Lazy Big Data Integration

Prof. Dr. Andreas Thor
Hochschule für Telekommunikation Leipzig (HfTL)

Martin-Luther-Universität Halle-Wittenberg
16.12.2016

Agenda

- Data Integration
  - Data analytics for domain-specific questions
  - Use cases: Bibliometrics & Life Sciences
- Big Data Integration
  - Techniques for efficient big data management
  - Exploiting cloud infrastructures (MapReduce, NoSQL data stores)
- Lazy Big Data Integration
  - (Efficient and) effective goal-oriented data integration
  - Integrated analytical approach for big data analytics

Use Case: Bibliometrics

- Does the peer review process actually work?
  Does it select the „best“ papers?

- Data from reviewing process (e.g., Easy Chair)
  - Bibliographic information (title, authors, …) of submitted papers
  - Review score(s) incl. editorial decision
- Data from bibliographic data sources (e.g., Google Scholar)
  - Accepted papers and rejected papers that are published elsewhere
  - Number of citations

- Determine **covariance** between review score(s) and #citations

Data Integration

- **Combining data** residing at different sources and providing the user with a **unified view** of this data
  - Added value by linking & merging data
  - Queries that can only be answered using multiple sources
- Schema Matching
  - Finding mappings of corresponding attributes
- **Entity Matching**
  - Finding equivalent data objects

```
@inproceedings{DBLP:conf/xsym/RahmTA07,
  author    = {Erhard Rahm and
               Andreas Thor and
               David Aumueller},
  title     = {Dynamic Fusion of Web Data},
  booktitle = {Database and XMLTechnologies, 5t
               XSym 2007, Vienna, Austria, Sept
  pages     = {14--16},
  year      = {2007},
  crossref  = {DBLP:conf/xsym/2007},
  url       = {http://dx.doi.org/10.1007/978-3-
  doi       = {10.1007/978-3-540-75288-2_2},
  timestamp = {Fri, 14 Sep 2007 09:12:45 +0200}
  biburl    = {http://dblp.uni-trier.de/rec/bib
  bibsource = {dblp computer science bibliograp
}
```

Dynamic fusion of web data
E Rahm, A **Thor**, D Aumueller - International XML Database Symposium, 2007 - Sprin
Abstract Mashups exemplify a workflow-like approach to dynamically integrate data an services from multiple web sources. Such integration workflows can build on existing services for web search, entity search, database querying, and information extraction a
Cited by 14 Related articles All 13 versions Cite Save

[PDF] **Dynamic Fusion of Web Data**: Beyond Mashups
E Rahm, A **Thor**, D Aumüller - Proc. of XSym07, 2007 - dbs.uni-leipzig.de
Page 1. **Dynamic Fusion of Web Data**: Beyond Mashups Erhard Rahm Andreas **Tho** David Aumüller http://dbs.uni-leipzig.de 24th September, 2007 Top VLDB '97 Pubs: Go Scholar's Top-5 Page 2. Google Scholar's Top-5 (2) ... more GS quality problems Dupli
Cited by 3 Related articles All 2 versions Cite Save More

[CITATION] **Dynamic Fusion of Web Data**, University of Leipzig, Ger
E Rahm, A **Thor**, D Aumueller - 2007
Cited by 3 Related articles Cite Save

Data Quality

- Can/should we use Google Scholar citations for ranking …
  - Papers
  - Researchers
  - Institutions
  - etc.

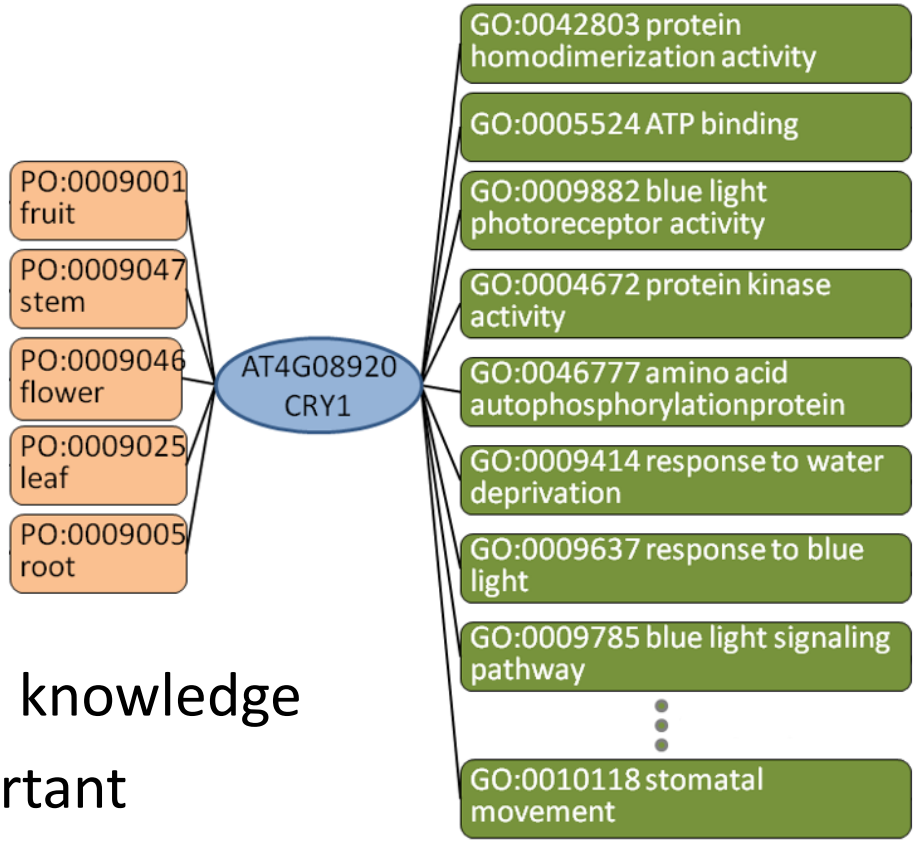
Citation indices	All	Since 2011
Citations	1202	970
h-index	19	18
i10-index	29	23

