DAGSTUHL
REPORTS

Volume 13, Issue 6, June 2023

Dagstuhl Reports, Vol. 13, Issue 6 ISSN 2192-5283

**ISSN 2192-5283**

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

*Publication date*
February, 2024

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

Digital Object Identifier: 10.4230/DagRep.13.6.i

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

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

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

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

Report from Dagstuhl Seminar 23241

# Scalable Analysis of Probabilistic Models and Programs

**Sebastian Junges*[1], Joost-Pieter Katoen*[2], Scott Sanner*[3], Guy Van den Broeck*[4], and Bahare Salmani†[5]**

1 **Radboud University Nijmegen, NL.** sebastian.junges@ru.nl
2 **RWTH Aachen, DE.** katoen@cs.rwth-aachen.de
3 **University of Toronto, CA.** ssanner@gmail.com
4 **UCLA, US.** guyvdb@cs.ucla.edu
5 **RWTH Aachen, DE.** salmani@cs.rwth-aachen.de

## Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 23241 "Scalable Analysis of Probabilistic Models and Programs". The seminar brought together researchers from probabilistic graphical models, verification of probabilistic programming languages, and probabilistic planning. The communities bring vastly different perspectives on the methods and goals of inference under uncertainty. In this seminar, we worked towards a common understanding of how the different angles yield subtle differences in the problem statements and how the different methods provide different strengths and weaknesses. The report describes the different areas, the activities during the seminar including hot topics that were vividly discussed, and an overview of the technical talks.

**Seminar** June 11–16, 2023 – https://www.dagstuhl.de/23241
**2012 ACM Subject Classification** Computing methodologies → Planning under uncertainty; Computing methodologies → Probabilistic reasoning; Theory of computation → Automated reasoning
**Keywords and phrases** model counting, probabilistic inference, probabilistic model checking, probabilistic planning, probabilistic programs
**Digital Object Identifier** 10.4230/DagRep.13.6.1

## 1 Executive Summary

*Sebastian Junges*
*Joost-Pieter Katoen*
*Scott Sanner*
*Guy Van den Broeck*

In this Dagstuhl Seminar, we brought together researchers from three different communities that all bring their own perspective on the role of probabilities in programs and models. To facilitate future collaborations, we saw a need to define common terminology. Therefore, we split this seminar into two parts: On the first two days, we surveyed the research areas (see Chapter 3) and on the second half, we had in-depth sessions. In the first half, we particularly discussed the following exemplary questions:

* Editor / Organizer
† Editorial Assistant / Collector

Scalable Analysis of Probabilistic Models and Programs, *Dagstuhl Reports*, Vol. 13, Issue 6, pp. 1–21
Editors: Sebastian Junges, Joost-Pieter Katoen, Scott Sanner, and Guy Van den Broeck
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

- *What are probabilistic circuits and why do they allow tractable inference?*
- *What is probabilistic model checking and what are temporal specifications?*
- *What are probabilistic programs and how are their semantics defined?*
- *How does one reason over probabilistic programs on the source-code level?*

We also spent time to identify common interests in problems, which led to some hot topics mentioned below. The second half of the seminar featured 30-minute technical talks (see Section 4) that provided in-depth discussions on recent research and informal discussions based on the technical talks and the hot topics.

**Hot discussion topics.** Our discussions led to a list of seven hot topics, summarized below, that were the basis for informal discussions later in the week.

- Unbounded executions in probabilistic programs, their use-cases, analysis techniques, and the downsides.
- Algebraic decision diagrams versus probabilistic circuits and their benefits for reasoning about reachability in graphs.
- The expressiveness and tractability borders between different model types.
- Adequate planning horizons in different scenarios and their influence on the effectiveness of various approaches.
- Inferring symbolic plans and policies via reinforcement learning and with logical constraints, in particular in the context of providing verifiable and explainable controllers.
- The limits of Boolean synthesis in the context of model counting.
- Learning models from data across communities, including perspectives on inference and active automata learning.
- Probabilistic models as distribution transformers and the verification of distributional properties.

**In-depth technical sessions.** We wanted to highlight the many in-depth discussions that happened mostly 24/7. These discussions were both on the hot topics listed above, as well as very technical, in-depth discussions. We believe that part of the success of this seminar were the various connections that were established on very technical levels. It became clear that the prevalent approaches for very similar problems are vastly different and that there was a common interest to learn about these methodologies.

**Challenges.** While we are proud of what we achieved, the different backgrounds required a significant effort in order to understand the problems that the different subcommunities find most pressing. As organizers, we would have loved to see time and room to also discuss application areas, but it was hard to find cross-community overlapping interests in those.

## 2 Table of Contents

# 3 Introduction

*Sebastian Junges, Joost-Pieter Katoen, Scott Sanner, Guy Van den Broeck*

Models and model-based reasoning allow reasoning about symbolic knowledge about a system. It is often convenient to represent such systems with probabilistic behavior. Such probabilistic behavior is a natural way to treat uncertainty or to abstract from behavior that appears probabilistic but is a consequence of complex interactions. Reasoning about these systems requires treating the probabilistic aspects as first-class citizens. Famously, humans are bad at reasoning under probabilistic uncertainty and thus the availability of scalable engines that support humans in understanding probabilistic systems and making decisions is essential. Naturally, many disciplines in and outside computer science investigate methods that lead to such engines.

A key challenge in the creation of such reasoning engines is a concise and clear modelling language. Historically many of the tools we had for reasoning and inference about the world were built on top of deterministic programming languages that pose a challenge for the representation of stochastic systems. Probabilistic programming – the notion that programs execute stochastically – and the analysis of such programs have caused a major shift and inference for probabilistic programming languages has already enabled various applications: They steer autonomous robots, are at the heart of the NET-VISA system adopted by the UN to detect seismological activities [2], are used in cognitive science [36], planning in AI, describe security mechanisms and naturally code up randomised algorithms. Probabilistic programming is a rapidly emerging field [18]. For almost all programming languages, there is a probabilistic variant by now, and main industrial players (e.g., Meta and Microsoft) have spent serious research efforts.

A recent trend goes towards thinking about any stochastic model as given by a probabilistic program (PP), which can be thought of as a computational specification of a probability distribution. Probabilistic programs can be used to describe complex distributions and the standard inference task is to understand this distribution. PPs can explicitly represent conditioning as part of a model by syntactic constructs that enable conditioning. PPs with such observation statements involve solving the inverse problem of inferring (the likelihood of) inputs matching a given evidence (aka: observation).

However, the analysis of PPs is and remains hard. Elementary questions such as "does a program terminate almost surely on a given input?" are undecidable. Lifting existing inference techniques to the level of programs is difficult, and reasoning about loops is harder than for classical programs. Techniques to analyse PPs in a symbolic and fully automated manner are currently a vibrant research topic and receive lots of attention in the various research fields, most notably: *probabilistic graphical models*, *probabilistic model checking/program analysis*, as well as *planning in AI*. This Dagstuhl Seminar brought together members of these communities to taxonomize existing research domains and progress in terms of a probabilistic programming lingua franca, identify areas for cross-pollination of ideas across fields, and understand major representational, inferential, and domain-specific challenges that the community (and groups of researchers from this seminar) may collectively tackle beyond the seminar.

## The Research Areas

**Probabilistic Graphical Models.** These models cover Bayesian networks, undirected Markov networks, and extensions thereof dealing with e.g., relational data. They have numerous applications in machine learning, computer vision, natural language processing, and computational biology. Graphical models bring together graph theory and probability theory, and provide a flexible framework for modeling large collections of random variables with complex interactions. Although exact inference is PP-complete in general, efficient algorithms exist for specific structures (e.g., join tree) and dedicated symbolic data structures have been developed to make inference efficient. Parameter and structure learning techniques enable synthesising values in conditional probability tables and full graph topologies. Recent work in exact discrete inference with probabilistic graphical models focuses on tractable models, where marginal probabilities, or the mode of the distribution, can all be computed efficiently in the size of the model. In particular, probabilistic circuits in the form of sum-product networks, arithmetic circuits, and other data structures, have received considerable attention in recent years [13]. Such new probabilistic models provide an opportunity to re-imagine the types of analysis that are possible, for example towards information-theoretic queries [44]. Another important frontier is to discover larger classes of distributions where the probabilistic inference analysis can be performed efficiently [46]. Probabilistic graphical models, either classical ones, or more modern tractable models, are often the target representation of compilers that start with a higher-level language – for example a probabilistic program [17, 24, 35], or even a quantum program [25].

**Verification of Probabilistic Models: Model Checking and Beyond.** Probabilistic model checking [4, 3, 27] is a verification technique that takes a probabilistic model and a (temporal) specification and asks whether the model satisfies the specification. The (underlying) models are typically Markov chains, Markov decision processes, or stochastic games, state-based models describing how the system evolves over time. These models are pivotal in various disciplines that involve process analysis such as performance evaluation, and sequential decision making, e.g., in reinforcement learning and robotics. To overcome the state-space explosion problem that limits the scalability of PMC algorithms, additional structure must be exploited. This structure can be found in the symbolic descriptions of models and are often defined using programs. The scalability of PMC is then boosted using symbolic data structures, but also clever model reduction techniques [32, 19, 43], and iterative abstraction techniques [22, 28, 6]. Specifications range from the more classical reachability queries and temporal logics to cost-bounded [20, 7], conditional [5] and multi-objective queries [11, 12]. Mature tools such as PRISM [31] and Storm [21] are applicable to finite-state probabilistic programs and are not limited to just the verification question. They can compute how much a specification is satisfied, extract strategies that satisfy the specification, and handle unknown probabilities. Model checkers are used often as back-ends to find plans: in the context of robotics, e.g., in [33] or to synthesise probabilistic programs [14, 1]. Beyond model checking, automated verification techniques for infinite-state Markov models such as bounded model checking, termination analysis (e.g., using deep neural networks), loop analysis, and $k$-induction have recently been made.

**Probabilistic Planning in AI.** (Classical) planning aims to find a policy or strategy to solve a decision making process in problems that can range from navigational path planning [34] to supply chain logistics [37] to the management of epidemic outbreaks [45]. Just as in model checking (MC), one is interested in plans that achieve a goal (in MC, to find a bug) or more generally to minimize a cumulative cost function or bound thereon. Contrary to MC, one is less interested in proving the absence of a solution (or bug). Despite these minor differences, there has been successful cross-fertilization partially initiated at earlier Dagstuhl Seminars. The need for planning under uncertainty has led to probabilistic planning methods that explicitly take this uncertainty into account. These plans (or policies) are typically computed within the Markov Decision Process (MDP) framework, and probabilistic planners are able to successfully find strategies in large MDPs using techniques such as lifting [41, 16] on Monte Carlo Tree Search with dead-end detection [29] as witnessed by International Probabilistic Planning Competitions (IPCs) [15, 42]. The probabilistic planning and probabilistic model checking communities have diverged in their symbolic representations with RDDL [40] being the quasi-standard in probabilistic planning for the object-oriented specification of concurrently evolving stochastic systems (which was in turn inspired by a lifted perspective of dynamic Bayesian networks). Furthermore, the planning community has long embraced partially observable settings as evidenced by partially observed MDP tracks of the IPC [15], whereas such extensions have only recently gained traction in the model checking community. Finally, it is important to note that probabilistic planning typically focuses on a class of algorithms particularly tailored for reachability analysis, namely highly scalable heuristic search techniques and with specialized domain analysis to support them [29].

## The Seminar Topics

In *Scalable Analysis of Probabilistic Models and Programs*, the aim is the development of methods, algorithms, and tools that reason in and about probabilistic uncertainty. Such reasoning is interesting in a variety of scientific areas both inside and outside of computer science. But how can we fundamentally boost the reasoning engines and make them more applicable beyond our own communities?

**Joint Context.** In the planning, and verification communities, program-like descriptions of models have a rich tradition (RDDL [40], Prism [31], Modest [8], JANI [10]). Compared to more general programming languages, these models typically have a variety of tailored but intricate constructs. In recent years, probabilistic programs have been heavily studied as a natural extension of probabilistic graphical models. These probabilistic programs are easier to learn, but a naive user may not obtain the necessary performance out of their inference engines. The holy grail are engines that are efficient in reasoning about easy-to-use probabilistic programs. While goals and perspectives differ across the research communities, there are plenty of similarities in automated analysis techniques in the aforementioned three research fields. For instance, symbolic techniques such as binary decision diagrams (BDDs) and satisfiability solver (SAT/SMT) techniques are commonly used and model counting has made substantial progress in recent years.

**Challenges.** While (general-purpose) probabilistic programming languages that extend (classical) programming languages are already a success for a significant class of applications, their analysis remains challenging: In particular, we highlight *the presence of discrete variables, conditional program flow, non-deterministic behavior, and unbounded loops.* Within the AI, planning, programming languages and verification community, various efforts aim to improve the analysis of probabilistic programs, all from their own perspective.

Despite this strong link between analysis in probabilistic planning and reachability analysis in PMC, the research in general, and the development of new algorithms, has happened largely independently in each community. However, first cross-fertilizations between verification and inference [38] and vice versa [23], and probabilistic planning and PMC [9, 30] have been established and show the potential of overcoming research community boundaries. Beyond standard inference, methods still seem orthogonal but may be unified: e.g., probabilistic program sketching approaches [1] seem orthogonal to structure learning techniques [26] for Bayesian networks and parameter synthesis in Markov models have potential for parameter synthesis in graphical models [39].

### References

1 Roman Andriushchenko, Milan Ceska, Sebastian Junges, Joost-Pieter Katoen, and Simon Stupinský. PAYNT: A tool for inductive synthesis of probabilistic programs. In *CAV (1)*, volume 12759 of *Lecture Notes in Computer Science*, pages 856–869. Springer, 2021.

2 Nimar S. Arora, Stuart Russell, and Erik Sudderth. NET-VISA: Network processing vertically integrated seismic analysis. *Bulletin of the Seismological Society of America*, 103(2A):709–729, 2013.

3 Christel Baier, Luca de Alfaro, Vojtech Forejt, and Marta Kwiatkowska. Model checking probabilistic systems. In *Handbook of Model Checking*, pages 963–999. Springer, 2018.

4 Christel Baier, Holger Hermanns, and Joost-Pieter Katoen. The 10, 000 facets of MDP model checking. In *Computing and Software Science*, volume 10000 of *Lecture Notes in Computer Science*, pages 420–451. Springer, 2019.

5 Christel Baier, Joachim Klein, Sascha Klüppelholz, and Sascha Wunderlich. Maximizing the conditional expected reward for reaching the goal. In *TACAS (2)*, volume 10206 of *Lecture Notes in Computer Science*, pages 269–285, 2017.

6 Kevin Batz, Sebastian Junges, Benjamin Lucien Kaminski, Joost-Pieter Katoen, Christoph Matheja, and Philipp Schröer. Pric3: Property directed reachability for mdps. In *CAV (2)*, volume 12225 of *Lecture Notes in Computer Science*, pages 512–538. Springer, 2020.

7 Frantisek Blahoudek, Tomás Brázdil, Petr Novotný, Melkior Ornik, Pranay Thangeda, and Ufuk Topcu. Qualitative controller synthesis for consumption Markov decision processes. In *CAV (2)*, volume 12225 of *Lecture Notes in Computer Science*, pages 421–447. Springer, 2020.

8 Henrik C. Bohnenkamp, Pedro R. D'Argenio, Holger Hermanns, and Joost-Pieter Katoen. MODEST: A compositional modeling formalism for hard and softly timed systems. *IEEE Trans. Software Eng.*, 32(10):812–830, 2006.

9 Tomás Brázdil, Krishnendu Chatterjee, Martin Chmelik, Vojtech Forejt, Jan Kretínský, Marta Z. Kwiatkowska, David Parker, and Mateusz Ujma. Verification of Markov decision processes using learning algorithms. In *ATVA*, volume 8837 of *Lecture Notes in Computer Science*, pages 98–114. Springer, 2014.

10 Carlos E. Budde, Christian Dehnert, Ernst Moritz Hahn, Arnd Hartmanns, Sebastian Junges, and Andrea Turrini. JANI: quantitative model and tool interaction. In *TACAS (2)*, volume 10206 of *Lecture Notes in Computer Science*, pages 151–168, 2017.

11 Krishnendu Chatterjee, Rupak Majumdar, and Thomas A. Henzinger. Markov decision processes with multiple objectives. In *STACS*, volume 3884 of *Lecture Notes in Computer Science*, pages 325–336. Springer, 2006.

12 Krishnendu Chatterjee, Mickael Randour, and Jean-François Raskin. Strategy synthesis for multi-dimensional quantitative objectives. *Acta Informatica*, 51(3-4):129–163, 2014.

13 YooJung Choi, Antonio Vergari, and Guy Van den Broeck. Probabilistic circuits: A unifying framework for tractable probabilistic models. oct 2020.

14 Philipp Chrszon, Clemens Dubslaff, Sascha Klüppelholz, and Christel Baier. Profeat: feature-oriented engineering for family-based probabilistic model checking. *Formal Aspects Comput.*, 30(1):45–75, 2018.

15 A. Coles, A. Coles, A. García Olaya, S. Jiménez, C. Linares López, S. Sanner, and S. Yoon. A survey of the seventh international planning competition. *Artificial Intelligence Magazine (AI Magazine)*, 33(1):83–88, 2012.

16 Hao Cui, Thomas Keller, and Roni Khardon. Stochastic planning with lifted symbolic trajectory optimization. In *ICAPS*, pages 119–127. AAAI Press, 2019.

17 Daan Fierens, Guy Van den Broeck, Joris Renkens, Dimitar Shterionov, Bernd Gutmann, Ingo Thon, Gerda Janssens, and Luc De Raedt. Inference and learning in probabilistic logic programs using weighted Boolean formulas. *Theory and Practice of Logic Programming*, 15:358–401, 5 2015.

18 Andrew D. Gordon, Thomas A. Henzinger, Aditya V. Nori, and Sriram K. Rajamani. Probabilistic programming. In *FOSE*, pages 167–181. ACM, 2014.

19 Henri Hansen, Marta Z. Kwiatkowska, and Hongyang Qu. Partial order reduction for model checking Markov decision processes under unconditional fairness. In *QEST*, pages 203–212. IEEE Computer Society, 2011.

20 Arnd Hartmanns, Sebastian Junges, Joost-Pieter Katoen, and Tim Quatmann. Multi-cost bounded tradeoff analysis in MDP. *J. Autom. Reason.*, 64(7):1483–1522, 2020.

21 Christian Hensel, Sebastian Junges, Joost-Pieter Katoen, Tim Quatmann, and Matthias Volk. The probabilistic model checker storm. *CoRR*, abs/2002.07080, 2020.

22 Holger Hermanns, Björn Wachter, and Lijun Zhang. Probabilistic CEGAR. In *CAV*, volume 5123 of *Lecture Notes in Computer Science*, pages 162–175. Springer, 2008.

23 Steven Holtzen, Sebastian Junges, Marcell Vazquez-Chanlatte, Todd D. Millstein, Sanjit A. Seshia, and Guy Van den Broeck. Model checking finite-horizon Markov chains with probabilistic inference. In *CAV (2)*, volume 12760 of *Lecture Notes in Computer Science*, pages 577–601. Springer, 2021.

24 Steven Holtzen, Guy Van den Broeck, and Todd Millstein. Scaling exact inference for discrete probabilistic programs. *Proc. ACM Program. Lang. (OOPSLA)*, nov 2020.

25 Yipeng Huang, Steven Holtzen, Todd Millstein, Guy Van den Broeck, and Margaret R. Martonosi. Logical abstractions for noisy variational quantum algorithm simulation. In *Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2021.

26 Tommi S. Jaakkola, David A. Sontag, Amir Globerson, and Marina Meila. Learning Bayesian network structure using LP relaxations. In *AISTATS*, volume 9 of *JMLR Proceedings*, pages 358–365. JMLR.org, 2010.

27 Joost-Pieter Katoen. The probabilistic model checking landscape. In *LICS*, pages 31–45. ACM, 2016.

28 Mark Kattenbelt, Marta Z. Kwiatkowska, Gethin Norman, and David Parker. A game-based abstraction-refinement framework for Markov decision processes. *Formal Methods Syst. Des.*, 36(3):246–280, 2010.

29 Thomas Keller and Patrick Eyerich. PROST: Probabilistic planning based on UCT. In *International Conference on Automated Planning and Scheduling*, pages 119–127, 2012.

30 Michaela Klauck, Marcel Steinmetz, Jörg Hoffmann, and Holger Hermanns. Bridging the gap between probabilistic model checking and probabilistic planning: Survey, compilations, and empirical comparison. *J. Artif. Intell. Res.*, 68:247–310, 2020.

31 Marta Z. Kwiatkowska, Gethin Norman, and David Parker. PRISM 4.0: Verification of probabilistic real-time systems. In *CAV*, volume 6806 of *Lecture Notes in Computer Science*, pages 585–591. Springer, 2011.

32 Marta Z. Kwiatkowska, Gethin Norman, David Parker, and Hongyang Qu. Assume-guarantee verification for probabilistic systems. In *TACAS*, volume 6015 of *Lecture Notes in Computer Science*, pages 23–37. Springer, 2010.

33 Bruno Lacerda, Fatma Faruq, David Parker, and Nick Hawes. Probabilistic planning with formal performance guarantees for mobile service robots. *Int. J. Robotics Res.*, 38(9), 2019.

34 Maxim Likhachev, Dave Ferguson, Geoff Gordon, Anthony Stentz, and Sebastian Thrun. Anytime dynamic A*: An anytime, replanning algorithm. In *Proceedings of the Fifteenth International Conference on International Conference on Automated Planning and Scheduling*, ICAPS'05, page 262–271. AAAI Press, 2005.

35 Andrew McCallum, Karl Schultz, and Sameer Singh. Factorie: Probabilistic programming via imperatively defined factor graphs. *Advances in Neural Information Processing Systems*, 22:1249–1257, 2009.

36 Desmond C. Ong, Harold Soh, Jamil Zaki, and Noah D. Goodman. Applying probabilistic programming to affective computing. *IEEE Trans. Affect. Comput.*, 12(2):306–317, 2021.

37 David Pardoe and Peter Stone. Predictive planning for supply chain management. In *ICAPS*, pages 21–30. AAAI, 2006.

38 Bahare Salmani and Joost-Pieter Katoen. Bayesian inference by symbolic model checking. In *QEST*, volume 12289 of *Lecture Notes in Computer Science*, pages 115–133. Springer, 2020.

39 Bahare Salmani and Joost-Pieter Katoen. Fine-tuning the odds in Bayesian networks. In *ECSQARU*, volume 12897 of *Lecture Notes in Computer Science*, pages 268–283. Springer, 2021.

40 S. Sanner. Relational Dynamic Influence Diagram Language (RDDL): Language description. Unpublished Manuscript, Australian National University, 2010.

41 S. Sanner and C. Boutilier. Practical solution techniques for first-order MDPs. *Artificial Intelligence Journal (AIJ)*, pages 748–788, April 2009. Recipient of the 2014 Artificial Intelligence Journal (AIJ) Prominent Paper Award.

42 M. Vallati, L. Chrpa, M. Grzes, T. L. McCluskey, M. Roberts, and S. Sanner. The 2014 international planning competition: Progress and trends. *Artificial Intelligence Magazine (AI Magazine)*, 36(3):90–98, 2015.

43 Tom van Dijk and Jaco van de Pol. Multi-core symbolic bisimulation minimisation. *Int. J. Softw. Tools Technol. Transf.*, 20(2):157–177, 2018.

44 Antonio Vergari, YooJung Choi, Anji Liu, Stefano Teso, and Guy Van den Broeck. A compositional atlas of tractable circuit operations for probabilistic inference. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, dec 2021.

45 Shan Xue. *Scheduling and Online Planning in Stochastic Diffusion Networks.* PhD thesis, Oregon State University, 2020.

46 Honghua Zhang, Brendan Juba, and Guy Van den Broeck. Probabilistic generating circuits. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, jul 2021.

# 4 Overview of Talks

Below, we provide abstracts of the technical talks presented at the seminar, ordered lexicographically by the presenter.

## 4.1 Markov Decision Processes as Distribution Transformers: Decidability and Strategy Synthesis for Safety Objectives

*S. Akshay (Indian Institute of Technology Bombay – Mumbai, IN)*

**Joint work of** S. Akshay, Krishnendu Chatterjee, Tobias Meggendorfer, Djordje Zikelic

Markov decision processes can be viewed as transformers of probability distributions. This view is useful to reason about trajectories of distributions, but even basic reachability and safety problems turn out to be computationally intractable (Skolem-hard!). The issue is further complicated by the question of how much memory is allowed: even for simple examples, strategies for safety objectives over distributions can require infinite memory and randomization.

In light of this, we ask what one can do to tackle these problems in theory and in practice. After taking a look at some theoretical insights, we adopt an over-approximation route to approach these questions. Inspired by the success of invariant synthesis in program verification, we develop a framework for template-based synthesis of certificates as affine distributional and inductive invariants for safety objectives in MDPs. We show the effectiveness of our approach as well as explore limitations and future perspectives.

## 4.2 A Framework for Transforming Specifications in Reinforcement Learning

*Suguman Bansal (Georgia Institute of Technology – Atlanta, US)*

**Joint work of** Suguman Bansal, Rajeev Alur, Kishor Jothimurugan, Osbert Bastani
**Main reference** Rajeev Alur, Suguman Bansal, Osbert Bastani, Kishor Jothimurugan: "A Framework for Transforming Specifications in Reinforcement Learning", in Proc. of the Principles of Systems Design – Essays Dedicated to Thomas A. Henzinger on the Occasion of His 60th Birthday, Lecture Notes in Computer Science, Vol. 13660, pp. 604–624, Springer, 2022.
**URL** https://doi.org//10.1007/978-3-031-22337-2_29

Reinforcement Learning (RL) algorithms are designed to learn an optimal policy when the transition probabilities of the MDP are unknown but require the user to associate local rewards with transitions. The appeal of high-level temporal logic specifications has motivated research to develop RL algorithms for the synthesis of policies from specifications. To understand the techniques, and nuanced variations in their theoretical guarantees, in the growing body of resulting literature, we develop a formal framework for defining transformations among RL tasks with different forms of objectives. We define the notion of sampling-based reduction to transform a given MDP into another one that can be simulated even when the transition probabilities of the original MDP are unknown. We formalize the notions of preservation of optimal policies, convergence, and robustness of such reductions. We then use our framework to restate known results, establish new results to fill in some gaps, and identify open problems. In particular, we show that certain kinds of reductions from LTL specifications to reward-based ones do not exist, and prove the non-existence of RL algorithms with PAC-MDP guarantees for safety specifications.

## 4.3 Approximate Weighted Model Counting using Universal Hashing

*Supratik Chakraborty (Indian Institute of Technology Bombay – Mumbai, IN)*

**Joint work of** Supratik Chakraborty, Daniel Fremont, Kuldeep Meel, Sanjit Seshia, Moshe Vardi
**Main reference** Supratik Chakraborty, Daniel J. Fremont, Kuldeep S. Meel, Sanjit A. Seshia, Moshe Y. Vardi: "Distribution-Aware Sampling and Weighted Model Counting for SAT", in Proc. of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada, pp. 1722–1730, AAAI Press, 2014.
**URL** https://doi.org//10.1609/aaai.v28i1.8990

Given a system of propositional constraints, (unweighted) model counting concerns counting the number of satisfying assignments of the constraints. If every assignment is associated with a non-negative weight, the problem of finding the total weight of all satisfying assignments is called weighted model counting. This is a fundamental problem in computer science, with diverse applications in probabilistic inference, quantitative information flow, and partition function estimation among others. Unfortunately, model counting (even the unweighted variant) is computationally intractable – Valiant showed that this is #P-complete. Hence, it is unlikely that efficient algorithms exist for solving this problem. This has spurred a lot of interest in approximate weighted model counting. While there has been a lot of theoretical work in this domain that has yielded algorithms with strong guarantees of approximation, and also a lot of work that uses heuristics to scale to large problem instances without providing strong approximation guarantees, marrying scalability in practice with strong approximation guarantees is significantly hard. In this talk, we present a technique based on universal hash functions for solving the weighted model counting with PAC-style guarantees, yielding an approximate counter that scales well to large problem instances. The core idea of our technique is to partition the set of all assignments into cells of "small enough" and "almost equal" weights using universal hash functions, and then to count the total weight of solutions in a randomly chosen cell. The resulting weight is then multiplied by the total number of cells in the partition to obtain an estimate of the overall weighted model count.

We define a parameter called the "tilt" of weights of assignments, and show how the above idea leads to an algorithm that makes polynomially (in tilt, number of variables, and PAC approximation parameters) many calls to an NP oracle to yield an estimate of the weighted model count with PAC guarantees. Experiments show that this algorithm scales very well in practice when the tilt is bounded by a small constant. We then discuss an extension of our algorithm to deal with problem instances where the tilt may be large, and where assignment weights are the product of literal weights. The extended algorithm requires solving linear (in the number of variables) pseudo-boolean constraints. In practice, pseudo-boolean satisfiability solvers (including those that reduce to propositional satisfiability) are not as efficient in practice as propositional satisfiability solvers on large problem instances. Therefore, weighted model counting using universal hashing doesn't scale as well in practice when the tilt of weights is large, compared to the case of small tilt. Future advances in pseudo-boolean satisfiability solving are likely to directly impact the ability of weighted model counters to solve problem instances with large tilt.

## 4.4 Tractable Inference with Probabilistic Circuits

*YooJung Choi (Arizona State University – Tempe, US)*

Probabilistic circuits (PCs) are a family of models that guarantee exact inference of various probabilistic queries in polynomial (often linear) time. In this talk, we introduce the syntax and semantics of probabilistic circuits and study the structural properties that enable linear-time inference of marginal and MAP queries. We then discuss how we can perform inference on other probabilistic models such as Bayesian networks and probabilistic programs by compiling to circuits, in particular by reducing probabilistic inference to the task of weighted model counting/integration which can be performed tractably on certain circuits. Lastly, we showcase some recent works in complex reasoning using PCs. For instance, by representing queries as pipelines of atomic circuit operations, we show how we can systematically derive tractability conditions and inference algorithms for various information-theoretic entropies and divergences. This talk is based on the joint tutorial with Antonio Vergari, Robert Peharz, and Guy Van den Broeck.

## 4.5 Bit Blasting Probabilistic Programs

*Poorva Garg (UCLA, US), Steven Holtzen (Northeastern University – Boston, US), and Guy Van den Broeck (UCLA, US)*

**Joint work of** Poorva Garg, Steven Holtzen, Todd Millstein, Guy Van den Broeck

Probabilistic programming languages (PPLs) have emerged as a prominent area of research due to their ability to democratize probabilistic modeling. One of the key tasks in building a PPL is to design a generalizable probabilistic inference algorithm. Weighted model counting (WMC) is a popular exact inference algorithm for discrete probabilistic programs with much success. Can we extend the advantages of WMC to a wider class of probabilistic programs with both discrete and continuous distributions? Discretization of continuous distribution seems to be an obvious choice. However, it either leads to exhaustive enumeration or compromises on accuracy. LexBit (Language for Exact Bit blasting) is a non-trivial core language, with discrete and continuous constructs, that does not suffer from the limitations of discretization. It bit blasts exactly and scalably. We bit blast the continuous distributions outside this language using linear piece-wise distributions. Once all the continuous distributions in the probabilistic program are bit blasted, we harness the power of existing discrete PPLs to perform exact inference on the new discrete probabilistic program. Case studies and experiments on existing benchmarks show that this approach of bit blasting is competitive with existing probabilistic inference algorithms.

## 4.6 Compositional Probabilistic Model Checking with String Diagrams of MDPs

*Ichiro Hasuo (National Institute of Informatics – Tokyo, JP)*

**Joint work of** Kazuki Watanabe, Clovis Eberhart, Kazuyuki Asada, Ichiro Hasuo
**Main reference** Kazuki Watanabe, Clovis Eberhart, Kazuyuki Asada, Ichiro Hasuo: "Compositional Probabilistic Model Checking with String Diagrams of MDPs", in Proc. of the Computer Aided Verification – 35th International Conference, CAV 2023, Paris, France, July 17-22, 2023, Proceedings, Part III, Lecture Notes in Computer Science, Vol. 13966, pp. 40–61, Springer, 2023.
**URL** https://doi.org//10.1007/978-3-031-37709-9_3

We present a compositional model checking algorithm for Markov decision processes, in which they are composed in the categorical graphical language of string diagrams. The algorithm computes optimal expected rewards. Our theoretical development of the algorithm is supported by category theory, while what we call decomposition equalities for expected rewards act as a key enabler. Experimental evaluation demonstrates its performance advantages.

## 4.7 Introduction to Probabilistic Programming Inference

*Steven Holtzen (Northeastern University – Boston, US)*

**URL** https://www.khoury.northeastern.edu/home/sholtzen/CS7480Fall21/

How do we effectively run probabilistic programs in order to reason automatically about their behavior? In particular, how do we efficiently execute them in order to compute the probability that the program will have a particular behavior as efficiently as possible? In this talk, we go over the foundations of probabilistic program semantics and inference. We built a simple probabilistic programming language from scratch and described how to run it in order to evaluate queries. This talk was based on the introduction to a course taught on probabilistic programming at Northeastern University; the link is in the description.

## 4.8 Intelligent and Dependable Decision-Making Under Uncertainty

*Nils Jansen (Radboud University Nijmegen, NL)*

**Main reference** Nils Jansen: "Intelligent and Dependable Decision-Making Under Uncertainty", in Proc. of the Formal Methods – 25th International Symposium, FM 2023, Lübeck, Germany, March 6-10, 2023, Proceedings, Lecture Notes in Computer Science, Vol. 14000, pp. 26–36, Springer, 2023.
**URL** https://doi.org//10.1007/978-3-031-27481-7_3

This talk highlights our vision of foundational and application-driven research toward safety and dependability in artificial intelligence (AI). We take a broad stance on AI that combines formal methods, machine learning, and control theory. As part of this research line, we study problems inspired by autonomous systems, planning in robotics, and industrial applications.

We consider reinforcement learning (RL) as a specific machine learning technique for decision-making under uncertainty. RL generally learns to behave optimally via trial and error. Consequently, and despite its massive success in the past years, RL lacks mechanisms to ensure safe and correct behavior. Formal methods, in particular formal verification, is a research area that provides formal guarantees of a system's correctness and safety based on rigorous methods and precise specifications. Yet, fundamental challenges have obstructed the effective application of verification to reinforcement learning. Our main objective is to devise novel, data-driven verification methods that tightly integrate with RL. In particular, we develop techniques that address real-world challenges to the safety of AI systems in general: Scalability, expressiveness, and robustness against the uncertainty that occurs when operating in the real world. The overall goal is to advance the real-world deployment of reinforcement learning.

The talk is mainly based on the following references: [1, 2, 3, 4, 5].

### References

1 Nils Jansen. Intelligent and dependable decision-making under uncertainty. In *FM*, volume 14000 of *Lecture Notes in Computer Science*, pages 26–36. Springer, 2023.

2 Thom S. Badings, Thiago D. Simão, Marnix Suilen, and Nils Jansen. Decision-making under uncertainty: beyond probabilities. *Int. J. Softw. Tools Technol. Transf.*, 25(3):375–391, 2023.

3 Steven Carr, Nils Jansen, Sebastian Junges, and Ufuk Topcu. Safe reinforcement learning via shielding under partial observability. In *AAAI*, pages 14748–14756. AAAI Press, 2023.

4 Thom S. Badings, Licio Romao, Alessandro Abate, David Parker, Hasan A. Poonawala, Mariëlle Stoelinga, and Nils Jansen. Robust control for dynamical systems with non-gaussian noise via formal abstractions. *J. Artif. Intell. Res.*, 76:341–391, 2023.

