-based tractogram filtering

Daniel Jörgens (KTH Royal Institute of Technology – Stockholm, SE)

Joint work of Daniel Jörgens, Philippe Poulin, Rodrigo Moreno, Pierre-Marc Jodoin, Maxime Descoteaux

Diffusion magnetic resonance imaging (dMRI) provides the opportunity to non-invasively obtain measures that relate to the human brain tissue microstructure. A common approach to analyze the global white matter architecture based on this modality is tractography. The basic idea of this technique is to create trajectories which are aligned with the local diffusion measurements at each point. The set of the trajectories (or streamlines) derived in this way is usually referred to as a tractogram.

Despite the existence of a variety of tractography methods, all these suffer from inherent limitations due to the relatively low spatial and angular resolution of dMRI as well as the generally ill-posed nature of the inverse problem they aim to solve. It has been shown that this often results in a large number of anatomically implausible streamlines.

Several approaches for 'cleaning' tractograms have been proposed whose aim is to classify streamlines as being anatomically plausible or implausible. In general, these are based on different features derived either from the streamlines or from samples of the diffusion data along them. Inspired by recent proposals which successfully employed machine learning in the step-wise creation of streamlines, we propose to use a machine learning-based, binary classifier for this task. In this talk, I will present our preliminary results in this context. In particular, we investigate different settings to train a binary classifier which is able to separate a tractogram into sets of plausible and implausible streamlines. In this work, we assess the performance of a) convolutional vs. recurrent neural network architecture, b) relying solely on dMRI as input features, and c) training on a single set of ground truth labels vs. training on a composition of several sets.

3.13 A path to process general matrix fields

Andreas Kleefeld (Jülich Supercomputing Centre, DE)

Joint work of Bernhard Burgeth, Andreas Kleefeld

A general framework is presented that allows for transferring data-processing algorithms for scalar to arbitrary matrix fields. That means to find analogues for fundamental operations such as linear combinations and maximum/minimum in this setting. Furthermore, we aim to process fields consisting of certain subsets such as the symmetric, skew-symmetric, Hermitian, and the general and orthogonal group. Some numerical examples concerning the special orthogonal group and the general linear group connected to Moebius transforms in hyperbolic geometry are presented.

3.14 Mapping scalar and tensor magnetic susceptibility of biological tissues

Chunlei Liu (University of California at Berkeley, US)

Magnetic susceptibility of biological tissues is intrinsically of tensor nature. In some cases, it can be approximated as a scalar quantity. Methods and challenges will be discussed for quantifying magnetic susceptibility based on MRI.

3.15 A Deep Learning Approach to Identifying Shock Locations in Turbulent Combustion Tensor Fields

Timothy Luciani (University of Illinois – Chicago, US)

Joint work of Mathew Monfort, Jonathan Komperda, Brian Ziebart, Farzad Mashayek, G. Elisabeta Marai
Main reference Matthew Monfort, Timothy Luciani, Jonathan Komperda, Brian Ziebart, Farzad Mashayek, G. Elisabeta Marai: "Deep learning features of interest from turbulent combustion tensor fields", Modeling, Analysis, and Visualization of Anisotropy, Springer, pp. 375-392, Feb., 2017.
URL https://doi.org/10.1007/978-3-319-61358-1_16

We introduce a deep learning approach for the identification of shock locations in large scale tensor field datasets. Such datasets are typically generated by turbulent combustion simulations. In this proof of concept approach, we use deep learning to learn mappings from strain tensors to Schlieren images which serve as labels. The use of neural networks allows for the Schlieren values to be approximated more efficiently than calculating the values from the density gradient. In addition, we show that this approach can be used to predict the Schlieren values for both two-dimensional and three-dimensional tensor fields, potentially allowing for anomaly detection in tensor flows. Results on two shock example datasets show that this approach can assist in the extraction of features from reacting flow tensor fields.

