Dutch-Belgian Database Day
– Past and Present of Entity Resolution

Keynote Information

date: Tuesday, 7 December 2021, 13:00-13:45

presenter: Ekaterini Ioannou http://www.eioannou.nl

Department of Management, University of Tilburg

- Introduction
- Time evolution of challenges
- Generations based on the challenges

  ↳ latest developments & directions

Presentation based on

- G. Papadakis, E. Ioannou, T. Palpanas: *Entity Resolution: Past, Present & Yet-to-Come,* EDBT, 2020. https://research.tilburguniversity.edu/en/projects/4ger
- G. Papadakis, E. Ioannou, E. Thanos, T. Palpanas: *The Four Generations of Entity Resolution*, Morgan & Claypool Publishers, 2021. https://entityresolution4g.com

Entities

people

companies

products

projects

events

...

...

locations

- Encode a large part of our knowledge
- Valuable asset for numerous current applications and (Web) systems

TILBURG UNIVERSITY

A. Matching, Linkage, Reconciliation, etc.

- Many names, descriptions, or IDs (URIs) are used for the same real-world objects
- Example:

London 런던 ލަންޑަން लंडन लंदन લંડન ለንደን ロンドン লন্ডন ລອນດອນ இலண்டன் ლონდონი Llundain Londain Londe Londen Londen Londen Londinium London Londona Londonas Londoni Londono Londra Londres Londrez Londyn Lontoo Loundres Luân Đôn Lunden Lundúnir Lunnainn Lunnon لندن لندن لندن لوندون לאנדאן לונדון Λονδίνο Лёндан Лондан Лондон Лондон Лондон Լոնդոն 伦敦 …

capital of UK, host city of the IV Olympic Games, host city of the XIV Olympic Games, future host of the XXX Olympic Games, city of the Westminster Abbey, city of the London Eye, the city described by Charles Dickens in his novels, …

http://sws.geonames.org/2643743/
http://en.wikipedia.org/wiki/London
http://dbpedia.org/resource/Category:London
…

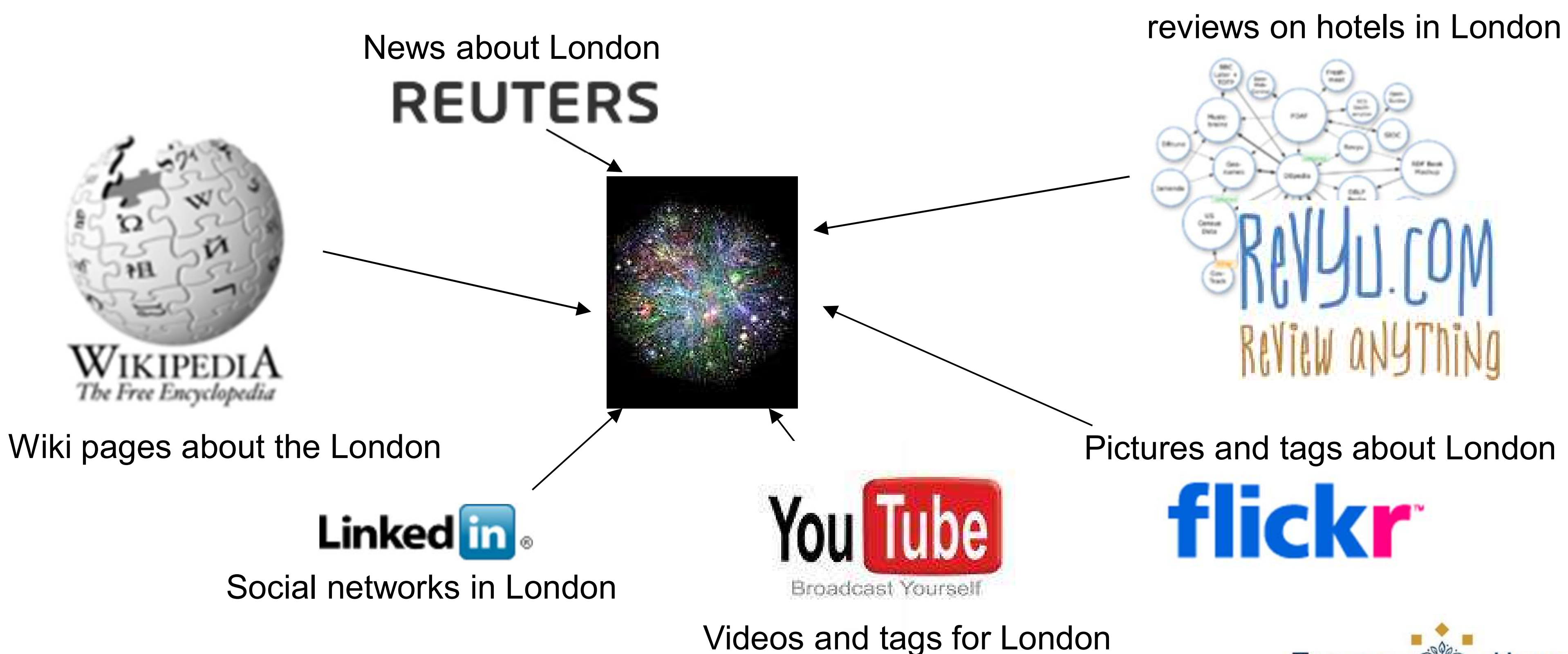
B. Disambiguation, Deduplication, etc.

- Plethora of different objects have the same name
- Examples:
  - London, KY
  - London, Laurel, KY
  - London, OH
  - London, Madison, OH
  - London, AR
  - London, Pope, AR
  - London, TX
  - London, Kimble, TX
  - London, MO

  - London, Jack
2612 Almes Dr
Montgomery, AL
(334) 272-7005
  - London, Jack R
2511 Winchester Rd
Montgomery, AL 36106-3327
(334) 272-7005
  - London, Jack
1222 Whitetail Trl
Van Buren, AR 72956-7368
(479) 474-4136
  - London, Jack
7400 Vista Del Mar Ave
La Jolla, CA 92037-4954
(858) 456-1850
  - …

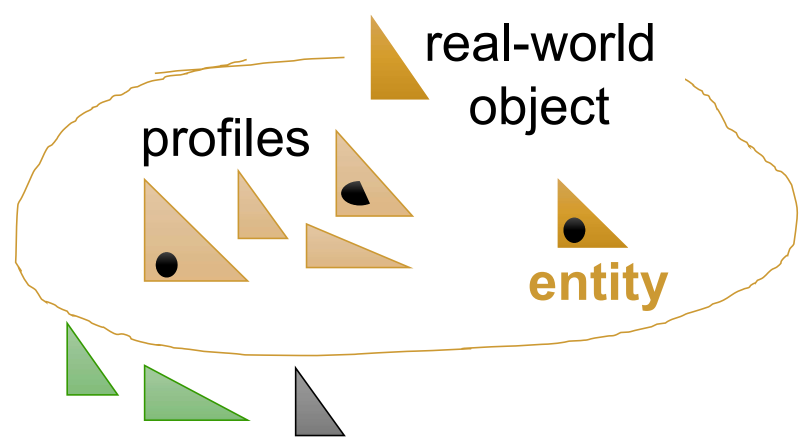
Today's situation

- Content providers offering valuable information describing (part of) real-world objects
- Information is useful for data integration, link discovery, query processing, searching, etc.



Entity Resolution

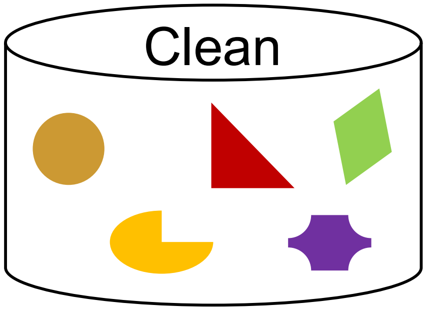
- Task that identifies and aggregates the different profiles that describe the same real-world objects [1, 2, 3, 4, 5]
- Primary usefulness:
  - Improves data quality and integrity
  - Fosters re-use of existing data sources
- Example application domains:
  - Linked Data
  - Building Knowledge Graphs
  - Census data
  - Price comparison portals



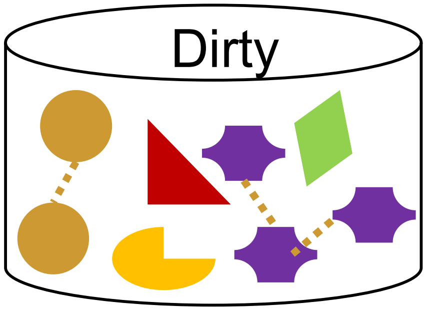
Types of Entity Resolution

- Data collections can be of two types:

  clean + dirty [3, 5, 6]



1. Clean:
   - Duplicate-free data
   - E.g., DBLP, ACM Digital Library, Wikipedia, Freebase

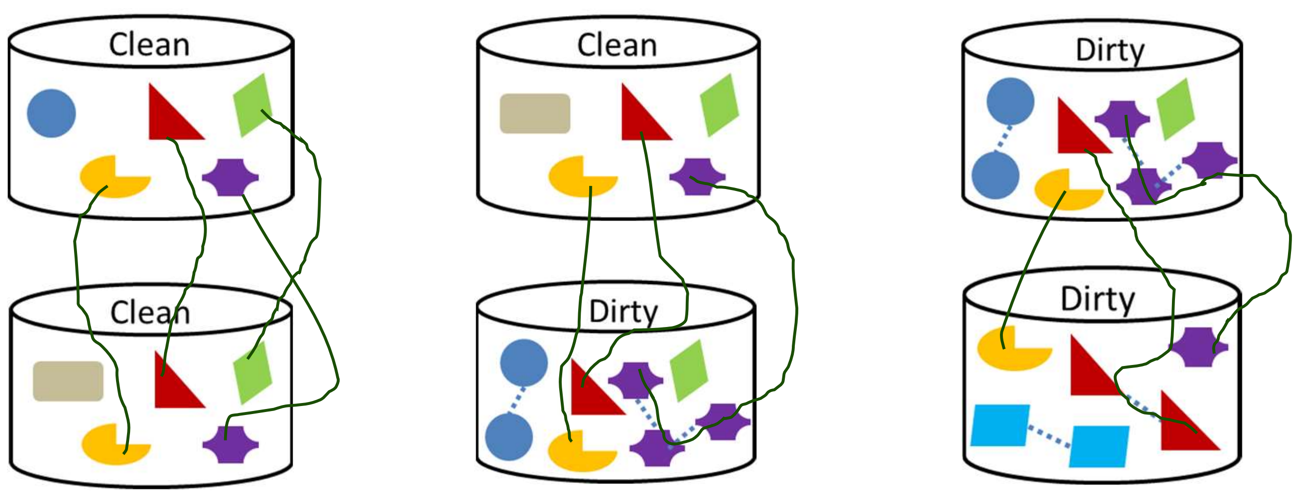


2. Dirty:
   - Contain duplicate profiles
   - E.g., Google Scholar, CiteseerX

Types of Entity Resolution

- Based on the quality of input, we distinguish entity resolution into 3 sub-tasks:
  1. Clean-Clean ER, a.k.a. *Record Linkage* in databases
  2. Dirty-Clean ER
  3. Dirty-Dirty ER

  (2 and 3) equivalent to *Dirty ER*, a.k.a. *Deduplication* in databases



References

1. X. L. Dong, D. Srivastava. Big Data Integration. Synthesis Lectures on Data Management, Morgan & Claypool Publishers 2015.
2. A. K. Elmagarmid, P. G. Ipeirotis, V. S. Verykios. Duplicate Record Detection: A Survey. TKDE 2007.
3. V. Christophides, V. Efthymiou, K. Stefanidis. Entity Resolution in the Web of Data. Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool Publishers 2015.
4. P. Christen. Data Matching – Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Data-Centric Systems and Applications 2012.
5. P. Christen. A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. TKDE 2012.
6. G. Papadakis, E. Ioannou, T. Palpanas. Entity Resolution: Past, Present and Yet-to-Come. EDBT 2020.

- Introduction
- Time evolution of challenges
- Generations based on the challenges

  ↳ latest developments & directions

Overview of Challenges!

title	A Blocking Framework for ER …
conference	TKDE 2013
author	George Papadakis
author	Ekaterini Ioannou
author	Themis Palpanas
author	Claudia Niederée
author	Wolfgang Nejdl

Overview of Challenges!

title	A Blocking Framework for ER …
conference	TKDE 2013
author	G. Papadakis
author	E. Ioannou
author	T. Palpanas
author	C. Niederée
author	W. Nejdl

title	A Blocking Framework for ER …
conference	TKDE 2013
author	George Papadakis
author	Ekaterini Ioannou
author	Themis Palpanas
author	Claudia Niederée
author	Wolfgang Nejdl

heterogeneity: abbreviations, misspellings, acronym, initials, etc.

Overview of Challenges!

title	A Blocking Framework for ER …
conference	TKDE 2013
author	G. Papadakis
author	E. Ioannou
author	T. Palpanas
author	C. Niederée
author	W. Nejdl

publication	A Blocking Framework for ER …
venue	TKDE 2013
researcher	George Papadakis
researcher	Ekaterini Ioannou
researcher	Themis Palpanas
researcher	Claudia Niederée
researcher	Wolfgang Nejdl

heterogeneity, schema variations

Overview of Challenges!

title	A Blocking Framework for ER …
conference	TKDE 2013
author	G. Papadakis
author	E. Ioannou
author	T. Palpanas
author	C. Niederée
author	W. Nejdl

?

title	A Blocking Framework for ER …
conference	TKDE 2013
author	George Papadakis
author	Ekaterini Ioannou
author	Themis Palpanas
author	Claudia Niederée
author	Wolfgang Nejdl

→ alternatives matches increase the belief

title	Entity Resolution: Past, Present and Yet-to-Come
conference	EDBT 2020
author	George Papadakis
author	Ekaterini Ioannou
author	Themis Palpanas

heterogeneity, schema variations, collective

Overview of Challenges!

title	A Blocking Framework for ER …
conference	TKDE 2013
author	G. Papadakis
author	E. Ioannou
author	T. Palpanas
author	C. Niederée
author	W. Nejdl

title	A Blocking Framework for ER …
conference	TKDE 2013
author	George Papadakis
author	Ekaterini Ioannou
author	Themis Palpanas
author	Claudia Niederée
author	Wolfgang Nejdl

→ Propagate information of a detected match

title	Entity Resolution: Past, Present and Yet-to-Come
conference	EDBT 2020
author	George Papadakis
author	Ekaterini Ioannou
author	Themis Palpanas

heterogeneity, schema variations, collective

Overview of Challenges!

A Blocking Framework for ER …
TKDE 2013
Papadakis
George
Ioannou
Ekaterini
Ioannou
Palpanas
Themis
Niederée
Claudia
Nejdl
Wolfgang

Entity Resolution: Past, Present and Yet-to-Come
EDBT 2020
G. Papadakis
E. Ioannou
T. Palpanas

heterogeneity, schema variations, collective, unstructured

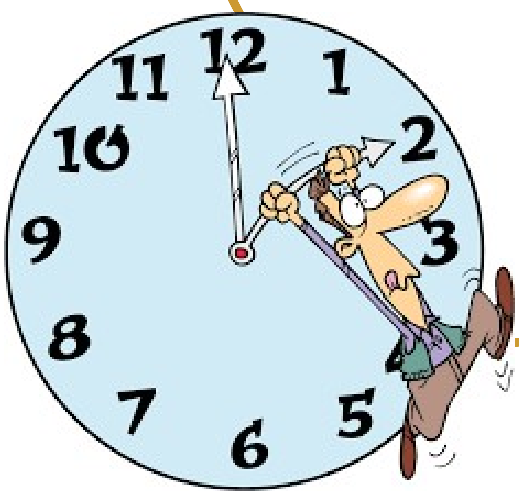
Overview of Challenges!

Jacqueline Lee Bouvier

…

Jacqueline Kennedy

…



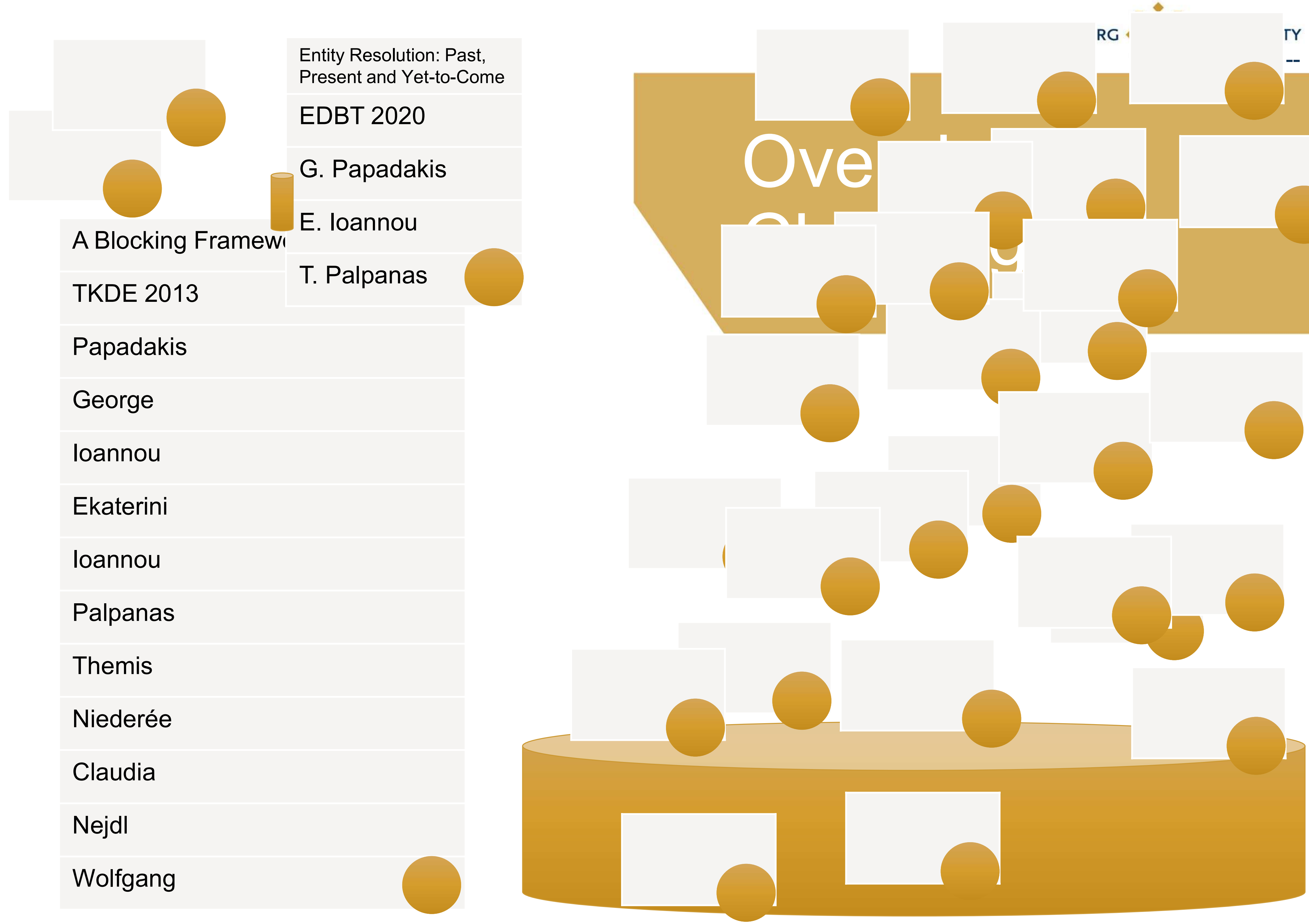
First Lady of the United States

…

Jackie Onassis

…

heterogeneity, schema variations, collective, unstructured, volatility

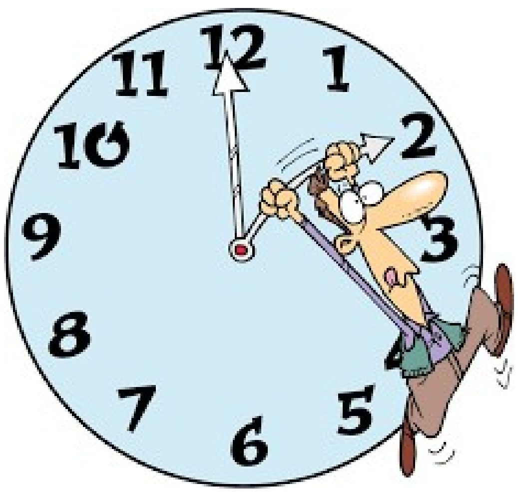


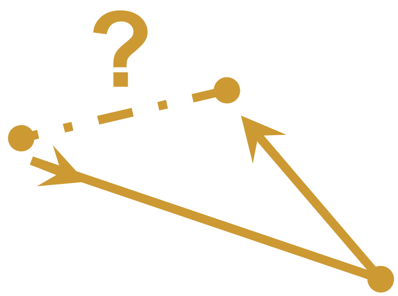
heterogeneity, schema variations, collective, unstructured, volatility, Big collections, …

Overview of Challenges!

