UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

SCALABLE AND PRIVACY-PRESERVING DATA INTEGRATION

ERHARD RAHM,
UNIVERSITY OF LEIPZIG

www.scads.de

UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

UNIVERSITY OF LEIPZIG

- Founded in 1409
- Now about 30.000 students in 14 faculties
- Computer science
  - 13 professorships and 2 junior professors
  - 150 PhD students and postdocs (120 by third party funding)

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)

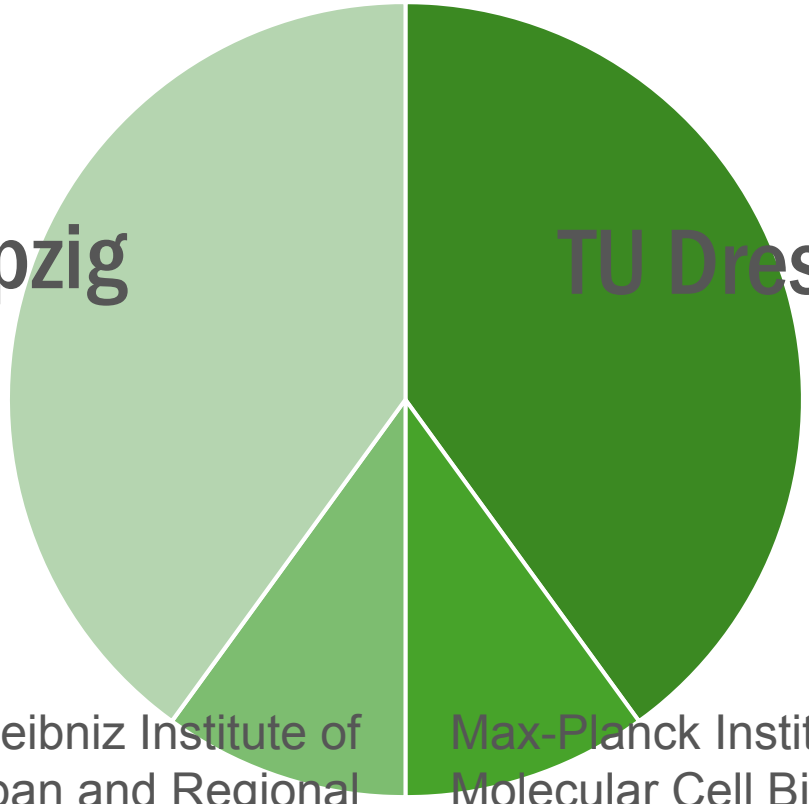
- scientific coordinators: Nagel (TUD), Rahm (UL)
- start: Oct. 2014
- duration: 4 years (option for 3 more years)
- initial funding: ca. 5.6 Mio. Euro

ScaDS DRESDEN LEIPZIG GOALS

- Bundling and advancement of existing expertise on Big Data
- Development of Big Data Services and Solutions
- Big Data Innovations

ScaDS DRESDEN LEIPZIG

FUNDED INSTITUTES



UNIVERSITÄT LEIPZIG

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

UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

RESEARCH PARTNERS

- Data-intensive computing W.E. Nagel
- Data quality / Data integration E. Rahm
- Databases W. Lehner, E. Rahm
- Knowledge extraction/Data mining
  C. Rother, P. Stadler, G. Heyer
- Visualization
  S. Gumhold, G. Scheuermann
- Service Engineering, Infrastructure
  K.-P. Fähnrich, W.E. Nagel, M. Bogdan

UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

APPLICATION COORDINATORS

- Life sciences G. Myers
- Material / Engineering sciences M. Gude
- Environmental / Geo sciences J. Schanze
- Digital Humanities G. Heyer
- Business Data B. Franczyk

UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

BIG DATA SUMMER SCHOOL 2016 IN LEIPZIG

Date: 11th – 15th of July
Courses:

- Storage/ NoSQL
- Processing (Spark/Flink)
- Graph Analytics
- Data Integration

Supervised hands-on sessions – three domains (Text, Bio, Finance)

Online Courses for preparation

Prerequisites

- good Java programming skills (for hands-on sessions)
- good English skills

Fees and registration: **www.scads.de/summerschool-2016**

ScaDS DRESDEN LEIPZIG

OPEN POSITIONS IN LEIPZIG

PhD students (m/f) in Big Data Center ScaDS (www.scads.de)

Topics:

- Big Data Integration/ Graph-based Data Integration
- Big Data Matching & Big Data Quality
- Privacy-preserving data mining

Requirements:

- excellent Master/Diploma degree in computer science
- very good database skills
- very good Englisch skills (speech & writing)
- research interest

ScaDS DRESDEN LEIPZIG AGENDA

- ScaDS Dresden/Leipzig
- Big Data Integration
  - Scalable entity resolution / link discovery
  - Large-scale schema/ontology matching
  - Holistic data integration
- Privacy-preserving record linkage
  - Privacy for Big Data
  - PPRL basics
  - Scalable PPRL
- Graph-based data integration and analytics
  - Introduction
  - Graph-based data integration / business intelligence (BIIIG)
  - Hadoop-based graph analytics (GRADOOP)

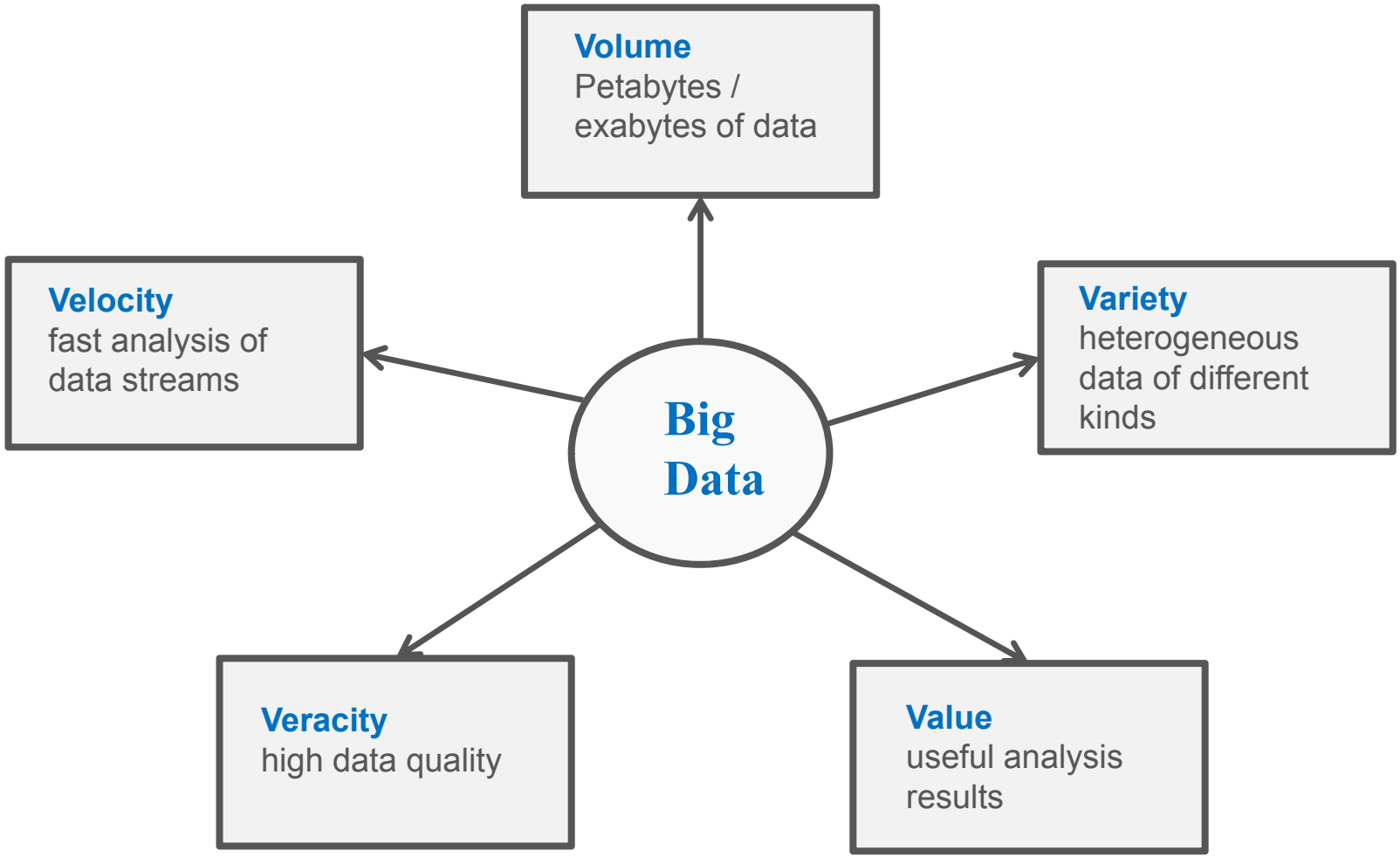
ScaDS DRESDEN LEIPZIG

AGENDA PART I (BIG DATA INTEGRATION)

- Introduction
  - Big Data
  - Data Quality
- Scalable entity resolution / link discovery
  - Introduction
  - Comparison of ER frameworks
  - Comparison of Frameworks for Link Discovery
  - Use case: Matching of product offers
  - Hadoop-based entity resolution (Dedoop)
  - Load balancing to deal with data skew
- Large-scale schema/ontology matching
- Holistic data integration
- Summary

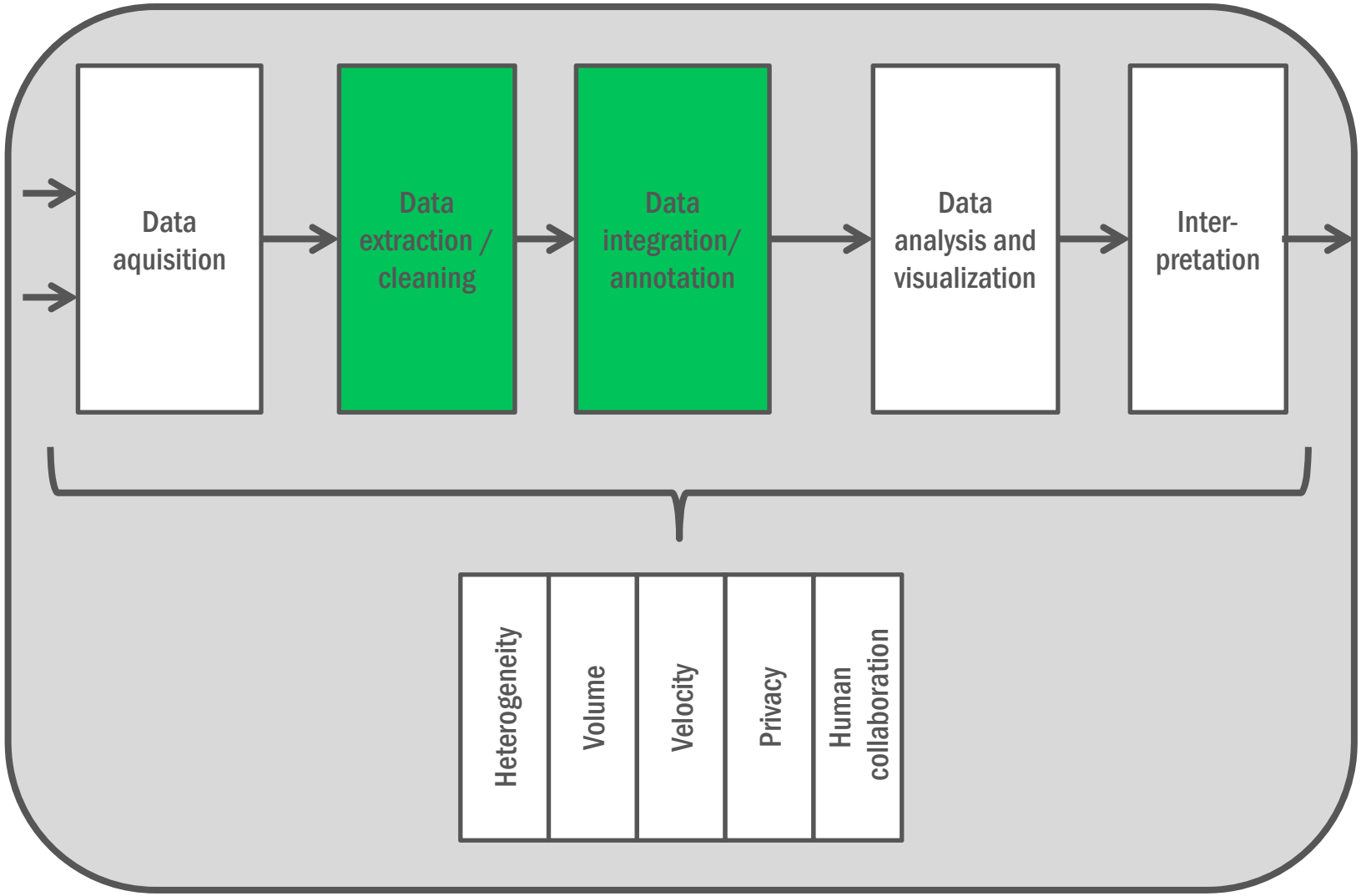
ScaDS DRESDEN LEIPZIG

BIG DATA CHALLENGES



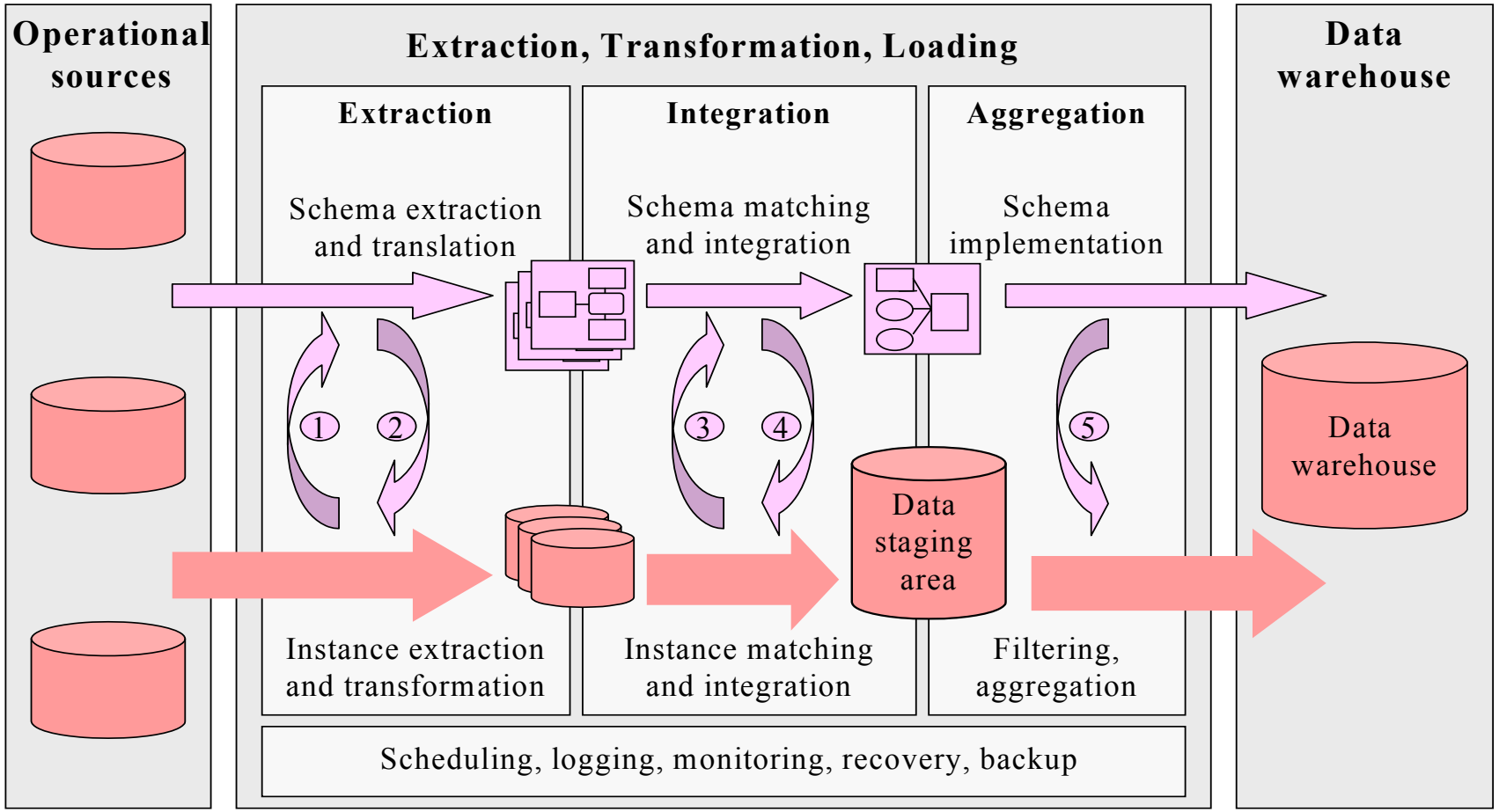
ScaDS DRESDEN LEIPZIG

BIG DATA ANALYSIS PIPELINE



ScaDS DRESDEN LEIPZIG

ETL PROCESS FOR DATA WAREHOUSES



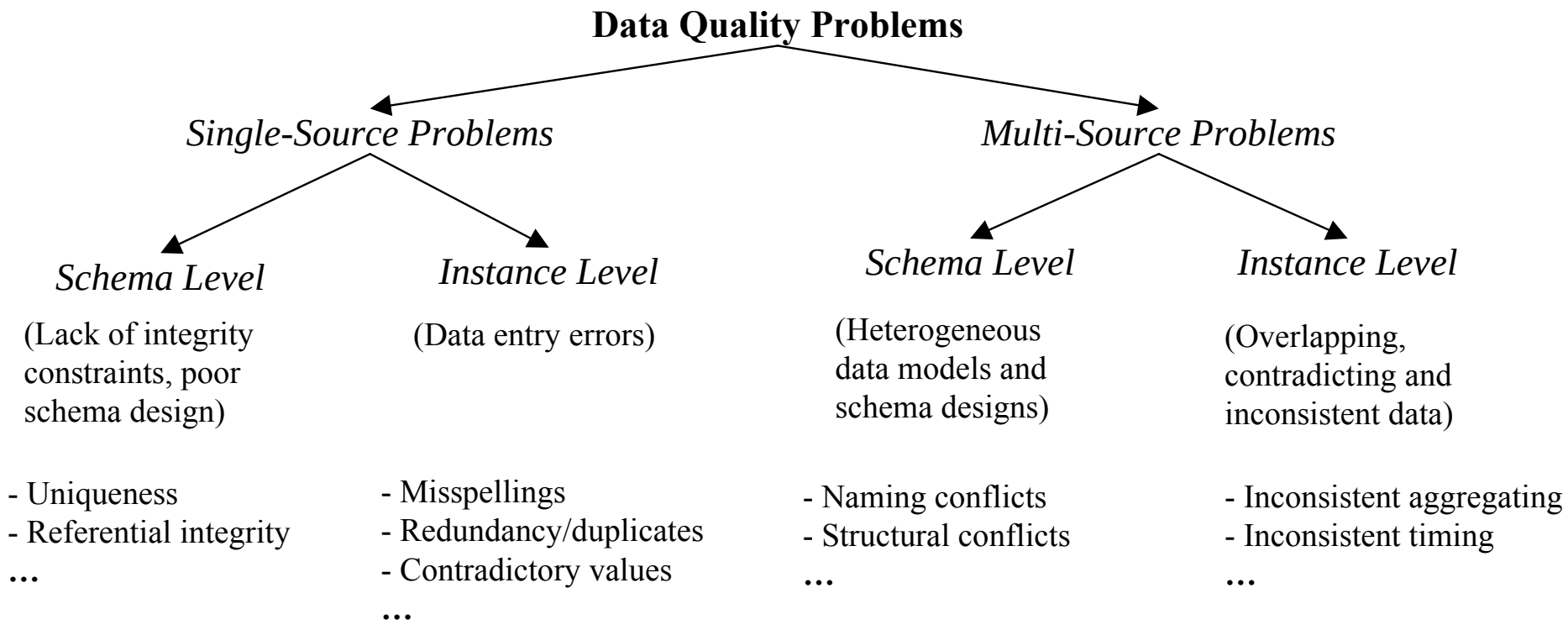
E. Rahm, H. H. Do: *Data Cleaning: Problems and Current Approaches.* IEEE Techn. Bulletin on Data Engineering, 2000

ScaDS DRESDEN LEIPZIG

2 LEVELS OF DATA INTEGRATION

- Metadata (schema) level
  - *Schema Matching:* find correspondences between source schemas and data warehouse schema
  - *Schema Merge:* integrate new source schemas into data warehouse schemas
- Instance (entity) level
  - transform heterogeneous source data into uniform representation prescribed by data warehouse schema
  - identify and resolve data quality problems
  - identify and resolve equivalent instance records): *object matching / deduplication / entity resolution*

CLASSIFICATION OF DATA QUALITY PROBLEMS



E. Rahm, H. H. Do: *Data Cleaning: Problems and Current Approaches.* IEEE Techn. Bulletin on Data Engineering, 2000

ScaDS DRESDEN LEIPZIG

AGENDA PART I (BIG DATA INTEGRATION)

- Introduction
  - Big Data
  - Data Quality
- Scalable entity resolution / link discovery
  - Introduction
  - Comparison of ER frameworks
  - Comparison of Frameworks for Link Discovery
  - Use case: Matching of product offers
  - Hadoop-based entity resolution (Dedoop)
  - Load balancing to deal with data skew
- Large-scale schema/ontology matching
- Holistic data integration
- Summary

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 1141 - Related articles - All 37 versions

Hai Do H.: Data Cleaning: Problems and Current approches *
E Rahm - Bulletin of the Technical Committee on Data ..., 2000
Cited by 5 - Related articles

Data Cleaning: Problems & Current Approaches *
D Hang-Hai, R Erhard - IEEE bulletin of the technical committee on Data ..., 2000
Cited by 5 - Related articles

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

ti Hal Do. Data Cleaning: Problem and Current Approaches *
R Erhard - IEEE Techn Bulletin Data Engineering, 2000
Cited by 3 - Related articles

Do Hong Hai," Data Cleaning: Problems and Current Approaches *
R Erhard - IEEE Bulletin of the Technical Committee on Data ..., 2000
Cited by 2 - Related articles

