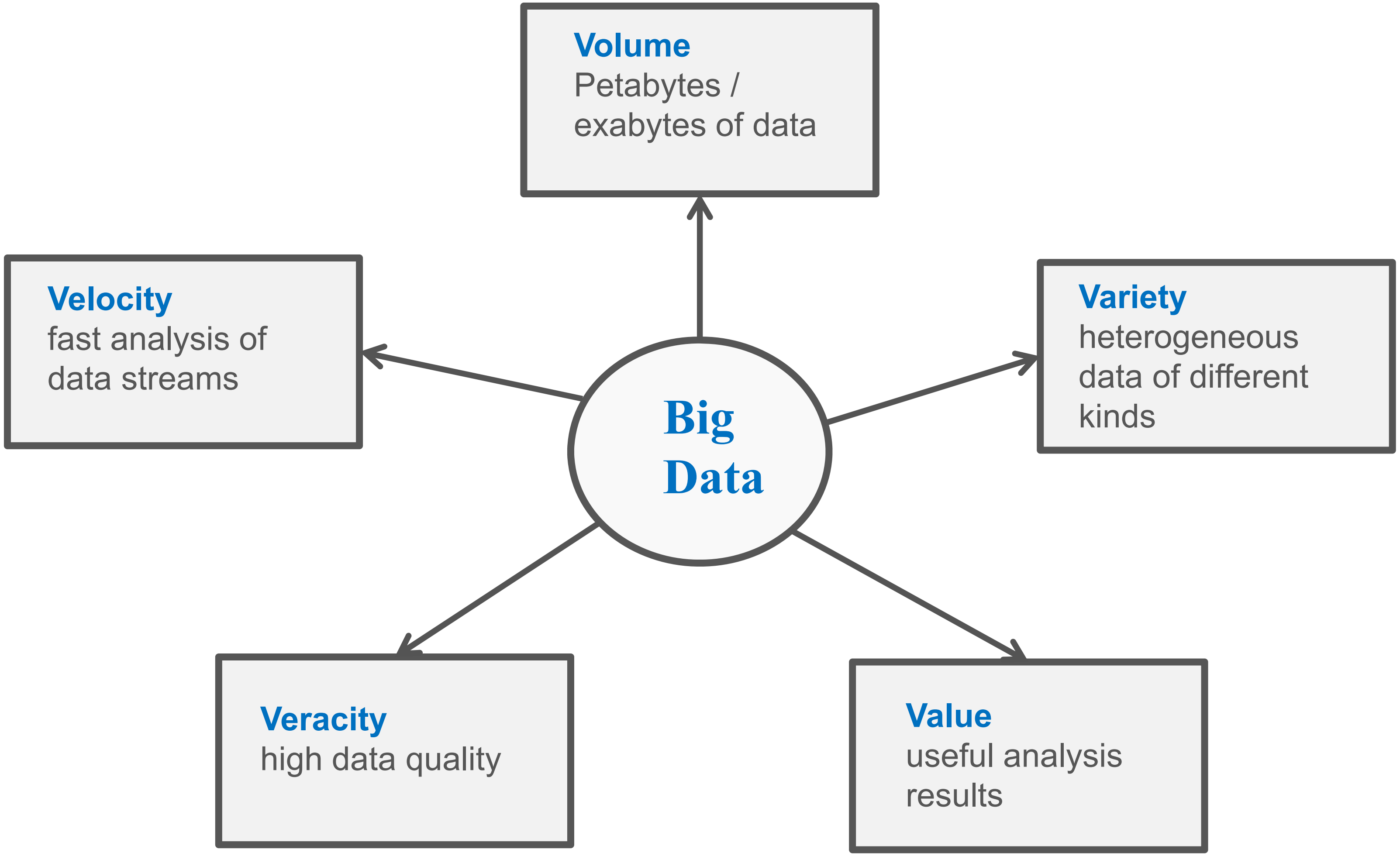
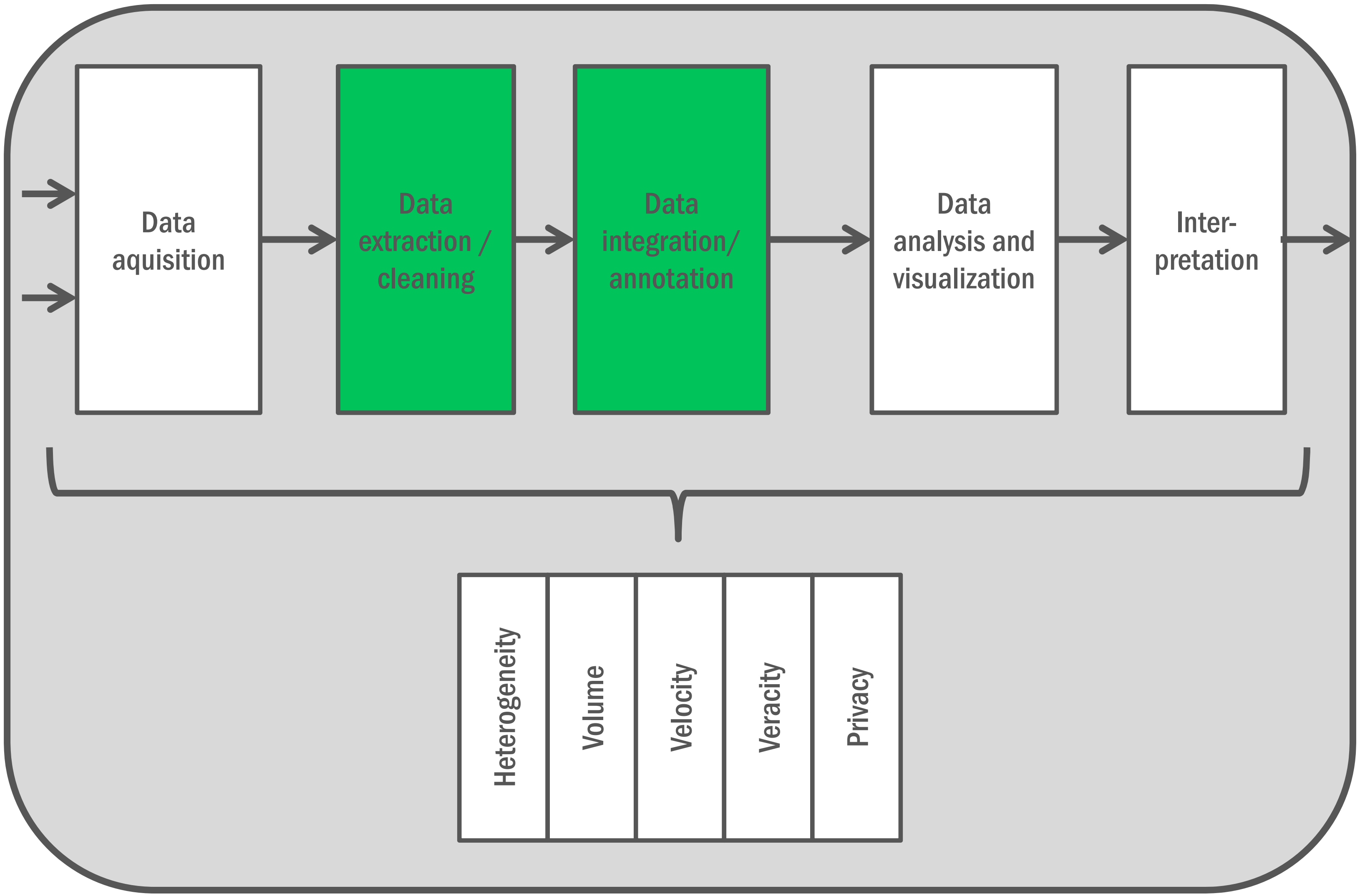
BIG DATA INTEGRATION RESEARCH AT SCADS

Erhard Rahm
Eric Peukert
Alieh Saeedi
Marcel Gladbach

BIG DATA CHALLENGES



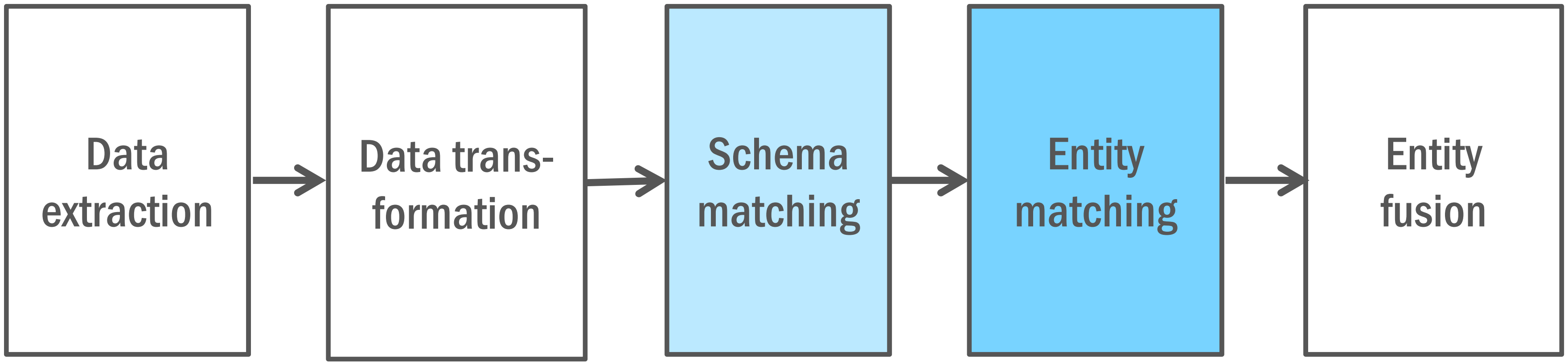
BIG DATA ANALYSIS PIPELINE



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

KEY STEPS IN DATA INTEGRATION



ENTITY MATCHING

- also called entity resolution, record linkage, deduplication …
- identification of semantically equivalent entities
  - 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

WEB ENTITIES: PRODUCT OFFERS

Canon VIXIA HF S10 Camcorder - 1080p - 8.59 MP - 10 x optical zoom
Flash card, 32 GB, 1y warranty, F/1.8-3.0
The VIXIA HF S10 delivers brilliant video and photos through a Canon exclusive 8.59 megapixel CMOS image sensor and the latest version of Canon's advanced image processor, ...
12 reviews - Add to Shopping List

$975 new
from 52 sellers
Compare prices

Canon (VIXIA) HF S10 iVIS Dual Flash Memory Camcorder
Canon HF S10 iVIS Dual Flash Memory CamcorderSPECIAL SALE PRICE: $899 Display both English/Japanese + we supplu all English manuals in English as PDF. **...**.
Add to Shopping List

$899.00 new
Made in Japan Online

Canon VIXIA HF S10
Dual Flash Memory High Definition Camcorder The Next Step Forward in HD Video Canon has a well-known and highly-regarded reputation for optical excellence, **...**.
Add to Shopping List

$999.00 new
Performance Audio
2 seller ratings

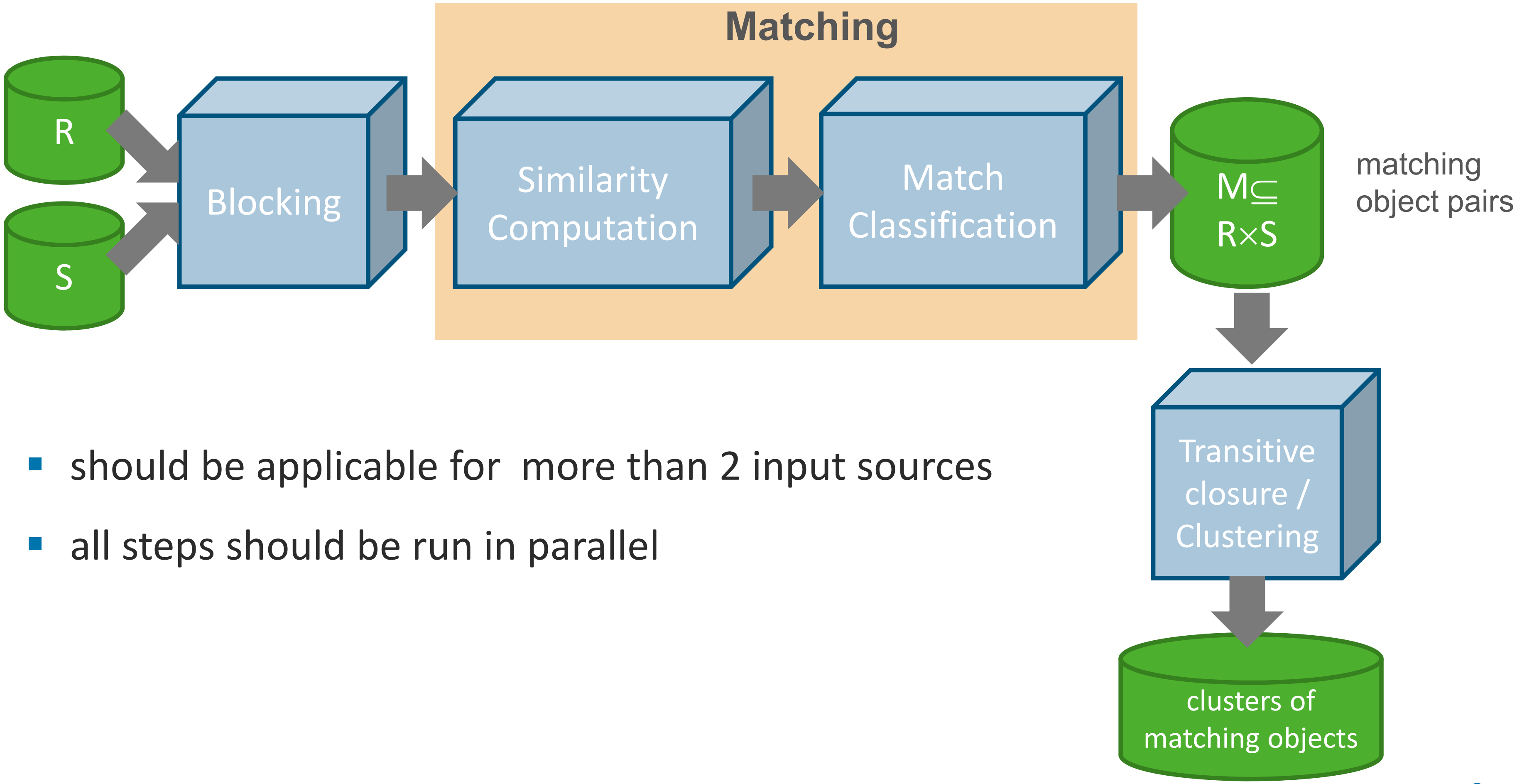
Canon VIXIA HF S100 Flash Memory Camcorder
***Canon Video HF S100 Instant Rebate Receive $200 with your purchase of a new Canon VIXIA HF S100 Flash Memory Camcorder. (Price above includes $200 **...**.
Add to Shopping List

$899.95 new
Arlingtoncamera.com
5 seller ratings

Canon Vixia Hf S10 Care & Cleaning
Care & Cleaning Digital Camera/Camcorder Deluxe Cleaning Kit with LCD Screen Guard Canon VIXIA HF S10 Camcorders Care & Cleaning.
Add to Shopping List

$2.99 new
shop.com
38 seller ratings

ENTITY MATCHING WORKFLOW



- should be applicable for more than 2 input sources
- all steps should be run in parallel

ScaDS DRESDEN LEIPZIG

KEY DATA INTEGRATION CHALLENGES

UNIVERSITÄT LEIPZIG

- Data quality
  - unstructured, semi-structured sources
  - need for data cleaning and enrichment
- Large-scale matching
  - reduce search space, e.g. utilizing blocking techniques
  - massively parallel processing (Hadoop clusters, GPUs, etc.)
- Holistic data integration
  - support for many data sources, not only 1 or 2
  - binary integration approaches do not scale -> clustering
- Graph-based data integration
  - integrate entities of multiple types and their relationships, e.g. within knowledge graphs
  - Support for graph analytics
- Privacy for sensitive data
  - privacy-preserving record linkage and data mining

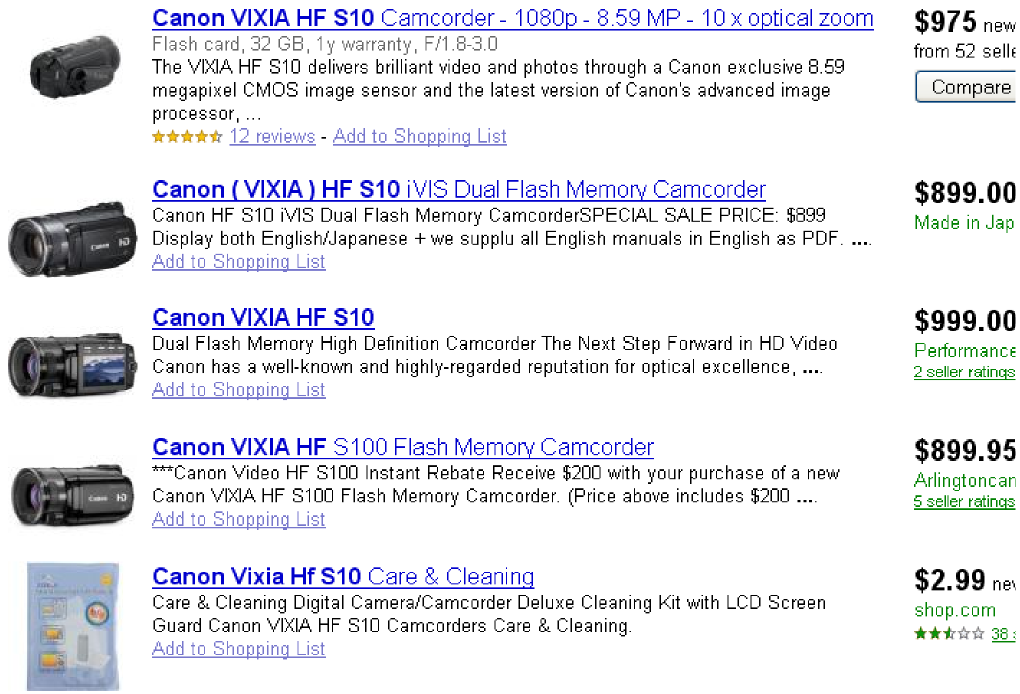
AGENDA

- Introduction
- **Scalable / holistic / graph-based matching (Rahm)**
  - Use case: Matching of product offers
  - Hadoop-based entity resolution (Dedoop)
  - Holistic data integration
  - Gradoop approach for graph-based data integration/analysis
- Demo Gradoop Service (Peukert)
- Holistic entity matching with FAMER (Saeedi)
- Privacy-preserving record linkage (Gladbach)