5 Steven Carr, Nils Jansen, and Ufuk Topcu. Task-aware verifiable rnn-based policies for partially observable markov decision processes. *J. Artif. Intell. Res.*, 72:819–847, 2021.

## 4.9 Deductive Verification of Probabilistic Programs: Loops and Recursion

*Joost-Pieter Katoen (RWTH Aachen, DE)*

Probabilistic programs describe recipes on how to infer conclusions about big data from a mixture of uncertain data and real-world observations. Bayesian networks, a key model in decision-making, are simple instances of such programs. Probabilistic programs are used to steer autonomous robots and self-driving cars, are key to describe security mechanisms, and naturally encode randomised algorithms. Due to their learning ability, they are rapidly encroaching on AI and probabilistic machine learning.

This talk focuses on syntax-based verification of discrete probabilistic programs. We will show how weakest pre-condition style reasoning can be used to determine quantitative program properties such as the probability of divergence, bounds on the expected outcomes of program expressions, or the program's expected run-time. Complementary to Holtzen's talk on straight-line code, we focus primarily on how to treat possibly unbounded loops and recursion.

We will present automated methods such as k-induction and how to find loop invariants in a CEGIS-like fashion. An outlook will be given of some alternative automated techniques for program equivalence and how to exploit model checking for obtaining lower bounds on loops in probabilistic programs.

## 4.10 Scalable Learning of Probabilistic Circuits

*Anji Liu (UCLA, US)*

**Joint work of** Anji Liu, Guy Van den Broeck, Honghua Zhang, Antonio Vergari, YooJung Choi, Stephan Mandt
**Main reference** Anji Liu, Honghua Zhang, Guy Van den Broeck: "Scaling Up Probabilistic Circuits by Latent Variable Distillation", in Proc. of the The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net, 2023.
**URL** https://openreview.net/pdf?id=067CGykiZTS

Probabilistic Circuits (PCs) are a unified framework for tractable probabilistic models that support efficient computation of various probabilistic queries (e.g., marginal probabilities). One key challenge is to scale PCs to model large and high-dimensional real-world datasets: we observe that as the number of parameters in PCs increases, their performance immediately plateaus. This phenomenon suggests that the existing optimizers fail to exploit the full expressive power of large PCs. We propose to overcome such bottleneck by latent variable distillation: we leverage the less tractable but more expressive deep generative models to provide extra supervision over the latent variables of PCs. Specifically, we extract information from Transformer-based generative models to assign values to latent variables of PCs, providing guidance to PC optimizers. Experiments on both image and language modeling benchmarks (e.g., ImageNet and WikiText-2) show that latent variable distillation substantially boosts the performance of large PCs compared to their counterparts without latent variable distillation. In particular, on the image modeling benchmarks, PCs achieve competitive performance against some of the widely-used deep generative models, including variational autoencoders and flow-based models, opening up new avenues for tractable generative modeling.

## 4.11 How to Make Logics Neurosymbolic

*Luc De Raedt (KU Leuven, BE)*

**Joint work of** Luc De Raedt, Giuseppe Marra, Robin Manhaeve, Thomas Winters, Vincent Derkinderen, Sebastijan Dumancic
**Main reference** Giuseppe Marra, Sebastijan Dumancic, Robin Manhaeve, Luc De Raedt: "From Statistical Relational to Neural Symbolic Artificial Intelligence: a Survey", CoRR, Vol. abs/2108.11451, 2021.
**URL** https://arxiv.org/abs/2108.11451

Neurosymbolic AI (NeSy) is regarded as the third wave in AI. It aims at combining knowledge representation and reasoning with neural networks. Numerous approaches to NeSy are being developed and there exists an "alphabet soup" of different systems, whose relationships are often unclear. I will discuss the state-of-the-art in NeSy and argue that there are many similarities with statistical relational AI (StarAI).
Taking inspiration from StarAI, and exploiting these similarities, I will argue that Neurosymbolic AI = Logic + Probability + Neural Networks. I will also provide a recipe for developing NeSy approaches: start from a logic, add a probabilistic interpretation, and then turn neural networks into "neural predicates". Probability is interpreted broadly here and is necessary to provide a quantitative and differentiable component to the logic. At the semantic and the computation level, one can then combine logical circuits (ako proof structures) labeled with probability, and neural networks in computation graphs.

I will illustrate the recipe with NeSy systems such as DeepProbLog, a deep probabilistic extension of Prolog, and DeepStochLog, a neural network extension of stochastic definite clause grammars (or stochastic logic programs).

The key references of the talk are as follows: [1, 2, 3].

### References

1 Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Deepproblog: Neural probabilistic logic programming. In *NeurIPS*, pages 3753–3763, 2018.

2 Thomas Winters, Giuseppe Marra, Robin Manhaeve, and Luc De Raedt. Deepstochlog: Neural stochastic logic programming. In *AAAI*, pages 10090–10100. AAAI Press, 2022.

3 Giuseppe Marra, Sebastijan Dumančić, Robin Manhaeve, and Luc De Raedt. From statistical relational to neural symbolic artificial intelligence: a survey. *arXiv preprint arXiv:2108.11451*, 2021.

## 4.12 Tutorial: Probabilistic Model Checking

*David Parker (University of Oxford, GB)*

Probabilistic model checking is an automated technique for the formal verification of stochastic systems. This tutorial will provide an introduction to some of the key ingredients of this technique, giving a particular focus on the similarities and differences with some of the other fields represented at the seminar, such as probabilistic programming, probabilistic circuits and, probabilistic planning.
I will cover: (i) the types of probabilistic models typically used; (ii) the use of temporal logic to formalise quantitative behavioural specifications, in particular for models such as Markov chains and Markov decision processes; (iii) the solution techniques usually used by probabilistic model checking tools, and the approaches taken to improve scalability and efficiency; and (iv) modelling languages for probabilistic verification. In the final part of the talk, I will discuss how this framework has been extended to support multi-agent systems modelled as stochastic games.

## 4.13 Mixing formal methods and learning to tackle (too) large MDPs

*Jean-Francois Raskin (UL – Brussels, BE)*

In a recent series of works, we investigate optimizing strategies in Markov decision processes (MDPs) using Monte Carlo Tree Search (MCTS). We introduce symbolic advice to enhance MCTS, biasing its selection and simulation strategies while maintaining its theoretical guarantees. Efficient implementation of symbolic advice is achieved using QBF and SAT solvers. Additionally, we integrate formal methods and deep learning to produce superior receding horizon policies in large MDPs. Model-checking techniques guide MCTS for high-quality decision sampling, which subsequently trains a neural network to imitate the sampled

policy. This network can guide low-latency MCTS online searches or serve as an independent policy when quick responses are vital. Statistical model checking identifies areas needing extra samples and highlights discrepancies between the neural network and the offline policy. Our methodologies are validated using the Pac-Man and Frozen Lake environments, benchmarks in evaluating reinforcement-learning algorithms. The results outperform standard MCTS and human players.
The main related papers to the talk are as follows: [1, 2, 3].

### References

1 Debraj Chakraborty, Damien Busatto-Gaston, Jean-François Raskin, and Guillermo A. Pérez. Formally-sharp dagger for MCTS: lower-latency monte carlo tree search using data aggregation with formal methods. In Noa Agmon, Bo An, Alessandro Ricci, and William Yeoh, editors, *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023, London, United Kingdom, 29 May 2023 – 2 June 2023*, pages 1354–1362. ACM, 2023.

2 Damien Busatto-Gaston, Debraj Chakraborty, and Jean-François Raskin. Monte carlo tree search guided by symbolic advice for mdps. In Igor Konnov and Laura Kovács, editors, *31st International Conference on Concurrency Theory, CONCUR 2020, September 1-4, 2020, Vienna, Austria (Virtual Conference)*, volume 171 of *LIPIcs*, pages 40:1–40:24. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020.

3 Gilles Geeraerts, Shibashis Guha, and Jean-François Raskin. Safe and optimal scheduling for hard and soft tasks. In Sumit Ganguly and Paritosh K. Pandya, editors, *38th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, FSTTCS 2018, December 11-13, 2018, Ahmedabad, India*, volume 122 of *LIPIcs*, pages 36:1–36:22. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018.

## 4.14 Automatically Finding the Right Probabilities in Bayesian Networks

*Bahare Salmani (RWTH Aachen, DE) and Joost-Pieter Katoen (RWTH Aachen, DE)*

**Main reference** Bahare Salmani, Joost-Pieter Katoen: “Automatically Finding the Right Probabilities in Bayesian Networks”, J. Artif. Intell. Res., Vol. 77, pp. 1637–1696, 2023.
**URL** https://doi.org//10.1613/jair.1.14044

Parametric Bayesian networks (pBNs) are extensions of Bayesian networks that allow conditional probability tables (CPTs) to include unknown parameters rather than concrete probabilities. We present in this talk alternative techniques to find the correct values for the parameters with respect to a given constraint. The key is to translate (a) pBNs to parametric Markov chains (pMCs) and (b) pBN constraints to reachability constraints. This allows exploiting the state-of-the-art parameter synthesis techniques for pMCs to target pBN problems including sensitivity analysis, (minimal-change) parameter tuning, and parameter space partitioning. We address pBNs with an arbitrary number of parameterized CPTs and with arbitrary dependencies between the parameters. This lifts the limitations of the existing pBN techniques. Our experimental results indicate that our techniques scale up to 800 unknown parameters for large Bayesian networks with $\sim 100$ random variables.

## 4.15 Tutorial on Planning with Probabilistic Programming Languages

*Scott Sanner (University of Toronto, CA)*

Planning aims to find sequences of actions that achieve a goal or optimize a cost-based objective given an initial starting state. Modern planning methods seek to leverage the structure in symbolic domain specification languages to improve the efficiency of the search process. The Relational Dynamic Influence Diagram Language (RDDL) is a probabilistic programming language that has been developed to compactly model real-world stochastic planning problems, i.e., Markov Decision Processes (MDPs), and specifically factored MDPs with highly structured transition and reward functions. In this tutorial, we will cover the basics of RDDL and present recent language extensions and capabilities through the incremental development and extension of running examples based on real-world domains. We will also introduce a range of planning methodologies that leverage RDDL structure covering Monte Carlo Tree Search (MCTS), mathematical programming, gradient-based optimization, and symbolic methods.

## 4.16 Deterministic stream-sampling for probabilistic programming: semantics and verification

*Alexandra Silva (Cornell University – Ithaca, US)*

**Joint work of** Fredrik Dahlqvist, Alexandra Silva, William Smith
**Main reference** Fredrik Dahlqvist, Alexandra Silva, William Smith: "Deterministic stream-sampling for probabilistic programming: semantics and verification", in Proc. of the LICS, pp. 1–13, 2023.
**URL** https://doi.org//10.1109/LICS56636.2023.10175773

Probabilistic programming languages rely fundamentally on some notion of sampling, and this is doubly true for probabilistic programming languages which perform Bayesian inference using Monte Carlo techniques. Verifying samplers – proving that they generate samples from the correct distribution – is crucial to the use of probabilistic programming languages for statistical modelling and inference. However, the typical denotational semantics of probabilistic programs is incompatible with deterministic notions of sampling. This is problematic, considering that most statistical inference is performed using pseudorandom number generators. We present a higher-order probabilistic programming language centred on the notion of samplers and sampler operations. We give this language an operational and denotational semantics in terms of continuous maps between topological spaces. Our language also supports discontinuous operations, such as comparisons between reals, by using the type system to track discontinuities. This feature might be of independent interest, for example in the context of differentiable programming. Using this language, we develop tools for the formal verification of sampler correctness. We present an equational calculus to reason about equivalence of samplers, and a sound calculus to prove semantic correctness of samplers, i.e. that a sampler correctly targets a given measure by construction.

## 4.17 E-MCTS: Deep Exploration in Model-Based Reinforcement Learning by Planning with Epistemic Uncertainty

*Matthijs Spaan (TU Delft, NL)*

**Joint work of** Matthijs Spaan, Wendelin Böhmer, Yaniv Oren
**Main reference** Yaniv Oren, Matthijs T. J. Spaan, Wendelin Böhmer: "Planning with Uncertainty: Deep Exploration in Model-Based Reinforcement Learning", CoRR, Vol. abs/2210.13455, 2022.
**URL** https://doi.org//10.48550/arXiv.2210.13455

One of the most well-studied and highly performing planning approaches used in Model-Based Reinforcement Learning (MBRL) is Monte-Carlo Tree Search (MCTS). Key challenges of MCTS-based MBRL methods remain dedicated deep exploration and reliability in the face of the unknown, and both challenges can be alleviated through principled epistemic uncertainty estimation in the predictions of MCTS. We present two main contributions: First, we develop methodology to propagate epistemic uncertainty in MCTS, enabling agents to estimate the epistemic uncertainty in their predictions. Second, we utilize the propagated uncertainty for a novel deep exploration algorithm by explicitly planning to explore. We incorporate our approach into variations of MCTS-based MBRL approaches with learned and provided models, and empirically show deep exploration through successful epistemic uncertainty estimation achieved by our approach. We compare to a non-planning-based deep-exploration baseline, and demonstrate that planning with epistemic MCTS significantly outperforms non-planning based exploration in the investigated setting.

## Participants

- S. Akshay
Indian Institute of Technology Bombay – Mumbai, IN
- Suguman Bansal
Georgia Institute of Technology – Atlanta, US
- Kevin Batz
RWTH Aachen, DE
- Milan Ceska
Brno University of Technology, CZ
- Supratik Chakraborty
Indian Institute of Technology Bombay – Mumbai, IN
- YooJung Choi
Arizona State University – Tempe, US
- Cassio de Campos
TU Eindhoven, NL
- Luc De Raedt
KU Leuven, BE
- Rayna Dimitrova
CISPA – Saarbrücken, DE
- Poorva Garg
UCLA, US
- Vibhav Gogate
University of Texas at Dallas – Richardson, US
- Ichiro Hasuo
National Institute of Informatics – Tokyo, JP
- Holger Hermanns
Universität des Saarlandes – Saarbrücken, DE
- Steven Holtzen
Northeastern University – Boston, US
- Manfred Jaeger
Aalborg University, DK
- Nils Jansen
Radboud University Nijmegen, NL
- Sebastian Junges
Radboud University Nijmegen, NL
- Amir Kafshdar Goharshady
HKUST – Kowloon, HK
- Benjamin Kaminski
Universität des Saarlandes – Saarbrücken, DE
- Joost-Pieter Katoen
RWTH Aachen, DE
- Samuel Kolb
KU Leuven, BE
- John Li
Northeastern University – Boston, US
- Anji Liu
UCLA, US
- Christoph Matheja
Technical University of Denmark – Lyngby, DK
- Chih-Hao Luke Ong
Nanyang TU – Singapore, SG
- David Parker
University of Oxford, GB
- Jean-Francois Raskin
UL – Brussels, BE
- Jurriaan Rot
Radboud University Nijmegen, NL
- Bahare Salmani Barzorki
RWTH Aachen, DE
- Scott Sanner
University of Toronto, CA
- Alexandra Silva
Cornell University – Ithaca, US
- Matthijs Spaan
TU Delft, NL
- Guy Van den Broeck
UCLA, US
- Antonio Vergari
University of Edinburgh, GB

Report from Dagstuhl Seminar 23242

# Privacy Protection of Automated and Self-Driving Vehicles

**Frank Kargl***[1], **Ioannis Krontiris***[2], **Jason Millar***[3], **André Weimerskirch***[4], **and Kevin Gomez**†[5]

1 **Universität Ulm, DE.** frank.kargl@uni-ulm.de
2 **Huawei Technologies – München, DE.** ioannis.krontiris@huawei.com
3 **University of Ottawa, CA.** jmillar@uottawa.ca
4 **Lear Corporation, US.** aweimerskirch@lear.com
5 **TH Ingolstadt, DE.** Kevin.Gomez@carissma.eu

## Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 23242 "Privacy Protection of Automated and Self-Driving Vehicles". While privacy for connected vehicles has been considered for many years, automated and autonomous vehicles (AV) technology is still in its infancy and the privacy and data protection aspects for AVs are not well addressed. Their capabilities pose new challenges to privacy protection, given the large sensor arrays that collect data in public spaces and the integration of AI technology.

During the seminar, several keynote presentations highlighted the research challenges from different perspectives, i.e. legal, ethical, and technological. It was also discussed extensively why vehicles need to make dynamic assessments of trust as an enabling factor for the secure communication and data sharing with other vehicles, but without increasing any privacy risks.

Then, the main objective of the seminar was to produce a research road-map to address the major road-blockers in making progress on the way to deployment of privacy protection in automated and autonomous vehicles. First, the group identified six common scenarios of Cooperative, Connected and Automated Mobility (CCAM) during development and product life-cycle, and analyzed the privacy implications for each scenario. Second, it formulated the need to have a methodology to determine the cost-benefit trade-offs between privacy and other criteria like financial, usability, or safety. Third, it identified existing tools, frameworks, and PETs, and potential modifications that are needed to support the automotive industry and automotive scenarios. Finally, the group explored the interplay between privacy and trust, by elaborating on different trust properties based on performance, on ethical aspects, and on user acceptance.

**Seminar** June 11–16, 2023 – https://www.dagstuhl.de/23242

**2012 ACM Subject Classification** Security and privacy → Human and societal aspects of security and privacy; Security and privacy → Privacy protections; Security and privacy → Privacy-preserving protocols

**Keywords and phrases** Automotive Security and Privacy, Privacy and Data Protection, Cooperative Connected and Automated Mobility

**Digital Object Identifier** 10.4230/DagRep.13.6.22

* Editor / Organizer
† Editorial Assistant / Collector

Privacy Protection of Automated and Self-Driving Vehicles, *Dagstuhl Reports*, Vol. 13, Issue 6, pp. 22–54
Editors: Frank Kargl, Ioannis Krontiris, Jason Millar, André Weimerskirch, and Kevin Gomez
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

# 1 Executive Summary

*Frank Kargl (Universität Ulm – Ulm, DE)*
*Ioannis Krontiris (Huawei Technologies – München, DE)*
*Jason Millar (University of Ottawa, CA)*
*André Weimerskirch (Lear Corporation – Ann Arbor, US)*

Cooperative, connected and automated mobility (CCAM) has the potential to drastically reduce accidents, travel time, and the environmental impact of road travel. To achieve these goals, automated vehicles (AV) will require a range of sensors and communication devices that receive and read extensive data from the vehicle's environment, as well as machine learning algorithms that process this data. This immediately raises the concern of privacy for AVs. A first Dagstuhl Seminar was held virtually January 23–28, 2022 [1], and identified four main challenges: (1) How to encourage stakeholders to follow proper ethics and responsible behaviour, (2) how regulation needs to evolve for CCAM systems, (3) the commercial limitations to develop and implement proper privacy protection, and (4) availability of privacy-enhancing technologies for CCAM systems. The Dagstuhl Seminar at hand was then held in person June 11–16, 2023, with the goal to approach those main challenges.

This seminar was organized in a number of expert presentations, and then the group split into four working groups. The expert presentations covered many relevant aspects around regulation and governance, cloud-based support infrastructure, and technology. The four working groups roughly map to the main challenges:

1. Scenarios, Risks, Impacts, and Collected Data in CCAM: This group identified six common CCAM scenarios during development and product life-cycle, and analyzed the privacy implications for each scenario. Some of these scenarios are unique to CCAM privacy and set it apart from other areas. The results can now be used as foundation for further general in-depth privacy research.
2. Privacy Tensions for Connected Automated Vehicles: It is believed that privacy comes at a cost, whether it is a financial cost, reduced usability, or reduced safety. It is essential to understand how to find the acceptable trade-off between privacy and the considered criteria. However, today we have no proper methodology in place to determine proper trade-off points, and therefore this group worked on developing such a methodology. Additionally, this group will identify the technology readiness level of privacy enhancing technologies (PET) to support the trade-off points. The working group plans to describe details in an upcoming scientific publication.
3. Automotive Privacy Engineering: Privacy engineering provides the underlying tools, frameworks, and technologies to develop privacy protecting CCAM. This working group focused on identifying existing tools, frameworks, and PETs that could support our use-case, potential modifications that are needed to support CCAM, and gaps. The group emphasized the need to match the privacy engineering to users' privacy and usability expectations. The group identified and discussed six questions that addressed the major aspects, and derived various action items for the automotive privacy research community.
4. Interplay between Privacy and Trust: One of the most important milestones in order to achieve the shared vision on the deployment of Cooperative Intelligent Transport Systems (C-ITS) towards cooperative, connected and automated mobility (CCAM), is to allow

participating entities to assess dynamically the trustworthiness of the shared information, in order to be able to rely on it and coordinate their actions [2]. In addressing this complex issue, it's paramount to strike a balance between enhancing trust and ensuring the privacy and security of users' personal information and data. The group explored the interplay between privacy and trust, by elaborating on different trust properties based on performance, on ethical aspects, and on user acceptance.

We conclude that more solution-oriented research and development is required to establish privacy modeling tools and privacy engineering specifically for CCAM, and we hope that the results and papers coming from this seminar will support the journey to privacy protecting CCAM. Shortly after the seminar, the Mozilla Foundation's Privacy Not Included [3, 4] reviewed 25 major car brands for consumer privacy and gave all of them failing marks for consumer privacy, and we hope that this seminar's solutions also improve the privacy of next generation passenger vehicles.

## References

1 Frank Kargl, Ioannis Krontiris, Nataša Trkulja, André Weimerskirch, and Ian Williams, Privacy Protection of Automated and Self-Driving Vehicles (Dagstuhl Seminar 22042), Dagstuhl Reports, Vol. 12, Issue 1, pp. 83–100, https://doi.org/10.4230/DagRep.12.1.83.

2 EU Project "CONNECT: Continuous and Efficient Cooperative Trust Management for Resilient CCAM", [ONLINE] https://horizon-connect.eu/

3 Mozilla Foundation, "Privacy Nightmare on Wheels": Every Car Brand Reviewed By Mozilla – Including Ford, Volkswagen and Toyota – Flunks Privacy Test, [ONLINE] https://foundation.mozilla.org/en/blog/privacy-nightmare-on-wheels-every-car-brand-reviewed-by-mozilla-including-ford-volkswagen-and-toyota-flunks-privacy-test/

4 Mozilla Foundation, Privacy Not Included, [ONLINE] https://foundation.mozilla.org/en/privacynotincluded/categories/cars/

## 2 Table of Contents

**Executive Summary**

**Overview of Talks**

**Working Groups**

# 3 Overview of Talks

## 3.1 Exploring the Costs of AVs and Privacy

*Adam Henschke (University of Twente, NL)*

I introduced a range of ethical issues about AVs (autonomous vehicles) and privacy that arise in relation to insurance. Tesla vehicles in some US states now offer a "safety score" which impacts their Tesla provided insurance. This seems good as it incentivises safer driving and reduced insurance premiums. However, there are problems like phantom breaking, in which false information (AV breaking when it does not need to) has an effect of unfairly raising driver's insurance premiums. This application highlights that AVs present a unique set of privacy risks and challenges. For instance, there are economic incentives to collect behavioural data on drivers. Second, this approach to individualising/personalising insurance costs runs the risk of "responsibilisation", in which individuals are held responsible for systemic issues, like poorly maintained road: An individual's safety score may drop if they drive on poorly maintained roads, even if they are not the cause of those poorly maintained roads and can do nothing individually to repair them. By looking at safety scores and insurance, we have a useful way to think about a wide range of privacy issues when considering AVs.

## 3.2 A Quick Intro to AD Regulations

*Ben Brecht (Berlin, DE)*

With the adoption of (EU) 2022/1426 [1] and (EU) 2022/2236 [2] as an amendment to the EU Whole Vehicle Type Approval Framework, type approval of an SAE Level 3 or 4 autonomous vehicle is possible for the first time in Europe. Type Approval is not sufficient to operate an autonomous vehicle in Europe. This requires an adapted national framework, as the EU has no legislative competence for the registration of vehicles and thus for the approval of an operating area. In Germany, this has been achieved through adjustments to the Road Traffic Act (StVG)[3], the Compulsory Insurance Act, the Vehicle Registration Ordinance (FZV)[4] and the creation of the Autonomous Vehicles Approval and Operation Ordinance (AFGBV)[5]. Specifically, the AFGBV contains the rules for the operating area permit, which is a mandatory requirement for road registration of L3/L4 vehicles after the revision of the FZV. For the permission to transport passengers, a concession according to the Passenger Transport Act (PBefG)[6] is additionally required.

All these regulations impose requirements on the generation, storage, processing or sharing of data (e.g. with authorities). A deeper look at the requirements from (EU) 2022/1426 with extensions of the list of data to be stored in the Event Data Recorder according to Article 6 of Regulation (EU) 2019/2144[7] reveals a rather traditional approach, which ignores data from cameras, lidars and radars. These data are also not taken into account in the requirements for reporting to authorities, although they are indispensable for the analysis of safety-relevant incidents. Only in the requirements for a safety management system (SMS), there is a very broad scope for the storage and processing of data, but – after a first rough technical analysis – it seems restrictions on purpose or use of such data might be missing.

## References

1 Commission Implementing Regulation (EU) 2022/1426 of 5 August 2022 laying down rules for the application of Regulation (EU) 2019/2144 of the European Parliament and of the Council as regards uniform procedures and technical specifications for the type-approval of the automated driving system (ADS) of fully automated vehicles. *Official Journal, L 221, 26.8.2022, p. 1–64.*

2 Commission Delegated Regulation (EU) 2022/2236 of 20 June 2022 amending Annexes I, II, IV and V to Regulation (EU) 2018/858 of the European Parliament and of the Council as regards the technical requirements for vehicles produced in unlimited series, vehicles produced in small series, fully automated vehicles produced in small series and special purpose vehicles, and as regards software update. *Official Journal, L 296, 16.11.2022, p. 1–176.*

3 Road Traffic Act (Straßenverkehrsgesetz) [ONLINE] `https://www.gesetze-im-internet.de/stvg/BJNR004370909.html`

4 Verordnung über die Zulassung von Fahrzeugen zum Straßenverkehr (Fahrzeug-Zulassungsverordnung – FZV) [ONLINE] `https://www.gesetze-im-internet.de/fzv_2023/BJNR0C70B0023.html`

5 Verordnung zur Genehmigung und zum Betrieb von Kraftfahrzeugen mit autonomer Fahrfunktion in festgelegten Betriebsbereichen (Autonome-Fahrzeuge-Genehmigungs-und-Betriebs-Verordnung – AFGBV) [ONLINE] `https://www.gesetze-im-internet.de/afgbv/BJNR098610022.html`

6 Personenbeförderungsgesetz (PBefG) [ONLINE] `https://www.gesetze-im-internet.de/pbefg/BJNR002410961.html`

7 Regulation (EU) 2019/2144 of the European Parliament and of the Council of 27 November 2019 on type-approval requirements for motor vehicles and their trailers, and systems, components and separate technical units intended for such vehicles, as regards their general safety and the protection of vehicle occupants and vulnerable road users [ONLINE] `https://eur-lex.europa.eu/eli/reg/2019/2144/oj`

## 3.3 The Automotive Industry under Worldwide Data Protection Regulations: A Technical Perspective

*Alaa Al-Momani (Ulm University, DE)*

The recent adoption of data protection regulations is necessary to regulate how and for what purpose consumers' data is collected, processed, and shared. Generally, organisations that collect, process, or share personal information of data subjects are required to comply with one (or more) data protection regulations. In the case of the automotive industry and its various services, collecting as well as processing and sharing (sensitive) personal information is highly likely, including identifiers and geolocation of end-users. In this talk, we compare the data protection regulations in major automotive industry markets around the world, ie, the European Union (EU), the United States of America (US), and Japan. In particular, we look at the General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), and the Japanese Act on the Protection of Personal Information (APPI), respectively, and discuss the impact of these regulations on automotive services. We consider an autonomous taxicab service as an example of an automotive service and investigate how such a service can be designed in compliance with the previous regulations. We further highlight the challenge that a worldwide service provider faces when complying with all of the previous regulations at once as they may substantially differ in some aspects. Furthermore, we take a look at the road ahead and highlight the challenges when it comes to integrating machine learning models and artificial intelligence within automotive services.

## 3.4 Demystifying the Tension between Trust and Privacy in CCAM

*Thanassis Giannetsos (UBITECH Ltd. – Athens, GR)*

Modern vehicles are no longer mere mechanical devices; they comprise dozens of digital computing platforms coordinated by an in-vehicle network, and have the potential to significantly enhance the digital life of individuals on the road. While this transformation has driven major advancements in road safety and transportation efficiency, significant work remains to be done to capture the strict security, privacy, and trust requirements of all involved stakeholders.

For instance, driving on the road requires trust in others and the environment, but in reality, we never completely trust – not us, not other drivers or what is ahead of us. Therefore, how can we be sure about the data integrity and level of trust in connected cars that cooperatively need to execute a safety-critical function? At the same time, the integration of such integrity and assurance controls might impede with the privacy profile of the vehicles which, in turn, might affect the level of user acceptance of such systems – user perceived trust is greatly affected by the system's capability to preserve the privacy of the driver.

In this presentation, we had a deeper look into the details of trust management vs. privacy and why vehicles need to make dynamic assessments of trust as an enabling factor for the secure communication and data sharing with other CCAM entities, but without increasing any privacy risks. EU Project CONNECT has shown a complete framework how this is technically possible. At the end we achieve the end-goal of combing a vehicle's systems with information available externally (from multiple sources), in a way that expand the knowledge on the environment that is required for decision-making, in a trustworthy but also privacy-friendly way. This increases the safety of the overall CCAM ecosystem and unlocks future applications.

## 3.5 Privacy Challenges in Vehicle Security Operation Centers – From a CCAM perspective

*Kevin Gomez (Technical University Ingolstadt, DE)*

The SELFY project develops a toolbox for the CCAM environment. We (THI) develop a Vehicle Security Operation Center (VSOC) for the SELFY toolbox and CCAM environments. The SELFY VSOC can be considered a meta VSOC that collects data from various endpoints within the vehicle ecosystem and provides services to the ecosystem and vehicle manufacturers. Those services include the detection of anomalies, distribution of information (e.g., security scenarios and MITRE ATT&CK matrix), updates, and trust scores.

One of the main challenges within the VSOC is the trustworthiness of OEMs. Why should an OEM trust the SELFY VSOC? What is the benefit of sharing information with the SELFY VSOC? And how can we, as the SELFY VSOC, trust data from the endpoints and OEMs?

One solution to tackle the challenge could be differential privacy. Here, data can be shared while individuals without identifying the OEM as a participant in the dataset. This characteristic would address challenges by OEMs with external systems such as the SELFY VSOC. An OEM could share data with the SELFY VSOC without being identified as a participant and potentially leaking information on their security scenarios, vulnerabilities, and used technologies. However, privacy does not come without additional costs. In differential privacy, the degree of privacy depends on a factor, usually referred to as epsilon. "How much data does one need for effective different privacy in a VSOC?", "Is the distrust from OEMs solved by differential privacy?" and "What epsilon should be chosen for which data types?".

## 3.6 PQC Impacts on V2X

*Takahito Yoshizawa (KU Leuven, BE) and Brigitte Lonc (IRT SystemX, FR)*

Vehicular communication, or Vehicle-to-Everything (V2X) communication uses digital signature called Elliptic Curve Digital Signature Algorithm (ECDSA) to verify the message's integrity and sending vehicle's authenticity. Its signature and public key lengths are 64 and 32 bytes, respectively. At the same time, Due to the evolving capabilities of Quantum Computers (QC), public-key cryptography (e.g. RSA, ECC) are expected to be broken when the QC of sufficient computing power become available. According to several articles, QCs as large as 100,000 qubits may become available by 2030 [1, 2]. If this becomes a reality, some of the expressed concerns may become an imminent issue [3, 4].

To address this issue, US National Institute of Standards and Technology (NIST) started standardizing work of Post-Quantum Cryptography (PQC) in 2016. As of today, 3 PQC signature algorithms are standardized [5]. However, all of them have large signature and/or public key size which saturates the V2X message size [6, 7]. In light of this situation, NIST invited call for proposal (CFP) for additional PQC signature algorithm that has characteristics of short signature and fast verification [8]. According to their current schedule, we may have viable solution(s) identified in early 2025, which may be suitable for real-time systems such as V2X communication.

### References

1 Wikipedia.org, "Timeline of quantum computing and communication", [ONLINE] https://en.wikipedia.org/wiki/Timeline_of_quantum_computing_and_communication.
2 MIT Technology Review, "IBM wants to build a 100,000-qubit quantum computer", [ONLINE] https://www.technologyreview.com/2023/05/25/1073606/ibm-wants-to-build-a-100000-qubit-quantum-computer/
3 J. Proos, C. Zalka, "Shor's discrete logarithm quantum algorithm for elliptic curves," Quant. Inf. Comput., vol. 3, no. 4, pp. 317–344, Jul. 2003.
4 NIST IR 8105, "Report on Post-Quantum Cryptography", Apr. 2016

5 NIST, IR8413 “Status Report on the Third Round of the NIST Post-Quantum Cryptography Standardization Process”, https://nvlpubs.nist.gov/nistpubs/ir/2022/NIST.IR.8413-upd1.pdf.
6 B. Lonc, X. Aubry, H. Bakhti, M. Christofi, H.A. Mehrez, “Feasibility and Benchmarking of Post-Quantum Cryptography in the Cooperative ITS Ecosystem”, IEEE Vehicular Networking Conference (VNC 2023), April. 2023.
7 T. Yoshizawa, B. Preneel, “Post-Quantum Impacts on V2X Certificates – Already at The End of The Road”, IEEE Vehicular Technology Conference (VTC 2023), June 2023.
8 NIST, “Post-Quantum Cryptography: Digital Signature Schemes”, [ONLINE] https://csrc.nist.gov/projects/pqc-dig-sig

## 3.7 Technology, Data, and Enforcement in Service to Autonomy and Community

*Bryant Walker Smith (University of South Carolina – Columbia, US)*

Privacy is about power and, as a means rather than an end, should be discussed in relation to the twin societal goals of human autonomy and community.

As an initial matter: While the state of automated driving has for years been overhyped, currently it is underappreciated. Today there is much reason for technical optimism, even as economic success is challenging. But technical or even economic success is not the same as social or policy success. For example, automated driving could become widespread even if it is not privacy-protecting – an outcome which may or may not be socially desirable.

These policy discussions are happening now. For example, dozens of countries in the UN’s Global Forum for Road Traffic Safety are currently exploring a new international instrument on automated driving. Privacy has been a flash point in this effort and its predicates. Some of the especially interesting issues arise from cross-border dynamics of the vehicle or its automated driving system, relevant data, and notions of “control.” I describe some of these specific scenarios.

To my thesis: Privacy is an incomplete frame. Some people prioritize acceptance, belonging, validation, or legacy over privacy. There are also tradeoffs within as well as with privacy: The victim of a drunk driver has a privacy interest in not being hit, stripped naked and touched for emergency surgery, and dependent on assistance for the remainder of their life. More fundamentally, privacy is a negative right based on antagonistic relationships: “The right to be left alone.” It is defensive rather than optimistic.

The twin goals of human autonomy and community are more inclusive and inspiring than privacy alone. Autonomy is the freedom to discover oneself, be true to oneself, and live one’s own life; it requires some, but not complete, privacy. Community is connection with others, which often depends on “frictions” in life, such as interacting with strangers. While autonomy and community are in some tension, they are both part of happiness in the sense of leading a good life (eudaimonia). And so it may be more helpful to ask what promotes these goals, which might require data protection or data sharing. This is often a question of power, and in general power and privacy should be inversely related: The more powerful a person or a company is, the fewer privacy protections they should enjoy in law.

Several additional points from my talk last year can ground our discussions. First, focus these discussions by considering what is and is not unique about automated driving, what distinctions with respect to data collection and use might be helpful, and the various actors against whom one might assert a privacy interest. Second, safety offers a useful analogy for privacy, especially because a key question for both is whether the company behind a given technology is trustworthy (rather than whether the technology itself is trusted). Third, one of the key policy choices is who or what will be empowered: individuals, governments, companies, or other collectives – that is, communities. Artificial intelligence has a role here, especially in connecting and magnifying those with common interests. These points are further explained in last year's abstract.