3.16 Tensor Approximation for Multidimensional and Multivariate Data

Renato Pajarola (Universität Zürich, CH)

Tensor decomposition methods and multilinear algebra are powerful emerging tools to cope with current trends in computer graphics, image processing and data visualization, in particular with respect to compact representation and processing of increasingly large-scale, high-dimensional and multivariate data sets. Initially proposed as an extension of the concept of matrix rank for 3 and more dimensions, tensor decomposition methods have found applications in a remarkably wide range of disciplines. We will briefly review the most successful tensor decomposition models and their applications in graphics and visualization, as well as describe specific benefits and features exploited for visual data compression, signal processing and interactive data manipulation. Furthermore, we will include a first outlook on porting these techniques to multivariate data such as vector and tensor fields.

3.17 Detecting and Describing Ultra Diffuse Galaxies and Faint Galaxy Streams

Reynier Peletier (University of Groningen, NL)

Recent astronomical surveys are so deep that many objects, which up to now were invisible, can be detected. One detects for example large numbers of Ultra Diffuse Galaxies, and also around larger galaxies often stellar streams are found that are remnants of previous galaxy-galaxy interactions. I am interested in developing methods that detect these faint features, but also to describe them in an objective way, so that they can be compared with galaxy formation simulations to study how galaxies formed. This work is done in collaboration with astronomers and computer scientists across Europe.

3.18 Exploiting the Signal's Spherical Mean to Calculate the Minimal Anisotropic Kernel

Marco Pizzolato (EPFL – Lausanne, CH)

Main reference Marco Pizzolato, Demian Wassermann, Rachid Deriche, Jean-Philippe Thiran, Rutger Fick: "Orientation-Dispersed Apparent Axon Diameter via Multi-Stage Spherical Mean Optimization," in CDMRI, 2018.
URL https://infoscience.epfl.ch/record/257227

The spherical mean of the diffusion MRI signal, process also known as "powder averaging", has revealed powerful for estimating the "microscopic" kernel, irrespective of its alignment in the 3d space. While tensorial kernels have been used due to the availability of explicit solutions for their spherical mean, through approximation it is possible to calculate generic micro-kernels. Moreover, the dimensionality reduction in the parameter space granted by powder averaging can be exploited to improve stability, speed, and limit degeneracy during parameter estimation of functions that convolve such kernels with 3d spatial distributions. Finally, leveraging our control over MRI acquisition parameters we may infer on further explorable properties of these kernels, such as their dependence on curvature.

3.19 Tensor Visualization Using Fiber Surfaces of Invariant Space

Gerik Scheuermann (Universität Leipzig, DE)

Symmetric second order tensor fields appear in medicine, astronomy, geometry, as well as mechanics. If one is interested in the invariants, one is left with a three dimensional space. For a typical application in medicine or mechanics, and sometimes in geometry or astronomy, the domain is also three-dimensional. Therefore, a symmetric second-order tensor field over a three-dimensional domain defines a mapping from its domain into three-dimensional space. This allows for interactive exploration. For this purpose, we define the preimage of a surface in the invariant space as fiber surface and demonstrate an algorithm to compute these surfaces. So far, the algorithm assumes a tetrahedral grid with linear interpolation of the invariants to keep things simple. In the following, we indicate how this construction may allow for a formulation of a topology of tensor invariant map which complements the classical tensor field topology in a strict sense.

3.20 Explicit and Implicit Prior Knowledge in fODF Estimation: The Best of Both Worlds

Thomas Schultz (Universität Bonn, DE)

Joint work of Ankele, Michael; Groeschel, Samuel; Lim, Lek-Heng; Patel, Kanil; Schultz, Thomas
Main reference Kanil Patel, Samuel Groeschel, Thomas Schultz: "Better Fiber ODFs From Suboptimal Data With Autoencoder Based Regularization", in Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI), 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III LNCS, vol. 11072, pp. 55–62, Springer, 2018.
URL https://doi.org/10.1007/978-3-030-00931-1_7

Estimating fiber orientation distribution functions (fODFs) from diffusion MRI is an ill-conditioned problem, and especially challenging when dealing with sub-optimal data that has been acquired in a clinical setting. We improve results by investing two types of prior knowledge: First, a positive definiteness constraint enforces that fODFs will not contain any negative contributions. Second, we regularize the estimation from sub-optimal data based on learning the distribution of plausible fODFs from high-quality data. We combine both strategies in a way that preserves their respective advantages, and eliminates the need for frequent re-training.