BIG DATA INTEGRATION USE CASE

Integration of product offers in comparison portal

- Thousands of data sources (shops/merchants)
- Millions of products and product offers
- Continous changes
- Many similar, but different products
- Low data quality



ScaDS DRESDEN LEIPZIG

SOLUTION APPROACH

UNIVERSITÄT LEIPZIG

Input:

- new product offers
- existing product catalog with associated products and offers

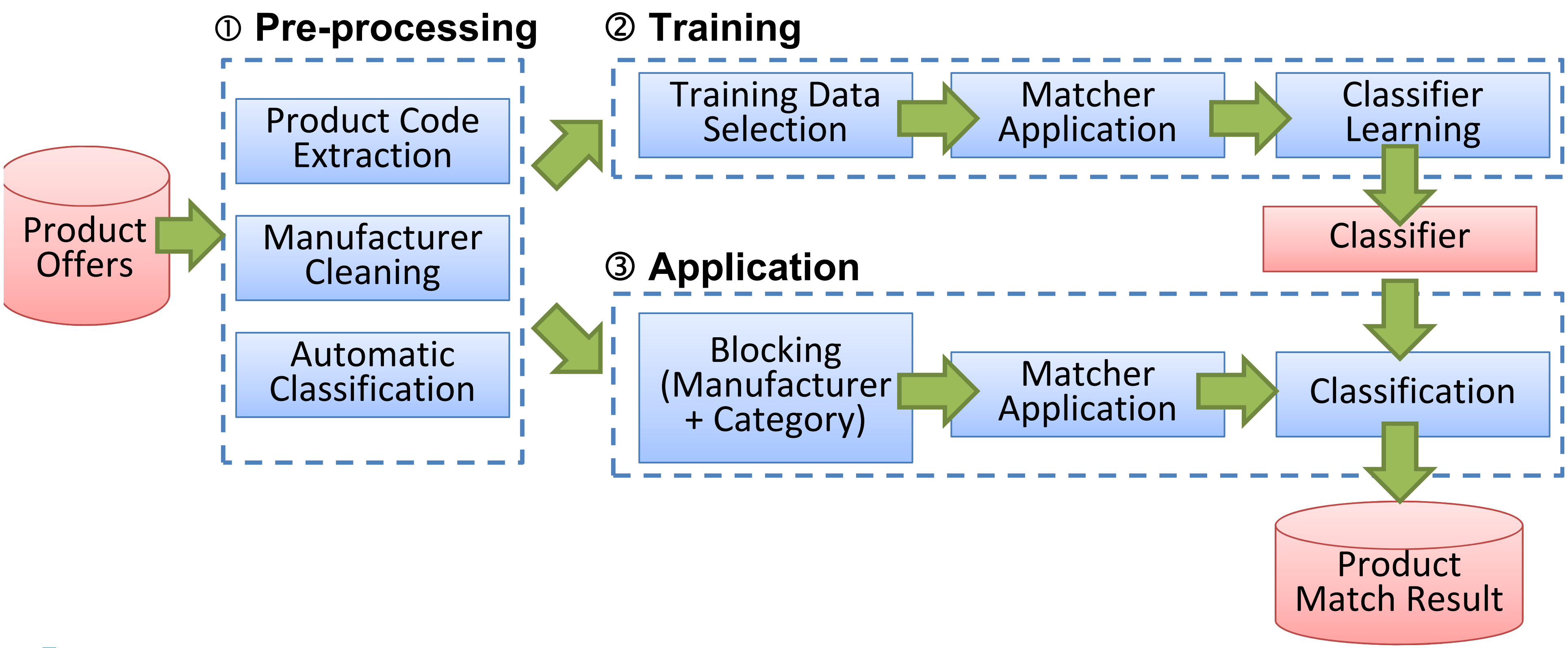
Preprocessing/ Data Cleaning:

- extraction and consolidation of manufacturer info
- extraction of product codes

Canon VIXIA HF S100 Camcorder - 1080p - 8.59 MP

Hahnel HL-XF51 7.2V 680mAh for Sony NP-FF51

LEARNING-BASED MATCH APPROACH



H. Köpcke, A. Thor, S. Thomas, E. Rahm: *Tailoring entity resolution for matching product offers*. Proc. EDBT 2012

ScaDS DRESDEN LEIPZIG

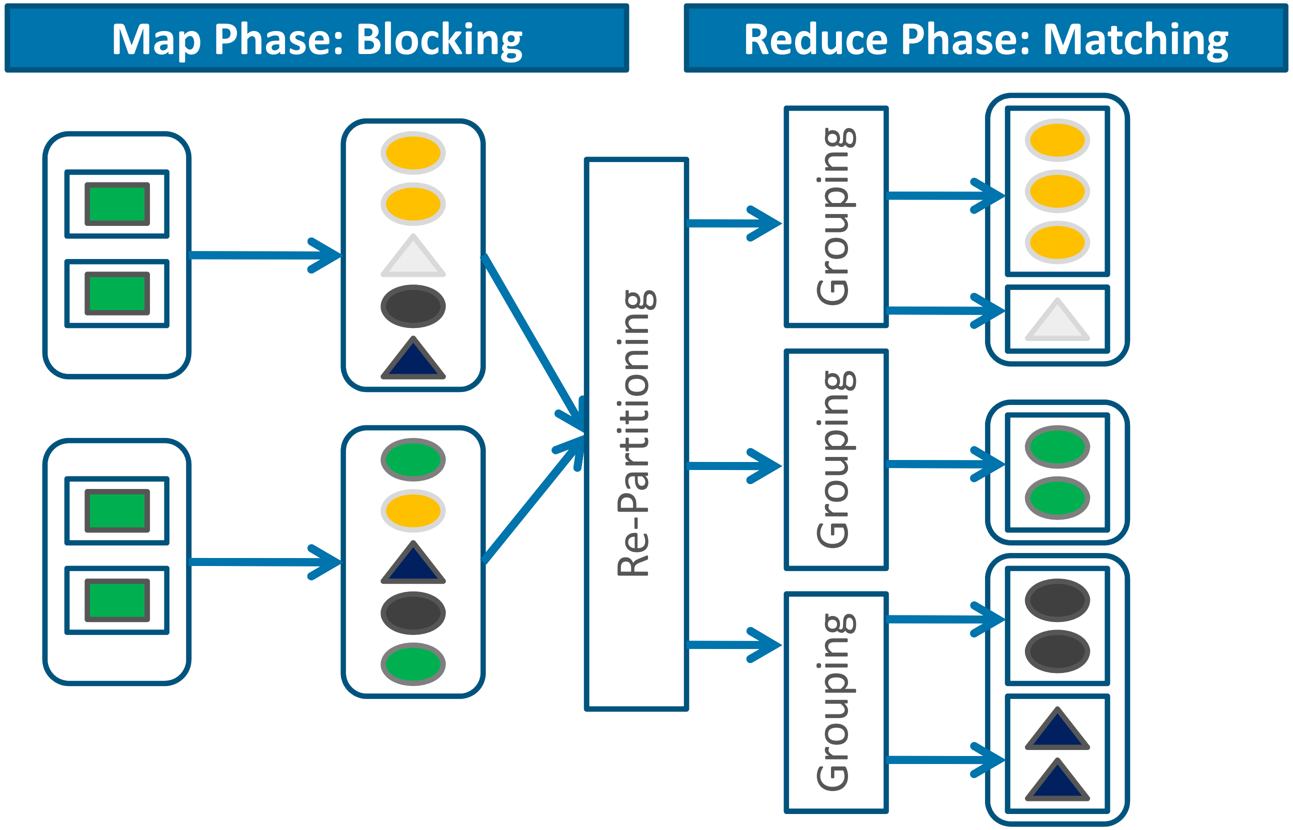
HOW TO SPEED UP ENTITY MATCHING?

UNIVERSITÄT LEIPZIG

- **Blocking** to reduce search space
  - group similar objects within blocks based on *blocking key,* e.g. manufacturer or name prefix
  - restrict matching to entities from the same block

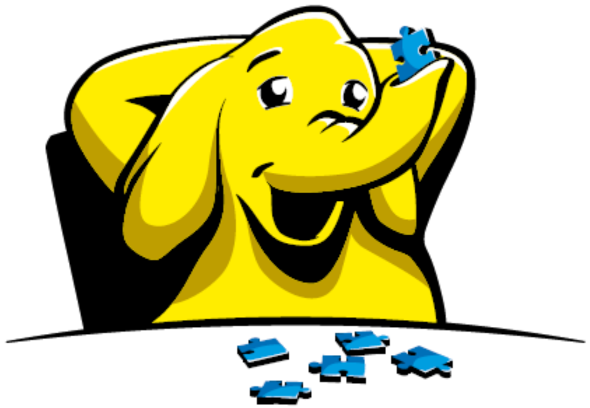
- **Parallelization**
  - split match computation in sub-tasks to be executed in parallel
  - exploitation of Big Data infrastructures such as Hadoop Map/Reduce, Apache Spark or Apache Flink

USE OF MAP/REDUCE



DEDOOP: HADOOP-BASED MATCHING

- parallel execution of data integration/ entity match workflows with Hadoop
- powerful library of match and blocking techniques
- learning-based configuration
- GUI-based workflow specification
- automatic generation and execution of Map/Reduce jobs on different clusters
- automatic load balancing for optimal scalability
- iterative computation of transitive closure



"This tool by far shows the most mature use of MapReduce for data deduplication"
www.hadoopsphere.com

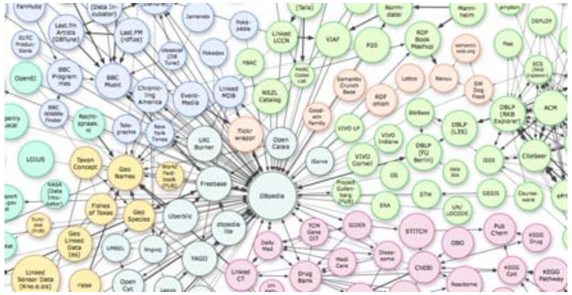
L. Kolb, A. Thor, E. Rahm: *Dedoop: Efficient Deduplication with Hadoop*. PVLDB 5(12), 2012

ScaDS DRESDEN LEIPZIG

HOLISTIC DATA INTEGRATION

UNIVERSITÄT LEIPZIG

- Scalable approaches for integrating N data sources (N >>2)
  - pairwise matching does not scale
  - 200 sources -> 20.000 mappings
- Increasing need due to numerous sources, e.g., from the web
  - many thousands of web shops
  - hundreds of LOD sources (Linked Open Data)
  - millions of web tables



- Large open data /metadata/mapping repositories
  - *dataset collections*: data.gov, datahub.io, www.opensciencedatacloud.org, web-datacommons.org

E. Rahm: *The Case fo Holistic Data Integration*. Proc. ADBIS, LNCS 9809, 2016

HOLISTIC DATA INTEGRATION: USE CASES

- Entity search engines
  - clustering of matching entities (publications, product offers)
  - physical data integration
  - thousands of data sources

Google Scholar

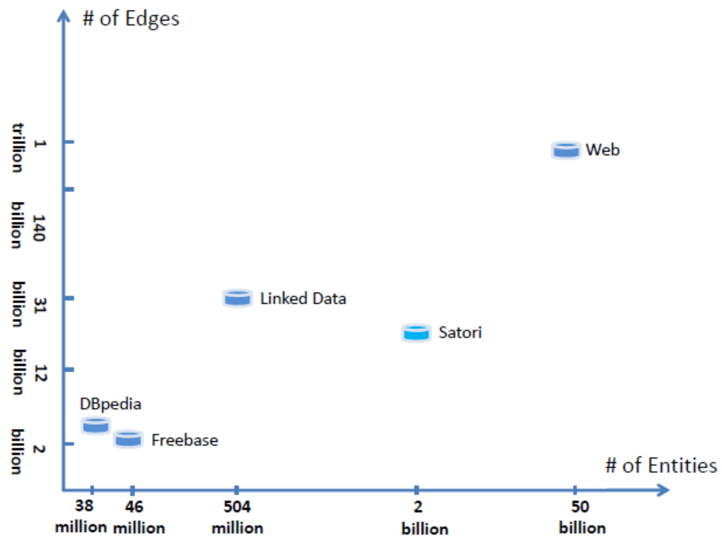
[PDF] **Data cleaning**: Problems and current approaches
E Rahm, HH Do - IEEE **Data** Eng. Bull., 2000 - academia.edu
We classify **data** quality problems that are addressed by **data cleaning** and provide an overview of the main solution approaches. **Data cleaning** is especially required when integrating heterogeneous **data** sources and should be addressed together with schema ...
☆ 99 Cited by 1654 Related articles All 35 versions

Google | Shopping

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 …

The Scale of Knowledge Graphs



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

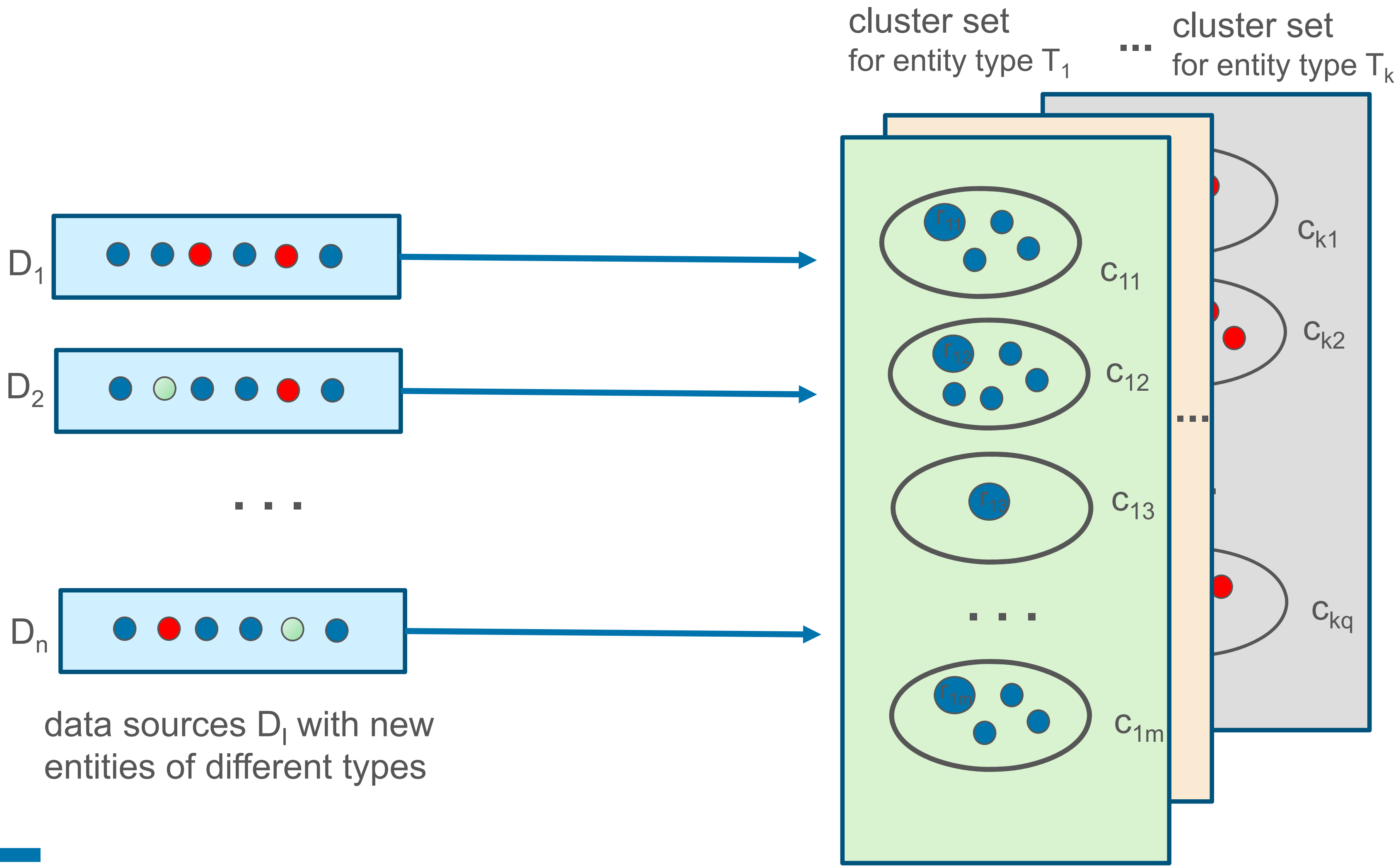
ScaDS DRESDEN LEIPZIG

HOLISTIC ENTITY RESOLUTION

UNIVERSITÄT LEIPZIG

- requirements
  - scalability to many data sources and high data volumes
  - dynamic addition of new sources /entities
  - support for many entity types
  - high match quality
  - little or no manual interaction
- binary match approaches not sufficient
- clustering-based approaches
  - represent matching entities from k sources in single cluster
  - determine cluster representative for further processing/matching
  - incremental addition/clustering of sources, e.g., starting with the largest data source
  - utilize blocking to restrict number of clusters to match with

