DAGSTUHL
REPORTS

Volume 12, Issue 4, April 2022

ISSN 2192-5283

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

Publication date
November, 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 `https://dnb.d-nb.de`.

Digital Object Identifier: 10.4230/DagRep.12.4.i

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

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

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

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

Report from Dagstuhl Seminar 22141

Symmetric Cryptography

Nils Gregor Leander[*1], Bart Mennink[*2], María Naya-Plasencia[*3], Yu Sasaki[*4], and Eran Lambooij[†5]

1 **Ruhr-Universität Bochum, DE.** gregor.leander@rub.de
2 **Radboud University Nijmegen, NL.** b.mennink@cs.ru.nl
3 **INRIA – Paris, FR.** maria.naya_plasencia@inria.fr
4 **NTT – Tokyo, JP.** yu.sasaki.sk@hco.ntt.co.jp
5 **University of Haifa, IL.** eranlambooij@gmail.com

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 20041 "Symmetric Cryptography". The seminar was held on April 3-8, 2022 in Schloss Dagstuhl – Leibniz Center for Informatics. This was the eigth seminar in the series "Symmetric Cryptography". Previous editions were held in 2007, 2009, 2012, 2014, 2016, 2018, and 2022. Participants of the seminar presented their ongoing work and new results on topics of (quantum) cryptanalysis and provable security of symmetric cryptographic primitives. In this report, a brief summary of the seminar is given followed by the abstracts of given talks.

Seminar April 3–8, 2022 – http://www.dagstuhl.de/22141
2012 ACM Subject Classification Security and privacy → Cryptanalysis and other attacks; Security and privacy → Symmetric cryptography and hash functions
Keywords and phrases block ciphers, cryptography, hash functions, stream cipers, symmetric cryptography
Digital Object Identifier 10.4230/DagRep.12.4.1

1 Executive Summary

Nils Gregor Leander (Ruhr-Universität Bochum, DE)
Bart Mennink (Radboud University Nijmegen, NL)
Maria Naya-Plasencia (INRIA – Paris, FR)
and Yu Sasaki (NTT – Tokyo, JP)

IT Security plays an increasingly crucial role in everyday life and business. Virtually all modern security solutions are based on cryptographic primitives. Symmetric cryptography deals with the case where both the sender and the receiver of a message use the same key. Due to their good performance, symmetric cryptosystems are highly relevant not only for academia, but also for industrial activities.

We identified the following areas as some of the most important topics on symmetric cryptography at the moment.

Lessons Learnt from NIST Lightweight Cryptography Project. The US National Institute of Standards and Technology (NIST) acknowledged in 2013 the real-world importance of lightweight cryptography, and announced an initiative for standardization. It is expected that the new lightweight standard will not only be used in the US, but rather worldwide.

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

New Design Strategies. This area deals with the development of symmetric cryptographic primitives and modes that must operate for specific applications, such as STARKs, SNARKs, fully homomorphic encryption, and multi-party computation. These novel applications lead to a paradigm shift in design criteria that we are just starting to understand, both in terms of possible optimizations as well as security impacts.

Quantum-Safe Symmetric Cryptography. For symmetric cryptography, it is short-sighted to expect that cryptanalysis will not improve with the help of quantum computers in the future. It is of importance to understand both the possibility to quantize existing classical attacks, as well as the possibility to perform new types of cryptanalytic attacks using a quantum computer.

Understanding Security Implications from Ideal and Keyless Primitives. Permutation-based cryptography has gained astounding popularity in the last decade, and security proofs are performed in an ideal permutation model. Partly as a consequence of this, the concrete security analysis of the involved primitives has become more difficult. One challenge is to understand (i) to what extent distinguishers impact the security of cryptographic schemes and (ii) what non-random properties of permutations seem likely to be translated into an attack on the full scheme.

Seminar Program

The seminar program consisted of a few short presentations and in-depth group meetings. Presentations were about the above topics and other relevant areas of symmetric cryptography, including state-of-the-art cryptanalytic techniques and new designs. Below one can find the list of abstracts for talks given during the seminar.

The research groups were on various topics in symmetric cryptography, all related to one of the above points in one way or another. On the last day of the seminar, the leaders of each group gave brief summaries of achievements. Some teams continued working on the topic after the seminar and started new research collaborations. Here are the summaries of the five groups:

- Group 1 worked and discussed on various problems of provable security, roughly corresponding to one project per person. For three of the projects, the groups had preliminary discussions, and the next step will be to perform the remaining research and investigate the details offline. For two problems, namely improved unforgeability of certain MAC constructions and generic analysis of PRF's and MAC's on 2 public permutations, they advanced quite well and the details will be written down soon after the seminar.
- Group 2 worked on several topics that they plan to continue after the seminar. One was to find good algorithms for detecting the optimal trees of some Boolean functions in the context of improved key-recovery attacks, and figuring out if we actually need trees, or if we could find or use better partitions that do not correspond to a tree and yet improve the complexity. They also worked on building two attacks on the HALFLOOP construction. They solved the problem of finding structures in linear layers and of decomposing them, and they applied this to Streebog. They also continued developing a new cryptanalysis family; differential meet-in-the-middle attacks. They figured out how to correctly combine it with bicliques, and started working on an application on the construction of SKINNY, which should be comparable if not better than the best known attacks.

- Group 3 worked on several topics related to cryptanalysis, that they plan to continue after the seminar. The studied Tweakable Twine, a tweakable variant of Twine proposed in 2019. They looked at impossible differential distinguishers, but unfortunately they were not able to cover more rounds than in previous work. They also looked at the differential propagation of the cipher. They were able to find a distinguisher that would be established with a probability of 2^{-61}, and they rediscovered a 24-round zero correlation attack in Twine. They have also pointed out several observations on TinyJAMBU, including a method to break the P_b permutation (for 384 rounds) if one can observe collisions during the processing phase. They looked at a paper from 2016 on KATAN that searches for extended boomerang distinguishers. They are implementing the attacks to observe the impact of the middle-round dependencies experimentally. Finally, they looked at (free-start) collisions on Romulus-H and tried to find differential characteristics that are suitable to be used in two SKINNY invocations. One idea would be to use the dependencies to have a collision of a higher probability.
- Group 4 has worked on integral distinguishers on big finite fields. After looking at different topics, this group focused in the following problem: can we find integral distinguishers from the knowledge of some properties of the univariate representation of a function $F : \mathbb{F}_{2^n} \to \mathbb{F}_{2^n}$? In other words, they wanted to find some coefficients $(\lambda_0, \ldots, \lambda_{2^n-1})$ in $\mathbb{F}_{2^n}$ such that $\forall x, \sum_{i=0}^{2^n-1} \lambda_i F(\alpha^i X) = 0$. In the particular case where all $\lambda_i \in \mathbb{F}_{2^n}$, this corresponds to finding sets of inputs such that the corresponding outputs sum to zero. They proved that $\sum_{i=0}^{2^n-1} \lambda_i F(\alpha^i X)$ does not contain any term of degree ℓ if and only if $A_\ell = 0$ or $P(\alpha^\ell) = 0$ where $F(X) = \sum_{i=0}^{2^n-1} A_i X^i = 0$. Therefore, they aimed at finding polynomials P which vanish on all α^ℓ when i varies in a given set, and which have the smallest possible number of terms. Indeed, the number of terms of P is the data complexity of the distinguisher. When the only information we have on F is that $A_i = 0$ for all i of weight $\geq d$, then the polynomial P with binary coefficients and with the smallest weight corresponds to the usual distinguisher obtained with higher-order differentials, i.e., $rot(P) = 2^d$. However, if we have more information on A_i, then we can obtain distinguishers with lower data complexity than expected.
- Group 5 looked at a few different topics, quite unrelated to each other. One of them was how to sample binary words of fixed weight (say 200) and length (say 40000) efficiently and in "cryptographic constant time". A possible approach is to use format-preserving encryption, but this turns out to be quite slow compared to alternatives. They eventually slightly revisited an existing method that oversamples w' indices uniformly and independently such that at least w of them are unique with high probability, by proposing a possibly novel and simple constant-time algorithm to extract such a subset of w indices uniformly: write a list (v_i, i) of the w' samples; sort with respect to v_i, mark any duplicate by setting i to infinity; sort with respect to i and keep the w first entries. Another topic was the study of the exact differential probability of 1/4 round of Salsa, or rather computing exactly the probability of any 1/4 round differential. A "promising" approach would be to use finite automata to parameterize the space of solutions to part of a round, and then iteratively propagate this through the successive steps thereof. They have not implemented this, but one could in principle at least partially rely on some existing tools for the first part. Whether the parameterization would be sufficiently compact to also allow an efficient propagation is not clear yet.

2 Table of Contents

3 Overview of Talks

3.1 New Directions in Cryptanalysis

Orr Dunkelman (University of Haifa, IL)

Joint work of Orr Dunkelman, Itai Dinur, Nathan Keller, Eyal Ronen, Adi Shamir

A central problem in cryptanalysis is to find all the significant deviations from randomness in a given n-bit cryptographic primitive. When n is small (e.g., an 8-bit S-box), this is easy to do, but for large n, the only practical way to find such statistical properties was to exploit the internal structure of the primitive and to speed up the search with a variety of heuristic rules of thumb. However, such bottom-up techniques can miss many properties, especially in cryptosystems which are designed to have hidden trapdoors.

In this paper we consider the top-down version of the problem in which the cryptographic primitive is given as a structureless black box, and reduce the complexity of the best known techniques for finding all its significant differential and linear properties by a large factor of $2^{n/2}$. Our main new tool is the idea of using *surrogate differentiation*. In the context of finding differential properties, it enables us to simultaneously find information about all the differentials of the form $f(x) \oplus f(x \oplus \alpha)$ in all possible directions α by differentiating f in a single arbitrarily chosen direction γ (which is unrelated to the α's). In the context of finding linear properties, surrogate differentiation can be combined in a highly effective way with the Fast Fourier Transform. For 64-bit cryptographic primitives, this technique makes it possible to automatically find in about 2^{64} time all their differentials with probability $p \geq 2^{-32}$ and all their linear approximations with bias $|p| \geq 2^{-16}$; previous algorithms for these problems required at least 2^{96} time. Similar techniques can be used to significantly improve the best known time complexities of finding related key differentials, second-order differentials, and boomerangsa In addition, we show how to run variants of these algorithms which require no memory, and how to detect such statistical properties even in trapdoored cryptosystems whose designers specifically try to evade our techniques.

3.2 Review of the NIST Modes of Operation: Status Update and Standardization of a New Mode?

Nicky Mouha (NIST – Gaithersburg, US)

The Crypto Publication Review Board was established by NIST to identify cryptography standards and other publications to be reviewed. Currently, the NIST-recommended modes of operation (NIST SP 800-38 Series) are undergoing review.

At this time of writing, the Crypto Publication Review Project website (`https://csrc.nist.gov/Projects/crypto-publication-review-project`) lists the following modes of operation as subject to review: SP 800-38A (ECB, CBC, CFB, OFB, CTR), SP 800-38A Addendum (three ciphertext stealing variants for CBC), SP 800-38D (GCM and GMAC), and SP 800-38E (XTS).

In this presentation, we gave a technical overview of the NIST-recommended modes of operation, giving insights into the functionality of the algorithms, and an overview of the public comments received.

Less than two weeks before the presentation, NIST had made an announcement related to the review of these modes of operation. This gave an opportunity to provide a status update, and to collect feedback for NIST from the attendees of this talk at the Dagstuhl Symmetric Cryptography Seminar.

3.3 Simplified MITM Modeling for Permutations: New (Quantum) Attacks

André Schrottenloher (CWI – Amsterdam, NL)

Joint work of André Schrottenloher, Marc Stevens
Main reference André Schrottenloher, Marc Stevens: "Simplified MITM Modeling for Permutations: New (Quantum) Attacks", IACR Cryptol. ePrint Arch., p. 189, 2022.
URL https://eprint.iacr.org/2022/189

Meet-in-the-middle (MITM) is a general paradigm where internal states are computed along two independent paths ("forwards" and "backwards") that are then matched. Over time, MITM attacks improved using more refined techniques and exploiting additional freedoms and structure, which makes it more involved to find and optimize such attacks. This has led to the use of detailed attack models for generic solvers to automatically search for improved attacks, notably a MILP model developed by Bao et al. at EUROCRYPT 2021.

In this paper, we study a simpler MILP modeling combining a greatly reduced attack representation as input to the generic solver, together with a theoretical analysis that, for any solution, proves the existence and complexity of a detailed attack. This modeling allows to find both classical and quantum attacks on a broad class of cryptographic permutations. First, Present-like constructions, with the permutations of the Spongent hash functions: we improve the MITM step in distinguishers by up to 3 rounds. Second, AES-like designs: despite being much simpler than Bao et al.'s, our model allows to recover the best previous results. The only limitation is that we do not use degrees of freedom from the key schedule. Third, we show that the model can be extended to target more permutations, like Feistel networks. In this context we give new Guess-and-determine attacks on reduced Simpira v2 and Sparkle. Finally, using our model, we find several new quantum preimage and pseudo-preimage attacks (e.g. Haraka v2, Simpira v2 ...) targeting the same number of rounds as the classical attacks.

3.4 Triplex: an Efficient and One-Pass Leakage-Resistant Mode of Operation

Yaobin Shen (University of Louvain, BE)

Joint work of Thomas Peters, Yaobin Shen, François-Xavier Standaert

This talk introduces and analyzes Triplex, a leakage-resistant mode of operation based on Tweakable Block Ciphers (TBCs) with $2n$-bit tweaks. Triplex enjoys beyond-birthday ciphertext integrity in the presence of encryption and decryption leakage in a liberal model where all intermediate computations are leaked in full and only two TBC calls operating a long-term secret are protected with implementation-level countermeasures. It provides beyond-birthday confidentiality guarantees without leakage, and standard confidentiality guarantees with leakage for a single-pass mode embedding a re-keying process for the bulk of its computations (i.e., birthday confidentiality with encryption leakage under a bounded leakage assumption). Triplex improves leakage-resistant modes of operation relying on TBCs with n-bit tweaks when instantiated with large-tweak TBCs like Deoxys-TBC (a CAESAR competition laureate) or Skinny (used by the Romulus finalist of the NIST lightweight crypto competition). Its security guarantees are maintained in the multi-user setting.

4 Working groups

4.1 Exact Differential Analysis of One Round of Salsa20

Orr Dunkelman (University of Haifa, IL), Antonio Florez-Gutierrez (INRIA – Paris, FR), Pierre Karpman (Université Grenoble Alpes – Saint Martin d'Hères, FR), Eram Lambooij (University of Haifa, IL), and Nicky Mouha (NIST – Gaithersburg, US)

In Salsa20, one round consists of four parallel quarterround functions on independent inputs. These quarterround functions transform a 128–bit by adding two 32–bit inputs (modulo 2^{32}), rotating the output over a fixed amount of bits and XORing it with a third 32–bit input. This operation is performed four times within a quarterround.

It seems to be an open problem to compute the exact differential probability for one quarterround of Salsa20. It has been shown that theoretical estimates of the probability may not correspond to estimates obtained by experimental verification [1, 2].

The goal of this research group was to explore some methods to calculate the exact differential probability for one quarterround of Salsa20, and confirm these with experiments on a small-scale variant of Salsa20. Because the four quarterround functions are independent of each other, a method to determine the exact differential probability for one quarterround, would also lead to a result for one round of Salsa20.

We explored the problem from both a theoretical and experimental point of view, and arrived at various new insights that will be helpful to find an elegant and efficient solution to this problem.

References

1 Nicky Mouha and Bart Preneel. *Towards Finding Optimal Differential Characteristics for ARX: Application to Salsa20*. Cryptology ePrint Archive, Report 2013/328, 2013.
2 Nicky Mouha. *On Proving Security against Differential Cryptanalysis*. CFAIL 2019, A Conference for Failed Approaches and Insightful Losses in Cryptology, 2019.

4.2 A simple quasi-linear constant-time algorithm for sampling fixed-sized supports

Pierre Karpman (Université Grenoble Alpes – Saint Martin d'Hères, FR), Orr Dunkelman (University of Haifa, IL), Antonio Florez–Gutierrez (INRIA – Paris, FR), Eram Lambooij (University of Haifa, IL), and Nicky Mouha (NIST – Gaithersburg, US)

In this short note, we present a simple algorithm for uniformly sampling a subset of [[0, N – 1]] of size w, for some integers N and w. The cost of our algorithm is quasi- linear in w, assuming a constant cost for arithmetic and random sampling of integers less than N. It is also amenable to "cryptographic constant-time" implementations, that is whose running time and memory accesses neither depend on the random coins used in the sampling. Such an algorithm and implementation find applications in certain code-based cryptosystems.

4.3 Research group on the cryptanalysis of recent primitives

Virginie Lallemand (LORIA – Nancy, FR), Xavier Bonnetain (LORIA & INRIA Nancy, FR), Maria Eichlseder (TU Graz, AT), Daniël Kuijsters (Radboud University Nijmegen, NL), Clara Pernot (INRIA – Paris, FR), Shahram Rasoolzadeh (Radboud University Nijmegen, NL), Yu Sasaki (NTT – Tokyo, JP), and André Schrottenloher (CWI – Amsterdam, NL)

We worked and discussed several topics related to cryptanalysis, that we plan to continue after the seminar:

Tweakable Twine: We studied this cipher which is a tweakable version of Twine proposed in 2019. We looked at impossible differentials distinguishers but unfortunately were not able to cover more rounds than in the previous work. We also looked at the differential properties of the cipher, and were able to find a $2^{-60.21}$ distinguisher on 17 rounds. We (re-)discovered a 24-round zero-correlation on Twine.

Tiny-Jambu: We have several observations, including a method to break the P_k permutation (for 384 rounds) if we can observe collisions during the AD processing phase.

Katan: We looked at a paper from 2016 that searches extended boomerang distinguishers. We are implementing the attack to observe the impact of the middle-round dependencies experimentally.

Romulus-H: (Skinny-Hirose) We looked at free-start and (real) collisions and tried to find differential characteristics that are suitable to be used in two Skinny invocations in Hirose's mode. One idea would be to use the dependencies in the states to have a collision of higher probability.

4.4 Provable Security Research Group

Bart Mennink (Radboud University Nijmegen, NL), Ritam Bhaumik (INRIA – Paris, FR), Aldo Gunsing (Radboud University Nijmegen, NL), Ashwin Jha (CISPA – Saarbrücken, DE), and Yaobin Shen (University of Louvain, BE)

The aim of the provable security group within Dagstuhl was to analyze generic security of modes, either by proving security under certain assumptions or by mounting generic attacks. The provable security group, consisting of Ritam Bhaumik, Aldo Gunsing, Ashwin Jha, Bart Mennink, and Yaobin Shen, worked on various topics in provable security. We discussed five topics in total, one corresponding to each group member. For three problems, we postponed the continuation until after Dagstuhl: it was required that each group member would read certain relevant papers offline, and only then we could continue solving the problem. For two problems we advanced quite well. The first problem was the unforgeability of a strengthened version of the Wegman-Carter-Shoup authenticator. Although this strengthened version only achieves birthday bound PRF-security, we observed that its provable unforgeability is better, and we drafted the proof ideas. The second problem was about a generic description and analysis of PRFs based on two public permutations, and a generic description and analysis of MAC functions based on two public permutations. We described the generic classification and filtered out the "sets" of functions that achieve high security.

4.5 Workgroup 1

Maria Naya-Plasencia (INRIA – Paris, FR), Christof Beierle (Ruhr-Universität Bochum, DE), Christina Boura (University of Versailles, FR), Patrick Derbez (University of Rennes, FR), Patrick Felke (FH Emden, DE), Nils Gregor Leander (Ruhr-Universität Bochum, DE), and Sondre Rønjom (University of Bergen, NO)

We worked on several topics, that we plan to continue after the seminar.

1. Find good algorithms for detecting the optimal trees of some boolean functions in the context of improved key-recovery attacks. Do we need trees? Could we find/use better cases with partitions that do not correspond to a tree?
2. A new type of attack: Differential MitM. We continue to develop its theoretical complexities, adding this to apply the byclique extension and work on building an application on the block cipher Skinny.
3. We built two attacks on the construction HAL + LOOP.
4. We solved how to find structures in linear layers, how to decompose them, and how to apply it to Streeborg.

4.6 Univariate Integral Distinguishers

Yann Rotella (University of Versailles, FR), Subhadeep Banik (University of Lugano, CH), Clémence Bouvier (INRIA – Paris, FR), Anne Canteaut (INRIA – Paris, FR), Margot Funk (University of Versailles, FR), Daniël Kuijsters (Radboud University Nijmegen, NL), Patrick Neumann (Ruhr–Universität Bochum, DE), Léo Perrin (INRIA – Paris, FR), Christian Rechberger (TU Graz, AT), Markus Schofnegger (TU Graz, AT), and Tyge Tiessen (Technical University of Denmark – Lyngby, DK)

Recent surge in development of advanced cryptographic protocols (such as multi-party computation, zero-knowledge proofs) created interest in specialized symmetric-key cryptographic primitives including block ciphers, stream ciphers, hash functions. The new setting favors *algebraic* constructions based on relatively large finite fields, since it leads to lesser costs in the protocols. This contrasts with classic symmetric-key cryptography, where operations are typically bit-oriented and are optimized for performance on common CPUs.

The new paradigm demands for exploring new cryptanalysis methods. In this work, we focused on adapting the *integral* attacks, which before were typically developed in the *binary multivariate* setting.

Integral attacks on classic symmetric primitives are well understood and state-of-the-art includes many tools both for finding and exploiting integral distinguishers. On the other hand, integral attacks on the algebraic constructions are not yet well studied and do not seem to fully exploit the algebraic properties. Initial work in this direction was made in recent works [1, 2, 3, 5]. In this working group, we aimed to advance this direction by exploring and systemizing methods of *searching for* and *exploiting* integral distinguishers in the *univariate* setting. More precisely, we studied which linear combinations of the outputs of a function are constant, given a set of missing monomials in the function's univariate representation.

The working group achieved several interesting results.

1. We briefly studied methods of bounding the degree in large fields and attempted to generalize standard methods based on tracking maximum variable degrees or division property [4]. We reached to conclusion that, in large characteristic, naive approaches seem to often provide the exact degree and thus no significant improvements can be done.
2. We developed a method of studying univariate integral distinguishers based on *function operators*, which act predictably on the univariate coefficients. We considered several operators, such as operators reducing the coefficients to their field trace or trace-based filtering of coefficients, operators summing over a multiplicative subgroup.
3. We discovered a simple operator resembling a composition of polynomials, which includes most previously mentioned operators as special cases. It also has interesting mathemathical properties such as commutativity of operators.
4. We studied a few concrete examples, such as: a single monomial missing, a single cyclotomic class missing, a (multivariate) algebraic degree is bounded – and proved optimal distinguishers for these cases.

References

1 Tim Beyne, Anne Canteaut, Itai Dinur, Maria Eichlseder, Gregor Leander, Gaëtan Leurent, María Naya-Plasencia, Léo Perrin, Yu Sasaki, Yosuke Todo, Friedrich Wiemer. *Out of Oddity – New Cryptanalytic Techniques Against Symmetric Primitives Optimized for Integrity Proof Systems*. CRYPTO, 2020.

2 Maria Eichlseder, Lorenzo Grassi, Reinhard Lüftenegger, Morten Øygarden, Christian Rechberger, Markus Schofnegger, Qingju Wang. *An Algebraic Attack on Ciphers with Low-Degree Round Functions: Application to Full MiMC*. ASIACRYPT, 2020.

3 Carlos Cid, Lorenzo Grassi, Aldo Gunsing, Reinhard Lüftenegger, Christian Rechberger, Markus Schofnegger. *Influence of the Linear Layer on the Algebraic Degree in SP-Networks*. TosC, 2022.

4 Yosuke Todo. *Structural Evaluation by Generalized Integral Property*. EUROCRYPT, 2015.

5 Clémence Bouvier, Anne Canteaut, Léo Perrin. *On the Algebraic Degree of Iterated Power Functions*. EPRINT, 2022.

Participants

- Subhadeep Banik
University of Lugano, CH
- Christof Beierle
Ruhr-Universität Bochum, DE
- Ritam Bhaumik
INRIA – Paris, FR
- Xavier Bonnetain
LORIA & INRIA Nancy, FR
- Christina Boura
University of Versailles, FR
- Clémence Bouvier
INRIA – Paris, FR
- Anne Canteaut
INRIA – Paris, FR
- Patrick Derbez
University of Rennes, FR
- Orr Dunkelman
University of Haifa, IL
- Maria Eichlseder
TU Graz, AT
- Patrick Felke
FH Emden, DE
- Antonio Florez-Gutierrez
INRIA – Paris, FR
- Margot Funk
University of Versailles, FR
- Aldo Gunsing
Radboud University
Nijmegen, NL
- Ashwin Jha
CISPA – Saarbrücken, DE
- Pierre Karpman
Université Grenoble Alpes –
Saint Martin d'Hères, FR
- Daniël Kuijsters
Radboud University
Nijmegen, NL
- Virginie Lallemand
LORIA – Nancy, FR
- Eran Lambooij
University of Haifa, IL
- Nils Gregor Leander
Ruhr-Universität Bochum, DE
- Bart Mennink
Radboud University
Nijmegen, NL
- Nicky Mouha
NIST – Gaithersburg, US
- Maria Naya-Plasencia
INRIA – Paris, FR
- Patrick Neumann
Ruhr-Universität Bochum, DE
- Clara Pernot
INRIA – Paris, FR
- Léo Perrin
INRIA – Paris, FR
- Shahram Rasoolzadeh
Radboud University
Nijmegen, NL
- Christian Rechberger
TU Graz, AT
- Yann Rotella
University of Versailles, FR
- Sondre Rønjom
University of Bergen, NO
- Yu Sasaki
NTT – Tokyo, JP
- Markus Schofnegger
TU Graz, AT
- André Schrottenloher
CWI – Amsterdam, NL
- Yaobin Shen
University of Louvain, BE
- Tyge Tiessen
Technical University of Denmark
– Lyngby, DK
- Aleksei Udovenko
University of Luxembourg, LU

Report from Dagstuhl Seminar 22161

Recent Advancements in Tractable Probabilistic Inference

Priyank Jaini*[1], Kristian Kersting*[2], Antonio Vergari*[3], and Max Welling*[4]

1 **Google – Toronto, CA.** jaini.priyank1@gmail.com
2 **TU Darmstadt, DE.** kersting@cs.tu-darmstadt.de
3 **University of Edinburgh, GB.** aver@cs.ucla.edu
4 **University of Amsterdam, NL.** welling.max@gmail.com

Abstract

In several real-world scenarios, decision making involves advanced reasoning under uncertainty, i.e. the ability to answer probabilistic queries. Typically, it is necessary to compute these answers in a *limited amount of time*. Moreover, in many domains, such as healthcare and economical decision making, it is crucial that the result of these queries is reliable, i.e. either *exact* or comes with approximation guarantees. In all these scenarios, tractable probabilistic inference and learning are becoming increasingly important.

Research on representations and learning algorithms for tractable inference embraces very different fields, each one contributing its own perspective. These include automated reasoning, probabilistic modeling, statistical and Bayesian inference and deep learning.

Among the many recent emerging venues in these fields there are: tractable neural density estimators such as autoregressive models and normalizing flows; deep tractable probabilistic circuits such as sum-product networks and sentential decision diagrams; approximate inference routines with guarantees on the quality of the approximation.

Each of these model classes occupies a particular spot in the continuum between tractability and expressiveness. That is, different model classes might offer appealing advantages in terms of efficiency or representation capabilities while trading-off other of these aspects.

So far, clear connections and a deeper understanding of the key differences among them have been hindered by the different languages and perspectives adopted by the different "souls" that comprise the tractable probabilistic modeling community.

This Dagstuhl Seminar brought together experts from these sub-communities and provided the perfect venue to exchange perspectives, deeply discuss the recent advancements and build strong bridges that can greatly propel interdisciplinary research.

