Traceability and securing of results as essential challenges of research in the digital age

April 8-9, 2019

Hörsaalruine

Berlin Museum of Medical History
Charitéplatz 1, 10117 Berlin

About this conference

One currently pivotal global challenge for scientific research in the digital age concerns the potential contradiction between (largely) automatized processing of an ever-growing amount of data and the need for validating, verifying and securing results. This two-day conference will illustrate how these essential challenges regarding data provenance, collection, storage, processing and interpretation are tackled in a number of different disciplines such as physics, bioinformatics, materials science and the digital humanities as represented by computational linguistics. In addition to gathering state-of-the-art facts and insights from these different subjects, the conference aims at promoting exchange and reflection from a broader, interdisciplinary perspective. The focus will thereby lie on methodological issues and deliberately refrain from addressing -equally essential- ethical and legal aspects.

The conference is part of the Digital Turn in the Sciences and Humanities project of the Head Office of the German Research Foundation (DFG). The DFG Head Office has launched this project to assess the current situation of the sciences and humanities in the digital age. The aim of this work is to put the DFG in a better position to respond to key developments from a scientific point of view and, if necessary, adapt its funding policies and practices.
For further information please visit http://www.dfg.de/en/research_funding/principles_dfg_funding/digital_turn/index.html.

Programme

April 8, 2019

From 9:30	Arrival and registration
10:30	Conference opening by **Peter Strohschneider**, President of the DFG
10: 45	*Traceability, reproducibility, replicability…What it means for Computational Linguistics* **Nancy Ide** (Vassar)
11:45	*Towards (more) transparent Natural Language Processing technologies: How teaching others about our tools forces us to ask the right questions* **Antske Fokkens** (VU Amsterdam)
12:45	Lunch break
14:00	*Data provenance* **Peter Buneman** (University of Edinburgh)
15:00	*FAIR data: The European Galaxy server* **Björn Grüning** (Freiburg University)
16:00	Coffee break
16:30	*Can knowledge inform Machine Learning?* **Christian Bauckhage** (Fraunhofer IAIS/University of Bonn)
17:30	End of day 1

April 9, 2019

9:00	*Robust and reliable machine learning* **Matthias Hein** (University of Tübingen)
10:00	*Towards Reproducibility in Machine Learning and AI* **Kristian Kersting** (TU Darmstadt)
11:00	Coffee break
11:30	*Traceability in materials design: A use case from molecular simulation* **Chandler Becker** (National Institute of Standards and Technology)
12:30	Lunch break
14:00	*Automizing work flows in computational materials design* **Jörg Neugebauer** (Max-Planck-Institut für Eisenforschung)
15:00	*What is a measurement record?* **Michaël-Andreas Esfeld** (Lausanne University)
16:00	Coffee break
16:30	*Mastering complex data processing procedures: from particle detector measurements via machine learning algorithms to physics results* **Martin Erdmann** (RWTH Aachen)
17:30	End of the conference

Abstracts

Nancy Ide

Traceability, reproducibility, replicability…What it means for Computational Linguistics

We see terms such as "traceability", "reproducibility", and "replicability" frequently these days, due to increasing awareness of the need for more robust research reporting practices. However, these terms are not necessarily used consistently, and where attempts have been made to come up with precise definitions, there is not always complete agreement. Furthermore, the terms may have somewhat different meanings and relevance depending on the discipline in question.

In this talk, I will first consider the various definitions of these terms and attempt to establish their meanings, at least in order to provide context for the remainder of the presentation. I will then discuss their relevance for the field of computational linguistics, which differs from most science disciplines in the nature of its data and results. Finally, I will describe historical practice in computational linguistics and outline recent attempts to deal with the traceability problem in response to the growing outcry for better practices.

Nancy Ide is Professor in Computer Science at the Department of Computer Science at Vassar College. She is also a founding member of the Text Encoding Initiative, for which she was awarded the Antonio Zampolli Prize of the Alliance of Digital Humanities Organisations. Her research interests and activities cover a wide range of topics in computational linguistics with a special focus on annotation and the development of the open-access, web service platform for Natural Language Processing (NLP) research and development called “The Language Application Grid”.

Antske Fokkens

Towards (more) transparent Natural Language Processing technologies: How teaching others about our tools forces us to ask the right questions

Computational linguistics has a rich tradition of using benchmark datasets and shared tasks for evaluating their methods. These resources have been valuable, allowing us to compare methods and assess the quality of our methods. A downside of the wide-spread approach of comparing results on a gold dataset is that it is relative common practice to draw conclusions based purely on which system performs best on some evaluation set. However, scores alone do not show what a method captures correctly and what it (still) fails at. These insights become particularly relevant when natural language processing tools are used in digital humanities and digital social science: if other researchers want to base their results on automatically analyzed text, we need to know whether the outcome of our analyses reflect the properties of the data correctly, or whether they are the result of biases in training data or random effects in our tools.