	Google Scholar	Web of Science
Coverage	Huge	Limited
Data quality	Medium (fully automatic)	High (manually curated)
Costs	Free	Expensive

- **Convergent validity** of citation analyses?
  - Comparison of analysis results for source overlap

Use Case: Life Sciences

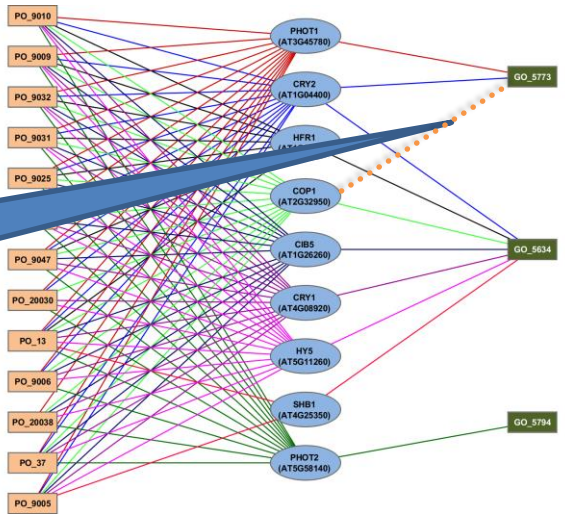
- Gene Annotation Graph
  - Genes are annotated with Gene Ontology (GO) and Plant Ontology (PO) terms



- Links form a **graph** that captures meaningful biological knowledge
- Sense making of graph is important

- **Prediction** of new annotations
  - hypothesis for wet lab experiments

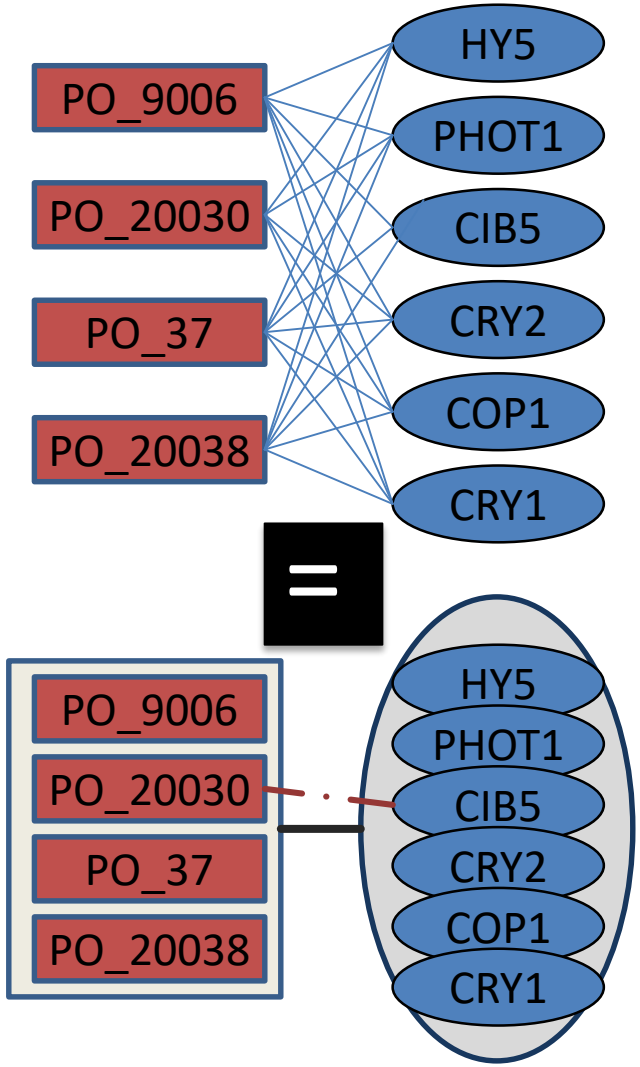
Is this annotation likely to be added in the future?



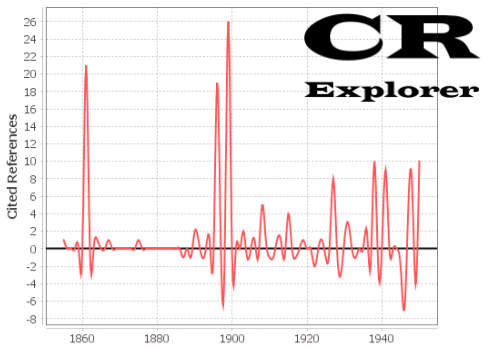
Graph Summarization + Link Prediction

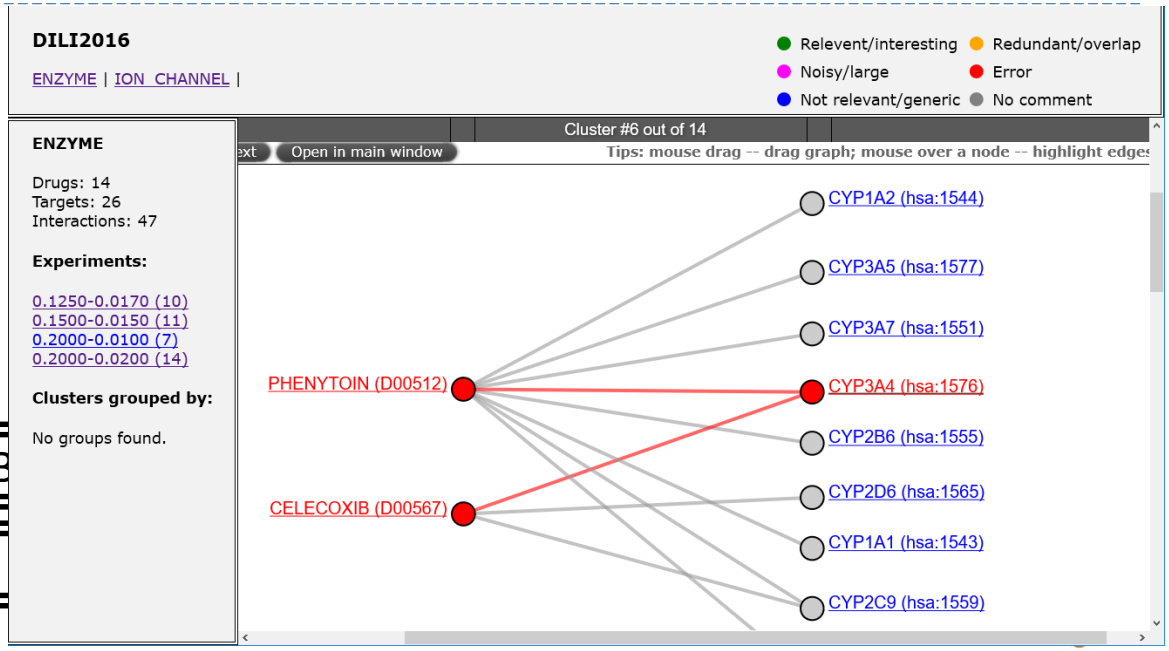
- Graph summary = Signature + Corrections
- Signature: graph pattern / structure
  - Super nodes = partitioning of nodes
  - Super edges = edges between super nodes = all edges between nodes of super nodes
- Corrections: edges e between individual nodes
  - Additions: e $\in$ G but e $\notin$ signature
  - Deletions: e $\notin$ G but e $\in$ signature

