Database Group Leipzig
Department of Computer Science

UNIVERSITÄT LEIPZIG

# Research Report 2022/2023

https://dbs.uni-leipzig.de

## Overview


| | | |
|---|---|---|
| 1 | **Staff** | 2 |
| 2 | **Highlights** | 3 |
| 3 | **Research Topics and Projects** | 6 |
| 4 | **Publications and Theses** | 20 |
| 5 | **Talks** | 26 |



Database Group in June 2023. r.t.l.: Christopher Rost, Andre Wille, Florens Rohde, Prof. Dr. Erhard Rahm, Karim Rakia, Dr. Christian Martin (top), Julius Ellermann (top), Marie-Sophie von Braun, Dr. Victor Christen, Daniel Obraczka, Lucas Lange, Adrian Kuhn, Maximilian Heykeroth, Marvin Hofer, Mouna Ammar, Martin Franke. Persons missing: Dr. Thomas Burghardt, Martin Grimmer, Aruschka Kramm, Jonas Kreusch, Anja Neumann, Dr. Eric Peukert, Maja Schneider, Matthias Täschner, Benjamin Uhrich.

# 1 Staff

| | |
|---|---|
| Prof. Dr. Rahm, Erhard | Professor |
| Hesse, Andrea | Secretary |
| Ammar, Mouna (since May 2023) | Research associate |
| Dr. Burghardt, Thomas | Postdoctoral Researcher (ScaDS.AI) |
| Carnot, Miriam Louise (since April 2023) | Research associate |
| Dr. Christen, Victor | Postdoctoral Researcher |
| Dr. Ewald, Jan | Postdoctoral Researcher (ScaDS.AI) |
| Franke, Martin | Research associate |
| Gomez, Kevin (until Jun. 2022) | Research associate |
| Grimmer, Martin | Research associate |
| Hannemann, Anika (until Dec. 2022) | Research associate |
| Raphael Hildebrandt (until Dec. 2022) | Research associate |
| Hofer, Marvin | Research associate |
| Dr. Köpcke, Hanna (until Dec. 2022) | Postdoctoral Researcher |
| Kramm, Aruscha | Research associate |
| Kreusch, Jonas | Research associate |
| Lange, Lucas (since March 2022) | Research associate |
| Leipnitz, Alexander (until January 2022) | Research associate |
| Dr. Martin, Christian | Postdoctoral Researcher (ScaDS.AI) |
| Neumann, Anja | Research associate |
| Obraczka, Daniel | Research associate |
| Peter, Lucas (since April 2023) | Research associate |
| Pollack, Jacob (since June 2023) | Research associate |
| Pogany, Gergely (until Dec. 2022) | Research associate |
| Dr. Peukert, Eric | Postdoctoral Researcher (ScaDS.AI) |
| Rohde, Florens | Research associate |
| Rost, Christopher | Research associate |
| Dr. Saeedi, Alieh (until May 2022) | Postdoctoral Researcher |
| Schneider, Maja | Research associate |
| Schuchart, Jonathan (until July 2022) | Research associate |
| Täschner, Matthias | Research associate |
| Uhrich, Benjamin | Research associate |
| Von Braun, Marie-Sophie | Research associate |
| Wilke, Moritz (until January 2022) | Research associate |
| Prof. Dr. Thor, Andreas (HfTL Leipzig) | Associated team member |
| Prof. Dr. Anika Groß (HS Anhalt) | Associated team member |

## 2 Highlights

There have been several highlights in 2022 and 2023:

1. Prof. Rahm acted as vice president and treasurer of the German Informatics (GI) society and was re-elected by the GI members at the end of 2023 for another two years (2024/25). In March 2023, the GI Board met for a closed meeting at Leipzig University.

Board meeting of the German Informatics Society at ScaDS.AI Leipzig. In the back from left: Daniel Krupka (managing director), Christine Regitz (president), Ulrike Lucke (vice president), Jörg Desel, Katherina Weitz, Front: Cornelia Winter (managing director), Erhard Rahm (vice president)

2. In July 2022, ScaDS.AI Dresden/Leipzig has organized together with partners and in cooperation with the city of Leipzig the Data Week Leipzig 2022 that included the 5th ScaDS.AI Big Data and AI in Business Workshop (BDIB). The Workshop featured keynote speaker Dr. Volker Gruhn as well as presentations by company representatives on their Big Data and AI projects.

Director of ScaDS.AI Dresden/Leipzig Prof. Erhard Rahm at BDIB workshop.

3. From July 11-15, 2022, the German Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI Dresden/Leipzig) hosted its 8th International Summer School on AI and Big Data. After two years with restrictions due to the Corona pandemic, this summer school could finally take place in attendance again. It was held on the premises of the Radisson Blu Hotel, right next to the Augustusplatz campus of Leipzig University.

8th International Summer School on AI and Big Data in Leipzig

4. On September 5, 2022, ScaDS.AI Dresden/Leipzig (Center for Scalable Data Analytics and Artificial Intelligence) celebrated its consolidation as a permanant national competence center for research into artificial intelligence (AI), data science and big data. ScaDS.AI is headed by Prof. Dr. Wolfgang Nagel (TU Dresden) and Prof. Dr. Erhard Rahm (Leipzig University) and is supported by both universities. Thanks to the now permanent funding, the center receives almost 20 million euros per year from the federal and state governments.

Kick-off participants l.t.r.: Prof. Dr. Erhard Rahm (Leipzig University), Prof. Dr. Eva Inès Obergfell (Leipzig University), Minister of Science Sebastian Gemkow , Dr. Christoph March (BMBF), Prof. Ursula M. Staudinger (TU Dresden), Prof. Dr. Wolfgang Nagel (TU Dresden).

5. In March 2023 Maja Schneider, Victor Christen and Martin Franke visited Prof. Dr. Peter Christen, Dr. Anushka Vidanage and Charini Nanayakkara at the Australian National University for some joint research activities. This cooperation led to a research paper about (privately) estimating the linkage quality in record linkage application that was accepted for the EDBT

conference in 2024.

6. From July 26-30, 2023, the Institute for Applied Informatics (InfAI) e.V., ScaDS.AI Dresden/Leipzig, eccenca GmbH and the Open Geospatial Consortium (OGC) have organized the Data Week at the New City Hall Leipzig together with the Departement "Digitale Stadt" of the City of Leipzig as well as collaborators of the CUT – Connected Urban Twin project. The focus of the event week will be on the topics of digitization, data, artificial intelligence, and urban development challenges. In particular, the importance of international data spaces for climate protection strategies and measures will be highlighted.

7. The external group seminar took place in June 2022 in Zingst (for the 17th time) and in May 2023 in Darlingerode (Harz).

8. In October 2023 the database group launched a new website.

9. From April 2023, the new research project *HyGraph: Querying and Analytics on Hybrid Graphs* started in cooperation with the Lyon 1 University in France. This French-German collaboration is founded as a joint ANR and DFG research project for a duration of three years. Besides Prof. Dr. Angela Bonifati as a PI of Lyon 1 and Prof. Dr. Erhard Rahm as a PI of Leipzig University, Eric Peukert and Christopher Rost supervise the project, too. Our new colleague Mouna Ammar already started her work on the HyGraph project as a PhD student in May 2023. After an online kickoff meeting in June 2023, the whole team met in a first in-person project meeting in Lyon in November 2023 for two days.


Kick-off participants l.t.r.: Prof. Dr. Erhard Rahm (Leipzig University), Mouna Ammar (Leipzig University), Prof. Dr. Angela Bonifati (Lyon 1 University), Dr. Riccardo Tommassini (INSA Lyon), Dr. Remy Cazabet (Lyon 1 University), Christopher Rost (Leipzig University).

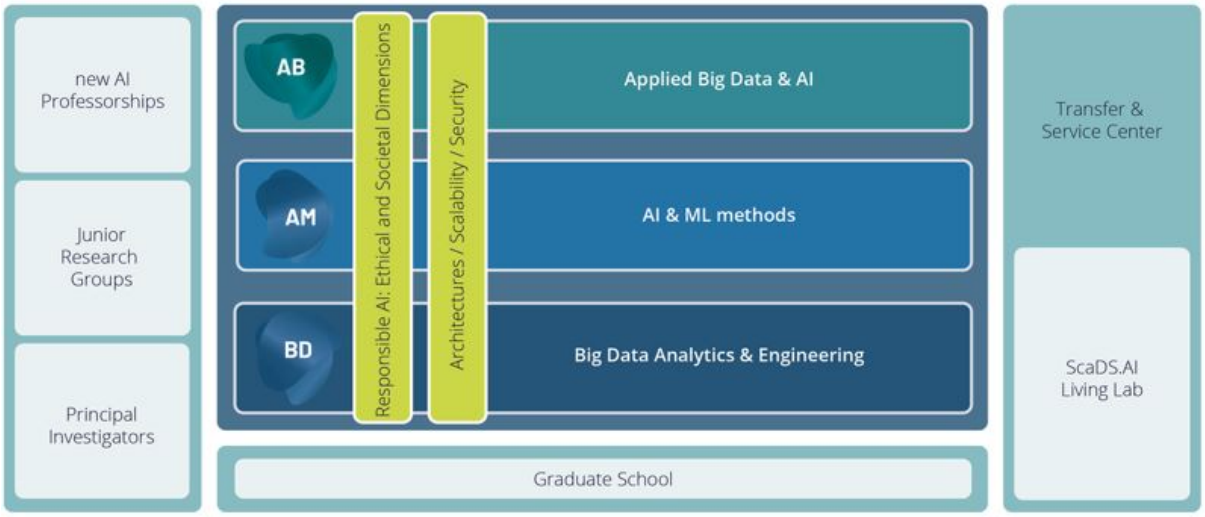
10. Several additional third-party funded projects could be secured and started in 2022/23, especially MIRACLE, DiGURaL and Come2Data.

