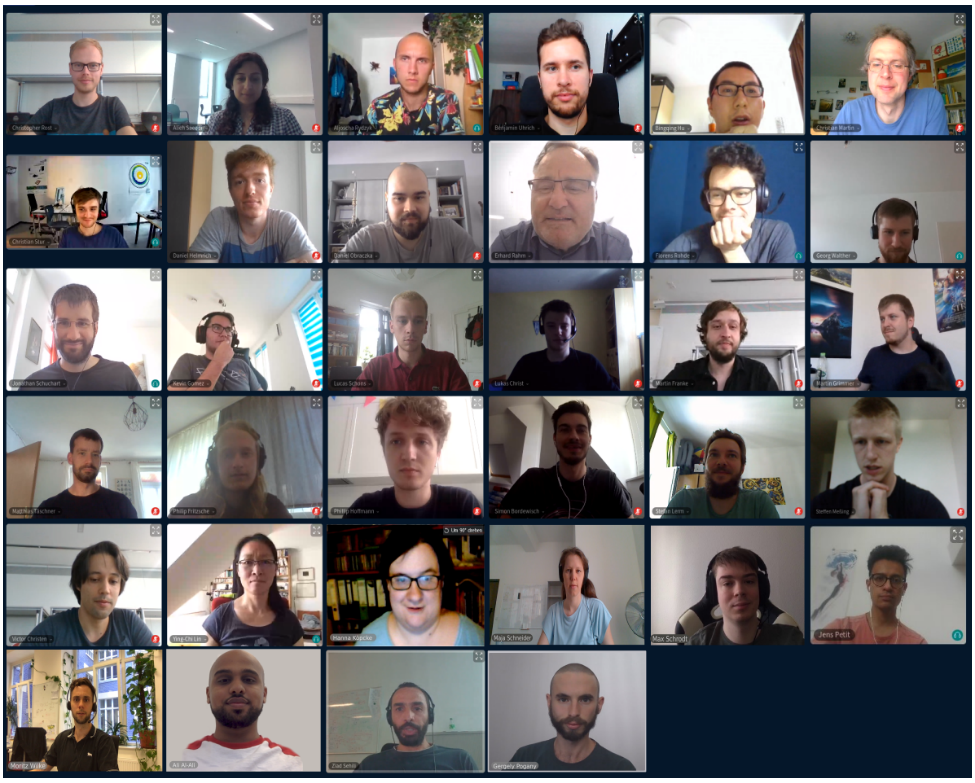
Database Group Leipzig
Department of Computer Science

UNIVERSITÄT LEIPZIG

Research Report 2020/2021

https://dbs.uni-leipzig.de

Overview



Database Group in June 2020. r.t.l.: Christopher Rost, Dr. Alieh Saeedi, Aljoscha Rydzyk, Benjamin Uhrich, Bingqing Hu, Dr. Christian Martin, Christian Stur, Daniel Helmrich, Daniel Obraczka, Prof. Dr. Erhard Rahm, Florens Rohde, Georg Walther, Jonathan Schuchart, Kevin Gomez, Lucas Schons, Lukas Christ, Martin Franke, Martin Grimmer, Matthias Täschner, Philip Fritzsche, Philip Hoffmann, Simon Bordewisch, Stefan Lerm, Steffen Meßing, Dr. Victor Christen, Dr. Ying-Chi Lin, Dr. Hanna Köpcke, Maja Schneider, Max Schrodt, Jens Petit, Moritz Wilke, Ali Al-Ali, Ziad Sehili, Gergely Pogany

1 Staff

Prof. Dr. Rahm, Erhard	Professor
Hesse, Andrea	Secretary
Alkamel, Abdalrahman (May 2020 - April 2021)	Research associate
Alkhouri, Georges (until Feb. 2020)	Research associate (BMWI)
Prof. Dr. Christen, Peter (Apr. 2021 - Sept. 2021)	Leibniz Professorship
Dr. Christen, Victor	Postdoctoral Researcher
Franke, Martin	Research associate
Gomez, Kevin	Research associate
Grimmer, Martin	Research associate (BMBF)
Hannemann, Anika (since Dec. 2021)	Research associate
Hofer, Marvin (since April 2021)	Research associate
Leipnitz, Alexander (since Jan. 2021)	Research associate
Dr. Lin, Ying-Chi	Research associate (DFG)
Dr. Köpcke, Hanna (since April 2020)	Postdoctoral Researcher
Kramm, Aruscha (since Dec. 2021)	Research associate
Kreusch, Jonas (since Jan. 2021)	Research associate
Dr. Martin, Christian	Postdoctoral Researcher
Neumann, Anja (since Jan. 2021)	Research associate
Obraczka, Daniel	Research associate
Petit, Jens (May 2020 - Jan. 2021)	Research associate
Pogany, Gergely	Research associate
Dr. Peukert, Eric	Postdoctoral Researcher (BMBF)
Rohde, Florens	Research associate
Rost, Christopher	Research associate
Dr. Saeedi, Alieh	Postdoctoral Researcher
Schneider, Maja (since April 2020)	Research associate
Schuchart, Jonathan	Research associate
Sehili, Ziad	Research associate
Täschner, Matthias	Research associate
Uhrich, Benjamin (since April 2020)	Research associate
Wilke, Moritz	Research associate
Prof. Dr. Thor, Andreas (HfTL Leipzig)	Associated team member
Dr. Burghardt, Thomas	Postdoctoral Researcher (since July 2020)

2 Highlights

There have been several highlights in 2020 and 2021:

1. Prof. Rahm has initiated a new master degree course of study on Data Science that started in April 2020. By the end of 2021 almost 100 students are enrolled in this program.
2. The AI and data science center ScaDS.AI, co-directed by Prof. Rahm, organized its 6th international summer school in July 2020. For the first time it took place in virtual format due to the Covid pandemic.
3. ScADS.AI Leipzig moved in Dec. 2020 to a new building in Humboldstr. 25 with a capacity of about 2500 square meters and 110 work places. Its new living lab as well as the graduate school were kicked off in a ceremony on Oct. 7, 2021, together with the science minister of Saxony, Sebastian Gemkow.

Opening Ceremony of ScaDS.AI Living Lab and Graduate School with Sebastian Gemkow.

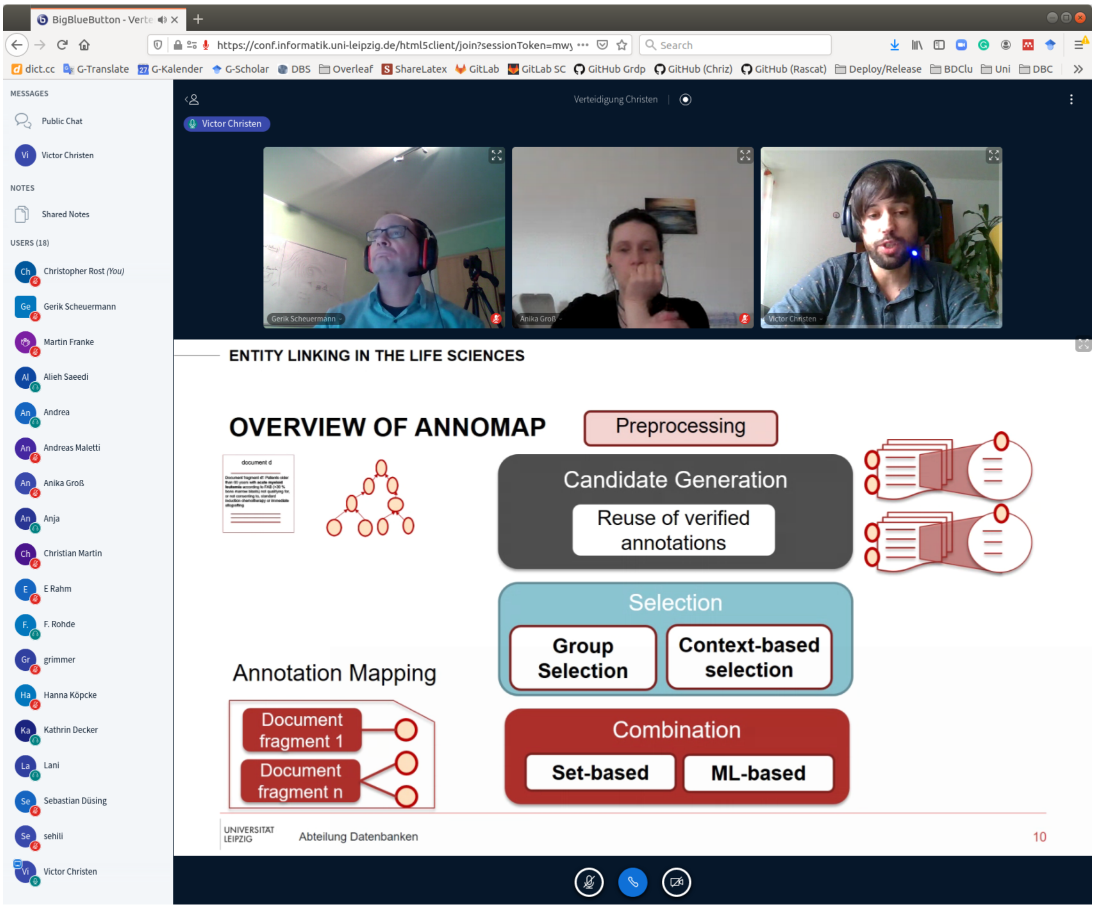
4. Prof. Peter Christen (ANU) has been named the Leibniz Professor of the University of Leipzig in the summer semester 2021 and has visited the database group and ScaDS.AI during this time.
5. Markus Nentwig, Victor Christen and Alieh Saeedi successfully defended their Ph.D. theses.
6. The paper "Enhancing Cross-lingual Semantic Annotations Using Deep Network Sentence Embeddings" has won the Best Paper Award in the HEALTHINF 2021 (BIOSTEC) conference.
7. Prof. Rahm was elected Vice President of the German Informatics Society (Gesellschaft für Informatik e.V.)
8. The application for permanent funding of ScaDS.AI has been positively evaluated in the second half of 2021. The institutional phase will start in July 2022.
9. The DFG project ELISA has been successfully ended in 2021. Funding for several new projects could be secured: IOTTest, AMPL, CUT and K-M-I. Furthermore we are a cooperation partner in the new Saxocell consortium.