A key example of these power dynamics is law enforcement. Far too many people die and are injured on our roads. This has led to a "safe systems approach" to traffic safety. Enforcement of traffic laws is also highly flawed, including in ways that demand a "safe and just approach" to this enforcement. Automated enforcement is part of this approach. It also raises a key question: Is there a difference between ideal enforcement and perfect (that is, complete) enforcement? I discuss case studies including graduated licensing, intoxication detection, and automated enforcement.

Private enforcement in combination with automation is especially striking. Imagine an automated driving company that punishes a wayward pedestrian by restricting that person's access to the company's automated driving services or even its other services. More generally, companies with evidence of a legal violation will choose how to interact with law enforcement – by sharing such data with police or the public at large without prompting, by demanding some private or public process for release of these data, or by seeking to make such release impossible. They may also be skeptical or deferential in their posture toward legal processes to order such release. A particular concern of power is when governments and companies "team up" with each other against individuals on matters of privacy. Some antagonism between these potentially powerful actors is helpful!

The safe systems approach can inspire and inform a "grand unified theory of enforcement." It involves thinking systematically and strategically, enforcing upstream, cultivating norms, and designing for feedback. I discuss each of these principles at length.

# 4 Working Groups

## 4.1 Working Group on Scenarios, Risks, Impacts, and Collected Data in CCAM

*Kevin Gomez (Technische Hochschule Ingolstadt, DE)*
*Adam Henschke (University of Twente, NL)*
*Bryant Walker Smith (University of South Carolina – Columbia, US)*
*Brigitte Lonc (IRT SystemX, FR)*
*Ben Brecht (Berlin, DE)*
*Stefan Gehrer (Robert Bosch LLC, Pittsburgh, US)*
*Christos Papadopoulos (University of Memphis, US)*

Our working group was interested in exploring the following set of issues related to Continuous Connected Automated Mobility/Autonomous Vehicles (CCAM/AVs)[1]: presenting scenarios to recognize and express the risks and potential impacts of collected data, examining whether there are unique features of *automotive privacy* that set it apart from other areas where information technologies may impact privacy, and discuss how law enforcement and international geopolitical issues distinguish CCAM/AVs from traditional discussions on legacy automotive vehicles and systems. The group's primary focus was on identifying scenarios, risks, impacts, and collected data in CCAM/AVs, which will serve as a foundation for further in-depth research to better understand the relationship between privacy and CCAM, considering the highlighted characteristics.

### 4.1.1 Discussed Problems

Our aim was to engage in a series of structured discussions that would facilitate the development of scenarios to identify and explain privacy issues in the context of CCAM/AV. We followed the following method: Each participant presented a set of questions to structure the subsequent analysis. We then presented various cases or issues that highlighted privacy-relevant concerns. In this process, we identified key stakeholders who might be vulnerable to privacy issues, whose activities could impact privacy, or who would be responsible for responding to such issues. Subsequently, we created a generalized scenario that outlined the key steps in CCAM/AV development and public release, allowing us to provide specific detailed examples relevant to privacy concerns. These examples were then organized into thematic clusters, as presented in Table 1. Building on this, we offer suggestions on how this analysis and thematic clustering can aid in identifying and communicating privacy issues in CCAM/AVs.

[1] A note on nomenclature: We use here the terms CCAM/AVs to capture issues in both connected automated driving and autonomous vehicles. While these terms may refer to different sets of technologies, they are also sometimes used interchangeably. Given that they are used differently and can refer to different sets of technologies whilst also referring to overlapping technologies, we consider that CCAM/AVs is a valuable way of encompassing the broader sets of discussions in which both CCAM and AVs present challenges for privacy. Note, however, that in connected versus autonomous driving, a lot of CCAM infrastructure might gather data that is independent of/orthogonal to AVs, i.e. CCTV, licence matching, intersection management, etc.

Following, we identified various structuring questions:

- Should there be any surveillance tools built into CCAM/AVs that actively help law enforcement?
- Should there be any restrictions on what LEOs (Law Enforcement Officers/Organisations) can access from CCAM/AVs?
- Who/what groups own vehicles and/or data?
- What sort of data is being discussed? Such as
  - Data in vehicle
  - Data outside the vehicle
  - Publicly collected data
- How to recognize and understand new sorts of data that can be gathered in CCAM/AVs, specifically medical/health data?
- What are the cybersecurity issues presented by CCAM/AVs? Such as
  - Misbehaviour detection
  - OEM outsourced services and cloud services
  - OEMs doing cloud services,
- What are the impacts on informed consent/privacy/confidentiality etc?
- Are CCAM/AVs considered critical infrastructure?
  - Noting that: If CCAM/AVs are designated as critical infrastructure, there may then be a need to share information with relevant national security institutions and LEOs, including the sharing of data in industrial control systems
- How to balance between the needs of LEO and the needs of customers or individuals?
- What role should an OEM play in this?
- (Why) is automotive privacy special from traditional driving with legacy vehicles and/or distinct from other information-gathering devices like smartphones?

Before delving into scenarios, risks, impacts, and collected data, the working group meticulously identified stakeholders in CCAM. We initiated the list with stakeholders for automotive digital forensics, as defined by Gomez Buquerin and Hof [1], considering that CCAM provides a relevant environment for such practices. However, we recognized that additional stakeholders were pertinent to this context beyond those identified by Gomez Buquerin and Hof. Thus, the final list of stakeholders is provided below: [list of stakeholders].

- Insurer (e.g., DEKRA, Allianz)
- Approval authority (e.g., UNECE approval entity)
- Business car owner (e.g., Telecom, Qualcom)
- Criminal (e.g., cybercriminal, state-sponsored attacker group)
- OEM (e.g., Volkswagen, Toyota, General Motors)
- Legal institution (e.g., police)
- Researchers (e.g., research institutes, universities)
- Supplier (e.g., Bosch, Continental)
- Tuner (e.g., MTM, Brabus)
- Private car owner
- Mobility provider (e.g., Uber, Lyft), including renters/fleet managers
- Road infrastructure
- Government:
  - Bureaucratic
  - Regulatory
  - Investigative

- Third parties:
  - Pedestrians
  - Passengers
  - Emergency responders
  - Road construction
  - Any individuals/groups that provide remote assistance/remotely facilitate (automated) driving,
  - End-of-life issues (particularly concerning remnant data),
  - Lawyers
  - Cyber-security actors
  - People borrowing a car
  - Leased car (owned by the bank)

The government as a stakeholder sparked discussions within the working group. Since the government acts on behalf of the population (in democratic countries), they should not have their own stake in privacy aspects of CCAM. However, the working group decided to add the government as a stakeholder due to its involvement in bureaucratic, regulatory, and investigative aspects.

Based on the stakeholders, we defined various scenarios where privacy is impacted or at risk in CCAM. Those are:

- In Germany, BMW as a police entity, change the vehicles for a police car. New vehicles are now not being rebuilt; how can ex-police vehicles safely delete information when selling after service?
- Capacity to use data in the wheel sensor TPMS IDs global IDs from the sensor in the tire.
- Privacy for autonomous vehicles is unique because of the scale of data – time, amount, and range of data (more potent than NEST), and the computing resources have significant storage, communication, and analytic power.
- Ownership of vehicles/data derived from it. Issue of fleets/transportation as a service.
- Are CCAM/AVs to be understood as a product or a service?
- Insurance – Are you insuring the driver or the system? Insurance companies push for the owner of data as being the owner of the data so that they can get ownership over that information.

### 4.1.2 Possible Approaches

Consider now a case of a safety incident involving a test vehicle on a public street. As part of the forensic investigation, the data gathered by the test vehicle may be requested to identify the factors around the safety incident and perhaps to identify individuals or groups who may be held culpable or deserving of some redress. A range of potential stakeholders would be affected by/involved in the privacy analysis. They would include:

- OEM/Research team; those gathering data, those storing the data and/or those with access to the data
- Government; investigators such as local police
- Third parties; Pedestrians, Emergency responders, other road users, lawyers,

The information being accessed would include the following:

- Visual data
- (Raw) sensor data
- Internal bus data
- Personally identifiable information (PII)

Consider now a case of a cybersecurity incident in which PII gathered as part of testing and evaluation was at risk of either being exfiltrated or altered. Misbehavior detection tools have identified that a database of test data had been the target of the attack, and now it must be established if this cyberattack was successful and if there are any privacy implications arising from it. A range of potential stakeholders would be affected by/involved in the cybersecurity investigation and any subsequent privacy concerns. They would include:

- OEM/Research team; those gathering data, those storing the data and/or those with access to the data
- Government; investigators such as cybersecurity forensic investigators/CERTs etc.
- Third parties; drivers involved in the vehicle testing, pedestrians, other road users, lawyers

Consider now a case where a CCAM/AVs service provider unintentionally identifies criminal behaviour in a fleet car/shared car. As part of this service, in-vehicle cameras and microphones monitor abnormal behaviour and will record events inside the vehicle to either recognize health events (such as a heart attack) or serious safety risks (violent activity between passengers). In this case, two vehicle occupants pretend to wrestle, activating the in-vehicle surveillance. Once activated, the occupants are recorded consuming controlled substances and discussing where they got them. It turns out that the substances were legal in one state, but now the vehicle has crossed state lines, and the substances are now illegal. The in-vehicle surveillance and recording would affect a range of potential stakeholders here. They would include:

- Third parties; vehicle passengers, lawyers
- Government; investigators including LEOs
- OEM; does the OEM that is monitoring non-private vehicles have a responsibility to report this illegal activity? Are they permitted to volunteer this information? What level of privacy should occupants in “non-private but not public spaces reasonably expect?
- What, if any role, does the passage from one jurisdiction to another play in the expectations of the passengers, the responsibility of the investigators, and the permissions of the OEM?

### 4.1.3 Conclusions

We identified various privacy implications on the different phases of the development and product life-cycle of modern vehicles, their functions, and services. The following tables (Table 1, Table 2, Table 3, Table 4, Table 5, Table 6) summarizes those.

#### References

1 Kevin Gomez Buquerin; Hans-Joachim Hof, *Identification of automotive digital forensics stakeholders*, SECURWARE 2021, The Fifteenth International Conference on Emerging Security Information, Systems and Technologies, p. 8-13 , 2021.

**Table 1** Identified privacy issues in test fleet data collection.

| | |
|---|---|
| Test fleet data collection | Is it in public or not? |
| | Chilling effects/experimentation |
| | Informality/rush to market/security as a later/afterthought |
| | Foreign data leakage/national security |
| | Where is the data stored/location/entity |
| | Should we collect everything – gauge engineering inclination to collect all |
| | UN regulation 155, must follow more than security best practice |
| | Need safety compliance to meet standing orders of NITSA, including reporting on prototypes, safety incidents |
| | Risks of shared/open data sets |
| | Public's first impressions/foundational artefact |

**Table 2** Identified privacy issues in in-car data collection and back-end communication.

| | |
|---|---|
| In-car data collection and back-end communication | Where is the data stored and how safe is it? |
| | Unauthorized access (entities, individuals etc.) |
| | Authorized access (entities, individuals etc.) |
| | Data retention policy including storage period |
| | Choices about what to track during processing |
| | Retention of raw data versus retention of processed data |
| | Potential for anonymization (when, where, and whether) |
| | Issues in security of communication of information/transfer/data accessibility means and methods |
| | Integrity of data set: Use and manipulation of data/data integrity/chain of custody/documentation |
| | Explainability/transparency/trust |

**Table 3** Identified privacy issues in training and testing algorithms.

| | |
|---|---|
| Training and testing algorithms | Pitfalls in machine learning, things like bias (sampling, parameters, inappropriate baseline) |
| | Inappropriate assumptions (i.e. threat model) |
| | Data insecurity: Find out what data was used to train the ML, black box or white box, you can then use that to infer the training data, model inversion from data leaking, remnant personally identifiable information, membership |
| | Deanonymisation |
| | Use of synthetic versus real data (privacy versus safety/accuracy) |
| | Imperfect approaches to anonymization |
| | Third party partners, especially labelling/post-processing |
| | Limitations of anonymization |
| | Digital twinning |
| | Lack of understanding of privacy implications of privacy (i.e. gait/walk might be privacy revealing) |
| | Overuse of data (might as well use, might need to use given algorithm, design pathways) |
| | Metadata |

**Table 4** Identified privacy issues in making and producing products or services.

| | |
|---|---|
| Making and producing products or services | Meeting/not meeting privacy requirements (e.g., GDPR), cybersecurity requirements etc., noting that they are different by jurisdiction |
| | Products of services that are incidental to AD (i.e., health monitoring) |
| | Product or services that necessarily involve PII (i.e., health monitoring) |
| | Pressures inherent in low profit margins |
| | Seeing/setting privacy as a goal or as a constraint, new limitations or design criteria, trade-offs between different components (given need to save money/cents per unit) |
| | Conflicts/trade-offs between marketers, managers, and engineers, i.e. overpromising (WRT privacy, promising services that are/potentially in conflict with privacy |
| | Privacy requirements (i.e. GDPR) that may be different given different jurisdiction |
| | Need to have inactivated privacy features (anticipated for a particular jurisdiction) |
| | Supply chain complexities and data disputes (e.g. android auto) |
| | Over the air updates/right to repair introducing/perpetuating vulnerabilities, hardware dependencies |
| | Open versus closed systems |
| | "Software defined cars": subscription features/subscription model of service provision, in vehicle marketing |
| | Legacy systems – both privacy and cybersecurity vulnerabilities |
| | Factory/location of production locations, i.e. issues of supply chain integrity |
| | Simplicity versus complexity – is the solution to security/privacy to create really complicated systems or really simple systems |

**Table 5** Identified privacy issues in after market operations.

| | |
|---|---|
| After market operations | Data leftovers – shared car/leased car (i.e. police example) arising from multiple users |
| | Oversharing by parents/friends/relatives etc., |
| | Road user privacy and |
| | Changing/disrupted concepts of privacy |
| | Inside vehicle (driver/owner, passengers), other road users in vehicles, other road users not in vehicles, PII from other vehicles |
| | No more service – who is responsible for ongoing privacy after OEM responsibility ends |
| | Buying privacy/monetizing privacy (inequities/fairness) |
| | Non vehicle privacy invasive infrastructure (i.e. intersection management, verification of ongoing operational safety), V2V, V2X communications etc. |
| | Used vehicles changing jurisdictions – vehicles systems designed for the privacy demands of one country moving to another country |
| | Updates |
| | General operations |
| | Re-purposing data, look for additional models/sources, desperation to monetize, corporate end of life |
| | Metadata and data creep |
| | Reporting and investigations of safety incidents |
| | Change of ownership |
| | Particular (public spaces and semi-public transport) |
| | Outsourcing the monitoring |
| | Companies cooping privacy i.e. good faith arguments of privacy are exploited, to resist the sharing of information, bad faith, NY goes to Uber wants to understand where people are travelling through the day, but Uber says no because they want to sell it, but use privacy as the excuse not to sell it |
| | Who owns the data |
| | Geopolitics – i.e., gathering national security significant surveillance data, economic competitions |
| | New privacy regulations that come up that need to be fulfilled/met |
| | Ongoing responsibility for OEMs, is the driver/owner responsible to update or not? |
| | Complex supply chains and streams of commerce |
| | Privacy related incidents and who is responsible? OEM, supplier, individual |
| | Mandatory reporting for cybersecurity events |
| | Cross border travel/enforcement |
| | Cyber-security incidents |
| | Cyber-security versus privacy trade-offs |
| | Vehicles as attractive targets, fleets at scale |

**Table 6** Identified privacy issues in selling and releasing products or services.

| | |
|---|---|
| Selling and releasing products or services | Over promising and hyping |
| | Pressures to take product to market, we can fix it later (issues more for engineers for early release) |
| | Downstream confusion (dealers etc.), Upselling of privacy risking components/-salesperson incentives |
| | Challenges for customer – mass and length of privacy policies |
| | Informed consent and ongoing responsibilities |
| | Legalistic notion of informed consent shifting responsibility to customer |
| | One off consent model versus dynamic and ongoing consent models, ongoing notion of consent |
| | Adam's conspiracy theories/trust issues |
| | Point of sale – When a vehicle is being constantly updated, when is the point of sale? (Software defined vehicles) |
| | Who is the customer/customers? |
| | Issues around fleet models – who is the customer/who is setting the privacy expectations (user, employer, business etc.) |
| | Disruption in the responsibilities of the salesperson – can a salesperson/should a sales person have to assess the competence of a customer's capacity to give informed consent – medical bioethics model |
| | Risk of information overload and decisions/consent fatigue |
| | Marketing pressures and the desire/incentive to say that everything is fine – competitive pressures on marketers |
| | Broad notion of supply chain integrity |
| | Tricky to sell privacy (because it is risk based rather than reward based) |

## 4.2 Working Group on Privacy Tensions for Connected Automated Vehicles

*Jonathan Petit (Qualcomm, USA)*
*Jason Millar (University of Ottawa, CA)*
*Sarah Thorton (NURO, USA)*
*Michael Buchholz (Ulm University, DE)*
*Zoltan Mann (University of Amsterdam, NL)*

### 4.2.1 Context and Objective

A common belief is that privacy comes at a cost, and hence, a tension exists between achieving full privacy and full "performance". This tension can be seen as finding the acceptable trade-off between privacy and the considered value/criteria. Let us consider the case of an automated vehicle driving in a city. To ensure road safety, an automated vehicle needs to detect pedestrians with high accuracy. Moreover, for accurate prediction and planning, the vehicle should be able to differentiate between different types of pedestrian, e.g., children, adult, impaired users. Indeed, depending on the user type, the motion model used by the behaviour prediction algorithm is adjusted. However, a privacy goal is to minimize those attributes. Therefore, a privacy-enhancing technology could only allow to output the coarse object "pedestrian", thus conflicting with the "required" granularity.

In order to comprehensively discuss the privacy tensions, the working group identified the lack of a common methodology. The objective of the group was then to propose a methodology.

### 4.2.2 Methodology

The objective of the methodology is twofold. First, to capture and rate the privacy tensions (positive or negative). Second, because identifying the privacy tensions is an iterative process, it is useful to understand the current coverage of the analysis. So the methodology should also output a coverage/completion value.

The proposed methodology follows an 5-steps approach. Note that this is a work-in-progress and will be refined in a future publication.

1. Specify use case: describe the scenario, its objective(s).
2. Identify stakeholders: list direct and indirect user/actors.
3. For each stakeholder, list respective assets, privacy needs and performance objectives.
4. Rate impact of dimension on privacy.
5. Assess current coverage/completion of the analysis for each dimension.

### 4.2.3 Conclusion

Creating a methodology to analyze privacy tensions (or synergies) is paramount to capture each dimensions (e.g., security, ethics, efficiency) and identify the technology readiness level of appropriate PETs. In a forthcoming publication, we will refine the methodology and validate it by applying it to case studies such as automated delivery service.

## 4.3 Automotive Privacy Engineering

*Ala'a Al-Momani (Ulm University, DE)*
*David Balenson (USC Information Sciences Institute, US)*
*Christoph Bösch (Robert Bosch GmbH, DE)*
*Kyusuk Han (Technology Innovation Institute, Abu Dhabi, AE)*
*Mario Hoffmann (ARRK Engineering GmbH, DE)*
*Sebastian Pape (Continental Automotive Technologies, DE)*
*Nataša Trkulja (Ulm University, DE)*
*Takahito Yoshizawa (KU Leuven, BE)*

Privacy engineering forms one of the core aspects to develop privacy-friendly products and services. Many components of privacy engineering and privacy-by-design have been introduced lately, yet their applicability to the automotive domain is still in its early stage. One example of privacy-by-design in vehicular communication is the 5GAA's whitepaper [1].

Our overarching goal in this working group is to identify existing or needed tools and frameworks to help commercial entities comply with regulations by embedding privacy into their products, and how the available tools can be adapted and tailored toward the automotive industry. A major part of this includes investigating how privacy engineering can be embedded in the software or product development life cycle. Furthermore, we aim to explore whether privacy strategies, as introduced by Hoepman [3] as well as privacy patterns [4] are applicable in the automotive industry and automotive scenarios in a straightforward manner, or whether certain modification is necessary to better suit special requirements in such scenarios. Moreover, we look into privacy enhancing technologies (PETs) and investigate whether these – once implemented – could achieve the system's functionality without introducing negative consequences due to specific automotive use cases, and at what cost PETs can be incorporated in a company's vision or implemented in its products.

To this end, it is of ultimate necessity to combine this technological and strategic privacy engineering effort with users' expectations and their behaviour when it comes to privacy protection.

This includes how a vehicle driver and passengers can be educated about vehicle data usage and, more importantly, its privacy implications, and how other road users and non-vehicle entity's privacy such as pedestrians' can be addressed. Intuitive design of user interfaces (UIs) needs to be used to ensure informed decisions about certain options for driver's or passengers' privacy. This may go beyond traditional consent forms that graphically appear on, e.g., webpages or mobile apps, to include other forms of human machine interaction (HMI) such as audio. The goal of such design should maximize transparency and avoid privacy dark patterns [5] in any UI design.

In the following, we discuss the questions addressed in this working group followed by our recommendations to enhance the applicability of privacy engineering in the automotive domain.

### 4.3.1 Discussion Questions

In this working group, we identify and discuss six questions that address some of the major aspects of privacy engineering in the automotive domain.

Q1. **What tools and frameworks are there to help commercial entities comply with regulations and embed privacy?**

Our goal here is to identify privacy tools that have already been applied to automotive, in addition to identifying and investigating the applicability of general privacy engineering tools to automotive.

We begin by pointing out that privacy in automotive is distinctive due to several factors such as the limited storage and processing power, specific protocols related to vehicle connectivity and communication (e. g, CAN bus), and the direct interaction with human safety among other trade-offs. In addition, automotive scenarios often rely on certain sensitive data and information such as location data and driving behaviour requiring privacy protection. We refer the reader to [2] for additional information about automotive data. A noteworthy point to consider here is that the entire automotive ecosystem is very complex. Modern architecture of automotive systems include vehicles, mobile devices, communications, and third-party connected services such as infotainment services. Furthermore, vehicles come with a complex supply chain consisting of multiple layers, for which privacy must be considered with different responsibility, accountability, and liability. To illustrate this, we consider an ASIL-like certification program made for privacy. It is yet unclear whether such a program would be available at different levels, OEM-level, Tier-1, Tier-2 (Privacy Certificate for suppliers) or it would be for OEMs only.

In the privacy engineering landscape, there exist various tools and frameworks that have the potential to be used in the automotive domain. Examples of such include LINDDUN [13] as a privacy threat modelling framework, privacy design strategies, privacy patterns [4], and various privacy-enhancing technologies. Conducting a threat assessment and risk assessment (TARA) in the automotive domain may require considering different scopes of OEMs and suppliers along with the need to have a hierarchical analysis with a privacy impact assessment (PIA) on vehicle level supported by PIAs on component levels. This could also lead to a split of responsibility and tasks. For example, on a vehicle-level there needs to be a data strategy while on a component-level the task is more focused on implementing a PET.

One source of inspiration for the automotive privacy engineering community to consider are adjacent domains such as IT, mobile systems, and e-health. Ideas may be inspired by investigating the privacy engineering challenges of those domains and investigate if the solutions addressed the challenges in those domains can be transferred.

In the next questions, we will dig deeper into certain tools and frameworks of privacy engineering and investigate how to adapt and tailor them towards the automotive industry.

Q2. **How to tailor privacy patterns to automotive industry, including what privacy patterns are there for automotive and leveraging architecture patterns?**

In this question, we want to particularly investigate whether current privacy strategies and privacy patterns [4] are directly applicable to the automotive domain.

In order to address this question thoroughly, we recommend the community to identify a set of automotive use cases and assess the applicability of existing patterns and define how to adapt such applicable patterns. We note that it may also be necessary to develop new patterns specific to vehicles, taking into account vehicle architectural patterns and the privacy challenges in future use cases, e. g., for automated driving and shared robo-taxis.

In this working group, we aim to provide an initial assessment of the applicability of patterns in the automotive domain. We randomly choose three privacy patterns from the pattern repository [4] and, on a high-level, investigate their applicability to certain automotive scenarios.

**Awareness Feed** This pattern states the importance of providing information to the end user concerning their privacy. However, in the automotive scenarios and particularly, the direct applicability of such a pattern in the vehicle to inform the user is challenging as this may divert the driver's attention and thus may lead to an unsafe situation in cases of not fully automated driving scenarios.

To this end, it is important to identify when it is a good time to show the drivers certain privacy information in a way that does not interrupt their attention on driving. This may include innovative and unorthodox methods of informing end users, such as vibrating the steering wheel, or usage of audio channels.

Applying this pattern is also challenging if we consider informing other users, such as passengers of a car/taxi/bus. How to ensure such users are aware, are able to give consent, or select certain privacy preferences remain unsolved. The topic is even more complex for other road users, which can be communicated with even harder.

**Informed Secure Passwords** This pattern requests to ensure end users select strong and long passwords with various characters for different services and applications. However, given that the UIs available in cars nowadays do not match those of mobile phones and keyboards, it is very challenging and critical to apply this pattern directly in the vehicle. It is generally hard to type passwords in current vehicle interfaces. This opens the door to different solutions to enter passwords in vehicles, or even to question the usability of passwords as authentication method in vehicles. The automotive privacy community needs to investigate other forms of authentication such as physical authentication tokens (e. g. Yubikeys), or bio-metric information, and assess whether these are better fit to the automotive use cases. Other challenges include identity management and having different profiles/roles for a certain identity, e. g., a business and a private profile.

**Location Granularity** This pattern deals with location data and states that precise locations should be used with less granularity, e. g., street, ZIP code or city name, if precise location is not needed. On a first glance, we foresee a direct applicability for automotive scenarios, but we point out that the applicability is context-dependent, i. e., what level of granularity is needed for the context and thereafter it needs to be set. One example is a difference in the granularity requirements for finding a nearby restaurant in comparison to ordering a taxi for pickup. In the first example, the rough area is sufficient to receive information from a service about restaurants. The latter requires a more specific rendezvous point to make sure passenger and taxi can meet.

Q3. **How can privacy engineering be embedded into the product or software development life cycle (SDLC) including threat modelling and verification of implementation?**

The privacy engineering process defined by Hoepman [11] should be applied and integrated into the SDLC. Furthermore, extended SDLC models for privacy such as the W-model [18] and the $\sigma$-model [19] should be applied. Such models extend classical SDLC used in automotive industry like the V-model to include privacy-by-design phases. Another more challenging problem is the integration of privacy-enhancing technologies into agile environments [20, 21], since PETs are hard to compose. There seems to be a lack of privacy methodologies in the right-hand side of the V-model, i. e., the testing phases. We note that formulating testing and evaluation is challenging for privacy aspects, but not only in the automotive domain. Often, there is no external data available at the time of testing a certain privacy objective, and this heavily depends on the context and on the runtime environment [10].

Q4. **Can PETs achieve system's functionality without blocking a certain functionality and at what cost can PETs be implemented in a company or a product?**

Different PETs have been proposed in the literature, to name a few:

**Multi-Party Computation (MPC):** Several parties in a system collaboratively compute an agreed upon function where respective inputs are secret.

**(Fully) Homomorphic Encryption ((F)HE):** FHE supports the computation on encrypted data without the need to decrypt it.

**Differential Privacy (DP):** By adding noise into the analysis output, it formalizes and measures how much privacy is brought relative to losing utility.

**K-anonymity:** Masking data such that every record is indistinguishable from $k - 1$ other records in a dataset.

**Trusted Execution Environments (TEEs):** Areas in processing units with secure interaction with the rest of the system where data is encrypted outside but decrypted inside this environment.

**Attribute-Based Credentials (ABC):** Credentials, based on attributes instead of identities, allowing anonymous credentials for role-based access control (RBAC) or attribute-based access control (ABAC). Selective disclosure of attributes forms a key to achieving anonymous authentication.

**Zero-Knowledge Proofs (ZKPs):** Proving whether a statement is correct or not without leaking more information.

We note that this is only a subset of PETs, further PETs that might be used in automotive scenarios are listed by Garrido *et al.* [7] [Tab. II, p. 3; Tab. III,p. 4]. Garrido *et al.* discuss PETs in automotive use cases and conclude that there is still the need for a deep understanding of the use cases and the proposed PETs. A noteworthy remark is that in some cases, more than one PET needs to be applied to achieve an overarching privacy goal.

It is ultimately necessary to investigate whether PETs can be used generally in automotive scenarios without reducing functionality. More particularly, without impacting trust, safety, and other distinctive aspects of the automotive domain.

Let's consider another possible scenario: cooperative intersection management using mobile edge computers. In this scenario, the edge, ideally, needs to know who you are, where you want to go, and your current position. In this scenario, the vehicle's position is needed but not a link to the driver's identity. **HE** may be used to compute the positions and directions in an encrypted fashion, but as a time-critical application, this may greatly impact the flow of traffic and the safety of road users. Considering that multiple vehicles are present in the vicinity with a known function to calculate, **MPC** might be useful.

Another solution to this scenario could be based on **ABC**, or privacy-preserving signature schemes, while using the minimal set of data needed in cleartext, e.g., the current location, velocity, steering angles, vehicle size category (e.g., car, bus, or truck with trailer), and the desired direction at the intersection (i.e., left, straight, right, or u-turn). This data could be used anonymously, i.e., without the vehicle, driver, and passenger identities. It would only be required to verify that the data truly originates from a valid vehicle, e.g., through the use of attribute-based credentials, or privacy-preserving signature schemes. However, it is worth noting that repetitive usage of accurate vehicle's data such as the vehicle's exact dimensions and direction of movement over a long period of time may increase the possibility of re-identification of that vehicle and/or driver [8].

To sum up, it is challenging to determine whether PETs may or may not be directly applicable in automotive scenarios and which one to use [7, 9]. Applying a PET in an application or a scenario requires clearly defining functionality requirements, privacy requirements, threat models, the bigger context of the systems and protocols involved in automotive scenarios.

Q5. **What are people's privacy behaviours and expectations: how can vehicle users be educated about vehicle data usage and its implication? What about the privacy of extravehicular entities (pedestrians, etc.)?**

In order to foster widespread adoption of privacy-friendly automotive features and scenarios, creating a demand from customers is crucial. This requires understanding the current people's perception of privacy when it comes to the automotive domain. Consequently, it is necessary to raise the awareness of privacy among customers to create the needed demand that will foster the deployment of privacy technologies into automotive scenarios.

The automotive privacy community needs to take into account the differences in people's perception of privacy that could stem from, e.g., cultural aspects based on regions or countries, which may impact how privacy options and controls can be designed and aligned with users' expectations. Therefore, we recommend conducting user studies on privacy expectations and actions specifically in the automotive space while taking into account cultural differences.

Additionally, we discussed the uniqueness of the automotive space over other domains such as mobile phones. This uniqueness may include further privacy-related factors such as driving behaviours and (in-cabin) cameras. Such additional factors would necessitate creating awareness of what data is collected, what PII is, and what seemingly is not PII, but can still be used to identify individuals or their driving patterns. The latter may be used to determine insurance rates or for re-identification of individuals using different vehicles in order to create and observe movement patterns. Thus, user awareness is essential in a transparent manner. When considering autonomous driving, different perceptions to the ones we discussed so far could be based on the level of autonomy. For example, at levels 4 and 5 there may not be any human driving pattern. However, privacy can be still an issue in new ways, such as user profiling in the car including, e.g., eye gazes, looking out the window, touching the steering wheel, etc.

Also, we need a comprehensive understanding of people's privacy perceptions under different use cases, that could be when users use their own car or using a shared car. The case of passengers' privacy needs also to be addressed adequately, that is, how passengers' privacy can be protected. To this end, this would require novel approaches of HMI considering the specific and uniqueness of the automotive domain. Enhancing customer awareness may take various forms and approaches. For example, the Vehicle Privacy Report website [12], provides privacy car facts for free by searching your VIN. It is necessary to study the best ways to increase user awareness and provide them with tools to learn about, understand, and adjust their privacy settings. One possibility is the "app" used for the scenario of shared car service. Another example that may be helpful to develop and use is the standardized privacy labels analogous to the energy-efficiency rating of consumer products, such as TV, fridge, etc. This links to the idea of a privacy score [15].

In fact, it is the tendency to collect data by default at any time when a new technology or service is deployed. Service providers will collect any and all data until they are told otherwise, for example, by regulators. On the other hand, regulating privacy can be very difficult. Even with privacy by default, providers will find a way around the regulation. Also, demands from the people themselves are needed to lead regulators to take action, bringing non-functional requirements like privacy as close to functional requirements and regulations.

We figure out the criteria for a user-friendly website, which includes "Easy to use", "intuitive", and "non-deception" (i. e., dark patterns [5, 6]). We need to tailor and apply such criteria to the automotive space, e. g. [14].

In summary, we consider that just relying on traditional privacy techniques may not be sufficient for automotive use cases. We encounter new challenges and need new creative ways for HMI, consent, informing, setting regulations, and preventing unwanted capture. Ultimately, we want people to drive knowing what's allowed without being deceived or tricked. Privacy labels and scores might be helpful to consider in this case.

Q6. **How can privacy controls (options) be designed in a way that maximizes transparency and avoids dark patterns in the UI?**

Given the specifics in the automotive context we discussed previously, we understand that providing information with long text should be avoided, instead, informative videos could be used. We need to consider providing alternatives to match different users, as some people prefer text, while others may prefer pictures or videos.

Informing the user versus legally binding may be two different things. Does it need to be text to be legally binding? Should the legal agreement be done at the same time when the user is informed about data handling? As of today, it seems to be the practice that information and legally binding consent are handled together. Thus, these texts are often written by lawyers and are hard to understand for the users. If those two actions are split, who would write the "understandable" text or create the video? Still Lawyers? Or, the Marketing team or Independent parties? Eventually, this leads to lots of questions about this process.

It was also unclear when would be the best time to inform the users respectively get their consent. When the car is purchased? Every time the door is opened? When setting up the car and occasionally repeated as a reminder? When there are updates? In general, users get tired of seeing the same notices and warnings and eventually start to ignore them, as we are currently observing with the cookie banners on web pages. How would the passengers be informed or give their consent?

Clearly, we need an independent authority to define some kind of privacy metrics and scores. Ideally, manufacturers (and customers) should respond to these metrics and define their privacy controls accordingly.

Lengthy and detailed information may not be available at the moment the personal data is being used. Also, people may not understand how the data is being processed and what can be derived from it, even if detailed explanations are provided. (Will transparency still apply here?)

Which data is really needed for a service? Given the current usage of "Legitimate interest", where service providers claim the need for all kinds of data for service improvement, marketing, and other analyses, it would be interesting to measure and identify the "bare minimum" of data needed: Given a certain scenario, based on the current state of the art what would be the minimum of data needed to provide a specific service?

### 4.3.2 Conclusion

Our working group sought to identify existing or needed tools and frameworks to help commercial entities comply with regulations by embedding privacy into their products, and how the available tools can be adapted and tailored toward the automotive industry. We explored six questions that address some of the major aspects of privacy engineering in the automotive domain, including identifying available tools and frameworks, tailoring privacy patterns, embedding privacy engineering in the SDLC, applying PETs to achieve needed functionality, learning about user privacy behaviour and expectation, and maximising transparency and avoiding dark patterns.

