UNIVERSITÄT
LEIPZIG

ScaDS
DRESDEN LEIPZIG

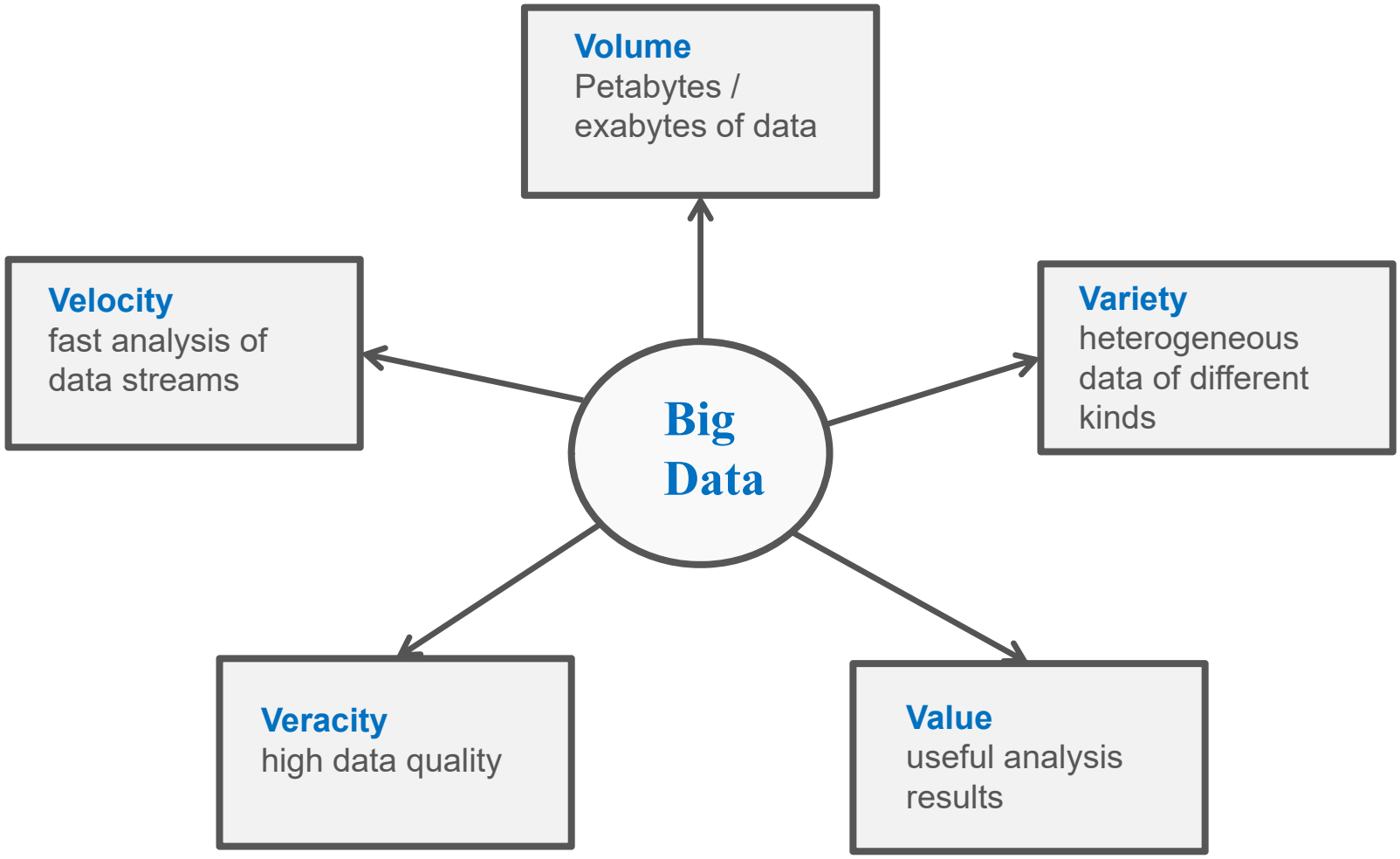
BIG DATA INTEGRATION RESEARCH AT SCADS

Erhard Rahm
Eric Peukert
Alieh Saeedi
Marcel Gladbach

ScaDS DRESDEN LEIPZIG

BIG DATA CHALLENGES

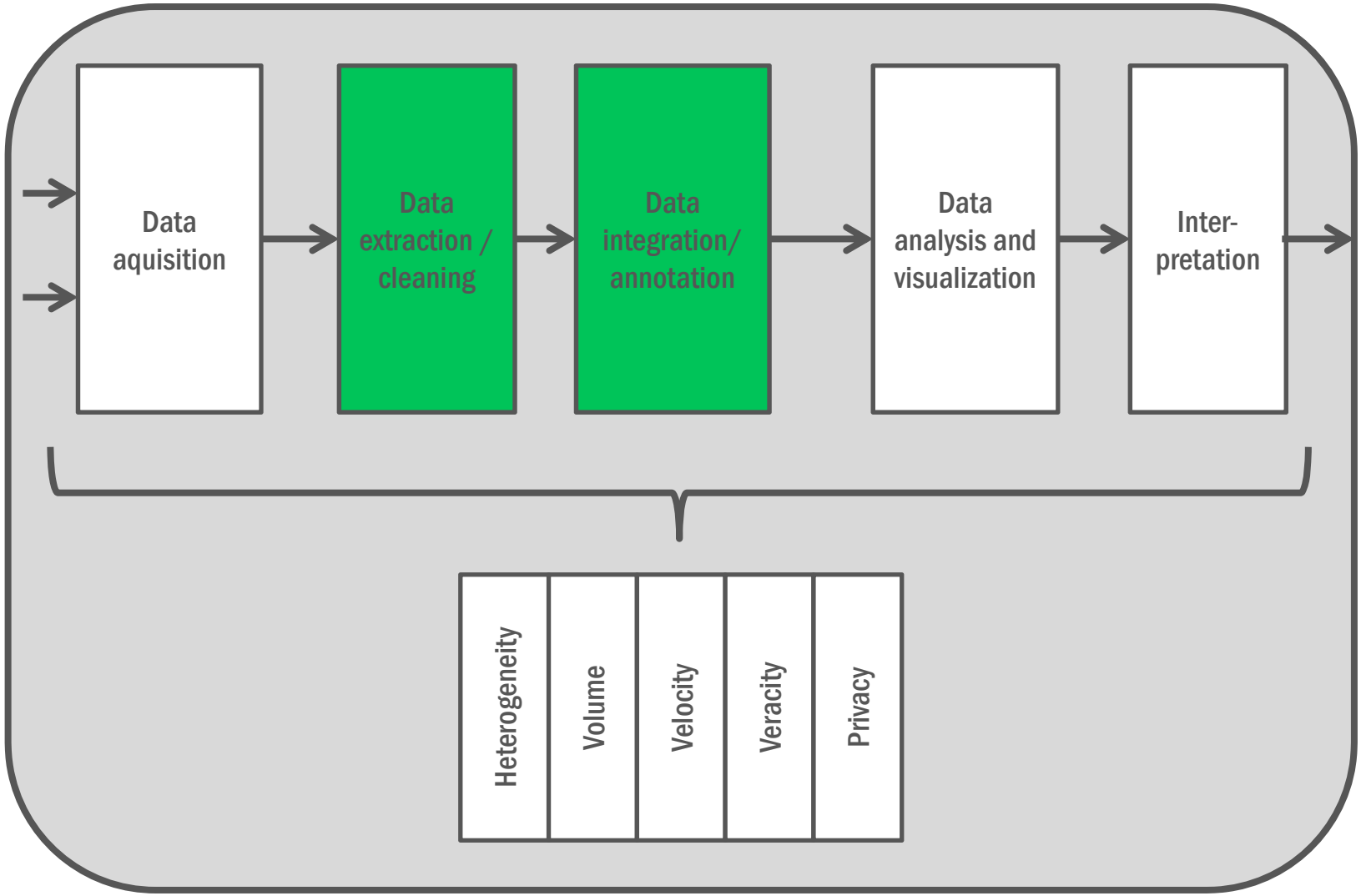
UNIVERSITÄT LEIPZIG



ScaDS DRESDEN LEIPZIG

BIG DATA ANALYSIS PIPELINE

UNIVERSITÄT LEIPZIG



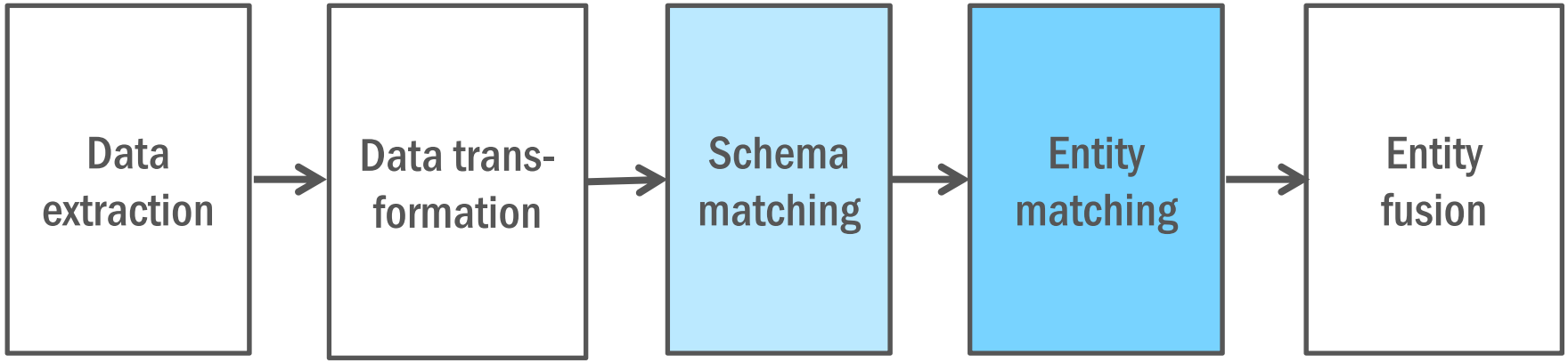
ScaDS DRESDEN LEIPZIG

DATA INTEGRATION

UNIVERSITÄT LEIPZIG

- 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, ...
★★★★½ 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

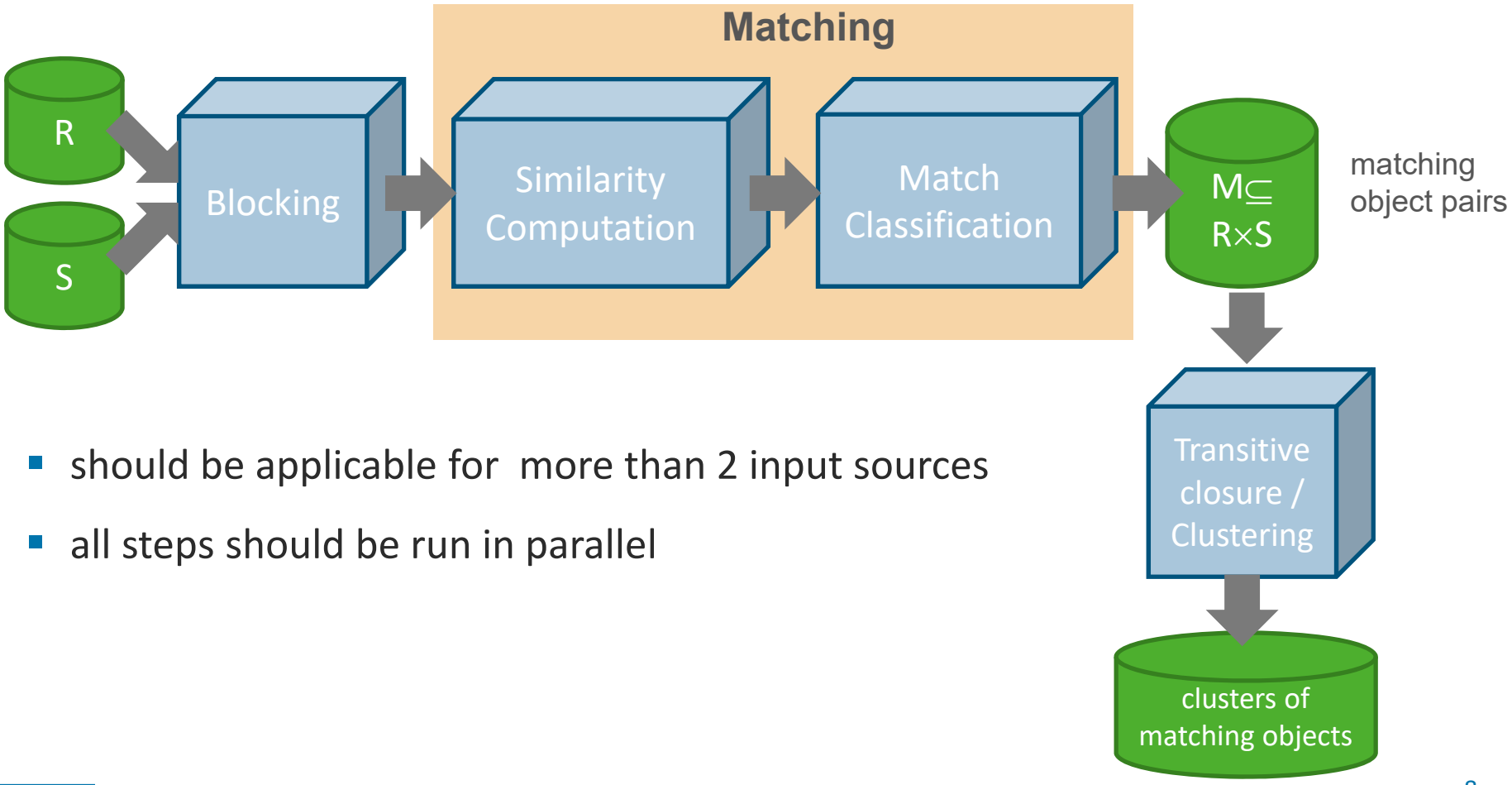