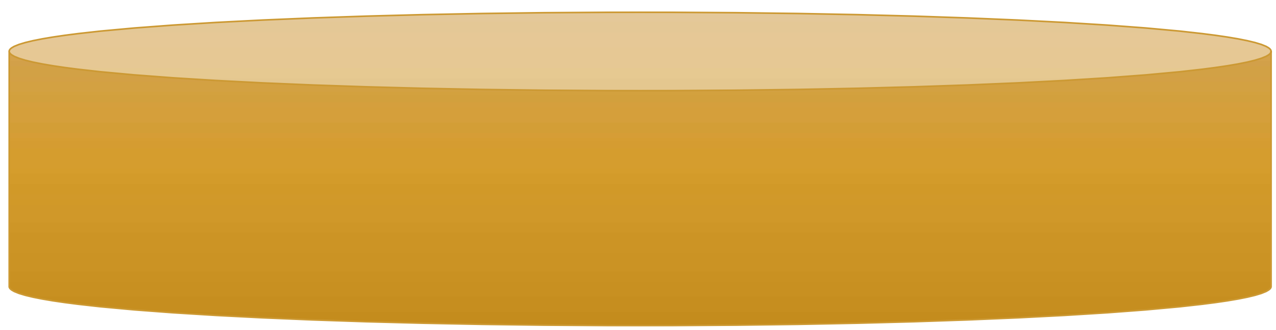
title	A Blocking Framewo
venue	TKDE 2013
author	George Papadakis
author	Ekaterini Ioannou
author	Themis Palpanas
author	Claudia Niederée
author	Wolfgang Nejdl

Entity Resolution: Past, Present and Yet-to-Come
EDBT 2020
G. Papadakis
E. Ioannou
T. Palpanas





D
A
T
A



heterogeneity, schema variations, collective, unstructured, volatility, Big collection, . . .

Challenges

- ER challenges arise from the application settings
- Examples:
  - Data characteristics
  - System and resources
  - Time restrictions
  - …
- Evolving nature of the application settings implies:
  - Constant modification of the challenges
  - Plethora of resolution methods

- Introduction
- Time evolution of challenges
- Generations based on the challenges

  ↳ latest developments & directions

additional material at the end of the presentation

Tilburg University

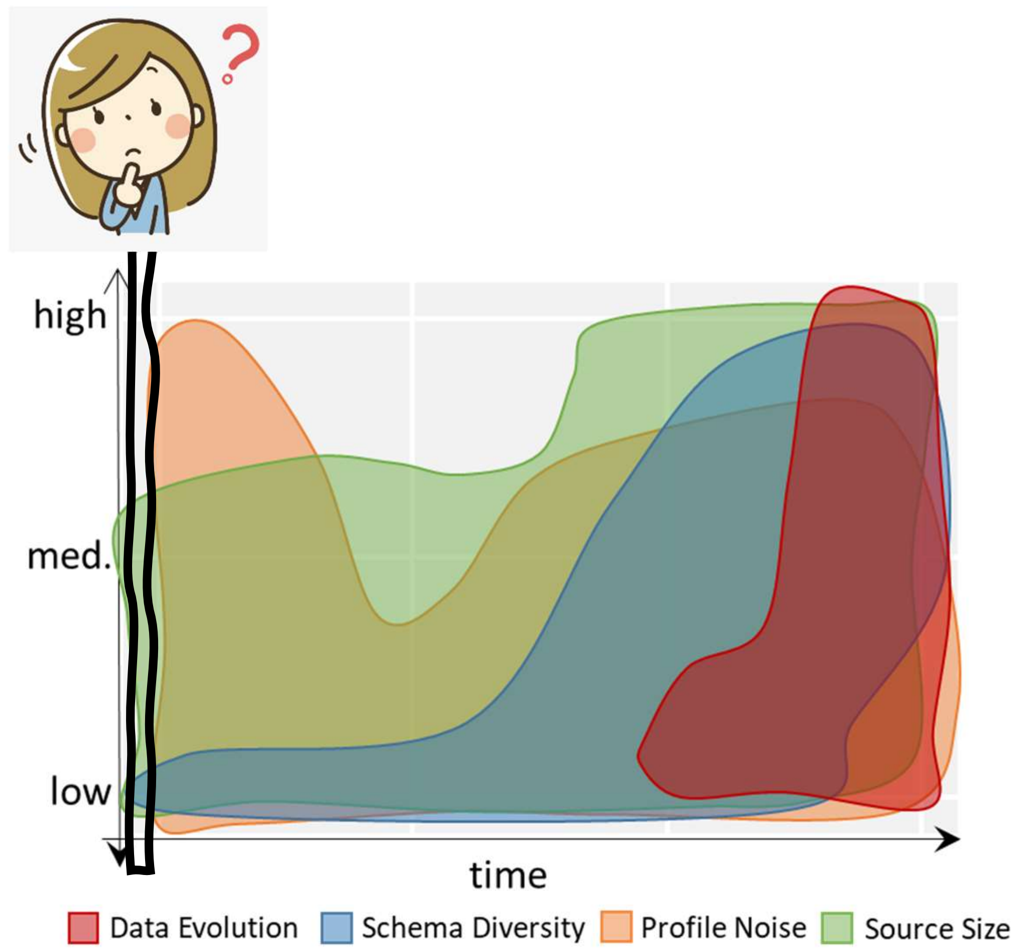
Veracity

+Volume

+Variety

+Velocity

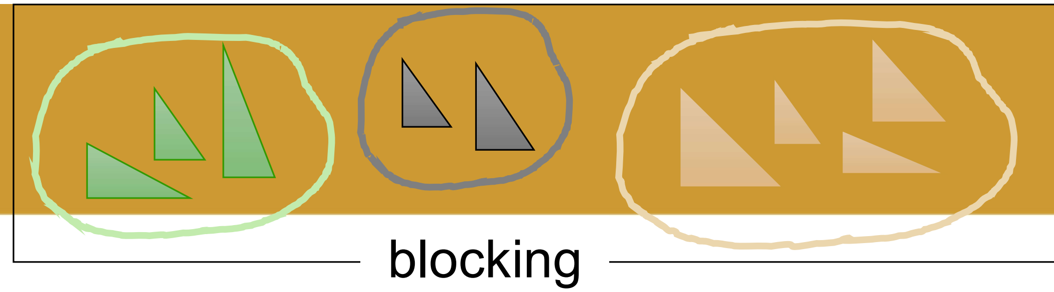
ER **Methods**

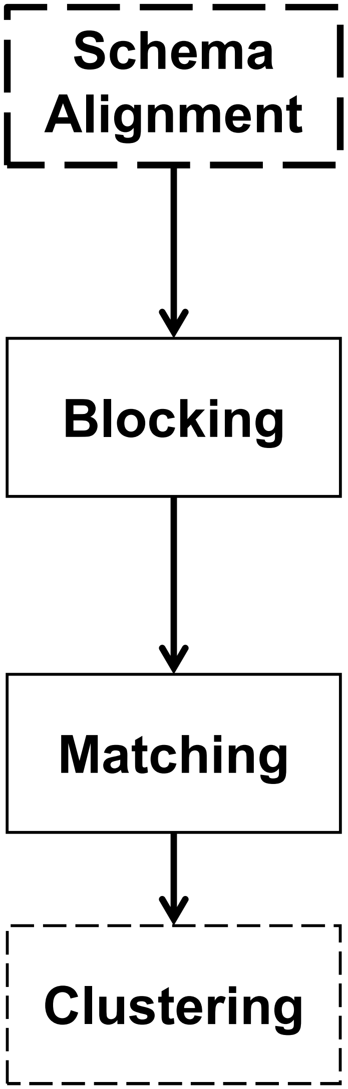


Earliest Resolution Methods (Generation 1)

- Scope → Veracity
  - Structured data with known semantics and quality, e.g., small relational databases
  - Dealing with high levels of profile noise
- Goal:
  - Achieve high accuracy despite inconsistencies, noise, or errors in profiles
- Assumption is a “known schema”

Generation 1

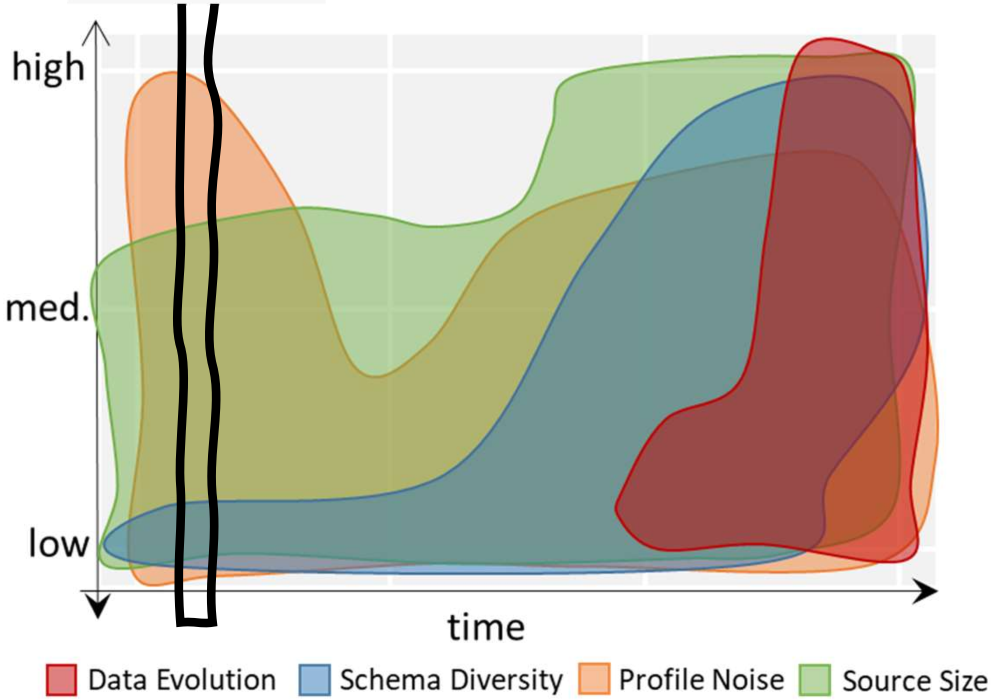




- Creating mappings between equivalent attributes of the two schemata, e.g., profession ≡ job
- Grouping similar profiles into blocks
  - All profiles in one block might be the same
  - Profiles of different blocks cannot be the same
- Estimating the similarity among the candidate matches
- Partitioning the matched pairs into equivalence clusters, i.e., sets describing the same real-world object

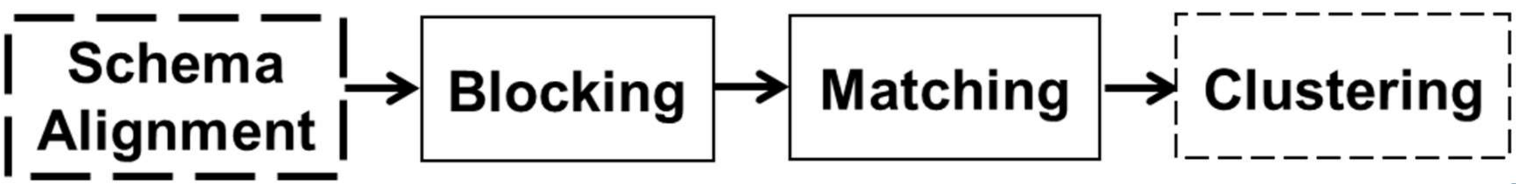
Veracity
+Volume
+Variety
+Velocity

ER Methods



Generation 2

- Scope → Volume & Veracity
  - (tens of) millions of structured profiles
- Goals:
  - High accuracy despite noise
  - High time efficiency despite the size of data
- Assumptions:
  - Known schema → custom, schema-based solutions
- Workflow remains the same

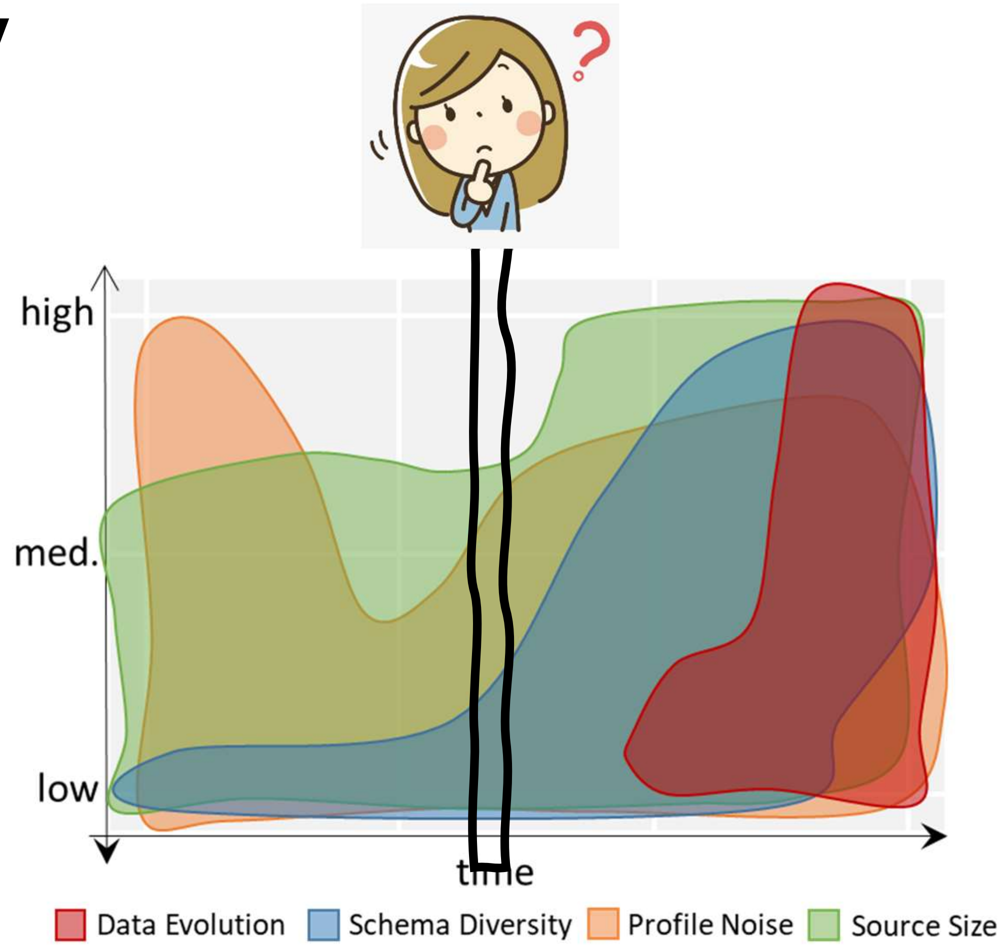


Solution: Parallelization

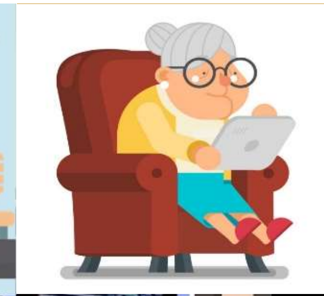
- Type A :: Multi-core parallelization
  - Single system → shared memory
  - Distribute processing among available CPUs
- Type B :: Massive parallelization
  - Cluster of independent systems
  - Map-Reduce paradigm [1]
    - Data partitioned across the nodes of a cluster
    - Map Phase: transforms a data partition into (key, value) pairs
    - Reduce Phase: processes pairs with the same key
- Additional material (end of presentation)
  - Parallelization method for each workflow step

