Multi-Source Data Matching and Clustering

Erhard Rahm

German AI Centers

5 new, permanent German AI centers (in addition to DFKI) :

- Berlin (BIFOLD)
- Dortmund / Bonn (ML2R)
- Dresden / Leipzig (ScaDS.AI)
- München (MCML)
- Tübingen (tuebingen.ai)

www.humboldt-foundation.de

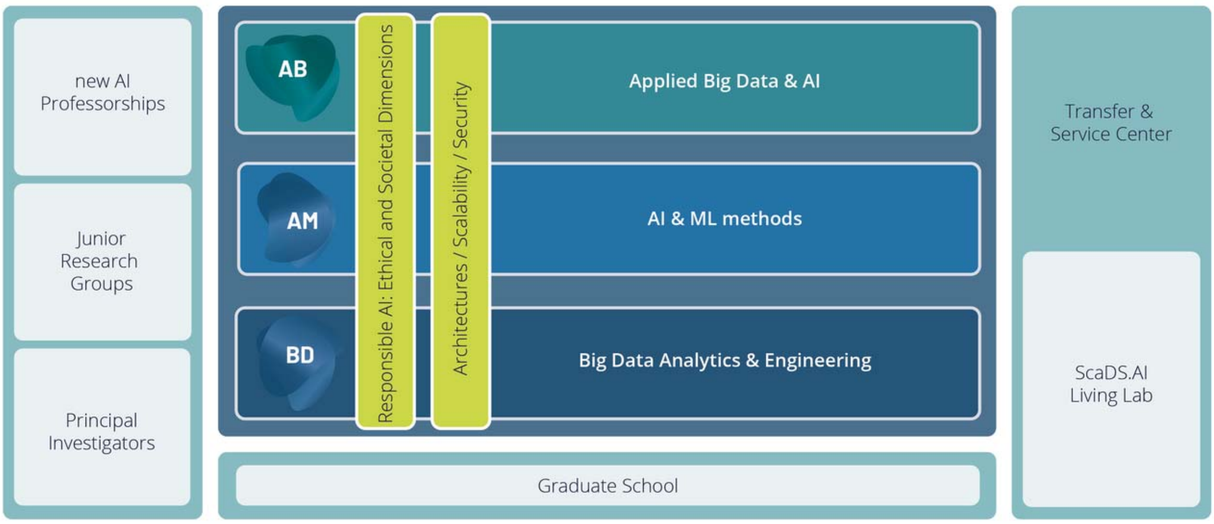
ScaDS.AI

- **SCADS.AI:** Center for **Sca**lable **D**ata Analytic**S** and **A**rtificial **I**ntelligence
- extends previous Big Data center ScaDS Dresden/Leipzig (est. 2014)
- since 2019: AI / Data Science center ScaDS.AI
- July 2022: institutional funding starts
  - co-financed by BMBF and state of Saxony

ScaDS.AI DRESDEN LEIPZIG

UNIVERSITÄT LEIPZIG

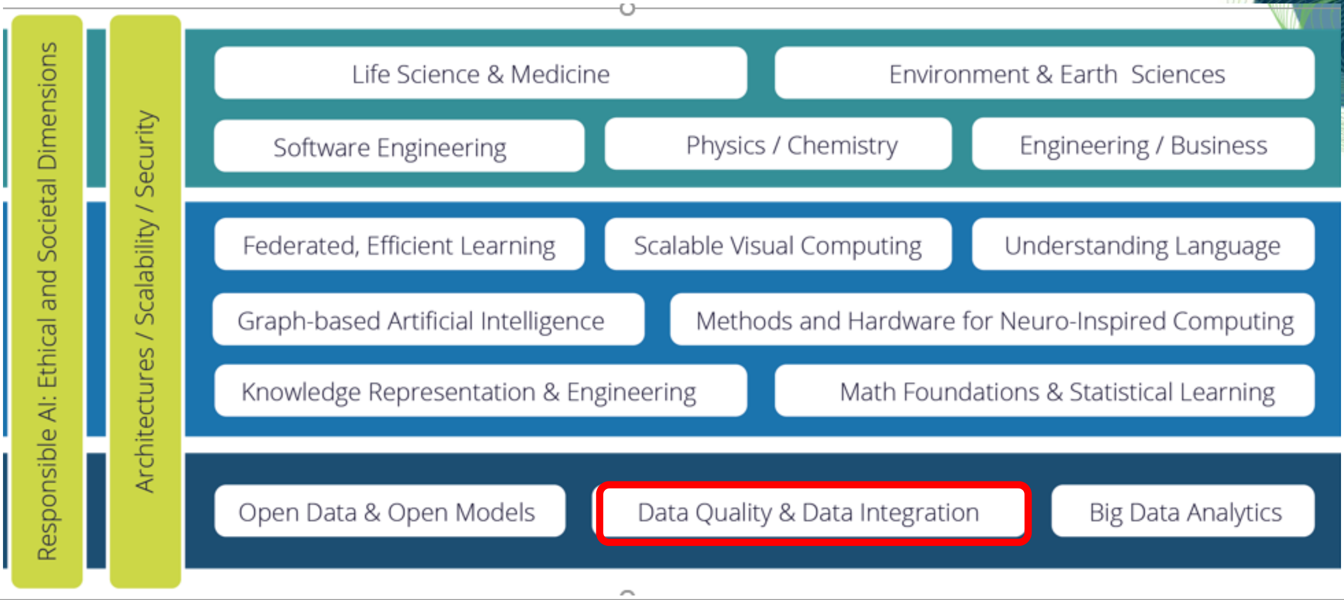
ScaDS.AI: Overall structure



ScaDS.AI DRESDEN LEIPZIG

UNIVERSITÄT LEIPZIG

Research Areas



Data Integration

Provision of uniform access to data originating from multiple, autonomous sources

Physical data integration

- original data is combined within a new dataset / database for access and analysis
- approach of data warehouses, knowledge graphs and most Big Data applications

Virtual data integration

- data is accessed on demand in their original data sources, e.g. based on an additional query layer
- approach of federated databases and linked data

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

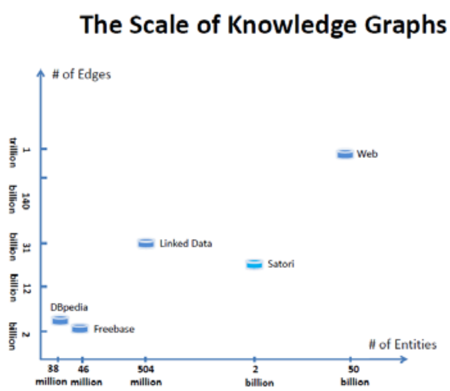
Instance (entity, data) level

- transform heterogeneous source data into uniform representation
- identify and resolve data quality problems
- identify and resolve equivalent instance records: *link discovery / data matching / entity resolution* …
- fusion of matching objects