# 3 Research Topics and Projects

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

*E. Peukert, T. Burghardt, J. Ewald, C. Martin, D. Obraczka, J. Schuchart, M. Täschner, M. Wilke, E. Rahm*

ScaDS.AI Dresden/Leipzig, the Center for Scalable Data Analysis and Artificial Intelligence, led by Prof. Nagel (TU Dresden) and Prof. Rahm (Leipzig University), is one of the new national competence centers for Artificial Intelligence (AI) that are being funded under the federal government's AI strategy and established as permanent research facilities. With its two locations in Leipzig and Dresden, ScaDS.AI Dresden/Leipzig combines the AI and Data Science expertise of Leipzig University, and TU Dresden as well as of ten further research institutions. It expands the previous competence center for Big Data — ScaDS Dresden/Leipzig — founded in 2014. At the University of Leipzig as well as the Technische Universität Dresden, up to 12 AI professorships will be established to increase research excellence and to place applied AI research on a broad foundation. In 2022, the first ScaDS.AI professorhip as well as a Humboldt Professorship for AI (Sayan Mukherjee) could be established at ScaDS.AI Leipzig. In 2023, the first junior research groups and associated memberships could be added.




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 transfer of research results into industry with an increasing number of cooperation projects with companies. This aims at strengthening the innovation power and competitiveness of the participating companies. 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, the service center offers trainings in the area of Data Science, Big Data, AI, and High-Performance Computing (HPC) and enables partners with the use of AI and HPC resources. Besides the regular trainings offered, additional courses were held as part of events such as the Data Week Leipzig in collaboration with InfAI and the City of Leipzig. Here, the basics and advanced knowledge of data visualization and the application of AI algorithms were taught. In October 2023, a one-week AI workshop was held with the Mathematical Modeling and Simulation Network (MMS) of Leibniz Society. 40 participants had the opportunity to learn the basics and advanced applications of AI and ML, gain practical experience in using common tools and working on HPC resources, and talk to experts from ScaDS.AI Dresden/Leipzig about their research questions.

### *Living Lab – Science Communication*

Digital voice assistants, ChatGPT, friend and media suggestions in social networks and streaming services: AI is becoming increasingly important in many areas. However, AI-based systems are still difficult to understand for many users. For this reason, ScaDS.AI Dresden/Leipzig aims to promote the possibilities of knowledge transfer in the fields of AI and data science. The public can experience AI on the premises of the ScaDS.AI Living Labs in Leipzig and Dresden. With the help of demonstrators and practical application examples, topics such as machine learning, process monitoring in 3D printing, data protection and image recognition are made understandable. Visitors can interact with the technologies playfully using sample applications.

In 2022/2023, the Living Labs in Leipzig has hosted 64 events and welcomed approximately 800 visitors from a wide range of ages and backgrounds, including politics, business, science, and the public. The ScaDS.AI Living Lab opens its door for events such as the Long Night of Science, Girls' Day or various individual events such as discussion groups, workshops, trainings, hackathons, thematic lectures, and guided tours. In addition, the latest research results are presented in the Living Lab Hands-On Demonstrator Series and the Living Lab Lecture Series. For the transfer to science and industry, public events for exchange and networking take place regularly.

### *Graduate Qualification*

The ScaDS.AI Graduate School operated regularly, introduced additional activities and, near the end of 2023, continues to refine its processes. Firstly, the school's established qualification elements in terms of its local seminar series for doctoral researchers were complemented by the ScaDS.AI colloquium series, a publicly open lecture series that delivers input from and networking opportunities with invited international and national guests. Eight hybrid lectures in Leipzig and one online session were realized since October 2022. Secondly, two four-day seminar trips in October 2022 and 2023, respectively, kicked off an annual series of retreats that brings together Graduate School members from Leipzig and Dresden. Both events were well-attended, with 33 participants in 2022 and 42 in 2023. Moreover, the cooperation with the Graduate Academy Leipzig (restructured from the former Research Academy Leipzig in March 2023) was continued, concerning its course program for doctoral students as well as through exchange at the level of coordinators of structured graduate programs at or in connection with Leipzig University. Concerning resources, the Graduate School website was set up in 2022 as well as a wiki area, which is constantly being expanded. As of December 2023, the Graduate School accompanies and supports 51 doctoral researchers in Leipzig in pursuing their research.

## Analysis and Processing of Temporal Graphs and Graph Streams

*C. Rost, K. Gomez, E. Rahm*

Analyzing highly connected data as graphs becomes increasingly essential in many different domains. Prominent examples are social networks, e.g., Facebook and Twitter, and information networks like the World Wide Web or biological networks. With the objective of analyzing 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 built on the distributed dataflow framework Apache Flink and uses the provided APIs to implement the TPGM and its operators. The system is publicly available (`https://github.com/dbs-leipzig/gradoop`) and gets code contributions from other institutes and companies. A demo application, namely the Temporal Graph Explorer, shows the usage and result of three selected temporal graph operators: snapshot, difference, and time-dependent grouping. The application is open-source (`https://github.com/dbs-leipzig/temporal_graph_explorer`) and the corresponding articles published.

Graph metrics, such as the simple but popular vertex degree and others based on it, are well-defined for static graphs. However, adapting static metrics for temporal graphs is still part of current research. In a recently published paper at the BTW 2023, we propose a set of temporal extensions of four degree-dependent metrics and aggregations like minimum, maximum, and average degree of (i) a vertex over a time interval and (ii) a graph at a specific point in time. We show why using the static degree can lead to wrong assumptions about the relevance of a vertex in a temporal graph and highlight the need to include *time* as a dimension in the metric.

Our ongoing work focuses on processing and analyzing dynamic graphs and further graph streams, which represent the continuous addition and removal of vertices and edges and frequent changes in their attributes. In a recent publication accepted for EDBT 2024, we proposed the syntax and semantics of a continuous query language for property graph streams. We further research discovering temporal patterns within a graph stream or summarizing graph structures in a windowed form. The HyGraph project (below) seamlessly joins the research on dynamic graphs and investigates the combination of temporal graphs with time series data and graph streams.

## HyGraph: Querying and Analysis of Hybrid Graphs

*M. Ammar, C. Rost, E. Peukert, E. Rahm*

Graphs are simple yet highly expressive data structures for modeling and analyzing relationships between real-world objects. As the structure and content of graphs are continuously changing, e.g., in social networks or transport and mobility networks, novel data models and analysis mechanisms are needed. The fusion of such temporal graphs with time-series data, as well as the high-frequency updating by graph streams, is a significant challenge, which so far has only been partially enabled by means of distinct models and analytical systems.

In this project called *HyGraph*, funded by a collaboration of the German DFG and French ANR, Leipzig University and Lyon 1 University are working together on a new data model for hybrid graph querying and analysis. We aim to develop a novel hybrid data model that seamlessly combines temporal graphs with time series and enables high-frequency updates through graph streams. This combination in a unified hybrid model paves the way to novel, unprecedented query, analysis, data mining, and machine learning tasks. By means of a planned operator concept,

both queries and analyses can be executed on the hybrid graphs and powerful data mining algorithms, such as frequent pattern mining or clustering, enabled by the concatenation operators. The overall system will be prototypically implemented, and its applicability will be demonstrated for at least one use case. More details are available on the project website: `https://hygraph.net`. The project started in April 2023, and the duration is 3 years.

In May 2023, Mouna Ammar started as a PhD student at the database group of Leipzig University, working full-time on the goals of that project.

## Knowledge Graph Construction

*D. Obraczka, M. Hofer, A. Saeedi, H. Köpcke, V. Christen, E. Rahm*

Knowledge Graphs (KGs) have become a popular data structure to organize information. Automating the creation of KGs involves connecting a plethora of research questions.