Veracity
+Volume
+Variety
+Velocity

ER Methods



Generation 3



Generation 3

Scope → Variety & Volume & Veracity

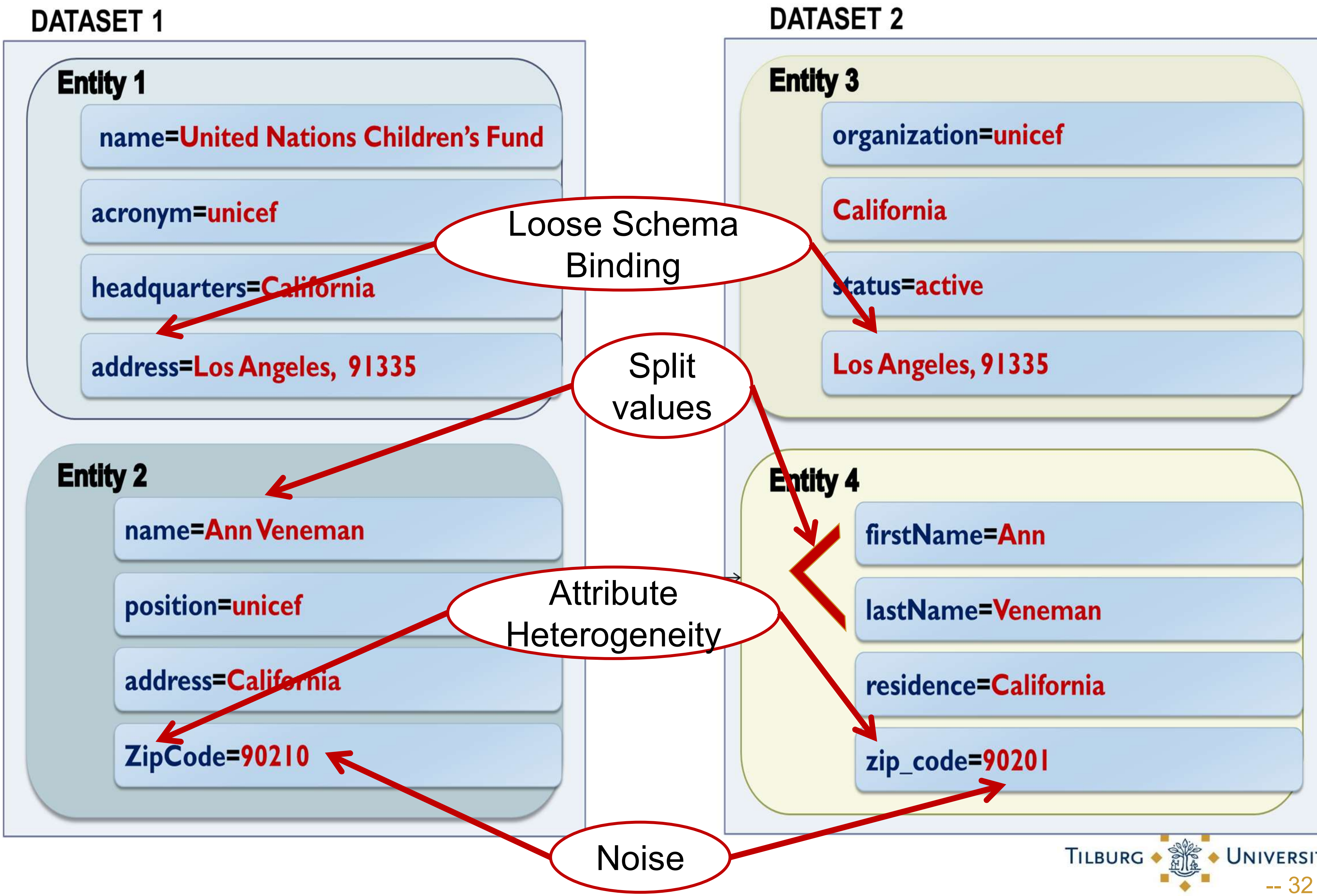
- User-generated Web Data
- Users are free to add attribute values and/or attribute names

  → unprecedented levels of schema heterogeneity

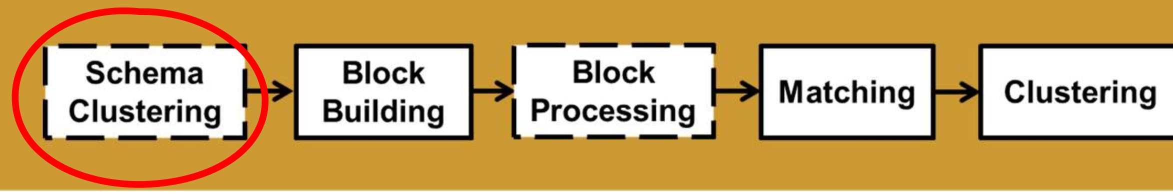
  - Google Base: 100,000 schemata for 10,000 profile types
  - BTC09:136,000 attribute names
- Voluminous, (semi-)structured datasets
  - BTC09: 1.15 billion triples, 182 million profiles
- Several datasets produced by automatic information extraction techniques → noise, tag-style values

(BTC: billion triple challenge database)

Example of Web Data

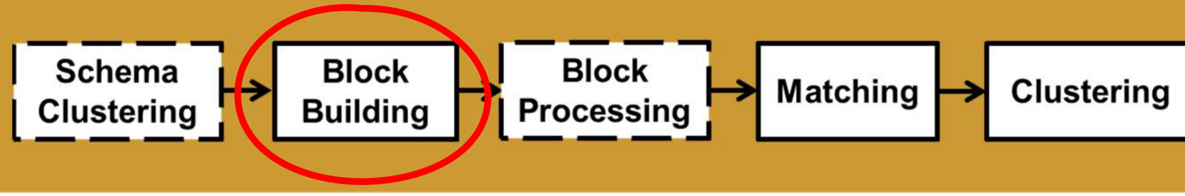


Schema Clustering



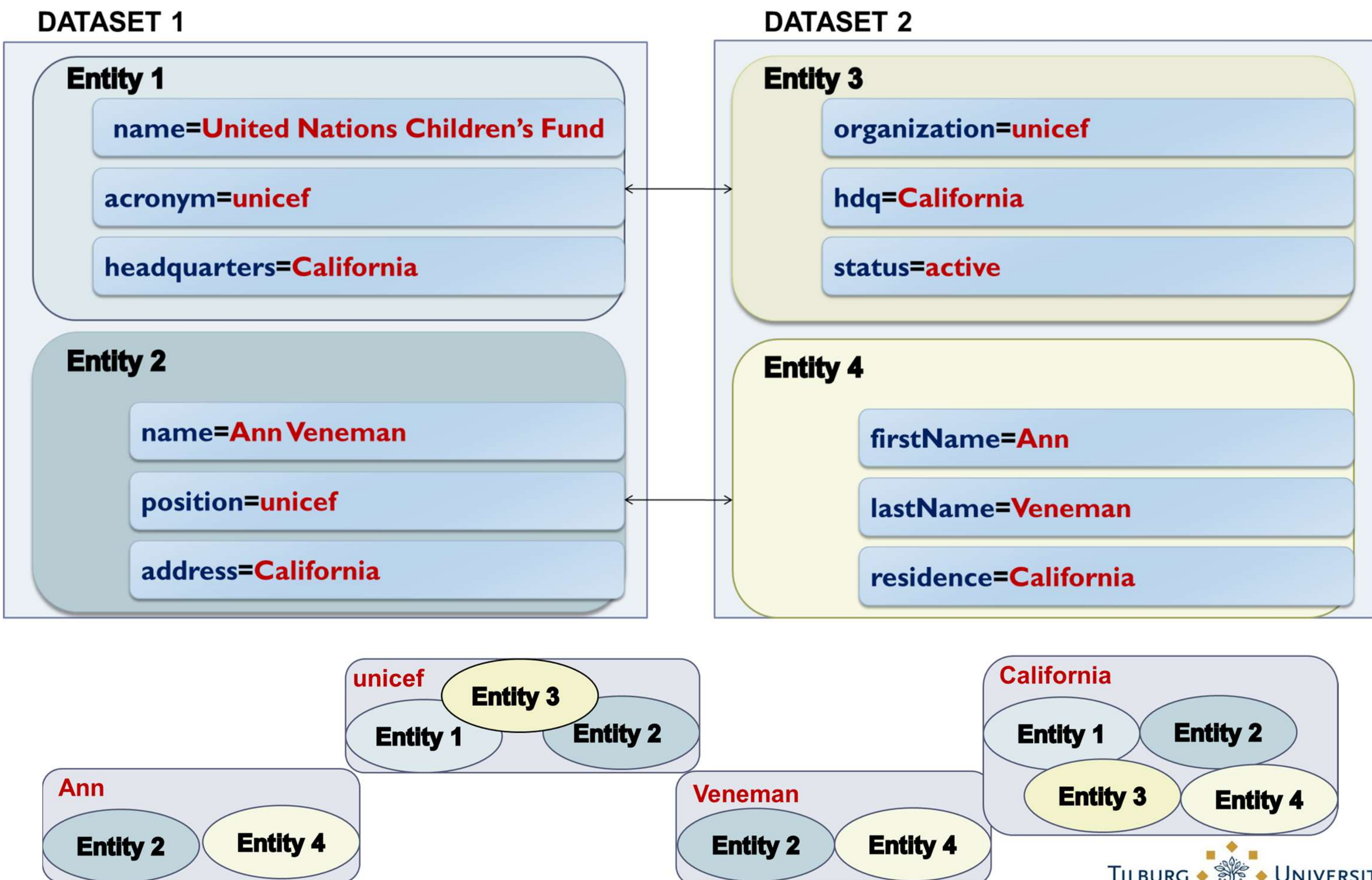
- Schema Matching → not applicable (too many alternatives)
- Instead, partition attributes according to their syntactic similarity, regardless of their semantic relation
- Goal: Facilitate next steps
- Both Clean-Clean and Dirty ER
- Attribute clustering using graphs

Block Building

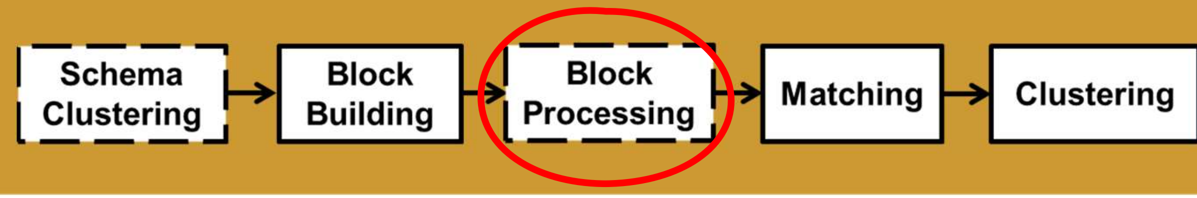


- Considers all attribute values and completely ignores all attribute names → schema-agnostic functionality
- Core approach: Token Blocking
  1. Given a profile, extract all tokens that are contained in its attribute values
  2. Create one block for every distinct token with frequency > 2 → each block contains all profiles with the corresponding token
     Pros:
     - Parameter-free
     - Efficient
     - Unsupervised

Example of Token Blocking

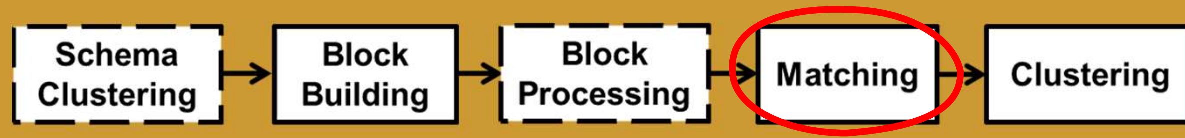


Block Processing



- Block Building creates a huge number of blocks
- Results in an additional step in the workflow
  - Goal: restructure the original blocks in order to increase precision at no significant cost in recall
- Focus on reducing / removing comparisons:
  1. Redundant comparisons, i.e., comparing profiles that were already compared in a previous block
  2. Superfluous comparisons, i.e., high number of comparisons between irrelevant profiles

Entity Matching



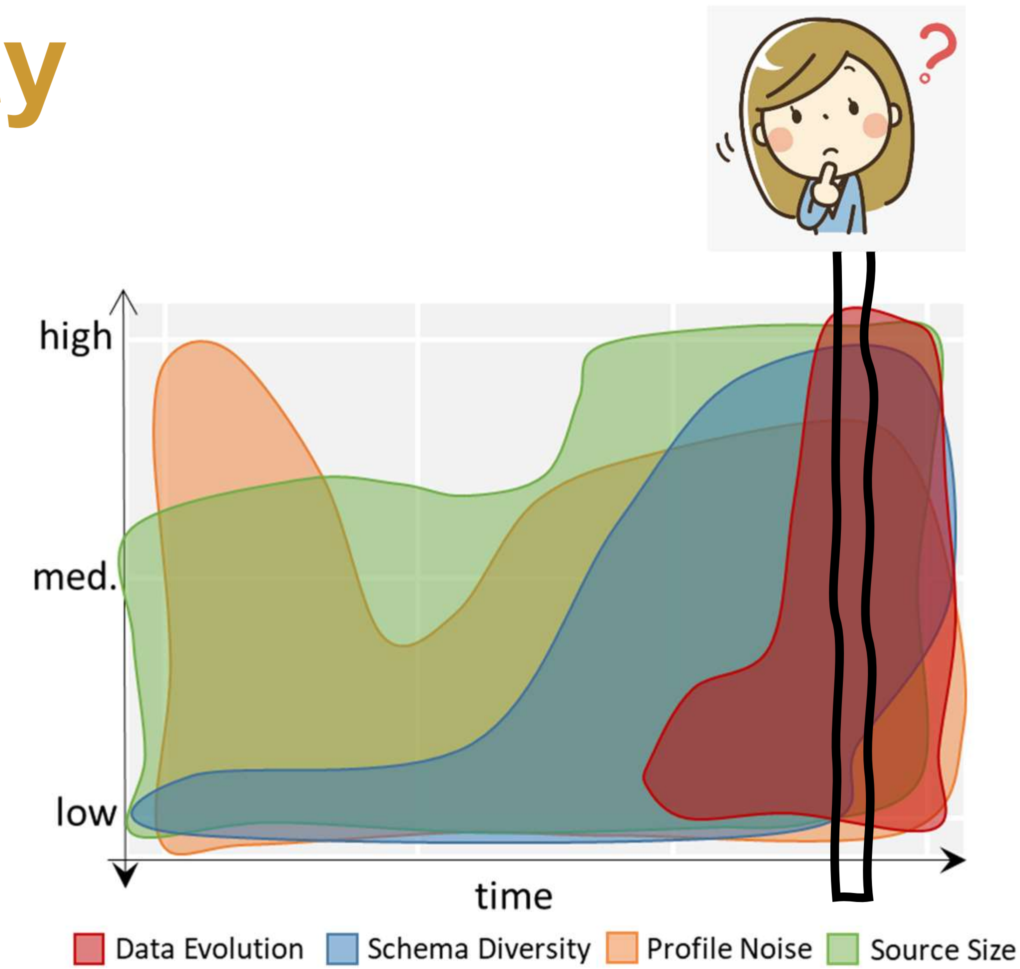
- Collective approaches to tackle Variety
- Most methods crafted for Clean-Clean ER
- General outline:
  - Iterative process starts with a few reliable seed matches
  - Propagate initial matches to neighbors
  - Order candidate matches in descending overall similarity
  - Recompute the similarity of the neighbors
  - Update candidate matches order
- Alternative is to perform a specific number of steps, rather than iterating until convergence

Entity Clustering

- Methods from previous generations are still applicable
- Only difference:
  - Similarity scores extracted in a schema-agnostic fashion, not from specific attributes

Veracity
+Volume
+Variety
+Velocity

ER Methods

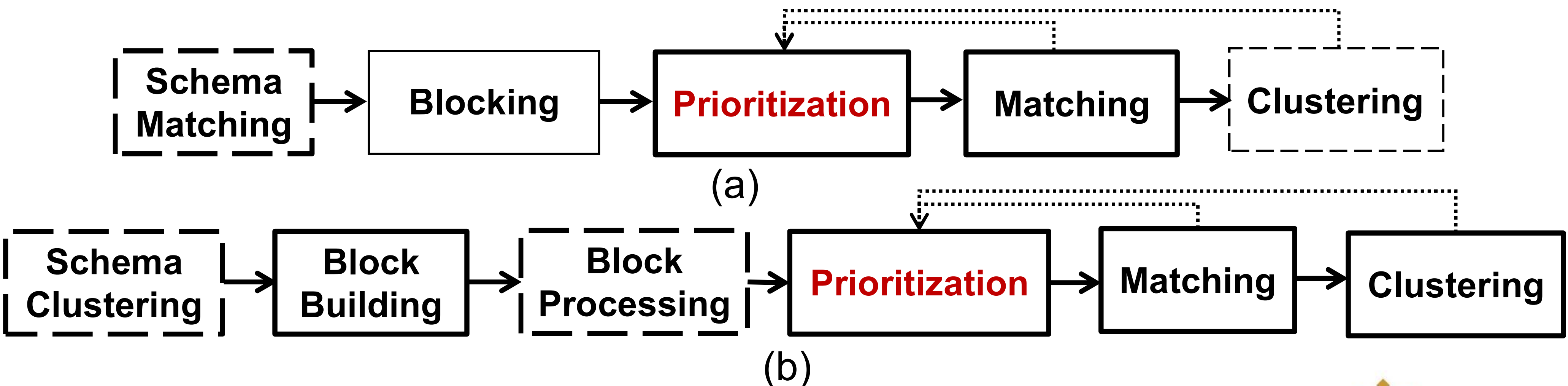


Generation 4

Scope → Velocity & Variety & Volume & Veracity

- Applications with increasing data volume & time constraints
  - Loose ones (e.g., minutes, hours) ⟶ Progressive ER
  - Strict ones (i.e., seconds) ⟶ Real-time (On-line) ER

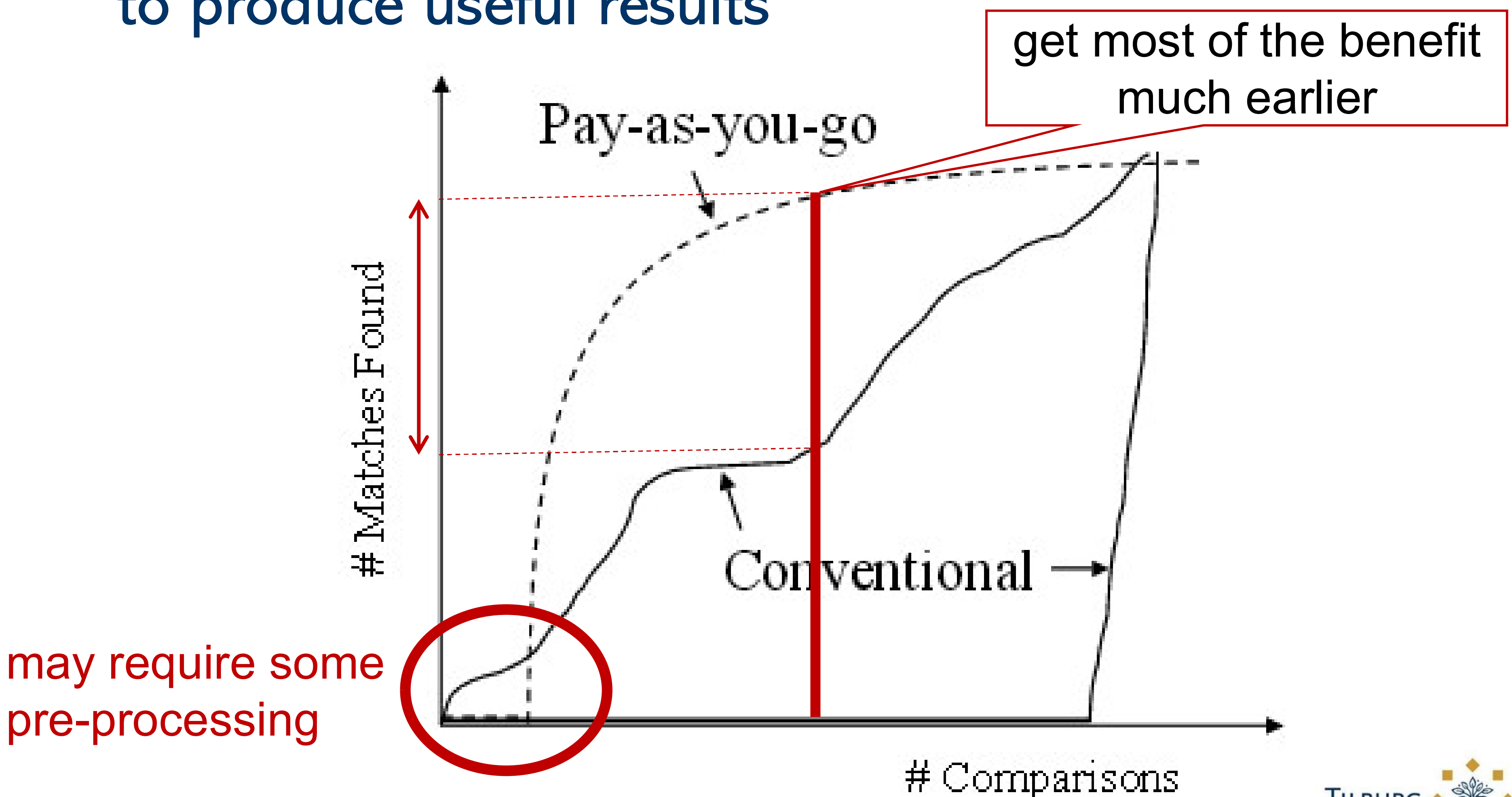
- End-to-end workflows for Progressive ER:



Progressive Entity Resolution

- Unprecedented, increasing volume of data

→ applications can compromise with partial solutions to produce useful results

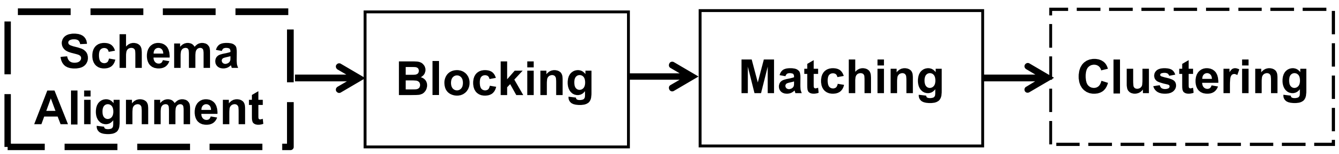


Outline Progressive ER

- Requires:
  - Improved early quality
  - Same eventual quality
- Prioritization:
  - Defines optimal processing order for a set of entities
  - Static methods [1, 2]:
    - Guide which records to compare first (independently of ER matching results)
  - Dynamic methods [3]:
    - If a duplicate is found, then check neighbors as well
    - Assumption: oracle for entity matching

Real-time (also Query-based, Online, Incremental)

Same workflow as original two workflows:



Different goal:

- resolve each query over a large dataset in the shorted possible time (and with the minimum memory footprint)

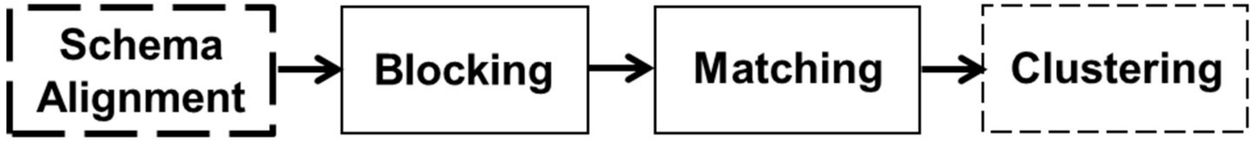
Same scope (so far):

- structured data

Different input:

- stream of query profiles

Techniques per workflow step



Incremental Blocking

- Maintain a dynamic set of blocks
- I.e., block contents are updated as new profiles arrive
- Examples: DySimII [1], F-DySNI [2, 3], (S)BlockSketch [4]

Incremental Matching

- QDA [5] - SQL-like selection queries over a single dataset
- QuERy [6] - complex join queries over multiple, overlapping, dirty DSs
- Simple or statistical queries over possible worlds (i.e., resolutions) [7, 8]
- Evolving matching rules [9]

Incremental Clustering

- Maintain the entities detected by clustering
- Examples: Incremental Correlation Clustering [10]

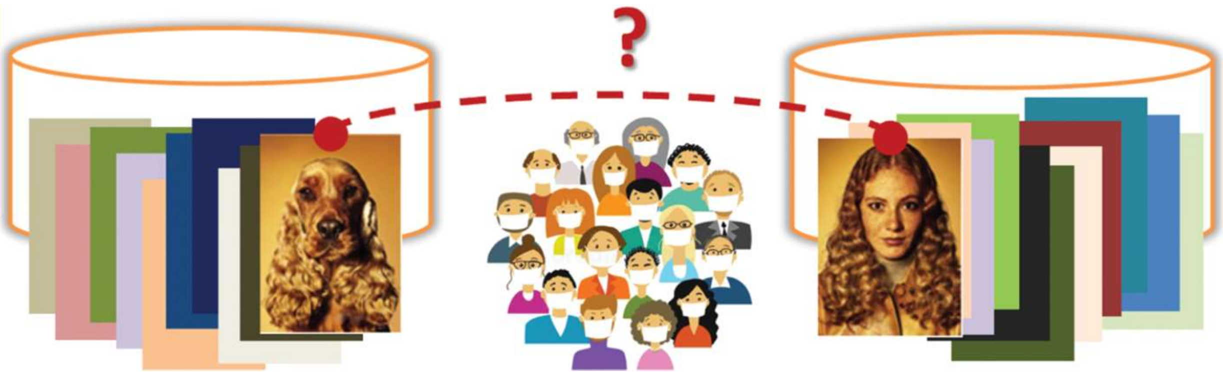
Progressive ER References

1. S. E. Whang, D. Marmaros, and H. Garcia-Molina. Pay-as-you-go entity resolution. TKDE, 25(5):1111–1124, 2013.
2. T. Papenbrock, A. Heise, and F. Naumann. Progressive duplicate detection. TKDE, 27(5):1316–1329, 2015.
3. G. Simonini, G. Papadakis, T. Palpanas, S. Bergamaschi. Schema-Agnostic Progressive Entity Resolution. IEEE Trans. Knowl. Data Eng. 31(6): 1208-1221 (2019)
4. Y. Altowim and S. Mehrotra. Parallel progressive approach to entity resolution using mapreduce. In ICDE, pages 909–920, 2017.

Incremental ER References

1. B. Ramadan and P. Christen, H. Liang, and R. W. Gayler, and D. Hawking. Dynamic similarity-aware inverted indexing for real-time entity resolution. In PAKDD Workshops, pages 47–58, 2013.
2. B. Ramadan and P. Christen. Forest-based dynamic sorted neighborhood indexing for real-time entity resolution. In CIKM, pages 1787–1790, 2014.
3. B. Ramadan and P. Christen, H. Liang, and R. W. Gayler. Dynamic sorted neighborhood indexing for real-time entity resolution. J. Data and Information Quality, 6(4):15:1–15:29, 2015.
4. D. Karapiperis, A. Gkoulalas-Divanis, V. S. Verykios. Summarization Algorithms for Record Linkage. EDBT 2018: 73-84.
5. H. Altwaijry, D. V. Kalashnikov, and S. Mehrotra. QDA: A query-driven approach to entity resolution. TKDE, 29(2):402–417, 2017.
6. H. Altwaijry, S. Mehrotra, and D. V. Kalashnikov. Query: A framework for integrating entity resolution with query processing. PVLDB, 9(3):120–131, 2015.
7. E. Ioannou, W. Nejdl, C. Niederée, and Y. Velegrakis. On-the-fly entity-aware query processing in the presence of linkage. PVLDB, 3(1): 429–438, 2010.
8. E. Ioannou, and M. Garofalakis. Query Analytics over Probabilistic Databases with Unmerged Duplicates. TKDE, 27(8):2245-2260, 2015.
9. S. E. Whang and H. Garcia-Molina. Entity resolution with evolving rules. PVLDB, 3(1):1326–1337, 2010.
10. A. Gruenheid, X. L. Dong, and D. Srivastava. Incremental record linkage. Proc. VLDB Endow., 7(9):697–708, May 2014. ISSN 2150-8097.

Last 1-2 years



- Crowdsourcing
  - E.g.: how do maximize accuracy while minimize cost?
- Deep Learning
- Explainability for matches and non-matches
- Advance solutions for evolving data
  - E.g., automatic configuration of workflows
- Algorithmic Bias
- Benchmarks
- …

Questions

– additional material & citations

ER Methods

ER Challenges

Veracity

- Structured data with known semantics and quality
- Dealing with high levels of profile noise

+ Volume

- Very large number of profiles

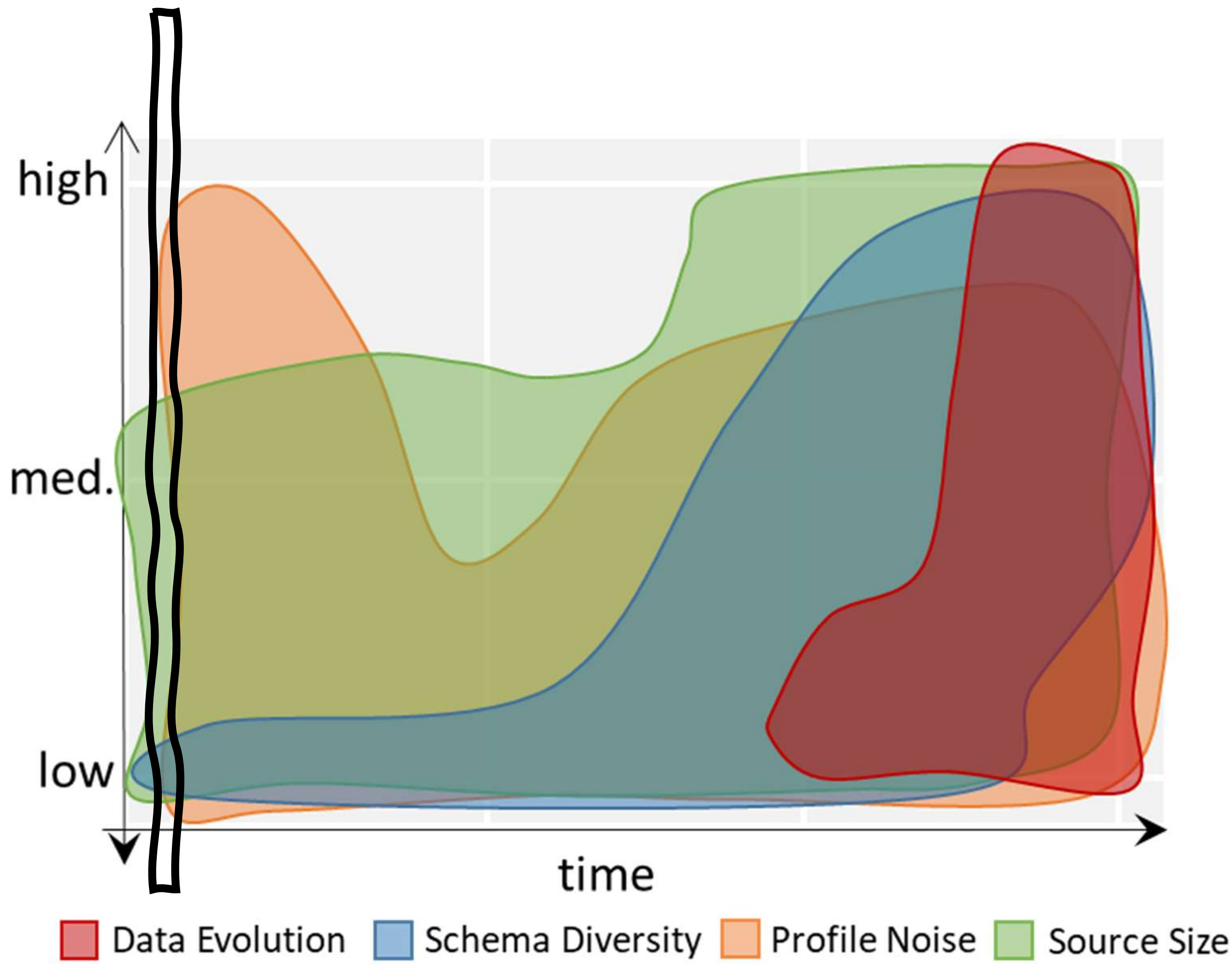
+ Variety

- Large volumes of semi-structured, unstructured or highly heterogeneous structured data

+ Velocity

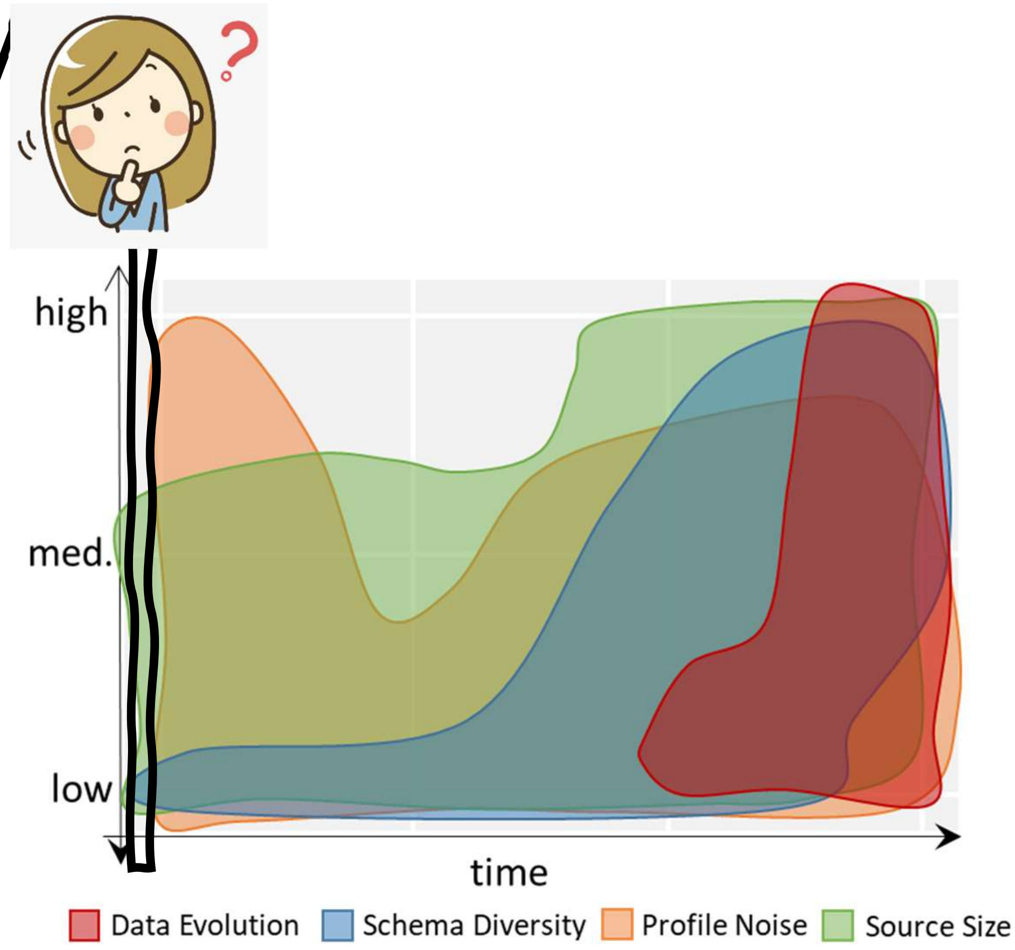
- Increasing volume of available data

ER Challenges in time



Veracity
+Volume
+Variety
+Velocity

ER **Methods**



Veracity

Schema Alignment → Blocking → Matching → Clustering

- Earliest ER methods
- Scope:
  - Structured data (e.g., small relational databases)
- Goal:
  - Achieve high accuracy despite inconsistencies, noise, or errors in profiles
- Assumptions:
  - Known schema → custom, schema-based solutions

Step 1: Schema Alignment

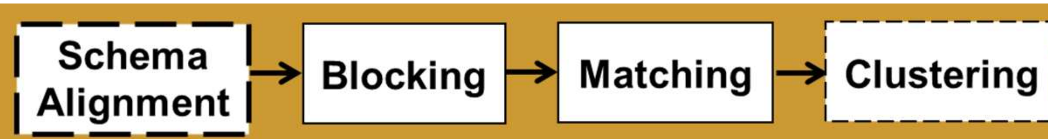
- Scope:
  - Record Linkage
- Goal:
  - Create mappings between equivalent attributes of the two schemata, e.g., *profession* ≡ *job*
- Types of Solutions:
  - Instance-based: use profiles to learn mappings, transformations, rules (e.g., merge of two values)
  - Structure-based: use schema information by converting it to trees or graphs
  - Hybrid

Step 1: Schema Alignment

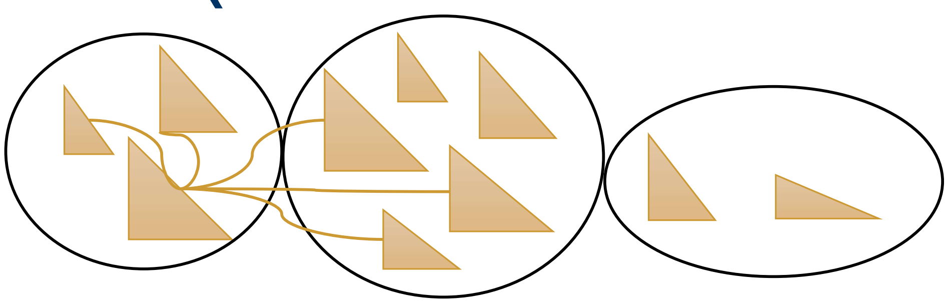
- Taxonomy of Main Schema Matching Methods (in chronological order)

Method	Category	Type of Evidence
Cupid [1]	Structure-based	Name similarity, Constraints, Contextual similarity
Similarity Flooding [2]	Structure-based	Name similarity, Contextual similarity
COMA [3]	Hybrid	Name similarity, Constraints, Contextual similarity
Distribution-based [4]	Instance-based	Value distribution