Data, COVID, and LSD
(or how LSD can help combat COVID-19)
Peter Christen
Leibniz Professor, University of Leipzig
Professor, School of Computing,
College of Engineering and Computer Science,
The Australian National University, Canberra
Contact: peter.christen@anu.edu.au
Homepage: http://users.cecs.anu.edu.au/~christ

Inaugural Lecture of Leibniz professor Peter Christen with Rector Beate Schücking (June 2021).

Defense of Markus Nentwig.

Handing over of the doctoral hat of Alieh Saeedi.



MESSAGES

Public Chat

Victor Christen

NOTES

Shared Notes

USERS (18)

Christopher Rost *(You)*

Gerik Scheuermann

Martin Franke

Alieh Saeedi

Andrea

Andreas Maletti

Anika Groß

Anja

Christian Martin

E Rahm

F. Rohde

grimmer

Hanna Köpcke

Kathrin Decker

Lani

Sebastian Düsing

sehili

Victor Christen

Verteidigung Christen

Victor Christen

Gerik Scheuermann

Anika Groß

Victor Christen

ENTITY LINKING IN THE LIFE SCIENCES

OVERVIEW OF ANNOMAP

Preprocessing

Candidate Generation

Reuse of verified annotations

Selection

Group Selection

Context-based selection

Combination

Set-based

ML-based

Annotation Mapping

Document fragment 1

Document fragment n

UNIVERSITAT LEIPZIG

Abteilung Datenbanken

10

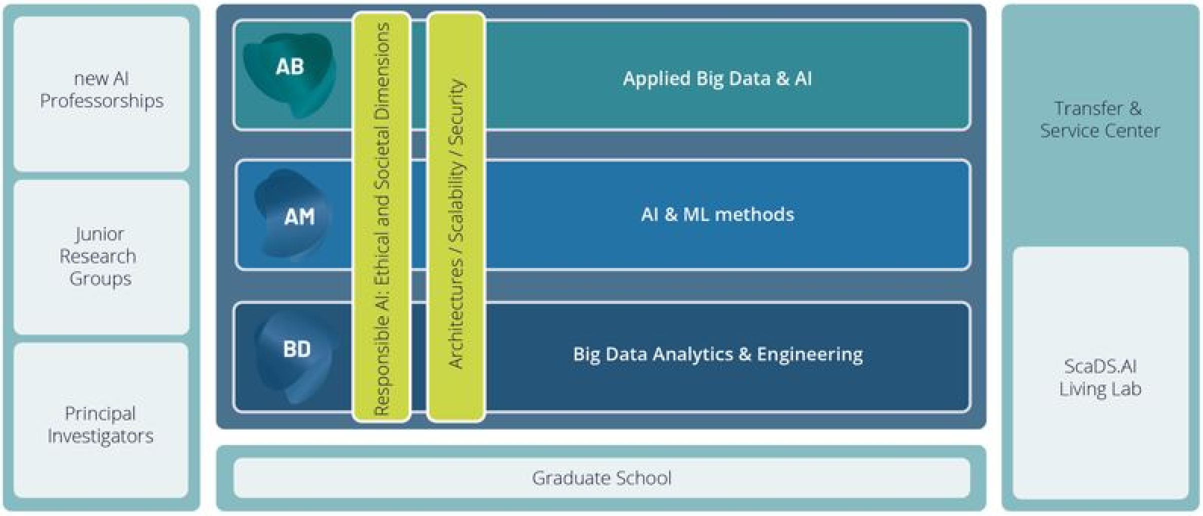
Defense of Victor Christen.

3 Research Topics and Projects

ScaDS.AI - Center for Scalable Data Analytics and Artifical Intelligence

E. Peukert, C. Martin, D. Obraczka, J. Schuchart, M. Täschner, M. Wilke, E. Rahm

ScaDS.AI (Center for Scalable Data Analytics and Artificial Intelligence) Dresden/Leipzig lead by Prof. Nagel from the TU Dresden and Prof. Rahm from the University of Leipzig is one of the German centers for artificial intelligence (AI), which is funded as part of the federal government's AI strategy by the Federal Ministry of Education and Research (FKZ: 01IS18026B) and the Free State of Saxony with the establishment of 4 new AI professorships at both locations. The research is running at two locations, Dresden and Leipzig, by the partners Dresden University of Technology, Leipzig University, Max Planck Institute for Molecular Cell Biology and Genetics, Leibniz Institute for Ecological Spatial Planning, Helmholtz Center for Environmental Research, Leipzig and the Helmholtz Center Dresden Rossendorf.ScaDS.AI Dresden/Leipzig has become a success story at the University of Leipzig and its partner institutions: ScaDS.AI moved in Dec. 2020 to a new building in Humboldstr. 25 with a capacity of about 2500 square meters and 110 work places. Due to the positive evaluation of the application for continuation by the federal and state governments the structure of ScaDS.AI will be expanded for another 7 years, collaborations will be deepened and new Principal Investigators will be recruited.


new AI Professorships
Junior Research Groups
Principal Investigators
AB
AM
BD
Responsible AI: Ethical and Societal Dimensions
Architectures / Scalability / Security
Applied Big Data & AI
AI & ML methods
Big Data Analytics & Engineering
Graduate School
Transfer & Service Center
ScaDS.AI Living Lab

Gross structure of ScaDS.AI including main research areas (middle part).

Main research areas

ScaDS.AI investigates the need of AI applications for high quality data and formalized knowledge to achieve valid and reliable prediction and analytical results. Therefore it combines research on new Fundamental AI methods not only with Big Data research on data integration and data quality, but also with new methods for data acquisition and visualization to support data-driven AI. In addition, AI methods need to be systematically integrated into scientific analysis workflows, which can accelerate research progress in many areas. In addition, trust, transparency, and traceability of AI-driven decisions and processes are key. Finally, privacy and informational self-determination remain largely unresolved issues that we will tackle with research on privacy-preserving machine learning.

Transfer

ScaDS.AI Dresden/Leipzig fosters the fast and efficient transfer of research results into industry. In an increasing number of cooperation projects with companies, state-of-the-art AI and Data Science methods are put into practice. Therefore, the innovation power and competitiveness of the participating companies, and thus Saxony as a whole as a center of business, are strengthened. The Transfer and Service Center is an integral part of ScaDS.AI in this regard. The employees identify relevant solutions and concepts from the research areas of the center and address research requests from industry partners in the area of Big Data and AI in joint pilot projects. Furthermore, in addition to scientific consulting, the service center offers training in the area of AI, Big Data, and High Performance Computing (HPC) and enables partners the use of AI and HPC resources.

