

Data Cleaning for Data Integration

Advanced School on Data Exchange,
Integration, and Streams (DEIS)

Ekaterini Ioannou

Tuesday, 9th of Nov. 2010, Schloss Dagstuhl

Data integration:

- Combine data from various sources/applications
- Merge into a single database
- Requires a unified view over the data → cleaning

Challenges:

- Handling the various incoming schemata
- Dealing with the missing data values
- **Entity Resolution**
 - combine the various descriptions or references for the same real world objects

Reasons for Various Descriptions

- Text variations:
 - Misspellings
 - Acronyms
 - Transformations
 - Abbreviations
 - etc.

Welcome to ICDE 2011

The IEEE International Conference on Data Engineering results and advanced data-intensive applications and dis The mission of the conference is to share research soluti identifv new issues and directions for future research anc

The Journal of Web Semantics is an interdisciplinary various subject areas that contribute to the devel service Web. These areas include: knowledge tec semantic grid, obviously disciplines like ... click h



Enrico Minack, Kaluca Paul-Alexandru Chirita, Wolfgang Nejdl: Leveraging personal metadata system J. Web Sem. 8(1): 37-54 (2010)

Reasons for Various Descriptions

- Text variations
- Local knowledge:
 - Each source uses different formats
e.g., person from publication vs. person from email
 - Lack of global coordination for identifier assignment

On-the-Fly Entity in the P

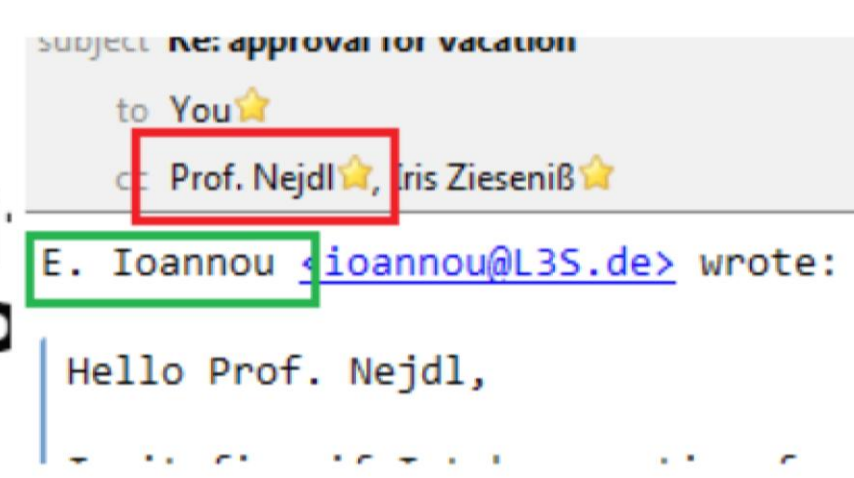
Ekaterini Ioannou
L3S Research Center
Hannover, Germany
ioannou@L3S.de

Wolfgang Nejdl
L3S Research Center
Hannover, Germany
nejdl@L3S.de

Claudia Niederee
L3S Research Center
Hannover, Germany
niederee@L3S.de

ABSTRACT
Entity linkage is central to almost every data integration and data

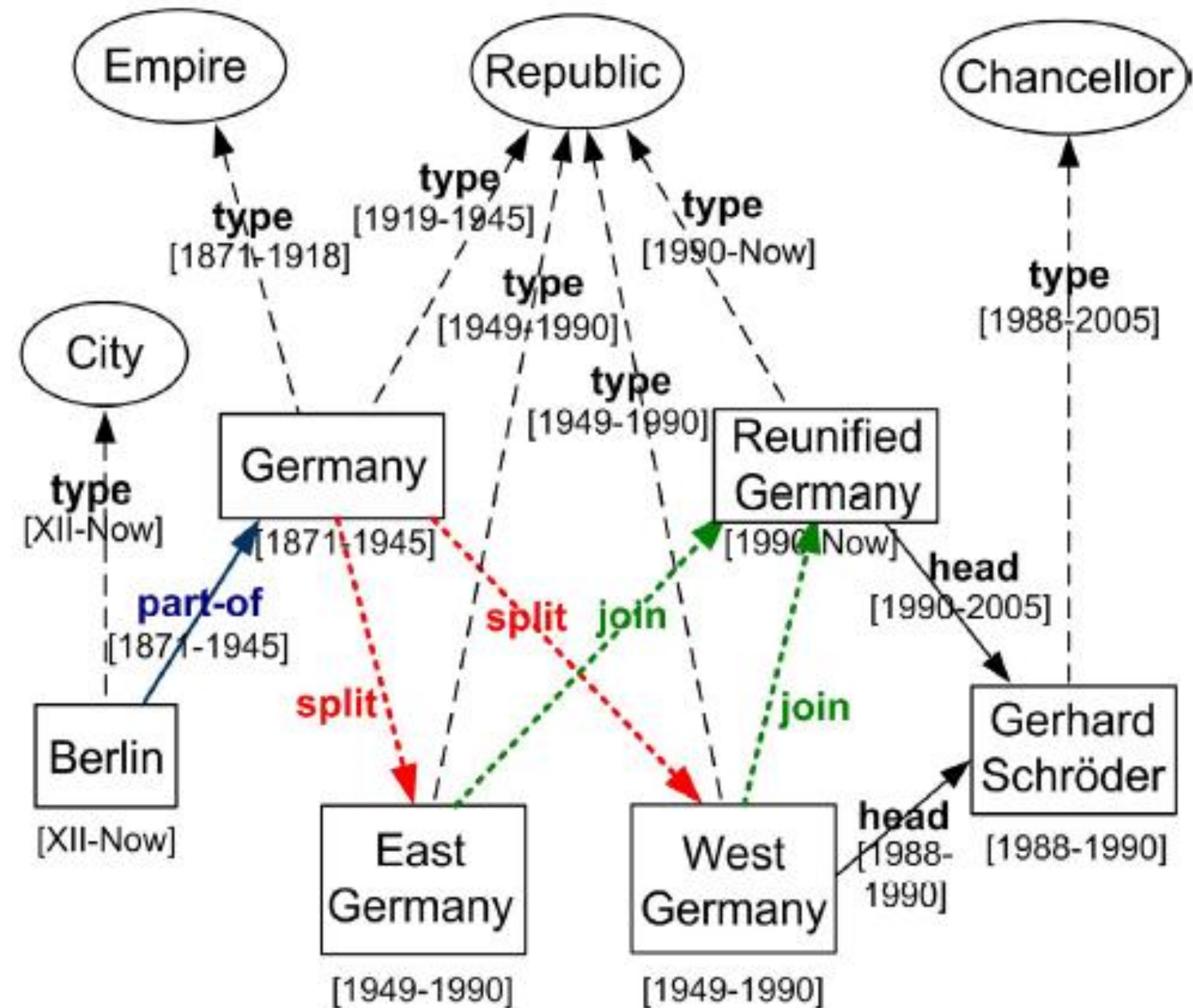
needs to identify these pi
single artifact that will be



subject: re: approval for vacation
to: You ★
cc: Prof. Nejdl ★, Iris Zieseniß ★
E. Ioannou <ioannou@L3S.de> wrote:
Hello Prof. Nejdl,

figure from [RVMB09]

- Text variations
- Local knowledge
- Evolving nature of data:
 - Entity alternative names appearing in time
 - Updates in entity data



Jacqueline Lee Bouvier

Alternate Names: Jackie Bouvier | Jackie Kennedy | Mrs. John F. Kennedy | Jackie Onassis | Jacqueline Kennedy Onassis | Jacqueline Onassis

- Text variations
- Local knowledge
- Evolving nature of data
- **New functionality:**
 - Web page extraction
e.g., Calais, Cogito
 - Import data collections from various applications
e.g., Wikipedia data used in Freebase
 - Mashups for easy and fast integration from various source
e.g., yahoo pipes

Entity Resolution typical methodology:

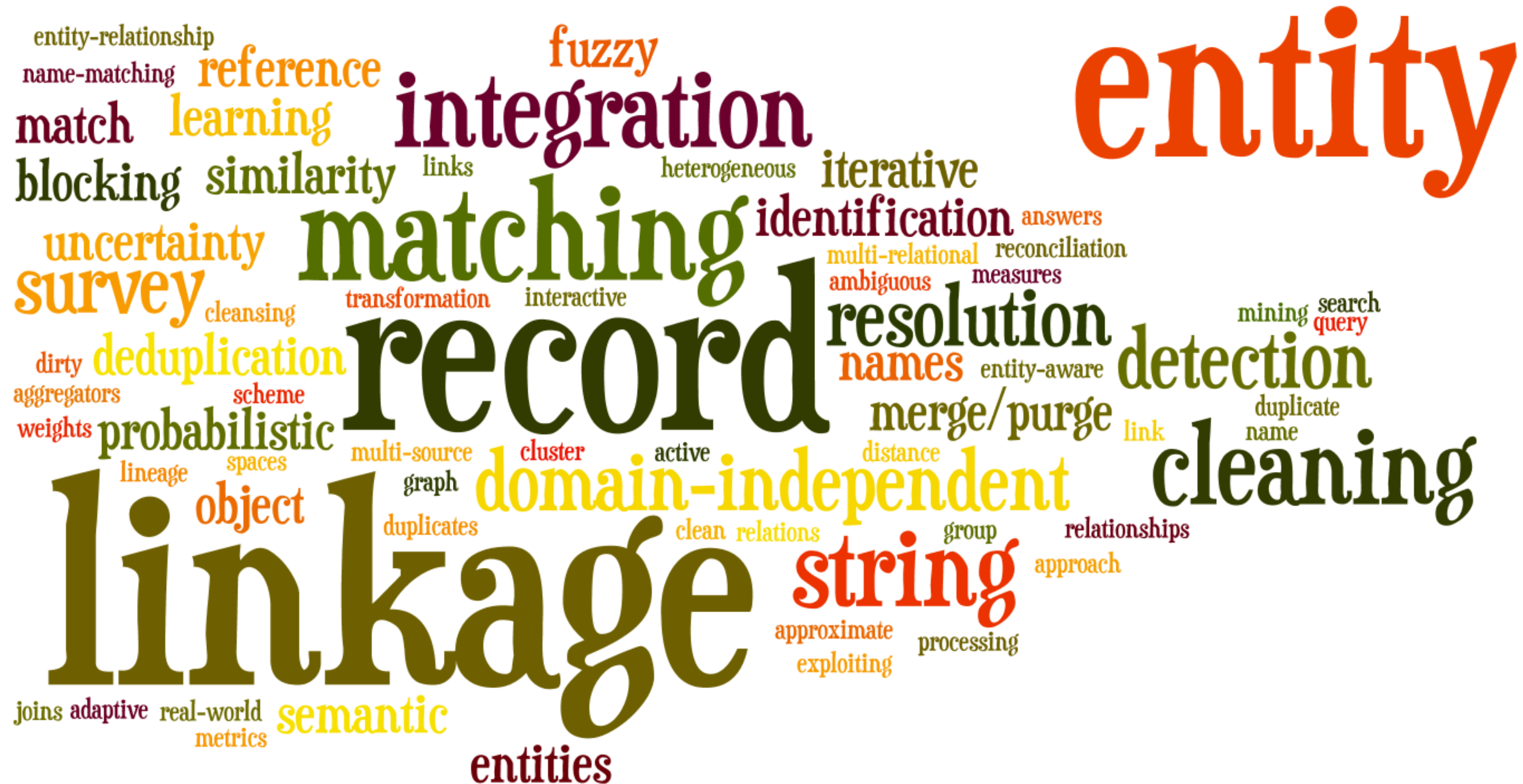
- Identify data describing the same real-world objects
- Decide how to merge the data
- Update the data collection

Solutions following various directions

We present them through four categories:

1. Atomic similarity methods
2. Similarity methods for sets
3. Facilitating inner-relationships
4. Methods in uncertain data

Alternative names for Entity Resolution



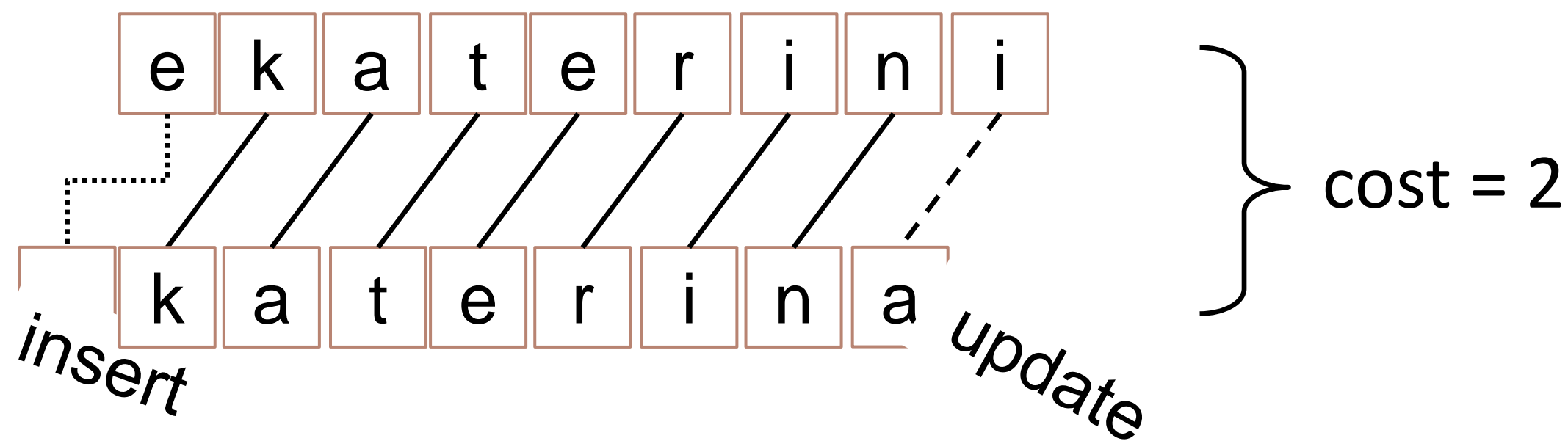
1. Motivation: Entity Resolution
- 2. Atomic similarity methods**
3. Similarity methods for sets
4. Facilitating inner-relationships
5. Methods in uncertain data
6. Conclusions

Examples of targeting cases:

- Publication authors: “John D. Smith” vs. “J. D. Smith”
- Journal names: “Transactions on Knowledge and Data Engineering” vs. “Trans. Knowl. Data Eng.”

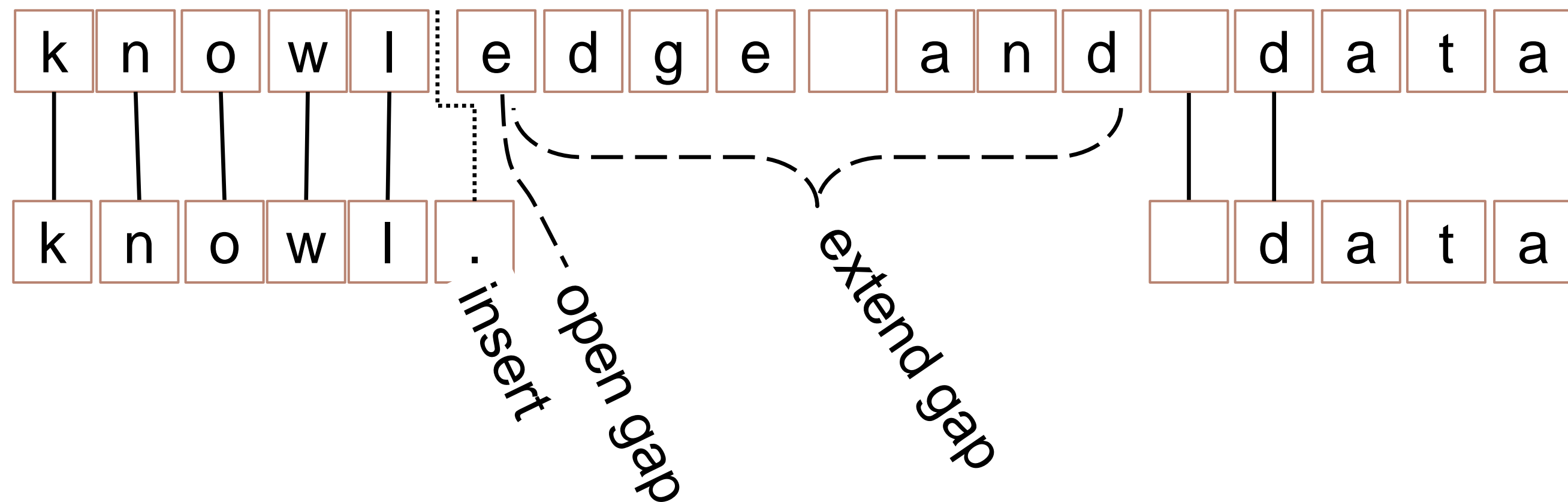
Edit Distance:

- Number of operations to convert from 1st to 2nd string
- Operations in Levenstein distance [\[Lev66\]](#)
 → delete, insert, and update a character with cost 1



Gap Distance:

- Overcome limitation of edit distance with shortened strings
- Considers two extra operations [\[Nav01\]](#)
 - open gap, and extend gap (with small cost)



$$\text{cost} = 1 + o + 8e$$

Jaro similarity [\[Jar89\]](#):

- Small string, e.g., first and last names

$$\text{JaroSim}(s_1, s_2) = \frac{1}{3} \left(\frac{C}{|s_1|} + \frac{C}{|s_2|} + \frac{C-T}{C} \right)$$

$C \rightarrow$ common characters in s_1 and s_2

$T \rightarrow$ transpositions/2 transposition is a k in which $s_1[k] \neq s_2[k]$

Example: “DEIS” vs. “DESI”

$$C=4, T=2/2, \text{JaroSim} = \frac{1}{3} \left(\frac{4}{4} + \frac{4}{4} + \frac{4-1}{4} \right) = 0.9167$$

Jaro-Winkler similarity [\[Win99\]](#):

- Extension that gives higher weight to matching prefix
- Increasing it's applicability to names

Soundex:

- Coverts each word into a phonetic encoding by assigning the same code to the string parts that sound the same
- Similarity between the corresponding phonetic encodings

Remarks:

- Surveys: [\[CRF03\]](#), [\[Win06\]](#)
- Existing API with these methods:
 - SecondString: <http://secondstring.sourceforge.net/>
 - SimMetrics: <http://www.dcs.shef.ac.uk/~sam/simmetrics.html>

1. Motivation: Entity Resolution
2. Atomic similarity methods
3. **Similarity methods for sets**
4. Facilitating inner-relationships
5. Methods in uncertain data
6. Conclusions

Database community:

- Each record is an entity
- A simple example:

	<u>Name</u>	<u>Email</u>	<u>Journal</u>
e1	John D. Smith	smith@uni.edu	Transactions on Knowledge and Data Engineering
e2	Smith, J.	smith@uni.edu	IEEE Trans. Knowl. Data Eng.

Merge-purge [\[HS95\]](#), [\[HS98\]](#):

- Idea: same entities will share information
- Create a key for each record (e.g., email)
- Sort records according to key
- Compare only a limited set of records in each iteration

Using transformations [\[TKM02\]](#):

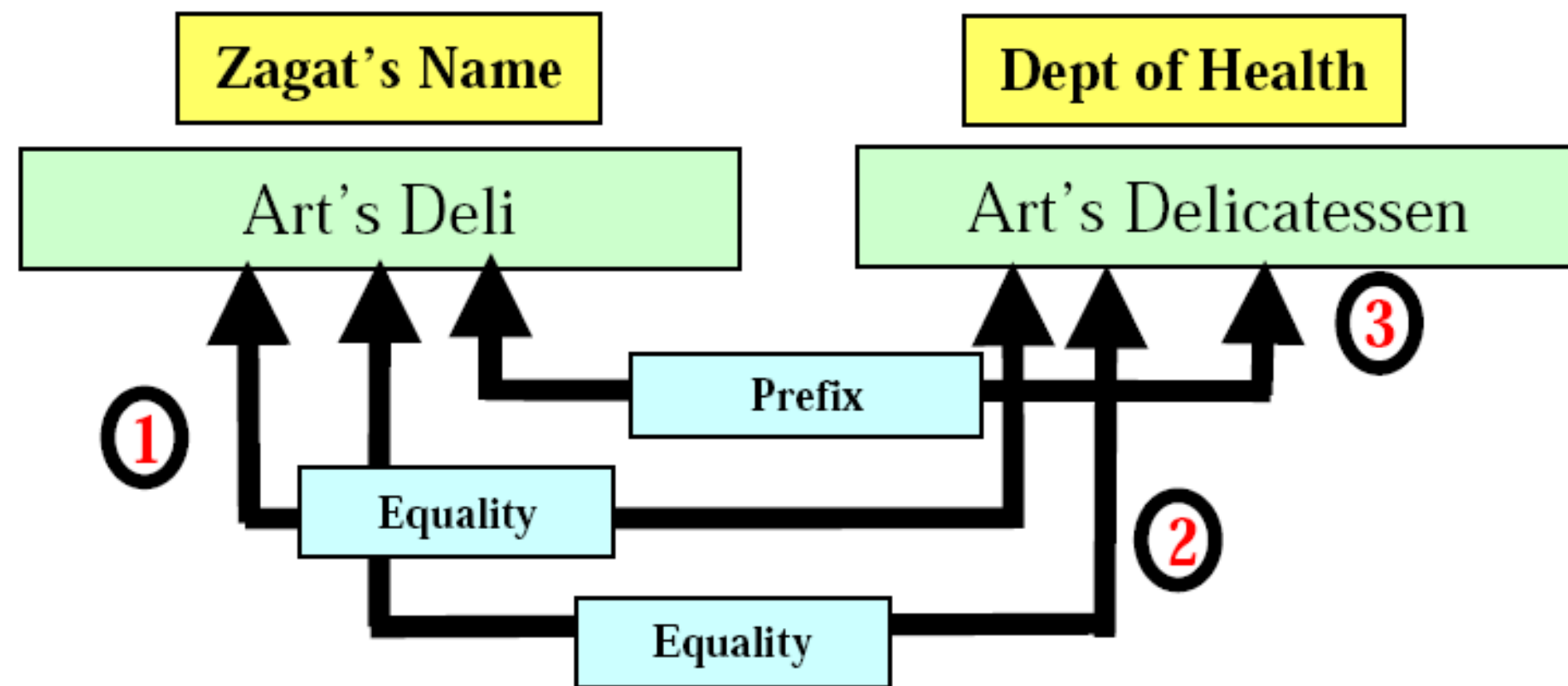
1. Analyze data to generate transformations

■ Unary transform:

- Equality, Stemming, Soundex, Abbreviation (e.g., 3rd or third)

■ N-ary transformations:

- Initial, Prefix, Suffix, Substring
Acronym, Abbreviation, Drop



2. Calculate transformation weights

3. Apply on candidate mappings

Group Linkage [\[OKLS07\]](#):

- Considers groups of relational records
 - not individual relational records
- Groups match when:
 1. High similarity between data of individual records
 2. Large fraction of matching records, i.e., no. 1

Some additional methods

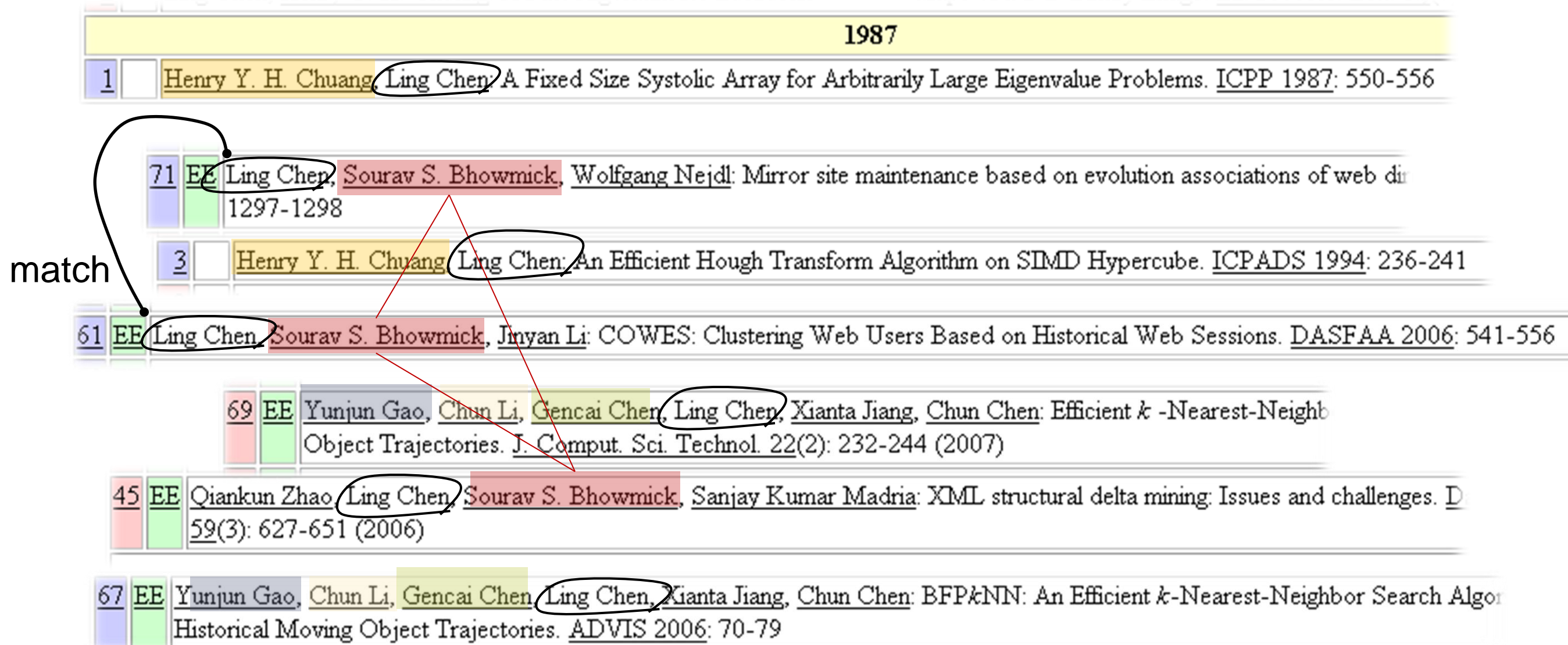
→ [\[DLLH03\]](#)

