



DAGSTUHL REPORTS

Volume 12, Issue 10, October 2022

Computer Science Methods for Effective and Sustainable Simulation Studies (Dagstuhl Seminar 22401) <i>Wentong Cai, Christopher Carothers, David M. Nicol, and Adelinde M. Uhrmacher</i>	1
Foundations for a New Perspective of Understanding Programming (Dagstuhl Seminar 22402) <i>Madeline Endres, André Brechmann, Bonita Sharif, Westley Weimer, and Janet Siegmund</i>	61
Theory and Practice of SAT and Combinatorial Solving (Dagstuhl Seminar 22411) <i>Olaf Beyersdorff, Armin Biere, Vijay Ganesh, Jakob Nordström, and Andy Oertel</i>	84
Intelligent Security: Is “AI for Cybersecurity” a Blessing or a Curse (Dagstuhl Seminar 22412) <i>Nele Mentens, Stjepan Picek, and Ahmad-Reza Sadeghi</i>	106
Security of Decentralized Financial Technologies (Dagstuhl Seminar 22421) <i>Arthur Gervais and Marie Vasek</i>	129
Developmental Machine Learning: From Human Learning to Machines and Back (Dagstuhl Seminar 22422) <i>Pierre-Yves Oudeyer, James M. Rehg, Linda B. Smith, and Sho Tsuji</i>	143
Data-Driven Combinatorial Optimisation (Dagstuhl Seminar 22431) <i>Emma Frejinger, Andrea Lodi, Michele Lombardi, and Neil Yorke-Smith</i>	166
Towards a Unified Model of Scholarly Argumentation (Dagstuhl Seminar 22432) <i>Khalid Al-Khatib, Anita de Waard, Dayne Freitag, Iryna Gurevych, Yufang Hou, and Harrisen Scells</i>	175
Optimization at the Second Level (Dagstuhl Seminar 22441) <i>Luce Brotcorne, Christoph Buchheim, Dick den Hertog, and Dorothee Henke</i>	207
Toward Scientific Evidence Standards in Empirical Computer Science (Dagstuhl Seminar 22442) <i>Timothy Kluthe, Brett A. Becker, Christopher D. Hundhausen, Ciera Jaspán, Andreas Stefik, and Thomas Zimmerman</i>	225

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

May, 2023

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <https://dnb.d-nb.de>.

License

This work is licensed under a Creative Commons Attribution 4.0 International license (CC BY 4.0).



In brief, this license authorizes each and everybody to share (to copy, distribute and transmit) the work under the following conditions, without impairing or restricting the authors' moral rights:

- Attribution: The work must be attributed to its authors.

The copyright is retained by the corresponding authors.

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.

Editorial Board

- Elisabeth André
- Franz Baader
- Daniel Cremers
- Goetz Graefe
- Reiner Hähnle
- Barbara Hammer
- Lynda Hardman
- Oliver Kohlbacher
- Steve Kremer
- Rupak Majumdar
- Heiko Mantel
- Albrecht Schmidt
- Wolfgang Schröder-Preikschat
- Raimund Seidel (*Editor-in-Chief*)
- Heike Wehrheim
- Verena Wolf
- Martina Zitterbart

Editorial Office

Michael Wagner (*Managing Editor*)
Michael Didas (*Managing Editor*)
Jutka Gasiorowski (*Editorial Assistance*)
Dagmar Glaser (*Editorial Assistance*)
Thomas Schillo (*Technical Assistance*)

Contact

Schloss Dagstuhl – Leibniz-Zentrum für Informatik
Dagstuhl Reports, Editorial Office
Oktavie-Allee, 66687 Wadern, Germany
reports@dagstuhl.de

Digital Object Identifier: 10.4230/DagRep.12.10.i

<https://www.dagstuhl.de/dagrep>

Computer Science Methods for Effective and Sustainable Simulation Studies

Wentong Cai^{*1}, Christopher Carothers^{*2}, David M. Nicol^{*3}, and Adelinde M. Uhrmacher^{*4}

1 Nanyang TU – Singapore, SG. aswtcai@ntu.edu.sg

2 Rensselaer Polytechnic Institute – Troy, US. chrisc@cs.rpi.edu

3 University of Illinois – Urbana Champaign, US. dmmicol@illinois.edu

4 Universität Rostock, DE. adelinde.uhrmacher@uni-rostock.de

Abstract

This report documents the program and the (preliminary) outcomes of Dagstuhl Seminar 22401 “Computer Science Methods for Effective and Sustainable Simulation Studies”. The seminar has been dedicated to addressing central methodological challenges in conducting effective and sustainable simulation studies. Lightning talks provided the opportunity for participants to present their current research and ideas to advance methodological research in modeling and simulation. However, the lion’s share of the seminar was dedicated to working groups. One working group investigated how machine learning and modeling and simulation can be effectively integrated (Intelligent Modeling and Simulation Lifecycle). Another working group focused on methodological challenges to support policy via simulation (Policy by simulation: seeing is believing for interactive model co-creation and effective intervention). A third working group identified 4 challenges closely tied to the quest for sustainable simulation studies (Context, composition, automation, and communication – towards sustainable simulation studies) thereby, focusing on the role of model-based approaches and related methods.

Seminar October 3–7, 2022 – <http://www.dagstuhl.de/22401>

2012 ACM Subject Classification Computing methodologies

Keywords and phrases Modeling, simulation, high performance computing, machine learning, visual analytics

Digital Object Identifier 10.4230/DagRep.12.10.1

1 Executive Summary

Adelinde M. Uhrmacher (Universität Rostock, DE)

Wentong Cai (Nanyang TU – Singapore, SG)

Christopher Carothers (Rensselaer Polytechnic Institute – Troy, US)

David M. Nicol (University of Illinois – Urbana Champaign, US)

License © Creative Commons BY 4.0 International license

© Adelinde M. Uhrmacher, Wentong Cai, Christopher Carothers, and David M. Nicol

Motivation. Simulation becomes more and more important in application areas, establishing itself as the third way of science in addition to theory and (real) experiments. To answer research questions, simulation studies form increasingly intricate processes that intertwine the design and execution of various, often calculation-intensive simulation experiments, the generation and refinement of simulation models, and steps of analysis.

* Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Computer Science Methods for Effective and Sustainable Simulation Studies, *Dagstuhl Reports*, Vol. 12, Issue 10, pp. 1–60

Editors: Wentong Cai, Christopher Carothers, David M. Nicol, and Adelinde M. Uhrmacher



DAGSTUHL Dagstuhl Reports

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

	Monday, 10/03	Tuesday, 10/04	Wednesday, 10/05	Thursday, 10/06	Friday, 10/07
08 a.m.		Breakfast	Breakfast	Breakfast	Breakfast
09 a.m.		Introduction incl. identifying challenges	Working groups	Intermediate results	Summary/Work.G.
10 a.m.		30 min. Coffee	30 min. Coffee	30 min. Coffee	30 min. Coffee
11 a.m.		Model-based approaches for Modeling and Simulation	Working groups	Working groups	Final summary
12 noon		Lunch	Lunch	Lunch	Lunch
01 p.m.		Visual analytics, experiment design, statistical model checking	Lightning talks	Working groups	Departure
02 p.m.		10 min. Coffee			
03 p.m.		High Performance Computing: exploiting new architectures			
04 p.m.	Arrival	30 min. Coffee	30 min. Coffee	30 min. Coffee	
05 p.m.		Collecting working group ideas	Hiking	Working groups	
06 p.m.		Dinner	Dinner	Dinner	

■ **Figure 1** Schedule of the Dagstuhl Seminar *Computer Science Methods for Effective and Sustainable Simulation Studies*.

The Dagstuhl Seminar has been dedicated to addressing central methodological challenges in supporting the conduction of effective and sustainable simulation studies. Thereby, the seminar focused on problems and solutions related to improving:

- Effectiveness: the usage of resources, including computing infrastructure and data, and the assistance of humans throughout a simulation study.
- Sustainability: continuing a simulation study into the future through support for reusing or building upon its central products, such as simulation model, data, and processes as well as the software used.

The last decades have seen a wide range of methodological developments in computer science that are likely to be instrumental in achieving effective and sustainable simulation studies. However, those efforts are scattered across different computer science fields that include high-performance computing, (modeling) language design, operations research, visual analytics, workflows, provenance, and machine learning, as well as modeling and simulation. The seminar brought participants with diverse computer science backgrounds together to enhance the methodological basis for conducting simulation studies.

Organization and results. Being one day shorter than typical seminars, the seminar started on Tuesday with a short round of introduction and continued with collecting ideas about achievements and challenges of modeling and simulation from the participants on 2 pinboards (see Figure 1). 3 talks and partly extensive discussions followed, one focusing on modeling and model-based approaches applied to simulation studies, one on high-performance computing for simulation, and one on analysis and experiment designs. In the late afternoon, the information gathered on the pinboards was revisited. In the end 3 working groups formed to work towards state-of-the-art and open-challenges papers on the following topics:

- Intelligent Modeling and Simulation Lifecycle
- Policy by simulation: seeing is believing for interactive model co-creation and effective intervention
- Context, composition, automation, and communication: towards sustainable simulation studies

Among the application fields as diverse as cell biological systems, traffic systems, or computer networks, one application dominated the discussions, i.e., Covid 19 simulation. The Covid pandemic showed the importance of modeling and simulation studies being conducted in an efficient, reliable manner, and, accordingly, of comprehensive, intelligent computer support for these studies, it revealed limitations, including those referring to communicating effectively modeling and simulation studies and their results to decision-makers. The results of the working groups are included as short summaries in this report. Wednesday afternoon, the participants presented their current research work and ideas in a series of lightning talks whose abstracts are also included in the report. However, most of the time was dedicated to the working groups. Plenary sessions on Thursday and finally on Friday allowed the participants to catch up with ideas and the progress made in the different working groups.

2 Table of Contents

Executive Summary

Adeline M. Uhrmacher, Wentong Cai, Christopher Carothers, and David M. Nicol 1

Overview of Talks

Towards Differentiable Agent-Based Simulation <i>Philipp Andelfinger</i>	6
Parametric verification of stochastic model using stochastic variational inference <i>Luca Bortolussi</i>	7
Beyond DDDAS and symbiotic simulation <i>Wentong Cai</i>	8
Towards a new facility for model-based design and evaluation of sustainable complex systems <i>Rodrigo Castro</i>	10
Challenges for Sustainable Twinning <i>Joachim Denil and Stijn Bellis</i>	10
The ASTRÉE Analyzer <i>Jérôme Feret</i>	12
Fighting COVID-19 with Simulation <i>Peter Frazier</i>	13
Virtual Time Integration of Emulation and Simulation Systems for Smart Grid Application Testing and Evaluation <i>Dong (Kevin) Jin</i>	13
Towards an Open Repository for Reproducible Performance Comparison of Parallel and Distributed Discrete-Event Simulators <i>Till Köster, Philipp Andelfinger, and Adeline M. Uhrmacher</i>	15
Simulation Based Analysis of Social Systems – Models, Data and Policy <i>Michael Lees</i>	16
Parallel Simulation – What Worked and What Not <i>Jason Liu</i>	17
Simulation at the Edge <i>Margaret Loper</i>	19
Bottlenecks of using simulation for policy making <i>Fabian Lorig</i>	21
Supporting Transparent Simulation Studies: The Role of Provenance Information <i>Bertram Ludäscher</i>	22
Interactive Visual Analysis for Simulation <i>Kresimir Matkovic</i>	24
A logic-based approach to reason about large-scale spatially-distributed systems <i>Laura Nenzi</i>	26
On the Attractiveness of Speculative PDES: Challenges and Pitfalls <i>Alessandro Pellegrini</i>	26

Methods for Integrated Simulation – from Data Acquisition to Decision Support <i>Niki Popper</i>	29
Using the Adaptable I/O System (ADIOS) for Effective and Sustainable Simulation Studies <i>Caitlin Ross</i>	32
A Simulation Architecture to Study Diffusion Processes in Multiplex Networks <i>Cristina Ruiz-Martin</i>	34
Model-based Software and Systems Engineering for Digital Twins <i>Bernhard Rumpe</i>	36
Data Farming: Better Decisions Via Inferential Big Data <i>Susan Sanchez</i>	37
Decision Making using Reinforcement Learning in Contested and Dynamic Environments <i>Claudia Szabo</i>	38
Simulation-based Inference for Automatic Model Construction <i>Wen Jun Tan</i>	39
Models and Specifications within the Modeling and Simulation Life Cycle <i>Adelinde M. Uhrmacher and Claudia Szabo</i>	40
A discrete-event approach for the study of sustainability in buildings <i>Gabriel A. Wainer</i>	40
Automatically Generating Simulation Experiments based on Provenance <i>Pia Wilsdorf</i>	43
Working groups	
Intelligent Modeling and Simulation Lifecycle <i>Wentong Cai, Philipp Andelfinger, Luca Bortolussi, Christopher Carothers, Dong (Kevin) Jin, Till Köster, Michael Lees, Jason Liu, Margaret Loper, Alessandro Pellegrini, Wen Jun Tan, and Verena Wolf</i>	44
Policy by simulation: seeing is believing for interactive model co-creation and effective intervention <i>Rodrigo Castro, Joachim Denil, Jérôme Feret, Kresimir Matkovic, Niki Popper, Susan Sanchez, and Peter Slood</i>	47
Context, composition, automation and communication: towards sustainable simulation studies <i>Adelinde M. Uhrmacher, Peter Frazier, Reiner Hähnle, Franziska Klügl, Fabian Lorig, Bertram Ludäscher, Laura Nenzi, Cristina Ruiz-Martin, Bernhard Rumpe, Claudia Szabo, Gabriel A. Wainer, and Pia Wilsdorf</i>	53
Participants	60

3 Overview of Talks

3.1 Towards Differentiable Agent-Based Simulation

Philipp Andelfinger (Universität Rostock, DE)

License © Creative Commons BY 4.0 International license

© Philipp Andelfinger

Main reference Philipp Andelfinger: “Towards Differentiable Agent-Based Simulation”, ACM Trans. Model. Comput. Simul., Vol. 32(4), Association for Computing Machinery, 2023.

URL <https://doi.org/10.1145/3565810>

Agent-based simulation models aiding the understanding, design, and optimization of systems reside on a spectrum with two extremes [5, 8]: *mechanistic models* are constructed manually by formalizing domain knowledge about the structure and behavior of the system under study, whereas *data-driven models* are generated by fitting generic parametric models against empirical observations of the system. While mechanistic models are typically also parameterized, adjusting the parameters does not alter the fundamental model logic.

The parameter synthesis for model calibration or optimization takes different forms depending on the model category: data-driven models usually permit the computation of the partial derivatives of the model output with respect to the parameters using automatic differentiation algorithms such as backpropagation [9]. Hence, gradient-based methods can be used to steer the parameter combination towards local optima in the model’s response surface. In contrast, mechanistic agent-based models tend to incorporate discrete decision-making logic [5], which can lead to response surfaces dominated by discontinuities and plateaus, thus largely preventing the fruitful use of gradient-based methods. As a consequence, most simulation optimization efforts using agent-based models employ metaheuristics [6] such as genetic algorithms or metamodeling approaches [4, 3], which generate an approximative surrogate of the original model.

We explore methods to make agent-based models involving discontinuous building blocks amenable to the automatic computation of gradients, under the hypothesis that the directed local search afforded by gradient-based methods may exhibit better convergence behavior than the existing approaches. Further, our aim is to enable the integration of the domain knowledge encoded in mechanistic agent-based models with the flexibility of data-driven models. By capturing the behavior of such a combined model, the computed gradients can serve to swiftly identify high-quality solutions for problems in calibration, optimization, control, and reinforcement learning.

Previously, we showed that when weighting the effects of different branches in the logic of agent-based models using a smoothing function, gradients computed using automatic differentiation can be used to accelerate the progress in traffic light control problems using microscopic traffic simulations [1, 2]. Moreover, we showed that the integration of the differentiable simulation model with a neural network enables the gradient-based training of a neural traffic light controller. The weighted execution of branches can be regarded as an approximation of an exact probabilistic program semantics, which is prohibitively expensive in practical cases [7]. However, using simplifications, the computational costs can be reduced to an acceptable level.

In the future, to avoid the need for modelers to manually apply smoothing to their agent-based models, languages or APIs are required that allow simulation models to be executed in their original or in a smoothed form. As models must be expected to vary severely in their suitability for smoothing and their potential for improvements in optimization progress, e.g., depending on the presence of continuous model elements, a categorization of models according to such properties would be beneficial.

By providing a natural way to unify mechanistic and data-driven models and the gradient-based methods used for optimization, we hope for this work to reduce the gap between the communities focused on mechanistic and data-driven modeling.

References

- 1 Philipp Andelfinger. Differentiable agent-based simulation for gradient-guided simulation-based optimization. In *Proceedings of the 2021 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, pages 27–38, 2021.
- 2 Philipp Andelfinger. Towards differentiable agent-based simulation. *ACM Trans. Model. Comput. Simul.*, 2022.
- 3 Russell R Barton. Tutorial: metamodeling for simulation. In *2020 Winter Simulation Conference (WSC)*, pages 1102–1116. IEEE, 2020.
- 4 Atharv Bhosekar and Marianthi Ierapetritou. Advances in surrogate based modeling, feasibility analysis, and optimization: A review. *Computers & Chemical Engineering*, 108:250–267, 2018.
- 5 Eric Bonabeau. Agent-based modeling: methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, 99(suppl 3):7280–7287, 2002.
- 6 Benoît Calvez and Guillaume Hutzler. Automatic tuning of agent-based models using genetic algorithms. In *International Workshop on Multi-Agent Systems and Agent-Based Simulation*, pages 41–57. Springer, 2005.
- 7 Swarat Chaudhuri and Armando Solar-Lezama. Smoothing a program soundly and robustly. In *International Conference on Computer Aided Verification*, pages 277–292. Springer, 2011.
- 8 Hamdi Kavak, Jose J Padilla, Christopher J Lynch, and Saikou Y Diallo. Big data, agents, and machine learning: towards a data-driven agent-based modeling approach. In *Proceedings of the Annual Simulation Symposium*, pages 1–12, 2018.
- 9 David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

3.2 Parametric verification of stochastic model using stochastic variational inference

Luca Bortolussi (University of Trieste, IT)

License © Creative Commons BY 4.0 International license
© Luca Bortolussi

Joint work of Luca Bortolussi, Francesca Cairolì, Ginevra Carbone, Paolo Pulcini
Main reference Luca Bortolussi, Francesca Cairolì, Ginevra Carbone, Paolo Pulcini: “Scalable Stochastic Parametric Verification with Stochastic Variational Smoothed Model Checking”, arXiv, 2022.
URL <https://doi.org/10.48550/ARXIV.2205.05398>

Parametric verification for stochastic models can be expressed as checking the satisfaction probability of a certain property as a function of the parameters of the model. Smoothed model checking (smMC) [1] aims at inferring the satisfaction function over the entire parameter space from a limited set of observations obtained via simulation. As observations are costly and noisy, smMC is framed as a Bayesian inference problem so that the estimates have an additional quantification of the uncertainty. In [1] the authors use Gaussian Processes (GP), inferred by means of the Expectation Propagation algorithm. This approach provides accurate reconstructions with statistically sound quantification of the uncertainty. However, it inherits the well-known scalability issues of GP. In this paper, we exploit recent advances in probabilistic machine learning to push this limitation forward, making Bayesian inference of smMC scalable to larger datasets, enabling its application to models with high dimensional


parameter spaces. We propose Stochastic Variational Smoothed Model Checking (SV-smMC), a solution that exploits stochastic variational inference (SVI) to approximate the posterior distribution of the smMC problem. The strength and flexibility of SVI make SV-smMC applicable to two alternative probabilistic models: Gaussian Processes (GP) and Bayesian Neural Networks (BNN). The core ingredient of SVI is a stochastic gradient-based optimization that makes inference easily parallelizable and it enables GPU acceleration. In this paper, we compare the performances of smMC [1] against those of SV-smMC by looking at the scalability, the computational efficiency and at the accuracy of the reconstructed satisfaction function.

References

- 1 Bortolussi, L., Milios, D., Sanguinetti, G.: Smoothed model checking for uncertain continuous-time markov chains. *Information and Computation* 247, 235–253 (2016)

3.3 Beyond DDDAS and symbiotic simulation

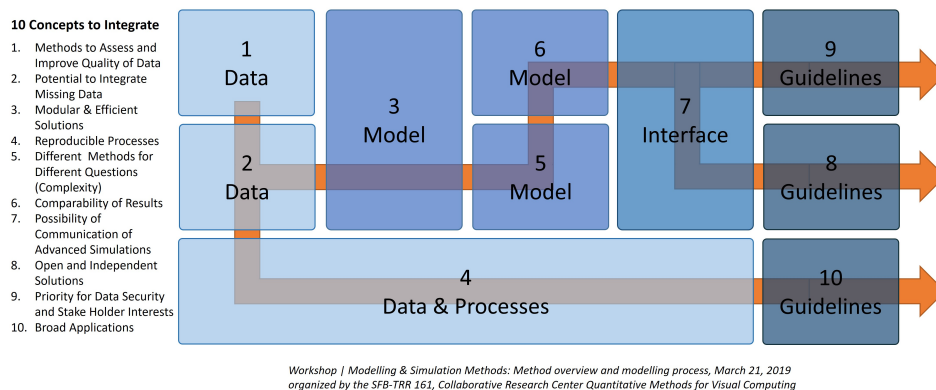
Wentong Cai (Nanyang TU – Singapore, SG)

License  Creative Commons BY 4.0 International license
© Wentong Cai

Darema in 2000 as a paradigm where measurement data from an operational system is dynamically incorporated into an execution model of that system, and computational results from the model are then used to guide the measurement process. Independently, Symbiotic Simulation [2] paradigm proposed in Dagstuhl Seminar on Grand Challenges for Modelling and Simulation in 2002 solves the what-if problem by having the simulation system and the physical system interact in a mutually beneficial manner. Since their inception, many techniques have been developed to support DDDAS and symbiotic simulations. Digital Twin, originally proposed by Michael Grieves in 2002 and popularized in recent years due to the rise of IoT and AI, is “a real-time virtual representation of a real-world physical system or process that serves as the indistinguishable digital counterpart of it for practical purposes, such as system simulation, integration, testing, monitoring, and maintenance” [3]. Digital twin is a concept. DDDAS and symbiotic simulation techniques form the basis for the realization of digital twins. They emphasize on the interaction between the virtual and physical systems.

Moving beyond DDDAS and symbiotic simulation, data-driven and machine learning (ML) techniques should be integrated in various stages of M&S. For instance, in addition to use sensor data from physical system to calibrate simulation models, ML techniques can be used to extract useful knowledge and insight from the data to facilitate model development. Data analytics and ML techniques can also be used to manipulate or steer simulation experiments on the fly (see Figure 2).

Some examples of our recent works along this direction include: Our PADS’21 paper [4] is about how to use ML approach to create a car-following model (instead of using traditional physics-based model) and how to dynamically calibrate the model using online data. Our recent research work focuses more on using data-driven approach to improve performance of simulation execution and simulation-based optimization. Our PADS’20 & PADS’22 papers [5, 6] are about dynamically analyzing simulation state to determine level of details to be used in the model of a simulation entity during simulation execution. The objective is to reduce the simulation runtime while maintaining accuracy of the simulation results. We applied the approach to a semi-conductor manufacturing simulation. And our WSC’18



■ **Figure 2** Beyond DDDAS – Integration of machine learning.


paper [7] uses an approach to dynamically predict the usefulness of a simulation run. If the results of a simulation run won't contribute to the overall optimization objective, then the simulation run can be terminated early. In this way, the total number of simulation runs required in a simulation-based optimization process will be reduced.

References

- 1 Dynamic Data Driven Applications System. (2022), https://en.wikipedia.org/wiki/Dynamic_Data_Driven_Applications_System, [Online; accessed 23-Sept-2022]
- 2 R. Fujimoto, D. Lunceford, E. Page, and A. Uhrmacher. Grand challenges for modeling and simulation. *Schloss Dagstuhl*. 350 (2002)
- 3 Digital twin. (2022), https://en.wikipedia.org/wiki/Digital_twin, [Online; accessed 23-Sept-2022]
- 4 Htet Naing, Wentong Cai, Nan Hu, Tiantian Wu, and Liang Yu. Data-driven microscopic traffic modelling and simulation using dynamic lstm. *Proceedings Of The 2021 ACM SIGSIM Conference On Principles Of Advanced Discrete Simulation*. pp. 1-12 (2021)
- 5 Moon Gi Seok, Chew Wye Chan, Wentong Cai, Daejin Park, and Hessam S. Sarjoughian. Adaptive abstraction-level conversion framework for accelerated discrete-event simulation in smart semiconductor manufacturing. *IEEE Access*. 8 pp. 165247-165262 (2020)
- 6 Moon Gi Soon, Wen Jun Tan, and Wentong Cai. Hyperparameter Tuning in Simulation-based Optimization for Adaptive Digital-Twin Abstraction Control of Smart Manufacturing System. *Proceedings Of The 2022 ACM SIGSIM Conference On Principles Of Advanced Discrete Simulation*. pp. 61-68 (2022)
- 7 Philipp Andelfinger, Sajeev Udayakumar, David Eckhoff, Wentong Cai, and Alois Knoll. A. Model preemption based on dynamic analysis of simulation data to accelerate traffic light timing optimisation. *2018 Winter Simulation Conference (WSC)*. pp. 652-663 (2018)

3.4 Towards a new facility for model-based design and evaluation of sustainable complex systems

Rodrigo Castro (*University of Buenos Aires, AR*)

License  Creative Commons BY 4.0 International license
© Rodrigo Castro

Problems involving societies and their interactions with cybernetic systems in the context of physical restrictions represent a paradigmatic case of Complex Adaptive Systems (CAS) involving emergent behavior and micro-macro loops.

The study of dynamic CAS lack analytical solutions thus requiring simulation as the only means for quantitative research.

If we add a layer of goal-seeking governance to inform real-world policy-making, legitimately contradictory worldviews must be factored in. This paves the way to reaching what has been termed as Wicked Problems, those that do not accept “correct” definitions nor “optimal” solutions, but rather require discursive processes to reach a consensus about the models themselves.

We postulate Planning for Sustainable Egalitarian Development as an embracing, flagship case study that pushes the envelope of sustainable simulation, thus challenging the state of the art and practice of several computer science disciplines (HPC, model checking, validation and verification, visualization, to name a few)


The proposal includes building on control theory to include simulation models in the loop of decision-making processes, creating a simulation-assisted arena to experiment with interventions, thus obtaining “feasible future developments” and analyzing them with advanced visual analytics.

We envision a framework within which planners (human intelligence) and algorithms (artificial intelligence) inform each other to obtain better strategies to use the simulation models as demonstrators of feasible paths of development.

An advanced facility, such as an interactive and immersive visualization room for complex data and reactive simulations shall integrate and boost capabilities for participatory model-based design and evaluation of sustainable complex systems.

3.5 Challenges for Sustainable Twinning

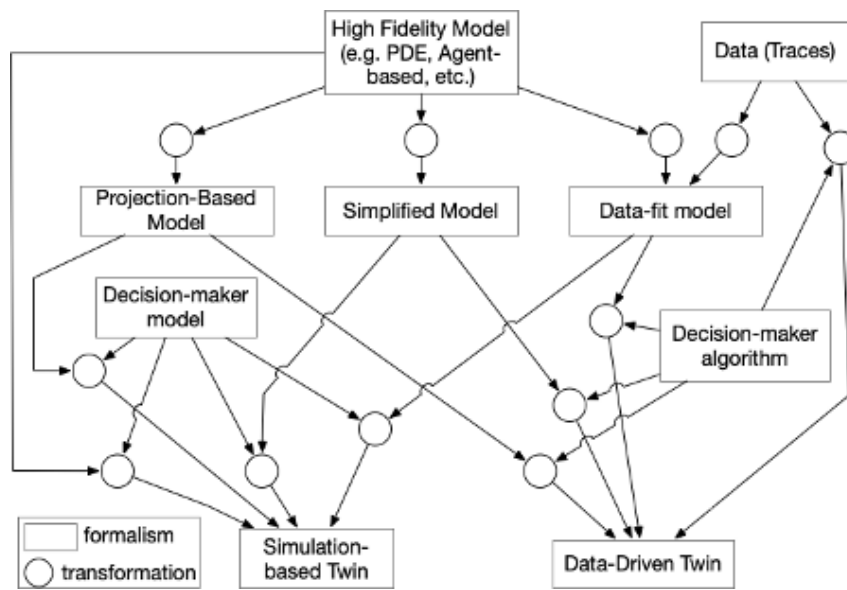
Joachim Denil (*University of Antwerp, BE*) and Stijn Bellis

License  Creative Commons BY 4.0 International license
© Joachim Denil and Stijn Bellis

Main reference Stijn Bellis, Joachim Denil: “Challenges and possible approaches for sustainable digital twinning”, in Proc. of the 25th International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings, MODELS 2022, Montreal, Quebec, Canada, October 23-28, 2022, pp. 643–648, ACM, 2022.

URL <https://doi.org/10.1145/3550356.3561551>

Advances in digital twin technology are creating value for many companies. We consider, from a sustainability perspective, how these digital twins can be better developed. At first glance, the energy consumption of the twin during its life cycle can be described as follows: $E_{total} = E_{design} + E_{local} + E_{networking} + E_{cloud} + E_{update}$. Decomposing this formula allows us to see several challenges in the design and operation of a digital twin.



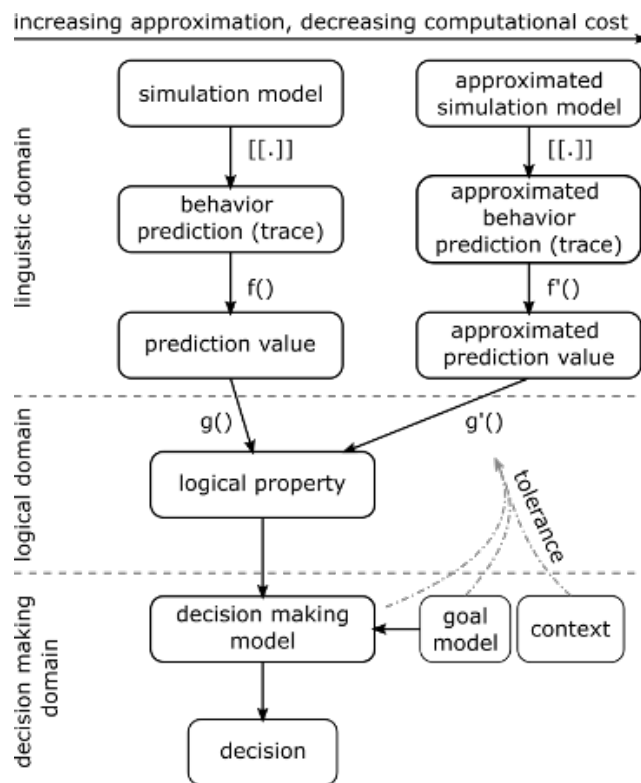
■ **Figure 3** Transformation of Models for a Digital Twin Architecture.

The choice of formalism. When developing a twin from detailed physics models (typically modelled using partial differential equations), different choices concerning the formalism(s) must be made. In Figure 3, we show some simplified paths to obtaining a digital twin. Each operation takes a certain amount of energy while running a simulation using a specific formalism during the twin lifecycle also takes energy. This results in trade-offs that need to be managed. For example, creating a lumped model takes a lot of engineering time while training a neural net takes a lot of energy.

The value proposition of the twin. From the perspective of simulation engineering, the purpose has a huge influence on the engineering of the simulation model. Figure 4 shows how simulation is used to decide on properties of interest. It also shows how approximated models provide value for their users. Estimating how much uncertainty the decision maker can tolerate for gaining enough value from the digital twin is a difficult problem. Furthermore, once we know the allowable tolerance, we still need to include this in the formalism selection process.

The evolution of the system and the twin. To allow for the long and continuous operation of the digital twin, we need insight into the range of validity of the model in combination with insights into the system's evolution. Including the dimensions of evolution within the formalism selection process is needed. Having a good estimate of these evolutions in frequency and severity helps determine the needed boundary conditions and validity of the model. If not considered, a new model needs to be used, possibly a model that consumes more energy.

The deployment of the twin. The final challenge is to reason about the deployment choices related to the deployment architecture used for the digital twin. Most choices are impacted by the system's requirements, which in turn depend on the value proposition of the twin. Some examples of choices include (a) where to run (parts of) the twin: local, edge or cloud? (b) networking and telemetry choices, (c) cloud infrastructure and storage options.



■ **Figure 4** Approximations and their Effect on Validity.

3.6 The ASTRÉE Analyzer

Jérôme Feret (ENS – Paris, FR)

License © Creative Commons BY 4.0 International license

© Jérôme Feret

Joint work of Bruno Blanchet, Patrick Cousot, Radhia Cousot, Jérôme Feret, Laurent Mauborgne, Antoine Miné, David Monniaux, Xavier Rival

Main reference Bruno Blanchet, Patrick Cousot, Radhia Cousot, Jérôme Feret, Laurent Mauborgne, Antoine Miné, David Monniaux, Xavier Rival: “A static analyzer for large safety-critical software”, in Proc. of the ACM SIGPLAN 2003 Conference on Programming Language Design and Implementation 2003, San Diego, California, USA, June 9-11, 2003, pp. 196–207, ACM, 2003.

URL <https://doi.org/10.1145/781131.781153>

The ASTRÉE analyzer [2, 3, 4] aims at statically proving the absence of Run-Time errors in critical embedded software. Following the abstract interpretation framework [1], it is based on a formal semantics of C. Then various abstract domains are available to abstract this semantics. The correctness of the approach is proven by construction. ASTRÉE has been used successfully to certify the absence of run-time errors in the primary flight control software of the A340 (2003), the primary flight control software of the A380 (2005), and a C version of the animatic docking software of the Jules Vernes ATV (for the International Space Station). Since 2009, it is commercialized by Absint Angewandte Informatik (Saarbrücken) and is used mainly in automotive software, but also, in avionic and nuclear software.

References

- 1 Cousot, P., Cousot, R.: Abstract interpretation: A unified lattice model for static analysis of programs by construction or approximation of fixpoints. In: Graham, R.M., Harrison, M.A., Sethi, R. (eds.) POPL’77: Proceedings of the 4th Symposium on Principles of Programming Languages. pp. 238–252. ACM (1977)

- 2 Blanchet, B., Cousot, P., Cousot, R., Feret, J., Mauborgne, L., Miné, A., Monniaux, D., Rival, X.: Design and implementation of a special-purpose static program analyzer for safety-critical real-time embedded software. In: Mogensen, T.Æ., Schmidt, D.A., Sudborough, I.H. (eds.) *The Essence of Computation, Complexity, Analysis, Transformation. Essays Dedicated to Neil D. Jones [on occasion of his 60th birthday]*. Lecture Notes in Computer Science, vol. 2566, pp. 85–108. Springer (2002). https://doi.org/10.1007/3-540-36377-7_5
- 3 Blanchet, B., Cousot, P., Cousot, R., Feret, J., Mauborgne, L., Miné, A., Monniaux, D., Rival, X.: A static analyzer for large safety-critical software. In: Cytron, R., Gupta, R. (eds.) *Proceedings of the ACM SIGPLAN 2003 Conference on Programming Language Design and Implementation 2003, San Diego, California, USA, June 9-11, 2003*. pp. 196–207. ACM (2003). <https://doi.org/10.1145/781131.781153>
- 4 Cousot, P., Cousot, R., Feret, J., Mauborgne, L., Miné, A., Monniaux, D., Rival, X.: The astreé analyzer. In: Sagiv, S. (ed.) *Programming Languages and Systems, 14th European Symposium on Programming, ESOP 2005, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2005, Edinburgh, UK, April 4-8, 2005, Proceedings*. Lecture Notes in Computer Science, vol. 3444, pp. 21–30. Springer (2005). https://doi.org/10.1007/978-3-540-31987-0_3

3.7 Fighting COVID-19 with Simulation

Peter Frazier (Cornell University – Ithaca, US)

License © Creative Commons BY 4.0 International license
© Peter Frazier

Joint work of Peter I. Frazier, J. Massey Cashore, Ning Duan, Shane G. Henderson, Alyf Janmohamed, Brian Liu, David B. Shmoys, Jiayue Wan, Yujia Zhang
Main reference Peter I. Frazier, J. Massey Cashore, Ning Duan, Shane G. Henderson, Alyf Janmohamed, Brian Liu, David B. Shmoys, Jiayue Wan, Yujia Zhang: “Modeling for COVID-19 college reopening decisions: Cornell, a case study”, *Proceedings of the National Academy of Sciences*, Vol. 119(2), p. e2112532119, 2022.

URL <https://doi.org/10.1073/pnas.2112532119>

Universities faced a difficult decision in summer 2020: whether to reopen for in-person instruction despite the pandemic and how to protect students, employees, and the surrounding community if they did. Simulation was critical to this decision at Cornell University in the USA, which successfully reopened for in-person instruction in fall 2020 under the protection of an asymptomatic screening that tested all undergraduate students twice per week. This talk discusses key factors that helped simulation modelers earn credibility and influence: transparency, designing a testing strategy that was robust to the unknown, explaining analysis in a simple clear way, understanding stakeholders’ incentives, responsiveness to stakeholder questions, and a focus on providing value.

3.8 Virtual Time Integration of Emulation and Simulation Systems for Smart Grid Application Testing and Evaluation

Dong (Kevin) Jin (University of Arkansas – Fayetteville, US)

License © Creative Commons BY 4.0 International license
© Dong (Kevin) Jin

Modern energy systems are increasingly adopting Internet technology to boost control efficiency, which unfortunately opens up a new security frontier. As a result, extensive applications have been proposed to enhance the cyber resilience and security of those critical

infrastructures. Incorporation of new technologies in such systems is very challenging because of strong real-time requirements, continuous system availability, and many resource-constrained legacy devices. Therefore, a testing platform targeting such cyber-physical systems is strongly needed for the research community to evaluate the new application and system designs and their impact on the power systems before the real deployment.

We develop a unique testbed, DSSnet [1], combining container-based network emulation and power system simulation using a novel Linux-kernel-based virtual time system. DSSnet enables the modeling of a modern power distribution system and simulates the Intelligent Electrical Devices that make it up. DSSnet also enables high-fidelity analysis by allowing real networking applications to run in the network emulator and interact with the power simulator. DSSnet is composed of the following main components: (1) a container-based network emulator, Mininet [2] that allows execution of real software and communication network applications, (2) an electrical power distribution system simulator, OpenDSS that enables power flow simulation studies [3], (3) a unique Linux-kernel-based virtual time system [4] for synchronization of the two sub-systems, which significantly enhances the temporal fidelity issues in ordinary co-simulation or hardware-in-the-loop testbeds; (4) two coordinators for interfacing with the cyber- and physical-side modules and the virtual time system; and (5) a distributed software-defined networking (SDN) control environment, ONOS [5] that provides high-level abstractions and APIs for power grid control applications to manage, monitor and program the emulated communication network.

One key challenge is synchronizing the execution of the power simulator and the container-based emulator. This is because all the processes in the emulator execute real programs and use the system clock to advance experiments, while the simulator executes models to advance experiments with respect to its simulation virtual clock. To address this issue, we developed an independent and lightweight middleware in the Linux kernel to support virtual time for Linux container [4]. Our system transparently provides the virtual time to processes inside the containers, while returning the ordinary system time to other processes. No change is required in applications. Next, we expanded the capability of the testbed with a distributed virtual time system [6] that enables processes and their clocks to be paused, resumed, and dilated across embedded Linux devices through the use of hardware interrupts and a common kernel module. The distributed system architecture uniquely consists of a common virtual time Linux kernel module and three communication channels, one for virtual time synchronization using general-purpose-input-and-output (GPIO) hardware interrupts, one for connecting the embedded Linux devices, and one for interfacing with the physical system simulation that performs an offline computation. Additionally, we modeled and analyzed the temporal error during non-CPU operations, such as disk I/O, network I/O, and GPU computation, and developed a barrier-based time compensation mechanism to enable accurate virtual time advancement with precise I/O time measurement and compensation [7].

In summary, we present DSSnet, a testing platform that combines an electrical power system simulator and a communication network emulator using a virtual time system. DSSnet can be used to model and simulate power flows, communication networks, and smart grid control applications, and to test and evaluate the effect of network applications on the smart grid.

References

- 1 Christopher Hannon, Jiaqi Yan, and Dong Jin. DSSnet: A smart grid modeling platform combining electrical power distribution system simulation and software-defined networking emulation. ACM SIGSIM-PADS, 2016.
- 2 Mininet: An Instant Virtual Network on your Laptop (or other PC). <http://mininet.org/>

- 3 OpenDSS: An Electrical Power System Simulation Tool for Distribution Systems, Elect. Power Res. Inst. <http://smartgrid.epri.com/SimulationTool.aspx>
- 4 Jiaqi Yan and Dong Jin. A Virtual Time System for Linux-container-based Emulation of Software-defined Networks. ACM SIGSIM-PADS, 2015.
- 5 ONOS: Open Network Operating System. <https://opennetworking.org/onos/>
- 6 Christopher Hannon, Jiaqi Yan, Dong Jin, and Yuan-An Liu. A Distributed Virtual Time System on Embedded Linux for Evaluating Cyber-Physical Systems. ACM SIGSIM-PADS, 2019.
- 7 Gong Chen, Zheng Hu, and Dong Jin. Integrating I/O Time to Virtual Time System for High Fidelity Container-based Network Emulation. ACM SIGSIM-PADS 2022.

3.9 Towards an Open Repository for Reproducible Performance Comparison of Parallel and Distributed Discrete-Event Simulators

Till Köster (Universität Rostock, DE), Philipp Andelfinger (Universität Rostock, DE), and Adelinde M. Uhrmacher (Universität Rostock, DE)

License © Creative Commons BY 4.0 International license

© Till Köster, Philipp Andelfinger, and Adelinde M. Uhrmacher

Main reference Till Köster, Adelinde M. Uhrmacher, Philipp Andelfinger: “Towards an Open Repository for Reproducible Performance Comparison of Parallel and Distributed Discrete-Event Simulators”, in Proc. of the SIGSIM-PADS ’22: SIGSIM Conference on Principles of Advanced Discrete Simulation, Atlanta, GA, USA, June 8 – 10, 2022, pp. 31–32, ACM, 2022.

URL <https://doi.org/10.1145/3518997.3534989>

Performance is one of the core motivations in the field of parallel and distributed simulation. Contributions for new methods and optimizations frequently rely on custom models, parametrizations, and baseline implementations. This makes a direct comparison between methods and approaches difficult. We present our vision and initial steps towards COMPADS, a benchmark model and repository for reproducibly comparing the performance of parallel and distributed simulators and their respective algorithms. COMPADS[1] is short for COMparing Parallel And Distributed Simulators. The first results include a novel deterministic-by-design synthetic benchmark model inspired by PHOLD and La-pdes. The benchmark output is a checksum that attests to the correctness of an implementation and its execution. So far, implementations exist for the simulators ROOT-Sim and ROSS.

References

- 1 Till Köster, Adelinde Uhrmacher, and Philipp Andelfinger. 2022. Towards an Open Repository for Reproducible Performance Comparison of Parallel and Distributed Discrete-Event Simulators. In Proceedings of the 2022 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation (SIGSIM-PADS ’22). Association for Computing Machinery, New York, NY, USA, 31–32. <https://doi.org/10.1145/3518997.3534989>

3.10 Simulation Based Analysis of Social Systems – Models, Data and Policy

Michael Lees (*University of Amsterdam, NL*)

License © Creative Commons BY 4.0 International license
© Michael Lees

Joint work of Michael Lees, Eric Dignum, Andreas Flache, Willem Boterman

Main reference Eric Dignum, Efi Athieniti, Willem Boterman, Andreas Flache, Michael Lees: “Mechanisms for increased school segregation relative to residential segregation: a model-based analysis”, *Comput. Environ. Urban Syst.*, Vol. 93, p. 101772, 2022.

URL <https://doi.org/10.1016/j.compenvurbsys.2022.101772>

The role of modelling and simulation in the scientific domains has had a long history and *computational X* is now a well-established area in Physics, Chemistry, Biology and more. During the last decade with the increase in detailed data, and the development of novel modelling techniques, the application of modelling and simulation has become more commonplace in the social sciences. In some cases, these models provide scientists and policymakers with unique ways to reason about sociological challenges (e.g., Polarization, Segregation and Inequality).

In this talk, I present a sample of current work [1] in which we develop models to understand the process of primary school choice and school segregation with the Municipality of Amsterdam. In the talk, I present an agent-based model where households face residential decisions depending on neighbourhood compositions and make school choices based on distance and school compositions. Using a global sensitivity analysis we demonstrate that the observed excess (the level of school segregation compared to residential segregation) segregation in schools occurs for a wide range of parameters and demonstrate that asymmetric preferences (residential vs. school selection) are not a requirement for excess school segregation.

Using this case study I highlight a number of challenges and opportunities for modelling and simulation within the social sciences. Firstly, the use of models for theory building and testing in social sciences, in particular how simple models with clear assumptions can demonstrate potential causes for social phenomena. Secondly, using techniques from simulation-based inference I demonstrate how novel calibration methods offer promising solutions to calibrate city-scale models of social dynamics using microdata. Finally, I present some real-world cases where a “digital twin” of the city can be used to answer important policy questions for the municipality of Amsterdam and help them statistically estimate the likely outcomes for different interventions.

I conclude the talk by highlighting a number of initiatives [2] within Amsterdam and the Netherlands where new computational infrastructure presents unique opportunities to conduct computational social science and modelling simulation at a city and country-wide scale.

References

- 1 Dignum, E., Athieniti, E., Boterman, W., Flache, A., & Lees, M. (2022). Mechanisms for increased school segregation relative to residential segregation: A model-based analysis. *Computers, Environment and Urban Systems*, 93, 101772.
- 2 Emery, T., Braukmann, R., Wittenberg, M., van Ossenbruggen, J., Siebes, R., & van de Meer, L. (2020). The ODISSEI Portal: Linking Survey and Administrative Data.

3.11 Parallel Simulation – What Worked and What Not

Jason Liu (Florida International University – Miami, US)

License © Creative Commons BY 4.0 International license
© Jason Liu

Parallel discrete-event simulation (PDES) refers to the class of techniques and tools for efficiently running discrete-event simulation on parallel and distributed computing platforms. Applications of PDES include performance modeling and simulation of large systems. Simulation of large systems exacts high computational demand, and successful PDES efforts must be able to cope with both the scale and complexity of the target systems on modern parallel computing architectures. In this talk, we summarize some of our research efforts for developing and applying parallel simulation of various systems.

Traditional PDES research has been largely focused on examining efficient parallel synchronization algorithms and incorporating those in simulation tools for general applications. Previously, we developed DaSSF [1], a simulator implemented in C and C++ that incorporates a composite parallel synchronization algorithm [2]. The algorithm was extended to run on both shared and distributed memory architectures and was implemented in a simulator called MiniSSF [3]. We have also explored the use of scripting languages (including Python, LUA, and Javascript) to simplify model development. The simulator, called Simian, was shown to be able to achieve good, and in some cases, even superior performance by taking advantage of just-in-time compilation [4]. One latest effort was the development of a simulator with full-fledged support for the process-oriented simulation world-view in Python for fast development cycle [5]. Abraham Maslow once said: *“If you have a hammer, everything looks like a nail.”* These and other simulators have been used for simulations of large-scale computer systems and networks, including the Internet [6], mobile ad-hoc networks [7], high-performance computing interconnection networks [8], and parallel files systems [9].

PDES not only enables large-scale simulation of complex systems, but can also be incorporated with real-time and interactive simulation of large systems due to its superior simulation performance. We previously designed and implemented a real-time network simulator to run on parallel and distributed platforms [10]. With real-time simulation, simulated network protocols, such as TCP, can seamlessly interact with real network entities in a hybrid simulation and emulation setting [11]. One can also control and steer the network experiments in real time – for example, by injecting network events and observing the results via a remote dashboard during the live experiment [12].

Many large complex systems feature a huge number of components and processes that may inter-operate across multiple layers of the system hierarchy and at different time granularity. A fine-grained simulation may not be able to scale up to the required size even if PDES could achieve linear speedup. Solving problems in many cases does not require brute force. George Box once said: *“All models are wrong but some are useful.”* In this case, one must be able to use multi-resolution models to balance the trade-off between simulation performance and accuracy. A case in point is the hybrid network traffic modeling, which combines fluid traffic models (e.g., based on ordinary differential equations) and packet-oriented simulation to achieve faster-than-real-time performance for large-scale network simulation [13, 14]. Another example is network simulation and emulation symbiosis. To allow high-fidelity high-performance large-scale network experimentation, one can run a full-scale detailed network model on high-performance parallel computing platforms, and an emulation system, which executes unmodified applications in a virtual machine environment configured to represent the target system. Both systems need to represent the same traffic behavior. We

applied model reduction techniques to scale down the model complexity both in emulation to improve its performance and in data exchange between the two systems to reduce the synchronization and communication overhead [15, 16].

We observe PDES research has evolved by leaps and bounds over the last three decades. Many PDES techniques and tools have matured in various domains, although the community continues to discover new techniques, tools, and applications, many coinciding with the emergence of big data systems, machine learning techniques, and data-driven applications. We predict PDES will continue to play an important, and sometimes indispensable, role in modeling dynamic and complex systems, in many cases though combining with new techniques in order to provide its unique capability in solving problems.

References

- 1 J. Liu, D. Nicol, B. J. Premore, and A. L. Poplawski, "Performance prediction of a parallel simulator," in *Proceedings of the thirteenth workshop on Parallel and distributed simulation*. IEEE Computer Society, 1999, pp. 156–164.
- 2 D. M. Nicol and J. Liu, "Composite synchronization in parallel discrete-event simulation," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 13, no. 5, pp. 433–446, 2002.
- 3 J. Liu and R. Rong, "Hierarchical composite synchronization," in *Principles of Advanced and Distributed Simulation (PADS), 2012 ACM/IEEE/SCS 26th Workshop on*. IEEE, 2012, pp. 3–12.
- 4 N. Santhi, S. Eidenbenz, and J. Liu, "The simian concept: parallel discrete event simulation with interpreted languages and just-in-time compilation," in *Proceedings of the 2015 Winter Simulation Conference*. IEEE Press, 2015, pp. 3013–3024.
- 5 J. Liu, "Simulus: easy breezy simulation in python," in *2020 Winter Simulation Conference (WSC)*. IEEE, 2020, pp. 2329–2340.
- 6 D. M. Nicol, J. Liu, M. Liljenstam, and G. Yan, "Simulation of large scale networks using ssf," in *Simulation Conference, 2003. Proceedings of the 2003 Winter*, vol. 1. IEEE, 2003, pp. 650–657.
- 7 J. Liu, Y. Yuan, D. M. Nicol, R. S. Gray, C. C. Newport, D. Kotz, and L. F. Perrone, "Simulation validation using direct execution of wireless ad-hoc routing protocols," in *Parallel and Distributed Simulation, 2004. PADS 2004. 18th Workshop on*. IEEE, 2004, pp. 7–16.
- 8 K. Ahmed, J. Liu, S. Eidenbenz, and J. Zerr, "Scalable interconnection network models for rapid performance prediction of HPC applications," in *2016 IEEE 18th International Conference on High Performance Computing and Communications (HPCC)*, Dec 2016, pp. 1069–1078.
- 9 M. Erazo, T. Li, J. Liu, S. Eidenbenz *et al.*, "Toward comprehensive and accurate simulation performance prediction of parallel file systems," in *Dependable Systems and Networks (DSN), 2012 42nd Annual IEEE/IFIP International Conference on*. IEEE, 2012, pp. 1–12.
- 10 J. Liu, "A primer for real-time simulation of large-scale networks," in *Proceedings of the 41st Annual Simulation Symposium (ANSS'08)*, 2008, pp. 85–94.
- 11 M. A. Erazo, Y. Li, and J. Liu, "SVEET! a scalable virtualized evaluation environment for TCP," in *Proceedings of the 5th International Conference on Testbeds and Research Infrastructures for the Development of Networks & Communities and Workshops (TRIDENTCOM'09)*, 2009, pp. 1–10.
- 12 N. Van Vorst, M. Erazo, and J. Liu, "PrimoGENI: Integrating real-time network simulation and emulation in GENI," in *Proceedings of the 2011 IEEE Workshop on Principles of Advanced and Distributed Simulation (PADS)*, 2011, pp. 1–9.
- 13 J. Liu, "Packet-level integration of fluid TCP models in real-time network simulation," in *Proceedings of the 2006 Winter Simulation Conference (WSC'06)*, December 2006, pp. 2162–2169.

- 14 J. Liu, “Parallel simulation of hybrid network traffic models,” in *Proceedings of the 21st Workshop on Principles of Advanced and Distributed Simulation (PADS’07)*, June 2007, pp. 141–151.
- 15 M. A. Erazo, R. Rong, and J. Liu, “Symbiotic network simulation and emulation,” *ACM Trans. Model. Comput. Simul.*, vol. 26, no. 1, pp. 2:1–2:25, 2015.
- 16 R. Rong and J. Liu, “Distributed mininet with symbiosis,” in *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1–6.

3.12 Simulation at the Edge

Margaret Loper (Georgia Institute of Technology – Atlanta, US)

License © Creative Commons BY 4.0 International license
© Margaret Loper

Mobile networks evolve roughly every ten years, each generation bringing new services and capabilities. By 2030, 6G will bring more capable, intelligent, reliable, scalable, and power-efficient communications. This will enable new applications such as holographic telepresence, collaborative autonomous driving and personalized body area networks [3]. The 6G era will help realize the hyper-connected world of people, data and things – the Internet of Everything (IoE). It will also enable a focus on small data, generated by the plethora of edge devices embedded in our everyday world. The IoE will create a need for efficient processing of massive amounts of small data and edge intelligence, a process where data is collected, analyzed, and insights produced near the end user. The promise of 6G and the emergence of edge intelligence will provide users with actionable real-time information. An extension of this concept is to also provide users with actionable real-time predictions. The intersection of edge computing, sensor networks, artificial intelligence and online predictive simulations enable a new vision called Simulation at the Edge.

Simulation Paradigms. Paradigms for sensor-driven simulations first emerged in the mid-2000s with concepts like symbiotic, ad hoc and data driven adaptive simulation systems. Symbiotic simulation is a paradigm which refers to a close relationship between a simulation system and a physical system. It was defined at the Dagstuhl Seminar on Grand Challenges for Modeling and Simulation in 2002 [5]. The simulation system benefits from real-time measurements about the physical system which are provided by sensors, and the physical system may benefit from decisions made by the simulation [1]. An important concept in symbiotic simulation is the “what-if” analysis, where multiple experiments investigate alternative scenarios based on data from the physical system. Symbiotic simulation does not refer to a specific type of simulation (e.g., real-time, discrete event), it is an umbrella term which refers to independent simulations employed to analyze alternative scenarios regarding a physical system.

The Dynamic Data-Driven Application Systems (DDDAS) concept utilizes online data to drive simulation computations, and the results are then used to optimize the system or adapt the measurement process [2]. For example, live sensor data and analytics can be used to construct or infer the current state of a system and faster-than-real-time simulation can then be used to project the system’s future state. Also, simulation can be used to control an operational system, e.g., data from a real system are fed directly into the simulation model which analyzes alternate options and produces recommended courses of action.

An ad hoc distributed simulation is a collection of autonomous on-line simulations, each modeling some portion of a larger physical system, that are brought together to predict future states of the overall system [4]. In a conventional distributed simulation, the system being modeled is partitioned into non-overlapping elements (e.g., geographic regions) in a top-down fashion. By contrast, ad hoc simulations are constructed bottom-up, resulting in multiple simulators modeling common, overlapping portions of the physical system. In other words, it is constructed in an “ad hoc” fashion, in much the same way a collection of mobile radios join together to form an ad hoc wireless network. Ad hoc simulations are on-line simulation programs, meaning they are able to capture the current state of the system through measurement, and then execute forward as rapidly as possible to project a future state of the system.

Edge Intelligence and Simulation. These simulation paradigms have some similarities in their approach, but different levels of success in accomplishing their vision. For example, they are all data-driven using real-time sensor input and the simulations run faster-than-real-time to predict future state, look at “what if” scenarios or steer measurements. Where they differ include their interaction or feedback with the physical system, as well as the number of simulations in use. For example, DDDAS can be tightly coupled with the real system, steering the measurement process, where symbiotic and ad hoc do not. Further, ad hoc uses more than one simulation to model different parts of a system, whereas symbiotic and DDDAS are focused on a single, central simulation. DDDAS has been quite successful modeling large scale structural systems, urban water systems and transportation systems; and symbiotic simulations have had success in semiconductor manufacturing, business process optimization, and control of unmanned aerial vehicles. While research into ad hoc simulation has been limited to transportation and queueing systems [6], it has so much more potential in the era of 6G and IoE.

The proliferation of mobile computing and IoE, edge computing is an emerging paradigm that pushes computing tasks and services from the network core to the network edge. Edge computing is widely recognized as a promising solution for processing the “zillion” bytes of data generated by IoE devices [7]. It has also attracted attention for its promise to reduce latency, save bandwidth, improve availability, and to keep data secure. At the same time, a proliferation of AI algorithms and models which accelerate the deployment of intelligence in edge devices has emerged. These trends, called Edge Intelligence, can power the evolution of ad hoc simulation to Simulation at the Edge.

Embedding simulations within edge intelligence brings the simulation closer to the data, lessening the need to aggregate sensor data in order to reduce communication bandwidth requirements. It also has the potential to be more resilient to failures, especially communication failures, as portions of the system could be managed under local control. The application of intelligent edge devices embedded with predictive simulations are varied and diverse. They could be used to monitor transportation systems, or rerouting vehicle traffic after a severe accident; track the spread of wild fires, floods, and pollution; optimize emergency responses, such as evacuations during floods or tornadoes; provide self-optimizing communication networks, by reconfiguring the physical network to improve performance and avoid bottlenecks; or respond to breakdowns within a manufacturing system.

Despite the decades of research in symbiotic, DDDAS and ad hoc simulation, research challenges remain. These include: compact representation of system state, fault tolerant and robust systems, multi-resolution modeling, and automatic validation [5]. Like its predecessor, distributed Simulation at the Edge raise a number of intriguing questions. Can such a distributed simulation make sufficiently accurate predictions of future system states to be useful? Can they incorporate new information and revise projections more rapidly and/or effectively than conventional approaches, e.g., global, centralized simulations? How would Simulation at the Edge be organized and operate? The power of edge intelligence and 6G could be the catalyst to help answer these questions.

References

- 1 Heiko Aydt, Stephen John Turner, Wentong Cai, and Malcolm Yoke Hean Low. Research issues in symbiotic simulation. In *Proceedings of the 2009 winter simulation conference (WSC)*, pages 1213–1222. IEEE, 2009.
- 2 F Darema, M Rotea, M Goldberg, DH Newlon, JC Cherniavsky, JE Figueroa, JE Hudson, C Friedman, P Lyster, and R Bohn. Dddas: dynamic data driven applications systems. URL: <http://www.nsf.gov/pubs/2005/nsf05570/nsf05570.htm> (Accessed 20 September 2013), 2005.
- 3 Chamitha De Alwis, Anshuman Kalla, Quoc-Viet Pham, Pardeep Kumar, Kapal Dev, Won-Joo Hwang, and Madhusanka Liyanage. Survey on 6g frontiers: Trends, applications, requirements, technologies and future research. *IEEE Open Journal of the Communications Society*, 2:836–886, 2021.
- 4 Richard Fujimoto, Michael Hunter, Jason Sirichoke, Mahesh Palekar, Hoe Kim, and Wonho Suh. Ad hoc distributed simulations. In *21st International Workshop on Principles of Advanced and Distributed Simulation (PADS'07)*, pages 15–24. IEEE, 2007.
- 5 Richard Fujimoto, WH Lunceford Jr, Ernst H Page, and Adelinde Uhrmacher. Grand challenges for modelling and simulation (dagstuhl seminar 02351). Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2002.
- 6 Ya-Lin Huang, Christos Alexopoulos, Michael Hunter, and Richard M Fujimoto. Ad hoc distributed simulation methodology for open queueing networks. *Simulation*, 88(7):784–800, 2012.
- 7 Zhi Zhou, Xu Chen, En Li, Liekang Zeng, Ke Luo, and Junshan Zhang. Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proceedings of the IEEE*, 107(8):1738–1762, 2019.

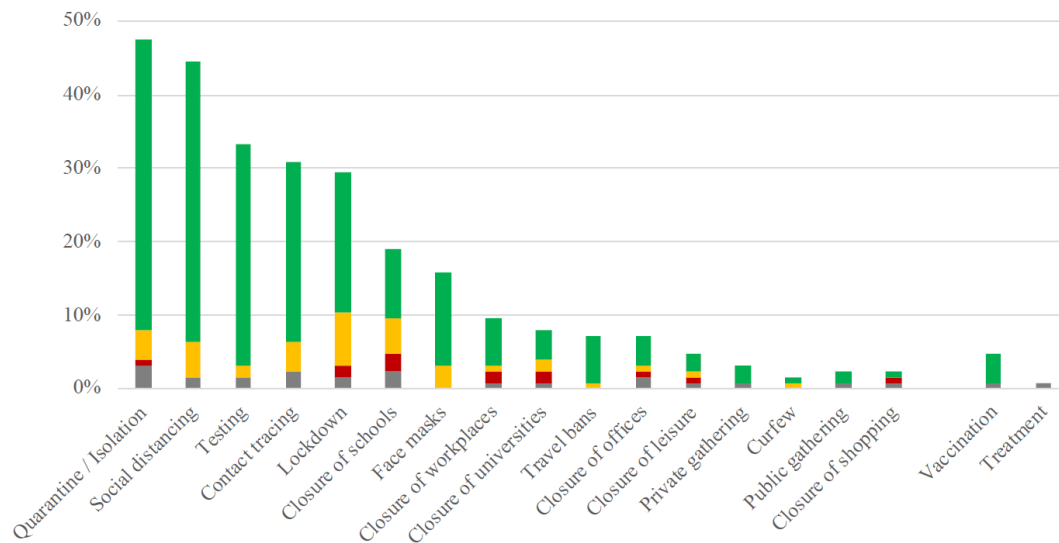
3.13 Bottlenecks of using simulation for policy making

Fabian Lorig (Malmö University, SE)

License © Creative Commons BY 4.0 International license
© Fabian Lorig

Computer simulation has become an established method for assessing uncertainty by conducting what-if analyses across a variety of disciplines and domains. It allows us to observe the behavior of a system under different circumstances and to investigate the potential consequences of different actions or interventions without actually interfering with the system we want to analyze. During the Covid-19 pandemic, for instance, we could see that a great number of models was developed already in the first months after the initial outbreak of the crisis [1]. Figure 5 shows that models were developed for all different kinds of potential interventions and to investigate how they possibly might affect transmission processes and the overall dynamics of the pandemic.

In practice, however, many researchers experienced that their models and the results generated by their simulation studies were not considered by policy makers when deciding upon which interventions to pursue and implement. In addition to this, other shortcomings in the development and application of simulation models and the conducting of simulation studies could be observed. Most researchers, for instance, did not reuse existing simulations and instead have chosen to start developing new models and to conduct new studies, which affects responsiveness in crisis situations. Potential reasons might be the limited availability of models, ambiguous documentations of the models and previously conducted studies, or



■ **Figure 5** Number of models that support the simulation of non-pharmaceutical and pharmaceutical interventions [1].

a general lack of trust in simulations studies conducted by others. This leads to a series of questions regarding why simulations sometimes fail to support what they were actually intended for namely to provide valuable insights and to facilitate decision making processes.

In this regard, when we discuss how to improve the effectiveness and sustainability of simulation studies, it seems that there might be different perspectives on this very issue. A simulation that we as developers and modellers consider highly insightful, comprehensive, and trustworthy might not be convincing, helpful, or plausible for the decision makers the study was intended for. Therefore, when discussing the effectiveness and sustainability of simulation studies, we should not forget about the stakeholders that might be involved in the process of a simulation study. How can we make sure to develop and generate what decision makers actually need? Why do we prefer to conduct own (new) studies instead of reusing the models and results of others? And how can we improve the rigor and trustworthiness of our studies?

References

- 1 Lorig, F., Johansson, E., & Davidsson, P. *Agent-based Social Simulation of the COVID-19 Pandemic: A Systematic Review*. JASSS: Journal of Artificial Societies and Social Simulation, 24(3).

3.14 Supporting Transparent Simulation Studies: The Role of Provenance Information

Bertram Ludäscher (University of Illinois at Urbana-Champaign, US)

License © Creative Commons BY 4.0 International license
© Bertram Ludäscher

Computational science experiments and simulation studies often require significant computational and human resources, making it impractical or impossible to repeat (i.e., reimplement and/or re-execute) these studies for reproducibility purposes. Transparency, on the other

hand, is arguably always desirable [8] and can be achieved by revealing and laying open the assumptions, general approach, computational methods, codes, parameter settings, runtime environments, and – last not least – the data sources used to conduct a study. Regarding the latter, precisely identifying the correct subsets of data that were used in a computational experiment is already a difficult challenge [3, 12].

Provenance (a.k.a. lineage) information captures the origin and processing history of data products [5, 7, 6], i.e., it is a form of metadata that provides the technical means to support transparency [10]. While the existing W3C standard PROV [11] provides a minimal baseline for exchanging provenance information, domain-specific extensions need to be developed by and for the community to capture more of the application context, the semantics of data and parameters, and assumptions of simulation models and scientific workflows [13, 16].

Transparent research objects [1, 15, 4, 10, 9, 14], with their data, computational, and provenance artifacts, and research papers (which tell the “science story” of a paper) will continue to co-evolve and should ultimately converge towards open, transparent, and reproducible research tales [2]. – *Declaremus et calculemus!*


References

- 1 Bechhofer, S., De Roure, D., Gamble, M., Goble, C., Buchan, I.: Research Objects: Towards Exchange and Reuse of Digital Knowledge. *Nature Precedings* (Jul 2010). <https://doi.org/10.1038/npre.2010.4626.1>
- 2 Brinckman, A., Chard, K., Gaffney, N., Hategan, M., Jones, M.B., Kowalik, K., Kulasekaran, S., Ludäscher, B., Mecum, B.D., Nabrzycki, J., Stodden, V., Taylor, I.J., Turk, M.J., Turner, K.: Computing environments for reproducibility: Capturing the “Whole Tale”. *Future Gener. Comput. Syst.* **94**, 854–867 (2019). <https://doi.org/10.1016/j.future.2017.12.029>
- 3 Buneman, P., Christie, G., Davies, J.A., Dimitrellou, R., Harding, S.D., Pawson, A.J., Sharman, J.L., Wu, Y.: Why data citation isn’t working, and what to do about it. *Database* (Jan 2020). <https://doi.org/10.1093/databa/baaa022>
- 4 Chard, K., Gaffney, N., Jones, M.B., Kowalik, K., Ludäscher, B., McPhillips, T., Nabrzycki, J., Stodden, V., Taylor, I., Thelen, T., Turk, M.J., Willis, C.: Application of BagIt-Serialized Research Object Bundles for Packaging and Re-Execution of Computational Analyses. *15th Intl. Conf. on eScience* pp. 514–521 (Sep 2019). <https://doi.org/10.1109/eScience.2019.00068>
- 5 Freire, J., Koop, D., Santos, E., Silva, C.T.: Provenance for Computational Tasks: A Survey. *Computing in Science Engineering* **10**(3), 11–21 (May 2008). <https://doi.org/10.1109/MCSE.2008.79>
- 6 Herschel, M., Diestelkämper, R., Lahmar, H.B.: A survey on provenance: What for? What form? What from? *The VLDB Journal* **26**(6), 881–906 (Dec 2017). <https://doi.org/10.1007/s00778-017-0486-1>
- 7 Ludäscher, B.: A Brief Tour Through Provenance in Scientific Workflows and Databases. In: Lemieux, V.L. (ed.) *Building Trust in Information*, pp. 103–126. Springer Proceedings in Business and Economics, Springer International Publishing (2016). https://doi.org/10.1007/978-3-319-40226-0_7
- 8 McPhillips, T., Ludäscher, B., Goble, C., Willis, C., Bowers, S.: Workshop Report – T7 Workshop on Provenance for Transparent Research. Tech. Rep., Zenodo, Provenance Week 2021 (Aug 2021). <https://doi.org/10.5281/zenodo.5301583>
- 9 McPhillips, T.M., Thelen, T., Willis, C., Kowalik, K., Jones, M.B., Ludäscher, B.: CPR-A Comprehensible Provenance Record for Verification Workflows in Whole Tale. *Provenance and Annotation of Data and Processes*. LNCS, Springer International Publishing, (2021). https://doi.org/10.1007/978-3-030-80960-7_23

- 10 McPhillips, T.M., Willis, C., Gryk, M.R., Corrales, S.N., Ludäscher, B.: Reproducibility by other means: Transparent research objects. 15th Intl. Conf. on eScience pp. 502–509. (Sep 2019). <https://doi.org/10.1109/eScience.2019.00066>
- 11 Moreau, L., Groth, P., Cheney, J., Lebo, T., Miles, S.: The rationale of PROV. Web Semantics: Science, Services and Agents on the World Wide Web **35**(Part 4), 235–257 (Dec 2015). <https://doi.org/10.1016/j.websem.2015.04.001>
- 12 Rauber, A., Gößwein, B., Zwölf, C.M., Schubert, C., Wörister, F., Duncan, J., Flicker, K., Zettsu, K., Meixner, K., McIntosh, L.D., Jenkyns, R., Pröll, S., Miksa, T., Parsons, M.A.: Precisely and Persistently Identifying and Citing Arbitrary Subsets of Dynamic Data. Harvard Data Science Review **3**(4) (Nov 2021). <https://doi.org/10.1162/99608f92.be565013>
- 13 Ruschinski, A., Wilsdorf, P., Dombrowsky, M., Uhrmacher, A.M.: Capturing and Reporting Provenance Information of Simulation Studies Based on an Artifact-Based Workflow Approach.: Proc. ACM SIGSIM Conference on Principles of Advanced Discrete Simulation. pp. 185–196. (2019). <https://doi.org/10.1145/3316480.3325514>
- 14 Soiland-Reyes, S., Sefton, P., Crosas, M., Castro, L.J., Coppens, F., Fernández, J.M., Garijo, D., Grüning, B., La Rosa, M., Leo, S., Ó Carragáin, E., Portier, M., Trisovic, A., RO-Crate Community, Groth, P., Goble, C.: Packaging research artefacts with RO-Crate. Data Science **5**(2), 97–138 (Jan 2022). <https://doi.org/10.3233/DS-210053>
- 15 Ton That, D.H., Fils, G., Yuan, Z., Malik, T.: Sciunits: Reusable Research Objects. 13th Intl. Conf. on E-Science (eScience). pp. 374–383 (Oct 2017). <https://doi.org/10.1109/eScience.2017.51>
- 16 Wilsdorf, P., Wolpers, A., Hilton, J., Haack, F., Uhrmacher, A.M.: Automatic Reuse, Adaptation, and Execution of Simulation Experiments via Provenance Patterns. ACM Transactions on Modeling and Computer Simulation (Sep 2022). <https://doi.org/10.1145/3564928>

3.15 Interactive Visual Analysis for Simulation

Kresimir Matkovic (VRVis – Wien, AT)

License  Creative Commons BY 4.0 International license
© Kresimir Matkovic

Visualization and interactive visual analysis have been used to explore and analyze simulation data for a long time. At the beginning, results of single run simulation have been visualized. With advancement of storage and computing technologies, the models became more and more complex and, at the same time, ensemble summation or simulation experiments – multiple simulation runs using variations of the same model – became possible. In case of computational fluid dynamics simulation the ensembles contain a relatively small number of members. Due to large and complex spatial-temporal data and long computing time it is not feasible to compute many models. Wang et al. [6] provide a recent survey on visualization of such ensemble simulations. In cases where simulation can be computed fast it is possible to compute hundreds, thousands or even tens of thousands of simulation runs. Matković et al. [4] provide an overview of such approaches.

The main idea here is to deploy coordinated multiple views that visualize multidimensional parameter space and complex simulation results at once. Interaction is used to support exploration and analysis. The user can brush, i.e. interactively select a subset of data in any view, and the items correspond to the brushed subset will be highlighted in all views. The idea functions if the number of parameters does not exceed half a dozen, or so. A

parameter space of a higher dimensionality requires many runs in order to be sufficiently covered. Interactive simulating ensemble steering can be used in such cases. A kind of automatic analysis can be integrated in order to guide the expert.

In the case of multi-model simulation, there is little available support from the visualization community. Simultaneous exploration of simulation results computed with models of different levels of detail remains a challenge for the visualization. Large data, hierarchical data structure, and a need for fluent switching of context depending on level (and maybe even providing the overall context across the levels at the same time) is subject of future research. However, based on successful deployment of interactive visualization in the simulation community up to know, it is plausible to reason that interactive visualization can become a key interface in navigating in complicated multi-level models and simulation results, and a great support in comprehending underlying phenomena for different stake-holders. Scalability often represents a serious problem in visualization for simulation. In case of multi-model simulation, scalability will also represent a challenge. Finally, provenance tracking will gain in importance, as user actions across multiple levels need to be stored and recalled on demand. Finally, Dimara and Stasko [1] recently reported on the missing link between user tasks in visualisation taxonomies (e.g., sensemaking) and the high-level task of decision-making. One of the key causes of this mismatch is a lack of interdisciplinary approaches. A close collaboration with simulation experts represents a valid approach to establishing the missing link.


In order to support future requirements from the simulation community we need further advances in interactive visualization. We expect the novel approaches to be inspired by several directions of interactive visualization, depending on the simulation methods and corresponding models, as well as on the identified explorative tasks. Besides coordinated multiple views, the promising pillars of future research are certainly focus and context techniques [3], comparative visualization [2], tree [5], networks and graphs visualization, as well as plethora of existing visualization for simulation methods and provenance tracking.