Throughout our discussion, we identified various action items that we believe are necessary to be addressed by the automotive privacy research and engineering community:

a. There is a need to confirm, adapt, reject existing privacy strategies and patterns. (cf. Question 2) in terms of their applicability to the automotive domain.
   i. It is necessary to identify appropriate privacy strategies (i.e., Minimize, Hide, Separate, Abstract, Inform, Control, Enforce, Demonstrate) and confirm that the existing eight strategies are appropriate and complete in the automotive space.
   ii. Within this context, it is necessary to determine which existing privacy patterns are valid and usable in the automotive context, which ones have to be adapted, and which ones are missing and need to be added.
   iii. Once the strategies and patterns are properly adapted to automotive context, then there is a need to map the privacy patterns to the steps of the development or product life cycle. (cf. Question 3)

b. Identify a set of automotive scenarios, including connected cars and autonomous driving scenarios, and analyze which data is the *bare minimum* needed for the functionality of the service, including the use of PETs. Additionally, identify and derive upper and lower bounds where possible. (cf. Question 4 and Question 6)

c. Investigate the possibility to test and evaluate implementations and adjust configuration of PETs in a way that fulfils their purposes. In other words, define methodologies that address the verification and the validation of privacy enhancing technologies in the automotive context. (cf. Question 3)

d. Investigate the necessary knowledge and skills for all stakeholders involved in the software and product development life cycle. (cf. Question 3)

e. Conduct user studies on privacy expectations and behaviours specifically in the automotive space, taking into account cultural differences. (cf. Question 5)

f. Utilize HMI in vehicles for information and awareness on data processing and consent requests with regard to the specific situation of the driver (and passengers) in the vehicle. This may need new ideas for HMI. (cf. Question 5 and Question 6)

### References

1 5GAA Automotive Association. *Privacy by Design Aspects of C-V2X, 2020.* https://5gaa.org/wp-content/uploads/2020/11/5GAA_White-Paper_Privacy_by_Design_V2X.pdf

2 European Data Protection Board. *Guidelines 1/2020 on processing personal data in the context of connected vehicles and mobility related applications, 2020.* https://edpb.europa.eu/system/files/2021-03/edpb_guidelines_202001_connected_vehicles_v2.0_adopted_en.pdf

3 Jaap-Henk Hoepman. *Privacy Design Strategies. CoRR*, vol. abs/1210.6621, 2012.

4 *Privacy Patterns (website).* https://privacypatterns.org/

5 Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. *Tales from the dark side: privacy dark strategies and privacy dark patterns. Proc. Priv. Enhancing Technol.*, 2016(4):237–254, 2016.

6 *Dark Privacy Patterns* dark.privacypatterns.eu/

7 Gonzalo Munilla Garrido, Kaja Schmidt, Christopher Harth-Kitzerow, Johannes Klepsch, Andre Luckow, and Florian Matthes. *Exploring privacy-enhancing technologies in the automotive value chain.* In *2021 IEEE International Conference on Big Data (Big Data).* IEEE, dec 2021. https://arxiv.org/pdf/2209.05085.pdf

8 Ezzini, S., Berrada, I. & Ghogho, M. *Who is behind the wheel? Driver identification and fingerprinting.* J Big Data 5, 9 (2018). https://doi.org/10.1186/s40537-018-0118-7

9 Sascha Löbner, Frédéric Tronnier, Sebastian Pape and Kai Rannenberg. *Comparison of De-Identification Techniques for Privacy Preserving Data Analysis in Vehicular Data Sharing.* In *CSCS '21: ACM Computer Science in Cars Symposium, Ingolstadt, Germany, November 30th, 2021.*

10 Blagovesta Kostova, Seda Gürses, and Carmela Troncoso. *Privacy engineering meets software engineering. On the challenges of engineering privacy by design*, 2020. https://arxiv.org/pdf/2007.08613.pd

11 Jaap-Henk Hoepman. *Privacy Design Strategies (The Little Blue Book)* https://www.cs.ru.nl/~jhh/publications/pds-booklet.pdf

12 Privacy4Cars, Inc. *Vehicle Privacy Report* https://vehicleprivacyreport.com

13 https://linddun.org/

14 The European Data Protection Board. *Guidelines 3/2022 on Dark patterns in social media platform interfaces: How to recognise and avoid them* https://edpb.europa.eu/system/files/2022-03/edpb_03-2022_guidelines_on_dark_patterns_in_social_media_platform_interfaces_en.pdf

15 https://privacyscore.org/

16 U.S. Department of Transportation. *Security Credential Management System (SCMS)* https://www.its.dot.gov/resources/scms.htm

17 Gürses, Seda and Troncoso, Carmela and Diaz, Claudia *Engineering privacy by design reloaded* Amsterdam Privacy Conference, 2015

18 Ala'a Al-Momani, Frank Kargl, Robert Schmidt, Antonio Kung, Christoph Bösch. *A Privacy-Aware V-Model for Software Development.* IEEE Security and Privacy Workshops (SPW), 2019.

19 Ala'a Al-Momani, Frank Kargl, Robert Schmidt, Antonio Kung, Christoph Bösch. *Poster: Towards A Reliable Privacy-Enhanced V-Model For Software Development.* IEEE Security and Privacy (SP), 2019.

20 Gürses, S., & Van Hoboken, J. *Privacy after the Agile Turn.* In E. Selinger, J. Polonetsky, & O. Tene (Eds.), The Cambridge Handbook of Consumer Privacy (Cambridge Law Handbooks, pp. 579-601). Cambridge: Cambridge University Press, 2018.

21 Carmela Troncoso. *Privacy technologies need to go to the gym: on the challenges of privacy engineering in an Agile world.* Keynote at IWPE19. San Francisco, US. May 2019.

## 4.4 Interplay between Privacy and Trust

*Thanassis Giannetsos (UBITECH Ltd., GR)*
*Frank Kargl (Universität Ulm, DE)*
*Ioannis Krontiris (Huawei Technologies – München, DE)*
*Francesca Bassi (IRT SystemX – Palaiseau, FR)*
*Anje Gering (Volkswagen AG – Wolfsburg, DE)*

The participants of the Dagstuhl Seminar recognized the challenge of converging privacy protection of vehicle data with the need to establish trust amongst the involved actors. For that reason, a dedicated Working Group was formed with the goal to dive deeper into the interplay between privacy and trust.

Privacy and trust concerns are of utmost importance in the CCAM field, particularly with respect to vehicles and all other road users (specifically vulnerable road users). Standardized protocols, like the vehicular PKI, that function as base for establishing trust in V2X communication have effectively addressed the privacy-respecting identity management of vehicles, incorporating measures such as digital certificates and robust authentication mechanisms. Nevertheless, the issue goes beyond the mere tracing of cars and encompasses the privacy and trust implications arising from a broader set of data and the interactions between all actors (vehicles, roadside users, MEC infrastructure, etc.).

Broadly speaking, a trust relationship is a directional relationship between two trust objects that can be called trustor and a trustee. The trustor is the “source” trust object as part of a trust relationship for which trust is assessed (one who trusts, the “thinking entity”, the assessor). The trustee is a “sink” trust object as part of a trust relationship for which trust is assessed (one who is trusted). Trustworthiness then can be defined as the measure of the likelihood of the trustee to fulfill the expectations of the trustor in a given context. One way to evaluate this likelihood is by assessing whether the trustee exhibits the right and relevant set of properties that enable it to meet the trustor’s expectations in a given trust relationship. For example, consider a trust relationship between a zonal controller within a vehicle and a camera ECU during a Cooperative Adaptive Cruise Control (CACC) function where the zonal controller is a trustor that relies on the camera ECU, the trustee, to deliver non-compromised camera data to it. Here, the camera ECU needs to exhibit, among others, the property of reliability. So, assessing whether the camera ECU is reliable in passing on its data to the ECU can give positive evidence of its trustworthiness.

An indicative set of properties that are relevant for evaluation of trustworthiness of systems in C-ITS and their components can be found in sources such as documentation on standards (such as ISO/IEC TS 5723:2022, ISO/IEC 22624:2020 and Recommendation ITU-T Y.3057), existing literature on autonomous vehicle systems and trustworthiness (such as [1]), and existing documentation on CCAM (Cooperative, Connected and Automated Mobility [2]). EU Project CONNECT [3] has processed these sources and came up with a elaborated list in Deliverable D2.1, focusing specifically on the CCAM domain.

Figure 1 illustrates a list of trustworthiness properties used to evaluate a trust relationship between a Trustor and a Trustee. During the discussions, the group elaborated on those properties and categorized them in three broad categories: based on performance, based on ethical aspects and based on user acceptance.

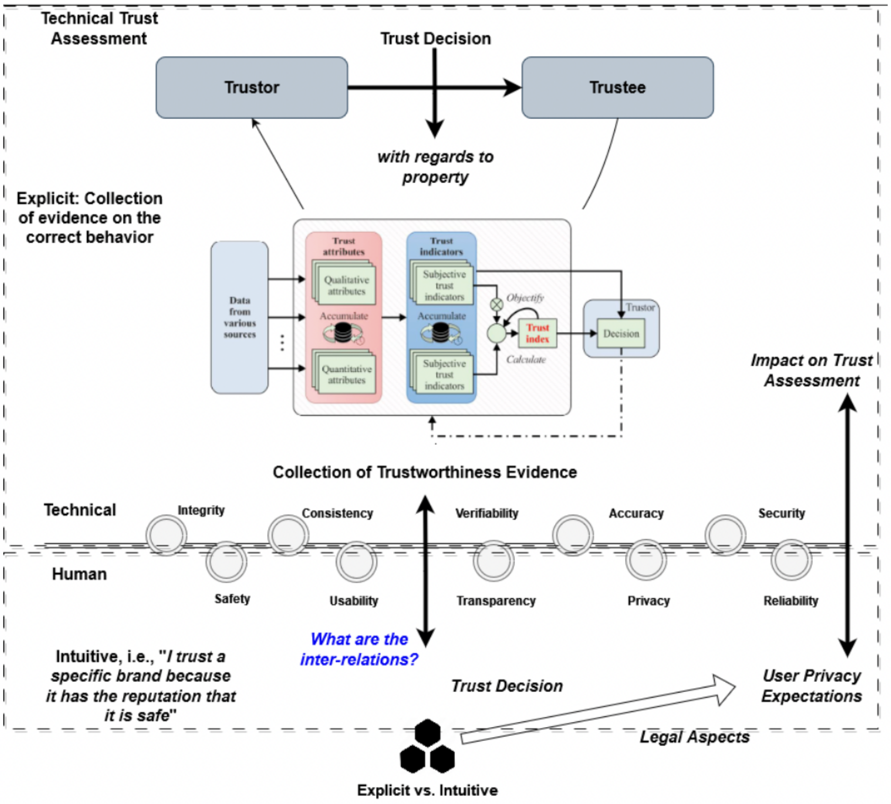

Technical Trust Assessment
Trust Decision
Trustor
Trustee
with regards to property
Explicit: Collection of evidence on the correct behavior
Data from various sources
Trust attributes
Qualitative attributes
Accumulate
Quantitative attributes
Trust indicators
Subjective trust indicators
Accumulate
Subjective trust indicators
Objectify
Calculate
Trust index
Trustor
Decision
Impact on Trust Assessment
Collection of Trustworthiness Evidence
Technical
Integrity
Consistency
Verifiability
Accuracy
Security
Human
Safety
Usability
Transparency
Privacy
Reliability
What are the inter-relations?
Intuitive, i.e., "I trust a specific brand because it has the reputation that it is safe"
Trust Decision
User Privacy Expectations
Legal Aspects
Explicit vs. Intuitive

**Figure 1** Assessing trustworthiness based on different evidence.

### 4.4.1 Trust Assessment Based on Performance

Trust and trustworthiness will play an increasingly important role as we shift towards higher levels of automation, because we need to rely on external data to facilitate partially automated or fully automated driving functions. In this context, the integrity and trustworthiness of external data sources, such as external sensor information, maps, and positioning data, becomes paramount. If the integrity of this data is compromised or not provided with the expected quality, the building blocks of the automated operation functions will use incorrect data to control the vehicle. There is a broad set of security attacks that have consequences on the trustworthiness of the data and data sources. The dependability and resilience of CCAM systems can be seriously affected by these attacks at run-time. Furthermore, there are many sources and reasons that can negatively impact dependability and safety that are not related to security. Properties like reliability, accuracy, and robustness are critical for providing consistent and dependable performance, while a property like resilience is essential for adaptability to various real-world scenarios, fostering user trust.

The issue of trust in CCAM extends beyond the realm of data and data sources. ETSI introduced the term Multi-access Edge Computing (MEC) [4] with the goal to bring processing power near the vehicle to meet ultra-low-latency requirements, and to reduce network traffic towards a centralized data-center. However, it is essential to acknowledge that such Edge Computing environments possess inherent characteristics of a complex and highly heterogeneous ecosystem due to the involvement of multiple vendors, suppliers, and stakeholders. In this context, several entities that belong to different trust domains must interact with each other to exchange privacy-sensitive data in order to enable safety-critical collaborative services. However, if these interactions are not properly managed, it can be the cause of privacy leaks. 5GAA published a report recently describing an in-depth analysis of the trust related threats of MEC in the automotive context [5].

EU Project CONNECT [2] addresses the above challenges by building a trust assessment framework for CCAM, which can measure and manage levels of trust under uncertainty, based on incomplete and/or subjective information provided by potentially untrustworthy sources. Furthermore, this framework can accommodate dynamically changing trust relationships due to the high level of mobility exhibited during the operational time of the systems at run-time.

### 4.4.2 Trust Assessment Based on Ethical Aspects

Ethics-based properties are of paramount importance when considering trustworthiness due to their direct impact on public perception and societal implications. For example, the trustworthiness property of accountability emphasizes the importance of data controllers and processors taking responsibility for their actions in managing personal data. This aligns with the ethical principle of accountability, which is key in building trust as it shows that an organization is willing to be answerable for its data practices. Similarly, privacy principles require organizations to be transparent about their data practices, including data collection, processing, and sharing. When individuals can easily understand and access information about how their data is used, it fosters trust in the organization's integrity. Explainability ensures that system actions are interpretable to users and regulators, addressing concerns about the "black-box" nature of AI (or components based on AI-based technologies).

In summary, adhering to privacy and data protection principles helps organizations demonstrate ethical behavior in their data handling practices, ultimately leading to increased trust from individuals and stakeholders. By upholding strong ethical principles, CCAM systems can build a foundation of trust with users and society, promoting widespread adoption and contributing to the safe and responsible advancement of autonomous mobility technologies.

### 4.4.3 Trust Assessment Based on User Acceptance

When discussing the notion of trust in CCAM, we cannot ignore the dimension of human trust from the side of the passenger that will eventually make use of the AV. In that respect, trust of people to the technology is a factor directly affecting the acceptance and adoption of AVs. Research has already demonstrated that the level of trust influences the acceptance of AVs [6].

One compelling interpretation of trust revolves around the sense of vulnerability experienced by individuals inside a vehicle due to the loss of control. In that sense, trust is defined as "the extent to which drivers willingly become vulnerable when using an AV" [7]. Another interpretation of trust is from the perspective of the existence of functionality, i.e. the degree of confidence drivers and passengers have in the predictability and functionality of the vehicle [8].

In order to better understand the human aspect of trust, Kenesei et al. [9] break down trust into three categories as follows: i) trust in the performance of the AV, ii) trust in the manufacturers of the AV, and iii) trust in the institutions responsible for regulating AVs. These dimensions of trust have been elaborated in previous research as well. Eiser et al. [10] point out that people might reject an innovation even if the technology is trustworthy, simply because the organisations behind the technology are not themselves considered as trustworthy. Liu et al. [11] adds another aspect to this dimension, raising the aspect of trust in jurisdiction. Hence, the concept of competence goes beyond mere ability, as it also includes the element of trust in governmental bodies tasked with formulating and implementing laws and regulations that assess the proficiency of these companies. These regulatory authorities grant certificates to brands that exhibit consistent adherence to the specified regulations. For instance, individuals can readily determine the extent to which different businesses adhere to the GDPR, highlighting the importance of proficiency and adherence to regulations as crucial factors in establishing confidence in the domain of data protection and adoption of technology.

Kenesei et al. [9] also explore the intricate interplay between trust and perceived risk. Indeed, when using an AV, the user should have sufficient trust that reduces the perceived risk of potential failure and misuse. More specifically, the authors examine two dimensions of risk: i) the perceived risk of the performance and hence security of the AV, and ii) the risk of misuse of the personal data that is exposed during use, which intersects with privacy protection considerations. Interestingly, their results indicate that privacy risk is influenced by trust in OEMs: trust in the manufacturer decreases the perceived risk of incorrect data handling.

In this light, it becomes evident that the policies governing how OEMs manage and process the collected data, should be considered. At the same time, the societal dimension, intricately linked to user acceptance, assumes a pivotal role in shaping trust perceptions. It becomes apparent though that the policy-making and the societal factors are intertwined, characterized by strong interdependencies that warrant thorough investigation. It is imperative to comprehend how these intertwined factors can influence users' perceived trust, and consequently, their acceptance of emerging technologies.

## References

1 European Commission, Joint Research Centre, Fernández Llorca, D., Gómez, E., Trustworthy autonomous vehicles – Assessment criteria for trustworthy AI in the autonomous driving domain, Publications Office of the European Union, 2021, https://data.europa.eu/doi/10.2760/120385

2 European Commission, Cooperative, connected and automated mobility (CCAM), [ONLINE] https://transport.ec.europa.eu/transport-themes/intelligent-transport-systems/cooperative-connected-and-automated-mobility-ccam_en

3 EU Project "CONNECT: Continuous and Efficient Cooperative Trust Management for Resilient CCAM", [ONLINE] https://horizon-connect.eu/

4 ETSI MEC ISG, "Multi-access Edge Computing (MEC); Framework and Reference Architecture," ETSI GS MEC 003 V2.1.1, January 2019

5 5GAA Automotive Association. *Cybersecurity for Edge Computing, 2023.* https://5gaa.org/content/uploads/2023/04/gmec4auto-cybersecurity-for-edge-computing.pdf

6 Shariff, A., Bonnefon, JF., Rahwan, I. Psychological roadblocks to the adoption of self-driving vehicles. Nat Hum Behav 1, 694–696 (2017), https://doi.org/10.1038/s41562-017-0202-6.

7 S. S. Man, W. Xiong, F. Chang and A. H. S. Chan, "Critical Factors Influencing Acceptance of Automated Vehicles by Hong Kong Drivers," in IEEE Access, vol. 8, pp. 109845-109856, 2020, https://doi.org/10.1109/ACCESS.2020.3001929.

8 Bernd Herrenkind, Alfred Benedikt Brendel, Ilja Nastjuk, Maike Greve, Lutz M. Kolbe, Investigating end-user acceptance of autonomous electric buses to accelerate diffusion, Transportation Research Part D: Transport and Environment, Volume 74, 2019.

**9** Zsófia Kenesei, Katalin Ásványi, László Kökény, Melinda Jászberényi, Márk Miskolczi, Tamás Gyulavári, Jhanghiz Syahrivar, Trust and perceived risk: How different manifestations affect the adoption of autonomous vehicles, Transportation Research Part A: Policy and Practice, Volume 164, 2022, `https://doi.org/10.1016/j.tra.2022.08.022`.

**10** Eiser, J. R., Miles, S., Frewer, L. J. (2002). Trust, perceived risk, and attitudes toward food technologies. Journal of Applied Social Psychology, 32(11), 2423–2433.

**11** Peng Liu, Zhigang Xu, Xiangmo Zhao, Road tests of self-driving vehicles: Affective and cognitive pathways in acceptance formation, Transportation Research Part A: Policy and Practice, Volume 124, 2019,

## Participants

- Ala'a Al-Momani
Ulm University, DE
- David Balenson
USC Information Sciences Institute – Marina del Rey, US
- Francesca Bassi
IRT SystemX – Palaiseau, FR
- Christoph Bösch
Robert Bosch GmbH – Renningen, DE
- Benedikt Brecht
Volkswagen AG – Berlin, DE
- Michael Buchholz
Universität Ulm, DE
- Stefan Gehrer
Robert Bosch LLC – Pittsburgh, US
- Anje Gering
Volkswagen AG – Wolfsburg, DE
- Thanassis Giannetsos
UBITECH Ltd. – Athens, GR
- Kevin Gomez
TH Ingolstadt, DE
- Kyusuk Han
Technology Innovation Institute – Abu Dhabi, AE
- Adam Henschke
University of Twente, NL
- Mario Hoffmann
ARRK Engineering GmbH, DE
- Frank Kargl
Universität Ulm, DE
- Ioannis Krontiris
Huawei Technologies – München, DE
- Brigitte Lonc
Nanterre, FR
- Zoltán Mann
University of Amsterdam, NL
- Jason Millar
University of Ottawa, CA
- Christos Papadopoulos
University of Memphis, US
- Sebastian Pape
Continental Automotive Technologies – Frankfurt, DE
- Jonathan Petit
Qualcomm, US
- Sarah Thornton
Nuro – Mountain View, US
- Natasa Trkulja
Universität Ulm, DE
- Bryant Walker Smith
University of South Carolina – Columbia, US
- Takahito Yoshizawa
KU Leuven, BE

Report from Dagstuhl Seminar 23251

# Challenges in Benchmarking Optimization Heuristics

**Anne Auger*[1], Peter A. N. Bosman*[2], Pascal Kerschke*[3], Darrell Whitley*[4], and Lennart Schäpermeier†[5]**

1 **INRIA Saclay – Palaiseau, FR.** anne.auger@inria.fr
2 **CWI – Amsterdam, NL.** peter.bosman@cwi.nl
3 **TU Dresden, DE.** pascal.kerschke@tu-dresden.de
4 **Colorado State University – Fort Collins, US.** darrell.whitley@gmail.com
5 **TU Dresden, DE.** lennart.schaepermeier@tu-dresden.de

## Abstract

This report documents the program and outcomes of the Dagstuhl Seminar 23251 "Challenges in Benchmarking Optimization Heuristics". In the domain of optimization heuristics, a stable basis for fairly evaluating the performance of optimization algorithms and other solution approaches – commonly referred to as "benchmarking" – is fundamental to ensuring steady scientific progress. Although many pitfalls are well known in the community, the development of sound benchmarking protocols is slow, and the adoption of community-wide recognized and implementable standards requires lasting and joint efforts among research groups. This seminar brought together people from diverse backgrounds and fostered discussions among different optimization communities, focusing on how to cope with "horse racing papers", landscape analysis techniques for understanding problem instances, and discussions about the overarching goals of benchmarking.

**Seminar** June 18–23, 2023 – https://www.dagstuhl.de/23251

**2012 ACM Subject Classification** General and reference → Empirical studies; Computing methodologies → Search methodologies

**Keywords and phrases** benchmarking, design of search heuristics, optimization, real-world applications, understanding problem complexity

**Digital Object Identifier** 10.4230/DagRep.13.6.55

## 1 Executive Summary

*Pascal Kerschke (TU Dresden, DE)*
*Anne Auger (INRIA Saclay – Palaiseau, FR)*
*Peter A. N. Bosman (CWI – Amsterdam, NL)*
*Darrell Whitley (Colorado State University – Fort Collins, US)*

### Motivation

The overall objective of the seminar was to explore the possibilities of defining how one can ensure that benchmarking is used to fundamentally advance the field of computational heuristics.

More often than not, in current practice, benchmarks are used to suit the needs of specific authors. That means that benchmark problems, including specific settings for benchmarks – such as how long to run the heuristics or what target performance(s) to achieve – are

[*] Editor / Organizer
[†] Editorial Assistant / Collector

cherry-picked by authors to make the tested algorithm look good. Moreover, the algorithms used for comparison are often cherry-picked too, and are not always considered state-of-the-art in the field. On the outskirts of our fields, as a consequence, one finds a proliferation of algorithms that have little basis in assumptions on problem structure that may be exploited but rather are vaguely based on biological or physical phenomena.

While benchmarking cannot stop this from happening, it can contribute to what is considered good practice at the core of the field, creating a stable basis for algorithmic advances and ensuring sensible comparisons.

## Seminar Structure

The organization of the seminar consisted of a mix of talks based on proposals from participants, discussions organized along breakout sessions, together with presentations that were encouraged by the organizers of the seminar in order to define common grounds among participants from different research fields.

## Outcome

In the various breakout sessions, ways to ensure good practices – in both theory and practice – were discussed for different (types of) problems that one often encounters. For some scenarios, such as the classic single-objective, non-expensive optimization case, advanced discussions led to definitions of ground rules for experimental studies on what is sometimes called “horse racing” algorithms. In other scenarios, where arguably comparisons are more difficult, such as multi-objective expensive optimization, discussions were more exploratory, yet key takeaways were formulated to be expanded upon in the future.

Related to this, in various talks, the related concept of landscape analysis has been discussed, potentially providing insights into why certain algorithms work well on certain problems, another hallmark of what we try to achieve through benchmarking. Whereas many of the former aspects are related more to the engineering aspect of algorithmic design, these aspects are more closely related to the scientific aspect of algorithmic design, increasing our understanding of what can and cannot be computed in a certain amount of time. On both sides of the coin, advances were made during the seminar, and bridges were built.

Overall, the seminar brought people closer together, advancing efforts on benchmarking from the fundamental (*how*) and the importance (*why*) perspective. The audience’s interdisciplinary nature helped define the palette of problems to create benchmarks for and understand different views on the same problem. From the various sessions, it became clear that there are lessons learned already that can inform the future creators of benchmarks to ensure that new benchmarks have added value and help to truly advance the field of optimization heuristics.

## 2 Table of Contents

**Executive Summary**

**Overview of Talks**

**Working groups**

# 3 Overview of Talks

## 3.1 The Role of Software in Benchmarking

*Thomas Bäck (Leiden University, NL)*
*Carola Doerr (Sorbonne University – Paris, FR)*
*Diederick Vermetten (Leiden University, NL)*
*Hao Wang (Leiden University, NL)*

At its core, benchmarking optimization algorithms might seem easy: we run some algorithms on some problems, collect data and analyze this result. However, the benchmarking pipeline can quickly become more complex when practical concerns are integrated. Software can be used to deal with these complexities by providing connections between parts of the benchmarking pipeline.

We discuss how to ensure these tools are extensible and how they contribute to the standardization of the experimental procedures. We also discuss how software facilitates adherence to benchmarking best practices. Finally, we focus on how the "barrier to entry" can be lowered.

## 3.2 What about the p-value – How and when should we "Test" reproducibility?

*Nikolaus Hansen (INRIA Saclay – Palaiseau, FR)*

We discuss the many ways how scientific publications can be false and remark that not each and every source of error is avoidable. Hence, readers of scientific literature always need to estimate (implicitly or explicitly) the likelihood that a conclusion is in essence false. The statistical p-value is a multiplier (Bayes factor) that decreases the odds ratio for H0 to be true. A small p-value is necessary, however not sufficient to reckon that H0 is (probably) false; to conclude the latter, we also need the prior odds of H0 or at least some plausible estimate thereof. When in doubt, any single (first) paper should be considered rather as hypothesis generating instead of hypothesis testing/confirming work.

## 3.3 Inconvenient Truths on Algorithm Competitions and Ways of Improving on Known Weaknesses

*Holger H. Hoos (RWTH Aachen, DE)*

Progress in solving challenging problems in artificial intelligence, computer science at large and beyond is driven, to a significant extent, by competition – regular algorithm competitions as well as comparative performance evaluation against state-of-the-art methods from the literature. A prominent example for this is the satisfiability problem in propositional logic (SAT), an NP-hard problem that not only lies at the foundations of computer science, but also plays a key role in many real-world applications, notably in ensuring the correctness of hard- and software. Unfortunately, these types of competitive evaluations suffer from a range of fundamental weaknesses; as a result, they can (and often do) produce an incorrect picture regarding the true state of the art in solving a given problem and, worse, create incentives misaligned with improvements thereof. Among these weaknesses are noise and low statistical significance, unfair and out-of-context tuning, and a focus on broad-spectrum performance achieved by single solvers. Therefore, new methods and approaches are required to analyse competition outcomes, to assess the strength of solvers, rather than the degree of tuning, and to exploit performance complementarity between different solvers for the same problem. Fortunately, as demonstrated in this presentation, such methods are now available; however, more work has to be done to enable and ensure their broad adoption.

## 3.4 Reproducibility in Optimization Research

*Manuel López-Ibáñez (University of Manchester, GB)*

**Main reference** Manuel López-Ibáñez, Jürgen Branke, Luís Paquete: "Reproducibility in Evolutionary Computation", ACM Trans. Evol. Learn. Optim., Vol. 1(4), pp. 14:1–14:21, 2021.
**URL** https://doi.org//10.1145/3466624

This talk discussed the topic of reproducibility in the context of optimization research. From a scientific perspective, reproducibility & falsifiability is how the scientific community reaches consensus. In addition, reproducibility has practical benefits in terms of error correction and building upon the work of others. The terminology around reproducibility may be confusing. ACM has proposed a terminology that is perhaps too general for optimization research. Recently, we have published a paper at ACM TELO, where we propose a more fine-grained classification of reproducibility levels. Each level has different purposes and not all of them are equally important. We discuss as well the cultural obstacles to reproducibility and how to overcome them.

## 3.5 Evolution of Benchmark Suites

*Olaf Mersmann (TH Köln, DE)*

The breakout group "Evolution of Benchmark Suites" focused on continuous optimization and specifically within the context of the COCO/BBOB benchmark suites, while highlighting their applicability to various domains. There was consensus that there is no one-size-fits-all approach and that alternative experimental regimes should be explored (as has been done by some groups). The group agreed that it is important to incentivise the design of novel benchmark suites. Questions raised include the consideration of precision for benchmark suites (e.g., float16/float32), the advantages and disadvantages of allowing competitors to submit functions/instances, the impact of evaluation cost depending on the number of decision variables changed, and strategies for collecting new benchmark functions. Benchmark suites should diversify and cater to different communities' needs, such as Neural Architecture Search (NAS) and Operations Research (OR), by introducing (artificial) real-world problems. This then led to discussions of the tradeoffs for implementers in terms of dependencies and runtime to ensure accessibility for casual users.

## 3.6 Synthetic vs. Real-World Landscapes: A Local Optima Networks View

*Gabriela Ochoa (University of Stirling, GB)*

Local optima networks (LONs) are a compressed model of landscapes where nodes are local optima according to given a neighbourhood and edges account for possible search transitions among optima (adjacency of attraction basins). LONs capture the connectivity pattern of local optima, and are thus useful to analyse and visualise the landscapes global structure and characterise funnels. This talk uses LONs to contrast the global structure of easy and hard instances as well as of synthetic functions against those of real-world problems. With an emphasis on visualisation, we show case studies in combinatorial and continuous optimisation, including hyper-parameter search spaces. We observe that hard instances have multi-funnel structures. Real-world problems have neutrality and symmetries that are generally absent in synthetic benchmarks.

## 3.7 Instance Space Analysis for Assessing and Generating Benchmark Instances for "Stress-testing" Algorithms

*Kate Smith-Miles (The University of Melbourne, AU)*

This talk provided an overview of Instance Space Analysis & the online tool MATILDA (`matilda.unimelb.edu.au`). A number of case studies were presented from combinatorial optimization (timetabling) & continuous optimization (BBO) to show how to create instance spaces & various strategies to evolve new benchmark test instances within the instance space boundary were discussed. Finally, the library of existing instance spaces in MATILDA were shown, spanning various problems in optimization, machine learning & model fitting.

## 3.8 RW-Benchmarking & Nevergrad

*Olivier Teytaud (Meta AI Research – Tournon-sur-Rhone, FR)*

We present the benchmarking suite in Nevergrad, which contains many published test functions. We have both:

- real-world functions;
- noisy optimization;
- discrete domains;

- low-dimensional to high-dimensional (we range from 2 to hundreds of thousands);
- continuous domains;
- multi-objective or single-objective;
- sequential or parallel optimization.

In addition, the platform contains many optimization methods, so that a user can reproduce all the runs and modify the context as she prefers.

During the seminar, some people pointed out how much it is risky to use all benchmarks which have been overfitted by so many people (Nevergrad is updated frequently and contains many benchmarks which did not exist 10 years ago), and that using a platform co-developped with an optimization method might lead to biased result (Nevergrad remains independent of any specific method and focuses on aggregated them and allowing modifications by whoever proposes a pull request). Various suggestions during the seminar have been taken into account; Gomea is already present in a branch, some chainings of Cobyla and ES have been added, and a cleaner export of results as a PDF file is now automatically generated. Applications to StableDiffusion, control of an AI player at Doom, and others have been discussed and (as of Sept. 5th) collaborations are in progress, in particular with the birth of NgIoh, a powerful black-box optimization wizard merged in Nevergrad 0.12.0.

## 3.9 Are Tree Decomposition Mk Landscapes Useful Benchmarks?

*Dirk Thierens (Utrecht University, NL)*

First, we reflect on what aspects define a good benchmark problem. Next, we discuss the CliqueTreeMk algorithm to construct tree decomposition TDMk Landscapes, whose global optimum can be computed efficiently using dynamic programming. In a Gray-box setting – this is, when the optimization algorithm knows the structural information of the tree decomposition, or equivalently, the problem variables interaction graph – TDMk Landscapes are too easy to solve to serve as benchmark problem for heuristic optimization algorithms. However, when used in a black-box setting – this is, when the heuristic optimization algorithm does not know the structural information – TDMk Landscapes are very well suited for benchmarking heuristic optimization algorithms that aim to learn dependencies between the problem variables while searching.

As an illustration, we discuss experimental results of the LinkageTree-GOMEA optimization algorithm on TDMk Landscapes with increasing overlap between the k-bounded subfunctions, that are unknown to the optimization algorithm.

## 3.10 Some Issues in Benchmarking Multiobjective Optimization Algorithms

*Tea Tusar (Jozef Stefan Institute – Ljubljana, SI)*

**Joint work of** Dimo Brockhoff, Tea Tusar

If we want to use benchmarking to support finding the best algorithm for a particular real-world problem, we need to construct a "knowledge database" on how various algorithms perform on a diverse set of problems. This has implications on the desired properties of benchmark problem suites as well as the used benchmarking methodology. We list some issues with how benchmarking is currently performed in multiobjective optimization and provide better alternatives to most of them. One very important remaining open question is how to construct a suitable suite of benchmark problems in a way that is resistant to overfitting.

## 3.11 Challenges in Optimizing Quantum Algos

*Hao Wang (Leiden University, NL)*

The quantum cost function induced from variational quantum algorithms brings an additional optimization challenge – the Barren Plateaus Problem – which essentially states that the variance of the partial derivative of the cost function diminishes w.r.t. the number of qubits.

## 3.12 101 Questions About Benchmarking Optimization Solvers

*Stefan M. Wild (Lawrence Berkeley National Laboratory, US)*

Intentionally provocative, we pose 101 literal questions, without offering a single answer. We begin with existential questions about the intentions, aims, and implications about benchmarking before specializing to settings such as heuristics/nonheuristics, randomized solvers, stochastic optimization, constrains, multi-objective, parallel computing, and machine learning. We conclude with questions about sociological & ethical considerations about benchmarking. Selection of the questions was naturally biased and rigorous discussion regarding missed questions followed.

# 4 Working groups

## 4.1 Breakout Session on Benchmarking in the Expensive Multi-Objective Optimization Setting

*Peter A. N. Bosman (CWI – Amsterdam, NL)*
*Mariapia Marchi (ESTECO SpA – Trieste, IT)*

### 4.1.1 Summary

This breakout session focused on what benchmarking should look like in case the problem at hand is expensive and multi-objective, which happens in several real-world scenarios. Most established benchmark problems however are single-objective and not necessarily expensive, but may sometimes be treated as such (by having a lower budget in terms of time or evaluations). The question arises whether such benchmarks are useful and whether we would not need better benchmarks that better reflect reality. This comes with several questions that we tried to answer during the 2 breakout sessions we had.

