DAGSTUHL
REPORTS

Volume 9, Issue 11, November 2019

Biggest Failures in Security (Dagstuhl Seminar 19451)
Frederik Armknecht, Ingrid Verbauwhede, Melanie Volkamer, and Moti Yung 1

Machine Learning Meets Visualization to Make Artificial Intelligence Interpretable (Dagstuhl Seminar 19452)
Enrico Bertini, Peer-Timo Bremer, Daniela Oelke, and Jayaraman Thiagarajan . 24

Conversational Search(Dagstuhl Seminar 19461)
Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein 34

BOTse: Bots in Software Engineering (Dagstuhl Seminar 19471)
James D. Herbsleb, Carolyn Penstein Rosé, Alexander Serebrenik, Margaret-Anne Storey, and Thomas Zimmermann 84

Composing Model-Based Analysis Tools (Dagstuhl Seminar 19481)
Francisco Durán, Robert Heinrich, Diego Pérez-Palacín, Carolyn L. Talcott, and Steffen Zschaler 97

Diversity, Fairness, and Data-Driven Personalization in (News) Recommender System (Dagstuhl Seminar 19482)
Abraham Bernstein, Claes De Vreese, Natali Helberger, Wolfgang Schulz, and Katharina A. Zweig 117

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
March, 2020

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.9.11.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 19451

Biggest Failures in Security

Edited by

Frederik Armknecht[1], Ingrid Verbauwhede[2], Melanie Volkamer[3], and Moti Yung[4]

1 **Universität Mannheim, DE,** armknecht@uni-mannheim.de
2 **KU Leuven, BE,** ingrid.verbauwhede@esat.kuleuven.be
3 **KIT – Karlsruher Institut für Technologie, DE,** melanie.volkamer@kit.edu
4 **Columbia University – New York, US,** moti@cs.columbia.edu

Abstract

In the present era of ubiquitous digitalization, security is a concern for everyone. Despite enormous efforts, securing IT systems still remains an open challenge for community and industry. One of the main reasons is that the variety and complexity of IT systems keeps increasing, making it practically impossible for security experts to grasp the full system. A further problem is that security has become an interdisciplinary challenge. While interdisciplinary research does exist already, it is mostly restricted to collaborations between two individual disciplines and has been rather bottom-up by focusing on very specific problems.

The idea of the Dagstuhl Seminar was to go one step back and to follow a comprehensive top-down approach instead. The goal was to identify the "biggest failures" in security and to get a comprehensive understanding on their overall impact on security. To this end, the Dagstuhl Seminar was roughly divided into two parts. First, experienced experts from different disciplines gave overview talks on the main problems of their field. Based on these, overlapping topics but also common research interests among the participants have been identified. Afterwards, individual working groups have been formed to work on the identified questions.

Seminar November 3–8, 2019 – http://www.dagstuhl.de/19451
2012 ACM Subject Classification Security and privacy, Social and professional topics
Keywords and phrases Cryptography, Hardware, Security engineering, Software engineering, Usability, Human Computer interaction (HCI), Human and societal aspects of security and privacy, Usable security or human factors in security, Security evaluation and certification
Digital Object Identifier 10.4230/DagRep.9.11.1

1 Executive Summary

Frederik Armknecht
Ingrid Verbauwhede
Melanie Volkamer
Moti Yung

General Introduction

In the present era of ubiquitous digitalization, security is a concern for everyone. Consequently, it evolved as one of the most important fields in computer science. However, one may get the impression that the situation is hopeless. Nearly on a daily basis, reports of new security problems and cyberattacks are published. Thus, one has to admit that despite the huge

efforts continuously invested since many decades, securing IT systems remains an open challenge for community and industry.

One of the main reasons is that the variety and complexity of IT systems keeps increasing, making it practically impossible for security experts to grasp the full system. This results into the development of independent and isolated security solutions that at best can close some specific security holes. Summing up, security requires to solve an increasing number of inter- and intradisciplinary challenges while current approaches are not sufficiently effective. The aim of this seminar was to gain an interdisciplinary view on security and to identify new strategies for comprehensively securing IT systems.

Goals

The goals of the seminar was to address the following main challenges and to commonly discuss solution strategies:

Challenge 1: Interdisciplinarity The topic of security is getting more and more complex and already understanding the state-of-the-art within one discipline is highly challenging. This makes it practically impossible to understand the problems and constraints from other disciplines. Moreover, different disciplines often have their own methods and "culture". From our experience, working with colleagues from other disciplines requires at the beginning an enormous effort to understand each other. The complexity grows even further when more than two disciplines are involved.

Challenge 2: Variety of Problems In each discipline, a variety of problems do exist. Naturally, researchers have to single out specific problems that they work on instead of aiming for comprehensive solutions. The selection of problems usually depends on several factors, e.g., background of the researcher, topicality of the subject, etc. Most often, researchers aim for solving very specific problems rather than coming up with more comprehensive solutions. Moreover, the selection is driven by interdisciplinary factors.

For sure, interdisciplinary research does exist already. However, it is mostly restricted to address very few disciplines and has been rather bottom-up by focusing on very specific problems. Instead, the scope of the seminar was to aim for a *broad top-down approach*. To this end, the focus was on the following questions:

- What are the main recurring reasons within disciplines why security solutions fail, i.e., the biggest failures? (Top View)
- How do these failures impact solutions developed in other sub-disciplines? (Broad View)
- What are possible strategies to solve these problems?

Structure

The seminar was structured accordingly. Before the seminar, a survey was conducted where the participants have been asked, what they consider to be biggest failures in security. The list of participants was composed of experts from different, selected sub-fields who were encouraged to explain the main challenges in their field to the audience. Here, ample opportunities for discussions have been provided. That is, instead of having many different talks back-to-back, we had several overview talks from different fields within the first few days. Afterwards, the whole audience commonly identified three topics to be further investigated in separate working groups:

1. The process and role of certifications
2. The human factor in security
3. The education of the society in security

These subgroups met in parallel and worked on specific questions. The remaining days were composed of workgroup meetings and individual talks. At the end of the seminar, the workgroups reported to the whole audience their findings.

This report summarizes the finding of the survey (Section 3), the topics of the individual talks (Section 4), and also the findings of the individual workgroups (Section 5).

2 Table of Contents

Executive Summary
Frederik Armknecht, Ingrid Verbauwhede, Melanie Volkamer, and Moti Yung . . . 1

Survey Results . . . 6

Overview of Talks . . . 8

DDoS Still Challenging 20 Years Later
Sven Dietrich . . . 8

Research Directions for a Safer Europe
Fabio di Franco . . . 9

Attacker Models and Assumption Coverage
Felix Freiling, Frederik Armknecht . . . 10

Values in Computing – a Short Talk
Lucy Hunt . . . 11

DRM and Security – A Big Failure?
Stefan Katzenbeisser . . . 11

Failures in TLS Implementations
Olivier Levillain . . . 12

Human Involvement in Highly Automated Systems: Human System Integration in Security
Joachim Meyer . . . 12

The Biggest Failures to "Protect" You in the Internet
Vasily Mikhalev . . . 12

Relation of Business Models to Security (Failures)
Sebastian Pape . . . 13

Memory Corruption Vulnerability Exploitation and Mitigations
Michalis Polychronakis . . . 13

Trusted Computing: The Biggest Failure or Opportunity?
Ahmad-Reza Sadeghi . . . 13

Challenges of Regulating Security
Christoph Sorge . . . 14

Fantastic Embedded Security Failures and Where to Find Them.
Lennert Wouters . . . 14

Layers of Abstraction and Layers of Obstruction
Moti Yung . . . 15

Working groups . . . 15

Certification Working Group
Felix Freiling and Begül Bilgin . . . 15

Education Working Group
Lucy Hunt, Magnus Almgren, Hervé Debar, Fabio di Franco, Sven Dietrich, Daisuke Fujimoto, Youngwoo Kim, Gabriele Lenzini, Olivier Levillain, Lennert Wouters, and Moti Yung 19

Human Factors Working Group
Joachim Meyer, Robert Biddle, Sebastian Pape, Kazue Sako, Martina Angela Sasse, Stephan Somogyi, Borce Stojkovski, Ingrid Verbauwhede, and Yuval Yarom 22

Participants 23

3 Survey Results

In order to prepare and to kick-off the seminar, an online survey was distributed to all participants. It mainly contained two questions:

1. What is the one biggest failure in security? Please explain why you selected this one as the biggest one.
2. Which other failures in security should be considered?

The Survey was filled out by 17 participants (3 from industry and 14 from research institutions). Participants have on average 21 years of past experience in security (with min. 13 and max. 36 years).

The open-ended text answers were analysed by two researchers. The answers were clustered and six main and five smaller themes were identified. For the analyses, it was decided not to distinguish between answers of both categories as several participants provided more than one failure in their response to question 1 and some provided more than two failures in their answers to question 2. Though, in the following when we provide quotes, those in italic are those taken from answer to question 2.

In the following the identified main themes are introduced and quotes are provided:

Theme 1: Lack of Holistic Approach for Complex Systems

Several answers were related to various aspects of (not) ideal approaches taken throughout the development of systems which need to be secured against attacks. Example quotes are:

- ... without adequate consideration of the importance of holistic design ...
- ... [systems] are too complex to be well-understood ...
- ... boundaries of a system get more and more fuzzy ...
- ... mechanism provides a solution for a very dedicated security challenge, one can often not exclude the existence of ... other security holes ...
- ... involve multitude of disciplines ...
- ... across disciplinary boundaries ...
- ... quality of risk modeling ... as a whole is ... poor
- ... list of assumptions for the overall system are not clear
- Making tradeoffs that overfocus on providing security to undifferentiated large scale groups rather than numerically smaller demographics

Theme 2: Lack of Usability

Several participants mentioned human related aspects wrt. security mechanisms. Note, the number of answers assigned to this one was higher than for all the other themes. Example quotes are:

- ... Not designing security with the Human Factor in mind – solutions with too much workload, complexity. Users are being set up to fail, ...
- ... Implementing more ideal security features with complicated procedures rather than usability ...
- ... usability is another central issue ... security mechanisms should operate "invisible" ... mechanisms complicated or impacting usability negatively ...
- Overload of IT users, e.g. requesting to memorize > 10 passwords.
- ... failure of organizations to appreciate the interplay between usability and security, driving usability underground, and compromising security ...
- ... which leads to the ... question of usability of security mechanisms ...

- ... why is security sometimes at perceived as trading off usability ... ?
- ... Lack of empirical testing of effectiveness of security measures.
- Lack of user friendly identity management infrastructure.
- The lack of ... unobservable communications usable by normal citizens

Theme 3: Not Learning from Past Mistakes

Several participants provided answers indicating that the community does not learn from past failures. Example quotes are:

- ... how we do not seem to learn from our mistake ...
- ... never seeming to learn from old mistakes. ...
- .. Incapability or inconsequence to learn from failures sustainably ...
- ... We patch it and learn about it on one system ... but when there is a shift to something new, similar ... vulnerability pops up again ...
- ... but many mistakes by programmers are long known and could easily be prevented ...
- ... lack of education where a new generation is doing the old mistakes ...
- ... we continue doing things just because that's the way we've always done them ...

Theme 4: Decision Makers Not Taking (appropriate) Actions

Several participants mentioned various types of decision makers (related to law and politics) in the failures they see. Example quotes are:

- ... we have been slow to update laws to reflect our technology, and slow to appreciate the impact of technology on legal protections ...
- .. Governments take a hands-off approach, and let organizations scale up until it becomes difficult to change ...
- ... lack of attention by decision makers, until sth. major happens ...
- ... Companies are rarely rewarded for building reliable systems ...
- ... [accept] convenient and cheap solutions that lead to major ... problems later.
- ... it seems to widely accepted that companies have outsourced security updates to the users. Users need to spend time and sometimes money ... to fix shortcomings of the systems they are using.
- ... lack of regulations from the onset. Anyone can write, publish/sell an app – other sectors require a clear process ...

Theme 5: Lack of Appropriate Certification Concepts

Answers related to certification and standardisation were assigned to this theme. Example quotes are:

- ... lack of certification concepts for the security and privacy of products and services that scale to the needs of agile development and cloud delivery ...
- ... the question of suitable criteria for cloud based, agile software is not addressed at all in the discussions ...
- ... failure of standards bodies ... to make certificate infrastructure work properly ...
- Not understanding the degree of accuracy required, leading to high failure rates.

Theme 6: Lack of End User Protection

Several answers focused on the general inability of protection end user / consumers adequately. Example quotes are:

- ... lack of protection of consumers against malware ...
- ... lack of robust online identities ...
- The inability ... to provide consumers with a reasonable &reasonably ICT device for day-to-day tasks –the digital ... Golf to use a car-market analogy
- Protecting humans from bad decisions. Why are systems designed in a way that a user can damage the whole system just by opening a link or an attachment of a mail? ...
- "Solutions" which place the risk at the weak parties ...

Smaller but more specific themes

The following specific failures have been mentioned (note, only those provided by at least two participants are mentioned)

- Unsecure programming language (3)
- Phishing is still among the major causes of breaches (2)
- Passwords are still around (2)
- Issues related to machine learning (2)
- Web browsers becoming an execution environment for everything (2)

Overall, the result of this survey allowed us to make all seminar participants aware of the wide range and level of abstraction of failures one can think of. The result helped us also to group in working groups.

4 Overview of Talks

4.1 DDoS Still Challenging 20 Years Later

Sven Dietrich (City University of New York, US)

We provide an overview of the fundamental flaws that have contributed to allowing the distributed denial-of-service (DDoS) phenomenon [1, 2] to happen over the last 20 years. This includes design flaws for the Internet and its protocols, management decisions, and sometimes faulty defensive stances. We show that the imprecision of the DDoS problem itself contributed (and still contributes) to the difficulty in responding to it. Incremental fixes have only created good albeit partial solutions to subproblems of the DDoS phenomenon. Defense mechanisms have varied from attack source identification, volumetric attack detection, network puzzles, pushback from target-resident detection, and command-and-control detection, and graph-based analysis for botnets [6]. The migration of attack sources over the years from government or university owned computers, to broadband-connected home computer systems and most recently Internet-of-Things devices shows the active continuation of the DDoS phenomenon and our inability to completely suppress the problem [7]. Repeated calls for an overhaul of the Internet, allowing for improvement and better flexibility in addressing the DDoS problem, have been stalled over the years, even though some good starting points for next-generation network infrastructures do exist [4, 3], but many challenges remain to be solved.

References

1 Jelena Mirković, Sven Dietrich, David Dittrich, and Peter Reiher. *Internet Denial of Service: Attack and Defense Mechanisms*. Pearson, USA, 2004.
2 CERT. *Results of the Distributed Systems Intruder Tools Workshop*. Published December 7, 1999.
3 Marc C. Dacier, Sven Dietrich, Frank Kargl, Hartmut König. Dagstuhl Seminar 16361: Network Attack Detection and Defense: Security Challenges and Opportunities of Software-Defined Networking, September 2016.
4 Marc Dacier, Hartmut König, Radoslaw Cwalinski, Frank Kargl, Sven Dietrich. *Security Challenges and Opportunities of Software-Defined Networking*. IEEE Security & Privacy 15(2):96–100 (2017).
5 David Dittrich, Sven Dietrich. *P2P as botnet command and control: a deeper insight*, in Proceedings of the 3rd International Conference on Malicious and Unwanted Software (Malware), pp. 46–63, October 2008.
6 Baris Coskun, Sven Dietrich, Nasir Memon. *Friends of an enemy: identifying local members of peer-to-peer botnets using mutual contacts*. In Proceedings of the 26th Annual Computer Security Applications Conference (ACSAC), pp 131-140. Austin, TX, December 2010.
7 Sven Dietrich. *Cybersecurity and the Future*. IEEE Computer 50(4): 7 (2017)

4.2 Research Directions for a Safer Europe

Fabio di Franco (ENISA – Attica, GR)

Main reference Fabio Di Franco: "Analysis of the European R&D priorities in cybersecurity", ENISA, December 2018.
URL https://www.enisa.europa.eu/publications/analysis-of-the-european-r-d-priorities-in-cybersecurity/at_download/fullReport

Europe should become "a global leader in cybersecurity by 2025, in order to ensure the trust, confidence and protection of our citizens, consumers and enterprises online and to enable a free and law-governed internet", as stated at the Tallinn Digital Summit in September 2017. The focus of the report is to highlight and recommend how focussed R&D can address emerging challenges that might pose a severe risk to our society. A key element is the recognition that the world is moving digitally and fast. The speed of adoption of new technologies has a potentially huge benefit resulting in increasing productivity, but at the same time may also pose risks if the technology where used against the best interests of society. Social norms take dozens of years to develop and the digital transformation is creating an increasingly blurred distinction between the digital and the physical world. In the digital world, a small number of corporations, popularly referred to as Internet giants , are increasingly required to service the societies of the 21st century. However, this requires a barter between the user's data and the internet giants' services: the users allow the digital platforms to track their location, record their interests and monitor their online activities in return for a wide series of services demanded by the users. In almost all cases, there is no direct user interaction in the bartering system, or only to the extent that the user understands the meaning of data collection notices. orted):

References

1 Fabio Di Franco *Analysis of the European R&D priorities in cybersecurity*. ENISA, Dccember 2018

4.3 Attacker Models and Assumption Coverage

Felix Freiling (Friedrich-Alexander-Universität Erlangen-Nürnberg, DE), Frederik Armknecht (Universität Mannheim, DE)

In his seminal paper on "failure mode assumptions and assumption coverage" [1], David Powell defines several central concepts:

1. The notion of *failure mode assertions*, i.e., precise statements about the way in which certain components may fail in the time domain and the value domain.
2. The *failure mode implication graph*, i.e., a lattice induced by the combination of failure modes defining the partial order between different composed failure modes.
3. The notion of *assumption coverage*, i.e., the probability that the assertion defining the assumed behavior of a component proves to be true in practice conditioned by the fact that the component has failed [1, p. 391].

The goal of this discussion session was to reflect on the similarities and differences between safety and security regarding attacker assumptions and assumption coverage and to ask whether any related work and concepts exist. Safety was understood here as the area of fault-tolerance and dependability, whereas security was understood as the area of cryptography. The connection to the title of this Dagstuhl seminar was the fact, that one of the biggest failures in security appears to be the fact that we do not learn sufficiently from other areas.

Regarding the concept of attacker assumptions, our observation was that in safety attacker assumptions are usually fixed for a specific scenario and in this scenario often empirically measurable. Examples are failure rates of components or maximum frequency of bitflips on communication lines or in memory. The mechanism, with which a component attempts to tolerate these problems, has no influence on the assumption coverage.

In security, the attacker assumption is usually determined by a domain expert and must be regularly checked whether it is still correct. It can even change spontaneously. In circumstances where this is expected to happen, issues of *risk management* arise. Furthermore, security mechanisms can have an effect on attacker behavior:

- either a strong mechanism deters attackers and makes the system uninteresting compared to others,
- or a weak mechanism is circumvented easily with minimal effort.

In safety we have concepts like *graceful degradation* and *stabilization*. On the one hand, graceful degradation means that the level of violation of specification is proportional to the strength of failure behavior. On the other hand, stabilization refers to a temporary violation of a safety property if attacker assumption is violated, and a return to safety property when attacker assumption is satisfied.

In security, the attacker assumption is usually a worst-case attacker assumption. Intermediate levels of attackers are unusual. Also switching between different security mechanisms is unusual and it is unclear on what basis the switch should occur since many violations of confidentiality and integrity are undetectable.

In the discussion, people from security admitted that worst-case assumptions usually are prefered, but often also weaker assumptions are used, so the cryptography community does not really live up to this claim of always choosing worst-case assumptions.

It was also mentioned that *testing* has strong similarities to transient attacks that try to throw a single machine off the tracks, and that stabilization has similarities to the mechanisms used to tolerate denial-of-service attacks.

References

1 David Powell. Failure mode assumptions and assumption coverage. In *Digest of Papers: FTCS-22, The Twenty-Second Annual International Symposium on Fault-Tolerant Computing, Boston, Massachusetts, USA, July 8-10, 1992*, pages 386–395, 1992.

4.4 Values in Computing – a Short Talk

Lucy Hunt (Lancaster University, GB)

Joint work of Emily Winter, Stephen Forshaw, Lucy Hunt, Maria Angela Ferrario
Main reference Emily Winter, Stephen Forshaw, Lucy Hunt, Maria Angela Ferrario: "Towards a systematic study of values in SE: tools for industry and education", in Proc. of the 41st International Conference on Software Engineering: New Ideas and Emerging Results, ICSE (NIER) 2019, Montreal, QC, Canada, May 29-31, 2019, pp. 61–64, IEEE / ACM, 2019.
URL https://doi.org/10.1109/ICSE-NIER.2019.00024

Values in Computing (ViC) is about understanding how human values influence software production and transforming the way values are considered in software industry practices, policy making and education. With the increasing number of high impact technology breaches and failures, we need computing professionals equipped to understand what human values are and what social responsibility means. To this end, we need to help create more resilient, secure and less vulnerable software systems that are mindful of the wider ethical, social and human impact of what their technology does or could do. ViC has a body of research establishing a framework for the systematic investigation of human values in software production and a website to disseminate our work (www.valuesincomputing.org).

How can software (security) incident story-telling be used to improve SE industry and education practices?

4.5 DRM and Security – A Big Failure?

Stefan Katzenbeisser (Universität Passau, DE)

In the talk we discuss the evolution of Digital Rights Managenent techniques, which were proposed to secure online content. The key idea was to encrypt content and transmit the encryption key in a special license. The failure of DRM can be tracked down to technical issues (such as the absence of trusted hardware at that time), changes in the business model (such as the uprising of flatrate streaming media) and usability problems. Media security tried to fill this gap by marking distributed media invisibly. Still, the fundamental different nature of analog signals led to numerous problems (such as robustness issues and conflicts in dispute resolution). Nevertheless, the techniques developed in the area of media security nowadays play a significant role in the construction of covert channels.

4.6 Failures in TLS Implementations

Olivier Levillain (Télécom SudParis – Evry, FR)

In the recent years, we saw a lot of implementation bugs in SSL/TLS stacks, ranging from classical programming errors to parsing bugs, cryptographic issues and state machine flaws. In many cases, similar problems were found in different independent implementations. Maybe the root cause of the problem is not only the developers' lack of skills. On the contrary, it might be time to use better languages and better development methodologies, as well as to improve the specifications we produce. Regarding this last point, we discuss what TLS 1.3 can/will bring to improve the situation.

4.7 Human Involvement in Highly Automated Systems: Human System Integration in Security

Joachim Meyer (Tel Aviv University, IL)

The security of systems depends to a large extent on the actions of the humans who interact with the technology. Human actions can introduce threats, but they can also help to mitigate risks. Humans are often supported by automation that provides them with advice on decisions, guides their actions, blocks alternatives, and may automatically perform acts that are deemed necessary. The talk addresses the question of human-systems integration in the context of automation, presenting four different ways in which humans can be involved in systems (humans receive advice, humans supervise automation and intervene in certain cases, humans supervise automation and set parameters, and maintaining "meaningful human involvement" without specifying its nature). Quantitative models and empirical results for the different types of involvement are shown, and some implications for system design are discussed.

4.8 The Biggest Failures to "Protect" You in the Internet

Vasily Mikhalev (Universität Mannheim, DE)

Today, personal data is among the most important resources which is being collected by some governments and big organizations. This data can be used for many different purposes including targeted advertising and even targeted propaganda. The existing technologies which are based on the combination of data science methods together with better understanding of human brain allow for "hacking" human beings using their personal data collected from the internet and for manipulating people's emotions. In this talk, we discuss the "protection" measures that Russian government has implemented in order to increase the security of citizens and why most of these measures appeared to be the biggest failures.

4.9 Relation of Business Models to Security (Failures)

Sebastian Pape (Goethe-Universität Frankfurt am Main, DE)

URL https://pape.science/files/talks/1911_Pape_Dagstuhl.pdf

When looking at the usability of current systems, we can note systems often leave the users in (potentially) dangerous situations. In theory, it should not be possible to brick a system or get infected by malware when reading mails or working on office documents. Many of the features are used by a small number of users or not appropriate for the tool leaving users in a state with lots of rules what they should do (do not click on embedded links, do not open attachments, do not activate macros, ...). As a consequence, users are used to do 'strange things' for the sake of security. This can be exploited by dark patterns and companies make use of it by their business models. For example when companies blame hackers for outage or simply security failures, outsource consequences of bad security (e.g. malwertising, insecure IoT devices) and effort (correction of false positives, e.g. in malware detection), and obfuscate business goals with security (e.g. when asking for phone numbers for two factor authentication, but inadvertently used them for advertising).

The result of that is a downward spiral where users have the feeling that they need to do 'strange things' for the sake of security which can be exploited by companies to ask them to obey 'strange orders' pretending to improve the users security. Which again increases the users feeling that they need to do 'strange things' for security reasons.

4.10 Memory Corruption Vulnerability Exploitation and Mitigations

Michalis Polychronakis (Stony Brook University, US)

In this talk I will present our work on generating self-specializing software that i) reduces its attack surface by removing unneeded code and logic according to mission-specific or end-point-specific configurations and dependencies, and ii) shields itself against exploitation by retrofitting specialized protection mechanisms, such as code randomization and data isolation. Endpoint-specific specialization is facilitated by a novel binary code transformation framework that relies on compiler-rewriter cooperation to enable fast and robust fine-grained code transformation on endpoints, while achieving transparent deployment by maintaining compatibility with existing software distribution models.

4.11 Trusted Computing: The Biggest Failure or Opportunity?

Ahmad-Reza Sadeghi (TU Darmstadt, DE)

After years of research in hardware security, we are still missing adequate solutions to protect modern computing platforms. Deployed hardware solutions like PUFs, TPMs, and Trusted Execution Environments (TEEs) are lacking widespread usage, or have been attacked

through various side-channels. Additionally, we are witnessing a shift towards cross-layer attacks, exploiting hardware vulnerabilities from software, also remotely, as demonstrated recently by attacks like CLKScrew, Meltdown, and Spectre, which affect even systems with advanced defenses such as (Control Flow Integrity (CFI). Moreover, the Hack@DAC 2018 hardware security competition revealed a protection gap for current chip designs, since existing verification approaches may fail to detect certain classes of vulnerabilities in RTL code. In this talk will provide an overview of hardware-assisted security. We will discuss the impact of deployed solutions, their strengths and shortcomings, as well as new research directions.

4.12 Challenges of Regulating Security

Christoph Sorge (Universität des Saarlandes, DE)

Can legislation help mitigate the "Biggest Failures in Security"? Laws can obviously influence behaviour, and provide incentives to prioritize security. However, IT security legislation is hard due to conflicting goals.

Unspecific laws are not very useful. They lead to uncertainty, and even companies trying to abide by the laws risk fines or civil liability. Too specific regulation is quickly outdated, and can only cover individual sectors. Instead of detailed ex-ante regulation, liability rules could be considered as an alternative. Liability, however, also requires an understanding of obligations (and does not replace this understanding). As a consequence, IT security regulation usually has a limited scope. The focus can be on a specific sector, or on specific aspects such as security management and processes.

The German communication platform used for communication between lawyers and courts (beA) may serve as an example for a failure in security. Its security issues were, in part, caused by a problematic regulation approach and a resulting lack of requirements engineering.

To conclude, security legislation may work, as long as its scope is limited, and there are ways to adapt the legal requirements due to technical innovation. The technical community, however, should not wish for a detailed and overarching security regulation.

4.13 Fantastic Embedded Security Failures and Where to Find Them.

Lennert Wouters (KU Leuven, BE)

Main reference Lennert Wouters, Eduard Marin, Tomer Ashur, Benedikt Gierlichs, Bart Preneel: "Fast, Furious and Insecure: Passive Keyless Entry and Start Systems in Modern Supercars", IACR Trans. Cryptogr. Hardw. Embed. Syst., Vol. 2019(3), pp. 66–85, 2019.
URL https://doi.org/10.13154/tches.v2019.i3.66-85

During this talk we discuss common security issues encountered in embedded devices.

We take a look at issues ranging from the use of broken cryptographic primitives to insecure firmware updates and backend API issues. All of these issues are discussed using several real world examples ranging from vacuum cleaners to high-end cars. The goal of this talk is to spark discussion on how these issues came to be and how we can prevent them in the future.

4.14 Layers of Abstraction and Layers of Obstruction

Moti Yung (Columbia University – New York, US)

In this work we argue that what has made the field of "computer science" and its realization in real life as the "Information Technology Industry" successful, in fact, makes security hard! The success of computer science evolves around its evolution as a field where "complexity is controlled", namely, the ability to abstract sub-problems and sub-fields, solve problems in the abstracted domain and then apply it to the entire system in the right layer. The ability to solve concrete specific problems within a layer extends to sub-area, which enables the splitting of computer science into well defined courses: one can study hardware, computer organization, architecture, software, operating systems, databases, computer languages, algorithms, etc. in separate courses, yet in reality computation as a field employs all areas. Well defined API's and other mechanisms to connect layers enable also separate companies to deal with a subarea: hardware, database management system, cloud infrastructure, application package, which again, in reality work together.

When it comes to the area of security, we have to deal with an external threat, typically described as a threat model or an adversary. The adversary is an external entity to the system, hence it does not obey the layering assumption and design methodology: it is going to attack across layers! Thus, to defend systems practically, the notion of ethical hacking (white hat and red hat teams) that mimc attacks and itself performs attacks and observations across layers is typically employed.

We examine how the layers of abstraction, most often obstruct design of security. We ask: Is practical white hat monitoring and examination the only way to remedy the situation, or can design be updated to include some cross layers security considerations? We attempt to examine by example the latter.

The example we use is the development of the "Universal Second Factor" (U2F) by looking at the small example of servers, additional servers, user, device, and second factor device. By showing that considering attacks of different elements in the system, and further measures that are taken to cope with it, a design which is more robust and foils more attacks can be achieved. It demonstrates a possible refinement methodology, which adds attacks on other layers, as part of refining a design of a layer, thus being much more robust than merely considering each layer by itself.

5 Working groups

5.1 Certification Working Group

Felix Freiling (Universität Erlangen-Nürnberg, DE) and Begül Bilgin (Rambus – Rotterdam, NL & KU Leuven, BE)

The topic of this working group started out rather fuzzy as a discussion involving "certification, quantification, liability, responsibility, etc." in the context of avoiding security failures in the future, and so the working group started by collecting and sorting out the issues that

had motivated the participants to join the group. Participants were asked to provide specific questions which were subsequently grouped into three main categories:

1. Technical aspects of certification, e.g., how to integrate different perspectives and needs into the certification process,
2. understanding certificates, e.g., how to formulate the essence of the certified security properties so that relevant stakeholders can understand them, and
3. the big picture, on how responsibility, liability and regulation work together, e.g., the usefulness of certification in the absence of quantified risk models, possible civil or criminal liability for bad security products, or economic incentives for certification.

We aimed to go top down from category 1 to 3, but since a lot of questions from category 1 had been discussed in the talk by Volkmar Lotz on "Security Product Certification" the discussion started from category 2 with frequent side steps into other categories.

5.1.1 Understanding Certificates

Certification is often confused with penetration testing, a common technique to exhibit the security of a system in practice. While both topics are related, certification usually consists of a fixed set of tests that are performed more in the direction of a a checklist, while in penetration testing a skilled attacker tries to find vulnerabilities with defined resources. Security certification in terms of Common Criteria, however, is very close to penetration testing since it required independent analysis, repeatability, and a definition of attackers' resources.

What is also often confused is that a certificate for some part of a system does not necessarily imply the security of that part of the system. It always depends on the scope of the certification and the commitment of the involved parties. For example, if a specific security parameter (e.g., the ECC curve choice) was not included in the certificate, then plugging in the wrong security mechanism (the wrong curve) makes the system vulnerable. In the ideal process of *committed certification* all stakeholders try to honestly and with true interest try to raise the security level of a system or product through certification. But in practice, it is often not clear whether certification is applied in this way. This is exhibited by the often fierce battle of stakeholders about the scope of certification. A trait often seen in practice and termed *creative certification* is to formulate the certification goals in such a way that they sound good and appear to capture the essence of what is to be proven, but at second sight fail to follow the spirit of certification. Certifying a product will therefore often follow the letter of law but lead to no clear increase in security. Even worse, *fraudulent certification* tries to misuse the certification process to make certain stakeholders like the public believe in a security property which was never actually intended to hold.

In this context it is important to understand the concept of a *protection profile* as defined in the Common Criteria, which is a carefully crafted statement of the security targets and the associated resource bounds (cost, etc.) for attackers tailored to a specific class of systems. The discussions frequently referred to the example of a protection profile for electronic voting systems developed in Germany by the Federal Information Security Agency (BSI) which took about 4 years to develop. This is also a general problem in certification: certification documents must be precise, but they still should be understandable. Today, many certification documents are dominated by rather mechanical language that is hard to understand by people who are not from the regulation field. For researchers, for example, it is often easier to read and understand an evaluation report from an independent evaluator or white hat hacker that is written more like a research paper.

Looking at certification in terms of Common Criteria, it was mentioned that certification appears to work better for hardware than for software. The reasons for this were conjectured to be (1) the higher complexity of big software systems in contrast to big hardware systems and (2) the need (or maybe trend) of commercial software for frequent functionality updates. It was also mentioned that in the context of safety systems, systems are only allowed to operate in a certified state. The discovery of a security vulnerability puts system operators in a conflict between safety and security: they may either keep the safety certification of the system and risk successful attacks, or violate safety considerations due to security updates. This is a fundamental and still unsolved goal conflict.

5.1.2 Technical Aspects of Certification

As discussed above, the scope of certification is important and is usually described in the protection profile. In a certain sense, it defines what is "sufficient" to call a system secure. Perfect security, i.e., the ability to withstand all attacks, if often not the aim. For certain attacks, other security behaviors can be acceptable. With respect to data protection issues it was asked whether we can get inspiration from other application areas about what happens when a software component does not function as it is supposed to or when there are usability problems, e.g., for a customer to claim the money used to purchase the product.

The newly introduced European data protection legislation GDPR states explicitly that "state-of-the-art" security evaluation has to be perfomed, but it also mentions the cost factor. It is often not so clear and debated what this means in practice, and this is also an issue where the research community needs to take a stand.

Another example is the legislation involving critical infrastructures where also state-of-the-art certification is often referred to. Such infrastructures are geeting increasingly large. As an example the infrastrucure to manage millions of autonomous vehicles in the future was mentioned. There are, however, already examples from this area that were discussed. For example, trains and the railway system have a long tradition of safety and (partly also) security evaluation. There, a large system (a complete train system) is divided into pieces (physical or logical) which should have the same security level and which should follow standardized functional requirements almost to the point of having a checklist. Problems arise in the interconnections of these systems because composibility of security properties and checking for composed requirements are known hard problems. It was discussed in what way the division into parts could help in the case of updates for already certified devices. One could try to use isolation properties to update certain parts of the system without affecting others.

On a technical side, it was asked whether formal verification could not be used to a larger extent in certification. It was stated that to a certain extent, formal verification is already part of many certification processes, but in the end the input and the output of a certification is a document in human language and not in a formal language which could be used as a basis for formal verification. So generally, a first step in using formal verification in certification is to formalize the set of security targets appropriately, a hard task in its own.