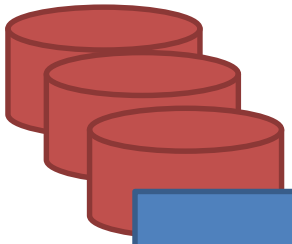
- *p*(PO_20030, CIB5) $\approx$ 0.96
  - High prediction score because it is the "only missing piece" to a "perfect 4x6 pattern"



(Big) Data Analytics Pipeline





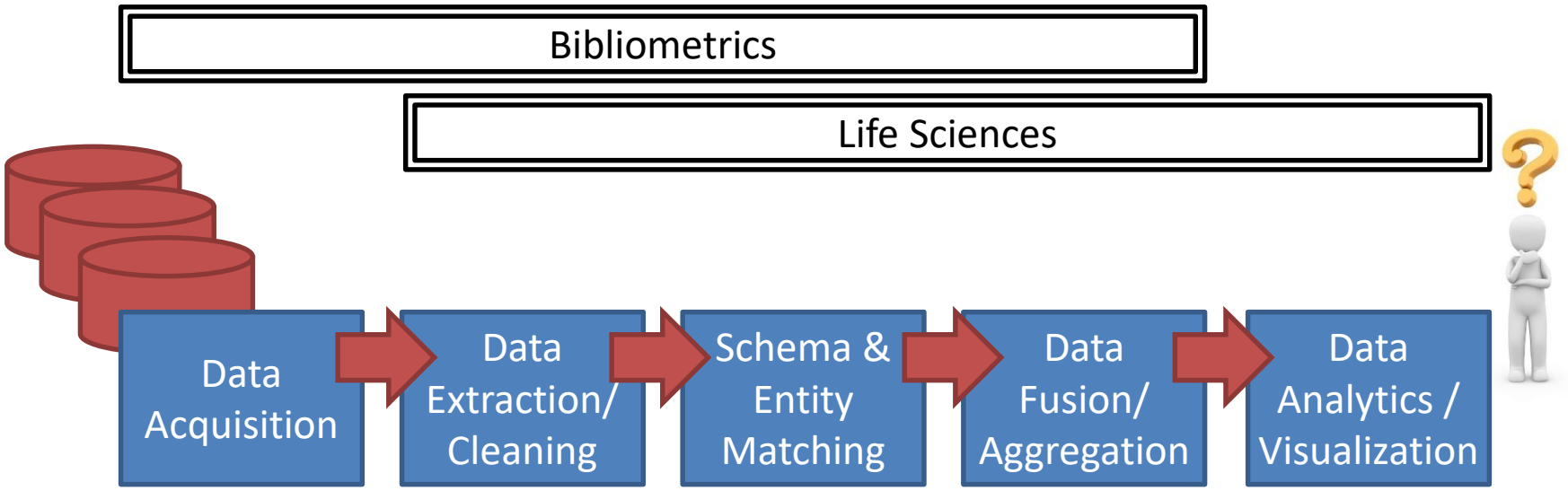


Data Acquisition → Data Extraction/ Cleaning → Schema & Entity Matching → Data Fusion/ Aggregation → Data Analytics / Visualization

Agenda

- Data Integration
  - Data analytics for domain-specific questions
  - Use cases: Bibliometrics & Life Sciences
- Big Data Integration
  - Techniques for efficient big data management
  - Exploiting cloud infrastructures (MapReduce, NoSQL data stores)
- Lazy Big Data Integration
  - (Efficient and) effective goal-oriented data integration
  - Integrated analytical approach for big data analytics

