Data Integration for Knowledge Graphs

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
- **since July 2022: institutionally funded**
  - co-financed by BMBF and state of Saxony

Research Areas

Topic Areas

Applied AI & Big Data

- Life Science & Medicine
- Environment & Earth Sciences
- Software Engineering
- Physics / Chemistry
- Engineering / Business

AI Algorithms & Methods

- Understanding Language
- Methods and Hardware for Neuro-Inspired Computing
- Graph-based Artificial Intelligence
- Knowledge Representation & Engineering
- Scalable Visual Computing
- Federated, Efficient Learning
- Math Foundations & Statistical Learning

Big Data Analytics & Engineering

- Big Data Analytics
- Open Data & Open Models
- Data Quality & Data Integration

Crosscutting Topics

- Responsible AI: Ethical and Societal Dimensions
- Architectures / Scalability / Security

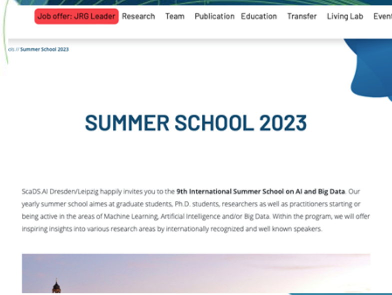
ScaDS.AI DRESDEN LEIPZIG

UNIVERSITÄT LEIPZIG

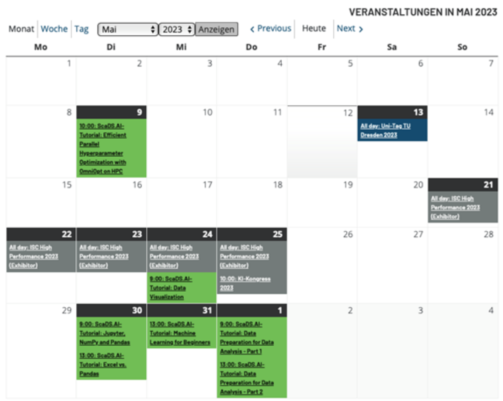
Building up the center

- >150 employees
  - graduate school with about 100 Ph.D. students
  - service & transfer center with living labs in both Leipzig and Dresden
- 8+ **new AI/data science professorships**
- new **junior research groups** (5 so far)
- many additional 3rd-party projects and industry collaborations
- many events

EVENTS / OUTREACH



2019 2020 2021 2022 2023



ScaDS.AI DRESDEN LEIPZIG

UNIVERSITÄT LEIPZIG

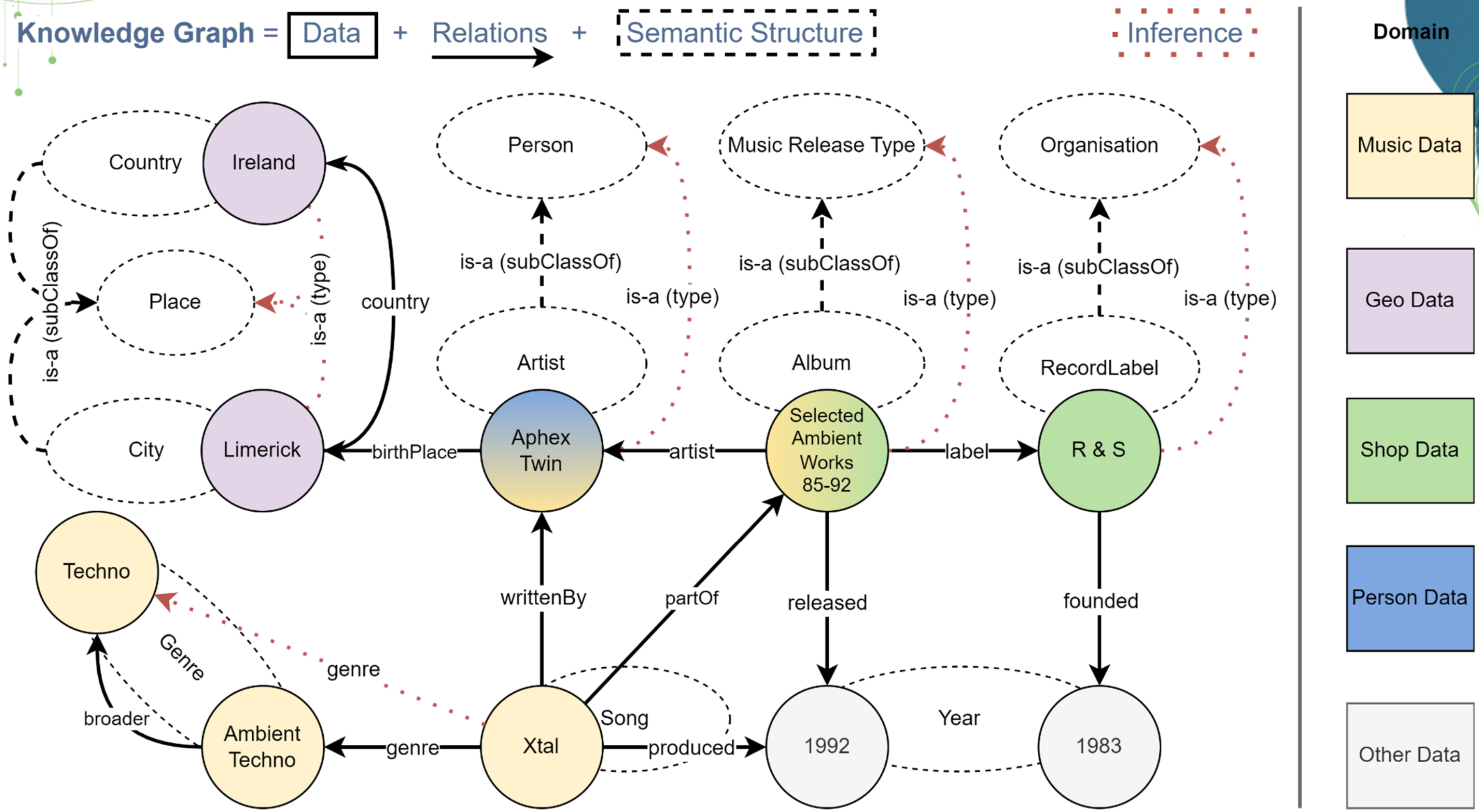
AGENDA

- ScaDS.AI Dresden/Leipzig
- **Construction of Knowledge Graphs**
  - KG intro
  - requirements for KG construction
  - processing steps
  - comparison of existing approaches
  - open challenges
- Entity resolution / matching
  - ER intro
  - Entity clustering and incremental ER (Famer)
  - embedding-based matching of KGs
- Conclusions

Knowledge Graph Key Characteristics

*A **graph of data** consisting of **semantically described entities and relations of different types** that are **integrated from different sources**.*

- a graph (network) of "real world" entities
- high number of entity and relation types
- a formal semantic representation of things (e.g., using a KG ontology)


Knowledge Graph = Data + Relations + Semantic Structure
Inference
Domain
Music Data
Geo Data
Shop Data
Person Data
Other Data
Country
Ireland
Place
City
Limerick
is-a (subClassOf)
is-a (type)
country
Person
Artist
Aphex Twin
Music Release Type
Album
Selected Ambient Works 85-92
Organisation
RecordLabel
R & S
birthPlace
artist
label
writtenBy
partOf
released
founded
Techno
Genre
broader
Ambient Techno
genre
Xtal
Song
produced
1992
Year
1983

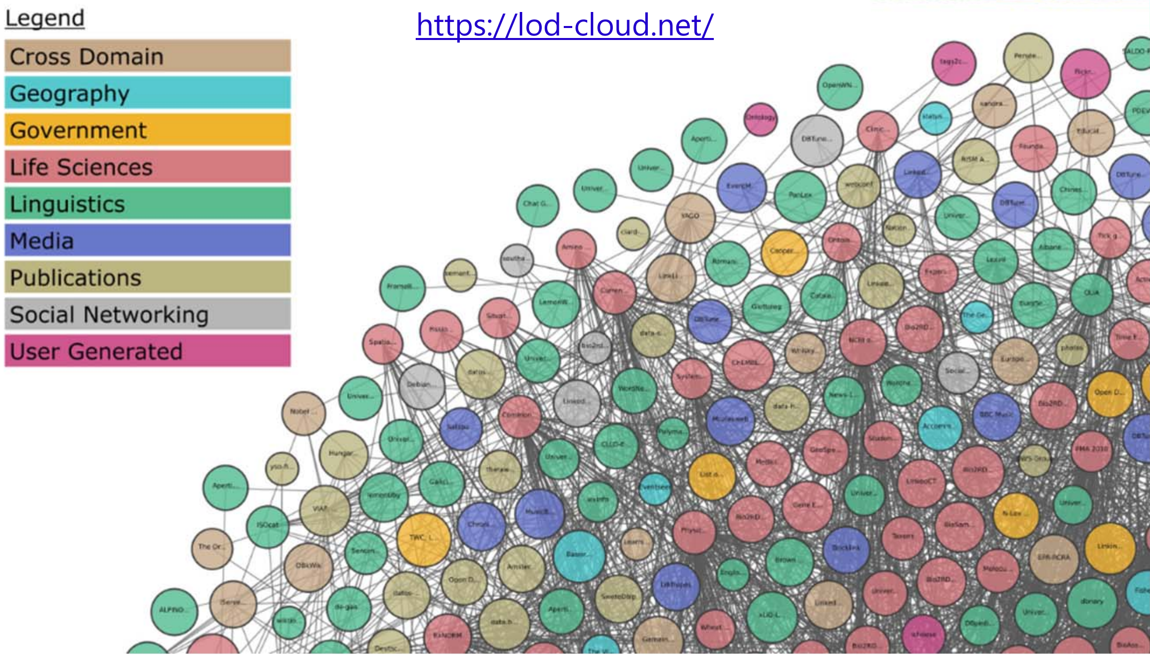
Importance of Knowledge Graphs

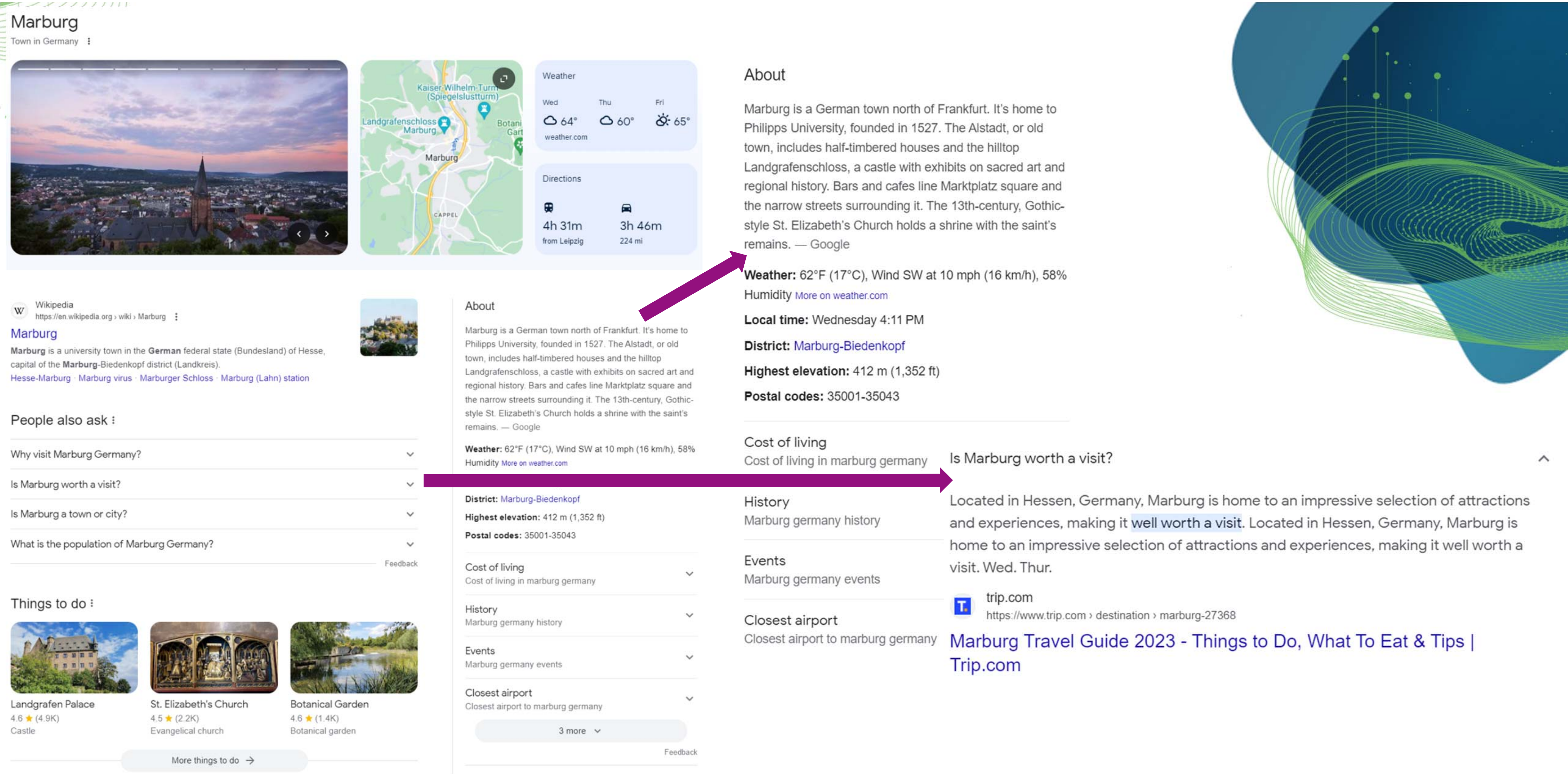
- background knowledge
- semantic search
- QA
- recommender systems