Surveys for methods in this category

→ [\[DH05\]](#), [\[EIV07\]](#), [\[OS99\]](#)

Remarks:

- Methods do not consider semantics of data
- Currently used as a first step of Entity Resolution



1987

1 Henry Y. H. Chuang, Ling Chen: A Fixed Size Systolic Array for Arbitrarily Large Eigenvalue Problems. *ICPP 1987*: 550-556

71 EE Ling Chen, Sourav S. Bhowmick, Wolfgang Nejdl: Mirror site maintenance based on evolution associations of web dir
1297-1298

3 Henry Y. H. Chuang, Ling Chen: An Efficient Hough Transform Algorithm on SIMD Hypercube. *ICPADS 1994*: 236-241

61 EE Ling Chen, Sourav S. Bhowmick, Jinyan Li: COWES: Clustering Web Users Based on Historical Web Sessions. *DASFAA 2006*: 541-556

69 EE Yunjun Gao, Chun Li, Gencai Chen, Ling Chen, Xianta Jiang, Chun Chen: Efficient k -Nearest-Neighbor
Object Trajectories. *J. Comput. Sci. Technol.* 22(2): 232-244 (2007)

45 EE Qiankun Zhao, Ling Chen, Sourav S. Bhowmick, Sanjay Kumar Madria: XML structural delta mining: Issues and challenges. *D*
59(3): 627-651 (2006)

67 EE Yunjun Gao, Chun Li, Gencai Chen, Ling Chen, Xianta Jiang, Chun Chen: BFP k NN: An Efficient k -Nearest-Neighbor Search Algo
Historical Moving Object Trajectories. *ADVIS 2006*: 70-79

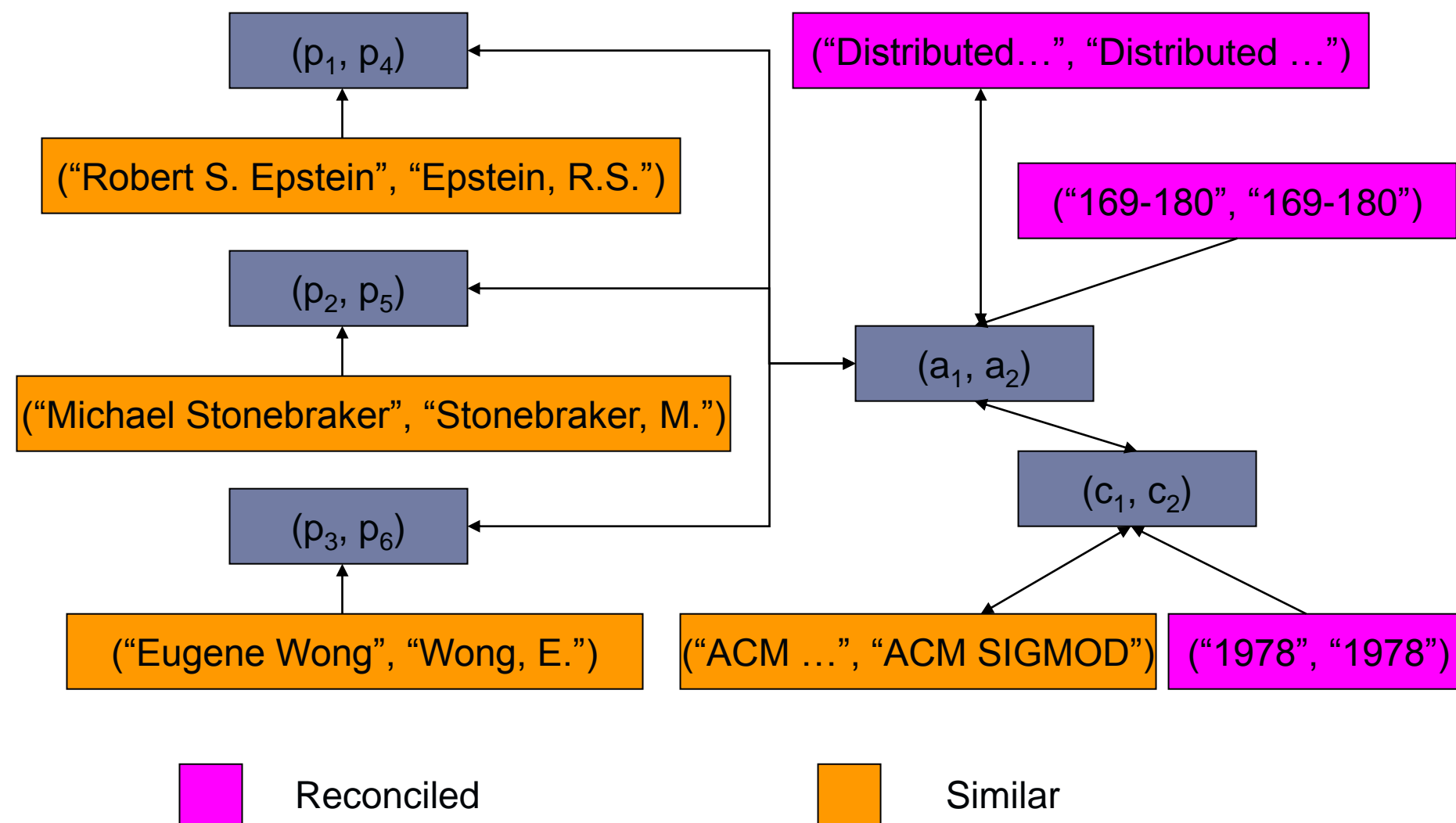
1. Motivation: Entity Resolution
2. Atomic similarity methods
3. Similarity methods for sets
4. **Facilitating inner-relationships**
5. Methods in uncertain data
6. Conclusions

General idea

- Heterogeneous data
 - Lack of schema information
 - Variations in entity descriptions
 - Incomplete or missing values
- Improve effectiveness by considering data semantics
- Example → Reference Reconciliation

