ScaDS.AI
DRESDEN LEIPZIG

UNIVERSITÄT
LEIPZIG

RECORD LINKAGE AND PPRL with Clustering of Matches

Erhard Rahm, Martin Franke

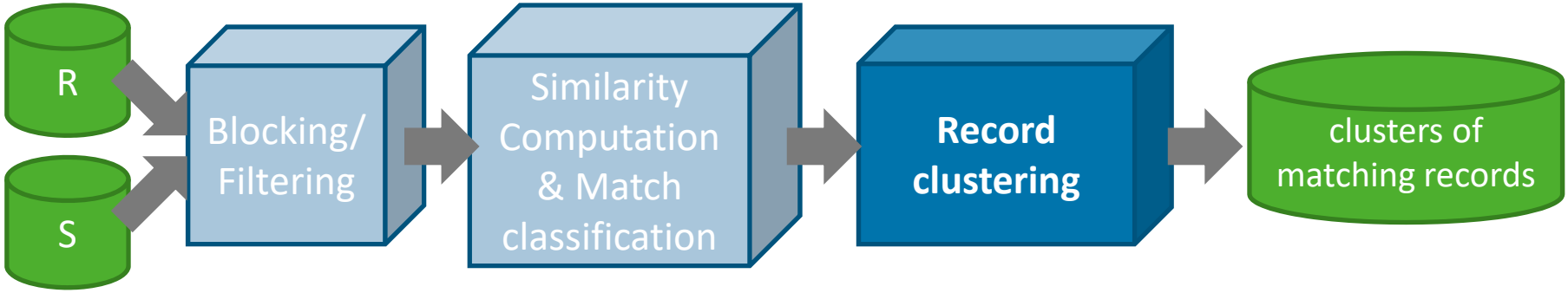
AGENDA

UNIVERSITÄT
LEIPZIG

- Record Linkage with Clustering
  - FAMER Tool
- Privacy-Preserving Record Linkage (PPRL)
  - PRIMAT Tool
- Summary

RECORD LINKAGE WORKFLOW

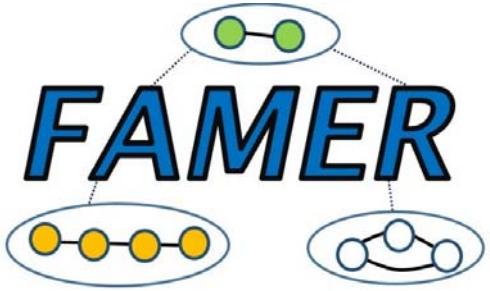
UNIVERSITÄT LEIPZIG



- input: 1, 2 or n data sources
- clustering can improve match quality
  - additional matches
  - removal of wrong matches
- compact representation of matches
  - cluster with k records corresponds to k(k-1)/2 match pairs
  - e.g. 10 elements instead of 45 match pairs

FAMER TOOL

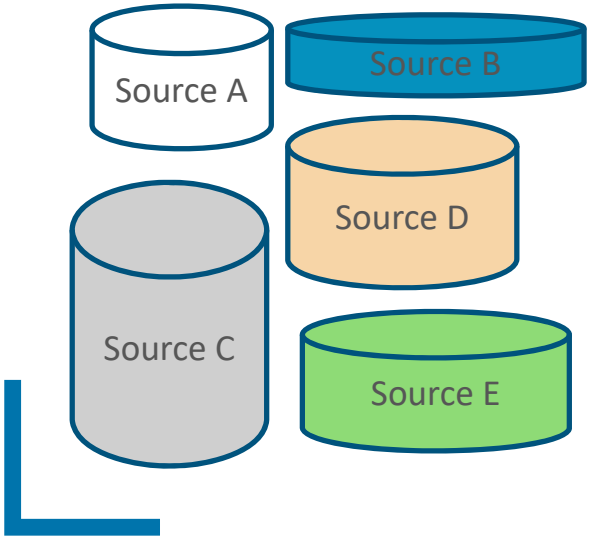
UNIVERSITÄT LEIPZIG



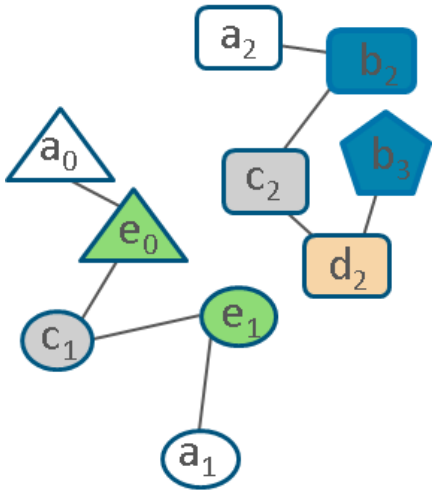
FAst **M**ulti-source **E**ntity **R**esolution System