(Big) Data Analytics Pipeline



Query Strategies

Product Code Extraction

Load Balancing

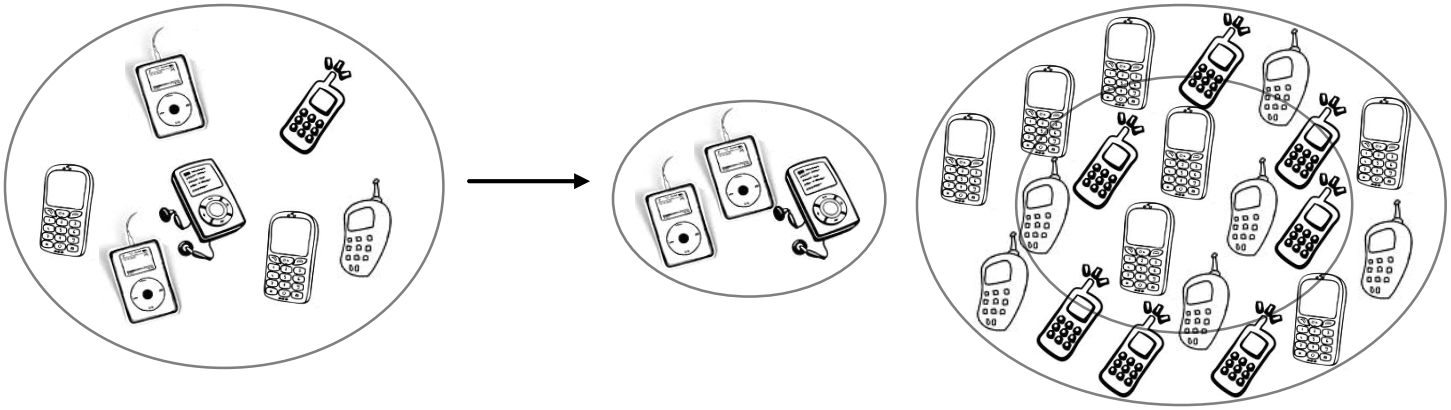
NoSQL Sharding

AutoShard

Dedoop (Deduplication with Hadoop)

How to speed up entity matching?

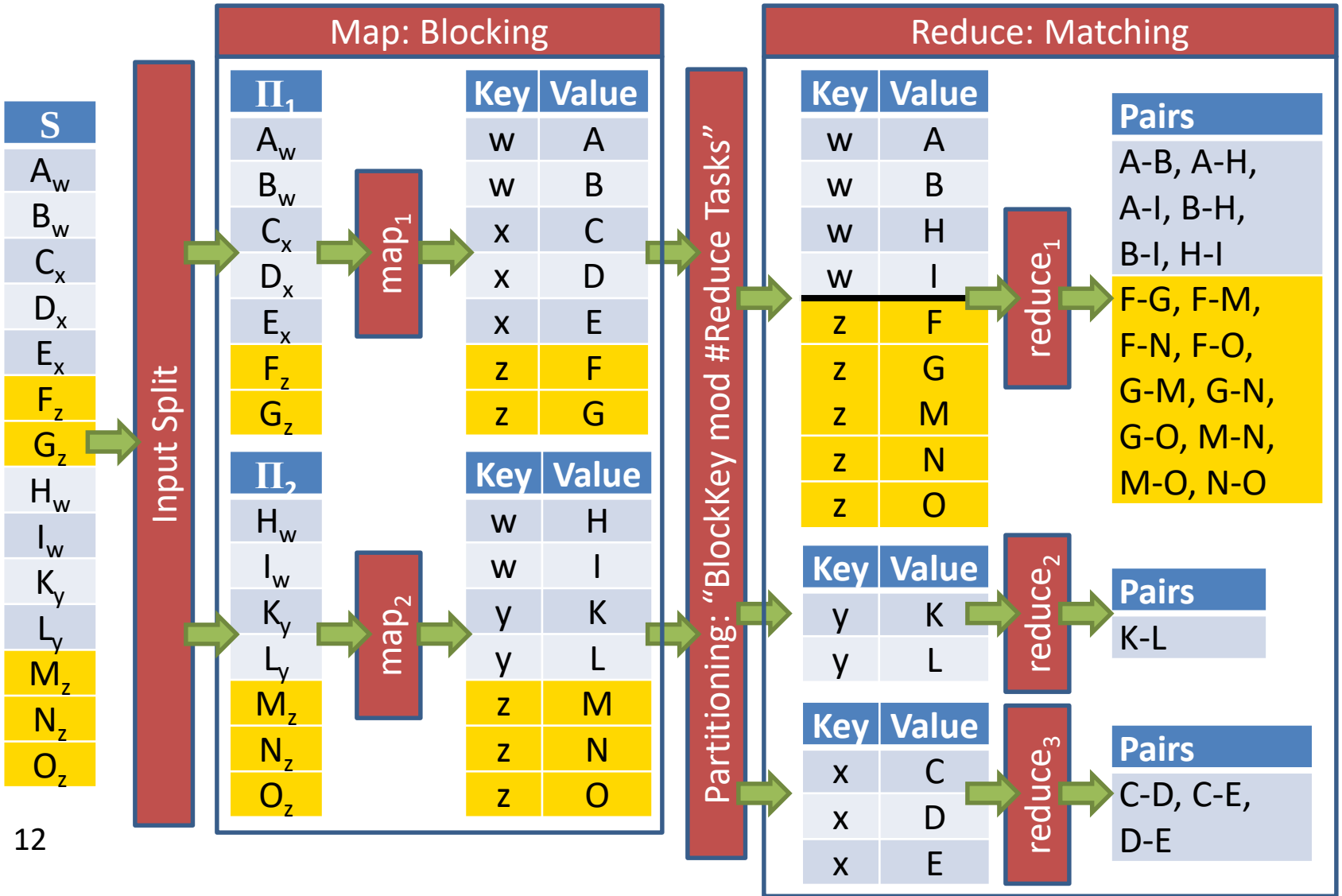
- Entity matching is expensive (due to pair-wise comparisons)
- Blocking to reduce search space
  - Group similar entities within blocks based on blocking key
  - Restrict matching to entities from the same block



- Parallelization
  - Split match computation in sub-tasks to be executed in parallel
  - Exploitation of cloud infrastructures and frameworks like MapReduce

