UNIVERSITÄT LEIPZIG

ScaDS
DRESDEN LEIPZIG

# HOLISTIC DATA INTEGRATION FOR BIG DATA

ERHARD RAHM,
UNIVERSITY OF LEIPZIG,
AUGUST 2016

www.scads.de

ScaDS DRESDEN LEIPZIG

# GERMAN CENTERS FOR BIG DATA

Two Centers of Excellence for Big Data in Germany

- ScaDS Dresden/Leipzig
- Berlin Big Data Center (BBDC)

ScaDS Dresden/Leipzig:
Competence Center for
Scalable Data Services and Solutions
Dresden/Leipzig

- scientific coordinators: Nagel (TUD), Rahm (UL)
- start: Oct. 2014
- duration: 4 years (option for 3 more years)

ScaDS DRESDEN LEIPZIG

# STRUCTURE OF THE CENTER

Life sciences

Material and Engineering sciences

Environmental / Geo sciences

Digital Humanities

Business Data

Service center

Big Data Life Cycle Management and Workflows

Data Quality / Data Integration

Knowledge Extraktion

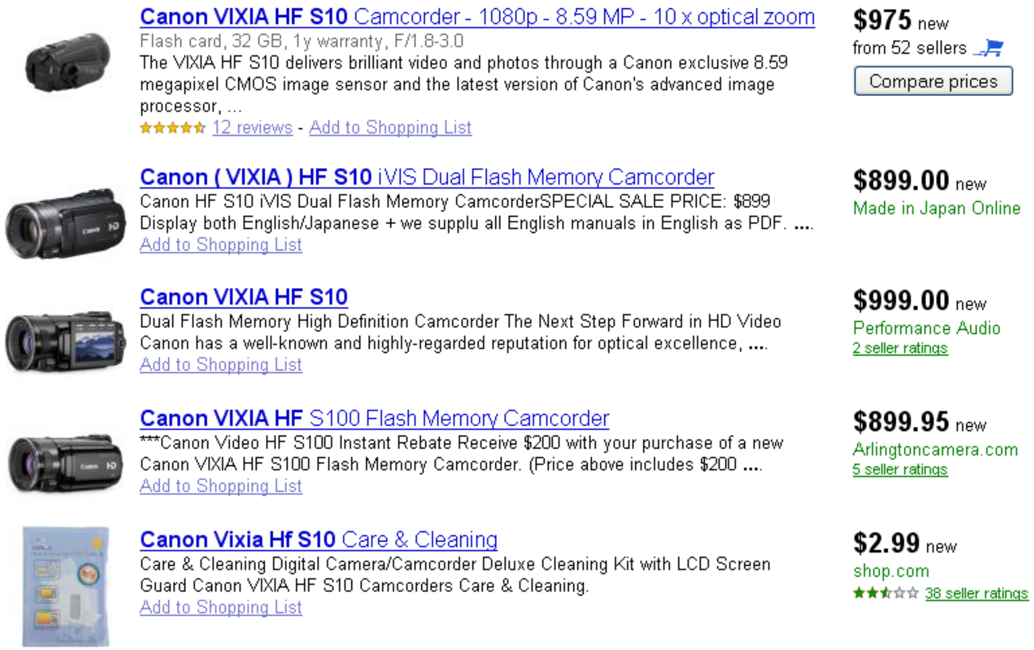
Visual Analytics

Efficient Big Data Architectures

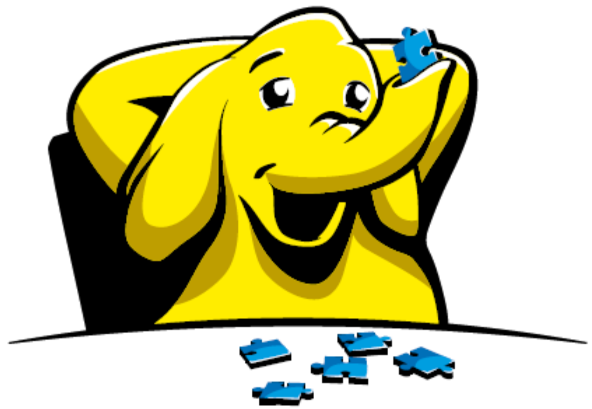
ScaDS DRESDEN LEIPZIG

# DATA QUALITY AND INTEGRATION

- Parallel execution of comprehensive data integration workflows
- Learning based configuration of integration workflows




continuous changes in thousands of data sources

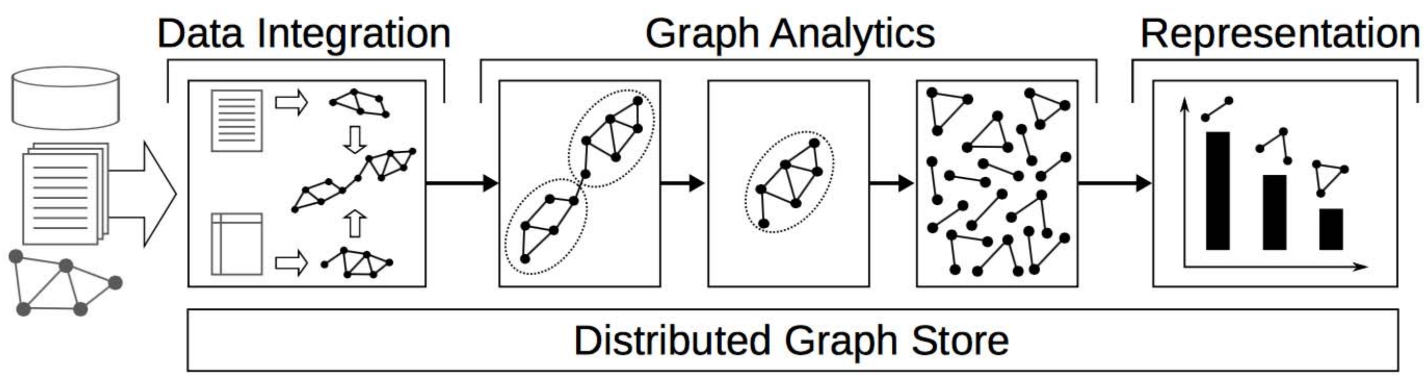


*"This tool (Dedoop) by far shows the most mature use of MapReduce for data deduplication"*
*www.hadoopsphere.com*

ScaDS DRESDEN LEIPZIG

# LARGE SCALE GRAPH ANALYTICS

- GRADOOP framework
  - End-to-end graph data management and analytics
  - Extended property graph data model (EPGM)
  - powerful graph operators for data integration and graph analytics
  - Analytics language GRALA
  - parallel execution on Hadoop clusters
  - open-source: www.gradoop.org



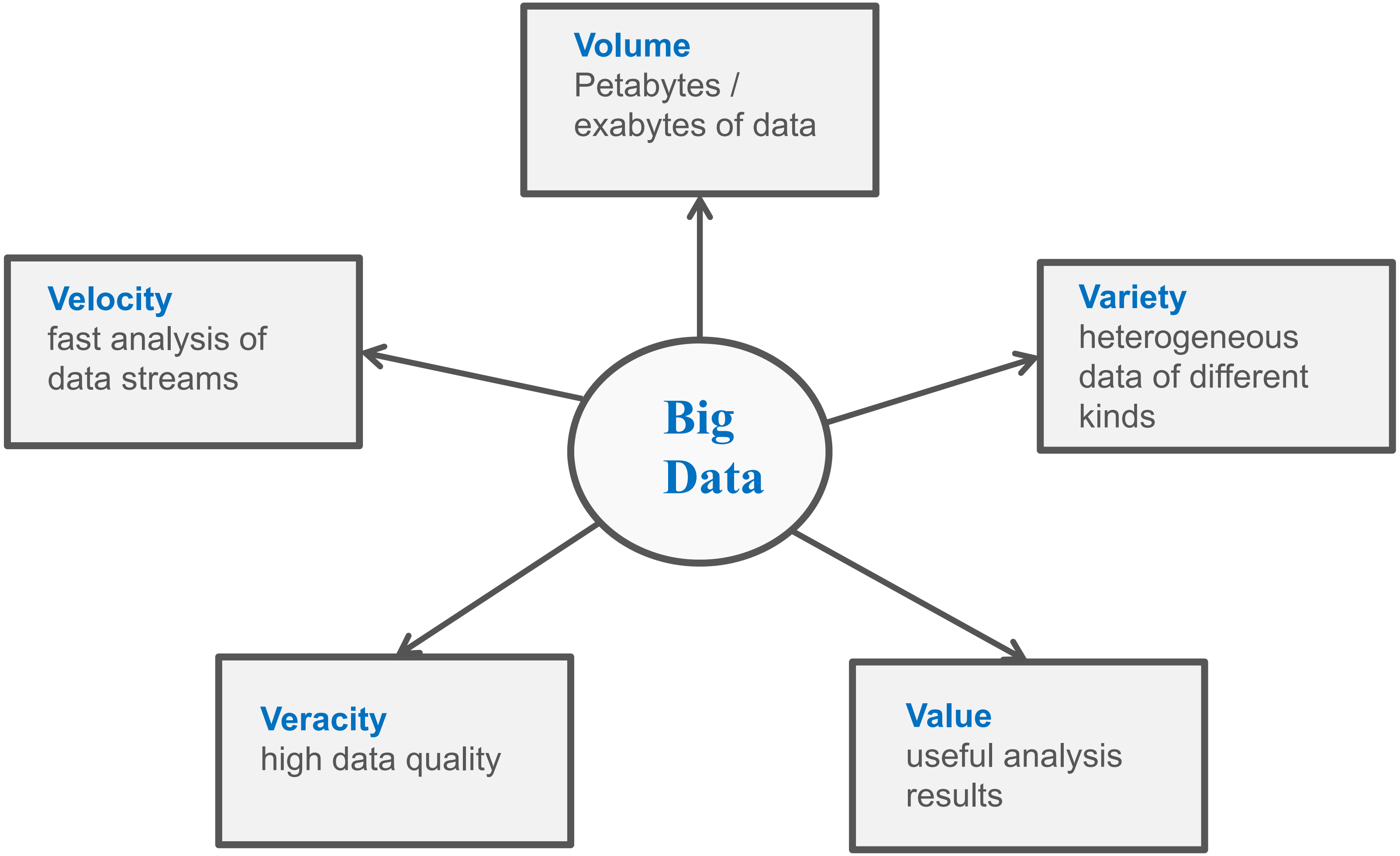
ScaDS DRESDEN LEIPZIG

# AGENDA

- Introduction
  - Big Data
  - Scalable data integration
- Holistic data integration: use cases
- Holistic entity resolution for Linked Data
- Summary