- scalable linking & clustering for many sources

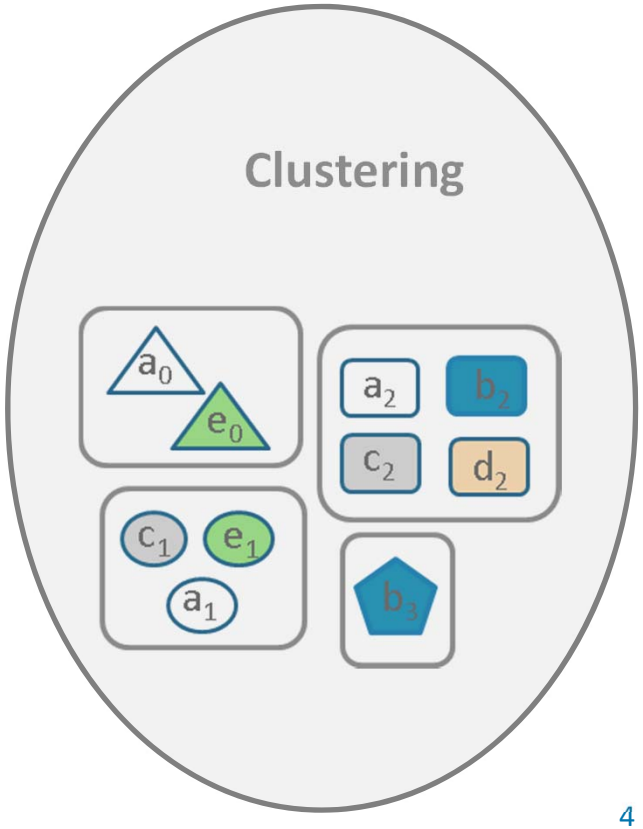
Input



Linking: Similarity Graph

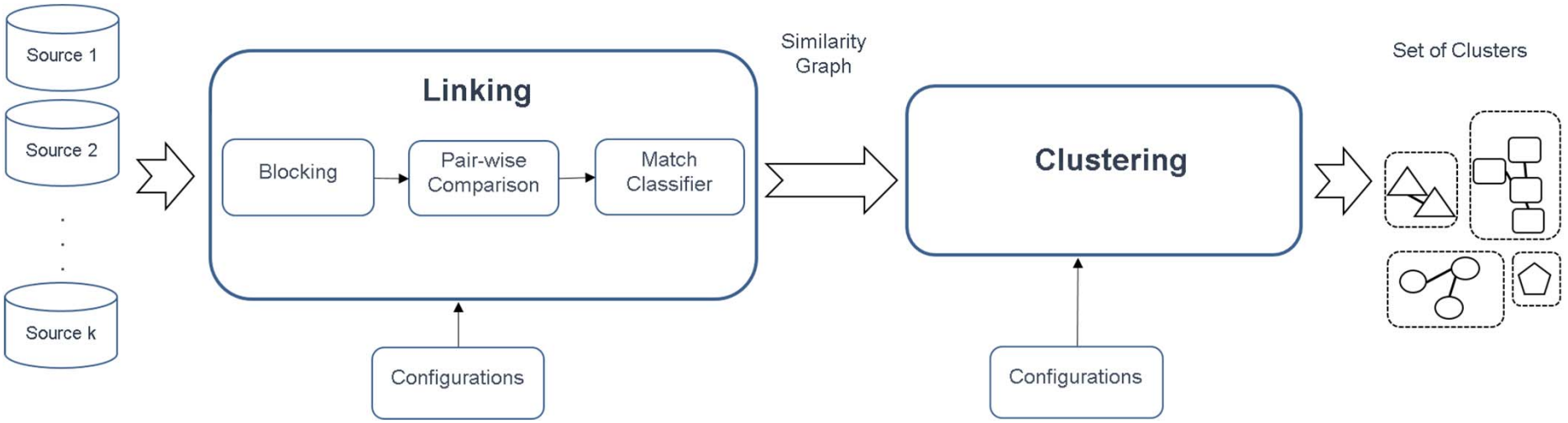


Clustering



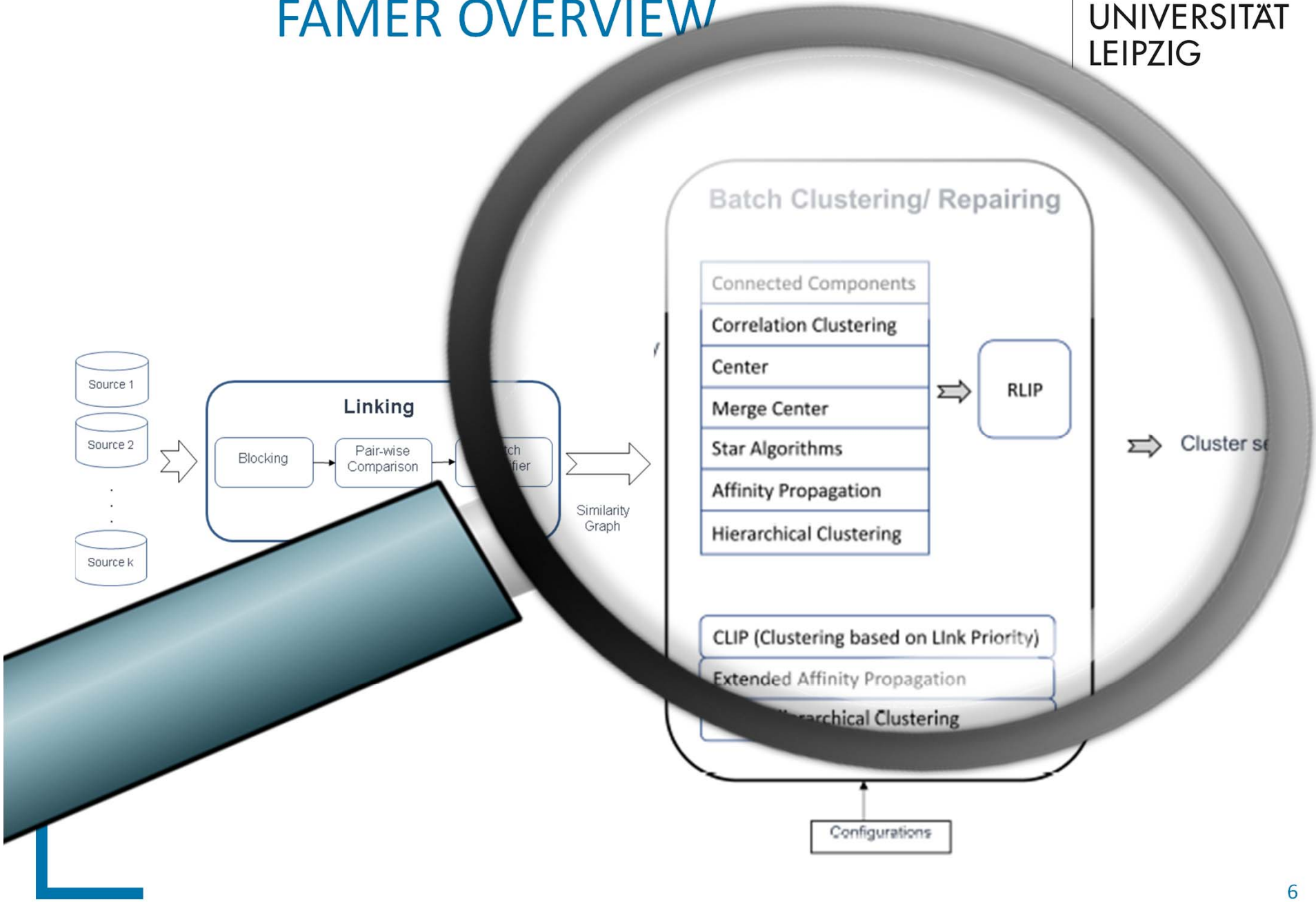
FAMER OVERVIEW

UNIVERSITÄT LEIPZIG



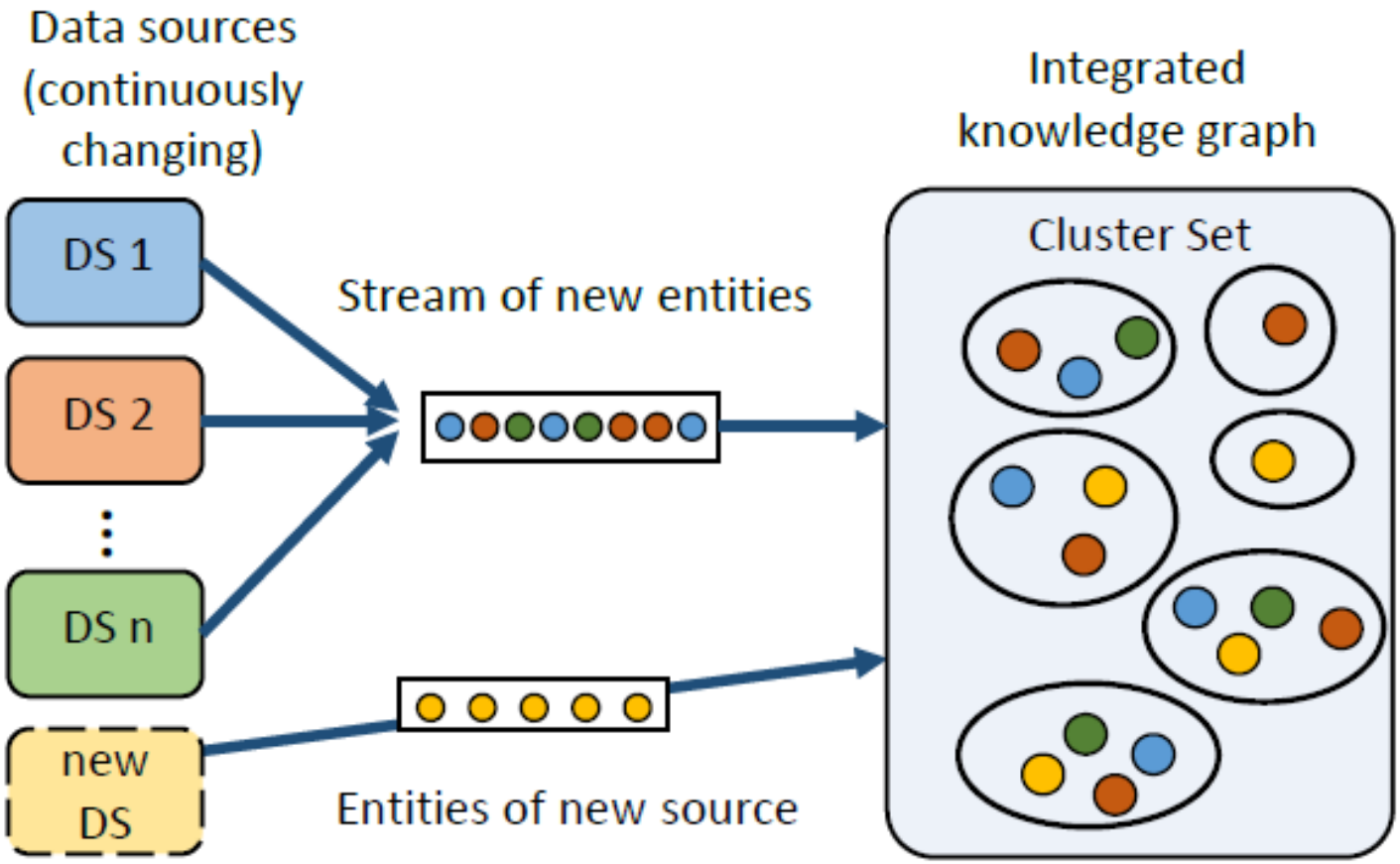
FAMER OVERVIEW

UNIVERSITÄT LEIPZIG



INCREMENTAL MATCHING & CLUSTERING

UNIVERSITÄT LEIPZIG



AGENDA

UNIVERSITÄT LEIPZIG

- Record Linkage mit Clustering
  - FAMER Tool
- Privacy-Preserving Record Linkage (PPRL)
  - PRIMAT Tool
- Zusammenfassung