Knowledge Graphs

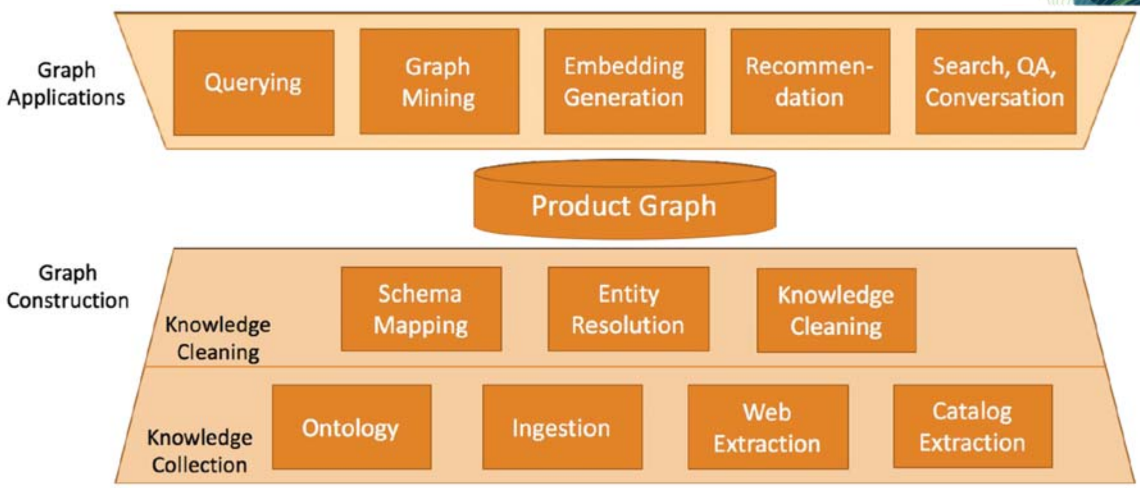
uniform representation and semantic categorization of entities of different types

- examples: DBPedia, Yago, Wikidata, Google KG, MS Satori, Facebook, …
- entities often extracted from other resources (Wikipedia, Wordnet etc.) or web pages, documents, web searches etc.
- Knowledge Graphs provide valuable background knowledge for enhancing entities (based on prior *entity linking),* improving search results …



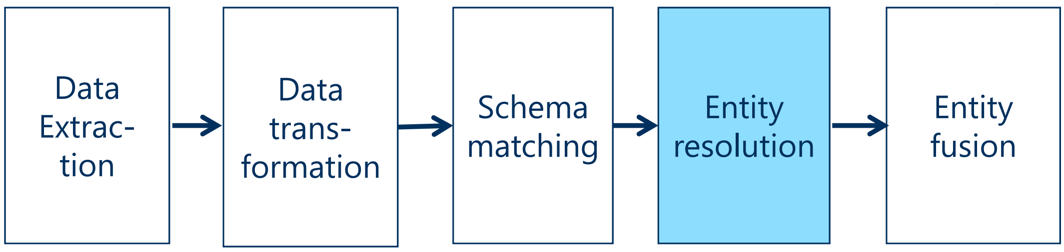
Shao, Li, Ma (Microsoft Asia): Distributed Real-Time Knowledge Graph Serving (slides, 2015)

Example: Product Knowledge Graph



from: Dong. KDD2018

Main steps in data integration



DATA INTEGRATION CHALLENGES 1

- Data quality
  - unstructured, semi-structured sources
  - need for data cleaning and enrichment
- Large-scale data integration
  - large data/metadata volume or/and many sources
  - improve runtime by reducing search space (e.g. with blocking) and parallel processing (Hadoop clusters, GPUs, etc.)
  - many sources require *holistic data integration*: clustering of schema elements and entities, not only binary matching
- High match quality
  - needs effective combination of several similarities
  - use of supervised ML approaches
  - representation learning (embeddings) can provide improved data input

DATA INTEGRATION CHALLENGES 2

- Support for evolution and change
  - addition of new sources and new entities without having to integrate everything again
  - incremental / dynamic vs batch / static data integration
- Graph-based data integration, e.g. to create knowledge graphs
  - integrate entities of multiple types and their relationships
  - requires holistic and incremental data integration
- Privacy for sensitive data
  - privacy-preserving record linkage and data mining

Holistic Data Integration*

scalable approaches for integrating N data sources (N >>2)

increasing need due to numerous sources, e.g., from the web
- many thousands of web shops
- data lakes with thousands to millions of tables

pairwise matching/linking does not scale
- 200 sources -> 20.000 mappings

clustering-based approaches
- represent matching entities from k sources in single cluster
- determine *cluster representative* for further processing/matching
- new entities are only compared with clusters rather than entities of all sources

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

AGENDA

- Introduction to Data Integration
- Entity resolution and clustering
  - introduction / ER workflow / tools
  - FAMER
  - entity clustering for clean and mixed sources (CLIP, MSCD-HAP)
- Incremental entity clustering / repair
- Summary and outlook

DATA MATCHING / ENTITY RESOLUTION

- Identification of semantically equivalent objects
  - within one data source or between different sources

DUPLICATE PUBLICATION ENTRIES

Data cleaning: Problems and current approaches
E Rahm, HH Do - IEEE Data Eng. Bull., 2000
Cited by 2456 Related articles All 24 versions

Data Cleaning: Problems & Current Approaches*
D Hang-Hai, R Erhard - IEEE bulletin of the technical committee on Data ..., 2000
Cited by 8 Related articles

Problems and Current Approaches*
E Rahm, DC Do HH - IEEE Bulletin on Data Engineering.-2000.-23 (4), 2015
Cited by 6 Related articles

Data cleaning: Problems and current approaches' IEEE Data Eng. Bull., 2000*
E Rahm, HH Do - 2000
Cited by 5 Related articles

Hong Hai Do*
E Rahm - IEEE Bulletin of the Technical Committee on Data ..., 2000
Cited by 4 Related articles

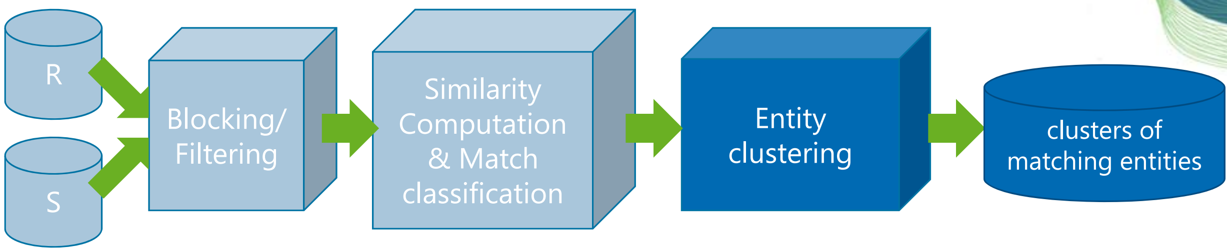
Do. H. 2000. Data cleaning: Problems and current approaches*
E Rahm - IEEE Data Engineering Bulletin
Cited by 4 Related articles

Do. H. 2000*
E Rahm, H Do - Data cleaning: Problems and current approaches, 2011
Cited by 3 Related articles

Data engineering—Special issue on data cleaning*
E Rahm, HH Do - Data Engineering, 2000
Cited by 3 Related articles

