UNIVERSITÄT LEIPZIG

ScaDS.AI DRESDEN LEIPZIG

# SCADS.AI TOPICS OF THE DATABASE GROUP

**Erhard Rahm**

ScaDS.AI DRESDEN LEIPZIG

## GERMAN AI CENTERS

UNIVERSITÄT LEIPZIG

- 5 new national AI centers in Germany (in addition to older DFKI)
  - Berlin (BIFOLD)
  - Dortmund / Bonn (ML2R)
  - Dresden / Leipzig (ScaDS.AI)
  - Munich (MCML)
  - Tübingen (tuebingen.ai)

- **SCADS.AI:** Center for **Sca**lable **D**ata Analytic**S** and **A**rtificial **I**ntelligence
- extends the Big Data center ScaDS Dresden/Leipzig, started in 2014
- since Nov. 2019: AI center ScaDS.AI

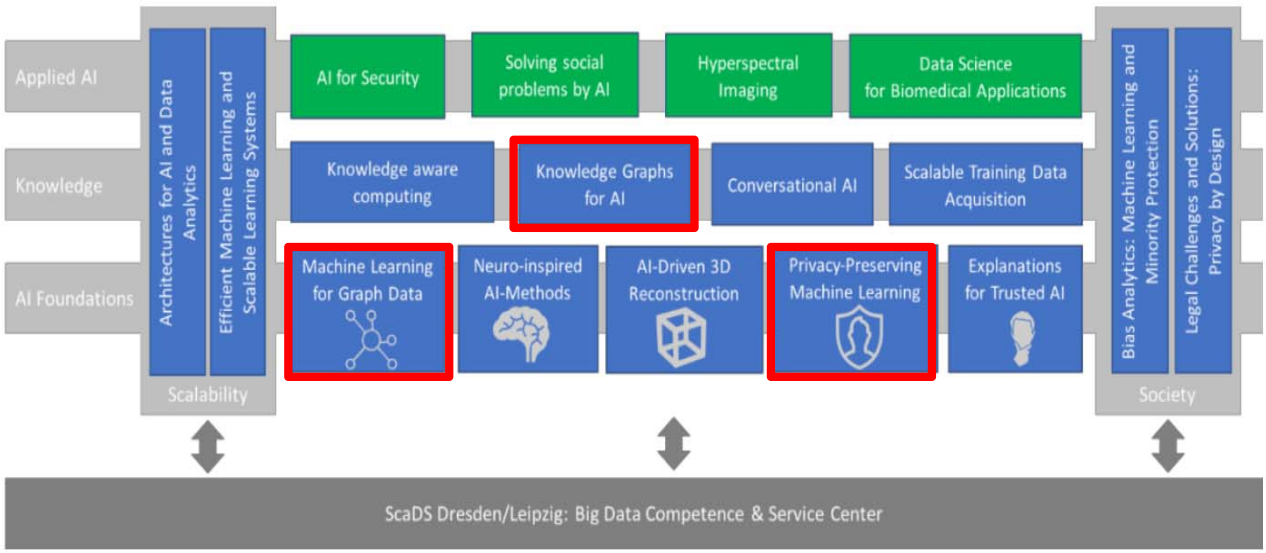
- **Highlights ScaDS.AI**
  - 8 new AI professorships (4 at Univ. of Leipzig)
  - graduate school for Ph.D. students
  - Demo and Living Lab





# SCADS.AI RESEARCH TOPICS



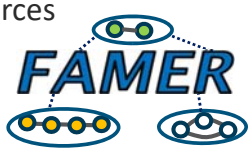


ScaDS.AI DRESDEN LEIPZIG

# TOPICS OF DATABASE GROUP

UNIVERSITÄT LEIPZIG

- Knowledge graphs
  - refinement of large knowledge graphs from many data sources
  - incremental matching and clustering of new entities
  - research based on prototype FAMER




- Privacy-preserving machine learning
  - combination with privacy-preserving record linkage (PPRL)
  - use of anonymisation techniques like differential privacy
  - based on PPRL prototype PRIMAT


- ML on dynamic graph data
  - based on graph analysis platform GRADOOP
  - utilization of context knowledge (e.g., graph embeddings) for improved predictions, also for data streams

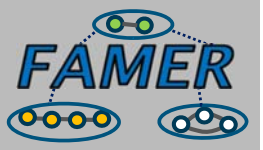



ScaDS.AI DRESDEN LEIPZIG

# AGENDA

UNIVERSITÄT LEIPZIG

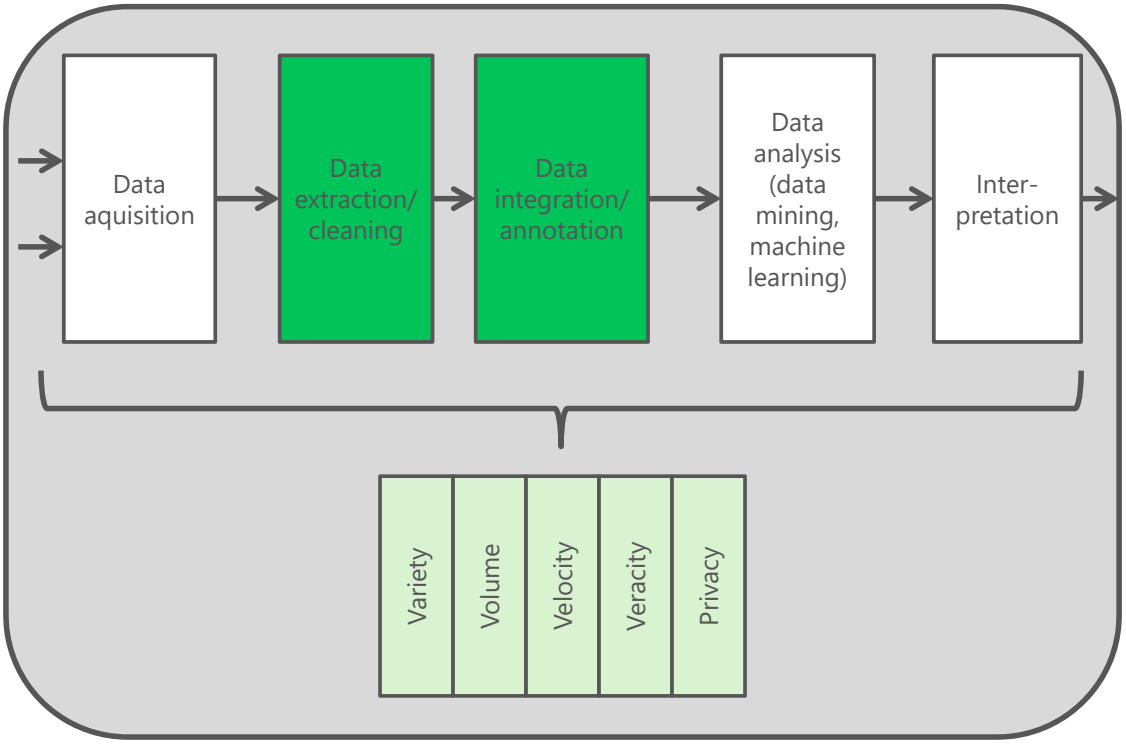
- Introduction
- Holistic entity resolution for knowledge graph completion
  - data integration challenges
  - entity resolution / FAMER
  - incremental entity resolution




- Privacy-related research


- Graph data analysis

# BIG DATA ANALYSIS/ML PIPELINE





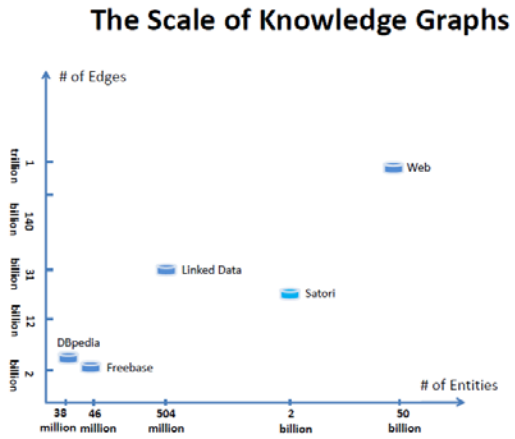

# DATA INTEGRATION

- provision of uniform access to data originating from multiple, autonomous sources
- **physical data integration**
  - original data is combined within a new dataset / database for access and analysis
  - approach of data warehouses, knowledge graphs and most Big Data applications
- **virtual data integration**
  - data is accessed on demand in their original data sources, e.g. based on an additional query layer
  - approach of federated databases and linked data

# KNOWLEDGE GRAPHS



- uniform representation and semantic categorization of entities of different types
  - examples: DBPedia, Yago, Wikidata, Google KG, MS Satori, Facebook, …
  - entities often extracted from other resources (Wikipedia, Wordnet etc.) or web pages, documents, web searches etc.
  - knowledge graphs provide valuable background knowledge for enhancing entities (based on prior *entity linking),* improving search results …