References

- 1 Evanthia Dimara and John Stasko. A critical reflection on visualization research: Where do decision making tasks hide? *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1128–1138, 2022.
- 2 Michael Gleicher, Danielle Albers, Rick Walker, Ilir Jusufi, Charles D. Hansen, and Jonathan C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, oct 2011.
- 3 Helwig Hauser. Generalizing focus+context visualization. In Georges-Pierre Bonneau, Thomas Ertl, and Gregory M. Nielson, editors, *Scientific Visualization: The Visual Extraction of Knowledge from Data*, pages 305–327, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.
- 4 Krešimir Matković, Denis Gračanin, and Helwig Hauser. Visual analytics for simulation ensembles. In *Proceedings of the 2018 Winter Simulation Conference, WSC '18*, page 321–335. IEEE Press, 2018.
- 5 Tamara Munzner, François Guimbretière, Serdar Tasiran, Li Zhang, and Yunhong Zhou. Treejuxtaposer: Scalable tree comparison using focus+context with guaranteed visibility. *ACM Trans. Graph.*, 22(3):453–462, jul 2003.
- 6 Junpeng Wang, Subhashis Hazarika, Cheng Li, and Han-Wei Shen. Visualization and visual analysis of ensemble data: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 25(9):2853–2872, 2019.

3.16 A logic-based approach to reason about large-scale spatially-distributed systems

Laura Nenzi (University of Trieste, IT)

License  Creative Commons BY 4.0 International license
© Laura Nenzi

From the reliability in a wireless sensor network to the formation of traffic jams, spatio-temporal patterns are key in understanding how complex behaviors can emerge in a network of locally interacting dynamical systems. One of the most important and intriguing questions is how to describe such behaviors in a formal and human-understandable specification language. A possible approach consists in using formal methods and in particular logic languages. In this talk, we show how a logic specification can be used to specify and analyse complex behaviours of large-scale spatially-distributed systems. Furthermore, we briefly show how we can use the logic for parameter synthesis, anomaly detection and the automatic feature extraction from spatio-temporal data.

3.17 On the Attractiveness of Speculative PDES: Challenges and Pitfalls

Alessandro Pellegrini (University of Rome “Tor Vergata”, IT)

License  Creative Commons BY 4.0 International license
© Alessandro Pellegrini

Introduction: The last 40 years have witnessed an explosion of methodologies and techniques related to speculative PDES [8]. Looking at the history of this research line, while performance aspects have always been paramount, each decade has seen its hot points. Roughly speaking, in the first decade, algorithmic solutions were proposed that enabled significant speedups with relatively little effort (see, e.g., [12]). The second decade began to propose solutions for the adaptivity of particular simulation aspects (see, e.g., [6, 10]). The third decade focused on effectively supporting models’ programming by hiding the complexity of Speculative PDES (see, e.g., [23]). The fourth decade showed the technique’s robustness even on new hardware/software architectures [2, 13, 21].

We can conclude that Speculative PDES brings together effective solutions and methodologies for executing many classes of models, offering significant speedups and making tractable problems that otherwise would not be.

However, these results do not appear to be generally exploited. Market penetration of these techniques is severely limited and often, with some exceptions, confined to the academic sphere. The research community has already shown that there are various scenarios in which exploiting speculative PDES can bring benefits (e.g. Agent-Based Models [1], Spiking Neural Networks, SNN [20, 16], Traffic [11] and Hardware Architecture Simulations [25]). However, the most widely used simulation software does not consider this methodology at all. To give a few examples, in traffic simulation, SUMO [3] is single-threaded; in SNN simulations, most simulators are sequential (Brian [22]) or, if parallel, are based on conservative/time stepped algorithms (NEST [9]). Architecture simulators, such as gem5 [4], are also sequential.

Over the past few years, we have been wondering what the reasons might be for this poor uptake of methodologies that the scientific community has proven to be effective in many fields. From these motivations, we have tried to identify some challenges that we believe are relevant to make Speculative PDES more attractive to other fields of research and industry.

Challenges. In general, using speculative PDES to support model execution is difficult. However long and deep the problem has been studied, it is still not wholly possible to hide the complexity of algorithms such as Time Warp from model developers. While some solutions allow the technical complexities to be hidden, it is still true that a model not explicitly developed for Time Warp may provide inadequate performance. Work on self-tuning and self-optimization has shown that it is possible to improve overall performance by optimizing specific parts of the algorithm (e.g., the GVT computation). However, a significant challenge could be to build an optimization approach capable of mixing different, more or less optimistic algorithms that can provide no worse performance than a sequential simulation.

Attractiveness to other domains must necessarily come through real-world models, which has also recently been recognized as vital [19]. The scientific community has often used synthetic benchmarks (see, for example, [7, 17]) to study Time Warp behaviour. However, scientists in other fields or industrial settings may be unable to map these results to their use cases. An important challenge could be to design and implement a real-world benchmark suite to show how effective speculative PDES can be in these areas. Approaches of this kind have already been followed in other research fields (see [15]) and could be successful for more widespread adoption of speculative PDES.

Another problem that makes using Speculative PDES difficult is that among the various implementations built by research groups (see, e.g., [5, 18, 14]), there is no uniformity of interface. PDES researchers have done much work on abstract model representations (e.g., [24]), but an agreement on interfaces among developers of runtime environments would also be desirable. Moreover, this would greatly help in the cross-fertilization of approaches used in the runtimes and would more easily allow experimental studies to be carried out, also bringing benefit to the scientific community dealing with PDES.

Another problem we encountered is related to the quality of support libraries for PDES runtime environments. Many tools rely on MPI interfaces for message exchange, which is a correct choice to support deployment on supercomputers. Unfortunately, it has been observed that the most modern MPI features are not widely used in the High Performance Computing world, especially those related to asynchronous execution. The result is that the most common implementations often suffer from correctness bugs¹. Moreover, MPI libraries do not consider highly concurrent and asynchronous usage patterns. Thus, even if the implementations are correct, they do not hold up to the message rate, especially if the runtime is highly optimised. This makes the work done on runtime environments unusable by researchers in other fields. The experience gained in developing runtime environments for speculative simulations could be used to create extremely fast message exchange libraries geared toward speculative PDES simulations. In this way, the usability of research results could be significantly improved.

References

- 1 Sameera Abar, Georgios K Theodoropoulos, Pierre Lemarinier, and Gregory M P O'Hare. Agent based modelling and simulation tools: A review of the state-of-art software. *Computer Science Review*, 24:13–33, 2017.
- 2 Peter D Barnes, Christopher D Carothers, David R Jefferson, and Justin M LaPre. Warp speed: executing time warp on 1,966,080 cores. In *Proceedings of the 1st ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, SIGSIM PADS '13, pages 327–336, New York, NY, USA, May 2013. ACM.

¹ Some issues that we identified while developing a high-performance PDES runtime [18] can be found following the links [here](#), [here](#), and [here](#).

- 3 Michael Behrisch, Laura Bieker, Jakob Erdmann, and Daniel Krajzewicz. SUMO—simulation of urban mobility: an overview. In *Proceedings of SIMUL 2011, The Third International Conference on Advances in System Simulation*. elib.dlr.de, 2011.
- 4 Nathan Binkert, Bradford Beckmann, Gabriel Black, Steven K Reinhardt, Ali Saidi, Arkaprava Basu, Joel Hestness, Derek R Hower, Tushar Krishna, Somayeh Sardashti, Rathijit Sen, Korey Sewell, Muhammad Shoaib, Nilay Vaish, Mark D Hill, and David A Wood. The gem5 simulator. *SIGARCH Comput. Archit. News*, 39(2):1–7, August 2011.
- 5 Christopher D Carothers, David Bauer, and Shawn Pearce. ROSS: A high-performance, low-memory, modular time warp system. *Journal of parallel and distributed computing*, 62(11):1648–1669, November 2002.
- 6 Josef Fleischmann and Philip A Wilsey. Comparative analysis of periodic state saving techniques in time warp simulators. In *Proceedings of the 9th workshop on Parallel and Distributed Simulation*, PADS '95, pages 50–58, Piscataway, NJ, USA, July 1995. IEEE Computer Society.
- 7 Richard M Fujimoto. Performance of time warp under synthetic workloads. In David Nicol, editor, *Proceedings of the SCS Multiconference on Distributed Simulation*, pages 23–28, San Diego, CA, USA, 1990. Society for Computer Simulation International.
- 8 Richard M Fujimoto, Rajive Bagrodia, Randal E Bryant, K Mani Chandy, David Jefferson, Jayadev Misra, David Nicol, and Brian Unger. Parallel discrete event simulation: The making of a field. In *2017 Winter Simulation Conference (WSC)*, pages 262–291, December 2017.
- 9 Marc-Oliver Gewaltig and Markus Diesmann. *NEST (NEural Simulation Tool)*, volume 2, chapter 4. Scholarpedia, 2007.
- 10 D W Glazer and Carl Tropper. On process migration and load balancing in time warp. *IEEE Transactions on Parallel and Distributed Systems*, 4:318–327, 1993.
- 11 Masatoshi Hanai, Toyotaro Suzumura, Georgios Theodoropoulos, and Kalyan S Perumalla. Exact-Differential Large-Scale traffic simulation. In *Proceedings of the 3rd ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, SIGSIM PADS '15, pages 271–280, New York, NY, USA, June 2015. Association for Computing Machinery.
- 12 David R Jefferson. Virtual time. *ACM Transactions on Programming Languages and Systems*, 7(3):404–425, July 1985.
- 13 Xihu Liu and Philipp Andelfinger. Time warp on the GPU: Design and assessment. In *Proceedings of the 2017 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, SIGSIM-PADS '17, pages 109–120, New York, NY, USA, May 2017. ACM.
- 14 D E Martin, T J McBrayer, and P A Wilsey. WARPED: a time warp simulation kernel for analysis and application development. In *Proceedings of the 29th Hawaii International Conference on System Sciences*, volume 1 of *HICSS*, pages 383–386 vol.1, Piscataway, NJ, USA, January 1996. IEEE Computer Society.
- 15 Chi Cao Minh, Jaewoong Chung, Christos Kozyrakis, and Kunle Olukotun. STAMP: Stanford transactional applications for Multi-Processing. In *2008 IEEE International Symposium on Workload Characterization*, pages 35–46, September 2008.
- 16 Quang Anh Pham Nguyen, Philipp Andelfinger, Wentong Cai, and Alois Knoll. Transitioning spiking neural network simulators to heterogeneous hardware. In *Proceedings of the 2019 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, SIGSIM-PADS, pages 115–126, New York, NY, USA, May 2019. ACM.
- 17 Eunjung Park, Stephan Eidenbenz, Nandakishore Santhi, Guillaume Chapuis, and Bradley Settlemyer. Parameterized benchmarking of parallel discrete event simulation systems: Communication, computation, and memory. In *2015 Winter Simulation Conference (WSC)*, pages 2836–2847. ieeexplore.ieee.org, December 2015.

- 18 Alessandro Pellegrini, Roberto Vitali, and Francesco Quaglia. The ROME OpTimistic simulator: Core internals and programming model. In *Proceedings of the 4th International ICST Conference on Simulation Tools and Techniques*, SIMUTOOLS, pages 96–98, Brussels, Belgium, April 2012. ICST.
- 19 Kalyan Perumalla, Maximilian Bremer, Kevin Brown, Cy Chan, Stephan Eidenbenz, K Scott Hemmert, Adolfo Hoisie, Benjamin Newton, James Nutaro, Tomas Opielstrup, and Others. Computer science research needs for parallel discrete event simulation (PDES). Technical report, USDOE Office of Science (SC)(United States), 2022.
- 20 Adriano Pimpini, Andrea Piccione, Bruno Ciciani, and Alessandro Pellegrini. Speculative distributed simulation of very large spiking neural networks. In *Proceedings of the 2022 SIGSIM Conference on Principles of Advanced Discrete Simulation*, SIGSIM PADS, New York, NY, USA, 2022. ACM.
- 21 Shafiqur Rahman, Nael Abu-Ghazaleh, and Walid Najjar. PDES-A: a parallel discrete event simulation accelerator for FPGAs. In *Proceedings of the 2017 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, SIGSIM-PADS '17, pages 133–144, New York, NY, USA, May 2017. Association for Computing Machinery.
- 22 Marcel Stimberg, Romain Brette, and Dan F M Goodman. Brian 2, an intuitive and efficient neural simulator. *eLife*, 8(e47314):e47314, August 2019.
- 23 Roberto Toccaceli and Francesco Quaglia. DyMeLoR: Dynamic memory logger and restorer library for optimistic simulation objects with generic memory layout. In *Proceedings of the 22nd Workshop on Principles of Advanced and Distributed Simulation*, PADS, pages 163–172, Piscataway, NJ, USA, June 2008. IEEE Computer Society.
- 24 Bernard P Zeigler, Tag Gon Kim, and Herbert Praehofer. *Theory of Modeling and Simulation*. Academic Press, January 2000.
- 25 L Zhu, G Chen, B K Szymanski, C Tropper, and T Zhang. Parallel logic simulation of million-gate VLSI circuits. In *13th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems*, pages 521–524. ieeexplore.ieee.org, September 2005.

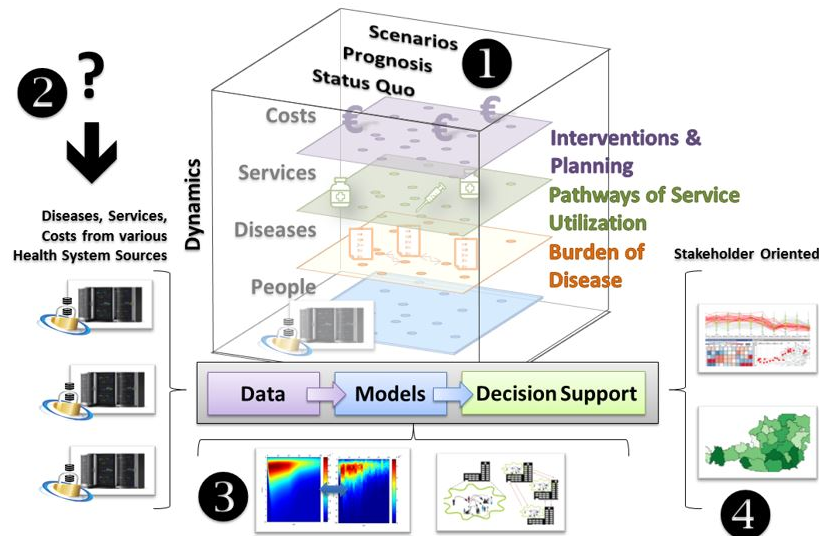
3.18 Methods for Integrated Simulation – from Data Acquisition to Decision Support

Niki Popper (*Technische Universität Wien, AT*)

License © Creative Commons BY 4.0 International license
© Niki Popper

Diversity and Heterogeneity of man-made systems increase rapidly and so do the costs spent for them. Measuring efficiency and effectiveness becomes more and more elaborate but is an urgent need. Development of new methods, models and technologies is needed to support analysing, planning and controlling. The quantity and quality of available data strongly increase and therefore facilitate the description and analysis of systems like health care. Bringing together necessary technologies is an enormous challenge.

Data-based Demographic models have to be combined with models for the spread of diseases. Dynamic modelling concepts must be parametrised with complex data sets from various sources. For system simulation, an important aspect is the possibility to implement changes inside the system, like interventions within the computer model, and analysing their effects. As a recent example see Covid-19 Modelling at TU Wien [1]).



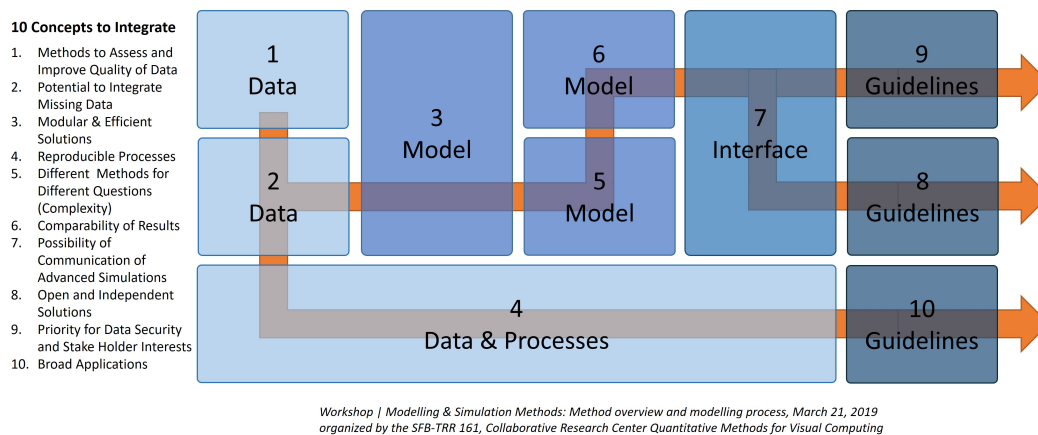
■ **Figure 6** Schematic overview of DEXHELPP infrastructure and process.

On basis of experiences of the Austrian DEXHELPP Competence Centre for Decision Support in Health Policy and Planning [2], which started in 2014 a concept was developed how large, interdisciplinary teams can handle these complex processes in the future and what are similarities and differences between health systems and other complex man-made DEXHELPP developed. DEXHELPP developed an innovative research infrastructure with (1) a flexible virtualised health system, (2) methods to cope with data, (3) an adaptive analysis and simulation methods pool and (4) stakeholder-oriented interfaces to enable researchers and other stakeholders to share data and methods for research and decision making. (see Figure 6).

“Ten Concepts to Integrate” were identified and first presented at Invited Talks at the University of Rostock [3] and University of Stuttgart [4]. The ten concepts are (1) Methods to Assess and Improve Quality of Data, (2) Potential to Integrate Missing Data, (3) Modular & Efficient Solutions, (4) Reproducible Processes, (5) Different Methods for Different Research Questions, (6) Comparability of Results, (7) Possibility of Communication of Advanced Simulations, (8) Open and Independent Solutions, (9) Priority for Data Security and Stakeholder Interests and (10) Broad Applications like Health System, Energy, Industry, Energy, Mobility and Infrastructure Planning and Usage (see Figure 7).

“Methods to Assess and Improve data Quality of Data” and “Potential to Integrate Missing Data” (summarized as “Data” in Figure 7) offer the possibility to find wrong data and correct them, ideally also during the simulation process. For this purpose, methods of interactive visualisation and statistics are used, among others, to preprocess data collected unilaterally, e.g. sensor data or reimbursement data or to link data that are unstructured or have different structures without direct linkage, e.g. [5].

To develop “Modular & Efficient Solutions” using “Different Methods for Different Research Questions” and maintain “Comparability of Results” (summarised as “Model” in Figure 7) includes sustainable, modular models that can be quickly adapted to new problems and concepts for comparing, combining and linking models (qualitatively and quantitatively) in order to demonstrate the benefits and limitations of the concrete model. As an example



■ **Figure 7** 10 Concepts to integrate into simulation processes.

the DEXHELPP Population Model GEPOC (Generic Population Concept) was used for Covid-19 Modelling in order to guarantee rapid and quality-assured implementation, e.g. [6] also to compare it in the ECDC Covid-19 Scenario Hub [7].

“Reproducible Processes” and the “Possibility of Communication of Advanced Simulations” (Blocks 4 and 7 in Figure 7) are essential to guarantee the credibility and usability of the models. We need tools to manage and share data (e.g. [8]) and models and to communicate not only the simulation results, but also the modelling process and model construction.

Last but not least, guidelines, standards that go beyond the concrete implementation are crucial (Blocks 8-10 in Figure 7). Here, the concepts of “Open and Independent Solutions”, “Priority for Data Security and Stake Holder Interests” and “Broad Applications” are crucial. The possibility of publication is limited, for example, by (justified) economic or data protection interests, which, however, leads to a lack of comparability of different models and thus jeopardises quality. This requires fundamental regulations such as those addressed in the General Data Protection Regulation and Data Governance Act. And clear and transparent processes are necessary for every project (even before the start of a simulation development) as well as the reuse of models is necessary to ensure quality and sustainability over time.

It is planned to publish the process description and examples as a White Paper within the European Umbrella Organisation of Simulation Societies EUROSIM [9].


References

- 1 Covid-19 Simulation in Austria, <http://www.dexhelpp.at/en/appliedprojects/tmp/covid-19/>, [Online; accessed 4.11.2022]
- 2 DEXHELPP, <http://www.dexhelpp.at/en/project-description/state-of-the-art/>, [Online; accessed 4.11.2022]
- 3 Abstract Kolloquiumsvortrag 14.6.2018 “Sharing Data, Methods, and Simulation Models – New Opportunities for Digital Health Care”, <https://www.informatik.uni-rostock.de/veranstaltungen/detailansicht-des-events/n/kolloquiumsvortrag-von-nikolas-popper-director-dexhelpp-wien-coordinator-centre-for-computational-complex-systems-tu-wien-46337/>, [Online; Accessed 4.11.2022]
- 4 Abstract Talk 25.3.2019 “Integrated Processes for Modelling & Simulation”, https://www.sfbtrr161.de/newsandpress/events_sfbtrr161/pastevents/, [Online; Accessed 4.11.2022]

- 5 N. Popper, B. Glock, F. Endel and G. Endel. Deterministic Record Linkage of Health Data as Preparatory Work in Modelling and Simulation-Use Case: Hospitalizations in Austria. *Proceedings of the 6th International Workshop on Innovative Simulation for Health Care*. pp. 44-49 (2017)
- 6 M. Bicher, C. Rippinger, C. Urach, D. Brunmeir, U. Siebert, N. Popper. Evaluation of Contact-Tracing Policies against the Spread of SARS-CoV-2 in Austria: An Agent-Based Simulation. *Medical Decision Making*, 41-8. pp. 1017-1032 (2021)
- 7 ECDC Covid-19 Scenario Hub, <https://covid19scenariohub.eu/>, [Online; accessed 4.11.2022]
- 8 N. Popper, M. Zechmeister, D. Brunmeir, C. Rippinger, N. Weibrecht, C. Urach, M. Bicher, G. Schneckenreither and A. Rauber. Synthetic Reproduction and Augmentation of COVID-19 Case Reporting Data by Agent-Based Simulation. *Data Science Journal*, 20(1). pp. 16ff (2021)
- 9 EUROSIM Technical Committee “Data Driven System Simulation”, <https://www.eurosim.info/tcs/tc-ddss/>, [Online; accessed 4.11.2022]

3.19 Using the Adaptable I/O System (ADIOS) for Effective and Sustainable Simulation Studies

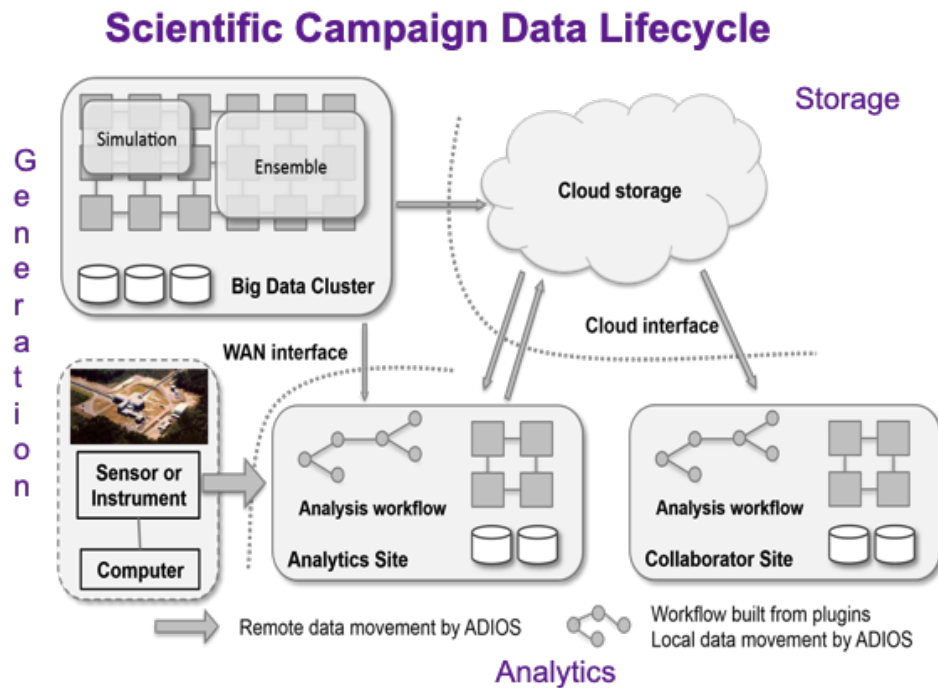
Caitlin Ross (Kitware – Clifton Park, US)

License  Creative Commons BY 4.0 International license
© Caitlin Ross

With the large amount of data written by large-scale scientific simulations, efficient parallel I/O is critical. However, leadership class systems have a number of factors that make developing applications difficult, from complex storage hierarchies to different parallel filesystems with different configurations. The Adaptable I/O System (ADIOS) [1] aims to simplify I/O, while achieving high performance on leadership class supercomputers and providing a simple API for developers. ADIOS is an open source C++ library that also provides bindings in C, Fortran, and Python. ADIOS is used by applications in various scientific fields, such as plasma physics, earth science, and aerospace engineering. Figure 8 shows the data lifecycle of a large-scale scientific campaign. ADIOS can be used to transfer data from simulations or sensors to persistent storage, couple simulations to exchange data, provide checkpoint/restart capabilities, or transfer data to other systems for analysis workflows.

ADIOS has a single, simple API to perform I/O, whether data is to be written to disk or streamed over a network. ADIOS provides this capability through different engines that can be selected at runtime. File engines include a self-describing format called binary packed (BP), as well as a HDF5 engine. There are a number of engines that stream data over a network, such as the Sustainable Staging Transport (SST) engine which can be used for exchanging data between loosely coupled jobs running on either the same compute cluster or over a wide area network. There is also an Inline engine that provides in process coupling of simulation and analysis/visualization for in situ analysis. Regardless of the engine chosen, it is selected in an XML configuration file and ingested by ADIOS at runtime, so data can be handled in different ways without needing to change the source code or using different API calls.

Recent work with ADIOS has involved integrating it into visualization tools such as ParaView [2] and the Visualization Toolkit (VTK) [3]. The Fides library was developed for this purpose and provides a data model schema which enables users to describe their data in a simple JSON format mapping ADIOS variables to the mesh and field characteristics of a



■ **Figure 8** The data generation, transfer, storage, and processing stages of a scientific campaign. ADIOS can be used for efficient transfer of data locally in an analysis workflow, as well as remotely over wide area network interfaces.

high-level data model. With ADIOS and Fides, simulations can use ParaView out of the box to visualize their data with post hoc, in situ, or in transit methods, without needing to write specialized adaptors or having a deep understanding of the VTK data model.

To the best of our knowledge, ADIOS is not currently being used in parallel discrete event simulations (PDES), but it could be useful in several ways for PDES. For instance, a recent US Department of Energy roundtable report [4] discusses the need for interoperability of multiple simulators that is portable and efficient on emerging platforms. ADIOS can be used for coupling simulations and provides portability and efficient parallel I/O.

The roundtable report also discusses incorporating machine learning (ML) frameworks into PDES engines/models. Some future work planned by the ADIOS team is to extend Fides metadata and provide data transformation utilities for ML data, which would enable ADIOS codes to integrate ML methods with their simulations. In addition to extending Fides metadata for ML data, we would also like to extend it for non-spatial data such as performance data and provide Python bindings. This would enable the use of other visualization and analysis tools besides scientific visualization tools like ParaView.

References

- 1 W. Godoy, N. Podhorszki, R. Wang, C. Atkins, G. Eisenhauer, J. Gu, P. Davis, J. Y. Choi, K. Germaschewski, K. Huck, A. Huebl, M. Kim, J. Kress, T. Kurc, Q. Liu, J. Logan, K. Mehta, G. Ostrouchov, M. Parashar, and S. Klasky, "ADIOS 2: The adaptable input output system. A framework for high-performance data management," *SoftwareX*, vol. 12, p. 100561, 07 2020.

- 2 U. Ayachit, The ParaView Guide. Kitware, Inc., 2015, 978-1-930934-30-6. [Online]. Available: <http://www.paraview.org>
- 3 Kitware Inc., “The Visualization Tool Kit (VTK),” <http://www.vtk.org/>, September 2021, [Online; accessed 08-October-2021].
- 4 K. Perumalla, M. Bremer, K. Brown, C. Chan, S. Eidenbenz, K. S. Hemmert, A. Hoisie, B. Newton, J. Nutaro, T. Opielstrup et al., “Computer science research needs for parallel discrete event simulation (PDES),” USDOE Office of Science (SC)(United States), Tech. Rep., 2022.

3.20 A Simulation Architecture to Study Diffusion Processes in Multiplex Networks

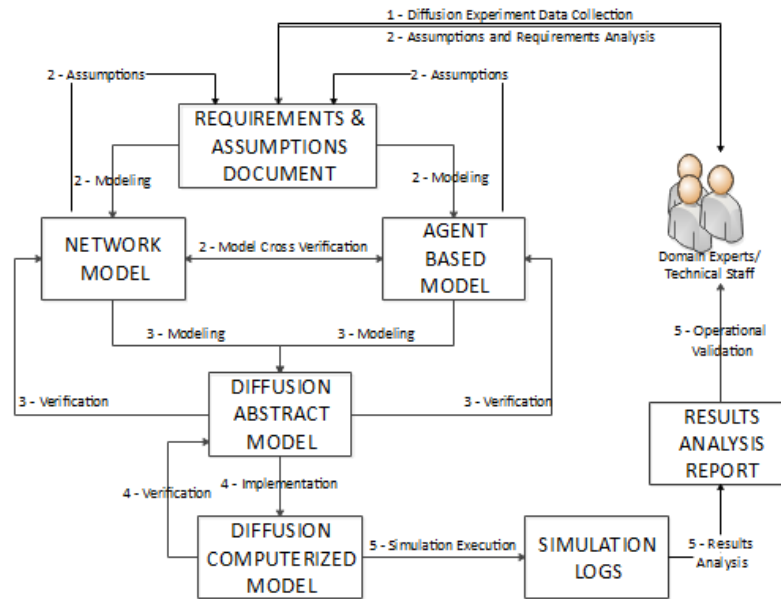
Cristina Ruiz-Martin (Carleton University – Ottawa, CA)

License © Creative Commons BY 4.0 International license
© Cristina Ruiz-Martin

Main reference Cristina Ruiz Martin, Gabriel A. Wainer, Adolfo López-Paredes: “Discrete-Event Modeling and Simulation of Diffusion Processes in Multiplex Networks”, *ACM Trans. Model. Comput. Simul.*, Vol. 31(1), pp. 6:1–6:32, 2021.

URL <https://doi.org/10.1145/3434490>

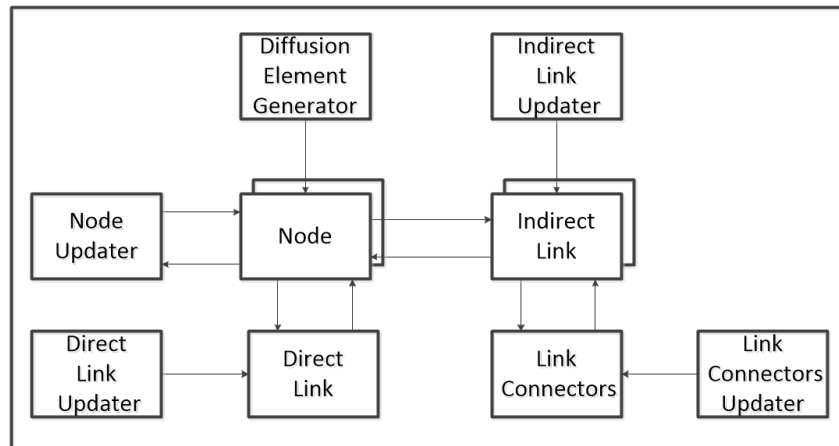
A variety of phenomena (such as the spread of diseases, pollution in rivers, etc.) can be studied as diffusion processes over networks (i.e. the diffusion of the phenomenon over a set of interconnected entities). There are different methods for studying diffusion processes, but most are based on various kinds of entities that are interconnected (networks). Two main approaches have been used to model the diffusion process over the network: Differential Equations (DE) and Agent-Based Modeling (ABM). The main advantage of nonlinear DE is that they can include a wide range of feedback effects (i.e. how the current value of a parameter of the system affects its future value, as in closed-loop systems). However, when they are used to study diffusion processes, one typically needs to aggregate nodes into fewer states or categories. Instead, ABM uses different attributes in each category, and different nodes in the same category may have different behavior. The network structure is clearly defined, and the behavior of each node is modeled individually at an increased computational cost. Likewise, neither of these two approaches provides well-established modeling and simulation (M&S) mechanisms for incorporating diffusion algorithms into multiplex dynamic networks and running simulations. The limitations of DE and ABM to studying diffusion processes pose the following question: how can we study diffusion processes in Multiplex Dynamic Networks to overcome the limitations of DE and ABM? Is there a framework that allows us to do this? To answer these questions, we introduced a method to study diffusion processes in multiplex dynamic networks, maintaining separations of concerns in all phases of modeling, implementation, and experimentation. The results include: a) an Architecture to study Diffusion Processes in Multiplex dynamic networks (ADPM); b) a systematic Process to define, implement and simulate diffusion processes over such networks. We use Network Theory formal specifications to define the topology of the diffusion process, Agent-Based Modeling (ABM) to define the behavior of the entities involved, and a formal specification of both for simulation modeling. This research uses the Discrete Event System Specification formalism (DEVS) to define the formal simulation model. The Architecture (and development process) to simulate Diffusion Processes in Multiplex dynamic networks (ADPM), is presented in Figure 9. The architecture is generic and can be used to study several types of diffusion processes.



■ **Figure 9** ADPM organization and workflow.

ADPM includes the Requirements and Assumptions Document of the problem; a Network model of the relations among components; an Agent-Based model of behavior; the Diffusion Abstract Model (DAM) (Figure 10), a formal representation based on the Network and Agent models; a Diffusion Computerized Model (DCM) of the DAM; the Simulation Logs and a Results Analysis Report.

The DAM is defined using a formal specification, DEVS, in our case. This solves some of the limitations of Network Theory, such as the lack of a formally verified simulator to simulate a diffusion process over the network. By including ABM, we can also define the behavior of both the nodes and links in the network. The formal DAM confers ABM rigor while separating modeling, verification, and experimentation. The main advantage of combining Network Theory, ABM, and DEVS is that we can use the most appropriate method to model the various aspects of the problem. Network Theory is well suited to model the relations between components; ABM is better suited to model behavior. DEVS provides a formal specification to define the whole model as components with hierarchical and modular specifications, which can be executed using well-established abstract simulation algorithms, which are proven to execute models correctly. This combination allows us to separate concerns and clearly differentiate each part of the problem, as well as separating models from simulation engines and experiments that are independent software entities.



■ **Figure 10** The DAM structure.

3.21 Model-based Software and Systems Engineering for Digital Twins

Bernhard Rumpe (RWTH Aachen, DE)

License © Creative Commons BY 4.0 International license
© Bernhard Rumpe

Main reference Manuela Dalibor, Malte Heithoff, Judith Michael, Lukas Netz, Jérôme Pfeiffer, Bernhard Rumpe, Simon Varga, Andreas Wortmann: “Generating customized low-code development platforms for digital twins”, *J. Comput. Lang.*, Vol. 70, p. 101117, 2022.

URL <https://doi.org/10.1016/j.cola.2022.101117>

Modeling is an important technique in many engineering disciplines. Modern modeling languages and tools allow developers to early concentrate on key aspects of the product and thus frontload quality assurance e.g. through simulation.

Equally important, explicit models of system requirements, technology independent function and architecture models, as well compact software and system component models enable reuse and variability management. Composition, tracing and refactoring assist evolutionary development and simulative quality assurance in a way that greatly reduces development cost for products of all kinds.

Using explicit models in appropriate modeling languages, like SysML or even domain specific languages, and an integrated, highly automated tool chain for construction, analysis and simulation is key to integrate all forms of system components. A holistic development approach needs a clear decomposition and decoupling and thus also well defined integration techniques.

Furthermore, models are a good basis for the construction of a digital twin, because a digital twin shares a lot of characteristics with a model. A digital twin is like a real twin: it is an active instance that interacts with the real system, allows to share real operative data, but also to simulate what the real twin would do and thus predict a systems behavior.

We examine the current state and problems of modeling for cyberphysical systems. In particular, we discuss how to make use of models in large development projects, where a set of heterogeneous models of different languages needs is developed and needs to fit together. A model based development process (both with UML/SysML as well as a domain specific modeling language) heavily relies on modeling core parts individually and composing those through generators to early and repeatedly cut simulative and productive code as well as a digital twin infrastructure from these models.

3.22 Data Farming: Better Decisions Via Inferential Big Data

Susan Sanchez (Naval Postgrad. School – Monterey, US)

License © Creative Commons BY 4.0 International license
© Susan Sanchez

Recently, the ready availability of “big data” has led to the adoption of data mining methods by organizations around the globe, as they seek to sift through massive volumes of data to find interesting patterns that are, in turn, transformed into actionable information. Yet a key drawback to the big data paradigm is that it relies on observational data, limiting the types of insights that can be gained.

The simulation world is different. Simulation models are integral to modern scientific research, national defense, industry and manufacturing, and in public policy debates. These models tend to be extremely complex, often with hundreds of thousands of potential factors (inputs or embedded parameters) and many sources of uncertainty. A “data farming” metaphor captures the notion of purposeful data generation from simulation models using efficient designed experiments.

Efficient design of experiments are required because it is literally impossible to examine even moderate numbers of factors by brute force. When the first supercomputer with petaflop performance (a quadrillion operations per second) was launched in 2008, the New York Times stated that machines such as this had “*the potential to fundamentally alter science and engineering*” by letting researchers “*ask questions and receive answers virtually interactively*” and “*perform experiments that would previously have been impractical*” [3]. Fourteen years later, the “Frontier” leads the world as the first exaflop machine [8], capable of over 10^{18} or a quintillion operations per second. But let us take a closer look at the practicality of a brute-force approach. Suppose a simulation has 100 factors, each factor has two levels (low and high) of interest, and we decide to look at all combinations of these 100 factors. A single replication of this brute-force experiment would take over 40 millenia on the Frontier, even if each of the 2^{100} simulation runs consisted of a single machine instruction! Efficient design of experiments can break this curse of dimensionality where expensive hardware cannot; experiments involving 100 factors can be completed in hours to weeks even for simulations with runtimes of minutes or hours [7].

With a data farming mindset, we can achieve tremendous leaps in the breadth, depth, and timeliness of the insights yielded by simulation. Large-scale experiments let us grow the simulation output efficiently and effectively. Modern statistical and visual analytic methods help us explore massive input spaces, uncover interesting features of complex simulation response surfaces, and explicitly identify cause-and-effect relationships [7, 2, 4]. Yet despite the current benefits offered by data farming, opportunities remain for advancing both the practice and research of simulation studies. A brief list of a few opportunities and challenges follows: see [1, 6, 5] for more.

From the practical standpoint, decision makers in many areas (climate change, economics, public health, to name a few) face increasingly complex problems where computational models are better than simple analytic models at capturing the complexity of the underlying system. At the same time, decision makers are increasingly comfortable with computerized and computer-based decisions based on observational big data, such as machine learning, artificial intelligence, and the use of digital twins. The rapid evolution of data science means that a greater number of simulation and non-simulation professionals are becoming more adept at scripting, modeling, graphical and statistical displays. Decision makers may, similarly, be less likely to shy away from using *model-driven* data to inform their decisions

if they are made aware of the potential benefits – particularly as they seek solutions to complex problems. By varying inputs in carefully chosen ways and exploring or building metamodels of the input/output relationships, we use simulation as an inferential (rather than descriptive) decision support tool. Rather than simply answering ‘what is?’ and ‘what if?’ questions, we can explore ‘what matters?, how?, and why?’

From a research perspective, the portfolio of methods suitable for addressing complex questions needs to be expanded. Further research is needed on multi-objective procedures; exploitation of parallel computing; adaptive sequential design methods; and methods that leverage the structure of inferential big data, rather than observational big data, for analysis and visualization tools. Regarding simulation optimization and other adaptive search techniques, it may be that we should be doing optimization on metamodels, rather than on the simulations themselves – and that we need automated ways of reoptimizing as these metamodels evolve over time. Adaptive sequential procedures are particularly important if a simulation study is viewed as an ongoing process, rather than a terminating one.

References

- 1 Elmegreen, B. E., S. M. Sanchez, and A. S. Szalay. 2014. The Future of Computerized Decision Making. In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk et al., 943–949. Piscataway, New Jersey: IEEE.
- 2 Feldkamp, N., S. Bergmann, and S. Strassburger. 2020. Knowledge discovery in simulation data. *ACM Transactions on Modeling and Computer Simulation*, 30(4): Article 22, 1–25.
- 3 Markoff, J. 2008. Military Supercomputer Sets Record. *New York Times*, June 9.
- 4 Matković, K., D. Gračanin, and H. Hauser. 2018. Visual analytics for simulation ensembles. In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe et al., 321–335. Piscataway, New Jersey: IEEE.
- 5 Sanchez, S. M. 2020. Data farming: Methods for the present, opportunities for the future. *ACM Transactions on Modeling and Computer Simulation*, 30(4): Article 22, 1–30.
- 6 Sanchez, S. M., and P. J. Sanchez. 2017. Better big data via data farming experiments. In *Advances in Modeling and Simulation: Seminal Research from 50 Years of Winter Simulation Conferences*, edited by A. Tolk, J. Fowler, G. Shao, and E. Yücesan, 159–179. Cham, Switzerland: Springer International Publishing.
- 7 Sanchez, S. M., P. J. Sanchez, and H. Wan. 2021. Work smarter, not harder: A tutorial on designing and conducting simulation experiments. In *Proceedings of the 2021 Winter Simulation Conference*, edited by S. Kim et al., 15 pages. Piscataway, New Jersey: IEEE.
- 8 TOP500 News Team 2022, May 30. ORNL’s Frontier first to break the exaflop ceiling. <https://www.top500.org/news/ornls-frontier-first-to-break-the-exaflop-ceiling/>.

3.23 Decision Making using Reinforcement Learning in Contested and Dynamic Environments

Claudia Szabo (University of Adelaide, AU)

License  Creative Commons BY 4.0 International license
© Claudia Szabo

Systems operating in military operations and crisis situations usually do so in contested and dynamic environments with poor and unreliable network conditions. Individual nodes within these systems usually have an incomplete, local and changing view of the system and its operating environment, and as such optimizing how nodes communicate in order to improve decision making is critical. Reinforcement learning approaches have been very

successful at solving problems in dynamic environments and in some cases when dealing with incomplete information. In this talk, we discuss the challenges of training and executing reinforcement learning approaches within such environments, in particular when considering communication middlewares.

3.24 Simulation-based Inference for Automatic Model Construction

Wen Jun Tan (Nanyang TU – Singapore, SG)

License © Creative Commons BY 4.0 International license
© Wen Jun Tan

Simulation-based inference aims to address the question of linking simulation models with empirical data by designing statistical inference procedures that can be applied to simulators [1]. Data assimilation is an inference method, in which the observed data are assimilated into the model to produce a time sequence of estimated system states [2]. Data assimilation was initially developed in the field of numerical weather prediction. However, simply inserting point-wise measurements into the dynamic weather models results in large instabilities in the simulation, rendering the forecasts meaningless. Hence, data assimilation is developed to initialize a model using observation data while making sure to maintain stability in the model by iteratively selecting the best system estimation.

The initial concept of data assimilation relies on the initialization of an existing model. However, changes in the observation data may also result from dynamic structural changes in the system. This means that the original simulation model built from the past system will be invalid; initialization of an outdated model will render the simulation results useless. Instead of assimilating the observation data into an existing model, we propose to build the model concurrently with the data assimilation procedure.

Considering a smart factory with Internet-of-Things (IoT) sensors to monitor physical events occurring in the factory. Process mining is frequently used to analyze operational processes based on event logs. In process mining, process discovery aims at constructing a process model as an abstract representation of an event log [3]. The goal is to build a model (e.g., a Petri net) that provides insight into the behavior captured in the log. A petri net is a class of discrete event dynamic system, which can be used to simulate discrete event systems, such as the manufacturing process in a factory.

Due to dynamic changes in the customers' orders, there will be frequent changes to the production operations in the factory. A new approach to combine data assimilation with process mining is proposed to accurately model the dynamic manufacturing processes. By monitoring the sequence of physical events, these events are captured into snapshots of the event lists for each event's arrival. Concurrently for each event, the real system performance is also measured, e.g., factory throughput. Process discovery will be performed on these snapshots to obtain the process models for each snapshot. Process enhancement will be used to enrich these process models by mining additional operational behaviours from the event logs, e.g., queuing or batching operations, to achieve a more realistic discrete event model of the factory. Each process model is simulated to obtain the system performance. Data assimilation will be performed by comparing the simulation performance with the real system performance and selecting the process model with the best estimated system states. By iterating discovering new process models and selecting the most accurate process model, the proposed approach is able to adapt to changes in the system structure while producing a calibrated simulation model at the same time.

References

- 1 Cranmer, K., Brehmer, J. & Louppe, G. The frontier of simulation-based inference. *Proceedings Of The National Academy Of Sciences*. **117**, 30055-30062 (2020)
- 2 Bouttier, F. & Courtier, P. Data assimilation concepts and methods March 1999. *Meteorological Training Course Lecture Series. ECMWF*. **718** pp. 59 (2002)
- 3 Van Dongen, B., Medeiros, A. & Wen, L. Process mining: Overview and outlook of petri net discovery algorithms. *Transactions On Petri Nets And Other Models Of Concurrency II*. pp. 225-242 (2009)

3.25 Models and Specifications within the Modeling and Simulation Life Cycle

Adelinde M. Uhrmacher (Universität Rostock, DE) and Claudia Szabo (University of Adelaide, AU)

License © Creative Commons BY 4.0 International license
© Adelinde M. Uhrmacher and Claudia Szabo

Modeling means organizing knowledge about a system of interest. In the talk, we will extend this citation, which has been attributed to Bernhard Zeigler, to the various artifacts of the modeling and simulation life cycle, as well as the life cycle itself. Thus, we will consider not only simulation models, but also models about simulation experiments, simulation methods, or entire simulation studies. With the subject of modeling the means of modeling varies as well, and includes approaches as diverse as domain-specific languages, formalisms, meta-models, ontologies, and logics. We will exemplify the interrelations between the subject of the model and the modeling approach. The second part of the talk is dedicated to the challenges that developing and applying model-based approaches face in modeling and simulation studies. Examples of such challenges including expressivity trade-offs, computational efficiency, reusability, accessibility, explainability, and evaluation are discussed.

3.26 A discrete-event approach for the study of sustainability in buildings

Gabriel A. Wainer (Carleton University – Ottawa, CA)

License © Creative Commons BY 4.0 International license
© Gabriel A. Wainer

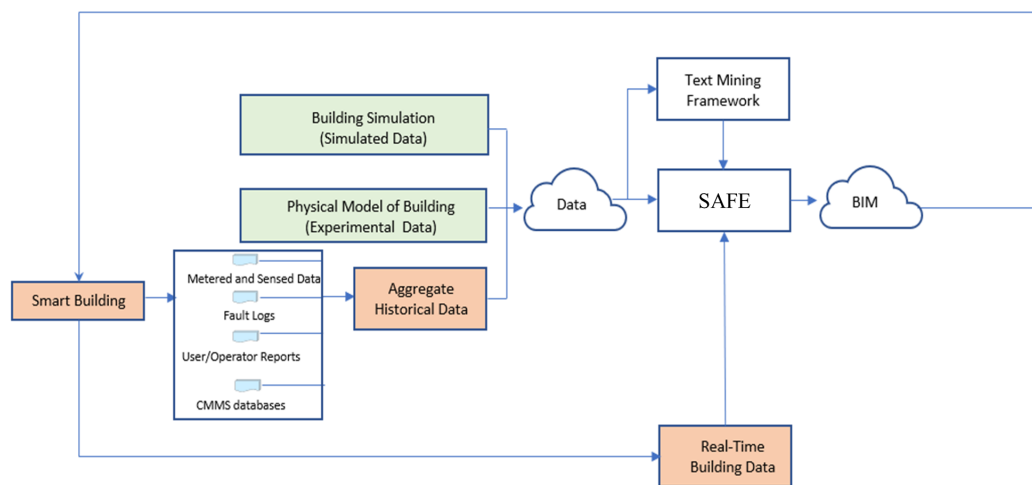
Joint work of Vinu Subashini Rajus, Joseph Boi-Ukeme, Rishabh Sudhir Jiresal, Nicolas Arellano Risopatrón, Pedram Nojedehi, Gabriel A. Wainer, Liam O'Brien, Stephen Fai

Main reference Vinu Subashini Rajus, Joseph Boi-Ukeme, Rishabh Sudhir Jiresal, Nicolas Arellano Risopatrón, Pedram Nojedehi, Gabriel A. Wainer, Liam O'Brien, Stephen Fai: Measured Data Reliability for Building Performance and Maintenance. *IEEE Instrum. Meas. Mag.* 25(1): 55-61 (2022)

URL <https://ieeexplore.ieee.org/document/9693445>

Indoor factors like thermal, visual, acoustic, and chemical exposure all impact occupants' wellbeing. One of the main approaches to improve wellbeing is to have continuous monitoring of the building. Similarly, building design issues or technical flaws in the building software can have negative impact on occupants' health [1]. Modern building systems include sensors, actuators, and control devices including a tight coupling of hardware and software features. Advanced building systems collect data to improve building performance, operation, and maintenance. Building systems today use a variety of state-of-the-art equipment, such as

embedded hardware, wide-area connectivity, and software for decision making [2]. We use cloud computing to increase collaboration among various building stakeholders and to display real-time data from buildings systems for performance and maintenance analysis. Using cloud services, designers can store large Building Information Models (BIM), historical data or simulated data. BIM allows managing the digital representation of building components, their digital geometry, metadata, relationships, and the parametric rules to manipulate them [3]. We propose a system architecture and a workflow called SUSTAIN (Sensor-based Unified Simulation Techniques for Advanced In-building Networks). SUSTAIN integrates fault tolerance models, a text-mining framework and BIM to improve building system reliability. The overall software architecture is shown in Fig.11.

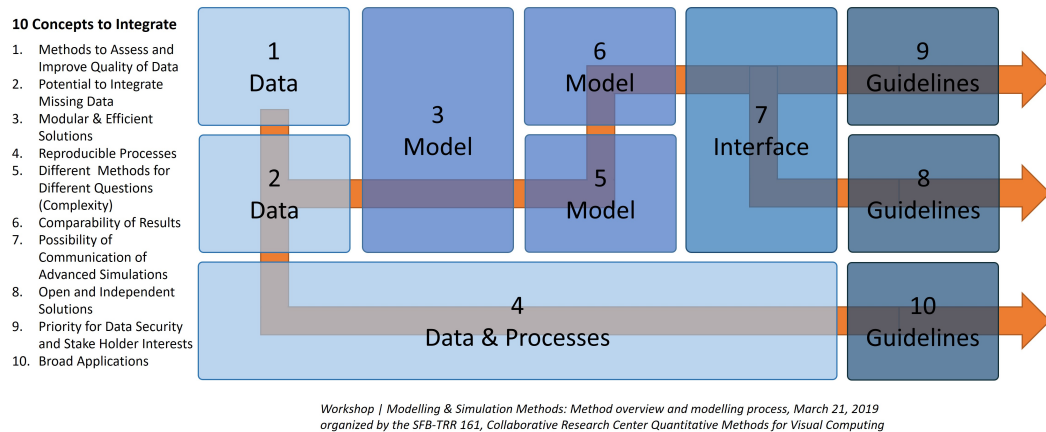


■ **Figure 11** SUSTAIN Architecture and Workflow.

The data collected from SAFE is integrated into BIM through Autodesk Forge, a cloud-based platform. Autodesk Forge allows to store and access various BIM and includes an integrated visualization component to visualize the real-time data through various web services. The BIM displays the sensor locations and ties the timeline data to its location. The aggregated data is represented as a shader to the room volume depending on the data parameters like temperature, humidity, or CO₂.

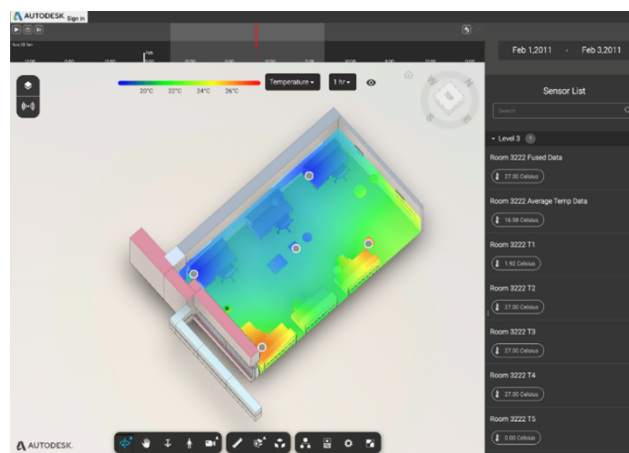
The VSIM BIM is a replica of the actual building with all the building elements (windows, doors, desks, etc.) and building systems (sensors, HVAC, etc.). We show a case study using temperature sensors with an operating range 0°C – 50°C (with an accuracy of $\pm 2\%$). We conduct simulation studies with different temperature readings within the measuring range were used. To evaluate the SAFE framework, faults were deliberately injected in both simulation and experimentation using different scenarios. The BIM of the lab was used to fabricate a maquette at a scale of 1:20 using laser cutting on acrylic sheets and adding control hardware (Fig. 12).

This physical model mimics the real-world setting of the research lab, which consists of six workstations, a fixed window, an operable blind, automated HVAC, and lighting controls. Fig. 13 shows the integration of the various sensors. Although the physical model does not behave as the real-world setting (e.g., the temperature in the maquette becomes steady faster than in the real building; however, when they are both in steady state, their behavior is similar), the objective here is different. SUSTAIN is focused on the development of the



■ **Figure 12** Physical prototype for experimentation [4].

building control software, the integration with 3D visualization, simulation and support in a cloud environment. The physical model permits conducting a variety of experiments useful during the software development cycle, and this can be done safely and without affecting the actual building operations.



■ **Figure 13** Integration of data with BIM using Autodesk Forge [5].

References

- 1 M. Arif, M. Kafatygiotou, A. Mazroei, A. Kaushik, and E. Elsarrag. *Impact of indoor environmental quality on occupant well-being and comfort: A review of the literature*. International Journal of Sustainable Built Environment, 5(1), pp.1-11, 2016.
- 2 V. Reppa, M.M. Polycarpou, and C. G. Polycarpou. *Sensor Fault Diagnosis*. Foundations and Trends in 2020 Spring Simulation Conference (SpringSim), (virtual) pp.1-12. IEEE, 2020.
- 3 S.A. Viktoros, M.K. Michael, and M.M Polycarpou. *Compact Fault Dictionaries for Efficient Sensor Fault Diagnosis in IoT-enabled CPSs*. In 2020 IEEE International Conference on Smart Internet of Things (SmartIoT), IEEE, 2020.
- 4 J. Boi-Ukeme, and G. Wainer. *A framework for the extension of DEVS with sensor fusion capabilities*. Systems and Control, 3(1-2), pp.1-248, 2016.
- 5 Autodesk Inc. *Dynamo*. <https://dynamobim.org/>. 2022.

3.27 Automatically Generating Simulation Experiments based on Provenance

Pia Wilsdorf (Universität Rostock, DE)

License © Creative Commons BY 4.0 International license
© Pia Wilsdorf

Main reference Pia Wilsdorf, Anja Wolpers, Jason Hilton, Fiete Haack, Adelinde M. Uhrmacher: “Automatic Reuse, Adaption, and Execution of Simulation Experiments via Provenance Patterns”, *ACM Trans. Model. Comput. Simul.*, Association for Computing Machinery, 2022.

URL <https://doi.org/10.1145/3564928>

Simulation experiments play a vital role during a simulation study, be it for calibration, validation, or exploration [1]. Many simulation experiments are used repeatedly throughout a simulation study or across related simulation studies. For instance, models are successively calibrated and validated after each major model refinement. Similarly, if a model is built by extension of a previous study, cross-validation experiments are carried out. And in the case where different models represent alternative hypotheses about a system, or the same model has been implemented for different modeling and simulation platforms, again slightly adapted simulation experiments are used to compare these model alternatives.

The Reuse and Adapt framework for Simulation Experiments (RASE) supports modelers in conducting these repeated experiments in a more systematic, effective, and efficient manner [2]. Based on predefined activity patterns and inference rules, key user activities are automatically detected (e.g., model refinements), whereupon suitable, previously executed simulation experiments (e.g., validation experiments) are selected, adapted, and executed in the context of the new simulation model, possibly reusing other information such as data or requirements. The framework is founded on the notion of provenance, i.e., information about the various entities (the research questions, simulation models, simulation experiments, simulation data, input data, requirements, qualitative model, assumptions, theories, etc.) and how they participated in or were generated by the diverse activities of modeling and simulation [3]. This information needs to be captured during a simulation study and stored as a directed acyclic graph based on the PROV-DM standard [4, 5] to be accessible to the pattern detection mechanism of the framework.

Currently, a new approach for capturing the provenance of simulation studies on-the-fly with minimal user involvement is being developed [6]. However, also the means for formalizing the context of a simulation model (provenance meta-data) need to be enhanced for provenance to be fully machine-interpretable. Here, we might take advantage of existing model-based approaches, such as signal temporal logic for specifying requirements [7], whereas, for the specification of other entities, such as assumptions or theories, no approaches yet exist.

References

- 1 Robert G. Sargent. An introduction to verification and validation of simulation models. In *2013 Winter Simulations Conference (WSC)*, pages 321–327, 2013.
- 2 Pia Wilsdorf, Anja Wolpers, Jason Hilton, Fiete Haack, and Adelinde M. Uhrmacher. Automatic reuse, adaption, and execution of simulation experiments via provenance patterns. *ACM Transactions on Modeling and Computer Simulation*, 2022.
- 3 Kai Budde, Jacob Smith, Pia Wilsdorf, Fiete Haack, and Adelinde M. Uhrmacher. Relating simulation studies by provenance—developing a family of wnt signaling models. *PLOS Computational Biology*, 17(8):1–27, 2021.
- 4 Andreas Ruschewski, Pia Wilsdorf, Marcus Dombrowsky, and Adelinde M. Uhrmacher. Capturing and reporting provenance information of simulation studies based on an artifact-

based workflow approach. In *Proceedings of the 2019 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, SIGSIM-PADS '19, page 185–196, 2019.

- 5 Khalid Belhajjame, Reza B'Far, James Cheney, Sam Coppens, Stephen Cresswell, Yolanda Gil, Paul Groth, Graham Klyne, Timothy Lebo, Jim McCusker, et al. PROV-DM: The PROV data model. *W3C Recommendation*, 14:15–16, 2013.
- 6 Anja Wolpers. *Lightweight Provenance Capturing for Simulation Studies*. PhD thesis, University of Rostock, 2022.
- 7 Laura Nenzi and Luca Bortolussi. Specifying and monitoring properties of stochastic spatio-temporal systems in signal temporal logic. *EAI Endorsed Transactions on Cloud Systems*, 1(4), 2015.

4 Working groups

4.1 Intelligent Modeling and Simulation Lifecycle

Wentong Cai (Nanyang TU – Singapore, SG), Philipp Andelfinger (Universität Rostock, DE), Luca Bortolussi (University of Trieste, IT), Christopher Carothers (Rensselaer Polytechnic Institute – Troy, US), Dong (Kevin) Jin (University of Arkansas – Fayetteville, US), Till Köster (Universität Rostock, DE), Michael Lees (University of Amsterdam, NL), Jason Liu (Florida International University – Miami, US), Margaret Loper (Georgia Institute of Technology – Atlanta, US), Alessandro Pellegrini (University of Rome “Tor Vergata”, IT), Wen Jun Tan (Nanyang TU – Singapore, SG), and Verena Wolf (Universität des Saarlandes – Saarbrücken, DE)

License © Creative Commons BY 4.0 International license

© Wentong Cai, Philipp Andelfinger, Luca Bortolussi, Christopher Carothers, Dong (Kevin) Jin, Till Köster, Michael Lees, Jason Liu, Margaret Loper, Alessandro Pellegrini, Wen Jun Tan, and Verena Wolf

Modeling and simulation (M&S) has shown to be effective at conducting what-if analyses of complex scenarios. However, the current societal and technical challenges require building increasingly complex models and carrying out larger scale simulation experiments, which call for more efficient and intelligent approaches in all aspects of simulation studies, from model creation to model execution and experimentation.

In his state-of-the-art and open-challenges (STAROC) presentation on “High Performance Computing (HPC): Exploiting New Architectures”, Christopher Carothers introduced an array of new emerging hardware that supports effective and sustainable simulation studies, including ultra-dense chips, fast analog computing, wafer scale processor, low precision computing, and brain inspired chips. He went on to discuss the demand for Artificial Intelligence (AI) / Machine Learning (ML) and how it is driving the development of hardware and applications in the HPC sector. The trend of growing system complexity is also accelerating the development of software tools and hardware platforms specifically developed for AI/ML [1]. In his presentation, Carothers posed the following questions:

- Can AI/ML help M&S be more sustainable?
- Can M&S help improve AI/ML predictions?

To enable efficient utilization of the HPC resources, M&S needs to harness these new emerging computing hardware and platforms developed for AI/ML. In addition, AI/ML also opens up new and interesting opportunities to enhance and enrich M&S [2].

Initially, there were two topics identified for the working group: i) AI/ML + Simulation, and ii) Enabling Models to Run Efficiently on Heterogeneous Hardware. The first topic involves the integration of AI/ML techniques with M&S; and the second topic focuses on the efficient utilization of heterogeneous hardware to execute simulation models. As seen from Carothers' presentation, the current development of HPC systems and applications are "riding the AI/ML wave", i.e. the new emerging hardware developed for AI/ML. As a result, the working group decided to change the questions to incorporate the notion of the AI/ML wave: i) how can the emerging hardware be efficiently utilized for M&S? and ii) can M&S also ride on the AI/ML wave to make it more efficient and sustainable?

In the last decade, research challenges on exploiting emerging computing platforms and using heterogeneous hardware to accelerate the execution of discrete event simulations have been addressed in several research papers:

- 2015 Paper on "Grad Challenges for Modelling and Simulation" [3],
- 2016 Richard Fujimoto's STAROC paper on "Parallel and Distributed Simulation (PADS)" [4],
- 2016 NSF Workshop on "Future Research in Modelling and Simulation" [5],
- 2019 Survey paper on "Agent-based Simulation Using Hardware Accelerators" [6], and
- 2022 DOE Roundtable report on "Computer Science Research Needs for Parallel Discrete Event Simulation (PDES)" [7].

Particularly, Perumalla et al presented a computer science-oriented view of research challenges in PDES and identified priority research opportunities in advanced computing for PDES [7].

Given this prior research, the working group decided to focus the discussion on a merged topic: Intelligent Modeling and Simulation Lifecycle. The group believed the STAROC paper should focus on how modelling and simulation can benefit from recent advances in AI/ML techniques as well as the emerging powerful and pervasive hardware, computing paradigms and systems. In the proposed paper, we will examine the existing and emerging hardware/systems and AI/ML techniques in the context of modeling and simulation lifecycle (e.g., model creation, calibration, execution, and experimentation). We will also identify the major challenges and opportunities faced in M&S applications and outline important research directions for intelligent modeling and simulation, to improve speed, accuracy, and capabilities, by taking advantage of AI/ML and computing paradigms.

After intense discussion and brainstorming, the working group developed the following structure for the proposed STAROC paper:

- The first section of the paper will provide the motivations and introduce the main thesis of this article. We will focus on how advanced AI/ML and existing/emerging computing hardware/platforms will improve M&S at different stages of the M&S lifecycle.
- The second section "M&S Lifecycle, Challenges, and Demands" will provide an introduction to the M&S lifecycle: creation, calibration, execution, simulation experiments. These stages shall be defined clearly with respect to the traditional M&S pipeline: input modeling, concept model & validation, computational model & verification, experiment design, model execution, and output analysis. We will identify major M&S applications, including transportation and mobility, power grid, networking and cybersecurity, HPC, and epidemic modeling, in order to identify challenges and demands at different M&S stages that can be fulfilled by AI/ML and new emerging computing hardware and platforms.

- The third section “Emerging AI/ML-Assisted Approaches for M&S” will review the existing prominent AI/ML techniques and methodologies that have been applied for M&S. This will be achieved by categorizing these approaches to the various stages of the M&S lifecycle, and thereby identifying the gaps that can be explored in future research efforts.
- The fourth section “Heterogeneous and Emerging Computing Platforms for M&S” will categorize the diverse computing systems, platforms and emerging computing hardware that can support the M&S mission. First, we will introduce the existing computing platforms, e.g., HPC, cloud, edge, accelerators (GPUs, FPGAs). Next, we will explore the newer emerging hardware that are specially developed for AI/ML computing, e.g., neuromorphic devices, low-precision AI chips. At the same time, we will also examine the different computational paradigms: serverless computing, micro-services, virtualization, containerization, quantum computing or analog computing.
- The fifth section “Research Challenges and Roadmap for Intelligent M&S Lifecycle” will identify specific research directions and outline a roadmap for applying AI/ML and emerging hardware/computing platforms for the M&S lifecycle. We will create a table to illustrate the gaps or challenges at each stage of the M&S lifecycle and list the corresponding techniques and research directions to be taken to address these challenges. In the subsequent subsections, we will elaborate on these techniques and research directions.
- Finally, we will conclude the article by pointing out that we are focusing on using AI/ML for M&S. However, there are other research directions, such as applying M&S for improving AL/ML (such as explainable AI).

Tentatively, the working group plans to complete and submit the paper by mid-Jan 2023.

References

- 1 Daniel A Reed and Jack Dongarra. Exascale computing and big data. *Communications Of The ACM*. 58(7): 56-68, July 2015.
- 2 Alexander Lavin, Hector Zenil, Brooks Paige, David Krakauer, Justin Gottschlich, Tim Mattson, Anima Anandkumar, et al. Simulation intelligence: Towards a new generation of scientific methods. *ArXiv Preprint ArXiv:2112.03235*. 2021.
- 3 Simon J. E. Taylor, Azam Khan, Katherine L. Morse, Andreas Tolk, Levent Yilmaz, Justyna Zander, and Pieter J. Mosterman. Grand challenges for modeling and simulation: simulation everywhere—from cyberinfrastructure to clouds to citizens. *Simulation: Transactions of the Society for Modeling and Simulation International* . 91(7):648-665, 2015.
- 4 Richard Fujimoto. Research challenges in parallel and distributed simulation. *ACM Transactions On Modeling And Computer Simulation (TOMACS)*. 26(4):22:1 -22:29, May 2016.
- 5 Christopher Carothers, Alois Ferscha, Richard M. Fujimoto, David Jefferson, Margaret Loper, Madhav Marathe, P. Mosterman, Simon Taylor, and H. Vakilzadian. Computational challenges in modeling and simulation, in *Research Challenges In Modeling And Simulation For Engineering Complex Systems*. pp. 45-74, 2017.
- 6 Jiajian Xiao, Philipp Andelfinger, David Eckhoff, Wentong Cai, and Alois Knoll. A survey on agent-based simulation using hardware accelerators. *ACM Computing Surveys (CSUR)*. 51(6):131:1 – 131:35, Nov 2019.
- 7 Kalyan Perumalla, Maximilian Bremer, Kevin Brown, Cy Chan, Stephan Eidenbenz, Scott K. Hemmert, Adolfo Hoisie, Benjamin Newton, James Nutra, Tomas Ooppelstrup, Robert Ross, Markus Schordan, Nathan Urban. Computer Science Research Needs for Parallel Discrete Event Simulation (PDES US Department of Energy, Advanced Scientific Computing Research, Roundtable Report, May 2022. (DOR: 10.2172/1855247)

4.2 Policy by simulation: seeing is believing for interactive model co-creation and effective intervention

Rodrigo Castro (University of Buenos Aires, AR), Joachim Denil (University of Antwerp, BE), Jérôme Feret (ENS – Paris, FR), Kresimir Matkovic (VRVis – Wien, AT), Niki Popper (Technische Universität Wien, AT), Susan Sanchez (Naval Postgrad. School – Monterey, US), and Peter Slood (University of Amsterdam, NL)

License © Creative Commons BY 4.0 International license
© Rodrigo Castro, Joachim Denil, Jérôme Feret, Kresimir Matkovic, Niki Popper, Susan Sanchez, and Peter Slood

Motivation

We understand that policy makers are currently much more willing (than they were 10 years ago, for example) to participate in simulation-assisted planning and decision-making processes. We must therefore be well prepared so we are able to seize the opportunity.

Simulation is particularly useful in decision support when bifurcations can occur, i.e., when a system under study takes qualitatively different courses depending on the input parameters or random events as the system dynamics evolve, a distinctive feature of complex systems. Interventions represent a special example that by definition leads to discrete changes in the dynamics that may not be represented with classical methods alone [13].

Two key aspects of policy by simulation are that the systems being addressed are inevitably complex (they include overlapping facets such as social, physical, biological, cybernetic, etc.) and also that the interventions being applied in the real world must ultimately and inevitably be understood and agreed upon by humans.

As there are usually several stakeholders from different domains, it is essential to provide means which will efficiently support specifications of interventions as well as communication of results which do not rely solely on mathematical-computational models.

In this context, dynamic and reactive visualization emerges as a key element for human understanding of interventions in complex systems.

While the technologies of modeling, simulation and visualization (M&S&V) have made enormous progress during the past several decades, several challenges remain. We need to provide efficient and sustainable mechanisms that engage interdisciplinary teams of decision makers with simulation-assisted processes.

Meanwhile, recent work highlights missing links between user tasks in visualization taxonomies (e.g., sensemaking) and the high-level task of decision-making (Dimara and Stasko [3]). They identified a lack of interdisciplinary approaches as one of the key causes of this mismatch. The latter suggests that a joint approach of modeling, simulation and visualization experts stands as a valid approach to reduce this gap.

We therefore envision “policy by simulation”, a comprehensive co-creation framework that exploits state-of-the-art modeling, simulation, experimentation, visualization and AI to engage policy makers and stakeholders from multiple domains concerned with a common system under study. The goal is an efficient and sustainable framework of interoperable M&S&V-based tools that allow stakeholders to model potential interventions into simulation worlds and deliver timely insights about complex systems for the purpose of planning efficient and effective real-world interventions.

These envisaged capabilities are shared by several initiatives such as the POLDER Simulation Center (POLicy Decision-support and Evidence-based Reasoning) [15], the DEXHELPP Project (DEcision support for HEalth Policy and Planning) [16], the SEED Center (Simulation Experiments and Efficient Designs) for Data Farming [17], and the Simulation and Immersive Visualization Lab [18].

We first present the framework we envision, and then describe some of the open research challenges that must be addressed to fully realize its potential.

A Policy by Simulation Framework

To implement an integrated decision-making process that covers both simulation aspects and stakeholder needs for modeling feasible interventions, we envision an iterative three-level structure. The ultimate aim is to unveil potentials for action and to outline a basic structure as well as minimum requirements without claiming to be exhaustive.

- **Conceptual level.** Generate problem statements, identify mental models, create a narrative definition of goals, identify areas of conflict, and discuss attitudes toward risk. Identify the timeline for making decisions about whether and how to intervene. This will be revisited as external circumstances change, insights arise, or new stakeholders join the team.
- **Logical level.** Determine data availability or specify data model specifications and assumptions. Describe correlations, causations, causal loops, anchor points, levers, and potential interventions. Discuss the key response measures and aggregation levels. Prioritize features to include in the next round of executable models. Delegate responsibility for creating embedded submodels or suitable linkages between stand-alone submodels and the overarching model.
- **Assessment level.** Create, refine, and explore executable models. This can include models at different levels of granularity, from the coarse grain level to a highly detailed level. Types of models could include systems dynamics models (SD), partial differential equations (PDE), agent based models (ABM), discrete event simulation (DES), complex networks (CN), and more. Policy stakeholders will not be involved in the actual model implementation, but are vital for scoping the experimental region(s) of interest. Efficient experiments can be used to grow data that can be jointly and visually explored and assessed, spawning further experimentation as insights are gained and new questions arise.

Visualization techniques shall encompass all stages, providing visual coherence and continuity across all levels of information, connecting input data, simulation models, and simulation results across levels as the study evolves. Visualization is a key enabler for model-based tradeoff discussions among stakeholders with competing priorities and world views.

On the one hand, these visualization techniques should cover all the necessary needs in each stage. On the other hand, they should provide sufficient freedom, especially in early phases, so as not to impose restrictions that could lead to mapping errors and ultimately reduce the quality of the process and the intervention decisions.

Research Challenges: Bridging the Credibility Gap

The overall policy by simulation framework can be implemented using existing tools and methods. However, its full potential will not be realized without additional research in a variety of areas.

Credibility Gap: Even though we can create sophisticated simulation models for policy intervention, we often fall short in communicating their structure and results in transparent and understandable ways. For example, causal loop or process flow diagrams are very understandable, but more detailed models may be difficult for some stakeholders to interpret.

Past experience with model-based approaches (or lack thereof) may also affect stakeholder credibility in different ways.

A priori: The question of which aspects are important depends not only on the subjective view of the stakeholders, but also on the non-assessability of dynamic effects. These are to be investigated either by data analysis or by methods of causal analysis. In this case, the estimation of potentials is often low.

A posteriori: Real-world validation is difficult in the sense of the prevention paradox, interventions that are set cannot be compared with interventions that are not set—at least in real-world settings. Also, for some modeling situations and paradigms the system boundaries must be narrowly chosen in order to be able to create meaningful models.

We describe four threads we feel are key for bridging the credibility gap: model co-creation, visualization, flexibility, and efficiency.

Co-creation

There is a lack of robust methodologies for co-creating a palette of simulation models [12, 6] that can refocus across several levels of model detail, from conceptual high-level models down to very detailed model components. Similar methods are also needed for co-creating and co-exploring disparate models that capture different aspects of the system of interest, regardless of their level of detail, as part of a “many model thinking” paradigm.