  …

- ML support
  - training data
  - Classification
  - improved explainability …

https://lod-cloud.net/

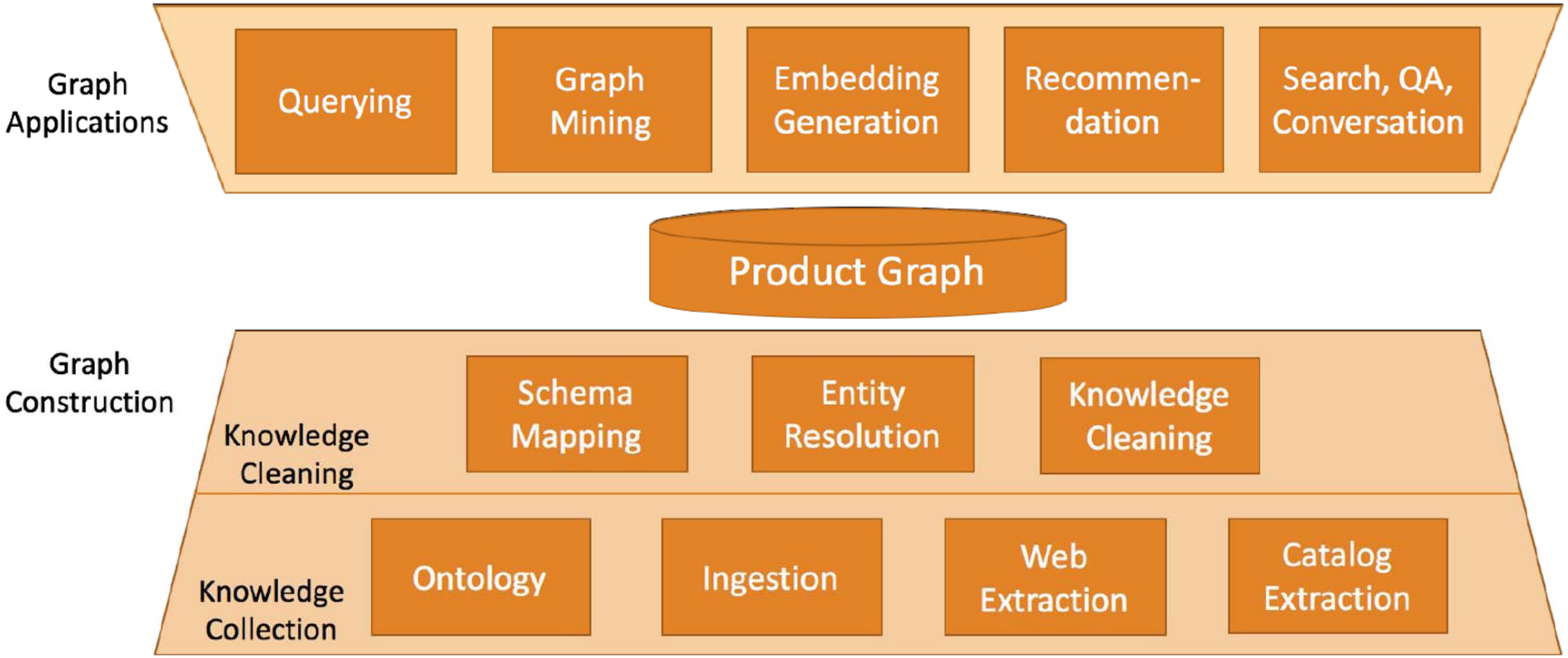




ScaDS.AI DRESDEN LEIPZIG

UNIVERSITÄT LEIPZIG

Example: Product Knowledge Graph



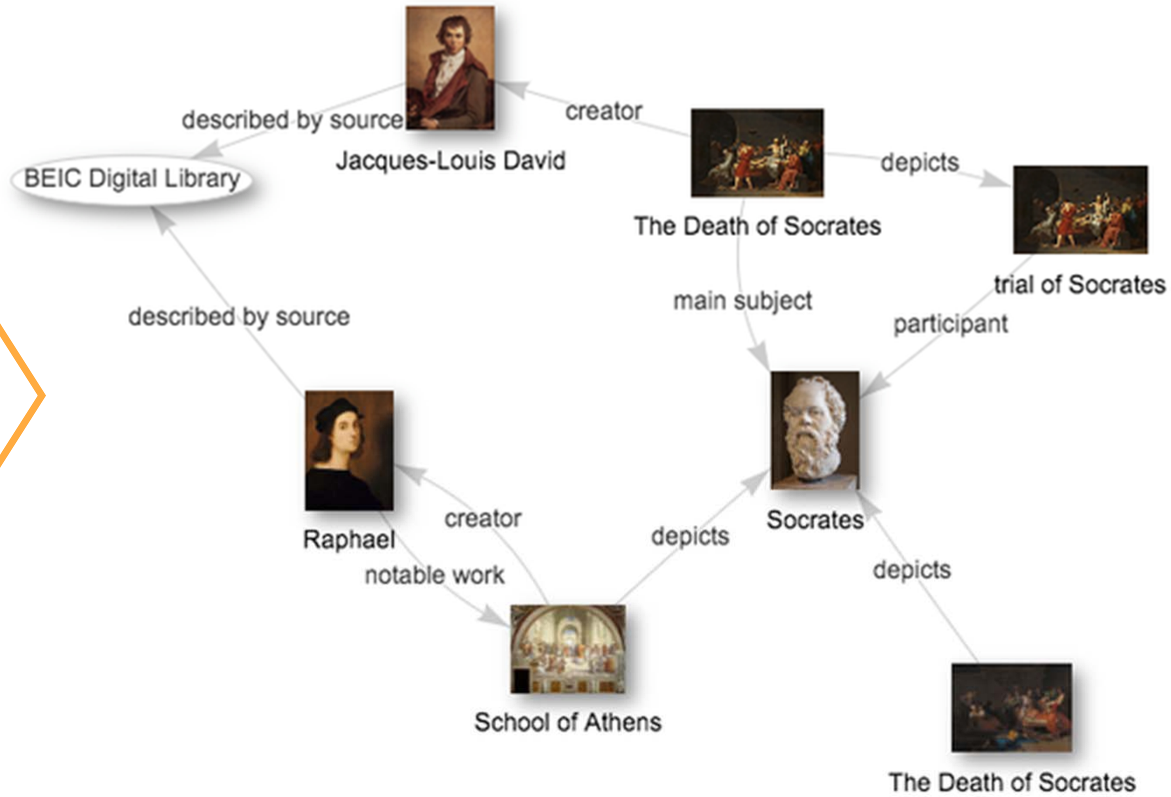
from: Dong. KDD2018

Knowledge Graph Construction

unstructured (TEXT)
or multimodal data
(audio, images, videos)

semi-structured
(e.g., JSON, CSV)

structured
(RDB, KGs)



arXiv preprint: Hofer, M., Obraczka, D., Saeedi, A., Köpcke, H., & Rahm, E. (2023). Construction of Knowledge Graphs: State and Challenges. *ArXiv,* [abs/2302.11509](abs/2302.11509).

Construction of Knowledge Graphs: State and Challenges

Marvin Hofer [b,*], Daniel Obraczka [b], Alieh Saeedi [a,b], Hanna Köpcke [a] and Erhard Rahm [a,b]

[a] *Dept. of Computer Science, Leipzig University, Germany*

[b] *Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) Dresden/Leipzig, Germany*

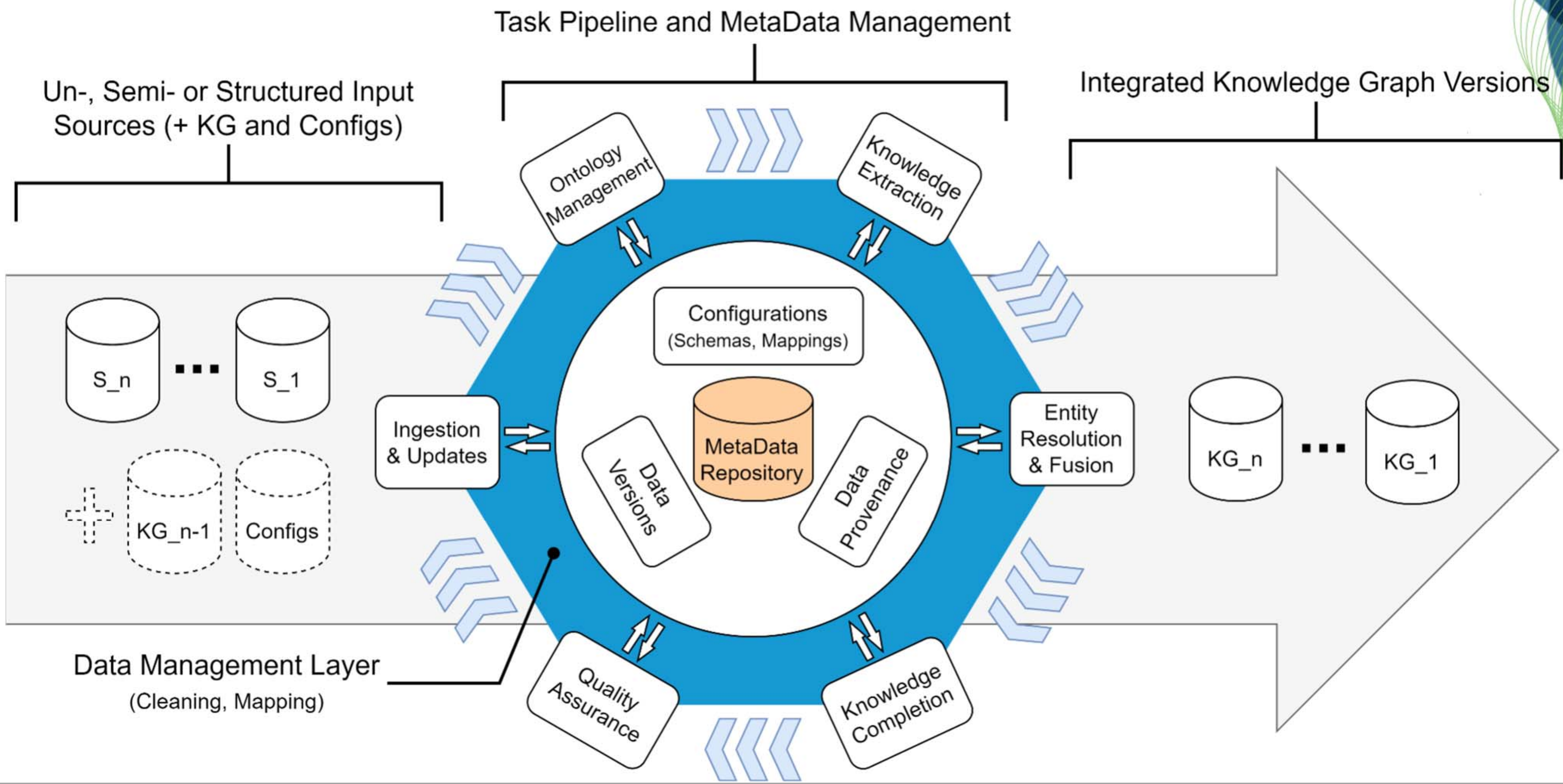
Abstract. With knowledge graphs (KGs) at the center of numerous applications such as recommender systems and question answering, the need for generalized pipelines to construct and continuously update such KGs is increasing. While the individual steps that are necessary to create KGs from unstructured (e.g. text) and structured data sources (e.g. databases) are mostly well-researched for their one-shot execution, their adoption for incremental KG updates and the interplay of the individual steps have hardly been investigated in a systematic manner so far. In this work, we first discuss the main graph models for KGs and introduce the major requirement for future KG construction pipelines. Next, we provide an overview of the necessary steps to build high-quality KGs, including cross-cutting topics such as metadata management, ontology development, and quality assurance. We then evaluate the state of the art of KG construction w.r.t the introduced requirements for specific popular KGs as well as some recent tools and strategies for KG construction. Finally, we identify areas in need of further research and improvement.