PPRL REQUIREMENTS

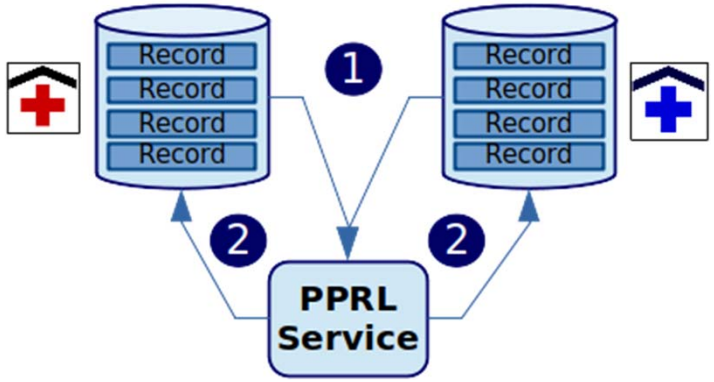
UNIVERSITÄT LEIPZIG

- high degree of privacy
  - no forwarding of unencoded quasi-identifiers to other institutions
  - minimal attack risk, e.g. no forwarding of person data to multiple sites or no central storage of (encoded) patient lists
- High performance and scalability
  - Many institutions, many persons/patients
- high match quality on encoded data
  - robustness against data errors/deviations, e.g. changed last name
  - support for match groups (clusters), not only match pairs