5.1.3 Liability, Resposibility, and Regulation

We finally turned to the third aspect of the discussion, the big picture involving liability and regulation. The interplay between these issues and certification was a frequent initial question since not every product needs security but it appears that everything connected to the Internet could need some minimal form of security. In this context issues of *negligence*

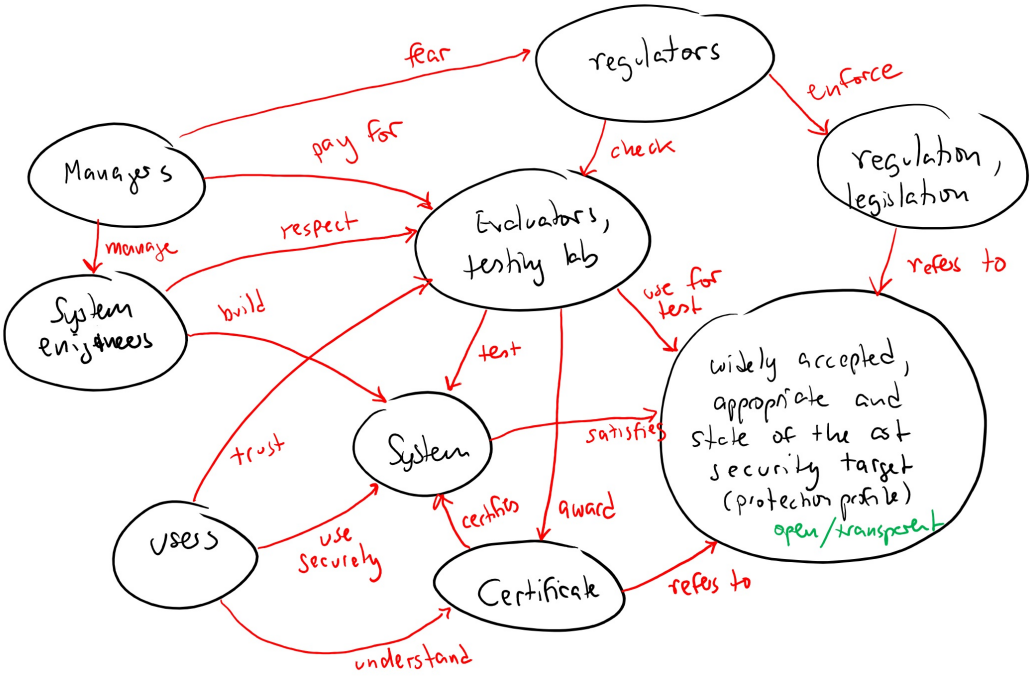

Figure 1 The "ideal world" of certification for security.

were discussed, a term coming from regulation but often introduced into discussions by people from the research community. However, the question is where does negligence start? The problem is that there is no common consensus from the security community. It is important that the security community attempts to interact with the law/regulation community to have more concrete orientation points.

In practice the incentives for certification are various, some involve regulation necessities (like critical infrastructures or GDPR) and risks of law suits agains a company, others involve the risks of bad press and the general problem of naming and shaming that appears to work sometimes at least.

The certification process as a whole was also questioned: Would it be better in certain circumstances to not certify a system at all so as to not create any false expectations? Should we have something more lightweight in-between certification and no certification that is a bit faster but still understandable? It is not so clear what this could be, alltough it would surely be better than performing no security evaluation at all. It is however important to raise no false expectations, as with regular certification. It boils down to knowledge of different levels of assurance, the target of evaluation and the assumptions that come with the evaluation, and an understandable message to the end user and other stakeholders.

5.1.4 Summary

When preparing the results of the discussions during the workshop, the authors of this summary felt that it was easier to summarize the discussed issued based on an understanding of the ideal world in certification for security (see Figure 1, taken from the presentation). In this ideal world, there exist widely accepted, appropriate and state-of-the-art security targets for the system in question that are also openly accessible. Certificates, that are issued by

trusted and independent testing labs or evaluators, can then refer to these targets to test the system. Regulators in turn check the practices of the evaluators to avoid fraudulent certification. In the end, users can then use the system securely.

Obviously, there are many open issues in which the real world differs from the ideal world. Most critically, a set of "widely accepted, appropriate and state-of-the-art security targets" does not exist for most systems, especially security targets that contain measures based on empirical evidence. Furthermore, such security targets are necessary for regulators to define negligence and enforce liability (and motivate managers to pay for certification). In turn, certification standardizes such security targets and can be used for "branding" security, but they still can be misused in the sense of creative and/or fraudulent compliance and misunderstood. Lastly, independent, professional evaluators with high work ethics are needed for trustworthy evaluation. This statement is true even independently of certification.

At the end of the discussions we collected a final round of statements on what participants had taken from the discussions. Here is an unsorted list, that still gives a good insight into the final mental state of the group:

- We need certification but it is unclear how to do this for complex systems.
- We need to define meaningful certifications.
- Currently, certifications are a marketing thing.
- Hardware certification is different from software certification.
- I am skeptical about certification.
- We need a Dagstuhl seminar on certification for security.
- Expectations on certification are too high.
- Certification is better than doing nothing.
- Certification has limited scope, but what is the scope?
- Certificates are necessary.
- Certificates shouldn't lead to blame shifting.

It seems that discussions must continue.

5.2 Education Working Group

Lucy Hunt (Lancaster University, GB), Magnus Almgren (Chalmers University of Technology – Göteborg, SE), Hervé Debar (Télécom SudParis, FR), Fabio di Franco (ENISA – Attica, GR), Sven Dietrich (City University of New York, US), Daisuke Fujimoto (Nara Institute of Science and Technology, JP), Youngwoo Kim (Nara Institute of Science and Technology, JP), Gabriele Lenzini (University of Luxembourg, LU), Olivier Levillain (Télécom SudParis – Evry, FR), Lennert Wouters (KU Leuven, BE), and Moti Yung (Columbia University – New York, US)

5.2.1 Introduction and Approach

Despite enormous efforts, securing IT systems remain an open challenge for both community and industry. Prior to the Dagstuhl seminar, participants identified key security failures and challenges through a "Biggest Failures in Security" survey. From the presented survey results, the group decided on three strategy areas to explore in smaller working groups:

- Certification
- Education
- Human Factors

Working group 2, made up of 11 people, explored education as a strategy to the identified challenges. We reflected that IT security is a multi-disciplinary field, raising questions about our understanding of the population and diversity of engineers – who are the key stakeholders that need educating about security? To make an impact through education we have to understand the audience and effectiveness of channels for sharing and maintaining usable system security mechanisms, knowledge and best practices. We identified educational goals for three stakeholder groups:

- Formally educated engineers
- Non-formally educated engineers
- Industry and the general public

We need to better educate and communicate security knowledge to engineers (software, firmware, hardware, electrical, networking, Internet of Things etc.) that study at university or take formal training. We also need to find ways to identify and reach non-formally educated (e.g. self-trained) engineers – there are many more people than before coding (or such), whose code may have an effect on security. The overarching goal is to make security a good (and easy) thing to do – from the usability for end users to the security design decisions engineers make while building systems. Alongside this, we identified the need for societal change where there is a better understanding, demand and willingness to pay for secure devices, products and services.

5.2.2 Engineers

For engineers, we discussed the need to:

1. Implement incentives and support for educators to demonstrate secure practices and behaviors. This means pointing to secure coding standards and verified (or at least vetted) best practices. In terms of education best practices, successful channels allowed interaction with incident response teams (source: CERT).
2. Identify and train (retrain) IT professionals in security best practices – in particular people that haven't recently or ever been through formal software engineering education, are self-taught software engineers or come in from different fields. All have a need for security resources, training and support. We identified examples of local security education initiatives (CyberEdu in France, Seccap in Japan), the challenge is how to scale globally and so reach larger audiences. We reflected on certification schemes and organisational responsibility for security – as working group 1 were looking at this we parked that discussion.
3. Find ways to attract and train new people into security roles. Security practitioners are sometimes seen as "getting in the way" of the software development life cycle, if security has not been properly integrated into the process previously.
4. Develop better code re-use opportunities, to take advantage of engineer laziness (cut and paste of code samples). To make good practices more accessible, while trying to eradicate bad examples from the publicly available online resources.

We identified a number of solutions, focusing on helping engineers to identify best practices (rather than common practices):

1. Creating tools (e.g. compilers) and methods to make it easier to do the good/right thing. We need software development tools that make security an integral part. Both the creation of more secure code as well as providing feedback to the programmer (software engineer) to transparently move forward with enhancing security aspects are needed.

2. Designing better security mechanisms for engineers, that improve usability for the users. Engineers need to work closely with usability experts to allow better interaction of the users with the hardware devices and their associated operating systems and application software.
3. Teaching engineers how to design user interfaces that help get security concepts across to users. Early studies going back to 1999 showed challenges and confusion when it comes to security concepts. An interdisciplinary approach is needed in training engineers to convey the right ideas. Sample best practices and positive feedback would help reinforce this approach.
4. Enhancing code sharing platforms (such as the web site Stack Overflow): new voting for "best practice secure" answers and code samples so that less experienced programmers understand the choices they make when copying and pasting shared code. By providing accessible and vetted code samples, designs, or approaches, best secure practices would be promoted, while making sure the engineers understand why that choice was made.
5. Industry incentives: "follow these practices – we will rank your app higher". A reward mechanism with software repositories, such as Apple App Store or Google Play Store for mobile and desktop software, could issue a higher ranking for developers that adhere to best practices.

5.2.3 Industry and Public

For industry and the wider public, we need to:

- capture and share IT failures and consequences – to exploit failures and raise awareness
- find approaches to better demonstrate security – to experts, industry and wider society
- motivate people to value and prioritize security requirements

We discussed initiatives for the wider engagement and awareness raising within society including better publicity of vulnerabilities and associated real life failure and success stories – how do we capture and learn from our mistakes? Can the need for security be compared to the climate change movement – can we use society to drive changes?

5.2.4 Further Questions

Other questions raised for further discussion:

1. What has been the impact of GDPR (General Data Protection Regulation) on secure coding practices?
2. Regarding workplace incentives and measures – what metrics are there for secure practices?
3. How has the population and diversity of (software) engineers changed?
4. Is education really failing us in security – how to measure the success and impact of security education?

5.3 Human Factors Working Group

Joachim Meyer (Tel Aviv University, IL), Robert Biddle (Carleton University – Ottawa, CA), Sebastian Pape (Goethe-Universität Frankfurt am Main, DE), Kazue Sako (NEC – Kawasaki, JP), Martina Angela Sasse (Ruhr-Universität Bochum, DE), Stephan Somogyi (Google Inc. – Mountain View, US), Borce Stojkovski (University of Luxembourg, LU), Ingrid Verbauwhede (KU Leuven, BE), and Yuval Yarom (University of Adelaide, AU)

The seminar provided a group of participants with different academic and employment backgrounds with the opportunity to learn and to reflect about what might be the greatest failures and threats in security today. Our specific group dealt with the roles humans and human behavior have in security. The work is based on the premise that the introduction of threats into systems often results from human actions, which may be inadvertent (e.g., the opening of an infected email attachment) or may be deliberate risk taking (e.g., the override of a certificate warning about a site). The group spent several hours discussing ways to address the issue of human involvement in threats. It became clear that this is a complex, multilayered problem, that still warrants a comprehensive conceptual analysis. The group started to discuss the possibility of writing a "cybersecurity harm-reduction manifesto" that would be a synthesis of the different positions brought by the members of the group. In particular, the idea would be to apply ideas from public health by making efforts at a broad level to reduce real harm, rather than offloading the responsibility onto individual users, stigmatizing human behavior and blaming users for any failures.

Major points that arose in the discussions include:

1. Humans are involved in systems in various, often very different capacities (developers, system administrators, security experts, end users, etc.). The knowledge, preferences, and attitudes towards security issues may differ greatly between these groups.
2. The dealing with threats can take various forms, and it is not clear under which conditions, which approach might be best. For instance, one can aim to design out the possibility of threats materializing, one can lower the harm that may be done if a threat materialized, one can train people to detect and cope intelligently with threats, etc. It is not clear how realistic the adoption of different approaches will be to deal with different threats.
3. We still lack well-substantiated knowledge about the effectiveness of different risk-reduction methods. Intuitive approaches (e.g., force users to have very long passwords, which need to be changed every few weeks) often fail to provide the desired results.
4. Security-related behavior is part of a person's interaction with the system. The person's perceptions of risks and the adequacy of different behaviors, the estimates of costs and benefits of different outcomes, and the user's mental model of the system and its security all affect the user's actions and choices. The design of secure systems will also require the design of the interactions that support secure behavior.
5. We still lack theoretical tools to predict the effects, changes in the system, the threats, the environment or the user may have on risk-related behaviors. A major challenge for the scientific work in this field will be to develop and validate such tools.

These points demonstrate the wealth of topics that were discussed and that need to be considered when dealing with the human aspects of security threats and failures. The Dagstuhl seminar can serve as a starting point for discussions and the development of joint research on this broad topic.

Participants

- Tigist Abera
TU Darmstadt, DE
- Magnus Almgren
Chalmers University of Technology – Göteborg, SE
- Frederik Armknecht
Universität Mannheim, DE
- Daniel J. Bernstein
University of Illinois – Chicago, US
- Sarani Bhattacharya
KU Leuven, BE
- Robert Biddle
Carleton University – Ottawa, CA
- Begül Bilgin
Rambus – Rotterdam, NL & KU Leuven, BE
- Dominik Brodowski
Universität des Saarlandes, DE
- Marc C. Dacier
EURECOM – Sophia Antipolis, FR
- Hervé Debar
Télécom SudParis, FR
- Fabio di Franco
ENISA – Attica, GR
- Sven Dietrich
City University of New York, US
- Felix Freiling
Universität Erlangen-Nürnberg, DE
- Daisuke Fujimoto
Nara Institute of Science and Technology, JP
- Lucy Hunt
Lancaster University, GB
- Ghassan Karame
NEC Laboratories Europe – Heidelberg, DE
- Stefan Katzenbeisser
Universität Passau, DE
- Florian Kerschbaum
University of Waterloo, CA
- Youngwoo Kim
Nara Institute of Science and Technology, JP
- Tanja Lange
TU Eindhoven, NL
- Gabriele Lenzini
University of Luxembourg, LU
- Olivier Levillain
Télécom SudParis – Evry, FR
- Volkmar Lotz
SAP Labs France – Mougins, FR
- Michael Meier
Universität Bonn, DE
- Joachim Meyer
Tel Aviv University, IL
- Vasily Mikhalev
Universität Mannheim, DE
- Christian Müller
Universität Mannheim, DE
- Sebastian Pape
Goethe-Universität Frankfurt am Main, DE
- Michalis Polychronakis
Stony Brook University, US
- Kai Rannenberg
Goethe-Universität Frankfurt am Main, DE
- Ahmad-Reza Sadeghi
TU Darmstadt, DE
- Kazue Sako
NEC – Kawasaki, JP
- Martina Angela Sasse
Ruhr-Universität Bochum, DE
- Stephan Somogyi
Google Inc. – Mountain View, US
- Christoph Sorge
Universität des Saarlandes, DE
- Borce Stojkovski
University of Luxembourg, LU
- Ingrid Verbauwhede
KU Leuven, BE
- Melanie Volkamer
KIT – Karlsruher Institut für Technologie, DE
- Edgar Weippl
SBA Research – Wien, AT
- Lennert Wouters
KU Leuven, BE
- Yuval Yarom
University of Adelaide, AU
- Moti Yung
Columbia University – New York, US

Report from Dagstuhl Seminar 19452

Machine Learning Meets Visualization to Make Artificial Intelligence Interpretable

Edited by

Enrico Bertini[1], Peer-Timo Bremer[2], Daniela Oelke[3], and Jayaraman Thiagarajan[4]

1 **NYU – Brooklyn, US,** `enrico.bertini@nyu.edu`
2 **LLNL – Livermore, US,** `bremer5@llnl.gov`
3 **Siemens AG – München, DE,** `daniela.oelke@siemens.com`
4 **LLNL – Livermore, US,** `jjayaram@llnl.gov`

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 19452 "Machine Learning Meets Visualization to Make Artificial Intelligence Interpretable".

Seminar November 3–8, 2019 – http://www.dagstuhl.de/19452
2012 ACM Subject Classification Human-centered computing → Visualization, Computing methodologies → Artificial intelligence, Computing methodologies → Machine learning
Keywords and phrases Visualization, Machine Learning, Interpretability
Digital Object Identifier 10.4230/DagRep.9.11.24

1 Executive Summary

Enrico Bertini (NYU – Brooklyn, US, enrico.bertini@nyu.edu)
Peer-Timo Bremer (LLNL – Livermore, US, bremer5@llnl.gov)
Daniela Oelke (Dep. of Informatics, Siemens AG – München, DE, daniela.oelke@siemens.com)
Jayaraman J. Thiagarajan (LLNL – Livermore, US, jayaram@llnl.gov)

The recent advances in machine learning (ML) have led to unprecedented successes in areas such as computer vision and natural language processing. In the future, these technologies promise to revolutionize everything ranging from science and engineering to social studies and policy making. However, one of the fundamental challenges in making these technologies useful, usable, reliable and trustworthy is that they are all driven by extremely complex models for which it is impossible to derive simple (closed-format) descriptions and explanations. Mapping decisions from a learned model to human perceptions and understanding of that world is very challenging. Consequently, a detailed understanding of the behavior of these AI systems remains elusive, thus making it difficult (and sometimes impossible) to distinguish between actual knowledge and artifacts in the data presented to a model. This fundamental limitation should be addressed in order to support model optimization, understand risks, disseminate decisions and findings, and most importantly to promote trust.

While this grand challenge can be partially addressed by designing novel theoretical techniques to validate and reason about models/data, in practice, they are found to be grossly insufficient due to our inability to translate the requirements from real-world applications into tractable mathematical formulations. For example, concerns about AI systems (e.g., biases) are intimately connected to several human factors such as how information is perceived,

Machine Learning Meets Visualization to Make Artificial Intelligence Interpretable, *Dagstuhl Reports*, Vol. 9, Issue 11, pp. 24–33
Editors: Enrico Bertini, Peer-Timo Bremer, Daniela Oelke, and Jayaraman Thiagarajan

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

cognitive biases, etc. This crucial gap has given rise to the field of *interpretable machine learning*, which at its core is concerned with providing a human user better understanding of the model's logic and behavior. In recent years, the machine learning community, as well as virtually all application areas, have seen a rapid expansion of research efforts in interpretability and related topics. In the process, visualization, or more generally interactive systems, have become a key component of these efforts since they provide one avenue to exploit expert intuition and hypothesis-driven exploration. However, due to the unprecedented speed with which the field is currently progressing, it is difficult for the various communities to maintain a cohesive picture of the state of the art and the open challenges; especially given the extreme diversity of the research areas involved.

The focus of this Dagstuhl Seminar was to convene various stakeholders to jointly discuss needs, characterize open research challenges, and propose a joint research agenda. In particular, three different stakeholders were engaged in this seminar: application experts with unmet needs and practical problems; machine learning researchers who are the main source of theoretical advances; and visualization and HCI experts that can devise intuitive representations and exploration frameworks for practical solutions. Through this seminar, the group of researchers discussed the state of practice, identified crucial gaps and research challenges, and formulated a joint research agenda to guide research in interpretable ML.

Program Overview

The main goal of this Dagstuhl seminar was to discuss the current state and future research directions of interpretable Machine Learning. Because two different scientific communities met, the Machine Learning community and the Visualization community, we started the seminar by discussing and defining important terms and concepts of the field. Afterwards, we split up into working groups to collect answers to the following questions: "*Who needs interpretable machine learning? For what task is it needed? Why is it needed?*". This step was then followed by a series of application lightning talks (please refer to the abstracts below for details).

On the second day, we had two overview talks, one covering the machine learning perspective on interpretability, and the other one the visualization perspective on the topic. Afterwards, we built working groups to collect research challenges from the presented applications and beyond.

The third day was dedicated to clustering the research challenges into priority research directions. The following priority research directions were identified:

- Interpreting Learned Features and Learning Interpretable Features
- Evaluation of Interpretability Methods
- Evaluation and Model Comparison with Interpretable Machine Learning
- Uncertainty
- Visual Encoding and Interactivity
- Interpretability Methods
- Human-Centered Design

On Thursday, the priority research directions were further detailed in working groups. We had two rounds of working groups in which 3, respectively 4, priority research challenges were discussed in parallel by the groups according to the following aspects: problem statement, sub-challenges, example applications, and related priority research directions. Furthermore, all research challenges were mapped into descriptive axes of the problem space and the solution space.

On the last day, we designed an overview diagram that helps to communicate the result to the larger scientific community.

2 Table of Contents

Executive Summary
Enrico Bertini, Peer-Timo Bremer, Daniela Oelke and Jayaraman J. Thiagarajan 24

Overview of Talks

Understanding Generative Physics Models with Scientific Priors
Rushil Anirudh . . . 27

VIS Perspectives on Interactive and Explainable Machine Learning
Mennatallah El-Assady . . . 28

Modernizing Supercomputer Monitoring via Artificial Intelligence
Elisabeth Moore . . . 28

Interpretability Applications: Materials Discovery and Recidivism Prediction
Sorelle Friedler . . . 28

Human in the loop ML
Nathan Hodas . . . 29

Application Scenarios for Explainable AI in an Industrial Setting
Daniela Oelke . . . 29

Explainable AI for Maritime Anomaly Detection and Autonomous Driving.
Maria Riveiro . . . 29

Ada Health GmbH: ExAI in Digital Health
Sarah Schulz . . . 30

XAI for insurance
Jarke J. van Wijk . . . 31

Open problems

Interpretability for Scientific Machine Learning
Peer-Timo Bremer . . . 31

Open Questions and Future Directions in Interpretability Research
Sebastian Lapuschkin . . . 31

Explainability for affected users. The role of Information Design
Beatrice Gobbo . . . 32

Participants . . . 33

3 Overview of Talks

3.1 Understanding Generative Physics Models with Scientific Priors

Rushil Anirudh (LLNL – Livermore, US)

Joint work of Rushil Anirudh, Jayaraman J. Thiagarajan, Peer-Timo Bremer, Brian K. Spears
Main reference Rushil Anirudh, Jayaraman J. Thiagarajan, Shusen Liu, Peer-Timo Bremer, Brian K. Spears: "Exploring Generative Physics Models with Scientific Priors in Inertial Confinement Fusion", CoRR, Vol. abs/1910.01666, 2019.
URL https://arxiv.org/abs/1910.01666

Modern neural networks are highly effective in modeling complex, multi-modal data and thus have raised significant interested in exploiting these capabilities for scientific applications. In particular, the ability to directly ingest multi-modal, non-scalar data, i.e. images, energy spectra, etc., has proven to be a significant advantage over more traditional statistical approaches. One common challenge for such systems is to properly account for various invariants and constraints to guarantee physically meaningful results, i.e. positive energy, mass conservation, etc. Existing approaches either integrate the physical laws, or rather the corresponding partial differential equations, directly into the training process or add the constraints into the loss function. However, this only works for known constraints that can be explicitly formulated as some differentiable equation in order to be integrated into the neural network training. In practice, not all constraints are known or can be formulated in this manner and explicitly enforcing some constraints while ignoring others is likely to bias the resulting system. Furthermore, constraints are often based on unrealistic assumptions, i.e. physical relationships under some idealized condition, which are not satisfied in the real data. Consequently, strictly enforcing such constraints may produce incorrect results.

In this talk, I explored a few ways in which we can explore, evaluate, and understand the behavior of generative models for scientific datasets. By directly incorporating all known constraints into the loss function, evaluating the constraints post-hoc becomes a self-fulfilling prophecy with the compliance driven largely by the choice of weights in the loss function and a significant potential to over-correct the results. At the same time, most existing metrics are either designed for traditional computer vision problems like Inception scores, FID-scores, or they rely on other global metrics like manifold alignment, which may have little significance in the scientific context. Instead, we propose to use the constraints to evaluate a generative model and show how exploring the data distribution in latent space, i.e. the physics manifold, through the lense of the constraint can provide interesting insights. In particular, we use Inertial Confinement Fusion (ICF) as a testbed problem, with multi-modal data generated from a 1D semi-analytic simulator.

3.2 VIS Perspectives on Interactive and Explainable Machine Learning

Mennatallah El-Assady (Universität Konstanz, DE)

Main reference Thilo Spinner, Udo Schlegel, Hanna Schäfer, Mennatallah El-Assady: "explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning", IEEE Trans. Vis. Comput. Graph., Vol. 26(1), pp. 1064–1074, 2020.
URL https://doi.org/10.1109/TVCG.2019.2934629

Interactive and explainable machine learning can be regarded as a process, encompassing thee high-level stages: (1) understanding machine learning models and data; (2) diagnosing model limitations using explainable AI methods; (3) refining and optimizing models interactively.

In my talk, I review the current state-of-the-art of visualization and visual analytics techniques by grouping them into the three stages. In addition, I argue for expanding our approach to explainability through adapting concepts like metaphorical narratives, verbalization, as well as gamification.

I further introduce the explAIner.ai framework for structuring the process of XAI and IML, as well as operationalizing it through a TensoBoard plugin.

Lastly, to derive a robust XAI methodology, I present a survey on XAI strategies and mediums, transferring knowledge and best practices gained from other disciplines to explainable AI.

3.3 Modernizing Supercomputer Monitoring via Artificial Intelligence

Elisabeth Moore (Los Alamos National Laboratory, US)

This talk is an overview of recent advances at Los Alamos National Laboratory regarding the use of machine learning / artificial intelligence to improve management of datacenters and large-scale computing facilities. Three primary projects will be discussed: (1) Anomaly detection in computer-generated text logs, (2) Natural language processing for job outcome prediction, and (3) Effectiveness of telemetry data for predicting node failures.

3.4 Interpretability Applications: Materials Discovery and Recidivism Prediction

Sorelle Friedler (Haverford College, US)

I present two applications where interpretability is important. First, in materials discovery, the goal is to predict the outcome of chemical experiments. Specifically, the problem is framed as a classification problem where the goal is to predict whether a given set of reactants, at specific masses, temperature, and other experimental conditions, will produce a crystal or not. The goal of the chemists involved in the project is to develop and test scientific hypotheses, i.e., to learn as much as possible about science from the machine learning models. Second, in recidivism predictions, the goal is to reduce the number of people detained pre-trial in the U.S. by releasing more defendants determined to be "low risk". The interpretability goals for this task focus on both understanding each step in a model's prediction and understanding potential unfairness (both racism and sexism) in the machine learning models; both are necessary for defense lawyers to best do their job.

3.5 Human in the loop ML

Nathan Hodas (Pacific Northwest National Lab. – Richland, US)

For few-shot learning, the user specifies a small training set (1-5 images or data points) and the system looks for matches. With only a few data points, this allows for ambiguity in the task. In this case, the user needs to "explain" to the computer what the task is (what does it mean to make a good match?). Similarly, the computer needs to explain to the user how it is making decisions, so the user can alter their explanations, in turn.

Sharkzor is used by scientists and other non-data scientists to conduct ML in real-time without any code, so any solution needs to leverage strong human-in-the-loop analytics and minimal friction for interaction. Taken together, HITL explanations and few-shot learning will become increasingly important for non-ML experts to benefit from advanced Machine Learning.

3.6 Application Scenarios for Explainable AI in an Industrial Setting

Daniela Oelke (Siemens AG – München, DE)

In my talk I gave three examples for industrial applications with a need for making machine learning models transparent. In the first example XAI is needed to get a proof that the employed machine learning model takes the right decision in all potential situations of a safety-critical scenario. The second example showcased an application in which the decisions of an anomaly detection system had to be explained. Finally, I presented a use case from the domain of energy management in which the need for calibrated trust and validation was on the focus.

3.7 Explainable AI for Maritime Anomaly Detection and Autonomous Driving.

Maria Riveiro (Univ. of Skövde, SE & Univ. of Jönköning, SE)

Main reference Maria Riveiro: "Evaluation of Normal Model Visualization for Anomaly Detection in Maritime Traffic", TiiS, Vol. 4(1), pp. 5:1–5:24, 2014.
URL https://doi.org/10.1145/2591511
Main reference Tove Helldin, Göran Falkman, Maria Riveiro, Staffan Davidsson: "Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving", in Proc. of the Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI '13, Eindhoven, The Netherlands, October 28-30, 2013, pp. 210–217, ACM, 2013.
URL http://dx.doi.org/10.1145/2516540.2516554

The aim of this talk is to present two application scenarios where visual explanations were provided in order to support users' decision-making processes.

The first scenario, maritime anomaly detection [1], concerns the analysis of spatio-temporal data to find anomalous behavior in maritime traffic. In this case, machine learning methods

were used to create normal behavioral models of different types of vessels. We studied how to present and explain the models created (for understanding and improvement) and the detected anomalies to various stakeholders.

The second scenario, autonomous driving [2], concerns how to present the capability of an autonomous vehicle to drive safely, and the effects that such visual explanations have on driver's performance, acceptance and trust.

These scenarios showcase specific challenges in explainable AI and interpretable machine learning, for instance: (1) constraints related to the limited time to understand the explanations provided, (2) level of detail and content of the explanations given user's goals and tasks, (3) model improvement by domain experts, (4) design for trust calibration and system acceptance, (5) how to represent and visualize normal behavioral models and anomalies and, finally, (6) evaluation metrics and methods for users using explainable AI-systems over time.

References

1 Riveiro, M. (2014). *Evaluation of normal model visualization for anomaly detection in maritime traffic.* ACM Transactions on Interactive Intelligent Systems (TiiS), 4(1), 5.

2 Helldin, T., Falkman, G., Riveiro, M. and Davidsson, S. (2013). *Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving.* Proc. 5th Int. Conf. on Automotive User Interfaces and Interactive Vehicular Applications (Automotive'UI 13), Eindhoven, The Netherlands.

3.8 Ada Health GmbH: ExAI in Digital Health

Sarah Schulz (Ada Health – Berlin, DE)

Ada Health GmbH develops a system that is meant to be a health companion. It is created by doctors, scientists, and industry pioneers to bring the future of personalized health to everyone. As digital health is clearly a sector which has to deal with the fact that there might be consequences to decisions made by AI systems, explainability and transparency of machine behaviour and output is inevitable. At Ada Health there are essentially two stages where explanations are needed:

- Ada's knowledge base is manually curated by medical experts. In order to support and accelerate this process, we apply Natural Language Processing methods to extract relevant medical information from unstructured text. To enable the medical expert to refuse or accept a suggestion made by the system they need (visual) explanations to make a decision in a given context.
- Since Ada aims at providing access to medical information to everyone and empowering people to understand their health better, the factors that led to the suggested diagnoses have to be transparent and comprehensible for non-expert users.

3.9 XAI for insurance

Jarke J. van Wijk (TU Eindhoven, NL)

Joint work of Dennis Collaris, Leon Vink, Jarke van Wijk
Main reference Dennis Collaris, Leo M. Vink, Jarke J. van Wijk: "Instance-Level Explanations for Fraud Detection: A Case Study", CoRR, Vol. abs/1806.07129, 2018.
URL http://arxiv.org/abs/1806.07129

I first told a story about transparency, based on my experience with a fine I got for a red light. Fortunately, the evidence showed the light was green, and hence this was fixed easily. Next, I described our experience with fraud detection work for an insurance company. My MSc student Dennis Collaris has worked hard on that, with somewhat puzzling results: different methods give different explanations, and also, practioners did not seem to care [1].

References

1 Dennis Collaris, Leon M. Vink, Jarke J. van Wijk. *Instance-Level Explanations for Fraud Detection: A Case Study*. ICML Workshop on Human Interpretability in Machine Learning, 28-33, 2018.

4 Open problems

4.1 Interpretability for Scientific Machine Learning

Peer-Timo Bremer (LLNL – Livermore, US)

The ability of data driven models to ingest complex, multimodal data types has enabled a new generation of surrogate modeling in many scientific and engineering applications going far beyond previous scalar response functions. However, the black box nature of these models make it challenging to derive actionable insights even from highly accurate and well-tuned models. As a result, interpretability has been recognized as one of the key capabilities to exploit the full power of modern machine learning for scientific discovery.

4.2 Open Questions and Future Directions in Interpretability Research

Sebastian Lapuschkin (Fraunhofer-Institut – Berlin, DE)

Within the last decade, neural network based predictors have demonstrated impressive – and at times super-human – capabilities. This performance is often paid for with an intransparent prediction process, hindering wide-spread adoption of modern machine learning techniques due to scepticism, safety concerns and distrust, or legal demands (see the European Union's extended General Data Protection Regulation act), e.g. in healtcare and industry.

Recognizing the demand for novel and appropriate solutions to the interpretability problem in ML, the explainable artificial intelligence (XAI) community has proposed numerous methods and solutions in recent years. Here, it is essential to note, that each existing approach answers a different aspect of the interpretability question, and consequently no method constitutes a comprehensive solution to the problem as a whole. In addition to that, most approaches are only applicable effectively under specific conditions in terms of data domain, model architecture and model task.

With a plethora of options to choose from (including future developments), and the fact that not every stakeholder is also an XAI domain expert it is important to ask and ultimately answer the following questions (among others):

1 Which methods do the right thing for one's intent, model and application? (I.e., which kind of information does the method provide, and does it synergize well with the model, e.g. wrt. model architecture and task)
2 Can we define a catalogue of (measurable) quality criteria for XAI methods, considering [1] ?
3 How can we generate explanations for non-domain-experts, which includes domain-specific knowledge (to avoid improper interpretation of explanations)?
4 How can we bridge the gap from explanations for individual model predictions to explanations truly characterizing the general model behavior?

4.3 Explainability for affected users. The role of Information Design

Beatrice Gobbo (Politecnico di Milano – Milano, IT)

Purposes of interpretable and explainable machine learning range from debugging models to raise awareness about their social impact, especially when these models are wrong or biased. However, if visual analytics and information visualiSation have been largely used for addressing problems as explainability for the debugging processes, the same means and tools have scarcely been used for raising awareness of machine learning miscalculations among lay users. Taking into account the ethical role of data visualiSation and how much abstraction or approximation could be used when representing inner workings of complex machine learning models, the communication and information designer, together with other professional figures such as computer scientists, can design artifacts able to funnel perception of reliance and doubt of results of these technologies.

