DAGSTUHL
REPORTS

Volume 12, Issue 8, August 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
March, 2023

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

Digital Object Identifier: 10.4230/DagRep.12.8.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 22331

Visualization and Decision Making Design Under Uncertainty

Nadia Boukhelifa*[1], Christopher R. Johnson*[2], and Kristi Potter*[3]

1 **INRAE – Palaiseau, FR.** nadia.boukhelifa@inrae.fr
2 **University of Utah – Salt Lake City, US.** crj@sci.utah.edu
3 **NREL – Golden, US.** kristi.potter@nrel.gov

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 22331 "Visualization and Decision Making Design Under Uncertainty". The seminar brought together 33 researchers and practitioners from different domains concerned with visualization and decision making under uncertainty including visualization, visual analytics, human-computer interaction, artificial intelligence, climate research, geography and geology. The programme was organized in two parts: In the first part which lasted two days, participants gave short talks where they discussed current practices and the uncertainty visualization challenges they encountered in their own research. At the end of day two, participants brainstormed collectively around the main uncertainty visualization research challenges across domains and applications. In the second part, participants voted for the following three main challenges they wished to discuss for the remainder of the seminar (one and a half days): applications, human-centered uncertainty visualization, a design process for uncertainty visualization. Thus three break-out groups were formed to discuss these challenges. Abstracts for the individual talks and the break-out group activities are included in this report.

Seminar August 15–19, 2022 – http://www.dagstuhl.de/22331
2012 ACM Subject Classification Human-centered computing → Visualization
Keywords and phrases Decision making, Uncertainty visualization, Visual Analytics, Visualization
Digital Object Identifier 10.4230/DagRep.12.8.1

1 Executive Summary

Nadia Boukhelifa (INRAE – Palaiseau, FR)
Christopher R. Johnson (University of Utah – Salt Lake City, US)
Kristi Potter (NREL – Golden, US)

Uncertainty is an important aspect to data understanding. Without awareness of the variability, error, or reliability of a data set, the ability to make decisions on that data is limited. However, practices around uncertainty visualization remain domain-specific, rooted in convention, and in many instances, absent entirely. Part of the reason for this may be a lack of established guidelines for navigating difficult choices of when uncertainty should be added, how to visualize uncertainty, and how to evaluate its effectiveness. Unsurprisingly, the inclusion of uncertainty into visualizations is a major challenge to visualization [1]. As work concerned with uncertainty visualization grows, it has become clear that simple visual

* Editor / Organizer

Visualization and Decision Making Design Under Uncertainty, *Dagstuhl Reports*, Vol. 12, Issue 8, pp. 1–19
Editors: Nadia Boukhelifa, Christopher R. Johnson, and Kristi Potter
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

additions of uncertainty information to traditional visualization methods do not appropriately convey the meaning of the uncertainty, pose many perceptual challenges, and, in the worst case, can lead a viewer to a completely wrong understanding of the data.

The goal of this Dagstuhl Seminar was to bring together experts with diverse knowledge of uncertainty visualization and comprehension toward building a foundation of accessible, practical knowledge that practitioners and researchers alike can rely on in addressing challenges related to uncertainty. Specifically, this seminar brought together leaders in the field of uncertainty visualization and communication, along with experts on quantification and practitioners and domain experts dealing with uncertainty on a daily basis. Drawing on the knowledge of the participants, the seminar worked toward goals of synthesizing disparate findings and approaches from across computer science and related literature, noting current practices surrounding uncertainty, and identifying unsolved problems in common workflows, and areas needing further study.

As a major result from the seminar, the following challenges and research topics in visualization and decision making under uncertainty have been identified:

- Applications,
- Human-centered uncertainty visualization (including how to support "feeling uncertain"),
- A design process for uncertainty visualization,
- Defining terms related to uncertainty,
- Algorithms and uncertainty quantification,
- Software dissemination,
- User studies,
- Ethics of uncertainty (when to include uncertainty information),
- Surveys of uncertainty-aware visual analytics, and
- Teaching uncertainty visualization.

The top three challenges were discussed in depth during this intensive three and a half days Dagstuhl Seminar as part of the break-out groups, and are further discussed in this report. In particular, the break-out groups examined uncertainty visualization research challenges from three complementary perspectives: from an application viewpoint looking at how uncertainty visualization and assessment are used in many domains; from a human-centered perspective considering the needs and information of the viewer; and from a more theoretical stand focusing on the problem space for designing uncertainty visualization.

The seminar ended with a presentation from each group and discussions on the next steps. Interesting research questions and potential solutions were identified during the discussions, and plans were made to continue the collaboration. Details of the individual talks and break-out group discussions are provided in this report.

References

1 Chris R. Johnson and Allen R. Sanderson. *A next step: Visualizing errors and uncertainty*. IEEE Computer Graphics and Applications, 2003, vol. 23, no 5, p. 6-10.

2 Table of Contents

3 Overview of Talks

3.1 Statistical Analysis for Uncertainty Quantification and Visualization of Scientific Data

Tushar Athawale (Oak Ridge National Laboratory, US)

Joint work of Tushar Athawale, Chris R. Johnson

Data visualization has become indispensable for efficient interpretation of complex data generated across diverse scientific domains, such as biomedical imaging and meteorology. Many critical decisions directly rely on the quality of data visualizations. Inaccuracies in visualizations cannot be averted due to uncertainties inherent in underlying data and non-linear transformations of data caused by the stages of the visualization pipeline. The uncertainty in the final visualizations can adversely impact the decision-making process. The accurate quantification of uncertainties in data visualizations has, therefore, been recognized as the top research challenge for minimizing risks associated with scientific decisions.

In this talk, I will present the abstract statistical methods for uncertainty visualization and a few uncertainty visualization applications. My main topics of discussion are as follows: 1) Need for uncertainty visualizations, 2) abstract statistical methods for uncertainty quantification, 3) a few applications of uncertainty visualization to key scientific visualization techniques, such as fiber surfaces and Morse complexes, and domain-specific data, e.g., biomedical imaging, 4) open research challenges in uncertainty visualization. Our experimental results relevant to uncertainty visualizations confirm the significance or need for incorporating statistical error analysis into computational models for visualization applications.

3.2 A Tentative List of Uncertainty Visualization Research Challenges

Nadia Boukhelifa

Uncertainty visualization research has made considerable progress leading to a variety of techniques, algorithms, systems, frameworks and user studies. The goal of this talk is to provide a preliminary list of open problems and challenges that our visualization community has been focused on in the last 20 years. I present findings from a literature survey of 17 papers from 2002 -2022, covering multiple domains including scientific visualisation, information visualization and visual analytics. I focus on surveys, state-of-the-art reports, viewpoint articles and position papers rather than on papers on specific techniques, algorithms, systems or user studies.

The results of this survey shows eight main areas of open challenges related to conceptualisation, evaluation, formalisation and theory, quantification, representation, training and dissemination, uncertainty-aware tools, and user Interaction. Some of the found challenges may have already been solved, and new ones may not yet have been fully documented. There is a need to review progress of the field of uncertainty visualization across domains, and to highlight success stories, long-standing problems as well as emerging and new ones.

3.3 Visualization of Climate Simulation Data and related Uncertainty

Michael Böttinger (DKRZ Hamburg, DE)

Climate models simulate the most important processes governing the climate system, i.e. the coupled system of atmosphere, ocean, sea ice, land-biosphere and ocean-biogeochemistry. Simulations result in 3D time-dependent multivariate data sets, characterized by high variability at various time scales. Internal variability of the coupled climate system additionally contributes to this noise. However, the high variability reduces the signal-to-noise ratio, thus makes it hard to detect climate change signals. Analyzing and visualizing climate change in the presence of noise is challenging, but with ensemble simulations, the signal-to-noise ratio can be enhanced and the internal climate variability assessed. I present examples from climate change research that show the visualization of robustness in the presence of a highly variable field. However, with respect to the climate change to 2100, the largest uncertainty is in the range of possible evolutions of the socio-economic system. Furthermore, I show visualizations of the CMIP6 multi model ensemble of simulations conducted globally with regard to the 6th IPCC report that capture this range through a range of scenarios describing different socio-economic development pathways. Finally, I briefly present recent collaborative work with Gerik Scheuermann's group to highlight the challenges in the visualization of uncertain topology-based features for highly variable complex phenomena such as the North Atlantic Oscillation and its evolution in a changing climate.

References

1 Vietinghoff, D., Heine, C., Böttinger, M., Maher, N., Jungclaus J., and Scheuermann, G. *Visual Analysis of Spatio-Temporal Trends in Time-Dependent Ensemble Data Sets on the Example of the North Atlantic Oscillation*. 2021 IEEE 14th Pacific Visualization Symposium (PacificVis), 2021, pp. 71-80, doi:10.1109/PacificVis52677.2021.00017

2 Vietinghoff, D., Böttinger, M., Scheuermann, G. and Heine, C *Detecting Critical Points in 2D Scalar Field Ensembles Using Bayesian Inference* IEEE 15th Pacific Visualization Symposium (PacificVis), 2022, pp. 1-10, doi:10.1109/PacificVis53943.2022.00009.

3.4 The Impossibility of Zero: Effects of Individual Differences in Medical Decision Making

Remco Chang (Tufts University – Medford, US)

Making decisions that might affect a person's long-term physical wellbeing can be difficult and stressful. As most medical diagnosis contains some amount of uncertainty (including type I and type II errors), it is often up to a patient to assess their own comfort level with different treatment options. In this talk, I present three challenges relating to medical decision making through the perspective of the patients, namely risk communication, reasoning with conditional probability, and visualization design for decision making.

First, I present a design study of a visualization tool for communicating a patient's prostate cancer risk. After interviewing 6 prostate cancer patients and two urologists, we iteratively designed the visualization based on the participants' feedback. Our takeaways

from this design study include: (1) prostate cancer patients (who tend to be older men) have trouble using even basic visualizations (e.g. bar chart, stacked area chart, scatterplot, etc.). Text explanations that accompany the visualizations are a must. (2) Emotion and stress can affect a patient's ability to reason about their diagnosis. After receiving a positive diagnosis, a patient often has limited cognitive capacity to think through the diagnosis rationally. (3) Most Patients' first question after receiving a positive diagnosis is "how much time do I have left," suggesting that there's an order to the presenting of information that can best meet the patients' decision-making needs.

Second, I present an experimental study on people's ability to reason about their diagnosis as conditional probabilities. Most screening tests contain some amount of uncertainty, in particular as type I and type II errors. When a patient is told that they have a positive diagnosis for a disease, it is often up to the patient to reason through these probabilities to assess what their "true risks" are. In our experiment, we tested 6 visualization designs that were accompanied by text explanations. Our initial analysis of the results found no statistical significance between the effectiveness of the 6 visualizations. However, when the participants were stratified based on their spatial ability scores (as measured using the paper-folding test), we found that some of the visualizations are very effective (near 100% accuracy) for the participants with high spatial abilities. Unfortunately, we found no visualization that was helpful for the participants with low spatial abilities.

Lastly, I discuss the challenges in designing visualizations for helping patients make difficult medical decisions. For example, a patient might not perceive any difference between a diagnosis with 30% or 31% of having a disease. However, when the difference is between 0% and 1% chance of having a disease, the same difference of 1% becomes more significant to a patient as it represents "not having a disease" versus "possibly having a disease." A visualization will need to incorporate individuals' risk perception and risk tolerance utility curves to best support their decision making process.

3.5 Underthinking Uncertainty Visualization

Michael Correll (Tableau Software – Seattle, US)

Uncertainty visualization is viewed as a hard problem. Sources of these difficulties include complexity and disagreement around how uncertainty is modeled or quantified and clashes between between idealized forms of decision-making and the actual behavior of human beings. It is true that these are problems. But we can't wait for statisticians and psychologists to settle all of their internal disputes on these topics; people have decisions to make today. What we can do, however, is find solutions that are likely to be generally good enough for many practical purposes.

In this talk I will introduce a framework for ways to address uncertainty without having to think too hard, specifically around leveraging the ability of people to estimate statistical properties in visualizations without additional scaffolding, and the ability of visualization designers to "nudge" these estimates to align with statistical models of decision-making without being dogmatic or domineering. This is good news for uncertainty visualization as a discipline in that it does not require either designers or viewers of visualizations to be perfectly rational statistical deities to get their work done, but perhaps bad news in that

we now have to do much more work as a field to build a deeper understanding of graphical perception for "fuzzier" tasks, take stronger stances around desired behavior from viewers of visualizations, and to better integrate statistical models, models of inference, and rhetorical goals into our design thinking.

3.6 Uncertainty in Public Policy Decision Making

Stephanie Deitrick (Arizona State University – Tempe, US)

Public policy decision makers leverage both qualitative and quantitative data as part of their decision-making processes. With increased interests in science-based information and leveraging data for their decisions, agencies are often expanding their workforce to include more data scientists and partnering with researcher on a variety of topics. While policy makers understand that data are uncertain at some level, that may not be something they explicitly consider as part of how they currently leverage data.

Since data are often communicated through visualization, such as maps and charts, should uncertainty be part of that communication? Would it produce better or more informed decisions?

3.7 Uncertainty-aware Visual Analytics

Christina Gillmann (Universität Leipzig, DE)

Visual analytics has been successfully applied to a variety of applications also in terms of uncertainty analysis. Unfortunately, the visual analytics process does not include a mechanism to systematically handle uncertainty. In order to solve this issue, we developed the concept of uncertainty-aware visual analytics. Therefore, an extension of the classic visual analytics cycle is achieved that includes the quantification of uncertainty in each component, the exchange of analysis and visualization approaches in general by uncertainty-aware options and the introduction of provenance to monitor the accumulation and propagation of uncertainty throughout the visual analytics cycle. In order to create uncertainty-aware visual analytics cycles for particular applications, we determined a workflow that consists of 5 steps that constructs an uncertainty-aware visual analytics cycle starting from the classic approach. The procedure is based on a developed taxonomy of uncertainties that allow to understand the nature of different uncertainty events and their effect on the visual analytics cycle.

3.8 Visualizing uncertainty in digital geologic map databases

Amy Gilmer (USGS – Denver, US) and Kathleen Warrell (UCAR – Boulder, US)

The geologic map remains the primary tool geologists use to model and communicate what we know about Earth's surface. All geologic models contain some level of uncertainty, but this uncertainty is rarely incorporated in traditional geologic maps, potentially limiting application by decision makers. Even as our geological depictions have migrated to digital geologic map databases, our map symbology has largely remained the same as that used on traditional paper maps. While varying dash length for contacts and faults may convey a relative sense of uncertainty to experienced users, it does not convey meaning to the nonexpert user. Cartographic uncertainty visualizations are an effective way to communicate how well we know what and where something is.

The adoption of the Geologic Map Schema (GeMS) standard for geologic maps has enabled geologists to capture feature-level metadata, including location uncertainty, as well as feature identity and existence confidence. To visually communicate the underlying locational uncertainty in the USGS Intermountain West geological framework database, we have developed an ArcGIS Python toolbox that extracts existing location confidence data from feature attributes, and then buffers and aggregates the uncertainty across a tessellation grid. The tessellation grid can then be visualized by any of the statistical fields generated. This toolbox can be applied to any geologic map database adhering to the GeMS format to produce visualizations summarizing uncertainty. While there is still much we can do to refine how we quantify uncertainty in mapping geologic features, this type of visualization, when provided alongside the geologic map data, summarizes the uncertainty without requiring the user to understand the nuances of traditional map cartography. Additionally, this quantitative approach can help identify areas characterized by high levels of uncertainty, potentially a result of low-resolution map data, that can be used for geologic mapping needs assessments and to better inform end users to limit improper use of the map data.

3.9 Summarization, Uncertainty, Estimation. . . : Models as a basis for visualization

Michael Gleicher (University of Wisconsin-Madison, US)

Models are part of most (if not all) use of data. However, they are often hidden or implicit. We often expect viewers to figure out what they are, estimate their parameters, and apply them correctly to achieve their goals. I argue that models should be a first class citizen in how we help people work with data. Many problems, including uncertainty, seem to be made worse because the models are hidden. Many concepts, such as summarization, estimation, and uncertainty are often conflated, especially when models are hidden. My conjecture is that by having a better way to include models in our thinking and by de-conflating the key terms, we can better discuss, design, and evaluate tools to help people work with data.

3.10 Uncertainty in Definition

Hans-Christian Hege (Zuse-Institute Berlin, DE)

Facts about the world are mainly represented linguistically. The building blocks of language are concepts, both concrete and abstract. Concepts help us humans to organize, understand and explain the world. We use them in cognitive processes such as categorization, reasoning, and decision-making, as well as in explanation and communication. An important task of data science/visualization is to connect the world of data and the world of concepts by finding equivalents of the concepts in the data.

However, concepts are only defined in language and often imprecisely. This leads to "uncertainty of definition". Metaphorically speaking, concepts are not points in conceptual space, but rather regions with blurred boundaries. Examples: What exactly is a vortex in a flow? What exactly is the spatial extent of a vortex? Which patients are considered to have died from COVID-19 as opposed to patients who died with COVID-19? What exactly is an epidemic wave and what is not? Which atmospheric phenomenon is a hurricane and which is not? Almost every statistic or visualization is preceded by such questions. Different answers are possible, but they lead to different results: the definition uncertainties propagate into the results. If we take into account the uncertainties in definition, we get ensembles of results.

We should be aware of this type of uncertainty, capture it and its propagation into the results, communicate it, and reduce it. Visualization can help with the latter by showing the variance that results from different definitions of the concepts.

References

1 Natalia Mikula, Tom Dörffel, Daniel Baum, Hans-Christian Hege. *An Interactive Approach for Identifying Structure Definitions.* Computer Graphics Forum, 41:3, pp. 321-332 (2022), DOI: 10.1111/cgf.14543

3.11 Visualization and Analysis of XCT Data – Decision Making under Uncertainty

Christoph Heinzl (Universität Passau, DE)

Visualization and analysis of "rich" X-ray computed tomography (XCT) data has become highly attractive for boosting research endeavors in the materials science domain. On the one hand, XCT allows to generate detailed and cumulative data of the specimens under investigation in a non-destructive way. On the other hand, through the conception, the development, and the implementation of novel, tailored analysis and visualization techniques, in-depth investigations of complex material systems turned into reality.

This talk presents contributions to computer science in terms of design studies, methods, and techniques, which are advancing visual analysis and visualization for enabling insights into "rich" XCT data. The introduced methods and techniques focus on three distinct technical areas of visual analysis and visualization of XCT data, which are interactive visualization of spatial and quantitative data, visual parameter space analysis of respective data processing and visualization pipelines, uncertainty and sensitivity analysis. For each area, the problem statements, important research questions to be solved as well as some of the author's contributions thereto are discussed.

References

1 A. Reh, C. Gusenbauer, J. Kastner, E. Gröller, C. Heinzl. *MObjects – A Novel Method for the Visualization and Interactive Exploration of Defects in Industrial XCT Data*. IEEE Transactions on Visualization and Computer Graphics, Vol. 19, No. 12, 2013, pp. 2906-2915

2 A. Reh, A. Amirkhanov, J. Kastner, E. Gröller, C. Heinzl. *Fuzzy Feature Tracking: Visual Analysis of Industrial 4D XCT Data*, In Computers & Graphics vol. 53, Part B, 2015, pp. 177–184

3 B. Fröhler, T. Möller, C. Heinzl. *GEMSe: Visualization-Guided Exploration of Multi-channel Segmentation Algorithms*. Computer Graphics Forum, vol. 35, 2016, pp. 191-200

4 B. Fröhler, T. Elberfeld, T. Möller, HC Hege, J. Weissenböck, J. De Beenhouwer, J. Sijbers, J Kastner, C. Heinzl. *A Visual Tool for the Analysis of Algorithms for Tomographic Fiber Reconstruction in Materials Science*. Computers Graphics Forum vol. 38 (3), 2019, pp. 273-283

5 J. Weissenböck, B. Fröhler, E. Gröller, J. Kastner, C. Heinzl. *Dynamic Volume Lines: Visual Comparison of 3D Volumes through Space-filling Curves*. IEEE Transactions on Visualization and Computer Graphics, vol. 25 (1), 2019, pp. 1040-1049

6 A. Amirkhanov, C. Heinzl, C. Kuhn, J. Kastner, E. Gröller. *Fuzzy CT Metrology: Dimensional Measurements on Uncertain Data*. Proceedings of the 29th Spring Conference on Computer Graphics, 81-90, 2013

7 B. Fröhler, T. Elberfeld, T. Möller, H.C. Hege, J. Maurer, C. Heinzl. *Sensitive vPSA– Exploring Sensitivity in Visual Parameter Space Analysis*. arXiv preprint arXiv:2204.01823

3.12 Designing, de-"bias"ing, and de-probabilizing uncertainty visualization

Matthew Kay (Northwestern University – Evanston, US)

I discuss three challenges in uncertainty visualization: (1) how do we design uncertainty visualizations systematically? (2) how do we (and should we) de-bias uncertainty visualizations? (3) how do we visualize possibilistic and qualitative forms of uncertainty?

3.13 Centering Uncertainty on People

Miriah Meyer (Linköping University, SE)

From a perspective of data as a situated perspective – one that is inherently partial and incomplete – knowledge about the shortcomings of data is often known by domain experts. In recent work we propose a framing of this knowledge as data hunches, and argue that hunches are a source of qualitative uncertainty. Acknowledging and valuing the hunches people bring to visual analysis opens new opportunities to design visualization tools that support people in externalizing and communicating their hunches.

3.14 Actionable Uncertainty Visualization

Kristi Potter (NREL – Golden, US)

Ensemble simulations capture the variability present in the predictions of future states by combining multiple runs of a computational model with different parameter settings. Datasets derived from ensemble simulations are often quite large and complex, making it hard to create visualizations that facilitate decisions, particularly for people not intimately involved with the scientific domain or the creation of the dataset. An example comes from the renewable energy space, where improvements to the electrical grid to facilitate the large-scale reduction of carbon emissions involves making decisions on highly complex systems. Traditional methods for uncertainty visualizations primarily focus on the challenge of visually presenting large-scale, high-dimensional datasets in an exploratory manner. However those approaches do not facilitate decision making by non-experts, such as policy-makers, who may not know enough about the computational system to appropriately choose appropriate parameter settings to achieve a desirable outcome. In this talk I will discuss ideas for distilling down the parameter space by importance, annotating contextual information needed for better understanding, and designing a visualization tool that is streamlined for decision-making.

3.15 A design theory for uncertainty visualization?

Maria Riveiro (Jönköping University, SE)

Despite the large volume of research on uncertainty visualization, we do not fully understand the impact of uncertainty visualization on decision-making. There is evidence of both positive and negative effects of visually depicted uncertainty on decision-making.

This talk presents examples of evaluations carried out with practitioners in various application areas, including autonomous driving, air traffic risk assessment and maritime surveillance. I summarise the effects of the uncertainty visualizations provided on the users and their decision-making processes in these evaluations.

Finally, we discuss the need for a design theory/space of uncertainty visualization and elaborate on the multiple dimensions/variables that such a design space should have.

References

1 Helldin, T., Falkman, G., Riveiro, M. and Davidsson, S. (2013) Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. Proc. 5th Int. Conf. on Automotive User Interfaces and Interactive Vehicular Applications (Automotive'UI 13), Eindhoven, The Netherlands.
2 Riveiro, M., Helldin, T., Falkman, G. and Lebram, M. (2014) Effects of visualizing uncertainty on decision-making in a target identification scenario, Computers & Graphics, Volume 41, Pages 84-98, Elsevier, ISSN 0097-8493.
3 Riveiro, M. (2016). Visually supported reasoning under uncertain conditions: Effects of domain expertise on air traffic risk assessment. Spatial Cognition & Computation: An Interdisciplinary Journal, vol. 16(2), pp. 133-153. Taylor & Francis.

4 Kinkeldey, C., MacEachren, A. M., Riveiro, M., & Schiewe, J. (2017). Evaluating the effect of visually represented geodata uncertainty on decision-making: systematic review, lessons learned, and recommendations. Cartography and Geographic Information Science, 44(1), 1-21.

3.16 Critical Points of an uncertain Scalar field

Gerik Scheuermann (Universität Leipzig, DE)

Critical points like extrema or saddles are a well established concept in (deterministic) scalar field visualization. There is strong practical interest, a clear mathematical concept in continuous and discrete settings, and corresponding algorithms, including implementations in commercial systems. Looking at uncertain scalar fields, formally described as smooth stochastic processes, practically often given as ensembles over a common grid, the situation changes. The concept of a "critical point" is not exactly defined. Two major definitions are "critical point of the (deterministic) mean field" or "probability distribution of critical points in a sample from the stochastic process/ensemble". The choice of concept definition has effects on the visualization and its interpretation, like "a maximum being multiple (significant) maxima in one sample and no significant maximum in another case". Also, the sampling quality of the ensemble should somehow be integrated into the visualization. The talk concern these issues. Looking at the distribution definition, I show how to infer critical point distributions from ensembles using Bayesian Inference. Looking at the mean field definition, I will discuss how bootstrapping allows to reason about the sampling quality to derive significant results. Finally, the talk shows how this allows to decide two practical questions regarding the future of the north atlantic oscillation (NAO) depending on Climate Change. (NAO describes they interplay between Iceland Low and Azore High – which is the most dominant factor in European winter weather.) We derive that the centers of action of both pressure systems move substantially depending on the global warming, and that the IceLand Low will most likely see a split into two centers of action in the extreme scenarios. The work was done with Dominik Vietinghoff, Christian Heine, and Michael Böttinger.

References

1 Vietinghoff, Heine, Böttinger, Maher, Jungclaus, Scheuermann. *Visual analysis of spatio-temporal trends in time-dependent ensemble data sets on the example of the north atlantic oscillation*. IEEE PacificVis 2021 Proceedings, 71-80, 2021
2 Vietinghoff, Heine, Böttinger, Scheuermann. *An Extension of Empirical Orthogonal Functions for the Analysis of Time-Dependent 2D Scalar Field Ensembles*. IEEE PacificVis 2021 Short Papers, 46-50, 2021.
3 Vietinghoff, Böttinger, Scheuermann, Heine. *Detecting Critical Points in 2D Scalar Field Ensembles Using Bayesian Inference*. IEEE PacificVis 2022 Proceedings, 1-10, 2022.

3.17 Uncertainty in Time Series and Geographic Data

Johanna Schmidt (VRVis – Wien, AT)

URL https://www.digi-hydro.com/

Design decisions must be made to make data visual, and modifications to the data are needed. Data modification includes reconstruction, resampling, filtering, and aggregation. In one of our projects, we have to deal with time series data recorded from sensors installed in hydropower machines. The project's purpose is to better understand which sensors can give information about the current state of the hydropower machine. This needs to be done with exploratory data analysis. It is not yet known to the mechanical engineers which sensors will be descriptive for detecting certain stages during machine operation. However, the data is large (approximately 30 TB of data), and it is impossible to analyze the raw data in this case. We, therefore, need to apply resampling and filtering to the data, which introduces uncertainty in the analysis the mechanical engineers should be informed about. In the case of geological data, reconstruction (of point cloud data) and 3D rendering introduce uncertainty in the data representation. When performing analyses, methods like plane fitting are also not wholly accurate. This uncertainty in the data and how it is presented to the users needs to be communicated, as both user groups (mechanical engineers and geologists) highly depend on detailed analysis results.

3.18 Quantifying and Visualizing Uncertainty in Medical Image Segmentation

Thomas Schultz (Universität Bonn, DE)

Joint work of Shekoufeh Gorgi Zadeh, Thomas Schultz
Main reference Shekoufeh Gorgi Zadeh, Maximilian W. M. Wintergerst, Thomas Schultz: "Intelligent interaction and uncertainty visualization for efficient drusen and retinal layer segmentation in Optical Coherence Tomography", Comput. Graph., Vol. 83, pp. 51–61, 2019.
URL http://dx.doi.org/10.1016/j.cag.2019.07.001

Neural networks have greatly increased the accuracy in many medical image segmentation tasks, and have been successfully deployed for large-scale image analysis. However, fully automated results are still not reliable enough to be trusted blindly in applications where segmentation quality might be critical to the well-being of individuals. Using an application example in ophthalmology, we demonstrate that visualizing the uncertainty in neural network based segmentations, and providing uncertainty-aware tools for segmentation editing, can make it more time efficient to identify and correct remaining segmentation errors. We also discuss the important open question of reliable uncertainty quantification in an out-of-distribution setting, for example when processing images that have been acquired with a different scanner, and we mention strategies for approaching that problem.

3.19 Visualizing the Uncertainty in Image Analysis – Previous work and new opportunities

Brian Summa (Tulane University – New Orleans, US)

In this talk, I give an overview of my research in the visualization of uncertainty in scientific data, while highlighting new opportunities for uncertainty quantification in topological data analysis (TDA) or in accounting for uncertainty due to human variability.

3.20 Uncertainty and Trustworthy AI

Stefan Hagen Weber (Siemens – München, DE)

Joint work of Daniela Oelke, Stefan H. Weber.
Main reference Daniela Oelke, Stefan H. Weber: "Line Density Plots – Visualizing uncertainty in forecast ensembles". Talk on IEEE VIS 2018 VisInPractice event.

Visual representations of density under uncertainty have been explored for geographic data, scatterplots, line charts or parallel coordinates. Ensemble forecasting is a widely known application for uncertainty visualization. Often end users have specific requirements and tasks for the visualization, e.g.

- each forecast (line) should be visible and interactable,
- the resulting chart should not be overcrowded
- the uncertain space between ensembles should be filled by upsampling
- the uncertainty should be made visible
- Identifying outliers is as important as spotting the main trend

All these requirements can be realized by a novel technique for generating density representations for line charts that is visually and computationally scalable with respect to the number of lines that are shown. In contrast to alternative kernel-based density representations, it also keeps the course of the lines visible unless the local density is very high. Points are on top of each other (or crossing lines) represent the (un)certainty (density surface) and are mapped to color. A smoothed representation with an upsampling effect is done by adding a "glow" around the lines. This glow is implemented by decreasing the alpha value with increasing distance from the line. The amount of glow at a certain distance is determined by the shape of the specific kernel function. The kernel width determines the extension of the glow around the line. Some considerable effort was spent to design and implement the visualization and integrate it into a commercial system (linking & brushing), to fulfill the end user's requirements. The result was evaluated together with the end user who wanted to gain more insight into their ensemble forecasts to answer the question "When is the best time to buy oil?". The end user first inspected the overall distribution pattern in time over all ensemble members. They used the median to separate the higher half of the distribution from the lower. They got immediate insights regarding the trend and distribution. The visualization was more explored, and the end user provided very positive feedback. Our expectation was of course that the result will be used from now on. However, the opposite happened. The end user so far was not sure if he can trust the ensemble forecast method. With our visualization he gained trust in the AI after a few hours. From that point

on he was fine with a simple KPI: "Just tell me when to buy oil". Lessons learned: A lot of effort was spent for a visualization that was only used a few times. You might argue that it was not worth the effort. However, it turned out that the initial task of understanding the uncertainty aspect was only the first step. The final effect was that the user increased his trust in the AI. Trustworthy AI is a valuable asset. It might be a frustrating experience but increasing trust in AI is a huge long-term benefit. Even if the visualization was only one short part of the journey. Showing uncertainly in a proper way can increase trust.

3.21 Overcoming Uncertainties in Molecular Visualization

Thomas Wischgoll (Wright State University – Dayton, US)

Joint work of Thomas Wischgoll, Christina Gillmann, Robin Maack, Matthew Marangoni
URL https://avida.cs.wright.edu

Uncertainties are difficult if not impossible to avoid. Capturing data from the analog world almost always results in some form of uncertainty. The amount of uncertainty depends on the method of measurement and its accuracy. When visualizing data that has some associated uncertainty, it is essential to properly process and convey such uncertainty and especially the amount of uncertainty keeping in mind that additional processing steps can amplify the uncertainty. There are various sources of uncertainty, such as numerical limitations or limitations of the capture device. However, there are other sources of uncertainty. Some of these uncertainties stem from model assumptions or limitations of how we translate natural specimens to 3D representations. Molecular structures are one example of this. This talk will illustrate this further and point to some of the solutions.

3.22 Uncertainty Visualization of Health Data

Liang Zhou (Peking University, CN)

Health science relies on a wide range of different types of data. There, uncertainty is ubiquitous and is aware by health science experts. Uncertainty visualization is, therefore, important and could potentially aid decision making. In this talk, I will introduce my own research work on new visualization techniques for representative health data. These examples focus on uncertainty visualization of ensemble medical imaging data, local correlation and subspace visualization for multidimensional data, and perceptual enhancement for visualization images. I will also discuss works on visual analytics of health data with uncertainties from missing data. Finally, I will discuss uncertainty challenges that I identified in the various types of health data.