### 4.1.2 Key Considerations

1. What is expensive? This is not a priori clear in general, so it should be part of the benchmark. Generally, the consensus is that it means that the number of evaluations available is less than +/- 100d and that d is typically in the order of tens of variables, not more.
2. Should the global optimum be known? As it is not to be expected that we can find the global optimum within the restricted budget available, this is generally perceived as something that is not required.
3. Should there be a pre-phase (design of experiments) defined? In practice, often, there is a phase in which one would get a few trials first before running a large-scale experiment or real optimization run. For this reason, it is generally assumed that it would be good/realistic if a pre-phase is allowed. However, it is generally agreed that it is probably best if the benchmark provides a few evaluated solutions for the sake of repeatability and fairness.
4. Should the benchmark problems themselves be expensive? It is generally agreed that this should not be the case, otherwise the benchmark will likely not find its way into use by researchers. It is probably best therefore to use surrogates of real-world problems for benchmark purposes.
5. Should objectives be evaluable separately? In practice, expensive optimization may arise in situations where a simulator is involved. In such cases, objectives can often not be evaluated separately. Moreover, in situations where you can do this, the benchmark would be different, especially when one objective is much cheaper than another. Therefore, it is advisable to categorize such problems into distinct classes of benchmark problems.
6. Should the benchmark problems have constraints? It is generally agreed that if the benchmark problems are to be representative of real-world problems, there should always be constraints. There are, however, different types of constraints that could distinguish different classes of benchmark problems. In particular:
   - Algebraic/simulation-based constraints
   - Quantifiable/unquantifiable constraint violations
   - Relaxable/unrelaxable constraints
   - Hidden constraints

Beyond this, simulation-based benchmarks should integrate a probability of failure, that could either be systematic or random. The probability setting used should then be reported as part of the benchmark setting.

### 4.1.3 Other Considerations

Other considerations discussed within the breakout sessions were existing benchmarks, such as EXPOBench (which is only single objective), having different evaluation times for different objectives that make the problems difficult in other ways (i.e., without the complexity class of the objectives/problems themselves changing), multi-objectivization of single-objective expensive optimization problems, and the fact that comparisons between algorithms are very difficult for various reasons. Firstly, comparisons between multi-objective optimization algorithms are in general tricky (what indicator(s) to choose). Secondly, several expensive optimization scenarios can make comparisons even more difficult. E.g., if objectives can be evaluated separately, and one objective is cheaper to evaluate than others, spending more evaluation (or time) budget on the cheaper objective may give very different results and overly positive values for indicators, whereas final approximation fronts are skewed, actually. For this reason, additional descriptions are needed for benchmark problems, e.g., of how evaluations can be spent.

### 4.1.4 General Recommendations

1. For expensive optimization benchmarks, do not compare optimizers the same way as in “standard” optimization benchmarking (do not race horses in expensive races).
2. To make fairer comparisons, explicitly take into account the cost of evaluations rather than only using the more classic numbers of evaluations.

## 4.2 Breakout Session on Competitions vs Empirical Analysis

*Tobias Glasmachers (Ruhr-Universität Bochum, DE)*
*Emma Hart (Edinburgh Napier University, GB)*
*Holger H. Hoos (RWTH Aachen, DE)*
*Manuel López-Ibáñez (University of Manchester, GB)*
*Kate Smith-Miles (The University of Melbourne, AU)*

Topics:

- different needs for benchmarks for competition vs. empirical analysis
- terminology: how to clearly disentangle the two?
- do’s and don’ts for horse-racing papers

Participants:

- Tuesday: Tobias, Konstantinos, Kate, Katharina, Lennart, Anne, Pascal, Emma, Holger, Manuel
- Wednesday: Carolin, Kate, Katharina, Emma, Carola, Lennart, Pascal, Holger, Anne, Konstantinos, Tobias, Manuel
- Thursday Session 1: Manuel, Konstantinos, Kate, Carola, Emma, Holger, Carolin, Pascal, Lennart
- Thursday Session 2: Carola, David, Carolin, Kate, Katharina, Emma, Holger, Manuel, Pascal, Lennart, Lars, Konstantinos

Terminology:

- "horse racing" describes taking a performance snapshot.
- "benchmarking" is pretty much the same as horse racing.
- "empirical analysis" pursues different goals.

### 4.2.1 Horse Racing

These types of studies have many problems:

- Differences are rarely statistically significant, effect sizes are small. Drawing strong conclusions should be avoided.
- Bias is unavoidable, sometimes even desired. Must be made explicit.
- Competitions can be extremely motivating and drive relevant progress. But they are not "scientific".
- Competitions and horse racing rarely highlight contributions made in relevant niches.

### 4.2.2 Recommendations

During its final session, the working group fixed a comprehensive list of minimal and optional criteria for quality horse-racing papers. The list will be finalized after the Dagstuhl Seminar. It is intended to support the review process of top journals in the field in the future.

The preliminary list of necessary requirements agreed on by the breakout session participants is the following:

1. METRICS: Clearly define and justify metrics that you compare on (performance, budget, variance, worst-case performance, etc.); in particular, deviations from commonly used performance metrics (especially those used in the literature on the state of the art) must be described.
2. SELECTION OF PROBLEMS/INSTANCES: Benchmark instance selection needs to be defensible (e.g., benchmarks widely used in the recent literature in combination with a similar metric) and not biased towards making the new algorithm look better than it is (no cherry-picking); if you deviate from standardized benchmark collections by using only a subset of it, explain why you have decided to deviate and how the problems/instances/data sets have been chosen as well as the rationale behind this choice; if you create a new dataset need to clearly explain properties and why existing benchmarks are not suitable
3. BASELINES: Compare against reasonable baseline algorithms (i.e., state of the art, as documented in the literature or known from competitions – the way of determining the state of the art needs to be explained; simpler baselines, such as random search, Latin Hypercube sampling or similar can also be used if there is demonstrable value in it)
   - explain how the state-of-the-art has been identified
   - what, if no state of the art exists so far? (rare case, but could happen) [Then, focus on "simple" baselines such as random search or some naive local search?]
   - What if the state-of-the-art is not available as open-source?
4. EXECUTION ENVIRONMENT: When running times (CPU/GPU times, wall-clock times), time-outs or mem-outs are reported, the execution environment needs to be specified (including information such as CPU/CPU make and model, number of cores, speed, cache size, RAM size, OS version).
5. TUNING: Unfair tuning and performance optimisation (carried out manually or automatically) must be avoided whenever possible, otherwise, a compelling explanation must be given; this includes choice of programming language, degree of parallelisation, compiler optimisation, configuration and parameter tuning. If tuning and performance optimisation

is performed, it should be reported and done equally for all algorithms equally, i.e., same tuning instances, budget / effort spent; specifically, baselines should be tuned in the same way as the new algorithm; when using automated configuration, the initial configurations should include the default configuration (and configurations recommended by the original authors for similar problems).

6. STATISTICAL VALIDITY: Statistical validity of claims should be assessed and reported (using statistical tests, confidence bounds, or any other widely accepted method capable of detecting lack of validity in observed performance differences)
7. FRAMING THE CLAIMS IN THE CONTEXT OF THE EXPERIMENT: Conclusions drawn must be carefully stated in terms of the experimental setting considered by the horse race, and broader generalisations that are not yet supported should be avoided (unless many of the desirable criteria have been met for a more insightful experimental analysis enabling broader conclusions about a new algorithm's power).
8. REPRODUCIBILITY: Results should be reproducible (in the sense captured in well-established reproducibility checklists, e.g.,those from AAAI, AutoML conf, NeurIPS, JAIR – to be released, GECCO tutorial checklist, ACM Artifact Review and Badging . . . ), and limitations to reproducibility must be stated and justified.

## 4.3 Breakout Session on Reinforcement Learning for Grey-Box Evolutionary Computation

*Vanessa Volz (modl.ai – Copenhagen, DK)*
*Tobias Glasmachers (Ruhr-Universität Bochum, DE)*
*Boris Naujoks (TH Köln, DE)*
*Mike Preuß (Leiden University, NL)*

**Main reference** Erik A. Meulman, Peter A. N. Bosman: "Toward self-learning model-based EAs", in Proc. of the Genetic and Evolutionary Computation Conference Companion, GECCO 2019, Prague, Czech Republic, July 13-17, 2019, pp. 1495–1503, ACM, 2019.
**URL** https://doi.org//10.1145/3319619.3326819

### 4.3.1 Motivation

In evolutionary computation and consequently, in related benchmarking setups, we most commonly target a black-box optimisation scenario, i.e. the problem needs to be solved without prior knowledge or prior training / tuning. However, in practice, there are many scenarios that instead allow some insight into the problem. Take, for example, the optimisation of a medical treatment plan [6]. While the exact instance of the problem might not repeat for different patients, the problems certainly have similarities that the algorithm could be trained to exploit. Another scenario with similar properties is designing the floorplan for the physical layout of a computer chip [3].

In our breakout sessions, we aimed to investigate with a small experiment how black-box optimisation problems can be formulated in a manner that allows for training across different instances, i.e. problems of similar nature. For such recurring problems, techniques from the domain of reinforcement learning seem to be suitable, as they learn policies across different but similar problems. We thus devised some initial experiments towards expressing these described grey-box problems in a benchmark, with baselines from RL/EC hybrids.

### 4.3.2 Related Work

While there are different approaches for framing an environment for an evolutionary algorithm in this context, we chose to focus on a dynamic algorithm configuration setting. This is because step-size adaptation in evolutionary computation has been shown to be beneficial [1], but is still an open problem as demonstrated by the fact that a benchmark was proposed recently [2].

Additionally, reinforcement learning has been shown to work well in a setting where it is responsible for dynamic algorithm configuration. A framework for applying reinforcement learning to train model-based evolutionary algorithms (MBEAs) has been proposed in [4]. Further, dynamic step-size adaptation for CMA-ES has been demonstrated to outperform manual configuration in [5]. The authors further show that the trained policies can be applied to different function classes as well as higher dimensions.

Backed by these successful results in different settings, in these breakout sessions, we were aiming instead to target a simplified setup in order to be able to investigate general and theoretical hypotheses.

### 4.3.3 Experiment

We therefore chose the sphere function along with various transformations as our problem class. We then tried different ways of formulating an environment suitable for reinforcement learning (RL) agents. Concretely, we set up an OpenAI gym environment [7] specifically to target the sphere function in continuous space. Even if we assume that the environment frames the interaction as there being a single agent with a specific position in search space, there are still many options for defining the action space.

In this case, we chose to imitate an $(1, \lambda)$ evolution strategy with our setup. The only action the algorithm can take is to choose the variance used for generating lambda new individuals around the previous location. The best offspring is chosen automatically.

In our small experiment, we specified:

1. Action: Action $a$ results in variance $v$ for generation of offspring, where $v = 10^a$ and $a \in [-10, -1]$
2. Observation: Observation $o$ is the log of the distance from the current fitness value $f_t$ to the optimal one $f^*$, so $o = \min(9, \lceil \log(f_t - f^*) \rceil)$
3. Reward: Fitness improvement $f_{t-1} - f_t$ of the chosen action in log-scale, so $\log(f_{t-1} - f_t)$

In order to allow for simple RL approaches, the values above are discretised by using the log. This formulation further encodes domain knowledge about optimising sphere functions by grouping states with a similar distance to the goal together. In our experiment, we then applied a simple Q-Learning algorithm to the problem, as well as a baseline Proximal Policy Optimisation (PPO) algorithm.

### 4.3.4 Results and Discussion

As expected, the agent is able to learn to reduce the step-size the closer it gets to the known optimum. It is able to reach the optimum (up to a specified precision) in a similar timeframe as CMA-ES for an unseen problem instance.

However, in this experiment setup, we made several assumptions that benefit the RL agent.

1. The action, observation and reward space make use of the fact that we are working with a sphere function, as they are basically discretising the values by assigning them to a concentric band around the known optimum.
2. We assume that we know the optimum for the reward.

After the initial setup as described above, we are aiming to start our investigation first on the sphere function, and later potentially other problem classes as well. We are specifically going to investigate different environment formulations and their effects on the algorithm performance. For example, formulations with and without known optima should be compared. However, this knowledge may not be as important as first thought as in RL, we do not have to provide an *exact* reward but can e.g. go with just indicating a reward whenever an improvement has been reached. We thus hypothesize that assumption 2 can be circumvented.

Overall, we are aiming to determine general recommendations that can then be transferred to more complex and practical problem settings and evolutionary algorithms.

### References

1 B. Doerr, C. Doerr and T. Kötzing. *Provably Optimal Self-adjusting Step Sizes for Multi-valued Decision Variables*. Parallel Problem Solving from Nature – PPSN XIV., pp. 782-791, 2016.

2 T. Eimer et al. *DACBench: A Benchmark Library for Dynamic Algorithm Configuration*. International Joint Conference on Artificial Intelligence, IJCAI 2021, pp. 1668-1674, 2021.

3 A. Mirhoseini et al. *A graph placement methodology for fast chip design*. Nature 594(7862), pp. 207-212, 2021.

4 E. Meulman and P. Bosman *Toward self-learning model-based EAs*. Genetic and Evolutionary Computation Conference (GECCO) Companion, pp. 1495-1503, 2019.

5 G. Shala et al. *Learning step-size adaptation in CMA-ES*. Parallel Problem Solving from Nature – PPSN XVI., pp. 691-706, 2020.

6 N. Luong et al. *Application and benchmarking of multi-objective evolutionary algorithms on high-dose-rate brachytherapy planning for prostate cancer treatment*. Swarm and Evolutionary Computation 40, pp. 37-42, 2018

7 G. Brockman et al. *OpenAI Gym*. arXiv:1606.01540, 2016

## 4.4 The Concept of Generalization for Optimization Algorithms

*Hao Wang (Leiden University, NL)*
*Thomas Bäck (Leiden University, NL)*
*Gabriela Ochoa (University of Stirling, GB)*
*Dirk Thierens (Utrecht University, NL)*
*Sebastien Verel (Calais University, FR)*
*Diederick Vermetten (Leiden University, NL)*

Training, testing, overfitting, and generalization are all well known concepts in the domain of machine learning. We propose to develop similar concepts for optimization heuristics, to train (tune) an algorithm for a set of problem instances, to test it on problem instances that are “similar enough”, and thereby to demonstrate that the tuned algorithm can generalize to other problem instances that are “similar enough”. We contrive to provide a first definition of the necessary concepts such as *similarity of problem instances* and *generalization*.

### 4.4.1 Introduction to the Related Breakout Sessions

We were discussing the concept of "generalizability" in optimization theory, by which we intuitively mean the idea that if an optimization algorithm $\mathcal{A}$ performs well on an optimization problem instance $f_1$, it should also perform well on a sufficiently similar problem instance $f_2$.

However, there are many open questions/loose ends in the above intuition. For instance, what we mean exactly by "similar problems (instances)" in the context of optimization.

- The concept of *instance similarity* could potentially be measured by *distance in feature space*, for which we would need features that describe instance characteristics appropriately.
- We can distinguish between instance features (those that can be extracted from the instance data) and landscape features (which require sampling the fitness landscape involving neighborhood operators).
- Interpretable features are important for experts/user to develop/understand the concept of generalizability, if possible.
- A large number of fitness landscape features have been already defined [2]: A single feature can not explain the whole search difficulty, several features can be linearly correlated, and on the contrary, a combination of features can be meaningful. Indeed, optimal relevant set of features is an open problem, and suppose to be problem domain dependent.
- Feature normalization is important for developing such a distance measure.
- It makes a big difference, whether we consider combinatorial or continuous/numerical optimization problems.
- Information content, derived from random walk data on the decision space, was theoretically proven to be strongly related to the fluctuations of the gradient field of a continuous objective function [5].
- We discussed the idea of whether neural networks could be used to automatically extract features. Some works are dedicated to this direction [6].
- For continuous space with black-box optimization scenario, a sampling a search space is a way to discretize the continuous space. Based on this sampling, a neighborhood relation between sampled points can be defined in order to have discrete fitness landscape which approximate the original continuous one, and extract standard combinatorial fitness landscape features. Several sampling techniques can be used: static one such DoE [11], or adaptive one [9].

There are three main application domains for the features, namely (i) for *algorithm selection/algorithm performance prediction*, (ii) for defining *instance similarity*, and (iii) for defining *instance hardness*. The underlying assumption is that sufficiently similar problem instances would imply that an algorithm also yields similar performance on these instances. Based on that, we could come up, potentially, with a definition of "generalization".

More (somewhat random) ideas that relate to these concepts:

- If we assume we have a set of training instances, we also have a set of baseline functions and could use a metric between sets of functions as a means to measure the similarity, e.g., Hausdorff distance. The most straightforward way to measure the similarity between two functions is the $L^p$ norm.
- Notice that a measure of similarity based on features will depend on the scaling of the features values. So, it would be important to normalize the feature values (according to problem dimension, variance of values, maximum/minimum, quantity of information, etc.) to improve the meaningful of similarity measure.
- We need to develop a cross-validation analogy with machine learning: enumerate or randomize all/many possible 80/20 splits.

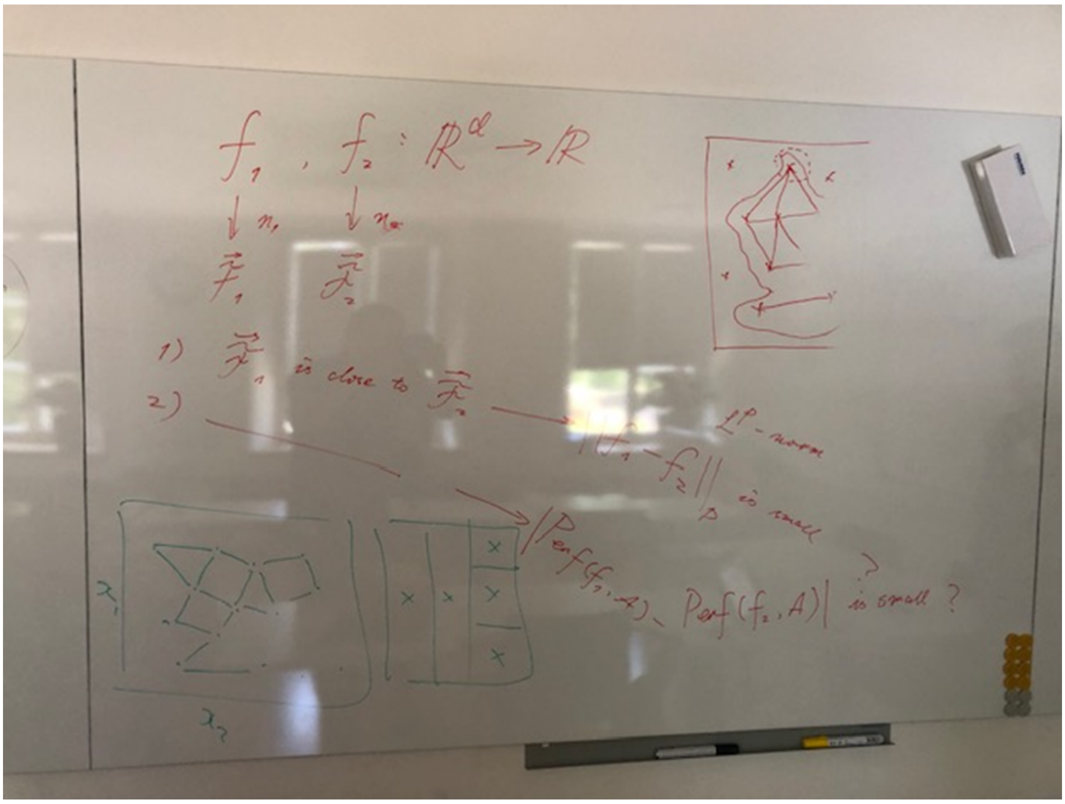


**Figure 1** Being close in feature vector space implies the underlying problem instances are similar, and the difference in performance of algorithm $\mathcal{A}$ on these instances is similar.

- Further open questions:
  - Which features should be used?
  - Which performance measures should be used?
  - Example in the combinatorial domain: QUBO.
  - Example in the continuous domain: BBOB (or a subset thereof).
- (Probably approximately correct) PAC learning analogy:
  - Tuning hyperparameters of optimizers to problems.
  - Tuning hyperparameters of ML algorithms to data.

### 4.4.2 Feature-based generalizability

We assume the set $\mathcal{L}^p(\mathbb{X},\mu)$ of measurable functions (w.r.t. Borel sets on $\mathbb{X}$) $f$ from $\mathbb{X}$ to $\mathbb{R}^k$, where the domain $\mathbb{X}=\mathbb{R}^d$ for continuous black-box functions and $\mathbb{X}=\{0,1\}^d$ for pseudo-Boolean functions (similar story for combinatorial problems). We shall assume the single-objective scenarios here ($k=1$). Naturally, the $p$-norm is defined for functions $f_1, f_2 \in \mathcal{L}^p(\mathbb{X},\mu)$ as follows:

$$\|f_1 - f_2\|_p = \left(\int_{\mathbb{X}} |f_1 - f_2|^p \,\mathrm{d}\mu\right)^{1/p}.$$

We also consider a set of black-box optimization algorithms $\mathcal{A}=\{A_i\}_i$ and an empirical performance measure $\mathrm{Perf}\colon \mathcal{A}\times\mathcal{L}^p(\mathbb{X},\mu)\to\mathbb{R}$, subject to maximization. Note that the empirical measure is essentially a random variable since it uses a finite set of independent runs/executions of an algorithm on a function to quantify the empirical performance.

For continuous black-box optimization problems/functions (which are infinite-dimensional objects), the $p$-norm can only be computed with Monte Carlo method (convergence rate: $\mathcal{O}(m^{-1/2})$ according to CLT; $m$ is the number of function evaluations/data samples), which can be costly and unreliable. Hence, it is desirable to define some sample-efficient *landscape features* that are consistent with the $p$-norm. Denoting by $\tilde{\mathbf{f}}$ the numerical features of a function $f$, we have the following intuitive criteria on assessing the appropriateness of numerical features:

1. *Consistency:* $\tilde{\mathbf{f}}_1$ is close to $\tilde{\mathbf{f}}_2 \implies \|f_1 - f_2\|_p$ is small: distances in function space is bounded by distances in feature space.
2. *Usefulness:* $\tilde{\mathbf{f}}_1$ is close to $\tilde{\mathbf{f}}_2 \implies |\operatorname{Perf}(A, f_1) - \operatorname{Perf}(A, f_2)|$ is small: performance difference is bounded by feature difference;
3. *Effectiveness:* for almost every function $f$ in $\mathcal{L}^p(\mathbb{X}, \mu)$, the convergence rate of $\tilde{\mathbf{f}}$ (considered a statistical estimator) should not be slower than $\mathcal{O}(m^{-1/2})$ according to CLT, where $m$ is the number of function evaluations.

Going beyond this, one could be even more optimistic and assume that, if the two functions are similar, even the best performing algorithms $A_i^*$ (or at least hyperparameter settings for a given algorithm) for these two functions should be similar (e.g., in terms of their code, assuming they are programmed in the same programming language). This results in the following requirement (see also figure 2):

$$\tilde{\mathbf{f}}_1 \text{ is close to } \tilde{\mathbf{f}}_2 \Rightarrow \operatorname{dist}(A_1^*, A_2^*) \text{ is small.} \tag{1}$$

(although defining distance metric among algorithms is also a nontrivial task) Here, we assume that

$$A_i^* = \arg\max_{A \in \mathcal{A}} \operatorname{Perf}(A, f_i). \tag{2}$$

A close enough goal, to start with, would be to say that the algorithm is not different, but for the same algorithm we are assuming the distance between their optimal hyperparameter configurations $\theta_i$ is small, i.e., $\|\theta_1^* - \theta_2^*\|$ is small and

$$\theta_i^* = \arg\max_{\theta \in \Theta} \operatorname{Perf}(A(\theta), f_i) \ . \tag{3}$$

### 4.4.3 Feature-free definition

Another formulation attempt, as in figure 4, is based on the idea that we can use a training set $F_{\text{train}}$ of functions to "train" an algorithm $A$ and a test set $F_{\text{test}}$ to "test" whether $A$ generalizes thereto. In that case, we would require something like

$$\exists L < \infty, \quad \frac{|\operatorname{Perf}(A, F_{\text{train}}) - \operatorname{Perf}(A, F_{\text{test}})|}{D_H(F_{\text{train}}, F_{\text{test}})} \leq L \ , \tag{4}$$

where $D_H$ is the Hausdorff metric between the training and testing sets.

$$D_H(F, G) = \max\left\{\sup_{f \in F} \inf_{g \in G} \|f - g\|_p, \sup_{g \in G} \inf_{f \in F} \|f - g\|_p\right\} . \tag{5}$$

#### 4.4.3.1 Example

To make things clearer, we were then trying to define an approach to test things in reality, both for the continuous domain $\mathbb{R}^d$ and the binary domain $\{0,1\}^d$. As test problem domains, we could use, say, training set $F_{\text{train}} = \{f_i\}_i$ to be 50 instances selected u.a.r. from BBOB, and test set $F_{\text{test}} = \{g_i\}_i$ to be 10 instances from BBOB, with AUC under the ECDF curve being the performance measure. Likewise, for the binary domain we had the idea to use QUBO problem formulations with a tree width parameter. A formulation following equation (4) would then, loosely formulated, look like

$$R(A, F_{\text{train}}, F_{\text{test}}) = \frac{|\,\text{Perf}(A, F_{\text{train}}) - \text{Perf}(A, F_{\text{test}})|}{D_H(F_{\text{train}}, F_{\text{test}})} \,. \tag{6}$$

Now, imagine both $F_{\text{train}}$ and $F_{\text{test}}$ are sampled from training $\mathcal{F}$ and testing $\mathcal{T}$ function families, respectively. Then, we can compute the above ratio $R$ for multiple test sets, which are generated/sampled randomly from the testing family $\mathcal{T}$ of functions. The *empirical generalizability* of algorithm $A$ from training family $\mathcal{F}$ to $\mathcal{T}$ can be calculated as $\sup\{R_1, R_2, \ldots\}$, where $R_1, R_2, \ldots$ are the ratio values obtained on multiple testing sets.

Since for a (infinite) function family, the above ratio $R$ can only be computed via a finite subset of functions, and therefore this ratio becomes a random variable. In this sense, it is natural/beneficial to provide a probabilistic formulation of generalizability (PAC-learning like):

$$\Pr(R(A, \mathcal{F}, T) \geq \delta) \leq U(\delta) \,, \tag{7}$$

for some upper bound function $U$, to be developed in the future.

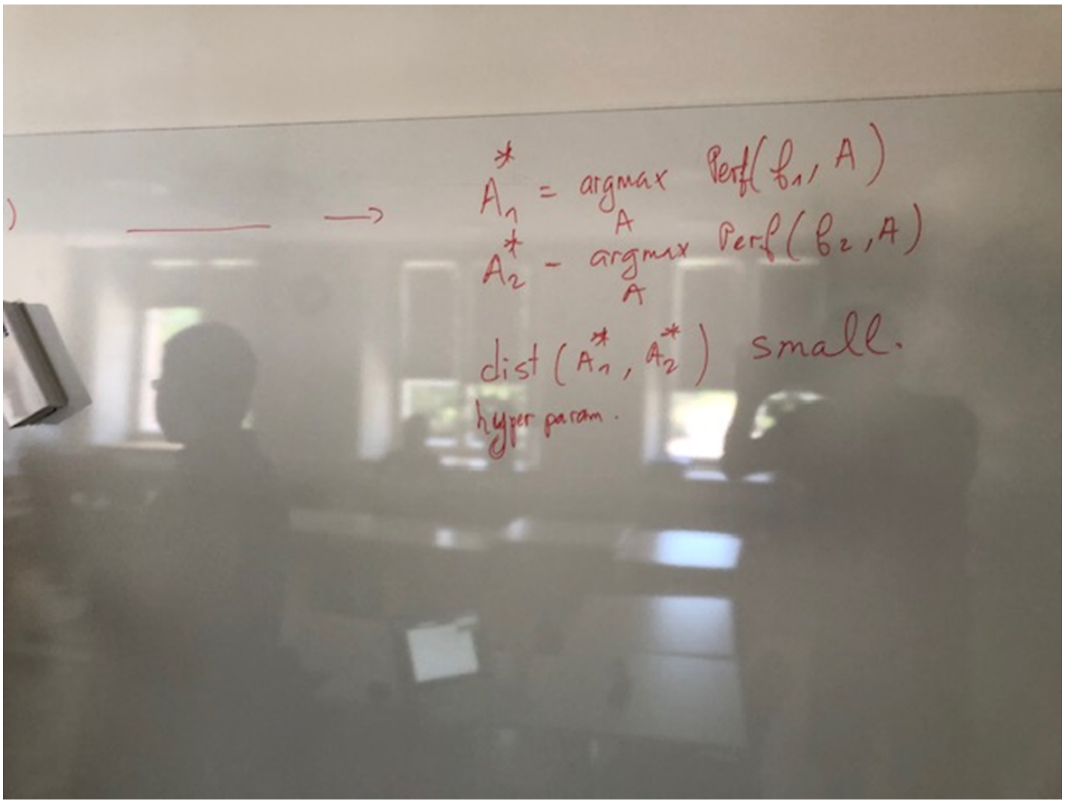

$A_1^* = \underset{A}{\text{argmax}}\ \text{Perf}(f_1, A)$
$A_2^* - \underset{A}{\text{argmax}}\ \text{Perf}(f_2, A)$
$\text{dist}(A_1^*, A_2^*)$ small.
hyper param.

**Figure 2** ... and even maybe that the best performing algorithms are similar, e.g. at least in terms of hyperparameter settings (maybe in terms of "code similarity").

### 4.4.4 Papers of Interest and Related Work

- Survey of fitness landscape features (in both discrete and continuous optimization): [1, 2]
- Features for combinatorial multi-objective problems: [7]
- Nearly the same features for continuous multi-objective problems: [8, 11, 9]
- An adaptive way to sample continuous single objective problems to create some possible features: [9]
- MA-BBOB paper: [10]
- Hyper-heuristics and cross-domain optimization [3, 4], are approaches that seek to increase the level of generality of optimization algorithms. They are practical algorithmic methods to solve complex combinatorial problems, and have not devoted much effort to quantifying the notion of generality of solvers.

#### References

1 Malan, K. & Engelbrecht, A. A survey of techniques for characterising fitness landscapes and some possible ways forward. *Inf. Sci.*. **241** pp. 148-163 (2013), https://doi.org/10.1016/j.ins.2013.04.015

2 Malan, K. A Survey of Advances in Landscape Analysis for Optimisation. *Algorithms.* **14**, 40 (2021), https://doi.org/10.3390/a14020040

3 Burke, E., Gendreau, M., Hyde, M., Kendall, G., Ochoa, G., Özcan, E. & Qu, R. Hyper-heuristics: a survey of the state of the art. *J. Oper. Res. Soc.*. **64**, 1695-1724 (2013), https://doi.org/10.1057/jors.2013.71

4 Ochoa, G., Hyde, M. & Others HyFlex: A Benchmark Framework for Cross-Domain Heuristic Search. *Evolutionary Computation In Combinatorial Optimization (EvoCOP).* **7245** pp. 136-147 (2012), https://doi.org/10.1007/978-3-642-29124-1%5C_12

5 Pérez-Salinas, A., Wang, H. & Bonet-Monroig, X. Analyzing variational quantum landscapes with information content. *ArXiv Preprint ArXiv:2303.16893.* (2023)

6 Stein, B., Long, F., Frenzel, M., Krause, P., Gitterle, M. & Bäck, T. DoE2Vec: Deep-learning Based Features for Exploratory Landscape Analysis. *Companion Proceedings Of The Conference On Genetic And Evolutionary Computation, GECCO 2023, Companion Volume, Lisbon, Portugal, July 15-19, 2023.* pp. 515-518 (2023), https://doi.org/10.1145/3583133.3590609

7 Liefooghe, A., Daolio, F., Vérel, S., Derbel, B., Aguirre, H. & Tanaka, K. Landscape-Aware Performance Prediction for Evolutionary Multiobjective Optimization. *IEEE Trans. Evol. Comput.*. **24**, 1063-1077 (2020), https://doi.org/10.1109/TEVC.2019.2940828

8 Liefooghe, A., Vérel, S., Lacroix, B., Zavoianu, A. & McCall, J. Landscape features and automated algorithm selection for multi-objective interpolated continuous optimisation problems. *GECCO '21: Genetic And Evolutionary Computation Conference, Lille, France, July 10-14, 2021.* pp. 421-429 (2021), https://doi.org/10.1145/3449639.3459353

9 Derbel, B., Liefooghe, A., Vérel, S., Aguirre, H. & Tanaka, K. New features for continuous exploratory landscape analysis based on the SOO tree. *Proceedings Of The 15th ACM/SIGEVO Conference On Foundations Of Genetic Algorithms, FOGA 2019, Potsdam, Germany, August 27-29, 2019.* pp. 72-86 (2019), https://doi.org/10.1145/3299904.3340308

10 Vermetten, D., Ye, F., Bäck, T. & Doerr, C. MA-BBOB: Many-Affine Combinations of BBOB Functions for Evaluating AutoML Approaches in Noiseless Numerical Black-Box Optimization Contexts. *CoRR.* **abs/2306.10627** (2023), https://doi.org/10.48550/arXiv.2306.10627

11 Liefooghe, A., Verel, S., Lacroix, B., Zăvoianu, A. & McCall, J. Landscape features and automated algorithm selection for multi-objective interpolated continuous optimisation problems. *Proceedings Of The Genetic And Evolutionary Computation Conference.* pp. 421-429 (2021)