Participants

- Rushil Anirudh
LLNL – Livermore, US
- Enrico Bertini
NYU – Brooklyn, US
- Alexander Binder
Singapore University of Technology and Design, SG
- Peer-Timo Bremer
LLNL – Livermore, US
- Mennatallah El-Assady
Universität Konstanz, DE
- Sorelle Friedler
Haverford College, US
- Beatrice Gobbo
Polytechnic University of Milan, IT
- Nikou Guennemann
Siemens AG – München, DE
- Nathan Hodas
Pacific Northwest National Lab. – Richland, US
- Daniel A. Keim
Universität Konstanz, DE
- Been Kim
Google Brain – Mountain View, US
- Gordon Kindlmann
University of Chicago, US
- Sebastian Lapuschkin
Fraunhofer-Institut – Berlin, DE
- Heike Leitte
TU Kaiserslautern, DE
- Yao Ming
HKUST – Kowloon, HK
- Elisabeth Moore
Los Alamos National Laboratory, US
- Daniela Oelke
Siemens AG – München, DE
- Steve Petruzza
University of Utah – Salt Lake City, US
- Maria Riveiro
Univ. of Skövde, SE & Univ. of Jönköning, SE
- Carlos E. Scheidegger
University of Arizona – Tucson, US
- Sarah Schulz
Ada Health – Berlin, DE
- Hendrik Strobelt
MIT-IBM Watson AI Lab – Cambridge, US
- Simone Stumpf
City, University of London, GB
- Jayaraman Thiagarajan
LLNL – Livermore, US
- Jarke J. van Wijk
TU Eindhoven, NL

Report from Dagstuhl Seminar 19461

Conversational Search

Edited by

Avishek Anand[1], Lawrence Cavedon[2], Hideo Joho[3], Mark Sanderson[4], and Benno Stein[5]

1 **Leibniz Universität Hannover, DE,** anand@kbs.uni-hannover.de
2 **RMIT University – Melbourne, AU,** lawrence.cavedon@rmit.edu.au
3 **University of Tsukuba – Ibaraki, JP,** hideo@slis.tsukuba.ac.jp
4 **RMIT University – Melbourne, AU,** mark.sanderson@rmit.edu.au
5 **Bauhaus-Universität Weimar, DE,** benno.stein@uni-weimar.de

Abstract

Dagstuhl Seminar 19461 "Conversational Search" was held on 10-15 November 2019. 44 researchers in Information Retrieval and Web Search, Natural Language Processing, Human Computer Interaction, and Dialogue Systems were invited to share the latest development in the area of Conversational Search and discuss its research agenda and future directions. A 5-day program of the seminar consisted of six introductory and background sessions, three visionary talk sessions, one industry talk session, and seven working groups and reporting sessions. The seminar also had three social events during the program. This report provides the executive summary, overview of invited talks, and findings from the seven working groups which cover the definition, evaluation, modelling, explanation, scenarios, applications, and prototype of Conversational Search. The ideas and findings presented in this report should serve as one of the main sources for diverse research programs on Conversational Search.

Seminar November 10–15, 2019 – http://www.dagstuhl.de/19461
2012 ACM Subject Classification Computing methodologies → Artificial intelligence, Computer systems organization → Robotics, Information systems → Information retrieval, Human-centered computing → Human computer interaction (HCI)
Keywords and phrases discourse and dialogue, human-machine interaction, information retrieval, interactive systems, user simulation
Digital Object Identifier 10.4230/DagRep.9.11.34
Edited in cooperation with Khalid Al-Khatib, Jurek Leonhardt, Johanne Trippas

1 Executive Summary

Avishek Anand (Leibniz Universität Hannover, DE)
Lawrence Cavedon (RMIT University – Melbourne, AU)
Hideo Joho (University of Tsukuba – Ibaraki, JP)
Mark Sanderson (RMIT University – Melbourne, AU)
Benno Stein (Bauhaus-Universität Weimar, DE)

Background and Motivation

The Conversational Search Paradigm promises to satisfy information needs using human-like dialogs, be it in spoken or in written form. This kind of "information-providing dialogs" will increasingly happen enpassant and spontaneously, probably triggered by smart objects with which we are surrounded such as intelligent assistants such as Amazon Alexa, Apple Siri,

Google Assistant, and Microsoft Cortana, domestic appliances, environmental control devices, toys, or autonomous robots and vehicles. The outlined development marks a paradigm shift for information technology, and the key question(s) is (are):

What does Conversational Search mean and how to make the most of it–given the possibilities and the restrictions that come along with this paradigm?

Currently, our understanding is still too limited to exploit the Conversational Search Paradigm for effectively satisfying the existing diversity of information needs. Hence, with this first Dagstuhl Seminar on Conversational Search we intend to bring together leading researchers from relevant communities to understand and to analyze this promising retrieval paradigm and its future from different angles.

Among others, we expect to discuss issues related to interactivity, result presentation, clarification, user models, and evaluation, but also search behavior that can lead into a human-machine debate or an argumentation related to the information need in question.

Moreover, we expect to define, shape, and formalize a set of corresponding problems to be addressed, as well as to highlight associated challenges that are expected to come in the form of multiple modalities and multiple users. Correspondingly, we intend to define a roadmap for establishing a new interdisciplinary research community around Conversational Search, for which the seminar will serve as a prominent scientific event, with hopefully many future events to come.

Seminar Program

A 5-day program of the seminar consisted of six introductory and background sessions, three visionary talk sessions, one industry talk session, and nine breakout discussion and reporting sessions. The seminar also had three social events during the program. The detail program of the seminar is available online. [1]

Pre-Seminar Activities

Prior to the seminar, participants were asked to provide inputs to the following questions and request:

1. What are your ideas of the "ultimate" conversational search system?
2. Please list, from the perspective of your research field, important open questions or challenges in conversational search.
3. What are the three papers a PhD student in conversational search should read and why?

From the survey, the following topics were initially emerged as interests of participants. Many of these topics were discussed at length in the seminar.

- Understanding nature of information seeking in the context of conversational agents
- Modelling problems in conversational search
- Clarification and explanation
- Evaluation in conversational search systems
- Ethics and privacy in conversational systems
- Extending the problem space beyond the search interface and Q/A

Another outcome of the above pre-seminar questions was a compilation of recommended reading list to gain a solid understanding of topics and technologies that were related to the research on Conversational Search. The reading list is provided in Section 5 of this report.

[1] https://www.dagstuhl.de/schedules/19461.pdf

Invited Talks

One of the main goals and challenges of this seminar was to bring a broad range of researchers together to discuss Conversational Search, which required to establish common terminologies among participants. Therefore, we had a series of 18 iinvited talk throughout the seminar program to facilitate the understanding and discussion of conversational search and its potential enabling technologies. The main part of this report includes the abstract of all talks.

Working Groups

In the afternoon of Day 2, initial working groups were formed based on the inputs to the pre-seminar questionnaires, introductory and background talks, and discussions among participants. On Day 3, the grouping was revisited and updated, and, eventually, the following seven groups were formed to focus on topics such as the definition, evaluation, modelling, explanation, scenarios, applications, and prototype of Conversational Search.

- Defining Conversational Search
- Evaluating Conversational Search
- Modeling in Conversational Search
- Argumentation and Explanation
- Scenarios that Invite Conversational Search
- Conversation Search for Learning Technologies
- Common Conversational Community Prototype: Scholarly Conversational Assistant

We have summarized the working groups' outcomes in the following. Please refer to the main part of this report for the full description of the findings.

Defining Conversational Search

This group aimed to bring structure and common terminology to the different aspects of conversational search systems that characterise the field. After reviewing existing concepts such as Conversational Answer Retrieval and Conversational Information Seeking, the group offers a typology of Conversational Search systems via functional extensions of information retrieval systems, chatbots, and dialogue systems. The group further elaborates the attributes of Conversational Search by discussing its dimensions and desirable additional properties. Their report suggests types of systems that should not be confused as conversational search systems.

Evaluating Conversational Search

This group addressed how to determine the quality of conversational search for evaluation. They first describe the complexity of conversation between search systems and users, followed by a discussion of the motivation and broader tasks as the context of conversational search that can inform the design of conversational search evaluation. The group also surveys 12 recent tasks and datasets that can be exploited for evaluation of conversational search. Their report presents several dimensions in the evaluation such as User, Retrieval, and Dialog, and suggests that the dimensions might have an overlap with those of Interactive Information Retrieval.

Modeling Conversational Search

This group addressed what should be modeled from the real world to achieve a successful conversational search and how. They explain why a range of concepts and variables such as capabilities and resources of systems, beliefs and goals of users, history and current status of process, and search topics and tasks should be considered to advance understanding between systems and users in the context of Conversational Search. The group points out that the options the current search engines present to users can be too broad in conversational interaction. They suggest that a deeper modeling of users' beliefs and wants, development of reflective mechanisms, and finding a good balance between macroscopic and microscopic modeling are promising directions for future research.

Argumentation and Explanation

Motivated by inevitable influences made to users due to the course of actions and choices of search engines, this group explored how the research on argumentation and explanation can mitigate some of potential biases generated during conversational search processes, and facilitate users' decision-making by acknowledging different viewpoints of a topic. The group suggests a research scheme that consists of three layers: a conversational layer, a demographics layer, and a topic layer. Also, their report explains that argumentation and explanation should be carefully considered when search systems (1) select, (2) arrange, and (3) phrase the information presented to the users. Creating an annotated corpus with these elements is the next step in this direction.

Scenarios for Conversational Search

This group aimed to identify scenarios that invite conversational search, given that natural language conversation might not always be the best way to search in some context. Their report summarises that modality and task of search are the two cases where conversational search might make sense. Modality can be determined by a situation such as driving or cooking, or devices at hand such as a smartwatch or AR/VR systems. As for the task, the group explains that the usefulness of conversational search increases as the level of exploration and complexity increases in tasks. On the other hand, simple information needs, highly ambiguous situations, or very social situations might not be the bast case for conversational search. Proposed scenarios include a mechanic fixing a machine, two people searching for a place for dinner, learning about a recent medical diagnosis, and following up on a news article to learn more.

Conversation Search for Learning Technologies

This group discussed the implication of conversational search from learning perspectives. The report highlights the importance of search technologies in lifelong learning and education, and the challenges due to complexity of learning processes. The group points out that multimodal interaction is particularly useful for educational and learning goals since it can support students with diverse background. Based on these discussions, the report suggests several research directions including extension of modalities to speech, writing, touch, gaze, and gesturing, integration of multimodal inputs/outputs with existing IR techniques, and application of multimodal signals to user modelling.

Common Conversational Community Prototype: Scholarly Conversational Assistant

This group proposed to develop and operate a prototype conversational search system for scholarly activities as academic resources that support research on conversational search. Example activities include finding articles for a new area of interest, planning sessions to attend in a conference, or determining conference PC members. The proposed prototype is expected to serve as a useful search tool, a means to create datasets, and a platform for community-based evaluation campaigns. The group outlined also a road map of the development of a Scholarly Conversational Assistant. The report includes a set of software platforms, scientific IR tools, open source conversational agents, and data collections that can be exploited in conversational search work.

Conclusions

Leading researchers from diverse domains in academia and industries investigated the essence, attributes, architecture, applications, challenges, and opportunities of Conversational Search in the seminar. One clear signal from the seminar is that research opportunities to advance Conversational Search are available to many areas and collaboration in an interdisciplinary community is essential to achieve the goal. This report should serve as one of the main sources to facilitate such diverse research programs on Conversational Search.

2 Table of Contents

Executive Summary
Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, Benno Stein . . 34

Overview of Talks

What Have We Learned about Information Seeking Conversations?
Nicholas J. Belkin 41

Conversational User Interfaces
Leigh Clark 41

Introduction to Dialogue
Phil Cohen 42

Towards an Immersive Wikipedia
Bernd Fröhlich 42

Conversational Style Alignment for Conversational Search
Ujwal Gadiraju 43

The Dilemma of the Direct Answer
Martin Potthast 43

A Theoretical Framework for Conversational Search
Filip Radlinski 44

Conversations about Preferences
Filip Radlinski 44

Conversational Question Answering over Knowledge Graphs
Rishiraj Saha Roy 45

Ranking People
Markus Strohmaier 45

Dynamic Composition for Domain Exploration Dialogues
Idan Szpektor 46

Introduction to Deep Learning in NLP
Idan Szpektor 46

Conversational Search in the Enterprise
Jaime Teevan 47

Demystifying Spoken Conversational Search
Johanne Trippas 47

Knowledge-based Conversational Search
Svitlana Vakulenko 47

Computational Argumentation
Henning Wachsmuth 48

Clarification in Conversational Search
Hamed Zamani 48

Macaw: A General Framework for Conversational Information Seeking
Hamed Zamani 49

Working groups

Defining Conversational Search
Jaime Arguello, Lawrence Cavedon, Jens Edlund, Matthias Hagen, David Maxwell, Martin Potthast, Filip Radlinski, Mark Sanderson, Laure Soulier, Benno Stein, Jaime Teevan, Johanne Trippas, and Hamed Zamani 49

Evaluating Conversational Search
Rishiraj Saha Roy, Avishek Anand, Jens Edlund, Norbert Fuhr, and Ujwal Gadiraju 55

Modeling Conversational Search
Elisabeth André, Nicholas J. Belkin, Phil Cohen, Arjen P. de Vries, Ronald M. Kaplan, Martin Potthast, and Johanne Trippas 61

Argumentation and Explanation
Khalid Al-Khatib, Ondrej Dusek, Benno Stein, Markus Strohmaier, Idan Szpektor, and Henning Wachsmuth 63

Scenarios that Invite Conversational Search
Lawrence Cavedon, Bernd Fröhlich, Hideo Joho, Ruihua Song, Jaime Teevan, Johanne Trippas, and Emine Yilmaz 65

Conversational Search for Learning Technologies
Sharon Oviatt and Laure Soulier 69

Common Conversational Community Prototype: Scholarly Conversational Assistant
Krisztian Balog, Lucie Flekova, Matthias Hagen, Rosie Jones, Martin Potthast, Filip Radlinski, Mark Sanderson, Svitlana Vakulenko, and Hamed Zamani 74

Recommended Reading List 80

Acknowledgements 82

Participants 83

3 Overview of Talks

3.1 What Have We Learned about Information Seeking Conversations?

Nicholas J. Belkin (Rutgers University – New Brunswick, US)

Main reference Nicholas J. Belkin, Helen M. Brooks, Penny J. Daniels: "Knowledge Elicitation Using Discourse Analysis", International Journal of Man-Machine Studies, Vol. 27(2), pp. 127–144, 1987.
URL http://dx.doi.org/10.1016/S0020-7373(87)80047-0

From the Point of View of Interactive Information Retrieval: What Have We Learned about Information Seeking Conversations, and How Can That Help Us Decide on the Goals of Conversational Search, and Identify Problems in Achieving Those Goals?

This presentation describes early research in understanding the characteristics of the information seeking interactions between people with information problems and human information intermediaries. Such research accomplished a number of results which I claim will be useful in the design of conversational search systems. It identified functions performed by intermediaries (and end users) in these interactions. These functions are aimed at constructing models of aspects of the user's problem and goals that are needed for identifying information objects that will be useful for achieving the goal which led the person to engage in information seeking. This line of research also developed formal models of such dialogues, which can be used for driving/structuring dialog-based information seeking. This research discovered a tension between explicit user modeling and user modeling through the participants' direct interactions with information objects, and relates that tension to both the nature and extent of interaction that's appropriate in such dialogues. Two examples of relevant research are [1] and [2]. On the basis of these results, some specific challenges to the design of conversational search systems are identified.

References

1 N. J. Belkin, H. M. Brooks, and P. J. Daniels. Knowledge Elicitation Using Discourse Analysis. *International Journal of Man-Machine Studies*, 27(2):127–144, 1987.
2 S. Sitter and A. Stein. Modelling the Illocutionary Aspects of Information-Seeking Dialogues. *Information Processing & Management*, 28(2):165–180, 1992.

3.2 Conversational User Interfaces

Leigh Clark (Swansea University, GB)

Joint work of Leigh Clark, Philip R. Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew P. Aylett, João P. Cabral, Cosmin Munteanu, Justin Edwards, Benjamin R. Cowan, Christine Murad, Nadia Pantidi, Orla, Cooney
Main reference Leigh Clark, Philip R. Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew P. Aylett, João P. Cabral, Cosmin Munteanu, Justin Edwards, Benjamin R. Cowan: "The State of Speech in HCI: Trends, Themes and Challenges", Interacting with Computers, Vol. 31(4), pp. 349–371, 2019.
URL http://dx.doi.org/10.1093/iwc/iwz016

Conversational User Interfaces (CUIs) are available at unprecedented levels though interactions with assistants in smart speakers, smartphones, vehicles and Internet of Things (IoT) appliances. Despite a good knowledge of the technical underpinnings of these systems, less is known about the user side of interaction – for instance how interface design choices impact

on user experience, attitudes, behaviours, and language use. This talk presents an overview of the work conducted on CUIs in the field of Human-Computer Interaction (HCI) and highlights from the 1st International Conference on Conversational User Interfaces (CUI 2019). In particular, I highlight aspects such as the need for more theory and method work in speech interface interaction, consideration of measures used to evaluated systems, an understanding of concepts like humanness, trust, and the need for understanding and possibly reframing the idea of conversation when it comes to speech-based HCI.

3.3 Introduction to Dialogue

Phil Cohen (Monash University – Clayton, AU)

This talk argues that future conversational systems that can engage in multi-party, collaborative dialogues will require a more fundamental approach than existing "intent + slot"-based systems. I identify significant limitations of the state of the art, and argue that returning to the plan-based approach o dialogue will provide a stronger foundation. Finally, I suggest a research strategy that couples neural network-based semantic parsing with plan-based reasoning in order to build a collaborative dialogue manager.

3.4 Towards an Immersive Wikipedia

Bernd Fröhlich (Bauhaus-Universität Weimar, DE)

Joint work of Bernd Fröhlich, Alexander Kulik, André Kunert, Stephan Beck, Volker Rodehorst, Benno Stein, Henning Schmidgen
Main reference Stephan Beck, André Kunert, Alexander Kulik, Bernd Froehlich: "Immersive Group-to-Group Telepresence", IEEE Trans. Vis. Comput. Graph., Vol. 19(4), pp. 616–625, 2013.
URL http://dx.doi.org/10.1109/TVCG.2013.33

It is our vision that the use of advanced Virtual and Augmented Reality (VR, AR) in combination with conversational technologies can take the access to knowledge to the next level. We are researching and developing procedures, methods and interfaces to enrich detailed digital 3D models of the real world with the complex knowledge available on the Internet, in libraries and through experts and make these multimodal models accessible in social VR and AR environments through natural language interfaces. Instead of isolated interaction with screens, there will be an immersive and collective experience in virtual space –, in a kind of walk-in Wikipedia – where knowledge can be accessed and acquired through the spatial presence of visitors, their gestures and conversational search.

References

1 Stephan Beck, André Kunert, Alexander Kulik, Bernd Froehlich: Immersive Group-to-Group Telepresence. *IEEE Transactions on Visualization and Computer Graphics*, 19(4): 616-625, 2013.

3.5 Conversational Style Alignment for Conversational Search

Ujwal Gadiraju (Leibniz Universität Hannover, DE)

Joint work of Sihang Qiu, Ujwal Gadiraju, Alessandro Bozzon
Main reference Panagiotis Mavridis, Owen Huang, Sihang Qiu, Ujwal Gadiraju, Alessandro Bozzon: "Chatterbox: Conversational Interfaces for Microtask Crowdsourcing", in Proc. of the 27th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2019, Larnaca, Cyprus, June 9-12, 2019, pp. 243–251, ACM, 2019.
URL http://dx.doi.org/10.1145/3320435.3320439
Main reference Sihang Qiu, Ujwal Gadiraju, Alessandro Bozzon: "Understanding Conversational Style in Conversational Microtask Crowdsourcing", 7th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2019), 2019.
URL https://www.humancomputation.com/assets/papers/130.pdf

Conversational interfaces have been argued to have advantages over traditional graphical user interfaces due to having a more human-like interaction. Owing to this, conversational interfaces are on the rise in various domains of our everyday life and show great potential to expand. Recent work in the HCI community has investigated the experiences of people using conversational agents, understanding user needs and user satisfaction. This talk builds on our recent findings in the realm of conversational microtasking to highlight the potential benefits of aligning conversational styles of agents with that of users. We found that conversational interfaces can be effective in engaging crowd workers completing different types of human-intelligence tasks (HITs), and a suitable conversational style has the potential to improve worker engagement. In our ongoing work, we are developing methods to accurately estimate the conversational styles of users and their style preferences from sparse conversational data in the context of microtask marketplaces.

3.6 The Dilemma of the Direct Answer

Martin Potthast (Universität Leipzig, DE)

A direct answer characterizes situations in which a potentially complex information need, expressed in the form of a question or query, is satisfied by a single answer–i.e., without requiring further interaction with the questioner. In web search, direct answers have been commonplace for years already, in the form of highlighted search results, rich snippets, and so-called "oneboxes" showing definitions and facts, thus relieving the users from browsing retrieved documents themselves. The recently introduced conversational search systems, due to their narrow, voice-only interfaces, usually do not even convey the existence of more answers beyond the first one.

Direct answers have been met with criticism, especially when the underlying AI fails spectacularly, but their convenience apparently outweighs their risks.

The dilemma of direct answers is that of trading off the chances of speed and convenience with the risks of errors and a reduced hypothesis space for decision making.

The talk will briefly introduce the dilemma by retracing the key search system innovations that gave rise to it.

3.7 A Theoretical Framework for Conversational Search

Filip Radlinski (Google UK – London, GB)

Joint work of Filip Radlinski, Nick Craswell
Main reference Filip Radlinski, Nick Craswell: "A Theoretical Framework for Conversational Search", in Proc. of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR 2017, Oslo, Norway, March 7-11, 2017, pp. 117–126, ACM, 2017.
URL http://dx.doi.org/10.1145/3020165.3020183

This talk presented a theory and model of information interaction in a chat setting. In particular, we consider the question of what properties would be desirable for a conversational information retrieval system so that the system can allow users to answer a variety of information needs in a natural and efficient manner. We study past work on human conversations, and propose a small set of properties that taken together could measure the extent to which a system is conversational.

3.8 Conversations about Preferences

Filip Radlinski (Google UK – London, GB)

Joint work of Filip Radlinski, Krisztian Balog, Bill Byrne, Karthik Krishnamoorthi
Main reference Filip Radlinski, Krisztian Balog, Bill Byrne, Karthik Krishnamoorthi: "Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences", Proc. of 20th Annual SIGdial Meeting on Discourse and Dialogue, pp. 353–360, 2019.
URL https://doi.org/10.18653/v1/W19-5941

Conversational recommendation has recently attracted significant attention. As systems must understand users' preferences, training them has called for conversational corpora, typically derived from task-oriented conversations. We observe that such corpora often do not reflect how people naturally describe preferences.

We present a new approach to obtaining user preferences in dialogue: Coached Conversational Preference Elicitation. It allows collection of natural yet structured conversational preferences. Studying the dialogues in one domain, we present a brief quantitative analysis of how people describe movie preferences at scale. Demonstrating the methodology, we release the CCPE-M dataset to the community with over 500 movie preference dialogues expressing over 10,000 preferences.

3.9 Conversational Question Answering over Knowledge Graphs

Rishiraj Saha Roy (MPI für Informatik – Saarbrücken, DE)

Joint work of Philipp Christmann, Abdalghani Abujabal, Jyotsna Singh, Gerhard Weikum
Main reference Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, Gerhard Weikum: "Look before you Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion", in Proc. of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019, pp. 729–738, ACM, 2019.
URL http://dx.doi.org/10.1145/3357384.3358016

Fact-centric information needs are rarely one-shot; users typically ask follow-up questions to explore a topic. In such a conversational setting, the user's inputs are often incomplete, with entities or predicates left out, and ungrammatical phrases. This poses a huge challenge to question answering (QA) systems that typically rely on cues in full-fledged interrogative sentences. As a solution, in this project, we develop CONVEX: an unsupervised method that can answer incomplete questions over a knowledge graph (KG) by maintaining conversation context using entities and predicates seen so far and automatically inferring missing or ambiguous pieces for follow-up questions. The core of our method is a graph exploration algorithm that judiciously expands a frontier to find candidate answers for the current question. To evaluate CONVEX, we release ConvQuestions, a crowdsourced benchmark with 11,200 distinct conversations from five different domains. We show that CONVEX: (i) adds conversational support to any stand-alone QA system, and (ii) outperforms state-of-the-art baselines and question completion strategies.

3.10 Ranking People

Markus Strohmaier (RWTH Aachen, DE)

The popularity of search on the World Wide Web is a testament to the broad impact of the work done by the information retrieval community over the last decades. The advances achieved by this community have not only made the World Wide Web more accessible, they have also made it appealing to consider the application of ranking algorithms to other domains, beyond the ranking of documents. One of the most interesting examples is the domain of ranking people. In this talk, I highlight some of the many challenges that come with deploying ranking algorithms to individuals. I then show how mechanisms that are perfectly fine to utilize when ranking documents can have undesired or even detrimental effects when ranking people. This talk intends to stimulate a discussion on the manifold, interdisciplinary challenges around the increasing adoption of ranking algorithms in computational social systems. This talk is a short version of a keynote given at ECIR 2019 in Cologne.

3.11 Dynamic Composition for Domain Exploration Dialogues

Idan Szpektor (Google Israel – Tel-Aviv, IL)

We study conversational exploration and discovery, where the user's goal is to enrich her knowledge of a given domain by conversing with an informative bot. We introduce a novel approach termed dynamic composition, which decouples candidate content generation from the flexible composition of bot responses. This allows the bot to control the source, correctness and quality of the offered content, while achieving flexibility via a dialogue manager that selects the most appropriate contents in a compositional manner.

3.12 Introduction to Deep Learning in NLP

Idan Szpektor (Google Israel – Tel-Aviv, IL)

Joint work of Idan Szpektor, Ido Dagan

We introduced the current trends in deep learning for NLP, including contextual embedding, attention and self-attention, hierarchical models, common task-specific architectures (seq2seq, sequence tagging, Siamese towers) and training approaches, including multitasking and masking. We deep dived on modern models such as the Transformer and BERT and discussed how they are being evaluated.

References

1 Schuster and Paliwal. 1997. Bidirectional Recurrent Neural Networks.
2 Bahdanau et al. 2015. Neural machine translation by jointly learning to align and translate.
3 Lample et al. 2016. Neural Architectures for Named Entity Recognition.
4 Serban et al. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models.
5 Das et al. 2016. Together We Stand: Siamese Networks for Similar Question Retrieval.
6 Vaswani et al. 2017. Attention Is All You Need.
7 Devlin et al. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
8 Yang et al. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding.
9 Lan et al. 2019: ALBERT: a lite BERT for self-supervised learning of language representations.
10 Zhang et al. 2019. HIBERT: document level pre-training of hierarchical bidirectional transformers for document summarization.
11 Peters et al. 2019. To tune or not to tune? Adapting pretrained representations to diverse tasks.
12 Tenney et al. 2019. BERT Rediscovers the Classical NLP Pipeline.
13 Liu et al. 2019. Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets.

3.13 Conversational Search in the Enterprise

Jaime Teevan (Microsoft Corporation – Redmond, US)

As a research community we tend to think about conversational search from a consumer point of view; we study how web search engines might become increasingly conversational, and think about how conversational agents might do more than just fall back to search when they don't know how else to address an utterance. In this talk I challenge us to also look at conversational search in productivity contexts, and highlight some of the unique research challenges that arise when we take an enterprise point of view.

3.14 Demystifying Spoken Conversational Search

Johanne Trippas (RMIT University – Melbourne, AU)

Joint work of Johanne Trippas, Damiano Spina, Lawrence Cavedon, Mark Sanderson, Hideo Joho, Paul Thomas

Speech-based web search where no keyboard or screens are available to present search engine results is becoming ubiquitous, mainly through the use of mobile devices and intelligent assistants. They do not track context or present information suitable for an audio-only channel, and do not interact with the user in a multi-turn conversation. Understanding how users would interact with such an audio-only interaction system in multi-turn information-seeking dialogues, and what users expect from these new systems, are unexplored in search settings. In this talk, we present a framework on how to study this emerging technology through quantitative and qualitative research designs, outline design recommendations for spoken conversational search, and summarise new research directions [1, 2].

References

1 J. R. Trippas. *Spoken Conversational Search: Audio-only Interactive Information Retrieval. PhD thesis*, RMIT, Melbourne, 2019.
2 J. R. Trippas, D. Spina, P. Thomas, H. Joho, M. Sanderson, and L. Cavedon. Towards a model for spoken conversational search. *Information Processing & Management*, 57(2):1–19, 2020.

3.15 Knowledge-based Conversational Search

Svitlana Vakulenko (Wirtschaftsuniversität Wien, AT)

Joint work of Svitlana Vakulenko, Axel Polleres, Maarten de Reijke

Conversational interfaces that allow for intuitive and comprehensive access to digitally stored information remain an ambitious goal. In this thesis, we lay foundations for designing conversational search systems by analyzing the requirements and proposing concrete solutions for automating some of the basic components and tasks that such systems should support. We describe several interdependent studies that were conducted to analyse the design

requirements for more advanced conversational search systems able to support complex human-like dialogue interactions and provide access to vast knowledge repositories. Our results show that question answering is one of the key components required for efficient information access but it is not the only type of dialogue interactions that a conversational search system should support [1].

References

1 Svitlana Vakulenko. *Knowledge-based Conversational Search. PhD thesis*, TU Wien, 2019.

3.16 Computational Argumentation

Henning Wachsmuth (Universität Paderborn, DE)

Argumentation is pervasive, from politics to the media, from everyday work to private life. Whenever we seek to persuade others, to agree with them, or to deliberate on a stance towards a controversial issue, we use arguments. Due to the importance of arguments for opinion formation and decision making, their computational analysis and synthesis is on the rise in the last five years, usually referred to as *computational argumentation*. Major tasks include the mining of arguments from natural language text, the assessment of their quality, and the generation of new arguments and argumentative texts. Building on fundamentals of argumentation theory, this talk gives a brief overview of techniques and applications of computational argumentation and their relation to conversational search. Insights are given into our research around args.me, the first search engine for arguments on the web [1].

References

1 Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59, 2017.

3.17 Clarification in Conversational Search

Hamed Zamani (Microsoft Corporation, US)

Joint work of Hamed Zamani, Susan T. Dumais, Nick Craswell, Paul Bennett, Gord Lueck

Search queries are often short, and the underlying user intent may be ambiguous. This makes it challenging for search engines to predict possible intents, only one of which may pertain to the current user. To address this issue, search engines often diversify the result list and present documents relevant to multiple intents of the query. However, this solution cannot be applied to scenarios with "limited bandwidth" interfaces, such as conversational search systems with voice-only and small-screen devices. In this talk, I highlight clarifying question generation and evaluation as two major research problems in the area and discuss possible solutions for them.

3.18 Macaw: A General Framework for Conversational Information Seeking

Hamed Zamani (Microsoft Corporation, US)

Joint work of Hamed Zamani, Nick Craswell
Main reference Hamed Zamani, Nick Craswell: "Macaw: An Extensible Conversational Information Seeking Platform", CoRR, Vol. abs/1912.08904, 2019.
URL http://arxiv.org/abs/1912.08904

Conversational information seeking (CIS) has been recognized as a major emerging research area in information retrieval. Such research will require data and tools, to allow the implementation and study of conversational systems. In this talk, I introduce Macaw, an open-source framework with a modular architecture for CIS research. Macaw supports multi-turn, multi-modal, and mixed-initiative interactions, for tasks such as document retrieval, question answering, recommendation, and structured data exploration. It has a modular design to encourage the study of new CIS algorithms, which can be evaluated in batch mode. It can also integrate with a user interface, which allows user studies and data collection in an interactive mode, where the back end can be fully algorithmic or a wizard of oz setup.

4 Working groups

4.1 Defining Conversational Search

Jaime Arguello (University of North Carolina – Chapel Hill, US), Lawrence Cavedon (RMIT University – Melbourne, AU), Jens Edlund (KTH Royal Institute of Technology – Stockholm, SE), Matthias Hagen (Martin-Luther-Universität Halle-Wittenberg, DE), David Maxwell (University of Glasgow, GB), Martin Potthast (Universität Leipzig, DE), Filip Radlinski (Google UK – London, GB), Mark Sanderson (RMIT University – Melbourne, AU), Laure Soulier (UPMC – Paris, FR), Benno Stein (Bauhaus-Universität Weimar, DE), Jaime Teevan (Microsoft Corporation – Redmond, US), Johanne Trippas (RMIT University – Melbourne, AU), and Hamed Zamani (Microsoft Corporation, US)

4.1.1 Description and Motivation

As the theme of this Dagstuhl seminar, it appears essential to define conversational search to scope the seminar and this report. With the broad range of researchers present at the seminar, it quickly became clear that it is not possible to reach consensus on a formal definition. Similarly to the situation in the broad field of information retrieval, we recognize that there are many possible characterizations. This breakout group thus aimed to bring structure and common terminology to the different aspects of conversational search systems that characterize the field. It additionally attempts to take inventory of current definitions in the literature, allowing for a fresh look at the broad landscape of conversational search systems, as well as their desired and distinguishing properties.

4.1.2 Existing Definitions

Conversational Answer Retrieval

Current IR systems provide ranked lists of documents in response to a wide range of keyword queries with little restriction on the domain or topic. Current question answering (Q/A) systems, on the other hand, provide more specific answers to a very limited range of natural language questions. Both types of systems use some form of limited dialogue to refine queries and answers. The aim of conversational is to combine the advantages of these two approaches to provide effective retrieval of appropriate answers to a wide range of questions expressed in natural language, with rich user-system dialogue as a crucial component for understanding the question and refining the answers. We call this new area conversational answer retrieval. The dialogue in the CAR system should be primarily natural language although actions such as pointing and clicking would also be useful. Dialogue would be initiated by the searcher and proactively by the system. The dialogue would be about questions and answers, with the aim of refining the understanding of questions and improving the quality of answers. Previous parts of the dialogue, such as previous questions or answers, should be able to be referred to in the dialogue, also with the aim of refining and understanding. Dialogue, in other words, should be used to fill the inevitable gaps in the system's knowledge about possible question types and answers [1].

Conversational Information Seeking

Conversational Information Seeking (CIS) is concerned with a task-oriented sequence of exchanges between one or more users and an information system. This encompasses user goals that include complex information seeking and exploratory information gathering, including multi-step task completion and recommendation. Moreover, CIS focuses on dialog settings with various communication channels, such as where a screen or keyboard may be inconvenient or unavailable. Building on extensive recent progress in dialog systems, we distinguish CIS from traditional search systems as including capabilities such as long term user state (including tasks that may be continued or repeated with or without variation), taking into account user needs beyond topical relevance (how things are presented in addition to what is presented), and permitting initiative to be taken by either the user or the system at different points of time. As information is presented, requested or clarified by either the user or the system, the narrow channel assumption also means that CIS must address issues including presenting information provenance, user trust, federation between structured and unstructured data sources and summarization of potentially long or complex answers in easily consumable units [2].

Radlinski and Craswell [4] define a conversational search system as a system for retrieving information that permits a mixed-initiative back and forth between a user and agent, where the agent's actions are chosen in response to a model of current user needs within the current conversation, using both short- and long-term knowledge of the user. Further, they argue that such a system can be characterized as having five key properties. The first two characterize learning, specifically user revealment (that is, the system assisting the user to learn about their actual need) and system revealment (that is, the system allowing the user to learn about the system's abilities). The remaining three refer to functionality: Supporting the mixed-initiative, possessing memory (including the ability for the user to reference past conversational steps), and the ability for it to reason about sets of items [4].

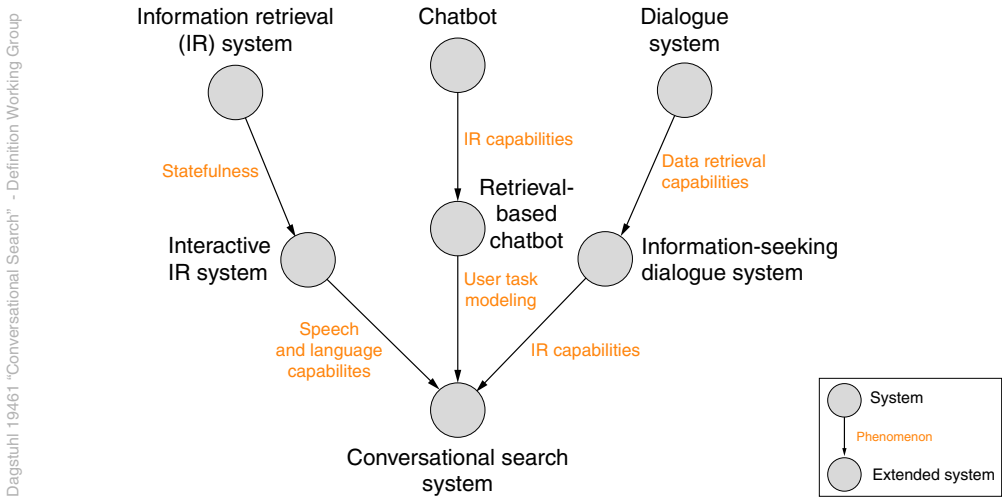

Figure 1 The Dagstuhl Typology of Conversational Search defines conversational search systems via functional extensions of information retrieval systems, chatbots, and dialogue systems.