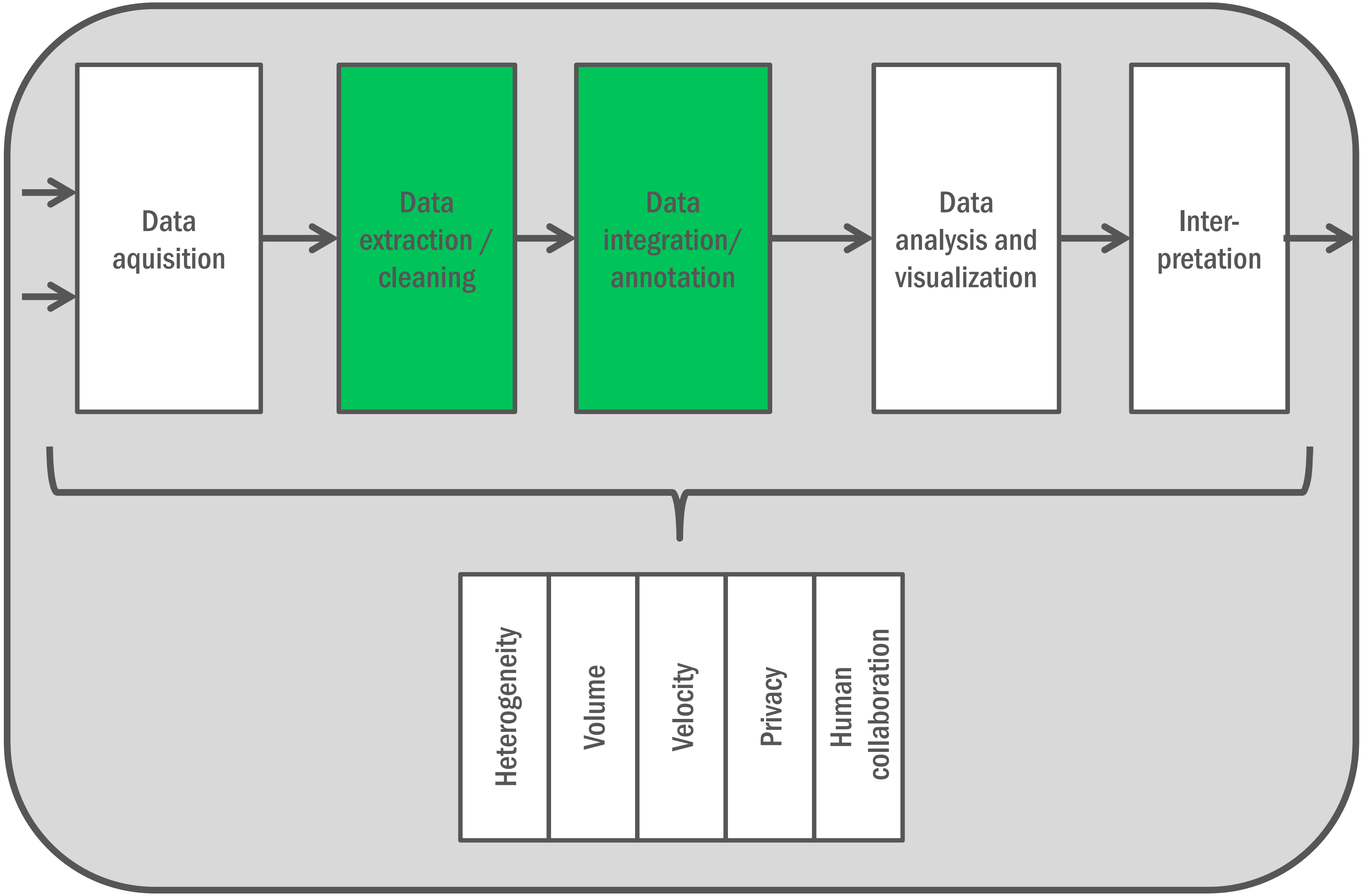
ScaDS DRESDEN LEIPZIG

# BIG DATA CHALLENGES



ScaDS DRESDEN LEIPZIG

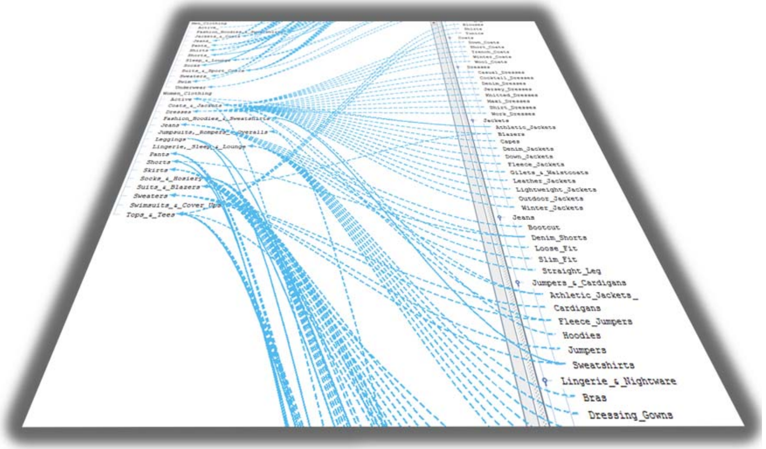
# BIG DATA ANALYSIS PIPELINE



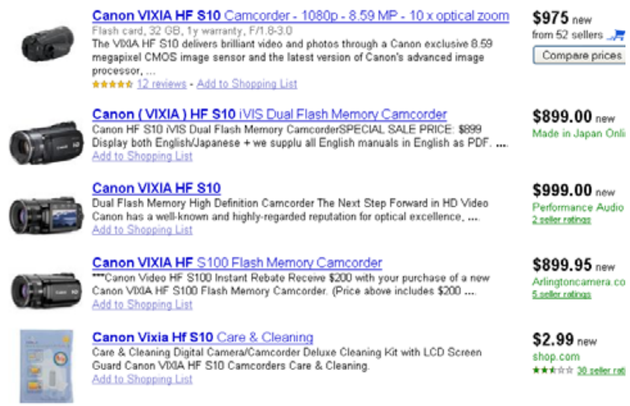
ScaDS DRESDEN LEIPZIG

# 2 LEVELS OF DATA INTEGRATION

- Metadata (schema/ontology) level
  - *Schema Matching:* find correspondences between source schemas and target schema
  - *Schema Merge:* combine source schemas into integrated target schema




- Instance (entity) level
  - transform heterogeneous source data into uniform representation
  - identify and resolve data quality problems
  - identify and resolve equivalent instance records: *object matching / entity resolution* / link discovery
  - Fusion of matching objects



ScaDS DRESDEN LEIPZIG

# OBJECT MATCHING (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 |

ScaDS DRESDEN LEIPZIG

# DUPLICATE WEB ENTITIES: PUBLICATIONS

Data cleaning: Problems and current approaches
E Rahm, HH Do - IEEE Data Eng. Bull., 2000
Cited by 1234 - Related articles - All 37 versions

Data cleaning: Problems and current approaches *
ERHH Do, E Rahm - IEEE Data Engineering Bulletin, 2000
Cited by 12 - Related articles

Hai Do H.: Data Cleaning: Problems and Current approches *
E Rahm - Bulletin of the Technical Committee on Data …, 2000
Cited by 7 - Related articles

Problems and Current Approaches *
D Cleaning - Erhard Rahm, Hong Hai Do. IEEE Data Engineering …, 2000
Cited by 6 - Related articles

Hong Hai Do *
E Rahm - IEEE Bulletin of the Technical Committee on Data …, 2000
Cited by 5 - Related articles

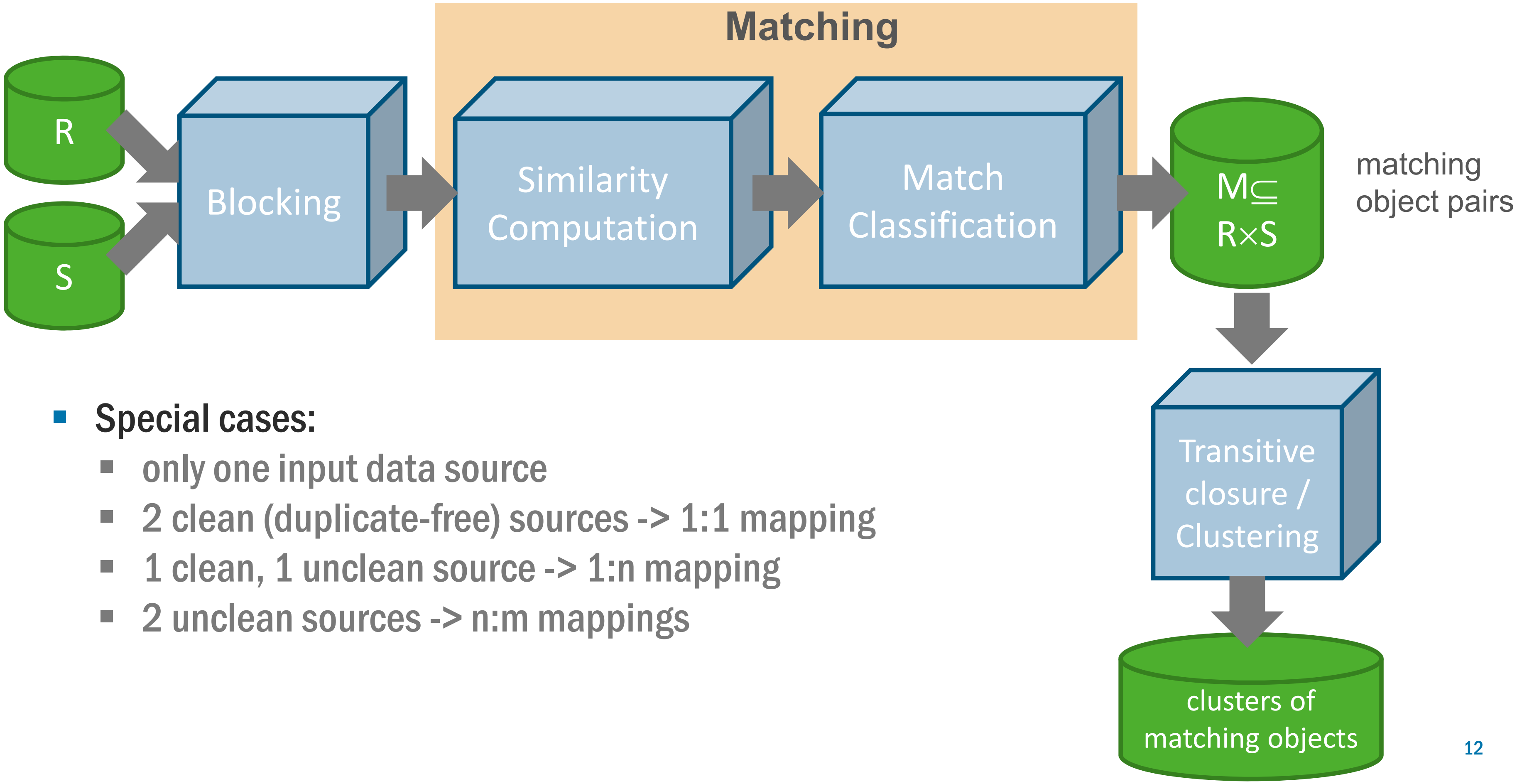
Do Hon g hai. Data Cleaning: Problems and Current Approaches *
E Rahrn - IEEE Data Enginering Bulletin, 2000
Cited by 4 - Related articles

Data Cleaning: Problems & Current Approaches *
D Hang-Hai, R Erhard - IEEE bulletin of the technical committee on Data …, 2000
Cited by 4 - Related articles

Data Cleaning: Problems and Current Approaches. IEEE Techn *
E Rahm, HH Do - Bulletin on Data Engineering, 2000
Cited by 3 - Related articles