Seminar April 18–22, 2022 – http://www.dagstuhl.de/22161
2012 ACM Subject Classification Computing methodologies → Artificial intelligence; Computing methodologies → Machine learning
Keywords and phrases approximate inference with guarantees, deep generative models, probabilistic circuits, Tractable inference
Digital Object Identifier 10.4230/DagRep.12.4.13

* Editor / Organizer

Recent Advancements in Tractable Probabilistic Inference, *Dagstuhl Reports*, Vol. 12, Issue 4, pp. 13–25
Editors: Priyank Jaini, Kristian Kersting, Antonio Vergari, and Max Welling

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

1 Executive Summary

Priyank Jaini (Google – Toronto, CA)
Kristian Kersting (TU Darmstadt, DE)
Antonio Vergari (University of Edinburgh, GB)
Max Welling (University of Amsterdam, NL)

ML models and systems to enable and support decision making in real-world scenarios need to robustly and effectively *reason in the presence of uncertainty* over the configurations of the world that can be observed. *Probabilistic inference* provides a principled framework to carry on this reasoning process, and enables probabilistic *modeling*: a collection of principles to design and learn from data models that are capable of dealing with uncertainty. The main purpose for these models, once learned or built, is to answer *queries* – posed by humans or other autonomous systems – concerning some aspects of the represented world and quantifying some form of uncertainty over it. That is, that is computing some quantity of interest of the probability distribution that generated the observed data. For instance, the mean or the modes of such a distribution, the marginal or conditional probabilities of events, expected utilities of our policies, or decoding most likely assignments to variables (also known as MAP inference, cf. the Viterbi algorithm). Answering these queries reliably and efficiently is more important than ever: we need ML models and systems to perform inference based on well-calibrated uncertainty estimates throughout all reasoning steps, especially when informing and supporting humans in decision making processes in the real world.

For instance, consider a ML system learned from clinical data to support physicians and policy makers. Such a system would need to support *arbitrary queries* posed by physicians, that is, questions that are not known a priori. Moreover, these queries might involve complex probabilistic reasoning over possible states of the world, for instance involving maximization of some probabilities and the ability to marginalize over unseen or not available (missing) attributes like "At what age is a patient with this X-ray but no previous health record most likely to show *any* symptom of COVID-19?", or counting and comparing sub-populations "What is the probability of there being more cases with fever given a BMI of 25 in this county than in the neighboring one?". At the same time, it should guarantee that the uncertainty in its answers, modeled as probabilities, should be faithful to the real-world distribution as uncalibrated estimates might greatly mislead the decision maker.

Recent successes in machine learning (ML) and particularly deep learning have delivered very expressive probabilistic models and learning algorithms. These have proven to be able to induce exceedingly richer models from larger datasets but, unfortunately, at an incredible cost: these models are vastly *intractable* for all but the most trivial of probabilistic reasoning tasks, and they have been demonstrated to provide unreliable uncertainty estimations. In summary, their applicability to real-world scenarios, like the one just described, is very limited.

Nevertheless all these required "ingredients" are within the grasp of several models which we group together under the umbrella name of *tractable probabilistic models*, the core interest of this seminar. *Tractability* here guarantees answering queries *efficiently* and *exactly*. Tractable probabilistic models (TPMs) have a long history rooted in several research fields such as classical probabilistic graphical models (low-treewidth and latent variable models), automated reasoning via knowledge compilation (logical and arithmetic circuits) and statistics (mixture models, Kalman filters). While these classical TPMs are known to

be limited in expressiveness, several recent advancements in deep tractable models (sum-product networks, probabilistic sentential decision diagrams, normalizing flows and neural autoregressive models) are inverting the trend and promising tractable probabilistic inference with little or no compromise when compared to the deep generative models discussed above. It becames then more and more important to have a seminar on these recent successes of TPMs bringing together the different communities of tractable probabilistic modeling at the same table to propel collaborations by defining the goals and the agenda for future research.

These are the major topics around which the seminar brought up the aforementioned discussion:

- Advanced probabilistic query classes
- Deep tractable probabilistic modeling
- Robust and verifiable probabilistic inference
- Exploiting symmetries for probabilistic modelling and applications in science.

Advanced probabilistic query classes

Probabilistic inference can be reduced as computing probabilistic queries, i.e., functions whose output are certain properties of a probability distribution (e.g., its mass, density, mean, mode, etc.) as encoded by a probabilistic model. Probabilistic queries can be grouped into classes when they compute the same distributional properties and hence share the same computational effort to be answered. Among the most commonly used query classes there are complete evidence (EVI), marginals (MAR), conditionals (CON) and maximum a posteriori (MAP) inference. While these classes have been extensively investigated in theory and practice, they constitute a small portion of the probabilistic inference that might be required to support complex decision making in the real-world.

In fact, one might want to compute the probabilities of logical and arithmetic constraints, of structured objects such as rankings, comparing the likelihood and counts of groups of events or computing the expected predictions of discriminative model such as a classifier or regression w.r.t. some feature distribution. Tracing the exact boundaries of tractable probabilistic inference for these advanced probabilistic query classes and devising probabilistic models delivering efficient and reliable inference for them is an open challenge.

Deep tractable probabilistic modeling

A probabilistic model falls under the umbrella name of tractable probabilistic models (TPMs) if it guarantees exact and polytime inference for certain query classes. As different model classes can be tractable representations for different query classes, a *spectrum* of tractable inference emerges. Typically, this create a tension with the extent of a model class supporting a larger set of tractable query classes, and its *expressive efficiency*, i.e., the set of functions it can represent compactly.

Recent deep generative models such as generative adversarial networks (GANs), regularized and variational autoencoders (VAEs) fall out of the TPM umbrella because they either have no explicit likelihood model or computing even the simplest class of queries, EVI, is hard in general. In fact, despite their successes, their inference capabilities are severely limited and one has to recur to approximations. However, the approximate inference routines available so far (such as the evidence lower bound and its variants) do not provide sufficiently strong guarantees on the quality of the approximation delivered to be safely deployed in real-world scenarios.

On the other hand, classical TPMs from the probabilistic graphical model community support larger classes of tractable queries comprising MAR, CON and MAP (to different extents based on the model class). Among these there are: i) low or bounded-treewidth probabilistic graphical models that exchange expressiveness for efficiency; ii) determinantal point processes which allow tractable inference for distributions over sets; iii) graphical models with high girth or weak potentials, that provide bounds on the performance of approximate inference methods; and iv) exchangeable probabilistic models that exploit symmetries to reduce inference complexity.

A different prospective on tractability is brought by models compiling inference routines into efficient computational graphs such as arithmetic circuits, sum-product networks, cutset networks and probabilistic sentential decision diagrams have advanced the state-of-the-art inference performance by exploiting context-specific independence, determinism or by exploiting latent variables. These TPMs, as well as many classical tractable PGMs as listed above, can be cast under a unifying framework of probabilistic circuits (PCs), abstracting from the different graphical formalisms of each model. PCs with certain structural properties support tractable MAR, CON, MAP as well as some of the advanced query classes touched in the previous topic item. Guy Van den Broeck gave a long talk on the first day of the seminar to set the stage for participants for viewing tractable probabilistic models from the lens of probabilistic circuits.

More recently, the field of neural density estimators has gained momentum in the tractable probabilistic modeling community. This is due to model classes such as normalizing flows and autoregressive models. Autoregressive models and flows retain the expressiveness of GANs and VAEs, by levering powerful neural representations for probability factors or invertible transformations, while overcoming their limitations and delivering tractable EVI queries. As such, they position themselves in the spectrum of tractability in an antithetic position w.r.t. PCs: while the latter support more tractable query classes, the former are generally more expressive. On the first day of the seminar, Marcus Brubaker introduced these models to the seminar participants in a long talk. It is an interesting open challenge to combine TPM models from different regions of such a spectrum to leverage the "best of different worlds", i.e., increase a model class expressive efficiency while retaining the largest set of supported tractable query classes as possible. The first day subsequently ended with a lively open discussion on the differences between TPMs and Neural Generative Models and what advantages and lessons they can provide the other models.

Robust and verifiable probabilistic inference

Along exactness and efficiency, one generally requires inference routines to be robust to adversarial conditions (noise, malicious attacks, etc.) and to be allow exactness and efficiency to be formally provable. This is crucial to deploy reliable probabilistic models in real-world scenarios (cf. other topic). Recent advancements in learning tractable and intractable probabilistic models from data have raised the question if the learned models are just exploiting spurious correlations in input space, thus ultimately delivering an unfaithful image of the probability distribution they try to encode. This raises several issues, as in tasks like anomaly detection and model comparison, which rely on correctly calibrated probabilities, one can be highly mislead by such unfaithful probabilistic models. Furthermore, one might want to verify a priori or ex-post (e.g., in presence of adversarial interventions) if one probabilistic inference algorithm truly guarantees exact inference. Questions like this have just very recently been tackled in a formal verification setting, where proofs of the correctness of inference can be verified with less resources than it takes to execute inference.

Over the course of the seminar, through informal discussions and formal talks by the participants discussed the above mentioned issues in tractable probabilistic inference through topics such as Bayesian Deep Learning, Incorporating symmetries in probabilistic modelling using equivariance with applications in sciences, explainable AI etc.

Overall, the seminar produced numerous insights into how efficient, expressive, flexible, and robust tractable probabilistic models can be built. Specially, the discussions and talks at the seminar spurred a renewed interest in the community to:

- develop techniques and approaches that bring together key ideas from several different fields that include deep generative models, probabilistic circuits, knowledge compilation, and approximate inference.
- create bridges between researchers in these different fields and identify ways in which enhanced interaction between the communities can continue.
- generate a set of goals, research directions, and challenges for researchers in these field to develop robust and principled probabilistic models.
- provide a unified view of the current undertakings in these different fields towards probabilistic modelling and identifying ways to incorporate ideas from several fields together.
- develop a new systematic and unified set of development tools encompassing these different areas of probabilistic modelling.

2 Table of Contents

3 Overview of Talks

3.1 Causality and Tractable Probabilistic Models

Alessandro Antonucci (IDSIA – Manno, CH)

Probabilistic sentential decision diagrams (PSDDs) are a popular class of probabilistic circuits intended to implement generative models consistent with a propositional knowledge base. We discuss a number of results related to these models. This includes: the sensitivity analysis of the inferences with respect to perturbations in the local probabilistic parameters of the circuit; a structural learning algorithm for these models based on a relaxation of the closed-world assumption for the training data; and a discussion on the benefits and the challenges related to the embedding of knowledge bases in ML tasks.

3.2 A tutorial on Normalizing Flows

Marcus A. Brubaker (York University – Toronto, CA)

Normalizing flows (NFs) offer an answer to a long-standing question in computer vision: How can one define faithful probabilistic models for complex high-dimensional data like natural images? NFs solve this problem by means of non-linear bijective mappings from simple distributions (e.g. multivariate normal) to the desired target distributions. These mappings are implemented with invertible neural networks and thus have high expressive power and can be trained by gradient descent in the usual way. Thanks to bijectivity, NFs can work forward and backward, serving as both discriminative and generative models alike, and are especially suitable for inverse problems. This tutorial will explain the theoretical underpinnings of NFs, show various practical implementation options, clarify their relationships with GANs, VAEs, and non-linear ICA. Particular emphasis will be given to successful applications in the field of computer vision.

3.3 Solving Marginal MAP Exactly by Probabilistic Circuit Transformations

YooJung Choi (UCLA, US)

Joint work of YooJung Choi, Antonio Vergari, Guy Van den Broeck

Probabilistic circuits (PCs) are a class of tractable probabilistic models that allow efficient, often linear-time, inference of queries such as marginals and most probable explanations (MPE). However, marginal MAP, which is central to many decision-making problems, remains a hard query for PCs unless they satisfy highly restrictive structural constraints. In this paper, we develop a pruning algorithm that removes parts of the PC that are irrelevant to a marginal MAP query, shrinking the PC while maintaining the correct solution. This pruning

technique is so effective that we are able to build a marginal MAP solver based solely on iteratively transforming the circuit–no search is required. We empirically demonstrate the efficacy of our approach on real-world datasets.

3.4 Towards Robust Classification with Deep Generative Forests

Cassio de Campos (TU Eindhoven, NL)

Joint work of Alvaro H. C. Correia, Robert Peharz, Cassio de Campos
Main reference Alvaro H. C. Correia, Robert Peharz, Cassio P. de Campos: "Towards Robust Classification with Deep Generative Forests", CoRR, Vol. abs/2007.05721, 2020.
URL https://arxiv.org/abs/2007.05721

Decision Trees (DTs) and Random Forests (RFs) are powerful discriminative learners and tools of central importance to the everyday machine learning practitioner and data scientist. Due to their discriminative nature, however, they lack principled methods to process inputs with missing features or to detect outliers, which requires pairing them with imputation techniques or a separate generative model. In this paper, we demonstrate that DTs and RFs can naturally be interpreted as generative models, by drawing a connection to Probabilistic Circuits, a prominent class of tractable probabilistic models. This reinterpretation equips them with a full joint distribution over the feature space and leads to Generative Decision Trees (GeDTs) and Generative Forests (GeFs), a family of novel hybrid generative-discriminative models. This family of models retains the overall characteristics of DTs and RFs while additionally being able to handle missing features by means of marginalisation. Under certain assumptions, frequently made for Bayes consistency results, we show that consistency in GeDTs and GeFs extend to any pattern of missing input features, if missing at random. Empirically, we show that our models often outperform common routines to treat missing data, such as K-nearest neighbour imputation, and moreover, that our models can naturally detect outliers by monitoring the marginal probability of input features.

3.5 Exploiting Symmetries for Probabilistic Generative Modelling

Priyank Jaini (Google – Toronto, CA)

Joint work of Priyank Jaini, Lars Holdijk, Max Welling
Main reference Priyank Jaini, Lars Holdijk, Max Welling: "Learning Equivariant Energy Based Models with Equivariant Stein Variational Gradient Descent", CoRR, Vol. abs/2106.07832, 2021.
URL https://arxiv.org/abs/2106.07832

Symmetries play a crucial role in Physics and Mathematics. In this talk, I will explore generative models for efficient sampling and inference by incorporating inductive biases in the form of symmetries. I will begin by introducing Equivariant Stein Variational Gradient Descent (SVGD) algorithm — an equivariant sampling method based on Stein's identity for sampling from symmetric distributions. Subsequently, I will discuss training equivariant energy based models using Equivariant-SVGD to model invariant probability distributions with applications in many-body particle systems and molecular structure generation.

3.6 Equivariant Probabilistic Models for Physics

Danilo Jimenez Rezende (Google DeepMind – London, GB)

The study of symmetries in physics has revolutionized our understanding of the world. Inspired by this, the development of methods to incorporate internal (Gauge) and external (space-time) symmetries into machine learning models is a very active field of research. We will present our work on invariant generative models and its applications to lattice-QCD and molecular dynamics simulations. In the molecular dynamics front, we'll talk about how we constructed permutation and translation-invariant normalizing flows on a torus for free-energy estimation. In lattice-QCD, we'll present our work that introduced the first U(N) and SU(N) Gauge-equivariant normalizing flows for pure Gauge simulations and its extensions to incorporate fermions.

3.7 Predictive Complexity Priors

Eric Nalisnick (University of Amsterdam, NL)

Specifying a Bayesian prior is notoriously difficult for complex models such as neural networks. Reasoning about parameters is made challenging by the high-dimensionality and over-parameterization of the space. Priors that seem benign and uninformative can have unintuitive and detrimental effects on a model's predictions. To help cope with these problems, I will describe our work on predictive complexity priors: a prior that is defined by comparing the model's predictions to those of a reference model.

3.8 Extracting context specific independencies from sum product networks

Sriraam Natarajan (University of Texas – Dallas, US)

Joint work of Sriraam Natarajan, Athresh Karanam, Saurabh Sanjay Mathur, Predrag Radivojac, Kristian Kersting

I present the problem of explaining a class of tractable deep probabilistic model, the Sum-Product Networks (SPNs). First, I motivate how knowledge as qualitative constraints could be extracted from SPNs and then present an algorithm EXSPN to generate explanations. To this effect, I define the notion of a context-specific independence tree(CSI-tree) and present an iterative algorithm that converts an SPN to a CSI-tree. The resulting CSI-tree is both interpretable and explainable to the domain expert. We achieve this by extracting the conditional independencies encoded by the SPN and approximating the local context specified by the structure of the SPN. Our extensive empirical evaluations on synthetic, standard, and real-world clinical data sets demonstrate that the resulting models exhibit superior explainability.

3.9 Implicit MLE: Backpropagating Through Discrete Exponential Family Distributions

Mathias Niepert (Universität Stuttgart, DE)

Joint work of Mathias Niepert, Pasquale Minervini, Luca Franceschi
Main reference Mathias Niepert, Pasquale Minervini, Luca Franceschi: "Implicit MLE: Backpropagating Through Discrete Exponential Family Distributions", CoRR, Vol. abs/2106.01798, 2021.
URL https://arxiv.org/abs/2106.01798

Combining discrete probability distributions and combinatorial optimization problems with neural network components has numerous applications in learning and reasoning but poses several challenges. We propose Implicit Maximum Likelihood Estimation (I-MLE), a framework for end-to-end learning of models combining discrete exponential family distributions and differentiable neural components. I-MLE is widely applicable as it only requires the ability to compute the most probable states and does not rely on smooth relaxations. The framework encompasses several approaches such as perturbation-based implicit differentiation and recent methods to differentiate through black-box combinatorial solvers. We introduce a novel class of noise distributions for approximating marginals via perturb-and-MAP. Moreover, we show that I-MLE simplifies to maximum likelihood estimation when used in some recently studied learning settings that involve combinatorial solvers. Experiments on several datasets suggest that I-MLE is competitive with and often outperforms existing approaches which rely on problem-specific relaxations. Lastly we discuss potential connections with more sophisticated reasoning scenarios with tractable models.

3.10 Rapid Adaptation in Robot Learning

Deepak Pathak (Carnegie Mellon University – Pittsburgh, US)

Generalization, i.e., the ability to adapt to novel scenarios, is the hallmark of human intelligence. While we have systems that excel at cleaning floors, playing complex games, and occasionally beating humans, they are incredibly specific in that they only perform the tasks they are trained for and are miserable at generalization. One of the fundamental reasons is that, unlike humans, most of these artificial agents start tabula-rasa without any prior knowledge and learn only towards a fixed goal. Could actually optimizing towards fixed external goals be hindering the generalization instead of aiding it? In this talk, I will present our initial efforts toward endowing artificial agents with an ability to generalize in diverse scenarios. The main insight is to first allow the agent to learn general-purpose skills in a completely self-directed manner, without optimizing for any external goal. These skills are then later repurposed to perform complex tasks. I will discuss how this framework can be instantiated to develop curiosity-driven agents (virtual as well as real) that can learn to play games, learn to walk, and learn to perform real-world object manipulation without any rewards or supervision. These curious robotic agents, after exploring the environment, can generalize to find their way in office environments, tie knots using rope and rearrange object configuration.

3.11 Exact and Efficient Adversarial Robustness with Decomposable Neural Networks

Robert Peharz (TU Graz, AT)

Joint work of Robert Peharz, Pranav Shankar Subramani, Antonio Vergari, Gautam Kamath
Main reference Pranav Shankar Subramani, Antonio Vergari, Gautam Kamath, Robert Peharz: "Exact and Efficient Adversarial Robustness with Decomposable Neural Networks", in Proc. of the The 4th Workshop on Tractable Probabilistic Modeling, 2021.
URL https://openreview.net/forum?id=5E7V1tCwLq

As deep neural networks are notoriously vulnerable to adversarial attacks, there has been significant interest in defenses with provable guarantees. Recent solutions advocate for a randomized smoothing approach to provide probabilistic guarantees, by estimating the expectation of a network's output when the input is randomly perturbed. As the convergence of the estimated expectations depends on the number of Monte Carlo samples, and hence network evaluations, these techniques come at the price of considerable additional computation at inference time. We take a different route and introduce a novel class of deep models – decomposable neural networks (DecoNets) – which are hierarchical multi-linear functions over non-linear input features. DecoNets can compute the expectation over the outputs in closed form in a *single network evaluation*, thus providing *exact* smoothing guarantees. Our empirical analysis shows the promising nature of DecoNets: they achieve the same or better certified accuracy in comparison to models of equivalent size on benchmark datasets, while providing exact guarantees one or two orders of magnitude faster.

3.12 Probabilistic Circuits: Representations, Inference, Learning and Applications

Guy Van den Broeck (UCLA, US)

Joint work of Antonio Vergari, Guy Van den Broeck
URL https://web.cs.ucla.edu/ guyvdb/talks/IJCAI20-tutorial/

Exact and efficient probabilistic inference and learning are becoming more and more mandatory when we want to quickly take complex decisions in presence of uncertainty in real-world scenarios where approximations are not a viable option. In this tutorial, we will introduce probabilistic circuits (PCs) as a unified computational framework to represent and learn deep probabilistic models guaranteeing tractable inference. Differently from other deep neural estimators such as variational autoencoders and normalizing flows, PCs enable large classes of tractable inference with little or no compromise in terms of model expressiveness. Moreover, after showing a unified view to learn PCs from data and several real-world applications, we will cast many popular tractable models in the framework of PCs while leveraging it to theoretically trace the boundaries of tractable probabilistic inference.

3.13 Conditional Generative Models and Where to Apply Them

Max Welling (University of Amsterdam, NL)

I talked about how we can use flow and diffusion models to generate data from the equilibrium distribution, but that it seems much harder to generate from conditional generative models of the form $\boldsymbol{F} : (\boldsymbol{z}, \boldsymbol{x}) \rightarrow \boldsymbol{y}$ with $\boldsymbol{z} \sim p(\boldsymbol{z})$ and $\boldsymbol{x}$ some conditioning statement. These models are important for searching through chemical space, for proposing moves in a MCMC algorithm, for modeling domain shifts, etc. This talk will be mostly asking questions: why is this problem hard (harder than sampling from the unconditional distribution $\boldsymbol{F} : \boldsymbol{z} \rightarrow \boldsymbol{y}$?

3.14 Bayesian Deep Learning and a Probabilistic Perspective of Model Construction

Andrew G. Wilson (New York University, US)

Main reference Andrew Gordon Wilson, Pavel Izmailov: "Bayesian Deep Learning and a Probabilistic Perspective of Generalization", CoRR, Vol. abs/2002.08791, 2020.
URL https://arxiv.org/abs/2002.08791

The key distinguishing property of a Bayesian approach is marginalization, rather than using a single setting of weights. Bayesian marginalization can particularly improve the accuracy and calibration of modern deep neural networks, which are typically underspecified by the data, and can represent many compelling but different solutions. We show that deep ensembles provide an effective mechanism for approximate Bayesian marginalization, and propose a related approach that further improves the predictive distribution by marginalizing within basins of attraction, without significant overhead. We also investigate the prior over functions implied by a vague distribution over neural network weights, explaining the generalization properties of such models from a probabilistic perspective. From this perspective, we explain results that have been presented as mysterious and distinct to neural network generalization, such as the ability to fit images with random labels, and show that these results can be reproduced with Gaussian processes. We also show that Bayesian model averaging alleviates double descent, resulting in monotonic performance improvements with increased flexibility. Finally, we provide a Bayesian perspective on tempering for calibrating predictive distributions.

Participants

- Alessandro Antonucci
IDSIA – Manno, CH
- Michael Chertkov
University of Arizona – Tucson, US
- YooJung Choi
UCLA, US
- Alvaro Correia
TU Eindhoven, NL
- Priyank Jaini
Google – Toronto, CA
- Kristian Kersting
TU Darmstadt, DE
- Stefan Mengel
University of Artois/CNRS – Lens, FR
- Eric Nalisnick
University of Amsterdam, NL
- Sriraam Natarajan
University of Texas – Dallas, US
- Mathias Niepert
Universität Stuttgart, DE
- Robert Peharz
TU Graz, AT
- Xiaoting Shao
TU Darmstadt, DE
- Guy Van den Broeck
UCLA, US
- Antonio Vergari
University of Edinburgh, GB
- Andrew G. Wilson
New York University, US

Remote Participants

- Marcus A. Brubaker
York University – Toronto, CA
- Cassio de Campos
TU Eindhoven, NL
- Nicola Di Mauro
University of Bari, IT
- Laurent Dinh
Montreal, CA
- Danilo Jimenez Rezende
Google DeepMind – London, GB
- Mikko Koivisto
University of Helsinki, FI
- Sara Magliacane
University of Amsterdam, NL
- Lilith Francesca Mattei
IDSIA – Lugano, CH
- Denis D. Mauá
University of Sao Paulo, BR
- Karthika Mohan
Oregon State University, US
- David Montalvan Hernandez
TU Eindhoven, NL
- Deepak Pathak
Carnegie Mellon University – Pittsburgh, US
- Tahrima Rahman
University of Texas – Dallas, US
- Jakub Tomczak
VU University Amsterdam, NL
- Aki Vehtari
Aalto University, FI
- Max Welling
University of Amsterdam, NL
- Yaoliang Yu
University of Waterloo, CA
- Han Zhao
University of Illinois – Urbana-Champaign, US

Report from Dagstuhl Seminar 22162

Urban Mobility Analytics

David Jonietz[*1], Monika Sester[*2], Kathleen Stewart[*3], Stephan Winter[*4], Martin Tomko[*5], and Yanan Xin[†6]

1 **HERE – Zürich, CH.** david.jonietz@here.com
2 **Leibniz Universität Hannover, DE.** monika.sester@ikg.uni-hannover.de
3 **University of Maryland – College Park, US.** stewartk@umd.edu
4 **The University of Melbourne, AU.** winter@unimelb.edu.au
5 **The University of Melbourne, AU.** tomkom@unimelb.edu.au
6 **ETH – Zürich, CH.** yanxin@ethz.ch

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 22162 "Urban Mobility Analytics". The seminar brought together researchers from academia and industry who work in complementary ways on urban mobility analytics. The seminar especially aimed at bringing together ideas and approaches from deep learning research, which is requiring large datasets, and reproducible research, which is requiring access to data.

Seminar April 18–22, 2022 – http://www.dagstuhl.de/22162
2012 ACM Subject Classification Applied computing → Transportation; Computing methodologies → Planning under uncertainty; Information systems → Geographic information systems
Keywords and phrases data analytics, Deep learning, Reproducible research, urban mobility
Digital Object Identifier 10.4230/DagRep.12.4.26

1 Executive Summary

Monika Sester (Leibniz Universität Hannover, DE)
Martin Tomko (The University of Melbourne, AU)
Stephan Winter (The University of Melbourne, AU)

Seminar 22162 addressed recent trends in urban mobility analytics that are shaping the information available to transport planners, operators, and travellers. Seminar participants were particularly discussing how information can be provided that supports the critical transformation of urban mobility towards climate neutrality and other sustainability goals, i.e, that supports to change mobility behaviour.

The trends identified for this seminar were, on one hand, the rise of deep learning methods for massive data analytics, and on the other hand the emerging digital divide between those having massive data and those who haven't, which, in short, forms the challenges of academia for reproducible research. Massive data on urban mobility is collected by industry and transport authorities, with limited access outside, for various reasons. Also, the research and development capacity behind the closed doors of large transnational companies – especially in the platform economy – is arguably faster than the typical PhD process.

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

These challenges and opportunities were shaping the discussions where participants split into working groups on (a) *ethics and the social good* – how can information trigger change in mobility behaviour; (b) *methods and explainability*; (c) *benchmarking and datasets*; and (d) *applications*.

The seminar had quite a diversity of participants, which was inspiring in all the discussions. Participants from industry gave talks about what happened behind their 'closed doors', and further tutorials were introducing datasets, the principle of reproducible research, and European funding opportunities.

The industry partners showed great interest in collaboration with academia, however, the problem of data sharing was still considered as paramount. There are trends to open certain kinds of data, e.g. in aggregated form, or simulated data, or only based on contracts with certain institutions. Still, open data sharing remains to be a challenge.