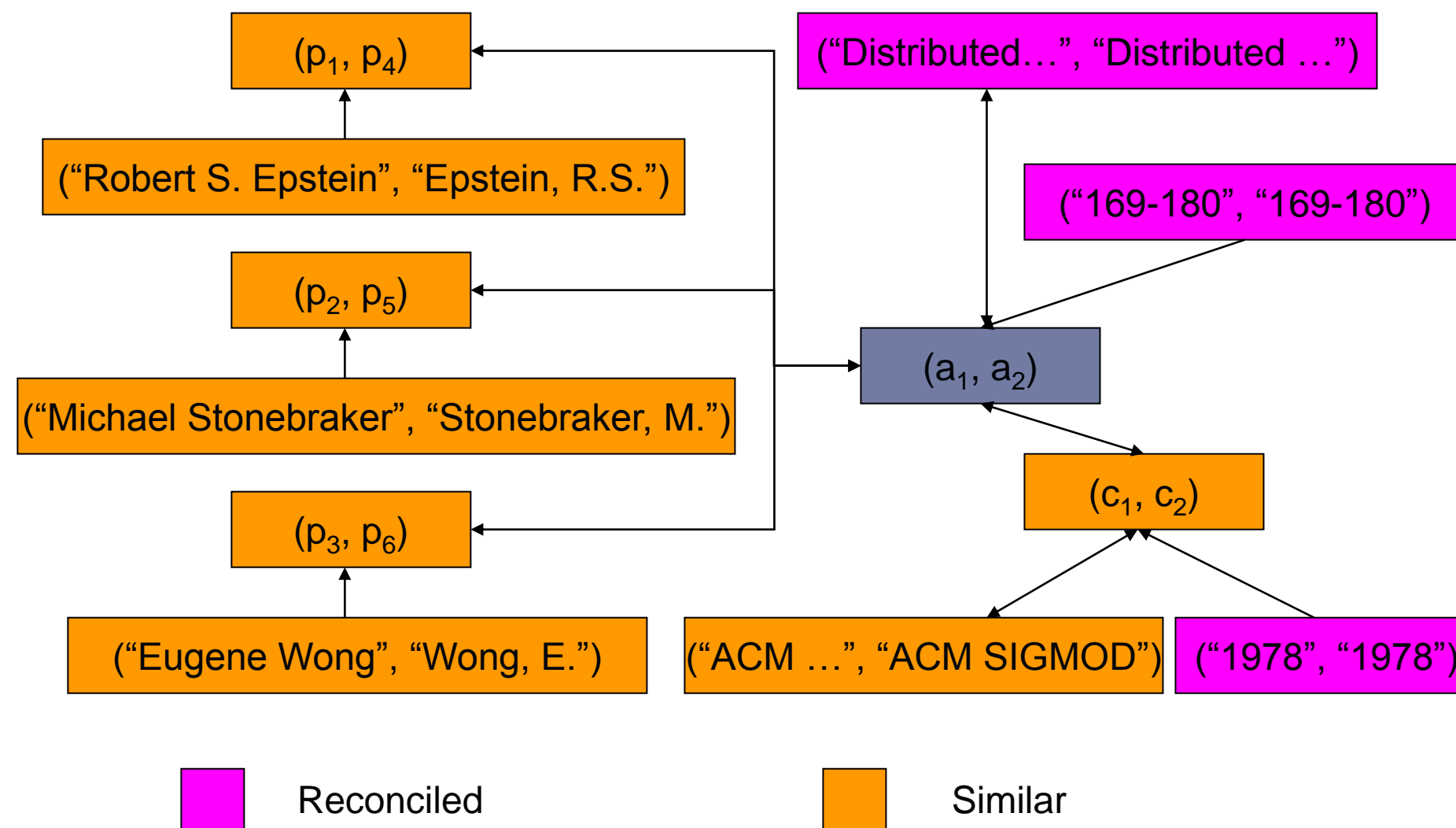
Reference Reconciliation [\[DHM05\]](#)

1. Build a dependency graph



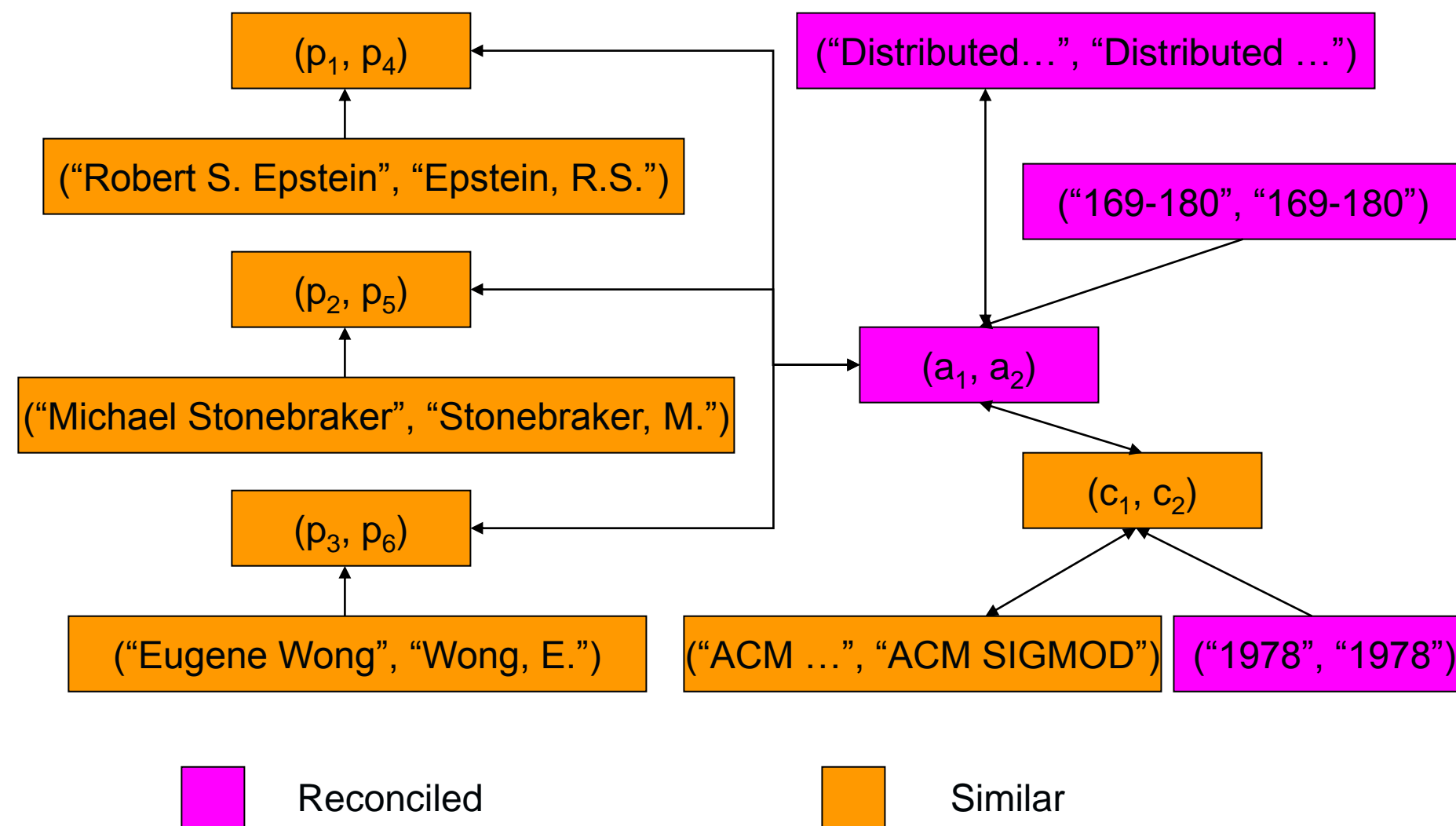
Reference Reconciliation [\[DHM05\]](#)

1. Build a dependency graph
2. Exploit information and relationships



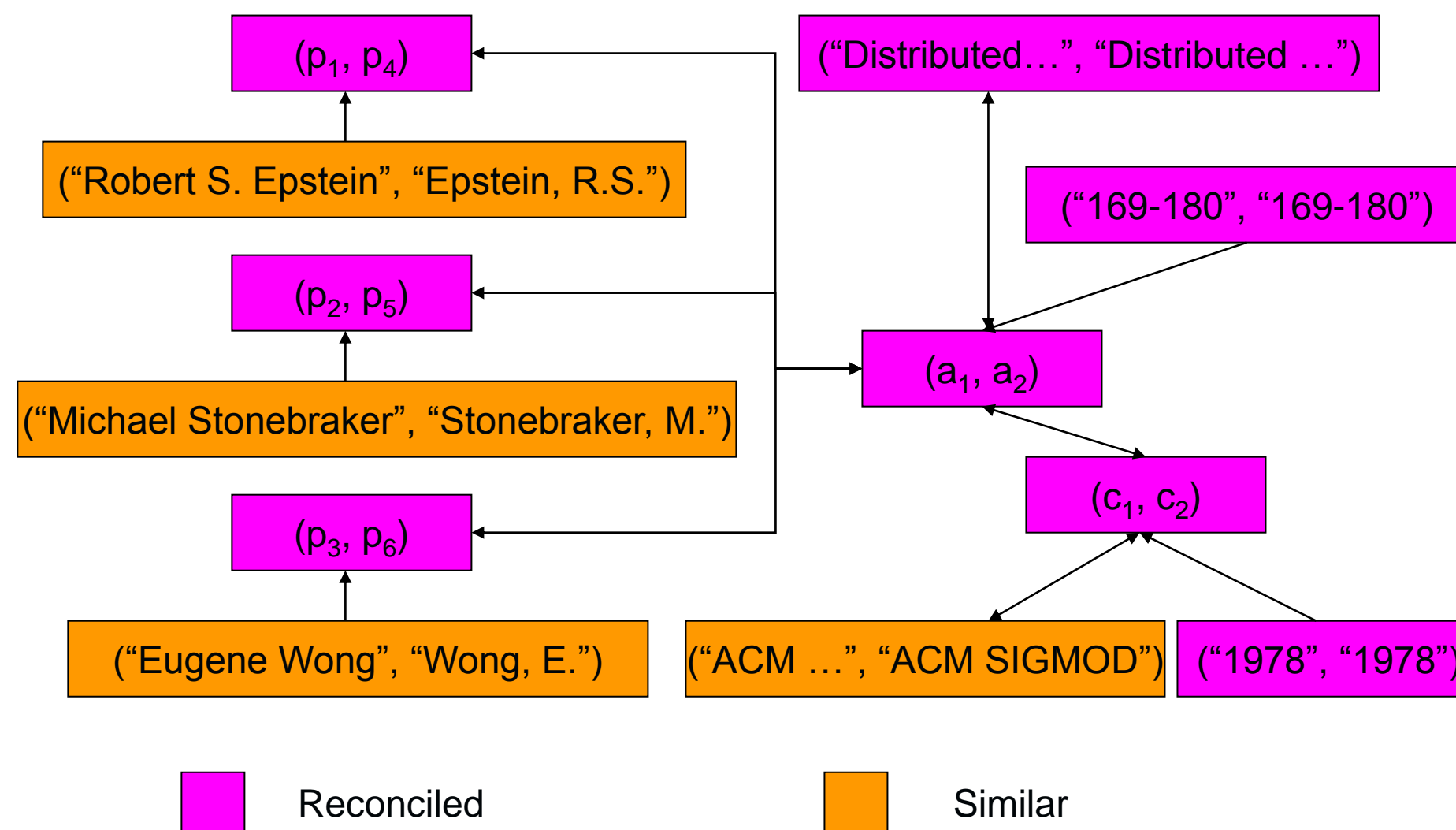
Reference Reconciliation [\[DHM05\]](#)