Vakulenko [7] define conversational search as a task of retrieving relevant information using a conversational interface, where a conversation is understood as a sequence of natural language expressions (utterances) made by several conversation participants in turns [7].

Trippas [6] define a spoken conversational system (SCS) as a broad term for any system which enables users to interact over speech (i.e., voice) in a conversational manner. Likewise she defines *spoken* conversational search as a process concerning open domain multi-turn verbal natural language exchanges between the user(s) and the system. They refine the requirements of SCS systems as follows: An SCS system supports the users' input which can include multiple actions in one utterance and is more semantically complex. Moreover, the SCS system helps users navigate an information space and can overcome standstill-conversations due to communication breakdown by including meta-communication as part of the interactions. Ultimately, the SCS multi-turn exchanges are mixed-initiative, meaning that systems also can take action or drive the conversation. The system also keeps track of the context of individual questions, ensuring a natural flow to the conversation (i.e., no need to repeat previous statements). Thus the user's information need can be expressed, formalized, or elicited through natural language conversational interactions [6].

4.1.3 The Dagstuhl Typology of Conversational Search

In this definition, we derive conversational search systems from well-known and widely studied notions of systems from related research fields. Figure 1 shows "The Dagstuhl Typology of Conversational Search" (the conversational Ψ).

Usage

The typology captures the diversity of systems that can be expected from the conflation of the two research fields most related to conversational search, information retrieval, and dialogue systems. Dependent on the base system on which a conversational search system is built, and consequently the background of its makers, the following statements can be made:

1. An interactive information retrieval system with speech and language capabilities is a conversational search system.

2. A retrieval-based chatbot that models a user's tasks is a conversational search system.
3. An information-seeking dialogue system with information retrieval capabilities is a conversational search system.

These statements are useful when existing systems are to be classified. More often, however, the term "conversational search (system)" needs to be defined. But simply reversing one of the above statements would exclude the other alternatives. We hence recommend to write something like this:

- A conversational search system can be based on ...
- Our conversational search system is based on ...
- We build our conversational search system based on ...

If a fully-fledged written definition is desired (e.g., as an opening statement for a related work section), and there is no room to include the above figure, the following can be used:

> *A conversational search system is either an interactive information retrieval system with speech and language processing capabilities, a retrieval-based chatbot with user task modeling, or an information-seeking dialogue system with information retrieval capabilities.*

All of the above, including Figure 1, are free to be reused.

Background

Clearly, the number and kinds of properties that can be distinguished in a real-world instance of any of the aforementioned systems are manifold as well as overlapping. The purpose of this definition is neither to capture every last aspect nor to perfectly separate every conceivable instance of each of the aforementioned systems, but rather to outline the most salient differences that, in the eye of a domain expert, help to structure the space of possible systems. In particular, this definition serves as a straightforward way to teach students making their first steps in information retrieval or dialogue system in general, and conversational search in particular, since this definition is much easier to be recollected compared to lists of must-have and can-have properties.

4.1.4 Dimensions of Conversational Search Systems

We consider important dimensions of conversational search systems and relate them to "classical" IR systems (see Figures 2 and 3). To these dimensions belong among others the interactivity level, the state of the search session, the engagement of the user, and the engagement of the system (partly inspired by [5]).

- User intent/engagement towards the conversation: This dimension measures the level and the form of the conversation engaged by the user. For instance, a low engagement would be characterized by a behavior in which the user is only focused on his information need without awareness of the system understanding (or at least its ability to understand). On the contrary, a high engagement from the user would lead to clarification and sense-making exchange to be sure being understandable for the system, maximizing the task achievement. This dimension is correlated to the user's awareness of system abilities.
- System engagement: This dimension is system-centered and allows to distinguish the interaction way of systems. It ranges from passive systems that only aim to acting as users required (e.g., retrieving documents from a user query, whether contextualized or not) to pro-active systems that aim at maximizing and anticipating the task achievement

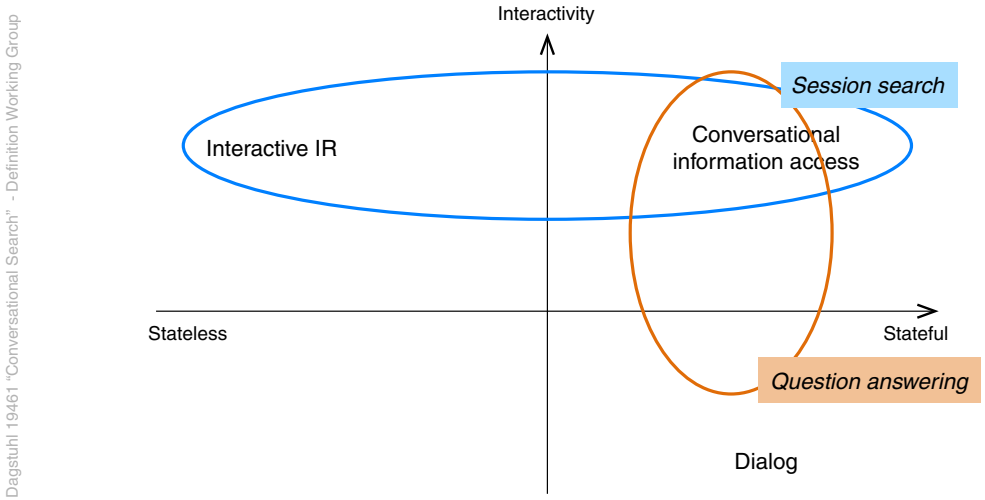

Figure 2 Dimensions of conversational search systems and their relation to "classical" IR systems (Part I).

and the user satisfaction. The system proactivity engenders a total awareness from the system side of users' actions and search directions to identify any drift or anticipate useless actions.

- Concurrency: This dimension expresses the temporal span of a conversation (immediate or delayed). In conversational search, the user expects an immediate response but the task achievement might be delayed due to the sense-making process.
- Usage of information: The information flow between a user and a system will vary depending on the objective. We distinguish information exchange/supply in which the process is only focused on answering a question (as in a Q/A setting or chit-chat bots) from sense-making process in which both users and systems are engaged in a cooperation with the objective to satisfy a goal (as in search-oriented conversational systems).
- Interaction naturalness: This dimension considers the way of communication. We distinguish interactions driven by structured language (e.g., keywords in classic IR) from interactions in natural language (as in conversational systems) for which the system has to figure out the intention with an intermediary level of language understanding.
- Statefulness: This dimension is it related to system/user engagement and the notion of awareness.
- Interactivity level: This dimension related to the number and the type of interactions as well as the interaction mode.

Desirable Additional Properties

From our point of view, there exists a set of properties that ideal conversational search systems are expected to have:

- User revealment: The system helps the user express (potentially discover) their true information need, and possibly also long-term preferences [4].
- System revealment: The system reveals to the user its capabilities and corpus, building the user's expectations of what it can and cannot do [4].
- Mixed initiative (be able to take dialogue and/or task control): Horvitz defined mixed-initiative interaction as a flexible interaction strategy in which each agent (human or computer) contributes what it is best suited at the most appropriate time [3]. Mixed

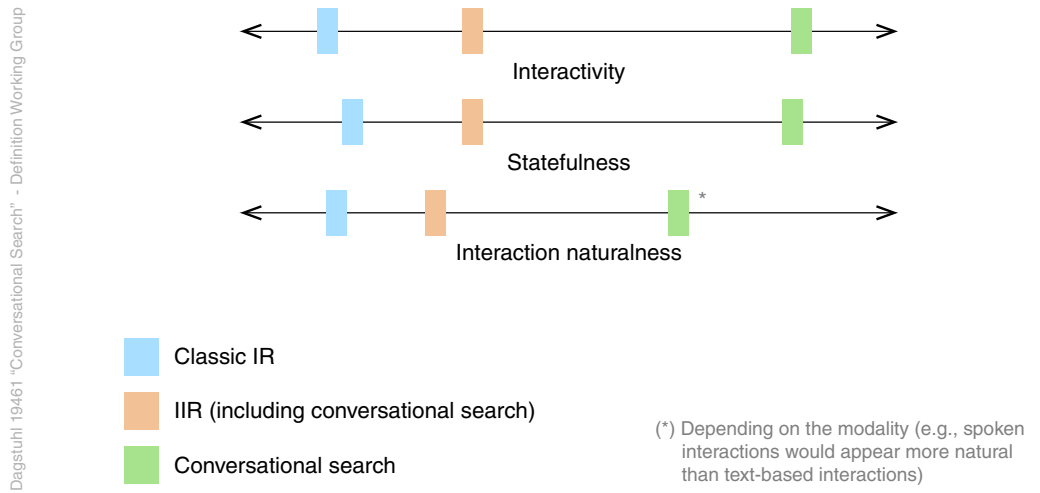

Figure 3 Dimensions of conversational search systems and their relation to "classical" IR systems (Part II).

initiative systems can take control of the communication either at the dialogue level (e.g., by asking for clarification or requesting elaboration) or at the task level (e.g., by suggesting alternative courses of action).

- Memory of interactions (indexing and access to history): The user can reference past statements, which implicitly also remain true unless contradicted [4].
- Recovering from communication breakdowns: A conversational search system can recover from communication breakdowns and ambiguity by asking clarification. Clarification can be simply in the form of "asking for repeat" or more advanced and intelligent form of clarification (e.g., "asking for disambiguation and explanation").
- Representation generation: Conversational search systems should be able to generate new (and useful) representations that are shared between a user and system. These may include new commands and/or shortcuts that are derived from action/reaction pairs present in past interactions.
- Multimodality: Conversational search systems may involve multiple modalities in terms of input (e.g., touchscreen, gesture-based, spoken dialogue) and output (visual, spoken dialogue). Multimodal output may be valuable for the system to elicit information in the context of an information item.
- Speech: Conversational search system may involve speech-based input and output, but may also support text-based input and output.
- Reasoning about sets and shortlists: Conversational search systems may benefit from the ability to inquire about characteristics of sets of potentially relevant items. Reasoning about sets includes inferring common attributes along which the sets can be differentiated and/or prioritized.
- Analyzing conversations for support (synchronously or asynchronously): Conversational search systems may include systems that can analyze human-human conversations and intervene to provide contextually relevant information.
- Understanding and reasoning about user limitations (speech is a particularly revealing modality): Dialogue is a means of communication that may allow a system to infer more information about a specific user (e.g., cognitive abilities and styles, domain knowledge). In turn, gaining insights about users may help systems to provide more personalized information and interactions.

Other Types of Systems that are not Conversational Search

We also chose to define conversational search systems by what explicating they are not. In particular, we discussed types of systems that may involve conversation but themselves are not conversational search:

- Systems that facilitate conversations between people (by eavesdropping and providing relevant information)
- Collaborative conversational search systems (multiple searchers)
- Speech-based Q/A systems
- Searching conversational corpora
- PIM conversational search
- Conversational access to structured data sources
- IBM Project Debater

References

1 James Allan, Bruce Croft, Alistair Moffat, and Mark Sanderson (eds.), Frontiers, Challenges, and Opportunities for Information Retrieval: *Report from SWIRL 2012. SIGIR Forum*, 46(1):2-32, 2012.

2 J. Shane Culpepper, Fernando Diaz, Mark D. Smucker (eds), Research Frontiers in Information Retrieval: *Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018)*, SIGIR Forum 51(1):34-90, 2018.

3 Eric Horvitz, Principles of Mixed-Initiative User Interfaces, *SIGCHI conference on Human Factors in Computing Systems*, 1999.

4 Radlinski, F. and Craswell, N. A Theoretical Framework for Conversational Search, *CHIIR* 2017.

5 Chirag Shah. Collaborative information seeking, *Journal of the Association for Information Science and Technology* 65(2):215-236, 2014.

6 Johanne R. Trippas. Spoken Conversational Search: Audio-only Interactive Information Retrieval. *RMIT University*, 2019.

7 Svitlana Vakulenko. Knowledge-based Conversational Search. *PhD thesis*. TU Wien. 2019.

4.2 Evaluating Conversational Search

Rishiraj Saha Roy (MPI für Informatik – Saarbrücken, DE), Avishek Anand (Leibniz Universität Hannover, DE), Jens Edlund (KTH Royal Institute of Technology – Stockholm, SE), Norbert Fuhr (Universität Duisburg-Essen, DE), and Ujwal Gadiraju (Leibniz Universität Hannover, DE)

4.2.1 Introduction

A key challenge for conversational search is in determining the quality of the search and/or system, and whether one search/system is better than another. So, what makes a good conversational search (CS)? And what makes a good conversational search system (CSS)? This is an open challenge.

Let's consider the following example where a user (U) interacts with a conversational search system (S):

- S: Hi, K, how can I help you?
- U: I would like to buy some running shoes.

The system may respond in a variety of ways depending on how well it has understood the request, or depending on the system's affordances.

- S1: OK, so you would like to buy funny shoes.
- S2: OK, so you would like to buy running shoes.
- S3: Great, what did you have in mind?
- S4: There are lots of different types of running shoes out there–are you interested in running shoes for cross fitness, road or trail?

S1-S4 are only a handful of possible responses. Here, S1 has misinterpreted the user's request. S2 appears to have interpreted the user's request correctly, and provides the user with confirmation–and could be followed by S3, S4 or some follow up question or response (i.e. listing shoes, etc.). S3 acknowledges the request and asks a open-ended follow up question, while S4 acknowledges the request and selects a possible facet (type of shoe) that may help in directing the conversation.

Clearly, S1 is not desirable and similarly other errors in communication and intent are not either. However, things become more complicated when considering the other possible responses. S2 elongates the conversation by providing a confirmation, while, S3 acknowledges, but assumes the intent. And S4, provides confirmation while drilling into a particular aspect. So which direction should the conversation take, and what would lead to resolving the conversational search in the most effective, efficient, experiential, etc. manner [1]?

A key challenge will be in balancing the trade-off between topic explorations and topic exploitation i.e. finding information directly useful for the task at hand versus finding information about the topic and domain in general [1].

4.2.2 Why would users engage in conversational search?

An important consideration in both the design and evaluation of conversational search is to understand users' goals for engaging with a conversational search system. As with other IIR and HCI evaluation, understanding users' goals and the context of their use is a very important aspect of designing appropriate evaluations.

First, the user's broader work task and information seeking should be considered. Information seekers make choices about the types of information interactions and information systems they interact with in order to try to satisfy their information needs. Thus, an important question for CSS is to consider *why* users might choose to engage with a conversational search system rather than some other information source or system (e.g., a web search engine, a book, talking to a colleague or friend, etc.).

CSS differs from traditional query-response retrieval systems (e.g., search engines) in several important ways. In a traditional SE interaction, the user controls the process, issuing queries to the system and scanning/selecting which items on the SERP to attend to, and in what order. When using a SE, users have a lot of control (initiative) in the interaction between user and system.

However, in a CSS, users relinquish some of this control in exchange for some other perceived benefit. The CSS interaction is likely to involve a more mixed-initiative style of interaction, which implies different possibilities and expectations from the user about the type of interaction which will occur (as opposed to the query-response paradigm of SEs).

Thus, we can ask, what perceived benefits or differences in interaction a user might expect by engaging with a CSS? This impacts how we evaluation overall success of a CSS, user satisfaction, and even component-level evaluation.

People choose to engage in human-to-human information seeking conversations for a variety of reasons, including to get guidance, seek advice, to consult an expert, to get a summary or synthesis of complex topics, and to get information from a trusted authority (among others). It seems reasonable that information seekers may have similar expectations for engaging with a conversational search system.

There may be other reasons for engaging with a CSS. For example, users may be engaged in a primary task and need information in a hands-busy and/or eyes-busy situation (e.g.., while cooking, driving, walking, performing a complex task such as fixing a dishwasher), and are able to engage with a CSS through speech.

Another area where CSS may be of benefit is in the context of searching to learn about a topic–where the user may learn more about the topic through a narrative i.e. conversational search as learning.

Conversational search may also be useful to assist conversations between two or more users. This may be to query a specific talking point in interaction (e.g. multi-user talk in a pub or cafe [5]) or engaging with a system that is embedded in the social interaction between users (e.g. searching for an interactive group game with an intelligent personal assistant [4]).

4.2.3 Broader Tasks, Scenarios, & User Goals

The goals of engaging in conversational search can be broadly categorised, but not necessarily limited to, the five areas described below. These categories may overlap in definition, and interactions may include several different categories as the interaction unfolds.

Sequential topic-based questions: A sequence of user-directed questions that are focused on a specific topic, with the subsequent questions emerging from the initial query and engagement with the conversational system.

- U: What are some good running shoes?
- S: . . .
- U: Tell me about the Nike Pegasus shoes?
- S: . . .
- U: How much are they?

Learning about a topic: A less-directed or possibly undirected exploration of a topic initiated by a user can lead to a conversational "search as learning" task. And so depending on the user's level of expertise the starting query will vary from broad to specific, and the expectation is that through the conversation the user will learn more about the topic.

- U: Tell me about different styles of running shoes.
- S: . . .
- U: What kinds of injuries do runners get?

Seeking Advice or guidance: Another scenario may involve learning more specifically about a topic to glean advice that is personally relevant to the information seeker. Using the above examples, this may be to query such things as product differences, comparing items, diagnosing a problem, resolving an issue, etc.

- U: What are the main differences between road and trail shoes?
- U: How can I improve my running style to avoid ankle pain?

Planning an Activity: A more task oriented but potentially less directed scenario arises in the case of planning activities where a user may have something in mind, or whether they need to explore the space of possibilities.

- U: OK, I'd like to go running this weekend.
- U: I'm travelling to Dagstuhl and like to know where I can go running.

Making a Decision: More transactional in nature are scenarios where the user engages the CSS in order to make a specific decision such as purchasing products, voting, etc. where a decision results in a transaction.

- U: I'd like to find a pair of good running shoes?

4.2.4 Existing Tasks and Datasets

Several tasks have been proposed as important milestones towards the goal of conversational search. They each were designed to solve a particular sub-problem of conversational search, though it may also be argued that some exist in their current form because we have large-scale data sources available and we are able to provide clear-cut evaluations for them. While it is difficult to properly evaluate a conversational search system end-to-end, particular sub-components can be evaluated by reporting precision, recall, accuracy and other similarly easy-to-compute metrics. Let's now look at existing tasks and datasets.

Conversation response ranking (e.g., [8]): Here, the problem of a conversational system responding to a user utterance is formulated as a retrieval problem. Given a conversation up to a particular user utterance, rank a given set of potential responses. Typically between 5-50 potential responses are provided and test collections are designed in a way that the correct response (there is assumed to be just one) is part of the potential response set. While this setup allows us to experiment and design a range of retrieval algorithms, the setup is artificial: (i) in an actual conversational search system there is no guarantee that a correct response exists in the historical corpus of conversations, (ii) more than one possible/accurate responses may exist (as seen in the initial example of this section), and, (iii) ranking potentially hundreds of millions of historic responses in a meaningful manner is beyond our current ranking capabilities (and thus the preselection of a handful of responses to rank).

Dialogue act prediction (e.g., [6]): Given an utterance of an information-seeking conversation, we are here interested in labeling it with a particular dialogue act label (specific to conversational search) such as Clarifying-Question, Further-Details, Potential-Answer and so on. It is to some extent an open question how this information can then be employed in the conversational search pipeline.

Next question prediction (e.g., [9]): This task is set up to predict the next user question, and is setup/evaluated in a similar manner to conversation response ranking. Thus, a similar critical point remains: we need a more realistic evaluation setup.

Sub-goals prediction (e.g., [3]): This task is also known as task understanding: given a user query (the task to complete), the system predicts the set of sub-goals/sub-tasks that are required to complete the task.

Sequential question answering (e.g., [2]): Here, instead of the standard question answering task (each question is treated separately), we are interested in answering a series of interrelated questions (e.g. Q1: What are the best running shoes? Q2: Where can I buy them? Q3: How much are they?).

While the creation of datasets and benchmarks is a fruitful avenue of research/publication in the NLP/DS communities, the IR community has been less receptive and thus many conversational datasets are proposed elsewhere. We note here that many of the currently existing corpora for CSS are based on human-to-human conversations. However, this includes much knowledge that is outside the current scope of retrieval systems. As human-to-human conversations differ from human-to-machine conversations it is an open question to what

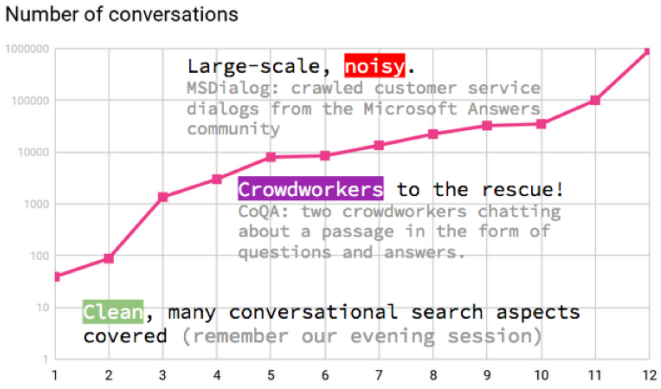

Figure 4 Overview of the dataset sizes of 12 recently introduced conversational datasets that are multi-turn, non-chit-chat and human-to-human.

extent corpora of human-to-human conversations are our best option to train conversational search systems. We argue that (at least in the near future) we should optimize conversational search systems based on human-machine conversations that are grounded in current retrieval systems and technologies (one instantiation of how to collect such a dataset can be found in Trippas et al. [7]).

A particular challenge of conversational search datasets is to meaningfully collect and build large-scale datasets (required for neural net-based training regimes). Consider Figure 4 where we plot the number of conversations across 12 recently introduced conversational datasets (such as MSDialog, UDC, CoQA, Frames, SCS and others). Even the largest dataset has fewer than a million conversations, while the smallest ones have fewer than 100 conversations. Importantly, the larger datasets are usually crawls of large fora (e.g. Stack Overflow or other technical fora) with little to no additional labelling to enable a range of conversational tasks. At the other end of the spectrum we have very small, but also very clean and well-annotated datasets that are very useful to analyze conversations but not sufficient to train today's machine learning algorithms.

4.2.5 Measuring Conversational Searches and Systems

In Figure 5, we have enumerated a number of different dimensions in which we may wish to evaluate CS/CSS by, whether they are mainly user-focused, retrieval-focused or dialogue-focused. Lab-based and A/B testing will typically involve a complete (or simulated) system setup and thus facilitate end-to-end (e2e) evaluation. However, given the highly interactive nature of CS it is unlikely that a reusable test collection will be able to be developed to support any serious e2e evaluations–test collections should be able to support component level evaluation.

Ideally, the measures used should scale. That is, if the measure is used at the component level, then it should inform as to how that measure would change the e2e experience.

Note that in the table ticks indicate that this measure can be done using test collection, lab-based or A/B testing, while indicates that it might be possible or could be done via a proxy.

The different dimensions suggest that many trade-offs are likely to arise during the conversational search. For example, higher effort may be indicative of a poor CS experience, but could equally be indicative of a good conversational search experience – as it depends on how much the user gains from the experience in terms of how much they learn about the

		Test Collection	Lab-Based	A/B
		Comp / (~E2E)	E2E / Comp	E2E / Comp
User	Credibility / Trust		✓	
User	Cognitive Load		✓	~
User	Engagement		✓	~
User	Satisfaction		✓	~
User	Info gain (w.r.t. task & domain)		✓	
User	Information gain (about system)	~	✓	✓
User	Effort / Time	✓	✓	✓
User	Success / Decision / Outcome	~	✓	
User/Retrieval	Utility / Usefulness	~	✓	~
Retrieval/Dialog	Information gain (about user)	~	✓	✓
Retrieval/Dialog	Robustness / Error Recovery	✓	✓	✓
Retrieval	Completeness / Coverage	✓	✓	

Figure 5 A summary of evaluation criteria and evaluation methodologies for component-based and/or end-to-end evaluation of conversational search systems.

topic, the domain (and the search space) and the system (and it's affordances). However, for longer term measures such as trust, it is dependent on the cumulative experiences and the successes/decisions/outcomes that result from the conversations. For example, if K buys the Nike's but finds them later for a lower price, or buys them and finds out that they are not as comfortable as described–then they may be be subsequently unhappy, and thus have less trust in the system.

From Figure 5, it is clear that the measures are not different from those used in interactive information retrieval – however, depending on the form of conversational search, certain dimensions are likely to be more important than others.

References

1 Leif Azzopardi, Mateusz Dubiel, Martin Halvey, and Jffery Dalton. Conceptualizing agent-human interactions during the conversational search process. In *Second International Workshop on Conversational Approaches to Information Retrieval*, 2018.

2 Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), page 1821–1831, Vancouver, Canada, 2017. ACL.

3 Evangelos Kanoulas, Emine Yilmaz, Rishabh Mehrotra, Ben Carterette, Nick Craswell, and Peter Bailey. Trec 2017 tasks track overview. In *Text REtrieval Conference*, 2017.

4 Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. Voice interfaces in everyday life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, New York, NY, USA, 2018. ACM.

5 Martin Porcheron, Joel E. Fischer, and Sarah Sharples. Do animals have accents?": Talking with agents in multi-party conversation. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 207–219, NewYork, NY, USA, 2017. ACM.

6 Chen Qu, Liu Yang, W. Bruce Croft, Yongfeng Zhang, Johanne R. Trippas, and Minghui Qiu. User intent prediction in information-seeking conversations. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, CHIIR '19, page 25–33, NewYork, NY, USA, 2019. ACM.
7 Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. Informing the design of spoken conversational search: Perspective paper. *CHIIR '18*, page 32–41, New York, NY, USA, 2018. ACM.
8 Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W. Bruce Croft, Jun Huang, and Haiqing Chen. Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. In The *41st International ACM SIGIR Conference on Research Development in Information Retrieval*, SIGIR '18, page 245–254, New York, NY, USA, 2018. ACM.
9 Liu Yang, Hamed Zamani, Yongfeng Zhang, Jiafeng Guo, and W. Bruce Croft. Neural matching models for question retrieval and next question prediction in conversation. *CoRR*, *abs/1707.05409*, 2017.

4.3 Modeling Conversational Search

Elisabeth André (Universität Augsburg, DE), Nicholas J. Belkin (Rutgers University – New Brunswick, US), Phil Cohen (Monash University – Clayton, AU), Arjen P. de Vries (Radboud University Nijmegen, NL), Ronald M. Kaplan (Stanford University, US), Martin Potthast (Universität Leipzig, DE), and Johanne Trippas (RMIT University – Melbourne, AU)

4.3.1 Description and Motivation

An information-seeking system cannot carry out a two-way conversation to make a search more effective unless it maintains interpretable models of its own capabilities and resources, its beliefs about the goals and capabilities of the user, the history and current state of the search process, the context of the search, and other strategies and sources that might satisfy the user's information need. The reflection and self-awareness that these models support enable conversations that help the system and user come to a common understanding of the user's underlying objectives and help the user understand what the system can and cannot do. This should result in a shared plan for executing a successful search. The models are refined or reconstructed through the course of the conversational interaction, as intermediate results are presented and discussed, the search mission is clarified, and new goals and constraints come to light. Importantly, the system's strategic behavior is guided by its ability to inspect the explicit representations of intents, capabilities, and history that the evolving models encode.

In order for a conversational system to talk about a topic, it needs to have a model of that topic. Current deeply learned systems that are trained from prior conversational interactions about arbitrary topics incorporate latent topic models. However, training such a system would require a huge amount of conversational data about that topic, an effort that would be infeasible for conversational search tasks. Rather, a more fruitful approach may be a factored model that separately models conversation, as applied to information-seeking tasks. Thus, systems would learn how to talk separately from the specific content.

Conversational search systems should be collaborative in the sense that they attempt to satisfy the user's information seeking goals. However, people do not often state what their motivating information-seeking goals are, and their specific information requests may not literally state what they are looking for. The conversational search system of the future should interact collaboratively with the user to narrow down the interpretation of the user's desires, especially in the face of search failures, vague descriptions, unstructured digital information, non-digital information, and non-federated information sources, such as a museum's archives.

Thus, in order for a conversational system to be helpful, it needs a model of the task that motivates the information-seeking request. Such a model would enable the conversational system to find alternative approaches to achieving the higher-level motivating goal when a failure occurs. Additionally, the conversational system would need a model of the user, especially if the information-seeking task is extended over time, in order that the system does not tell the user what it believes the user already knows. The user model should contain models of what the user knows, is intending to do or come to know, what s/he has already done, etc. Such models could be derived from general background knowledge and from prior interactions with the system. Among the elements of the user model should be a model of what the user thinks the system can do, what it contains/knows, etc. The conversational search system will need to reveal its capabilities during interaction because it cannot display all its capabilities as menu items. The system will also need a model of itself and models of other non-federated systems, in order that it be able to provide information that it is incapable of handling a request, but the user should inquire with another system that may contain the desired information. During the conversation, the user may state, or the system may request, information about the task or goal that is motivating the user's information need. In order to understand the user's natural language response, the system will need to build its own model of the user's goals, intentions, tasks, and planned actions. Such a model will need to be precise enough to inform the search system, but not require such precision and certainty that it cannot handle vague user responses. Indeed, part of the conversational search system's collaborative task is to gradually elicit such information and in order to narrow down such vague requests. The model of the task should at least provide parameters and actions that the information system can use to perform such sharpening.

4.3.2 Proposed Research

Humans have the ability to infer information about the user's beliefs and wants based on the situative and conversational context and consider this information when performing search tasks with others. For example, we might tell somebody leaving the house where to find an umbrella even when it is currently not raining, but considering that it might rain according to the weather forecast. Current search engines tend to take a macroscopic view and present the users with a number of options they might be interested in. For example, one of the authors of this abstract was provided with suggestions of hotels in cities she has visited before even though she had no intention to visit most of the cities again. While such an unsolicited collection might inspire people to explore new ideas, there are situations where users expect more selective results based on a specific search request. To accomplish this task, a system requires a deeper understanding of the user's desires, beliefs and intentions as well as the situational and conversational context. In the area of cognitive sciences, such an ability is called "Theory of Mind". In many applications, such as the medical domain, it is critical to know how a system retrieved its search results, how confident it is about their sources and how results from different sources have been integrated. A system that is able to explain its behaviors is likely to increase user trust. Thus in addition to a model of the user's wants and

beliefs, an explicit representation of the system's self-model is required. An explicit model of the people's and system's wants and beliefs is a necessary prerequisite for collaborative conversational search where the system, for example, asks for additional information from the user or refers to third parties to accomplish the user's initial search request.

Despite significant attempts to formalize models of the users' and the system's belief and wants for dialogue systems, this research has found surprisingly little attention in conversational search. We do not argue that all applications require deep models and explanations. In particular, users might feel overwhelmed by a system revealing too many details on its inner workings.

1. Investigate how conversational search may be enhanced by a model of the users' beliefs and wants
2. Enhance conversational search by a reflective mechanism that explains the applied search mechanism and the accessed sources
3. Explore techniques to find a good balance between macroscopic and microscopic modeling and explanation

4.4 Argumentation and Explanation

Khalid Al-Khatib (Bauhaus-Universität Weimar, DE), Ondrej Dusek (Charles University – Prague, CZ), Benno Stein (Bauhaus-Universität Weimar, DE), Markus Strohmaier (RWTH Aachen, DE), Idan Szpektor (Google Israel – Tel-Aviv, IL), and Henning Wachsmuth (Universität Paderborn, DE)

4.4.1 Description

Search, in a broader sense, means to satisfy an information need of a person. Conversational search, in particular, restricts the exchange of information to achieve this goal to natural language primarily (in contrast to having access to powerful display, for instance). Although a conversation may be pleasant to the information seeker, it usually implies a reduction in bandwidth: Which of the possibly many search refinement criteria should be asked first by the system? When to get what piece of information from the information seeker? Which retrieved search result should be shown first?

A conversational search system definitely introduces a bias when choosing among questions and results, and it may frame the entire information seeking process. This raises the need for a conversational search system to explain its decisions. Even more, the conversational search system may implicitly tell the information seeker what are the important concepts related to the information need and may change the seeker's beliefs on the topic. Argumentation technology provides the means to address these and related issues.

4.4.2 Motivation

Argumentation and explanation are required for different purposes in conversational search. They can be essential to justify each move the system takes in the conversation, especially if the information seeker explicitly requests such information. Furthermore, argumentation is a fundamental mechanism to acknowledge different viewpoints of a discussed topic. Accordingly,

argumentation technology may be used for result diversification or aspect-based search within conversational settings.

An exemplary conversational search scenario where argumentation plays a key role is scholarly research. When an information seeker attempts, e.g., to search for the best venue to submit a paper to or aims to find the most influential studies for a concrete research topic, it is highly beneficial that the system explains its answers during the conversation and even supports them with high-quality evidence.

4.4.3 Proposed Research

To build new computational models of argumentative conversational search, appropriate training data is required first. We propose to start with existing datasets with conversational argumentative content, such as debate portals and forum discussions (e.g., debate.org, Reddit ChangeMyView, Wikipedia talk pages, or news comments) and community question answering platforms, such as Quora [2]. However, these datasets need to be filtered to focus on search scenarios only. We believe that this can be done (semi-)automatically by following the role and engagement of the seeker in the debate. Additional non-search data as well as data from wiki-like debate portals (e.g., idebate.org) can be used later to improve argumentation capabilities of the models.

To further understand the topic and to support more efficient model training, we propose developing a specific annotation scheme related to conversational search, building upon works of [3], [1], and [4]. This scheme should roughly include the following layers:

- **Conversational layer.** Argumentative relations, speech acts, rhetorical moves.
- **Demographics layer.** Socio-demographic indicators of participants as far as available, involvement of the seeker.
- **Topic layer.** Specific domain concepts, frames.

Furthermore, the annotation should clarify why and how each specific conversation relates to search and to a conversational need as well as why argumentation or explanation are needed to satisfy this need. As the immediate next step, we propose to run a small-scale annotation pilot study which will result in a theoretical analysis of argumentation strategies in conversational search and in data annotation guidelines tested for annotator agreement.

4.4.4 Research Challenges

When providing information within the conversation between a system and an information seeker, the system needs to incrementally decide upon three basic questions matching concepts from research on rhetoric and argumentation synthesis [5]:

1. **Selection.** How to select information, i.e., what to convey to the seeker?
2. **Arrangement.** How to arrange the information, i.e., what to say first and what later?
3. **Phrasing.** How to phrase the information, i.e., what linguistic style to use?

A question arising specifically in argumentative contexts is whether the way the system provides the information should be personalized towards the profile of a specific seeker or should stay general to all seekers. A related issue is the possibility and extent of learning from user-provided information and user feedback. Also, there is a trade-off between the conciseness and the comprehensiveness of the arguments and explanations given for certain information or for the behavior of the system.

As indicated above, however, the most immediate challenge is that no corpora are available so far that sufficiently allow carrying out the research that we propose. We therefore argue that the first challenges to be tackled are the following:

- **Data.** The acquisition of a corpus for studying argumentation in conversational search.
- **Annotation.** The annotation of the corpus towards the scheme outlined above.

