DAGSTUHL
REPORTS

Volume 11, Issue 9, October 2021

ISSN 2192-5283

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

Publication date
April, 2022

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

Digital Object Identifier: 10.4230/DagRep.11.9.i

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

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

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

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

Report from Dagstuhl Seminar 21401

Visualization of Biological Data – From Analysis to Communication

Edited by

Karsten Klein[1], Georgeta Elisabeta Marai[2], Kay Katja Nieselt[3], and Blaz Zupan[4]

1 **Universität Konstanz, DE,** karsten.klein@uni-konstanz.de
2 **University of Illinois Chicago, US,** gmarai@uic.edu
3 **Universität Tübingen, DE,** kay.nieselt@uni-tuebingen.de
4 **University of Ljubljana, SI,** blaz.zupan@fri.uni-lj.si

Abstract

Technological advancements in biology allow us to collect and generate a large quantity of data and pose a significant challenge to data interpretation and understanding. Addressing this challenge requires a blend of methodology from data visualization, bioinformatics, and biology. This methodology encompasses perception and design knowledge, algorithm design, techniques for analyzing and visualizing big data, statistical approaches, and specific domain knowledge for different application problems. In particular, it is essential to develop robust and integrative visualization methods combined with computational analytical techniques and approaches to communicate the outcomes visually. The purpose of Dagstuhl Seminar 21401, "Visualization of Biological Data – From Analysis to Communication," was to bring together researchers from various fields to discuss the state of the art, to debate means of advancing science in the field of visualization of biological data, and to foster the development of our international community.

Seminar October 3–8, 2021 – http://www.dagstuhl.de/21401
2012 ACM Subject Classification Applied computing → Bioinformatics; Applied computing → Life and medical sciences; Human-centered computing → Visualization
Keywords and phrases Bioinformatics, biology, Imaging, interdisciplinarity, Omics, Visual analytics, visualization
Digital Object Identifier 10.4230/DagRep.11.9.1
Edited in cooperation with Beauxis-Aussalet, Emma

1 Executive Summary

Kay Katja Nieselt (Universität Tübingen, DE)
Karsten Klein (Universität Konstanz, DE)
Georgeta Elisabeta Marai (University of Illinois Chicago, US)
Blaz Zupan (University of Ljubljana, SI)

Advances in technology have turned biology into data-driven research. High-throughput and high-resolution techniques help us generate and collect vast amounts of data to be explored, analyzed, and turned into knowledge or actionable models. This abundance of biological data creates a substantial challenge in processing, analysis, and modeling. A popular way to address this challenge is through visual representation and analysis of the data.

Creating compelling visualizations of biological data requires combining data visualization, bioinformatics, statistics, and computational biology. The field of biological visualization is interdisciplinary and involves collaboration between researchers from different areas.

Our aim with this Dagstuhl Seminar was to bring together researchers from multiple disciplines to discuss how to continue the interdisciplinary dialogue and foster the development of an international community concerned with biological visualization. We aimed to examine the state of the art and find areas to advance the research that might benefit from the joint efforts of all groups involved.

Our initial aim, expressed in the seminar proposal, was to explore the following four topics:

- data abstraction to support building custom visual tools of biological data,
- interactive analysis for biological data exploration,
- collaboration and communication through new tools,
- curriculum for teaching visualization in bioinformatics.

We have discussed these topics in the first two days of our five-day seminar and gradually came out with the following six working groups:

- facilitating cross-expertise exploration in explainable AI for multi-omics via visualization,
- visions for the lab notebook of the future,
- visual analytics of multilayer networks representing knowledge graphs,
- recommendations for designing visual, interpretable, and deep learning-based analytics pipelines in medical imaging,
- semantically enabled biomedical cartooming,
- a curriculum for the future of biological visualization.

Notice that with this new set of topics for the working groups, our seminar still closely followed our initial aim to explore the space of data abstractions, interactive analysis, and design of tools to support collaborations.

We have developed the schedule for the seminar based on our experience and expertise in previous successful Dagstuhl seminars. We aim to emphasize the balance between prepared talks and panels and breakout groups for less structured discussions focused on a selection of highly relevant topics. Three types of plenary presentations were available to participants who had indicated an interest in presenting during the seminar:

- Overview talks (20 minutes plus 10 minutes for questions)
- Regular talks (10 minutes plus 5 minutes for questions)
- Panel presentations (5 minutes per speaker followed by a 20 – 25 minute discussion)

The breakout groups met multiple times for several hours during the week and reported to the overall group on several occasions. This format successfully brought bioinformatics and visualization researchers onto the same platform and enabled researchers to reach a shared, deep understanding through their questions and answers. It also stimulated fruitful discussions that all participants deeply appreciated.

We have organized the seminar during the COVID-19 pandemic. Due to various regulations and quarantines, slightly above half of the participants attended in person, while the other participants attended online. The meeting took the hybrid form, and we thank Dagstuhl for equipping the seminar rooms with suitable hardware. Still, we found the organization of the hybrid meeting challenging, to say the least, as it imposed constraints on the discussion and engagement of everyone. An all-important part of Dagstuhl's experience is off-line meetings during meals, trips, or long evenings, which online participants miss. Thus, also following the responses in the participants' survey, we would endorse their recommendation that Dagstuhl should return to the previous, non-hybrid format of seminars once the pandemic stops.

This report describes in detail the outcomes of our meeting. At the present stage, the outcome includes a set of white papers summarizing the breakout sessions, overviews of the talks, and an emerging detailed curriculum for future biological and medical visualization eduction.

Acknowledgements

We would like to thank all participants of the seminar for their contributions and lively discussions; we also would like to thank the scientific directorate of Dagstuhl for providing us with the opportunity to organize this seminar. Finally, the seminar would not have been possible without the untiring help of the (scientific) staff of Dagstuhl, including Ms. Jutka Gasiorowski and Ms. Susanne Bach-Bernhard.

2 Table of Contents

3 Overview of Talks

3.1 Embracing the complexity of reality

Jan Aerts (Hasselt University – Diepenbeek, BE)

In this talk, we briefly go over what explainable AI could mean for data analysis, and explain how our own approach of topological data analysis (TDA) can help with that endeavor by providing contextual information. We illustrate this through two use cases: analysis of polysomnography data, and supporting patient/clinician discussion on cancer treatment.

3.2 Collaboration and Communication in Science (Using Data Visualization)

Andreas Bueckle (Indiana University – Bloomington, US) and Katy Börner (Indiana University – Bloomington, US)

As science becomes more complex, communication across interdisciplinary teams becomes ever more essential to solve complex problems arising in disease and health, the physical laws of nature, and the functioning of human societies. In this talk, we present three ongoing efforts that aim to empower researchers and domain experts to communicate and collaborate on the challenges of our time: First, we present a paper on visualizing big science projects (`https://doi.org/10.1038/s42254-021-00374-7`) using a dataset of 13,893 publications and 1,139 grants by 21,945 authors cited more than 333,722 times. The work aims to communicate the global and interdisciplinary reach and impact of big science projects and their evolution via distinct project phases. Second, we present user interfaces developed for the Human Biomolecular Atlas Program (HuBMAP), a multi-year, multi-million, international effort funded by the National Institutes of Health (NIH) with the goal "to develop an open and global platform to map healthy cells in the human body" (`https://commonfund.nih.gov/hubmap`). Specifically, we show how the Registration User Interface (`https://hubmapconsortium.github.io/ccf-ui/rui/`) and the Exploration User Interface (`https://portal.hubmapconsortium.org/ccf-eui`) have been developed as web-deployed, 3D tools for users to register and browse through human tissue blocks from more than a dozen organs in their spatial and semantic context. Finally, we discuss the Scalable Precision Medicine Knowledge Engine (SPOKE) Visualizer, which allows users to explore a complex network of 3 million nodes and 30 million edges involving food, genes, proteins, diseases, and symptoms, among others through a series of network and map visualizations (`https://cns-iu.github.io/spoke-vis`). The site was designed for novice users interested to understand the coverage and quality of SPOKE data and expert users interested to analyze and optimize the interlinked knowledge graphs in SPOKE.

3.3 Explainable AI – On the importance of domain awareness for the creation of visual feedback on black box classifier decisions on imaging data

Katja Bühler (VRVis – Wien, AT)

Joint work of Katja Bühler, David Major, Dimitrios Lenis, Maria Wimmer, Astrid Berg

Recent research in AI addresses architectures and network types providing a certain level of interpretability by design. This is opening up new avenues for future research on how this information can be employed for creating visual feedback to create trust and provide transparency on the decision and its reliability to end users. However, in particular for imaging data, many of these novel methods are not yet widely applied to clinical imaging data in a real-world context.

On the other hand, classical "black box" classifiers for imaging data based on ANNs are well established, but do not provide any insight into the reliability and the reasons behind the result. This is preventing in particular a straight forward solution to visually communicate the causality between input data and the classifiers' decision to end user. Existing methods take different strategies to attribute salient regions but generally fail either in creating focused and intelligible feedback tailored to the specific task or come with the danger to operate out of domain, making the classifiers decision, and consequently also the visualization building on it, unreliable. I will highlight the importance of domain awareness in designing task specific and trustworthy visual feedback and present a novel method for trustworthy XAI for pathology classification on medical images based on two of our recent papers [1, 2]

References

1 Major D., Lenis D., Wimmer M., Sluiter G., Berg A., and Bühler, K. (2020), *Interpreting medical image classifiers by optimization based counterfactual impact analysis*, IEEE International Symposium on Biomedical Imaging, Iowa City, 2020.

2 Lenis D., Major D., Wimmer M., Berg A., Sluiter G., Bühler K. (2020) *Domain Aware Medical Image Classifier Interpretation by Counterfactual Impact Analysis.*. In: Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol 12261. Springer, Cham.

3.4 Explainable and Interactive AI: Overview and Open Research Challenges

Mennatallah El-Assady (ETH Zürich, CH)

URL https://explainer.ai/

Interactive and explainable machine learning can be regarded as a process encompassing three high-level stages:

1. Understanding machine learning models and data,
2. Diagnosing model limitations using eXplainable AI (XAI) methods, and
3. Refining and optimizing models interactively.

In this talk, I review the current state-of-the-art of visualization and visual analytics techniques by grouping them into these three stages. In addition, I argue for expanding our approach to explainability by adapting concepts, such as metaphorical narratives, verbalization, and gamification. I further introduce the explAIner.ai framework for structuring the process of explainable artificial intelligence and interactive machine learning, and operationalizing it through a TensorBoard plugin. Lastly, to derive a robust XAI methodology, I present some first steps to extract XAI strategies and mediums by transferring knowledge and best practices from other disciplines.

3.5 Impact of Model Prediction on Human Decision Making Confidence

Carsten Görg (University of Colorado – Aurora, US)

We conducted a pilot study to assess the impact of introducing model prediction in an existing clinical workflow for classifying Adamantinomatous Craniopharyngioma (ACP). We designed an experiment with 3 conditions: (1) viewing images only, (2) viewing images plus a model prediction, and (3) viewing images providing a model prediction and using the What-If-Tool (WIT) for exploring conterfactuals. In each condition we used 28 cases, about half ACP and non-ACP. Most cases had either an MRI or a CT image, some cases had both. We enrolled 2 expert subjects, a neurosurgeon and a neuroradiologist. We collected measures of performance, confidence and difficulty. Performance was similar across conditions but different between subjects. The confidence and difficult measures show the expected pattern of high confidence for not difficult cases, but also less confidence for non-ACP cases. Both experts found the similarity feature in WIT useful, but the tool itself too difficult to use. Limitations include the small number of subjects and limited scope of the dataset.

3.6 Designing a Topic-Based Literature Exploration Tool – A Neuroscience Exploratory Study

Lynda Hardman (CWI – Amsterdam, NL)

Exploring an extensive body of literature can be facilitated through topic-based, rather than article-based, interaction. Neuroscience researchers are interested in exploring relations between topics, such as anatomical regions of the brain and diseases that affect them. Given the three-dimensional nature of the brain, we postulate that supporting the exploration of relations between neuroscience topics in Augmented Reality could improve and extend existing literature exploration workflows. We identify visualization and interaction design requirements for distant reading of neuroscience literature using a user-centered approach. Using an existing concept analysis of tens of thousands of neuroscience papers, we designed an Augmented Reality environment to support distant reading of relations between brain regions and brain diseases.

3.7 Abstraction for Bioviz

Lawrence Hunter (University of Colorado – Aurora, US)

Abstractions convey groupings of facts as a unit, by omitting "details". To express knowledge, a system must select which "facts" to present. This selection is based both on the particular task the visualization is to support, and on the shared understanding of scientists. One useful way to make progress is with formal representations of biomedical knowledge. These formal representations are often in the form of knowledge graphs where the nodes (and sometimes arcs) are drawn from computational ontologies. These Knowledge Graphs (KGs) can be either tools to visualize data structured by the graph, or to visualize subgraphs that are relevant to particular biomedical tasks. One possible concrete challenge would be tools for the generation of molecular biology "cartoon" summary figures.

3.8 From Graphs and Hypergraphs to Maps and MetroMaps

Stephen G. Kobourov (University of Arizona – Tucson, US)

Relational datasets are often modeled with graphs and hypergraphs: objects become vertices and relationships become edges and hyperedges. Algorithms for graph and hypergraph visualization aim to represent such data in an effective and aesthetically pleasing way. From a theoretical point of view, the underlying problems give rise to algorithmic and complexity questions. From a practical point of view, building functional visualization systems is associated with questions of scalability and usability. GMap and MetroMaps are general visualization frameworks for utilizing familiar data representations and metaphors, such as geographic maps and metro maps. Such representations are more intuitive, as people are already proficient with such maps and standard interactions via panning and zooming.

3.9 Reproducibility and Reusability in Interactive Visual Analytics

Alexander Lex (University of Utah – Salt Lake City, US)

Interactive analysis is an important part of the data science process. It enables analysts to directly interact with the data, exploring it with minimal effort. Unlike code, however, an interactive visualization session is ephemeral and cannot be easily shared, revisited or reused. In this talk, I will sketch approaches to "Literate Visual Analysis". I will show how we can leverage provenance data of an analysis session to create well-documented and annotated visualization stories that enable reproducibility and sharing. I will also introduce work on inferring analysis goals, which allows us to understand the analysis process at a higher level. Understanding analysis goals enables us to enhance interactive capabilities and even re-use visual analysis processes. I will conclude by demonstrating how this provenance data can be leveraged to bridge the gap between computational and interactive environments.

3.10 Collaboration in a Test of Time Project

Georgeta Elisabeta Marai (University of Illinois Chicago, US)

RuleBender is an open-source system for the integrated visualization, modeling and simulation of rule-based intracellular biochemistry. RuleBender has been remarkably stable over the past ten years, and has been used by at least hundreds of computational biologists to do real and high impact research. It has been adopted at over 40 institutions as a research and educational tool. We reflect on the lessons learned from the design, development, and deployment of this successful system. We particularly emphasize the activity-centered design paradigm and the close interaction with domain experts that allowed RuleBender to better serve the needs of the systems biology community.

3.11 Proteomics Data – the Next Generation

Lennart Martens (Ghent University, BE)

This talk was derived from an ad-hoc question posed to the Seminar Participants regarding any input they might have on the visualisation (and analysis) challenges of a large amount of newly generated proteomics results, based on a proteome-wide analysis of post-translational modifications derived from a large and heterogeneous set of public proteomics data for human and mouse samples.

Computational proteomics has evolved extensively over the past decade, introducing the first widely successful machine learning approach with the Percolator algorithm for post-processing of identification data [1], and slowly expanding the capabilities of predictive algorithms for analyte behaviour during liquid chromatography separation and fragmentation. Despite clear advances in these areas, however, the adoption of such predictions in identification pipelines was very slow to non-existent. This was primarily due to the already quite capable traditional identification algorithms, especially when complemented with Percolator post-processing.

However, renewed interest in the use of the prediction of analyte behaviour was created by the field's diversification into more complex analyses such as metaproteomics, proteogenomics, immunopeptidomics, and open modification searches, all of which revealed inherent limitations in existing, traditional search engines [2].

As it turns out, machine learning based predictions are highly effective at solving the issues encountered in such complex DDA proteomics approaches, and the production of highly performant algorithms has soared as a consequence.

Interestingly, the availability of vast amounts of public data have also enabled a breakthrough in the types of machine learning algorithms that can now be employed on proteomics data, which has seen a surge in the application of complex neural networks (so-called deep learning approaches) in the field. When provided with enough data, these deep learning algorithms deliver extremely good predictions, which will in turn fuel a more sensitive and more robust interpretation of already acquired data.

In this lecture, our latest, cutting-edge developments in machine learning based algorithms for computational proteomics have therefore been briefly presented, alongside our applications of these algorithms to the complete re-analysis of all publicly available human and mouse proteomics data.

Moreover, the lecture shows our first efforts at making the results from these data widely available, which takes two main forms: the Scop3P system to show modifications in context on linear sequence and 3D structure [3], and the Tabloid Proteome approach to derive protein association networks from co-occurring proteins across different experiments [4, 5].

The resulting discussion proved highly interesting and informative, including the possibility of including the results set as a challenge in the 2022 BioVis Conference, the analysis of the results using models originally derived for linguistics work, and the application of upset visualisations and interactive histograms.

References

1 Lukas Käll„ Jesse D. Canterbury, Jason Weston, William Stafford Noble, and Michael J. MacCoss. *Semi-supervised learning for peptide identification from shotgun proteomics datasets*. Nature methods 4, no.11, pp.923–925, 2007.

2 Niklaas Colaert, Sven Degroeve, Kenny Helsens, and Lennart Martens. *Analysis of the resolution limitations of peptide identification algorithms*. Journal of proteome research 10, no.12, pp.5555–5561, 2011.

3 Pathmanaban Ramasamy, Demet Turan, Natalia Tichshenko, Niels Hulstaert, Elien Vandermarliere, Wim Vranken, and Lennart Martens. *Scop3P: a comprehensive resource of human phosphosites within their full context*. Journal of Proteome Research 19, no. 8, pp.3478–3486, 2020.

4 Surya Gupta, Kenneth Verheggen, Jan Tavernier, and Lennart Martens. *Unbiased protein association study on the public human proteome reveals biological connections between co-occurring protein pairs*. Journal of proteome research 16, no.6, pp.2204–2212, 2017.

5 Surya Gupta, Demet Turan, Jan Tavernier, and Lennart Martens. *The online Tabloid Proteome: an annotated database of protein associations*. Nucleic acids research 46, no.1, pp.581–585, 2018.

3.12 Bio+Med+Vis Education

Torsten Möller (Universität Wien, AT)

Joint work of Torsten Möller, Johanna Beyer, Jan Byska, Ingrid Hotz, Bara Kozlikova, Renata Raidon, Noeska Smit, Hsiang-Yun Wu

URL https://biomedvis.github.io/

In this talk we are detailing our efforts toward a curriculum for a Bio+Med+Vis course and program. In the context of the 2018 BioVis Dagstuhl and the 2020 Shonan Meeting, a working group has formed that developed the topic matrix further, and held a first Spring School to fill the materials which are now accessible to everyone for free online. We would like to brainstorm on how to continue this effort and solicit feedback from the BioVis 2021 Dagstuhl participants.

3.13 Teach (from) Your Users

Bruno Pinaud (University of Bordeaux, FR)

As a visualization researcher, I agree with Jeffery Heer: "The ultimate subject of the visualization research community is people not pictures" (Eurovis, 2019). I would like to present some lessons learned as a Principal Investigator (PI) for a multidisciplinary project, collaborating with domain experts in biological and digital humanities along with computer science colleagues. It is a delicate work to collaborate with experts you do not know, and who do not know what you can bring to them. They need money to produce data, while I need data to visualize. That is why such project unfortunately starts around its end.

A challenge, thus, is to make data experts tell you what they want as early as possible in the project, to avoid having yet another unused tool at the end of the project. So, first of all each side has to teach the other a bit about its domain. For instance, in visualization, we have mantras. They are good starting points, especially "Details first, show context, overview last" [1]. To conclude, never forget the golden rule: "Keep it simple".

References

1 Timothy Luciani, Andrew Burks, Cassiano Sugiyama, Jonathan Komperda, and G. Elisabeta Marai. *Details-first, show context, overview last: supporting exploration of viscous fingers in large-scale ensemble simulations*. IEEE Transactions on Visualization and Computer Graphics. 25(1), pp.1225–1235, 2019.

3.14 Exigent Visualization: When Good Enough is Better than Better

William Ray (Ohio State University – Columbus, US)

The Vis community usually talks about the dichotomy between Exploratory Data Visualization and Explanatory Data Visualization, and the contrast between the data-representation needs for discovering new things in data versus communicating something known in the data. Critically, we focus on optimizing completeness and correctness of data communication in these different modalities.

As a result of a clinical collaboration, we have recently encountered an interactive data visualization task (estimation of burn size in burn trauma) that does not fit either of these categories. Moreover, several decades of attempts to optimize the existing representations for completeness and correctness has resulted in the actual results in practice, getting worse rather than better.

A careful categorization of the different types of error that creep into the results from this visualization task, suggests that the practical results can be optimized by discarding the canonical optimizations to the visualization tool. In fact, making the tool fundamentally much less accurate, significantly decreases downstream errors by clinicians and can result in improved care.

We propose that a similar analysis of the different contributions to the total error observed in other visualization tasks may lead to similar counterintuitive opportunities to improve results.

3.15 Collaboration in BioMedVis

Timo Ropinski (Universität Ulm, DE)

Establishing fruitful collaborations in the area of BioMedVis, where visualization researchers and domain experts work towards a common goal, can be a challenging task. Besides the actual topics to be addressed, also the mechanics of collaboration need to be determined. While the prime goal is often to work on a common project, involving one or more full time researchers, this setup can seldom be obtained from the beginning. It is rather beneficial to initiate collaborations by starting with student projects of different levels. Furthermore, a balance between "give and take" must be determined.

3.16 Interactive exploratory analysis with iSEE

Charlotte Soneson (FMI – Basel, CH)

The iSEE (interactive SummarizedExperiment Explorer) R/Bioconductor software package (`https://bioconductor.org/packages/iSEE`), built on the R/shiny framework, provides a general-purpose graphical interface for exploring any rectangular dataset with additional sample and feature annotations, for example single-cell RNA-seq data. Users can create, configure, and interact with the iSEE interface, enabling quick iterations of data visualization. This facilitates generation of new scientific hypotheses and insights into biological phenomena, and empowers a wide range of researchers to explore their data in depth. iSEE also guarantees the reproducibility of the analysis, by reporting the code generating all the output elements as well as the layout and configuration of the user interface. The combination of interactivity and reproducibility makes iSEE an ideal candidate to bridge and complement the expertise of researchers, who are able to design flexible, accessible, and robust dashboards that can also be directly shared and deployed in collaborative contexts – connecting large data collections to broad audiences, thus further increasing the value of generated research data.

3.17 The process of designing, interpreting, and registering complex transcriptomic biomarkers for clinical oncology

Miha Stajdohar (Genialis – Houston, US)

We have modeled four distinct phenotypic subtypes of cancer patients. These subtypes represent the interplay of key biological processes in the tumor microenvironment and are clinically actionable. This is the first complex transcriptomic biomarker that is currently being registered with the FDA (USA) as a clinical trial assay. It will guide patient enrollment in the upcoming clinical trials, and the development of a companion diagnostic test (CDx) is underway. In this presentation, we will present the data science of design and validation of the classifier, visualization and interpretation of its predictions, and discuss the hurdles that we had to overcome to make our way into the clinical decision making.

3.18 The Code Monkey and The Parachuter: Reflecting on the Values of Interdisciplinary Collaboration in Visualization

Danielle Szafir (University of North Carolina at Chapel Hill, US)

Visualization prides itself on being a highly interdisciplinary field. We draw on techniques from a wide range of disciplines to design, implement, and evaluate visualizations. However, all too often, we find ourselves in collaborations that fail to truly contribute to multiple disciplines: we either build something exclusively for the domain (acting as the code monkey) or we stop contributing once the core visualization research question is solved (acting as the parachuter). I will reflect on contributions from my past work to explore the types of bidirectional contributions that interdisciplinary work can foster (or fail to foster) and examine how we might consider a range of ways visualization can benefit and be benefitted by cross-disciplinary collaboration.

3.19 What do visualizations explain with/for Explainable AI

Cagatay Turkay (University of Warwick – Coventry, GB)

This talk will start by exploring what explainability in AI means and why it matters. We will then cover some of the roles that visualization can play as a facilitator and enabler for explainability, as such roles emerge within the broader visualization and AI literature. As well, we will make a case for the human-centered nature of visualization, and how that is critical for thinking about explainability. The talk will then focus on the intersection between Explainable AI and Biological Data Visualization and explore advances, successes, failures, challenges and opportunities in this space.

3.20 XAI through a combination of interactions and workflows

Blaz Zupan (University of Ljubljana, SI)

URL https://orangedatamining.com

Visual analytics, a technique that combines exploratory data analysis and interactive visualization, can be an excellent tool for explainable AI (XAI). Consider single cell gene expression data, and a set of marker genes, for example. We can visualize these cells in t-SNE, expose those that express marker genes, allow users to select a cluster, perform differential expression analysis, and do Gene Ontology (GO) enrichment for explanation. The challenge is to generalize such pipelines to any biomedical dataset, and find the minimal set of components to implement it.

4 Working groups

4.1 Visions for the Lab Notebook of the Future

Jan Aerts (Hasselt University – Diepenbeek, BE), Jian Chen (Ohio State University – Columbus, US), Mennatallah El-Assady (ETH Zürich, CH), Alexander Koch (BioLizard – Gent, BE), Alexander Lex (University of Utah – Salt Lake City, US), James Procter (University of Dundee, GB), William Ray (Ohio State University – Columbus, US), Charlotte Soneson (FMI – Basel, CH), Granger Sutton (J. Craig Venter Institute – Rockville, US), Danielle Szafir (University of North Carolina at Chapel Hill, US), Cagatay Turkay (University of Warwick – Coventry, GB), and Blaz Zupan (University of Ljubljana, SI)

Our group worked on the concept of provenance in interactive data visualizations and insight generation. We kicked off our discussions with a roundtable where everyone introduced themselves and explained why they were interested in this topic. The participants' interests and motivations ranged from high-level issues and abstract methods to practical applications:

- Interactive data visualizations are under-respected in the field of computer science and misunderstood outside of the vis community.
- How can we communicate how and why we created a particular visualization?
- How do users, such as, for example, patients and doctors, interact with the data visualization tools we create?
- What are the best practices for creating interactive visualizations? Can we compile a list of recommendations?
- How can we make interactive tools more reproducible?
- How do we build tools that work for/with people instead of replacing them (e.g., computer in the loop rather than human in the loop)?

The first point of discussion was whether we should keep the discussion biovis-specific or make it more general. We decided to keep the discussion (and any follow-up efforts) general but apply the ensuing general principles and practical implementations to biological data visualization.

The basis of our group's further discussions was the lack of provenance in interactive data visualizations. We formulated a list of relevant questions around interactive bio(medical) visualizations on which we could build during our brainstorming sessions. These included:

- How do users, such as patients and medical personnel, interact with the data visualization tools we create?
- How can we make interactive tools more reproducible?
- How can we capture, store, and share user interactions? What can we learn from these interactions?
- How can we document research processes and decision-making processes?
- How can we guarantee that explanations in reports and publications are not post hoc rationalizing a biased decision? This applies to both human and AI decisions.
- How can we visually represent joint human and non-human (AI/ML) decision making?

What initially started off as a discussion on how we can capture and keep track of user interactions in interactive visualizations, quickly expanded to a discussion on how we can track, store and share the ways we arrive at insights in general. Many of us in the data

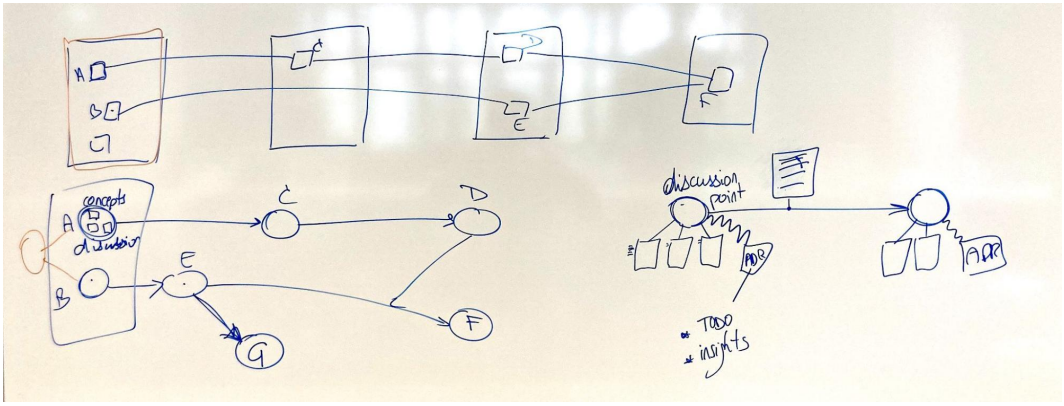


Figure 1 The provenance problem is portrayed as a network graph, in which the nodes represent decision points and the edges the processes that lead from one decision to the next (e.g. an algorithm or interactive notebook).

visualization field are accustomed to code-based notebooks, which are interactive, *in silico* equivalent of the biologist's lab notebook. We envisioned a tool that combines the best of both worlds: the interactivity, reproducibility, and scalability of the former and the flexibility and information density of the latter. We called this tool "the lab notebook of the future" (or, to lend it a certain futuristic air and for the sake of having a somewhat manageable deadline: "the lab notebook of 2030").

From a purely computational perspective, having provenance means that we need to record and store every step of the data visualization process. Any successful notebook must capture a sequence of views, insights, data points, and interactions that all together lead to a particular decision. There exist several technical approaches to this challenge, but we agreed that the flexibility of graph-based solutions makes them prime candidates (Figure 1). From a human perspective, having provenance means that we want to describe, explain, justify, adjust, and share the different steps taken to come to a particular visualization-driven insight.

High-quality provenance data allows for better retrospective exploration and analysis of provenance, which helps rescue the original intent of particular actions. It gives us a way to mine the cognitive trail created when using interactive tools. It also leaves an audit trail, which has several benefits in and of itself, from improving scientific reproducibility to boosting confidence in clinical decision making. Above all, provenance can help keep the agency in the visualization process. It ensures that a solution or visualization is arrived at actively, with intent. Interactive visualizations of (biological) data are inherently dependent on user input, as the visualization tool is not just delivering the same view regardless of user intent and actions.

We identified several hurdles that stand between us and perfect provenance. How do we, for example, deal with continuous, iterative analyses? Or layered hypotheses? Does provenance scale with the hierarchy and complexity of the data and knowledge, as in the case of biological data? Having provenance does not necessarily mean understanding it. How can we ensure that stakeholders have the same mental model when thinking about or discussing provenance? Depending on the complexity of an application, provenance data can quickly become unwieldy. They can also be difficult or impossible to interpret following system changes (e.g., software updates).

The end goal of introducing provenance in interactive data visualization tools is to empower the user. Achieving this goal will require a combination of human, algorithmic, and visualization solutions. The end goal of our Dagstuhl working group is to crystallize our discussions and ambitions into a (position) paper that describes the problem and details our proposed solution and to put them into practice by building "the lab notebook of 2030".