- dynamic PPRL: support for additional patients/ institutions
- flexible usability in different projects / scenarios
  - possible need to allow re-identification of patiens, e.g. for recruitment in new clinical study
- suitable tool support
  - flexible configurability, support for different usage forms
  - ease of use

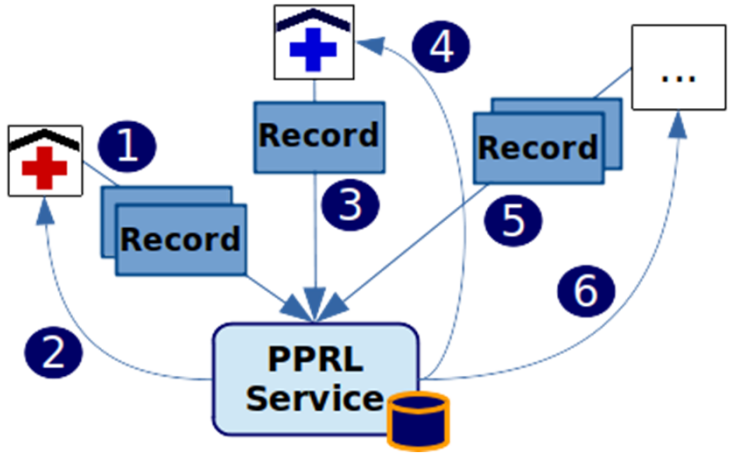
BATCH VS. INCREMENTAL MATCHING

UNIVERSITÄT LEIPZIG

- **Batch Matching**
  - linkage on fixed set of data
  - no storage of data and match resulzts at linkage unit
  - re-computation needed for changed data