References

1 Ankele, M., Lim, L. H., Groeschel, S., Schultz, T.: Versatile, robust, and efficient tractography with constrained higher-order tensor fODFs. Int'l J. of Computer Assisted Radiology and Surgery 12(8), 1257–1270 (2017)

3.21 Increasing the dimensionality and size of the MRI parameter space for microstructure imaging: cure or curse?

Chantal Tax (Cardiff University, GB)

Diffusion MRI (dMRI) is the preferred tool for studying tissue microstructure in vivo, with the dMRI literature growing exponentially. However, current technology does not always allow comprehensive assessment, and fundamental issues challenge interpretation of results. We have reached a hiatus in advancing tissue characterization by dMRI alone, advocating the combination of multiple MRI-modalities. Instead of separately acquiring different contrasts, their joint information can be optimally exploited by simultaneously varying multiple experimental variables, an approach common in physical-chemistry NMR. The latest ultra-strong gradient MRI technology facilitates the translation of multidimensional MRI to human brain by greatly extending the size of the available measurement space. However, as the dimensionality and size of this space increase, efficient data acquisition and representation become more challenging. This talk aims to debate how multidimensional MRI and ultra-strong gradients could help our understanding of brain microstructure, and to identify open challenges in experiment design and analysis that hamper clinical translation of comprehensive and efficient microstructural MRI.

3.22 Topological Structures and Comparative Measures for Tensor Fields

Bei Wang (University of Utah – Salt Lake City, US)

This talk summarizes some speculative thinking regarding the topological structures and comparative measures for tensor fields. Suppose we are given a number of tensor fields defined on a common domain, we are interested in developing comparative measures that capture the relationship between two or more tensor fields. Similarly to the study of scalar fields (and to some extent, vector fields), it seems natural to construct a topological structure (e.g., a contour tree, a Reeb graph or a Morse-Smale complex) as a summary for each tensor field and use such a topological summary for comparison. It is also possible to take a Morse-theoretic approach to study the relationship between two or more tensor fields via an analog of the Jacobi set. In this talk, we will discuss some of these possibilities and challenges in establishing mathematical and algorithmic foundations for the study of multiple tensor fields.

3.23 Cosmic Microwave Background and the very early Universe

Dong-Gang Wang (Leiden University, NL)

In this talk I will give a brief introduction to the physics and observation of Cosmic Microwave Background (CMB) anisotropies. CMB is the relic radiation of the Big Bang, which was formed when the Universe was 380000 year old. After 13.7 billion years, today the CMB photons have a thermal black body spectrum at a temperature of 2.73K. More importantly, this temperature has tiny fluctuations across the whole sky, which is believed to be generated in the very early stages of the Universe. Therefore, through the measurement of the CMB anisotropies, we can extract a lot of information about the Big Bang. In particular, I will focus on the leading scenario of the primordial Universe — inflationary cosmology, and how much information CMB can tell us about it. Although current CMB data fits the standard predictions of inflation very well, there are some future observations, like primordial gravitational waves, non-Gaussianities, features and anomalies, which could reveal some new physics beyond our current understanding. I may elaborate on one to-be-confirmed observational anomaly, hemispherical power asymmetry, and show why its explanation poses a challenge for theoretical cosmologists.

3.24 Multidimensional diffusion MRI

Carl-Fredrik Westin (Harvard Medical School – Boston, US)

We recently proposed a new multidimensional diffusion MR framework for imaging and modeling of microstructure that we call q-space trajectory imaging (QTI). QTI framework enables microstructure modeling that is not possible with the traditional pulsed gradient encoding as introduced by Stejskal and Tanner. In this talk I will review this multidimensional diffusion MRI framework, and present novel extensions related to multidimensional diffusion MR and introduce relaxation to this model. The approach is inspired by multidimensional correlation spectroscopy, which improves differentiation of heterogeneous media in the field of NMR. I will discuss a six order tensor cumulant expansion of the diffusion MRI signal and present intuitive scalar invariants that can be derived from the moments of the expansion.