4.2 Recommendations for the design of visual, interpretable, and deep learning-based analytics pipelines in medical imaging

Katja Bühler (VRVis – Wien, AT), Guadalupe Canahuate (University of Iowa, US), Carsten Görg (University of Colorado – Aurora, US), Helena Jambor (Universitätsklinikum TU Dresden, DE), Georgeta Elisabeta Marai (University of Illinois Chicago, US), Timo Ropinski (Universität Ulm, DE), and Hagit Shatkay (University of Delaware – Newark, US)

Recent years have seen a constant increase in utility and multitude of medical images in clinical practice. Modern deep learning techniques have been developed and studied extensively in this context to cope with the availability and multitude of medical images. They promise to accelerate their use in clinical routines, like the diagnosis process and intervention planning and increase the reproducibility of clinical decision-making.

While the accuracy of these techniques is unmet, their applicability also leads to novel challenges. One such challenge is the interpretability of the AI decision process. When medical doctors are confronted with decisions made by AI, they are often not able to interpret the reasons behind such decisions. While various explainable AI techniques have been proposed to address this challenge, it is largely unclear how these techniques can be optimally used in the medical decision process. Deep learning experts design artificial neural network-based solutions that only focus on technical aspects, missing the end users' need for interpretability.

Visualization is a powerful communication channel for explainable AI, but existing visualization research is mainly decoupled from existing technical constraints and knowledge on what information can be delivered by XAI. Therefore, within this paper, we aim to build a bridge between both domains, providing a framework for visual XAI pipelines for image-based diagnostic workflows. We will initially review domain terminology as medical doctors communicate findings from medical images. Based on this terminology, we will then lay out so-called micro-tasks, as they are often solved by modern AI technologies, and link them to the needs of the medical experts. Finally, we will review recent explainable AI techniques and show how they can address these needs by visually communicating the results of the relevant microtasks.

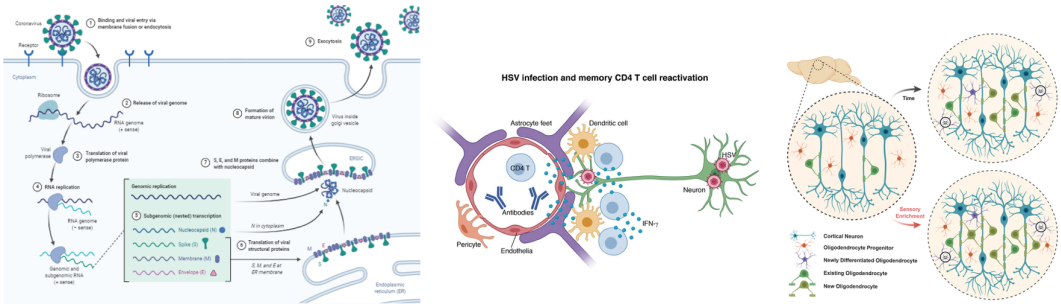


Figure 2 Examples of biomedical cartoons from `www.biorender.com`.

CARTOOMICS

Figure 3 Potential logo for the cartooning tool.

4.3 Semantically enabled biomedical cartooming

Lawrence Hunter (University of Colorado – Aurora, US), Emma Beauxis-Aussalet (VU University Amsterdam, NL), Nadezhda T. Doncheva (University of Copenhagen, DK), Lynda Hardman (CWI – Amsterdam, NL), and Martin Krzywinski (BC Cancer Research Centre – Vancouver, CA)

Originally constituted as a group to focus on semantic abstraction in bioviz, we split off from the main group to focus on a more specific task: Semantic support for drawing biomedical cartoons. Such cartoons are very common in molecular biology research publications and are used to summarize and communicate complex mechanistic hypotheses (Figure 2).

Existing drawing tools for producing such drawings and related interactive visualizations (e.g. Cytoscape) operate primarily on generic structures (e.g., networks, matrices) and provide domain-specific glyphs or templates (e.g., SBML or BioRender). Drawing operations are also generic (e.g., grouping or aligning drawing objects). They do not provide specific support to relate biomedical elements or concepts to the visual elements of the cartoons. Our idea is to exploit existing biomedical knowledge graphs to provide semantic support for a cartoon composition tool that would automatically link biomedical elements to templates of visual elements. We named such a tool Cartoomics (Figure 3).

Ontologically grounded biomedical knowledge graphs (e.g. PheKnowLator or SPOKE) open the potential for drawing tools to use existing knowledge graphs to help users create effective visualizations. We can use such visualization, for example, to explore the relationship of experimental results, compare it to prior knowledge, and communicate new findings. Starting from a set of experimental findings, the Cartoomics tools could exploit biomedical knowledge graphs to suggest visual elements to add to a cartoon ("semantic autocomplete"). Cartoomics could also allow the specification of semantically related entities to be manipulated as a group ("semantic templates", e.g., for aggregating or highlighting associated elements).

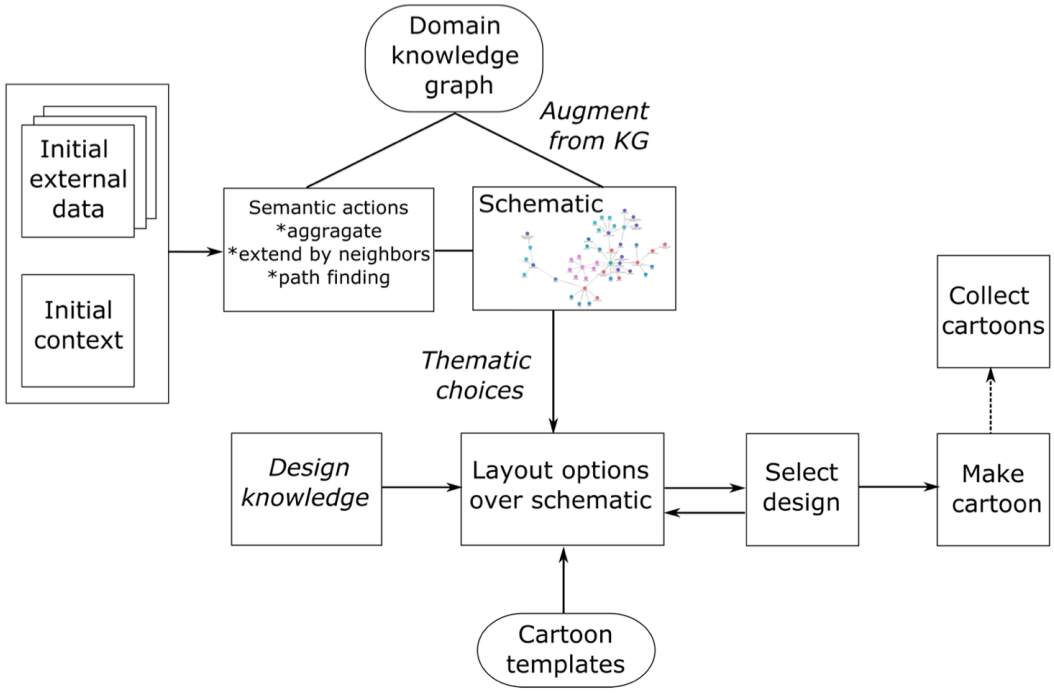


Figure 4 A conceptual diagram of a semantically enabled biomedical cartooning system.

Subsumption hierarchies in the ontologies that ground the knowledge graphs offer semantic zooming, e.g., for increasing or decreasing the level of detail. Other potential semantic actions include adding related concepts ("semantic neighbors"), following paths inferred from the knowledge graphs, embedding related concepts in aggregates, or exploring subgraphs with interactive queries.

We envisioned at least two possible application domains. In one, oncologists use genomic sequence data to make treatment decisions. Here, genomic data is combined with patient data and mapped to pathway diagrams, which are used to compare the likely effect of drugs. These diagrams vary from patient to patient but follow general patterns. The construction of these diagrams is time-consuming and requires expertise. Semantically enabled biomedical cartooning could potentially speed up the production of these diagrams and also improve their quality.

A second application domain would be interpreting transcriptomic, proteomic, or multi-omic experimental results. Here, a list of differentially expressed genes (e.g., proteins)needs to be understood (e.g., why these genes? how they relate to each other?). The investigation may combine contexts such as the description of an experimental manipulation (e.g. the addition of a compound). These elements need to be visually represented to communicate the main findings or hypotheses.

We drafted a conceptual design for such a system (Figure 4). The process begins from an existing domain knowledge graph and an initial input set of data and context (e.g., new experimental results). The tool would query the knowledge graph based on the initial input to produce a semantic schema, representing parts of interest in the existing knowledge graphs (e.g., with ontologically defined nodes and edges relevant to the inputs). The schema would be interactive and provide a variety of semantic affordances to modify the graph selected initially. Such modifications could include aggregation, extension by n-neighbors, pathfinding, subgraph matching, and even vector space aggregating neighbors using an embedding calculated from the graph.

When the user was happy with the semantic content of the selected schema and knowledge graphs, a second stage would solicit information about layouts of interest for the cartoons. We could draw the layouts from predefined cartoon templates or additional design knowledge graphs to recommend specific templates. The Cartoomics tool would allow the user to make thematic choices and to iterate through potential design alternatives. From there, automated processes could draw the final cartoons.

There was considerable enthusiasm among the participants for this idea. We plan to apply for grants, both in US and EU, to pursue the design and evaluation of the potential of such a cartooning system.

4.4 Visual Analytics of Multilayer Networks Representing Knowledge Graphs

Karsten Klein (Universität Konstanz, DE), Michael Behrisch (Utrecht University, NL), Andreas Kerren (Linköping University, SE), Stephen G. Kobourov (University of Arizona – Tucson, US), Michael Krone (Universität Tübingen, DE), Bruno Pinaud (University of Bordeaux, FR), and Falk Schreiber (Universität Konstanz, DE)

This working group originated from the discussions of the first breakout group on "data abstraction," which centered around abstraction for biomedical knowledge graphs. A *biomedical knowledge graph* is based on information about biological concepts and processes that are used to describe a complex biological system. It models corresponding biological relations between the involved entities and their effects on the system, such as protein interaction, functional associations between diseases and cellular processes, and gene expression. Due to the vast amount of information from life-science databases and publications, such a knowledge graph is usually huge, with millions of entities and relationships between them, resulting in the need for automated analysis and interactive exploration methods for knowledge extraction. Knowledge graphs can support a variety of biomedical applications, e.g., pharmaceutical or multi-omics applications [1].

The initial discussions of the "data abstraction" group centered around two different topics: (1) is a standard graph an appropriate abstraction for the underlying data, and (2) how to create the cartoon-like graphics common in life-science publications. Our subgroup first discussed general and conceptual aspects of data abstraction in the context of knowledge graph analysis to arrive at a common understanding. We defined a list of potential research projects that the group members are interested in, extending the original scope of data abstraction towards visual analytics of multilayer networks for the visual representation of knowledge graphs. The group consists primarily of visualization and network visualization researchers; hence our discussions were about appropriate abstractions for knowledge graphs, representation and visualization metaphors, network layouts, and algorithmic aspects. Notably, we agreed that multilayer networks (which capture different types of nodes and various types of relationships [2, 3]) might offer a more appropriate abstraction than standard graphs (made of one set of nodes and one set of links).

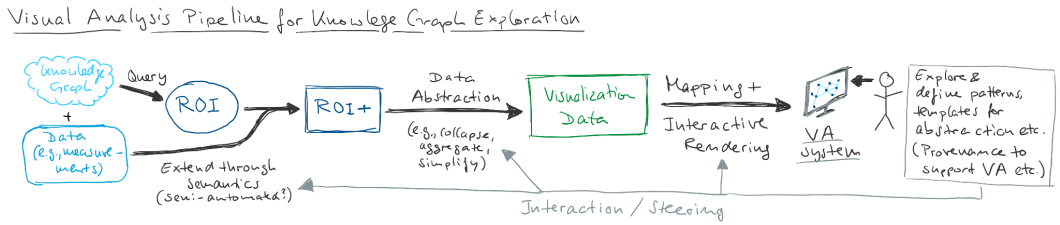


Figure 5 The visual analysis pipeline model used in the group's discussions.

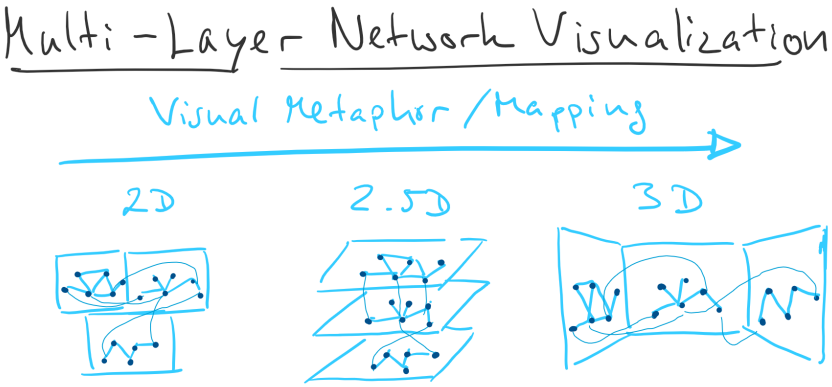


Figure 6 Visual metaphor classes for multilayer network visualization, comprising 2D, 2.5D, and 3D.

From the presentations of the domain experts, we derived a workflow model for knowledge graph analysis, which covers query-based subgraph extraction and semantic extension, abstraction, mapping on visualizations, and interactive exploration, see Figure 5 for getting an overview of this model. We defined several concrete research challenges that we briefly describe in the following paragraphs based on this workflow.

Challenge 1 – Visual Metaphors. This challenge investigates which visual metaphors are appropriate for representing multilayer networks, see Figure 6. Important characteristics of the metaphor include the dimensionality of the representation of the complete network (in 1D, 2D, 2.5D, or even 3D), the arrangement of layers, the encoding of entities, and abstractions. We want to design a human subjects study to evaluate different metaphors in combination with analysis tasks, which we can obtain from the existing task taxonomy described by McGee et al. [2, Chapter 4].

Challenge 2 – Interactive Exploration. The second challenge covers the impact of interaction and representation devices, e.g., immersive environments including 2D and stereo display, head-mounted display, and augmented/virtual reality [4]. Based on the results of our study for the first research challenge, we aim to compare different environment designs for one or more metaphors.

Challenge 3 – Network Analysis. The aspects of network analysis, which encompass centralities, motifs, aggregation, and clustering, and the need to incorporate the semantics, define our third challenge. Results here could inform the network visualization characteristics, including interactive layer arrangements, derivation, and filtering. Future work could include realizing a visual analytics system for knowledge graph analysis.

References

1 David N. Nicholson and Casey S. Greene. "Constructing knowledge graphs and their biomedical applications." Computational and structural biotechnology journal, 18, pp.1414-1428, 2020.
2 Fintan McGee, et al. *Visual Analysis of Multilayer Networks. Synthesis Lectures on Visualization.* 8(1), pp.1-150, 2021.
3 Falk Schreiber, Andreas Kerren, Katy Börner, Hans Hagen, and Dirk Zeckzer. *Heterogeneous Networks on Multiple Levels.* Multivariate Network Visualization: Dagstuhl Seminar #13201, Dagstuhl Castle, Germany, 2013.
4 Tim Dwyer, et al. *Immersive Analytics: An Introduction.* In Immersive Analytics, pp.1-23. Springer, Cham, 2018.

4.5 A curriculum for the future of bio+med-Vis

Torsten Möller (Universität Wien, AT), Emma Beauxis-Aussalet (VU University Amsterdam, NL), Helena Jambor (Universitätsklinikum TU Dresden, DE), Barbora Kozlíková (Masaryk University – Brno, CZ), Kay Katja Nieselt (Universität Tübingen, DE), and William Ray (Ohio State University – Columbus, US)

The communication role of visualization is well accepted in the bio+med communities. At the same time, it is essential to make biological and medical practitioners more aware of the value of visualization, not only as a dissemination tool but also as a data analysis tool. The development of a Curriculum for these communities, bio+med, and visualization, has thus a high priority and aims to familiarise students with modern visual analysis methodologies applied to biological and medical data and provide hands-on training.

While several visualization community members are teaching summer camps, tutorials, and workshops on biological and medical data visualization, many of these educational sessions take the form of an introduction to specific tools. We find ourselves handling similar questions: What is exploratory data visualization? What is visual analysis, which frameworks to think about visualization exist, how can we explore design space? And how can we visualise biological and medical data to gain insights into them so that hypotheses can be generated, explored, and validated, and further analyses can be targeted for future work?

Despite the increasing importance of visualization for the bio+med communities, there is currently a general lack of integration into curricula, and hence a lack of visual literacy. A useful and appropriate curriculum has not yet been developed. Therefore, in our group, we focused on addressing the following questions:

- What should a modern and seminal curriculum for visualization for the bio+med communities contain?
- With the different target audiences in mind, how far along in introducing visualization should such curriculum go while also integrating courses on the domain and topics on data?
- What are the essential topics for such a curriculum, and how can it provide comprehensive hands-on training?

The actions taken in this working group were shaped based on these questions and on the previous initiative of building a teaching platform for the bio+med+vis audience.

The latter developed into the first Bio+Med+Vis Spring School, held in 2021 (`https://biomedvis.github.io/`). The initial ideas for this initiative were drawn from the 2018 BioVis Dagstuhl seminar 18161 "Visualization of Biological Data – Crossroads", and from the 2020 Shonan seminar 167 on "Formalizing Biological and Medical Visualization".

Within our working group, we aimed to reflect on and review what has been done so far to extend and enhance the matrix of topics for the curriculum. The current status of the matrix can be found following this link.

We first revised the current content in designing the curriculum and discussed potentially missing areas or topics. We then extended the overall *Platform* table by the *Overview* row, which will serve for collecting introductory topics and prerequisites for individual methods (listed in table columns), without any specifics of scales (listed in table rows). We further added the *Visual Design* table with its corresponding content. We also added the *Contributors* section, where we asked all participants of the Dagstuhl seminar to express their willingness to participate in this initiative. Our goal was to identify enthusiastic collaborators and potential driving forces of the upcoming initiatives in this context. Our working group also identified the need for defining the learning outcomes for each module, as these are one of the driving forces for the content of the curriculum.

The next actions that will follow after the Dagstuhl seminar will be:

- Finalizing the matrix content by filling in the missing gaps in individual tables.
- Addressing experts in given fields to prepare the educational materials for topics listed in the matrix.
- Initiating the creation of the BioMedVis educational platform to collect the educational materials and make them publicly available.
- Addressing potential collaborators regarding the preparation of the following Spring School and discussing the date and content with them.
- Explicitly joining forces within the BioVis + VCBM organizational structures to bring the communities together and strengthen our research and educational efforts.

Finally, we would like to acknowledge the collaboration of our colleagues not attending the seminar but who have contributed to the results of our working group. These include Jan Byška, Renata Raidou, Noeska Natasja Smit, Ingrid Hotz, Johanna Beyer, and Hsiang-Yun Wu.

4.6 Facilitating cross-expertise exploration in XAI for multi-omics via visualization

Cagatay Turkay (University of Warwick – Coventry, GB), Jillian Aurisano (University of Cincinnati, US), Theresa Anisja Harbig (Universität Tübingen, DE), Raghu Machiraju (Ohio State University – Columbus, US), and Mathias Witte Paz (Universität Tübingen, DE)

With the emerging high-throughput methods in biology, we can systematically collect the data from multiple omics levels such as genomics, transcriptomics, proteomics, and metabolomics. Therefore, in addition to analyzing single entities, such as a small group of genes in one omics layer, researchers can now analyze larger sets of entities across multiple omics layers to increase their understanding of complex biological systems. However, the produced data

of all layers need to be analyzed simultaneously for cross-features such as (anti-)correlations or common clusters. Hence, there is an increasing interest in applying artificial intelligence (AI) methods to not just one but several combined "omics" datasets. Since data experts are not often familiar with AI methods, a further actor comes into play for method development. With this, a new challenge within the omics field arises since the domain experts must communicate their research approaches with non-experts, and model developers also need to share their knowledge. However, the combination of the data and the applied methods complicates the explainability of the results since both are already complex by themselves. We believe that we can use visualization techniques to facilitate the communication between actors.

Here, we first discuss the characteristics of multi-omics data that make the development of AI methods more difficult. We then identify the current stakeholders within this AI-multi-omics system and propose analyzing the existing visualization techniques via a survey.

Multi-Omics Data Characteristics

Multi-omics data have particular characteristics that can complicate AI model development and make it difficult to explain its results. First of all, multi-omics data is high-dimensional. Depending on the study, genomics, transcriptomics, proteomics, metabolomics, and/or epigenomics data can be included (among others). Analyzing and understanding these data types on their own has been a longstanding challenge for the field since high throughput measurements were introduced to molecular biology research, let alone understanding the combination of multiple omics data, or the results of an abstract AI model, or the AI model itself. AI methods have tremendous potential to enable researchers to make sense of complex data. However, the problem of helping domain experts understand these systems is complicated by the use of opaque AI methods.

Several challenges in multi-omics data creation and collection complicate model development and interpretation. First of all, multi-omics data sets can be very different. The number of samples and the analyzed omics levels can vary immensely, leading to different requirements for the model depending on the data. Moreover, missing data is a challenge for all omics data. For example, in genomics, not all genes might be annotated, or the functions of some genes might be missing, which can lead to an incomplete model or difficulties in understanding the model results. Moreover, missing data can lead to data imbalance. For example, when analyzing cancer data, certain cancer types are much more common than others and characterized more thoroughly. Therefore, when interpreting the results of an AI model or the model itself, it is always important to be aware of the quality of the input data. If the results do not match the expectations when using this input data, it does not necessarily mean that the model is flawed or that the expected signal could not be detected. It could simply mean that the input data was not sufficient.

Finally, several challenges arise when a researcher interprets the result of an AI model or the model itself. Multi-omics is an emerging field, and domain experts often specialize in one specific omics data type. Naturally, this can lead to an interpretation of the data and the artificial intelligence results' centered on a particular field of omics. Domain experts might attempt to apply AI techniques to increase the understanding of an already known phenomenon or to find an explanation for a hypothesis about the data. In contrast, non-domain experts might have a completely different, more unbiased view of the data and perform a purely data-driven analysis more often. However, they tend to have less background expertise and less knowledge about potentially missing data. This can lead to a challenge in communication between different stakeholders dealing with multi-omics data.

Stakeholders

From our experience in the multi-omics field, we have distinguished stakeholders with different goals and interests concerning the AI methods applied for the data analysis. Commonly known stakeholders from this field are the modelers and the domain experts. Even though this division might not always be as sharp as described here, it is common for these two actors to have disjoint backgrounds. Usually, the modelers tend to have less domain expertise from the applied field (e.g. microbiology). In contrast, domain experts might have little to no understanding of the underlying machine-learning models. This could lead to the different interests of both parties: the modelers might be primarily involved with the model's performance rather than with the underlying phenomena that the model captures. At the same time, the domain experts focus only on the model's final outputs since understanding the decision path would require time and background expertise they usually do not possess.

However, there might be situations where these two stakeholders require detailed knowledge on the complementary side of the system, especially when a problem with the model occurs, such as a difference in the expected output from the side of the domain expert or low performance of the model after being evaluated by the modeler. For this, they usually go to an explainer or assume such a role for having an accurate translation between model development, data expertise, and artificial intelligence methods. By integrating both sides of the project and encoding the missing knowledge in a simplified way, such as in a visualization, the person assuming the role of the explainer could also motivate the two stakeholders to be curious and explore the decision paths of the model. We believe more research is needed to develop compelling visualizations that facilitate the task of the explainer acting as a translator and motivate the stakeholders to explore the decision paths by themselves in more detail, even if no issue has been detected in the developed model. Such curious stakeholders might be able to find information that could result in breakthroughs for their corresponding field by accounting for the model's output or structure and considering the model's behavior.

Outlook

Though the previously described stakeholders were identified from our prior experiences in the multi-omics field, it is unclear how often the actors appear, especially the explainer. To clarify this, we propose to develop a survey on how often curiosity comes into play in a multi-omics project applying AI methods, especially concerning the decision paths taken by the model. We also expect to gather current visualization techniques used from the explainers' side to reduce the gap between the parties. The gathered insights will help develop future tools that apply machine-learning methods in multi-omics data and avoid curious actors requiring the help of an explainer.

Nevertheless, it might be possible that the number of curious actors is not high enough for us to collect sufficient materials. This would mean that, although theoretically reducing the gap between modelers and domain experts would provide a better insight into the studied phenomena, such approaches may not be feasible or desirable in reality. In such cases, our focus from our survey results would be different: we would start thinking of visualization techniques that motivate the stakeholders to think outside their boxes and consider the whole picture. We believe that by reducing such a gap, the dogmas of both sides might be questioned more intensely, and it might facilitate the multi-omics data to provide new perspectives into the applied fields.

Participants

- Jan Aerts
Hasselt University – Diepenbeek, BE
- Emma Beauxis-Aussalet
VU University Amsterdam, NL
- Michael Behrisch
Utrecht University, NL
- Katja Bühler
VRVis – Wien, AT
- Nadezhda T. Doncheva
University of Copenhagen, DK
- Mennatallah El-Assady
ETH Zürich, CH
- Carsten Görg
University of Colorado – Aurora, US
- Theresa Anisja Harbig
Universität Tübingen, DE
- Lynda Hardman
CWI – Amsterdam, NL
- Lawrence Hunter
University of Colorado – Aurora, US
- Helena Jambor
Universitätsklinikum TU Dresden, DE
- Andreas Kerren
Linköping University, SE
- Karsten Klein
Universität Konstanz, DE
- Stephen G. Kobourov
University of Arizona – Tucson, US
- Alexander Koch
BioLizard – Gent, BE
- Michael Krone
Universität Tübingen, DE
- Martin Krzywinski
BC Cancer Research Centre – Vancouver, CA
- Alexander Lex
University of Utah – Salt Lake City, US
- Lennart Martens
Ghent University, BE
- Torsten Möller
Universität Wien, AT
- Kay Katja Nieselt
Universität Tübingen, DE
- Bruno Pinaud
University of Bordeaux, FR
- Timo Ropinski
Universität Ulm, DE
- Falk Schreiber
Universität Konstanz, DE
- Cagatay Turkay
University of Warwick – Coventry, GB
- Blaz Zupan
University of Ljubljana, SI

Remote Participants

- Jillian Aurisano
University of Cincinnati, US
- Tanya Berger-Wolf
Ohio State University – Columbus, US
- Katy Börner
Indiana University – Bloomington, US
- Andreas Bueckle
Indiana University – Bloomington, US
- Guadalupe Canahuate
University of Iowa, US
- Jian Chen
Ohio State University – Columbus, US
- Nils Gehlenborg
Harvard University – Boston, US
- Barbora Kozlíková
Masaryk University – Brno, CZ
- Raghu Machiraju
Ohio State University – Columbus, US

- Georgeta Elisabeta Marai
University of Illinois
Chicago, US
- James Procter
University of Dundee, GB
- William Ray
Ohio State University –
Columbus, US
- Jos B.T.M. Roerdink
University of Groningen, NL
- Ryo Sakai
GSK – Uxbridge, GB
- Hagit Shatkay
University of Delaware –
Newark, US
- Charlotte Soneson
FMI – Basel, CH
- Miha Štajdohar
Genialis – Houston, US
- Marc Streit
Johannes Kepler Universität
Linz, AT
- Granger Sutton
J. Craig Venter Institute –
Rockville, US
- Danielle Szafir
University of North Carolina at
Chapel Hill, US
- Mathias Witte Paz
Universität Tübingen, DE

Report from Dagstuhl Seminar 21402

Digital Disinformation: Taxonomy, Impact, Mitigation, and Regulation

Edited by

Claude Kirchner[1] and Franziska Roesner[2]

1 **CNPEN/CCNE and Inria – Paris, FR,** `claude.kirchner@inria.fr`
2 **University of Washington – Seattle, US,** `franzi@cs.washington.edu`

Abstract

We report on the discussions and conclusions of a Dagstuhl seminar focused on digital mis- and disinformation, held in October of 2021. An international and interdisciplinary group of seminar participants considered key technical and societal topics including trustworthiness algorithms (i.e., how to build systems that assess trustworthiness automatically), friction as a technique in platform design (e.g., to slow down people's consumption of information on social media), the ethics of mis/disinformation interventions, and how to educate users. We detail these discussions and highlight questions for the future.

Seminar October 3–6, 2021 – http://www.dagstuhl.de/21402
2012 ACM Subject Classification Information systems → Collaborative and social computing systems and tools; Security and privacy → Human and societal aspects of security and privacy; General and reference → Verification
Keywords and phrases Information, disinformation, misinformation, fake news, deep fake, ethics, trustworthiness, friction, verification
Digital Object Identifier 10.4230/DagRep.11.9.28

1 Executive Summary

Claude Kirchner
Franziska Roesner

Dagstuhl Seminar #21402 on Digital Disinformation occurred on October 4-6, 2021. The seminar was initially planned by Claude Kirchner (CNPEN/CCNE & Inria), Ninja Marnau (CISPA), and Franziska Roesner (University of Washington), and it was then co-lead and this report was written by Kirchner and Roesner, with input from other seminar participants. The seminar had been originally planned for June of 2020 but was then postponed due to the COVID-19 pandemic. It was held in a hybrid format, with some participants on-site in Dagstuhl and most others joining remotely via the video conferencing system Zoom.

In order to maximize discussion and allow the interests of the group to drive the direction of the seminar, we did not plan for formal talks. Participants were asked to prepare a single slide, few-minute introduction about their research interests and methodologies related to digital disinformation, and a "burning question" they have in the space.

Digital Disinformation: Taxonomy, Impact, Mitigation, and Regulation, *Dagstuhl Reports*, Vol. 11, Issue 09, pp. 28–44
Editors: Claude Kirchner and Franziska Roesner
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Participants included the following individuals, spanning a range of expertise from computer science to law:

- Esma Aïmeur (University of Montréal, Canada)
- Jos Baeten (CWI Amsterdam, Netherlands)
- Asia Biega (Max Planck Institute for Security and Privacy, Germany)
- Camille Darche (CNPEN, Inria and Université Paris Nanterre, France)
- Sébastien Gambs (Université du Québec à Montréal, Canada)
- Krishna Gummadi (Max Planck Institute for Software Systems, Germany)
- Claude Kirchner (CNPEN/CCNE and Inria, France)
- Vladimir Kropotov (Trend Micro, Russia)
- Jean-Yves Marion (Lorraine University, France)
- Evangelos Markatos (University of Crete, Greece)
- Fil Menczer (Indiana University, USA)
- Trisha Meyer (Vrije Universiteit Brussel, Belgium)
- Franziska Roesner (University of Washington, USA)
- Kavé Salamatian (University of Savoie, France)
- Juliette Sénéchal (University of Lille, France)
- Dimitrios Serpanos (University of Patras, Greece)
- Serena Villata (CNRS, France)

Based on our preliminary discussions, we identified four topics of interest to many seminar participants: *trustworthiness algorithms* (i.e., how to build systems that assess trust automatically), *friction as a technique in platform design* (e.g., to allow for people to take time and a step back when consuming information on social media), *the ethics of interventions* (e.g., the ethics of blocking or content moderation), and *how to educate users* (e.g., without creating over-skepticism). We then structured the rest of the seminar around four deep-dive conversations on these topics, described in the subsequent sections of this report. Due to the relatively small size of the gathering, and most participants' broad interest in all four topics, we did not break out into smaller discussion groups but rather continued to discuss as a full group.