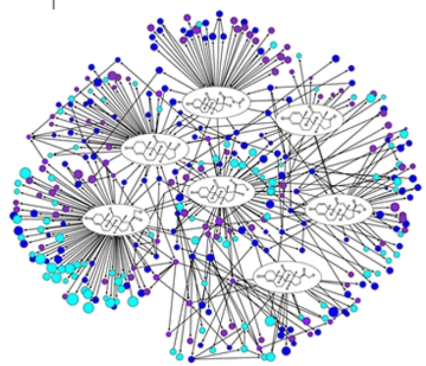
(INCREMENTAL) CLUSTERING-BASED ENTITY RESOLUTION



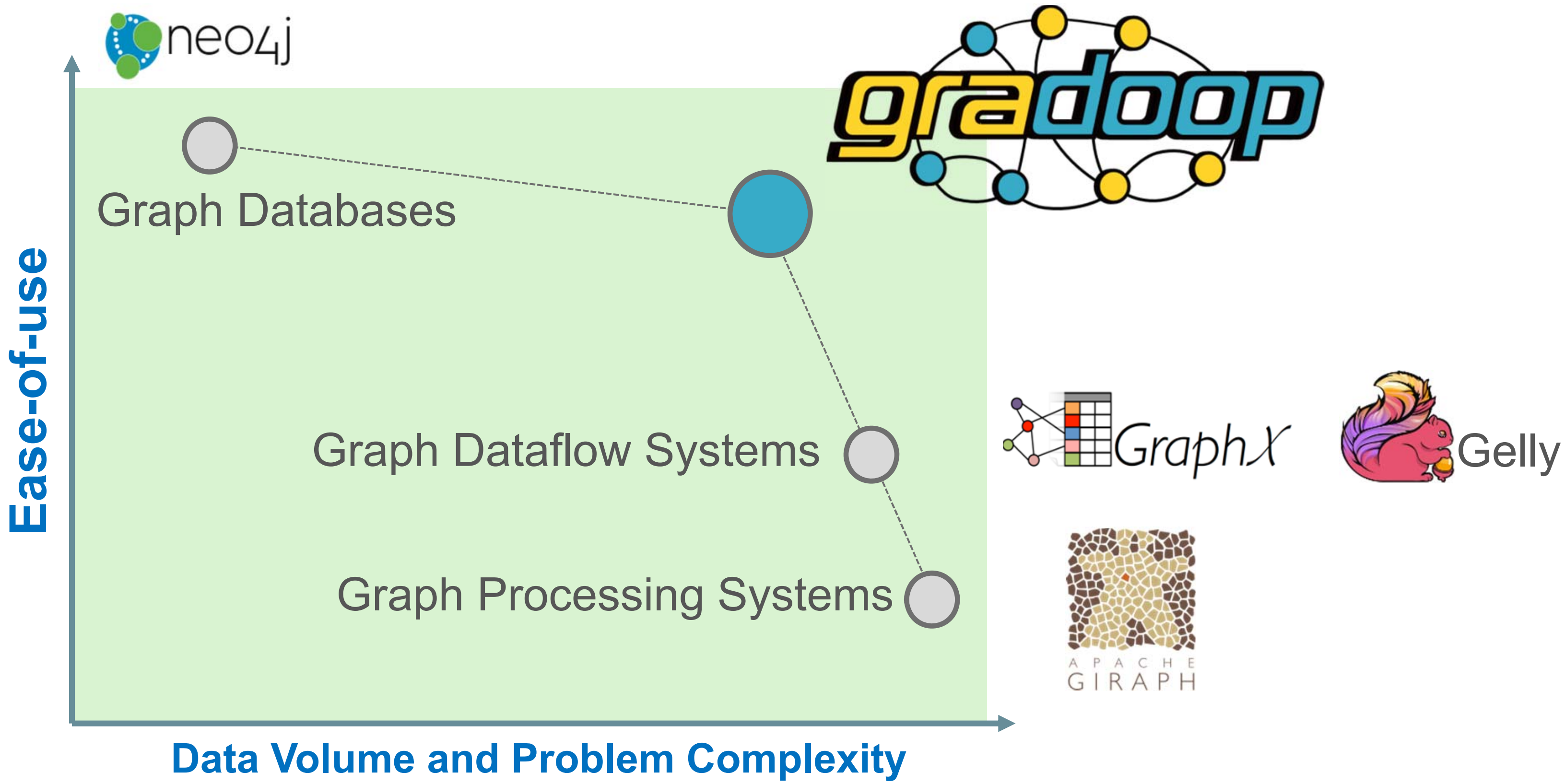
ScaDS DRESDEN LEIPZIG

GRAPH ANALYTICS

UNIVERSITÄT LEIPZIG



- advanced data analytics considering entities and their relationsships
- numerous use cases
  - social networks, bibliographic networks, bioinformatics, …
  - also useful for business intelligence
- requirements for „big“ graph analytics
  - semantically expressive graph data model supporting entites / relationships of different types, e.g. property graph model
  - powerful query and graph mining capabilities
  - high performance and scalability
  - support for graph-based data integration
  - support for versioning and evolution
  - comprehensive visualization support



neo4j
gradoop
Graph Databases
Graph Dataflow Systems
Graph Processing Systems
GraphX
Gelly
APACHE GIRAPH
Ease-of-use
Data Volume and Problem Complexity

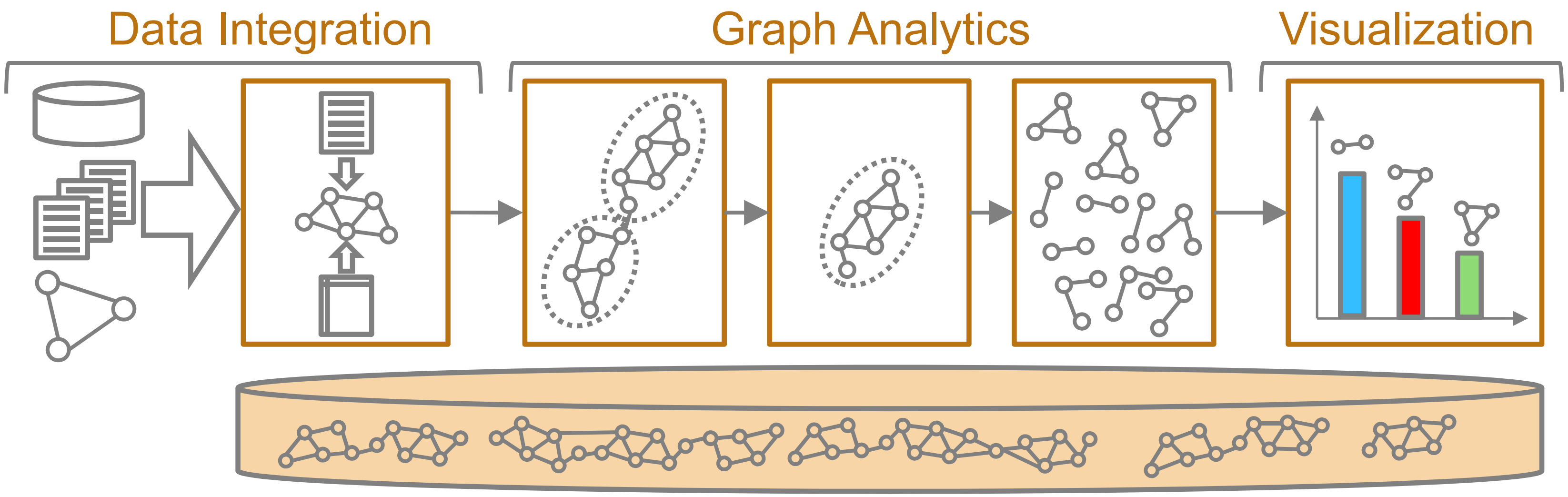
ScaDS DRESDEN LEIPZIG

GRADOOP CHARACTERISTICS

UNIVERSITÄT LEIPZIG

- Hadoop-based framework for graph data management and analysis
  - persistent graph storage in scalable distributed store (Hbase)
  - 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
- open-source implementation: www.gradoop.org

END-TO-END GRAPH ANALYTICS

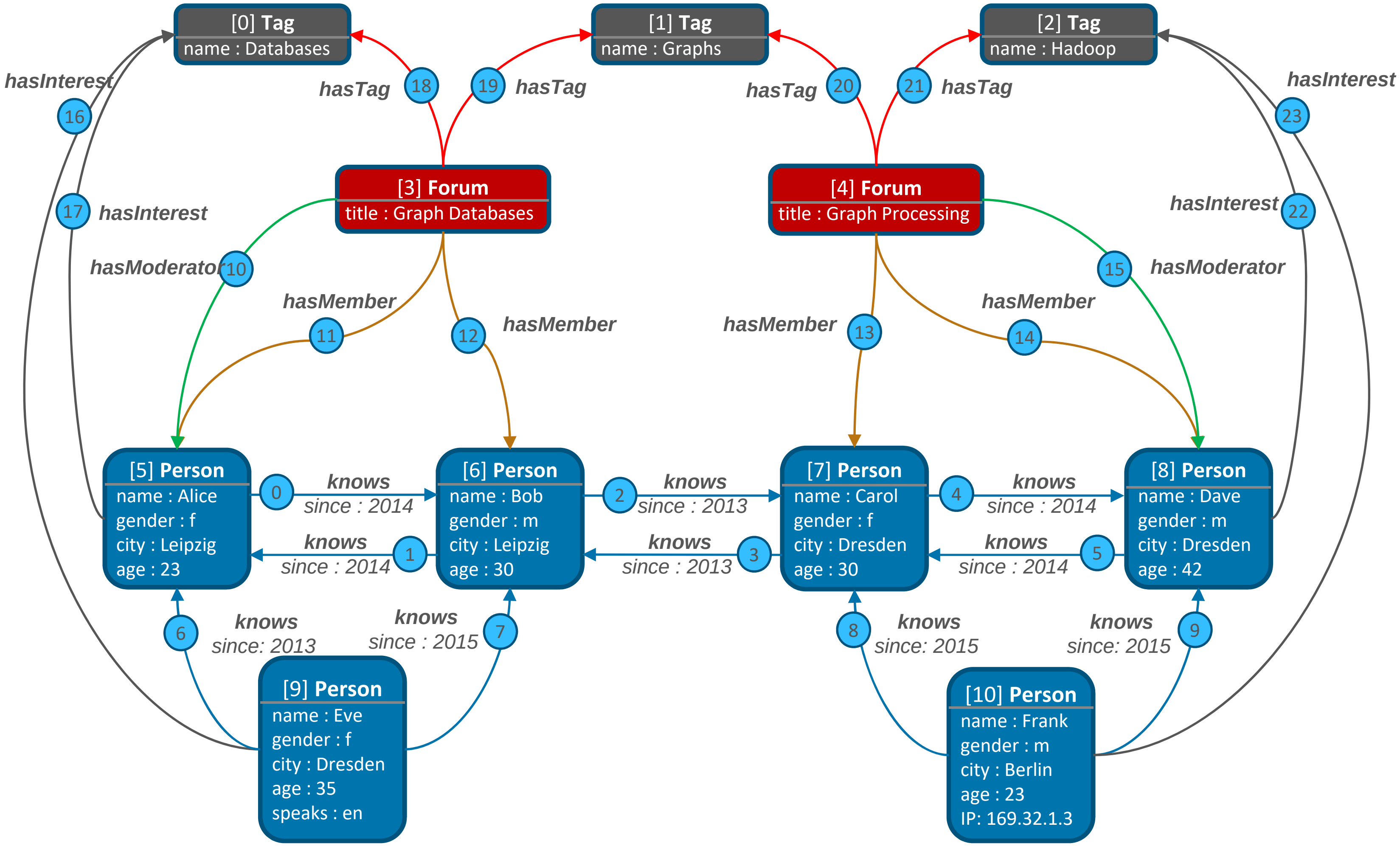


- **integrate data** from one or more sources into a dedicated **graph store** with **common graph data model**
- definition of **analytical workflows** from **operator algebra**
- result representation in **meaningful way**

	Operators: Unary	Operators: Binary	Algorithms
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 *		