Canon VIXIA HF S100 Flash Memory Camcorder
***Canon Video HF S100 Instant Rebate Receive $200 with your purchase of a new Canon VIXIA HF S100 Flash Memory Camcorder. (Price above includes $200 ....
Add to Shopping List

$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
★★½☆☆ 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

ScaDS DRESDEN LEIPZIG

AGENDA

UNIVERSITÄT LEIPZIG

- 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)

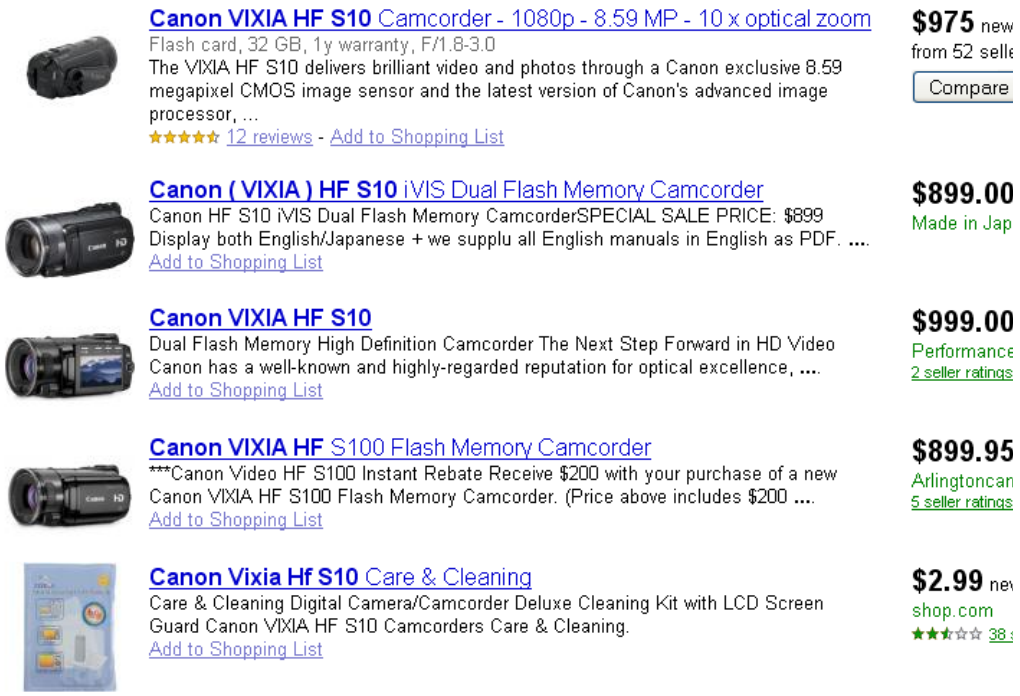
ScaDS DRESDEN LEIPZIG

BIG DATA INTEGRATION USE CASE

UNIVERSITÄT LEIPZIG

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



SOLUTION APPROACH

Input:

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

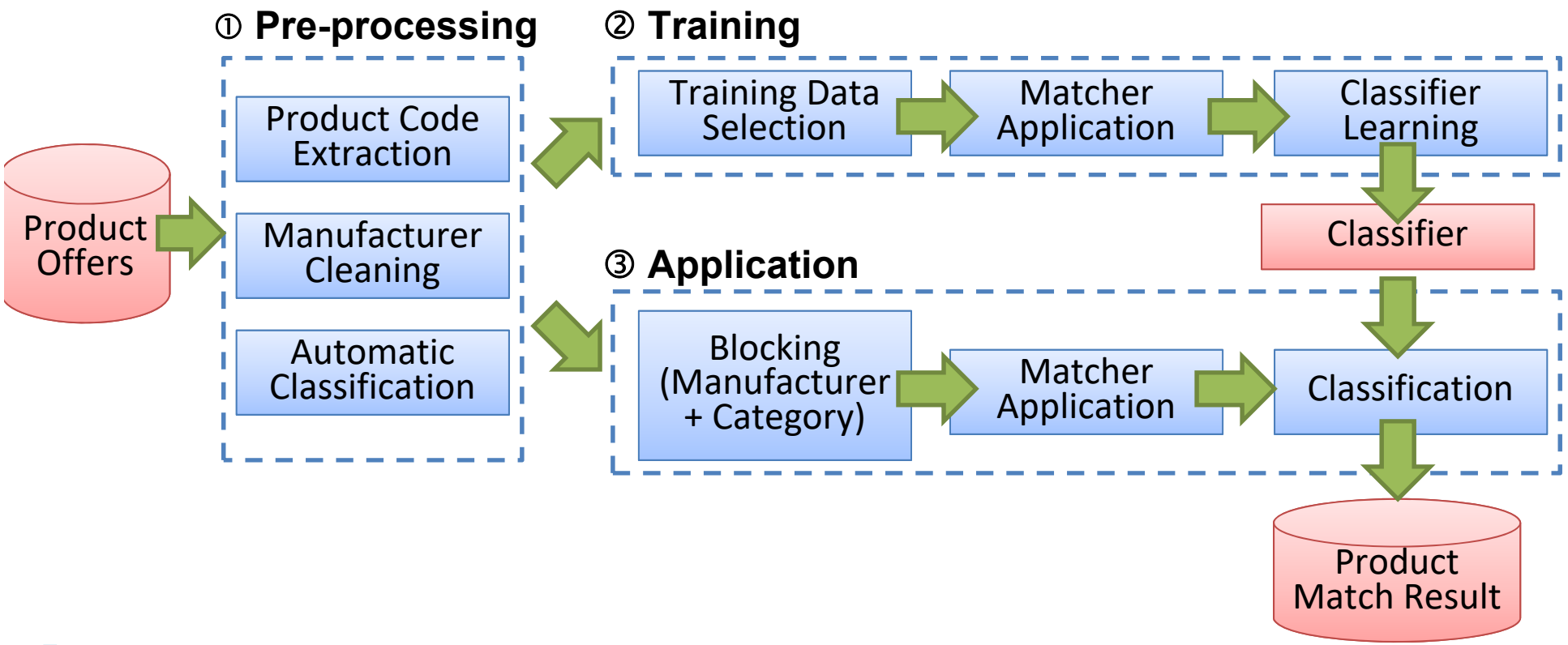
Preprocessing/ Data Cleaning:

- 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

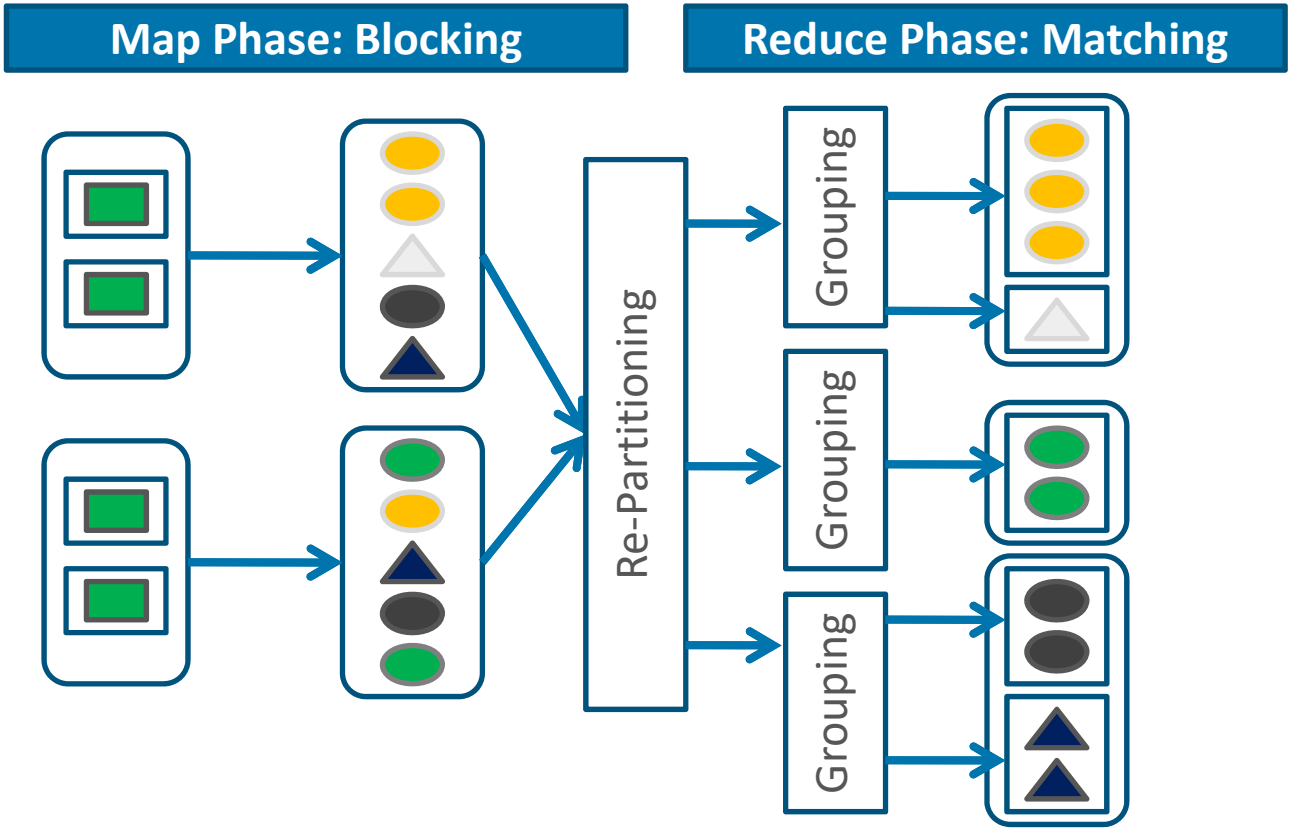
HOW TO SPEED UP ENTITY MATCHING?

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

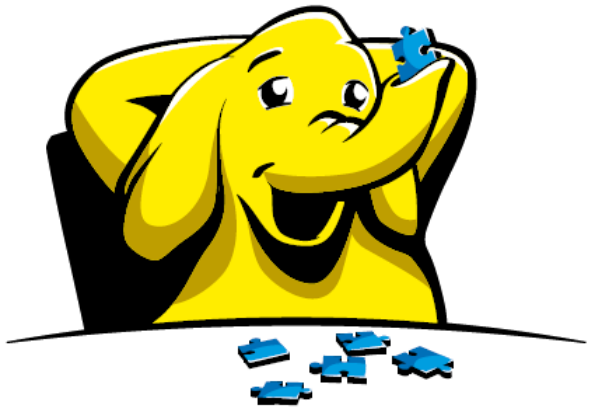
- **Parallelization**
  - split match computation in sub-tasks to be executed in parallel
  - exploitation of Big Data infrastructures such as Hadoop Map/Reduce, Apache Spark or Apache Flink

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

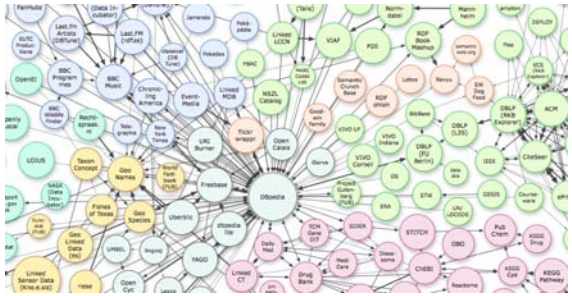
L. Kolb, A. Thor, E. Rahm: *Dedoop: Efficient Deduplication with Hadoop*. PVLDB 5(12), 2012