ScaDS DRESDEN LEIPZIG

DUPLICATE 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

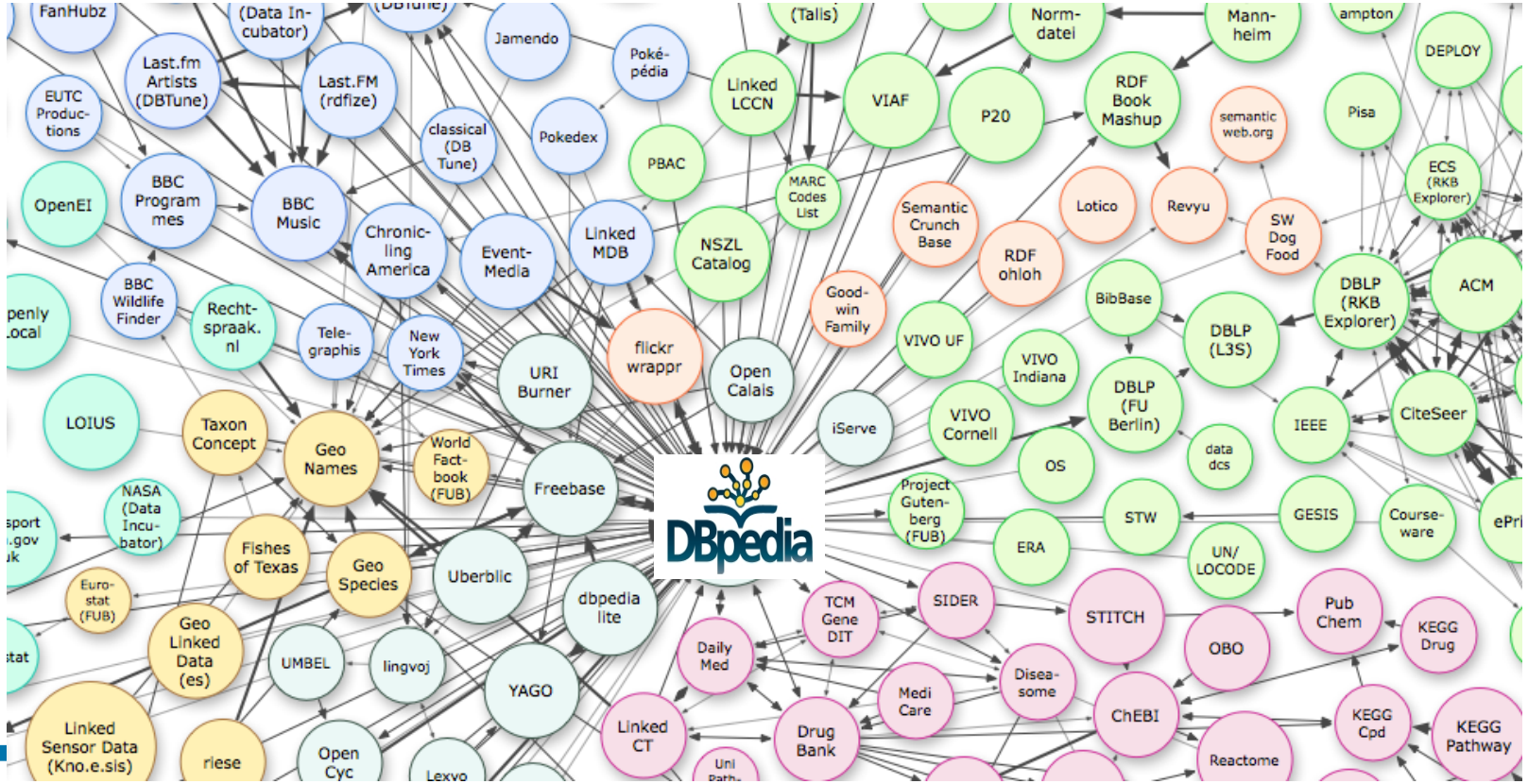
Canon Vixia Hf S10 Care & Cleaning
Care & Cleaning Digital Camera/Camcorder Deluxe Cleaning Kit with LCD Screen Guard Canon VIXIA HF S10 Camcorders Care & Cleaning.
Add to Shopping List

$2.99 new
shop.com
★★½☆☆ 38 seller ratings

ScaDS DRESDEN LEIPZIG

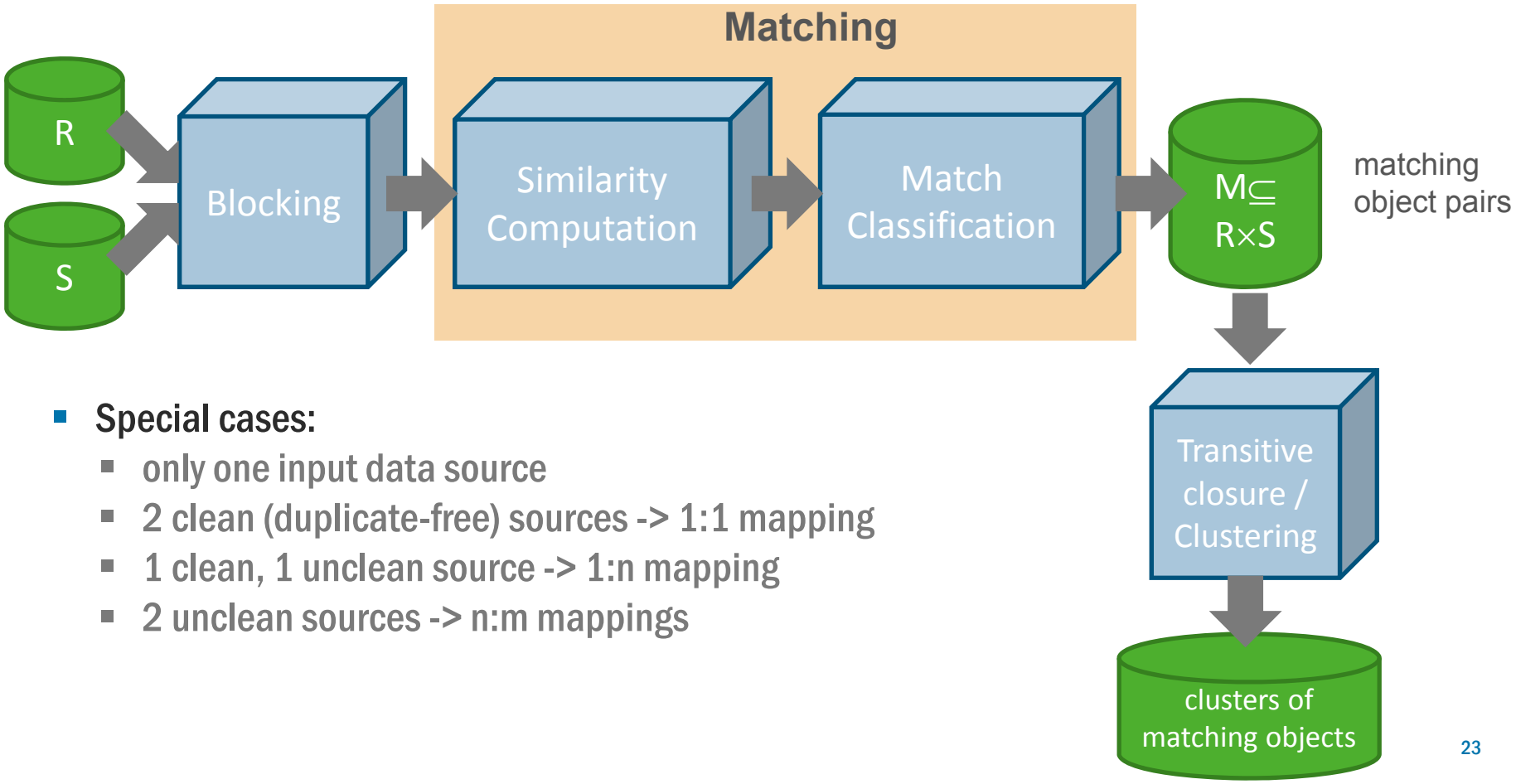
LINKED OPEN DATA

- Many linked data sources containing ontologies and associated instances
  - 31 billion triples (2011), 90 billion triples (2015, stats.lod2.eu)
- sameAs Links (mappings) between ontology concepts and instances for data integration



ScaDS DRESDEN LEIPZIG

GENERAL OBJECT MATCHING WORKFLOW

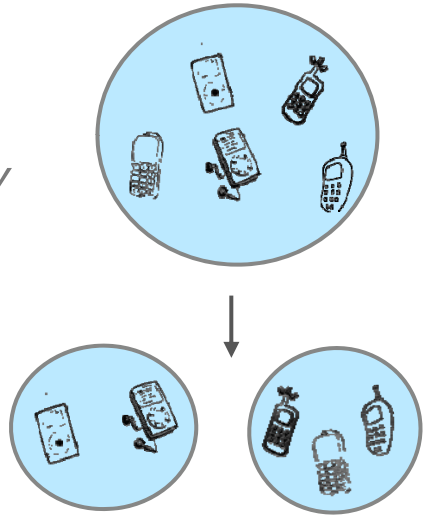


- Special cases:
  - only one input data source
  - 2 clean (duplicate-free) sources -> 1:1 mapping
  - 1 clean, 1 unclean source -> 1:n mapping
  - 2 unclean sources -> n:m mappings

ScaDS DRESDEN LEIPZIG

EXISTING OBJECT MATCHING APPROACHES

- Many tools and research prototypes
- **Blocking** to reduce search space
  - Group similar objects within blocks based on *blocking key*
  - Restrict object matching to objects from the same block
  - Alternative approach: Sorted Neighborhood
- Combined use of **several matchers**
  - Attribute-level matching
    based on generic or domain-specific similarity functions,
    e.g., string similarity (edit distance, n-gram, TF/IDF, etc.)
  - Context-based matchers
  - Learning-based or manual specification of matcher combination
- Optional: **transitive closure** of matches to identify indirect matches



ScaDS DRESDEN LEIPZIG

ER FRAMEWORKS 1 (NON-LEARNING)*

	BN	MOMA	SERF	DuDe	FRIL
Entity type	XML	relational	relational	relational	relational
Blocking key definition	-	-	-	manual	manual
partitioning disjoint overlapping	-	-	-	Sorted Neighborhood	Sorted Neighborhood
Matchers	attribute, context	attribute, context	attribute	attribute	attribute
Matcher combination	numerical	workflow	rules	workflow	workflow

* Koepcke, H.; Rahm, E.: *Frameworks for entity matching: A comparison*. Data & Knowledge Engineering, 2010

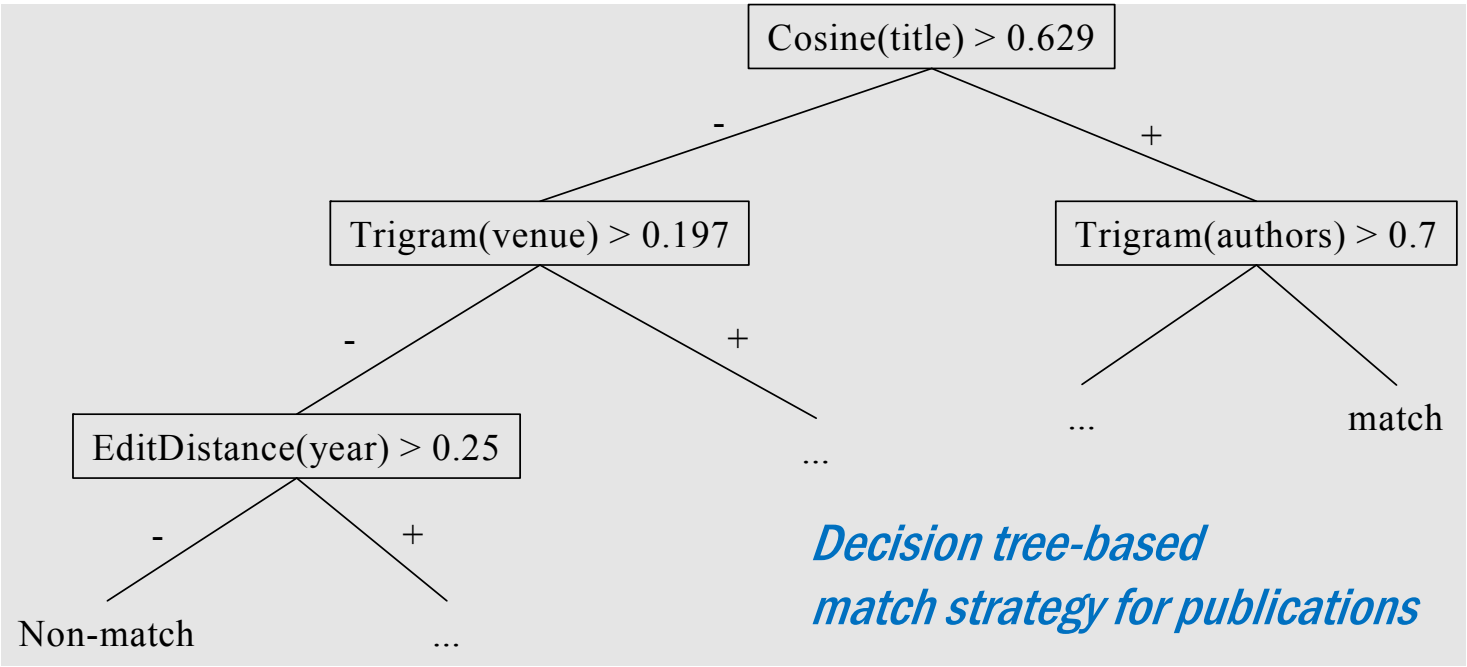
ScaDS DRESDEN LEIPZIG

ER FRAMEWORKS 2 (LEARNING-BASED)

	Active Atlas	MARLIN	Op. Trees	TAILOR	FEBRL	Context-b. F.work	FEVER
Entity type	relational	rel.	rel.	rel.	XML, rel.	rel.	rel.
Blocking key definition	manual	manual	manual	manual	manual	manual	manual
partitioning *disjoint* overlapping	hashing	canopy clustering	canopy cl.	*threshold* Sorted Neighb.	SN	canopy-like	*several*, SN, canopy
Matchers	attribute	attr.	attr.	attr.	attr.	attr., context	attr.
Matcher combination	rules	numerical, rules	rules	numerical, rules	numerical	numerical, rules	workflow
Learners	decision tree	SVM, dec. tree	SVM-like	probab. dec. tree	SVM	diverse	multiple,SVM, dec. tree, ..
Training selection	manual, semi-autom.	manual, semi-autom.	manual	manual	manual, automatic	manual	manual, semi-autom.

LEARNING-BASED MATCHING

- Supervised learning
  - use of training data (matching / non-matching pairs of entities) to find effective matcher combination and configuration
  - FEVER uses Decision Tree, Logistic Regression, SVM and multiple learner approach



Decision tree-based match strategy for publications

ScaDS DRESDEN LEIPZIG

OBSERVATIONS FROM [DKE'10]

- Numerous frameworks with similar functionality regarding blocking and matchers
  - Primarily attribute-level matching for relational sources
  - Manual selection of matchers / attributes
  - Manual specification of blocking keys
- Frequent use of training-based match strategies
  - Mostly manual training
  - Most popular learners: SVM, decision tree
- Heterogeneous, non-conclusive evaluations
  - Different datasets and methodologies
  - Missing specification details, e.g. on training
  - Unclear scalability to larger datasets

ScaDS DRESDEN LEIPZIG

COMPARATIVE EVALUATION: MATCH TASKS

Match task		Source size (#entities)		Mapping size (#correspondences)		
Domain	Sources	Source 1	Source 2	Full input mapping (cross product)	Reduced input mapping (blocking)	perfect match result
Bibliographic	DBLP-ACM	2,616	2,294	6 million	494,000	2224
	DBLP-Scholar	2,616	64,263	168.1 million	607,000	5343
E-commerce	Amazon-GoogleProducts	1,363	3,226	4.4 million	342,761	1300
	Abt-Buy	1,081	1,092	1.2 million	164,072	1097

Koepcke, Thor, Rahm: *Evaluation of entity resolution approaches on real-world match problems*. PVLDB 2010

Koepcke, Thor, Rahm: Comparative evaluation of entity resolution approaches with FEVER. PVLDB 2009

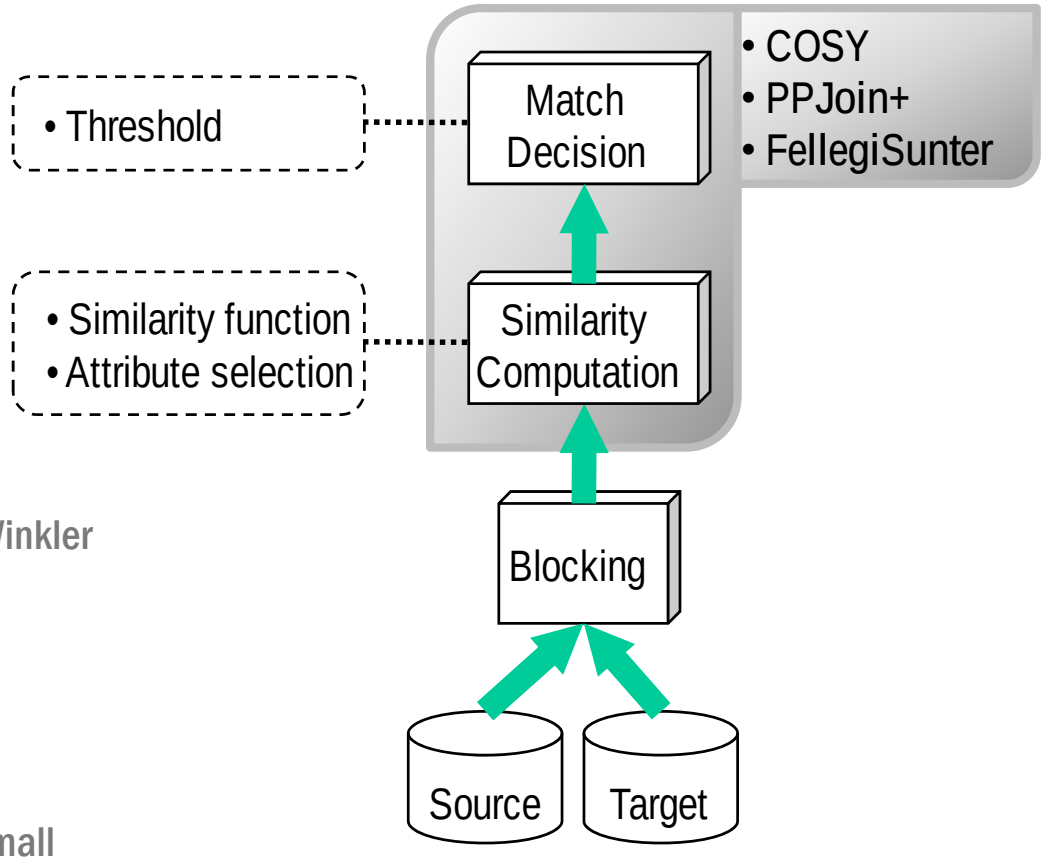
FEVER

Framework for EValuating Entity Resolution

ScaDS DRESDEN LEIPZIG

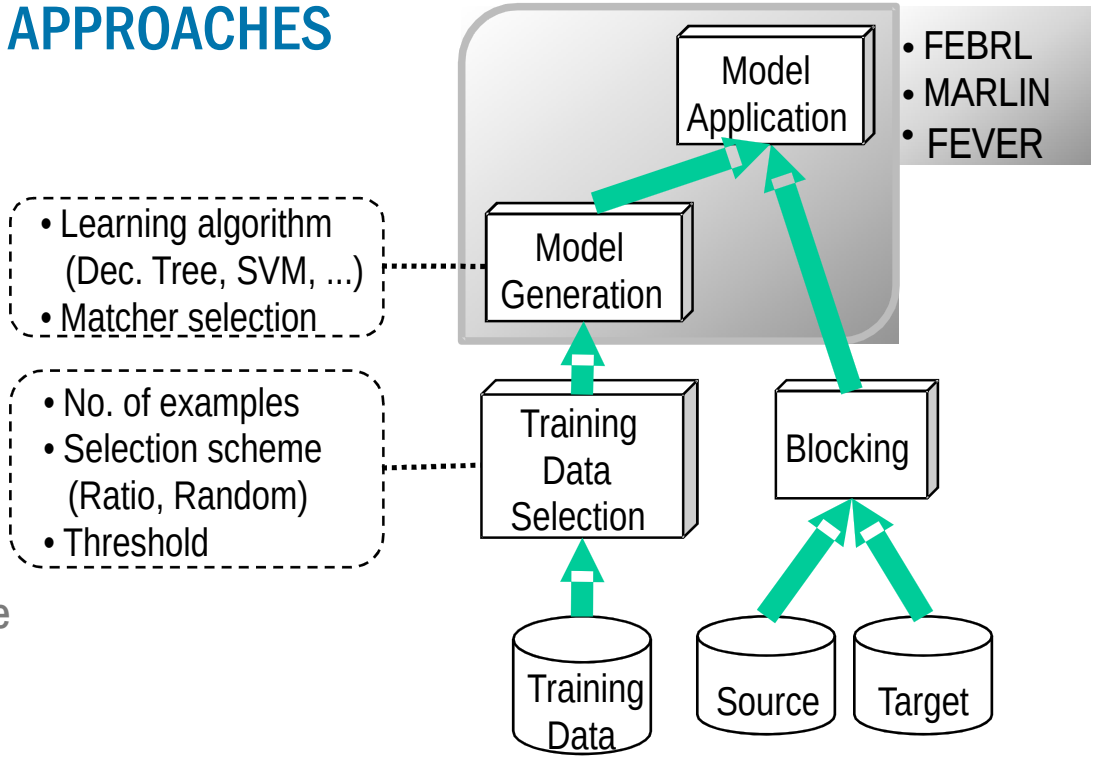
NON-LEARNING APPROACHES

- **COSY** (commercial system)
  - Black box similarity function
  - Overall and attribute level thresholds
- **PPJoin+**
  - Similarity functions: Cosine, Jaccard
  - Threshold
- **FellegiSunter** (FEBRL)
  - Similarity functions: TokenSet, Trigram, Winkler
  - Similarity threshold
- **Match configurations**
  - Use of 1 or 2 attributes
  - Use of FEVER to optimize thresholds for small subset of input data (500 object pairs)