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

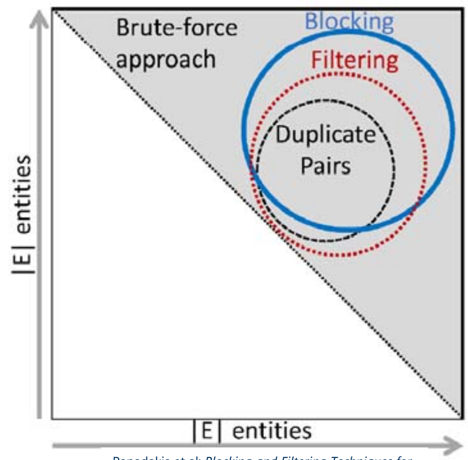
ENTITY RESOLUTION WORKFLOW



- mostly only 1 or 2 sources
- n>=2: duplicate-free (clean) sources or not
  - clean sources: at most one entity per cluster (cluster sizes <= n)

BLOCKING & FILTERING

- naïve: pairwise matching of all entities
  - quadratic complexity, not scalable
  - strong need to reduce match search space
- **Blocking**
  - group similar objects within blocks / partitions
  - only compare entities of the same block
  - many variations: Standard Blocking, LSH, Sorted Neighborhood, …
- **Filtering**
  - typically applied for *similarity joins* with fixed threshold *t*: **sim (e1, e2) ≥ *t***
  - utilizes characteristics of similarity function, e.g., for string similarity
  - can utilize triangle inequality for metric similarity/distance functions



Papadakis et al: *Blocking and Filtering Techniques for Entity Resolution: A Survey.* ACM CSUR 2020

BLOCKING TECHNIQUES

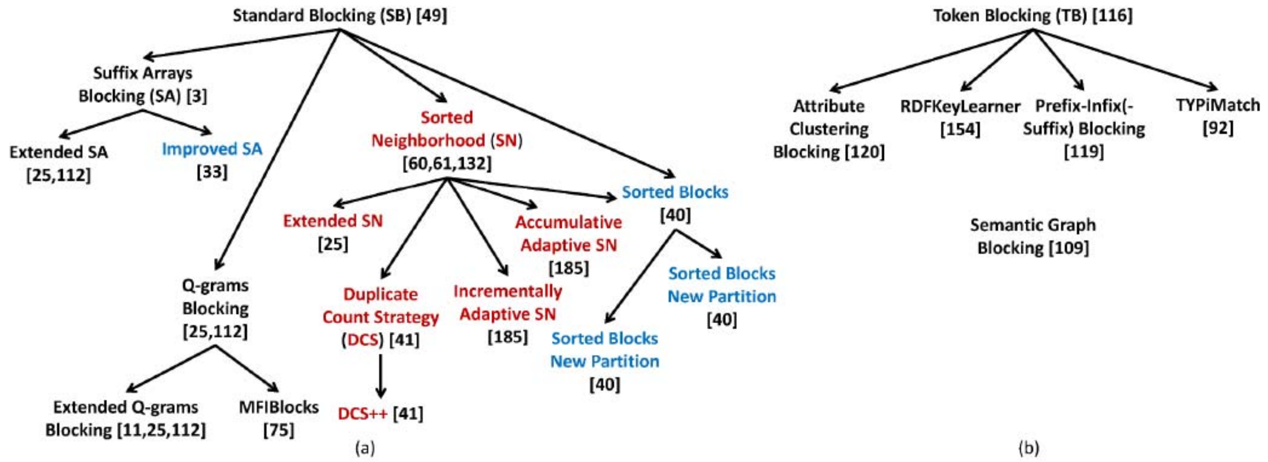


Fig. 3. The genealogy trees of nonlearning (a) schema-aware and (b) schema-agnostic Block Building techniques. Hybrid, hash-, and sort-based methods are marked in blue, black, and red, respectively.

Papadakis et al: *Blocking and Filtering Techniques for Entity Resolution: A Survey*. ACM CSUR 2020

MATCHING

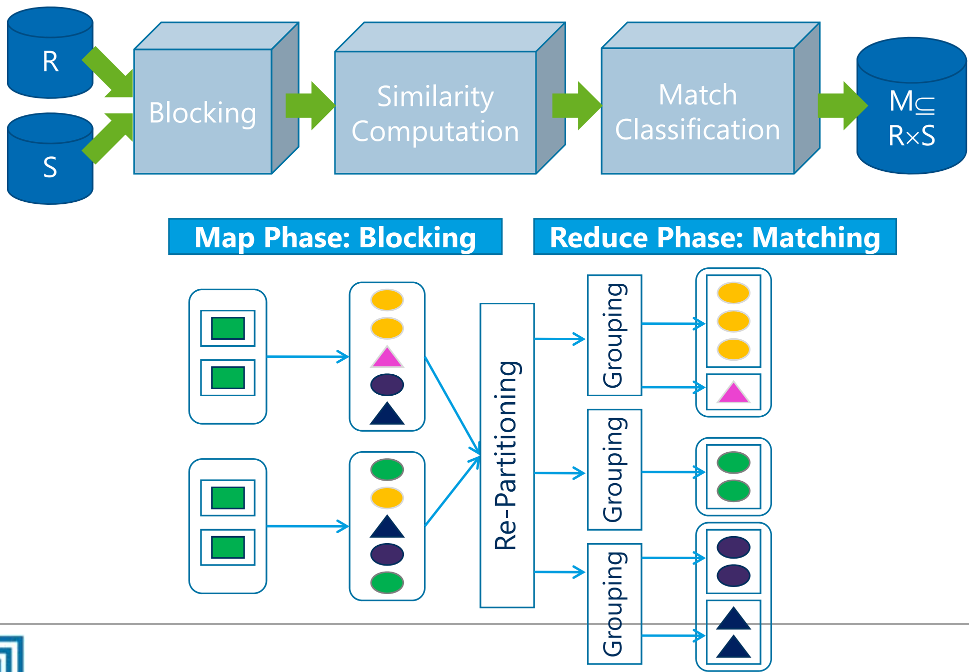
- combined use of several similarity values
  - attribute similarities, e.g. using numeric or string similarity measures
  - context-based matchers
- general match rules with multiple similarties
  - e.g. pubs match if *title sim.* $\geq$ *0.9 & author sim. > 0.4*
- learned/supervised match classification models
  - need suitable training data

DEDOOP: DEDUPLICATION WITH HADOOP

- Parallel execution of match workflows with Hadoop
- library of match and blocking techniques
- learning-based match configuration
- GUI-based workflow specification
- automatic generation and execution of Map/Reduce jobs on different clusters
- Automatic load balancing for optimal scalability

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

PARALLEL MATCHING WITH MAP/REDUCE



RECENT ER TOOLS

- Magellan
  - PyMatcher component provides several blocking and similarity algorithms to customize match approach
  - support for machine learning, including deep learning
- JedAI
  - supports matching for structured and unstructured data
  - plethora of methods for blocking, matching and clustering
  - provides GUI

JedAI

RECENT ER TOOLS 2

- FAMER
  - FAst Multi-source Entity Resolution system
  - built on Apache Flink
  - Blocking, linking and clustering module for multiple sources
  - many clustering approaches included for clean and dirty sources
  - support for incremental matching and clustering