In the paradigm of the fourth industrial revolution this involves incorporating smart manufacturing devices which share enormous amounts of data. A crucial step is therefore developing advanced data integration methods that are able to consolidate and combine heterogeneous data from multiple sources. In a book chapter we outline the use of knowledge graphs for data integration and provide an overview of proposed approaches to create and update such knowledge graphs, in particular for schema and ontology matching, data lifting and especially for entity resolution. Furthermore, we present data integration use cases for Industry 4.0 and discuss open problems.

In a more general survey paper we investigate the need for generalized pipelines to construct and continuously update KGs. While the individual steps that are necessary to create KGs from unstructured (e.g. text) and structured data sources (e.g. databases) are mostly well- researched for their one-shot execution, their adoption for incremental KG updates and the interplay of the individual steps have hardly been investigated in a systematic manner so far. We introduce the major requirement for future KG construction pipelines and provide an overview of the necessary steps to build high-quality KGs including cross-cutting topics such as metadata management, ontology development and quality assurance. We then evaluate the state of the art of KG construction w.r.t the introduced requirements for specific popular KGs as well as some recent tools and strategies for KG construction. Ares of future research and improvement are discussed as a concluding outlook

The flexibility of Knowledge Graphs to represent heterogeneous entities and relations of many types is challenging for conventional data integration frameworks. In order to address this challenge the use of Knowledge Graph Embeddings (KGEs) to encode entities from different data sources into a common lower-dimensional embedding space has been a highly active research field. It was recently discovered however that KGEs suffer from the so-called hubness phenomenon. If a dataset suffers from hubness some entities become hubs, that dominate the nearest neighbor search results of the other entities. Since nearest neighbor search is an integral step in the entity alignment procedure when using KGEs, hubness is detrimental to the alignment quality. We investigate a variety of hubness reduction techniques and (approximate) nearest neighbor libraries to show we can perform hubness-reduced nearest neighbor search at practically no cost w.r.t speed, while reaping a significant improvement in quality. We ensure the statistical significance of our results with a Bayesian analysis.

## 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 developed privacy-preserving methods that are applied while collecting sensor data to protect the privacy of drivers, e.g. in sensitive areas. Furthermore we build a tool to 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 build 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 instead of the plaintext address data when offering and searching data on the trading platform.

The project started in August 2019 and ended in December 2022.

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

*B. Uhrich, E. Peukert, E. Rahm*

TWIN focused on the development process for the production of metal prototypes (product creation phase) using 3D printing technologies. To this end, TWIN developed a digital product service system covering the entire value chain of additive manufacturing. The aim was to optimize and monitor complex industrial systems using machine learning methods to provide intelligent recommendations during the production process.

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. We have achieved several milestones in the field of physics-informed machine learning for predicting heat transfer in SLM (Selective Laser Melting). By using machine learning methods that incorporate physical domain knowledge, we have been able to achieve accurate and efficient condition monitoring of printing processes. These milestones were based on thermal image data documenting the temperature process during SLM. This image data could be integrated via cloud storage and then extracted via a suitable analysis pipeline. This led not only to automated error detection, but also to the optimization of individual 3D printing processes. As the project progressed, we decided to focus on extending intelligent real-time analysis to various part geometries in higher-dimensional model space. Our central concern was to identify and exploit the potential of the analysis techniques in a broader context. In collaboration with SIEMENS, we conducted a comprehensive analysis of thermomeasurement data for several prototype part geometries, including pyramids, turbine blades, tension rods and a benchmark part. Finally, we have focused on integrating heterogeneous sensor data (grayscale images and thermal images) into machine learning models to enable multimodal learning. This significantly extends the capabilities of digital twins, as they can now interpret and use different data sources simultaneously to make detailed and more accurate predictions and analyses. 3D printing with integrated quality control requires minimal manual intervention. In the future, it will be possible to intelligently monitor complex industrial systems such as 3D printing. As a result, production processes can be continuously optimized and there will no longer be the need for the level of individual machine monitoring that is necessary today.

The project was funded by the BMWi (FKZ: 02K18D055), started in October 2019 and ended in March 2023.

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

*G. Pogany (till Dec 22), D. Prascevic (start Apr 23), J. Ewald, 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 and was prolonged for 1 year due patient recruitment issues caused by the pandemic (new end of project May 2024). The database group is responsible for the data integration and is also supporting the analyses of heterogeneous molecular and medical data sources of the project. Clinical and biological partners are in the process of wrapping up their analyses and the final months of the project will be the focus of data analysis and integration to derive prognostic markers for treatment response in gastric cancer patients.

## MIRACLE – A Machine learning approach to Identify patients with Resected non-small-cell lung cAnCer with high risk of reLapsE

*M.-S. von Braun, C. Martin, J. Ewald, E. Rahm*

Early-stage non small cell lung cancer (ES-NSCLC) represents 20-30% of all NSCLC and is characterized by a high survival rate after surgery. However, there is variability in clinical outcomes among patients sharing the same disease stage, suggesting that other factors could determine the risk of relapse. Accurate and validated tools to stratify patients according to their risk of relapse are still lacking. Multiple factors may influence the prognosis of resected ES-NSCLC patients: tumor tissue and microenvironment (TME) characteristics, liquid biopsy, radiomics features and clinical-pathological factors could all be involved. The primary aim of the study is the development of a machine learning (ML) algorithm acting as a clinical decision support tool for disease free survival (DFS) prediction and patient stratification based on joint analysis of biological, clinical and radiologic features. The model will be trained on a training cohort of 220 resected ES-NSCLC patients and validated on an independent prospective cohort of 200 patients.

As the leading partner in algorithm development, our role at ScaDS.AI will be to (1) integrate the heterogeneous multimodal data and identify predictive features, (2) develop statistical, machine learning, and deep learning methods for survival analysis, and (3) support explainability for personalized medicine.

The MIRACLE project is funded by the European ERA PerMed Joint Call and is managed by a consortium of partners from Italy, France, Spain, and Germany. It started in June 2022 and will run for three years.

## 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 the years 22/23, further investigations were carried out together with the project partners, which were ultimately integrated into an overall evaluation system that can automatically detect security vulnerabilities.

## Privacy-preserving Record Linkage

*V. Christen, M. Franke, F. Rohde, 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 to be solved for 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.

*Bloom Filter Hardening using Autoencoders*

The majority of PPRL approaches utilize Bloom filters to encode plaintext data. However, Bloom filters suffer from cryptanalysis attacks where frequent bit patterns are recognized and are mapped to plaintext values such qgrams. Even hardening techniques do not guarantee complete protection against these attacks. We investigated an approach based on autoencoders that transform Bloom filters in the continuous vector space preventing cryptanalysis attacks. To guarantee the compatibility between the encoded Bloom filters generated by the local encoders from each data owner, we propose a protocol that enables the training of mapping function from one encoder space to the encoder space of the other data owner by using randomly generated data. The results have shown the effectiveness

of the proposed approach since the quality is comparable and even better regarding the Bloom filter-based linkage.

*Value-specific Weighting for Record-level Encodings in Privacy-Preserving Record Linkage*
A crucial step of PPRL is the classification of Bloom filter pairs as match or non-match based on computed similarities. In the context of record linkage, several weighting schemes and classification methods are available. The majority of weighting methods determine and adapt weights by applying the Fellegi-Sunter model for each attribute. In the PPRL domain, the attributes of a record are typically encoded in a joint record-level Bloom filter to impede cryptanalysis attacks so that the application of existing attribute-wise weighting approaches is not feasible. We studied methods that use attribute-specific weights in record-level encodings and integrate weight adaptation approaches based on individual value frequencies. The experiments on real-world datasets showed that frequency-dependent weighting schemes improve the linkage quality as well as the robustness with regard to threshold selection.

*(Privately) Estimating Linkage Quality for Record Linkage*
To evaluate the quality of record linkage approaches, the performance measures of precision, recall, and F-measure are commonly used. These measures require ground truth data that specifies known matches and non-matches. However, in practical linkage applications there typically is no such ground truth data available. Although linkage quality can be assessed manually by domain experts, such a clerical review process is time- and resource-consuming and generally not feasible when linking databases that are very large or that contain sensitive (personal) data. Therefore, we review existing and propose improved unsupervised approaches for estimating the quality of linkage results. We evaluate our approaches on multiple datasets from three different domains. This evaluation shows that our approaches outperform existing methods and lead to estimates that are close to the actual linkage quality. These estimates can be used in practical applications to identify suitable linkage methods and to optimize their parameters, such as the classification threshold.

Parts of this work were conducted as part of the research stay at the Australian National University (ANU) in collaboration with Prof. Dr. Peter Christen.

*Future Research Directions*
There are still several open research questions in the area of PPRL that need to be addressed. Hardening techniques aim to improve the privacy properties of encoding techniques in order to prevent certain types of attacks. Typically, only a single hardening technique is applied on the entire input databases to be linked. Therefore, the vulnerability of individuals or groups of encoded records is not considered. Since hardening techniques generally show a trade-off between linkage quality and privacy, a compelling research direction is to analyze which minimum combination of techniques is required to minimize the likelihood of success regarding different attacks while maintaining a high linkage quality.