4 Working groups

4.1 Applications

Tushar Athawale (Oak Ridge National Laboratory, US), Michael Böttinger (DKRZ Hamburg, DE), Amy Gilmer (USGS – Denver, US), Hans-Christian Hege (Zuse-Institute Berlin, DE), Christoph Heinzl (Universität Passau, DE), Christopher R. Johnson (University of Utah – Salt Lake City, US), Gerik Scheuermann (Universität Leipzig, DE), Johanna Schmidt (VRVis – Wien, AT), Thomas Schultz (Universität Bonn, DE), Jarke J. van Wijk (TU Eindhoven, NL), Stefan Hagen Weber (Siemens – München, DE), and Xiaoru Yuan (Peking University, CN)

Uncertainty visualization and assessment are used in many domains, including medical applications, non-destructive testing, industrial AI, geology, renewable energies, and climate research. The way uncertainty is used by users in these domains differs depending on the required tasks and the data used. As an outcome of this working group, we identified success stories of published or successfully applied in seven domains. Based on these success stories, we identified common open challenges and research questions that will be worth working on: (i) Uncertainty could be viewed from a mathematical point of view, looking at stochastic processes, statistics, correlations, and similar. This would also enable the quantification of uncertainty in different domains. (ii) The different sources of uncertainty need to be discussed – whether they are similar in different domains and to which degree they depend on tasks and data. Also, the terminology used in different domains to describe uncertainty differs. (iii) An interesting question is to differentiate between visualization applications where uncertainty visualization is needed and where not. It might depend on the task, and the types of questions users have, whether it makes sense to include uncertainty in a visual representation or not. (iv) Visualizing uncertainty also relates to perceptual issues, describing how well uncertainty can be perceived using different encodings. (v) As a wrap-up, it would be interesting to find out how far uncertainty visualization is already used in commercial software.

4.2 What's the point?: Focusing on the human in uncertainty vis

Nadia Boukhelifa (INRAE – Palaiseau, FR), Michael Correll (Tableau Software – Seattle, US), Stephanie Deitrick (Arizona State University – Tempe, US), Matthew Kay (Northwestern University – Evanston, US), Miriah Meyer (Linköping University, SE), Kristi Potter (NREL – Golden, US), Paul Rosen (University of Utah – Salt Lake City, US), and Regina Maria Veronika Schuster (Universität Wien, AT)

Current techniques around uncertainty visualization are often oriented around statistical models, with the efficacy of an uncertainty visualization viewed as either how accurately the viewer is able to retrieve or estimate specific model values, or how well the viewer's

decision-making aligns with that of some normative model of utility or decision quality. This model-driven rather than human-driven perspective introduces several key limitations when designing or evaluating uncertainty visualizations. For one, it elides many aspects of decision-making under uncertainty that are not amenable to tidy quantification, such as situated or implicit knowledge. For another, it ignores psychological, sociological, rhetorical, or ethical aspects of presenting uncertainty information. We propose a human-centered view of uncertainty visualization in which the needs and information of the viewer, rather than backing statistical or inferential models, are given precedence.

In the human-centered view of uncertainty visualization, viewers are neither rote reciters of p-values, nor conditioned to mimic the actions of a statistical test. Rather, they have many goals, including being able to audit or justify their decisions, build appropriate trust in the data source and designers, integrate their own mental models and domain knowledge with existing data, or even just walk away satisfied that they made a reasonable decision given the information they had to hand. In this paper, we show how existing frames around uncertainty visualization may fail to result in designs that accomplish these goals, and present both existing strategies for better integrating the human in the uncertainty visualization design process as well as open problems in visualization research.

4.3 A Problem Space for Designing (Uncertainty) Visualizations

Maria Riveiro (Jönköping University, SE), Remco Chang (Tufts University – Medford, US), Oliver Deussen (Universität Konstanz, DE), Christina Gillmann (Universität Leipzig, DE), Michael Gleicher (University of Wisconsin-Madison, US), and Tatiana von Landesberger (Universität Köln, DE)

Main reference Hans-Jorg Schulz, Thomas Nocke, Magnus Heitzler, and Heidrun Schumann. 2013. A Design Space of Visualization Tasks. IEEE Transactions on Visualization and Computer Graphics, 19(12):2366–2375.

Visualization researchers seek appropriate abstractions to help us design, analyze, organize, and evaluate the things we create. Information visualization literature has many task structures (taxonomies, typologies, etc.), design spaces, and related frameworks. In this working group, we discussed current frameworks for designing visualizations, and we considered developing a new problem space that complements the existing ones by focusing on the needs that a visualization is meant to solve. Briefly, the proposed problem space is based on the earlier work by [1], considering the 5Ws and H (who, why, what, where, when and how).

We believe that this problem space provides a valuable conceptual tool for designing and discussing visualizations, including uncertainty visualisations.

References

1 Hans-Jorg Schulz, Thomas Nocke, Magnus Heitzler, and Heidrun Schumann. 2013. A Design Space of Visualization Tasks. IEEE Transactions on Visualization and Computer Graphics, 19(12):2366–2375.

Participants

- Michael Böttinger
DKRZ Hamburg, DE
- Nadia Boukhelifa
INRAE – Palaiseau, FR
- Remco Chang
Tufts University – Medford, US
- Oliver Deussen
Universität Konstanz, DE
- Christina Gillmann
Universität Leipzig, DE
- Amy Gilmer
USGS – Denver, US
- Michael Gleicher
University of Wisconsin-Madison, US
- Hans-Christian Hege
Zuse-Institute Berlin, DE
- Christoph Heinzl
Universität Passau, DE
- Miriah Meyer
Linköping University, SE
- Kristi Potter
NREL – Golden, US
- Maria Riveiro
Jönköping University, SE
- Gerik Scheuermann
Universität Leipzig, DE
- Johanna Schmidt
VRVis – Wien, AT
- Thomas Schultz
Universität Bonn, DE
- Regina Maria Veronika Schuster
Universität Wien, AT
- Jarke J. van Wijk
TU Eindhoven, NL
- Tatiana von Landesberger
Universität Köln, DE
- Stefan Hagen Weber
Siemens – München, DE

Remote Participants

- Tushar Athawale
Oak Ridge National Laboratory, US
- Mehdi Chakhchoukh
INRAE Paris, FR
- Michael Correll
Tableau Software – Seattle, US
- Stephanie Deitrick
Arizona State University – Tempe, US
- Jake Hofman
Microsoft – New York, US
- Amit Jena
Indian Institute of Technology Bombay, IN
- Christopher R. Johnson
University of Utah – Salt Lake City, US
- Matthew Kay
Northwestern University – Evanston, US
- Robert Lempert
RAND – Santa Monica, US
- Paul Rosen
University of Utah – Salt Lake City, US
- Han-Wei Shen
Ohio State University – Columbus, US
- Brian Summa
Tulane University – New Orleans, US
- Kathleen Warrell
UCAR – Boulder, US
- Thomas Wischgoll
Wright State University – Dayton, US
- Xiaoru Yuan
Peking University, CN
- Liang Zhou
Peking University, CN

Report from Dagstuhl Seminar 22332

Differential Equations and Continuous-Time Deep Learning

David Duvenaud[*1], **Markus Heinonen**[*2], **Michael Tiemann**[*3], **and Max Welling**[*4]

1 **University of Toronto, CA.** duvenaud@cs.toronto.edu
2 **Aalto University, FI.** markus.o.heinonen@aalto.fi
3 **Robert Bosch GmbH – Renningen, DE.** michael.tiemann@de.bosch.com
4 **University of Amsterdam, NL.** welling.max@gmail.com

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 22332 "Differential Equations and Continuous-Time Deep Learning". Neural ordinary-differential equations and similar continuous model architectures have gained interest in recent years, due to the existence of a vast literature in calculus and numerical analysis. Thus, continuous models might lead to architectures with finer control over prior assumptions or theoretical understanding. In this seminar, we have sought to bring together researchers from traditionally disjoint areas – machine learning, numerical analysis, dynamical systems and their "consumers" – to try and develop a joint language about this novel modeling paradigm. Through talks & group discussions, we have identified common interests and we hope that this first seminar is but the first step on a joint journey.

Seminar August 15–19, 2022 – http://www.dagstuhl.de/22332
2012 ACM Subject Classification Computing methodologies → Machine learning; Computing methodologies → Philosophical/theoretical foundations of artificial intelligence; Mathematics of computing → Differential equations; Mathematics of computing → Solvers
Keywords and phrases deep learning, differential equations
Digital Object Identifier 10.4230/DagRep.12.8.20

1 Executive Summary

David Duvenaud
Markus Heinonen
Michael Tiemann
Max Welling

Deep models have revolutionised machine learning due to their remarkable ability to iteratively construct more and more refined representations of data over the layers. Perhaps unsurprisingly, very deep learning architectures have recently been shown to converge to differential equation models, which are ubiquitous in sciences, but so far overlooked in machine learning. This striking connection opens new avenues of theory and practice of continuous-time machine learning inspired by physical sciences. Simultaneously, neural networks have started to emerge as powerful alternatives to cumbersome mechanistic dynamical systems. Finally, deep learning models in conjecture with stochastic gradient optimisation has been used to numerically solve high-dimensional partial differential equations. Thus, we have entered a new era of continuous-time modelling in machine learning.

* Editor / Organizer

This change in perspective is currently gaining interest rapidly across domains and provides an excellent and topical opportunity to bring together experts in dynamical systems, computational science, machine learning and the relevant scientific domains to lay solid foundations of these efforts. On the other hand, as the scientific communities, events and outlets are significantly disjoint, it is key to organize an interdisciplinary event and establish novel communication channels to ensure the distribution of relevant knowledge.

Over the course of this Dagstuhl Seminar, we want to establish strong contacts, communication and collaboration of the different research communities. Let's have an exchange of each community's best practices, known pitfalls and tricks of the trade. We will try to identify the most important open questions and avenues forward to foster interdisciplinary research. To this end, this seminar will feature not only individual contributed talks, but also general discussions and "collaboration bazaars", for which participants will have the possibility to pitch ideas for break-out project sessions to each other. In the break-out sessions, participants may discuss open problems, joint research obstacles, or community building work.

2 Table of Contents

3 Overview of Talks

3.1 Differential Equations for Causal Inference in Complex Stochastic Biological Processes

Hananeh Aliee (Helmholtz Zentrum München, DE)

In my talk, I presented a sparsity-enforcing regularizer for continuous-time neural networks motivated by causality. Sparsification can help to identify the parameters of the differential equations and infer the causal interaction between variables. I also discussed an application of that in single-cell genomics for modeling gene dynamics and inferring gene regulatory networks using neural ODEs. Finally, I discussed some open problems and challenges in modeling complex stochastic biological processes and potential directions for future work.

3.2 Computation Theory for Continuous Time. Programming with Ordinary Differential Equations.

Olivier Bournez (Ecole Polytechnique – Palaiseau, FR)

In this talk, we will argue that computation theory for continuous time analog models did not develop at the level as the one for digital models. We will review some examples of such models, such as the General Purpose Analog Computer (GPAC) from Claude Shannon, proposed as a model of Differential Analyzers. We will show how this model can be programmed on several example. We will then discuss about how this model relates to classical models of computabiility such as Turing machines, both considering computability theory and complexity theory. We will show the close relation between this model and polynomial Ordinary Differential Equations (pODEs-. As a side effect of our constructions, we will see that one can program with pODEs and we will discuss applications.

3.3 Injecting Physics into Differential Equation based Deep Learning Models

Biswadip Dey (Siemens – Princeton, US)

This talk focused on demonstrating the usefulness of using a physics-informed inductive bias in differential equation based deep learning models and highlighted some open problems on this topic. We discussed Symplectic-ODENet and its extensions which encode energy conservation into the computation graph to improve model performance, efficiency, and interpretability. However, these models typically assumes that the systems states can be directly measured. This leads to the following open questions: (i) Can we learn a suitable latent representation from high-dimensional observations and then enforce physics (e.g., energy conservation) in the learned latent space? and (ii) Can we enforce physics even when only a subset of the system states can be directly measured?

3.4 Equivariant Deep Learning via PDEs

Remco Duits (TU Eindhoven, NL)

We consider PDE-based Group Convolutional Neural Networks (PDE-G-CNNs) that generalize Group equivariant Convolutional Neural Networks (G-CNNs). In PDE-G-CNNs a network layer is a set of PDE-solvers where geometrically meaningful PDE-coefficients become trainable weights. The underlying PDEs are morphological and linear scale space PDEs on the homogeneous space of positions and orientations to the roto-translation group SE(2). The PDEs provide a geometrical and probabilistic understanding of the network. The network is implemented by morphological convolutions with approximations to kernels solving nonlinear HJB-PDEs (for morphological α-scale spaces), and to linear convolutions solving linear PDEs (for linear α-scale spaces). In the morphological setting, the parameter α regulates soft max-pooling over Riemannian balls, whereas in the linear setting the cases $\alpha = 1/2$ and $\alpha = 1$ correspond to the Poisson and Gaussian semigroup. We prove that our practical analytic approximation kernels are accurate. In the morphological setting, we propose analytic approximations of (sub)-Riemannian balls on M(2) which carry the correct reflectional symmetries globally and we provide asymptotic error analysis. The analytic approximations allow for efficient, accurate training of fundamental neuro-geometrical association field models in the GPU-implementations of our PDE-G-CNNs. The equivariant PDE-G-CNN network implementation consists solely of linear and morphological convolutions with parameterized analytic kernels on M(d). Common mystifying nonlinearities in CNNs are now obsolete and excluded. We present blood vessel segmentation experiments in medical images that show clear benefits of PDE-G-CNNs compared to state-of-the-art G-CNNs: increase of performance along with a huge reduction in network parameters and training data.

3.5 Putting All of Modeling into Adaptive SPDE Solvers

David Duvenaud (University of Toronto, CA)

My talk presented a roadmap for building spatiotemporal models which can automatically introduce auxiliary variables. These auxiliary variables can be tuned jointly with the parameters of the model to find dynamics which are easy to integrate, either by encouraging approximate spatial factorization, or fast mixing temporally. I also introduced a scheme for stateless sampling from Brownian sheets.

3.6 Bayesian Calibration of Computer Models & Beyond

Maurizio Filippone (EURECOM – Biot, FR)

Bayesian calibration of computationally expensive computer models offers an established framework for quantification of uncertainty of model parameters and predictions. Traditional Bayesian calibration involves the emulation of the computer model and an additive model discrepancy term using Gaussian processes; inference is then carried out using Markov chain Monte Carlo. In this talk, I present a calibration framework where limited flexibility and scalability are addressed by means of compositions of Gaussian processes into Deep Gaussian processes and scalable variational inference techniques. This formulation can be easily implemented in development environments featuring automatic differentiation and exploiting GPU-type hardware. I then discuss identifiability issues and cases where the computer model implements ODEs/PDEs/SDEs. Finally, I draw connections with other inference frameworks, such as transfer learning, gradient matching for ODEs and SDEs, and Physics-informed priors for Bayesian deep learning.

3.7 High Order SDE Solvers in Machine Learning

James Foster (University of Oxford, GB)

From Markov Chain Monte Carlo to Neural SDEs and Score-based diffusions, there has been a recent uptick in the applications of SDEs in machine learning. However, SDEs have been studied by the mathematics community for decades and it has been well established that SDE solvers have fundamental limitations in their convergence rates. In this talk, we will review this theory and discuss how noise types influence convergence rates for SDEs solvers. This will naturally lead us to pose the following question:

"Can we construct SDEs that are easy to solve?"

By considering both kinetic Langevin and Score-based diffusions, two prominent examples of SDEs, we give a positive answer to this question and speculate that finding such "easy-to-solve" SDEs will be an area of opportunity in future research.

3.8 Interpretable Polynomial Neural Ordinary Differential Equations

Colby Fronk (University of California – Santa Barbara, US)

Neural networks have the ability to serve as universal function approximators, but they are not interpretable and don't generalize well outside of their training region. Both of these issues are problematic when trying to apply standard neural ordinary differential equations (neural ODEs) to dynamical systems. We introduce the polynomial neural ODE, which is a deep polynomial neural network inside of the neural ODE framework. We demonstrate the capability of polynomial neural ODEs to predict outside of the training region, as well as perform direct symbolic regression without additional tools such as SINDy.

3.9 Neural Differential Equations and Operator Learning

Jacob Seidman (University of Pennsylvania – Philadelphia, US)

My talk presented two categories of methods to learn maps between spaces of functions. The first is known as Neural PDEs/SDEs and parameterizes PDEs/SDEs to implicitly define operators through their solutions. The other category is typically known as Operator Learning and uses compositions of parameterized integral transformations, pointwise transformations, and function reconstructions from learned basis or nonlinear representations. I posed the question of which approach works better in different scenarios. This led to a discussion about the pros and cons of each approach in terms of properties such as expressivity, ability to encode prior information, and computational efficiency.

3.10 On Practical Inference and Learning in Dynamical Systems

Arno Solin (Aalto University, FI)

In general spatio-temporal systems, time takes a fundamentally different role from other (spatial) dimensions as observations can be ordered over time. This talk takes interest in challenges in online inference and learning problems, where the model admits the form of a stochastic differential equation (SDE) or stochastic partial differential equation (SPDE). These types of problems occur naturally in sensor fusion applications where the dynamics borrow from first principles but also include unknown (stochastic) effects. The talk presents open problems in designing principled approximate inference methods, with non-linear continuous-discrete inertial navigation as a practical example.

3.11 Partial Differential Equations and Deep Learning

Nils Thuerey (TU München, DE)

In my talk I focused on the combination of PDE for applications such as fluids and deep learning (DL). Despite success of integrating solvers as differentiable components in DL, many challenges for training remain. Interestingly, the regular gradient has some fundamental problems, as indicated by its mismatch in terms of units. I discussed potential avenues for alleviating these problems, such as using inverse solvers of partial inversions.

3.12 Dynamical Systems Cookbook (& their solvers, & their optimization)

Michael Tiemann (Robert Bosch GmbH – Renningen, DE)

In many areas of science and engineering, neural ordinary differential equations seem like natural candidates for extending limited first-principles models. However, retaining an interpretability in terms of preserved quantities of interest or system properties, such as volume invariances, preserved first integrals and other conservation laws, requires an algebra of models that represent a wide variety of dynamical systems, while guaranteeing the preservation of these quantities by construction. In this call for contributions, we hope to establish a grass-roots initiative that will contribute to cookbook of building blocks that represent a wide variety of potential applications, while working reliability "out-of-the-box" for the majority of modeling problems. Futhermore, this cookbook needs to consider not only the algebra of the ODE vector fields, but also that of the numerical discretizations and finally of their identification through means of optimization or other adaptation methods.

3.13 Graph-based Differential Equations, Continuum Limits, and Merriman-Bence-Osher schemes

Yves van Gennip (TU Delft, NL)

Ideas and methods from differential equations and variational methods on graphs can also play a role for neural networks (NN). In particular, we take a look at the Merriman–Bence–Osher (MBO) scheme and the family of semi-discrete implicit Euler (SDIE) schemes and see that they can be written as NN. We also discuss discrete-to-continuum limits at the variational and gradient flow levels and open questions.

4 Working groups

4.1 Brainstorm session

Yves van Gennip (TU Delft, NL), Olivier Bournez (Ecole Polytechnique – Palaiseau, FR), Joachim M. Buhmann (ETH Zürich, CH), Remco Duits (TU Eindhoven, NL), Sho Sonoda (RIKEN – Tokyo, JP), and Max Welling (University of Amsterdam, NL)

1. local-nonlocal interactions

We asked the question if the PDE models with local derivatives can be generalized to more general non-local (integral) operators. We believe this is possible and would lead to genuinely new models that would be better in modeling problems with highly nonlocal interactions.

2. multi-scale/renormalisation group
We asked if there is merit to introduce a scale-space into the representations. For instance, every layer can represent a full scale space, or the progression through the layers represents a coarse graining transformation. The former can be viewed as a special case of scale equivariance (a semi-group!), while the latter is more like a renormalization group transformation.

3. equivariance/symmetries*/local equivariance (*in continuum formulation and after discretisation)
We discussed if there are extensions to equivariance to non-group transformations (e.g. semi-groups see above). Also, if we formulate the NN as a PDE in the continuum limit, we can model symmetries also as a transformation with a generator that commutes with the Hamiltonian. Can we think of equivariance as simply finding a homomorphism in the hidden layers that forms a commuting diagram with the transformations in the input layer: transformation in input layer –> embedding to hidden layer = embedding to hidden layer –> transformation in hidden layer. We also discussed the role of local versus global symmetries: to what extent does a global equivariance also enforce local equivariance. Can this be formalized? Can this be generalized to diffeomorphisms?

4. quantum extensions (learning unitary operators in quantum computers)
The Schrodinger equation is also a PDE. We can extend the continuum PDE limit of a linear layer to a quantum layer by evolving an input quantum state using the SE. This maps to a model for optical quantum hardware. Is this beneficial, or more powerful for ML? Can we include symmetries?

5. conserved quantities/Noether's theorem (see also 3)
We discussed what could be at the basis of a general theory and thought the notion of conserved quantities to be a good candidate. We discussed how to apply Hamiltonian reduction by stages (Marsden) from classical mechanics to deep learning.

6. (geometrical flow?) interpretations of full networks (example: mean curvature flow)
We saw one particular example of a CNN that appears to be interpretable in terms of mean curvature flow. This raises a more general question regarding the possibility of interpreting NNs (not per layer, but as a whole) in terms of geometrical flows.

PDE-G-CNNs provide geometric interpretation of flows in the neural networks. PDE-G-CNNs, in contrast to CNNs, do not include ad-hoc nonlinearities, but only solutions to linear and nonlinear PDEs, both solved by equivariant convolutions over different semirings. The merging of association fields as visible in feature maps of PDE-G-CNNs requires algebraic geometry (Betti numbers). We discussed Lie group extensions of recent works of Creemers, but realized it is quite challenging.

7. new operators/integral operators
We discussed if we also want to consider nonlocal operators (besides classical "local" PDEs), such as non-local derivatives and fractional powers of semi-group evolutions, in the network layers we consider, or if nonlocalities are only allowed to appear as a result of interactions between many layers.

8. why is deep better than wide? (linear vs polynomial scaling of "influence" of neurons?)
Why do deep NNs perform better than wide ones? The initial thought is that interactions between layers scale nonlinearly in the number of neurons, whereas interactions within a layer scale linearly.

9. how to design nonlinearties?
We prefer HJB-PDEs that allow for morphological convolutions, not by (ad-hoc) ReLU's that are a non-optimal special case.

10. continuum limits
Techniques exist in the mathematical literature to find continuum limits of (loss) functionals, such as Gamma-convergence, and limits of gradient flows derived from those functionals. Can we employ those to prove relevant continuum limits for neural networks?

11. wishlist for PDEs (equivariance; semi-group structure; homogeneity in metric tensor per "unit")
We wish for an axiomatic approach to PDE-based equivariant deep learning with Lie-group domains and semi-ring co-domains.

We started investigations on that after the seminar and will continue to work on this thoroughly in the coming year.

12. PDE GCNNs on graphs.
We noted that equivariant networks on SE(3) in general require sparsification to become practical in view of memory management. The PDEs can enter by providing appropriate kernels for equivariant graph neural networks.

Participants

- Hananeh Aliee
Helmholtz Zentrum München, DE
- Jesse Bettencourt
University of Toronto, CA
- Olivier Bournez
Ecole Polytechnique – Palaiseau, FR
- Joachim M. Buhmann
ETH Zürich, CH
- Johanne Cohen
University Paris-Saclay – Orsay, FR
- Biswadip Dey
Siemens – Princeton, US
- Remco Duits
TU Eindhoven, NL
- David Duvenaud
University of Toronto, CA
- Maurizio Filippone
EURECOM – Biot, FR
- James Foster
University of Oxford, GB
- Colby Fronk
University of California – Santa Barbara, US
- Jan Hasenauer
Universität Bonn, DE
- Markus Heinonen
Aalto University, FI
- Patrick Kidger
Google X – Bay Area, US
- Diederik P. Kingma
Google – Mountain View, US
- Linda Petzold
University of California – Santa Barbara, US
- Jack Richter-Powell
University of Toronto, CA
- Lars Ruthotto
Emory University – Atlanta, US
- Jacob Seidman
University of Pennsylvania – Philadelphia, US
- Arno Solin
Aalto University, FI
- Sho Sonoda
RIKEN – Tokyo, JP
- Nils Thuerey
TU München, DE
- Michael Tiemann
Robert Bosch GmbH – Renningen, DE
- Filip Tronarp
Universität Tübingen, DE
- Yves van Gennip
TU Delft, NL
- Max Welling
University of Amsterdam, NL
- Verena Wolf
Universität des Saarlandes – Saarbrücken, DE
- Daniel Worrall
DeepMind – London, GB

Report from Dagstuhl Seminar 22341

Power and Energy-Aware Computing on Heterogeneous Systems (PEACHES)

Kerstin I. Eder[*1], Timo Hönig[*2], Daniel Mosse[*3], Max Plauth[*4], and Maja Hanne Kirkeby[†5]

1 **University of Bristol, GB.** cskie@bristol.ac.uk
2 **Ruhr-Universität Bochum, DE.** timo.hoenig@rub.de
3 **University of Pittsburgh, US.** mosse@cs.pitt.edu
4 **Hasso-Plattner-Institut, Universität Potsdam, DE.** max.plauth@hpi.de
5 **Roskilde University, DK.** majaht@ruc.dk

Abstract

This report documents the program and outcomes of the Dagstuhl Seminar 22341 – Power and Energy-Aware Computing on Heterogeneous Systems (PEACHES). The seminar was held on Aug 21 – Aug 26, 2022, and brought together 35 international experts from different domains across the entire system stack – from system designers to programmers and operators. We present the abstracts of 18 talks and 5 summaries of discussions and active sessions on the principal topic areas: Energy Transparency from Hardware to Software, Energy Optimisation and Management, Computing for Sustainability, Green Computing Hackathon, and Disruptive Paradigms.

Seminar August 21–26, 2022 – http://www.dagstuhl.de/22341

2012 ACM Subject Classification Computer systems organization → Heterogeneous (hybrid) systems; General and reference → Evaluation; General and reference → Design; General and reference → Measurement; General and reference → Metrics; Hardware → Power and energy; Software and its engineering → Operating systems

Keywords and phrases energy, heterogeneous computing, operating systems, power, systems

Digital Object Identifier 10.4230/DagRep.12.8.31

1 Executive Summary

Kerstin I. Eder (University of Bristol, GB)
Timo Hönig (Ruhr-Universität Bochum, DE)
Daniel Mosse (University of Pittsburgh, US)
Max Plauth (Hasso-Plattner-Institut, Universität Potsdam, DE)

More than ever, emissions, carbon footprint, and other related environmental concerns are at the forefront of society, from several different perspectives. There is an urgent need to understand how **computing** fits into the broader picture of our planet's energy consumption and what is the role of computing in reducing our carbon footprint worldwide. This requires new ways of thinking across different domains, and necessitates highly energy-efficient hardware and software designs that adapt to changing operating conditions to become more efficient. Collaboration is increasingly required across the entire system stack – from system designers to programmers and operators.

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

Power and Energy-Aware Computing on Heterogeneous Systems (PEACHES), *Dagstuhl Reports*, Vol. 12, Issue 8, pp. 31–59
Editors: Kerstin I. Eder, Timo Hönig, Daniel Mosse, Max Plauth, and Maja Hanne Kirkeby
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The Dagstuhl Seminar 22341 on Power and Energy-Aware Computing on Heterogeneous Systems (PEACHES) brought together experts from computer science and computer engineering that share a common vision towards reducing carbon emissions both using innovative designs for computing systems and techniques that bridge the gap between hardware and software, as well as using computing systems to manage other environment-influencing systems. Five principal topic areas were discussed in working groups during the meeting: Energy transparency from hardware to software, Energy optimisation and management, Sustainability in computing, "Green Computing" hackathons, and Disruptive paradigms.

This report documents the program and the outcomes of PEACHES.

2 Table of Contents

3 Overview of Talks

3.1 Adaptive Optimization of (some) Parallel Applications

Antonio Carlos Schneider Beck Filho (Federal University of Rio Grande do Sul, BR)

Joint work of Arthur F. Lorenzon, Charles C. de Oliveira, Jeckson D. Souza, Antonio Carlos S. Beck
Main reference Arthur Francisco Lorenzon, Charles Cardoso De Oliveira, Jeckson Dellagostin Souza, Antonio Carlos Schneider Beck: "Aurora: Seamless Optimization of OpenMP Applications", IEEE Trans. Parallel Distributed Syst., Vol. 30(5), pp. 1007–1021, 2019.
URL https://doi.org/10.1109/TPDS.2018.2872992

Efficiently exploiting thread-level parallelism has been challenging for software developers. As many parallel applications do not scale as the number of cores increases, the task of rightly choosing the ideal configuration (number of threads and/or DVFS level and/or thread/page mapping) to produce the best results in terms of performance and/or energy is not straightforward. In this talk, I show a solution that is transparent (does not demand changes in the original code) and is adaptive (automatically adjusts to applications at runtime). However, to achieve such levels of adaptability and transparency, our optimization is limited to some applications only: those implemented with OpenMP. This is the price to pay when it comes to adaptability and energy consumption.

3.2 Energy Automation: What do people want from?

Ruzanna Chitchyan (University of Bristol, GB)

Joint work of Jan Marc Schwidtal, Proadpran Piccini, Matteo Troncia, Ruzanna Chitchyan, Mehdi Montakhabi, Christina Francis, Anna Gorbatcheva, Timothy Capper, Mustafa A. Mustafa, Merlinda Andoni, Valentin Robu, Mohamed Bahloul, Ian Scott, Tanaka Mbavarira, Juan Manuel Espana, Lynne Kiesling
Main reference Jan Marc Schwidtal, Proadpran Piccini, Matteo Troncia, Ruzanna Chitchyan, Mehdi Montakhabi, Christina Francis, Anna Gorbatcheva, Timothy Capper, Mustafa A. Mustafa, Merlinda Andoni, Valentin Robu, Mohamed Bahloul, Ian Scott, Tanaka Mbavarira, Juan Manuel Espana, Lynne Kiesling: "Emerging Business Models in Local Energy Markets: A Systematic Review of Peer-To-Peer, Community Self-Consumption, and Transactive Energy Models" SSRN, 8 Mar 2022.
URL https://doi.org/10.2139/ssrn.4032760

The different Peer-To-Peer, Community Self-Consumption, and Transactive Energy Models gives rise to new configurations of business models for local energy trading among a variety of actors. Pragmatically, as software engineers, we must view social, technical, and environmental concerns as closely interrelated. Neither of these dimensions can be ignored in the software project and product. It is a challenge to develop software tools, methods, and applications that remedy the environmental impact of human activity while improving or maintaining the social and economic standing of the system stakeholders. Inevitably, this leads to a socio-technical systems engineering approach, where focus on the human and technical elements are equally important.

(Edited by: Maja Hanne Kirkeby, Collector).

3.3 Towards a hybrid (static and statistical) worst-case execution time and worst case energy consumption estimation

Liliana Cucu-Grosjean (INRIA – Paris, FR)

Joint work of Liliana Cucu-Grosjean, Marwan Wehaiba el Khazen, Hadrien Clarke, Kevin Zagalo, Adriana Gogonel, Yves Sorel, Avner Bar Hen, Yasmina Abdeddaim, Slim Ben Amor, Kossivi Kougblenou, Rihab Bennour
Main reference Marwan Wehaiba el Khazen, Kevin Zagalo, Hadrien Clarke, Mehdi Mezouak, Yasmina Abdeddaïm, Avner Bar-Hen, Slim Ben-Amor, Rihab Bennour, Adriana Gogonel, Kossivi Kougblenou, Yves Sorel, Liliana Cucu-Grosjean: "Work in Progress: KDBench – towards open source benchmarks for measurement-based multicore WCET estimators", in Proc. of the 28th IEEE Real-Time and Embedded Technology and Applications Symposium, RTAS 2022, Milano, Italy, May 4-6, 2022, pp. 309–312, IEEE, 2022.
URL https://doi.org/10.1109/RTAS54340.2022.00035

Since the work of Edgar and Burns in 2000s [1], the real-time community has showed increased interest in using statistical estimator for the problem of the worst-case execution time estimation (WCET) of programs on embedded processors. We start this talk by building a common vocabulary on the real-time notions like deadline, release and worst-case execution time while scheduling algorithms are underlined by priority-based. The presented figures are obtained from measurements of an open data benchmark calls PX4-RT [2], while the data are collected either real flights or Gazebo-based simulation. We remind that for the sake of the reproducibility such information is important and their lack is one identified drawback of existing work on statistical methods for the WCET estimation. We then present the two main classes of WCET estimators: static analysis based and measurement-based and we present a statistical WCET definition to allow hybridizing these two classes, while the common understanding of WCET as a bound stays coherent. We present a hybrid WCET estimator mixing statistical and static analyses. Results indicate that this facilitates the identification of relevant paths as well as the construction of a WCET bound for complex systems. This is explained by the fact that the probabilistic description of the WCET bound is an excellent basis for time composabillity.

We present the results of a commercial tool, RocqStat, implementing the presented hybrid estimator and conclude by providing our current stream of work of using hybrid estimators for the identification of relevant paths for the problem of worst-case energy consumption.