In this talk, I will show how this challenge in interdisciplinary data can be used to the advantage of computational linguistic research as addressing it appropriately can provide us new frameworks for evaluation.

Antske Fokkens is Assistant Professor at the Vrije Universiteit Amsterdam. She works in the area of pattern extraction, word embeddings and neural networks against the background of implicit information encoding in language. Her research carries a strong methodological focus, particularly on how Natural Language Processing (NLP) can be applied to Digital Humanities projects such as information identification in historic data.

Peter Buneman

Models of Provenance

As more and more information is available to us on the Internet, the understanding of its provenance – its source and derivation – is essential to the trust we place in that information. Provenance has become especially important to scientific research, which now relies on information that has been repeatedly copied, transformed and annotated. Provenance is also emerging as a topic of interest to many branches of computer science including probabilistic databases, data integration, file synchronization, program debugging and security. Efforts by computer scientists to characterize provenance have resulted in a somewhat bewildering variety of models which, although they have the same general purpose, appear to have little in common.

In this talk, I will attempt to survey these models, to describe why they were developed and to indicate how they can be connected. There has been a particularly rich effort in describing the provenance of data that results from database queries. I shall discuss the impact of this work in areas such as data citation and the potential impact in the social media.

Peter Buneman is Professor in Computer Science at the University of Edinburgh. He is a Fellow of the Royal Society, the Royal Society of Edinburgh and the Association for Computing Machinery. He was also a founder and Associate Director of Research of the UK Digital Curation Centre. His work in computer science has focused mainly on databases and programming languages. He has recently worked on issues associated with scientific databases such as data provenance, archiving, annotation and data citation.

Björn Grüning

FAIR data: The European Galaxy server

Tools and protocols in life-science are highly complex and changing all the times. The mountain of data that is produced is growing exponentially and the only way to scale the data analysis is to enhance its accessibility so everyone can participate in it. In this talk I will talk about the Galaxy project and how a data analysis platform can safeguard scientific results by guaranteeing data provenance, reproducibility and reusability of data.

Björn Grüning is a Bioinformatics scientist at the University of Freiburg and head of its Galaxy team. He also acts as head of the newly founded Steinbeis-Research Center 'Bioinformatics Services Freiburg'. His work aims to provide scientists simple access to data, tools and protocols in the field of Bioinformatics, but also across disciplines by Galaxy, an open source, web-based platform for data intensive biomedical research. In addition he is core-member of the Bioconda, Conda-Forge and BioContainer project.

Christian Bauckhage

Can knowledge inform Machine Learning?

Despite its great success over the past decade, there still are situations where Machine Learning does not or even cannot work well. Especially if training data is limited or not representative or severely biased, state of the art approaches will not be able to learn to generalize well and run the risk of making silly mistakes during application. Crucial questions therefore are if and how "common sense" could be integrated into the machine learning pipeline? Are there mechanisms that allow for informed learning and introspection? Could such approaches provide non-black box solutions whose internal computations are transparent and whose decisions are accountable? In short, are there approaches towards more explainable ML systems that could be deployed in situations where there is few data to learn from and traceable decisions are a necessity? These and similar questions will be addressed in this presentation.

Christian Bauckhage is Professor in Computer Science at the University of Bonn and Lead Scientist for Machine Learning at the Fraunhofer Institute for Intelligent Analysis and Information Systems. His research focuses on theory and practice of AI and machine learning.

Matthias Hein

Robust and reliable machine learning

Machine Learning and in particular, deep learning, is becoming increasingly important for scientific discoveries in complex scientific problems. However, at the moment neural networks are non-robust - small adversarial changes of the input change the decision of a classifier - and non-reliable: far away from the training data neural networks make high-confidence predictions. I will present our work towards provable robustness of neural networks and how to overcome the problem that neural networks don't know when they don't know.

Matthias Hein is Professor in Computer Science at the University of Tübingen and member of the Machine Learning research group. As computer scientist and mathematician, he is also interested in mathematical statistics and optimization. Furthermore, his research includes the application of machine learning to problems in bioinformatics, computer vision and other fields in computer science and the natural sciences.

Kristian Kersting

Towards Reproducibility in Machine Learning and AI

Have you ever tried to stand on another ML/AI researcher's work and not been able to repeat their empirical finding? Most likely, you are not alone. A 2016 survey presented in the journal *Nature*[1] argues that about "70% of researchers have tried and failed to reproduce another scientist's experiments." And reproducing ML and AI results is seldom straightforward either, as noted e.g. by Henderson et al.[2]. Thus, the democratization of ML and AI does not mean dropping the data on everyone's desk and saying, "good luck"! It means making ML and AI methods usable in such a way that people can easily instruct machines to have a "look" at data and help them to understand and act on it.