2 Table of Contents

3 Deep Dive 1: Trustworthiness Algorithms

Our first deep dive conversation focused on the question of how to build systems that assess trustworthiness automatically.

3.1 Reflections on Truth versus Trustworthiness

Helping people to understand the information they have access to in the digital world could be facilitated by algorithms designed to check veracity of statements, to detect incorrect statements, or to find inconsistencies in a given set of knowledge. While automated mis/disinformation algorithms are often framed in terms of such fact-checking, seminar participants repeatedly made the observation that assessing truth can be challenging.

Automating truth assertion is difficult since the level of specification of a statement to check can vary from formal (as in $2 + 2 = 4$) to nearly formal (as in "$2 + 2 =$ four") to imprecise and complex (such as "COVID-19 is a hoax"). Even for formal statements, we know since Turing and Post that some facts are even undecidable: for some (rather rare) facts there exists no algorithm able to decide if they are true or false. And for complex, informal statements, establishing truth can be hard or impossible, even if we could transform them into formalized statement by explicitly describing all needed details of the context under consideration. The context of particular statement may depend on time, culture, history, education, current availability of verified information or knowledge (e.g., as we have seen scientific knowledge evolving since the beginning of the COVID-19 pandemic), etc.

Moreover, misinformation is rarely something to which we can assign a binary truth value. Often we find a mix of truth with misleading suggestions, implications, attribution, or false context. There is significant nuance that makes it hard even for expert humans to fact-check some claims. Often the best a human fact-checker can do is to observe there there is no evidence in support of a claim, rather than conclude that the claim is false.

Thus, since establishing truth can be very hard (possibly undecidable), participants suggested that we could instead consider automated assessments of trustworthiness. As an example: not everything on Breitbart News or Occupy Democrats is "false", but these sources have low credibility according to fact-checking organizations.

3.2 Assessing Trustworthiness

Moving, then, from fact-checking or truth assessment to the assessment of trustworthiness, we discussed several aspects:

What is being evaluated? Approaches for assessing trustworthiness might target *content* (e.g., might this image be a deep fake?), content *producers* (e.g., does this website or social media account have a reputation for sharing truthworthy content?), or even content *consumers* (e.g., does this user have a history of re-sharing misinformation?).

Who is assessing trustworthiness? Trustworthiness assessments might be heavily impacted by factors such as the background, education, and biases of the annotators – either humans, or automated processes that may embed such bias in their design.

What makes something trustworthy? Or rather, what signals might a trustworthiness algorithm use to assess content or sources? Participants mentioned potential signals including:

- Adherence to journalistic standards
- Falsifiability (see Popper)
- Accountability, e.g., a history of correcting errors when they are confirmed as such
- Primary knowledge domain or level of expertise of a source being relevant to the topic at hand
- Confidence that an account is not compromised
- User feedback (though this may be manipulated [12])
- Behavior of users who are known to be good at distinguishing mis/disinformation
- Use of techniques (e.g., clickbait, sensationalism) known to play into human cognitive biases

There is a growing number of news and fact-checking organizations that compile credibility metric for sources according to the above criteria, e.g., `https://mediabiasfactcheck.com`, `https://www.newsguardtech.com`, `https://iffy.news`, and `https://factcheck.afp.com`.

Who is the audience for trustworthiness information? We can consider different audiences for the outputs of trustworthiness algorithms. One potential audience is a social media or other platform itself, for use in amplifying or de-amplifying certain content. Beyond this, the targets of such interventions are often end users themselves, e.g., labels on content to help them assess information as they see it. For any user-facing interventions, the designers must consider whether they are targeting users who are receptive to additional information (e.g., conscientious consumers of information trying to make sense of the information ecosystems) or not (e.g., people already convinced of a false conspiracy theory). And in addition to different kinds of people, there are quite different kinds of information. One size will obviously not fit everyone.

Building systems assessing trust automatically will rely (i) on the many factors listed above, but also on (ii) ways to allow people to have time to take a step back (e.g., using friction as described in Section 4) and (iii) on thinking about the many ethics issues concerning the platforms (Section 5).

3.3 Technical Approaches

Participants also brainstormed specific technical approaches that might be used to support trustworthiness assessment or other verification. The observation was made that online platforms are already doing some of this work, but not much about their techniques or algorithms is public or transparent; therefore the research community cannot assess the soundness and appropriate application of platform regulation based on these metrics.

Automated property proving could be used on sufficiently formal statements, possibly with specific pre-processing of them.

The pagerank technique is an early example, initially used by search engines. These types of algorithms could be used on Twitter graph data, for example, starting with initial trustworthiness labels from existing sources.

Cryptographic techniques can be used to trace the provenance of content, identify the sources of content, and/or attest to attributes of content (e.g., metadata like when and where a video was shot, or when and by whom it was published online). While

cryptographic techniques like digital signatures do allow attribution of content to a (possibly trusted) source, participants pointed out that there are limitations to these approaches because we still need a root of trust – e.g., we might be able to verify the provenance of a video, but we still need to know whether we can trust the source.

Deep fake detection is useful, but needs continuous adaptation to technical improvements in deep fake creation (i.e., an "arms race").

3.4 Reflections on Harms from Untrustworthy Content

As part of this deep dive conversation, participants also reflected on the potential harms of untrustworthy content – both what those harms are, as well as how they might be used to prioritize different types of sources or content for intervention.

Spread and targeting

The current social internet allows both brute force broadcasting (i.e., everyone receives the same (mis)information) as well as highly targeted (mis)information dissemination. Highly targeted information can be sent to many different people, and they will not see the same thing as it will be automatically tailored to their unique profile – a unique phenomenon in our new digital ecosystem. We observed that both mis/disinformation with broad reach, as well as highly targeted mis/disinformation, can create substantial harm, but may require different approaches to combat.

Assessing and prioritizing potential harms

Different types of mis/disinformation may have have different harms. For example, ill-founded distrust in vaccines has clear potential harms [18], while on its face it seems less problematic if some people believe that the earth is flat – though the distinctions are not so simple, as it has been show that beliefs in different conspiracy theories may be correlated. Moreover, different populations may be more vulnerable to certain harms. With a good characterization we could use that to prioritize limited resources. But characterizing and measuring harm in online content could be very hard and remains an active research area (e.g., [31, 21]). Even irrespective of harm prediction, any prioritization that makes a judgment on which harms to which populations are more "important" than others will have ethical impacts.

Long-term impacts of content despite awareness that source is untrustworthy

Even with trustworthiness information, people might not remember the provenance of information they have seen – they might assume something is true even though they were exposed to it from an untrustworthy source. For example, it has been shown in the context of ads [17] that people do not necessarily remember the sources of information. This observation raises a more general related question: how are people influenced by bad information they just see in their information ecosystem but do not engage much with, such as posts they scroll by or ads they think they ignore? Does that information (especially with source provenance later lost) make it into their worldviews? Is this effect worse when people read only catchy headlines, but not the underlying articles?

3.5 Challenging Trustworthiness

We also noted that trustworthiness is challenged in various ways. First, disinformation is fundamentally an attack on the human brain, taking advantage of aspects of human neuro-cognition, culture, behaviors, and potentially fault or non-rational heuristics. Simply surfacing information about trustworthiness to people will not, on its own, suffice to overcome these issues. A second challenge comes from traditional computer security: information maybe altered on purpose, e.g., via the criminal alteration of data stored in compromised computers in order to support disinformation and falsely justify fake news. For example, a participant pointed to recent targeted attacks on research against COVID-19 vaccines [23]. And finally, while AI and ML can provide tools to help combat disinformation (e.g., supporting trustworthiness assessment), their advancement may also increase the volume and the quality of fake news, as well as bringing new disinformation vectors. These challenges compound the complexity of understanding and controlling the context in which we would like to increase trustworthiness, and require for further research.

4 Deep Dive 2: Friction as a Technique in Platform Design

Our second deep dive conversation focused on the topic of "friction" as a technique in (social media or other information) platform design. The idea of friction involves slowing down some human interaction with the platform, either on the information producer side (e.g., limiting down how often someone can post) or on the information consumer side (e.g., limiting how much time someone spends on the platform).

4.1 Argument for Friction

Some participants presented evidence for why the idea of friction might be fruitful. One piece of evidence: while one might imagine that platform behavior might self-regulate, in the sense that higher quality content should become more popular, in fact, the correlation between quality and popularity is very weak as long as people have finite attention (i.e., are unable to see everything that is posted on the platform) [13]. This reality is problematic from the perspective of misinformation, because it makes us vulnerable to bad actors flooding the network with more information. Indeed, there is empirical evidence that adversaries flood the network with lots of volume of problematic content [25, 22].

At the same time, we know from many examples in other domains and in our own lives that what might be good for a person's well-being in general, and what they want or choose to do in a given moment, are not always aligned, making us vulnerable to addictive online platforms. Related to this, the question was raised: *is there an equivalent to the "privacy paradox" [3] for misinformation?* Seminar participants posited that the answer is likely yes: that people are not as conscientious about consuming information as they say they are or want to be (e.g., they may still click on clickbait content even though they recognize it as clickbait). However, just as the privacy paradox is not truly a paradox (because people are forced into the impossible choice between opting out of crucial online platforms and agreeing to their privacy ultimatums) [24], we posit that in the context of misinformation, people's actual choices and behavior are deeply influenced by the platforms. In other words, the situation is not so much a paradox as a failure of platforms to design for their users' well-being.

Towards that end, we thus considered "friction" as a potential tool in platform design to reduce mis/disinformation and increase well-being. In the extreme, as a thought experiment, a platform might have no friction (i.e., anyone can post or view anything) or only friction (i.e., everything is blocked). The latter case is clearly nonsensical, but as a practical matter, so is the former: participants stressed that *platform designs are never neutral*. Particularly because there is finite time and attention that an individual can spend on a platform, the design choices about who is shown what content in what way, and the affordances for how content is posted and shared, naturally shape people's consumption and production behaviors.

4.2 Examples and Proposals for Friction in Platform Design

We considered existing examples of friction, as well as our own new proposals, for different stakeholders.

Friction for Information Consumers

Consumers are users who use social media or other information platforms to read and/or view content. Examples of friction for consumers include:

- Most generally, the platform's algorithm for what is shown to whom when is a form of friction for certain types of content.
- Labeling content (e.g., with "false" labels if something was able to be explicitly fact-checked, or with other information, such as about the trustworthiness of the source as discussed in Section 3) as a nudge for people to think about it a second time.
- Minor changes in an interface can have impacts on how people consume information (e.g., the presence or absence of a search box might change people's ability to act on their preferences in content consumption [16]).
- Helping people limit how much they interact with platforms, e.g., how much time they spend on social media. Existing tools include screen time capping tools (e.g., smartphone apps or browser extensions), including tools with additional incentives (e.g., planting real or virtual trees). A proposal was to surface the climate impact for use of platforms per unit time, as a new type of incentive/motivation.
- *Proposal:* Some kind of platform guidance for helping shape discourse (e.g., like an ongoing project designing a tool to help local actors intervene in mis/disinformation related conversations [26]).
- *Proposal:* Design social media platforms for more intent-driven interactions. Today, people visiting social media sites are inundated with content on all sorts of topics interleaved on their feeds, which can create information overload, reduce motivation to assess information quality [10], and other potential harms. Imagine instead a platform that greets a user with an empty search box, asking them to take an active role in shaping the topics and types of content that they see. However, we note a potential privacy-related side effect here: active interactions with the platform will also provide more information about the user's interests and habits that could be exploited for targeted advertising or manipulation; thus, any such approach should be coupled with robust privacy protections.
- *Proposal:* Helping users filter the content they want to see, e.g., on Twitter, to engage on academic research topics but not U.S. politics.

Friction for Information Producers

Information producers are those who create new posts containing mis/disinformation or other low-quality or problematic content. Examples of friction for producers include:

- Twitter limits the number of posts per day (though posters may violate this by deleting things they posted earlier). One might also impose other limits, such as rate limits.
- Twitter used captchas to limit posting volume during the 2020 U.S. elections.
- Content moderation in general was considered a general form of friction (e.g., removing posts or account, shadow-banning). The Facebook Oversight Board was mentioned, which evaluates content removal or account banning decisions.
- Other researchers have highlighted the importance of banning "repeat offenders", accounts that frequently post or spread mis/disinformation [8].
- *Proposal:* On WhatsApp, a message that has spread beyond a certain number of people could be made public, to allow for scrutiny and content moderation (otherwise challenging on an end-to-end encrypted platform).
- *Proposal:* Some kind of monetary cost for posting (similar to what advertisers must do today, or similar to postage stamps for physical messages).
- *Proposal:* Some kind of "friction score" or "trustworthiness score" for information producers, where higher-quality information producers are able to post more, with more reach. The idea would be to create disincentives (cost) for bad actions, and incentives or rewards for responsible use. This proposal comes with many questions and challenges, including how such a score should be computed, by whom, the need to provide transparency about its implementation and enforcement, the privacy risks associated with how the score might be misused, etc. Some ideas for factors that might be incorporated into the score included: the quality of previously shared links, prior track record of posted content that has been flagged or taken down, certain behaviors (e.g., high-frequency posting, inauthentic/coordinated behavior), and audience diversity or other audience properties. Or could we translate some of the trust indicators developed for news organizations (e.g., as used by Newsguard) to individuals? Using crowdsourced data as part of the score was cautioned against, due to risks of coordinated activity or similar [30].
- *Proposal:* Time-based friction for posting, when critical events like elections are happening or shortly before a company goes for IPO, friction could be increased to minimize manipulations at the time when they are most impactful. This idea could apply to friction for information consumers and/or intermediaries.

Friction for Intermediaries

Intermediaries are users who knowingly or unknowingly re-share mis/ disinformation posted by others (e.g., retweeting on Twitter). Examples of friction for intermediaries include:

- WhatsApp limits the number of people to whom someone can forward a message.
- Facebook alerts you if you try to share a link or post that their fact-checkers have labeled as false.
- Twitter warns you if you retweet a link without having first clicked on it.
- *Proposal:* Temporarily disable reposting capability for users who ignore warnings with high frequency.
- *Proposal:* The above-mentioned "trustworthiness score" could be used for intermediaries as well.

- *Proposal:* Behaviorally identifying users who are good at identifying and/or avoiding mis/disinformation, and using that information to, for example, amplify or reduce the spread of certain types of content.

4.3 Practical and Ethics Considerations

Orthogonal to any specific friction proposals, participants voiced concerns about practical considerations in deploying such techniques. First, there are questions about the incentives of the platform designers, in terms of economic models and business incentives. Indeed, after the seminar, there was the example of a browser extension designed to limit people's Facebook use that was shut down by Facebook [20]. The corresponding observation was that these questions are not merely technical challenges but will require regulations to put any solutions into place.

Additionally, there are ethical, legal, and political questions around the power of platforms to make decisions like promoting or demoting certain types of content or blocking certain people. One participant also noted that the use of "friction" is, of course, not ubiquitously a good thing, pointing, for example, to work documenting the uses of friction and flooding as techniques by authoritarian governments to exert control over information in potentially harmful ways [19]. We discuss ethics more in Section 5.

4.4 Looking Ahead

Clearly, significant work remains to be done to design, implement, and evaluate the impacts of different types of friction proposals, compounded with the challenges of platform incentives and their limited transparency and oversight. Stepping back, we concluded our discussion on this topic with a crucial, overarching question: *How might we redesign platforms more radically?* For example, if we wanted to design a "public interest social media platform", what would that look like? Could it be done? Or, looking further into the future towards AR/VR or "metaverse" interactions: can such platforms be designed in ways that help move us more towards (imagined?) ideals of in-person instead of digitally-mediated interactions?

5 Deep Dive 3: Ethics of Interventions

On the second day, we chose to include one formal talk to help seed our discussion on ethics: Serena Villata gave a talk titled "Online disinformation: content moderation, ethical challenges, and future work directions" describing in particular the collective thinking held within the National Pilot Committee for Digital Ethics (Conseil National Pilote d'Ethique du Numérique – CNPEN) [4]. The abstract says:

> The purpose [...] is to identify the ethical issues and challenges arising from the widespread use of these different algorithms and tools, which are part of a complex phenomenon with wide-ranging implications. Among others, some questions arise: what does action or inaction in this domain mean in the context of COVID-19? Is it simply a quantitative shift, or are we seeing a more profound change in the nature of the digital solutions designed to fight online disinformation and misinformation?

More generally, how do we face the complexity of this phenomenon, which requires an analysis that seems to go beyond ethics, or even to challenge the notion of ethics itself? Indeed, ethical questions do not arise in the same way depending on whether one is dealing with actors who act consciously to deceive their target or, on the other hand, whether one is dealing with actors who simply get caught up in the flow of information in digital format and, in particular, participate in the virality of this information often in an unconscious way. In the first case, ethical reflection questions responsibility, while in the second case it consists mainly in moving towards awareness. In either case, the requirement is to identify – specifically in the digital domain – the economic, legal, social, political or philosophical dimensions of disinformation or misinformation.

The discussion opened by the presentation helped raise fundamental questions about ethics in the context of mis/disinformation.

5.1 Which ethics for which purpose?

Are there existing ethical frameworks/theories that could be applied in this context? Shall we develop and/or rely on existing philosophical frameworks? Can we take inspiration from the many AI governance propositions issued from the committees set up, for example, by the UE, IEEE or Unesco? How do we resolve some of the fundamental-seeming tensions (e.g., free speech vs content moderation for safety)?

A crucial point concerns the cultural aspects of information. Do we expect a global set of ethics accepted by everyone, or do we expect ethics to be regional? Clearly, norms and group behaviors are quite often following local customs and they can be influenced or offended by other behaviors in different regions. Related works like the Confucian approaches to tech ethics were mentioned [29].

When looking at ethics as defined by the French Academy [6] ("Reflection on human behaviour and the values on which it is based, carried out with a view to establishing a doctrine, a science of morality."), the reflection process may evolve over time and of course in different cultures. This highlights the difficulty to answer the question: should we (can we?) establish a common set of rules? And for which purpose?

As ethics precedes regulation and laws that are regional or national in many cases, we see the difficulty, but also the interest, to collaborate on these questions and to try to find or establish common views in the context of global digital platforms.

5.2 What values are behind moderation tools?

When asking which values are behind the moderation tools, the following points were raised.

Algorithmic fairness to avoid issues like algorithmic bias (e.g., the Amazon hiring example). Bias could in particular be introduced when detecting proactively vs noticing the impacts. Bias could be better understood and avoided when running simulations to help predict otherwise unforeseen side effects of algorithmic designs.

Transparency appears as a key value here again. For reference, a talk was mentioned from Oana Goga (based on work with Krishna Gummadi and others) auditing explanations provided by social platforms on why some content was shown, highlighting also how these explanations can be manipulated to appear "neutral" [11].

Consent is key as in bioethics, and it relies in part on thoughtful consent design and on education: how to ensure that people understand what they are consenting for? The role of nudging, with much prior work, was also highlighted by participants.

5.3 How can ethics committees help governing social media platforms?

We currently see the setting up of ethical committees close to the platform (e.g., Facebook Oversight Board), possibly acting independently. Can this help? Participants raised the following questions for/about the Facebook Oversight Board:

- What does Facebook do with the user-generated reports of Facebook takeout data (and of course with all the data)?
- What data does the Oversight Board actually get access to?
- Process, data, objectives, consequences (how seriously is this taken)?
- Membership of the board? Multi-stakeholder?
- How does Facebook's internal governance interact with the external committee?
- What about ethics-washing?

Participants also observed:

1. The role of the Oversight Board is limited to content decisions, not actually oversight of Facebook per se, despite its name.
2. Facebook's platform is not neutral: the design of the platform shapes not only consumption behaviors but has also shaped content (e.g., from the Facebook whistleblower interview with 60 Minutes [1], politicians saying they felt forced to take more extreme positions that will work well with Facebook's algorithm)

Every problem Facebook and others are facing is so complex, with conflicting needs from different stakeholders – could we actually do better, if we were there? Are there examples of ethics boards in such companies working well? We left the question open but we begin to think about what would be desirable.

Let's imagine new, different social media platforms

Such a platform might rely on an oversight board accessing information (what?) to make decisions (which?) to be applied. The ethics committee might use tools for being more transparent on data and models without releasing the data and the model, e.g., [9, 15]. And of course it is crucial to involve multiple, globally and otherwise diverse stakeholders.

6 Deep Dive 4: User Education

Finally, we discussed the potential role of and strategies for user education against mis/-disinformation. Indeed, many existing educational efforts exist in this space (e.g., several efforts at the University of Washington alone [28, 5], TrendMicro's internet safety educational resources [27], the Fakey game [14]); our goal here is not to suggest that all of the ideas below are new, but rather to summarize the discussion at the seminar about the role of and important considerations in educational interventions.

6.1 Role of User Education

When considering educational interventions, an important question is always to consider how much should the responsibility be on users to protect themselves versus how much should the responsibility be on companies and regulators to obviate the need for education and self-protection. Beyond responsibility, there is a question of how realistic or appropriate it is to expect users to develop deep understandings of the technical systems they use– indeed, many users do not have such understanding [7]. it seems clear to us that the burden cannot and should not be on end users alone. Nevertheless, educational interventions may have an important role to play in a multi-faceted fight against mis/disinformation, and can provide users with agency and empowerment even in the face of slow-moving or otherwise incentivized regulators and platforms.

6.2 Characterizing Education Efforts

Seminar participants considered different ways to characterize or scope educational interventions, and different options for their aims and implementations. These included:

Who is the audience of an education effort? This might include vulnerable populations, such as kids, teenagers, older adults, people with lower literacy. This might include particularly powerful individuals, such as politician, local influencers, C-level executives, business owners, or people in charge of platform design and implementation. And this might include everyone, continuously, who interacts with online information ecosystems.

What is the goal of an education effort? At first blush, the core goal of an anti-misinformation educational effort might seem to be to *help people learn how to identify mis/disinformation.* However, participants were quick to point out that one should avoid teaching only skepticism, which risks cultivating cynicism [2] and undermining trust in information ecosystems altogether. Instead, educational efforts should also include the opposite goal: *helping people identify trustworthy information and sources.*

Other potential educational goals articulated by seminar participants included: *helping people thinking critically* and *helping people learn situational awareness online*, as well as specifically *helping people identify manipulative techniques being used* in the content that they see (e.g., Cialdini's principles of inflence or persuasion). Another goal might be to *help people engage with others productively about misinformation* (e.g., teaching skills for how to interact with someone when they have posted something false).

Participants also observed that educational goals towards *online literacy in general* might have benefits regarding mis/disinformation, in terms of helping people understand how information flows online and how mis/disinformation may spread or be targeted at them. For example: educating people about how online systems and information ecosystems work in general; helping people understand how information about them is collected, processed, and used online; helping people understand what it means when they consent to online services.

The question was also raised about whether misinformation education efforts can take lessons from educational efforts in other domains, e.g., cybersecurity.

How to deliver educational interventions? Education might be (for example): embedded as part of platforms themselves (e.g. for a chatbot platform, by the chatbot itself); incorporated into existing courses across different levels of education; presented in stand-alone workshops; incorporated into existing or stand-alone websites; presented as public service announcements

(e.g., similar to public health messaging by public health agencies); included as part of entertainment media (e.g., TV shows); or delivered through educational games or apps (e.g., [5, 14]).

7 Conclusion and Looking Ahead

This Dagstuhl workshop, occurring right after the main assault of the COVID-19 pandemic, was held in hybrid mode. It was quite different from "standard" Dagstuhl meetings which benefit from more informal and direct in-person interactions. Still, thanks to our engaged participants and the quality of the technical video conferencing support provided by the Dagstuhl organisation, we found our conversations fruitful, with much involvement and interactions between geographically (and time zone) distant participants.

It is clear that there are more questions than answers about mis/disinformation in our online ecosystems at this time, and we add our voices to the growing and interdisciplinary research community in this space. We look forward to future – hopefully fully in-person – Dagstuhl seminars on this topic, as well as future collaborations between seminar participants and within the research community more broadly.

References

1. 60 Minutes. Facebook Whistleblower Frances Haugen: The 60 Minutes Interview, October 2021. https://www.youtube.com/watch?v=_Lx5VmAdZSI.
2. Mahmoudreza Babaei, Juhi Kulshrestha, Abhijnan Chakraborty, Elissa M. Redmiles, Meeyoung Cha, and Krishna P. Gummadi. Analyzing biases in perception of truth in news stories and their implications for fact checking. *IEEE Transactions on Computational Social Systems*, 2021.
3. Susanne Barth and Menno D.T.de Jong. The privacy paradox – investigating discrepancies between expressed privacy concerns and actual online behavior – a systematic literature review. *Telematics and Informatics*, 34(7):1038–1058, 2017.
4. Comité National Pilote d'Éthique du Numérique. Ethical issues in the fight against disinformation and misinformation. Ethics oversight bulletin 2, CNPEN/CCNE, July 2020. https://www.ccne-ethique.fr/sites/default/files/cnpen-dsinf_eng.pdf.
5. Chris Coward, Jin Ha Lee, Lindsay Morse, and Travis Windleharth. Misinformation escape room. https://tascha.uw.edu/projects/misinformation-escape-room/.
6. Dictionnaire de l'Académie française. éthique. https://www.dictionnaire-academie.fr/article/A9E2876.
7. doteverone. People, power and technology: the 2018 digital understanding report, April 2018. https://doteveryone.org.uk/2018/04/people-power-and-technology-the-2018-\digital-understanding-report/.
8. Election Integrity Partnership Team. Repeat Offenders: Voting Misinformation on Twitter in the 2020 United States Election, October 2020. https://www.eipartnership.net/rapid-response/repeat-offenders.
9. Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. Datasheets for datasets, 2020.
10. Christine Geeng, Savanna Yee, and Franziska Roesner. Fake news on facebook and twitter: Investigating how people (don't) investigate. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2020.
11. Oana Goga. Investigating ad transparency mechanisms in social media, October 2017. https://lig-membres.imag.fr/gogao/presentations/EJC-Montreal.pdf.

12 Lion Gu, Vladimir Kropotov, and Fyodor Yarochkin. The Fake News Machine: How Propagandists Abuse the Internet and Manipulate the Public. A TrendLabs Research Paper, 2017. https://documents.trendmicro.com/assets/white_papers/wp-fake-news-machine-how-propagandists-abuse-the-internet.pdf.

13 Filippo Menczer and Thomas Hills. Information overload helps fake news spread, and social media knows it. *Scientific American*, December 2020. https://www.scientificamerican.com/article/information-overload-helps-\fake-news-spread-and-social-media-knows-it/.

14 Nicholas Micallef, Mihai Avram, Filippo Menczer, and Sameer Patil. Fakey: A game intervention to improve news literacy on social media. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), April 2021.

15 Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Jan 2019.

16 Nuno Mota, Abhijnan Chakraborty, Asia J. Biega, Krishna P. Gummadi, and Hoda Heidari. On the desiderata for online altruism: Nudging for equitable donations. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), oct 2020.

17 Michel Tuan Pham and Gita Venkataramani Johar. Contingent Processes of Source Identification. *Journal of Consumer Research*, 24(3):249–265, 12 1997.

18 Francesco Pierri, Brea Perry, Matthew R. DeVerna, Kai-Cheng Yang, Alessandro Flammini, Filippo Menczer, and John Bryden. The impact of online misinformation on U.S. COVID-19 vaccinations. *CoRR*, abs/2104.10635, April 2021.

19 Margaret Earling Roberts. *Fear, Friction, and Flooding: Methods of Online Information Control*. PhD thesis, Harvard University, 2014.

20 Lucas Ropek. Facebook Banned the Creator of 'Unfollow Everything' and Sent Him a Cease and Desist Letter, October 2021. https://gizmodo.com/facebook-banned-the-creator-of-unfollow-everything-and-1847826505.

21 Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. A framework of severity for harmful content online. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), October 2021.

22 Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature Communications*, 9, 2018.

23 Jon Sharman. Hackers targeted University of Oxford's Covid vaccine research, cyber spies reveal, November 2021. https://www.independent.co.uk/news/uk/home-news/covid-vaccine-hack-cyber-oxford-b1959147.html.

24 Daniel J. Solove. The myth of the privacy paradox. 89 George Washington Law Review 1 (2021), GWU Legal Studies Research Paper No. 2020-10, GWU Law School Public Law Research Paper No. 2020-10, January 2021. https://ssrn.com/abstract=3536265.

25 Pablo Suárez-Serrato, Margaret E. Roberts, Clayton Davis, and Filippo Menczer. On the influence of social bots in online protests. In Emma Spiro and Yong-Yeol Ahn, editors, *Social Informatics*, pages 269–278. Springer International Publishing, 2016.

26 Nevin Thompson. Hacks/Hackers, Partners Awarded Funding to Participate in the 2021 National Science Foundation's Convergence Accelerator, September 2021. https://newsq.net/2021/09/22.

27 TrendMicro. Internet safety for kids & families. https://www.trendmicro.com/internet-safety/.

28 University of Washington Center for an Informed Public. Misinfoday. https://www.cip.uw.edu/get-involved/misinfoday/.

29 Pak-Hang Wong. google scholar page. `https://scholar.google.com/citations?hl=en&user=e4yJCwcAAAAJ&view_op=list_works&sortby=pubdate`.

30 Taha Yasseri and Filippo Menczer. Can the wikipedia moderation model rescue the social marketplace of ideas? *CoRR*, abs/2104.13754, 2021.

31 Eric Zeng, Tadayoshi Kohno, and Franziska Roesner. What makes a "bad" ad? user perceptions of problematic online advertising. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2021.

Participants

- Camille Darche
French Nat. Pilot Committee for Dig. Ethics – Paris, FR
- Lynda Hardman
CWI – Amsterdam, NL
- Claude Kirchner
INRIA – Le Chesnay, FR
- Jean-Yves Marion
CNRS – Nancy, FR

Remote Participants

- Esma Aimeur
University of Montreal, CA
- Jos Baeten
CWI – Amsterdam, NL
- Asia J. Biega
MPI-SP – Bochum, DE
- Sébastien Gambs
University of Montreal, CA
- Krishna Gummadi
MPI–SWS – Saarbrücken, DE
- Vladimir Kropotov
Trend Micro – Garching, DE
- Evangelos Markatos
FORTH – Heraklion, GR
- Filippo Menczer
Indiana University – Bloomington, US
- Trisha Meyer
Free University of Brussels, BE
- Franziska Roesner
University of Washington – Seattle, US
- Kavé Salamatian
University of Savoie – Annecy le Vieux, FR
- Juliette Sénéchal
University of Lille, FR
- Dimitrios Serpanos
ATHENA Research Center – Patras, GR
- Serena Villata
Université Côte d'Azur – Sophia Antipolis, FR

Report from Dagstuhl Seminar 21411

Machine Learning in Sports

Edited by

Ulf Brefeld[1], Jesse Davis[2], Martin Lames[3], and James J. Little[4]

1 **Universität Lüneburg, DE,** brefeld@leuphana.de
2 **KU Leuven, BE,** jesse.davis@cs.kuleuven.be
3 **TU München, DE,** martin.lames@tum.de
4 **University of British Columbia – Vancouver, CA,** little@cs.ubc.ca

Abstract

Data about sports have long been the subject of research and analysis by sports scientists. The increasing size and availability of these data have also attracted the attention of researchers in machine learning, computer vision and artificial intelligence. However, these communities rarely interact. This seminar aimed to bring together researchers from these areas to spur an interdisciplinary approach to these problems. The seminar was organized around five different themes that were introduced with tutorial and overview style talks about the key concepts to facilitate knowledge exchange among researchers with different backgrounds and approaches to data-based sports research. These were augmented by more in-depth presentations on specific problems or techniques. There was a panel discussion by practitioners on the difficulties and lessons learned about putting analytics into practice. Finally, we came up with a number of conclusions and next steps.