ScaDS DRESDEN LEIPZIG

# GENERAL OBJECT MATCHING WORKFLOW



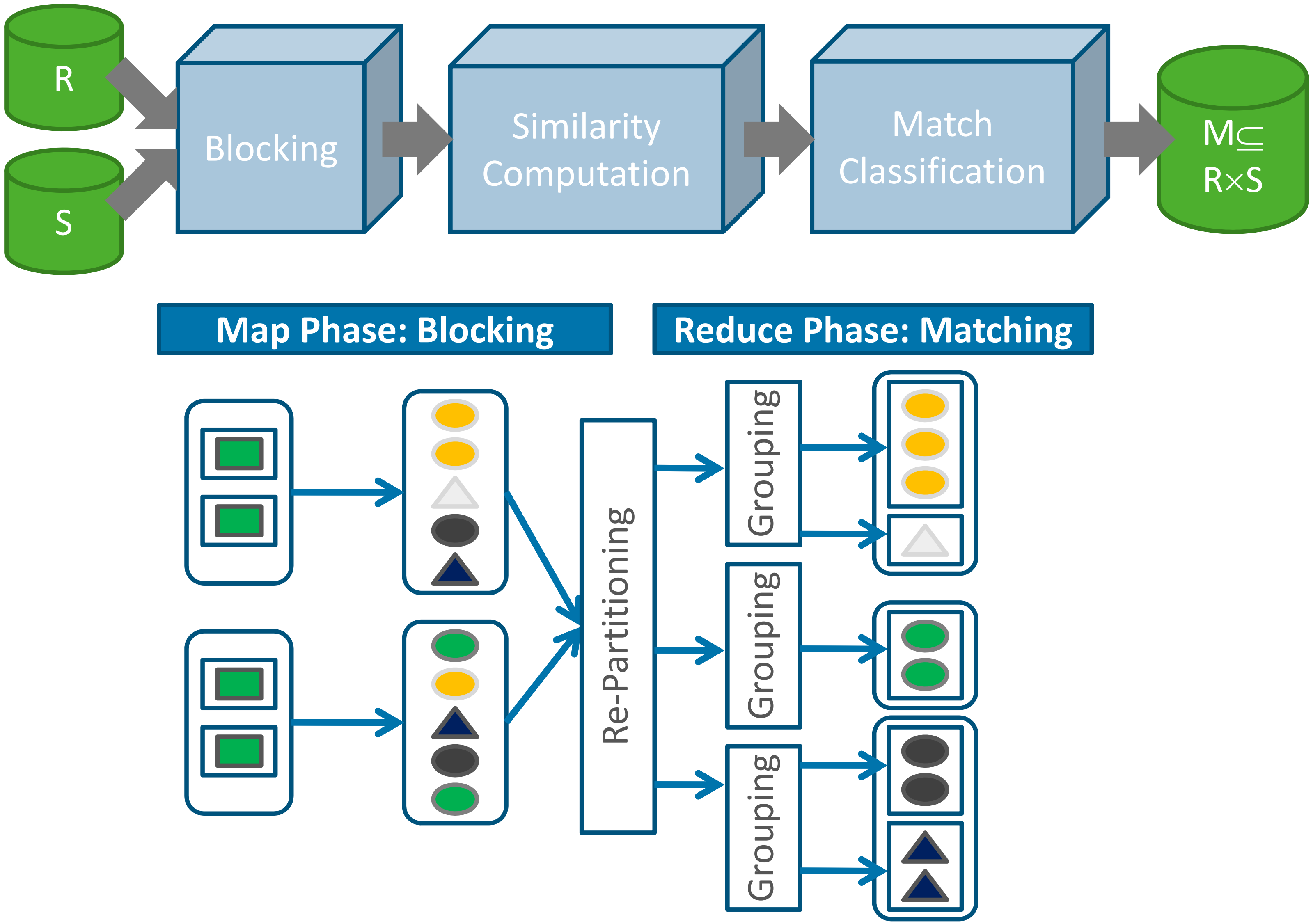

- **Special cases:**
  - only one input data source
  - 2 clean (duplicate-free) sources -> 1:1 mapping
  - 1 clean, 1 unclean source -> 1:n mapping
  - 2 unclean sources -> n:m mappings

# KEY CHALLENGES

- 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
- Data quality
  - unstructured, semi-structured sources
  - need for data cleaning and enrichment
- Privacy for sensitive data
  - privacy-preserving record linkage and data mining

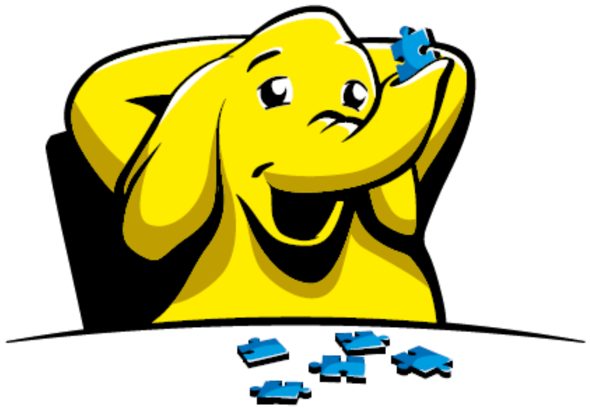
ScaDS DRESDEN LEIPZIG

# PARALLEL OBJECT MATCHING WITH MAP/REDUCE



# DEDOOP: EFFICIENT DEDUPLICATION WITH HADOOP

- Parallel execution of data integration/ 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*

ScaDS DRESDEN LEIPZIG

# AGENDA

- Introduction
- Holistic data integration: use cases
  - Linked Open Data (LOD)
  - Holistic schema matching / ontology integration
  - Web tables
  - Matching of product offers
  - Knowledge graphs
  - Comparison
- Holistic entity resolution for Linked Data
- Summary

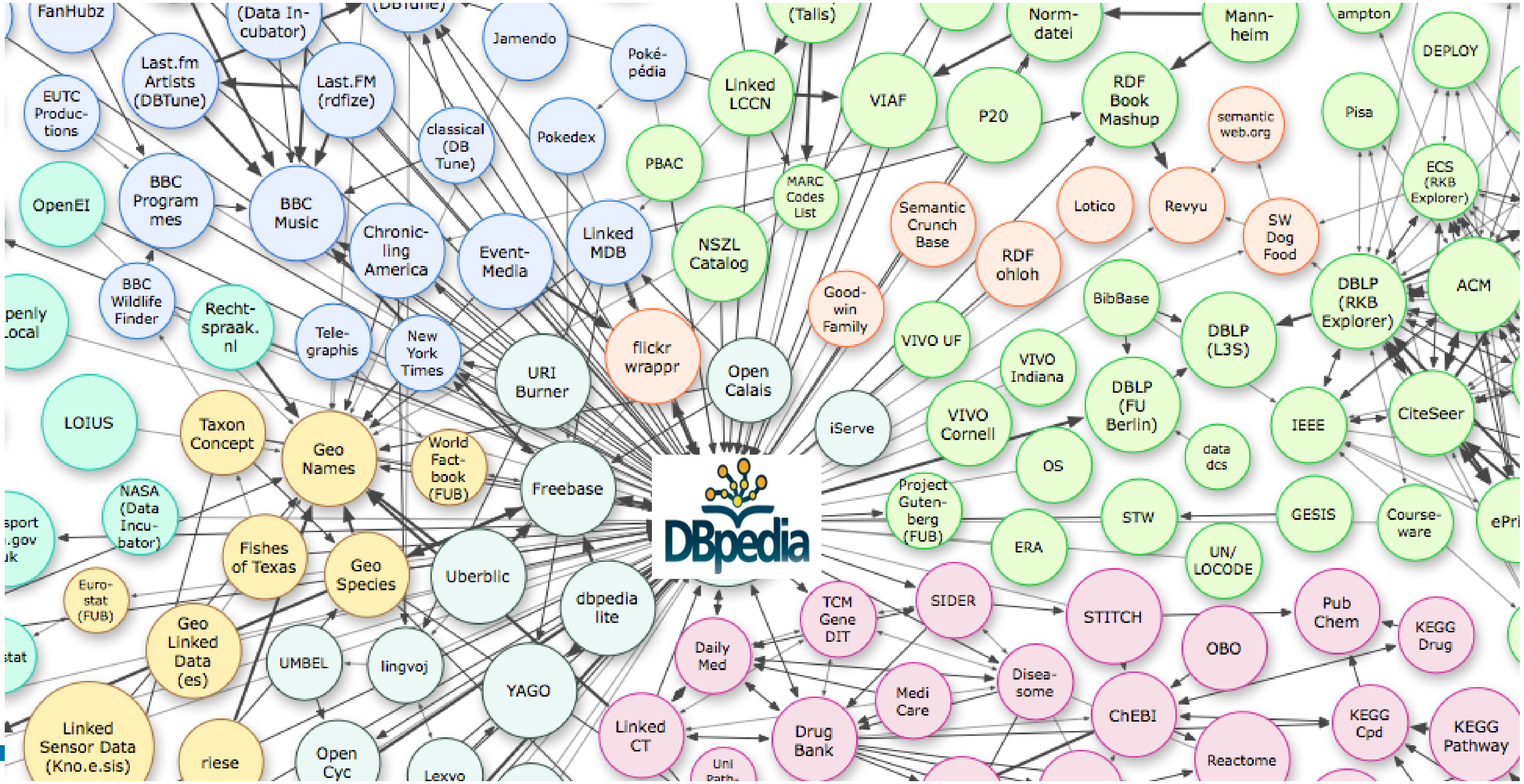
# HOLISTIC DATA INTEGRATION

- Scalable approaches for integrating N data sources (N >>2)
- Increasing need due to numerous sources, e.g., from the web
  - hundreds of LOD sources (Data Web)
  - many thousands of web shops
  - many millions of web tables
- Large open data /metadata/mapping repositories
  - *dataset collections*: data.gov, datahub.io, www.opensciencedatacloud.org, web-datacommons.org
- pairwise matching does not scale
  - 200 sources -> 20.000 mappings

ScaDS DRESDEN LEIPZIG

# LINKED OPEN DATA (LOD)

- many linked data sources containing ontologies and associated instances
- *sameAs* Links (mappings) between ontology concepts and instances for data integration
- federated SPARQL queries to traverse links and access data from different sources



# LINK DISCOVERY FRAMEWORKS

Considered LD tools (sorted by year of initial publication)

| System / initial publication | Year | Institution | Learning-based | OAEI IM participation | Support for pure ontology matching |
|---|---|---|---|---|---|
| RiMOM [66] | 2004 | Univ. of Tsinghua, China | | ✓ | ✓ |
| KnoFuss [46] | 2007 | Open Univ. Milton Keynes, UK | ✓ | | |
| AgreementMaker [9] | 2009 | Univ. of Illinois at Chicago, USA | | ✓ | ✓ |
| Silk [69] | 2009 | FU Berlin, Germany | ✓ | | |
| CODI [44] | 2010 | Univ. of Mannheim, Germany | | ✓ | ✓ |
| LIMES [39] | 2011 | Univ. of Leipzig, Germany | ✓ | | |
| LogMap [25] | 2011 | Univ. of Oxford, UK | | ✓ | ✓ |
| SERIMI [3] | 2011 | Delft Univ. of Techn., Netherlands | | ✓ | |
| Zhishi.links [48] | 2011 | Shanghai Jiao Tong Univ., China | | ✓ | |
| SLINT+ [43] | 2012 | Nat. Inst. of Informatics, Japan | | ✓ | |
| RuleMiner [47] | 2012 | Shanghai Jiao Tong Univ., China | ✓ | | |

M. Nentwig, M. Hartung, A. Ngonga, E. Rahm: *A Survey of Current Link Discovery Frameworks.* Semantic Web Journal 2016 (accepted for publication)

# NON-LEARNING LINK DISCOVERY TOOLS

Characteristics of proposed LD frameworks ("-" means not existing, "?" unclear from publication, "*" supported in respective ontology matching framework, [1] no answer on form submission)

| | RiMOM | AgreementMaker | CODI | LogMap | SERIMI | Zhishi.links | SLINT+ |
|---|---|---|---|---|---|---|---|
| Data Input | RDF, OWL | SPARQL | RDF, OWL | RDF, OWL | SPARQL | RDF | RDF |
| Supported link types | `owl:sameAs` | `owl:sameAs` | `owl:sameAs` | `owl:sameAs` | `owl:sameAs` | `owl:sameAs` | `owl:sameAs` |
| Configuration<br>- matcher combination | adaptive<br>weighted average | manual<br>weighted combination | manual<br>weighted average | manual<br>weighted average | adaptive<br>- | manual<br>weighted combination | adaptive<br>weighted average |
| Runtime optimization<br>- Blocking<br>- Filtering | <br>-<br>indexing | <br>-<br>indexing | <br>-<br>- | <br>-<br>indexing | <br>-<br>- | <br>-<br>indexing | <br>-<br>indexing |
| String similarity measures | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Further similiarty measures | - | - | - | - | - | geographical coordinates | inverted disparity |
| Structure matcher | - | semantic similarity | iterative anchor-based mapping generation | iterative anchor-based mapping generation | - | semantic similarity | - |
| Use of<br>- external dictionaries<br>- existing mappings | <br>?*<br>- | <br>?*<br>- | <br>-<br>- | <br>?*<br>- | <br>-<br>- | <br>-<br>- | <br>-<br>- |
| Post-processing | - | - | Coherence checks | Inconsistency repair | - | - | - |
| Parallel processing | - | - | - | - | - | MapReduce | - |
| GUI/web interface/API | - / - / - | ✓/ ? / - | - / - / - | ✓/ ✓/ - | - / - / - | - / - / - | - / - / - |
| Download Tool/Source<br>Open Source project | ✓/ -<br>- | -[1] / -<br>- | ✓/ ✓<br>✓ | ✓/ ✓<br>✓ | ✓/ ✓<br>✓ | ✓/ -<br>- | ✓/ -<br>- |