Step 2: Blocking



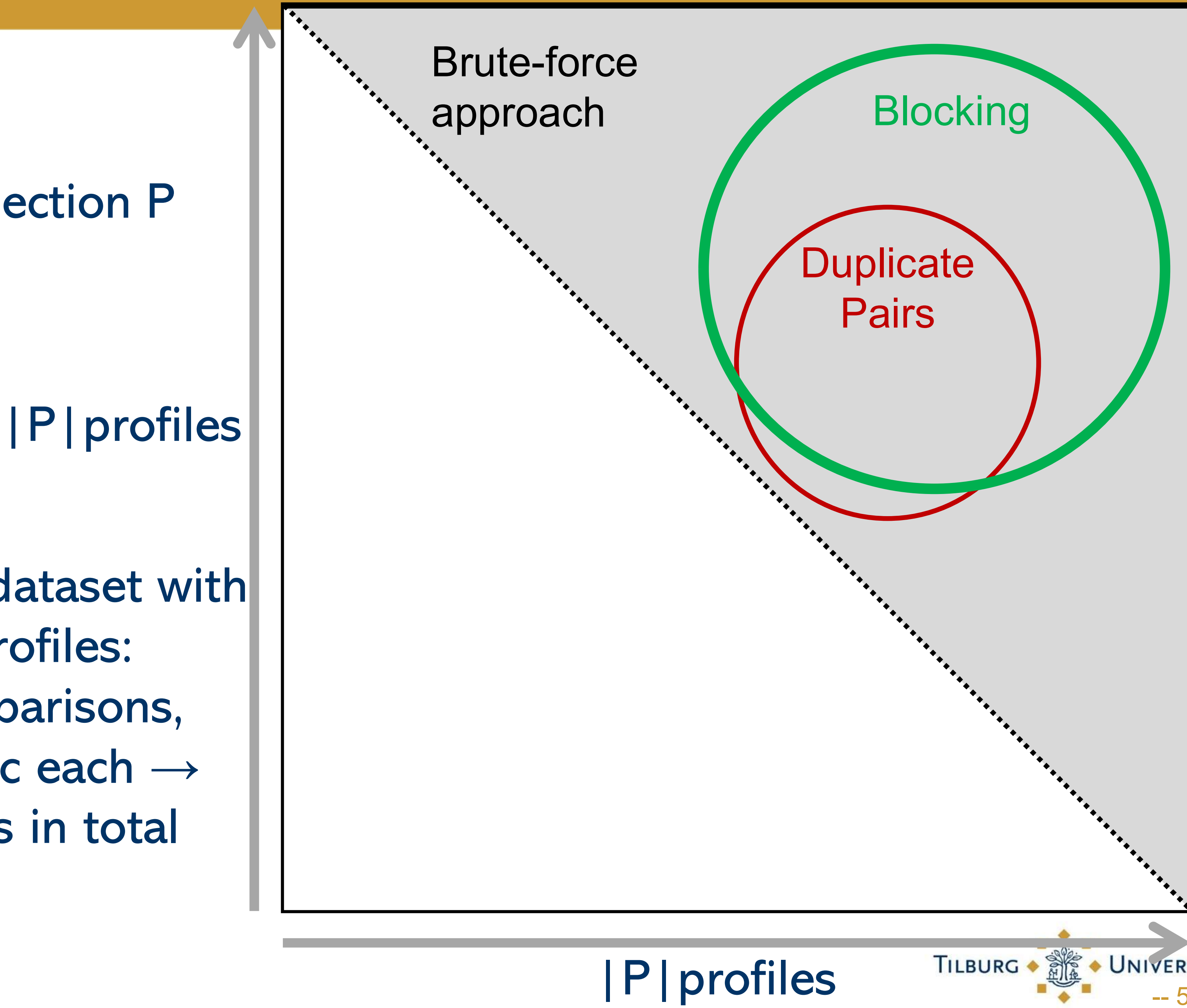
- Scope:
  - Both Deduplication and Record Linkage
- Goal:
  - Entity resolution is an inherently quadratic problem, i.e., $O(n^2)$ with every profile compared to all others
  - Blocking groups similar profiles into blocks
    - Comparisons executed only inside each block
    - Complexity is now quadratic to the size of the block (much smaller than dataset size!)



Computational cost

Input:
Profile Collection P

E.g.: For a dataset with 100,000 profiles: ~10^{10} comparisons, If 0.05 msec each → >100 hours in total



General Principles of Blocking

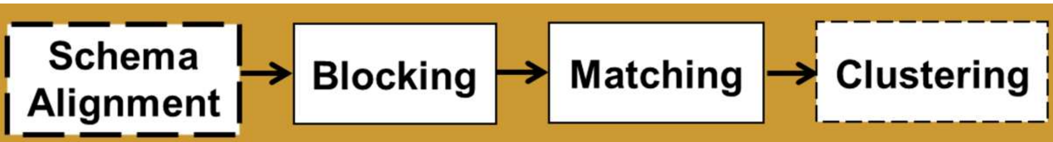
1. Represent each profile by *one or more* signatures called blocking keys
   - Focus on string values
2. Place into blocks all profiles having the *same or similar* blocking key
3. Two matching profiles can be detected as long as they co-occur in at least one block
   - Trade-off between recall and precision!

name	Ekaterini Ioannou
zip code	9876
address	…
..	…

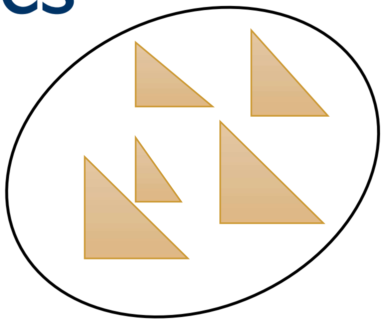
Taxonomy of Blocking Methods [1]

Method	Key Type	Matching awareness
Standard Blocking [2]	Hash-based	Static
Suffix Arrays [3] + [4,5]	Hash-based	Static
Q-grams Blocking [6] + [4]	Hash-based	Static
MFIBlocks [7]	Hash-based	Static
Sorted Neighborhood [9] + [4,10]	Sort-based	Static
Duplicate Count Strategy [11]	Sort-based	Dynamic
Sorted Blocks [12]	Hybrid	Static
ApproxDNF [13]	Hash-based	Static
Blocking Scheme Learner [14]	Hash-based	Static
CBlock [15]	Hash-based	Static
FisherDisjunctive [16]	Hash-based	Static

Step 3: Matching

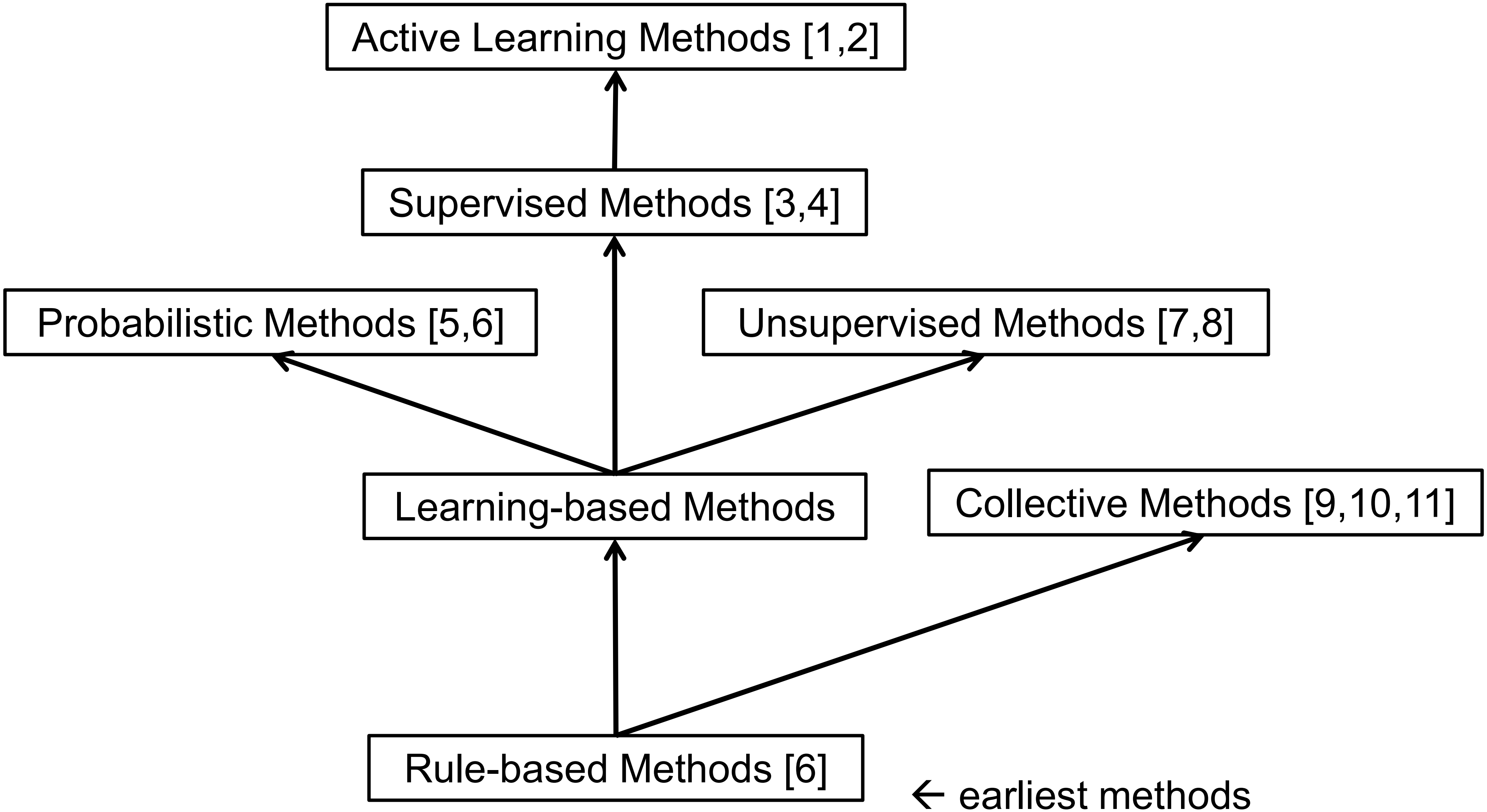


- Estimates the similarity of candidate matches



- Input: a set of blocks
  - Every distinct comparison in any block is a candidate match
- Output: similarity Graph
  - Nodes → profiles
  - Edges → candidate matches
  - Edge weights → matching likelihood (based on similarity score)

Evolution of Matching



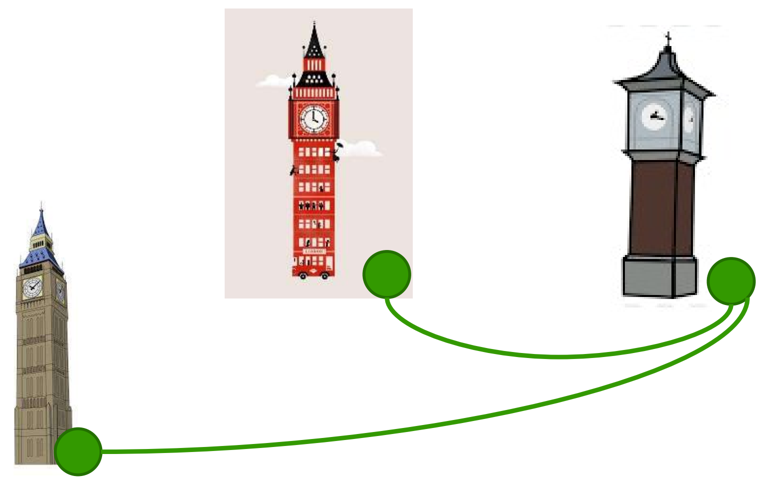
All are heavily based on string similarity measures [6]

Step 4: Clustering

Schema Alignment → Blocking → Matching → Clustering

- Partitions the matched pairs into equivalence clusters

  i.e., groups of profiles describing the same real-world object
- Input
  - Similarity Graph:
    - Nodes → profiles
    - Edges → candidate matches
    - Edge weights → matching likelihood (based on similarity score)



- Output
  - Equivalence Clusters

Clustering Algorithms for Record Linkage

Relies on 1-1 constraint

- One profile from source dataset matches one profile from the target dataset

1. Unique Mapping Clustering [1, 2]
   - Sorts all edges in decreasing weight
   - Starting from the top, each edge corresponds to a pair of duplicates, if:
     - None of the adjacent profiles have already been matched to other profiles, or
     - Predefined threshold < edge weight

Clustering Algorithms for Record Linkage

Relies on 1-1 constraint

- One profile from source dataset matches one profile from the target dataset

1. Unique Mapping Clustering
2. Row-Column Clustering [3]
   - Based on an efficient approximation of the Hungarian Algorithm

Clustering Algorithms for Record Linkage

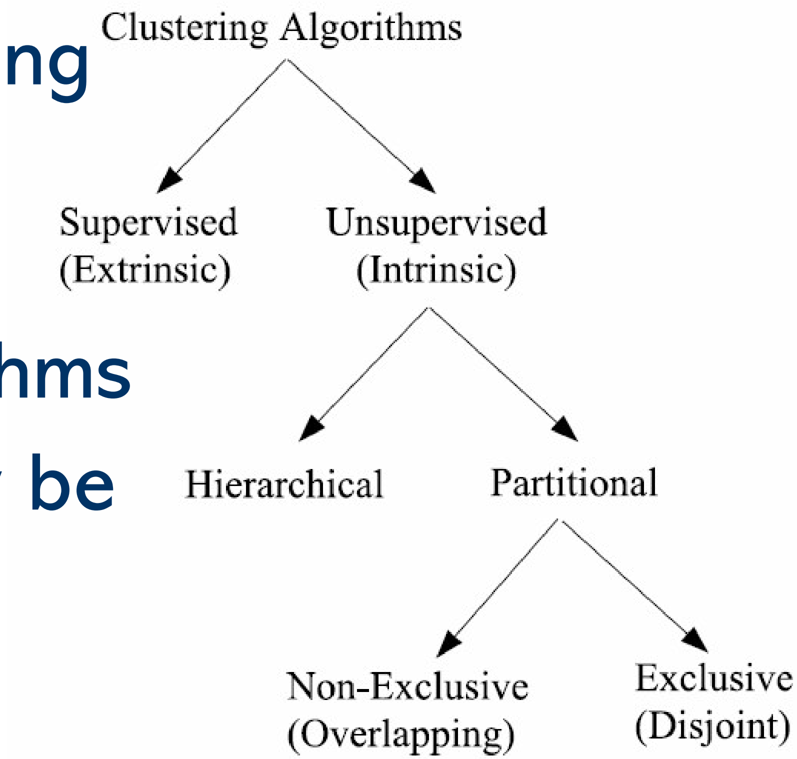
Relies on 1-1 constraint

- One profile from source dataset matches one profile from the target dataset

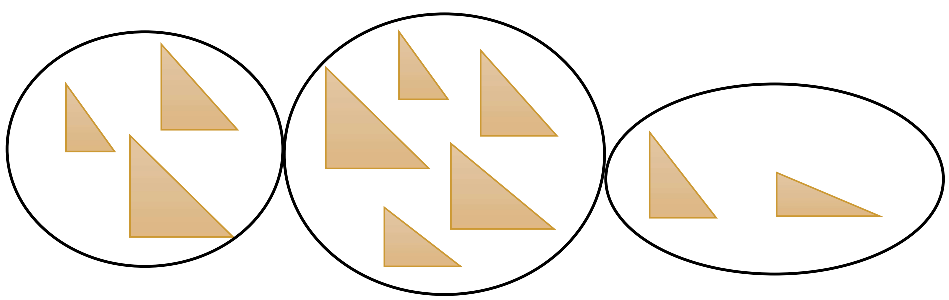
1. Unique Mapping Clustering
2. Row-Column Clustering
3. Best Assignment Clustering [4]
   - Efficient, heuristic solution to the assignment problem in unbalanced bipartite graphs

Clustering Algorithms for Deduplication

- A wealth of literature on clustering algorithms
- Requirements:
  - Partitional and disjoint algorithms
    - Sometimes overlapping may be desirable
  - Goal: Sets of clusters that
    - maximize the intra-cluster weights
    - minimize the inter-cluster edge weights



Classification of clustering methods [6]



(slide from O. Hassanzadeh)

Dirty ER Clustering Algorithms Characteristics [3]

- Most important feature "unconstrained algorithms", i.e.,
  - Algorithms need to be able to *predict* the correct number of clusters (do not require the number)
- Need to scale well
  - Time complexity < $O(n^2)$
- Need to be robust with respect to characteristics of the data
  - E.g., distribution of the duplicates
- Need to be capable of finding 'singleton' clusters
  - Different from many clustering algorithms
    - E.g., algorithms proposed for image segmentation

(slide from O. Hassanzadeh)

Summary of Experimental Results [3]

	Scalability (Current Implementations)	Ability to find the correct number of clusters	Robustness Against		
			Choice of threshold	Amount of Errors	Distribution of errors
Partitioning	High	Low	Low	Low	High
CENTER	High	High	Low	Low	High
MERGE CENTER	High	High	Low	Low	High
Star	Medium	High	Low	Low	High
SR	Low	Medium	High	High	Low
BSR	Low	Low	High	High	Low
CR	Low	High	Medium	High	High
OCR	Low	High	Medium	High	Low
Correlation Clustering	Low	High	Low	Low	High
Markov Clustering	High	High	Medium	Medium	High
Cut Clustering	Low	Low	Low	Low	High
Articulation Point	High	Medium	Low	Low	High

Schema Matching References

1. J. Madhavan, P. A. Bernstein, and E. Rahm. Generic schema matching with cupid. In VLDB, pages 49–58, 2001.
2. S. Melnik, H. Garcia-Molina, and E. Rahm. Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In ICDE, pages 117–128, 2002.
3. H.-H. Do and E. Rahm. COMA: a system for flexible combination of schema matching approaches. In VLDB, pages 610–621, 2002.
4. M. Zhang, M. Hadjieleftheriou, B. C. Ooi, C. M. Procopiuc, D. Srivastava. Automatic discovery of attributes in relational databases. In SIGMOD, pages 109–120, 2011.
5. H. W. Kuhn. The hungarian method for the assignment problem. Naval research logistics quarterly, 2(1-2):83–97, 1955.
6. L. Ramshaw and R. E. Tarjan. On minimum-cost assignments in unbalanced bipartite graphs. HP Labs, Palo Alto, CA, USA, Tech. Rep. HPL-2012-40R1, 2012.