3.25 Geometry and data representation

Hsiang-Yun Wu (TU Wien, AT)

Joint work of Kazuho Watanabe, Yusuke Niibe, Shigeo Takahashi, Issei Fujishiro
Main reference Kazuho Watanabe, Hsiang-Yun Wu, Yusuke Niibe, Shigeo Takahashi, Issei Fujishiro: "Biclustering multivariate data for correlated subspace mining", in Proc. of the 2015 IEEE Pacific Visualization Symposium, PacificVis 2015, Hangzhou, China, April 14-17, 2015, pp. 287–294, IEEE Computer Society, 2015.
URL http://dx.doi.org/10.1109/PACIFICVIS.2015.7156389

Large datasets collected through advanced measurement techniques lead to increasing computation costs on analysis, formulation, and verification processes. In this presentation, we discuss the possibilities and the challenge of the analysis of high-dimensional data since human intuition about the geometry of high dimensions often diverges. Some applications, such as correlation analysis was introduced to initialize the discussion. The present techniques include Parallel Coordinate Plots, which has become a standard tool to analyze high-dimensional data and its extension Many-to-Many Parallel Coordinate Plots being developed to investigate all pairs of relations between dimension axes. Except directly plotting numerical data on screen space using visual encoding, including positions, colors, and shapes, alternatively, adding text or image labels enable another representation of additional information to the target of interests.

3.26 Tensor Field Design in Volumes

Eugene Zhang (Oregon State University – Corvallis, US)

Joint work of Jonathan Palacios, Lawrence Roy, Prashant Kumar, Chen-Yuan Hsu, Weikai Chen, Chongyang Ma, Li-Yi Wei
Main reference Jonathan Palacios, Lawrence Roy, Prashant Kumar, Chen-Yuan Hsu, Weikai Chen, Chongyang Ma, Li-Yi Wei, Eugene Zhang: "Tensor field design in volumes", ACM Trans. Graph., Vol. 36(6), pp. 188:1–188:15, 2017.
URL http://dx.doi.org/10.1145/3130800.3130844

The design of 3D tensor fields is important in several graphics applications such as procedural noise, solid texturing, and geometry synthesis. Different fields can lead to different visual effects. The topology of a tensor field, such as degenerate tensors, can cause artifacts in these applications. Existing 2D tensor field design systems cannot handle the topology of 3D tensor fields. We present, to our best knowledge, the first 3D tensor field design system. At the core of our system is the ability to specify and control the type, number, location, shape, and connectivity of degenerate tensors. To enable such capability, we have made a number of observations of tensor field topology that were previously unreported. We demonstrate applications of our method in volumetric synthesis of solid and geometry texture as well as anisotropic Gabor noise.

3.27 Topological Features in Stress Tensor Fields over Hex Mesh Distribution

Yue Zhang (Oregon State University – Corvallis, US)

Having a quality mesh is important to numerical modeling of physical phenomena. Much research exists to define the quality of the input mesh for a valid simulation and more work is needed to guide the distribution of the quality over the entire mesh. Our analysis here shows the effect of hex meshing with different distribution of quality over a numerical sample on the calculated stress tensor fields. Different topological features are shown for these fields while the distribution of the quality of the mesh varies.

3.28 Anisotropy Measure for Multi-component Shapes

Jovisa Zunic (Mathematical Institute, Serbian Academy of Sciences – Belgrad, RS)