2 Table of Contents

3 Overview of Seminar

3.1 Urban Mobility Analytics Seminar

Monika Sester (Leibniz Universität Hannover, DE), Martin Tomko (The University of Melbourne, AU), Stephan Winter (The University of Melbourne, AU)

Transportation in cities is undergoing unprecedented change, such as by vehicle technology towards autonomous driving (disrupting mobility); massive real-time data and data analytics (smart cities, sensing cities, dashboards); sharing and integration platforms (ride-hailing, mobility-as-a-service) and urban logistics (changing shopping patterns). All this happens in parallel with an increasing willingness, and a sharp necessity, to change mobility behaviour in the face of human-induced climate change, where urban transport is a major contributor [1, 2, 3].

Critical to the success of transforming urban mobility towards climate neutrality is information provided to planners, operators, and travellers. Increasingly, this information can be produced based on data. This Dagstuhl Seminar on Urban Mobility Analytics addressed recent trends that are shaping the information derived from such urban mobility analytics:

- A prominent trend, not only in transportation research, is the rise of deep learning methods for massive data analytics [4, 5]. In the domain of urban mobility, this massive data emerges from a range of sensor platforms, from infrastructure (CCTV, induction loops, people counters, WiFi, smart cards, air quality) to vehicles (GPS, vision, LiDAR, radar) and smartphones (GPS, location-based apps, accelerometer, gyroscope, magnetometer), in volume, heterogeneity, velocity and veracity a prime application domain for deep learning.
- A second trend is the emerging digital divide between academia and industry and its challenges for reproducible research [6, 7], a trend that has been compared to digital feudalism [8]. While massive data on urban mobility is collected by industry and transport authorities, their access for academic research is limited by privacy concerns and also by commercial sensitivities. While reproducible research hinges on access to data (and the generation of benchmark datasets is costly and often limited to narrow use cases), the research and development capacity behind the closed doors of large transnational companies – especially in the platform economy – is arguably faster than the typical PhD process.
- Related to both trends above is the buzzword of Digital Twins (e.g., `https://muenchen.digital/twin/`. Since both data and data analytics become more often available sufficiently close to real-time, the information derived is less and less consumed in human decision making but in the self-regulation of cyber-physical-social systems. These systems will propel future urban mobility by autonomously driving vehicles and mobility-as-a-service [9, 10], however, their development and use involve many still-open research questions, such as reliability, trust, and the interaction of humans with these systems, let alone the bigger question of social or ethical engagement in data-driven solutions [11].

Accordingly, the seminar brought together researchers from academia and industry who work in complementary ways on urban mobility analytics such that they do not necessarily meet at the same conferences or refer to standardized discipline practices. Especially we aimed to bring together ideas and approaches from deep learning research, which is requiring large datasets, and reproducible research, which is requiring access to data.

Tuesday	Wednesday	Thursday	Friday
Welcome, aims, program, and short introductions	Session 1 of participants talks – relevant research	David Doerr: *Horizon Europe's calls on smart mobility*	From break-out discussions to outcomes
Industry stories: • Michael Nolting (VW) • Erik Nygren (SBB)	Session 2 of participants talks – relevant research	Break-out groups	Conclusions: plans ahead
Moritz Neun and Christian Eichenberger (IARAI): *IARAI data tutorial*	*Hike into the nearby woods*	Break-out groups, ctd.	
Data challenges and activities	Review of challenges and activities	Daniel Nüst: *Reproducible research*	

Figure 1 Seminar structure.

The seminar made also a deliberate effort to invite people from both sides of the digital divide (i.e., from academia and industry) to share their experiences, their approaches, and their challenges, and to explore more future collaboration. These dynamics at the seminar were also inspired by two available real-world, large traffic datasets, one sponsored by IARAI (`https://www.iarai.ac.at/traffic4cast/`), and one by SBB and others (`https://flatland.aicrowd.com`).

Since the seminar took place in the Easter week, it was only a four-day seminar, starting with Tuesday. The week had a recognizable structure (Figure 1):

- **Tuesday**, after short introductions of the participants, belonged completely to the industry. Two industry speakers, Michael Nolting from VWN and Erik Nygren from SBB, spoke about data capture in their commercial environments, and the use of this data for management and planning. Erik Nygren introduced also the Flatland challenge. Afterwards Moritz Neun and Christian Eichenberger (IARAI) presented the Traffic4cast dataset in detail. In the last session, participants discussed themes for the week, including a data challenge.
- **Wednesday** was filled with short talks by the participants, introducing their research interests and embedding them into the context of the seminar. The afternoon, after the traditional hike, was completed by a plenary session to rank the identified themes and to plan for break-out groups.
- **Thursday** was filled with these break-out groups, framed by an introduction to current and future Horizon Europe calls on smart mobility in the morning by David Doerr, and an introduction to reproducible research by Daniel Nüst.
- **Friday** collected the discussions of the break-out groups, discussed reporting, and ended with collecting concrete commitments for next steps.

In the following, the report will first give an overview of selected talks, followed by the summaries of the break-out groups. It concludes with a brief summary of the outcomes.

Due to the ongoing pandemic, the seminar took place in hybrid mode, with about two thirds of participants on site. While Dagstuhl's conferencing system and our two technical assistants Maya Santhira Sekeran and Ivan Majic were providing a smooth interaction, time zones were causing challenges for the online participants. Unfortunately, two of the original organizers – Kathleen Stewart and David Jonietz – were not able to participate at all, such that the remaining two original organizers – Monika Sester and Stephan Winter – were grateful for Martin Tomko to come on board in the last minute. The organizers acknowledge every person's contributions and commitment.

References

1 Boschmann, E. & Kwan, M. Toward Socially Sustainable Urban Transportation: Progress and Potentials. *International Journal Of Sustainable Transportation*. **2**, 138-157 (2008)

2 May, A. Urban Transport and Sustainability: The Key Challenges. *International Journal Of Sustainable Transportation*. **7**, 170-185 (2013)

3 OECD Decarbonising Urban Mobility with Land Use and Transport Policies: The Case of Auckland. (OECD Publishing,2020)

4 Li, S., Dragicevic, S., Anton, F., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A. & Cheng, T. Geospatial Big Data Handling Theory and Methods: A Review and Research Challenges. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **115**, 119-133 (2016)

5 Varghese, V., Chikaraishi, M. & Urata, J. Deep Learning in Transport Studies: A Meta-analysis on the Prediction Accuracy. *Journal Of Big Data Analytics In Transportation*. **2**, 199-220 (2020), https://doi.org/10.1007/s42421-020-00030-z

6 Ivie, P. & Thain, D. Reproducibility in Scientific Computing. *ACM Comput. Surv.*. **51**, Article 63 (2018), https://doi.org/10.1145/3186266

7 Nüst, D. & Pebesma, E. Practical Reproducibility in Geography and Geosciences. *Annals Of The American Association Of Geographers*. **111**, 1300-1310 (2021), https://doi.org/10.1080/24694452.2020.1806028

8 Jensen, J. The Medieval Internet: Power, Politics and Participation in the Digital Age. (Emerald Publishing,2020)

9 Ibrahim, M., Rassõlkin, A., Vaimann, T. & Kallaste, A. Overview on Digital Twin for Autonomous Electrical Vehicles Propulsion Drive System. *Sustainability*. **14** (2022), https://www.mdpi.com/2071-1050/14/2/601

10 Mahmoud, E., Darwish, A. & Hassanien, A. The Future of Digital Twins for Autonomous Systems: Analysis and Opportunities. *Digital Twins For Digital Transformation: Innovation In Industry*. pp. 187-200 (2022), https://doi.org/10.1007/978-3-030-96802-1_10

11 Charitonidou, M. Urban scale digital twins in data-driven society: Challenging digital universalism in urban planning decision-making. *International Journal Of Architectural Computing*. pp. 14780771211070005 (2022), https://doi.org/10.1177/14780771211070005

4 Overview of Talks

4.1 Vehicle Data Democratization at Volkswagen Commercial Vehicles

Michael Nolting (Volkswagen Nutzfahrzeuge – Hannover, DE)

In the near future, cars will become more than just mechanical objects bringing customers safely and quickly from point A to B. Currently, the focus is on mechanical improvements to safety features such as airbags or reducing fuel consumption and emissions. Soon, however, the focus will switch to the electrical and IT aspects of cars. The automobile will become a rolling computer, providing added value via autonomous driving and an infotainment system which makes the resulting leisure time more productive and enjoyable. Thanks to improvements in electric motor technology, the complexity regarding manufacturing has decreased significantly. This means that enterprises such as Google or Apple – companies that develop sophisticated computer, smart phone and infotainment platforms – will be more involved in important aspects of car production than traditional car manufacturers such as Volkswagen. As often stated, based on valuable information stemming from big data analytics, future cars will be electric, autonomous, connected, on the whole smart.

You can regularly read in newspapers about the progress car manufacturers and their suppliers are making regarding the first three points. There is not much conversation, however, regarding the coming smart car revolution. It is obvious that the industry has recognized the risk of losing future market share by failing to evolve into big data enterprises based on recent acquisitions, investments and partnerships they have made, and their hiring spree of data scientists. BMW, Audi, and Mercedes, for example, just bought Nokia's mapping service, and BMW has undertaken a massive recruitment of data scientists. Nevertheless, time is running out. If car manufacturers don't start to take the reins now, they will miss out on the smart car revolution of tomorrow.

This is the reason why car manufacturers have to transform into a software company. In order to achieve this, they have to increase their daily deployment frequency and reduce the overall lead time. This can be achieved by coping with the integration issues, which currently still exist. In his talk, Dr. Michael Nolting has shown ways how to overcome these integration issues and so paving the way how to transform into a software-driven company. In addition to this, the next steps would then be to become a data-driven company be democratizing data and AI within the company and find ways and solutions how to share this data with research institutes, which is still a challenging task.

4.2 Mobility Research at Swiss Federal Railway Company

Erik Nygren (Schweizerische Bundesbahnen – Bern, CH)

The Swiss Federal Railway Company (SBB) is an integral part of the Swiss public mobility network and transports over 1 million passengers and 200 thousand tons of goods each day. Current projections indicate that the demand for public urban mobility will increase by up to 40%. To be able to continue to offer a reliable public service many technical, social, and operational challenges must be overcome. SBB has identified 7 main topics of interest where applied research together with research institutes and other partners is being conducted.

1. Customer Oriented Railway: We focus on enhancing the attractiveness of railway through different incentives and improvements to our service. Social and behavioral research is needed to better understand the decision making of our customers.
2. Simplify Access to Railway: The focus lies in the seamless integration of railway into other modes of transport. Research focus is both on technical issues related to the interaction of different modes of transportation as well as social and behavioral aspects to understand the customer needs.
3. Flexible offers and production models: The change in mobility behavior both in passenger and freight transportation towards on-demand systems requires dynamic and real-time planning. Research is mainly conducted around network wide optimization and planning in real-time.
4. Resilience and Efficiency: Railway industry is asset heavy, and its reliable operations depend on well planned and efficient maintenance. We research the use of novel technologies and algorithms to improve both planning and execution of maintenance work on the infrastructure and rollingstock.
5. Long-term strengthening of SBB: What will the future of mobility look like and what will the role of railway be? This and many other questions are being investigated while considering both technological and social developments in society.

6. Environment and Sustainability: Railway today is already one of the most sustainable forms of mobility, we aim to further lower the impact of mobility by improving the efficiency of material usage and lower the energy consumption per travelled kilometre.
7. Optimized freight logistics: To facilitate the move from road-based transportation systems to railway, more dynamic freight logistics need to be offered. Together with academia, we are looking for more efficient and reliable planning and coordination algorithms as well as new business models to improve railway freight attractiveness.

SBB uses different forms of applied and academic research to tackle open questions around these 7 topics of interest. Research projects conducted in close collaboration with academic partners allow us to get a good grasp of future possibilities and plan our services accordingly. In addition, we use Open Data and Open Research to reach a larger group of experts from many different fields. Open Research allows us to compare different research results with each other and monitor the progress and performance of novel approaches.

This approach has proven valuable for traffic management systems research, where SBB has hosted an optimization competition called Flatland from 2018 until today. The competition consists of a simple railway traffic simulation and the objective to optimize traffic flow towards high punctuality while stochastic disturbances force frequent large-scale replanning. Over the years many novel and surprising solutions have been submitted by participants around the world and been evaluated. We observe that new algorithms from the field of Deep Learning still struggle to outperform classical optimization algorithms, but that their quality and performance is advancing each year. Throughout all different research projects conducted at our company we have learned that a close and bilateral collaboration on research topics is necessary to shorten the time to market for new ideas. Inspiration for new solutions often comes from academic research and industry partners can provide valuable insights into real-world challenges faced by mobility providers. We, therefore, invest in more interdisciplinary exchange and collaboration between industry and academia.

4.3 Traffic4cast Data Intro

Moritz Neun (IARAI – Zürich, CH) and Christian Eichenberger (IARAI – Zürich, CH)

Expectations on AI in mobilities are immense, targeting climate-neutral & smart cities. To achieve this goal there is an increasing need not only for models but also for metrics that help steer and influence these complex urban mobility systems. To give priority to people and not to transportation vehicles, cities must learn how to understand traffic as a whole for shaping the performance of a transportation system and for evaluating and reflecting changes in the real-world through an iterative data-driven setting (i.e. online model calibration; model, data fusion and metrics innovation; real-world change detection). Such a learned digital twin would help close the gap between traffic control systems and model-based planning tools by providing an integrated holistic feedback loop with more diverse data and a deeper embedding of traffic dependencies.

The Traffic4cast competition aims at providing such an understanding of traffic rules in a data-driven way, pushing the latest methods in modern machine learning to model complex spatial systems over time. The competition was part of the NeurIPS competition track in the last 3 years and provides an industrial-scale dataset with a high-resolution privacy-preserving

view of urban traffic. The dataset is derived from GPS trajectories (floating car data) of a large fleet of probe vehicles and covers 10 culturally diverse cities around the world in a time span of 2 years.

Floating car data from GPS is a typical source to be used in trajectory or origin-destination analysis. On the other hand, in traffic detection and prediction other data sources such as stationary traffic counters are used. While traffic counter data only observes traffic in certain locations (spatial bias), floating car data can observe traffic everywhere but usually sees only a tiny part of the total traffic "population" (car fleet bias of industrial providers and participation bias of crowd-sourced initiatives). With the Traffic4cast dataset we now do have a floating car dataset that has a sufficiently large volume to study those effects.

In Traffic4cast the floating car data is aggregated into the traffic map movie format using spatio-temporal cells – this facilitates processing and preserves privacy. Each cell corresponds to the area of approximately 100m x 100m. The GPS data is aggregated as vehicle count and average speed per cell, temporally in 5 minute time intervals and axially in 4 heading quadrants (NE, SE, SW, and NW). While this 4-level aggregation and compression scheme gives away some details in the data, at the same time it also allows it to handle much larger data volumes over longer periods in time. It also allows and encourages the direct use of the latest advances in analytics and ML, as has been successfully shown in the results of the past three competitions.

In 2022, Traffic4cast is moving from the grid based movies format to a graph based format which allows to combine loop counter and floating car data. The floating car data in the gridded map movie format will serve as input for deriving the speed and travel time ground truth labels on the road graph. Therefore the 5 minute time bins in our usual spatio-temporal data format (see [1]) are getting aggregated to 15 minutes time bins taking the average speed to our loop counter data. From the dynamic data, we then derive the ground truth labels, namely CongestionClass (CC; red/congested, yellow/warning, green/uncongested) for each segment in the road graph and Travel Time (ETA) for each super-segment.

We focus on the three cities London, Madrid and Melbourne where both floating car data as well as large and open traffic counter datasets are available. The underlying raw data is similar to the input for commercial traffic maps and routing products. Similar to the previous years, the data will be made available for download from HERE Technologies.

References

1 Christian Eichenberger, Moritz Neun, Henry Martin, Pedro Herruzo, Markus Spanring, Yichao Lu, Sungbin Choi, Vsevolod Konyakhin, Nina Lukashina, Aleksei Shpilman, Nina Wiedemann, Martin Raubal, Bo Wang, Hai L. Vu, Reza Mohajerpoor, Inhi Kim, Luca Hermes, Andrew Melnik, Riza Velioglu, Markus Vieth, Malte Schilling, Alabi Bojesomo, Hasan Al Marzouqi, Panos Liatsis, Jay Santokhi, Dylan Hillier, Yiming Yang, Joned Sarwar, Anna Jordan, Emil Hewage, David Jonietz, Fei Tang, Aleksandra Gruca, Michael Kopp, David Kreil, and Sepp Hochreiter. Traffic4cast at NeurIPS 2021 – temporal and spatial few-shot transfer learning in gridded geo-spatial processes. In Hugo Jair Escalante and Katja Hofmann, editors, *Proceedings of the NeurIPS 2021 Competition Track*, volume forthcoming of *Proceedings of Machine Learning Research*. PMLR, 2022. URL `https://arxiv.org/abs/2203.17070`.

4.4 Deep Learning of Road User Behavior

Hao Cheng (Leibniz Universität Hannover, DE)

Learning how road users behave is essential for developing many intelligent systems, such as traffic safety control, self-driving cars, and robot navigation systems. However, automated and accurate recognition of road users' behavior is still one of the bottlenecks in realizing such systems in urban traffic that is – compared to other types of traffic – especially dynamic and full of uncertainties. Some urban environments make detecting and predicting road users' behavior particularly challenging, e.g., temporarily shared spaces of intersections for vehicle turning or shared spaces as a traffic design. The former allows vehicles to turn and interact with other crossing road users, the latter is intended to make different types of road users share the space, therefore reducing the dominance of vehicles and improving pedestrian movement and comfort. Direct interactions between vehicles and vulnerable road users (VRUs, e.g., pedestrians and cyclists) lead to high uncertainty of traffic behavior. Their dynamic movements and mutual influence make their behavior multimodal, such as stopping, accelerating, and turning in different directions. A road user may have more than one option to choose such maneuvers even in the same traffic situation. Moreover, ambiguous traffic situations (e.g., road users negotiating usage of the road) make their behavior difficult to predict.

With the development of deep learning techniques and the availability of large-scale real-world traffic data, there is a high chance to automatically and accurately learn road users' behavior. The core question is how to leverage such traffic data captured by e.g., camera or LiDAR, or vector data like trajectories and GPS tracks, to train a deep learning model that can mimic the multimodality of road users' behavior in various traffic situations, especially at places as mentioned above where they inevitably confront each other?

This research project aims to investigate building smart intersections and shared spaces, with the ability to predict how different types of road users move and interact with each other. The following steps are a conceptual pipeline of learning road users' multimodal intent. First, traffic in the area of interest at a vehicle turning intersection or in a shared space is captured by using stationary camera sensors. Then, the image pre-processing steps, i.e., camera calibration, object detection, projection transformation, and object tracking, are carried out to extract trajectories from the image data. The well-established approaches from, e.g., benchmark multi-object tracking algorithms[1] will be applied for trajectory tracking. The main contributions of this research lie in the step of trajectory forecasting and intent prediction. The predicted behavior in space and time, e.g., intended trajectories, are analyzed for three major tasks: a) path planning of the ego agent [1, 2], b) safety analysis, including collisions and conflicts of different severity [3], and c) anomaly detection, classifying behavior patterns that do not conform to a well-defined notion of normal behavior [4]. In parallel to the predicted behavior, the further extracted behavior will serve as an observed reference for evaluating the performances of the prediction tasks.

[1] `https://motchallenge.net/`

References

1 Cheng, H., Liao, W., Tang, X., Yang, M. Y., Sester, M., and Rosenhahn, B. (2021). Exploring dynamic context for multi-path trajectory prediction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12795–12801.

2 Cheng, H., Liao, W., Yang, M. Y., Rosenhahn, B., and Sester, M. (2021). Amenet: Attentive maps encoder network for trajectory prediction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 172:253–266.

3 H. Cheng, H. Liu, T. Hirayama, F. Shinmura, N. Akai, and H. Murase, "Automatic interaction detection between vehicles and vulnerable road users during turning at an intersection," in *Proceedings of the 31st IEEE Intelligent Vehicles Symposium, Las Vegas, NV, USA*, vol. 19, 2020.

4 Koetsier, C., Fiosina, J., Gremmel, J. N., Müller, J. P., Woisetschläger, D. M., and Sester, M. (2022). Detection of anomalous vehicle trajectories using federated learning. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 4:100013.

4.5 Modeling the Interaction between Places and Human Mobility

Cheng Fu (Universität Zürich, CH)

Place and human mobility are the two sides of the coin in an urban system: Human activities, including mobility, reflect the actual usage of places. On the other hand, the spatial organization of places determines the origin and destination of the daily trips, e.g., commuting flows start from residential areas to commercial areas. The purposes of the trips are also associated with the amenities provided by the destination places. Such dual relationships between place and human mobility can happen at different spatial, temporal, and behavioral scales.

By modeling the influence of places on driving behaviors and trajectories, our research team found that certain types of places are good predictors for estimating car accident risks for drivers [1]. We also applied places (e.g., POIs) as the semantic context of trajectories for adaptive simplification so that the segments of a trajectory near dense POIs are simplified less while segments near sparse POIs are simplified more [2].

Our recent interests mainly focus on how places may influence older adults. Many countries are experiencing fast aging. Infrastructures and facilities in cities however are primarily planned for working-age commuters. To achieve healthy aging, we thus need to understand the mobility patterns of the older adults and model their interactions with the places. The Mobility, Activity, and Social Interaction Study (MOASIS) project overcomes the digital gap of older adults by collecting mobility data from customized GPS loggers. The collected data are small in terms of the number of participants but rich in their demographic details and their physical and mental health status. We supervised several MSc theses regarding whether older adults' visitation patterns to certain types of places can be good indicators to infer their physical and mental health.

For place modeling with big human mobility data, the conventional workflows model the activities in the place per se as the features for learning. The recent development of neural networks, and particularly deep learning networks, shows the capacity to embed complex contextual information. That provides more powerful computational models for more complex conceptual models on place modeling with mobility data.

References

1 Brühwiler, L., Fu, C., Huang, H., Longhi, L. & Weibel, R. Predicting individuals' car accident risk by trajectory, driving events, and geographical context. *Computers, Environment And Urban Systems*. **93**, 101760 (2022), https://doi.org/10.1016/j.compenvurbsys.2022.101760

2 Fu, C., Huang, H. & Weibel, R. Adaptive simplification of GPS trajectories with geographic context – a quadtree-based approach. *International Journal Of Geographical Information Science*. pp. 1-28 (2020), https://doi.org/10.1080/13658816.2020.1778003

4.6 Horizon Europe: Introduction to R&I Funding in the Field of Mobility

Stephan Winter (The University of Melbourne, AU) and David Doerr (TÜV Rheinland – Köln, DE)

URL https://ec.europa.eu/research/pdf/horizon-europe/annex-5.pdf

David Doerr, from *Nationale Kontaktstelle Klima, Energie, Mobilität für das EU-Rahmenprogramm für Forschung und Innovation "Horizont Europa"*, presented an introduction to mobility research funding available through Horizon Europe, the EU's key funding programme for research and innovation.

The EU Framework programmes are complementary to national funding and require European added value. They promote cooperation between researchers and innovators (such as companies, research centres). The priorities of the current Horizon Europe programme are *green – digital – innovative – open – resilient – participative*. Of the three pillars of the current framework, this talk focused on *Global Challenges and European Industrial Competitiveness: Climate, Energy and Mobility*, specifically Cluster 5: *Climate, Energy and Mobility*[2]. Actions 5 and 6 are about mobility.

Relevant for this seminar is *Destination 6: Safe, resilient transport and smart mobility services for passengers and goods*, where especially two topics are related to mobility data analytics: connected, cooperative, automated mobility (CCAM), and multimodal and sustainable transport systems.

Calls open in the Work Programme for 2021-2022 are:

- HORIZON-CL5-2022-D6-02-02: Urban logistics and planning: anticipating urban freight generation and demand including digitalisation of urban freight (Innovation Action)
- HORIZON-CL5-2022-D6-02-04: Accelerating the deployment of new and shared mobility services for the next decade (Innovation Action)
- HORIZON-CL5-2022-D6-02-05: Advanced multimodal network and traffic management for seamless door-to-door mobility of passengers and freight transport (Research and Innovation Action)

These calls are closing on 6 September 2022.

Related, with the same deadline, is a call in the *Missions Work Programme 2021-22 on Climate-neutral and Smart Cities*:

[2] https://ec.europa.eu/research/pdf/horizon-europe/annex-5.pdf

- HORIZON-MISS-2022-CIT-01-01: Designing inclusive, safe, affordable and sustainable urban mobility (Innovation Action)

For this action, a consortium must contain at least four cities as living labs, and additional four follower cities.

4.7 Reproducibility for Urban Mobility Analysis

Daniel Nüst (Universität Münster, DE)

Main reference Daniel Nüst: "Reproducibility for Urban Mobility Analysis", Zenodo, 2022.
URL https://doi.org/10.5281/zenodo.6477034

Open and reproducible research is a prerequisite for a sustainable and meaningful science. As I research in the context of geospatial sciences [1], the problems stemming from irreproducibility are substantial and solutions to improve reproducibility, albeit practical both regarding technology and culture, are too rarely implemented. If research is not reproducible it cannot be inspected, reused, or extended and thereby such works slow innovation and hinder the advancement of science. In my experience, challenges but also solutions from other disciplines can and should be transferred to mobility research to ensure the urban mobility researchers can help to solve societal challenges in the most transparent, effective and collaborative way. That is why I applaud the organisers of the Dagstuhl Seminar 22162 "Urban Mobility Analytics" [2] to provide time for the participants to learn about and discuss the foundations and challenges of reproducible research.

The session material is published at [3]. It covers theoretical and practical basics of reproducibility and draws from the large amount of open educational resources provided by the reproducible research and open science communities, e.g., [4, 5]. In the talk, I give advise on working reproducibly as individuals and present how communities can change their practices. For *individuals* to create reproducible research, I postulate the minimalistic motto *"Have a README: all else is details."*, which is inspired by Greg Wilson's first Rule of *Teaching Tech Together* and intents to make clear that reproducibility is an ideal to strive for with best efforts, not a binary property of a paper. I further stress the aspects of *craftspersonship* needed to realise computational reproducibility. The motto is extended to comprise different practices for integrating reproducible workflows into personal habits, such as good filenames, consciously managing the computational environment, using notebooks, publishing & citing code and data, and creating research compendia (cf. [6]). In *research groups or labs*, a very effective means to improve reproducibility-related skills and increase the quality of work are mutual reproductions amongst colleagues before submissions of manuscripts. For *community practices*, I argue that the establishing of code execution and reproductions as part of peer review is crucial to ensure adoption and proper credit. Two examples of successful initiatives for reproducibility reviews illustrate the feasibility of this approach [7, 8]. To convince the seminar participants to adopt these practices, to provide literature for further self-study, and to give material for sharing and convincing others, the session material includes an overview of relevant literature with a special eye on the individual benefits of working reproducibly.

Finally, I see the following *relevant topics for the urban mobility and transportation disciplines* with respect to open reproducible research. It can be noted that the topic is not widely discussed yet, with few mentions in the literature and no comprehensive studies on

the reproducibility of papers. The specific challenges are (i) the importance of (tailored, novel) hardware for research, e.g., for autonomous vehicles, which can not be shared digitally and is often expensive or unique, (ii) the high complexity of analyses, e.g., in the context of routing and networks, which often requires high performance computing methods, (iii) the lack of open, not proprietary, datasets and the bias of existing data towards specific modes of transport, i.e., cars, and (iv) the widespread use of machine learning and artificial intelligence (AI) approaches due to the complexity of the problems, which often are only shared as black boxes. Solutions for some these challenges do exist, such as synthetic data sharing and compiling new open datasets, or are an active field or research, such as explainable AI, but they to complicate matters and are likely to be used as an excuse to not conduct the steps for open reproducible research, which are still wrongly perceived as "extra work". The close collaboration with industry partners and public authorities, not the least because they have the (closed) data, also comes with opportunities. The skill set required for reproducible computational workflows, such as research software engineering (RSEng, [9]), is highly relevant for individuals to transition to industry jobs. Furthermore, reproducible data analyses have a proven functionality and high reusability that can greatly reduce the efforts for putting them into practice.