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

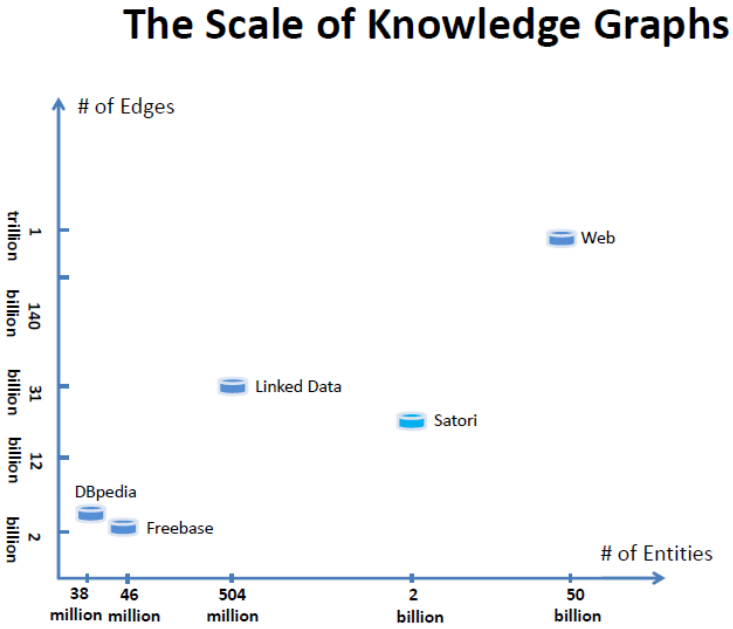
[PDF] **Data cleaning**: Problems and current approaches
E Rahm, HH Do - IEEE **Data** Eng. Bull., 2000 - academia.edu
We classify **data** quality problems that are addressed by **data cleaning** and provide an overview of the main solution approaches. **Data cleaning** is especially required when integrating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

Google | Shopping

€650,96 from 50+ shops
Apple iPhone 8 - 64 GB - Mattrot - Ohne SIM-Lock
★★★★½ (6.574)
Free shipping

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 …

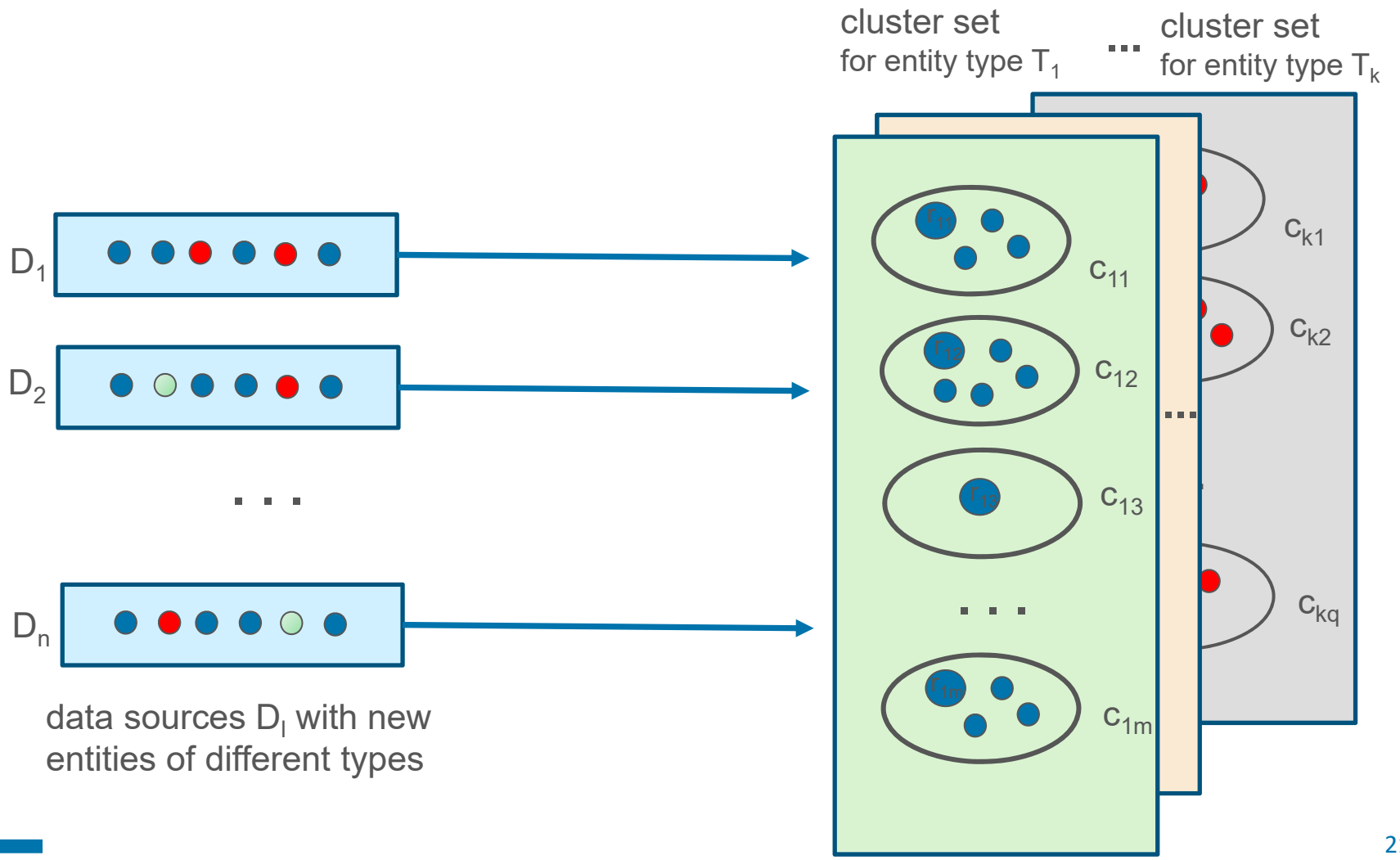


Shao, Li, Ma (Microsoft Asia): Distributed Real-Time Knowledge Graph Serving (slides, 2015)