# LEARNING-BASED LINK DISCOVERY TOOLS

Characteristics of learning-based LD frameworks. "-" means not existing, "*" investigated in [20], but not available in current release

| | KnoFuss | Silk | LIMES | RuleMiner |
|---|---|---|---|---|
| Data Input | RDF, SPARQL | RDF, SPARQL, CSV | RDF, SPARQL, CSV | RDF |
| Supported linktypes | `owl:sameAs` | `owl:sameAs`, user-specified others | `owl:sameAs`, user-specified others | `owl:sameAs` |
| Configuration | manual (match rules), unsupervised learning (genetic programming) | manual (match rules), supervised learning (genetic programming, active learning) | manual (match rules), supervised learning (genetic programming, active learning), unsupervised (genetic programming) | adaptive (match rules), supervised learning (expectation maximization) |
| Runtime optimization<br>- Blocking<br>- Filtering | <br>-<br>indexing | <br>multi-dimensional<br>- | <br>-<br>space tiling | <br>-<br>indexing |
| String similarity measures | ✓ | ✓ | ✓ | ✓ |
| Further similiarty measures | - | numeric, date equality | geographical coordinates, numeric, date equality | -<br>- |
| Structure matcher | - | - | - | semantic similarity |
| Use of<br>- external dictionaries<br>- existing mappings | <br>-<br>- | <br>-<br>- | <br>-<br>- | <br>-<br>- |
| Post-processing | one-to-one mapping | - | Stable marriage, hospital-resident | - |
| Parallel Processing | - | MapReduce | (MapReduce)* | MapReduce |
| GUI/web interface/API | - / - / - | ✓ / ✓ / ✓ | ✓ / ✓ / ✓ | - / - / - |
| Download Tool/Source<br>Open Source project | ✓ / ✓<br>✓ | ✓ / ✓<br>✓ | ✓ / -<br>- | - / -<br>- |

ScaDS DRESDEN LEIPZIG

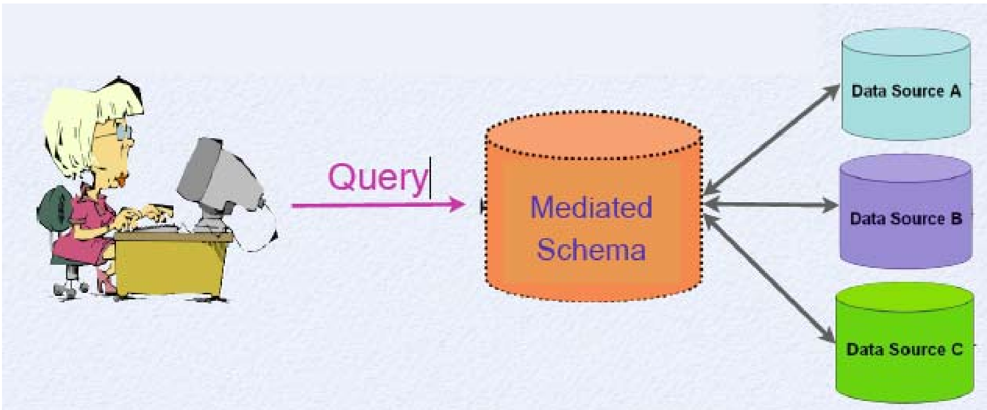
# LD TOOLS: OBSERVATIONS

- mostly simple (attribute/property-based) matchers
  - limited use of context and existing links / synonym information
- limited use of blocking
  - use of filtering/indexing to speed up string comparisons
- mostly support for semi-automatic configurations, e.g. utilizing learning-based techniques
- general availability of most tools
- need for more comparative evaluations, especially regarding large-scale datasets
- binary linking only: no support for holistic data integration

ScaDS DRESDEN LEIPZIG

# HOLISTIC SCHEMA MATCHING: META-SEARCH

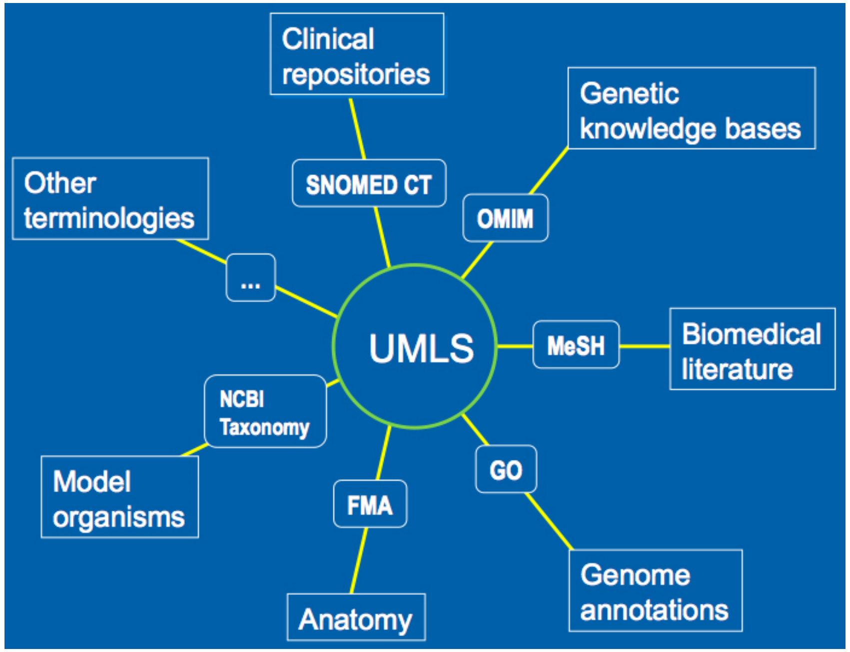
- creation of a mediated schema
  - holistic matching between N simple schemas, e.g., web forms
  - virtual data integration: meta-search
- holistic matching based on *clustering* of similar attributes
  - utilize high name similarity between schemas
  - similar names within a schema are mismatches (e.g. first name, last name)



ScaDS DRESDEN LEIPZIG

# HOLISTIC ONTOLOGY INTEGRATION

- Development of an integrated domain ontology (e.g., UMLS)
  - physical metadata integration
  - tens of source ontologies
  - clustering of synonymous concepts (synsets)
  - largely manual integration effort




source: Fiszman, 2012 “Natural Language Processing and the National Library of Medicine”

ScaDS DRESDEN LEIPZIG

# OPEN DATA REPOSITORIES

- huge data collections with (mostly unrelated) up to millions of files or tables from numerous domains

Google fusion tables beta

datahub
The easy way to get, use and share data

DWTC
Dresden Web Table Corpus

OSDC
OPEN SCIENCE DATA CLOUD

Web Data Commons

Open Data REPOSITORY

DATA.GOV

# WEB TABLES

- Web contains hundreds of millions tables
  - only 1% relational tables + vertical tables about one entity*
- several corpora with huge number of heterogenous tables

Longest rivers

| | River | Length (km) | Length (miles) | Drainage area (km²)[citation needed] | Average discharge (m³/s)[citation needed] | Outflow |
|---|---|---|---|---|---|---|
| 1. | Nile – Kagera[n 1] | 6,853 (6,650) | 4,258 (4,132) | 3,254,555 | 5,100 | Mediterranean |
| 2. | Amazon – Ucayali – Apurímac[n 1] | 6,992 (6,400) | 4,345 (3,976) | 7,050,000 | 219,000 | Atlantic Ocean |
| 3. | Yangtze (Chang Jiang) | 6,300 (6,418) | 3,917 (3,988) | 1,800,000 | 31,900 | East China Sea |
| 4. | Mississippi–Missouri–Jefferson | 6,275 | 3,902 | 2,980,000 | 16,200 | Gulf of Mexico |
| 5. | Yenisei–Angara–Selenge | 5,539 | 3,445 | 2,580,000 | 19,600 | Kara Sea |

Prague (Praha)

| | |
|---|---|
| Country | Czech Republic |
| Founded | c. 885 |
| Government | |
| • Mayor | Adriana Krnáčová (ANO) |
| Area[1] | |
| • Urban | 496 km² (192 sq mi) |
| Highest elevation | 399 m (1,309 ft) |
| Lowest elevation | 177 m (581 ft) |
| Population (2015-12-31)[3] | |
| • Capital city | 1,267,449 |
| • Metro | 2,156,097 [2] |
| Demonym(s) | Praguer |
| Time zone | CET (UTC+1) |
| • Summer (DST) | CEST (UTC+2) |
| Postal code | 100 00 – 199 00 |
| Vehicle registration | A |
| NUTS code | 2012 |
| - Total | €40 billion |
| - Per capita | €32,000([1]) |
| Website | praha.eu |

*Balakrishnan, S., Halevy, A., Harb, B., Lee, H., Madhavan, J., et al: Applying WebTables in Practice. Proc. CIDR 2015

# WEB TABLES (2)

- Query support
  - find related tables for keywords
  - augment tables by additional attributes
- Need to add semantics
  - attributes need to be annotated, e.g., with knowledge graph
  - table contents described in surrounding text
  - Vertical tables: identify key column vs. property column
- *Table augmentation*: find coherent attributes from other tables that can extend a given table

| Company |
|---|
| Bank of China |
| Banco do Brasil |
| Rogers Communications |
| China Mobile |
| AT&T |

Revenue

| | 2012 | 2013 |
|---|---|---|
| Bank of China | x1 | x2 |
| Deutsche Bank | y1 | y2 |
| Banco do Brasil | z1 | z3 |

Telco companies

| | Revenue 2014 |
|---|---|
| China Mobile | x |
| AT&T | y |

Eberius et al: *Top-k entity augmentation using consistent set covering*. Proc. SSDM 2015

# WEB TABLES (3)

- many open challenges to better utilize table/dataset corpora
  - more comprehensive domain categorization and attribute annotation
  - more sophisticated matching of attributes based on instances + metadata
  - clustering and fusing related tables/datasets
  - derivation of domain-specific knowledge graphs