For the next steps, I invite the seminar participants to initiate discussions on the importance of reproducible research across their different roles in academia, be they authors, reviewers, or editors, as much as I encourage them to try to improve their own habits. As a next step, understanding the state of reproducibility in urban mobility research seems necessary. However, the community may need to have a discourse about which of the goals for data-driven research are most relevant for them: *replicability*, albeit needing new data and being more work, or *robustness* checks through applying multiple methods on the same data? All of these goals help to unhide the data and code underlying computational and data-driven research in urban mobility and eventually ensure a recognition of their important contributions to science.

References

1 Nüst, D. 2022. Infrastructures and Practices for Reproducible Research in Geography, Geosciences, and GIScience. Zenodo. https://doi.org/10.5281/zenodo.4768096
2 Dagstuhl Seminar 22162. Urban Mobility Analytics. D. Jonietz, M. Sester, K. Stewart, S. Winter (Organizers), M. Tomko (Coordinator). https://www.dagstuhl.de/22162
3 Nüst, D. 2022. Reproducibility for Urban Mobility Analysis. Zenodo. https://doi.org/10.5281/zenodo.6477034
4 The Turing Way Community. 2021. The Turing Way: A handbook for reproducible, ethical and collaborative research (1.0.1). Zenodo. https://doi.org/10.5281/zenodo.5671094
5 Reproducible Research Support Service (R2S2) Knowledge Base. https://confluence.uni-muenster.de/display/r2s2/
6 Nüst, D. & E. Pebesma. 2021. Practical reproducibility in geography and geosciences. Annals of the American Association of Geographers, 111(5), 1300–1310. tps://doi.org/10.1080/24694452.2020.180602
7 Reproducible AGILE. https://reproducible-agile.github.io/
8 Nüst, D., & S. J. Eglen. 2021. CODECHECK: An Open Science initiative for the independent execution of computations underlying research articles during peer review to improve reproducibility. F1000Research, 10, 253. https://doi.org/10.12688/f1000research.51738; https://codecheck.org.uk/
9 Cohen, J., D. S. Katz, M. Barker, N. Chue Hong, R. Haines and C. Jay. 2021. The Four Pillars of Research Software Engineering. IEEE Software 38(1), 97-105. https://doi.org/10.1109/MS.2020.2973362

5 Working Groups

5.1 Group on Ethics / Social Good

Alexandra Millonig (AIT – Austrian Institute of Technology – Wien, AT), Ivan Majic (TU Graz, AT), Edoardo Neerhut (Meta – Burlingame, US), Moritz Neun (IARAI – Zürich, CH), Luca Pappalardo (CNR – Pisa, IT), Chiara Renso (ISTI-CNR – Pisa, IT), and Stephan Winter (The University of Melbourne, AU)

5.1.1 Background

Going by paper titles at relevant conferences in both the intelligent transportation sector and the computing sector, much of current research on urban mobility analytics is spent on efficiency improvements of current mobility solutions: improving the efficiency of individual decisions on mobility choices such as modes (how to go), routes (where to go), or times (when to go), or transport network optimizations. All this research engages in predicting demand, providing supply, or managing systems. There are a number of issues with such thinking:

- It is treating human decision making as rational. But behavioral economists characterize human decision making as intrinsically tied to overconfidence, loss aversion, limited attention, and cognitive biases, all leading to systematic errors in judgment [1, 2].
- It is treating the accumulation of optimal decisions of individuals as the best achievable, which is a fallacy [3, 4].
- Its promised efficiency gains are just too little or too slow to respond to the current climate crisis and climate goals [5].

Therefore, the group discussion returned to the widely promoted three lines of action towards sustainable urban mobility, *avoid – shift – improve*[3], which have been introduced in the early 1990s but have failed to succeed in the coming decades, particularly regarding the avoidance of transport and a significant shift from motorized to non-motorized forms of mobility. Therefore, the group questioned the solutions which have been developed in the past under this principle and asked whether we, as a scientific community, could (or even should), do more urban mobility analytics research on avoiding and shifting (motorized) mobility. For example, the *where to go* question seems to have ignored the option of choosing alternative destinations for certain activities that are closer in space or travel time, with few exceptions such as [6].

5.1.2 Challenges and Opportunities for Urban Mobility Analytics

If change of mobility behavior is required, it should be incentivized, either by material or immaterial recognition, to trigger the human internal reward system and reinforcement learning [7]. More generally, the group asked whether negatively framed expressions, such as emissions-based or climate-based (reduction of a threat) might be too vague to trigger loss aversion biases, but positively framed expressions such as gains in quality of life (and rewards for contributions) are more tangible. One tested way of rewards in transport demand management is gamification [8], although incentives can also be linked to intrinsic motivations. However, the group is also aware of the limitations that gamification and

[3] e.g., `https://en.wikipedia.org/wiki/Avoid-Shift-Improve`

nudging approaches have in inducing behavior change across modes or even avoiding trips, as habits are particularly difficult to break if the alternative is not well known or is subject to prejudice. Positive reinforcement therefore demands bespoke messages to individual behavior and consequently very good insight into current behavior.

In this regard, the group identified one challenge: To monitor (prove) a change in behaviour of an individual, in order to reward, for example in cases when (a) people choose to walk rather than take motorised forms of transportation, or (b) people deliberately swap a longer distance trip for a nearer one, sacrificing some original intentions or rewards. The group sees here a potential application for urban mobility analytics. While gamification approaches reward individual decisions from a system's perspective (e.g., a public transport ride outside peak time), they typically do not link their rewards to a change in an individual's behavior. These changes could happen in each category: avoid, shift, or improve. Incentivization of individuals, while contributing to the societal good, raise significant ethical questions of privacy though.

The second area identified by the group where urban mobility analytics can make a difference is in supporting sustainable living in cities. Information derived from urban mobility analytics (Geo AI, mobility AI [9]) can be used to:

- Track, predict, and manage the impact of mobility on climate, with an immediate goal of climate neutrality. Air quality, temperature (health), and coping of infrastructure with weather events (urban resilience) are main factors of livability. Examples of this approach are [10, 11]. In this regard, current AI systems are optimised to satisfy the need of the individual only, without caring about the collective effects on the city. We should design AI systems that optimise for both individual needs and collective societal goods. For example, we should ensure that routing suggestions from navigation apps (e.g., Google Maps, Waze) optimise for societal good, and that this routing is advantaged over AI suggestions that do not.
- The spatial distribution of venues (places for activities) in a city significantly impacts citizens' mobility needs. The reduction of the impact of human mobility on climate and city well-being should be achieved not only by improving the efficiency of our transportation means, but also by reshaping our cities in terms, e.g., of how we distribute venues within them. Modern cities, in which neighborhoods contain clusters of a variety of economic or social activities, may reduce mobility demand, equalise transportation disadvantages, open public space for forms of use other than transport and can hence decrease the negative impacts of motorized traffic [12]. The 15-minute city[4] – a city of nearly self-sufficient neighborhoods ('superblocks') – should be tested more rigorously by means of data analytics, and what-if analyses about the spatial distribution of venues and their impact on mobility should be conducted and integrated in urban planners' body of knowledge [13, 14] and tools.
- Support the resilience of a city to epidemics. The COVID-19 pandemic made evident how fragile our cities are in facing epidemics [15, 16]. Indeed, limiting human mobility was needed to reduce the diffusion of the COVID-19 pandemic in cities all around the world. Policies for the reduction of impact of mobility on climate and well-being should take into account also the epidemics dimension, so to avoid mobility solutions that are potentially dangerous in cases of epidemics spread.

[4] `https://en.wikipedia.org/wiki/15-minute_city`

5.1.3 Potential Contributions

In this regard, urban mobility analytics can be further developed to form a body of knowledge about the sensitivities of urban designs, and thus inform urban planners through novel decision-support systems.

Further references

- Carlos Moreno: The 15-minute city (`https://youtu.be/TQ2f4sJVXAI`)
- Andreas M. Dalsgaard, Jan Gehl: The Human Scale (`https://youtu.be/oA2eAQKkr-k`)

References

1 Gigerenzer, G. & Goldstein, D. Reasoning the Fast and Frugal Way: Models of Bounded Rationality. *Psychological Review*. **103**, 650-669 (1996)

2 Kahneman, D. Thinking, Fast and Slow. (Farrar, Straus,2011)

3 Tversky, A. & Kahneman, D. Judgement under Uncertainty: Heuristics and Biases. *Science*. **185**, 1124-1131 (1974)

4 Roughgarden, T. Selfish Routing and the Price of Anarchy. (MIT Press,2005)

5 IPCC Climate Change 2022: Impacts, Adaptation, and Vulnerability – Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. (Intergovernmental Panel on Climate Change,2022), `https://www.ipcc.ch/report/ar6/wg2/`

6 Wang, Y., Winter, S. & Tomko, M. Collaborative activity-based ridesharing. *Journal Of Transport Geography*. **72**, 131-138 (2018)

7 Schultz, W. Neuronal Reward and Decision Signals: From Theories to Data. *Physiological Reviews*. **95**, 853-951 (2015)

8 Yen, B., Mulley, C. & Burke, M. Gamification in transport interventions: Another way to improve travel behavioural change. *Cities*. **85** pp. 140-149 (2019)

9 Luca, M., Barlacchi, G., Lepri, B. & Pappalardo, L. A Survey on Deep Learning for Human Mobility. *ACM Computing Surveys*. **55**, Article 7 (2021)

10 Nyhan, M., Sobolevsky, S., Kang, C., Robinson, P., Corti, A., Szell, M., Streets, D., Lu, Z., Britter, R., Barrett, S. & Others Predicting vehicular emissions in high spatial resolution using pervasively measured transportation data and microscopic emissions model. *Atmospheric Environment*. **140** pp. 352-363 (2016)

11 Böhm, M., Nanni, M. & Pappalardo, L. Gross polluters and vehicles' emissions reduction. (arXiv,2021)

12 Millonig, A., Rudloff, C., Richter, G., Lorenz, F. & Peer, S. Fair mobility budgets: A concept for achieving climate neutrality and transport equity. *Transportation Research Part D: Transport And Environment*. **103** pp. 103165 (2022), `https://www.sciencedirect.com/science/article/pii/S1361920921004600`

13 Christaller, W. Die zentralen Orte in Süddeutschland. (Gustav Fischer,1933)

14 Jacobs, J. The death and life of great American cities. (Random House,1961)

15 Kraemer, M., Yang, C., Gutierrez, B., Wu, C., Klein, B., Pigott David, M., Group, O., Plessis, L., Faria, N., Li, R., Hanage, W., Brownstein, J., Layan, M., Vespignani, A., Tian, H., Dye, C., Pybus, O. & Scarpino, S. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*. **368**, 493-497 (2020)

16 Lucchini, L., Centellegher, S., Pappalardo, L., Gallotti, R., Privitera, F., Lepri, B. & De Nadai, M. Living in a pandemic: changes in mobility routines, social activity and adherence to COVID-19 protective measures. *Scientific Reports*. **11**, 1-12 (2021)

5.2 Group on Benchmarking and Datasets

Vanessa Brum-Bastos (Wroclaw University of Environmental and Life Sci., PL), Christian Eichenberger (IARAI – Zürich, CH), Cheng Fu (Universität Zürich, CH), Erik Nygren (Schweizerische Bundesbahnen – Bern, CH), and Maya Sekeran (TU München, DE)

Movement is a fundamental characteristic of life [1]. Humans move on a daily basis for the most diverse reasons, to multiple places and using varied transportation modes. The understanding of human mobility behaviour, i.e., "why", "where", "when" and "how" people move, is extremely relevant for urban planning, traffic engineering, policy making and other many applications [2]. Moreover, in face of climate changes and the recent COVID-19 pandemic, it even became pivotal to further understand human mobility in order to manage its related carbon footprint and epidemiological role.

5.2.1 Challenges and Problems

The recent technological advances in location-based services and devices have been providing researchers with unprecedented amounts of data to study human mobility [3]. However, the same vast amounts of data that brought new research opportunities in urban mobility analytics have also created a series of new challenges that still need to be addressed by the research community. Our group discussion focused on trying to identify and propose potential solutions to these problems.

1. **Benchmark datasets**: benchmark datasets are used for both training and testing of methodologies, but can also be helpful for evaluating the performance of unfamiliar datasets for implemented methods. The desire for benchmark datasets became clear during the discussions in breakout groups and the plenary. However, the multidisciplinary aspect of urban mobility analytics seems to call for not one but multiple benchmarking datasets that can attend the demands of specific applications within urban mobility analytics, such as traffic planning, routing or programming self-driving vehicles.
2. **Bias and representativeness of the data**: the massive amounts of data provided by modern location-based technologies come along with an increasing public and legal concern with individual's privacy. These privacy concerns have led to data anonymization, meaning that researchers do not have any metadata describing the population sample (e.g., gender, age, income) that generated the mobility datasets and therefore are not aware of how representative of the total population these data are. Understanding the bias in modern mobility datasets is especially critical for applied research on urban mobility analytics for solving real world problems.
3. **Data availability and access**: Even though massive amounts of data on human mobility are produced daily, that does not mean that these data are accessible to researchers and policy makers. Many of these datasets are produced, owned and sold by private companies, such as Near and Strava. The high fees for such data may pose a barrier to many researchers, especially the ones working at countries with less advantaged economies. There are also datasets owned by governmental agencies and public companies, however, even when those are available, many of them are not widely known by or not readily accessible to the research community.

5.2.2 Potential Paths for Future Solutions

We believe that a unified international data catalog could be a good starting point for addressing the aforementioned issues. The development of a unified international data catalog has been a strategy used by movement ecology (see [4]), a discipline that has also recently benefited from the new opportunities and challenges brought by location-based services and devices [5]. In fact, a discipline that has been working for decades on challenges that urban movement analytics researchers are just now facing.

We believe that unified international data catalogue would support and foster a growing research community as well as potentially bring the following benefits:

1. **Federated cataloguing data sets on human and human-related movement** as a centralised international index. The goal is to facilitate the access to and exploration of data sets that would otherwise be unknown, while supporting research collaboration at global, regional and local level. Moreover, this is also an opportunity to maximize the financial resources applied for data purchase by avoiding buying duplicate data from companies.
2. **Data reviewing across disciplines** The catalogue should be a forum for the community from different fields to put data sets into perspective with first-hand experience and related contextual data such as census data, fostering transparency on openness of data sets. It supports best practices for cross-research work. Data sets come with papers and code they have been used in, linking to contextual and supplementary material.
3. **Data collection waste reduction** The catalogue helps to foster research collaboration and to establish networks in order to reduce data collection redundancy and improve the use of the limited financial resources.
4. **Data fusion and multi-modality** The catalogue is a channel for data compilation where data sets on different transportation modes can be combined to answer specific research questions, while also encouraging multi-modal transportation research and planning.
5. **Methodological development** In its core, the catalogue is data and benchmark centered. However, it is supposed to become a central starting point for developing methodologies and state-of-the-art benchmarks and enhance reproducibility in human movement analytics studies across disciplines, in particular for bias detection. It supports discovering transferability opportunities from solutions derived from the data shared.

5.2.3 Mock-Ups

From the discussions held on how a mobility data catalog should look like, the overall motivation is to provide a mobility data platform which encourages open research and reproducibility.

Figure 2 show mock-up interfaces with different functionalities. Other than the typical search bar, a map with dataset locations and accessibility levels helps users to immediately have a sense of where mobility data is being collected, how to access them and the intensity of data collection activities in a certain location.

Upon clicking the dataset of interest, the next page offers a detailed description of the dataset which includes a summary, dataset attributes, benchmark, publications, comments and related datasets. In this way, users are able to understand how the dataset has been used for research and improvement opportunities from the publications that used the particular dataset. The related dataset offers possibility to validate findings and implement models or algorithms using other sources of data as well.

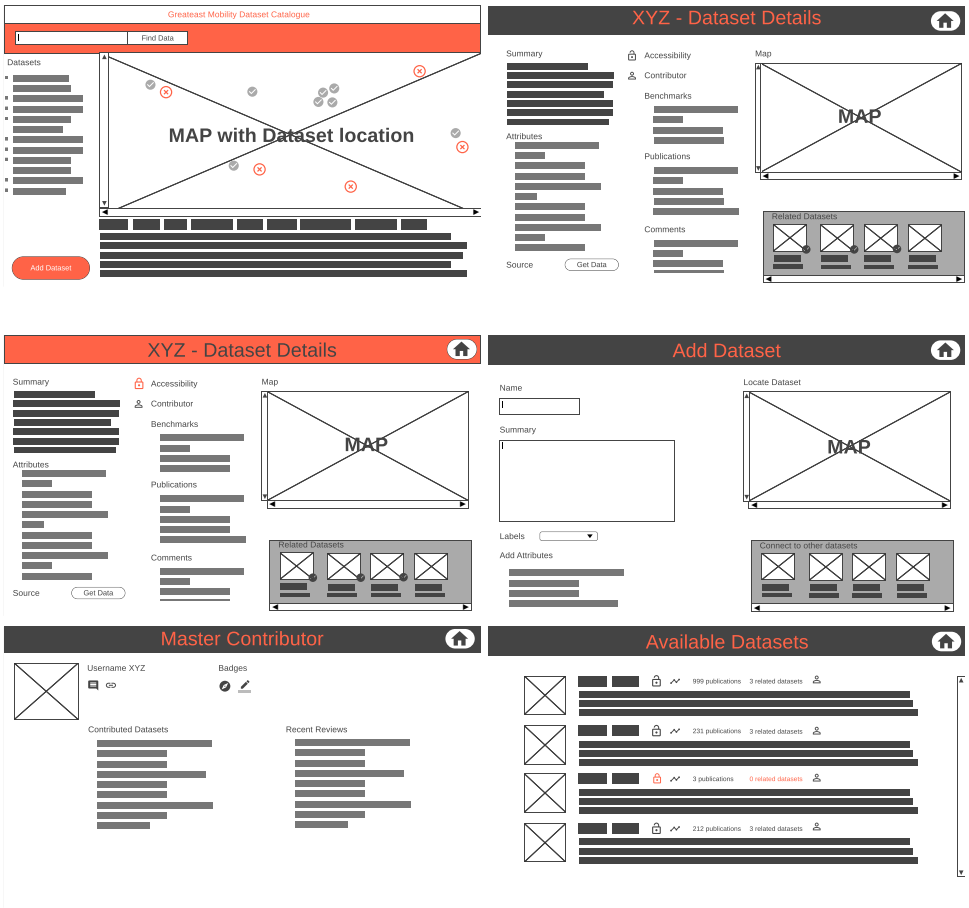

Figure 2 Mockup of Mobility Dataset Catalogue.

To encourage data contributions, contributor profiles include badges that acts as acknowledgements for offering the data and follow up badges for open research, etc.

Submitting data is easily done by a click of a button and a simple submission interface and finally an overview page provides a list of datasets with accessibility levels, number of views, number of publications, related datasets and a link to the contributor profile.

5.2.4 Discussion and Possible Extensions

Obviously, a catalogue is only as good as its users and curators. If it does not provide any insights or does not suit the way people work, it will be a non-starter. However, data repository and engineering workflow ("github for datasets") or an integrated analysis platform (see e.g. [6]) would mean a huge conceptual effort that requires engineering and platform resources, which is beyond our scope, hence, the reason why we envision the catalogue to be a more special-purpose combination of `https://paperswithcode.com/datasets` and `https://openreview.net/` known in the ML community.

References

1 Nathan, R., Getz Wayne, M., Revilla, E., Holyoak, M., Kadmon, R., Saltz, D. & Smouse Peter, E. A movement ecology paradigm for unifying organismal movement re-

search. *Proceedings Of The National Academy Of Sciences.* **105**, 19052-19059 (2008), https://doi.org/10.1073/pnas.0800375105

2 Brum-Bastos, V., Long, J. & Demšar, U. Weather effects on human mobility: a study using multi-channel sequence analysis. *Computers, Environment And Urban Systems.* **71** pp. 131-152 (2018)

3 Demšar, U., Long, J., Benitez-Paez, F., Brum-Bastos, V., Marion, S., Martin, G., Sekulić, S., Smolak, K., Zein, B. & Siła-Nowicka, K. Establishing the Integrated Science of Movement: bringing together concepts and methods from animal and human movement analysis. *International Journal Of Geographical Information Science.* **35**, 1273-1308 (2021)

4 Kranstauber, B., Cameron, A., Weinzerl, R., Fountain, T., Tilak, S., Wikelski, M. & Kays, R. The Movebank data model for animal tracking. *Environmental Modelling And Software.* **26**, 834-835 (2011)

5 Cagnacci, F., Boitani, L., Powell, R. & Boyce, M. Animal ecology meets GPS-based radiotelemetry: a perfect storm of opportunities and challenges.. *Philosophical Transactions Of The Royal Society Of London. Series B, Biological Sciences.* **365**, 2157-62 (2010,7)

6 Tomko, M., Bayliss, C., Galang, G., Greenwood, P., Koetsier, J., Mannix, D., Morandini, L., Nino-Ruiz, M., Pettit, C., Sarwar, M., Voorsluys, W., Widjaja, I., Stimson, R. & Sinnott, R. The AURIN e-Infrastructure: design, development and delivery. (2012)

5.3 Group on Methods and Explainability

Anita Graser (AIT – Austrian Institute of Technology – Wien, AT), Hao Cheng (Leibniz Universität Hannover, DE), Tao Cheng (University College London, GB), Ioannis Giannopoulos (TU Wien, AT), Daniel Nüst (Universität Münster, DE), Martin Tomko (The University of Melbourne, AU), and Yanan Xin (ETH Zürich, CH)

5.3.1 Introduction

Today's research and analytical methods, such as machine learning, rely heavily on tooling (software and scripts) capturing analysis as *code*, and on data. Any computational research poses strong challenges for effective communication and collaboration since both data and code are not well shared through the traditional method of scholarly communication, the scientific paper. If research outcomes cannot be shared, they can not transparently inform practices to improve mobility management in society.

The framework presented in Figure 3 framed the discussion in this group. Applicable to any computational research, this framework ties together the requirements on data (further discussed by a separate breakout group) and analysis.

Currently, many computational mobility analysis methods are de-facto *black boxes*. This may be either because the methods are only vaguely described in papers (often to an insufficient level of formality or because these methods are non-trivial to be then implemented in a computational environment) or because of the nature of the methods themselves, (i.e., the hard to explain and interpret computational procedures learned from specific datasets, such as in the case of deep neural networks). In cases where no human directly implemented assumptions and models, but a computer derived (learned) patterns from datasets, *explainability* must be especially considered to provide openness and transparency. Lack of explainability can result in erosion of trust from users who may question the results.

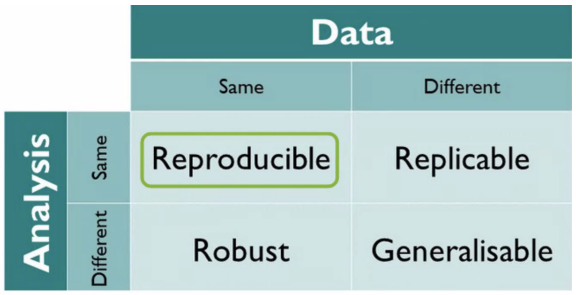

Figure 3 The reproducibility decision space.

This issue is also related to ethics because resulting recommendations may discriminate against certain groups due to biased training data or other issues. Furthermore, without explainability, theoretical frameworks cannot advance and therefore, scientific progress is slowed down.

We argued that a comprehensive publication of methods and their explainability is required to enable understanding, reuse, and extension of pieces of research so that societal challenges can be tackled effectively. Methods and the explainability of methods were discussed in this group along with possible avenues for improvement as a move towards better transparency: improvement in generalisability and robustness of methods, the ability to re-use the state of the art, and thus advance the discipline in a transparent manner.

5.3.2 Mobility Analytics Methods Research Agenda

The group collected topics broadly reflecting on current methods underpinning mobility analysis. This list naturally shows the breadth of the methodological toolkit needed to model and analyse mobility:

1. Explainability and transferability in deep learning models for human mobility [1];
2. Data integration, incl. geospatial encodings and embeddings for AI and challenges around vector vs. raster data;
3. How to separate *outlier* patterns of interest in large datasets from large amounts of *standard* situations;
4. Integrating physics-based models with data-driven models for analyzing mobility data to achieve more interpretable results;
5. Investigation of trajectory analytical methods for modeling travel behaviours using network-based trajectory data rather than grid-based mobility data;
6. The challenge of analysing interactions between modelled individuals, who are often assumed to be independent by mobility predictors and generators;
7. Models that capture the complexity of mobility more comprehensively, such as the complexities of dynamic structures of transport networks and multi-modal routing;
8. How to comprehensively model urban mobility across modalities (road traffic, public transport, pedestrian/bike), incl using deep learning (DL); and
9. Multi-modal networks performance optimization.

To identify missing methods that need to be developed, domain knowledge and context information are of vital importance. Overall, the design of methods should be further improved to satisfy the theoretical constraints of mobility patterns, instead of just relying on training complex models and large amounts of data to lead to usable outcomes.

5.3.3 Advancing the Scientific Mobility Analytics Toolbox

Participants with a background in scientific software development emphasized the need for a *consolidation in tool development* based on replicable and reusable software principles, thus enabling *reference implementations* of critical partial methods. This would enable us to compare, consolidate, and advance the discipline that is currently fragmented across tools and approaches[5].

A consolidation of tools and their open publishing is a prerequisite for more collaborative and effective development of tools, where tool extension and improvement to fix real world problems are valued more than marginal methodological improvements. Understanding and theory building should be the goal, i.e., scientific progress, not merely addressing an engineering challenge.

The group also discussed possible incentives to contribute to joint efforts enabling *methodological explainability*. Participants were interested in inspecting DL models and increasing the understanding of model behaviour. A key question is: what methods exist to explain the inner workings of DL models but also their inputs. Potential explainability approaches include: transfer of methods from computer vision, model comparison (especially for simple/simplified models), or visualisation of intermediate layers for understanding outputs.

Understanding and interpretability were also discussed on a more abstract level, that is: What kind of interpretability do we want for which type of analysis? Do we want to understand the mechanisms for generating the predictions, or make the actual model understandable to humans?

The participants share a quite critical view of the current state of methods in urban mobility analysis. There are concerns about the usefulness, usability, and relevance of methods, especially modern yet complex and possibly questionable methods based on AI and DL. These shared concerns motivate the following definition of challenges and recommendations.

5.3.4 Challenges in Developing Methods for Mobility Analysis

- **Challenge 1: defining appropriate objectives to evaluate machine learning models**
  High accuracy alone is often inadequate to ensure accountable and reliable deployment of models in practice. We also need to offer explanations to end-users as to why the model produces a certain prediction (such as a routing suggestion) to guarantee transparency, fairness, and reliability [2]. In these cases, explainability should also be considered as one of the objectives besides accuracy. But explainability is a fuzzy concept and can mean various things in different contexts. This makes it an extremely challenging objective to optimize for [3]. Whether an explanation achieves its desired goal also depends on who receives it, how the person perceives it, and what actions a person can take based on the explanation.
- **Challenge 2: bridging a theory-driven approach with a data-driven approach**
  It is still unclear how to effectively combine the data-driven approach with the theory-driven approach in mobility analysis. A key barrier to the integration is a lack of shared data, tools, and terminologies across different research communities working on mobility analysis (e.g., city planning, transportation engineering, geography, computer science, etc.).