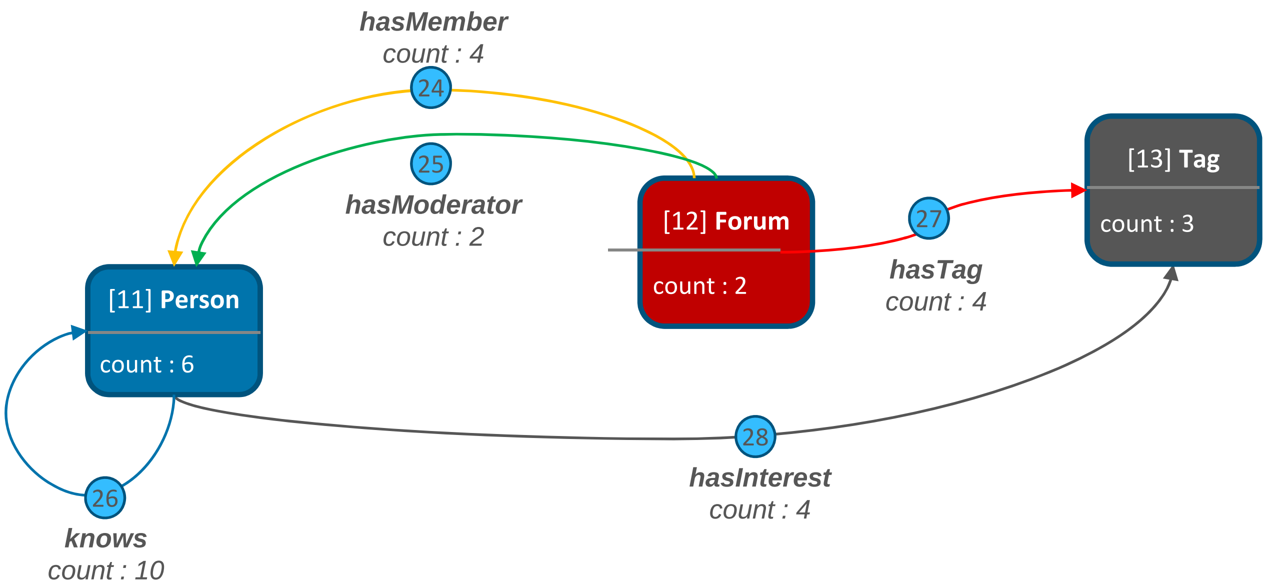
* auxiliary

SAMPLE GRAPH



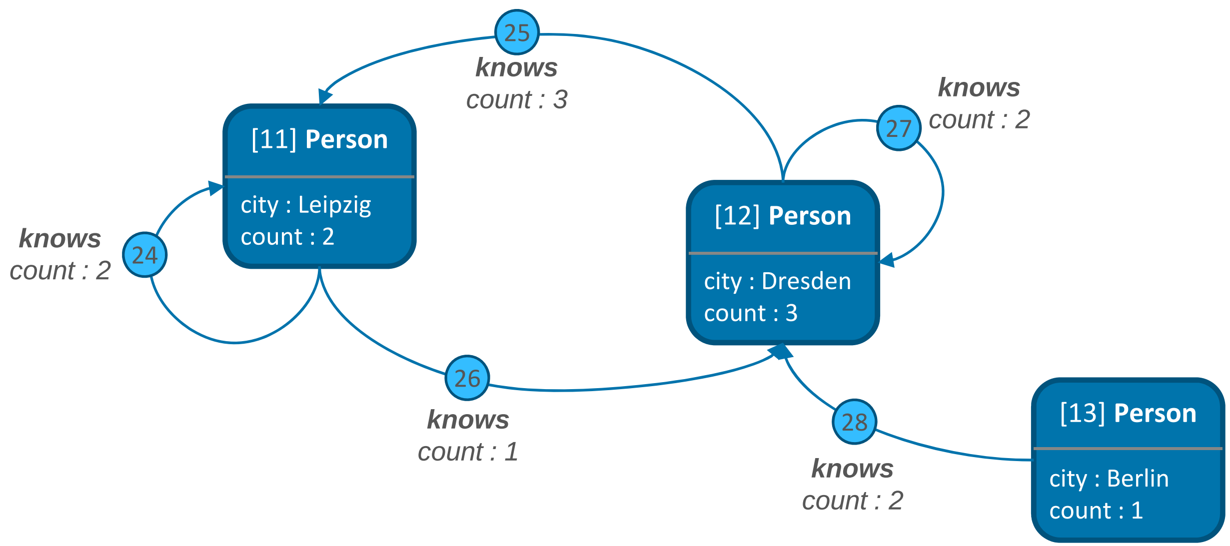
GROUPING: TYPE LEVEL *(SCHEMA GRAPH)*

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



GROUPING: PROPERTY-SPECIFIC

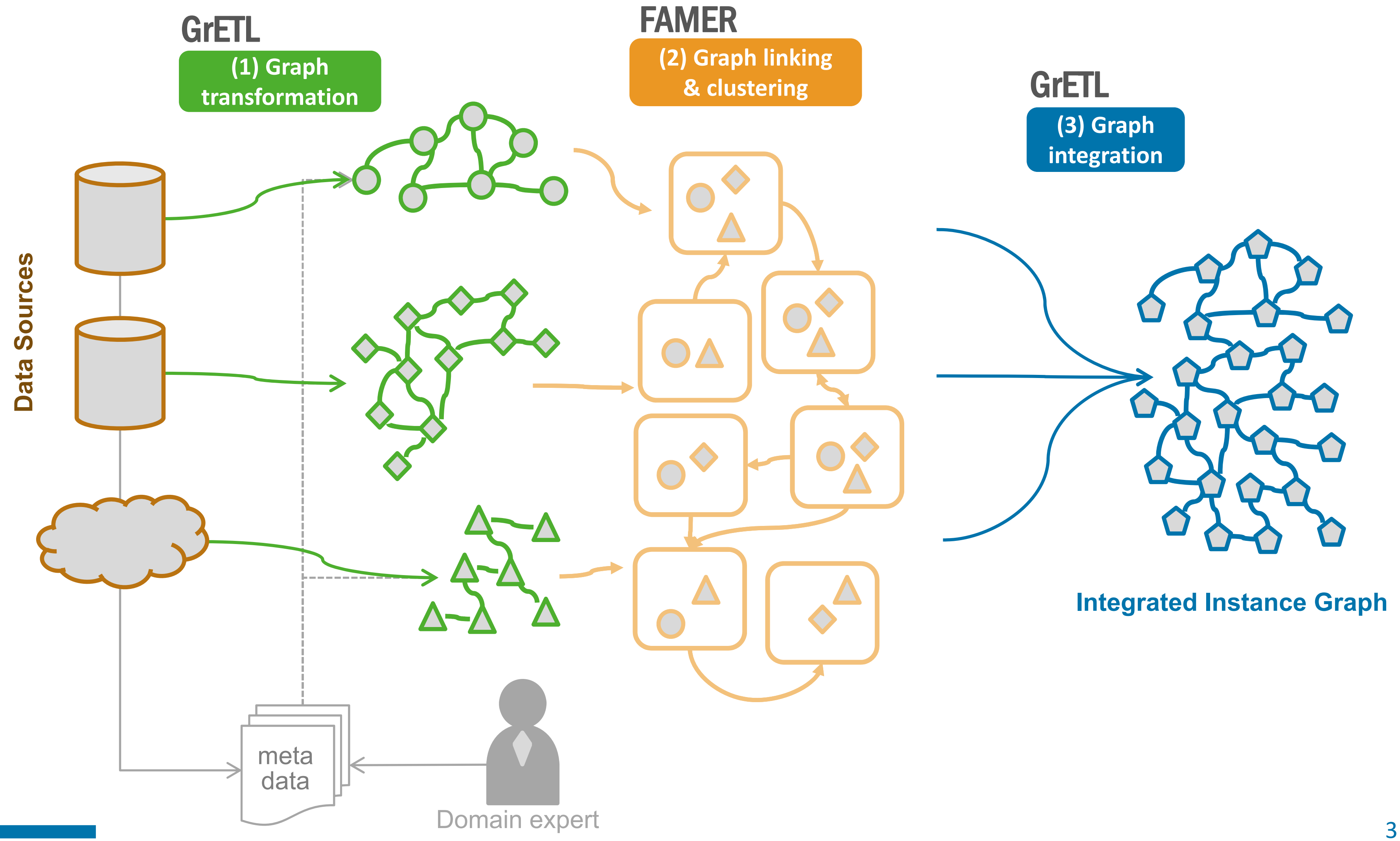
```
personGraph   = databaseGraph.subgraph((vertex => vertex[:label] == 'Person'),
                                       (edge => edge[:label] == 'knows'))
vertexGrKeys = [:label, "city"]
edgeGrKeys   = [:label]
sumGraph     = personGraph.groupBy(vertexGrKeys, [COUNT()], edgeGrKeys, [COUNT()])
```



GRAPH-BASED DATA INTEGRATION

- need to integrate diverse data from different sources (or from data lake) into semantically expressive graph representation
  - for later graph analysis
  - for constructing knowledge graphs
- traditional tasks for data acquisition, data transformation & cleaning, schema / entity matching, entity fusion, data enrichment / annotation
- most previous work for RDF data, but not for property graphs
- new challenges
  - many data sources (pairwise linking of sources not sufficient)
  - match and fuse both entities and relationships
  - several entity and relationship types
  - more complex preparatory data transformations to resolve structural heterogeneity in input sources/graphs

GRAPH DATA INTEGRATION WORKFLOW



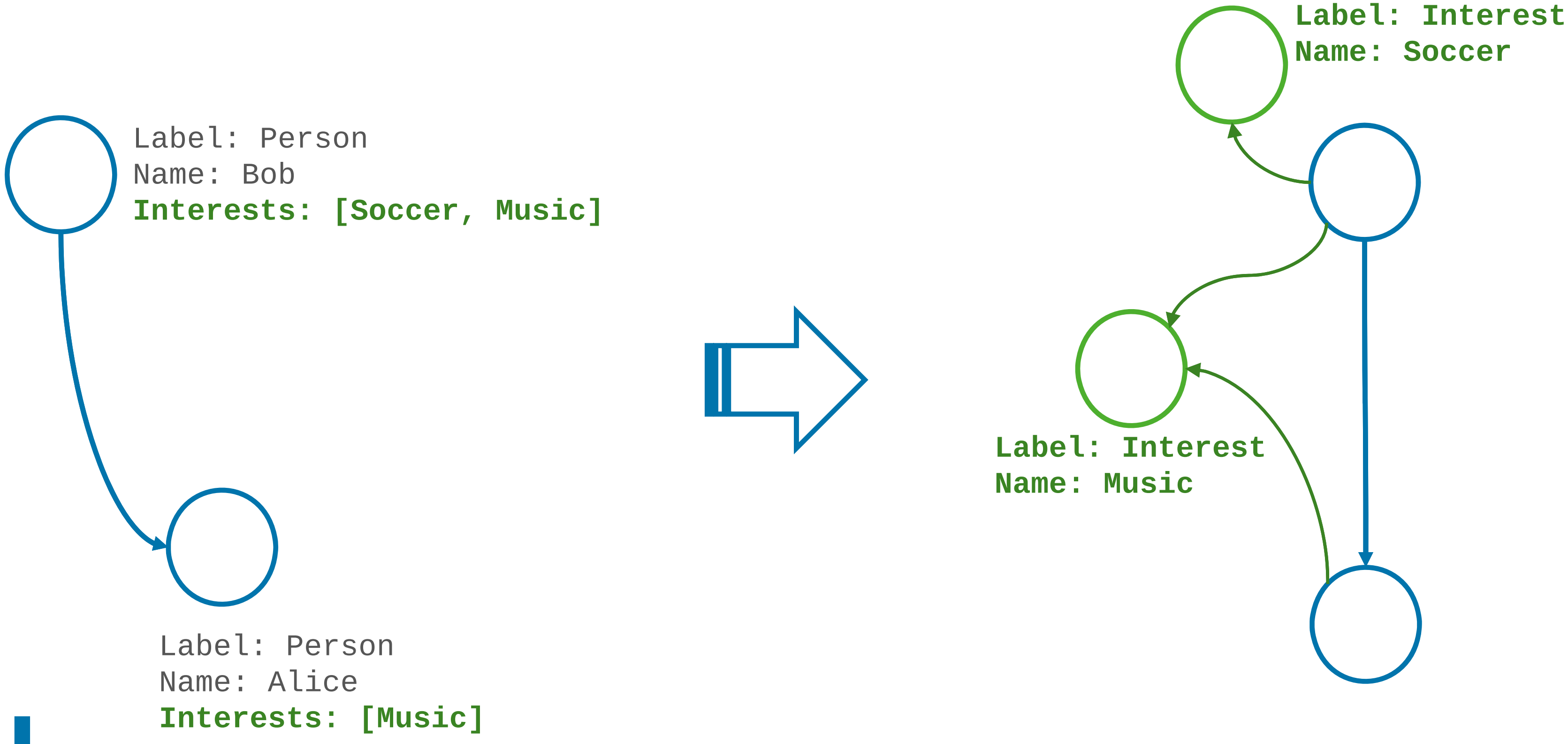
Structural Transformations

- Grouping
- Property to Vertex
  - simple deduplication
- Edge to Vertex
- Vertex to Edge
- Edges by Neighborhood
- Fuse Edges
- Cypher construct

PROPERTY TO VERTEX

Pseudocode:

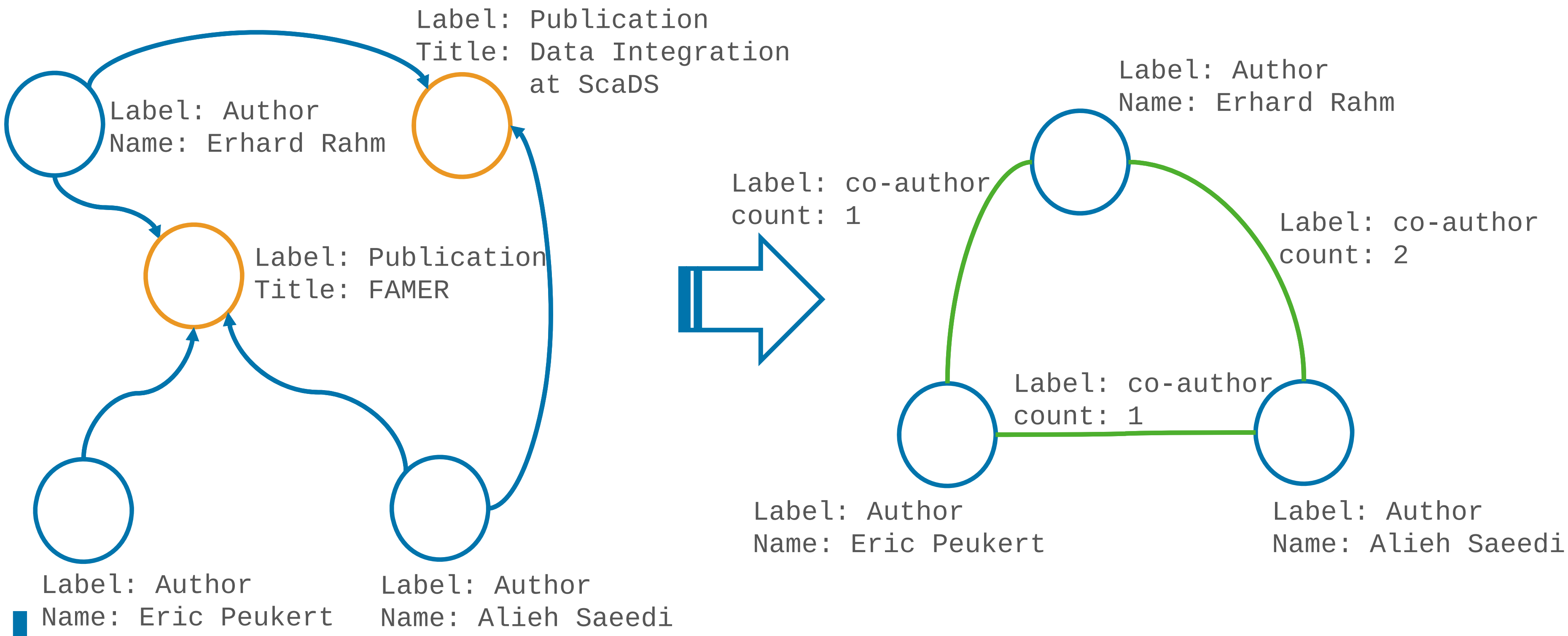
```
inputGraph
  .extractProperty(Person, Interests, Interest)
```



EDGES BY NEIGHBORHOOD

Pseudocode:

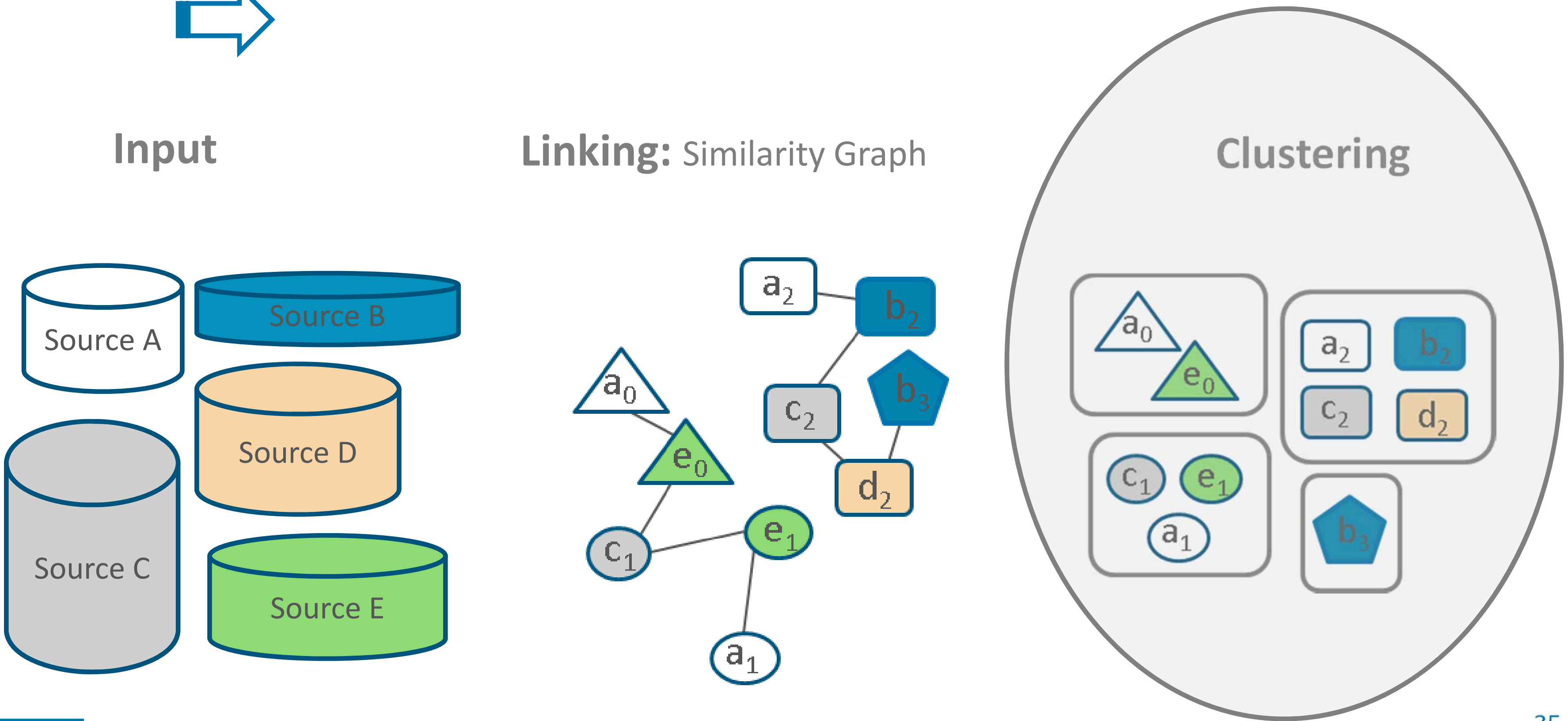
```
inputGraph
  .edgesByNeighborhood(Publication, Author, co-author)
  .fuseEdges(co-author, count, SUM)
  .vertexInducedSubgraph(ByLabel(Author))
```



ENTITY LINKING AND CLUSTERING

FAMER: scalable linking & clustering for many sources

see talk of Alieh Saeedi



ScaDS DRESDEN LEIPZIG

SUMMARY

UNIVERSITÄT LEIPZIG

- Challenges of Big Data Integration
  - Data quality and scalability
  - Many sources: need for holistic data integration / entity clustering
  - graph-based data integration for context-based matching of vertices and edges
  - Privacy-preserving record linkage (PPRL)
- Preprocessing and machine learning help to achieve high data quality (use case: matching of product offers)
- Parallel matching, e.g. based on MapReduce, Apache Spark/Flink (DEDOOP, FAMER)
- Graph-based data integration: work in progress (GRADOOP, GrETL)
  - graph-based data transformation
  - matching for multiple entity and relationship types

AGENDA

- Introduction
- Scalable / holistic / graph-based matching (Rahm)
  - Use case: Matching of product offers
  - Hadoop-based entity resolution (Dedoop)
  - Holistic data integration
  - Gradoop approach for graph-based data integration/analysis
- **Demo Gradoop Service (Peukert)**
- **Holistic entity matching with FAMER (Saeedi)**
- **Privacy-preserving record linkage (Gladbach)**

UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

FAMER

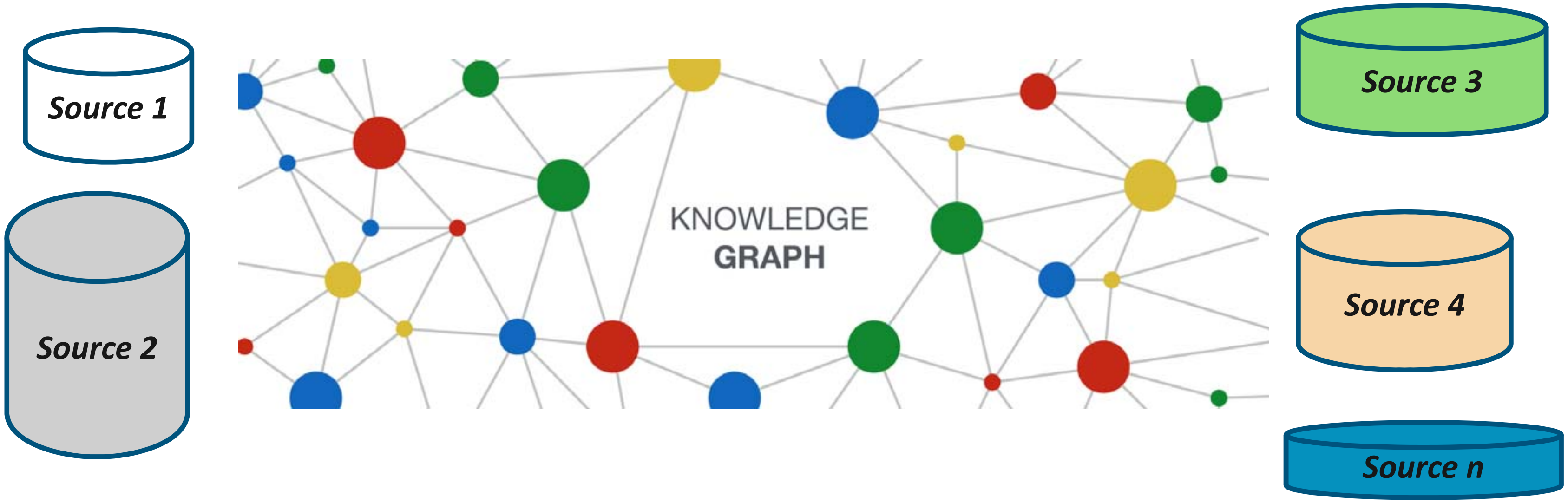
FAst Multi-source Entity Resolution System