ScaDS DRESDEN LEIPZIG

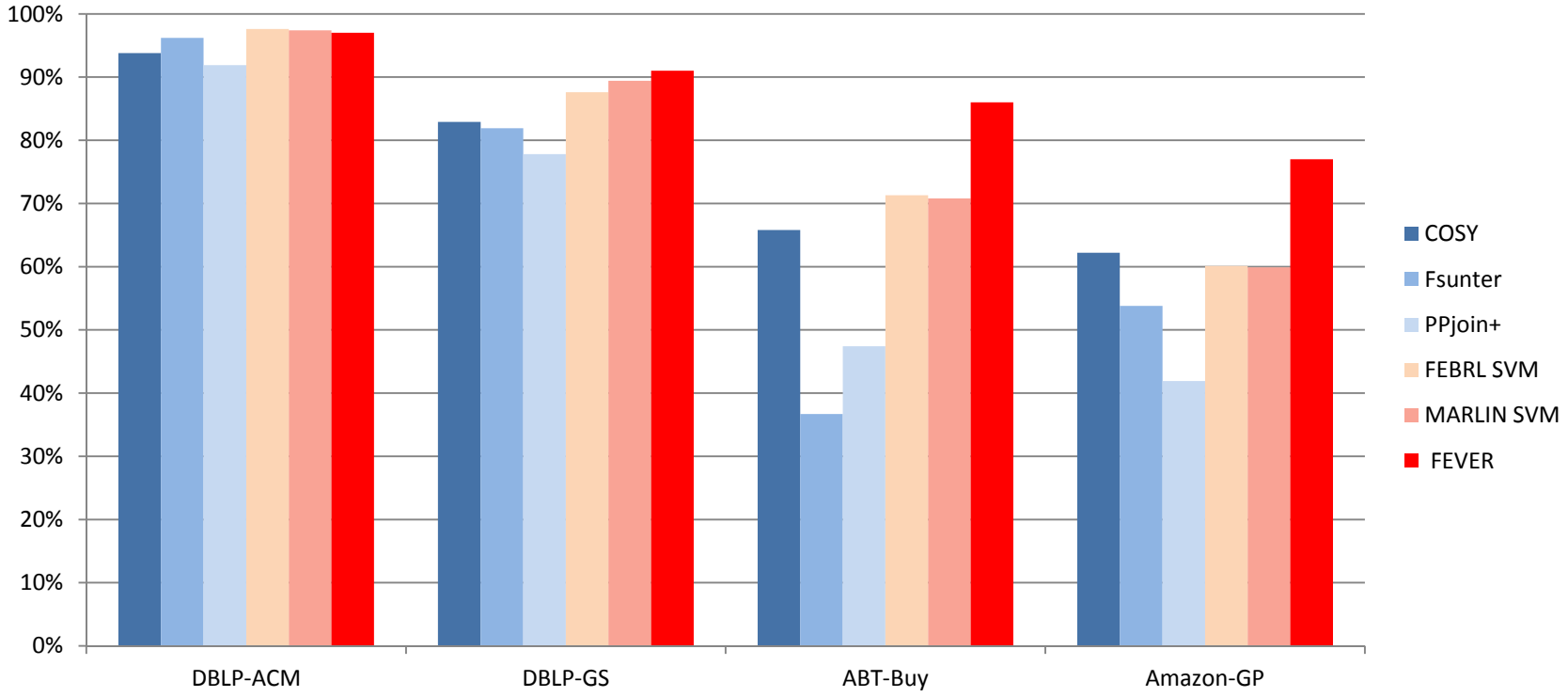
LEARNING-BASED APPROACHES



- **FEBRL**
  - 3 matchers: Winkler, Tokenset, Trigram
  - learning algorithm: SVM
- **MARLIN**
  - 2 matchers: Edit Distance, Cosine
  - learning algorithms: SVM , decision tree
  - single step vs. two level learning
- **FEVER**
  - Trigram and TF/IDF matchers
  - Majority consensus from 3 learners (SVM , decision tree, logistic regression)
- **Match configurations**
  - use of 1 or 2 attributes
  - small training size (max. 500 object pairs with balanced matches/non-matches)

ScaDS DRESDEN LEIPZIG

QUALITY (F-MEASURE) COMPARISON



- Bibliographic tasks are simpler than E-commerce tasks
- Learning-based approaches perform best, especially for difficult match problems
  - SVM most promising learner
  - FEVER benefits from majority consensus of 3 learners
- COSY relatively good / PPJoin+ limited to 1 attribute

ScaDS DRESDEN LEIPZIG

EFFICIENCY RESULTS

	Blocked (s)	Cartesian (s)
COSY	1 – 44	2– 434
FellegiSunter	2 – 2,800	17 – >500,000
PPJoin+	<1 – 3	<1 – 7
FEBRL SVM	99-480	1,400 – >500,000
MARLIN SVM	20-380	2,200 – >500,000

- PPJoin+ and COSY very fast, even for Cartesian product
- FellegiSunter slowest non-learning approach
- Learning-based approaches very slow
  - require blocking

ScaDS DRESDEN LEIPZIG

OBSERVATIONS

- Evaluations reveal big differences regarding match quality and execution times
- Effective approaches: Learning-based approaches,
  COSY (partly)
- Fast approaches: COSY, PPJoin+
- Weak points:
  - Combination of several attributes requires higher tuning/training effort
  - E-commerce tasks could not be effectively solved. More sophisticated methods are needed there
  - Scalability to large test cases needs to be better addressed

ScaDS DRESDEN LEIPZIG

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)

ScaDS DRESDEN LEIPZIG

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 - matcher combination	adaptive weighted average	manual weighted combination	manual weighted average	manual weighted average	adaptive -	manual weighted combination	adaptive weighted average
Runtime optimization - Blocking - Filtering	 - indexing	 - indexing	 - -	 - indexing	 - -	 - indexing	 - 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 - external dictionaries - existing mappings	 ?* -	 ?* -	 - -	 ?* -	 - -	 - -	 - -
Post-processing	-	-	Coherence checks	Inconsistency repair	-	-	-
Parallel processing	-	-	-	-	-	MapReduce	-
GUI/web interface/API	- / - / -	✓ / ? / -	- / - / -	✓ / ✓ / -	- / - / -	- / - / -	- / - / -
Download Tool/Source Open Source project	✓ / - -	-[1] / - -	✓ / ✓ ✓	✓ / ✓ ✓	✓ / ✓ ✓	✓ / - -	✓ / - -

ScaDS DRESDEN LEIPZIG

LEARNING-BASED LINK DISCOVERY TOOLS

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

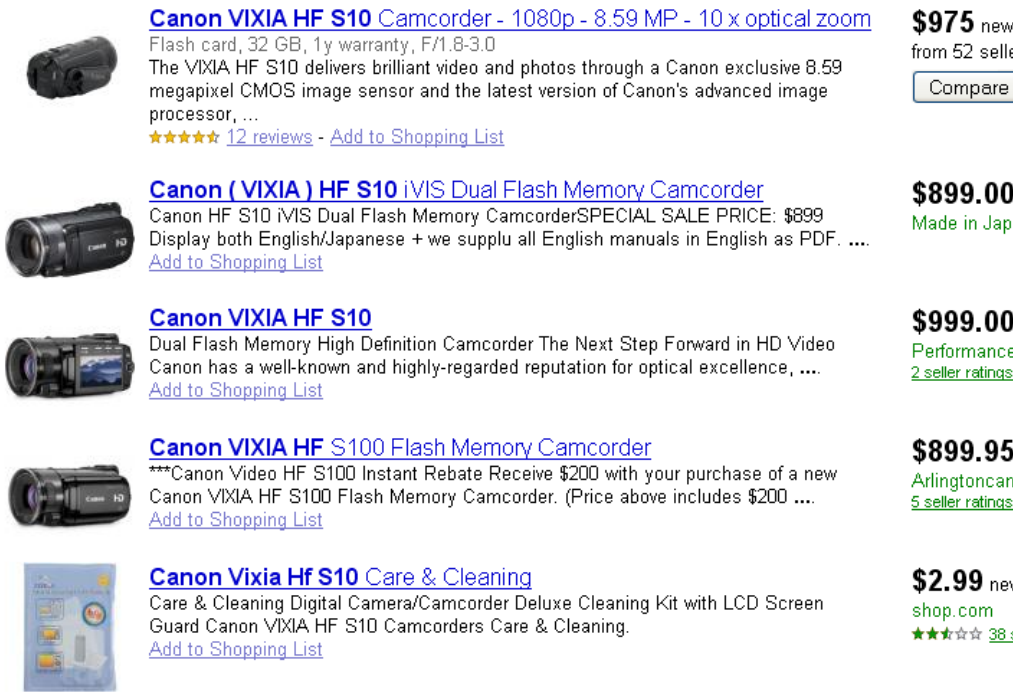
	KnoFuss	Silk	LIMES	RuleMiner
Data Input	RDF, SPARQL	RDF, SPARQL, CSV	RDF, SPARQL, CSV	RDF
Supported linktypes	`owl:sameAs`	`owl:sameAs`, user-specified others	`owl:sameAs`, user-specified others	`owl:sameAs`
Configuration	manual (match rules), unsupervised learning (genetic programming)	manual (match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

ScaDS DRESDEN LEIPZIG

BIG DATA INTEGRATION USE CASE

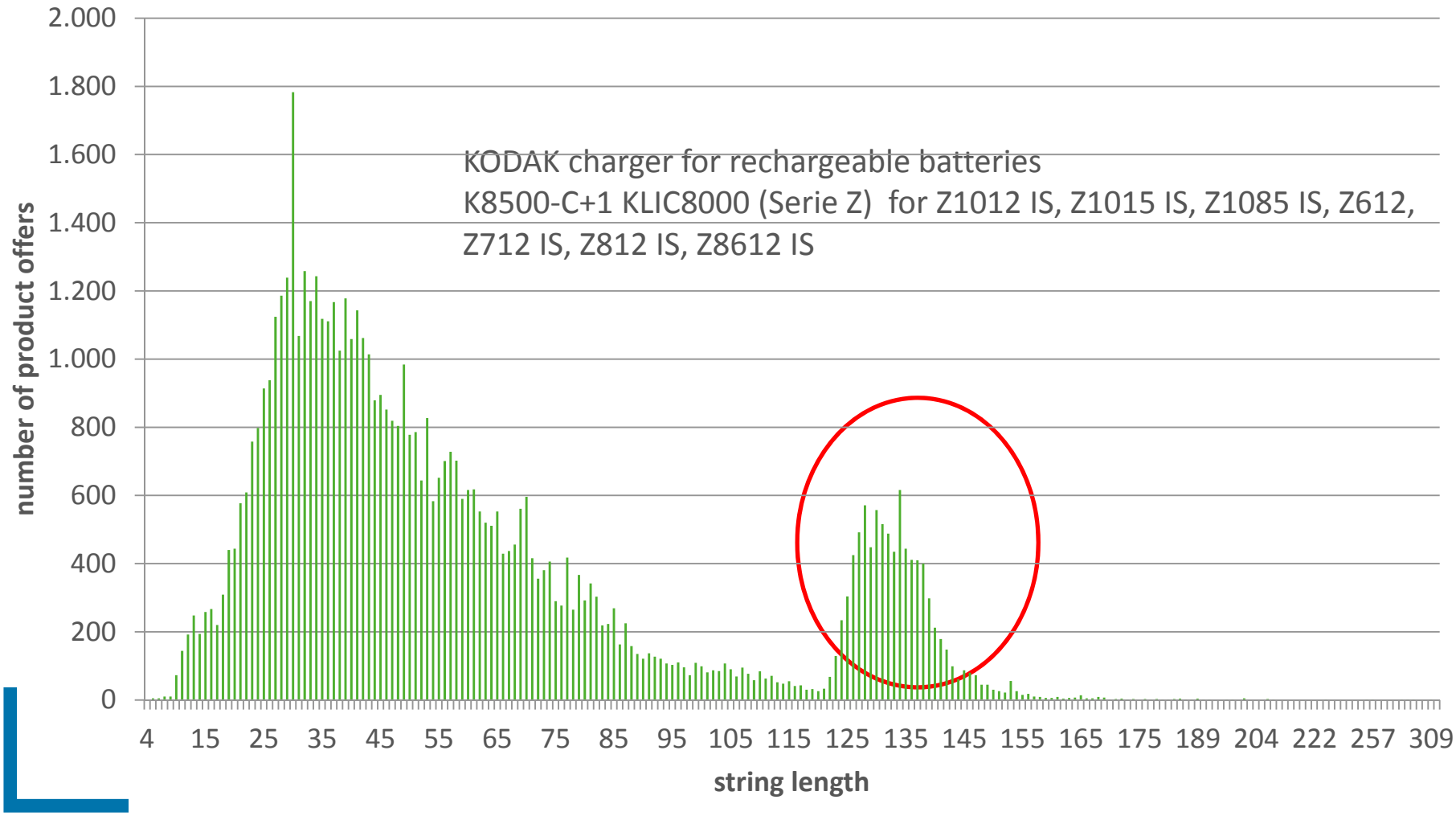
INTEGRATION OF PRODUCT OFFERS IN COMPARISON PORTAL

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



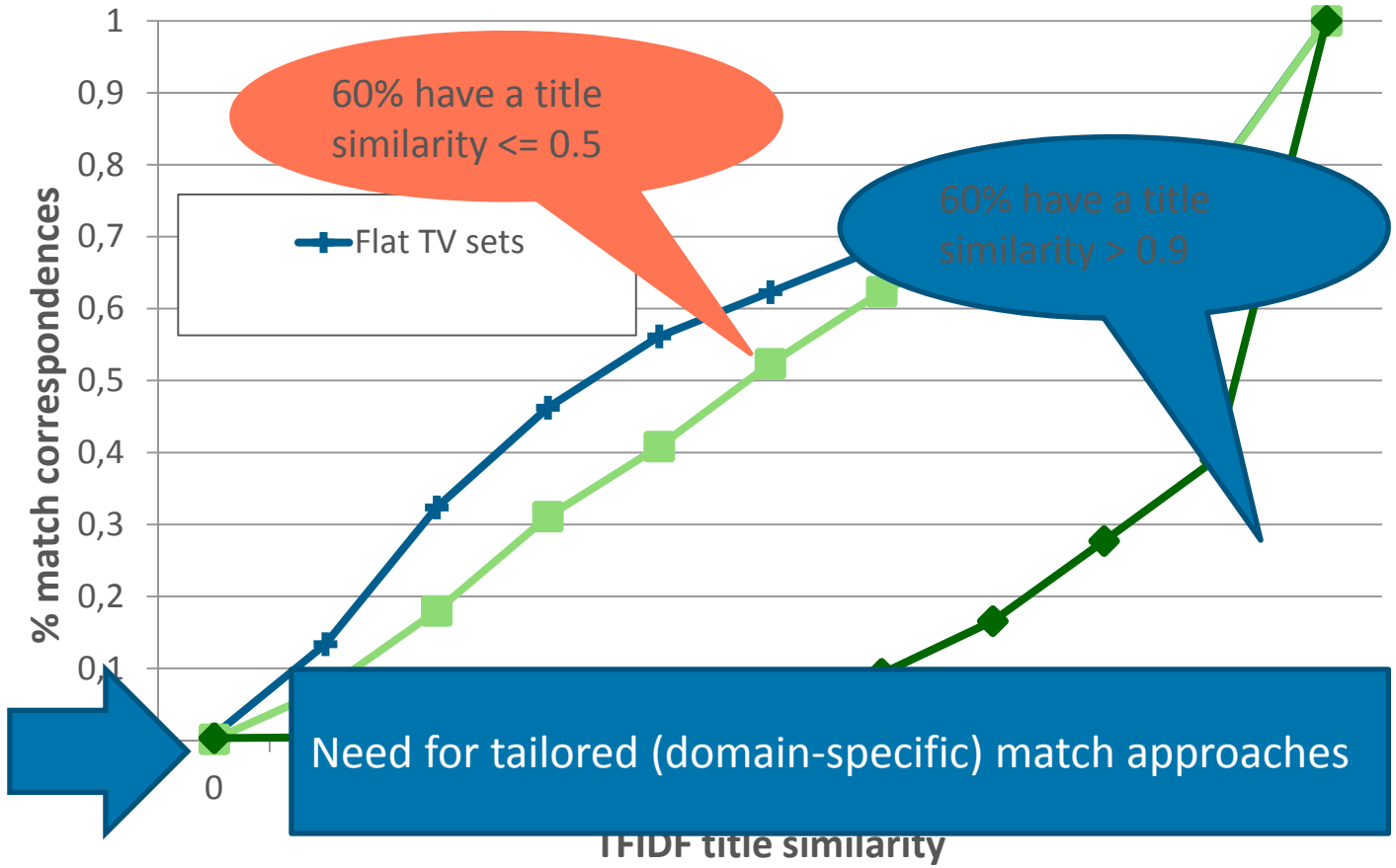
ScaDS DRESDEN LEIPZIG

HETEROGENEOUS AND VERBOSE STRINGS



ScaDS DRESDEN LEIPZIG

STANDARD STRING MATCHERS FAIL



ScaDS DRESDEN LEIPZIG

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

ScaDS DRESDEN LEIPZIG

USE OF PRODUCT CODES

- Frequent existence of specific product codes for certain products
- Product code = manufacturer-specific identifier
  - any sequence consisting of alphabetic, special, and numeric characters split by an arbitrary number of white spaces.

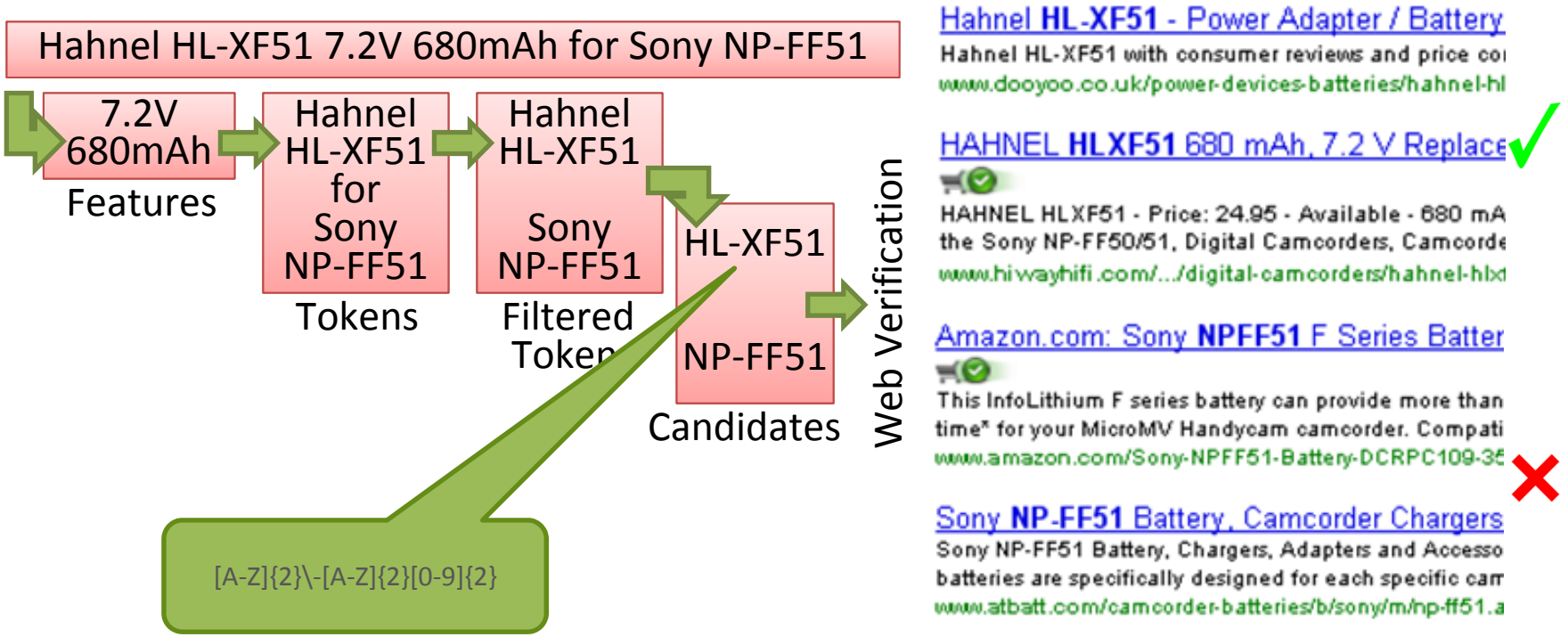
➢ Utilize to differentiate similar but different products.

Canon VIXIA HF S100 Camcorder - 1080p - 8.59 MP

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

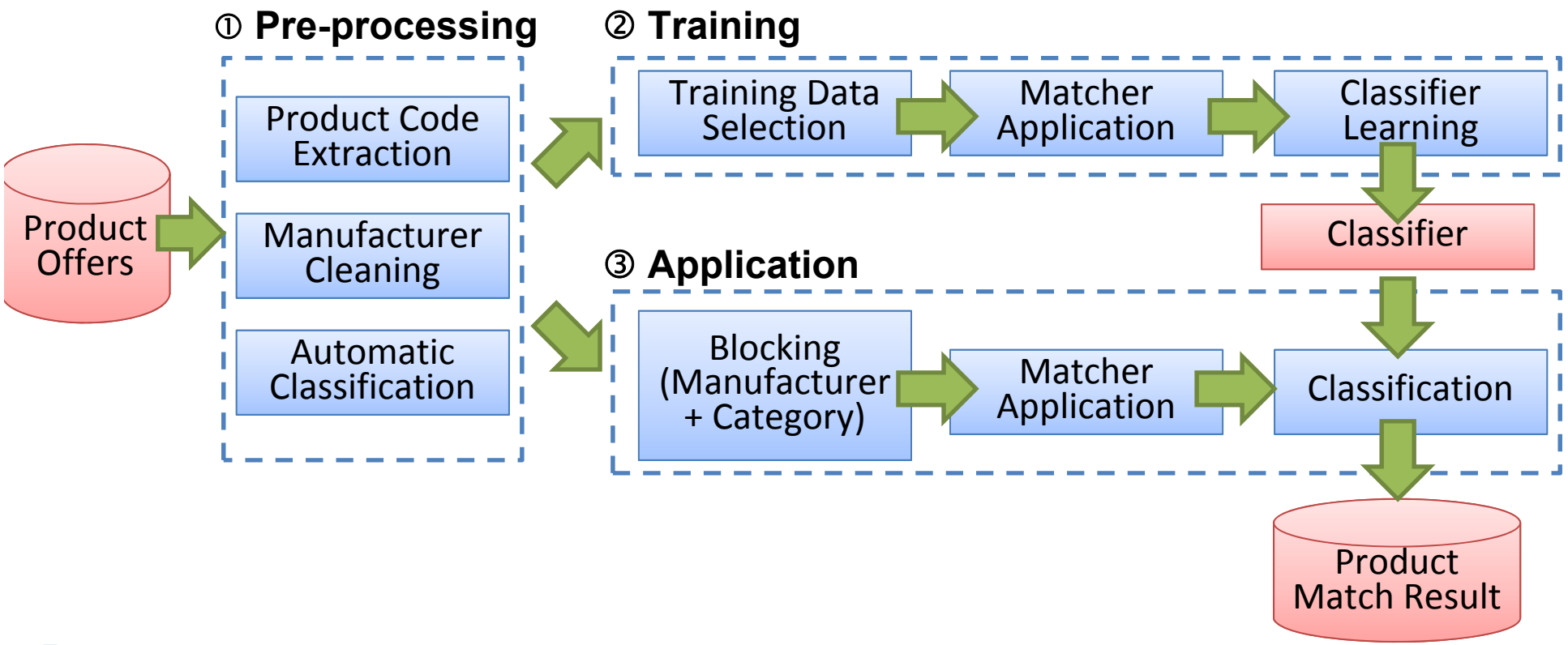
ScaDS DRESDEN LEIPZIG

PRODUCT CODE EXTRACTION



ScaDS DRESDEN LEIPZIG

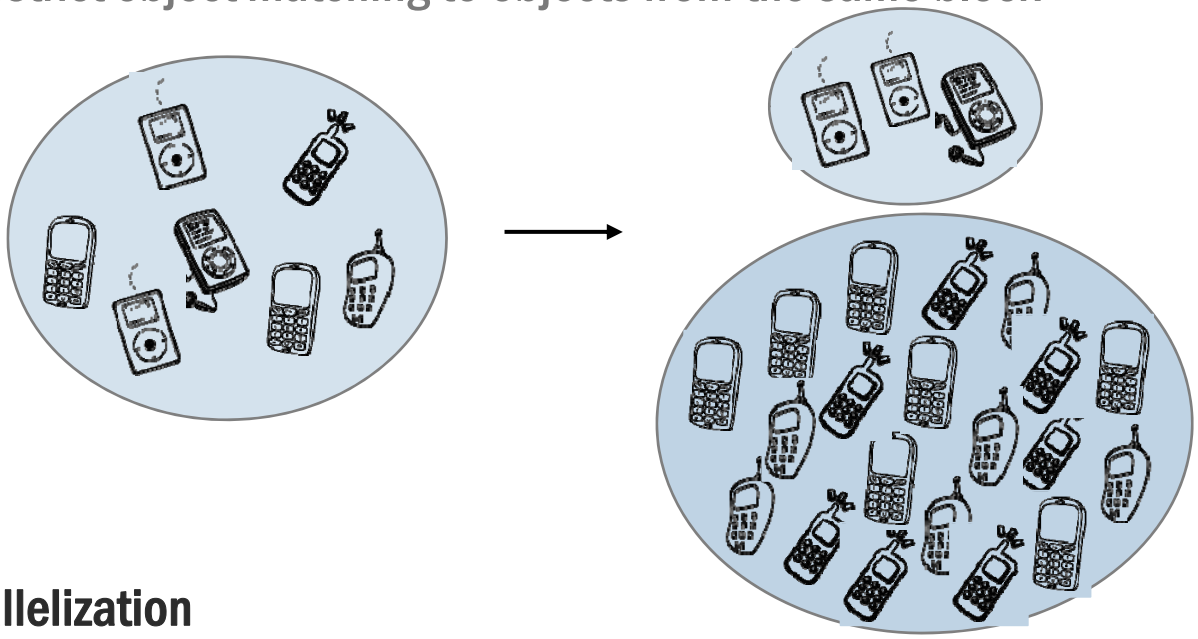
LEARNING-BASED MATCH APPROACH



ScaDS DRESDEN LEIPZIG

HOW TO SPEED UP OBJECT MATCHING?