Blocking + MapReduce: Naïve

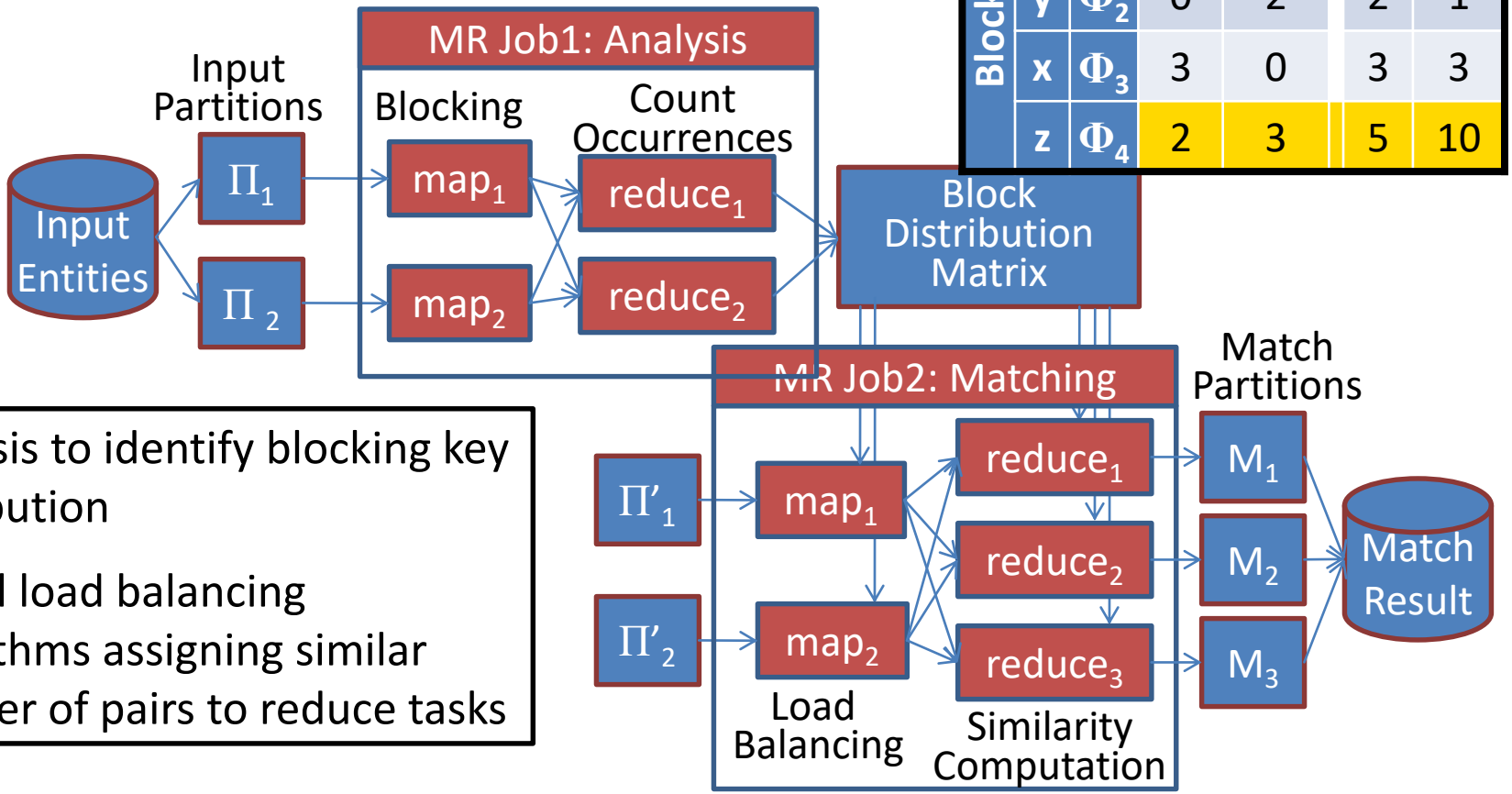
- **Data skew** leads to unbalanced workload



Load Balancing for MR-based EM

Partition	Π_1							Π_2						
Entity	A	B	C	D	E	F	G	H	I	K	L	M	N	O
Blocking Key	w	w	x	x	x	z	z	w	w	y	y	z	z	z

			Partition		Overall	
			Π_1	Π_2	#E	#P
Blocks	w	Φ_1	2	2	4	6
	y	Φ_2	0	2	2	1
	x	Φ_3	3	0	3	3
	z	Φ_4	2	3	5	10



Analysis to identify blocking key distribution

Global load balancing algorithms assigning similar number of pairs to reduce tasks

BlockSplit

- Large blocks split into m sub-blocks
  - according to m input partitions
  - large if $\#P_{Block} > \#P_{Overall} / \#Reducer$
- Two types of match tasks
  - Single (small blocks and sub-blocks)
  - Two sub-blocks
- Greedy load balancing
  - Sort match tasks by number of pairs in descending order
  - Assign match task to reducer with lowest number of pairs
- Example
  - r=3 reduce tasks, split Φ_4 in m=2 sub-blocks
  - Φ_4's match tasks: $\Phi_{4.1}$, $\Phi_{4.2}$, and $\Phi_{4.1\times2}$