[5] See a list at `https://github.com/anitagraser/movement-analysis-tools`

5.3.5 Recommendations

To address the above mentioned challenges, we have identified the following steps revolving around the development of shared resources for mobility analysis:

- **Recommendation 1: joint terminology**
  Fundamental terminological glossaries are needed to enable communication within the community. This terminological glossary would also relate to best practice about *methodological* translation of a mobility/transport concept (stop, home location, activity, behaviour) to the computational methods realising it, and an argument about their appropriateness in a given setting.
- **Recommendation 2: reference implementations**
  Standard, community-verified implementations (cross-linguistic, possibly, to enable implementation in the main programming languages) would enable contributors to build robust solutions.
- **Recommendation 3: standard dataset**
  A set of simple to more complex shared datasets (e.g., basic trajectories, semantic trajectories, related urban structure data) are needed to test and evaluate computational methods – see recommendations of Group 2.
- **Recommendation 4: standard data formats**
  A definition of a common data representation for movement data would enable an interoperable (incl multi-language) interface to the data (see, e.g., the OGD MovingFeatures standard or the standardisation on input formats in geoparquet for python/geopandas and R).
- **Recommendation 5: shared ML models**
  Reference implementations of basic ML methods and a platform for sharing benchmarks or established models built on top of these methods could become a community convergence point (similar to *Hugging Face*[6] *in the ML community).*

References

1 Luca, M., Barlacchi, G., Lepri, B. & Pappalardo, L. A Survey on Deep Learning for Human Mobility. *ACM Computing Surveys.* **55**, Article 7 (2021)

2 Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. *ArXiv Preprint ArXiv:1702.08608.* (2017)

3 Lipton, Z. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery.. *Queue.* **16**, 31-57 (2018)

[6] Hugging Face: `https://huggingface.co/`

5.4 Group on Applications

Martin Lauer (KIT – Karlsruher Institut für Technologie, DE), Andris Clio (Georgia Institute of Technology – Atlanta, US), Dirk Christian Mattfeld (TU Braunschweig, DE), and Monika Sester (Leibniz Universität Hannover, DE)

5.4.1 Goal

The goal of this discussion group was to generate ideas in which results of urban mobility analytics can be used in applications and how the respective application areas can be linked to urban mobility analytics. After a brainstorming phase the group concentrated on the fact that the knowledge of mobility behavior is of very much use in the domain of autonomous driving, in which autonomous vehicles interact with other traffic participant in urban traffic. To achieve an optimal behavior, the autonomous vehicles require knowledge about the behavior of other traffic participants to adapt their own behavior and to interact seamlessly with others. It is expected that for a long time human driven vehicles, autonomous vehicles, bicyclists, pedestrians and other traffic participants will share the same roads; therefore, models of pedestrians' behavior, bicyclists' behavior, and the behavior of human drivers is very much relevant. Furthermore, the behavior of traffic participants cannot be assumed to be the same worldwide but it differs from country to country. E.g. the driving style in Italy obviously differs from the driving style in Germany, the U.S. or in India or China. Similarly, pedestrians behave differently in those country. Hence, the autonomous vehicle must be able to understand the country specific behavior of other traffic participants, and also adapt their behaviour accordingly.

In the following, the aspects data, analysis methods, and necessary stakeholders were discussed, which would be needed to push forward such an application.

5.4.2 Data

The country-specific adaptation of behavior models can be considered as a typical application of machine learning methods, which are based on revealing this information from large amounts of observed data. So the question is how these data can be generated. Several data sources have been discussed in the group, including

- the usage of existing datasets like *highD*[7], *INTERACTION*[8], *LUMPI*[9]
- datasets for environmental perception (e.g. Kitti, nuScenes)
- recording own datasets from top view perspectives using drones or mounting cameras on high buildings
- recording own datasets from test fields, e.g. Testfeld Karlsruhe[10]

Besides these open data sets, commercial data would be very beneficial. Data schould be available in different European cities and potentially in further cities outside of Europe.

Since it is impossible to record and model the behavior of traffic participants in all possible traffic situations it would be necessary to start with limiting the recordings to some most interesting, and critical, scenarios. These could include

[7] highD: `https://www.highd-dataset.com/`
[8] INTERACTION: `http://interaction-dataset.com/`
[9] LUMPI: `https://data.uni-hannover.de/cs_CZ/dataset/lumpi`
[10] Testfeld Autonomes Fahren Baden-Württemberg `https://taf-bw.de/`

- merging into highway ramps
- shared spaces
- crossing intersections
- vehicle following

To model these scenarios it would be necessary to extract relevant features from the data including

- trajectory data (position over time, velocity, acceleration)
- relationship data (relative distances, relative velocities between traffic participants, interactions)
- movement patterns (e.g. groups)
- typical patterns at certain locations
- geometric data (shape of traffic participants, layout of road infrastructure, lanes, curbs)
- intentional data (face and body orientation of pedestrians, roll angle of bicyclists, indicator lights)

5.4.3 Data Analysis Methods

The analysis of the data requires automatic methods and processes, e.g. the extraction of the features mentioned above from the sensory data, the creation of behavior models for traffic participants, and the integration of those models into the decision making and trajectory planning system of the autonomous vehicle. Urban mobility analytics comes in play especially for the second category of methods that analyze the recorded mobility data. Various methods might be suitable, including deep learning methods, transformers, recurrent neural networks, inverse reinforcement learning. In addition, also methods from computational geometry, e.g. trajectory pattern analysis, are relevant. An important aspect is to include the notion of cooperation into the process, which allows the autonomous vehicle to also exploit information from other traffic participants, which enlarges its own perception range.

For the integration of behavior models to decision making and trajectory planning of the autonomous vehicle it would be useful to have a powerful simulation environment available, that also should be adapted to country specific behavior. Finally, for a real live demonstration the implementation of the behavior models in a real autonomous vehicle would be desirable.

5.4.4 Stakeholders

A joint project to extract and exploit country specific behavior patterns for autonomous vehicles could be relevant for various partners in research and industry. The following stakeholders could be included

- researchers in urban mobility analytics: analyze data, build behavior models, extract country specific parameters; develop models and concepts that allow for exploiting shared data, without the need to make it publicly available (e.g. using federated learning ([1])
- researchers in perception: record sensory data and extract relevant features
- researchers in traffic psychology: analyze and explain country specific behavior patterns
- researchers in autonomous driving: provide autonomous vehicle, integrate behavior models in decision making of autonomous vehicles
- automobile industry: integrate country specific behavior models into their vehicles, provide data from onboard sensors
- data and map companies: collect and analyze country specific data; specify map interfaces for country specific behavior patterns
- municipalities: provide access to infrastructure, equip certain areas with sensors, implement certain regulations (e.g. temporal speed limits)

References

1 Koetsier, C., Fiosina, J., Gremmel, J., Müller, J., Woisetschläger, D. & Sester, M. Detection of anomalous vehicle trajectories using federated learning. *ISPRS Open Journal Of Photogrammetry And Remote Sensing*. **4** pp. 100013 (2022), https://www.sciencedirect.com/science/article/pii/S2667393222000023

5.5 Outcomes

Monika Sester (Leibniz Universität Hannover, DE), Martin Tomko (The University of Melbourne, AU), and Stephan Winter (The University of Melbourne, AU)

The outcomes of the breakout group discussions were collected and shared on Friday, and further discussed.

In this session, the industry partners rated the cooperation and exchange with universities as very relevant to them. The caveat is usually, that collaboration attempts within industry are not considered as mainstream by all employees yet, but heavily depends on convinced individuals. This often leads to only ad-hoc projects – unless there is already an established long-standing cooperation based on contracts. It was also pointed out that industry partners could be very relevant for tool development, or that a Dagstuhl Seminar focusing on an innovative tool could be very attractive for industry.

As a way forward, the participants committed to:

- Extracting a vision paper out of this report, for a venue such as the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems 2022
- Developing tools and teaching material for a Summer School in Spring 2023
- A potential follow-up Dagstuhl Seminar closer focusing either on the core of mobility analytics or on the application of the emerging body of knowledge for climate-neutral urban mobility
- In addition, an number of individual collaboration themes were identified (internships, co-supervisions, visits).

Participants

- Vanessa Brum-Bastos
Wroclaw University of Environmental and Life Sci., PL
- Hao Cheng
Leibniz Universität Hannover, DE
- Tao Cheng
University College London, GB
- David Doerr
TÜV Rheinland – Köln, DE
- Christian Eichenberger
IARAI – Zürich, CH
- Cheng Fu
Universität Zürich, CH
- Martin Lauer
Karlsruher Institut für Technologie, DE
- Dirk Christian Mattfeld
TU Braunschweig, DE
- Alexandra Millonig
Austrian Institute of Technology – Wien, AT
- Edoardo Neerhut
Meta – Burlingame, US
- Moritz Neun
IARAI – Zürich, CH
- Daniel Nüst
Universität Münster, DE
- Erik Nygren
Schweizerische Bundesbahnen – Bern, CH
- Maya Sekeran
TU München, DE
- Monika Sester
Leibniz Universität Hannover, DE
- Martin Tomko
The University of Melbourne, AU
- Stephan Winter
The University of Melbourne, AU
- Yanan Xin
ETH Zürich, CH

Remote Participants

- Andris Clio
Georgia Institute of Technology – Atlanta, US
- Ioannis Giannopoulos
TU Wien, AT
- Anita Graser
Austrian Institute of Technology – Wien, AT
- Ivan Majic
TU Graz, AT
- Michael Nolting
Volkswagen Nutzfahrzeuge – Hannover, DE
- Luca Pappalardo
CNR – Pisa, IT
- Chiara Renso
ISTI-CNR – Pisa, IT
- Piyushimita Vonu Thakuriah
Rutgers University – New Brunswick, US

Report from Dagstuhl Seminar 22171

Digital Twins for Cyber-Physical Systems Security

Alvaro Cárdenas Mora[*1], Simin Nadjm-Tehrani[*2], Edgar Weippl[*3], and Matthias Eckhart[†4]

1 **University of California – Santa Cruz, US.** alacarde@ucsc.edu
2 **Linköping University, SE.** simin.nadjm-tehrani@liu.se
3 **Universität Wien, AT.** edgar.weippl@univie.ac.at
4 **SBA Research – Wien, AT.** meckhart@sba-research.org

Abstract

Cyber-physical systems (CPSs) may constitute an attractive attack target due to the increased networking of components that yields an expanded attack surface. If their physical control capabilities are compromised, safety implications may arise. Thus, it is vital that the CPSs being engineered are thoroughly tested and that adequate response measures can be realized upon detecting intruders during operation. However, security testing is hard to conduct due to expensive hardware, limited maintenance periods, and safety risks. Furthermore, the increased stealthiness of threat actors requires new intrusion detection and response methods. Interestingly, digital twins have become an important concept in industrial informatics to solve similar problems, yet with a non-security-related focus: Digital twins that virtually replicate the real systems provide cost-efficient modeling, testing, monitoring, and even predictive capabilities. However, until recently, the digital-twin concept has mainly focused on production optimizations or design improvements without considering its potential for CPS security. The Dagstuhl Seminar 22171 "Digital Twins for Cyber-Physical Systems Security" therefore aimed to serve as an interdisciplinary, open knowledge-sharing platform to investigate the benefits and challenges of applying the digital-twin concept to improve the security of CPSs.

Seminar April 24–29, 2022 – http://www.dagstuhl.de/22171
2012 ACM Subject Classification Security and privacy → Intrusion/anomaly detection and malware mitigation; Computer systems organization → Embedded and cyber-physical systems
Keywords and phrases cyber-physical systems, digital twins, information security, production systems engineering, SCADA, industrial control systems, Industry 4.0
Digital Object Identifier 10.4230/DagRep.12.4.54

1 Executive Summary

Matthias Eckhart (SBA Research – Wien, AT, meckhart@sba-research.org)
Alvaro Cárdenas Mora (University of California – Santa Cruz, US, alacarde@ucsc.edu)
Simin Nadjm-Tehrani (Linköping University, SE, simin.nadjm-tehrani@liu.se)
Edgar Weippl (University of Vienna & SBA Research – Wien, AT, edgar.weippl@univie.ac.at)

In the light of the increasing digitization and move toward Industry 4.0 [1], cyber security becomes more and more important for cyber-physical systems (CPSs). The advanced computation, communication, and control capabilities of CPSs lead to a wider attack surface and greater exposure to security flaws. Furthermore, the added complexity puts

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

a considerable burden on security professionals, who have to ensure that the CPSs are adequately protected against adversaries throughout the entire lifecycle. As a matter of fact, designing holistic security measures is a significant ongoing challenge for academia and industry alike. Thorough security testing during the engineering- and, particularly, the operation phase is often not feasible. The development of custom CPS testbeds is complicated, expensive, and time-consuming due to high hardware costs, space constraints, and complex dependencies between components [2]. Past attempts to conduct penetration tests directly on live systems led to unintended system behavior, putting human workers in significant danger and causing a disruption of production lines [3]. In addition to regular security testing, adequate countermeasures need to be implemented in response to newly discovered vulnerabilities that emerge during operation or if the CPS is already under attack. However, the steadily increasing sophistication of cyberattacks calls for more effective intrusion detection and prevention techniques. On top of that, new mechanisms to test and evaluate attack response strategies in a controlled setting are required.

A digital twin, that is, a virtual replica of a real system, was originally envisioned for similar, yet non-security-related purposes: The life of a spacecraft is virtually mirrored through high-fidelity simulations and sensor updates to detect anomalies and safely test mitigation options such that degradation can be reduced and damages prevented [4]. This idea was picked up by the industrial informatics community, whose members implemented the digital-twin concept in various CPS applications for monitoring, lifecycle management, and decision support [5, 6, 7]. In the past few years, researchers have also shown interest in utilizing digital twins for security-enhancing purposes [8, 9, 10, 11, 12, 13]. Although the definition of what constitutes a digital twin in the context of cybersecurity differs in the literature, its main application areas seem to be clear: Virtually replicated systems by means of emulation, simulation, and modeling technologies, coupled with real-time or historical data flows, might be used to improve security testing, intrusion detection, and attack recovery. However, fundamental research questions and challenges remain before digital twins can be applied for security-enhancing purposes. Furthermore, concerns have been raised about the potential security threats associated with the digital-twin concept [14].

Thus, the primary goal of this Dagstuhl Seminar was to lay the foundation for future interdisciplinary collaboration on digital-twin research for CPS security. The interdisciplinary character of this novel research area is reflected in its origin. As already indicated, the notion of using "twins" originally emerged from the space industry [6], gained wider adoption by the industrial informatics community [5, 6, 7], and was eventually applied with the objective of attaining security improvements [8, 9, 10, 11, 12, 13]. For this reason, the seminar has brought together 20 researchers with backgrounds in computer security, control theory, automation engineering, and data science. Inspired by the concept's promised security improvement potential, the seminar was structured along three different themes:

Foundations of Security-focused Digital Twins. This theme was motivated by the lack of clarity around the digital-twin concept. Therefore, the purpose of this theme was to develop a common understanding of what a digital twin in the context of security is, how it can be defined, and how it relates to existing concepts, such as cyber ranges, data-driven models, and honeypots. Closely tied to this theme were discussions on methods for digital-twin implementation, including (i) emulating systems and simulating physical processes, (ii) knowledge retrieval for digital-twin generation in greenfield and brownfield environments, and (iii) synchronizing digital twins with their physical counterparts.

Intrusion Detection. The objective of this theme was to explore intrusion detection as a potential use case for digital twins. Assuming that the digital twin is built from a benign specification such that legitimate behavior is exhibited when executed in sync with its counterpart, any deviations observed on the logic, network, and physics layers could indicate malicious activity. Building on this idea, participants discussed how digital twins can serve as a foundation for such behavior-specification-based intrusion detection systems (IDSs) that possess physics- and process-aware capabilities. Moreover, discussions touched on how digital twins can be used for data generation purposes to improve the training phase of (semi-)supervised learning approaches that are employed in behavior-based IDSs.

Attack Response Mechanisms. The last theme was associated with research questions on implementing proactive and reactive attack response strategies, which may represent another use case of digital twins. Proactive security measures can prevent cyber-physical attacks in the face of imminent threats when new vulnerabilities in the CPS are discovered. On the other hand, reactive responses to an attack can be initiated to control damage by ensuring that the physical system maintains a safe state. In this context, questions were raised about how the digital-twin concept can help in designing attack-resilient CPS architectures and response strategies for control systems. This theme highlighted the benefits and challenges of using digital twins to test countermeasures in a simulated environment and assess their effects.

The program started with a welcome session that provided an opportunity for participants to get to know one another. Furthermore, the organizers used this session to share information about the seminar program and explain key terms to participants who were not au fait with the terminologies used by different communities. Over the five days, 14 participants gave lightning talks that focused on the following topics:

- building blocks for digital-twin construction, including emulating and simulating CPS components, data-driven approaches and semantic technologies, synchronization mechanisms,
- reverse engineering programmable logic controllers, deception technology (e.g., honeypots), security testbeds,
- attack detection in CPSs, featuring physics-based, data-driven, and process-aware techniques,
- attack-resilient control using different tools for risk mitigation (viz., prevention, detection, and treatment),
- various aspects of dataset availability in CPS research (e.g., attack simulation, data collection, evaluation, and validation), and
- digital-twin use cases for the safety-related system development lifecycle.

The lightning talk sessions offered each speaker 15 minutes to present new perspectives and talk about current challenges in CPS security. The highly interdisciplinary setting and stimulating presentations given by participants resulted in active discussions, which were carried on in the breakout sessions.

The afternoons of Monday, Tuesday, and Wednesday were used for breakout sessions to give participants the opportunity to work together on research issues of common interest. Based on the discussions that took place on Monday after the session on bridging the disciplinary gap, we identified the following topics of interest to be explored by working groups: (i) conceptualization of the digital twin for cyber-physical systems security, and (ii) attack recovery for control systems. Participants who worked on the former topic discussed

characteristics that digital twins need to have to be useful for security applications, while those who focused on the latter topic investigated strategies in the context of control theory to respond to attacks in a reactive manner.

The seminar received very positive feedback from participants, who also expressed strong interest in future editions. In addition, several invitees, who were forced to cancel their participation at short notice due to the SARS-CoV-2 pandemic, have shown great interest in follow-up events. Thus, we believe that this Dagstuhl Seminar should be repeated in the future. A second edition would be worthwhile to investigate open problems concerning system emulation. These issues could be addressed in a future follow-up seminar if more participation from the embedded systems and systems security communities is achieved.

As the organizers, we would like to thank everyone who attended this seminar for their interesting talks, the thought-provoking questions, and the fruitful contributions that led to a highly collaborative atmosphere for scientific discussions. We also would like to express our sincere gratitude to the scientific and administrative staff of Schloss Dagstuhl for their outstanding support that made this seminar possible.

References

1 Henning Kagermann, Johannes Helbig, Ariane Hellinger, and Wolfgang Wahlster. Recommendations for implementing the strategic initiative INDUSTRIE 4.0 – securing the future of german manufacturing industry. Final report of the Industrie 4.0 working group, acatech – National Academy of Science and Engineering, München, April 2013.

2 Benjamin Green, Anhtuan Lee, Rob Antrobus, Utz Roedig, David Hutchison, and Awais Rashid. Pains, gains and PLCs: Ten lessons from building an industrial control systems testbed for security research. In *10th USENIX Workshop on Cyber Security Experimentation and Test (CSET 17)*, Vancouver, BC, 2017. USENIX Association.

3 David Duggan, Michael Berg, John Dillinger, and Jason Stamp. Penetration testing of industrial control systems. *Sandia National Laboratories*, 2005.

4 Mike Shafto, Mike Conroy, Rich Doyle, Ed Glaessgen, Chris Kemp, Jacqueline LeMoigne, and Lui Wang. Draft modeling, simulation, information technology & processing roadmap. *Technology Area*, 11, 2010.

5 Elisa Negri, Luca Fumagalli, and Marco Macchi. A review of the roles of digital twin in CPS-based production systems. *Procedia Manufacturing*, 11:939 – 948, 2017. 27th International Conference on Flexible Automation and Intelligent Manufacturing, FAIM2017, 27-30 June 2017, Modena, Italy.

6 Roland Rosen, Georg von Wichert, George Lo, and Kurt D. Bettenhausen. About the importance of autonomy and digital twins for the future of manufacturing. *IFAC-PapersOnLine*, 48(3):567 – 572, 2015. 15th IFAC Symposium on Information Control Problems in Manufacturing INCOM 2015.

7 Werner Kritzinger, Matthias Karner, Georg Traar, Jan Henjes, and Wilfried Sihn. Digital twin in manufacturing: A categorical literature review and classification. *IFAC-PapersOnLine*, 51(11):1016 – 1022, 2018. 16th IFAC Symposium on Information Control Problems in Manufacturing INCOM 2018

8 Matthias Eckhart and Andreas Ekelhart. *Digital Twins for Cyber-Physical Systems Security: State of the Art and Outlook*, chapter 14, pages 383–412. Springer International Publishing, Cham, 2019.

9 Mariana Segovia and Joaquin Garcia-Alfaro. Design, modeling and implementation of digital twins. *Sensors*, 22(14), 2022.

10 Marietheres Dietz and Gunther Pernul. Unleashing the digital twin's potential for ICS security. *IEEE Security & Privacy*, 18(4):20–27, July 2020.

11 David Holmes, Maria Papathanasaki, Leandros Maglaras, Mohamed Amine Ferrag, Surya Nepal, and Helge Janicke. Digital twins and cyber security – solution or challenge? In *2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, pages 1–8, September 2021.

12 Rajiv Faleiro, Lei Pan, Shiva Raj Pokhrel, and Robin Doss. Digital twin for cybersecurity: Towards enhancing cyber resilience. In Wei Xiang, Fengling Han, and Tran Khoa Phan, editors, *Broadband Communications, Networks, and Systems*, pages 57–76, Cham, 2022. Springer International Publishing.

13 Abhishek Pokhrel, Vikash Katta, and Ricardo Colomo-Palacios. Digital twin for cybersecurity incident prediction: A multivocal literature review. In *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops*, pages 671–678, New York, NY, USA, 2020. Association for Computing Machinery.

14 Cristina Alcaraz and Javier Lopez. Digital twin: A comprehensive survey of security threats. *IEEE Communications Surveys & Tutorials*, 2022.

2 Table of Contents

3 Overview of Talks

3.1 Dataset availability and requirements for CPS security research

Magnus Almgren (Chalmers University of Technology – Göteborg, SE)

Joint work of Magnus Almgren, Wissam Aoudi, Mikel Iturbe
Main reference Wissam Aoudi, Mikel Iturbe, Magnus Almgren: "Truth Will Out: Departure-Based Process-Level Detection of Stealthy Attacks on Control Systems", in Proc. of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018, Toronto, ON, Canada, October 15-19, 2018, pp. 817–831, ACM, 2018.
URL https://doi.org/10.1145/3243734.3243781

One of the challenges of CPS security research is validating the results, be it through a dataset or by using a real(-alistic) system. The first challenge is to find or create a system or dataset containing the indicators that are used in the algorithm. The second challenge is then to demonstrate different properties: true positives, false positives, true negative, false negatives. The third challenge is then to argue that the attacks or the system under study are realistic, preferably created by someone outside of the research group. One might also need to be able to show a certain set of robustness of the system. All of the above are challenges when it comes to any sort of validation, but more so when it concerns CPS of a societal value.

In the talk, I will outline these challenges by using as a case study the process we used for validating the system presented at CCS 2018.

3.2 Modelling in the Safety Lifecycle of Radiation Monitoring Systems at CERN

Katharina Ceesay-Seitz (CERN – Meyrin, CH)

Joint work of Katharina Ceesay-Seitz, Hamza Boukabache, Daniel Perrin, Gael Ducos, Sarath Kundumattathil-Mohanan, Amitabh Yadav

CERN, the European Organisation for Nuclear Research, operates the world's largest particle accelerator and many other high energy physics experiments. These experiments produce ionizing radiation, for example when particles hit stable matter. The radiation protection group is responsible for protecting humans from any unjustified radiation exposure. The CERN RadiatiOn Monitoring Electronics (CROME) are the new generation of instruments built for measuring ionizing radiation levels and triggering alarms and machine interlocks based on these measurements [1].

Models of subsystems are used throughout the safety lifecycle of CROME. Physics simulations were used to model the expected radiation levels in different zones. Based on these simulations and on the envisioned use cases the system's functional and safety requirements were defined. Models of subsystems were used throughout the design phase for interoperability and testing purposes.

The CROME Measuring and Processing Unit consists of a radiation detector and an electronic system for data communication and storage, signal processing and safety-related decision taking. It houses a Zynq-7000 System-on-Chip (SoC) consisting of a dual-core ARM processor and an FPGA section. The ARM cores execute an embedded Linux and an

application that receives around 150 parameters via a custom TCP/IP based communication library [2] form a SCADA system [3]. An independent test tool has been developed to model the librarie's functionalities [4]. It has been used to strengthen the robustness of the design by sending malformed messages to CROME and observing its response.

The parameters, which can be floating point variables or integers with ranges up to 64 bit, or others, are processed and sent to the FPGA, which performs all safety critical calculations and decision making. It calculates the radiation dose received in a given time as well as the dose rate from the input received from the radiation detector. Based on these measurements and the current parameter configuration, it can autonomously trigger alarms and machine interlocks. Models with different levels of abstraction are used to verify the functionality of the system. Constrained-random simulation has been used to simulate a large state space, which led to the discovery of several faults. Simulation only covers a subset of the possible states. Many additional faults have been found with formal verification, even in scenarios that were impossible to simulate due to the real-time nature of the system [5]. Formal verification has also been successfully used for the partial verification of a prototype of the future frontend of CROME, the ACCURATE 2 ASIC for ultra-low current measurement [7, 6].

This talk presents the different modelling approaches and discusses potential use cases for digital twins.

References

1 Hamza Boukabache, Michel Pangallo, Gael Ducos, Nicola Cardines, Antonio Bellotta, Ciarán Toner, Daniel Perrin, and Doris Forkel-Wirth. Towards a novel modular architecture for CERN radiation monitoring. *Radiation Protection Dosimetry*, 173(1-3):240–244, November 2016.

2 Amitabh Yadav, Hamza Boukabache, Katharina Ceesay-Seitz, Nicola Gerber, and Daniel Perrin. ROMULUSLib: An autonomous, TCP/IP-based, multi-architecture C networking library for DAQ and control applications. *Proceedings of the 18th International Conference on Accelerator and Large Experimental Physics Control Systems*, ICALEPCS2021:69–76, 2022.

3 Adrien Ledeul, Alexandru Savulescu, Gustavo Segura, Bartlomiej Styczen, and Daniel Vazquez Rivera. CERN supervision, control and data acquisition system for radiation and environmental protection. *Proceedings of the 12th Int. Workshop on Emerging Technologies and Scientific Facilities Controls*, PCaPAC2018:248–252, 2019.

4 Katharina Ceesay-Seitz, Hamza Boukabache, Marvin Leveneur, and Daniel Perrin. RomLibEmu: Network interface stress tests for the CERN radiation monitoring electronics (CROME). *Proceedings of the 18th International Conference on Accelerator and Large Experimental Physics Control Systems*, ICALEPCS2021:581–585, 2022.

5 Katharina Ceesay-Seitz, Hamza Boukabache, and Daniel Perrin. A functional verification methodology for highly parametrizable, continuously operating safety-critical FPGA designs: Applied to the CERN RadiatiOn monitoring electronics (CROME). In António Casimiro, Frank Ortmeier, Friedemann Bitsch, and Pedro Ferreira, editors, *Computer Safety, Reliability, and Security*, pages 67–81, Cham, 2020. Springer International Publishing.

6 Katharina Ceesay-Seitz, Sarath Kundumattathil Mohanan, Hamza Boukabache, Daniel Perrin, and Hamza Boukabache. Formal property verification of the digital section of an ultra-low current digitizer ASIC. In *Design and Verification Conference in Europe*, October 2021.

7 Sarath Kundumattathil Mohanan, Hamza Boukabache, Vassili Cruchet, Daniel Perrin, Stefan Roesler, and Ullrich R. Pfeiffer. An ultra low current measurement mixed-signal ASIC for radiation monitoring using ionisation chambers. *IEEE Sensors Journal*, 22(3):2142–2150, February 2022.

3.3 Digital Twins for CPS Security

Alvaro Cárdenas Mora (University of California – Santa Cruz, US)

In this talk we discuss the differences between IT and OT security, and how digital twins for physical systems are a natural component to address the new challenges of OT security.