Approaches that can support model co-creation, but require further research, include:

- **Immersion.** As simulation professionals, we have to show that we understand the context, the stakeholders and their interests, and how the simulation is embedded in the whole environment. System and simulation boundaries have to be discussed. These can be changed but have to be communicated in transparent ways. Similarly, stakeholders themselves must become immersed in the co-creation process. We need methods for eliciting their inputs and assumptions, engaging them with other stakeholders during the model development process, and engendering their trust in using a model-based framework to gain insights.
- **Formal methods.** Formal methods [4] for model relation and model reduction can help increase stakeholder confidence. They can relate models at different levels of abstraction and automatically derive reduced models while ensuring a formal relationship between the initial and the reduced ones. This can be done either analytically (which eases their application to a wide collection of models) or by language specific approaches (based on the structural properties of the description language). So derived model reduction may be exact [2] (as a reduction of the dimensionality of the model), or approximated [1] (including explicit bounds for numerical errors). Analytic approaches may require unwarranted assumptions and struggle to scale to the complexity of large models. Event graphs are modeling representations that can be implemented in a variety of languages, but are less widely used [11]. The development of language-based approaches is more time consuming since a given tool will work only for a dedicated modeling language. Further research is needed to address these challenges.

Visualization

Although visualization technologies and visual analytics techniques are established as exploratory methods and have been applied to numerous decision making and simulation problems, visualization and interaction with models at different levels of detail in an integrated framework remains an unsolved challenge. Further research on visualization techniques

that leverage the structure of data from designed simulation experiments is also needed. Comparison of different scenarios [5], provenance of information and tracking of interactions [7], efficient visualization for a plethora of stakeholders from different domains [3], all represent significant challenges for visualization research. Finally, as the number of stakeholders increases, there will be more and more scenarios to evaluate. Visualizing many scenarios represents a scalability challenge for visualization [8]. Limited screen space often does not allow to show all possibilities at once. At the same time important information has to be shown.

Flexibility

As demands and goals change over time, we need to be able to zoom in and out in the simulation scenarios across different levels of model detail. We also need the ability to focus on different model components to reveal the worldviews of different stakeholders.

- **Integration.** Integration for Exploration: We must build in flexibility in the form of software hooks from the beginning of the co-creation process. This includes flexibility in setting or modifying inputs and linking them to outputs, so that causal relationships within the model setting can be readily explored and identified. Input links should facilitate switching between real-world data (if available), diverse input data models, or changes driven by a mixture of approaches including interactive visual analytics, single-stage or adaptive experimental designs, or AI/ML-guided searches.
- **Flexible timing.** We need to provide partial insights in a timely way. Since this may involve assessments when the executable models are all at very high levels, input data are limited or unavailable, or only portions of the modeling are complete, best practices should be developed for effectively engaging stakeholders so the overall process provides value. How and when to revisit insights obtained early in the co-creation process, or how to determine the appropriate level for a particular intervention opportunity, remain open challenges.

Efficiency

We take a broad view of efficiency because the overarching goal of policy by simulation is that the stakeholders gain insights about effective interventions in a timely manner. Clearly, modeling efficiency and simulation run-time efficiency play their parts, and are often the focus of benchmark studies. We need metrics that capture the entire path to effective intervention by including person-hours (for modeling, analysis, visualization, documentation, discussions, etc.) as well as the computer cycles needed to arrive at insights. Harder (but extremely important) aspects to capture are the flexibility provided, the quality of the insights obtained, and the effectiveness of the intervention.

- **Sustainability.** There is an unmet need for automatic mechanisms that provide robust reusability of model libraries and experiments so that simulation projects can be properly evolved and adapted across longer periods of time. This includes accommodations for legacy models that continue to be used, as well as sunset clauses to handle models that should be reimplemented or retired as new modeling paradigms or languages are developed. Methods that reduce maintenance expenses for these “living models” are also of interest.
- **Model selection.** As a result of the zoom in/out requirement, we may need to use different methods for detailed modeling producing different outcomes. This imposes

different requirements for the underlying numerical and computational implementations. Over time, “best modeling approaches” might be revealed for certain problem domains. Such guidance would improve the modeling efficiency and so improve the timeliness of the overall process. This is another area that merits further research.

The above presented four threads –co-creation, visualization, flexibility, and efficiency– are interrelated. Even minor advances in any area may lead to dramatic improvements in the process as a whole.

Guiding principles

It is well known that in policy making, choosing not to intervene can be as decisive as enacting a particular intervention. Moreover, there are often windows of opportunity that require making the best possible decision in a given time frame, weighing the best sources of (always imperfect) information available at the time. We believe that the quality of such information can be greatly improved by following processes such as those presented herein, relying on simulation models and visualization techniques for intervention planning.

At early stages it is key to determine what kind of interventions could be feasibly made, identifying the lever points for controllable interventions in the complex system under study. This is a result of a discursive process with multiple stakeholders involved, including those with legitimately contradictory perspectives and worldviews of the overarching problem at hand.

The co-creation of simulation models implies including all necessary subsystems, variables, parameters and data or data models that capture the valid concerns and interests of heterogeneous stakeholders involved in the modeling process.

One important measure of the quality of the resulting models (or family of models) is the degree at which they allow the stakeholders to exercise relevant what-if analysis in an exploratory process. Emergent behavior is a key feature of complex systems [14] (basically anything unexpected can happen) therefore “potentially feasible paths of action” must be explored. While experimenting, the framework shall keep track of all interventions in the models along the exploratory process. To that aim, different alternatives are possible. If the computational time for single runs is long, checkpoints could be recorded enabling the team to “go back and rebranch” if a given path of interventions is rendered inefficient/undesirable. If storage requirements for checkpoints is prohibitive, random number seeds could be recorded enabling the team to initially output limited data (e.g., end-of-run or other summary information) but “go back and restart” to output detailed information or additional performance measures. The exploratory process itself becomes the result of a discussion with a group of “humans in the loop”. Varied data widgets should provide, simultaneously, multiple different views on a same system, both for simulation results and for the structure of model components. Yet, in parallel to human-based analysis, AI/ML-driven algorithms can automatically explore the parametric space (intelligent parameter sweeping) to prune and rule out model configurations that lead to undesirable results (e.g., violate restrictions on state variables). Other automatic analyses that can enrich the knowledge about the underlying model are parameter sensitivity analyses, which can provide a better understanding of the relative potential impacts of lever points. These sensitivity levels can be visually mapped into the model structure to guide the experience in a more efficient way. Large-scale designed experiments (“data farming”) can be used to identify the most impactful model inputs and interactions, uncover other interesting features about the I/O behavior, and assess how uncertainties propagate through a model or hierarchy of models [9, 10]. Such insights can help steer the interactive exploration toward relevant paths.

Related needs

The implementation of the proposed process has implications for other areas of simulation technology -not addressed here- which were dealt with in the Dagstuhl Seminar “Computer Science Methods for Effective and Sustainable Simulation Studies” in October 2022:

- Reduce simulation ecological footprint while providing increased simulation performance
- Provide efficient automatic documentation of models along with their ranges of validity
- Track the provenance of models, simulation experiments, and the associated input and output data
- Create better ways (standards?) to store and reuse model components across several projects with some common parts of their systems under study

Summary

We believe that the “policy by simulation” framework we propose can – and should – be used for identifying and implementing real-world interventions to address many complex problems currently faced by policy makers. If done effectively, this approach will improve the quality, timeliness, and effectiveness of the intervention decisions, will also enhance the insights stakeholders gain from the simulation models, and so improve the credibility of the collections of models used. On a broader level, as policy makers and other stakeholders become more familiar with this framework, we will have more opportunities to create a mindset that values and celebrates the exploration of interventions in the virtual world, rather than bemoaning the inability for a posteriori validation in the real world. We envision policy by simulation as an integrated, interactive, ongoing process, rather than a one-time product. Similarly, the framework itself will evolve as new tools and methods are developed to address the many research challenges identified above. Also, as we CS and simulation professionals, policy makers, and other stakeholders co-create the models and co-learn from analytic and visual exploration of their behavior, we expect other challenges and research opportunities to arise.

References

- 1 Andreea Beica, Jérôme Feret, and Tatjana Petrov. Tropical abstraction of biochemical reaction networks with guarantees. *Electronic Notes in Theoretical Computer Science*, 350:3–32, 2020. Proceedings of SASB 2018, the Ninth International Workshop on Static Analysis and Systems Biology, Freiburg, Germany – August 28th, 2018.
- 2 Vincent Danos, Jérôme Feret, Walter Fontana, Russell Harmer, and Jean Krivine. Abstracting the differential semantics of rule-based models: Exact and automated model reduction. In *2010 25th Annual IEEE Symposium on Logic in Computer Science*, pages 362–381, 2010.
- 3 Evanthia Dimara and John Stasko. A critical reflection on visualization research: Where do decision making tasks hide? *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1128–1138, 2022.
- 4 Patrick Cousot and Radhia Cousot. Abstract interpretation: A unified lattice model for static analysis of programs by construction or approximation of fixpoints. In Robert M. Graham, Michael A. Harrison, and Ravi Sethi, editors, *Conference Record of the Fourth ACM Symposium on Principles of Programming Languages, Los Angeles, California, USA, January 1977*, pages 238–252. ACM, 1977.
- 5 Michael Gleicher, Danielle Albers, Rick Walker, Ilir Jusufi, Charles D. Hansen, and Jonathan C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, Oct 2011.
- 6 Scott E. Page. *The Model Thinker: What You Need to Know to Make Data Work for You*. Basic Books, Inc., USA, 2018.

- 7 Eric D. Ragan, Alex Endert, Jibonananda Sanyal, and Jian Chen. Characterizing provenance in visualization and data analysis: An organizational framework of provenance types and purposes. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):31–40, 2016.
- 8 Gaëlle Richer, Alexis Pister, Moataz Abdelaal, Jean-Daniel Fekete, Michael Sedlmair, and Daniel Weiskopf. Scalability in visualization, arxiv.org/abs/2210.06562, 2022.
- 9 Susan M. Sanchez. Data farming: Methods for the present, opportunities for the future. *ACM Transactions on Modeling and Computer Simulation*, 30(4): Article 22.
- 10 Susan M. Sanchez, Paul J. Sanchez, Hong Wan. Work smarter, not harder: A tutorial on designing and conducting simulation experiments. In *Proceedings of the 2021 Winter Simulation Conference*, 15 pages, 2021.
- 11 Lee W. Schruben. Simulation modeling with event graphs. *Communications of the ACM*, 28(11):957-963, 2013.
- 12 Bernard P. Zeigler. Why should we develop simulation models in pairs? In *2017 Winter Simulation Conference (WSC)*, pages 2–2, 2017.
- 13 Rodrigo Castro. Chapter 24 – Open Research Problems: Systems Dynamics, Complex Systems. *Theory Of Modeling And Simulation (Third Edition)*. pp. 641-658 (2019)
- 14 Foguelman, D., Henning, P., Uhrmacher, A. and Castro, R. EB-DEVS: A formal framework for modeling and simulation of emergent behavior in dynamic complex systems. *Journal Of Computational Science*. **53** pp. 101387 (2021)
- 15 POLDER Simulation Center (POLicy Decision-support and Evidence-based Reasoning) <https://polder.center/>
- 16 DEXHELPP Project (DEcision support for HEalth Policy and Planning) <http://www.dexhelpp.at/>
- 17 SEED Center (Simulation Experiments and Efficient Designs) for Data Farming <https://harvest.nps.edu/>
- 18 Simulation and Immersive Visualization Lab <https://modsimu.exp.dc.uba.ar/>

4.3 Context, composition, automation and communication: towards sustainable simulation studies

Adelinde M. Uhrmacher (Universität Rostock, DE), Peter Frazier (Cornell University – Ithaca, US), Reiner Hähnle (TU Darmstadt, DE), Franziska Klügl (University of Örebro, SE), Fabian Lorig (Malmö University, SE), Bertram Ludäscher (University of Illinois at Urbana-Champaign, US), Laura Nenzi (University of Trieste, IT), Cristina Ruiz-Martin (Carleton University – Ottawa, CA), Bernhard Rumpe (RWTH Aachen, DE), Claudia Szabo (University of Adelaide, AU), Gabriel A. Wainer (Carleton University – Ottawa, CA), and Pia Wilsdorf (Universität Rostock, DE)

License © Creative Commons BY 4.0 International license

© Adelinde M. Uhrmacher, Peter Frazier, Reiner Hähnle, Franziska Klügl, Fabian Lorig, Bertram Ludäscher, Laura Nenzi, Cristina Ruiz-Martin, Bernhard Rumpe, Claudia Szabo, Gabriel A. Wainer, and Pia Wilsdorf

Motivation. Simulation has become a sine qua non in many areas. Particularly, the COVID-19 crisis has revealed the importance of simulation studies [6], but also limitations in terms of how fast we can develop useful models, how to interpret and build on results, and how to communicate the results to decision-makers. Therefore, new and better support for conducting simulation studies is needed, in particular

- to help simulation analysts who conduct or use simulation studies in building on their and others' results, improving the quality of their analyses, and making it easier for them to correctly interpret and reuse results.
- to facilitate the appropriate use of simulation by domain experts and decision-makers (e.g., policymakers in government, industry leaders), as a way to improve the quality of the decisions that they make.

Our goals. For comprehensive support of simulation studies, model-based approaches, in terms of formalisms, language-based approaches, logic, and meta-models play a central role, as they allow to make knowledge about a system explicit, computationally accessible, and interpretable. To move ahead, it is necessary to analyze possibilities already offered by current approaches and to identify challenges for future methodological research in model-based approaches to:

- ensure that simulation studies come with context. Context is important for helping technical experts correctly interpret results and for ensuring the quality of analyses based on the results of others. It is also important to correctly and confidently explain simulation results to decision-makers.
- improve composition and model re-use. Model re-use avoids building models from scratch. It thus saves time and, in addition, may improve analysis quality. Saving time may also increase access to simulation, increasing the number of decisions that it supports.
- support simulation automation. This would again save analyst time and may contribute to the analysis quality since automation facilitates the application of methods and thus may reduce errors in their application and, even, broaden the scope of analysis done with simulation models.
- facilitate communication both between simulation analysts and domain experts and between simulation analysts and decision-makers. We see communication as one of several limiting factors in the use of simulation for decisions. Better communication would also reduce the time required to achieve an impact.

A systematic application of model-based approaches for modeling and simulation should also allow us to harvest synergies with other areas of computer science such as language design, human-computer interaction, high-performance computing, visualization, or machine learning more effectively.

Starting point – the simulation study. Simulation studies that are aimed at developing and applying simulation models are intricate processes that combine different activities, such as model building, model refinement, model composition, calibration, analysis, and validation, and involve different types of entities, such as theories or data, and possibly products from other simulation studies. Conceptual models, requirements, assumptions, simulation models, qualitative models, simulation experiment specifications, and simulation data belong to the primary products of a simulation study, although not all of this information is explicitly represented (or documented). These ingredients can be distilled from work on modeling and simulation life cycles [22] and documentation guidelines for simulation studies [7]. For several of those, different modeling approaches have already been developed and are being applied (see Table 1).

In Table 1, simulation models refer to the class of discrete-event systems. We define requirements as expectations referring to the produced simulation outputs, which are sometimes stated in terms of constraints or inequalities. Requirements are often also given in terms of data (and a distance measure) whether or not the simulation outputs are sufficiently close to the data, or as formal expressions in a (spatial-)temporal logic which can be used for

■ **Table 1** Exemplary model-based approaches for specifying simulation models, simulation experiments, and requirements.

	Simulation Model	Simulation Experiment	Requirements
Formalisms	DEVS [31], Stochastic Petri Nets [1], Process algebras [9]	DEVS [4]	–
Programming languages, in particular DSLs	APIs [26], NetLogo [25], BioNetGen [2]	SESSL [27], NLRX [24]	FITS [17]
Logics	Ontologies for composing and reuse [18]	Ontologies of simulation algorithms [3]	Temporal logic [19]
Active objects	ABS [12, 13]	ABS runtime [30]	event-driven, cooperating processes
Metamodels	ATom3 for multi-formalism modeling [14]	meta-models for simulation experiments [29]	–

statistical model checking or runtime verification [19]. Ontologies are widely used, e.g., in the context of model composition so to map variables of different models consistently [18], in the context of simulation experiments, to classify the simulation algorithms used (and assess the approximation) [3], and referring to requirements again ontologies can be used to clarify what is meant with the variables used within a requirement. Metamodels are used for transformations to compose models defined in different modeling approaches and generate an executable model [14] or to generate simulation experiments for different modeling approaches as diverse as finite element analysis and stochastic discrete event simulation to be executed in different tools [29]. Whereas requirements refer to the model’s output, assumptions are some form of model input (as they influence how the simulation model looks like), and as such are essential for providing context and for semantically consistent model composition [18]. However, so far textual representations prevail, and little research has been done to model assumptions or theories. Not only the ingredients but also the simulation studies themselves have been subject to modeling, e.g., via workflows. Workflows have been used to support documenting and executing (also in terms of guidance) simulation studies [21, 23].

Towards Sustainable Simulation Studies. Referring to sustainability we adopted the definition given in the seminar, i.e., sustainability: continuing a simulation study into the future through support for reusing or building upon its central products, such as simulation model, data, and processes as well as the software used. We emphasized that this also applies to a simulation study currently being conducted (thus how to assist the modeler during a simulation study based on what has already been done and information gathered, see automation) and that the building upon is not restricted to modelers and domain experts but also needs to include decision-makers.

Context: The context of a simulation study is all information that helps in interpreting and reproducing the results achieved in a simulation study and makes up a large part of its documentation according to standardized guidelines [7] or its conceptual model [22, 28]. A recent empirical study shows that efforts invested in the reproducibility of a simulation study enhance its impact [10].

Effective adoption of provenance standards opens up new possibilities as essential processes, and the various sources that contributed to the results (also intermediate ones) become explicit.

Thereby, the interpretability, consistency, and reproduction of simulation studies can be enhanced, and the stored information can be used for user guidance and automation [11]. New research challenges arise, such as an unobtrusive collection of provenance information or on-the-fly abstraction to cater to the needs of different types of users. It should be noted, that the adoption of provenance standards does not alleviate the problem of how to unambiguously specify the different ingredients that make up the documentation, conceptual model, or provenance graph. However, the promise to have more comprehensive, automatic support for conducting simulation studies increases the incentive to tackle the problem.

Composition and reuse: Modeling languages should support composition in one way or the other to facilitate the development of the model and its reuse. The composition of simulation models is supported in formalisms such as DEVS by black-box composition via input and output ports, or process algebras inherently by the parallel composition of processes. Solutions that lie in between the rigidity of DEVS and the rather fluent composition of processes, and allow to flexibly define interaction points and the interface to evolve on demand, require further attention. In addition to these syntactical considerations, semantically valid reuse of simulation models still provides many challenges in modeling and simulation, although various approaches have been developed since its being stated as a grand challenge in 2002 [20]. A central question here is how the consistency of the respective context of the simulation models that shall be composed can be guaranteed. This leads us back to the question of how to represent the different ingredients that make up the context of a simulation study respectively simulation model in an unambiguous and (ideally) computationally accessible manner. A different aspect of reuse concerns variability: simulation models come in a plethora of related variants that are distinguished by differing assumptions, parameters, submodels, etc. It is essential to manage the commonality and variability embodied in these variants in a systematic manner.

Automation: Automation is one means to increase efficiency. As simulation studies are knowledge-intensive processes [5], any effort aimed at automation has to look at means for representing and evaluating the required knowledge computationally [11, 16]. Thus, model-based approaches are a central step towards at least (partly) automating simulation studies. There are various challenges referring to knowledge engineering, e.g., what information is needed to automate various activities such as model building, model refinement, calibrating, validating, or analyzing simulation models, which need to be addressed by the modeling and simulation community. Thereby, also new possibilities that are offered by applying, combining, and possibly revising methods from logic-based reasoning and machine learning have to be taken into account to automate simulation studies, e.g., in selecting methods to conduct specific experiments such as sensitivity analysis [15].

Communication: Much of the work on model-based approaches are aimed at more effective communication with different users. Despite the plethora of formalisms, domain-specific languages, etc., coming up with abstractions that are coherent to the modeling metaphors of domains, adequate for the problem, and even match the mental models of individuals remains a challenge. The research is also hampered by the difficulty to measure the effectiveness or potential impact of new methods and thus the progress made. New developments in visual analytics open up new possibilities to communicate the results

and context of simulation studies [8]. Discussing the various methods already used, their potential, limitations, and open challenges for visualization is the subject of another working group of the seminar. We expect challenges of mapping and adapting model representations of entities and processes involved in conducting a simulation study to different users and their respective needs to carry over to visualization. However, knowledge about the simulation study made explicit by model-based approaches will also further more effective visualization methods.

Conclusion. During the seminar, we identified four crucial areas of research for enhancing the sustainability of simulation studies, i.e., documenting the context of simulation studies, composition and reuse of simulation results, automation for conducting simulation studies more systematically, and communicating the results and processes of simulation studies with domain-experts and decision-makers. Central discussions revolved around whether and how concepts of software and programming languages can be or are already adopted in the field of modeling and simulation. There, the focus has been on composition, reuse, abstraction, and variability. We have still to hone in on the diversity of model-based approaches, the role they already play or might play in this endeavor, and the concrete methodological (and community) challenges associated.

References

- 1 Gianfranco Balbo. Introduction to stochastic petri nets. In *School organized by the European Educational Forum*, pages 84–155. Springer, 2000.
- 2 Michael L Blinov, James R Faeder, Byron Goldstein, and William S Hlavacek. Bionetgen: software for rule-based modeling of signal transduction based on the interactions of molecular domains. *Bioinformatics*, 20(17):3289–3291, 2004.
- 3 Mélanie Courtot, Nick Juty, Christian Knüpfer, Dagmar Waltemath, Anna Zhukova, Andreas Dräger, Michel Dumontier, Andrew Finney, Martin Golebiewski, Janna Hastings, et al. Controlled vocabularies and semantics in systems biology. *Molecular systems biology*, 7(1):543, 2011.
- 4 Joachim Denil, Stefan Klikovits, Pieter J Mosterman, Antonio Vallecillo, and Hans Vangheluwe. The experiment model and validity frame in m&s. In *Proceedings of the Symposium on Theory of Modeling & Simulation*, pages 1–12, 2017.
- 5 Claudio Di Ciccio, Andrea Marrella, and Alessandro Russo. Knowledge-intensive processes: characteristics, requirements and analysis of contemporary approaches. *Journal on Data Semantics*, 4(1):29–57, 2015.
- 6 Peter I Frazier, J Massey Cashore, Ning Duan, Shane G Henderson, Alyf Janmohamed, Brian Liu, David B Shmoys, Jiayue Wan, and Yujia Zhang. Modeling for covid-19 college reopening decisions: Cornell, a case study. *Proceedings of the National Academy of Sciences*, 119(2):e2112532119, 2022.
- 7 Volker Grimm, Jacqueline Augusiak, Andreas Focks, Béatrice M Frank, Faten Gabsi, Alice SA Johnston, Chun Liu, Benjamin T Martin, Mattia Meli, Viktoriia Radchuk, et al. Towards better modelling and decision support: documenting model development, testing, and analysis using trace. *Ecological modelling*, 280:129–139, 2014.
- 8 Jussi Hakanen, Kaisa Miettinen, and Krešimir Matković. Task-based visual analytics for interactive multiobjective optimization. *Journal of the Operational Research Society*, 72(9):2073–2090, 2021.
- 9 Jane Hillston. Process algebras for quantitative analysis. In *20th Annual IEEE Symposium on Logic in Computer Science (LICS'05)*, pages 239–248. IEEE, 2005.
- 10 Sebastian Höpfl, Jürgen Pleiss, and Nicole Radde. Bayesian hypothesis testing reveals that reproducible models in systems biology get more citations. 2022.

- 11 Pia Wilsdorf, Anja Wolpers, Jason Hilton, Fiete Haack, and Adelinde M. Uhrmacher. Automatic reuse, adaption, and execution of simulation experiments via provenance patterns. *ACM Transactions on Modeling and Computer Simulation*, 2022.
- 12 Einar Broch Johnsen, Reiner Hähnle, Jan Schäfer, Rudolf Schlatte, and Martin Steffen. ABS: A core language for abstract behavioral specification. In Bernhard K. Aichernig, Frank de Boer, and Marcello M. Bonsangue, editors, *Proc. 9th Intl. Symp. on Formal Methods for Components and Objects (FMCO 2010)*, volume 6957 of *LNCS*, pages 142–164. Springer, 2011.
- 13 Eduard Kamburjan, Stefan Mitsch, and Reiner Hähnle. A hybrid programming language for formal modeling and verification of hybrid systems. *Leibniz Transactions on Embedded Systems*, 8(1), 2022. Special Issue on Distributed Hybrid Systems.
- 14 Juan de Lara and Hans Vangheluwe. Atom 3: A tool for multi-formalism and meta-modelling. In *International Conference on Fundamental Approaches to Software Engineering*, pages 174–188. Springer, 2002.
- 15 Stefan Leye, Roland Ewald, and Adelinde M Uhrmacher. Composing problem solvers for simulation experimentation: a case study on steady state estimation. *PloS one*, 9(4):e91948, 2014.
- 16 Fabian Lorig. *Hypothesis-Driven Simulation Studies: Assistance for the Systematic Design and Conducting of Computer Simulation Experiments*. Springer Vieweg, Wiesbaden, 2019.
- 17 Fabian Lorig, Colja A Becker, and Ingo J Timm. Formal specification of hypotheses for assisting computer simulation studies. In *Proceedings of the Symposium on Theory of Modeling & Simulation*, pages 1–12, 2017.
- 18 Maxwell Lewis Neal, Matthias König, David Nickerson, Göksel Mısırlı, Reza Kalbasi, Andreas Dräger, Koray Atalag, Vijayalakshmi Chelliah, Michael T Cooling, Daniel L Cook, et al. Harmonizing semantic annotations for computational models in biology. *Briefings in bioinformatics*, 20(2):540–550, 2019.
- 19 Laura Nenzi, Luca Bortolussi, Vincenzo Ciancia, Michele Loreti, and Mieke Massink. Qualitative and quantitative monitoring of spatio-temporal properties. In *Runtime Verification*, pages 21–37. Springer, 2015.
- 20 C Michael Overstreet, Richard E Nance, and Osman Balci. Issues in enhancing model reuse. In *International Conference on Grand Challenges for Modeling and Simulation. San Antonio, Texas, USA*, 2002.
- 21 Judicaël Ribault and Gabriel Wainer. Using workflows and web services to manage simulation studies (wip). In *Proceedings of the 2012 Symposium on Theory of Modeling and Simulation-DEVS Integrative M&S Symposium*, pages 1–6, 2012.
- 22 Stewart Robinson. Conceptual modeling for simulation: issues and research requirements. In *Proceedings of the 2006 winter simulation conference*, pages 792–800. IEEE, 2006.
- 23 Andreas Ruschinski, Tom Warnke, and Adelinde M Uhrmacher. Artifact-based workflows for supporting simulation studies. *IEEE Transactions on Knowledge and Data Engineering*, 32(6):1064–1078, 2019.
- 24 Jan Salecker, Marco Sciaini, Katrin M Meyer, and Kerstin Wiegand. The nlr x r package: A next-generation framework for reproducible netlogo model analyses. *Methods in Ecology and Evolution*, 10(11):1854–1863, 2019.
- 25 Seth Tisue and Uri Wilensky. Netlogo: A simple environment for modeling complexity. In *International conference on complex systems*, volume 21, pages 16–21. Boston, MA, 2004.
- 26 Yentl Van Tendeloo and Hans Vangheluwe. The modular architecture of the python (p) devs simulation kernel. In *Proceedings of the 2014 Symposium on Theory of Modeling and Simulation-DEVS*, pages 387–392, 2014.
- 27 Tom Warnke and Adelinde M Uhrmacher. Complex simulation experiments made easy. In *2018 Winter Simulation Conference (WSC)*, pages 410–424. IEEE, 2018.

- 28 Pia Wilsdorf, Fiete Haack, and Adelinde M Uhrmacher. Conceptual models in simulation studies: Making it explicit. In *2020 Winter Simulation Conference (WSC)*, pages 2353–2364. IEEE, 2020.
- 29 Pia Wilsdorf, Jakob Heller, Kai Budde, Julius Zimmermann, Tom Warnke, Christian Haubelt, Dirk Timmermann, Ursula van Rienen, and Adelinde M Uhrmacher. A model-driven approach for conducting simulation experiments. *Applied Sciences*, 12(16):7977, 2022.
- 30 Peter Y. H. Wong, Elvira Albert, Radu Muschevici, José Proença, Jan Schäfer, and Rudolf Schlatte. The ABS tool suite: modelling, executing and analysing distributed adaptable object-oriented systems. *STTT*, 14(5):567–588, 2012.
- 31 Bernard P Zeigler, Alexandre Muzy, and Ernesto Kofman. *Theory of modeling and simulation: discrete event & iterative system computational foundations*. Academic press, 2018.

Participants

- Philipp Andelfinger
Universität Rostock, DE
- Luca Bortolussi
University of Trieste, IT
- Wentong Cai
Nanyang TU – Singapore, SG
- Christopher Carothers
Rensselaer Polytechnic Institute –
Troy, US
- Rodrigo Castro
University of Buenos Aires, AR
- Joachim Denil
University of Antwerp, BE
- Jérôme Feret
ENS – Paris, FR
- Peter Frazier
Cornell University – Ithaca, US
- Reiner Hähnle
TU Darmstadt, DE
- Dong (Kevin) Jin
University of Arkansas –
Fayetteville, US
- Franziska Klügl
University of Örebro, SE
- Till Köster
Universität Rostock, DE
- Michael Lees
University of Amsterdam, NL
- Jason Liu
Florida International University –
Miami, US
- Margaret Loper
Georgia Institute of Technology –
Atlanta, US
- Fabian Lorig
Malmö University, SE
- Bertram Ludäscher
University of Illinois at
Urbana-Champaign, US
- Kresimir Matkovic
VRVis – Wien, AT
- Laura Nenzi
University of Trieste, IT
- David M. Nicol
University of Illinois –
Urbana Champaign, US
- Alessandro Pellegrini
University of Rome “Tor
Vergata”, IT
- Niki Popper
Technische Universität Wien, AT
- Caitlin Ross
Kitware – Clifton Park, US
- Cristina Ruiz-Martin
Carleton University –
Ottawa, CA
- Bernhard Rumpe
RWTH Aachen, DE
- Susan Sanchez
Naval Postgrad. School –
Monterey, US
- Nadja Schlungbaum
Universität Rostock, DE
- Peter Sloot
University of Amsterdam, NL
- Claudia Szabo
University of Adelaide, AU
- Wen Jun Tan
Nanyang TU – Singapore, SG
- Adelinde M. Uhrmacher
Universität Rostock, DE
- Gabriel A. Wainer
Carleton University –
Ottawa, CA
- Pia Wilsdorf
Universität Rostock, DE
- Verena Wolf
Universität des Saarlandes –
Saarbrücken, DE



Foundations for a New Perspective of Understanding Programming

Madeline Endres^{*1}, André Brechmann^{†2}, Bonita Sharif^{†3},
Westley Weimer^{†4}, and Janet Siegmund^{†5}

1 University of Michigan – Ann Arbor, US. endremad@umich.edu

2 Leibniz-Institut für Neurobiologie – Magdeburg, DE.
brechmann@lin-magdeburg.de

3 University of Nebraska – Lincoln, US. bsharif@unl.edu

4 University of Michigan – Ann Arbor, US. weimerw@umich.edu

5 TU Chemnitz, DE. janet.siegmund@informatik.tu-chemnitz.de

Abstract

Software is created by people who think, feel, and express themselves to one another and their computers. For a long time, researchers have investigated how people read and write code on their computers and talk about code with one another. This way, researchers identified skills, education, and practices necessary to acquire expertise and perform software development duties. While these investigations are valuable, we have yet to devise and validate a scientific theory of *program comprehension*, which would be an important step in designing support for developers that is tailored to their cognitive needs. To succeed, we need techniques to shed more light on how programmers think. To this end, we need to look beyond computer science research.

Specifically, in the field of psychology and cognitive neuroscience, considerable progress has been made in building theories of cognitive processes. Important enabling technologies include eye tracking, functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and functional near infrared spectroscopy (fNIRS). These methods have revolutionized the understanding of cognitive processes and are routinely used in non-computing disciplines. Such techniques have the potential to also modernize classic approaches to program comprehension research by informing new experimental designs. However, the use of such technologies to study program comprehension is recent, and many of the challenges of this interdisciplinary field remain unexplored.

This report documents the program and the outcomes of Dagstuhl Seminar 22402, “Foundations for a New Perspective of Understanding Programming”, which explores these challenges. In total, 23 on-site participants attended the seminar along with two virtual keynote speakers. Participants engaged in intensive collaboration, including discussing past and current research, identifying gaps in the literature, and proposing future directions for improving the state of the art in program comprehension research.

Seminar October 3–7, 2022 – <http://www.dagstuhl.de/22402>

2012 ACM Subject Classification General and reference → General literature; General and reference → Empirical studies; Software and its engineering → Software design engineering; Human-centered computing → User studies

Keywords and phrases Programming Methodology, Programming Education, Program Comprehension, Neuro-imaging, Eye Tracking, Human Cognition, Human Computer Interaction, Software Engineering, Human Factors

Digital Object Identifier 10.4230/DagRep.12.10.61

* Editorial Assistant / Collector

† Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Foundations for a New Perspective of Understanding Programming, *Dagstuhl Reports*, Vol. 12, Issue 10, pp. 61–83
Editors: Madeline Endres, André Brechmann, Bonita Sharif, Westley Weimer, and Janet Siegmund



DAGSTUHL
REPORTS

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

1 Executive Summary


Madeline Endres (University of Michigan – Ann Arbor, US)

André Brechmann (Leibniz-Institut für Neurobiologie – Magdeburg, DE)

Bonita Sharif (University of Nebraska – Lincoln, US)

Westley Weimer (University of Michigan – Ann Arbor, US)

Janet Siegmund (TU Chemnitz, DE)

License  Creative Commons BY 4.0 International license

© Madeline Endres, André Brechmann, Bonita Sharif, Westley Weimer, Janet Siegmund

The goal of the seminar *Foundations for a New Perspective of Understanding Programming* was to address how to modernize the perspective on program comprehension and thus make progress regarding our understanding of it. We focused on two challenges: First, we discussed how to provide guidance on addressing methodological issues in interdisciplinary program comprehension research. Second, we aimed at defining a unifying enumeration of the dimensions of a neuroscientific perspective on program comprehension, such that researchers are able to more systematically investigate gaps in the literature.

Through individual participant presentations and the resulting group discussions, we identified several relevant aspects that we discussed further in dedicated working groups. These included discussing how to make better use of eye tracking (Section 4.1), identifying the role of readability for program comprehension (Section 4.2), and considering how machine learning could help to develop a model of program comprehension (Section 4.3). These working groups helped us address the first goal of the seminar by providing guidance on program comprehension research methodologies and identifying potential next steps. To conclude the seminar, participants discussed a possible taxonomy for program comprehension research (Section 5.1). This taxonomy identifies commonalities and differences across a broad research area, and addresses our second goal of helping to unify the program comprehension research space; it can serve as a starting point for future researchers to build on to develop an understanding of program comprehension. Also, for researchers entering this field, it is a first glimpse of the complexity of understanding and researching program comprehension.

Beyond identifying research problems in program comprehension research, the many collaborative sessions at this seminar generated numerous potential multi-institutional and interdisciplinary collaborations. We hope that, by making progress on the program comprehension research challenges, we help bring this new research direction one step closer to becoming standard in programming research and disseminating it to a wider audience.

2 Table of Contents

Executive Summary

Madeline Endres, André Brechmann, Bonita Sharif, Westley Weimer, Janet Siegmund 62

Overview of Talks

Brains on Code: A Neuroscientific Foundation on Program Comprehension <i>Sven Apel</i>	65
Shared Intentionality in Program Comprehension <i>Andrew Begel</i>	65
Insights into Program Comprehension with EEG and Eye tracking <i>Annabelle Bergum</i>	66
Linking fMRI research on sequential processing and category learning to understanding programming <i>André Brechmann</i>	66
Studying Eye Movements During Code Reading <i>Teresa Busjahn</i>	67
Using Physiological Measures to Identify Cognitive States <i>Martha E. Crosby and Jan Stelovsky</i>	67
Finding Flocus: Using Logs Data to Identify When Software Engineers Experience Flow or Focused Work <i>Sarah D'Angelo</i>	68
Tracking Eye Movements in Programming <i>Andrew Duchowski</i>	68
How Do New Programmers Understand Programs? <i>Madeline Endres</i>	68
Measuring Cognitive Effort During Programming: current methods and the cognitive offloading tools of the future <i>Sarah Fakhoury</i>	69
Sensing in the Wild: Increasing Productivity by Sensing Interruptibility <i>Thomas Fritz</i>	69
Predicting human reading comprehension from eye movements <i>Lena A. Jäger</i>	70
Investigating Programming Expertise With Event-Related Desynchronization <i>Timothy Kluthe</i>	71
Computational NeuroSE <i>Takatomi Kubo</i>	71
Safe and Secure Software Engineering – A Program Comprehension Perspective <i>Jürgen Mottok</i>	72
The logical reasoning network encodes algorithms even in programming novices reading plain-language description of programming functions <i>Yun-Fei Liu</i>	72

An Eye Tracking Analysis of Tracing and Debugging Collaboration among Programming Pairs <i>Maria Mercedes T. Rodrigo</i>	73
Exploring Common Code Reading Strategies in Debugging <i>Maria Mercedes T. Rodrigo and Christine Lourrine S. Tablatin</i>	73
Detecting Expertise in Developer Eye Movements <i>Bonita Sharif</i>	74
How does the Brain Change during Programming Learning? <i>Janet Siegmund</i>	74
Evidence-Based Programming and the “Quorum Project” <i>Andreas Stefik</i>	75
Making Novices More Like Experts? <i>Westley Weimer</i>	75
Could we please be a bit more explicit? <i>Marvin Wyrich</i>	75
Working Groups	
Eye Tracking Best Practices & Ideas <i>Teresa Busjahn, Martha E. Crosby, Maria Mercedes T. Rodrigo, Christine Lourrine S. Tablatin, Westley Weimer</i>	76
Readability <i>Andrew Begel, Annabelle Bergum, Madeline Endres, Sarah Fakhoury, Timothy Kluthe, Yun-Fei Liu, Bonita Sharif, Jan Stelovsky, Marvin Wyrich</i>	78
Statistics and machine learning to predict and model program comprehension <i>Sven Apel, André Brechman, Janet Siegmund</i>	79
Open problems	
A Taxonomy of Program Comprehension	81
Participants	83
Remote Participants	83

3 Overview of Talks

3.1 Brains on Code: A Neuroscientific Foundation on Program Comprehension

Sven Apel (Universität des Saarlandes – Saarbrücken, DE)

License  Creative Commons BY 4.0 International license
© Sven Apel

Research on program comprehension has a fundamental limitation: program comprehension is a cognitive process that cannot be directly observed, which leaves considerable room for misinterpretation, uncertainty, and confounders. In Brains On Code, we are developing a neuroscientific foundation of program comprehension. Instead of merely observing whether there is a difference regarding program comprehension (e.g., between two programming methods), we aim at precisely and reliably determining the key factors that cause the difference. This is especially challenging as humans are the subjects of study, and inter-personal variance and other confounding factors obfuscate the results.

The key idea of Brains On Code is to leverage established methods from cognitive neuroscience to obtain insights into the underlying processes and influential factors of program comprehension. Brains On Code pursues a multimodal approach that integrates different neuro-physiological measures as well as a cognitive computational modeling approach to establish the theoretical foundation. This way, Brains On Code lays the foundations of measuring and modeling program comprehension and offers substantial feedback for programming methodology, language design, and education. With Brains On Code, addressing longstanding foundational questions such as “How can we reliably measure program comprehension?”, “What makes a program hard to understand?”, and “What skills should programmers have?” comes into reach. Brains On Code does not only help answer these questions, but also provides an outline for applying the methodology beyond program code (models, specifications, requirements, etc.).

3.2 Shared Intentionality in Program Comprehension

Andrew Begel (Carnegie Mellon University – Pittsburgh, US)

License  Creative Commons BY 4.0 International license
© Andrew Begel

Observing communication is a revealing way to indicate comprehension about code at many different abstraction levels. Using an analytic lens from linguistics, we can precisely describe this communication and thus enable us to make inferences about a person’s comprehension of a program. Not only are the speaker’s actions important, but the agency of the listener is vital to establishing a desired states of shared attention (i.e., both parties are thinking about the same thing) and shared intentionality (i.e., the recursive knowledge that they both know the other is thinking about the same thing they are). When both speaker and listener communicate together, they can begin to comprehend code and take actions on it as a single distributed cognitive unit. The pair’s joint knowledge can be used to execute changes to the code that may have been difficult or impossible for each of them apart.

The effectiveness of this kind of communication is not robust however, when one or both members of the pair identify with physical or cognitive disabilities, e.g., a programmer is blind or low vision, or another has ADHD or dyslexia. In our research, we employ AI techniques in

computer vision, speech recognition, NLP, and physiological sensors to interpret, translate, and convey information between speaker and listener. This increases the likelihood that people of mixed abilities can successfully communicate about code, achieving a desired state of shared intentionality that illustrates their joint distributed comprehension and enables them to efficiently make changes together.

3.3 Insights into Program Comprehension with EEG and Eye tracking

Annabelle Bergum (Universität des Saarlandes – Saarbrücken, DE)

License © Creative Commons BY 4.0 International license
© Annabelle Bergum

Main reference Norman Peitek, Annabelle Bergum, Maurice Rekrut, Jonas Mucke, Matthias Nadig, Chris Parnin, Janet Siegmund, Sven Apel: “Correlates of programmer efficacy and their link to experience: a combined EEG and eye-tracking study”, in Proc. of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2022, Singapore, Singapore, November 14-18, 2022, pp. 120–131, ACM, 2022.

URL <https://doi.org/10.1145/3540250.3549084>

One of the main research questions in our research group is “How can we reliably measure program comprehension?”. To answer this question, fMRI and eye tracking studies were conducted. When I joined the team, in 2021, we enlarged the research field from our research group from fMRI to EEG. Since then, we conducted two EEG studies. The first study included the challenge of the new methodology which comes along with using EEG instead of fMRI. In the second study, we took a closer look at the influence of the baseline. We thereby adapted our study design to incorporate four different baselines. Therefore, we can afterwards investigate the effect of the baseline within one study, eliminating a lot of other influencing factors.

3.4 Linking fMRI research on sequential processing and category learning to understanding programming

André Brechmann (Leibniz-Institut für Neurobiologie – Magdeburg, DE)

License © Creative Commons BY 4.0 International license
© André Brechmann

When starting to work on understanding program comprehension using fMRI ten years ago in collaboration with Janet Siegmund, there was no blueprint how to approach the topic experimentally. First, I had to learn that empirical research was not very abundant at that time even though programming is such an important economic and societal topic. In the previous years, however, we have seen much progress in empirical research on programming, including brain imaging. Now it is about time for in depth discussions on how to pursue the topic further by teaming up with interested researchers from different disciplines and to start developing a theory of programming. In my talk (see attached slides) I contribute my neuroscientific perspective on program comprehension and discuss my view of sequential processing and cognitive sequencing as key component of programming and how to deal with the dynamics and individuality of program comprehension based on my experience from studying the dynamics of rule-based category learning.

3.5 Studying Eye Movements During Code Reading

Teresa Busjahn (Hochschule für Technik und Wirtschaft Berlin – Berlin, DE)

License © Creative Commons BY 4.0 International license
© Teresa Busjahn

Main reference Teresa Busjahn: “Empirical analysis of eye movements during code reading: evaluation and development of methods”, 2021.

URL <https://doi.org/10.17619/UNIPB/1-1118>

Studying eye movements during reading provides valuable insights into natural-language text comprehension and also lends itself to application in program comprehension. Using English and Java as exemplary languages, differences can be found between natural-language text and code reading, as well as in how novice and expert programmers read code. For instance, when reading natural-language text, a larger part of the text is looked at directly than when reading code. Moreover, during code reading, expert programmers attend to the main-method much sooner than novices. However, this line of research also brings about methodological challenges like event detection and correction of spatial errors. Addressing these is an ongoing effort.

3.6 Using Physiological Measures to Identify Cognitive States

Martha E. Crosby (University of Hawaii at Manoa – Honolulu, US) and Jan Stelovsky (University of Hawaii at Manoa – Honolulu, US)

License © Creative Commons BY 4.0 International license
© Martha E. Crosby and Jan Stelovsky

The way that humans interact and absorb information delivered through technology is of interest to researchers in many fields. The accurate assessment of cognitive states such as arousal, fatigue, stress, task difficulty is essential to identifying and defining cognitive processes and testing models of how cognitive processes operate and interact. The individual and the situation can affect the measurement of cognitive states from a single type of sensor. Measurements from multiple sensors, when combined, produce a more robust measurement of cognitive states (CS) such as mental overload. Adaptive filtering or other techniques of CS can then be used to potentially find misconceptions and potentially improve task performance. For the last several years, we have designed and executed experiments about individual differences of the way people perceive, search, and understand information presented in various multimedia environments. We have used a suite of passive physiological sensors (eye fixations, skin conductivity, body temperature, heart wave form, electroencephalography, and pressures on mouse) to better understand the processes that facilitate seeking, filtering, and shaping relevant information during tasks such as understanding computer programs. To understand and solve problems, programmers must have a level of program comprehension established. There are many variations and levels of program comprehension, dependent on individual programmers as well the specific code itself. Our research focuses on deriving changes in cognitive state information from the patterns of data acquired from the user from physiological sensors. If companies incorporate technology already available in Augmented Reality glasses and cell phones, CS information could potentially be used to give feedback in classroom or industry settings. For example, if many programmers show confusion or are misled by the code being reviewed, it may indicate there is a need to improve various aspects of it such as documentation or style.

3.7 Finding Flocus: Using Logs Data to Identify When Software Engineers Experience Flow or Focused Work

Sarah D’Angelo (Google – Mountain View, US)

License © Creative Commons BY 4.0 International license

© Sarah D’Angelo

Joint work of Sarah D’Angelo, Adam Brown, Ben Holtz, Ciera Jaspán, Collin Green

The concept of flow has been studied for decades across a wide variety of contexts from work to hobbies, and is a critical aspect of engineering productivity, however non-disruptively measuring flow has remained difficult. In this work, we take a mixed methods approach to understanding and measure how software engineers experience flow. We introduce a logs based metric called “flocus” that leverages machine learning and a comprehensive collection of logs data to identify periods of related actions (indicating focused behavior), and validate this metric against self-reported time in flow or focus using diary data and quarterly survey data. Our results indicate that we can determine when software engineers at a large technology company experience flow or focus using flocus. Extending this approach to incorporate other signals such as physiological data or eye tracking has the potential to get us closer to measuring a flow state.

3.8 Tracking Eye Movements in Programming

Andrew Duchowski (Clemson University, US)

License © Creative Commons BY 4.0 International license

© Andrew Duchowski

Main reference Krzysztof Krejtz, Andrew T. Duchowski, Katarzyna Wisiecka, Izabela Krejtz: “Entropy of Eye Movements While Reading Code or Text”, in Proc. of the 10th IEEE/ACM International Workshop on Eye Movements in Programming, EMIP@ICSE 2022, Pittsburgh, PA, USA, May 18-24, 2022, pp. 8–14, IEEE, 2022.

URL <https://ieeexplore.ieee.org/document/9808970>

The keynote, while focused on eye movements in programming, covers a wide variety topics, including: (1) basics of eye movements, (2) basic metrics, (3) advanced metrics, (4), cognitive load, (5) eye movements in programming, and (6) future challenges.

3.9 How Do New Programmers Understand Programs?

Madeline Endres (University of Michigan – Ann Arbor, US)

License © Creative Commons BY 4.0 International license

© Madeline Endres

Joint work of Madeline Endres, Zachary Karas, Xiaosu Hu, Ioulia Kovelman, Madison Fransher, Priti Shah, Westley Weimer

Main reference Madeline Endres, Zachary Karas, Xiaosu Hu, Ioulia Kovelman, Westley Weimer: “Relating Reading, Visualization, and Coding for New Programmers: A Neuroimaging Study”, in Proc. of the 43rd IEEE/ACM International Conference on Software Engineering, ICSE 2021, Madrid, Spain, 22-30 May 2021, pp. 600–612, IEEE, 2021.

URL <https://doi.org/10.1109/ICSE43902.2021.00062>

Understanding how novices reason about coding at a neurological level has implications for training the next generation of software engineers. I first briefly talk about our work using neuroimaging (fNIRS) to measure the neural activity associated with introductory programming. In this work, we relate brain activity when coding to that of reading natural

language or spatial visualization. In contrast to some studies with more expert programmers, we find that while programming, reading, and spatial visualization are all neurologically distinct for novices, there are more significant differences between prose and coding than between spatial visualization and coding. We also find a neural activation pattern predictive of programming performance 11 weeks later. I conclude with a discussion of future education-related directions I hope neuroimaging research explores going forward, including understanding how external factors such as native natural language, or learning disabilities (e.g., Dyslexia) impact program comprehension and learning.

References

- 1 Madeline Endres, Zachary Karas, Xiaosu Hu, Ioulia Kovelman, and Westley Weimer. *Relating Reading, Visualization, and Coding for New Programmers: A Neuroimaging Study*. International Conference on Software Engineering (ICSE), 2021.
- 2 Madeline Endres, Madison Fransher, Priti Shah, and Westley Weimer. *To Read or To Rotate? Comparing the Effects of Technical Reading Training and Spatial Skills Training on Novice Programming Ability*. Foundations of Software Engineering (ESEC/FSE), 2021.

3.10 Measuring Cognitive Effort During Programming: current methods and the cognitive offloading tools of the future

Sarah Fakhoury (Microsoft Research (MSR) – Redmond, US)

License © Creative Commons BY 4.0 International license
© Sarah Fakhoury

Programming tasks require inherent cognitive load, but the design of the tools and languages a programmer uses to complete their task can either increase mental burden, or optimize for it. I briefly talk about our work using simultaneous fNIRS and eye tracking to measure cognitive effort caused by programming antipatterns in the context of bug localization tasks. We observe that we cannot make assumptions about cognitive effort based on traditional metrics like correctness and time on task alone. Novel methods give us the ability to develop and test theories of how and why various factors influence comprehension.

Next I briefly touch on pain points related to tooling that the community can make joint strides in. Finally, I raise questions about the future of programming comprehension research in the age of AI coding assistants that aim to serve as cognitive offloading tools.

3.11 Sensing in the Wild: Increasing Productivity by Sensing Interruptibility

Thomas Fritz (Universität Zürich, CH)

License © Creative Commons BY 4.0 International license
© Thomas Fritz

The modern workplace is more demanding than ever before. Software developers have to work on a wide variety of cognitively demanding tasks, face constant context switches, work in distributed teams, and have blurred work-life boundaries. What does it mean to be productive in this context, and how can we best support developers in staying focused? To address these questions and better understand developers' cognitive and emotional states,

in our research, we employ a variety of sensors in a range of studies, from small controlled lab experiments to 8-week-long field studies. The results show that while it is not always feasible to gather fine-grained biometric data for highly accurate classifications, even with coarser-grained data, we can develop approaches that take into account developers' states and help boost their productivity.

3.12 Predicting human reading comprehension from eye movements

Lena A. Jäger (Universität Zürich, CH)

License © Creative Commons BY 4.0 International license

© Lena A. Jäger

Joint work of Lena A. Jäger, David R. Reich, Silvia Makowski, Ahmed Abdelwahab, Niels Landwehr, Tobias Scheffer, Paul Prasse, Frank Goldhammer

Main reference David Robert Reich, Paul Prasse, Chiara Tschirner, Patrick Haller, Frank Goldhammer, Lena A. Jäger: “Inferring Native and Non-Native Human Reading Comprehension and Subjective Text Difficulty from Scanpaths in Reading”, in Proc. of the ETRA 2022: Symposium on Eye Tracking Research and Applications, Seattle, WA, USA, June 8 – 11, 2022, pp. 23:1–23:8, ACM, 2022.

URL <https://doi.org/10.1145/3517031.3529639>

Eye movements in reading have long been known to reflect cognitive processes involved in reading. Abundant evidence from cognitive psychology and psycholinguistics demonstrates that, among many other factors, reading comprehension is a significant predictor of fixation durations. However, it has turned out that, conversely, predicting reading comprehension from eye movements is much more challenging. In my talk, I will present two different approaches to predict reading comprehension from eye-tracking data. I will first present a psychologically motivated generative model of eye movements in reading from which we derive a discriminative Fisher kernel to predict reading comprehension. Second, I will present a neural sequence model that processes raw scanpaths along with the read text and predicts the reader's comprehension. The proposed models outperform the previous state-of-the-art methods. Finally, I will discuss the current challenges that models aiming to predict reading comprehension from eye movements are facing.

References

- 1 David R. Reich, Paul Prasse, Chiara Tschirner, Patrick Haller, Frank Goldhammer, and Lena A. Jäger. *Inferring Native and Non-Native Human Reading Comprehension and Subjective Text Difficulty from Scanpaths in Reading*, ETRA 2022.
- 2 Silvia Makowski, Lena A. Jäger, Ahmed Abdelwahab, Niels Landwehr, and Tobias Scheffer. *A discriminative model for identifying readers and assessing text comprehension from eye movements*. ECML-PKDD 2018. In Brefeld et al. (eds). *Machine Learning and Knowledge Discovery in Databases*, Springer, Cham, Switzerland, 2019.

3.13 Investigating Programming Expertise With Event-Related Desynchronization

Timothy Kluthe (University of Nevada – Las Vegas, US)

License © Creative Commons BY 4.0 International license
© Timothy Kluthe

Joint work of Igor Crk, Timothy Kluthe, Andreas Stefik

Main reference Igor Crk, Timothy Kluthe, Andreas Stefik: “Understanding Programming Expertise: An Empirical Study of Phasic Brain Wave Changes”, *ACM Trans. Comput. Hum. Interact.*, Vol. 23(1), pp. 2:1–2:29, 2016.

URL <https://doi.org/10.1145/2829945>

With the recent resurgence of interest in applying cognitive science technologies in the study of computer science, we present some of our previous research on the topic. Using electroencephalography and Event-Related Desynchronization measurements, we investigated several sub-bands associated with various cognitive subprocesses and how they differ in programmers of varying experience levels when working through programming comprehension tasks. Currently, we are working on making data science accessible for everyone. To achieve this goal, we will be gathering empirical evidence on design and syntax choices using typical human factors methodologies such as surveys and usability studies. In addition to this, we will be designing neuroscience-based studies which look at the same problems from a different angle.

3.14 Computational NeuroSE

Takatomi Kubo (Nara Institute of Science and Technology, JP)

License © Creative Commons BY 4.0 International license
© Takatomi Kubo

Joint work of Takatomi Kubo, Takeshi D. Itoh, Yoshiharu Ikutani

Main reference Yoshiharu Ikutani, Takatomi Kubo, Satoshi Nishida, Hideaki Hata, Kenichi Matsumoto, Kazushi Ikeda, Shinji Nishimoto: “Expert Programmers Have Fine-Tuned Cortical Representations of Source Code”, *eNeuro*, Vol. 8(1), Society for Neuroscience, 2021.

URL <https://doi.org/10.1523/ENEURO.0405-20.2020>

NeuroSE is a research field in software engineering (SE) that makes use of neuroscientific methods and knowledge to better understand the software development process, as well as the software system itself as the outcome of the process. The neuroscience is expected to contribute to a better understanding of the SE process and to affect the software system itself positively as a consequence. The NeuroSE field is characterized by collaboration of researchers from various disciplines, and still relatively young. In the next decade, NeuroSE should advance to the next stage. One of the missing pieces in the current NeuroSE is “computational approach”.

In the neuroscience, computational neuroscience was advocated and is a branch of neuroscience which employs mathematical models, computer simulations and theoretical analyses with abstractions of the brain to understand the principles that govern the development, structure, and functions of the nervous system. In the history of computational neuroscience, David Marr offered a distinction of three levels: (i) computational theory, (ii) representation and algorithm, and (iii) hardware implementation.

These concepts should have high affinity to SE or its related fields since these terms are often used in them. From such background, the emergence of Computational NeuroSE should be natural. Computational NeuroSE will lead to unveiling the algorithm in the brain to understand the algorithms in the external world. I will also mention the potential interaction between Computational NeuroSE and AI4code/ML4code in the presentation.

3.15 Safe and Secure Software Engineering – A Program Comprehension Perspective

Jürgen Mottok (OTH Regensburg, Germany)

License © Creative Commons BY 4.0 International license
© Jürgen Mottok

Joint work of Lisa Grabinger, Florian Hauser, Jürgen Mottok

What is the difference between experts and novices in software engineering disciplines like requirements engineering, analysis and design, and implementation? Functional Safety and IT-Security demand highly qualified software engineers with a deep conceptual understanding of e.g. strength and weakness of programming techniques.

We are pursuing different experimental settings, including different reading techniques, scaffolding approaches, mixed model approaches or EMME to evaluate which techniques are useful to guide the transformation process from a novice to an expert in software engineering.

We are interested in replication studies and can provide an Eye-Tracking laboratory with 14 Tobii Pro Spectrum (600Hz) for field studies.

3.16 The logical reasoning network encodes algorithms even in programming novices reading plain-language description of programming functions

Yun-Fei Liu (Johns Hopkins Univ. – Baltimore, US)

License © Creative Commons BY 4.0 International license
© Yun-Fei Liu

Joint work of Yun-Fei Liu, Marina Bedny

Main reference Yun-Fei Liu, Judy Kim, Colin Wilson, Marina Bedny: “Computer code comprehension shares neural resources with formal logical inference in the fronto-parietal network”, *eLife*, Vol. 9, p. e59340, eLife Sciences Publications, Ltd, 2020.

URL <https://doi.org/10.7554/eLife.59340>

In a previous functional MRI study, we found the fronto-parietal logical reasoning network is engaged during code comprehension in programming experts. Additionally, we can use support vector machine (SVM) to classify FOR and IF algorithms using the spatial activation patterns in the regions in this network. In an ongoing project, we ask whether the fronto-parietal system processes the semantic content (i.e., the algorithms) regardless of the specific syntax in which the algorithms are presented – even in programming novices. During the MRI scan, Programming-naïve students read “pseudocode” passages, which are natural language descriptions of Python functions used in the previous expert study. Preliminary findings suggest that pseudocode reading also engaged the fronto-parietal logical reasoning network, and that FOR and IF algorithms (expressed with plain language rather than Python code) were also decodable in this network. Overall, the data suggest the logical reasoning system is recycled for code comprehension. This logical reasoning system represents algorithms all along, independent of the syntax of specific programming languages, and even in individuals with 0 programming experience.

3.17 An Eye Tracking Analysis of Tracing and Debugging Collaboration among Programming Pairs

Maria Mercedes T. Rodrigo (Ateneo de Manila University – Quezon City, PH)

License © Creative Commons BY 4.0 International license

© Maria Mercedes T. Rodrigo

Joint work of Maria Mercedes T. Rodrigo, Maureen Villamor

Main reference Maureen Villamor, Ma. Mercedes T. Rodrigo: “Gaze collaboration patterns of successful and unsuccessful programming pairs using cross-recurrence quantification analysis”, *Res. Pract. Technol. Enhanc. Learn.*, Vol. 14(1), p. 25, 2019.

URL <https://doi.org/10.1186/s41039-019-0118-z>

We make use of Cross-Recurrence Quantification Analysis (CRQA) to characterize tracing and debugging collaboration behaviors among programming students engaged in a pair programming task. We describe how successful and unsuccessful pairs significantly differ in their gaze patterns. We also describe how prior knowledge and acquaintanceship affect pair success.

References

- 1 Maureen Villamor, Ma. Mercedes T. Rodrigo: *Gaze collaboration patterns of successful and unsuccessful programming pairs using cross-recurrence quantification analysis*. *Res. Pract. Technol. Enhanc. Learn.* 14(1): 25 (2019)
- 2 Maureen Villamor, Ma. Mercedes T. Rodrigo: *Predicting Successful Collaboration in a Pair Programming Eye Tracking Experiment*. UMAP (Adjunct Publication) 2018: 263-268

3.18 Exploring Common Code Reading Strategies in Debugging

Maria Mercedes T. Rodrigo (Ateneo de Manila University – Quezon City, PH) and Christine Lourrine S. Tablatin (Pangasian State University, PH and Ateneo de Manila University – Quezon City, PH)

License © Creative Commons BY 4.0 International license

© Maria Mercedes T. Rodrigo and Christine Lourrine S. Tablatin

Main reference Christine Lourrine S. Tablatin, Maria Mercedes T. Rodrigo: “Identifying Code Reading Strategies in Debugging using STA with a Tolerance Algorithm”, *APSIPA Transactions on Signal and Information Processing*, Vol. 11(1), pp. –, 2022.

URL <https://doi.org/10.1561/116.00000040>

The purpose of this study was to identify the common code reading strategies of the high and low performing students engaged in a debugging task. Using Scanpath Trend Analysis (STA) with a tolerance on eye tracking data, common scanpaths of high and low performing students were generated. The common scanpaths revealed differences in the code reading patterns and code reading strategies of high and low performing students. High performing students follow a bottom-up code reading strategy when debugging complex programs with logical and semantic errors. A top-down code reading strategy is employed when debugging programs with simple control structures, few lines of code, and simple error types. These results imply that high performing students use flexible debugging strategies based on the program structure. The generated common scanpaths of the low performing students, on the other hand, showed erratic code reading patterns, implying that no obvious code reading strategy was applied. The identified code reading strategies of the high performing students could be explicitly taught to low performing students to help improve their debugging performance.

3.19 Detecting Expertise in Developer Eye Movements

Bonita Sharif (University of Nebraska – Lincoln, US)

License © Creative Commons BY 4.0 International license
© Bonita Sharif

Joint work of Bonita Sharif, Salwa Aljehane, Jonathan Maletic

Main reference Salwa Aljehane, Bonita Sharif, Jonathan I. Maletic: “Determining Differences in Reading Behavior Between Experts and Novices by Investigating Eye Movement on Source Code Constructs During a Bug Fixing Task”, in Proc. of the 2021 Symposium on Eye Tracking Research and Applications, ETRA 2020, Virtual Event, Germany, May 25-27, 2021, Short Papers, pp. 30:1–30:6, ACM, 2021.

URL <https://doi.org/10.1145/3448018.3457424>

What constitutes developer expertise? Could expertise be determined based on the nature of the task rather than by how many years a developer worked in the field? It is also more likely that expertise is not necessarily a binary decision: expert vs. non-expert, as there may be variations of expertise. We start to address these questions by investigating developer expertise prediction solely from a biometric data source namely, eye fixation data on source code elements. Which clustering similarity metrics work best for determining developer expertise on eye fixation sequences? What should the level of granularity be when looking at fixations and their sequences? Can we train a model to predict expertise with high confidence given eye tracking data for a particular task? One problem facing this research is the lack of enough eye tracking datasets on a variety of tasks from a diverse demographic. Another issue is the individual differences that exist even in how two experts solve a task. Given this, we are seeking to uncover some commonalities in how people read and navigate between code chunks/beacons when they have similar expertise.

References

- 1 Salwa Aljehane, Bonita Sharif, Jonathan I. Maletic: *Determining Differences in Reading Behavior Between Experts and Novices by Investigating Eye Movement on Source Code Constructs During a Bug Fixing Task*. ETRA Short Papers 2021: 30:1-30:6
- 2 Naser Al Madi, Cole S. Peterson, Bonita Sharif, Jonathan I. Maletic: *From Novice to Expert: Analysis of Token Level Effects in a Longitudinal Eye Tracking Study*. ICPC 2021: 172-183

3.20 How does the Brain Change during Programming Learning?

Janet Siegmund (TU Chemnitz, DE)

License © Creative Commons BY 4.0 International license
© Janet Siegmund

Learning programming has been a challenge for decades, despite many approaches to support students in mastering this important skill. The neuro-cognitive perspective of programming has shown that language-processing skills are essential during programming. Thus, we will be looking into how tapping into language learning can support students who are learning programming. To this end, we teach students an artificial language before they learn programming and evaluate whether this improves their performance. This additional step can be one piece of the puzzle to teach programming to everyone. In a long-term study, we want to observe how the programming skills of students evolve and how that will be reflected in their neuronal representation of programming. Similar to other skills that have an efficient neuronal representation with a high level of expertise, we evaluate whether we can find such a similar change for programming skills.

3.21 Evidence-Based Programming and the “Quorum Project”

Andreas Stefik (University of Nevada, Las Vegas, US)

License © Creative Commons BY 4.0 International license
© Andreas Stefik

In this talk, we will explore the Quorum project, an attempt to make the design programming languages more evidence based in regard to their human factors impact. In the process, we will discuss the history of evidence gathering, competing language designs, and several studies that document the impact of such designs on people at various ability levels and in different demographics

3.22 Making Novices More Like Experts?

Westley Weimer (University of Michigan – Ann Arbor, US)

License © Creative Commons BY 4.0 International license
© Westley Weimer

Joint work of Endres, Madeline; Huang, Yu; Leach, Kevin

Main reference Madeline Endres, Zachary Karas, Xiaosu Hu, Ioulia Kovelman, Westley Weimer: “Relating Reading, Visualization, and Coding for New Programmers: A Neuroimaging Study”, in Proc. of the 43rd IEEE/ACM International Conference on Software Engineering, ICSE 2021, Madrid, Spain, 22-30 May 2021, pp. 600–612, IEEE, 2021.

URL <https://doi.org/10.1109/ICSE43902.2021.00062>

Can we make novices more like experts through targeted training or neurostimulation? We discuss investigations of code comprehension, data structures, code writing and code review, including contextual and functional connectivity analyses. We conclude with a call to arms about the potential use of transcranial magnetic stimulation for causal analyses and behavioral improvements.

3.23 Could we please be a bit more explicit?

Marvin Wyrich (Universität Stuttgart, DE)

License © Creative Commons BY 4.0 International license
© Marvin Wyrich

Main reference Marvin Wyrich, Justus Bogner, Stefan Wagner: “40 Years of Designing Code Comprehension Experiments: A Systematic Mapping Study”, arXiv, 2022.

URL <https://doi.org/10.48550/ARXIV.2206.11102>

We looked at the study designs of code comprehension experiments from the past 40 years and summarized them in a systematic mapping study. We noticed that the primary studies do not yet name, define, and explain too well what construct they actually intend to measure. Before we go after code comprehension with new methods, perhaps we should pause for a moment and clarify whether we intend to measure the same thing. What could code comprehension be?

4 Working Groups

4.1 Eye Tracking Best Practices & Ideas


Teresa Busjahn (Hochschule für Technik und Wirtschaft Berlin – Berlin, DE)

Martha E. Crosby (University of Hawaii at Manoa – Honolulu, US)

Maria Mercedes T. Rodrigo (Ateneo de Manila University – Quezon City, PH)

Christine Lourrine S. Tablatin (Pangasian State University, PH)

Westley Weimer (University of Michigan – Ann Arbor, US)

License  Creative Commons BY 4.0 International license
© Teresa Busjahn, Martha E. Crosby, Maria Mercedes T. Rodrigo, Christine Lourrine S. Tablatin,
Westley Weimer

This working group discussed the current state of the art of eye tracking use for research in computing and gaps in the literature and eye tracking challenges.

4.1.1 Discussed Open Problems

- Programming-specific methodological innovation: The working group members discussed how elements of programming (e.g., navigating through multiple files) make some standard eye tracking approaches more challenging (e.g., static stimuli). The members of the working group agreed that there is currently a gap between the experimental capabilities of eye tracking and studying professional software development. As summarized by Dr. Crosby, “the bigger problems are too big. The experiments we can do [with eye tracking] are code snippets, not production code”.
- Individual differences and generalization: Working group members report that it is challenging to account for individual differences in eye tracking studies and to have confidence that results may generalize. As Dr. Busjahn stated, “Individual differences are a big issue. What’s hard for me might not be hard for you (e.g., if you’ve seen the algorithm before).” Additionally, working group members noted that individual differences can have a large impact on eye tracking results; even the amount of coffee you had while practicing vs. when participating in the study may matter.
- Recruitment: Several working group members noted that they find it difficult to recruit participants for eye tracking studies. Group members noted that for researchers in academia, it is often impossible to pay professional developers their standard hourly rate to participate in eye tracking studies, making the recruitment of professional developers challenging. Additionally, group members note that it can be difficult to recruit diverse populations of programmers, something that is especially important in eye tracking research as demographic factors can influence the results (e.g., if your native language is read from right to left or left to right).
- Standardization of analysis methodologies and empirical results reporting: All of the working group members agree that more standardization is needed for conducting, analyzing, and presenting eye tracking results in computer science research. In particular, concerns were raised regarding the use of parametric statistical tests when not appropriate or the lack of multiple comparison correction in some eye tracking studies.

4.1.2 Possible Approaches and Recommendations

For the problems discussed above, the members of this working group agreed on a set of possible approaches, recommendations, and talking points.

1. *Experimental Design*: Use an established metric (and established name for it) if you can (e.g., from the Holmqvist et al. book [1]). In addition, most CS papers using eye tracking have individual participants complete tasks that are too different. Avoid! “Think before you start the eye tracker.”
2. *Statistical Transparency*: Most CS papers using eye tracking do not indicate which tests they use (e.g., to assess normality of data and otherwise check assumptions). In addition, always report on basic measures (like fixation duration) even if you’re not using them statistically.
3. *Statistical Rigor*: Most CS papers using eye tracking do not check to see if parametric tests are appropriate. In addition, most papers do not consider false discovery rate or correcting for multiple comparisons. Be intentional and use non-parametric tests and correct for multiple comparisons when needed!
4. *Between-Subjects Comparisons*: Most CS papers using eye tracking do not correctly handle normalization.
5. *Good Starting Point Recommendations*:
 - eyecode (for automated AOI analysis for code and prose)
 - code2vec (for salient points in the code)
 - pre-testing for stimuli with a similar population to assess times and difficulty
 - stick to one specific thing to measure, resist the temptation to have five conditions (esp. given the noisy nature of eye tracking)
 - the guide on eye tracking studies in software engineering by Sharafi et al. [2] also provides valuable cues
6. *Anonymous Recommendation*: SE conference program committees should have an ‘on-staff’ person who has done behavioral science (social science) research to help assess claims. This person doesn’t have to write anything unless they spot something. We who are supervising students need to ensure that they have the knowledge.
7. *Longer-Term Question*: Do we have reason to suspect that our results will differ across populations?
 - Examples: Left-to-Right vs. Right-to-Left reading order, dyslexia, corrective lenses, shape of the eye (which may correlate with race/ethnicity), spatial ability.
 - However, we should be careful to consider individual differences: more variability in outcomes comes from the person, not the group membership.

4.1.3 Conclusions

Overall, the working group agrees that while eye tracking can be a powerful tool to study program comprehension, there remain many challenges facing the use of this technology in various research contexts. To help address these challenges, the members of this working group discussed the state of the art and recommend best practices to improve research quality and result standardization.

References

- 1 Kenneth Holmqvist, Marcus Nystrom, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, Joost Van De Weijer; Eye Tracking: *A comprehensive guide to methods and measures*. Oxford University Press, 2015
- 2 Zohreh Sharafi, Bonita Sharif, Yann-Gaël Guéhéneuc, Andrew Begel, Roman Bednarik, Martha E. Crosby: *A practical guide on conducting eye tracking studies in software engineering*. Empirical Software Engineering 25(5): 3128-3174 2020.

4.2 Readability

Andrew Begel (Carnegie Mellon University – Pittsburgh, US)

Annabelle Bergum (Universität des Saarlandes – Saarbrücken, DE)

Madeline Endres (University of Michigan – Ann Arbor, US)

Sarah Fakhoury (Microsoft Research (MSR) – Redmond, US)

Timothy Kluthe (University of Nevada – Las Vegas, US)

Yun-Fei Liu (Johns Hopkins Univ. – Baltimore, US)

Bonita Sharif (University of Nebraska – Lincoln, US)

Jan Stelovsky (University of Hawaii at Manoa – Honolulu, US)

Marvin Wyrich (Universität Stuttgart, DE)

License  Creative Commons BY 4.0 International license

© Andrew Begel, Annabelle Bergum, Madeline Endres, Sarah Fakhoury, Timothy Kluthe, Yun-Fei Liu, Bonita Sharif, Jan Stelovsky, Marvin Wyrich

This working group focused on defining *readability* in the context of code comprehension and programming human studies. To this end, the working group first brainstormed various aspects of readability, coming up with a preliminary taxonomy of related concepts and factors. In particular, the working group considered the definition of code readability (e.g., is readability the same thing as understandability or traceability), the impact of context and programming task on readability (e.g., code review, debugging, API use), metrics for measuring readability (e.g., subjective vs. objective metrics, binary vs. continuous), and potential experimental designs for readability studies (e.g., ecological validity, participant demographics impacts).

Following this brainstorming session, the working group discussed open questions and future directions concerning code readability, ultimately producing a list of 27 research questions of interest for the community.

4.2.1 Readability Open Questions and Research Directions

■ *Foundational and Definitional Questions*

1. How is readability connected to other foundational terms in code research such as comprehension, debuggability, usability, traceability, complexity, or maintainability?
2. What does it mean to be a good code reader?
3. How fast does the average developer read code?
4. Are faster readers better developers?
5. What do professional developers think readability is?
6. Is there a trade off between code readability, and code conciseness or code quality?
7. Does readable code have to be slower?

■ *Participant Demographics and Differences*

1. How do developer demographic factors correlate with or impact code readability?
2. How is readability impacted when comments and documentation are in a programmer's native language?
3. Does native natural language effect readability?
4. How does anticipated readability impact comprehension (e.g., self-efficacy)?
5. How well do metrics for sight reading code work for hearing code (e.g., for blind and visibly impaired programmers)?

■ *Readability, Biometrics, and Cognition*

1. Does readable code help you use less cognitive effort?
2. How does program difficulty impact brain activity when reading code?
3. Does reading other languages or music transfer to code reading?