Seminar October 10–15, 2021 – http://www.dagstuhl.de/21411
2012 ACM Subject Classification Computing methodologies → Machine learning; Computing methodologies → Computer vision
Keywords and phrases machine learning, artificial intelligence, sports science, computer vision, explanations, visualization, tactics, health, biomechanics
Digital Object Identifier 10.4230/DagRep.11.9.45

1 Executive Summary

Ulf Brefeld
Jesse Davis
Martin Lames
Jim Little

Sports has become an incredibly data rich field with the advent of data sources such as event data (e.g., time and locations of actions), tracking data (i.e., positional data), and athlete monitoring (e.g., bio-sensors, IMUs, GPS). These data are commonly and widely collected across multiple different sports, both on a professional and recreational level. The advent of such data raises the need to exploit the collected data both from the theoretical (e.g., sports modeling) as well as practical (e.g., training in top level sports) perspective. Problem-solving solutions can only be provided by an interaction between the sports science & informatics (S&I) and the machine learning (ML) communities. Machine learning is emerging as a powerful, new paradigm for sports analytics, as it provides novel approaches to making sense

Machine Learning in Sports, *Dagstuhl Reports*, Vol. 11, Issue 09, pp. 45–63
Editors: Ulf Brefeld, Jesse Davis, Martin Lames, and Jim Little
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

of the collected data. However, the S&I and ML communities are traditionally separate, each with its own agenda. The seminar aims to bring together top researchers and practitioners who are active in these two fields that can contribute to an assessment of their potential synergies.

We structured the seminar along five different themes, each of which was the focus of between half and a full day. Given the diversity of the participants' backgrounds in terms of discipline, each theme began with overview to get everyone on the same page. Then there were more detailed presentations. The five themes were:

Machine learning meets sports The goal of this session was to provide an overview of some of the machine learning techniques (predictive modeling, text mining) and how they can be applied in the sports context. The illustrative applications where ML can play a role included assessing the performances of teams and players, supporting sport broadcast, assessing fans reactions to rule changes and helping reduce the time burden on video analysts.

Sports science meets machine learning The goal was to provide an overview of basic concepts in sports science to inform researchers from machine learning. The basic concepts were the relation between competition, training and athlete's abilities, the structure of performance in different sports, and the demand for support in sports practice. In particular, the structure of team sports as dynamic interaction processes with emergent behavior was explained as this the most frequent application field for machine learning in sports.

Computer vision for sports The session aimed to expose the participants to the practice of gathering information about team sports through analysis of visual information. The session began with an overview of the general practice of computer vision for sports. Three of the presenters are from industry, representing companies with significant presence in the business of providing to analytics producers information on team sports such as basketball, football (soccer), and ice hockey. The fourth presenter, from academia, discussed material on camera planning and analytics and in addition has himself been involved in tech transfer of visual analytics methods from amateur sports. The overall goal, i.e., informing the participants regarding methods and applications of vision, was well met by the lectures of these experienced researchers.

Interdisciplinary view on tactics The session aimed to build a common understanding of tactics and their implementation in predictive/generative models. It is still an open question how to represent overarching long-term strategies in computer models and different ideas were discussed on the example of overview and contributed presentations.

Explaining, interpreting, and visualizing models and data for sports A key challenge is effectively conveying the results of machine learned models to domain experts, which is compounded by the black-box nature of many such models. This session highlighted a variety of techniques for meeting this objective, with illustrative examples arising from practice were shown for a variety of sports such as ice hockey, table tennis and football. This remains an active area of research and variety of lessons learned and ideas for improving the communication between domain-experts and technical-experts were discussed.

2 Table of Contents

3 Overview of Talks

3.1 Quantifying Offensive Actions, Tactics, and Strategies

Gabriel Anzer (Hertha BSC Berlin, DE)

Offensive performances in football have always been of great focus for fans and clubs alike as evidenced by the fact that nearly all Ballon d'Or winners have been forwards or midfielders. With the increase in availability of granular data, evaluating these performances on a deeper level than just goals scored or gut instinct has become possible. The domain of sports analytics has recently emerged, exploring how applying data science techniques or other statistical methods to sports data can improve decision making within sporting organizations. This thesis follows the footsteps of other sports like baseball or basketball where, at first, offensive performances were analyzed. It consists of four studies exploring various levels of offensive performance, ranging from basic actions to team-level strategy. For that, it uses a data set part of larger research program that also explores the automatic detection of tactical patterns. This dataset mainly consists of positional and event data from eight seasons of the German Bundesliga and German Bundesliga 2 between the seasons 2013/2014 and 2020/2021. In total this amounts to 4, 896 matches, with highly accurate player and ball positions for every moment of the match and detailed logs of every action that occurred, thus making it one of the largest football datasets to be analyzed at this level of granularity. In a first step, this thesis shows how the two different data sources can be synchronized. With this synchronized data it is possible to better quantify individual basic actions like shots or passes. For both actions new metrics (Expected Goals and Expected Passes) were developed, that use the contextual information to quantify the chance quality and passing difficulty. Using this improved quantification of individual actions, the subsequent studies evaluate offensive performance on a tactical pattern level (how goals are scored) and on a strategy level (what team formations are particular effective offensively). Besides their usage on the performance side, these metrics have also been adapted from broadcasters to enhance their data story telling: Expected goals and expected passes are shown during every Bundesliga match to a worldwide audience, thus bringing the field of sports analytics to millions of fans.

3.2 Automated Detection of Complex Tactical Patterns in Football

Pascal Bauer (Deutscher Fußball Bund – Frankfurt am Main, DE)

Football tactics is a topic of public interest, where decisions are predominantly made based on gut instincts from domain-experts. Sport science literature often highlights the need for evidence-based research on football tactics, however the limited capabilities in modelling the dynamics of football has prevented researchers from gaining usable insights. Recent technological advances have made high quality football data more available and affordable. Particularly, positional data providing player and ball coordinates at every instance of a match can be combined with event data containing spatiotemporal information on any event taking place on the pitch (e.g. passes, shots, fouls). On the other hand, the application of machine

learning methods to domain-specific problems yields a paradigm shift in many industries including sports. The need for more informed decisions as well as automating time consuming processes–accelerated by the availability of data–has motivated many scientific investigations in football analytics. This thesis is part of a research program combining methodologies from sports and data-science to address the following problems: the synchronization of positional and event data, objectively quantifying offensive actions, as well as the detection of tactical patterns. Although various basic insights from the overall research program are integrated, this thesis focuses primarily on the latter one. Specifically, positional and event data are used to apply machine learning techniques to identify eight established tactical patterns in football: namely high–/mid–/low–block defending, build-up/attacking play in the offense, counterpressing and counterattacks during transitions, and patterns when defending corner-kicks, e.g. player–/zonal– or post-marking. For each pattern, we consolidate definitions with football experts and label large amounts of data manually using video recordings. The inter-annotator reliability is used to ensure that each pattern is well-defined. Unsupervised techniques are used for the purpose of exploration, and supervised machine learning methods based on expert-labeled data for the final detection. As an outlook, semi-supervised methods were used to reduce the labeling effort. This thesis proves that the detection of tactical patterns can optimize everyday processes in professional clubs, and leverage the domain of tactical analysis in sport science by gaining unseen insights. Additionally, we add value to the machine learning domain by evaluating recent methods in supervised and semi-supervised machine learning on challenging, real-world problems.

3.3 From pixels to points: Using tracking data to measure performance in professional sports

Luke Bornn (Zelus Analytics, USA)

In this talk I will explore how players perform, both individually and as a team, on a basketball court. By blending advanced spatiotemporal models with geography-inspired mapping tools, we are able to understand player skill far better than either individual tool allows. Using optical tracking data consisting of hundreds of millions of observations, I will demonstrate these ideas by characterizing defensive skill and decision making in NBA players.

3.4 Understanding the women's football game using machine learning techniques

Lotte Bransen (SciSports – Amersfort, NL)

The women's football game has made huge advances in recent years. Despite its increased popularity, there has been less analysis of data arising from the women's game. This talk provides a summary of our initial work on analyzing the technical data being collected during professional women's football matches. Using event data covering a number of seasons from

the top women's leagues, this talk presents three analyses. First, we perform an exploratory analysis by computing several technical indicators (e.g., goal scoring rates over the season, conversion rates, shot locations) and then compare and contrast them to the indicators for comparable men's leagues and find several intriguing differences. Second, we assess whether expected goals (xG) models on one gender are applicable to data from a different gender. Third, we present some preliminary analyses about differences in where and how often women perform certain actions compared to men.

3.5 A computational model for quantifying Availability in soccer

Uwe Dick (Sportec Solutions – Ismanning, DE) and Daniel Link (TU Munich, DE)

The paper presents a computational approach to Availability of soccer players. Availability is defined as the probability that a pass reaches the target player without being intercepted by opponents. Clearly, a computational model for this probability grounds on models for ball dynamics, player movements, and technical skills of the pass giver. Our approach aggregates these quantities for all possible passes to the target player to compute a single Availability value. Empirically, our approach outperforms state-of-the-art competitors using data from 58 professional soccer matches. Moreover, our experiments indicate that the model can even outperform soccer coaches in assessing the availability of soccer players from static images.

3.6 AI videography for amateur hockey

James Elder (York University – Toronto, CA)

Research on computer vision for sports tends to focus on broadcast feeds of professional matches. However, computer vision and AI may have even greater application in amateur sport, for reasons of scale and budget. For hockey, there are 1,000 registered amateur players in North America for every player in the NHL, and amateur leagues typically cannot afford to hire a team of professional videographers for every game. These factors motivate the development of low-cost AI videography systems for amateur hockey; in this talk I will consider three research problems that are central to this goal. The first problem is spatial attention. While hockey is played over a large surface area, instantaneous play is highly localized and fans in the stands use their visual attention and oculomotor systems to continuously follow the play with their eyes. Here I will describe our research efforts to use spatial attention to create a comparable viewing experience for those watching from home. The resulting attentive computer vision system dynamically tracks the play in an 8K wide-field video, automatically extracting a dynamic HD zoomed video streamed to fans in real time. The second problem is temporal attention. While hockey games typically comprise 60 minutes of regulation play, stoppages mean that games take 140 minutes to complete. I will describe a temporal attention system that fuses visual and auditory cues to automatically extract the periods of active play from a recorded video, thus allowing a complete game to be enjoyed offline within a shorter timeframe. The third problem is automatic labeling of

players according to team, which is essential for computing play statistics. I will describe an unsupervised learning strategy that allows reliable team assignment even for novel teams, enabling our system to be deployed broadly without retraining.

3.7 AI for sports and health

Björn Eskoffier (FAU Erlangen-Nürnberg, DE)

Wearable computing systems play an increasingly important role in recreational and elite sports. They comprise of two parts. First, sensors for physiological (ECG, EMG, ...) and biomechanical (accelerometer, gyroscope, ...) data recording are embedded into clothes and equipment. Second, embedded microprocessors (e.g. in smartphones) are used for monitoring and analysis of the recorded data. Together, these systems can provide real-time information and feedback for scientific studies in real sports situations. In order to implement these systems, several challenges have to be addressed. Our work focuses on four of the most prevalent of these: (1) Integration: sensors and microprocessors have to be embedded unobtrusively and have to record a variety of signals. (2) Communication: sensors and microprocessors have to communicate in body-area-networks in a secure, safe and energy-saving manner. (3) Interpretation: physiological and biomechanical data have to be interpreted using signal processing and machine learning methods. (4) Simulation and modeling: understanding of sensor data is needed to model processes in sports more accurately, simulation methodologies help here to provide basic information to drive those models. Data mining concepts provide tools for analyzing the considerable amount of physiological and biomechanical data that is generated in sports science studies. Especially when using wearable computing systems, the number of participants and variety of measured data is unlimited in general. Traditional statistical analysis methods commonly cannot handle this amount of data easily. Thus, the analysis is often restricted to individual variables rather than multidimensional dependencies and a considerable amount of information is neglected. Moreover, the results are frequently biased by the expectation of the researcher. Here, the objective, data-driven methods from data mining can contribute by offering useful tools for the analysis tasks. These tools have the ability to deal with large data sets, to analyze multiple dimensions simultaneously, to work data-driven rather than hypothesis-driven, and to provide valuable insights into training effects and injury risks.

3.8 Labeling Situations in Soccer

Dennis Fassmeyer (Leuphana University of Lüneburg, DE)

We study the automatic annotation of situations in soccer games. At first sight, this translates nicely into a standard supervised learning problem. However, in a fully supervised setting, predictive accuracies are supposed to correlate positively with the amount of labeled situations: more labeled training data simply promise better performance. Unfortunately, non-trivially annotated situations in soccer games are scarce, expensive and almost always

require human experts; a fully supervised approach appears infeasible. Hence, we split the problem into two parts and learn (i) a meaningful feature representation using variational autoencoders on unlabeled data at large scales and (ii) a large-margin classifier acting in this feature space but utilize only a few (manually) annotated examples of the situation of interest. We propose four different architectures of the variational autoencoder and empirically study the detection of corner kicks, crosses and counterattacks. We observe high predictive accuracies above 90% AUC irrespective of the task.

3.9 Live acquisition of tracking data in soccer

Eric Hayman (ChyronHego – Stockholm, SE)

Tracab is a leading provider of tracking data in soccer, estimating the positions of all players, referees and the ball in real time at 25 times per second. In this talk I will describe our optical tracking system; first launched 15 years ago and which has seen continuous enhancements since. Given that the Dagstuhl seminar contains a number of participants from the sports science community, who consume centre-of-mass optical player tracking data in their work, I will place particular emphasis on the strengths and limitations of optical tracking systems. The motivation is to give those participants a richer understanding of the data they receive. I will highlight how modern deep learning techniques are exploited in our system to enhance both accuracy and automation, but will also emphasize the role of traditional computer vision methods. Moreover, I will describe how human intelligence, in the form of system operators, has been used to complement the strengths of computer algorithms, and how operator workload has been reduced as the technology matures. In the second part of the talk I will describe recent extensions of the system which go beyond centre-of-mass estimation. In particular I will describe how modern deep learning methods permit 3D limb tracking, and how this can be exploited to vastly improve the speed and accuracy of semi-automated offside decision systems to enhance the VAR workflow in soccer. Finally, I will discuss how machine learning techniques can provide a richer stream of player data based on body pose. In the future, this new data should prove beneficial for performance analysis.

3.10 Sport analytics: From pixels to useful metrics

Mehrsan Javan (Sportlogiq, Canada)

Sports analytics is about observing, understanding, and describing the game in an intelligent manner. In practice, most of the focus has been on visual perception to take the video data and extract tracking data and game events. However, turning the incomplete data into actionable insights for the clubs has always been a challenge. This talk focuses on the use of broadcast feed for sport analytics, covers the components of a vision system for data acquisition, provides examples of how Sportlogiq captures the data from broadcast videos and turns them into useful insights for the clubs to make better decisions.

3.11 Providing Personalized training insights in cycling for Team Jumbo Visma

Arno Knobbe (University of Leiden, NL)

In order to gauge how successful their training efforts are, elite sports teams would like to estimate often how fit their athletes are. There are a number of standardized tests that achieve just that, but these tests come with a number of downsides, having to do with the intrusive nature of the tests and the impact that a test has on the intended training schedule. For that reasons, coaches avoid doing tests more than, say, once per month. However, ideally you would be able to get an estimate over several key performance indicators about fitness on a daily basis. This talk describes attempts to achieve this in elite road cycling. Over the last years, power meters have become quite standard in cycling. Combined with the easily obtainable heart rate readings, a fairly detailed source of information becomes available about the work performed on the bike, and the response of the body to this exertion. Normally, one would expect a monotonic relationship between the power output and the heart rate, but with varying exertion over time (for example due to hilly terrain), this relationship is not so clear. We describe techniques to model this data more accurately by integration over time. We will demonstrate that using the right physiological model of the heart's response to external factors, the instantaneous heart rate can be modeled quite effectively. It turns out that parameters of this model fluctuate over time (on the scale of days to weeks), as a function of how tired or fit a rider is, and thus can be used as good indicators of fitness. We give examples of how the heart rate model and its (personalized) parameters can be used to answer several questions in the optimal training of riders.

3.12 Text mining and performance analysis

Otto Kolbinger (TU Munich, DE)

As a discipline originating from notational analysis and biomechanics, performance analysis is dominated by research based on metrical data like match-statistics or physiological measures. However, a lot of evaluations of performance are provided in the form of written text, for example, scouting reports or (pre-) match analysis. Innovative text mining techniques enable researchers to derive knowledge from such written texts in an effective manner but have hardly been applied in performance analysis so far. During my talk, I presented some promising applications of text mining in three areas: the evaluation of technological officiating aids, player scouting and predicting match outcome. For each, I intended to demonstrate how deriving information from text sources can contribute to the knowledge base in the respective area. However, the aim of this talk was twofold. Besides showing such promising approaches, I also pointed at the current lack of universally deployed evaluation standards for classifiers and shortcomings in reporting the performed approaches. An issue not just of text mining – but also of machine learning in applied science in general.

3.13 Sports science for machine learners

Martin Lames (TU Munich, DE)

The idea of this session is to support the interdisciplinary character of the seminar that addresses informaticians as well as sports scientists. Concretely it is about informing informaticians on theories and concepts of sport science. The general areas of training, athletes' capabilities and competition are introduced with a specific emphasis on their interactions. Here, the idea is expressed that support for training can be given by identifying the relevant capabilities responsible for performance in competition. The different structures of performance in the different groups of sports are mentioned, emphasizing that in game sports (team or net sports) there are specific opportunities for machine learning support, because there, we need to decode the relation between action plans and behavioral outcome. The general notion of game sports as dynamic interaction processes with emergent behavior is explained and underpinned with empirical results on the negotiation of match intensity, emergence of behavior, and the impact of chance on goal scoring. The differences between theoretical and practical performance analysis (TPA and PPA) are listed with respect to their main distinguishing aspects. TPA looks for lawlike structures of the sports, whereas in PPA the practical support is dominant. In the former, large data bases on positional and event data create an interesting experimenting ground for informatics, while in the later the challenge lies in the demand for integrating data from many sources in a knowledge base. Each field contains opportunities for machine learning applications if the respective problems are perceived and targeted at.

3.14 The next frontier in sports analytics: Collecting and utilizing tracking data from broadcast video

Patrick Lucey (Stats Perform – Chicago, USA)

Sports Analytics can be divided into two eras: i) the data-driven era, and ii) the AI-driven era. Due to the success of making data-driven decisions, the demand for more granular data and better insights has ushered in the AI-driven era, where computer vision systems are used to collect player/ball tracking data and machine learning to utilize these granular data-sources. Although effective, the AI-Driven era is not yet pervasive due to the hardware requirements of currently deployed computer vision systems (i.e., they need to be in-venue with fixed cameras), which limits the amount of tracking data that can be collected going forward but also historically. However, with the improvement of computer vision technology this is changing. In this talk, I will highlight our work in collecting tracking data from thousands of college basketball broadcast videos and how we can utilize the data collected to make predictions of future NBA talent. I will also highlight current challenges we are working on as we scale out our computer vision solutions. Additionally, I will showcase our recent work in soccer where we predict performance of players on other teams and leagues.

3.15 Diagnostics and training data at the IAT

Björn Mäurer (IAT – Leipzig, DE)

The Institute for Applied Training Science in Leipzig is the central research institute for German elite and junior competitive sport. The main tasks of the institute are performance diagnostics, training analysis and training research. In analysis, we pursue different approaches. Among other things, a 3D analysis of videos is useful and important to be able to analyze the athletes' movements precisely. To speed up the analysis (it is a time-consuming work), we have supplemented the current programme "Mess3d" with a pre-learned software. Here, the athletes' joint points are recorded automatically. Only in case of large deviations of the visual axes of the two cameras (skewed straight lines), a manual check by the scientist is necessary. In this way, a time saving of about 2/3 could be achieved. However, this is very dependent on the type of sport. To achieve a good analysis, you need as much data as possible. These are training, performance diagnostics and competition data. We have developed the IDA software to facilitate the exchange of the necessary data between athletes, coaches and scientists. It enables the institute to obtain the most comprehensive data possible from the athletes. But also to provide coaches and athletes with quick, detailed analyses. The entire digital exchange can thus be handled within one software. This also makes daily communication between all participants much easier. Currently there are about 30 instances of the software with about 5 million completely recorded events. The large amount of data, especially from top-level sport, enables a higher-level analysis. For this task, a machine-based approach is interesting. Therefore, we are currently trying to go down this path. We are always open for cooperation with other scientific institutions.

3.16 Computer vision and AI for sports boadcasting

Fabrizio Pece (Vzrt – Zürich, CH)

Vizrt is a leading provider of cutting-edge solutions for the broadcasting industry. Such solutions allow broadcasters to create and share visually compelling stories, including those concerning live televised sporting events. More specifically, Vizrt's sports production solutions enhance the abilities of broadcasters, coaches, and teams with tools to inform audiences – whether they are at home, in the stands, or in locker rooms. Said tools are built upon state of the art computer vision and machine learning algorithms, and are fused together to create powerful, but yet intuitive solutions to enrich sport event storytelling. In this talk, I will first introduce Vizrt sport production tools, illustrating how computer vision and machine learning algorithm can be used to perform interactive game analysis, real-time, augment-reality graphics insertion, as well as data integration and visualization. The second part of the talk will focus on the challenges of adapting research results to industrial settings, showing how the journey from research to production requires a deep domain understanding to solve domain-specific challenges which are otherwise seldomly tackled in academic settings.

3.17 Winning in sports with data and machine learning

Kostas Pelechrinis (University of Pittsburgh, USA)

Basketball is currently second only to baseball when it comes to integrating data in team operations, decision making and game preparation. While data and statistical analysis have been always part of basketball operations, the availability of detailed player tracking data as well as, additional contextual meta-data, have pushed the envelope further. In this talk, I will start with a specific case study facilitated by player tracking data. In particular, I will present an analysis of how the corner 3-point shots are created and what makes them the second most efficient shot type in the NBA. In the second part of the talk, I will present a more general framework that models the movements of players on the court and tracks the expected points to be scored in real-time. This type of model allows us to evaluate micro-actions such as screens, passes, etc.,that traditionally have been hard to evaluate. Finally, I will briefly discuss other applications of machine learning in basketball.

3.18 HI v. AI in athletic development

Martin Rumo (OYM – Cham, Switzerland)

This presentation argues from the perspective of performance optimization in Sports practice. It explores constraints to the use of machine learning or artificial intelligence concepts in sport practice and argues to formalize human understanding of competition and required skills to enhance the effectiveness of data analytics in sport. Sport is a complex and dynamic domain in which there is a clear goal to optimize nevertheless, the factors which lead to optimal performance are not clear cut. Since sport does not allow any explicit element of luck all factors involved in performance optimization can be considered as skill. Some skill sets, like developing physical strength at a fast rate and doing this repeatedly, are easily measurable and there are lab and field tests to determine those skills in athletes. Since those kinds of basic skills are easily measurable, there is plenty of data available but machine learning concepts do not really contribute to new insights here. Some sports like invasion games are more complex and coaches and athletes elaborate shared concepts of the competition that are not easily represented in data. This introduces a form of semantic gap between the concepts and the available data. This is an interesting area for Machine Learning concepts, but unfortunately the more complex those patterns of events representing the shared concepts become, the less standardized data is available. This is an impediment to training robust Machine Learning models. Furthermore, in sport practice decision making processes underlie an explore-exploit reality, where decisions are not necessarily optimized by exploiting available information, but rather they are motivated by exploring new ways and learning from the resulting outcome. Machine Learning has proven to be very efficient in data generation, especially in computer vision. But using machine learning algorithms in the realm of more abstract structure of the game is more difficult and formalizing human understanding of the underlying structure of the competition seems more effective in providing actionable information for sports practitioner. The presentation explores basic approaches to meet these challenges when delivering real value to the practitioner.

It is recommended to use the discussed approach when practitioners and computer scientists are co-creating valuable data products. Special care should be given to user experience and visualization, since communicating the information is as important in practice as generating it.

3.19 Pitch Control

William Spearman (Liverpool FC, GB)

The presentation introduces recently developed concepts for measuring the importance of space in football. Traditional concepts like controlled zones can be aggregated on a team level to show dominated areas (pitch control). Clearly, controlled space is not equally value. As a proxy, the positioning of the defending team may serve as the basis for computing a measure on space quality. Finally, both metrics can be combined to obtain a computational understanding of control in desired areas of the pitch.

3.20 Biomechanically inspired machine learning to improve performance in biomechanics

Benedicte Vanwanseele (KU Leuven, BE)

The aim of our research is to develop insights as well as tools to come to personalized adaptable rehabilitation and training regimens to enable every individual to physically perform to the best of their own ability. We know that the human body is a complex system where mental, cardiorespiratory and musculoskeletal system is loaded during sport and exercise. In this talk we will focus on the musculoskeletal system as we know that training adaptations to this system are slower compared to the cardiorespiratory system and therefore musculoskeletal overuse injuries are a big challenge in training and sports. To improve performance the human system needs to be load enough but not too much so without increasing the risk of injury. As the musculoskeletal load is the product of volume and magnitude but magnitude is a lot more important it is crucial that we develop method that are able to monitor the load magnitude and how this varies during training. In this talk we will discuss how we develop a method based on a single trunk-based accelerometer to monitor musculoskeletal loading during training and as well as changes within a training session. The second problem is that we cannot really share the code either. The first author, who also implemented everything, has left academia in summer. Unfortunately, the code is in no shape to think of a release in its current form. We will certainly help interested colleagues and may also send the code in private communication but a public release is unfortunately not an option.

3.21 Visual Analytics of Sports Data

Yincai Wu (Zhejiang University - Hangzhou, CN)

With the rapid development of sensing technologies and wearable devices, large sports data have been acquired daily. The data usually implies a wide spectrum of information and rich knowledge about sports. Visual analytics, which facilitates analytical reasoning by interactive visual interfaces, has proven its value in solving various problems. In this talk, I will discuss our research experiences in visual analytics of sports data and introduce several recent studies of our group of making sense of sports data through interactive visualization.

3.22 Visual Analysis of Table Tennis Tactics

Hui Zhang (Zhejiang Universitity – Hangzhou, CN)

Table tennis is a skillful sport, thus, techniques and tactics are the core factors for winning matches. Therefore, coaches, players, and researchers have always paid attention to technique innovations and tactic analysis. In recent years, we have conducted a series of studies on table tennis match data collection, analysis, evaluation, mining, simulation, and video augmentation, mainly including: 1) We proposed a data collection framework to facilitate interactive annotation of table tennis match videos with the support of computer vision algorithms; 2) To further analysis and visual explore the tactics of table tennis matches, a novel interactive visualization system was developed, which provides a holistic visualization of an entire match from three main perspectives, namely, time-oriented, statistical, and tactical analyses; 3) Stroke evaluation is critical for coaches to evaluate players' performance in table tennis matches. For this reason, we proposed an automatic stroke evaluation framework. In particular, to integrate analysts' knowledge into the machine learning model, we employed the latest effective framework named abductive learning, showing promising performance. Based on abductive learning, the system combines the state-of-the-art computer vision algorithms with analysts' knowledge to extract and embed stroke features for evaluation; 4) Tactical analysis in table tennis is challenging as the analysts can often be overwhelmed by the large quantity and high dimension of the data, to address these issues, we designed a visual analytics system to allow analysts to effectively analyze, explore, and compare tactics of multiple matches based on the advanced embedding and dimension reduction algorithms along with an interactive glyph; 5) Simulative analysis in competitive sports can provide prospective insights, which can help improve the performance of players in future matches. We propose a well-established hybrid second-order Markov chain model to characterize and simulate the competition process in table tennis. Compared with existing methods, our approach supports the effective simulation of tactics, which represent high-level competition strategies in table tennis; 6) Visualizing data in sports videos is gaining traction in sports analytics, given its ability to communicate insights and explicate player strategies engagingly. So, we design fast prototyping tool, to ease the creation of augmented table tennis videos by leveraging machine learning-based data extractors and design space-based visualization recommendations. With the system, analysts can create an augmented video by selecting the table tennis match data to visualize instead of manually drawing the graphical marks.

3.23 Using machine learning to assess and compare athletes in team sports

Albrecht Zimmermann (University of Caen, FR)

Using machine learning techniques (ML) to assess action (and derived from it player) quality is a recent and promising alternative to expert-based assessments. The big challenges for these approaches consist of data acquisition, data transformation and augmentation, which often require in-depth knowledge of the sport in question, as well as understanding which ML techniques are well-suited not only for final modeling but also for data preparation.

4 Panel discussions

We organized a panel on the challenges of putting data science insights into practice. Therefore, we asked a number of attendees with experience in this regard to serve as panelists: Luke Bornn (Zelus Analytics), Lotte Bransen (SciSports), Jan Van Haaren (FC Brugge), Mehrsan Javan (Sportlogic), Patrick Lucey (STATS), Benedicte Vanwanseele (KU Leuven).

Before the panel, we solicited topics from the attendees. Based on the response we posed the following questions to the panel:

- What are the key challenges and opportunities for applying machine learning to sports problems in sports beyond football?
- It is often argued that each athlete (or team) is unique, which indicates that a personalized modeling approach would be ideal. However, machine learning works best when there is lots of data, which suggests pooling data across athletes is the way to go. How do you navigate this trade-off between developing a personalized model vs. one model that serves all?
- What challenges did you face when communicating insights from data to practitioners? How did you overcome them?
- What are the key open opportunities for machine learning in physiological monitoring?
- What are the key open opportunities for machine learning in tactical analyses?
- Who do you see as key end users of the advanced analyses? For example fans, teams, coaches, media, ...?
- Particularly for physical monitoring, we have advanced techniques that yield very accurate measurements in a lab setting. However, measuring in the wild during competitions is more challenging. How do we know that what that we find in the lab translates to a more ecological setting?

5 Results

During the seminar, we identified the following actions and activities to establish a forum for future exchange and bringing the disciplines together:

1. We endeavor to participate in each other's regular conferences and meeting such as International Symposium on Computer Science in Sport, Machine Learning and Data Mining for Sports Analytics, Computer Vision for Sports, etc. We will try to facilitate combined events to continue the interaction.