Living Lab – Science Communication

On Oct. 7 2021, the new Living Lab at ScaDS.AI Leipzig was kicked off in a ceremony together with the science minister of Saxony, Sebastian Gemkow. The Living Lab will approach the main research topics of Artificial Intelligence, Big Data, and Data Science in a diverse and vibrant way. With the Living Lab experiment space, a user-centered ecosystem is being created that is designed to share current and topical research approaches and results of our center with visitors. The Living Lab appears as a multi-faceted venue, exhibition space, teaching and education center, and laboratory. It is a vehicle for communication, discovery, development, and evaluation of our research for the public. Practical and everyday problems as well as questions in dealing with new digital technologies shall be in the focus as well as finding innovative approaches for business and industry. New methods and research results need to be communicated with and evaluated by our visitors from different target groups, so that diverse opinions can be immediately included in our research and innovation process.

Graduate Qualification

On Oct. 7, 2021 the graduate school, lead by Jun.-Prof. Potthast, was kicked off in a ceremony together with the science minister of Saxony, Sebastian Gemkow. The graduate school contains among other things a qualification programme in cooperation with the Research Academy Leipzig, a selective recruitment process for PhD positions, an incorporation of running PhD-projects into the school as well as a "Book Club", a self-organized reading group (topics: linear algebra, Bayesian statistics, machine learning fundamentals, . . .). Despite the numerous challenges posed by the global pandemic, the graduate school and structured qualification programme operate according to the defined targets. At the end of the report period, in 12 / 2021, the number of Leipzig-based graduates in the graduate school, including interested prospective members that are actively participating in the joint activities, amounts to 30.

Masters' program Data Science

In April 2020 started the consecutive application and research-oriented Master's program, Data Science in cooperation with ScaDS.AI and the Institute of Computer Science at the University of Leipzig which was initiated by Prof. Rahm. The program aims to provide a scientifically founded, application-oriented education in the essential Data Science areas of scalable Data Management ("Big Data") and Data Analysis based on Data Mining and Machine Learning methods. The professorships involved in the ScaDS.AI research center play a major role in shaping the teaching of the Data Science course, thus enabling students to participate in current research projects.

Bitemporal Property Graphs to Organize Evolving Systems

C. Rost, K. Gomez, P. Fritzsche, L. Schons, T.Adameit, E. Rahm

The analysis of highly connected data as graphs becomes more and more important in many different domains. Prominent examples are social networks, e.g., Facebook and Twitter, as well as information networks like the World Wide Web or biological networks. One important similarity of these domain specific data is their inherent graph structure which makes them eligible for analytics using graph algorithms. In addition, datasets have other common features: 1) they are very large, making it difficult or even impossible to process them on a single machine, 2) they are heterogeneous in terms of the objects they represent and the data associated with them, and 3) they evolve over time, i.e., new entities, relationships, or properties are added or existing ones are modified or deleted. With the objective of analyzing these large-scale, heterogeneous and dynamic graphs, we continue developing a framework called "Gradoop" (Graph Analytics on Hadoop). Gradoop is built around the so called Temporal Property Graph Model (TPGM) which supports not only single but also collections of heterogeneous and temporal graphs and includes a wide range of combinable operators. These operators allow the definition of complex analytical programs as they take single temporal graphs or graph collections as input and result in either of those. Gradoop is build on top of the distributed dataflow framework Apache Flink, and makes use of the provided APIs to implement the TPGM and its operators. The system is publicly available (www.gradoop.com) and gets code contributions from other institutes and companies. A demo application, namely the Temporal Graph Explorer, showing the usage and resultset of three selected temporal graph operators: snapshot, difference and time-dependent grouping. The application is open-source available (`https://github.com/dbs-leipzig/temporal_graph_explorer`) and the corresponding articles published.

In 2020, a one-year cooperation has started with the industry partner Oracle Labs, a research and development organization within Oracle Inc., USA. The project "Temporal Property Graphs as Organizing Principles", funded by Oracle Labs, has focused on the development of a native property graph store with bi-temporal time management and querying via a graph query language like PGQL. Such a system can be used, for example, to organize multi-dimensional time-series of correlating sensor data as a graph structure. A prototypical implementation was created as a proof of concept and for experimental evaluation. The existing graph query language PGQL was extended to define temporal graph patterns, e.g. to find subgraphs that happened in the past or within a period of time. Not only the analysis of the graph history is in the foreground, but also queries that consider the continuous changes of the graph. The research results were published at the end of the project.

In our ongoing work, we continue focusing on the processing and analysis of temporal graphs and further graph streams, which represent the continuous addition and removal of vertices and edges as well as frequent changes in their attributes. We already started with the development of a data model for graph streams as well as some initial query methods and analytical operators. Searching and discovering temporal patterns within a graph stream or summarizing graph structures in a windowed form are typical examples of analyses that focus on the continuous representation of a graph as a stream.

Distributed Large-Scale Graph Data Integration

A. Saeedi, D. Obraczka, J. Schuchart, M. Hofer, E. Peukert, E. Rahm

With the rising complexity and interrelation of information knowledge graphs have become a popular data structure to organize knowledge. These graphs are used in a variety of tasks such as e.g. Question Answering or Recommendation Systems. In order to gain a holistic view of their data organizations and companies need to integrate information from a variety of different sources. This makes high-quality data integration tools a crucial precursor to any data analytics task. With the vast amount of data produced nowadays these tools need to be scalable as well.

While the conceptual flexibility of knowledge graphs makes them easily extendable this comes at a price in the data integration process. Conventional entity resolution systems are mainly build for tabular structures and are not easily appliable for integration of knowledge graphs. A recent trend in research hase therefore been the investigation of knowledge graph embeddings, that transform nodes in the graph into a lower-dimensional vector. These embeddings are constructed in such a way that similar entities from different data sources are close in the embedding space. In the EAGER approach we investigated whether combining such entity embeddings with attribute similarities as input for machine learning algorithms such as Multi-Layer-Perceptrons improves the alignment results, when compared to either using the embeddings or the attribute similarities alone. Our results suggest, that the combination of embeddings and attribute similarities significantly outperforms using either on their own, provided that the graphs that should be matched have a dense enough connectivity. The aforementioned entity embeddings usually lie in spaces with high-dimensionality (¿ 50). This comes with certain problems since high-dimensional spaces tend to suffer from a phenomenon known as "hubness". Data that suffers from hubness has some points that are nearest neighbors to most other points, while many points are nearest neighbors to no one. Since entity resolution with knowledge graph embeddings relies on nearest neighbor computations a high degree of hubness is detrimental to match quality. We therefore investigated several hubness reduction techniques in combination with (approximate) nearest neighbor approaches to find that hubness reduction with approximate nearest neighbor search tends to give significantly more accurate and faster results, than using no hubness reduction with exact nearest neighbor search. To make these results practicably usable the techniques are available in the open-source python library *kiez*.

DE4L - Data Economy for advanced Logistics

M. Schneider, J. Kreusch, F. Rohde, E. Peukert, E. Rahm

The DE4L project is pursuing the development of an intelligent ecosystem as part of a platform for data exchange for logistics service companies. This is to avoid high congestion on delivery vehicles, costs due to incorrect delivery and repeated delivery and pickup attempts. The so-called "last mile" of the supply chain, meaning the exact delivery and collection of parcels at the front door, offers a great deal of potential for increasing efficiency. With the platform DE4L strengthens the cooperation of the service companies and promotes the digitization of the information.

DE4L is a BMWI-funded cooperation (FKZ: 01MD19008D) with different partners from the logistics domain, the Fraunhofer IML and the data science center ScaDS.AI.

We are developing privacy-preserving methods that are applied while collecting sensor data to protect the privacy of drivers, e.g. in sensitive areas. Furthermore we are building tools that help data owners to investigate, visualise and assess the privacy risk and to choose appropriate privacy-enhancing techniques before sharing/selling their data.

Additionally, we are building an innovative Blockchain/Distributed Ledger-based trading platform for address-related data from logistics such as opening hours, preferred delivery locations etc. To protect the potentially sensitive address data we designed data flows based on privacy-preserving record linkage to assign pseudonymous global ids to addresses which are used when offering and searching data on the trading platform.