HOLISTIC ENTITY RESOLUTION

- 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

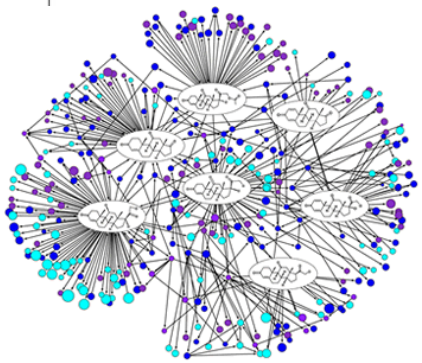
(INCREMENTAL) CLUSTERING-BASED ENTITY RESOLUTION



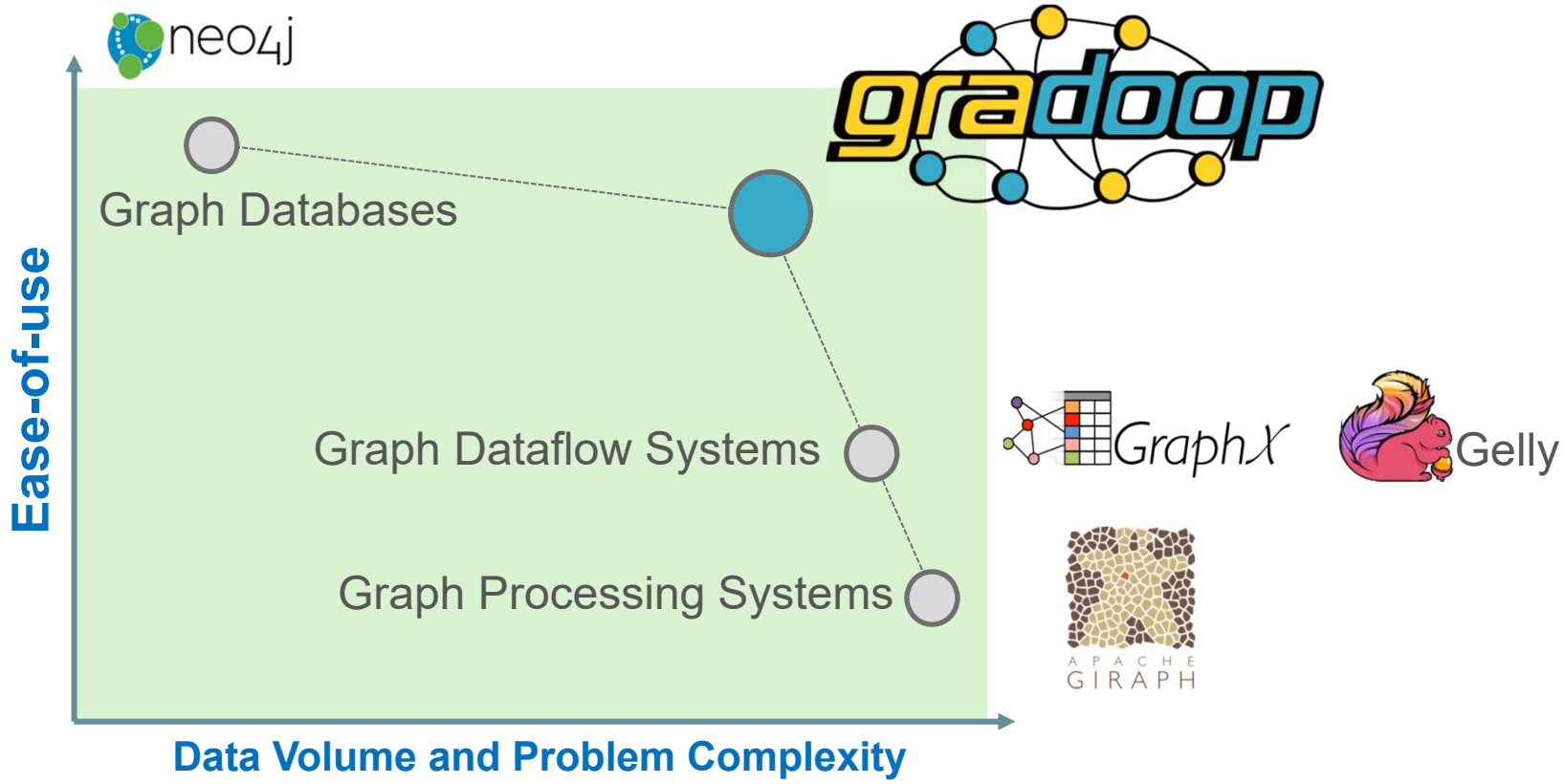
ScaDS DRESDEN LEIPZIG

UNIVERSITÄT LEIPZIG

GRAPH ANALYTICS



- 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

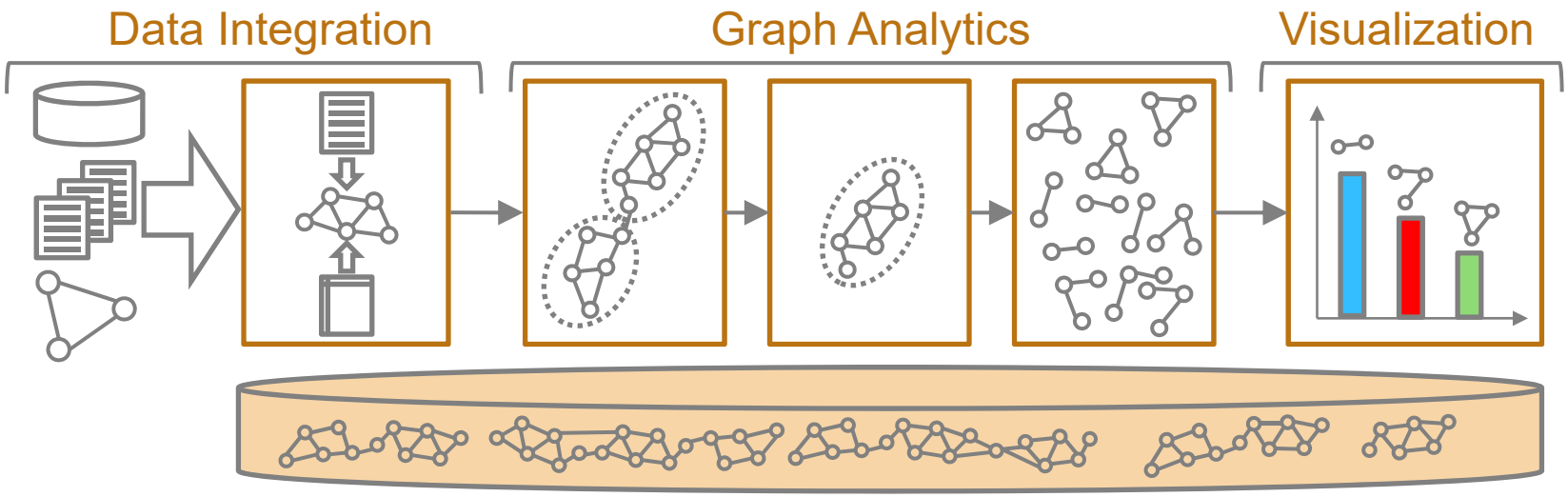
ScaDS DRESDEN LEIPZIG

GRADOOP CHARACTERISTICS

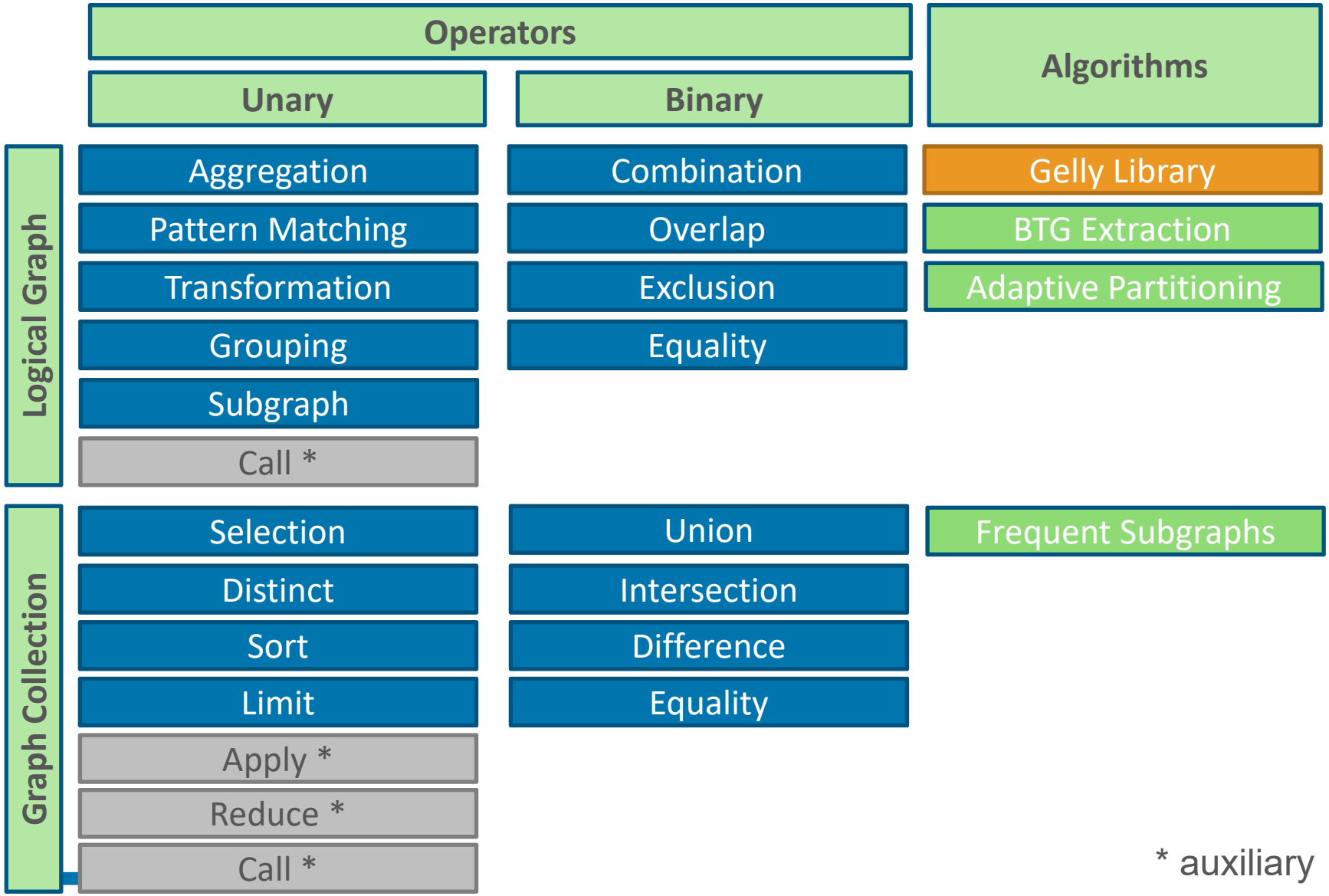
UNIVERSITÄT LEIPZIG

- 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**

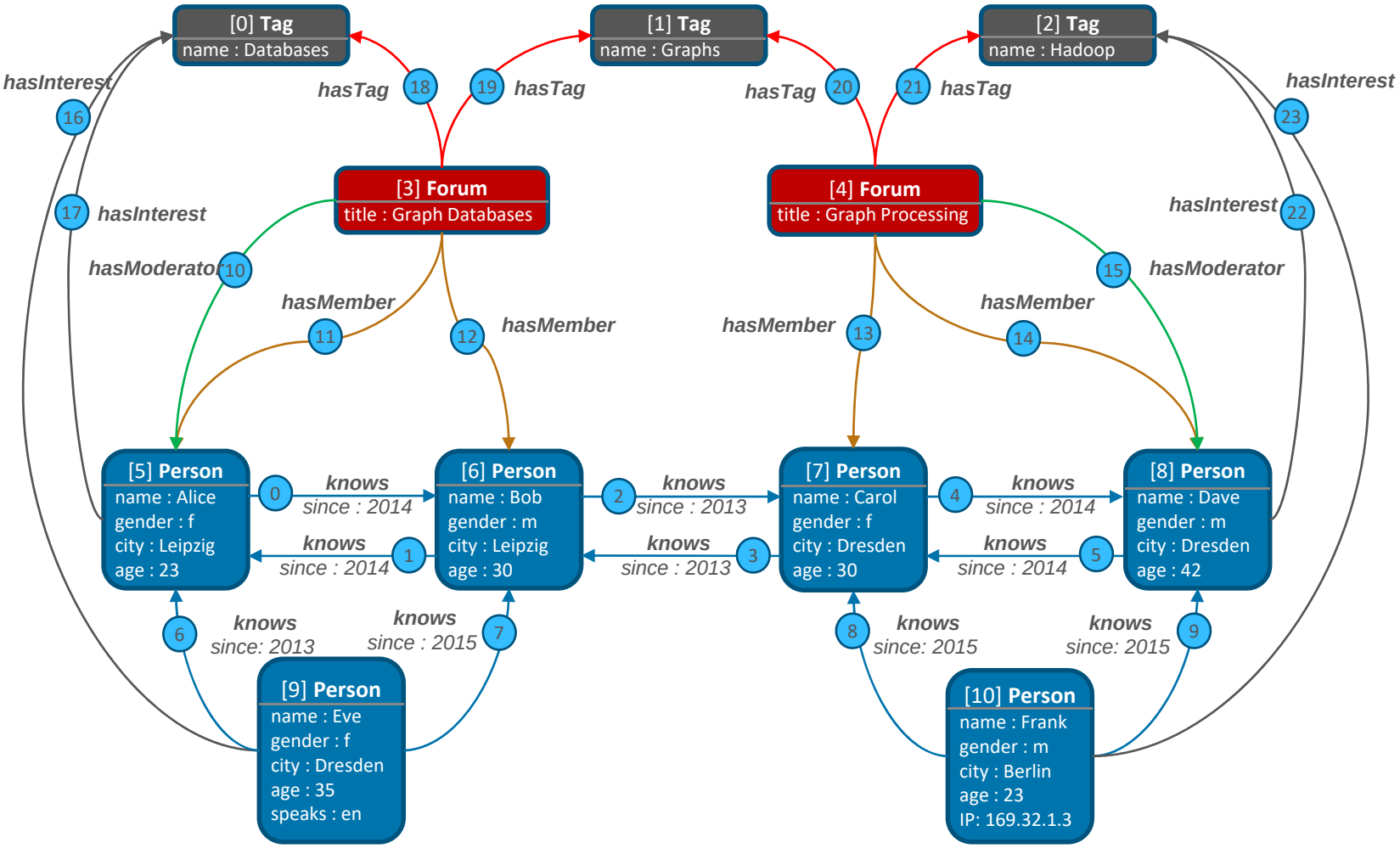

Operators
Unary