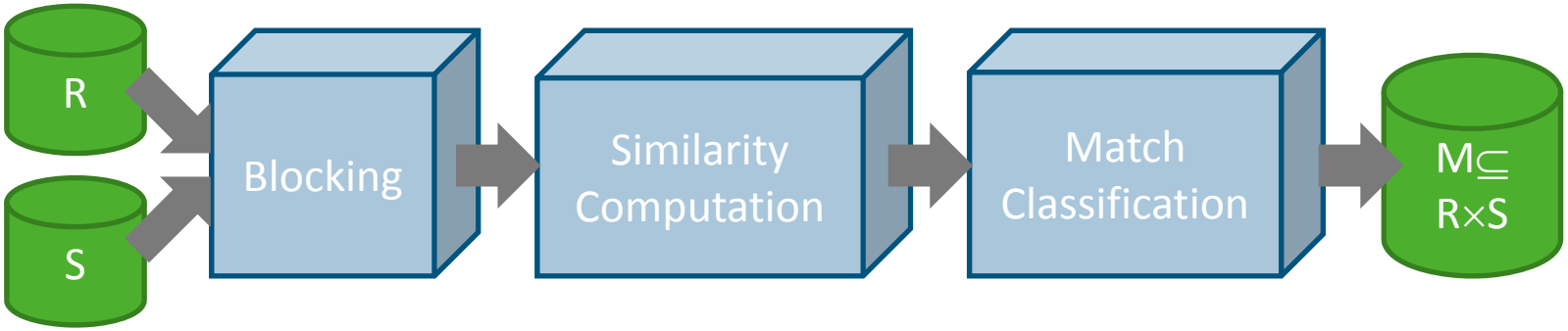
- **Blocking** to reduce search space
  - group similar objects within blocks based on *blocking key*
  - restrict object matching to objects from the same block

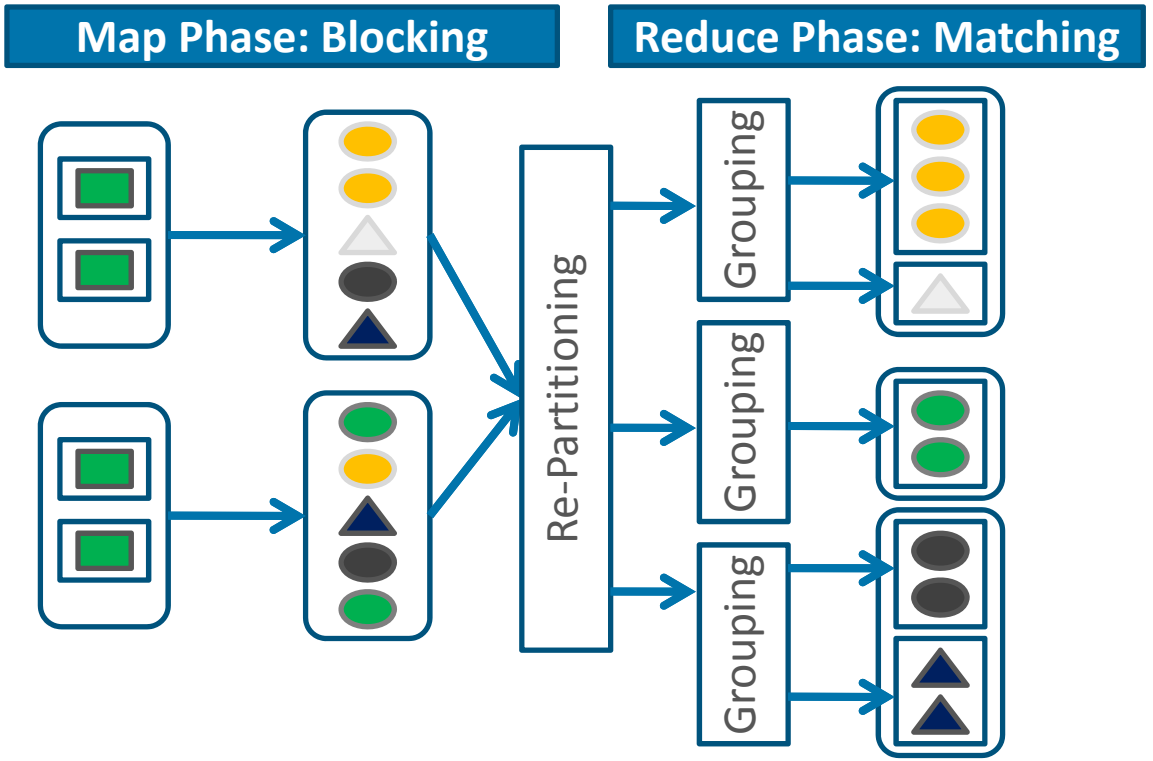


- **Parallelization**
  - split match computation in sub-tasks to be executed in parallel
  - exploitation of Big Data infrastructures such as Hadoop (Map/Reduce or variations)

ScaDS DRESDEN LEIPZIG

GENERAL OBJECT MATCHING WORKFLOW





ScaDS DRESDEN LEIPZIG

LOAD BALANCING

- Data skew leads to unbalanced workload
  - Large blocks prevent utilization of more than a few nodes
  - Deteriorates scalability and efficiency
  - Unnecessary costs (you also pay for underutilized machines!)

- Key ideas for load balancing
  - Additional MR job to determine blocking key distribution, i.e., number and size of blocks (per input partition)
  - Global load balancing that assigns (nearly) the same number of pairs to reduce tasks

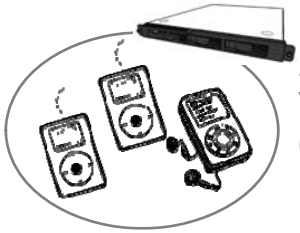
- Simplest approach : BlockSplit (ICDE2012)
  - split large blocks into sub-blocks with multiple match tasks
  - distribute the match tasks among multiple reduce tasks

ScaDS DRESDEN LEIPZIG

BLOCK SPLIT: 1 SLIDE ILLUSTRATION

- Example: 3 MP3 players + 6 cell phones → 18 pairs (1 time unit)
- Parallel matching on 2 (reduce) nodes

naiive approach

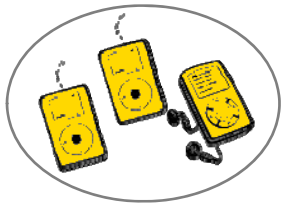


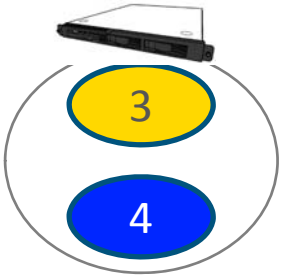
3 pairs (16%)

15 pairs (84%)

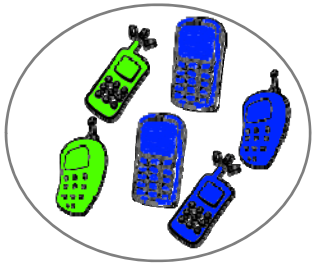
Speedup: 18/15=1.2

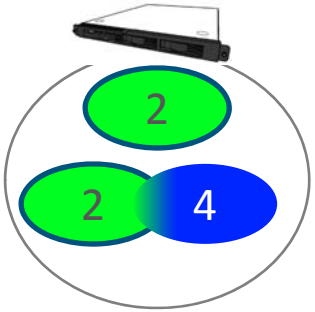
BlockSplit





3 pairs
6 pairs
9 pairs (50%)





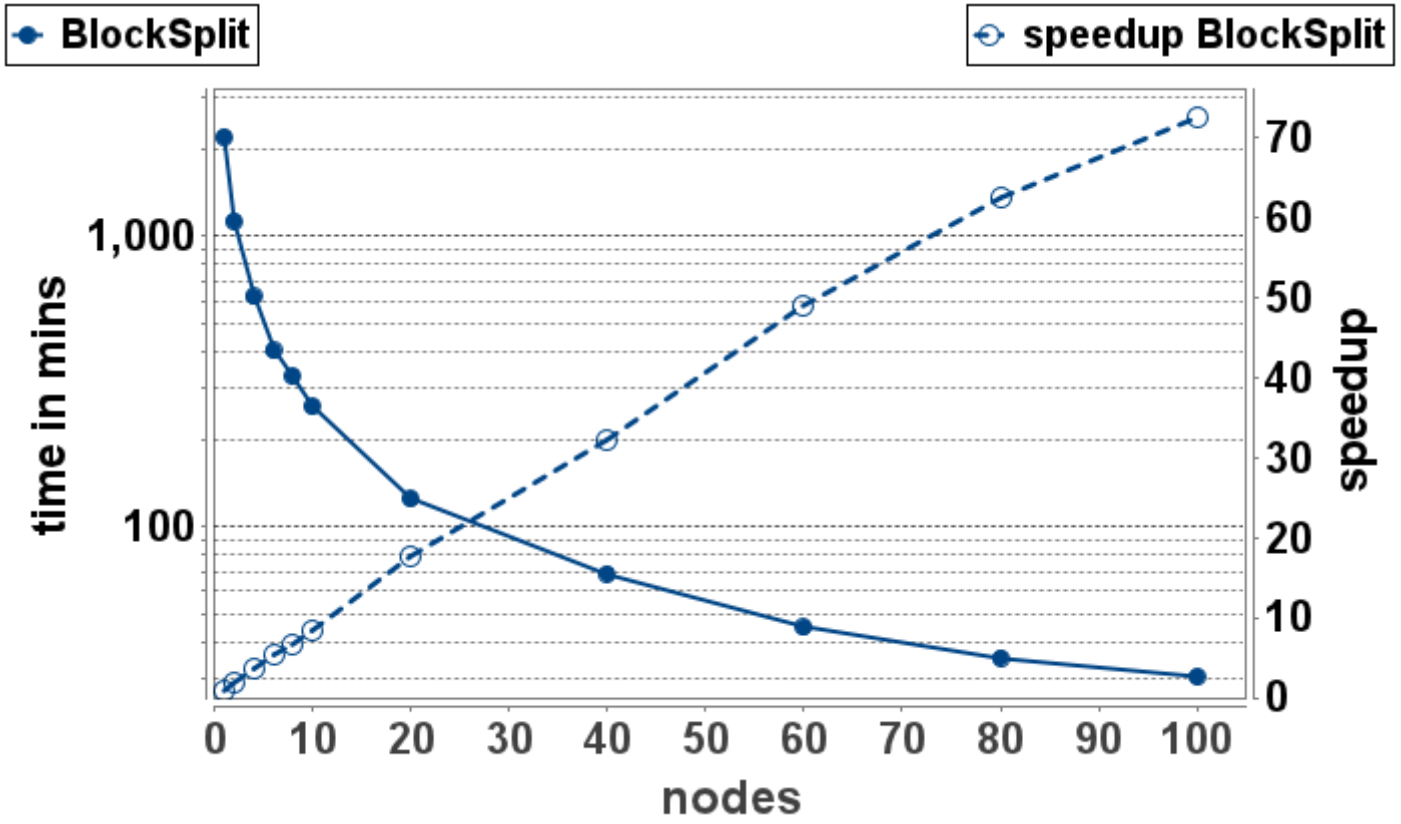
1 pair
8 pairs
9 pairs (50%)

Speedup: 2

ScaDS DRESDEN LEIPZIG

BLOCK SPLIT EVALUATION: SCALABILITY

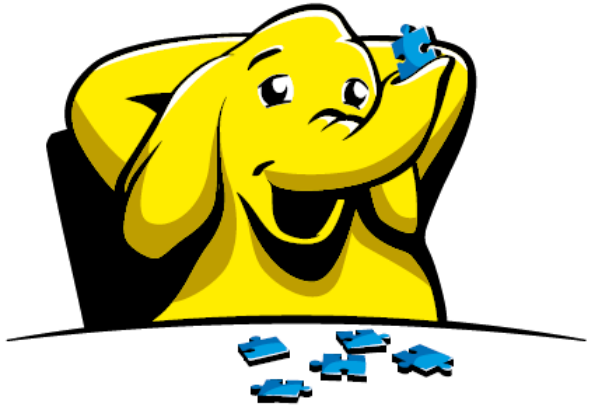
- Evaluation on Amazon EC infrastructure using Hadoop
- Matching of 114.000 product records



ScaDS DRESDEN LEIPZIG

DEDOOP: EFFICIENT DEDUPLICATION WITH HADOOP

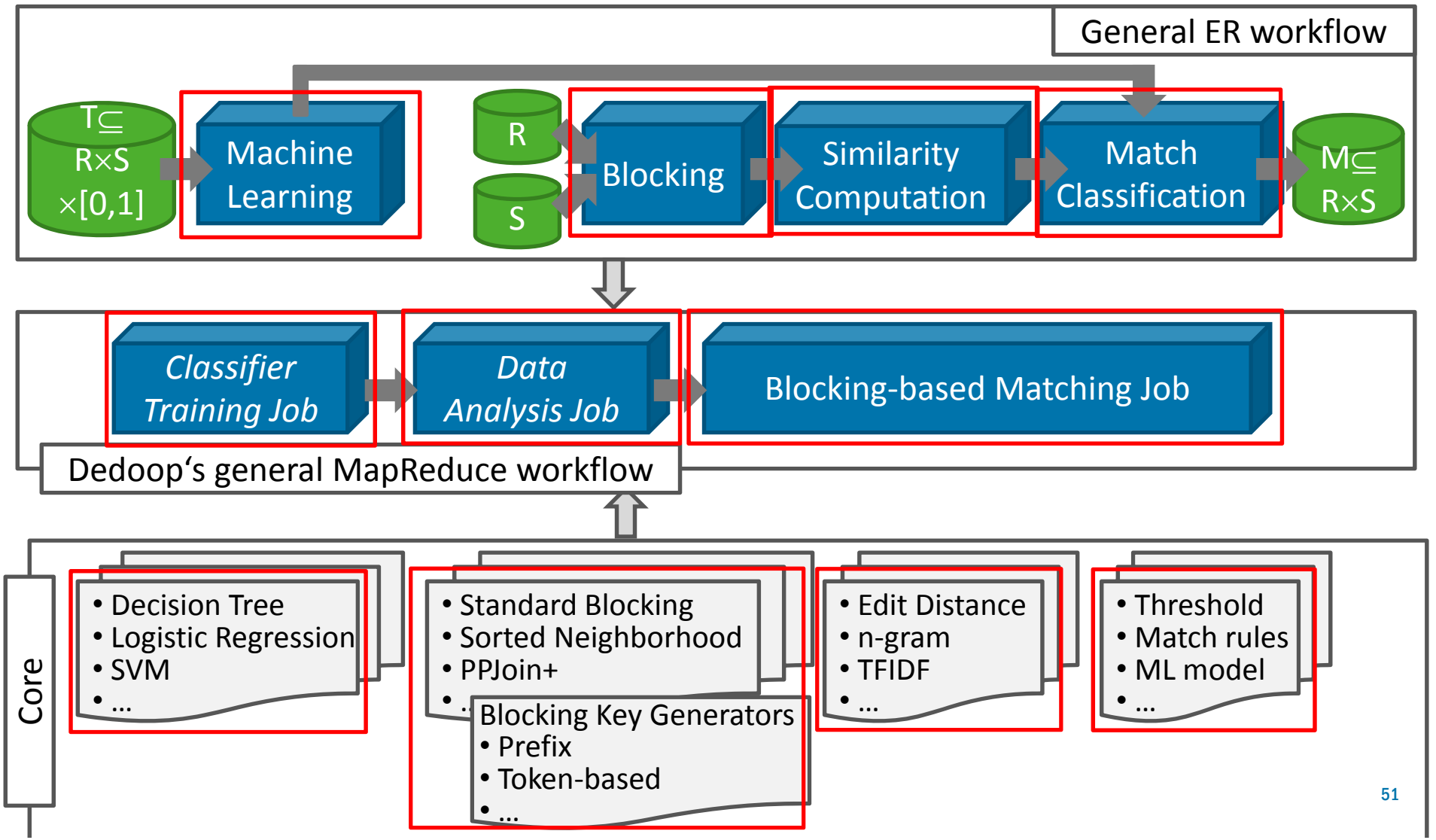
- Parallel execution of data integration/ match workflows with Hadoop
- Powerful library of match and blocking techniques
- Learning-based configuration
- GUI-based workflow specification
- Automatic generation and execution of Map/Reduce jobs on different clusters
- Automatic load balancing for optimal scalability
- Iterative computation of transitive closure



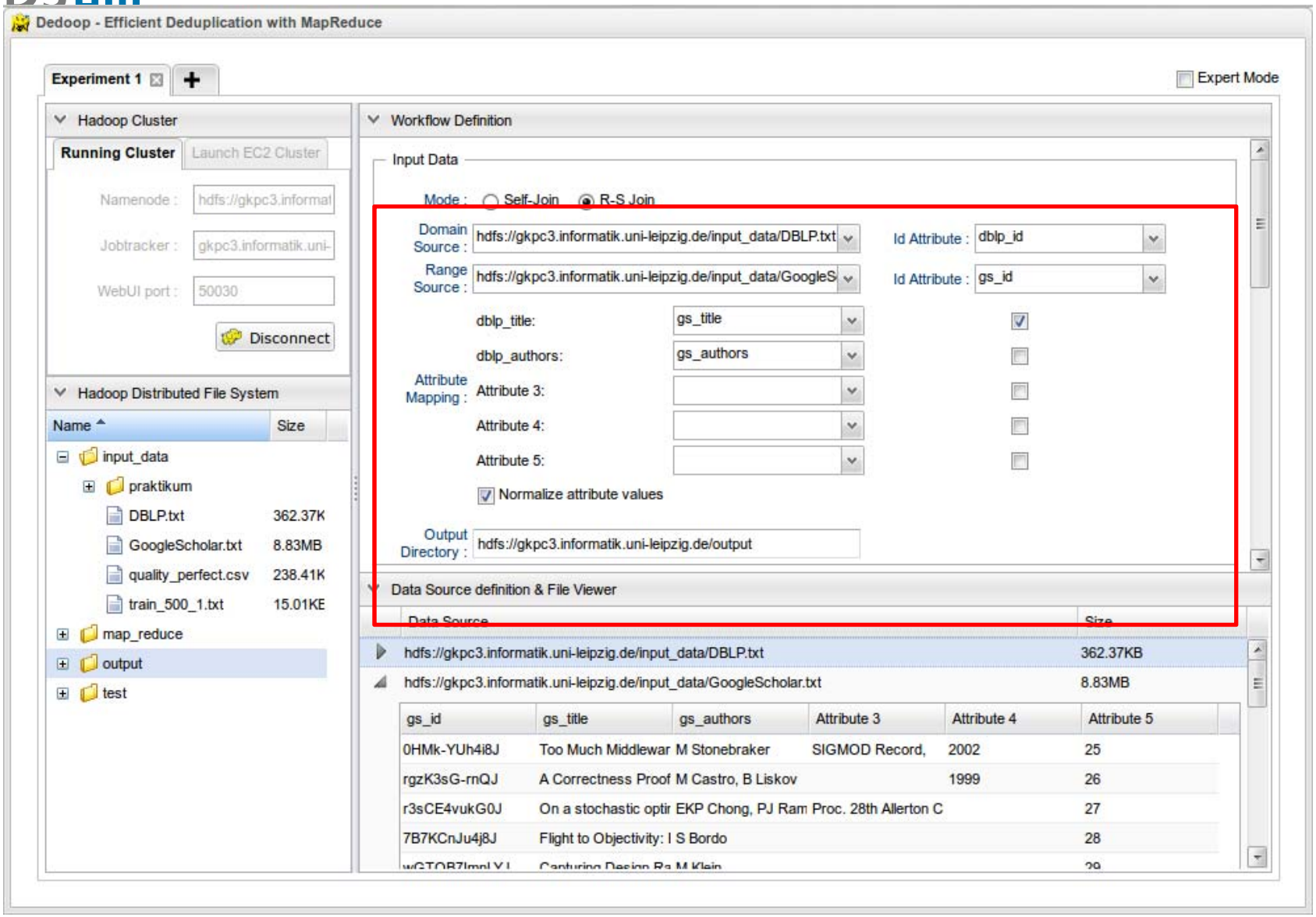
"This tool by far shows the most mature use of MapReduce for data deduplication"
www.hadoopsphere.com