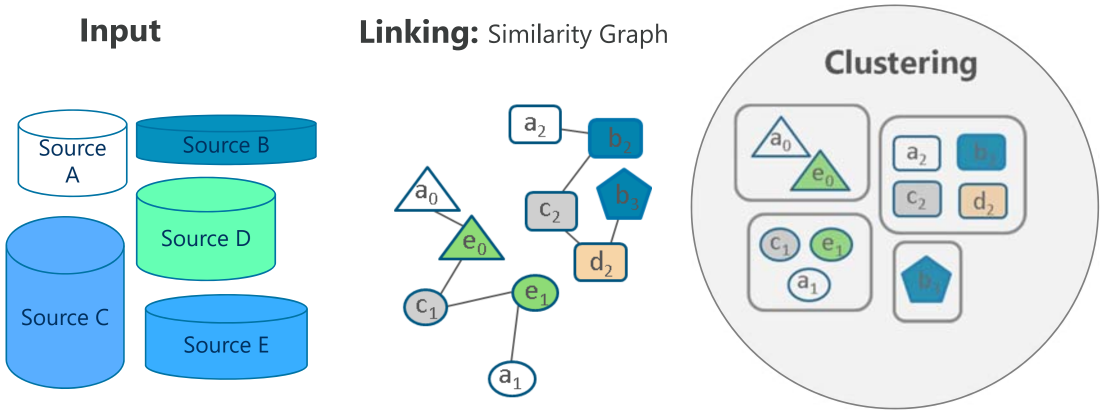
FAMER

TOOL COMPARISON

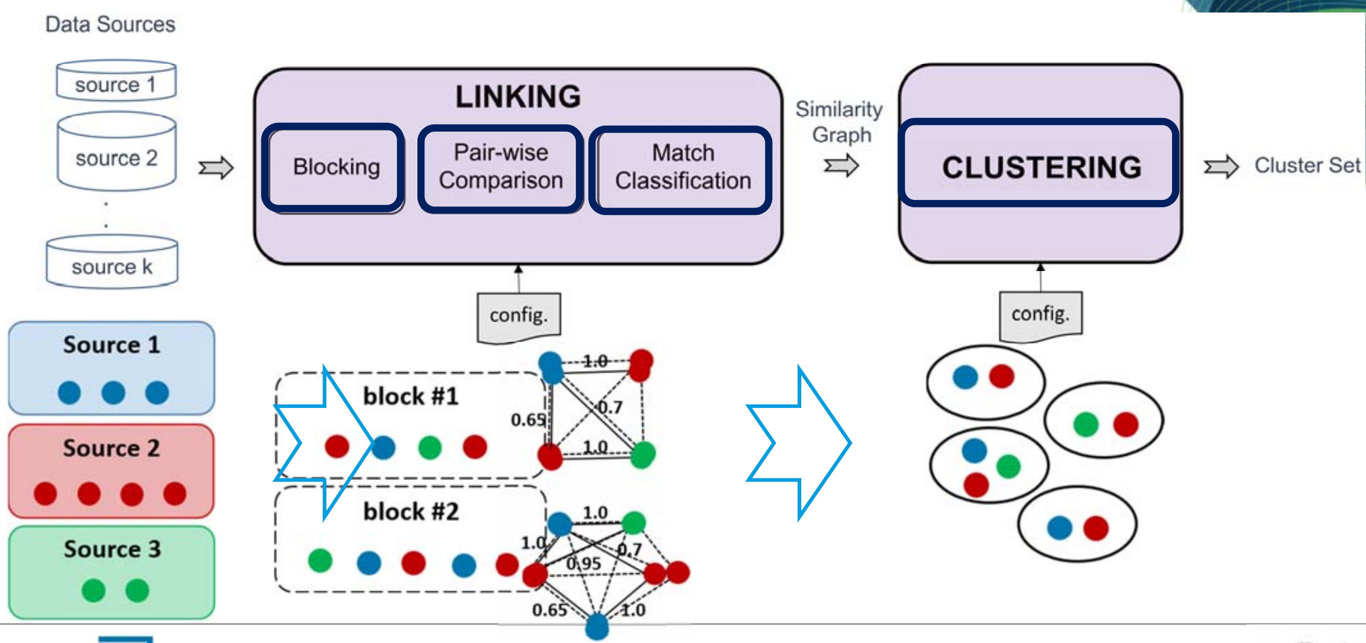
	Magellan	JedAI	FAMER
Blocking			
Matching			
Clustering			
Incremental ER			
GUI			
Big Data Architecture	only in commercial CloudMatcher		Apache Flink

FAMER TOOL

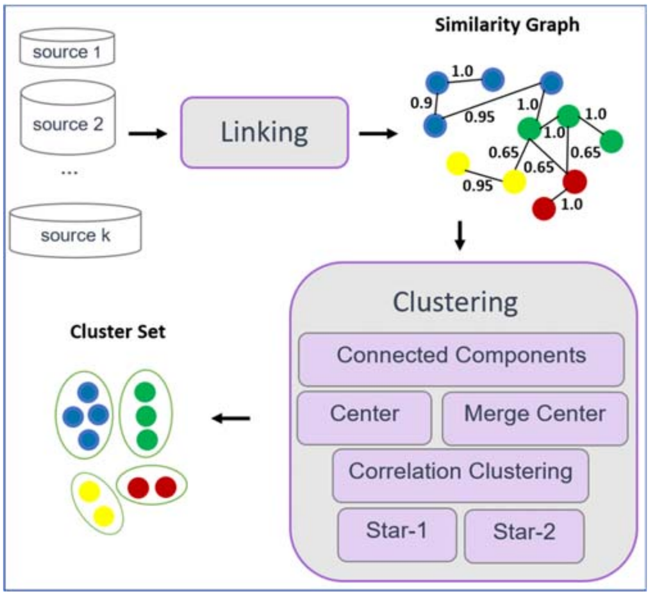
- **FA**st **M**ulti-source **E**ntity **R**esolution System
  - scalable linking & clustering for many sources