Alieh Saeedi, Eric Peukert, Erhard Rahm

www.scads.de

MOTIVATION

- Physical data integration
  - Knowledge graph: Store data from **multiple sources** in a **graph-like** structure

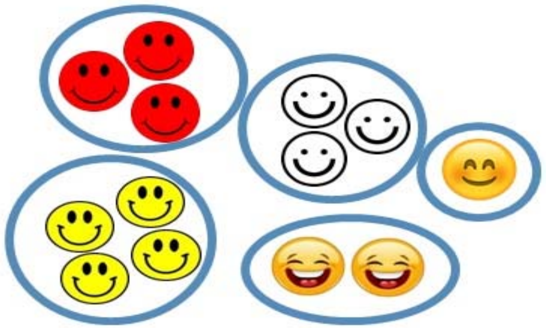


MOTIVATION

- Automatic construction & maintenance of KG: **data quality**
- Challenges for data quality
  - Entity Resolution: The task of **identifying** and **linking** entities that refer to the **same** real-world entity

2 sources:
Binary linking

N sources:
Clustering

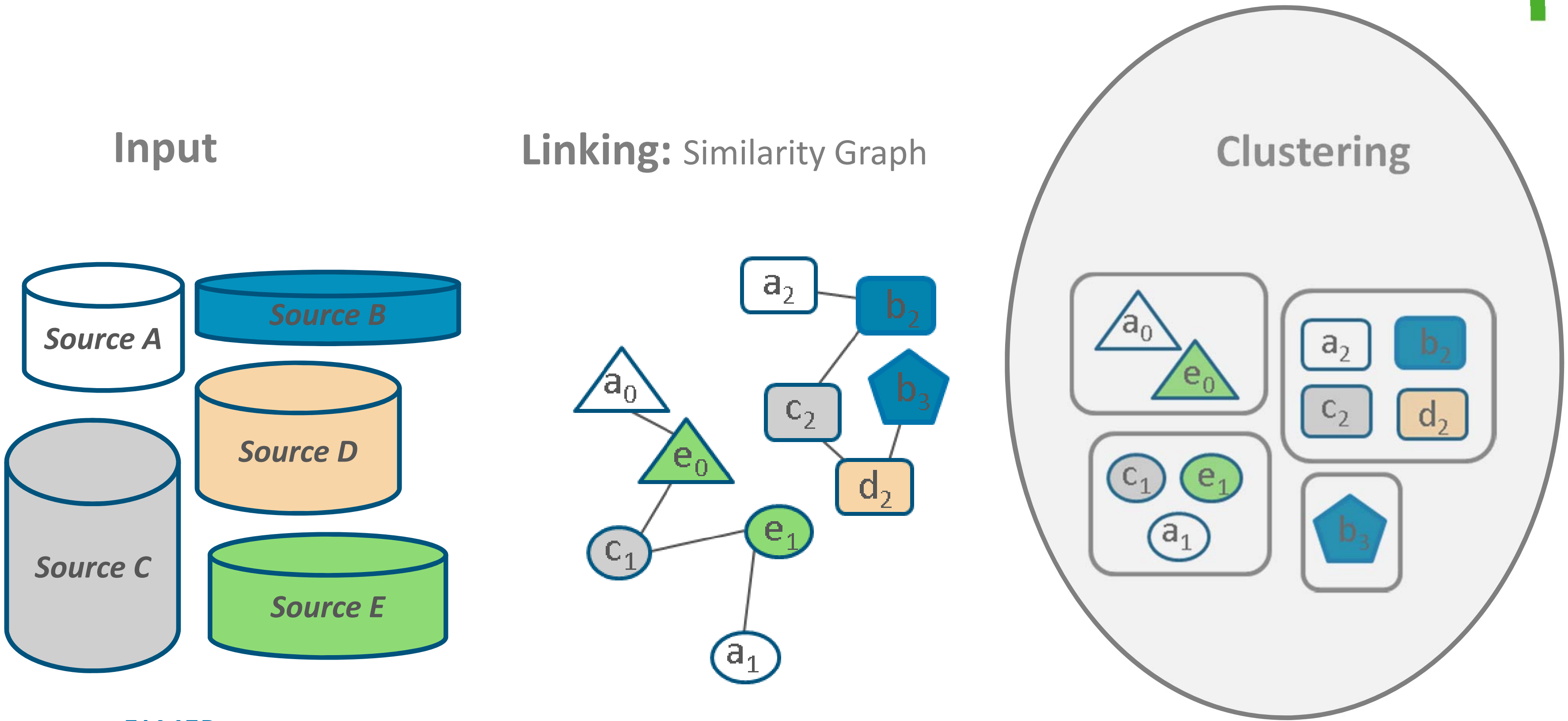


FAMER

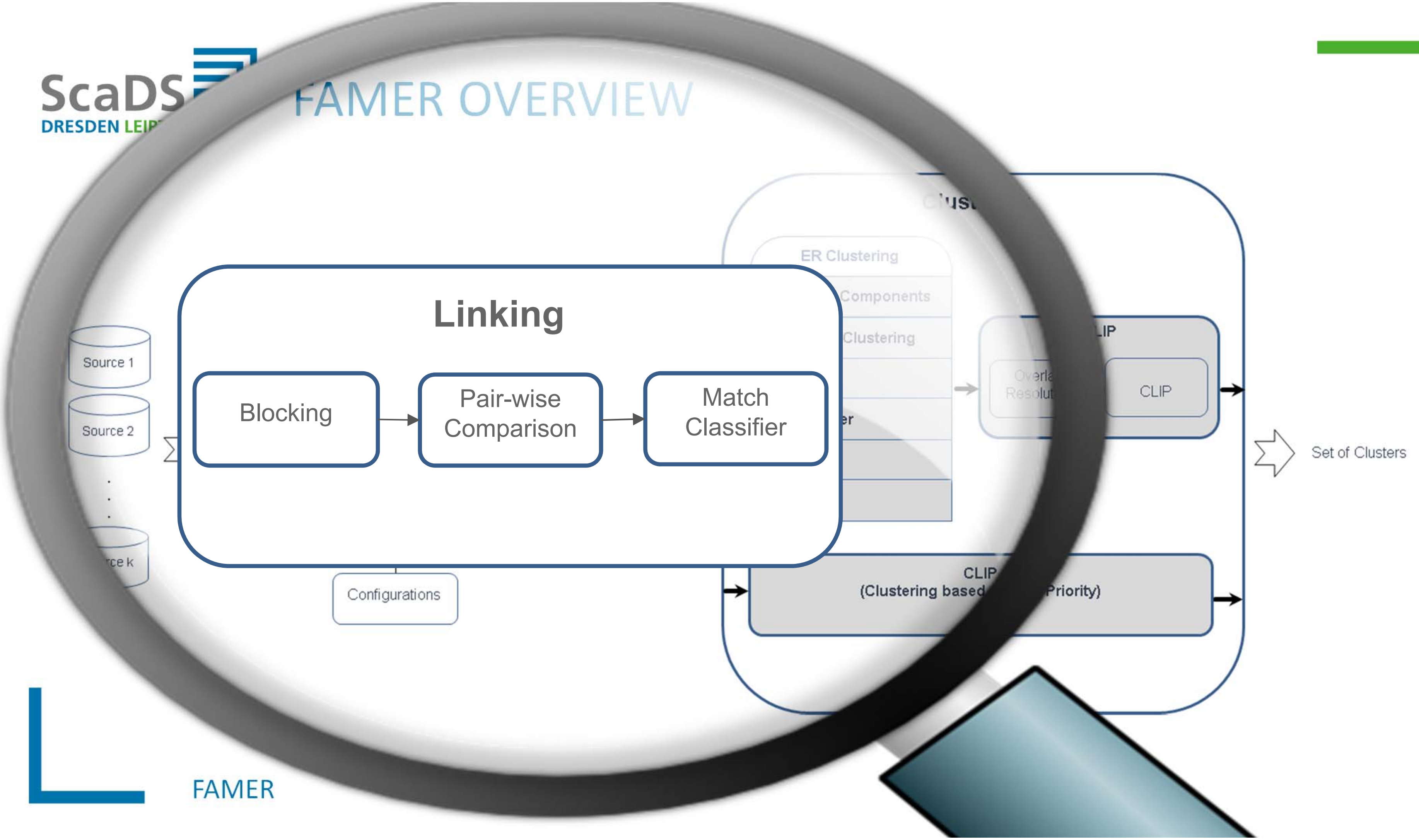
- **FAMER** (**FA**st **M**ulti-source **E**ntity **R**esolution system)
  - Scalable ER approaches for big data
    - Multiple data sources
    - Large volumes of data
  - Built on top of the distributed data flow framework Apache Flink and scalable graph analytics framework Gradoop
    - High scalability
    - Many machines

https://dbs.uni-leipzig.de/research/projects/object_matching/famer

FAMER OVERVIEW



FAMER OVERVIEW



FAMER LINKING

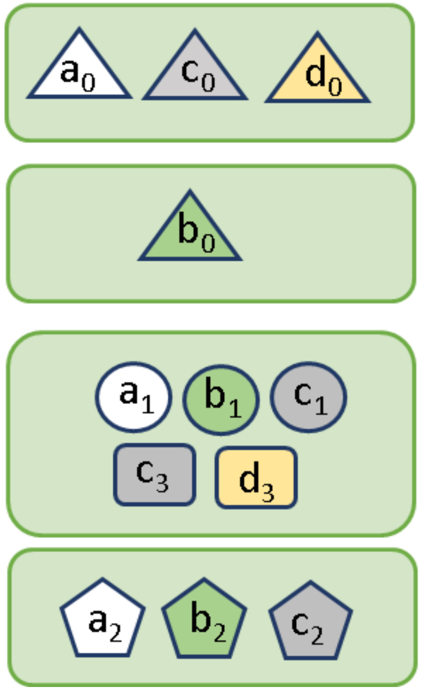
Blocking → Pair-wise Comparison → Match Classifier

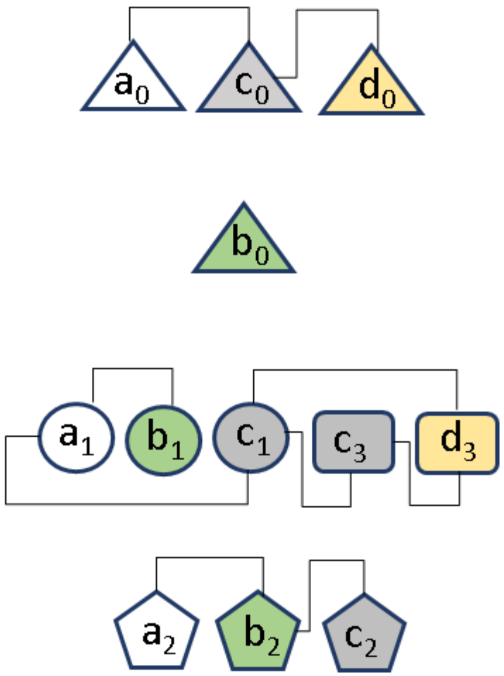
Id	Name	Surname	Suburb	Post code	SourceId
a_0	ge0rge	Walker	winston salem	271o6	Src A
b_0	George	Alker	winstom salem	27106	Src B
c_0	George	Walker	Winstons	27106	Src C
d_0	Geoahge	Waker	Winston	271oo	Src D
a_1	Bernie	Davis	pink hill	28572	Src A
b_1	Bernie	Daviis	Pinkeba	2787z	Src B
c_1	Bernii	Davs	pink hill	28571	Src C
a_2	Bertha	Summercille	Charlotte	28282	Src A
b_2	Bertha	Summeahville	Charlotte	2822	Src B
d_2	Brtha	Summerville	Charlotte	28222	Src D
c_3	Bereni	dan’lel	Pinkeba	27840	Src C
d_3	Bereni	Dasniel	Pinkeba	2788o	Src D

Id	key
a_0	wa
c_0	wa
d_0	wa
b_0	al
a_1	da
b_1	da
c_1	da
c_3	da
d_3	da
a_2	su
b_2	su
d_2	su