- *Study Design and Metrics*

1. How does the study task affect comprehension process or readability performance?
2. Should your readability metrics be customized for the task, experimental context, or participants? Are they robust to comprehension strategy?
3. How fast can you measure code readability?

- *Potential Study Tasks*

1. How does code presentation affect readability?
2. How do syntactic features of code affect readability?
3. Do formal style guides or formatting rules make code more readable?
4. What mechanisms help make difficult code easier to read (e.g., interactive tools in code editors)?
5. How do comments affect code readability?
6. Can two people read code better than one?
7. Do many eyes make bugs shallow?
8. How does readability of code or data affect the effectiveness of debugging?

4.2.2 Conclusions

Overall, everyone in the working group agrees that program readability is an important research concept to consider in future research. However, there is also a consensus that more work needs to be done as a community to define readability and understand how it contrasts, overlaps, or connects with other core concepts in program comprehension research such as understandability and debuggability. There was also the acknowledgement of the need for a literature review of work relating to code readability to help build consensus in the research community.

4.3 Statistics and machine learning to predict and model program comprehension

Sven Apel (Universität des Saarlandes – Saarbrücken, DE)

André Brechmann (Leibniz-Institut für Neurobiologie – Magdeburg, DE)

Janet Siegmund (TU Chemnitz, DE)

Lena Jäger (Universität Zürich, CH)

Andreas Stefik (University of Nevada, US)

Takatomi Kubo (Nara Institute of Science and Technology, JP)

License  Creative Commons BY 4.0 International license
© Sven Apel, André Brechman, Janet Siegmund

In this working group, the members discussed how the program comprehension research community could use statistics and machine learning combined with insights from psychology and cognitive science to predict and model program comprehension.

4.3.1 Discussion Summary

One foundational problem discussed by the working group is what is meant by a model of program comprehension. For example, is it a collection of inputs and outputs, such as a model that represents the activation in the brain when a particular stimulus is provided to a person?

The working group members note that one very important distinction is that when using linear models or other traditional machine learning techniques, we need specific features for code comprehension. These features would depend on the experimental designs used. Additionally, the members note that it is important that when modeling program comprehension, researchers should keep in mind the bigger picture and the impact of this work. For example, researchers should consider what *actions* to take or not take depending on the predictions of such a model (e.g., if a model predicts X about program comprehension, should practitioners change programming languages, computer science curricula, or something else?).

4.3.2 Open Questions and Suggestions

Beyond general discussion, the working group discussed open questions regarding what a model of program comprehension should look like. The questions the working group recommends the community to consider include:

- What should these models of program comprehension predict?
 - Suggested outputs include programmer time or accuracy on a task, programmer neural response, eye movement, or the overall impact of various programming language features (e.g., static typing, or syntax choices).
- What inputs should go into program comprehension models?
 - Suggested inputs include brain responses (e.g., fMRI, EEG), eye tracking information, response time and accuracy data, field data about programming usage in practice, corrective information, source code features, and subjective ratings about code or difficulty.
- What demographics should these program comprehension models be predictive for?
- How should we test program comprehension models?
- What is the societal impact of models of program comprehension?
 - Suggested impacts to consider include changing programming languages or predicting how they should change, and educational implications.
- How should we define a scientific process whereby we can create and validate such a model of program comprehension?
 - Suggestions from the working group members include taking existing experiments with known results and vary them based on a proposed model or breaking down the model in terms of basic operations of program comprehension. However, the working group members also note that the basic operations of program comprehension are not yet well defined.

4.3.3 Conclusions

The working group members suggest that as a community, we should decide what the goal of creating a model of program comprehension. Based on that, it will be possible to decide what kind of model to pursue. Additionally, the working group members discussed what existing cognitive models from other fields program comprehension researchers can build on. One such model could be *predictive coding* [1]. Language models extend on predictive coding that have been trained on several programming languages (such as C or Java). Other models from language that have a strong basis in psychology that could be interesting to consider for code include the ACT-R model of sentence processing [2].

References

- 1 Linda Ficco, Lorenzo Mancuso, Jordi Manuello, Alessia Teneggi, Donato Liloia, Sergio Duca, Tommaso Costa, Gyula Zoltán Kovacs, Franco Cauda; *Disentangling predictive processing in the brain: a meta-analytic study in favour of a predictive network*. Nature, 2021.
- 2 Lewis, R. L., Vasishth, S. (2013). An activation-based model of sentence processing as skilled memory retrieval. In Cognitive Science, 29(3), 375-419

5 Open problems

At the end of the seminar, all participants discussed open problems facing program comprehension research. One common theme that emerged was the need for clarity of definitions and experimental dimensions in the field. Many participants viewed doing so as a first step towards unifying as a research community and identifying the most important outstanding directions and questions for program comprehension researchers. As a result, the participants of the seminar worked together to build a preliminary list of concepts for an actionable taxonomy of program comprehension.

5.1 A Taxonomy of Program Comprehension

Participants were generally in consensus that the field of program comprehension currently lacks an explicit taxonomy codifying the research space. As a result, participants worked together to brainstorm categories and dimensions that should be included in such a taxonomy. The goal of this taxonomy was to give researchers a practical guide of dimensions to consider when designing a program comprehension study. The discussion was moderated by Dr. Andrew Begel. In the rest of this section, we present the preliminary taxonomy categories determined by the seminar participants.

- *Level of Comprehension*: Letters, Lexemes, Words, Program structure, Program semantics, Program intent, Program rationale
- *Reason for Comprehension*: Design, Knowledge retention, Immediate use, Understanding
- *Comprehension Success Criteria*: How well is the intent or idea of some code transferred to its audience? (e.g., more formally, code comprehension has succeeded when $Comprehension(Coding(original_idea)) = original_idea$)
- *Code Precision*: Pseudocode, Natural Language, Primitive Language, Symbolic Programming Language (like PERL or APL), Math
- *Amount, Type, or Organization of Code*: Code, Program, Snippet, Architecture, Software, Docs, Modules, Functions, Libraries, Comments, Error messages, Logs
- *Domain of the Code*: Production Software, Experimental or Prototype, Educational, Scientific Computing, Data Processing, Games, Art, Business, Systems, Security, Etc.
- *Code Language Aspects*: Templates or Generics, Classes, Strong or Dynamic Types
- *Modality*: Physical Code Representation (e.g., visual, audio, tactile, features to support presentation such as syntax highlighting), Mental Model Representation
- *Presentation (can do any of these as any modality)*: Text, Graph, Data Structure Graph, Blocks, Flow Charts
- *Experimental Task*: Reading, Writing, Explaining (consider the target audience – see code as a boundary object below), Communicating, Teaching, Tracing, Debugging, Reasoning about variable values, Dependencies, Judging design quality, Refactoring, Fixing Bugs or

Security Vulnerabilities, Adding Features, Code Review, Revising, Code Summarization, Writing Documentation

- *Experimental Interventions (Real time)*: Syntax highlighting, Font, Heatmap, Define/Use/Navigate, Structural Elision Collapse / hide parts of code, Provided code summary, code layout, error messages
- *Experimental Time Scale*: Session, Sprint, Milestone, Day, Week, Semester
- *Code as a Boundary object*: Between coder and computer, Between coder and coder, Between coder and tester, Between coder and user, Between coder and boss, Between student and teacher, Between computer and computer, Between coder and generic outside audience, Size of the audience (either code readers or code users), etc.
- *Number of Code Readers*: Solo, Pair, Mob, Team, Cohort, Replacement
- *Simultaneity*: Interactivity, Synchronous Communication, Asynchronous Communication, No access to code author, No access to friends
- *Person or Programmer In Your Study*: Demographics (e.g., gender, ethnicity, cultural background), Skills or Expertise, Job or Industry, Psychometric tests, Native natural languages, Cognitive difference (e.g., neurodiversity, spatial reasoning ability), Physical Disability, Pregnancy Brain, Mental Health, Identity
- *Experimental Metrics*: Time on task, Accuracy, Speed, Reading Distribution – Spatial/temporal (e.g., fault localization for debugging), Subjective ratings / self reporting, Perceived difficulty or understanding, Continuous or discrete, Cost, Affect, Motivation, Self-efficacy, Biometric and cognitive metrics (e.g., focus, cognitive load, gaze path, brain activity),
- *Code Metrics*: Code complexity Size Debuggability Maintainability Readability
- *Experimental Hypothesis*: from a given hypothesis, you can choose different elements from various categories in the taxonomy.

5.1.1 Conclusions

In this seminar, the participants proposed preliminary categories for a taxonomy of program comprehension research. We hope that the community will build on this work to create a more complete and formal taxonomy going forward.

Participants

- Sven Apel
Universität des Saarlandes –
Saarbrücken, DE
- Andrew Begel
Carnegie Mellon University –
Pittsburgh, US
- Annabelle Bergum
Universität des Saarlandes –
Saarbrücken, DE
- André Brechmann
Leibniz-Institut für Neurobiologie
– Magdeburg, DE
- Teresa Busjahn
Hochschule für Technik und
Wirtschaft Berlin – Berlin, DE
- Martha E. Crosby
University of Hawaii at Manoa –
Honolulu, US
- Sarah D'Angelo
Google – Mountain View, US
- Madeline Endres
University of Michigan –
Ann Arbor, US
- Sarah Fakhoury
Microsoft Research (MSR) –
Redmond, US
- Thomas Fritz
Universität Zürich, CH
- Lena A. Jäger
Universität Zürich, CH
- Timothy Kluthe
University of Nevada –
Las Vegas, US
- Takatomi Kubo
Nara Institute of Science and
Technology, JP
- Yun-Fei Liu
Johns Hopkins Univ. –
Baltimore, US
- Jürgen Mottok
OTH Regensburg, Germany
- Maria Mercedes T. Rodrigo
Ateneo de Manila University –
Quezon City, PH
- Bonita Sharif
University of Nebraska –
Lincoln, US
- Janet Siegmund
TU Chemnitz, DE
- Andreas Stefik
University of Nevada –
Las Vegas, US
- Jan Stelovsky
University of Hawaii at Manoa –
Honolulu, US
- Christine Lourrine S. Tablatin
Pangasian State University, PH
- Westley Weimer
University of Michigan –
Ann Arbor, US
- Marvin Wyrich
Universität Stuttgart, DE



Remote Participants

- Andrew Duchowski
Clemson University, US
- Russell Poldrack
Stanford University, US

Theory and Practice of SAT and Combinatorial Solving

Olaf Beyersdorff^{*1}, Armin Biere^{*2}, Vijay Ganesh^{*3},
Jakob Nordström^{*4}, and Andy Oertel^{†5}

1 Friedrich-Schiller-Universität Jena, DE. olaf.beyersdorff@uni-jena.de

2 Universität Freiburg, DE. armin.biere@gmail.com

3 University of Waterloo, CA. vijay.ganesh@uwaterloo.ca

4 University of Copenhagen, DK & Lund University, SE. jn@di.ku.dk

5 Lund University, SE. andy.oertel@cs.lth.se

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 22411 “*Theory and Practice of SAT and Combinatorial Solving*”. The purpose of this workshop was to explore the Boolean satisfiability (SAT) problem, which plays a fascinating dual role in computer science. By the theory of *NP*-completeness, this problem captures thousands of important applications in different fields, and a rich mathematical theory has been developed showing that all these problems are likely to be infeasible to solve in the worst case. But real-world problems are typically not worst-case, and in recent decades exceedingly efficient algorithms based on so-called conflict-driven clause learning (CDCL) have turned SAT solvers into highly practical tools for solving large-scale real-world problems in a wide range of application areas. Analogous developments have taken place for problems beyond *NP* such as SAT-based optimization (MaxSAT), pseudo-Boolean optimization, satisfiability modulo theories (SMT) solving, quantified Boolean formula (QBF) solving, constraint programming, and mixed integer programming, where the conflict-driven paradigm has sometimes been added to other powerful techniques.

The current state of the art in combinatorial solving presents a host of exciting challenges at the borderline between theory and practice. Can we gain a deeper scientific understanding of the techniques and heuristics used in modern combinatorial solvers and why they are so successful? Can we develop tools for rigorous analysis of the potential and limitations of these algorithms? Can computational complexity theory be extended to shed light on real-world settings that go beyond worst case? Can more powerful methods of reasoning developed in theoretical research be harnessed to yield improvements in practical performance? And can state-of-the-art combinatorial solvers be enhanced to not only solve problems, but also provide verifiable proofs of correctness for the solutions they produce?

This workshop gathered leading applied and theoretical researchers working on SAT and combinatorial optimization more broadly in order to stimulate an exchange of ideas and techniques. We see great opportunities for fruitful interplay between theory and practice in these areas, as well as for technology transfer between different paradigms in combinatorial optimization, and our assessment is that this workshop demonstrated very convincingly that a more vigorous interaction has potential for major long-term impact in computer science, as well for applications in industry.

Seminar October 9–14, 2022 – <http://www.dagstuhl.de/22411>

2012 ACM Subject Classification Theory of computation → Proof complexity; Theory of computation → Complexity theory and logic; Theory of computation → Logic and verification; Theory of computation → Automated reasoning; Theory of computation → Constraint and logic programming; Theory of computation → Discrete optimization; Security and privacy → Logic and verification

Keywords and phrases Boolean satisfiability (SAT), SAT solving, computational complexity, proof complexity, combinatorial solving, combinatorial optimization, constraint programming, mixed integer linear programming

Digital Object Identifier 10.4230/DagRep.12.10.84

* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Theory and Practice of SAT and Combinatorial Solving, *Dagstuhl Reports*, Vol. 12, Issue 10, pp. 84–105

Editors: Olaf Beyersdorff, Armin Biere, Vijay Ganesh, Jakob Nordström, and Andy Oertel



Dagstuhl Reports

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

1 Executive Summary

Olaf Beyersdorff

Armin Biere

Vijay Ganesh

Jakob Nordström

License © Creative Commons BY 4.0 International license
© Olaf Beyersdorff, Armin Biere, Vijay Ganesh, and Jakob Nordström

This event gathered leading researchers working on applied and theoretical aspects of the satisfiability (SAT) problem in areas like Boolean satisfiability (SAT) solving and proof complexity and computational complexity theory more broadly, as well as representatives from neighbouring areas such as, e.g., satisfiability modulo theories (SMT) solving, maximum satisfiability (MaxSAT) solving, pseudo-Boolean optimization, constraint programming, and mixed integer linear programming (MIP) on the applied side, and from other areas of computational complexity theory such as exact exponential-time algorithms and parameterized complexity on the theoretical side. This was meant to create an environment conducive to exchange of ideas and techniques between different fields of research. Among the goals of the workshop were to develop a better scientific understanding of real-world efficient computation in general and of the starkly different perspective between the theory and practice of *NP*-hard problems in particular, to explore new approaches for SAT and other challenging combinatorial problems that would have the potential to go beyond the current state of the art, and to stimulate a technology transfer between SAT and other related areas.

This workshop is part of a highly successful series starting at the Banff International Research Station (BIRS) in Canada in 2014, and with follow-up editions held at Schloss Dagstuhl in 2015 (Seminar 15171), at the Fields Institute in Toronto, Canada in 2016, and at the BIRS-affiliated Casa Matemática Oaxaca in Oaxaca, Mexico in 2018. After this fifth edition at Schloss Dagstuhl in October 2022, a sixth edition “*Satisfiability: Theory, Practice, and Beyond*” has already been organized at the Simons Institute for the Theory of Computing at UC Berkeley in April 2023 as part of an eponymous two-month scientific program.

Topic of the Workshop

What served as the point of departure of this workshop is one of the most significant problems in all of mathematics and computer science, namely that of proving logic formulas. This is a problem of immense importance both theoretically and practically. On the one hand, it is believed to be intractable in general, and deciding whether this is so is one of the famous million dollar Clay Millennium Problems (the *P* vs. *NP* problem). On the other hand, today so-called SAT solvers are routinely and successfully used to solve large-scale real-world instances in a wide range of application areas (such as hardware and software verification, electronic design automation, artificial intelligence research, cryptography, bioinformatics, and operations research, just to name a few examples).

During the last two decades there have been dramatic – and surprising – developments in SAT solving technology that have improved real-world performance by many orders of magnitude. However, while modern solvers can often handle formulas with millions of variables, there are also tiny formulas with just a few hundred variables that cause even the very best solvers to stumble. The fundamental question of when SAT solvers perform well or badly, and what underlying properties of the formulas influence performance, remains very

poorly understood. Other practical SAT solving issues, such as how to optimize memory management and how to exploit parallelization on modern multicore architectures, are even less well studied and understood from a theoretical point of view.

Perhaps even more surprisingly, the best SAT solvers today are still based on relatively simple methods from the early 1960s (though the introduction of so-called conflict-driven learning in the 1990s was a very important addition), searching for proofs in the so-called resolution proof system. Although other mathematical methods of reasoning are known that are much stronger than resolution in theory, in particular methods based on algebra (Gröbner bases) and geometry (cutting planes), attempts to harness the power of such methods have mostly failed to deliver significant improvements in practical performance for SAT solving. And while resolution is a fairly well-understood proof system, even very basic questions about these stronger algebraic and geometric methods remain wide open.

This is an interesting contrast to developments in neighbouring areas such as, e.g., constraint programming and mixed integer programming. There much more powerful methods of reasoning are successfully used to guide the search, but compared to SAT solving the attempts to employ conflict-driven learning have had much less of an impact. Also, while for SAT solvers it is at least possible to understand some aspects of their performance (by analysing proof systems such as resolution), a corresponding theoretical framework for constraint programming and mixed integer linear programming seems to be mostly missing.

In this workshop, we gathered leading researchers working on SAT and other challenging combinatorial optimization problems in order to stimulate an increased exchange of ideas between theoreticians and practitioners. As discussed above, previous editions of this workshop series at the Banff International Research Station, Schloss Dagstuhl, Fields Institute, and Casa Matemática Oaxaca have had a major impact on the involved communities and has helped to create bridges and stimulate the emergence of a joint research agenda. We are happy to report that the October 2022 workshop at Schloss Dagstuhl fully delivered on the expectation of serving as the valuable next step on this journey. During recent years we have already seen how computational complexity theory can shed light on the power and limitations on current and possible future techniques for SAT and other optimization problems, and that problems encountered on the applied side can spawn interesting new areas in theoretical research. We see great potential for continued interdisciplinary research at the border between theory and practice in this area, and believe that more vigorous interaction between practitioners and theoreticians could have major long-term impact in both academia and industry.

Goals of the Workshop

A strong case can be made for the importance of increased exchange between the two fields of SAT solving on the one hand and proof complexity (and more broadly computational complexity) on the other. Given how many questions that would seem to be of mutual interest, it is striking that the level of interaction had been so low until our workshop series started with the first two meetings in Banff in 2014 and Dagstuhl in 2015. Below, we outline some of the concrete questions that served as motivation for organizing our second Dagstuhl workshop in this series in 2022, and for broadening the scope from Boolean satisfiability to combinatorial solving and optimization in general. We want to stress that this inst is far from exhaustive, and we believe that one important outcome of the seminar was to uncover also other questions at the intersection of theoretical and applied research, and of different research areas within combinatorial solving and optimization.

What Makes Formulas Hard or Easy in Practice for Modern SAT Solvers?

The best SAT solvers known today are based on the DPLL procedure, augmented with optimizations such as conflict-driven clause learning (CDCL) and restart strategies. The propositional proof system underlying such algorithms, resolution, is arguably the most well-studied system in all of proof complexity.

Given the progress during the last decade on solving large-scale instances, it is natural to ask what lies behind the spectacular success of CDCL solvers at solving these instances. And given that there are still very small formulas that resist even the most powerful CDCL solvers, a complementary interesting question is if one can determine whether a particular formula is hard or tractable. Somewhat unexpectedly, very little turns out to be known about these questions.

In view of the fundamental nature of the SAT problem, and in view of the wide applicability of modern SAT solvers, this seems like a clear example of a question of great practical importance where the theoretical field of proof complexity could potentially provide useful insights. In particular, one can ask whether one could find theoretical complexity measures for formulas that would capture their practical hardness in some nice and clean way. Besides greatly advancing our theoretical understanding, answering such a question could also have applied impact in the longer term by clarifying the limitations, and potential for further improvements, of modern SAT solvers.

Can Proof Complexity Shed Light on Crucial SAT Solving Issues?

Understanding the hardness of proving formulas in practice is not the only problem for which more applied researchers would welcome contributions from theoretical computer scientists. Examples of some other possible practical questions that would benefit from a deeper theoretical understanding follow below.

- Firstly, we would like to study the question of memory management. One major concern for clause learning algorithms is to determine how many clauses to keep in memory. Also, once the algorithm runs out of the memory currently available, one needs to determine which clauses to throw away. These questions can have huge implications for performance, but are poorly understood.
- In addition to clause learning, the concept of restarts is known to have decisive impact on the performance on modern CDCL solvers. It would be nice to understand theoretically why this is so. The reason why clause learning increases efficiency greatly is clear – without it the solver will only generate so-called tree-like proofs, and tree-like resolution is known to be exponentially weaker than general resolution. However, there is still ample room for improvement of our understanding of the role of restarts and what are good restart strategies.
- Given that modern computers are multi-core architectures, a highly topical question is whether this (rather coarse-grained) parallelization can be used to speed up SAT solving. While there are some highly successful attempts in parallelizing SAT for solving theoretical problems in, e.g., extremal combinatorics, the speed-ups obtained for more applied problems are rather modest or sometimes non-existent. This is a barrier for further adoption of SAT technology already today, and will become a more substantial problem as thousands of cores and cloud computing are becoming the dominant computing platforms in the future. A theoretical understanding of if and how SAT can be parallelized will be essential to develop new parallelization strategies to adapt SAT to this new computing paradigm.

We believe that progress on any of these questions has the potential of influencing the further development of both theoretical and applied research, and to stimulate a further cross-pollination between these two areas.

Can we build SAT Solvers based on Stronger Proof Systems than Resolution?

Although the performance of modern CDCL SAT solvers is impressive, it is nevertheless astonishing, not to say disappointing, that the state-of-the-art solvers are still based on resolution. This method lies close to the bottom in the hierarchy of propositional proof systems, and there are many other proof systems based on different forms of mathematical reasoning that are known to be strictly stronger. Some of these appear to be natural candidates on which to build stronger SAT solvers than those using CDCL.

In particular, proof systems such as polynomial calculus (based on algebraic reasoning) and cutting planes (based on geometry) are known to be exponentially more powerful than resolution. While there has been some work on building SAT solvers on top of these proof systems, progress has been fairly limited. We believe it would be fruitful to discuss what the barriers are that stops us from building stronger algebraic or geometric SAT solvers, and what is the potential for future improvements. An important part of this work would seem to be to gain a deeper theoretical understanding of the power and limitations of these proof methods. Here there are a number of fairly long-standing open theoretical questions. At the same time, only in the last couple of years proof complexity has made substantial progress, giving hope that the time is ripe for decisive break-throughs in these areas.

Can Technology Be Transferred Between Different Combinatorial Optimization Paradigms?

Continuing the discussion of stronger methods of reasoning, it is natural to ask whether techniques from e.g., constraint programming (CP) and mixed integer linear programming (MIP) could be imported into SAT solving and vice versa. At a high level, the main loop of combinatorial solvers in all of these paradigms consists of two phases:

- During the *search phase*, the solver makes *decisions* (guesses) about variable assignments and propagates “obvious” consequences until it either finds a solution or discovers a violated constraint (a *conflict*).
- In case of a conflict, during the *backtracking phase* the solver analyses what went wrong and reverses some decision(s) to remove the violation, after which it switches to search again.

For CP and MIP solvers, significant effort is spent during the search phase on making intelligent decisions and deriving (sometimes not so obvious) consequences. In comparison, the backtracking phase is not so sophisticated. For SAT and pseudo-Boolean (PB) solvers it is exactly the other way around. The decisions during the search phase are done quite naively, and since the constraints are relatively simple, propagations cannot be too strong and are fairly easy to detect. Once a conflict is reached, however, an elaborate *conflict analysis* algorithm combines the constraints involved in this conflict to *learn* a new, globally valid constraint that is added to the input formula.

Could it be possible to make more efficient use of conflict analysis in MIP and CP solving? CDCL-style analysis has already been tried with some success, but since this approach boils down to reasoning with clauses it is provably exponentially weaker than what pseudo-Boolean techniques can offer. In the other direction, a quite tempting proposition is to integrate into SAT and PB solvers the vastly stronger propagation used in CP and MIP. Another change

of perspective would be to turn *core-guided MaxSAT solving* into a general pseudo-Boolean optimization technique, using PB conflict analysis to extract better cores, and letting these cores serve as a heuristic for introducing new variables in the spirit of *extended resolution*.

Organization of the Workshop

The scientific program of the seminar consisted of 29 presentations. Among these there were 14 50-minute surveys of different core topics of the workshop. These talks occupied most of the morning schedule Monday-Thursday, and were intended to make sure that the diverse audience would have a bit of a common background for the more technical talks reporting on recent research projects. The list of survey talks and speakers were as follows:

- SAT: interactions between theory and practice (Olaf Beyersdorff)
- SAT and computational complexity theory (Ryan Williams)
- Proof complexity and SAT solving (Jakob Nordström)
- Efficient proof search in proof complexity, a.k.a. automatability (Susanna de Rezende)
- Satisfiability modulo theories (SMT) solving (Nikolaj Bjørner)
- Quantified Boolean formula (QBF) solving and proof complexity (Meena Mahajan)
- Constraint programming (Ciaran McCreesh)
- Mixed integer linear programming (Ambros Gleixner)
- Algebraic methods for circuit verification (Daniela Kaufmann)
- First-order theorem proving (Laura Kovacs and Martin Suda)
- Automated planning (Malte Helmert)
- Formally verified combinatorial solvers (Mathias Fleury)
- Certifying solvers with proof logging (Armin Biere)
- Formally verified proof checking for certifying solvers (Yong Kiam Tan)

The rest of the talks were 25-minute presentations on recent research of the participants. The time after lunch each day was left for self-organized collaborations and discussions, and there was no schedule on Wednesday afternoon.

Based on polling of participants before the seminar week, it was decided to have an open problem session on Thursday afternoon. The poll also asked whether a panel discussion should be organized, but the support for this idea was weaker, and several participants emphasized that the workshop program should not be too dense and that the evenings should be left free of any program. Therefore, the organizers decided not to have a panel discussion. As a nice contribution, some of the participants of the workshop organized a music night on the last evening of the workshop.

2 Table of Contents

Executive Summary

<i>Olaf Beyersdorff, Armin Biere, Vijay Ganesh, and Jakob Nordström</i>	85
---	----

Overview of Talks

Clause Redundancy and Preprocessing in Maximum Satisfiability <i>Jeremias Berg</i>	92
Theory and practice of SAT solving <i>Olaf Beyersdorff</i>	92
Trusting SAT Solvers <i>Armin Biere</i>	92
In introduction to SMT with Proofs <i>Nikolaj S. Bjørner</i>	93
On Symmetries and Certification <i>Bart Bogaerts</i>	93
CDCL vs resolution: the picture in QBF <i>Benjamin Böhm</i>	94
Theoretical Barriers for Efficient Proof Search (a Survey) <i>Susanna de Rezende</i>	94
On Design Decisions of Extending CDCL with External Propagators <i>Katalin Fazekas</i>	94
Discussion: How to combine and compare options in solvers? <i>Mathias Fleury and Armin Biere</i>	95
Verifying Solvers: How Much Do You Want to Prove? <i>Mathias Fleury</i>	95
Algorithmic Mixed Integer Programming: Between Exactness and Performance in Theory and Practice <i>Ambros M. Gleixner</i>	95
The Packing Chromatic Number of the Infinite Square Grid is 15 <i>Marijn J. H. Heule</i>	96
Pseudo-Boolean Optimization by Implicit Hitting Sets <i>Matti Järvisalo</i>	96
Exploring Algebraic Methods for Circuit Verification <i>Daniela Kaufmann</i>	97
First-Order Theorem Proving – Theory and Practice <i>Laura Kovács</i>	97
Towards a Deeper Understanding of Modern CDCL SAT Solvers <i>Chunxiao (Ian) Li, Jonathan Chung, Vijay Ganesh, Antonina Kolokolova, Alice Mu, Soham Mukherjee, and Marc Vinyals</i>	98
Quantified Boolean Formulas: (Solving and) Proof Complexity <i>Meena Mahajan</i>	98

How Constraint Programming Isn't Like SAT <i>Ciaran McCreesh</i>	98
Certification of Samplers <i>Kuldeep S. Meel</i>	99
Proof complexity and SAT solving <i>Jakob Nordström</i>	99
Certified CNF Translations for Pseudo-Boolean Solving <i>Andy Oertel</i>	100
Scalable optimization with SAT-based local improvement (SLIM) <i>Andre Schödler, Friedrich Slivovsky, and Stefan Szeider</i>	100
On the Use of SAT Solvers in a Modern ATP <i>Martin Suda</i>	101
The Last Mile in Trustworthy Automated Reasoning <i>Yong Kiam Tan</i>	101
A SAT Solver + Computer Algebra Attack on the Minimum Kochen-Specker Problem <i>Vijay Ganesh</i>	102
Theoretical limits of UIP Learning <i>Marc Vinyals</i>	102
Exponential separations using guarded extension variables <i>Emre Yolcu</i>	102
Evaluation by Participants	103
Participants	105

3 Overview of Talks

3.1 Clause Redundancy and Preprocessing in Maximum Satisfiability

Jeremias Berg (University of Helsinki, FI)

License © Creative Commons BY 4.0 International license
© Jeremias Berg

Joint work of Ihalainen, Hannes; Järvisalo, Matti; Berg, Jeremias

Main reference Hannes Ihalainen, Jeremias Berg, Matti Järvisalo: “Clause Redundancy and Preprocessing in Maximum Satisfiability”, in Proc. of the Automated Reasoning – 11th International Joint Conference, IJCAR 2022, Haifa, Israel, August 8-10, 2022, Proceedings, Lecture Notes in Computer Science, Vol. 13385, pp. 75–94, Springer, 2022.

URL http://dx.doi.org/10.1007/978-3-031-10769-6_6

The study of clause redundancy in Boolean satisfiability (SAT) has proven significant in various terms, from fundamental insights into preprocessing and inprocessing to the development of practical proof checkers and new types of strong proof systems. I will present our recent work on liftings of the recently-proposed notion of propagation redundancy — based on a semantic implication relationship between formulas — in the context of maximum satisfiability (MaxSAT), where of interest are reasoning techniques that preserve optimal cost (in contrast to preserving satisfiability in the realm of SAT). We establish the strongest MaxSAT-lifting of propagation redundancy allows for changing in a controlled way the set of minimal correction sets in MaxSAT. This ability is key in succinctly expressing MaxSAT reasoning techniques and allows for obtaining correctness proofs in a uniform way for MaxSAT reasoning techniques very generally. I will also highlight some interesting directions for future work.

3.2 Theory and practice of SAT solving

Olaf Beyersdorff (Friedrich-Schiller-Universität Jena, DE)

License © Creative Commons BY 4.0 International license
© Olaf Beyersdorff

This talk provides a brief survey on the relations between proof complexity and SAT solving. What can proof complexity tell us about the strength and limitations of SAT solving? Why should practitioners be interested in proof complexity results and why should theorists study SAT solving? What have we achieved in the past 25 years and which problems remain open?

3.3 Trusting SAT Solvers

Armin Biere (Universität Freiburg, DE)

License © Creative Commons BY 4.0 International license
© Armin Biere

Many critical applications crucially rely on the correctness of SAT solvers. Particularly in the context of formal verification, the claim by a SAT solver that a formula is unsatisfiable corresponds to a safety or security property to hold, and thus needs to be trusted. In order to increase the level of trust an exciting development in this century was to let SAT solvers produce certificates, i.e., by tracing proofs of unsatisfiability, which can independently be

checked. In the last ten years this direction of research gained substantial momentum, e.g., solvers in the main track of the SAT competition are required to produce such certificates and industrial applications of SAT solvers require that feature too. In this talk we review this quarter of century of research in certifying the result of SAT solvers, discuss briefly alternatives, including testing approaches and verifying the SAT solver directly, mention exciting research on new proof systems produced in this context as well as how these ideas extend beyond formulas in conjunctive normal form.

3.4 In introduction to SMT with Proofs

Nikolaj S. Bjørner (Microsoft – Redmond, US)

License © Creative Commons BY 4.0 International license
© Nikolaj S. Bjørner

The talk provides an overview of selected current trends in SMT solving theories and techniques <https://z3prover.github.io/slides/proofs.html>. An active area of discussion in the SMT community is around proof formats for SMT solvers. I give an introduction to current approaches pursued in solvers such as Z3, CVC5, VeriT and SMTInterpol.

3.5 On Symmetries and Certification

Bart Bogaerts (VU – Brussels, BE)

License © Creative Commons BY 4.0 International license
© Bart Bogaerts

Joint work of Bogaerts, Bart; Devriendt, Jo; Gocht, Stephan; McCreesh, Ciaran; Nordström, Jakob
Main reference Bart Bogaerts, Stephan Gocht, Ciaran McCreesh, Jakob Nordström: “Certified Symmetry and Dominance Breaking for Combinatorial Optimisation”, in Proc. of the Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 – March 1, 2022, pp. 3698–3707, 2022.

URL <https://ojs.aaai.org/index.php/AAAI/article/view/20283>

In this talk, we take a deep-dive in the fascinating world of symmetry handling for Boolean Satisfiability, by reviewing three main classes of techniques: static symmetry breaking, dynamic symmetry breaking, and (dynamic) symmetric learning. We focus on proof logging techniques that have been developed for these symmetry handling methods, and in particular study our recent symmetry handling methods built on VeriPB. We end with some open problems and challenges.

3.6 CDCL vs resolution: the picture in QBF

Benjamin Böhm (Friedrich-Schiller-Universität Jena, DE)

License  Creative Commons BY 4.0 International license
 © Benjamin Böhm

Joint work of Peitl, Tomás; Beyersdorff, Olaf

Main reference Benjamin Böhm, Tomás Peitl, Olaf Beyersdorff: “Should Decisions in QCDCL Follow Prefix Order?”, 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. 11:1–11:19, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022.

URL <http://dx.doi.org/10.4230/LIPIcs.SAT.2022.11>

This talk will cover the relations between QBF resolution and QCDCL solving algorithms. Modelling QCDCL as proof systems we show that QCDCL and Q-Resolution are incomparable. We also introduce new versions of QCDCL that turn out to be stronger than the classic models. This talk is based on a couple of recent papers (joint with Olaf Beyersdorff and Tomas Peitl, which appeared in ITCS’21, SAT’21, SAT’22 and IJCAI’22).

3.7 Theoretical Barriers for Efficient Proof Search (a Survey)

Susanna de Rezende (Lund University, SE)

License  Creative Commons BY 4.0 International license
 © Susanna de Rezende

The proof search problem is a central question in automated theorem proving and SAT solving. Clearly, if a propositional tautology F does not have a short (polynomial size) proof in a proof system P , any algorithm that searches for P -proofs of F will necessarily take super-polynomial time. But can proofs of “easy” formulas, i.e., those that have polynomial size proofs, be found in polynomial time? This question motivates the study of automatability of proof systems. In this talk, we give an overview of known non-automatability results, focusing on the more recent ones, and present some of the main ideas used to obtain them.

3.8 On Design Decisions of Extending CDCL with External Propagators

Katalin Fazekas (TU Wien, AT)

License  Creative Commons BY 4.0 International license
 © Katalin Fazekas

Joint work of Katalin Fazekas, Armin Biere

Solving combinatorial problems often combines SAT solving with different reasoning techniques. An external propagator can interpret the partial assignment built by the SAT solver during search and, based on such different reasoning methods, can construct clauses that are propagating or conflicting under that assignment. The use of such external propagators allows to directly guide the search of the SAT solver into a preferred direction, which can make problem solving in several problem domains more efficient (consider e.g. dynamic symmetry breaking). However, both the efficient combination of external propagators with the complex features of modern SAT solvers (e.g. proof logging and inprocessing), and the theoretical understanding of such combined reasoning methods are open problems. This talk, based on current work in progress, presents some of the design decisions that must be considered when external propagation is combined with modern CDCL solvers. We describe

some challenges and formulate both practical and theoretical open questions about how to implement external propagation in CDCL in the presence of current SAT solver features such as proof logging, clause database reduction and inprocessing.

3.9 Discussion: How to combine and compare options in solvers?

Mathias Fleury (Universität Freiburg, DE) and Armin Biere (Universität Freiburg, DE)

License © Creative Commons BY 4.0 International license
© Mathias Fleury and Armin Biere

Joint work of Sakallah, Kareem; Biere, Armin; Fleury, Mathias

Comparing options between solvers is a complicated task. There are three main ways: runtime options (with the risk of not understanding requirements between features), compile-time option (with the issue of testing and making the code very complicated), or different versions of the source code. A second question is how to measure the performance without implementing the most advanced version. This (short) talk should serve as a basis for discussion on how to organize the development of a solver with various options.

3.10 Verifying Solvers: How Much Do You Want to Prove?

Mathias Fleury (Universität Freiburg, DE)

License © Creative Commons BY 4.0 International license
© Mathias Fleury

Joint work of Blanchette, Jasmin; Lammich, Peter; Weidenbach, Christoph; Fleury, Mathias

In this talk, I present the two main approaches to verify solvers: partial verification (usually bottom-up from code to the specification) and complete verification (usually top-down from the specification towards the code). The former approach present many similarities to verify checkers, whereas the latter starts with a full formalization of underlying algorithm. I compare the approaches and show where the main challenges are.

3.11 Algorithmic Mixed Integer Programming: Between Exactness and Performance in Theory and Practice

Ambros M. Gleixner (HTW - Berlin, DE)

License © Creative Commons BY 4.0 International license
© Ambros M. Gleixner

Joint work of Eifler, Leon; Gleixner, Ambros M.

Main reference Leon Eifler, Ambros M. Gleixner: “A computational status update for exact rational mixed integer programming”, *Math. Program.*, Vol. 197(2), pp. 793–812, 2023.

URL <http://dx.doi.org/10.1007/s10107-021-01749-5>


Today’s state-of-the-art solvers for the general class of mixed integer programs exhibit both exact and heuristic properties. Both aspects are crucial for their lasting relevance in academic and industrial practice. In this talk, we give an overview of methods implemented and successfully used in mixed-integer programming solvers and try to point out connections to satisfiability solving and pseudo-Boolean optimization. We conclude by outlining our efforts to address the ubiquitous use of floating-point arithmetic in virtually all fast mixed integer programming solvers and report advances in performant roundoff-error-free MIP solving with proof logging [1].

References

- 1 Leon Eifler and Ambros Gleixner. A computational status update for exact rational mixed integer programming. *Mathematical Programming*, 2022.

3.12 The Packing Chromatic Number of the Infinite Square Grid is 15

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

License  Creative Commons BY 4.0 International license
© Marijn J. H. Heule

Joint work of Marijn J. H. Heule, Bernardo Subercaseaux

A packing k -coloring of a graph $G = (V, E)$ is a mapping from V to $1, \dots, k$ such that any pair of vertices u, v that receive the same color c must be at distance greater than c in G . Arguably the most fundamental problem regarding packing colorings is to determine the packing chromatic number of the infinite square grid. Various works in the last 20 years improved the bounds of this problem. We finally solve it and show that the answer is 15. A crucial part of our solution is a novel encoding that reduces the runtime by a factor of 30. Moreover, we construct and validate a DRAT proof of unsatisfiability for the direct encoding of the problem. This proof includes the symmetry-breaking and reencoding techniques that we applied.

3.13 Pseudo-Boolean Optimization by Implicit Hitting Sets

Matti Järvisalo (University of Helsinki, FI)

License  Creative Commons BY 4.0 International license
© Matti Järvisalo

Joint work of Smirnov, Pavel; Berg, Jeremias; Järvisalo, Matti

Main reference Pavel Smirnov, Jeremias Berg, Matti Järvisalo: “Pseudo-Boolean Optimization by Implicit Hitting Sets”, 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. 51:1–51:20, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021.

URL <http://dx.doi.org/10.4230/LIPIcs.CP.2021.51>

Recent developments in applying and extending Boolean satisfiability (SAT) based techniques have resulted in new types of approaches to pseudo-Boolean optimization (PBO), complementary to the more classical integer programming techniques. In this talk, we outline a first approach to pseudo-Boolean optimization based on instantiating the so-called implicit hitting set (IHS) approach [2, 1], motivated by the success of IHS implementations for maximum satisfiability (MaxSAT). In particular, we harness recent advances in native reasoning techniques for pseudo-Boolean constraints, which enable efficiently identifying inconsistent assignments over subsets of objective function variables (i.e. unsatisfiable cores in the context of PBO), as a basis for developing a native IHS approach to PBO, and study the impact of various search techniques applicable in the context of IHS for PBO.

References

- 1 P. Smirnov, J. Berg, and M. Järvisalo. Improvements to the implicit hitting set approach to pseudo-boolean optimization. In *SAT 2022*.
- 2 P. Smirnov, J. Berg, and M. Järvisalo. Pseudo-boolean optimization by implicit hitting sets. In *CP 2021*.

3.14 Exploring Algebraic Methods for Circuit Verification

Daniela Kaufmann (TU Wien, AT)

License © Creative Commons BY 4.0 International license
© Daniela Kaufmann

Joint work of Kaufmann, Daniela; Biere, Armin; Kauers, Manuel; Fleury, Mathias; Beame, Paul; Nordström, Jakob
Main reference Daniela Kaufmann, Paul Beame, Armin Biere, Jakob Nordström: “Adding Dual Variables to Algebraic Reasoning for Gate-Level Multiplier Verification”, in Proc. of the 2022 Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 1431–1436, 2022.

URL <http://dx.doi.org/10.23919/DATE54114.2022.9774587>

Digital circuits are widely utilized in computers, because they provide models for various digital components and arithmetic operations. To avoid problems like the infamous Pentium FDIV bug, it is critical to ensure that these circuits are correct. Formal verification can be used to evaluate whether a circuit meets a given specification. Arithmetic circuits, in particular integer multipliers, pose a challenge to current verification approaches. Techniques that rely solely on SAT solving or decision diagrams appear incapable of tackling this problem in an acceptable period of time. In practice, circuit verification still requires a substantial amount of manual labor.

In this talk, we will demonstrate an automated verification technique that is based on algebraic reasoning and is currently considered to be one of the most successful verification methods for circuit verification. In this approach the circuit is modeled as a set of polynomial equations that is implied by the circuit. For a correct circuit, we must demonstrate that the specification is implied by the polynomial representation of the given circuit. However, some sections of the multiplier, such as final stage adders, are difficult to check using simply computer algebra. To address this issue, we will provide a hybrid solution that blends SAT and computer algebra.

But who verifies the verifier? The ability to independently generate and check proof certificates boosts confidence in the outcomes of automated reasoning tools. We present an algebraic proof calculus that allows us to obtain certificates as a by-product of circuit verification and that can be efficiently verified with our independent proof checking tools.

3.15 First-Order Theorem Proving – Theory and Practice

Laura Kovács (TU Wien, AT)

License © Creative Commons BY 4.0 International license
© Laura Kovács

First-order theorem proving is undergoing a rapid development thanks to its successful use in software analysis, formal verification, IT security, symbolic computation, theorem proving in mathematics, and other related areas. Breakthrough results in all areas of theorem proving have been obtained, including improvements in theory, implementation, and the development of powerful theorem proving tools.

This talk serves as a mini-tutorial on the theory and practice of first-order theorem proving. We introduce the core concepts of automating first-order theorem proving in first-order logic with equality. We will discuss the resolution and superposition calculus, introduce the saturation principle, present various algorithms implementing redundancy elimination, and demonstrate how these concepts are implemented in our Vampire theorem prover. We also survey practical considerations for making saturation efficient.

The talk will next be complemented with the presentation of Dr. Martin Suda, discussing applications of SAT solvers in the efficient automation of first-order reasoning.

3.16 Towards a Deeper Understanding of Modern CDCL SAT Solvers

Chunxiao (Ian) Li (University of Waterloo, CA), Jonathan Chung, Vijay Ganesh (University of Waterloo, CA), Antonina Kolokolova (University of Newfoundland, CA), Alice Mu, Soham Mukherjee, and Marc Vinyals (University of Newfoundland, CA)

License © Creative Commons BY 4.0 International license

© Chunxiao (Ian) Li, Jonathan Chung, Vijay Ganesh, Antonina Kolokolova, Alice Mu, Soham Mukherjee, and Marc Vinyals

Main reference Chunxiao Li, Jonathan Chung, Soham Mukherjee, Marc Vinyals, Noah Fleming, Antonina Kolokolova, Alice Mu, Vijay Ganesh: “On the Hierarchical Community Structure of Practical Boolean Formulas”, in Proc. of the Theory and Applications of Satisfiability Testing – SAT 2021 – 24th International Conference, Barcelona, Spain, July 5-9, 2021, Proceedings, Lecture Notes in Computer Science, Vol. 12831, pp. 359–376, Springer, 2021.

URL http://dx.doi.org/10.1007/978-3-030-80223-3_25

Understanding why state-of-the-art SAT solvers are empirically successful has been a long standing question since the beginning of solver research. An ideal answer to the question should be both empirically verifiable, and in the same time can be theoretically analyzed. To shed light on this difficult problem, I will present a novel concept for SAT formulas and proofs, namely the hierarchical community structure (HCS). Empirically we show that hierarchical community structure can be used to distinguish industrial formulas from random, crafted and crypto formulas. And theoretically, we prove size upper bounds parameterized in the HCS structure. I will also discuss some recent developments in the locality of proofs through the lens of HCS.

3.17 Quantified Boolean Formulas: (Solving and) Proof Complexity

Meena Mahajan (The Institute of Mathematical Sciences – Chennai, IN)

License © Creative Commons BY 4.0 International license

© Meena Mahajan

QBF solving brings many new challenges and has thrown up many innovative approaches and heuristics. QBF proof complexity explores the theoretical underpinnings of these approaches rigorously, explains relative strengths of different approaches, exposes limitations, and suggests new approaches. This talk will survey some of the developments in the area.

3.18 How Constraint Programming Isn’t Like SAT

Ciaran McCreesh (University of Glasgow, GB)

License © Creative Commons BY 4.0 International license

© Ciaran McCreesh

This talk provides an overview of modern constraint programming and how it differs from SAT solving, both in technology and terminology. I’ll give an introduction to how the CP community thinks and speaks, starting with modelling and reformulation; then constraints, propagation, and lazy clause generation; and finally, search. Next we’ll take a closer look at the all-different constraint: I’ll explain how it’s propagated and why CNF can’t do the same thing, and then we’ll look at whether stronger propagation is actually a good idea in practice. I’ll conclude with a look at three exciting research topics: proof logging, belief propagation, and parallel search.

3.19 Certification of Samplers

Kuldeep S. Meel (National University of Singapore, SG)

License © Creative Commons BY 4.0 International license
© Kuldeep S. Meel

Joint work of Chakraborty, Sourav; Golia, Priyanka; Soos, Mate; Meel, Kuldeep S.

Main reference Mate Soos, Priyanka Golia, Sourav Chakraborty, Kuldeep S. Meel: “On Quantitative Testing of Samplers”, in Proc. of the 28th International Conference on Principles and Practice of Constraint Programming, CP 2022, July 31 to August 8, 2022, Haifa, Israel, LIPICs, Vol. 235, pp. 36:1–36:16, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022.

URL <http://dx.doi.org/10.4230/LIPICs.CP.2022.36>

Given a Boolean formula F , the problem of constrained sampling is to generate solutions uniformly at random. Constrained sampling is a fundamental problem in computer science with various applications. While constrained sampling is a computationally challenging problem, the SAT revolution has paved the way for various approaches focused on designing scalable samplers. On one end of the spectrum, we have techniques that provide provably rigorous guarantees but often fail to scale to large instances. On the other hand, we have samplers that can scale to large instances, but theoretical tools are often insufficient to argue about the quality of these samplers.

I will discuss our project, Barbarik, which seeks to certify samplers [1]. Barbarik builds on the recent advances in distribution testing and can identify samplers whose distributions are not of high quality. The availability of Barbarik led us to investigate the possibility of test-driven development of constrained samplers, and our efforts yielded CMSGen [2], a sampler that augments the standard CDCL routine with *just enough* randomization so as not to be *caught* by Barbarik. Surprisingly, CMSGen is highly effective in practice and thus paving the way forward to a new design methodology for samplers [3].

References

- 1 Sourav Chakraborty and Kuldeep S. Meel. On testing of uniform samplers. In *Proc. of AAAI*, pages 7777–7784. AAAI Press, 2019.
- 2 Priyanka Golia, Mate Soos, Sourav Chakraborty, and Kuldeep S. Meel. Designing samplers is easy: The boon of testers. In *Proc. of FMCAD*, pages 222–230, 2021.
- 3 Mate Soos, Priyanka Golia, Sourav Chakraborty, and Kuldeep S. Meel. On quantitative testing of samplers. In Christine Solnon, editor, *Proc. of CP*, volume 235 of *LIPICs*, pages 36:1–36:16, 2022.

3.20 Proof complexity and SAT solving

Jakob Nordström (University of Copenhagen, DK & Lund University, SE)

License © Creative Commons BY 4.0 International license
© Jakob Nordström

This talk is intended to give an overview of proof complexity and connections to Boolean satisfiability (SAT) solving. The focus will be on proof systems (and corresponding algorithms) such as resolution (DPLL and conflict-driven clause learning), Nullstellensatz and polynomial calculus (linear algebra and Gröbner basis computations), and cutting planes (pseudo-Boolean solving and 0-1 integer linear programming). Time permitting, we will also discuss briefly proof systems such as Sherali-Adams and sums of squares (linear programming and semidefinite programming hierarchies), stabbing planes (0-1 ILP), and extended resolution (SAT pre- and inprocessing).

3.21 Certified CNF Translations for Pseudo-Boolean Solving

Andy Oertel (Lund University, SE)

License © Creative Commons BY 4.0 International license
© Andy Oertel

Joint work of Gocht, Stephan; Martins, Ruben; Nordström, Jakob; Oertel, Andy
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 <http://dx.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.

3.22 Scalable optimization with SAT-based local improvement (SLIM)

Andre Schidler (TU Wien, AT), Friedrich Slivovsky (TU Wien, AT), and Stefan Szeider (TU Wien, AT)

License © Creative Commons BY 4.0 International license
© Andre Schidler, Friedrich Slivovsky, and Stefan Szeider

SAT-based local improvement (SLIM) is an optimization metaheuristic. It repeatedly employs SAT-based solvers to local versions of the problem instance at hand, gradually improving a heuristically computed initial global solution. SLIM has been successfully instantiated for several problems, including graph decomposition and coloring, decision tree induction, Bayesian network structure learning, and circuit synthesis.

3.23 On the Use of SAT Solvers in a Modern ATP

Martin Suda (Czech Technical University – Prague, CZ)

License © Creative Commons BY 4.0 International license
© Martin Suda

Joint work of Reger, Giles; Suda, Martin

Main reference Giles Reger, Martin Suda: “The Uses of SAT Solvers in Vampire”, in Proc. of the 1st and 2nd Vampire Workshops, Vampire@VSL 2014, Vienna, Austria, July 23, 2014 / Vampire@CADE 2015, Berlin, Germany, August 2, 2015, EPiC Series in Computing, Vol. 38, pp. 63–69, EasyChair, 2015.

URL <http://dx.doi.org/10.29007/4w68>

A modern saturation-based automatic theorem prover for first-order logic (ATP) relies on the SAT technology at various places. We have the Inst-Gen calculus and global subsumption reduction rule, MACE-style finite model finding, and, most notably, the AVATAR architecture. I will give an overview of these pieces of technology and outline how they contribute to the performance of our ATP Vampire.

3.24 The Last Mile in Trustworthy Automated Reasoning

Yong Kiam Tan (Infocomm Research – Singapore, SG)

License © Creative Commons BY 4.0 International license
© Yong Kiam Tan

Joint work of Tan, Yong Kiam; Heule, Marijn J. H.; Myreen, Magnus O.

Main reference Yong Kiam Tan, Marijn J. H. Heule, Magnus O. Myreen: “cake_lpr: Verified Propagation Redundancy Checking in CakeML”, in Proc. of the Tools and Algorithms for the Construction and Analysis of Systems – 27th International Conference, TACAS 2021, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2021, Luxembourg City, Luxembourg, March 27 – April 1, 2021, Proceedings, Part II, Lecture Notes in Computer Science, Vol. 12652, pp. 223–241, Springer, 2021.


URL http://dx.doi.org/10.1007/978-3-030-72013-1_12

State-of-the-art automated reasoning tools are complex and highly-optimized pieces of software. This complexity can lead to an increased risk of soundness-critical bugs, which may affect the trustworthiness of automatically generated results. To remedy this state of affairs, many tools now generate proof logs (or proof certificates), which can be independently checked for correctness.

This talk is about the “last mile” in highly trustworthy automated reasoning – the development of efficient, formally verified proof checkers that are capable of soundly scrutinizing proof logs for various theories. I will survey theories, proof systems, and proof checkers that have been formalized in proof assistants, including my work with various collaborators on verifying proof checkers using HOL4 and CakeML. Looking ahead, I speculate that today’s foundational software verification tools are well-suited to handle tougher challenges in end-to-end verification of proof checkers. For example, 1) building common infrastructure to ease verification of new proof systems and/or efficient proof checkers; 2) developing a unified proof checker that seamlessly handles proofs from different theories; or 3) verifying proof checkers for proof systems that feature probabilistic or interactive elements.

3.25 A SAT Solver + Computer Algebra Attack on the Minimum Kochen-Specker Problem

Vijay Ganesh (University of Waterloo, CA)

License  Creative Commons BY 4.0 International license
© Vijay Ganesh

Joint work of Vijay Ganesh, Curtis Bright, Brian Li

One of the most fundamental results in the foundations of quantum mechanics is the Kochen–Specker (KS) theorem, a “no-go” theorem which states that contextuality is an essential feature of any hidden-variable theory. The theorem hinges on the existence of a mathematical object called a KS vector system. Although the existence of a KS vector system was first established by Kochen and Specker, the problem of the minimum size of such a system has stubbornly remained open for over 50 years. In this paper, we present a new method based on a combination of a SAT solver and a computer algebra system (CAS) to address this problem. We improve the lower bound on the minimum number of vectors in a KS system from 22 to 23 and improve the efficiency of the search by a factor of over 1000 when compared to the most recent computational methods. Finding a minimum KS system would simplify experimental tests of the KS theorem and have direct applications in quantum information processing, specifically in the security of quantum cryptographic protocols based on complementarity, zero-error classical communication, and dimension witnessing.

3.26 Theoretical limits of 1UIP Learning

Marc Vinyals (University of Newfoundland, CA)

License  Creative Commons BY 4.0 International license
© Marc Vinyals

Even though CDCL can reproduce resolution proofs with at most a polynomial overhead, it is not clear how large that overhead needs to be, or if one is needed at all. We investigate the role that learning schemes play in this simulation by focusing on syntactical properties of proofs generated by CDCL solvers that employ the standard 1UIP learning scheme. In particular we show that proofs of this kind can simulate resolution proofs with at most a linear overhead, but there also exist formulas where such overhead is necessary or, more precisely, that there exist formulas with resolution proofs of linear length that require quadratic CDCL proofs.

3.27 Exponential separations using guarded extension variables

Emre Yolcu (Carnegie Mellon University – Pittsburgh, US)

License  Creative Commons BY 4.0 International license
© Emre Yolcu

Joint work of Emre Yolcu, Heule, Marijn

I talked about the complexity of proof systems augmenting resolution with inference rules that allow, given a formula Γ in conjunctive normal form, deriving clauses that are not necessarily logically implied by Γ but whose addition to Γ preserves satisfiability. When the derived clauses are allowed to introduce variables not occurring in Γ , those systems

become equivalent to extended resolution. We are concerned with their versions “without new variables.” They are called BC^- , RAT^- , SBC^- , and GER^- , denoting respectively blocked clauses, resolution asymmetric tautologies, set-blocked clauses, and generalized extended resolution. Each of them formalizes some restricted version of the ability to make assumptions that hold “without loss of generality,” which is commonly used informally to simplify or shorten proofs. I described several new separations between those systems, which give an almost complete picture of their relative strengths.

4 Evaluation by Participants

In addition to the traditional Dagstuhl evaluation after the workshop, the organizing committee also arranged for a separate evaluation which specific questions about different aspects of the workshop. Below follows a summary of the answers collected in both evaluations.

In the Dagstuhl survey, the scientific quality of the workshop was ranked very highly, and the workshop also scored highly on the questions whether it inspired new ideas, led to insights from neighbouring fields or communities, and inspired new research. Some participants pointed out technical problems during the workshop with the beamer and pointers in the seminar room (which seem to have since been resolved) and with e-mail messages from organizers not getting through to participants via the seminar mailing list.

There were comments about how information about informal talks and gatherings can be spread more efficiently among all participants, and, in particular, about the fact that there was an unfortunate collision between an informal presentation and the music night on Thursday evening.

In the organizers survey (which was filled in by 23 participants – a similar number to the Dagstuhl survey), the decision to have an open problem session was uniformly assessed as good or very good. A majority of respondents to the organizer poll agreed that *not* having a panel discussion was good. A majority of respondents also said that the amount of scheduled activities in the workshop program overall was about right, but a large minority thought it was a bit too much. About an equally large majority found the balance between longer survey talks and shorter contributed talks in the program to be good, but again a large minority found the amount of surveys to be a bit too much, or even clearly too much. Overall, the large number of survey presentations were highlighted by different participants both as a positive and a negative aspect of the workshop in both surveys.

Since the workshop tried to cover a fairly large number of different research areas related to combinatorial solving and optimization, the organizer poll asked the participant whether there was a good balance between depth and breadth. Several participants commented that the coverage of many different areas was a good aspect of the workshop, but there was also a suggestion to maybe focus more clearly on one or two external research areas per workshop in order to be able to go deeper.

Among good aspects to keep for future editions in this workshop series the responses listed:

- keeping the mix of survey and research talks;
- keeping the mix of different research communities;
- setting aside lots of time for informal discussions;
- making sure to invite junior participants, including MSc students.

Some aspects that could be improved were considered to be:

- A bit more slack in the schedule, to allow for longer discussions when talks generated lots of interaction (maybe at the price of starting earlier in the afternoon).
- A better mode of communication than e-mail (or wiki) for spreading information during the workshop (e.g., about informal talks).
- More interaction between theoretical and applied researchers.

All in all, it seems fair to say that the the feedback from the participants was overwhelmingly positive, just as for the Dagstuhl workshop in 2015. When asked if they would come to a similar workshop again in Europe, 22 out of 23 respondents in the organizer poll replied that this is very likely. If such a workshop were instead to be held in North America, a clear majority would still want to come, but the enthusiasm in the responses went down slightly (perhaps reflecting that there was a clear bias of European participants in the workshop this time). For a workshop in East Asia or India, the responses were even more mixed, though there were still more positive than negative replies.

Participants

- Jeremias Berg
University of Helsinki, FI
- Olaf Beyersdorff
Friedrich-Schiller-Universität
Jena, DE
- Armin Biere
Universität Freiburg, DE
- Nikolaj S. Bjørner
Microsoft – Redmond, US
- Benjamin Böhm
Friedrich-Schiller-Universität
Jena, DE
- Bart Bogaerts
VU – Brussels, BE
- Jonas Conneryd
Lund University, SE
- Susanna de Rezende
Lund University, SE
- Katalin Fazekas
TU Wien, AT
- Mathias Fleury
Universität Freiburg, DE
- Vijay Ganesh
University of Waterloo, CA
- Mexi Gioni
Zuse Institut Berlin, DE
- Ambros M. Gleixner
HTW – Berlin, DE
- Malte Helmert
Universität Basel, CH
- Marijn J. H. Heule
Carnegie Mellon University –
Pittsburgh, US
- Matti Järvisalo
University of Helsinki, FI
- Mikoláš Janota
Czech Technical University –
Prague, CZ
- Daniela Kaufmann
TU Wien, AT
- Antonina Kolokolova
University of Newfoundland, CA
- Laura Kovács
TU Wien, AT
- Chunxiao (Ian) Li
University of Waterloo, CA
- Meena Mahajan
The Institute of Mathematical
Sciences – Chennai, IN
- Ciaran McCreesh
University of Glasgow, GB
- Kuldeep S. Meel
National University of
Singapore, SG
- Jakob Nordström
University of Copenhagen, DK &
Lund University, SE
- Andy Oertel
Lund University, SE
- Albert Oliveras
UPC Barcelona Tech, ES
- Pavel Pudlák
The Czech Academy of Sciences –
Prague, CZ
- Torsten Schaub
Universität Potsdam, DE
- Andre Schidler
TU Wien, AT
- Laurent Simon
University of Bordeaux, FR
- Friedrich Slivovsky
TU Wien, AT
- Martin Suda
Czech Technical University –
Prague, CZ
- Stefan Szeider
TU Wien, AT
- Yong Kiam Tan
Infocomm Research –
Singapore, SG
- Dieter Vandesande
VU – Brussels, BE
- Marc Vinyals
University of Newfoundland –
St. John's, CA
- Ryan Williams
MIT – Cambridge, US
- Emre Yolcu
Carnegie Mellon University –
Pittsburgh, US



Intelligent Security: Is “AI for Cybersecurity” a Blessing or a Curse

Nele Mentens^{*1}, Stjepan Picek^{*2}, and Ahmad-Reza Sadeghi^{*3}

1 Leiden University, NL. n.mentens@liacs.leidenuniv.nl

2 Radboud University Nijmegen, NL. stjepan.picek@cs.ru.nl

3 TU Darmstadt, DE. ahmad.sadeghi@trust.tu-darmstadt.de

Abstract

This report documents the outcomes of Dagstuhl Seminar 22412 “Intelligent Security: Is “AI for Cybersecurity” a Blessing or a Curse”. The seminar brought together 25 attendees from 10 countries (Canada, Croatia, Czech Republic, France, Germany, Netherlands, Singapore, Sweden, Switzerland, and the USA). There were 17 male and 8 female participants. Three participants were from the industry, and the rest were from academia.

The gathered researchers are actively working in the domains of artificial intelligence and cybersecurity, emphasizing hardware security, fuzzing, physical security, and network security. The seminar aims to foster sharing experiences and best practices between various cybersecurity applications and understand how and when certain approaches are transferable. The first two days were devoted to 20-minute self-introductions by participants to achieve these goals. At the end of the second day, we made a list of topics that were decided to be the focus of the seminar and that will be discussed in the groups in the next few days. On the third and fourth days, the work was conducted in four discussion groups where at the end of each day, all participants gathered to report the results from the discussion groups and to align the goals. On the last day, we again worked in one group to summarize the findings and foster networking among participants. A hike was organized in the afternoon of the third day. The seminar was a success. The participants actively participated in the working groups and the discussions and went home with new ideas and collaborators. This report gathers the abstracts of the presented talks and the conclusions from the discussion groups, which we consider relevant contributions toward better interdisciplinary research on artificial intelligence and cybersecurity.

Seminar October 9–14, 2022 – <http://www.dagstuhl.de/22412>

2012 ACM Subject Classification Security and privacy → Cryptography; Security and privacy → Intrusion/anomaly detection and malware mitigation; Security and privacy → Security in hardware; Security and privacy → Systems security; Computing methodologies → Artificial intelligence; Computing methodologies → Machine learning; Computer systems organization → Real-time systems

Keywords and phrases Cybersecurity, Artificial Intelligence, Hardware Security, Machine Learning, Explainability

Digital Object Identifier 10.4230/DagRep.12.10.106

* Editor / Organizer



1 Executive Summary

Stjepan Picek (Radboud University, NL)

License  Creative Commons BY 4.0 International license
© Stjepan Picek

In recent years, artificial intelligence (AI) has become an emerging technology to assess security and privacy. Moreover, we can see that AI does not represent “only” one of the options for tackling security problems but instead a state-of-the-art approach. Besides providing better performance, AI also brings automated solutions that can be faster and easier to deploy but are also resilient to human errors. We can only expect that future AI developments will pose even more unique security challenges that must be addressed across algorithms, architectures, and hardware implementations. While there are many success stories when using AI for security, there are also multiple challenges. AI is commonly used in the black-box setting, making the interpretability or explainability of the results difficult. Furthermore, research on AI and cybersecurity commonly look at the various sub-problems in isolation, mostly relying on best practices in the domain. As a result, we often see how techniques are “reinvented”, but also that strong approaches from one application domain are introduced to another only after a long time.

The Dagstuhl Seminar 22412 on Intelligent Security: Is “AI for Cybersecurity” a Blessing or a Curse brought together experts from diverse domains of cybersecurity and artificial intelligence with the goal of facilitating the discussion at different abstraction levels to uncover the links between scaling and the resulting security, with a special emphasis on the hardware perspective. The seminar started with two days of contributed talks by participants. At the end of the second day, every participant suggested topics to be discussed in more detail. From the initial pool of nine topics, we decided to concentrate on four topics on the third and fourth day of the seminar: 1) the explainability of AI for cybersecurity, 2) AI and implementation attacks, 3) AI and fuzzing, and 4) the security of machine learning. The first group approached the problem of the explainability of AI for cybersecurity. The discussion mainly revolved around scenarios where deep learning is used as the attack method, but explainability is necessary to understand why the attack worked and, more importantly, how to propose new defense mechanisms that will be resilient against such AI-based attacks. During the discussion, we considered two perspectives: a) understanding the features and b) understanding deep neural networks.

The second group focused on how AI can improve the performance of implementation attacks. More precisely, we discussed the side-channel analysis and fault injection. Most of the discussion aimed at usages of deep learning for side-channel analysis and evolutionary algorithms for fault injection. However, we also discussed how the lessons learned from one domain could be used in another one. The third group worked on the topic of security fuzzing. We discussed how techniques like evolutionary algorithms are used for evolving diverse mutations and mutation scheduling. At the same time, machine learning is (for now) somewhat less used, but there are many potential scenarios to explore. For instance, instead of using evolutionary algorithms, it should be possible to use reinforcement learning to find mutation scheduling. The fourth group discussed the topic of the security of machine learning. More precisely, it focused on backdoor attacks and federated learning settings. While both attack and defense perspectives were discussed, the discussion group emphasized the need for stronger defenses. Each group followed a cross-disciplinary setting where the participants exchanged groups based on their interests. We had one group switch per day to allow sufficient time to discuss a topic. At the end of each day, all participants joined a

meeting to discuss the findings and tweak the topics for the discussion groups. On the last day of the seminar, all participants worked together on fine-tuning the findings and discussing possible collaborations. The reports of the working groups, gathered in the following sections, constitute the main results from the seminar. We consider them the necessary next step toward understanding the interplay between artificial intelligence and cybersecurity, as well as the interplay among diverse cybersecurity domains using AI. Moreover, we expect that the seminar (and this report) will help better understand the main open problems and how to use techniques from different domains to tackle cybersecurity problems. This will encourage innovative research and help to start joint research projects addressing the issues.

2 Table of Contents

Executive Summary

<i>Stjepan Picek</i>	107
--------------------------------	-----

Overview of Talks

Can AI clone the microarchitecture of a microcontroller? <i>Ileana Buhan</i>	111
Deep Learning Application for Side-Channel Analysis and Fault Injection <i>Lukasz Chmielewski</i>	111
Backdoor Detection in Federated Learning via Deep Layer Predictions <i>Alexandra Dmitrienko</i>	112
Breaking cryptographic algorithms using power and EM side-channels <i>Elena Dubrova</i>	112
Blockchain tools for privacy-preserving machine learning <i>Oğuzhan Ersoy</i>	112
Mitigating Backdoor Attacks in Federated Learning (FL) using Frequency Analysis of the Local Model updates <i>Hossein Fereidooni and Ahmad-Reza Sadeghi</i>	113
Neural Networks: predators and prey <i>Fatemeh Ganji</i>	114
AI for Cybersecurity: a taste of things to come... or papers of future past? <i>Domagoj Jakobovic</i>	115
Hardware Security and Deep Learning <i>Dirmanto Jap</i>	116
AI for fault injection <i>Marina Krcek</i>	117
Assessing the Trustworthiness of AI Systems <i>Jesus Luna Garcia</i>	117
Use cases of side-channel data analysis <i>Damien Marion</i>	118
New Directions in AI-Based Cryptography <i>Luca Mariot</i>	118
High-throughput network intrusion detection based on deep learning <i>Nele Mentens</i>	119
Fuzz testing with machine learning <i>Irina Nicolae</i>	119
Explainability of deep learning-based side-channel analysis <i>Stjepan Picek</i>	120
Engineering Models versus Scientific Models <i>Patrick Schaumont</i>	120
Remote Electrical-Level Attacks on Cloud FPGAs: The Role of AI <i>Mirjana Stojilović</i>	121

AI-Assisted System-level Tamper Detection <i>Shahin Tajik</i>	121
Peek into the Black-Box: Interpretable Neural Network using SAT Equations in Side-Channel Analysis <i>Trevor Yap</i>	122
Working Groups	
Explainability of AI in Cybersecurity <i>Stjepan Picek, Nele Mentens</i>	122
AI for Implementation Attacks <i>Stjepan Picek, Nele Mentens</i>	124
Security Fuzzing <i>Stjepan Picek</i>	125
Security of Machine Learning <i>Stjepan Picek</i>	126
Participants	128

3 Overview of Talks

3.1 Can AI clone the microarchitecture of a microcontroller?

Ileana Buhan (Radboud University Nijmegen, NL)

License © Creative Commons BY 4.0 International license
© Ileana Buhan

Early attempts to create automated tooling and the recently increased efforts toward this purpose prove the appeal of leakage simulators. A leakage simulator translates a sequence of assembly instructions into a power trace. The challenge for the wide-scale adoption lies in the manual effort required to create a leakage simulator. ABBY is the first post-silicon leakage simulator, where we used deep learning to automate the profiling of the target.

3.2 Deep Learning Application for Side-Channel Analysis and Fault Injection

Lukasz Chmielewski (Radboud Universiteit Nijmegen, NL & Masaryk University – Brno, CZ)

License © Creative Commons BY 4.0 International license
© Lukasz Chmielewski

Joint work of Guilherme Perin, Lejla Batina, Stjepan Picek, Madura Shelton, Niels Samwel, Markus Wagner, Leo Weissbart, Yuval Yarom

Main reference Guilherme Perin, Lukasz Chmielewski, Lejla Batina, Stjepan Picek: “Keep it Unsupervised: Horizontal Attacks Meet Deep Learning”, IACR Trans. Cryptogr. Hardw. Embed. Syst., Vol. 2021(1), pp. 343–372, 2021.

URL <https://doi.org/10.46586/tches.v2021.i1.343-372>

Main reference Lukasz Chmielewski, Leo Weissbart: “On Reverse Engineering Neural Network Implementation on GPU”, in Proc. of the Applied Cryptography and Network Security Workshops – ACNS 2021 Satellite Workshops, AIBlock, AIHWS, AIoTS, CIMSS, Cloud S&P, SCI, SecMT, and SiMLA, Kamakura, Japan, June 21–24, 2021, Proceedings, Lecture Notes in Computer Science, Vol. 12809, pp. 96–113, Springer, 2021.

URL https://doi.org/10.1007/978-3-030-81645-2_7

Main reference Madura A. Shelton, Lukasz Chmielewski, Niels Samwel, Markus Wagner, Lejla Batina, Yuval Yarom: “Rosita++: Automatic Higher-Order Leakage Elimination from Cryptographic Code”, in Proc. of the CCS ’21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 – 19, 2021, pp. 685–699, ACM, 2021.

URL <https://doi.org/10.1145/3460120.3485380>

This presentation covers selected topics in Deep Learning (DL) applications to physical attacks, including Side-Channel Analysis (SCA) and Fault Injection (FI). The following topics are covered: horizontal attack against Elliptic Curve Cryptography (ECC) and RSA, XYZ hotspot selection (SCA & FI), attacks against hardware DL accelerators, and DL-based power simulators.

3.3 Backdoor Detection in Federated Learning via Deep Layer Predictions

Alexandra Dmitrienko (Universität Würzburg, DE)

License © Creative Commons BY 4.0 International license
© Alexandra Dmitrienko

Joint work of Phillip Rieger, Torsten Krauß, Markus Miettinen, Alexandra Dmitrienko, Ahmad-Reza Sadeghi
Main reference Phillip Rieger, Torsten Krauß, Markus Miettinen, Alexandra Dmitrienko, Ahmad-Reza Sadeghi: “Close the Gate: Detecting Backdoored Models in Federated Learning based on Client-Side Deep Layer Output Analysis”, CoRR, Vol. abs/2210.07714, 2022.
URL <https://doi.org/10.48550/arXiv.2210.07714>

This talk discusses the challenges of backdoor detection in federated learning (FL) related to adaptive attackers and non-independent and identically distributed (non-IID) data. It then presents an approach to identify backdoored local contribution of FL clients by analyzing local client predictions of deep learning layers and comparing those to predictions made by a global model. The approach can handle an extended non-IID scenarios compare to the related work and is resilient to adaptive adversaries.

3.4 Breaking cryptographic algorithms using power and EM side-channels

Elena Dubrova (KTH Royal Institute of Technology – Kista, SE)

License © Creative Commons BY 4.0 International license
© Elena Dubrova

Side-channel attacks are one of the most efficient physical attacks against implementations of cryptographic algorithms at present. They exploit the correlation between physical measurements (power consumption, electromagnetic emissions, timing) taken at different points during the algorithm’s execution and the secret key. In this talk, I will give an introduction to power and EM-based side-channel attacks and present some of our recent results.