2. To facilitate this initiative, we will set up a mailing list of those that are interested in conferences, seminars, job opportunities, etc. The list is: ml-ai-4sports@googlegroups.com
3. We will aim to setup a reoccurring online seminar series with sessions every four to six weeks.
4. We will explore the possibility of securing funding to establish a network. Arnold Baca has some experience in this regard and graciously offered to investigate it.

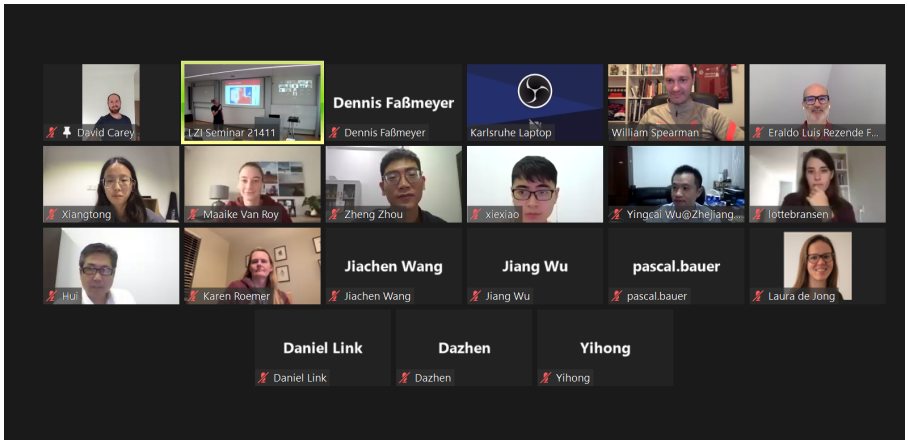
Participants

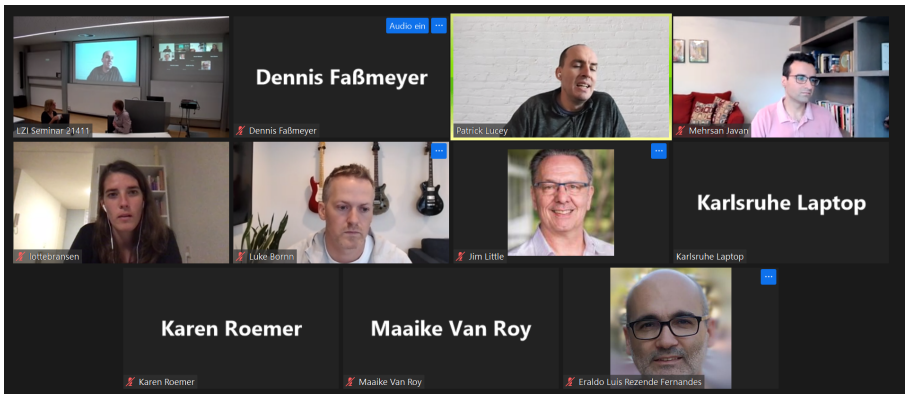
- Gabriel Anzer
Hertha BSC – Berlin, DE
- Arnold Baca
Universität Wien, AT
- Pascal Bauer
DFB – Frankfurt, DE
- Ulf Brefeld
Universität Lüneburg, DE
- Jesse Davis
KU Leuven, BE
- Björn Eskofier
Universität Erlangen-Nürnberg, DE
- Dennis Faßmeyer
Leuphana Universität Lüneburg, DE
- Eric Hayman
ChyronHego – Stockholm, SE
- Arno J. Knobbe
Leiden University, NL
- Otto Kolbinger
TU München, DE
- Philipp Kornfeind
Universität Wien, AT
- Martin Lames
TU München, DE
- Daniel Link
TU München, DE
- Björn Mäurer
IAT – Leipzig, DE
- Fabrizio Pece
Vizrt AG – Zürich, CH
- Martin Rumo
OYM AG – Cham, CH
- Tiago Guedes Russomanno
University of Brasilia, BR
- Marc Schmid
TU München, DE
- Jan Van Haaren
FC Bruges, BE
- Benedicte Vanwanseele
KU Leuven, BE
- Albrecht Zimmermann
Caen University, FR

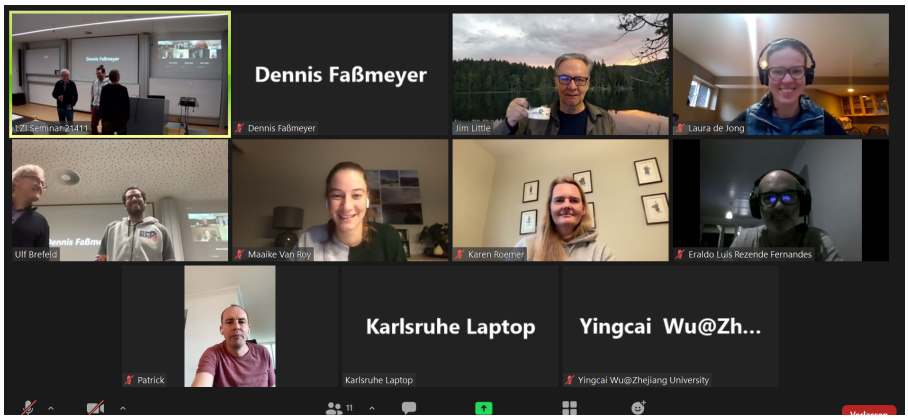
Remote Participants

- Luke Bornn
Simon Fraser University – Burnaby, CA
- Lotte Bransen
SciSports – Amersfoort, NL
- Mirjam Bruinsma
AFC Ajax – Amsterdam, NL
- David Carey
La Trobe University – Melbourne, AU
- Xiangtong Chu
Zhejiang University, CN
- Laura de Jong
Deakin University – Melbourne, AU
- Uwe Dick
Sportec Solutions AG – Ismaning, DE
- James Elder
York University – Toronto, CA
- Irfan A. Essa
Georgia Institute of Technology – Atlanta, US
- Mehrsan Javan
Sportlogiq – Montréal, CA
- Jim Little
University of British Columbia – Vancouver, CA
- Patrick Lucey
STATS Perform – Chicago, US
- Konstantinos Pelechrinis
University of Pittsburgh, US
- Eraldo Luis Rezende Fernandes
Leuphana Universität Lüneburg, DE

- Karen Roemer
Central Washington University – Ellensburg, US
- Yannick Rudolph
Leuphana Universität Lüneburg, DE & SAP SE – Berlin, DE
- Oliver Schulte
Simon Fraser University – Burnaby, CA
- William Spearman
Liverpool Football Club, GB
- Karl Tuyls
DeepMind – London, GB
- Maaike Van Roy
KU Leuven, BE
- Jiang Wu
Zhejiang University – Hangzhou, CN
- Yingcai Wu
Zhejiang University – Hangzhou, CN
- Hui Zhang
Zhejiang University – Hangzhou, CN
- Zhou Zheng
Zhejiang University – Hangzhou, CN







Report from Dagstuhl Seminar 21421

Quantum Cryptanalysis

Edited by

Stacey Jeffery[1], Michele Mosca[2], Maria Naya-Plasencia[3], and Rainer Steinwandt[4]

1 **CWI – Amsterdam, NL,** smjeffery@gmail.com
2 **University of Waterloo, CA,** michele.mosca@uwaterloo.ca
3 **INRIA – Paris, FR,** maria.naya_plasencia@inria.fr
4 **University of Alabama in Huntsville, US,** rs0141@uah.edu

Abstract

This seminar report documents the program and the outcomes of Dagstuhl Seminar 21421 *Quantum Cryptanalysis*. The seminar took place in a hybrid format in Fall 2021. The report starts out with the motivation and comments on the organization of this instance of the Dagstuhl Seminar series on Quantum Cryptanalysis, followed by abstracts of presentations. The presentation abstracts were provided by seminar participants.

Seminar October 17–22, 2021 – http://www.dagstuhl.de/21421
2012 ACM Subject Classification Hardware → Quantum technologies; Security and privacy → Cryptanalysis and other attacks; Theory of computation → Computational complexity and cryptography
Keywords and phrases computational algebra, post-quantum cryptography, quantum computing, quantum resource estimation
Digital Object Identifier 10.4230/DagRep.11.9.64
Edited in cooperation with André Schrottenloher

1 Executive Summary

Stacey Jeffery (CWI – Amsterdam, NL)
Michele Mosca (University of Waterloo, CA)
Maria Naya-Plasencia (INRIA – Paris, FR)
Rainer Steinwandt (University of Alabama in Huntsville, US)

Motivation and scope

Owing to the ongoing pandemic, this (sixth) installment of the Dagstuhl Seminar series on *Quantum Cryptanalysis* was held in a hybrid format. The focus of this seminar was on deployed schemes and more mature post-quantum cryptographic schemes, such as Round 3 candidates in NIST's standardization effort. For the technical program of the seminar, we encouraged research on

Quantum algorithmic innovations to attack cryptographic building blocks, leveraging state-of-the-art quantum computing. How can we leverage quantum algorithms to improve cryptanalytic capabilities, and how can we optimize the best available cryptanalytic results in meaningful quantum attack models?

Techniques and software tools to optimize and quantify resources for such attacks. Can we establish reasonably precise quantum resource counts for cryptanalytic attacks, especially for problem instances and parameter choices that are actually deployed or considered for standardization for future deployment?

Quantum attacks against today's RSA or elliptic-curve based cryptography and against modern block ciphers, which help us understand the urgency for transitioning to post-quantum solutions, fall in the seminar scope. As in the past, the seminar brought together researchers who work in the field of quantum computing with experts in classical cryptography, taking into account the latest advances in both fields. With 26 participants on site and 29 remote participants, Schloss Dagstuhl hosted a broad group of leading experts from across the globe.

Organization

The ongoing pandemic impacted the organization of the seminar, which for the first time was offered in a hybrid format. Thanks to the available technology at Schloss Dagstuhl and the efficient support of two volunteers (Shaun Kepley and Galina Pass), integrating remote presentations into the schedule worked smoothly.

The scheduling accounted for time zone differences and, as in the past, we left ample time for discussions and collaboration – for a typical day, we scheduled no more than four presentations. Following the Dagstuhl tradition and in line with prior seminars in the Quantum Cryptanalysis series, there was no technical program during Wednesday afternoon, leaving participants time for exploring the surroundings, spending time on research, or taking care of testing requirements for upcoming international travel.

Results and next steps

The collaboration between cryptographers and experts in quantum computing has come a long way, and it seems fair to say that this Dagstuhl Seminar series has contributed to this positive development. The quantum cryptanalytic progress in symmetric cryptography is very noticeable. This was evidenced by the number and quality of presentations on this subject offered by seminar participants. On the asymmetric side, the presentations demonstrated fascinating research progress on understanding computational problems related to lattices and codes. At the same time, a need remains to better quantify the potential of quantum algorithms for tackling hardness assumptions as used in state-of-the-art post-quantum proposals.

2 Table of Contents

3 Overview of Talks

3.1 Enumeration-based Lattice Reduction

Shi Bai (Florida Atlantic University – Boca Raton, US)

Joint work of Martin R. Albrecht, Shi Bai, Pierre-Alain Fouque, Paul Kirchner, Damien Stehlé, Weiqiang Wen
Main reference Martin R. Albrecht, Shi Bai, Pierre-Alain Fouque, Paul Kirchner, Damien Stehlé, Weiqiang Wen: "Faster Enumeration-Based Lattice Reduction: Root Hermite Factor $k^{1/(2k)}$ Time $k^{k/8+o(k)}$", in Proc. of the Advances in Cryptology – CRYPTO 2020 – 40th Annual International Cryptology Conference, CRYPTO 2020, Santa Barbara, CA, USA, August 17-21, 2020, Proceedings, Part II, Lecture Notes in Computer Science, Vol. 12171, pp. 186–212, Springer, 2020.
URL https://doi.org/10.1007/978-3-030-56880-1_7

Lattice reduction algorithms have received much attention in recent years due to their relevance to lattice-based cryptography. In this talk, we will discuss some of the recent developments on enumeration-based lattice reduction algorithms.

First, we will discuss a lattice reduction algorithm that achieves root Hermite factor $k^{(1/(2k))}$ in time $k^{(k/8+o(k))}$ and polynomial memory. This improves the previously best known enumeration-based lattice-reduction algorithms which achieve the same quality, but in time $k^{(k/(2e)+o(k))}$. The main idea is to conduct the preprocessing in a larger dimension than the enumerate dimension. Second, we discuss the usage of approximate enumeration oracles in BKZ, together with extended preprocessing ideas. In the end, we will illustrate some simulated results to demonstrate their practical behavior.

References

1 Martin R. Albrecht, Shi Bai, Pierre-Alain Fouque, Paul Kirchner, Damien Stehlé, Weiqiang Wen, Faster Enumeration-Based Lattice Reduction: Root Hermite Factor $k^{(1/(2k))}$ Time $k^{(k/8+o(k))}$. CRYPTO (2) 2020: 186-212

2 Martin R. Albrecht, Shi Bai, Jianwei Li, Joe Rowell, Lattice Reduction with Approximate Enumeration Oracles – Practical Algorithms and Concrete Performance. CRYPTO (2) 2021: 732-759

3.2 Quantum hardness of the code equivalence problem

Jean-François Biasse (University of South Florida – Tampa, US)

Joint work of Alessandro Barenghi, Jean-François Biasse, Edoardo Persichetti, Paolo Santini
Main reference Alessandro Barenghi, Jean-François Biasse, Edoardo Persichetti, Paolo Santini: "LESS-FM: Fine-Tuning Signatures from the Code Equivalence Problem", in Proc. of the Post-Quantum Cryptography – 12th International Workshop, PQCrypto 2021, Daejeon, South Korea, July 20-22, 2021, Proceedings, Lecture Notes in Computer Science, Vol. 12841, pp. 23–43, Springer, 2021.
URL https://doi.org/10.1007/978-3-030-81293-5_2

In this talk we introduce quantum algorithms for solving the Code Equivalence problem. Code Equivalence can serve as the hardness assumption of certain code-based signature schemes. This problem has been studied for a long time, but not in the context of (post-quantum) cryptography. In this presentation, we showed how to use quantum computers to speed up the classical algorithm due to Leon (and its subsequent improvements, in particular a recent work from Beulens), and the Support Splitting Algorithm (SSA) due to Sendrier.

References

1 Jean-François Biasse, Giacomo Micheli, Edoardo Persichetti, Paolo Santini, LESS is More: Code-Based Signatures Without Syndromes. AFRICACRYPT 2020: 45-65

2 Alessandro Barenghi, Jean-François Biasse, Edoardo Persichetti, Paolo Santini, LESS-FM: Fine-Tuning Signatures from the Code Equivalence Problem. PQCrypto 2021: 23-43

3.3 Quantum Linearization Attacks

Xavier Bonnetain (University of Waterloo, CA)

Joint work of Xavier Bonnetain, Gaëtan Leurent, María Naya-Plasencia, André Schrottenloher
Main reference Xavier Bonnetain, Gaëtan Leurent, María Naya-Plasencia, André Schrottenloher: "Quantum Linearization Attacks", in Proc. of the Advances in Cryptology – ASIACRYPT 2021 – 27th International Conference on the Theory and Application of Cryptology and Information Security, Singapore, December 6-10, 2021, Proceedings, Part I, Lecture Notes in Computer Science, Vol. 13090, pp. 422–452, Springer, 2021.
URL https://doi.org/10.1007/978-3-030-92062-3_15

Recent works have shown that quantum period-finding can be used to break many popular constructions (some block ciphers such as Even-Mansour, multiple MACs and AEs...) in the superposition query model. So far, all the constructions broken exhibited a strong algebraic structure, which enables to craft a periodic function of a single input block. Recovering the secret period allows to recover a key, distinguish, break the confidentiality or authenticity of these modes.

In this paper, we introduce the *quantum linearization attack*, a new way of using Simon's algorithm to target MACs in the superposition query model. Specifically, we use inputs of multiple blocks as an interface to a function hiding a linear structure. Recovering this structure allows to perform forgeries.

We also present some variants of this attack that use other quantum algorithms, which are much less common in quantum symmetric cryptanalysis: Deutsch's, Bernstein-Vazirani's, and Shor's. To the best of our knowledge, this is the first time these algorithms have been used in quantum forgery or key-recovery attacks.

Our attack breaks many parallelizable MACs such as LightMac, PMAC, and numerous variants with (classical) beyond-birthday-bound security (LightMAC+, PMAC) or using tweakable block ciphers (ZMAC). More generally, it shows that constructing parallelizable quantum-secure PRFs might be a challenging task.

3.4 Quantum Period Finding against Symmetric Primitives in Practice

Xavier Bonnetain (University of Waterloo, CA)

Joint work of Xavier Bonnetain, Samuel Jaques
Main reference Xavier Bonnetain, Samuel Jaques: "Quantum Period Finding against Symmetric Primitives in Practice", IACR Trans. Cryptogr. Hardw. Embed. Syst., Vol. 2022(1), pp. 1–27, 2022.
URL https://doi.org/10.46586/tches.v2022.i1.1-27

We present the first complete description of a quantum circuit for the offline Simon's algorithm, and estimate its cost to attack the MAC Chaskey, the block cipher PRINCE and the NIST lightweight candidate AEAD scheme Elephant. These attacks require a reasonable amount

of qubits, comparable to the number of qubits required to break RSA-2048. They are faster than other collision algorithms, and the attacks against PRINCE and Chaskey are the most efficient known to date. As Elephant has a key smaller than its state size, the algorithm is less efficient and ends up more expensive than exhaustive search.

We also propose an optimized quantum circuit for boolean linear algebra as well as complete reversible implementations of PRINCE, Chaskey, spongent and Keccak which are of independent interest for quantum cryptanalysis. We stress that our attacks could be applied in the future against today's communications, and recommend caution when choosing symmetric constructions for cases where long-term security is expected.

3.5 Lattice sieving via quantum random walks

André Chailloux (INRIA – Paris, FR)

Joint work of André Chailloux, Johanna Loyer
Main reference André Chailloux, Johanna Loyer: "Lattice Sieving via Quantum Random Walks", in Proc. of the Advances in Cryptology – ASIACRYPT 2021 – 27th International Conference on the Theory and Application of Cryptology and Information Security, Singapore, December 6-10, 2021, Proceedings, Part IV, Lecture Notes in Computer Science, Vol. 13093, pp. 63–91, Springer, 2021.
URL https://doi.org/10.1007/978-3-030-92068-5_3

Lattice-based cryptography is one of the leading proposals for post-quantum cryptography. The Shortest Vector Problem (SVP) is arguably the most important problem for the cryptanalysis of lattice-based cryptography, and many lattice-based schemes have security claims based on its hardness. The best quantum algorithm for the SVP is due to Laarhoven [1] and runs in (heuristic) time $2^{0.2653d+o(d)}$ where d is the dimension of the lattice. In this article, we present an improvement over Laarhoven's result and present an algorithm that has a (heuristic) running time of $2^{0.2570d+o(d)}$ We also present time-memory trade-offs where we quantify the amount of quantum memory and quantum random access memory of our algorithm. The core idea is to replace Grover's algorithm used in [1] in a key part of the sieving algorithm by a quantum random walk in which we add a layer of local sensitive filtering.

References

1 Thijs Laarhoven. *Search problems in cryptography, From fingerprinting to lattice sieving*. PhD thesis, Eindhoven University of Technology, 2016

3.6 Quantum Reduction of Finding Short Code Vectors to the Decoding Problem

Thomas Debris-Alazard (Ecole Polytechnique – Palaiseau, FR)

Joint work of Thomas Debris-Alazard, Maxime Remaud, Jean-Pierre Tillich
Main reference Thomas Debris-Alazard, Maxime Remaud, Jean-Pierre Tillich: "Quantum Reduction of Finding Short Code Vectors to the Decoding Problem", CoRR, Vol. abs/2106.02747, 2021.
URL https://arxiv.org/abs/2106.02747

We give a quantum reduction from finding short codewords in a random linear code to decoding for the Hamming metric. This is the first time such a reduction (classical or quantum) has been obtained. Our reduction adapts to linear codes Stehlé-Steinfield-Tanaka-Xagawa' re-interpretation of Regev's quantum reduction from finding short lattice vectors to

solving the Closest Vector Problem. The Hamming metric is a much coarser metric than the Euclidean metric and this adaptation has needed several new ingredients to make it work. For instance, in order to have a meaningful reduction it is necessary in the Hamming metric to choose a very large decoding radius and this needs in many cases to go beyond the radius where decoding is unique. Another crucial step for the analysis of the reduction is the choice of the errors that are being fed to the decoding algorithm. For lattices, errors are usually sampled according to a Gaussian distribution. However, it turns out that the Bernoulli distribution (the analogue for codes of the Gaussian) is too much spread out and can not be used for the reduction with codes. Instead we choose here the uniform distribution over errors of a fixed weight and bring in orthogonal polynomials tools to perform the analysis and an additional amplitude amplification step to obtain the aforementioned result.

3.7 Cryptanalysis of HFEv-

Jintai Ding (Tsinghua University – Beijing, CN)

Joint work of Chengdong Tao, Albrecht Petzoldt, Jintai Ding
Main reference Chengdong Tao, Albrecht Petzoldt, Jintai Ding: "Efficient Key Recovery for All HFE Signature Variants", in Proc. of the Advances in Cryptology – CRYPTO 2021 – 41st Annual International Cryptology Conference, CRYPTO 2021, Virtual Event, August 16-20, 2021, Proceedings, Part I, Lecture Notes in Computer Science, Vol. 12825, pp. 70–93, Springer, 2021.
URL https://doi.org/10.1007/978-3-030-84242-0_4

The HFE cryptosystem is one of the best known multivariate schemes. Especially in the area of digital signatures, the HFEv- variant offers short signatures and high performance. Recently, an instance of the HFEv- signature scheme called GeMSS was elected as one of the alternative candidates for signature schemes in the third round of the NIST Post Quantum Crypto (PQC) Standardization Project. In this paper, we propose a new key recovery attack on the HFEv- signature scheme. Our attack shows that both the Minus and the Vinegar modification do not enhance the security of the basic HFE scheme significantly. This shows that it is very difficult to build a secure and efficient signature scheme on the basis of HFE. In particular, we use our attack to show that the proposed parameters of the GeMSS scheme are not as secure as claimed.

3.8 On completely factoring any integer efficiently in a single run of an order-finding algorithm

Martin Ekerå (KTH Royal Institute of Technology – Stockholm, & Swedish NCSA, SE)

Main reference Martin Ekerå: "On completely factoring any integer efficiently in a single run of an order-finding algorithm", Quantum Inf. Process., Vol. 20(6), p. 205, 2021.
URL https://doi.org/10.1007/s11128-021-03069-1

In this talk, we present the recent paper [1]: Specifically, we show, for any integer N, that given the order of a single element selected uniformly at random from $\mathbb{Z}_N^*$, we can completely factor N efficiently classically with very high probability. The implication of this result, in the context of Shor's factoring algorithm, is that a single run of the quantum order-finding part is usually sufficient. All factors may then be recovered in a classical post-processing step.

The classical algorithm needed for this step is essentially a slightly modified randomized version of an algorithm due to Miller. For further details, interested readers are referred to the abstract and full text of [1].

References

1 Martin Ekerå, On completely factoring any integer efficiently in a single run of an order-finding algorithm. Quantum Inf. Process. 20(6): 205 (2021)

3.9 Limitations of the Macaulay matrix approach for using the HHL algorithm to solve multivariate polynomial systems

András Gilyén (Alfréd Rényi Institute of Mathematics – Budapest, HU)

Joint work of Jintai Ding, Vlad Gheorghiu, Sean Hallgren, Jianqiang Li
Main reference Jintai Ding, Vlad Gheorghiu, András Gilyén, Sean Hallgren, Jianqiang Li: "Limitations of the Macaulay matrix approach for using the HHL algorithm to solve multivariate polynomial systems", CoRR, Vol. abs/2111.00405, 2021.
URL https://arxiv.org/abs/2111.00405

Recently Chen and Gao (2017) proposed a new quantum algorithm for Boolean polynomial system solving, motivated by the cryptanalysis of some post-quantum cryptosystems. The key idea of their approach is to apply a Quantum Linear System (QLS) algorithm to a Macaulay linear system over $\mathbb{C}$, which is derived from the Boolean polynomial system. The efficiency of their algorithm depends on the condition number of the Macaulay matrix. In this paper, we give a strong lower bound on the condition number as a function of the Hamming weight of the Boolean solution, and show that in many (if not all) cases a Grover-based exhaustive search algorithm outperforms their algorithm. Then, we improve upon Chen and Gao's algorithm by introducing the Boolean Macaulay linear system over $\mathbb{C}$ by reducing the original Macaulay linear system. This improved algorithm could potentially significantly outperform the brute-force algorithm, when the Hamming weight of the solution is logarithmic in the number of Boolean variables. Furthermore, we provide a simple and more elementary proof of correctness for our improved algorithm using a reduction employing the Valiant-Vazirani affine hashing method, and also extend the result to polynomial systems over $\mathbb{F}_q$ improving on subsequent work by Chen, Gao and Yuan (2018). We also suggest a new approach for extracting the solution of the Boolean polynomial system via a generalization of the quantum coupon collector problem of Arunachalam, Belovs, Childs, Kothari, Rosmanis, and de Wolf (2020).

3.10 Quantum Collision Attacks on Reduced SHA-256 and SHA-512

Akinori Hosoyamada (NTT – Tokyo, JP)

Joint work of Akinori Hosoyamada, Yu Sasaki
Main reference Akinori Hosoyamada, Yu Sasaki: "Quantum Collision Attacks on Reduced SHA-256 and SHA-512", in Proc. of the Advances in Cryptology – CRYPTO 2021 – 41st Annual International Cryptology Conference, CRYPTO 2021, Virtual Event, August 16-20, 2021, Proceedings, Part I, Lecture Notes in Computer Science, Vol. 12825, pp. 616–646, Springer, 2021.
URL https://doi.org/10.1007/978-3-030-84242-0_22

In this talk, we present dedicated quantum collision attacks on SHA-256 and SHA-512. The attacks reach 38 and 39 steps, respectively, which significantly improve the classical attacks for 31 and 27 steps. Both attacks adopt the framework of the previous work that converts many semi-free-start collisions into a 2-block collision, and are faster than the generic attack in the cost metric of time-space tradeoff. We observe that the number of required semi-free-start collisions can be reduced in the quantum setting, which allows us to convert the previous classical 38 and 39 step semi-free-start collisions into a collision. The idea behind our attacks is simple and will also be applicable to other cryptographic hash functions. This talk is based on our paper of the same title presented at CRYPTO 2021.

3.11 Automatizing applications of Simon's algorithm to symmetric crypto

Nils Gregor Leander (Ruhr-Universität Bochum, DE)

Joint work of Federico Canale, Nils Gregor Leander, Lukas Stennes

We simplify the search for new applications of Simon's algorithm and thereby overcome the increasing complexity of the attacks in the literature.

More precisely, we present a generic algorithm that aims at finding, given a symmetric cryptographic scheme E, non-trivial periodic functions f, that can then be efficiently computed by a quantum computer.

Our approach here is to represent those functions f dependent on E by a class of circuits. Those circuits can make use of oracle gates for E and potentially further oracle gates for internal parts of the scheme E. We then automatically examine all circuits up to a certain number of gates and test each of them for periodicity, by instantiating the respective function on small dimensions. Of course, this means that many useless circuits, as well as many useless periods, are generated. The main technical contribution and work is aimed at addressing this problem, and keeping the process efficient.

Using our approach, we automatically find the first efficient key-recovery attacks against constructions like 5-round MISTY L-FK or 5-round Feistel-FK (with internal permutation) using Simon's algorithm.

3.12 Test of Quantumness with Small-Depth Quantum Circuits

François Le Gall (Nagoya University, JP)

Joint work of Shuichi Hirahara, François Le Gall
Main reference Shuichi Hirahara, François Le Gall: "Test of Quantumness with Small-Depth Quantum Circuits", in Proc. of the 46th International Symposium on Mathematical Foundations of Computer Science, MFCS 2021, August 23-27, 2021, Tallinn, Estonia, LIPIcs, Vol. 202, pp. 59:1–59:15, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021.
URL https://doi.org/10.4230/LIPIcs.MFCS.2021.59

Recently Brakerski, Christiano, Mahadev, Vazirani and Vidick (FOCS 2018) have shown how to construct a test of quantumness based on the learning with errors (LWE) assumption: a test that can be solved efficiently by a quantum computer but cannot be solved by a classical polynomial-time computer under the LWE assumption. This test has lead to several cryptographic applications. In particular, it has been applied to producing certifiable randomness from a single untrusted quantum device, self-testing a single quantum device and device-independent quantum key distribution.

In this paper, we show that this test of quantumness, and essentially all the above applications, can actually be implemented by a very weak class of quantum circuits: constant-depth quantum circuits combined with logarithmic-depth classical computation. This reveals novel complexity-theoretic properties of this fundamental test of quantumness and gives new concrete evidence of the superiority of small-depth quantum circuits over classical computation.

3.13 Quantum Time-Space Tradeoff for Finding Multiple Collision Pairs

Frédéric Magniez (CNRS – Paris, FR)

Joint work of Yassine Hamoudi, Frédéric Magniez
Main reference Yassine Hamoudi, Frédéric Magniez: "Quantum Time-Space Tradeoff for Finding Multiple Collision Pairs", in Proc. of the 16th Conference on the Theory of Quantum Computation, Communication and Cryptography, TQC 2021, July 5-8, 2021, Virtual Conference, LIPIcs, Vol. 197, pp. 1:1–1:21, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021.
URL https://doi.org/10.4230/LIPIcs.TQC.2021.1

We study the problem of finding K collision pairs in a random function $f : [N] \to [N]$ by using a quantum computer. We prove that the number of queries to the function in the quantum random oracle model must increase significantly when the size of the available memory is limited.

Classically, the same question has only been settled recently by Dinur [Eurocrypt'20], who showed that the Parallel Collision Search algorithm of van Oorschot and Wiener achieves the optimal time-space tradeoff.

Our result limits the extent to which quantum computing may decrease this tradeoff. Our method is based on a novel application of Zhandry's recording query technique [Crypto'19] for proving lower bounds in the exponentially small success probability regime.

As a second application, we give a simpler proof of the time-space tradeoff for sorting N numbers on a quantum computer, which was first obtained by Klauck, Špalek and de Wolf.

3.14 Boosting the hybrid attack on NTRU

Phong Q. Nguyen (INRIA & ENS Paris, FR)

Main reference Nguyen, Phong Q.: "Boosting the Hybrid Attack on NTRU: Torus LSH, Permuted HNF and Boxed Sphere." Third PQC standardization conference, 2021.
URL https://csrc.nist.gov/CSRC/media/Events/third-pqc-standardization-conference/documents/accepted-papers/nguyen-boosting-hybridboost-pqc2021.pdf

We revisit collision attacks on NTRU, namely Odlyzko's meet-in-the-middle attack and Howgrave-Graham's hybrid attack. We show how to simplify and improve these attacks with respect to effi- ciency, analysis and ease of implementation. Our main ingredients are randomization and geometry: we randomize the attacks by introducing torus variants of locality sensitive hashing (LSH) and new HNF-like bases of the NTRU lattice; and we establish a connection between the success probability of the hybrid attack and the intersection of an n-dimensional sphere with a random box. We provide mathematical and algorithmic bounds for such intersections, which is of independent interest. Our new analyses remain partially heuristic, but are arguably much more sound than previous analyses, and are backed by experiments. Our results show that the security estimates of the NTRU finalist in NIST's post-quantum standardization need to be revised.