# HOLISTIC INTEGRATION OF ENTITIES

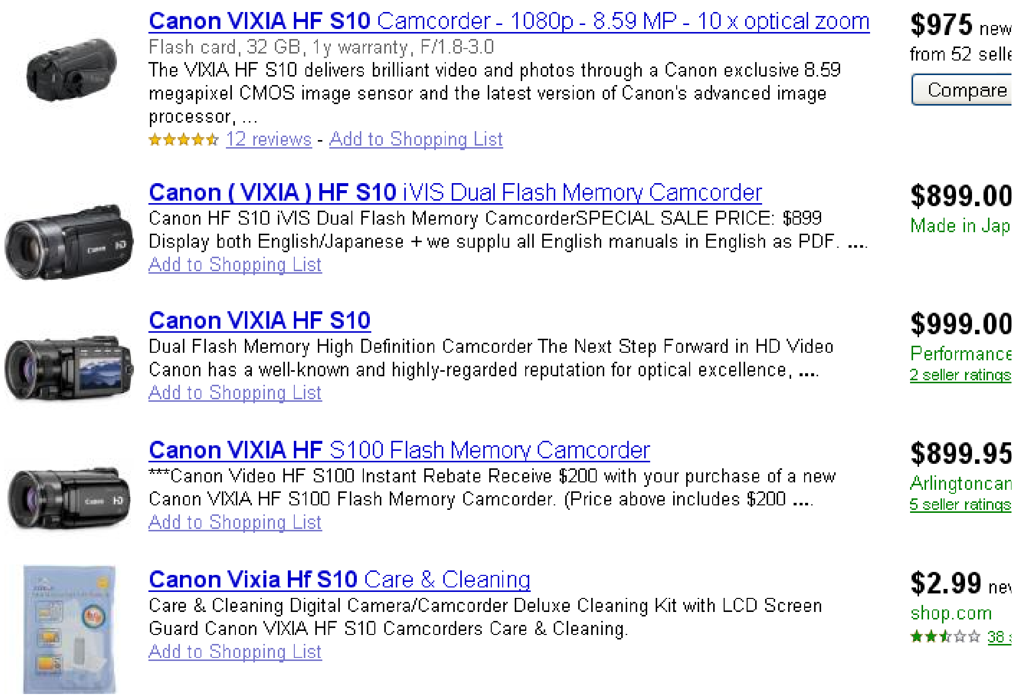
- Entity search engines
  - clustering of matching entities (publications, product offers)
  - physical data integration
  - thousands of data sources

- Comparison / booking portals
  - clustered offers within (integrated) taxonomy
  - physical or virtual data integration

ScaDS DRESDEN LEIPZIG

# HOLISTIC 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



# SOLUTION APPROACH*

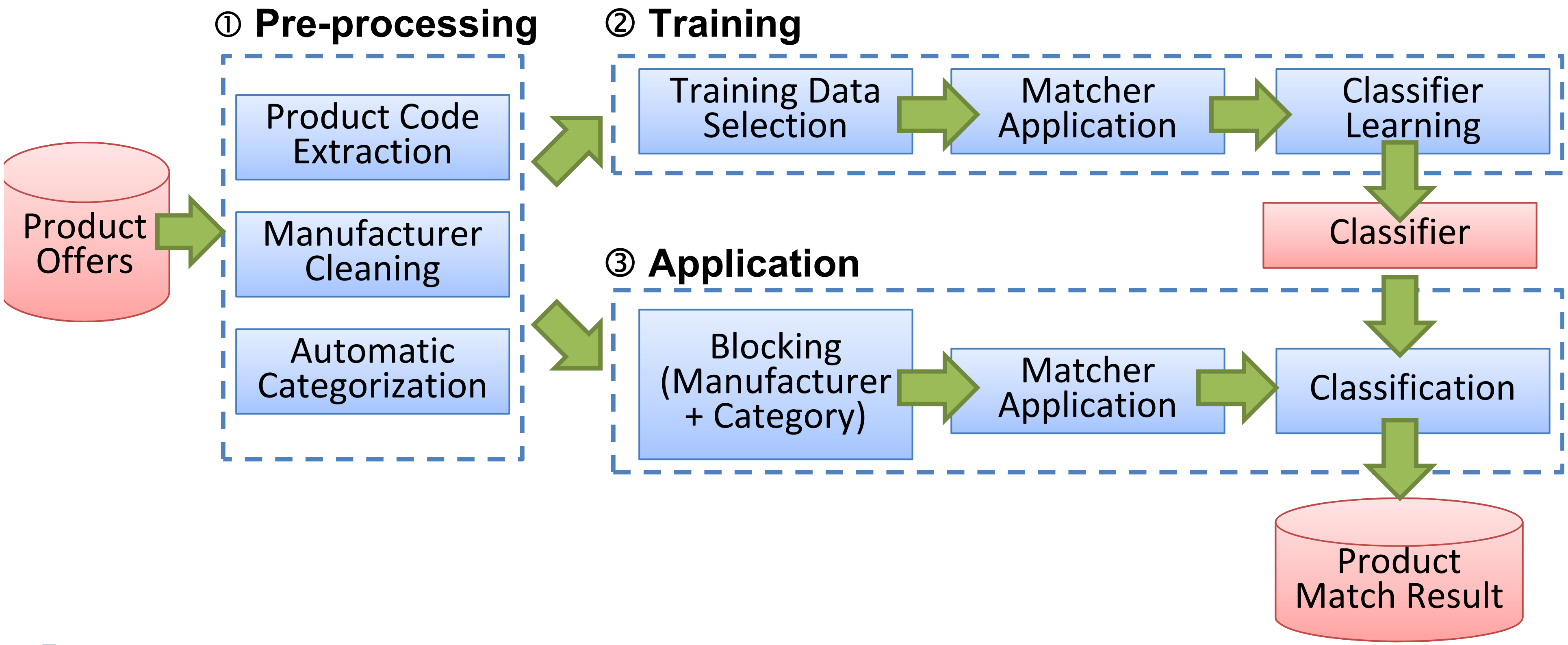
Input:

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

1. preprocessing/ data cleaning:
   - extraction and consolidation of manufacturer info
   - extraction of product codes
2. (learning-based) categorization of product offers
   - determine product (entity) type
3. (learning-based) matching of product offers per product category

* Koepcke, Thor, Thomas, Rahm: *Tailoring entity resolution for matching product offers*. Proc. EDBT, 2012

# LEARNING-BASED MATCH APPROACH



# KNOWLEDGE GRAPHS

- web-scale integration and categorization of heterogeneous entities

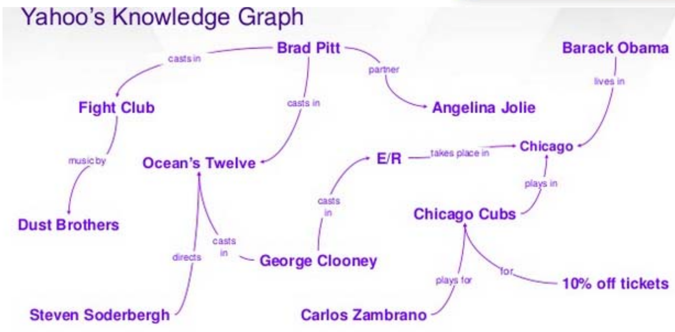
DBpedia

WIKIDATA

yago select knowledge

Google Knowledge Graph

悟 SATORI




Nicolas Torzec: Making knowledge reusable at Yahoo!:
a Look at the Yahoo! Knowledge Base (SemTech 2013)

YAHOO!

ScaDS DRESDEN LEIPZIG

# GOOGLE KNOWLEDGE GRAPH

- Enhanced search results

**Universität Leipzig**
www.uni-leipzig.de/ ▾ Translate this page
Offizieller Internetauftritt mit Vorstellung der Leipziger **Universität** mit umfangreichen Informationen zu Forschung und Lehre.

Results from uni-leipzig.de

**Studiengänge**
Kommunikations - Management Science - Psychologie - Medizin

**Bewerbung und Immatrikulat...**
Sie sind hier: Studium»; Bewerbung und ...

**Fakultäten**
14 Fakultäten der Universität. Theologische Fakultät ...

**International Study**
International students. Welcome. You are from abroad and you ...

**Leipzig University - Wikipedia, the free encyclopedia**
https://en.wikipedia.org/wiki/Leipzig_University ▾
Leipzig University (German: **Universität Leipzig**), located in Leipzig in the Free State of Saxony, Germany, is one of the oldest universities in the world and the ...

**Universität Leipzig – Wikipedia**
https://de.wikipedia.org/wiki/Universität_Leipzig ▾ Translate this page
Die **Universität Leipzig** – Alma Mater Lipsiensis (AML) – ist die größte Hochschule in Leipzig. Mit ihrem Gründungsjahr 1409 ist sie auf dem Gebiet der ...

**Universität Leipzig - bei Facebook**
https://de-de.facebook.com/unileipzig ▾ Translate this page
★★★★★ Rating: 4,4 - 259 votes
**Universität Leipzig**, Leipzig. 39.880 „Gefällt mir"-Angaben · 685 Personen sprechen darüber · 12.647 waren hier. Offizielle Facebook-Präsenz der...

**Universitätsmedizin Leipzig**
www.uniklinikum-leipzig.de/ ▾ Translate this page
Im Herzen der Stadt **Leipzig** gehört der Medizin-Campus an der Liebigstraße zu den modernsten in ganz Deutschland mit besten Bedingungen für ambulante ...

**Universität Leipzig (@UniLeipzig) | Twitter**


## Leipzig University

Website | Directions

University in Leipzig, Germany

Leipzig University, located in Leipzig in the Free State of Saxony, Germany, is one of the oldest universities in the world and the second-oldest university in Germany. Wikipedia

**Address:** Augustusplatz 10, 04109 Leipzig
**Enrollment:** 28,275 (2014)
**Customer service:** 0341 97108
**Founded:** December 2, 1409
**President:** Beate Schücking
**Founders:** William II, Margrave of Meissen, Frederick I, Elector of Saxony, Wilhelm Wundt

Profiles

LinkedIn

Notable alumni — View 45+ more

Angela Merkel | Johann Wolfgang von Goethe | Gottfried Wilhelm Leibniz | Richard Wagner | Gotthold Ephraim Lessing

# GOOGLE KNOWLEDGE GRAPH (2)

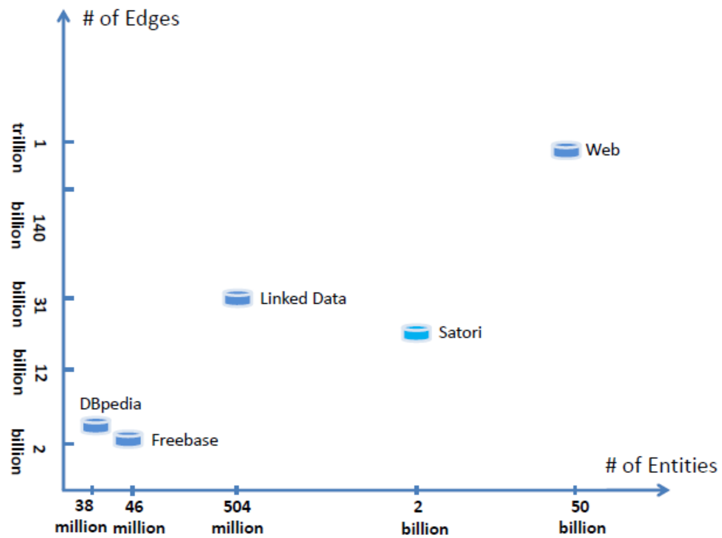
- combines knowledge from numerous sources
  - Freebase, Wikipedia, CIA World fact book, …
  - 2012: > 570 million entities, > 18 billion facts/relationships

TABLE II
SIZE OF SOME SCHEMA-BASED KNOWLEDGE BASES

| Knowledge Graph | Number of Entities | Number of Relation Types | Number of Facts |
|---|---|---|---|
| Freebase | 40 M | 35,000 | 637 M |
| Wikidata | 13 M | 1,643 | 50 M |
| DBpedia[1] | 4.6 M | 1,367 | 68 M |
| YAGO2 | 10 M | 72 | 120 M |
| Google Knowledge Graph | 570 M | 35,000 | 18,000 M |

Nickel, Murphy, Tresp, Gabrilovich: A review of relational machine learning for knowledge graphs (Arxiv, 2015)

## The Scale of Knowledge Graphs




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

ScaDS DRESDEN LEIPZIG

# KNOWLEDGE GRAPHS

- uniform representation and semantic categorization of entities of many domains (web-scale) or only one domain, one enterprise etc.
  - examples: DBPedia, Yago, Wikidata, Google KG, MS Satori, Facebook, ...
  - entities and their metadata often extracted from other resources such as Wikipedia, Wordnet etc. as well as from normal web pages, documents, web searches etc.
  - comprehensive taxonomies to categorize entities and their details
  - extreme entity heterogeneity (attributes + values) even within domains
  - very challenging data integration problems
- Knowledge Graphs provide valuable background knowledge for
  - enhancing entities (based on prior *entity linking* )
  - improving data integration (e.g., by utilizing additional information)
  - improving search results ...