References

1 Robert I. Davis and Liliana Cucu-Grosjean. *A Survey of Probabilistic Timing Analysis Techniques for Real-Time Systems*. Leibniz Trans. Embed. Syst., 6(1):03:1–03:60, 2019
2 Marwan Wehaiba el Khazen et al. *Work in Progress: KDBench – towards open source benchmarks for measurement-based multicore WCET estimators*. IEEE Real-time systems and applications symposium, 2022

3.4 Performance Isolation for Power-Limited CPUs

Mathias Gottschlag (KIT – Karlsruher Institut für Technologie, DE)

Joint work of Mathias Gottschlag, Philipp Machauer, Yussuf Khalil, Frank Bellosa
Main reference Mathias Gottschlag, Philipp Machauer, Yussuf Khalil, Frank Bellosa: "Fair Scheduling for AVX2 and AVX-512 Workloads", in Proc. of the 2021 USENIX Annual Technical Conference, USENIX ATC 2021, July 14-16, 2021, pp. 745–758, USENIX Association, 2021.
URL https://www.usenix.org/conference/atc21/presentation/gottschlag

Since the breakdown of Dennard scaling, modern CPUs have become power limited to the point where they have to reduce their frequency when executing power-intensive code. This behavior poses a problem for performance isolation: When one task executes power-intensive instructions such as AVX2 or AVX-512, the resulting frequency reduction often affects other less power-intensive tasks.

We propose modifying the CPU accounting of existing fair schedulers to counteract this performance impact. We allocate more CPU time to low-power tasks according to the frequency reduction during execution of these tasks. While the resulting scheduling policy greatly improves performance isolation for many workloads, some workloads present a challenge as the CPU does not provide sufficient information on the power characteristics of individual tasks.

3.5 Security costs Energy – because we're doing it wrong!

Daniel Gruss (TU Graz, AT)

Joint work of Jonas Juffinger, Lukas Lamster, Andreas Kogler, Maria Eichlseder, Moritz Lipp, Daniel Gruss
Main reference J. Juffinger, L. Lamster, A. Kogler, M. Eichlseder, M. Lipp, D. Gruss: "CSI:Rowhammer – Cryptographic Security and Integrity against Rowhammer", in Proc. of the 2023 2023 IEEE Symposium on Security and Privacy (SP) (SP), pp. 236–252, IEEE Computer Society, 2023.
URL https://doi.org/10.1109/SP46215.2023.00014

As we are running into a global energy crisis, saving energy in ICT is more important than ever. However, today we just patch security on top of system designs – an approach that inherently introduces performance and energy overheads. The Meltdown patch alone caused performance overheads of roughly 5%, meaning an overhead on greenhouse-gas emissions of up to 0.09% in 2018, a similar single patch in 2030, would cause an overhead of up to 0.4%. This is unsustainable and forces us to rethink security and question how we design systems. In the talk, we argue that the curve we use to optimize systems goes from reliable to unreliable to completely unusable where the system freezes and crashes so frequently that it is not useful for any task. We then argue that we need to rethink how we approach these problems and introduce cryptography-grade error detection combined with correction mechanisms to adjust this curve to make it continuous such that optimizing *too far* leads only to a loss in performance but never to an uncorrected system error or silent data corruption. If we achieve this, we can optimize for the sweet spot of system efficiency, that has been far out of our reach so far, and while increasing the security of the system.

3.6 Turning the knobs – Automatically determine energy-efficient process configurations and clock frequencies on Linux

Benedict Herzog (Ruhr-Universität Bochum, DE), Sven Köhler (Hasso-Plattner-Institut, Universität Potsdam, DE)

Operating systems offer numerous configuration parameters to tune the system behavior in general and the energy efficiency in particular. One keystone for an energy-efficient system is the right configuration tailored to the currently running application. Finding the right configuration, however, proves a challenging task. We independently pursued and developed two different approaches, Polar and memutil, to automatically find such an energy-efficient configuration.

Polar is based on a neural network receiving the application profile as input and provides an energy-efficient configuration as output. With this blackbox approach we found that the average energy efficiency (in terms of the energy delay-squared) can be improved by 11.5% for typical applications.

In an independent work, we implemented memutil – a Linux CPU frequency governor – automatically adapting each core's clock frequency based on live readings from performance measurement units. The most promising heuristic we found to cut idling cycles on high clock frequencies was the number of L2 cache misses. Using a linear frequency interpolation, memutil reduces reduce the energy demand of all tested benchmarks compared to the default governor at minimal execution time penalty of all tested benchmarks from the NPB suite compared to the default.

Although started with different assumptions and application scenarios Polar and memutil show comparable insights and foster future investigation and collaboration.

3.7 Energy-awareness Amplifies Side Channels

Henry Hoffmann (University of Chicago, US)

Joint work of Tejas Kannan, Henry Hoffmann
Main reference Tejas Kannan, Henry Hoffmann: "Protecting adaptive sampling from information leakage on low-power sensors", in Proc. of the ASPLOS '22: 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Lausanne, Switzerland, 28 February 2022 – 4 March 2022, pp. 240–254, ACM, 2022.
URL https://doi.org/10.1145/3503222.3507775

This talk examines recent work on energy-awareness in sensing systems. This includes both adaptive sampling and adaptive neural networks. Both techniques save energy by adapting execution to inputs or sequences of inputs. Adapting sampling reduces energy by intelligently determining when to take a sample, saving energy through reduced usage of both sensor and radio. Adaptive neural networks save energy by exiting the network early when additional computation is unlikely to affect accuracy.

Unfortunately, both approaches leak information about the inputs (i.e., to either the sensor or the neural network). This talk is meant to spur discussions about these privacy/energy tradeoffs and discuss potential solutions to the problem.

3.8 Interoperability of Energy-aware Systems

Henry Hoffmann (University of Chicago, US)

Main reference Henry Hoffmann: "JouleGuard: energy guarantees for approximate applications", in Proc. of the 25th Symposium on Operating Systems Principles, SOSP 2015, Monterey, CA, USA, October 4-7, 2015, pp. 198–214, ACM, 2015.
URL https://doi.org/10.1145/2815400.2815403

Energy-awareness has become an important topic and has been addressed by researchers working at various levels of the system stack. Energy-aware solutions adjust parameters at their level of the stack to meet energy constraints or optimize energy efficiency. It has recently been observed that solutions working at different levels can interfere with each other, reducing both performance and energy efficiency. One solution for this interference is preventing individual solutions and instead having a central energy-aware authority that manages the options available at all levels simultaneously.

This talk highlights the benefits of coordinating energy-aware solutions across the system stack. It also points out drawbacks of existing, centralized approaches (e.g., [2, 1, 3, 4]). It then highlights several challenges to be addressed to achieve a modular, interoperative, and composable energy-aware system stack.

References

1 H. Hoffmann, "JouleGuard," in Proceedings of the 25th Symposium on Operating Systems Principles, Oct. 2015, pp. 198–214, doi: 10.1145/2815400.2815403.
2 M. Maggio, A. V. Papadopoulos, A. Filieri, and H. Hoffmann, "Automated control of multiple software goals using multiple actuators," in Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, Aug. 2017, pp. 373–384, doi: 10.1145/3106237.3106247.
3 C. Wan, M. H. Santriaji, E. Rogers, H. Hoffmann, M. Maire, and S. Lu, "ALERT: Accurate Learning for Energy and Timeliness," in 2020 USENIX Annual Technical Conference, USENIX ATC 2020, July 15-17, 2020, 2020, pp. 353–369, [Online]. Available: https://www.usenix.org/conference/atc20/presentation/wan.
4 A. Filieri, H. Hoffmann, and M. Maggio, "Automated multi-objective control for self-adaptive software design," in Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering, Aug. 2015, pp. 13–24, doi: 10.1145/2786805.2786833.

3.9 Secure Energy-Aware Operating Systems

Henriette Hofmeier (Ruhr-Universität Bochum, DE), Benedict Herzog (Ruhr-Universität Bochum, DE), and Timo Hönig (Ruhr-Universität Bochum, DE)

Software mitigations of hardware vulnerabilities come with costs in terms of runtime and energy. Especially, when an application heavily interacts with the operating system, the mitigation-induced costs can exceed 25% for the Meltdown and Spectre vulnerabilities. Similar results can be observed when disabling SMT/Hyperthreading to mitigate, for example, Microarchitectural Data Sampling (MDS) cross-HyperThread attacks. Hence, it is worth to analyse on the one side whether an application requires the full protection by the mitigations and if not disable them dynamically. On the other side the appropriate mitigation can be

selected depending on the current hardware platform and system state for which we propose an approach. So questions that remain are:

- What are the potentials for such an approach?
- How to determine the protection requirements of an application?
- Who's in control?
- Is it beneficial to move the configuration from hardware into the operating system?

3.10 The (in)efficiency of the Internet

Romain Jacob (ETH Zürich, CH)

Joint work of Romain Jacob, Laurent Vanbever
Main reference Romain Jacob, Laurent Vanbever: "The Internet of tomorrow must sleep more and grow old". In HotCarbon, 2022.
URL https://hotcarbon.org/pdf/hotcarbon22-jacob.pdf

Today, the ICT industry has a massive carbon footprint (a few percent of the worldwide emissions) and one of the fastest growth rates. The Internet accounts for a large part of that footprint while being also energy inefficient; i.e., the total energy cost per byte transmitted is very high. Thankfully, there are many ways to improve on the current status; we discuss two relatively unexplored directions in this paper. Putting network devices to "sleep," i.e., turning them off, is known to be an efficient vector to save energy; we argue that harvesting this potential requires new routing protocols, better suited to devices switching on/off often, and revising the corresponding hardware/software co-design. Moreover, we can reduce the embodied carbon footprint by using networking hardware longer, and we argue that this could even be beneficial for reliability! We sketch our first ideas in these directions and outline practical challenges that we (as a community) need to address to make the Internet more sustainable.

3.11 Embodied Carbon, ICT's dirty little secret

Alex Jones (University of Pittsburgh, US)

First steps to address the challenge of sustainable computing are naturally to consider the energy efficiency of information and communication technologies (ICT) during their use phase (i.e., after they are deployed into service). This includes reducing energy consumption in processors, memory systems, peripheral devices, cooling systems and a host of other components that are used in deployed systems. However, for computing to be truly sustainable, all phases of the system life-cycle, from manufacturing to disposal, must be considered. In particular, there is limited awareness to the considerable fraction of the total life-cycle from "embodied" energy and carbon impacts of computing systems from the fabrication of the integrated circuits (ICs) that are used in those devices. These impacts can be predicted from "life-cycle assessment" techniques. Using tools like GreenChip it is possible to holistically evaluate energy consumption and other environmental impacts from computing. Using these tools, I show examples of how machine learning applications and in-memory compression can impact design choices for systems. From this we can find disruptive new ways to think about sustainability and even conservation when designing next generation ICT.

3.12 Task scheduling: What should we aim for in terms of reducing energy consumption?

Julia Lawall (INRIA – Paris, FR)

Joint work of Julia Lawall, Himadri Chhaya-Shailesh, Jean-Pierre Lozi, Baptiste Lepers, Willy Zwaenepoel, Gilles Muller
Main reference Julia Lawall, Himadri Chhaya-Shailesh, Jean-Pierre Lozi, Baptiste Lepers, Willy Zwaenepoel, Gilles Muller: "OS scheduling with nest: keeping tasks close together on warm cores", in Proc. of the EuroSys '22: Seventeenth European Conference on Computer Systems, Rennes, France, April 5 – 8, 2022, pp. 368–383, ACM, 2022.
URL https://doi.org/10.1145/3492321.3519585

The task scheduler decides when each task will run, and on which core, and thus can impact the machine energy consumption. In the context of large Intel servers (the case of a 2-socket Intel 5218 was illustrated in the talk), we first observe that making the machine fully idle saves more energy than leaving one thread running somewhere on the machine, due to the ensuing uncore costs. Thus, running an application faster, to finish sooner, may reduce energy consumption. We thus present the Nest scheduler, published at EuroSys 2022 [1], that concentrates tasks on a minimal number of cores. Nest make these cores show higher utilization and thus achieve higher core frequencies, causing the application to finish sooner. Finally, we next study schedutil, Linux's new power governor for reducing energy consumption, and illustrate how, on tasks that frequently sleep for short periods of time due to synchronization, it greatly increases execution time without saving energy. We then wonder how to move forward. Should the operating system impose strategies that impact energy consumption on the hardware, which does not seem to be very successful in the schedutil case, or should it try to work with hardware features, as in Nest?

References

1 Julia Lawall, Himadri Chhaya-Shailesh, Jean-Pierre Lozi, Baptiste Lepers, Willy Zwaenepoel, and Gilles Muller. OS scheduling with Nest: keeping tasks close together on warm cores. In *EuroSys*, pages 368–383. ACM, 2022.

3.13 Minimizing the energy consumption of Federated Learning on heterogeneous devices

Laércio Lima Pilla (University of Bordeaux, FR)

Main reference Laércio Lima Pilla: "Optimal Task Assignment for Heterogeneous Federated Learning Devices", in Proc. of the 35th IEEE International Parallel and Distributed Processing Symposium, IPDPS 2021, Portland, OR, USA, May 17-21, 2021, pp. 661–670, IEEE, 2021.
URL https://doi.org/10.1109/IPDPS49936.2021.00074

Federated Learning is a distributed machine learning technique focused on data privacy and security. In a nutshell, Federated Learning involves a group of heterogeneous devices working together to iteratively train a machine learning model under the coordination of a central server. The server chooses a subset of devices for training and sends them the model's weights for the round. These devices train the model with their own data (which is never shared) and send the updated weights to the server, which averages them before the next training round. The energy consumption of Federated Learning devices is a subject of interest

both for environmental reasons and due to the limited energy available on battery-powered devices. In this context, this talk discussed some of the efforts behind performance and energy optimizations on Federated Learning models, including a new idea for an optimal scheduling algorithm for minimizing the energy consumption of Federated Learning devices when controlling how much data each device should use for training.

3.14 Power-Aware Computing at Scale

Frank Mueller (North Carolina State University – Raleigh, US)

This talk focuses on power and energy control for cloud and high-performance computing facilities. It promotes hierarchical control systems dynamically adapting to application characteristics in a multi-tenant environment. Controls need to coordinate constraints at application/job level as well as center level to balance multiple objectives while exploiting heterogeneous memory and compute resources to the best of their ability. This leads to a number of challenges with open problems regarding trade-offs between objectives such as energy, performance, QoS, sustainability and profitability subject to future work.

References

1 *Uncore Power Scavenger: A Runtime for Uncore Power Conservation on HPC Systems* by Neha Gholkar, Frank Mueller, Barry Rountree, in Supercomputing (SC), Nov 2019.
2 *PShifter: Feedback-based Dynamic Power Shifting within HPC Jobs for Performance* by Neha Gholkar, Frank Mueller, Barry Rountree in High-Performance Parallel and Distributed Computing (HPDC), Jun 2018, pages 106-117.
3 *Power Tuning HPC Jobs on Power-Constrained Systems* by Neha Gholkar, Frank Mueller, Barry Rountree in International Conference on Parallel Architecture and Compilation Techniques (PACT), Sep 2016.
4 *PEARS: A Performance-Aware Static and Dynamic Framework for Heterogeneous Memory* by Onkar Patil, Latchesar Ionkov, Jason Lee, Frank Mueller, Michael Lang, in International Symposium on Memory Systems (MEMSYS), Oct 2021.
5 *NVM-based energy and cost efficient HPC clusters* by Onkar Patil, Latchesar Ionkov, Jason Lee, Frank Mueller, Michael Lang, in International Symposium on Memory Systems (MEMSYS), Oct 2021.
6 *Performance characterization of a DRAM-NVM hybrid memory architecture for HPC applications using Intel Optane DC Persistent Memory Modules* by Onkar Patil, Latchesar Ionkov, Jason Lee, Frank Mueller, Michael Lang, in International Symposium on Memory Systems (MEMSYS), Sep/Oct 2019.

3.15 Power Resilient NextG Data Centers

Simon Peter (University of Washington – Seattle, US)

This talk proposes to build and evaluate a power control plane (PCP) for NextG data centers that operate under tight and variable power envelopes. PCP can control power demand at a fine granularity and over short timescales by making it software-defined. The key is

to gracefully trade off power and quality of service over time. This allows PCP to shed or consolidate load to less power-intensive processors to conserve power during a power event. I show a prototype power resilient distributed file system (DFS) that establishes the viability of the idea. The DFS provides low-latency fail-over and recovery for load shedding events enabling fine-grained load control by PCP. I outline the important research questions that must be answered for PCP to become practical.

3.16 On energy awareness in NVRAM-based operating systems – NEON and PAVE

Wolfgang Schröder-Preikschat (Universität Erlangen-Nürnberg, DE), Timo Hönig (Ruhr-Universität Bochum, DE)

Joint work of Wolfgang Schröder-Preikschat, Timo Hönig, Jörg Nolte

This abstract deals with the recently launched projects NEON (non-volatility in energy-aware operating systems) and PAVE (power-fail aware byte-addressable virtual non-volatile memory), both of which have byte-addressable non-volatile main memory (NVRAM) at their core. Motivation is, on the NEON side, an energy-aware operation of a computing system by using modern NVRAM technology for the operating system itself, namely (1) to save power and (2) to survive power failures, and, on the PAVE side, an operating-system abstraction that should pave the way for a scalable and fail-safe execution of programs (esp. legacy software) virtually directly in NVRAM.

An advantage of the emerging NVRAM technology is the ability to eliminate the need for conventional persistence measures within an operating system and the thus resulting reduction in its space, time and energy requirements as well as a smaller attack surface. Further benefits of NVRAM are its higher speed compared to conventional storage, its rather high capacity compared to conventional DRAM and its ability to keep its state persistently without energy costs.

Unfortunately, today's CPU cannot (yet) do without volatile memory when registers and caches are considered. Furthermore, NVRAM is much slower than conventional main memory technology such as DRAM and changing the persistent state in NVRAM by writing results in more power consumption than corresponding state changes in DRAM [2]. In addition, fail-safe guarantees are now required from the system, since power failures when writing to the NVRAM, for example, can lead to control flows that unexpectedly convert a sequential process into a non-sequential one [3]. These problems can be solved by an operating system by (1) a suitable event-based, sporadically triggered checkpointing mechanism integrated in its exception- handling subsystem and (2) a suitable integration of NVRAM into the memory hierarchy via its virtual-memory subsystem.

The idea is that a trap or power failure interrupt (PFI) results in a micro-checkpoint request that is handled with strict time guarantee in the operating system: This checkpoint event basically preempts the running process from its volatile environment into the NVRAM. The specified *residual energy window* as a characteristic feature of the power-supply unit (PSU) determines the upper time limit for this procedure [1], the *worst-case execution time* (WCET) of which must never exceed it. In program areas where this mechanism cannot be used, particularly for the backup procedure itself, transactional programming comes into play.

A main problem are *non-maskable interrupt* (NMI) nesting and critical sections that block an *interrupt request* (IRQ): both endanger the timely handling of a possible checkpoint request. Respective sections are localised by static program analysis and then rearranged based on program transformation tools so that IRQ locks are either eliminated or at least canceled again in good time. This measure ultimately allows also the intended elimination of the persistence measures in the operating system that are superfluous due to NVRAM. For memory-hierarchy integration appropriate is a two-level hierarchy of software-managed caches that uses NVRAM as a buffer for data in conventional storage and DRAM as a buffer for NVRAM pages. The buffering of the pages in the DRAM is subject to strict time guarantees so that this logically persistent data can be reliably consolidated in the NVRAM again in exceptional cases. That is, in order to survive power failures, the maximum size of this DRAM page cache is to be aligned to the size of the remaining energy window in the power supply. Last but not least, energetic methods are designed that improve the mutual interaction of the operating system and NVRAM in such a way that the persistence properties of NVRAM are used to increase energy efficiency. Static NVRAM sleep modes are provided that actively reduce power consumption orthogonally to dynamic runtime improvements by NVRAM governors in the operating system.

In summary, the starting point of both projects is existing knowledge about NVRAM-based persistence of intermittently powered mobile devices on the one hand and persistence measures in stationary computing systems on the other. The basic assumption is that an NVRAM-based operating system manages computing more efficiently than a functionally identical DRAM-based twin. Subject of the study is the specific relationship between energy efficiency, computing power and latency in operating-system variants based on Linux, as far as NEON is concerned, and NVRAM-based capacity scaling and NVRAM virtualisation in FreeBSD, as PAVE contribution. Building on this, the general relationship in terms of applicability, transferability, and generalisation of the techniques developed for NEON and PAVE are examined.

This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Individual Research Grants 465958100 (NEON), 501993201 (PAVE), and 502228341 (Memento) as part of the Priority Programme on Disruptive Memory Technologies (SPP 2377).

References

1 Dushyanth Narayanan, Orion Hodson. *Whole-System Persistence*. ASPLOS XVII: Proceedings of the Seventeenth International Conference on Architectural Support for Programming Languages and Operating Systems, ACM, 2012

2 Ivy B. Peng, Maya B. Gokhale, Eric W. Green. *System Evaluation of the Intel Optane Byte-addressable NVM*. MEMSYS '19: Proceedings of the International Symposium on Memory Systems, ACM, 2019

3 Benjamin Ransford, Brandon Lucia. *Nonvolatile Memory is a Broken Time Machine*. MSPC '14: Proceedings of the workshop on Memory Systems Performance and Correctness, ACM, 2014

3.17 Heterogeneous, High-Performance Serverless Computing: Energy Implications and Opportunities

Devesh Tiwari (Northeastern University – Boston, US)

The next wave of cloud computing – the serverless computing model – is enjoying adoption at scale by different cloud computing vendors. The serverless computing model is already rapidly accelerating the development and deployment of enterprise applications. Unfortunately, the HPC community appears to be left behind in the revolution and it is not clear what are the energy implications/opportunities for HPC community if they adopted the serverless computing model. First, I'll discuss how HPC applications and workflows could attain higher performance and energy efficiency via hybrid execution [1, Mashup]. Second, I'll demonstrate how we can leverage server heterogeneity to reduce the overall energy consumption [2, IcBreaker]. Finally, I will discuss some predictive techniques to mitigate the bottlenecks of serverless computing [3, DayDream], and a brief introduction to a novel AI-workload dataset to enable carbon/aware, sustainable data center computing efforts [4].

References

1 Roy, Rohan Basu, Tirthak Patel, Vijay Gadepally, and Devesh Tiwari. "Mashup: making serverless computing useful for HPC workflows via hybrid execution." In Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP), pp. 46-60. 2022.

2 Roy, Rohan Basu, Tirthak Patel, and Devesh Tiwari. "IceBreaker: warming serverless functions better with heterogeneity." In Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), pp. 753-767. 2022.

3 Rohan Basu Roy, Tirthak Patel, and Devesh Tiwari. "DayDream: Executing Dynamic Scientific Workflows on Serverless Platforms with Hot Starts." In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC). 2022.

4 Li, Baolin, Rohin Arora, Siddharth Samsi, Tirthak Patel, William Arcand, David Bestor, Chansup Byun et al. "AI-Enabling Workloads on Large-Scale GPU-Accelerated System: Characterization, Opportunities, and Implications." In 2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA), pp. 1224-1237. IEEE, 2022.

3.18 How can we avoid ICT becoming the next Greenpeace target?

Samuel Xavier-de-Souza (Federal University of Rio Grande do Norte, BR)

Joint work of Samuel Xavier-de-Souza, Kerstin Eder

Information & Communication Technologies consume more than 10% of the energy produced on Earth, about twice as much as the aviation industry. According to an article from Nature, this number might rise above 20% by 2030. This talk discussed alternatives that might save the world from the digital revolution without stopping it. We discussed the historical facts that have led us to rely on energy-inefficient computing, the possible paths forward, and necessary actions to revert this. As takeaways we established that computing plays

a significant and active role in the present and future environmental challenges facing the world; that much of what we know and rely about computing changed in mid 2000's; that software, especially parallel software, is now even more important to the course of the digital revolution; and that software operation becomes a key concept to ensure optimisations are not wasted.

4 Working groups

4.1 Disruptive Paradigms

Michael Engel (Universität Bamberg, DE), Ahmed Ali-Eldin Hassan (Chalmers University of Technology – Göteborg, SE), Timo Hönig (Ruhr-Universität Bochum, DE), Alex Jones (University of Pittsburgh, US), Frank Mueller (North Carolina State University – Raleigh, US), and Wolfgang Schröder-Preikschat (Universität Erlangen-Nürnberg, DE)

Power and energy consumption depend on a large number of parameters on the hardware as well as software level of systems as well as on the interaction of hardware and software. With respect to disruptive paradigms, we can consider the impact of disruptions in hardware and software technology on power and energy and, conversely, the impact of focusing on a reduction of power and energy (on local device as well as global scale) on the design of technologies that form the foundations of future systems.

First, we identified a number of recent disruptions in hardware and software which have significant impact on power and/or energy. Trends such as the ubiquitous application of (deep) machine learning algorithms to all sorts of problems as well as the rise of digital currencies and blockchain technology have significantly increased power and energy requirements of typical applications. The impact of machine learning approaches can not only be identified on the training side, but also on the inference side [1]. Applications that make intensive use of machine learning inference, e.g. autonomous cars, exacerbate the problem since all on-board computing systems of autonomous vehicles compete for battery capacity with the electric drive of the unit. Here, it can also be observed that the complexity of related solutions is increasing rapidly, one example that was discussed is Google's TensorFlow software [2].

Especially the increase of software complexity has impacts not only on the local device level, e.g. on the battery runtime of a smartphone, but in turn on the global sustainability of devices. Since modern computing devices are no longer designed for upgradability (or repairability, though at least in the EU regulations related to repairability are currently in the making [5]), the plethora of incoming software updates results in devices being too "weak" – e.g. in terms of computing power or memory – for next year's version of the system and application software. In turn, users are forced to replace devices long before the end of their lifetime due to physical constraints. This is problematic on a global scale since a significant fraction of the CO2 (equivalent) emissions caused by electronic devices is generated by the *production* of devices, in many cases more than 50% of the overall impact. Accordingly, one simple way to reduce the CO2 impact of a device is to use it for a longer time [4].

The extended use of devices, however, is significantly constrained by the current mode in which software updates are handled by the device manufacturer. Devices connected to the Internet require frequent security updates to enable secure (and, in turn, safe) operation.

Even if updates are provided on a regular basis, there is often no way to obtain security updates independent of feature additions. Here, one *disruptive solution* would be to *require the separation of security and feature updates* and allow users to install only selected subsets of updates. However, this leads to an explosion of possible configurations which need to be tested, updated, and provided [6].

An even more disruptive approach is to start from scratch. Instead of trying to debloat software, as discussed elsewhere in this seminar, a renewed concentration on the development of *lean software*, as postulated by Niklaus Wirth already in 1995 [3], might be an approach to solve this side of the software crisis. This is also based on the observation that, as an extension of the Pareto principle, a significant fraction of the features of a software product are rarely or never used. For example, a DuPont study mentioned in an Agile 2002 keynote [7] concluded that "only 25% of a system's features are ever needed".

Another application of the Pareto principle was the basis of a followup discussion on power and energy optimization. Even though it is well known from software optimizations for performance, the question *which parts of the system to optimize* should be considered early on. Accordingly, the selection of optimizations to apply should not be based on the largest possible gain in the respective component, but on the frequency of use of a component. As with other non-functional properties, small gains in the reduction of energy consumption achieved on a hot path are more important than large gains that are only relevant for a small number of rarely executed corner cases.

When discussing the hardware-software interface, one central question is how to distribute the implementation of functionality between hardware and software components. Here, hardware can be more power and energy efficient due to customization to a specific problem as well as the larger grade of possible parallelism compared to software. Typical accelerators for often used functionality are in use in commodity systems for more than a decade, e.g. for video de- and encoding, general signal processing, cryptography, and neural networks. On the hardware-software interface layer, instruction set extensions provide a tradeoff between efficiency and flexibility, e.g. by the addition of DSP or vector instructions to conventional CPU instruction sets.

On the side of computer architecture, the possible power and energy impact of some recent developments was discussed. A significant contributor are embedded microcontroller chips. While the power and energy consumption of individual chips is low, the large number of deployed devices (6.7 billion ARM-based chips alone in the fourth quarter of 2020 [14]) make microcontrollers an interesting target for disruptive optimizations. The recently introduced Raspberry Pi Pico 2040 microcontroller does not follow the common approach to integrate flash memory for persistent storage on chip, but rather uses an external flash chip. Since integrating flash on-chip requires the use of relatively coarse semiconductor feature sizes, this also constrains the power scaling effects described by Dennard [15], whereas taking flash off-chip enables the downscaling of the remaining controller circuit. By coincidence, this also removed the reliance on specific manufacturing capacities that are in high demand during the current semiconductor supply crisis, making the Pico 2040 one of the few microcontrollers without significant supply constraints. Accordingly, the disaggregation of functionality to different chips (on a common carrier using chiplets or in separate packages) might be a – at first counterintuitive – approach to reduce power and energy consumption.

Disruptions related to *memory energy consumption* were also the focus of extended discussions. One significant contributor here is the requirement to periodically refresh DRAM contents. Several approaches to reduce the refresh overhead by using software and hardware approaches have been published recently [16]. A more disruptive approach is the use of

persistent, byte-addressable main memory, which is also the focus of the recently started German coordinated research project on *Disruptive Memory Technologies* [17] and was the subject of a presentation at this seminar [23]. Here, the discussion in the group diverged significantly from power- and energy-related topics to general questions about the hardware and software implications of persistent memory technologies, which are omitted here for brevity and focus.

However, one notable and possibly disruptive idea, related to lean software discussed above, was *cache-only computing*. Here, the idea is that CPU caches in today's systems are so large that significant parts of a software product (e.g., a microkernel of a system, see also the Pareto discussion above) can always remain in cache on a certain level, which could result in a significant reduction of memory traffic. Similar approaches to *cache-locking* of code have been employed in hard real-time systems for some time already [18]; investigating their energy impact will be interesting. Here, the discussion diverged into topics related to *predictable computing*, e.g. through the use of scratchpad memories instead of caches [19] and the creation of predictable computer systems out of unpredictable components, similar to von Neumann's early ideas on building dependable systems from unreliable components [20]. This, in turn, can be of relevance for improved worst-case energy consumption (WCEC) analyses [21], which e.g. are relevant to implement effective and efficient *intermittent computing* for IoT devices [22].

Another interesting *disruptive approach* could be the combination of accelerators and *approximate computing* [8]. It can be shown that in many applications, such as signal processing, perfect numerical precision is not required [9] – especially if the noise introduced by the signals to be processed is larger than the noise introduced by the computational approximation. The disruptive factor here could be to revisit analog computing components as parts of function-specific accelerators [10] and the introduction of hybrid digital/analog ("mixed signal") circuits. This approach can result in significant reductions of energy and power consumption, since certain computationally complex operations can be implemented by exploiting physical effects of electronic components. For example, integration is a numerically complex operation which is intrinsically implemented by a capacitor.

This move away from digital computing can be taken even further. A recent research direction is to attempt building computing systems that do not run on electricity, but instead are based on (synthetic) biological circuits [11]. For example, it is often stated that the human brain only consumes about 20 W of power while (sometimes) providing capabilities unmatched by any digital computer. First approaches to model brains using complex digital electronic systems are already in existence [12], while the development, programming, and integration of biological circuits into digital systems still seems to be a bit further away. Orthogonal to biological computing, implementing *storage using biological components such as DNA* [13] could be another disruptive approach.

In summary, the workgroup on disruptive paradigms discussed a large number of approaches, which in turn shows the vast size of the possible design space related to power and energy efficiency. Some of the discussed topics, such as software complexity and upgradability, also made the connection between localized, short-term power and energy effects and long-term global sustainability challenges. Overall, a single localized solution will not be sufficient – one major takeaway of the discussions in the related session is that a combination and cooperation of disruptive approaches on all levels of the hardware and software stack is required in order to have a positive impact on power and energy consumption. However, we have also noticed that "blind" optimizations without considering Pareto effects can be wasted efforts. Accordingly, power and energy are system-wide challenges which have to be optimized using carefully coordinated approaches.