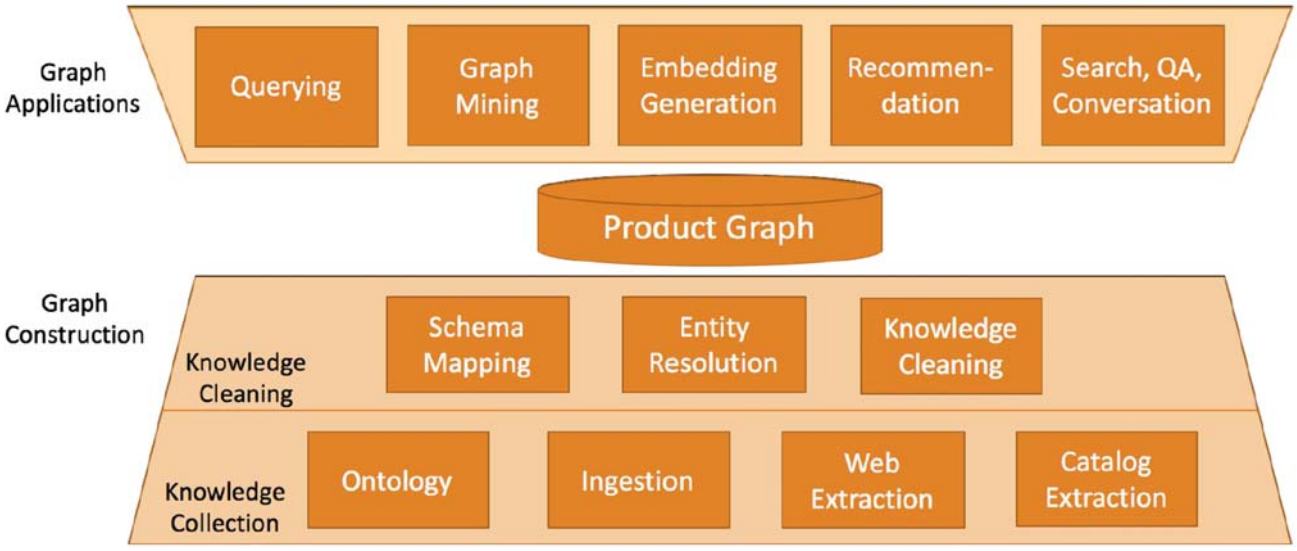

Shao, Li, Ma (Microsoft Asia): Distributed Real-Time Knowledge Graph Serving (slides, 2015)





# EXAMPLE: PRODUCT KNOWLEDGE GRAPH






from: Dong. KDD2018

# KEY DATA INTEGRATION CHALLENGES



- data quality
  - unstructured, semi-structured sources
  - need for data cleaning and enrichment
  - need for advanced matching and clustering of entities
- large-scale data integration
  - large data/metadata volume or/and many sources
  - improve runtime by reducing search space (e.g. with blocking) and parallel processing (Hadoop clusters, GPUs, etc.)
  - many sources require *holistic data integration*: clustering of schema elements and entities, not only binary matching
- support for evolution and change
  - addition of new sources and new entities without having to integrate everything again
  - incremental / dynamic vs batch / static data integration
- privacy for sensitive data
  - privacy-preserving record linkage and data mining





# ENTITY RESOLUTION (DEDUPLICATION)



- identification of semantically equivalent objects
  - within one data source or between different sources
- original focus on structured (relational) data, e.g. customer data

| *Cno* | *LastName* | *FirstName* | *Gender* | *Address* | *Phone/Fax* |
|---|---|---|---|---|---|
| 24 | Smith | Christoph | M | 23 Harley St, Chicago IL, 60633-2394 | 333-222-6542 / 333-222-6599 |
| 493 | Smith | Kris L. | F | 2 Hurley Place, South Fork MN, 48503-5998 | 444-555-6666 |

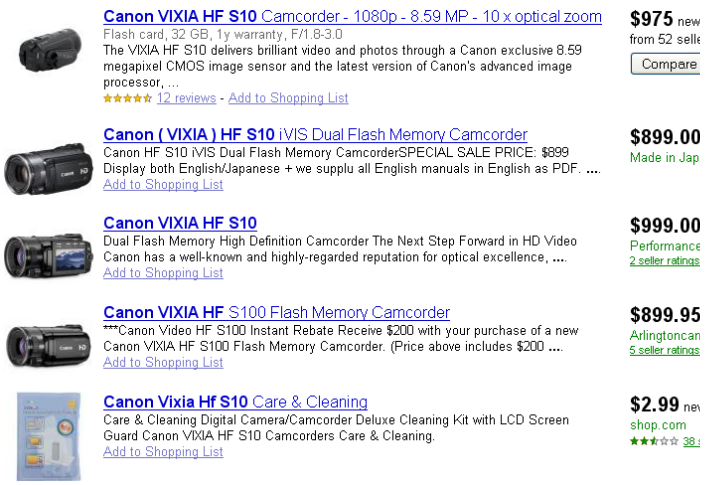
| *CID* | *Name* | *Street* | *City* | *Sex* |
|---|---|---|---|---|
| 11 | Kristen Smith | 2 Hurley Pl | South Fork, MN 48503 | 0 |
| 24 | Christian Smith | Hurley St 2 | S Fork MN | 1 |

# HOLISTIC DATA INTEGRATION USE CASE



## Integration of product offers in comparison portal

- thousands of data sources (shops/merchants)
- millions of products and product offers
- continuous changes
- many similar, but different products
- low data quality








# SAMPLE CAMERA OFFERS



| property | value |
|---|---|
| "35mm equivalent" | "25-300mm" |
| "<page title>" | "Nikon Coolpix S6800 Digital Camera (Black) \| UK Digital Cameras" |
| "brand" | "Nikon" |
| "camera resolution" | "16 Megapixels" |
| "colour" | "Black", |
| "features" | "Slimline" |
| "hd video" | "Full HD (1080P)" |
| "lcd size" | "3.0"" |
| "lens tele mm" | "300" |
| "lens wide mm" | "25" |
| "mpn" | "VNA520E1" |
| "optical zoom" | "23" |
| "optical zoom range" | "18x and higher" |

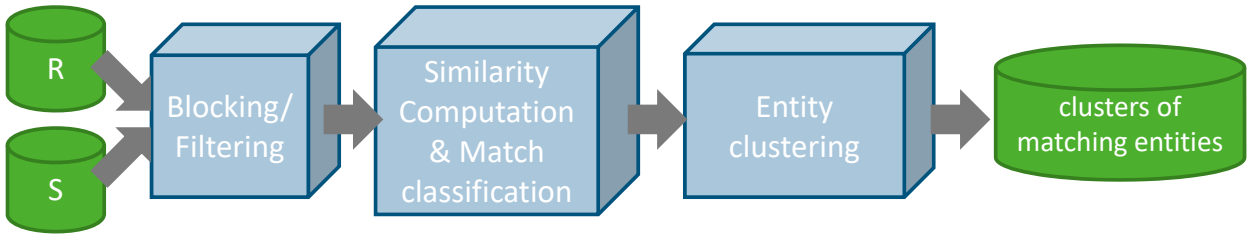
| property | value |
|---|---|
| "<page title>" | "Nikon Coolpix S6800 Price in India with Offers, Reviews & Full Specifications \| PriceDekho.com" |
| "color" | "Black", |
| "amazon" | "Infibeam Ebay Homeshop18 Snapdeal Flipkart" |
| "digital zoom" | "4x" |
| "bangalore" | "Hyderabad Chennai Mumbai Delhi Pune" |
| "approx resolution" | "16 MP" |
| "external memory" | "Yes" |
| "face detection" | "NA" |
| "gps" | "NA" |
| "hdmi" | "NA" |
| "maximum shutter speed" | "1/2000 sec" |
| "metering" | "NA" |
| "minimum shutter speed" | "1 sec" |
| "optical zoom" | "18x" |
| "screen size" | "3 Inches" |
| "usb" | "Yes", |
| "video display resolution" | "NA" |
| "wifi" | "Yes; Wi-Fi 802.11 b/g/n" |



D. Obraczka, A. Saeedi, A. E. Rahm, E.: *Knowledge Graph Completion with FAMER.* Proc. KDD workshop on Data Integration to Knowledge Graphs (DI2KG), 2019

# ENTITY RESOLUTION WORKFLOW





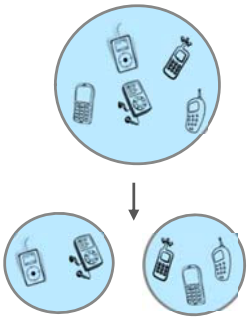

- input: 1, 2 or n data sources
- n>=2: duplicate-free (clean) sources or not
  - cluster sizes <= n für clean sources
  - at most one match per source





# BLOCKING





- naïve: pairwise matching of all entities
  - quadratic complexity, not scalable
  - strong need to reduce match search space
- **blocking**
  - group similar objects within blocks / partitions
  - only compare entities of the same block
  - many variations: Standard Blocking, LSH, Sorted Neighborhood, …
  - dirty data may require use of multiple blocking keys
- block size critical for runtime and quality

# MATCHING

- combined use of several similarity values
  - attribute similarities, e.g. using numeric or string similarity measures (e.g., edit distance, n-gram)
  - context-based matchers
- use of match rules
  - e.g. pubs match if *title sim. ≥ 0.9 & author sim. > 0.4*
  - special case: similarity joins ( sim (e1, e2) ≥ *t* )
  - strong dependency on threshold and not well scalable
- learned/supervised match classification models
  - requires suitable training data
  - e.g. decision tree, logistic regression, SVM, deep neural networks
  - use of representation learning (embeddings) to enhance matchability of attribute values