Joint work of Paul L. Rosin, Jovisa Zunic
Main reference Jovisa D. Zunic, Paul L. Rosin, Vladimir Ilic: "Disconnectedness: A new moment invariant for multi-component shapes", Pattern Recognition, Vol. 78, pp. 91–102, 2018.
URL http://dx.doi.org/10.1016/j.patcog.2018.01.010
Main reference Mohamed Ben Haj Rhouma, Jovisa D. Zunic, Mohammed Chachan Younis: "Moment invariants for multi-component shapes with applications to leaf classification", Computers and Electronics in Agriculture, Vol. 142, pp. 326–337, 2017.
URL http://dx.doi.org/10.1016/j.compag.2017.08.029
Main reference Paul L. Rosin, Jovisa D. Zunic: "Orientation and anisotropy of multi-component shapes from boundary information", Pattern Recognition, Vol. 44(9), pp. 2147–2160, 2011.
URL http://dx.doi.org/10.1016/j.patcog.2011.02.018
Main reference Jovisa D. Zunic, Paul L. Rosin: "An Alternative Approach to Computing Shape Orientation with an Application to Compound Shapes", International Journal of Computer Vision, Vol. 81(2), pp. 138–154, 2009.
URL http://dx.doi.org/10.1007/s11263-008-0149-1

We will discuss a recent concept of multi-component shapes, which is applicable to image processing and image analysis tasks. The domain of multi-component shapes is very diverse and includes shapes that correspond to a group of objects that act together (e.g. a fish shoal), natural components of a segmented object (e.g. cells in embryonic tissues), a set of shapes corresponding to the same object appearing at different times (e.g. human gait in an image sequence), and many more.

Multi-component shapes have their specific characteristic (i.e. shape descriptors). Such shape characteristics need to be evaluated numerically, for an easier computer supported processing and analysis. One of them is the anisotropy measure, introduced as a quantity that should evaluate how much the shape components are ordered consistently.

3.29 Neurite morphology and the orientationally-averaged diffusion MR signal

Evren Özarslan (Linköping University, SE)

Joint work of Evren Özarslan, Cem Yolcu, Magnus Herberthson, Hans Knutsson, Carl-Fredrik Westin
Main reference Evren Özarslan, Cem Yolcu, Magnus Herberthson, Hans Knutsson, Carl-Fredrik Westin: "Influence of the size and curvedness of neural projections on the orientationally averaged diffusion MR signal," Front. Phys., vol. 6, 17, 2018.
URL https://doi.org/10.3389/fphy.2018.00017

Anisotropy of the diffusion-attenuated magnetic resonance (MR) signal has been widely utilized in applications such as tractography for mapping the connections between distant regions of the brain. Getting rid of this anisotropy information by orientationally-averaging the signal provides valuable insight into tissue's microscopic anisotropy and structure. We present recent advances on the influence of the geometry of neural projections on the observed diffusion MR signal.

4 Panel discussions

4.1 Astronomy

Andrea Fuster (Eindhoven University of Technology, NL), Reynier Peletier (University of Groningen, NL)

At this session we discussed how techniques in computer science might help processing and analyzing astronomical data. One of the problems treated was how to classify the different components of the large scale structure of the Universe or cosmic web (clusters, filaments, voids and walls) by using data from astronomical surveys. The type of data depends on whether the survey is shallow or deep, i.e. a large part of the sky is covered with a sparse distribution of points such as the Sloan Digital Sky Survey (SDSS), or a small portion of the sky with a higher density of data points. Some suggestions were to classify filaments by computing ridges (one-dimensional curves traced by high-density regions) or using tensor-based measures, and to compare results obtained with each approach. Related topics were the use of interpolation and denoising methods for data from shallow surveys.

Astronomical data acquired at smaller spatial scales, corresponding to objects such as stars, galaxies and galaxy clusters, was also discussed. One of the goals of such studies is to (automatically) classify stars vs. galaxies. One of the problems is that faint galaxies are not detected by the current software tools. One of the suggestions was to attempt a classification based on spectral decomposition. Another problem is how to get rid of artifacts related to optical reflection in images of groups of galaxies. Algorithms for image deconvolution could be used for this purpose.