Keywords: Knowledge Graph, Data Integration, Data Science

Requirements for KG construction

- Input Data Requirements
  - support for many, large and heterogenous data sources
  - techniques for data acquisition, knowledge extraction, entity resolution/fusion
- Support for Incremental KG updates
  - process new input data in batches or continuously in a streaming manner
  - series of batch-created KG versions vs. incremental updates of changes/new sources
  - tradeoffs in simplicity vs. scalability /freshness
- Pipeline and Tools Requirements
  - tool support needed to simplify KG construction (creation of application-specific pipelines)
  - utilize existing, independently developed tools
  - simplified configuration of individiual steps
  - support for debugging and tuning
- Quality Assurance
  - ensure high data quality in individual pipeline steps and in resulting KG

Pipeline Blueprint



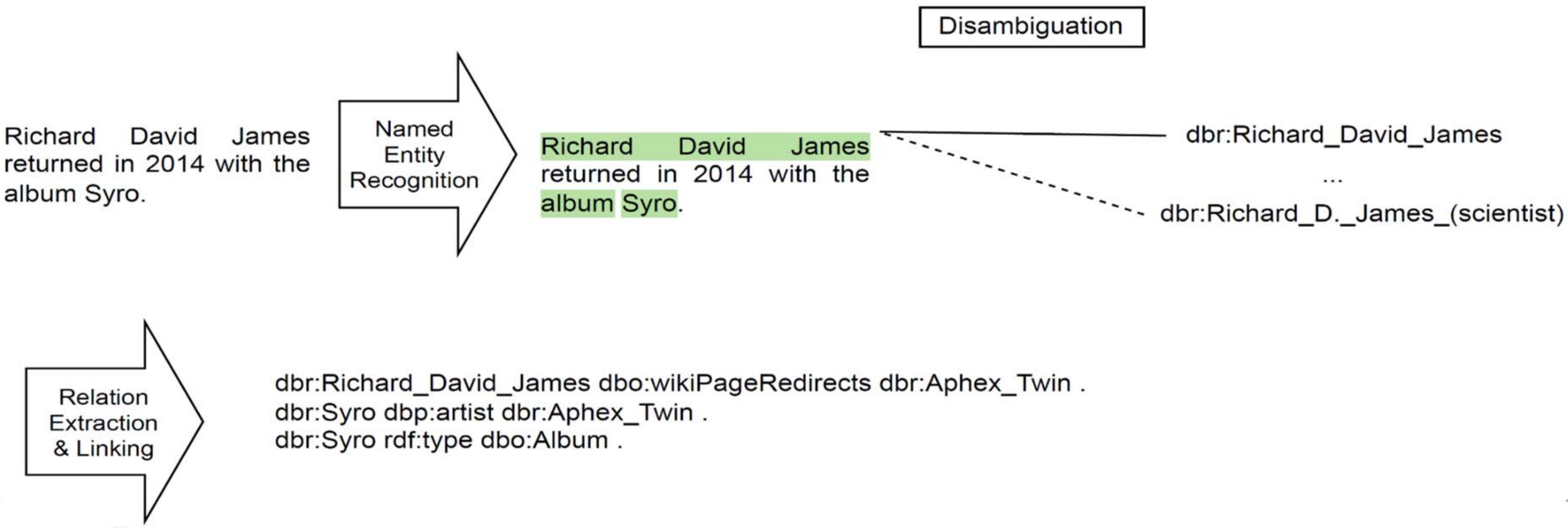
Overview of KG Construction Tasks

- **Initial KG construction:** manual crowdsourcing, sampling existing KG
- **Data preprocessing:** data acquisition, data cleaning and transformation
- ***Metadata management:** persistence, access, versioning, provenance
- ***Ontology development:** creation, evolution, integration
- **Knowledge extraction:** entity recognition, linking, relation extraction
- **Entity resolution:** entity matching, clustering, data fusion
- ***Quality assurance:** quality assessment, repair, debugging
- **Knowledge completion:** type-, link prediction, enrichment, polishing

*cross-cutting and special tasks

Knowledge Extraction

- bringing unstructured or semi-structured data to structured, machine-readable information
- subtasks: **Named-Entity Recognition (NER), Entity Linking (EL), and Relation Extraction (RE)**
- multi-modal KE: visual relation extraction from images



Quality Assurance

- high KG quality crucial for credibility and usability
- subtasks: **quality evaluation** (identifying issues) and **quality improvement** (fixing issues) / **KG completion**
- Quality evaluation
  - dimensions: accuracy, consistency, timeliness, completeness, trustworthiness, availability
  - manual checks (experts, crowd-sourcing), statistical analysis, semantic reasoning, comparison with external sources
- Quality improvement
  - Error correction, data cleaning, entity resolution and fusion
  - ontology evolution
- **Knowledge completion:** improve KG by new nodes, relations, properties
  - type completion: Assigning types to nodes lacking type information using node classification, logical reasoning, or statistical approaches.
  - link prediction: Identifying missing relations in KG, with techniques like distant supervision, embedding-based methods, or Graph Neural Networks.
  - data enrichment: add entity information from external knowledge bases, e.g. using persistent identifiers (ISBN, DOIs, ORCIDs …)

Exemplary Selection and Comparison

- Investigation of 23 specific KGs/construction approaches and toolsets
  - 3 closed KGs: Google, Diffbot, Amazon
  - 3 manually curated KGs: Freebase, Wikidata, ORKG
  - 10 open KGs: DBPedia, DBPedia-live, YAGO, NELL, ArtistKG, CovidKG, …
  - **7 toolsets** for KG construction: FlexiFusion, dstlr, XI, Autoknow, HKGB, SLOGERT; Saga
- selection based on relevance (popularity), novelty, existing paper/documentation, with multiple versions

	Year	Domain	Srcs.	Model	Entities	Relations	Types	R-Types	Vers.	Update
Closed KG										
Google KG [195]	2012	Cross,MLang	>>>1	Custom,RDF	1B	>100B	?	?	?	?
Diffbot.com	2019	Cross	>>>1	RDF	5.9B	>1T	?	?	?	?
Amazon PG [196]	2020	Products	>1	Custom	30M	1B	19K	1K	?	?
Open Access KG										
*Freebase [197]	2007	Cross	>>1	RDF	22M	3.2B	53K	70K	>1	2016
DBpedia [198]	2007	Cross,MLang	140	RDF	50M	21B	1.3K	55K	>20	2023
YAGO [199, 200]	2007	Cross	2-3	RDF(-Star)	67M	2B	10K	157	5	2020
NELL [201]	2010	Cross	⩾1	Custom,RDF	2M	2.8M	1.2K	834	>1100	2018
*Wikidata [202]	2012	Cross,MLang	>>>1	RDB/RDF	100M	14B	300K	10.3K	>100	2023
DBpedia-EN Live [203]	2012	Cross	1	RDF	7.6M	1.1B	800	1.3K	>>>1	2023
Artist-KG [204]	2016	Artists	4	Custom	161K	15M	>1	18	1	2016
*ORKG [205]	2019	Research	>>1	RDF	130K	870K	1.3K	6.3K	>1	2023
AI-KG [206]	2020	AI Science	3	RDF	820K	1.2M	5	27	2	2020
CovidGraph [207]	2020	COVID-19	17	PGM	36M	59M	128	171	>1	2020
DRKG [208]	2020	BioMedicine	>7	CSV	97K	5.8M	17	107	1	2020
VisualSem [209]	2020	Cross,MLang	2	Custom	90k	1.5M	(49K)	13	2	2020
WorldKG [210]	2021	Geographic	1	RDF	113M	829M	1176	1820	1	2021

*manually curated

				Consumed Data					(Meta)Data			Performed Construction Tasks							
Name of System	System Version/Year	Open Implementation	Incremental Integration	Unstructured Data	Semi-Structured Data	Structured Data	(Event-)Stream Data	Supplementary Input	Deep Provenance	Temporal Data	Additional Metadata	KG Initialization	Input Cleaning	Ontology Management	Knowledge Extraction	Entity Resolution	Entity/Value Fusion	Quality Assurance	Knowledge Completion
Dataset Specific																			
DBpedia	2019	✓			✓			✓	✓	✓	✓	○	●	○	○			●	○
YAGO4	2020	✓			✓	✓		✓	✓	✓		○	○	●				●	
DBpedia-Live	2012	✓	○		✓		✓	✓	✓	✓		○	●	○	○				
NELL	2018		●	✓	✓			✓	✓	✓		○		●	●			○	
Artist-KG	2016	✓	○		✓	✓						○	○	●		●			
AI-KG	2020		?		✓			✓	✓	✓		○			●			○	
CovidGraph	2020	✓	○	✓	✓	✓		✓	✓	?		○		?	●	○			
DRKG	2020	✓			✓	✓			✓			?			○		○		●
VisualSem	2020	✓		✓	✓	✓		✓				○	●		○				
WorldKG	2021	✓			✓			✓				●	●	●	○			○	
Toolset/Strategy																			
FlexiFusion [90]	2019				✓	✓		✓	✓		✓	○	○				●		
dstlr [137]	2019	✓	?	✓					✓			○			●			○	○
XI [50]	2020		?	✓	✓				?	?	?	○			●			?	
AutoKnow [196]	2020			✓	✓			✓				○	●	●	●				●
HKGB [211]	2020		○		✓			✓		✓		●		●	●	?		○	●
SLOGERT [212]	2021	✓			✓				✓	✓		○			●	?			○
SAGA [47]	2022		●	✓	✓	✓	✓	✓	✓	✓		?	●	○	●	●	●	●	●