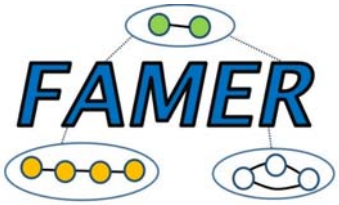


# ENTITY CLUSTERING

- grouping of matching entities
  - cluster/group can be used to fuse information from different variants and create golden record/cluster representative, e.g. for knowledge graph
  - for duplicate-free sources at most one entity per cluster similarity
- clustering approaches
  - input: match candidates with their similarity (similarity graph)
  - try to maximize similariy within cluster and minimize similarity between clusters
- baseline: connected component /trans. closure
  - good recall but can lead to huge clusters with low precision

# FAMER TOOL

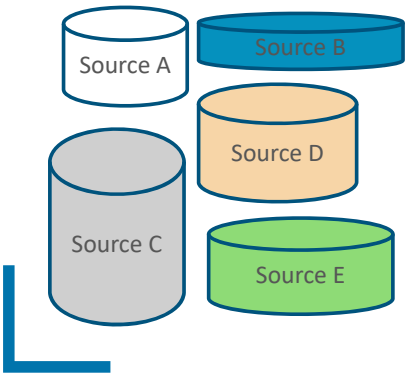




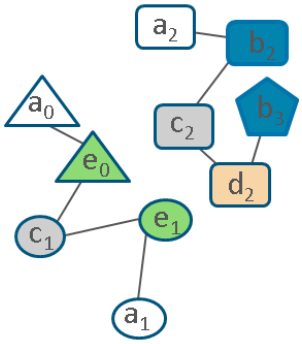

**FA**st **M**ulti-source **E**ntity **R**esolution System

- scalable linking & clustering for many sources

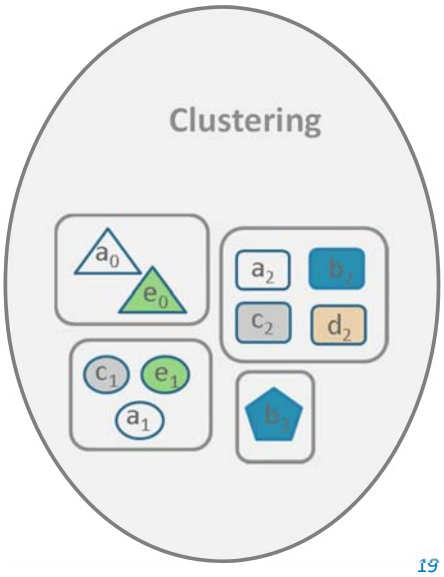
**Input**




**Linking:** Similarity Graph




**Clustering**

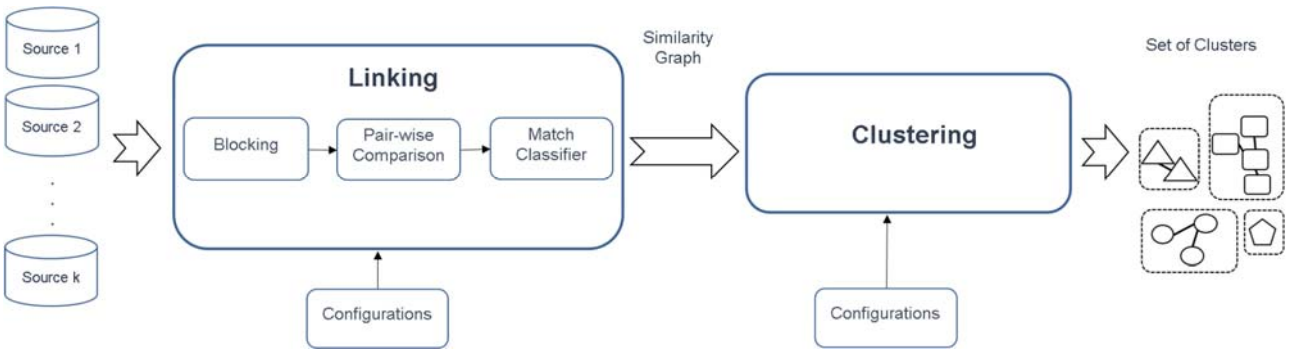




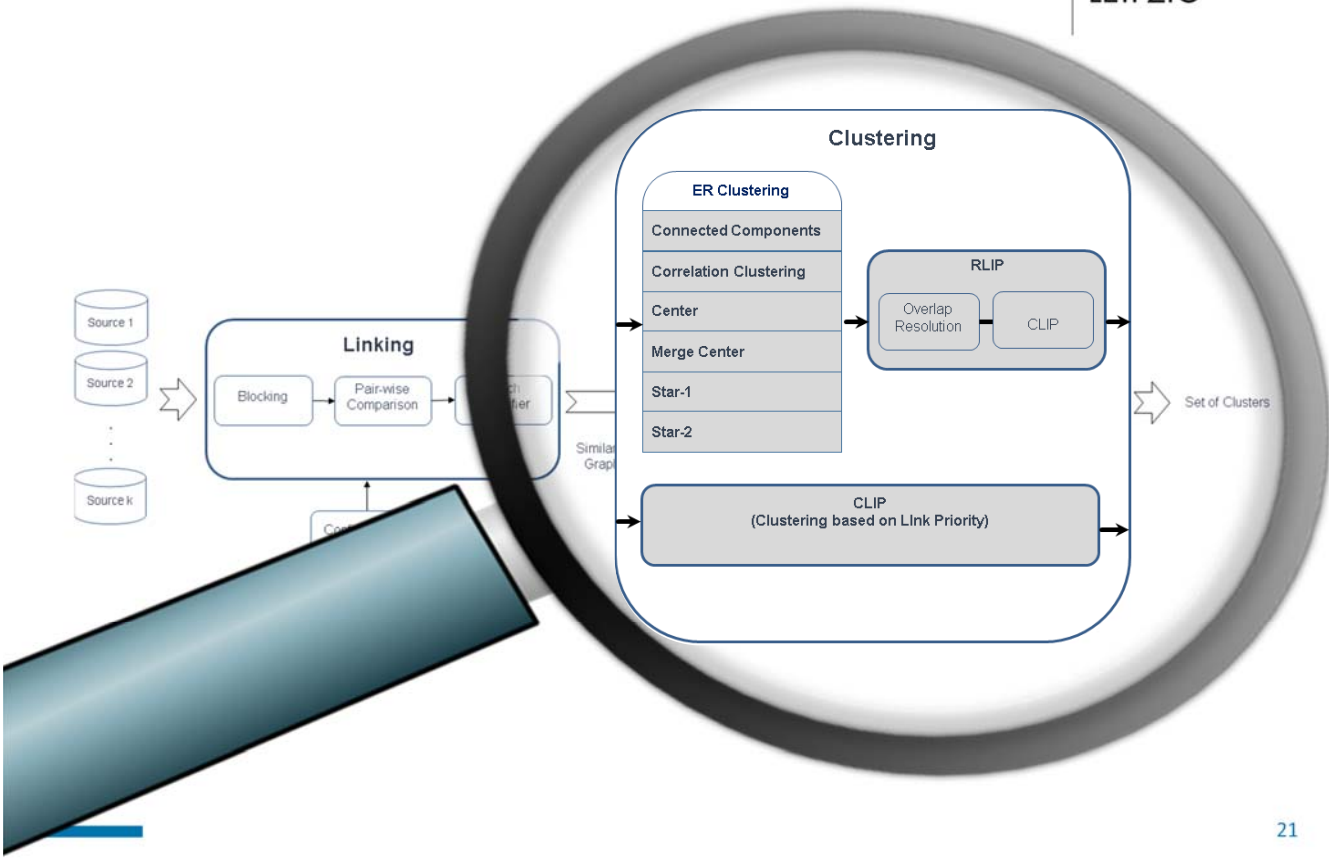



# FAMER OVERVIEW



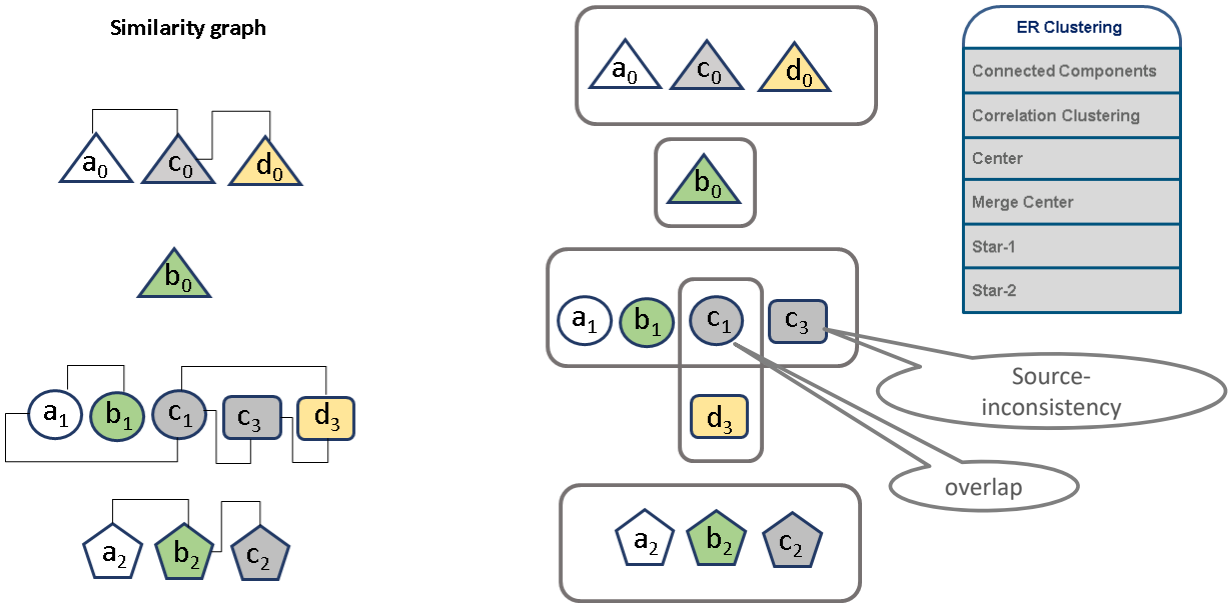


# FAMER OVERVIEW






# FAMER CLUSTERING

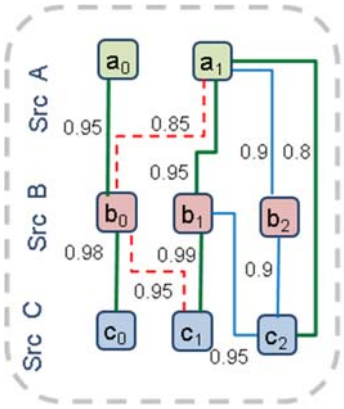


# LINK STRENGTH (USED IN CLIP)



– Link Strength

- **Strong**
- **Normal**
- **Weak**



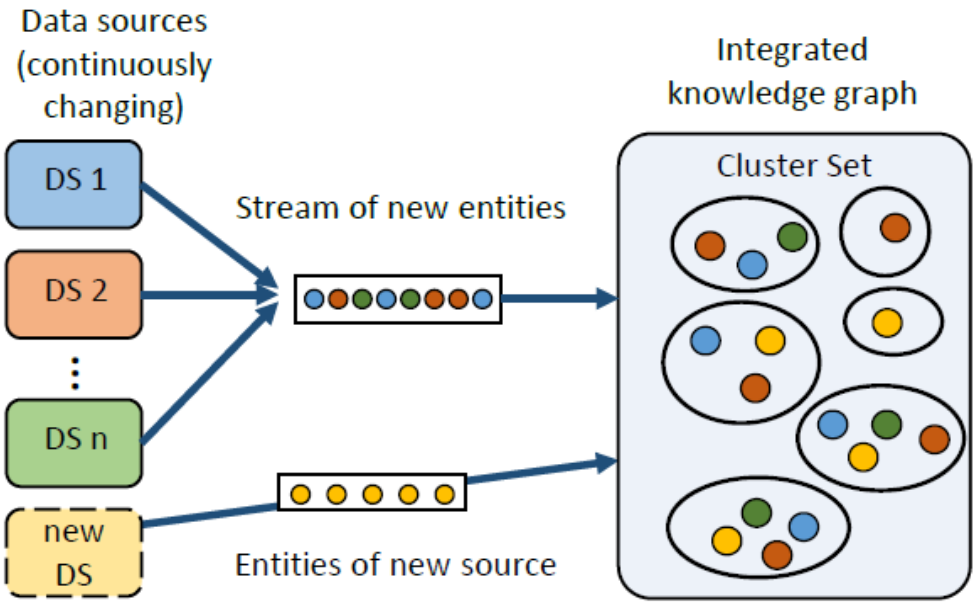

- **CLIP clustering**
  - weak links are ignored
  - strong links are preferred
  - normal links with high similarity are used as long as they do not lead to violation of source consistency



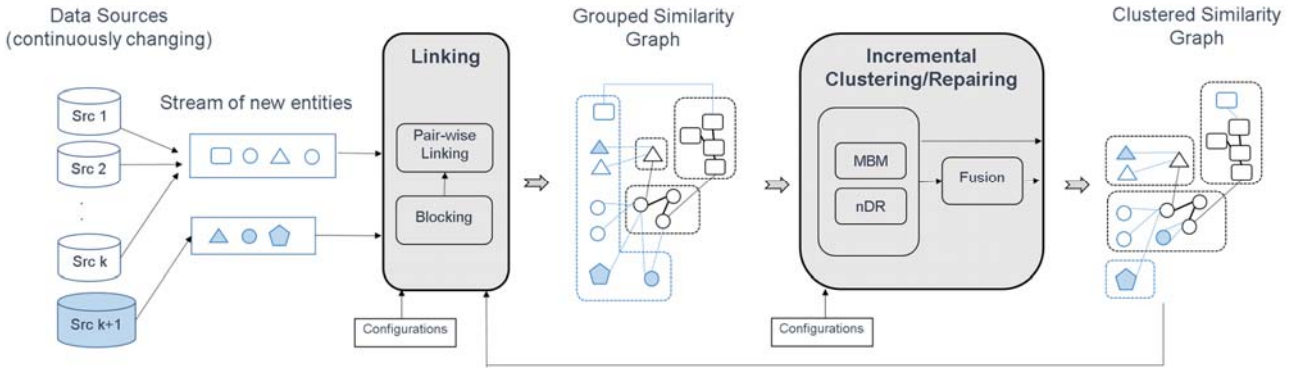