The project started in August 2019 and is running for three years.

TWIN - Transformation of complex product development processes into knowledge-based services for additive manufacturing

E. Peukert, E. Rahm

In TWIN, development processes of laser-based generative manufacturing (metal and plastic) and additive manufacturing processes are to be digitilized. For this purpose, TWIN is developing a digital product service system (with a digital twin as the core object) with the participation of the entire value chain of industrial additive manufacturing. Particular emphasis is placed on support for using machine learning processes.

The University of Leipzig is concerned with two main areas in the project: (1) integration and storage of heterogeneous sensor and process data as well as (2) the subsequent analysis and modeling. Data integration and analysis is implemented as an iterative process, i.e. the digital twin in this project progresses gradually and is expanded to include data sources and models.

The project is funded by the BMWi (FKZ: 02K18D055) and started in October 2019. It will run for three years.

GRAMMY - InteGRAtive analysis of tuMor, Microenvironment, immunitY and patient expectation for personalized response prediction in Gastric Cancer

G. Pogany, C. Martin, E. Rahm

Gastric cancer (CG) is a complex disease, the fifth most common malignant tumor in the world and the third leading cause of death from cancer. CG is very heterogeneous and affects twice as many men as women. Chemotherapy combined with surgery represents the standard of care for stage II to III CG, but the efficacy of such treatments is still limited for many patients. It is therefore imperative to develop an innovative approach aimed at identifying new predictive markers, including those deduced from taking into account the impact of the psychosocial and cultural environment of each patient. We defend the idea that the style of communication, the degree of acceptance of the treatment by the patient, as well as the doctor-patient interaction, can influence the response to treatment, with in particular differences in compliance. The integration of different levels of information, biological and psychosocial, is very promising, although it is particularly difficult, to identify the links between the specific biological characteristics of the disease, the patient's perception and the prognosis. The consortium consists of an number of European partners from Italy (lead), Greece and France.

The "GRAMMY" project is funded by the European ERA PerMed call (Antragsnummer-SAB: 100394103) and will run for 3 years until 2022/23. The database group is responsible for the data integration and is also supporting the analyses of heterogeneous medical data sources of the project.

MitSystemZumErfolg – IoTTest / Anomaliebasierte Angriffserkennung auf daten- und kontrollflussbasierten Sensoren

M. Grimmer, E. Rahm

The widespread networking of electronic devices offers industry the potential for new, innovative products and services. However, the accompanying increase in the attack surface poses new challenges. The increasing complexity of application scenarios significantly increases the effort required to secure the underlying systems against security vulnerabilities. At the same time, the time for adequate security tests is decreasing due to ever shorter release cycles and the closer integration of development and operation (DevOps). This makes tool-supported security testing necessary. A fully automated solution is being developed for this purpose in the project. This comprises the identification of test targets, test case generation and test evaluation, including the generation of reports. The starting point is attacks on IoT applications identified in the field, which are varied using genetic algorithms in order to identify similar vulnerabilities. This not only supports the efficient development of security patches, but also advances protection against previously unknown attacks. Automation brings an enormous time and cost advantage over manual investigations. In 20/21, research was conducted on anomaly detection algorithms and a paper was published on this topic.

ELISA - Evolution of Semantic Annotations

Y.-C. Lin, V. Christen, E. Rahm

Annotating documents or datasets using concepts of biomedical ontologies has become increasingly important. Such ontology-based semantic annotations can improve the interoperability and the quality of data integration in health care practice and biomedical research. For instance, PubMed, the search engine for MEDLINE database, uses MeSH (Medical Subject Headings) terms to retrieve more relevant results. The use of the hierarchy information within the ontologies can further expand the potential matches. Furthermore, annotating data across multiple disconnected databases using concepts from same ontologies enables data integration.

With the development of the medical knowledge, the ontologies are changing continuously. On the other hand, the documents to be annotated, such as medical forms, can also be revised into different versions or adapted into different languages. The ELISA (Evolution of Semantic Annotations) project aims to investigate the impacts of such changes, in both ontologies and the documents, on the semantic annotations. The project is a cooperation with the Luxembourg Institute of Science and Technology (LIST), the University of Paris-Sud.

We designed and implemented a (semi-)automatic approach to insure the validation of the semantic annotations when the underlying ontology is evolving. The maintenance framework considers rules that exploit the morphosyntactic form of terms denoting attribute values, such as split or merge. Secondly, it also includes further background knowledge such as additional biomedical terminologies to determine the correct update of the annotation. Finally, the framework adapts the new annotation using Semantic Change Patterns that regards the lexical and semantic similarities of the terms.

In our previous work, we investigated cross-lingual methods for annotating German forms. The approach utilized online translators to translate the questions of a form to English. The resulting translations are annotated with concepts from an English ontology. In our current work, we propose an annotation method utilizing pre-trained language models such as BERT, RoBERTa, etc. The approach determines for each question of a form and each concept a sentence embedding being compared using the cosine similarity. Our method improves the results compared to the translator-based annotation strategy.

Privacy-preserving Record Linkage

Z. Sehili, M. Franke, F. Rohde, V. Christen, E. Rahm

Record linkage aims at linking records that refer to the same real-world entity, such as persons. Typically, there is a lack of global identifiers, therefore the linkage can only be achieved by comparing available quasi-identifiers, such as name, address or date of birth. However, in many cases, data owners are only willing or allowed to provide their data for such data integration if there is sufficient protection of sensitive information to ensure the privacy of persons, such as patients or customers. Privacy-preserving Record Linkage (PPRL) addresses this problem by providing techniques to securely encode and match records. By combining data from different sources data analysis and research can be improved significantly. The linkage of person-related records is based on encoded quasi-identifiers while the data needed for analysis, e.g., health data, is excluded from the linkage.

PPRL is confronted with several challenges needing to be solved to ensure its practical applicability. In particular, a high degree of privacy has to be ensured by suitable encoding of sensitive data and organizational structures, such as the use of a trusted linkage unit. PPRL must achieve a high linkage quality by avoiding false or missing matches. Furthermore, a high efficiency with fast linkage time and scalability to large data volumes are needed. A main problem for performance is the inherent quadratic complexity of the linkage problem when every record of the first source is compared with every record of the second source. For better efficiency, the number of comparisons can be reduced by adopting blocking or filtering approaches. Furthermore, the matching can be performed in parallel on multiple processing nodes.

Multi-Party PPRL (MP-PPRL) introduces further challenges to be addressed. In particular, the number of record comparisons grows quadratically with the size and the number of sources making scalability a problem. Furthermore, a record may have matches in an arbitrary subset of the data sources not only in one data source. This asks for clustering matching records over multiple sources so that a cluster contains all matches for a specific person. This clustering should utilize that individual sources are often curated and duplicate-free so that every cluster should have at most one record for any data source. To address these challenges lately, we investigate several novel approaches for MP-PPRL and clustering of encoded records. First, we proposed an extension of the pivot-based metric space approach with a dynamic adaptation of the pivot selection. A dynamic pivot selection allows to deal with additional data sources and growing data volume. Besides, we investigated different clustering schemes for multiple parties that either cluster new data sources one after the other (early clustering) or that first determine similar record pairs over all sources before a final clustering is performed (late clustering). An extensive evaluation showed the high scalability and good quality of Max-Both as an early clustering method and SKB-S (Sort-Keep-Best using avg. Similarity) as a late clustering method.

Bloom Filter Hardening Techniques Recent PPRL approaches mainly focus on encoding techniques utilizing Bloom Filters as error-tolerant and privacy-preserving method to encode records containing sensitive information. While Bloom-filter-based encodings have become the quasi-standard in PPRL approaches, several studies analyzed weaknesses and implemented successful attacks on Bloom filters. In general, it was observed that Bloom filters carry a non-negligible re-identification risk because they are vulnerable to frequency-based cryptanalysis. In order to prevent such attacks, various Bloom filter hardening techniques were proposed. Such techniques aim at reducing patterns and frequency information that can be obtained by analyzing the frequency of individual Bloom filters or (co-occurring) 1-bits.