References

1 Radosvet Desislavov, Fernando Martínez-Plumed, and José Hernández-Orallo. *Compute and energy consumption trends in deep learning inference*. arXiv preprint arXiv:2109.05472, 2021

2 Zejun Zhang, Yanming Yang, Xin Xia, David Lo, Xiaoxue Ren, and John Grundy. *Unveiling the mystery of API evolution in deep learning frameworks: a case study of Tensorflow 2*. In Proceedings of the 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP '21). IEEE Press, 238–247. https://doi.org/10.1109/ICSE-SEIP52600.2021.00033

3 Niklaus Wirth. *A Plea for Lean Software*. IEEE Computer, vol. 28, no. 02, pp. 64-68, 1995. doi: 10.1109/2.348001

4 Peter Marwedel and Michael Engel. *Plea for a Holistic Analysis of the Relationship between Information Technology and Carbon-Dioxide Emissions*. 23th International Conference on Architecture of Computing Systems 2010, pp. 1-6

5 Sahra Svensson et al. *The emerging 'Right to repair' legislation in the EU and the US*. Proceedings from Going Green–Care Innovation (2018): 27-29

6 Reinhard Tartler, Daniel Lohmann, Julio Sincero, and Wolfgang Schröder-Preikschat, W. *Feature consistency in compile-time-configurable system software: Facing the Linux 10,000 feature problem*. Proceedings of the sixth conference on Computer systems. 2011

7 Martin Fowler – report from Agile 2002, quoting DuPont study `http://martinfowler.com/articles/xp2002.html`

8 Sparsh Mittal. *A survey of techniques for approximate computing*. ACM Computing Surveys (CSUR) 48.4 (2016): 1-33.

9 Andreas Heinig, Vincent J. Mooney, Florian Schmoll, Peter Marwedel, Krishna Palem, and Michael Engel. *Classification-based improvement of application robustness and quality of service in probabilistic computer systems*. In International Conference on Architecture of Computing Systems (pp. 1-12). Springer, Berlin, Heidelberg

10 Glenn ER Cowan, Robert C. Melville, and Yannis P. Tsividis. *A VLSI analog computer/digital computer accelerator*. IEEE Journal of Solid-State Circuits 41.1 (2005): 42-53

11 Stuart A. Kurtz et al. *Biological computing*. Complexity theory retrospective II (1997): 179-195

12 Steve B. Furber et al. *The Spinnaker project*. Proceedings of the IEEE 102.5 (2014): 652-665

13 Yiming Dong et al. *DNA storage: research landscape and future prospects*. National Science Review 7.6 (2020): 1092-1107

14 Arm, Inc. *The Arm ecosystem ships a record 6.7 billion Arm-based chips in a single quarter*. `https://www.arm.com/company/news/2021/02/arm-ecosystem-ships-record-6-billion-arm-based-chips-in-a-single-quarter`

15 Robert H. Dennard et al. *Design of ion-implanted MOSFET's with very small physical dimensions*. IEEE Solid-State Circuits Society Newsletter 12.1 (2007): 38-50

16 Ishwar Bhati et al. *DRAM refresh mechanisms, penalties, and trade-offs*. IEEE Transactions on Computers 65.1 (2015): 108-121

17 Olaf Spinczyk and Jörg Nolte. *Disruptive Memory Technologies – DFG Priority Program 2377*. `https://spp2377.uos.de`

18 Isabelle Puaut, and Alexis Arnaud. *Dynamic instruction cache locking in hard real-time systems*. Proc. of the 14th Int. Conference on Real-Time and Network Systems. 2006

19 Rajeshwari Banakar, Stefan Steinke, Bo-Sik Lee, M. Balakrishnan, and Peter Marwedel. *Scratchpad memory: A design alternative for cache on-chip memory in embedded systems*. In Proceedings of the Tenth International Symposium on Hardware/Software Codesign. CODES 2002 (IEEE Cat. No. 02TH8627) (pp. 73-78)

20 John von Neumann. *Probabilistic logics and the synthesis of reliable organisms from unreliable components*. Automata studies 34 (1956): 43-98

21 Ramkumar Jayaseelan, Tulika Mitra, and Xianfeng Li. *Estimating the worst-case energy consumption of embedded software*. 12th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS'06). IEEE, 2006

22 Brandon Lucia et al. *Intermittent computing: Challenges and opportunities*. 2nd Summit on Advances in Programming Languages (SNAPL 2017) (2017)

23 Timo Hönig and Wolfgang Schröder-Preikschat. *On Energy-Awareness in NVRAM-based Operating Systems – NEON and PAVE*. In Power and Energy-aware Computing on Heterogeneous Systems (PEACHES), Dagstuhl Seminar 22341, 2022

4.2 Energy Optimization and Management

Maja Hanne Kirkeby (Roskilde University, DK), Ahmed Ali-Eldin Hassan (Chalmers University of Technology – Göteborg, SE), Liliana Cucu-Grosjean (INRIA – Paris, FR), Benedict Herzog (Ruhr-Universität Bochum, DE), Henry Hoffmann (University of Chicago, US), Fiodar Kazhamiaka (Stanford University, US), Laércio Lima Pilla (University of Bordeaux, FR), Simon Peter (University of Washington – Seattle, US), George Porter (University of California – San Diego, US), and Samuel Xavier-de-Souza (Federal University of Rio Grande do Norte, BR)

In this break-out session we discussed the aspect of Energy optimization and management for Power and Energy-aware Computing on Heterogeneous Systems. Our discussion sought to answer a question: How can we create energy optimal or near optimal software solutions? We propose a classification of open problems and associated challenges to identify relevant energy models such that multiple objectives and constraints like price, security, accuracy, time, memory, number of kernels, flexibility, and energy budget are achieved.

The computer stack has several layers, e.g., a simple stack covers developed applications, compilers creating the binaries which are executed by the operating system managing the hardware. Considering the stack from hardware and upwards; the higher the layer, the greater the abstraction, or seeing it top-down: when a layer provides its results to lower layers, it gives control to the lower layers. While the layered structure reduces complexity of the individual task and allows great flexibility, the layers obfuscate how different actions in one layer cause changes in the system's energy consumption. In an effort to improve transparency it is increasingly common to have the lower layers provide parameters to higher layers of the stack. However, due to the traditional abstractions, the higher layer may not be aware or take advantage. The aspect of energy transparency over the layers is discussed by another breakout session, so in this session we focus on how we can consider energy optimization and management over the whole system and on identifying open challenges.

Different studies show that individual layers can optimize for energy, however no individual layer can provide energy optimal solutions. Additionally, trying to optimize from several layers at the same time can cause energy inefficiencies; as layers that work without knowledge of the others cause destructive interference [1, 2] . Objectives for the optimization are given by the context and the user; prioritizing wanted effect (the output) to the controller(s). Each layer has different optimization opportunities and while a layer may not be able to find an optimal solution by itself, it may expose "knobs", i.e., variables, and constraints, e.g.,

max frequency, to controllers. We expect these variables to be domain dependent and, here, we keep the definition of controllers to a semantic definition where a controller decides the values of the variables. A controller could be a full-stack controller, the next system layers (upwards or downwards), or another type of control that governs the values.

In the following, we propose a general formalization of the Energy Optimization and Management Problem. Let there be a set of metrics (non-functional properties) such as price, security, accuracy, time, latency, memory, flexibility, energy budget, green house gas emissions. Some of these metrics will be objectives and some constraints, and over time their role may change. The different layers of the system expose configuration variables and their constraints. We can then formulate the global optimization problem covering the entire stack in the general form of:

```
optimize f(metrics, t)        // objective
    subject to                //constraints
    G(variables, t) <= b(t)     // this is a system of inequalities
                               // that express the constraints on
                               // the non-functional properties
                               //  and variable values
```

In general, energy (or power) could appear as part of the objective function or part of the constraints. For example, embedded systems might want to minimize energy given a target latency [3], while an HPC system might have a power constraint (dictated by the facility) and work to maximize application throughput given that constraint [4, 5]. The decision variables represent options provided by different layers of the stack. For example, the embedded application could expose different configuration variables representing the algorithm to use for detecting targets in a signal, with more accurate algorithms requiring more energy. The HPC system could use voltage and frequency scaling as variables that govern power and performance tradeoffs.

This form is a classical optimization form and can be used to express situations where a single controller is given a complete system model and full knowledge of the system's knobs, or a distributed approach such as when each layer has a separate controller that optimizes over the limited set of knobs in its domain. It can also express problems that are solved once apriori for a workload, or that are frequently re-solved for dynamic workloads that benefit from the system being tuned over time. Given the complexity of modern computing systems, a centralized controller approach is likely to be intractable, although it can be made tractable by identifying and removing variables that have little or no impact on the objective function, being able to find optimal solutions considering a reduced vertical design space. Depending on the problem, the design space is reduced in different ways, i.e., different parts of the solution are fixed:

1. from hardware to solution: given a fixed hardware, how may I write my software to solve a specific task while optimizing for a specified objective?
2. from software to platform: given my software, how can I find the hardware and/or software system that enables me to solve my specific task while optimizing for a specified objective?
3. HW/SW co-design or co-optimization: given this general scenario for my software solving a class of tasks on a family of platforms, how can I optimize for a specified objective?

Formalizing the problem in this way can help us understand the computational complexity of the optimization problem, as well as reveal how objectives can be incorporated into the application development, e.g., accuracy, and a future challenge lies in finding the best ways to build software that can be optimized for energy.

To enable the use of this formulation in practice, we identified the following open challenges:

1. Identifying which variables are important to optimize for a given problem class, e.g., a given application or class of applications, or specific hardware, or family of hardware. The fewer variables exposed, the more tractable the problem. It is important that the exposed variables are the significant ones.
2. Developing effective interfaces that allow each layer to expose the variables and receive information from other layers.
3. Developing energy models that express the relationship between variables and objectives. We speculate that models could be both white-box models such as sets of equations or black-box learned models.

References

1 Henry Hoffmann. 2015. JouleGuard: energy guarantees for approximate applications. In Proceedings of the 25th Symposium on Operating Systems Principles (SOSP '15). Association for Computing Machinery, New York, NY, USA, 198–214. https://doi.org/10.1145/2815400.2815403

2 H. Hoffmann, "CoAdapt: Predictable Behavior for Accuracy-Aware Applications Running on Power-Aware Systems," 2014 26th Euromicro Conference on Real-Time Systems, 2014, pp. 223-232, doi: 10.1109/ECRTS.2014.32.

3 C. Imes, D. H. K. Kim, M. Maggio and H. Hoffmann, "POET: a portable approach to minimizing energy under soft real-time constraints," 21st IEEE Real-Time and Embedded Technology and Applications Symposium, 2015, pp. 75-86, doi: 10.1109/RTAS.2015.7108419.

4 Huazhe Zhang and Henry Hoffmann. 2019. PoDD: power-capping dependent distributed applications. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '19). Association for Computing Machinery, New York, NY, USA, Article 28, 1–23. https://doi.org/10.1145/3295500.3356174

5 Neha Gholkar, Frank Mueller, Barry Rountree, and Aniruddha Marathe. 2018. PShifter: feedback-based dynamic power shifting within HPC jobs for performance. In Proceedings of the 27th International Symposium on High-Performance Parallel and Distributed Computing (HPDC '18). Association for Computing Machinery, New York, NY, USA, 106–117. https://doi.org/10.1145/3208040.3208047

4.3 Results and Insights from the Green Computing Hackathon

Sven Köhler (Hasso-Plattner-Institut, Universität Potsdam, DE), Benedict Herzog (Ruhr-Universität Bochum, DE), Henriette Hofmeier (Ruhr-Universität Bochum, DE), Maja Hanne Kirkeby (Roskilde University, DK), Max Plauth (Hasso-Plattner-Institut, Universität Potsdam, DE), and Lukas Wenzel (Hasso-Plattner-Institut, Universität Potsdam, DE)

Despite energy consumption of software being an omnipresent topic of discussion, knowledge about means to measure and quantify that energy demand is less widely spread. Complementing the scientific talks at the PEACHES seminar, a series of three practical sessions was held over the course of several days.

In total, 36 persons took part in this "Green Computing Hackathon". Participants were given a choice of multiple hardware platforms (Laptops, PCs and various embedded boards), as well as example workloads by the organizers. Alternatively, they were encouraged and supported in using their own hardware and investigating self-provided problems from their own research domains.

The first session started with an introductory talk on measurement facilities, hardware counters and a showcase of software tools like "PinPoint" or "likwid" for retrieving energy and power readings on multiple platforms. Furthermore, good practices and avoidance of common pitfalls in energy measurements were discussed.

Among the workloads and benchmark suites investigated by the participants, the most popular was "heatmap", a simple round-based convolution simulation. Participants compared the influence of different compiler flags, optimisation levels, parallelization strategies, task sizes, clock frequencies, hardware platforms, and vector extensions, as well as the difference between CPU- and GPU-based implementations.

In an extended session, Alex K. Jones demonstrated the "GreenChip" tool, that estimates the carbon footprint for production and application phases in the lifecycle of chip designs. Taking into account factors such as chip area population, chip operational energy, manufacturing technology node, and power grid mix, the tool is focused on indifference and break-even analyses between alternative design choices.

The participants' feedback for all three hackathon sessions was overwhelmingly positive and we highly encourage future Dagstuhl Seminars to include comparable hands-on sessions.

Tools used:

- `https://github.com/osmhpi/pinpoint`
- `https://github.com/RRZE-HPC/likwid/wiki/Likwid-Powermeter`
- `https://github.com/Pitt-Jones-Lab/Greenchip`

List of hackathon organizers:
Maja Hanne Kirkeby (Roskilde Universitet), Benedict Herzog, Henriette Hofmeier (Ruhr-Universität Bochum), Max Plauth, Lukas Wenzel, Sven Köhler (Hasso Plattner Institute Potsdam)

4.4 Energy Transparency across the Hardware/Software Stack

George Porter (University of California – San Diego, US), Ahmed Ali-Eldin Hassan (Chalmers University of Technology – Göteborg, SE), Antonio Carlos Schneider Beck Filho (Federal University of Rio Grande do Sul, BR), Ruzanna Chitchyan (University of Bristol, GB), Kerstin I. Eder (University of Bristol, GB), Christian Eichler (Ruhr-Universität Bochum, DE), Mathias Gottschlag (KIT – Karlsruher Institut für Technologie, DE), Daniel Gruss (TU Graz, AT), Maja Hanne Kirkeby (Roskilde University, DK), Sven Köhler (Hasso-Plattner-Institut, Universität Potsdam, DE), Julia Lawall (INRIA – Paris, FR), Simon Peter (University of Washington – Seattle, US), Max Plauth (Hasso-Plattner-Institut, Universität Potsdam, DE), Sibylle Schupp (TU Hamburg, DE), and Samuel Xavier-de-Souza (Federal University of Rio Grande do Norte, BR)

Five to ten years in the future, the power landscape will have fundamentally changed. The world will have a high proportion of renewables, with volatile power generation. Impact from climate change will be felt much more directly, with power supply variability due to grid disruption sharply up. Information and communication technology (ICT) will also constitute a much larger fraction of the world's demand for power, due to the end of Dennard scaling and increasing demand for more computation due to machine learning. In this world, power must be a first-class design constraint for all aspects of the systems design process.

The working group discussed how to achieve better transparency of power supply and demand across the systems hardware and software stack. We ask the question: *How can we poke through the entire stack of layers in a compute system to allow relevant energy information to flow across the layers to where it is needed? How can lower layers provide energy use information? How can higher layers communicate performance requirements?*

4.4.1 Artifacts and grand challenges

A number of grand challenges and research artifacts that are necessary for power transparency were identified:

- Carbon/PowerTop at scale. This tool would identify top power consumers across a cluster of machines. It should be able to break down power consumption into increasingly fine grain consumers, including to virtual machines, processes, users, functions, and instructions, across user and kernel modes. It should also be able to break down consumption to hardware components, including network switches, peripherals, IO devices, accelerators, network links and interfaces.
- Top-Down analysis for power. Such a tool would help pinpoint sources of power consumption at the microarchitectural level. Similar to Intel's popular Top-Down microarchitectural analysis for performance (cf. https://github.com/andikleen/pmu-tools), this tool would descend the hierarchy of microarchitectural components for progressively finer-grain views into power consumption.
- PowerLint. This tool would help find power bugs and suggest fixes via source code analysis. A potentially interesting angle is to expand the scope of this tool to a planetary database of major power consumers. It could then suggest fixes also to code that is not a power center in isolation, but becomes so by being part of popular uses that collectively consume a large amount of power via long tail effects.
- McPowerAfee. A scanner for power viruses.

4.4.2 Intellectual challenge

All of these artifacts require power attribution. Providing it requires solving a number of challenges in hardware, OS, language, compiler, and toolchains. We developed a rough order of importance of intellectual challenges to be solved to attribute power at the required fidelity and granularity.

The first challenge is to achieve better fidelity and granularity in time and space at the hardware level. Existing technologies, such as Intel's RAPL, operate at relatively coarse granularities of roughly 1ms (some down to 50-150us), making it difficult to attribute power use to functions and instructions. To do so, we need power attribution down to nanoseconds. We also need models and instrumentation for power attribution to microarchitectural components, including TLBs, caches, memory banks, and IO lanes. For example, it was shown that storing a data structure across multiple DRAM banks uses more power. To reduce power usage, we have to be able to account for these effects. Finally, power attribution into the power utility infrastructure (AC, power distribution, power storage, ...) would be necessary to determine large-scale power draw. The breakout group believes that power usage effectiveness (PUE) is not enough to characterize the power draw of these components. While Hyperscalers are near PUE = 1, the edge is not. Further, PUE is usually reported as an average. In reality, it varies based on utilization. Actuation overheads also need to come down to be able to react appropriately. For example, the latency to switch among C states, enter/exit hibernation, currently have high hardware and especially software overheads.

The second challenge is how to achieve scale of attribution. We need to trace power use across clusters of machines, switches, storage devices, memory, etc. This requires scalable, low-overhead mechanisms and data structures to identify power centers at high fidelity and over short timescales.

The third challenge is to trace power use across virtualization and abstraction boundaries. To trace power across the system stack, we need APIs and execution environments to communicate power requirements and demand among power consumers and producers. Virtualization also implies dealing with power as a virtual resource. Power may be stranded if over-provisioned to a single VM that does not use all of the power. Power may be unavailable if over-committed to many VMs that collectively use more than is available.

The fourth challenge is quality of service under power constraints. Many service level agreements include slack that may be used to trade performance for power. To do so, power consumption models are needed. Program configuration changes may also be used to trade fidelity for power. For example, deep neural networks (DNNs) can turn off layers to save power, with a quantified quality loss. To do all of this, we need interfaces for software to specify its power-QoS tradeoff space. A few example uses are power-fair scheduling and power-proportionate computing.

The fifth challenge is automation. Many developers and users do not care or do not know how to optimize for power. We need compilers and programming language tools that help us automate power optimization and tracing across the systems stack. Power could become a non-functional specification of the design and build process of modern software. The specification could be used by compilers for algorithm selection and fitting of algorithms to hardware.

4.4.3 Carbon versus power

We also discussed operational carbon issues of power. However, the main conclusion was that carbon simply modulates power cost, so it can be easily modeled as part of a power variability problem.

4.4.4 Finding and fixing power bugs

We lack an intuition of what uses how much energy (missing conceptual model) and our expectations sometimes break.

power bug: *implementation or hardware caused increased power consumption unrelated to the algorithm.*

How to find power bugs? This requires an energy profiler (attribution of which operation/program section/part of the system consumes which amount of energy). However, an energy linter is what we actually want. The linter processes the (expectedly huge) amounts of profiler output and identifies what part we should look out for.

How to attribute power to code, in particular under presence of interactions with "intelligent" devices (e.g., DMA, Accelerators, Hard Drives) is a prerequisite for a profiler.

Measurement: Use RAPL (Resolution: CPU Package µJ per 1ms, DRAM µJ per 1ms, power planes µJ per 50µs, core voltage V per 150µs), also include off-chip energy consumption (DRAM activity as a result of what ever happens below the last cache level, diverse IO operations affecting devices potentially across the entire data center (stuff happening on the storage network, ...); for GPUs there are similar performance counters/registers.

Mapping: we need to attribute the RAPL measurements back to an instruction/basic block granularity. 1ms/50us translates to quite a large window of possible "culprits" in the original program. We could pad functions/basic blocks/... with nop/pause/... to full RAPL windows for measurement (will be slower than regular execution but allows for measurement).

We also should go beyond the processor for true end to end evaluation. This is even more difficult on the IO level, where packets/operations from multiple sources get aggregated and share the blame. What about indirect effects like spinning up fans just because of an unfortunate simultaneous placement of two independent computationally heavy threads, that even might have happened seconds ago (heat propagates slowly)?

We might also want to account for activity in remote machines triggered by a local action. From that, attribute what part of the remote power consumption was caused by the local action. This step is controversial. The execution time profiling community hasn't solved the issue and they have been working on this for a long time. Hence, this is an overly ambitious, risky step. To resolve this issue, we should get reliable estimates of which proportion of total energy for a workload execution is consumed by the processor alone. This would require a controlled setup in an otherwise quiet datacenter, but it might be illuminating.

How to detect power bugs? We need to establish a firm understanding of which patterns/behavior cause power bugs to identify them. We can do this in a variety of ways. Empirically: run extensive benchmark suites, analyze behavior and generalize appropriately. However, there is a wide variety of architectures and implementations, so it is challenging to have a comparable results collected. Conceptually: reason from underlying micro- to system architectures, which patterns are expected to have detrimental effects. Big Data: collect large data sets from systems in regular operation and analyze those. "Bugs" may also be observed from a divergence between the energy requirements (i.e., features/properties expected by the software stakeholder or owner who pays for the software) and the delivered implementation. E.g., if the customer wants "green star/Energy Efficiency" badge, but is not delivered the required energy behavior, that is a power bug.

4.5 Sustainable Computing

Andreas Schmidt (Universität des Saarlandes – Saarbrücken, DE), Henriette Hofmeier (Ruhr-Universität Bochum, DE), Alex Jones (University of Pittsburgh, US), Daniel Mosse (University of Pittsburgh, US), Frank Mueller (North Carolina State University – Raleigh, US)

In the "Sustainable Computing" breakout session, we looked beyond energy transparency, management, and optimization (which was the focus of the parallel groups). One of our first conclusion was that the footprint of computing services should become more transparent. With computing services, we also mean the physical products involved, such as servers or smart phones. In this case, footprint is not only limited to energy usage in operation, but extends to energy required to manufacture hardware as well as other sustainability metrics related to manufacturing. The latter includes, for instance, CO2 equivalents or effects on humans, e.g. carcinogens, disability-adjusted life years (DALY), or volatile organic compounds (VOC). Generally speaking, we believe that Life-cycle Assessment (LCA) methods should be more thoroughly applied to computing systems to improve transparency.

The *Planetary Health Diet*[1] is a diet that allows a certain population (10bn by the year 2050) to life without hunger, while respecting earth's natural resources. Similarly, we came up with an idea to develop a *Planetary Digital Diet* that governs what a sustainable (and healthy) consumption of digital services would be. Analogous to the grams per day for different food categories (e.g. red meat or vegetables), one could come up with minutes per day for different digital services (e.g. office software on desktop or mobile games). We also applied the "Five R's of Sustainability" (5Rs) to computing. The idea behind the 5Rs is to provide multiple steps at which we can do something with products before we, in the worst case, put them in a landfill or burn them. *Refuse* is about avoiding a footprint in the first place, while *reduce* aims for a lower footprint, if it cannot be avoided. When we *reuse*, we continue using a product in a way that is different from the original manufacturer's idea; either because it can no longer be used for its original purpose or we do not need it anymore. *Recycle* is the (usually lossy and energy-intensive) process of turning the product into pieces and using the pieces to eventually create a new product. Lastly, *rot* is "giving back to nature", thereby keeping the resources in our ecosystem. Applied to computing, we came up with:

- *Refuse*: How can users that care about sustainability refuse digital services that would increase their footprint? Which analogue solutions are more sustainable than digital equivalents? How can we stop providers from offering free-of-charge, unlimited services? How can data minimization policies, e.g. GDPR, have a positive sustainability impact?
- *Reduce*: How can we nudge users to reduce their consumption of digital services (and hence, reduce their footprint)? How can existing and upcoming technology be altered to be more sustainable?
- *Reuse*: How can we foster the reuse of software instead of reproducing it? How can software be easily reduced in footprint (debloated)? How can software be designed to be easily reusable? How can we foster the reuse of data and models, in particular via open science/source? How can we enable a second life for (more) hardware (analogous to car batteries being reused in stationary contexts)?

[1] `https://en.wikipedia.org/wiki/Planetary_health_diet`

- *Recycle*: How can computing hardware be changed to allow for better decomposition, allowing reparation and recycling of individual components?
- *Rot*: How can computing systems be build in a biodegradable way? How can renewable materials be used to build hardware?

In summary, we can conclude that there are various steps to be taken to make computing software and hardware more sustainable as well as educate users in sustainable consumption. These steps, in synergy with improving energy-usage that our digital services rely on, can help to create a more sustainable digital future.

Participants

- Ahmed Ali-Eldin Hassan
Chalmers University of Technology – Göteborg, SE
- Antonio Carlos Schneider Beck Filho
Federal University of Rio Grande do Sul, BR
- Ruzanna Chitchyan
University of Bristol, GB
- Liliana Cucu-Grosjean
INRIA – Paris, FR
- Julian De Hoog
The University of Melbourne, AU
- Kerstin I. Eder
University of Bristol, GB
- Christian Eichler
Ruhr-Universität Bochum, DE
- Michael Engel
Universität Bamberg, DE
- Mathias Gottschlag
KIT – Karlsruher Institut für Technologie, DE
- Daniel Gruss
TU Graz, AT
- Benedict Herzog
Ruhr-Universität Bochum, DE
- Timo Hönig
Ruhr-Universität Bochum, DE
- Henry Hoffmann
University of Chicago, US
- Henriette Hofmeier
Ruhr-Universität Bochum, DE
- Romain Jacob
ETH Zürich, CH
- Alex Jones
University of Pittsburgh, US
- Fiodar Kazhamiaka
Stanford University, US
- Maja Hanne Kirkeby
Roskilde University, DK
- Sven Köhler
Hasso-Plattner-Institut, Universität Potsdam, DE
- Julia Lawall
INRIA – Paris, FR
- Laércio Lima Pilla
University of Bordeaux, FR
- Tulika Mitra
National University of Singapore, SG
- Daniel Mosse
University of Pittsburgh, US
- Frank Mueller
North Carolina State University – Raleigh, US
- Simon Peter
University of Washington – Seattle, US
- Max Plauth
Hasso-Plattner-Institut, Universität Potsdam, DE
- George Porter
University of California – San Diego, US
- Andreas Schmidt
Universität des Saarlandes – Saarbrücken, DE
- Gunnar Schomaker
Universität Paderborn, DE
- Wolfgang Schröder-Preikschat
Universität Erlangen-Nürnberg, DE
- Sibylle Schupp
TU Hamburg, DE
- Jennifer Switzer
University of California – San Diego, US
- Devesh Tiwari
Northeastern University – Boston, US
- Lukas Wenzel
Hasso-Plattner-Institut, Universität Potsdam, DE
- Samuel Xavier-de-Souza
Federal University of Rio Grande do Norte, BR

Report from Dagstuhl Seminar 22342

Privacy in Speech and Language Technology

Simone Fischer-Hübner*[1], Dietrich Klakow*[2], Peggy Valcke*[3], and Emmanuel Vincent*[4]

1 **Karlstad University, SE.** simone.fischer-huebner@kau.se
2 **Saarland University – Saarbrücken, DE.** dietrich.klakow@lsv.uni-saarland.de
3 **KU Leuven, BE.** peggy.valcke@kuleuven.be
4 **Inria – Nancy, FR.** emmanuel.vincent@inria.fr

Abstract

This report documents the outcomes of Dagstuhl Seminar 22342 "Privacy in Speech and Language Technology". The seminar brought together 27 attendees from 9 countries (Australia, Belgium, France, Germany, the Netherlands, Norway, Portugal, Sweden, and the USA) and 6 distinct disciplines (Speech Processing, Natural Language Processing, Privacy Enhancing Technologies, Machine Learning, Human Factors, and Law) in order to achieve a common understanding of the privacy threats raised by speech and language technology, as well as the existing solutions and the remaining issues in each discipline, and to draft an interdisciplinary roadmap towards solving those issues in the short or medium term.

To achieve these goals, the first day and the morning of the second day were devoted to 3-minute self-introductions by all participants intertwined with 6 tutorials to introduce the terminology, the problems faced, and the solutions brought in each of the 6 disciplines. We also made a list of use cases and identified 6 cross-disciplinary topics to be discussed. The remaining days involved working groups to discuss these 6 topics, collaborative writing sessions to report on the findings of the working groups, and wrap-up sessions to discuss these findings with each other. A hike was organized in the afternoon of the third day.

The seminar was a success: all participants actively participated in the working groups and the discussions, and went home with new ideas and new collaborators. This report gathers the abstracts of the 6 tutorials and the reports of the working groups, which we consider as valuable contributions towards a full-fledged roadmap.

Seminar August 21–26, 2022 – https://www.dagstuhl.de/22342

2012 ACM Subject Classification Artificial Intelligence → Natural Language Processing; Security and Privacy → Human and Societal Aspects of Security and Privacy; Security and Privacy → Software and Application Security; Security and Privacy → Database and storage security

Keywords and phrases Privacy, Speech and Language Technology, Privacy Enhancing Technologies, Dagstuhl Seminar

Digital Object Identifier 10.4230/DagRep.12.8.60

* Editor / Organizer

Privacy in Speech and Language Technology, *Dagstuhl Reports*, Vol. 12, Issue 8, pp. 60–102
Editors: Simone Fischer-Hübner, Dietrich Klakow, Peggy Valcke, Emmanuel Vincent

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

1 Executive Summary

Simone Fischer-Hübner (Karlstad University, SE)
Dietrich Klakow (Saarland University – Saarbrücken, DE)
Peggy Valcke (KU Leuven, BE)
Emmanuel Vincent (Inria – Nancy, FR)

In the last few years, voice assistants have become the preferred means of interacting with smart devices and services. Chatbots and related language technologies such as machine translation or typing prediction are also widely used. These technologies often rely on cloud-based machine learning systems trained on speech or text data collected from the users. The recording, storage and processing of users' speech or text data raises severe privacy threats. This data contains a wealth of personal information about, e.g., the personality, ethnicity and health state of the user, that may be (mis)used for targeted processing or advertisement. It also includes information about the user identity which could be exploited by an attacker to impersonate him/her. News articles exposing these threats to the general public have made national headlines.

A new generation of privacy-preserving speech and language technologies is needed that ensures user privacy while still providing users with the same benefits and companies with the training data needed to develop these technologies. Recent regulations such as the European General Data Protection Regulation (GDPR), which promotes the principle of privacy-by-design, have further fueled interest. Yet, efforts in this direction have suffered from the lack of collaboration across research communities. This Dagstuhl Seminar was the first event to bring 6 relevant disciplines and communities together: Speech Processing, Natural Language Processing, Privacy Enhancing Technologies, Machine Learning, Human Factors, and Law.

After 6 tutorials given from the perspective of each of these 6 disciplines, the attendees gathered into cross-disciplinary working groups on 6 topics. The first group analyzed the privacy threats and the level of user control for a few case studies. The second group focused on anonymization of unstructured speech data and discussed the legal validity of the success measures developed in the speech processing literature. The third group devoted special interest to vulnerable groups of users in regard to the current laws in various countries. The fifth group tackled the design of privacy attacks against speech and text data. Finally, the sixth group explored the legal interpretation of emerging privacy enhancing technologies.

The reports of these 6 working groups, which are gathered in the following, constitute the major result from the seminar. We consider them as a first step towards a full-fledged interdisciplinary roadmap for the development of private-by-design speech and language technologies addressing societal and industrial needs.

2 Table of Contents

3 Overview of Talks

3.1 Speech privacy

Emmanuel Vincent (Inria – Nancy, FR) emmanuel.vincent@inria.fr

Large-scale collection, storage, and processing of speech data poses severe privacy threats. Indeed, speech encapsulates a wealth of personal data (e.g., age and gender, ethnic origin, personality traits, health and socio-economic status, etc.) which can be linked to the speaker's identity via metadata or via automatic speaker recognition. Speech data may also be used for voice spoofing using voice cloning software. In this tutorial, I provide an overview of privacy preservation solutions for speech data, with a focus on voice anonymization. I define the voice anonymization task and evaluation metrics, and outline solutions based on voice conversion and differential privacy. I also briefly mention federated learning, and conclude by stating open questions for future research.

3.2 Privacy-enhancing natural language processing

Pierre Lison (Norsk Regnesentral – Oslo, NO) plison@nr.no

This tutorial describes the main aspects of privacy-enhancing techniques developed in the field of Natural Language Processing. We first explain the main privacy risks that may arise from processing text or training natural language processing models. We then review a number of privacy-enhancing techniques, in particular text sanitization, text obfuscation, text rewriting and synthesis, and privacy-preserving training of natural language processing models. We also discuss a number of open challenges and research questions.

3.3 Privacy from a security perspective

Meiko Jensen (Karlstad University, SE) meiko.jensen@kau.se

From a technical or security perspective, privacy has specific connotations and definitions beyond legal or societal dimensions. Especially in the process of designing IT systems, such as AI-based natural language processing systems, these challenges must be addressed appropriately, based on a common understanding of the exact notions of each domain. In this tutorial, I provide some technical definitions of common privacy-related concepts (such as anonymity, or the difference between data and information), and I explain the approach of the protection goals for privacy engineering as an interdisciplinary effort to harmonize privacy considerations at the intersection of law, society, and information technology.

3.4 Privacy issues and mechanisms in machine learning

Olga Ohrimenko (University of Melbourne, AU) oohrimenko@unimelb.edu.au

Machine learnings models, including those that process text, can leak information about their training data. This has been demonstrated by several attacks (e.g., identifying whether a record in the training dataset, extraction of phrases). Algorithms and mechanisms for protecting training data can be grouped into those that can protect against a data collector and those that protect from a user of the trained model (e.g., for text generation). Secure hardware, cryptographic techniques and local differential privacy can be used for the former setting and have a set of tradeoffs in terms of guarantees, assumptions, and performance. The latter group includes central differential privacy. Though differential privacy is seeing adoption in practice, its applicability for text and speech is an open question and depends on a unit of privacy one is interested in protecting (e.g., a user, a phrase, an utterance, voice) that may be difficult to define.