# INCREMENTAL ENTITY CLUSTERING





# FAMER INCREMENTAL ER*

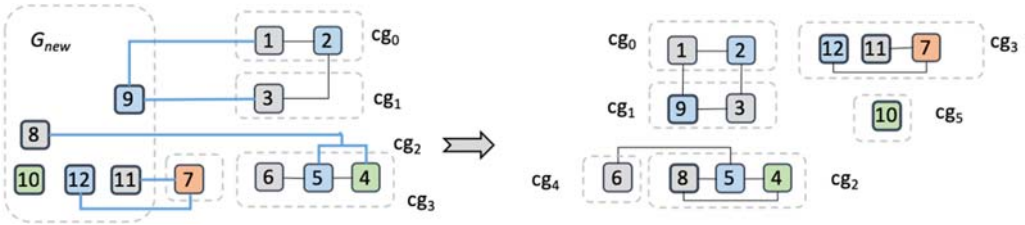



- based on existing clusters and similarity graph
- matching of new entities with existing similarity graph
  - optional matching among new entities
  - Max-both assignment between new entities and existing clusters
- reclustering possible to repair wrong clusters

*A. Saeedi, E. Peukert, E. Rahm: Incremental Multi-source Entity Resolution for Knowledge Graph Completion. Proc. ESWC 2020



# N-DEPTH RECLUSTERING (FAMER)




1-depth reclustering

- n=1 sufficient
- highly effective in evaluations
  - same quality after several source additions as for batch ER with all sources
  - different insert orders lead to same/similar result

# PLANNED RESEARCH ITEMS

- more general graph-based ER
  - multiple entity types
  - matching of relationships
  - Context/neigbhborhood-based match approaches
- learning-based methods
  - decision about blocking / matching / clustering strategies
  - determine order in which different entity types are considered
- strategies to deal with dirty sources with duplicates
- cope with all steps for knowledge graph completion
  - prediction of entity type for new entities
  - schema/property matching
  - entity resolution
  - fusion of entities / relations



# AGENDA

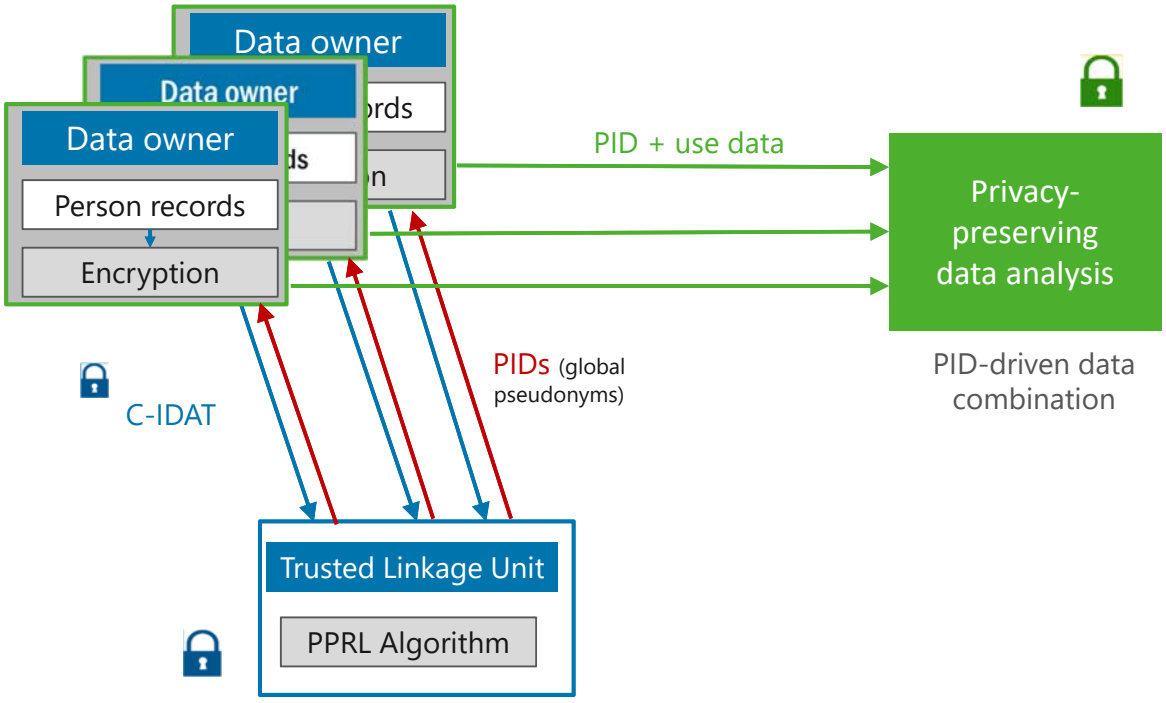
- Introduction
- Holistic entity resolution for knowledge graph completion
- Privacy-related research
  - Introduction
  - PPRL with Bloom filters / PRIMAT
  - PPML
- Graph data analysis