- **Incremental Matching**
  - encoded person records and matches are stored in database
  - incremental matching for new records
  - faster compared to complete re-computation

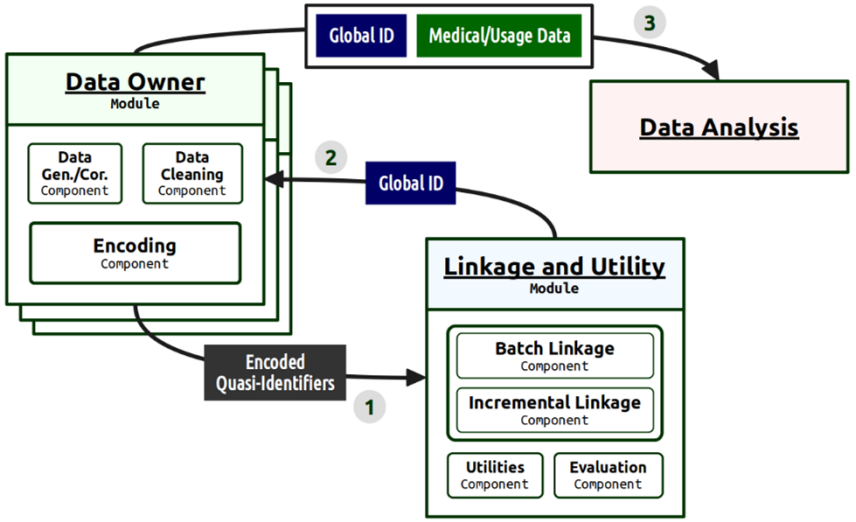


PRIMAT TOOLBOX

UNIVERSITÄT LEIPZIG

- **Pri**vate **Ma**tching **T**oolbox (Uni Leipzig)
- open-source PPRL-Tool for entire PPRL process
  - components for data owner and Linkage Unit
  - batch or incremental matching
- flexible configuration and execution of PPRL workflows
- high performance by blocking and parallel matching

PRIMAT
Private Matching Toolbox



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

COMPARISON OF PPRL TOOLS

UNIVERSITÄT LEIPZIG

	Mainzelliste (Lablans et. al.)	PRIMAT (U Leipzig)
Open source	✔	✔
data cleaning	(✔)	✔
Flexible encodings (Bloom Filter)	field-level Bloom Filter	✔
Hardening support	✗	✔
private blocking (with LSH)	✔	✔
Parallel matching	✗	(✔)
Match grouping / clustering	without transitivity	✔
Central or decentralized linkage	central	central
Batch + incremental Matching	Incremental	(✔)