3.5 Human factors in privacy

Zinaida Benenson (Friedrich-Alexander-Universität – Erlangen, DE) zinaida.benenson@fau.de

This tutorial discusses how people make decisions about sharing or withholding their data towards commercial organization, governmental organizations and individuals. Unfortunately, we cannot expect people to act in their best interest in this domain. Privacy decisions are subject to many psychological effects: they are heavily dependent on context (who asks, in which order, how the request is framed) and to well-known behavioral biases such as unrealistic optimism and immediate gratification. Moreover, people overestimate risks of terrorism and similar high-emotion threats, which makes them susceptible to the rhetoric of "surveillance for the greater good", no matter whether this surveillance actually reduces risk. Additionally, ubiquitous presence of IoT devices in private and public spaces raises new issues concerning interpersonal privacy: how to negotiate differing privacy preferences of different users, such as regular inhabitants of smart homes and bystanders.

3.6 Voice and speech: the perspective of legal scholars

Lydia Belkadi (KU Leuven, BE) lydia.belkadi@kuleuven.be
Abdullah Elbi (KU Leuven, BE) abdullah.elbi@kuleuven.be
Peggy Valcke (KU Leuven, BE) peggy.valcke@kuleuven.be
Els Kindt (KU Leuven, BE) els.kindt@kuleuven.be

The entanglement of different attributes within speech and text snippets raises important challenges from a legal perspective. For example, it is unclear how speech or text snippets should be defined from a legal perspective or how to apply existing legal definitions. Similarly, this entanglement implies considerable contradictions with data protection principles, such as data minimization and purpose limitation. In other words, snippets may reveal more data than is necessary for a given purpose (e.g., text and typing patterns in language processing). In addition, from a legal perspective, special attention must be given to the concept of vulnerability where the wide spread use of speech technologies may create new type of vulnerabilities. In some situations, the users' right to privacy may conflict with the voice technology company's legal requirements. For example, if the voice technology company collects speech or text data suggesting that a crime (e.g., child abuse) or a life-threatening danger (e.g., heart attack) has taken place, should it report it to the relevant authority, thereby violating the user's privacy? Questions identified during the law tutorial included:

- Are there practices that should be prohibited? What are red lines to the use of voice snippets? (in light of existing/possible safeguards) What are risky applications? (e.g., emotions – what is technically possible or not possible?)
- Can we work towards a common terminology / vocabulary to carry out risk assessments / Data Protection Impact Assessments?
- Should we consider "outliers" (i.e., people whose voice is more identifiable than others) as a new vulnerable group? Novel Speech and Language Technologies as creating new types of vulnerabilities?
- Ethical / moral questions: shall the staff of voice technology companies intervene in the situations when they pick up worrying situations while screening users' voice data? Shall large-scale users' speech and language data be used as legal evidence?

4 Working Groups

4.1 Case studies and user interaction

Zinaida Benenson (Friedrich-Alexander-Universität – Erlangen, DE) zinaida.benenson@fau.de
Abdullah Elbi (KU Leuven, BE) abdullah.elbi@kuleuven.be
Zekeriya Erkin (TU Delft, NL) z.erkin@tudelft.nl
Natasha Fernandes (Macquarie University – Sydney, AU) natasha.fernandes@mq.edu.au
Simone Fischer-Hübner (Karlstad University, SE) simone.fischer-huebner@kau.se
Ivan Habernal (TU Darmstadt, DE) ivan.habernal@tu-darmstadt.de
Els Kindt (KU Leuven, BE) els.kindt@kuleuven.be
Anna Leschanowsky (Fraunhofer IIS – Erlangen, DE) anna.leschanowsky@iis-extern.fraunhofer.de
Pierre Lison (Norsk Regnesentral – Oslo, NO) plison@nr.no
Christina Lohr (Friedrich-Schiller-Universität – Jena, DE) christina.lohr@uni-jena.de
Emily Mower Provost (University of Michigan – Ann Arbor, US) emilykmp@umich.edu
Jo Pierson (Free University of Brussels, BE) jo.pierson@vub.be
David Stevens (Gegevensbeschermingsautoriteit – Brussels, BE) david.stevens@apd-gba.be
Francisco Teixeira (Instituto Superior Técnico – Lisbon, PT) francisco.s.teixeira@tecnico.ulisboa.pt
Shomir Wilson (Pennsylvania State University – University Park, US) shomir@psu.edu

Two separate working groups were initially created on case studies, stakeholders, risks, and benefits on the one hand, and on user control on the other hand. After the first discussion session, they decided to merge. Hence we present their joint outcomes below.

4.1.1 Existing uses of speech and language technology

Speech and natural language are fundamental to human communication, and they serve as conduits for enormous amounts of personal information. Language technology users share information across a spectrum of levels of privacy sensitivity, from mild to acutely strong.

Uses of speech and language technologies emerged early in the era of digital computers and in recent years they have become ubiquitous. We list some currently existing technologies to motivate the discussion that follows. Many of these may involve a combination of spoken language, acoustics, or written language:

- call center monitoring, e.g., to evaluate the performance of call center agents,
- automated phone menu systems,
- medically-focused technologies, e.g., for diagnosis or tracking symptom severity,
- language learning, e.g., apps for learning to read or speak a second language,
- voice assistants, such as Amazon's Alexa and Apple's Siri,
- machine translation between natural languages,
- law enforcement and security, e.g., to detect malicious activity,
- web search, which (like many items in this list) could be text or speech,
- search specific to websites or services, such as on Amazon.com or Facebook,
- large-scale analysis of documents, such as legal documents like court records or laws,
- online social networks, such as Twitter and TikTok,
- writing support services, such as Grammarly.

4.1.2 Stakeholders

Stakeholders in speech and voice technology include:

- the individual, i.e., the person whose voice or language are being processed, also referred to as the data subject (in some cases, this individual might actually also be the user of a speech or language technology or only the data subject),
- other individuals, e.g., whose voices are incidentally included in speech audio recordings, or who may be the subject of text written by the individual,
- the first-party service provider, with whom the individual directly interacts,
- third parties (i.e., external to the user and the first party) that the first party shares an individual's data with to fulfill aspects of their service,
- third parties that the first party shares an individual's data with for nonessential purposes, e.g., marketing-focused data brokers,
- government entities, including public agencies and law enforcement,
- the individual's employer or school, if applicable,
- data protection authorities.

This list is not meant to be comprehensive and other stakeholders are likely to exist.

4.1.2.1 Data provenance

We specify three common categories of data sources, acknowledging that there may be more:

- *input data*, that is information disclosed through participation by the individual and provided by the individual to the speech and/or language application,
- *inferred data*, that is data created by the application automatically or manually by labels/annotations of the data received, where the labels/annotations were not obtained by the participation of the individual,
- *metadata*, that is technical information associated with either the input data or inferred data, e.g., time stamps, location data, etc.

Note: very recently (August 1st 2022) the Court of Justice of the European Union ruled that the level of protection is the same for sensitive data directly provided by the individual itself, as for other types of (non-sensitive) personal data from which "sensitive information" (e.g., political preference, sexual preference, etc., see Article 9 of the GDPR) can be inferred. Applied to voice technology, this means that the higher standards of protection (as sensitive data, e.g., "explicit consent" vs. "normal consent") would be applicable to all voice and language technologies.[1]

4.1.2.2 Preliminary categorization

As a next step, we have trimmed down the list of uses of speech and language technology to a more workable number of types of uses from a data protection risk-based perspective. In this respect, two criteria of risk seem particularly relevant. First, we take into account the situations in which the processing will take place (e.g., on-device). This allows us to describe risk in terms of the likelihood of information leakage. The second criterion we applied is the potential combination of data (because combinations of voice related data with other types of personal data are likely to be more problematic from a data protection point of view).

[1] https://curia.europa.eu/juris/document/document.jsf?text=&docid=263721&pageIndex=0&doclang=EN&mode=req&dir=&occ=first&part=1&cid=481514

Finally, we also consider the number of parties that can have access to the personal data as an indicator of increasing risk to the private sphere of the individual involved.

Applying these criteria, we identify the following three categories of situations in which speech and language *data can be processed*:

1. locally on a user device, also referred to as "on-device" processing, where input data and inferred data (see definition of terms) does not leave the device (maximum user control and most limited number of parties involved),
2. networked or connected services in which input data and/or inferred data are transmitted from the device that recorded input data (e.g., provided by a commercial service provider, for example online communication between users),
3. processing of data without active intervention or request of the individual (e.g., in the public domain by a public authority, for example usage of voice enabled cameras in public areas, or using voice technology in employer-employee context).

We are fully aware that our proposed categorization has limits. First, it presupposes the availability of a significant amount of information about the technical set-up of a product or a service. Such information might not always be easily or publicly accessible. Second, it is not unlikely that a particular speech or voice product or service might fall in more than one category (example 1: checking medical conditions might be done by a combination of processing locally on a device, while also processing some part of the data in a networked mode; example 2: the processing of wake-up commands by Alexa, both in a local and networked mode).

We identify physical scopes of *data storage*: on a local device (typically one the user interacts with directly) or on remote servers (including but not limited to cloud storage). A separate dimension is the intended scope of access, which may include an arbitrary subset of these options: the user only, the service provider, third parties that the user specifically designates, and the general public.

The case scenarios implementing speech and language technology are numerous. For the purposes of the discussion below, we identified three specific examples, which could stand for three different categories of use cases, based on factors such as user control, parties involved in the processing activities and power and information asymmetry:

Scenario. 1: Speech diagnosis by health practitioners: In a doctor–patient relationship, speech and language technology can be used to aid in the diagnosis of particular disorders, determination of treatment and/or monitoring any progress of medication and treatment.

Scenario. 2: Online language learning service: A mobile application ("App") that provides a user with a curriculum to learn to write and/or speak a new language.

Scenario. 3: Recording of voice and speech in public places: In the last decades, cameras have emerged in public areas. Recently, some cities are experimenting with the additional registration of audio by these devices in order to fight noise pollution[2] or for public safety or policing purposes (e.g., recognition of aggression in public spaces)[3]. The usage of voice enabled cameras in public contexts is a case study of particular concern.

In addition, we also discuss some specific needs of scientific research in the public interest, in particular the need for available data (both personal and non-personal data) such as for training speech and language models.[4] Societies have become data economies with increasing

[2] https://www.vrt.be/vrtnws/nl/2021/09/24/genk/

[3] https://www.ed.nl/eindhoven/netwerk-van-hypermoderne-camera-s-op-stratumseind-in-eindhoven-gaat-politie-helpen~a1e8acee/?referrer=https%3A%2F%2Fwww.google.com%2F

[4] See e.g., EU Commission, A European Strategy for data, COM/2020/66 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066.

needs for data, for the benefit of people, organizations, economy and society progress as a whole. Specific safeguards however are needed and are moreover legally required under the European data protection legislation to protect information about identified and identifiable individuals. The usual safeguards of anonymization and pseudonymization are relevant and briefly discussed hereunder, but also the limitation thereto.

4.1.3 User control and privacy threats

User control is at the core of data protection. Individuals shall be given the choice as for the collection of additional information and any consent shall be in a granular way.[5]

While individuals are given the option to agree (opt-in) with the collection and use of additional information extraction from the speech and language application, there is a profound risk that their choice will not be taken into account, because

- the algorithmic learning models may already have information about demographics, etc.,
- the company or entity uses different labels/annotations.

The latter issue may lead the company or entity to avoid or not acknowledging that specific inferred information is processed. This may seem problematic, but in the end, it will however remain the responsibility of the company/entity to label the inferred information correctly and to respect the choice of the individual. The first issue, however, remains problematic, especially in an increasingly "connected world" with dominant players. Cross-correlation of data from different platforms requires unambiguous consent.

Additionally, users might not be able to make informed choices due to misleading phrasing and confusing interfaces fraught with dark patterns, which is already happening on large scale with cookie consent notices [1]. The companies will be tempted to use dark patterns and nudging towards privacy-decreasing choices also in case of consent notices for language and speech processing, as their business models depend on this data, just like in the case of cookies.

At the same time, user control may not be sufficient in case of privacy interferences, when applications are invading in the "private sphere", such as in use case 3. Individuals are entitled to respect privacy even in public places, and even if they would be public persons. At the same time, "privacy is a broad term, not susceptible of a definition". It encompasses a wide array of interests, including the right to personal development and to engage in relationships, to meet and to engage with other people. Individuals also have (some degree of) privacy when conducting professional activities and are entitled to protect their identity. And – also very importantly – privacy may be needed to exercise fundamental rights, including the right to free speech or to protest. Privacy is therefore inherently linked with freedom.

Any risk of applications limiting privacy shall therefore be assessed at the design phase of each and any voice, speech and text application. The concerns shall be addressed hence before development, right from the start and, for example, by using PETs or organizational measures ("privacy and data protection by design"). If this would not be sufficient, only limited exceptions to the fundamental right to privacy are possible but only in as far as necessary ("is it the last measure that can be effective, e.g., to curb public threat") and proportionate ("is it in proportion with the legitimate goal to be reached?") in democratic societies, and a sufficiently precise law is adopted to allow the interference.

[5] See Article 29 Working Party, Opinion on consent, `https://en.wikipedia.org/wiki/Article_29_Data_Protection_Working_Party`.

4.1.3.1 User privacy in speech and language technology

We draw a distinction between input data and inferred data (see above). Inferences may include characteristics of an individual that can be automatically extracted from their input data, including, but not limited to, culture, race, age, gender identity, socioeconomic status, education, marital or parental status, health information, location, emotion, and stress. Inferred data does not have to be human interpretable. A more detailed discussion on this can be found in the section on PETs.

One way for a computing system to gather information about a user is to ask them directly. In that case, the terms of use guide how these characteristics are used and shared. However, when the input data include audio, speech, and text, and these characteristics are *inferred* rather than *disclosed*, it may become less clear how or if the inferred characteristics, the inferred data, can be reused.

One path to protect the consumer's non-disclosed information is to place protections around the inference of the characteristics, for example noting that emotion or gender identity should not be inferred. This is in line with the concept of sticky policies and privacy rights management, defined as "a form of digital rights management involving licenses to personal data". These policies describe what can and cannot be done with a given data resource. However, due to the complexity of machine learning algorithms, it is difficult to enforce this.

For example, consider an application designed to teach a user to speak a foreign language. It may be advantageous to understand how gender identity, culture, age, or many other demographic factors influence the types of errors that may be observed. Therefore, the company may be incentivized to train algorithms that learn to recognize errors (e.g., mistakes made in pronunciation, grammar, or word choice) and how those errors overlap with these demographic identifiers. To do so they would collect data that includes both errors and demographic identifiers and train a system to jointly predict both errors and the demographic identifiers (see Fig. 1, left). This would result in a predictive model and a high-dimensional description of the data (see the yellow box in Fig. 1).

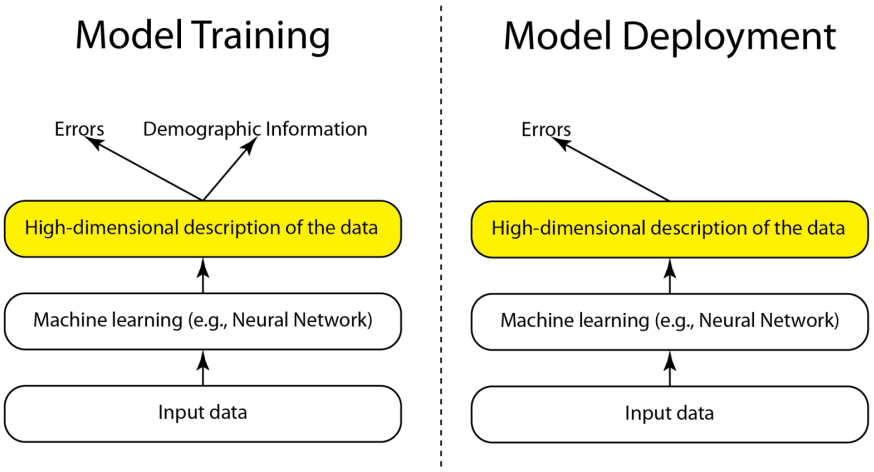


Figure 1 Model training and deployment.

When the model is deployed (see Fig. 1, right), in line with the consumer protections, it would not include the prediction of the demographic characteristics. Thus, it would not be inferring demographic characteristics because the demographic information classifier is not included. However, the same yellow embedding, the embedding that distills out the

demographic characteristics, would be generated when the model was deployed (note: this is true even when demographic information is not included as a classification target). As such, demographic characteristics would be included in the learned numeric representation of the data. These representations could then be automatically clustered (grouped) to identify similar users. Thus, although the exact information about their demographic characteristics is not known, inferences about these characteristics will be.

These inferred data have value. They can be aggregated across data sources to form detailed user profiles that may guide decision making ranging from advertising (which products should be displayed to which users, when?), insurance (who is at risk of serious, and expensive, illness?), mortgage loans (who is higher or lower risk), job hiring (who has characteristics that a company may find (un)desirable), law enforcement, and more. The question is then, what, if anything, should be done to control how these inferences are reused?

We highlight this challenge in Fig.2 using the example of a language learning app, one that takes in acoustic information and provides feedback to a user to promote the user's language mastery. We assume that the app requires audio information and the ability to extract speech-language information (note the red exclamation point in the matrix). The company would like to retain this information to improve the model's performance and the app's behavior. The company would also like to use this information to build a user profile, a mechanism that would allow the user to automatically advance through the app, given mastery. The company may desire text feedback, although this is not required. However, there are no mechanisms in place that safeguard the inference of the user's characteristics either within the functionality of the system itself or outside of the company, or organization, that has collected this information. We highlight this challenge in the matrix, using a box that notes "application of privacy regulations is unclear". We borrow inspiration for this matrix from prior work on consumer privacy nutrition labels [2].

4.1.3.2 User awareness and concerns about inferred information

As outlined above, highly sensitive information can be inferred from speech and language data: age, gender, ethnicity, geographical origin, emotional states, physical states (e.g., intoxication level), health-related information, intention to deceive [4]. Respective privacy threats can be roughly divided into impersonation and profiling. Impersonation refers to spoofing user identity, e.g., for authentication purposes, but also for spreading fake news and defamation. Profiling facilitates targeted advertising (including political marketing), but also discrimination, e.g., in language-based services such as call centers, or in job application processes. Additional privacy threats arise from language models for text and speech processing, as neural network language models can memorize the training data and reveal secrets from it. See more information in the section on possible attacks.

In user studies on privacy in smart homes, users generally express concerns about storage of their voice recordings by providers. For example, Malkin et al. [5] showed that unlimited storage of voice recordings, which is the default option for Amazon Alexa and Google Home, does not match well with users' expectation that this data should only be stored for short periods, and then deleted. At the same time, voice data was not considered to be particularly sensitive, and over 70% of participants reported that they have never had privacy concerns about their devices.

Yet, the general public seems to be poorly informed about possible inferences from text and voice processing and threats originating from these. To the best of our knowledge, Kröger et al. [3] were the first to explicitly investigate user awareness of and concerns about inferences from voice recordings. They asked a representative sample of the UK population

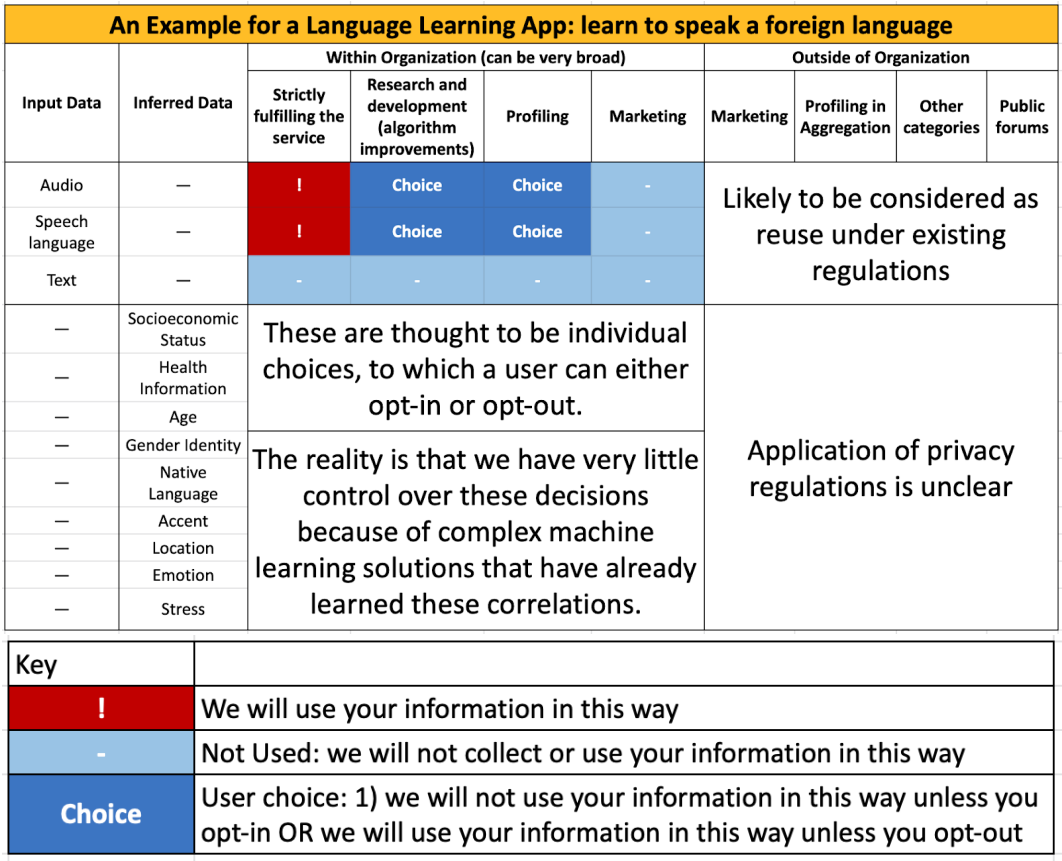


Figure 2 Example language learning app.

(n=683) to indicate how aware they are of three types of inferences: demographic data (age, gender, geographic origin), short- and medium-term states (e.g., intoxication, sleepiness, moods and emotions) as well as personal traits (mental and physical health, personality traits). Overall awareness level was quite low and depended on the inference type. Whereas awareness of the demographic inferences was the highest (almost 50% of respondents reported to be at least somewhat aware of it), only around 20% of respondents reported at least some awareness of the personal trait's inferences, with the awareness of short- und medium-term states inferences being in-between. Concern level about inferences was mixed, with around 40% of participants reporting to be concerned, and approximately the same percentage reporting to be unconcerned. When asked to justify their concern level, participants provided free-text answers that indicated, e.g., well-known privacy misconceptions such as "I've got nothing to hide" [6], a lack of knowledge about possible misuse of inferred data, but also the perception that benefits of voice-based technologies outweigh their dangers.

4.1.3.3 Moving forward

User awareness and control are very complex and subject to well-known behavioral biases. For example, Acquisti et al. [7] showed in a series of experiments that users can be manipulated towards greater information disclosure by distractions such as small delays. Furthermore, they showed that increased perceived control over the release of information also increases risky behavior, leading to higher information disclosure. As a result, awareness may have

only limited (or even adverse!) impact on safeguarding users' speech and language data. Yet, users must receive this information in a manner that is comprehensible and devoid of nudging and dark patterns. They should be able to know what happens with the data and what can be inferred. Further, regulating bodies should be made aware, or increasingly aware, of the complexities in this space. However, users' and policy makers' awareness alone will not solve the problem. We must identify additional regulations around the reuse of inferred data when these data contain personally identifiable information or otherwise personal data.

References

1. Lorrie Faith Cranor. Cookie monster. *Communications of the ACM*, 65(7):30–32, 2022.
2. Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, pages 1–9, 2009.
3. Jacob Leon Kröger, Leon Gellrich, Sebastian Pape, Saba Rebecca Brause, and Stefan Ullrich. Personal information inference from voice recordings: User awareness and privacy concerns. *Proceedings on Privacy Enhancing Technologies*, (1):6–27, 2022.
4. Jacob Leon Kröger, Otto Hans-Martin Lutz, and Philip Raschke. Privacy implications of voice and speech analysis–information disclosure by inference. In *IFIP International Summer School on Privacy and Identity Management*, pages 242–258. Springer, Cham, 2019.
5. Nathan Malkin, Joe Deatrick, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. Privacy attitudes of smart speaker users. *Proceedings on Privacy Enhancing Technologies*, (4):250–271, 2019.
6. Daniel J. Solove. I've got nothing to hide and other misunderstandings of privacy. *San Diego L. Rev.*, 44:745, 2007.
7. Alessandro Acquisti, Idris Adjerid, and Laura Brandimarte. Gone in 15 seconds: The limits of privacy transparency and control. *IEEE Security & Privacy*, 11(4):72–74, 2013.

4.2 Metrics for anonymization of unstructured datasets

Lydia Belkadi (KU Leuven, BE) lydia.belkadi@kuleuven.be
Martine De Cock (University of Washington – Tacoma, US) mdecock@uw.edu
Natasha Fernandes (Macquarie University – Sydney, AU) natasha.fernandes@mq.edu.au
Katherine Lee (Google Brain & Cornell University – Ithaca, US) katherinelee@google.com
Christina Lohr (Friedrich-Schiller-Universität – Jena, DE) christina.lohr@uni-jena.de
Andreas Nautsch (Avignon Université, FR) andreas.nautsch@univ-avignon.fr
Laurens Sion (KU Leuven, BE) laurens.sion@kuleuven.be
Natalia Tomashenko (Avignon Université, FR) natalia.tomashenko@univ-avignon.fr
Marc Tommasi (University of Lille, FR) marc.tommasi@univ-lille.fr
Peggy Valcke (KU Leuven, BE) peggy.valcke@kuleuven.be
Emmanuel Vincent (Inria – Nancy, FR) emmanuel.vincent@inria.fr

4.2.1 Introduction

Article 32 of the GDPR requires data controllers and processors to implement "appropriate technical and organizational measures to ensure a level of security appropriate to the risk". Such measures may include pseudonymization, encryption, the ability to ensure the ongoing confidentiality, integrity, availability and resilience of processing systems and

services. Reference is also made to processes for "regularly testing, assessing and evaluating the effectiveness of technical and organizational measures for ensuring the security of the processing".

Notions like confidentiality and integrity have been borrowed from the fields of security and privacy engineering, and their meaning is often hard to grasp for lawyers when implementing the law or assessing compliance of systems and applications. This underlines the need to bridge technical and legal vocabularies and methods, something which is also particularly important in the context of Article 25 of the GDPR (data-protection-by-design).

To facilitate this "translation", Data Protection Authorities have issued guidance documents specifying what data protection entails: which privacy and security goals does it intend to achieve, and what do these mean in terms of risks and metrics? One example of this are the six protection goals developed by the German Data Protection Authoritys or the Article 29 Working Party Opinion 05/2014 on Anonymisation Techniques.[6]

In the latter, the Article 29 Working Party (the predecessor of the European Data Protection Board) offers guidance on key notions on the context of pseudonymization and anonymization techniques, to assess whether they can provide appropriate privacy guarantees. It is, however, clear from the use of "database" throughout the document that Opinion 05/2014 has been written with structured data in mind. Given that text and speech are mainly unstructured data, the question arises to what extent the parameters used in existing risk assessment frameworks are still appropriate. What do notions like *singling out* or *linkability* mean in a speech and text context? Are they still relevant to capture and measure risks to privacy? What are the shortcomings? What are possible alternative notions?

The question is particularly relevant given that the GDPR applies only to the processing of personal data, defined as "any information relating to an identified or identifiable natural person ("data subject"); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person". By contrast, Recital 26 of the GDPR considers anonymous information as "information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable". To determine whether a natural person is identifiable, account should be taken of all the means *reasonably likely* to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.

4.2.2 Structured vs. unstructured data

To guide the discussion below, we first explain the difference between structured and unstructured data. Structured data is typically stored in databases where each row of the database contains the data of an individual, and such data is structured according to named

[6] https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

attributes (columns) which have a clear (identified) meaning. Unstructured data, including text, speech and images, is stored in such a way that attributes of the data are not explicitly identified aside from (possibly structured) metadata associated with the data snippet.

The language of data protection appears to assume that data is stored in a structured format, such that individuals can be identified with attributes which are explicitly stored in the data. Words associated with structured data include "records", "attributes" and "databases". Moreover, structured data assumes the concept of a data subject or individual, which corresponds to the person identified by a row of the database.

We think the reason why these notions become increasingly complex is related to the fact that there are several layers in speech/text that need legal protection: 1° the content of the speech/text (which can contain personal data that isn't necessarily exclusive to the author/speaker); 2° the identity of the author/speaker (derived from the physiological and/or behavioral characteristics of the voice or writing style); 3° other characteristics that you can derive from the voice/writing style (like gender, mental state, etc.).

In unstructured data such as speech and text, the notion of an individual or data subject is ambiguous, as it may refer to the individual who produced the data, the individuals mentioned in the data or even other individuals whose identity can be inferred through the data. In addition, attributes are not explicitly recorded but are implicitly embedded in the data. Extracting these attributes is itself an ongoing research task. Finally, the attributes used to link individuals may not be ones that can be explicitly described in human terms; for example, a machine learning system may use attributes of speech that are not necessarily explainable to a person to infer the identity of an individual.

4.2.2.1 Singling out, linkability and inference

The traditional meaning of singling out, linkability and inference has been described by the Article 29 Working Party in its Opinion 05/2014 on Anonymisation Techniques:

- *singling out* corresponds to the "possibility to isolate some or all records which identify an individual in the dataset";
- *linkability* is the ability to link, at least, two records concerning the same data subject or a group of data subjects (either in the same database or in two different databases). If an attacker can establish (e.g., by means of correlation analysis) that two records are assigned to a same group of individuals but cannot single out individuals in this group, the technique provides resistance against "singling out" but not against linkability;
- *inference* is the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes.

These definitions leave some room for interpretation, even in the case of structured data. For instance, the notion of "group of data subjects" in the definition of linkability should be restricted to groups that are small enough to "nearly" identify the subject or meaningfully make inferences about the subject as opposed to groups such as "male" or "above 25" which are huge. Most forms of data processing rely on classification/grouping hence maintaining the utility of the anonymized data requires that some grouping is allowed according to the intended usage. Similarly, reasonable inferences about anonymized data should be allowed such that the data remains useful for the intended usage.

In addition, the term "inference" has various meanings in different communities; for example, in natural language processing it typically refers to the process of logical deduction, close to the meaning implied in the above. In the statistics community, inference refers to conclusions reached based on probabilistic estimates, different from the meaning above. In

the context of this document (and particular to unstructured data) it may be the case that inferences (in the statistical sense) are required in order to draw inferences (in the meaning implied above). This confusion of terminology should be clarified.

Since speech and text data are unstructured with entangled information, the processes of singling out, linking, and inference emerge as part of one overarching process that cannot be clearly distinguished into the three processes that are separable for structured data. If only unstructured data is given, to link attributes, one needs to first infer them (to know which values should be linked); to single out one needs to use inference (to know what to filter by); thus, to single out, one needs to do the opposite of linking, thus linking methodology is implied (discarding a hypothetical association/linkage requires to assess the strength of that relation/linkability). For structured data (e.g., processed outcomes only) and mixed structured and unstructured data (e.g., annotated utterances), singling out, linking, and inference can be however different processes. It is only when one starts with nothing but unstructured data, that the three processes become part of one larger coherent process.[7]

Let's try to translate these notions to the text and speech context:

- *Singling out*: Here "individual" could refer either to the creator of the data (the speaker, the author, etc.) or to subjects explicitly referred to in the data (eg. the person who was photographed), or subjects implicitly mentioned in the data (e.g., from background noise). The data protections should apply to all.
  - Speech and text data make it possible to interrogate the data with different questions such that the filtering process of structured data becomes inherently easy using derived data from that speech and text data.
  - It is unclear at this stage how this concept can be quantified. The notion of Predicate Singling Out [1] defined for structured data is not widely agreed-upon yet, it is limited to datasets with a single recording per individual, and its extension to unstructured data raises the question of assessing singling out based on relevant attributes of the data subject as opposed to irrelevant attributes of the data (e.g., the value of the N-th sample in an audio recording) that allow records to be singled-out in the Predicate Singling Out sense but do not identify the data subject in any way.
- *Linkability*: This is the capacity to interrelate (unstructured) data sources, (unstructured) models trained on those data sources, derived (structured) information/attributes, and (even spurious) prior connections to further develop structured records concerning (groups of) individuals.
  - There is further complexity in the speech and language context due to the various interpretations of the "data subject", who could be the speaker/author, or the subject spoken about, or (in speech) a subject identifiable in the background. Any of these "data subjects" could be linkable to other data records in a data collection. Moreover, background acoustics allow for the characterization of the recording environment.
  - In the framework of the VoicePrivacy challenge[8], different empirical metrics were compared and assessed: the zero-evidence biometrics recognition assessment "ZEBRA" framework assesses expected and worst-case privacy disclosure motivated from information theory, forensic sciences, and secure communication (cryptography); the unlinkability metric targets the local and global divergence between two protected

[7] Note for machine learning experts: when structured data is involved, inference implies a deduction of unknown elements by applying patterns known from another source; this relates, e.g., to super-interpolation/reconstruction tasks and to joint-intersecting databases.