4.4.5 Broader Impact

Integrating argumentation and explanation in conversational search will help elevate the retrieval of information from providing documents in a search interface to providing contextual information about sources, viewpoints, potential biases, and conventions in a more natural and dialogue-oriented way. Having explicit structures for argumentation and explanation in search allows information seekers to ask clarification and justification questions. Also, it can help the seekers to build better mental models of the underlying information retrieval processes. This will also enable to navigate different perspectives of controversial debates and thereby has the potential to overcome some of the pressing challenges of search today including filter bubbles, bias in information provision, or misinformation.

References

1. Khalid Al Khatib, Henning Wachsmuth, Kevin Lang, Jakob Herpel, Matthias Hagen, and Benno Stein. Modeling Deliberative Argumentation Strategies on Wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers*, pages 2545-2555, 2018.
2. Adi Omari, David Carmel, Oleg Rokhlenko, and Idan Szpektor. Novelty Based Ranking of Human Answers for Community Questions. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 215-224, 2016.
3. Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. A Conversational Search Transcription Protocol and Analysis. In *Proceedings of ACM SIGIR Workshop on Conversational Approaches for Information Retrieval*, 2017.
4. Svitlana Vakulenko, Kate Revoredo, Claudio Di Ciccio, and Maarten de Rijke. QRFA: A Data-Driven Model of Information-Seeking Dialogues. In *Advances in Information Retrieval. Proceedings of the 41st European Conference on Information Retrieval*, pages 541-556, 2019.
5. Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al Khatib, Maria Skeppstedt, and Benno Stein. Argumentation Synthesis following Rhetorical Strategies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3753-3765, 2018.

4.5 Scenarios that Invite Conversational Search

Lawrence Cavedon (RMIT University – Melbourne, AU), Bernd Fröhlich (Bauhaus-Universität Weimar, DE), Hideo Joho (University of Tsukuba – Ibaraki, JP), Ruihua Song (Microsoft XiaoIce- Beijing, CN), Jaime Teevan (Microsoft Corporation – Redmond, US), Johanne Trippas (RMIT University – Melbourne, AU), and Emine Yilmaz (University College London, GB)

Our working group identified scenarios that invite conversational search. What emerged is (1) no other modality available (or best modality is different), (2) the task invites

conversation. In this document, we motivate these key scenarios and propose research around prototypical tasks in this space. The associated key research challenges were identified in collecting, constructing and representing the rich multimodal contextual information of conversational search, summarizing and presenting the results in speech-only scenarios, design of conversational strategies and in evaluating the dialogue and search systems. Collaborative conversational search adds further challenges that consider the potentially highly interactive, multimodal and synchronous communication between humans and agents.

4.5.1 Motivation

Natural language conversation is not always the best way for a person to search. Conversational search makes the most sense when (1) the situation requires that a person uses an interaction modality that is better suited to conversational interaction than conventional input and output methods, or (2) when the task requires significant context and interaction. In this section we expand on scenarios related to these two cases, and also explore when conversational search might not be the right approach.

Interaction and Device Modalities that Invite Conversational Search

Conversational search is particularly useful when a person's search interactions will be via a modality other than the traditional screen, keyboard, and mouse. This may be because people do not have immediate access to a conventional computer (e.g., they are driving or cooking), are unable to use one (e.g., due to impaired vision or literacy constraints) or they might be simply not very proficient in typing. It may also be because other form factors that are more readily available that lend themselves to conversation e.g. a smartwatch. Furthermore, many modern form factors, like smart speakers, earbuds, or AR/VR systems, have no keyboard and are designed around speech in- and output. Because speech lends itself to far-field interaction, it enables a person to search without actually going to the device and makes it easy for multiple people to simultaneously interact with the system.

Tasks that Invite Conversational Search

Search tasks currently supported by non-conventional modalities tend to be simple and fact-finding in nature (e.g., "Cortana, what is the weather in Frankfurt?"). However, we expect these systems starting to address more complex tasks (i.e., tasks where different information units need to be inspected and compared) as conversational search capabilities improve. Furthermore, conversation is good for building shared context and common ground, and tasks that require much contextual information – on the part of one or more searchers, the system, or shared between them – invite conversational search even when someone is using conventional modalities.

For this reason, conversational search is likely to be particularly useful for exploratory search tasks where the searcher wants to learn about an area. Such tasks typically require clarification of the searcher's need, and the search process may be so complex that it needs to be decomposed into pieces. Conversation can help guide this process while maintaining the larger picture. Conversational search can also be useful where sense-making is required to understand the content the system provides. In contrast to exploratory search, with casual information seeking the searcher does not have a particular goal and just wants to be entertained in a similar way as when browsing a news feed. As an example, a news article might serve as a starting point which sparks interest in further information about some mentioned facts which could be verbally expressed without the need of going to a search engine. In such scenarios, users are often looking to cognitively and affectively make sense

of how the world works and why or they might want to relate some provided information to their personal environment and life. Conversational search may also be useful when a balanced view is important to understand a particular issue and come up with solutions to the issue.

Finally, conversational search makes much sense in contexts where multiple people are involved and there is a shared context. People communicate with each other via conversation, in meetings, via email and text chat, and even through things like comments in documents. A conversational search system is likely to be a good way to address information needs that come up in the course of these conversations, and conversational search tasks seem particularly likely to be collaborative.

Scenarios that Might *not* Invite Conversational Search

Conversational search is not always a good idea and can add overhead for simple information needs where existing channels already work well. Conversations carry cognitive load and offer limited bandwidth. The traditional keyword search paradigm thus probably makes more sense than conversation when a person's modality is not constrained, it is easy for them to describe their information need via querying, and the task requires high bandwidth output that is well served by a ranked list. This may be particularly true for highly ambiguous situations where quick iteration is useful, as people often have a hard time understanding the limits of conversational systems, and recovering from failure in natural language can be hard. Speech based systems can also be problematic in social situations where they can disrupt others or unintentionally expose private information.

4.5.2 Proposed Research

We propose that conversational search research focus on addressing these modalities and tasks. Prototypical scenarios that look at interaction and modalities that invite conversational search often include speech, and must handle noise, address distraction and errors, and be aware of social context. Some examples include:

- Mechanic fixing a machine, wants to know something to help them do a better job.
- Two people searching for a place to eat dinner via speech while driving. The system asks for their preferences and mediates their discussion of the options.

Prototypical scenarios that address tasks that invite conversational search are ones that require significant exploration, interaction, and clarification. Examples include:

- Learning about a recent medical diagnosis. Includes the person asking for general information, the system asking clarifying questions and providing some context, and then dealing with follow up questions from the person.
- Following up on a news article to learn more about the topic and get additional closely or loosely related facts.

4.5.3 Research Challenges and Opportunities

Various research questions arise due to the multimodal aspect of conversational search, as well as due to the importance of considering the context for conversational search. Some issues particularly important in speech-based conversational systems in general also apply to conversational search such as the personality of the system as well as privacy and security issues which we do not discuss here.

Context in Conversational Search

With the multimodality and richer scenarios for conversational search in mind, a variety of contextual aspects need to considered including task context, personal context (affect, cognitive load, etc.), spatial context (location, environment), or social context. General research questions regarding the context in conversational search might include: What are the contextual factors where conversational search systems are reliable to collect and process and what are not? What are effective mechanisms and models for collecting, constructing this contextual information? Are (personal) knowledge graphs and knowledge bases sufficient for representing this information? How could the system incorporate these additional sources of information into the search process?

Result presentation

Speech-only communication is a not an uncommon modality for conversational systems, and this raises specific challenges in the case of output from Conversational Search Systems, which can provide information-rich output that may be difficult to process by human consumers, due to cognitive and memory limitations. The temporally-linear and ephemeral nature of speech also limits the ability to "scan" results: strategies for overcoming such limitations needs to be devised, possibly including:

- Designing methods to present result summaries, or of result categories, to facilitate discussion and clarification of results of specific interest;
- Designing techniques to facilitate "tagging" of results for later reference;
- Designing techniques to highlight specific aspects of results to indicate their relevance.

Conversational strategies and dialogue

New conversational strategies that support information seeking behaviours need to be designed: The conversational structure implemented by a system should mirror and/or support information seeking behaviour, which raises various questions such as:

- How to detect and model information seeking behaviours that should be supported?
- What do the corresponding conversational structures/operations look like: e.g., what conversational operations support identifying the user's uncompromised information need?

Conversational search can provide opportunities to ask users clarifying questions to obtain more information about their search task, work tasks and personal condition (e.g. medical condition) for a better understanding of the users' needs, to personalise the responses to an individual user or to recover from errors. What is the structure of clarifying questions that help better understand end-users search tasks and work tasks? What are effective mechanisms for constructing such clarification questions? What level of personification is desirable in conversational search tasks?

Evaluation

Availability of different modalities would also require the design of new evaluation methodologies for conversational search which should consider implicit and explicit satisfaction signals present in responses from users including affect, tone of voice and cognitive load. In a dialogue we can also explicitly ask for feedback or implicitly provoke conversational responses that inform the evaluation.

Collaborative Conversational Search

Person-to-person communication scenarios are a particularly promising application field of speech-based conversational search since the need for search might naturally emerge from a conversation. Here, the general challenge is to augment unobtrusively a potentially highly interactive, multimodal and synchronous communication of humans being co-located or at different locations (e.g., Skype). Conversational agents need to be aware of the roles of the users and social context of the communication. Furthermore, when multiple people are involved, conflicts, different points of view and different goals and interests are an inherent part of the conversational search process.

Particular research challenges for collaborative scenarios include the identification of prototypical, collaborative information seeking processes, the extraction of an information need from a conversation happening between people and the construction of a corresponding representation of the information seeking task. Work on research questions such as how personal knowledge graphs of individual users can be merged into a group knowledge graph or how to design effective multi-party NLP systems can provide the necessary building blocks for collaborative conversational search systems.

4.6 Conversational Search for Learning Technologies

Sharon Oviatt (Monash University – Clayton, AU) and Laure Soulier (UPMC – Paris, FR)

Conversational search is based on a user-system cooperation with the objective to solve an information-seeking task. In this report, we discuss the implication of such cooperation with the learning perspective from both user and system side. We also focus on the stimulation of learning through a key component of conversational search, namely the multimodality of communication way, and discuss the implication in terms of information retrieval. We end with a research road map describing promising research directions and perspectives.

4.6.1 Context and background

What is Learning?

Arguably, the most important scenario for search technology is lifelong learning and education, both for students and all citizens. Human learning is a complex multidimensional activity, which includes procedural learning (e.g., activity patterns associated with cooking, sports) and knowledge-based learning (e.g., mathematics, genetics). It also includes different levels of learning, such as the ability to solve an individual math problem correctly. It also includes the development of meta-cognitive self-regulatory abilities, such as recognizing the type of problem being solved and whether one is in an error state. These latter types of awareness enable correctly regulating one's approach to solving a problem, and recognizing when one is off track by repairing momentary errors as needed. Later stages of learning enable the generalization of learned skills or information from one context or domain to others– such as applying math problem solving to calculations in the wild (e.g., calculation of garden space, engineering calculations required for a structurally sound building).

Human versus System Learning

When people engage an IR system, they search for many reasons. In the process they learn a variety of things about search strategies, the location of information, and the topic about which they are searching. Search technologies also learn from and adapt to the user, their situation, their state of knowledge, and other aspects of the learning context [4]. Beyond adaptation, the engagement of the system impacts the search effectiveness: its pro-activity is required to anticipate user's need, topic drift, and lower the cognitive load of users [10]. For example, when someone is using a keyboard-based IR system of today, educational technologies can adapt to the person's prior history of solving a problem correctly or not, for example by presenting a harder problem next if the last problem was solved correctly, or presenting an easier problem if it was solved incorrectly.

Based on conversational speech IR systems, it is now possible for a system to process a person's acoustic-prosodic and linguistic input jointly, and on that basis a system can adapt to the person's momentary state of cognitive load. The ideal state for engaging in new learning would be a moderate state of load, whereas detection of very high cognitive load might suggest that the person could benefit from taking a break for some period of time or address easier subtopics to decomplexify the search task [3].

4.6.2 Motivation

How is Learning Stimulated?

Based on the cognitive science and learning sciences literature, it is well known that human thought is spatialized. Even when we engage in problem-solving about temporal information, we spatialize it [5]. Since conversational speech is not a spatial modality, it is advantages to combine it with at least one other spatial modality. For example, digital pen input permits handwriting diagrams and symbols that convey spatial location and relations among objects. Further, a permanent ink trace remains, which the user can think about. Tangible input like touching and manipulating objects in a virtual world also supports conveying 3D spatial information, which is especially beneficial for procedural learning (e.g., learning to drive in a simulator). Since learning is embodied and enhanced by a person's physical activity, touch, manipulation, and handwriting can spatialize information and result in a higher level of interactivity, producing more durable and generalizable learning. When combined with conversational input for social exchange with other people, such input supports richer multimodal input.

Based on the information-seeking point of view, the understanding of users' information need is crucial to maintain their attention and improve their satisfaction. As of now, the understanding of information need has been evaluated using relevant documents, but it implies a more complex process dealing with information need elicitation due to its formulation in natural language [2] and information synthesis [6, 11]. There is, therefore, a crucial need to build information retrieval systems integrating human goals.

How Can We Benefit from Multimodal IR?

Multimodality is the preferred direction for extending conversational IR systems to provide future support for human learning. A new body of research has established that when a person can use multimodal input to engage a system, all types of thinking and reasoning are facilitated, including (1) convergent problem solving (e.g., whether a math problem is solved correctly); (2) divergent ideation (e.g., fluency of appropriate ideas when generating science

hypotheses); and (3) accuracy of inferential reasoning (e.g., whether correct inferences about information are concluded or the information is overgeneralized) [9]. It is well recognized within education that interaction with multimodal/multimedia information supports improved learning. It also is well recognized that this richer form of information enables accessibility for a wider range of diverse students (e.g., blind and hearing impaired, lower-performing, non-native speakers) [9].

For these and related reasons, the long-term direction of IR technologies would benefit by transitioning from conversational to multimodal systems that can substantially improve both the depth and accessibility of educational technologies. With respect to system adaptivity, when a person interacts multimodally with an IR system, the system now can collect richer contextual information about his or her level of domain expertise [8]. When the system detects that the person is a novice in math, for example, it can adapt by presenting information in a conceptually simpler form and with fewer technical terms. In contrast, when a person is detected to be an expert, the system can adapt by upshifting to present more advanced concepts using domain-specific terminology and greater technical detail. This level of IR system adaptivity permits targeting information delivery more appropriately to a given person, which improves the likelihood that he or she will comprehend, reuse, and generalize the information in important ways. The more basic forms of system adaptivity are maintained, but also substantially expanded by the integration of more deeply human-centered models of the person and their existing knowledge of a particular content domain.

Apart from the greater sophistication of user modeling and improved system adaptivity, multimodal IR systems would benefit significantly by becoming more robust and reliable at interpreting a person's queries to the system, compared with a speech-only conversational system [7]. This is because fusing two or more information sources reduces recognition errors. There are both human-centered and system-centered reasons why recognition errors can be reduced or eliminated when a person interacts with a multimodal system. First, humans will formulate queries to the IR system using whichever modality they believe is least error-prone, which prevents errors. For example, they may speak a query, but switch to writing when conveying surnames or financial information involving digits. In addition, when they encounter a system error after speaking input, they can switch to another modality like writing information or even spelling a word–which leads to recovering from the error more quickly. When using a speech-only system, instead the person must re-speak information, which typically causes them to hyperarticulate. Since hyperarticulate speech departs farther from the system's original speech training model, the result is that system errors typically increase rather than resolving successfully [7].

How can user learning and system learning function cooperatively in a multimodal IR framework?

Conversational search needs to be supported by multimodal devices and algorithmic systems trading off search effectiveness and users' satisfaction [10]. Figure 6 illustrates how the user, the system, and the multimodal interface might cooperate. The conversation is initiated by users who formulate their information need through a modality (voice, text, pen, etc). The system is expected to be proactive by fostering both (1) user revealment by eliciting the information need and (2) system revealment by suggesting what actions are available at the current state of the session [1]. In response, users are able to clarify their need and the span of the search session, providing them a deeper knowledge with respect to their information need. The relevant features impacting both users and system's actions include (1) users' intent, (2) users' interactions, (3) system outputs, and (4) the context of the session

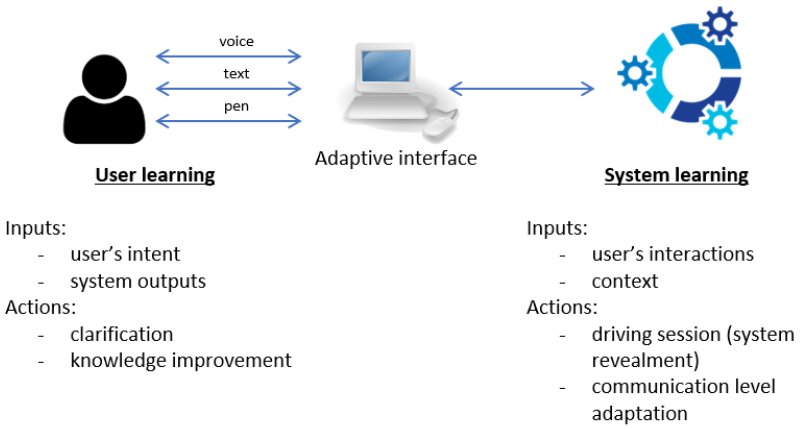

Figure 6 User Learning and System Learning in Conversational Search.

(communication modality, spatial and temporal information, etc.). Several advantages of the user and system cooperation might be noticed. First, based on past interactions, the system is able to learn from right and wrong past actions. It is, therefore, more willing to target IR pieces of information that might be relevant to users. This straightforward allows reducing interactions between users and systems and lower the cognitive effort of users. Second, users being driven by increasing their knowledge acquisition experience, the system should be able to learn users' satisfaction and therefore bolster new information in the retrieval process. Altogether, these advantages advocate for a more sophisticated and a deeper user modeling regarding both knowledge and retrieval satisfaction.

4.6.3 Research Directions and Perspectives

Proposed Research and Challenges: Directions for the Community and Future PhD Topics. Among the key research directions and challenges to be addressed in the next 5-10 years in order to advance conversational search as a more capable learning technology are the following:

- Transforming existing IR knowledge graphs into richer multi-dimensional ones that currently are used in multimodal analytic research — which supports integrating information from multiple modalities (e.g., speech, writing, touch, gaze, gesturing) and multiple levels of analyzing them (e.g., signals, activity patterns, representations).
- Integration of multimodal input and multimedia output processing with existing IR techniques
- Integration of more sophisticated user modeling with existing IR techniques, in particular ones that enable identifying the user's current expertise level in the content domain that is the focus of their search and leveraging the span of the search session.
- Conversely, integrating analytics that enable the user to identify the authoritativeness of an information source (e.g., its level of expertise, its credibility or intent to deceive).
- Development of more advanced multimodal machine learning methods that go beyond audio-visual information processing and search. Development of more advanced machine learning methods for extracting and representing multimodal user behavioral models.

Broader Impact. The research roadmap outlined above would result in major and consequential advances, including in the following areas:

- More successful IR system adaptivity for targeting user search goals.

- IR systems that function well based on fewer and briefer interactions between user and system.
- IR system that are more reliable and robust at processing user queries. Expansion of the accessibility of IR technology to a broader population.
- Improved focus of IR technology on end-user goals and values, rather than commercial for-profit aims.
- Improvement of powerful machine learning methods for processing richer multimodal information and achieving more deeply human-centered models.
- Acceleration of the positive impact of lifelong learning technologies on human thinking, reasoning, and deep learning.

Obstacles and Risks.

- Establishing and integrating more deeply human-centered multimodal behavioral models to advance IR technologies risks privacy intrusions that must be addressed in advance.
- Establishing successful multidisciplinary teamwork among IR, user modeling, multimodal systems, machine learning, and learning sciences experts will need to be cultivated and maintained over a lengthy period of time.
- Mutually adaptive systems risk unpredictability and instability of performance, and must be studied to achieve ideal functioning.
- New evaluation metrics will be required that substantially expand those used by IR system developers today.

Acknowledgements

We would like to thank the Schloss Dagstuhl and the seminar organizers Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein for this week of research introspection and networking. We also thank the ANR project SESAMS (Projet-ANR-18-CE23-0001) which supports Laure Soulier's work on this topic.

References

1 Leif Azzopardi, Mateusz Dubiel, Martin Halvey, and Jeff Dalton. Conceptualizing agent-human interactions during the conversational search process, 2018.
2 Wafa Aissa, Laure Soulier, and Ludovic Denoyer. A reinforcement learning-driven translation model for search-oriented conversational systems. In *Proceedings of the 2nd International Workshop on Search-Oriented Conversational AI*, SCAI@EMNLP 2018, Brussels, Belgium, October 31, 2018, pages 33–39, 2018.
3 Ahmed Hassan Awadallah, Ryen W. White, Patrick Pantel, Susan T. Dumais, and Yi-Min Wang. Supporting complex search tasks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM 2014, Shanghai, China, November 3-7, 2014, pages 829–838, 2014.
4 Kevyn Collins-Thompson, Preben Hansen, and Claudia Hauff. Search as Learning (Dagstuhl Seminar 17092). *Dagstuhl Reports*, 7(2):135–162, 2017.
5 Philip N. Johnson-Laird. Space to think. In L. Nadel P. Bloom, M. Peterson and M. Garrett, editors, Language and Space, pages 437–462. The MIT Press, Cambridge MA., 1999.
6 Gary Marchionini. Exploratory search: from finding to understanding. Commun. ACM, 49(4):41–46, 2006.
7 Sharon L. Oviatt and Philip R. Cohen. The Paradigm Shift to Multimodality in Contemporary Computer Interfaces. *Synthesis Lectures on Human-Centered Informatics*. Morgan & Claypool Publishers, 2015.

8 Sharon Oviatt, Joseph Grafsgaard, Lei Chen, and Xavier Ochoa. The handbook of multimodal-multisensor interfaces. *chapter Multimodal Learning Analytics: Assessing Learners' Mental State During the Process of Learning*, pages 331–374. Association for Computing Machinery and Morgan & Claypool, New York, NY, USA, 2019.
9 S. L. Oviatt. The Design of Future of Educational Interfaces. Routledge Press, New York, 2013.
10 Zhiwen Tang and Grace Hui Yang. Dynamic search – optimizing the game of information seeking. *CoRR, abs/1909.12425*, 2019.
11 Ryen W. White and Resa A. Roth. Exploratory Search: Beyond the Query-Response Paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*. Morgan & Claypool Publishers, 2009.

4.7 Common Conversational Community Prototype: Scholarly Conversational Assistant

Krisztian Balog (University of Stavanger, NOR), Lucie Flekova (Technische Universität Darmstadt, DE), Matthias Hagen (Martin-Luther-Universität Halle-Wittenberg, DE), Rosie Jones (Spotify, US), Martin Potthast (Leipzig University, DE), Filip Radlinski (Google, UK), Mark Sanderson (RMIT University, AUS), Svitlana Vakulenko (University of Amsterdam, NL), and Hamed Zamani (Microsoft, US)

4.7.1 Description

This working group discussed the potential for creating academic resources (tools, data, and evaluation approaches) to support research in conversational search, by focusing on realistic information needs and conversational interactions. Specifically, we propose to develop and operate a prototype conversational search system for scholarly activities. This Scholarly Conversational Assistant would serve as a useful tool, a means to create datasets, and a platform for running evaluation challenges by groups across the community.

4.7.2 Motivation

Conversational search is a newly emerging research area that aims to provide access to digitally stored information by means of a conversational user interface, that is, a dialogue-based interaction inspired and informed by human communication processes [5, 15, 18]. The major goal of a conversational search system is to effectively retrieve relevant answers to a wide range of questions expressed in natural language, with rich user-system dialogue as a crucial component for understanding the question and refining the answers [1]. The respective dialogue comprises of a sequence of exchanges between one or more users and a conversational search system, which can enable multi-step task completion and recommendation [6]. Several theoretical frameworks that further specify various components and requirements for an effective conversational search system have recently been proposed [14, 2, 16, 19, 17].

It is commonly recognized that only few natural conversational search corpora exist. Rather, corpora are often created through imagined needs (often in task-oriented Wizard-of-Oz studies), are inspired by logs, or come from crawls of community fora. This leads to significant research effort being planned around existing biased data and metrics, rather than data and metrics being constructed to support the most impactful research. While

there have been instances of the research community interaction enabling research, such as at ECIR 2019,[2] this is relatively rare. One of our key motivations is to produce a system and corpus that contains and supports real user needs.

Simultaneously, our community has common unsatisfied needs that appear very well suited to conversational search. Some common tasks are performed by researchers repeatedly without providing any community research value in terms of data and feedback collection, despite being relevant to many published experiments. Examples of these tasks include PC selection or finding interest profiles in EasyChair, or identifying the most relevant sessions in the Whova conference app. The collective time spent (arguably inefficiently) by our community on such tasks may far surpass the cost of creating a system that also supports research progress while providing this *community value*.

4.7.3 Proposed Research

We propose to develop and operate a prototype conversational search system (Scholarly Conversational Assistant) that would serve as

- a useful search tool,
- a means to create datasets for further academic research,
- and a platform for running evaluation challenges by groups across the community.

In particular, the Scholarly Conversational Assistant would allow our research community to perform a range of research-related activities. In extensive discussions, we settled on this domain for a number of reasons: (1) The data that is involved (such as papers authored, conferences/talks attended, PC memberships) is generally considered less private. Indeed most such data is already public albeit difficult to search. (2) The system is one that the members of our community would be using ourselves, giving an active knowledgeable participant base, who could contribute improvements and publish papers based on interactions observed. (3) It caters to a broad range of information needs (see below) that are currently not supported well by existing systems. (4) The relevant research groups could avoid competing with commercial providers.

A number of other possible domains were discussed, including movies, music, news, and podcasts. They have a significantly larger potential audience, yet potentially compete with commercial providers. In determining our plan, it became clear that some participants also consider interests in these areas to be highly sensitive or personal. As a critical constraint, privacy of relevant data is key (having impacted, for example, the Living Labs research [10] despite significant effort).

4.7.4 Research Challenges

The aim of the Scholarly Conversational Assistant system would be to enable a wide variety of research in conversational search by covering example information needs like:

- "What should I read?"–Find research on a new area of interest.
- "Help me plan my attendance"–Plan what sessions to attend and whom to talk to at a conference. (Conference organizers could also use that information for optimizing room allocations.)
- "Whom should I invite?"–Find conference PC, SPC, session chairs, invite speakers, etc.

[2] http://ecir2019.org/sociopatterns/

Importantly, the system would log all interactions such that classes of information needs that have potential for study may be identified over time. People may evaluate the system by filling out a questionnaire, with the option of free text feedback, after each conversation (and possibly leave comments behind for individual system utterances).

Connection to Knowledge Graphs

The system would operate on a *personal research graph* (PKG) [3], more specifically, the portion of the PKG that the user wants to share with the system. The PKG could include, among other information:

- Authorship information (which may be connected to a public citation graph),
- Conference committee membership, awards, etc.,
- Talks given anywhere public,
- Attendance of conferences, sessions, etc.,
- (in the private part) Annotations of papers, notes on talks, etc.

First Steps

The project is ambitious, but we think it can be grown incrementally:

- A starting point would be to get one ore more graduate students to start coding a tool and check it in to GitHub. It is likely that students will be able to build on top of existing infrastructure. In order for this to work, it will be necessary for a research team to own the decisions who (believes they will) get value out of such work. With a prototype system in place, one could establish a shared task at a workshop or conduct a lab study at scale. One might also design a challenge at TREC/CLEF to make use of the skeleton.
- One might alternatively start by collecting evidence that such a system is something the community actually wants. Here, a sample of dialogues or information needs (that one might want to support) could be gathered.

4.7.5 Broader Impact

The organization of shared tasks has a long tradition in information retrieval as well as natural language processing and the dialogue community within it. In conversational search, these two communities will collaborate to build search systems that have a natural language interface as well as conversational capabilities. The breadth of potential tasks that are due to this confluence of research fields–as also identified in Dagstuhl Seminar 19461–is large. As such, developing common infrastructure and shared tasks would have high value for the community.

In particular, the outcome of shared tasks are typically large corpora and performance measures that, together, form reusable benchmarks. For example, the Cranfield-style evaluation frameworks that were adapted by TREC, or the corpora developed for the CoNLL shared tasks have had a broad impact on their respective communities at large. We expect that a conversational search challenge, too, will help to align and shape the community.

Moreover, by developing specific shared tasks in the form of living labs [9, 10], we see the opportunity to apply early conversational search systems in practice as soon as possible. Here, the application domain of scholarly search, while allowing for a wide range of basic and advanced evaluation setups, may ideally transfer directly into new prototypes to enhance research itself, for instance, impacting the productivity of managing one's personal conferences schedules.

4.7.6 Obstacles and Risks

A variety of systems for storing and accessing research publications, reviews and conference attendance already exist. For the Scholarly Conversational Assistant to be successful, it must either be more useful than these, or potentially integrate with them. Some of the existing systems include: dblp, semantic scholar, ACM library, Google scholar, ACL anthology, open review, arXiv, Athena conference chatbot, Citeseer, Arnetminer, and arXivDigest (more on these in related reading).

Risks involved in operationalizing our envisaged conversational search system include:

- *Privacy and data retention rules.* Ideally, the Scholarly Conversational Assistant would allow the logging of user interactions including voice input. For all personal data, the system would require a process for data access, retention and deletion as well as logging, in compliance with local regulations. Even the use of third-party speech recognizers may be sensitive depending on the location of data storage.
- *Opinions != facts in indexing.* Some information that could be collected is likely to be expressed opinions rather than facts (e.g., tweets about papers). Thus, we may want to allow verification of such information before use for search and recommendation, or present it in a separate clearly-marked format with the potential for correction or deletion. Others may wish to combine private information (such as a user's personal opinions about papers), without this information being propagated.
- *Speech recognition.* The use of third-party speech recognizers may be sensitive depending on the location of data storage. In addition, in the Scholarly Conversational Assistant case, the corpus contains many proper names and technical terms. A speech recognizer may require a custom language model integrating this corpus to perform well.
- *Personal Knowledge Graph implementation.* We would need a design that allows both cloud- and client-side storage of personal data. We need to make sure that private parts of the PKG remain private and also that users have full control over what is stored in their PKG. In case an offline dataset is created and shared, there needs to be an agreement in place that ensures that personal data would need to be removed upon request. (It should be noted that there is no way to enforce this, and "unauthorized" access may only be spotted if people publish using that data.)
- *Usage volume.* Low user participation is a concern. Beyond ensuring that the system is useful, other ways to mitigate this could include rewarding (paying) users or incentivizing them through gamification (e.g., at conferences to use the system).
- *Implementation.* The underlying system would require a significant effort to implement. As this would likely be contributions from different practitioners at various stages in their careers over an extended time, the contributors would naturally change. To alleviate some associated risk, a strong modularization would be beneficial, with clear interfaces and documentation. Moreover, the design of the initial prototype should be as simple as possible, with agreement of how the system's continued development is ensured during operation. The live service would also need coordination, for example, of how live experiments are planned and executed.
- *Operation.* Past academic systems have often been deployed on individual servers without redundancy, and potentially lacking resources for scalability. This project would likely wish to consider for this project to identify possible sponsorship from a cloud provider or host institution with significant cluster resources. The hosting decision should likely take into account long-term commitment.
- *Stability and reproducibility.* If used for online challenges where participants submit code that runs live, this would need to be of suitable quality to be widely used. Care would

need to be taken in designing common APIs that minimize the risks involved where a component does not behave as expected.

4.7.7 Suggested Readings and Resources

In the following, we list a set of resources (data and tools) that might be useful in building such a system.

Software platforms:

- Macaw: A conversational information seeking platform implemented in Python which supports multiple interfaces and modalities [21].
- TIRA Integrated Research Architecture [13] (a modularized platform for shared tasks).

Scientific IR tools:

- ArXivDigest: A personalized scientific literature recommendation framework based on arXiv articles.[3]
- GrapAL: Querying Semantic Scholar's literature graph [4] (web-based tool for exploring scientific literature, e.g., finding experts on a given topic).[4]

Open-source scholarly conversational agents:

- UKP-ATHENA: A scientific conversational agent [12] (early prototype for assisting ACL* conference attendees and answering basic ACL Anthology queries).[5]

Data collections suitable to be incorporated in the Scholarly Conversational Assistant include, but are not limited to:

- Open Research Knowledge Graph[6] (ORKG) [11]: Semantic annotations of scientific publications
- Semantic Scholar: Articles in a broad range of fields
- ACM DL: A subset of computer science articles
- dblp: A clean list of computer science articles
- ACL Anthology: A public collection of ACL* articles
- Open Review: A small subset of conference articles with public reviews
- Other sources include: Google Scholar, Citeseer, Arnetminer, and Conference attendance apps (e.g., Whova)

Other related work:

- [8]: Recupero: Conference Live: Accessible and Sociable Conference Semantic Data
- [7]: Vote Goat: Conversational Movie Recommendation
- [20]: Aminer: Search and mining of academic social networks (researcher-centric IR)

References

1 J. Allan, B. Croft, A. Moffat, and M. Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from swirl 2012 the second strategic workshop on information retrieval in lorne. *SIGIR Forum*, 46(1):2–32, May 2012. ISSN 0163-5840. 10.1145/2215676.2215678. URL http://doi.acm.org/10.1145/2215676.2215678.

[3] https://github.com/iai-group/arxivdigest
[4] https://allenai.github.io/grapal-website/
[5] http://athena.ukp.informatik.tu-darmstadt.de:5002/
[6] http://orkg.org

2 L. Azzopardi, M. Dubiel, M. Halvey, and J. Dalton. Conceptualizing agent-human interactions during the conversational search process. In *The Second International Workshop on Conversational Approaches to Information Retrieval*, July 2018. URL https://strathprints.strath.ac.uk/64619/.

3 K. Balog and T. Kenter. Personal knowledge graphs: A research agenda. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, pages 217–220, New York, NY, USA, 2019. ACM. URL http://doi.acm.org/10.1145/3341981.3344241.

4 C. Betts, J. Power, and W. Ammar. Grapal: Querying semantic scholar's literature graph. *arXiv preprint arXiv:1902.05170*, 2019.

5 B. R. Cowan and L. Clark, editors. *Proceedings of the 1st International Conference on Conversational User Interfaces, CUI 2019, Dublin, Ireland, August 22-23, 2019*, 2019. ACM.

6 J. S. Culpepper, F. Diaz, and M. D. Smucker. Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (swirl 2018). *SIGIR Forum*, 52(1):34–90, Aug. 2018. ISSN 0163-5840. 10.1145/3274784.3274788. URL http://doi.acm.org/10.1145/3274784.3274788.

7 J. Dalton, V. Ajayi, and R. Main. Vote goat: Conversational movie recommendation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 1285–1288, New York, NY, USA, 2018. ACM. URL http://doi.acm.org/10.1145/3209978.3210168.

8 A. L. Gentile, M. Acosta, L. Costabello, A. G. Nuzzolese, V. Presutti, and D. Reforgiato Recupero. Conference live: Accessible and sociable conference semantic data. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 1007–1012, New York, NY, USA, 2015. ACM. URL http://doi.acm.org/10.1145/2740908.2742025.