FAMER BATCH PIPELINE



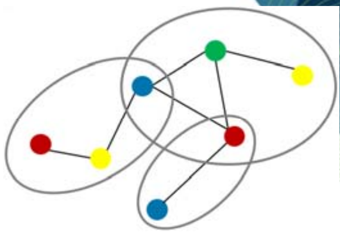
EXISTING CLUSTERING ALGORITHMS*



* Hassanzadeh et al.: *Clustering for Duplicate Detection. VLDB 2009*

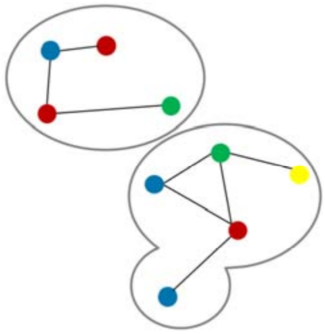
PROBLEMS

overlapping clusters

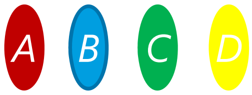


source-inconsistent clusters for clean (duplicate-free) sources

each cluster should not have more than one entity per source

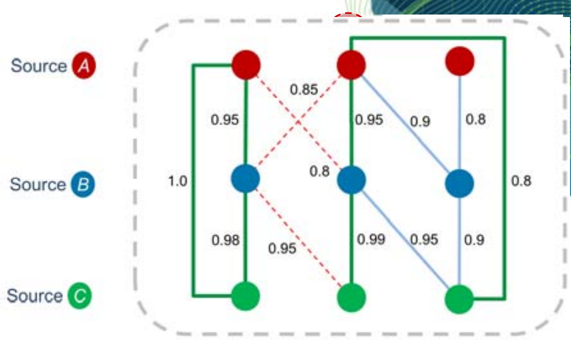


sources



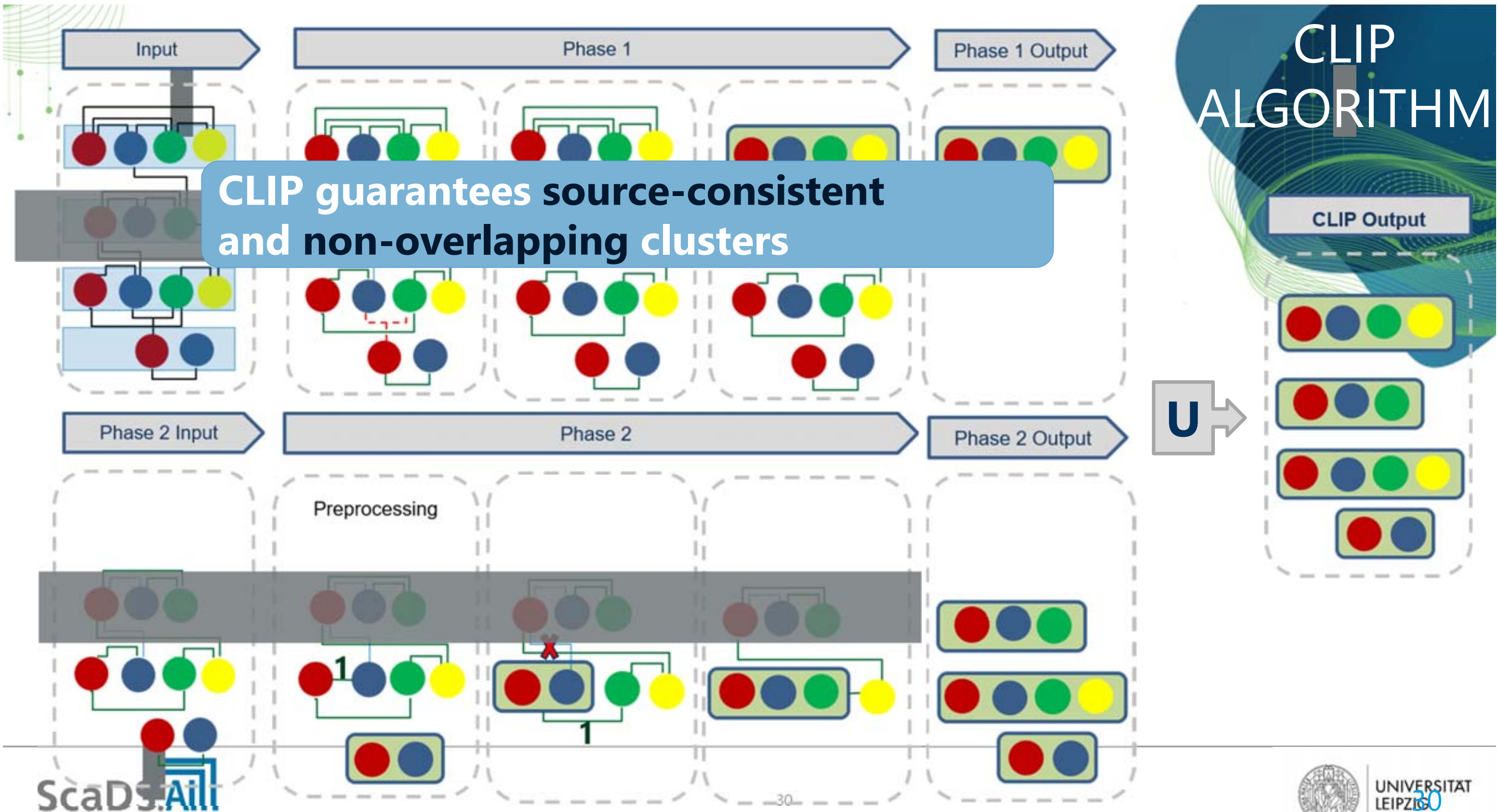
CLIP APPROACH (ESWC BEST RESEARCH PAPER)

- CLIP (CLustering based on Link Priority) uses **link strength**
  - strong: maximum link from **both** ends
  - normal: maximum link from **one** end
  - weak: maximum link from **no** end

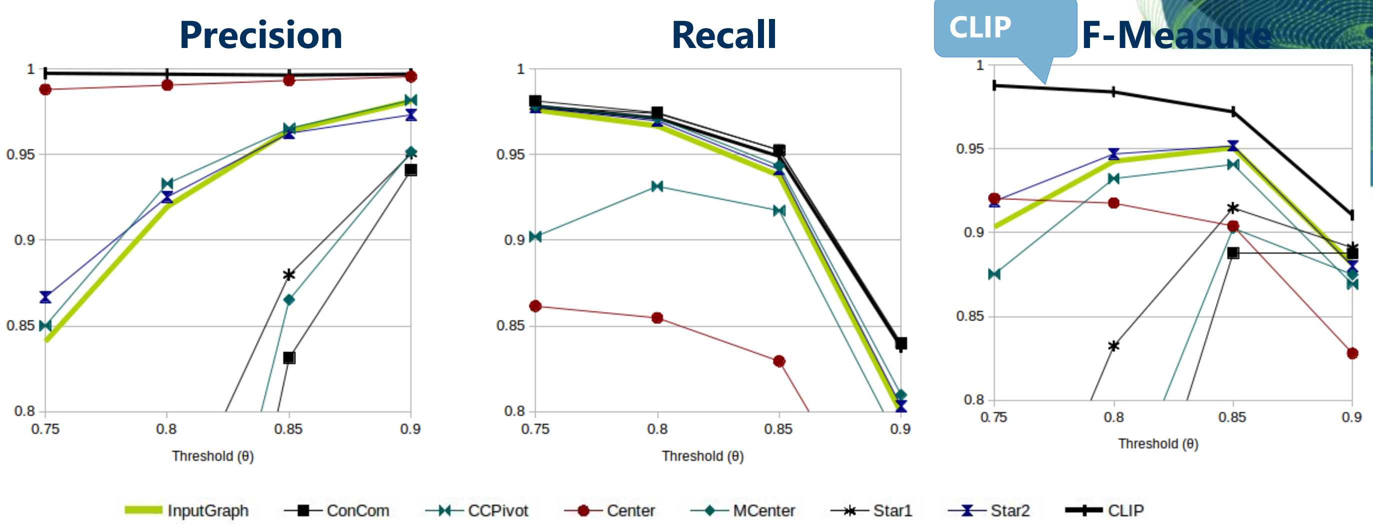


- CLIP
  - ignores weak links
  - focusses on strong links
  - also considers normal links

CLIP ALGORITHM



EVALUATION: GEO. DATASET



RUNTIME AND SPEED-UP

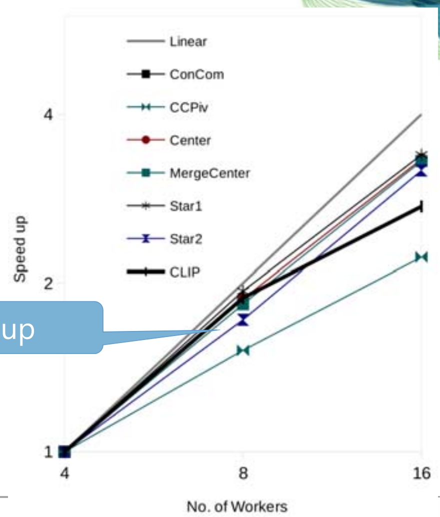
- Experiments based on Hadoop and Apache Flink (16 machines)

North Carolina Voters (10 mill.) runtimes on 16 workers - th = 0.8	
Connected Components	79 sec.
Star1/2	197/173 sec.
CLIP	228 sec.
Center	423 sec.
Merge Center	695 sec.
CCPiv	1303 sec.