3.15 Scalable Methods and Tools for (Very Large) Quantum Circuits

Alexandru Paler (Aalto University, FI)

Quantum circuits compilation and optimisation are an important building block of the quantum computing software stacks. Quantum hardware is a very scarce resource, and the successful execution of the first practical quantum computations, error-corrected or not, depends on squeezing logical circuits on the available quantum hardware. Moreover, in the context of fault-tolerant quantum computations, the execution of the circuits requires online, real-time compilation and optimisation methods. Consequently, scalability is more than a desirable characteristic of the methods and tools forming the software stacks. This talk follows a top-down description of the steps involved in the preparation of fault-tolerant quantum circuit executions. We present state-of-the-art tools and analyse their scalability with respect to very large instances of practical quantum circuits. We also discuss methods for verifying the correctness of resulting optimised circuits, as well as preliminary results on applying machine learning techniques for circuit optimisation.

3.16 Fast factoring integers by SVP algorithms

Claus Peter Schnorr (Goethe-Universität – Frankfurt am Main, DE)

Main reference Claus-Peter Schnorr:"Fast Factoring Integers by SVP Algorithms, corrected. IACR Cryptol. ePrint Arch. 2021: 933 (2021)
URL https://eprint.iacr.org/2021/933

To factor an integer N we construct n triples of p_n-smooth integers $u, v, |u - vN|$ for the n-th prime p_n. Denote such triple a fac-relation. We get fac-relations from a nearly shortest vector of the lattice $\mathcal{L}(\mathbf{R}_{n,f})$ with basis matrix $\mathbf{R}_{n,f} \in \mathbb{R}^{(n+1)\times(n+1)}$ where $f\colon [1,n] \to [1,n]$ is a permutation of $[1,2,\ldots,n]$ and $(f(1),\ldots,f(n),N' \ln N)$ is the diagonal and $(N' \ln p_1,\ldots,N' \ln p_n, N' \ln N)$ for $N' = N^{\frac{1}{n+1}}$ is the last line of $\mathbf{R}_{n,f}$. An independent permutation f' yields an independent fac-relation. We find sufficiently short lattice vectors by strong primal-dual reduction of $\mathbf{R}_{n,f}$. We factor $N \approx 2^{400}$ by $n = 47$ and $N \approx 2^{800}$ by $n = 95$. Our accelerated strong primal-dual reduction of [1] factors integers $N \approx 2^{400}$ and $N \approx 2^{800}$ by $4.2 \cdot 10^9$ and $8.4 \cdot 10^{10}$ arithmetic operations, much faster then the quadratic sieve and the number field sieve and using much smaller primes p_n. This destroys the RSA cryptosystem.

References

1 N. Gama and P.Q. Nguyen, Finding Short Lattice Vectors within Mordell's Inequality. Proc. of the 2008 ACM Symposiumon on Theory of Computing, pp. 208-216, 2008

3.17 Beyond quadratic speedups in quantum attacks on symmetric schemes

André Schrottenloher (CWI – Amsterdam, NL)

Joint work of Xavier Bonnetain, André Schrottenloher, Ferdinand Sibleyras
Main reference Xavier Bonnetain, André Schrottenloher, Ferdinand Sibleyras: "Beyond quadratic speedups in quantum attacks on symmetric schemes", CoRR, Vol. abs/2110.02836, 2021.
URL https://arxiv.org/abs/2110.02836

In symmetric cryptography, doubling the sizes of keys is often assumed to be a sufficient protection against quantum adversaries. This is because Grover's quantum search algorithm, which can be used for generically recovering the key, is limited to a quadratic speedup.

In this talk, we will study this key-recovery problem for block cipher constructions in the ideal model, such as the Even-Mansour or FX ciphers. In the superposition setting (a strong model of quantum attackers), some of these constructions are completely broken, even though they have classical security proofs. But attacks using only classical queries, which are deemed more realistic, have remained much more limited to date. As they were reaching quadratic speedups at most, they confirmed so far the intuition that security levels should just be doubled in general.

We will show that this is not always the case, by presenting a symmetric block cipher design with: 1. a security bound of $2^{2.5n}$ against classical adversaries, and: 2. a quantum attack in time roughly 2^n, that uses classical queries only. This gives, for the first time, a proven 2.5 speedup on a quantum attack in the classical query model.

3.18 Provable quantum algorithms for SVP

Yixin Shen (Royal Holloway University of London, GB)

Joint work of Divesh Aggarwal, Yanlin Chen, Rajendra Kumar, Yixin Shen
Main reference Divesh Aggarwal, Yanlin Chen, Rajendra Kumar, Yixin Shen: "Improved (Provable) Algorithms for the Shortest Vector Problem via Bounded Distance Decoding", CoRR, Vol. abs/2002.07955, 2020.
URL https://arxiv.org/abs/2002.07955

The most important computational problem on lattices is the Shortest Vector Problem (SVP). We present new algorithms that improve the state-of-the-art for provable quantum algorithms for SVP, ie, we present a quantum algorithm that runs in time $2^{0.953n+o(n)}$ and requires $2^{0.5n+o(n)}$ classical memory and poly(n) qubits. In the Quantum Random Access Memory (QRAM) model our algorithm takes only $2^{0.873n+o(n)}$ time and requires a QRAM of size $2^{0.1604n+o(n)}$, poly(n) qubits and $2^{0.5n}$ classical space. This improves over the previously fastest classical (which is also the fastest quantum) algorithm due to [1] that has a time and space complexity $2^{n+o(n)}$. The time complexity of our quantum algorithms are obtained using a known upper bound on a quantity related to the lattice kissing number which is $2^{0.402n}$. We conjecture that for most lattices this quantity is a $2^{o(n)}$. Assuming that this is the case, our quantum algorithm runs in time $2^{0.750n+o(n)}$ and our quantum algorithm in the QRAM model runs in time $2^{0.667n+o(n)}$.

References

1 Divesh Aggarwal, Daniel Dadush, Oded Regev, and Noah Stephens-Davidowitz. Solving the shortest vector problem in 2^n time using discrete gaussian sampling: Extended abstract. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 733–742, 2015.

3.19 NIST status update on the 3rd round

Daniel C. Smith-Tone (NIST – Gaithersburg, US)

NIST provided historical information, current status updates and future timeline information on the 3rd Round PQC standardization process.

Participants

- Marco Baldi
Polytechnic University of Marche, IT
- Jean-François Biasse
University of South Florida – Tampa, US
- Xavier Bonnetain
University of Waterloo, CA
- André Chailloux
INRIA – Paris, FR
- Thomas Debris-Alazard
Ecole Polytechnique – Palaiseau, FR
- Yfke Dulek
CWI – Amsterdam, NL
- Martin Ekerå
KTH Royal Institute of Technology – Stockholm, & Swedish NCSA, SE
- Stacey Jeffery
CWI – Amsterdam, NL
- Antoine Joux
CISPA – Saarbrücken, DE
- Stavros Kousidis
BSI – Bonn, DE
- Nils Gregor Leander
Ruhr-Universität Bochum, DE
- Frédéric Magniez
CNRS – Paris, FR
- Maria Naya-Plasencia
INRIA – Paris, FR
- Phong Q. Nguyen
INRIA & ENS Paris, FR
- Alexandru Paler
Aalto University, FI
- Galina Pass
CWI – Amsterdam, NL
- Edoardo Persichetti
Florida Atlantic University – Boca Raton, US
- Stephanie Reinhardt
BSI – Bonn, DE
- Paolo Santini
Polytechnic University of Marche, IT
- Claus Peter Schnorr
Goethe-Universität – Frankfurt am Main, DE
- André Schrottenloher
CWI – Amsterdam, NL
- Nicolas Sendrier
INRIA – Paris, FR
- Yixin Shen
Royal Holloway University of London, GB
- Jana Sotáková
University of Amsterdam, NL
- Rainer Steinwandt
University of Alabama in Huntsville, US
- Jean-Pierre Tillich
INRIA – Paris, FR

Remote Participants

- Andris Ambainis
University of Latvia – Riga, LV
- Shi Bai
Florida Atlantic University – Boca Raton, US
- Aleksandrs Belovs
University of Latvia – Riga, LV
- Daniel J. Bernstein
University of Illinois – Chicago, US
- Jintai Ding
Tsinghua University – Beijing, CN
- Philippe Gaborit
University of Limoges, FR
- András Gilyén
Alfréd Rényi Institute of Mathematics – Budapest, HU
- Maria Isabel González Vasco
King Juan Carlos University – Madrid, ES
- Akinori Hosoyamada
NTT – Tokyo, JP
- Tetsu Iwata
Nagoya University, JP
- Samuel E. Jaques
University of Oxford, GB
- Floyd Johnson
Florida Atlantic University – Boca Raton, US
- Elena Kirshanova
Immanuel Kant Baltic Federal Univ.- Kaliningrad, RU
- Péter Kutas
University of Birmingham, GB
- Tanja Lange
TU Eindhoven, NL
- François Le Gall
Nagoya University, JP
- Dustin Moody
NIST – Gaithersburg, US
- Michele Mosca
University of Waterloo, CA
- Ludovic Perret
Sorbonne University – Paris, FR
- Rachel Player
Royal Holloway University of London, GB
- Thomas Pöppelmann
Infineon Technologies AG – Neubiberg, DE
- Angela Robinson
NIST – Gaithersburg, US
- Yu Sasaki
NTT – Tokyo, JP
- John M. Schanck
Portland, US
- Daniel C. Smith-Tone
NIST – Gaithersburg, US
- Fang Song
Portland State University, US
- Adriana Suárez Corona
University of León, ES
- Dániel Szabó
University Paris Diderot, FR
- Bo-Yin Yang
Academia Sinica – Taipei, TW

Report from Dagstuhl Seminar 21431

Rigorous Methods for Smart Contracts

Edited by

Nikolaj S. Bjørner[1], Maria Christakis[2], Matteo Maffei[3], and Grigore Rosu[4]

1 **Microsoft – Redmond, US,** nbjorner@microsoft.com
2 **MPI-SWS – Kaiserslautern, DE,** maria@mpi-sws.org
3 **TU Wien, AT,** matteo.maffei@tuwien.ac.at
4 **University of Illinois – Urbana-Champaign, US,** grosu@illinois.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 21431 "Rigorous Methods for Smart Contracts". Blockchain technologies have emerged as an exciting field for both researchers and practitioners focusing on formal guarantees for software. It is arguably a "once in a lifetime" opportunity for rigorous methods to be integrated in audit processes for parties deploying smart contracts, whether for fund raising, securities trading, or supply-chain management.

Smart contracts are programs managing cryptocurrency accounts on a blockchain. Research in the area of smart contracts includes a fascinating combination of formal methods, programming-language semantics, and cryptography. First, there is vibrant development of *verification and program-analysis techniques* that check the correctness of smart-contract code. Second, there are emerging designs of *programming languages and methodologies* for writing smart contracts such that they are more robust by construction or more amenable to analysis and verification. Programming-language abstraction layers expose low-level cryptographic primitives enabling developers to design high-level *cryptographic protocols*. *Automated-reasoning mechanisms* present a common underlying enabler; and the specific needs of the smart-contract world offer new challenges.

This workshop brought together stakeholders in the aforementioned areas related to advancing reliable smart-contract technologies.

Seminar October 24–29, 2021 – http://www.dagstuhl.de/21431
2012 ACM Subject Classification Security and privacy → Logic and verification; Software and its engineering → Formal language definitions; Software and its engineering → Software verification and validation
Keywords and phrases automated reasoning, cryptographic protocols, program verification, programming languages, smart contracts
Digital Object Identifier 10.4230/DagRep.11.9.80
Edited in cooperation with Schindler, Tanja

1 Executive Summary

Nikolaj S. Bjørner (Microsoft – Redmond, US)

The seminar attracted 22 on-site and approximately as many off-site participants. The hybrid mode presented an opportunity for collaborators, particularly students, of invitees to participate remotely and contribute to the discussions. Remote participation spanned all

time zones which attested to their involvement. The on-site participants had the benefit of extended interactions and relation building so crucial for advancing scientific activities.

The technical program was organized around first day of tutorial presentations on the main topics covered by the seminar. These topics were *static analysis techniques*, *program verification methods*, *protocol design for decentralized ledgers*, and *semantic-based tools*.

The following days provided for in-depth sessions around these topics. Static analysis techniques spanned using Horn clause solvers, Datalog engines, and abstract interpretation frameworks in a mixture of academic and industrial settings. Program verification techniques, likewise, were pursued both by academic and industry participants. The seminar offered an excellent forum for the scientific and commercial community around smart contracts to exchange experiences and develop ideas.

For the social program, we hiked for two hours during a beautiful October afternoon to Landgasthof Paulus & Der Laden for a delightful dinner.

2 Table of Contents

3 Overview of Talks

3.1 The GASOL project: a GAS Optimization tooLkit

Elvira Albert (Complutense University of Madrid, ES) and Albert Rubio (Complutense University of Madrid, ES)

Joint work of Elvira Albert, Pablo Gordillo, Alejandro Hernández-Cerezo, Albert Rubio

Super-optimization is a compilation technique that searches for the optimal sequence of instructions semantically equivalent to a given (loop-free) initial sequence. This talks overviews our approach for super-optimization of smart contracts based on Max-SMT which is split into two main phases:(i) the extraction of a functional specification from the basic blocks of the smart contract, which is simplified using rules that capture the semantics of the arithmetic, bit-wise, relational operations, etc. and(ii) the synthesis of optimized blocks which, by means of an efficient Max-SMT encoding, finds the bytecode blocks with minimal cost (according to the selected optimization criteria) and whose functional specification is equal (modulo commutativity) to the extracted one. Our experiments on randomly selected real contracts achieve important gains in gas and in bytes-size over code already optimized by solc.

References

1 Elvira Albert, Pablo Gordillo, Albert Rubio, Maria Anna Schett: Synthesis of Super-Optimized Smart Contracts Using Max-SMT. CAV (1) 2020: 177-200

3.2 Formally Verifying Ethereum Smart Contracts by Overwhelming Horn Solvers

Leonardo Alt (Ethereum – Berlin, DE)

Ethereum smart contracts hold billions of USD, have (usually) immutable logic, and are (also usually) open source. Therefore, ensuring that the programs are bug free is essential for this ecosystem. Formal verification, particularly, has seen a successful application in smart contracts also due to their rather small complexity compare to other types of systems, since the contract size is limited and complex code often implies higher computation costs. In this work we present a model checker for Solidity smart contracts based on Constrained Horn Clauses [1]. The Solidity programs are encoded as systems of Horn clauses where verifying a safety clause consists of Horn satisfiability. We show how the encoding is performed following [2], and the special behaviors from smart contracts that lead to the specific Horn encoding for these problems from [1]. We also show experiments on a small scale, focusing on specific features, as well as large real-world instances, demonstrating how properties that are part of large systems can also be solved automatically. Finally, we present data comparing how the two backend Horn solvers used by the tool, Spacer and Eldarica, compare in the different instances that are part of the experiments.

References

1 Matteo Marescotti, Rodrigo Otoni, Leonardo Alt, Patrick Eugster, Antti E. J. Hyvärinen, Natasha Sharygina: Accurate Smart Contract Verification Through Direct Modelling. ISoLA (3) 2020: 178-194

2 Nikolaj Bjørner, Arie Gurfinkel, Kenneth L. McMillan, Andrey Rybalchenko: Horn Clause Solvers for Program Verification. Fields of Logic and Computation II 2015: 24-51

3.3 Smart contracts in Bitcoin and BitML

Massimo Bartoletti (University of Cagliari, IT)

Joint work of Massimo Bartoletti, Roberto Zunino
Main reference Massimo Bartoletti, Roberto Zunino: "BitML: A Calculus for Bitcoin Smart Contracts", in Proc. of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018, Toronto, ON, Canada, October 15-19, 2018, pp. 83–100, ACM, 2018.
URL https://doi.org/10.1145/3243734.3243795

Although Bitcoin is mainly used to exchange crypto-currency, its blockchain and consensus mechanism can also be exploited to execute smart contracts, allowing mutually untrusted parties to exchange crypto-assets according to pre-agreed rules. To this purpose, Bitcoin features a non Turing-complete script language, which is used to specify the redeem conditions of transactions. This is a simple language of expressions, without loops or recursion. To write complex smart contracts, one needs to suitably combine transactions: in this approach, executing a contract amounts to appending sequences of transactions in a given order.

A drawback of this approach is that the complexity of writing smart contracts grows quickly in the number of transactions needed to implement it. Reasoning about the correctness of these contracts is even harder: one would have to consider computational adversaries who interact with the blockchain, only being constrained to use PPTIME algorithms. To overcome these issues we have proposed BITML [1], a high-level DSL for smart contracts with a computationally sound compiler to Bitcoin transactions.

The computational soundness property allows us to reason about contracts at the symbolic level of the BitML semantics. We exploit this possibility to investigate a landmark property of contracts, called liquidity, which ensures that funds never remain frozen within a contract. Liquidity is a relevant issue, as witnessed by a recent attack to the Ethereum Parity Wallet, which has frozen 160M USD within the contract, making this sum unredeemable by any user.

We develop a static analysis for liquidity of BitML contracts. This is achieved by first devising a finite-state, safe abstraction of infinite-state semantics of BitML, and then model-checking this abstraction.

We conclude by discussing a few open issues: in particular, how to enhance the expressiveness of Bitcoin contracts via minor extensions of the Bitcoin script language, and how to reduce the cost of executing contracts.

References

1 Massimo Bartoletti, Roberto Zunino: BitML: A Calculus for Bitcoin Smart Contracts. CCS 2018: 83-100

3.4 On Supporting Smart Contract Verification in Z3

Nikolaj S. Bjørner (Microsoft – Redmond, US)

We give an overview of current activities and features in Z3 that are aimed to make reasoning about smart contracts more efficient. The use of SMT solvers for Smart Contract analysis spans a range of scenarios, noteworthy symbolic execution, extended static analysis, symbolic model checking through solving satisfiability of Horn clauses, to program verification style reasoning. Thus, there is a fair range of reasoning capabilities that are relevant for Smart Contracts. Based on current experiences from users of z3 the talk presents ongoing work and extensions that may be of use for advancing reasoning about smart contracts.

Native Large Bit-width reasoning Verification conditions seem from Certora include code paths from EVM that involve bit-vectors with 256 bits each. When translating bit-vector reasoning to integers, the large bit-widths result in formulas with numerals that require expensive representations of large numerals. Integer reasoning does not assume fixed width numerals and has to take into account that integers can be unbounded. Integer reasoning is furthermore limited when it comes to non-linear arithmetic, as generally even quantifier-free non-linear integer satisfiability is undecidable (Hilbert's 10'th problem). Z3's native bit-vector reasoning engine converts bit-vector reasoning to propositional SAT. The overhead of representing bit-vector multiplication and division for large bit-widths makes propositional bit-vector reasoning impractical. With Jakob Rath at TU Vienna we are developing an new word level bit-vector reasoning engine in Z3 called *PolySAT*. PolySAT builds on and extends ideas developed by [1] to handle constraints that involve polynomial inequalities over bit-vectors. The main innovations in PolySAT include a generalization of conflict detection to linear inequalities, not only linear inequalities with unit coefficients. Conflict detection is complemented by an on-demand *saturation* phase to generalize infeasible cores.

Refinement Sorts Uses of Z3 at Meta (Facebook) suggest the relevance of integrating refinement sorts to the input formalism of SMT solvers. For example, the sort of natural numbers is a refinement sort of integers that are non-negative. Each natural number is an integer that is also non-negative. We illustrate how refinement sorts can be supported as theory that lazily instantiates axioms required to enforce refinement constraints.

Code as Constraints A new capability in Z3 is (re)exposing a capability to encode on-demand propagators outside of the solver. This enables users to encode properties that may require a bloated axiomatization.

References

1 Stéphane Graham-Lengrand and Dejan Jovanovic and Bruno Dutertre, *Solving Bitvectors with MCSAT: Explanations from Bits and Pieces*, in Automated Reasoning – 10th International Joint Conference, IJCAR 2020, Paris, France, July 1-4, 2020, Proceedings, Part I

3.5 Resource-Aware Session Types for Digital Contracts

Ankush Das (Amazon – Cupertino, US)

Joint work of Ankush Das, Stephanie Balzer, Jan Hoffmann, Frank Pfenning, Ishani Santurkar
Main reference Ankush Das, Stephanie Balzer, Jan Hoffmann, Frank Pfenning, Ishani Santurkar: "Resource-Aware Session Types for Digital Contracts", in Proc. of the 34th IEEE Computer Security Foundations Symposium, CSF 2021, Dubrovnik, Croatia, June 21-25, 2021, pp. 1–16, IEEE, 2021.
URL https://doi.org/10.1109/CSF51468.2021.00004

Programming digital contracts comes with unique challenges, which include (i) expressing and enforcing protocols of interaction, (ii) controlling resource usage, and (iii) preventing the duplication or deletion of a contract's assets. This talk presents the type-theoretic foundation and implementation of Nomos, a programming language for digital contracts that addresses these challenges. To express and enforce protocols, Nomos is based on shared binary session types. To control resource usage, Nomos employs automatic amortized resource analysis. To prevent the duplication or deletion of assets, Nomos uses a linear type system. A monad integrates the effectful session-typed language with a general-purpose functional language. Nomos' prototype implementation features linear-time type checking and efficient type reconstruction that includes automatic inference of resource bounds via off-the-shelf linear optimization. The effectiveness of the language is evaluated with case studies about implementing common smart contracts such as auctions, elections, and currencies. Nomos is completely formalized, including the type system, a cost semantics, and a transactional semantics to instantiate Nomos contracts on a blockchain. The type soundness proof ensures that protocols are followed at run-time and that types establish sound upper bounds on the resource consumption, ruling out re-entrancy and out-of-gas vulnerabilities.

3.6 Rich Specifications for Ethereum Smart Contract Verification

Marco Eilers (ETH Zürich, CH)

Joint work of Christian Bräm, Marco Eilers, Peter Müller, Robin Sierra, Alexander J. Summers
Main reference Christian Bräm, Marco Eilers, Peter Müller, Robin Sierra, Alexander J. Summers: "Rich specifications for Ethereum smart contract verification", Proc. ACM Program. Lang., Vol. 5(OOPSLA), pp. 1–30, 2021.
URL https://doi.org/10.1145/3485523

The verification of smart contracts poses challenges that rarely arise in other domains due to their typical use case (manipulating and transferring resources) and the necessity to interact with adversarial outside code. In this talk, I present a novel specification methodology, tailored to the domain of smart contracts, which enables (1) sound and precise reasoning in the presence of unverified code and arbitrary re-entrancy, (2) modular reasoning about collaborating smart contracts, and (3) domain-specific specification of resources and resource transfers, expressing a contract's behavior in intuitive and concise ways and excluding typical errors by default. I also briefly show the implementation of our technique in 2vyper, an SMT-based automated verification tool for Ethereum smart contracts written in Vyper, demonstrating its effectiveness for verifying real-world contracts.

3.7 Verifying Lighting in Why3

Grzegorz Fabianski (University of Warsaw, PL)

Joint work of Grzegorz Fabianski, Rafał Stefański

Lighting Network is an off-chain payment protocol working over bitcoin (and arguably the biggest application of scripting capabalities of bitcoin). As such, it uses complicated logic on the client-side to circumvent limited capabalities of bitcoin scripting language. In this talk, I present the status of ongoing work about verifying the Lighting network in the Why3 system. I will describe techniques that will enable us to verify randomized protocol (like Lighting) using deterministic Hoare Logic. Then I'll explain an overview of the project architecture and used abstractions.

3.8 Consensus for Decentralized Ledgers

Bryan Ford (EPFL Lausanne, CH)

URL https://drive.google.com/file/d/1M0xEtOT7lNBQFE7DCzWhBDMFrSKXnOey/view?usp=sharing

Blockchain and distributed ledger technology, as popularized by Bitcoin, has reinvigorated the classic computer science topic of consensus algorithms and protocols, and sent research in this space in many new directions. This talk summarizes a few of these developments in consensus for decentralized ledger systems. While classic "permission" consensus mechanisms such as Paxos assume a fixed set of a few consensus nodes, cryptocurrencies like Bitcoin established new expectations: to be open to "permissionless" participation; to scale to thousands or millions of participants; and to ensure that Byzantine security increases as participation increases and diversifies. Bitcoin's "Nakamoto consensus" is slow and has many other costs and limitations, however. Research on improving blockchain consensus has introduced "hybrid" schemes such as Byzcoin that achieve the best properties of Bitcoin-style permissionless consensus and PBFT-style Byzantine consensus. Sharding schemes such as Omniledger allow a blockchain's processing capacity to increase via horizontal scalability as the number of participants grows, potentially without bound, without sacrificing Byzantine security or the ability to execute transactions atomically across shards. Random beacon protocols such as RandHound/RandHerd and drand are instrumental to enabling secure sharding and other advanced blockchain consensus schemes. New asynchronous consensus algorithms inspired by blockchain systems, while still not yet deployed in practice, promise greater resilience to potentially-adversarial network conditions such as denial-of-service attacks once they become truly practical. New structuring concepts such as threshold logical clocks (TLC) may help make asynchronous consensus both more practical and more understandable. Proof of Stake offers an alternative permissionless participation foundation to proof of work, offering much lower energy waste, but still suffers from potential (re-)centralization or "rich get richer" effects. Proof of Personhood schemes, such as pseudonym parties, offer a more egalitarian path towards inclusive permissionless participation, attempting to ensure "one person, one vote" or "one person, one unit of stake" in permissionless blockchain consensus.

References

1 Eleftherios Kokoris-Kogias, Philipp Jovanovic, Linus Gasser, Nicolas Gailly, Ewa Syta, Bryan Ford: OmniLedger: A Secure, Scale-Out, Decentralized Ledger via Sharding. IEEE Symposium on Security and Privacy 2018: 583-598.

2 Eleftherios Kokoris-Kogias, Philipp Jovanovic, Nicolas Gailly, Ismail Khoffi, Linus Gasser, Bryan Ford: Enhancing Bitcoin Security and Performance with Strong Consistency via Collective Signing. USENIX Security Symposium 2016: 279-296

3 Ewa Syta, Philipp Jovanovic, Eleftherios Kokoris-Kogias, Nicolas Gailly, Linus Gasser, Ismail Khoffi, Michael J. Fischer, Bryan Ford: Scalable Bias-Resistant Distributed Randomness. IEEE Symposium on Security and Privacy 2017: 444-460

4 Kirill Nikitin, Eleftherios Kokoris-Kogias, Philipp Jovanovic, Nicolas Gailly, Linus Gasser, Ismail Khoffi, Justin Cappos, Bryan Ford: CHAINIAC: Proactive Software-Update Transparency via Collectively Signed Skipchains and Verified Builds. USENIX Security Symposium 2017: 1271-1287

5 Maria Borge, Eleftherios Kokoris-Kogias, Philipp Jovanovic, Linus Gasser, Nicolas Gailly, Bryan Ford: Proof-of-Personhood: Redemocratizing Permissionless Cryptocurrencies. EuroS&P Workshops 2017: 23-26

6 Bryan Ford: Threshold Logical Clocks for Asynchronous Distributed Coordination and Consensus. CoRR abs/1907.07010 (2019)

7 Bryan Ford, Philipp Jovanovic, Ewa Syta: Que Sera Consensus: Simple Asynchronous Agreement with Private Coins and Threshold Logical Clocks. CoRR abs/2003.02291 (2020)

8 Bryan Ford: Identity and Personhood in Digital Democracy: Evaluating Inclusion, Equality, Security, and Privacy in Pseudonym Parties and Other Proofs of Personhood. CoRR abs/2011.02412 (2020)

3.9 Program analysis tools for software auditors

Diego Garbervetsky (University of Buenos Aires, AR)

Joint work of Diego Garbervetsky, Javier Godoy, Juan Pablo Galeotti, Sebastian Uchitel

In this talk I will summary part of the collaboration work with OpenZeppelin. In particular our quest for tools that can help auditors to be more productive. In this collaboration we first explore how expert perform software auditors, we then participated in audits trying to see how to formalize some of the properties found by auditors. We also performed a survey of existing tools and analyze which tools could fit OpenZeppelin's audit process. Finally we discuss a new approach to understand and validate smart contracts based on abstractions of behavioral models.

3.10 Modular verification of memory-manipulating programs

Isabel Garcia-Contreras (IMDEA Software – Madrid, ES)

Joint work of Isabel Garcia-Contreras, Arie Gurfinkel, Jorge A. Navas

In SMT-based model-checking (SMT-MC) the correctness of a program is determined through the satisfiability of logical verification conditions (VCs) expressing the program semantics. A popular approach to model memory is to encode memory accesses using array store or select terms. When generating modular (i.e., per function) VCs, functions modifying memory take arrays as parameters, and their summaries consist then of two sub-formulae: one expressing memory changes and the other, called the frame, expressing the unmodified parts. Due to the unbounded nature of arrays, the frame is often expressed by quantified formulae.

In this talk, we focus on the problem of discovering automatically inductive invariants and function summaries. We contribute to the generation of modular VCs, using Constrained Horn Clauses (CHCs), which are more amenable for SMT-MC. We first propose a new static analysis that infers the finite memory footprint of a function. That is, the memory regions that may be only accessed in a bounded number of locations. Second, we encode finite memory using finite maps, eliminating the need of quantifiers to express frame axioms. We propose a theory of finite maps adapted to CHCs and an algorithm to check satisfiability of CHCs over integers, arrays and finite maps.

3.11 On the Just-In-Time Discovery of Profit-Generating Transactions in DeFi Protocols

Arthur Gervais (Imperial College London, GB)

Joint work of Liyi Zhou, Kaihua Qin, Antoine Cully, Benjamin Livshits, Arthur Gervais
Main reference Liyi Zhou, Kaihua Qin, Antoine Cully, Benjamin Livshits, Arthur Gervais: "On the Just-In-Time Discovery of Profit-Generating Transactions in DeFi Protocols", CoRR, Vol. abs/2103.02228, 2021.
URL https://arxiv.org/abs/2103.02228

In this paper, we investigate two methods that allow us to automatically create profitable DeFi trades, one well-suited to arbitrage and the other applicable to more complicated settings. We first adopt the Bellman-Ford-Moore algorithm with DEFIPOSER-ARB and then create logical DeFi protocol models for a theorem prover in DEFIPOSER-SMT. While DEFIPOSER-ARB focuses on DeFi transactions that form a cycle and performs very well for arbitrage, DEFIPOSER-SMT can detect more complicated profitable transactions. We estimate that DEFIPOSER-ARB and DEFIPOSER-SMT can generate an average weekly revenue of 191.48ETH (76,592USD) and 72.44ETH (28,976USD) respectively, with the highest transaction revenue being 81.31ETH(32,524USD) and22.40ETH (8,960USD) respectively. We further show that DEFIPOSER-SMT finds the known economic bZx attack from February 2020, which yields 0.48M USD. Our forensic investigations show that this opportunity existed for 69 days and could have yielded more revenue if exploited one day earlier. Our evaluation spans 150 days, given 96 DeFi protocol actions, and 25 assets.