Then we discuss our work on how digital twins can help in security by:

- Deploy new defenses such as attack recovery
- Understand the consequences of attacks and risks of CPS
- Interact with the adversary (Through honeypots or by executing malware in a contained setting)
- Finding new attacks in a principled manner (e.g., fuzzing the physical system).

3.4 A Roadmap Toward a Digital-Twin Framework for Cyber-Physical Systems Security: Vision, Recent Progress, and Open Challenges

Matthias Eckhart (SBA Research – Wien, AT)

Joint work of Matthias Eckhart, Andreas Ekelhart

The term "digital twin" is one of the latest technology buzzwords that has emerged along with the digital transformation that is taking place in CPS domains. Since there is no generally accepted definition of this term yet, the understanding of the digital-twin concept is often limited to the notion that a cyber-physical system is replicated in a digitally-enhanced way. This talk provides one interpretation of digital twins by breaking down the concept into four components that are required to implement them, viz., i) system emulation or via system containers, including I/O simulation, ii) network emulation, iii) interactive, real-time simulation of the physical process, and iv) synchronization with the physical counterparts. After putting the digital-twin concept into context, we present our current progress on developing a framework named CPS Twinning that integrates these four components for the purpose of generating such digital twins, so that security applications (e.g., intrusion detection) can be built on top. The talk concludes with an overview of open challenges and research opportunities in this area.

References

1 Matthias Eckhart and Andreas Ekelhart. *Digital Twins for Cyber-Physical Systems Security: State of the Art and Outlook*, chapter 14, pages 383–412. Springer International Publishing, Cham, 2019.
2 Matthias Eckhart and Andreas Ekelhart. A specification-based state replication approach for digital twins. In *Proceedings of the 2018 Workshop on Cyber-Physical Systems Security and PrivaCy*, CPS-SPC '18, pages 36–47, New York, NY, USA, 2018. ACM.
3 Matthias Eckhart, Andreas Ekelhart, and Roland Eisl. Digital twins for cyber-physical threat detection and response. *ERCIM News*, 2021(127), 2021.
4 M. Eckhart, A. Ekelhart, and E. Weippl. Enhancing cyber situational awareness for cyber-physical systems through digital twins. In *2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, pages 1222–1225, Sep. 2019.

5 Matthias Eckhart and Andreas Ekelhart. Towards security-aware virtual environments for digital twins. In *Proceedings of the 4th ACM Workshop on Cyber-Physical System Security*, CPSS '18, pages 61–72, New York, NY, USA, 2018. ACM.
6 Matthias Eckhart and Andreas Ekelhart. Securing cyber-physical systems through digital twins. *ERCIM News*, 2018(115), 2018.

3.5 Towards Semantically Enhanced Digital Twins

Helge Janicke (Cyber Security CRS – Joondalup, AU)

Joint work of Helge Janicke, David Holmes, Surya Nepal
Main reference David Holmes, Maria Papathanasaki, Leandros A. Maglaras, Mohamed Amine Ferrag, Surya Nepal, Helge Janicke: "Digital Twins and Cyber Security – solution or challenge?", in Proc. of the 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference, SEEDA-CECNSM 2021, Preveza, Greece, September 24-26, 2021, pp. 1–8, IEEE, 2021.
URL https://doi.org/10.1109/SEEDA-CECNSM53056.2021.9566277

Digital twin technology today is diverse and emerging and its full potential is not yet widely understood. The concept of a digital twin allows for the analysis, design, optimisation and evolution of systems to take place fully digital, or in conjunction with a cyber-physical system to improve speed, accuracy and efficiency when compared to traditional engineering approaches. Digital Twin technology is mainly used today as a digital replica of a physical system with the generated and observed data being used for applications such as predictive maintenance, fault analysis and optimisation. This is predominantly a data-driven approach that uses modern machine learning technologies to maximise the benefit of the available data. This talk proposes the semantic markup of digital twins to unlock the benefits of other aspects of Artificial Intelligence, namely semantic reasoning, to broaden the application facilitate deeper analysis of systems and their properties than can be achieved by analysing their data and behaviours through observation. The talk will explore potential synergies and barriers that need to be overcome for this approach to unlock future digital twin applications.

3.6 Detection of Cyber-Physical Attacks with IIoT data

Marina Krotofil (Maersk – Aarhus, DK)

Novel IIoT architectures such as NOA (NAMUR Open Architecture) allow for delivery of raw or high-resolution IIoT data via dedicated data highways. This data is used for various purposes such as developing digital twin models, predictive maintenance and augmented reality applications, etc. These data can also be used as a source of forensic artefacts or even evidence when investigating cyber-physical attacks. In this talk we will show a specific example of how IIoT data is used to detect an ongoing attack on an industrial pump and determine its root cause. We will leave the audience with an open question about the requirement to the collection, transport and storage of IoT data to ensure their utility to incident response and admissibility as legal evidence.

3.7 Control-theoretical Analysis of Systems under CPU Starvation Attacks

Martina Maggio (Universität des Saarlandes – Saarbrücken, DE)

Joint work of Martina Maggio, Martin Gunnarsson, Nils Vreman

Embedded systems and cyber-physical controllers have been proven vulnerable to security attacks of various nature, including man-in-the-middle attacks that alter sensor data and actuator commands, and attacks that disrupt the calculation of the control signals. While attack detection has been widely studied, countermeasures are scarce at best. We propose and implement a defence technique, based on executing the controller code in a trusted execution environment.

3.8 RICSel21: Data Collection from Attacks in a Virtual Power Grid

Simin Nadjm-Tehrani (Linköping University, SE)

Joint work of Simin Nadjm-Tehrani, Chih-Yuan Lin, August Fundin, Eric Westring, Tommy Gustavsson
Main reference Chih-Yuan Lin, August Fundin, Erik Westring, Tommy Gustafsson, Simin Nadim-Tehrani: "RICSel21 Data Collection: Attacks in a Virtual Power Network", in Proc. of the IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids, SmartGridComm 2021, Aachen, Germany, October 25-28, 2021, pp. 201–206, IEEE, 2021.
URL https://doi.org/10.1109/SmartGridComm51999.2021.9632328

In this talk I give an overview of the work done in one of the three tracks within the Swedish research centre on Resilient Information and Control Systems (RICS) [2]. The three tracks involve a) Data emulation b) Attack modelling and risk analysis, and c) Anomaly detection. The work on the Data emulation part has resulted in a national virtual testbed RICS-el for Supervisory Control and Data Acquisition (SCADA) security analysis in an electricity distribution network with a commercial SCADA software, some 20 emulated substations connected with wide area networks, OT, DMZ and IT segments. It has so far been exposed in two published works in collaboration with several colleagues [3, 1]. This talk focuses on the latest publication where 12 attacks were performed in the testbed and the outcomes documented. The dataset from the attack scenarios and the baseline (no-attack) counterpart is available for sharing.

References

1 Chih-Yuan Lin, August Fundin, Erik Westring, Tommy Gustafsson, and Simin Nadim-Tehrani. RICSel21: Data collection: Attacks in a virtual power network. In *2021 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pages 201–206, October 2021.
2 Simin Nadjm-Tehrani, Mathias Ekstedt, and Magnus Almgren. RICS: Research centre on resilient information and control systems, 2022. Availble: `https://www.rics.se/`
3 Magnus Almgren, Peter Andersson, Gunnar Björkman, Mathias Ekstedt, Jonas Hallberg, Simin Nadjm-Tehrani, and Erik Westring. RICS-el: Building a national testbed for research and training on SCADA security (short paper). In Eric Luiijf, Inga Žutautaitė, and Bernhard M. Hämmerli, editors, *Critical Information Infrastructures Security*, pages 219–225, Cham, 2019. Springer International Publishing.

3.9 Building High Fidelity Replicas for Cyber-Physical Systems Security Research – Lessons from a Testbeds Programme

Awais Rashid (University of Bristol, GB)

Joint work of Awais Rashid, Rob Antrobus, Barnaby Craggs, Joseph Gardiner, Benjamin Green, David Hutchison, Anhtuan Lee, Utz Roedig
Main reference Joseph Gardiner, Barnaby Craggs, Benjamin Green, Awais Rashid: "Oops I Did it Again: Further Adventures in the Land of ICS Security Testbeds", in Proc. of the ACM Workshop on Cyber-Physical Systems Security & Privacy, CPS-SPC@CCS 2019, London, UK, November 11, 2019, pp. 75–86, ACM, 2019.
URL https://doi.org/10.1145/3338499.3357355

Digital twins aim to provide an extensive and scalable means to model and evaluate properties of real-world systems. Developing such digital twins for cyber-physical systems is non-trivial even more so at a high enough fidelity in order to suitably replicate behaviours of real-world systems when compromised or under attack. In this talk, I will reflect on experiences of over 8 years of research building cyber-physical systems security testbeds particularly those to support security analyses of industrial control systems. I will discuss challenges arising from the need to represent a diversity of devices, networking mechanisms and software platforms as well as scalability of experimentation and managing the complexity of the testbed environment itself. I will reflect on what research on digital twins can learn from these experiences and the potential for "physical" testbed environments to work in tandem with digital twins.

References

1 Joseph Gardiner, Barnaby Craggs, Benjamin Green, and Awais Rashid. Oops i did it again: Further adventures in the land of ICS security testbeds. In *Proceedings of the ACM Workshop on Cyber-Physical Systems Security & Privacy*, CPS-SPC'19, pages 75–86, New York, NY, USA, 2019. Association for Computing Machinery.
2 Benjamin Green, Anhtuan Lee, Rob Antrobus, Utz Roedig, David Hutchison, and Awais Rashid. Pains, gains and PLCs: Ten lessons from building an industrial control systems testbed for security research. In *10th USENIX Workshop on Cyber Security Experimentation and Test (CSET 17)*, Vancouver, BC, 2017. USENIX Association.
3 Awais Rashid, Joseph Gardiner, Benjamin Green, and Barnaby Craggs. Everything is awesome! or is it? cyber security risks in critical infrastructure. In Simin Nadjm-Tehrani, editor, *Critical Information Infrastructures Security*, pages 3–17, Cham, 2020. Springer International Publishing.

3.10 Integrated distributed SCADA security in power grids

Anne Remke (Universität Münster, DE)

Joint work of Anne Remke, Verena Menzel, Johannn Hurinnk
Main reference Verena Menzel, Johann L. Hurink, Anne Remke: "Securing SCADA networks for smart grids via a distributed evaluation of local sensor data", in Proc. of the IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids, SmartGridComm 2021, Aachen, Germany, October 25-28, 2021, pp. 405–411, IEEE, 2021.
URL https://doi.org/10.1109/SmartGridComm51999.2021.9632283

Within smart grids the safe and dependable distribution of electric power highly depends on the security of Supervisory Control and Data Acquisition (SCADA) systems and their underlying communication protocols. Existing network-based intrusion detection systems for

Industrial Control Systems (ICS) are usually centrally applied at the SCADA server and do not take the underlying physical process into account. A recent line of work proposes an additional layer of security via a process-aware approach applied locally at the field stations. Currently, we broaden the scope of process-aware monitoring by considering the interaction between neighboring field stations, which facilitates upcoming trends of decentralized energy management (DEM). Local security monitoring is lifted to monitoring neighborhoods of field stations, therefore achieving a broader grid coverage w.r.t. security. We provide a distributed monitoring algorithm of the generated sensory readings for this extended setting. The feasibility of the approach is shown via a prototype simulation testbed and a scenario with two subgrids.

3.11 Attack-resilient control using model- and data-based intrusion detection

Henrik Sandberg (KTH Royal Institute of Technology – Stockholm, SE)

Joint work of Henrik Sandberg, Kaveh Paridari, Niamh O'Mahony, Alie El-Din Mady, Rohan Chabukswar, Menouer Boubekeur, David Umsonst
Main reference Kaveh Paridari, Niamh O'Mahony, Alie El-Din Mady, Rohan Chabukswar, Menouer Boubekeur, Henrik Sandberg: "A Framework for Attack-Resilient Industrial Control Systems: Attack Detection and Controller Reconfiguration", Proc. IEEE, Vol. 106(1), pp. 113–128, 2018.
URL https://doi.org/10.1109/JPROC.2017.2725482
Main reference David Umsonst, Henrik Sandberg: "On the confidentiality of controller states under sensor attacks", Autom., Vol. 123, p. 109329, 2021.
URL https://doi.org/10.1016/j.automatica.2020.109329

In this talk, we discuss two aspects of model- and data-based intrusion detection. First, we show how a centralized model- and data-based intrusion detector in an industrial control system can use analytical redundancy to first detect and then reconstruct attacked signals in local feedback loops, to achieve resilience. We discuss pros and cons of the model- and data-based detection schemes. Second, we discuss a necessary and sufficient condition for an adversary with access to sensor data to replicate the state of the control system, and in extension the intrusion detection system. Advanced adversaries use such state information to launch stealthy attacks, and our condition gives insights as to how to block such attacks. The condition also provides insights on the possibilities for adversaries to replicate and synchronize with the state of digital twins.

The talk is based on the following papers: [1, 2].

References

1 Kaveh Paridari, Niamh O'Mahony, Alie El-Din Mady, Rohan Chabukswar, Menouer Boubekeur, and Henrik Sandberg. A framework for attack-resilient industrial control systems: Attack detection and controller reconfiguration. *Proceedings of the IEEE*, 106(1):113–128, January 2018.
2 David Umsonst and Henrik Sandberg. On the confidentiality of controller states under sensor attacks. *Automatica*, 123:109329, 2021.

3.12 Through the Looking Glass, and What We Found There

Nils Ole Tippenhauer (CISPA – Saarbrücken, DE)

Main reference Daniele Antonioli, Nils Ole Tippenhauer: "MiniCPS: A Toolkit for Security Research on CPS Networks", in Proc. of the First ACM Workshop on Cyber-Physical Systems-Security and/or PrivaCy, CPS-SPC 2015, Denver, Colorado, USA, October 16, 2015, pp. 91–100, ACM, 2015.
URL https://doi.org/10.1145/2808705.2808715

In this talk, we reflect on our research journey in the area of cybersecurity for industrial control systems. During our work on GPS spoofing [1], we noted two main challenges for precise GPS spoofing: i) the attacker needs to accurately create spoofed GPS signals (i.e., their signal strength, timing, etc), and ii) the attacker needs to carefully start the attack to slowly divert the victim's state estimation (assuming prior synchronization to legitimate GPS signals) from the legitimate to the manipulated state. Such challenges that introduce control theoretic approaches to cybersecurity motivated us to further investigate cybersecurity for general Cyber-Physical Systems, in particular industrial control systems. To understand and experiment with such systems, we built several testbeds at SUTD in Singapore [2], and designed the MiniCPS framework [3] to emulate those environments. The resulting datasets turned out to be very useful for training and evaluation of process-aware attack detection systems [4, 5]. We also realized that tools such as MiniCPS could enable the construction of to *digital twins* – for example to be used as Honeynets, reference in anomaly detection, and for attack development and verification.

References

1 Nils Ole Tippenhauer, Christina Pöpper, Kasper Bonne Rasmussen, and Srdjan Capkun. On the requirements for successful GPS spoofing attacks. In *Proceedings of the 18th ACM Conference on Computer and Communications Security*, CCS '11, pages 75–86, New York, NY, USA, 2011. Association for Computing Machinery.

2 A. P. Mathur and N. O. Tippenhauer. SWaT: A water treatment testbed for research and training on ICS security. In *2016 International Workshop on Cyber-physical Systems for Smart Water Networks (CySWater)*, pages 31–36, April 2016.

3 Daniele Antonioli and Nils Ole Tippenhauer. MiniCPS: A toolkit for security research on CPS networks. In *Proceedings of the First ACM Workshop on Cyber-Physical Systems-Security and/or PrivaCy*, CPS-SPC '15, pages 91–100, New York, NY, USA, 2015. ACM.

4 Riccardo Taormina, Stefano Galelli, Nils Ole Tippenhauer, Elad Salomons, Avi Ostfeld, Demetrios G. Eliades, Mohsen Aghashahi, Raanju Sundararajan, Mohsen Pourahmadi, M. Katherine Banks, B. M. Brentan, Enrique Campbell, G. Lima, D. Manzi, D. Ayala-Cabrera, M. Herrera, I. Montalvo, J. Izquierdo, E. Luvizotto, Sarin E. Chandy, Amin Rasekh, Zachary A. Barker, Bruce Campbell, M. Ehsan Shafiee, Marcio Giacomoni, Nikolaos Gatsis, Ahmad Taha, Ahmed A. Abokifa, Kelsey Haddad, Cynthia S. Lo, Pratim Biswas, M. Fayzul K. Pasha, Bijay Kc, Saravanakumar Lakshmanan Somasundaram, Mashor Housh, and Ziv Ohar. Battle of the attack detection algorithms: Disclosing cyber attacks on water distribution networks. *Journal of Water Resources Planning and Management*, 144(8):04018048, 2018.

5 Alessandro Erba, Riccardo Taormina, Stefano Galelli, Marcello Pogliani, Michele Carminati, Stefano Zanero, and Nils Ole Tippenhauer. Constrained concealment attacks against reconstruction-based anomaly detectors in industrial control systems. In *Annual Computer Security Applications Conference*, ACSAC '20, pages 480–495, New York, NY, USA, 2020. Association for Computing Machinery.

4 Working Groups

4.1 Conceptualization of the Digital Twin for Cyber-Physical Systems Security

Matthias Eckhart (SBA Research – Wien, AT)

Joint work of Matthias Eckhart, David Allison, Magnus Almgren, Katharina Ceesay-Seitz, Andreas Ekelhart, Helge Janicke, Simin Nadjm-Tehrani, Awais Rashid, Edgar Weippl, Mark Yampolskiy

The objective of this working group was to (i) analyze the potential characteristics of digital twins, (ii) identify security-relevant purposes, and (iii) create a mapping between the two to inform security researchers and practitioners about the characteristics that are required to implement a certain purpose. The first breakout session kicked off with a brainstorming exercise to decompose the research problem at hand into a set of questions, namely:

- In the context of the barest definition of the term, what would qualify as a digital twin?
- How does a digital twin differ from a digital representation of a physical entity that may be implemented as a data-driven model, 3D visual model, or simulation?
- How can the fidelity of a digital twin be defined and measured?
- On which CPS layers should digital twins function?
- What does synchronization in the context of digital twins mean?
- To what extent is synchronization between the digital twin and its counterpart necessary?
- How can a synchronization mechanism be implemented that covers the physics, application, network, and user layers?
- For which cases would a bidirectional connection between the CPS and the digital twin(s) be necessary?
- How would the time and methodology of digital-twin construction differ for certain activities within the CPS lifecycle?
- What is the value of a digital twin in terms of improving the security of CPSs?
- How do digital twins differ from honeypots and cyber ranges (i.e., security testbeds)?

The rationale behind asking these questions was to explore and identify different characteristics that define security-focused digital twins. During the breakout sessions, the participants engaged in vivid discussions that generated an initial draft of definitions. The group then assigned those characteristics to security-relevant purposes, indicating which features a digital twin should possess to be useful for addressing well-known cybersecurity challenges. A summary of the results is currently in preparation and will be submitted for peer review in the upcoming months.

4.2 Attack Recovery for Control Systems

Martina Maggio (Universität des Saarlandes – Saarbrücken, DE)

Joint work of Martina Maggio, Ali Abbasi, Alvaro Cárdenas Mora, Marina Krotofil, Miroslav Pajic, Awais Rashid, Francesco Regazzoni, Anne Remke, Henrik Sandberg, Anne-Kathrin Schmuck, Nils Ole Tippenhauer

In this working group, we discussed *digital-twin approved recovery* strategies. Suppose that an attack is ongoing and has been detected, the discussion centred around "what kind of manoeuvres are safe after an attack has been identified"?

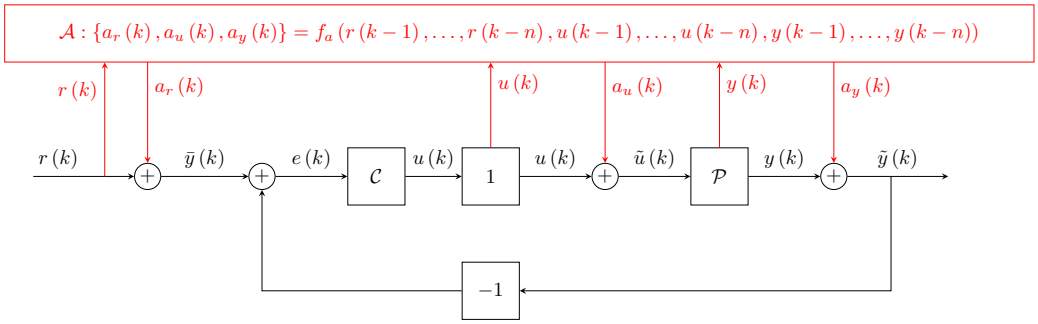

Figure 1 General attack model.

Generally speaking, we identified different goals for attack recovery:

- Recovery
- Resilience (long-term recovery)
- Safe shutdown or graceful degradation
- Survivability (we could test on the digital twin that the system would survive a catastrophic event)
- Mission completion

In this context, we moved onto discussing the actual possible actions that can be taken as a response to the attack and a potential modelling of the attack.

In Figure 1, we identify how a control system and its digital twin would look like. The variable k counts time iterations. A setpoint $r(k)$ is provided to the system (a drone should reach a given point in a tri-dimensional space). This setpoint can be intercepted and attacked using a signal $a_r(k)$ that is summed to the actual given setpoint (this models any replacement of the setpoint value). The controller then receives $\bar{y}(k)$ and calculates an error signal $e(k)$ that determines the current distance from the setpoint. This value is used by the controller to calculates a control signal $u(k)$, that is then sent to a plant. An attacker can intercept the sensor data and modify the control signal. This is modelled using a value $a_u(k)$ that is calculated by the attacker and summed to the received control signal, forming $\tilde{u}(k)$, which is fed to the plant. The plant then executes and physical values $y(k)$ are sensed. Sensors can also be attacked, via a signal $a_y(k)$, generated by the attacker.

The blocks $\mathcal{C}$, 1, and -1 can be augmented with knowledge from the plant (for example: typical execution delays, typical network delays, typical probability of not receiving packets over the network, etc). The block $\mathcal{P}$ can be augmented with knowledge from the physics (for example: acceptable values for friction and stiction coefficients). This knowledge augments the blocks forming the *digital twin*, and can be exploited by the recovery mechanism to detect and react to unusual situations. For example, if the controller execution time is longer than expected, the digital twin can suspect an attack.

A consideration that emerged is that while normally the controller closes the loop around a physical system, during the recovery period the system runs in open loop and can and must not trust the input data it receives from the sensors, because they would be compromised. In this situation, the detection of the attack could lead us to understand and estimate when the attack started and hence when the last reliable data was received by the controller. The digital twin could than be used to fast forward the execution of the controller and estimate the state of the actual system that received control signals that were calculated based on

compromised data. The digital twin could also be used to understand what are good control signal to apply while the system is running in open loop. From the control perspective, this can for example be done running a model predictive control algorithm.

Participants

- Ali Abbasi
Ruhr-Universität Bochum, DE
- David Allison
AIT – Austrian Institute of Technology – Wien, AT
- Magnus Almgren
Chalmers University of Technology – Göteborg, SE
- Alvaro Cárdenas Mora
University of California – Santa Cruz, US
- Katharina Ceesay-Seitz
CERN – Meyrin, CH
- Matthias Eckhart
SBA Research – Wien, AT
- Andreas Ekelhart
SBA Research – Wien, AT
- Helge Janicke
Cyber Security CRS – Joondalup, AU
- Marina Krotofil
Maersk – Aarhus, DK
- Martina Maggio
Universität des Saarlandes – Saarbrücken, DE
- Simin Nadjm-Tehrani
Linköping University, SE
- Miroslav Pajic
Duke University – Durham, US
- Awais Rashid
University of Bristol, GB
- Francesco Regazzoni
University of Amsterdam, NL & Università della Svizzera italiana, CH
- Anne Remke
Universität Münster, DE
- Henrik Sandberg
KTH Royal Institute of Technology – Stockholm, SE
- Anne-Kathrin Schmuck
MPI-SWS – Kaiserslautern, DE
- Nils Ole Tippenhauer
CISPA – Saarbrücken, DE
- Edgar Weippl
University of Vienna & SBA Research – Wien, AT
- Mark Yampolskiy
Auburn University, US

Report from Dagstuhl Seminar 22172

Technologies to Support Critical Thinking in an Age of Misinformation

Tilman Dingler*[1], Benjamin Tag*[2], and Andrew Vargo*[3]

1 **The University of Melbourne, AU.** tilman.dingler@unimelb.edu.au
2 **The University of Melbourne, AU.** benjamin.tag@unimelb.edu.au
3 **Osaka Metropolitan University, JP.** awv@omu.ac.jp

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 22172 "Technologies to Support Critical Thinking in an Age of Misinformation". This seminar brought together experts from computer science, behavioural psychology, journalists, and policy makers to examine and define the challenges of misinformation and fake news in the internet and social networks. This included discussions of what constitutes misinformation, technological advances for both spreading and mitigating misinformation, and discussions around policies that can be created and implemented to address propagators, both active and passive, of misinformation. The goal of this report is to summarize and present the various challenges and options for the development and implementation of technologies to support critical thinking.

Seminar April 24–27, 2022 – http://www.dagstuhl.de/22172
2012 ACM Subject Classification Human-centered computing → Human computer interaction (HCI); Human-centered computing → Social networks
Keywords and phrases Cognitive Security, Misinformation, Bias Computing
Digital Object Identifier 10.4230/DagRep.12.4.72

1 Executive Summary

Andreas Dengel
Laurence Devillers
Tilman Dingler
Koichi Kise
Benjamin Tag

The Dagstuhl Seminar on "Technologies to Support Critical Thinking in an Age of Misinformation" ran over a course of three days in April 2022. Each day focused on one specific aspect of the problem of Misinformation and the role technologies play in its worsening and mitigation.

Day 1 put the overall seminar goals and an introduction to the topic into its focus. All participants introduced themselves and gave a concrete example of an important challenge they have identified. The collected challenges were organized and later used as core challenges for group work activities, here Regulations/Policies, Human Factors and Platforms, and Critical Thinking. Over the course of the three days three groups worked on defining challenge statements (Day 1), ideas to solve the issue (Day 2), and concrete Research Questions and Project/Collaboration proposals (Day 3).

* Editor / Organizer

Technologies to Support Critical Thinking in an Age of Misinformation, *Dagstuhl Reports*, Vol. 12, Issue 4, pp. 72–95
Editors: Andreas Dengel, Laurence Devillers, Tilman Dingler, Koichi Kise, and Benjamin Tag
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The theoretical underpinnings of all group discussions and activities were provided by a series of presentations that were topically organized. Day 1 was centered around how the problem of misinformation has evolved and why misinformation is so successful these days. A historical overview was given by keynote speaker Prof. Emma Spiro, which concluded with the key insights that Networks and platforms shape information flow and that attention dynamics matter. The second keynote talk of the day was given by Prof. Andreas Dengel that put light on the crucial role that images and their power to convey information that is tainted with emotional information, and how technology (e.g, CNNs) can be used to detect those, classify them, and can potentially correct them.

On day 2, the participants zeroed in on the role technology plays. Session 1, started with a keynote by Prof. Niels van Berkel on the role of Artificial Intelligence, and Human-AI interaction. Looking at Technology, Society, and Policy on a larger scale, van Berkel identified the core issue that there exists a lack of literacy on the tech side as well as on the regulatory side, a potential consequence of the lack of qualified tech personnel on regulatory bodies. Keynote 2, by Prof. Laurence Devillers, looked at how technology is used to misinform, deceive, and change public opinion, while proposing solutions, such as Nudging and Boosting techniques, how Human-Ai interaction should be better understood, and how research and industry must work together to mitigate the problem of lacking literacy. In session 2 of the day, Prof. Albrecht Schmid led an open, provocative discussion that served as a brainstorming session for the upcoming group work, mainly focussing on the role of platforms and technology. The third keynote was given by Prof. Stephen Lewandowsky who gave a detailed account of the role of human cognition and the larger impact of misinformation on democratic societies. He identified pressure points and proposes countermeasures that are effective but need to be scaled up through improved and coordinated cross-country regulation. Day 2 ended with a Misinformation Escape Room group activity (demo), led by Dr. Chris Coward, which aims at teaching players the power of misinformation and the complexity of the problem.