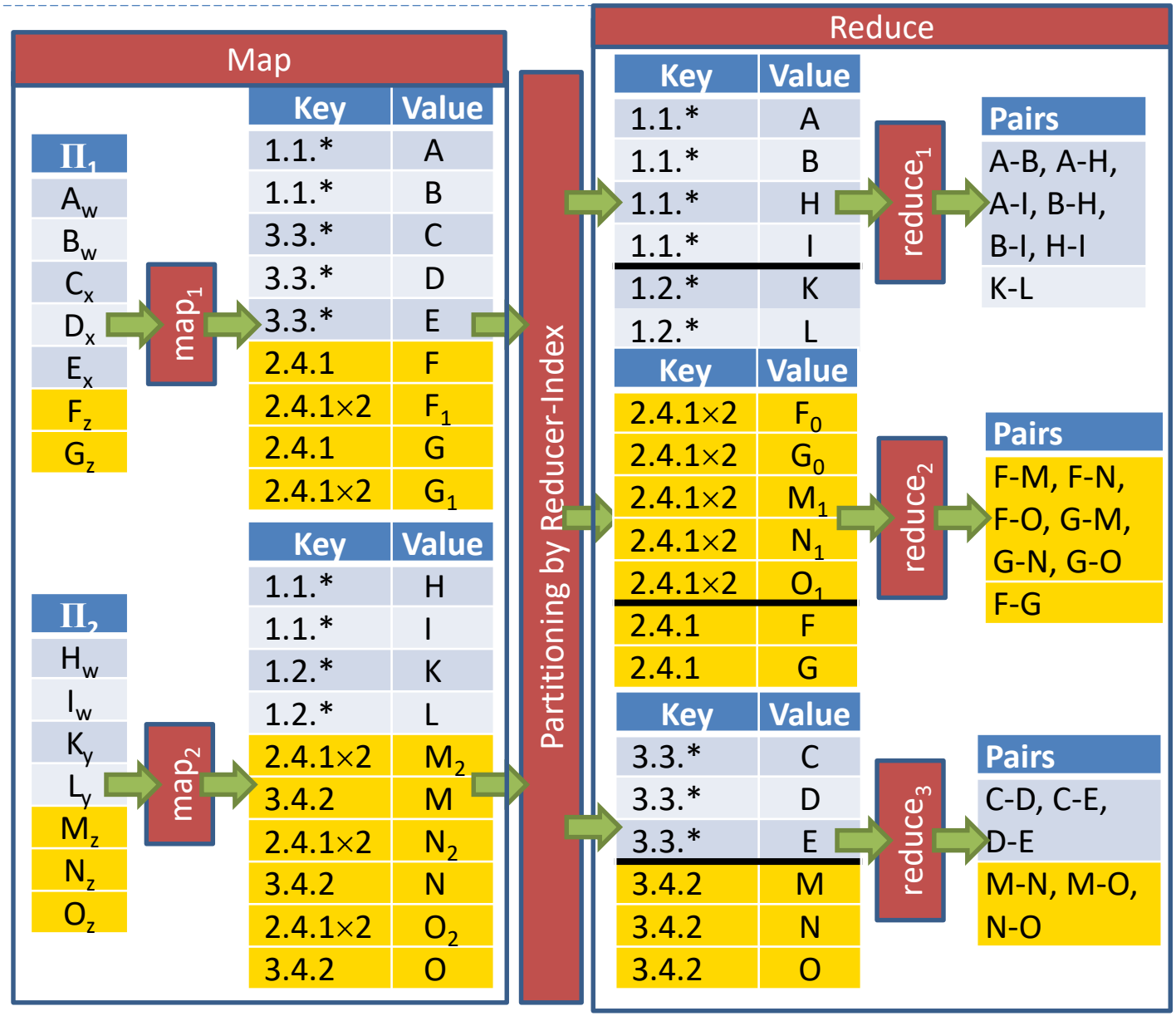
Blocks			Partition Π_1	Partition Π_2	Overall #E	Overall #P
	w	Φ_1	2	2	4	6
	y	Φ_2	0	2	2	1
	x	Φ_3	3	0	3	3
	z	Φ_4	2	3	5	10

Match Tasks	#P	Reducer
Φ_1	6	
$\Phi_{4.1\times2}$	6	
Φ_3	3	
$\Phi_{4.2}$	3	
Φ_2	1	
$\Phi_{4.1}$	1	