✓ *supported/provides*

○ *simple/manual*

● *sophisticated/semi-automatic*

? *unclear implementation*

Name of System	System Version/Year	Open Implementation	Incremental Integration	Consumed Data: Unstructured Data	Consumed Data: Semi-Structured Data	Consumed Data: Structured Data	Consumed Data: (Event-)Stream Data	Consumed Data: Supplementary Input	(Meta)Data: Deep Provenance	(Meta)Data: Temporal Data	(Meta)Data: Additional Metadata	Performed Construction Tasks: KG Initialization	Performed Construction Tasks: Input Cleaning	Performed Construction Tasks: Ontology Management	Performed Construction Tasks: Knowledge Extraction	Performed Construction Tasks: Entity Resolution	Performed Construction Tasks: Entity/Value Fusion	Performed Construction Tasks: Quality Assurance	Performed Construction Tasks: Knowledge Completion
Dataset Specific																			
DBpedia	2019	✓			✓			✓	✓	✓	✓	○	●	○	○			●	○
YAGO4	2020	✓			✓	✓		✓	✓	✓		○	○	●				●	
DBpedia-Live	2012	✓	○		✓		✓	✓	✓	✓		○	●	○	○				
NELL	2018		●	✓	✓			✓	✓	✓		○		●	●			○	
Artist-KG	2016	✓	○		✓	✓						○	○	●		●			
AI-KG	2020		?		✓			✓	✓	✓		○			●			○	
CovidGraph	2020	✓	○	✓	✓	✓		✓	✓	?		○		?	●	○			
DRKG	2020	✓			✓	✓			✓			?			○		○		●
VisualSem	2020	✓		✓	✓	✓		✓				○	●		○				
WorldKG	2021	✓			✓			✓				●	●	●	○			○	
Toolset/Strategy																			
FlexiFusion [90]	2019				✓	✓		✓	✓		✓	○	○				●		
dstlr [137]	2019	✓	?	✓					✓			○			●			○	○
XI [50]	2020		?	✓	✓				?	?	?	○			●			?	
AutoKnow [196]	2020			✓	✓			✓				○	●	●	●				●
HKGB [211]	2020		○		✓			✓		✓		●		●	●	?		○	●
SLOGERT [212]	2021	✓			✓				✓	✓		○			●	?			○
SAGA [47]	2022		●	✓	✓	✓	✓	✓	✓	✓		?	●	○	●	●	●	●	●

simple matching, no fusion

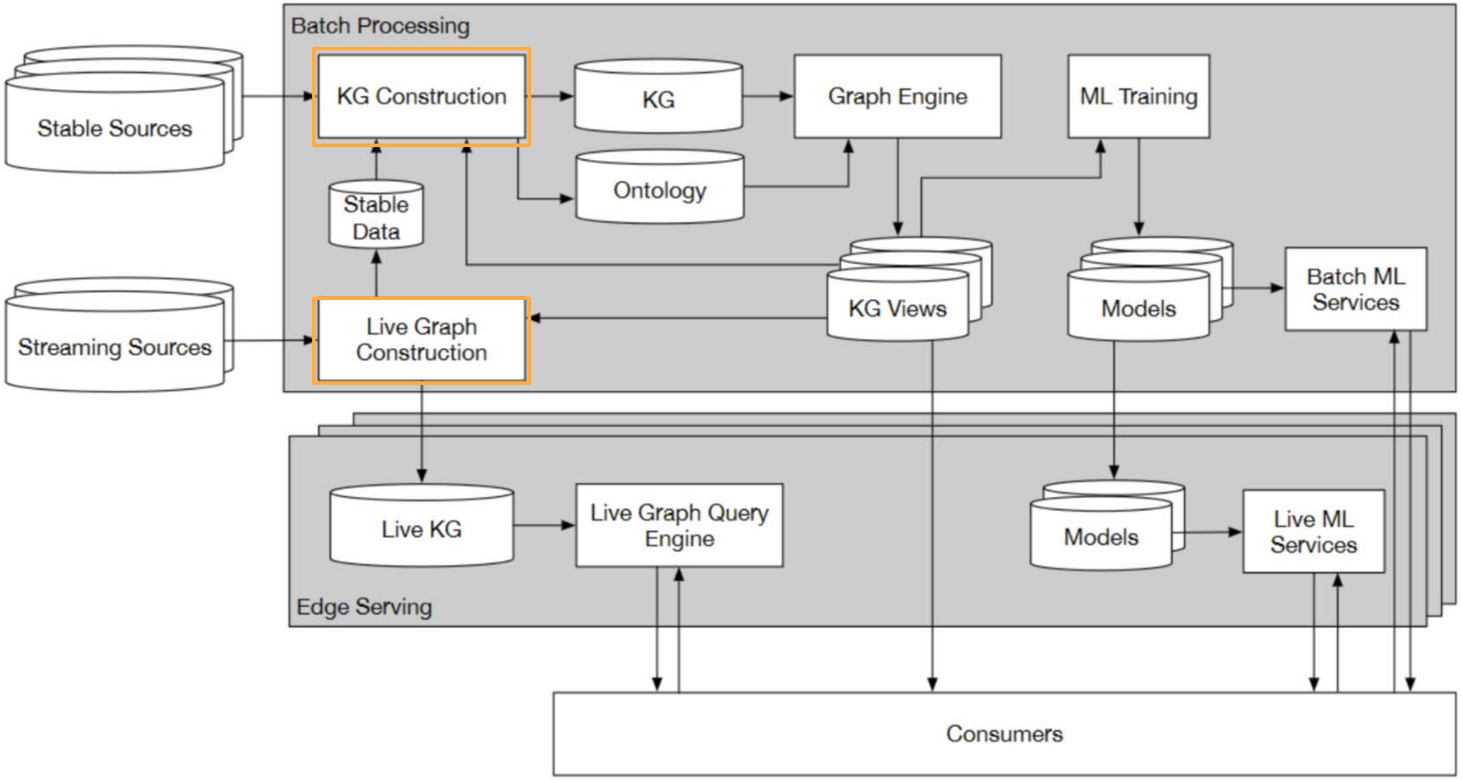
✓ *supported/provides*

○ *simple/manual*

● *sophisticated/semi-automatic*

? *unclear implementation*

SAGA tool (Apple, Ilyas et al., Sigmod 2022)



Open challenges in KG construction

- better support for **incremental KG construction**
  - batch-like KG re-creation has limited scalability and out-of-date information
  - more complex: change detection in sources and incremental pipeline
- **lack of open tools** for KG construction
- toolset for defining different KG construction pipelines with different implementations for certain tasks (extensible, modular approach needed)
- more comprehensive approaches needed for **metadata management** and KG **quality assurance**
- **evaluation of KG construction approaches**
  - so far only benchmarks for single tasks (extraction, matching, completion)
  - not sufficient to evaluate/compare different end-to-end construction approaches
- **use of Large Language Models (LLMs)** for KG construction

AGENDA

- ScaDS.AI Dresden/Leipzig
- Construction of Knowledge Graphs
  - KG intro
  - requirements for KG construction
  - processing steps
  - comparison of existing approaches
  - open challenges
- **Entity resolution / matching**
  - ER intro
  - entity clustering and incremental ER (Famer)
  - embedding-based matching of KGs
- Conclusions

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 2790 Related articles All 29 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 7 Related articles

Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23 (4), 3–13 *
E Rahm, H Do - 2000
Cited by 7 Related articles

Data engineering—Special issue on data cleaning *
E Rahm, HH Do - Data Engineering, 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 5 Related articles

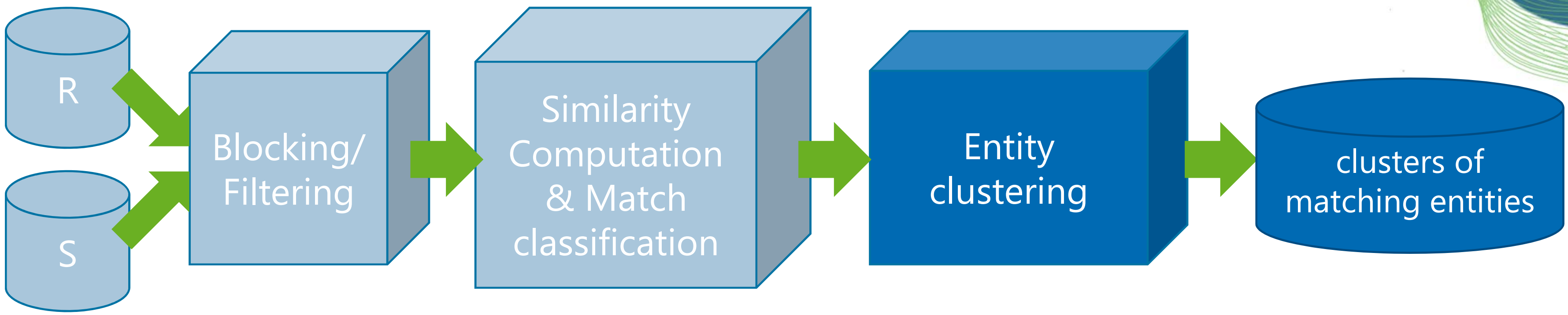
Data cleaning: Problems and current approaches' IEEE Data Eng. Bull., 2000 *
E Rahm, HH Do - 2000
Cited by 5 Related articles

Do. H. 2000. Data cleaning: Problems and current approaches *
E Rahm, HAI HONG - IEEE Data Engineering Bulletin
Cited by 5 Related articles

ER CHALLENGES

- Scalability
  - large data volume or/and many sources
  - need to reduce search space (e.g. with blocking) + parallel processing
- High match quality
  - low quality input data (unstructured, semi-structured sources)
  - needs effective combination of several techniques
  - use of supervised ML approaches
  - use of entity embeddings
- Support for evolution and change
  - addition of new sources and new entities without having to integrate everything again
  - incremental / dynamic vs batch / static ER

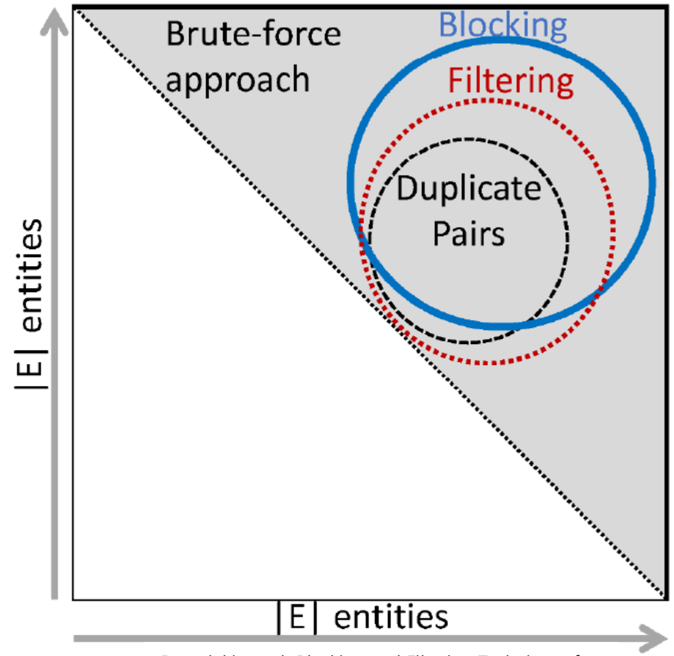
ENTITY RESOLUTION WORKFLOW



- 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
  - for embeddings: only consider nearest neighbors