ScaDS DRESDEN LEIPZIG

# USE CASES FOR HOLISTIC DATA INTEGRATION

| Use case | Data integration type | | #domains | #sources | Clustering? | degree of automated data integration |
|---|---|---|---|---|---|---|
| **Meta-search** | virtual | metadata | 1 | low - medium | attributes | medium |
| **Open data, web tables** | physical collection | primarily metadata | many | very high | (possible) | high, but limited integration |
| **Integrated ontology** | physical | metadata | 1+ | low - medium | concepts | low - medium |
| **Entity search engines** | physical | data (+ metadata) | 1 | very high | entities | high |
| **Booking portals** | physical | data + metadata | 1+ | high | entities | high |
| **Knowledge graphs** | physical | data + metadata | many | low - high | entities + concepts/ attributes | medium - high |

ScaDS DRESDEN LEIPZIG

# OBSERVATIONS

- Most scalable approaches are based on
  - Physical data integration
  - Integration of instance data rather than metadata integration
- Clustering instead of mappings
  - cluster of k matching objects represents $k^2/2$ correspondences
  - cluster size limited by #sources (for duplicate-free sources)
  - simplified fusion of corresponding objects
  - incremental data integration: additional sources/objects only need to be matched with clusters instead of all other sources

-> need for clustering-based concept matching and instance matching

ScaDS DRESDEN LEIPZIG

# AGENDA

- Introduction
- Holistic data integration: use cases
- Holistic entity resolution for Linked Data
  - High-level clustering approach
  - Clustering for linked data
- Summary

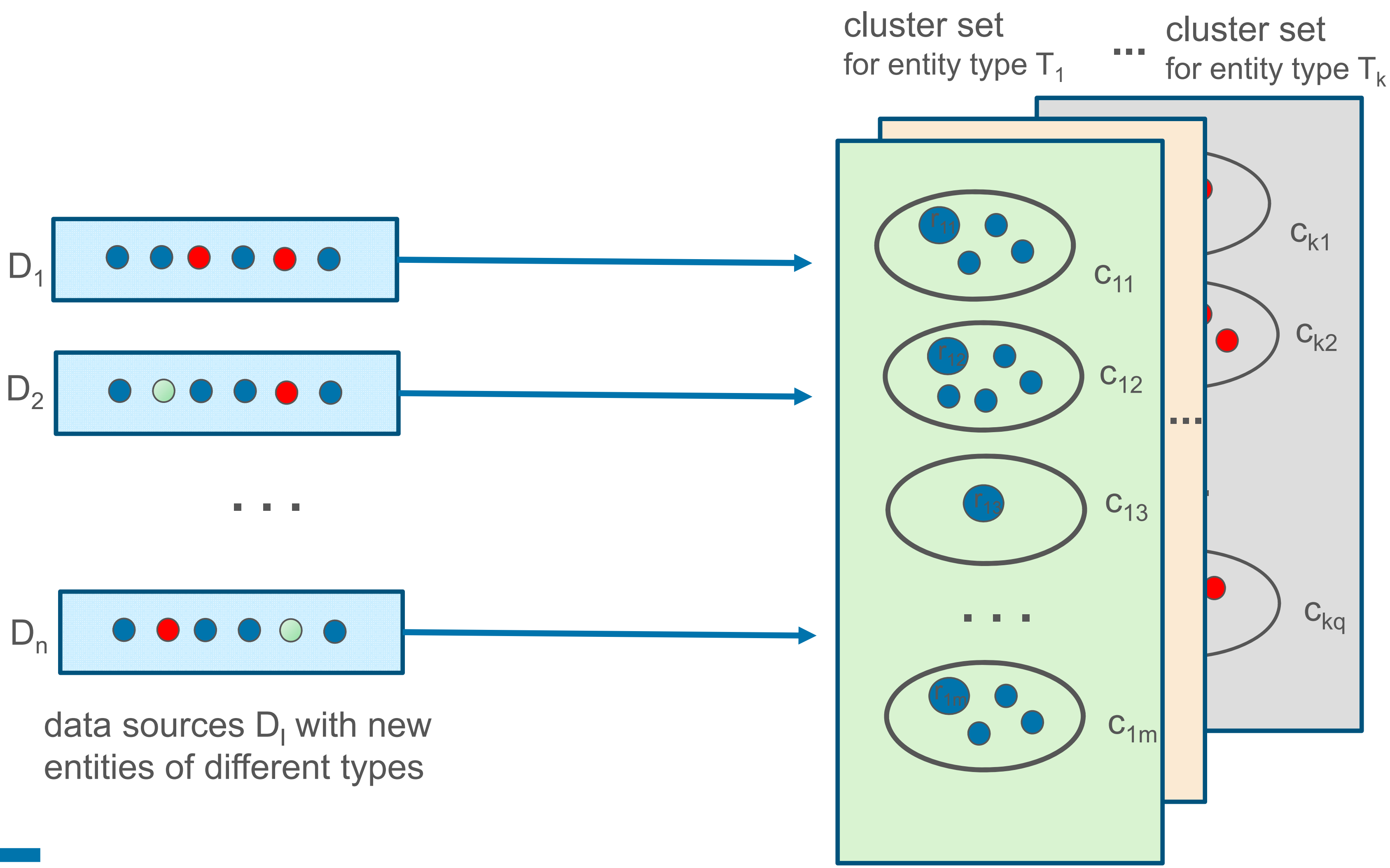
# HOLISTIC ENTITY RESOLUTION

- requirements
  - scalability to many data sources and high data volumes
  - dynamic addition of new sources / entities (data streams)
  - support for many entity types
  - high match quality
  - little or no manual interaction
- binary match approaches not sufficient

# HOLISTIC ENTITY RESOLUTION (2)

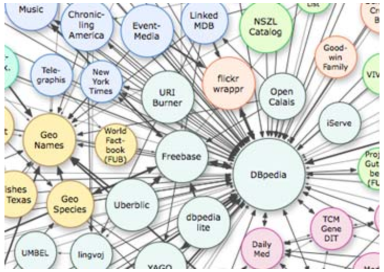
- clustering-based approaches
  - represent matching entities from k sources in single cluster
  - determine *cluster representative* for further processing/matching
- incremental addition/clustering of sources, e.g., starting with the largest data source
- distinguish clusters by entity type (product categories, geographical entities, etc.)
  - determine entity type for new entities (preprocessing)
- utilize blocking to restrict number of clusters to match with

# CLUSTERING-BASED ENTITY RESOLUTION



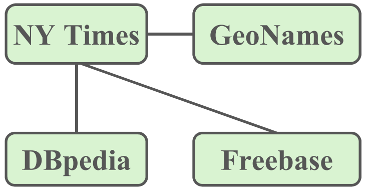
ScaDS DRESDEN LEIPZIG

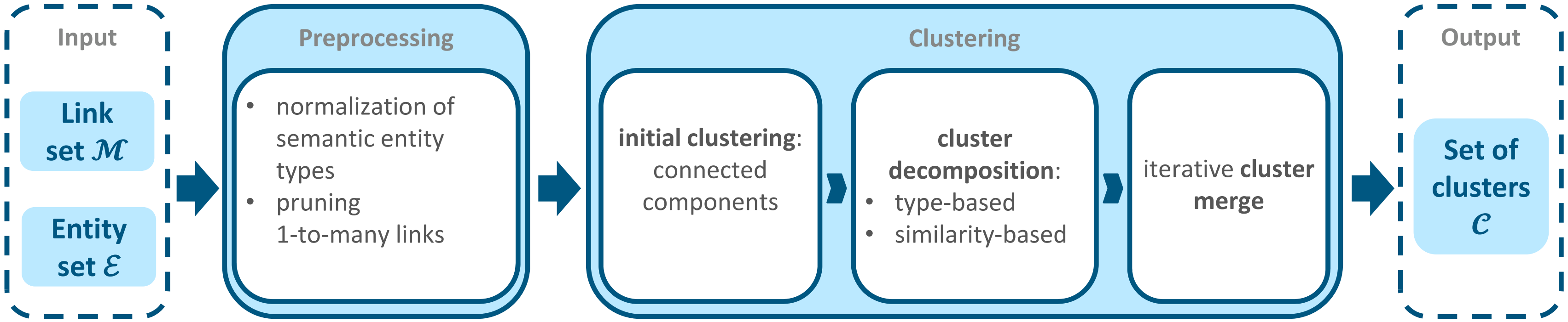
# HOLISTIC ENTITY RESOLUTION FOR LINKED DATA



- characteristics
  - entity clusters instead of binary sameAs links
  - utilization of existing (possibly noisy) link mappings
  - support for different entity types
- found clusters represent many additional links and fused knowledge of improved quality
  - can be used to build knowlede graphs

Use case: cluster geographical entities from 4 sources

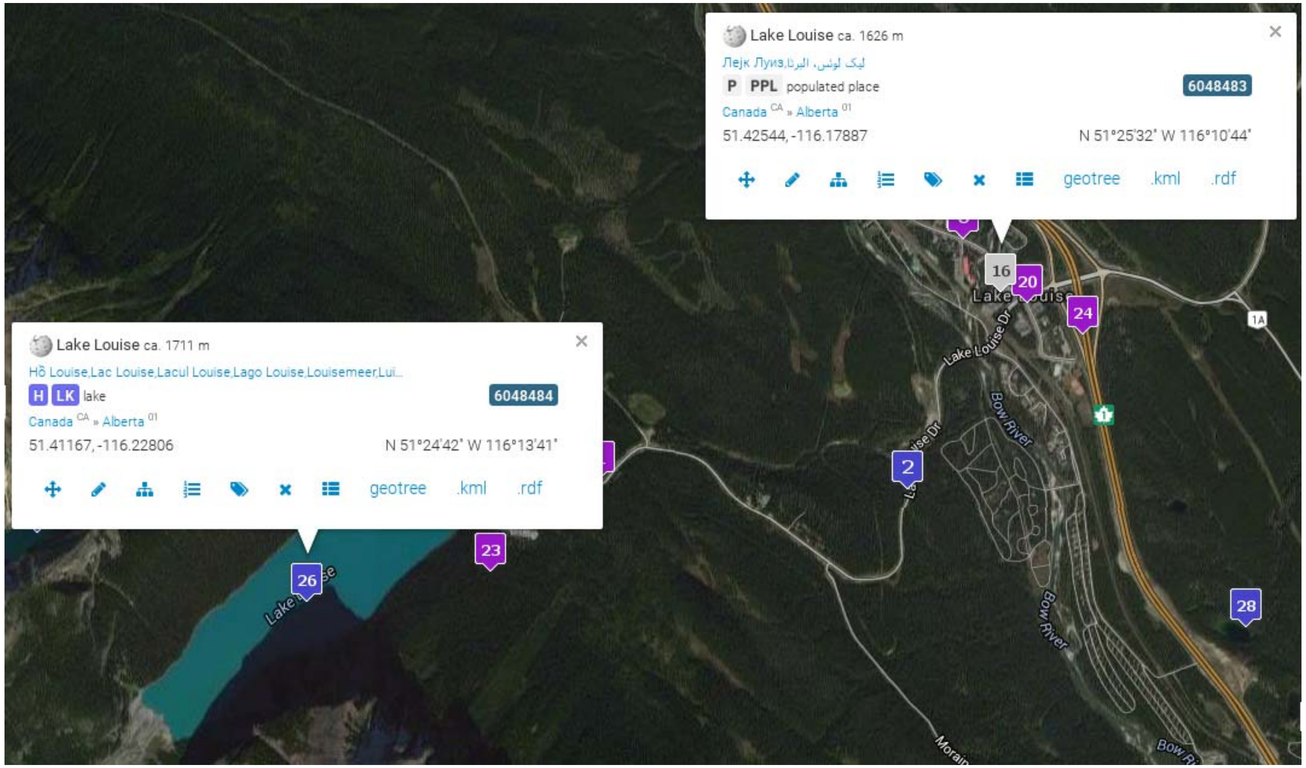


ScaDS DRESDEN LEIPZIG

# LINKING ERRORS

- existing sameAs links partially wrong
  - equal name, similar geographic coordinates
  - but different type (lake vs. city)