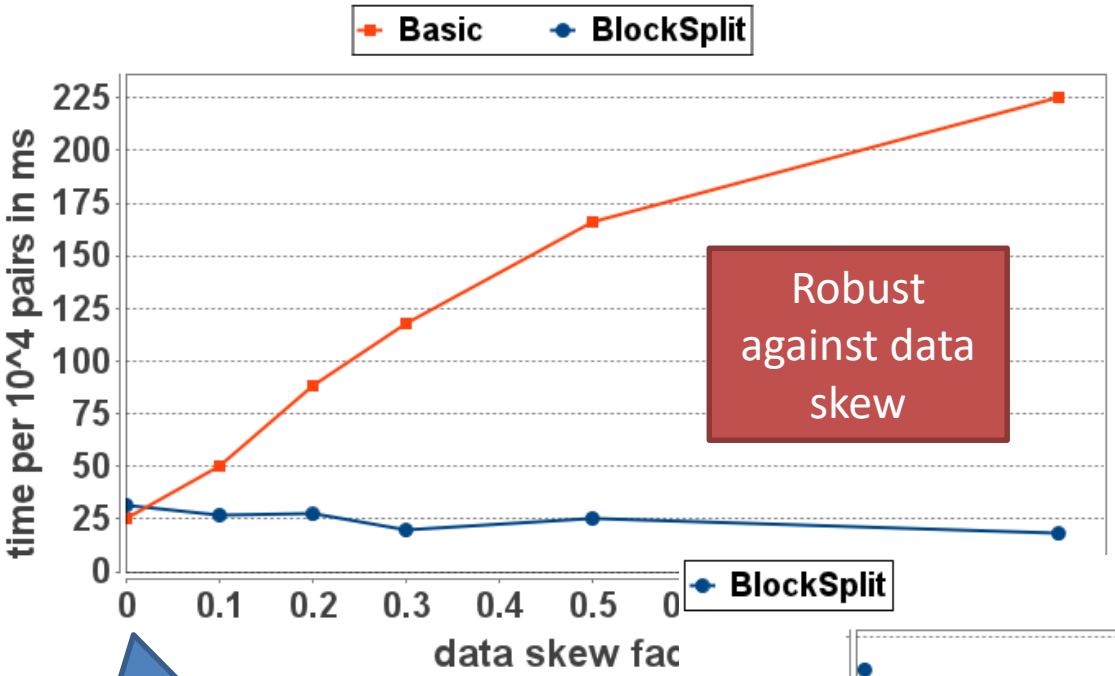
BlockSplit: MR-Dataflow

MapReduce Techniques

- MapKey = ReducerIndex + MatchTask
- Replicate entities of sub-blocks

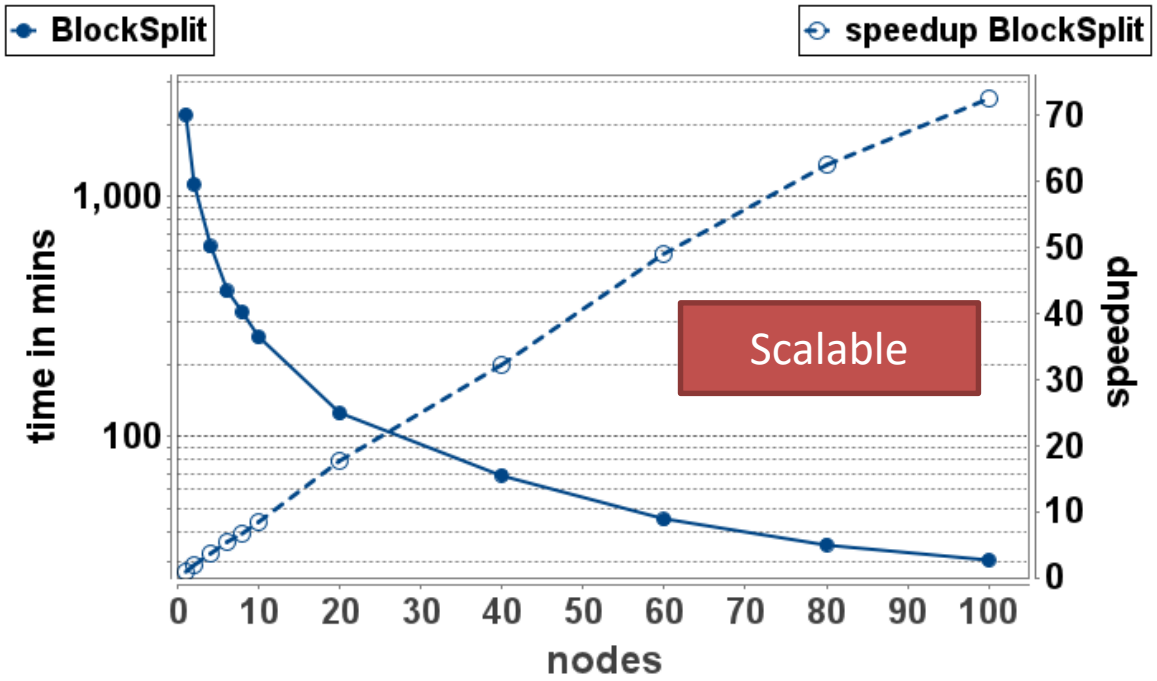


Evaluation: Robustness + Scalability



Robust against data skew

„Uniform distribution“

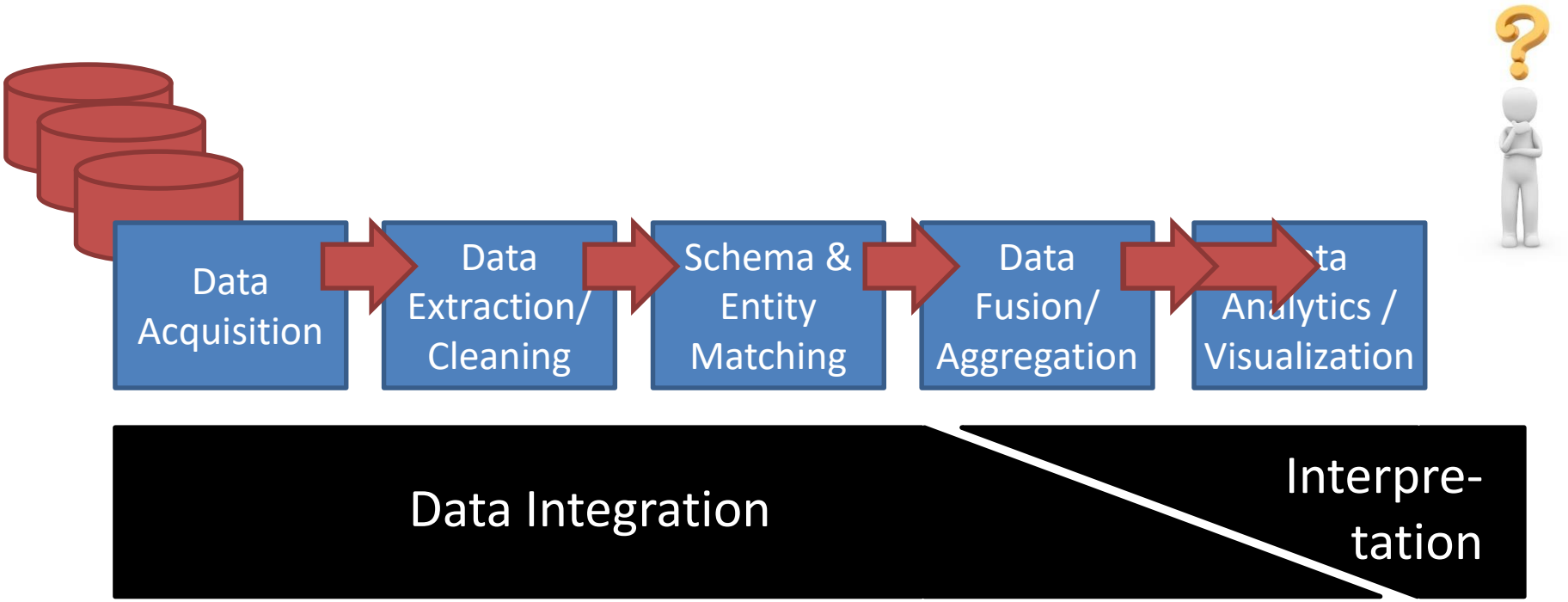


Scalable

Agenda

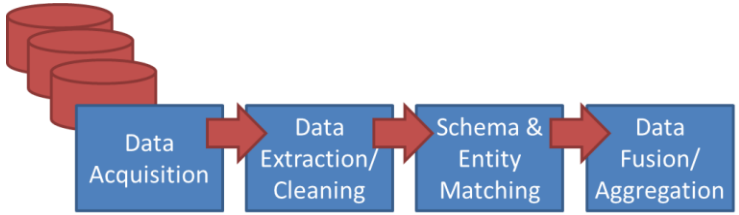
- Data Integration
  - Data analytics for domain-specific questions
  - Use cases: Bibliometrics & Life Sciences
- Big Data Integration
  - Techniques for efficient big data management
  - Exploiting cloud infrastructures (MapReduce, NoSQL data stores)
- Lazy Big Data Integration
  - (Efficient and) effective goal-oriented data integration
  - Integrated analytical approach for big data analytics