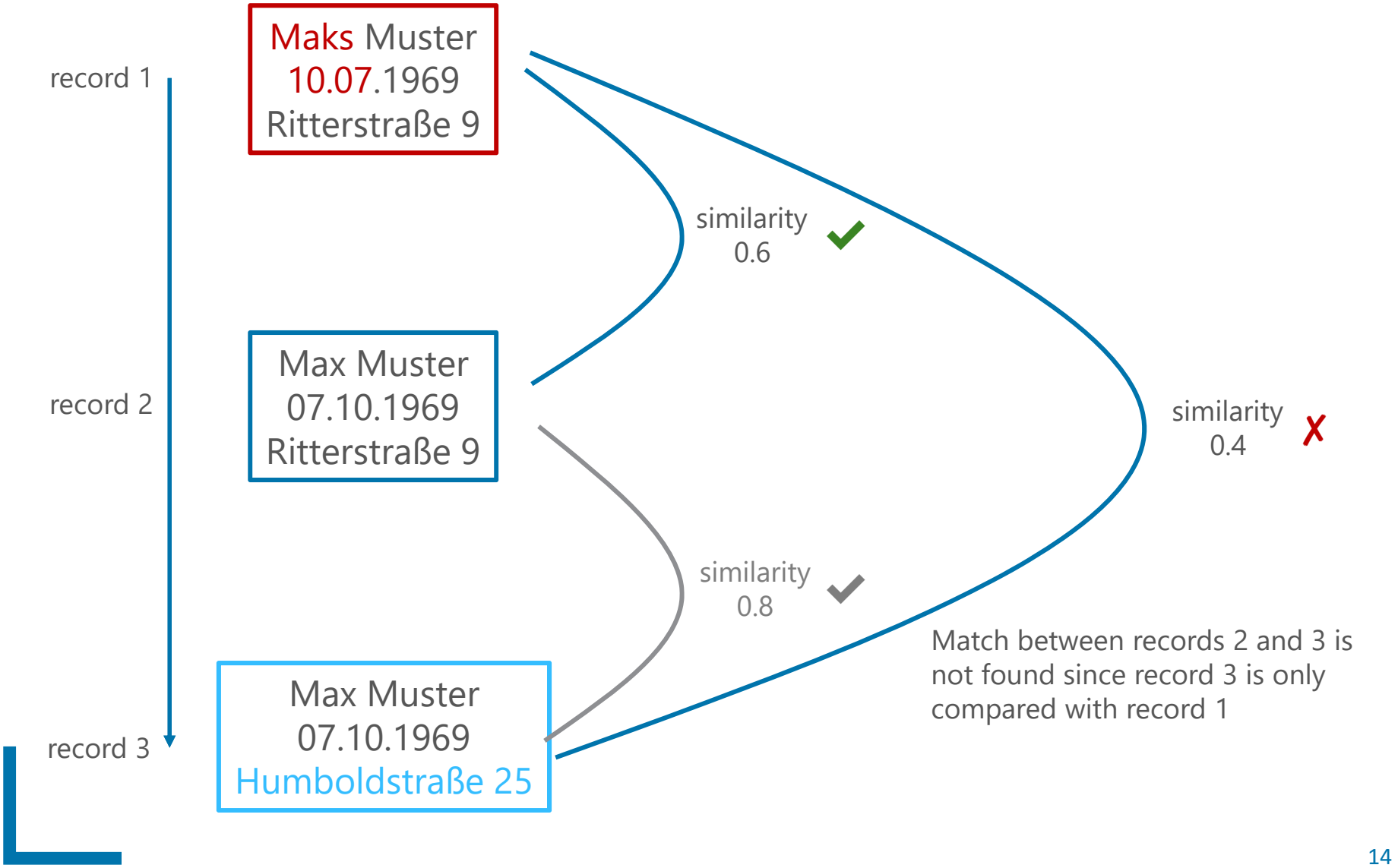
CLUSTERING IN MAINZELLISTE

UNIVERSITÄT LEIPZIG

- Mainzelliste tool used in many projects in Germany
- Cluster is represented by oldest (first) record
- new record is assigned to cluster with highest similarity (above threshold)
- inserts are processed one record at a time (slow, results can depend on order of inserts)
- **Problems:**
  - poor qualiy of cluster representant (first record) can reduce match quality
  - fixed cluster representant cannot deal with changes in address, name etc.
  - joint insertion of multiple records or entire sources can improve performance and quality