To address errors and inconsistencies in real-world data, PPRL encoding techniques support the calculation of similarities between encoded records. As a consequence, the similarities between encoded records can be compared to the similarities of plaintext records. Recent graph-based similarity attacks exploit this relationship by analyzing and matching the similarity graphs of the encoded and the plaintext records. So far, only a few works have focused on hardening or novel encoding techniques that aim at avoiding attacks that exploit the similarity of encodings.

## Privacy-preserving Machine Learning

*L. Lange, M. Schneider, E. Rahm*

Privacy-preserving Machine Learning (PPML) is a field at the intersection of machine learning and privacy, aiming to develop methodologies that enable the training and deployment of models without

compromising the confidentiality of sensitive data. With the proliferation of data-driven technologies, concerns about privacy breaches have become paramount. Research in this domain employs cryptographic techniques such as homomorphic encryption and secure multi-party computation, anonymization methods, such as k-anonymity, and Differential Privacy (DP) to design algorithms that operate on encrypted or perturbed data. Additionally, federated learning, a decentralized approach, allows models to be trained across multiple devices without raw data leaving individual devices. The overarching goal is to strike a balance between the increasing demand for sophisticated machine learning models and the imperative to protect the privacy of individuals contributing to the training data.

A fundamental consideration in PPML is the inherent trade-off between utility and privacy. As protective measures, such as encryption or noise injection, are implemented to preserve privacy, they concurrently introduce challenges to the accuracy and effectiveness of machine learning models. Striking the right balance becomes crucial; aggressive privacy measures may result in a loss of model utility, diminishing its predictive power or generalization capabilities. Current techniques grapple with optimizing this trade-off, aiming to implement privacy-preserving techniques that mitigate disclosure risks while maintaining the requisite level of model performance.

*Privacy Risk in Medical Applications with and without Differential Privacy Protection*
Machine learning can help fight pandemics like COVID-19 by enabling rapid screening of large volumes of images. To perform data analysis while maintaining patient privacy, we create machine learning models that satisfy DP. Previous works exploring private COVID-19 models are in part based on small datasets, provide weaker or unclear privacy guarantees, and do not investigate practical privacy. We suggest improvements to address these open gaps. We account for inherent class imbalances and evaluate the utility-privacy trade-off more extensively and over stricter privacy budgets. Our evaluation is supported by empirically estimating practical privacy through black-box Membership Inference Attacks (MIAs). The introduced DP should help limit leakage threats posed by MIAs, and our practical analysis is the first to test this hypothesis on the COVID-19 classification task. Our results indicate that needed privacy levels might differ based on the task-dependent practical threat from MIAs. The results further suggest that with increasing DP guarantees, empirical privacy leakage only improves marginally, and DP therefore appears to have a limited impact on practical MIA defense. Our findings identify possibilities for better utility-privacy trade-offs, and we believe that empirical attack-specific privacy estimation can play a vital role in tuning for practical privacy.

*Analyzing Social Media Sentiment with Guaranteed Privacy*
Sentiment analysis is a crucial tool to evaluate customer opinion on products and services. However, analyzing social media data raises concerns about privacy violations since users may share sensitive information in their posts. In this work, we propose a privacy-preserving approach for sentiment analysis on Twitter data using DP. We first implement a non-private baseline model and assess the impact of various settings and preprocessing methods. We then extend this approach with DP under multiple privacy parameters $\varepsilon = \{0.1, 1, 10\}$ and finally evaluate the usability of the resulting private models. Our results show that DP models can maintain high accuracy for the studied task. We contribute to the development of privacy-preserving machine learning for customer opinion analysis and provide insights into trade-offs between privacy and utility. The proposed approach helps protect sensitive information while still allowing for valuable insights to be gained from social media data.

*Detecting Stress from Smartwatches without Compromising Sensitive Health Data*
We present the first privacy-preserving approach for stress detection from wrist-worn wearables based on the Time-Series Classification Transformer (TSCT) architecture and incorporating DP to ensure provable privacy guarantees. The non-private baseline results prove the TSCT to be an effective model for the given task. Our DP experiments then show that the private models suffer from reduced utility but can still be used for reliable stress detection depending on the application. Our proposed approach has potential applications in smart health, where it can be used to monitor smartwatch users'

stress levels without compromising their privacy and provide timely interventions or suggestions to prevent adverse health outcomes. Another primary contribution is our evaluation, which studies and shows negative effects of DP regarding model training. The results of this work provide perspectives for future research and applications whenever the fields of stress detection and data privacy intervene.

*Point of Interest Detection in Privacy-Sensitive Trajectories*
Data collected through mobile sensors on private and commercial devices can give valuable insights into mobility patterns and facilitate applications such as urban planning or traffic forecasting. At the same time, such data can carry immense privacy risks for the data producers. Stop detection approaches can reveal a person's points of interest (POI) by clustering temporal and spatial features, uncovering private attributes such as home or work addresses. Privacy-preserving mechanisms aim at hiding these POIs, for example via speed smoothing approaches that are able to preserve high data utility. We showed that such smoothing approaches are not sufficient to protect POIs in trajectories as they represent an anomaly in the typical movement pattern. We presented a novel attack *D-TOUR* that reveals POIs based on deviations from the optimal route, which can detect POIs with higher accuracy than state of the art stop detection algorithms.

*Tuning the Utility-Privacy Trade-Off in Trajectory Data*
Trajectory data, often collected on a large scale with mobile sensors in smartphones and vehicles, are a valuable source for realizing smart city applications, or for improving the user experience in mobile apps. But such data can also leak private information, such as a person's whereabouts and their points of interest (POI). These in turn can reveal sensitive information, for example a person's age, gender, religion, or home and work address. Location privacy preserving mechanisms (LPPM) can mitigate this issue by transforming data so that private details are protected. But privacy-preservation typically comes at the cost of a loss of utility. It can be challenging to find a suitable mechanism and the right settings to satisfy privacy as well as utility. We presented an interactive open-source framework Privacy Tuna, able to visualize trajectory data and intuitively estimate data utility and privacy while applying various LPPMs. The tool makes it easy for data owners to investigate the value of their data, choose a suitable privacy-preserving mechanism and tune its parameters to achieve a good utility-privacy trade-off.

*Distributed, Privacy-Aware Location Data Aggregation*
Multiple use cases require the aggregation of location data that is linked with a sensitive attribute. For example, in disaster management, healthcare resource allocation, or disease hotspot detection, the health status of users and their location needs to be processed to obtain an aggregated analysis, necessary to derive political measures or economic decisions. In this context not only the location of a person but also their health status is especially privacy-sensitive and needs to be protected. In search of a good utility-privacy trade-off we present a new anonymization technique *DIPALDA* that builds on the well established concept of k-anonymity, and introduce further privacy parameters, that can be individually controlled by the person sharing their data and is more interpretable than state of the art methods, such as Differential Privacy. By following a decentralized approach using cryptographic techniques we make sure that no trusted third party is required in the analysis process.

*Future Research Directions*
Regarding the privacy attributes of smartwatch health data, we want to investigate the impeding risk of identifying individuals in a health dataset collected from wearables. Constructing an attack that allows an adversary to break currently employed anonymization techniques in such data collections, would underline the need for privacy-preserving approaches. If an attack fails, it would on the other hand show, how privacy might be overbilling the utility in for these applications. As a defense mechanism we could test noise injection methods to hide an individual's identity inside the data. Other considerations include the generation of private synthetic health data. By replacing or enlarging

the currently available small-sized datasets for smartwatch scenarios, we would be able to better understand and improve on shortcomings in accuracy and limitations in scalability of current models.

## CUT - 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 ("Teilprojekte") called T1 to T5 and the supervision of these lies within different cities. The task of T1 is the administration of the digital twin, including writing a requirement catalog and planning the twin's architecture. T2'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 T3 is working on. T5's task is to make the generated knowledge public and share it with other cities and citizens. T4 is the measure that ScaDS.AI is a part of. Under the headline "experimental & transformative research", T4 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 T4 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. Further projects include analysis of large text data such as text or contributions generated in participation projects. The objective here is to find overall topics in all contributions to make evaluation for the responsible person easier. Another use case is the search function of the "Ratsinformationssystem" (ALLRIS). This system concludes all files and protocols, inquiries and answers of city council meetings. Within these files it can be difficult and confusing to find a specific topic. Our use case is working towards making the search within these files more efficient and user-friendly. Contact persons within the ScaDS.AI for the CUT project are Eric Peukert and Aruscha Kramm.