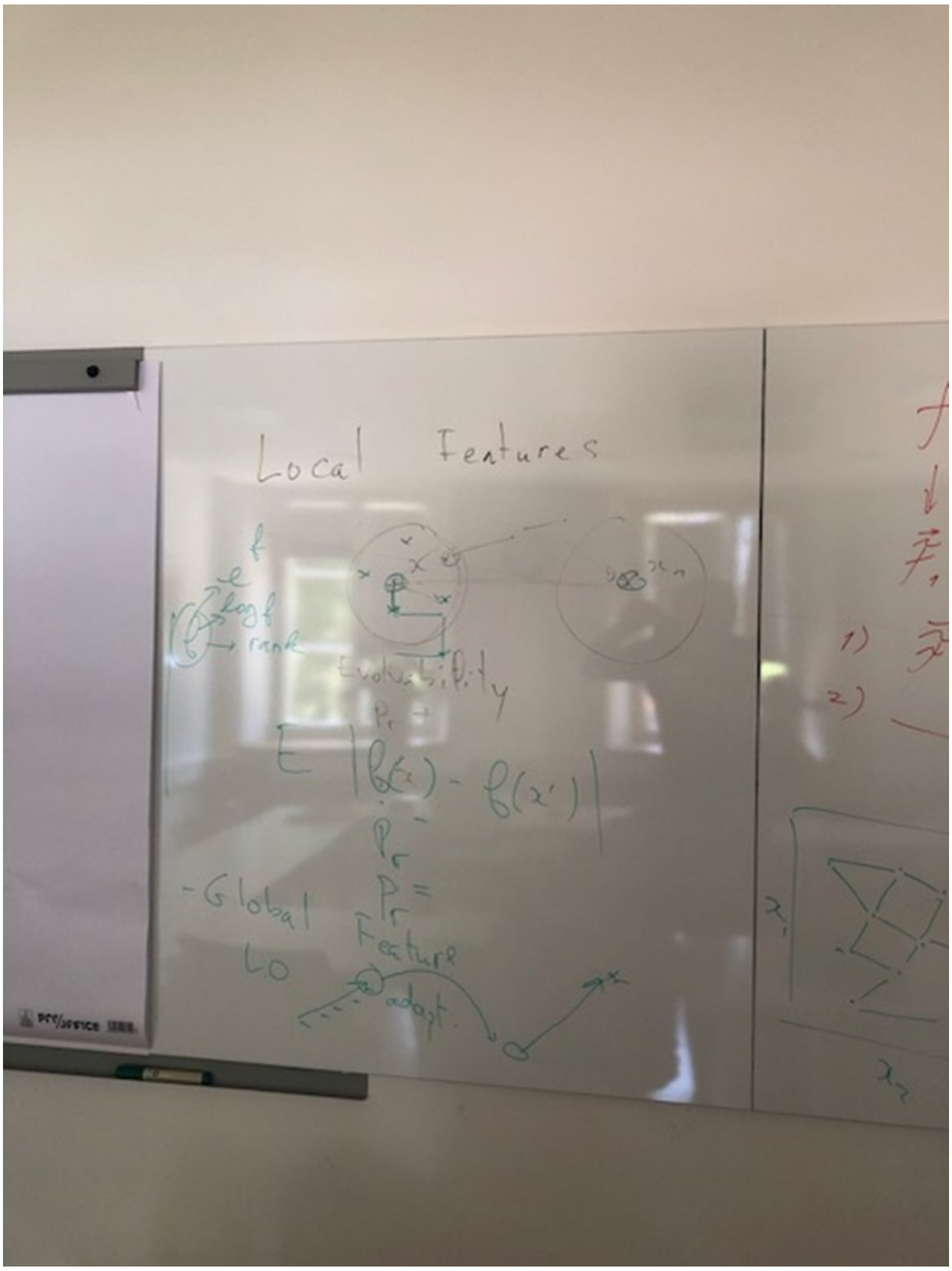


**Figure 3** Local feature computation.

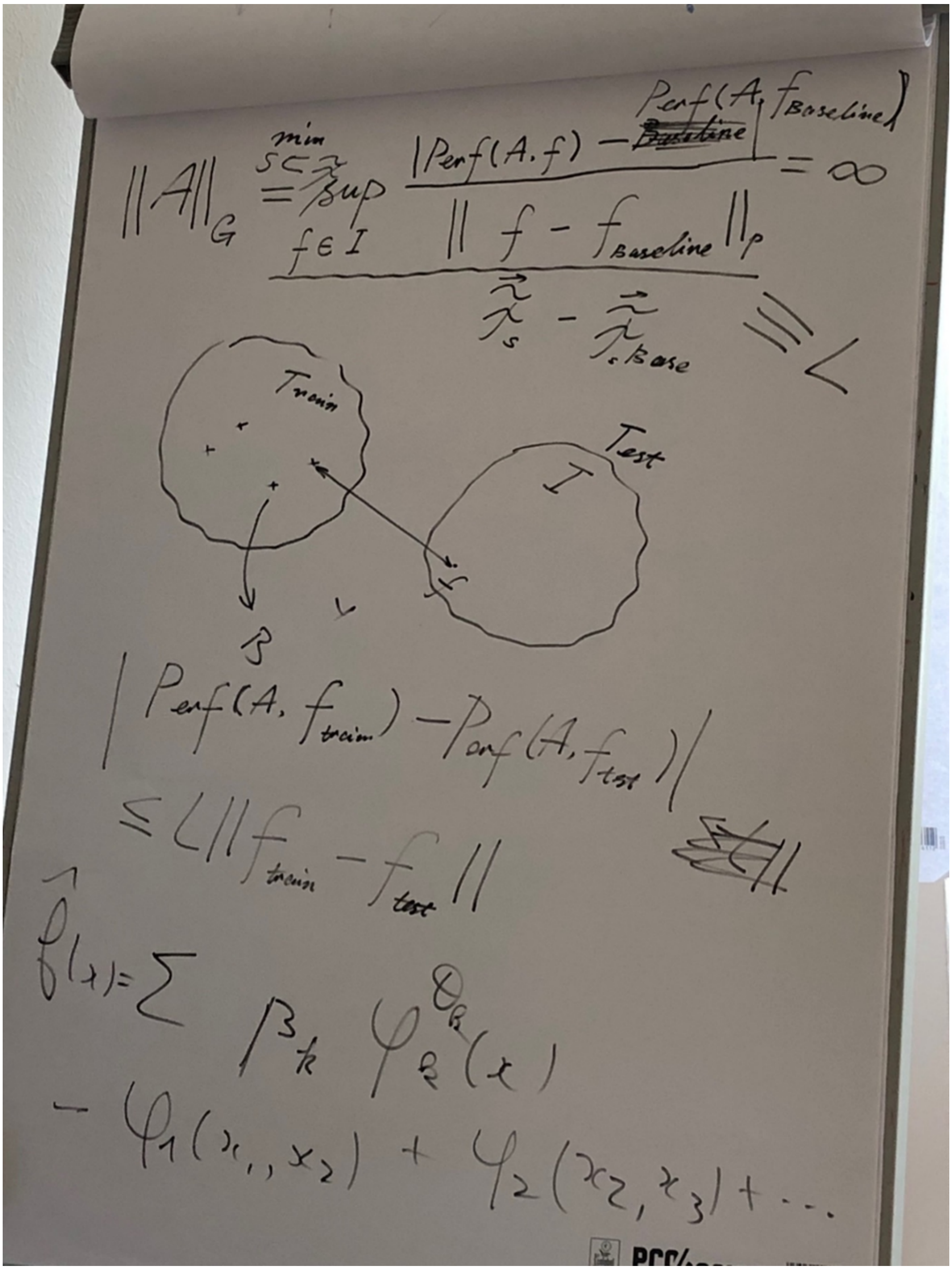

$\|A\|_G = \min_{S \subset \vec{x}} \sup_{f \in I} \dfrac{\dfrac{|Perf(A.f) - Perf(A, f_{Baseline})|}{\| f - f_{Baseline} \|_p}}{\vec{x}_S - \vec{x}_{Base}} = \infty$
~~Baseline~~
$\leq L$
Train
Test
$I$
$B$
$|Perf(A, f_{train}) - Perf(A, f_{test})|$
$\leq L \| f_{train} - f_{test} \|$
$\hat{f}(x) = \sum \beta_k \varphi_k^{\theta_k}(x)$
$- \varphi_1(x_1, x_2) + \varphi_2(x_2, x_3) + \dots$

**Figure 4** Trying to formalize the concept of "generalization". Sup links it to worst-case instance.

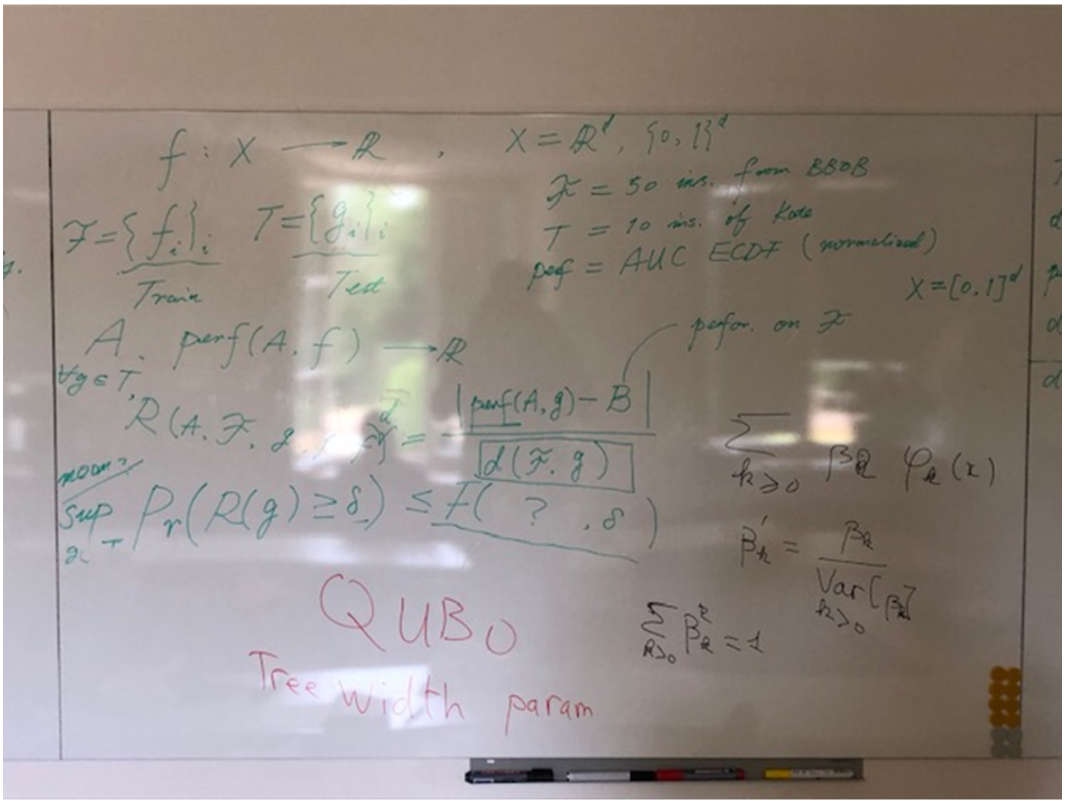
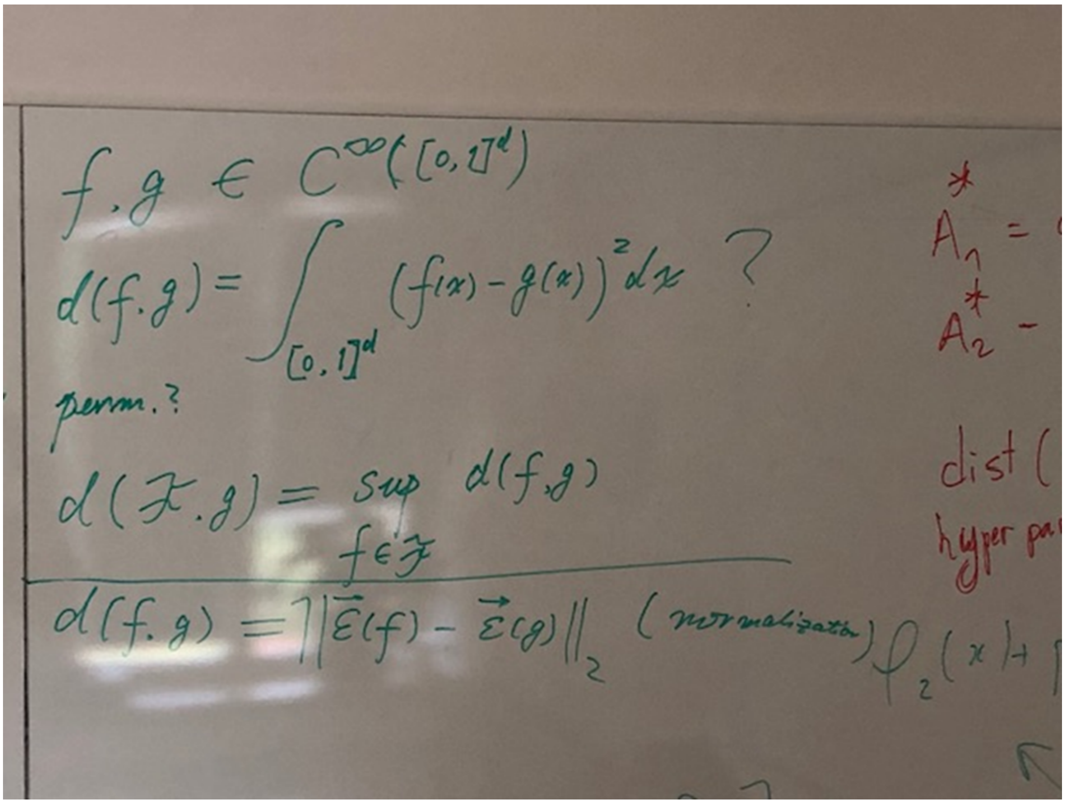

$f: X \longrightarrow \mathbb{R}$, $X = \mathbb{R}^d, \{0,1\}^d$
$\mathcal{F} = 50$ ins. from BBOB
$T = 10$ ins. of Kate
perf = AUC ECDF (normalized)
$X = [0,1]^d$
$\underbrace{\mathcal{F} = \{f_i\}_i}_{Train}$ $\underbrace{T = \{g_i\}_i}_{Test}$
$A$. $perf(A, f) \longrightarrow \mathbb{R}$
perfor. on $\mathcal{F}$
$\forall g \in T$, $R(A, \mathcal{F}, g) = \frac{|perf(A,g) - B|}{\boxed{d(\mathcal{F}, g)}}$
moon?
$\sup_{x \sim T} P_r(R(g) \geq \delta) \leq F(\ ?\ , \delta)$
$\sum_{k \geq 0} \beta_k \varphi_k(x)$
$\beta'_k = \frac{\beta_k}{\mathrm{Var}_{k>0}[\beta_k]}$
$\sum_{k \geq 0} \beta_k^2 = 1$
QUBO
Tree width param

Figure 5 Trying to make definition clearer.


f.g ∈ C^∞([0,1]^d)
d(f.g) = ∫_{[0,1]^d} (f(x) - g(x))^2 dx ?
perm.?
d(F.g) = sup_{f∈F} d(f,g)
d(f.g) = ||E(f) - E(g)||_2 (normalization) f_2(x)+
A_1^* =
A_2^* -
dist (
hyper pa

■ **Figure 6** Instance distance measures.

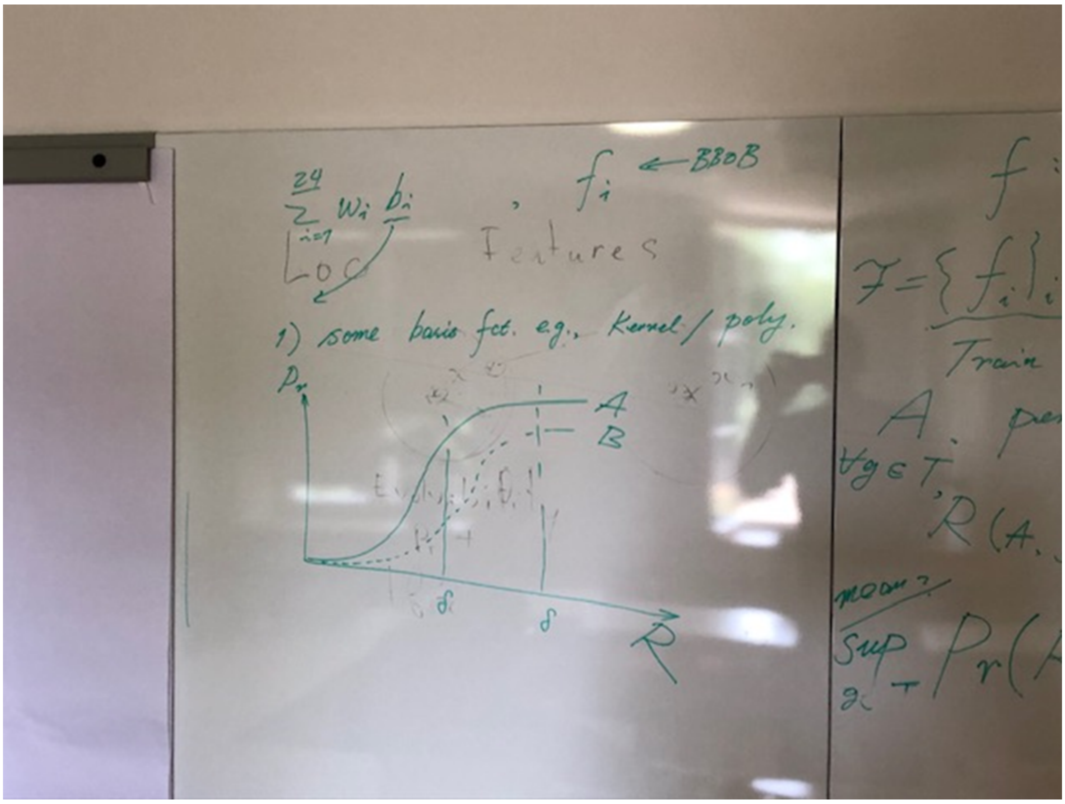


Figure 7 PAC-type formulation.

## Participants

- David L. Applegate
  Google – New York, US
- Anne Auger
  INRIA Saclay – Palaiseau, FR
- Thomas Bäck
  Leiden University, NL
- Carolin Benjamins
  Leibniz Universität
  Hannover, DE
- Peter A. N. Bosman
  CWI – Amsterdam, NL
- Carola Doerr
  Sorbonne University – Paris, FR
- Katharina Eggensperger
  Universität Tübingen, DE
- Tobias Glasmachers
  Ruhr-Universität Bochum, DE
- Nikolaus Hansen
  INRIA Saclay – Palaiseau, FR
- Emma Hart
  Edinburgh Napier University, GB
- Holger H. Hoos
  RWTH Aachen, DE
- Pascal Kerschke
  TU Dresden, DE
- Lars Kotthoff
  University of Wyoming –
  Laramie, US
- Manuel López-Ibáñez
  University of Manchester, GB
- Mariapia Marchi
  ESTECO SpA – Trieste, IT
- Olaf Mersmann
  TH Köln, DE
- Boris Naujoks
  TH Köln, DE
- Gabriela Ochoa
  University of Stirling, GB
- Gorjan Popovski
  Jozef Stefan Institute –
  Ljubljana, SI
- Mike Preuß
  Leiden University, NL
- Lennart Schäpermeier
  TU Dresden, DE
- Ofer M. Shir
  Tel-Hai College –
  Upper Galilee, IL
- Kate Smith-Miles
  The University of Melbourne, AU
- Olivier Teytaud
  Meta AI Research –
  Tournon-sur-Rhone, FR
- Dirk Thierens
  Utrecht University, NL
- Tea Tusar
  Jozef Stefan Institute –
  Ljubljana, SI
- Konstantinos Varelas
  Athens, GR
- Sebastien Verel
  Calais University, FR
- Diederick Vermetten
  Leiden University, NL
- Vanessa Volz
  modl.ai – Copenhagen, DK
- Hao Wang
  Leiden University, NL
- Darrell Whitley
  Colorado State University –
  Fort Collins, US
- Stefan M. Wild
  Lawrence Berkeley National
  Laboratory, US

Report from Dagstuhl Seminar 23252

# Inclusive Data Visualization

**Bongshin Lee*[1], Kim Marriott*[2], Danielle Szafir*[3], and Gerhard Weber*[4]**

1 **Microsoft Research – Redmond, US.** bongshin@microsoft.com
2 **Monash University – Caulfield, AU.** kim.marriott@monash.edu
3 **University of North Carolina at Chapel Hill, US.** dnszafir@cs.unc.edu
4 **Technische Universität Dresden, DE.** gerhard.weber@tu-dresden.de

## Abstract

Data plays an increasingly important role in our lives, and data visualization pervades our world as a means not only to analyze and explore data but also to identify and communicate insights. Most existing data visualizations, however, remain out of reach for many people with disabilities as they are designed on implicit assumptions about people's sensory, cognitive, and motor abilities. With an aim to tackle the significant equity issues posed by inaccessible data and data visualization, this Dagstuhl Seminar brought together researchers and practitioners from relevant fields, including visualization, accessibility, human-computer interaction, and health informatics. Five – both remote and in-person – invited talks gave participants an opportunity to understand barriers and challenges people with various disabilities currently face. With lightning talks and demos, participants shared their experiences and research relevant to inclusive data visualization. In addition, through brainstorming and discussion in break-out sessions combined with short report back presentations, participants identified research challenges and opportunities for inclusive data visualization.

**Seminar** June 18–23, 2023 – https://www.dagstuhl.de/23252
**2012 ACM Subject Classification** Human-centered computing → Visualization; Human-centered computing → Accessibility; Human-centered computing → Human computer interaction (HCI); Human-centered computing → Interaction design
**Keywords and phrases** Data Visualization, Information Visualization, Accessibility, Data Access, Human-Computer Interaction, Human-Data Interaction
**Digital Object Identifier** 10.4230/DagRep.13.6.81

## 1 Executive Summary

*Bongshin Lee (Microsoft Research – Redmond, US, bongshin@microsoft.com)*
*Kim Marriott (Monash University – Caulfield, AU, kim.marriott@monash.edu)*
*Danielle Szafir (University of North Carolina-Chapel Hill, US, dnszafir@cs.unc.edu)*
*Gerhard Weber (Technische Universität Dresden, DE, gerhard.weber@tu-dresden.de)*

We live in a data-driven world, where copious data are generated and captured by computing devices and sensors on and around us, and critical decisions are made based on data. Experts and lay individuals alike have access to a large amount of data, and understanding data and sharing insights have become a core part of information work. Data visualization is a powerful means not only to analyze and explore data but also to identify and communicate data-driven insights. Most existing data visualizations, however, are designed on implicit assumptions

* Editor / Organizer

Inclusive Data Visualization, *Dagstuhl Reports*, Vol. 13, Issue 6, pp. 81–105
Editors: Bongshin Lee, Kim Marriott, Danielle Szafir, and Gerhard Weber
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

about people's sensory, cognitive, and motor abilities. A lack of access to visualizations and their underlying data resulting from the differences in such abilities impacts people's education, work, lifestyle, and health, posing a significant equity issue. To address this important issue, visualization, accessibility, and other HCI researchers should work together to develop guidelines, methods, and techniques for increasing visualization accessibility.

One of the main goals of this Dagstuhl Seminar was to build partnerships and develop a shared understanding of this important research topic. This seminar helped us bring together researchers and practitioners from relevant fields, including data visualization, accessibility and assistive technologies, mobile and tangible interaction, human-computer interaction, health informatics, and vision science. It also provided opportunities to hear from representatives from disability support organizations and people with lived experience of disability. Furthermore, the unique setting of Schloss Dagstuhl helped us have an interactive dialog, facilitating the exchange of information and experiences, stimulating discussion and brainstorming, kickstarting collaborations, and identifying novel ideas around inclusive data visualization.

The main outcomes from the activities and discussions in this five-day seminar, which will be described below (The Week at a Glance), are:

- The establishment of a community around inclusive data visualization.
- The identification of open problems and challenges required to establish a rigorous foundation for inclusive data visualization.
- The development of an initial research agenda and plans for future activities in inclusive data visualization.
- The collaborations across different disciplines, including increased awareness of accessibility in the data visualization community and expanded awareness of data visualization in the accessibility community.

These outcomes will provide the impetus for a critical overarching goal: making data and visualization accessible to a broad range of people.

## The Week at a Glance

**Monday.** The seminar started with a brief introduction by the organizers about the main goals, topics, and structure of the seminar. This introduction was followed by the first invited talk, which was scheduled before participants' self-introduction to accommodate the timezone of Louisa Willoughby, a remote speaker from Monash University in Australia. The organizers planned the total of five invited talks on Monday to help seminar participants understand the main challenges people with different disabilities (e.g., cognitive, motor, and vision impairments) face. Afterwards, participants introduced themselves describing their main interests and expertise along with their aims for the seminar, with a short single-slide presentation (Figure 1 top left), and then had the morning coffee break. The morning session ended with two invited talks: one by Kirsten Ellis from Monash University and the other by Eun Kyoung Choe from University of Maryland at College Park.

After having lunch and a group of participants taking a walk outside (Figure 1 bottom left), the afternoon started with a brainstorming activity, where we – all participants – discussed seminar goals and outcomes we wanted to achieve together. This was followed by the fourth invited talk by JooYoung Seo from University of Illinois at Ulbana-Champaign. After having the afternoon coffee break, we continued the brainstorming activity for 30 minutes. Then, the last invited talk was given by Shea Tanis, a remote speaker from University of Kansas in

**Figure 1** Some of the activities our seminar participants engaged: Self-introduction and taking a walk on Monday (left); Excursion to Trier and a Winery on Wednesday (center); and Grand challenge discussion on Friday (right).

the United States. Finally, the first day ended with three invited demos (Figure 3): one by Ather Sharif, a remote presenter from University of Washington in the United States, the second by Arvind Satyanarayan from MIT, and the third by John Thompson from Microsoft Research at Redmond.

**Tuesday.** The organizers called for volunteers who would like to share interesting work, including research outcomes (artifacts, systems, study findings, etc.) and viewpoints, relevant to inclusive data visualization through short lightening demos and talks. The Tuesday morning session started with four demos and talks:

- *Haptification of Maps and Data* by Gerhard Weber and Meinhardt Branig
- *Audio-tactile Access to Floor Plans* by Karin Müller
- *Tactile Graphic Formats through Time and Their Varied Affordances for Inclusive Data Visualisation* by Leona Holloway
- *Designing Accessible Visualizations for People with Intellectual and Developmental Disabilities (IDD)* by Keke Wu

Before having the morning coffee break, we finalized the brainstorming activity on the seminar goals and outcomes, and then initiated the process of identifying topics for breakout groups. After the coffee break, we discussed and decided breakout groups along with the logistics and plans for working group activities. We initially identified eight topics, but considering participants' availability, schedule, and relation between topics, we later merged two of them into others, resulting in six topics. The rest of the day was devoted to breakout group discussions on the first three topics: user needs and abilities, authoring and tools, and technologies and information displays.

**Wednesday.** Wednesday morning was mainly devoted to discussion in breakout groups to enable participants to further delve into their discussion topics, which was followed by the report back from each of the three groups. Right before lunch, Jason Dykes shared thought-provoking viewpoints to consider visualization as Visual Data Design: with this lens, other materials, modes, and approach can be considered as various Data Designs, such as Textual, Haptic, Audio, and Edible Data Design.

Following the Dagstuhl Seminar tradition, Wednesday afternoon was set aside for social activities. After having a guided group tour in Trier (Figure 1 center top), we visited the "von Nell" Winery to have a winery tour (Figure 1 center bottom) followed by wine tasting and dinner. This social event facilitated personal conversations and fostered the discussion on research collaboration opportunities and deeper networking in a more relaxed setting.

**Thursday.** The Thursday morning session started with the following six lightning demos and talks:

- *Accessibility in the Context of India* by Anirudha Joshi
- *Action Audio* by Cagatay Goncu
- *Soundception: Multimodal Access to Depth in Images for Blind People* by Helen Petrie
- *3D Printing to Support Access to Graphical Content by People Who are Blind or Have Low Vision* by Matthew Butler
- *Physicalization Platforms as Possible Media for Accessible Data Representation* by Danielle Szafir
- *Making for All: Including People Living with Disabilities* by Kirsten Ellis

For the remainder of the day, we devoted the time for breakout group discussions, for the remaining three topics: data representations, education and literacy, and research methods. The day ended with the report back from each of the three groups.

**Friday.** Friday started with one lightning talk, *Machine-learning based Dysgraphia Detection in Children Handwritings* by Simone Marinai. Next, for the most of the morning, we worked as a group with an aim to identify the 10 grand challenges for inclusive data visualization. Then before lunch we briefly discussed next steps and how to build the community. This includes developing a website containing resources for inclusive data visualization and holding workshops and panels at relevant conferences. The day ended at lunch as participants left to make their way back home.

## 2 Table of Contents

**Working Groups**

# 3 Overview of Invited Talks and Demos

To provide an overview of the main challenges people with different disabilities (e.g., cognitive, motor, and vision impairments) face on the first day, the organizers planned five invited talks (Figure 2). In addition, to highlight some of the recent achievements in inclusive data visualization for blind or low-vision people, the organizers invited three demos of accessible visualization experiences for screen reader users (Figure 3). Three of the presenters could not attend the seminar in-person, and thus gave a talk via Zoom.

Two Perspective of Technology
Utopia – New Methods empower a broader range of people
Horrorscape – With each new technique more people are excluded
Stroke Survivors' Data Needs
Descriptive
Self-Reflective
Suggestive
Motivational
Laptop KL
Kim Marriott
Cognitive Disability Most Prevalent in United States
12.1%
12.8%
7.2%
6.1%
4.8%
3.6%

**Figure 2** Three of the five invited talks: Accessible Visualization for Physical Disabilities (top left); Stroke Care: A Rich Canvas for HCI Research (top right); and Cognitive Disability (bottom).

## 3.1 Introduction to Deafblind Communication

*Louisa Willoughby (Monash University – Clayton, AU, louisa.willoughby@monash.edu)*

In this talk, I will discuss the various kinds of deafblindness and the ways in which people who are deafblind communicate. I will also discuss two barriers deafblind people are currently facing. The first is the independent use of technology such as computers. The second is gaining information about the objects in their immediate environment and navigating through this space. Finally, I will discuss how tactile maps drawn on the body are used to provide spatial information.

## 3.2 Accessible Visualization for Physical Disabilities

*Kirsten Ellis (Monash University – Clayton, AU, kirsten.ellis@monash.edu)*

Accessible visualization for physical disability is an under-researched area that is often invisible in discussions of making data visualizations accessible. Physical disabilities include fine and gross motor movement and can be congenital, acquired, progressive or temporary so methods for accessibility may need to be adapted dynamically. There are three roles that people can take in the visualization process: passive viewer, active user and creator. In the role of passive viewer people with physical disabilities face minimal barriers to access but as soon as interactions are required to access or create content the barriers can be significant. The modalities used to design and use data visualization can significantly impact the ability of people with physical disabilities experience with visualization but research has not been conducted that establishes best practice for this group.

## 3.3 Stroke Care: A Rich Canvas for HCI Research

*Eun Kyoung Choe (University of Maryland – College Park, US, choe@umd.edu)*

Stroke, an injury to the brain from disrupted blood flow, often leads to lasting impairments such as speech difficulties (aphasia), one-sided weakness (hemiparesis), and cognitive issues. These changes can complicate daily tasks. A large body of work exists for stroke rehabilitation, designed to enhance stroke survivors' functional recovery. Of particular interest to the visualization community is mobile self-tracking interventions that show personal data to motivate people to engage in rehabilitation. Wearable sensors capture limb performance, providing metrics like use intensity, active arm use duration, and use ratio. Displaying such data in conjunction with personalized goals may encourage patients to utilize the affected limb in their everyday living. I propose multimodal feedback that transcends descriptive data, integrating self-reflective questions, suggestions, and motivational messages, presented via visual and audio narratives. This multimodal approach could enhance an understanding of the data and provide therapeutic support for stroke survivors.

## 3.4 Perceiving Beyond Vision: My Journey Through Dreamscapes, Numerical Cognition, and a Blind Critique of Inclusive Data Visualization

*JooYoung Seo (University of Illinois Urbana-Champaign, US, jseo1005@illinois.edu)*

In this personal narrative, I offer a unique perspective on life beyond the visual realm, delving into the nature of dreams, the intricacies of mental math, and the concept of "inclusive data visualization" through my lived experience as a blind individual. I invite you to explore the uncharted landscapes of my dreams, where the absence of sight gives rise to a distinct, immersive experience that challenges typical perceptions. I further share my unique approach to mental arithmetic, demonstrating the adaptive, versatile nature of human cognition in the absence of visual cues. From my vantage point, I critically assess the prevailing concept of "inclusive data visualization," questioning its true inclusivity for the visually impaired community. Join me in this journey, as we rethink the boundaries of perception and inclusivity in our predominantly visual world.

## 3.5 Cognitive Disability

*Shea Tanis (University of Kansas – Lawrence, US, tanis@ku.edu)*

In 2023, the United States Centers for Disease Control and Prevention, identified cognitive disability as the most prevalent disability across the nation, touching 12.8% of the population surpassing mobility disabilities (12.1%). As datafication of our world proliferates, understanding equity in knowledge translation and access to data visualizations becomes increasingly important. In 2018, the State of the States in Intellectual and Developmental Disabilities Ongoing Longitudinal Data Project of National Significance, partnered with the VisuaLab to validate anecdotal evidence and understand further cognitive accessibility of data visualizations. This research, the first of its kind, established guidelines for making visualizations more meaningful to users with cognitive disabilities. The presentation provided an overview of the partnership, community demands for equity, user-design approaches, and future research topics.

## 3.6 VoxLens: An Interactive JavaScript Library to Make Online Data Visualizations Accessible to Screen-Reader Users

*Ather Sharif (University of Washington – Seattle, US, asharif@cs.washington.edu)*

JavaScript visualization libraries are widely used to create online data visualizations but provide limited access to their information for screen-reader users. Building on prior findings about the experiences of screen-reader users with online data visualizations, in this demonstration, we present VoxLens, an open-source JavaScript plug-in that – with a single line of code – improves the accessibility of online data visualizations for screen-reader users using a multimodal approach. Specifically, VoxLens enables screen-reader users to obtain a holistic summary of presented information, play sonified versions of the data, and interact with visualizations in a "drill-down" manner using voice-based information querying.

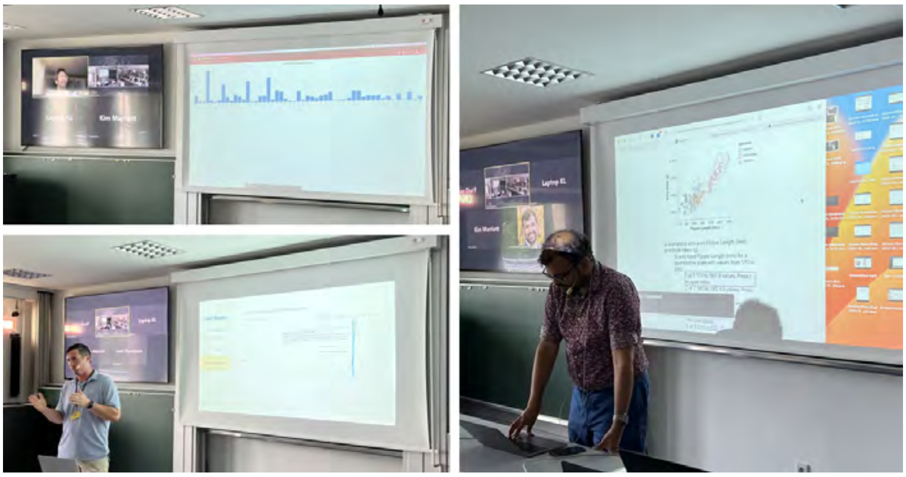


**Figure 3** Invited demos of accessible visualization experiences for screen reader users: VoxLens (top left); Olli (right); and Chart Reader (bottom left).

## 3.7 Olli: An Extensible Visualization Library for Screen Reader Accessibility

*Arvind Satyanarayan (MIT – Cambridge, US, arvindsatya@mit.edu)*

Olli, an open source library that converts visualizations into a keyboard-navigable structure accessible to screen readers. Using an extensible adapter design pattern, Olli is agnostic to the specific toolkit used to author the visualization. Olli renders a chart as an accessible tree view following the HTML Accessible Rich Internet Applications (ARIA) standard. The fields participating in the visualization serve as branches of the tree, and levels of the tree correspond to different granularities of data (e.g., major axis regions, minor axis regions, individual data values). Users can navigate up and down the tree using the up/down arrow keys, or move between sibling nodes using the left/right arrow keys. Users can also jump to specific positions in the tree via a series of drop down menus, or press the "T" key to invoke a data table view for more traditional row-by-row, column-by-column navigation.

## 3.8 Chart Reader: Accessible Visualization Experiences Designed with Screen Reader Users

*John Thompson (Microsoft Research – Redmond, US, johnthompson@microsoft.com)*

We demonstrate Chart Reader, an accessibility engine that renders web visualizations optimized for screen reader access. By designing and developing Chart Reader during a five-month iterative co-design study with 10 blind or low vision people, we aim to improve accessible visualization experiences. Our approach, realized through three sequentially designed and developed prototypes, allows users to interrogate visualizations using keyboard interactions, resulting in multimodal audio (announcements and sonification) of the chart. The web-based accessibility engine generates bar charts, stacked bar charts, and single-/multi-series line charts.

# 4 Overview of Lightning Demos and Talks

Encouraged and inspired by the invited talks and demos given on Monday, some of the seminar participants wanted an opportunity to share their work and research. The organizers thus called for volunteers who would like to share interesting work and ideas, including research outcomes (artifacts, systems, study findings, etc.) and viewpoints, relevant to inclusive data visualization through short lightening demos and talks.