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

BLOCKING TECHNIQUES

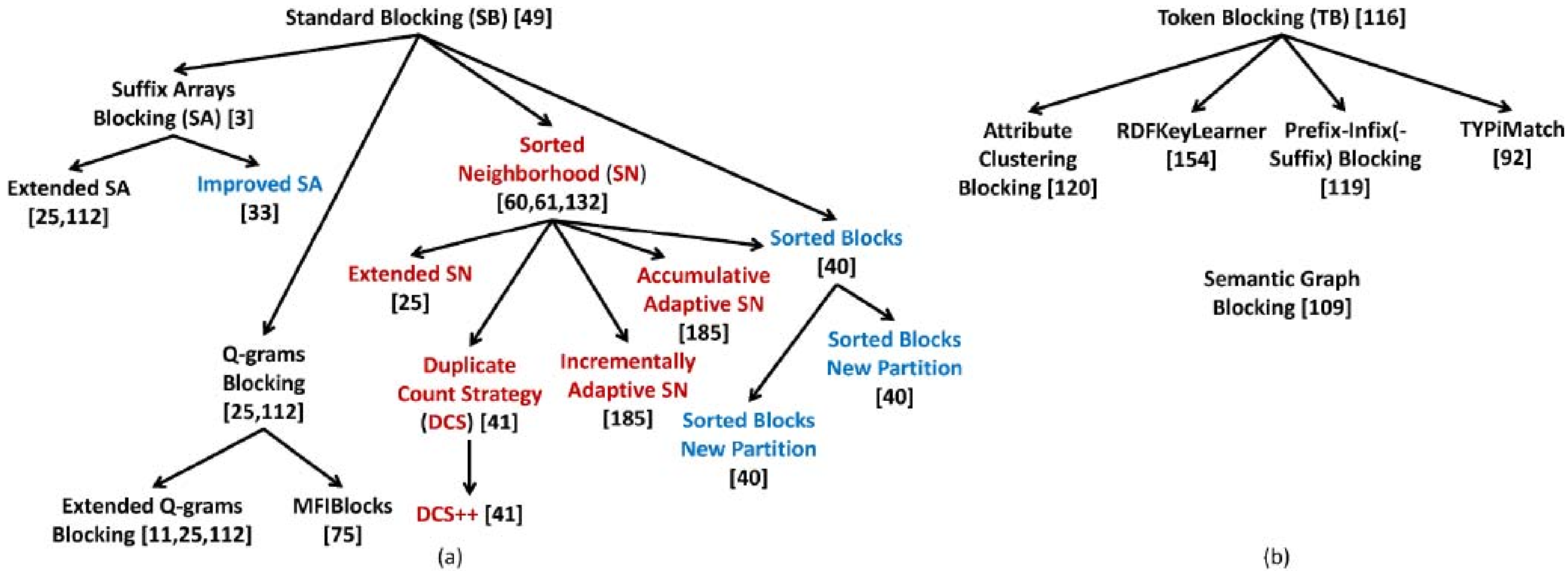


Fig. 3. The genealogy trees of nonlearning (a) schema-aware and (b) schema-agnostic Block Building techniques. Hybrid, hash-, and sort-based methods are marked in blue, black, and red, respectively.

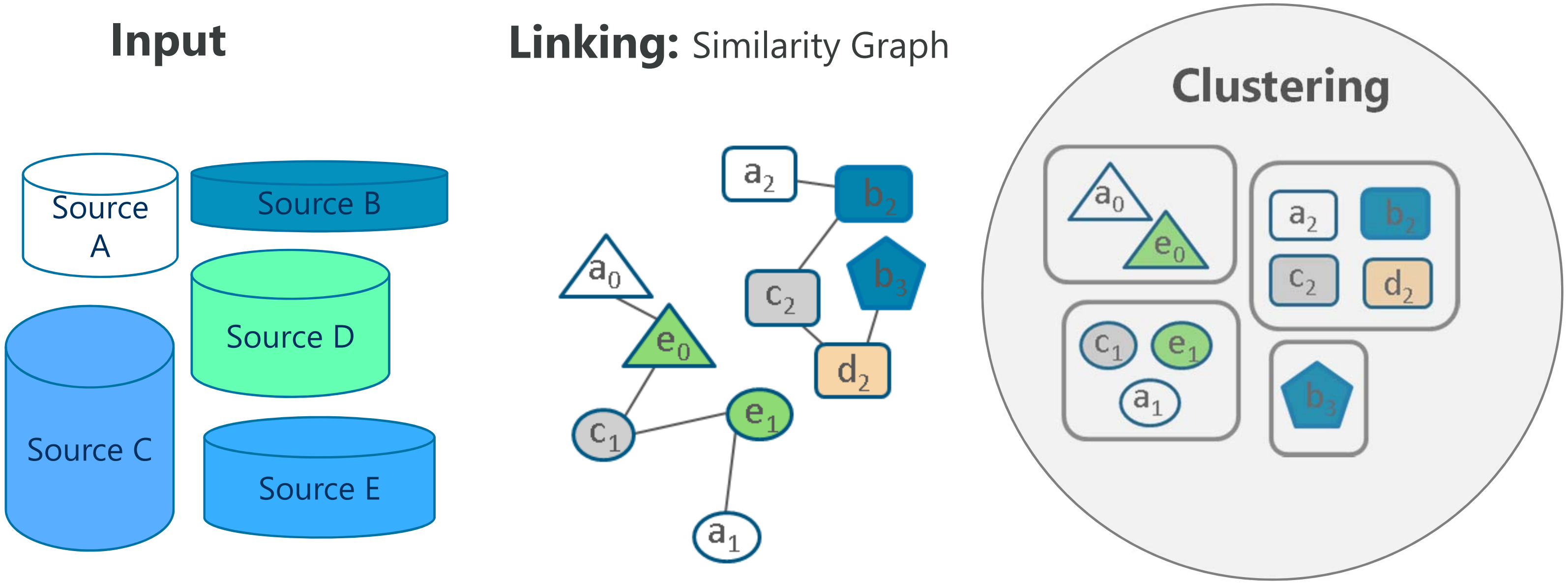
Papadakis et al: *Blocking and Filtering Techniques for Entity Resolution: A Survey*. ACM CSUR 2020

MATCHING

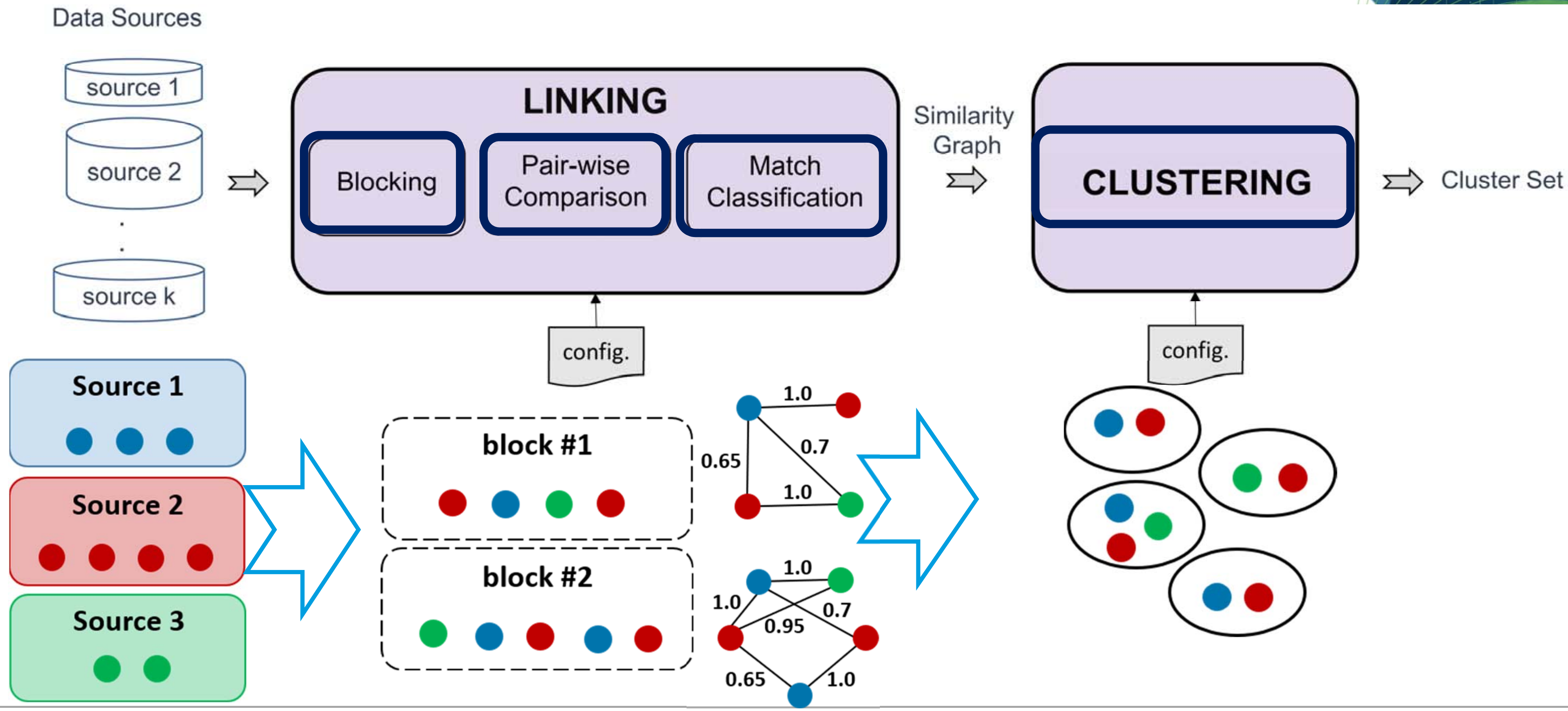
- combined use of several similarity values
  - attribute similarities, e.g. using numeric or string similarity measures
  - context-based matchers
- general match rules with multiple similarties
  - e.g. pubs match if *title sim.* $\geq$ *0.9 & author sim. > 0.4*
- learned/supervised match classification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