3.5 Blockchain tools for privacy-preserving machine learning

Oğuzhan Ersoy (TU Delft, NL)

License © Creative Commons BY 4.0 International license
© Oğuzhan Ersoy

In recent years, blockchain technology get the attention of both industry and academia. Thanks to the interest, there are several cryptographic tools developed for decentralized systems that can be used in other domains including machine learning. Among these tools, VDF, VRF, and adaptor signatures are mentioned in this talk. Firstly, VDFs (Verifiable Delay Functions) allow a prover to show a verifier that a certain amount of time running a function was spent. In a machine learning setting, VDFs can be used to limit the number of queries on a machine learning model. Specifically, by requesting parties to provide VDF proofs when they query the model, we can restrict the number of queries sent to the system. Compared to proof-of-work-based techniques [1], VDF-based query limitations would also guarantee that the adversary cannot parallelize the VDF challenge. Secondly, VRFs (Verifiable Random

Functions) are used to generate random numbers that can be verifiable by all parties involved. In collaborative machine learning, this can be used, for example, cryptographic sortition and leader selection [2]. In this selection, an adversary would not be able to predict the leader in advance. Finally, adaptor signatures allow parties to embed a condition into the signature. It has been used to improve the fungibility of transactions in the blockchain domain. However, it is yet an open question how to utilize adaptor signatures in the machine learning domain.

References

- 1 Adam Dziedzic; Muhammad Ahmad Kaleem; Yu Shen Lu; Nicolas Papernot, *Increasing the Cost of Model Extraction with Calibrated Proof of Work*, International Conference on Learning Representations (ICLR), 2022.
- 2 Rui Wang; Oğuzhan Ersoy; Hangyu Zhu; Yaochu Jin; Kaitai Liang, *FEVERLESS: Fast and Secure Vertical Federated Learning based on XGBoost for Decentralized Labels*, IEEE Transactions on Big Data, 1–15, 2022.

3.6 Mitigating Backdoor Attacks in Federated Learning (FL) using Frequency Analysis of the Local Model updates

Hossein Fereidooni (TU Darmstadt, DE) and Ahmad-Reza Sadeghi (TU Darmstadt, DE)

License  Creative Commons BY 4.0 International license

© Hossein Fereidooni and Ahmad-Reza Sadeghi

Joint work of Hossein Fereidooni, Alessandro Pegoraro, Phillip Rieger, Ahmad-Reza Sadeghi

Federated learning (FL) is a distributed machine learning technique enabling participating clients to collaboratively learn a shared global model without sharing their potentially private data. Despite its benefits (i.e., communication efficiency and reduced requirements for hardware), federated learning has been shown to be vulnerable to adversarial threats such as backdoor attacks where the adversary stealthily manipulates the global model so that adversary-selected inputs result in adversary-selected outputs. Although there are multiple defense mechanisms proposed by previous works, the backdoor attacks with sophisticated hiding techniques still pose a threat to FL. Existing defense solutions cannot fully mitigate backdoor attacks and have a number of deficiencies such as unrealistic assumptions for data distributions and attack strategies. The core idea of this talk is that backdoored model might be related to frequency analyses of neural networks. We are going to we investigate a relationship between backdoor and frequency components of the model parameters (i.e., weights) that can be used for model filtering during the aggregation process in FL to implement backdoor attack defense. More specifically, we set up the FL process and implement state-of-the-art backdoor attacks (i.e., Semantic attack, Stealthy Model Poisoning, etc.) and then transform tensor weights (i.e., local model updates) to the frequency domain and apply frequency analysis (i.e., Discrete Cosine Transform – DCT) to find a relationship between backdoor patterns and frequency components of the weights.

References

- 1 E. Bagdasaryan, Andreas Veit, Yiqing Hua, D. Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In AISTATS, 2020.
- 2 A. Bhagoji, Supriyo Chakraborty, Prateek Mittal, and S. Calo. Analyzing federated learning through an adversarial lens. In ICML, 2019.
- 3 Clement Fung, Chris J. M. Yoon, and Ivan Beschastnikh. Mitigating sybils in federated learning poisoning. ArXiv, abs/1808.04866, 2018.

- 4 Tianyu Gu, Brendan Dolan-Gavitt, and S. Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. ArXiv, abs/1708.06733, 2017.
- 5 Hongyi Wang, Kartik K. Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. ArXiv, abs/2007.05084, 2020.
- 6 Chen Wu, Xiangwen Yang, Sencun Zhu, and P. Mitra. Mitigating backdoor attacks in federated learning. ArXiv, abs/2011.01767, 2020.
- 7 Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. ArXiv, abs/1807.01251, 2019.
- 8 Yi Zeng, Won Park, Z. M. Mao, and R. Jia. Rethinking the backdoor attacks’ triggers: A frequency perspective. ArXiv, abs/2104.03413, 2021.

3.7 Neural Networks: predators and prey

Fatemeh Ganji (Worcester Polytechnic Institute, US)

License © Creative Commons BY 4.0 International license
© Fatemeh Ganji

Joint work of Fatemeh Ganji, Domenic Forte, Rabin Acharya, Mohammad Hashemi, Steffi Roy

Main reference Rabin Yu Acharya, Fatemeh Ganji, Domenic Forte: “Information Theory-based Evolution of Neural Networks for Side-channel Analysis”, IACR Trans. Cryptogr. Hardw. Embed. Syst., Vol. 2023(1), pp. 401–437, 2023.

URL <https://doi.org/10.46586/tches.v2023.i1.401-437>

Main reference Mohammad Hashemi, Steffi Roy, Domenic Forte, Fatemeh Ganji: “HWGN²: Side-Channel Protected NNs Through Secure and Private Function Evaluation”, in Proc. of the Security, Privacy, and Applied Cryptography Engineering – 12th International Conference, SPACE 2022, Jaipur, India, December 9-12, 2022, Proceedings, Lecture Notes in Computer Science, Vol. 13783, pp. 225–248, Springer, 2022.

URL https://doi.org/10.1007/978-3-031-22829-2_13


This talk covers two main topics relevant to how neural networks (NNs) have become a powerful tool to assess the security of cryptographic primitives and how NNs themselves have been targeted to extract their assets. The first part of the talk is devoted to NN-enabled side-channel analysis (SCA), in particular, profiled SCA that leverages leakage from cryptographic implementations to extract the secret key. It is known that when combined with advanced methods in NNs, profiled SCA can successfully attack even crypto-cores with protection devised to impair the effectiveness of SCA. Similar to other machine learning tasks, a range of questions have remained unanswered about NN-enabled SCA, namely: how to choose an NN with an adequate configuration, how to tune the NN’s hyperparameters, when to stop the training, etc. This talk introduces “InfoNEAT,” which tackles these issues in a natural way. InfoNEAT relies on the concept of neural structure search (NAS), enhanced by information-theoretic metrics to guide the evolution, halt it with novel stopping criteria, and improve time-complexity and memory footprint. Besides the considerable advantages regarding the automated configuration of NNs, InfoNEAT demonstrates significant improvements over other approaches for effective key recovery in terms of the number of epochs and the number of attack traces compared to both MLPs and CNNs, as well as a reduction in the number of trainable parameters compared to MLPs. Furthermore, through experiments, it is demonstrated that InfoNEAT’s models are robust against noise and desynchronization in traces.

In the second part of the talk, SCA against NNs has been taken into account. In fact, recent work has highlighted the risks of intellectual property (IP) piracy of deep learning (DL) models from the side-channel leakage of DL hardware accelerators. In response, fundamental cryptographic approaches, specifically built upon the notion of multi-party computation,

could potentially improve the robustness against side-channel leakage. To examine this and weigh the costs and benefits, we introduce hardware garbled NN (HWGN²), a DL hardware accelerator implemented on FPGA. HWGN² also provides NN designers with the flexibility to protect their IP in real-time applications, where hardware resources are heavily constrained, through a hardware-communication cost trade-off. Concretely, we apply garbled circuits, implemented using a MIPS architecture that achieves up to 62.5× fewer logical and 66× less memory utilization than the state-of-the-art approaches at the price of communication overhead. Further, the side-channel resiliency of HWGN² is demonstrated by employing the test vector leakage assessment (TVLA) test against both power and electromagnetic side channels.

3.8 AI for Cybersecurity: a taste of things to come... or papers of future past?

Domagoj Jakobovic (University of Zagreb, HR)

License  Creative Commons BY 4.0 International license
© Domagoj Jakobovic

Designing a secure system requires a lot of expertise in the security domain. In that process, some of the tasks can be automated with the help of Artificial Intelligence (AI). The use of AI methods does not aim to replace the human designer; rather, they can help in the design optimization process, where standardized algorithms can be readily applied to increase the efficiency. As long as a complex system design task can be decomposed into simpler elements, AI methods can substantially facilitate the optimization of individual components. Furthermore, most methods can be used to optimize an arbitrary (set of) design criteria.

However, although there are problems that can be efficiently solved with AI techniques, it is not always obvious *which* AI technique or optimization algorithm should be applied. In practice, a bit of knowledge in both domains is needed to select the appropriate method and to efficiently apply it to the problem at hand. Even then, for many AI methods there are no formal guarantees of efficiency, which is especially evident for obscure machine learning models such as deep neural networks.

Ideally, the AI component should provide *explainability*, so the decision making process can be justified at each step. We may even employ less efficient but explainable models to evaluate obscure models which bring performance. There are use cases in which a part of a black-box model may be replaced with an equivalent white-box component offering the same level of performance. Additionally, different optimization algorithms may be used to prune “fat” models, either to provide insight into their functionality or to reduce application complexity. In this regard, neuroevolution methods may be used to design and optimize the structure and hyperparameters of deep neural models.

The application of the above techniques can be found in model building efforts in various domains; the usual goals are knowledge representation, model parameter optimization, feature extraction and selection, etc. Some of the efficient examples of this paradigm are already evident in cryptology and security where different AI techniques, most notably evolutionary algorithms, have been applied. Here, the focus was mainly on the design of different cryptography primitives, such as Boolean functions, S-boxes and pseudo-random number generators. Successful applications also include fault injection, intrusion detection, hyper-parameter optimization etc. Recently, evolutionary algorithm methods have also been applied to fuzzing, where they obtained competitive performance in a target-based comparison with commonly used solutions.

3.9 Hardware Security and Deep Learning

Dirmanto Jap (Nanyang TU – Singapore, SG)

License © Creative Commons BY 4.0 International license
© Dirmanto Jap

Joint work of Yoo-Seung Won, Xiaolu Hou, Dirmanto Jap, Jakub Breier, Shivam Bhasin, Soham Chatterjee, Arindam Basu Leijla Batina, Stjepan Picek

Main reference Yoo-Seung Won, Xiaolu Hou, Dirmanto Jap, Jakub Breier, Shivam Bhasin: “Back to the Basics: Seamless Integration of Side-Channel Pre-Processing in Deep Neural Networks”, *IEEE Trans. Inf. Forensics Secur.*, Vol. 16, pp. 3215–3227, 2021.

URL <https://doi.org/10.1109/TIFS.2021.3076928>

Main reference Yoo-Seung Won, Soham Chatterjee, Dirmanto Jap, Arindam Basu, Shivam Bhasin: “DeepFreeze: Cold Boot Attacks and High Fidelity Model Recovery on Commercial EdgeML Device”, in *Proc. of the IEEE/ACM International Conference On Computer Aided Design, ICCAD 2021, Munich, Germany, November 1-4, 2021*, pp. 1–9, IEEE, 2021.

URL <https://doi.org/10.1109/ICCAD51958.2021.9643512>

Main reference Leijla Batina, Shivam Bhasin, Dirmanto Jap, Stjepan Picek: “CSI NN: Reverse Engineering of Neural Network Architectures Through Electromagnetic Side Channel”, in *Proc. of the 28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019*, pp. 515–532, USENIX Association, 2019.

URL <https://www.usenix.org/conference/usenixsecurity19/presentation/batina>

In this presentation, we provided the discussion on two main direction on the area of hardware security and deep learning (DL). First, we discussed about the use of feature extraction or pre-processing techniques, which could help improving the performance of DL based side-channel attacks (SCA). In most of the research works done, the main goal is towards the direction of designing an efficient network that can provide the best attacks against each side-channel trace dataset. On the other hand, little work has been done to investigate the possibility of strengthening DL architecture with the capability of integrating existing side-channel pre-processing or filtering techniques, which have been thoroughly investigated over the past decades. As such, one of the aim is to minimize the necessity for architecture adjustments while enabling seamlessly integration of pre-processing method for attack. In our work, we propose to incorporate feature extraction and classification in a single framework by using a multi-branch model. The experimental results indicated that the model can perform better than the benchmark model even though it is not specifically tailored for the dataset. These show that it is an inherent property of MCNN which allows it to learn more feature representations and result in better attacks. As for the potential future direction, we discussed the possibility of using other DL based approach as a way to further automate the feature pre-processing method.

Next, we discussed about the vulnerability of DL implementation on physical device against side-channel and fault attacks. Due to the rapid growth of DL application, more and more efforts are being allocated to build and train critical DL models. These DL models have then become valuable Intellectual Properties (IPs) that cost companies lots of time and resources, which inadvertently attract malicious parties to steal them. We presented the work on model extraction and reverse engineering of the neural networks model through electromagnetic (EM) side-channel leakage. We also presented alternative work for reverse engineering of neural network models through cold boot attacks. The work is then conducted targeting edge AI hardware accelerators, Intel Neural Compute Stick 2 (NCS2). It is based on the observation that the model architecture and parameters have to be loaded to Intel NCS2 before the inference, and thus, by performing cold boot attack on host device, it is possible to recover the information, albeit with correction required. As for potential future direction, we proposed to investigate different target devices or more complex architectures. We also discussed on possible countermeasures for the implementation as well as the security evaluation of these countermeasures.

3.10 AI for fault injection

Marina Krcek (TU Delft, NL)

License © Creative Commons BY 4.0 International license
© Marina Krcek

Joint work of Marina Krcek, Thomas Ordas, Daniele Fronte, Stjepan Picek

Main reference Marina Krcek, Thomas Ordas, Daniele Fronte, Stjepan Picek: “The More You Know: Improving Laser Fault Injection with Prior Knowledge”, in Proc. of the Workshop on Fault Detection and Tolerance in Cryptography, FDTC 2022, Virtual Event / Italy, September 16, 2022, pp. 18–29, IEEE, 2022.

URL <https://doi.org/10.1109/FDTC57191.2022.00012>

Fault injection types such as laser FI, electromagnetic FI, or voltage glitching have different parameters to define. Nevertheless, the parameter search space becomes large for all types because of many parameters and possibilities. Since the search space is large, commonly used methods like grid and random search lead to suboptimal performance/results. We use AI techniques discussed in this talk to improve the efficiency of the search. Specifically, genetic and memetic algorithms from evolutionary computation were shown to find more parameter combinations that lead to erroneous outputs compared to random search [1]. Additionally, hyperparameter tuning methods like successive halving and reinforcement learning from the machine learning domain were also shown to be quite successful [2, 3]. On the other hand, machine learning can be helpful for transferability issues in fault injection. As discussed during the talk, we can use prior knowledge from tested devices and parameter combinations generalized with decision trees to find more vulnerabilities on a new target or bench in the same amount of tested parameters.

References

- 1 Maldini, Antun; Samwel, Niels; Picek, Stjepan; Batina, Lejla, Genetic algorithm-based electromagnetic fault injection. In: 2018 Workshop on Fault Diagnosis and Tolerance in Cryptography (FDTC). IEEE, 2018. p. 35-42.
- 2 Werner, Vincent; Maingault, Laurent; Potet, Marie-Laure, Fast calibration of fault injection equipment with hyperparameter optimization techniques. In: Smart Card Research and Advanced Applications: 20th International Conference, CARDIS 2021, Lübeck, Germany, November 11–12, 2021, Revised Selected Papers. Cham: Springer International Publishing, 2022. p. 121-138.
- 3 Moradi, Mehrdad; Oakes, Bentley James; Saraoglu, Mustafa; Morozov, Andrey; Janschek, Klaus; Denil, Joachim, Exploring fault parameter space using reinforcement learning-based fault injection. In: 2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W). IEEE, 2020. p. 102-109.

3.11 Assessing the Trustworthiness of AI Systems

Jesus Luna Garcia (Robert Bosch GmbH – Stuttgart, DE)

License © Creative Commons BY 4.0 International license
© Jesus Luna Garcia

Joint work of Parmar, Manojkumar Somabhai; Serna, Jetzabel

Main reference European Commission, On Artificial Intelligence – A European approach to excellence and trust. 2020.

URL https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

Despite the topic of AI (cyber)security has received lots of academic and industrial attention in recent years, these communities have started to realize the need for a holistic approach related to this topic. We do not mean only from a system perspective, where the different

implementation layers (e.g., cloud) also contribute to the security (and even to the attack surface) of the AI-application, but also from equally important features like privacy, transparency/explainability, bias, and safety (to name just a few). Extrapolating relevant security research to this much needed holistic approach is critical for the uptake of trusted AI system. This talk discusses some relevant industrial and regulation-related aspects on the field of AI trustworthiness, along with few identified challenges which are being tackled from an EU perspective. One of the main points relates to the need of developing a framework for supporting the assessment of AI systems for cybersecurity certification purposes. The referred framework should be able to leverage realistic levels of automation which can pave the road for continuous (automated) certification. It is expected that such a framework might provide support for accelerating the uptake of relevant standards and regulations like the EU AI Act.

3.12 Use cases of side-channel data analysis

Damien Marion (IRISA – Rennes, FR)

License  Creative Commons BY 4.0 International license
© Damien Marion

Joint work of Damien Marion, Duy-Phuc Pham, Annelie Heuser

Abstract. In this talk, we went through different use cases of side-channel analysis for different security purposes. The first use case was the analysis of micro-architectural leakage, in order to address the gap between leakage and unknown micro-architectures. The second use case was the usage of electromagnetic leakage to classify and/or detect malware and rootkits[1, 2]. Then the talk quickly tackled some problems of securing PQ-cryptography from side-channel point of view. From a more general point of view, side-channel analysis could be viewed as a subpart of data analysis for security. How to extract or quantify sensitive information present in huge amounts of noise data, this where IA (or machine learning) can solve existing issues.

References

- 1 Duy-Phuc Pham, Damien Marion, and Annelie Heuser, ULTRA: Ultimate Rootkit Detection over the Air. 25th International Symposium on Research in Attacks, Intrusions and Defenses, RAID 2022, Limassol, Cyprus, October 26-28, 2022 (2022)
- 2 Duy-Phuc Pham, Damien Marion, Matthieu Mastio, and Annelie Heuser, Obfuscation Revealed: Leveraging Electromagnetic Signals for Obfuscated Malware Classification. ACSAC '21: Annual Computer Security Applications Conference, Virtual Event, USA, December 6 – 10, 2021 (2021)

3.13 New Directions in AI-Based Cryptography

Luca Mariot (Radboud University Nijmegen, NL)

License  Creative Commons BY 4.0 International license
© Luca Mariot

Main reference Luca Mariot, Domagoj Jakobovic, Thomas Bäck, Julio Hernandez-Castro: “Artificial Intelligence for the Design of Symmetric Cryptographic Primitives”, pp. 3–24, 2022.

URL https://doi.org/10.1007/978-3-030-98795-4_1

In this talk, we give a general overview of AI methods and computational models to design cryptographic primitives. These include the use of bio-inspired optimization techniques (particularly evolutionary algorithms) to construct symmetric primitives with good cryptographic properties, like Boolean functions and S-boxes. The approach leverages also on the

use of AI computational models like Cellular Automata (CA) as an efficient representation technique for such primitives. In the second part of the talk, new directions of research are illustrated based on the experience gained with regard to the above AI methods and models. In particular, we focus on the use of evolutionary algorithms to design algebraic constructions of symmetric primitives, to evolve differential distinguishers for small symmetric ciphers, and to explore the space of adversarial examples in machine learning models. Particular emphasis is given to the inherent interpretability and explainability of the solutions provided by evolutionary algorithms, specifically in the case of Genetic Programming (GP).

3.14 High-throughput network intrusion detection based on deep learning

Nele Mentens (Leiden University, NL)

License © Creative Commons BY 4.0 International license
© Nele Mentens

Joint work of Nele Mentens, Laurens Le Jeune

The evolution of our digital society relies on networks that can handle an increasing amount of data, exchanged by an increasing number of connected devices at an increasing communication speed. With the growth of the online world, criminal activities also extend onto the Internet. Network Intrusion Detection Systems (NIDSs) detect malicious activities by analyzing network data. While neural network-based solutions can effectively detect various attacks in an offline setting, it is not straightforward to deploy them in high-bandwidth online systems. This talk elucidates why Field-Programmable Gate Arrays (FPGAs) are the preferred platforms for online network intrusion detection, and which challenges need to be overcome to develop FPGA-based NIDSs for Terabit Ethernet networks.

3.15 Fuzz testing with machine learning

Irina Nicolae (Robert Bosch GmbH, Bosch Center for AI – Stuttgart, DE)

License © Creative Commons BY 4.0 International license
© Irina Nicolae

Joint work of Maria-Irina Nicolae, Max Eisele, Andreas Zeller

Fuzzing – testing software and hardware with randomly generated inputs – has gained significant traction due to its success in exposing program vulnerabilities automatically. Machine learning has increasingly been applied to different parts of the fuzzing loop, with the goal of improving fuzzing efficiency. In this talk, we examine *neural program smoothing* for fuzzing, a family of methods that approximate the tested program with a neural network for novel test case generation. We uncover fundamental and practical limitations of neural program smoothing, which prevent it from reaching its advertised performance and limit its practical interest.

3.16 Explainability of deep learning-based side-channel analysis

Stjepan Picek (Radboud University, NL)

License © Creative Commons BY 4.0 International license
© Stjepan Picek

Joint work of Guilherme Perin, Lichao Wu, Stjepan Picek

Main reference Guilherme Perin, Lichao Wu, Stjepan Picek: “I Know What Your Layers Did: Layer-wise Explainability of Deep Learning Side-channel Analysis”, IACR Cryptol. ePrint Arch., p. 1087, 2022.

URL <https://eprint.iacr.org/2022/1087>

Deep learning-based side-channel analysis is an extremely powerful option as it can work without feature engineering and defeats various hiding and masking countermeasures. Still, from the evaluator’s perspective, even after a successful evaluation (attack), a crucial detail is missing: how did the neural network break the target? Thus, the explainability of deep learning-based side-channel analysis becomes an important issue. Unfortunately, up to now, there are only sporadic attempts to understand how neural network defeats countermeasures and none that gives the complete answer. Some early explored techniques include SVCCA [1] and ablation [2]. While good first steps, these techniques do not provide enough information to understand how countermeasures are circumvented. This talk concentrated on a recent approach to explaining the deep learning-based side-channel attack: layer-wise explainability and its comparative advantages over previous approaches.

References

- 1 Daan van der Valk, Stjepan Picek, Shivam Bhasin: Kilroy Was Here: The First Step Towards Explainability of Neural Networks in Profiled Side-Channel Analysis. COSADE 2020: 175-199.
- 2 Lichao Wu, Yoo-Seung Won, Dirmanto Jap, Guilherme Perin, Shivam Bhasin, Stjepan Picek: Explain Some Noise: Ablation Analysis for Deep Learning-based Physical Side-channel Analysis. IACR Cryptol. ePrint Arch. 2021: 717 (2021).

3.17 Engineering Models versus Scientific Models

Patrick Schaumont (Worcester Polytechnic Institute, US)

License © Creative Commons BY 4.0 International license
© Patrick Schaumont

Cybersecurity implementations, in hardware or software, are created from engineering models, not from scientific models. Scientific models reflect the laws of nature in formulae, while engineering models aim at the opposite: we use the laws of nature to mimic an abstraction.

The observations of secure implementations in the real world are noisy distortions from the ideal, noiseless engineering models. However, we *know* that the ground truth corresponds to the engineering model, which is noiseless and undistorted.

This has an important consequence on machine learning applications. We can use simulation (of engineering models) to create a ground truth to improve inference on measured, distorted implementation. For example, using simulated data, we can build attacks on real-world systems that outperform real-world measurements [1].

References

- 1 Dillibabu Shanmugam, Patrick Schaumont, “Improving Side-channel Leakage Assessment using Pre-silicon Leakage Models,” 14th International Workshop on Constructive Side-channel Analysis and Secure Design (COSADE 2023), Munch, Germany, April 2023.

3.18 Remote Electrical-Level Attacks on Cloud FPGAs: The Role of AI

Mirjana Stojilović (EPFL – Lausanne, CH)

License © Creative Commons BY 4.0 International license
© Mirjana Stojilović

Main reference Ognjen Glamocanin, Louis Coulon, Francesco Regazzoni, Mirjana Stojilovic: “Are Cloud FPGAs Really Vulnerable to Power Analysis Attacks?”, in Proc. of the 2020 Design, Automation & Test in Europe Conference & Exhibition, DATE 2020, Grenoble, France, March 9-13, 2020, pp. 1007–1010, IEEE, 2020.

URL <https://doi.org/10.23919/DATE48585.2020.9116481>

Field-programmable gate arrays (FPGAs) have made their way into the cloud, allowing users to gain remote access to the state-of-the-art reconfigurable fabric and implement their custom accelerators. As FPGAs are large enough to accommodate multiple independent designs, the multi-tenant user scenario may soon be prevalent in cloud computing environments. However, shared FPGAs are vulnerable to remote power-side channel and fault-injection attacks [1, 3, 4]. Machine learning (ML) further broadens the attack space: (1) ML accelerators may be the targets of remote attacks, (2) ML techniques can be used to infer the type of workloads or the computations the FPGA is running [2], and (3) ML can help detecting malicious circuits in FPGA bitstreams. This talk has two parts: In the first, the techniques enabling remote electrical-level attacks on cloud FPGAs are explained. In the second, the opportunities for using ML for detecting and locating malicious activity, or for guiding the cloud hypervisors in managing the FPGA users in a security-aware manner are discussed.

References

- 1 Ognjen Glamočanin; Louis Coulon; Francesco Regazzoni; Mirjana Stojilović, *Are Cloud FPGAs Really Vulnerable to Power Analysis Attacks?*, in Proceedings of the Design, Automation and Test in Europe Conference and Exhibition (DATE), 1007–10, 2020.
- 2 Ognjen Glamočanin; Hajira Bazaz; Mathias Payer; Mirjana Stojilović, *Temperature Impact on Remote Power Side-Channel Attacks on Shared FPGAs*, in Proceedings of the Design, Automation and Test in Europe Conference and Exhibition (DATE), 1–6, 2023.
- 3 Dina G. Mahmoud; Samah Hussein; Vincent Lenders; Mirjana Stojilović, *FPGA-to-CPU Undervolting Attacks*, in Proceedings of the Design, Automation and Test in Europe Conference and Exhibition (DATE), 999–1004, 2022.
- 4 Dina G. Mahmoud; Mirjana Stojilović, *Timing Violation Induced Faults in Multi-Tenant FPGAs*, in Proceedings of the Design, Automation and Test in Europe Conference and Exhibition (DATE), 1745–50, 2019.

3.19 AI-Assisted System-level Tamper Detection

Shahin Tajik (Worcester Polytechnic Institute, US)

License © Creative Commons BY 4.0 International license
© Shahin Tajik

Joint work of Shahin Tajik, Tahoura Mosavirik, Patrick Schaumont

To mount physical attacks adversaries might need to place probes in the proximity of the integrated circuits (ICs) package, create physical connections between their probes/wires and the system’s PCB, or physically tamper with the PCB’s components, chip’s package, or substitute the entire PCB to prepare the device for the attack. In this talk, inspired by methods known from the field of power integrity analysis, we show how the impedance

characterization of the system’s power distribution network (PDN) using an on-chip circuit-based network analyzer can detect various categories of tamper events. By analyzing the frequency response of the system different classes of tamper events from board to chip level are revealed. Using the Wasserstein Distance as a metric, we demonstrate that we can confidently detect tamper events. We demonstrate that even environment-level tampering activities, e.g., proximity of contactless EM probes to the IC package or slightly polished IC package, can be detected using on-chip impedance sensing.

3.20 Peek into the Black-Box: Interpretable Neural Network using SAT Equations in Side-Channel Analysis

Trevor Yap (Nanyang TU – Singapore, SG)

License © Creative Commons BY 4.0 International license
© Trevor Yap

Joint work of Trevor Yap, Adrien Benamira, Shivam Bhasin, Thomas Peyrin

Main reference Trevor Yap, Adrien Benamira, Shivam Bhasin, Thomas Peyrin: “Peek into the Black-Box: Interpretable Neural Network using SAT Equations in Side-Channel Analysis”, 2022.

URL <https://eprint.iacr.org/2022/1247>

Deep neural networks (DNN) have become a significant threat to the security of cryptographic implementations with regards to side-channel analysis (SCA), as they automatically combine the leakages without any preprocessing needed, leading to a more efficient attack. However, these DNNs for SCA remain mostly black-box algorithms that are very difficult to interpret. Benamira *et al.* recently proposed an interpretable neural network called Truth Table Deep Convolutional Neural Network (TT-DCNN), which is both expressive and easier to interpret. In particular, a TT-DCNN has a transparent inner structure that can entirely be transformed into SAT equations after training. This talk gives a brief outline of why we need explainability, and on what TT-DCNN is. The talk also presented a way to analyse the SAT equations of TT-DCNN and show some results. Furthermore, we give a possible direction to analyse this paper.

4 Working Groups

4.1 Explainability of AI in Cybersecurity

Stjepan Picek (Radboud University, NL)

Nele Mentens (Leiden University, NL)

License © Creative Commons BY 4.0 International license
© Stjepan Picek, Nele Mentens

The explainability of AI in cybersecurity represents an important problem since often, it is not sufficient to only have a successful solution. Still, we also must explain why that solution works. For instance, in side-channel analysis, from the perspective of a security evaluator, it is important to know how secure a target is. But, if the target gets broken, a necessary step is to report back to the implementation designers and explain what went wrong (e.g., how a countermeasure got broken). Unfortunately, while deep learning can break various targets, the explainability part is still very much unexplored and vague [6, 4, 7]. For instance, in deep learning-based side-channel analysis, the state-of-the-art approaches can easily break

implementations protected with various countermeasures (masking, hiding, or a combination of masking and hiding). At the same time, understanding why the attack works is based on intuition or general terms from the machine learning domain, e.g., desynchronization is defeated due to the spatial invariance of convolutional neural networks.

Furthermore, deep learning has recently been shown to be a very powerful option in mounting cryptanalysis attacks where the neural networks serve as distinguishers. More precisely, the differential-neural distinguishers are based on distinguishing ciphertext-pairs that belong to a fixed plaintext difference from random ones. While the approach works well, and for several ciphers, the researchers managed to find attacks that are at least competitive with classical approaches. Unfortunately, even after the successful attack, the question remains why the attack works and how to fix the cipher to make it more secure. Works addressing such issues are sparse and far from conclusive [2, 1, 3, 5].

The discussion centered on two questions we consider at the core of explainability. Finally, the discussion from this group was also connected with other discussion groups since explainability is of relevance whenever applying AI in cybersecurity.

- Why?
 - To improve the model: more efficient implementation, more powerful in solving the intended task (e.g., getting the key, increasing the performance metrics, lowering the number of false alarms), more efficient test cases for fuzzing.
 - To improve the security of the implementation against attacks (e.g., SCA, crypto): understand the vulnerabilities of the implementation under attack, fix the implementation based on the position of the leakage, and fix the countermeasure based on the discovered vulnerabilities.
 - To improve trust in the model: important in intrusion detection systems, lower the number of false positives and false negatives, enable application in online systems.
 - To contribute to the security of AI: discover which parts are weak against backdoors, etc.
- How?
 - Understand the features:
 - * Feature visualization: activation maximization, code inversion.
 - * Feature attributions: LIME, occlusion, delivery maps, Shapley values.
 - * Rule extraction: DeepRed, scalability challenges (data, model).
 - Understand the neural network:
 - * ablation.
 - * SVCCA.
 - * layer-wise explainability for side-channel analysis.

References

- 1 Aron Gohr, Gregor Leander, Patrick Neumann: An Assessment of Differential-Neural Distinguishers. *IACR Cryptol. ePrint Arch.* 2022: 1521 (2022).
- 2 Aron Gohr: Improving Attacks on Round-Reduced Speck32/64 Using Deep Learning. *CRYPTO (2) 2019*: 150-179.
- 3 Adrien Benamira, David Gérard, Thomas Peyrin, Quan Quan Tan: A Deeper Look at Machine Learning-Based Cryptanalysis. *EUROCRYPT (1) 2021*: 805-835.
- 4 Trevor Yap, Adrien Benamira, Shivam Bhasin, Thomas Peyrin: Peek into the Black-Box: Interpretable Neural Network using SAT Equations in Side-Channel Analysis. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2023(2), 24–53. <https://doi.org/10.46586/tches.v2023.i2.24-53>.

- 5 Nicoleta-Norica Bacuieti, Lejla Batina, Stjepan Picek: Deep Neural Networks Aiding Cryptanalysis: A Case Study of the Speck Distinguisher. *ACNS 2022*: 809-829.
- 6 Lichao Wu, Yoo-Seung Won, Dirmanto Jap, Guilherme Perin, Shivam Bhasin, Stjepan Picek: Explain Some Noise: Ablation Analysis for Deep Learning-based Physical Side-channel Analysis. *IACR Cryptol. ePrint Arch. 2021*: 717 (2021).
- 7 Guilherme Perin, Lichao Wu, Stjepan Picek: I Know What Your Layers Did: Layer-wise Explainability of Deep Learning Side-channel Analysis. *IACR Cryptol. ePrint Arch. 2022*: 1087 (2022).

4.2 AI for Implementation Attacks

Stjepan Picek (Radboud University, NL)

Nele Mentens (Leiden University, NL)

License  Creative Commons BY 4.0 International license
© Stjepan Picek, Nele Mentens

Implementation attacks aim at the weaknesses of the implementation and not the algorithm. The most common options for implementation attacks are side-channel attacks and fault injection attacks. In both domains, AI is used extensively. In side-channel attacks, it is common to use machine learning in the profiling attack scenario. There, the attacker has a copy of the device to be attacked under control and uses it to build a model of a device. Later, the model is used to attack the target and obtain secret information. Machine learning attacks in such a setup have been used for over a decade and show excellent attack performance. More recent results with deep learning provide even better attack performance against targets protected with countermeasures and with no need to conduct feature engineering [3]. Still, multiple open issues need to be resolved. For instance, the attacks assume that the attacker has access to a copy of a device to be attacked, which is often not a realistic assumption. As such, one of the big challenges to be solved is how to mount non-profiling deep learning-based attacks [2]. Next, leakage assessment is important as it provides the first information on whether the target has secure implementation or if there is some leakage. The results with deep learning are promising but sparse [4]. Mounting an attack once the device is produced is a common setup but results in large expenses for manufacturers once security vulnerabilities are detected. As such, it is important to understand whether we can use various simulation-based approaches and techniques to construct synthetic measurements to assess the security of devices even before they are produced [5, 8]. Finally, as previously discussed, the explainability perspective is important for side-channel attacks. While most of the AI-based approaches for side-channel analysis use machine (deep) learning, there are also some efforts in feature engineering or hyperparameter tuning [1, 7]. More open challenges discussed during the workshop can be found in [6].

On the other hand, in fault injection, AI is mostly used to allow fast characterization of the target (cartography). In that context, various evolutionary and local search algorithms are used [11, 9]. More recently, deep learning is also used to predict if a point on a target will result in a faulty response [10]. We identified the research gaps in making the approaches more stable and maintaining the balance between exploring various regions of the target and fast convergence to a region with many faulty responses. Finally, research rarely explores how to use the located faults in mounting the attacks (which could help understand if all located faults are equally important).

Finally, implementation attacks can be used to attack machine learning, connecting this topic with the security of AI [12, 13].

References

- 1 Karlo Knezevic, Juraj Fulir, Domagoj Jakobovic, Stjepan Picek, Marko Đurasevic: Neuro-SCA: Evolving Activation Functions for Side-Channel Analysis. *IEEE Access* 11: 284-299.
- 2 Lichao Wu, Guilherme Perin, Stjepan Picek: Hiding in Plain Sight: Non-profiling Deep Learning-based Side-channel Analysis with Plaintext/Ciphertext. *IACR Cryptol. ePrint Arch.* 2023: 209 (2023).
- 3 Guilherme Perin, Lichao Wu, Stjepan Picek: Exploring Feature Selection Scenarios for Deep Learning-based Side-channel Analysis. *IACR Trans. Cryptogr. Hardw. Embed. Syst.* 2022(4): 828-861 (2022).
- 4 Thorben Moos, Felix Wegener, Amir Moradi: DL-LA: Deep Learning Leakage Assessment A modern roadmap for SCA evaluations. *IACR Trans. Cryptogr. Hardw. Embed. Syst.* 2021(3): 552-598 (2021).
- 5 Omid Bazangani, Alexandre Iooss, Ileana Buhan, Lejla Batina: ABBY: Automating the creation of fine-grained leakage models. *IACR Cryptol. ePrint Arch.* 2021: 1569 (2021).
- 6 Stjepan Picek, Guilherme Perin, Luca Mariot, Lichao Wu, Lejla Batina: SoK: Deep Learning-based Physical Side-channel Analysis. *ACM Computing Surveys* Volume 55 Issue 11 Article No.: 227pp 1–35.
- 7 Unai Rioja, Lejla Batina, Jose Luis Flores, Igor Armendariz: Auto-tune POIs: Estimation of distribution algorithms for efficient side-channel analysis. *Comput. Networks* 198: 108405 (2021)
- 8 Naila Mukhtar, Lejla Batina, Stjepan Picek, Yinan Kong: Fake It Till You Make It: Data Augmentation Using Generative Adversarial Networks for All the Crypto You Need on Small Devices. *CT-RSA 2022*: 297-321
- 9 Marina Krcek, Thomas Ordas, Daniele Fronte, Stjepan Picek: The More You Know: Improving Laser Fault Injection with Prior Knowledge. *FDTC 2022*: 18-29.
- 10 Lichao Wu, Gerard Ribera, Noemie Beringuier-Boher, Stjepan Picek: A Fast Characterization Method for Semi-invasive Fault Injection Attacks. *CT-RSA 2020*: 146-170.
- 11 Rafael Boix Carpi, Stjepan Picek, Lejla Batina, Federico Menarini, Domagoj Jakobovic, Marin Golub: Glitch It If You Can: Parameter Search Strategies for Successful Fault Injection. *CARDIS 2013*: 236-252.
- 12 Lejla Batina, Shivam Bhasin, Dirmanto Jap, Stjepan Picek: CSI NN: Reverse Engineering of Neural Network Architectures Through Electromagnetic Side Channel. *USENIX Security Symposium 2019*: 515-532.
- 13 Jakub Breier, Xiaolu Hou, Dirmanto Jap, Lei Ma, Shivam Bhasin, Yang Liu: Practical Fault Attack on Deep Neural Networks. *CCS 2018*: 2204-2206.

4.3 Security Fuzzing

Stjepan Picek (Radboud University, NL)

License © Creative Commons BY 4.0 International license
© Stjepan Picek

Vulnerabilities caused by programming errors are a major threat to today's programs. For instance, memory corruption vulnerabilities can lead to uncontrolled behavior in the program, which attackers can often abuse. A modern strategy to uncover such programming errors is automated software testing using fuzz testing (fuzzing). Fuzzing automatically generates inputs from testcases and feeds them to the program under test while monitoring it. If a programming error has been reached, the fuzzer notices that the program hangs or crashes. Mutational fuzzing requires a set of program inputs (seeds) that can be obtained from

testcases or real inputs. The process of mutation can be influenced by 1) the location in the input that gets mutated and 2) the mutation that is applied, with the selection done randomly or guided by a heuristic. A common option is to use evolutionary algorithms for such goals [1]. While the approach works well, there are issues. Due to a wide number of available evolutionary algorithms, selecting what algorithm to use and how to customize it for the task is not trivial. Moreover, since evolutionary algorithms are guided through an objective function, appropriate evaluations should be done. Machine learning is also used in fuzzing for various tasks like seed file generation, testcase generation, or mutation operator selection [2]. It is important to understand whether evolutionary algorithms or machine learning produce better results for tasks that can be achieved by both (e.g., mutation operator selection) and in what scenarios to select a specific AI technique. For instance, finding the states in stateful fuzzing is not easy, and machine learning could be used for this task.

References

- 1 Patrick Jauernig, Domagoj Jakobovic, Stjepan Picek, Emmanuel Stapf, Ahmad-Reza Sadeghi: DARWIN: Survival of the Fittest Fuzzing Mutators. CoRR abs/2210.11783 (2022).
- 2 Wang Y, Jia P, Liu L, Huang C, Liu Z (2020) A systematic review of fuzzing based on machine learning techniques. PLoS ONE 15(8): e0237749. <https://doi.org/10.1371/journal.pone.0237749>.

4.4 Security of Machine Learning

Stjepan Picek (Radboud University, NL)

License  Creative Commons BY 4.0 International license
© Stjepan Picek

Machine (deep) learning found its place in various real-world applications, where many applications have security requirements. Unfortunately, as these systems become more pervasive, understanding how they fail becomes more challenging. There are several failure modes in machine learning, but one category received significant attention in the last few years: backdoor attacks. Backdoor attacks aim to make a model misclassify some of its inputs to a preset-specific label while other classification results behave normally. This misclassification is activated when a specific property is included in the model input. This property is called the trigger and can be anything the targeted model understands. Deep learning is evaluated in either a centralized or distributed setting. While the centralized one is simpler, it poses privacy concerns due to the need to have the training data available (and, for instance, shared in the case of online training). Then, a common option is to use federated learning as a distributed learning paradigm that works on isolated data. In federated learning, clients can collaboratively train a shared global model under the orchestration of a central server while keeping the data decentralized. Multiple backdoor attacks and defenses exist on machine learning systems (centralized and distributed) and for diverse data types: computer vision (e.g., images, video), sound, text, and graph data. While many observations can be transferred from one setup to another, unique characteristics also require detailed experimentalism [1, 2]. We need more systematic evaluations of diverse attack factors in different domains and with larger (more realistic) datasets and neural network models. Finally, more effort must be given to designing powerful, transferable, and efficient defenses [4, 3].

References

- 1 Gorka Abad, Oguzhan Ersoy, Stjepan Picek, Aitor Urbieta: Sneaky Spikes: Uncovering Stealthy Backdoor Attacks in Spiking Neural Networks with Neuromorphic Data. CoRR abs/2302.06279 (2023)
- 2 Jing Xu, Rui Wang, Stefanos Koffas, Kaitai Liang, Stjepan Picek: More is Better (Mostly): On the Backdoor Attacks in Federated Graph Neural Networks. ACSAC 2022: 684-698.
- 3 Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, Farinaz Koushanfar, Ahmad-Reza Sadeghi, Thomas Schneider: FLAME: Taming Backdoors in Federated Learning. USENIX Security Symposium 2022: 1415-1432.
- 4 Kavita Kumari, Phillip Rieger, Hossein Fereidooni, Murtuza Jadliwala, Ahmad-Reza Sadeghi: BayBFed: Bayesian Backdoor Defense for Federated Learning. CoRR abs/2301.09508 (2023).

Participants

- Ileana Buhan
Radboud University
Nijmegen, NL
- Lukasz Chmielewski
Radboud University Nijmegen,
NL & Masaryk University –
Brno, CZ
- Alexandra Dmitrienko
Universität Würzburg, DE
- Elena Dubrova
KTH Royal Institute of
Technology – Kista, SE
- Oguzhan Ersoy
TU Delft, NL
- Hossein Fereidooni
TU Darmstadt, DE
- Fatemeh Ganji
Worcester Polytechnic
Institute, US
- Houman Homayoun
University of California –
Davis, US
- Domagoj Jakobovic
University of Zagreb, HR
- Dirmanto Jap
Nanyang TU – Singapore, SG
- Florian Kerschbaum
University of Waterloo, CA
- Marina Krcek
TU Delft, NL
- Jesus Luna Garcia
Robert Bosch GmbH –
Stuttgart, DE
- Damien Marion
IRISA – Rennes, FR
- Luca Mariot
Radboud University
Nijmegen, NL
- Nele Mentens
Leiden University, NL
- Irina Nicolae
Bosch Center for AI –
Renningen, DE
- Stjepan Picek
Radboud University
Nijmegen, NL
- Jeyavijayan Rajendran
Texas A&M University –
College Station, US
- Ahmad-Reza Sadeghi
TU Darmstadt, DE
- Patrick Schaumont
Worcester Polytechnic
Institute, US
- Matthias Schunter
INTEL ICRI-SC –
Darmstadt, DE
- Mirjana Stojilović
EPFL – Lausanne, CH
- Shahin Tajik
Worcester Polytechnic
Institute, US
- Trevor Yap
Nanyang TU – Singapore, SG



Security of Decentralized Financial Technologies

Arthur Gervais*¹, and Marie Vasek*²

1 Imperial College London, GB. arthur@gervais.cc

2 University College London, GB. m.vasek@ucl.ac.uk

Abstract

The decentralized finance (DeFi) sector has grown to a 13+ billion USD economy, encompassing various financial activities. The non-custodial nature of DeFi requires users to take responsibility for managing their assets, but it also provides them more control over their assets. The Dagstuhl Seminar brought researchers together to examine the security, privacy, and financial properties of DeFi and explore ways to protect users. The seminar aimed to reconcile the conflicting demands of security, usability, and performance in DeFi and outline best practices. Despite progress made in the DeFi sector, there is still much to be explored and improved, such as user education, regulatory compliance, and the scalability and performance limitations of decentralized ledgers. To build a secure and user-friendly DeFi ecosystem, continued collaboration among experts is needed.

Seminar October 16–21, 2022 – <http://www.dagstuhl.de/22421>

2012 ACM Subject Classification Information systems → Digital cash; Security and privacy → Cryptography; Security and privacy → Distributed systems security

Keywords and phrases blockchain technology, decentralized finance (DeFi), distributed consensus protocols, security economics, security foundations

Digital Object Identifier 10.4230/DagRep.12.10.129

1 Summary

Arthur Gervais

Marie Vasek

License © Creative Commons BY 4.0 International license
© Arthur Gervais and Marie Vasek

Trusted intermediaries have been the backbone of economic transactions for centuries. However, with the rise of decentralized ledgers like Bitcoin and Ethereum, individuals now have the opportunity to trade and interact without relying on a centralized authority. In 2020, the decentralized finance (DeFi) sector grew to become a 13+ billion USD economy, encompassing exchanges, borrowing/lending, margin trading, derivatives, and more.

The non-custodial nature of decentralized ledgers gives individuals more control over their assets, but it also requires them to take greater responsibility for managing their private keys and assets. Cryptographers expect DeFi users to have a deep understanding of the security properties and guarantees of the protocols, but in reality, it is challenging to keep users informed about these complexities. Therefore, there is a pressing need for more research to clarify user comprehension of DeFi properties. Additionally, decentralized ledgers face a number of technical limitations, such as scalability issues and potential vulnerabilities to pseudonymous malicious actors.

* Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Security of Decentralized Financial Technologies, *Dagstuhl Reports*, Vol. 12, Issue 10, pp. 129–142

Editors: Arthur Gervais and Marie Vasek



Dagstuhl Reports

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

To address these challenges, the Dagstuhl Seminar brought together researchers with expertise in various subfields of DeFi to jointly examine the security, privacy, and financial properties of decentralized finance. The primary objective of the seminar was to explore how to protect DeFi users. The seminar aimed to reconcile the conflicting demands of security, usability, and performance in DeFi, and outline best practices for users to remain safe while engaging in DeFi activities. Finally, the seminar aimed to apply its recommendations to the growing DeFi ecosystem.

During the seminar, participants presented talks on a wide range of topics, including active attacks on the DeFi ecosystem, proposed cryptographic schemes for enhancing the security of cryptocurrencies, and network insights on cryptocurrencies. The seminar also featured productive discussions across working groups, bringing together researchers from diverse perspectives to achieve the common goal of securing the DeFi landscape.

Given the rapid growth of the DeFi sector, it is important to keep exploring ways to improve its security and user-friendliness. One way to do this is through collaboration and information-sharing among researchers, developers, and users. The Dagstuhl Seminar was an important step in this direction, but there is still much work to be done.

One area of focus could be on improving user education and awareness. This could include developing easy-to-understand guides and tutorials, as well as increasing the transparency of DeFi protocols and the risks associated with using them. Additionally, there is a need for more research into the scalability and performance limitations of decentralized ledgers, as well as finding ways to mitigate security risks such as smart contract vulnerabilities.

Another important aspect to consider is the regulatory landscape for DeFi. Currently, many DeFi protocols operate in a regulatory gray area, and it is important to ensure that they comply with relevant laws and regulations while also protecting user privacy and security. This may require more collaboration between DeFi developers and regulators to establish clear guidelines and standards.

Despite the progress made in the DeFi sector, there are still many unknowns that need to be explored. For example, there is limited understanding of how the Ethereum Proof-of-Stake (PoS) security mechanism works, and what guarantees it provides. This is a crucial aspect of the DeFi landscape as Ethereum is the most widely used blockchain for DeFi applications. Further research is needed to understand the security properties of Ethereum PoS and how it can be improved to better protect users. Additionally, there are other areas in DeFi that require further investigation, such as the scalability and performance limitations of decentralized ledgers, and the trade-offs between privacy and security. By exploring these unknowns, we can gain a better understanding of the DeFi ecosystem and find ways to improve its security and user-friendliness.

In conclusion, the DeFi sector is still in its early stages, and there is much room for growth and improvement. By continuing to bring together experts from various fields and encouraging collaboration, we can help to build a secure and user-friendly DeFi ecosystem that benefits everyone.

2 Table of Contents

Summary

<i>Arthur Gervais and Marie Vasek</i>	129
---	-----

Overview of Talks

Concrete bounds for PoW <i>Rainer Böhme</i>	132
Miner Extractable Value (MEV) and Flash Freezing Flash Boys (F3B) <i>Bryan Ford</i>	132
What can we learn from four years of attacks on Decentralized Finance? <i>Arthur Gervais</i>	133
Ethereum P2P Network Topology <i>Lucianna Kiffer</i>	133
ROAST: Robust Asynchronous Schnorr Threshold Signatures <i>Tim Ruffing</i>	134
State of Signatures in Bitcoin <i>Tim Ruffing</i>	134
Suboptimality in DeFi <i>Aviv Yaish</i>	135

Working groups

Human Aspects of DeFi and Cryptocurrencies <i>Svetlana Abramova and Markus Dürmuth</i>	135
Thwarting Long-Range Attacks with Peacock Mantis Shrimp Checkpoints <i>Sarah Azouvi, George Danezis, Bryan Ford, Philipp Jovanovic, Pedro Moreno-Sanchez, Joachim Neu, and Tim Ruffing</i>	136
Cross-Chain Privacy <i>Jens Ernstberger and Fan Zhang</i>	136
Longest Chain Consensus Under Low Bandwidth <i>Joachim Neu and Lucianna Kiffer</i>	138
Stability in DeFi <i>Aviv Yaish, Alex Biryukov, Rainer Böhme, Arthur Gervais, Lioba Heimbach, Aljosha Judmayer, Ben Livshits, Marie Vasek, and Roger Wattenhofer</i>	140

Participants	142
-------------------------------	-----

3 Overview of Talks

3.1 Concrete bounds for PoW

Rainer Böhme (*Universität Innsbruck, AT*)

License © Creative Commons BY 4.0 International license
© Rainer Böhme

Joint work of Rainer Böhme, Patrik Keller

Main reference Patrik Keller, Rainer Böhme: “Parallel Proof-of-Work with Concrete Bounds”, CoRR, Vol. abs/2204.00034, 2022.

URL <https://doi.org/10.48550/arXiv.2204.00034>

We review the succession of work leading to concrete bounds for the failure probability of Bitcoin’s proof-of-work mechanism in adversarial synchronous networks. While Bitcoin uses proof-of-work sequentially, we propose to study concrete bounds for state replication protocols using non-sequential proof-of-work. Numerical analyses suggest that after the typical interval of 10 minutes, a novel parallel proof-of-work protocols offers two orders of magnitude more security than sequential proof-of-work. This means that state updates could be sufficiently secure to support commits after one block (i.e., after 10 minutes), removing the risk of double-spending in many applications.

3.2 Miner Extractable Value (MEV) and Flash Freezing Flash Boys (F3B)

Bryan Ford (*EPFL Lausanne, CH*)


License © Creative Commons BY 4.0 International license
© Bryan Ford

Joint work of Bryan Ford, Haoqian Zhang, Louis-Henri Merino, Vero Estrada-Galiñanes

Front-running attacks, which benefit from advanced knowledge of pending transactions, have proliferated in the blockchain space since the emergence of decentralized finance. Front-running causes devastating losses to honest participants and continues to endanger the fairness of the ecosystem. We present Flash Freezing Flash Boys (F3B), a blockchain architecture that addresses front-running attacks using threshold cryptography. In F3B, a user generates a symmetric key to encrypt their transaction, and once the underlying consensus layer has committed the transaction, a decentralized secret-management committee reveals this key. F3B mitigates front-running attacks because an adversary can no longer read the content of a transaction before commitment, thus preventing the adversary from benefiting from advanced knowledge of pending transactions. Unlike other threshold-based approaches where the user encrypts their transaction based on the key of a future block, F3B enables the user to generate their key for each transaction. This feature ensures the confidentiality that all uncommitted transactions are not revealed, even if they are delayed. F3B addresses front-running at the execution layer; thus, our solution is agnostic to the underlying consensus algorithm and compatible with existing smart contracts. We evaluated F3B based on Ethereum, demonstrating a 0.05% transaction latency overhead with a secret-management committee of 128 members, indicating our solution is practical at a low cost.

3.3 What can we learn from four years of attacks on Decentralized Finance?

Arthur Gervais (Imperial College London, GB)

License  Creative Commons BY 4.0 International license
© Arthur Gervais

Within just four years, the blockchain-based Decentralized Finance (DeFi) ecosystem has accumulated a peak total value locked (TVL) of \$253 billion USD. Unfortunately, this increase in DeFi's popularity has been accompanied by a number of attacks that have cost at least \$3.24 billion USD between 2018 and 2022. In this talk, we offer a method for measuring, analyzing, and comparing DeFi attacks. By presenting cutting-edge defense strategies that go beyond the conventional smart contract code auditing approaches, we also hope to summarize the insights discovered to strengthen DeFi security.

3.4 Ethereum P2P Network Topology

Lucianna Kiffer (ETH Zürich, CH)

License  Creative Commons BY 4.0 International license
© Lucianna Kiffer

Blockchain protocols' primary security goal is consensus: one version of the global ledger that everyone in the network agrees on. Their proofs of security depend on assumptions on how well their peer-to-peer (P2P) overlay networks operate. Further, the Defi ecosystem built on top of these protocols also explicitly and inexplicably make similar assumptions. Yet, surprisingly, little is understood about what factors influence the P2P network properties. In this talk, I present work where we extensively study the Ethereum P2P network's connectivity and its block propagation mechanism. We gather data on the Ethereum network by running the official Ethereum client, geth, modified to run as a "super peer" with many neighbors. We run this client in North America for over seven months, as well as shorter runs with multiple vantages around the world. Our results expose an incredible amount of churn, and a surprisingly small number of peers who are actually useful (that is, who propagate new blocks). We also find that a node's location has a significant impact on when it hears about blocks, and that the precise behavior of this has changed over time (e.g., nodes in the US have become less likely to hear about new blocks first). Our results motivate questions on how these open systems can be manipulated and whether we should move to more structured/purposeful networks.

3.5 ROAST: Robust Asynchronous Schnorr Threshold Signatures

Tim Ruffing (Blockstream – Victoria, CA)

License  Creative Commons BY 4.0 International license
© Tim Ruffing

Joint work of Tim Ruffing, Viktoria Ronge, Elliott Jin, Jonas Schneider-Bensch, Dominique Schröder
Main reference Tim Ruffing, Viktoria Ronge, Elliott Jin, Jonas Schneider-Bensch, Dominique Schröder: “ROAST: Robust Asynchronous Schnorr Threshold Signatures”, in Proc. of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022, pp. 2551–2564, ACM, 2022.
URL <https://doi.org/10.1145/3548606.3560583>

Bitcoin and other cryptocurrencies have recently introduced support for Schnorr signatures whose cleaner algebraic structure, as compared to ECDSA, allows for simpler and more practical constructions of highly demanded “t-of-n” threshold signatures. However, existing Schnorr threshold signature schemes still fall short of the needs of real-world applications due to their assumption that the network is synchronous and due to their lack of robustness, i.e., the guarantee that honest signers are able to obtain a valid signature even in the presence of other malicious signers who try to disrupt the protocol. This hinders the adoption of threshold signatures in the cryptocurrency ecosystem, e.g., in second-layer protocols built on top of cryptocurrencies.

In this work, we propose ROAST, a simple wrapper that turns a given threshold signature scheme into a scheme with a robust and asynchronous signing protocol, as long as the underlying signing protocol is semi-interactive (i.e., has one preprocessing round and one actual signing round), provides identifiable aborts, and is unforgeable under concurrent signing sessions. When applied to the state-of-the-art Schnorr threshold signature scheme FROST, which fulfills these requirements, we obtain a simple, efficient, and highly practical Schnorr threshold signature scheme.

3.6 State of Signatures in Bitcoin

Tim Ruffing (Blockstream – Victoria, CA)

License  Creative Commons BY 4.0 International license
© Tim Ruffing

Support for Schnorr signatures has been activated in Bitcoin in the as part of “Taproot” softfork. This talk sheds light on the motivation behind this technical change, namely a better provably security as compared to ECDSA, improved efficiency, and most importantly the possibility to construct more practical variants of advanced signature protocols such as multisignatures, threshold signature and blind signatures.

We then give an overview of the state-of-art in these areas, touching upon recent results in the area of Schnorr multisignatures signatures (e.g., MuSig2, FROST, ROAST) as well as blind signatures (e.g., Fuchsbauer, Plouviez and Seurin 2020). We also discuss open research questions in this area, e.g., how multisignatures and threshold signatures can be nested (with a tree-style key setup) while maintaining security under concurrent sessions and privacy, and whether the practicality of distributed key-generation protocols can be improved.

3.7 Suboptimality in DeFi

Aviv Yaish (The Hebrew University of Jerusalem, IL)

License  Creative Commons BY 4.0 International license
© Aviv Yaish

Joint work of Aviv Yaish, Maya Dotan, Kaihua Qin, Aviv Zohar, Arthur Gervais

The Decentralized Finance (DeFi) ecosystem has proven to be immensely popular in facilitating financial operations such as lending and exchanging assets, with Ethereum-based platforms holding a combined amount of more than 30 billion USD. The public availability of these platforms' code together with real-time data on all user interactions and platform liquidity has given rise to sophisticated automatic tools that recognize profit opportunities on behalf of users and seize them.

In this work, we formalize three core DeFi primitives which together are responsible for a daily volume of over 100 million USD in Ethereum-based platforms alone: (1) lending and borrowing funds, (2) liquidation of insolvent loans using swaps, and (3) using *flashswaps* to close arbitrage opportunities between cryptocurrency exchanges. The profit which can be made from each primitive is then cast as an optimization problem that can be solved.


We use our formalization to analyze several case studies for each primitive, showing that popular platforms and tools which promise to automatically optimize profits for users, actually fall short. In specific instances, the profits can be increased by more than 100%, with the highest amount of “missed” revenue by a single suboptimal action equal to 428.14 ETH, or roughly 517K USD.

Finally, we show that many missed opportunities to make a profit do not go unnoticed by other users. Indeed, suboptimal transactions are sometimes immediately followed by “trailing” back-running transactions which extract additional profits using similar actions. By analyzing a subset of these events, we uncover that users who frequently create such trailing transactions are heavily tied to specific miners, meaning that all of their transactions appear only in blocks mined by one miner in particular. As a portion of the backrun non-optimal transactions are private, we hypothesize that the users who create them are, in fact, miners (or users collaborating with miners) who use inside information known only to them to make a profit, thus gaining an unfair advantage.

4 Working groups

4.1 Human Aspects of DeFi and Cryptocurrencies

Svetlana Abramova (Universität Innsbruck, AT), Markus Dürmuth (Leibniz Universität Hannover, DE)

License  Creative Commons BY 4.0 International license
© Svetlana Abramova and Markus Dürmuth

In this discussion group, we considered decentralized finance systems and cryptocurrencies from the point of view of a (human) user. We identified a number of interesting topics that can guide future research, as well as some related challenges.

Trust of users in the system seems crucial for their participation. A tentative list of factors influencing trust may include “fairness” of transaction ordering, which influences who can buy rare goods (such as NFTs). The presence of MEVs, and the fact that currently mostly powerful market players can utilize those, could be adverse for the trust; and obviously

“stability” as discussed in another group. Another topic may be privacy related concerns, and the question with regards to which entities users wish privacy of financial transactions. Governance and decision making in the DeFi & cryptocurrency space is another interesting factor in itself, and may again influence trust in a system. The usability of crypto-wallets provides some very interesting use-case for user authentication, due to their pronounced requirements in availability.

We also identified a number of challenges that need to be overcome to conduct research. Recruitment of users for surveys or user studies is not easy as central methods to directly contact such users are rare, and for services claiming to sample from blockchain users it’s not easy to (non-intrusively) verify those claims. This is additionally complicated by the high heterogeneity of the user-base of cryptocurrencies/DeFi (found by previous studies) and wrong mental models. In many fields, recruitment of decision makers as research subjects (here miners or developers) is even more difficult. It is quite unclear at the moment how a sensible sample of miners could be recruited. This is related to similar problems for sampling decision makers in software design.

4.2 Thwarting Long-Range Attacks with Peacock Mantis Shrimp Checkpoints

Sarah Azouvi (Protocol Labs – Edinburgh, GB), George Danezis (University College London, GB), Bryan Ford (EPFL Lausanne, CH), Philipp Jovanovic (University College London, GB), Pedro Moreno-Sanchez (IMDEA Software Institute – Madrid, ES), Joachim Neu (Stanford University, US), Tim Ruffing (Blockstream – Victoria, CA)

License © Creative Commons BY 4.0 International license
© Sarah Azouvi, George Danezis, Bryan Ford, Philipp Jovanovic, Pedro Moreno-Sanchez, Joachim Neu, and Tim Ruffing

In this work, we propose Peacock Mantis Shrimp, a checkpointing mechanism onto Bitcoin for any PoS consensus scheme. It supports PoS schemes with an arbitrary number of validators, and has an efficient checkpoint verification requiring auditors to download only a small number of Bitcoin full blocks. Peacock Mantis Shrimp achieves this by randomly sampling validators into subgroups which then commit a previously agreed-upon checkpoint onto the Bitcoin blockchain using specially crafted threshold signed transactions. Peacock Mantis Shrimp improves on the state-of-the-art that either suffers from scaling constraints, supporting only a limited number of validators, or requires auditors to examine the full Bitcoin chain. We analyze parametrizations and show the overall failure probability can be driven as low as desired.

4.3 Cross-Chain Privacy

Jens Ernstberger (TU München, DE) and Fan Zhang (Yale University – New Haven, US)

License © Creative Commons BY 4.0 International license
© Jens Ernstberger and Fan Zhang

Cross-Chain Communication received a lot of attention recently due to growing interoperability needs in DeFi and major security flaws in existing protocols that had caused significant financial loss for their users. This session came forth due to a recently published work,

zkBridge [5], that improves the safety of cross-chain communication by replacing trusted committees (a single point failure) with zero-knowledge proofs (ZKP). While improving safety is crucial, the privacy implication of bridges received much less attention. Most deployed systems do not provide privacy guarantees and in particular allow an observer to *link* bridged assets to the original ones. Such linkage could affect the frangibility of assets (since minted coins inherit their history from the source chain) and even erode user privacy (e.g., deanonymization attacks on one chain could now impact other chains through bridges).

This article provides a summary proposed solutions for privacy-preserving, and discusses the opportunities and challenges. The discussion evolved around *(i)* the current solution space for privacy in cross-chain solutions, *(ii)* unique use-cases for private cross-chain communication, *(iii)* potential pitfalls in privacy preserving cross-chain solutions with regard to interoperability of private and public blockchains and *(iv)* alternatives to zkSNARK based solutions for private cross-chain bridges.

Generally, cross-chain exchanges of assets can be facilitated by either atomic swaps or bridges. (Here we leave sidechains (such as [6]) out of scope since we target solutions that can bridge two existing blockchains.) Depending on whether the source/destination blockchain provides native privacy guarantees, such exchanges can happen in one of the following scenarios:

1. public \rightarrow public
2. public \rightarrow private
3. private \rightarrow public
4. private \rightarrow private

Further, we find that the following (rather informal) privacy goals are essential – *(i)* hiding the fact that a swap/bridge of an asset takes place, *(ii)* hiding the amount / type of the involved asset and *(iii)* ensuring unlinkability between participants. We elaborate on privacy-preserving approaches to cross-chain communication that use atomic swaps and bridges, and in which scenarios each of them are applicable as well as sufficiently researched, in the following.

In an atomic swap, Alice intends to exchange X tokens A (native to blockchain A) for Y tokens B (native to blockchain B), such that the asset exchange is included atomically in both blockchains. Simply, this exchange can be achieved with HTLCs. However, HTLCs do not provide privacy, such that recent work proposed adaptor signatures to atomically release secrets whilst assuring privacy [1, 4]. However, we find that applying an atomic swap based on adaptor signatures inherently depends of the confidentiality of the underlying blockchain. Further, we find that such a construction only works if blockchain A is public (i.e. the transaction where the first transaction is included, case 1 + 2). If the sender blockchain is private (e.g., for shielded addresses in Zcash), there is no way to guarantee atomic inclusion in existing constructions. [4] suggests that that their method can be extended to shielded coins with 2PC generation of SNARKs, though details are not specified.

While atomic swaps allows a pair of users to exchange assets, a bridge can enable arbitrary message passing between two chains (thus atomic swaps can be seen as a specific application of a bridge). Typically bridges either depends on *(i)* a committee or *(ii)* a relay network that relays the block header. Alternative approaches also provide cross-chain capabilities through TEEs and MPC [2, 3]. As the current state-of-the-art converges on a construction based on a relay network, we discussed potential extensions to bridges in this domain. As a result, we find that bridges relying on a relay network with an updater contract are inherently incapable of obfuscating the fact that a bridging of assets took place. However, by applying a blockchain mixer on both the source and receiver chain, one can hide the amount transferred

as well as provide unlinkability of sender and receiver accounts. Note, that a single mixer on the source chain is sufficient to ensure unlinkability, whereas a second mixer on the receiving chain can ensure confidentiality of the received amount by the receiver. Note that a private bridge is currently only possible for cross-chain communication between two public chains (Case 1), due to non-existent deployments of privacy-preserving smart contract enabled blockchains (which may change in the future).

In comparison, existing proposals for both private atomic swaps and private bridges face unique limitations that are partially exclusive. We also noted that performing generalized, privacy preserving smart contract function calls, where the invoked function resides on a different chain and the result of the function call needs to be returned to the invoker, can be especially challenging and is an equally unsolved problem, even in a case that involves no privacy. In general, it depends on the use-case at hand, whether one needs to apply a privacy preserving atomic swap or bridge. We deem further investigation of hybrid approaches, that leverage the benefits of both privacy preserving atomic swaps and bridges, an interesting area of future work.

References

- 1 A. Deshpande and M. Herlihy. *Privacy-preserving cross-chain atomic swaps*. In International Conference on Financial Cryptography and Data Security, 2020.
- 2 I. Leontiadis. *Private Blockchain Bridge*. Published in <https://hackmd.io/@EwN07cCvQvy1Tn3mdYW0PQ/rk-r3kZ0q>.
- 3 Y. Lan, J. Gao, Y. Li, K. Wang, Y. Zhu, and Z. Chen. *Trustcross: Enabling confidential interoperability across blockchains using trusted hardware*. In 4th International Conference on Blockchain Technology and Applications, 2021.
- 4 S. Thyagarajan, K. Aravinda, G. Malavolta, and P. Moreno-Sanchez. *Universal atomic swaps: Secure exchange of coins across all blockchains*. In IEEE Symposium on Security and Privacy, 2022.
- 5 T. Xie, J. Zhang, Z. Cheng, F. Zhang, Y. Zhang, Y. Jia, D. Boneh, and D. Song. *zkBridge: Trustless Cross-chain Bridges Made Practical*. Archiv, 2022.
- 6 F. Baldimtsi, I. Miers, and X. Zhang. *Anonymous Sidechains*. In Data Privacy Management, Cryptocurrencies and Blockchain Technology, 2022.

4.4 Longest Chain Consensus Under Low Bandwidth

Joachim Neu (Stanford University, US), Lucianna Kiffer (ETH Zürich, CH)

License © Creative Commons BY 4.0 International license
© Joachim Neu and Lucianna Kiffer

Traditionally, Nakamoto’s longest chain (LC) consensus protocol is analyzed and proven secure in the synchronous adversarial Δ -bounded-delay network model. Specifically, analyses such as [1, 2, 3, 4, 5] exhibit the tradeoff between block production rate λ , adversarial resilience β , and delay upper bound Δ . Thus, these analyses examine ‘how much honest mining rate is lost’ because honest nodes mine on ‘old’ chains because they have not yet heard of the most recent chains due to the Δ delay.

However, the Δ -bounded-delay network model assumes that consensus messages travel between honest nodes with at most Δ delay, *irrespective of network load*. Thus, the model neglects important aspects of real communication networks such as congestion and queuing delays caused by limited bandwidth. Consequently, the prior analyses do not capture “how much honest mining rate is lost” because honest nodes mine on “old” chains while they are busy downloading more recent chains.

Earlier work [6] provides a network model that captures the fact that every node has limited bandwidth of C block content downloads per time, and analyzes proof-of-stake (PoS) Nakamoto consensus in that setting. For PoS LC, an analysis capturing a bandwidth constraint was particularly interesting, because in PoS the adversary can produce an infinite number of “valid” blocks for each block production opportunity, then spam the network with these equivocating blocks, and thus induce congestion in an attempt to attack the protocol. Though proof-of-work (PoW) LC naturally throttles the spamming ability of the adversary through the necessity of producing valid “work”, the problem of congestion and block download delay remains relevant. In particular, we observe this when bandwidth is low (i.e., when target consensus throughput is close to the bandwidth limit).

Unfortunately, the analysis of [6] is rather pessimistic, in the sense that it analyses the worst-case amount of outstanding block downloads and provisions sufficiently high bandwidth C to always be able to complete outstanding downloads promptly. Consequently, the provisioned bandwidth C is asymptotically higher than average-case block download requirement based on the blockchain’s throughput. As a result, the basic LC variant of [6] requires vanishing throughput for security, a situation that [6] only improves upon by proposing a parallel composition of multiple instances of the basic LC variant.


In contrast, in our group work we aimed to improve upon [6] and to show that both PoW and PoS Nakamoto consensus can be made secure for low (i.e., constant) bandwidth and thus for non-vanishing throughput. In the PoW setting, the global limit on block production rate provided by PoW can be used to strengthen the analysis of [6]. For PoS LC, further changes to the protocol are necessary to ensure that per block production opportunity, honest nodes need to download at most one block content. Specifically, in a new protocol, honest nodes could use the consensus protocol to agree on “proofs of equivocation” to consistently blank out the contents of equivocating blocks from the block tree and thus obviate the need to download multiple equivocating blocks. Thus, as compared to [6], a new protocol should maintain the structure of Nakamoto consensus, while providing security under non-vanishing throughput proportional to the bandwidth constraint.

References

- 1 R. Pass, L. Seeman, and a. shelat. *Analysis of the blockchain protocol in asynchronous networks*. In Annual International Conference on the Theory and Applications of Cryptographic Techniques, 2017.
- 2 R. Pass and E. Shi. *The sleepy model of consensus*. In International Conference on the Theory and Application of Cryptology and Information Security, 2017.
- 3 L. Kiffer, R. Rajaraman, and a. shelat. *A better method to analyze blockchain consistency*. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 2018.
- 4 P. Gaži, A. Kiayias, and A. Russell. *Tight consistency bounds for bitcoin*. In Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, 2020.
- 5 A. Dembo, S. Kannan, E. Tas, D. Tse, P. Viswanath, X. Wang, and O. Zeitouni. *Everything is a race and nakamoto always wins*. In Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, 2020.
- 6 J. Neu, S. Sridhar, L. Yang, D. Tse, and M. Alizadeh. *Securing Proof-of-Stake Nakamoto Consensus Under Bandwidth Constraint*. In AFT ’22: 4th ACM Conference on Advances in Financial Technologies, 2022.

4.5 Stability in DeFi

Aviv Yaish (The Hebrew University of Jerusalem, IL), Alex Biryukov (University of Luxembourg, LU), Rainer Böhme (Universität Innsbruck, AT), Arthur Gervais (Imperial College London, GB), Lioba Heimbach (ETH Zürich, CH), Aljosha Judmayer (Universität Wien & SBA Research – Wien), Ben Livshits (Imperial College London, GB), Marie Vasek (University College London, GB), Roger Wattenhofer (ETH Zürich, CH)

License  Creative Commons BY 4.0 International license
 © Aviv Yaish, Alex Biryukov, Rainer Böhme, Arthur Gervais, Lioba Heimbach, Aljosha Judmayer, Ben Livshits, Marie Vasek, and Roger Wattenhofer

Notions of Stability and Risk

Stability in traditional financial systems is not well-defined, though a common definition used by the European Central Bank [1] and the World Bank [2] treats economic stability as the ability of an economic ecosystem to sustain shocks while still continuing to function and providing financial services as usual.

This definition could be imported to the *Decentralized Finance (DeFi)* ecosystem. Although such definitions are still relevant, they ignore certain inherent properties of DeFi protocols and the underlying blockchain infrastructure. Certain platforms, such as *Constant Product Automated Market Makers (CPAMMs)* and utilization-based lending pools can continue functioning in times of duress, though in an unstable manner [3] where market prices violently oscillate in short periods of time [4]. Although users are traditionally viewed as promoting a more efficient market, their self-interested actions actually might cause price instability [5].

Types of DeFi Risk

From a technical viewpoint, there are three types of high level DeFi risks:

1. Ones that lead to the collapse or instability of one DeFi ecosystem or token.
2. Instability which starts with one DeFi ecosystem and propagates to another.
3. The risk that a collapse or instability propagates to the traditional financial system [7].

How Can DeFi Risks and Stability be Measured?

An encompassing definition of stability, although perhaps slightly imprecise, would be the ecosystem’s proximity to a price equilibrium for all assets and financial services contained within the ecosystem [6].

When quantifying stability in the context of DeFi, due to the wildly varying mechanisms involved [11], one can take a per-platform approach. A potential avenue to explore is the requirements to trigger a “bank run” on lending protocols and how close these are to such a collapse. If cryptocurrency prices were to fall quickly and liquidations could no longer execute in time, lending pools that could no longer meet contractual obligations to repay lenders would face a liquidity crisis [12].