## DiGuRaL - Digital development of the urban space Leipzig (Digitale Gestaltung des urbanen Raums Leipzig)

*E. Peukert, M.L. Carnot, E. Rahm*

The smart city project DiGuRaL was launched in March 2023 with a planned duration of three years. ScaDS.AI is working together with the City of Leipzig, the city's municipal cleaning service, the Aufbauwerk as the project management and the companies CCC and CyFace. The overall goal is to equip the cleaning vehicles with cameras and sensors in order to ensure the quality of the streetscape with up-to-date data and to understand the development over time. All 24 city cleaning vehicles will be equipped with a smartphone that takes pictures and sensors that record environmental data such as particulate matter. One pilot vehicle will also be equipped with industrial cameras and a LiDAR sensor to record 3D point clouds. Data protection issues are an important aspect of the project. To this end, several anonymization methods will be used and a data protection report will be prepared by InfAI Infinity GmbH. The first use case agreed upon is the analysis of the condition of road signs. All signs are extracted from the captured images and classified according to their category (stop sign,

give way sign, etc.) and their condition (glued, faded, not visible, etc.). The information gained will help the city administration to tend to emerging problems faster and thus enhance roadway safety. Another application is the analysis of the clearance height above roads. According to the law, there must be a clearance of four meters to ensure safe passage. The recorded 3D point clouds are used to check whether trees are growing within this distance. The same applies to the clearance above footpaths and cycling paths, where the legally required clearance height is 2.50 meters. In the future, the list of applications will likely be expanded to include other use cases.

## 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) was originally planned to run for 30 months from 01/2021 to 06/2023. As part of a cost-neutral extension until 12/2023, the developed components were refined during extensive test runs and a comprehensive data set was generated for a final evaluation.

## K-M-I (Artificial and Human Intelligent)

*E. Rahm, M. Täschner, C. Augenstein (IWI), L. Peter, J. Pollack*

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. As part of a project extension, funds were raised for two additional pilot projects, which started in 11/2023.

### Come2Data - Competence Center for Interdisciplinary Data Sciences

*M. Täschner, E. Rahm*

As a data competence center (DKZ), Come2Data pursues a Saxon regional approach to imparting practice-oriented data competencies to science, but also to the areas of administration, society and, in the long term, to the economy. In doing so, it bundles existing data science training and support services as well as expertise in research data management, NFDI, high-performance computing and analysis methods for data-intensive interdisciplinary research applications, such as artificial intelligence (AI) and data modeling. Here, the diverse activities existing in Saxony are consolidated and merged into a sustainable offering. The basis is a comprehensive training and support offer in the fields of data integration, management, analysis, and publication. Come2Data creates an open research, support, networking and learning venue across Saxon locations to provide the training, support and knowledge offer to researchers, teachers, and learners as well as to the public via a central virtual platform. A sustainable support site is created, which will provide support by means of a knowledge base and the agency of experts. In addition, Come2Data creates a space in which solutions for real data and teaching problems are provided and made available in a didactically processed form. The project partners are TU Dresden, TU Chemnitz, the Saxon State and University Library Dresden (SLUB), and Leipzig University. As a project partner, Leipzig University will focus on knowledge and teaching offers for data competencies, will participate in the development of the technical infrastructure of the DKZ and will establish and expand the necessary target group and expert networks as well as the scientific communication of the DKZ. Furthermore, real data and teaching problems will be worked on within the framework of pilot projects based in Leipzig, and the resulting findings will be continuously incorporated into the knowledge and teaching offerings of Come2Data. Come2Data is funded by the BMBF (Funding reference: 16DKZ2044C) and will run for 3 years from 11/2023 till 11/2026.

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

*M. Joas, 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). In the first funding phase the platform focused on the development and optimization of a specific production process of mesenchymal stem cells which have a broad spectrum of applicability in cell therapy and current production heavily relying on manual and labor intensive steps. For the optimization of production by machine learning and AI, we focused with project partners on the automated confluence estimation (cell density) based on live-cell imaging and confluence

determination in 3D cell cultures by cell impedance measurements. To reduce labor intensive labeling of cells and provide tools to quickly adapt to other cell types, we studied and implemented active learning strategies and employed foundation models such as "Segment Anything" by Meta to learn robust AI-models to predict cell confluence in images. Further, these tools for AI-based process optimization is envisioned to be embedded in an digital documentation and dashboard system developed of project partners (ICCAS Leipzig, Prof. Neumuth). The first funding period will end in September 2024 and second phase funding (additional 3 years) is in the process of preparation.

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

*M. Joas, 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, data management and decentralized data analysis. As a result a data sharing and analysis concept was developed including several levels of multi-omics data analysis support. A cornerstone is the establishment of data analysis server making use of the Galaxy-community to provide a self-hosted analysis server at ScaDS.AI with a graphical user interface to perform standardized bioinformatic analyses. In the future we envision to provide additional levels of support and to attract more SaxoCell projects using the services as well as to establish close collaborations. Additionally, the established data analysis server can be of use beyond the SaxoCell cluster and depending on demand analysis server is scalable and decentralized by the possibility of an easy roll to more Galaxy-instances. The first funding period will end in September 2024 and second phase funding (additional 3 years) is in the process of preparation.

# 4 Publications and Theses

## Conference/Workshop Publications and Book Chapters


[1] Markus Bauer, Benjamin Uhrich, Martin Schäfer, Oliver Theile, Christoph Augenstein, and Erhard Rahm. "Multi-Modal Artificial Intelligence in Additive Manufacturing: Combining Thermal and Camera Images for 3D-Print Quality Monitoring." In: *Proceedings of the 25th International Conference on Enterprise Information Systems - Volume 1: ICEIS*. 2023, pp. 539–546. URL: https://www.scitepress.org/Link.aspx?doi=10.5220/0011967500003467.

[2] Erik Buchmann and Andreas Thor. "Online Exams in the Era of ChatGPT." In: *21. Fachtagung Bildungstechnologien (DELFI)*. Bonn: Gesellschaft für Informatik e.V., 2023, pp. 79–84. ISBN: 978-3-88579-732-6. DOI: 10.18420/delfi2023-15.

[3] Martin Franke, Victor Christen, Peter Christen, Florens Rohde, and Erhard Rahm. "(Privately) Estimating Linkage Quality For Record Linkage." In: *Proceedings of the 27td International Conference on Extending Database Technology, EDBT 2024, Paestum, Italy, March 25 - March 28, 2024*. accepted to appear. OpenProceedings.org, 2024.

[4] Johannes Frey, Marvin Hofer, and Sebastian Hellmann. "Studying Linked Data Accessibility Healthiness for the Long Tail of the Data Web." In: *9th Workshop on Managing the Evolution and Preservation of the Data Web*. Nov. 2023, pp. 55–64. URL: https://ceur-ws.org/Vol-3565/MEPDaW2023-paper2.pdf.

[5] M. Grimmer, T. Kaelble, E. Schulze, T. Rucks, and E. Rahm. "Extended Abstract: LID-DS 2021." In: *The 17th International Conference on Critical Information Infrastructures Security (CRITIS) 2022*. 2022.

[6] Martin Grimmer, Tim Kaelble, Felix Nirsberger, Emmely Schulze, Toni Rucks, Jörn Hoffmann, and Erhard Rahm. "Dataset Report: LID-DS 2021." In: *Critical Information Infrastructures Security*. Cham: Springer Nature Switzerland, 2023, pp. 63–73. ISBN: 978-3-031-35190-7.

[7] Victor Jüttner, Martin Grimmer, and Erik Buchmann. "ChatIDS: Explainable Cybersecurity Using Generative AI." In: *SECURWARE 2023 : The Seventeenth International Conference on Emerging Security Information, Systems and Technologies*. best paper award. 2023. URL: https://www.thinkmind.org/index.php?view=article&articleid=securware_2023_1_20_30025.

[8] Aruscha Kramm, Julia Friske, and Eric Peukert. "Detecting Floors in Residential Buildings." In: *KI 2023: Advances in Artificial Intelligence: 46th German Conference on AI, Berlin, Germany, September 26–29, 2023, Proceedings*. Berlin, Germany: Springer-Verlag, 2023, pp. 130–143. ISBN: 978-3-031-42607-0. DOI: 10.1007/978-3-031-42608-7_11. URL: https://doi.org/10.1007/978-3-031-42608-7_11.

[9] Lucas Lange, Borislav Degenkolb, and Erhard Rahm. "Privacy-Preserving Stress Detection Using Smartwatch Health Data." In: *4. Interdisciplinary Privacy & Security at Large Workshop, INFORMATIK 2023*. Sept. 2023.

[10] Lucas Lange, Maja Schneider, Peter Christen, and Erhard Rahm. "Privacy in Practice: Private COVID-19 Detection in X-Ray Images." In: *20th International Conference on Security and Cryptography (SECRYPT 2023)*. SciTePress, July 2023, pp. 624–633. ISBN: 978-989-758-666-8. DOI: 10.5220/0012048100003555. URL: https://www.scitepress.org/PublicationsDetail.aspx?ID=UnmHswKdjFk=.