Day 3 featured the keynote by Roger Taylor which strongly focussed on the way misinformation is regulated globally, and how regulatory frameworks (Digital Service Act) and effective regulation can help to mitigate the misinformation problem. As an advisor to the UK government, and an expert in responsible AI programs and data ethics, Roger Taylor put a light on pain points in the bureaucracy and the misaligned aims of technology development and research, and politics.

2 Table of Contents

3 Overview of Talks

3.1 Misinformation Escape Room: A gamified approach to building resilience to misinformation

Chris Coward (University of Washington – Seattle, US)

URL www.lokisloop.org

"While facts make an impression, they just don't matter for our decision-making, a conclusion that has a great deal of support in the psychological sciences" [3]. This statement poses a fundamental challenge for the field of media and information literacy (MIL), a largely rationalist approach to learning that presumes the underlying problem to be solved is a skills deficit. The emergence of disinformation has not dislodged this conviction for many MIL scholars and practitioners. As one meta review summarizes, there is a prevailing belief that "the bulk of disinformation on the Internet could be combated with basic evaluation skills" [2]. At the same time we have witnessed a growing chorus questioning this conviction, with the most significant shortcoming concerning the psychological dimensions of disinformation, including the role of personal beliefs, social identity, emotion, confirmation bias, motivated reasoning, and epistemic beliefs [1].

In response to these observations and interviews with librarians with front-line experience helping patrons navigate misinformation, a research team led by Chris Coward at the University of Washington designed a misinformation escape room as an immersive, social, and active learning environment. In this Dagstuhl session we will play the escape room, followed by a discussion of the project's goals and research findings.

References

1 Lewandowsky, S. The "Post-Truth' World, Misinformation, and Information Literacy: A Perspective From Cognitive Science. In S. Goldstein (Ed.), Informed Societies (1st ed., pp. 69 – 88). 2019. Facet. https://doi.org/10.29085/9781783303922.006
2 Sullivan, M. C. Why librarians can't fight fake news. Journal of Librarianship and Information Science, 096100061876425. 2018. https://doi.org/10.1177/0961000618764258
3 Wardle, C. and Derakhshan, H. Information Disorder: Toward an interdisciplinary framework for research and policy making. Council of Europe report. 2019. Retrieved from: https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

3.2 What the world thinks: Trending Topics and Multimedia Opinion Mining

Andreas Dengel (DFKI – Kaiserslautern, DE)

The Internet is full of opinion hidden under tons of irrelevant and unstructured data. From micro-blogging platforms like Twitter to video repositories such as YouTube the users express sentiments about products, brands, institutions and governments. Moreover, users tag other users' opinions: they submit comments in the same micro-blog, link them with other media

content, submit brief comments to the videos or just click the "like" option. In the past few years, there has been a huge increment of interest in the analysis of this type of content by opinion consumers, such as companies and media organizations. Among others, companies aim at mining this collective opinion in order to know what people think and how they feel about products and services. The main drawback of current solutions is that they only consider the textual content, ignoring other sources and modalities of opinion and its cross-media relationships.

This talk addresses the challenge of opinion mining of multimedia content from the Web. This comprises a multi-modal analysis of social media streams and their underlying network dynamics considering different media channels such as Twitter, YouTube, Flickr, Google, and Wikipedia. It specifically proposes solutions for:

1. The detection of trending topics from a large set of dynamic data streams. These trending topics were able to be clustered, tracked and aggregated over social media channel and over time. In particular the combination of statistical methods with linked open data was of help to achieve this goal (see [3, 4, 6]).
2. The content-based multimedia analysis on single modalities and combination of multi-modalities. Here, the early shift from traditional approaches towards deep learning proofed to be in particular successful. Key element of analysis were Adjective Noun Pairs (ANP) providing a mid-level representation for visual content and foundation for sentiment analysis. The idea of ANPs was further extended towards Verb-Noun-Pairs to capture temporal concepts present in audio and video streams (see [1, 2, 5, 7])

References

1 M. Al-Naser, S. M. Chanijani, S. S. Bukhari, D. Borth, and A. Dengel. What makes a Beautiful Landscape beautiful: Adjective Noun Pairs Attention by Eye-Tracking and Gaze Analysis. In ACM Workshop on Affect and Sentiment in Multimedia (ASM), 2015.

2 B. Bischke, D. Borth, and A. Dengel. Large-Scale Social Multimedia Analysis. In Vrochidis Stefanos and Huet Benoit and Chang Ed and Kompatsiaris Ioannis, editor, Big Data Analytics for Large-Scale Multimedia Search. Wiley & Sons, Ltd., 2018. (to appear).

3 S. Elkasrawi, H. Elwy, S. Bauman, C. Reuschling, and A. Dengel. Prediction of Social Trends Using Nearest Neighbours Time Series Matching and Semantic Similarity. In Advances in Data Mining, 16th Industrial Conference, ICDM 2016, Poster Proceedings, 2016. (to appear).

4 S. Fuchs, D. Borth, and A. Ulges. Trending topic aggregation by news-based context modeling. In Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz), pages 162 – 168. Springer, 2016.

5 J. Folz, C. Schulze, D. Borth, and A. Dengel. Aesthetic Photo Enhancement using Machine Learning and Case-Based Reasoning. In ACM Workshop on Affect and Sentiment in Multimedia (ASM), 2015.

6 A. Koochali, S. Kalkowski, A. Dengel, D. Borth, and C. Schulze. Which languages do people speak on flickr?: A language and geo-location study of the yfcc100m dataset. In Proceedings of the 2016 ACM Workshop on Multimedia COMMONS, pages 35 – 42. ACM, 2016.

7 S. Kalkowski, C. Schulze, A. Dengel, and D. Borth. Real-time Analysis and Visualization of the YFCC100m Dataset. In ACM Multimeda MMCOMMONS Workshop, 2015.

3.3 AI-enhanced nudging mechanism using affective computing: ethical issues

Laurence Devillers (CNRS – Orsay, FR & Sorbonne University – Paris, FR)

Ethics, Goals, and Societal impact have always been central subjects since the early days of the field of research on artificial intelligence such as affective computing. But currently, the new uses of social robots, affective conversational agents (chatbots), and, more generally, the so-called "affectively intelligent" digital environments in fields as diverse as health, education, insurance, transport, or economics reflect a phase of significant change in human-machine relations, amplify the necessity to keep great attention in ethical dimensions of these systems. What ethical issues arise from the development of affective computing with chatbot/robot interaction? Does it raise the crucial issue of trust? How will humans co-learn, co-create and co-adapt with the Machine? Notably, how will vulnerable people be protected against potential threats of the machine? During an interaction, we adapt our linguistic behaviors but also our prosodic and gestural behaviors and our conversational strategies. This multi-level adaptation can have several functions: reinforcing engagement in interaction, emphasizing our relationship with others, and showing empathy. Anthropomorphism introduces many challenges, among them ethical, uncanny valley, practical implementation, and user mind-reading problems. The anthropomorphic goal of "just like a human-to-human conversation". The designers of conversational agents seek for many to imitate, simulate the dialogical behavior of humans, and users spontaneously anthropomorphize the conversational agents" capacities and lend them human understanding. Thus, the Dilemma of the researchers is, on the one hand, to achieve the highest performance with conversational virtual agents and robots (close to or even exceeding human capabilities) but on the other hand, to demystify these systems by showing that they are "only machines".

Conversational agents and social robots using autonomous learning systems and affective computing will change the game around ethics. We need to build long-term experimentation to survey Human-Machine Co-evolution and to build "ethics by design" chatbots and robots. In the chair HUMAAINE (head: L. Devillers, LISN-CNRS, France), we aim to study the Human-Machine Affective interactions and relationships, in order to audit and measure the potential influence of intelligent and affective systems on humans, and finally to go towards a conception of "ethical systems", by design or not and to propose evaluation measures. For this purpose, the planned scientific work focuses on the detection of social emotions in a human voice, and on the study of audio and spoken language "nudges" [1, 2], intended to induce changes in the behavior of the human interlocutor.

Nudging is an ethically highly problematic topic. A digital nudge is an almost imperceptible incentive in the design of a digital system to drive behavior that is supposed to improve personal or collective well-being. Digital nudges use personal data and biometric sensors to profile and encourage people to take unintended actions while using familiar online technologies such as email, pop-ups, SMS, web interfaces, smart watches, mobile apps, IoT, home appliances, smart cars, chatbots, robots, etc. However, when a digital nudge is enhanced by Artificial Intelligence systems (so-called AI-enhanced nudge) using machine learning and affective computing technologies based on cognitive biases and behavioral science, its potential is immense. While its usage can be beneficial for an individual or the society, the AI-enhanced nudge persuasive power and intrusive capacity can also cause subliminal manipulations and profound and long-lasting changes in the behavior of users,

especially children, and vulnerable people. If AI-enhanced Nudges are (intentionally or not) misused, they may become dangerous and raise serious ethical issues that can undermine the level of distrust in AI-enhanced nudging systems. AI-enhanced Nudging is already a reality influencing the actions and behaviors of thousands of people in the fields of education, health, gaming, gambling, hospitality, smart cities, security, justice, etc. However, this "soft" manipulation of behavior and emotions raises ethical questions to which no standard today provide direct answers. As AI-enhanced Nudging systems are flourishing in the market, it is widely believed that their design and the use of them, in the short or long term, ought to establish human and social responsibility through auditable behaviors under a typical set of conditions. This could help to build trustworthiness within a sustainable market. There is an urgent need to create a shared terminology and consensual processes and methodologies to mitigate and ethically adjust the enormous ability of people's manipulation provided by AI technologies such as affective computing to digital nudge [3].

References

1 H. Ali Mehenni, S. Kobylyanskaya, I. Vasilescu, L. Devillers, Nudges with a conversational agent or social robot: a first experiment with children at a primary school, IWSDS 2020
2 N. Kalashnikova, S. Pajak, F. Le Guel, I. Vasilescu, G. Serrano, and L. Devillers, *Corpus Design for Studying Linguistic Nudges in Human-Computer Spoken Interactions*, Language Resources and Evaluation Conference 2022 (LREC 2022), Marseille, France 2022, June 2022.
3 L. Devillers, E. Panai, Ad-hoc group 6: AI-Enhanced nudges (AFNOR/CEN-CENELEC/JTC21)

3.4 From Cognition-Aware to Bias-Aware Systems

Tilman Dingler (The University of Melbourne, AU)

Joint work of Tilman Dingler, David A. Eccles, Martin Pielot, Benjamin Tag

With advancements in sensing and processing power and more sophisticated machine-learning methods, computing systems can increasingly detect and monitor human activities. Computers that consider the context in which they are used can support their users according to their current location, activities, and intent, a field coined context-aware computing [9]. In our work, we have extended this notion to also include the user's cognitive context to build systems that help users increase their ability to effectively process information according to their current mental state. By utilising phone sensor data, for example, we have trained machine-learning algorithms to detect when people are attentive to their phones [3] and seek stimulation [8]. Insights into when a person is bored or focused can provide us with a better understanding of when people are more productive and when downtimes occur: during highly focused states, devices in the user's environment can be advised to prevent interruptions in order to help people focus better. Beyond in-situ assessments of cognitive states, we have developed tools and methods to elicit users' circadian rhythms of alertness [1, 2, 10], which describe systematic cognitive performance fluctuations throughout the day. Awareness of these rhythms opens up a whole range of opportunities to suggest content, schedule a day full of work, or generally recommend activities whose cognitive requirements match the user's current state [5]. The resulting tools and algorithms give insights into the user's internal body clock, which helps people to better schedule, for example, learning sessions or generally

activities that require high focus. In recent years, however, it has become apparent that effectively dealing with information is not necessarily a matter of consuming more in less time but of the quality of the information processing itself. Society is transitioning into an era where computing pervades all aspects of people's lives, with humans and their cognitive processes at the centre of it. The rise of fake news and the interplay between bad actors, fast dissemination through social media, and people's receptivity to emotionally charged content present an ever-growing challenge to individuals, society, and our democratic institutions [7]. When looking at what can be done about its reception, we have identified several preventative interventions to bolster people against fake news. These include media literacy training, psychological inoculation, and transaction cost economics [6]. Further, receptivity to fake news often comes from the prevalence of cognitive biases. They play an important role in how information is perceived and processed, a fact that can be both utilised and exploited by computing systems. A prominent example of a cognitive bias is the confirmation bias, i.e., the tendency to seek out information that confirms our existing perspectives and notions. We have recently established a strand of research to use sensors to detect the occurrence of cognitive biases. Computing systems capable of detecting cognitive biases, which we call bias-aware systems, can help people increase their awareness of and mitigate their effects as well as inform recommender systems to introduce a more balanced news diet. One of the main challenges is the collection of ground truth, i.e., ensuring we can successfully induce and measure the occurrence of cognitive biases for observational and experimental research. Therefore, we developed a tool to collect ground truth on people's implicit preferences that can be adapted to any thematic issue, such as opinions on climate change, feminism, or political ideologies [4]. The Dagstuhl Seminar on "Technologies to Support Critical Thinking in an Age of Misinformation" was born out of the realisation that the problem of fake news can only be addressed in a truly interdisciplinary fashion as it involves the technology through which fake news spread, the human who creates, receives and shares it, and the regulatory bodies who are looking for ways of reeling in its spread. Throughout the 3-day seminar, our team sat down with behavioural psychologists, government advisers, and technologists to discuss the human element in this triangle of technology, human, and government. The goal of our bias detection research is to allow people to increase their awareness of their innate biases and allow systems to help mitigate them. Media literacy training, on the other hand, can help current and future generations of technology users critically process online information. Future computing systems need to be designed responsibly to consider people's cognitive biases and help them bolster against their cognitive vulnerabilities. Technology is thus the ouroboros of fake news, i.e., its enabler and mitigator.

References

1. Dingler, Tilman, Albrecht Schmidt, and Tonja Machulla. *Building cognition-aware systems: A mobile toolkit for extracting time-of-day fluctuations of cognitive performance.* Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1, no. 3 (2017): 1-15.
2. Dingler, Tilman, Ken Singer, Niels Henze, and Tonja-Katrin Machulla. *Extracting Daytime-Dependent Alertness Patterns from Mobile Game Data.* In 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 1-6. 2020.
3. Dingler, Tilman, and Martin Pielot. *I'll be there for you: Quantifying Attentiveness towards Mobile Messaging.* In Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 1-5. 2015.
4. Dingler, Tilman, Benjamin Tag, David A. Eccles, Niels van Berkel, and Vassilis Kostakos. *Method for Appropriating the Brief Implicit Association Test to Elicit Biases in Users.* In

CHI Conference on Human Factors in Computing Systems, pp. 1-16. 2022.
5 Dingler, Tilman, Dominik Weber, Martin Pielot, Jennifer Cooper, Chung-Cheng Chang, and Niels Henze. *Language learning on-the-go: opportune moments and design of mobile microlearning sessions.* In Proceedings of the 19th international conference on human-computer interaction with mobile devices and services, pp. 1-12. 2017.
6 Eccles, David A., Sherah Kurnia, Tilman Dingler, and Nicholas Geard. *Three Preventative Interventions to Address the Fake News Phenomenon on Social Media.* In Proceedings of ACIS, 2021.
7 Lazer, David MJ, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger et al. *The science of fake news.* Science 359, no. 6380 (2018): 1094-1096.
8 Pielot, Martin, Tilman Dingler, Jose San Pedro, and Nuria Oliver. *When attention is not scarce-detecting boredom from mobile phone usage.* In Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing, pp. 825-836. 2015.
9 Schilit, Bill, Norman Adams, and Roy Want. *Context-aware computing applications.* In 1994 first workshop on mobile computing systems and applications, pp. 85-90. IEEE, 1994.
10 Tag, Benjamin, Andrew W. Vargo, Aman Gupta, George Chernyshov, Kai Kunze, and Tilman Dingler. *Continuous alertness assessments: Using EOG glasses to unobtrusively monitor fatigue levels In-The-Wild.* In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1-12. 2019.

3.5 Technology and Democracy: Cognitive Remedies

Stephan Lewandowsky (University of Bristol, GB)

Democracy is in retreat or under pressure worldwide. Even in countries with strong democracies, polarization is increasing, and the public sphere is awash in misinformation and conspiracy theories. Many commentators have blamed social media and the lack of platform governance for these unfortunate trends, whereas others have celebrated the Internet as a tool for liberation, with each opinion being buttressed by supporting evidence. One way to resolve this paradox is by identifying some of the pressure points that arise between the architecture of human cognition and the online information landscape, and their fallout for the well-being of democracy. Two such pressure points arise from the algorithmic curation of content and the prevalence of misinformation and disinformation on social medial. Virtually everything users see on the internet is curated by intelligent algorithms (e.g., the newsfeed on Facebook or Twitter). These algorithms are designed by platforms without public accountability or auditing, with the primary intent of keeping users engaged longer by satisfying their presumed preferences. While preference satisfaction by itself is not a threat to democracy, it can become problematic if extremist or conspiratorial content keep a person engaged longer, because the platforms are then incentivized to prevent more and more potential harmful content. Algorithms may thus at least indirectly imperil our democracy when people are being radicalized or are presented with misinformation not because they want to, but because platforms are making money by facilitating it. Algorithmic content curation can additionally be problematic if people's personal data are used to identify sensitive attributes, such as their personality or sexual orientation, which political operatives can then exploit by presenting messages to people that exploit their personal vulnerabilities. This process is known as

microtargeting and it comes with a number of attributes that may imperil democracy. In the absence of any regulation, one possible countermeasure involves "boosting" people's ability to detect on their own when they might be targeted by manipulative messages. A existence proof of boosting showed that once people were given information about their own personality along the introversion-extraversion spectrum, they were better able to identify advertisements that were aimed at them based on their personality. Similar "boosting" approaches can also equip people to become resilient to misinformation and disinformation online. This approach is known as inoculation and it entails warning people ahead of time that they might be misled, and providing them with information about the misleading rhetorical techniques they are likely to encounter. Inoculation has been shown to be effective in numerous different domains, from anti-vaccination messages to radicalization attempts and conspiracy theories. In all cases, people's ability to detect when they are manipulated was significantly enhanced by inoculation. Notwithstanding the success of such cognitive countermeasures, they are insufficient to counter the immense asymmetry in power between platforms and users that currently exists and that gives rise to the pressure points between cognition and technology. It requires deep structural change and smart regulation to create a new Internet with democratic credentials.

3.6 Regulation of Misinformation

Roger Taylor (Open Data Partners – London, GB)

Different countries are taking very different approaches to regulation of social media to combat misinformation. Singapore has passed a law against telling lies which allows the government to order the take down of material regarded as untrue. China is seeking to register anyone who comments publicly about key political issues online. The European Union is bringing in regulation that makes social media responsible for harms which include harms to democracy and civic discourse as well as harms to fundamental rights. The UK is proposing more limited regulation that focuses on immediate harm to individuals rather than harm to society (but which still might capture medical misinformation). The US is adopting a more laissez-faire attitude based on giving primacy of freedom of speech. However, within the US, individual platforms are implementing their own governance mechanisms in recognition of public pressure for change.

These regulatory strategies do not specifically call for action on critical thinking. However, platforms may respond to the European regulatory proposals by adopting measures such as misinformation vaccination. (Also, there are, in some territories, complementary strategies on media literacy alongside regulatory proposals – the EU strategy on disinformation).

Key limitations to the successful implementation of regulation are: A lack of social consensus around the meanings of the words used (e.g. "harm to public discourse" "psychological harm"). Lack of agreed mechanisms that are capable of determining whether such harm has occurred. Lack of mechanisms for determining responsibility. (Regulations require platforms to balance rights to free speech against risk of harm. The issue for regulators is whether they have found the right balance. This requires a determination of whether or not they could have done better in balancing these risks which, in turn, requires an understanding of what is possible in order to make a sound assessment of responsibility.)

Each of these issues is exacerbated by the rapidly changing and hugely heterogenous nature of the harms being addressed, as well as the complexity of the environment that is being regulated.

Regulators will likely have to adopt an approach based on identifying the most egregious harms and using rough and ready measures to assess the responsibility of the platform. The degree to which this will significantly impact disinformation and online harms is uncertain.

Regulators would be wise to adopt a strong stance at the outset with regard to the data access provisions in the EU regulations. They should set out a long term strategy to establish

- relatively objective/consensual approaches to categorising and monitoring misinformation;
- research methods to understand the impact of misinformation on individuals and on democracy;
- and mechanisms for understanding the relative impact of different types of remedy including media literacy and critical thinking.

3.7 Move Slow and Fix Things: Algorithms, Systems, and Design

Niels van Berkel (Aalborg University, DK)

The efforts toward technologies to support critical thinking in an age of misinformation require a collaborative effort across the fields of Technology, Society, and Policy. In this appetiser talk, I will outline some of the primary challenges faced in each area, pointing to promising research that indicates opportunities for moving forward.

TECHNOLOGY
Challenges faced within the technology field include biases, biased algorithms, and black box decision-making. Therefore, it is critical to recognise the real-world consequences of algorithmic systems. Examples include disparities in AI skin cancer diagnoses between different skin colours and discrimination built into the design of a fraud detection system of the Dutch tax authorities. Recent work by Huszár et al. analyses the amplification of tweets by elected legislators from major political parties in seven countries [1]. Their results show that the mainstream political right enjoys higher algorithmic amplification. While highlighting the possibility of assessing the impact of algorithmic-driven recommendation systems, it also raises new questions, including; Should distribution always be a perfect 50/50 split? Are politics as black and white as left / right? What can we do about this technological bias?

SOCIETY
Disinformation, filter bubbles, and an increased polarisation in politics and beyond are amongst the challenges currently faced in society. Current events, such as the Russian invasion of Ukraine, bring to the front the societal challenges related to disinformation. Disinformation also plays a significant role domestically, with the US being a famous example of the growing ideological divide between Democrats and Republicans. In democratic countries with a multi-party democratic system, such as The Netherlands, polarisation can take a different form – with traditional parties finding it increasingly challenging to distinguish themself from one another. Recent work by Broockman and Kalla studies the effect of paying Fox News viewers to regularly watch CNN (two politically opposed media channels) [2]. Compared to a control group of Fox News viewers, the study finds more nuanced political

beliefs and knowledge of current events. The authors highlight how the skewing of media has had a broader and negative impact on how US society functions. Highlighting the opportunity for viewers to obtain more nuanced viewpoints once presented with an alternative media source, the study raises new relevant questions, including the potential for long-term effects and the impact of removing the financial incentives offered in the study.

POLICY
With the growing impact of technology in an increasingly unstable world, policy is often looked at as the instrument to bring back some stability. Simultaneously, policy can be perceived as a slow-moving instrument which struggles in dealing with local versus global issues. In this context, we increasingly see the global impact of national and international governmental organisations. In particular, the US and the EU are at the forefront of developing AI policies and research plans. A recent review by Jobin et al. studies the global landscape of AI ethics guidelines [3]. Their results highlight eleven unique ethical principles that are discussed within these guidelines, including "transparency", "justice", and "privacy". Their study shows an apparent interest of both governmental organisations and industry in developing ethics guidelines while also presenting questions for further research. For example; How do we implement these guidelines in products and services? Can we match policy with outcomes? How do these, primarily developed in Europe and North America, impact the rest of the world?

References

1 Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). *Algorithmic amplification of politics on Twitter*. Proceedings of the National Academy of Sciences, 119(1).
2 Broockman, D., & Kalla, J. (2022). *The manifold effects of partisan media on viewers' beliefs and attitudes: A field experiment with Fox News viewers*. OSF Preprints.
3 Jobin, A., Ienca, M., & Vayena, E. (2019). *The global landscape of AI ethics guidelines*. Nature Machine Intelligence, 1(9), 389-399.

4 Working groups

On the first day of the seminar, participants jointly collected and discussed a range of challenges that misinformation on digital platforms presents. Throughout the day, we arrived at six common themes:

1. **Critical Thinking**: what are critical thinking skills students and, more generally, online users need to have to navigate online platforms? How can critical reflection be prompted? How can we design platforms to help users break down filter bubbles?
2. **Regulation and Policies**: How does an empirical approach towards research and policy development with regard to misinformation look like? How can we systematically gather evidence of the impact on interventions on false beliefs? Which measures (algorithm policing, platform policies, education) are most impactful? How do we regulate online discourse without breaking the fundamentals of a pluralistic society?
3. **Discourse *vs.* Censorship**: How can policing be done without silencing non-wanted voices? What does *good* or *healthy* online discourse look like? Research mixed with activism can be problematic. Certain opinions are difficult to express at universities. Does the notion of safe spaces lead to places where contrary opinions are not expressed anymore? What is the cost of not talking to each other?

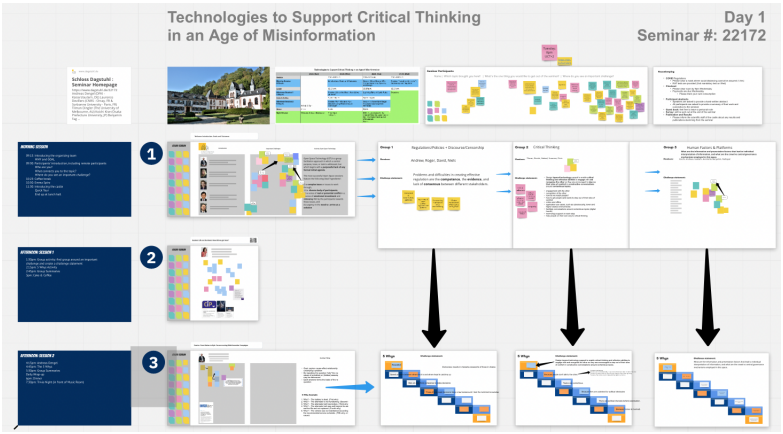

Figure 1 The Miro Digital Whiteboard Used to Facilitate Interaction and Group-Work for the In-Person and Remote Participants.

4. **Human Factors**: how can we model human cognition/emotion to make reliable predictions of behaviour? What is a meaningful cognitive architecture to enable students to think critically? And to what extend can fake information be used to achieve/do more, *i.e.*, taking benevolent advantage of it?
5. **Platforms**: how can we control misinformation across different platforms? What happens when platforms intervene but fail? Can research lead to an early detection of where technology contributes and where technology goes wrong? What role do algorithms play, and how can we design *better* ones?
6. **Bad Information**: a common problem seems to be that people don't make decisions based on facts. How can we reconcile this and bring facts back into the decision-making process?

After some discussion, we identified some overlaps and merged the *Human Factors* with *Platforms* and *Regulation and Policies* with *Discourse vs. Censorship*. The *Bad Information* scheme seemed to be ubiquitous, hence we settled on three final themes, around which we formed the following three working groups.

4.1 A Governance Framework for faster Technology Regulation

David Eccles (The University of Melbourne, AU)

Joint work of Eccles, David; Taylor, Roger; van Berkel, Niels; Vargo, Andrew.

4.1.1 Challenge Statement

Problems and difficulties in creating effective regulation are the **competence**, the **evidence**, and **lack of consensus** between different stakeholders.