4.2 Time-varying anisotropy

Jana Hutter (King's College London, GB)

In this session, relevant applications of time-dependency in tensors, and the inherent link with anisotropy, were discussed. Potential interests in time-dependency arises for instance in structural mechanics, where the stress tensor is time-dependent even though this information is relevant in specific cases. Deformation is brought as example of a continuous process for which time could be relevant, as well as of periodic loading assessment that is often performed for risks prevention in artificial structures or to test the resistance of machines. Another important field of application of time-varying tensors is computational fluid dynamics, where they are commonly used to identify local density, pressure, and velocity related to shock waves. Medicine is probably the field where data has an inherent time-dependent nature, as it is the case of pathology progression monitoring. However, in this field, the concept of time-dependency is not only explicit (like in the case of the data) but can be implicitly adopted in some processing techniques where time appears as an "artificial" mechanism involved in the generation of heat maps distributions in imaging, or in denoising, or even in the acquired data such as the concept of *diffusion time* in diffusion MRI. The last field of interest discussed in this session was geometrical meshing. A representative example of a time-varying application is computer animation, where a good deformation of meshes during time is the at the core of a successful result.

4.3 Theoretical tools

Magnus Herberthson (Linköping University, SE)

This session aimed at discussing a subset of common mathematical tools involved in the topics discussed in the seminar, with the purpose of becoming more familiar with them, finding cross-field overlaps, and posing fundamental questions to be solved.

The first tool discussed was the Bloch-Torrey equation, which describes the effects of water proton diffusion on the NMR/MRI relaxation phenomena. Particular emphasis was placed on the problem of determining the relaxation times T_1 (longitudinal) and T_2 (transverse) in the presence of complications arising from the presence of diffusion phenomena.

While the direct problem is well-understood and described by the previously mentioned equation, the inverse problem of estimating the parameters has revealed difficulties, with no satisfactory results in the community, up to now. The discussion aimed at tackling the estimation issues from different angles, while suggesting that it is quite likely that this ill-posed inverse problem falls within known classes of "parameter estimation of PDEs." If so, it was noted that there is already a machinery that could be employed, which would involve Tikhonov-regularization, Landweber-iteration, and Frechet-derivative and its adjoint.

Higher order tensors were the focus of the second main point of discussion. Many topics were touched upon, mainly from a theoretical perspective. These included tensor versus matrix formalisms (lower order tensor alternatives), the relation with continuum mechanics, various decomposition approaches, relevant invariants, and the issue of higher order SVD (singular value decomposition).

4.4 Visualization

Ingrid Hotz (Linköping University, SE)

The breakout session on visualization was originally guided by four topics: tensor topology applications in science, (3D) astronomy applications, multi-parameter diffusion imaging, and uncertainty visualization. A first round collecting topics of interest resulted at first in a very broad list, however over the course of the meeting common interests became clearer and are summarized below. One of the key outcomes was stating the need of close interdisciplinary projects to develop solutions that are targeted to the specific needs, ideally developed in a participatory manner between users and visualization experts.

1) There is a large gap between practitioners and the visualization community. The gap between the communities becomes already clear when considering common state-of-the-art visualization methods. While every community seems to have its own domain-specific solutions covering some basic visualization methods, more advanced methods are, to a large extent, unknown to the domain scientists. This seems to be especially true for explorative and analytic methods. For instance, tensor fields are still mostly visualized using derived scalar fields which might be suboptimal.

As one reason for this gap, an insufficient technology transfer and financial support of software development have been identified. There is a large difference between freely available tools and research software, the latter often being not easy-to-use and remaining at a proof-of-concept implementation stage. There exist several state-of-the-art articles, which are, however, targeted mostly towards visualization experts. Available tools include OpenSpace for astronomy data, MRtrix for diffusion MRI, as well as ParaView, VTK, VisIt, and AMIRA (not free) as generic tools for simulation data. Several of them have Python interfaces. It was identified that tensor field visualization tools are few. Some general challenges for visualization:

- specifying what is really needed, often it is unclear what should be visualized (this is often the main part of the problem);
- visualization should simplify the task without hiding information;
- standardization is needed;
- tutorials and learning packages for specific applications are needed in the from of "Cookbooks."

2) Special topics of interest for many participants working with tensor data. Despite the different application areas there were also some issues of joined interest, here listed.