Burning the World Down: Destabilizing DeFi

Besides exploring and identifying already existing stability risks in DeFi protocols, another question is how can one design attacks on DeFi or use DeFi to execute attacks that amplify already existing problems such that suboptimality and inefficiencies are more likely to lead to instability.

Such attacks could be performed using a mixture of technical and financial means, for example Distributed Denial-of-Service (DDoS) attacks at the network level [10], or preventing oracle price updates by creating congestion [9].

Another attack strategy would be to synchronize the actions of multiple entities by technical means. Crowdfunded attacks could be executed directly on the consensus layer, as suggested in [8]. But also new attacks which utilize Blockscan messaging and smart contracts to manipulate interest rates, as demonstrated by [3] can be envisioned.

References

- 1 European Central Bank. *Financial stability and macroprudential policy*. In <https://www.ecb.europa.eu/ecb/tasks/stability/html/index.en.html>, last accessed 2022.
- 2 World Bank. *Financial stability*. In <https://www.worldbank.org/en/publication/gfdr/gfdr-2016/background/financial-stability>, last accessed 2022.
- 3 A. Yaish, M. Dotan, K. Qin, A. Zohar, and A. Gervais. *Suboptimality in DeFi*, 2022.
- 4 M. Friedman. *A monetary and fiscal framework for economic stability*. In *Essential Readings in Economics*, Springer: 345-365, 1995.
- 5 N. Kaldor. *Speculation and economic stability*. In *The Review of Economic Studies*, Wiley-Blackwell, 7:1-27, 1939.
- 6 P. Samuelson. *Spatial price equilibrium and linear programming*. In *The American Economic Review*, JSTOR, 42:293-303. 1952.
- 7 S. Aramonte, W. Huang, and A. Schrimpf. *DeFi risks and the decentralisation illusion*. 2021.
- 8 A. Judmayer, N. Stifter, A. Zamyatin, I. Tsabary, I. Eyal, P. Gazi, S. Meiklejohn, and E. Weippl. *Pay to win: Cheap, crowdfundable, cross-chain algorithmic incentive manipulation attacks on PoW cryptocurrencies*. In *Cryptology ePrint Archive*, 2019.
- 9 B.Liu, P. Szalachowski, and J. Zhou. *A first look into defi oracles*. In *IEEE International Conference on Decentralized Applications and Infrastructures*, 2021.
- 10 R. Chaganti, R. Boppana, V. Ravi, K. Munir, M. Almutairi, F. Rustam, E. Lee, and I. Ashraf. *A Comprehensive Review of Denial of Service Attacks in Blockchain Ecosystem and Open Challenges*. In *IEEE Access*, 10:96538-96555, 2022.
- 11 F. Schär. *Decentralized Finance: On Blockchain-and Smart Contract-based Financial Markets*. In *SSRN 3571335*, 2021.
- 12 L. Gudgeon, D. Perez, D. Harz, B. Livshits, and A. Gervais. *The Decentralized Financial Crisis*. In *Crypto Valley Conference on Blockchain Technology*, 2020.

Participants

- Svetlana Abramova
Universität Innsbruck, AT
- Sarah Azouvi
Protocol Labs – Edinburgh, GB
- Alex Biryukov
University of Luxembourg, LU
- Rainer Böhme
Universität Innsbruck, AT
- Stefanos Chaliasos
Veridise – London, GB
- George Danezis
University College London, GB
- Markus Dürmuth
Leibniz Universität
Hannover, DE
- Jens Ernstberger
TU München, DE
- Bryan Ford
EPFL Lausanne, CH
- Arthur Gervais
Imperial College London, GB
- Lioba Heimbach
ETH Zürich, CH
- Philipp Jovanovic
University College London, GB
- Aljosha Judmayer
Universität Wien & SBA
Research – Wien
- Ghassan Karame
Ruhr-Universität Bochum, DE
- Lucianna Kiffer
ETH Zürich, CH
- Ben Livshits
Imperial College London, GB
- Pedro Moreno-Sanchez
IMDEA Software Institute –
Madrid, ES
- Joachim Neu
Stanford University, US
- Tim Ruffing
Blockstream – Victoria, CA
- Florian Tschorsch
TU Berlin, DE
- Marie Vasek
University College London, GB
- Roger Wattenhofer
ETH Zürich, CH
- Aviv Yaish
The Hebrew University of
Jerusalem, IL
- Fan Zhang
Yale University – New Haven, US
- Liyi Zhou
Chainlink Labs – London, GB
- Aviv Zohar
The Hebrew University of
Jerusalem, IL



Developmental Machine Learning: From Human Learning to Machines and Back

James M. Rehg^{*1}, Pierre-Yves Oudeyer^{*2}, Linda B. Smith^{*3},
Sho Tsuji^{*4}, Stefan Stojanov^{†5}, and Ngoc Anh Thai^{†6}

- 1 Georgia Institute of Technology – Atlanta, US. rehg@gatech.edu
- 2 INRIA – Bordeaux, FR. pierre-yves.oudeyer@inria.fr
- 3 Indiana University – Bloomington, US. smith4@indiana.edu
- 4 University of Tokyo, JP. tsujish@gmail.com
- 5 Georgia Institute of Technology – Atlanta, US. sstojanov@gatech.edu
- 6 Georgia Institute of Technology – Atlanta, US. athai6@gatech.edu

Abstract

This interdisciplinary seminar brought together 18 academic and industry computer science researchers in artificial intelligence, computer vision and machine learning with 19 researchers from developmental psychology, neuroscience and linguistics. The objective was to catalyze connections between these communities, through discussions on both how the use of developmental insights can spur advances in machine learning, and how computational models and data-driven learning can lead to novel tools and insights for studying child development. The seminar consisted of tutorials, working groups, and a series of talks and discussion sessions. The main outcomes of this seminar were 1) The founding of DevelopmentalAI (<http://www.developmentalai.com>), an online research community to serve as a venue for communication and collaboration between developmental and machine learning researchers, as well as a place collect and organize relevant research papers and talks; 2) Working group outputs – summaries of in-depth discussions on research questions at the intersection of developmental and machine learning, including the role of information bottlenecks and multimodality, as well as proposals for novel developmentally motivated benchmarks.

Seminar October 16–21, 2022 – <http://www.dagstuhl.de/22422>

2012 ACM Subject Classification Computing methodologies → Artificial intelligence

Keywords and phrases developmental psychology, human learning, machine learning, computer vision, language learning

Digital Object Identifier 10.4230/DagRep.12.10.143


1 Executive Summary

Pierre-Yves Oudeyer

James M. Rehg

Linda B. Smith

Sho Tsuji

License  Creative Commons BY 4.0 International license
© Pierre-Yves Oudeyer, James M. Rehg, Linda B. Smith, Sho Tsuji

Recent advances in artificial intelligence, enabled by large-scale datasets and simulation environments, have resulted in breakthrough improvements in areas like object and speech recognition, 3D navigation, and machine translation. In spite of these advances, modern

* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Developmental Machine Learning: From Human Learning to Machines and Back, *Dagstuhl Reports*, Vol. 12, Issue 10, pp. 143–165

Editors: Pierre-Yves Oudeyer, James M. Rehg, Linda B. Smith, and Sho Tsuji



DAGSTUHL
REPORTS Dagstuhl Reports

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

artificial learning systems still pale in comparison to the competencies of young human learners. The differences between human learning and the paradigms that currently guide machine learning are striking. For example, children actively identify both the concepts to be learned and the data items used for learning, they infer the labels for learning from ambiguous perceptual data, and they learn from continuous streams of percepts without storing and curating large datasets. Artificial intelligence researchers are increasingly looking to developmental science for ideas and inspiration to improve machine learning, while developmental scientists are adopting tools from data science and machine learning to analyze large datasets and gain insights into developmental processes.

This seminar created new connections between the developmental and machine learning research communities by bringing together researchers in linguistics, psychology, cognitive science and neuroscience with investigators working in computer vision, machine learning and robotics. The seminar focused on three research questions:

1. What are the key computational problems and challenges that need to be addressed in creating a developmentally-inspired machine learner? Existing machine learning methods are built on a set of canonical problem formulations such as supervised learning and reinforcement learning. At the same time, decades of research in developmental science have produced an increasingly detailed characterization of learning in children. How can we leverage these insights to create new and more powerful machine learners and revise standard ML problem formulations?
2. What criteria are necessary for agent-based simulation models of development to advance machine learning and provide useful tests of developmental hypotheses? Advances in computer graphics and physics simulation have made it possible to create synthetic environments for training reinforcement learning agents to perform developmentally-relevant cognitive tasks such as navigating 3D space and manipulating objects. Can such computational experiments serve as useful tests of developmental hypotheses?
3. How can data-driven computational models be used to advance developmental science? It is increasingly feasible to collect dense sensor data that captures the perceptual inputs children receive (e.g. via wearable cameras and eye trackers), their behaviors during naturalistic interactions, and a variety of contextual variables relevant to cognitive tasks. These rich datasets, in conjunction with advances in deep learning have created the opportunity to create machine learning models which can “solve” certain developmental tasks such as object recognition. Given that such deep models do not speak directly to mechanisms of human learning, how can such research advance developmental science?

Through a seminar program consisting of tutorials, talks, working group meetings, and an early career mentorship sessions, we gained interdisciplinary insights into these core research questions. Attendees discussed the potential research directions that different research disciplines can benefit from each other, as well as collaboration opportunities and future development of the community. As the initial step, we aim to connect interested researchers online through social media and provide a common repository for relevant literature.

2 Table of Contents

Executive Summary

Pierre-Yves Oudeyer, James M. Rehg, Linda B. Smith, Sho Tsuji 143

Overview of Talks

Studying visual object learning with egocentric computer vision <i>David J. Crandall</i>	148
Using machine learning in early language acquisition research: Examples from long-form audio-recordings <i>Alejandrina Cristia</i>	148
Can Machine Learning Inform the Science of Infant Development, and Vice-Versa? <i>Rhodri Cusack</i>	149
Towards embracing complexity to understand atypical development: the case of Down syndrome <i>Hana D'Souza</i>	149
Simulating early language acquisition using self-supervised learning <i>Emmanuel Dupoux</i>	149
“ML as a tool” vs. “ML as a model” for the study of child development in the wild <i>Abdellah Fourtassi</i>	150
Predictive models of early language learning <i>Michael C. Frank</i>	150
Visual affordances from video: learning to interact by watching people <i>Kristen Grauman</i>	150
The First 1,000 Days Project <i>Uri Hasson, Casey Lew-Williams</i>	151
How language can help machines to acquire general intelligence? <i>Felix Hill</i>	151
The Impact of Dataset Bias on Model Learning <i>Judy Hoffman</i>	152
Truth, lies, and misinformation during cognitive development <i>Celeste Kidd</i>	152
Enhancement of cues and the oddball effect in child-directed speech <i>Eon-Suk Ko</i>	152
Studying infant-like visual category generalization using the Toybox dataset <i>Maithilee Kunda</i>	153
Learning Vision for Walking <i>Jitendra Malik</i>	153
Does Affective communication increase the relation between children with ASD and their mothers? <i>Atsushi Nakazawa</i>	154
Language and Culture Internalization for Autotelic Human-Like AI <i>Pierre-Yves Oudeyer</i>	154

The Never-Ending Visual classification Stream (NEVIS) 1.0 <i>Marc'Aurelio Ranzato</i>	155
Connecting 3D Shape Learning and Object Categorization <i>James M. Rehg</i>	155
Human infants' brains are specialized for social functions <i>Rebecca Saxe</i>	155
Towards Teachable Autonomous Agents: How can developmental psychology help? <i>Olivier Sigaud</i>	156
Why self-generated behavior has more radical consequences than you might originally think <i>Linda B. Smith</i>	156
Rethinking the developmental pathway of early infant language learning <i>Daniel Swingley</i>	156
SCALa: A blueprint for computational models of language acquisition in social context <i>Sho Tsuji</i>	157
Visual attention development in infancy <i>Ingmar Visser</i>	157
Temporal patterns in vocal even sequences produced by human infants and computational vocal learning models <i>Anne Warlaumont</i>	158
Curiosity in infants and computational models <i>Gert Westermann</i>	158
Magnifying Time and Space: New Ways of Studying Early Development and Learning from the Infant's Point of View <i>Chen Yu</i>	159
Audio-visual self-supervised learning <i>Andrew Zisserman</i>	159

Working Groups

Group 1.1: The power of informational processing bottlenecks <i>Rhodri Cusack, Uri Hasson, Celeste Kidd, Marc'Aurelio Ranzato, Stefan Stojanov, Anh Thai</i>	160
Group 1.2: Developmental AI Benchmark <i>Emmanuel Dupoux, James M. Rehg, Daniel Swingley, Anne Warlaumont, Gert Westermann, Chen Yu</i>	161
Group 2.1: Embodied Intention Prediction Challenge <i>Michael Frank, Naithilee Kunda, Marvin Lavechin, Pierre-Yves Oudeyer, Rebecca Saxe, Maureen de Seyssel, Ingmar Visser</i>	162
Group 2.2: ML for Causal Theories of Child Development? <i>David Crandall, Alejandrina Cristia, Hana D'Souza, Abdellah Fourtassi, Clément Romac, Olivier Sigaud</i>	163

Group 3: Multimodality in babies and machines
*Thomas Carta, Hiromichi Hagihara, Felix Hill, Judy Hoffman, Eon-Suk Ko, Casey
Lew-Williams, Atsushi Nakazawa, Jelena Suvevic* 163

Participants 165

Remote Participants 165

3 Overview of Talks

3.1 Studying visual object learning with egocentric computer vision

David J. Crandall (Indiana University – Bloomington, US)

License  Creative Commons BY 4.0 International license
© David J. Crandall

While early work in computer vision was inspired by studies of human perception, most recent work has focused on techniques that work well in practice but probably have little biological basis. But low-cost, lightweight wearable cameras and gaze trackers can now record people’s actual fields of view as they go about their everyday lives. Such first-person, “egocentric” video contains rich information about how people see and interact with the world around them, potentially helping us better understand human perception and behavior while also yielding insights that could improve computer vision. I’ll describe a recent interdisciplinary project (with Chen Yu and Linda Smith) in which we used computer vision to try to characterize the properties of childrens’ egocentric views as they interact with objects – the “training data” of the child’s learning system – and then showed that injecting similar properties into the training data of computer vision algorithms could improve the algorithms’ accuracies as well.

3.2 Using machine learning in early language acquisition research: Examples from long-form audio-recordings

Alejandrina Cristia (LSCP – Paris, FR)

License  Creative Commons BY 4.0 International license
© Alejandrina Cristia

In 2022, we may not have hoverboards, but we have seen artificial intelligences beat humans at go, write in the style of Shakespeare, and generate novel continuations to incomplete spoken sentences. Feats like these have, in part, been due to the rise of self-supervision machine learning techniques, in which systems are trained with vast amounts of unlabeled data. In this talk, I argue that such techniques are useful to infant researchers working in under-described languages and cultures in two key ways: First, to create classifiers that describe and annotate the vast amounts of infant-centered data we can now easily collect; and second, to build systems that potentially learn like infants do. I draw from recent work using audio recordings collected with wearables to illustrate these two avenues of work in the description of children’s spoken language environment, while highlighting both opportunities and challenges, including saliently ethical and legal ones.

3.3 Can Machine Learning Inform the Science of Infant Development, and Vice-Versa?

Rhodri Cusack (Trinity College Dublin, IE)

License  Creative Commons BY 4.0 International license
© Rhodri Cusack

It can be difficult for psychologists and neuroscientists to conceptualise how cognitive functions emerge in the infant brain. Computational models may help, by providing a way to quantify the statistics of the environment, and to test the efficacy of proposed learning objectives and inductive biases. I will describe our ongoing work using deep neural networks to model the development of diverse aspects of the visual system, and the neuroimaging experiments we are running to evaluate them. Finally, I will discuss the potential for translation in the opposite direction, by reviewing how what we have learned about the development of infant cognition might inform the next generation of unsupervised machine learning.

3.4 Towards embracing complexity to understand atypical development: the case of Down syndrome

Hana D'Souza (Cardiff University, GB)

License  Creative Commons BY 4.0 International license
© Hana D'Souza

Development is a complex process, involving interactions between various domains across levels of description. Yet, many of our traditional developmental paradigms aim to isolate domains. The domains measured from various tasks are then correlated in order to understand how they are connected. However, everyday experiences emerge through the complex interactions of various domains – such as motor ability, attention allocation, and the actions of other social agents. Thus, in order to understand typical and atypical development, it is crucial to embrace complexity by putting these interactions at the very core of our research. Findings from studies using this approach have been challenging fundamental assumptions about typical development. I will introduce some of our initial steps in applying this approach to atypical development (Down syndrome) and explain why it has the potential to reconceptualise our understanding of neurodevelopmental disorders.

3.5 Simulating early language acquisition using self-supervised learning


Emmanuel Dupoux (LSCP – Paris, FR)

License  Creative Commons BY 4.0 International license
© Emmanuel Dupoux

Recent progress in self supervised learning opens up the way to learn probabilistic models of language from raw audio signals. We propose to use these models as a proof of feasibility of the 'statistical learning hypothesis', which states that infants bootstrap into language primarily by extracting regularities from the speech input. Similarities and differences between the developmental curves in the models and infants are presented and discussed.

3.6 “ML as a tool” vs. “ML as a model” for the study of child development in the wild


Abdellah Fourtassi (Aix-Marseille University, FR)

License  Creative Commons BY 4.0 International license
© Abdellah Fourtassi

Recent improvements in Machine Learning (ML) promise to transform research in developmental psychology by allowing the quantitative study of children’s behavior outside the lab. ML can help achieve this goal in two steps: 1) automatic annotation of a target behavior from naturalistic data, and 2) quantitative prediction of this behavior from complex (possibly causal) factors. These two steps are obviously related but they diverge in the nature of the ML they call upon. In the first, ML is a “tool” whose purpose is to overcome the limitations of manual labor. In the second, ML is considered a “model” whose purpose is to mimic the child’s behavior given a similarly rich input/stimuli. In this brief talk, I will illustrate – based on ongoing research in our team about children’s early conversational development – how “ML as a tool” and “ML as a model” can be articulated to help build quantitative theories of child development in the wild.

3.7 Predictive models of early language learning

Michael C. Frank (Stanford University, US)

License  Creative Commons BY 4.0 International license
© Michael C. Frank

How can we create mechanistic models of children’s early language learning? One key problem is the availability of data to train and evaluate such models. I’ll present our approach to combining data from large numbers of children – inputs from CHILDES, outcomes from Wordbank – to model early vocabulary acquisition across languages. A simple regression approach allows us to combine both descriptive and model-based predictors, holding the promise of more integrative, data driven theories.

3.8 Visual affordances from video: learning to interact by watching people

Kristen Grauman (University of Texas – Austin, US)

License  Creative Commons BY 4.0 International license
© Kristen Grauman

First-person or “egocentric” vision requires understanding the video that streams to a wearable camera. It offers a special window into the camera wearer’s attention, goals, and interactions, making it an exciting avenue for perception in augmented reality and robot learning. I will present our recent work using passive observations of human activity to inform active robot behaviors – such as learning object affordances from video to shape dexterous robot manipulation, transforming video into a human-centric topological map of the physical space and the activities it supports, or discovering compatible objects to shortcut visual semantic planning. We show how reinforcement learning agents that prefer

human-like interactions can successfully accelerate their task learning and generalization. Finally, I will overview Ego4D, a massive new egocentric video dataset and benchmark built by a multi-institution collaboration that offers a glimpse of daily life activity around the world.

3.9 The First 1,000 Days Project

Uri Hasson, Casey Lew-Williams (Princeton University, US)

License © Creative Commons BY 4.0 International license
© Uri Hasson, Casey Lew-Williams

How do natural, everyday statistics in infants' environments give rise to learning? We will introduce a big-data project, the First 1,000 Days Project at Princeton University, inspired by prior video corpora, including the Human Speechome Project and the SAYCam corpus. Our dataset is designed to video-record 20 families for 1,000 days, beginning when the family returns home after birth. Each house is wired with eight cameras and four microphones that will record for 12 hours per day. Our team is deploying (and developing) machine learning tools for automated analysis of objects, people, space, proximity, and language, including a 'baby detector' and a pipeline that can analyze 300+ years of raw video and audio data. We have completed the development and automation of the research pipeline, and data collection has started with eight families in New Jersey and eastern Pennsylvania, with five additional families waiting to start once their babies are born. Our goal is to recruit a final sample of 20 families that represents the diversity of U.S. demographics.

3.10 How language can help machines to acquire general intelligence?


Felix Hill (Google DeepMind – London, GB)

License © Creative Commons BY 4.0 International license
© Felix Hill

Having and using language makes humans as a species better learners and better able to solve hard problems. I'll present three results that demonstrate how this can also be the case for artificial models of general intelligence. First, I'll show that agents with access to visual and linguistic semantic knowledge explore their environment more effectively than non-linguistic agents, enabling them to learn more about the world around them. Second, I'll demonstrate how an agent embodied in a simulated 3D world can be enhanced by learning from explanations – answers to the question “why?” expressed in language. Agents that learn from both classical reinforcement and explanations solve harder cognitive challenges than those trained from RL alone. Finally, I'll present evidence that the skewed and bursty distribution of natural language may explain how large language models can be prompted to rapidly acquire new skills or behaviours. This suggests how modelling language can make a neural network better able to acquire new cognitive capacities quickly, even when those capacities are not necessarily explicitly linguistic.

3.11 The Impact of Dataset Bias on Model Learning


Judy Hoffman (Georgia Institute of Technology – Atlanta, US)

License  Creative Commons BY 4.0 International license
© Judy Hoffman

Computer vision relies on learning from collections of data. The mechanisms used for collecting, curating, and annotating visual data results in datasets with distinct forms of bias. In turn, models that are trained using biased data, then perpetuate that bias into their learned representations. As the world changes the particular visual appearance bias of the initial data collection may not well represent the appearances the model is expected to operate on. This discrepancy leads to reduced performance and reliability of the learned model. In strong contrast, people are able to experience a biased sample of the world yet generalize (under certain conditions) to alternative world views, like a child who can recognize an elephant at the zoo after being shown cartoon drawings of an elephant. This talk will discuss two key challenges towards producing generalizable visual learning: 1) how can we leverage the learning process to help us identify bias in our data and 2) how can we mitigate bias through modified learning protocols or by adapting to new observations as they appear?

3.12 Truth, lies, and misinformation during cognitive development

Celeste Kidd (University of California – Berkeley, US)

License  Creative Commons BY 4.0 International license
© Celeste Kidd

I will talk about our lab's current work-in-progress exploring interventions designed to give children a greater ability to discern truth from falsity. I will discuss some of the foundational empirical studies in progress on two types of interventions designed to facilitate children's ability to discern fact from fiction. The first set of interventions target factors external to the child relating to the information ecosystems in which they are making judgements. The second set of interventions involve investigating internal mechanisms children may have available for helping them detect misinformed opinions. Both sets of work build off the lab's previous behavioral experiments and computational models about how children sample subsets of information from the world based on their uncertainty in order to form their beliefs and guide their subsequent sampling decisions. I will briefly provide some background on how our new work is building off of our prior papers.

3.13 Enhancement of cues and the oddball effect in child-directed speech

Eon-Suk Ko (Chosun University, KR)


License  Creative Commons BY 4.0 International license
© Eon-Suk Ko

People adapt their way of speaking when addressing children, and this speech register called Child-Directed Speech (CDS) is considered to provide features beneficial for infants' language learning. I present some of these features based on Korean mothers' interaction with their

children. I then raise the question about the mechanism of how such features might benefit infants' learning given their small proportions provided in the input. I suggest that infants' novelty-driven learning and the oddball effect might help us understand aspects of such a mechanism.

3.14 Studying infant-like visual category generalization using the Toybox dataset

Maithilee Kunda (Vanderbilt University, US)

License  Creative Commons BY 4.0 International license
© Maithilee Kunda

Infants can generalize from a small number of object instances within a category to novel instances. For computer vision, this problem can be posed as a domain adaptation problem, i.e., where the distribution of data in the training dataset differs from the distribution seen at test time. However, current domain adaptation tasks and datasets do not target learning across this particular type of distribution shift. We have used our lab's Toybox dataset of handheld object manipulation videos to create a new task that mimics this learning scenario, and I will present initial work on examining how existing domain adaptation models perform on this challenging new task. I will also briefly describe two other projects that investigate how agents might learn spatial reasoning skills and theory of mind reasoning skills.

3.15 Learning Vision for Walking


Jitendra Malik (University of California – Berkeley, US)

License  Creative Commons BY 4.0 International license
© Jitendra Malik

As a child interacts with the world around her, there is a barrage of sensory information – proprioception, tactile, audition, vision – together with knowledge of her own commanded actions via the efference copy. In AI and robotics, the cross-modal supervision that this would enable has been quite under-exploited. In recent work, <https://arxiv.org/abs/2211.03785> (to appear at ICRA 2023), we showed a concrete example of how this might work by learning a visual walking policy for a quadruped legged robot. We train a visual module in the real world to predict the upcoming terrain with our proposed algorithm Cross-Modal Supervision (CMS). CMS uses time-shifted proprioception to supervise vision and allows the policy to continually improve with more real-world experience. We evaluate our vision-based walking policy over a diverse set of terrains including stairs (up to 19cm high), slippery slopes (inclination of 35 degrees), curbs and tall steps (up to 20cm), and complex discrete terrains. We achieve this performance with less than 30 minutes of real-world data. Finally, we show that our policy can adapt to shifts in the visual field with a limited amount of real-world experience.

3.16 Does Affective communication increase the relation between children with ASD and their mothers?

Atsushi Nakazawa (Kyoto University, JP)

License  Creative Commons BY 4.0 International license
© Atsushi Nakazawa

Affective communication has the function of facilitating smooth communication. Our group have been studying a French-originated affective communication method “Humanitude” which was originally developed for the nursing of dementia care. The Humanitude consists of face-to-face communication (eye contact and facial expressions), touching, and talking, but there have been no studies quantifying the elements. Using computational behavioral science methods, our group have detected and analyzed the skill elements including eye contact, face-to-face communication using image recognition from first and third person video, developed and used the state-of-the-arts whole-body tactile sensor for the touch communication analysis, and developed a novel mobile facial myoelectric for facial expression recognition. As the result, our group revealed the skill elements of the methodology. Moreover, we developed the training system of the Humanitude using Augmented Reality (AR) technology which outperformed the existing communication trainings method. We will also introduce our efforts to apply this technique to improve parent-child relationships in ASD. While the experiment is preliminary, their eye contact and physical communication significantly increased after the intervention.

3.17 Language and Culture Internalization for Autotelic Human-Like AI

Pierre-Yves Oudeyer (INRIA – Bordeaux, FR)

License  Creative Commons BY 4.0 International license
© Pierre-Yves Oudeyer

Building autonomous artificial agents able to grow open-ended repertoires of skills is one of the fundamental goals of AI. To that end, a promising developmental approach recommends the design of intrinsically motivated agents that learn new skills by generating and pursuing their own goals – autotelic agents. However, existing algorithms still show serious limitations in terms of goal diversity, exploration, generalization or skill composition. This perspective calls for the immersion of autotelic agents into rich socio-cultural worlds. We focus on language especially, and how its structure and content may support the development of new cognitive functions in artificial agents, just like it does in humans. Indeed, most of our skills could not be learned in isolation. Formal education teaches us to reason systematically, books teach us history, and YouTube might teach us how to cook. Crucially, our values, traditions, norms and most of our goals are cultural in essence. This knowledge, and some argue, some of our cognitive functions such as abstraction, compositional imagination or relational thinking, are formed through linguistic and cultural interactions. Inspired by the work of Vygotsky, we suggest the design of Vygotskian autotelic agents able to interact with others and, more importantly, able to internalize these interactions to transform them into cognitive tools supporting the development of new cognitive functions. This perspective paper proposes a new AI paradigm in the quest for artificial lifelong skill discovery. It justifies the approach by uncovering examples of new artificial cognitive functions emerging from interactions between language and embodiment in recent works at the intersection of deep reinforcement learning and natural language processing. Looking forward, it highlights future opportunities and challenges for Vygotskian Autotelic AI research. This presentation will be an overview of some of the ideas in this paper: <https://arxiv.org/pdf/2206.01134.pdf>.

3.18 The Never-Ending Visual classification Stream (NEVIS) 1.0

Marc'Aurelio Ranzato (DeepMind – London, GB)

License © Creative Commons BY 4.0 International license
© Marc'Aurelio Ranzato

Intelligent agents need to constantly adapt to change; for instance they need to adapt to change in the environment or change in the computation versus accuracy trade-off. Even modern large-scale models such as large vision and language models need to constantly adapt. They not only need to adapt to the current task but also use that experience to better learn future tasks. Unfortunately, there does not exist any benchmark today which is useful to investigate the question of how to efficiently adapt and consolidate knowledge over time and at scale. In this talk, I will provide an overview of NEVIS, a new benchmark which consists of a stream of very challenging and diverse visual classification tasks. I will then discuss the preliminary results we obtained using a variety of baseline approaches. NEVIS will be released in about a month, and it is meant to motivate researchers working in continual learning, meta-learning and auto-ml to join forces and to make strides together towards the development of robust systems that can become more apt and efficient over time.

3.19 Connecting 3D Shape Learning and Object Categorization

James M. Rehg (Georgia Institute of Technology – Atlanta, US)

License © Creative Commons BY 4.0 International license
© James M. Rehg

A classical topic in computer vision and psychology is the link between knowledge of 3D object shape and the ability to categorize objects. In this talk we revisit this link in two machine learning contexts that are connected to development: few-shot learning and continual learning. We show that learning a representation of 3D shape in the form of dense local descriptors provides a surprisingly powerful cue for rapid object categorization. Our shape-based approach to low-shot learning outperforms state-of-the-art models trained on category labels. We also present the first investigation of continual learning of 3D shape and demonstrate significant differences relative to continual category learning, finding that 3D shape learning does not suffer from catastrophic forgetting.

3.20 Human infants' brains are specialized for social functions


Rebecca Saxe (MIT – Cambridge, US)

License © Creative Commons BY 4.0 International license
© Rebecca Saxe

In this talk, I will argue that human infants have distinct social representations and motivations. Infants' learning about, and representations of, other people are not just a downstream consequence of generic processes that promote learning in the nonsocial environment, nor are they built by gradual, bottom-up adjustment to the statistics of visual experience. On the contrary, infants' attention to people depends on specific inferences about their social relevance; and is related to activity in distinctively social brain regions.

3.21 Towards Teachable Autonomous Agents: How can developmental psychology help?


Olivier Sigaud (Sorbonne University – Paris, FR)

License  Creative Commons BY 4.0 International license
© Olivier Sigaud

As a developmental AI researcher, I will outline a research program where we try to endow autotelic agents (agents who learn to represent, pursue and reach their own goals) with a teachability property, so that we can influence their goals through social interactions. With such agents, we can mimic guided play interactions with children, where they learn both on their own and from the guidance of a tutor or caregiver. Then I will show that such a research program faces the language grounding problem and that a central issue is the acquisition of language-sensitive sensorimotor representations. I will question existing lines of AI research related to this challenge and conclude by showing that developmental psychology research can bring a lot to address it, by providing relevant concepts, models and experimental data about it.

3.22 Why self-generated behavior has more radical consequences than you might originally think

Linda B. Smith (Indiana University – Bloomington, US)

License  Creative Commons BY 4.0 International license
© Linda B. Smith

Humans, including toddlers, are adept at taking knowledge from past experiences and using it in compelling new ways. Learning and generalization depend on both the learning machinery and the training data on which the machinery operates. This talk will highlight findings from studies of toddler's self-generated experiences. The main point is that everyday experiences occur in time-extended episodes. Each unique episode is characterized by a suite of coherence statistics. I propose that these statistics are the secret ingredient to innovative intelligence. Moreover, they provide novel insights into the internal processes that learn, generalize and innovate.

3.23 Rethinking the developmental pathway of early infant language learning

Daniel Swingley (University of Pennsylvania, US)


License  Creative Commons BY 4.0 International license
© Daniel Swingley

Prominent empirical results of the 1980s and 1990s in which infants were revealed to have learned aspects of their language's system of phonetic categories (consonants and vowels) contributed to a standard theoretical model in which infants first learn to perceive speech sounds, then aggregate these into possible words, and then seek to identify meanings for those words while grasping at regularities caused by grammar. Modeling approaches that are based on this pathway have shown how simple statistical heuristics computed over phoneme

sequences could help point infants to the early vocabulary. I will argue that this pathway is probably wrong and that current quantitative psychological models of infant word-form discovery are misguided. I will show that infant-directed speech is too variable and too unclear for such models to be plausible characterizations, and will sketch what an alternative looks like.

3.24 SCALa: A blueprint for computational models of language acquisition in social context

Sho Tsuji (University of Tokyo, JP)

License  Creative Commons BY 4.0 International license
© Sho Tsuji

Different views on language acquisition suggest a range of cues are used, from structure found in the linguistic signal, to information gleaned from the environmental context or through social interaction. Technological advances make it now possible to collect large quantities of ecologically valid data from young children's environment, but we still lack frameworks to extract and integrate such different kinds of cues from the input. SCALa (Socio-Computational Architecture of Language Acquisition) proposes a blueprint for computational models that makes explicit the connection between the kinds of information available to the social early language learner and the computational mechanisms required to extract language-relevant information and learn from it. SCALa further allows us to make precise recommendations for future large-scale empirical research.

3.25 Visual attention development in infancy


Ingmar Visser (University of Amsterdam, NL)

License  Creative Commons BY 4.0 International license
© Ingmar Visser

Eye-movements are a valuable source of information, next to responses and response times, for inferring cognitive states and processes. Infant research depends on eye-movements to a large extent as other behavioral response modalities are hard to use in this population. Eye-movement data comes with many challenges, many basic properties are not well known or understood. Optimal methods for defining fixations and saccades are still under much discussion. Free viewing presents a good way to study infant visual attention development and provides robust developmental trends for a number of phenomena that together form an interesting target for computational modeling.

3.26 Temporal patterns in vocal even sequences produced by human infants and computational vocal learning models

Anne Warlaumont (UCLA, US)

License  Creative Commons BY 4.0 International license
© Anne Warlaumont

In recent years, my collaborators and I have analyzed the timings of when over the course of a day human infants produce vocalizations. These patterns tend to have a somewhat fractal structure, wherein vocalizations occur in clusters within clusters within clusters in time. More recently we have begun to identify relationships between how close two consecutive infant vocal events are in time and how similar they are acoustically. And we are finding that infant vocalizations also tend to be more likely to occur in quick succession in the aftermath of hearing vocalizations produced by adults. We are developing some hypotheses for why these patterns may be important for infant vocal learning. An increase in infant vocalization rate following a reward (either social or intrinsic) may be a mechanism through which human infants can gain additional practice making specific sound types, capitalizing on the current state of the relevant neural and vocal apparatus. In other words, vocalization rate is potentially a pathway to achieving acoustically targeted vocal exploration. This pathway may be particularly useful given that infants' voluntary vocal control is limited; it may be a mechanism for bootstrapping vocal motor learning. Most computational models of vocal learning do not concern themselves with when vocalization occurs in the first place, and also don't consider vocalization-to-vocalization patterns. I expect that some modeling approaches will be better suited than others to addressing these aspects of human vocal learning. These temporal patterns may provide a useful dimension for comparing models to human data, and prioritizing a fit along this dimension may turn out to favor more biologically realistic architectures.

3.27 Curiosity in infants and computational models


Gert Westermann (Lancaster University, GB)

License  Creative Commons BY 4.0 International license
© Gert Westermann

Much of what we know about infants' cognitive development comes from studies in which infants are passive recipients of information presented to them on a computer screen in an order and duration determined by the experimenter. While this body of work has provided us with many insights about infants' learning and their cognitive abilities, these methods ignore a fundamental aspect of real-life learning: outside the lab, infants are actively involved in their learning through exploring their environment and engaging with information in the order and duration they choose. In our lab we investigate infants' information seeking using behavioural, eye tracking, EEG and computational modelling methods. I will give a very brief overview of the methods and studies currently going on in my lab, and then describe a simple auto-encoder neural network model used to simulate intrinsically-motivated exploration that is based on maximizing in-the-moment learning progress. This model learns a stimulus set used in seminal studies of infant category learning as well as a non-curious model embedded in an optimally structured external environment.

3.28 Magnifying Time and Space: New Ways of Studying Early Development and Learning from the Infant's Point of View

Chen Yu (University of Texas – Austin, US)

License  Creative Commons BY 4.0 International license
© Chen Yu

The three primary research goals in my lab are 1) to quantify the statistical regularities in the real world; 2) to examine the underlying learning mechanisms operated on the statistical data; and 3) to discover developmental pathways in a complex and multi-causal system. Toward the first goal, we have collected a corpus of infant-perspective visual scenes and infant gaze data as they play with their parent in a home-like environment. We have analyzed visual properties of infant-perspective scenes and quantified the ambiguity/transparency of individual parent naming events using infant gaze. We have also fed egocentric video to deep learning models to examine the quantity and quality of the statistical data that lead to successful learning. Toward the second goal, we have used the corpus of scenes that co-occur with parent naming to construct lab experiments which are composed of different mixes of high and low ambiguity naming events. Infants were trained and tested in multiple experimental conditions, varying in terms of the ambiguity of training trials and also in the composition and order of those trials to test specific hypotheses about statistical learning mechanisms. Toward the third goal, we have examined the social effects of joint attention in the development of the infant's own sustained attention and identified the potentially malleable pathway through which social interactions influence the self-regulation of sustained attention. I will conclude my talk by discussing developmental dependencies among motor development, visual perception, sustained attention, joint attention, and language learning.

3.29 Audio-visual self-supervised learning

Andrew Zisserman (University of Oxford, GB)

License  Creative Commons BY 4.0 International license
© Andrew Zisserman

Lesson 1 from the classic paper “The Development of Embodied Cognition: Six Lessons from Babies” is ‘Be Multimodal’. This talk explores how recent work in the computer vision literature on audio-visual self-supervised learning addresses this challenge. The aim is to learn audio and visual representations and tasks directly from the audio-visual data stream of a video (without providing any manual supervision of the data) – much as an infant could learn from the correspondence and synchronization between what they see and hear. It is shown that a neural network that simply learns to synchronize audio and visual streams is able to localize the faces that are speaking (active speaker detection). It is shown that a network that simply learns from the correspondence of faces and voice is able to cluster speakers according to their identity, and so be able to recognize the person from their face or voice.

4 Working Groups

Overview


We split the participants into five working groups to explore three main open-ended research questions related to developmental and machine learning as detailed below:

1. What are the connections between current computational research in self-supervised, weakly-supervised, and continual machine learning, and analogous developmental learning processes in humans and animals?
2. What is the role of computational models of learning (e.g., object recognition, machine perception, and reinforcement learning) in advancing developmental science? Can computational tools enable new developmental research questions? What kinds of data should developmental scientists produce that would be the most useful for computational approaches?
3. What is the role of multimodal learning (learning from diverse signal types such as visual, audio, touch, force, etc.) in development? What are the challenges and opportunities in multimodal machine learning?

Specifically, there are two groups investigated question 1 (Group 1.1 and 1.2), two groups did question 2 (Group 2.1 and 2.2) and one group examined question 3 (Group 3).

4.1 Group 1.1: The power of informational processing bottlenecks

Rhodri Cusack, Uri Hasson, Celeste Kidd, Marc'Aurelio Ranzato, Stefan Stojanov, Anh Thai

License  Creative Commons BY 4.0 International license

© Rhodri Cusack, Uri Hasson, Celeste Kidd, Marc'Aurelio Ranzato, Stefan Stojanov, Anh Thai

This working group explored the relationships between information processing bottlenecks in biological systems and machine learning techniques. Attention bottlenecks are pervasive in biological systems, for example in humans: 1) visual input streams are actively sampled by fixating only on one area of the field of view at a time, often in a context dependent way; 2) working memory is constrained to a few items and features; 3) humans are embodied and physically constrained to only perform one task at a time.

Information bottlenecks can be regarded both as a bug, because they force information to be thrown away, limit parallel processing, or turn an inherently multi-modal task into a unimodal one, or a feature, because they force abstraction, encourage generalization and reduce computational cost. Attention mechanisms have been developed in machine learning, in the form of transformers, LSTM networks. Further, the idea of core sets is concerned with removing data that is informative for learning by finding semantic redundancies. Last, specialised bottlenecked representations have been proposed e.g. variational autoencoders and sparse coding. This group identified that working to obtain high accuracy machine learning systems under constraints such as wall clock time, instantaneous compute, memory, hardware time, bandwidth, is potentially key to artificial general intelligence, in addition to the current trend of scaling model capacity and dataset size.

References

- 1 Cartwright-Finch, Ula, and Nilli Lavie. “The role of perceptual load in inattentive blindness.” *Cognition* 102.3 (2007): 321-340
- 2 Alvarez, George A., and Steven L. Franconeri. “How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism.” *Journal of vision* 7.13 (2007): 14-14.

- 3 Asplund, Christopher L., et al. “Surprise-induced blindness: a stimulus-driven attentional limit to conscious perception.” *Journal of Experimental Psychology: Human Perception and Performance* 36.6 (2010): 1372.
- 4 Feigenson, Lisa, and Justin Halberda. “Conceptual knowledge increases infants’ memory capacity.” *Proceedings of the National Academy of Sciences* 105.29 (2008): 9926-9930.

4.2 Group 1.2: Developmental AI Benchmark

Emmanuel Dupoux, James M. Rehg, Daniel Swingley, Anne Warlaumont, Gert Westermann, Chen Yu

License © Creative Commons BY 4.0 International license
 © Emmanuel Dupoux, James M. Rehg, Daniel Swingley, Anne Warlaumont, Gert Westermann, Chen Yu

This group conceptualized a new developmental AI benchmark focusing on speech articulation. The outcome of successfully accomplishing this challenge would be an artificial speech system that faithfully reproduces children’s speech production learning trajectories. The challenge is separated into three rounds.

The first round consists of learning a control model to articulate speech. More specifically, given an articulatory tract model [1](a model that simulates controllable muscles and vocal tract physics that can synthesize vocalizations including speech sounds), train a control system that can imitate heard speech, producing intelligible words. After training, the model will be tested on reproducing spoken English words and non-words. Such a system would be constrained by the physical properties of the human organs that are used to produce speech sounds, such that the control problem is dynamic and nonlinear, making it highly non-trivial.

The second round would consist of developing general learning and control algorithm such that a single model can operate multiple articulatory models provided (e.g. simulating individual and/or maturational differences among babies), requiring models to show the generality and adaptability characteristic of human learners.

The third round of this challenge is modeling child development trajectories. By providing realistic, recorded child input as a resource for model training, do similar child phonology phenomena appear in the trained models as they learn as in the children? Specific evaluation criteria may include matching children’s speech errors and matching prelinguistic vocal milestone sequences.


The group members will work on creating this challenge, which will involve creating an articulatory speech synthesizer and an API, creating training and evaluation datasets, and organizing the challenge.

References

- 1 Boersma, Paul. *Functional phonology*. Netherlands Graduate School of Linguistics, 1998.

4.3 Group 2.1: Embodied Intention Prediction Challenge

Michael Frank, Naithilee Kunda, Marvin Lavechin, Pierre-Yves Oudeyer, Rebecca Saxe, Maureen de Seyssel, Ingmar Visser

License  Creative Commons BY 4.0 International license

© Michael Frank, Naithilee Kunda, Marvin Lavechin, Pierre-Yves Oudeyer, Rebecca Saxe, Maureen de Seyssel, Ingmar Visser

The goal of this working group was to conceptualize a challenge that could be promoted to the AI community to facilitate the study of computational models that can learn about relationships between infants' attention and the language of their caregivers in naturalistic settings.

Investigating the relationship between language and attention is one way to operationalize larger problems around the role of multi-modal input in language development. For example, when a caregiver says "Look over there!", what is the response of the infant's visual attention, that is, where is the infant looking? Currently this problem is challenging to study because the mutual information between language and attention is hard to quantify and datasets are hard to annotate. Further, lab experiments on attention and intent may not generalize to naturalistic settings. We hoped that the challenge format would provide a focal point for bringing together annotated datasets and teams interested in bringing new models to bear.

This group proposed two complementary multi-modal prediction challenges. In both challenges, the model would have access to caregiver language, either in transcript or raw audio form, and video from an infant's head mounted camera. For the first challenge, the goal is to predict the infant's visual attention in a set of future frames, given the past caregiver language and infant's visual attention. For the second challenge, the goal is to predict the future caregiver language, from previous infant attention and caregiver language. The suggested input data length was 10 seconds, and then the language or attention prediction would be done over the next two seconds. The output for the visual attention challenge would be a vector indicating where the attention will move in the next 2 seconds, whereas for language it would be the words that the parent will say in the next 2 seconds.

Potential datasets that can be used for this challenge are the SAYCam [1] and SEED-Lings [2] datasets. The models can be evaluated in two testing regimes, within-subject (training and testing done on data from only one child) or between-subject (training done on pooled data from multiple infants, and tested on data from other infants not seen during training).

A further approach to test such a model would be to use it in a closed-loop time-extended manner to drive attention of a learning agent in an environment where it would interact with a human, and test whether interaction is structured and coordinated in a way that reproduces high-level properties of similar child-caretaker interactions.

References

- 1 Sullivan, Jessica, et al. "SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective." *Open mind* 5 (2021): 20-29.
- 2 Bergelson, Erika, and Richard N. Aslin. "Nature and origins of the lexicon in 6-mo-olds." *Proceedings of the National Academy of Sciences* 114.49 (2017): 12916-12921.

4.4 Group 2.2: ML for Causal Theories of Child Development?

David Crandall, Alejandrina Cristia, Hana D'Souza, Abdellah Fourtassi, Clément Romac, Olivier Sigaud

License © Creative Commons BY 4.0 International license
© David Crandall, Alejandrina Cristia, Hana D'Souza, Abdellah Fourtassi, Clément Romac, Olivier Sigaud

This group studied using machine learning (ML) in understanding causal theories of child development. Machine learning can be used for studying child development in three main ways, based on the purpose of the study: ML for annotation, ML for modeling, and ML for simulation.

Machine learning tools can be used to aid automatic annotation for both observational and experimental data in various settings and for different purposes (e.g., object detection, social cues transcription). Data shareability is identified as a potential issue, including privacy concerns and implementation infrastructure needed to host and process data.

Machine learning models are further helpful for predicting child target behaviors or understanding learning mechanisms. However, current techniques to understand the mechanistic process of machine learning models remain superficial.

Another utility of machine learning tools is to study emerging behaviors in artificial contexts. Designing the context and initializing the agent's biases and properties should be carefully considered in order to obtain meaningful results from the simulation.

This group further proposed different ways in which machine learning and developmental learning communities can collaborate and provide helpful scientific insights for each other regarding computational tools and naturalistic data.

4.5 Group 3: Multimodality in babies and machines

Thomas Carta, Hiromichi Hagihara, Felix Hill, Judy Hoffman, Eon-Suk Ko, Casey Lew-Williams, Atsushi Nakazawa, Jelena Suvevic

License © Creative Commons BY 4.0 International license
© Thomas Carta, Hiromichi Hagihara, Felix Hill, Judy Hoffman, Eon-Suk Ko, Casey Lew-Williams, Atsushi Nakazawa, Jelena Suvevic

This working group investigated the role of multimodal inputs in learning in both machines and babies. Multimodality is perceived differently by machine learning and developmental learning communities. For example, modality in human learning is a wide range of human senses from vision to social cues and motion awareness while the machine learning community mostly focuses on a much smaller set of modals such as vision and language or audio. One of the significant questions raised in the discussion concerns if we can bridge “the gap” between cross-disciplinary communities regarding learning with multimodal inputs.

Perceiving multimodal inputs is an inherent part of the human perception system [1, 2]. Different sensors can provide beneficial “redundant” information and create clearer learning moments that support one-shot learning where different sensors are complementary with each other, e.g. some certain toys might make some certain unique sounds. Another difference between human and machine is that multimodality learning cues in humans adapt depending on developmental needs [3, 4, 5]. For example, infants initially focus on faces when they are young but shift to hands which guide their attention to objects [4] when they are older. In contrast, in machines there is no such dynamic present in the integration of multimodal inputs.

Another distinction between learning in machines and babies lies in the processing efficiency of the underlying network structures (layered – machine vs dense – brain connection), synapses pruning and attention mechanisms.

An attempt to bridge the gap between human learning and machine learning is to construct data for learning from a wide-range of modalities for machines. Existing multimodal datasets such as SAYCam [6] and Databrary [7] provide more sensor data for machines than current standard datasets. This group proposed a means to obtain rich multimodality data for machines that is similar to play behavior observed in children.

References

- 1 Kuhl, Patricia K., and Andrew N. Meltzoff. "The bimodal perception of speech in infancy." *Science* 218.4577 (1982): 1138-1141.
- 2 Rosenblum, Lawrence D., Mark A. Schmuckler, and Jennifer A. Johnson. "The McGurk effect in infants." *Perception & psychophysics* 59.3 (1997): 347-357.
- 3 Mlinec, Miranda M., et al. "Posture matters: Object manipulation during the transition to arms-free sitting in infants at elevated vs. typical likelihood for autism spectrum disorder." *Physical & Occupational Therapy In Pediatrics* 42.4 (2022): 351-365.
- 4 Fausey, Caitlin M., Swapna Jayaraman, and Linda B. Smith. "From faces to hands: Changing visual input in the first two years." *Cognition* 152 (2016): 101-107.
- 5 Ko, E.-S., Abu-Zhaya, R., Kim, E.-S., Kim, T., On, K.-W., Kim, H., Zhang, B.-T., and Seidl, A. "Mothers' use of touch across infants' development and its implications for word learning: Evidence from Korean dyadic interactions", *Infancy* (2023). DOI: 10.1111/infa.12532
- 6 Sullivan, Jessica, et al. "SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective." *Open mind* 5 (2021): 20-29.
- 7 R. O. Gilmore, K. E. Adolph and D. S. Millman, "Curating identifiable data for sharing: The databrary project," 2016 New York Scientific Data Summit (NYSDS), New York, NY, USA, 2016, pp. 1-6, doi: 10.1109/NYSDS.2016.7747817.

Participants

- Thomas Carta
INRIA – Bordeaux, FR
- David J. Crandall
Indiana University –
Bloomington, US
- Alejandrina Cristia
LSCP – Paris, FR
- Rhodri Cusack
Trinity College Dublin, IE
- Hana D’Souza
Cardiff University, GB
- Maureen de Seyssel
INRIA & ENS Paris, FR
- Emmanuel Dupoux
LSCP – Paris, FR
- Abdellah Fourtassi
Aix-Marseille University, FR
- Michael C. Frank
Stanford University, US
- Hiromichi Hagihara
University of Tokyo, JP
- Uri Hasson
Princeton University, US
- Felix Hill
Google DeepMind – London, GB
- Judy Hoffman
Georgia Institute of Technology –
Atlanta, US
- Celeste Kidd
University of California –
Berkeley, US
- Eon-Suk Ko
Chosun University, KR
- Maithilee Kunda
Vanderbilt University, US
- Marvin Lavechin
Meta AI – Paris, FR
- Casey Lew-Williams
Princeton University, US
- Atsushi Nakazawa
Kyoto University, JP
- Pierre-Yves Oudeyer
INRIA – Bordeaux, FR
- Marc’Aurelio Ranzato
DeepMind – London, GB
- James M. Rehg
Georgia Institute of Technology –
Atlanta, US
- Clement Romac
INRIA – Bordeaux, FR
- Rebecca Saxe
MIT – Cambridge, US
- Olivier Sigaud
Sorbonne University – Paris, FR
- Stefan Stojanov
Georgia Institute of Technology –
Atlanta, US
- Jelena Sucevic
University of Oxford, GB
- Daniel Swingley
University of Pennsylvania, US
- Ngoc Anh Thai
Georgia Institute of Technology –
Atlanta, US
- Ingmar Visser
University of Amsterdam, NL
- Anne Warlaumont
UCLA, US
- Gert Westermann
Lancaster University, GB
- Chen Yu
University of Texas – Austin, US



Remote Participants

- Kristen Grauman
University of Texas – Austin, US
- Jitendra Malik
University of California –
Berkeley, US
- Linda B. Smith
Indiana University –
Bloomington, US
- Sho Tsuji
University of Tokyo, JP
- Andrew Zisserman
University of Oxford, GB

Data-Driven Combinatorial Optimisation

Emma Frejinger^{*1}, Andrea Lodi^{*2}, Michele Lombardi^{*3}, and
Neil Yorke-Smith^{*4}

1 University of Montreal, CA. emma.frejinger@umontreal.ca

2 Cornell Tech – New York, US. andrea.lodi@cornell.edu

3 University of Bologna, IT. michele.lombardi2@unibo.it

4 TU Delft, NL. n.yorke-smith@tudelft.nl

Abstract

Machine learning’s impressive achievements in the last decade have urged many scientific communities to ask if and how the techniques developed in that field to leverage data could be used to advance research in others. The combinatorial optimisation community is one of those, and the area of data-driven combinatorial optimisation has emerged. The motivation of the seminar and its design and development have followed the idea of making researchers both in academia and industry belonging to different communities – from operations research to constraint programming, from artificial intelligence to machine learning – communicate, establish a shared language, and ultimately (try to) set the roadmap for the development of the field.

Seminar 23–28 October 2022 – <http://www.dagstuhl.de/22431>

2012 ACM Subject Classification Theory of computation → Constraint and logic programming; Computing methodologies → Machine learning; Theory of computation → Mathematical optimization; Theory of computation → Reinforcement learning

Keywords and phrases combinatorial optimisation, constraint programming, machine learning, Mixed integer programming, operations research, Reinforcement learning

Digital Object Identifier 10.4230/DagRep.12.10.166

1 Executive Summary

Emma Frejinger (University of Montreal, CA)

Andrea Lodi (Cornell Tech – New York, US)

Neil Yorke-Smith (TU Delft, NL)

License  Creative Commons BY 4.0 International license
© Emma Frejinger, Andrea Lodi, and Neil Yorke-Smith

In the last five years, an area now being influenced in a new way by machine learning (ML) is combinatorial optimisation (CO). Combinatorial optimisation is studied for both its importance in theory, since CO problems are NP-hard problems, and for its importance in real-world decisions, for example, planning drivers and routes for a fleet of delivery vehicles. CO problems are studied in operations research (OR) and also traditionally in symbolic artificial intelligence (AI) such as constraint programming (CP) and satisfiability modulo theories.

This Dagstuhl Seminar built on the fast-growing interest in combining ML with ‘traditional’ AI methodologies like CP, and with OR more generally [1, 2]. Surveying the scattered initiatives, the seminar had the ambition to set the agenda for constraint-based ‘Combinatorial Optimisation 2.0’. Historically, several communities have focussed on different approaches to CO, mostly in a disjoint manner. This division between, on the one hand, the OR and

* Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Data-Driven Combinatorial Optimisation, *Dagstuhl Reports*, Vol. 12, Issue 10, pp. 166–174

Editors: Emma Frejinger, Andrea Lodi, Michele Lombardi, and Neil Yorke-Smith



DAGSTUHL REPORTS Dagstuhl Reports

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

symbolic AI communities, and on the other, the ML and functional AI communities, is historically strong. While in recent years a dialogue between symbolic and functional AI communities has emerged, there remains too little connection between the discrete OR and ML communities.

The seminar was organised by Emma Frejinger (Canada), Andrea Lodi (USA), Michele Lombardi (Italy) and Neil Yorke-Smith (Netherlands). Michele was unable to attend in person, due to last minute circumstances, and joined plenary parts of the seminar online. Similarly, it was necessary for Pierre-Luc Bacon to give his tutorial remotely.

Seminar Overview

The seminar opened with four tutorials, whose abstracts are given in this report, on the topics of CP (by Tias Guns), mixed (non)-linear integer programming (MIP) (by Ruth Misener), end-to-end ML for CO (by Ferdinando Fioretto), and reinforcement learning (RL) (by Pierre-Luc Bacon).

The seminar included a set of informal short introductory and topical talks, and sessions of collaborative planning. The overarching questions that structured this planning are, on the one hand, (1) how ML can help in modelling or solving CO problems – or both modelling and solving – and in particular constraint-based models and solving; and on the other hand, (2) how CO can help in tasks approached using ML, including ML training and algorithms. Then, (3) what problems and tasks can be addressed (only) by the synergistic combinations of these methodologies?

Through discussions, the participants identified jointly six topics to be approached in smaller working groups: (i) self-supervised representation learning for combinatorial optimisation, (ii) uncertainty, prediction, optimisation and decision-focussed learning, (iii) OR for ML, (iv) vehicle routing and the role of ML, (v) ML-augmented MIP solvers, and (vi) fairness. The groups discussed challenges, existing work and identified open research questions with promising future avenues at the intersection between OR and ML. The working groups are summarised below.

The outcomes of the seminar in furthering the development of a community at the intersection of OR and ML are expected to be felt in the coming couple of years. Already, however, there are tangible outcomes in terms of roadmap ideas, an open online discussion forum (Slack)¹, multiple new collaborations, and a research grant submitted. A special issue of the journal *Frontiers in Applied Mathematics and Statistics* is organised by one of the participants.

The scientific programme was beautifully facilitated by the surroundings and academic services of Schloss Dagstuhl. Further, on the opening evening of the seminar, volunteers among the participants took part in “slide bingo”. During this humorous session they improvised presenting the slides of others. On Wednesday afternoon, the participants took a walk to a nearby village in unexpectedly fine sunny weather for October.

¹ On Slack, ML<>CO

Reflections on the Week

All communities present at the seminar found benefit from the interactions and discussions. Meeting in person at the scale of a Dagstuhl Seminar was much appreciated! Participants were aware of the differing emphases, mindsets, and publication practices of different communities. In general, it was felt that strengthening the connection between ML and OR helps in bridging the gap between predictive and prescriptive analytics, which can benefit industrial or government actors and citizens alike.

Working Groups

In this section, we briefly summarise the discussions in the six working groups.

Self-supervised representation learning for combinatorial optimisation

This working group enjoyed lively discussions around the concept of a “foundational model” for CO. The motivation is to avoid retraining from scratch when there is a relatively small change. The group discussed transfer learning in terms of problem formulation, downstream task and instance distributions.

The group identified open questions:

- What is the equivalent of saving models/checkpoints in ML or natural language processing (NLP) in data-driven CO (DDCO)?
- Can we share pre-trained models to generate SAT/CP/MIP embeddings without training again?
- NLP has the concept of a tokeniser that preprocesses the text before it gets fed into the network; in DDCO we would need similar pre processors that transfer the problem instances into the model’s expected (graph) structure.
- What is the equivalent of “large” aspect from large language NLP models for DDCO?
- Is there a (super) GLUE benchmark equivalent for DDCO?

Uncertainty, prediction, optimisation and decision-focussed learning

Decision-focussed learning aims at training prediction models against a loss reflecting the quality of decisions instead of a classic prediction loss. This working group weighed up the questions: when is decision-focussed learning (DFL) better than (traditional) alternatives? The group gave energy into thinking about stochastic formulations, data perturbation and interpretability of DFL. The group also identified a connection with RL, in particular contextual bandits (single-stage decision). Since there has been some confusion around the terminology, the group recommended to use “decision-focussed learning” instead of “predict and optimise” or “predict + optimise”.

The group wrote down example problems for three settings of decision focussed learning for CO. First, unknown parameters in the objective. This is the most studied case and there are several applications in the literature. Second, unknown parameters in the right-hand side of the constraints. For example, transport network planning where demand predictions occur in capacity constraints. Third, unknown parameters in the left-hand side of the constraints. For example, healthcare scheduling problems where treatment durations are predicted and should not exceed a given schedule length.

Operations research for machine learning

This working group was provoked by the feeling in the OR community that the ML community is seldom happy with discrete optimisation. In other words, how can CO have an influence on problems that generally come from the ML community? Those from OR background in the group expressed that they want to make real contributions to ML.

The group proceeded to outline three major obstacles:

- Scalability. OR methods tend to be limited when dealing with extremely large datasets that are often associated with ML applications.
- Optimality. OR methods have been designed in general to provide guarantees. Computing confidence bounds for ML applications would be excellent but one major obstacle is that the loss function is computed over samples from an unknown distribution. In other words, there is uncertainty with respect to the real objective function, which would require a re-interpretation of what confidence bounds are.
- Software. Whereas the ML community is used to working with self-installing open-source software, the OR community uses more cumbersome, often commercial software.

The recommendation was for research into a new generation of optimisation-based heuristics. The group identified four areas in which discrete optimisation methods are likely relevant: (i) optimal transport problems, (ii) neural-network verification, (iii) the broad area of fairness, explainability and interpretability, and (iv) training for Gaussian processes with tree kernels.

Vehicle routing and the role of machine learning

This working group ascertained exciting research on using ML to help solve routing problems. Two main aspects are, first, that the current state of (deployed) routing software has no idea whether it has seen a problem before, the types of problems being solved, and so forth; and second, anticipating the future – for instance dynamic settings, demand estimation, service time estimation, and so forth – would make the solution of routing problems even more relevant in practice. The working group felt that leveraging ML in both aspects could lead to significant improvements.

The group identified open questions:

- Does the ML model output individual actions or does it output an instance-specific heuristic?
- Can we learn insights about the problem from the predictions?
- Where does or could deep RL work best?
- Can we make a unified routing model (a ‘foundation model’)?

Machine learning augmented MIP solvers

This working group started from the general questions: which is the big challenge in MIP solving? And, will ML-augmented MIP solvers be ever significantly better than improved versions of the current solvers?

The group found that one significant motivation to go for ML-augmented MIP is ‘democratisation’: ML could allow the more general use of MIP technology by automatising some steps of MIP development and solution that depend on the specific a class of problems in hand without requiring the intervention of experts in the loop. Such a democratisation would require the definition of a robust pipeline on how to learn – from data – tasks like branching, cutting, preprocessing, etc. on the specific class of instances in hand, i.e., characteristic

of the application one wants to solve. Such a robust pipeline does not exist yet though there is strong evidence of successful stories where ML-augmented tasks are performed more efficiently than in classical MIP solvers. Some of those successful examples have been already integrated into the solvers, even commercial ones.

The group identified a number of interesting directions, as yet unexplored in the fast-moving subfield of ML-for-MIP. Among these are hypothesis generation, defining appropriate performance metrics, learning for cutting plane generation and selection. The group also discussed benchmark libraries.

Fairness

This working group sought to learn more about data-driven CO models for fairness. The group recognised that an issue with fairness is already in its definition. While in the ML community, the concept of fairness is related to a tradeoff between overall accuracy and group accuracy, in CO for decision making, there is not a clear definition of fairness.

The group discussed an application in the online scheduling of radiologists and neuro-radiologists of CT scans. In this context, because of the scarcity of the resources, fairness is associated with their correct and ‘fair’ use.

Fairness at large is also related to explainability and interpretability and the working group discussed the use of classical CO methods that tend to be more interpretable of the ML ones that are often perceived as black-boxes. Further, a potential important area in this context is that of integrating ML and OR to achieve a higher level of explainability, for example by improving methods like decision trees and ML classification algorithms.

References

- 1 Yoshua Bengio, Andrea Lodi, Antoine Prouvost: *Machine learning for combinatorial optimization: A methodological tour d’horizon*. Eur. J. Oper. Res. 290(2): 405-421 (2021). <https://doi.org/10.1016/j.ejor.2020.07.063>
- 2 James Kotary, Ferdinando Fioretto, Pascal Van Hentenryck, Bryan Wilder: *End-to-End Constrained Optimization Learning: A Survey*. IJCAI 2021: 4475-4482. <https://doi.org/10.24963/ijcai.2021/610>

2 Table of Contents

Executive Summary

Emma Frejinger, Andrea Lodi, and Neil Yorke-Smith 166

Overview of Talks

End-to-end constrained optimization learning

Ferdinando Fioretto 172

Data-driven combinatorial optimisation, with a CP flavour

Tias Guns 172

Machine learning for mathematical optimization and mathematical optimization
for machine learning

Ruth Misener 173

An overview of reinforcement learning and learning for control

Pierre-Luc Bacon 173

Participants 174

3 Overview of Talks

3.1 End-to-end constrained optimization learning

Ferdinando Fioretto (Syracuse University, US)

License © Creative Commons BY 4.0 International license
© Ferdinando Fioretto

Main reference James Kotary, Ferdinando Fioretto, Pascal Van Hentenryck, Bryan Wilder: “End-to-End Constrained Optimization Learning: A Survey”, in Proc. of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021, pp. 4475–4482, ijcai.org, 2021.

URL <https://doi.org/10.24963/ijcai.2021/610>

This tutorial reviews the recent advancements made in using constrained optimization to incorporate structural information and domain knowledge into machine learning models. We start by reviewing how to convert optimization into differentiable layers to use in machine learning models. Such integration enables to enforce structural information and domain knowledge into machine learning models. Next, we focus on extending this setting by integrating constrained optimization to enforce structure in the outputs of learned embeddings, leading to end-to-end decision-focused learning, that trains models to directly optimize the performance in targeted applications. Finally, we review techniques to learn constrained optimization surrogates by leveraging a distribution of optimization problems and their solutions leading to enhanced optimization modeling technology for operations research decision tasks. The tutorial concludes with a discussion of challenges and open questions.

3.2 Data-driven combinatorial optimisation, with a CP flavour

Tias Guns (KU Leuven, BE)

License © Creative Commons BY 4.0 International license
© Tias Guns

Joint work of Tias Guns, Rocs Canoy, Jayanta Mandi, Maxime Mulamba, Victor Bucarey Lopez, Emilio Gamba, Ignace Bleukx, Michelangelo Diligenti, Michele Lombardi, Bart Bogaerts

Main reference Mohit Kumar, Samuel Kolb, Tias Guns: “Learning Constraint Programming Models from Data Using Generate-And-Aggregate”, in Proc. of the 28th International Conference on Principles and Practice of Constraint Programming, CP 2022, July 31 to August 8, 2022, Haifa, Israel, LIPICs, Vol. 235, pp. 29:1–29:16, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022.

URL <https://doi.org/10.4230/LIPICs.CP.2022.29>

We first provided a general overview of different constraint solving technologies, including constraint programming. We then discussed the general trend of using machine learning to either learn (part of) the model, the problem specification, or to learn how to obtain solutions faster. Both topics were well covered in the seminar overall.

The rest of the talk then focussed on recent work in learning part of the model, more specifically learning the constraints using passive or (inter)active constraint learning, as well as learning the objective using pre-training networks and integrating constraint solving in the inference, with examples in visual constraint solving, learning preferences in vehicle routing and more.

The tutorial ended with how using learning part of the model also increases the need for explainable models, both at the machine learning and the constraint solving side.

3.3 Machine learning for mathematical optimization and mathematical optimization for machine learning

Ruth Misener (Imperial College London, GB)

License © Creative Commons BY 4.0 International license
© Ruth Misener

Joint work of Ruth Misener, Calvin Tsay, Francesco Ceccon, Jordan Jalving, Joshua Haddad, Alexander Thebelt, Carl Laird, Miten Mistry, Jan Kronqvist, Radu Baltean, Pierre Bonami, Andrea Tramontani

Main reference Francesco Ceccon, Jordan Jalving, Joshua Haddad, Alexander Thebelt, Calvin Tsay, Carl D. Laird, Ruth Misener: “OMLT: Optimization & Machine Learning Toolkit”, CoRR, Vol. abs/2202.02414, 2022.

URL <https://arxiv.org/abs/2202.02414>

We consider how machine learning can be used for expediting mathematical optimization solvers. We also consider how mathematical optimization can contribute to machine learning. This presentation is biased towards the kind of optimization I understand, so it mostly concerns mixed-integer nonlinear optimization.

3.4 An overview of reinforcement learning and learning for control

Pierre-Luc Bacon

License © Creative Commons BY 4.0 International license
© Pierre-Luc Bacon

Main reference Evgenii Nikishin, Romina Abachi, Rishabh Agarwal, Pierre-Luc Bacon: “Control-Oriented Model-Based Reinforcement Learning with Implicit Differentiation”, in Proc. of the Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 – March 1, 2022, pp. 7886–7894, AAAI Press, 2022.

URL <https://ojs.aaai.org/index.php/AAAI/article/view/20758>

The success of the field of reinforcement learning hinges upon its multidisciplinary nature. This tutorial highlights significant contributions from other disciplines, such as optimal control, operations research and simulation, to build a more robust theoretical understanding of modern algorithms. We begin our overview by studying the class of temporal difference learning algorithms as an application of the stochastic approximation method. We then see how sample average approximation and the ‘stochastic counterpart’ method in stochastic optimization can offer insights into the class of fitted value methods behind the latest advances in deep reinforcement learning. Using the tools for sensitivity analysis in simulation, we then explore how ideas in derivative estimation give rise to the class of policy gradient methods in reinforcement learning. Finally, we conclude our tour d’horizon by broadening our perspective on reinforcement learning by merging optimization techniques in optimal control with learning through end-to-end or decision-aware methods.

Participants

- Karen Aardal
TU Delft, NL
- Claudia D’Ambrosio
Ecole Polytechnique –
Palaiseau, FR
- Bistra Dilkina
USC – Los Angeles, US
- Ferdinando Fioretto
Syracuse University, US
- Emma Frejinger
University of Montreal, CA
- Maxime Gasse
Polytechnique Montréal, CA
- Stefano Gualandi
University of Pavia, IT
- Oktay Gunluk
Cornell University – Ithaca, US
- Tias Guns
KU Leuven, BE
- Serdar Kadioglu
Brown University –
Providence, US
- Lars Kotthoff
University of Wyoming –
Laramie, US
- Hoong Chuin Lau
SMU – Singapore, SG
- Pierre Le Bodic
Monash University –
Clayton, AU
- Andrea Lodi
Cornell Tech – New York, US
- Marco Lübbecke
RWTH Aachen, DE
- Sofia Michel
NAVER Labs Europe –
Meylan, FR
- Andrea Micheli
Bruno Kessler Foundation –
Trento, IT
- Ruth Misener
Imperial College London, GB
- Laurent Perron
Google – Paris, FR
- Sebastian Pokutta
Zuse Institut Berlin, DE
- Louis-Martin Rousseau
Polytechnique Montréal, CA
- Helge Spieker
Simula Research Laboratory –
Oslo, NO
- Kevin Tierney
Universität Bielefeld, DE
- Pashootan Vaezipoor
University of Toronto, CA
- Pascal Van Hentenryck
Georgia Institute of Technology –
Atlanta, US
- Stefan Voß
Universität Hamburg, DE
- Neil Yorke-Smith
TU Delft, NL
- Yingqian Zhang
TU Eindhoven, NL



Towards a Unified Model of Scholarly Argumentation

Khalid Al-Khatib^{*1}, Anita de Waard^{*2}, Dayne Freitag³,
Iryna Gurevych^{*4}, Yufang Hou^{*5}, and Harrison Scells^{†6}

- 1 University of Groningen, NL. khalid.alkhatib@rug.nl
- 2 Elsevier – Jericho, US. a.dewaard@elsevier.com
- 3 SRI International, US. daynefreitag@sri.com
- 4 TU Darmstadt, DE. gurevych@cs.tu-darmstadt.de
- 5 IBM Research – Dublin, IE. bnuxiaofang@gmail.com
- 6 Universität Leipzig, DE. harry.scells@uni-leipzig.de

Abstract

This report summarizes the outcomes of the Dagstuhl Seminar 22432: “Towards a Unified Model of Scholarly Argumentation.” The purpose of this Seminar was to enable robust advances in argumentation technology by collecting and collaborating on use cases in scholarly and biomedical discourse and working on a foundational model for argumentation in science and healthcare. Most importantly, the seminar served to develop a multidisciplinary, international research community devoted to building and maintaining principles, tools, and models for studying scholarly argumentation. Over the course of the seminar week, the seminar laid the foundation of a shared formalism, illuminated important scholarly use cases for argumentation modeling, and identified directions for future exploration.

Seminar October 23–28, 2022 – <http://www.dagstuhl.de/22432>

2012 ACM Subject Classification Computing methodologies → Artificial intelligence; Theory of computation; Computing methodologies → Machine learning

Keywords and phrases Argument mining, Argument modeling, Scholarly discourse

Digital Object Identifier 10.4230/DagRep.12.10.175

1 Executive Summary

Khalid Al-Khatib (University of Groningen, NL, khalid.alkhatib@rug.nl)

Anita de Waard (Elsevier-Jericho, US, a.dewaard@elsevier.com)

Iryna Gurevych (TU Darmstadt, DE, iryana.gurevych@tu-darmstadt.de)

Yufang Hou (IBM Research-Dublin, IE, yhou@ie.ibm.com)

License  Creative Commons BY 4.0 International license

© Khalid Al-Khatib, Anita de Waard, Iryna Gurevych, and Yufang Hou

Background

Argumentation is prevalent in scientific discourse and critical to scientific progress. Recent efforts have attempted to identify and model argumentative structures in scholarly discourse from different perspectives. Within the domain of scientific literature analysis, computational approaches to argumentation have followed the route of discourse modeling by identifying relations between spans and clauses encoding rhetorical structures (e.g., premises and conclusions), or as typed turns in community debate (e.g., supports or attacks). Another thread of research, often applied to biomedical literature, focuses on capturing functional

* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Towards a Unified Model of Scholarly Argumentation, *Dagstuhl Reports*, Vol. 12, Issue 10, pp. 175–206

Editors: Khalid Al-Khatib, Anita de Waard, Dayne Freitag, Iryna Gurevych, Yufang Hou, and Harrison Scells



Dagstuhl Reports

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

discourse at different levels of granularity, such as objectives, methods, results or scientific claims, and their relation to reported evidence. Most work adopts a corpus perspective, either highlighting the role of sentences or phrases within the scientific discourse or aligning claims across documents, and using citations to construct claim-evidence networks that summarize the state of knowledge in a field. Within the health sciences, argumentative structures have been used to automate the production of systematic reviews by identifying key actionable knowledge elements from collections of clinical reviews, case studies, and research papers. For an overview of previous work, see e.g. [1].