ScaDS DRESDEN LEIPZIG

DEDOOP OVERVIEW



BROWSER-BASED CONFIGURATION



ScaDS DRESDEN LEIPZIG

COMPARATIVE EVALUATION OF ER METHODS

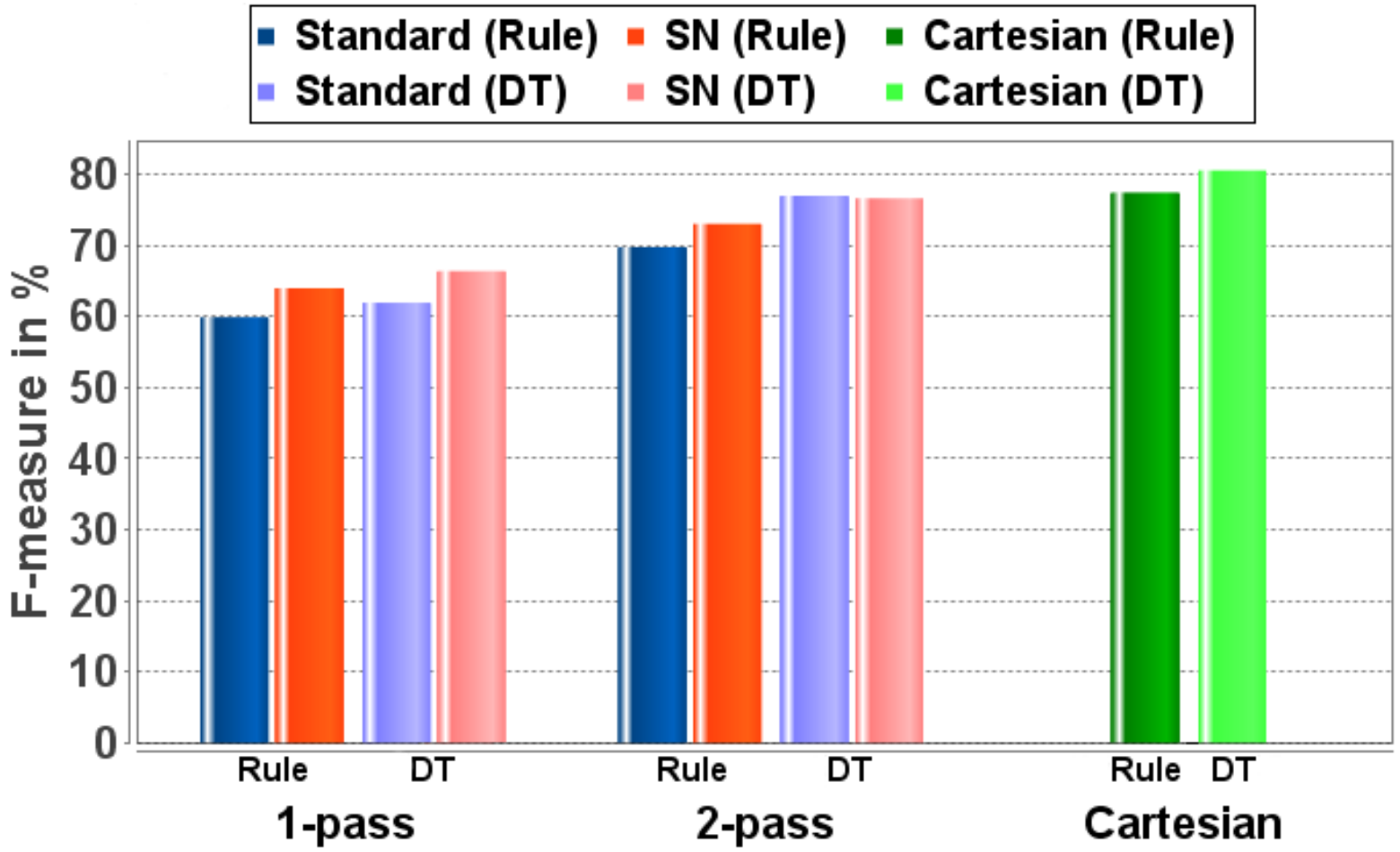
- using Dedoop to compare match quality and runtime for parallel blocking + matching*
  - Blocking: standard or sorted neighborhood (one or two passes)
  - Matching on 1 or 2 attributes (title, author)
  - manually specified, rule-based approaches
  - learning-based approaches (SVM, decision tree)
- bibliographic evaluation for relatively unclean Google Scholar dataset (65 K entities)
- training data: 500 labeled entity pairs
- 20 machines (Amazon EC)

Match Classification	Input Similarity Features	Match Criterium
$Rule_1$	3-gram(title)	sim ≥ 0.8
$Rule_2$	3-gram(title) 3-gram (authors)	sim (title)≥ 0.6 and sim (author)≥ 0.4
SVM	3-gram(title) 3-gram (authors)	SVM (WEKA LibSVM, -K 0 -C 10)
DT	3-gram(title) 3-gram (authors)	Decision Tree (WEKA LMT, default config)

*Kolb, Rahm: *Parallel Entity Resolution with Dedoop*. 2013

ScaDS DRESDEN LEIPZIG

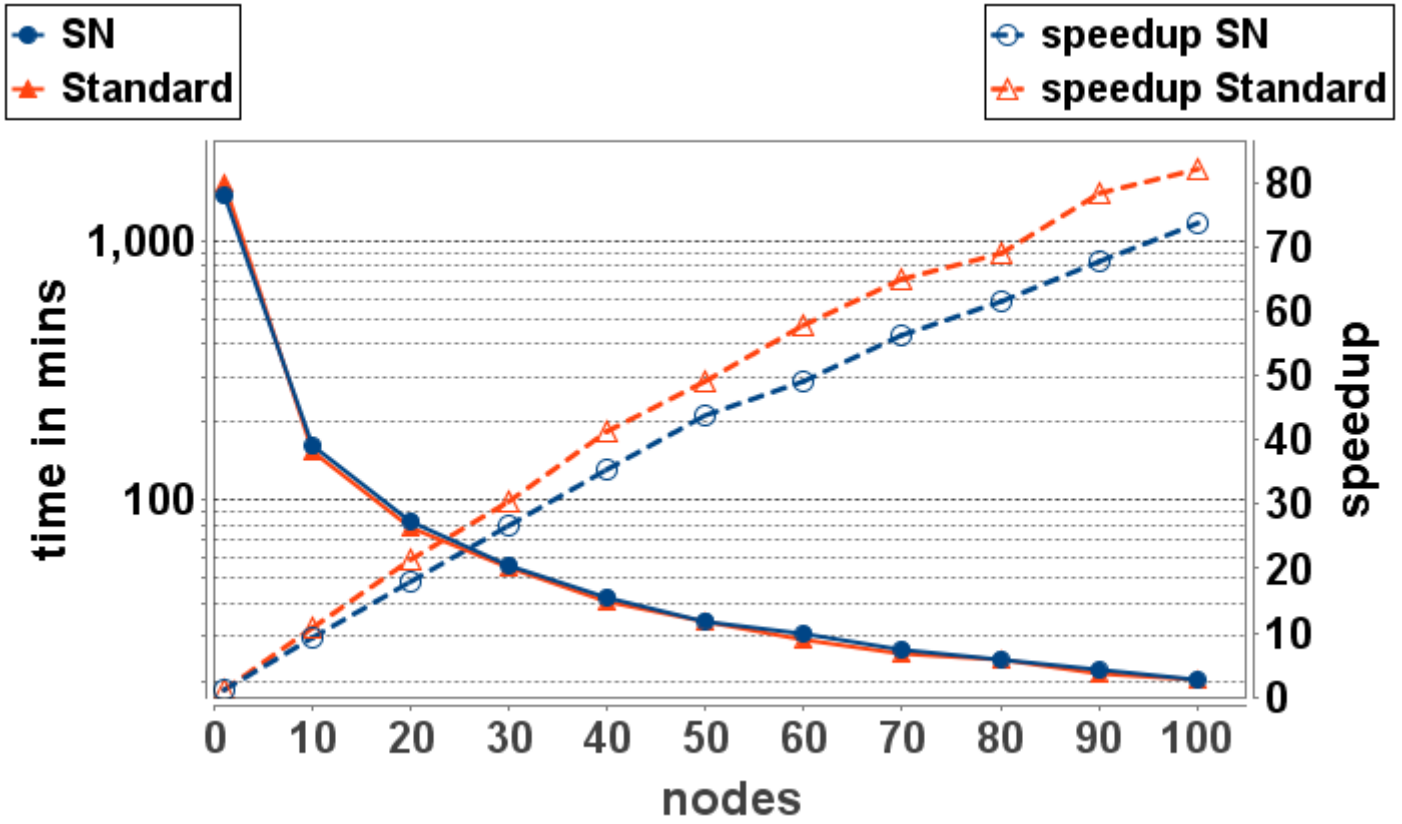
F-MEASURE RESULTS



- SVM slightly worse than Decision tree
- Sorted neighborhood (SN) includes additional matches found by transitive closure

ScaDS DRESDEN LEIPZIG

SCALABILITY RESULTS

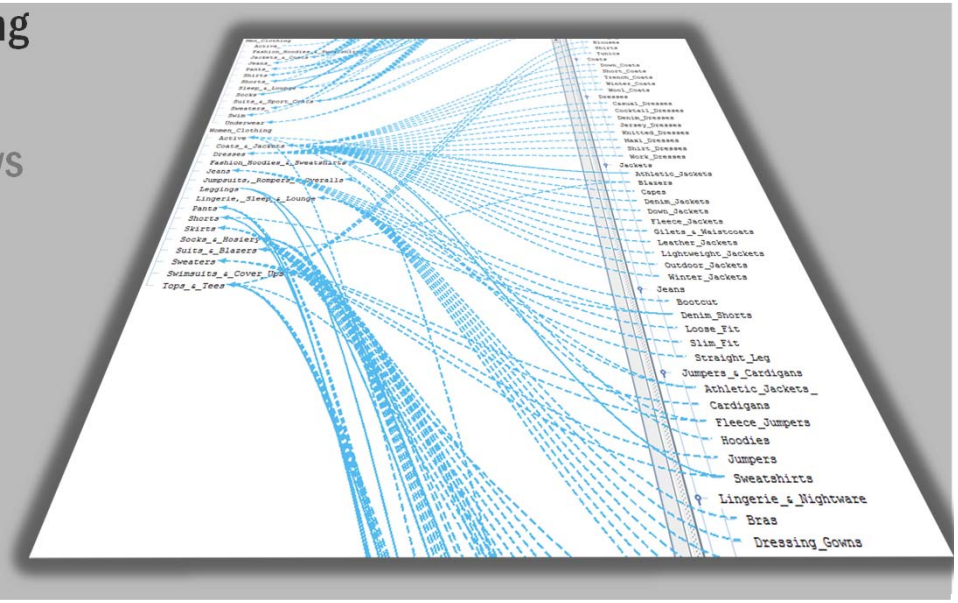


Citeseer dataset: ca 1.4 million publications

ScaDS DRESDEN LEIPZIG

AGENDA PART I (BIG DATA INTEGRATION)

- Introduction
- Scalable entity resolution / link discovery
- Large-scale schema/ontology matching
  - Introduction
  - Basic match techniques / workflows
  - Large-Scale Matching
  - Self-tuning match processes
  - Reuse-oriented matching
  - Match prototypes and products
  - Semantic Matching, Ontology Merging

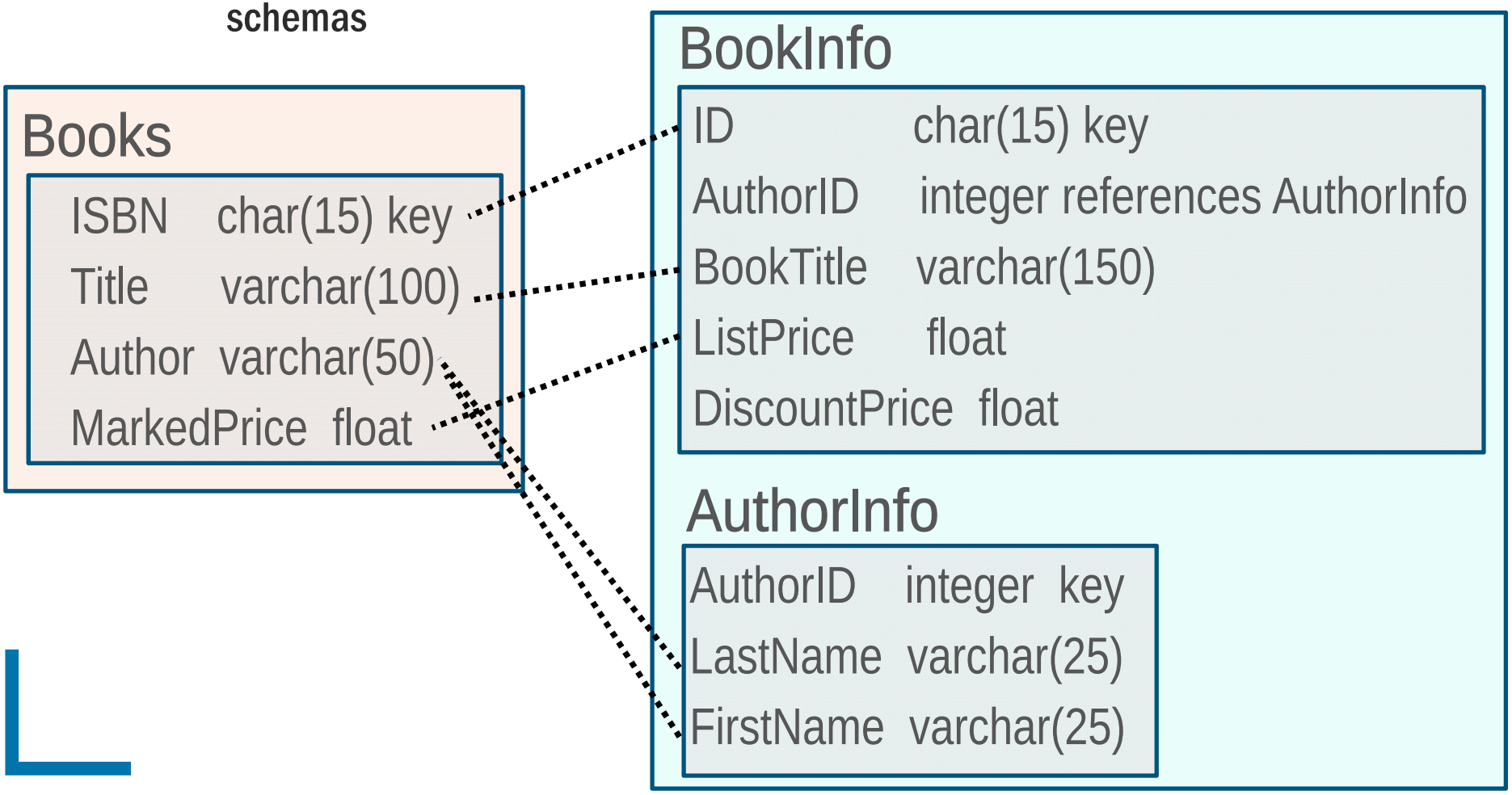


- Holistic data integration
- Summary

ScaDS DRESDEN LEIPZIG

THE SCHEMA MATCHING PROBLEM

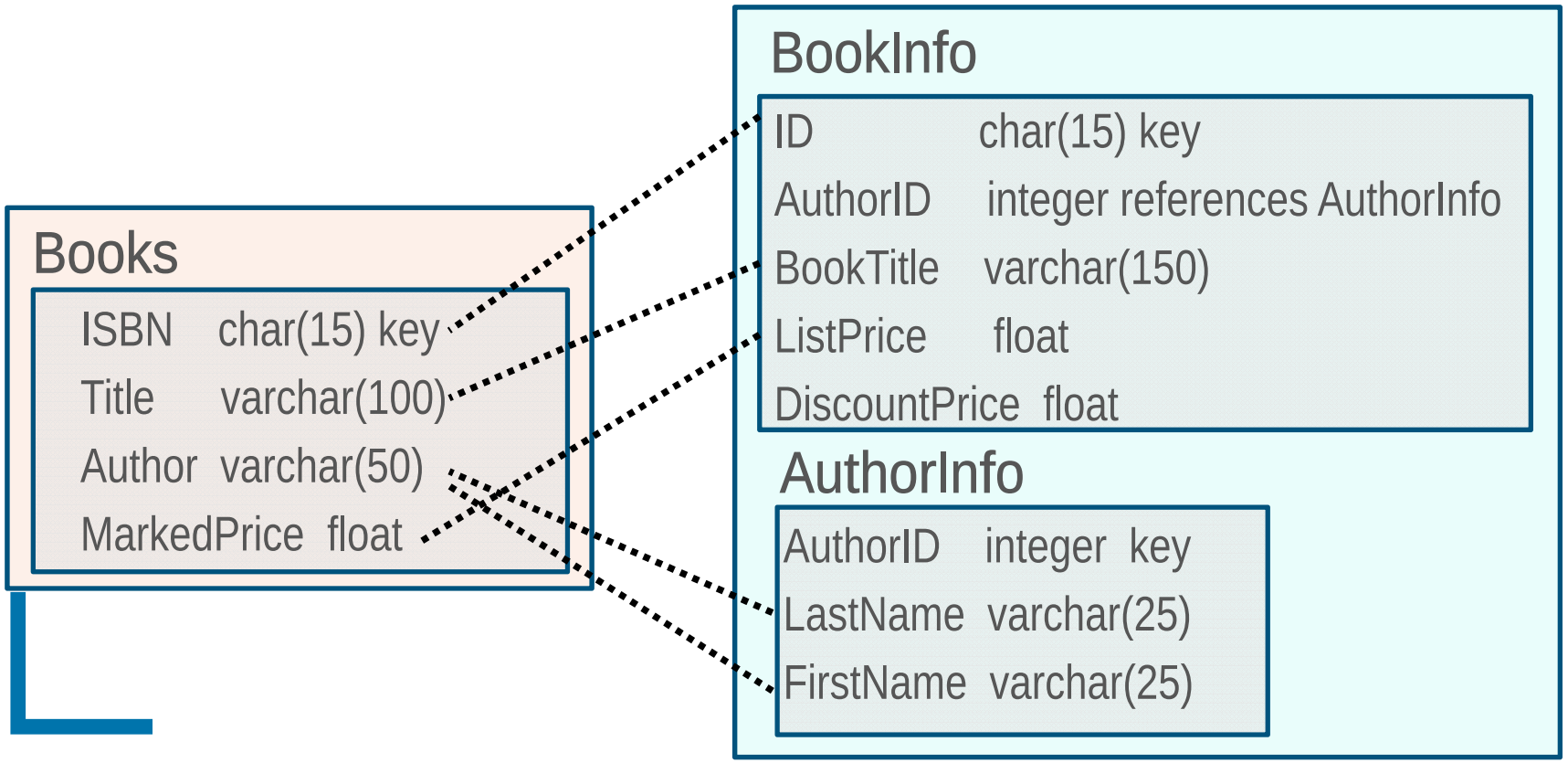
- The problem of generating correspondences between elements of two schemas



ScaDS DRESDEN LEIPZIG

BASIC INPUTS TO MATCHING TECHNIQUES

- Element names
- Schema structure
- Constraints: data type, keys, nullability



ScaDS DRESDEN LEIPZIG

OTHER INPUTS TO BASIC MATCHING

- Synonyms
  - Code = Id = Num = No
  - Zip = Postal [code]
  - Node = Server
- Acronyms
  - PO = Purchase Order
  - UOM = Unit of Measure
  - SS# = Social Security Number
- Data instances
  - Elements match if they have similar instances or value distributions

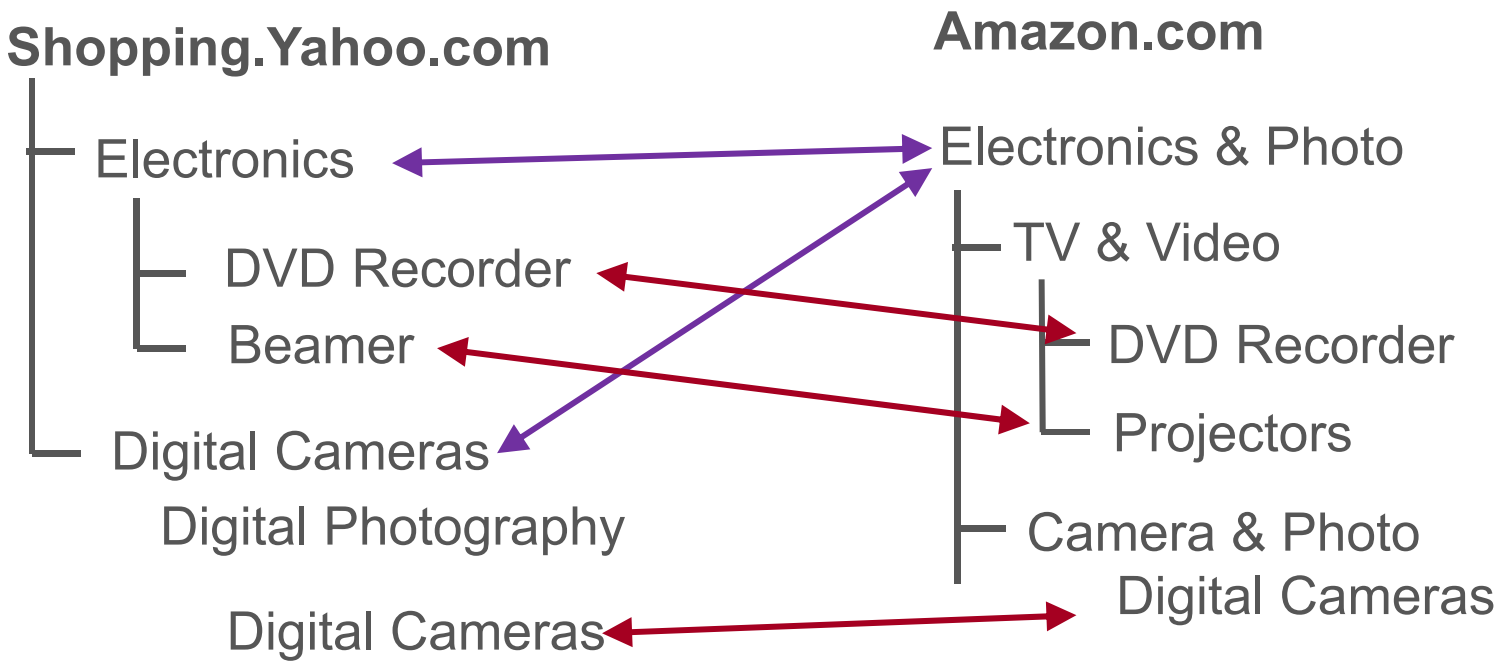
ScaDS DRESDEN LEIPZIG

MANY APPS NEED CORRESPONDENCES

- Data translation
  - Object-to-relational mapping
  - XML message translation
  - Data warehouse loading (ETL)
- Data integration
- ER design tools
- Schema evolution