1. Build a dependency graph
2. Exploit information and relationships



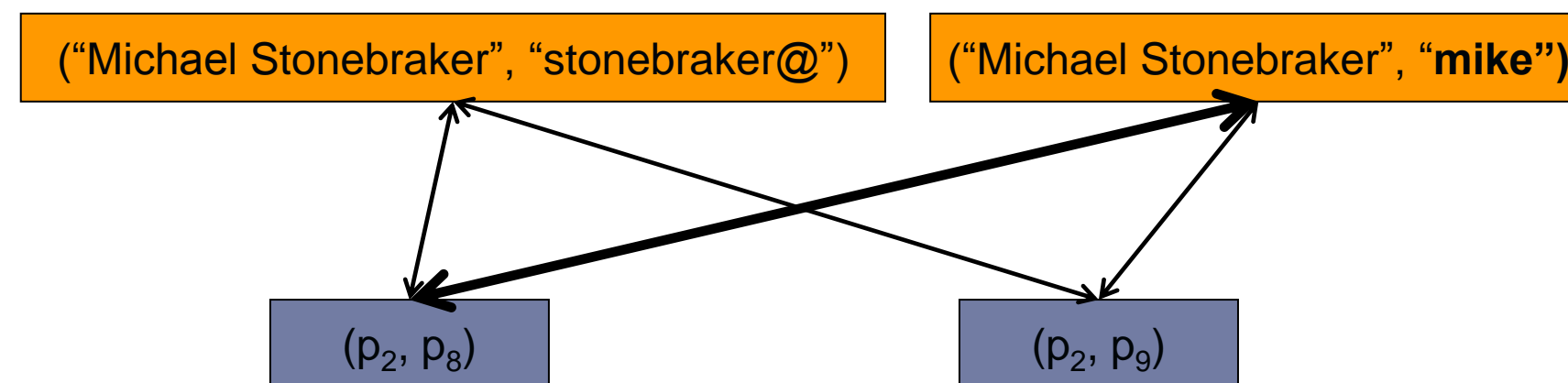
Reference Reconciliation [\[DHM05\]](#)

1. Build a dependency graph
2. Exploit information and relationships



Reference Reconciliation [\[DHM05\]](#)

1. Build a dependency graph
2. Exploit information and relationships
3. Propagate information → enrich relationships



Facilitating inner-relationships

Analysis of entity-relationship graph [\[KM06\]](#), [\[KMC05\]](#):

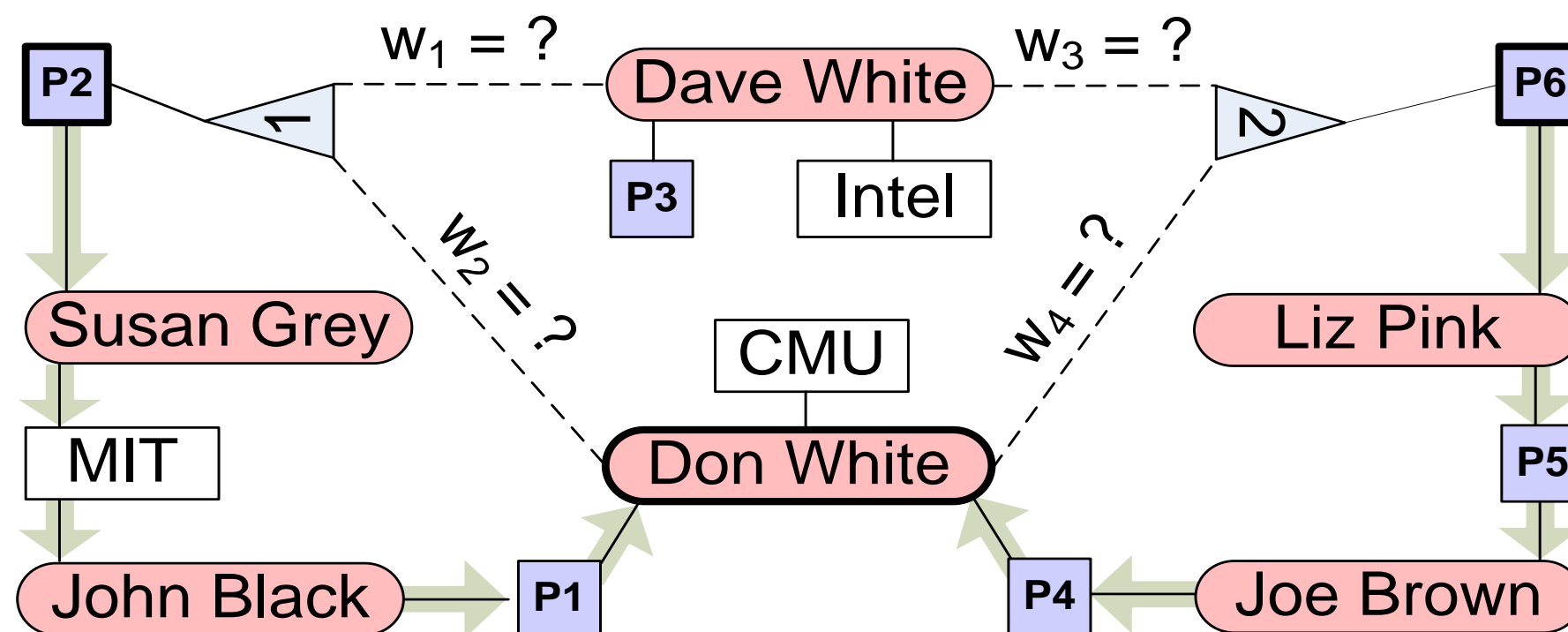
? Author table (clean)	<u>Publication table (to be cleaned)</u>
⟨A1, 'Dave White', 'Intel'⟩	⟨P1, 'Databases . . .', 'John Black', 'Don White'⟩
⟨A2, 'Don White', 'CMU'⟩	⟨P2, 'Multimedia . . .', 'Sue Grey', 'D. White' ⟩
⟨A3, 'Susan Grey', 'MIT'⟩	⟨P3, 'Title3 . . .', 'Dave White'⟩
⟨A4, 'John Black', 'MIT'⟩	⟨P4, 'Title5 . . .', 'Don White', 'Joe Brown'⟩
⟨A5, 'Joe Brown', unknown⟩	⟨P5, 'Title6 . . .', 'Joe Brown', 'Liz Pink'⟩
⟨A6, 'Liz Pink', unknown⟩	⟨P6, 'Title7 . . .', 'Liz Pink', 'D. White' ⟩

Facilitating inner-relationships

Analysis of entity-relationship graph [\[KM06\]](#), [\[KMC05\]](#):