- *Quantification and visualization of uncertainties* of the entire data analysis pipeline. Challenges lie in the integration of global and local measures as well in visualization of uncertainty in its entire complexity, e.g. covariances. This requires the visualization of tensors of higher order. Furthermore, the challenge of finding a good representation of uncertainty without overwhelming the user was discussed.
- *Visualization of higher order or higher dimensional tensors.* In many applications, the interest goes beyond simple second order tensors in three dimensions, e.g., gradients of tensors or tensors in Finsler geometry.
- *Visualization of ensembles allowing comparison* of tensors or groups of tensors. Related applications often involve high dimensional parameter spaces which require explorative visualization at different abstraction levels.
- *Multi-scale visualization.* Topology could be an interesting tool for abstracting data. However, there is a communication gap here as well, since it is not clear what the purpose is and when it should be used.
- *Simple and easy-to-understand visualizations*, such as cartoon-like automatic illustrations.

3) Visualization for different purposes. Some concepts are listed.

- *Visualization for non-experts of generic mathematical concepts* is required, and with different levels of abstraction: for instance, the visualization for clinicians must be different than that for DTI researchers.
- *Visualization for scientific communication*, e.g., 3D rendering on a dome was brought as an example.
- *Specific needs* should be taken into account in the design of visualization techniques: visualization for astronomy was brought as an example, due to its high complexity and the amount of information that needs to be displayed.

4.5 Diffusion MRI

Chantal Tax (Cardiff University, GB), Tom Dela Haije (University of Copenhagen, DK)

This session was mainly oriented towards bringing up topics of interest in Diffusion MRI that could be tackled by participants of the seminars given their knowledge of tools, and their different perspectives on the subject. A relevant macro-topic was tractography, i.e., the generation of a neuroanatomically accurate representation of the three-dimensional geometry of brain white-matter tracts that connect different cerebral regions. The resulting tractogram is typically inferred by processing tensor fields or related variants. The relevance of tractography spans from neurosurgery to the discovery of brain connectivity with serious implications for investigations in neuroscience.

The main challenge with tractography is its ill-posed nature. Every method leads to different results and there is no single ground-truth solution. This problem is deleterious for all applications, especially for delicate matters involving surgery. Because of this, tractography has had limited applicability, and those practitioners who do not rely on tractography need to be well-trained in order to correctly interpret the results.

Despite the abundance of methods present in the literature, none achieves satisfactory results, so that when ranking methods based on synthetic numerical or physical phantoms, the criterion adopted is often the compromise between false positives and false negatives. Indeed, the typical outcome of tractography entails the detection of tracts that do not exist and the lack of others that are known to exist. This is still an open question. Additional or new type of data could help improve the sensitivity and specificity of tractography together with the clever injection of prior knowledge.

As a case example, geodesic tractography was described. This allows for reconstructing the most likely paths of molecules by processing tensor fields according to a Riemannian/Finslerian metric. One of the issues, however, is the presence of many isotropic areas in the tensor field, rendering the main direction uncertain. Another important point is that the tractogram, while having to be geometrically accurate, has to properly describe the connections between different brain regions, and these connections often do not coincide with the most likely paths.

To add to the challenges, it was also mentioned that at the moment there is no real means of knowing the precise brain anatomy and connectivity in order to assess the accuracy of the tractograms. Connectivity, for instance, is not constant across lifetime, and some areas can take over the function of others, such as in the case of post-stroke recovery. These concerns lead to the problem of tractography validation. As a remark on this, it was highlighted that even a technique like polarized light imaging of *ex vivo* samples suffers from registration problems between slices.

To get around this, it was proposed to avoid giving a strict anatomical interpretation to the tractograms, and to develop methods that compute and visualize the possible solutions in a clever manner. Probabilistic approaches can play a great role in this endeavor, provided that neurosurgeons, neuroscientists, and other practitioners are satisfied with such information, although it is well-known that different experts would likely have different needs and opinions. Along a similar line, another approach discussed was making the many parameters behind the generation of tractography explicit. In some sense, by changing one or more parameters in the tractography pipeline (regularization over the tensor field, tract propagation, stopping criteria, etc.), the outcome changes as well. It would be important to define signal-based optimal criteria to avoid basing such choices on empiricism. Moreover, tractography depends on several physical parameters related to the acquisition and, as of now, no attempt to incorporate such information has been done.