Blocking References – Part I

1. George Papadakis, Dimitrios Skoutas, Emmanouil Thanos, Themis Palpanas: A Survey of Blocking and Filtering Techniques for Entity Resolution. CoRR abs/1905.06167 (2019)
2. I. P. Fellegi and A. B. Sunter. A theory for record linkage. Journal of the American Statistical Association, 64(328):1183–1210, 1969.
3. A. N. Aizawa and K. Oyama. A fast linkage detection scheme for multi-source information integration. In WIRI, pages 30–39, 2005.
4. P. Christen. A survey of indexing techniques for scalable record linkage and deduplication. IEEE TKDE, 24(9):1537–1555, 2012.
5. T. de Vries, H. Ke, S. Chawla, and P. Christen. Robust record linkage blocking using suffix arrays. In CIKM, pages 305–314, 2009
6. R. Baxter, P. Christen, and T. Churches. A comparison of fast blocking methods for record linkage. In KDD Workshops, 2003.
7. B. Kenig and A. Gal. Mfiblocks: An effective blocking algorithm for entity resolution. Inf. Syst., 38(6):908–926, 2013.
8. M. A. Hernández and S. J. Stolfo. The merge/purge problem for large databases. In SIGMOD, pages 127–138, 1995.
9. S. Yan, D. Lee, M. Kan, and C. L. Giles. Adaptive sorted neighborhood methods for efficient record linkage. In JCDL, pages 185–194, 2007.

Blocking References – Part II

11. U. Draisbach, F. Naumann, S. Szott, and O. Wonneberg. Adaptive windows for duplicate detection. In ICDE, pages 1073–1083, 2012.
12. U. Draisbach and F. Naumann. A generalization of blocking and windowing algorithms for duplicate detection. In ICDKE, pages 18–24, 2011
13. M. Bilenko, B. Kamath, and R. J. Mooney. Adaptive blocking: Learning to scale up record linkage. In ICDM, pages 87–96, 2006
14. M. Michelson and C. A. Knoblock. Learning blocking schemes for record linkage. In AAAI, pages 440–445, 2006
15. A. D. Sarma, A. Jain, A. Machanavajjhala, and P. Bohannon. An automatic blocking mechanism for large-scale de-duplication tasks. In CIKM, pages 1055–1064, 2012.
16. M. Kejriwal and D. P. Miranker. An unsupervised algorithm for learning blocking schemes. In ICDM, pages 340–349, 2013.

Matching References

1. K. Qian, L. Popa, P. Sen. Active Learning for Large-Scale Entity Resolution. CIKM 2017.
2. J. Fisher, P. Christen, Q. Wangl. Active Learning Based Entity Resolution Using Markov Logic. PAKDD (2) 2016: 338-349
3. Reyes-Galaviz, O.F., Pedrycz, W., He, Z., Pizzi, N.J. A supervised gradient-based learning algorithm for optimized entity resolution. Data Knowl. Eng. 112, 106–129 (2017)
4. P. Christen. Automatic record linkage using seeded nearest neighbour and support vector machine classification. KDD 2008: 151-159.
5. A. Rasch, R. Schulze, W. Gorus, J. Hiller, S. Bartholomäus, S. Gorlatch. High-performance probabilistic record linkage via multi-dimensional homomorphisms. SAC 2019: 526-533.
6. A. K. Elmagarmid, P. G. Ipeirotis, V. S. Verykios. Duplicate Record Detection: A Survey. IEEE Trans. Knowl. Data Eng. 19(1): 1-16 (2007)
7. A. Jurek, J. Hong, Y. Chi, W. Liu. A novel ensemble learning approach to unsupervised record linkage. Inf. Syst. 71: 40-54 (2017)
8. A. Jurek, Deepak P. It Pays to Be Certain: Unsupervised Record Linkage via Ambiguity Minimization. PAKDD (3) 2018: 177-190.X
9. L. Dong, A. Y. Halevy, J. Madhavan. Reference Reconciliation in Complex Information Spaces. SIGMOD Conference 2005: 85-96.O

10.. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, J. Widom. Swoosh: a generic approach to entity resolution. VLDB J. 18(1): 255-276 (2009).

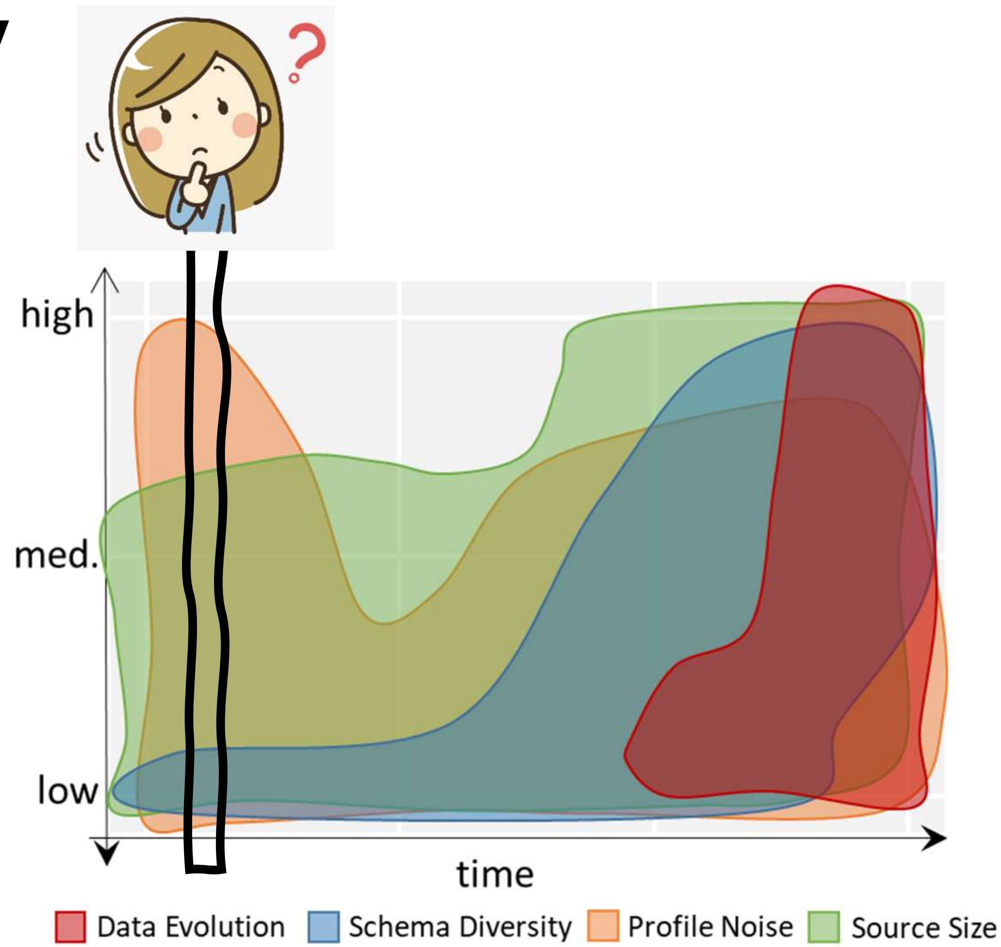
11.I. Bhattacharya, L. Getoor. Collective entity resolution in relational data. TKDD 1(1): 5 (2007).

Clustering References

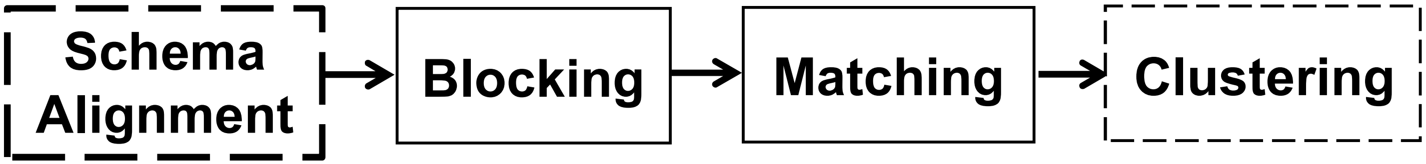
1. S. Lacoste-Julien, K. Palla, A. Davies, G. Kasneci, T. Graepel, Z. Ghahramani. SIGMa: simple greedy matching for aligning large knowledge bases. KDD 2013: 572-580
2. F. M. Suchanek, S. Abiteboul, P. Senellart. PARIS: Probabilistic Alignment of Relations, Instances, and Schema. PVLDB 5(3): 157-168 (2011)
3. O. Hassanzadeh, F. Chiang, R. J. Miller, H. C. Lee. Framework for Evaluating Clustering Algorithms in Duplicate Detection. PVLDB 2(1): 1282-1293 (2009)
4. H. W. Kuhn. The hungarian method for the assignment problem. Naval research logistics quarterly, 2(1-2):83–97, 1955.
5. L. Ramshaw and R. E. Tarjan. On minimum-cost assignments in unbalanced bipartite graphs. HP Labs, Palo Alto, CA, USA, Tech. Rep. HPL-2012-40R1, 2012.
6. A. Jain, R. Dubes, “Algorithms for Clustering Data”, Prentice Hall, 1988.

Veracity
+Volume
+Variety
+Velocity

ER Methods



Volume & Veracity



- Workflow remains the same
- Scope:
  - (tens of) millions of structured profiles
- Goals:
  - High accuracy despite noise
  - High time efficiency despite the size of data
- Assumptions:
  - Known schema → custom, schema-based solutions

Solution: Parallelization

Two types:

- Multi-core parallelization
  - Single system $\longrightarrow$ shared memory
  - Distribute processing among available CPUs
- Massive parallelization
  - Cluster of independent systems
  - Map-Reduce paradigm [1]
    - Data partitioned across the nodes of a cluster
    - Map Phase: transforms a data partition into (key, value) pairs
    - Reduce Phase: processes pairs with the same key

Mechanisms per Workflow step

Schema Alignment → Blocking → Matching → Clustering

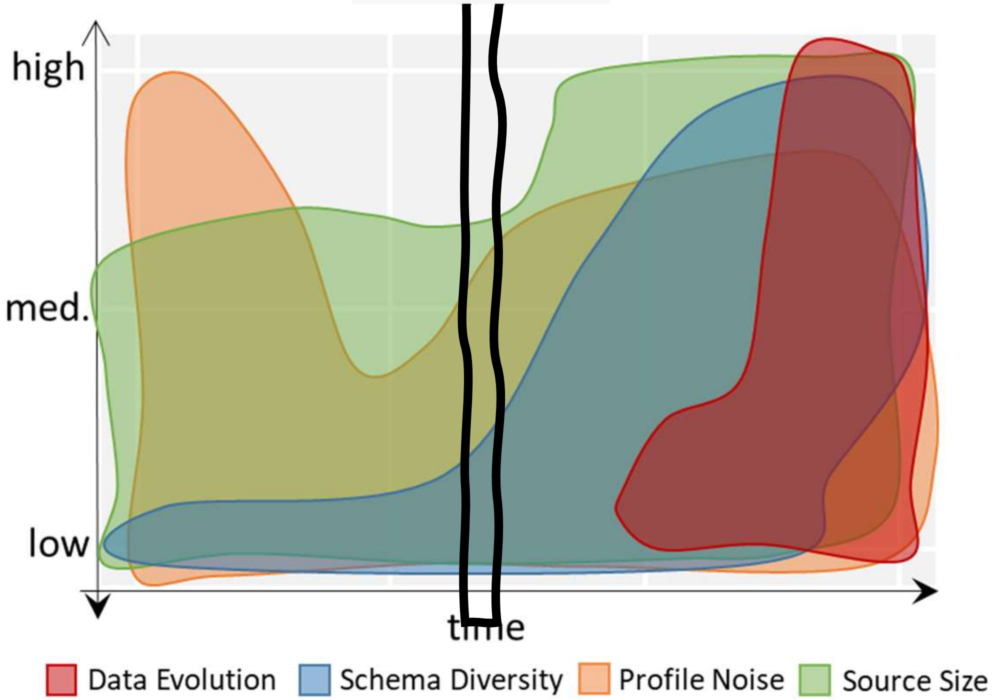
- Blocking based on map reduce
  - Dedoop [2]
  - MapReduce-based Sorted Neighborhood [3]
- Matching
  - Multi-core approaches [7, 8]
  - MapReduce-based: Emphasis on load balancing
    - BlockSplit & PairRange [4, 5]
    - Dis-Dedup [6]
    - Message-passing framework [9]
- Clustering
  - Fast Multi-source ER (FAMER) framework [10, 11]

References

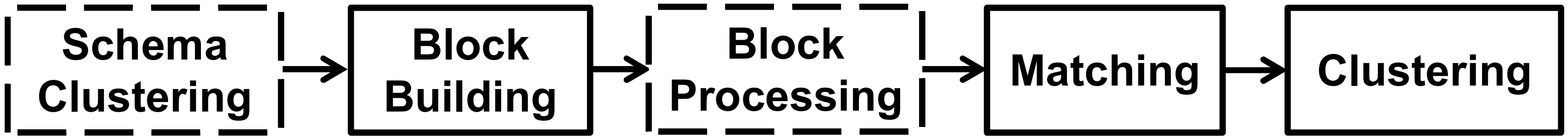
1. J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. Commun. ACM, 51(1):107–113, 2008.
2. L. Kolb, A. Thor, and E. Rahm. Dedoop: Efficient deduplication with hadoop. PVLDB, 5(12):1878–1881, 2012.
3. L. Kolb, A. Thor, and E. Rahm. Multi-pass sorted neighborhood blocking with mapreduce. Computer Science - R&D, 27(1):45–63, 2012.
4. L. Kolb, A. Thor, and E. Rahm. Load balancing for mapreduce-based entity resolution. In ICDE, pages 618–629, 2012.
5. W. Yan, Y. Xue, and B. Malin. Scalable load balancing for mapreduce-based record linkage. In IPCCC, pages 1–10, 2013.
6. X. Chu, I. F. Ilyas, and P. Koutris. Distributed data deduplication. PVLDB, 9(11):864–875, 2016.
7. O. Benjelloun, H. Garcia-Molina, H. Gong, H. Kawai, T. E. Larson, D. Menestrina, and S. Thavisomboon. D-swoosh: A family of algorithms for generic, distributed entity resolution. In ICDCS, page 37, 2007.
8. Hung-sik Kim and Dongwon Lee. Parallel linkage. In CIKM, pages 283–292, 2007.
9. V. Rastogi, N. N. Dalvi, and M. N. Garofalakis. Large-scale collective entity matching. PVLDB, 4(4):208–218, 2011.
10. A. Saeedi, E. Peukert, and E. Rahm. Comparative evaluation of distributed clustering schemes for multi-source entity resolution. In ADBIS, pages 278–293, 2017.
11. A. Saeedi, M. Nentwig, E. Peukert, and E. Rahm. Scalable matching and clustering of entities with FAMER. CSIMQ, 16:61–83, 2018.

Veracity
+Volume
+Variety
+Velocity

ER Methods



Variety & Volume & Veracity

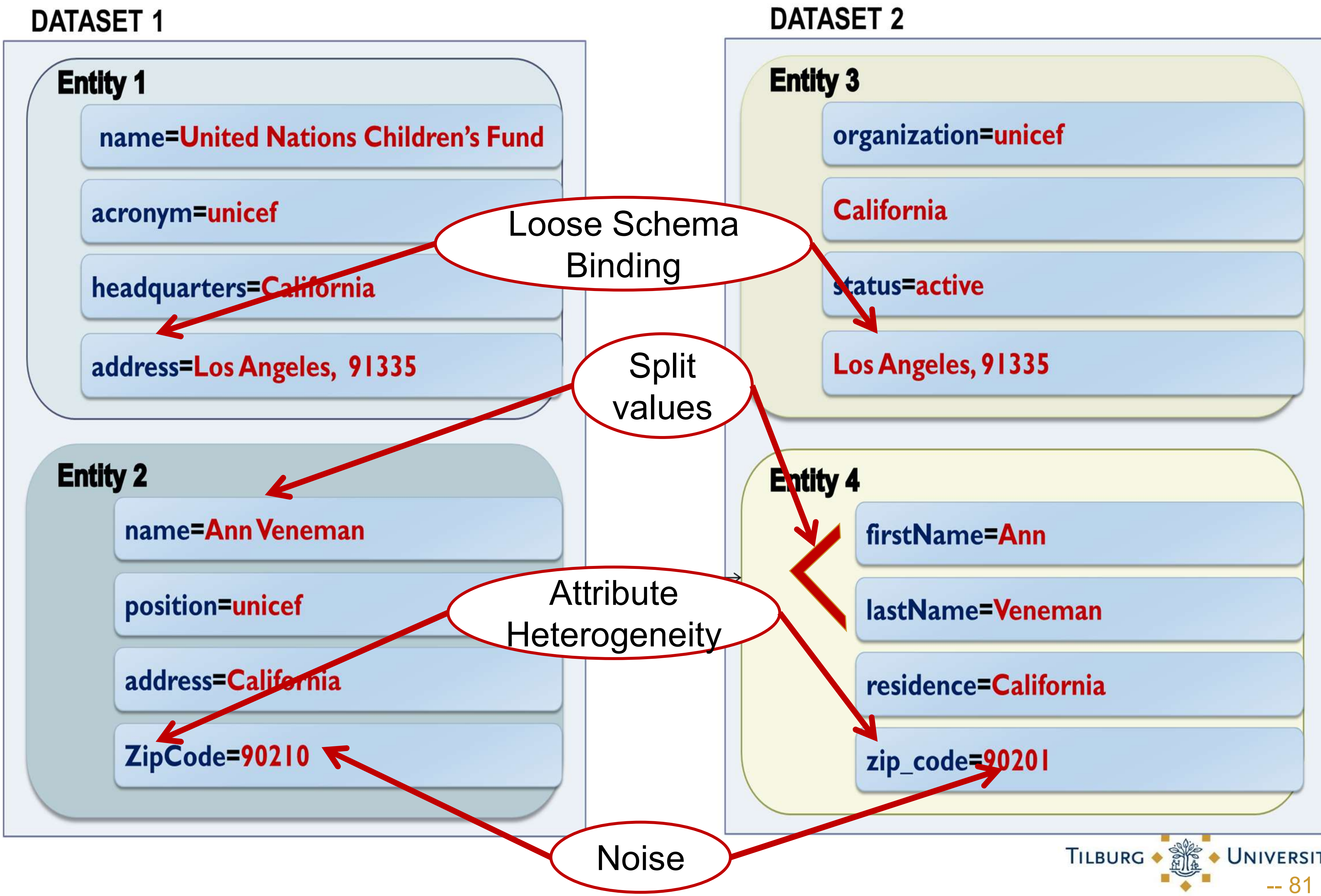


Scope (e.g., user-generated Web Data):

- Voluminous, (semi-)structured datasets
  - BTC09: 1.15 billion triples, 182 million profiles
- Users are free to add attribute values and/or attribute names

  → unprecedented levels of schema heterogeneity
  - Google Base: 100,000 schemata for 10,000 profile types
  - BTC09:136,000 attribute names
- Several datasets produced by automatic information extraction techniques → noise, tag-style values

Example of Web Data



Schema Clustering

- Schema Matching → not applicable (too many alternatives)
- Instead, partition attributes according to their syntactic similarity, regardless of their semantic relation
- Goal: Facilitate next steps
- Scope: Both Clean-Clean and Dirty ER
- Attribute Clustering [1, 2, 3]
  - Create a graph, with nodes representing attributes
  - For each node n_i
    - Find the most similar node n_j
    - If $sim(n_i,n_j) > 0$, add an edge $<n_i,n_j>$
  - Extract connected components
  - Put all singleton nodes in a “glue” cluster