1. Dataset modeled as a graph

Author table (clean)	Publication table (to be cleaned)
<ul style="list-style-type: none"> ⟨A1, 'Dave White', 'Intel'⟩ ⟨A2, 'Don White', 'CMU'⟩ ⟨A3, 'Susan Grey', 'MIT'⟩ ⟨A4, 'John Black', 'MIT'⟩ ⟨A5, 'Joe Brown', unknown⟩ ⟨A6, 'Liz Pink', unknown⟩ 	<ul style="list-style-type: none"> ⟨P1, 'Databases . . .', 'John Black', 'Don White'⟩ ⟨P2, 'Multimedia . . .', 'Sue Grey', 'D. White'⟩ ⟨P3, 'Title3 . . .', 'Dave White'⟩ ⟨P4, 'Title5 . . .', 'Don White', 'Joe Brown'⟩ ⟨P5, 'Title6 . . .', 'Joe Brown', 'Liz Pink'⟩ ⟨P6, 'Title7 . . .', 'Liz Pink', 'D. White'⟩

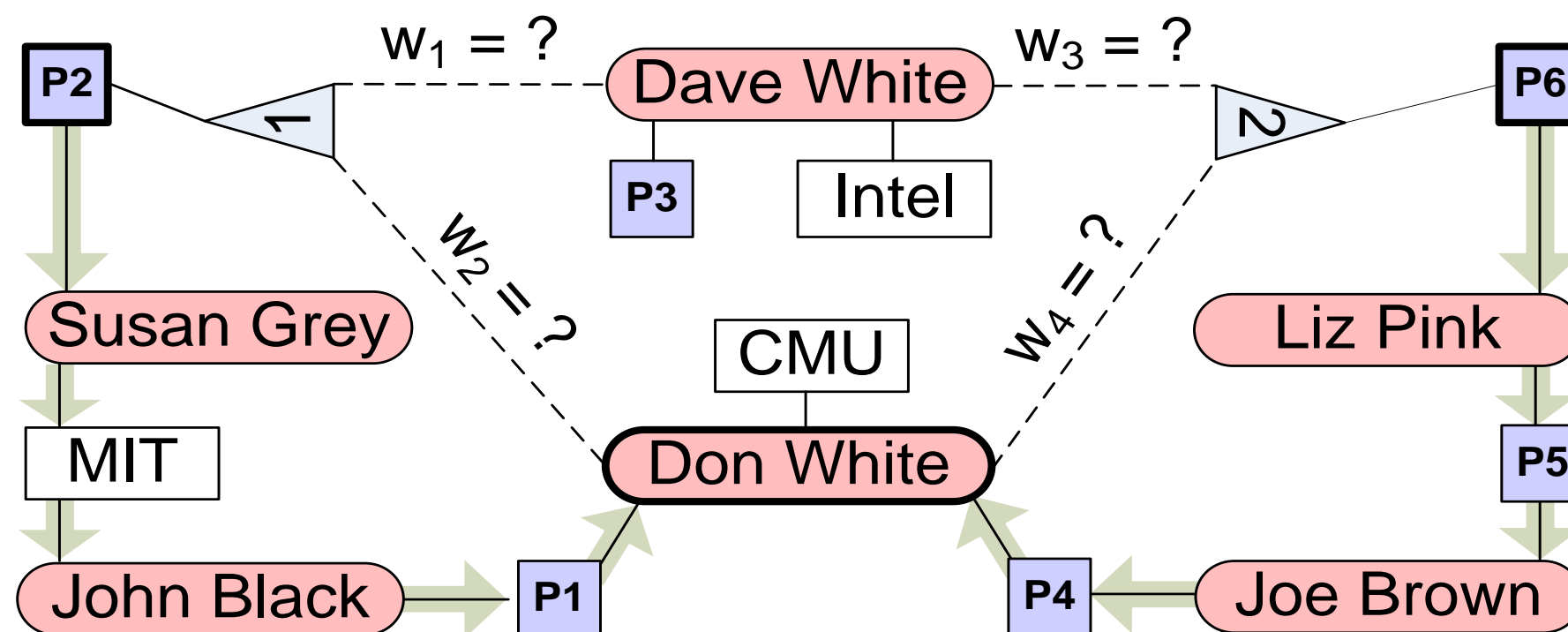


Facilitating inner-relationships

Analysis of entity-relationship graph [\[KM06\]](#), [\[KMC05\]](#):

1. Dataset modeled as a graph
2. Data more strongly connected when sharing relationships

Author table (clean)	Publication table (to be cleaned)
<p>?</p> <p>⟨A1, 'Dave White', 'Intel'⟩</p> <p>⟨A2, 'Don White', 'CMU'⟩</p> <p>⟨A3, 'Susan Grey', 'MIT'⟩</p> <p>⟨A4, 'John Black', 'MIT'⟩</p> <p>⟨A5, 'Joe Brown', unknown⟩</p> <p>⟨A6, 'Liz Pink', unknown⟩</p>	<p>⟨P1, 'Databases . . .', 'John Black', 'Don White'⟩</p> <p>⟨P2, 'Multimedia . . .', 'Sue Grey', 'D. White'⟩</p> <p>⟨P3, 'Title3 . . .', 'Dave White'⟩</p> <p>⟨P4, 'Title5 . . .', 'Don White', 'Joe Brown'⟩</p> <p>⟨P5, 'Title6 . . .', 'Joe Brown', 'Liz Pink'⟩</p> <p>⟨P6, 'Title7 . . .', 'Liz Pink', 'D. White'⟩</p>

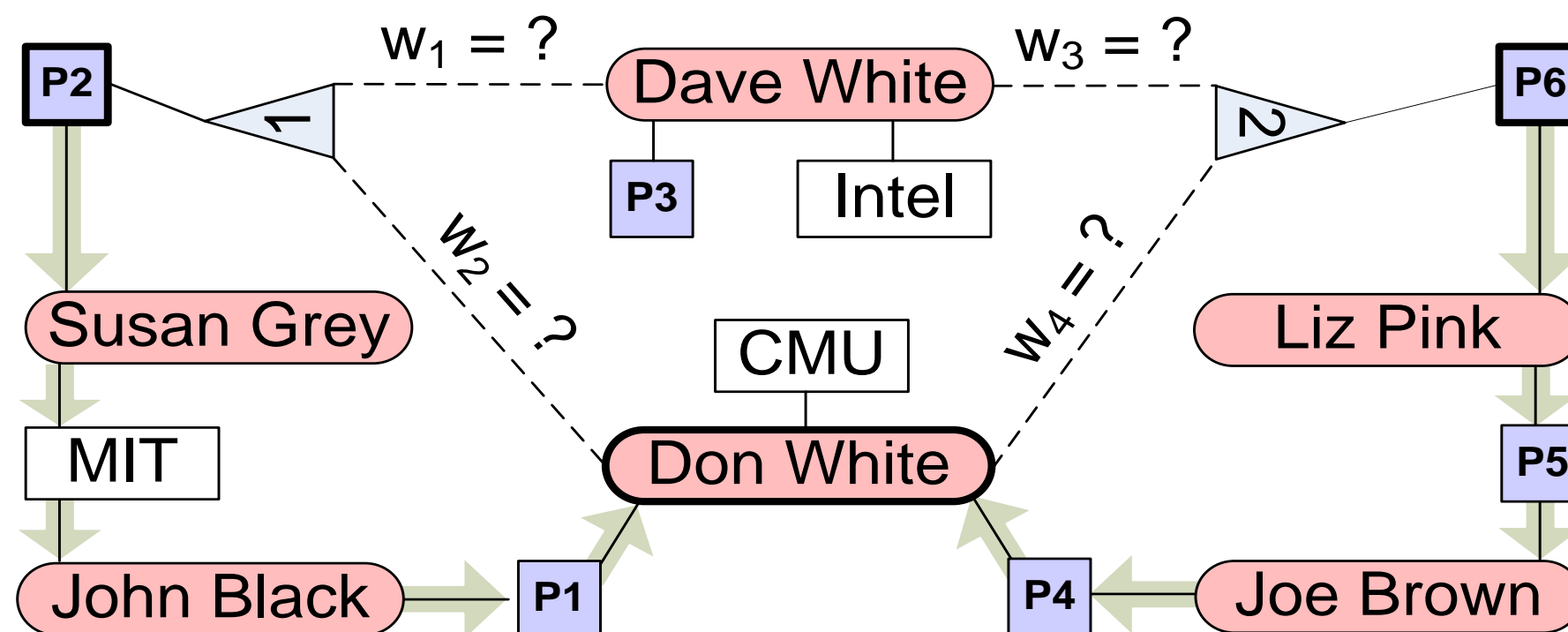


Facilitating inner-relationships

Analysis of entity-relationship graph [\[KM06\]](#), [\[KMC05\]](#):