9 F. Hopfgartner, A. Hanbury, H. Müller, I. Eggel, K. Balog, T. Brodt, G. V. Cormack, J. Lin, J. Kalpathy-Cramer, N. Kando, M. P. Kato, A. Krithara, T. Gollub, M. Potthast, E. Viegas, and S. Mercer. Evaluation-as-a-service for the computational sciences: Overview and outlook. *J. Data and Information Quality*, 10(4):15:1–15:32, Oct. 2018. URL http://doi.acm.org/10.1145/3239570.

10 F. Hopfgartner, K. Balog, A. Lommatzsch, L. Kelly, B. Kille, A. Schuth, and M. Larson. Continuous evaluation of large-scale information access systems: A case for living labs. In N. Ferro and C. Peters, editors, *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*, pages 511–543. Springer, 2019. URL https://doi.org/10.1007/978-3-030-22948-1_21.

11 M. Y. Jaradeh, A. Oelen, K. E. Farfar, M. Prinz, J. D'Souza, G. Kismihók, M. Stocker, and S. Auer. Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture*, K-CAP '19, pages 243–246, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-7008-0. 10.1145/3360901.3364435. URL http://doi.acm.org/10.1145/3360901.3364435.

12 M. Mesgar, P. Youssef, L. Li, D. Bierwirth, Y. Li, C. M. Meyer, and I. Gurevych. When is acl's deadline? a scientific conversational agent. *arXiv preprint arXiv:1911.10392*, 2019.

13 M. Potthast, T. Gollub, M. Wiegmann, and B. Stein. Tira integrated research architecture. In *Information Retrieval Evaluation in a Changing World*, pages 123–160. Springer, 2019.

14 F. Radlinski and N. Craswell. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 117–126, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4677-1. 10.1145/3020165.3020183. URL http://doi.acm.org/10.1145/3020165.3020183.

15 J. R. Trippas. Spoken Conversational Search: Audio-only Interactive Information Retrieval. *PhD thesis*, RMIT University, 2019.

16 J. R. Trippas, D. Spina, L. Cavedon, and M. Sanderson. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 325–328, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4677-1. 10.1145/3020165.3022144. URL http://doi.acm.org/10.1145/3020165.3022144.
17 J. R. Trippas, D. Spina, P. Thomas, M. Sanderson, H. Joho, and L. Cavedon. Towards a model for spoken conversational search. *Information Processing & Management*, 57(2): 102162, 2020.
18 S. Vakulenko. Knowledge-based Conversational Search. *PhD thesis*, TU Wien, 2019.
19 S. Vakulenko, K. Revoredo, C. D. Ciccio, and M. de Rijke. QRFA: A data-driven model of information-seeking dialogues. In *Advances in Information Retrieval – 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings, Part I*, pages 541–557, 2019.
20 H. Wan, Y. Zhang, J. Zhang, and J. Tang. Aminer: Search and mining of academic social networks. *Data Intelligence*, 1(1):58–76, 2019.
21 H. Zamani and N. Craswell. Macaw: An extensible conversational information seeking platform. *arXiv preprint arXiv:1912.08904*, 2019.

5 Recommended Reading List

These publications were recommended by the seminar participants via the pre-seminar survey. Please also refer to the reading list available in individual reports of working groups.

- Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for Human-AI Interaction, *CHI 2019*. https://doi.org/10.1145/3290605.3300233
- Aliannejadi, Mohammad, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. Asking clarifying questions in open-domain information-seeking conversations. *SIGIR 2019*. https://doi.org/10.1145/3331184.3331265
- Belkin, Nicholas J., Colleen Cool, Adelheit Stein, and Ulrich Thiel. Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert systems with applications*, 9 (3), 1995. https://doi.org/10.1016/0957-4174(95)00011-W
- Timothy Bickmore, Justine Cassell. Social dialogue with embodied conversational agents. Advances in natural multimodal dialogue systems, 2005. https://doi.org/10.1007/1-4020-3933-6_2
- Daniel Braun, Adrian Hernandez-Mendez, Florian Matthes, Manfred Langen. Evaluating natural language understanding services for conversational question answering systems. *SIGdial 2017*. https://doi.org/10.18653/v1/W17-5522
- Andrew Breen, et al. Voice in the User Interface. Interactive Displays: Natural Human-Interface Technologies, 2014, https://doi.org/10.1002/9781118706237.ch3.
- Brennan, Susan E., and Eric A. Hulteen. Interaction and feedback in a spoken language system: A theoretical framework. *Knowledge-based systems*, 8 (2-3), 1995. https://doi.org/10.1016/0950-7051(95)98376-H
- Harry Bunt. Conversational principles in question-answer dialogues. *Zur Theorie der Frage*, pages 119-141.
- Justine Cassell. Embodied conversational agents. *AI Magazine*, 22(4), 2001. https://doi.org/10.1609/aimag.v22i4.1593

- Justine Cassell, Joseph Sullivan, Elizabeth Churchill, Scott Prevost. Embodied conversational agents. MIT Press, 2000.
- Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, Luke Zettlemoyer. QuAC: Question Answering in Context. *EMNLP 2018*. https://dx.doi.org/10.18653/v1/D18-1241
- Christakopoulou, Konstantina, Filip Radlinski, and Katja Hofmann. Towards conversational recommender systems. *SIGKDD 2016*. https://doi.org/10.1145/2939672.2939746
- Leigh Clark, Phillip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Benjamin Cowan. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers*, 31 (4), 2019. https://doi.org/10.1093/iwc/iwz016
- Mark Core and James Allen. Coding dialogs with the DAMSL annotation scheme. *AAAI Fall Symposium on Communicative Action in Humans and Machines*, 1997.
- Paul Grice. Studies in the Way of Words. Harvard University Press, 1989.
- Haider, Jutta, and Olof Sundin. Invisible Search and Online Search Engines: The ubiquity of search in everyday life. Routledge, 2019.
- Ben Hixon, Peter Clark, Hannaneh Hajishirzi. Learning knowledge graphs for question answering through conversational dialog. *NAACL 2015*. http:s//dx.doi.org/10.3115/v1/N15-1086
- Mohit Iyyer, Wen-tau Yih, Ming-Wei Chang. Search-based neural structured learning for sequential question answering. *ACL 2017*. https://dx.doi.org/10.18653/v1/P17-1167
- Diane Kelly and Jimmy Lin. Overview of the TREC 2006 ciQA task. *SIGIR Forum* 41(1), 2007. https://doi.org/10.1145/1273221.1273231
- Liu, Bei, Jianlong Fu, Makoto P. Kato, and Masatoshi Yoshikawa. Beyond narrative description: Generating poetry from images by multi-adversarial training. *ACM Multimedia*, 2018. https://doi.org/10.1145/3240508.3240587
- Dominic W. Massaro, Michael M. Cohen, Sharon Daniel, Ronald A Cole. Developing and evaluating conversational agents. *Human performance and ergonomics*, 1999. https://doi.org/10.1016/B978-012322735-5/50008-7
- McTear, Michael F. Spoken dialogue technology: enabling the conversational user interface. *ACM Computing Surveys*, 34 (1), 2002. https://doi.org/10.1145/505282.505285
- Oddy, Robert N. Information retrieval through man-machine dialogue. Journal of documentation, 33 (1), 1977. https://doi.org/10.1108/eb026631
- Filip Radlinski, Nick Craswell. A theoretical framework for conversational search. CHIIR 2017. https://doi.org/10.1145/3020165.3020183
- Siva Reddy, Danqi Chen, Christopher D. Manning. CoQA: A conversational question answering challenge. *TACL, 7, 2019*. https://doi.org/10.1162/tacl_a_00266
- Ren, Gary, Xiaochuan Ni, Manish Malik, and Qifa Ke. Conversational query understanding using sequence to sequence modeling. *The Web Conference*, 2018. https://doi.org/10.1145/3178876.3186083
- Zsófia Ruttkay, Catherine Pelachaud. From brows to trust: Evaluating embodied conversational agents. *Springer Science & Business Media*, 2004. https://doi.org/10.1007/1-4020-2730-3
- Shum, Heung-Yeung, Xiao-dong He, and Di Li. From Eliza to XiaoIce: challenges and opportunities with social chatbots. Frontiers of Information Technology & Electronic Engineering, 19 (1), 2018. https://doi.org/10.1631/FITEE.1700826
- Adelheit Stein, Elisabeth Maier. Structuring Collaborative Information-Seeking Dialogues. *Knowledge-Based Systems*, 8(2-3), 1995. https://doi.org/10.1016/0950-7051(95)98370-L

- Oriol Vinyals, Quoc Le. A neural conversational model. ICML Deep Learning Workshop, 2015. https://arxiv.org/abs/1506.05869
- Marylyn Walker, Dianne Litman, Candace Kamm, Alicia Abella. PARADISE: A framework for evaluating spoken dialogue agents. *ACL 1997*. https://dx.doi.org/10.3115/976909.979652
- Weston, J., Bordes, A., Chopra, S., Rush, A.M., van Merriënboer, B., Joulin, A. and Mikolov, T. Towards ai-complete question answering: A set of prerequisite toy tasks. https://arxiv.org/abs/1502.05698
- Wu, Wei, and Rui Yan. Deep Chit-Chat: Deep Learning for Chit-Chat. *SIGIR 2019*. https://doi.org/10.1145/3331184.3331388
- Zhang, Yongfeng, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. Towards conversational search and recommendation: System ask, user respond. *CIKM 2018*. https://doi.org/10.1145/3269206.3271776
- Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. A dataset for document grounded conversations. *EMNLP 2018*. https://dx.doi.org/10.18653/v1/D18-1076
- Zhou, Li, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of XiaoIce, an empathetic social chatbot. *Computational Linguistics*, 2020. https://doi.org/10.1162/coli_a_00368

6 Acknowledgements

The seminar organisers would like to thank all participants and speakers of invited talks for their active contributions. We also thank the staff of Schloss Dagstuhl for providing a great venue for a successful seminar. The organisers were in part supported by JSPS KAKENHI Grant Number 19H04418. Any opinions, findings, and conclusions described here are the authors and do not necessarily reflect those of the sponsors.

Participants

- Khalid Al-Khatib
Bauhaus University Weimar, DE
- Avishek Anand
Leibniz Universität
Hannover, DE
- Elisabeth André
University of Augsburg, DE
- Jaime Arguello
University of North Carolina at
Chapel Hill, US
- Leif Azzopardi
University of Strathclyde –
Glasgow, GB
- Krisztian Balog
University of Stavanger, NO
- Nicholas J. Belkin
Rutgers University –
New Brunswick, US
- Robert Capra
University of North Carolina at
Chapel Hill, US
- Lawrence Cavedon
RMIT University –
Melbourne, AU
- Leigh Clark
Swansea University, UK
- Phil Cohen
Monash University –
Clayton, AU
- Ido Dagan
Bar-Ilan University –
Ramat Gan, IL
- Arjen P. de Vries
Radboud University
Nijmegen, NL
- Ondrej Dusek
Charles University –
Prague, CZ
- Jens Edlund
KTH Royal Institute of
Technology – Stockholm, SE
- Lucie Flekova
Amazon R&D – Aachen, DE
- Bernd Fröhlich
Bauhaus University Weimar, DE
- Norbert Fuhr
University of Duisburg–
Essen, DE
- Ujwal Gadiraju
Leibniz Universität
Hannover, DE
- Matthias Hagen
Martin Luther University
Halle–Wittenberg, DE
- Claudia Hauff
TU Delft, NL
- Gerhard Heyer
University of Leipzig, DE
- Hideo Joho
University of Tsukuba –
Ibaraki, JP
- Rosie Jones
Spotify – Boston, US
- Ronald M. Kaplan
Stanford University, US
- Mounia Lalmas
Spotify – London, GB
- Jurek Leonhardt
Leibniz Universität
Hannover, DE
- David Maxwell
University of Glasgow, GB
- Sharon Oviatt
Monash University –
Clayton, AU
- Martin Potthast
University of Leipzig, DE
- Filip Radlinski
Google UK – London, GB
- Rishiraj Saha Roy
MPI for Computer Science –
Saarbrücken, DE
- Mark Sanderson
RMIT University –
Melbourne, AU
- Ruihua Song
Microsoft XiaoIce – Beijing, CN
- Laure Soulier
UPMC – Paris, FR
- Benno Stein
Bauhaus University Weimar, DE
- Markus Strohmaier
RWTH Aachen University, DE
- Idan Szpektor
Google Israel – Tel Aviv, IL
- Jaime Teevan
Microsoft Corporation –
Redmond, US
- Johanne Trippas
RMIT University –
Melbourne, AU
- Svitlana Vakulenko
Vienna University of Economics
and Business, AT
- Henning Wachsmuth
University of Paderborn, DE
- Emine Yilmaz
University College London, UK
- Hamed Zamani
Microsoft Corporation, US

Report from Dagstuhl Seminar 19471

BOTse: Bots in Software Engineering

Edited by

Margaret-Anne Storey[1], Alexander Serebrenik[2], Carolyn Penstein Rosé[3], Thomas Zimmermann[4], and James D. Herbsleb[5]

1 **University of Victoria, CA,** mstorey@uvic.ca
2 **Eindhoven University of Technology, NL,** a.serebrenik@tue.nl
3 **Carnegie Mellon University – Pittsburgh, US,** cprose@cs.cmu.edu
4 **Microsoft Corporation – Redmond, US,** tzimmer@microsoft.com
5 **Carnegie Mellon University – Pittsburgh, US,** jim.herbsleb@gmail.com

Abstract

This report documents the program and the outcomes of the Dagstuhl Seminar 19471 "BOTse: Bots in Software Engineering". This Dagstuhl seminar brought researchers and practitioners together from multiple research communities with disparate views of what bots are and what they can do for software engineering. The goals were to understand how bots are used today, how they could be used in innovative ways in the future, how the use of bots can be compared and synthesized, and to identify and share risks and challenges that may emerge from using bots in practice. The report briefly summarizes the goals and format of the seminar and provides selected insights and results collected during the seminar.

Seminar November 17–22, 2019 – http://www.dagstuhl.de/19471
2012 ACM Subject Classification Human-centered computing, Software and its engineering
Keywords and phrases automated software development, bots, chatbots, collaborative software development, cscw, devops, nlp, software engineering
Digital Object Identifier 10.4230/DagRep.9.11.84
Edited in cooperation with Nathan Cassee

1 Executive Summary

James D. Herbsleb
Carolyn Penstein Rosé
Alexander Serebrenik
Margaret-Anne Storey
Thomas Zimmermann

This Dagstuhl seminar brought researchers and practitioners together from multiple research communities with disparate views of what bots are and what they can do for software engineering. The goals were to understand how bots are used today, how they could be used in innovative ways in the future, how the use of bots can be compared and synthesized, and to identify and share risks and challenges that may emerge from using bots in practice.

BOTse: Bots in Software Engineering, *Dagstuhl Reports*, Vol. 9, Issue 11, pp. 84–96
Editors: James D. Herbsleb, Carolyn Penstein Rosé, Alexander Serebrenik, Margaret-Anne Storey, and Thomas Zimmermann
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Bots, often called chatbots, are considered by some to be computer programs that provide a conversational style interface for interacting with software services, while others consider bots to be any semi-autonomous software service that may or may not take on a human-like persona.

Regardless of the definition of what makes a bot a bot, bots are found in many domains such as shopping, entertainment, education, and personal productivity. In software development, bots are rapidly becoming a *de facto* interface for developers and end users to interact with software services in a myriad of ways: e.g., bots are used to fetch or share information, extract and analyze data, detect and monitor events and activities in communication and social media, connect developers with key stakeholders or with other tools, and provide feedback and recommendations on individual and collaborative tasks.

Through this Dagstuhl Seminar, we aimed to gain important insights on how bots may play a role in improving software development productivity and in enhancing collaborative software development. In particular we discussed how bots, with or without a conversational UI, may play a prominent role in software practice. We gathered literature and resources on how bots can have an impact on development processes, software quality, and on end users. The goal was to channel previously siloed communities and through this confluence forge a common vision and plot next steps that might leverage the variety of expertise and push forward both the research and the practices related to bots. The activities were meant to surface the difficult questions and tensions that arise when one looks beyond what at first blush appears to be a superficial distinction, but in fact touches upon core values and driving questions that define the boundaries between fields.

2 Table of Contents

Executive Summary
James D. Herbsleb, Carolyn Penstein Rosé, Alexander Serebrenik, Margaret-Anne Storey, and Thomas Zimmermann 84

Seminar Format 87

Bots in Software Engineering: Insights and Outlook 87

Definitions: Software Bots and Software Bot Ecosystems 87

Developer acceptance 88

To bot or not 89

Ethical concerns 90

Diversity and inclusion 91

Bots to support collaboration 92

Follow up work 93

Overview of Invited Talks 94

Bots in Wikipedia – Brief review of selected research
Claudia Müller-Birn 94

Overview of Natural Language Dialogue Systems
David R. Traum 94

Highlights of the first International Workshop on Bots in Software Engineering (BotSE)
Emad Shihab and Stefan Wagner 94

Bots – The hidden side of software development
Bogdan Vasilescu 95

Participants 96

3 Seminar Format

In this seminar, we brought together researchers and practitioners with diverse backgrounds in software engineering (engineering tools, empirical research), natural language processing (NLP), artificial intelligence (AI), data science, machine learning, human computer interaction (HCI), computer-supported cooperative work (CSCW), computer-supported collaborative learning (CSCL), social computing, affective computing and cognitive computing to discuss, explore and recommend how bots could be used in software engineering.

In advance of the seminar, we conducted a short survey to collect relevant questions to be addressed in the seminar and to form break out group topics. At the seminar, we arranged large and small group discussions, breakout activities, short and long talks and bot demos. These activities helped (a) to foster vibrant discussion, (b) to identify and address relevant questions on how bots are or could be used in software engineering, and to (c) foster interaction and collaborations between attendees.

In the next section, we summarize some of the selected insights discussed during the seminar concerning bots in software engineering. We conclude with a list of abstracts from the invited talks presented at the seminar.

4 Bots in Software Engineering: Insights and Outlook

4.1 Definitions: Software Bots and Software Bot Ecosystems

A discussion theme throughout the seminar involved defining what is or is not a bot. An agreed upon definition could assist researchers when designing, evaluating and conducting research that may involve or refer to bots.

From one breakout session, the following ideas emerged. First, it was seen as easier to define what is not a bot. Simple scripts and badges on GitHub were not considered as bots, whereas software that meets some or all of the following criteria may be seen as a bot:

- Automates one or more feature(s)
- Performs one or more function(s) that a human may do
- Interacts with a human or other agents

Furthermore, a bot may additionally exhibit these features:

- Acts autonomously as an independent actor or agent
- Appears intelligent (and may learn)
- Supports feedback loops
- Personifies a human or is human-like
- Has a name, and 2nd person pronoun
- Exhibits emotions and feelings

In a later breakout session, the concept of what is a **software bot** (in general terms, not just within the context of software engineering) was discussed once again but in the context of a faceted taxonomy developed by Carly Lebeuf as part of her Masters' thesis[1]. The definition of software bot given by Lebeuf is: "A software bot is an interface that connects users with software services and provides additional value to the user (by way of its interaction style, automation, and anthropomorphism)". Lebeuf claims it is the additional

[1] https://dspace.library.uvic.ca//handle/1828/10004

value that bots add on top of software's basic capabilities which distinguish bots from non-bot scripts and programs. Indeed, in this view, software bots are seen as special cases of software scripts and programs that bring this additional value in terms of automation, consolidation of multiple services, interaction mechanisms and anthropomorphic features. Chatbots and agents are more specific kinds of software bots, where chatbots are bots with natural language capabilities and agents are bots that can sense/act upon their environments and may be intelligent, autonomous and social.

The faceted taxonomy we discussed has three main dimensions: 1) Environment dimension; 2) Intrinsic dimension; and 3) Interaction dimension. To guide this discussion, we considered the Travis bot and how we could define Travis using this framework. This group discussion led to a number of suggested changes to the faceted taxonomy including refinement of some of the facets (such as scope, dynamism, and predictability) defined for the environment dimension. This framework was seen as a good start but would require much more input from a wider set of researchers with more diverse backgrounds (beyond a software engineering background) to agree on a new version of the framework. Revisiting the framework after the seminar is one of our future goals.

In another breakout we discussed the concept of **Bot Ecosystem**. We arrived at three different definitions as follows:

- A bot ecosystem is a set of bots working on the same or related projects
- A bot ecosystem is the set of APIs provided by a given platform supporting a set of bots using those APIs
- A bot ecosystem is a place where humans and bots can cooperate and communicate, enabled by conventions and tools

Throughout the seminar we also referred extensively to a list of software engineering bots being maintained by Mairieli Wessel[2].

4.2 Developer acceptance

A theme that emerged during the seminar as important was the acceptance of bots among developers, that is, what are important aspects that bots that must be satisfied for bots to be included in the daily work of developers. A breakout group identified a list of do's and don'ts:

- Don't: Repetitive notifications
- Don't: Wrong answers
- Don't: Hide the identity as a bot from the users
- Do: Automation tasks that users don't want to do
- Do: Provide actionable recommendations. Explain the reasoning.
- Do: Functionality for the user to perform the action (if possible).
- Do: Allow users to provide feedback and have bots learn over time.
- Do: Consistency in the task being done
- Do: Make bots more adaptive to user needs.
- Do: Be context aware.

[2] https://github.com/mairieli/awesome-se-bots

An important topic that the group discussed was **trust**. Bots need to build trust over time, for example, by focusing on trust initially rather than recall, not spamming the users with recommendations that are inaccurate. Context was mentioned as being important for trust. Making accurate recommendations can depend on individual developers, their experience and background, the current task, the rest of the team. This emphasizes the importance of adaptive bots that learn over time and are assessed with multi-dimensional benchmarks with respect to effectiveness. From an evaluation perspective, there are also many points of view, for example how effective bots are for individual developers, teams, the entire company, or even the society.

The group also identified several topics for follow-up discussions:

- Can bots eventually take over tasks from humans?
- What are the appropriate levels of intrusiveness?
- What about the maintenance cost of bots?
- What about exploration cost? Is there a tradeoff in the time that it takes to develop bots?
- How to integrate context and knowledge for bots?
- How to identify the right robot for the task?

4.3 To bot or not

Related to what is a bot or not, is the question of when to create a bot or not for a given use case. To explore this common design conundrum (and to also help understand more about what is or is not a bot) we brainstormed some use cases for which we could design bot and non-bot designs. This discussion set the stage for consideration of deeper challenges related to bringing multiple fields together.

The use cases we discussed in breakout groups were as follows: Potential aims of such an assistant included:

- Code review discussion support (to reduce noise in code reviews and related discussions)
- Privacy awareness support during development
- Pull request ranking support (provide support during pull request management)
- Contextual documentation (support for generating and assessing API documentation)
- Support to alleviate burnout and stress
- Transforming Data Science with Interactive Support for Feature Engineering and Model Adaptation
- Onboarding assistant (to lower the barriers to newcomers wishing to contribute to open source projects).

For example, in the breakout on a bot to support contextual documentation, discussion began by considering bots for documentation, and through this discussion specific challenges emerged about how software developers use documentation. Specifically, we focused on use cases from two perspectives: a) from a developer who is writing code to be used (API developers) and b) developers who are using or updating someone else's code (API users).

The functionalities of contextual documentation considered in the discussion included:

- Generating non-existing documentation from scratch
- Assessing documentation
- Modifying/improving/updating existing documentation (triggered by code updates)
- Generating of task-oriented documentation
- Asking refining questions

Throughout our documentation bot use case discussion we returned to two ideas. The first was identifying scenarios where bots were different from scripts. We resolved that this could be dictated by how they are triggered or even how they wait upon a user's reply. Waiting on a user's reply could be a function of communication styles of probing for information or waiting for another person or agent to prompt for dialogue.

The second open discussion we returned to several times was that the usefulness of the bot was something that mattered most, which begs the question of how bots are situated within a broader community and the roles and responsibilities of bots change the impact the roles and responsibilities of humans and their interactions with one another within that environment. More concretely, with the introduction of bots within the environment, who might be helped? Who might be displaced? How do individuals within the environment feel about their participation in the resulting sociotechnical system?

As a concrete example, we compared bot to non bot versions when trying to access documentation examples. In particular, we considered how getting relevant examples for how to use a new API has trade offs of copying an example from online vs. gaining offline social capital of discussing work in person.

When considering the big questions around what should be the roles and responsibilities of bots, and what should be the nature of their interaction with humans within the environment, difficult questions are raised. Some of these questions relate to ethical concerns, which are discussed in the next section. With the lack of clarity surrounding the extent of humanness desired, or what that humanness is meant to accomplish, or whether it is humanness itself that accomplishes those goals best, the difficult question was raised of what specific joint endeavors between the software engineering community and the NLP community would be mutually beneficial. A keynote talk and supplementary talk by David Traum offering an overview of the history of the field of dialogue systems and some pointers to available technologies provided some common ground to begin this substantive exploration. However, questions related to paths forward remain open, especially with regard to what new challenges for dialogue system technology would be interesting for researchers in NLP and would provide capabilities that would be valued within the field of software engineering going forward in the face of open questions.

4.4 Ethical concerns

One of the primary ethical concerns related to bots is whether the bots are allowed to impersonate humans. Such behavior might be considered unethical (as it might compromise privacy) and indeed it is explicitly prohibited by such platforms as Wikipedia. At the same time, previous studies (Murgia et al. 2016)[3] have shown the communities might be more likely to accept bots impersonating humans as opposed to bots disclosing that they are bots. During the seminar, participants identified additional advantages and concerns related to bots impersonating as humans:

- On the one hand, the primary advantage of designing bots impersonating humans is that humans know how to interact with other humans, so it might be easier for them to interact with human-impersonating bots. Moreover, by the same argument human-like communication might be more effective at communicating to developers and triggering

[3] Alessandro Murgia, Daan Janssens, Serge Demeyer, Bogdan Vasilescu: Among the Machines: Human-Bot Interaction on Social Q&A Websites. CHI Extended Abstracts 2016: 1272-1279

the desired action or response from them. Finally, the evolution from "humans using bots" to "egalitarian collaboration between humans and bots" is more likely to be achieved if the difference between humans and bots is less marked.

- On the other hand, bots posing as humans trigger a question of responsibility: who is to blame if the bot is "misbehaving"? Typically bots have owners that take responsibility of the bot's actions and can intervene in their behaviour if needed. If a bot is impersonating a human, there seems to be no obvious way to escalate or complain about poor behavior, or ultimately intervene. Furthermore, the question of impersonation is related to the ethics of lie: should the bots be allowed to pretend to be humans but be required to disclose their bot status if explicitly asked?

Another aspect related to ethics is the question of what cultural values are reflected in determining what kind of behaviour of bots should be considered ethical. Indeed, values of a software project that might be interested in adopting the bot, an ecosystem the project belongs to (e.g., "Python ecosystem") and the platform hosting the project or the ecosystem (e.g., GitHub/wikipedia) do not necessarily share the same cultural values.

Based on the aforementioned discussion, the seminar participants recommend that projects who use bots may adopt a *Manifesto of BOT ethics* that should include the following information:

- What behavior guidelines (e.g., contributor guidelines, code of conduct) are the bots required to follow?
- What rules pertain to bots impersonating humans:
  - Bots should not pose as humans
  - Or if asked directly it should respond
- What additional rules should the bots follow, e.g., what should the bots not do
  - Steal code
  - Modify with timeline/ project history
- What are the accountability rules for bots?
  - Who is accountable for the bot's actions?
  - Who is accountable/ have rights to materials created by Bots
  - When bots become autonomous and has a human persona, then what happens to its accountability?
- What cultural norms should the bots reflect in their behavior?

4.5 Diversity and inclusion

As diversity in software development teams is known to be beneficial for software development projects, the seminar participants discussed how bots can be used to encourage creation of a diverse and inclusive environment. During the discussion we identified several possible role for bots:

- Supporting newcomers. An example of a technical solution that can benefit newcomers would be a bot capable of identifying development tasks suited for newcomers or designing such tasks, e.g., by splitting more complex ones. Those microtasks might be better suited not only for novice developers but might also engage developers working on less powerful computing platforms such as mobile phones. Moreover, to facilitate onboarding bots can partner with newcomers working together on software development task: in this way bots can remove social stigma associated with asking for help. Finally, the bots can identify mentors.

- Support inclusive communication. Here bots can be used to implement a broad spectrum of actions, ranging from policing (e.g., checking for presence of toxic speech or code of conduct violations) to clarifying (e.g., flagging possible misunderstandings or clarifying communicative intentions), and monitoring the discussion to ensure that all voices are heard (e.g., reporting the percentage of the comments posted by the members of the dominant group).
- Support through personalisation: bots can be used to encourage self-reflection e.g., by applying mindfulness techniques, to deliver developers information in a way corresponding to their cognitive styles, or act as proxies of individual developers capable of communicating mental models to other bots/developers.

While the applications envisioned above focus on improving the software development process, bots can be also deployed to improve software products, e.g., by checking different facets of GenderMag[4].

At the same time, designers of bots should be aware that while aiming at encouraging diversity and inclusion, when not designed correctly bots might reinforce stereotypes and toxic cultures, appear patronizing or insincere, target minorities without changing community culture or increasing belonging, as well as shutting down communication channels by removing voices from the table that some people might find unsavory.

4.6 Bots to support collaboration

One important application of bots is to help developers collaborate effectively. Initially, several questions arose which made it difficult to understand the group's charter: What is a bot versus other sorts of tools, automation, scripts, etc.? Second, what do we mean by collaboration? For example, is the author of a library collaborating in some sense with a user who imports the library later, perhaps without the library author's knowledge? And third, is "collaboration support" too large a space to identify overarching research questions?

To make progress in the face of these threshold questions, the breakout identified three important dimensions along which such bots would likely deliver, and which might help distinguish importantly different types of bots and different research questions. There is a long and rich history in psychology about different forms of leadership, with two fundamentally different kinds of leaders providing very different ways of supporting group work. It was argued that while bots may not be group leaders, they could potentially help groups collaborate by supporting these two distinct functions. One dimension is Task-focused (helping the users complete the task faster/better) vs socio-emotional-focused (group well being, motivation, morale, cohesion). A second key dimension is collaboration in the large (e.g., Wikipedia, GitHub) versus collaboration in the small (people working at the same time and place). Finally, there seems to be an important distinction between "automation" agents that simply execute some standard task, versus dialog agents that interact with users to understand and be guided through some task, and support users through a natural language interface.

The group discussed how to understand the interactions between task focus versus socio-emotional focus for collaboration in the small versus collaboration in the large for dialog agents and for automation agents. The breakout concluded by identifying open questions and new ideas for bots.

[4] https://gendermag.org/

Open questions:

- Dialogue agents:
  - How to interpret user intent (opportunity: in SE this can be narrowed down to typical tasks in this area)
  - Bots may not be very generalizable (task-specific is too specific?)
  - Communication in the face of differences in vocabulary usage between subcommunities. That is, some commands / interactions may require specialized domain knowledge, that users may not have. How do bridge this gap?
- To what extent should we allow bots to act autonomously?
- To what extent do users trust bots?
- Rules of engagement – when should the bot jump in?
  - What's an appropriate role for a bot in a collaboration?
  - How to handle exceptions / unanticipated requests?
- Explainability / Transparency. Where are you now in the process? If you made an error, what were you trying?
- Bots are a socio-technical rather than technical system. Must be designed taking into account the human interaction
- How to coordinate code review checks (e.g., style guidelines) to maximize learning for PR submitter, rather than just providing a laundry list of things wrong with your contribution?
- How to **adapt to the user's background** (experience of the PR submitter)?

New Bot ideas:

- Canary releases: Bot goes through a checklist of things to check before releasing on a larger scale (building confidence)
- Ecosystem-level bots: Integrate information from the whole ecosystem (e.g., there is a new library available, how other people experienced the new library)
  - Which info to extract?
  - How to extract and aggregate it?
  - How to present it?
- Detect communication breakdown and intervene to prevent it. How can we design interventions to reduce the likelihood of these?

5 Follow up work

Following the vibrant discussions we had during the seminar, several collaborations were formed to continue research, such as the development of a software bot framework. Furthermore, a second edition of the BotSE workshop at ICSE is being planned for this May 2020 in South Korea[5]. Links to related documents from the seminar will be shared online. These links include more detailed notes from the breakout groups, bot design documents, a list of software bots to support software engineering and a bibliography.

[5] http://botse.org/

6 Overview of Invited Talks

6.1 Bots in Wikipedia – Brief review of selected research

Claudia Müller-Birn (Freie Universität Berlin, DE)

Over the last ten years, the Wikipedia community has gained manifold experiences in dealing with bots. They are primarily programmed to automate existing activities, for example, they inject data into Wikipedia content from public databases, monitor and curate Wikipedia content, extend Wikipedia's functionality, or protect from malicious activity. Wikipedia operates with a system of algorithmic governance that describes the interdependency between human user, bots and the technical infrastructure. From these experiences, we can provide a set of guidelines for the design or usage of bot in other settings, such as software engineering. Bots, for example, can be identified as such in Wikipedia. The community has developed a bot policy which is regularly adapted to the changing needs. Moreover, the community implemented a public approval procedure for bots. In summary, bots are always part of a social system, therefore should be treated as a socio-technical system and, therefore, designed as such.

6.2 Overview of Natural Language Dialogue Systems

David R. Traum (USC – Playa Vista, US)

Dialogue is defined as communication including multiple contributions, coherent interaction and multiple participants. An overview is presented of common types of automated dialogue systems that communicate with people in natural language. The most common types are task-oriented assistants and social chat, but also role-play of human roles and other types of systems have been built. A number of examples of different roles and systems filling these roles were presented as well as common dialogue system architectures. Finally research topics and resources for the area were introduced.

6.3 Highlights of the first International Workshop on Bots in Software Engineering (BotSE)

Emad Shihab (Concordia University – Montreal, CA) and Stefan Wagner (Universität Stuttgart, DE)

We organized the first workshop in the area of bots for software engineering as well as engineering bots this may in Montreal, Canada, co-located with ICSE. We reported on the workshop in general, the presentation topics as well as the discussions. We were impressed by the breadth of topics and noted that there was still an ongoing discussion about what a bot in SE is. There will be a BotSE again next year with ICSE.

6.4 Bots – The hidden side of software development

Bogdan Vasilescu (Carnegie Mellon University – Pittsburgh, US)

As automation agents, bots have become popular with the advent of the DevOps movement. In trying to maintain software quality and improve developer productivity while building software at a faster and faster pace, a myriad of automation agents have emerged; for continuous integration, testing, code coverage analysis, or dependency management, just to name a few. How do software engineers use such bots? Are the bots effective? Can we detect in archival data that bots are being used? And can we use such data to empirically study the effects of using bots?

In this talk I go over some recent results from my research group, focusing on how to detect bots in data from open-source software repositories, and what impact the introduction of automation agents has had on project outcomes. I also go over a recent experiment with a bit that answers programming questions automatically on Stack Overflow.