Despite these varied efforts and the clear practical importance of the work, there is lack of consensus on how scientific argumentation should be formalized. For instance, it remains unclear whether formalisms popular in non-scientific domains apply to scientific discourse, and whether a single formalism can adequately support argumentation research in diverse disciplines such as biology, chemistry, materials science, medical research and computer science. This lack of consensus manifests in a dearth of shared reference corpora, which are needed to advance research into computational treatments of scientific argumentation. It has also led to the absence of an operational theory for defining argumentative components in scholarly text.

Goals

Our Dagstuhl Seminar, titled *Towards a Unified Model for Scholarly Argumentation*, sought to further the emergence of this missing consensus. Specifically, the seminar objectives included:

- Enabling robust advances in argument technology by collecting and working on use cases in scholarly and medical discourse;
- Starting the development of a foundational model for argumentation in science and healthcare;
- Laying the groundwork for a multidisciplinary community devoted to building and maintaining principles, tools, and models to identify key components in scholarly argumentation.

Outcomes

The seminar was attended by scientists at different levels of seniority and from a variety of research backgrounds. Some participants have made the computational modeling of argumentation or the scholarly literature the central focus of their careers. Others were drawn to the seminar through their work on applications in adjacent problem areas. Ultimately, all emerged with a sense that important bonds of shared interest had formed, fostered by several seminar outcomes.

Knowledge Baseline

A shared understanding of the problem space was obtained, through a series of keynotes and panel discussions on theory, models, tools, and available corpora. These are described in greater detail in this report, in Section 3. In particular, two introductory talks summarized the state of the art in argument modeling (3.1) and computational argument mining (3.2).

Five further plenary talks described different use cases where argument identification can support NLP tasks:

- using scientific discourse to understand and measure the impact of scholarly contributions (3.3);
- using argument modeling to generate discourse (3.4);
- generating scholarly documents using argument structures (3.5);
- interpreting a fortiori arguments (3.6);
- synthesizing evidence from text to support public policy (3.7).

A series of eleven flash talks covered a host of other efforts, presenting corpora, tools, and relevant applications, such as document understanding, extracting high-level claims, and identifying fallacious and persuasive elements in scholarly texts (Section 4).

Problem Elucidation

At the beginning of the workshop, the group identified several important focus areas that then became the subject of breakout group deliberation over the course of the week. All materials, including the full program, slides, summaries of the breakout sessions and code and corpora submitted can be found on the workshop Google Drive at <https://bit.ly/TUMSA22>.

- *Foundations* (Section 5.1). A subgroup of participants discussed a shared argumentation model, based on the various proposals presented during the plenary sessions. The group debated and wrote a first-order consensus of these varying views, which can be used for further development of a foundational model of scholarly argumentation.
- *Domains* (Section 5.2). This working group pursued a comparison of argumentation in different scholarly domains. A methodology was delineated for how to annotate argumentation across domains while reducing the need for domain experts.
- *Argument Quality* (Section 5.3). This working group explored how argumentation quality can be evaluated, and defined a series of questions to assess this. Additionally, the group members contributed an open-source tool to perform the evaluation of argumentation quality, which can be further developed to support this task.
- *Community Dialogue* (Section 5.4). This working group looked at how argument structure can support an important editorial task, namely to decide on an accept or reject decision for a submitted manuscript, based on a number of peer reviews. The group developed a corpus of dialogues that simulate how a meta-reviewer asks questions about a document that has received a number of reviews, which can be used in future work in this domain.

Community Formation

Building on the connections developed during the seminar, a series of collaborations have been fostered, and thoughts on how to proceed with this work through a multidisciplinary lens have been put forth. Multiple new collaborations have been formed as an outcome of this week, in some cases centered on new tools and research corpora first conceived in the workshop.

Next Steps

This Dagstuhl Seminar brought together a multi-disciplinary, international, and diverse community of researchers from academia and industry to discuss scholarly argumentation. Much argumentation occurred, during and after presentations, in breakout groups, during

the social events spread out through the week, and long into the night. Necessarily, this is only the beginning of a conversation that will unfold over the coming years, one that will ultimately produce a shared model of scholarly argumentation and a set of concrete research tasks and important new use cases.

We hope that this seminar was the first in a series of events devoted to this topic, that this inaugural event proves pivotal in the formation of a cohesive research community addressing a problem with large practical ramifications. This report can hopefully contribute to accelerate work in this area, by offering a summary of current efforts, and a number of interesting problems to work on.

References

- 1 Khalid Al Khatib, Tirthankar Ghosal, Yufang Hou, Anita de Waard, and Dayne Freitag. 2021. Argument Mining for Scholarly Document Processing: Taking Stock and Looking Ahead. In Proceedings of the Second Workshop on Scholarly Document Processing, pages 56–65, Online. Association for Computational Linguistics.

2 Table of Contents

Executive Summary

Khalid Al-Khatib, Anita de Waard, Iryna Gurevych, and Yufang Hou 175

Introductory Talks

An introduction to Models of Argumentation
Graeme Hirst and Chris Reed 181

Computational Argumentation in Scholarly Discourse
Khalid Al-Khatib and Henning Wachsmuth 181

Towards Automatically Understanding and Measuring the Contributions of Scientific Work
Maria Liakata 181

The Role of Text Generation in Argumentation
Smaranda Muresan 182

InterText: Modeling Text as a Living Object in Cross-Document Context
Iryna Gurevych 182

Formalizing and Generating the Structure of Scholarly Papers
Eduard Hovy 183

Towards Automatic Interpretation of A Fortiori Arguments
Simone Teufel 183

Using the Claim Framework to Inform Public Policy
Ryan Wang 184

Flash Talks

Narrative Structures in Scientific Documents
Wolf-Tilo Balke 184

Argumentation in Biochemistry Articles
Robert Mercer 185

Linking Computational Argumentation to Information Quality
Davide Ceolin 186

Building Computational Models to Understand Scholarly Documents
Yufang Hou 186

PEER – Collaborative Lightweight Argument Annotation
Nils Dycke 187

Towards Constructive Conversations
Andreas Vlachos 187

Expressing High-Level Scientific Claims with Formal Semantics
Davide Ceolin 188

Argumentation, Persuasion, Propaganda, and More
Preslav Nakov 188

Fallacies in Political Argumentation
Serena Villata 189

Communicating Scientific Work with the Public through Dialogue Initiative <i>Milad Alshomary and Smaranda Muresan</i>	189
BAM: Benchmarking Argument Mining on Scientific Documents <i>Florian Ruosch</i>	190
Working Groups	
Foundations of Scholarly Argumentation <i>Elena Cabrio, Graeme Hirst, Eduard Hovy, Maria Liakata, Robert Mercer, Smaranda Muresan, Preslav Nakov, Chris Reed, Florian Ruosch, Simone Teufel, Serena Villata</i>	190
Cross-domain Argumentation Model for Scholarly Argumentation <i>Khalid Al-Khatib, Fengyu Cai, Dayne Freitag, Daniel Garijo, Benno Stein, Henning Wachsmuth</i>	196
Evaluation of Argument Quality <i>Yufang Hou, Tobias Mayer, Domenic Rosati, Harrisen Scells, Ferdinand Schlatt, Simone Teufel, Ryan Wang</i>	200
Scholarly Argumentation as a Community Dialogue <i>Wolf-Tilo Balke, Andreas Vlachos, Davide Ceolin, Milad Alshomary, Nils Dycke, Sukannya Purkayastha, Iryna Gurevych, Anne Lauscher, Tilman Beck</i>	202
Participants	206

3 Introductory Talks

3.1 An introduction to Models of Argumentation

Graeme Hirst (University of Toronto, CA) and Chris Reed (University of Dundee, UK)

License © Creative Commons BY 4.0 International license
© Graeme Hirst and Chris Reed

We reviewed the fundamental concepts of arguments and argumentation, including the basic elements of arguments, the types of argument structures, and the types of attacks on arguments. We introduced the idea of argumentation as a dialogue game, and the conditions required of a well-formed argument. We outlined the Toulmin model of argumentation, and explain the concept of argumentation schemes as templates for arguments.

3.2 Computational Argumentation in Scholarly Discourse

Khalid al-Khatib (University of Groningen, NL) and Henning Wachsmuth (Leibniz Universität Hannover, DE)

License © Creative Commons BY 4.0 International license
© Khalid Al-Khatib and Henning Wachsmuth

Computational argumentation deals with the computational analysis and synthesis of natural language arguments. In this tutorial talk, we provided an overview of computational argumentation from a natural language processing (NLP) perspective, and we reviewed the state of the art of computational argumentation in scholarly discourse. Starting from the basics of human argumentation, the first part of the talk introduced the central tasks of argument mining, argument assessment, and argument generation. We then looked at the latest trends for these tasks considering audience-specific argument quality assessment and knowledge encoding during argument generation. In the second part, we concentrated on scholarly discourse organizing existing research based on the domains being tackled and the argument models built on. Most existing work addresses the creation of new corpora for scholarly documents and the mining of their argumentative structure. We discussed the main envisioned applications of computational argumentation in scholarly discourse and the challenges towards these.

3.3 Towards Automatically Understanding and Measuring the Contributions of Scientific Work

Maria Liakata (The Alan Turing Institute – London, UK & Queen Mary University of London, UK)

License © Creative Commons BY 4.0 International license
© Maria Liakata

Researchers have been working on the automatic extraction of information from scientific articles for over two decades. A key aspect in this line of research is capturing how scientists discuss their work, the scientific discourse. In my talk I gave a brief overview of early work on identifying the scientific discourse and how this can improve downstream tasks involving

the extraction of information from the scientific literature. I then showed a number of neural approaches to capturing scientific argumentation in a multi-task learning setting. I also presented recent work on the relation between the scientific discourse and the way it is represented in the news through cross-document cross-domain coreference between scientific articles and news and press releases that refer to the scientific articles, as a step towards understanding the more comprehensive (non-academic) impact of scientific work.

3.4 The Role of Text Generation in Argumentation

Smaranda Muresan (Columbia University – New York City, USA)

License  Creative Commons BY 4.0 International license
© Smaranda Muresan

Large-scale language models based on transformer architectures, such as GPT-3 or BERT, have advanced the state of the art in Natural Language Understanding and Generation. However, even though these models have shown impressive performance for a variety of tasks, they often struggle with reasoning and modeling implicit meaning, which are required for understanding and generating argumentative text. In this talk, I presented some of our recent work on text generation models for argumentation. There are several challenges we have to address to make progress in this space: 1) the need to model commonsense knowledge; 2) the lack of large training datasets. I discussed our proposed theoretically-grounded knowledge-enhanced text generation models for enthymeme reconstruction and for recognizing argument fallacies. I concluded by discussing opportunities and remaining challenges for neural text generation systems for argumentation.

3.5 InterText: Modeling Text as a Living Object in Cross-Document Context

Iryna Gurevych (Technical University Darmstadt, DE)

License  Creative Commons BY 4.0 International license
© Iryna Gurevych

The ability to find and interpret cross-document relations is crucial in many fields of human activity, from social media to collaborative writing. While natural language processing has made tremendous progress in extracting information from single texts, a general NLP framework for modeling interconnected texts including their versions and related documents is missing. The talk reported on our ongoing efforts to establish such a framework. We addressed several challenges related to this. First, NLP has an acute need for diverse data to model cross-document tasks. We discussed our new, ethically sound data acquisition strategies and present unique cross-document datasets in the scientific domain, along with a generic data model that can capture text structure and cross-document relations in heterogeneous documents. Second, we reported on a study that instantiates our framework in the domain of scientific peer reviews. Finally, we highlighted our vision for cross-document computational argument analysis instantiating the InterText framework for analyzing arguments across documents. Our results pave the way to move NLP forward towards more human-like interpretation of text in the context of other texts.

3.6 Formalizing and Generating the Structure of Scholarly Papers

Eduard Hovy (Carnegie Mellon University – Pittsburgh, USA & University of Melbourne, AU)

License © Creative Commons BY 4.0 International license
© Eduard Hovy

As robust single-sentence generation in response to a prompt is more or less a solved issue now, and the controlled production of a coherent longer text is very much under investigation, one can wonder: what would it take to automatically generate a scholarly paper? In this talk I described (1) the representation in a structured form of the scholarly content; (2) the genre-oriented information required in scholarly discourse; (3) how to compose the first kind of information with the second using a typical modern neural network approach to argumentation structure. Topic (1) describes frameworks that can serve as templates for scholarly information; topic (2) outlines some rhetorical functions that information must be cast into to produce the appropriately structured scholarly genre; and topic (3) surveys various approaches and architectures to perform the requisite text planning.

3.7 Towards Automatic Interpretation of A Fortiori Arguments

Simone Teufel (University of Cambridge, UK)

License © Creative Commons BY 4.0 International license
© Simone Teufel

In this talk, I reported on work by my PhD student Olesya Razuvaevskaya. Her starting point was the restoration of premises in mini-arguments, whereby we wanted the generated premise to be guaranteed to be logically valid, as well as objectively explainable. We concentrated on the phenomenon of A fortiori logic, a logically valid reasoning pattern that has been known since ancient times and that is very frequent in day-to-day language use. Starting from sentences containing the phrase “let alone”, our analysis uses the fact that two situations are described and compared in terms of their likelihood. This simple fixed structure allows us to isolate the underlying logic to a single principle per argument, with just a few parameters necessary for explaining each case. The cases we consider are a) two quantities are concerned; b) the difference in likelihood concerns specificity; c) one of the situations described is a precondition of the other, and d) some underlying resource not mentioned in the text is required to explain the difference in likelihood. The d) cases require deeper reasoning. I also described key points of Olesya’s implementation of a system for the automatic partial interpretation of a system for a fortiori interpretation. The implementation uses standard neural sequence analysers and masked and unmasked transformers to provide a modular, pipelined analysis of three core aspects of the analysis.

3.8 Using the Claim Framework to Inform Public Policy

Ryan Wang (*University of Illinois Urbana-Champaign – Urbana, USA*)

License  Creative Commons BY 4.0 International license
© Ryan Wang

In this talk, I discussed the Claim Framework [1] and its application in evidence-based policymaking. The Claim Framework is concerned with the identification and representation of five kinds of scientific claims: the explicit claim, the implicit claim, observation, correlation, and comparison. An explicit claim consists of two entities connected by a relationship term that indicates a change observed in the experiment. An implicit claim similarly has two entities but the relationship between those is expressed in a more implicit manner. Unlike explicit and implicit claims, an observation identifies a change and the entity impacted by the change while leaving out the entity that causes the change. A claim that describes a correlation between two entities is a correlation. Finally, a comparison is a comparative construction where two entities are compared on a common ground. Taken together, the Claim Framework offers a principled means of extracting and organizing scientific claims that can be of great value to policymakers. [2] provides an example that uses the Claim Framework to automatically extract supporting, neutral, and refuting evidence of cell death and proliferation from biomedical abstracts with the aim of accelerating the otherwise time-consuming process of chemical risk assessment.

References

- 1 Catherine Blake. 2010. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *J. Biomed. Inform.* 43, 2 (April 2010), 173–189. <https://doi.org/10.1016/j.jbi.2009.11.001>
- 2 Catherine Blake and Jodi A. Flaws. 2021. Using semantics to scale up evidence-based chemical risk-assessments. *PLoS ONE* 16, 12, Article e0260712 (15 Dec. 2021), 24 pages. <https://doi.org/10.1371/journal.pone.0260712>

4 Flash Talks

4.1 Narrative Structures in Scientific Documents

Wolf-Tilo Balke (*TU Braunschweig, DE*)

License  Creative Commons BY 4.0 International license
© Wolf-Tilo Balke

From early on, narratives have been used as an essential means to convey information and knowledge in a form that is close to human communication and sense making. Moreover, references to archetypical narratives, such as David vs. Goliath, can also transport a set of connotations beyond the actual story allowing for a framing of information in the sense of speech acts. Facing today's flood of data and scientific results, data-driven narratives are thus an ideal way to make complex topics comprehensible, to make sense of certain events, or to assess the plausibility of given narratives or lines of arguments. However, these features are rarely used in information systems today. In particular, most of the current work on narratives is limited to representing structural properties such as story or plot graphs/plot units, event chains, or representations of entities and events without exploiting the deeper meaning of narratives. We explore narratives in the sense of logical overlays over

heterogeneous knowledge repositories, such as knowledge graphs, linked open data sources, document collections, or even concrete datasets. In its simplest form, a narrative then is a directed graph consisting of entities, events, and literals as nodes. Narrative edges describe the flow of the modeled events, i.e. on the one hand the semantic interaction between events and entities and on the other hand the respective types of interaction by suitable edge labels (e.g., in the causal or temporal sense). Essential for the expressive power of this overlay model is that edges of a narrative must always be bound against underlying knowledge repositories. In particular, this allows the plausibility of each edge to be evaluated against a given set of trusted repositories. Of course, this also means that the information in the underlying repositories needs to be carefully extracted with respect to classical dimensions of data quality, such as correctness, completeness, or validity.

4.2 Argumentation in Biochemistry Articles

Robert Mercer (University of Western Ontario – London, CA)

License © Creative Commons BY 4.0 International license
© Robert Mercer

Joint work of Eli Moser, Robert Mercer

Main reference Eli Moser, Robert E. Mercer: “Use of Claim Graphing and Argumentation Schemes in Biomedical Literature: A Manual Approach to Analysis”, in Proc. of the 7th Workshop on Argument Mining, pp. 88–99, Association for Computational Linguistics, 2020.

URL <https://aclanthology.org/2020.argmining-1.10>

This talk presented our contributions to argumentation in the experimental life sciences, scholarly biochemistry articles, in particular. Biomedical articles found in PubMed divide naturally into two classes: clinical and experimental. In the experimental class two types of articles have been or are being studied: genetics and biochemistry. With evidence from five biochemistry articles, the argumentation schemes that Green [2] has proposed for genetics articles transfer to biochemistry. We have studied the argumentation graphs that can be produced from the premises and claims in these articles and suggest an argumentation scheme hierarchy that is found therein. Biochemistry articles are structured in the IMRaD style (Introduction, Methods, Results, and Discussion). In work that is complementary to the well-known Argumentation Zoning model [4], Kanoksilapathum [3] has proposed rhetorical moves for each of these four sections. Providing computational models to identify these moves is ongoing work. In addition to the argumentation structure that exists in the main body of an article, titles with finite verbs strongly indicate the main claim of the article [1]. And structured abstracts provide similarly organized summaries of each of the four IMRaD sections. Work proceeds to connect Rhetorical Structure Theory to the argumentation schemes found in scholarly biochemistry articles with the ultimate goal of automating the identification of the schemes.

References

- 1 Heather Graves, Roger Graves, Robert E. Mercer, and Mahzereen Akter. 2014. Titles that announce argumentative claims in biomedical research articles. In *Proceedings of the First Workshop on Argumentation Mining*, pages 98–99.
- 2 Nancy Green. 2015. Identifying argumentation schemes in genetics research articles. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 12–21.
- 3 Budsaba Kanoksilapatham. 2005. Rhetorical structure of biochemistry research articles. *English for Specific Purposes*, 24(3):269–292.
- 4 Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

4.3 Linking Computational Argumentation to Information Quality

Davide Ceolin (Centrum Wiskunde & Informatica, Amsterdam, NL)

License  Creative Commons BY 4.0 International license
© Davide Ceolin

The logical and argument structure of information items can be an indicator of their information quality. In this talk, we presented a transparent pipeline to automatically mine and reason on arguments from information items [1, 2]. We evaluate how such argument-based analyses reflect on information quality by comparing argument-based assessments with quality assessments considering diverse aspects of quality (e.g., veracity, precision, completeness). The pipeline we propose combines diverse components based on machine learning, symbolic reasoning, and human computation. We evaluate the impact of diverse implementations of these components and test the pipeline on a dataset of product reviews. We plan to extend this pipeline to analyze scholarly documents in the future.

References

- 1 Davide Ceolin, Giuseppe Primiero, Jan Wielemaker, and Michael Soprano. 2021. Assessing the Quality of Online Reviews Using Formal Argumentation Theory. In *Web Engineering: 21st International Conference, ICWE 2021, Biarritz, France, May 18–21, 2021, Proceedings*. Springer-Verlag, Berlin, Heidelberg, 71–87. https://doi.org/10.1007/978-3-030-74296-6_6
- 2 Ceolin, D, Primiero, G, Soprano, M, & Wielemaker, J. (2022). Transparent assessment of information quality of online reviews using formal argumentation theory. *Information Systems*, 110, 102107.1–102107.14. doi:10.1016/j.is.2022.102107

4.4 Building Computational Models to Understand Scholarly Documents

Yufang Hou (IBM Research Europe – Dublin, IE)

License  Creative Commons BY 4.0 International license
© Yufang Hou

The accumulated scientific knowledge is the foundation upon which informed decision making is built, with huge impact across a wide range of critical applications. In this talk, I gave a short overview of my recent work on information extraction and natural language generation on scholarly documents, including interactive document2slides generation [1], scientific leaderboards construction [2], NLP TDM knowledge graph construction [3, 4]. Finally, I talked about our recent work on diachronic analysis of the NLP research areas, in which we developed a model to analyse NLP research areas and answer the following questions: (1) What is the general trend of a research area? (2) How is a research area influenced by other research concepts? (3) How do researchers argue about a specific research concept?


References

- 1 Edward Sun, Yufang Hou, Dakuo Wang, Yunfeng Zhang and Nancy X.R. Wang. D2S: Automated Slide Generation With Query-based Text Summarization From Documents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2021)*, Online, 6–11 June 2021

- 2 Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, Debasis Ganguly. Identification of Tasks, Datasets, Evaluation Metrics, and Numeric Scores for Scientific Leaderboards Construction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), Florence, Italy, 27 July-2 August 2019
- 3 Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, Debasis Ganguly. TDMSci: A Specialized Corpus for Scientific Literature Entity Tagging of Tasks Datasets and Metrics. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021), Online, 19-23 April 2021
- 4 Ishani Mondal, Yufang Hou, Charles Jochim. End-to-End Construction of NLP Knowledge Graph. In Proceedings of Findings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021 Findings), Online, 1-6 August 2021 Association for Computational Linguistics.

4.5 PEER – Collaborative Lightweight Argument Annotation

Nils Dycke (Technical University Darmstadt, DE)

License  Creative Commons BY 4.0 International license
© Nils Dycke

In this talk we introduced PEER, a collaborative, light-weight annotation tool for scholarly documents. The wide range of commercial tools for highlighting and commenting in PDFs (e.g. Google Docs, hypothesis, ...) cannot be used for scientific annotation studies at scale: they require uploading of confidential and potentially sensitive research data to public servers, and offer no mechanisms to manage, export or import annotation data. On the other hand, classical data annotation tools from the NLP community (e.g. Inception) require significant effort to set up for scholarly documents and, while being very feature-rich, they can be overwhelming to non-experts. To close this gap, we propose the PDF-annotation tool PEER, which unites the ease-of-use of highlighting and commenting software with the ease-of-access to NLP researchers of classical annotation tools. PEER offers a test bed for rapid prototyping different span annotation schemata and a lean study management interface. Users engage in their habitual process of highlighting and commenting in the annotation interface without the need for extensive familiarization with the tool. PEER comes as a ready-to-use web application and can be set up on local servers quickly. Hereby, we contribute towards the creation of new annotated datasets in the scholarly argumentation research.

4.6 Towards Constructive Conversations

Andreas Vlachos (University of Cambridge, UK)

License  Creative Commons BY 4.0 International license
© Andreas Vlachos

Joint work of Christine De Kock, Youmna Farag, Georgi Karadzhov, Tom Stafford, Andreas Vlachos

In this talk I presented our work motivated by the question “What makes conversations among humans more constructive and how can we intervene to make them happen”. First, I discussed group decision-making in the context of the Wason Card Selection task [1], where we find that groups perform better than individuals, and, more interestingly, can reach a correct decision even if no one had it in the beginning of the conversation [2]. Following

this, I presented the Wikipedia disputes dataset [3] which has allowed us to examine how disagreements are resolved in the context of Wikipedia, the most successful large-scale collaborative project. Finally, I described our work on developing and evaluating a dialogue agent for exposing people to the opposing side of an argument [4].

References

- 1 Peter C Wason. 1968. Reasoning about a rule. *Quarterly journal of experimental psychology*, 20(3):273281.
- 2 Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2022. DeliData: A dataset for deliberation in multi-party problem solving. <https://arxiv.org/abs/2108.05271>
- 3 Christine De Kock and Andreas Vlachos. 2021. I Beg to Differ: A study of constructive disagreement in online conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5601–5613, Dublin, Ireland. Association for Computational Linguistics.
- 4 Farag, Youmna; Brand, Charlotte; Amidei, Jacopo; Piwek, Paul; Stafford, Tom; Stoyanchev, Svetlana and Vlachos, Andreas (2022). Opening up Minds with Argumentative Dialogues. In: *Findings of EMNLP (Empirical Methods in Natural Language Processing)*

4.7 Expressing High-Level Scientific Claims with Formal Semantics

Davide Ceolin (Centrum Wiskunde & Informatica, Amsterdam, NL)

License  Creative Commons BY 4.0 International license
© Davide Ceolin


In this talk, we presented a method to express the content of high-level scientific claims using formal semantics in a systematic way [1]. Leveraging existing semantic formalisms, we developed the concept of “superpattern”, i.e., a formal representation of scientific claims corresponding to a conditional probability over logical formulas. Through this formalism, we can enable a full machine-understandable representation of scientific claims. The effectiveness of superpatterns has been evaluated both by effectively representing multiple claims from diverse scientific outlets, and by performing a user study that shows a high level of agreement among experts employing this technique.

References

- 1 Bucur, C-I, Kuhn, T, Ceolin, D & Ossenbruggen, JV 2021, “Expressing High-Level Scientific Claims with Formal Semantics”, arXiv, pp. 233-240. <https://doi.org/10.1145/3460210.3493561>

4.8 Argumentation, Persuasion, Propaganda, and More

Preslav Nakov (Mohamed bin Zayed University of Artificial Intelligence – Abu Dhabi, AE)

License  Creative Commons BY 4.0 International license
© Preslav Nakov

We described the connection between argumentation, persuasion, and propaganda: what their goals are and what techniques they use. We presented a specific inventory of propaganda techniques and we show that they do appear in scholarly articles. We further discussed framing as well as the role of figures and citances in scholarly articles, esp. in the life sciences. Finally, we discussed ways to use text summarization techniques with the aim of producing a layman’s summary of a scholarly article.

4.9 Fallacies in Political Argumentation

Serena Villata (Université Côte d'Azur, CNRS, Inria, I3S, France)

License © Creative Commons BY 4.0 International license
© Serena Villata

Joint work of Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, Serena Villata
Main reference Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, Serena Villata:
“Fallacious Argument Classification in Political Debates”, in Proc. of the Thirty-First International
Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022,
pp. 4143–4149, ijcai.org, 2022.
URL <https://doi.org/10.24963/ijcai.2022/575>

First, I presented a novel annotated resource of 31 political debates from the U.S. Presidential Campaigns, where we annotated six main categories of fallacious arguments (i.e., ad hominem, appeal to authority, appeal to emotion, false cause, slogan, slippery slope) leading to 1628 annotated fallacious arguments. Second, I introduced this novel task of fallacious argument classification and I presented the neural architecture based on transformers we proposed. Our results show the important role played by argument components and relations in this task.

4.10 Communicating Scientific Work with the Public through Dialogue Initiative

Milad Alshomary (Leibniz Universität Hannover, DE) and Smaranda Muresan (Columbia University – New York City, USA)

License © Creative Commons BY 4.0 International license
© Milad Alshomary and Smaranda Muresan

The gap between the scientific community and the public is growing. AI hype and distrust in science have become challenging issues nowadays. In our work, we aim to bridge this gap by first encouraging authors of scientific works to communicate their work to the public (e.g., journalists, non-experts, etc.). Instead of producing lay summaries, we hypothesize that the best form of communication is through dialogues, giving a space for both the authors and the public to construct an explanation and understanding of the subject matter jointly. Second, by studying the dialogical communication between these two parties, we can potentially provide assistant tools that can help authors sharpen their communication skills and (semi) automate the process of explaining scientific work to the public.

4.11 BAM: Benchmarking Argument Mining on Scientific Documents

Florian Ruosch (Universität Zürich, CH)

License © Creative Commons BY 4.0 International license
© Florian Ruosch

Joint work of Florian Ruosch, Cristina Sarasua, Abraham Bernstein

Main reference Florian Ruosch, Cristina Sarasua, Abraham Bernstein: “BAM: Benchmarking Argument Mining on Scientific Documents”, in Proc. of the Workshop on Scientific Document Understanding co-located with 36th AAAI Conference on Artificial Intelligence, SDU@AAAI 2022, Virtual Event, March 1, 2022, CEUR Workshop Proceedings, Vol. 3164, CEUR-WS.org, 2022.

URL <http://ceur-ws.org/Vol-3164/paper5.pdf>

I presented BAM, a unified Benchmark for Argument Mining (AM): a method to homogenize both the evaluation process and the data to provide a common view in order to ultimately produce comparable results. Built as a four stage and end-to-end pipeline, the benchmark allows for the direct inclusion of additional argument miners to be evaluated. First, the system pre-processes a ground truth set used both for training and testing. Then, the benchmark calculates a total of four measures to assess different aspects of the mining process. To showcase an initial implementation of our approach, the procedure is applied and evaluates a set of systems on a corpus of scientific publications. With the obtained comparable results, we can homogeneously assess the current state of AM in this domain.

5 Working Groups

5.1 Foundations of Scholarly Argumentation

Elena Cabrio (Université Côte d’Azur – Sophia Antipolis, FR)

Graeme Hirst (University of Toronto, CA)

Eduard Hovy (Carnegie Mellon University – Pittsburgh & University of Melbourne, AU)

Maria Liakata (The Alan Turing Institute – London, UK & Queen Mary University of London, UK)

Robert Mercer (University of Western Ontario, London, CA)

Smaranda Muresan (Columbia University – New York City, USA)

Preslav Nakov (Mohamed bin Zayed University of Artificial Intelligence – Abu Dhabi, AE)

Chris Reed (leader) (University of Dundee, UK)

Florian Ruosch (Universität Zürich, CH)

Simone Teufel (University of Cambridge, UK)

Serena Villata (Université Côte d’Azur – Sophia Antipolis, FR)

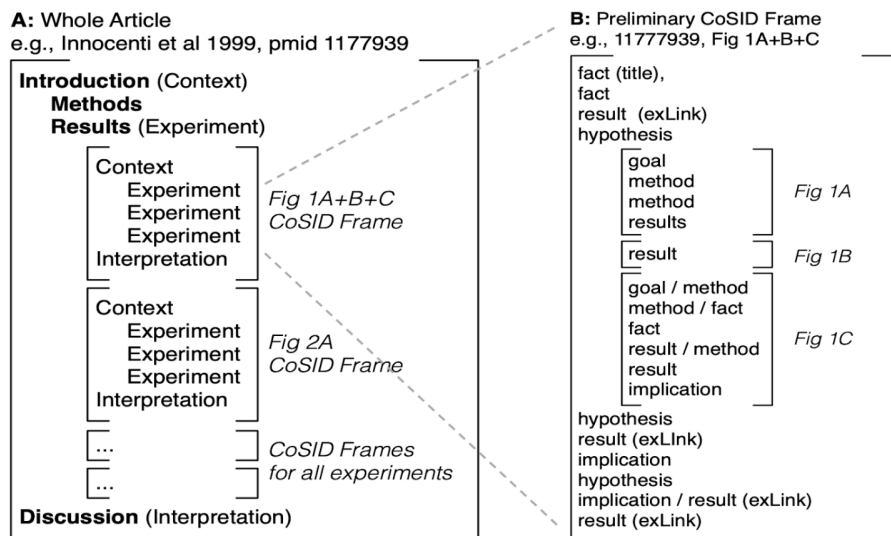
License © Creative Commons BY 4.0 International license

© Elena Cabrio, Graeme Hirst, Eduard Hovy, Maria Liakata, Robert Mercer, Smaranda Muresan, Preslav Nakov, Chris Reed, Florian Ruosch, Simone Teufel, Serena Villata

5.1.1 Introduction

We develop a framework to represent scholarly argumentation presented in research papers. We attempt to make the framework compatible with as much existing work as feasible and strive not to introduce novelties that still need to be defined, verified, and generally accepted.

We take the approach that there exist different “genres” of scholarly papers, such as *Experiment Report*, *Mathematical Proof*, and *Research Survey*, among others. Each paper genre has a characteristic stereotypical structure. For example, an *Experiment Report*



■ **Figure 1** Frames describing a set of Experiments from [3].

paper includes the description of an experiment containing a hypothesis, methods employed, measurement procedure, and measured results, while a *Mathematical Proof* paper includes a claim and its proof. Different disciplines tend to prefer different genres.

Regardless of genre, a scholarly paper is an artifact containing text, images, and possibly data or software in which the author makes an argument in support of one or more claims. In addition to the core claim, an argument includes text to support the claim and text to refute contradictory claims. The internal structure of arguments consists of text blocks of various types that recursively contain smaller blocks, ending with (approximately) a clause as the basic unit. Typically each block fulfills a discourse function, such as *introduce* or *prove*. Blocks are related to units or other blocks in various ways, for example through coreference among units.

In this section of the report, we describe the most common blocks and their composition into typical scholarly paper structures. The bulk of the section provides sets of labels that characterize the types of blocks and the types of relations that hold between them.

5.1.2 Frames

We represent the internal (sub)structure of argument blocks using frames. A frame is a list of smaller blocks, each supporting a specific discourse function within the larger block. The label names the block's function. For example, the *Experiment Report* is the frame (the sequence of blocks) Hypothesis + Experiment + Conclusion, where Experiment consists of the frame Method + Measurement and Conclusion consists of Interpretation + Claim. One elaboration of an *Experiment Report* frame for Biomedicine was developed in [3], see Figure 1.

5.1.3 Annotation Layers

To define frames, we have to define their building elements: the labels. Each textual unit in a frame carries one or several labels. The “smallest” textual unit, the simple proposition, is approximately a clause.

At the outset, we note that some units in papers refer solely to domain objects and actions, which exist in the world independently of the author’s beliefs or argumentation, while others, including claims, hypotheses, proofs, etc., include the author’s beliefs and are used by the author to build the argument. We call the former the *Domain World* and the latter the *Rhetorical World*. In general, propositions from the latter world reflect (explicitly or implicitly) some aspect of the author’s opinion (beliefs about factuality or attitudes about desirability), while propositions from the former have no such connotation. Of course, additional worlds of annotation exist, notably the Evaluation World to capture readers’ assessments of the argument in the paper. This Evaluation World is the focus of the Evaluation Group (see Section 5.3).

Typically, a clause has one or more labels from each world, plus perhaps linkages to units elsewhere in the paper. Corresponding to these worlds, we annotate a paper at two separate layers, each world providing its own set of labels and assigning additional information to a textual unit. Layer 1 (the narrative of the paper) is locutionary.¹ Layer 2 (the argumentation layer) is illocutionary. Hypotheses are mapped from the locutionary to the illocutionary layer. Evidence to support hypotheses stems from the locutionary layer and can consist of individual textual units representing observations, results, conclusions, and background claims.

► **Definition 1. Layer 1** (Domain World): The “semantic” layer that reflects the underlying domain information. Typical labels are *Domain Entity*, *Domain Relationship*, *Method*, or *Measurement*. The precise semantics of each label requires definition, and is probably going to differ in different annotation schemes.

► **Definition 2. Layer 2** (Rhetorical World): The “rhetorical” layer that reflects the argumentation of the paper. This necessarily includes the author in some way, for example as holder of an opinion or observer of some fact. Typical labels are *Claim*, *Hypothesis*, *Motivation*, *Purpose*, *Observation*, *Related Work*, *Experiment*, *Model*, *Background*, or *Conclusion*. The precise semantics of each label requires a definition, and is probably going to differ in different annotation schemes. Note that the same Layer 1 units can be rearranged into different arguments by different Layers 2.

5.1.4 An Example Labelset

We have in mind a modular, core annotation scheme that is domain-independent and can be extended as needed with domain specificities. Many people have worked on the components and functions of argumentation, from Aristotle [1] to Toulmin [9] and Walton et al. [10]. We do not propose a preferred set of labels as the “correct” one; we merely draw from previous work as an illustration. By adopting (some of) these labels and adding more as needed for any specific task, anyone using this framework would make their work available to others doing the same. Our labelset is drawn primarily from the following three sources.

The first source [6] includes eleven main categories, considered by the authors as the Core Scientific Concepts (CoreSC). They are listed in Table 1, including the distinction of a finer-grained classification that gives details about the properties of objects and methods mentioned in the paper.

¹ Following Austin [2], locutionary acts are our statements with their immediate and direct meanings. Illocutionary acts derive from the performance of our statements, like asserting, hypothesising, or performing. Perlocutionary acts affect the hearer indirectly after inference; for example, someone being persuaded or insulted.

■ **Table 1** Categories from the CoreSC Annotation scheme [6].

Category	Description
Hypothesis	A statement not yet confirmed rather than a factual statement
Motivation	The reasons behind an investigation
Background	Generally accepted background knowledge and previous work
Goal	A target state of the investigation where intended discoveries are made
Object-New	An entity which is a product or main theme of the investigation
Object-New-Advantage	Advantage of an object
Object-New-Disadvantage	Disadvantage of an object
Method-New	Means by which authors seek to achieve a goal of the investigation
Method-New-Advantage	Advantage of a Method
Method-New-Disadvantage	Disadvantage of a Method
Method-Old	A method mentioned pertaining to previous work
Method-Old-Advantage	Advantage of a Method
Method-Old-Disadvantage	Disadvantage of a Method
Experiment	An experimental method
Model	A statement about a theoretical model or framework
Observation	The data/phenomena recorded in an investigation
Result	Factual statements about the outputs of an investigation
Conclusion	Statements inferred from observations & results relating to research hypothesis

For the next source, we use the Argument Interchange Format (AIF) [8]. It is tailored towards modeling argumentation as a graph, but is not specific to scholarly papers. There are various types of nodes that represent ontological concepts:

- Information: The *I-node* contains the utterance (also called proposition), the minimal building block without any other rhetorical semantics.
- Anchor: The *YA-node* is about all speech acts, such as assert, hypothesize, or claim, among others. It links two *I-nodes* and represents the illocutionary forces.
- Applications of Rules of Inference or Conflict: The *RA-node* is used for connecting two *I-nodes* with inference, while the *CA-node* does the same but for conflict.
- Rephrase: The *MA-node* is for restating a proposition and includes purposes such as generalization, specification, or exemplification.
- Transition: The *TA-node* indicates the transition between two *I-nodes* and can coexist between the same two propositions parallel to another relation. They contain the dialogue relations (e.g., between a question and an answer).

Inspired by the work of Moser and Mercer [7] and Green [5], we find that the following labels are used in experimental science papers: *Premise*, *Inference*, and *Claim*. A list and taxonomy can be found in [4].

Even without formal definitions, the labelset overlaps and differences are obvious.

5.1.5 Definitions of Example Labels

This section lists a set of fairly generic accepted labels with definitions for each. Most exist in the above mentioned Rhetorical World (not the Domain World or the Evaluation World).

► **Definition 3. Assertion:** A simple proposition (typically a clause) that states something. In fact, every *Assertion* is a *Claim* since the implicit assumption (unless otherwise stated) is that the author believes the proposition (except perhaps for the awkward case of null hypotheses). However, we differentiate *Assertions*, which for us carry no implicit connotation that the author believes them to be true, from *Claims*, for which the author’s epistemic (truth) judgment must be given. Thus, we use a narrow interpretation of “*Claim*”. In AIF, this is called *I-Node*.

► **Definition 4. Claim:** A frame that consists of an author or a speaker (called claimer), *Claim* content (an *Assertion*), the epistemic status (which can be true, false, maybe, desired, unknown, . . .), and a set of links to support or opposition frames. In AIF, this corresponds to *YA-nodes*.

► **Definition 5. Support:** A link that connects two other frames. To be able to associate additional information with it, we reify the link and state it as a frame consisting of a *Claim*, which may even appear in another paper, and Evidence (a set of *Assertions* or *Claims*). This is called *RA-node* in AIF.

► **Definition 6. Oppose:** As *Support*, mutatis mutandis, and corresponds to *CA-node*.

► **Definition 7. Hypothesis:** A frame, which is almost identical to a *Claim* but whose epistemic status is unknown or desired. It usually appears without *Support* or *Oppose* links. In AIF, this is expressed using *YA-nodes*.

► **Definition 8. Motivation or Goal:** A frame that expresses the desired target state after an experiment has been executed consisting of a holder (a person with the goal, usually the author) and a desired state (usually a *Hypothesis*, but with its epistemic status being proved). This is included in the *YA-nodes*.

► **Definition 9. Step:** A single action (in the Domain World) performed on domain objects (from the Domain World). This corresponds to a clause and involves an actor (usually, someone from the author’s team). There are different kinds of *Step*, depending on the nature of the domain. Most experiments include a measurement (see *Assay* below), one or more observations, and one or more conclusions.

► **Definition 10. Method:** A frame of an ordered series of *Steps*.

► **Definition 11. Assay:** A frame (a more specific *Step*) consisting of an actor, a measurement (a *Method*), a metric (a measuring unit accepted in the Domain World), and a result (a number determined by the *Method* expressed in the *metric*).

► **Definition 12. Experiment:** A frame with a local *Hypothesis* (i.e., restricted to one aspect being studied), a *Method*, an *Assay*, and a result, which is a *Claim* frame whose epistemic status is proved.

► **Definition 13. Interpretation:** A frame that draws together several *Assays* into a single *Claim*. It is made up of *Experiments* (a list of *Assays*, or perhaps their *Experiments*) and conclusions, which are a set of *Claims* or *Hypotheses* with the epistemic status of proved.

► **Definition 14. Restatement:** A link that connects two other frames that have the “same” (semantic) meaning. To be able to associate additional information with it, we reify the link and state it as a frame consisting of Version 1 and Version 2 (of an *Assertion*). These are propositions, which may even appear in another paper. In AIF, this is represented by the *MA-nodes*.

► **Definition 15. (Research) Question:** A question about a proposition which is a frame consisting of a questioner (a person, usually the author) and the focus of a question (a proposition). This is included in AIF's *YA-nodes*.

► **Definition 16. Dialogue Relation:** A link that connects two other frames and expresses a dialogue function, such as a question and an answer or a full form and a summary. Typically, *Dialogue Relations* coexist between two frames that are also related using another relation in parallel. To be able to associate additional information with it, we reify the link and state it as a frame made up of the dialogue prior (which is any frame, e.g., a *Question*) and the dialogue posterior (any frame, e.g., the proposition that answers it). In AIF, this is expressed using *TA-nodes*.

5.1.6 Next Steps

As mentioned in the outset, we do not propose a finalized framework of frames and sets of labels. But we hope that the frames and labels listed here may serve most purposes and encourage standardization across research. It is left for future work to flesh out both the labelset(s) and the relations.

Furthermore, the development of a typology of paper genres is necessary in order to apply the framework. At the minimum, the different structures used in the genres Experiment Reports, Mathematical Proofs, and Surveys should be elaborated.

The proposed framework of genres, frames, and labelsets will be best tested by the creation of example annotation datasets.

The other three sections of this report are compatible with the framework proposed here. Section 5.3 on Evaluation develops an additional layer of labels.

References

- 1 Aristotle (1954). *The Rhetoric and the Poetics of Aristotle (translated by W. Rhys Roberts)*. Random House, New York.
- 2 Austin, J. (1975). *How to Do Things with Words*. Harvard University Press.
- 3 Burns, G. A. P. C., de Waard, A., Dasigi, P., and Hovy, E. H. (2016). Cycles of scientific investigation in discourse – machine reading methods for the primary research contributions of a paper. In Jaiswal, P., Hoehndorf, R., Arighi, C. N., and Meier, A., editors, *Proceedings of the Joint International Conference on Biological Ontology and BioCreative, Corvallis, Oregon, United States, August 1-4, 2016*, volume 1747 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- 4 Green, N. L. (2018a). Proposed method for annotation of scientific arguments in terms of semantic relations and argument schemes. In Slonim, N. and Aharonov, R., editors, *Proceedings of the 5th Workshop on Argument Mining, ArgMining@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 105–110. Association for Computational Linguistics.
- 5 Green, N. L. (2018b). Towards mining scientific discourse using argumentation schemes. *Argument Comput.*, 9(2):121–135.
- 6 Liakata, M., Saha, S., Dobnik, S., Batchelor, C. R., and Rebholz-Schuhmann, D. (2012). Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinform.*, 28(7):991–1000.
- 7 Moser, E. and Mercer, R. E. (2020). Use of claim graphing and argumentation schemes in biomedical literature: A manual approach to analysis. In *Proceedings of the 7th Workshop on Argument Mining*, pages 88–99.
- 8 Rahwan, I. and Reed, C. (2009). The argument interchange format. In Simari, G. R. and Rahwan, I., editors, *Argumentation in Artificial Intelligence*, pages 383–402. Springer.

- 9 Toulmin, S. E. (2008). *The Uses of Argument, Updated Edition*. Cambridge University Press.
- 10 Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation Schemes*. Cambridge University Press.

5.2 Cross-domain Argumentation Model for Scholarly Argumentation

Khalid Al-Khatib (University of Groningen, NL)


Fengyu Cai (TU Darmstadt, DE)

Dayne Freitag (Artificial Intelligence Center, SRI International – Menlo Park, USA)

Daniel Garijo (Universidad Politécnica de Madrid, ES)

Benno Stein (Bauhaus-Universität Weimar, DE)

Henning Wachsmuth (Leibniz Universität Hannover, DE)

License  Creative Commons BY 4.0 International license

© Khalid Al-Khatib, Fengyu Cai, Dayne Freitag, Daniel Garijo, Benno Stein, Henning Wachsmuth

Although all scholarly discourse shares a common set of goals that can be easily articulated – the increase of human knowledge, the achievement of consensus among scholars, etc. – it encompasses a huge variety of disciplines and objectives. It is not immediately clear that a model developed to explain argumentation in one domain, like computational linguistics, can be applied to the scholarly literature on seismology or clinical psychology. A unified view of scholarly argumentation is clearly desirable, potentially increasing the speed with which new scholarly domains can be modeled computationally. The *Domains* working group sought to investigate the feasibility of such a universal framework. Rather than approaching this question based on first principles, as in the *Foundations* working group, we adopted a comparative approach, anchoring our inquiry in a close reading of two papers from widely different domains.

5.2.1 Objective

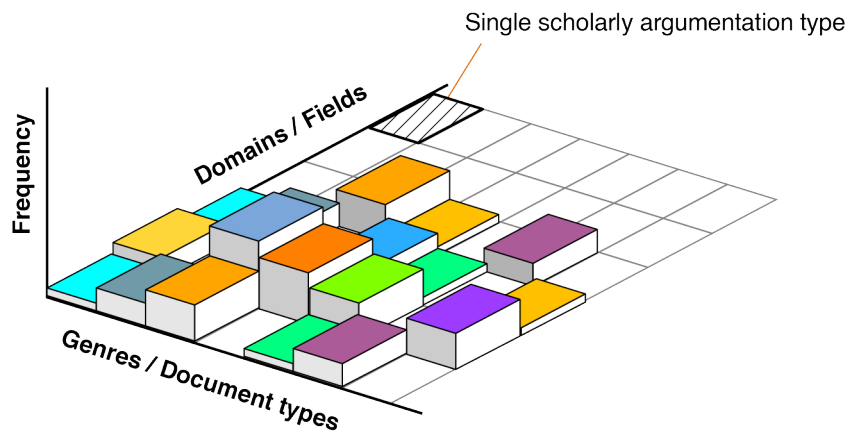
The basic goal of scholarly argumentation is to add knowledge to the existing knowledge of a domain/field.² The way this knowledge is added (the kind of scholarly argumentation) follows the specific rules and traditions of the field. The definition of these specific rules and traditions is what we refer to as an argumentation type.

We attempt to identify different scholarly argumentation types, organized based on the domain (i.e., area of expertise) and genre (i.e., document type) of the scientific publication. Figure 2 illustrates the idea of prevalent argumentation types across different domains and different document types.

If we can acquire conceptual and empirical knowledge about the distribution illustrated in Figure 2, we will identify usage patterns across domains, and at least partially, decouple topics, domains, and argumentation types. This orthogonality can justifiably be seen as the identification of argumentation strategies.

Starting from an anecdotal study with the review of two different publications, we discuss the main gaps when annotating argumentative sentences in scientific papers.

² In addition to this primary purpose, we recognize that papers are also written self-expression, career reasons, or other reasons.



Domains / Fields (exemplary, from DFG scheme) :

- Humanities
- Sport
- Law, Economics, and Social Sciences
- Mathematics, Natural Sciences
- Human Medicine / Health Sciences
- Agriculture, Forestry and Nutritional Science, Veterinary Medicine
- Engineering
- Art, Art Theory

Genres / Document types (exemplary) :

- Blog posts
- Debates
- Essays
- Law texts
- News
- Political speeches
- Reviews
- Scientific articles
- Wikipedia

■ **Figure 2** Each cell corresponds to a single argumentation type, where same/similar colors hint same/similar types. There are argumentation types that are used across all domains and genres, but also domain- and genre-specific types.

5.2.2 Anecdotal Study

Argumentation in scientific articles may be modeled at different levels of granularity, from the macro-level discourse structure of an entire article (e.g., in terms of elements such as model, experiments, and discussion) to the micro-level argumentative structures of individual clauses, sentences, and paragraphs. As an initial basic study, all members of the breakout group annotated the sentence-level structure of the introductions of two scientific articles from different disciplines, identifying which sentences comprise the *claims* and the *premises* of the authors' arguments. Here, we considered a claim to be an assertion that the authors aim to sell as new, true, or similar, and a premise as a reason supporting either the claim directly or another premise.

In particular, we considered one paper each from two domains reflecting two different types of papers, namely, a corpus paper from computational linguistics [1] and an experiment paper from medical chemistry [2]. For each paper, we first annotated its introduction individually, and then we compared and discussed the results. Here, we report only on some noteworthy findings that we made.

First, we observed that these two texts share a similar argumentative agenda and structure, despite the wide divergence in subject matter and lexical content. As shown in Table 2, this rough similarity can be exposed by comparing key sentences drawn from different locations in a paper's introductory section. These sentences exhibit intents – which we characterized

■ **Table 2** The argumentative agendas and structures of two papers, one from computational linguistics (a corpus paper) and one from medical chemistry (an experimental paper).

Location (<i>Audience</i>)	Computational Linguistics	Medical Chemistry
title (<i>reviewer</i>)	“Mama Always Had a Way of Explaining Things So I Could Understand”: A Dialogue Corpus for Learning to Construct Explanations	Protein-Structure Assisted Optimization of 4,5-Dihydropyrimidine-6-Carboxamide Inhibitors of Influenza Virus Endonuclease
lead sentence (<i>sponsor</i>)	Explaining is one of the most pervasive communicative processes in everyday life....	Influenza is an infectious disease associated with 500,000 deaths and 3–4 million severe illnesses annually....
main claim (<i>lead researcher</i>)	We argue that a better understanding of how humans explain in dialogues is needed, so that XAI can learn to interact with humans.	Our overarching approach has been to apply structure-based design, while optimizing inhibitors. . . in order to proactively develop lead inhibitors that are less likely to rapidly develop clinical resistance.
proximal claim (<i>junior researcher</i>)	In this paper, we present a first corpus for computational research on....	Here, we describe the further optimization of such a series of new endonuclease inhibitors....

in terms of putatively different audiences – that vary with their position in the discourse and are shared across the two target domains. Titles must succinctly summarize a paper’s content and, depending on the domain, may include features intended to draw interest from potential reviewers. Lead sentences typically state the overriding concern an entire field addresses, often in a language digestible by a general audience. Sentences expressing claims vary in their specificity and concreteness, ranging from concrete contribution to central insight.

As shown in the table, we distinguished between *main claims* (claims that the paper’s author presumably deemed most important) and *proximal claims* (*pro forma* claims that provide useful context). We found that the ability to distinguish these two types of claims relies substantially on domain expertise. For both domains, we observed that the introduction contains only very few real claims in the sense of assertions the authors aim to convince the reader of – about one to three depending on the annotator.

Initially, there was notable disagreement in the group, none of whom has extensive chemistry expertise, about which statements in the medical chemistry paper constituted claims and which of these was the main claim. In contrast, the group’s annotations of the computational linguistics paper showed considerable agreement. The only exception was the annotations of a group member with less background in computational linguistics. This member chose as the main claim a sentence that all other members viewed as proximal.

This result clearly established the importance of domain expertise for certain types of argumentative analysis. In particular, determining which is the *main* claim requires the reader to assess the scientific significance of a statement, an assessment that may require extensive knowledge of an area of research. However, based on our interdisciplinary discussion, we reached an agreement in most cases, even in our analysis of the chemistry paper, suggesting that the automated modeling of the argumentative structure of scientific articles is feasible, in principle. The key question is how much domain knowledge the analysis of scientific argumentation in a given discipline requires.

5.2.3 A New Approach: Towards Reducing Reliance on Domain Experts

Our anecdotal study and discussion suggest that having domain experts for all paper types and domains may be a costly and inefficient process. Instead, we identified a new research challenge: how to accelerate the interaction with domain experts to speed up cross-domain argumentation annotation? This research challenge spans new research questions such as:

- Can we probe experts with specific portions of text instead of having them read the whole publication?
- Can we identify a specific vocabulary and use it in customized domain-specific annotation platforms?
- Can we identify a set of questions for domain experts to help guide other users in the annotation process? Examples of these questions include identifying the section where the main claim is, which are the main sections to look at first when analyzing a paper in a particular domain, what are the main types of evidence in a publication or typical lexical cues to identify claims or evidence in a given domain.

Following our anecdotal study, we explored some of these questions with our two papers, as shown in Tables 3. We believe these are initial examples that should be expanded in order to identify a wider range of commonalities in scientific literature.

■ **Table 3** Examples of three questions that domain experts can answer to assist non-experts in the annotation process.

<i>Where is the main claim? in which section can we find it?</i>					
Publication domain	Introduction	Method	Experiments	Results	Discussion
NLP	X	X			
Chemistry	X			X	
<i>What are the main sections that we should look at first?</i>					
Publication domain	Introduction	Method	Experiments	Results	Discussion
NLP	1	2	4	3	5
Chemistry	1	5	4	2	3
<i>What are the main types of evidence in your domain?</i>					
Publication domain	Anecdote	Statistics	Testimony	Analogy	Figure/table
NLP		X			X
Chemistry		X		X	

5.2.4 Next Steps

Our working group discussed the *Introduction* section of two conference papers from computational linguistics and medical chemistry as examples to explore the discrepancy in argumentation between domains. Through manual annotation and discussion, we came to find that scientific argumentation varies among different domains noticeably. Further work may extend this analytical and comparative paradigm on scholarly argumentation to other domains, genres, and parts in the publication. After discussing the results of our anecdotal study, we believe that more research is needed to accelerate domain expert interaction for annotating argumentative sentences in different domains. Instead of asking domain experts to directly help with various task-specific works required by non-experts, their contribution would be more efficient and influential by helping summarize the universal features of argumentation for one specific genre, domain, and part. For example, experts could annotate a feasible scale of representative papers, extract lexical hints, etc.

Key questions in this area include how to structure the interaction with the domain expert for lightweight knowledge elicitation, and how to abstract, represent, and inject the features that encapsulate the knowledge required for accurate models of a given domain’s argumentation. Meanwhile, without sacrificing models’ performance, the minimum degree of domain knowledge elicitation from experts is also worth studying.

References

- 1 Henning Wachsmuth and Milad Alshomary. “mama always had a way of explaining things so I could understand”: A dialogue corpus for learning to construct explanations. In Proceedings of the 29th International Conference on Computational Linguistics, pages 344–354, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.
- 2 Diane Beylkin, Gyanendra Kumar, Wei Zhou, Jaehyeon Park, Trushar Jeevan, Chandraiah Lagiseti, Rhodri Harfoot, Richard J Webby, Stephen W White, and Thomas R Webb. Protein-structure assisted optimization of 4, 5-dihydroxypyrimidine-6-carboxamide inhibitors of influenza virus endonuclease. *Scientific reports*, 7(1):1–12, 2017.

5.3 Evaluation of Argument Quality

Yufang Hou (IBM Research Europe – Dublin, IE)

Tobias Mayer (Technical University Darmstadt – Darmstadt, DE)

Domenic Rosati (scite.ai – Halifax, CA)

Harrisen Scells (Leipzig Universität DE)

Ferdinand Schlatt (Universität Halle-Wittenberg – Halle, DE)

Simone Teufel (University of Cambridge, UK)

Ryan Wang (University of Illinois Urbana-Champaign – Urbana, USA)

License © Creative Commons BY 4.0 International license

© Yufang Hou, Tobias Mayer, Domenic Rosati, Harrisen Scells, Ferdinand Schlatt, Simone Teufel, Ryan Wang

5.3.1 Introduction

This working group focused on the evaluation of argumentation quality. We wanted to take an alternative approach to typical text synthesis evaluation. We also wanted to develop an evaluation framework that is general enough that can be applied to the numerous argumentation schemes that exist.

We settled on an approach that would determine the argumentation quality of a text through the interrogation by a Question Answering (QA) system about the argumentation within. For each genre of text, one would need to develop a series of diagnostic (and increasingly specific) questions that would reflect the quality of the reasoning or argumentation of the paper. Each genre of text would have a different set of characteristics. For example, the process for systematic reviews would ask questions about the comprehensiveness of the review, while for technical or experimental papers, questions would be related to the description of the model, and how or why it brings about an improvement in the domain.

The envisioned evaluation system would take a text to be evaluated and a question bank organised by the different genres. Depending on the genre the appropriate questions would be applied to the text and their answers rated. Note that the framework is flexible such that answers could be rated by humans and later automatically. Depending on the effectiveness of the QA system and the nature of the answers, a combination of human and automatic answer rating can be used.

■ **Table 4** Example questions that could be posed to a QA system, and descriptions of answers.

Example Question	Description of Answer
What is the typology of the paper?	E.g., empirical research, position paper, theoretical.
Which questions apply to which genres?	E.g., empirical science may be more focused on cogency while philosophy and mathematics may be more focused on reasonableness.
What are the general properties of argumentation of interest in a specific paper genre, for making questions?	
What are the domain-specific properties of argumentation of interest?	E.g., empirical science may be more focused on ...? while philosophy and mathematics may be more focused on ...?
What is the main claim of this paper?	One or two sentences from the text that should contain the claim.
What is the proof that the paper's proposed technique is better?	Two rows extracted from a table, one for the state of the art and the other for the system, containing two numbers, the system's being better.

■ **Table 5** Taxonomy of question types and their explanations.

Question Type	Explanation
Document-level assessment vs. corpora-level assessment	Intra-document vs. inter-document
Extractive evaluation vs. reasoned evaluation	Extractive: questions that can be answered through passages in the text; Reasoned: questions that must be answered through reasoning
Content vs. Form	Content: how well are the arguments presented? Form: how well are the arguments structured?

5.3.2 Evaluation Framework

We developed several initial question banks for academic papers. In Table 4 we provide a sample of what we believe to be the kinds of questions that should be asked. However, how would one judge the answers? Ideally questions would be simple and easy to judge automatically. In reality the argument is nuanced and complex. Therefore, assessing the quality of answers is likely to be a human task, though as answers become more formalised (even simplified to just yes or no questions) automated assessment becomes more feasible.

We note that it seems natural that the assessor might want to record caveats, concerns, or other thoughts. We allow the assessor to include such comments as motivation for why they assign the score they do.

To guide the development of questions, we also devised a taxonomy of question types. Table 5 contains our initial taxonomy of question types. However, in addition to question types, it is also necessary to define how answer to questions will be evaluated.

Thus finally, we devised a hierarchy of evaluation. Each level corresponds to a different interrogation method for probing argumentation quality.

1. First level of evaluation: model evaluation as retrieval
 - Input: Open-ended questions
 - Output: “retrieval unit” i.e., sentence/snippet/etc.
2. Second level of evaluation: model evaluation as a checklist
 - Input: Yes/No questions
 - Output: Yes/No

3. Third level of evaluation: multiple-choice QA
 - Input: multiple-choice questions
 - Output: selection of one answer from set of answers

5.3.3 Next Steps

We have already begun the development of a tool for the community to perform offline evaluation of argumentation quality. We are developing the tool as an open source project, and is available at https://github.com/hscells/arg_eval. We plan to continue to develop this tool to support the various question types and levels of evaluation. Once we have laid the groundwork with a proper evaluation tool and expanded upon the framework proposed here, the next logical step is the development of a QA system. The first version of the QA system will focus on a small subset of the possible question types and perhaps only one level of evaluation. This will demonstrate the viability of a QA system to evaluation argumentation quality and will set a clear direction for further expansion of the QA system.

5.4 Scholarly Argumentation as a Community Dialogue

Wolf-Tilo Balke (TU Braunschweig, DE)

Andreas Vlachos (University of Cambridge, UK)

Davide Ceolin (Centrum Wiskunde & Informatica, Amsterdam, NL)

Milad Alshomary (Leibniz Universität Hannover, Germany))

Nils Dycke (TU Darmstadt, DE)

Sukannya Purkayastha (TU Darmstadt, DE)

Iryna Gurevych (TU Darmstadt, DE)

Anne Lauscher (Universität Hamburg, DE)

Tilman Beck (TU Darmstadt, DE)

License © Creative Commons BY 4.0 International license

© Wolf-Tilo Balke, Andreas Vlachos, Davide Ceolin, Milad Alshomary, Nils Dycke, Sukannya Purkayastha, Iryna Gurevych, Anne Lauscher, Tilman Beck

5.4.1 Motivation

In science, peer reviewing is the deliberation process where members of a scientific community with diverse levels of experience decide if a scholarly work provides a valuable, scientific contribution [1, 2]. In the process, the actors of the community (i.e. authors, reviewers, meta-reviewers, and possibly others, e.g., chairs) exchange arguments about the strengths and weaknesses of a particular scientific contribution within multiple, direct and indirect dialogues (review, rebuttal, decision-making).

Usually, the decision-making process begins with the reviewers writing their reviews and optionally the authors responding to the reviews (i.e. rebuttal). Here, the **meta-reviewer has to arrive at a decision about the promotion of acceptance of the paper**. This process is mainly about weighing the arguments raised by the reviewers and happens under time constraints. To provide more efficient and effective access to (a) the content of the paper, and (b) the many arguments raised by the individual reviewers, we envision an intelligent dialogue system which answers questions of the meta-reviewer.

From an NLP perspective, this is more challenging than other domain-specific task-oriented dialog system scenarios [9], as the meta-reviewer’s needs underpinning these questions can vary from information retrieval and exploration (e.g. “What datasets did the authors use?”)

to combining information from multiple sources (e.g. “According to the reviews, what are the main weaknesses of the paper?”) and summarization tasks (e.g. “Please briefly summarize the paper?”).

The goal of this breakout group was to evaluate the feasibility of collecting (training) data for such a system and to refine the task definition along the way. Having such a dataset could provide an interesting basis for studying both various facets of argumentation, like quality and convincingness of reviews or implicit ranking of the value system employed by the meta-reviewer. Furthermore, different dialogue strategies can be analyzed, like the way of gathering information in order to come to a decision.

5.4.2 Summary and Conclusions

The breakout group defined the goal of the sessions as **formulating the decision-making process for a scholarly paper as a dialogue** and conducted a first annotation round using an Oxford-style inspired debate format. Two groups (debaters and judges) were involved in the decision-making process. Given a paper and its reviews, the debaters discussed the pros and cons of the paper and a decision was formed by the judges. To study the relation between arguments extracted from the reviews or the underlying paper, all turns required explicit grounding in the respective documents. For instance, an argument in favor of acceptance should be substantiated by the review passage (*As reviewer 1 says . . .*) from which it was derived.

After reviewing the annotation process, it became clear that the task needs to be better aligned with the actual review process of the respective research discipline (in our example: Natural Language Processing) and in such a way that the data collected will be useful for a real-world system. There was a consensus that the debate format is obstructive as it forces dialogue partners to defend a position which might be different from their own. Additionally, the coarse granularity of groundings in natural dialogue – i.e. referring to the entire document instead of sentences or paragraphs – limited the study of the relations between argumentative units in the reviews and papers.

We revised the system’s purpose as a decision-making support system for the meta-reviewer after reviews (and rebuttals) are collected. Therefore, the dialogue involves two parties (meta-reviewer, intelligent support system) with the meta-reviewer questioning the system to inform their final decision, and the system as an oracle with knowledge of the paper, the reviews and optionally other related work. To resemble a real-world situation, a time limit is imposed on the meta-reviewer which enforces limited exposure to the reviews and paper. It is important to note that such a system will be most beneficial for papers where the decision is difficult (i.e. so-called *borderline papers*). Further, the system will support in weighing the reviews as there exist different levels of reviewing expertise.

Finally, we conducted another round of data collection by pairing the senior members of the group (meta-reviewers) with the junior members (imitating the dialogue system). The junior members prepared themselves by reading the papers and reviews in detail. Before the dialogue, meta-reviewers had five minutes to study the reviews. At the end of the dialogue, the meta-reviewer had to make a statement about the acceptance or rejection of the paper. We collected 16 dialogues (in English) about 4 papers involving 4 meta-reviewers and 4 system agents. The conversations were transcribed using the OpenAI Whisper [10] model which is known to have good transcription quality. However, manual post-processing was necessary as the model output is not separated based on the speaker.

In summary, we formalized the idea of a decision-support system for meta-reviewers during peer reviewing as a dialogue system. We designed and evaluated a protocol to collect dialogue data for such a system. As a result we created a dataset of 16 high-quality question & answer dialogues between a meta-reviewer and a system agent which is knowledgeable about the paper and reviews.

5.4.3 Challenges And Next Steps

There exist several open questions about the future course of this project. State-of-the-art NLP models require a certain amount of data for training. However, as the data collection procedure requires the participation of expert-level reviewers (i.e. meta-reviewers), it cannot be scaled easily. One way could be to align the discussion format between reviewers and meta-reviewer during peer-reviewing with the dialogue format proposed in this group. Similar to the first pilot annotation, the study of grounding of the assistants' turns in the review texts is one important future step towards modeling such alignments. This step might be facilitated by adding structured annotations to the reviews, indicating, for example, the targets of the comments (comments regarding specific parts of the paper, the experimental settings, etc.) or the severity of the issues raised by the reviewers [3].

Another question is whether the data collection procedure can be generalized to different use-cases. Here, the first step should be to separate task-specific components (e.g. meta-reviewer role) from the more general aspects. Further, an additional layer of annotation would help specify general and domain-specific dialogue acts. The usability of different dialogue argumentation schemata [4] needs to be assessed. Annotation can be conducted using off-the-shelf tools, like INCEPTION [5] or PEER³.

A crucial issue is the evaluation of the success of the conversation. As stated above, the goal of the system is to inform the meta-reviewer's final decision. This is rather difficult to quantify and can be biased by other influencing factors, e.g. low-quality reviews. Conducting user studies is a possible direction but it is costly and time-consuming. Another approach could be to assess whether individual questions have been answered satisfactorily by the system rather than directly evaluating the overall conversation. While we successfully transcribed the audio data collected in this group, we recommend data collection via text-based input methods [6] to overcome the need for post-hoc manual speaker identification and enabling data collection in an online setup. Also, we point out that neither the rebuttals nor the official meta-reviews are included in the peer review dataset [7] due to the complicated licensing situation with peer-reviewing data [8], but they would be another useful resource.

References

- 1 Tom Jefferson and Elizabeth Wager Frank Davidoff, *Measuring the quality of editorial peer review*. JAMA, American Medical Association, pp.2786–2790, 2002.
- 2 Aliaksandr Birukou and Joseph R. Wakeling and Claudio Bartolini and Fabio Casati and Maurizio Marchese and Katsiaryna Mirylenka and Nardine Osman and Azzurra Ragone and Carles Sierra and Aalam Wassef, *Alternatives to Peer Review: Novel Approaches for Research Evaluation*. Frontiers Computational Neuroscience, p.56, 2011.
- 3 Cristina-Iulia Bucur and Tobias Kuhn and Davide Ceolin, *Peer Reviewing Revisited: Assessing Research with Interlinked Semantic Comments*. K-CAP, pp.179–187, 2019.

³ The PEER tool is an annotation tool designed for the domain of scholarly articles permitting span annotations and commenting directly inside an article's PDF. This tool is under development at the UKP Lab, Technical University of Darmstadt; please refer to <https://intertext.ukp-lab.de/> for updates on its release

- 4 Georgi Karadzhov and Tom Stafford and Andreas Vlachos, *DeliData: A dataset for deliberation in multi-party problem solving*. arXiv preprint 2108.05271, 2021.
- 5 Jan-Christoph Klie and Michael Bugert and Beto Boullosa and Richard Eckart de Castilho and Iryna Gurevych, *The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation*. COLING, pp.5–9, 2018.
- 6 Lorenz Stangier and Ji-Ung Lee and Yuxi Wang and Marvin Müller and Nicholas Frick and Joachim Metternich and Iryna Gurevych, *TexPrax: A Messaging Application for Ethical, Real-time Data Collection and Annotation*. ACL/IJCNLP, pp.9–16, 2022.
- 7 Nils Dycke and Ilia Kuznetsov and Iryna Gurevych, *NLPeer: A Unified Resource for the Computational Study of Peer Review*. arXiv preprint 2211.06651. 2022.
- 8 Nils Dycke and Ilia Kuznetsov and Iryna Gurevych, *Yes-Yes-Yes: Donation-based Peer Reviewing Data Collection for ACL Rolling Review and Beyond*. arXiv preprint 2201.11443, 2022.
- 9 Chia-Chien Hung and Anne Lauscher and Simone Paolo Ponzetto and Goran Glavas, *DS-TOD: Efficient Domain Specialization for Task-Oriented Dialog*. ACL, pp.891–904, 2022.
- 10 Alec Radford and Jong Wook Kim and Tao Xu and Greg Brockman and Christine McLeavey and Ilya Sutskever, *Robust Speech Recognition via Large-Scale Weak Supervision*. arXiv preprint 2212.04356. 2022.

Participants

- Khalid Al-Khatib
University of Groningen, NL
- Milad Alshomary
Leibniz Universität
Hannover, DE
- Wolf-Tilo Balke
TU Braunschweig, DE
- Tilman Beck
TU Darmstadt, DE
- Elena Cabrio
Université Côte d'Azur –
Sophia Antipolis, FR
- Fengyu Cai
TU Darmstadt, DE
- Davide Ceolin
CWI – Amsterdam, NL
- Anita de Waard
Elsevier – Jericho, US
- Nils Dycke
TU Darmstadt, DE
- Dayne Freitag
SRI – Menlo Park, US
- Daniel Garijo
Polytechnic University of
Madrid, ES
- Iryna Gurevych
TU Darmstadt, DE
- Graeme Hirst
University of Toronto, CA
- Yufang Hou
IBM Research – Dublin, IE
- Eduard H. Hovy
Carnegie Mellon University,
Pittsburgh, US & University of
Melbourne, AU
- Anne Lauscher
Universität Hamburg, DE
- Maria Liakata
Queen Mary University of
London, GB
- Tobias Mayer
TU Darmstadt, DE
- Robert Mercer
University of Western Ontario –
London, CA
- Smaranda Muresan
Columbia University –
New York, US
- Preslav Nakov
MBZUAI – Abu Dhabi, AE
- Sukannya Purkayastha
TU Darmstadt, DE
- Chris Reed
University of Dundee, GB
- Domenic Rosati
scite – Halifax, CA
- Florian Ruosch
Universität Zürich, CH
- Harrison Scells
Universität Leipzig, DE
- Ferdinand Schlatt
Universität Halle-
Wittenberg, DE
- Benno Stein
Bauhaus-Universität Weimar, DE
- Simone Teufel
University of Cambridge, GB
- Serena Villata
Université Côte d'Azur –
Sophia Antipolis, FR
- Andreas Vlachos
University of Cambridge, GB
- Henning Wachsmuth
Universität Paderborn, DE
- Ryan Wang
University of Illinois
Urbana-Champaign, USA



Optimization at the Second Level

Luce Brotcorne^{*1}, Christoph Buchheim^{*2}, Dick den Hertog^{*3}, and Dorothee Henke^{†4}

1 INRIA Lille, FR. luce.brotcorne@inria.fr

2 TU Dortmund, DE. christoph.buchheim@tu-dortmund.de

3 University of Amsterdam, NL. d.denhertog@tilburguniversity.edu

4 TU Dortmund, DE. dorothee.henke@math.tu-dortmund.de

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 22441 “Optimization at the Second Level”. The seminar was held on October 30 – November 4, 2022 in Schloss Dagstuhl – Leibniz-Zentrum für Informatik. Participants gave overview talks and presented recent results in bilevel, robust, and stochastic optimization. These three areas have in common that they typically deal with optimization problems which are contained in the second level of the polynomial hierarchy. The goal of the seminar was to bring together experts of bilevel, robust, and stochastic optimization in order to connect and work towards new insights and approaches for such problems. During the seminar, the relationships between these different areas of optimization were intensively discussed and interesting connections were identified.

Seminar October 30 – November 4, 2022 – <http://www.dagstuhl.de/22441>

2012 ACM Subject Classification Theory of computation → Complexity classes; Theory of computation → Complexity theory and logic; Theory of computation → Mathematical optimization

Keywords and phrases bilevel optimization, robust optimization, stochastic optimization, computational complexity, algorithmics

Digital Object Identifier 10.4230/DagRep.12.10.207

1 Executive Summary

Luce Brotcorne (INRIA Lille, FR)

Christoph Buchheim (TU Dortmund, DE)

Dick den Hertog (University of Amsterdam, NL)

License © Creative Commons BY 4.0 International license
© Luce Brotcorne, Christoph Buchheim, and Dick den Hertog

Topic of the Seminar

The second level of the polynomial hierarchy contains a variety of problems that allow natural simple formulations with one existential and one universal quantifier. For instance, a typical problem in robust optimization asks whether there EXISTS some production plan that performs reasonably well under ALL possible price scenarios for electricity in the coming two years. A typical problem in bilevel optimization asks whether there EXISTS a way of setting taxes so that ALL possible behaviors of the citizens generate a reasonable tax revenue. A typical problem in Stackelberg games asks whether there EXISTS a starting move for the first player that wins the game against ALL possible counter-moves of the second player.

* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Optimization at the Second Level, *Dagstuhl Reports*, Vol. 12, Issue 10, pp. 207–224

Editors: Luce Brotcorne, Christoph Buchheim, Dick den Hertog, and Dorothee Henke



Dagstuhl Reports

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

Problems of this type are usually complete for the class Σ_2^P and hence are most likely not contained in the class NP. For that reason, the methodologies that have been developed for NP-complete problems over the last 50 years do not directly apply to robust and/or bilevel optimization problems. Up to the current moment, most of the work on such problems is purely computational and without any deeper theoretical understanding. Most approaches simply try to carry over the well-developed machinery from integer programming to concrete robust problems and bilevel problems. We will need to develop new techniques, new tricks, new insights, new algorithms, and new theorems to get a grip on this area.

The goal of this Dagstuhl Seminar was to bring together experts in theoretical computer science and experts in combinatorial optimization, and to work towards the following goals:

- summarize the status quo of robust optimization and bilevel optimization,
- identify central research lines on the computational and implementational front,
- identify central research lines in theoretical computer science, as for instance in parameterized complexity and approximability.

The list of participants perfectly reflected these goals, it included experts in complexity theory as well as researchers interested in the development of effective algorithms and the practical solution of real-world bilevel or robust optimization problems.

Implementation and Conclusions

In this seminar we brought together, for the first time, leading researchers from three different communities (robust optimization, stochastic optimization, and bilevel optimization) in order to bridge the gap between these fields from a theoretical and practical point of view.

Considering the different backgrounds of participants, we scheduled several talks with an introductory character in the first half of the week: Marc Goerigk and Frauke Liers gave overview talks on robust optimization, from a combinatorial and continuous perspective, respectively. Martine Labbé presented the state of the art of bilevel optimization, while Martin Schmidt combined both topics in his talk on bilevel optimization under uncertainty. Bernardo Pagnoncelli gave an introduction into the related topic of stochastic optimization. These presentations laid a common foundation for all further presentations and discussions and were thus a crucial prerequisite for the success of the seminar.

The contributed talks covered a wide range of topics, including complexity theoretic results for (certain or uncertain) bilevel optimization, new models and new methods for bilevel or robust optimization, and new approaches for solving bilevel or robust optimization problems arising in practice. Apart from the exciting contents of these talks, a particularly positive aspect was the large representation of young researchers. In fact, seven out of 18 contributed talks were given by PhD students.

The whole seminar was marked by a very open and constructive atmosphere and by an extraordinarily interactive approach: many presentations quickly turned into lively discussions involving many different participants, often making the original schedule obsolete, but with the benefit of a better common understanding and often new insights. One of the recurrent topics arising in many of these discussions was the connection between bilevel and robust optimization, the two main subjects of the seminar. The fact that both problem classes lead to potentially Σ_2^P -hard problems, as mentioned above, yields a connection of rather theoretical nature. From a more concrete point of view, the discussion was about whether one of the two problem classes can be seen (or modeled) as a special case of the other. Additionally, interesting links to game theory and stochastic optimization were identified.

Even though a conclusive answer to these questions could not be given (and probably does not exist), one of the main insights of the seminar was that bilevel and robust optimization, though being investigated in separate communities, share many structural and algorithmic

properties. It is worth studying these connections and sharing the knowledge of both communities in order to profit from one another. The Dagstuhl Seminar on “Optimization at the Second Level” was a first significant step into this direction, which is hopefully followed by further progress.

The seminar was a big success, it stimulated new and very fruitful collaborations. We got laudatory feedback from many participants who were already thinking of organizing another seminar on a the same topic in the future.

In memoriam

At this point, our thoughts go out to our late colleague and friend Gerhard J. Woeginger. It was his idea to organize a seminar on “Optimization at the Second Level”, and without him the seminar would never have become real. Unfortunately, he could no longer witness how his idea was put into practice and how fruitful it turned out to be.

References

- 1 Gerhard J. Woeginger: “The trouble with the second quantifier”, 4OR – A Quarterly Journal of Operations Research, Vol. 19(2), pp. 157–181, 2021.

2 Table of Contents**Executive Summary**

Luce Brotcorne, Christoph Buchheim, and Dick den Hertog 207

Overview of Talks


On a Computationally Ill-Behaved Bilevel Problem with a Continuous and Nonconvex Lower Level <i>Yasmine Beck</i>	212
A pricing and routing problem for last-mile delivery <i>Martina Cerulli</i>	212
Reformulation-Perspectification Technique for nonconvex (robust) optimization <i>Danique de Moor</i>	213
Two-stage robust optimization with objective uncertainty <i>Boris Detienne</i>	213
Integer Programming Games (and Robust optimization) <i>Gabriele Dragotto</i>	214
Results and Challenges in Robust Combinatorial Optimization <i>Marc Goerigk</i>	214
The Robust Bilevel Selection Problem <i>Dorothee Henke</i>	215
Quadratic Regularization of Unit-Demand Envy-Free Pricing Problems and Application to Electricity Markets <i>Quentin Jacquet</i>	215
A comparison of k -adaptability and min-max-min robustness: new results and insights <i>Jannis Kurtz</i>	216
Bilevel Optimization with a focus on the linear case <i>Martine Labbé</i>	217
Incentivizing Truthfulness in Core-Stable Intermediated Combinatorial Exchanges with Budget Constraints <i>Christina Liepold</i>	217
Robust Optimization, including Data, Nonlinearity, and Current Research Questions <i>Frauke Liers</i>	218
SOCP-Based Disjunctive Cuts for a Class of Integer Nonlinear Bilevel Programs <i>Ivana Ljubić</i>	218
Oracle-base methods to solve multi-stage adjustable robust optimization with right-hand-side uncertainty <i>Ahmadreza Marandi</i>	219
On the complexity of bilevel bottleneck assignment problem <i>Komal Muluk</i>	219
Stochastic optimization: a tutorial to establish connections <i>Bernardo Pagnoncelli</i>	220


Projections are the new binaries <i>Jean Pauphilet</i>	220
K -adaptability with few recourse solutions <i>Michael Poss</i>	221
80–20 optimization under uncertainty <i>Krzysztof Postek</i>	221
Duality and Value Functions <i>Ted Ralphs</i>	222
Some recent results and thoughts on bilevel optimization under uncertainty <i>Martin Schmidt</i>	222
A Reverse Stackelberg Game Model for Grid Usage Pricing with Local Energy Markets <i>Juan Sepúlveda</i>	223
Dynamic pricing for Public Cloud Computing <i>Nathalia Wolf</i>	223
Participants	224

3 Overview of Talks

3.1 On a Computationally Ill-Behaved Bilevel Problem with a Continuous and Nonconvex Lower Level

Yasmine Beck (Universität Trier, DE)

License  Creative Commons BY 4.0 International license

 Yasmine Beck

Joint work of Yasmine Beck, Daniel Bienstock, Martin Schmidt, Johannes Thürauf


Main reference Yasmine Beck, Daniel Bienstock, Martin Schmidt, Johannes Thürauf: “On a Computationally Ill-Behaved Bilevel Problem with a Continuous and Nonconvex Lower Level”, arXiv, 2022.


URL <https://doi.org/10.48550/ARXIV.2202.01033>

It is well known that bilevel optimization problems are hard to solve both in theory and practice. In this talk, we highlight a further computational difficulty when it comes to solving bilevel problems with continuous but nonconvex lower levels. Even if the lower-level problem is solved to ε -feasibility regarding its nonlinear constraints for an arbitrarily small but positive ε , the obtained bilevel solution as well as its objective value may be arbitrarily far away from the actual bilevel solution and its actual objective value. This result even holds for bilevel problems for which the nonconvex lower level is uniquely solvable, for which the strict complementarity condition holds, for which the feasible set is convex, and for which Slater’s constraint qualification is satisfied for all feasible upper-level decisions. Since the consideration of ε -feasibility cannot be avoided when solving nonconvex problems to global optimality, our result shows that computational bilevel optimization with continuous and nonconvex lower levels needs to be done with great care. Finally, we show that the nonlinearities in the lower level are the key reason for the observed bad behavior by proving that this behavior cannot appear for linear bilevel problems.

3.2 A pricing and routing problem for last-mile delivery

Martina Cerulli (ESSEC Business School – Cergy Pontoise, FR)

License  Creative Commons BY 4.0 International license

 Martina Cerulli

Joint work of Claudia Archetti, Martina Cerulli, Elena Fernandez, Ivana Ljubić

The Profitable Tour Problem (PTP) belongs to the class of Vehicle Routing Problems with profits. In PTP, a vehicle, starting from a central depot, can visit a subset of the available customers, collecting a specific revenue whenever a customer is visited. The objective of the problem is the maximization of the net profit, i.e., the total collected revenue minus the total route cost. Most of the literature in this field considers only one decision maker. However, in several real-world routing applications, and in particular in the last-mile delivery, there are different involved agents with conflicting goals. If the decisions are made in a hierarchical order, this problem can be modeled with bilevel programming, with the PTP at the lower level. In this talk, we consider a company, which acts as a leader and offers disjoint subsets of a given set of items to a set of independent drivers. Each driver solves a PTP communicating to the company the items she accepts to serve. Both the company and the drivers aim at maximizing their net profit, which is calculated differently in the two levels. We propose a bilevel formulation that models this interaction allowing the leader not only to anticipate the best followers’ response, but also to find the optimal pricing scheme for each carrier. The value function reformulation of this bilevel program is considered, and further

reformulated by projecting out some of the lower-level variables. We find exact solutions to these models using a branch and cut approach, leveraging on an alternative reformulation of the lower-level problem, which benefits from a particular monotonicity property.

3.3 Reformulation-Perspectification Technique for nonconvex (robust) optimization

Danique de Moor (University of Amsterdam, NL)

License © Creative Commons BY 4.0 International license
© Danique de Moor

Joint work of Jianzhe Zhen, Danique de Moor, Dick den Hertog

Main reference Jianzhe Zhen, Danique de Moor, and Dick den Hertog: “An extension of the reformulation-linearization technique to nonlinear optimization”, *Optimization Online*, Vol. 17221, 2021.

URL <https://optimization-online.org/?p=17221>

In this talk, a new technique is introduced, called the Reformulation-Perspectification Technique (RPT), to obtain convex approximations of nonconvex continuous (robust) optimization problems. RPT consists of two steps, those are, a reformulation step and a perspectification step. The reformulation step generates redundant nonconvex constraints from pairwise multiplication of the existing constraints. The perspectification step then convexifies the nonconvex components by using perspective functions. The proposed RPT extends the existing Reformulation-Linearization Technique (RLT) in two ways. First, it can multiply constraints that are not linear or not quadratic, and thereby obtain tighter approximations than RLT. Second, it can also handle more types of nonconvexity than RLT. A numerical experiment on convex maximization problems demonstrates the effectiveness of the proposed approach.

3.4 Two-stage robust optimization with objective uncertainty

Boris Detienne (University of Bordeaux, FR)

License © Creative Commons BY 4.0 International license
© Boris Detienne

Joint work of Ayşe Nur Arslan, Boris Detienne, Henri Lefebvre, Enrico Malaguti, Michele Monaci

Main reference Boris Detienne, Henri Lefebvre, Enrico Malaguti, Michele Monaci: “Adaptive robust optimization with objective uncertainty”, 2021.

URL <https://hal.inria.fr/hal-03371438>

In this talk, we study optimization problems where some cost parameters are not known at decision time and the decision flow is modeled as a two-stage process within a robust optimization setting. We address general problems in which all constraints (including those linking the first and the second stages) are defined by convex functions and involve mixed-integer variables. We first give a short proof that the special case of the problem with linear constraints is NP-complete. We then show how the general problem can be reformulated using Fenchel duality, allowing to derive an enumerative exact algorithm, for which we prove asymptotic convergence in the general case, and finite convergence for cases where the first-stage variables are all integer. An implementation of the resulting algorithm, embedding a column generation scheme, is then computationally evaluated on a variant of the Capacitated Facility Location Problem with unknown transportation costs, using instances that are derived from the existing literature.

3.5 Integer Programming Games (and Robust optimization)

Gabriele Dragotto (Princeton University, US)

License  Creative Commons BY 4.0 International license
 Gabriele Dragotto

Joint work of Gabriele Dragotto, Rosario Scatamacchia
Main reference Gabriele Dragotto, Rosario Scatamacchia: “The ZERO Regrets Algorithm: Optimizing over Pure Nash Equilibria via Integer Programming”, arXiv, 2021.

URL <https://doi.org/10.48550/ARXIV.2111.06382>

In this talk, we briefly survey Integer Programming Games (IPGs), i.e., simultaneous one-shot non-cooperative games where each player solves a parametrized integer program. After discussing a few motivating examples, we introduce the concept of Nash equilibrium for IPGs, namely a stable solution where no single agent has the incentive to defect from it profitably. Although the concept of Nash equilibrium has a natural and simple interpretation, determining if an IPG instance has a Nash equilibrium is generally a Σ_2^P -complete problem. We present ZERO Regrets, an algorithm to compute, enumerate and optimize over the Nash equilibria of IPGs. We propose some ideas for its extension to generalized IPGs, i.e., IPGs where each player’s feasible set depends on their opponent’s variables. Finally, we sketch the connection between Nash equilibria and robust optimization. In particular, we argue that robust combinatorial optimization and IPGs are equivalent problems, and the robust solution is indeed a Nash equilibrium.

3.6 Results and Challenges in Robust Combinatorial Optimization

Marc Goerigk (Universität Siegen, DE)

License  Creative Commons BY 4.0 International license
 Marc Goerigk

Robust optimization over combinatorial problems often behave quite differently to their more general counterparts. In many settings, we only consider uncertainty in the objective and assume that the combinatorial structure is known (of course, other settings are considered as well). In this talk I try to give a general introduction to this topic, covering min-max, min-max regret, two-stage and multi-stage problems with different types of uncertainty (discrete, budgeted, intervals), as well as pointing out some of the more recent results and challenges. In particular, many complexity questions remain open (e.g, the approximability of min-max regret problems with interval uncertainty, the complexity of two-stage assignment with continuous budgeted uncertainty, and many multi-stage or explorable problems). Furthermore, for experiments, often random data is generated, but the algorithm performance on random instances is poorly understood. To help a systematic approach towards experimental work, we are currently building a benchmark website for datasets, see <https://robust-optimization.com/>

3.7 The Robust Bilevel Selection Problem

Dorothee Henke (TU Dortmund, DE)

License © Creative Commons BY 4.0 International license
© Dorothee Henke

In bilevel optimization problems, there are two players, the leader and the follower, who make their decisions in a hierarchy, and both decisions influence each other. Usually one assumes that both players have full knowledge also of the other player's data and objective. Although bilevel problems are often already NP- or even Σ_2^P -hard to solve under this assumption, a more realistic model might sometimes be needed. Uncertainty can be quantified, for example, using the robust optimization approach: Assume that the leader does not know the follower's objective function precisely, but only knows an uncertainty set of potential follower's objectives, and the leader's aim is to optimize the worst case of the corresponding scenarios. Now the question arises how the computational complexity of bilevel optimization problems changes under the additional complications of this type of uncertainty.

We make a step towards answering this question by investigating a simple bilevel problem that can be solved in polynomial time without uncertainty. In the bilevel selection problem, the leader and the follower each select a number of items, while a common number of items to select in total is given, and each of the two players maximizes the total value of the selected items, according to different sets of item values. We compare several variants of this problem and show that all of them can be solved in polynomial time. We then investigate the complexity of the robust version of the problem. If the item sets controlled by the leader and by the follower are disjoint, it can still be solved in polynomial time in case of a finite uncertainty set or interval uncertainty. If the two item sets are not disjoint, the robust problem version becomes NP-hard, even for a finite uncertainty set, which shows that uncertainty can indeed add additional complexity to a bilevel optimization problem.

3.8 Quadratic Regularization of Unit-Demand Envy-Free Pricing Problems and Application to Electricity Markets

Quentin Jacquet (EDF – Paris, FR)

License © Creative Commons BY 4.0 International license
© Quentin Jacquet

Joint work of Quentin Jacquet, Wim van Ackooij, Clémence Alasseur, and Stéphane Gaubert

Main reference Quentin Jacquet, Wim van Ackooij, Clémence Alasseur, Stéphane Gaubert: "Quadratic Regularization of Unit-Demand Envy-Free Pricing Problems and Application to Electricity Markets", arXiv, 2021.

URL <https://doi.org/10.48550/ARXIV.2110.02765>

We consider a profit-maximizing model for pricing contracts as an extension of the unit-demand envy-free pricing problem: customers aim to choose a contract maximizing their utility based on a reservation bill and multiple price coefficients (attributes). A classical approach supposes that the customers have deterministic utilities; then, the response of each customer is highly sensitive to price since it concentrates on the best offer. A second approach is to consider logit model to add a probabilistic behavior in the customers' choices. To circumvent the intrinsic instability of the former and the resolution difficulties of the latter, we introduce a quadratically regularized model of customer's response, which leads to a quadratic program under complementarity constraints (QPCC). This allows to robustify the deterministic model, while keeping a strong geometrical structure. In particular, we

show that the customer's response is governed by a polyhedral complex, in which every polyhedral cell determines a set of contracts which is effectively chosen. Moreover, the deterministic model is recovered as a limit case of the regularized one. We exploit these geometrical properties to develop an efficient pivoting heuristic, which we compare with implicit or non-linear methods from bilevel programming. These results are illustrated by an application to the optimal pricing of electricity contracts on the French market.

3.9 A comparison of k -adaptability and min-max-min robustness: new results and insights

Jannis Kurtz (University of Amsterdam, NL)

License  Creative Commons BY 4.0 International license
© Jannis Kurtz

Main reference Jannis Kurtz: "New complexity results and algorithms for min-max-min robust combinatorial optimization", arXiv, 2021.

URL <https://doi.org/10.48550/ARXIV.2106.03107>

Robust optimization is one of the most successful concepts to tackle optimization problems under uncertainty. In this talk we compare two popular robust optimization models, namely k -adaptability [1] and *min-max-min robust optimization* [2], both under objective uncertainty. While k -adaptability was introduced to approximate two-stage robust problems by calculating k second-stage solutions already in the first stage, the concept of min-max-min robustness was introduced as a less conservative version of the classical robust optimization problem where the user can calculate k instead of a single solution. This provides more flexibility, better performance for future uncertain scenarios and a small set of solutions which, in contrast to the two-stage robust problem, do not change in the future. This is especially useful for applications from disaster management, where a team has to be trained on an ideally small set of solutions.

While at a first glance both problems are very similar, it turns out they significantly differ in terms of computational complexity. However recent results derived for min-max-min robust optimization can partly be extended to k -adaptability problems and give new insights into the latter problem class.


We analyze the three cases where k , the number of calculated (second-stage) solutions, is a) larger than the dimension of the problem, b) a small value, c) smaller but close to the dimension of the problem. It turns out that, if the underlying deterministic problem is tractable, then the min-max-min robust problem is tractable in case a), often tractable in case c), and NP-hard in case b). On the other hand the k -adaptability problem can be NP-hard for all cases and provides the exact optimal value of the two-stage robust problem in case a). We provide an efficient heuristic algorithm which calculates good solutions for both problems for every k and give theoretical and computational proof that the optimality gap of these heuristic solutions decreases if k increases. We incorporate this heuristic into a branch & bound framework which finds optimal solutions for both problems for any k . Our computations show that the larger k the more instances of both problems can be solved during the timelimit which is mainly due to the decreasing optimality gap of our heuristic. This gives rise to the conclusion that solving the exact two-stage robust optimization problem is computationally more efficient than approximating it by the k -adaptability approach in case of objective uncertainty.

References

- 1 Grani A. Hanasusanto, Daniel Kuhn, and Wolfram Wiesemann: “ K -Adaptability in Two-Stage Robust Binary Programming”, *Operations Research*, Vol. 63(4), pp. 877–891, 2015.
- 2 Christoph Buchheim and Jannis Kurtz: “Min-max-min Robust Combinatorial Optimization”, *Mathematical Programming*, Vol. 163(1), pp. 1–23, 2016.

3.10 Bilevel Optimization with a focus on the linear case

Martine Labbé (UL – Brussels, BE)


License  Creative Commons BY 4.0 International license
© Martine Labbé

A bilevel optimization problem consists in an optimization problem in which some of the constraints specify that a subset of variables must be an optimal solution to another optimization problem. This paradigm is particularly appropriate to model competition between agents, a leader and a follower, acting sequentially.

In this talk I focus on the simplest bilevel problems, those that are linear. I present the main characteristics, properties and algorithms for these problems. Then, I discuss some recent results showing that these problems are already extremely challenging. Finally, I show how linear bilevel optimisation may help to improve the resolution of stochastic optimisation problems whose objective involve the Value at Risk (VaR) or quantile.

3.11 Incentivizing Truthfulness in Core-Stable Intermediated Combinatorial Exchanges with Budget Constraints

Christina Liepold (TU München, DE)

License  Creative Commons BY 4.0 International license
© Christina Liepold
Joint work of Christina Liepold, Maximilian Schiffer

Servitization markets that center their business around operationally connecting buyers and suppliers of goods, services, or capacities can be modeled as combinatorial exchanges. Usually, the markets employ a profit-oriented intermediary to coordinate the process. Particularities of the servitization market are that (I) potential buyers face budget constraints, which provide an incentive to overstate valuations, resulting in exchanges that are not welfare-maximizing and (II) core-stability of the market is essential to foster participation and, thus, the attractiveness of the heterogeneous exchange. However, it has been proven that no truthful mechanism exists under private budget constraints in combinatorial exchanges. We propose the introduction of a profit-oriented intermediary, which allows for the incentivization of truthfulness in these markets. We utilize the properties of Mixed Integer Bilevel Linear Problems to formulate the allocation and pricing problem for core-stable intermediated combinatorial exchanges under private budget constraints. We find that intermediated combinatorial exchanges with a profit-oriented intermediary lead to higher social welfare than truthful non-intermediated benchmarks while restricting the matching of untruthful buyers.

3.12 Robust Optimization, including Data, Nonlinearity, and Current Research Questions

Frauke Liers (Universität Erlangen-Nürnberg, DE)

License © Creative Commons BY 4.0 International license
© Frauke Liers

Joint work of Dennis Adelhütte, Martina Kuchlbauer, Frauke Liers, Michael Stingl

Main reference Martina Kuchlbauer, Frauke Liers, Michael Stingl: “Outer Approximation for Mixed-Integer Nonlinear Robust Optimization”, *J. Optim. Theory Appl.*, Vol. 195(3), pp. 1056–1086, 2022.

URL <https://doi.org/10.1007/s10957-022-02114-y>

In this talk, I first reviewed robust convex optimization, focussing on duality-based reformulations of robust counterparts to finite and algorithmically tractable problems as well as cutting plane approaches. For constraints that are convex in the decisions and concave in the uncertainty, Fenchel duality can be used. For combinatorial optimization problems with nonlinear objective, I showed new reformulations that combine Gamma-robust counterparts with Fenchel duality (This part is joint work with Dennis Adelhütte, FAU). A novel approach (joint with Martina Kuchlbauer and Michael Stingl, both FAU) for mixed-integer nonlinear robust optimization was summarized. Here, an adaptive bundle method uses nonconvex adversarial problems that are solved up to any given error, e.g., via piecewise linear relaxations. For problems that are convex in the uncertainty and potentially nonconvex in the uncertainty, the bundle method is embedded in an outer approximation scheme. I then continued with reviewing distributional robustness, focussing on reformulations and the construction of ambiguity sets based on historical data. I concluded by pointing out current research questions.

3.13 SOCP-Based Disjunctive Cuts for a Class of Integer Nonlinear Bilevel Programs

Ivana Ljubić (ESSEC Business School – Cergy Pontoise, FR)

License © Creative Commons BY 4.0 International license
© Ivana Ljubić

Joint work of Elisabeth Gaar, Jon Lee, Ivana Ljubić, Markus Sinnl, Kübra Tanınmış

Main reference Elisabeth Gaar, Jon Lee, Ivana Ljubić, Markus Sinnl, Kübra Tanınmış: “On SOCP-based disjunctive cuts for solving a class of integer bilevel nonlinear programs”, arXiv, 2022.

URL <https://doi.org/10.48550/ARXIV.2207.05014>

We study a class of integer bilevel programs with second-order cone constraints at the upper level and a convex-quadratic objective function and linear constraints at the lower level. We develop disjunctive cuts (DCs) to separate bilevel-infeasible solutions using a second-order-cone-based cut-generating procedure. We propose DC separation strategies and consider several approaches for removing redundant disjunctions and normalization. Using these DCs, we propose a branch-and-cut algorithm for the problem class we study, and a cutting-plane method for the problem variant with only binary variables. We present an extensive computational study on a diverse set of instances, including instances with binary and with integer variables, and instances with a single and with multiple linking constraints. Our computational study demonstrates that the proposed enhancements of our solution approaches are effective for improving the performance. Moreover, both of our approaches outperform a state-of-the-art generic solver for mixed-integer bilevel linear programs that is able to solve a linearized version of our binary instances.

3.14 Oracle-base methods to solve multi-stage adjustable robust optimization with right-hand-side uncertainty

Ahmadreza Marandi (TU Eindhoven, NL)

License © Creative Commons BY 4.0 International license
© Ahmadreza Marandi

Joint work of Ahmadreza Marandi, Ali Borumand, Geert-Jan van Houtum, Zumbul Atan

In this talk, we address multi-stage adjustable robust optimization with right-hand-side uncertainty. We consider both continuous cases and mixed integer cases. For continuous cases, we show that the problem can be solved by enumerating the vertices of the uncertainty set. So, for a polyhedral uncertainty set, we develop a simplex-type method that starts from a vertex and moves to an adjacent vertex with a worse objective value. For the mixed integer case, we developed an iterative approach that converges to the optimal solution. The algorithm is based on partitioning the uncertainty set into smaller sets and obtaining a lower and upper bound for each subset. Since the algorithm itself is an exhaustive search, we also developed a heuristic method based on the limited discrepancy search, where we develop a branching tree and construct waves to explore the tree. We show that both algorithms work in practical problems, as it is based on an oracle that solves the deterministic problem in each iteration.

3.15 On the complexity of bilevel bottleneck assignment problem

Komal Muluk (RWTH Aachen, DE)

License © Creative Commons BY 4.0 International license
© Komal Muluk

Joint work of Dennis Fischer, Komal Muluk, Gerhard Woeginger

Main reference Dennis Fischer, Komal Muluk, Gerhard J. Woeginger: “A note on the complexity of the bilevel bottleneck assignment problem”, 4OR, Vol. 20(4), pp. 713–718, 2022.

URL <https://doi.org/10.1007/s10288-021-00499-6>

Bilevel assignment problem is an example of bilevel optimization problems. In an instance of bilevel assignment problem, we are given a bipartite graph $G = (V, E)$, and the edge set E is partitioned into two sets E_ℓ and E_f . The set E_ℓ is controlled by the leader and the set E_f is controlled by the follower. Additionally, leader and follower have their own weight functions w_ℓ and w_f defined on the edge set E . Both decision makers have their own objective functions c, d for the leader and the follower respectively. We consider the settings in which the objective functions of the leader and the follower can be of sum or bottleneck type. Moreover, the follower can behave according to the optimistic or pessimistic rule. The goal of the leader is to find $L \subseteq E_\ell$ which minimizes her objective function subject to an optimal reaction of the follower $F \subseteq E_f$ such that $L \cup F$ forms a perfect matching in the graph G . We give a unified NP-hardness proof for the eight variants of bilevel assignment problem; the eight variants arise from the combinations of sum or bottleneck type of objective functions c and d , and from the pessimistic or optimistic behavior of the follower. Consequently, we show non-existence of an approximation algorithm for this problem.

3.16 Stochastic optimization: a tutorial to establish connections

Bernardo Pagnoncelli (SKEMA Business School – Lille, FR)

License © Creative Commons BY 4.0 International license
© Bernardo Pagnoncelli

In this presentation I will give a broad introduction to stochastic optimization, focusing on static problems. I will start with the classic paradigm of minimizing expected costs, and then discuss why the inclusion of risk measures is desirable in many practical situations. I will exemplify the effect of risk aversion in the newsvendor problem, using the CVaR as the risk measure.

I will briefly introduce the dynamic setting, and the different algorithms to solve the problems in sequential decision-making. I will conclude with the current challenges in end-to-end learning, which consist of problems that integrate prediction and prescription. Bilevel optimization is an important tool in this integration, and I will define precisely how it can build the bridge between machine learning and optimization.

3.17 Projections are the new binaries

Jean Pauphilet (London Business School, GB)

License © Creative Commons BY 4.0 International license
© Jean Pauphilet

Joint work of Dimitris Bertsimas, Ryan Cory-Wright, Jean Pauphilet

Main reference Dimitris Bertsimas, Ryan Cory-Wright, Jean Pauphilet: “Mixed-Projection Conic Optimization: A New Paradigm for Modeling Rank Constraints”, *Operations Research*, Vol. 70(6), pp. 3321–3344, 2022.

URL <https://doi.org/10.1287/opre.2021.2182>

Product recommendation, matrix factorization, or Euclidean embedding can be formulated as optimization problems with a constraint on the rank of the decision variables (a matrix). These optimization problems constitute hard non-convex optimization problems that cannot be modeled via convex mixed-integer optimization. Alternatively, we propose to introduce projection matrices (i.e., matrices that satisfy $P^2 = P$) to encode for the span of the decision variables and linearize the rank constraint. Projection matrices are not only a generalization of binary variables (i.e., scalar roots of the equation $z^2 = 2$) but we show that the resulting optimization framework, which we coin mixed-projection optimization, can effectively be used to solve low-rank optimization problems to provable optimality. In particular, some of the most successful ideas from mixed-integer optimization such as big-M formulations, perspective cuts, relax-then-round strategies generalize to low-rank problems by using projection matrices.

3.18 K -adaptability with few recourse solutions

Michael Poss (University of Montpellier & CNRS, FR)

License © Creative Commons BY 4.0 International license
© Michael Poss

Joint work of Ayşe Nur Arslan, Michael Poss, Marco Silva

Main reference Ayşe N. Arslan, Michael Poss, Marco Silva: “Min-Sup-Min Robust Combinatorial Optimization with Few Recourse Solutions”, *INFORMS J. Comput.*, Vol. 34(4), pp. 2212–2228, 2022.

URL <https://doi.org/10.1287/ijoc.2021.1156>

We present a variant of adaptive robust combinatorial optimization problems where the decision maker can prepare K solutions and choose the best among them upon knowledge of the true data realizations. We suppose that the uncertainty may affect the objective and the constraints through functions that are not necessarily linear.

In the first part of the talk, we propose a new exact algorithm for solving these problems when the feasible set of the nominal optimization problem does not contain too many good solutions. Our algorithm enumerates these good solutions, generates dynamically a set of scenarios from the uncertainty set, and assigns the solutions to the generated scenarios using a vertex p -center formulation, solved by a binary search algorithm. Our numerical results on adaptive shortest path and knapsack with conflicts problems show that our algorithm compares favorably with the methods proposed in the literature. We additionally propose a heuristic extension of our method to handle problems where it is prohibitive to enumerate all good solutions. This heuristic is shown to provide good solutions within a reasonable solution time limit on the adaptive knapsack with conflicts problem. Finally, we illustrate how our approach handles non-linear functions on an all-or-nothing subset problem taken from the literature.

In the second part of the talk, we discuss extensions of these algorithms to K -adaptability, where the decision maker must set first-stage optimization variables in addition to the preparation of the K second-stage solutions. We reformulate again the problem along the lines of p -center location problems, this time including additional coupling constraints. We discuss how subtle pricing and adversarial problems may lead to efficient branch-and-cut-price algorithms for solving the problem optimally.

3.19 80–20 optimization under uncertainty

Krzysztof Postek (TU Delft, NL)

License © Creative Commons BY 4.0 International license
© Krzysztof Postek

In this talk, we consider the design choices one has to make when considering optimization under uncertainty. Expanding the model form useful for solving the deterministic problem into an uncertainty-including form is often hard. This calls, in line with the seminar description, for completely new ways of modelling Σ_2^p problems. We illustrate one heuristic idea (evaluating incumbent solutions of a deterministic MILP for a given problem with stochastic simulations and picking the best one among those) on the fleet assignment problem in airline optimization, which performs similarly to a massive, exact MILP reformulation of the stochastic problem.

3.20 Duality and Value Functions

Ted Ralphs (Lehigh University – Bethlehem, US)

License  Creative Commons BY 4.0 International license
© Ted Ralphs

Duality is a central concept in optimization from which optimality conditions arise, among other important implications. It is duality-based optimality conditions that enable the reformulations that are the basis for many algorithms for addressing multistage optimization problems in which the problem to be solved at each stage is parametric in the decisions and information revealed at previous stages. Examples include multilevel optimization, robust optimization, and multistage stochastic programming with recourse. By replacing later stage problems with their optimality conditions, it is possible to reformulate these multistage problems as traditional mathematical optimization problems, albeit ones involving complex value function. While it is often stated that there is no strong duality for nonconvex optimization problems, such as mixed integer linear optimization problems (MILPs), we discuss in this talk that this is indeed not the case. We describe a duality theory that generalizes the classical LP/convex duality theory to the MILP setting and show that it can be derived from first principles beginning from each of three seemingly disparate starting points. The first is the traditional notion based on bounding of the value function, the second is based on a generalized Farkas lemma that leads to a complexity-theoretic interpretation, and the third is based on reformulation via projection, which leads to a generalized Benders decomposition. The relationship between these three different derivations and the generalized notions that arise shed light on the nature of the duality relations at the core of optimization theory and enable duality principles to be applied to a much wider range of problems.

3.21 Some recent results and thoughts on bilevel optimization under uncertainty

Martin Schmidt (Universität Trier, DE)

License  Creative Commons BY 4.0 International license
© Martin Schmidt

Joint work of Yasmine Beck, Ivana Ljubić, Martin Schmidt

Main reference Yasmine Beck, Ivana Ljubić, Martin Schmidt: “A Survey on Bilevel Optimization Under Uncertainty”, *European Journal on Operational Research*, forthcoming, 2023.

URL <https://optimization-online.org/?p=19047>

Bilevel optimization is a very active field of applied mathematics. The main reason is that bilevel optimization problems can serve as a powerful tool for modeling hierarchical decision making processes. This ability, however, also makes the resulting problems challenging to solve – both in theory and practice. In this talk, we focus on bilevel optimization problems under uncertainty. First, we give an overview about this rather young field. We particularly show that the sources of uncertainty are much richer in bilevel optimization when compared to “usual”, i.e., single-level, optimization. The reason is that, besides data uncertainty, one can also face uncertainties w.r.t. the decisions of the other player. Second, we briefly present two recent papers on uncertain bilevel optimization. One in which we develop a branch-and-cut framework for bilevel knapsack interdiction, where the lower-level is a Gamma-robustified model, and another one, in which we try to model the situation of a follower that is uncertain w.r.t. the leader’s decision. Third and finally, we sketch some open problems in the field of bilevel optimization under uncertainty to propel some further discussions during the seminar and beyond.

3.22 A Reverse Stackelberg Game Model for Grid Usage Pricing with Local Energy Markets

Juan Sepúlveda (INRIA Lille, FR)

License © Creative Commons BY 4.0 International license
© Juan Sepúlveda

Joint work of Juan Sepúlveda, Luce Brotcorne, H el ene Le Cadre

In the context of the massive penetration of renewables and the shift of the energy system operation towards more consumer-centric approaches, local energy markets are seen as a promising solution for prosumers' empowerment. Various local market designs have been proposed that often ignore the laws of physics ruling the underlying distribution network's power flows. This may compromise the power system's security or lead to operational inefficiencies. Therefore, including the distribution network in clearing the local market arises as a challenge. We propose using grid usage prices (GUPs) as an incentive mechanism to drive the system towards an economically and operationally efficient market equilibrium, subject to security constraints. Our approach requires expressing the incentive policies as affine functions of the prosumers' active and reactive power outputs. This setting falls into the category of reverse Stackelberg games, where we look for the optimal policy in the space of affine functions [2]. This approach takes advantage of controllability guarantees for the problem's unconstrained setting, which hopefully will enable the DSO to influence the output of the market towards an optimally determined target point. Market-related properties of the policy, such as economic efficiency, individual rationality, incentive compatibility, and fairness, will be rigorously studied. Two alternative solution approaches are proposed: The first is based on reformulating the resulting bilevel problem into a single-level problem employing the KKT conditions of the underlying lower-level problem. Then, the resulting quadratically constrained quadratic problem is solved applying a trust-region method. The second solution approach is based on first obtaining a team-problem solution and solving a feasibility problem to induce it [1]. Finally, extensive computational experiments will be carried out on different IEEE test feeders to assess the performance of the proposed approach statistically.

References

- 1 Tamer Ba sar and Hasan Selbuz: "Closed-loop stackelberg strategies with applications in the optimal control of multilevel systems", IEEE Transactions on Automatic Control, Vol. 24(2), pp. 166–179, 1979.
- 2 Noortje Groot, Bart De Schutter, and Hans Hellendoorn: "Optimal Affine Leader Functions in Reverse Stackelberg Games: Existence Conditions and Characterization", Journal of Optimization Theory and Applications, Vol. 168(1), pp. 348–374, 2016.

3.23 Dynamic pricing for Public Cloud Computing

Nathalia Wolf (INRIA Lille, FR)

License © Creative Commons BY 4.0 International license
© Nathalia Wolf

Joint work of Bernard Fortz, Luce Brotcorne, Nathalia Wolf

In this work, we present a cloud sharing system that increases the utilization rate of public cloud resources. The system is modeled as a mixed integer bilevel problem with two independent follower level problems. The leader interacts with the followers by (i) offering rewards to long-term consumers to lease their resources and (ii) setting prices to short-term consumers to access available resources. We solve the model using an optimal value function algorithm.

Participants

- Yasmine Beck
Universität Trier, DE
- Luce Brotcorne
INRIA Lille, FR
- Christoph Buchheim
TU Dortmund, DE
- Martina Cerulli
ESSEC Business School –
Cergy Pontoise, FR
- Claudia D’Ambrosio
CNRS & Ecole Polytechnique –
Palaiseau
- Danique de Moor
University of Amsterdam, NL
- Dick den Hertog
University of Amsterdam, NL
- Boris Detienne
University of Bordeaux, FR
- Gabriele Dragotto
Princeton University, US
- Marc Goerigk
Universität Siegen, DE
- Dorothee Henke
TU Dortmund, DE
- Felix Hommelsheim
Universität Bremen, DE
- Quentin Jacquet
EDF – Paris, FR
- Jannis Kurtz
University of Amsterdam, NL
- Martine Labbé
UL – Brussels, BE
- Christina Liepold
TU München, DE
- Frauke Liers
Universität Erlangen-
Nürnberg, DE
- Ivana Ljubic
ESSEC Business School –
Cergy Pontoise, FR
- Ahmadreza Marandi
TU Eindhoven, NL
- Komal Muluk
RWTH Aachen, DE
- Bernardo Pagnoncelli
SKEMA Business School –
Lille, FR
- Jean Pauphilet
London Business School, GB
- Michael Poss
University of Montpellier &
CNRS, FR
- Krzysztof Postek
TU Delft, NL
- Ted Ralphs
Lehigh University –
Bethlehem, US
- Martin Schmidt
Universität Trier, DE
- Juan Pablo Sepulveda
Adriazola
INRIA Lille, FR
- Shimrit Shtern
Technion – Haifa, IL
- Nathalia Wolf
INRIA Lille, FR
- Pawel Zielinski
Wroclaw University of
Technology, PL



Toward Scientific Evidence Standards in Empirical Computer Science

Timothy Kluthe^{*1}, Brett A. Becker^{†2}, Christopher D. Hundhausen^{†3},
Ciera Jaspán^{†4}, Andreas Stefik^{†5}, and Thomas Zimmermann^{†6}

1 University of Nevada – Las Vegas, US. tjkluthe@gmail.com

2 University College Dublin, IE. brett.becker@ucd.ie

3 Oregon State University – Corvallis, US. hundhaus@wsu.edu

4 Google – Mountain View, US. ciera@google.com

5 University of Nevada – Las Vegas, US. stefika@gmail.com

6 Microsoft Corporation – Redmond, US. tzimmer@microsoft.com

Abstract

Many scientific fields of study use formally established evidence standards during the peer review and evaluation process, such as Consolidated Standards of Reporting Trials (CONSORT) in medical research, the What Works Clearinghouse (WWC) used in education in the United States, or the APA Journal Article Reporting Standards (JARS) in psychology. The basis for these standards is community agreement on what to report in empirical studies. Such standards achieve two key goals. First, they make it easier to compare studies, facilitating replications, through transparent reporting and sharing of data, which can provide confidence that multiple research teams can obtain the same results. Second, they establish community agreement on how to report on and evaluate studies using different methodologies. The discipline of computer science does not have formalized evidence standards, even for major conferences or journals. This Dagstuhl Seminar has three primary objectives:

1. To establish a process for creating or adopting an existing evidence standard for empirical research in computer science.
2. To build a community of scholars that can discuss what a general standard should include.
3. To kickstart the discussion with scholars from software engineering, human-computer interaction, and computer science education.

In order to better discuss and understand the implications of such standards across several empirical subfields of computer science and to facilitate adoption, we brought together participants from a range of backgrounds; including academia and industry, software engineering, computer-human interaction and computer science education, as well as representatives from several prominent journals.

Funding: This material is based upon work supported by the National Science Foundation under Grant Numbers NSF HCC: 2106392 and NSF I-TEST: 2048356.

Seminar October 30–November 4, 2022 – <http://www.dagstuhl.de/22442>

2012 ACM Subject Classification General and reference → Empirical studies; Human-centered computing → Empirical studies in HCI; Social and professional topics → Computing education; General and reference → Reliability

Keywords and phrases Community evidence standards, Human factors

Digital Object Identifier 10.4230/DagRep.12.10.225

* Editorial Assistant / Collector

† Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Toward Scientific Evidence Standards in Empirical Computer Science, *Dagstuhl Reports*, Vol. 12, Issue 10, pp. 225–240

Editors: Timothy Kluthe, Brett A. Becker, Christopher D. Hundhausen, Ciera Jaspán, Andreas Stefik, and Thomas Zimmermann




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

1 Executive Summary

Timothy Kluthe University of Nevada – Las Vegas, US, tjkluthe@gmail.com

Andreas Stefik University of Nevada – Las Vegas, US, stefika@gmail.com

License  Creative Commons BY 4.0 International license
© Timothy Kluthe and Andreas Stefik

The goals of the seminar *Toward Scientific Evidence Standards in Empirical Computer Science* were to establish a process for introducing evidence standards in computer science, build a community of scholars that discuss what a general standard would include and have enough diversity of background to have a good basis for the breadth of community needs across a range of computer science-related venues.

Over the first few days, we conducted a series of breakout groups and larger group discussions. In these, to introduce people to evidence standards, we reviewed several, including: APA JARS[1], WWC[2], and CONSORT[3]. The purpose was introductory and to scaffold for discussions on what could work across the breadth of computer science or in subareas. We also conducted a session looking at existing papers and noted the changes that would need to be made to fit the APA JARS standards. This exercise in particular was found to be useful by participants, as it made it clear that the conversion is not particularly difficult, although it is aided by advanced planning for what might need to be collected during a study.

During the Dagstuhl, we also had several talks. These included an introductory talk by Andreas Stefik on evidence standards as a whole, telling the story of the well-known Tolbutamide drug and its influence on the medical field in regard to evidence standards. Christopher Hundhausen provided a talk on his experience with introducing reporting standards at ACM's Transactions on Computing Education (TOCE) (Section 3.2). Paul Ralph presented on the problems in scholarly peer review and how evidence standards could be a solution, along with a reviewing tool that he has developed (Section 3.3). Neil Ernst covered registered reports, their benefits to the transparency and quality of research, and his experience with introducing them at Mining Software Repositories (MSR) and Empirical Software Engineering (EMSE) (Section 3.4). Lastly, Kate Sanders et al. discussed a review on reviews, which spanned a variety of the computer science subfields. This included their observations on review criteria, ethical concerns in the peer review process and excerpts from interviews with conference chairs and journal editors that were relevant to the subject of the seminar (Section 3.5). Each of these gave insights into the process of adopting an evidence standard and some of the potential impacts of the status quo and potential changes (positive or negative).

Finally, after discussion, we identified four topics for breakout groups to brainstorm potential avenues toward actionable progress on goals: a deeper dive into how to write guidelines for more complex experiments like mixed-methods studies (Section 4.5), how can we measure the effects that evidence standards have both in reference in paper quality and community satisfaction (Section 4.6), what are the first steps towards community engagement as far as introducing the topic and adopting it (Section 4.7) and how to operationalize these standards in a way that is open source to allow for community control (Section 4.8). A final working group session went through some of the first steps could be made at conferences and a dissemination plan for how to start information the community about the topic (Section 4.9).

Overall, the seminar brought a range of computer science stakeholders up to speed on the state of evidence standards in the field, what could be gained by moving towards a domain-wide guidelines and started a discussion on how to spark the conversation in

various communities. A set of next steps on where and what to recommend and talk about with communities were set in motion, as well as plans for a collaborative position paper to introduce the topic to a wider audience.

References

- 1 American Psychological Association. APA Style Journal Article Reporting Standards (APA Style JARS). Accessed on December 12, 2022 from <https://apastyle.apa.org/jars>.
- 2 National Center for Education Evaluation and Regional Assistance. WWC | Find What Works. Accessed on December 12, 2022 from <https://ies.ed.gov/ncee/wwc/>.
- 3 The CONSORT Group. CONSORT Transparent Reporting of Trials. Accessed on December 12, 2022 from <https://www.consort-statement.org/>.

2 Table of Contents

Executive Summary

Timothy Kluthe and Andreas Stefik 226

Overview of Talks

Seminar Goals and a Brief Introduction to Evidence Standards

Andreas Stefik 229

TOCE's Journey into Reporting Standards (and a Flirtation with Evidence Standards)

Christopher D. Hundhausen 229

Revolutionizing Peer Review with Empirical Standards

Paul Ralph 230

Registered Reports in Computer Science: Why Bother?

Neil Ernst 231

Criteria and Scrutiny: Computing Education Research

Kate Sanders, Joseph Maguire, and Monica McGill 232

Working Groups

APA JARS – Quantitative 233

APA JARS – Qualitative 234

What Works Clearinghouse 235

Take a Paper and Convert It 236

Breakout: Mixed-Methods 237

Breakout: Measuring Effects 237

Breakout: Community Engagement 238

Breakout: Operationalizing 238

Next steps: What Form of Actions Will We Take? 239

Participants 240

3 Overview of Talks

3.1 Seminar Goals and a Brief Introduction to Evidence Standards

Andreas Stefik (University of Nevada – Las Vegas, US)

License © Creative Commons BY 4.0 International license
© Andreas Stefik

The concept of checking our assumptions through the use of independently verifiable evidence has a long tradition in the sciences. While this is well known, throughout the 20th century, especially, external stakeholders to science have pushed the community toward using increasingly rigorous evidence. This is in part because science can have implications for the public or for public policy. In this talk, we will briefly introduce the topic of the Dagstuhl: evidence standards. In doing so, we will review an exemplar of why evidence standards came about and provide an overview of the week's activities.

3.1.1 Discussion

In this talk, there was a large focus on CONSORT[1], which is the set of evidence standards used in the medical community. The first point of discussion was to note that the standards required in the medical community are often centered around life or death problems which may not be quite the level of what is faced in computer science. A counterpoint was made that while computer science may not be facing life or death decisions, often times the papers and research can have an impact on public policy and economics (e.g., do we adopt a new programming language). Bad software design can be life threatening (e.g., self-driving cars), and social media algorithms can impact worldviews, which in turn impacts politics. Thus, unless we conclude computer science research does not matter for the real-world, any work that actually matters should plausibly be held to a more scientifically rigorous set of guidelines. Put succinctly, the greater the impact of the research, the more rigorous it needs to be to ensure the impact is desirable.

References

- 1 The CONSORT Group. CONSORT Transparent Reporting of Trials. Accessed on December 12, 2022 from <https://www.consort-statement.org/>.

3.2 TOCE's Journey into Reporting Standards (and a Flirtation with Evidence Standards)

Christopher D. Hundhausen (Oregon State University – Corvallis, US)

License © Creative Commons BY 4.0 International license
© Christopher D. Hundhausen

The ACM Transactions on Computing Education (TOCE) is a premier journal for computing education research. In 2019, the journal convened a task force to explore the possibility of adopting evidence standards for the journal. In 2021, ACM TOCE became the first journal in the field of computing to adopt reporting standards. In this talk, I chronicle TOCE's two year development of an evidence standards proposal; its discussion with the TOCE editorial board; and the TOCE editorial board's ultimate decision to recommend, but not require, the use of the APA JARS reporting standard for new submissions. Based on TOCE's experience, I identify barriers to the adoption of evidence standards in empirical computer science and consider possible ways forward.

3.2.1 Discussion

Some of the key lessons learned from introducing reporting standards at TOCE were:

- Include more people in the task force.
- Do not call it an evidence standard (most prefer reporting standard).
- Focus on how it makes the papers easier to review, read and search.

A lot of the pushback when introducing the CONSORT, and to some extent APA JARS, standards at TOCE was the fear that some might be excluded. Both of these standards have a heavy focus on quantitative research, and therefore left some concern that qualitative research may start to be rejected. It was noted that qualitative research is especially common in education, as compared to medical and psychology where these standards were established, and it may have been a closer fit to introduce them at a computer science conference or journal that has a heavier percentage of quantitative research, such as ACM’s Conference on Human Factors in Computing Systems (CHI). Ultimately, the choice to introduce the standards at an education-focused journal was The fear of rejection is very real, but in reality, rejection already happens and the exclusion / inclusion criteria is not explicitly talked about. The counterpoint was made that by having a set of guidelines, there is more transparency in what gets excluded and it can be left to the community to adjust that line in the sand rather than leave it to the opinions of a few people behind closed doors.

3.3 Revolutionizing Peer Review with Empirical Standards

Paul Ralph (Dalhousie University – Halifax, CA)

License © Creative Commons BY 4.0 International license
© Paul Ralph

Main reference Paul Ralph, Sebastian Baltés, Domenico Bianculli, Yvonne Dittrich, Michael Felderer, Robert Feldt, Antonio Filieri, Carlo Alberto Furia, Daniel Graziotin, Pinjia He, Rashina Hoda, Natalia Juristo, Barbara A. Kitchenham, Romain Robbes, Daniel Méndez, Jefferson Seide Molléri, Diomidis Spinellis, Mirosław Staron, Klaas-Jan Stol, Damian A. Tamburri, Marco Torchiano, Christoph Treude, Burak Turhan, Sira Vegas: “ACM SIGSOFT Empirical Standards”, CoRR, Vol. abs/2010.03525, 2020.

URL <https://arxiv.org/abs/2010.03525>

Scholarly peer review is the practice of subjecting a scholarly work (e.g. a manuscript) to the scrutiny of one or more experts (e.g. to decide whether to accept the manuscript for publication). Empirical research consistently demonstrates that scholarly peer review is ineffective, unreliable, and prejudiced. In principle, the solution is to move from contemporary, unstructured, essay-like reviewing to evaluating an artifact against an unambiguous, standard using a checklist. Therefore, a task force of over 50 leading scholars created natural-language models—called “empirical standards”—of the software engineering research community’s expectations for different popular methodologies (e.g. case study, controlled experiment). These living documents, which should be continuously revised to reflect evolving consensus around research best practices, will improve research quality and make peer review more effective, reliable, transparent and fair. This talk will include a demonstration of reviewing tools developed based on the empirical standards.

3.3.1 Discussion

There are many problems with peer review as it exists today; many of which revolve around individuals making their own set of criteria or introducing biases to the process. The seemingly random nature of acceptance and rejection can impact on people’s lives and even

sway how decisions are made in public policy. This can plausibly be detrimental if a potentially impactful paper is rejected or less rigorous paper is accepted, but came to a fallacious conclusion.

The presented reviewing tools act as a sort of “checklist” for authors and reviewers to determine if the paper meets a set of guidelines. The attributes within the guidelines are broken down into essential and desirable attributes which can be established on the backend and it has the flexibility to change based on the type of methodologies being used. Much of the discussion was centered around the flexibility of the system and potential features and changes.

Overall, this was a fairly central talking point throughout the seminar, as the tool provides an open source option for introducing customizable guidelines in a straightforward manor that could be fit into a review process.

3.4 Registered Reports in Computer Science: Why Bother?

Neil Ernst (University of Victoria, CA)

License  Creative Commons BY 4.0 International license
© Neil Ernst

Main reference Neil A. Ernst, Maria Teresa Baldassarre: “Registered Reports in Software Engineering”, arXiv, 2023.
URL <https://doi.org/10.48550/ARXIV.2302.03649>

Registered reports are scientific publications which begin the publication process by first having a detailed research protocol, including key research questions, reviewed and approved by peers. Subsequent analysis and results are published with minimal additional review, even if there was no clear support for the underlying hypothesis, as long as the approved protocol is followed. Registered reports can prevent several questionable research practices and give early feedback on research designs. In software engineering research, registered reports were first introduced in the International Conference on Mining Software Repositories (MSR) in 2020. They are now established in three conferences and two pre-eminent journals. We explain the motivation for registered reports, outline the way they have been implemented in software engineering, and outline some ongoing challenges for addressing high quality software engineering research.

3.4.1 Discussion

Registered reports are a common part of the submission and review process in the medical and psychology fields, and has recently started to be established in a few computer science-related conferences and journals: MSR / EMSE, Transactions on Software Engineering and Methodology (TOSEM) and a special issue of Computer Science Education (CSE)). This method starts before any data has been collected and has the benefit of getting some feedback at an early stage of the research which can save time, money and be very helpful in getting insights and opinions outside of the authors’ research groups. The other benefit is in reducing or eliminating under-powered or selectively reported findings. The current design involves two phases: Phase 1 where you declare what you would like to do and then Phase 2 where you report on the findings.

Some of the raised concerns with this process involved problems like a paper being published as Phase 1 and then being scooped before the results are published or a common sentiment is that only positivist philosophies matter (e.g., significance testing, falsification). This is not true in fields like medicine and is specifically in place so that results cannot just

be cherry picked. Similarly, without a requirement to declare methodological design and analysis ahead of time, there is a risk of focusing on novelty and significance over whether the research is sound. Computer science as a field is currently at risk of various questionable research practices and these are sometimes common at many venues.

In the presented case, the Phase 1 review process happened at MSR and then Phase 2 reviews were done in the EMSE system. Some pitfalls that came up during this were:

- Reviewers at MSR had the burden of accepting for a top journal (EMSE).
- Hard to manage reviewer discussion between multiple systems.
- Reviewer continuity between phases.

Something similar is being introduced at TOSEM, but both phases would be done under the same journal which would simplify many of the above issues. One of the major concerns is the additional burden this places on reviewers and whether or not there is the proper bandwidth to keep up with the increased capacity of reviews. This would effectively introduce two rounds of review per paper. On the other hand, much of what is in Phase 2 was already written and reviewed in Phase 1. For example, in the medical community, often times the pre-registration paper simply leaves the results section blank. This highlights the importance of reviewer continuity, as it would be quicker to review the Phase 2 submission without having to catch back up about all of the details that were already reviewed.

Another potential problem is with judgements on if a paper is important early in the research design phase. It becomes a balancing act where one side of the scale is to only let in “important” papers, which are subjective opinions and can lead to problems with bias or gatekeeping, and the other side is to allow in papers based on “soundness”, which has potential to create a flood of “sound” but uninteresting, via some subjective criterion, research. For conferences, there are much harder limits on the number of submissions and page count, which makes this a problem somewhat unique to computer science’s conference journal model as compared to some of the other fields that have implemented evidence standards.

3.5 Criteria and Scrutiny: Computing Education Research

Kate Sanders (Rhode Island College – Providence, US), Joseph Maguire (University of Glasgow, GB), Monica McGill (CSEdResearch.org – Peoria, US)

License © Creative Commons BY 4.0 International license

© Kate Sanders, Joseph Maguire, and Monica McGill

Main reference Marian Petre, Kate Sanders, Robert McCartney, Marzieh Ahmadzadeh, Cornelia Connolly, Sally Hamouda, Brian Harrington, Jérémie O. Lumbroso, Joseph Maguire, Lauri Malmi, Monica M. McGill, Jan Vahrenhold: “Mapping the Landscape of Peer Review in Computing Education Research”, in Proc. of the Working Group Reports on Innovation and Technology in Computer Science Education, ITiCSE-WGR 2020, Trondheim, Norway, June 15-19, 2020, pp. 173–209, ACM, 2020.

URL <https://doi.org/10.1145/3437800.3439207>

In 2020, a working group was convened at the Innovation and Technology in Computer Science Education (ITiCSE) conference, led by Marian Petre, Kate Sanders and Robert McCartney on Mapping the Landscape of Peer Review in Computing Education Research (CER). The working group considered 17 venues, including CER conferences and journals, as well as overlapping conferences in Software Engineering and Human Factors. In this presentation, we consider some of the common review criteria observed across venues as well as some of the ethical concerns that emerged in peer-review and the process itself. In the

present talk, these elements are considered through the lens of excerpts and vignettes drawn from conference chairs and journal editors interviewed by the working group that reflect aspects of the conversations and debates that have happened during week at the seminar.

3.5.1 Discussion

After a retrospective on the criteria, ethical concerns and scrutiny involved in the review, much of the discussion was about continued topics like importance versus soundness, what types of comments should and should not be allowed and should we have a system for reviewing reviewers. The general sentiment is that more of the review process with evidence standards in place should be focused on the soundness, rather than off the cuff opinions on importance from reviewers with varying degrees of expertise and bias. If reviews were done in a fashion closer to Paul Ralph's work (Section 3.3), there would be more emphasis on determining if a paper contained the essential and desirable attributes for acceptance and less room for potentially ethically interactions between reviewers and authors. Lastly, it was discussed whether that reviewing tool could be enhanced to include a way to review reviewers, as that may provide some valuable data and create transparency in the quality of reviews.

4 Working Groups

4.1 APA JARS – Quantitative

In the first breakout session, participants formed small groups and reviewed the APA JARS guidelines for quantitative research. APA JARS provides a concise guide on what needs to be reported in each section. The overall goal is to improve the scientific rigor of peer-reviewed papers by providing requirements that support clear and transparent research. They can also work as a useful learning tool for novice researchers.

4.1.1 Discussion

After going through the APA JARS sections on quantitative guidelines in breakout groups, there was a larger discussion with the whole group to tie everything together.

First, it was discussed the types of audiences that would be impacted by the inclusion of reporting standards. These include, but are not limited to: authors, decision makers (e.g., reviewers and editors), novice researchers, readers, machine readers, lawmakers and the public. While there may be some initial turbulence in training those closely working on the writing and reviewing process to adhere to the guidelines, there are benefits to a wide audience by producing work that has a higher degree of clarity and transparency.

Overall, it was felt that there were pros and cons to this type of reporting standard. Some of the concerns were in needing to expand the standards further to account for the various types of research done in computer science and computer science education. APA JARS is centered around human studies, but standards in CS would need to be broadened to include things like graphics papers with timings or benchmarks. Others disagreed, as benchmarks are clearly quantitative, so differences with existing standards may not be large or valuable.

Furthermore, a major concern is that reporting standards will be used as a checklist to reject a paper. There may still be value in a paper even if it does not follow the standards perfectly. For example, in the medical community, they track the impact that CONSORT

has had overtime; analyzing what got better, rather than cutting off strict requirements. Introducing standards at a venue as optional with a plan in place to collect metrics on its adoption could be a good way to ease the community into using standards and down the road report on how and what it has changed.

The structured aspect of APA JARS was a hot topic of discussion. There were a handful of pros and cons mentioned about strict structure in the standards. Some participants felt that having strict requirements would hinder the style of the paper. APA JARS has guidelines on sections and what should be included in each, as well as requirements on a structured abstract which has very specific requirements on labeling and what is reported in it. The discussion on structured abstracts noted that it is very efficient and having the findings clearly labeled upfront is useful, but there was some concern that you miss out on telling a story that is helpful in attracting the reader. Some journals have gone the route of having both structured and general public abstracts so that detailed and easy to read options are available, which can help to broaden the audience that can consume the information. Some of the other benefits of having requirements on structure include the ease of finding information between different papers, as it is always reported in a specific section. Additionally, requirements on structure can be helpful in mining information across a larger set of papers.

4.1.2 Conclusions

Overall, the consensus seemed to be that reporting standards could be a beneficial addition to the field, although they may need to be adapted to better align with each community. The main takeaway was the importance of how it is introduced to the community. This will be vital in not ostracizing anyone while still uplifting the quality of reporting for the community as a whole.

4.2 APA JARS – Qualitative

Next, participants worked with their breakout groups to go over the qualitative and mixed-methods sections of the APA JARS guidelines before having a larger discussion with the whole group.

4.2.1 Discussion

The first note was the difference in the qualitative guidelines providing separate sections for guidance to reviewers and authors. Many interpreted this as sort of a “defensive mechanism” to instruct reviewers to not just reject a paper. There seems to be more fear of not meeting the standards, and when compared to the quantitative guidelines, the depth of detail was much shallower. It could be that the qualitative side is still being built out, which is plausible given the qualitative standards are new, and in the interim it was easier to ask reviewers to take more consideration. Another note was that there may be a problem with cross-paradigm reviewing where reviewers with a background in quantitative research were making judgements about qualitative submissions without fully understanding it.

While the quantitative side was broken down into many different methodology guidelines, the qualitative standards were condensed to just one. Qualitative research is very common in computer science, especially in computer science education, and more care will be necessary

when developing the standards than what is in APA JARS. In particular, there are a wide variety of approaches to inquiry, some include:

- Positivist: Makes hypotheses and gathers evidence to support or refute them. Restricts yourself to empirical/observable data. No cognitive processes because you cannot observe them. Rely on behavioral data.
- Post-positivist: Accepts that reality can only be known imperfectly. There's observed bias.
- Feminist: Capturing perspectives of marginalized peoples. Acknowledges complexity of social life.
- Intersectionality: People's experiences are different with combinatorial composition of their demographic attributes.
- Postmodern: Personal perspectives over truth.
- Constructivist: People construct their own reality through inquiry.
- Critical: Recognizing that the default system in society is biased toward able-bodied, white males. The system must be dismantled before we can be equal.

Each of these have foundations, some participants claimed, are different and the quality criteria for judging them differ significantly. From this perspective, the standards will need to be broken down into a set of guidelines that more closely matches each, and there needs to be more guidelines on how a reviewer should approach criticism for these papers. For example, they will need to have the expertise to recognize the philosophy being used and not make judgements on the paper based on a different philosophy. Others, however, disagreed. Just because a paper meets a philosophical, and subjective category, has no bearing on whether that paper should have an introduction or state research questions. Many commonalities can and do exist, despite paradigmatic claims.

If many templates are necessary to accommodate such a variety, there is a problem with who controls the standards. It could be an issue that those in charge could be seen as "gatekeeping" if they do not have standards for a particular philosophy. There needs to be a mechanism for introducing additional templates and allowing the community to give feedback and adjust the guidelines. In contrast, if paradigms exist at different levels of rigor, free discussion needs to be had for whether that rigor is sufficient for publication. For example, if a paper simply claimed to be part of a paradigm that did not require evidence at all, most scientists would find this lacking. How to manage this natural balance and tension was not clear.

4.2.2 Conclusions

Overall, the APA JARS standards for qualitative research were useful in providing a gap analysis of what it was missing compared to what the computer science subfields would need. They were a bit more lenient than the quantitative standards, but that may have been due to being relatively new and still under construction. A lot of good points were made about how these standards would need to, arguably, be broken down and maintained if they were to be adopted at computer science venues.

4.3 What Works Clearinghouse

The next breakout session consisted of forming small groups, similar to the previous sessions, and going through various sections of the What Works Clearinghouse (WWC) standards. These were established by U.S. Department of Education as a way to identify studies that

meet specific thresholds of evidence. This helps educators, policymakers, researchers and the public to understand the effectiveness of education programs and interventions, and ultimately plays a role in determining grant funding.

4.3.1 Discussion

When comparing APA JARS to WWC, it becomes clear that they are meant for different purposes. One is reporting standards for publications in general and the other is related to public policy and grants. The degree of detail and breakdown of tiers of evidence is not a part of computer science culture.

Most of the discussion centered around how would we introduce a set of standards like these into the field. Some of the suggestions included forming a working group or workshop just to give people a place to try it out and get feedback. Introducing standards as required, but not a cause of rejection during a startup period is another option. This would give time for authors to get feedback on their papers to understand what they didn't meet on the requirements and improve, as well as giving reviewers time to get used to the guidelines before having an impact on acceptance. Some discussed the route of making the standards optional, which is the direction that TOCE went. Both have their benefits, completely optional at the start is easier to get the community on board with, and then the expectation is that as more people choose to go in that direction, if people like that style of paper they will adopt it into their practices as well. The requirement with no impact at the start is probably going to be harder to get approval from the community and risks pushing people away, but it also forces authors and reviewers to become familiar with the standards while not rejecting them initially.

4.4 Take a Paper and Convert It

In this session, seminar attendees split up into breakout groups and were tasked with selecting one of their already published works and going through the relevant APA JARS guidelines. While doing this, they were tasked with making several classifications:

1. Information that was required in the guidelines was present and in the required section.
2. Information that was required in the guidelines was present but in a different section.
3. Information that was required in the guidelines was not present but was recorded and could have been included.
4. Information that was required in the guidelines was not present and was not recorded.

4.4.1 Discussion

This activity gave a more hands on interaction beyond reading through the guidelines. By putting the requirements in the APA JARS guidelines through the lens of existing works that had been published by the attendees, it became a bit more apparent which parts were essential and which parts might not be necessary for inclusion. This is similar to how the guidelines for APA JARS were formed by first looking at CONSORT as a model and fitting it to the needs of the psychology field. Some of the main takeaways were:

- It simplifies the process of having to remember what needs to be included because there's a list to remind you.
- It would not be too difficult to have written the papers with these standards if they were asked to.

- The standards would work well for teaching novice researchers about writing for publication.
- Not everything listed in APA JARS seems necessary to be in the actual narrative of the paper.
- The requirements on section content and section headers may be met with argument of reducing freedom of expression.

The paper structure requirements were heavily discussed. The two competing viewpoints were largely that the structure will take away academic freedom, which was also a common argument historically in other fields. The other is the benefit of consistency between papers. While the heavy structural requirements of a standard like CONSORT may seem a bit jarring (e.g., some subsections breakdown to headers with one sentence), it can have the benefit of increasing the speed and comprehensibility when reading through many papers if the relevant content is exactly where you expect it to be. Further, standardized sections ease automated analysis. Using this activity as an example, many found it that it was time consuming as a reviewer to take the checklist of requirements and sift through the paper looking for each piece, whereas with a structured requirement it would have been simple to run down the list.

4.5 Breakout: Mixed-Methods

While focus has been placed on quantitative and qualitative studies, there's the additional problem of how to handle mixed-methods studies. There are several types of mixed-methods studies that were taken into consideration. For example, there are mixed-methods studies that collect quantitative data in an A/B experiment followed by a survey and interview to gather qualitative data about a participant's experience. Also, there are triangulation studies which are common in the formative design of tools which follow a waterfall approach of interviews, surveys, and validation with user studies.

The main concerns were about the amount of page bloat that would come from reporting on every detail in a series of experiments in just one paper and the complexity of the guidelines becoming too much of a hurdle or barrier to entry. The group felt that a structured appendix or supplemental material might be a good location for the required information while keeping the page length short and still allowing the author to have some freedom with the narrative style of the paper. It was suggested to build upon this with a fork of Paul Ralph's work (Section 3.3); this would allow for the flexibility to create minimum viable templates based on the community's unique needs. That system could then be used by the author to input the information that was not included in the paper itself and then the structured appendix could be auto-generated. While this would create another set of documents that reviewers would need to go through, they would already need to verify that the information is included and it would all be in one place.

4.6 Breakout: Measuring Effects

The topic of this breakout was on how to measure the effect of implementing evidence standards. If a venue were to make use of a checklist like from Paul Ralph's talk (Section 3.3), even if the standards were introduced as an optional recommendation, we could measure the percentage of papers that passed that test of rigor and compare that year to year to see if the optional recommendations were sufficient to increase adoption. Another suggestion was to

try to get a measurement of “happiness factor” or sense of belonging in the community from authors and reviewers to see how that changes over time. If part of the reason for enforcing standards is to improve people’s trust in the system, then getting a measure before and after could be beneficial. This could range from satisfaction with the submissions process, their experience with the guidelines or whether they felt the reviews they received were fair. Another metric for reviews could be to have conferences release review data and report on inter-rater reliability. This would provide transparency to the community and there could be a bit more retrospective year to year on what changes were made at the conference and the impact it may have had.

4.7 Breakout: Community Engagement

Getting the community on board is a key piece of successfully implementing evidence standards. Compensation and motivation were a big factor. Providing cash incentives for training reviewers and paying them for the work they put in. One option for training is Designing Empirical Education Research Studies (DEERS) [1]. For authors, it could help motivate them to adhere to optional standards if a badge system was put in place to mark papers that follow the standards. And for venues, one path to motivating them to change could be to focus efforts at highest-quality venues first and show whether it has an impact on things like quality, clarity or transparency of work, and if the community likes the changes. Smaller venues may pick up on that and be motivated to change as well.

Reducing opposition is important so that members of the community do not feel like they no longer belong. It is important to be transparent in the messaging and get information about the changes out in various ways, such as running panels at venues and writing op-eds. Beyond spreading information, allowing the community to be able to contribute to and have some control over the standards is essential in making sure they work well for the types of papers at their venue. Deadlines for adoption should be far enough in the future and are stair-stepped to allow for adaptation, so that year to year, the standards can be contributed to and changed to be a proper fit.

References

- 1 Carver, Jeffrey C. and Heckman, Sarah and Sherriff, Mark. Designing Empirical Education Research Studies (DEERS). Accessed on December 12, 2022 from <http://empiricalcsed.org/>.

4.8 Breakout: Operationalizing

There needs to be a focus on implementing things like peer review into technology. Ideally, this could be done in an existing technology, like OpenReview[1], as long as it is open sourced. Unfortunately, many of the existing systems that are commonly used, like EasyChair[2], are closed and do not support the types of things from Paul Ralph’s talk (Section 3.3). An open source system would allow for easier adoption between different communities and provide the ability to include data gathering elements such as the previously discussed “happiness factors” or inter-rater reliability measures for reviews (Section 4.6).

References

- 1 OpenReview. Accessed on December 12, 2022 from <https://github.com/orgs/openreview/repositories>.
- 2 EasyChair. Accessed on December 12, 2022 from <https://easychair.org/>.

4.9 Next steps: What Form of Actions Will We Take?

The expectation is that evidence standards will not be something that can be put into place across the entire field quickly, nor should that be the goal. Instead, one goal is to make some movement toward positive changes that may help with acceptance incrementally. Toward this goal, the group discussed a variety of conferences in computer science, software engineering and computer science education where some of these changes could be brought up to program chairs, steering committees or town hall meetings. Following a similar path to what was done at TOCE, most recommendations will involve offering optional guidelines as a first step to get authors', reviewers' and editors' a chance to trial the changes and adjust the requirements. Additionally, a dissemination plan was discussed to try to give some information on evidence standards to a broader audience.

4.9.1 Conferences and Journals

- Koli: The information discussed at the seminar will be brought up at the next PC meeting. Koli has two tracks, systems and tools, and the tools track might be a good candidate for introducing a rubric using a fork of Paul Ralph's work. This could help to build up this style of review in the community and get some feedback.
- ETRA: Similarly, this information will be brought up at a town hall next year to see what the community thinks. This could be a good way to introduce it here first, and then use it as an example when bringing it up at other CHI sponsored conferences that might be interested in this style of standards.
- TOSEM: This will be presented at a board meeting with the intention of proposing a special topic on Human Factors in Software Engineering which could be used to test the reporting standards.
- ICER: It will be proposed to adapt the structure of the reviewing form to use Paul Ralph's work and have a subset of papers go through this process.
- CSE: Currently has structured abstracts and recently added a track for registered reports. As a next step, there will be a suggestion added to the instructions for authors to look at JARS and encourage them to consider following it where appropriate.

4.9.2 Dissemination

- Position Papers: All of the participants at the seminar were interested in contributing on a position paper that would be submitted to Communications of the ACM (CACM).
- Panels: Plans were discussed to set up a CHI panel on this subject. Initial ideas include presenting something similar to the activity from Section 4.4, which converted an already published paper into APA JARS format, and then let the audience look through the differences then discuss concerns with the panel.

Participants

- Brett A. Becker
University College Dublin, IE
- Andrew Begel
Carnegie Mellon University –
Pittsburgh, US
- Michelle Craig
University of Toronto, CA
- Andrew Duchowski
Clemson University, US
- Neil Ernst
University of Victoria, CA
- Arto Hellas
Helsinki University of
Technology, FI
- Christopher D. Hundhausen
Oregon State University –
Corvallis, US
- Ciera Jaspan
Google – Mountain View, US
- Timothy Kluthe
University of Nevada –
Las Vegas, US
- Juho Leinonen
Aalto University, FI
- Joseph Maguire
University of Glasgow, GB
- Monica McGill
CSEdResearch.org – Peoria, US
- Brad Myers
Carnegie Mellon University –
Pittsburgh, US
- Andrew Petersen
University of Toronto
Mississauga, CA
- Mauro Pezzè
University of Lugano, CH
- Paul Ralph
Dalhousie University –
Halifax, CA
- Kate Sanders
Rhode Island College –
Providence, US
- Andreas Stefik
University of Nevada – Las
Vegas, US
- Claudia Szabo
University of Adelaide, AU
- Jan Vahrenhold
Universität Münster, DE
- Titus Winters
Google – New York, US
- Aman Yadav
Michigan State Universit –
East Lansing, US