# PRIVACY FOR DATA SCIENCE

- need for comprehensive privacy support (“privacy by design”)
- privacy-preserving publishing of datasets
  - anonymization of datasets
- privacy-preserving record linkage (PPRL)
  - combine data about persons (e.g., patients) without revealing privacy
- privacy-preserving data mining / machine learning
  - analysis of anonymized data without re-identification



# PPRL + DATA ANALYSIS

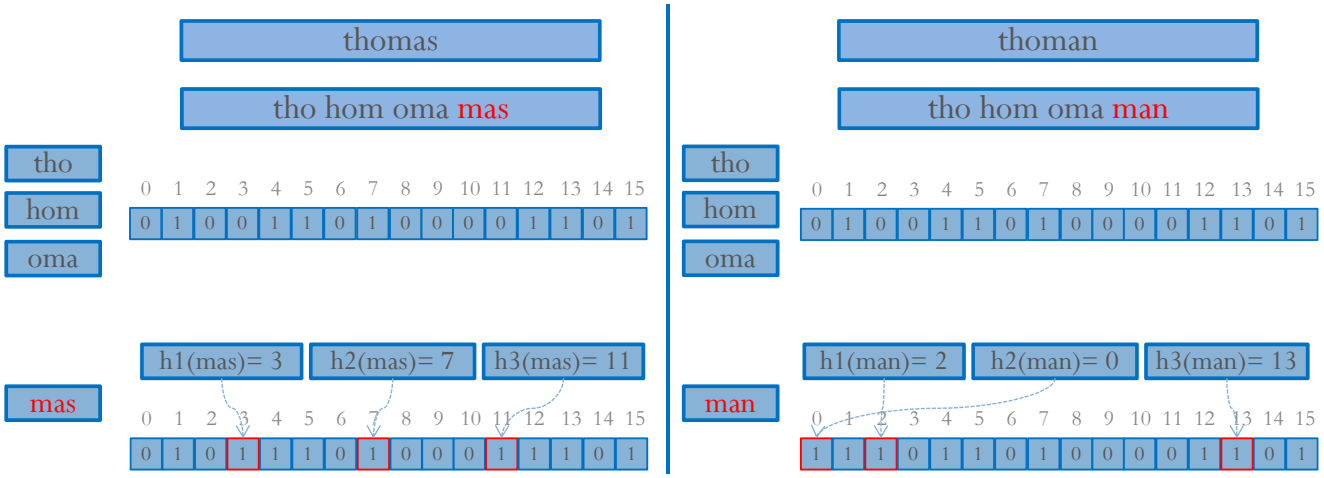


# PPRL WITH BLOOM FILTERS

- effective and simple encoding uses bloom filters (Schnell et al., 2009)
- tokenize match-relevant quasi-identifiers, e.g. using bigrams or trigrams
  - typical attributes: first name, last name (at birth), sex, date of birth, country of birth, place of birth
- map each token with a family of one-way hash functions to fixed-size bit vector (fingerprint)
  - original data cannot be reconstructed
- match of bit vectors (e.g., using Jaccard similarity) is good approximation of true match result



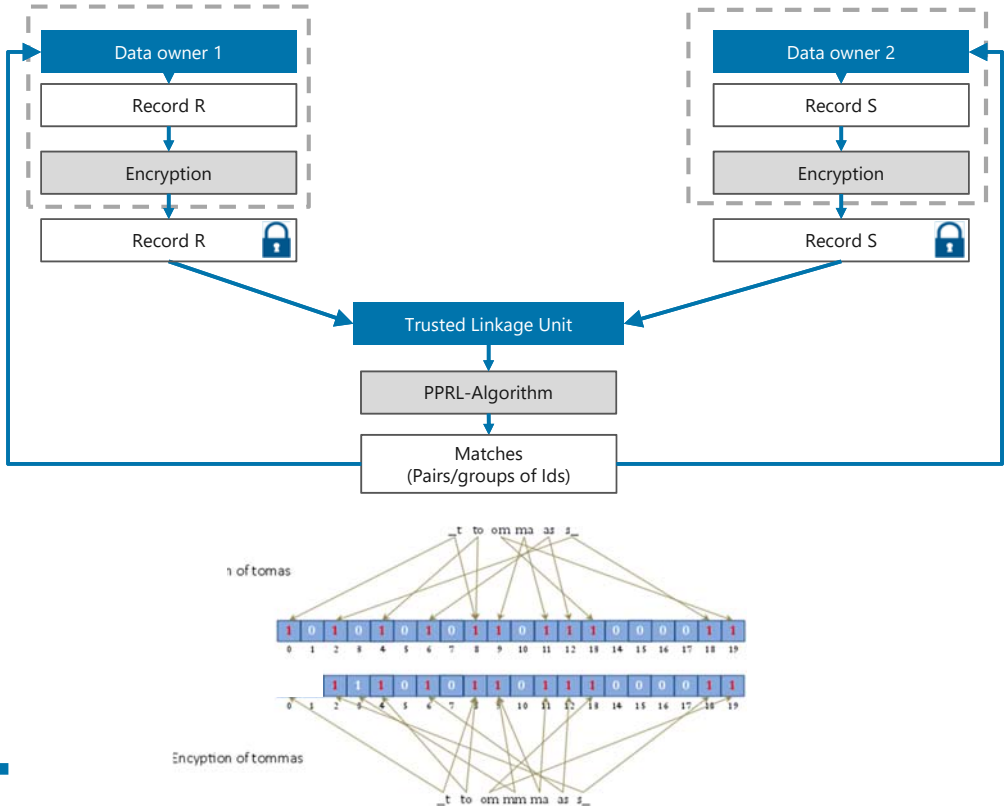
# SIMILARITY COMPUTATION – EXAMPLE




$$Sim_{Jaccard}(r1, r2) = (r1 \wedge r2) / (r1 \vee r2)$$

$$Sim_{Jaccard}(r1, r2) = 7/11$$

# PPRL PROTOCOL





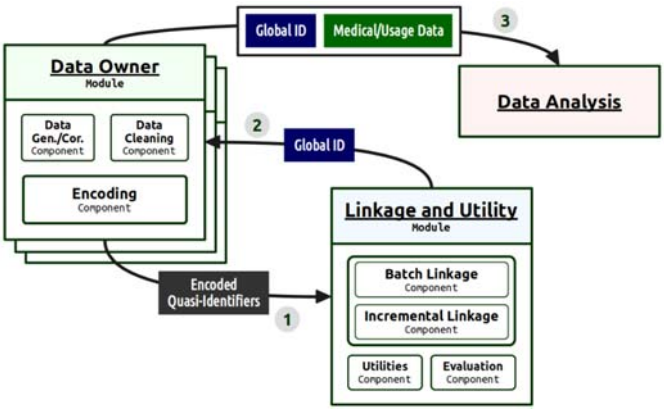

# SCALABLE PPRL

- filtering for specific similarity metrics / thresholds to reduce number of comparisons
  - privacy-preserving PPJoin (P4Join)
  - metric space: utilize triangular inequality
- (private) blocking approaches
  - partition datasets such that only records from same partition (block) need to be matched with each other
  - blocking at data owner on unencoded data (e.g., soundex) or at LU on bloom filters (e.g., LSH)
- parallel linkage
  - GPU-based matching of bit vectors
  - parallel matching on clusters, e.g. using Apache Spark/Flink

Franke, M et al.: *ScaDS Research on Scalable Privacy-preserving Record Linkage.* Datenbank-Spektrum 2019

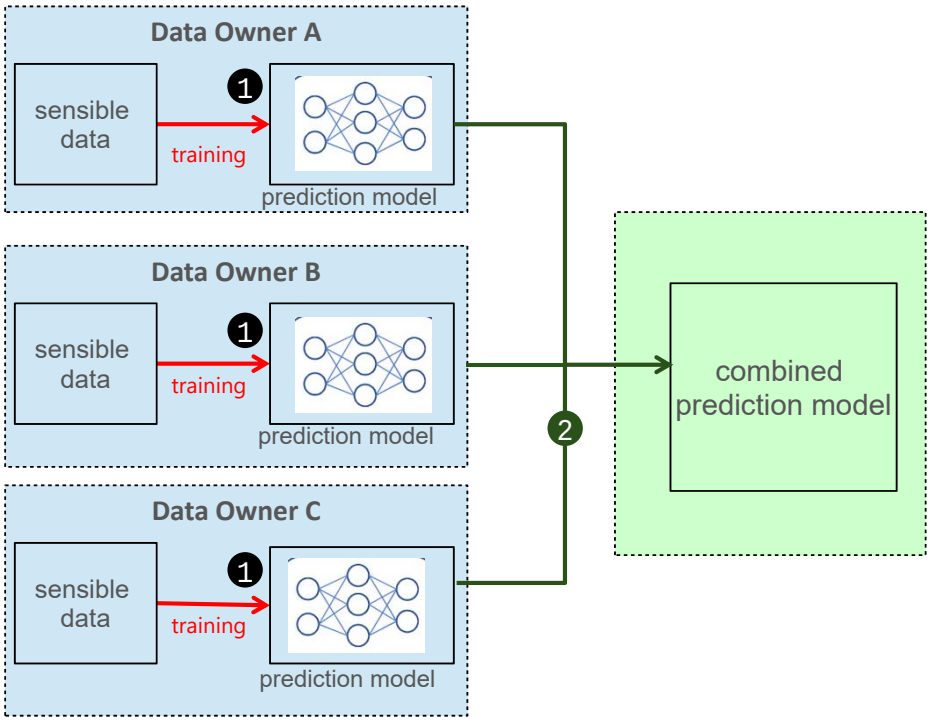
# PRIMAT TOOLBOX

- **Pri**vate **Ma**tching **T**oolbox (Uni Leipzig)
- open-source PPRL Tool for entire PPRL process
- separate modules for data owner and linkage unit
- flexible configuration and execution of PPRL workflows
- high performance by blocking and parallel execution
- comparative evaluation of different PPRL approaches/configurations on test datasets