Binary
Algorithms
Logical Graph
Aggregation
Pattern Matching
Transformation
Grouping
Subgraph
Call *
Combination
Overlap
Exclusion
Equality
Gelly Library
BTG Extraction
Adaptive Partitioning
Graph Collection
Selection
Distinct
Sort
Limit
Apply *
Reduce *
Call *
Union
Intersection
Difference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

* auxiliary

ScaDS DRESDEN LEIPZIG

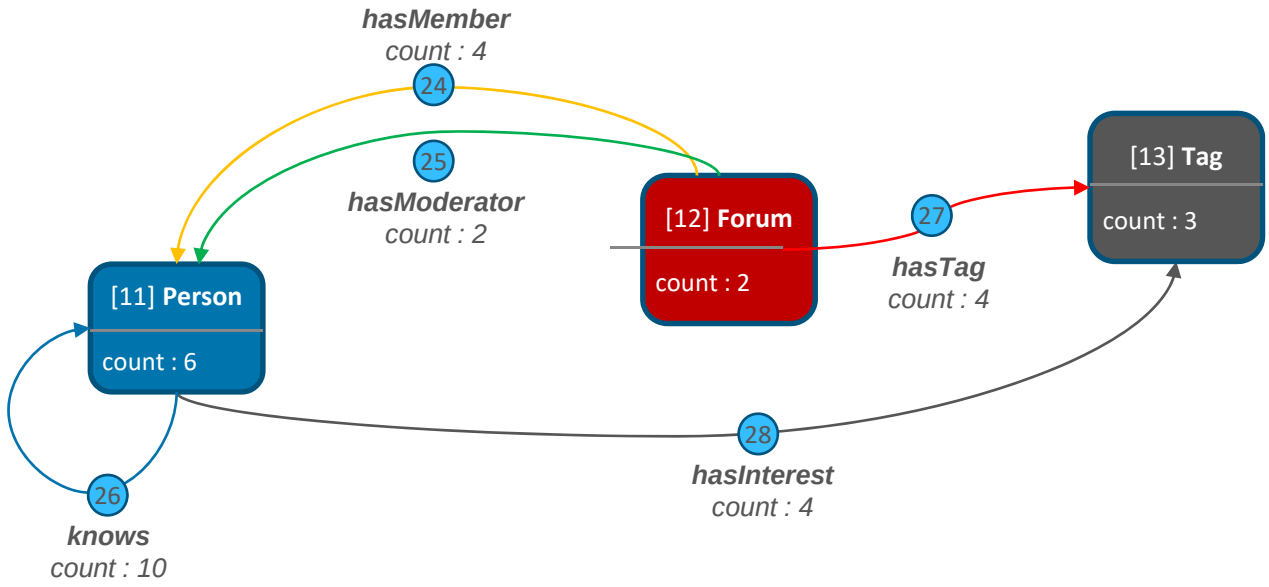
SAMPLE GRAPH

UNIVERSITÄT LEIPZIG



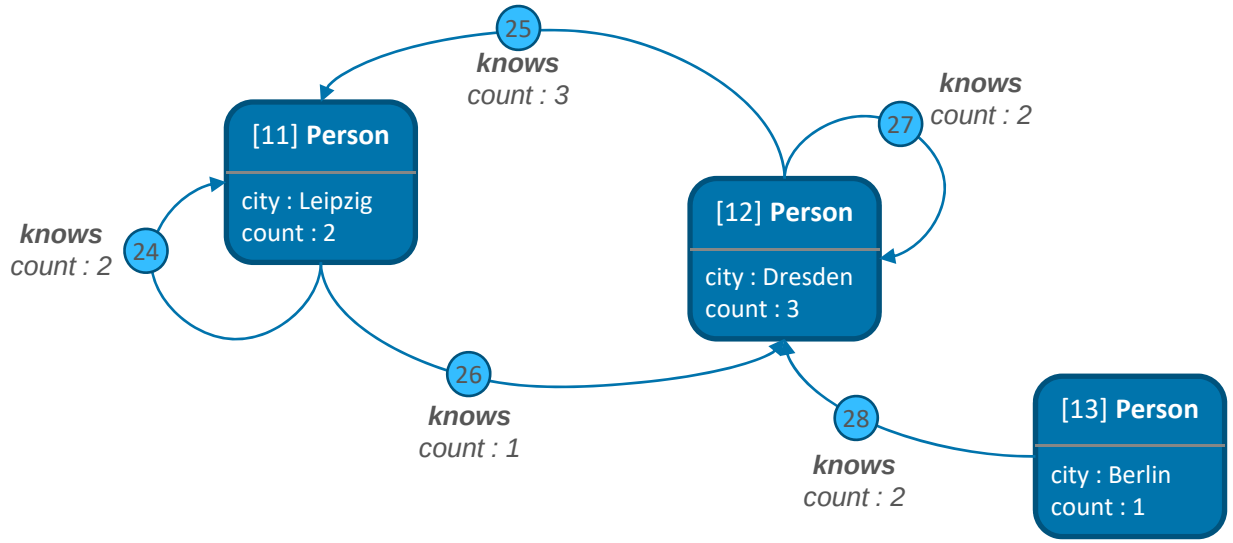
GROUPING: TYPE LEVEL *(SCHEMA GRAPH)*

```
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()])
```