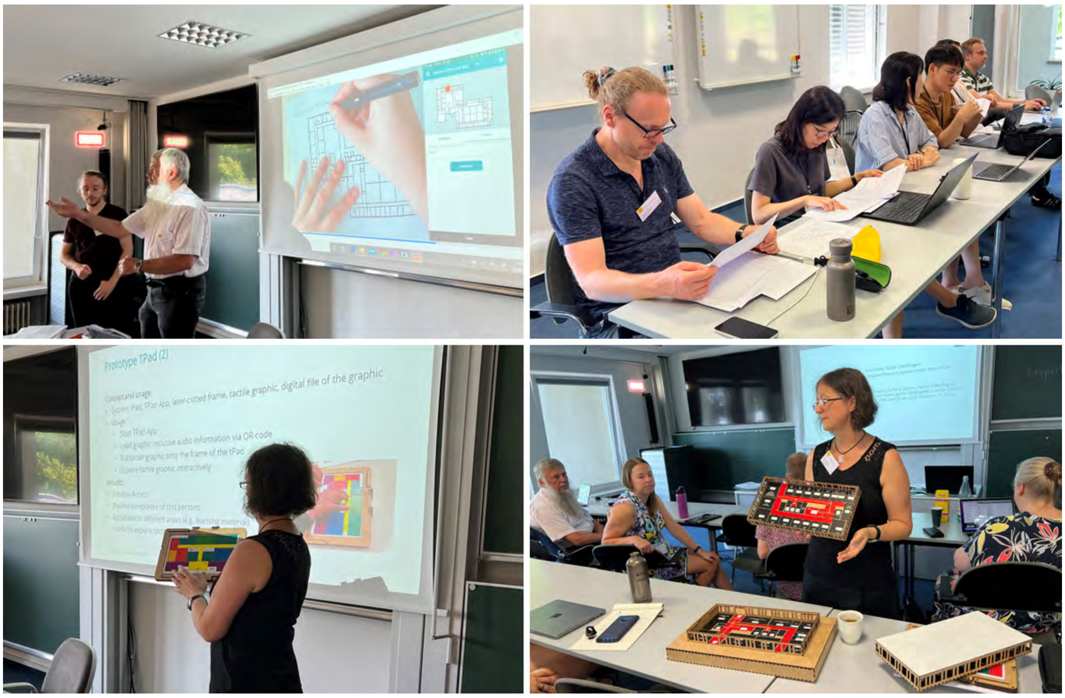


**Figure 4** A demo of haptification of maps and data (left top); seminar participants experience printed tactile graphics (right top); and a demo of Audio-tactile Access to Floor Plans (bottom).

## 4.1 Haptification of Maps and Data

*Gerhard Weber (TU Dresden, DE, gerhard.weber@tu-dresden.de)*
*Meinhardt Branig (TU Dresden, DE, meinhardt.branig@tu-dresden.de)*

Tactile media can be produced in various ways including embossers for braille and dynamic tactile displays. We demonstrate embossed tactile renderings of SVGPlot, a tool we developed and is itself accessible. In our evaluations bar graphs and scatter plots are identified as being suitable for exploration by blind people. Embossed tactile graphics can be turned into multimodal systems by a low cost pen with audio feedback (TipToi). We demonstrate a new version of an affordable Hyberbraille dynamic tactile display (portable, lowered height and reduced weight). As an extension for OpenStreetMap we show as a result of project AccessibleMaps maps of indoor buildings that are rendered both tactile aiming at blind people as well as for people with low vision through high contrast colors. The dynamic display is touch sensitive and also provides audio feedback for room names and barriers we identified in user surveys.

## 4.2 Audio-tactile Access to Floor Plans

*Karin Müller (Karlsruhe Institute of Technology, DE, karin.e.mueller@kit.edu)*

Visualization of information is a way to present data in a compact and clear way. People with blindness have access to this visual information only if it is provided by alternative forms of presentation such as audio-tactile. The TPad is a standard hardware, i.e., an iPad pro integrated in a frame made with a laser cutter. By using the frame a tactile graphic can be fixed. An app enables access to the additional digital information stored on the iPad via the audio-tactile system. A study with users with blindness showed that the system is useful to explore audio-tactile building plans and allows an intuitive access to this information.

## 4.3 Tactile Graphic Formats through Time and Their Varied Affordances for Inclusive Data Visualisation

*Leona Holloway (Monash University – Clayton, AU, leona.holloway@monash.edu)*

Tactile graphics, also known as raised line drawings, are best practice for the representation of 2D graphics to convey spatial relationships for people who are blind or have low vision. Tactile graphics can be created using a variety of methods, each with its own advantages and disadvantages. These include pressed paper, collage and other handcrafting, thermoform, swell or microcapsule paper, refreshable tactile displays and 3D printing on paper. Graphics must be simplified and redesigned for tactile reading, however little research has been conducted on the design of the many different data visualisations and their presentation using the various tactile graphic methods.

**Figure 5** Some of the lightening talks and demos.

## 4.4 Designing Accessible Visualizations for People with Intellectual and Developmental Disabilities (IDD)

*Keke Wu (University of North Carolina at Chapel Hill, US, kekewu@cs.unc.edu)*

Visualization amplifies cognition and helps a viewer see the trends, patterns, and outliers in data. However, conventional visualization tools and guidelines do not actively consider the unique needs and abilities of people with Intellectual and Developmental Disabilities (IDD), leaving them excluded from data-driven activities and vulnerable to ethical issues in everyday life. This work explores the challenges and opportunities of cognitively accessible visualization. Through mixed-method approaches and close collaboration with people with IDD, our group ran experiments and developed guidelines to improve current visualizations. We interviewed people with IDD and gained initial understandings of their daily data experiences, and we are currently in the process of running a participatory design workshop to create accessible visualizations for and with this population. We hope to further expand our knowledge of cognitively accessible visualization, translating what we have learned into a graphical user interface that supports people with IDD with better data analytics, and finally make this population more visible in the inclusive data visualization space.

## 4.5 Visual Data Design

*Jason Dykes (City, Univerity of London, UK, J.Dykes@city.ac.uk)*

I showed some knitted Selbu mittens from Norway. Noeska Smit used these to inspire her Data Knitualization, in which she uses wool (material) and knitting needles (technology) to encode information in physical form. She knits abstract representations – line graphs, and woollen models of real structures (body organs). I did so to differentiate between material, technology and data artefact, which we might consider to be a technology enabled manipulation of material that encodes data. This process is subjective, and so I like the notion of exposing this with the explicit use of the term "Design." It's a human process that involves intent. I forget who described "Design" as imagining a better World and doing something about it. But I like that perspective. I think it's what we try to do. So I wonder whether knitualization is PHYSICAL Data Design? Or perhaps we think about the material and consider this to be WOOL Data Design? So is visualisation VISUAL Data Design, and can we then think about alternative materials, modes, approaches and characterise these as: TEXTUAL, HAPTIC, EDIBLE, OLFACTORY, AUDIO, etc. Data Design? All of these can be interactive, collaborative, and combined in artefacts and analytical environments.

It feels to me as though this perspective has some advantages:

- it exposes the SUBJECTIVITY of the encoding and the artefact
- it reduces the cultural DOMINANCE of visualization among other forms of data depiction & representation
- it encourages MULTIMODAL designs and alternative encodings – why not do PHYSICAL WOOL Data Design or EDIBLE AUDIO Data Design?

These seem to be good things for reliable (consistent?) interpretation and accessible data. Many modes, many representations, many senses, many perspectives. But I look forward to hearing about disadvantages. Subsequently, I wonder whether Visual Data Design is actually a process in which we use light, and technologies that manipulate it to help us engage in Light Data Design to Data Design with Light. Maybe I have discovered that I am a Light Designer, and I use materials and technology that interact with light to depict data that represent aspects of the World. I hadn't thought of that before. Useful?

I also wonder whether Inclusive Data Visualization is achievable or even what we want to do. What we may really want to do is provide inclusive (diverse, comparable) Access to Data, and perhaps what this is really about is Access to Decision Making Processes and Influence, Empowerment.

## 4.6 Accessibility in the Context of India

*Anirudha Joshi (IIT Bombay, IN, anirudha@iitb.ac.in)*

I presented a study and showed two demos in the context of visually impaired Indian users. The first was a study that aimed to improve the accessibility of bar graphs. In this study we compared the speeds and accuracy of four techniques of auditory bar graphs of two lengths, namely Parallel-Tone, Parallel-Speech, Serial-Tone and Serial-Speech. The study included both sighted and visually impaired users. I also showed another demo of a 3D printable, modular tool that visually impaired people can use to both create and consume line charts. Lastly, I showed a demo of an accessible interaction technique that enables visually impaired users to enter numerical passwords without the need for using headphones. The technique was found to be shoulder-surfing-proof and can allow visually impaired users to confidently enter passwords in a public place. I skipped a demo on an accessible text input mechanism in Indian languages because that was a bit off-topic for this seminar, but it is available in the slides.

## 4.7 Action Audio

*Cagatay Goncu (Tennis Australia – Melbourne, AU, cagatay.goncu@gmail.com)*

Billions of people in the world watch sport media broadcasts to follow their favourite sport. They enjoy the actions captured by cameras and microphones on and around the fields. They socialise with other fans at home, at a local club, in a stadium and on social media. However, if you are blind or have low-vision (BLV), your overall experience is limited. While TV is providing the state of the art experience for sighted people, BLV need to tune in to radio broadcasts. Although used widely by BLV, radio can not provide all the actions in real time, in particular the movement of the objects such as the ball, puck, and players. Action Audio is a world-first system designed for the BLV to watch games in a broadcasting environment that is augmented with 3D sound. It provides alternative modalities to allow BLV access all the actions on sports broadcasting as well as live events in sport venues.

## 4.8 Soundception: Multimodal Access to Depth in Images for Blind People

*Helen Petrie (University of York, UK, helen.petrie@york.ac.uk)*

**Joint work of** Helen Petrie, Anna Bramwell-Dicks

We conducted an exploratory study with four blind people about the possibility of representing depth in images through variations in pitch and loudness of a tone. The blind participants were able to explore an image on the touchscreen of an iPad. Three sound conditions to represent depth were investigated: pitch variation; loudness variation; a fusion of pitch and loudness. All four blind participants strongly preferred the fusion option. As this was an initial exploratory study, more objective measures such as time and errors were not taken. However, these results are very encouraging and we will continue with this line of research.

## 4.9 3D Printing to Support Access to Graphical Content by People Who are Blind or Have Low Vision

*Matthew Butler (Monash University – Clayton, AU, matthew.butler@monash.edu)*

Access to visual information is compromised for people who are blind or have low vision. This impacts not only information access but general engagement with day-to-day activities that most take for granted. 3D printing has the potential to provide access to content that can be difficult with traditional tactile graphics. This demonstration provides an overview of work undertaken exploring how 3D printing can be used to convey traditionally visual content in the contexts of education, orientation and mobility, and cultural institutions. It hopes to inspire data visualisation designers to think about how this technology can be useful in creating accessible data visualisations.

## 4.10 Physicalization Platforms as Possible Media for Accessible Data Representation

*Danielle Szafir (University of North Carolina at Chapel Hill, US, dnszafir@cs.unc.edu)*

This talk presents two project, led by Sandra Bae in collaboration with Ellen Do, Michael Rivera, and Danielle Szafir, that offer potentially interesting physical platforms for accessible data representation. The first, the Data-Is-Yours toolkit, is a toolkit made from everyday materials (paper, mirrors, and cardboard) coupled with a cell phone to create basic interactive hybrid physical-digital visualizations. The goal of the toolkit is to leverage constructionist principles to inform data literacy in children through making; however, the hybrid physical-digital platform may also support accessible literacy programs as well as collaborative sensemaking through data. The second, sensing networks, provides a integrated pipeline for fabricating capacitive touch responsive node-link diagrams using 3D printing. The work enables people to readily construct physical models which can be immediately integrated into existing visualizations to provide a tangible input device grounded in data.

## 4.11 Making for All: Including People Living with Disabilities

*Kirsten Ellis (Monash University – Clayton, AU, kirsten.ellis@monash.edu)*

Makerspaces enable people to create digital items for themselves. They also provide the opportunity to build creative thinking and problem solving skills. Unfortunately people with disabilites are often excluded from participating because they are inaccessible. Research into how to make Makerspaces more inclusive is required for people with a range of different skills and abilities. An example of an inclusive circuit making activity is TapeBlocks which consists of chunky foam blocks wrapped in conductive tape with electronic components inserted under or on top. People with physical disabilities can push them together; blind people can feel the vibration and fan versions and people with intellectual disabilities can learn how to make them. TronicBoard are a flat version of TapeBlocks that enable a wider range of activities because they are easier to build into artifacts. There are lots of opportunities to make Making more inclusive but we need research to create and evaluate accessible tools.

## 4.12 Machine-learning Based Dysgraphia Detection in Children Handwritings

*Simone Marinai (University of Florence, IT, simone.marinai@unifi.it)*

The most common approach to identify dysgraphia in children is based on an interview with the pupil made by an expert in the field, who manually analyzes handwritten sentences. To this purpose, one widely adopted evaluation is based on the Brave Handwriting Kinder (BHK) test that takes into account features of handwriting produced by children that are manually annotated by an examiner. One important factor to consider is that during the test execution, the examiner can take into account the way the handwriting is produced (e.g., the posture of the pupils or how they handle the pen). However, the analysis of the handwritten specimen can be time-consuming and subjective, posing challenges in accurate and efficient diagnosis. In our recent research, we used smart-pens to perform the test and machine-learning based approaches to assess the dysgraphia level. The smart-pen allows us to capture features related to the speed of writing and pressure on the pen, in addition to the writing trajectory. Concerning the handwriting analysis, we implemented an algorithmic version of the BHK test and compared its performances with those achieved by an approach relying on deep-learning architectures. The children's handwritings have been also analyzed and scored, according to their potential level of dysgraphia by elementary school teachers.

# 5 Working Groups

One of the main activities during the seminar was to have in-depth discussions around the key topics through breakout groups. We first identified topics of interests through a group discussion and voting. We then had a series of break out group sessions for the six major topics we identified.

**Figure 6** Some sessions (one for each topic) from the series of breakout group sessions conducted to have in-depth discussions around six main topics.

## 5.1 Understanding the Needs and Challenges for People with Disabilities to Use and Create Data Visualizations

*John Thompson (Microsoft Research – Redmond, US, johnthompson@microsoft.com)*
*Eun Kyoung Choe (University of Maryland – College Park, US, choe@umd.edu)*
*Soyoung Choi (University of Illinois at Urbana-Champaign, US, soyoung@illinois.edu)*
*Kirsten Ellis (Monash University – Clayton, AU, kirsten.ellis@monash.edu)*
*Cagatay Goncu (Tennis Australia – Melbourne/Monash University – Clayton, AU, chatai.goncu@tennis.com.au)*
*Bongshin Lee (Microsoft Research – Redmong, US, bongshin@microsoft.com)*
*Helen Petrie (University of York, UK, helen.petrie@york.ac.uk)*
*JooYoung Seo (University of Illinois at Urbana-Champaign, US, jseo1005@illinois.edu)*
*Stephanie Wilson (City, University of London, UK, s.m.wilson@city.ac.uk)*
*Keke Wu (University of North Carolina at Chapel Hill, US, kekewu@cs.unc.edu)*

This research direction delves into the crucial intersection of user needs and abilities for crafting inclusive data visualizations. It addresses a comprehensive spectrum of disabilities, spanning visual, cognitive, physical, and auditory impairments, while also considering broader user preferences and individual lived experiences. Future research should recognize the inadvertent biases and exclusions in current practices that often arise from an assumption that individuals with disabilities are solely consumers, rather than potential producers of data and visualizations. Future research should address these problems by focusing on the following key dimensions: (1) Scrutinize unique and common requirements of diverse disability groups, aiming to uncover shared needs and barriers across the spectrum. Research is needed to identify tasks users want to accomplish and the challenges they encounter, thereby fostering an understanding of potential synergies among user groups. (2) Explore where and how individuals with disabilities interact with data, eliciting meaningful tasks while addressing systemic barriers. (3) Amplify user participation by advocating for the involvement of diverse user groups through co-design methods and community outreach. Research should also balance the broader needs of user groups with individual lived experiences by incorporating personalization.

## 5.2 Technologies for Inclusive Data Visualizations

*Kim Marriott (Monash University – Caulfield, AU, kim.marriott@monash.edu)*
*Jason Dykes (City, University of London, UK, J.Dykes@city.ac.uk)*
*Christian Frisson (Carleton University – Ottawa, CA, christianfrisson@cunet.carleton.ca)*
*Leona Holloway (Monash University – Clayton, AU, leona.holloway@monash.edu)*
*Karin Müller (Karlsruher Institut für Technologie, DE, karin.e.mueller@kit.edu)*
*Arvind Satyanarayan (MIT – Cambridge, US, arvindsatya@mit.edu)*
*Danielle Szafir (University of North Carolina at Chapel Hill, US, dnszafir@cs.unc.edu)*
*Tetsuya Watanabe (University of Niigata, JP, t2.nabe@eng.niigata-u.ac.jp)*
*Gerhard Weber (TU Dresden, DE, gerhard.weber@tu-dresden.de)*

The working group took some time to finalize the scope of this working group as the initial discussion focused on modalities and associated sensory channels/variables which overlapped with the working group on Representation. One interesting point was that variables may be multisensory, e.g. smoothness is a mix of haptic and visual [1]. We then decided to focus more on the underlying technologies.

We identified the following presentation technologies to support more inclusive data visualization: speech, sonification, magnification/image enhancement, dynamic tactile displays, tactile graphics, 3D models and data physicalization inc dynamic and shape-changing materials, force-feedback/vibrotactile/ultrasound haptic interfaces, tactiles and 3D models with audio-labels, making with everyday materials. The interaction technologies were: gesture/touch, speech, keyboard, mouse/joystick, multimodal combinations.

We discussed the requirements/characteristics of technology and perception (resolution both spatial and temporal, transparency, dynamicity, dimensionality (1-, 2-, 3-D) and the factors impacting technology use/choice (individual abilities, social acceptability, cost, availability, use context).

We recognized that lots had been done in sonification, physiology of haptic perception, color vision deficiencies but still much to do in these areas. The open questions/grand challenges list was huge ranging from how to provide sign language labels for Deaf users, choice of language for IDD users to exploring shape-changing technologies, use of making and dynamic tactile displays. A particular focus was multimodality. Concrete research questions that could/should be addressed now included:

- What is a “grammar of graphics” but for multimodal representations? (What’re the defaults, what is shared between modalities, what is modality specific)?
- How can multimodal representations help sighted and PWDs collaboratively analyze data.
- How do we understand the crossmodal perceptual trade-offs necessary to support effective sensemaking? How might these trade-offs map to different technologies?
- Do physical representations of (personal) data help people with IDD express themselves|think with data?
- How do we design interactions for dynamic tactile displays (zooming, filtering etc) that preserve mental model and make changes salient

### References

1 Yvonne Jansen et al., *Opportunities and Challenges for Data Physicalization*, In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI ’15), 2015. Association for Computing Machinery, New York, NY, USA, 3227–3236.

## 5.3 Accessible Authoring Tools and Production Methods for Inclusive Visualisation

*Matt Butler (Monash University – Clayton, AU, matthew.butler@monash.edu)*
*Catie Baker (Creighton University – Omaha, US, catherinebaker@creighton.edu)*
*Meinhardt Branig (TU Dresden, DE, meinhardt.branig@tu-dresden.de)*
*Leona Holloway (Monash University – Clayton, AU, leona.holloway@monash.edu)*
*Anirudha Joshi (IIT Bombay, IND, anirudha@iitb.ac.in)*
*Simone Marinai (University of Florence, IT, simone.marinai@unifi.it)*
*Karin Müller (Karlsruher Institut für Technologie, DE, karin.e.mueller@kit.edu)*
*Arvind Satyanarayan (MIT – Cambridge, US, arvindsatya@mit.edu)*
*Danielle Szafir (University of North Carolina at Chapel Hill, US, dnszafir@cs.unc.edu)*
*Benjamin Weyers (Trier University, DE, weyers@uni-trier.de)*

Traditional authoring tools and production methods for data visualisations often do not consider accessibility and inclusion. This may be with regard to the format of the visualisation, agency of the end user of the visualisation as part of the design process, and how accessible design tools are to people with different disabilities. As a result, people with disabilities have greatly reduced access to data and visualisations that are an increasing part of everyday life, as well as not developing key data literacy skills or being involved in data analysis activities.

This is an incredibly complex and challenging area. It encompasses process and tools, and requires expertise from the visualisation and accessibility communities. This paper seeks to capture the current state of data visualisation design and production, along with the tools used, and consider them in the context of inclusion and accessibility. The scope is articulated through an abstraction of the production process for accessible materials, along with the roles of key stakeholders in the design and production process. From here, grand challenges and key research questions are presented, including how people with different abilities can produce data visualizations for consumption by all, how to facilitate "born accessible" creation of inclusive data visualizations, and how emerging technologies can facilitate faster translation from existing graphics to inclusive data visualizations.

## 5.4 Beyond Visual Representations for Accessible Data Analysis and Communication

*Arvind Satyanarayan (MIT – Cambridge, US, arvindsatya@mit.edu)*
*Meinhardt Branig (TU Dresden, DE, meinhardt.branig@tu-dresden.de)*
*Eun Kyoung Choe (University of Maryland – College Park, US, choe@umd.edu)*
*Soyoung Choi (University of Illinois at Urbana-Champaign, US, soyoung@illinois.edu)*
*Jason Dykes (City, University of London, UK, J.Dykes@city.ac.uk)*
*Christian Frisson (Carleton University – Ottawa, CA, christianfrisson@cunet.carleton.ca)*
*Cagatay Goncu (Tennis Australia – Melbourne/Monash University – Clayton, AU, chatai.goncu@tennis.com.au)*
*Bongshin Lee (Microsoft Research – Redmond, US, bongshin@microsoft.com)*
*Kim Marriott (Monash University – Caulfield, AU, kim.marriott@monash.edu)*
*Helen Petrie (University of York, UK, helen.petrie@york.ac.uk)*
*JooYoung Seo (University of Illinois at Urbana-Champaign, US, jseo1005@illinois.edu)*
*Danielle Szafir (University of North Carolina at Chapel Hill, US, dnszafir@cs.unc.edu)*
*John Thompson (Microsoft Research – Redmond, US, johnthompson@microsoft.com)*

This research direction investigates how to formalize the study and development of alternate (non-visual) representations of data for analysis and communication. Participants began by drawing two analogies to visualization: top-down and bottom-up formalisms.

First, participants identified that "visual idioms" (also known as chart types) play a formative role in how everyday people conceptualize the visualization design space, and remain perhaps the most commonly used mechanism for creating visualizations (e.g., through tools such as Microsoft Excel, Google Sheets, etc.). Participants noted that these idioms shape the mental models people have about visualizations—including what sorts of hypotheses a given visualization is suitable for answering, and the types of interactive analysis one can perform on a visualization. Thus, participants wondered what an equivalent set of idioms would be for non-visual modalities. At a more fundamental level, participants pointed to visualization's grammatical formalisms that break idioms down into more atomic components: graphical shapes called "marks" whose properties, often called "visual variables" are determined by data. While participants identified that much prior work has identified candidate non-visual primitives (e.g., pitch, volume, timbre, etc. for sound or frequency, intensity, magnitude for haptics, etc.), it remains unclear how these primitives should be composed together. Moreover, while visualization grammars have yielded a large body of graphical perception studies (starting with Cleveland & McGill's seminal paper), there is a dearth of similar studies for non-visual modalities.

Participants brainstormed methods for answering these questions, identifying that co-design workshops were perhaps the most compelling approach. Such workshops would invite people with disabilities to create, design, and manipulate audio, haptic, visual, and tactile artifacts to accomplish a series of analysis and communicative goals. However, participants also noted that there were several seemingly foundational questions that are entangled with how such a workshop would be run. These questions include what should be the goal of non-visual data representations: should they replicate the affordances of visual representations, or should they focus on only specific tasks and be part of a multimodal ensemble? Similarly, what is the role of interaction in non-visual representation: should it maintain parallelism with its visual counterparts, and should that parallelism be maintained at the level of the "how" (i.e., the mechanics/operations) or the "what" (i.e., the insights that people gain as a result of performing the interaction).

## 5.5 Teaching and Learning How to Design and Make Sense of Inclusive Visualisations

*Catie Baker (Creighton University – Ottawa, US, catherinebaker@creighton.edu)*
*Kirsten Ellis (Monash University – Clayton, AU, Kirsten.Ellis@monash.edu)*
*Anirudha Joshi (IIT Bombay, IN, anirudha@iitb.ac.in)*
*Simone Marinai (University of Florence, IT, simone.marinai@unifi.it)*
*Karin Müller (KIT – Karlsruher Institut für Technologie, DE, karin.e.mueller@kit.edu)*
*Gerhard Weber (TU Dresden, DE, gerhard.weber@tu-dresden.de)*
*Benjamin Weyers (Trier University, DE, weyers@uni-trier.de)*
*Keke Wu (University of North Carolina at Chapel Hill, US, kekewu@cs.unc.edu )*

We are a group of visualization and accessibility researchers from different countries, disciplines and generations. We discussed "inclusive information visualisation" in the context of teaching and learning in visualization to identify best practices and materials for three main purposes: (i) teaching designers to create inclusive visualizations, (ii) teaching end users to make sense of inclusive visualizations and (iii) provide material and curriculum for "teaching teachers to teach". Such empowerment of learners was also the aim of a previous Dagstuhl Seminar on learning of and teaching about data visualisation [1].

Beyond people without a disability, inclusive information visualization addresses people to the widest extent possible. Several human abilities and needs have to be considered. In various contexts of learning such as school, university, or daily activities, learner's existing data literacy sets the starting point for achieving more advanced competences for analysis of data through multimodal (sequentially or in parallel) visualisations, both with respect as a designer and as a consumer of visualisations.

We discussed as an example the learned competences needed to design tactile graphics, verbalisations and auditory labels for bar charts as demonstrated by some of the participants earlier in the week, and how to become competent in designing with these technologies more abstract representations of visualisations such as box plots. We agreed, simulations of disability do not help to understand the needs of the targeted groups precisely, but for sensitising learners, it might help with the use of simulations to become competent in understanding consumer needs. Inspired by this process, we created scenarios that help visualization educators to assess success of learners. More such scenarios need to be developed and made available to educators showing good and bad approaches.

Consumers (blind people, people with low vision, deaf people, and neurodivergent people) can learn to interpret such scenario-based inclusive design of visualisations for the analysis of tabular data by specifying tasks already well known from earlier work on data visualisation, but which may require different ways to represent data e.g. by assistive technologies such as screen readers, sign language labels, sonification, or for instance by data videos, comics 3D visualizations and finding new ways to convey information.

We agreed, teaching inclusive data visualisation requires to solve grand challenges in each of the three main purposes and planned to develop a joint publication to identify them more clearly.

### References

1 Benjamin Bach et al., *Visualisation Empowerment: How to Teach and Learn Data Visualization*, Dagstuhl Reports, 12:6, 83–111, 2023.

## 5.6 Research Methods for Inclusive Data Visualization

*Leona Holloway (Monash University – Clayton, AU, leona.holloway@monash.edu)*
*Matt Butler (Monash University – Clayton, AU, matthew.butler@monash.edu)*
*Cagatay Goncu (Tennis Australia – Melbourne, AU, cagatay.goncu@gmail.com)*
*Tetsuya Watanabe (Niigata University, JP, t2.nabe@eng.niigata-u.ac.jp)*
*Stephanie Wilson (University of London, UK, S.M.Wilson@city.ac.uk)*

While both are rooted in the broad field of Human-Computer Interaction, the data visualization and accessibility research communities have differed in their use of research methods as a result of being driven by differences in aims and user populations. In bringing the two fields of data visualization and accessibility together to address the need for inclusive data visualizations that meet the needs of people with disabilities, there is now a need to examine these differences in methodologies and expectations to find a unified path forward. Here, we consider what we already know about accessibility and data visualization research, any gaps or differences, and the key priorities for a shared understanding. We also provide information and examples of existing best practice to assist researchers entering the new field of inclusive data visualization. Ultimately, the methodologies for inclusive data visualization research must aim to achieve rigor whilst also maintaining the principles of respect and inclusion.

# 6 Grand Challenges

On the last day, our main focus was to identify 10 grand challenges for inclusive data visualization. We broke into four groups, each of which was tasked to identify their top three challenges: after each participant individually came up with the top three challenges (Figure 7), each group synthesized the group's top three or four from the grand challenges their group members identified. Fourteen challenges were identified which could be grouped under six themes described below. We note that the boundaries of these themes are not necessarily clear. For challenges that could belong to multiple themes, we put them under the theme that has the strongest relation.

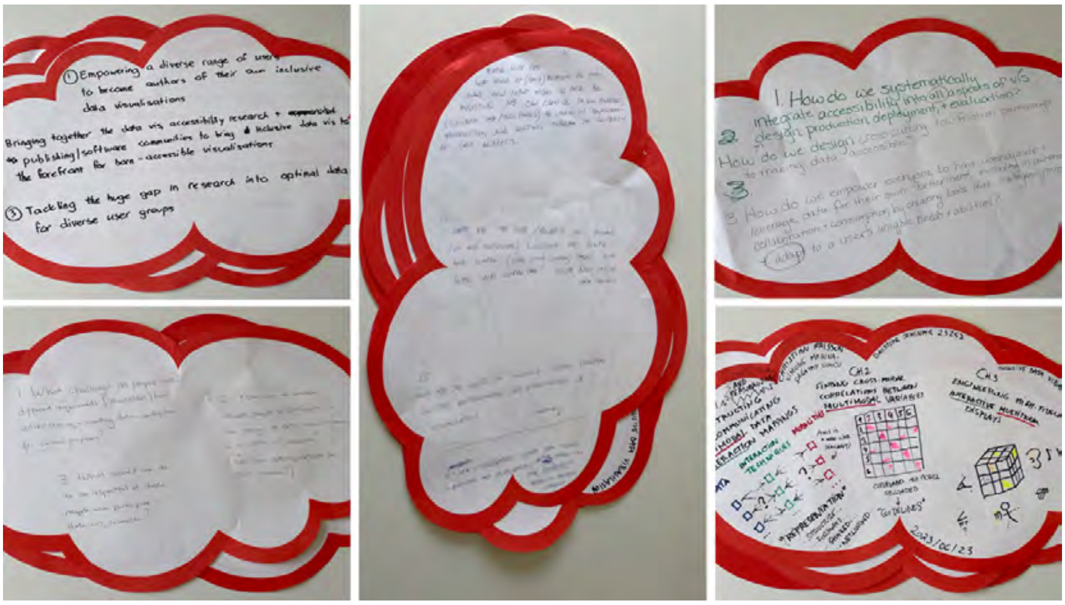

① Empowering a diverse range of users to become authors of their own inclusive data visualisations
Bringing together the data vis, accessibility research + ~~[illegible]~~ publishing/software communities to bring inclusive data vis to the forefront for born-accessible visualisations
③ Tackling the huge gap in research into optimal data for diverse user groups
1. How do we systematically integrate accessibility into all aspects of vis design, production, deployment, + evaluation?
2. How do we design [illegible] to making data accessible?
3. How do we empower everyone to [illegible] leverage data for their own betterment [illegible] adapt to a user's unique needs + abilities?
CH2
CH3
2023/06/23

Figure 7 Some of the top three challenges identified individually by participants.

## Needs

As a first step the inclusive data visualization community needs to better understand the real-world needs of people with disabilities. Currently, we lack an understanding of when people with disabilities would like to use data visualization and for what tasks as well as the current challenges.

- What are the actual data related needs, challenges, and difficulties that people with disabilities face to achieve tasks?
- Finding common needs and solutions that serve diverse abilities and draw upon diverse senses and datasets?
- What do diverse people value about data visualization and what challenges do they face when creating and using data visualizations?

## Accessible Design

We do not yet know how to design accessible data visualizations.

- What is the design space for accessible multimodal data visualizations?
- Investigating the optimal visualization techniques for different tasks, disabilities, and combinations, including the introduction of new disruptive techniques.

## Empowerment

Inclusive data visualization so far has been mainly focusing on providing access to other people's visualizations, but this is not enough. People with disabilities must be able to create their own data visualizations.

- How can end users of accessible representations be the primary designers and creators of accessible data vis, through supporting methods and tools?
- How can we empower people with a diverse range of abilities to become authors of their own inclusive data visualizations? This will include the need to improve data literacy in people with disabilities?

### Technology and Tools

We need better, cheaper accessible display technologies and tools that allow anybody to create accessible data visualizations that cater for individual requirements and preferences.

- Engineering high-fidelity interactive multimodal "displays"
- Tools that make it easier for everyone to create born accessible multimodal data visualizations that are accessible to people with a wide range of abilities
- Constructing and communicating and personalizing multimodal data interaction mappings ("Representation")

### Education

It is important to ensure that people of all abilities could learn the skills needed to create and understand accessible visualizations.

- How do we expand the education of creators and consumers with different abilities so that they have appropriate skills to create and use accessible visualizations?
- This [empowering people with diverse abilities to become authors] will include the need to improve data literacy in people with disabilities.

### Community Building

Accessible data visualization requires collaboration between the data visualization and accessibility research communities.

- How can we bring together the data visualization and accessibility communities to pursue sustainable action research?
- How can we align best practices and guidelines from accessibility and visualization communities?

## 7 Summary

Through this 5-day Dagstuhl Seminar, we increased awareness of the importance of inclusive data visualization research, facilitated the exchange of ideas and experiences, and discussed several important topics that for inclusive data visualization. Recognizing this is just a successful first step, we will continue to build a community around inclusive data visualization. All of our participants now joined to the Inclusive DataVis Slack workspace, `inclusivedatavis.slack.com`.

Another important outcome of the seminar is several possible next steps. We plan to create a website as a digital hub for inclusive data visualization. In addition to sharing relevant materials from the seminar, we aim to collect and propagate useful resources from the broader community, including actionable guidance for visualization designers, developers, and researchers. We also want to refine and share the outcome from group activities. For example, we desire to refine and publish the grand challenges to encourage and inspire the community to pursue. We hope to continue the effort and momentum through follow-up workshops or panels, as well as a special issue in a journal.

### Acknowledgements

The organizers would like to thank Dagstuhl for hosting our seminar and providing their great facilities, and three remote speakers for their time, effort, and great presentations.

## Participants

- Catie Baker
  Creighton University –
  Omaha, US
- Meinhardt Branig
  TU Dresden, DE
- Matthew Butler
  Monash University, AU
- Eun Kyoung Choe
  University of Maryland –
  College Park, US
- Soyoung Choi
  University of Illinois
  Urbana-Champaign, US
- Jason Dykes
  City, Univerity of London, UK
- Kirsten Ellis
  Monash University –
  Clayton, AU
- Christian Frisson
  Carleton University –
  Ottawa, CA
- Cagatay Goncu
  Tennis Australia –
  Melbourne, AU
- Leona Holloway
  Monash University –
  CLayton, AU
- Anirudha Joshi
  IIT Bombay, IN
- Bongshin Lee
  Microsoft Research –
  Redmond, US
- Simone Marinai
  University of Florence, IT
- Kim Marriott
  Monash University –
  Caulfield, AU
- Karin Müller
  Karlsruhe Institute of
  Technology, DE
- Helen Petrie
  University of York, UK
- Arvind Satyanarayan
  MIT – Cambridge, US
- Jooyoung Seo
  University of Illinois
  Urbana-Champaign, US
- Danielle Szafir
  University of North Carolina at
  Chapel Hill, US
- John Thompson
  Microsoft Research –
  Redmond, US
- Tetsuya Watanabe
  Niigata University, JP
- Gerhard Weber
  TU Dresden, DE
- Benjamin Weyers
  University of Trier, DE
- Stephanie Wilson
  City, Univerity of London, UK
- Keke Wu
  University of North Carolina at
  Chapel Hill, US

## Remote Participants

- Ather Sharif
  University of Washington –
  Seattle, US
- Shea Tanis
  University of Kansas –
  Lawrence, US
- Louisa Willoughby
  Monash University –
  Clayton, AU

Report from Dagstuhl Seminar 23261

# SAT Encodings and Beyond

**Marijn J. H. Heule*[1], Inês Lynce*[2], Stefan Szeider*[3], and Andre Schidler†[4]**

1 **Carnegie Mellon University - Pittsburgh, US.** marijn@cmu.edu
2 **University of Lisbon, PT.** ines.lynce@tecnico.ulisboa.pt
3 **TU Wien, AT.** stefan@szeider.net
4 **TU Wien, AT.** aschidler@ac.tuwien.ac.at

## Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 23261 "SAT Encodings and Beyond." The seminar facilitated an intense examination and discussion of current results and challenges related to encodings for SAT and related solving paradigms. The seminar featured presentations and group work that provided theoretical, practical, and industrial viewpoints. The goal was to foster more profound insights and advancements in encoding techniques, which are pivotal in enhancing solvers' efficiency.