We comprehensively reviewed hardening techniques proposed in the literature and evaluated their effectiveness in terms of achieving high privacy (security) and linkage quality. We therefore also proposed and analyzed measures that allow to quantify the privacy properties of different Bloom filter variants. These measures are based solely on a set of Bloom filters and do not need any reference dataset or other information. Our evaluation showed that multiple hardening techniques drastically reduce linkage quality and are thus not suitable in real-world applications. However, hardening techniques using salting and xor-folding can drastically reduce any frequency information while maintaining high linkage quality. As a consequence, these techniques will make any frequency-based cryptanalysis very unlikely to be successful.

Furthermore, we investigated hardening techniques based on autoencoders. The aim is to prevent frequency-based cryptanalysis by mapping Bloom filters to a continuous vector space. For guaranteeing the comparability of different trained autoencoders, we proposed a mapping method to transform the encoded Bloom filters from own data owner to the space of the vector space of the other data owner. The evaluation showed that we were able to achieve comparable results in terms of linkage quality and to decrease the risk of privacy attacks.

Attacks on PPRL The development of attacks supports to assess if an encoding or hardening technique is secure or not. The goal of an attack is to reveal clear-text values from the encoded records. In previous work, the similarity graphs of clear text values and Bloom filters were used to generate graph-based features. The alignment method computes the similarity between clear text records and Bloom filter records utilizing the features. In addition to the graph-based alignment approach, we investigated methods for generating embeddings for nodes of both graphs. The aim is to increase contextual information of each record so that the alignment of clear-text value records and encoded records is more precise.

PRIMAT In 2019 we demonstrated the first version of our PPRL toolbox, named PRIMAT, at the VLDB. Since then we extensively refactored and extended the initial codebase to simplify usage and to improve extensibility and maintainability. Starting in December 2021, we will incrementally add new features related to our ongoing research and continuously release new versions of PRIMAT. Most recently, we added new components for incremental matching and analysis utilities, e.g. to analyze records and error types.

PPRL in Practice One application area for PPRL is medical research, since the investigation of many scientific questions is only possible by merging distributed patient data where privacy and data protection are essential requirements as medical information is very sensitive personal data. We therefore contribute our experience to the SMITH consortium within the Medical Informatics Initiative in order to build an infrastructure that can be used in various data linkage projects. We also contributed to an upcoming NFDI4Health white paper on methods and best practises in (privacy-preserving) record linkage.

In 2019 we conducted an evaluation of the record linkage facilities of the so-called Mainzelliste, an open-source software for identity and pseudonym management of patients which is used in various medical projects. In 2020 we successfully published our evaluation results and the performance improvements we contributed to the open-source project.

Future Research Directions Together with Prof. Dr. Peter Christen we start working on several joint research projects. First, we continue to focus on privacy aspects of currently used encoding techniques for PPRL. Therefore, we designed a new iterative graph-based attack to break Bloom-Filter-based encodings. Furthermore, we want to refine our previously proposed privacy measures that rely on the uniformity of the encoding, e.g. 1-bit distribution of Bloom filters. Since uniformity does not imply fully pattern-less outcomes, more sophisticated privacy measures are needed. Ideally, a perfect encoding should produce outputs that are indistinguishable from fully random sequences.

Another aim is to develop linkage protocols that use multiple rounds where more flexible field-level encoding techniques are used in later stages to make better decisions on uncertain match candidates. A high level of privacy protection is maintained as only a small share of records is using these encodings making attacks based on frequency alignment very difficult.

VIP – Visual Product Matching

M. Wilke, E. Peukert, E. Rahm

A very useful application of record linkage techniques is the growing field of e-commerce. Linking product offers from different vendors allows to compare these and to gain valuable insight into the market. Unfortunately data from the web is very heterogeneous and not easy to integrate. E.g. an equal product can be described with a very differing level of detail, different attributes, description text and so on. Worse yet the description of two non-equal products can be identical (often when its verbose).

Recording linkages approaches as of today are largely based on columnar and textual data. However, online product data typically also consists of images and in some domains (e.g. fashion) this visual information is much more reliable and relevant for the user. The goal of the VIP project is to explore whether the additional information provided by the images can be used to improve the results of existing matching systems. To achieve this, a variety of image similarity metrics from computer vision and deep learning shall be investigated. Furthermore it is to examine how the image matching approaches can be integrated into current record linkage systems to allow the matching of heterogeneous, multi-modal data.

The VIP project started in October 2019, it is a joint project with the company Web Data Solutions. It is funded by the Sächsische Aufbaubank (SAB)

Connected Urban Twin

E. Peukert, A. Kramm, E. Rahm

The „Connected Urban Twin“ project, short CUT, started in 2021 and has a duration of 5 years. It assembles members of the administration of the three German cities Leipzig, Hamburg and Munich along with various research partners namely HafenCity University/ CityScienceLab Hamburg, the TU Munich or ScaDS.AI (Leipzig University). While the overall topic of the project can be labelled as smart city and the main objective is to create a digital twin of the cities, this objective can be split into smaller and more detailed objectives, that are intended to be achieved on different levels: The project is divided into measures ("Maßnahmen") called M1 to M5 and the supervision of these lies within different cities. The task of M1 is the administration of the digital twin, including writing a requirement catalog and planning the twin's architecture. M2's aim is to find use cases to show how technology and smart data can be used for decision making and governance within the cities. How to involve the citizens in planning processes is a task that M3 is working on. M5's task is to make the generated knowledge public and share it with other cities and citizens. M4 is the measure that ScaDS.AI is a part of. Under the headline "experimental & transformative research", M4 is working together with the cities and their data, to extract new knowledge from the data using artificial intelligence and simulations. A big focus is lying on scalability and transferability to be able to deploy developed methods and procedures to other regions or cities. Ongoing M4-projects are the usage of VR to simulate traffic situations, including sensor technology into city models, e.g. the integration of sensors on traffic lights. Within the ScaDS.AI, we are currently working together with the city of Leipzig on two datasets: data from the local transport service (LVB) and image data of the city similar to Google Street View. The image data is used to generate knowledge of the buildings the city does not yet have, such as the number of stories or the level of restoration. Contact persons within the ScaDS.AI for the CUT project are Eric Peukert and Aruscha Kramm.

AMPL - Automatic Meta Data Profiling and Lineage for Integrating Heterogeneous Data Sources

M. Täschner, E. Rahm, M. Miazga, D. Abitz

Efficiently managing and merging many heterogeneous, dynamic data sources has become a critical success factor for financial institutions. However, with increasing heterogeneity and dynamic data, it is becoming increasingly difficult to keep track of historically collected and exponentially growing data pots. This has already led to significant macroeconomic damage, including the global financial crisis of 2007 and 2008, the scale of which could have been contained with real-time transparency and thus a better overview of risk and metadata. Unfortunately, there is currently no solution for financial institutions that allows flexible integration of heterogeneous data sources while providing intuitive metadata preparation. AMPL aims to develop a new tool for structuring, analyzing, and exploring large volumes of heterogeneous, dynamic data sources. For this purpose, the tool computes comprehensive data profiles consisting of statistics, correlations and complex provenance information (lineage). Machine learning assisted methods help in schema mapping (schema matching, ontology matching) between data sources as well as new methods for scalable and incremental computation of data profiles. These will be developed based on current preliminary work of the project partners and recent research results in the area of graph analysis, SQL-based data integration and incremental record linkage (entity resolution) on dynamic and heterogeneous data sources. The data profiles are then presented in a novel web-based visual front-end that greatly simplifies data interaction and exploration. By breaking down existing silos and merging innovative technologies with the requirements of market participants, AMPL thus allows to completely rethink data and metadata management. The AMPL project is funded by the BMBF (Funding reference: 01IS20084B) and will run for 30 months from 01/2021 till 06/2023.

K-M-I (Artificial and Human Intelligent)

E. Rahm, M. Täschner, C. Augenstein (IWI)

The competence center K-M-I (Artificial and Human Intelligent) connects industry players with experts from the science locations Leipzig, Chemnitz and Zwickau and thus supports the sustainable structural change of the region by building up competence with regard to the use of artificial intelligence (AI) methods. The use of AI enables companies to establish new forms of work, develop new business models, and make work more efficient and humane. At the center of the project is the establishment of a competence center, which initially links four scientific partners, three technical companies, ten application partners and one network partner. On the basis of a broad requirements survey, a framework for the design of artificially and humanly intelligent systems will be developed, which forms the core of K-M-I. Based on the methods and process models of this framework, realization scenarios in the form of pilot applications for the use of AI in companies will be developed and implemented within the framework of the project together with the application partners. The bandwidth for the use of AI in the pilots ranges from data development and networking to approaches for the design of intra- and inter-company information flows and knowledge management to the data-based simulation of scenarios and the analysis of workloads. Here, Leipzig University focuses on developing individual solution approaches for AI-based data management and data analysis. The evaluation of the entire process not only ensures the long-term usability of the results for practice, but also provides extensive knowledge about the transfer between science and practice, which contributes to the continuous competence development of the K-I-M and flows into the consulting of other companies in the Central German lignite mining area. The K-M-I project is funded by the BMBF (Funding reference: 02L19C503) and will run for 5 years from 12/2021 till 11/2026.