ScaDS DRESDEN LEIPZIG

MATCHING OF PRODUCT CATALOGS

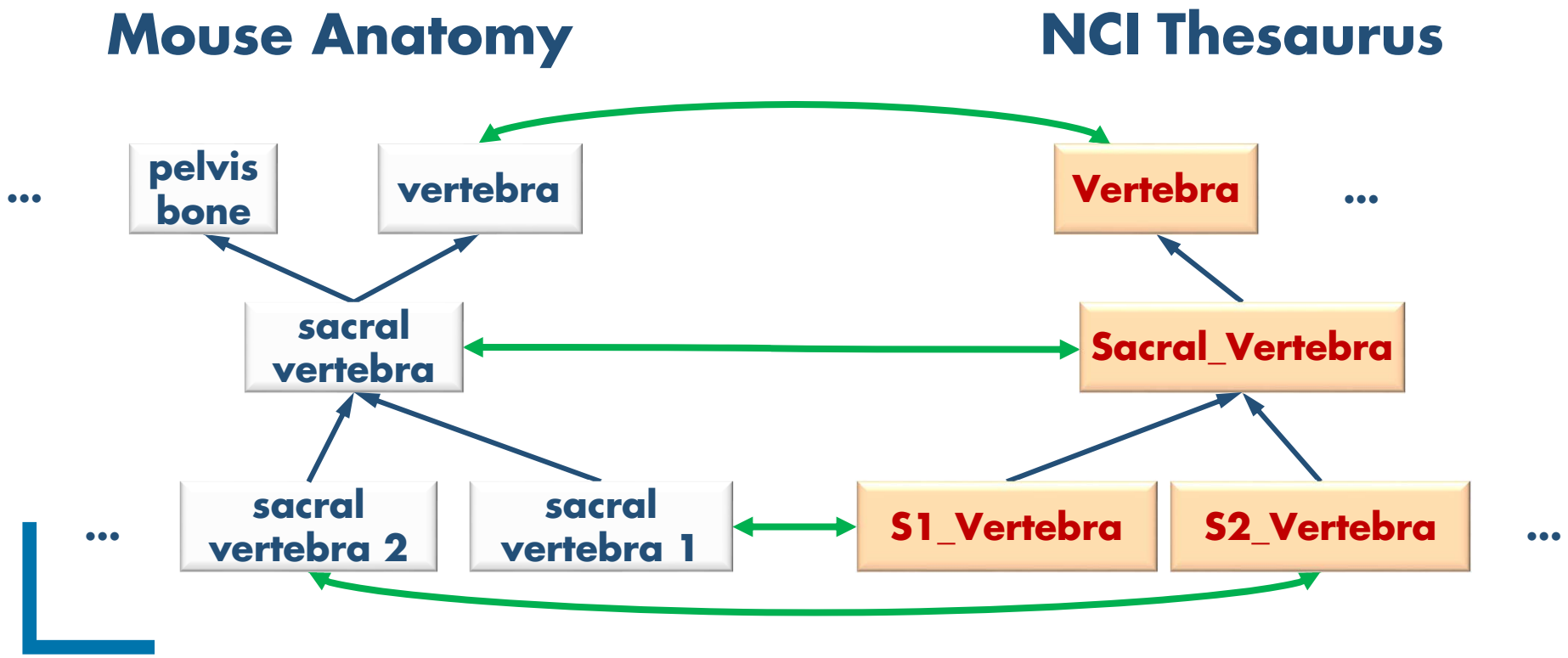


- Ontology mappings useful for
  - Improving query results, e.g. to find specific products across sites
  - Merging catalogs

ScaDS DRESDEN LEIPZIG

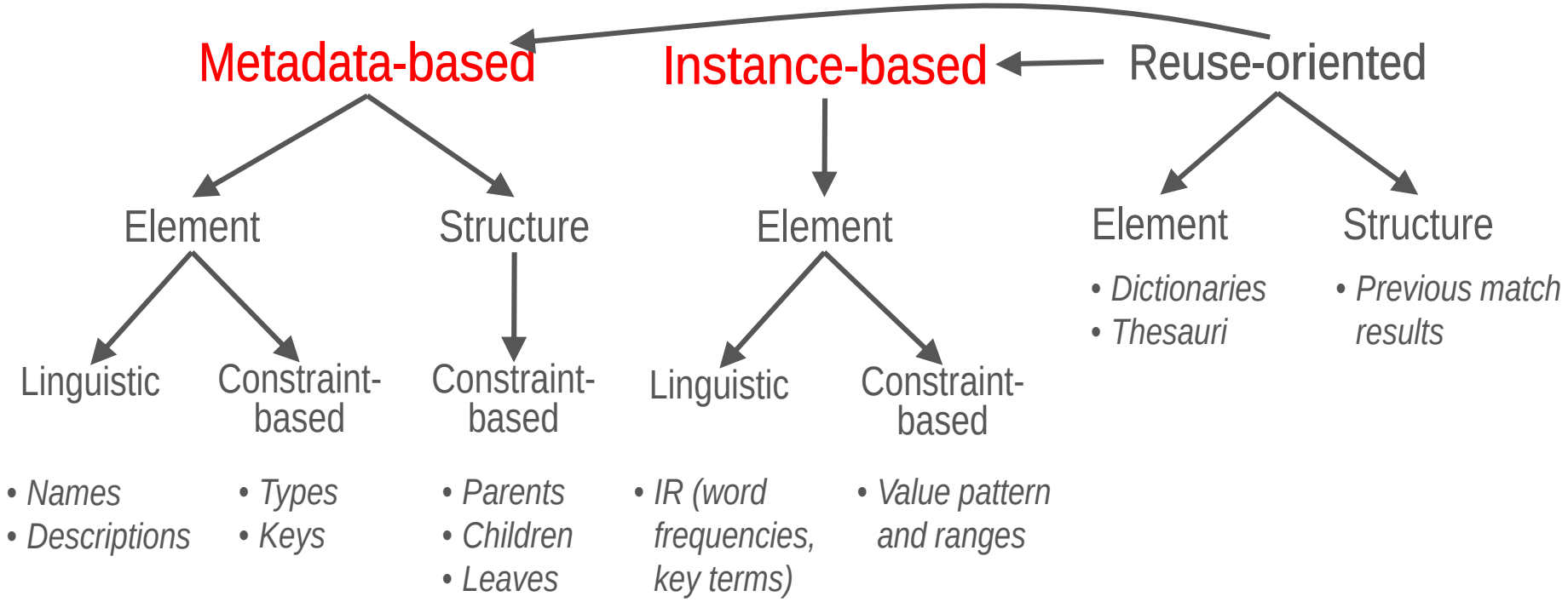
MATCHING LIFE SCIENCE ONTOLOGIES

- many large biomedical ontologies
  - used to annotate / enrich objects (genes, proteins ...) or documents (publications, electronic health records ...)



ScaDS DRESDEN LEIPZIG

AUTOMATIC MATCH TECHNIQUES*



- Matcher combinations
  - Hybrid matchers, e.g., considering name + type similarity
  - Composite matchers

* Rahm, E., P.A. Bernstein: *A Survey of Approaches to Automatic Schema Matching.* VLDB Journal 10(4), 2001

ScaDS DRESDEN LEIPZIG

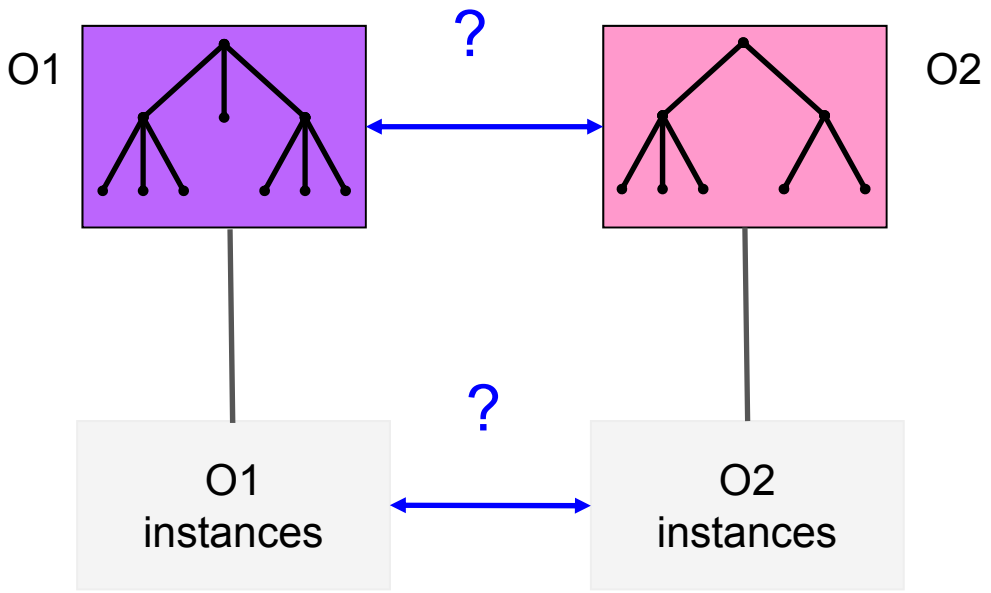
MATCH TECHNIQUES

- Linguistic matchers
  - (string) similarity of concept/element names
  - use of dictionaries/thesauri, e.g., WordNet / UMLS
- Structure-based matchers
  - consider similarity of ancestors/descendants
  - Graph-based matching (e.g., Similarity Flooding (Melnik, ICDE2002)
- Instance-based matchers
  - concepts with similar instances/annotated objects should match
  - consider all instances of a concept as a document and utilize document similarity (e.g., TF/IDF) to find matching concepts

ScaDS DRESDEN LEIPZIG

INSTANCE-BASED ONTOLOGY MATCHING

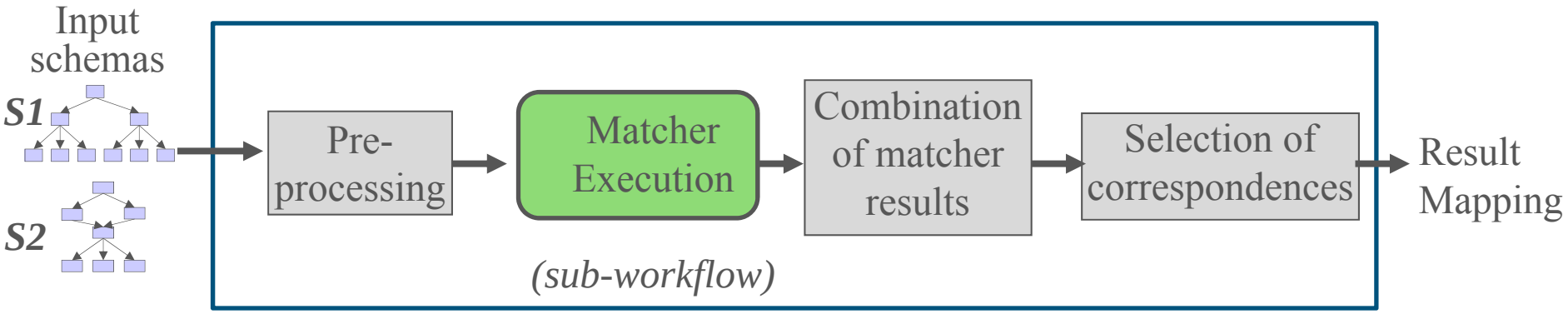
- concepts with most similar instances should match
  - requires shared/similar instances for most concepts
- mutual treatment of entity resolution (instance matching) and ontology matching
- promising for link discovery in the Linked Open Web of Data



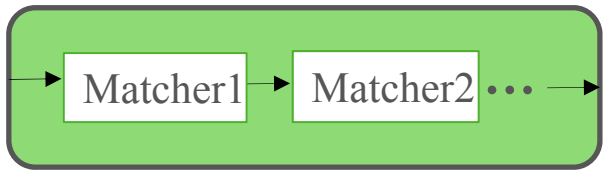
ScaDS DRESDEN LEIPZIG

SCHEMA MATCHING IS A MULTI-STEP PROCESS

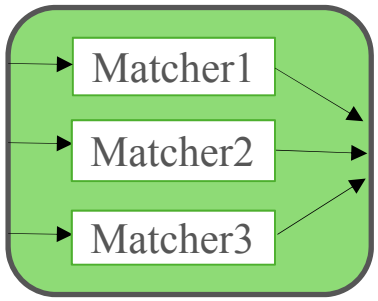
General workflow (COMA, ...)



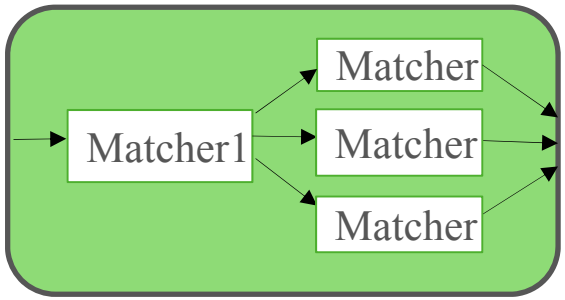
Matcher sub-workflows



Sequential matchers



Parallel (independent) matchers



Mixed strategy

ScaDS DRESDEN LEIPZIG

LARGE-SCALE MATCHING

Very large ontologies / schemas (>10.000 elements)

- quadratic complexity of evaluating the cartesian product (match efficiency)
- difficult to find all right correspondences (match quality)
- support for user interaction

Many (>>2) ontologies/schemas

- holistic ontology/schema matching
- clustering of equivalent concepts/elements or linking to some hubs

ScaDS DRESDEN LEIPZIG

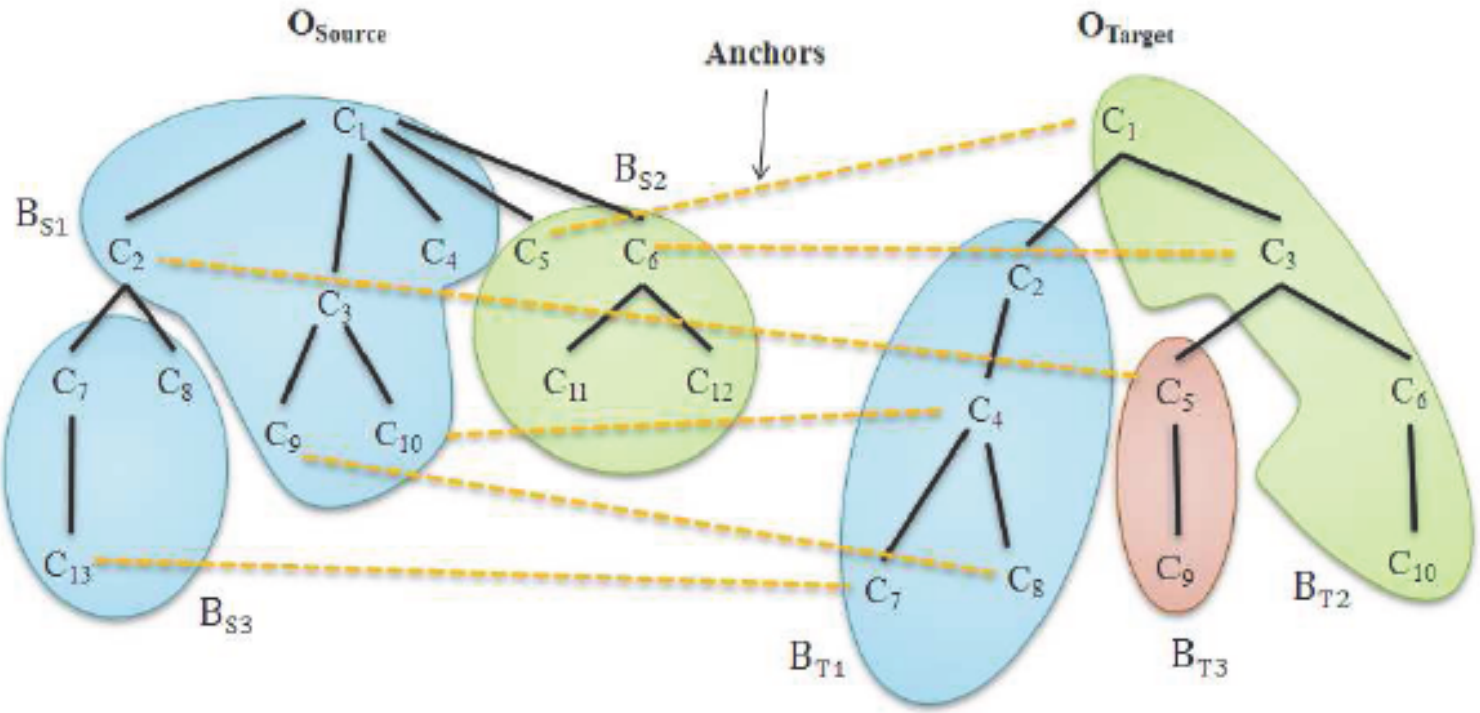
MATCH TECHNIQUES FOR LARGE SCHEMAS

- Low-level optimizations
  - optimized string matching
  - space-efficient similarity matrices
- Parallel matching
  - inter-matcher and intra-matcher parallelism
- Partition-based matching (COMA++, Falcon-AO)
  - reduced search space by matching only similar schema partitions/fragments
  - light-weight search for similar schema fragments

ScaDS DRESDEN LEIPZIG

PARTITION-BASED MATCHING IN FALCON-AO

- initially determine highly similar element pairs called “anchors”
- only partitions that share at least one anchor are matched



[Hamdi et al, 2009]

ScaDS DRESDEN LEIPZIG

SELF-TUNING MATCH WORKFLOWS (1)

- Semi-automatic configuration
  - Selection and ordering of matchers
  - Combination of match results
  - Selection of correspondences (top-k, threshold, ...)
- Initial tuning frameworks: Apfel, eTuner, YAM
- Use of supervised machine learning
  - need previously solved match problems for training
  - difficult to support large schemas

ScaDS DRESDEN LEIPZIG

SELF-TUNING MATCH WORKFLOWS (2)

- **Heuristic approaches**
  - use linguistic and structural similarity of input schemas to select matchers and their weights (RiMOM)
  - favor matchers giving higher similarity values in the combination of matcher results (QOM, PRIOR+, OpenII)
- **Rule-based approach (Peukert/Rahm, ICDE2012)**
  - comprehensive rule set to determine and tune match workflow
  - use of schema features and intermediate match results

ScaDS DRESDEN LEIPZIG

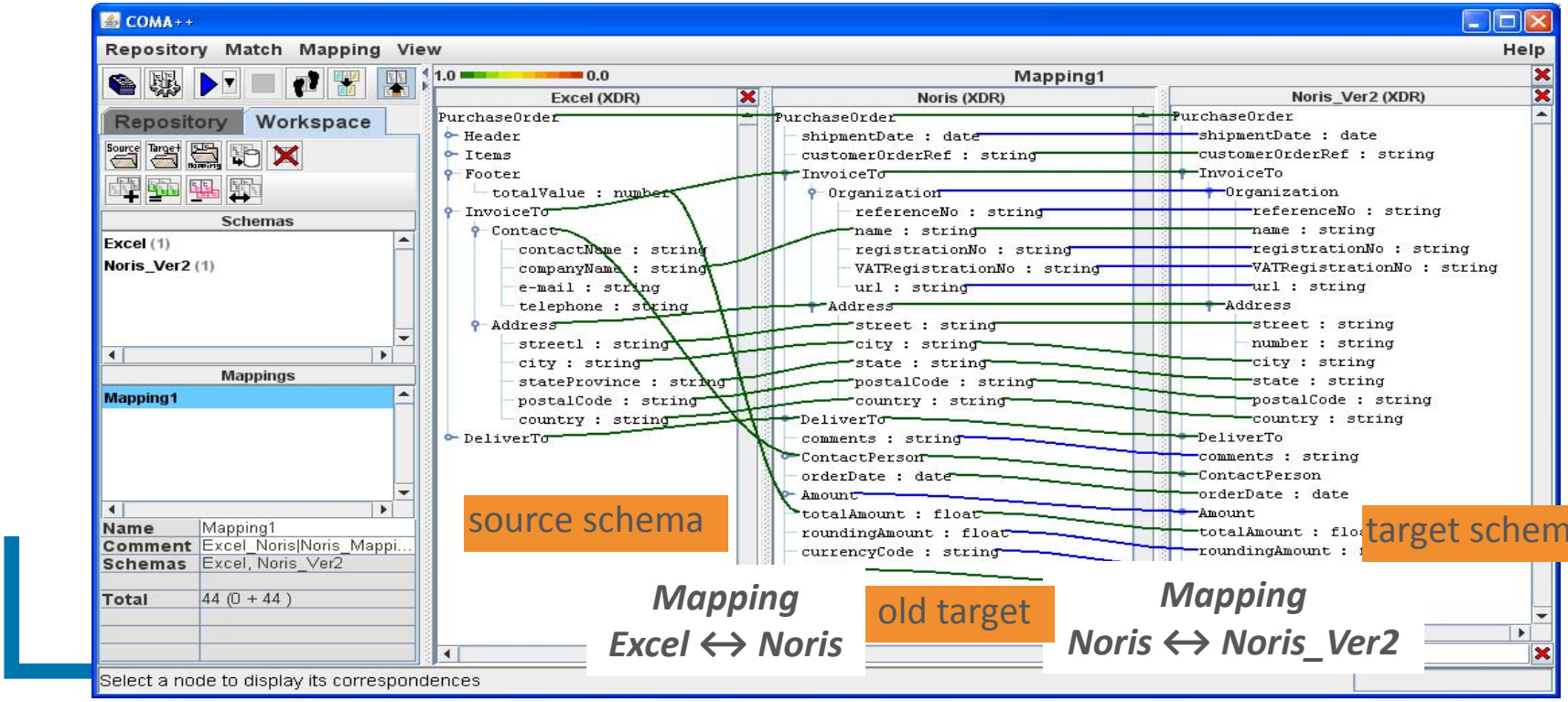
REUSE-ORIENTED MATCHING

- Many similar match tasks → reuse previous matches
  - can improve both efficiency and match quality
- Repository needed
  - store previously matched schemas/ontologies and obtained mappings
  - identify and apply reusable correspondences
- First proposals for reuse at 3 mapping granularities
  - reuse *individual element correspondences,* e.g. synonyms
  - reuse *complete mappings*, e.g. after schema/ontology evolution
  - reuse *mappings between schema/ontology fragments* (e.g., common data elements / CDE)

ScaDS DRESDEN LEIPZIG

MAPPING REUSE

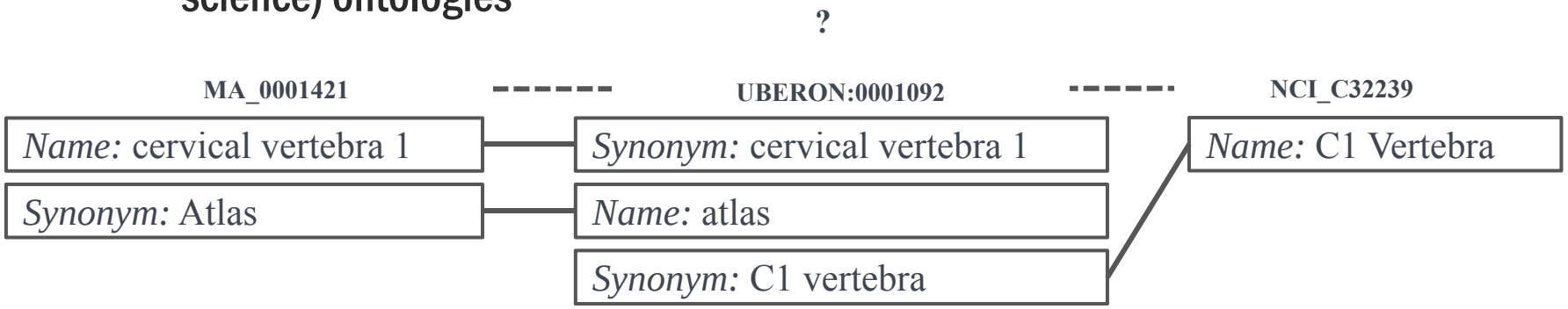
- Supported in match prototypes COMA and Gomma (Leipzig Univ.)
- Example: reuse match results after schema evolution
  - compose previous match result S—T with mapping T-T' to solve new match task S—T'