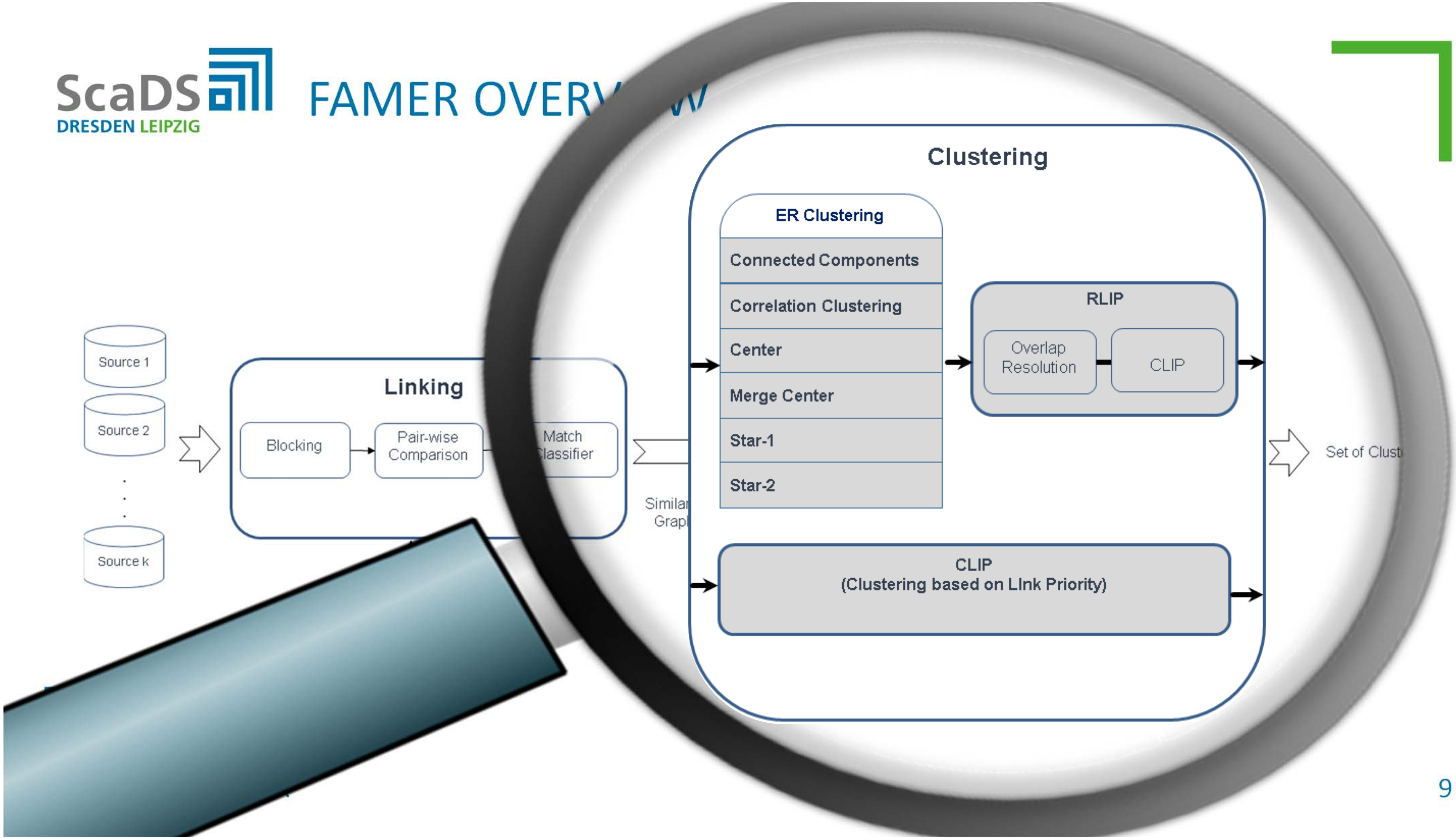
FAMER LINKING

Blocks

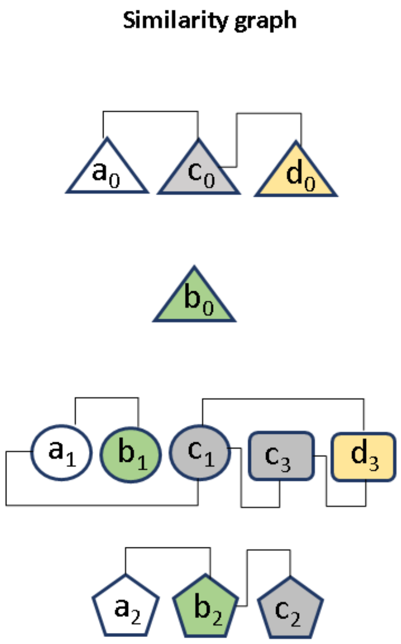


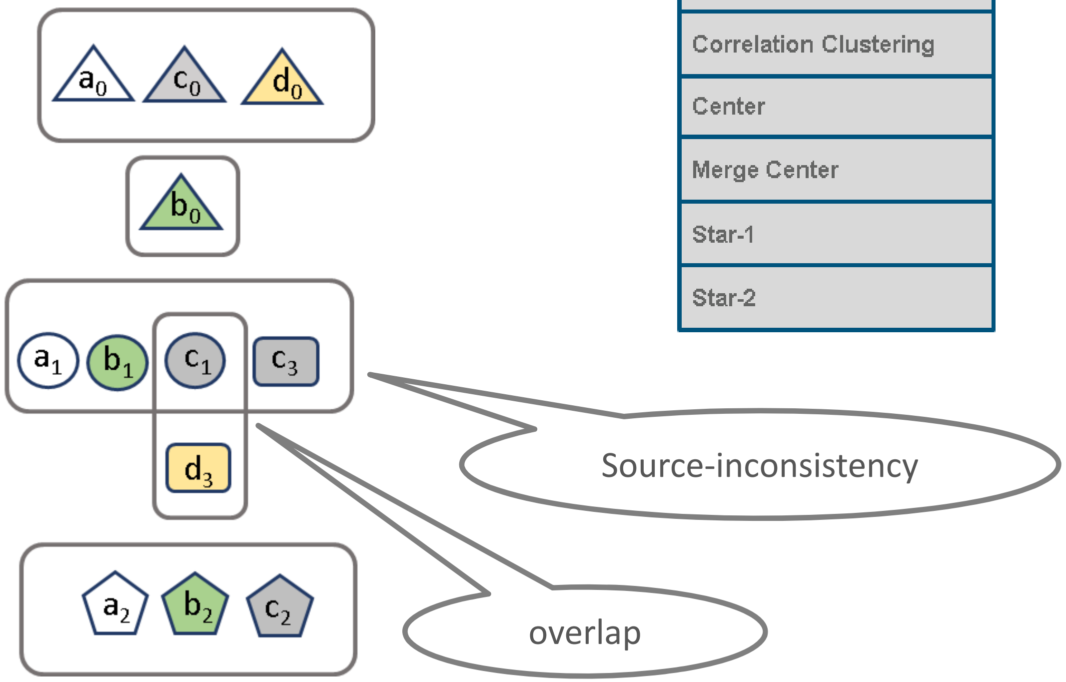


FAMER OVERV W



FAMER CLUSTERING





ER Clustering

Connected Components

Correlation Clustering

Center

Merge Center

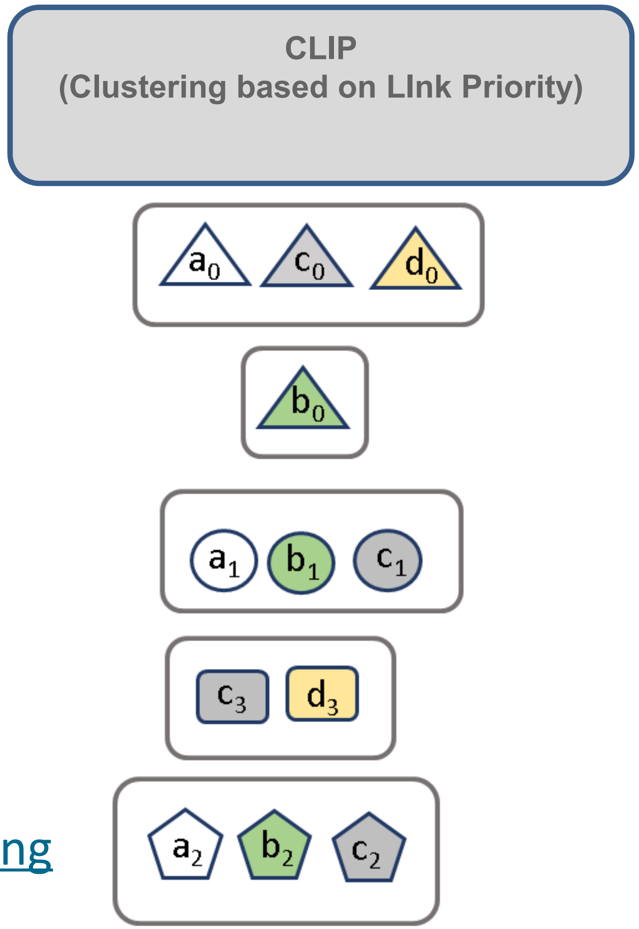
Star-1

Star-2

FAMER CLUSTERING

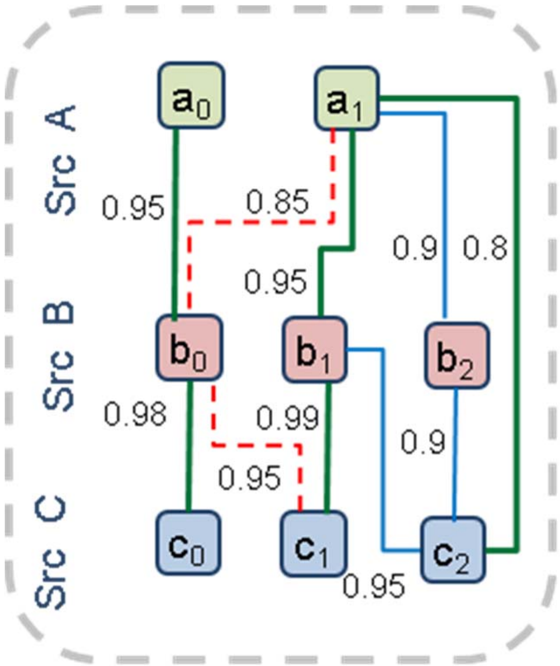
- Prioritize links based on
  - Link strength
    - Strong, Normal, Weak
  - Link degree
  - Similarity value
- produces
  - Source-consistent clusters
  - No overlap