Block Building

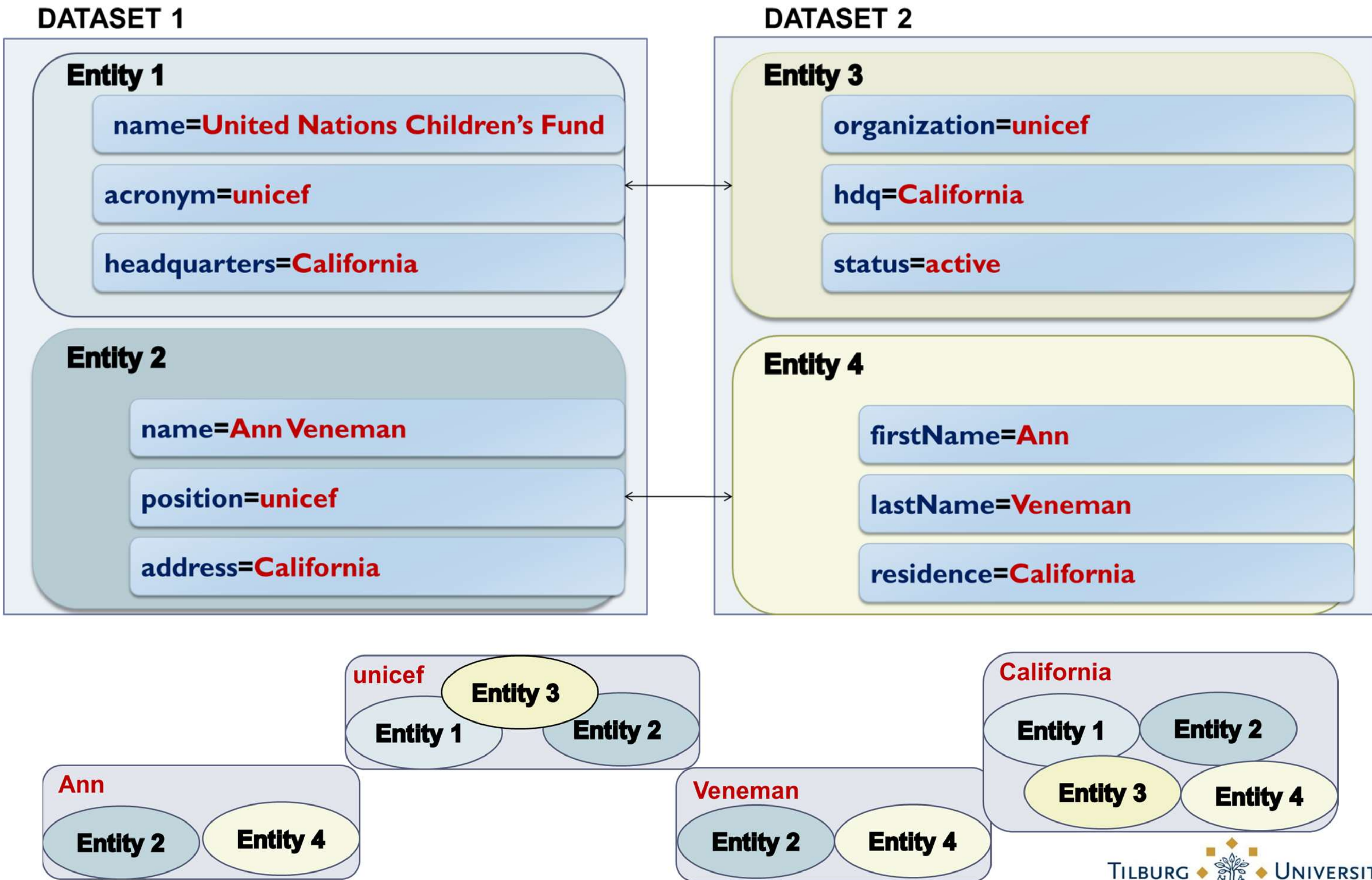
Schema Clustering → Block Building → Block Processing → Matching → Clustering

- Considers all attribute values and completely ignores all attribute names → schema-agnostic functionality

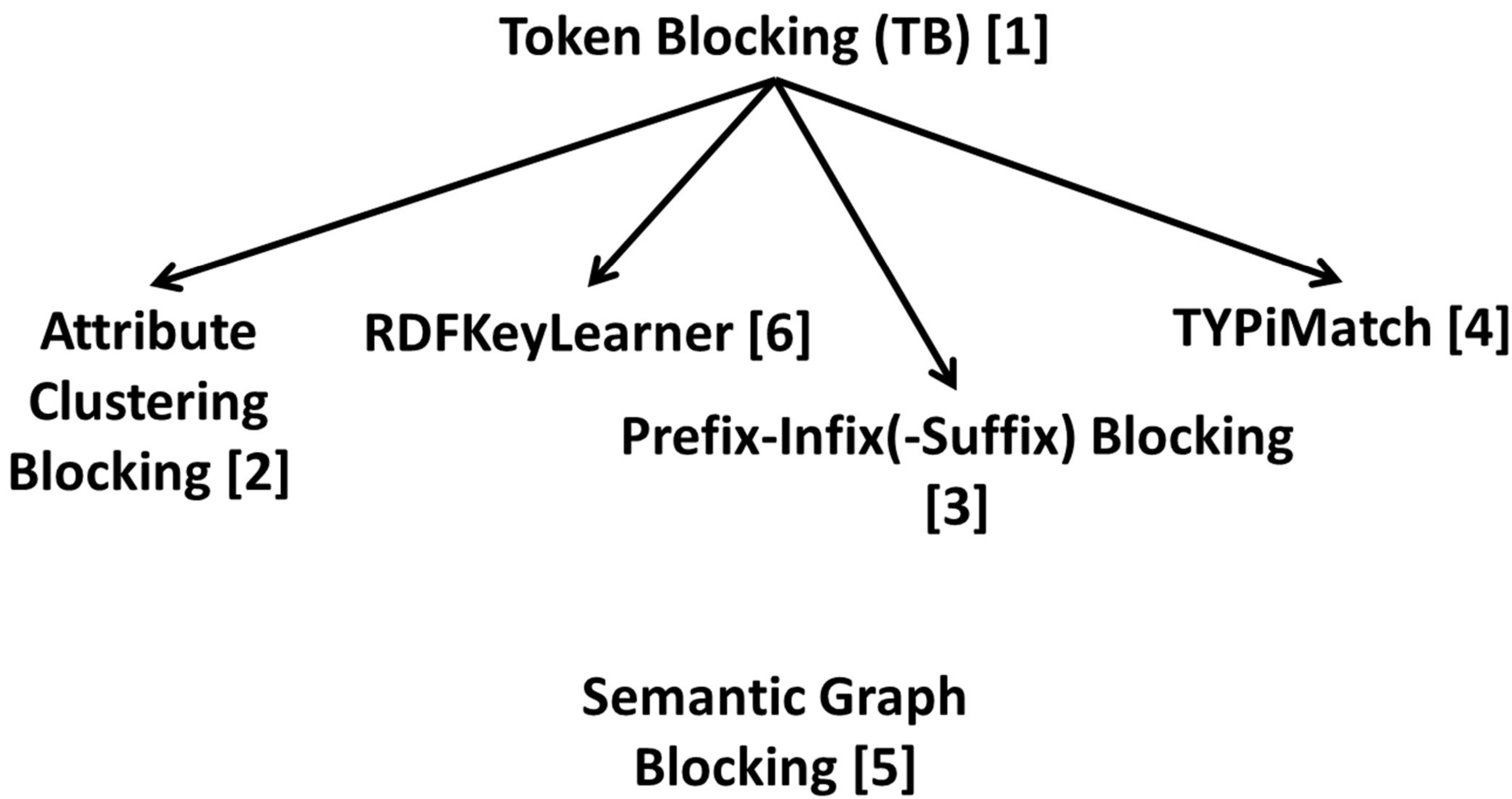
- Core approach: Token Blocking [1]
  1. Given a profile, extract all tokens that are contained in its attribute values
  2. Create one block for every distinct token with frequency > 2 → each block contains all profiles with the corresponding token

     Pros:
     - Parameter-free
     - Efficient
     - Unsupervised

Example of Token Blocking

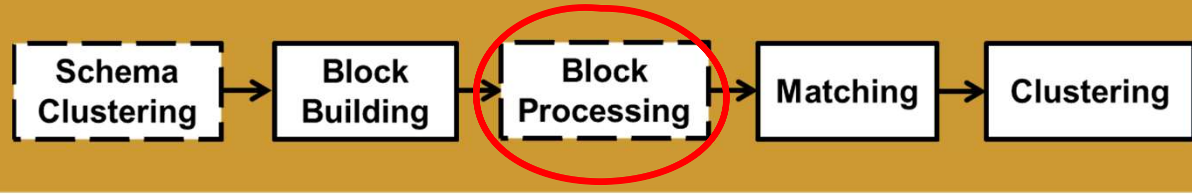


Genealogy of Block Building Techniques [8]



MapReduce-based parallelizations in [7]

Block Processing

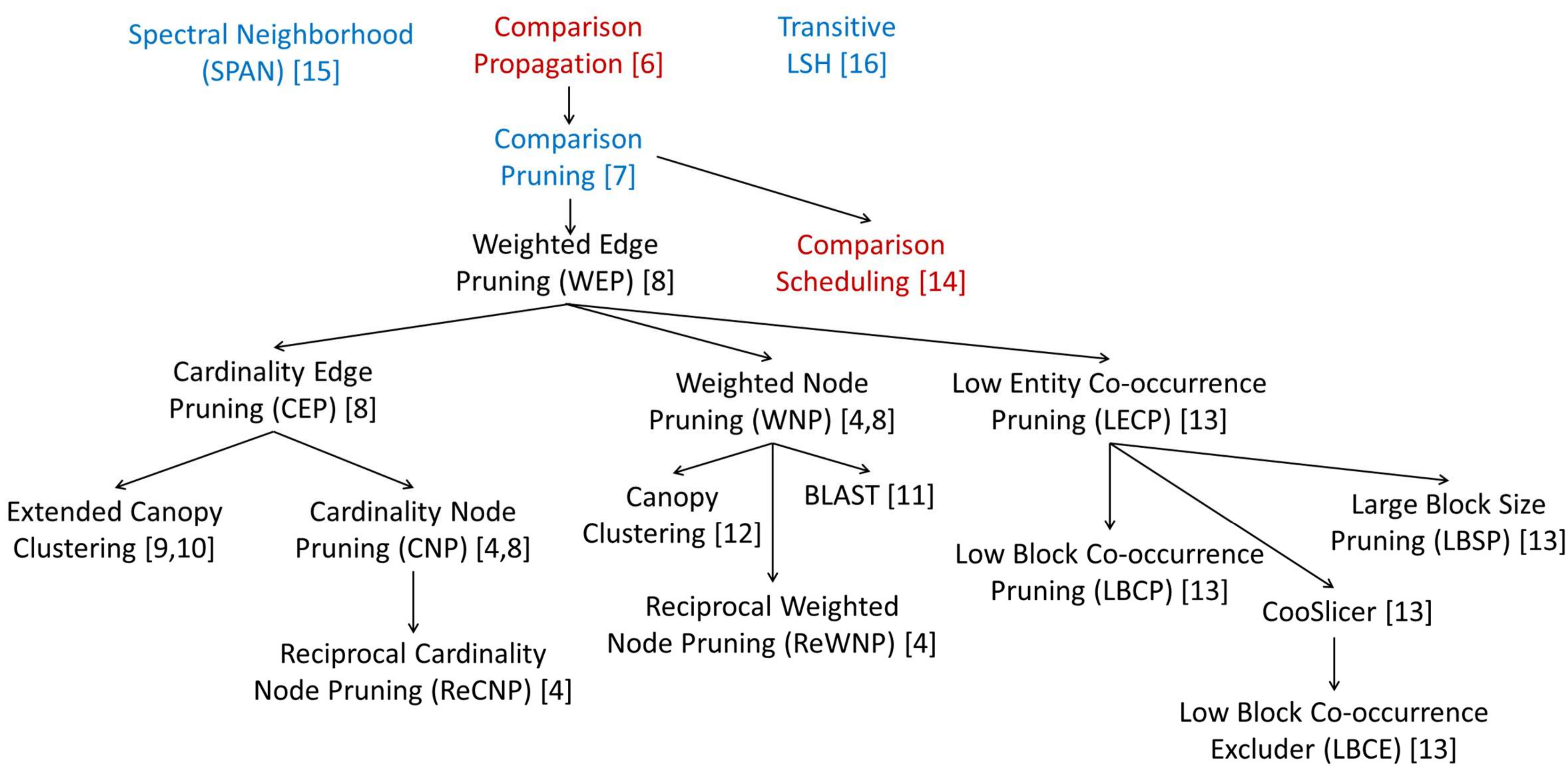


- Block Building resulted into huge number of blocks
- Results in addition of new step in the workflow
- Goal:
  - Restructure the original blocks in order to increase precision at no significant cost in recall
- Focus on reducing / removing comparisons:
  1. Redundant comparisons, i.e., comparing profiles that were already compared in a previous block
  2. Superfluous comparisons, i.e., high number of comparisons between irrelevant profiles

Block Processing Techniques

- Block Clustering:
  - Operate at the level of entire blocks
  - Methods
    - Block Purging [1,2,3]
    - Block Filtering [4]
    - Block Clustering [5]
- Comparison Cleaning:
  - Methods (next slide)

Comparison Cleaning Methods [17]



Entity Matching

Schema Clustering → Block Building → Block Processing → Matching → Clustering

- Collective approaches to tackle Variety
- Most methods crafted for Clean-Clean ER
- General outline of SiGMa[1], PARIS[2], LINDA[3], RiMOM-IM[4,5]
  - Iterative process starts with a few reliable seed matches
  - Propagate initial matches to neighbors
  - Order candidate matches in descending overall similarity
  - Recompute the similarity of the neighbors
  - Update candidate matches order
- MinoanER [6] performs a specific number of steps, rather than iterating until convergence

Entity Clustering

- Methods discussed before are still applicable
  - Only difference: similarity scores extracted in a schema-agnostic fashion, not from specific attributes
- SplitMerge [1]
  - Inherently capable of handling heterogeneous semantic types

[1] M. Nentwig, A. Groß, and E. Rahm. Holistic entity clustering for linked data. In ICDM Workshops, 2016.

Schema Clustering References

1. G. Papadakis, E. Ioannou, T. Palpanas, C. Niederée, W. Nejdl. A Blocking Framework for Entity Resolution in Highly Heterogeneous Information Spaces. IEEE Trans. Knowl. Data Eng. 25(12): 2665-2682 (2013)
2. G. Simonini, S. Bergamaschi, H. V. Jagadish. BLAST: a Loosely Schema-aware Meta-blocking Approach for Entity Resolution. PVLDB 9(12): 1173-1184 (2016)
3. G. Papadakis, L. Tsekouras, E. Thanos, G. Giannakopoulos, T. Palpanas, M. Koubarakis. The return of JedAI: End-to-End Entity Resolution for Structured and Semi-Structured Data. PVLDB 11(12): 1950-1953 (2018)

Block Building References

1. G. Papadakis, E. Ioannou, C. Niederée, P. Fankhauser. Efficient entity resolution for large heterogeneous information spaces. WSDM 2011: 535-544
2. G. Papadakis, E. Ioannou, T. Palpanas, C. Niederée, W. Nejdl. A Blocking Framework for Entity Resolution in Highly Heterogeneous Information Spaces. IEEE Trans. Knowl. Data Eng. 25(12): 2665-2682 (2013)
3. G. Papadakis, E. Ioannou, C. Niederée, T. Palpanas, W. Nejdl. Beyond 100 million entities: large-scale blocking-based resolution for heterogeneous data. WSDM 2012: 53-62
4. Y. Ma, T. Tran. TYPiMatch: type-specific unsupervised learning of keys and key values for heterogeneous web data integration. WSDM 2013: 325-334
5. J. Nin, V. Muntés-Mulero, N. Martínez-Bazan, and J. Larriba-Pey. On the use of semantic blocking techniques for data cleansing and integration. In IDEAS, pages 190–198, 2007.
6. D. Song and J. Heflin. Automatically generating data linkages using a domain-independent candidate selection approach. In ISWC, pages 649–664, 2011.
7. V. Christophides, V. Efthymiou, K. Stefanidis. Entity Resolution in the Web of Data. Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool Publishers 2015.
8. G. Papadakis, D. Skoutas, E. Thanos, T. Palpanas: Blocking and Filtering Techniques for Entity Resolution: A Survey. ACM Computing Surveys 2020.

Block Processing References – Part I

1. G. Papadakis, E. Ioannou, C. Niederée, P. Fankhauser. Efficient entity resolution for large heterogeneous information spaces. WSDM 2011: 535-544
2. G. Papadakis, E. Ioannou, T. Palpanas, C. Niederée, W. Nejdl. A Blocking Framework for Entity Resolution in Highly Heterogeneous Information Spaces. IEEE Trans. Knowl. Data Eng. 25(12): 2665-2682 (2013)
3. G. Papadakis, E. Ioannou, C. Niederée, T. Palpanas, W. Nejdl. Beyond 100 million entities: large-scale blocking-based resolution for heterogeneous data. WSDM 2012: 53-62
4. G. Papadakis, G. Papastefanatos, T. Palpanas, M. Koubarakis. Scaling Entity Resolution to Large, Heterogeneous Data with Enhanced Meta-blocking. EDBT 2016: 221-232
5. J. Fisher, P. Christen, Q. Wang, E. Rahm. A Clustering-Based Framework to Control Block Sizes for Entity Resolution. KDD 2015: 279-288
6. G. Papadakis, E. Ioannou, C. Niederée, T. Palpanas, W. Nejdl. Eliminating the redundancy in blocking-based entity resolution methods. JCDL 2011: 85-94.
7. G. Papadakis, E. Ioannou, C. Niederée, T. Palpanas, W. Nejdl. To compare or not to compare: making entity resolution more efficient. SWIM 2011: 3.
8. G. Papadakis, G. Koutrika, T. Palpanas, W. Nejdl. Meta-Blocking: Taking Entity Resolutionto the Next Level. IEEE Trans. Knowl. Data Eng. 26(8): 1946-1960 (2014).
9. P. Christen. A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. IEEE Trans. Knowl. Data Eng. 24(9): 1537-1555 (2012).