Participants

- Shivali Agarwal
IBM India – Bangalore, IN
- Ireti Amojo
Freie Universität Berlin, DE
- Ivan Beschastnikh
University of British Columbia – Vancouver, CA
- Kelly Blincoe
University of Auckland, NZ
- Nick Bradley
University of British Columbia – Vancouver, CA
- Fabio Calefato
University of Bari, IT
- Nathan Cassee
Eindhoven University of Technology, NL
- Jacek Czerwonka
Microsoft Research – GRedmond, US
- Antske Fokkens
Free University Amsterdam, NL
- Denae Ford
Microsoft Research – Redmond, US
- Thomas Fritz
Universität Zürich, CH
- Marco Gerosa
Northern Arizona University – Flagstaff, US
- Daniel Graziotin
Universität Stuttgart, DE
- Sonia Haiduc
Florida State University – Tallahassee, US
- James D. Herbsleb
Carnegie Mellon University – Pittsburgh, US
- Abram Hindle
University of Alberta – Edmonton, CA
- Akinori Ihara
Wakayama University, JP
- Minha Lee
Eindhoven University of Technology, NL
- Philipp Leitner
Chalmers University of Technology – Göteborg, SE
- Marin Litoiu
York University – Toronto, CA
- Walid Maalej
Universität Hamburg, DE
- Christoph Matthies
Hasso-Plattner-Institut – Potsdam, DE
- Marie-Francine Moens
KU Leuven, BE
- Martin Monperrus
KTH Royal Institute of Technology – Stockholm, SE
- Claudia Müller-Birn
Freie Universität Berlin, DE
- Nicole Novielli
University of Bari, IT
- Ayushi Rastogi
TU Delft, NL
- Paige Rodeghero
Clemson University, US
- Carolyn Penstein Rosé
Carnegie Mellon University – Pittsburgh, US
- Anita Sarma
Oregon State University – Corvallis, US
- Andreas Schreiber
German Aerospace Center – Köln, DE
- Alexander Serebrenik
Eindhoven University of Technology, NL
- Emad Shihab
Concordia University – Montreal, CA
- Arfon Smith
STSci – Baltimore, US
- Igor Steinmacher
Northern Arizona University – Flagstaff, US
- Margaret-Anne Storey
University of Victoria, CA
- David R. Traum
USC – Playa Vista, US
- Christoph Treude
University of Adelaide, AU
- Bogdan Vasilescu
Carnegie Mellon University – Pittsburgh, US
- Stefan Wagner
Universität Stuttgart, DE
- Mairieli Wessel
University of Sao Paulo, BR
- Jie Zhang
University College London, GB
- Thomas Zimmermann
Microsoft Corporation – Redmond, US

Report from Dagstuhl Seminar 19481

Composing Model-Based Analysis Tools

Edited by

Francisco Durán[1], Robert Heinrich[2], Diego Pérez-Palacín[3], Carolyn L. Talcott[4], and Steffen Zschaler[5]

1 **University of Málaga, ES,** duran@lcc.uma.es
2 **KIT – Karlsruhe, DE,** robert.heinrich@kit.edu
3 **Linnaeus University – Växjö, SE,** diego.perez@lnu.se
4 **SRI – Menlo Park, US,** clt@csl.sri.com
5 **King's College London, GB,** steffen.zschaler@kcl.ac.uk

Abstract

This report documents the program and the outcomes of the Dagstuhl Seminar 19481 "Composing Model-Based Analysis Tools". The key objective of the seminar was to provide more flexibility in model-driven engineering by bringing together representatives from industry and researchers in the formal methods and software engineering communities to establishing the foundations for a common understanding on the modularity and composition of modeling languages and model-based analyses.

Seminar November 24–29, 2019 – http://www.dagstuhl.de/19481
2012 ACM Subject Classification General and reference → General literature, General and reference
Keywords and phrases Modelling, Simulation, Semantics, Formal Methods, Software Engineering
Digital Object Identifier 10.4230/DagRep.9.11.97

1 Executive Summary

Francisco Durán
Robert Heinrich
Diego Pérez-Palacín
Carolyn L. Talcott
Steffen Zschaler

Quality properties like performance and dependability are key for today's systems. Several techniques have been developed to effectively model quality properties, which allow analyzing these systems. However, the very different nature of these properties has led to the use of different techniques and mostly independent tools. In addition, different tools and techniques can be used for modelling quality depending on the size and complexity of the systems and the available details. For example, for modeling dependability techniques like Fault Trees, Markov Chains, and Reliability Block Diagrams are available. Similarly, a range of analysis techniques are available, including simulations, using numerical, analytical or graphical techniques, and analytical methods.

Composing Model-Based Analysis Tools, *Dagstuhl Reports*, Vol. 9, Issue 11, pp. 97–116
Editors: Francisco Durán, Robert Heinrich, Diego Pérez-Palacín, Carolyn L. Talcott, and Steffen Zschaler

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Although it is worth exploring other techniques and methodologies, model-driven engineering (MDE) seems a promising technique to efficiently design and reason about behavior and quality of systems in various domains. Indeed, it has been very successfully applied to improve the efficiency of software development and analysis in various domains.

Moreover, recent innovations, like the Internet of Things, production automation, and cyber-physical systems, combine several domains such as software, electronics and mechanics. Consequently, also the analyses for each of these individual domains need to be combined to predictively analyze the overall behavior and quality. The composition of systems and their analyses is a challenging but unavoidable issue for today's complex systems. Existing MDE approaches to modeling and analysis are not sufficient to compose modular analyses combining domain-specific languages. First attempts towards composable modular models have been developed in recent years, attempting to compose, not only the structure of models and domain specific modeling languages (DSMLs), but also their dynamic aspects (behavior and semantics). These indeed may be good foundations for building composable modular analyses. However, much work remains ahead.

In this Dagstuhl Seminar, we target more flexibility in MDE by discussing how to modularize and compose models and analyses. This provokes questions from the theoretical computer science and formal methods community – for example, on validity, uncertainties, behavior and property protection/preservation/reflection, and termination of analyses. Traditionally, research on these topics is conducted in the formal methods community isolated from the MDE community. A key objective for bringing together representatives from industry and researchers in the formal methods and software engineering communities is to make progress towards establishing the foundations for a common understanding.

2 Table of Contents

Executive Summary
Francisco Durán, Robert Heinrich, Diego Pérez-Palacín, Carolyn L. Talcott, and Steffen Zschaler . . . 97

Introduction
Main Purpose of the Seminar . . . 101
Structure and Organization of the Seminar . . . 102
Viewpoint Talks . . . 104

Overview of Challenge Statements . . . 104

On the Relation between Language Composition and Analysis Composition 106

Orchestration of Analysis Tools . . . 108

Continuous Model Analysis (CMA) . . . 110

Creating Value from Analysis Results . . . 111

Composition of Models and Analysis affect Uncertainty . . . 112

Conclusions and Next Steps . . . 113

Participants . . . 116

3 Introduction

Quality properties like performance and dependability are key for today's systems. Ensuring these properties is a major concern for design engineers. Several techniques have been developed to effectively model and analyze characteristics like performance or failure of systems. However, the very different nature of these properties has led to the use of different techniques and mostly independent tools for their different aspects. For instance, while some of the properties (e.g., performance, reliability and availability) are quantitative, other ones (e.g., confidentiality and safety) are essentially qualitative.

Depending on the size and complexity of the system and the available details, there exists different techniques for modelling quality. For example, for modeling dependability techniques like Fault Trees [29], Markov Chains [11], and Reliability Block Diagrams [5] are available. Similarly, a range of techniques are available for dependability analysis, including simulations, using numerical, analytical or graphical techniques, and analytical methods. Several surveys have analyzed different aspects of the state of the art, including [17, 28, 7, 16, 25, 3, 19, 21, 1]. Although methods and procedures are not standardized for most industries and there are several open questions (exemplified in this report), known techniques both for the modelling and analysis are successfully used in cases such as defense, transportation, and space industries.

Where rigorous and precise methods are required, different formal methods have been used to provide mathematical reasoning, so that once the system's intended behavior is modelled, one can construct a proof that the given system satisfies its requirements. For dependability analysis, we can find proposals using Petri nets, model checking and higher-order logic theorem proving. See [1] for a recent survey on the use of formal methods for dependability modelling and analysis.

In model-driven engineering (MDE), models are created and applied to efficiently design and reason about behavior and quality of systems in various domains. For representing systems in the form of a model, a modeling language (for example defined through a metamodel) is required. Recent innovations like the Internet of Things, production automation and cyber-physical systems combine several domains such as software, electronics and mechanics. To successfully develop reliable systems for these contexts, analyses for the single domains need to be combined to estimate the overall behavior and quality. Usual simulation-based, analytical, or graph-based solvers can then be used to analyze models. An interesting case is the one of executable domain-specific modeling languages (xDSML), on which different techniques can be used for their analysis, including graph-based analyses, which has been extensively used in existing work [20].

The composition of systems and their analyses is a challenging but unavoidable issue for today's complex systems. However, existing approaches to modeling and analysis are not sufficient to compose modular analyses over xDSMLs. First attempts at composable modular models came up in recent years, attempting to compose, not only their structure (metamodels), but also their dynamic aspects (behavior) [8]. These indeed may be good foundations for building composable modular analyses. For example, a trace-based semantics of xDSMLs may lead to composable and decomposable traces, possibly inspired by existing notions such as trace slices or trace superposition.

Furthermore, since models are abstractions of reality, they are not a faithful representation of the system but they contain uncertainties [18, 30]. Identifying and handling these uncertainties is a challenge for the research community [10, 22] that is, at present, only partially addressed. The combination of models from different domains and usage perspectives

may exacerbate the effect of such uncertainties by creating, for instance, model inconsistencies, incoherence, mismatches in granularities of models, mismatches in the underlying assumptions made when creating the different models, etc. The study of the existence, quantification and management of the new uncertainties created during the combination of models is an unaddressed task that should be tackled to trust the results of the subsequent model analysis.

Our aim is to get more flexibility in MDE by discussing how to modularize and compose models and analyses based on modular models. Of course, composing modular analyses provokes questions from theoretical computer science and formal methods community – for example, on validity, uncertainties, behavior and property protection/preservation/reflection, and termination of analyses. Traditionally, research on these topics is mostly conducted in the formal methods community isolated from the MDE community. Bringing together representatives from industry and researchers in the formal methods and software engineering communities may lead to the establishment of the foundations for a common understanding.

3.1 Main Purpose of the Seminar

MDE has proven to be able to provide a good set of tools and techniques for the development of models and tools for the manipulation of these models. In the context of performance and dependability analysis, tools like Palladio, Modellica or AADL, are good examples of the possibilities. Indeed, a tool like Palladio already provides a modelling language for the analysis of performance, reliability and maintainability of systems [23]. However, it is a monolithic tool, making extension to new properties challenging. Thus, its internal structure eroded over time [27]. Furthermore, there is no way to verify the non-interference between the analyses provided. On the other hand, we have witnessed interesting advances in some of these issues, for example in the fields of graph rewriting, algebraic specification or tree automata. The seminar was organized with the believe that sharing specific problems and advances in some of these fields might lead to fruitful discussions, and possibly to new approaches and alternative views so some of these problems may be tackled.

With this goal in mind, we envision an environment in which xDSMLs for the different quality properties are independently provided, and where one can pick up the desired ones at will. In addition to being able to perform such a composition of models and analysis tools, the combined analysis performed by the composed system would allow us to analyze the tradeoffs between different properties (e.g., performance vs. security). Furthermore, we would like to be able to share the analysis effort between computation resources as much as possible. This led us to interesting questions relevant for research and industry like:

- Given the great costs of such analyses, if two different properties are analyzed using a graph-based simulation, how to combine the simulations so that the performance of the tool can be dramatically improved?
- Can analytical analyses performed on Markov chains or Petri nets be similarly composed?
- What are the limits of modular and composable model analysis?
- Is it possible to identify the uncertainties that spring from a composition of models? Is it possible to quantify the criticality of such uncertainties or inconsistencies? Is it possible to handle these possible uncertainties with robust methods so that the analysis still produces trustworthy results?

Composition of analyses over xDSMLs for different domains is an ambitious goal. In the past, it was assumed to be unattainable as models and analyses for different properties were considered to be too different. Recent research advances, however, lead to the conclusion that

commonalities can be identified, based on which foundational concepts can be elaborated as argued in detail by the following bullet points.

- Advances in behavior-parametrized modular specifications [9] together with advanced support for xDSMLs with graph-based operational semantics give reason for optimism in enabling the specification and composition of modular analyses.
- Research on formalization, measures, and metrics of single quality properties [2, 4, 12] resulted in much deeper and clearer understanding of what the corresponding analyses depend on, which is foundation for modularizing the analyses.
- Research on the analysis of mutual quality impact (e.g., performance [14] and maintainability [24, 13]) between different domains provides starting points for composing modular analyses.
- A first reference architecture for metamodels to tailor quality modeling and analysis [15] is starting point to more generic investigation on the topic.

3.2 Structure and Organization of the Seminar

The organization of a successful seminar poses a number of challenges, which are possibly consequence of:

(1) the selection of participants and their availability to participate,
(2) the attractiveness of the discussions and the topics around which these are going to happen along five days, and
(3) the involvement of the participants in such discussions and in the generation of summaries and documentation.

In order to organize the discussions during the seminar around common interests and challenges, before the seminar, participants were asked to share a short statement on their main interests (related to the topic under discussion). These challenges were analyzed and classified, and served as a starting point for the organization of the discussions around specific topics. The analysis of these challenges and how it led to the different working groups is explained in Section 4.

Although some of the participants knew each other before the seminar, we wanted them to introduce themselves to facilitate the interactions as soon as possible. To avoid spending too much time on introductions – 39 researchers participated in the seminar – we collected one slide from each participant with their core data and set up the presentation so that a new slide was on every two minutes. The presentations covered research interests, background, current research of the participant and topics to discuss at the seminar. Participants were asked to prepare their presentation slides before the seminar.

As discussed in Section 4, the analysis of the challenges, and the posterior discussion led to break-up groups in which the identified topics were discussed. The agenda for the week can be seen in Figure 1. The breakout groups were created to discuss these topics in smaller groups and create first results and plans for follow-up activities during the seminar. The breakout groups were suggested to produce paper projects and follow-up activities like workshop proposals or other community activities. Each breakout group started writing papers during the seminar using Overleaf, Google doc or other collaborative tools.

To share the discussions in the break-up groups with the rest of the participants, and to get feedback from them, there were presentations from each of the groups in the main room on the advances on their discussions. In addition to the consequent discussions that followed these presentations, and given the tight relationships between the different discussions,

<table>
<tr><th>Timeslot</th><th>Monday</th><th>Tuesday</th><th>Wednesday</th><th>Thursday</th><th>Friday</th></tr>
<tr><td>09:00-09:30</td><td>Opening and seminar objectives</td><td>Wrap up and planning of the day</td><td>Wrap up and planning of the day (and reorganizations of groups)</td><td>Presentation of Wednesday results, wrap up and planning of the day (and reorganizations of groups)</td><td>Seminar summary and planning for next steps</td></tr>
<tr><td>09:30-11:00</td><td>Introduction of participants</td><td>Breakout groups</td><td>Breakout groups</td><td>Cross-cut discussions</td><td>Seminar summary and planning for next steps</td></tr>
<tr><td>11:00-11:15</td><td colspan="5">Coffee break</td></tr>
<tr><td>11:15-12:15</td><td>Extended talks</td><td>Breakout groups</td><td>Breakout groups</td><td>Breakout groups</td><td>Seminar summary and planning for next steps / closing</td></tr>
<tr><td>12:15-13:30</td><td colspan="4">Lunch</td><td></td></tr>
<tr><td>13:30-13:45</td><td rowspan="2">Extended talks (13:30-14:00)

Introduction of participants (14:00-15:00)</td><td>Short wrap up of breakout groups</td><td rowspan="2">Excursion</td><td>Short wrap up of breakout groups</td><td rowspan="8">Departure</td></tr>
<tr><td>13:45-15:00</td><td>Breakout groups</td><td>Breakout groups</td></tr>
<tr><td>15:00-15:15</td><td colspan="4">Coffee break</td></tr>
<tr><td>15:15-17:00</td><td>Summary of submitted challenges (15:15-15:45)

Identification of topics for breakout groups and allocation of participants to groups (15:45-17:00)</td><td>Breakout groups</td><td>Excursion</td><td>Breakout groups</td></tr>
<tr><td>17:00-17:15</td><td colspan="4">Coffee break</td></tr>
<tr><td>17:15-18:00</td><td>Breakout groups</td><td>Presentation of results of the day</td><td>Excursion</td><td>Breakout groups</td></tr>
<tr><td>18:00-19:15</td><td colspan="4">Dinner</td></tr>
<tr><td>19:30-</td><td colspan="3">Come together</td><td>Tool demos</td></tr>
</table>

Figure 1 Seminar's schedule.

cross-cutting topics were identified in Thursday's first session (after a list of topics prepared by each group the previous day). These cross-cutting topics were discussed in the following session in groups in which there were representatives of each of the main break-up groups. These participants share these discussions back in their groups in the following break-up sessions.

Since the development of tools was present in many of the discussions, and many of the participants have themselves developed tools for the modelling and analysis of systems, a tool demo session was organized. This session was scheduled on the evening of the Thursday, and participants were asked to show their interest in delivering demos. The following tools were demonstrated:

- Shadow Models – incremental model transformation and lifting of error messages back to the original model, by Markus Voelter.
- FASTEN – a stack of DSLs and analyses for (formal) system level specification and assurance (SMV, tabular specification, contract-based design, UI-modeling, requirements specification, GSN, STPA), by Daniel Ratiu.
- MBEDDR – code-level analyses based on CBMC, Model-Driven Code Checking DSLs and analysis based on Spin, feature models consistency based on Sat4J, by Daniel Ratiu.
- ASMETA – a toolset for the Abstract State Machines (model specification, animation, simulation, verification, reviewing, ect), by Patrizia Scandurra and Elvinia Riccobene.
- The GEMOC Studio – a Language Workbench providing generic components through Eclipse technologies for the development, integration, and use of heterogeneous executable modeling languages, by Erwan Bousse, Steffen Zschaler and Benoit Combemale.
- GROOVE – a graph transformation tool for flexibly modelling any system whose states have a graph-like structure, and subsequently exploring and model checking the ensuing state space, by Arend Rensink.

- Palladio: A software architecture simulator for performance and reliability, by Robert Heinrich.
- Timed Rebeca Model Checking Tool – Afra, a tool for model checking and debugging a timed actor-based language, by Marjan Sirjani.
- Horus, for Business Process / Enterprise Modelling, by Arthur Vetter.
- AToMPM and Modelverse, by Hans Vangheluwe.

3.3 Viewpoint Talks

Although most participants might have common interests, each of us was approaching the general problem from different perspectives and using different techniques. Any of us could have told his/her story, but we did not want to spend the whole week with talks from the participants. We wanted to use the time to discuss and find new synergies between participants. We decided however to start sharing the focus with presentations from a selection of participants. From the pre-seminar challenge statements, we picked three of them that were providing three alternative views:

- Arthur Vetter, from KIT (Karlsruher Institut für Technologie), Germany, discussed on the evolution of systems and their analysis,
- Kenneth Johnson, from the Auckland University of Technology, New Zealand, focused on large scale complex systems, and shared his experience in the verification of large scale complex systems, with a cyberphysical perspective, and
- Fiona Polack, from Keele University, UK, presented an interesting discussion about *purpose*.

They were asked to deliver 15-minutes talks on their points of view. Moreover, since we were assuming that each of us works on different projects, for which we use different techniques, formalisms, and tools, they were specifically asked to avoid in their presentations details on these specifics, and try to provide a broader view from their specific perspectives. The goal was not about surveying on what can be done or how can be solved, that was the purpose of the rest of the week, we wanted to have a general discussion on what the problem is and what they would like to get as a result of the seminar.

4 Overview of Challenge Statements

Before the start of the seminar, all invited participants were asked to submit brief challenge statements, summarising what they felt were the key challenges in the area of composing model-based analysis tools. Overall, we received 27 such challenge statements. In preparation of the seminar, recurring themes from these challenge statements were clustered using a mindmapping technology. The resulting clusters can be seen in Figure 2. Through this exercise, we identified the following five top-level challenges:

1. *Composition of formalisms / languages.* Key sub-challenges here were mentioned as
   - Composition of semantics
   - Composition of different formalisms
   - Composition of analysis method vs composition of analysis results
   - Composition in space and time (variants vs versions)

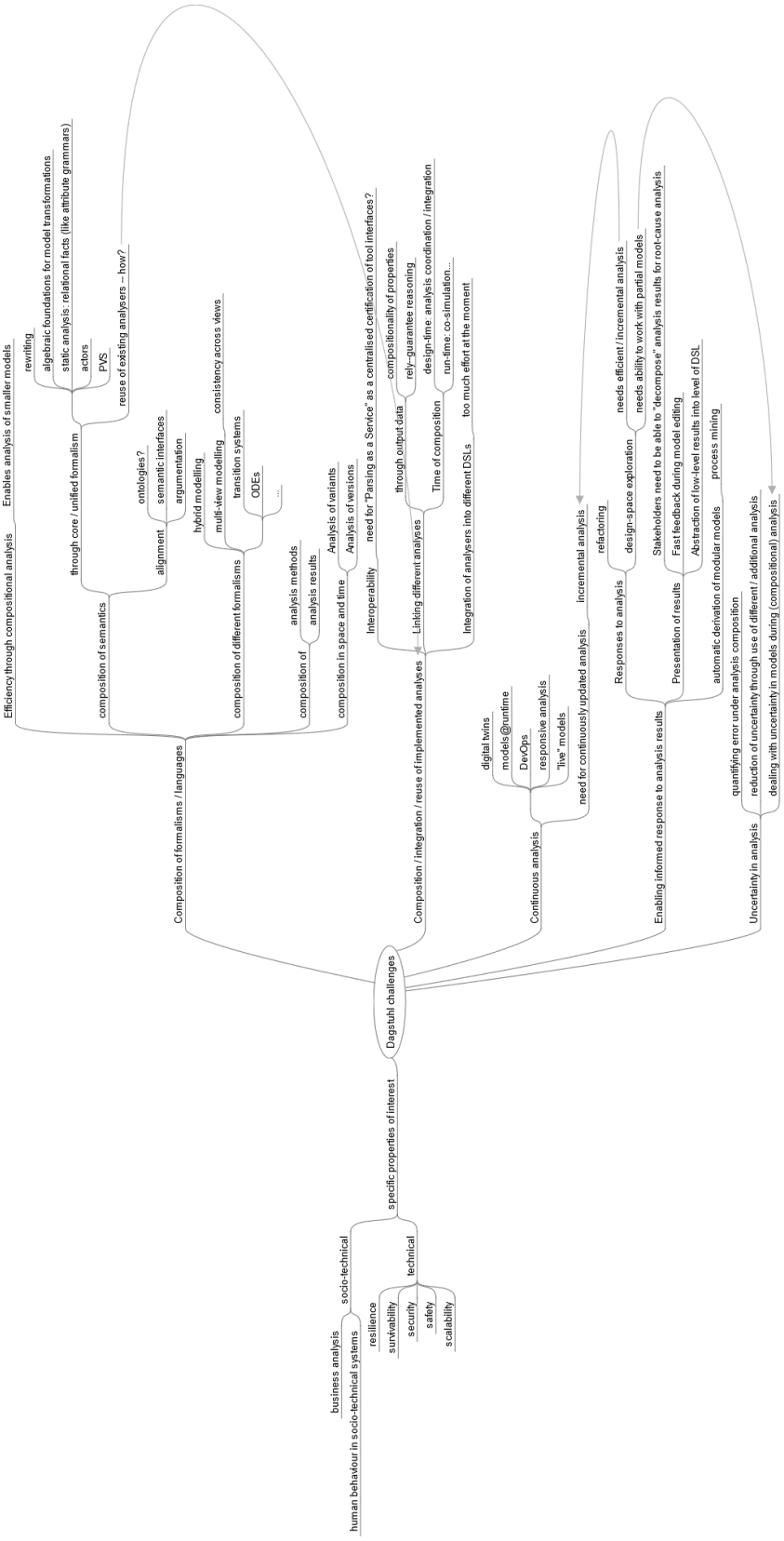

Figure 2 Mindmap of challenge clusters.

2. *Composition / integration / reuse of implemented analyses / tools.* Key sub-challenges here were mentioned as
   - Interoperability
   - Linking different analysers
   - Integration of analysers into different DSLs
3. *Continuous analysis.* Key sub-challenges here were mentioned as
   - Uses in different areas such as digital twins, models@runtime, DevOps, and responsive analysis
   - Achieving incremental analysis
4. *Enabling informed response to analysis results.* Key sub-challenges here were mentioned as
   - Presentation of results
   - Automatically obtaining modular models for efficient analysis
5. *Uncertainty in analysis.* Key sub-challenges here were mentioned as
   - Quantifying error under analysis composition
   - Combining different analyses to reduce overall uncertainty
   - Handling uncertainty / incompleteness in underlying models

To ensure we had correctly interpreted participants' challenge statements and to give an opportunity for all participants to contribute to the list of challenges, we undertook a separate clustering activity with all participants on the first day of the seminar. Here, participants were asked to write their key challenges on a post-it and to then physically cluster them on a pinboard. The resulting clustering can be seen in Figure 3. These clusters were very close to our original clustering and were used as the basis for the formation of breakout groups, with each group discussing one of the challenges in more depth. The following sections provide brief summaries of results provided by each group.

5 On the Relation between Language Composition and Analysis Composition

This group consisted of Carolyn Talcott, Ralf Reussner, Bernhard Rumpe, Hans Vangheluwe, Patrizia Scandurra, Kyungmin Bae, Séverine Sentilles, Narges Khakpour, Mark Hills, and Sofia Ananieva.

The group *On the Relation between Language Composition and Analysis Composition* was concerned with understanding the forms of compositions, especially the compositionality of analyses in relation to composition of underlying models and composition of the system. They also identified the different classes of compositionality and specific conditions of composition.

The group define analysis as answering questions about properties of a system under study. Next to the model of the system, they therefore also need a model of the context (actually the assumptions on the context) and a model of the property we want to analyse. For models, we follow the definition of Stachowiak [26] (i.e., abstraction, isomorphism, and pragmatics). Models can relate to each other via (a) *abstraction / refinement*, (b) *view projection / view merge* and (c) *architectural composition / decomposition*. They do not distinguish between modelling artifacts and models.

The group realised that considering the context of analysis is important for compositionality. Consequently, the definition of *what context is* depends on the kind of analysis. For structural (syntactic) analysis, the context is given by the meta-model / language definition. For behavioural analysis, the context is given through a semantic definition for the

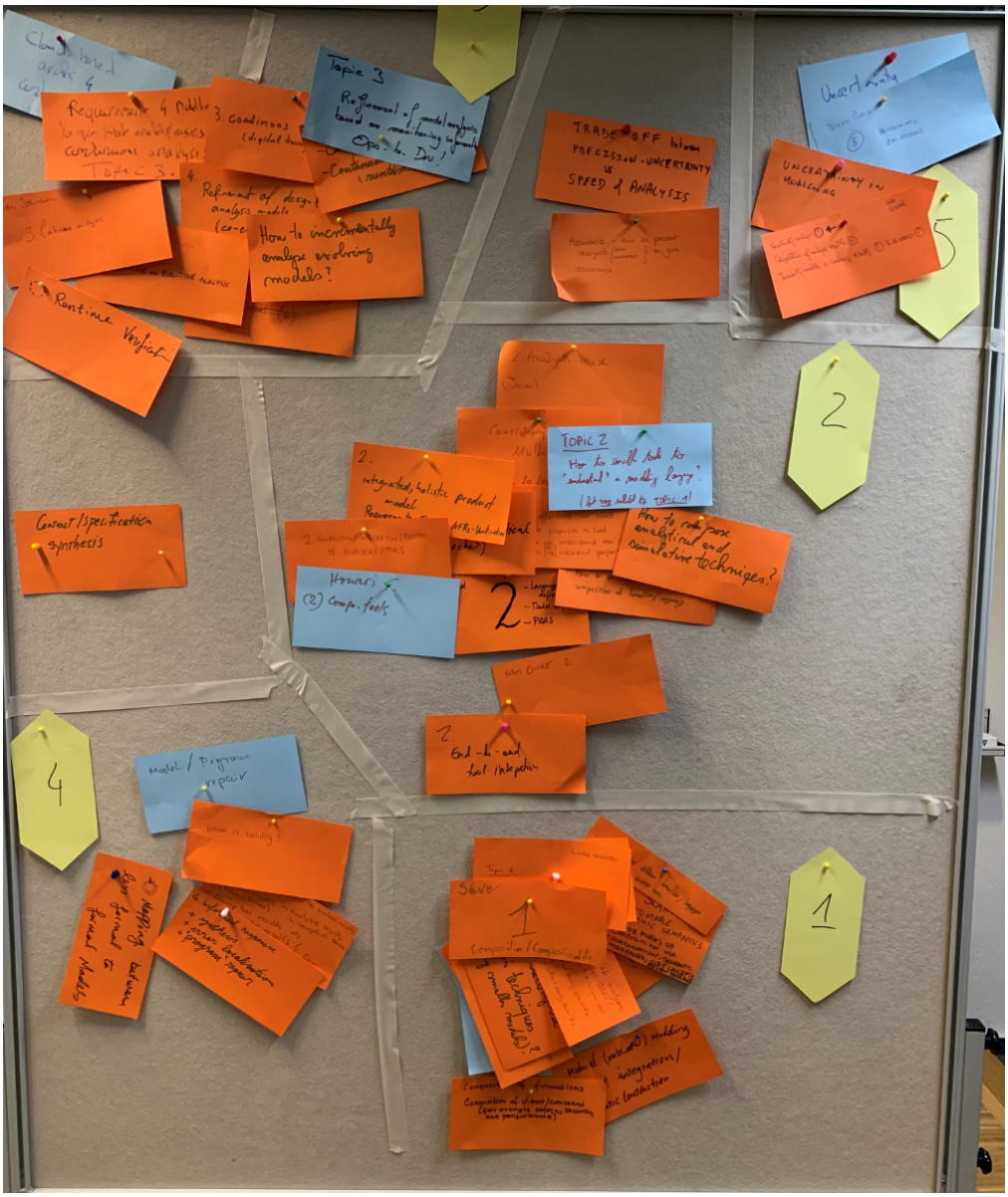

Figure 3 Clustering of participants cards.

system model and the specification of the system and its semantics. For the analysis of extra-functional properties (e.g., performance, reliability or security) the context is given through the model of usage profile, the model of the execution environment, and a model of the external services. If models of different semantic domains are involved, lifting of the analysis results back to the system model under manipulation is harder. Therefore, we need to understand the relationship between the involved contexts for composing analyses.

We require workflows to model the relation between activities using, changing and creating models. A workflow is a partially ordered set of basic activities (either human or computer based) or composed activities which take modelling artifacts as input and produce modelling artifacts as output. An activity can be refined into a workflow. The orchestrated execution of activities forms a workflow. Workflows can lead to variability in space (variants) and time (versions) for modelling artifacts.

Finally, we conceived three classes of composition:
1. System Model Composition (*white-box* composition)
2. Result Composition (*black-box* composition)
3. Analysis Composition (*grey-box* composition)

In analysis composition, we orchestrate the steps of the two analyses to be composed. Mathematically, the three cases are described as follows:
Let A be Analysis,
A_i be Analysis i,
A_i^j be the step j of Analysis i,
S_i be System model i,
C_i be Context model i,
Q_i be Question model i,
$\times$ be Composition operator, and
K be Orchestration model

System Model composition (white-box composition)
$A(S_1 \times S_2, C_1 \times C_2, Q_1 \times Q_2)$

Result composition (black-box composition)
$A_1(S_1, C_1, Q_1) \times A_2(S_2, C_2, Q_2)$

Analysis composition (grey-box composition)
$K((A_1^1(S_1, C_1, Q_1), A_1^2(S_1, C_1, Q_1), \ldots, A_2^1(S_2, C_2, Q_2), \ldots), C_1 \times C_2, Q_1 \times Q_2)$

6 Orchestration of Analysis Tools

This group consisted of Erwan Bousse, Robert Heinrich, Sandro Koch, Daniel Ratiu, Elvinia Riccobene, Markus Voelter, Marjan Sirjani, and Sam Owre.

Sophisticated engineering tools, often based on the principles of Model Driven Engineering (MDE) and Software Language Engineering (SLE), are becoming more and more ubiquitous– i.e., more and more disciplines rely on such tools. These tools become all the more valuable if they provide deep insights into the correctness or fitness for purpose of the created models. At the same time there is a community of analysis tool builders who distill mathematical and logic experience into analysis tools that rely on formalisms such as Satisfiability Modulo Theories (SMT) formulae, transition systems or discrete events. Many of these analyses can be used beneficially in the aforementioned engineering tools if they are suitable integrated. In practice this usually means that user-facing models must be translated to the input formalism of the analysis tool, and the results must be lifted back to the domain level. In addition, there are many use cases where multiple existing analyses must be orchestrated to deliver value in the context of the engineering tool.

The group's vision to tackle these challenges is to enable the definition of an architecture for the integration and or composition of existing analysis tools with a given domain-specific language. At the core of such an architecture is the orchestrator, a component both responsible for interacting with analysis tools, and for interacting with the domain expert willing to perform analyses. This orchestrator follows some orchestration logic that defines which analysis tools should be used for a given analysis task, in which order these tools should be used, and how the analysis results they produced should be combined or exchanged. The orchestrator relies on a set of tool drivers that each defines how to make use of a specific

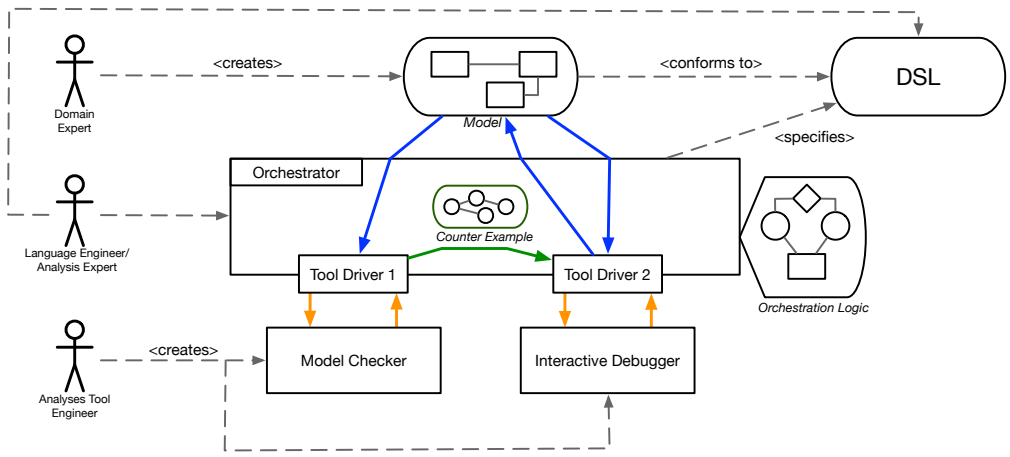

Figure 4 Vision applied to the same example.

analysis tool, including how to translate the domain-specific model into a valid input for the tool, how to lift back the analysis result into a form that makes sense at the abstraction level of the domain model, as well as the protocol to exchange messages and information with the tool. In addition, for such architectures to work, a set of requirements must be satisfied by the considered analysis tools: both the input and output languages of the tool but be explicitly defined (which excludes loosely structured output formats), and the protocol to use these tools must be explicitly defined and exploitable by the orchestrator. Figure 4 illustrates this overall vision with a simple case where a model checker is used to analyse a model, and where the counter-example produced by this model checker is injected in an interactive debugger for further investigation. In summary, the expected outcomes of such a research work are:

- A metamodel allowing one to define:
  - The overall architecture of a particular tool integration and composition scenario,
  - Tools drivers that each knows how to integrate the DSL with an analysis tool,
  - The orchestration logic (eg. the model is model checked when asked by the user, and the produced counter example (if any) is sent to an interactive debugger for investigation).
- A set of requirements for analysis tools, such as :
  - A tool must have a well-defined and explicit input language,
  - A tool must have a well-defined and explicit output language,
  - A tool must have a well-defined and explicit protocol.
- A set of case studies that demonstrates the relevance and applicability of the abstractions defined in this paper.