SaxoCellSystems: Establishment of AI-driven technologies to support automated ATMP manufacturing processes Made in Saxony.

C. Martin, J. Ewald, S. Fricke, U. Blache, E. Rahm

The overall goal of SaxoCell is to develop cell and gene therapeutics (ATMPs) for affordable and safe treatment of patients suffering from previously untreatable diseases. The SaxoCellSystems project is an essential part of the SaxoCell cluster to create a sustainable infrastructure in Saxony. Process optimization and automation has the great potential to support safety, efficiency and cost reduction of ATMP manufacturing and thus to make broad clinical application sustainable (collaboration with SaxoCellClinics). This essential contribution to the implementation of the SaxoCell strategy is additionally underpinned by the further training of GMP personnel with the created training opportunities (together with colleagues from the ICP of the SaxoCellHub). The developed platform technology is partially scalable and can be configured for the indications of different cluster partners. In later implementation phases, the automation platform will be concretized and expanded together with the clinical partners developing specific ATMPs in SaxoCell (see the different disciplines CAR-T cells, NK/CAR-NK cells, ATMPS for regenerative medicine, ZGT-modulating).

SaxoCellOmics: Technology and competence platform for efficient and harmonized evaluation of cell and gene therapies

C. Martin, J. Ewald, B. Ezio, K. Reiche, E. Rahm

The SaxoCell region has benefited from a concerted technology investment over the past 20 years. This has resulted in an offering of state-of-the-art cellular and molecular measurement techniques and associated data processing and interpretation. This regionally unique combination is being adapted for scientific and commercial challenges by SaxoCell and brought together in a common structure, SaxoCellOmics. SaxoCellOmics acts as a platform for multidimensional cellular, molecular and imaging measurement methods as well as structured data collection, integration and interpretation. Thus, we ensure an optimal and early support of the development and production of gene and cell therapeutics. In the first funding period, SaxoCellOmics will accompany one to two selected SaxoCell projects. The processes thus established are directly transferable to new products and industrial collaborations in further periods. The ScaDS.AI supports SaxoCellOmics in the efficient use of biological mass data with methods from AI as well as the management of knowledge by means of knowledge graphs and can draw on know-how from extensive preliminary work. In particular, Prof. Rahm's group is working on the annotation of medical documents using ontology concepts and AI [e.g., R10, R11, R12] and data integration in the context of the SMITH project [R13]. This is complemented by research on privacy preserving record linkage (PPRL) [R14] and traceability of AI-based decisions.

4 Publications and Theses

Conference/Workshop Publications and Book Chapters

[1] Daniel Ayala, Inma Hernández, David Ruiz, and Erhard Rahm. “Towards the smart use of embedding and instance features for property matching.” In: *ICDE*. IEEE. 2021, pp. 2111–2116.

[2] Simone Braun, Georges Alkhouri, and Eric Peukert. “KOBRA: Praxisfähige lernbasierte Verfahren zur automatischen Konfiguration von Business-Regeln in Duplikaterkennungssystemen.” In: Gesellschaft für Informatik, Bonn, 2021.

[3] Martin Franke, Ziad Sehili, Florens Rohde, and Erhard Rahm. “Evaluation of Hardening Techniques for Privacy-Preserving Record Linkage.” In: *EDBT*. 2021, pp. 289–300.

[4] Kevin Gomez, Matthias Täschner, M Ali Rostami, Christopher Rost, and Erhard Rahm. “Graph Sampling with Distributed In-Memory Dataflow Systems.” In: *BTW*. Gesellschaft für Informatik, Bonn, 2021.

[5] Martin Grimmer, Tim Kaelble, and Erhard Rahm. “Improving Host-based Intrusion Detection Using Thread Information.” In: *Symposium on Emerging Information Security and Applications (EISA)*. 2021.

[6] Marios Karapanos, Andreas Thor, and Heinz-Werner Wollersheim. “Itempool-Management mit Microsoft Excel.” In: *Workshop Gemeinschaften in Neuen Medien (GeNeMe) 2020*. TUDpress. 2020.

[7] Stefan Lerm, Alieh Saeedi, and Erhard Rahm. “Extended Affinity Propagation Clustering for Multi-source Entity Resolution.” In: *BTW*. Gesellschaft für Informatik, Bonn, 2021.

[8] Ying-Chi Lin, Victor Christen, Anika Groß, Toralf Kirsten, Silvio Domingos Cardoso, Cédric Pruski, Marcos Da Silveira, and Erhard Rahm. “Evaluating Cross-lingual Semantic Annotation for Medical Forms.” In: *HEALTHINF*. 2020, pp. 145–155.

[9] Ying-Chi Lin, Phillip Hoffmann, and Erhard Rahm. “Enhancing Cross-lingual Semantic Annotations using Deep Network Sentence Embeddings.” In: *HEALTHINF*. 2021, pp. 188–199.

[10] A. Meissner R.and Thor. “Creation and Utilisation of Domain Specific Knowledge Graphs (DSKG) for E-Learning.” In: *19. Fachtagung Bildungstechnologien der Gesellschaft für Informatik (DELFI)*. 2021.

[11] R. Meissner, D. Jenatschke, and A. Thor. “Evaluation of Approaches for Automatic E-Assessment Item Annotation with Levels of Bloom’s Taxonomy.” In: *ICWL*. 2020.

[12] R. Meissner and A. Thor. “Flexible Educational Software Architecture: at the example of EAs.LiT 2.” In: *Proc. 1st International Workshop on Intelligence Support for Mentoring Processes in Higher Education (IMHE 2020)*. 2020.

[13] Daniel Obraczka and Erhard Rahm. “An Evaluation of Hubness Reduction Methods for Entity Alignment with Knowledge Graph Embeddings.” In: *IC3K*. 2021.

[14] Daniel Obraczka, Jonathan Schuchart, and Erhard Rahm. “Embedding-Assisted Entity Resolution for Knowledge Graphs.” In: *ESWC Workshop on Knowledge Graph Construction (KGCW)*. 2021.

[15] Christopher Rost, Kevin Gomez, Philip Fritzsche, Andreas Thor, and Erhard Rahm. “Exploration and Analysis of Temporal Property Graphs.” In: *EDBT*. 2021, pp. 682–685.

[16] Alieh Saeedi, Lucie David, and Erhard Rahm. “Matching Entities from Multiple Sources with Hierarchical Agglomerative Clustering.” In: *IC3K*. 2021.

[17] Ziad Sehili, Florens Rohde, Martin Franke, and Erhard Rahm. “Multi-Party Privacy Preserving Record Linkage in Dynamic Metric Space.” In: *BTW*. Gesellschaft für Informatik, Bonn, 2021.

[18] Jingyu Shao, Qing Wang, Asiri Wijesinghe, and Erhard Rahm. “ErGAN: Generative Adversarial Networks for Entity Resolution.” In: *ICDM*. IEEE. 2020, pp. 1250–1255.

[19] Moritz Wilke and Erhard Rahm. “Towards Multi-modal Entity Resolution for Product Matching.” In: *GI-Workshop on Foundations of Databases / Grundlagen von Datenbanken (GVDB 21)*. 2021.

Journal Publications

[1] Daniel Ayala, Inma Hernández, David Ruiz, and Erhard Rahm. “Leapme: Learning-based property matching with embeddings.” In: *Data and Knowledge Engineering* 137 (2022), p. 101943.

[2] Axel-Cyrille Ngonga Ngomo, Mohamed Ahmed Sherif, Kleanthi Georgala, Mofeed Mohamed Hassan, Kevin Dreßler, Klaus Lyko, Daniel Obraczka, and Tommaso Soru. “LIMES: A Framework for Link Discovery on the Semantic Web.” In: *KI-Künstliche Intelligenz* (2021), pp. 1–11.