Block Processing References – Part II

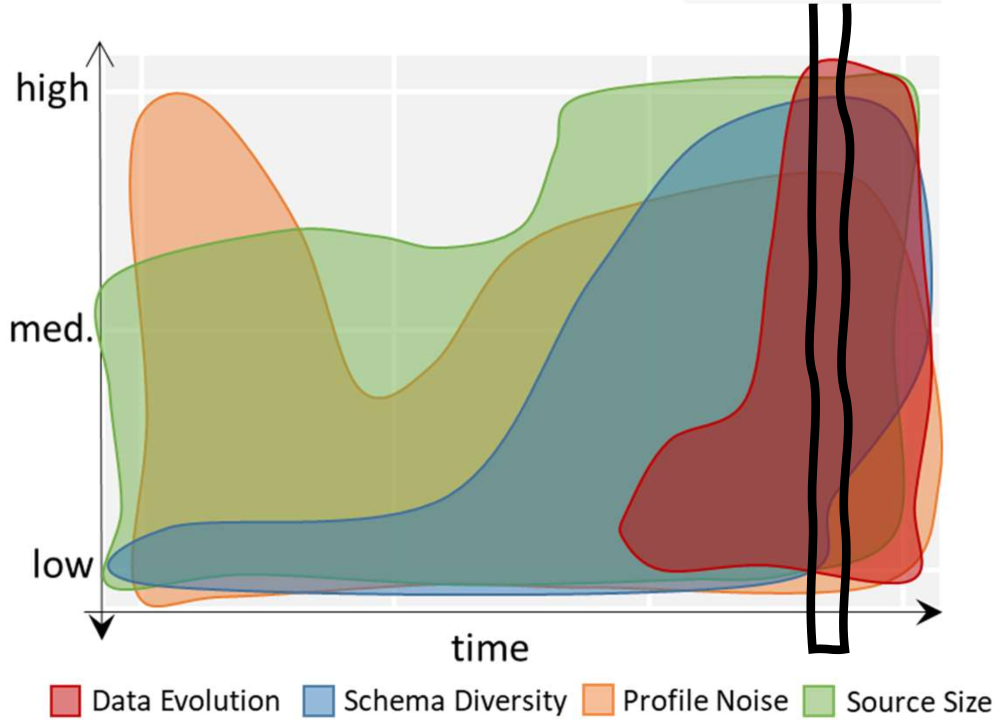
10. G. Papadakis, G. Alexiou, G. Papastefanatos, G. Koutrika. Schema-agnostic vs Schema-based Configurations for Blocking Methods on Homogeneous Data. PVLDB 9(4): 312-323 (2015).
11. G. Simonini, S. Bergamaschi, H. V. Jagadish. BLAST: a Loosely Schema-aware Meta-blocking Approach for Entity Resolution. PVLDB 9(12): 1173-1184 (2016)
12. A. McCallum, K. Nigam, L. H. Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. KDD 2000: 169-178.
13. D. C. Nascimento, C. E. S. Pires, and D. G. Mestre. Exploiting block co-occurrence to control block sizes for entity resolution. Knowledge and Information Systems, pages 1–42, 2019.
14. G. Papadakis, E. Ioannou, T. Palpanas, C. Niederée, and W. Nejdl. A blocking framework for entity resolution in highly heterogeneous information spaces. IEEE TKDE, 25(12):2665–2682, 2013.
15. L. Shu, A. Chen, M. Xiong, and W. Meng. Efficient spectral neighborhood blocking for entity resolution. In ICDE, pages 1067–1078, 2011.
16. R. C. Steorts, S. L. Ventura, M. Sadinle, and S. E. Fienberg. A comparison of blocking methods for record linkage. In Privacy in Statistical Databases, pages 253–268, 2014.
17. George Papadakis, Dimitrios Skoutas, Emmanouil Thanos, Themis Palpanas: A Survey of Blocking and Filtering Techniques for Entity Resolution. CoRR abs/1905.06167 (2019)

Entity Matching References

1. S. Lacoste-Julien, K. Palla, A. Davies, G. Kasneci, T. Graepel, Z. Ghahramani. SIGMa: simple greedy matching for aligning large knowledge bases. KDD 2013: 572-580
2. F. M. Suchanek, S. Abiteboul, P. Senellart. PARIS: Probabilistic Alignment of Relations, Instances, and Schema. PVLDB 5(3): 157-168 (2011)
3. C. Böhm, G. de Melo, F. Naumann, and G. Weikum. LINDA: distributed web-of-data-scale entity matching. In CIKM, pages 2104–2108, 2012.
4. J. Li, J. Tang, Y. Li, and Q. Luo. Rimom: A dynamic multistrategy ontology alignment framework. TKDE, 21(8):1218–1232, 2009.
5. C. Shao, L. Hu, J. Li, Z. Wang, T. L. Chung, and J.-B. Xia. Rimom-im: A novel iterative framework for instance matching. J. Comput. Sci. Technol., 31(1):185–197, 2016.
6. V. Efthymiou, G. Papadakis, K. Stefanidis, and V. Christophides. MinoanER: Schema-agnostic, non-iterative, massively parallel resolution of web entities. In EDBT, pages 373–384, 2019.

Veracity
+Volume
+Variety
+Velocity

ER Methods

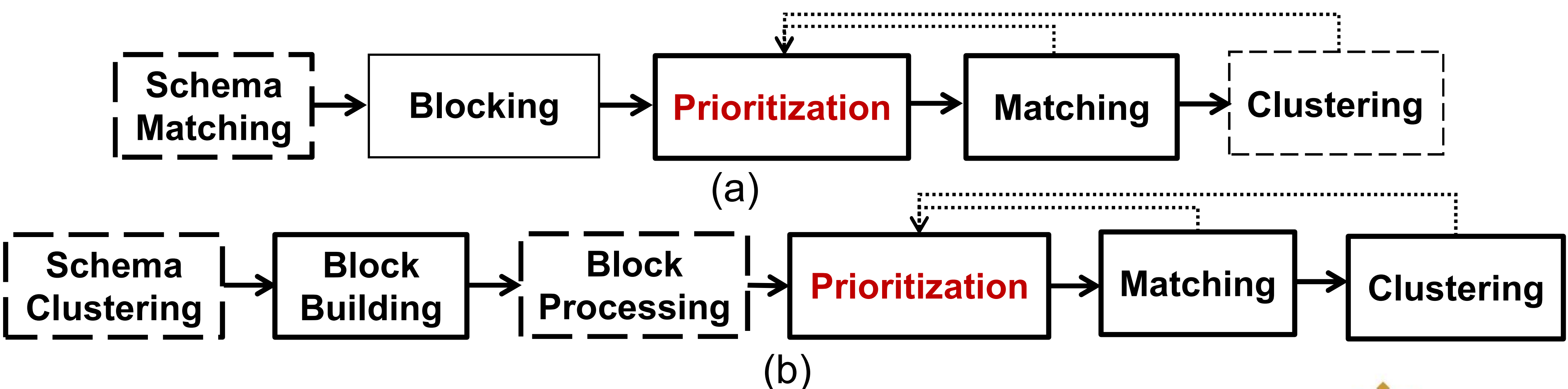


Velocity & Variety & Volume & Veracity

Scope:

- Applications with increasing data volume & time constraints
  - Loose ones (e.g., minutes, hours) → Progressive ER
  - Strict ones (i.e., seconds) → Real-time (On-line) ER

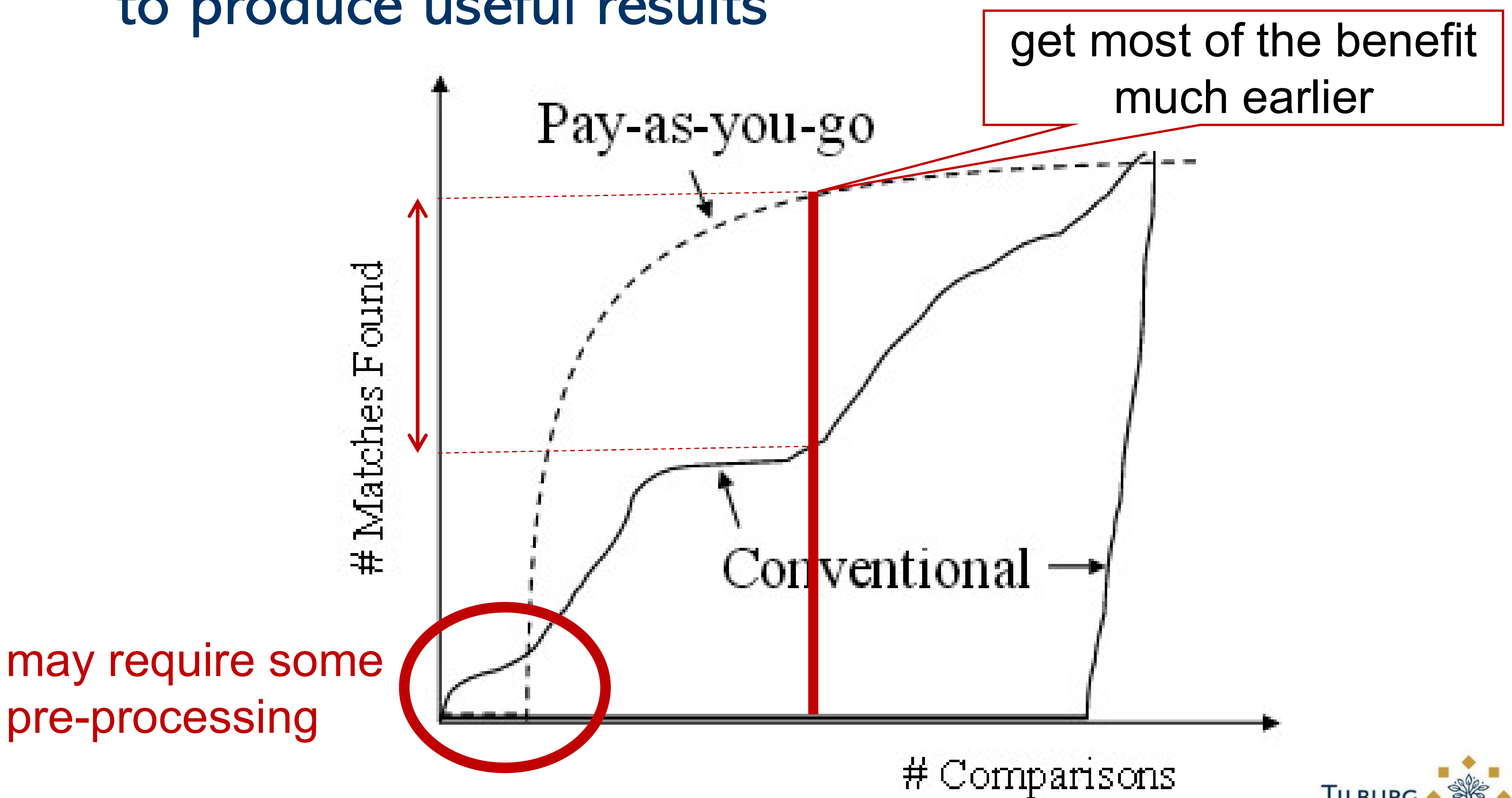
- End-to-end workflows for Progressive ER:



Progressive Entity Resolution

- Unprecedented, increasing volume of data

→ applications can compromise with partial solutions to produce useful results

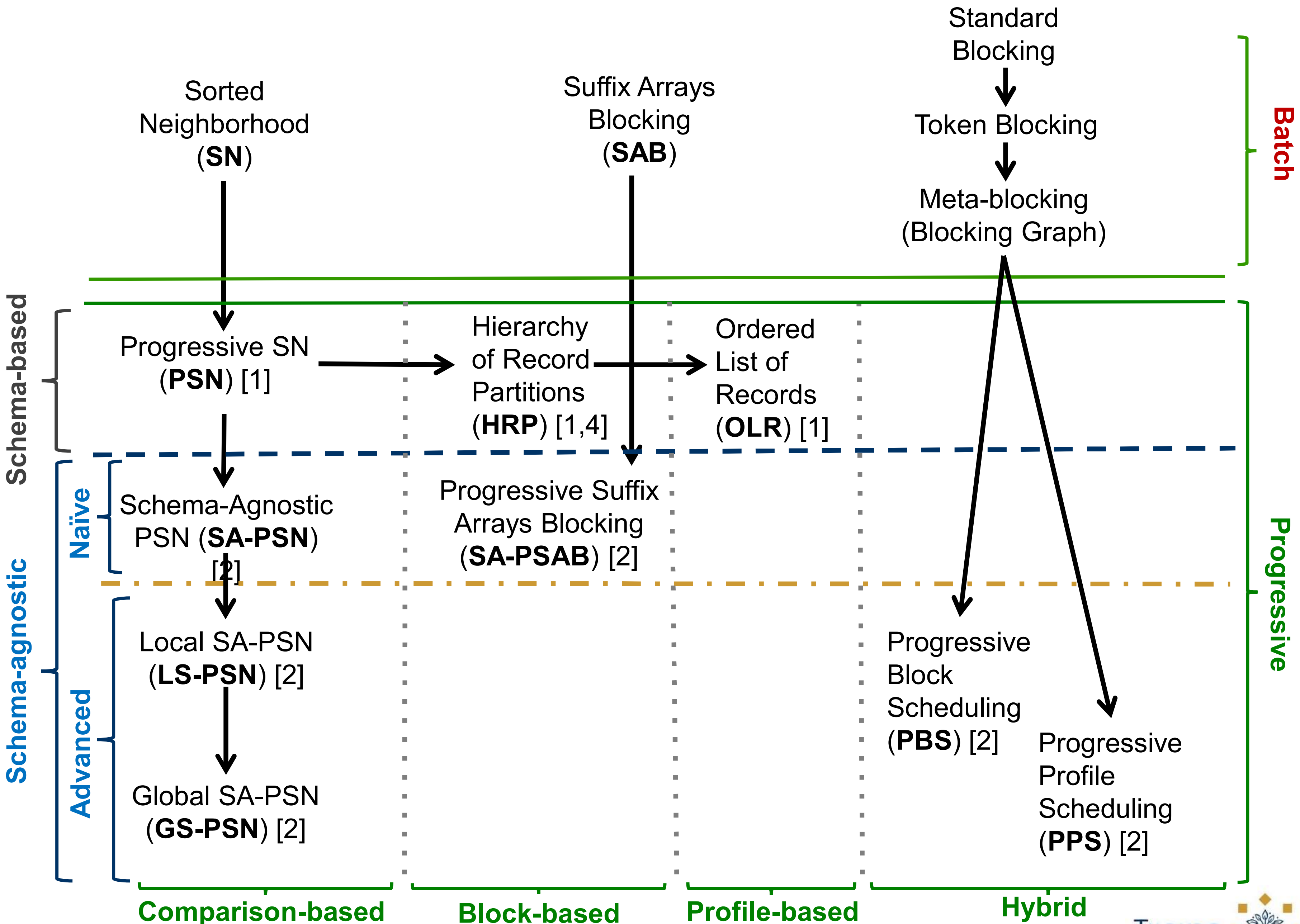


Outline Progressive ER

- Requires:
  - Improved early quality
  - Same eventual quality

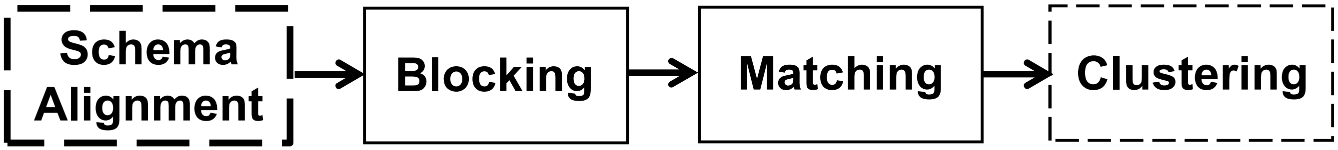
- Prioritization
  - Defines optimal processing order for a set of entities
  - Static methods [1, 2]:
    - Guide which records to compare first (independently of ER matching results)
  - Dynamic methods [3]:
    - If a duplicate is found, then check neighbors as well
    - Assumption: oracle for entity matching

Taxonomy of Static Prioritization Methods



Real-time Entity Resolution

Same workflow as original two workflows:



Same scope (so far):

- Structured data

Different input:

- stream of query profiles

Different goal:

- resolve each query over a large dataset in the shorted possible time (& with the minimum memory footprint)

Techniques per workflow step

Incremental Blocking

- DySimII [1] - extends Standard Blocking
- F-DySNI [2,3] - extends Sorted Neighborhood
- (S)BlockSketch [4] - bounded matching time, constant memory footprint

Incremental Matching

- QDA [5] - SQL-like selection queries over a single dataset
- QuERy [6] - complex join queries over multiple, overlapping, dirty DSs
- EAQP [7] - queries under data
- Evolving matching rules [8]

Incremental Clustering

- Incremental Correlation Clustering [9]

Progressive ER References

1. S. E. Whang, D. Marmaros, and H. Garcia-Molina. Pay-as-you-go entity resolution. TKDE, 25(5):1111–1124, 2013.
2. T. Papenbrock, A. Heise, and F. Naumann. Progressive duplicate detection. TKDE, 27(5):1316–1329, 2015.
3. G. Simonini, G. Papadakis, T. Palpanas, S. Bergamaschi. Schema-Agnostic Progressive Entity Resolution. IEEE Trans. Knowl. Data Eng. 31(6): 1208-1221 (2019)
4. Y. Altowim and S. Mehrotra. Parallel progressive approach to entity resolution using mapreduce. In ICDE, pages 909–920, 2017.

Incremental ER References

1. B. Ramadan and P. Christen, H. Liang, and R. W. Gayler, and D. Hawking. Dynamic similarity-aware inverted indexing for real-time entity resolution. In PAKDD Workshops, pages 47–58, 2013.
2. B. Ramadan and P. Christen. Forest-based dynamic sorted neighborhood indexing for real-time entity resolution. In CIKM, pages 1787–1790, 2014.
3. B. Ramadan and P. Christen, H. Liang, and R. W. Gayler. Dynamic sorted neighborhood indexing for real-time entity resolution. J. Data and Information Quality, 6(4):15:1–15:29, 2015.
4. D. Karapiperis, A. Gkoulalas-Divanis, V. S. Verykios. Summarization Algorithms for Record Linkage. EDBT 2018: 73-84.
5. H. Altwaijry, D. V. Kalashnikov, and S. Mehrotra. QDA: A query-driven approach to entity resolution. TKDE, 29(2):402–417, 2017.
6. H. Altwaijry, S. Mehrotra, and D. V. Kalashnikov. Query: A framework for integrating entity resolution with query processing. PVLDB, 9(3):120–131, 2015.
7. E. Ioannou, W. Nejdl, C. Niederée, and Y. Velegrakis. On-the-fly entity-aware query processing in the presence of linkage. PVLDB, 3(1): 429–438, 2010.
8. S. E. Whang and H. Garcia-Molina. Entity resolution with evolving rules. PVLDB, 3(1):1326–1337, 2010.
9. A. Gruenheid, X. L. Dong, and D. Srivastava. Incremental record linkage. Proc. VLDB Endow., 7(9):697–708, May 2014. ISSN 2150-8097.