Franke, M.; Sehili, Z.; Rahm, E.: *PRIMAT: A Toolbox for Fast Privacy-preserving Matching*. PVLDB 2019

# DISTRIBUTED PPML WITHOUT PPRL



# PLANNED RESEARCH ITEMS

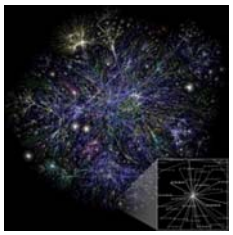
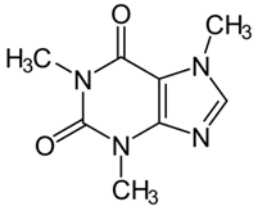
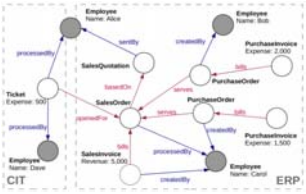
- comparison of PPML with and without PPRL
  - suitable applications, e.g. in medicine
  - different degrees of overlap between data sources
- PPRL on training data only ?
- investigation of advanced data mining / ML on anonymized data, e.g., graph based pattern mining
  - use case: money laundring with hidden money transfer bewtween many persons/participants



# AGENDA

- Introduction
- Holistic entity resolution for knowledge graph completion
- Privacy-related research
- Graph data analysis
  - requirements
  - graph analysis with GRADOOP
  - support for temporal graphs

# „GRAPHS ARE EVERYWHERE“

| Social science | Engineering | Life science | Information science |
|---|---|---|---|
|  |  |  |  |
|  |  |  |  |

Facebook
ca. 1.3 billion users
ca. 340 friends per user
Twitter
ca. 300 million users
ca. 500 million tweets per day

Internet
ca. 2.9 billion users

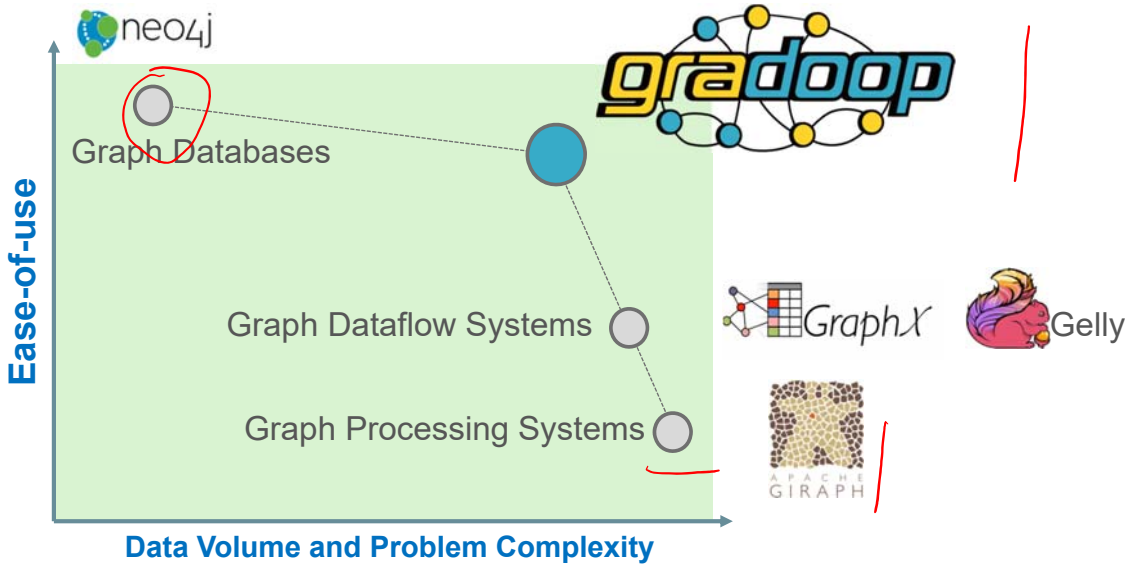
Gene (human)
20,000-25,000
ca. 4 million individuals
Patients
> 18 millions (Germany)
Illnesses
> 30.000

World Wide Web
ca. 1 billion Websites
LOD-Cloud
ca. 90 billion triples



# GRAPH DATA ANALYTICS: REQUIREMENTS

- powerful but easy to use graph data model
- interactive, declarative graph queries
- scalable graph mining and machine learning
- high performance and scalability
- graph-based integration of many data sources
- versioning and evolution (dynamic /temporal graphs)
- comprehensive visualization support





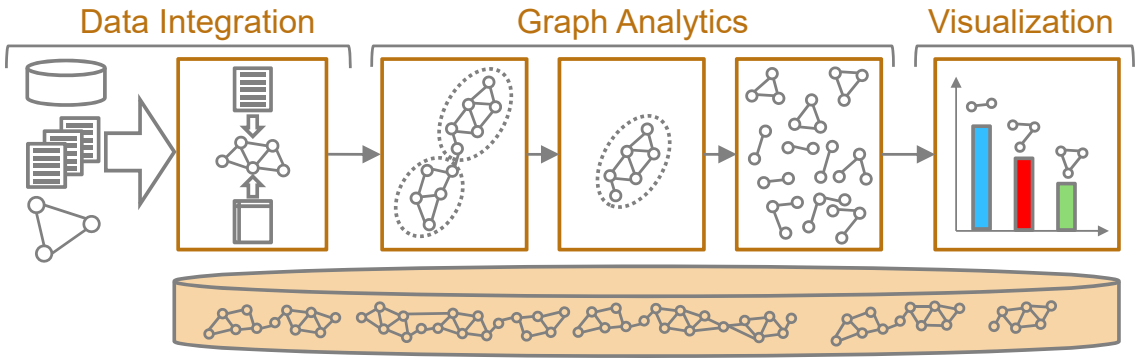



# GRADOOP CHARACTERISTICS

- open-source Hadoop-based framework for graph data management and analysis, www.gradoop.org
  - utilization of powerful dataflow system (Apache Flink) for parallel, in-memory processing
- Extended property graph data model (EPGM)
  - operators on graphs and sets of (sub) graphs
  - support for semantic graph queries and mining
- declarative specification of graph analysis workflows
  - Graph Analytical Language - GrALa
- end-to-end functionality
  - graph-based data integration, data analysis and visualization

# END-TO-END GRAPH ANALYTICS





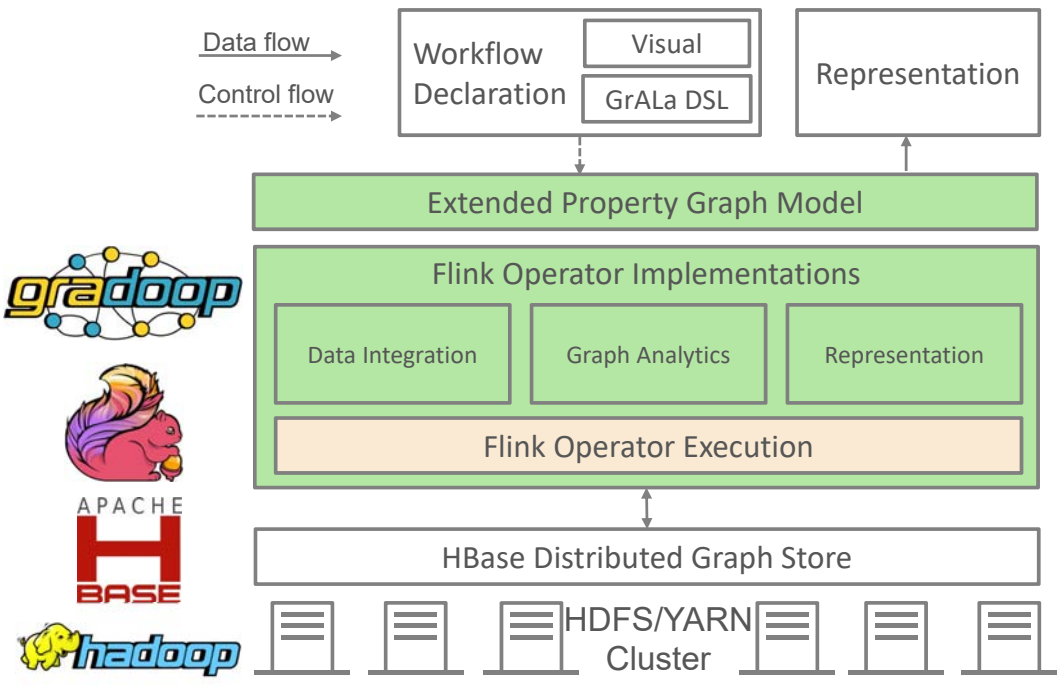

- **integrate data** from one or more sources into a dedicated **graph store** with **common graph data model**
- definition of **analytical workflows** from **operator algebra**
- result representation in **meaningful way**