Looking beyond the financial gains mentioned above, forks deteriorate the blockchain consensus security, as they increase the risks of double-spending and selfish mining. We explore the implications of DEFIPOSER-ARB and DEFIPOSER-SMT on blockchain consensus.

Specifically, we show that the trades identified by our tools exceed the Ethereum block reward by up to 874x. Given optimal adversarial strategies provided by a Markov Decision Process (MDP), we quantify the value threshold at which a profitable transaction qualifies as Miner ExtractableValue (MEV) and would incentivize MEV-aware miners to fork the blockchain. For instance, we find that on Ethereum, a miner with a hash rate of 10% would fork the blockchain if an MEV opportunity exceeds 4x the block reward.

3.12 Gigahorse: A Declaratively-Specified EVM Binary Lifter

Neville Grech (University of Malta – Msida, MT)

Joint work of Neville Grech, Lexi Brent, Bernhard Scholz, Yannis Smaragdakis
Main reference Neville Grech, Lexi Brent, Bernhard Scholz, Yannis Smaragdakis: "Gigahorse: thorough, declarative decompilation of smart contracts", in Proc. of the 41st International Conference on Software Engineering, ICSE 2019, Montreal, QC, Canada, May 25-31, 2019, pp. 1176–1186, IEEE / ACM, 2019.
URL https://doi.org/10.1109/ICSE.2019.00120

Smart contracts on blockchain platforms (e.g. Ethereum) represent a software domain with critical correctness needs. Smart contract users and security auditors can greatly benefit from a mechanism to recover the original structure of contracts, as evident from past work: many security analyses of smart contracts begin with a decompilation step.

In this talk, we present the Gigahorse framework, which is at the core of the contract-library.com service. Contract-library.com contains the most complete, high-level decompiled representation of all Ethereum smart contracts, with security analyses applied to these in realtime. The Gigahorse framework is a decompilation and security analysis framework that natively supports Ethereum Virtual Machine (EVM) bytecode. Its internal intermediate representation of smart contracts makes implicit data- and control-flow dependencies of the EVM bytecode explicit. Using this framework we have developed and adapted several advanced high-level client analyses, including MadMax and Ethainter. All our client analyses benefit from high-level domain-specific concepts (such as "dynamic data structure storage" and "safely resumable loops") and achieve high precision and scalability.

One such client analysis, MadMax, flags contracts with a current monetary value in the $B range. (Manual inspection of a sample of flagged contracts shows that 81% of the sampled warnings do indeed lead to vulnerabilities.)

3.13 solc-verify: A Modular Verifier for Solidity Smart Contracts

Ákos Hajdu (Budapest Univ. of Technology & Economics, HU)

Joint work of Hajdu, Ákos; Jovanović, Dejan; Ciocarlie, Gabriela
Main reference Ákos Hajdu, Dejan Jovanovic: "solc-verify: A Modular Verifier for Solidity Smart Contracts", in Proc. of the Verified Software. Theories, Tools, and Experiments – 11th International Conference, VSTTE 2019, New York City, NY, USA, July 13-14, 2019, Revised Selected Papers, Lecture Notes in Computer Science, Vol. 12031, pp. 161–179, Springer, 2019.
URL https://doi.org/10.1007/978-3-030-41600-3_11

Solc-verify [1] is a source-level verification tool for Ethereum smart contracts. It takes smart contracts written in Solidity and discharges verification conditions using modular program analysis and SMT solvers. Built on top of the Solidity compiler, solc-verify reasons at the

level of the contract source code. This enables solc-verify to effectively reason about high-level contract properties while modeling low-level language semantics precisely. The contract properties, such as contract invariants, loop invariants, function pre- and post-conditions, and event specifications [1, 3] can be provided as annotations in the code by the developer. This enables automated, yet user-friendly formal verification for smart contracts.

A distinguishing feature of solc-verify is its memory model [2], which is based on a formalization that covers all features of the language related to managing state and memory. In addition, the formalization is effective: all but few features can be encoded in the quantifier-free fragment of standard SMT theories. This enables precise and efficient reasoning about the state of smart contracts. The formalization is implemented in solc-verify and we provide an extensive set of tests that covers the breadth of the required semantics. We also provide an evaluation on the test set that validates the semantics and shows the novelty of the approach compared to other Solidity-level contract analysis tools.

References

1 Ákos Hajdu and Dejan Jovanović. *solc-verify: A Modular Verifier for Solidity Smart Contracts*. VSTTE 2019
2 Ákos Hajdu and Dejan Jovanović. *SMT-Friendly Formalization of the Solidity Memory Model*. ESOP 2020
3 Ákos Hajdu, Dejan Jovanović and Gabriela Ciocarlie. *Formal Specification and Verification of Solidity Contracts with Events (short paper)*. FMBC 2020

3.14 Smart contract = contract + control + settlement

Fritz Henglein (University of Copenhagen, DK)

Joint work of Fritz Henglein, Christian Kjær Larsen, Agata Murawska
Main reference Fritz Henglein, Christian Kjær Larsen, Agata Murawska: "A Formally Verified Static Analysis Framework for Compositional Contracts", in Proc. of the Financial Cryptography and Data Security – FC 2020 International Workshops, AsiaUSEC, CoDeFi, VOTING, and WTSC, Kota Kinabalu, Malaysia, February 14, 2020, Revised Selected Papers, Lecture Notes in Computer Science, Vol. 12063, pp. 599–619, Springer, 2020.
URL https://doi.org/10.1007/978-3-030-54455-3_42

We present a smart contract architecture where a smart contract is decomposed into a declarative contract (composed from subcontracts specifying obligations and permission), contract manager (a generic program that, for any given contract, monitors or controls contract events to be consistent with the contract's semantics), and ra esource manager (a system that maintains ownership state of user-definable resource types, which accepts only resource-preserving transfers and thus guarantees nonduplication of any resource). This separation of concerns facilitates expressing contracts in a declarative domain-specific language for expressing commercial contracts and financial instruments, with formally specified denotational semantics and support for mechanized static analysis [1].

References

1 Fritz Henglein, Christian Kjær Larsen, Agata Murawska. *A Formally Verified Static Analysis Framework for Compositional Contracts*. Proc. 4th Workshop on Trusted Smart Contracts, February 2020

3.15 Speculative Smart Contracts

Jing Chen (Algorand Inc, US)

Joint work of Jing Chen, Maurice Herlihy, John Jannotti, Victor Luchangco, Liuba Shrira

Existing smart contract architectures suffer from a bottleneck problem: smart contract calls result in user code being executed in the ledger's critical path, potentially delaying simple payments and transfers. We describe an alternative smart contract structure that executes user code speculatively away from the blockchain's critical path. A secure committee validates and votes on the results of each such execution, certifying the execution's preconditions and its effects, and forwarding the certified results to a distinct consensus committee that manages access to the ledger itself.

3.16 Testing Cosmos applications with TLA+ and Apalache

Igor Konnov (Informal Systems – Wien, AT)

TLA+ is a language for formal specification of all kinds of computer systems. System designers use this language to specify concurrent, distributed, and fault-tolerant protocols, which are traditionally presented in pseudo-code. At Informal Systems, we are using TLA+ to specify and reason about the protocols that are implemented in the Tendermint blockchains and Cosmos ecosystem. To this end, we run Apalache, our symbolic model checker for TLA+.

In this talk, we show how to leverage TLA+ and Apalache to produce tests for blockchain applications. In our approach, verification engineers are incrementally writing TLA+ specifications and their expected properties. By running the model checker, they produce sequences of transactions, to be tried against the test environment. While this approach can be used for testing Cosmos applications as a black box, we find it to be the most effective when verification engineers have access to the source code.

3.17 Towards Automated Verification of Smart Contract Fairness

Yi Li (Nanyang TU – Singapore, SG)

Joint work of Ye Liu, Yi Li, Shang-Wei Lin, Rong Zhao
Main reference Ye Liu, Yi Li, Shang-Wei Lin, Rong Zhao: "Towards Automated Verification of Smart Contract Fairness", in Proc. of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2020, p. 666–677, Association for Computing Machinery, 2020.
URL https://doi.org/10.1145/3368089.3409740

Smart contracts are computer programs allowing users to define and execute transactions automatically on top of the blockchain platform. Many of such smart contracts can be viewed as games. A game-like contract accepts inputs from multiple participants, and upon ending, automatically derives an outcome while distributing assets according to some predefined rules. Without clear understanding of the game rules, participants may suffer

from fraudulent advertisements and financial losses. In this paper, we present a framework to perform (semi-)automated verification of smart contract fairness, whose results can be used to refute false claims with concrete examples or certify contract implementations with respect to desired fairness properties. We implement FairCon, which is able to check fairness properties including truthfulness, efficiency, optimality, and collusion-freeness for Ethereum smart contracts. We evaluate FairCon on a set of real-world benchmarks and the experiment result indicates that FairCon is effective in detecting property violations and able to prove fairness for common types of contracts.

3.18 Practical and Provably Sound Static Analysis of Ethereum Smart Contracts

Matteo Maffei (TU Wien, AT)

Joint work of Matteo Maffei, Clara Scheidewind, Markus Scherer, Ilya Grishchenko
Main reference Clara Schneidewind, Ilya Grishchenko, Markus Scherer, Matteo Maffei: "eThor: Practical and Provably Sound Static Analysis of Ethereum Smart Contracts", in Proc. of the CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020, pp. 621–640, ACM, 2020.
URL https://doi.org/10.1145/3372297.3417250

Ethereum has emerged as the most popular smart contract development platform, with hundreds of thousands of contracts stored on the blockchain and covering a variety of application scenarios, such as auctions, trading platforms, and so on. Given their financial nature, security vulnerabilities may lead to catastrophic consequences and, even worse, they can be hardly fixed as data stored on the blockchain, including the smart contract code itself, are immutable. An automated security analysis of these contracts is thus of utmost interest, but at the same time technically challenging for a variety of reasons, such as the specific transaction-oriented programming mechanisms, which feature a subtle semantics, and the fact that the blockchain data which the contract under analysis interacts with, including the code of callers and callees, are not statically known.

In this talk, I will present eThor, the first sound and automated static analyzer for EVM bytecode, which is based on an abstraction of the EVM bytecode semantics based on Horn clauses. In particular, our static analysis supports reachability properties, which we show to be sufficient for capturing interesting security properties for smart contracts (e.g., single-entrancy) as well as contract-specific functional properties. Our analysis is proven sound against a complete semantics of EVM bytecode and an experimental large-scale evaluation on real-world contracts demonstrates that eThor is practical and outperforms the state-of-the-art static analyzers.

This talk is based on a paper with the same title presented at CCS 2020.

3.19 What we do at Certora

Alexander Nutz (Certora – Berlin, DE)

Joint work of All members of Certora

Certora's mission is "ensuring smart contract security". To achieve this we are developing a specification language for EVM smart contracts with the goal of being accessible to programmers with only basic preexisting knowledge of formal verification. In addition, specifications should be as portable as possible. In this context it is crucial to strike a balance between minimally invasive specifications (which is important for understandability and portability) and specs that are amenable to automatic proving techniques. We are also developing a tool to automatically check these specifications, which works by translating correctness queries to SMT formulas. There are numerous challenges in having these formulas solved by today's SMT solvers. We apply a variety of simplifications and abstractions to make this feasible. Static analysis is an important enabler to make these transformations sound. When running SMT solvers we are using a portfolio approach; depending on the input, we can translate to various different encodings as well as running different solvers and configurations. We also work closely with SMT solver developers to solve remaining problems. In particular the fragment of large bit vectors (256 bit) combined with nonlinear arithmetic is crucial for our efforts while having gotten comparatively little attention in the past.

3.20 Off-Chain Protocols meet Game Theory

Sophie Rain (TU Wien, AT)

Joint work of Sophie Rain, Zeta Avarikioti, Laura Kovács, Matteo Maffei
Main reference Sophie Rain, Zeta Avarikioti, Laura Kovács, Matteo Maffei: "Towards a Game-Theoretic Security Analysis of Off-Chain Protocols", CoRR, Vol. abs/2109.07429, 2021.
URL https://arxiv.org/abs/2109.07429

On-chain protocols constitute one of the most promising approaches to solve the inherent scalability issue of blockchain technologies. The core idea is to let parties transact on-chain only once to establish a channel between them, leveraging later on the resulting channel paths to perform arbitrarily many peer-to-peer transactions on-chain. While significant progress has been made in terms of proof techniques for on- chain protocols, existing approaches do not capture the game-theoretic incentives at the core of their design, which led to overlooking significant attack vectors like the Wormhole attack in the past. This work introduces the first game-theoretic model that is expressive enough to reason about the security of on-chain protocols. We advocate the use of Extensive Form Games EFGs and introduce two instances of EFGs to capture security properties of the closing and the routing of the Lightning Network. Specifically, we model the closing protocol, which relies on punishment mechanisms to disincentivize the uploading on-chain of old channel states, as well as the routing protocol, thereby formally characterizing the Wormhole attack, a vulnerability that undermines the fee-based incentive mechanism underlying the Lightning Network.

3.21 Formal Methods in Zero-Knowledge Protocols: Challenges in the circom Programming Language

Albert Rubio (Complutense University of Madrid, ES)

Joint work of Elvira Albert, Jordi Baylina, Marta Belles-Muñoz, Hermenegildo García-Navarro, Miguel Isabel-Márquez, José Manuel Muñoz-Tapia, Clara Rodríguez-Núñez, Albert Rubio

The most widely studied language in the context of Zero-Knowledge (ZK) proofs is arithmetic circuit satisfiability. In this talk we present circom, a programming language and a compiler that allows the programmer to provide a low-level description of the arithmetic circuit together with an effective way to execute it. We will introduce challenging safety properties to be checked in circom programs and show the need of improving existing techniques to analyse and simplify the nonlinear arithmetic constraints generated by the compiler.

3.22 Sharding Smart Contracts

Ilya Sergey (National University of Singapore, SG)

Joint work of George Pîrlea, Amrit Kumar, Ilya Sergey
Main reference George Pîrlea, Amrit Kumar, Ilya Sergey: "Practical smart contract sharding with ownership and commutativity analysis", in Proc. of the PLDI '21: 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation, Virtual Event, Canada, June 20-25, 2021, pp. 1327–1341, ACM, 2021.
URL https://doi.org/10.1145/3453483.3454112

Sharding is a popular way to achieve scalability in blockchain protocols. Existing approaches for blockchain sharding, however, do not scale well when concurrent transactions alter the same replicated state component–a common scenario in Ethereum-style smart contracts.

I will outline a novel approach for efficiently sharding such transactions. It is based on a folklore idea: state-manipulating atomic operations that commute can be processed in parallel, with their cumulative result defined deterministically, while executing non-commuting operations requires one to own the state they alter. We developed a static program analysis that soundly infers ownership and commutativity summaries for smart contracts and translates those summaries to sharding signatures that are used by the blockchain protocol to maximise parallelism. Our evaluation shows that using the analysis introduces negligible overhead to the transaction validation cost, while the inferred signatures allow the system to achieve a significant increase in transaction processing throughput for real-world smart contracts.

3.23 Smart Contract Vulnerabilities and Analysis

Yannis Smaragdakis (University of Athens, GR) and Neville Grech (University of Malta – Msida, MT)

Joint work of Yannis Smaragdakis, Neville Grech, Sifis Lagouvardos, Konstantinos Triantafyllou, Ilias Tsatiris
Main reference Yannis Smaragdakis, Neville Grech, Sifis Lagouvardos, Konstantinos Triantafyllou, Ilias Tsatiris: "Symbolic value-flow static analysis: deep, precise, complete modeling of Ethereum smart contracts", Proc. ACM Program. Lang., Vol. 5(OOPSLA), pp. 1–30, 2021.
URL https://doi.org/10.1145/3485540

In this talk, I give a quick introduction to the kinds of vulnerabilities that often appear in Ethereum smart contract coding, and discuss a static analysis infrastructure that has led to multiple high-profile vulnerability disclosures in the past year.

The main analysis architecture is "symbolic value-flow" (symvalic) analysis: a technique that reasons about the program both symbolically and with concrete values, while abstracting away from the program's control flow. Precision is being maintained through a set of "dependencies" between inferred values. This analysis architecture represents an attempt to defeat the state-explosion problem (as in model checking or concrete execution) by sacrificing a small amount of precision.

3.24 Accounts vs UTXO

Philip Wadler (University of Edinburgh, GB)

Joint work of Manuel M. T. Chakravarty, James Chapman, Kenneth MacKenzie, Orestis Melkonian, Michael Peyton Jones, Philip Wadler
Main reference Manuel M. T. Chakravarty, James Chapman, Kenneth MacKenzie, Orestis Melkonian, Michael Peyton Jones, Philip Wadler: "The Extended UTXO Model", in Proc. of the Financial Cryptography and Data Security – FC 2020 International Workshops, AsiaUSEC, CoDeFi, VOTING, and WTSC, Kota Kinabalu, Malaysia, February 14, 2020, Revised Selected Papers, Lecture Notes in Computer Science, Vol. 12063, pp. 525–539, Springer, 2020.
URL https://doi.org/10.1007/978-3-030-54455-3_37

The talk offered a brief description of two approaches to tracking balances, *accounts* as used by Ethereum and *UTxO* (Unspent Transaction Outputs) as used by Bitcoin and Cardano. The strengths and weaknesses of the two are contrasted. One strength of UTxO as compared with contracts is that the precise cost of running the smart contract can be calculated in advance – there are never any surprises where, due to a change on the blockchain that occurred between submitting the transaction and running the transaction, the cost of running the transaction has changed.

References

1 Manuel M. T. Chakravarty, James Chapman, Kenneth MacKenzie, Orestis Melkonian, Michael Peyton Jones, Philip Wadler: The Extended UTXO Model. Financial Cryptography Workshops 2020: 525-539.
2 Manuel M. T. Chakravarty, James Chapman, Kenneth MacKenzie, Orestis Melkonian, Jann Müller, Michael Peyton Jones, Polina Vinogradova, Philip Wadler: Native Custom Tokens in the Extended UTXO Model. ISoLA (3) 2020: 89-111

3.25 Formal Verification of Smart Contracts with the Move Prover

Wolfgang Grieskamp (Facebook – Bellevue, US)

Joint work of David Dill, Wolfgang Grieskamp, Junkil Park, Shaz Qadeer, Meng Xu, Emma Zhong
Main reference David L. Dill, Wolfgang Grieskamp, Junkil Park, Shaz Qadeer, Meng Xu, Jingyi Emma Zhong: "Fast and Reliable Formal Verification of Smart Contracts with the Move Prover", CoRR, Vol. abs/2110.08362, 2021.
URL https://arxiv.org/abs/2110.08362

The Move Prover (MVP) is a formal verifier for smart contracts written in the Move programming language. MVP has an expressive specification language, and is fast and reliable enough that it can be run routinely by developers and in integration testing. Besides the simplicity of smart contracts and the Move language, three implementation approaches are responsible for the practicality of MVP: (1) an alias-free memory model, (2) fine-grained invariant checking, and (3) monomorphization. The entirety of the Move code for the Diem blockchain has been extensively specified and can be completely verified by MVP in a few minutes. Changes in the Diem framework must be successfully verified before being integrated into the open source repository on GitHub.

3.26 Checking Properties of Smart Contract Systems

Valentin Wüstholz (ConsenSys – Kaiserslautern, DE) and Maria Christakis (MPI-SWS – Kaiserslautern, DE)

Joint work of Dimitar Bounov, Maria Christakis, Arie Gurfinkel, Joran J. Honig, Jorge A. Navas, Richard J. Trefler, Scott Wesley, Valentin Wüstholz
Main reference Valentin Wüstholz, Maria Christakis: "Harvey: a greybox fuzzer for smart contracts", in Proc. of the ESEC/FSE '20: 28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Virtual Event, USA, November 8-13, 2020, pp. 1398–1409, ACM, 2020.
URL https://doi.org/10.1145/3368089.3417064

Ensuring the correctness of smart contracts is becoming increasingly challenging. We provide an overview of how to check custom correctness properties for complex systems of smart contracts. We introduce the Harvey fuzzer which is used in two industrial analysis services for smart contracts. We also provide a brief overview of the Scribble specification language that we use to instrument contracts with runtime checks. Finally, we introduce SmartACE, a new analysis framework for Solidity contracts.

References

1 Valentin Wüstholz, Maria Christakis: *Harvey: a greybox fuzzer for smart contracts.* ESEC/SIGSOFT FSE 2020: 1398-1409
2 Valentin Wüstholz, Maria Christakis: *Targeted greybox fuzzing with static lookahead analysis.* ICSE 2020: 789-800
3 Scott Wesley, Maria Christakis, Jorge A. Navas, Richard J. Trefler, Valentin Wüstholz, Arie Gurfinkel: *Compositional Verification of Smart Contracts Through Communication Abstraction.* SAS 2021: 429-452
4 Scott Wesley, Maria Christakis, Jorge A. Navas, Richard J. Trefler, Valentin Wüstholz, Arie Gurfinkel: *Verifying Solidity Smart Contracts Via Communication Abstraction in SmartACE.* VMCAI 2022 (to appear)

3.27 Int-blasting

Yoni Zohar (Bar-Ilan University – Ramat Gan, IL)

Joint work of Ahmed Irfan, Makai Mann, Aina Niemetz, Andres Noetzli, Mathias Preiner, Andrew Reynolds, Clark Barrett, Cesare Tinelli, Yoni Zohar

The state of the art for bit-precise reasoning in the context of Satisfiability Modulo Theories (SMT) is a SAT-based technique called bit-blasting where the input formula is first simplified and then translated to an equisatisfiable propositional formula. The main limitation of this technique is scalability, especially in the presence of large bit-widths and arithmetic operators.

In this talk we introduced an alternative technique, which we call *int-blasting*, based on a translation to an extension of integer arithmetic rather than propositional logic. We present several alternative translations, discuss their differences, and evaluate them on benchmarks that arise from verification of rewrite rule candidates for bit-vector solving, as well as benchmarks from SMT-LIB. We also provide preliminary results on 35 benchmarks that arise from smart contract verification. The evaluation shows that this technique is particularly useful for benchmarks with large bit-widths and can solve benchmarks that the state of the art cannot.

4 Working groups

4.1 Specification Languages for Smart Contracts (Group Discussion)

discussion participants

The discussion was centered around specification languages for smart contracts and broadly touched on various related topics, such as expressiveness (e.g., safety properties, liveness properties, hyperproperties) and accessibility to programmers without background in formal methods.

4.2 Verifying Arithmetic Circuits from Zero Knowledge Applications

Leo Alt and Nikolaj Bjørner

Zero knowledge cryptography enables private computation in the form of zkSNARKs [1], that is, you can use an application without revealing some private input. The ZCash blockchain pioneered private cryptocurrency transactions, and nowadays several different applications use the same or similar technology in different ways. These computations that are performed on private data are represented as polynomials in the deepest layer of the proof system. However, initially, they are represented by arithmetic circuits. The small but huge difference from the machine circuits we may be used too is that the arithmetic on these circuits is performed over a very large prime field, since that is where the cryptographic primitives operate.

We studied how the ZoKrates [2] compiler translates high level program statements into arithmetic circuits. Because of the nature of the proof system and the verification process in the zkSNARK workflow (which I will not expand here), it is useful for such a compiler to not translate the statements intro precise constraints, but use rather weaker constraints to save circuit and proof size. In some cases, this may lead to nondeterminism in the circuit, which may or may not be harmful. This is under control for the statements that the compiler itself generates, but may become problematic once a developer starts writing custom constraints, as is common in Circom [3] and TurboPLONK [4].

We discussed about safety properties over these circuits, but converged on the smaller problem of detecting nondeterminism of a circuit. The research question now is, how can we modify/extend the tools we currently have in order to check the stated problem? If we just feed the circuit into an SMT solver that has Nonlinear Arithmetic support with extra modulo operations on all sums and multiplications, we know that this is quickly going to explode. So we need a specific approach that works fundamentally better on prime fields. Moreover, we also know that these primes are very large (254 bits), so this also needs to be taken into account when designing an algorithm for that. There are algorithms for finding roots of a system of polynomials, also on prime fields, but their complexity is prohibitively high in practice. However, we do not need full roots solving, we would be happy with satisfiability. So how can we design something in the middle?

References

1 https://zcash.github.io/halo2/concepts/proofs.html
2 https://zokrates.github.io/
3 https://github.com/iden3/circom
4 https://zcash.github.io/halo2/concepts/arithmetization.html

Participants

- Elvira Albert
Complutense University of Madrid, ES
- Leonardo Alt
Ethereum – Berlin, DE
- Nikolaj S. Bjørner
Microsoft – Redmond, US
- Maria Christakis
MPI-SWS – Kaiserslautern, DE
- Marco Eilers
ETH Zürich, CH
- Grzegorz Fabianski
University of Warsaw, PL
- Josselin Feist
Trail of Bits Inc. – New York, US
- Bryan Ford
EPFL Lausanne, CH
- Diego Garbervetsky
University of Buenos Aires, AR
- Isabel Garcia-Contreras
IMDEA Software – Madrid, ES
- Arthur Gervais
Imperial College London, GB
- Neville Grech
University of Malta – Msida, MT
- Wolfgang Grieskamp
Facebook – Bellevue, US
- Fritz Henglein
University of Copenhagen, DK
- Matteo Maffei
TU Wien, AT
- Alexander Nutz
Certora – Berlin, DE
- Sophie Rain
TU Wien, AT
- Albert Rubio
Complutense University of Madrid, ES
- Tanja Schindler
Universität Freiburg, DE
- Yannis Smaragdakis
University of Athens, GR
- Valentin Wüstholz
ConsenSys – Kaiserslautern, DE

Remote Participants

- Massimo Bartoletti
University of Cagliari, IT
- Andreea Buterchi
MPI-SWS – Kaiserslautern, DE
- Jing Chen
Stony Brook University, US
- Shuo Chen
Microsoft Research Asia – Beijing, CN
- Ankush Das
Amazon – Cupertino, US
- Stefan Dziembowski
University of Warsaw, PL
- Ákos Hajdu
Budapest Univ. of Technology & Economics, HU
- Aniket Kate
Purdue University – West Lafayette, US
- Markulf Kohlweiss
University of Edinburgh, GB
- Igor Konnov
Informal Systems – Wien, AT
- Yi Li
Nanyang TU – Singapore, SG
- Victor Luchangco
Algorand – Boston, US
- Anastasia Mavridou
NASA – Moffett Field, US
- Noam Rinetzky
Tel Aviv University, IL
- Grigore Rosu
University of Illinois – Urbana-Champaign, US
- Giulia Scaffino
TU Wien, AT
- Gerardo Schneider
University of Gothenburg, SE
- Ilya Sergey
National University of Singapore, SG
- Zhong Shao
Yale University – New Haven, US
- Philip Wadler
University of Edinburgh, GB
- Yoni Zohar
Bar-Ilan University – Ramat Gan, IL

Report from Dagstuhl Seminar 21432

Probabilistic Numerical Methods – From Theory to Implementation

Edited by

Philipp Hennig[1], Ilse C.F. Ipsen[2], Maren Mahsereci[3], and Tim Sullivan[4]

1 **Universität Tübingen, DE,** philipp.hennig@uni-tuebingen.de
2 **North Carolina State University – Raleigh, US,** ipsen@ncsu.edu
3 **Universität Tübingen, DE,** maren.mahsereci@uni-tuebingen.de
4 **University of Warwick – Coventry, GB,** t.j.sullivan@warwick.ac.uk

Abstract

Numerical methods provide the computational foundation of science, and power automated data analysis and inference in its contemporary form of machine learning. Probabilistic numerical methods aim to explicitly represent uncertainty resulting from limited computational resources and imprecise inputs in these models. With theoretical analysis well underway, software development is now a key next step to wide-spread success. This seminar brought together experts from the forefront of machine learning, statistics and numerical analysis to identify important open problems in the field and to lay the theoretical and practical foundation for a software stack for probabilistic numerical methods.

Seminar October 24–29, 2021 – http://www.dagstuhl.de/21432
2012 ACM Subject Classification Computing methodologies → Machine learning; Mathematics of computing → Numerical analysis
Keywords and phrases Machine learning, Numerical analysis, Probabilistic numerics
Digital Object Identifier 10.4230/DagRep.11.9.102
Edited in cooperation with Jonathan Wenger

1 Executive Summary

Philipp Hennig (Universität Tübingen, DE)
Ilse C.F. Ipsen (North Carolina State University – Raleigh, US)
Maren Mahsereci (Universität Tübingen, DE)
Tim Sullivan (University of Warwick – Coventry, GB)
Jonathan Wenger (Universität Tübingen, DE)

Probabilistic Numerical algorithms frame a numerical task as a statistical inference problem, expressed in the language of probabilistic inference. The key advantage of this approach is that it allows quantification of uncertainty arising from finite computational resources, and to combine thus with other forms of uncertainty, in particular those arising from model misspecification, finite observational data, and measurement errors. In recent years, algorithms arising from this formalism have repeatedly shown that they can enrich and improve upon classic methods in tasks where

Probabilistic Numerical Methods – From Theory to Implementation, *Dagstuhl Reports*, Vol. 11, Issue 09, pp. 102–119
Editors: Philipp Hennig, Ilse C.F. Ipsen, Maren Mahsereci, and Tim Sullivan

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

- hyperparameter adaptation is not straightforward;
- computational stochasticity and low precision play a prominent role;
- limited data make uncertainty quantification a key functionality;
- related problems have to be solved repeatedly;
- and where extreme scale or tight budgets call for rough approximations at low cost.

Probabilistic Numerics lies at the intersection of machine learning within computer science and numerical analysis within applied mathematics. This interdisciplinary nature raises an exciting and challenging set of viewpoints with regards to goals and challenges of the field. The first goal of this seminar was to rekindle our community following two years of pandemic lockdown, to provide an opportunity to update others on one's own research, and to discuss new directions and ideas together. We were lucky to assemble – both in-person and remote – a diverse group of people from computer science, machine learning, from statistics, optimization, and from numerical analysis.

The second key goal of this seminar was to take the next step in the development of probabilistic numerical methods by focusing on their implementation. From `Lapack` to `SciPy` to `PyTorch`, open-source software libraries have driven scientific advancement in their respective domains. Such libraries accelerate research, enable benchmarking and promote the development of new methods via rapid prototyping. Most importantly, they are a necessary step towards their use in applications. While considerable advances in the theoretical understanding of probabilistic numerical methods have been made, the lack of high-quality implementations is holding back their adoption. In response, we recently started a community effort to develop an open-source framework named `ProbNum` (`http://probnum.org`).