ScaDS DRESDEN LEIPZIG

GRAPH-BASED DATA INTEGRATION

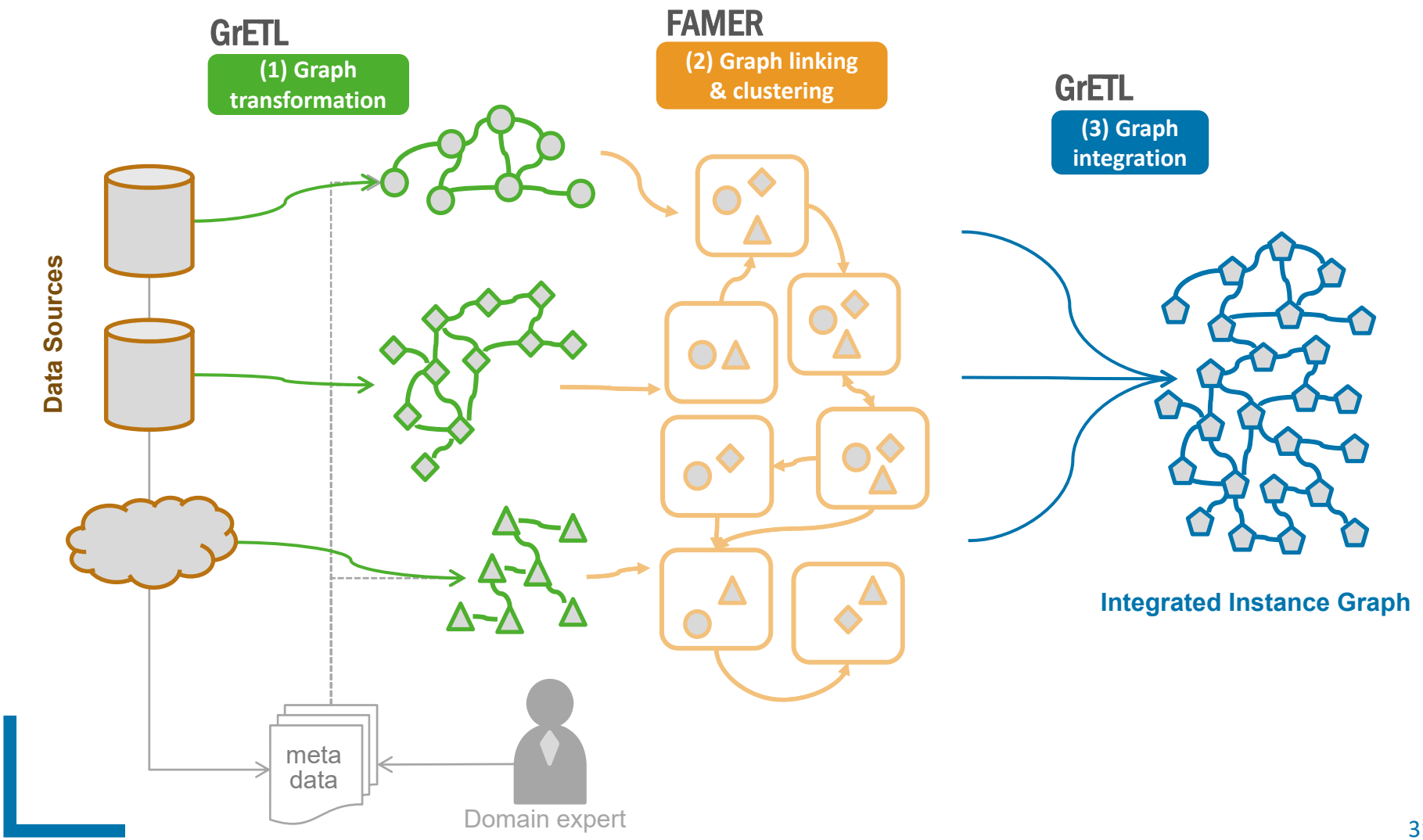
UNIVERSITÄT LEIPZIG

- 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

ScaDS DRESDEN LEIPZIG

GRAPH DATA INTEGRATION WORKFLOW

UNIVERSITÄT LEIPZIG



GRAPH TRANSFORMATION IN GRETL/GRADOOP

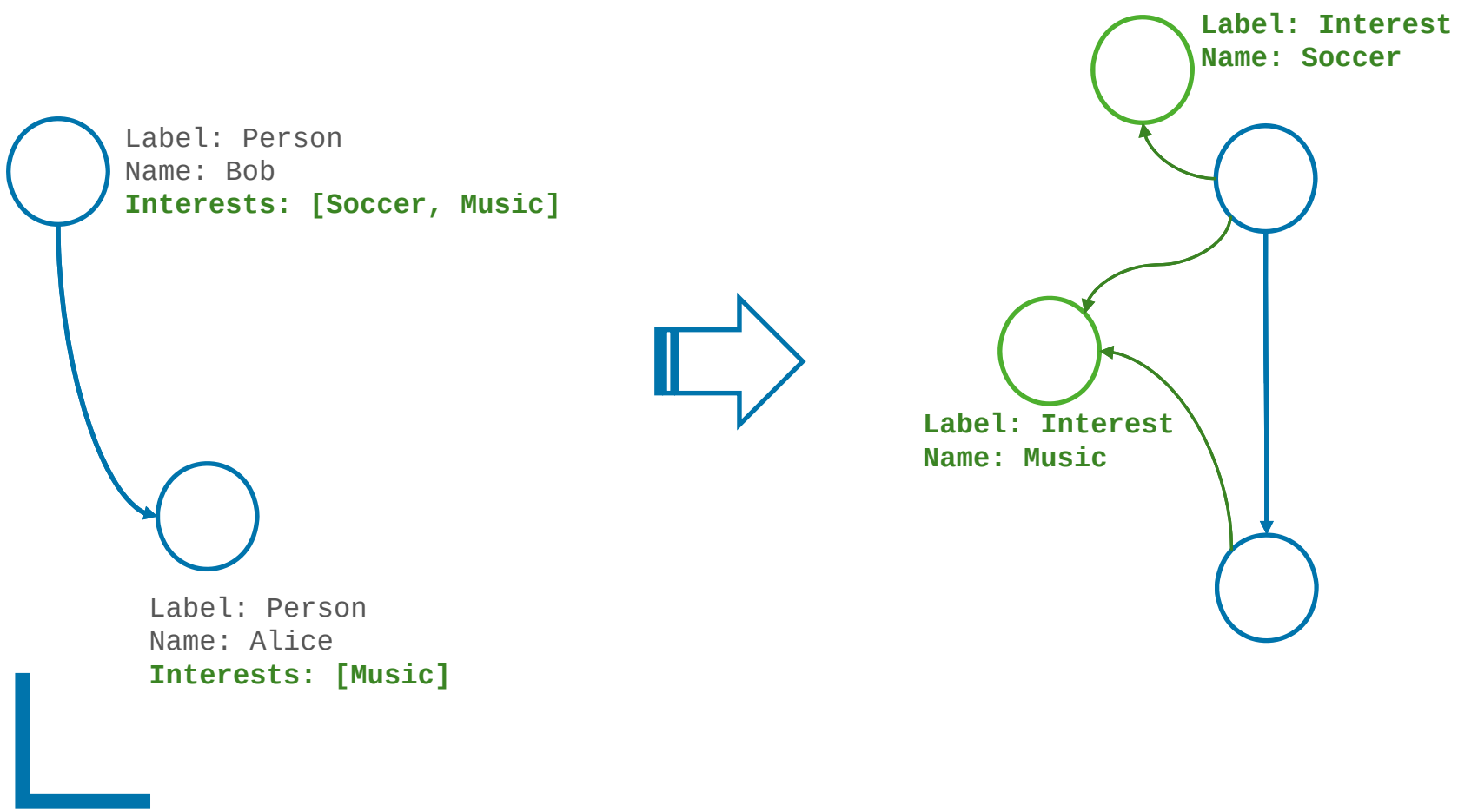
Structural Transformations

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

PROPERTY TO VERTEX

Pseudocode:

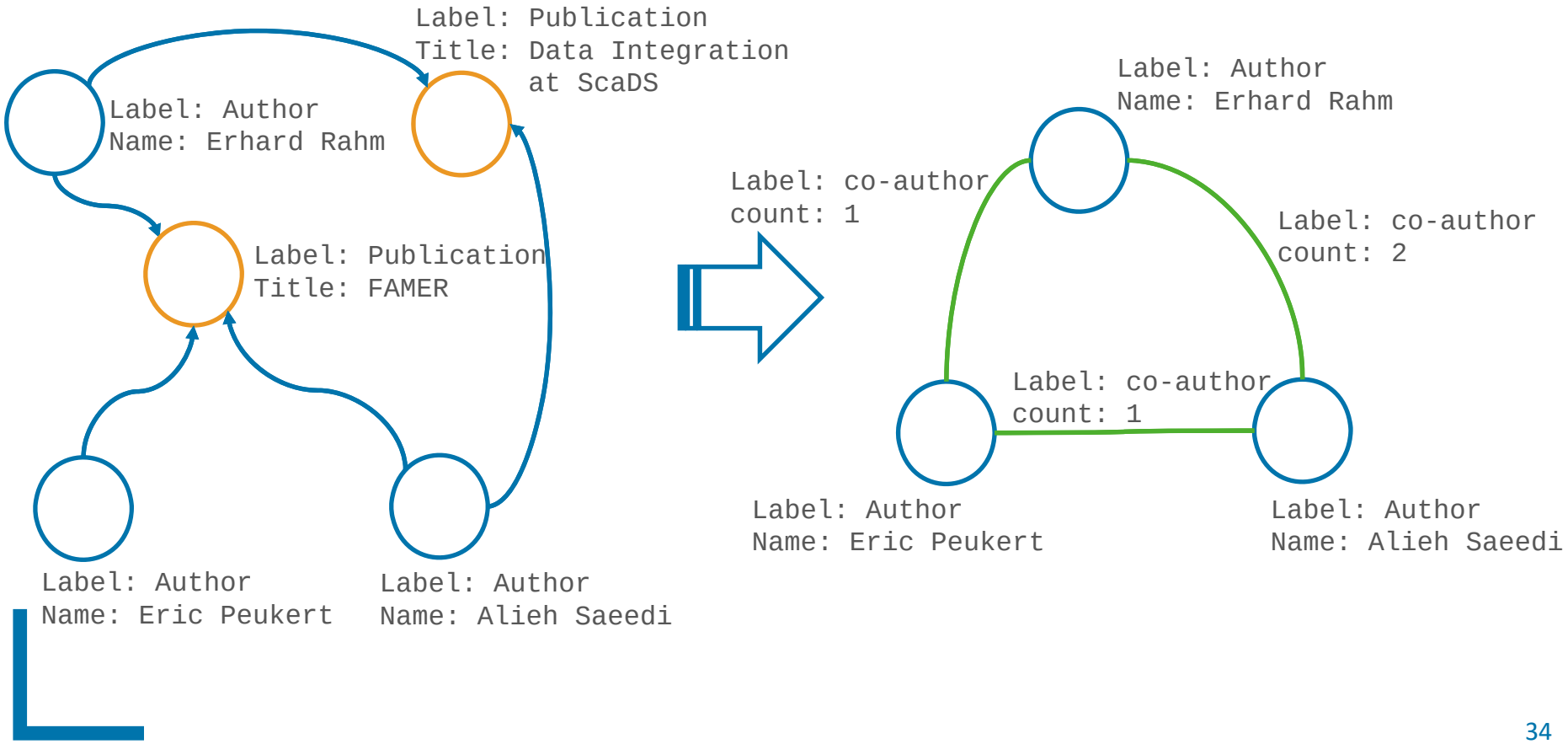
```
inputGraph
  .extractProperty(Person, Interests, Interest)
```



EDGES BY NEIGHBORHOOD

Pseudocode:

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

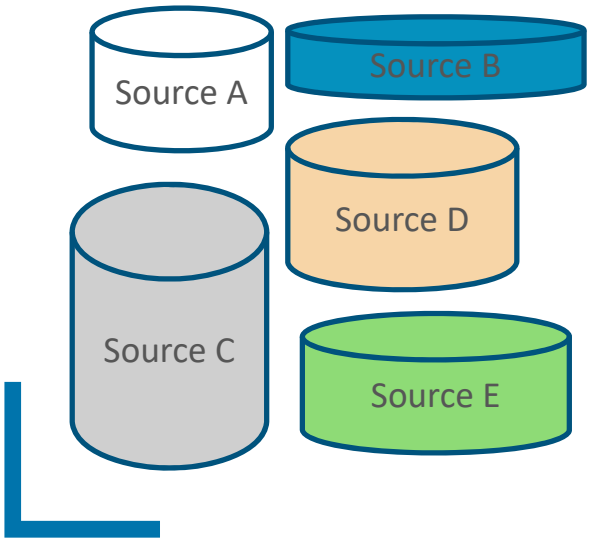


ENTITY LINKING AND CLUSTERING

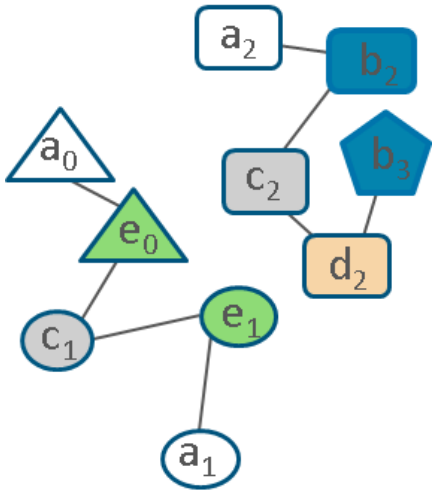
FAMER: scalable linking & clustering for many sources

see talk of Alieh Saeedi

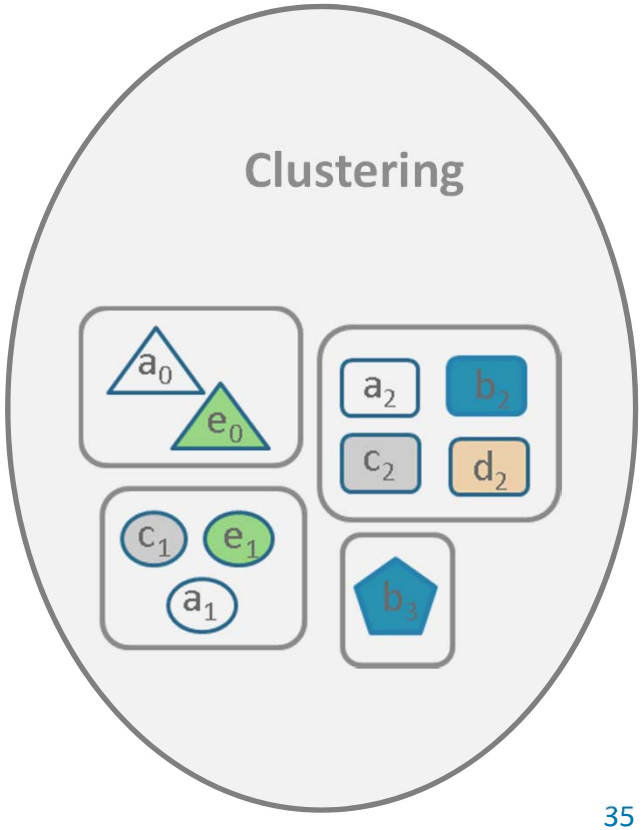
Input



Linking: Similarity Graph



Clustering



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

ScaDS DRESDEN LEIPZIG

AGENDA

UNIVERSITÄT LEIPZIG

- 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)**

ScaDS DRESDEN LEIPZIG

REFERENCES

UNIVERSITÄT LEIPZIG

- 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