CLUSTERING MAINZELLISTE PROBLEM SCENARIO

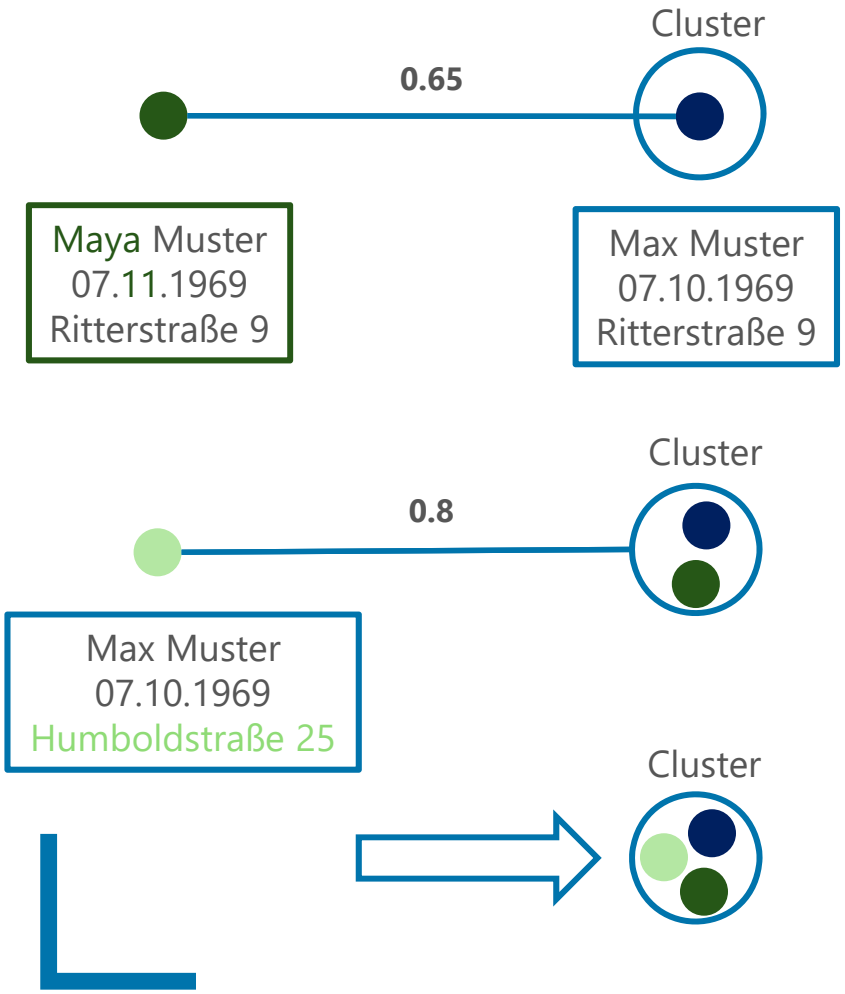
UNIVERSITÄT LEIPZIG



CLUSTERING MAINZELLISTE PROBLEM II

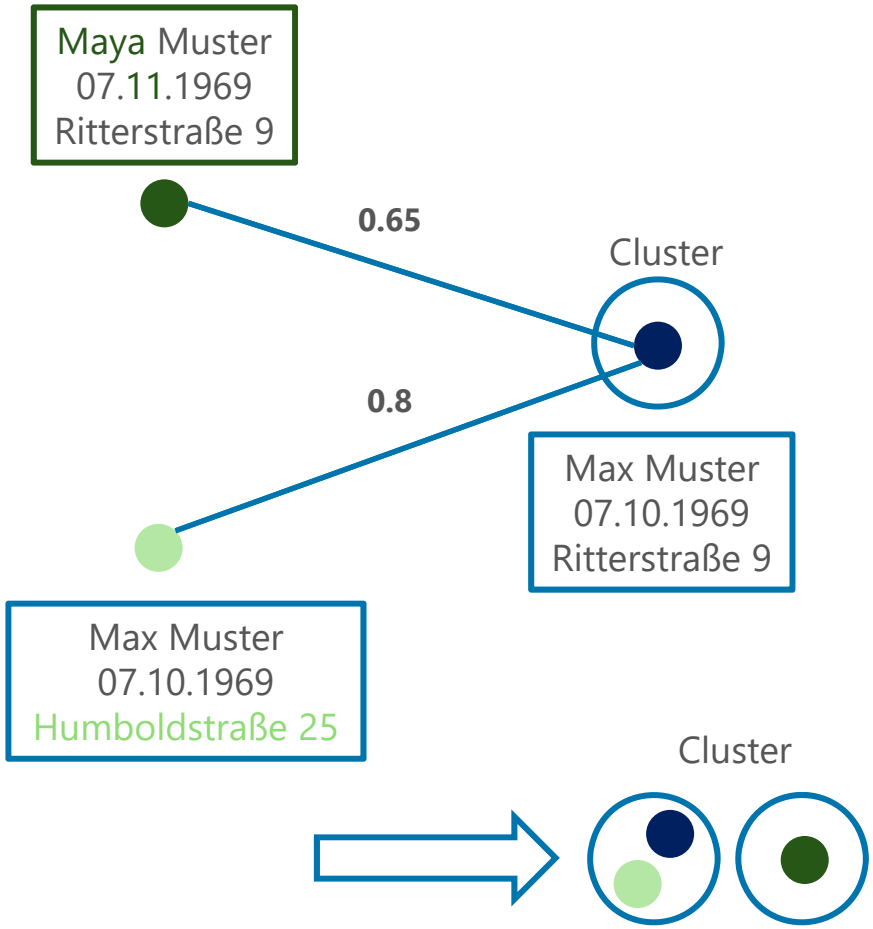
UNIVERSITÄT LEIPZIG

record-wise matching



joint consideration of multiple records

From duplicate-free source



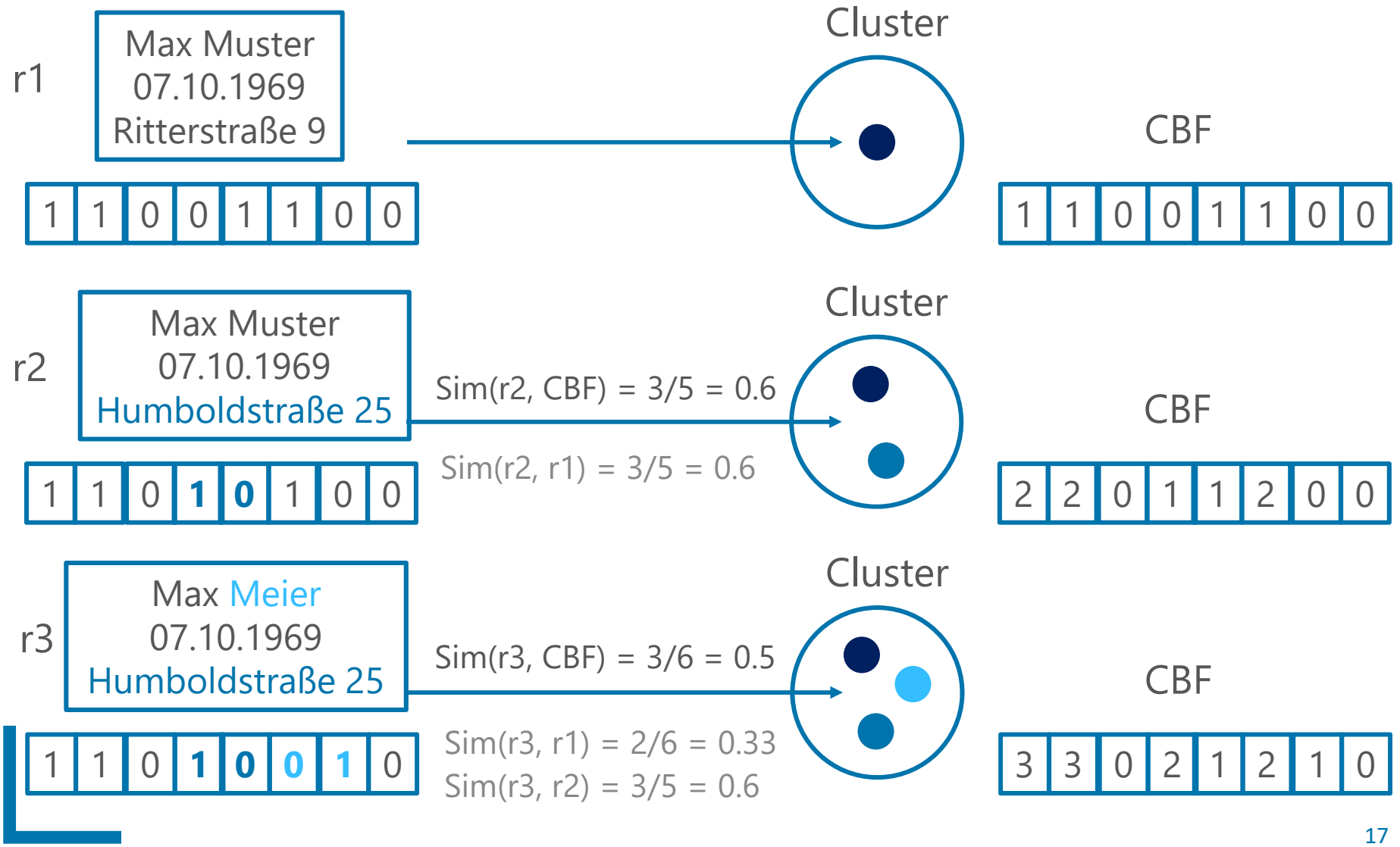
CLUSTERING EXTENSIONS

UNIVERSITÄT LEIPZIG

- Record for sources whether they are duplicate-free
  - optimized cluster assignment (per cluster at most one record of duplicate-free source)
- Comparison with all cluster members or only 1 representant
  - trade-off between quality and scalability
- Dynamic selection of cluster representant
  - youngest (newest) record
  - record with highest average similarity to all other cluster members
  - ‚virtual' cluster representant by aggregating cluster members, e.g. with Counting Bloom Filter

COUNTING-BLOOM-FILTER AS DYNAMIC CLUSTER REPRESENTANT

UNIVERSITÄT LEIPZIG



SUMMARY

UNIVERSITÄT LEIPZIG

- Record Linkage: pipeline with blocking, matching and clustering
- PPRL: many, partiallly project-specific requirements
- flexible configurability needed
  - batch and/or incremental matching
  - preprocessing, encoding, blocking, matching, clustering
- current PPRL tools provide no or insufficient support for clustering
  - need grows with more data sources and more possible duplicates
- PRIMAT: configurable toolbox with high flexibility and performance

REFERENCES RECORD LKINKAGE

UNIVERSITÄT LEIPZIG