https://github.com/dbs-leipzig/FAMER_Clustering

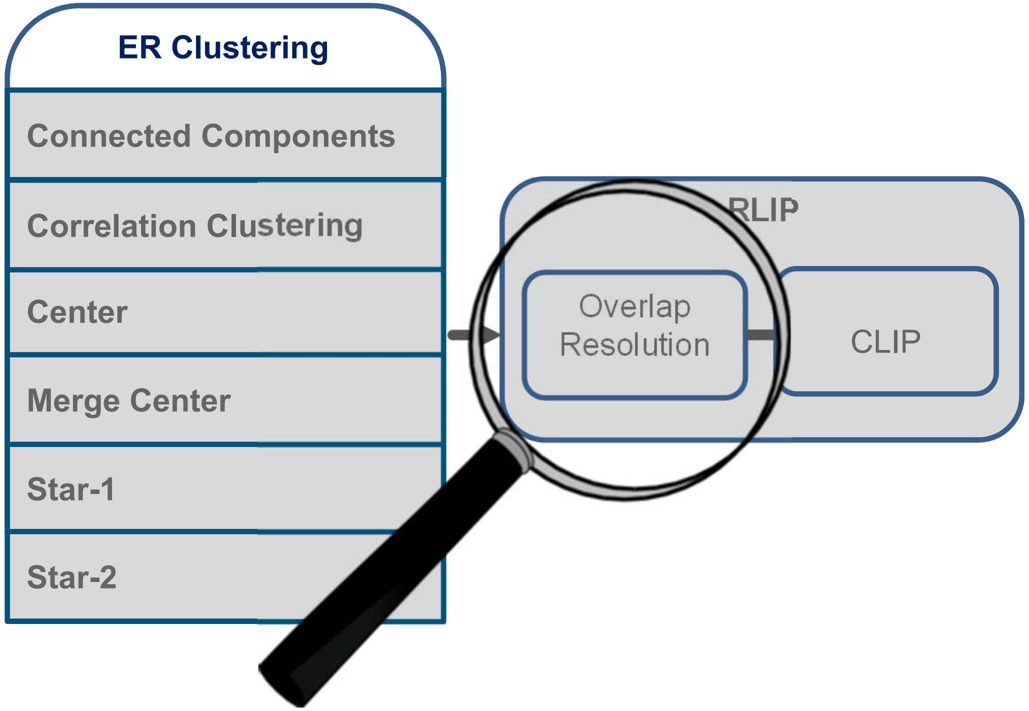


CLIP-LINK STRENGTH

- Link Strength
  - **Strong**
  - **Normal**
  - **Weak**

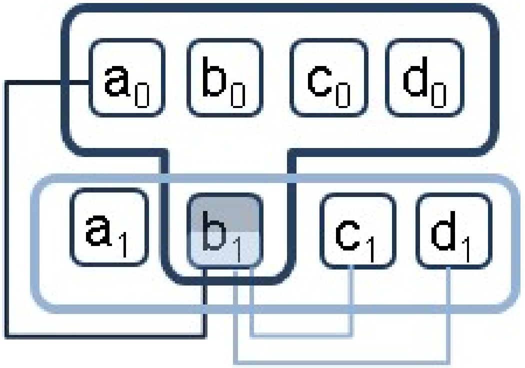


FAMER CLUSTERING

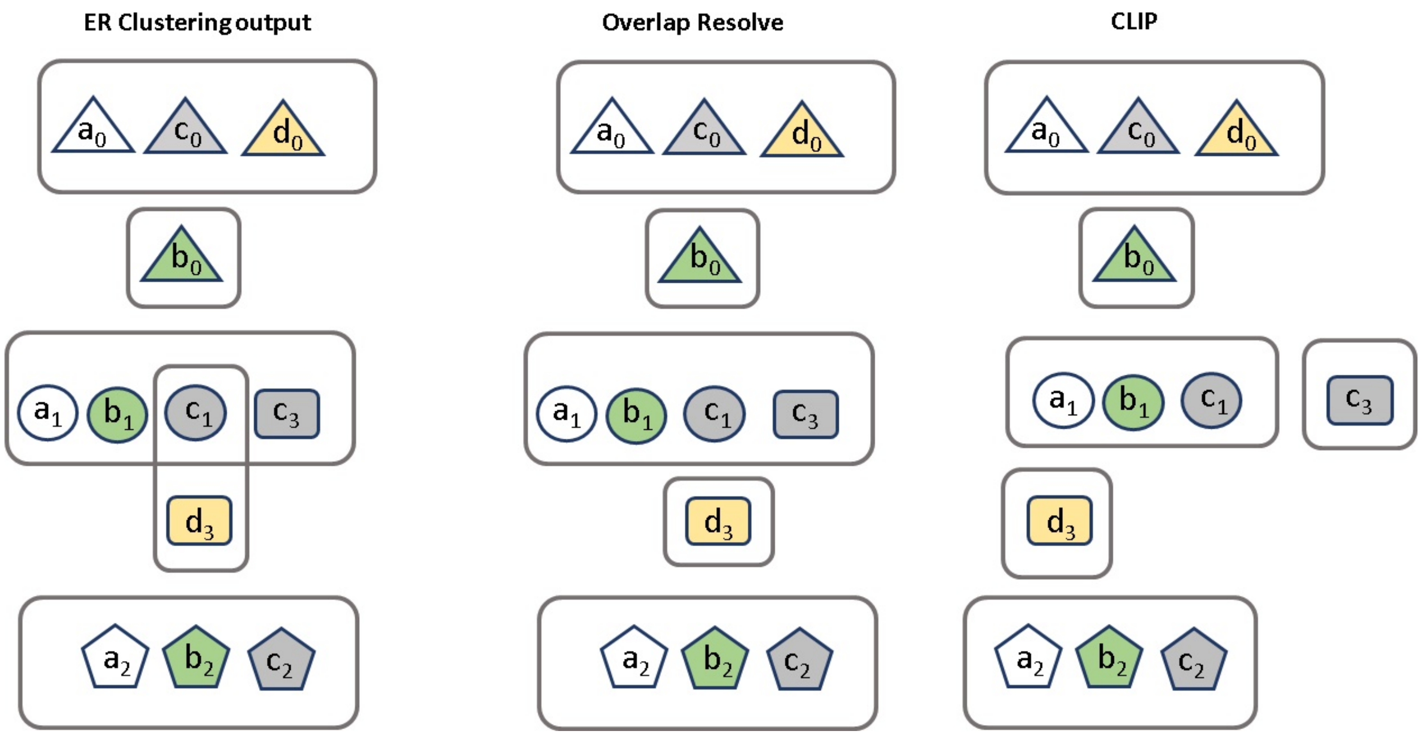


OVERLAP RESOLUTION

- Concepts
  - Link strength
  - Cluster association degree

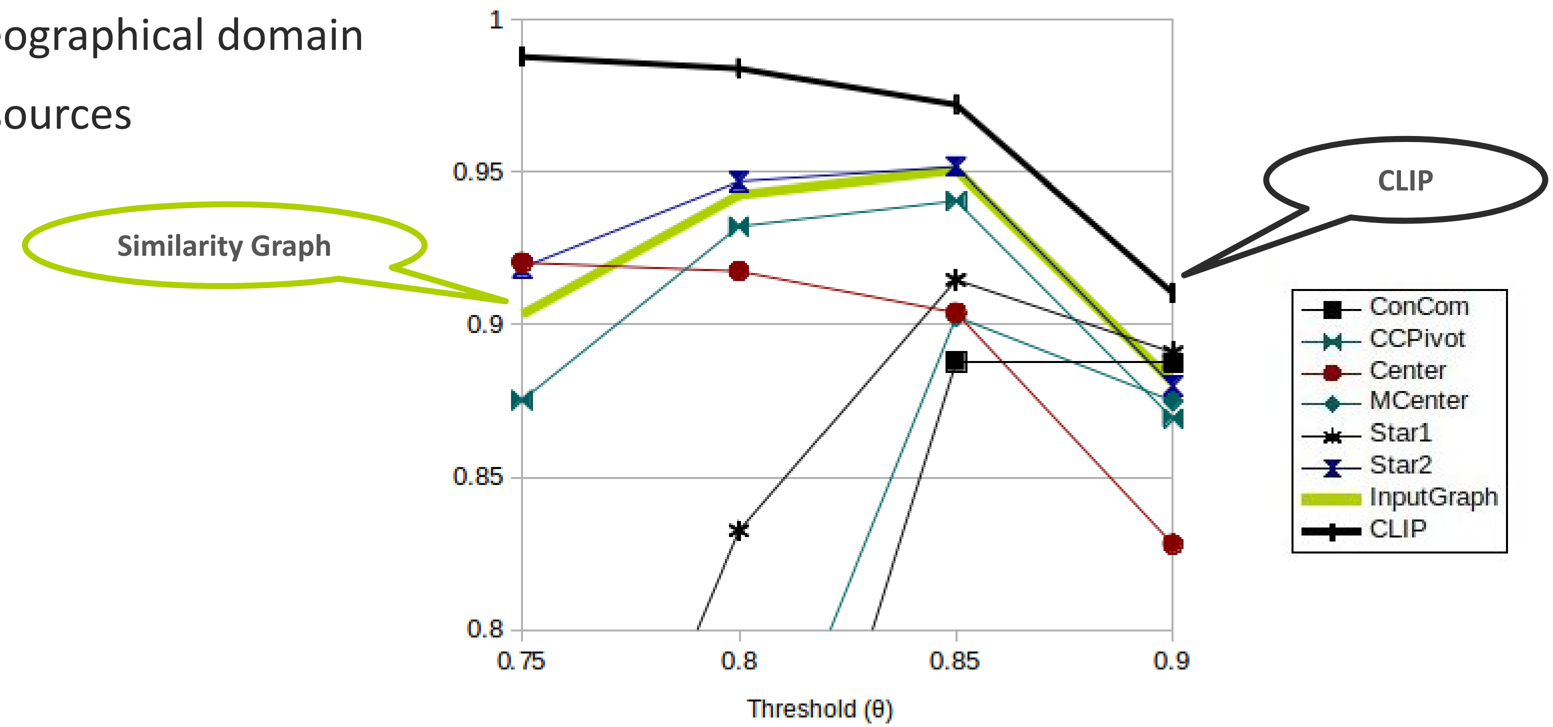


FAMER CLUSTER REPAIR

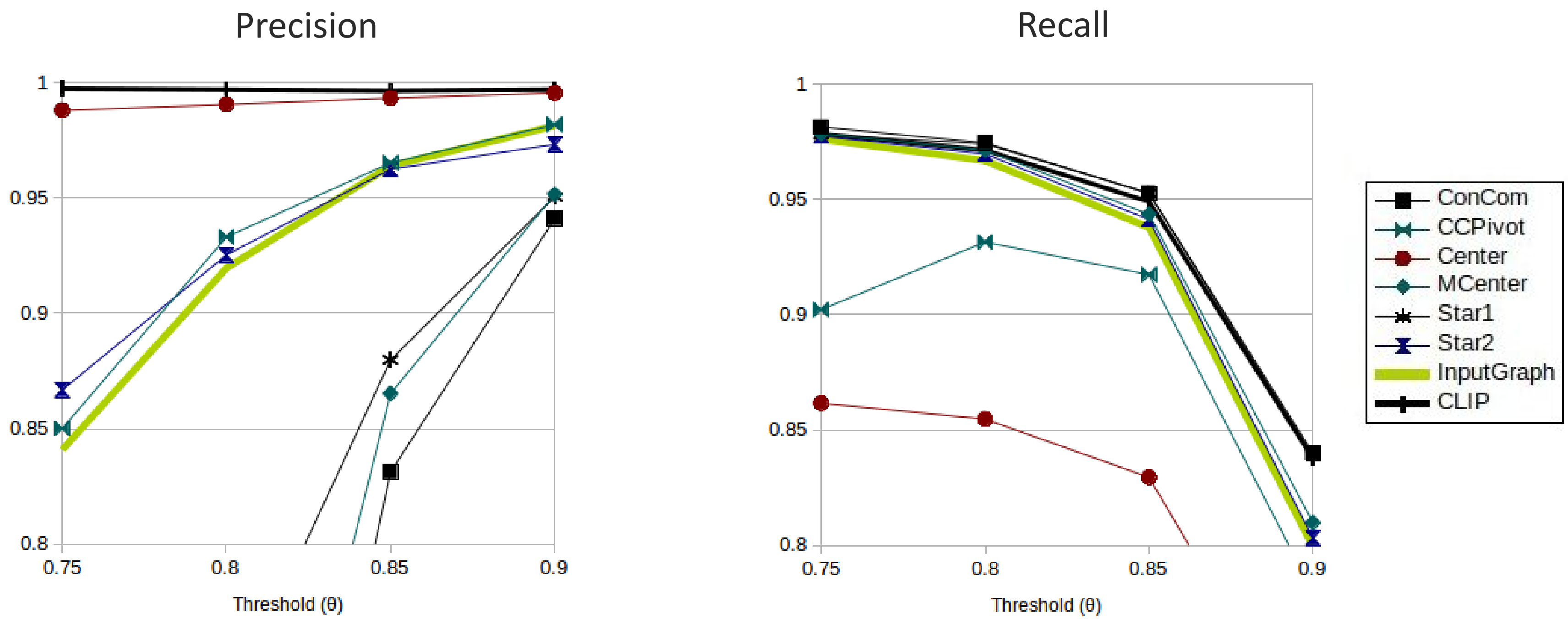


CLIP QUALITY- FMEASURE

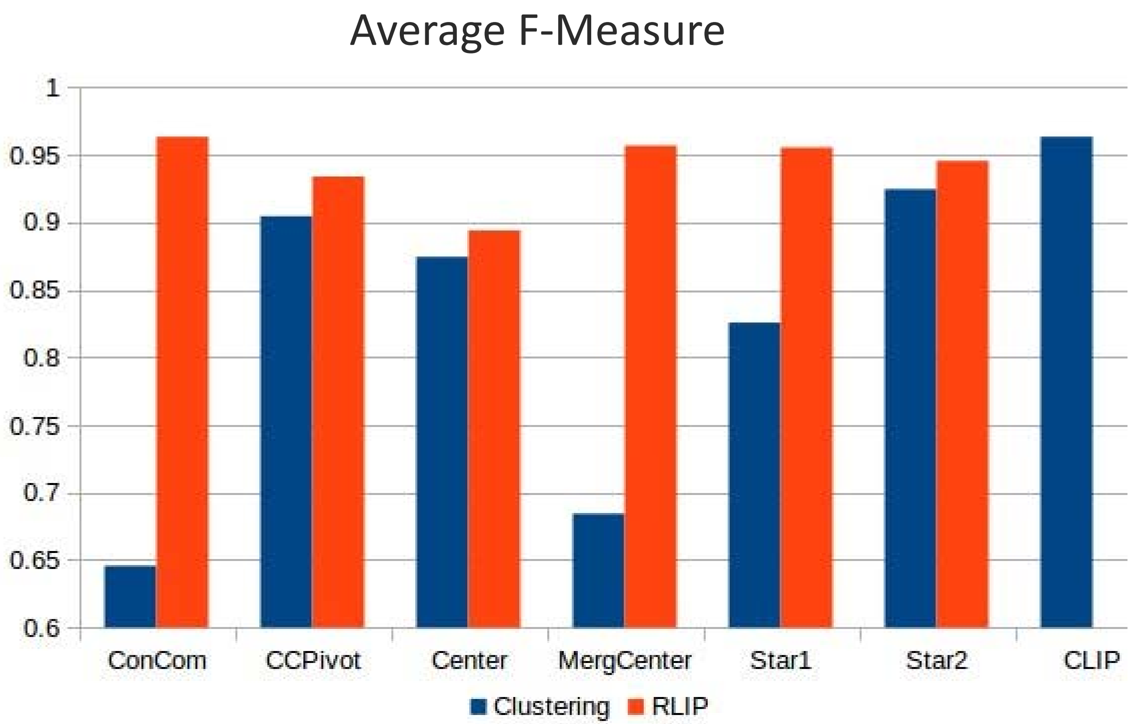
- Geographical domain
- 4 sources



CLIP QUALITY



RLIP QUALITY



RUNTIMES

- Flink cluster of 16 workers
  - 5 party: **69 seconds (~ 1 min)**
  - 10 party: **228 seconds (< 4 min)**

5 sources – 5,000,000 entities

	#workers		
Clustering	4	8	16
ConCom	51	57	55
CCPiv	1530	10008	688
Center	390	208	117
Merge Center	640	349	194
Star-1	288	149	85
Star-2	214	124	67
CLIP	**190**	**101**	**69**

RUNTIMES

5 sources – 5,000,000 entities

method	Run times (sec)			
	Clustering	RLIP		Sum
ConCom	55	69		124
CCPiv	688	46		734
Center	117	46		163
Merge Center	194	49		243
Star-1	85	55	53	193
Star-2	67	61	52	180
CLIP	**69**	-		**69**

SUMMARY

- Parallel execution of ER workflows using the Big Data framework Apache Flink
- ER for multi-source datasets
- Parallel clustering
  - A new clustering approach called **CLIP** (Clustering based on LInk Priority)
  - An approach called **RLIP** (cluster Repair based on LInk Priority)

OUTLOOK

- Improving Famer linking component
- Developing incremental ER strategies
- Scalability and quality test with more number of sources

Marcel Gladbach, *Martin Franke, Ziad Sehili, Erhard Rahm*

PRIVACY-PRESERVING RECORD LINKAGE

4th International Summer School for Big Data and Machine Learning, Leipzig, 05.07.2018

University of Leipzig
Faculty of Mathematics and Computer Science
Institute of Computer Science
Database Group

RECORD LINKAGE (RL)

– finding records from different data sources
  → referring to the same real-world entity *(usually persons or products)*
– typically no global identifiers

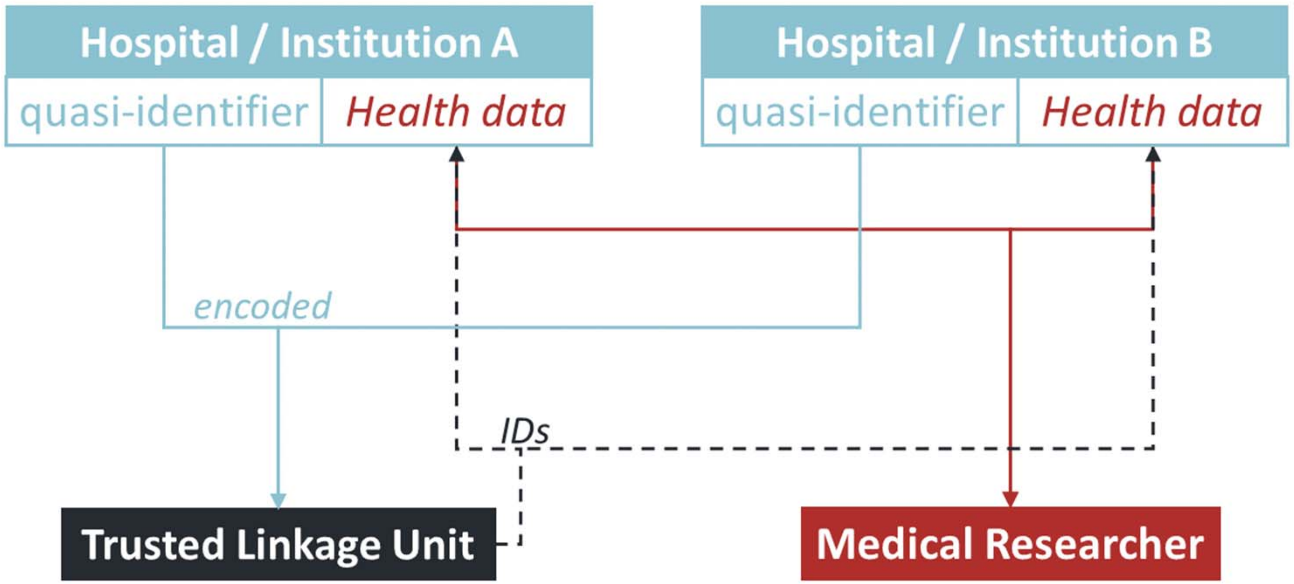
→ **linkage** by comparing *quasi-identifiers (name, address, etc.)*

PID	Last_name	First_name	Age	Address	Sex	Pressure	Stress	Reason
P1209	roberts	peter	41	16 Main St 2617	m	140/90	high	chest pain
P4204	miller	amelia	39	49 Aplecross Road 2415	f	120/80	high	headache
P4894	sieman	jeff	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

ID	Given_name	Surname	DOB	Gender	Address	Loan_type	Balance
6723	peter	robert	20.06.72	M	16 Main Street 2617	Mortgage	230,000
8345	smith	roberts	11.10.79	M	645 Reader Ave 2602	Personal	8,100
9241	amelia	millar	06.01.74	F	49E Applecross Rd 2415	Mortgage	320,750

PRIVACY PRESERVING RECORD LINKAGE (PPRL)

Linkage on encoded data to not reveal the identity of persons in the process



OUTLINE

USE CASES

Medical Domain

data from hospitals, physicians, insurance companies, studies, …

– central registries for certain diseases
– clinical studies to optimize treatments
– social studies

Criminalistics

data from banks, credit card companies, email service providers, authorities, …

– money laundry detection
– detection of criminal online activities

ScaDS DRESDEN LEIPZIG

CHALLENGES

Quality

- support of fuzzy matches
- avoiding false matches (Precision)
- avoiding missing matches (Recall)

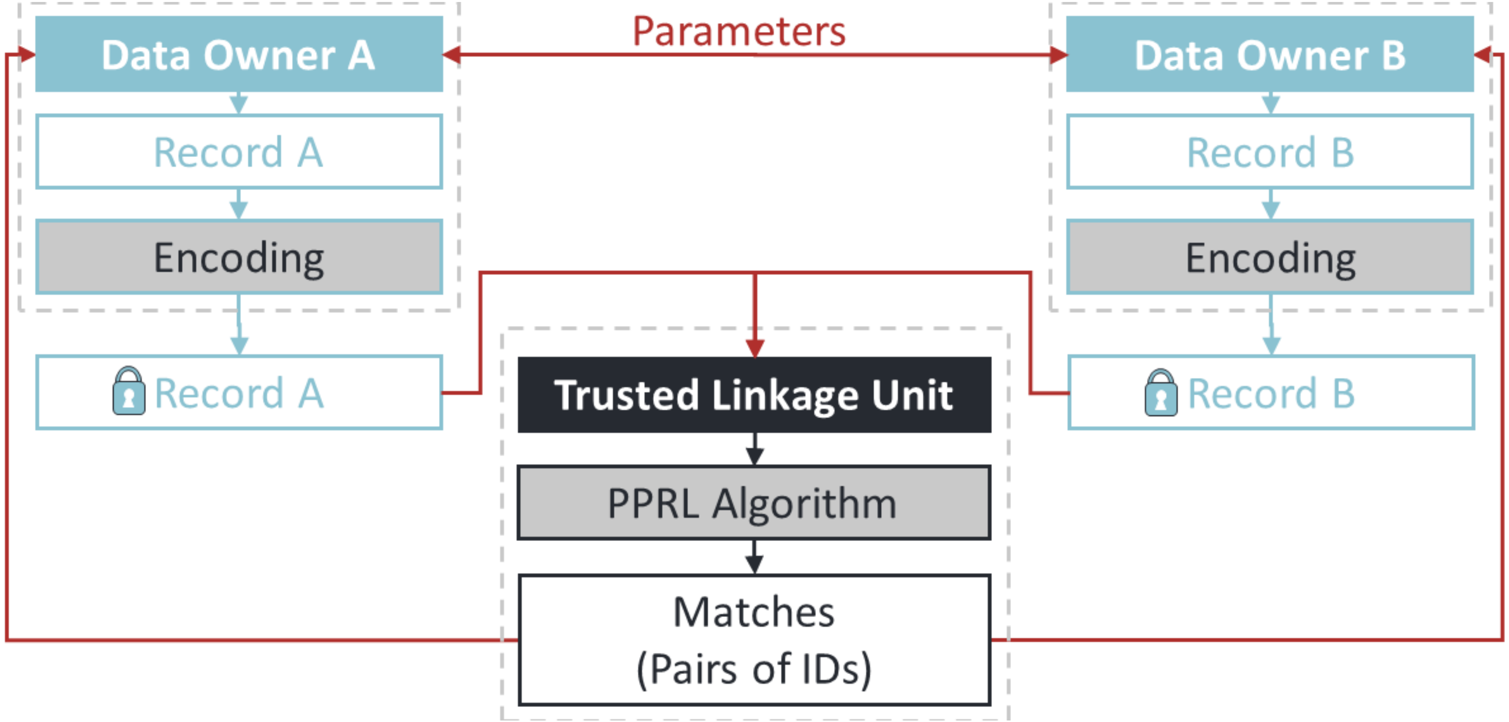
Privacy

- suitable encoding
- organizational structures
- protection against attacks

Performance

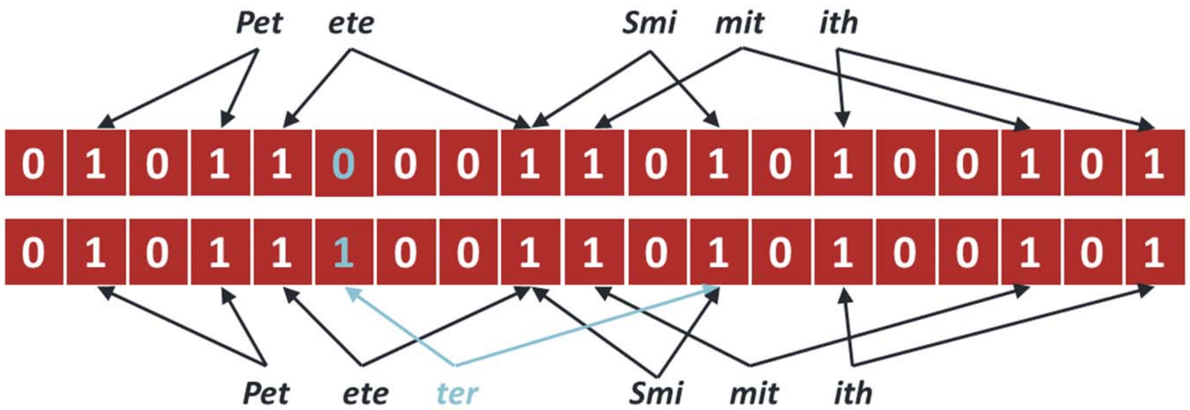
- scalability to large datasets and many parties
- filter and blocking techniques

PROTOCOLS


Parameters
Data Owner A
Record A
Encoding
Record A
Data Owner B
Record B
Encoding
Record B
Trusted Linkage Unit
PPRL Algorithm
Matches
(Pairs of IDs)

ENCODING

- *quasi-identifiers* are tokenized into a set of ***q-grams***
- ***q-grams*** are mapped with ***k hash functions*** into a bit vector with fixed length L: **Bloom filter**



typical parameters: q=3; L=1000; k=20

MATCHING

- **distance Function**, e.g. *Hamming distance*

  $d_h(a,b) = |a \vee b| - |a \ \wedge b| = |a\ XOR\ b|$ for bit vectors

- **similarity Function**, e.g. *Jaccard similarity*

  $$sim_j(a,b) = \frac{|a \wedge b|}{|a \vee b|}$$

a	0	1	0	1	1	0	0	0	1	1	0	1	0	1	0	0	1	0	1
b	0	1	0	1	1	1	0	0	1	1	0	1	0	1	0	0	1	0	1

$$d_h(a,b) = 10 - 9 = 1$$

$$sim_j(a,b) = \frac{9}{10} = 0.9$$

- pairs of records with similarity above a given threshold $t \in [0,1]$ are considered as matches

PERFORMANCE

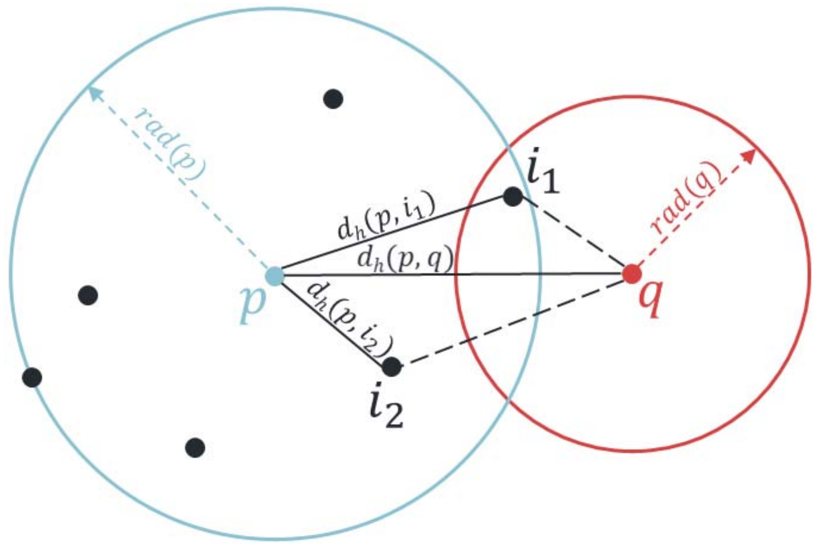
2 parties: inherent **quadratic complexity** of underlying linkage problem

Solutions:

- filter technique: **Metric Space approach**
- blocking technique: **Locality Sensitive Hashing (LSH)**
- **parallel PPRL**