[3] Florens Rohde, Martin Franke, Ziad Sehili, Martin Lablans, and Erhard Rahm. “Optimization of the Mainzelliste software for fast privacy-preserving record linkage.” In: *Journal of Translational Medicine* 19.1 (2021), pp. 1–12.

[4] Christopher Rost, Kevin Gomez, Matthias Täschner, Philip Fritzsche, Lucas Schons, Lukas Christ, Timo Adameit, Martin Junghanns, and Erhard Rahm. “Distributed temporal graph analytics with GRADOOP.” In: *VLDB Journal* (2021).

[5] Sherif Sakr, Angela Bonifati, Hannes Voigt, Alexandru Iosup, Khaled Ammar, Renzo Angles, Walid Aref, Marcelo Arenas, Maciej Besta, Peter A Boncz, et al. “The future is big graphs: a community view on graph processing systems.” In: *Communications of the ACM* 64.9 (2021), pp. 62–71.

[6] S. Scherzinger, A. Thor, and T. Härder. “Editorial Schwerpunktthema Digitale Lehre im Fachgebiet Datenbanksysteme (II).” In: *Datenbankspektrum 21(2)* (2021).

[7] Maja Schneider and Oliver Jokisch. “Towards a Robust Analysis and Classification of Dog Barking.” In: *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2020* (2020), pp. 117–124.

[8] A. Thor, L. Bornmann, R. Haunschild, and L. Leydesdorff. “Which are the influential publications in the Web of Science subject categories over a long period of time?” In: *Journal of Information Science 47(3)* (2021).

[9] A. Thor and T. Kirsten. “Das E-Assessment-Tool DMT.” In: *Datenbankspektrum 21(1)* (2021).

[10] Dinusha Vatsalan, Peter Christen, and Erhard Rahm. “Incremental clustering techniques for multi-party Privacy-Preserving Record Linkage.” In: *Data & Knowledge Engineering* 128 (2020), p. 101809.

arXiv Reports

[1] Daniel Obraczka, Jonathan Schuchart, and Erhard Rahm. *EAGER: Embedding-Assisted Entity Resolution for Knowledge Graphs*. 2021. arXiv: 2101.06126.

[2] Christopher Rost, Philip Fritzsche, Lucas Schons, Maximilian Zimmer, Dieter Gawlick, and Erhard Rahm. *Bitemporal Property Graphs to Organize Evolving Systems*. 2021. arXiv: 2111.13499.

Ph. D. Theses

[1] Victor Christen. “Advanced Methods for Entity Linking in the Life Sciences.” Ph.D. Leipzig University, 2020.

[2] Markus Nentwig. "Scalable Data Integration for Linked Data." Ph.D. Leipzig University, 2020.

[3] Alieh Saeedi. "Clustering Approaches for Multi-source Entity Resolution." Ph.D. Leipzig University, 2021.

Bachelor and Master Theses

[1] Thomas Abel. "Security im Internet der Dinge (IoT)." B.Sc. Leipzig University, 2021.

[2] Tinsaye Abye. "TF-IDF for Entity Resolution in huge Knowledge Graphs." B.Sc. Leipzig University, 2021.

[3] Timo Adameit. "Gradoop on Apache Spark." M.Sc. Leipzig University, 2020.

[4] Ali Al-Ali. "Implementierung und Evaluierung von effizienten Datenstrukturen für Image Matching." M.Sc. Leipzig University, 2021.

[5] Abdalrahman Alkamel. "Distributed Pattern Matching on Graph Streams." M.Sc. Leipzig University, 2020.

[6] Alwine Balfanz. "Erstellung eines Physical Distancing Graphen basierend auf Bluetooth Low Energy Signalstärken." B.Sc. Leipzig University, 2021.

[7] Frauke Beccard. "Development of a data-based model for multi-day pattern sampling based on single-day data." M.Sc. Leipzig University + Bosch GmbH, 2020.

[8] Stefan Berger. "Feature Selection Methods to Predict Wafer Thickness in Chip Manufacturing." M.Sc. Leipzig University + Global Foundries, 2021.

[9] Nadine Biedermann. "Untersuchung einer IoT-Plattform." B.Sc. Leipzig University, 2021.

[10] Simon Bordewisch. "End-To-End Stream Graph Analytics." B.Sc. Leipzig University, 2020.

[11] Simon Bordewisch. "Flexible Graphstream-Analyse mittels Apache Flinks DataStream-API." M.Sc. Leipzig University, 2021.

[12] Lukas Christ. "Time-dependent Graph Pattern Matching with Gradoop." M.Sc. Leipzig University, 2020.

[13] Lucie David. "Distributed Hierarchical Clustering Algorithm for Multi-source Entity Resolution." B.Sc. Leipzig University, 2021.

[14] Maurice Eisenblätter. "Implementation of an In-Memory data structure for the EPGM." B.Sc. Leipzig University, 2021.

[15] Philip Fritzsche. "Relationale Speicherung und Verarbeitung für temporale Graphdaten." M.Sc. Leipzig University, 2021.

[16] Xiaofan Guo. "Verteilte Graph-Stream Visualisierung mit Apache Flink." M.Sc. Leipzig University, 2021.

[17] Tim Häntschel. "Evaluation of Autoencoders for encrypting Bloom-Filters." B.Sc. Leipzig University, 2021.

[18] Daniel Helmrich. "Comparing Anomaly-Based Network Intrusion Detection Approaches Under Practical Aspects." B.Sc. Leipzig University, 2021.

[19] Philipp Hofmann. "Cross-lingual Semantic Annotation Using Sentence Embeddings for Medical Forms." B.Sc. Leipzig University, 2020.

[20] Rainer Hofmann. "Entwicklung eines Softsensors zur Bestimmung der Raumluftqualität in Büroräumen." B.Sc. Leipzig University, 2021.

[21] Bingqing Hu. "Postprocessing mit Knotenembeddings." B.Sc. Leipzig University, 2020.

[22] Jonathan Huthmann. "Vorhersage der Kontrastmittelanreicherung in 3D-Mamma MRT-Aufnahmen durch Einsatz von Deep Learning." M.Sc. Leipzig University, 2021.

[23] Julius Kluge. "ML für Schema Mapping auf heterogenen Datenquellen." M.Sc. Leipzig University, 2021.

[24] Michael Koch. "Privatsphäre-erhaltende Vorhersage der Überlebenszeiten von Magenkrebs-Patient*innen." M.Sc. Leipzig University, 2021.

[25] Aruscha Kramm. "Privatsph¨are-erhaltende Analyse des Fahrradklimas in Leipzig." M.Sc. Leipzig University, 2021.

[26] Jonas Kreusch. "Parallelisierung von Meta-Blocking Ans¨atzen in FAMER." M.Sc. Leipzig University, 2020.

[27] Dennis Kreußel. "Optimal Evasion Attacks and the Behavior of Machine Learning and Anomaly based Intrusion Detection Systems in Adversarial Environments." M.Sc. Leipzig University, 2021.

[28] Vasylyna Kuibida. "Ein Vorgehensmodell zur kontinuierlichen Lieferung von Datenbankänderungen." B.Sc. Leipzig University + PROMOS GmbH, 2021.

[29] Lennart Laverenz. "Untersuchung alternativer Kodierungsverfahren f¨ur PPRL." B.Sc. Leipzig University, 2021.

[30] Stefan Lerm. "Verteiltes Clustering für Multi-Source Entity Resolution gemischter Datensammlungen aus duplikatfreien und duplikatbehafteten Quellen." M.Sc. Leipzig University, 2020.

[31] Lea Löffelmann. "Optimierung einer Tourenplanung auf Basis realer Messdaten eines Kurierdienstes." B.Sc. Leipzig University, 2021.

[32] Tim Matzeck. "Attacks on PPRL using node embedding techniques." M.Sc. Leipzig University, 2021.

[33] Steffen Meßing. "LOTS-Erweiterungen: Trainer f¨ur Relationenalgebra und Cypher." B.Sc. Leipzig University, 2020.

[34] Clemens Morgenstern. "Analyse von BLE-Signalst¨arken und Ableitung von Handlungsempfehlungen für Demonstratoren." B.Sc. Leipzig University, 2021.

[35] Caroline Mösler. "Host-Based Intrusion Detection durch Systemcall Analysen mit Autoencodern." M.Sc. Leipzig University, 2020.