ScaDS DRESDEN LEIPZIG

MAPPING REUSE IN GOMMA

- comprehensive use of mapping composition to indirectly match (life science) ontologies

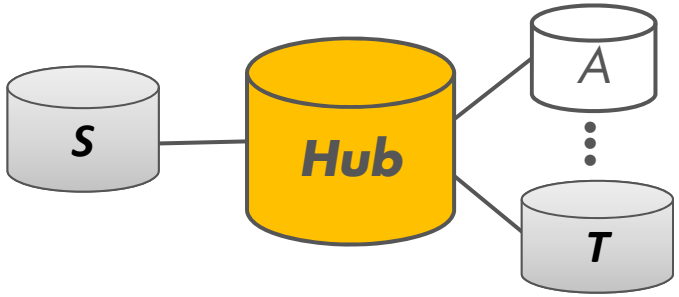


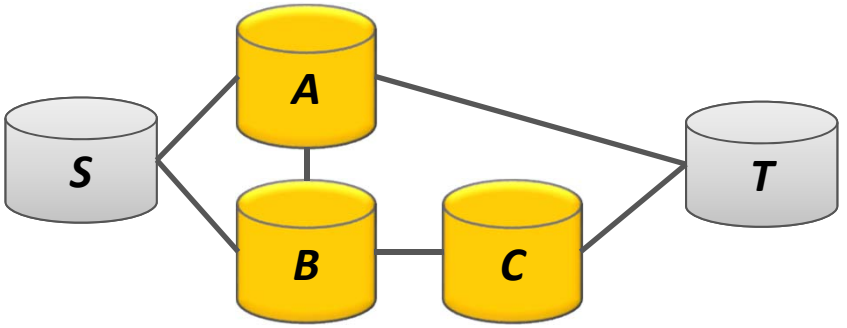
- utilizes both previous mappings and synonyms
- very fast and effective by reusing previously confirmed correspondences

ScaDS DRESDEN LEIPZIG

MAPPING REUSE IN GOMMA (2)

- effective exploitation of existing mappings and „hub“ ontologies (e.g. UMLS or Uberon in biomedicine)
- methods to determine most promising mappings and composition paths to reuse (and combine)





- additional direct matching of „uncovered“ concepts that are not mapped by previous mappings
- indirect matching helped to achieve very good results in OAEI contest (e.g., 92% F-Measure for anatomy)

A. Gross et al.: Mapping Composition for Matching Large Life Science Ontologies. Proc. Int. Conf. on Bio-Ontologies 2011

ScaDS DRESDEN LEIPZIG

FRAGMENT-LEVEL REUSE

- Related ontologies / schemas mostly overlap in some portions
  - standard match approaches try to map everything
  - reuse at level of entire mappings of limited help
- Reuse of ontology/schema fragments helps to reduce heterogeneity
  - e.g. CDE on customer, purchase orders, ...
  - reuse of correspondences at fragment level
- Most complex reuse approach
  - populate repository by most relevant fragments/CDE and their mappings
  - analyze schemas to be matched for fragment pairs in the repository
  - assemble and complement fragment mappings

UNIVERSITÄT LEIPZIG

ScaDS

RESEARCH MATCH PROTOTYPES

ScaDS DRESDEN LEIPZIG

MATCH PROTOTYPE COMPARISON*

		Cupid	COMA++	Falcon	Rimom	Asmov	Agr.Maker	OII Harmony
year of introduction		2001	2002/2005	2006	2006	2007	2007	2008
Input	*relational*	√	√	-	-	-	-	√
schemas	*XML*	√	√	-	-	-	(√)	√
	ontologies	-	√	√	√	√	√	√
OAEI participation		-	√	√	√	√	√	-
compreh. GUI		-	√	(√)	?	?	√	√
Matchers	*linguistic*	√	√	√	√	√	√	√
	structure	√	√	√	√	√	√	√
	Instance	-	√	-	√	√	√	-
use of ext.dictionaries		√	√	?	√	√	√	√
schema partitioning		-	√	√	-	-	-	-
parallel matching		-	-	-	-	-	-	-
dyn. matcher selection		-	-	-	√	-	-	-
mapping reuse		-	√	-	-	-	-	-

*Rahm, E.: Towards large-scale schema and ontology matching. In: Schema Matching and Mapping, Springer-Verlag, 2011

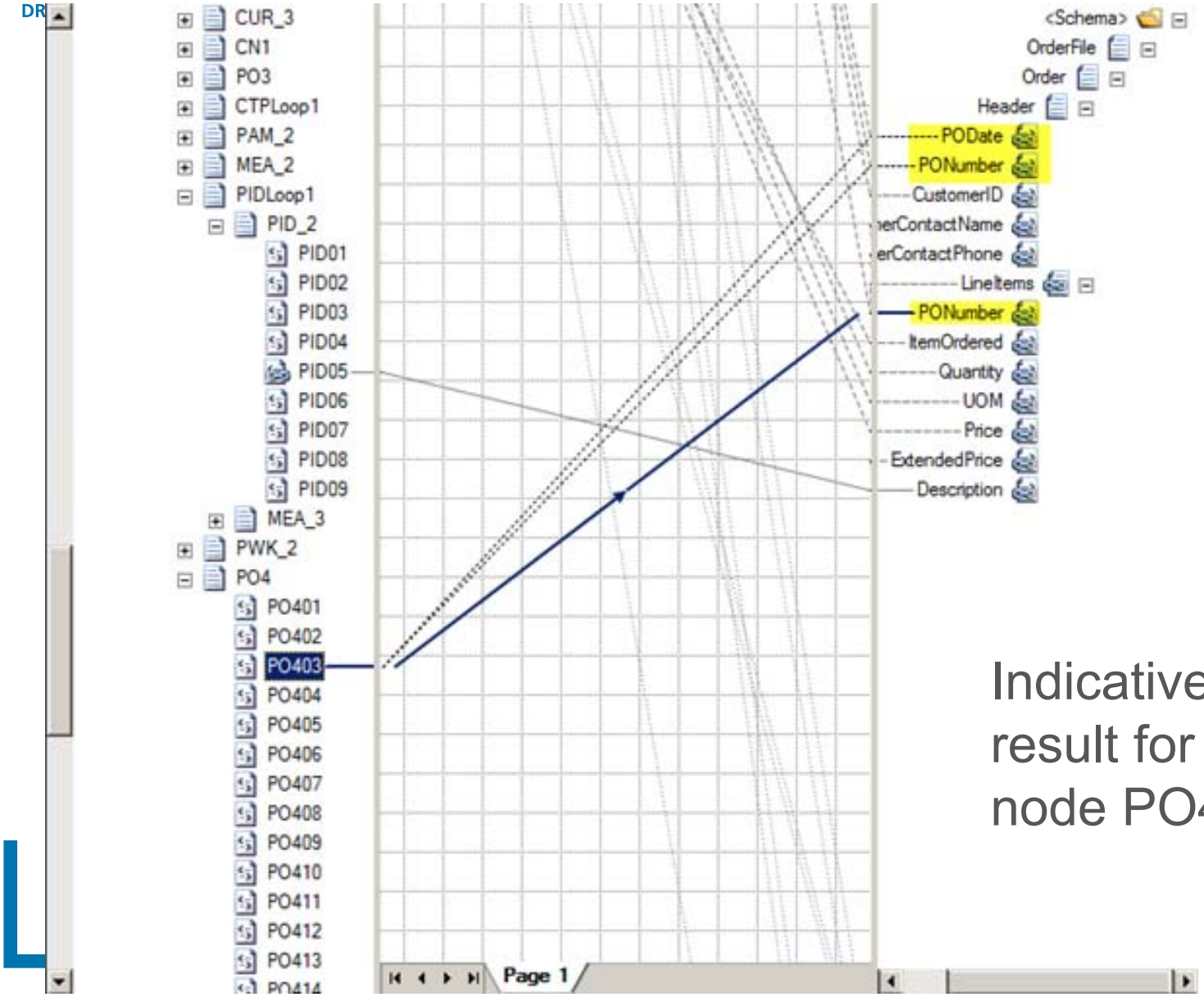
ScaDS DRESDEN LEIPZIG

COMMERCIAL SCHEMA MATCHING TOOLS

- Many GUI-based mapping editors to manually specify correspondences and mappings
- Initial support for automatic matching, in partiular linguistic matching
  - Altova MapForce
  - MS BizTalk Server
  - SAP Netweaver
  - IBM Infosphere
- Many further improvements possible
  - Structural / instance-based matching
  - Advanced techniques for large schemas

ScaDS DR

BIZTALK SCREENSHOT



Indicative match result for selected node PO403

ScaDS DRESDEN LEIPZIG

SEMANTIC MATCHING

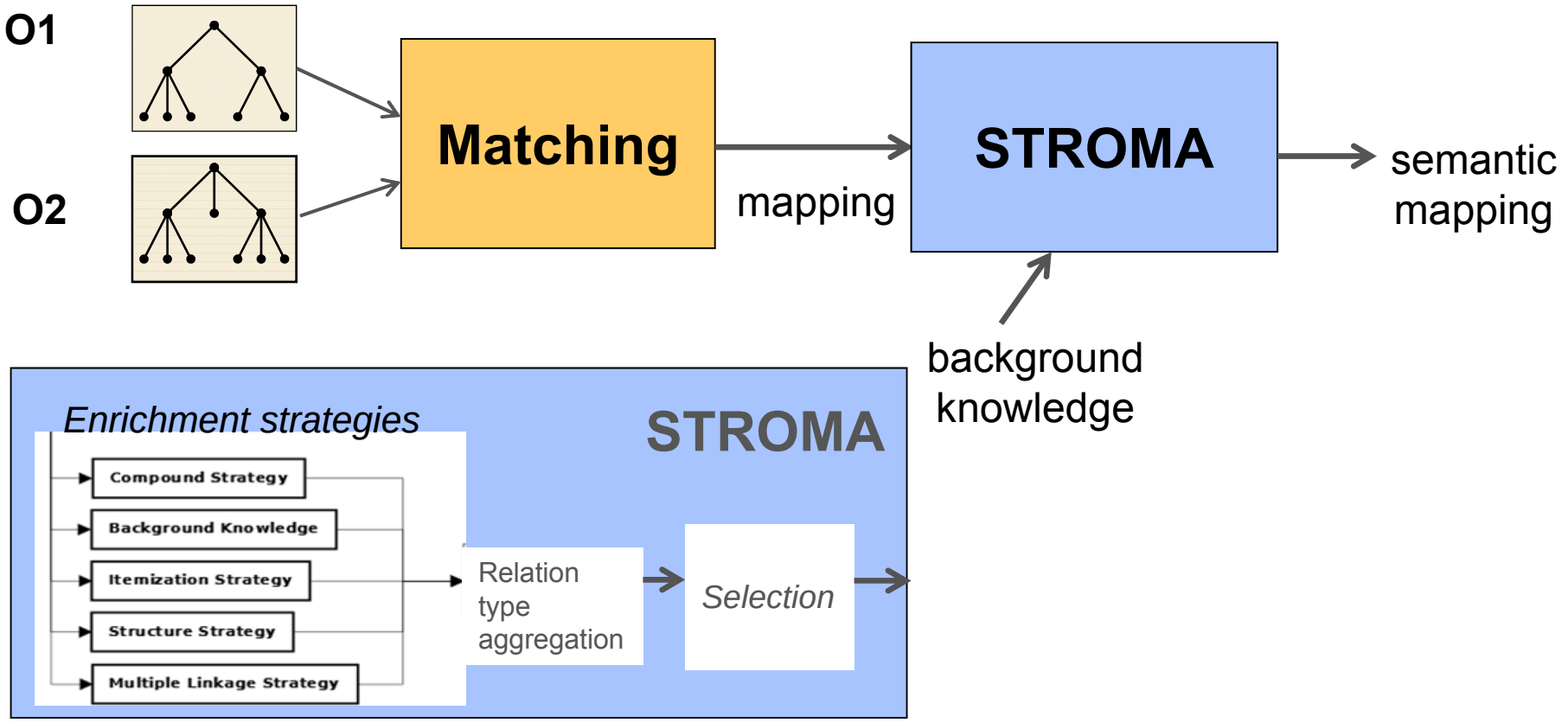
- Correspondences with semantic relationships equality, less general (is-a)/more general, part-of/has, disjointness
  - tablet *is-a* portable computer
  - computer *has* memory

	S-Match	TaxoMap	Aroma	ASMOV	STROMA
Architecture	1-step	1-step	1-step	1-step	2-step
Supported types	equal, is-a, related	equal, is-a, related	equal, is-a. disjoint	equal, is-a	equal, is-a, part-of, related
Background sources	WordNet	WordNet		WordNet	WordNet, UMLS, OpenThesaurus
Primary techniques	linguistic	linguistic	probabilistic, instance-based	linguistic, structural, instance-based	linguistic, structural

Arnold, P.; Rahm, E.: **Enriching Ontology Mappings with Semantic Relations**. Data and Knowledge Engineering, 2014

ScaDS DRESDEN LEIPZIG

STROMA: SEMANTIC REFINEMENT OF MAPPINGS

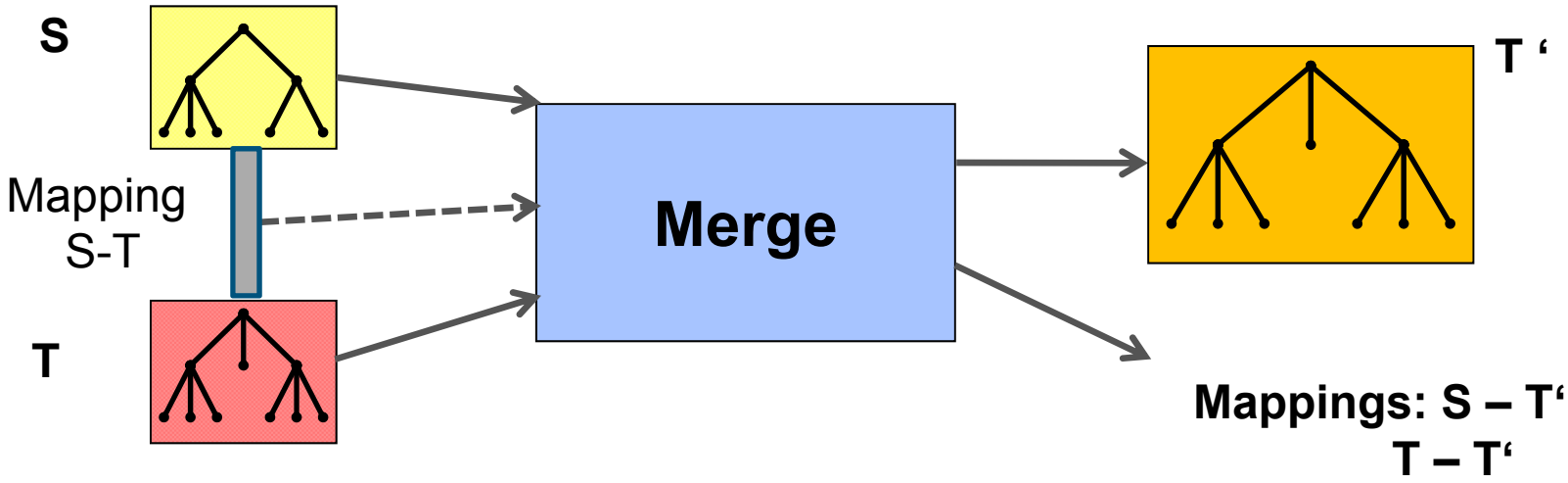


- Compound strategy: portable computer *is-a* computer
- Composition: (laptop, is-a, computer), (computer, has, memory)
  -> (laptop, has, memory)

ScaDS DRESDEN LEIPZIG

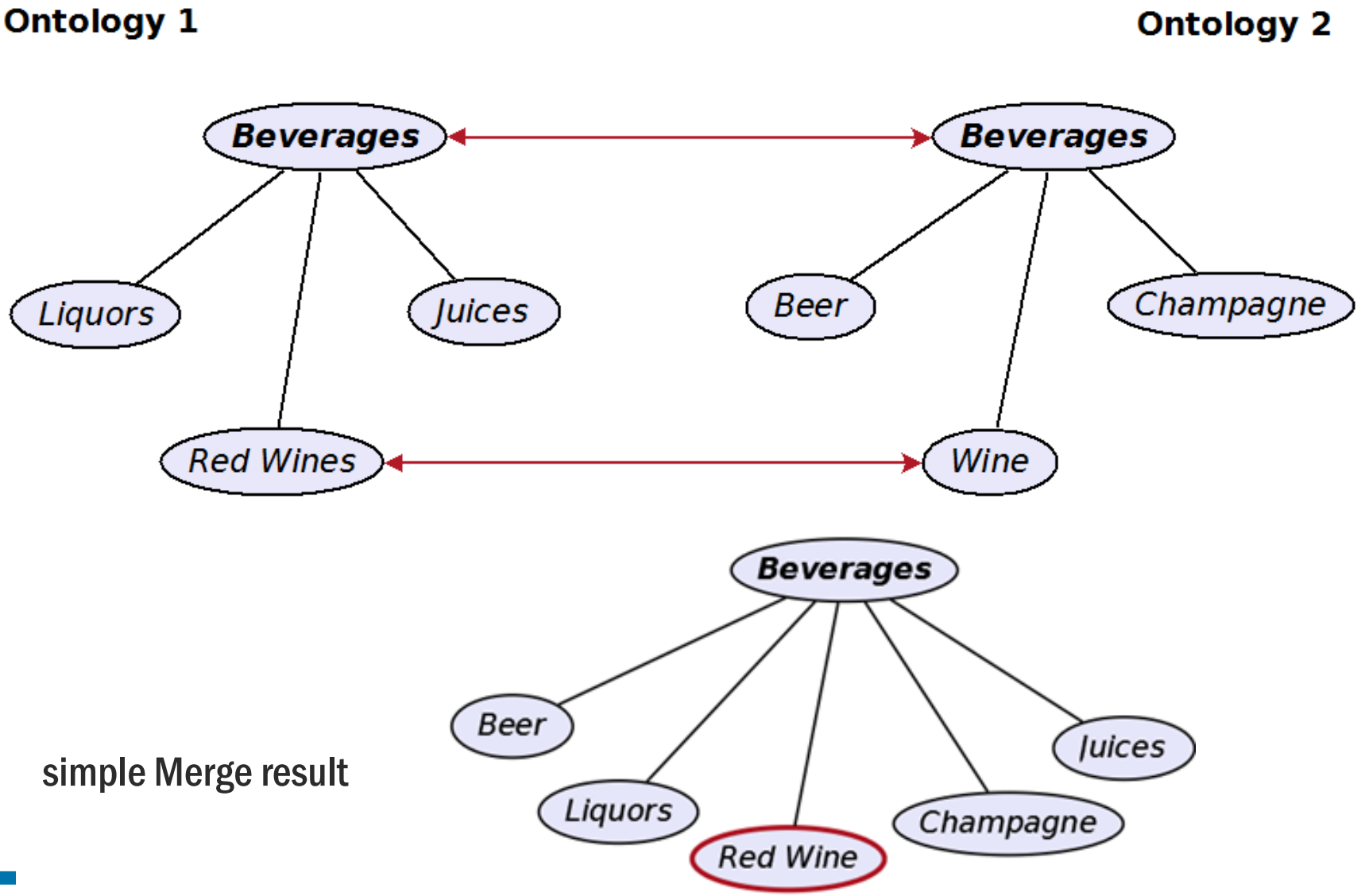
ONTOLOGY MERGING

- Process of merging input ontologies into integrated ontology
  - symmetric merge or
  - asymmetric, target-driven merge
- optional use of (simple or semantic) match mapping between input ontologies



ScaDS DRESDEN LEIPZIG

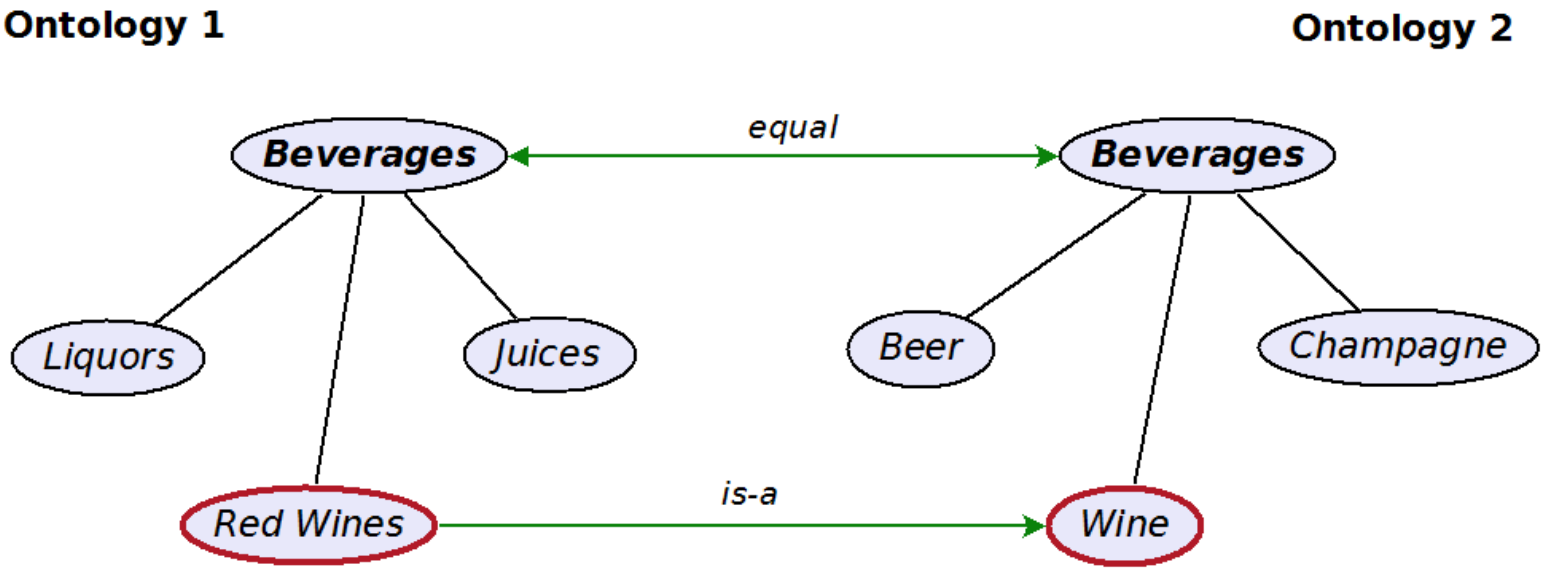
SIMPLE MATCH + MERGE



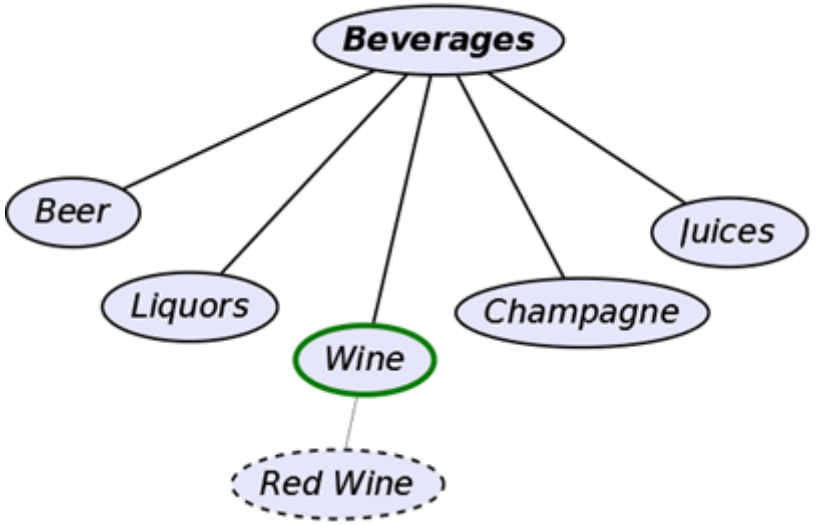
simple Merge result

ScaDS DRESDEN LEIPZIG

SEMANTIC MATCH + MERGE



improved
Merge result



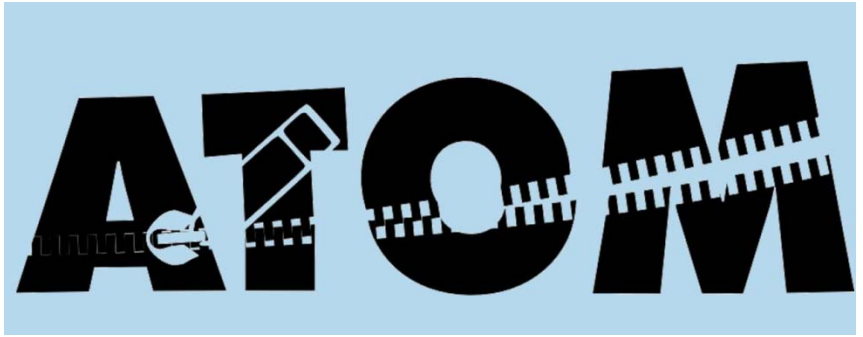
ScaDS DRESDEN LEIPZIG

PREVIOUS WORK ON MERGE

- huge amount of work on schema integration
  - mostly addressed both matching and merging
  - complex solutions with high degree of manual interaction
- more recent schema merging approaches based on predetermined match mapping
  - [Pottinger and Bernstein 2003], [Pottinger and Bernstein 2008]
  - [Chiticariu, Kolaitis, Popa 2008], [Radvan, Popa , Stanoi, Younis 2009] …
- relatively few approaches for ontology merging
  - PROMPT (1999-2000), Chimaera (2000), FCA-Merge (2001), …
  - combined approaches for match and merge
  - high degree of user intervention needed
  - symmetric merge (full preservation of both input ontologies)