FAMER TOOL

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

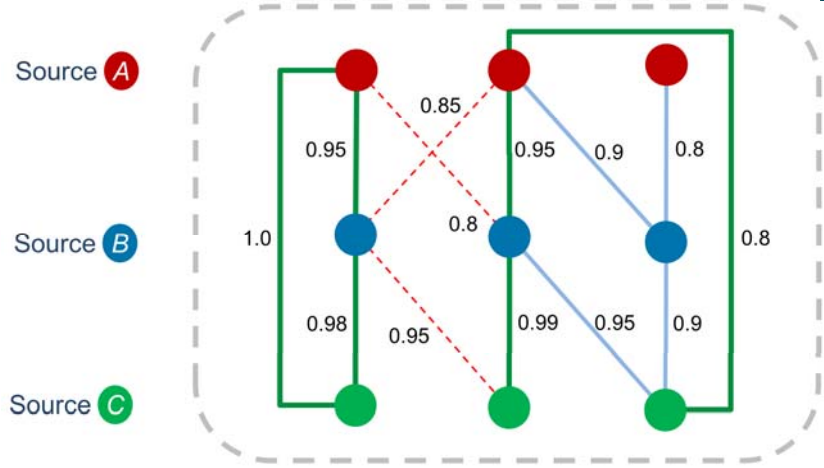


FAMER BATCH PIPELINE

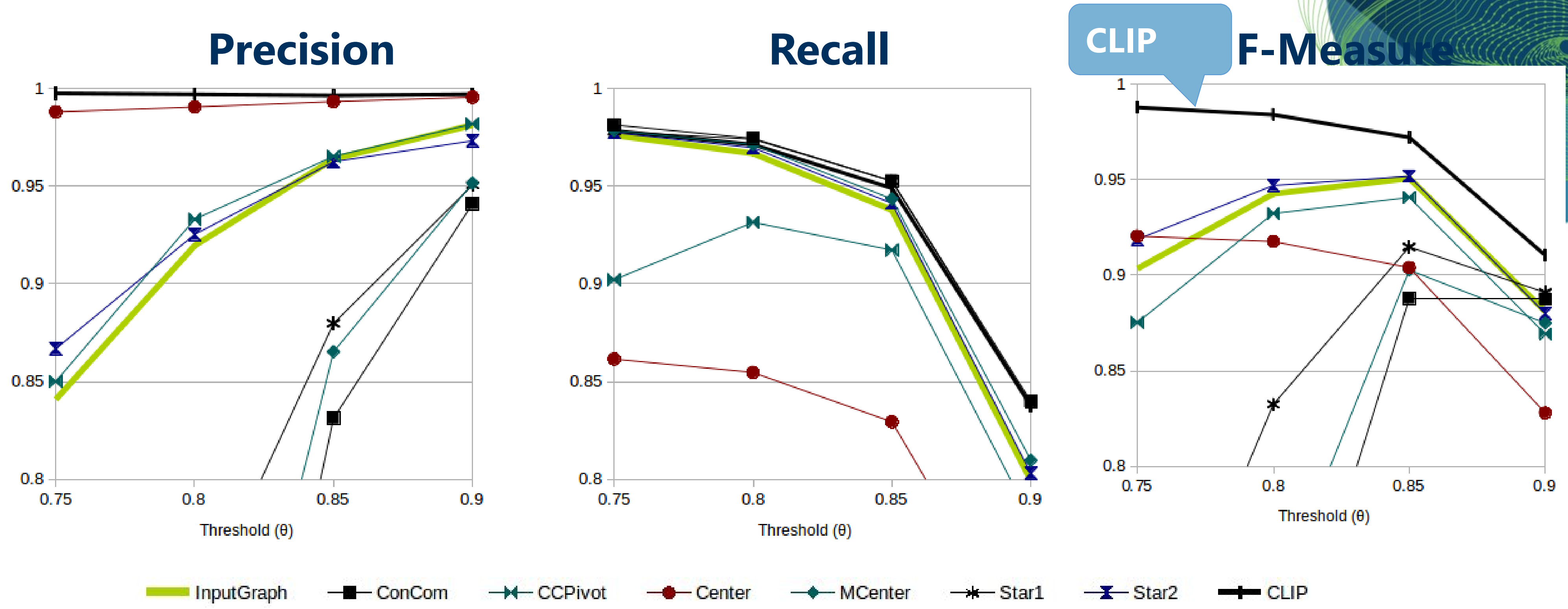


CLIP APPROACH (ESWC BEST RESEARCH PAPER)

- optimized for clean sources
- 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



EVALUATION: GEO. DATASET



MULTI-SOURCE CLEAN/DIRTY CLUSTERING

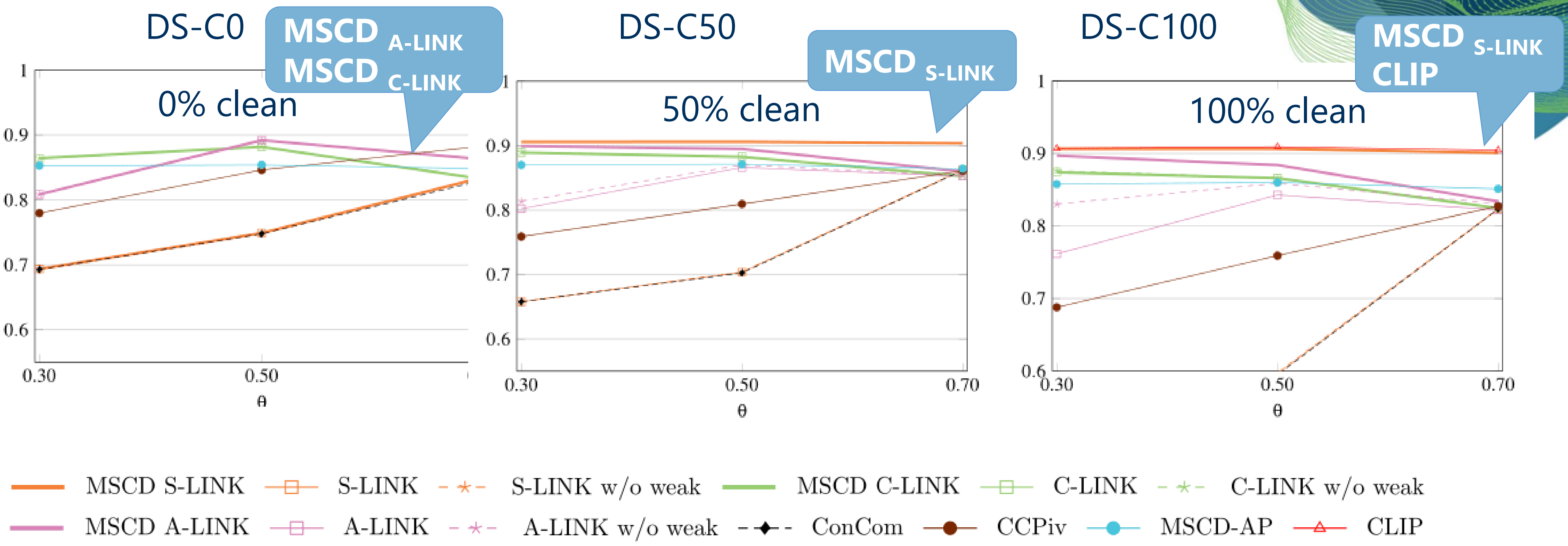
- previous assumption: data sources are duplicate-free
- more realistic assumption: some sources are dirty
  - solution: first deduplicate dirty sources
  - problem: requires immense effort and perhaps not completely successful
- solution: **MSCD approaches**
  - approaches that can deal with dirty sources
  - 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
  - most promising: hierarchical agglomerative clustering (HAC)

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 $NN(c_i) = c_j$
    - observe that 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)}

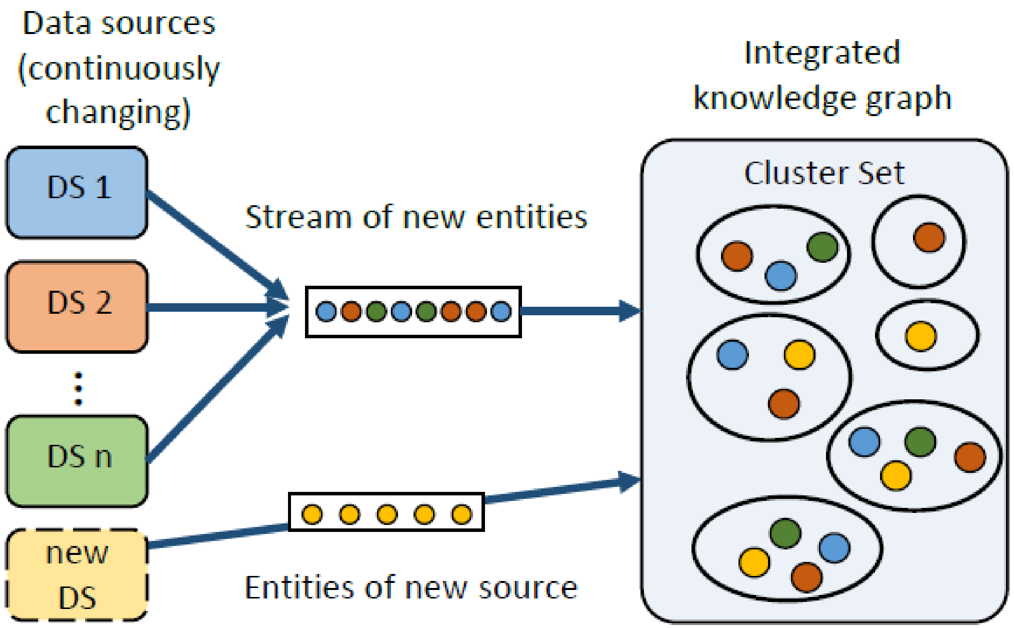
F-MEASURE: CAMERA DATASET

match threshold = merge threshold (θ)

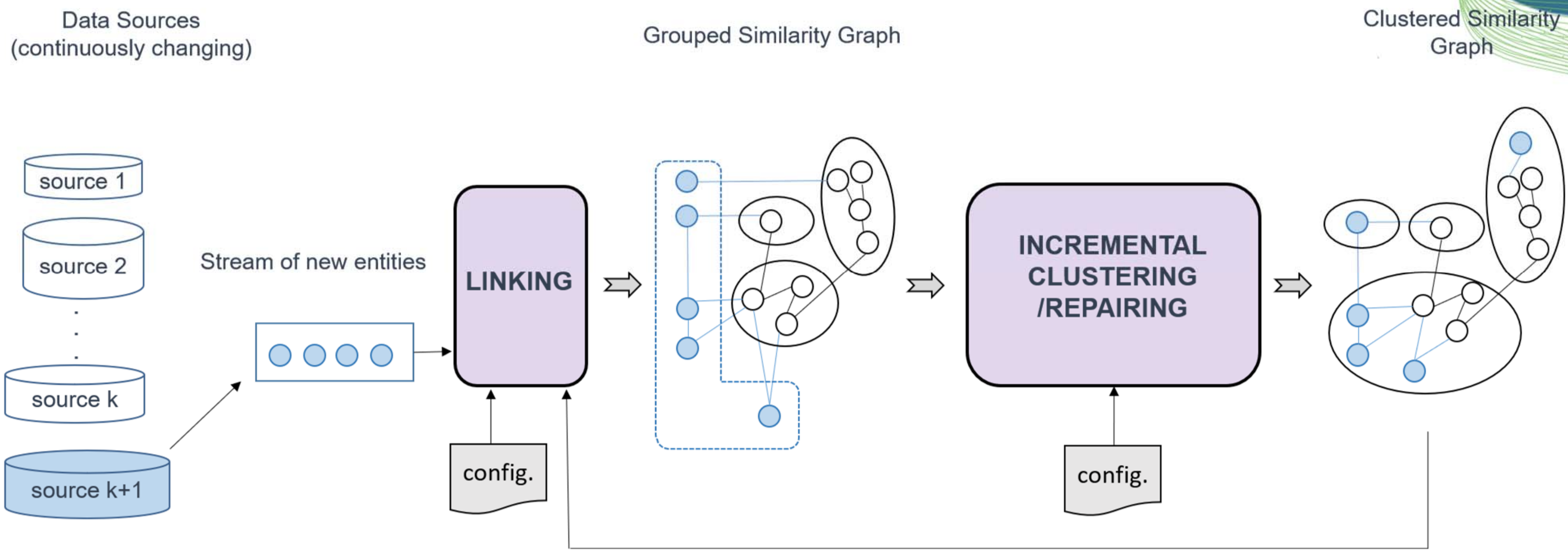


MOTIVATION

- 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

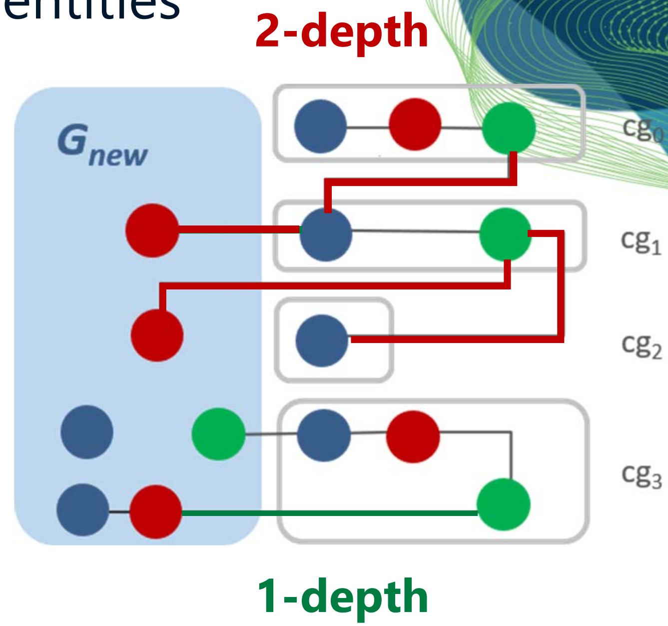


FAMER INCREMENTAL PIPELINE



FAMER N-DEPTH RECLUSTERING

- requires to keep similarity graphs for clustered entities
- recluster new entities in G_{new} with their neighbors
  - can repair old cluster decisions
  - limits amount of reclustering for efficiency
  - reduce dependence on order of entity additions
- evaluation results
  - incremental approaches are much faster and similarly effective than batch ER
  - quality of nDR does not depend on the order in which new entities are added