**Seminar** June 25–30, 2023 – https://www.dagstuhl.de/23261
**2012 ACM Subject Classification** Theory of computation → Constraint and logic programming; Theory of computation → Discrete optimization
**Keywords and phrases** constraint propagation, lower and upper bounds, problem formulation, propositional satisfiability, symmetry breaking
**Digital Object Identifier** 10.4230/DagRep.13.6.106

## 1 Executive Summary

*Marijn J. H. Heule (Carnegie Mellon University - Pittsburgh, US)*
*Inês Lynce (University of Lisbon, PT)*
*Stefan Szeider (TU Wien, AT)*

The propositional satisfiability problem (SAT) is one of the most fundamental problems in computer science. As the first problem shown to be NP-complete by the Cook-Levin Theorem, SAT remains a fundamental benchmark problem for complexity theory. In contrast to its theoretical hardness, research over the last 20 years has successfully designed and engineered powerful algorithms for the SAT problem, called SAT solvers, that are surprisingly efficient on problem instances that arise from real-world applications. However, to solve a problem with a SAT solver or a related tool, one must first formulate the problem in terms of propositional logic to be digestible by the solver. This translation from the original problem to propositional logic is often called a SAT encoding. The encoding itself is often the crucial part that determines whether the solver can solve the problem efficiently, making the encoding techniques at least as important as the solving techniques. Hence, much effort has been put into researching efficient encoding techniques.

[*] Editor / Organizer
[†] Editorial Assistant / Collector

SAT Encodings and Beyond, *Dagstuhl Reports*, Vol. 13, Issue 6, pp. 106–122
Editors: Marijn J. H. Heule, Inês Lynce, Stefan Szeider, and Andre Schidler
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Other previous scientific meetings primarily focused on solving techniques, not on encodings. Hence, this Dagstuhl Seminar provided an overdue opportunity for an in-depth discussion of the state-of-the-art of encodings and future challenges and research avenues. When planning the seminar, we identified the following five critical topics.

**The Effectiveness of Encodings.** Current challenging research questions are new encodings for global constraints, theoretical lower and upper bounds on encoding size for global constraints, and new methods for symmetry breaking. Topics of interest are general principles of problem reformulations and their impact on the effectiveness of encodings.

**The Complexity of Encodings.** Although state-of-the-art SAT solvers can deal with millions of clauses and variables, the size of the original instance must be significantly smaller since the encoding often causes a polynomial (often cubic or worse) size blow-up. Which methods can overcome these limitations?

**Encoding Tools.** To fully exploit the power of SAT solvers, researchers have designed high-level languages that are amenable to describing constraints and developed compilers for converting constraints into CNF. Exciting topics for discussions include the questions of how to obtain an optimal hybridization of encodings and how to decompose global constraints.

**Lazy Encodings.** An interesting approach to SAT encodings is to start with an incomplete under-constrained encoding and add clauses to it once a solution has been found that violates properties that are not considered by the encoding. SAT modulo Theories and Lazy Clause Generation are among the approaches utilizing eager encodings.

**Verifying Encodings.** Trust in the correctness of SAT-solving results increased significantly in the last couple of years as all top-tier solvers can produce proofs of unsatisfiability that can be validated using efficient and formally verified tools. An interesting topic is how the encoding part of the toolchain can be sufficiently validated.

**Beyond SAT.** The success of SAT solving has spawned the development of efficient solvers for problems that are more general than SAT, including MaxSAT, QBF-SAT, PB, ASP, and CP. These more general problems require new encoding techniques.

We invited key researchers to cover these topics and were happy that most of the people we wished for accepted the invitation. Hence, we could approach participants individually to solicit longer survey talks to cover these topics by top experts. Shorter, focused talks complemented these longer survey-like talks. The talks covered various encoding aspects for particular solving paradigms, including SAT, CP, ASP, MaxSAT, and QBF.

We were delighted to have the *industrial perspective* covered by Andreas Falkner (Siemens AG), who presented challenges in industrial product configuration.

Other talks were devoted to symmetry-breaking techniques that boost SAT-based combinatorial search, which included a live demo of the SMS tool; another focus of several talks was the verification of results obtained via encodings. Some talks explored the theoretical limits of encodings and the connection between computer algebra systems and SAT encodings.

In addition to the talks, we had an *open-problems and challenges* session and dedicated time for group work. The posed problems asked for desirable properties for proof logging, how encodings can ensure that propagation on a high level implies propagation on a low level, how encodings for enumeration and counting can be established, how one can measure the usefulness of auxiliary variables in encodings, how to verify that an encoding is correct, and the exact computational complexity of minimal resolution proof length (in binary). Also, efficient encodings for several concrete problems were posed, including Golumb Rulers, the Connect-4 game, the metric dimension of hypercubes, the independent configuration problem, problems related to Steiner Triples, line arrangements with a limited number of triangles, and block designs that appear in product configuration. We formed working groups to tackle some of these problems and had a brief session where progress on these problems was reported and discussed.

Overall, we are pleased with the outcome of the seminar. We have met our objectives and started a highly stimulating discussion and exchange of ideas, covering the state of the art and future challenges. Still, it also became clear that encodings are a far-reaching topic that leaves many challenging open questions for future work. So, a follow-up Dagstuhl Seminar in the future is strongly indicated.

## 2 Table of Contents

**Executive Summary**

**Overview of Talks**

# 3 Overview of Talks

## 3.1 SAT and Computer Algebra

*Curtis Bright (University of Windsor, CA)*

**Main reference** Curtis Bright, Ilias S. Kotsireas, Vijay Ganesh: "When satisfiability solving meets symbolic computation", Commun. ACM, Vol. 65(7), pp. 64–72, 2022.
**URL** https://doi.org//10.1145/3500921

Combining satisfiability (SAT) solvers with computer algebra systems (CASs) progress has enabled progress on problems requiring search and sophisticated mathematics [1]. In this talk, I will outline problems I have worked on in which the SAT+CAS method outperformed pure SAT or pure CAS approaches by orders of magnitude. For example, the SAT+CAS method found the first Williamson matrices of order 70 [2], certified the nonexistence of finite projective planes of order 10 [3], demonstrated a Kochen–Specker vector system in three dimensions must have size at least 24 [4, 5], and has improved certain side-channel attacks on integer factorization [6].

### References

1 C. Bright, I. Kotsireas, V. Ganesh. *When Satisfiability Checking Meets Symbolic Computation*. Communications of the ACM, 2022.
2 C. Bright, I. Kotsireas, V. Ganesh. *Applying Computer Algebra Systems with SAT Solvers to the Williamson Conjecture*. Journal of Symbolic Computation, 2020.
3 C. Bright, K. Cheung, B. Stevens, I. Kotsireas, V. Ganesh. *A SAT-based Resolution of Lam's Problem*. AAAI 2021.
4 Z. Li, C. Bright, V. Ganesh. *An SC-Square Approach to the Minimum Kochen–Specker Problem*, SC-Square Workshop, 2022.
5 Z. Li, C. Bright, V. Ganesh. *A SAT Solver + Computer Algebra Attack on the Minimum Kochen–Specker Problem*, arXiv:2306.13319, 2023.
6 Y. Ajani, C. Bright. *A Hybrid SAT and Lattice Reduction Approach for Integer Factorization*, SC-Square Workshop, 2023.

## 3.2 Verified encodings for SAT solvers

*Cayden Codel (Carnegie Mellon University - Pittsburgh, US) and Marijn J. H. Heule (Carnegie Mellon University - Pittsburgh, US)*

**Joint work of** Cayden R. Codel, Jeremy Avigad, Marijn J. H. Heule
**Main reference** Cayden R. Codel, Jeremy Avigad, Marijn J. H. Heule: "Verified Encodings for SAT Solvers," to appear in FMCAD 2023.
**URL** http://crcodel.com/research/fmcad2023.pdf

SAT is a powerful tool for solving a wide array of problems, but many problems are not expressed in propositional logic and must instead be encoded into SAT. These encodings are often subtle, and implementations are error-prone. Formal correctness proofs are needed to ensure that implementations are bug-free.

In this talk, we present a library for formally verifying SAT encodings, written using the Lean interactive theorem prover. Our library currently contains verified encodings for the parity, at-most-one, and at-most-k constraints. It also contains methods of generating fresh variable names and combining sub-encodings to form more complex ones, such as one for encoding a valid Sudoku board. The proofs in our library are general, and so this library serves as a basis for future encoding efforts.

## 3.3 Breaking Symmetries when Solving Hard Combinatorial Problems

*Michael Codish (Ben Gurion University - Beer Sheva, IL)*

Many hard combinatorial problems involve huge numbers of symmetries which derive from isomorphic representations of objects in the search space. Restricting search to avoid symmetries – aka "symmetry breaking" – makes a big difference when trying to solve such problems.

Symmetry breaking in constraint programming is often achieved by introducing symmetry breaking constraints which are satisfied by at least one member of each isomorphism class. Complete symmetry breaking constraints are satisfied by exactly one member from each class and other symmetry breaks are called partial.

In this talk I will focus mainly on breaking symmetries in graph search problems. The search for complete symmetry breaking constraints for graph search problems is itself a hard problem and it is unknown if there exists a complete symmetry breaking constraint that is polynomial in the size of the graph.

In computer science when the problem is hard – we typically follow one or more from three alternatives: (1) clever brute force computation, (2) approximation algorithms, or (3) identifying special cases where the problem is easier.

This talk will focus on how each of these three alternatives comes into play when solving hard graph search problems.

## 3.4 FPT-reductions to SAT – And SAT encodings for problems in FPT

*Ronald de Haan (University of Amsterdam, NL)*

In this talk, we will discuss some results and some research directions that connect the theory of parameterized complexity and the theory of encodings and solvers for SAT and related problems such as ASP.

The talk can be divided into roughly two parts. The first part addresses the (im)possibility of fixed-parameter tractable encodings into SAT. This makes sense for problems whose (classical) complexity is beyond NP, and so one does not hope for polynomial-time encodings. For suitable choices of parameters, the problem could be encoded in fpt-time to SAT. We will give a few examples of cases where this is possible, and we present a parameterized complexity toolbox that can be used to assess the (im)possibility of fpt-time encodings into SAT.

The second part addresses an ongoing research direction, that revolves around encoding fpt-time solvable problems into SAT in such a way that CDCL solvers are guaranteed to run in fixed-parameter tractable time. For some problems, one can do this in such a way that this works with any branching heuristic, and for some problems the choice of branching heuristic makes a difference. We will present some examples illustrating this (for SAT and ASP), and then we raise some open research questions in this arena.

## 3.5 Challenges in industrial product configuration

*Andreas Falkner (Siemens AG - Wien, AT)*

**Main reference** Andreas A. Falkner, Alois Haselböck, Gerfried Krames, Gottfried Schenner, Herwig Schreiner, Richard Taupe: "Solver Requirements for Interactive Configuration", J. Univers. Comput. Sci., Vol. 26(3), pp. 343–373, 2020.
**URL** http://www.jucs.org/jucs_26_3/solver_requirements_for_interactive
**Main reference** Andreas A. Falkner, Gerhard Friedrich, Konstantin Schekotihin, Richard Taupe, Erich Christian Teppan: "Industrial Applications of Answer Set Programming", Künstliche Intell., Vol. 32(2-3), pp. 165–176, 2018.
**URL** https://doi.org//10.1007/S13218-018-0548-6
**Main reference** Andreas A. Falkner, Gerhard Friedrich, Alois Haselböck, Gottfried Schenner, Herwig Schreiner: "Twenty-Five Years of Successful Application of Constraint Technologies at Siemens", AI Mag., Vol. 37(4), pp. 67–80, 2016.
**URL** https://doi.org//10.1609/AIMAG.V37I4.2688
**Main reference** Andreas A. Falkner, Anna Ryabokon, Gottfried Schenner, Kostyantyn M. Shchekotykhin: "OOASP: Connecting Object-Oriented and Logic Programming", in Proc. of the Logic Programming and Nonmonotonic Reasoning - 13th International Conference, LPNMR 2015, Lexington, KY, USA, September 27-30, 2015. Proceedings, Lecture Notes in Computer Science, Vol. 9345, pp. 332–345, Springer, 2015.
**URL** https://doi.org//10.1007/978-3-319-23264-5_28
**Main reference** Deepak Dhungana, Andreas A. Falkner, Alois Haselböck: "Generation of conjoint domain models for system-of-systems", in Proc. of the Generative Programming: Concepts and Experiences, GPCE'13, Indianapolis, IN, USA - October 27 - 28, 2013, pp. 159–168, ACM, 2013.
**URL** https://doi.org//10.1145/2517208.2517224
**Main reference** Andreas A. Falkner, Alois Haselböck, Gottfried Schenner, Herwig Schreiner: "Modeling and solving technical product configuration problems", Artif. Intell. Eng. Des. Anal. Manuf., Vol. 25(2), pp. 115–129, 2011.
**URL** https://doi.org//10.1017/S0890060410000570
**Main reference** Deepak Dhungana, Andreas A. Falkner, Alois Haselböck: "Configuration of Cardinality-Based Feature Models Using Generative Constraint Satisfaction", in Proc. of the 37th EUROMICRO Conference on Software Engineering and Advanced Applications, SEAA 2011, Oulu, Finland, August 30 - September 2, 2011, pp. 100–103, IEEE Computer Society, 2011.
**URL** https://doi.org//10.1109/SEAA.2011.24

Product configuration has been among the first successful applications of symbolic AI, e.g. translating feature models with cross-tree constraints to SAT encodings and finding consistent solutions. Despite the high maturity of state-of-the-art tools, encoding remains challenging in practice: dynamic size (i.e. unbounded multiplicities of variables and constraints), OO-like inheritance (for clearer knowledge representation), open domains, merging of subsystem encodings (from distributed authors), multi-dimensional optimization, reconfiguration and knowledge evolution, explanations and recommendations, debugging, etc.

## 3.6 Reasoning-Enabling Encodings

*Marijn J. H. Heule (Carnegie Mellon University - Pittsburgh, US)*

A common approach in automated reasoning is to encode a given problem into propositional logic and then solve the resulting formula with a satisfiability (SAT) solver. As the quality of the encoding has a big impact on solver performance, it is no coincidence that solvers are highly successful in the field of hardware verification: digital electronic circuits have a direct encoding into propositional logic which is often adequate for solving. However, the same is not true for many other applications. Sophisticated encodings may be required to efficiently solve some problems using SAT solvers. This talk will focus on sophisticated encodings of some hard-combinatorial problems for which a straightforward encoding is ineffective. We will first describe general techniques to produce high-quality encodings. Afterward, we will present encodings for specific problems: edge-matching puzzles, Hamiltonian cycles, and packing colorings.

## 3.7 SAT-Based Judgment Aggregation

*Matti Järvisalo (University of Helsinki, FI)*

**Joint work of** Ari Conati, Andreas Niskanen, Matti Järvisalo
**Main reference** Ari Conati, Andreas Niskanen, Matti Järvisalo: “SAT-Based Judgment Aggregation”, in Proc. of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS ’23, p. 1412–1420, International Foundation for Autonomous Agents and Multiagent Systems, 2023.
**URL** https://dl.acm.org/doi/abs/10.5555/3545946.3598792

Judgment aggregation (JA) offers a generic formal logical framework for modeling various settings where agents must reach joint agreements through aggregating the preferences, judgments, or beliefs of individual agents by social choice mechanisms. In this work, we develop practical JA algorithms for outcome determination by harnessing Boolean satisfiability (SAT) based solvers as the underlying reasoning engines, leveraging on their ability to efficiently reason over logical representations incrementally. Concretely, we provide algorithms for outcome determination under a range of aggregation rules, using natural choices of SAT-based techniques adhering to the computational complexity of the problem for the individual rules. We also implement and empirically evaluate the approach using both synthetic and PrefLib data, showing that the approach can scale significantly beyond recently proposed alternative algorithms for JA.

## 3.8 SAT encodings from a Contraint Programming perspective: Why and Why not

*George Katsirelos (INRAE - Palaiseau, FR)*

Encoding constraints to CNF is an attractive option for CP solvers, especially in the context of clause learning. However, it is not always a preferable or even feasible option, depending on our requirements. In this talk, I will go over some cases where SAT encodings have been used successfully in CP solvers, as well as some cases where it does not work out as well. I will point out some theoretical work for why this is the case, as well as some practical reasons for it.

## 3.9 Combining SAT and Computer Algebra for Circuit Verification

*Daniela Kaufmann (TU Wien, AT)*

Verifying multiplier circuits is an important problem which in practice still requires substantial manual effort. In this talk, I will demonstrate that encoding the entire problem into SAT is not the ideal strategy, nor is using a pure algebraic encoding. We use a combination of SAT and computer algebra in our method to significantly improve automated verification of integer multipliers.

## 3.10 Isomorph-Free Generation of Combinatorial Objects With SAT Modulo Symmetries

*Markus Kirchweger (TU Wien, AT) and Stefan Szeider (TU Wien, AT)*

**Main reference** Markus Kirchweger, Stefan Szeider: "SAT Modulo Symmetries for Graph Generation", in Proc. of the 27th International Conference on Principles and Practice of Constraint Programming, CP 2021, Montpellier, France (Virtual Conference), October 25-29, 2021, LIPIcs, Vol. 210, pp. 34:1–34:16, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.
**URL** https://doi.org//10.4230/LIPICS.CP.2021.34

SAT modulo Symmetries (SMS) is a framework for the exhaustive isomorph-free generation of combinatorial objects with a prescribed property. SMS relies on the tight integration of a CDCL SAT solver with a custom dynamic symmetry-breaking algorithm that iteratively refines an ordered partition of the generated object's elements. SMS supports DRAT proofs for the SAT solver's reasoning and offline verification of the symmetry breaking clauses, and thus provides an additional layer of confidence in the obtained results. This talk will discuss the basic concepts of SMS and review some of its applications on graphs, digraphs, hypergraphs, and matroids. At the end of the talk, we will give a live demo of the tool.

## 3.11 Automatic Tabulation in Constraint Models

*Zeynep Kiziltan (University of Bologna, IT)*

**Joint work of** Özgür Akgün, Ian P. Gent, Christopher Jefferson, Zeynep Kiziltan, Ian Miguel, Peter Nightingale, András Z. Salamon, Felix Ulrich-Oltean
**Main reference** Özgür Akgün, Ian P. Gent, Christopher Jefferson, Zeynep Kiziltan, Ian Miguel, Peter Nightingale, András Z. Salamon, Felix Ulrich-Oltean: "Automatic Tabulation in Constraint Models", CoRR, Vol. abs/2202.13250, 2022.
**URL** https://arxiv.org/abs/2202.13250

The performance of a constraint model can often be improved by converting a sub-problem into a single table constraint, which is referred to as tabulation. In this talk, I will describe an automatic tabulation approach, implemented in Savile Row which is a constraint model reformulation tool. Savile Row takes as input a model described in the high-level solver-independent modelling language Essence Prime and has backends for CP, SAT and ILP solvers. Our approach to automatic tabulation deploys heuristics to discover opportunities for tabulation and uses a specific propagator or an encoding for the generated table constraint depending on the chosen backend solver.

## 3.12 An iterative university course timetabling tool with MaxSAT

*Inês Lynce (University of Lisbon, PT)*

**Joint work of** Inês Lynce, Alexandre Lemos, Pedro T. Monteiro
**Main reference** Alexandre Lemos, Pedro T. Monteiro, Inês Lynce: "Introducing UniCorT: an iterative university course timetabling tool with MaxSAT", J. Sched., Vol. 25(4), pp. 371–390, 2022.
**URL** https://doi.org//10.1007/S10951-021-00695-6

This work describes the UniCorT tool designed to solve university course timetabling problems specifically tailored for the 2019 International Timetabling Competition (ITC). The proposed approach includes pre-processing, the use of a maximum satisfiability (MaxSAT) solver and a local search procedure. The impact of a handful of techniques in the quality of the solution and the execution time is evaluated. We take into account different pre-processing techniques and CNF encodings, as well as the combination with a local search procedure. The success of our tool is attested by having been ranked among the five finalists of the ITC 2019 competition.

## 3.13 Some connections between encodings and circuits

*Stefan Mengel (CNRS, CRIL - Lens, FR)*

In the literature, there are several results making SAT-encodings and Boolean circuits. In particular, it is known that different classes of circuits correspond tightly to encodings with specific properties, e.g. restricted (tree/clique-)width or propagation strength. In this talk, I will survey some of these connections and point out some open questions.

## 3.14 Certified CNF Translations for Pseudo-Boolean Solving

*Andy Oertel (Lund University, SE)*

**Joint work of** Stephan Gocht, Ruben Martins, Jakob Nordström, Andy Oertel
**Main reference** Stephan Gocht, Ruben Martins, Jakob Nordström, Andy Oertel: "Certified CNF Translations for Pseudo-Boolean Solving", in Proc. of the 25th International Conference on Theory and Applications of Satisfiability Testing, SAT 2022, August 2-5, 2022, Haifa, Israel, LIPIcs, Vol. 236, pp. 16:1–16:25, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.
**URL** https://doi.org//10.4230/LIPICS.SAT.2022.16

The dramatic improvements in Boolean satisfiability (SAT) solving since the turn of the millennium have made it possible to leverage state-of-the-art conflict-driven clause learning (CDCL) solvers for many combinatorial problems in academia and industry, and the use of proof logging has played a crucial role in increasing the confidence that the results these solvers produce are correct. However, the fact that SAT proof logging is performed in conjunctive normal form (CNF) clausal format means that it has not been possible to extend guarantees of correctness to the use of SAT solvers for more expressive combinatorial paradigms, where the first step is an unverified translation of the input to CNF. In this work, we show how cutting-planes-based reasoning can provide proof logging for solvers that translate pseudo-Boolean (a.k.a. 0-1 integer linear) decision problems to CNF and then run CDCL. To support a wide range of encodings, we provide a uniform and easily extensible framework for proof logging of CNF translations. We are hopeful that this is just a first step towards providing a unified proof logging approach that will also extend to maximum satisfiability (MaxSAT) solving and pseudo-Boolean optimization in general. This is joint work with Stephan Gocht, Ruben Martins and Jakob Nordström published at SAT'22.

## 3.15 Co-Certificate Learning with SAT Modulo Symmetries

*Tomáš Peitl (TU Wien, AT), Markus Kirchweger (TU Wien, AT), and Stefan Szeider (TU Wien, AT)*

**Main reference** Markus Kirchweger, Tomás Peitl, Stefan Szeider: "Co-Certificate Learning with SAT Modulo Symmetries", in Proc. of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China, pp. 1944–1953, ijcai.org, 2023.
**URL** https://doi.org//10.24963/IJCAI.2023/216

We present a new SAT-based method for generating all graphs up to isomorphism that satisfy a given co-NP property. Our method extends the SAT Modulo Symmetry (SMS) framework with a technique that we call co-certificate learning. If SMS generates a candidate graph that violates the given co-NP property, we obtain a certificate for this violation, i.e., a "co-certificate" for the co-NP property. The co-certificate gives rise to a clause that the SAT solver serving as SMS's backend learns as part of its CDCL procedure. We demonstrate that SMS plus co-certificate learning is a powerful method that allows us to improve the best-known lower bound on the size of Kochen-Specker vector systems, a problem that is central to the foundations of quantum mechanics and has been studied for over half a century. Our approach is orders of magnitude faster and scales significantly better than a recently proposed SAT-based method.

## 3.16 Exact resolution complexity

*Tomáš Peitl (TU Wien, AT) and Stefan Szeider (TU Wien, AT)*

**Main reference** Tomás Peitl, Stefan Szeider: "Finding the Hardest Formulas for Resolution", J. Artif. Intell. Res., Vol. 72, pp. 69–97, 2021.
**URL** https://doi.org//10.1613/JAIR.1.12589

This talk is based on two papers about computing shortest resolution proofs of formulas in CNF, in which we investigate encodings to compute shortest proofs of minimally unsatisfiable formulas, and of hitting formulas in particular, we compute the hardest formulas (and the hardest hitting formulas) with a small number of clauses, and discuss related questions. The abstracts of the two papers follow.

1. A CNF formula is harder than another CNF formula with the same number of clauses if it requires a longer resolution proof. In this paper, we introduce resolution hardness numbers; they give for m=1,2,... the length of a shortest proof of a hardest formula on m clauses. We compute the first 10 resolution hardness numbers, along with the

corresponding hardest formulas. To achieve this, we devise a candidate filtering and symmetry breaking search scheme for limiting the number of potential candidates for hardest formulas, and an efficient SAT encoding for computing a shortest resolution proof of a given candidate formula.

2. Hitting formulas, introduced by Iwama, are an unusual class of propositional CNF formulas. Not only is their satisfiability decidable in polynomial time, but even their models can be counted in closed form. This stands in stark contrast with other polynomial-time decidable classes, which usually have algorithms based on backtracking and resolution and for which model counting remains hard, like 2-SAT and Horn-SAT. However, those resolution-based algorithms usually easily imply an upper bound on resolution complexity, which is missing for hitting formulas. Are hitting formulas hard for resolution?
   In this paper we take the first steps towards answering this question. We show that the resolution complexity of hitting formulas is dominated by so-called irreducible hitting formulas, first studied by Kullmann and Zhao, that cannot be composed of smaller hitting formulas. However, by definition, large irreducible unsatisfiable hitting formulas are difficult to construct; it is not even known whether infinitely many exist. Building upon our theoretical results, we implement an efficient algorithm on top of the Nauty software package to enumerate all irreducible unsatisfiable hitting formulas with up to 14 clauses. We also determine the exact resolution complexity of the generated hitting formulas with up to 13 clauses by extending a known SAT encoding for our purposes. Our experimental results suggest that hitting formulas are indeed hard for resolution.

## 3.17 SAT-based Local Improvement Method

*Vaidyanathan Peruvemba Ramaswamy (TU Wien, AT), Andre Schidler (TU Wien, AT), Stefan Szeider (TU Wien, AT)*

**Joint work of** Peruvemba Ramaswamy Vaidyanathan, Franz Xaver Reichl, Andre Schidler, Friedrich Slivovsky, Stefan Szeider

**Main reference** Vaidyanathan Peruvemba Ramaswamy, Stefan Szeider: "Learning Fast-Inference Bayesian Networks", in Proc. of the Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, pp. 17852–17863, 2021.
**URL** https://proceedings.neurips.cc/paper/2021/hash/94e70705efae423efda1088614128d0b-Abstract.html

**Main reference** Vaidyanathan Peruvemba Ramaswamy, Stefan Szeider: "Turbocharging Treewidth-Bounded Bayesian Network Structure Learning", in Proc. of the Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pp. 3895–3903, AAAI Press, 2021.
**URL** https://doi.org//10.1609/AAAI.V35I5.16508

**Main reference** Vaidyanathan Peruvemba Ramaswamy, Stefan Szeider: "Learning large Bayesian networks with expert constraints", in Proc. of the Uncertainty in Artificial Intelligence, Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, 1-5 August 2022, Eindhoven, The Netherlands, Proceedings of Machine Learning Research, Vol. 180, pp. 1592–1601, PMLR, 2022.
**URL** https://proceedings.mlr.press/v180/peruvemba-ramaswamy22a.html

**Main reference** Franz-Xaver Reichl, Friedrich Slivovsky, Stefan Szeider: "Circuit Minimization with QBF-Based Exact Synthesis", in Proc. of the Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023, pp. 4087–4094, AAAI Press, 2023.
**URL** https://doi.org//10.1609/AAAI.V37I4.25524

**Main reference** André Schidler, Stefan Szeider: "SAT-based Decision Tree Learning for Large Data Sets", in Proc. of the Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pp. 3904–3912, AAAI Press, 2021.
**URL** https://doi.org//10.1609/AAAI.V35I5.16509

**Main reference** André Schidler, Stefan Szeider: "SAT-Boosted Tabu Search for Coloring Massive Graphs", ACM J. Exp. Algorithmics, Vol. 28, Association for Computing Machinery, 2023.
**URL** https://doi.org//10.1145/3603112

The SAT-based Local Improvement Method (SLIM) framework has yielded several competitive heuristics for a wide variety of problems such as treewidth, branchwidth, treedepth, decision trees, graph coloring, circuit minimization, etc. SLIM starts off with an initial heuristic solution and then repeatedly replaces small local parts with an improved version. The improved version is found by solving a SAT/MaxSAT/SMT encoding of the local part. This encoding must also ensure that the improved local part is still compatible with the rest of the global solution. We call this property 'replacement consistency', and this is the key challenge in each SLIM instantiation. SLIM capitalizes on the scalability of the initial heuristic algorithm and the power of modern SAT solvers to produce heuristic solutions of higher quality than simpler local search techniques. In this talk, we give an overview of the SLIM framework and then discuss some case-studies demonstrating the application of SLIM.

## 3.18 Structures from Combinatorial Geometry and their Encodings

*Manfred Scheucher (TU Berlin, DE)*

**Main reference** Manfred Scheucher: "Two disjoint 5-holes in point sets", Comput. Geom., Vol. 91, p. 101670, 2020.
**URL** https://doi.org//10.1016/J.COMGEO.2020.101670

Point and lines are fundamental entities from geometry. We discuss Erdös-Szekeres type problems on point sets and the underlying combinatorics of point configurations and their dual structure: arrangements of lines. By slightly relaxing the geometric restrictions ("lines" dont have to be straight), we obtain so-called pseudopoint configurations and arrangements of pseudolines. While the original settings cannot be axiomized via finitely many forbidden subconfigurations unless P=NP=ETR, there are indeed purely combinatorial descriptions for "pseudo" settings which allow to make investigations using computer assistance, and in particular, using SAT.

## 3.19 Solutions of Quantified Boolean Formulas

*Martina Seidl (Johannes Kepler Universität Linz, AT) and Sibylle Möhle (MPI für Informatik - Saarbrücken, DE)*

**Joint work of** Martina Seidl, Sibylle Möhle, Andreas Plank

In this talk, we will have a closer look at solutions of quantified Boolean formulas (QBFs), i.e., the tree models of true QBFs and the tree counter-models of false QBFs and their representations as Boolean functions. These models and counter-models are of practical interest as they contain the solutions to the application problems encoded as QBFs. We will discuss how well-understood concepts from SAT like model enumeration and model counting transfer to QBF and their (counter-)models.

## 3.20 Encoding MiniZinc for SAT, MaxSAT and QUBO

*Guido Tack (Monash University - Clayton, AU)*

**Joint work of** Guido Tack, Peter J.Stuckey, Jip J. Dekker, Hendrik Bierlee

The MiniZinc modelling language lets users express their constraint satisfaction and optimisation problems in a high-level, solver-independent way. MiniZinc supports a range of decision variable types (integer, Boolean, set, float), container types (sets, arrays, tuples, records), and a large number of pre-defined predicates and functions for typical problem domains such as scheduling, packing, rostering, network problems and many others. A MiniZinc program (usually called a "model") typically represents an entire problem class, which can be turned into a concrete problem instance by supplying values for the parameters defined by the program. MiniZinc can translate problem instances into input suitable for a variety of back-end solving formalisms, including CP (Constraint Programming), MIP (Mixed Integer Linear Programming), and SAT/MaxSAT (Boolean Satisfiability). The MiniZinc system consists of a generic, back-end independent compiler/interpreter implemented in C++, and back-end specific libraries of encodings expressed in the MiniZinc language itself. This talk will cover the basic architecture of the MiniZinc translation process, and then focus on the encodings for SAT, MaxSAT and QUBO, before giving a brief outlook on the next major version of MiniZinc that is currently under development.

## 3.21 Encodings of Collatz-like problems into termination of string rewriting

*Emre Yolcu (Carnegie Mellon University - Pittsburgh, US)*

**Joint work of** Emre Yolcu, Scott Aaronson, Marijn J. H. Heule
**Main reference** Emre Yolcu, Scott Aaronson, Marijn J. H. Heule: "An Automated Approach to the Collatz Conjecture", J. Autom. Reason., Vol. 67(2), p. 15, 2023.
**URL** https://doi.org//10.1007/S10817-022-09658-8

I will describe two different ways of encoding Collatz-like problems into termination of string rewriting: one using a unary representation of integers and another using a mixed-base representation. When integers are represented in unary, the termination problem that corresponds to the Collatz conjecture (or even its simpler variants) does not admit proofs via natural matrix interpretations, a method widely used in proving termination of rewriting. I will sketch a proof of this impossibility result and then show a few instances where simply changing the encoding results in the termination problem becoming easy to solve (even automatically) via matrix interpretations.

## Participants

- Carlos Ansotegui
University of Lleida, ES
- Jeremias Berg
University of Helsinki, FI
- Olaf Beyersdorff
Friedrich-Schiller-Universität Jena, DE
- Armin Biere
Universität Freiburg, DE
- Curtis Bright
University of Windsor, CA
- Cayden Codel
Carnegie Mellon University – Pittsburgh, US
- Michael Codish
Ben Gurion University – Beer Sheva, IL
- Ronald de Haan
University of Amsterdam, NL
- Emir Demirovic
TU Delft, NL
- Andreas Falkner
Siemens AG – Wien, AT
- Johannes Klaus Fichte
Linköping University, SE
- María Andreína Francisco Rodríguez
Uppsala University, SE
- Marijn J. H. Heule
Carnegie Mellon University – Pittsburgh, US
- Matti Järvisalo
University of Helsinki, FI
- Mikoláš Janota
Czech Technical University – Prague, CZ
- George Katsirelos
INRAE – Palaiseau, FR
- Daniela Kaufmann
TU Wien, AT
- Markus Kirchweger
TU Wien, AT
- Zeynep Kiziltan
University of Bologna, IT
- Inês Lynce
University of Lisbon, PT
- Vasco Manquinho
INESC-ID – Lisbon, PT
- Valentin Mayer-Eichberger
Isotronic – Berlin, DE
- Ciaran McCreesh
University of Glasgow, GB
- Stefan Mengel
CNRS, CRIL – Lens, FR
- Sibylle Möhle
MPI für Informatik – Saarbrücken, DE
- Jakob Nordström
University of Copenhagen, DK & Lund University, SE
- Andy Oertel
Lund University, SE
- Sebastian Ordyniak
University of Leeds, GB
- Tomáš Peitl
TU Wien, AT
- Vaidyanathan Peruvemba Ramaswamy
TU Wien, AT
- Jussi Rintanen
Aalto University, FI
- Torsten Schaub
Universität Potsdam, DE
- Manfred Scheucher
TU Berlin, DE
- Andre Schidler
TU Wien, AT
- Martina Seidl
Johannes Kepler Universität Linz, AT
- Carsten Sinz
Hochschule Karlsruhe, DE
- Takehide Soh
Kobe University, JP
- Stefan Szeider
TU Wien, AT
- Guido Tack
Monash University – Clayton, AU
- Hélène Verhaeghe
KU Leuven, BE
- Hai Xia
TU Wien, AT
- Emre Yolcu
Carnegie Mellon University – Pittsburgh, US
- Tianwei Zhang
TU Wien, AT