[11] Erik Morawetz, Nadine Hahm, and Andreas Thor. "Automatisierte Bewertung und Feedback-Generierung für grafische Modellierungen und Diagramme mit FeeDI." In: *21. Fachtagung Bildungstechnologien (DELFI)*. Bonn: Gesellschaft für Informatik e.V., 2023, pp. 97–102. ISBN: 978-3-88579-732-6. DOI: 10.18420/delfi2023-18.

[12] Erik Morawetz, Nadine Hahm, and Andreas Thor. "E-Assessment für Entity-Relationship-Diagramme mit FeeDI." In: *21. Fachtagung Bildungstechnologien (DELFI)*. Bonn: Gesellschaft für Informatik e.V., 2023, pp. 275–276. ISBN: 978-3-88579-732-6. DOI: 10.18420/delfi2023-53.

[13] Daniel Obraczka, Alieh Saeedi, Victor Christen, and Erhard Rahm. "Big Data Integration for Industry 4.0." In: *Digital Transformation - Core Technologies and Emerging Topics from a Computer Science Perspective*. Ed. by Birgit Vogel-Heuser and Manuel Wimmer. Springer, 2022, pp. 247–268. DOI: 10.1007/978-3-662-65004-2_10. URL: https://doi.org/10.1007/978-3-662-65004-2%5C_10.

[14] Martin Petersohn, Konrad Schöbel, and Andreas Thor. "DMT-Magic: Interaktives E-Assessment in der Datenbank-Lehre mit Jupyter Notebooks." In: *21. Fachtagung Bildungstechnologien (DELFI)*. Bonn: Gesellschaft für Informatik e.V., 2023, pp. 263–264. ISBN: 978-3-88579-732-6. DOI: 10.18420/delfi2023-47.

[15] Martin Petersohn, Konrad Schöbel, and Andreas Thor. "Kopplung von Jupyter Notebooks mit externen E-Assessment-Systemen am Beispiel des Data Management Testers." In: *21. Fachtagung Bildungstechnologien (DELFI)*. Bonn: Gesellschaft für Informatik e.V., 2023, pp. 85–90. ISBN: 978-3-88579-732-6. DOI: 10.18420/delfi2023-16.

[16] Florens Rohde, Martin Franke, Victor Christen, and Erhard Rahm. "Value-specific Weighting for Record-level Encodings in Privacy-Preserving Record Linkage." In: *Datenbanksysteme für Business, Technologie und Web (BTW 2023), 20. Fachtagung des GI-Fachbereichs, Datenbanken und Informationssysteme" (DBIS), 06.-10, März 2023, Dresden, Germany, Proceedings*. 2023, pp. 439–460. DOI: 10.18420/BTW2023-21.

[17] Christopher Rost, Kevin Gómez, Peter Christen, and Erhard Rahm. "Evolution of Degree Metrics in Large Temporal Graphs." In: *Datenbanksysteme für Business, Technologie und Web (BTW 2023), 20. Fachtagung des GI-Fachbereichs „Datenbanken und Informationssysteme" (DBIS), 06.-10, März 2023, Dresden, Germany, Proceedings*. 2023, pp. 485–507. DOI: 10.18420/BTW2023-23. URL: https://doi.org/10.18420/BTW2023-23.

[18] Christopher Rost, Riccardo Tommasini, Angela Bonifati, Emanuele Della Valle, Erhard Rahm, Keith W. Hare, Stefan Plantikow, Petra Selmer, and Hannes Voigt. "Seraph: Continuous Queries on Property Graph Streams." In: *Proceedings of the 27td International Conference on Extending Database Technology, EDBT 2024, Paestum, Italy, March 25 - March 28, 2024*. accepted to appear. OpenProceedings.org, 2024.

[19] Maja Schneider, Lukas Gehrke, Peter Christen, and Erhard Rahm. "D-TOUR: Detour-based point of interest detection in privacy-sensitive trajectories." In: *Proc. 52th annual conf. of German Informatics society INFORMATIK 2022*. Vol. P-326. LNI. 2022, pp. 219–230. DOI: 10.18420/inf2022_20.

[20] Maja Schneider, Jonathan Schneider, Lea Löffelmann, Peter Christen, and Erhard Rahm. "Tuning the Utility-Privacy Trade-Off in Trajectory Data." In: *26th Int. Conf. on Extending Database Technology (EDBT)*. 2023.

[21] Benjamin Uhrich, Nikolai Hlubek, Tim Häntschel, and Erhard Rahm. "Using Differential Equation-inspired Machine Learning for Valve Faults Prediction." In: *2023 IEEE 21st International Conference on Industrial Informatics (INDIN)*. Aug. 2023, pp. 1–8. URL: https://ieeexplore.ieee.org/abstract/document/10217897.

[22] Benjamin Uhrich, Martin Schäfer, Oliver Theile, and Erhard Rahm. "Using Physics-informed Machine Learning to Optimize 3D Printing Processes." In: *Proceedings of the 2nd International Conference on Progress in Digital and Physical Manufacturing (ProDPM'21)*. June 2023, pp. 206–221. URL: https://link.springer.com/chapter/10.1007/978-3-031-33890-8_18.

[23] Felix Vogel and Lucas Lange. "Privacy-Preserving Sentiment Analysis on Twitter." In: *SKILL 2023*. Sept. 2023.

### Journal Publications

[1] Daniel Ayala, Inma Hernández, David Ruiz, and Erhard Rahm. "LEAPME: Learning-based Property Matching with Embeddings." In: *Data & Knowledge Engineering* 137 (2022), p. 101943. ISSN: 0169-023X. DOI: https://doi.org/10.1016/j.datak.2021.101943.

[2] Daniel Ayala, Inma Hernández, David Ruiz, and Erhard Rahm. "Multi-source dataset of e-commerce products with attributes for property matching." In: *Data in Brief* 41 (2022), p. 107884. ISSN: 2352-3409. DOI: https://doi.org/10.1016/j.dib.2022.107884.

[3] Victor Christen, Tim Häntschel, Peter Christen, and Erhard Rahm. "Privacy-Preserving Record Linkage using Autoencoders." In: *Int. Journal of Data Science and Analytics* 15.4 (2023), pp. 347–357. DOI: 10.1007/S41060-022-00377-2.

[4] Nadine Hahm, Erik Morawetz, and Andreas Thor. "8. Die Materialität analog-digitaler Schnittstellen: Usability-Testung Stift-basierter Eingabegeräte für E-Assessment-Szenarien." In: *Digitale Lehre im Rahmen der Grundlagenausbildung in MINT-Fächern an Hochschulen* (2023), p. 127.

[5] Toralf Kirsten, Frank A Meineke, Henry Loeffler-Wirth, Christoph Beger, Alexandr Uciteli, Sebastian Stäubert, Matthias Löbe, René Hänsel, Franziska G Rauscher, Judith Schuster, et al. "The Leipzig Health Atlas—An Open Platform to Present, Archive, and Share Biomedical Data, Analyses, and Models Online." In: *Methods of Information in Medicine* 61.S 02 (2022), e103–e115.

[6] Y.C. Lin, P. Hoffmann, and E Rahm. "Enhancing Cross-lingual Biomedical Concept Normalization Using Deep Neural Network Pretrained Language Models." In: *SN COMPUT. SCI.* 3.387 (2022). DOI: https://doi.org/10.1007/s42979-022-01295-7.

[7] Daniel Obraczka and Erhard Rahm. "Fast Hubness-Reduced Nearest Neighbor Search for Entity Alignment in Knowledge Graphs." In: *SN Comput. Sci.* 3.6 (2022), p. 501. DOI: 10.1007/S42979-022-01417-1. URL: https://doi.org/10.1007/s42979-022-01417-1.

[8] Christopher Rost, Kevin Gómez, Matthias Täschner, Philip Fritzsche, Lucas Schons, Lukas Christ, Timo Adameit, Martin Junghanns, and Erhard Rahm. "Distributed temporal graph analytics with Gradoop." In: *VLDB J.* 31.2 (2022), pp. 375–401. DOI: 10.1007/S00778-021-00667-4. URL: https://doi.org/10.1007/s00778-021-00667-4.

[9] Georg Walther, Christian Martin, Amelie Haase, Ulf Nestler, and Stefan Schob. "Machine Learning for Rupture Risk Prediction of Intracranial Aneurysms: Challenging the PHASES Score in Geographically Constrained Areas." In: *Symmetry* 14.5 (2022). ISSN: 2073-8994. DOI: 10.3390/sym14050943. URL: https://www.mdpi.com/2073-8994/14/5/943.

[10] Richard A Williams, Lutz Bornmann, and Andreas Thor. "Panel Data and Multilevel Analyses of Academic Publishing Success Paper." In: *SSRN* (2022). DOI: 10.2139/ssrn.4093415.