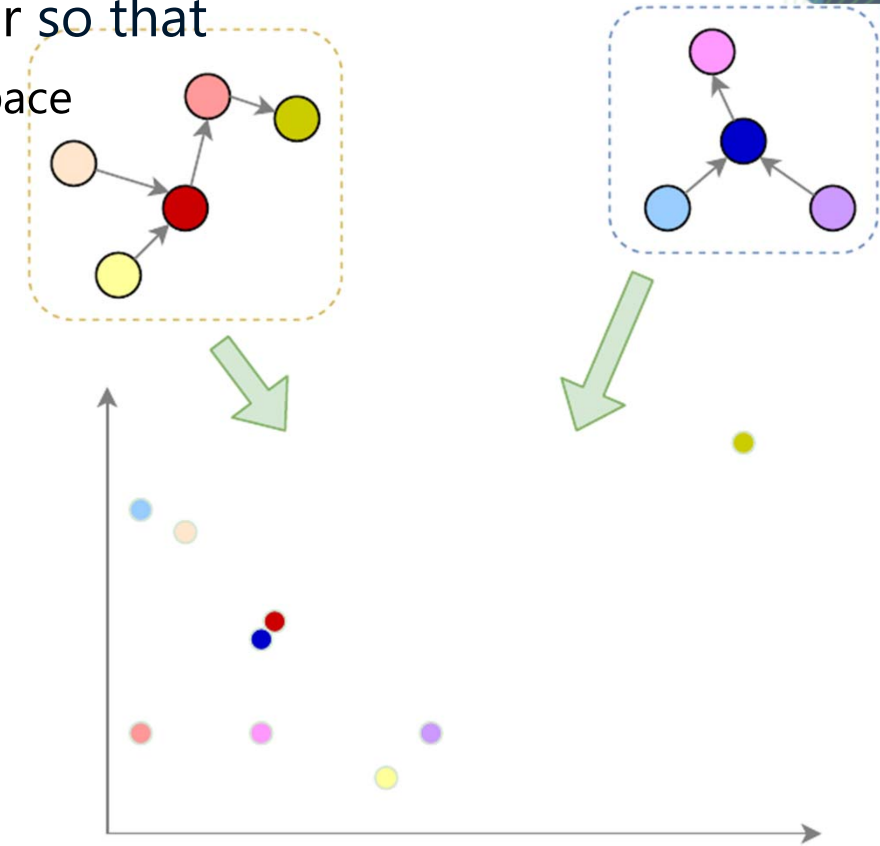
ENTITY RESOLUTION ON KNOWLEDGE GRAPHS

- similar ER challenges as discussed
  - large KGs (e.g., 100 million entities in Wikidata)
- **ER for many interrelated entity types needed**
  - standard ER assumes only 1 entity type

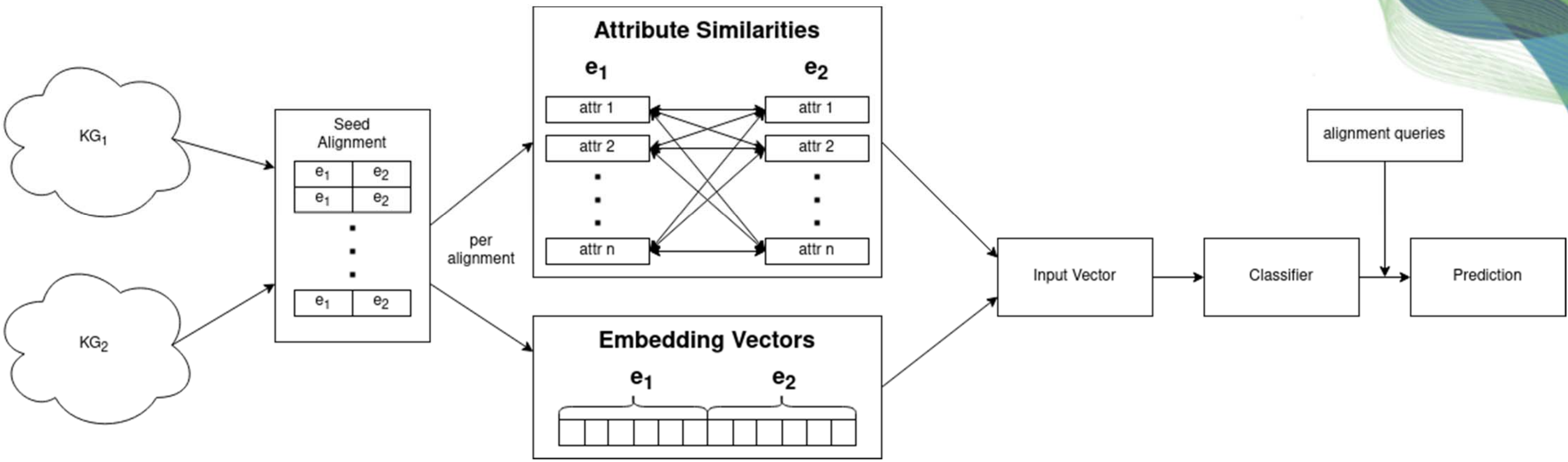
- **Key idea**: map entities of input KGs into embedding space and determine matches based on nearest neighborhood
  - word embeddings for properties/attribute values
  - graph embeddings to consider neighboring entities in KG

KNOWLEDGE GRAPH EMBEDDINGS (KGE)

- transform entities into a dense vector so that
  - similar entities close in the embedding space
  - relational information is retained
- many possible approaches
  - translational KGEs for triples <h,r,t> (e.g. MultiKE, BootEA)
  - Graph Neural Network approaches (e.g. RDGCN, CG-MuAlign) based on aggregated entity neighborhood in KG



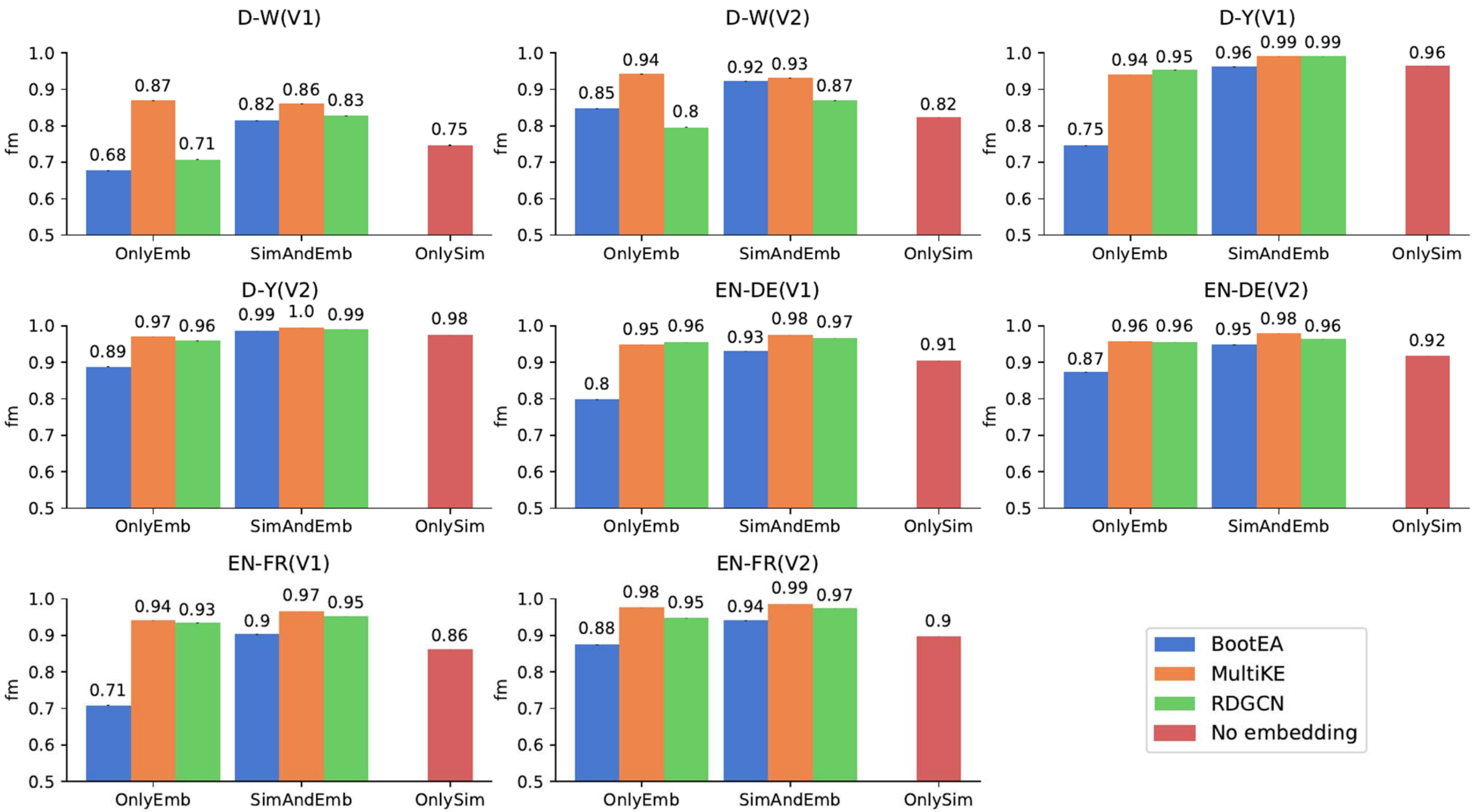
EAGER: EMBEDDING-ASSISTED ENTITY RESOLUTION FOR KG



Obraczka, Schuchart, and Rahm, "Embedding-Assisted Entity Resolution for Knowledge Graphs", 2021

EXPERIMENTAL EVALUATION

- 16 alignment tasks
  - KG subsets from DBpedia, Wikidata, YAGO
  - different densities, sizes and even cross-lingual settings
- 3 KG embedding approaches (BootEA, MultiKGE, RDGCN)
  - best performing approaches from *Sun et al: "A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs", 2020*
- comparison of 3 approaches
  - OnlyEmb – only graph embeddings are used
  - OnlySim: only attribute similarities are used
  - SimAndEmb: use both



Results for 100K datasets (using MLP as classifier)

ScaDS.AI DRESDEN LEIPZIG

UNIVERSITÄT LEIPZIG

PROBLEMS WITH EMBEDDINGS

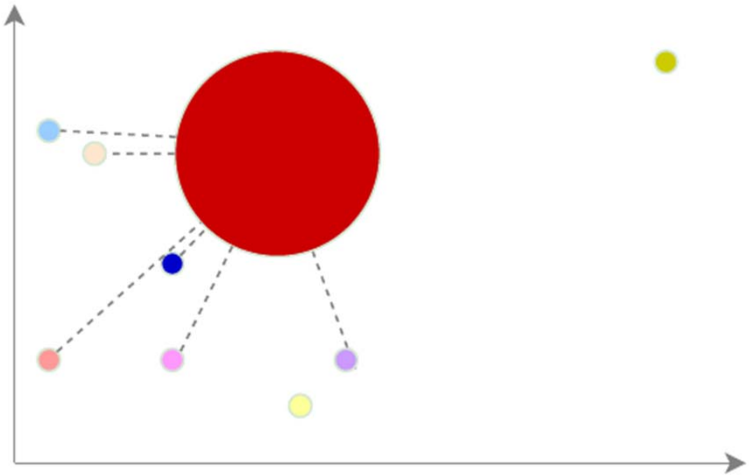
- Problems with runtime and quality für larger and more diverse KGs
- blocking approaches not applicable to speed-up matching
- exact nearest-neighbor algorithms become slow
  - need to apply faster approximate nearest neighbor (ANN) algorithms, e.g. Annoy, Faiss …
- but ANN algorithms lose some matches (reduced recall)
  - embeddings are relatively high-dimensional (> 200)
  - **“hubness”** of embedded entities