(Big) Data Analytics Pipeline



Citation Analysis Pipeline

- For a given set of Bibtex entries
  - Find matching Google Scholar entries
  - Determine aggregated citation counts



- Analytical questions for a researcher
  - Complete publication list + #citations
  - Top-5 publications
  - H-Index, Average Number of citations
- Analytical questions for comparing
  - Institutions
  - Research fields
  - …

```
@inproceedings{DBLP:conf/xsym/RahmTA07,
  author    = {Erhard Rahm and
               Andreas Thor and
               David Aumueller},
  title     = {Dynamic Fusion of Web Data},
  booktitle = {Database and XMLTechnologies, 5th International {XML}
               XSym 2007, Vienna, Austria, September 23-24, 2007, Pr
  pages     = {14--16},
  year      = {2007},
  crossref  = {DBLP:conf/xsym/2007},
  url       = {http://dx.doi.org/10.1007/978-3-540-75288-2_2},
  doi       = {10.1007/978-3-540-75288-2_2},
  timestamp = {Fri, 14 Sep 2007 09:12:45 +0200},
  biburl    = {http://dblp.uni-trier.de/rec/bib/conf/xsym/RahmTA07},
  bibsource = {dblp computer science bibliography, http://dblp.org}
}
```

Dynamic fusion of web data
E Rahm, A Thor, D Aumueller - International XML Database Symposium, 2007 - Springer
Abstract Mashups exemplify a workflow-like approach to dynamically integrate data and services from multiple web sources. Such integration workflows can build on existing services for web search, entity search, database querying, and information extraction and
Cited by 14 Related articles All 13 versions Cite Save

[PDF] **Dynamic Fusion of Web Data**: Beyond Mashups
E Rahm, A Thor, D Aumüller - Proc. of XSym07, 2007 - dbs.uni-leipzig.de
Page 1. **Dynamic Fusion of Web Data**: Beyond Mashups Erhard Rahm Andreas **Thor**, David Aumüller http://dbs.uni-leipzig.de 24th September, 2007 Top VLDB '97 Pubs: Google Scholar's Top-5 Page 2. Google Scholar's Top-5 (2) ... more GS quality problems Duplicates
Cited by 3 Related articles All 2 versions Cite Save More

[CITATION] **Dynamic Fusion of Web Data**, University of Leipzig, Germany
E Rahm, A Thor, D Aumueller - 2007
Cited by 3 Related articles Cite Save

„Lazy Machine“: Effectiveness

- Do the right thing! Do only things that are needed!
  - Priorization / filtering of data objects to be processed
- Example: Top-5 publications of a researcher
  - Entity Matching for highly cited Google Scholar entries
  - Cutoff data that does not contribute to the analytical result (anymore)
  - „does not“ → „is not likely to“

```
@inproceedings{DBLP:conf/xsym/RahmTA07,
  author    = {Erhard Rahm and
               Andreas Thor and
               David Aumueller},
  title     = {Dynamic Fusion of Web Data},
  booktitle = {Database and XMLTechnologies,
               XSym 2007, Vienna, Austria, Sep
  pages     = {14--16},
  year      = {2007},
  crossref  = {DBLP:conf/xsym/2007},
  url       = {http://dx.doi.org/10.1007/978-3
  doi       = {10.1007/978-3-540-75288-2_2},
  timestamp = {Fri, 14 Sep 2007 09:12:45 +0200
  biburl    = {http://dblp.uni-trier.de/rec/bi
  bibsource = {dblp computer science bibliogra
}
```

Dynamic fusion of web data
E Rahm, A Thor, D Aumueller - International XML Database Symposium, 2007 - Spr
Abstract Mashups exemplify a workflow-like approach to dynamically integrate data a
services from multiple web sources. Such integration workflows can build on existing
services for web search, entity search, database querying, and information extraction
Cited by 14 Related articles All 13 versions Cite Save

[PDF] **Dynamic Fusion of Web Data**: Beyond Mashups
E Rahm, A Thor, D Aumüller - Proc. of XSym07, 2007 - dbs.uni-leipzig.de
Page 1. **Dynamic Fusion of Web Data**: Beyond Mashups Erhard Rahm Andreas Th
David Aumüller http://dbs.uni-leipzig.de 24th September, 2007 Top VLDB '97 Pubs:
Scholar's Top-5 Page 2. Google Scholar's Top-5 (2) ... more GS quality problems Du
Cited by 3 Related articles All 2 versions Cite Save More

[CITATION] **Dynamic Fusion of Web Data**, University of Leipzig, Ge
E Rahm, A Thor, D Aumueller - 2007
Cited by 3 Related articles Cite Save

- Pipeline stages
  - Data Akquisition: query strategies
  - Data Extraction: on-demand
  - Data Matching: relevant entities only

„Lazy User“: Data Quality

- Automatic data integration does not give 100% data quality
  - Data acquisition might miss relevant data
  - Matching is imperfect (precision, recall)
  - …
- Pipeline & integrated result should effectively point the user to the “weak points”
- Examples
  - What (non-)matching pairs have the most effect on the analytical result?
  - Outlier detection → What pipeline stage caused the effect?

Lazy Big Data Integration

- Integrated approach for both
  - Data integration workflow
  - Analytical query

- Current work based on Gradoop (Graph Analytics on Hadoop)
  - Graph model + operators for analytical pipelines
  - Efficient execution in distributed environment

- Next steps
  - Operators for complex analytical queries / statistics (e.g., h-index)
  - Data provenance model for measuring the impact of data objects to specific results

Summary

- Data Integration
  - Data analytics for domain-specific questions
  - Use cases: Bibliometrics & Life Sciences
- Big Data Integration
  - Techniques for efficient big data management
  - Exploiting cloud infrastructures (MapReduce, NoSQL data stores)
- Lazy Big Data Integration
  - (Efficient and) effective goal-oriented data integration
  - Integrated analytical approach for big data analytics

Thank you!