### Technical/arXiv Reports

[1] Marvin Hofer, Daniel Obraczka, Alieh Saeedi, Hanna Köpcke, and Erhard Rahm. *Construction of Knowledge Graphs: State and Challenges*. 2023. DOI: 10.48550/ARXIV.2302.11509. arXiv: 2302.11509 [cs]. URL: https://doi.org/10.48550/arXiv.2302.11509.

[2] Timm Intemann, Knut Kaulke, Dennis-Kenji Kipker, Vanessa Lettieri, Christoph Stallmann, Carsten Schmidt, Lars O. Geidel, Martin Bialke, Christopher Hampf, Dana Stahl, Martin Lablans, Martin Franke, Klaus Peter Kraywinkel, Joachim Kieschke, Sebastian Bartholomäus, Anatol-Fiete Näher, Galina Tremper, Mohamed Lambarki, Stefanie March, Fabian Prasser, Anna Christine Haber, Johannes Drepper, Irene Schlünder, Toralf Kirsten, Iris Pigeot, Ulrich Sax, Benedikt Buchner, Wolfgang Ahrens, and Sebastian Claudius Semler. *White Paper - Verbesserung des Record Linkage für die Gesundheitsforschung in Deutschland*. 2023. DOI: 10.4126/FRL01-006461895.

[3] Lucas Lange, Maja Schneider, Peter Christen, and Erhard Rahm. *Privacy in Practice: Private COVID-19 Detection in X-Ray Images (Extended Version)*. Apr. 2023. DOI: 10.48550/arXiv.2211.11434. arXiv: 2211.11434 [cs]. URL: http://arxiv.org/abs/2211.11434.

[4] Lucas Lange, Tobias Schreieder, Victor Christen, and Erhard Rahm. *Privacy at Risk: Exploiting Similarities in Health Data for Identity Inference*. Aug. 2023. DOI: 10.48550/arXiv.2308.08310. arXiv: 2308.08310 [cs]. URL: http://arxiv.org/abs/2308.08310.

### Ph. D. Theses

[1] Martin Franke. "Scalable, Accurate and Secure Privacy-Preserving Record Linkage (submitted)." PhD. Leipzig University, Oct. 2023.

[2] Christopher Rost. "Scalable Management and Analysis of Temporal Property Graphs (submitted)." PhD. Leipzig University, Dec. 2023.

### Bachelor and Master Theses

[1] Karim Abrik. "Performanceanalyse der Commit-Historie von Gradoop." B.Sc. Leipzig University in Cooperation with URZ, 2023.

[2] Daniel Alker. "Establishment of a Machine Learning Pipeline for Battery Electric Trucks Mission Profiles." M.Sc. Leipzig University in Cooperation with MAN Truck & Bus, 2022.

[3] Malek Bakleh. "PPRL mit Multi-Feature-Klassifikation von Record-level Bloom Filtern." B.Sc. Leipzig University, 2023.

[4] Simon Bordewisch. "Flexible Graphstream-Analyse mittels Apache Flinks DataStream-API." M.Sc. Leipzig University, 2022.

[5] Lukas Burges. "Konzeption und Implementierung eines Iterativen Abfragebasierten Privacy-Preserving Record Linkage Protokolls." B.Sc. Leipzig University, 2022.

[6] Manh Le Dac. "Extension of the record linkage benchmark tool Snowman for tag-based linkage result analysis." B.Sc. Leipzig University, 2023.

[7] Duc Dung Dao. "Konzeption und Realisierung einer Anwendung zur Generierung von personenbezogenen Daten." B.Sc. Leipzig University, 2022.

[8] Borislav Degenkolb. "Entwicklung eines Privatsphäre-erhaltenden Transformers zur Stresserkennung anhand von Smartwatch-Health-Daten." M.Sc. Leipzig University, 2023.

[9] Duong Trung Duong. "Scalable and Accurate Decision Tree Learning for Entity Resolution." M.Sc. Leipzig University, 2022.

[10] Niklas Ehrlich. "Analyse des Fahrverhaltens von Versicherungskunden im Zusammenhang mit besonderen Ereignissen anhand von Telematik-Daten der HUK-COBURG." M.Sc. Leipzig University in Cooperation with HUK-COBURG, 2023.

[11] Maurice Eisenblätter. "Optimierung des GRADOOP I/O mittels Columnar Store Apache Parquet." B.Sc. Leipzig University, 2022.

[12] Christiane Ernst. "Prädiktion von Auslastung und Warteschlangen an Ladesäulen." M.Sc. Leipzig University in Cooperation with Porsche AG, 2023.

[13] Merlin Flach. "Methods to Cross the simulation-to-reality gap." B.Sc. Leipzig University, 2023.

[14] Manuel Friedrich. "Secure encoding for PPRL with shared secrets." B.Sc. Leipzig University, 2022.

[15] Martin Frühauf. "Neuer Webauftritt der Datenbankgruppe." B.Sc. Leipzig University, 2022.

[16] Daniel Grohmann. "Kolorieren Historischer Fotos der Stadt Leipzig mit Deep-Learning-Techniken." B.Sc. Leipzig University, 2023.

[17] Tim Häntschel. "Evaluation of Autoencoders for encrypting Bloom-Filters." B.Sc. Leipzig University, 2022.

[18] Lukas Hempel. "Performance-Verbesserung von HIDS durch Rückspielen menschlicher Auswertungen." M.Sc. Leipzig University, 2023.

[19] Alexander Hinz. "Anomalieerkennung in Zeitreihen mithilfe von unüberwachtem maschinellem Lernen." M.Sc. Leipzig University in Cooperation with UFZ, 2023.

[20] Jonathan Huthmann. "Vorhersage der Kontrastmittelanreicherung in 3D-Mamma MRT-Aufnahmen durch Einsatz von Deep Learning." M.Sc. Leipzig University, 2022.

[21] Mohammad Issa. "Detection of Segmentation of Lanes and Directions on Smartphone." M.Sc. Leipzig University (in coperation with ASAP GmbH), 2023.

[22] Tim Kaelble. "Host-Based Intrusion Detecion mit LSTMs." M.Sc. Leipzig University, 2022.

[23] Mehdi Karbalai. "Realisierung einer Android-App zur Extraktion der Metainformationen von Klausuren." B.Sc. Leipzig University, 2023.

[24] Felix Kirchgäßner. "Von gewöhnlichen und partiellen Differentialgleichungen inspirierte neuronale Netze." Diplom. Leipzig University in Cooperation with HUK-COBURG, 2023.

[25] Julius Kluge. "Machine Learning für Schema Mapping auf heterogenen Datenquellen." M.Sc. Leipzig University, 2022.

[26] Michael Koch. "Privatsphäre-erhaltende Vorhersage der Überlebenszeiten von Magenkrebspatienten." M.Sc. Leipzig University, 2022.

[27] Viktor Koch. "Building an Overlay Search Application for the XITASO Knowledge Ecosysteme." B.Sc. Leipzig University in Cooperation with XITASO GmbH, 2023.

[28] Jonas König. "Model training of a simulated self-driving vehicle using an evolution-based neural network approach." B.Sc. Leipzig University, 2022.

[29] Andreas Kretschmer. "Sports Analytics im Handball: Erstellen eines Modells zum Berechnen von erwarteten Toren auf Basis von Raum-Zeit Sensordaten." M.Sc. Leipzig University, 2023.

[30] Adrian Kuhn. "Semantic Classification of Datasets - An Approach to Enriching Metadata on the DBpedia Databus." B.Sc. Leipzig University, 2023.

[31] Lucas Lange. "Privacy-Preserving Detection of COVID-19 in X-Ray Images." M.Sc. Leipzig University, 2022.

[32] Kai Lanzendorf. "Studies on threshold selection methods in LID-DS." B.Sc. Leipzig University, 2023.

[33] Steven Lehmann. "Link Prediction als Möglichkeit zur automatisierten Personaleinsatzplanung." M.Sc. Leipzig University, 2023.

[34] Xiaoxi Li. "Comparison of single-machine and distributed calculation of temporal degree metrics." B.Sc. Leipzig University, 2023.

[35] Lea Löffelmann. "Tabu Search Algorithmus für das Vehicle Routing Problem eines Logistikunternehmens." B.Sc. Leipzig University, 2022.

[36] Maximilian Martin. “Evaluation von Graph Stream Analyse Pipelines mit Apache Flink.” B.Sc. Leipzig University, 2023.

[37] Anna Matusevich. “Untersuchung der Entwicklung der Performanz von Tomcat.” B.Sc. Leipzig University in Cooperation with URZ, 2023.

[38] Schaller Maximilian. “Investigating Training an Agent to Drive a Vehicle in a Simulated Environment Using Reinforcement Learning.” M.Sc. Leipzig University, 2023.