METRIC SPACE

- similarity join with threshold t ∈ [0,1] (same results as nested loop)
- filter technique using pivot elements
- utilizing triangle inequality to reduce search space
- distributed implementation using FLINK

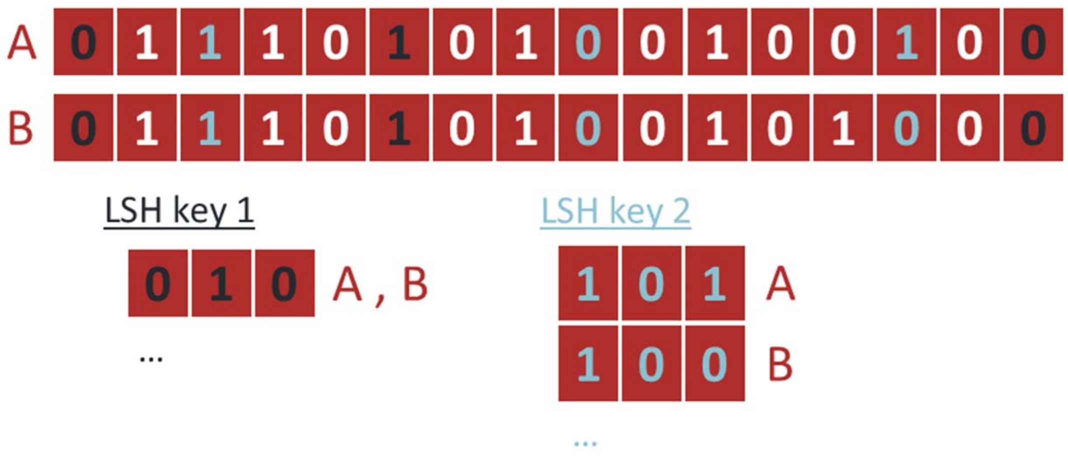


- usually saves > 95% of comparisons compared to nested loop

[research by Ziad Sehili and Marcel Gladbach]

LOCALITY-SENSITIVE HASHING (LSH)

- probabilistic blocking using locality sensitive hash functions for dimensionality reduction
- concatenation of hash values as blocking key
- multiple keys to deal with dirty data

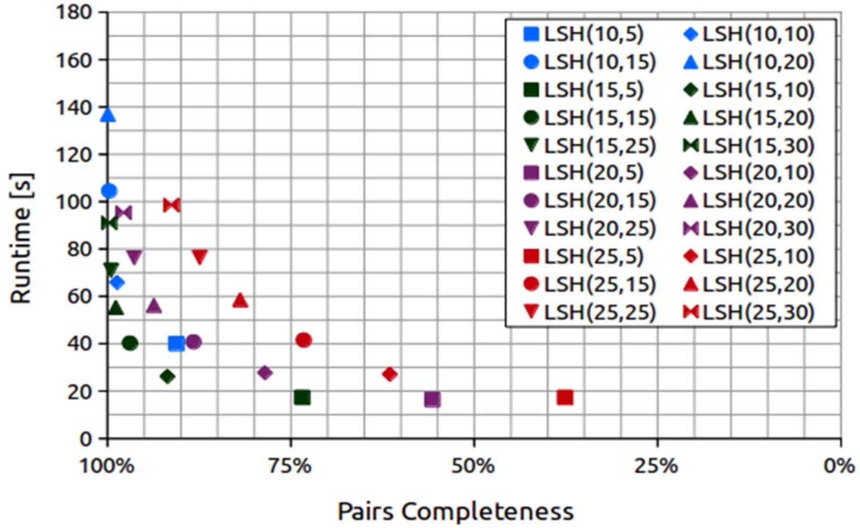


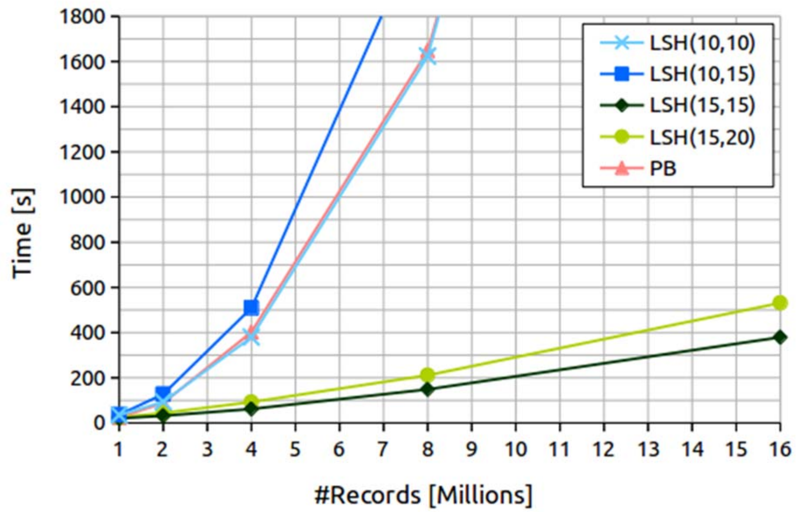
- loss of matching pairs (regarding t) during blocking possible
- but much faster runtimes than filtering techniques

[research by Martin Franke]

DISTRIBUTED LSH

– example results (1 million synthetic records 800K-200K on 16-node cluster with FLINK implementation):





[research by Martin Franke]

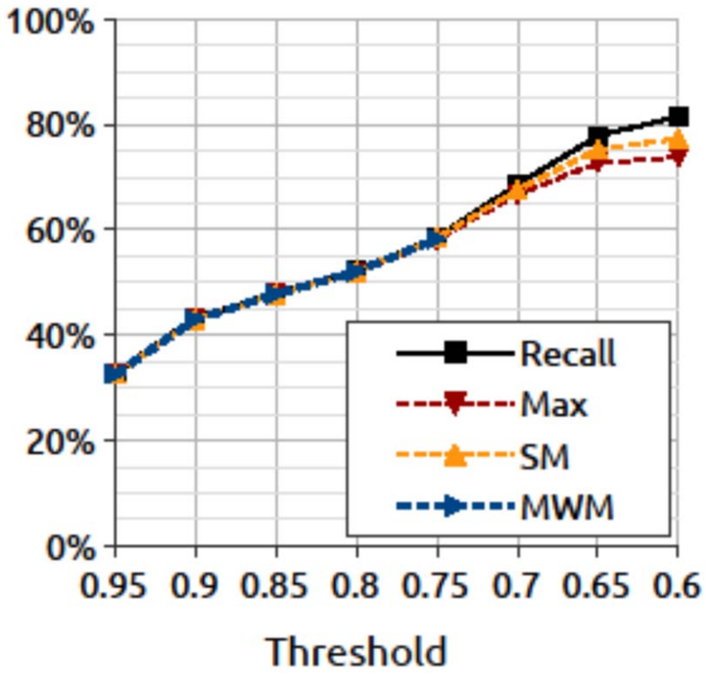
POSTPROCESSING

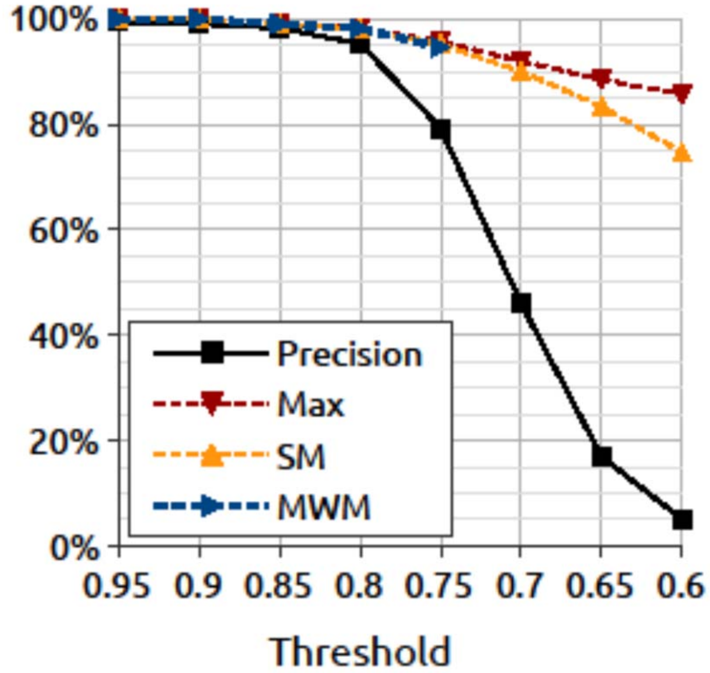
- linkage results often contain multi-links (**1:n**)
- but with assumed deduplicated databases: only **1:1** links are expected

- **post-processing** for multi-link cleaning
- evaluated methods:
  - Best Match Selection Strategy (Max1-both)
  - Stable Matching (SM)
  - Maximum Weight Matching (MWM) = Hungarian Algorithm

[research by Martin Franke]

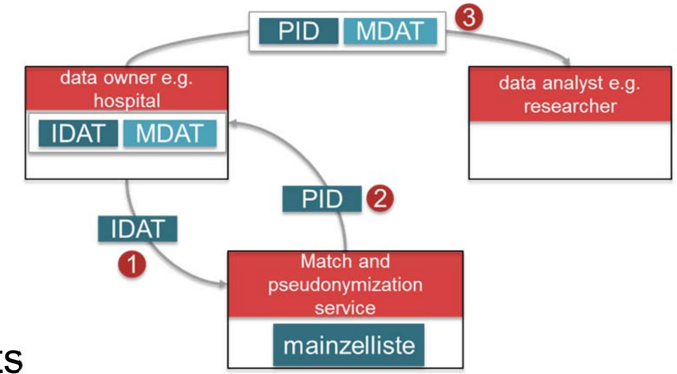
POSTPROCESSING





[research by Martin Franke]

MAINZELLISTE

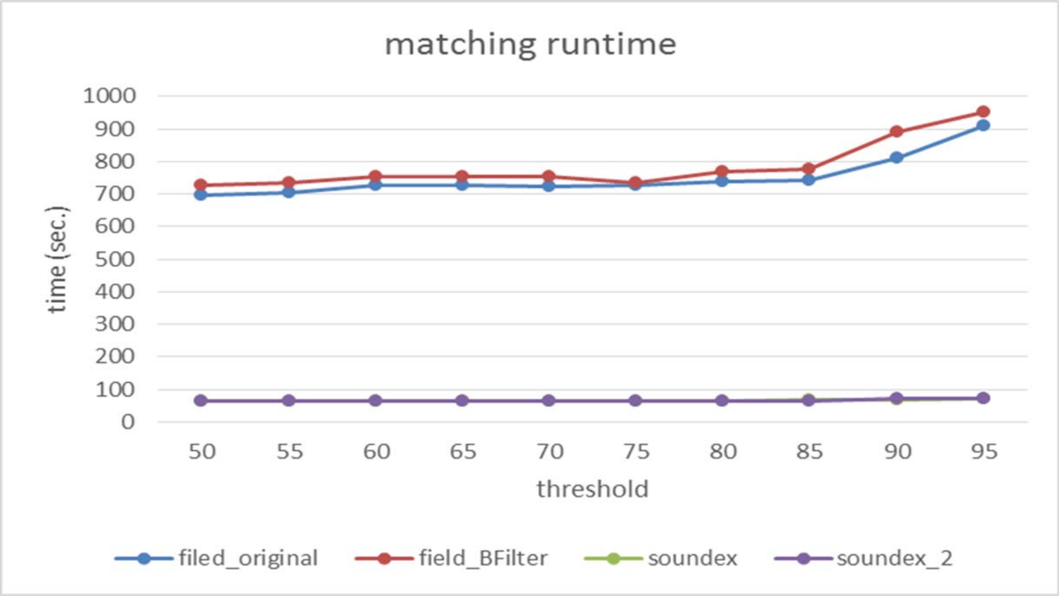


- web-based **pseudonymization service**
- centralized trusty unit
  - management of patients records
  - creation of non-descriptive PID for patients
  - RL on original fields or Bloom filters to assign same PID to same patient
- goal: supply users (e.g. researchers in health care) with data which underlie privacy policies and possibly spread over several institutions
- problem: long runtime for linkage due to the lack of a blocking strategy

[research by Ziad Sehili]

EVALUATION OF MAINZELLISTE

– evaluation of quality and runtime with different synthetic datasets with German names



simple Soundex blocking
(equality on first or last name)

– runtime improvement of a factor 10 to 30
– without reducing result quality

implementation of LSH planned

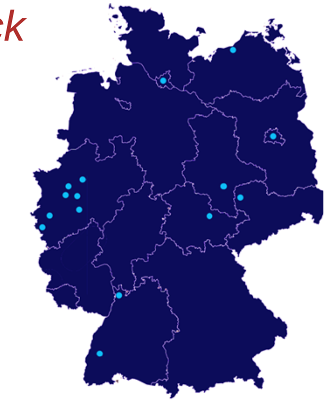
[research by Ziad Sehili]

THE SMITH CONSORTIUM

smith
Smart Medical Information Technology for Healthcare

ScaDS DRESDEN LEIPZIG

- part of the **Medical Informatics Initiative** Germany (BMBF funded)
- initiated by university hospitals *Leipzig, Jena, Aachen*
- complemented by UKs *Halle, Hamburg, Essen, Bonn, Düsseldorf, Rostock*
- industrial partners: *SAP, März / Tiani, Fraunhofer ISST and more*
- started with 4-year development and networking phase in 01/2018



Goals

- advancing and harmonizing IT infrastructure in participating sites
- enable data exchange for healthcare and for research
- establishment of data integration centers (**DIC**s) to support structured medical and study documentations in clinical and research IT

[research by Marcel Gladbach]

PPRL WITHIN SMITH

- apply PPRL for linkage and data exchange between different sites (DICs)
- development of PPRL components for ID Management
  - onsite Coding Service and central Matching Service

Features

- project-specific generation of bit vectors possible
  - flexibility regarding quasi-identifiers
- continuous matching
  - initial matching between two DICs
  - matching new patients without linkage of complete source
- multiparty matching
  - building clusters of matches of more than two sources

[research by Marcel Gladbach]

CONCLUSION

Research results

- performance improvements
  - distributed filter and blocking approaches
- postprocessing methods
  - resolution of 1:n matches
- multiparty approaches
  - linking of more than two sources

Application Projects

- Mainzelliste
- SMITH consortium

need for PPRL in praxis

OUTLOOK

Future work: transfer PPRL approaches into practical application

- considerable aspects for PPRL in practice
  - **continuous / incremental matching**
  - **multiparty matching**
- focus on **privacy** and **quality** of PPRL
- develop a **PPRL toolbox**
  - use in applications
  - comparative evaluation

REFERENCES

- P. Christen: *Data Matching*. Springer, 2012
- X.L. Dong, D. Srivastava: *Big Data Integration*. Synthesis Lectures on Data Management, Morgan & Claypool 2015
- H. Köpcke, A. Thor, E. Rahm: *Learning-based approaches for matching web data entities*. IEEE Internet Computing 14(4), 2010
- H. Köpcke, A. Thor, E. Rahm: *Evaluation of entity resolution approaches on real-world match problems*. Proc. 36th Intl. Conference on Very Large Databases (VLDB) / PVLDB 3(1), 2010
- 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)
- L. Kolb, A. Thor, E. Rahm: *Dedoop: Efficient Deduplication with Hadoop*. PVLDB 5(12), 2012
- L. Kolb, A. Thor, E. Rahm: *Load Balancing for MapReduce-based Entity Resolution*. ICDE 2012: 618-629
- M. Nentwig, M. Hartung, A. Ngonga, E. Rahm: *A Survey of Current Link Discovery Frameworks*. Semantic Web Journal, 2016
- E. Rahm, H. H. Do: *Data Cleaning: Problems and Current Approaches*. IEEE Techn. Bulletin on Data Engineering, 2000
- E. Rahm: *Towards large-scale schema and ontology matching*. In: Schema Matching and Mapping, Springer 2011

REFERENCES HOLISTIC DATA INTEGRATION

- M. Nentwig, A. Groß, E. Rahm: *Holistic Entity Clustering for Linked Data*. IEEE Int. Conf. on Data Mining Workshop, ICDMW 2016 2016
- M. Nentwig, A. Groß, Anika; M. Möller, E. Rahm: *Distributed Holistic Clustering on Linked Data.* Proc. OTM 2017 - LNCS 10574, pp 371-382
- E. Rahm: *The case for holistic data integration.* Proc. ADBIS, 2016
- 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**)

REFERENCES PPRL

- M. Franke, Z. Sehili, E. Rahm: *Parallel Privacy-Preserving Record Linkage using LSH-based blocking*. Proc. 3rd Int. Conf. on Internet of Things, Big Data and Security (IoTBDS), pp. 195-203, 2018
- M. Gladbach, Z. Sehili, T. Kudraß, P. Christen, E. Rahm: *Distributed Privacy-Preserving Record Linkage using Pivot-based Filter Techniques*. Proc. IEEE Inte. Conf. on Data Engineering Workshops (ICDE-W), pp. 33-38, 2018
- Z. Sehili, L. Kolb, C. Borgs, R. Schnell, E. Rahm: Privacy Preserving Record Linkage with PPJoin. Proc. 16th Conf. on Databases for Business, Technology and Web (BTW), 2015
- Z. Sehili, E. Rahm: Speeding up Privacy Preserving Record Linkage for Metric Space Similarity Measures. Datenbankspektrum 16, pp. 227-236, 11/2016
- D. Vatsalan, P. Christen, E. Rahm: *Scalable privacy-preserving linking of multiple databases using Counting Bloom filters*. Proc ICDM workshop on Privacy and Discrimination in Data Mining (PDDM), 2016
- D. Vatsalan, Z. Sehili, P. Christen, E. Rahm, Erhard: *Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges*. In: Handbook of Big Data Technologies (eds.: A. Zomaya, S. Sakr) , Springer 2017