ScaDS DRESDEN LEIPZIG

ATOM APPROACH*



- **A**utomatic **T**arget-Driven **O**ntology **M**erging
- asymmetric, target-driven merge approach
  - preserves target ontology but may drop source concepts and relationships that would introduce redundancy in the merge result
- utilization of input match mapping
  - base version: equivalence correspondences
  - improved version: semantic correspondences
- automatic generation of default solution(s)
  - result may interactively be adapted by users if needed

* Raunich, S., Rahm, E.: *Target-driven Merging of Taxonomies with ATOM.* Information Systems, 2014

ScaDS DRESDEN LEIPZIG

AGENDA PART I (BIG DATA INTEGRATION)

- Introduction
- Scalable entity resolution / link discovery
- Large-scale schema/ontology matching
- Holistic data integration
  - Introduction
  - Use cases
  - Holistic schema matching
  - Knowledge graphs
  - Web tables
- Summary

ScaDS DRESDEN LEIPZIG

HOLISTIC DATA INTEGRATION

- 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
  - hundreds of LOD sources
  - many thousands of web shops
  - many millions of web tables
- Large open data /metadata/mapping repositories
  - data.gov, datahub.io, www.opensciencedatacloud.org, web-datacommons.org
  - schema.org, medical-data-models.org
  - BioPortal, LinkLion

ScaDS DRESDEN LEIPZIG

HOLISTIC DATA INTEGRATION: USE CASES (1)

- Query mediator, e.g., for LOD query access (e.g., FedX system)
  - virtual data integration
  - with or without global schema
  - few sources
- Mediated web queries (e.g., MetaQuerier)
  - mediated schema (schema clustering)
  - virtual data integration
  - tens of data sources
- Integrated domain ontology (e.g., UMLS)
  - physical metadata integration
  - tens of source ontologies
  - clustering of synonymous concepts (synsets)
  - largely manual integration effort

ScaDS DRESDEN LEIPZIG

HOLISTIC DATA INTEGRATION: USE CASES (2)

- Entity search engines (Google scholar, Google shopping)
  - clustering of matching entities (publications, product offers)
  - physical data integration
  - thousands of data sources
- Comparison / booking portals (pricegrabber.com, booking.com …)
  - clustered offers within (integrated) taxonomy
  - physical or virtual data integration
- Web-scale knowledge graphs (Google, Facebook, Microsoft)
  - physical integration of data and metadata
  - highly automated
  - challenging data quality issues
- Web table repository (e.g., Google fusion tables, WDC web table corpora)
  - physical data collection with milions of tables
  - little integration (domain categorization, attribute linking)

ScaDS DRESDEN LEIPZIG

USE CASE CHARACTERISTICS

Use case	Data integration type		#domains	#sources	Clustering?	degree of automated data integration
Query mediator	virtual	metadata	1+	small	-	low
Meta web query	virtual	metadata	1	small	attributes	medium
Integrated ontology	physical	metadata	1+	small	concepts	low
Entity search engines	physical	data	1	very high	entities	very high
Booking portals	physical	data + metadata	1+	high	entities	high
Knowledge graphs	physical	data + metadata	many	medium	entities + concepts/ attributes	high
Web table corpus	physical	primarily data	many	very high	possible	very high, but limited integration

ScaDS DRESDEN LEIPZIG

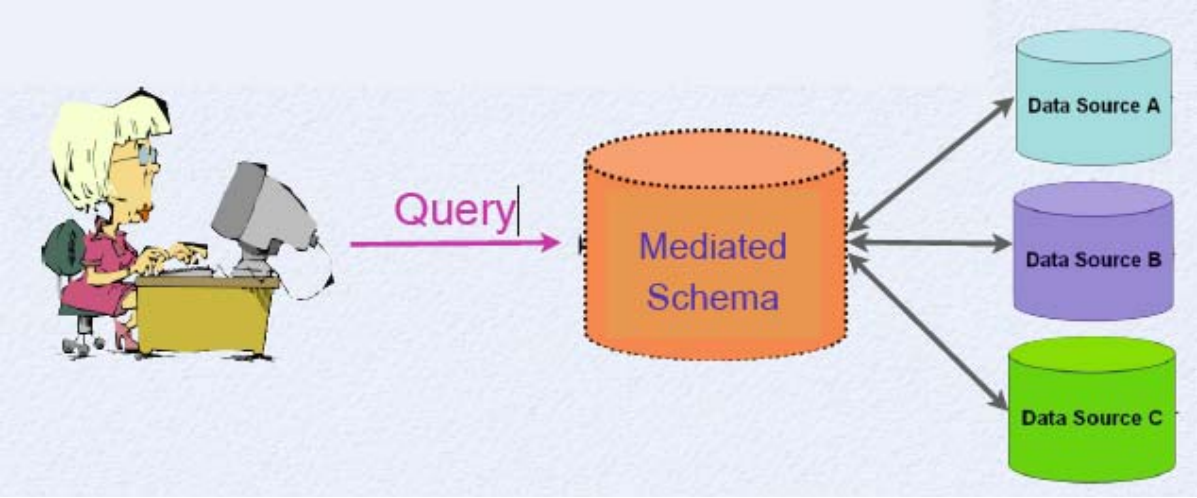
OBSERVATIONS

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

ScaDS DRESDEN LEIPZIG

HOLISTIC (COLLECTIVE) SCHEMA MATCHING

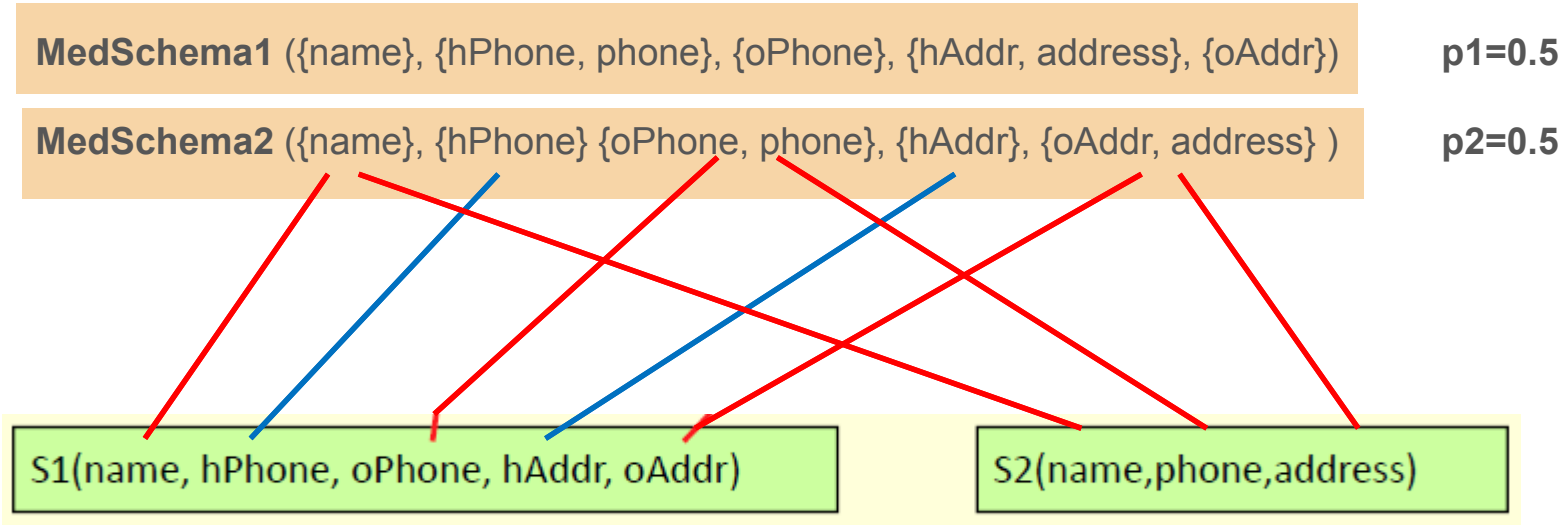
- Matching between N schemas, e.g. web forms
  - mostly simple schemas
- Typical use case: creation of a mediated schema
- Holistic matching based on clustering of similar attributes (Wise-Integrator, DCM, HSM, ...)
  - utilize high name similarity between schemas
  - similar names within a schema are mismatches (e.g. first name, last name)



ScaDS DRESDEN LEIPZIG

HOLISTIC (COLLECTIVE) SCHEMA MATCHING

- Probabilistic mediated schemas [Das Sarma et al., SIGMOD 2008]
  - first determine several *probabilistic mappings*
  - determine and rank attribute clusters
  - use all mediated schemas to answer queries and rank query results
  - fully automatic approach



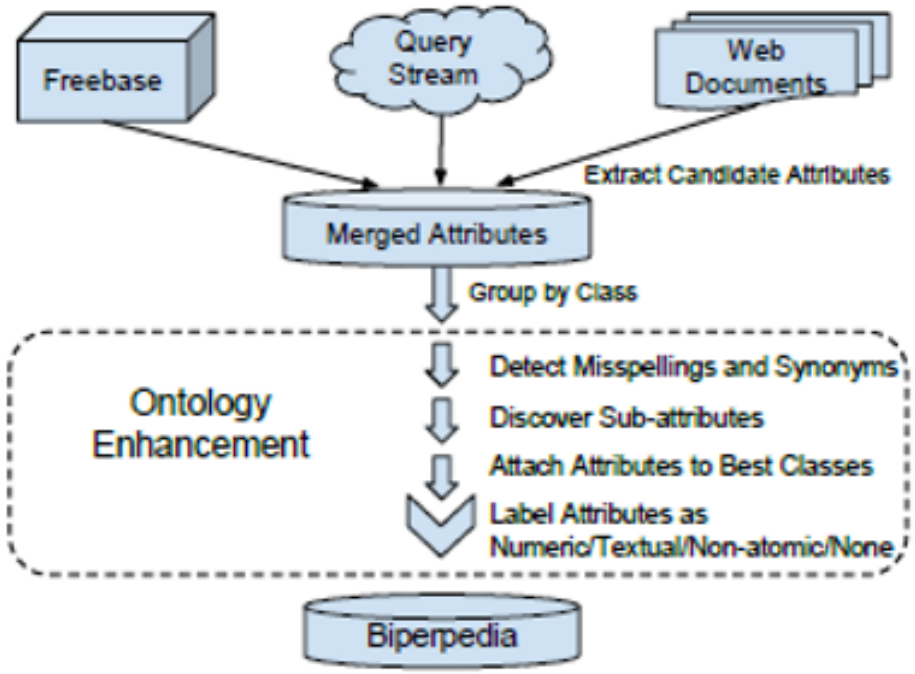
ScaDS DRESDEN LEIPZIG

KNOWLEDGE GRAPHS

- representation of background knowledge (data + metadata) for
  - enhancing entities (based on prior *entity linking*)
  - improving data integration (e.g., by utilizing additional information)
  - improving search results …
- comprehensive taxonomies to categorize entities and their details
  - extreme entity heterogeneity (attributes + values) even within domains
- construction of knowledge graphs is itself challenging data integration problem
  - use of existing knowledge sources, e.g., DBpedia, Freebase, Yago, bio-ontologies
  - extractions from websites and text documents

ScaDS DRESDEN LEIPZIG

SAMPLE KNOWLEDGE GRAPH: BIPERPEDIA



Attribute: CAPITAL (***Class***: COUNTRIES, ***Type***: atomic-textual)
Primarily attached to: LOCATIONS
Synonyms: CAPITALS
Misspells: CAPITAI, CAPITA, CAPTAL, CAPTEL, CAPIT, ...
Sub relations: CITY CAPITAL, FORMER CAPITAL, FASHION CAPITAL, ...
Provenance:

Source	*InstanceCount*	*QueryCount*	*Entities with this attribute*
Query Stream	317	11M	vietnam, turkey, romania, ...
Text	441	3M	afghanistan, iraq, pakistan, ...
Freebase	n/a	n/a	n/a

Text forms: "Hanoi is the **capital** of **Vietnam**.", "Beijing, the **capital** of **China**, is ..."
Query forms: "**capital Brazil**", "What is the **capital** of **Australia**?"

Figure 1: The elements of a Biperpedia attribute

- extends Freebase by information extracted from search queries and documents
- focus on finding additional attribute synonyms, misspellings, subattributes (A is-a B), type (numerical, atomical text, non-atomic) for existing concepts
- more than 10K classes, 67K unique attributes, 1.6 M class-attribute pairs

R. Gupta, A. Halevy, X. Wang, S. Whang, F. Wu: Biperpedia: An Ontology for Search Applications. PVLDB 2014

ScaDS DRESDEN LEIPZIG

GOOGLE KNOWLEDGE GRAPH

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_Univ**ersity ▾
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
★★★★½ 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.**uni**klinikum-**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
https://twitter.com/**unileipzig** ▾ Translate this page

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

ScaDS DRESDEN LEIPZIG

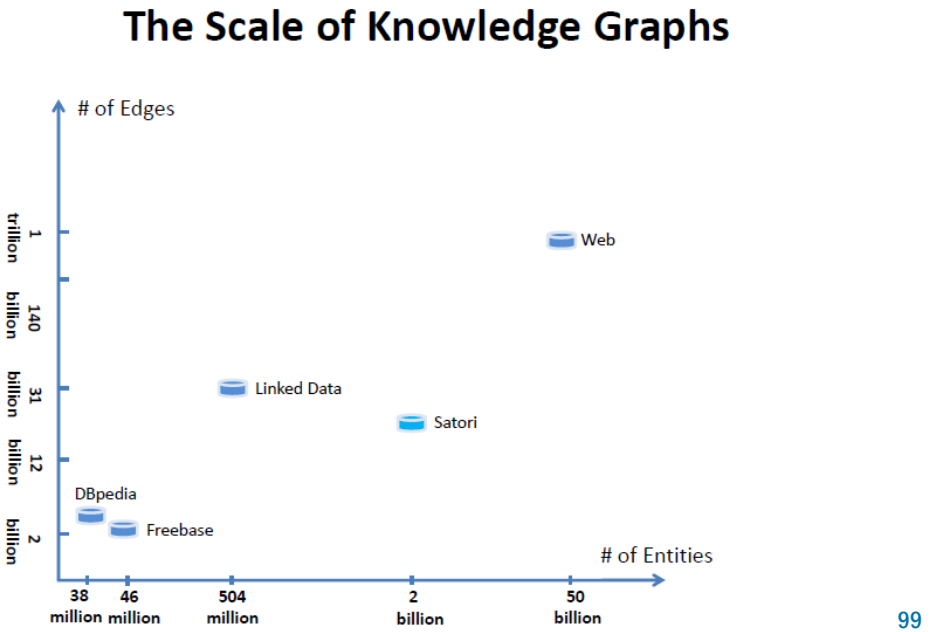
GOOGLE KNOWLEDGE GRAPH (2)

- 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)



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

ScaDS DRESDEN LEIPZIG

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

	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

Longest rivers

Bilbao

Country	Spain
Autonomous community	Basque Country
Province	Biscay
Comarca	Greater Bilbao
Founded	15 June 1300
Government	
• **Type**	Mayor-Council
• **Mayor**	Juan María Aburto (PNV)
Area	
• **Municipality**	41.50 km² (16.02 sq mi)
• **Urban**	18.22 km² (7.03 sq mi)
• **Rural**	23.30 km² (9.00 sq mi)
Elevation	19 m (62 ft)
Highest elevation	689 m (2,260 ft)
Lowest elevation	0 m (0 ft)
Population (2014)[1]	
• **Municipality**	346,574
• **Density**	8,400/km² (22,000/sq mi)
• **Metro**	950,155

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

ScaDS DRESDEN LEIPZIG

WEB TABLES (2)

- Need to add semantics
  - table contents described in surrounding text
  - identify key column vs. property column for vertical tables
  - attributes need to be annotated, e.g., with knowledge graph
- Integration tasks
  - cluster tables by domains
  - link or cluster equivalent attributes
- *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

ScaDS DRESDEN LEIPZIG

AGENDA PART I (BIG DATA INTEGRATION)

- Introduction
- Scalable entity resolution / link discovery
- Large-scale schema/ontology matching
- Holistic data integration
- Summary

ScaDS DRESDEN LEIPZIG

SUMMARY

- ScaDS Dresden/Leipzig
  - research focus on data integration, knowledge extraction, visual analytics
  - broad application areas (scientific + business-related)
- Big Data Integration
  - Big Data poses new requirements for data integration (variety, volume, velocity, veracity)
  - comprehensive data preprocessing and cleaning
  - Hadoop-based approaches for improved scalability, e.g. Dedoop
  - usability: machine-learning approaches, GUI, ...

ScaDS DRESDEN LEIPZIG

SUMMARY (2)

- Large-scale schema matching
  - combined use of linguistic, structural an instance-based techniques
  - performance techniques for fast match execution
  - utilization of background knowledge and reuse of previous matches are key to high match quality
- Holistic data integration
  - combined integration of many sources (metadata + instances)
  - clustering-based rather than mapping-based approaches
  - construction of and linking to large knowledge graphs
  - many research opportunities

ScaDS DRESDEN LEIPZIG

SOME OPEN CHALLENGES

- Parallel execution of more diverse data integration workflows for text data, image data, sensor data, etc.
  - learning-based configuration to minimize manual effort (active learning, crowd-sourcing)
- Improved reuse of large-scale schema matching
- Semi-automatic merging of large schemas / ontologies
- Holistic integration of many data sources (data + metadata)
  - clustering-based entity resolution for many sources
  - n-way merging of related ontologies (e.g. product taxonomies, domain-specific knowledge graphs)
  - improved utilization of large data collections, e.g. web tables

ScaDS DRESDEN LEIPZIG

REFERENCES

- A. Algergawy, S. Massmann, E. Rahm: *A Clustering-based Approach For Large-scale Ontology Matching*. Proc. ADBIS, 2011
- P. Arnold, E. Rahm: *Enriching Ontology Mappings with Semantic Relations*. Data and Knowledge Engineering, 2014
- 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.A. Bernstein, J. Madhavan, E. Rahm: *Generic Schema Matching, Ten Years Later*. PVLDB, 2011 (VLDB 10 Year Best Paper Award Paper)
- L. Chiticariu, P. G. Kolaitis, L. Popa: *Interactive generation of integrated schemas*. Proc. SIGMOD 2008
- P. Christen: *Data Matching*. Springer, 2012
- A. Das Sarma, X. Dong, A. Halevy: *Bootstrapping pay-as-you-go data integration systems*. Proc. SIGMOD 2008
- 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*. Prpoc. SSDM 2015
- H. Elmeleegy, J. Madhavan, A.Y. Halevy: *Harvesting Relational Tables from Lists on the Web*. PVLDB 2009
- A. Gross, M. Hartung, T. Kirsten, E. Rahm: *Mapping Composition for Matching Large Life Science Ontologies.* Proc. Int. Conf. on Bio-Ontologies 2011
- R. Gupta, A. Halevy, X.Wang, S. Whang, F. Wu: *Biperpedia: An Ontology for Search Applications*. PVLDB 2014

ScaDS DRESDEN LEIPZIG

REFERENCES (2)

- T. Kirsten, A. Gross, M. Hartung, E. Rahm: *GOMMA: a component-based infrastructure for managing and analyzing life science ontologies and their evolution.* J. Biomedical Semantics, 2(6), 2011
- 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
- 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)

ScaDS DRESDEN LEIPZIG

REFERENCES (3)

- S. Melnik, H. Garcia-Molina, E. Rahm: *Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching*. Proc. 18th Int. Conf. on Data Engineering (ICDE), San Jose, 2002
- 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
- G. Papadakis, Ge. Koutrika, T. Palpanas, W. Nejdl: *Meta-blocking: taking entity resolution to the next level*. TKDE 2013
- E. Peukert, J. Eberius, E. Rahm: *A Self-Configuring Schema Matching System*. Proc. ICDE, 2012
- R. Pottinger: *Mapping-Based Merging of Schemas*. In: Schema Matching and Mapping, Springer 2011
- E. Rahm, W.E. Nagel: *ScaDS Dresden/Leipzig: Ein serviceorientiertes Kompetenzzentrum für Big Data*. Proc. GI-Jahrestagung 2014: 717
- E. Rahm, P.A. Bernstein: *A Survey of Approaches to Automatic Schema Matching*. VLDB Journal 10 (4) 2001
- 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
- S. Raunich, E. Rahm: *Target-driven Merging of Taxonomies with ATOM*. Information Systems, 2014