[39] Marlene Mertens. “Classification of Cervical Cancer using Deep Neural.” M.Sc. Leipzig University, 2023.

[40] Elias Messner. “Robustness of Privacy Preserving Record Linkage against Dataset Variation.” B.Sc. Leipzig University, 2023.

[41] Maximilian Mischinger. “Clustering von Alarmen hostbasierter Intrusion Detection Systeme.” B.Sc. Leipzig University, 2022.

[42] Felix Nirsberger. “Entwicklung eines HIDS auf Basis des LID-DS unter Verwendung einer Self-Organizing Map.” M.Sc. Leipzig University, 2022.

[43] Anja Ohlhäuser. “Anomaliebasierte Angriffserkennung auf einem Hostsystem basierend auf Systemcalls und Netzwerkdaten.” M.Sc. Leipzig University, 2023.

[44] Abed Alaziz Owidat. “Entwicklung einer Android-App zur Erfassung und Visualisierung von ornithologischen Daten (Backend).” B.Sc. Leipzig University, 2023.

[45] Abed Alaziz Owidat. “Entwicklung einer Android-App zur Erfassung und Visualisierung von ornithologischen Daten (Frontend).” B.Sc. Leipzig University, 2023.

[46] Nils Pfeifer. “Modeling Heat Transport in Selective Laser Melting with Physics-informed Neural Networks.” M.Sc. Leipzig University, 2023.

[47] Julian Pielmaier. “Visualisierung zeitlicher Graph Metriken.” B.Sc. Leipzig University, 2023.

[48] Jacob Pollack. “Product matching using multimodal deep learning.” M.Sc. Leipzig University, 2023.

[49] Leonie Preker. “Evaluating embedding methods for genomic data.” M.Sc. Leipzig University, 2022.

[50] Noah Rasp. “A Study on the Impact of Class Imbalance on CNNs for Bee Health Detection.” B.Sc. Leipzig University, 2022.

[51] Lukas Reinhardt. “HIDS zur Identifizerung von Hardwaremanipulation.” B.Sc. Leipzig University, 2022.

[52] Stefan Rosenlund. “Extending Peass to Detect Performance Changes of Tomcat.” B.Sc. Leipzig University in Cooperation with URZ, 2022.

[53] Toni Rucks. “Host-based Intrusion Detecion unter der Verwendung von Systemcall-Entity-Graphen.” M.Sc. Leipzig University, 2023.

[54] Axel Schuster. “Organizing Time Series Data by Temporal Property Graphs.” M.Sc. Leipzig University, 2023.

[55] Leo Seeger. “Benchmarking of Federated Learning Tools for the Analysis of Gene Expression Data.” M.Sc. TU Berlin, in cooperation with ScaDS.AI, 2023.

[56] Elena Senger. “Entity Matching unter Verwendung von Geodaten am Beispiel von Daten der Deutschen Bahn.” M.Sc. Leipzig University in Cooperation with Deutsche Bahn AG, 2022.

[57] Felix Vogel. “Privatsphäre-erhaltende Sentiment Analyse auf Twitter.” B.Sc. Leipzig University, 2023.

[58] Robert Weiske. “Vorhersage der Clusterqualität unter Verwendung von Ähnlichkeitsgraph-Charakteristiken.” M.Sc. Leipzig University, 2023.

[59] Nils Wenzlitschke. "Privacy-Preserving Smartwatch Health Data Generation For Stress Detection Using GANs." M.Sc. Leipzig University, 2023.

[60] Jonathan Weske. "Standzeitvorhersage innerhalb eines Fahrradverleihsystems." B.Sc. Leipzig University in Cooperation with Nextbike, 2022.

[61] Andre Wille. "Erstellung einer Heatmap auf Grundlage temporaler Graphen." B.Sc. Leipzig University, 2023.

[62] Konstantin Wilson. "Temporal Graph Metrics." B.Sc. Leipzig University, 2023.

[63] Sung Geun Yun. "Windowed Graph Stream Pattern Matching using Subgraph-Isomorphism." B.Sc. Leipzig University, 2023.

[64] Johanna Zitt. "Denoising Cryoseismological Distributed Acoustic Sensing Data Using an U-net Autoencoder." M.Sc. Leipzig University, 2023.

# 5 Talks

[1] Martin Grimmer. *Alarmanlagen für Server – wie können Angriffe auf IT-Systeme frühzeitig erkannt werden?* IT-Kongress „Deutschland – Zukunft – Digital", Leipzig, Germany. Sept. 2022. URL: https://it-mitteldeutschland.de/event/it-kongress-deutschland-zukunft-digital-tag/.

[2] Martin Grimmer. *Dataset Report : LID-DS 2021*. CRITIS 2022 (The 17th International Conference On Critical Information Infrastructures Security), Munich, Germany. Sept. 2022. URL: https://critis2022.comtessa.org/program.

[3] Martin Grimmer. *Wie schützen Unternehmen ihre Server gegen Zero-Day-Angriffe?* Webinar of the week at ComConsult, Online. Feb. 2023. URL: https://www.comconsult.com/wie-schuetzen-unternehmen-ihre-server-gegen-zero-day-angriffe/.

[4] Marvin Hofer. *Bootstrapping Knowledge Graphs using DBpedia's Ecosytem - AKSW.org KG 2.0*. DBpedia Day @ SEMANTICS 2022. Sept. 2022. URL: https://www.dbpedia.org/events/dbpedia-day-semantics-2022/.

[5] Marvin Hofer. *Studying Linked Data Accessibility Healthiness for the Long Tail of the Data Web*. MEPDaW @ ISWC 2023. Nov. 2023. URL: https://mepdaw-ws.github.io/2023/.

[6] Daniel Obraczka. *Blocking Methods for Entity Resolution on Knowledge Graphs*. DBpedia Day @ SEMANTICS 2023. Sept. 2023. URL: https://www.dbpedia.org/events/dbpedia-day-semantics-2023/.

[7] Daniel Obraczka. *Connecting the Right Dots: Entity Resolution on Knowledge Graphs*. ScaDS.AI Summer School 2022. July 2022. URL: https://scads.ai/education/summer-schools/summer-school-2022/.

[8] Daniel Obraczka. *Fast Hubness-Reduced Nearest Neighbor Search for Entity Alignment in Knowledge Graphs*. DBpedia Day @ SEMANTICS 2022. Sept. 2022. URL: https://www.dbpedia.org/events/dbpedia-day-semantics-2022/.

[9] Erhard Rahm. *Data Integration for Knowledge Graphs.* Keynote LWDA Conf., Marburg. Oct. 2023.

[10] Erhard Rahm. *Multi-Source Data Matching and Clustering.* Virtual Lecture Series on Database Research, Hasso Plattner Institute (HPI) Potsdam. Jan. 2022.

[11] Erhard Rahm. *Research and Transfer at Data Science Center ScaDS.AI.* Invited Talk, 13. Digitalisierungskonferenz, Köthen. May 2022.

[12] Florens Rohde. *Value-specific Weighting for Record-level Encodings in Privacy-Preserving Record Linkage*. BTW Conf., Dresden. Mar. 2023.

[13] Christopher Rost. *Distributed temporal graph analytics with Gradoop*. Invited talk at 6th Joint Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA), co-located with SIGMOD 2023, Seattle, United States. June 2023. URL: https://gradesnda.github.io.

[14] Christopher Rost. *Evolution of Degree Metrics in Large Temporal Graphs*. BTW Conf., Dresden. Mar. 2023.

[15] Christopher Rost. *Graph Stream Zoomer: A window-based graph stream grouping system based on Apache Flink*. FOSDEM 2023, Brussels, Belgium. Feb. 2023. URL: https://archive.fosdem.org/2023/schedule/event/graph_grouping_zoomer/.

[16] Maja Schneider. *D-TOUR: Detour-based point of interest detection in privacy-sensitive trajectories*. 3. Interdisciplinary Privacy & Security at Large Workshop / Privacy&Security@Large. Sept. 2022. URL: https://pretalx.com/informatik-2022/talk/WTBWCD/.

[17] Maja Schneider. *Tuning the Utility-Privacy Trade-Off in Trajectory Data*. Demonstration. Mar. 2023. URL: http://edbticdt2023.cs.uoi.gr/.

[18] Benjamin Uhrich. *Multi-Modal Artificial Intelligence in Additive Manufacturing: Combining Thermal and Camera Images for 3D-Print Quality Monitoring*. ICEIS 2023 (The 25th International Conference On Enterprise Information Systems), Prague, Czech Republic. Apr. 2023. URL: https://iceis.scitevents.org/?y=2023.

[19] Benjamin Uhrich. *Using Differential Equation Inspired Machine Learning for Valve Faults Prediction*. INDIN 2023 (IEEE 21st International Conference on Industrial Informatics), Lemgo, Germany. July 2023. URL: https://2023.ieee-indin.org/.