- A. Saeedi, E. Peukert, E. Rahm: *Comparative Evaluation of Distributed Clustering Schemes for Multi-source Entity Resolution*. Proc. ADBIS 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, M. Nentwig, E. Peukert, E. Rahm: *Scalable Matching and Clustering of Entities with FAMER.* Complex Systems Informatics and Modeling Quarterly (CSIMQ), 2018
- A. Saeedi, E. Peukert, E. Rahm: *Incremental Multi-source Entity Resolution for Knowledge Graph Completion.* Proc. ESWC 2020
- S. Lerm, A. Saeedi, E. Rahm: *Extended Affinity Propagation Clustering for Multi-source Entity Resolution*. Proc. BTW 2021

REFERENCES PPRL

UNIVERSITÄT LEIPZIG

- M. Franke, Z. Sehili, E. Rahm: *Parallel Privacy Preserving Record Linkage using LSH-based blocking*. Proc. IoTBDS, 2018
- M. Franke, Z. Sehili, E. Rahm: *PRIMAT: A Toolbox for Fast Privacy-preserving Matching*. PVLDB 2019
- M. Franke et al.: *Post-processing Methods for High Quality Privacy-Preserving Record Linkage*. Proc, DPM, LNCS 2018
- M. Franke et al.: *ScaDS Research on Scalable Privacy-preserving Record Linkage.* Datenbank-Spektrum 2019
- M. Franke, Z. Sehili, F. Rohde, E. Rahm: *Evaluation of Hardening Techniques for Privacy-Preserving Record Linkage*. Proc. 24th EDBT Conf., 2021
- F. Rohde, M. Franke, Z. Sehili, M. Lablans, E. Rahm: *Optimization of the Mainzelliste software for fast privacy-preserving record linkage*. BMC Journal of Translational Medicine 2021
- Z. Sehili, F. Rohde, M. Franke, E. Rahm: *Multi-Party Privacy Preserving Record Linkage in Dynamic Metric Spac*e, Proc. BTW, 2021
- D. Vatasalan, P. Christen, E. Rahm: *Scalable privacy-preserving linking of multiple databases using Counting Bloom filters*. Proc Privacy and Discrimination in Data Mining (PDDM), 2016
- 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