# HIGH LEVEL ARCHITECTURE





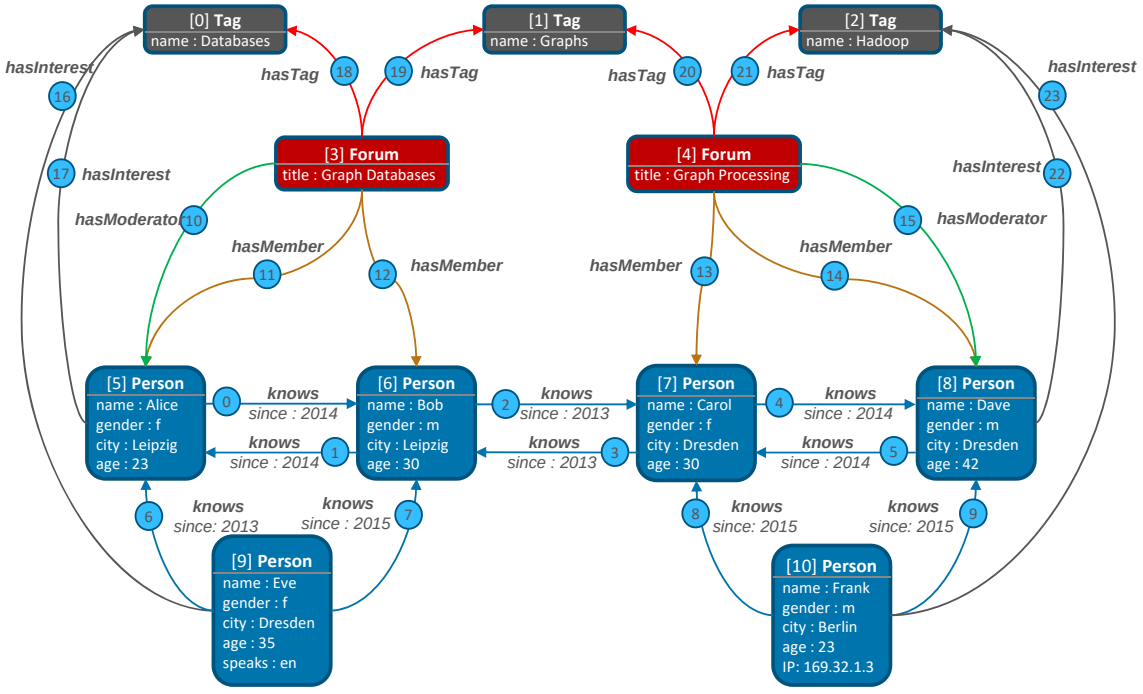
| | Operators | | Algorithms |
|---|---|---|---|
| | Unary | Binary | |
| Logical Graph | Aggregation | Combination | Gelly Library |
| | Pattern Matching | Overlap | BTG Extraction |
| | Transformation | Exclusion | Adaptive Partitioning |
| | Grouping | Equality | |
| | Subgraph | | |
| | Call * | | |
| Graph Collection | Selection | Union | Frequent Subgraphs |
| | Distinct | Intersection | |
| | Sort | Difference | |
| | Limit | Equality | |
| | Apply * | | |
| | Reduce * | | |
| | Call * | | |

* auxiliary



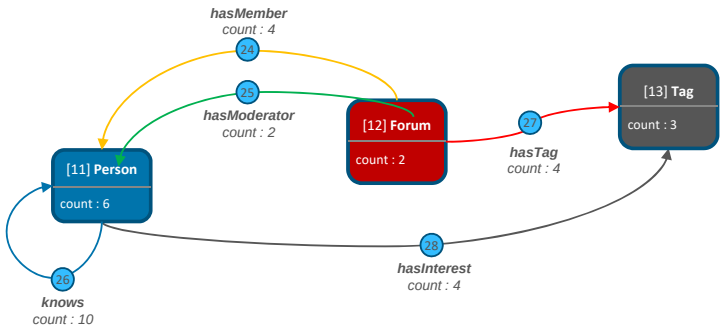


# SAMPLE GRAPH



# GROUPING: TYPE LEVEL *(SCHEMA GRAPH)*

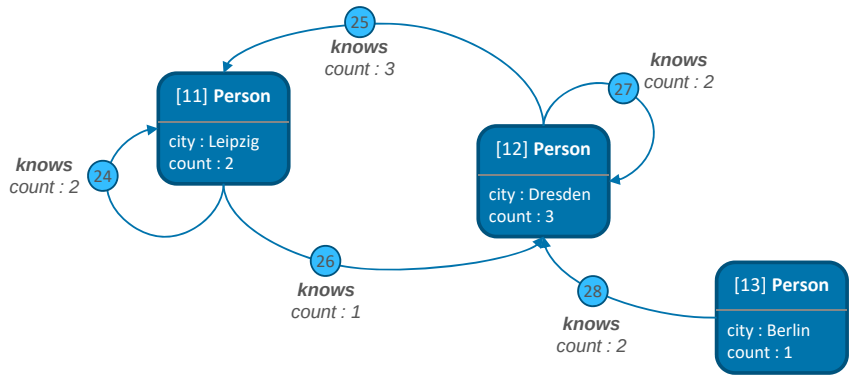
```
vertexGrKeys = [:label]
edgeGrKeys   = [:label]
sumGraph     = databaseGraph.groupBy(vertexGrKeys, [COUNT()], edgeGrKeys, [COUNT()])
```






# GROUPING: PROPERTY-SPECIFIC

```
personGraph  = databaseGraph.subgraph((vertex => vertex[:label] == 'Person'),
                                      (edge => edge[:label] == 'knows'))
vertexGrKeys = [:label, "city"]
edgeGrKeys   = [:label]
sumGraph     = personGraph.groupBy(vertexGrKeys, [COUNT()], edgeGrKeys, [COUNT()])
```



# DYNAMIC GRAPHS



- most graph databases and graph processing systems focus on static graphs
- real graphs like social networks, citation networks, road networks etc change over time
- analytical questions are often time-related
  - as-of queries on past states (snapshots)
  - change/evolution analysis …
  - need to efficiently update/refresh analysis results (graph metrics, communities/clusters, …)
- need of scalable approaches for managing and analyzing temporal graphs and graph data streams

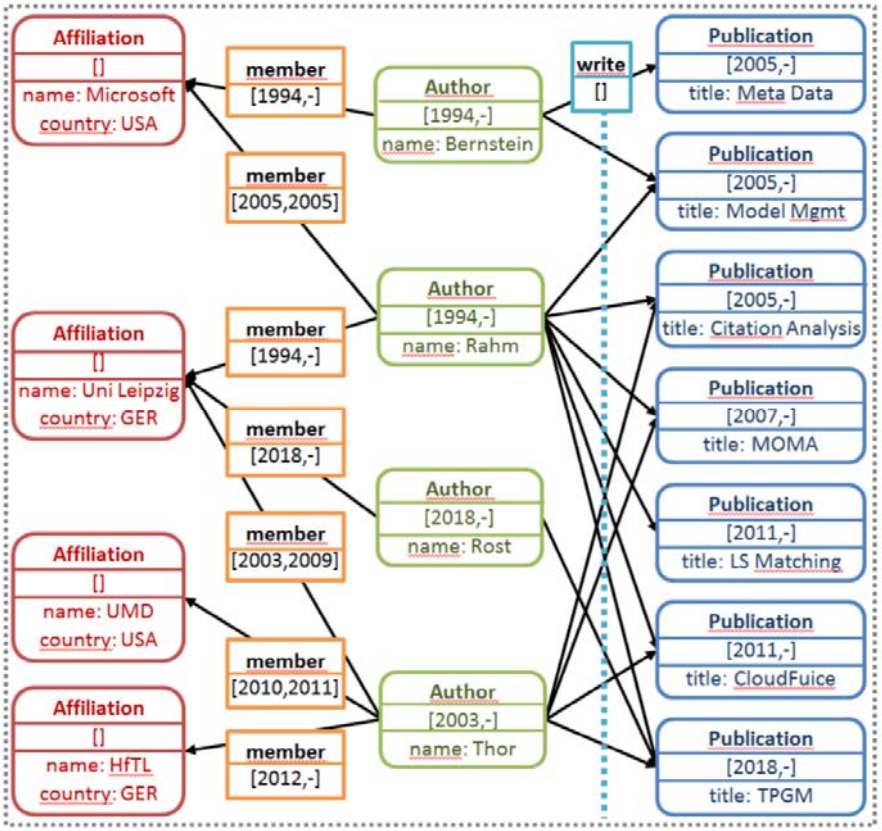




# TEMPORAL GRAPHS IN GRADOOP



- support for bitemporal graphs
  - time intervals for valid time and transaction time for vertices, edges and graphs
  - *valid time* provided by user, *tx time* is system-provided
- changes to existing operators
  - time predicates (as-of, between, overlap, precedes/suceeds …) for subgraph, pattern matching, grouping …
- new operators
  - snapshot extraction (as-of subgraph)
  - graph diff (between two snaphshots)