# HOLISTIC CLUSTERING WORKFLOW: EXAMPLE

Pre processing → Initial Clustering → Cluster Decomposition → Cluster Merge

| id | label | source | type | latitude | longitude |
|---|---|---|---|---|---|
| 0 | Lake Louise (Canada) | NYTimes | - | 51.42 | -116.23 |
| 1 | Lake Louise | GeoNames | gn:H.LK | 51.41 | -116.23 |
| 2 | Lake Louise, Alberta | DBpedia | db:Settlement | - | - |
| 3 | lake louise alberta | FreeBase | fb:location.citytown | 51.43 | -116.16 |
| 4 | Lake Louise (Alberta) | DBpedia | db:BodyOfWater | 51.41 | -116.23 |
| 5 | lake louise | FreeBase | fb:geography.lake | 51.41 | -116.23 |
| 6 | Mystic (Conn) | NYTimes | - | 41.35 | -71.97 |
| 7 | N17632379615920 | FreeBase | - | 41.35 | -71.97 |
| 8 | Mystic | GeoNames | gn:P.PPL | 41.35 | -71.97 |
| 9 | Black Hill | GeoNames | gn:T.HLL | 53.54 | -1.89 |
| 10 | Black Hill | DBpedia | db:Mountain | 53.53 | -1.88 |
| 11 | Black Hill | LinkedGeoData | lgd:Peak | 53.96 | -1.85 |
| 12 | Black Hill | LinkedGeoData | lgd:Peak | 54.69 | -2.15 |
| 13 | katmandu (nepal) | NYTimes | - | 27.72 | 85.32 |
| 14 | Kathmandu | GeoNames | gn:P.PPLC | 27.7 | 85.32 |
| 15 | Kathmandu | DBpedia | db:Settlement | 27.7 | 85.33 |
| 16 | Kathmandu | FreeBase | fb:location.citytown | 27.7 | 85.37 |

ScaDS DRESDEN LEIPZIG

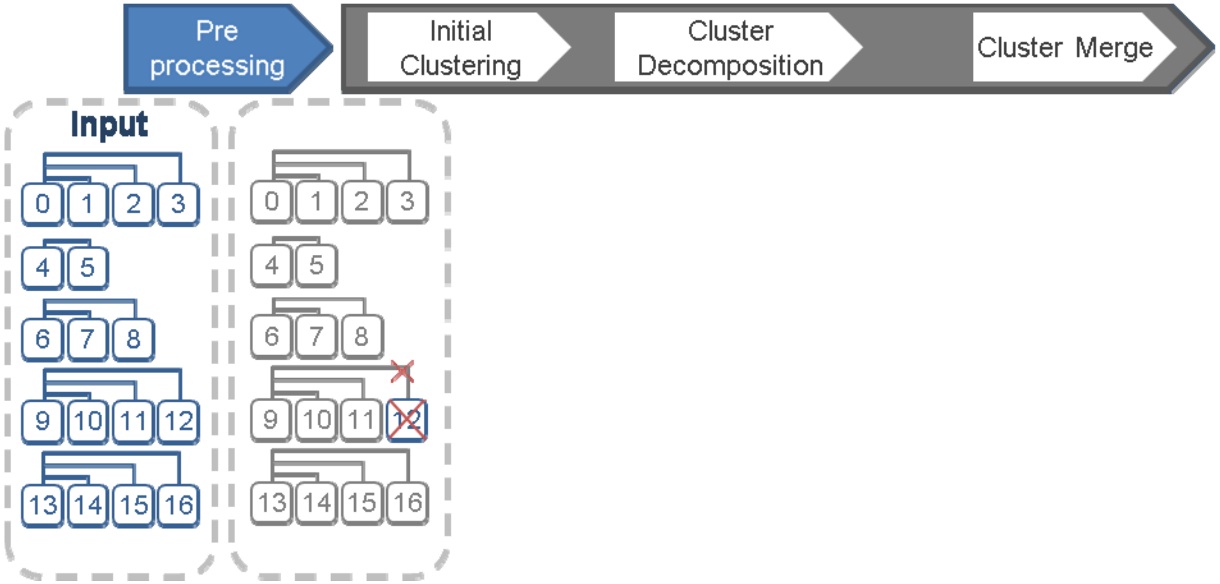
# HOLISTIC CLUSTERING (1): TYPE HARMONIZATION


- Each sources uses different entity types or no types
  - e.g., Settlement, AdministrativeRegion, Park, Country, Island
- Type harmonization using background knowledge
- Remove links for incompatible types
  - BodyOfWater != Settlement

| id | label | source | type |
|---|---|---|---|
| 0 | Lake Louise (Canada) | NYTimes | - |
| 1 | Lake Louise | GeoNames | BodyOfWater |
| 2 | Lake Louise, Alberta | DBpedia | Settlement |
| 3 | lake louise alberta | FreeBase | Settlement |
| 4 | Lake Louise (Alberta) | DBpedia | BodyOfWater |
| 5 | lake louise | FreeBase | BodyOfWater |
| 6 | Mystic (Conn) | NYTimes | - |
| 7 | N17632379615920 | FreeBase | - |
| 8 | Mystic | GeoNames | Settlement |
| 9 | Black Hill | GeoNames | Mountain |
| 10 | Black Hill | DBpedia | Mountain |
| 11 | Black Hill | LinkedGeoData | Mountain |
| 12 | Black Hill | LinkedGeoData | Mountain |
| 13 | katmandu (nepal) | NYTimes | - |
| 14 | Kathmandu | GeoNames | Settlement |
| 15 | Kathmandu | DBpedia | Settlement |
| 16 | Kathmandu | FreeBase | Settlement |

ScaDS DRESDEN LEIPZIG

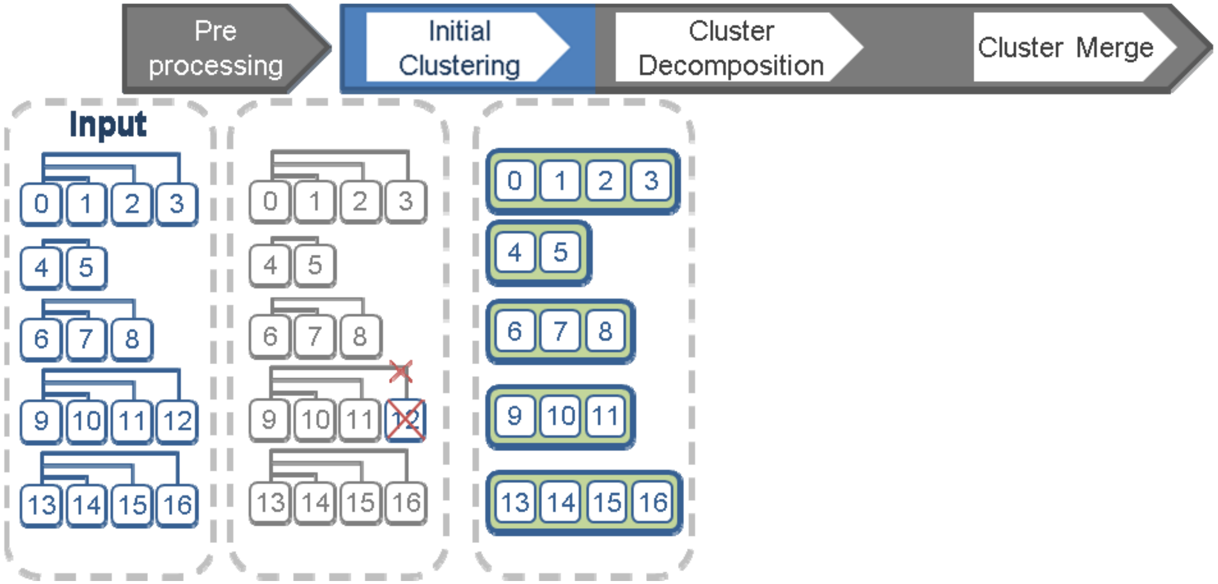
# HOLISTIC CLUSTERING (1): CHECK DUPLICATES




| | | | | | |
|---|---|---|---|---|---|
| 9 | Black Hill | GeoNames | Mountain | 53.54 | -1.89 |
| 10 | Black Hill | DBpedia | Mountain | 53.53 | -1.88 |
| 11 | Black Hill | ✓ LinkedGeoData | Mountain | 53.96 | -1.85 |
| 12 | Black Hill | ✗ LinkedGeoData | Mountain | 54.69 | -2.15 |

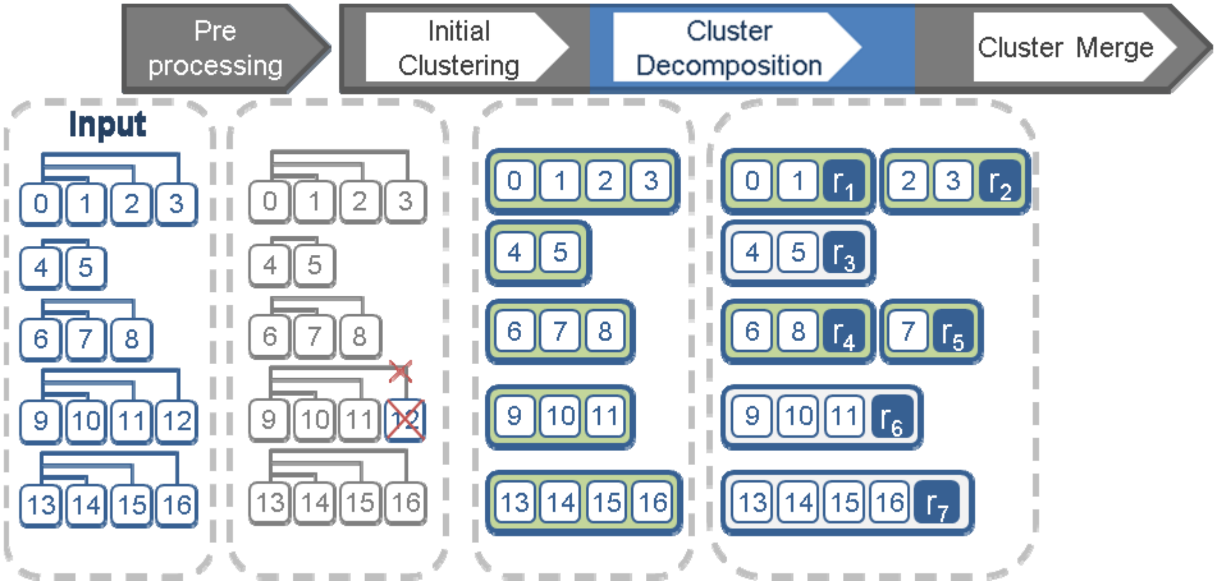
ScaDS DRESDEN LEIPZIG

# HOLISTIC CLUSTERING (2): CONN. COMPONENTS



ScaDS DRESDEN LEIPZIG

# HOLISTIC CLUSTERING (3): CLUSTER SPLITS




- Two split strategies for cluster refinement
  - Type-based grouping: ensure same entity type per cluster
  - Similarity-based refinement: eliminate entities with partially low intra-cluster similarity
- Determine cluster representatives with unified properties