A central theme of Dagstuhl seminars is the open, collaborative atmosphere with a focus on new ideas and tangible outcomes as opposed to existing work. The seminar stimulated multiple focussed discussions around software and additions to ProbNum. Examples included how to best include automatic differentiation functionality, or how to expand the package's Bayesian quadrature functionalities. It also set the starting point for potential new research collaborations on probabilistic linear solvers and probabilistic numerical methods for PDEs. Even at this point, shortly after the seminar's conclusion, two tangible products are already available: the seminar's participants jointly created a Probabilistic Numerics Wikipedia page `https://en.wikipedia.org/wiki/Probabilistic_numerics`, and the implementation sessions culminated with a preprint for the community library `ProbNum` [1].

As the organizers we want to thank all participants, both physical and virtual, for their interesting talks, the stimulating discussions and the collaborative overall atmosphere. We also want to thank Schloss Dagstuhl for their technical support that made the challenging hybrid format possible.

References

1 Jonathan Wenger, Nicholas Krämer, Marvin Pförtner, Jonathan Schmidt, Nathanael Bosch, Nina Effenberger, Johannes Zenn, Toni Karvonen Alexandra Gessner, François-Xavier Briol, Maren Mahsereci, and Philipp Hennig. ProbNum: Probabilistic numerics in Python. arxiv preprint, 2021. URL `http://arxiv.org/abs/2112.02100`

2 Table of Contents

3 Overview of Talks

3.1 Approximate Gaussian Process Regression as an Early Stopping Problem

Simon Bartels (University of Copenhagen, DK)

Joint work of Simon Bartels, Pablo Moreno-Munoz, Kristoffer Stensbo-Smidt, Wouter Boomsma, Jes Frellsen, Soren Hauberg

This talk is about a method to fit Gaussian process regression models to large datasets from only a subset of the data. The novelty of this approach is that the size of the subset is selected on the fly during inference with little computational overhead. This is achieved by monitoring probabilistic bounds on the model evidence that tighten as more data is processed. Remarkably, these bounds are largely composed of terms that appear in intermediate steps of the standard Cholesky decomposition, allowing to adaptively stop the decomposition once enough data have been observed.

3.2 ProbNumDiffEq.jl: Fast and Practical ODE Filters in Julia

Nathanael Bosch (Universität Tübingen, DE)

Joint work of Nathanael Bosch, Filip Tronarp, Philipp Hennig

ProbNumDiffEq.jl provides an implementation of ODE filters in Julia, building on top of the established DifferentialEquations.jl ecosystem. In this session, we discuss how software for probabilistic numerics (PN) can benefit from existing non-PN code to obtain higher performance, a larger set of features, and to reach a wider audience. This is facilitated by the composability of the Julia programming language, and the modular structure of DifferentialEquations.jl. We demonstrate the speed and ease of use of the resulting ODE solvers in a live code demo. Additionally, we show how ODE filters can be extended to include additional knowledge about the problem structure (such as second-order ODEs or energy conservation) and demonstrate how to solve such problems in practice.

3.3 Iterative Unbiased Linear Solvers for Gaussian Processes

Maurizio Filippone (EURECOM – Biot, FR)

Joint work of Maurizio Filippone, Raphael Engler
Main reference Maurizio Filippone, Raphael Engler: "Enabling scalable stochastic gradient-based inference for Gaussian processes by employing the Unbiased LInear System SolvEr (ULISSE)", in Proc. of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, JMLR Workshop and Conference Proceedings, Vol. 37, pp. 1015–1024, JMLR.org, 2015.
URL http://proceedings.mlr.press/v37/filippone15.html

In applications of Gaussian processes where quantification of uncertainty is of primary interest, it is necessary to accurately characterize the posterior distribution over covariance parameters. This paper proposes an adaptation of the Stochastic Gradient Langevin Dynamics algorithm

to draw samples from the posterior distribution over covariance parameters with negligible bias and without the need to compute the marginal likelihood. In Gaussian process regression, this has the enormous advantage that stochastic gradients can be computed by solving linear systems only. A novel unbiased linear systems solver based on parallelizable covariance matrix-vector products is developed to accelerate the unbiased estimation of gradients. The results demonstrate the possibility to enable scalable and exact (in a Monte Carlo sense) quantification of uncertainty in Gaussian processes without imposing any special structure on the covariance or reducing the number of input vectors.

3.4 Bayesian Cubature with Low Discrepancy Sequences in QMCPy

Fred J. Hickernell (Illinois Institute of Technology – Chicago, US)
Jagadeeswaran Rathinavel (Illinois Institute of Technology – Chicago, US)
Aleksei Sorokin (Illinois Institute of Technology – Chicago, US)

Bayesian cubature proceeds by constructing credible intervals for the integral based on a prior distribution for the integrand combined with sampled integrand values. We explain this approach has been developed using low discrepancy (digital net and lattice) sampling and matching covariance kernels to expedite the computation to be much faster than $\mathcal{O}(n^3)$, where n is the sample size. We tune the hyper-parameters of our covariance kernels using function data to increase the chance that the integrand is not an outlier. We also show how we have implemented these Bayesian cubature algorithms in QMCPy, a community supported Python library for quasi-Monte Carlo calculations. Some preliminary results also call into question the Bayesian assumption. This matter requires further study.

3.5 BayesCG: A probablistic numeric linear solver

Ilse C.F. Ipsen (North Carolina State University – Raleigh, US)

Joint work of Ilse C.F. Ipsen, Tim W. Reid, Jon Cockayne, Chris J. Oates
Main reference Tim W. Reid, Ilse C. F. Ipsen, Jon Cockayne, Chris J. Oates: "A Probabilistic Numerical Extension of the Conjugate Gradient Method", CoRR, Vol. abs/2008.03225, 2020.
URL https://arxiv.org/abs/2008.03225

We present the probabilistic linear solver BayesCG, an extension of the Conjugate Gradient method (CG) that relies on probability distributions to capture uncertainty due to early termination when solving linear systems with real symmetric positive definite coefficient matrices. We present a CG-based implementation of BayesCG with a structure-exploiting prior distribution. The BayesCG output consists of CG iterates and posterior covariances that can be propagated to subsequent computations. The covariances are low-rank and maintained in factored form. This allows easy generation of accurate samples to probe uncertainty in subsequent computations. We discuss the choice of efficient prior distributions, and end with speculation on how to propagate uncertainty through computational pipelines.

3.6 bayesquad: Bayesian quadrature in ProbNum

Toni Karvonen (University of Helsinki, FI)
Alexandra Gessner (Universität Tübingen, DE)

We present the current state of the implementation of Bayesian quadrature in ProbNum and discuss features that are to be included in the future. The talk contains examples of the use of the two main Bayesian quadrature functions, `bayesquad` and `bayesquad_from_data`, and their comparison to the corresponding SciPy functions.

3.7 Fun with ODE filters

Peter Nicholas Krämer (Universität Tübingen, DE)

Joint work of Peter Nicholas Krämer, Nathanael Bosch, Jonathan Schmidt, Philipp Hennig
Main reference Nicholas Krämer, Philipp Hennig: "Stable Implementation of Probabilistic ODE Solvers", CoRR, Vol. abs/2012.10106, 2020.
URL https://arxiv.org/abs/2012.10106

This talk surveys recent advances of probabilistic solvers for differential equations. At first, stable implementation of probabilistic, filtering-based ODE solvers is discussed; both, in low- and high-dimensional settings. Then, efficient probabilistic solvers for ODE boundary value problems and partial differnetial equations are explained.

3.8 ProbNum: Probabilistic Numerical Methods in Python

Maren Mahsereci (Universität Tübingen, DE)
Jonathan Wenger (Universität Tübingen, DE)

Joint work of Jonathan Wenger, Nicholas Krämer, Marvin Pförtner, Jonathan Schmidt, Nathanael Bosch, Nina Effenberger, Johannes Zenn, Alexandra Gessner, Toni Karvonen, François-Xavier Briol, Maren Mahsereci, Philipp Hennig
URL http://probnum.org

ProbNum is a Python library that provides probabilistic numerical solvers to a wider audience. In the talk, we describe the current state and functionality of ProbNum and highlight some benefits of open source collaboration for students and for the community. The second part of the talk contains a live demonstration of some of the ProbNum solvers.

3.9 Solving and Learning Differential Equations with Gaussian Processes

Houman Owhadi (California Institute of Technology – Pasadena, US)

Joint work of Houman Owhadi, Yifan Chen, Boumediene Hamzi, Bamdad Hosseini, Romit Maulik, Yoo Gene Ryan, Florian Schäfer, Clint Scovel, Andrew Stuart

Main reference Yifan Chen, Bamdad Hosseini, Houman Owhadi, Andrew M. Stuart: "Solving and learning nonlinear PDEs with Gaussian processes", J. Comput. Phys., Vol. 447, p. 110668, 2021.
URL https://doi.org/10.1016/j.jcp.2021.110668

Main reference B. Hamzi, R. Maulik, H. Owhadi: "Simple, low-cost and accurate data-driven geophysical forecasting with learned kernels", Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, Vol. 477(2252), p. 20210326, 2021.
URL https://doi.org/10.1098/rspa.2021.0326

Main reference Boumediene Hamzi, Houman Owhadi: "Learning dynamical systems from data: A simple cross-validation perspective, part I: Parametric kernel flows", Physica D: Nonlinear Phenomena, Vol. 421, p. 132817, 2021.
URL https://doi.org/10.1016/j.physd.2020.132817

Main reference Houman Owhadi, Clint Scovel: "Operator-Adapted Wavelets, Fast Solvers, and Numerical Homogenization: From a Game Theoretic Approach to Numerical Approximation and Algorithm Design", Cambridge University Press, 2019.
URL https://doi.org/10.1017/9781108594967

Main reference Houman Owhadi, Gene Ryan Yoo: "Kernel Flows: From learning kernels from data into the abyss", J. Comput. Phys., Vol. 389, pp. 22–47, 2019.
URL https://doi.org/10.1016/j.jcp.2019.03.040

We present a simple, rigorous, and unified framework for solving and learning (possibly nonlinear) differential equations (PDEs and ODEs) using the framework of Gaussian processes/kernel methods. For PDEs the proposed approach: (1) provides a natural generalization of collocation kernel methods to nonlinear PDEs and Inverse Problems; (2) has guaranteed convergence for a very general class of PDEs, and comes equipped with a path to compute error bounds for specific PDE approximations; (3) inherits the state-of-the-art computational complexity of linear solvers for dense kernel matrices. For ODEs, we illustrate the efficacy of the proposed approach by extrapolating weather/climate time series obtained from satellite data and highlight the importance of using adapted/learned kernels.

3.10 Implementing BayesCG Under The Krylov Prior

Timothy Reid (North Carolina State University – Raleigh, US)

Joint work of Timothy Reid, Ilse C.F. Ipsen, Jon Cockayne, Chris J Oates

Main reference Tim W. Reid, Ilse C. F. Ipsen, Jon Cockayne, Chris J. Oates: "A Probabilistic Numerical Extension of the Conjugate Gradient Method", CoRR, Vol. abs/2008.03225, 2020.
URL https://arxiv.org/abs/2008.03225

We present solutions to the computational challenges associated with implementing BayesCG under the Krylov prior. We also have a discussion of possible solutions to open questions related to implementing probabilistic numerical linear solvers.

3.11 A Probabilistic State Space Model for Joint Inference from Differential Equations and Data

Jonathan Schmidt (Universität Tübingen, DE)

Joint work of Jonathan Schmidt, Nicholas Krämer, Philipp Hennig
Main reference Jonathan Schmidt, Nicholas Krämer, Philipp Hennig: "A Probabilistic State Space Model for Joint Inference from Differential Equations and Data", CoRR, Vol. abs/2103.10153, 2021.
URL https://arxiv.org/abs/2103.10153

This talk shows how different sources of information – mechanistic knowledge and empirical observations – can be combined to solve differential equations that are assumed to underlie observed data. The mismatch between both sources of knowledge is captured by introducing a latent force acting on the vector field. The method is showcased by fitting an SIRD-model to COVID-19 case counts.

3.12 A Probabilistic View on Sparse Cholesky Factorization

Florian Schäfer (Georgia Institute of Technology – Atlanta, US)

Joint work of Jiong Chen, Mathieu Desbrun, Jin Huang, Matthias Katzfuss, Houman Owhadi, Florian Schäfer, Tim J. Sullivan
Main reference Florian Schäfer, Houman Owhadi: "Sparse recovery of elliptic solvers from matrix-vector products", CoRR, Vol. abs/2110.05351, 2021.
URL https://arxiv.org/abs/2110.05351
Main reference Florian Schäfer, Timothy John Sullivan, Houman Owhadi: "Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity", CoRR, Vol. abs/1706.02205, 2017.
URL http://arxiv.org/abs/1706.02205
Main reference Florian Schäfer, Matthias Katzfuss, Houman Owhadi: "Sparse Cholesky factorization by Kullback-Leibler minimization", CoRR, Vol. abs/2004.14455, 2020.
URL https://arxiv.org/abs/2004.14455
Main reference Jiong Chen, Florian Schäfer, Jin Huang, Mathieu Desbrun: "Multiscale cholesky preconditioning for ill-conditioned problems", ACM Trans. Graph., Vol. 40(4), pp. 81:1–81:13, 2021.
URL https://doi.org/10.1145/3450626.3459851

The guiding theme of this talk is the probabilistic interpretation of numerical linear algebra, in particular of the Cholesky factorization of kernel matrices and discretized elliptic partial differential equations. By using this interpretation, we derive simple, fast solvers with state-of-the-art complexity vs. accuracy guarantees for general elliptic differential- and integral equations. We furthermore derive an algorithm that allows the reconstruction, or learning, of elliptic solution operators from a number of solution pairs that scales only polylogarithmically in the target accuracy. Our methods come with rigorous error estimates, are easy to parallelize, and show good performance in practice.

3.13 Convergence and Robustness of Gaussian Process Regression

Aretha Teckentrup (University of Edinburgh, GB)

Main reference Aretha L. Teckentrup: "Convergence of Gaussian Process Regression with Estimated Hyper-Parameters and Applications in Bayesian Inverse Problems", SIAM/ASA J. Uncertain. Quantification, Vol. 8(4), pp. 1310–1337, 2020.
URL https://doi.org/10.1137/19M1284816

We are interested in the task of estimating an unknown function from data, given as a set of point evaluations. In this context, Gaussian process regression is often used as a Bayesian inference procedure, and we are interested in the convergence as the number of data points goes to infinity. Hyper-parameters appearing in the mean and covariance structure of the Gaussian process prior, such as smoothness of the function and typical length scales, are often unknown and learnt from the data, along with the posterior mean and covariance. We work in the framework of empirical Bayes', where a point estimate of the hyper-parameters is computed, using the data, and then used within the standard Gaussian process prior to posterior update. Using results from scattered data approximation, we provide a convergence analysis of the method applied to a fixed, unknown function of interest.

3.14 Black Box Probabilistic Numerics

Onur Teymur (University of Kent, GB)

Joint work of Onur Teymur, Christopher N. Foley, Philip G. Breen, Toni Karvonen, Chris. J. Oates
Main reference Onur Teymur, Christopher N. Foley, Philip G. Breen, Toni Karvonen, Chris J. Oates: "Black Box Probabilistic Numerics", CoRR, Vol. abs/2106.13718, 2021.
URL https://arxiv.org/abs/2106.13718

In many numerical algorithms, intermediate numerical outputs are nonlinearly related to the quantity of interest, rendering the proper conditioning of random variables in the probabilistic numerics paradigm difficult and limiting the range of numerical tasks that can be addressed by existing approaches. In this presentation we introduce an idea to construct probabilistic numerical methods based only on the final output from a traditional method. A convergent sequence of approximations to the quantity of interest constitute a dataset, from which the limiting quantity of interest can be extrapolated, in a probabilistic analogue of Richardson's deferred approach to the limit. This black box approach (1) massively expands the range of tasks to which probabilistic numerics can be applied, (2) inherits the features and performance of state-of-the-art numerical methods, and (3) enables provably higher orders of convergence to be achieved. We present several proof-of-concept applications, such as for nonlinear ordinary and partial differential equations, as well as for eigenvalue problems – the latter a setting for which no probabilistic numerical methods have yet been developed.

3.15 The MAP for ODEs

Filip Tronarp (Universität Tübingen, DE)

Joint work of Filip Tronarp, Simo Särkkä, Philipp Hennig
Main reference Filip Tronarp, Simo Särkkä, Philipp Hennig: "Bayesian ODE solvers: the maximum a posteriori estimate", Stat. Comput., Vol. 31(3), p. 23, 2021.
URL https://doi.org/10.1007/s11222-021-09993-7

There is a growing interest in probabilistic numerical solutions to ordinary differential equations. In this talk, the maximum a posteriori estimate is studied under the class of ν times differentiable linear time-invariant Gauss-Markov priors. The maximum a posteriori estimate corresponds to an optimal interpolant in the reproducing kernel Hilbert space associated with the prior, which in the present case is equivalent to a Sobolev space of smoothness ν. Subject to mild conditions on the vector field, convergence rates of the maximum a posteriori estimate are then obtained via methods from nonlinear analysis and scattered data approximation.

3.16 Priors in Probabilistic Numerics

Zi Wang (Google – Cambridge, US)

Joint work of Zi Wang, George E. Dahl, Kevin Swersky, Chansoo Lee, Zelda Mariet, Zackary Nado, Justin Gilmer, Jasper Snoek, Zoubin Ghahramani
Main reference Zi Wang, Caelan Reed Garrett, Leslie Pack Kaelbling, Tomás Lozano-Pérez. "Learning compositional models of robot skills for task and motion planning." Int. J. Robotics Res. 40(6-7) (2021).
Zi Wang, George E. Dahl, Kevin Swersky, Chansoo Lee, Zelda Mariet, Zackary Nado, Justin Gilmer, Jasper Snoek, and Zoubin Ghahramani. "Automatic prior selection for meta Bayesian optimization with a case study on tuning deep neural network optimizers." arXiv preprint arXiv:2109.08215 (2021).

How do we understand priors? Surrounding this question, I discuss some of my thoughts on different forms of priors and how they impact the design and performance of algorithms. In particular, I group priors into 3 categories: priors in nature, engineering priors and data priors. Data priors are very relevant to how we may be able to set priors better in probabilistic numerics. I use Bayesian optimization to illustrate how priors estimated from multi-task data can lead to better performance than hand-selected priors.

4 Working groups

4.1 Probabilistic Numerics Wikipedia Page

François-Xavier Briol (University College London, GB)

This working group focused on developing a Wikipedia page for the field of probabilistic numerics. The aim was to help raise the visibility of the field, but also to serve as a brief introduction for non-experts. Overall, this was very successful; less than a month after it was first created, the page has been viewed around 2000 times and has 30 distinct authors. A second page for the sub-field of Bayesian quadrature is also currently undergoing the Wikipedia approval process.

4.2 ProbNum Library Scope

Jonathan Wenger (Universität Tübingen, DE)

ProbNum implements probabilistic numerical methods in Python. Such methods solve numerical problems from linear algebra, optimization, quadrature and differential equations using probabilistic inference. This session discusses the exact use cases of a PN library and defines the separation from other libraries for classic numerics, probabilistic programming and uncertainty quantification.

4.3 From (Prior) Information to Usable Algorithm

Jonathan Wenger (Universität Tübingen, DE)

Prior information is often touted as a prime example of why probabilistic numerical methods have an advantage over classical methods. But what types of prior information are available for different numerical problems and what can we actually encode?

5 Panel discussions

5.1 Bayesian Optimization

Roman Garnett (Washington University – St. Louis, US)

Main reference Roman Garnett: "Bayesian Optimization", Cambridge University Press, 2022.
URL https://bayesoptbook.com

I discussed a number of issues related to Bayesian optimization (BayesOpt) and themes behind several of the success stories in that field. I identified one major theme in the BayesOpt literature – a focus on expensive functions – which has allowed research in that field consider correspondingly expensive approaches to modeling and policy design. In probabilistic numerics (PN), on the other hand, there seems to be a general trend away from this "expensive regime." I concluded by posing three questions to the group:

- To what extent is the fidelity of modeling in PN? Is the "expensive regime" of interest?
- To what extent is the sophistication/myopia of algorithms an issue in PN? Is the "expensive regime" of interest?
- How can we improve PN pipelines through user interaction?

An engaging discussion ensued.

5.2 Scientific Software Development

Andrei Paleyes (University of Cambridge, GB)
Maren Mahsereci (Universität Tübingen, DE)
Michael McKerns (Los Alamos National Laboratory, US)
Masha Naslidnyk (Amazon Research Cambridge, GB)
Geoff Pleiss (Columbia University – New York, US)
Jonathan Wenger (Universität Tübingen, DE)

Recent years saw the high amount of scientific software produced by research groups in academia and industry. A good software package can be an asset of a research group, increasing productivity of its members and enabling faster experimentation. However, development and maintenance of a high quality software can be a challenging task. Recent trends and challenges of scientific software development were discussed by the panel at the Dagstuhl seminar 21432 "Probabilistic Numerical Methods – From Theory to Implementation". This document is the summary of the discussion.

5.2.1 Introduction

Over the past 6 years the proportion of papers in machine learning that are released along with the codebase used to produce the results tripled, growing from under 10% in 2015 to nearly 30% in 2021[1]. While some of these codebases can remain unchanged since their release, there is also a growing tendency to turn them into reusable software components. This decision can bring benefits, but also comes at a cost. The panel discussed challenges of developing and maintaining software tools for academic and industry teams, associated trade offs and past experience panelists had with their software packages.

5.2.2 Should scientists worry about developing software or just use existing tools?

While development of a new scientific tool does not need to be a goal of each research project, it may bring certain benefits worth pursuing. Software that is intended to be reused over multiple research projects or over several years tends to be well tested, thus further ensuring quality of the scientific results produced. Long-lived software also tends to be well documented, thus enabling others to reproduce results from prior work, and to directly compare their own new results produced with the software. Finally, good scientific software can accelerate research and the development of other software produced by a research group. A research group can leverage good scientific software to provide a solid reusable software foundation to extend, with the benefit of several years of development, testing, and validation, rather than starting from scratch on each new project.

Since research groups are normally built around a certain research direction, a software package may also emerge naturally, even though the group never had such a goal, by accumulating reusable components from multiple projects done by the group. A research group may decide to produce and maintain their own software, however an equally viable (and potentially more fruitful) choice is to adopt an existing open source package, and contribute to it's development. Interestingly, the decision to contribute to an existing codebase as opposed to developing a new software package from scratch is unfortunately rare.

[1] According to `https://paperswithcode.com/trends`, accessed 21.11.2021

5.2.3 What are challenges in maintaining scientific software?

Scientific code does not always have to be production ready, and therefore does not always need to uphold high standards in terms of software design or test coverage. In that sense, the academic environment can be liberating. However, when this code is turned into a reusable toolbox, a certain level of quality is anticipated. This can be a challenge for researchers, as most of them do not have software engineering experience, and therefore might not be aware of best practices. Having access to a trained software developer might help, but again not every lab can afford hiring one. For that reason a lot of researchers in machine learning have to learn engineering skills on the fly in addition to producing research in their primary topic. This can be a steep and challenging learning curve.

In years past, a researcher learning good software development in an academic environment was often at the cost of an academic career, as the generation of software products was not seen as part of countable academic achievements. However, the ability to produce good software concurrent with academic research is now a highly-desired skill, even though a record of software contributions is still generally measured unfavorably against a record of journal publications. As journals and funding agencies have begun to expect research groups to develop and extend software when they produce new research results, the ability for an academic environment to attract and retain researchers with good software dvelopment skills has increased.

Maintaining software beyond the original funding source, or the original developer, can be challenging. Long-lived research software like mystic [1, 2] sustain steady development by being central to new proposals and new research funding. Software tied to project funding can pay researchers to serve as developers as part of their jobs, as opposed to requiring additional efforts outside of their funded work.

5.2.4 Can you hope to compete against industrial research labs when writing a software package?

At the first glance it may seem impossible for an academic research group to compete with industry when it comes to software development. Teams in industry are often well funded, have access to better infrastructure and attract significant software engineering talent. In fact, the competition is not as severe as it may seem. Companies normally only fund engineering effort that is able to support their business goals. As such, software engineering projects created by industrial researchers are aimed at deploying well established research ideas in practical applications. Groups in universities and research centers, on the other hand, are working on the cutting edge research ideas, and therefore are more likely to create a software project in a direction, which is not ready for production yet.

National research laboratories have traditionally been a source of reusable scientific software, as researchers often work on multiple well-funded projects with expected deliverables and shortened timelines, and thus have similar conditions to those found in industry. Projects at national laboratories often have industry partners, where the groups at the national laboratory are tasked with providing high-quality cutting-edge research tools.

5.2.5 How do you tread the line between spending too much time on maintenance, implementing features for others and moving your own research forward?

Navigating the balance between maintenance and novel development is always a trade off. However, while maintenance will always demand certain time investment from the package owner, there are several general rules that can alleviate the load.

Most importantly, time invested in software design always pays off. Simple high level API means less errors due to an improper use. An extensible software architecture means more straightforward contributions that are easier to review. Good test coverage means fewer bugs. All of these are good engineering practices that also help reduce the maintenance load. As an example, the author of dill[2] and mystic maintains a range of open source scientific packages, and the overall load is only manageable because each package was developed with accordance to good engineering practices where the reusability of the software enables it to be directly tied to new research funding.

Successful software packages can attract a strong community of users and contributors. Such a community can, to a certain extent, share the load of maintenance, as is the case now with GPyTorch [3]. Clearly such a community requires time and effort to gather, and therefore can be seen as a long-term investment. Even after the initial release of Emukit [4] its authors still continue to promote the package both in academia and in industry, which helps growing the community of its users.

However there can be cases when authors of the package are no longer interested in spending their time and effort in maintaining it. It is important to detect such situations when they happen, and identify the best way forward, which can be even discontinuing maintenance of the package, as is now the case with GpyOpt [5].

5.2.6 How do you scale a project from a small group to an open-source project with outside contributors?

As already mentioned, scientific packages can become a primary tool for a an research group. This link can also work in the other direction, where a successful research idea can be implemented as a new functionality of the package. This setup works well particularly for student and intern projects, and may even have an amplifying effect, where a scientist who contributed to a project joins another research group and may attract further users. Declaring contribution as an additional goal of a research project helps to accelerate growth during the early stages.

Once the package gains a certain level of maturity, it may attract interest from the industry partners. They can benefit by applying the scientific methods implemented in the package to their business problems. The level of contribution that companies can bring back varies widely, from feature requests and bug reports, to the contribution of individual features, to financial support for the library and its authors.

An ideal scenario is finding industrial collaborators who wish to build software on top of your package. For example, Facebook built the BoTorch Bayesian optimization package [6] on top of the GPyTorch package. As a result, the BoTorch team has a vested interest in ensuring that GPyTorch is maintained, yet ownership still ultimately resides with the original GPyTorch team. It can be challenging to find these collaborations, and companies are more likely to reach out if the software package is actively maintained by responsive developers.

However, at each stage of development it is important to define the purpose and goals of the package and therefore clearly separate what functionality is in- and out of scope.

References

1 M. McKerns, P. Hung, and M. Aivazis. mystic: highly-constrained non-convex optimization and UQ. `http://pypi.python.org/pypi/mystic`, 2009.

[2] https://dill.readthedocs.io/en/latest/dill.html

2 M. McKerns, L. Strand, T. J. Sullivan, A. Fang, and M. Aivazis. Building aframework for predictive science. In Proceedings of the 10th Python in Science Conference, pages 67-78, 2011. http://arxiv.org/pdf/1202.1056.
3 J. Gardner, G. Pleiss, K. Q. Weinberger, D. Bindel, and A. G. Wilson. GPyTorch: Blackbox matrix-matrix gaussian process inference with GPU acceleration. Advances in Neural Information Processing Systems, 31, 2018.
4 A. Paleyes, M. Pullin, M. Mahsereci, N. Lawrence, and J. González. Emulation of physical processes with Emukit. In Second Workshop on Machine Learning and the Physical Sciences, NeurIPS, 2019.
5 The GPyOpt authors. GPyOpt: A Bayesian optimization framework in Python. http://github.com/SheffieldML/GPyOpt, 2016
6 M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson,and E. Bakshy. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In Advances in Neural Information Processing Systems, 2020.

6 Open problems

6.1 Calibration of PN methods

Giacomo Garegnani (EPFL – Lausanne, CH)

One of the main goals underlying the development of probabilistic numerical methods is that they help to quantify the uncertainty due to approximate computations. In order for this purpose to be fulfilled, probabilistic methods should yield consistent global or local information on the error, or in other words they should be well calibrated. In this session, I will try to define what a well-calibrated probabilistic numerical methods should be, in particular by highlighting the connections with traditional a posteriori error estimators.

Participants

- Simon Bartels
University of Copenhagen – Copenhagen, DK
- Nathanael Bosch
Universität Tübingen, DE
- François-Xavier Briol
University College London, GB
- Maurizio Filippone
EURECOM – Biot, FR
- Giacomo Garegnani
EPFL – Lausanne, CH
- Roman Garnett
Washington University – St. Louis, US
- Alexandra Gessner
Universität Tübingen, DE
- Philipp Hennig
Universität Tübingen, DE
- Toni Karvonen
University of Helsinki, FI
- Peter Nicholas Krämer
Universität Tübingen, DE
- Maren Mahsereci
Universität Tübingen, DE
- Katharina Ott
Bosch Center for AI – Renningen, DE
- Marvin Pförtner
Universität Tübingen, DE
- Jonathan Schmidt
Universität Tübingen, DE
- Tomas Teren
TU Dresden, DE
- Filip Tronarp
Universität Tübingen, DE
- Jonathan Wenger
Universität Tübingen, DE
- Stephen Wright
University of Wisconsin – Madison, US

Remote Participants

- Oksana Chkrebtii
Ohio State University – Columbus, US
- Jon Cockayne
University of Southampton, GB
- Yuhan Ding
Illinois Institute of Technology – Chicago, US
- Matthew Fisher
Newcastle University, GB
- Fred J. Hickernell
Illinois Institute of Technology – Chicago, US
- Nick Higham
Manchester University, GB
- Ilse C.F. Ipsen
North Carolina State University – Raleigh, US
- Motonobu Kanagawa
EURECOM – Biot, FR
- Hans Kersting
INRIA – Paris, FR
- Tadashi Matsumoto
University of Warwick – Coventry, GB
- Michael McKerns
Los Alamos National Laboratory – New Mexico, US
- Masha Naslidnyk
Amazon Research Cambridge, GB
- Chris Oates
Newcastle University, GB
- Michael A. Osborne
University of Oxford„ GB
- Houman Owhadi
California Institute of Technology – Pasadena, US
- Andrei Paleyes
University of Cambridge, GB
- Kamran Pentland
University of Warwick – Coventry, GB

- Geoff Pleiss
Columbia University – New York, US
- Jagadeeswaran Rathinavel
Illinois Institute of Technology – Chicago, US
- Timothy Reid
North Carolina State University – Raleigh, US
- Simo Särkkä
Aalto University – Helsinki, FI
- Florian Schäfer
Georgia Institute of Technology – Atlanta, US
- Aleksei Sorokin
Illinois Institute of Technology – Chicago, US
- Tim Sullivan
University of Warwick – Coventry, GB
- Aretha Teckentrup
University of Edinburgh, GB
- Onur Teymur
University of Kent – Canterbury, GB
- Zi Wang
Google – Cambridge, US