[36] Paul Muschiol. "Development of an Audio-Classifier for Urban Sounds under consideration of the PATE framework." M.Sc. Leipzig University, 2021.

[37] Rana Noureddin. "Distributed Grouping of Property Graph Streams." M.Sc. Leipzig University, 2020.

[38] Nils Pfeifer. "Modellierung der Wärmeströmung in der additiven Fertigung von Metallbauteilen." M.Sc. Leipzig University, 2021.

[39] Leonie Preker. "Evaluating embedding methods for genomic data." M.Sc. Leipzig University, 2021.

[40] Jeremy Puchta. "Interactive Exploration of Embedding Spaces." M.Sc. Leipzig University, 2021.

[41] Lukas Reinhardt. "HIDS zur Identifizerung von Hardwaremanipulation." B.Sc. Leipzig University, 2021.

[42] Toni Rucks. "Verbesserung des LID-DS unter Verwendung einer Multi-Container-Docker-Umgebung zum Erfassen von Daten für host- und netzwerkbasierte Angriffserkennung." B.Sc. Leipzig University, 2021.

[43] Aljoscha Rydzyk. "Visualization-driven graph data reduction." M.Sc. Leipzig University, 2021.

[44] Maximilian Scheiber. "Datenanalyse von standortbezogenen Twitter-Daten." B.Sc. Leipzig University, 2021.

[45] Stefan Schmidt-Dichte. “Implementierung eines Product-Matching-Systems (extern bei HTWK).” B.Sc. Leipzig University, 2021.

[46] Lucas Schons. “Distributed Graph Layouting on Graph Collections.” B.Sc. Leipzig University, 2020.

[47] Max Schrodt. “Konzeption und Implementierung eines Tools zur visuellen Maskierung beim interaktiven Record Linkage.” B.Sc. Leipzig University, 2021.

[48] Till Schultz. “Performance Optimierung eines Big Data Analytic Systems.” B.Sc. Leipzig University, 2021.

[49] Dominik Schwabe. “Erschließung und Evaluierung von Matching-Datensätzen mit personenbezogenen Daten.” B.Sc. Leipzig University, 2020.

[50] Elena Senger. “Entity Matching unter Verwendung von Geodaten am Beispiel von Daten der Deutschen Bahn.” M.Sc. Leipzig University, 2021.

[51] Greta Staskewitsch. “Anomaliebasierte Host Intrusion Detection mittels Sequenz- und Parameteranalysen von Systemcalls.” B.Sc. Leipzig University, 2020.

[52] Marcus Stelzer. “Realisierung eines Frameworks zur forensischen Gutachtenerstellung und Auswertung der Browser-Historie.” B.Sc. Leipzig University, 2020.

[53] Georg Walther. “Detection of aneurysms and prediction of rupture.” B.Sc. Leipzig University, 2020.

[54] Rico Warnke. “Prozessoptimierung durch Digitalisierung im Sozialwesen am Beispiel einer datenbankgestützten Verwaltungssoftware für die stationäre Kinder- und Jugendhilfe.” B.Sc. Leipzig University, 2021.

[55] Robert Weiske. “Cluster Quality Prediction for Entity Resolution.” M.Sc. Leipzig University, 2021.

[56] Robert Weiske. “Incremental Entity-Resolution using cluster vector representations.” B.Sc. Leipzig University, 2020.

[57] Jonathan Weske. “Fahrrad-zentrische Analyse von Stehzeiten in einem Fahrradverleihsystem.” B.Sc. Leipzig University, 2021.

[58] Konstantin Wilson. “Temporal Graph Metrics.” B.Sc. Leipzig University, 2020.

[59] Maximilian Zimmer. “Kontinuierliche Graph-Query Notifikationen auf Basis eines RDBMS.” B.Sc. Leipzig University, 2021.

5 Talks

[1] Timo Adameit and Alexander Leipnitz. *Social Distancing mit Kamera und KI*. Meetup, Online, 19.05.2021. 2021.

[2] Thomas Burghardt. *Mit Kamera und Machine Learning Social Distancing unterstützen*. microTEC Südwest Clusterkonferenz 2022, Online. 2021. URL: https://www.microtec-suedwest.de/news-termine/cluster%5C%5Ckonferenz/archiv-clusterkonferenz/clusterkonferenz-2021.

[3] Martin Franke. *Evaluation of Hardening Techniques for PPRL*. EDBT, Online. 2021.

[4] Kevin Gomez and Christopher Rost. *Gradoop Tutorial - Scalable Graph Analytics on Apache Flink*. BOSS’20, VLDB2020, Online. 2020. URL: https://boss-workshop.github.io/boss-2020/#program.

[5] Kevin Gomez and Matthias Täschner. *Graph Sampling with Distributed In-Memory Dataflow Systems*. BTW, Online. 2021. URL: https://sites.google.com/view/btw-2021-tud/programm.

[6] Martin Grimmer. *Improving Host-based Intrusion Detection Using Thread Information*. EISA. 2021.

[7] Tobias Jagla. *The Secret Secrets Of The Robotic Hive: Schwarmintelligente Roboter, die ihre Umgebung zum Leuchten bringen*. Lange Nacht der Wissenschaften 2021, Online. 2021. URL: https://www.wissen-in-leipzig.de/programm-nach-einrichtungen/scads-ai-zentrum-fuer-skalierbare-datenanalyse-und-kuenstliche-intelligenz.

[8] Stefan Lerm. *Extended Affinity Propagation Clustering for Multi-source Entity Resolution*. BTW, Online. 2021.

[9] Ying-Chi Lin. *Enhancing Cross-lingual Semantic Annotations using Deep Network Sentence Embeddings*. HEALTHINF. 2021.

[10] Ying-Chi Lin. *Evaluating Cross-lingual Semantic Annotation for Medical Forms*. HEALTHINF. 2020.

[11] Daniel Obraczka. *An Evaluation of Hubness Reduction Methods for Entity Alignment with Knowledge Graph Embeddings*. IC3K. 2021.

[12] Daniel Obraczka. *Embedding-Assisted Entity Resolution for Knowledge Graphs*. ESWC Workshop KGCW. 2021.

[13] Erhard Rahm. *Querschnittsthemen in der NFDI*. Forschungsdaten-Kolloquium, Univ. Leipzig. 2021.

[14] Erhard Rahm. *SCADS.AI –Zentrum Für KI und Data Science*. Parlamentarischer Abend der Univ. Leipzig. 2020.

[15] Erhard Rahm. *ScaDS.AI Research Topics*. 6th International ScaDS Summer school on AI and Big Data. 2020.

[16] Florens Rohde. *Privacy-Preserving Record Linkage*. ScaDS.AI Living Lab Lecture Series, Online. 2021. URL: https://scads.ai/living-lab-en/online-lecture-series/6-privacy-preserving-record-linkage/.

[17] Christopher Rost, Timo Adameit, and Kevin Gomez. *Was wir aus nextbike Daten über die Leipziger Fahrradkultur erfahren können oder Was ist eigentlich ein Temporaler Graph?* Lange Nacht der Wissenschaften 2021, Online. 2021. URL: https://www.wissen-in-leipzig.de/programm-nach-einrichtungen/scads-ai-zentrum-fuer-skalierbare-datenanalyse-und-kuenstliche-intelligenz.

[18] Christopher Rost and Kevin Gomez. *Temporal Graph Analysis*. ScaDS.AI Living Lab Lecture Series, Online. 2021. URL: https://scads.ai/living-lab-en/online-lecture-series/temporal-graph-analysis/.

[19] Christopher Rost and Kevin Gomez. *Temporal Graph Analytics with GRADOOP*. FOSDEM'20, Brussels, Belgium. 2020. URL: https://archive.fosdem.org/2020/schedule/event/graph_temporal_gradoop/.

[20] Alieh Saeedi. *Matching Entities from Multiple Sources with Hierarchical Agglomerative Clustering*. IC3K. 2021.

[21] Maja Schneider. *Herausforderungen und Schutz der Privatsphäre im Projekt DE4L*. Workshop der Begleitforschung zu den Technologieprogrammen Smarte Datenwirtschaft (SDW) und KI-Innovationswettbewerb (KI-IW), Online. 2021.

[22] Ziad Sehili. *Multi-Party Privacy Preserving Record Linkage in Dynamic Metric Space*. BTW, Online. 2021.

[23] Moritz Wilke. *Towards Multi-modal Entity Resolution for Product Matching*. GVDB. 2021.