[8] `https://www.voiceprivacychallenge.org/`

biometric reference datasets when being compared to one another (designed to apply to the ISO/IEC 24745 standard on biometric information protection), and the equal-error rate (easiest explainable metric; explicitly deprecated in ISO/IEC 19795-1:2021 for performance reporting).

- *Inference* is the possibility to deduce, with even weak probability, the (not necessarily correct) value of a (structured) attribute from the values of a set of other attributes of unstructured data (speech, text, ...).
  - Here attributes are derived (implicit) from the data rather than described explicitly as in the case of structured data. It is already known that some speech and text attributes can be measured with a certain level of accuracy. Surprisingly, no attempt has yet been made to assess how well various combinations of such existing attributes identify individuals. In addition, identifying these implicit attributes in the data is still an ongoing research question. This makes it difficult to quantify protections against such inferences in the future.
  - Note that adversaries might collect lots of weakly deduced attributes under aggregated strength to shortlist automatically, and proceed then manually to single-out eventually.

The metrics mentioned in the above discussion are empirical metrics, which rely on attack models and depend on the property of the evaluation datasets. So, risk evaluation depends on the power of these attack models. It therefore depends on continuous assessment to ensure that technology progress in producing attacks is continuously monitored and risks mitigated.

The above discussion also highlights that several metrics have been proposed for singling out and linkability and that they are different, i.e., they do not always agree. While this did make sense for structured data, this does not make sense in the situation when only unstructured data is given. In such a situation, since singling out, linking, and inference are parts of a single overarching process, a single overarching metric would be desirable. Since the adversary remains unknowable, three metrics for structured data (singling-out; linking; inference) and one for unstructured data are relevant to performance reporting; for mixed data, an attacker might use structured data only to validate information gathered from unstructured data, without any attempt of linking structured data explicitly.

As an alternative to the above empirical metrics, formal privacy guarantees may also be considered for data protection impact assessment. The law does not impose one method or the other. Instead, the data controller proposes the method and the Data Protection Authority has responsibility to verify if the assessment is sufficient. Most existing formal guarantees (e.g., differential privacy, k-anonymity, etc.) were also designed with structured data in mind. Hence it is still an open question how to apply formal methods to privacy assessment in unstructured data such as speech and language.

4.2.3 Privacy disclosure: risk assessment

It is generally admitted that a Data Protection Impact Assessment should include a determination of the likelihood of a privacy breach (e.g., based on the probability of success) in addition to an impact assessment (made by human judgment) which determines the severity of impact for the data subjects concerned and the society at large. The impact cannot easily be quantified, hence technical means of assessing risk are expected to focus on quantifying the likelihood of the privacy breach. In the case of anonymization, this is the likelihood that the data subject (or a meaningful group of subjects) is re-identified according to the criteria above.

We recommend that the *likelihood* of a privacy breach should be assessed from both the worst-case and average-case perspectives. The likelihood varies depending on the data subject and the actual data. For instance, in the case of speech, some individuals may be

more easily re-identifiable than others, depending on their voice and on the spoken contents [2]. Assessing the worst case means quantifying the maximum breach probability across all subjects and all expected contents, while assessing the average case means quantifying the average breach probability across all subjects and all expected contents. These assessments may have to be done empirically rather than through a formal analysis.

This likelihood analysis should be combined with an impact assessment to determine an appropriate course of action. For example, if the severity of a breach is considered high, then data subjects with a high to moderate likelihood of breach should be removed or appropriate protections put in place to mitigate their risks. In the case of low impact breaches, it may be deemed appropriate to allow individuals with higher likelihood of risk to remain in the data collection. This judgment is expected to be made in coordination with the Data Protection Authority.

While existing methods that are widely used provide a strong support to landscape, navigate, and steer within data protection – for the new world it is to the most –, they are not mandatory by law. The GDPR framework allows for the emergence of new technologies and methodologies that aid better privacy and risk assessment. It should go without saying yet might benefit to be voiced: summaries of the status quo give an impression of how far a community got; one can go beyond.

4.2.4 Renewability revisited. Continuous countermeasure upgrading.

In the biometric information protection standard, "renewability" is defined via revocability: "revocation is required to prevent the attacker from future (or continued) unauthorized access". There, renewability is used as a security requirement. The goal is to adopt, e.g., a new cryptographic measure without needing to recapture/reacquire data. Here, we extend this notion to information protection for speech and text.

The state-of-the-art in attacking systems continually improves, and we must continue to adopt new countermeasures to unstructured data to continue to offer appropriate protections. The core idea behind the concept word "renewability" is to adopt such new measures to keep up with adversaries getting stronger over time:

- With structured data, the so-far view is to use a better encryption algorithm, enlarge the key size, etc. without needing to reveal or to recapture the unprotected data.
- For unstructured data, especially for speech, no recapture is exactly alike: repeated utterances vary, e.g., a slight change in timbre and background noise changing.
- Regarding upgrading a countermeasure for unstructured data that manipulates the data: data has had been manipulated before. An update that relies on transformation should ensure that there is no inverse function of that transform (such an inverse makes it possible to remove the update); or that there is a considerable amount of effort necessary to obtain a functional inverse. The update process should not be reversible.
- If unstructured data is not stored – processed only–, countermeasures can be uncomplicated in comparison to full data anonymization, e.g., if on-premise computing and access control are sufficient as parts of Data Protection Impact Assessment.
- Continuous upgrading suggests a regularity; not eternal waiting. Different events can be indicators to investigate upgrade strategies: a regular testing and evaluating of Article 32 of the GDPR (timespan to be defined in the Data Protection Impact Assessment); publication of related exploits; any related auditing taking place.
- There needs to be a continuing conversation with Data Protection Authorities for high-risk situations.

- Continuous assessments of the privacy risk of analysis on unstructured data is critical because we do not have formal guarantees for assessment.
- Formal methods always require a set of assumptions and scope. We can grow and improve the scope of formal protections, however, for any data as unstructured as speech and language, continuous assessment of privacy risk is highly recommended.

4.2.5 Defining the "accuracy" of algorithms

Terminology regarding AI is expected to be defined in the Proposal for the Regulation of Artificial Intelligence systems ("AI Act").[9] To aid the ability of communicating here, we describe contexts towards "accuracy". Notably, the harmonization of biometric systems as of ISO/IEC 2382-37:2017 does not define "accuracy" whereas a plethora of other metrics is defined to harmonize across the biometric standardization projects. Expert discussions in this ISO/IEC vocabulary harmonization group reached the conclusion that accuracy is not definable – other efforts were reported which attempted the definition of accuracy for over a decade, without success. In our discussions at Dagstuhl, we arrived at the perspective that legal communities (as well as natural language/text processing) operate via formal methods of philosophical reasoning using math, the classical deductive method. When following statistical paradigms, regardless of their type of data (nominal, categorical, etc.), these toolsets are less available, since uncertainty is unavoidable. One can through AI reduce uncertainty regarding associating a specific value to an attribute, but only so much so. To express remaining uncertainty is what "accuracy" implies. The ways of quantifying remaining uncertainty are shaped by the extent to which the entire process chain is covered from sensor capture to policy and governance. AI experts stop at performance reporting, leaving policy thereafter prone. When bringing the disciplines together, "accuracy" could also be reflective about how much of the remaining uncertainty disrupts policy makers to take decisions, or conversely how much information is explainable to decision makers and beneficial for their task at hand.

4.2.6 Assumptions: to promote appropriate uses of privacy methods, state assumptions clearly

Privacy methods all make assumptions about how models will be trained and deployed and how adversaries will act. They are only exhaustive to the extent as their underlying model is fully reflective of the world: countermeasures that have only gone so far but not to the extent that an adversary could go, (formal/logical) loopholes in countermeasure design are readily exploitable. However, to ensure that protections are maximized, privacy method developers must clearly state their assumptions to enable dataset creators and model developers to appropriately meet assumptions or understand the risks of relaxations of the assumptions.

References

1 Aloni Cohen and Kobbi Nissim. Towards formalizing the GDPR's notion of singling out. *Proceedings of the National Academy of Sciences*, 117(15):8344–8352, 2020.
2 George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. Technical report, DTIC Document, 1998.

[9] https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A52021PC0206

4.3 Vulnerable groups and legal considerations

Lydia Belkadi (KU Leuven, BE) lydia.belkadi@kuleuven.be
Meiko Jensen (Karlstad University, SE) meiko.jensen@kau.se
Dietrich Klakow (Saarland University – Saarbrücken, DE) dietrich.klakow@lsv.uni-saarland.de
Katherine Lee (Google Brain & Cornell University – Ithaca, US) katherinelee@google.com
Olga Ohrimenko (University of Melbourne, AU) oohrimenko@unimelb.edu.au
Jo Pierson (Free University of Brussels, BE) jo.pierson@vub.be
Emmanuel Vincent (Inria – Nancy, FR) emmanuel.vincent@inria.fr

4.3.1 Biometric systems

From a legal perspective, biometric verification (that is verifying the identity of a speaker) systems are often deemed to be not as risky as biometric identification systems (who out of a larger set of known speakers is speaking). Under data protection laws, legal scholars have discussed the definition and legal nature of biometric data. Indeed, Articles 4(14) of the GDPR and 3(13) of the Law Enforcement Directive define biometric data as "personal data resulting from specific technical processing [...] which allow or confirm the unique identification" of an individual. In particular, the definition seems to directly refer to biometric identification (i.e., "allow" the unique identification) and verification (i.e., "confirm" the unique identification) [1]. Article 9 of the GDPR further specifies that only biometric data "processed for the purpose of uniquely identifying" an individual are considered sensitive. In other words, the GDPR does not consider all processing of biometric data as sensitive and excludes verification purposes [2].

This distinction between identification and verification further permeates the risk assessment performed by the European Commission under the AI Act. This regulation aims to set out rules for the development, marketing, and use of AI systems. It further aims to steer AI uptake to reach a high level of protection of public interests (e.g., health, safety, fundamental rights). The AI Act relies on a risk-based framework spanning from unacceptable to minimal risks to support this approach. Accordingly, AI practices entailing severe risks to public interests are prohibited or more strictly regulated.

The current draft of the AI Act considers that biometric verification always entails "minimal risks", except in the context of migration, asylum and border control management. In particular, AI systems used to verify the authenticity of travel documents and check their security features are considered high-risk (Annex III). This exclusion means that providers, users, and other third parties involved in the supply chain would, in principle, not be subjected to the obligations set out in Articles 16 to 29 (e.g., taking corrective actions in case of non-conformity, information and cooperation with national competent authorities, etc.).

Furthermore, only high-risk AI practices are required to comply with a set of requirements related to the establishment of a risk management system, data governance, technical documentation, record-keeping, transparency and provision of information to users, human oversight and accuracy, robustness and cybersecurity (Articles 9 to 15). These requirements would be applicable to biometric verification systems only on a voluntary basis, through the adoption of codes of conduct (Article 69).

From a technical perspective, linking the risks of biometric verification only to the number of individuals enrolled in a database is criticizable. Indeed, risks still arise even when the database contains a single individual. First, storing biometric identifiers in the cloud as

opposed to the user's device implies that they may be more easily stolen, or that the user might be identified in a situation when they don't want to. Second, the "vocal signature" has been shown to contain a lot more information than biometric identity, which might be inferred [3]. The same risk arises with, e.g., typing patterns associated with text. Third, the boundary between verification and identification is not always clear, e.g., when a smart speaker is used by 5 members of a family, running speaker verification against 5 "vocal signatures" could qualify as a form of identification. The risk should therefore be quantified depending on the usage context, the location where the identifiers are stored, and whether the user is willing to be identified.

4.3.2 Beyond identity

Speech and text snippets are complex sources of information conveying more than (biometric) identity. For example, they may reveal speakers' emotional states or health conditions. It is not always possible to dissociate and isolate different attributes captured from individuals, entailing the collection of a wide scope of sensitive personal data. Over time, such collections may also enable the constitution of extensive (e.g., personality) profiles.

Many technical and legal distinctions may be drawn to determine the sensitivity of the collection and processing of speech and text. For example, the collection of a single instance or aggregates of emotional states would have different impacts on concerned individuals. Similarly, the use of aggregates of speech and text snippets for profiling would have distinct risks and benefits depending on the context (e.g., commercial or medical uses). Accordingly, a blanket prohibition of the extraction of specific attributes of speech and text snippets may not be desirable.

At the same time, the entanglement of different attributes within snippets raises important challenges from a legal perspective. For example, it is unclear how speech or text snippets should be defined from a legal perspective or how to apply existing legal definitions. This difficulty was well illustrated in recent legislative debates over the legal concept of "biometric data" under the upcoming AI Act. In particular, the European Parliament is discussing the opportunity to distinguish the concept of "biometric data" and "biometric-based data" to account for processing beyond biometric recognition (e.g., emotion recognition).[10]

Similarly, this entanglement implies considerable contradictions with data protection principles, such as data minimization and purpose limitation. In other words, snippets may reveal more data than is necessary for a given purpose (e.g., text and typing patterns in language processing).

The coexistence of these different attributes is important when determining the sensitivity of speech and text snippets and determining the legal basis to be used. In particular, it would require taking into account overlapping legal categories of data (e.g., data concerning health, biometric data). In turn, this overlap may mandate the performance of risk assessments that consider the complex nature of speech and text snippets, and the different attributes revealed (e.g., biometric and health attributes).

This challenge has become even more relevant after the Court of Justice of the European Union's ruling *OT v Vyriausioji tarnybinės etikos komisija*[11]. In previous years, the question to what extent data protection laws, and in particular the GDPR, offer protection against

[10] See for example the following study commissioned by the European Parliament: "Biometric Recognition and Behavioural Detection" (2021) p.96.

[11] Court of Justice of the European Union, Judgment of 1 August 2022, *(OT v Vyriausioji tarnybinės etikos komisija)*, C-184/20, ECLI:EU:C:2022:601: "*[...] Article 9(1) of Regulation 2016/679 must be interpreted as meaning that the publication, on the website of the public authority responsible for collecting and checking the content of declarations of private interests, of personal data that are liable to disclose indirectly the sexual orientation of a natural person constitutes processing of special categories of personal data, for the purpose of those provisions*

sensitive inferences (Article 9[12]) or remedies to challenge inferences or important decisions based on them (Article 22(3)) has been discussed in legal scholarship. Wachter et al., for instance, have pointed to significant shortcomings in this regard and concluded that individuals are granted little control and oversight over how their personal data is used to draw inferences about them [4]. In the ruling *OT v Vyriausioji tarnybinės etikos komisija*, the Court had the opportunity to illuminate the question whether Article 9 of the GDPR applies in the situation where special categories of personal data are not explicitly made public (more notably, in online declarations of interests by persons working in the public service as required under Lithuanian anti-corruption law), but Internet users may nevertheless infer certain sensitive information about the declarants, including their political opinions or sexual orientation. In other words, the personal data that needs to be published according to the Lithuanian anti-corruption law are not, inherently, sensitive data in the sense of the GDPR. However, it was possible to deduce from the name-specific data relating to the spouse, cohabitee or partner of the declarant certain information concerning the sex life or sexual orientation of the declarant and his or her spouse, cohabitee or partner. The question to be answered by the Court was, consequently, whether data that are capable of revealing the sexual orientation of a natural person by means of thinking (e.g., involving comparison or deduction) fall within the special categories of personal data, for the purpose of Article 9(1) of the GDPR. The Court confirmed the Advocate General's opinion from December 2021, namely that Article 9(1) must effectively be interpreted as meaning that the processing of special categories of personal data includes publishing the content of the declaration of interests on the website of the controller in question. In other words, the Court interprets the scope of Article 9 of the GDPR to include sensitive inferences, something advocated for by Wachter et al. [4].

Risk assessments may also need to be performed taking into account the impact of the processing on fundamental rights [5]. For example, under European data protection laws, controllers are obliged to carry out Data Protection Impact Assessments. Article 35 of the GDPR mandates such assessment where a type of processing is "likely to result in a high risk to the rights and freedoms of natural persons". Similarly, Article 7 of the upcoming AI Act expects the European Commission to consider the risks to individuals' fundamental rights when amending the list of high-risk AI systems. In relation to speech and language technologies, what would these obligations mean for data controllers when considering the principle of non-discrimination and the right to freedom of speech? Would new fundamental rights be necessary (e.g., right to freedom of emotions)?

4.3.3 Vulnerable groups

From a legal perspective, special attention must be given to the concept of vulnerability. Under the upcoming AI Act, vulnerability will be introduced under two key provisions. Firstly, the impact on vulnerable individuals or groups is a determining factor to qualify certain AI practices as unacceptable practices. For example, Article 5 prohibits the use of AI systems that exploit "any of the vulnerabilities" of a specific group of persons due

[12] Article 9(1) of the GDPR (previously Article 8(1) Directive 95/46) provides for the prohibition, inter alia, of processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of data concerning a natural person's sex life or sexual orientation. According to the heading of those articles, these are special categories of personal data, and such data are also categorized as "sensitive data" in recital 34 of Directive 95/46 and Recital 10 of the GDPR.

to their age, physical or mental disability when such use would distort their behavior in a manner that causes or is likely to cause physical or psychological harm. Similarly, the use by public authorities of AI systems to evaluate or classify the trustworthiness of individuals based on social behavior, known or predicted personal or personality characteristics are also prohibited, under certain conditions, when it leads to detrimental or unfavorable treatment of certain individuals or groups.

Additionally, the concept of vulnerability is also used as a factor to be assessed by the European Commission when amending the list of high-risk AI systems. Under Article 7, the European Commission needs to consider:

- the extent of harm or adverse impact of AI systems in terms of intensity and ability to affect a plurality of persons and
- whether impacted persons would be in a vulnerable position, particularly due to an imbalance of power, knowledge, economic or social circumstances, or age.

At the same time, the concepts of vulnerability and vulnerable groups in relation to language technologies raise many questions from an inter-disciplinary perspective.

When speech recognition or natural language processing systems are utilized on a broad scale, these systems at some point will interact with individuals from the so-called vulnerable groups. This broad term typically includes humans with conditions that require special consideration, both in a technical and legal dimension. We can distinguish three types of vulnerable groups of relevance here:

1. individuals with special characteristics of their voice or language,
2. individuals that are not themselves able to utilize their human rights, and
3. individuals that belong to discriminated groups due to special personal characteristics like sexual orientation, ethnicity, or religious or political position.

In the first group, people with speaking issues like stuttering, aphonia, or amnesic aphasia clearly become relevant. The so-called "Doddington zoo" effect [6] also means that some people's voices are more easily identifiable than others for reasons that cannot be traced back to a specific characteristic. As discussed previously, AI-based speech recognition works with training based on a large set of speech examples, which may or may not have contained people with these specific conditions. If present, the trained AI might be able to cope with (and hide) the specific type of speech characteristics, but if the training dataset did not contain such examples, it might work less well when confronted with speech or language examples from such individuals. Hence, one challenge lies in the proper and non-biased selection of training data, as inclusion of all possible speech- or language-specific abnormalities in the training dataset tends to raise discriminatory real-world issues in itself. As an example, consider an advertisement explicitly asking for stutterers to join a training dataset recording. The resulting dataset would be biased towards favoring stutterers to other speech issues, and the real-world discriminatory effects of such an advertisement could be socially challenging as well.

The second group requires close attention, especially from the legal point of view. Transfer of self-responsibility to another human is a severe and highly sensitive issue, and should only be done in cases that have no alternative. Children are especially vulnerable in this case, as they cannot oversee the consequences of their actions sufficiently, so their parents or legal guardians have to approve decisions or even make decisions themselves for the children. In terms of speech-based interaction technology, this dependency of a child towards its custodian makes the former especially vulnerable, as audio surveillance of sleeping babies is a common and mostly socially accepted scenario. However, this raises a lot of open issues when it comes to questions of secondary use of the voice data created by children, e.g., towards advertising or psychological analysis by third parties – especially in the long term, when these children grow up to be adults of the same personality.

Another example of the second group type is people with diseases like dementia or mental disorders. Even if these may at some point decide to e.g., utilize smart speakers in their homes, or consent to having their language in a social media chat app get analyzed by a research institution, this decision may not stay aware to them. Hence, subsequently, when confronted with the ongoing voice surveillance of the smart speaker, or receiving the feedback from the research institutions, such individuals may suffer from severe trauma. On the other hand, availability of such technical surveillance or assistance systems might be very beneficial towards these individuals, especially for those also suffering from physical deficiencies like inability to type or utilize other input devices for a computer.

The third group is special in a large variety of possible ways, ranging from sexual orientations that are considered illegal in some countries of the world to social discrimination or even physical frays based on skin color, nationality, or political opinions expressed. In all of those cases, speech and language processing systems to some extent may be able to identify such conditions, based on what was said or how it was said in specific contexts (e.g., lie detection when confronted directly).

In general, belonging to a vulnerable group is no explicit act, and the definitions of what substantiates a vulnerable group differ largely.

What is common to them is that speech and language processing systems have to be designed in a way that they are either reliably agnostic to these conditions or consider them appropriately in the design and behavior of the system in consideration. Here, privacy-enhancing technologies may help, and should be considered wherever possible.

4.3.4 Confidentiality vs. duty to rescue

In some situations, the users' right to privacy may conflict with the voice technology company's legal requirements. For example, if the voice technology company collects speech or text data suggesting that a crime (e.g., child abuse) or a life-threatening danger (e.g., heart attack) has taken place, should it report it to the relevant authority, thereby violating the user's privacy? Is it enough to report cases that have been incidentally found or should the company be required to automatically analyze the data to find all possible cases and have them screened by a human operator, which is a form of systematic surveillance? When answering these questions, it is important to realize that legal requirements regarding "duty to rescue" vary from one jurisdiction to another.[13] In most jurisdictions under civil law (Europe, Latin America) and in some US states, it is a legal duty for citizens to assist in such cases unless this would put them in danger, with some exceptions (e.g., if the citizen is a priest or a lawyer hearing a person confess a crime, the confidentiality obligation is stronger). The duty to rescue does not apply to companies in these jurisdictions nor to citizens or companies in other jurisdictions, which implies that such cases can be reported but it's not an obligation. Nevertheless, some companies have been requested by law enforcement agencies to automatically screen for, e.g., child pornography in personal image data. This raises three open questions. From a societal point of view, should companies be requested, allowed, or forbidden to perform large-scale automatic screening in the speech and text data they collect? If this is requested or allowed, what should be the territorial extent (e.g., would it apply to a European company processing data from an American citizen) and which legal safeguards should be put in place to preserve fundamental human rights regarding censorship (what can or cannot be uploaded) and massive surveillance? Also, from a technical point of view, could this screening be performed on-device in a privacy-preserving way?

[13] https://en.wikipedia.org/wiki/Duty_to_rescue

References

1 Catherine Jasserand. Legal nature of biometric data: From generic personal data to sensitive data. *European Data Protection Law Review*, 2(3):304, 2016.
2 Els Kindt. Having yes, using no? About the new legal regime for biometric data. *Computer Law & Security Review*, 34(3):523–538, 2018.
3 Desh Raj, David Snyder, Daniel Povey, and Sanjeev Khudanpur. Probing the information encoded in x-vectors. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 726–733. IEEE, 2019.
4 Sandra Wachter and Brent Mittelstadt. A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. *Columbia Business Law Review*, 2019(2), 2019.
5 Dara Hallinan and Nicholas Martin. Fundamental rights, the normative keystone of DPIA. *European Data Protection Law Review*, 6(3):178–193, 2020.
6 George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. Technical report, DTIC Document, 1998.

4.4 Privacy attacks

Abdullah Elbi (KU Leuven, BE) abdullah.elbi@kuleuven.be
Anna Leschanowsky (Fraunhofer IIS – Erlangen, DE) anna.leschanowsky@iis-extern. fraunhofer.de
Pierre Lison (Norsk Regnesentral – Oslo, NO) plison@nr.no
Andreas Nautsch (Avignon Université, FR) andreas.nautsch@univ-avignon.fr
Olga Ohrimenko (University of Melbourne, AU) oohrimenko@unimelb.edu.au
Laurens Sion (KU Leuven, BE) laurens.sion@kuleuven.be
Marc Tommasi (University of Lille, FR) marc.tommasi@univ-lille.fr

4.4.1 Context and motivation

One way to assess the strength of privacy-enhancing techniques (and the data protection they provide) is to conduct so-called *privacy attacks*. In our context, a privacy attack is a process which, given a particular input or model, seeks to uncover personal data that should be or should have been concealed. Privacy attacks can be employed as part of privacy risk assessments (including Data Protection Impact Assessments) or as an evaluation method in the development of privacy-enhancing techniques.

It is, however, important to stress that privacy attacks can usually only provide lower bounds when it comes to assessing the privacy risk associated with a given output or model. Privacy attacks are by construction not exhaustive and can only explore a limited region of the risk space. In other words, they can only demonstrate the presence of a privacy risk and not their absence. Although we can make assumptions about possible attackers and the background knowledge those attackers may have access to, those assumptions may very well turn out to be invalid. Attackers may also rely on other attack strategies than the ones that have been explicitly tested.

Although the present section focuses specifically on privacy attacks (i.e., attacks designed to uncover personal data), it is worth noting that security attacks (i.e., attacks targeting the confidentiality, integrity, or availability of an IT system) may also lead to privacy breaches. In particular, it has been shown that one can infer the hidden values of a black-box machine

learning model based on requests made on this model, but we consider such attacks as being primarily a security issue, although they might also lead to privacy risks. Another example is poisoning a model, which could lead to adverse decisions or inferences being made regarding another individual.

4.4.2 Core concepts

Let us assume a function f used to transform some raw data D into another data or model D', as illustrated in Fig. 3. This transformation may correspond to a sanitization/anonymization process or to the training of a machine learning model. Examples of such data include:

- raw speech recordings (prior to speech anonymization),
- initial text documents (prior to text sanitization),
- documents or speech recordings employed to train a machine learning/natural language processing model.

The result D' of the transformation $f(D)$ may take a variety of forms. It may correspond to a "sanitized" or anonymized version of D, but may also correspond to a trained model. We assume that both the transformation f and the outcome D' are known to the attacker (but not the personal data D).

We also assume that the raw data D contains (potentially sensitive) personal data. This ranges from data explicitly stated in the raw data (e.g., phrases or spoken words, location) to implicit data that can be learnt from raw data (e.g., speaker identity, author, emotional state, age, country of origin, gender, background noise including third parties) and metadata. This information may also be sensitive and belong to "special categories" of personal data in the GDPR such as health data, religious beliefs, or ethnic origin.

The goal of the attacker is then to infer some personal data, based on the observation of the transformed data/model D', function f, and background knowledge K.

The attacker can target different stages of the data processing pipeline, including:

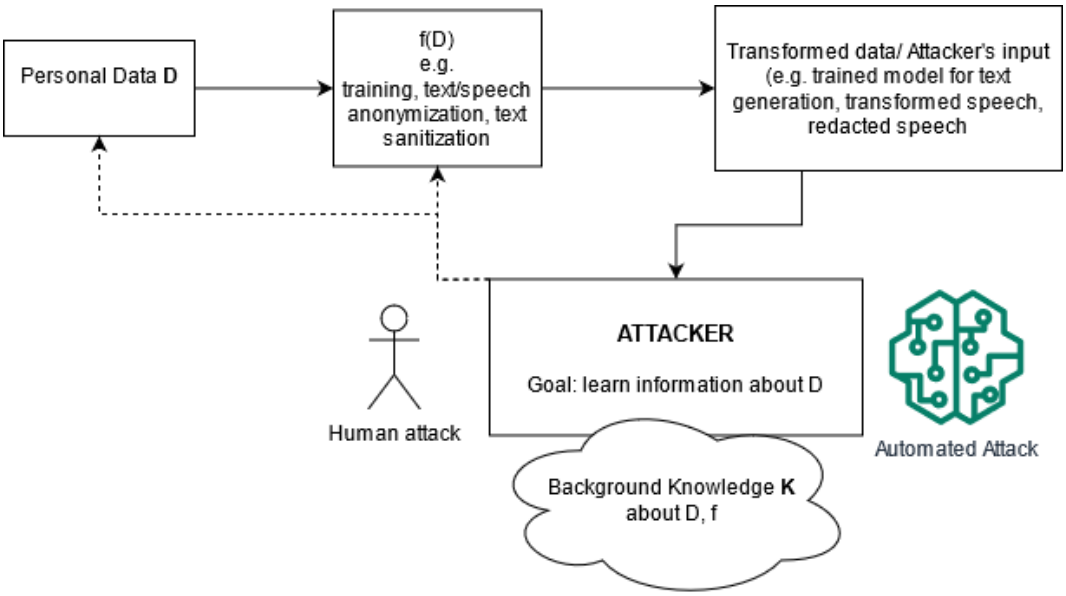


Figure 3 Core concepts of privacy attacks.

- the *training phase*, where the model is trained based on a training dataset which may include personal data,
- the *deployment/operational phase*, where the model is deployed to achieve identified objectives (for example, verifying the identity of individuals in the case of biometric verification).

4.4.2.1 Attack goal

The attacker may seek to retrieve various types of personal data, such as:
- the identity of the speaker (or author of a document),
- the identity of a person mentioned in an audio recording or text document (who may be distinct from the speakers/authors),
- (potentially sensitive) personal attributes associated with those individuals, such as their gender, age, emotional state, or country of origin,
- Whether a person is included/described in the raw dataset D.

Various mechanisms have been developed to conduct and demonstrate such types of attacks.

4.4.2.2 Attacker role

The attacker role considers whether the attacker has a privileged role in or access to the system. It distinguishes between attackers that are outsiders without any type of access and attackers that are involved in the data processing as a (trusted) third party, or attackers that have insider access to the system. Insider access can be the result of a consumer-service provider relationship (e.g., a voice assistant company storing and analyzing user data), an interpersonal relationship (e.g., a spouse that knows a user's passwords), an authoritative relationship (e.g., an employer or school), and possibly other relationships as mechanisms.

The nature of the attacker role may therefore have an impact on the attacker capabilities, the attacker's knowledge on the system, or the attacker's background knowledge we describe below.

4.4.2.3 Attacker capabilities

The attacker's capabilities can be different while having the same intention. We distinguish between those that can manipulate the raw data, or raw data transformation to those that cannot. For example, the former can intervene by contributing raw data into the processes (e.g., inject data) or manipulate the data transformation process (e.g., affecting the training code to memorize the data).

Attacker capabilities should also take into account the amount of computational resources the attacker has access to. Greater computational resources can increase the threat an attacker poses.

4.4.2.4 System knowledge of the attacker

Different attackers may have different degrees of knowledge about the system of interest. Some attackers may not have any information about the system at all and only have access to the outputs of the system. Others may have full access to the system and its source code and can leverage this information as part of their attacks. In the case of an attack in the training phase of a model using federated learning, the attacker may have access to a set of gradients, a set of models or a set of training losses.

4.4.2.5 Background knowledge of the attacker

Attacker models typically rely on assumptions regarding the information or tools that may be available to an adversary seeking to uncover the information that should be protected in the research or deployment phases of the machine learning tools. For instance, an adversary seeking to determine the identity of the speaker of a given speech segment may be assumed to have access to audio recordings of various potential candidates. Similarly, an adversary seeking to find out the identity of a person mentioned in a sanitized text will typically have access to public information sources (like web data) about potential individuals. This background knowledge may take various forms, and may also correspond to machine learning models or computer software. Attack models should take into account as much background knowledge as possible to ensure the attacks are sufficiently strong.

4.4.2.6 Attack mechanisms

Privacy attacks can be implemented through a range of possible techniques. Attacks can be conducted through automated inferences, and/or by humans seeking to uncover some unintended information based on various knowledge sources. For example, the attacker can train a model on an auxiliary dataset, resembling the structure of the model it is attacking. The adversary may also have access to multiple versions of the model (e.g., original and updated/fine-tuned model) and try to extract information by accessing the two in parallel. For such automated attacks, one also needs to make assumptions regarding the computational power that we expect to be available to a motivated attacker (see attacker capabilities).

4.4.2.7 Who conducts the attack and how it is evaluated

A final dimension to consider is who is practically responsible for designing and conducting the attack. This can be the organization developing the system, a third party, or a data protection authority. Privacy attacks also have a success rate that varies depending on attributes or personal or protected information of an individual (e.g., members of a vulnerable group maybe more susceptible than others). To this end, privacy attacks should be carried out on wide data distribution.

4.4.3 Concrete example attacks

We list the following example attacks and information they can extract. Some of these attacks should be adapted to speech and text (e.g., what does membership mean in this context) as they were first proposed on tabular data.