7 Continuous Model Analysis (CMA)

This group consisted of Christel Baier, Olivier Barais, Kenneth Johnson, Dániel Varró, Arthur Vetter, Marc Zeller.

The current pressure to ensure that software systems remain available, dependable, preferment at all times despite changes in their operating context, the addition or evolution of features, the increasing integration with other systems has led to the implementation of so-called continuous deployment techniques (i.e. DevOps) and self-healing mechanisms. Behind such mechanisms lies the need to continuously analyse a system against a number of properties. The combined use of a number of model-based analysis tools on abstract representations (models) of a running system has therefore become common.

Several challenges arise then:

1. How to understand and reason about this composition of analysis tools?
2. How to orchestrate these analyses?
3. How to minimize the analyses to be performed each time there is a change in the system specification, in the context of the system's execution or on the system itself?
4. How to validate compositionality of CMAs?
5. Could we use the same analysis tool at runtime and at design time?
6. What happens when we have uncertain knowledge of the system? What is the minimal set of information to carry out reusable analysis?

The fields of application are very diverse:

- Incremental verification of system component models at runtime
- DevOps pipeline
- Safety-critical systems development
- Self-adaptive systems

The working group sought to define a general conceptual framework for continuous model analysis. They keep to simple mathematical concepts to describe behaviours and relationships between modules. Notions of timing and change are important.

A key challenge is to formulate and validate compositionality of CMAs

CMA modules:

- System meta-model (the set S)
- System meta-model Requirements (the set Q)
- Models (Models of systems, requirements) (the set M) and their properties (the set P)
- Analysis tasks (the set A)
- Results output from the analysis (the set V)
- Monitoring and notifications from the actual system + requirements

Mathematically,

- $\alpha : S \to M$ modelling process from system meta-models to their semantic meaning. Will be input for analysis
- $\beta : Q \to P$ modelling process from system meta-model requirements to their properties. Will be input for analysis
- α and β are used in many ways to form analysis tasks in A. Let the process be modelled by $\gamma : [S \to M] \times [Q \to P] \to A$. This may simply be $a = ((s), (q))$ where verification analysis task a is comprised of a model and property.
- The analysis is modelled by $\rho : A \to V$ which takes as input an analysis task and computes some sort of result value(s) in V

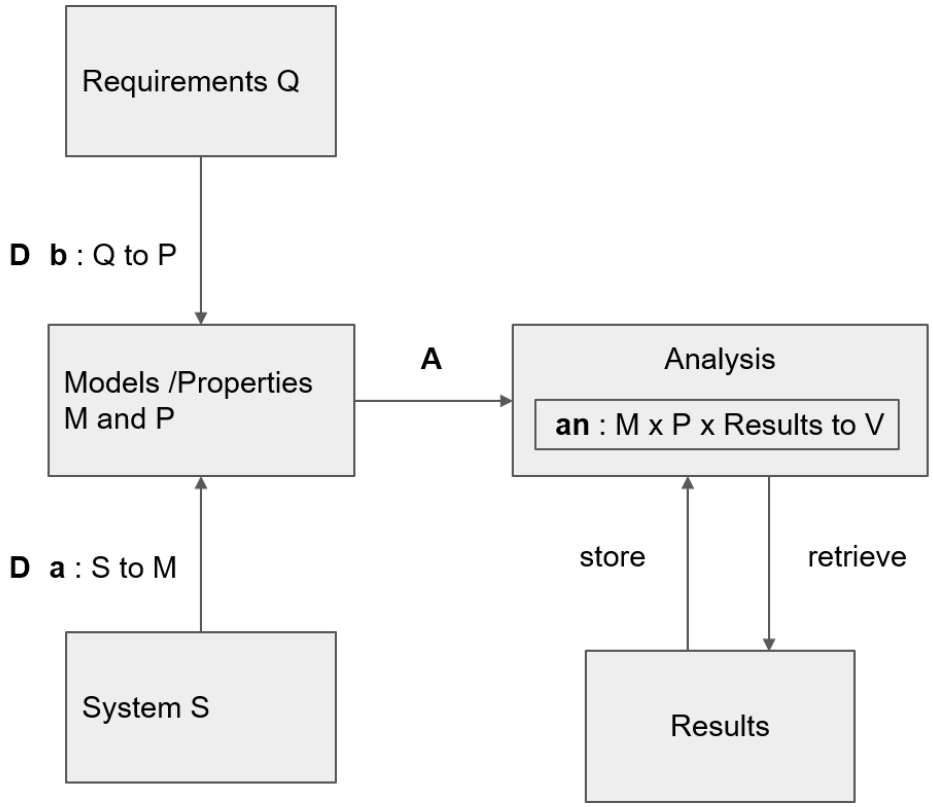

Figure 5 General conceptual framework for continuous model analysis (CMA).

- Note that we need to extend ρ to take as input all existing analysis tasks and results. We can extend this by $\rho : [A \to V] \times A \to V$

A key technical challenge is to minimise the amount of computation that needs to be performed.

8 Creating Value from Analysis Results

This group consisted of Steffen Zschaler, Houari Sahraoui, Esther Guerra, Martin Gogolla, Francisco Durán, and Juan de Lara.

The group focused on activities taking place once analysis has been completed: how are analysis results turned into value for the users of the analysis? Discussions were driven by diverse example cases of specific analyses, ranging from static analysis of programs written by novice programmers via business-process deadlock analysis to analysis of object churn. As a result of discussing these different analyses, the group identified a conceptual model of three pathways for result usage:

1. *Result presentation.* Analysis results can first be used to help developers or domain experts explore the system model further—for example to identify the root cause of a problem or to better understand a scientific hypothesis. Challenges the group identified in this pathway include:
   a. Lifting low-level analysis results back up to the domain level so that they can be understood by domain stakeholders.

   b. Selecting what analysis results are most important / useful to present in a given situation–this is closely connected to the original purpose for which the analysis was undertaken.
   c. Enabling users to drill down, possibly interactively, into the analysis results–for example to undertake root-cause analysis.
   d. Interpretation of the results by the end user–this may develop over time as users learn to correlate result presentation and their own understanding of the system and its properties.
2. *Result exploitation.* Analysis results can secondly be exploited to directly improve systems or their specifications / models. Examples of this include model / program repair, refactoring, or refinement. The changes to a system or its specification, in response to some analysis, can be done automatically (*e.g.,* search-based software engineering) as well as manually. Challenges identified in this pathway include:
   a. How far can this be automated for different properties of interest?
   b. Is there a generic automatic mechanism or does each property require its own mechanism?
   c. Can we learn automated exploitation mechanisms by observing how expert domain users respond to different types of analysis results?
   d. Is it possible to undertake repair or similar in relation to multiple properties of interest at the same time (*i.e.,* can repair be composed)?
3. *Analysis improvement.* Analysis results can be used to improve the analysis itself. For example, by asking it to focus on a particular aspect of the system model in more detail or by learning an analysis from a set of expert-provided examples of inputs and expected results. Challenges identified in this pathway include:
   a. How to enable users to understand the analysis results and provide suitable feedback to the learning algorithm (see also Pathway 2)?
   b. How to model such feedback so that it can be effectively used for improving the analysis?
   c. How to automate the learning process; to which extent is this even possible?

Overarching these pathways, there is a challenge of how to choose properties, analysis pathways, and combinations thereof to form an overall argument of fitness-for-purpose of the system as a whole. Goal-Question-Metric, safety cases, goal-oriented modelling (e.g., [6]) appear to have building blocks for answering this challenge, but as far as the group members were aware there is currently no integrated approach.

The group then proceeded to identifying similarities and differences between the example cases and how they covered each of the three pathways. This resulted in a detailed feature model, which can serve as the starting point of a systematic survey of analysis techniques, planned as one of the next steps to be undertaken by the group.

9 Composition of Models and Analysis affect Uncertainty

This group consisted of Simona Bernardi, Michalis Famelis, Jean-Marc Jézéquel, Raffaela Mirandola, Diego Perez-Palacin, Fiona Polack, and Catia Trubiani.

This group discussed the modelling and management of uncertainty, beginning by revisiting the various definitions and taxonomies of uncertainty in the literature. Following that, they discussed how uncertainty is localized in modelling and analysis, identifying four phases: model Definition, model Construction, QoS Analysis, and Validation.

To exercise the existence and location of uncertainties in the four phases, they constructed a simple example consisting of a set of models for a file sharing system based on a Peer-To-Peer protocol (PtPp). The group elicited uncertainty in the different modeling views of the protocol (state, class, deployment, and object diagrams, performance models) and discussed the effect of composition to the various uncertainties. To better understand this, they created a model (flow chart) of the process of development and performance analysis of PtPp to get awareness of the occurrence of uncertainty over time.

After that, the group explored, using a few scenarios, how uncertainty flows through the process model, also speculating that this kind of analysis could lead to a re-conceptualization of DevOps as an iterative approach for the reduction of some types of uncertainty.

In the context of the seminar at large, the group realized that there is a need for more popularisation in our community of existing theories, taxonomies, nomenclature about uncertainty.

The group agreed that more research is needed in deepening the understanding of the occurrence and evolution of uncertainty in the creation of systems, especially by exploring uncertainties in different kinds of quality analysis such as reliability, availability, and security. Finally, they planned joint publications and further academic events, starting with the submission of a Vision paper and a Workshop proposal to the MODELS 2020 conference, respectively.

10 Conclusions and Next Steps

This report summarized the structure, organisation and outcome of the Dagstuhl seminar 19481. We reported about discussions, group work and results of the seminar. Before the seminar, participants were asked to share a short statement on their main interests from which we identified topics for breakout groups. The breakout groups were created to discuss these topics in smaller groups and create first results and plans for follow-up activities during the seminar. We had five breakout groups that worked on the topics relation between language composition and analysis composition, orchestration of analysis tools, continuous model analysis, creating value from analysis results, and composition of models and analysis affect uncertainty, respectively.

The breakout groups individually produced plans for paper projects and follow-up activities like workshop proposals or other community activities. As a joint result of the seminar we agreed to produce an edited book summarising the discussions at the seminar, bringing together the thinking of the community, and demonstrating, through case studies, some of the important challenges and exemplary solutions in the field.

References

1 Waqar Ahmad, Osman Hasan, and Sofiène Tahar. Formal dependability modeling and analysis: A survey. In Michael Kohlhase, Moa Johansson, Bruce R. Miller, Leonardo de Moura, and Frank Wm. Tompa, editors, *9th International Conference Intelligent Computer Mathematics, CICM*, volume 9791 of *Lecture Notes in Computer Science*, pages 132–147. Springer, 2016.

2 Steffen Becker, Lucia Happe, Raffaela Mirandola, and Catia Trubiani. Towards a methodology driven by relationships of quality attributes for QoS-based analysis. In Seetharami Seelam, Petr Tuma, Giuliano Casale, Tony Field, and José Nelson Amaral, editors, *ACM/SPEC International Conference on Performance Engineering, ICPE*, pages 311–314. ACM, 2013.

3 Simona Bernardi, José Merseguer, and Dorina C. Petriu. Dependability modeling and analysis of software systems specified with UML. *ACM Comput. Surv.*, 45(1):2:1–2:48, 2012.

4 F. Brosch, H. Koziolek, B. Buhnova, and R. Reussner. Architecture-based reliability prediction with the palladio component model. *IEEE Transactions on Software Engineering*, 38(6):1319–1339, 2012.

5 Marko Čepin. *Reliability Block Diagram*, pages 119–123. Springer, 2011.

6 Lawrence Chung, Brian A. Nixon, Eric Yu, and John Mylopoulos. The Kluwer international series in software engineering. Kluwer Academic Publishers Group, Dordrecht, Netherlands, 1999.

7 Liliana Dobrica and Eila Niemelä. A survey on software architecture analysis methods. *IEEE Trans. Software Eng.*, 28(7):638–653, 2002.

8 Francisco Durán, Antonio Moreno-Delgado, Fernando Orejas, and Steffen Zschaler. Amalgamation of domain specific languages with behaviour. *J. Log. Algebr. Meth. Program.*, 86(1):208–235, 2017.

9 Francisco Durán, Fernando Orejas, and Steffen Zschaler. Behaviour protection in modular rule-based system specifications. In Narciso Martí-Oliet and Miguel Palomino, editors, *Recent Trends in Algebraic Development Techniques, 21st International Workshop, WADT, Revised Selected Papers*, volume 7841 of *Lecture Notes in Computer Science*, pages 24–49. Springer, 2013.

10 Naeem Esfahani and Sam Malek. Uncertainty in self-adaptive software systems. In Rogério de Lemos, Holger Giese, Hausi A. Müller, and Mary Shaw, editors, *Software Engineering for Self-Adaptive Systems II - International Seminar, Dagstuhl Castle, Germany, October 24-29, 2010 Revised Selected and Invited Papers*, volume 7475 of *Lecture Notes in Computer Science*, pages 214–238. Springer, 2013.

11 W. Gilks. *Markov chain Monte Carlo*. 2005.

12 Robert Heinrich. *Aligning Business Processes and Information Systems – New Approaches to Continuous Quality Engineering*. Springer, 2014.

13 Robert Heinrich, Sandro Koch, Suhyun Cha, Kiana Busch, Ralf Reussner, and Birgit Vogel-Heuser. Architecture-based change impact analysis in cross-disciplinary automated production systems. *Journal of Systems and Software*, 146:167 – 185, 2018.

14 Robert Heinrich, Philipp Merkle, Jörg Henss, and Barbara Paech. Integrating business process simulation and information system simulation for performance prediction. *Software and Systems Modeling*, 16(1):257–277, 2017.

15 Robert Heinrich, Misha Strittmatter, and Ralf Heinrich Reussner. A layered reference architecture for metamodels to tailor quality modeling and analysis. *IEEE Transactions on Software Engineering*, 2019.

16 Anne Immonen and Eila Niemelä. Survey of reliability and availability prediction methods from the viewpoint of software architecture. *Software and Systems Modeling*, 7(1):49–65, 2008.

17 Allen Johnson and Miroslaw Malek. Survey of software tools for evaluating reliability, availability, and serviceability. *ACM Comput. Surv.*, 20(4):227–269, 1988.

18 M. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society*, (63):425—464, 2001.

19 S. Shanmugavel L. Venkatesan and C. Subramaniam. A survey on modeling and enhancing reliability of wireless sensor network. *Wireless Sensor Network*, 5(3):41–51, 2013.

20 Antonio Moreno-Delgado, Francisco Durán, Steffen Zschaler, and Javier Troya. Modular dsls for flexible analysis: An e-motions reimplementation of palladio. In Jordi Cabot and Julia Rubin, editors, *Modelling Foundations and Applications – 10th European Conference,*

ECMFA 2014, volume 8569 of *Lecture Notes in Computer Science*, pages 132–147. Springer, 2014.

21 K. Pardeepkumar, P. Dahmani, and D. Narula. RAM analysis of some process industries: A critical literature review. *Int. J. Mech. Eng. & Rob. Res.*, 3(3), 2014.

22 Diego Perez-Palacin and Raffaela Mirandola. Uncertainties in the modeling of self-adaptive systems: a taxonomy and an example of availability evaluation. In Klaus-Dieter Lange, John Murphy, Walter Binder, and José Merseguer, editors, *ACM/SPEC International Conference on Performance Engineering, ICPE*, pages 3–14. ACM, 2014.

23 Ralf H. Reussner, Steffen Becker, Jens Happe, Robert Heinrich, Anne Koziolek, Heiko Koziolek, Max Kramer, and Klaus Krogmann. *Modeling and simulating software architectures: The Palladio approach*. MIT Press, 2016.

24 Kiana Rostami, Robert Heinrich, Axel Busch, and Ralf H. Reussner. Architecture-based change impact analysis in information systems and business processes. In *2017 IEEE International Conference on Software Architecture, ICSA*, pages 179–188. IEEE Computer Society, 2017.

25 S. Saraswat and G. Yadava. An overview on reliability, availability, maintainability and supportability (RAMS) engineering. 25(3):330–344, 2008.

26 Herbert Stachowiak. *Allgemeine Modelltheorie*. Springer, 1973.

27 Misha Strittmatter, Georg Hinkel, Michael Langhammer, Reiner Jung, and Robert Heinrich. Challenges in the evolution of metamodels: Smells and anti-patterns of a historically-grown metamodel. In Tanja Mayerhofer, Alfonso Pierantonio, Bernhard Schätz, and Dalila Tamzalit, editors, *Proceedings of the 10th Workshop on Models and Evolution*, volume 1706 of *CEUR Workshop Proceedings*, pages 30–39. CEUR-WS.org, 2016.

28 Kishor Trivedi and Manish Malhotra. Reliability and performability techniques and tools: A survey. In B. Walke and O. Spaniol, editors, *Messung, Modellierung und Bewertung von Rechen- und Kommunikationssystemen*, pages 27–48. Springer, 1993.

29 W. E. Vesely, F. F. Goldberg, N. H. Roberts, and D. F. Haasl. Fault tree handbook. Technical Report NUREG-0492, U.S. Nuclear Regulatory Commission, 1981.

30 W. Walker, P. Harremoës, J. Rotmans, J. van der Sluijs, M. van Asselt, P. Janssen, and M. Krayer von Krauss. Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support. *Integrated Assessment*, 4(1):5–17, 2003.

Participants

- Sofia Ananieva
FZI – Karlsruhe, DE
- Kyungmin Bae
Postech – Pohang, KR
- Christel Baier
TU Dresden, DE
- Olivier Barais
IRISA – University of Rennes, FR
- Simona Bernardi
University of Zaragoza, ES
- Erwan Bousse
University of Nantes, FR
- Benoit Combemale
University of Toulouse, FR
- Juan De Lara
Autonomous University of Madrid, ES
- Francisco Durán
University of Málaga, ES
- Michalis Famelis
Université de Montréal, CA
- Martin Gogolla
Universität Bremen, DE
- Esther Guerra
Autonomous University of Madrid, ES
- Robert Heinrich
KIT – Karlsruhe, DE
- Mark Hills
East Carolina University – Greenville, US
- Jean-Marc Jézéquel
IRISA – University of Rennes, FR
- Kenneth Johnson
Auckland University of Technology, NZ
- Narges Khakpour
Linnaeus University – Växjö, SE
- Sandro Koch
KIT – Karlsruhe, DE
- Raffaela Mirandola
Polytechnic University of Milan, IT
- Sam Owre
SRI – Menlo Park, US
- Diego Pérez-Palacín
Linnaeus University – Växjö, SE
- Fiona A. C. Polack
Keele University – Straffordshire, GB
- Daniel Ratiu
Siemens AG – München, DE
- Arend Rensink
University of Twente, NL
- Ralf H. Reussner
FZI – Karlsruhe, DE
- Elvinia Riccobene
University of Milan, IT
- Bernhard Rumpe
RWTH Aachen, DE
- Houari Sahraoui
Université de Montréal, CA
- Patrizia Scandurra
University of Bergamo – Dalmine, IT
- Séverine Sentilles
Mälardalen University – Västerås, SE
- Marjan Sirjani
Mälardalen University – Västerås, SE
- Carolyn L. Talcott
SRI – Menlo Park, US
- Catia Trubiani
Gran Sasso Science Institute, IT
- Hans Vangheluwe
University of Antwerp, BE
- Dániel Varró
McGill University – Montreal, CA
- Arthur Vetter
KIT – Karlsruher Institut für Technologie, DE
- Markus Völter
Völter Ingenieurbüro – Stuttgart, DE
- Marc Zeller
Siemens AG – München, DE
- Steffen Zschaler
King's College London, GB

Report from Dagstuhl Perspectives Workshop 19482

Diversity, Fairness, and Data-Driven Personalization in (News) Recommender System

Edited by

Abraham Bernstein[1], Claes De Vreese[2], Natali Helberger[3], Wolfgang Schulz[4], and Katharina A. Zweig[5]

1 **Universität Zürich, CH,** bernstein@ifi.uzh.ch
2 **University of Amsterdam, NL,** c.h.devreese@uva.nl
3 **University of Amsterdam, NL,** n.helberger@uva.nl
4 **Universität Hamburg, DE,** w.schulz@hans-bredow-institut.de
5 **TU Kaiserslautern, DE,** zweig@cs.uni-kl.de

Abstract

As people increasingly rely on online media and recommender systems to consume information, engage in debates and form their political opinions, the design goals of online media and news recommenders have wide implications for the political and social processes that take place online and offline. Current recommender systems have been observed to promote personalization and more effective forms of informing, but also to narrow the user's exposure to diverse content. Concerns about echo-chambers and filter bubbles highlight the importance of design metrics that can successfully strike a balance between accurate recommendations that respond to individual information needs and preferences, while at the same time addressing concerns about missing out important information, context and the broader cultural and political diversity in the news, as well as fairness. A broader, more sophisticated vision of the future of personalized recommenders needs to be formed–a vision that can only be developed as the result of a collaborative effort by different areas of academic research (media studies, computer science, law and legal philosophy, communication science, political philosophy, and democratic theory). The proposed workshop will set first steps to develop such a much needed vision on the role of recommender systems on the democratic role of the media and define the guidelines as well as a manifesto for future research and long-term goals for the emerging topic of fairness, diversity, and personalization in recommender systems.

Seminar November 24–29, 2019 – http://www.dagstuhl.de/19482

2012 ACM Subject Classification Information systems → Information retrieval diversity, Applied computing → Psychology, Human-centered computing → Empirical studies in HCI, Applied computing → Sociology, Information systems → Digital libraries and archives, Human-centered computing → HCI theory, concepts and models, Applied computing → Economics, Information systems → Web services

Keywords and phrases News, recommender systems, diversity

Digital Object Identifier 10.4230/DagRep.9.11.117

Diversity, Fairness, and Data-Driven Personalization in (News) Recommender System, *Dagstuhl Reports*, Vol. 9, Issue 11, pp. 117–124
Editors: Abraham Bernstein, Claes De Vreese, Natali Helberger, Wolfgang Schulz, and Katharina A. Zweig

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Abraham Bernstein (Universität Zürich, CH)
Claes De Vreese (University of Amsterdam, NL)
Natali Helberger (University of Amsterdam, NL)
Wolfgang Schulz (Universität Hamburg, DE)
Suzanne Tolmeijer (Universität Zürich, CH)
Katharina A. Zweig (TU Kaiserslautern, DE)

The Dagstuhl Perspectives Workshop 19482 on Diversity, Fairness, and Data-Driven Personalization in (News) Recommender Systems,[1] took place from November 24 to November 29 at Schloss Dagstuhl in Germany. The goal of the workshop was to bring together researchers from the various disciplines relevant to news recommender systems (computer, communications, legal, and political science) to (1) develop a joint understanding of the issues arising for society with regards to the diversity and fairness of recommender systems, (2) identify the gaps in science, practice and regulation with regards to these topics, and (3) to compile a set of recommendations–in the form of a manifesto–that outlines needed steps from all actors involved to address the societal issues at hand.

Workshop Schedule

The workshop was organized in the following phases:

Welcome and introductions This first phase introduced the workshop goal to the participants and then offered each of them five minutes to introduce their research activities, expertise, their interest in the topic, and research directions they see as relevant to the workshop's topic.

Impulse presentations Given the diversity of the backgrounds of the participants, eight brief stage setting presentations where given. The goal of these was to establish a common ground in terms of relevant questions and common vocabulary.[2]

Topical breakout group discussions Based on the introducing presentations and impulse presentations, the next phase of the workshop was organized around topical breakout groups. Topics discussed included relating fairness to diversity, user desiderata and characteristics, wider societal implications, governance, data requirements, and clustering of research gaps.

Writing sessions The next phase was focused on jointly drafting the manifesto that incorporated recommendations developed from discussions so far and compiling them into a coherent document.

The remainder of this text provides the abstracts of the impulse presentations. The insights resulting from our discussions can be found in the manifesto document, which will be published in due course.

[1] See workshop home page at https://www.dagstuhl.de/19482
[2] Brief abstracts of these talks can be found in this document.

2 Table of Contents

Executive Summary

Abraham Bernstein, Claes De Vreese, Natali Helberger, Wolfgang Schulz, Suzanne Tolmeijer, and Katharina A. Zweig 118

Overview of Talks

Bringing Diversity to News Recommender Algorithms
Abraham Bernstein 120

News Recommender Systems (NRS) – A communication science perspective
Claes De Vreese 120

Algorithmic Accountability and Fairness – A computer scientist's perspective
Marc Hauer 121

Democratic theory and Recommendations
Natali Helberger 121

Legal media policy
Wolfgang Schulz 121

Toward Measuring Viewpoint Diversity in News Consumption
Nava Tintarev 122

Measuring diversity in news recommendations – Or, at least, an attempt
Sanne Vrijenhoek 123

Computer science perspective: Measures as models of society
Katharina A. Zweig 123

Participants 124

3 Overview of Talks

3.1 Bringing Diversity to News Recommender Algorithms

Abraham Bernstein (Universität Zürich, CH)

Joint work of Abraham Bernstein, Bibek Paudel, Suzanne Tolmeijer
Main reference Bibek Paudel, Fabian Christoffel, Chris Newell, Abraham Bernstein: "Updatable, Accurate, Diverse, and Scalable Recommendations for Interactive Applications", TiiS, Vol. 7(1), pp. 1:1–1:34, 2017.
URL https://doi.org/10.1145/2955101

Recommender systems have become a backbone of consumption. They combine information about items and previous behavior of users to personalize the user's experience when reading the news, buying goods, or choosing what to watch in the evening. This talk issuccinctly introduces how recommender systems work to establish the technical underpinnings for all workshop attendees and suggests various approaches for how diversity can be added to them as as an additional target measure.

3.2 News Recommender Systems (NRS) – A communication science perspective

Claes De Vreese (University of Amsterdam, NL)

Main reference Nicholas Diakopoulos: "Towards a Design Orientation on Algorithms and Automation in News Production", Digital Journalism, Vol. 7(8), pp. 1180–1184, Routledge, 2019.
URL https://doi.org/10.1080/21670811.2019.1682938
Main reference Judith Möller, Damian Trilling, Natali Helberger, Bram van Es: "Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity", Information, Communication & Society, Vol. 21(7), pp. 959–977, Routledge, 2018.
URL https://doi.org/10.1080/1369118X.2018.1444076
Main reference Neil Thurman, Seth C. Lewis, Jessica Kunert: "Algorithms, Automation, and News", Digital Journalism, Vol. 7(8), pp. 980–992, Routledge, 2019.
URL https://doi.org/10.1080/21670811.2019.1685395

In this Introduction talk, NRS are contextualized as part of a larger development towards the role of data and automated decision making both in the production, dissemination, and consumption of news, in a changing media ecosystem. It is highlighted that communication science research often focuses on the user and effects on the user, but that in the space of NRS there is still a relative paucity of empirical research in this area. Recent publications have called for more attention to the design and features of NRS and the implications for user agency and effects on users' knowledge, attitudes and behavior. The diversity notion has been central in communication science for decades, and there is a clear need to expand diversity research in NRS beyond topical diversity to also include medium, device, outlet and content (e.g., tone, frame, actors) diversity. The talk concludes with a number of emerging topics in communication science research on NRS, such as the role of conversational agenda, NRS and platforms like YouTube, the role of NRS in journalistic production routines, and the potentially unintended consequences on diversity in NRS.

3.3 Algorithmic Accountability and Fairness – A computer scientist's perspective

Marc Hauer (TU Kaiserslautern, DE)

The talk gave a short outline about three notions of algorithmic accountability, the Algorithm Accountability Lab of TU Kaiserslautern is currently working on, namely how to assign responsibilities in the development of ADM-systems, the various and incompatible measures of fairness, and a regulation approach that has been included into the final report of the German Datenethikkomission.

References

1 Alexander Filipociv, Christopher Koska, Claudia Paganini. Ethik für Algorithmiker. Bertelsmann Stiftung, 2018, https://doi.org/10.11586/2018033.
2 Tobias D. Krafft, Katharina A. Zweig. Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse. Bundesverband Verbraucherzentrale, 2019, https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22_zweig_krafft_transparenz_adm-neu.pdf.

3.4 Democratic theory and Recommendations

Natali Helberger (University of Amsterdam, NL)

Main reference Natali Helberger: "On the Democratic Role of News Recommenders", Digital Journalism, Vol. 7(8), pp. 993–1012, Routledge, 2019.
URL https://doi.org/10.1080/21670811.2019.1623700

The argument that this presentation made is that diversity in the media is a concept with a mission: to further the values democratic societies are grounded in. Building on a brief discussion of four selected democratic theories of the media (liberal, participatory, deliberative and critical/antagonistic) and the growing body of literature about the digital turn in journalism, the presentation offered a conceptual framework for assessing the threats and opportunities around the democratic role of news recommenders. The talk concluded with developing a typology of different "democratic recommenders".

3.5 Legal media policy

Wolfgang Schulz (Universität Hamburg, DE)

The talk gives a legal perspective on diversity and the recent challenges for the concept. In German broadcasting regulation (like in many other jurisdictions) "diversity" appears as a main goal, meaning – according to the Federal constitutional court – that diversity of existing opinions should be presented in broadcasting as broadly and comprehensively as possible. In consequence, Public service broadcasters are required to promote diversity,

media regulators govern the distribution of broadcasting programs to maximize diversity. The recent draft of an amended Interstate Treaty on Broadcasting tries to extend diversity regulation to intermediaries. They should not discriminate among pieces of media content. However, regulatory concepts reach their limits if they want to apply diversity regulation to media in an information ecosphere where media content is just one among many types that also fulfil information needs of the users.

3.6 Toward Measuring Viewpoint Diversity in News Consumption

Nava Tintarev (TU Delft, NL)

Main reference Dimitrios Bountouridis, Jaron Harambam, Mykola Makhortykh, Mónica Marrero, Nava Tintarev, Claudia Hauff: "SIREN: A Simulation Framework for Understanding the Effects of Recommender Systems in Online News Environments", in Proc. of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019, pp. 150–159, ACM, 2019.
URL http://dx.doi.org/10.1145/3287560.3287583

The growing volume of digital data stimulates the adoption of recommender systems in different socioeconomic domains, including news industries. While news recommenders help consumers deal with information overload and increase their engagement and satisfaction, their use also raises an increasing number of societal concerns, such as "Matthew effects", "filter bubbles", and an overall lack of transparency. Considerable recommender systems research has been conducted on balancing diversification of content with relevance, however this work focuses specifically on topical diversity. For readers, diversity of *viewpoint* on a topic in news is however more relevant. This allows for measures of diversity that are multi-faceted, and not necessarily driven by previous consumption habits. This talk introduced preliminary work together with several Dutch news organizations (e.g., Blendle, Persgroep, and FDMediagroep), aiming to find ways to help users explore viewpoint diversity. This talk also explored transparency for content-providers, and introduced a simulation framework that allows content providers to (i) select and parameterize different recommenders and (ii) analyze and visualize their effects with respect to two diversity metrics. Consequently, this talk introduced first steps toward informing diverse content selection in a way that is meaningful and understandable, to both content providers and news readers.

References

1 Nava Tintarev, Emily Sullivan, Dror Guldin, Sihang Qiu, and Daan Odjik. "Same, same, but different: algorithmic diversification of viewpoints in news". In UMAP workshop on Fairness in User Modeling, Adaptation and Personalization, in association with UMAP'18. 2018.
2 Dimitrios Bountouridis, Jaron Harambam, Mykola Makhortykh, Monica Marrero, Nava Tintarev, and Claudia Hauff. "SIREN: a simulation framework for understanding the effects of recommender systems in online news environments". In ACM Conference on Fairness, Accountability, and Transparency (FAT*). 2019.
3 Feng Lu and Nava Tintarev. "A diversity adjusting strategy with personality for music recommendation". In Recsys workshop on Interfaces and Decision Making in Recommender Systems. 2018.

3.7 Measuring diversity in news recommendations – Or, at least, an attempt

Sanne Vrijenhoek (University of Amsterdam, NL)

Joint work of Sanne Vrijenhoek, Nadia Metoui, Judith Möller, Daan Odijk, Natali Helberger
URL https://github.com/svrijenhoek/dart/tree/master/dart

The University of Amsterdam, in collaboration with RTL News and funded by the SIDN Fonds, has started the development of an open source tool that enables data scientists at media companies to measure diversity in their news recommendations. In this talk we describe the setup of this project and the process of bridging the gap between normative notions of diversity, founded in democratic theory, and computationally viable methods. We identified a set of metrics approaching a subset of characteristics of different models of democracy, and evaluate them by comparing performance between a set of baseline recommender approaches.

References

1 Helberger, Natali. *"On the democratic role of news recommenders."* Digital Journalism (2019): 1-20.

3.8 Computer science perspective: Measures as models of society

Katharina A. Zweig (TU Kaiserslautern, DE)

Main reference Katharina Anna Zweig: "Network Analysis Literacy – A Practical Approach to the Analysis of Networks", Springer, 2016.
URL https://doi.org/10.1007/978-3-7091-0741-6
Main reference Isadora Dorn, Andreas Lindenblatt, Katharina A. Zweig: "The Trilemma of Network Analysis", in Proc. of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), ASONAM '12, p. 9–14, IEEE Computer Society, 2012.
URL https://doi.org/10.1109/ASONAM.2012.12

The talk first reviewed the idea of using centrality indices in complex network analysis and provided a solution of why there are so many of them. This is explained by a suggestion of Borgatti who stated that for every network flow process there is one centrality index that predicts which of the nodes is most heavily used by the network flow process. He characterized network flow processes by only a few characteristics, e.g., the type of paths used in the network or the distribution mode. Thus, each centrality index is tied to a network flow process and vice versa. In other words, centrality indices contain a *model* of a social process to which they can be applied to. This well-understood relationship between a certain class of indices or measures and a social process can be generalized to all kinds of operationalizations of social concepts, e.g., diversity of a news recommender system. If all measures and indices that are supposed to quantify a social term contain a model of a social process or a cultural perspective, it is 1) important to make these implicit assumptions as explicit as possible and 2) vital to only apply any measure to those kind of data and research questions that match with the implicit assumptions.

Participants

- Christian Baden
The Hebrew University of Jerusalem, IL
- Michael Beam
Kent State University, US
- Abraham Bernstein
Universität Zürich, CH
- Claes De Vreese
University of Amsterdam, NL
- Marc Hauer
TU Kaiserslautern, DE
- Lucien Heitz
Universität Zürich, CH
- Natali Helberger
University of Amsterdam, NL
- Pascal Jürgens
Johannes Gutenberg-Universität Mainz, DE
- Christian Katzenbach
Institute for Internet & Society – Berlin, DE
- Benjamin Kille
TU Berlin, DE
- Beate Klimkiewicz
University Jagiellonski – Krakow, PL
- Wiebke Loosen
Universität Hamburg, DE
- Judith Möller
University of Amsterdam, NL
- Goran Radanovic
MPI-SWS – Saarbrücken, DE
- Wolfgang Schulz
Universität Hamburg, DE
- Guy Shani
Ben Gurion University – Beer Sheva, IL
- Nava Tintarev
TU Delft, NL
- Suzanne Tolmeijer
Universität Zürich, CH
- Wouter van Atteveldt
VU University Amsterdam, NL
- Sanne Vrijenhoek
VU University Amsterdam, NL
- Theresa Züger
Institute for Internet & Society – Berlin, DE
- Katharina A. Zweig
TU Kaiserslautern, DE