ScaDS DRESDEN LEIPZIG

# HOLISTIC CLUSTERING (3): TYPE-BASED GROUPING


- Connected components may contain entities of different types, e.g., via intermediate entities without type
- Create sub-clusters for each type
- Assign entities without type to cluster of best-matching entity

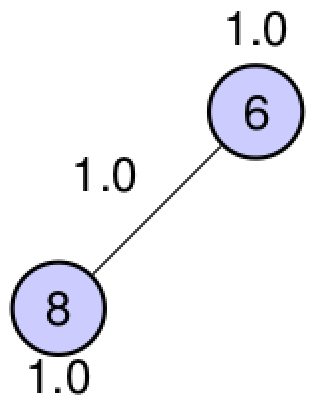
| | | | | | |
|---|---|---|---|---|---|
| 0 | Lake Louise (Canada) | NYTimes | - | 51.42 | -116.23 |
| 1 | Lake Louise | GeoNames | BodyOfWater | 51.41 | -116.23 |
| 2 | Lake Louise, Alberta | DBpedia | Settlement | - | - |
| 3 | lake louise alberta | FreeBase | Settlement | 51.43 | -116.16 |

ScaDS DRESDEN LEIPZIG

# HOLISTIC CLUSTERING (3): SIMILARITY-BASED SPLIT

Pre processing → Initial Clustering → Cluster Decomposition → Cluster Merge

- Iterative process to exclude low similarity entities from clusters
- Remove entity with lowest similarity from cluster if below threshold

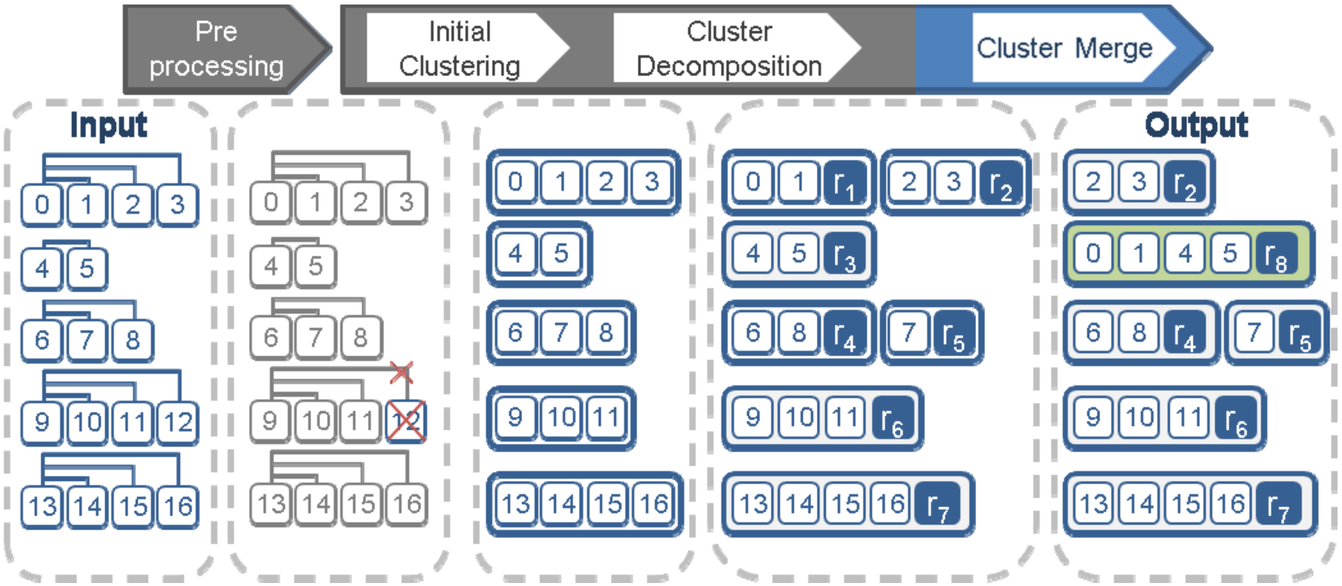



| | | | | | |
|---|---|---|---|---|---|
| 6 | Mystic (Conn) | NYTimes | - | 41.35 | -71.97 |
| 7 | N17632379615920 | FreeBase | - | 41.35 | -71.97 |
| 8 | Mystic | GeoNames | Settlement | 41.35 | -71.97 |

UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

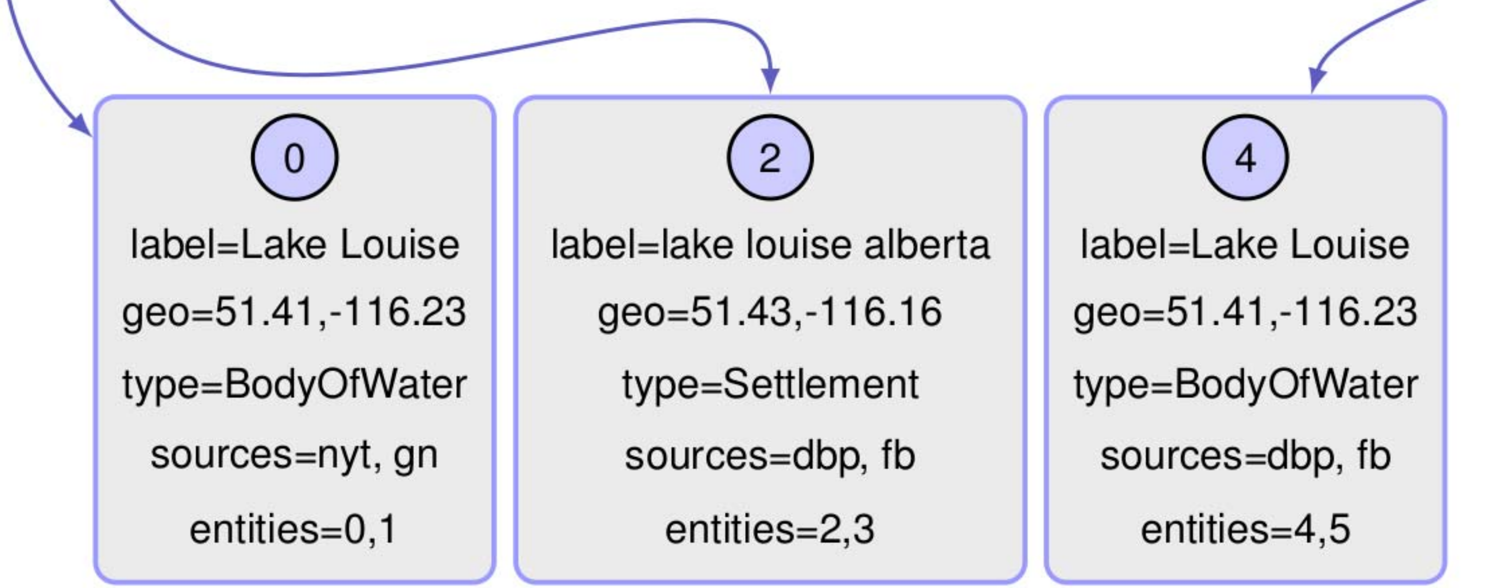
# HOLISTIC CLUSTERING (4): CLUSTER MERGE




- Iterative merging of similar clusters
  - Similarity of cluster representatives above threshold
  - only consider pairs of clusters of the same type and with entities from different sources

# HOLISTIC CLUSTERING (4): SAMPLE CLUSTER MERGE

| | | | | | |
|---|---|---|---|---|---|
| 0 | Lake Louise (Canada) | NYTimes | - | 51.42 | -116.23 |
| 1 | Lake Louise | GeoNames | BodyOfWater | 51.41 | -116.23 |
| 2 | Lake Louise, Alberta | DBpedia | Settlement | - | - |
| 3 | lake louise alberta | FreeBase | Settlement | 51.43 | -116.16 |
| 4 | Lake Louise (Alberta) | DBpedia | BodyOfWater | 51.41 | -116.23 |
| 5 | lake louise | FreeBase | BodyOfWater | 51.41 | -116.23 |



ScaDS DRESDEN LEIPZIG

# AGENDA

- Introduction
- Holistic data integration: use cases
- Holistic entity resolution for Linked Data
- Summary

ScaDS DRESDEN LEIPZIG

# SUMMARY

- Big Data Integration
  - Big Data poses new requirements for data integration (variety, volume, velocity, veracity)
  - blocking and parallel matching for improved scalability, e.g. utilizing Hadoop-based approaches such as Dedoop
- Holistic data integration
  - combined integration of many sources (metadata + instances)
  - clustering-based rather than mapping-based approaches
  - utilization of corpora with datasets, schemas, and mappings
  - construction of and linking to large knowledge graphs
  - many research opportunities

ScaDS DRESDEN LEIPZIG

# REFERENCES


- S. Balakrishnan, A. Halevy, et al: *Applying WebTables in Practice*. Proc. CIDR 2015
- Z. Bellahsene, A. Bonifati, E. Rahm (eds.). *Schema Matching and Mapping*. Springer-Verlag,2011
- P. Christen: *Data Matching*. Springer, 2012
- A. Doan, A. Y. Halevy, Z.G. Ives: *Principles of Data Integration*. Morgan Kaufmann 2012
- X.L. Dong, D. Srivastava: *Big Data Integration*. Synthesis Lectures on Data Management, Morgan & Claypool 2015
- J. Eberius, M. Thiele, K. Braunschweig, W. Lehner: *Top-k entity augmentation using consistent set covering*. Proc. SSDM 2015
- H. Elmeleegy, J. Madhavan, A.Y. Halevy: *Harvesting Relational Tables from Lists on the Web*. PVLDB 2009
- R. Gupta, A. Halevy, X.Wang, S. Whang, F. Wu: *Biperpedia: An Ontology for Search Applications*. PVLDB 2014
- H. Köpcke, A. Thor, E. Rahm: *Comparative evaluation of entity resolution approaches with FEVER*. Proc. 35th Intl. Conference on Very Large Databases (VLDB), 2009
- H. Köpcke, E. Rahm: *Frameworks for entity matching: A comparison*. Data & Knowledge Engineering 2010
- 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) / Proceedings of the VLDB Endowment 3(1), 2010
- H. Köpcke, A. Thor, S. Thomas, E. Rahm: *Tailoring entity resolution for matching product offers*. Proc. EDBT 2012: 545-550

ScaDS DRESDEN LEIPZIG

# REFERENCES (2)


- 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
- L. Kolb, A. Thor, E. Rahm: *Multi-pass Sorted Neighborhood Blocking with MapReduce*. Computer Science - Research and Development 27(1), 2012
- L. Kolb, A. Thor, E. Rahm: *Don't Match Twice: Redundancy-free Similarity Computation with MapReduce*. Proc. 2nd Intl. Workshop on Data Analytics in the Cloud (DanaC), 2013
- L. Kolb, Z. Sehili, E. Rahm: *Iterative Computation of Connected Graph Components with MapReduce*. Datenbank-Spektrum 14(2): 107-117 (2014)
- M. Nentwig, T. Soru, A. Ngonga, E. Rahm: *LinkLion: A Link Repository for the Web of Data*. Proc ESWC 2014
- M. Nentwig, M. Hartung, A. Ngonga, E. Rahm: *A Survey of Current Link Discovery Frameworks*. Semantic Web Journal, 2016
- M. Nentwig, A. Groß, E. Rahm: *Holistic entity resolution for Linked Data*. Tech. report, 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
- E. Rahm: *The case for holistic data integration*. Proc. ADBIS conf., LNCS, 2016
- S. Raunich, E. Rahm: *Target-driven Merging of Taxonomies with ATOM*. Information Systems, 2014