Another important point discussed in this session was assessing the role of visualization in tractography. For instance, the need of visualizing all sources of uncertainty in the tractogram while still guaranteeing its intelligibility emerged as a potential goal.

4.6 Geometry

Eugene Zhang (Oregon State University, USA)

The following conclusions and open research questions emerged from discussion in the breakout session on geometry processing:

- Meshes play an important role in shape representation, but now the field is wider with simulation being the main driving force behind high-quality meshes.
- Main challenges include how to define the quality of a mesh and how to generate meshes knowing the quality measure.
- For simulation, the quality measures can be the speed of convergence, numerical accuracy and stability. They will likely be application-dependent.
- The high-quality quad and hex meshes for shape representation are well-researched, with the ability to interact with the mesh generation process through design. However, this does not address the needs for simulations.
- Can we adapt the mesh shop idea from shape representation and 3D printing to simulation? Can we define the price tag for a mesh vs. its utility?
- How to generate anisotropic meshes, which should be better suited for applications involving anisotropy such as processing blood flow data?
- How to make open-source code for high-quality meshing available to practitioners?

5 Schedule

	Monday	Tuesday	Wednesday	Thursday	Friday
9:00	Introductions	Evren Özarslan	Dong-Gang Wang	Bernhard Burgeth	Aasa Feragen
		Luc Florack	Reynier Peletier	Andreas Kleefeld	Tom Dela Haije
		Daniel Jörgens	Andrea Fuster	Timothy Luciani	Jovisa Zunic
10:30	Coffee	Coffee	Coffee	Coffee	Coffee
11:00	Chunlei Liu	Chantal Tax	Bei Wang	Thomas Schultz	Wrap-up
	Ali Can Demiralp	Jana Hutter	Hsiang-Yun Wu	Shekoufeh Gorgi Zadeh	
12:15	Lunch	Lunch	Lunch	Lunch	Lunch
14:00	Yue Zhang	Marco Pizzolato	Social event	Ingrid Hotz	
	David Bommes	C.-F. Westin		Renato Pajarola	
	Eugene Zhang	Magnus Herberthson		Gerik Scheuermann	
15:30	Coffee	Coffee		Coffee	
16:00	Breakout	Breakout		Panel + book	
18:00	Dinner	Dinner		Dinner	

Figure 1 The meeting schedule.

Participants

- David Bommes
Universität Bern, CH
- Bernhard Burgeth
Universität des Saarlandes, DE
- Tom Dela Haije
University of Copenhagen, DK
- Ali Can Demiralp
RWTH Aachen, DE
- Aasa Feragen
University of Copenhagen, DK
- Luc Florack
TU Eindhoven, NL
- Andrea Fuster
TU Eindhoven, NL
- Shekoufeh Gorgi Zadeh
Universität Bonn, DE
- Hans Hagen
TU Kaiserslautern, DE
- Magnus Herberthson
Linköping University, SE
- Ingrid Hotz
Linköping University, SE
- Jana Hutter
King's College London, GB
- Daniel Jörgens
KTH Royal Institute of Technology – Stockholm, SE
- Andreas Kleefeld
Jülich Supercomputing Centre, DE
- Chunlei Liu
University of California at Berkeley, US
- Timothy Luciani
University of Illinois – Chicago, US
- Evren Özarslan
Linköping University, SE
- Renato Pajarola
Universität Zürich, CH
- Reynier Peletier
University of Groningen, NL
- Marco Pizzolato
EPFL – Lausanne, CH
- Gerik Scheuermann
Universität Leipzig, DE
- Thomas Schultz
Universität Bonn, DE
- Chantal Tax
Cardiff University, GB
- Bei Wang
University of Utah – Salt Lake City, US
- Donggang Wang
Leiden University, NL
- Carl-Fredrik Westin
Harvard Medical School – Boston, US
- Hsiang-Yun Wu
TU Wien, AT
- Eugene Zhang
Oregon State University – Corvallis, US
- Yue Zhang
Oregon State University – Corvallis, US
- Jovisa Zunic
Mathematical Institute SANU – Belgrad, RS