This is the vision of high-level programming languages for ML and AI. High-level features such as relations, quantifiers, functions, and procedures provide clarity and succinct characterisations of the machine learning problem at hand. What is even more important, high-level descriptions improve the credibility of past and future ML and AI research. By making easier-to-understand code available, researchers can more easily reproduce and verify the results claimed in scientific publications. High-level ML and AI code also makes it easier for engineers to transition academic research to industrial applications. Together with web-based platforms and containerization, it paves the way to creating more easily reproducible, lightweight ML and AI environments.

Kristian Kersting is Professor in Computer Science at the Technical University of Darmstadt and its Centre for Cognitive Science. He is heading the Machine Learning Lab, where various facets of Machine Learning and, more generally, Artificial Intelligence are investigated, both theoretically and reaching out to life science applications. His research deals with the issue of how to make machine learning results understandable (or at least, plausible) for humans, among other issues.

[1] M. Baker: 1,500 scientists lift the lid on reproducibility. Nature, 2016 May 26;533(7604):452-4. doi: 10.1038/533452.
[2] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, D. Meger: Deep Reinforcement Learning That Matters. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI), 2018) at AAAI 2018.

Chandler Becker

Traceability in materials design: A use case from molecular simulation

Materials design is the process of creating new materials that have improved performance over existing materials, hopefully faster than has been done historically by using computation and digital data in addition to experiments. However, for many applications, reliability and trustworthiness of the material and its associated data are essential considerations in determining whether a material will be used in a real product that has performance and safety requirements. We will discuss how a relatively focused problem in materials design, selecting a model for how atoms interact in a simulation, becomes representative of larger discussions around digital data, scientific traceability, and research confidence.

Chandler Becker is a research scientist at the U.S. National Institute of Standards and Technology, where she works in the Material Measurement Laboratory's Office of Data and Informatics. Her activities are focused in the space where materials science and data science overlap, particularly as related to how people can reliably use materials data and results generated from it. As Informatics and Analytics Lead, Dr. Becker is, among other things, interested in the development of tools and methods to facilitate sharing and re-use of experimental and simulation data occurring in materials science.

Jörg Neugebauer

Automizing work flows in computational materials design

Jörg Neugebauer is Director at the Max-Planck-Institut für Eisenforschung in Düsseldorf and head of its Computational Materials Design Department. He has strongly influenced the field of materials modelling with a new generation of simulation techniques, working ab initio, i.e., starting with the fundamental laws of quantum physics and chemistry. Applying these laws to span a wide range of different scales is not immediate; his contributions have been a significant step towards materials analysis solely by means of simulations.

Michael Esfeld

What is a measurement record?

In this talk, I'll argue for the following three claims:

(1) All measurement records are classical phenomena that are recorded as the positions of discrete objects and that are records of positions of discrete objects. In other words, measurements records have to be classical in contrast to quantum phenomena.

(2) The records concern not single data, but phenomena: these are salient patterns of data. What is recorded are regularities in the distribution and evolution of objects. Phenomena in this sense are the basis for laws and explanations in science.

(3) All records provide information about the world only as interpreted records.

I will illustrate (1) and (2) notably by examples from quantum physics and draw on Wittgenstein on rule following to establish (3).

Michael Esfeld is Full Professor of Philosophy of Science at the University of Lausanne since 2002. His main areas of research are the metaphysics of science, in particular physics, and the philosophy of mind and language.

Martin Erdmann

Mastering complex data processing procedures: from particle detector measurements via machine learning algorithms to physics results

Experimental physics at large-scale research facilities has always been confronted with large data volumes. Here, physics laws are extracted from a high image rate. Thereby image figuratively stands for all instances of measurement data. For new discoveries, individual processes have to be filtered from the many measurement data. This requires algorithms, compute power, storage media, whereby the exabyte scale of storage has long been reached. In the digitization era, we expect new challenges from at least one order of magnitude more data, more responsible use of data life cycles, and by machine learning. A portfolio of actions will be necessary to successfully bring the German research landscape into this new phase of digitization. This contribution describes the research field and its challenges as well as a portfolio of recommended measures. They range from human resources for coping with the multiple challenges through hardware investments to the adaptation of curricula of physics at the university.

Martin Erdmann is Professor for High Energy Physics at RWTH Aachen. He is the current head of the special interest group for artificial intelligence in the German Physical Society. He is involved in major (elementary particle and astroparticle physics) experiments like CMS at the Large Hadron Collider (LHC) at CERN. Through his work with complex data sets from these experiments, he has recently turned to machine learning techniques, notably the application of deep learning to data analysis problems in physics.