Increasing

Near linear speed up

North Carolina Voters (5 Mi)



MULTI-SOURCE CLEAN/DIRTY CLUSTERING

- previous assumption: data sources are duplicate-free
- more realistic assumption: some sources are dirty
  - solution: first deduplicate dirty sources
  - problem: requires immense effort and perhaps not completely successful [7]
- solution: **MSCD approaches**
  - approaches that can deal with dirty sources
  - only a fraction (possibly 0%) of sources have to be clean
  - goal: achieve better match quality than general clustering scheme while avoiding limitation of requiring duplicate-free sources
  - two approaches added to FAMER based on hierarchical agglomerative clustering (HAC) and affinity propagation (AP)

MSCD-HAC

- modify **H**ierarchical **A**gglomerative **C**lustering ->MSCD-HAC
- iterative approach
  - initially each entity forms a cluster
  - continuously determine most similar pair of clusters (c_i, c_j) as long as minimal merge sim. threshold is exceeded. Merge clusters c_i, c_j only when
    - they are *Reciprocal Nearest Neighbours* (RNN), i.e. $NN(c_j) = c_i$ and $NN(c_i) = c_j$
    - merge results in *source-consistent clusters*, i.e., at most one entity of a clean source in a cluster
- 3 approaches to determine cluster similarity sim (c_i, c_j)
  - *Single linkage (S-LINK):* sim c_i, c_j = max {sim(e_m, e_n)}
  - *Complete linkage (C-LINK)* : sim c_i, c_j = min {sim(e_m, e_n)}
  - *Average linkage (A-LINK)* : sim c_i, c_j = avg {sim(e_m, e_n)}

EVALUATION SUMMARY

- **camera dataset*** (23 sources, ~21 K entities)
  - combination of clean and dirty sources
- all approaches are experimented on all MSC and MSCD datasets
- MSCD clustering schemes MSCD-HAC and MSCD-AP are compared with
  - generic clustering schemes
  - CLIP

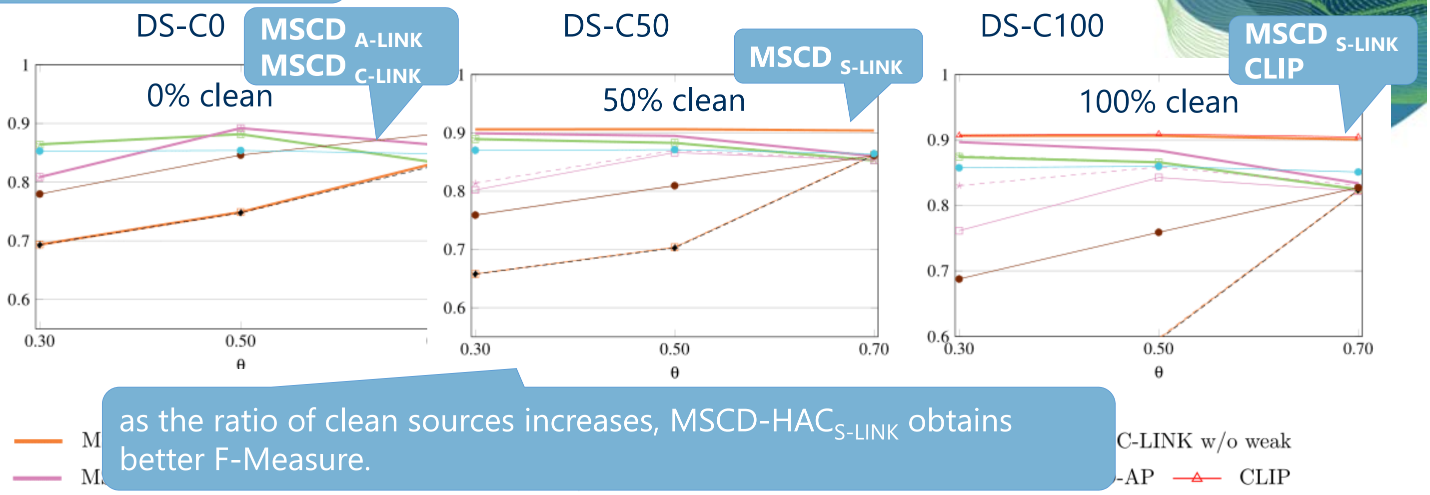
MSCD dataset	%entities from clean sources
DS-C0	0%
DS-C26	26%
DS-C32	32%
DS-C50	50%
DS-C62	62%
DS-C80	80%
DS-C100	100%

* ACM Sigmod programming contest 2020

F-MEASURE: CAMERA DATASET

high recall of $MSCD_{S\text{-}LINK}$
high precision of $MSCD_{S\text{-}LINK}$

match threshold = merge threshold (θ)



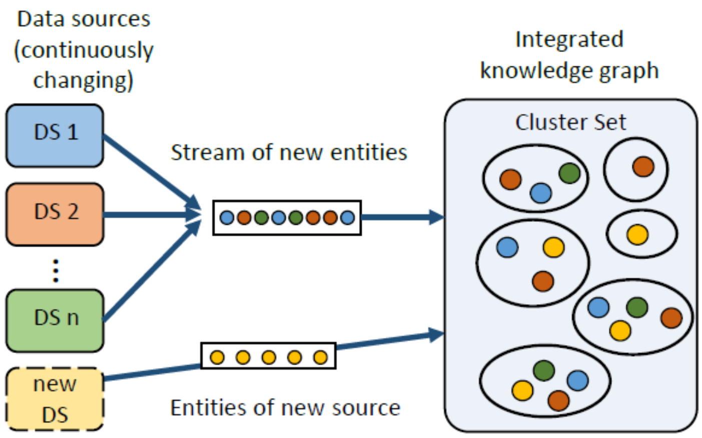
as the ratio of clean sources increases, $MSCD\text{-}HAC_{S\text{-}LINK}$ obtains better F-Measure.