HUBNESS REDUCES ALIGNMENT QUALITY

with increasing dimensionality:

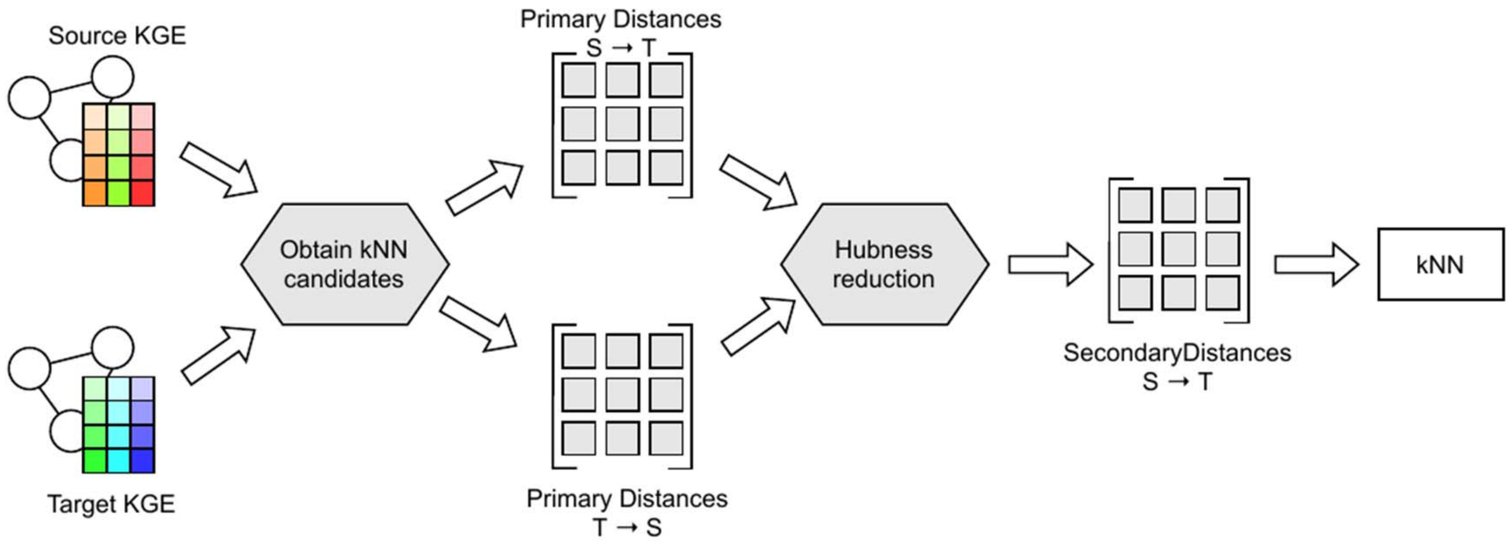
- few points are nearest neighbors (NN) of many points
- many points are NN of no points

⇒ hubness negatively affects alignment quality



kiez

open-source python library (github.com/dobraczka/kiez)
for hubness-reduced nearest neighbor search
(for entity alignment with knowledge graph embeddings)



Obraczka and Rahm, "An Evaluation of Hubness Reduction Methods for Entity Alignment with Knowledge Graph Embeddings", 2021

kiez

Open-source python library (github.com/dobraczka/kiez)
for hubness-reduced nearest neighbor search
(for entity alignment (with knowledge graph embeddings))

(Approximate) Nearest Neighbor Method:

- Sci-kit learn *Pedregosa et al., 2011*
  - BallTree *Omohundro, 1989*
  - KDTree *Bentley, 1975*
  - Bruteforce
- NMSLIB: HNSW *Malkov, 2018*
- NGT *Iwasaki, 2016*
- Annoy *(github.com/spotify/annoy)*
- Faiss *Johnson, Douze, and Jégou, 2017*

Hubness reduction methods:

- Local Scaling *Schnitzer et al., 2012*
- NICDM *Schnitzer et al., 2012*
- CSLS *Lample et al., 2018*
- Mutual Proximity *Schnitzer et al., 2012*
- DisSimLocal *Hara et al., 2016*

EVALUATION RESULTS

- hubness reduction improves alignment results
- using ANN algorithms (Faiss) with hubness reduction approach (NICDM) gives improvements at virtually no cost w.r.t speed

⇒ hubness reduction largely offsets decrease in alignment quality when using *approximate* nearest neighbor algorithm while still retaining speed advantage

FUTURE DIRECTIONS FOR KGE-BASED METHODS

- more realistic evaluations[1]
  - differently sized KGs, not only 1:1 matches, …
- better scalability of KGE-based methods
  - blocking-like approaches not yet explored
- dealing with unseen entities is almost unexplored[2]
- unsupervised KGE approaches, e.g. for clustering

[1] *Leone et al., "A Critical Re-evaluation of Neural Methods for Entity Alignment", 2022*
[2] *Wang et. al.,"Facing Changes: Continual Entity Alignment for Growing Knowledge Graphs",2022*

SUMMARY

- largely automatic creation/refinement of large **knowledge graphs** is still difficult
  - open toolsets needed supporting all major steps with easy configuration
  - better approaches needed for incremental updates, quality assurance, ontology evolution, multi-modal KGs …
  - holistically evaluating KG construction approaches is challenging
- **Entity resolution**
  - huge amount of previous work mostly on structured and static data for single kind of entities
  - need for incremental approaches for KGs with many entity types
  - use of KG embeddings promising but with need for improvements

References Knowledge Graphs

- Al-Aswadi, F.N., Chan, H.Y., & Gan, K.H. (2019). Automatic ontology construction from text: a review from shallow to deep learning trend. Artificial Intelligence Review, 53, 3901 - 3928.
- Christophides, V., Efthymiou, V., Palpanas, T., Papadakis, G., & Stefanidis, K. (2020). An Overview of End-to-End Entity Resolution for Big Data. ACM Computing Surveys (CSUR), 53, 1 - 42.
- Dong et al. (2020). AutoKnow: Self-Driving Knowledge Collection for Products of Thousands of Types. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.
- Ehrlinger, L., & Wöß, W. (2016). Towards a Definition of Knowledge Graphs. International Conference on Semantic Systems.
- Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P., & Santos, C.T. (2011). Ontology Alignment Evaluation Initiative: Six Years of Experience. J. Data Semant., 15, 158-192.
- Hofer, M., Obraczka, D., Saeedi, A., Köpcke, H., & Rahm, E. (2023). Construction of Knowledge Graphs: State and Challenges. ArXiv
- Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Gayo, J.E., Kirrane, S., Neumaier, S., Polleres, A., Navigli, R., Ngomo, A.N., Rashid, S.M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2020). Knowledge Graphs. ACM Computing Surveys (CSUR), 54, 1 - 37.
- Huang, X., Zhang, J., Li, D., & Li, P. (2019). Knowledge Graph Embedding Based Question Answering. Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining.
- Ilyas, I.F., Rekatsinas, T., Konda, V.V., Pound, J., Qi, X., & Soliman, M.A. (2022). Saga: A Platform for Continuous Construction and Serving of Knowledge at Scale. Proc. SIGMOD
- Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P.S. (2020). A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. IEEE Transactions on Neural Networks and Learning Systems, 33, 494-514.

References

- Li, J., Sun, A., Han, J., & Li, C. (2018). A Survey on Deep Learning for Named Entity Recognition. IEEE Transactions on Knowledge and Data Engineering, 34, 50-70.
- Noy, N. (2001). Ontology Development 101: A Guide to Creating Your First Ontology.
- Paulheim, H. (2016). Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web, 8, 489-508.
- Rekatsinas, T., Chu, X., Ilyas, I.F., & Ré, C. (2017). HoloClean: Holistic Data Repairs with Probabilistic Inference. ArXiv, abs/1702.00820.
- Suchanek, F.M., Abiteboul, S., & Senellart, P. (2011). PARIS: Probabilistic Alignment of Relations, Instances, and Schema. Proc. VLDB Endow., 5, 157-168.
- Tamašauskaitė, G., & Groth, P. (2022). Defining a Knowledge Graph Development Process Through a Systematic Review. ACM Transactions on Software Engineering and Methodology, 32, 1 - 40.
- Van Assche, D., Delva, T., Haesendonck, G., Heyvaert, P., De Meester, B., & Dimou, A. (2022). Declarative RDF graph generation from heterogeneous (semi-)structured data: A systematic literature review. J. Web Semant.
- Weikum, G., Dong, X. L., Razniewski, S., & Suchanek, F. (2021). Machine knowledge: Creation and curation of comprehensive knowledge bases. Foundations and Trends® in Databases, 10(2-4), 108-490.
- West, R., Gabrilovich, E., Murphy, K.P., Sun, S., Gupta, R., & Lin, D. (2014). Knowledge base completion via search-based question answering. Proc. 23rd WWW conf.
- Wilkinson et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data
- Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2015). Quality assessment for Linked Data: A Survey. Semantic Web, 7, 63-93.

References Data Integration / Entity resolution

- D. Ayala, I. Hernández, D. Ruiz, Rahm, Erhard: LEAPME: Learning-based Property Matching with Embeddings. Data & Knowledge Engineering 2022
- Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. Advances in Neural Information Processing Systems.
- 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

References

- 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
- Leone, M., Huber, S., Arora, A., García-Durán, A., & West, R. (2022). A Critical Re-evaluation of Neural Methods for Entity Alignment. Proc. VLDB Endow., 15(8), 1712–1725.
- 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, M. Hartung, A. Ngonga, E. Rahm: A Survey of Current Link Discovery Frameworks. Semantic Web Journal, 2017
- G. Papadakis et al: Blocking and Filtering Techniques for Entity Resolution: A Survey. ACM CSUR 2020
- Obraczka, D., & Rahm, E. (2022). Fast Hubness-Reduced Nearest Neighbor Search for Entity Alignment in Knowledge Graphs. SN Comput. Sci.
- 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
- Qi, Z., Zhang, Z., Chen, J., Chen, X., Xiang, Y., Zhang, N., & Zheng, Y. (2021). Unsupervised Knowledge Graph Alignment by Probabilistic Reasoning and Semantic Embedding. In Z.-H. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 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

References

- 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
- Suchanek, F. M., Abiteboul, S., & Senellart, P. (2011). PARIS: Probabilistic Alignment of Relations, Instances, and Schema. Proc. VLDB Endow., 5(3), 157–168
- Sun, Z., Zhang, Q., Hu, W., Wang, C., Chen, M., Akrami, F., & Li, C. (2020). A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs. Proc. VLDB Endow., 13(11), 2326–2340.
- Wang, Y., Cui, Y., Liu, W., Sun, Z., Jiang, Y., Han, K., & Hu, W. (2022). Facing Changes: Continual Entity Alignment for Growing Knowledge Graphs. Proc. ISWC 2022
- M. Wilke, E. Rahm: Towards Multi-modal Entity Resolution for Product Matching. GVDB 2021
- Wu, Y., Liu, X., Feng, Y., Wang, Z., Yan, R., & Zhao, D. (2019). Relation-aware entity alignment for heterogeneous knowledge graphs. IJCAI International Joint Conference on Artificial Intelligence, 2019
- Zhang, Q., Sun, Z., Hu, W., Chen, M., Guo, L., & Qu, Y. (2019). Multi-view knowledge graph embedding for entity alignment. IJCAI International Joint Conference on Artificial Intelligence, 2019
- Zhang, R., Trisedya, B. D., Li, M., Jiang, Y., & Qi, J. (2022). A benchmark and comprehensive survey on knowledge graph entity alignment via representation learning. VLDB J.
- Zhu, Q., Wei, H., Sisman, B., Zheng, D., Faloutsos, C., Dong, X. L., & Han, J. (2020). Collective Multi-type Entity Alignment Between Knowledge Graphs. Proc. WWW ’20