- *Membership inference attack*: The adversary tries to find out whether a certain data record (e.g., a data record can be a piece of text contributed by a user) was present in the personal data fed to the transformation function f [1]. In relation to speech or text, this may correspond to whether a certain person contributed to the corpus with text or voice that the attacker holds and where participation in the dataset could be considered personal or sensitive (e.g., written descriptions of patients' medical condition). A more general attack could be a *presence attack* where the attacker tries to infer if a given person has contributed to the dataset from examples of his voice or his writings whether or not these examples are present in the dataset. These attacks for text models can be used to see if user's data was used in training..
- *Re-identification attack*: The adversary attempts to determine an individual's identity. For example, in the case of transformed or anonymized text whether an attacker can

determine the identity of the person or narrow down to a small enough sample of users. This is an instance of attribute inference where the attribute is an identifier of a person.
- *Attribute inference*: Given a trained model as the output of the transformation function f, and some information about non-sensitive attributes of the individual whose privacy we want to attack, the adversary tries to reconstruct the sensitive attributes. Applicability of such an attack to speech and language data is questionable due to 1) the non-tabular nature of the data, and 2) the potential lack of agreement on what sensitive attributes of text and speech are, also taking into an account their availability in the training data (e.g., labeling such attributes beyond those "easily" determinable, such as age or occupation).
- *Data extraction (aka model inversion attack)*: The adversary tries to extract verbatim training data from a trained model [2, 3].
- *Update data extraction*: The adversary tries to extract the data that was used to update or fine-tune a model based on interaction with several models [4].

Liu et al. [5] provide a survey of privacy attacks on machine learning models.

4.4.4 Open questions and challenges

As mentioned earlier, privacy attacks can only explore a limited number of possible attacks. Due to this empirical nature of the evaluation, we believe that is important to vary the types of attackers, their knowledge, and to develop metrics and guarantees that can be given based on the attack success rate, for example, in terms of confidence intervals.

It remains unclear how to conduct a privacy attack on arbitrary types of language data, in particular for the case when the person to protect is not the author of the speech recording or text document, but corresponds to a third-party mentioned in those. An interesting strategy would be to use reinforcement learning or a GAN-inspired framework to help the attack model learn if personal data is leaked through the transformation f. As re-identification often proceeds in a sequence of reasoning steps, the use of "chain of thought" prompting [6] also constitutes a promising approach to re-identify individuals from speech or text data.

Most automated attack mechanisms have been developed by researchers. We believe creating a framework for automatically generating attacks would help streamline the process and potentially identify new attacks. This framework could for instance take the form of an open-source library of attacks/implementation, structured according to the attack mechanism, analogously to MITRE's CAPEC[14]. Alternatively, one could also make available a standard API to test your model against privacy attacks.

References

1 Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (S&P), pages 3-8, 2017.

2 Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. Machine learning models that remember too much. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 587–601, 2017.

3 Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.

4 Santiago Zanella-Béguelin, Lukas Wutschitz, Shruti Tople, Victor Rühle, Andrew Paverd, Olga Ohrimenko, Boris Köpf, and Marc Brockschmidt. Analyzing information leakage of

[14] Common Attack Pattern Enumeration and Classification, MITRE, `https://capec.mitre.org`

updates to natural language models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 363–375, 2020.

5 Chao Liu, Xin Xia, David Lo, Cuiyun Gao, Xiaohu Yang, and John Grundy. Opportunities and challenges in code search tools. *ACM Computing Surveys (CSUR)*, 54(9):1–40, 2021.

6 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

4.5 Privacy enhancing technologies

Martine De Cock (University of Washington – Tacoma, US) mdecock@uw.edu
Zekeriya Erkin (TU Delft, NL) z.erkin@tudelft.nl
Simone Fischer-Hübner (Karlstad University, SE) simone.fischer-huebner@kau.se
Meiko Jensen (Karlstad University, SE) meiko.jensen@kau.se
Dietrich Klakow (Saarland University – Saarbrücken, DE) dietrich.klakow@lsv.uni-saarland.de
Francisco Teixeira (Instituto Superior Técnico – Lisbon, PT) francisco.s.teixeira@tecnico.ulisboa.pt)

Privacy-enhancing technologies (PETs) provide technical building blocks for achieving privacy by design and can be defined as technologies that embody fundamental data protection goals [13] including the goals of unlinkability, interveneability, transparency and the classical CIA (confidentiality, integrity, availability) security goals by minimizing personal data collection and use, maximizing data security, and empowering individuals.

The privacy by design principle of a positive sum for speech and language technologies should enable users to benefit from the rich functions of these technologies while protecting the users' privacy at the same time. The fundamental question is how to achieve privacy by design for speech and language technology without hampering the services. To achieve this goal, different PETs exist that can be utilized for this purpose. Below, we first discuss what type of personal data are accessible via speech and text and should be the target of protection by PETs. Then, we provide an overview of PETs that can provide protection and discuss their limitations and challenges that arise when used for speech and language technologies.

4.5.1 Possible private content

When discussing PETs' capabilities and limitations, the first important dimension is the possible private content contained within speech and text. The different steps in the processing pipeline and the different levels of information contained might make the use of different PETs necessary.

For this we can look at the standard processing pipeline of a speech signal. We start by recording speech with one or more microphones; this is usually followed by some form of audio pre-processing, such as denoising, enhancement or separation. The life cycle of the speech signal then becomes dependent on the target application. If the target application is related with purely acoustic characteristics of the signal, speech can be fed directly into the

corresponding processing pipeline. If the end goal is to analyse and process text, then speech must first be transcribed either by automatic or manual means. Of course, text can also be produced by other means and will also be considered here.

Below we provide a summarized categorization of the levels of information contained within speech and text, to allow for a more detailed analysis on the necessity and application of PETs in the context of the processing of speech and text.

4.5.1.1 Speech

Being both a means for communication and a biometric signal, speech is formed by multiple dimensions, each containing different levels and types of information.

From a technological point of view, to analyze speech, we need to first record it. This means that a recording will contain voice, environmental acoustics, and information about the recording setup.

By voice, we mean any acoustic phenomenon produced by an individual's vocal tract, which can include verbal and non-verbal dimensions. The verbal dimension corresponds to lexical or phonetic content, that has a direct correspondence to language and communication. The non-verbal dimension can be further divided in two sub-dimensions, one that is non-verbal but communicative, and a second that is non-verbal and non-communicative. More formally, we within voice we have:

- Lexical information – also called verbal [3] or linguistic content [30], represents phonetic content associated with language that is meant to be communicative.
- Paralinguistic information [22] – acoustic factors related with communication which provide, consciously or otherwise, additional information to the listeners (e.g., emotion and intent, accent).
- Extra-linguistic Information[22] – acoustic factors not directly related to communication, corresponding to speaker characteristics [3] which arise from mental, biological, and physical traits of the speaker, such as the speaker's age, gender and health status. This information may also be characterized as paralinguistic [30].

Environmental acoustics include other background sounds, which may provide information about the speaker's location, surroundings, and communication context. Moreover, the type(s) of microphone(s) or device(s) used to create the recording will also be reflected in the recording itself, making it possible to even identify the specific device used to create it.

4.5.1.2 Text

Text can contain private information irrespective of whether it is produced from speech or typed directly. Possible sources for text from speech can be transcription tasks, proper dialog systems or any human-to-human communication that is transcribed after the prime use case (e.g., speech messages sent over WhatsApp). Possible sources of text are medical reports, legal course rulings, messages or e-mails.

Information derived from writing style is given away involuntarily. Here are some examples:

- Gender is often given away by the use of short function words e.g., the usage of "would" and "may" [12].
- Age: like with gender, word usage can give away the age of the writer. This could also be captured in a very single language model [12].
- Mother tongue: some studies have shown that the mother tongue is given away e.g., by the usage (or non-usage) of determiners.

- Literacy: statistical features measured by stylometric measurements are also detectable and can be grouped by
  - Length-based measurements: average length of words, sentence length in words and characters;
  - N-grams: the frequency of groups of two, three or more words;
  - Lexical richness measurements: values scored by frequency of the vocabulary and all occurred tokens (words and punctuations), in detailed, e.g., Giraud's R, Herdan's C, Dugast's k, Maas' a^2, Tuldava's LN, Brunet's W, Summer's S, Horoné's H, Sichel's S, Michéa's M, Entropy, Yule's K, Simpson's D, and Herdan's V_m;
  - Readability and formality scores: values scored by syllables, frequency of often-used words, counts of numeral words given by, e.g., Flesch's reading ease formula [9], Flesch-Kincaid formula [10], Dale-Chall scored by a list of 3000 words [7], McLaughlin's SMOG formula [24], FORCAST formula (US military, 1973), Gunning fog index [11], Automated readability index (ARI) [31], or Heylighen formality score [15];
  - Syntactical features: measured by scores depending on dependency and/or constituency grammar parsers, e.g., occurrences of elements from parse trees;
  - Semantic features: scores measured in connection with word wise request on semantic networks, e.g., WordNet, GermaNet (only for German language) or biomedical terminology networks, e.g., UMLS and/or SNOMED CT.
- Information derived from semantics: the prime purpose of text is to convey an explicit message. The open problem is to decide what the core content of a text is and which part should be protected. For practical purposes so-called named entities are often used as proxies. Those are person names, organization names or locations. Often dates or monetary amounts are also detected by the same software.

4.5.2 Training or inference

In machine learning applications for speech and language technology, personal data is typically used in two different stages.

Training. During the model development phase, a model is induced from training instances. The training instances consist of personal speech and/or text data from users that may be labeled (supervised learning) or not (unsupervised learning). One may need to protect:

- the training instances themselves [14] (this is sometimes referred to as input privacy),
- the resulting trained model, as the model itself can leak information about the training instances [4] (output privacy).

Furthermore, one may also want to protect model integrity, for instance to prevent malicious actors from poisoning the training data to deliberately hinder the model's performance or to make it more vulnerable to attaks [18].

Inference. After a model is trained, it can be deployed and used in the inference phase during which a trained model is used to deliver a service, such as emotion recognition from speech. One may need to protect [28, 34]:

- the model that is used to make the inferences – such as the emotion recognition model – as the model itself can leak information about the training instances (input privacy),
- the query instance – such as the snippet of speech to be classified – as it reveals information about the speaker (input privacy),
- the result of the inference – such as the inferred emotion – as it leaks information about the model and the query instance (output privacy).

Beyond preventing leakage of personal user data from the trained model, one may also want to protect the intellectual property (IP) of the owner/developer of the model. The latter

is particularly relevant when the trained model constitutes a competitive advantage, or in security applications such as spam or hate speech detection, where knowledge of the model would help adversaries to develop strategies for evading detection.

Orthogonal to the above, machine learning is also used for privacy-preserving release of speech and text data, in particular to recognize and mask personal data in text documents (such as court cases and medical records) and in speech signals.[15]

4.5.3 Training/inference locally, centrally or collaboratively

A relevant issue for both the training and inference phases is how the data is distributed, and where and by whom it is processed.

Training of machine learning models is traditionally done in a *central* way, assuming that one entity has access to all the training instances and uses it to induce a model. In many applications however, the data naturally originates from multiple entities who may be capable of each training their own model *locally*, or who may want to *collaborate* in training a joint model however without sending their raw data to each other or to a central entity.

Inference, namely classifying or scoring a new query instance with a trained machine learning model, inherently involves two entities, namely the model owner and the query instance owner (a.k.a. the user). Inference can be done *locally* by the model owner (which requires the user's query to be sent to the model owner); or locally by the user (which requires the model to be deployed on the user's end); or jointly by both in a *collaborative* manner. It can also be *outsourced* to yet another entity in the cloud.

4.5.4 The privacy enhancing techniques

Hoepman [16] introduced eight strategies for privacy: minimize, hide, separate, aggregate, inform, control, enforce and demonstrate. In the following, we use these strategies as a means for classifying existing PETs for speech and language technologies.

- "Minimize": This strategy enforces the basic privacy principle of data minimization by limiting the amount of personal data that is processed, e.g. by avoiding data collection from the start of by using anonymization or pseudonymization techniques.
  - Pseudonymization can be used to render text pseudonymous by replacing identifiers with pseudonyms. Text pseudonymization is often used in the context of clinical text de-identification, where for instance patient name are to be replaced by pseudonyms (see e.g. [26] for clinical text pseudonymization based on deep learning). Another example for an automated tool that recognizes and pseudonymizes privacy-relevant text parts in private communication (email) is provided by [8].
  - Typical anonymization techniques use data generalization or suppression (e.g. k-anonymization or variants) or data perturbation by adding statistical noise to aggregated data (e.g., differential privacy).
  - For text processing, k-anonymity has a number of applications. Summarizing text while hiding identifying attributes is a simple approach used in practice. In natural language processing, there are extensions for k-anonymity, e.g., t-plausibility[2] for generalizing words to desensitize text or c-sanitized [29], an information theoretic approach based on finding sensitive terms that have a high mutual information with an entity. These terms can be taken alone or in combination. After, they can be redacted or replaced with non-sensitive substitutes.

[15] https://www.voiceprivacychallenge.org/

  - Differential privacy (DP) is a data perturbation technique that can be applied locally on the user's device (local DP) or centrally for federated learning. Local DP approaches exist that are rewriting a text until a certain privacy guarantee (epsilon) is reached (see, e.g., [17]).
  - Further minimization methods that are specific to speech include voice anonymization techniques (see, e.g., [6, 35]).
  - A systematic review of deep learning methods for privacy-preserving natural language processing, including data minimization PETs categorized into de-identification (corresponding to pseudonymization) methods, anonymization methods and differential privacy methods, is provided by [32].
- "Hide": This strategy aims at disallowing access to personal data in plain view or restricting access to only authorized users, meaning that precautions such as transmitting data securely and storing data in the encrypted form should be considered. Furthermore, it would also be difficult for adversaries to observe the data flow by deploying mix-nets, onion routing and similar technologies. Cryptographic tools and techniques such as homomorphic encryption (HE – protecting input data and the result of computation) or functional encryption (FE – which in contract to HE only protects the input data while the computation result is made available in cleartext), as well as secure multiparty computation (MPC) and secure distance-preserving hashing (DPH) or randomized binarization also provide data hiding.
- "Separate": This strategy includes data or process separation by processing data including metadata in a distributed manner if possible. The data should be kept in a separate form: tables and datasets should be divided into different tables and if possible different datasets and locations.
  - A good example of data separation is using additive secret sharing, a.k.a. multi-party computation, where data is split into random parts and each part is kept in different servers. In this way, the attacker cannot deduce any meaningful information about the data unless she has a certain number of the shares. Secret data shares can still be processed using MPC techniques.
  - Federated learning (FL), which is used for many natural language processing applications, is also based on separation, but is not sufficiently privacy-preserving, as still information can leak. For this reason, FL is more and more used in combination with other PETs, such as DP and HE or MPC.
  - Slicing is another example for enhancing speech privacy via separation. After speech is anonymized, the speech signal is split into small chunks when storing the data such that it becomes harder for an adversary to re-identify the original speaker [23].
- "Aggregate": This strategy is applying data aggregation for concealing data. It is the step where the amount of personal information is restricted at a group level. By doing so, it is difficult for the adversary to identify a certain individual within the group. Data aggregation via statistics or building machine learning models is however not sufficiently protecting data, as personal information could leak via inference attacks, i.e., via correlation of statistics, and therefore complementary privacy-protecting measures such as DP are needed. To some extent, the use of synthetic data instead of data taken from real human individuals also falls into this category, as the typical characteristics of the real-world data are derived, in aggregated form, and utilized to generate randomized new data that follows the same statistical distribution. Hence, on an aggregated level, the original information on characteristics of the real-world data is still contained in the synthetic data, just on an aggregated level.

- "Inform": This strategy aims at providing transparency to data subjects when their data are processed. Different types of Transparency enhancing tools (TETs) exist.
  - Ex ante TETs can provide transparency before personal data is disclosed/processed for enabling informed decisions/consent. Privacy product labels on packages, e.g., for IoT (Internet of Things) or natural language processing products, can for instance inform customers about privacy practices when products are purchased (see also [19, 27]). Concepts for usable policies, e.g. as part of consent forms, multi-layered policies, automated policy assistants exist in general but have not specifically been applied for speech and natural language processing. Ex ante TETs explaining how AI and automated decision making is in principle working is still an area of research.
  - Ex post TETs are providing transparency about how data have been processed. Also ex post TETs for explainable AI, which are explaining how decisions have been made, e.g., via attention maps or feature selection, are a current research area. Other general ex post TETs include privacy dashboards for displaying what data has been processed by whom, for tracking data usages and flows or for data export, as well as privacy notification tools (e.g., informing about breaches or risks) (see [25] for an overview).
- "Control": This strategy should provide data subjects with control over their data. Different intervenability tools have been researched or developed for enhancing control, even though they have been hardly applied to speech and natural language processing yet. These tools include for instance:
  - privacy policy tools/languages exist for negotiating privacy policies between data subjects and data controllers,
  - privacy dashboards which allow data subjects to conduct or request (from the data controllers) data deletions or corrections or data export for data portability,
  - tools for data subjects to object to automated decision making,
  - consent management tools which enable data subjects to provide and to easily revoke any time informed consent.
- "Enforce": This strategy should guarantee that a privacy policy that complies with data protection/privacy laws is enforced. Access control systems, e.g., based on attribute or role-based access control, can provide means for technically enforcing privacy policies.
- "Demonstrate": This strategy requires the controller to demonstrate compliance and enable accountability. General approaches and tools for supporting this strategy include:
  - logging and auditing tools and privacy intrusion detection systems which allow to detect privacy breaches as a means for making attackers accountable and to demonstrate compliance,
  - privacy certification of PETs in regard to their privacy functionality and assurance by independent certification bodies as a means for demonstrating compliance – the certification results could be (certified) privacy seals (e.g., EuroPrise seal[16]) that mediate privacy levels of products to users/customers,
  - consent management tools which allow data controllers to easily keep proofs for the data subjects that consent has been provided.

4.5.5 Challenges for PETs

In general, PETs are very valuable tools for implementing privacy into speech and language technology. However, there also are several challenges to consider in the design phase:

[16] https://www.euprivacyseal.com/EPS-en/Home

- Difficulty to determine all personal attributes: it is hard to guarantee that all possible different types of personal information from an input are removed or protected by data hiding or access control, because we simply don't know what information might actually be contained in an audio or voice snippet.
- Privacy-performance tradeoffs: PETs, especially homomorphic encryption or MPC, may have severe efficiency tradeoffs, which may not be acceptable if fast responses are needed.
- Privacy-utility tradeoffs: data minimization, aggregation, and data hiding may conceal information or use data perturbation by adding statistical noise, which creates utility tradeoffs.
- Tradeoffs between privacy protection goals: for machine learning models, there is also a tradeoff between protection against member inference attacks and fairness of decision making [5]. Moreover, data minimization and transparency are privacy protection goals that are typically in conflict with each other.
- Usability: PETs based on "crypto-magic" operations are often counterintuitive to users and hard to comprehend and trust.
- The privacy guarantees of PETs rely on an attacker model and its assumptions: for instance, secret sharing and MPC require multiple non-colluding entities that act as independent organizations.

In general, it is a challenge to select and configure the right combination of PETs for addressing privacy trade-offs mentioned above, for sufficiently protecting all personal data and metadata items, and for fulfilling legal requirements posed by the GDPR and its privacy by default and by design principle, the AI act or other regulations.

Beyond these general issues, each specific technique or strategy implementation approach has its very unique challenges when applied to speech and language data, some of which are listed next:

- *Minimize*:
  - K-anonymity is not straightforward to apply on speech data. For example, t-plausibility [2] has strong assumptions that may not be realistic, as they lead to a too high utility loss.
  - For redacting text, deployment heavily relies on the type of the data: while it is possible to redact text for court cases, it might cause problems to do so for medical data. For c-sanitized, the computation of mutual information is not easy and for instance relies on google searches. For text redaction in eHealth application, a tradeoff between privacy and safety and accountability needs to be taken into consideration (see [1]).
  - For redacting speech, deployment may rely on manual redaction, or on auxiliary technologies that allow the transcription of the speech signal, or that perform keyword spotting.
  - Local DP, which is used for approaches of text re-writing, usually requires a high epsilon for still retaining sufficient data utility, and thus cannot provide very good privacy guarantees. For central DP applied for federated learning, there is no confidentiality of the data against the central aggregator that needs to be trusted. (see also discussion of limitation of DP for natural language processing in [21]).
- *Hide:*
  - The interactive nature of MPC might require bandwidth requirements and involvement of the parties to be online for processing. Particularly, in the case of malicious security model, the overhead introduced will be a significant performance burden overall. While this may be acceptable for training of MDATAL models, it can be problematic in inference applications where the responsiveness of the system (i.e., short response

times) matters. MPC does not inherently require adaptations to the system, meaning that for already trained models, no utility loss is introduced.
  - For HE computational burden is usually very high, but is supported by the server. Bandwidth costs are relatively low. HE also has limitations in terms of the type and amount of supported operations, which may make the implementation of certain technologies infeasible. The adaptations required by HE can cause utility loss.
  - Functional encryption has limitations in the type of operations that may be performed, and also suffers from performance restrictions. Applications of FE are different from HE as the processing party in FE is the party who learns the result.
- *Separate:*
  - As we discussed above, FL is not fully privacy-preserving. Information from the clients may leak to the central server, and to other clients. For this reason, FL is more and more used in combination with other PETs, such as DP and HE/MPC.
  - Another problem of FL is that it requires the clients to have sufficient training data. This is especially challenging if one expects that the training data is annotated, i.e., for supervised learning. While unsupervised FL with text data has been studied, to the best of our knowledge, unsupervised FL for speech is underexplored.
- *Aggregate:* [33] have recently shown that synthetic data does not provide a better trade-off between privacy and utility than traditional anonymization techniques, which is additionally even harder to predict than for traditional techniques.
- *Inform:*
  - Explainable machine learning: There is an inherent tension between privacy and explainability, as providing the user with an explanation of how an machine learning system reached an outcome inevitably entails leaking some information about the machine learning model and possibly the underlying training data.
  - Explaining the protection functionality of PETs remains a usability issue. For instance, [20] revealed several common misconceptions that lay users develop if confronted with metaphors for differential privacy that are commonly used by media outlets.
  - In general, there is a lack of TETs and of control tools for speech and language technology that users can use in practice for exercising their data subject rights.
- *Control:* Intervenabilty tools that major natural language processing providers, such as Google and Apple, do not support fine-grained controls in the form of deletions or corrections and only limited insights or controls for derived/inferred data (e.g., data portability is usually not provided for inferred data).
- *Enforce:* Defining fine-grained access control policies that also sufficiently protect not only content data but also metadata that can be inferred from speech and language processing, remains a research challenge.

References

1 Ala Sarah Alaqra, Simone Fischer-Hübner, and Erik Framner. Enhancing privacy controls for patients via a selective authentic electronic health record exchange service: qualitative study of perspectives by medical professionals and patients. *Journal of Medical Internet Research*, 20(12):e10954, 2018.

2 Balamurugan Anandan, Chris Clifton, Wei Jiang, Mummoorthy Murugesan, Pedro Pastrana-Camacho, and Luo Si. t-plausibility: Generalizing words to desensitize text. *Trans. Data Priv.*, 5(3):505–534, 2012.

3 Anton Batliner, Simone Hantke, and Björn Schuller. Ethics and good practice in computational paralinguistics. *IEEE Transactions on Affective Computing*, 13(3):1236–1253, 2020.

4 Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, 2019.

5 Hongyan Chang and Reza Shokri. On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 292–303. IEEE, 2021.

6 Alice Cohen-Hadria, Mark Cartwright, Brian McFee, and Juan Pablo Bello. Voice anonymization in urban sound recordings. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2019.

7 Edgar Dale and Jeanne S Chall. A formula for predicting readability: Instructions. *Educational Research Bulletin*, pages 37–54, 1948.

8 Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. Code alltag 2.0–a pseudonymized german-language email corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4466–4477, 2020.

9 Rudolf Flesch. A readability formula in practice. *Elementary English*, 25(6):344–351, 1948.

10 Rudolph Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221, 1948.

11 Robert Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.

12 Yaakov HaCohen-Kerner. Survey on profiling age and gender of text authors. *Expert Systems with Applications*, page 117140, 2022.

13 Marit Hansen, Meiko Jensen, and Martin Rost. Protection goals for privacy engineering. In *2015 IEEE Security and Privacy Workshops*, pages 159–166. IEEE, 2015.

14 Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

15 Francis Heylighen and Jean-Marc Dewaele. Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel*, 4, 1999.

16 Jaap-Henk Hoepman. Privacy design strategies. In *IFIP International Information Security Conference*, pages 446–459. Springer, 2014.

17 Timour Igamberdiev, Thomas Arnold, and Ivan Habernal. Dp-rewrite: Towards reproducibility and transparency in differentially private text rewriting. *arXiv preprint arXiv:2208.10400*, 2022.

18 Bargav Jayaraman, Esha Ghosh, Huseyin Inan, Melissa Chase, Sambuddha Roy, and Wei Dai. Active data pattern extraction attacks on generative language models. *arXiv preprint arXiv:2207.10802*, 2022.

19 Johanna Johansen, Tore Pedersen, Simone Fischer-Hübner, Christian Johansen, Gerardo Schneider, Arnold Roosendaal, Harald Zwingelberg, Anders Jakob Sivesind, and Josef Noll. A multidisciplinary definition of privacy labels. *Information & Computer Security*, (ahead-of-print), 2022.

20 Farzaneh Karegar, Ala Sarah Alaqra, and Simone Fischer-Hübner. Exploring user-suitable metaphors for differentially private data analyses. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 175–193, 2022.

21 Oleksandra Klymenko, Stephen Meisenbacher, and Florian Matthes. Differential privacy in natural language processing: The story so far. *arXiv preprint arXiv:2208.08140*, 2022.

22 John Laver, editor. *Principles of Phonetics*. Cambridge University Press, 1994.

23 Mohamed Maouche, Brij Mohan Lal Srivastava, Nathalie Vauquier, Aurélien Bellet, Marc Tommasi, and Emmanuel Vincent. Enhancing speech privacy with slicing. In *Interspeech 2022-Human and Humanizing Speech Technology*, 2022.

24 G. Harry Mc Laughlin. Smog grading – a new readability formula. *Journal of Reading*, 12(8):639–646, 1969.

25 Patrick Murmann and Simone Fischer-Hübner. Tools for achieving usable ex post transparency: a survey. *IEEE Access*, 5:22965–22991, 2017.

26 Jihad S. Obeid, Paul M. Heider, Erin R. Weeda, Andrew J. Matuskowitz, Christine M. Carr, Kevin Gagnon, Tami Crawford, and Stéphane M. Meystre. Impact of de-identification on clinical text classification using traditional and deep learning classifiers. *Studies in Health Technology and Informatics*, 264:283, 2019.

27 Alexandr Railean. *Improving IoT Device Transparency by Means of Privacy Labels*. PhD thesis, Georg-August-Universität Göttingen, 2022.

28 Devin Reich, Ariel Todoki, Rafael Dowsley, Martine De Cock, and Anderson C. A. Nascimento. Privacy-preserving classification of personal text messages with secure multi-party computation. *Advances in Neural Information Processing Systems*, 32, 2019.

29 David Sánchez and Montserrat Batet. C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology*, 67(1):148–163, 2016.

30 Björn Schuller and Anton Batliner. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. John Wiley & Sons, 2013.

31 RJ Senter and Edgar A Smith. Automated readability index. Technical report, Cincinnati Univ OH, 1967.

32 Samuel Sousa and Roman Kern. How to keep text private? a systematic review of deep learning methods for privacy-preserving natural language processing. *Artificial Intelligence Review*, pages 1–66, 2022.

33 Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data–anonymisation groundhog day. *arXiv preprint arXiv:2011.07018*, 2021.

34 Francisco Teixeira, Alberto Abad, Bhiksha Raj, and Isabel Trancoso. Towards end-to-end private automatic speaker recognition. *arXiv preprint arXiv:2206.11750*, 2022.

35 Henry Turner, Giulio Lovisotto, and Ivan Martinovic. Speaker anonymization with distribution-preserving x-vector generation for the voiceprivacy challenge 2020. *arXiv preprint arXiv:2010.13457*, 2020.

4.6 Uncertain legal interpretation(s) for emerging PETs

Lydia Belkadi (KU Leuven, BE) lydia.belkadi@kuleuven.be
Peggy Valcke (KU Leuven, BE) peggy.valcke@kuleuven.be

Cryptographic-based PETs (e.g., MPC, HE) rely on secret data representations, or modified views of the data, which allow the hiding of the original form of the data. This means that the original form can only be reconstructed by having access to additional sources of data, such as an encryption key, that are held by designated parties. For instance, in HE only authorized parties should have access to the encryption key. In contrast, in MPC several parties hold different, partial representations of the data which, when put together allow for the reconstruction of the original form.

4.6.1 On the legal understanding of information protected by PETs

Within the European Union, data protection rules reflect the scope and aims of a set of fundamental rights, and in particular the rights to privacy and data protection.[17] Indeed, these rights are not absolute, and may suffer limitations so long adequate legal and technical safeguards are adopted.[18] Against this background, EU data protection regimes rely on a broad definition of personal data, while establishing flexibility mechanisms that consider the specific circumstances at stake. This dual objective is reflected in the legal tests data controllers must carry out to determine whether and to what extent data protection rules apply. In legal terms, these tests are defined as the material and territorial scope of the law.

Under EU law, data protection relies on a dynamic conceptual construction between the definitions of personal data, pseudonymized data and anonymized data. For example, Article 2 of the GDPR defines the scope of the law as applying to

- processing of personal data wholly or partly by automated means,
- and processing other than by automated means of personal data which form part of a filing system.

Hence, data protection frameworks only apply to "personal data", defined as "any information relating to an identified or identifiable natural person" (Article 4(1) of the GDPR). The Article 29 Working Party explains in its Opinion 4/2007 on the Concept of Personal Data that the notion of identification is constructed in a broad way to encompass both direct and indirect identification and identifiability (e.g., name, identifiers, unique combinations of factors).[19]

However, the law also considers the effects of two types of transformations on the legal nature of personal data. On the one hand, anonymous data is explicitly excluded from the material scope of the GDPR. Recital 26 of the GDPR explains that anonymous data is defined negatively, as information which does not fall under the definition of personal data. In other words, anonymous data is information that does not relate to an identified or identifiable individual (i.e., information that is intrinsically anonymous from a data protection perspective) or personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable (i.e., transformed data). The Article 29 Working Party further explains in its Opinion 05/2014 on Anonymisation Techniques that such data transformations should be irreversible. To determine whether a technique amounts to an anonymization processing in the legal sense, the Article 29 Working Party underlines three factors which need to be assessed: singling out, linkability and inference.[20] On the other hand, the GDPR defines pseudonymization as a processing performed in such a manner that personal data cannot be attributed to an individual without the use of additional information (Article 4(5) of the GDPR). In its Opinion 05/2014, the Article 29 Working Party explicitly categorizes encryption with secret keys as a type of pseudonymization processing.[21] The GDPR further requires that additional information should be kept separately and subjected to technical and organizational measures.

PETs have emerged as tools to support controllers in complying with their legal obligations.

---

[17] The right to private life is enshrined in the Universal Declaration of Human Rights (Article 12), the European Convention on Human Rights (Article 8), and the European Charter of Fundamental Rights (Article 7). Under EU law, a separate right to data protection is also recognized in the European Charter of Fundamental Rights (Article 8).

[18] See the articles referenced above and Article 52 of the European Charter of Fundamental Rights regarding the legal tests that must be performed to restrict the rights to privacy and data protection. In the context of biometric processing, see [1].

[19] Article 29 Working Party, Opinion 4/2007 on the Concept of Personal Data, p. 13.

[20] Article 29 Working Party, Opinion 05/2014 on Anonymisation Techniques p. 7-19.

[21] Ibid p. 20.

Nonetheless, there remains substantial uncertainties as to how these technologies interact with and are regulated by data protection rules. This dual aspect of PETs raises important research questions regarding the legal nature of (a) new processing frameworks developed and implemented, and (b) the information protected by PETs.

4.6.2 On the need for specific risk assessments for PETs

In turn, this dual legal nature of PETs also raises important challenges to established legal methodologies to assess data processing risks. In particular, PETs are characterized by their variety and their composite nature. In other words, PETs seek to address specific privacy objectives (e.g., minimization, obfuscation, etc.), and can be used in combination. As a result, the residual risks arising from such assemblages are highly context-dependent. However, there remains a significant gap in interdisciplinary scholarship regarding the analysis and development of tailored risk assessments that would consider the nature and effects of PETs. In particular, there is a significant need to further research venues for interdisciplinary concept-building and suitable flexibility mechanisms.

References

1 Els Kindt. *Privacy and Data Protection Issues of Biometric Applications. A Comparative Legal Analysis*. Springer, 2013.

5 Conclusion

As can be seen from the above findings, the domains of voice, speech, and natural language processing provide a lot of opportunities, but also challenges, when it comes to privacy and data protection. Multiple aspects of privacy have to be considered and incorporated in the design of such processing systems, irrespective of being a smart speaker, voice assistant, or chatbot. The most relevant domains identified in the endeavor of this report are:

- legal considerations (GDPR, EU AI Act, and other applicable laws must be considered),
- human factors (usability and transparency must be addressed, vulnerable groups must be considered),
- technical aspects (voice anonymization techniques and privacy-enhancing technologies in general should be considered whenever possible).

In this report, we provided a first outline of these challenges from an interdisciplinary point of view.