AGENDA

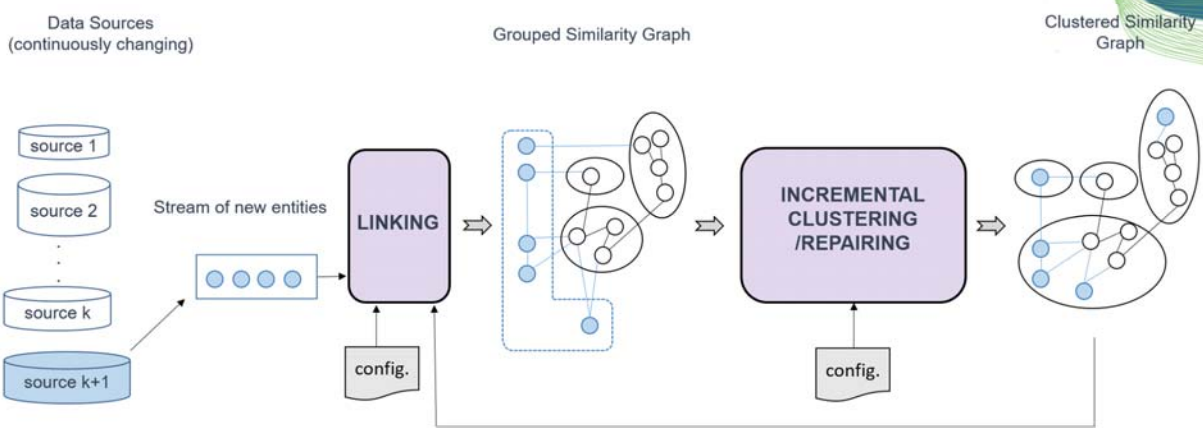
- Introduction to Data Integration
- Entity resolution and clustering
  - introduction / ER workflow / tools
  - FAMER
  - entity clustering for clean and mixed sources (CLIP, MSCD-HAP)
- Incremental entity clustering / repair
- Summary and outlook

MOTIVATION

- static one-time matching and clustering insufficient
- need for incremental approaches
  - data sources change over time
  - new relevant data sources are added continuously
- expensive re-computation of similarity graph /clusters to be avoided
- order in which new entities are added should have minimal impact
  - need to repair wrong clusters



FAMER INCREMENTAL PIPELINE



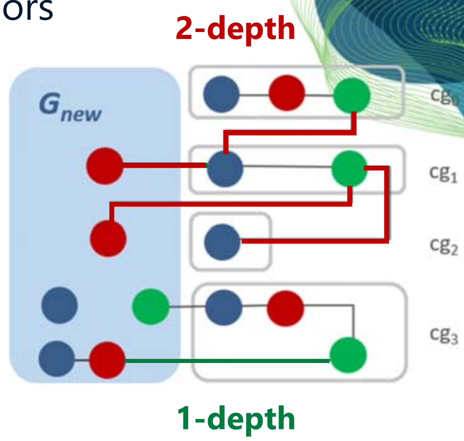
MAX-BOTH MERGE (MBM)

pre-cluster new entities
If a cluster pair (c_{new}, c_{old}) is linked via a max-both link
    if source-consistent (c_{new}, c_{old})
        Merge (c_{new}, c_{old})

- MBM inserts new entity either into existing cluster or forms a new cluster out of it
  - merging only for *max-both (strong) links* and when source-consistency constraint is met (at most one entity per clean source)

N-DEPTH RECLUSTERING

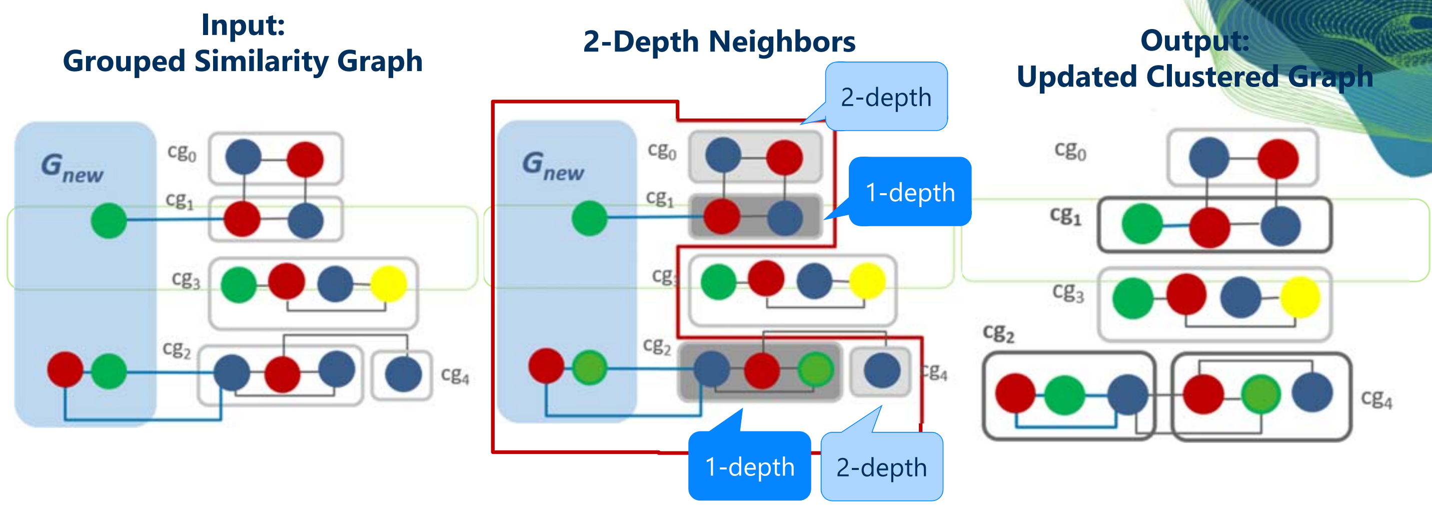
- reclusters new entities in G_{new} with their neighbors
  - can repair old cluster decisions
  - limits the amount of reclustering for the sake of efficiency
  - independent from order of source/entity additions



n-depth neighbors:
1-depth neighbors of the n-1-depth neighbors

1-depth neighbors:
directly linked groups

2-DEPTH RECLUSTERING: EXAMPLE



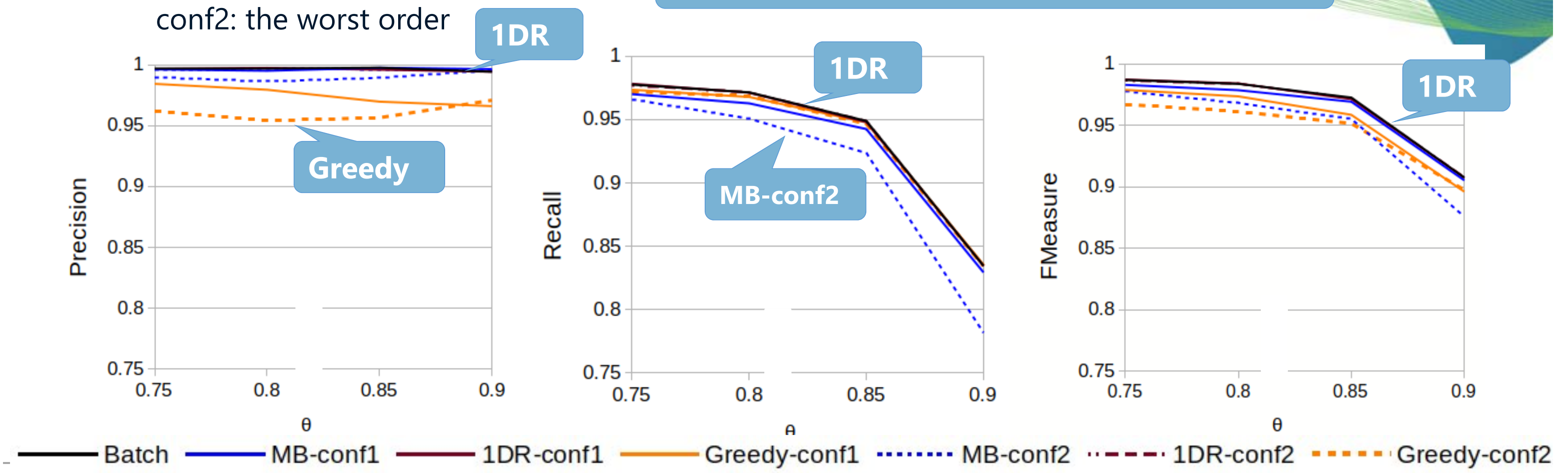
EVALUATION

- Geo. dataset

conf1: the best order
conf2: the worst order

Comparison with base approach: Greedy
[Incremental Record Linkage (Gruenheid et al., *VLDB 2014*)]