4.1.2 Working Group

Our role as an information systemist and human computer interaction researchers is on the intersection between people, their processes, data, and technology. We are interested in empowering individuals to better understand how their data and the technology they use reinforces an information and knowledge asymmetry. I spent most of my time at the Dagstuhl

Seminar in the governance stream as the threat to democratic processes and institutions is evident from my research on fake news on social media phenomenon. Residing in a country which has compulsory voting for all citizens and permanent residents over the age of 18, it has an impact on elections as political parties have to move to the centre not the extremes of liberal and conservative issues to win elections and govern [5]. Democracy is under direct threat from social media platform technologies implicit and explicit role in the fake news phenomenon. Today's world wide web is no longer the place of free and open exchange of information and ideas as envisioned by its creators [2]. The world wide web and in particular social media platforms that have arisen because of the internet's capabilities in their current form and function represent a threat to democratic processes such as elections and democratic institutions fraying the separation of powers (legislatures, executive, and judiciary) [1]. Arguments can readily be made for legislative intervention using individual, societal, and government reasons [3]. A regulatory intervention for economic reasons is rarely justified except when there is a deformity in the function of a free market. Examples of market failure include its devolution to a private exchange excluding new entrants, situations where there is extreme information and or data asymmetry in the relationship between parties creating an unfair commercial and / or negotiating advantage, or an oligarchy, monopoly or duopoly exists being able to manipulate market price through collusion in supply and demand [6]. The overwhelming dominance of five key vendors Meta (formerly Facebook), Apple, Google, Microsoft and Amazon in the world wide web advertising and social media platforms demonstrates an economic failure of the market. Interventions of this kind are not new (e.g., Standard Oil, Bell Corporation), however, any legislative intervention in markets is not without unintended outcomes. Interventions may bring benefit, may bring negative, or even catastrophic unforeseen consequences [4]. In the existing market of social media platforms and the world wide web this can be demonstrated by the unintended consequences of the EU's GDPR legislation. We propose a governance model of individual's sensitive private user internet data that separates user data from its application and use. We believe this model would be a minimum imposition on technology vendors, restore the market imbalance allowing for greater competition and new entrants, and lesson the data and information asymmetry between social media users and vendors. In this model data at rest would not reside with any commercial vendor but with a statutory authority as does user statistics for bodies such as the U.S. Census Bureau, and is already in place for sensitive personal data as in the case of the Australian Digital Health Agency.

References

1 Allcott, H., & Gentzkow, M. Social Media and Fake News in the 2016 Election. 2017. The Journal of Economic Perspectives, 31(2), 211 – 235.
2 Hern, A. Tim Berners-Lee on 30 years of the world wide web: 'We can get the web we want', Interview. 2019. The Guardian Retrieved from https://www.theguardian.com/technology/2019/mar/12/tim-berners-lee-on-30-years-of-the-web-if-we-dream-a-little-we-can-get-the-web-we-want
3 House of Commons Digital. Disinformation and "fake news": Final Report. Westminster House of Commons (UK Government). 2019. Retrieved from https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf
4 Merton, R. K. The Unanticipated Consequences of Purposive Social Action. 1936. American Sociological Review, 1(6), 894 – 904.
5 Swire, T. B., Ecker, U. K. H., Lewandowsky, S., & Berinsky, A. J. They Might Be a Liar But They're My Liar: Source Evaluation and the Prevalence of Misinformation. 2020. Political Psychology, 41(1), 21 – 34.
6 Williamson, O. E. The economic institutions of capitalism: firms, markets relational contracting. 1987. Free Press.

4.2 Working Group on Critical Thinking

Tilman Dingler (University of Melbourne, AU)

4.2.1 Challenge Statement

How can we design layered technology support to enable critical thinking and reflection abilities to engage with and recognise the other so they are encouraged to step out of their silos of comfort in constructive conversations around contentious topics?

4.2.2 Working Group

This working group focused on what can be done to foster critical thinking abilities in online users and ways of bringing media literacy education into young people's curricula.

Other than relying on government regulation, platform policies, or technological interventions, the idea was to strengthen the critical thinking abilities of online users and prevent cognitive vulnerabilities. One of the premises of acknowledging or accepting other people's viewpoints is to engage with *the other* in the first place. The group discussed the necessity of stepping out of people's comfort zone and making an effort to get to know and understand other people's perspectives. Social media and also universities more generally tend to form silos of comfort where like-minded people discuss topics from one congruent angle. But the online world is made up of a plethora of perspectives, which reflects the plurality of our society. The risk of these silos or echo chambers [1] is that people with differing viewpoints do not engage with and, as a result, further alienate each other. Democracy is built on a pluralistic society where viewpoints need to be discussed in the open to reach an agreement or compromise. Being exposed to other viewpoints and the ability to critically engage, understand and find compromise are crucial for members of our society. To build and foster this ability, critical thinking skills and media literacy should be systematically integrated into all levels of our education system. We should encourage students to leave their comfort zones in an attempt to understand, learn, and re-evaluate their own standpoints.

The group collated a range of techniques and interventions to teach critical thinking abilities, including:

- Inoculation: the goal is to build psychological immunity against misinformation. This technique has famously been applied by Roozenbeek *et al.* [2] in their *Bad News* game, where players have to apply the tricks of the trade of misinformation to get their fictional message out as wide as possible. The idea is that by getting exposed to common mechanisms of creating and spreading misinformation, receivers of inoculation training acquire the ability to spot and resist manipulation.
- Debunking / Prebunking: This is the attempt to correct misinformation after (debunking) or prior to (prebunking) exposure. This can be as simple as uncovering argument-supporting facts or finding flaws in the argumentation itself. Cook and colleagues [3], however, discuss in detail why people often struggle with correcting misinformation and inaccurate beliefs and why debunking misinformation is not always straight forward. For one, there is the risk of "backfire effects," which arises when, rather than being refuted, people double down on their preconceived notions. And second, there is the role of worldviews in accentuating the persistence of misinformation. Lewandowsky *et al.* created a series of writings on how to apply debunking effectively [4].

- Building Empathy: the idea to put oneself into the shoe of another and walk a thousand miles in it. Seeing the world from someone else's perspective is crucial to understanding where the other is coming from. Connecting and empathising are the key to meaningful conversations, especially around critical topics. Only when we understand the other we can have a generative dialogue, i.e., allow the other person to contribute to our knowledge and understanding of the world around us.
- Moderation: to facilitate contentious conversations, moderation might be necessary to bring people with different views to the table. Philosopher Jürgen Habermas formulated the *ideal speech situation*, a set of basic rules that are based on reason and evidence.

Efforts to integrate critical thinking education effectively into curricula of all levels need to be based in pedagogy and philosophy. For online learning and discourse, this requires a research agenda around measuring the effectiveness of different approaches. Which skills are more successful in fighting off misinformation? How can we implement digital interventions that build spaces to meet people with opposing opinions to allow them to learn about the existence of other opinions and help them value those? And how can we measure the effectiveness of these interventions? These questions built the basis for solution proposals that members of the working group would like to take forward, such as:

1. Collect existing interventions and gather empirical evidence for their effectiveness. The *Prosocial Design Network*[1] is an online space where ideas and empirical evidence supporting those is being gathered. The group would use the platform to collect ideas and inspiration for future studies.
2. Opportunistic education: social media platforms, such as Twitter or Facebook, already flag potential mis- or harmful information. The effect of such flags on users' critical thinking skills should be assessed. This requires, however, that platforms will collaborate with researchers on conducting such experiments and data collections. Such collaborations could be open the door to changes in regulation.
3. Tools and Methods to test interventions: we need standardised ways to test the effectiveness of new interventions to allow benchmarking and comparison.
4. Integration of critical thinking and media literacy programs into all levels of the schooling systems. While media literacy training has made its way into middle school education, the quickly changing nature of the online discourse requires a frequent revisiting of these contents and techniques. In schools, universities and vocational training.

References

1 Pariser, Eli. The filter bubble: What the Internet is hiding from you. penguin UK, 2011.
2 Roozenbeek, Jon, Sander van der Linden, and Thomas Nygren. "Prebunking interventions based on the psychological theory of "inoculation" can reduce susceptibility to misinformation across cultures." Harv Kennedy Sch Misinformation Rev 2020b 1 (2020).
3 Cook, John, Ullrich Ecker, and Stephan Lewandowsky. "Misinformation and how to correct it." Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource (2015): 1-17.
4 Lewandowsky, Stephan, John Cook, Ullrich Ecker, Dolores Albarracin, Michelle Amazeen, Panayiota Kendou, Doug Lombardi et al. The debunking handbook 2020. 2020.

[1] https://www.prosocialdesign.org/

4.3 Human Factors and Platforms

Benjamin Tag (The University of Melbourne, AU)

4.3.1 Challenge Statement

What are the **information and presentation factors** that lead to individual interpretation of information? What are the **crowd vs. central** governance mechanisms employed in this space?

4.3.2 Working Group

The members of work group "Human-Factors and Platforms" were researchers in Human-Computer Interaction, Misinformation, and directors of research for commercial entities. The expertise of the group members covers document analysis, social media analysis, artificial intelligence, as well as research in trust, safety, and algorithmic responsibility. The work group focused on the question: "What are the information and presentation factors that lead to individual interpretation of information, and what are the crowd vs central governance mechanisms employed in this space." During the initial discussion session, the members identified a series of problems underlying and deriving from this core challenge. The main challenge that all members identified as crucial is that academic research and industry partners have to collaborate better, i.e., more openly. Misinformation mostly spreads through platforms built, maintained, and promoted by a relatively small group of companies. However, while these platforms are connected to certain degrees, it is extremely difficult to control the cross-platform migration of misinformation. As it will be difficult to design regulations that satisfy the needs of different platforms as well as that of public and regulators, the group agreed that the human factor rather than the infrastructural aspect of the spread of misinformation should be put into the research focus, here especially a better understanding of human cognitive architecture, and the development of tools to support critical thinking. Because the most powerful way to stop misinformation is arming people against them. This, however, should be supported by technical solutions, such as providing meta-data, automated fact-checking, and providing warnings. Following this initial definition of the problem statement, the work group used the 5-Whys technique to probe the causes of why there is no solution yet to clearly identifying what is false and what is right. One problem is that we cannot measure truth, therefore, we cannot fully trust and rely on sources, links, and services. These often lack full transparency, and tend to hide a purpose or motivation, e.g., biasing information in favor of large sponsors. The final conclusion of this exercise is that people or companies often try to come ahead of others, e.g., to gain advantages in funding through advertising, making them more powerful. The philosophical conclusion of this exercise was that you can create power by creating your own reality, as many recent political campaigns have shown.

Based on these findings, the group started to dive into the solution exploration through a reverse thinking exercise. When it comes to Human Factors, the most promising solutions to understanding why misinformation is read, believed, and distributed by humans, are "silent" solutions. This means that researchers should use unobtrusive and non-invasive sensing solutions to make sense of human cognition, without altering the human's environment and context, which potentially leads to altered behavior, e.g., through the Hawthorn Effect. Many of the necessary sensors and technologies are already integrated in computers, smartphones,

and wearable devices, such as smart watches. It is therefore not necessary to develop new sensing solutions. Rather, are researchers in academia and industry required to collaborate and better identify mutual needs and rights. Here, especially the access to information collected by platforms, e.g., Facebook, is deemed extremely helpful to researchers, allowing them to create insights on human behavior. The group also discussed the importance of these collaborations intensively. Academic and industry researchers in the group agreed that both sides have to better communicate timelines (industry and academia differ substantially), the creation of necessary output, and the creation of IP. Finally, the work group summarized these discussions in two research questions:

1. How can we teach people to have a healthy mix of trust and mistrust towards machines/machine output?
2. How can physiological data be used to make interactions with platforms more intuitive/simple – while also protecting privacy?

These research questions informed the group's last task, that of defining a set of short-, mid-, and long-term projects that help tackling the identified problems. The short term projects aim at sensing the platform impact on the user. A core issue is to develop methods that help to anonymize (physiological) user data without making them useless in order to protect privacy, which increases trust in the platforms, while enabling the full data analysis spectrum to create actionable insights. Mid term projects shall take advantage of the insights, and help researchers and developers to build intervention and nudging systems that help users follow a more balanced information diet, while allowing for data to be shared with, e.g., coaches or data analysis tools. To tackle the question of quantifying truth and identifying truth, the members discussed the idea to crowd-source information from events that make the news. Today, the majority of users carries smart devices that allow for filming, reporting, and so on. However, the big challenge here is to protect the privacy of these citizen journalists to protect them from becoming victims of targeted campaigns. Image processing and smart algorithms (NLP) will allow for an analysis of these large amounts of data. Based on the short term project, an effective way to better understand the impact platforms have on humans, is the development of an affective middle-ware. This can not only be used for better understanding of the impact, but also as a source for individualizing apps, news distribution and provide a more balanced information intake. Last but not least, the group members agreed, that in the long run, we have to aim at understanding, i.e., quantifying, how people create truth out of information, and how they decide what information should be prioritized over other.

5 Open problems

5.1 Intellectual humility: a virtue worth pursuing in public discourse?

Nabeel Gillani (MIT – Cambridge, US)

Intellectual humility is the recognition that what we believe might, in fact, be wrong. What would it mean to have more intellectual humility in our online public discourse? Could it improve how we communicate with, perceive, and ultimately treat one another? How might we design for greater intellectual humility (e.g. through changes / additions to online discourse platforms)? This lightning talk poses these questions and offers examples of how

we might design tools and systems for fostering greater intellectual humility in order to foster group discussion about the potential merits and pitfalls of more intellectual humility in our online lives.

5.2 Open problems I found at the seminar

Koichi Kise (Osaka Prefecture University, JP)

The issue of misinformation and disinformation is not just technical but related to different factors such as human cognition, and social sciences. What are correct and fake are not always clearly defined, nor shared by all people. They are often relative, depending on the standpoints of people. Some people can be easily affected by the given (mis/dis)information, but some cannot change their way of thinking even if it is better. This seminar was a good starting point for me to think about the issue with the help of talented participants from a wide variety of fields. It is mandatory to discuss the issue with such people to avoid tunnel vision. At the beginning of the seminar, it was sometimes difficult to understand well what speakers from different fields say. But the seminar provided me with many ways to find out the solutions, for example, by having meals together, the short excursion, and the game called the escape room. Some open problems I found interesting during the seminar are:

1. mechanism of human cognition that produces, being affected by mis- and dis- information.
2. computational models that reveal justifications of human beliefs about information. It would be a good starting point to accept the fact that no recognition is possible without prejudice, as the "ugly duckling theorem" tells us.

5.3 My Background and Work on Critical Online Reasoning

Dimitri Molerov (Universität Mainz, DE)

I attended the Dagstuhl Seminar "Technologies to Support Critical Thinking in an Age of Misinformation" by recommendation from Prof. Andreas Dengel's office. His collaboration with my supervisor Prof. Olga Zlatkin-Troitschanskaia (economics education (educational assessment) JGU Mainz) on the cross-university initiative Positive Learning in the Age of Information (`https://www.plato.uni-mainz.de/`) had led to two interdisciplinary Springer volumes [1], including contributions by seminar participants. The "Age of Misinformation" phrase from prior presentations may have inspired the seminar's title, too. The focus was on scoping improvements to learning in higher education in the face of increasing self-directed online learning, as well as a need for more evidence-based reasoning in regard to the Internet (Asking not only what you know, but how the discipline found out about it). Two aspects have made it a priority topic in our research group. A) Media use surveys in the initiative showed students' use of online sources for learning surpassed their use of offline resources (e.g., scripts, textbooks) (this was even pre-pandemic)[2]; B) the Internet as an uncurated space for learning inputs (and mostly ignored space for educational research, apart from work on curated e-learning) with all its high- and low-quality information and its preselection and

distribution, attention-grabbing, addiction-reinforcing, and polarization mechanisms partly opposing learning preconditions (see HumaneTech). The general idea has been that students as Internet users need a specific skill-set for successfully acquiring reliable knowledge – the umbrella for the collaboration has been on how to model, measure, technologically support and foster necessary skills for students' self-directed learning online.

In our research group, we have meanwhile specified a skill-set for assessment as Critical Online Reasoning [3]. The concept follows known phase models, e.g., Information Problem Solving on the Internet [4, 5], including a search/information acquisition facet, and a critical evaluation facet (modeled as identifying cues to credibility or deficiency in online information), but also specifies a critical reasoning facet (weighing evidence, drawing conclusions) and expanding on the "activation" and monitoring (when do we even apply critical reflection, given that it takes mental effort and we are all cognitive misers). In essence, the additional thinking and behavior one has to undertake to ascertain information quality when one suspects that perceived information and consulted sources may not be entirely dependable. Another framing would be skills for discriminating dependable information from misinformation online. Various assessments exist; a novel approach has been to adapt the Civic Online Reasoning Assessment [6], which features the actual Internet and (sometimes dubious) websites to be searched and evaluated, and thereby affords quick and ecologically valid creation of test item stimuli. My PhD research revolves around the adaptation, modeling, and test item design for such skills and connection to critical thinking skills. It was very heartening to meet colleagues who are systematizing the various approaches to fostering skills.

In this vein, the work of the computer-science and developer community has been complementary. For one, applications are built to implement educational theories and models into learning support tools, and critical thinking online is supported in many other ways, e.g., automatic detection of misinformation in specific media formats. As I mentioned, to detect misinformation, "someone has to do the thinking", the computer or the human user; and our job may be to (re)negotiate the share of each, as the information landscape keeps evolving, and help optimise human-computer interaction. The study on image reverse detection and automatic emotional labeling was an impressive presentation.

I came to Dagstuhl also to gauge interest in the following project, i.e., if someone would like to digitize and gamify labels for (mis)information. The inspiration is John Cook's work on FLICC – who collected arguments by climate deniers, distilled them into common persuasion techniques (e.g., argument patterns, tropes, fallacies) [7], optimised them didactically using graphic icons, and also used the labels in an educational multiple-choice quiz game (Cranky Uncle game) [8]. One next step could be to take such labels to actual social or news media and either offer a computational pre-labeling, or more interestingly, enable users to label their own and/or peer's content. Having assigned "epistemic labels", a user could review their often evasive initial judgments of a piece of content or even single statements made in a chat (e.g., reminding themselves of initial vague irritation) at a later point and come back to reflect on it more thoroughly.

Social uses are envisioned, as well, from learning games to live collaboration on information evaluation, such as in crowd-sourced fact-checking. Epistemic labels (defined in a publicly accessible scheme or library) can go beyond verifiable facts in also highlighting undesirable or baseless persuasion techniques, which are not strictly falsifiable, but still say misleading and would warrant a warning (e.g., a "citation needed"/reflection needed flag from Wikis or a friendly bias reminder). Epistemic labels can be implemented as emojis, but rather than emotional responses, they would represent cognitive judgment snippets. Here, a future design challenge can be to define epistemically grounded, generative labels and set fair rules for

their interpretations, e.g., to gamify error culture or allow some space for people to learning democracies (Here, philosophers and logicians have mapped out a good part of the agenda and formalized reasonable discourse and truth conditions. The inquiry into truth conceptions is far from resolved, but a minimal consensus around wanting to be internally consistent and avoid basic fallacies can already be enough for developing tools and making progress in public discourse).

Disinformation campaigns are a severe risk on one end of a spectrum; on the other end, we find censorship, national/ally/block cultural media bubbles – which on a global scale can be as polarizing, as well: exchanging national for cross-national polarization). Equally some participants reflected back to me the apparently not so rare we-know-it-all-and-will-teach-you-the-right-way attitude or trap that we as designers of cognitive training, such as the inoculation approach, can fall into (and which I try to address more thoroughly in the PhD – the short response can be external bias checks, non-domination in education, and stressing user's capacity development as inherent self-interest). As designers, we may possibly ignore our own biographical, cultural, method biases – e.g., shouldn't it seem too one-sided to safeguard against Russian disinformation only, but remain blissfully unaware of own embedding in other national media diets and forget about past, recent (and maybe unknown) present propagandist efforts by governments, militaries, international business conglomerates, one's administration, and unwittingly participating compatriots. As one of four-five schools of critical thinking research – apart from logical (syllogisms, fallacies focus; recently computational argument), psychological (biases, emotions), educational (mix and content focus), and media scientific insights – the Frankfurt school of critical thinking has been strong in the humanities, and, e.g., with spin offs in critical pedagogy, offering criticisms of surrounding societal power structures that shape discourses. An integration into technology support and assessment seems to be still pending, e.g., in the form of a decision aid when to think critically about a range of granularities from the small everyday mental operations to large global systems (where algorithmic biases come up at pain points) to the metacognitive reflection of when not to overthink. Higher critical thinking requirements, involving criticism of self or own culture are difficult to assess within a government-dependent educational system and risky to design for responsibly. Do we need to acknowledge neo-Imperialism, as Noam Chomsky and Ray Dalio will have us, and a consequential imbalance of consumed cultural content, perhaps even embrace it as unavoidable anthropological evolution resulting of a human drive for power or excellence that aggregates, or do we reject domination attempts in the political, and particularly the digital sphere, as artifacts of last-century public administrative personnel that limits current human development? Reaching the big questions has been easy at Dagstuhl. Coming back out with organized and fun research designs was the grittier, but well-scaffolded part of the seminar.

How do we approach our task exactly? Researchers' work might also stretch beyond just picking a favorite approach between creating knowledge for the privileged few and to educate someone who does not know better. If we take a society-wide view, the work can be about strengthening mental capacities within one another (and building socially reinforcing systems), leaving meetings with the best available knowledge and skills within the largest possible N (including accepting a remainder group), and validating whether the individual reasoners' autonomy is preserved and strengthened and they feel their concerns have been truly addressed – which affords them some relaxation and emotional safety to approach more daunting questions of truth in information. As social psychology indicates, motives for misinformation consumption are often not cognitive, but a symptom of social and psychological conditions (e.g., power distance, lacking self-efficacy); however, pathologizing misinformation consumption would take away the individual's autonomy and the opportunity

to tap into their resources, as well as blanket their possibly legitimate concerns. Telling non-scientists to "trust the science" ignores the many cases of misinformation in science, the somewhat fewer scandals, the somewhat large paradigm shifts under way in a given set of disciplines at any time, basic research failures, interest, and incentive structures, and the inability of outsiders to quantify the magnitudes. The confusion may seem daunting to resolve, but can be more easily referenced by distinguishing disciplinary/within-method knowledge gain (critical question: am I being methodologically rigorous, avoiding thought traps), from interdisciplinary/cross-method knowledge gain (critical thinking: does my method apply to the problem? How much do I gain from different approaches?).

Not only preventing misinformation intake, supporting reevaluation of contaminated mindware. It is illusory (and possibly limiting) to safeguard users from being exposed to or rejecting any and all encountered online misinformation; some of it will find a way to seep in. Perhaps, the discussion needs to shift to (tolerable) percentages and thresholds of within- and between person misinformation. It seems equally important to admit that users have already been confused from different sources and support them in learning to regularly reevaluate their acquired misconceptions and "contaminated mindware"[9]. Debiasing and debunking are two successful approaches discussed.

Another still undervalued bundle of approaches is highlighting and insisting on positive conversation and evidence standards and strengthening virtuous communication techniques and patterns. How can these be supported technologically needs further discussion? This one has the advantages of refocusing conversation from problems to existing communicative solutions, being less threatening to the ego, and addressing prevalent cultural skepticism in other's intellectual rigor with grounding.

Overall, the visit to Dagstuhl helped me better understand some of the concepts of prior work on interventions against misinformation (e.g., inoculation theory, pre/debunking, debiasing), get a glimpse of what is being done on the technology and legal side, and meet important proponents and seasoned experts (e.g., Stephan Lewandowsky), while fleshing out project ideas. The sense of community-building and not having to tackle huge challenges alone was as nourishing as the Dagstuhl menu, beautiful nature, and the deep calm of the information scientist. The Outing was a special treat – pondering on media consumption habits and intellectual humility – while having your (lack of) misinformation (trivia knowledge) handed back to you. Much appreciated!

At the seminar, our work group agreed that labeling of information quality at several layers was one important goal for future projects, though precise frameworks are still scarce.

My proximate contribution going forward has been to collect and attempt to classify types of dis- and misinformation, together with example cues, and the search process phase they occur in. I aim to provide this in wiki format. Feedback on how to make classes easily machine-referencable will be much appreciated.

References

1 Zlatkin-Troitschanskaia, O. (Ed.). Frontiers and Advances in Positive Learning in the Age of InformaTiOn (PLATO). Cham: Springer International Publishing. 2020

2 Maurer, M., Quiring, O., & Schemer, C. Media Effects on Positive and Negative Learning. in Zlatkin-Troitschanskaia, O., Wittum, G., & Dengel, A. (Eds.). Positive Learning in the Age of Information. Wiesbaden: Springer Fachmedien Wiesbaden. 2018.

3 Molerov D., Zlatkin-Troitschanskaia O., Nagel M., Brückner S., Schmidt S., Shavelson R.J. Assessing University Students' Critical Online Reasoning Ability: A Conceptual and Assessment Framework With Preliminary Evidence. Frontiers in Education. 2020. https://doi.org/10.3389/feduc.2020.577843

4 Brand-Gruwel, S., Wopereis, I., Vermetten, Y. Information problem solving by experts and novices: analysis of a complex cognitive skill. Computers in Human Behavior. 2005. https://doi.org/10.1016/j.chb.2004.10.005
5 Brand-Gruwel, S., Wopereis, I., Walrave, A. A descriptive model of information problem solving while using internet. Computers & Education. 2009. https://doi.org/10.1016/j.compedu.2009.06.004
6 Wineburg, S., McGrew, S. Evaluating information: The cornerstone of civic online reasoning. Stanford History Education Group. 2016. https://apo.org.au/node/70888.
7 Cook, J. Deconstructing climate science denial. Research handbook on communicating climate change, 62-78. 2020. https://doi.org/10.4337/9781789900408.00014
8 Cook, J. Cranky uncle vs. climate change: How to understand and respond to climate science deniers. Citadel Press. 2020.
9 Stanovich, K. E. The comprehensive assessment of rational thinking. Educational Psychologist, 51(1), 23-34. 2016.

5.4 Critical Thinking and Misinformation in Academic Research

Andrew Vargo (Osaka Prefecture University – Sakai, JP)

Most of the research regarding misinformation that is spread online typically focuses on news and fake news (generally of a political or societal nature). While this is certainly an important and interesting topic, my focus is in knowledge-sharing in technical domains. One of the most fruitful aspects of the Seminar was the wide-ranging discussions held about trust, authority, and misinformation in academic research. In an age in which numerous papers use opaque data mining techniques and questionable data science, it is difficult to assign veracity to each research article. Doing so requires both technical (the academic field may vary) and domain expertise (the area on which the data is extracted may require specialized information) from a reader to call into question what is likely true and what is possibly not. The group discussion was very interesting and participants explored their experiences with problematic research areas and claims. It seems that what we think we know is often a product of repetition. If venues publish and promote research that have questionable conclusions based on flawed methodology, this can perpetuate more research using the same flawed techniques. This eventually creates something that we think as objective truth in the field.

This has spurred an interest in investigating the network relationship between published papers we can identify as having questionable methodologies and conclusions and their impact on the wider research community. This hopefully will uncover how deeply misinformation spreads in academic research and allow us to develop critical thinking tools for academics.

Participants

- Chris Coward
University of Washington – Seattle, US
- Henriette Cramer
Spotify – San Francisco, US
- Andreas Dengel
DFKI – Kaiserslautern, DE
- Tilman Dingler
The University of Melbourne, AU
- David Eccles
The University of Melbourne, AU
- Nabeel Gillani
MIT – Cambridge, US
- Koichi Kise
Osaka Prefecture University, JP
- Dimitri Molerov
Universität Mainz, DE
- Albrecht Schmidt
LMU München, DE
- Gautam Kishore Shahi
Universität Duisburg-Essen, DE
- Benjamin Tag
The University of Melbourne, AU
- Roger Taylor
Open Data Partners – London, GB
- Niels van Berkel
Aalborg University, DK
- Andrew Vargo
Osaka Prefecture University – Sakai, JP
- Eva Wolfangel
Stuttgart, DE

Remote Participants

- Susanne Boll
Universität Oldenburg, DE
- Nattapat Boonprakong
The University of Melbourne, AU
- Laurence Devillers
CNRS – Orsay, FR & Sorbonne University – Paris, FR
- Stephan Lewandowsky
University of Bristol, GB
- Philipp Lorenz-Spreen
MPI for Human Development-Berlin, DE
- Emma Spiro
University of Washington – Seattle, US