Participants

- Lydia Belkadi
KU Leuven, BE
- Zinaida Benenson
Friedrich-Alexander-Universität – Erlangen, DE
- Martine De Cock
University of Washington – Tacoma, US
- Abdullah Elbi
KU Leuven, BE
- Zekeriya Erkin
TU Delft, NL
- Natasha Fernandes
Macquarie University – Sydney, AU
- Simone Fischer-Hübner
Karlstad University, SE
- Ivan Habernal
TU Darmstadt, DE
- Meiko Jensen
Karlstad University, SE
- Els Kindt
KU Leuven, BE
- Dietrich Klakow
Saarland University – Saarbrücken, DE
- Katherine Lee
Google Brain & Cornell University – Ithaca, US
- Anna Leschanowsky
Fraunhofer IIS – Erlangen, DE
- Pierre Lison
Norsk Regnesentral – Oslo, NO
- Christina Lohr
Friedrich-Schiller-Universität – Jena, DE
- Emily Mower Provost
University of Michigan – Ann Arbor, US
- Andreas Nautsch
Avignon Université, FR
- Olga Ohrimenko
University of Melbourne , AU
- Jo Pierson
Free University of Brussels, BE
- Laurens Sion
KU Leuven, BE
- David Stevens
Gegevensbeschermingsautoriteit – Brussels, BE
- Francisco Teixeira
Instituto Superior Técnico – Lisbon, PT
- Natalia Tomashenko
Avignon Université, FR
- Marc Tommasi
University of Lille, FR
- Peggy Valcke
KU Leuven, BE
- Emmanuel Vincent
Inria – Nancy, FR
- Shomir Wilson
Pennsylvania State University – University Park, US

Report from Dagstuhl Seminar 22351

Interactive Visualization for Fostering Trust in ML

Polo Chau[*1], Alex Endert[*2], Daniel A. Keim[*3], and Daniela Oelke[*4]

1 **Georgia Institute of Technology – Atlanta, US.** polo@gatech.edu
2 **Georgia Institute of Technology – Atlanta, US.** endert@gatech.edu
3 **Universität Konstanz, DE,.** keim@uni-konstanz.de
4 **Hochschule Offenburg, DE.** daniela.oelke@hs-offenburg.de

Abstract

The use of artificial intelligence continues to impact a broad variety of domains, application areas, and people. However, interpretability, understandability, responsibility, accountability, and fairness of the algorithms' results – all crucial for increasing humans' trust into the systems – are still largely missing. The purpose of this seminar is to understand how these components factor into the holistic view of trust. Further, this seminar seeks to identify design guidelines and best practices for how to build interactive visualization systems to calibrate trust.

Seminar August 28–September 2, 2022 – http://www.dagstuhl.de/22351
2012 ACM Subject Classification Computing methodologies → Artificial intelligence; Computing methodologies → Machine learning; Human-centered computing → Visualization
Keywords and phrases accountability, artificial intelligence, explainability, fairness, interactive visualization, machine learning, responsibility, trust, understandability
Digital Object Identifier 10.4230/DagRep.12.8.103

1 Executive Summary

Polo Chau (Georgia Institute of Technology – Atlanta, US)
Alex Endert (Georgia Institute of Technology – Atlanta, US)
Daniel A. Keim (Universität Konstanz, DE)
Daniela Oelke (Hochschule Offenburg, DE)

Artificial intelligence (AI), and in particular machine learning (ML) algorithms, are of increasing importance in many application areas. However, interpretability, understandability, responsibility, accountability, and fairness of the algorithms' results – all crucial for increasing humans' trust into the systems – are still largely missing. All major industrial players, including Google, Microsoft, and Apple, have become aware of this gap and recently published some form of Guidelines for the Use of AI. While it is clear that the level of trust in AI systems does not only depend on technical but many other factors, including sociological and psychological factors, interactive visualization is one of the technologies that has strong potential to increase trust into AI systems. In our Dagstuhl Seminar, we discussed the requirements for trustworthy AI systems including sociological and psychological aspects as well as the technological possibilities provided by interactive visualizations to increase human trust in AI. As a first step, we identified the factors influencing the organizational and sociological as well as psychological aspects of AI and partitioned them into relationship-based and evidence-based aspects. Next, we collected measures that may be used to approximate

[*] Editor / Organizer

Interactive Visualization for Fostering Trust in ML, *Dagstuhl Reports*, Vol. 12, Issue 8, pp. 103–116
Editors: Polo Chau, Alex Endert, Daniel A. Keim, and Daniela Oelke
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

these aspects, such as interaction logs, eye tracking, and EEG. We also discussed the mechanisms to calibrate trust and their potential misuse. Finally, we considered the role that visualizations play in increasing trust in AI systems. This includes questions such as: Which mechanisms exist to make AI systems trustworthy? How can interactive visualizations contribute? Under which circumstances are interactive visualizations the decisive factor for enabling responsible AI? And what are the research challenges that still have to be solved – in the area of machine learning or interactive visualization – to leverage this potential in real world applications?

The seminar started with four keynote talks by experts in cognitive psychology, sociology, AI, and visualization, to provide participants with diverse perspectives that helped seed discussion topics. Then, the group decided to build 6 smaller groups to discuss the individual topics that should be worked on during the rest of the week. The six groups collectively came up with a longer list of potential topics surrounding the aspects of trust and machine learning. This list was voted on the plenum to distill it to the following four breakout groups: (1) Good practices and evil knobs in machine learning; (2) Evaluation, measures and metrics for trust in ML; (3) Interaction, expectations and dimension reduction; and (4) Definitions, taxonomy and relationships of trust in ML.

The outcome of this seminar is a better understanding of which aspects of trust have to be considered in fostering trust in AI systems and how interactive visualizations can help foster trust in artificial intelligence systems by making them more understandable and responsible. This will encourage innovative research and help to start joint research projects tackling the issue. Concrete outcomes are drafts of position papers describing the findings of the seminar and in particular, the research challenges identified in the seminar.

2 Table of Contents

3 Overview of Talks

3.1 Visual, Interactive, and Explainable AI: Perspectives on Trust-Building through Explainability

Mennatallah El-Assady (ETH Zürich, CH)

URL https://el-assady.com/

Interactive, mixed-initiative machine learning promises to combine the efficiency of automation with the effectiveness of humans for collaborative decision-making and problem-solving process. This can be facilitated through co-adaptive visual interfaces.

In the first part of this talk, I recapped the definitions of mixed-initiative analysis, arguing for the need for effective explanations, both from the side of the human as well as the AI agents. Next, I summarized visual, interactive, and explainable AI approaches along the process of understanding, diagnosis, and refinement of models.

Second, I reviewed human-centered evaluations in human-centered AI, focusing on how the trustworthiness of AI models and the trustworthiness of explanations were evaluated in previous works. Following up on this, I presented the output of the survey on enhancing trust in machine learning through visualization.

Lastly, I ended the talk with some reflections and open questions relating to trust-building through explainability.

3.2 Cognitive perspectives on visualization and trust

Brian D. Fisher (Simon Fraser University – Surrey, CA)

Joint work of Brian D. Fisher, Samar Al-Hajj, Richard Arias-Hernández, Linda T. Kaastra
Main reference B. Fisher & D. Kasik (2023) Pair Analytics in a Visual Analytics Context. Proceedings of the 56th Annual Hawaii International Conference on System Sciences. IEEE Digital Library, to appear.

Highly Integrated Basic and Application- Responsive (HIBAR) research approaches try to find optimal ways to bridge the knowledge-creation of basic science with the design and engineering of advanced technologies. Here I discuss ways to apply this approach to build interactive visualization systems that might be used to establish trust in machine learning processes that are grounded in basic research in the cognitive science of visually-enabled reasoning and agent-agent communication and at the same time to contribute to knowledge about those processes. I begin with a discussion of various kinds of trust and ways in which they are cognitively processed. In order to reduce the complexity of this analysis I use David Marr's triune approach from Vision (1984) of implementation, algorithm, and operational requirements and its extension by Poggio to include perspectives on the evolution and development of expertise in a given cognitive task.

This analysis is helpful in design of visualization for many complex cognitive processes that are supported by visual information in structured environments, such as our new project on cancer diagnosis through medical image analysis using machine learning. In order to address issues of trust, we must build a parallel understanding of how people are able to coordinate their behaviour with that of other agents. I build this from H.H. Clark's psycholinguistic pragmatic approach to human-human coordination in Joint Activities. Here

too we see benefits from Marr's Triune approach, with D'Andrade's Cognitive Anthropology and Hutchins' approach to Cognitive Ethnography as examples of ways in which groups of people, their technologies, and channels of communication interact to produce extended and distributed cognitive systems, with examples from work in my laboratory using our Pair Analytics and Group Analytics approach to safety and health decision-making, and our recent collaboration on Decision Intelligence approached pioneered by Lorien Pratt.

My final topic is pragmatic– how can we most effectively build HIBAR research programs that bridge real-world applications and creation of scientific knowledge. I briefly discuss new approaches to technoscience and creative design collaboration that require a rethinking of the structures that define research organizations in the university and industry. In order to build these systems in a reflective manner we must take the lens we used to understand other organizations as distributed systems and apply it to our own organizations.

3.3 Visualizing Deep Networks

Barbara Hammer (Universität Bielefeld, DE)

Joint work of Alexander Schulz, Fabian Hinder, Barbara Hammer
Main reference Alexander Schulz, Fabian Hinder, Barbara Hammer: "DeepView: Visualizing Classification Boundaries of Deep Neural Networks as Scatter Plots Using Discriminative Dimensionality Reduction", in Proc. of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, pp. 2305–2311, ijcai.org, 2020.
URL http://dx.doi.org/10.24963/ijcai.2020/319

In this spotlight talk, a pipeline for visualizing the classification boundary of deep networks together with example data has been presented. The technology relies on two tricks: using discriminative dimensionality reduction to shape the otherwise ill-posed problem of dimensionality reduction reduction from high dimensional spaces according to the task t hand, and sampling in the projection instead of the data space for efficiency. For deep networks, the first part can be approximated using one backpropagation loop only. The algorithm can be substantiated by convergence guarantees, and it is available as code https://github.com/LucaHermes/DeepView

3.4 When Does Seeing Become Believing? Potential Impacts of Model Characteristics and Visual Cues on Human Decisions

Laura Matzen (Sandia National Labs – Albuquerque, US)

Joint work of Laura Matzen, Breannan C. Howell, Zoe Gastelum, Kristin M. Divis, Michael C. Trumbo

How can interactive visualizations foster trust in machine learning (ML)? Trust is an extremely complex issue. A user may fail to trust a model when they could benefit from doing so, or they might trust too much, complying with a model's outputs even when they are erroneous. How can we support people in developing appropriate levels of trust in an ML tool, so that the human-machine system can reach the highest possible level of performance? In this talk, I discuss some of the cognitive issues that come into play when humans are asked to make decisions based on visualizations and other representations of information. I present two

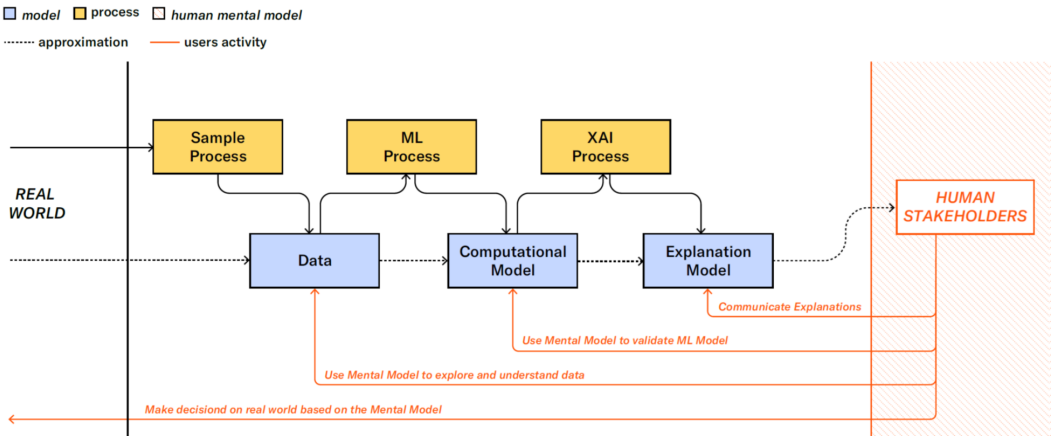
lines of research, one focused on the impact of ML errors on human decision making and one focused on visualizations of state uncertainty. Across these two sets of experiments, we found that different representations of the same information can lead to different patterns of decisions. People's ability to detect ML errors is impacted by the overall error rate of the system. Their tolerance of risk in a decision making task is impacted by the way in which information about risk is presented. In addition, individual differences in cognition and domain expertise influence participants' interpretation of and trust in ML outputs. The results of these experiments illustrate some of the many factors that can influence users' trust and decisions. Additional research at the intersection of cognitive science, data science, data visualization, and visual analytics will be required to develop a systematic understanding of these factors and the interactions between them.

3.5 Relations between models, trust and processes involved in Machine Learning

Daniela Oelke (Hochschule Offenburg, DE)

Joint work of the participants of the Dagstuhl Seminar 19452 "Machine Learning Meets Visualization to Make Artificial Intelligence Interpretable"

About three year ago a related seminar entitled "Machine Learning Meets Visualization to Make Artificial Intelligence Interpretable" took place in Dagstuhl. This talk introduces one of the outputs of this Dagstuhl Seminar, a diagram illustrating the relations between the different types of model and processes involved in the ML process and its explanation including the interactions of the human stakeholders with the models. The diagram does not yet address issues with trust, but could be adapted towards this direction.



3.6 Expectations, trust, and evaluation

Maria Riveiro (Jönköping University, SE)

Main reference Maria Riveiro, Serge Thill: ""That's (not) the output I expected!" On the role of end user expectations in creating explanations of AI systems", Artif. Intell., Vol. 298, p. 103507, 2021.
URL http://dx.doi.org/10.1016/j.artint.2021.103507

This talk focuses on the role of expectations in designing explanations from Artificial Intelligence/Machine Learning (AI/ML) -based systems. Explanations are crucial for system understanding that, in turn, are very relevant to supporting trust and trust calibration in such systems. I discuss the connections between expectations, explanations and trust in human-AI/ML system interaction.

I present two recent studies ([1, 2]) investigating if expectations modulate what people want to see and when from an AI/ML system when carrying out analytical tasks.

We found out that,

- For matched expectations, an explanation is often not required at all, while if one is, it is of the factual type
- For mismatched expectations, the picture is less clear, primarily because there does not seem to be a unique strategy, although mechanistic explanations are requested more often than other types

Overall, user expectations are a significant variable in determining the most suitable content of explanations (including whether an explanation is needed at all). More research is needed to investigate the relationship between expectations and explanations, and how they support trust calibration.

References

1 Riveiro, M., and Thill, S. (2021). That's (not) the output I expected! On the role of end user expectations in creating explanations of AI systems. Artificial Intelligence, 298, 103507.
2 Riveiro, M., and Thill, S. (2022). The challenges of providing explanations of AI systems when they do not behave like users expect. In Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (pp. 110-120).

3.7 Nonlinear dimensionality reduction – visualization with machine learning

Michel Verleysen (University of Louvain, BE)

Joint work of Michel Verleysen, John A. Lee, Cyril de Bodt, Pierre Lambert

Nonlinear dimensionality reduction (NLDR) is a branch of the wide machine learning field. NLDR is essentially unsupervised, which means that it is used to find something (information) in data, but we do not know in advance what kind of information. Consequently, even if it is easy to agree on the general principle of reducing the dimension of data without losing too much information, it is very difficult to agree on a scientific measure of how this loss is evaluated, leading to hundreds of NLDR methods. None of them can be objectively considered as better than others; they all reflect a specific user's point of view. Popular methods are often those that come with an efficient implementation, rather than being chosen for the quality (seen by the user) of their outputs.

This talk insists on one of the users' point of views, often underestimated in the literature: the compromise between a global and a local mapping of the data. We show that multiscale methods may produce much more interesting representations, with an additional computational cost that can be limited with the development of fast yet accurate algorithms.

3.8 A Computer Scientist's Existential Crisis

Emily Wall (Emory University – Atlanta, US)

This talk began with the observation that trust in AI systems encompasses trust in (1) the model – that it is accurate and just, (2) the decision maker – that the human-in-the-loop standing between you and the model decision or prediction has your best interest in mind, and (3) oneself – that you are equipped to make informed decisions. The technical aspects of trust in AI systems are actually only a small part of the story. Improving model accuracy and quantifying and mitigating bias in models can serve to calibrate trust. For visualization researchers in particular, this begs the question: what is our role? Where can we have impact? The talk asserts that this is a challenging socio-technical problem, requiring a suite of methodologies and frameworks that are not especially common in our community. I conclude the talk with a reference to a paper[1] that leverages a valuable socio-technical perspective to coin the concept "human-centered explainable AI" abbreviated HCXAI. This paper introduces important frameworks (including critical technical practice, reflective design, value-sensitive design) that can serve as a starting point for visualization researchers to expand their tool belts to include critical socio-technical frameworks to inform next steps addressing trust in AI through interactive visualization.

References

1 Upol Ehsan and Mark O. Reidl (2020). *Human-centered explainable ai: Towards a reflective sociotechnical approach.* International Conference on Human-Computer Interaction.

4 Working groups

4.1 Definitions, taxonomy and relationships

Emma Beauxis-Aussalet (VU University Amsterdam, NL), Peer-Timo Bremer (LLNL – Livermore, US), Steffen Koch (Universität Stuttgart, DE), Jörn Kohlhammer (Fraunhofer IGD – Darmstadt, DE), and Daniela Oelke (Hochschule Offenburg, DE)

A major point of discussion was the definition of trust. It quickly became clear that we needed to differentiate between (1) trust (as historically defined by philosophy) that denotes the relationship between humans and between humans and organizations, and (2) trust that concerns technical artifacts, especially machine learning components. A first attempt was the separation of trust on the human side and confidence on the technical side. While this worked to structure the terms that are related to trust, the current (different) use of these terms in the

various communities led to reconsiderations. A strong second candidate was the separation into subjective trust and objective trust. However, this did not transport well, that trust is inherently between humans, not a state of mind of a single human. Also, the objectivity of several aspects on the technical side, when it comes to trusting a technical artifact, was not adequately covered. An extensive research of terms by Emma Beauxis-Aussalet led us to the final two terms that we now use in this Dagstuhl seminar: relationship-based trust on the human side, and evidence-based trust on the technical side. The material that we prepared also contains a diagram showing how these two sides relate to each other during human decision making.

Detailed discussion of the definition of the term "trust"

Prior art [2, 1, 3] suggests that reliance on relationships is the core framework of trust, whether it concerns trust in another human, in an institution, in oneself, or in technology – the latter three being modeled after the first. This reliance can be warranted, based on evidence, but also arises from the necessity to depend on another party, which is potentially fallible. Evidence may eliminate the risks of such dependency (e.g., *well-grounded trust*) or only reduce them and demonstrate the value of such dependency (e.g., *justified trust*).

A body of evidence can inform the decision to enter or exit a relationship of dependency with another party, human or object. But such evidence does not only concern technical information, e.g., about the reliability of a ML system. It also concerns organisational and societal considerations that are external to the trustworthiness of technology. It is thus essential to consider the relationships that arise from integrating technology in human organisations and societies.

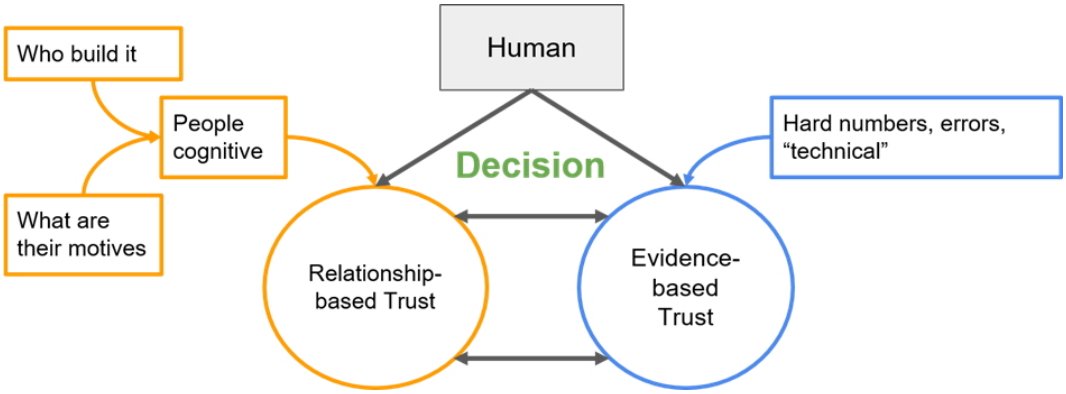
In essence, **trust may not be achieved solely by providing evidence on technical risks**: Trust can be established or withdrawn for reasons other than the quality of a technology, for instance due to material necessities or power dynamics. Trust in technology is also relationship-based because i) it relies on the relationships between a larger body of actors who have agency on the technology or its usage, ii) humans also establish relationships of reliance to the technology. The latter is arguably about trust, as the motives of the trustee is an essential component of trust [2] and technology itself has no motives. However, we can acknowledge that anthropomorphism occurs, and includes the illusory attribution of motives to AI and automated systems. Hence, computer science and visualization research aiming to support trust in machine learning must consider the cognitive, emotional, and societal aspects that are inherent to relationship-based trust.

Recent works also discuss trust by persons or organisations in AI and Machine Learning [4, 5]. While these works describe facets that can be used to make a distinction regarding relation-based and evidence based trust, this is not discussed explicitly.

Evidence-based trust is built on factual information such as error metrics, uncertainty estimates, and all the information available to reach a decision or complete a task. This information is measurable and verifiable. Yet humans are not entirely objective and base their decisions on a combination of subjective and objective aspects of a situation: trust in humans, organizations, or technology may not be fully informed by evidence, but combine relationship-based and evidence-based trust. Furthermore, a single human is not able to fully assess all possible evidence. Thus relationship-based trust is necessary to mediate the complexity of technology by delegating their assessment and management to a network of specialised actors.

Relationship-based trust is built on practical cooperations between actors with specific skills, motives, and will. In machine learning, the parties and roles of relationship-

based trust are largely evolving, same as the role of webmaster at the beginning of the web eventually evolved to a network of specialists. However evolutive the relationships, to support relationship-based trust it is essential to identify the goals, tasks, information needs, and profile of each actor.



References

1 Aneil K. Mishra. "Organizational Responses to Crisis: The Centrality of Trust." In: Trust in Organizations. Ed. by Roderick M. Kramer and Thomas Tyler. Newbury Park, California, USA: Sage, 1996
2 Carolyn McLeod, "Trust", The Stanford Encyclopedia of Philosophy (Fall 2021 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2021/entries/trust/
3 Rotter, Julian B. "A new scale for the measurement of interpersonal trust" Journal of personality 35.4 (1967): 651-665.
4 Toreini, Ehsan, et al. "The relationship between trust in AI and trustworthy machine learning technologies." Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020.
5 Ashoori, Maryam, and Justin D. Weisz. "In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes." arXiv preprint arXiv:1912.02675 (2019).

4.2 A human-centered perspective on trust in AI-driven socio-technical systems

Peer-Timo Bremer (LLNL – Livermore, US), Emma Beauxis-Aussalet (VU University Amsterdam, NL), Polo Chau (Georgia Institute of Technology – Atlanta, US), David S. Ebert (University of Oklahoma – Norman, US), Daniel A. Keim (Universität Konstanz, DE), Steffen Koch (Universität Stuttgart, DE), and Daniela Oelke (Hochschule Offenburg, DE)

Trust in the information provided is often cited as one of the key challenges to fully integrate AI-driven systems into high consequence decisions. However, there rarely exist a clear definition of the concept, what can be done to influence trust, and even whether the implicit goal to increase trust is always appropriate. The HCI community is often asked to mediate between the various stakeholders and visualization in particular is seen as the ideal conduit

to convey both subjective and objective information. However, the amalgamation of human and social aspects with more technical concerns on correctly conveying unbiased information makes formulating a clear research perspective difficult. Here we argue that the general notion of "trust" should be thought of as two related but fundamentally different concepts: relationship-based trust and evidence-based trust. A typical example of the former is the trust one places in an car functioning, which is rooted in the believe that an engineer certified the system rather than in any personal knowledge of the mechanics. Conversely, evidence-based trust is based in factual information, i.e., statistics on past performance or uncertainty bounds, analyzed directly. In both cases, the overarching goal should be to correctly calibrate trust to avoid both unfounded over-trust, for example, based on the social network echo chamber, as well as unjustified skepticism. This perspective will first define both concepts, show how they implicitly or explicitly align with prior arguments, and that they lead to fundamentally different research challenges. Subsequently, the perspective discusses priority research directions aimed at calibrate both form of trust in AI-driven system, how the different notions interact, and most importantly areas where unfettered research may raise ethical concerns. While this perspective grew out of discussions in the HCI community addressing many of the challenges will require convergent research in cognitive science, machine learning, ethics, visualization, and many others.

4.3 Trust Junk and Evil Knobs: Duality of Trust-Calibration Design Measures

Alex Endert (Georgia Institute of Technology – Atlanta, US), Rita Borgo (King's College London, GB), Polo Chau (Georgia Institute of Technology – Atlanta, US), Mennatallah El-Assady (ETH Zürich, CH), Laura Matzen (Sandia National Labs – Albuquerque, US), Adam Perer (Carnegie Mellon University – Pittsburgh, US), Harald Schupp (Universität Konstanz, DE), Hendrik Strobelt (MIT-IBM Watson AI Lab – Cambridge, US), and Emily Wall (Emory University – Atlanta, US)

Many AI systems make claims that specific design choices enhance trust or serve to calibrate trust. However, interface design choices are not neutral with respect to trust. There is inherent duality – that the same design choice may enhance trust in some cases, while simultaneously detracting in others. This group conceptualized "trust junk," analogous to "chart junk" in visualization, i.e., design choices intended to enhance trust without any specific connection to data, model, or the nature of the decision.

Consider AIs that utilize social information, e.g., "5 others in the organization have accepted a recommendation today." This choice may increase trust in the AI having social endorsement by others; however, this may also be used to create unfair social pressure and manipulate choices of the user that serve the interface creators. The group expanded these examples to consider different kinds of "knobs," which represent different design choices that can be made in the creation of an AI system. These include choices about which datasets are modeled, how the outputs of the system are represented to users, and what options users have for interacting with the outputs, the model, or the data. Turned in one direction in a given context, these knobs may enhance trust in the system. When considered from an adversarial perspective, these choices could also be used to mislead users or to promote

unwarranted levels of trust in a system. We construct a framework of these knobs and assert that understanding this space necessitates a sociotechnical approach; concrete generalizable interface guidelines may not yet be made.

4.4 Trust Evaluation

Maria Riveiro (Jönköping University, SE), Michael Behrisch (Utrecht University, NL), Simone Braun (Hochschule Offenburg, DE), David S. Ebert (University of Oklahoma – Norman, US), Daniel A. Keim (Universität Konstanz, DE), Tobias Schreck (TU Graz, AT), and Hendrik Strobelt (MIT-IBM Watson AI Lab – Cambridge, US)

Trust assessment during the data analysis process is a challenging task. Only if stakeholders have trust in the used data, algorithmic, and visual-interactive components, the results will be accepted, relied on and applied. In an ideal system, the trust of the user could be observed (measured), and the system could be adapted to increase trust where it is lacking, e.g., by providing additional information, explanations, or summarizations. However, trust is dynamic and emerges due to many influencing factors, both from the data analysis system designs, as well as personal factors. Additionally, learning effects, change in the user tasks, or socio-cultural influences complicate trust calibration. To date, several trust scales exist, but it is hard to assess how technological aspects of ML/AI systems affect trust and how to carry out empirical evaluations that isolate the effects that changes in technology have on trust.

Related work in Human-centered AI and Human-computer interaction suggests frameworks for compartmentalizing trust into cognition-based and affect-based trust. We combine two earlier frameworks from Madsen & Gregor (2000) [1] and Gulati et al. (2019) [2] and add socio-cultural trust factors that were not considered in the discussion. We scrutinize these frameworks considering particularities of AI/ML, Visualization and Interaction. Overall, we propose the following aspects/constructs of trust:

- Cognition-Based Trust
  - Perceived risk
  - Benevolence
  - Competence
  - Reciprocity
- Affect-Based Trust
  - Faith/Confidence
  - Personal Attachment
  - Personality (locus of control, risk-averseness, etc.)
  - Experience (past experiences)
- Social and cultural trust (environmental and contextual factors)
  - Peer Influence
  - Crowd Behavior
  - Cultural Norms

We complement our work by discussing which possible proxies, indirect measures that allow us to approximate trust scores, we can use to assess these aspects, and elaborate on which proxies are most suitable for each trust aspect/construct. The measures outlined are

questionnaires, EEG, Eye-Tracking, Face expressions, movement, recognition (e.g. FACS), Galvanic skin response (e.g. glove), Interaction logs, Feedback from users, Voice recognition, changes (e.g. shimmer, jitter, pitch, balance via GeMAPS), Model questioning users, Games and A/B-Testing. Future work has to prove their applicability for this challenging task.

We finalize with a discussion on how scalable and how good proxies these measures are for trust, and we outline the next steps in utilizing this knowledge for carrying out empirical evaluations and during the design process of Vis/ML/AI-based systems.

References

1 Madsen, M. and Gregor, S., 2000. Measuring human-computer trust. In 11th Australasian conference on information systems (Vol. 53, pp. 6-8). Brisbane, Australia: Australasian Association for Information Systems.

2 Gulati, S., Sousa, S. and Lamas, D., 2019. Design, development and evaluation of a human-computer trust scale. Behaviour & Information Technology, 38(10), pp.1004-1015.

4.5 The Flow of Trust: An Interactive Visualization Framework for Externalizing, Exploring and Explaining Trust in ML Applications

Stef Van den Elzen (TU Eindhoven, NL), Gennady Andrienko (Fraunhofer IAIS – Sankt Augustin, DE), Natalia V. Andrienko (Fraunhofer IAIS – Sankt Augustin, DE), Brian D. Fisher (Simon Fraser University – Surrey, CA), Rafael M. Martins (Linnaeus University – Växjö, SE), Jaakko Peltonen (Tampere University of Technology, FI), Alexandru C. Telea (Utrecht University, NL), and Michel Verleysen (University of Louvain, BE)

Currently, trust in Machine Learning applications is an implicit process that takes place in the mind of the user. As a result there is no method of feedback or communication of trust that can be acted upon. Trust differs from mere ability to inspect the model and from a model's claimed confidence in its predictions. frameworks that support such aspects are not sufficient to support trust. We argue that trust needs to be considered as a first-class citizen in the workflow of developing and using machine learning models. We present a formalization of trust flow and externalization as part of interactive machine learning workflows for analysis and for decision-making. The formalization differentiates several user roles in exploring machine learning models at different workflow stages and their corresponding opportunities for explicit communication of trust targeted at these stages. The formalization enables construction of interaction modes and interfaces to help users to efficiently build and communicate trust in ways that are appropriate for a given stage in the analytic process. Moreover, the formalization differentiates the roles of model exploration and trust communication of the user as well as differentiating user trust from a model's internal probabilistic representations. We formulate several research questions and directions arising from our framework which include

(a) typology/taxonomy of trust objects, trust issues, and possible reasons for (mis)trust;
(b) formalisms to represent trust in machine-readable form;
(c) ways for users to express their state of trust;
(d) ways to explore and develop trust over models and their different aspects using visual interactive techniques;
(e) ways to facilitate the user's expression and communication of the state of trust using visual interactive techniques.

Participants

- Gennady Andrienko
Fraunhofer IAIS – Sankt Augustin, DE
- Natalia V. Andrienko
Fraunhofer IAIS – Sankt Augustin, DE
- Emma Beauxis-Aussalet
VU University Amsterdam, NL
- Michael Behrisch
Utrecht University, NL
- Rita Borgo
King's College London, GB
- Simone Braun
Hochschule Offenburg, DE
- Peer-Timo Bremer
LLNL – Livermore, US
- Polo Chau
Georgia Institute of Technology – Atlanta, US
- David S. Ebert
University of Oklahoma – Norman, US
- Mennatallah El-Assady
ETH Zürich, CH
- Alex Endert
Georgia Institute of Technology – Atlanta, US
- Brian D. Fisher
Simon Fraser University – Surrey, CA
- Barbara Hammer
Universität Bielefeld, DE
- Daniel A. Keim
Universität Konstanz, DE
- Steffen Koch
Universität Stuttgart, DE
- Jörn Kohlhammer
Fraunhofer IGD – Darmstadt, DE
- Rafael M. Martins
Linnaeus University – Växjö, SE
- Laura Matzen
Sandia National Labs – Albuquerque, US
- Daniela Oelke
Hochschule Offenburg, DE
- Jaakko Peltonen
Tampere University of Technology, FI
- Adam Perer
Carnegie Mellon University – Pittsburgh, US
- Maria Riveiro
Jönköping University, SE
- Tobias Schreck
TU Graz, AT
- Harald Schupp
Universität Konstanz, DE
- Hendrik Strobelt
MIT-IBM Watson AI Lab – Cambridge, US
- Alexandru C. Telea
Utrecht University, NL
- Stef Van den Elzen
TU Eindhoven, NL
- Michel Verleysen
University of Louvain, BE
- Emily Wall
Emory University – Atlanta, US