nDR approach is robust against source order



EVALUATION: RUNTIME

with less resources Batch runtime is significantly higher

- North Carolina Voters, 10 Mill. entities

incremental approaches are faster than Batch

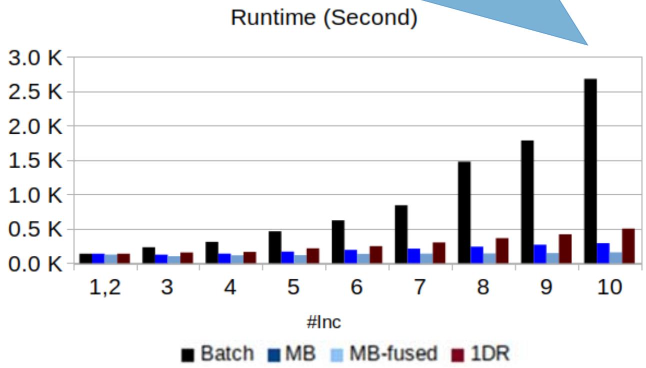
accumulated runtimes (s) for source-wise ER

#worker	Batch	MB	1DR
4	117,852	5,648	21,179
8	33,791	2,178	4,283
16	8,542	1,778	2,513

threshold (θ): 0.7

MB is faster than nDR

for 10th increment, batch runtime is more than **five** times higher than 1DR



INCREMENTAL METHODS CONTRIBUTIONS

- incremental approaches are much faster and similarly effective than batch ER
- reclustering approach nDR achieves same quality than batch ER while being much faster
- quality of nDR does not depend on the order in which new entities are added

SUMMARY

- Data integration still faces many challenges
  automation, data quality, efficiency/scalability, privacy support, continious change …
- need for multi-source entity resolution with clustering
- FAMER integrates new and effective approaches for
  - consideration of duplicate-free (clean) data sources
  - support for incremental matching/clustering and cluster repair

OPEN RESEARCH PROBLEMS

- Largely automatic creation/refinement of large-scale knowledge graphs
- requires tackling of several tasks / challenges
  - development and evolution of KG ontology
  - initial population of KG
  - data acquisition / extraction / cleaning for new data to be integrated
  - learning-based classification of new entities
  - incremental schema/property matching for many entity types
  - **incremental entity resolution/clustering for many entity types**
  - entity fusion …
- Multi-modal data integration

REFERENCES (1)

- D. Ayala, I. Hernández, D. Ruiz, E. Rahm, Erhard: *LEAPME: Learning-based Property Matching with Embeddings*. Data & Knowledge Engineering 2022
- P. Christen: *Data Matching*. Springer 2012
- **L. Dong: *Challenges and Innovations in Building a Product Knowledge Graph*. Tutorial, KDD 2018**
- J. Fisher, P. Christen, Q. Wang, E. Rahm: *A clustering-based framework to control block sizes for entity resolution*. Proc. KDD 2015
- **A. Gruenheid et al.: *Incremental record linkage*. VLDB 2014**
- **O. Hassanzadeh et al.: *Clustering for Duplicate Detection*. VLDB 2009**
- 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*. PVLDB 2010
- H. Köpcke, A. Thor, S. Thomas, E. Rahm: *Tailoring entity resolution for matching product offers*. Proc. EDBT 2012: 545-550
- **L. Kolb, A. Thor, E. Rahm: *Dedoop: Efficient Deduplication with Hadoop*. PVLDB 5(12), 2012**
- L. Kolb, E. Rahm: *Parallel Entity Resolution with Dedoop*. Datenbank-Spektrum 13(1): 23-32 (2013)
- L. Kolb, A. Thor, E. Rahm: *Load Balancing for MapReduce-based Entity Resolution*. ICDE 2012: 618-629
- **S. Lerm, A. Saeedi, E. Rahm: *Extended Affinity Propagation Clustering for Multi-source Entity Resolution*. BTW 2021**
- S. Mudgal et al.: *Deep learning for entity matching: A design space exploration*. SIGMOD 2018.
- M. Nentwig, A. Groß, E. Rahm: *Holistic Entity Clustering for Linked Data*. IEEE ICDMW 2016 2016
- M. Nentwig, A. Groß, Anika; M. Möller, E. Rahm: *Distributed Holistic Clustering on Linked Data*. LNCS 10574, 2017, pp 371-382
- M. Nentwig, M. Hartung, A. Ngonga, E. Rahm: *A Survey of Current Link Discovery Frameworks*. Semantic Web Journal, 2017

REFERENCES (2)

- **G. Papadakis et al.: *The return of jedAI: end-to-end entity resolution for structured and semi-structured data*. PVLDB 2018**
- **G. Papadakis et al: Blocking and Filtering Techniques for Entity Resolution: A Survey. ACM CSUR 2020**
- D. Obraczka, A. Saeedi, A. E. Rahm, E.: *Knowledge Graph Completion with FAMER*. Proc. KDD DI2KG, 2019
- D. Obraczka, J. Schuchart, E. Rahm: *Embedding-Assisted Entity Resolution for Knowledge Graphs*. Proc. ESWC KGCW, 2021
- E. Rahm, H. H. Do: Data Cleaning: Problems and Current Approaches. IEEE Techn. Bulletin on Data Engineering, 2000
- **E. Rahm: The case for holistic data integration. Proc. ADBIS, 2016**
- **A. Saeedi, L. David, E. Rahm, E: *Matching Entities from Multiple Sources with Hierarchical Agglomerative Clustering*. KEOD 2021**
- **A. Saeedi, M. Nentwig, E. Peukert, E. Rahm: *Scalable matching and clustering of entities with FAMER*. CSIM Quarterly 2018**
- **A. Saeedi, E. Peukert, E. Rahm: *Comparative Evaluation of Distributed Clustering Schemes for Multi-source Entity Resolution*. Proc. ADBIS, LNCS 10509, 2017**
- **A. Saeedi, E. Peukert, E. Rahm: *Using Link Features for Entity Clustering in Knowledge Graphs*. ESWC 2018**
- **A. Saeedi, E. Peukert, E. Rahm: *Incremental Multi-source Entity Resolution for Knowledge Graph Completion*. ESWC 2020**
- J. Shao, Q. ; Wang, A. Wijesinghe, E. Rahm: *ERGAN: Generative Adversarial Networks for Entity Resolution*. ICDM 2020
- M. Wilke, E. Rahm: Towards Multi-modal Entity Resolution for Product Matching. GVDB 2021