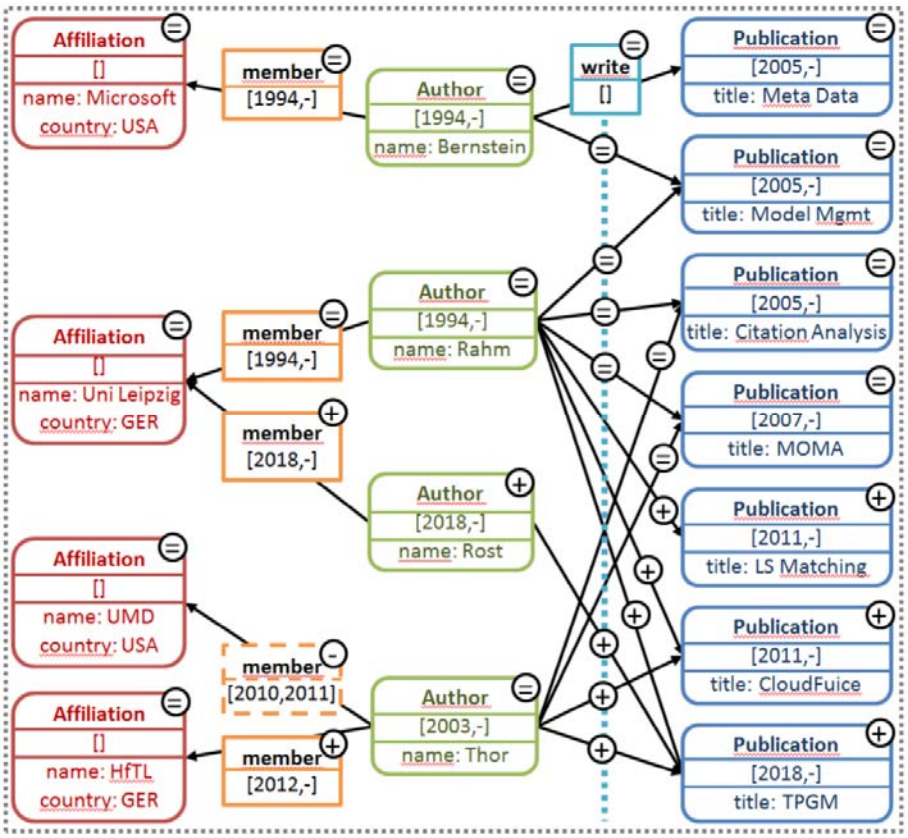



# EXAMPLE QUERIES

- affiliation memberships with a duration of less than 3 years
  - graph.subgraph (null, e -> e.label = 'member' AND YEAR(e.to)-YEAR(e.from) < 3
- authors who had an US affiliation in 2017 (temporal pattern matching)
  - (a:Author)-[m:member]->(f:Affiliation country : USA) WHERE m.asOf(2017)
- graph snapshot as of 2010
  - graph.snapshot (asOf(2010))
- graph difference
  - graph.diff (asOf(2010),asOf(2019))

# GRAPH DIFF BETWEEN 2010 AND NOW










# PLANNED RESEARCH ITEMS



- real-time graph analysis on data streams
  - consideration of limited time intervals
- graph-based machine learning
  - utilization of context knowledge for improved predictions
- graph embeddings
  - comparative evaluations of different graph embeddings
- protoytpe developement
  - application to and evaluation for diverse use cases

# SUMMARY

- Knowledge graph research
  - better automatic ways to integrate data and update knowledge graphs
  - incremental entity resolution (FAMER)
  - more complete and more learning-based approaches needed
- Privacy-preserving data analysis
  - usable approaches for PPRL have been developed (PRIMAT)
  - will explore use cases requiring both PPRL and PPML
  - advanced analysis approaches needed, e.g., privacy-preserving graph pattern mining
- ML on graph data
  - powerful framework for analysis of static and temporal graphs (GRADOOP)
  - need to support machine learning on graphs and streams of graph data



# REFERENCES ON DATA INTEGRATION


- P. Christen: *Data Matching*. Springer 2012
- H. Köpcke, A. Thor, S. Thomas, E. Rahm: *Tailoring entity resolution for matching product offers*. Proc. EDBT 2012: 545-550
- L. Kolb, E. Rahm: *Parallel Entity Resolution with Dedoop*. Datenbank-Spektrum 13(1): 23-32 (2013)
- M. Nentwig, A. Groß, E. Rahm: *Holistic Entity Clustering for Linked Data*. IEEE Int. Conf. on Data Mining Workshop, ICDMW 2016 2016
- D. Obraczka, A. Saeedi, A. E. Rahm, E.: *Knowledge Graph Completion with FAMER.* Proc. KDD workshop on Data Integration to Knowledge Graphs (DI2KG), 2019
- E. Rahm: *The case for holistic data integration.* Proc. ADBIS, 2016
- A. Saeedi, M. Nentwig, E. Peukert, E. Rahm: *Scalable matching and clustering of entities with FAMER*. Complex Systems Informatics and Modeling Quarterly 2018
- A. Saeedi, E. Peukert, E. Rahm: *Comparative Evaluation of Distributed Clustering Schemes for Multi-source Entity Resolution*. Proc. ADBIS, LNCS 10509, 2017
- A. Saeedi, E. Peukert, E. Rahm: *Using Link Features for Entity Clustering in Knowledge Graphs.* Proc. ESWC 2018 (**Best research paper award**)
- A. Saeedi, E. Peukert, E. Rahm: *Incremental Multi-source Entity Resolution for Knowledge Graph Completion.* Proc. ESWC 2020

# REFERENCES ON PPRL


- M. Franke, Z. Sehili, E. Rahm: *PRIMAT: A Toolbox for Fast Privacy-preserving Matching*. PVLDB 2019
- M. Franke, M. Gladbach, Z. Sehili, F. Rohde, E. Rahm *ScaDS Research on Scalable Privacy-preserving Record Linkage.* Datenbank-Spektrum 2019
- M. Franke, Z. Sehili, E. Rahm: *Parallel Privacy-Preserving Record Linkage using LSH-based blocking*. Proc. IoTBDS 2018
- M. Franke, Z. Sehili, M. Gladbach, E. Rahm: Post-processing Methods for High Quality Privacy-Preserving Record Linkage, LNCS 11025, 2018
- D. Vatsalan, P. Christen, E. Rahm: *Incremental Clustering Techniques for Multi-Party Privacy-Preserving Record Linkage*. Data & Knowledge Engineering 2020
- D. Vatasalan, Z. Sehili, P. Christen, E. Rahm: *Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges*. In: Handbook of Big Data Technologies, Springer 2017




# REFERENCES ON GRAPH ANALYSIS


- M. Junghanns, M. Kießling, A. Averbuch, A. Petermann, E. Rahm: *Cypher-based Graph Pattern Matching in Gradoop*. Proc. ACM SIGMOD workshop on Graph Data Management Experiences and Systems (GRADES), 2017
- M. Junghanns, M. Kießling, N. Teichmann, K. Gomez, A. Petermann, E. Rahm: *Declarative and distributed graph analytics with GRADOOP*. PVLDB 2018
- M. Junghanns, A. Petermann, M. Neumann, E. Rahm: *Management and Analysis of Big Graph Data: Current Systems and Open Challenges*. In: Big Data Handbook (eds.: S. Sakr, A. Zomaya) , Springer, 2017
- M. Junghanns, A. Petermann, E. Rahm: *Distributed Grouping of Property Graphs with GRADOOP.* Proc. BTW, 2017
- M. Junghanns, A. Petermann, N. Teichmann, K. Gomez, E. Rahm: *Analyzing Extended Property Graphs with Apache Flink*. Proc. ACM SIGMOD workshop on Network Data Analytics (NDA), 2016
- M. Kricke, E. Peukert, E. Rahm: *Graph transformations in Gradoop*. Proc. BTW 2019
- A. Petermann, M. Junghanns, S. Kemper, K. Gomez, N. Teichmann, E. Rahm: *Graph Mining for Complex Data Analytics.* Proc. ICDM 2016
- A. Petermann, M. Junghanns, R. Müller, E. Rahm: *Graph-based Data Integration and Business Intelligence with BIIIG.* PVLDB 2014
- A. Petermann, M. Junghanns, E. Rahm: *DIMSpan - Transactional Frequent Subgraph Mining with Distributed In-Memory Dataflow Systems.* Proc. BDCAT 2017
- E. Rahm et al.: *Big Data Competence Center ScaDS Dresden/Leipzig: Overview and selected research activities*. Datenbank-Spektrum 2019
- C. Rost, A. Thor, E. Rahm: *Temporal graph analysis using Gradoop*. Proc. BTW workshops, 2019
- C. Rost, A. Thor, E. Rahm: *Analyzing temporal graphs with Gradoop*. Datenbankspektrum 2019
- C. Rost, A. Thor, P. Fritzsche, K. Gomez, E. Rahm, E.: *Evolution Analysis of Large Graphs with Gradoop*. ECML PKDD 2019 workshops, Springer 2020
- M.A. Rostami, M. Kricke, E. Peukert, S. Kühne, M. Wilke, S. Dienst, E. Rahm: *BIGGR: Bringing Gradoop to Applications.* Datenbank-Spektrum 2019