1. Dataset modeled as a graph
2. Data more strongly connected when sharing relationships
3. Measure the connection strengths (details in paper)

Author table (clean)	Publication table (to be cleaned)
<p>?</p> <p>⟨A1, 'Dave White', 'Intel'⟩</p> <p>⟨A2, 'Don White', 'CMU'⟩</p> <p>⟨A3, 'Susan Grey', 'MIT'⟩</p> <p>⟨A4, 'John Black', 'MIT'⟩</p> <p>⟨A5, 'Joe Brown', unknown⟩</p> <p>⟨A6, 'Liz Pink', unknown⟩</p>	<p>⟨P1, 'Databases . . .', 'John Black', 'Don White'⟩</p> <p>⟨P2, 'Multimedia . . .', 'Sue Grey', 'D. White'⟩</p> <p>⟨P3, 'Title3 . . .', 'Dave White'⟩</p> <p>⟨P4, 'Title5 . . .', 'Don White', 'Joe Brown'⟩</p> <p>⟨P5, 'Title6 . . .', 'Joe Brown', 'Liz Pink'⟩</p> <p>⟨P6, 'Title7 . . .', 'Liz Pink', 'D. White'⟩</p>



Some additional methods:

- Relationship-based clustering [\[BG04a\]](#), [\[BG04b\]](#):
 - Common references for a match increase our belief
 - For this we need to identify common references
 - Iterative process: common matches → identifying additional matches
- Incremental & adaptive [\[INN08\]](#), [\[MPC+10\]](#):
 - Targets data that are constantly changing and evolving
 - Bayesian network to model entities, relationships, and evidences (possible linkages)
 - Enables flexible update of the network

Surveys for methods in this category

→ [\[GD05\]](#), [\[KSS06\]](#)

1. Motivation: Entity Resolution
2. Atomic similarity methods
3. Similarity methods for sets
4. Facilitating inner-relationships
5. **Methods in uncertain data**
6. Conclusions

General idea:

- Keep conflicting relations, e.g., [\[AFM06\]](#), [\[RDS07\]](#), [\[DS07a\]](#), [\[DHY07\]](#)
 - Lack of resolution rules to correctly resolve and merge relations
 - No merging, but maintain results in the database
 - Relations are alternative representations of the same real world object
- Entity representation with probability – indicates...
 - Reliability of the source
 - Output of the matching process
 - Etc.

customer				
	<u>custId</u>	name	income	prob
s_1	c1	John	\$120K	0.9
s_2	c1	John	\$80K	0.1
s_3	c2	Mary	\$140K	0.4
s_4	c2	Marion	\$40K	0.6

Clean answers over dirty databases [\[AFM06\]](#):

- Dirty database represents several possible databases
- Result set for queries should include the entity resolution results
- Query rewriting mechanism with efficient computation of probability for each answer

order	id	orderId	custFk	clIdFk	quantity	prob
t_1	o1	11	m1	c1	3	1
t_2	o2	12	m2	c1	2	0.5
t_3	o2	13	m3	c2	5	0.5

customer	id	custId	name	balance	prob
t_4	c1	m1	John	\$20K	0.7
t_5	c1	m2	John	\$30K	0.3
t_6	c2	m3	Mary	\$27K	0.2
t_7	c2	m4	Marion	\$5K	0.8

$$D_1^{cd} = \{t_1, t_2, t_4, t_6\}$$

$$D_2^{cd} = \{t_1, t_2, t_4, t_7\}$$

$$D_3^{cd} = \{t_1, t_2, t_5, t_6\}$$

$$D_4^{cd} = \{t_1, t_2, t_5, t_7\}$$

$$D_5^{cd} = \{t_1, t_3, t_4, t_6\}$$

$$D_6^{cd} = \{t_1, t_3, t_4, t_7\}$$

$$D_7^{cd} = \{t_1, t_3, t_5, t_6\}$$

$$D_8^{cd} = \{t_1, t_3, t_5, t_7\}$$

Clean answers over dirty databases [\[AFM06\]](#):

- Query rewriting

```
select  $A_1, \dots, A_n$ 
from  $R_1, \dots, R_m$ 
where  $W$ 
→ select  $A_1, \dots, A_n, \text{sum}(R_1.\text{prob} * \dots * R_m.\text{prob})$ 
from  $R_1, \dots, R_m$ 
where  $W$ 
group by  $A_1, \dots, A_n$ 
```

- Groups the result by the attributes
- For each group: sums the product of relation probabilities
- (applicable only to rewritable queries)

Entity-Aware querying over prob. linkages [\[INN10\]](#):

- Not merging the entities using threshold
- Keep probabilistic linkages alongside the original data
- Use them during query processing

Query:

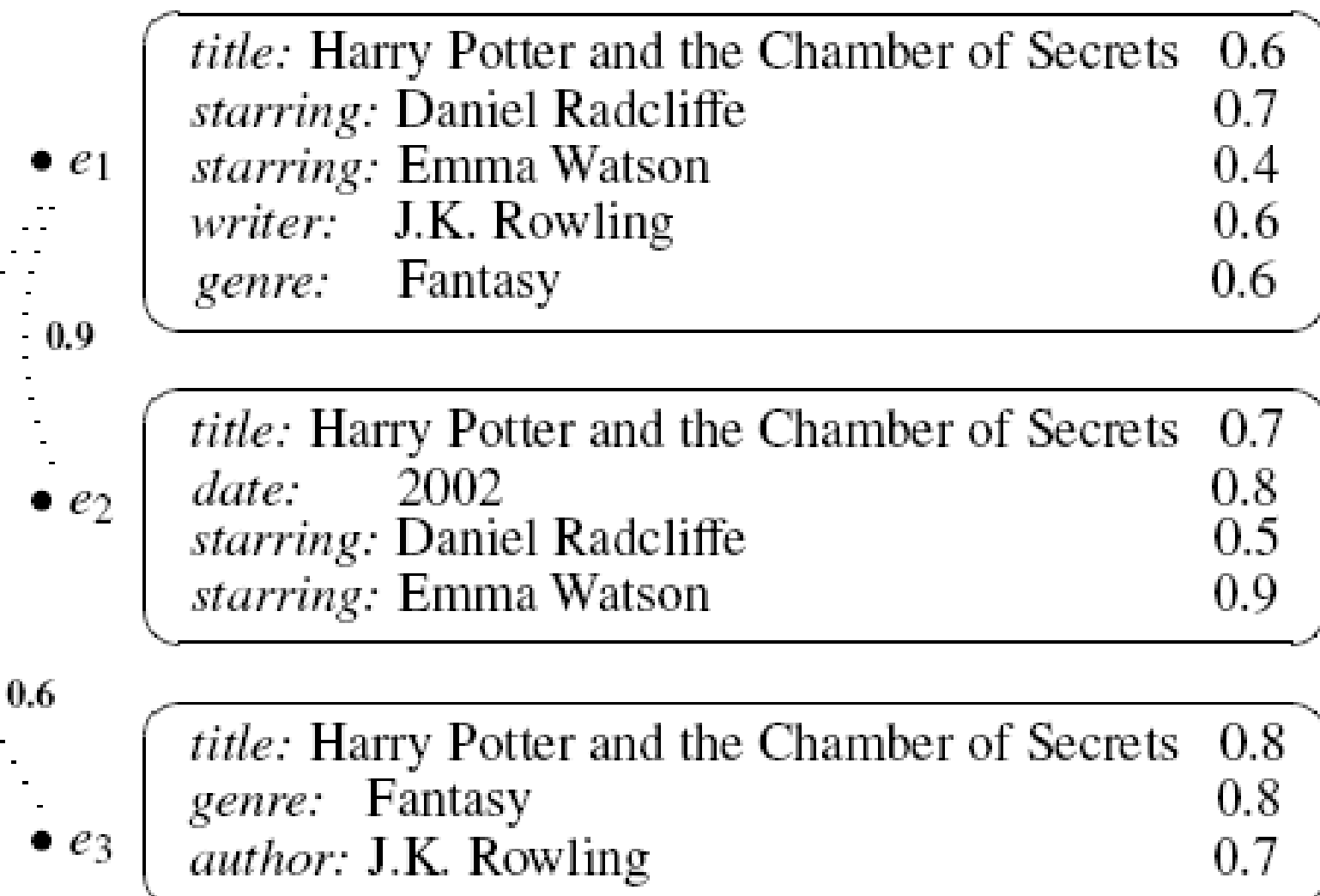
- “J. K. Rowling” movies in “2002”

Assume no linkages:

- zero results

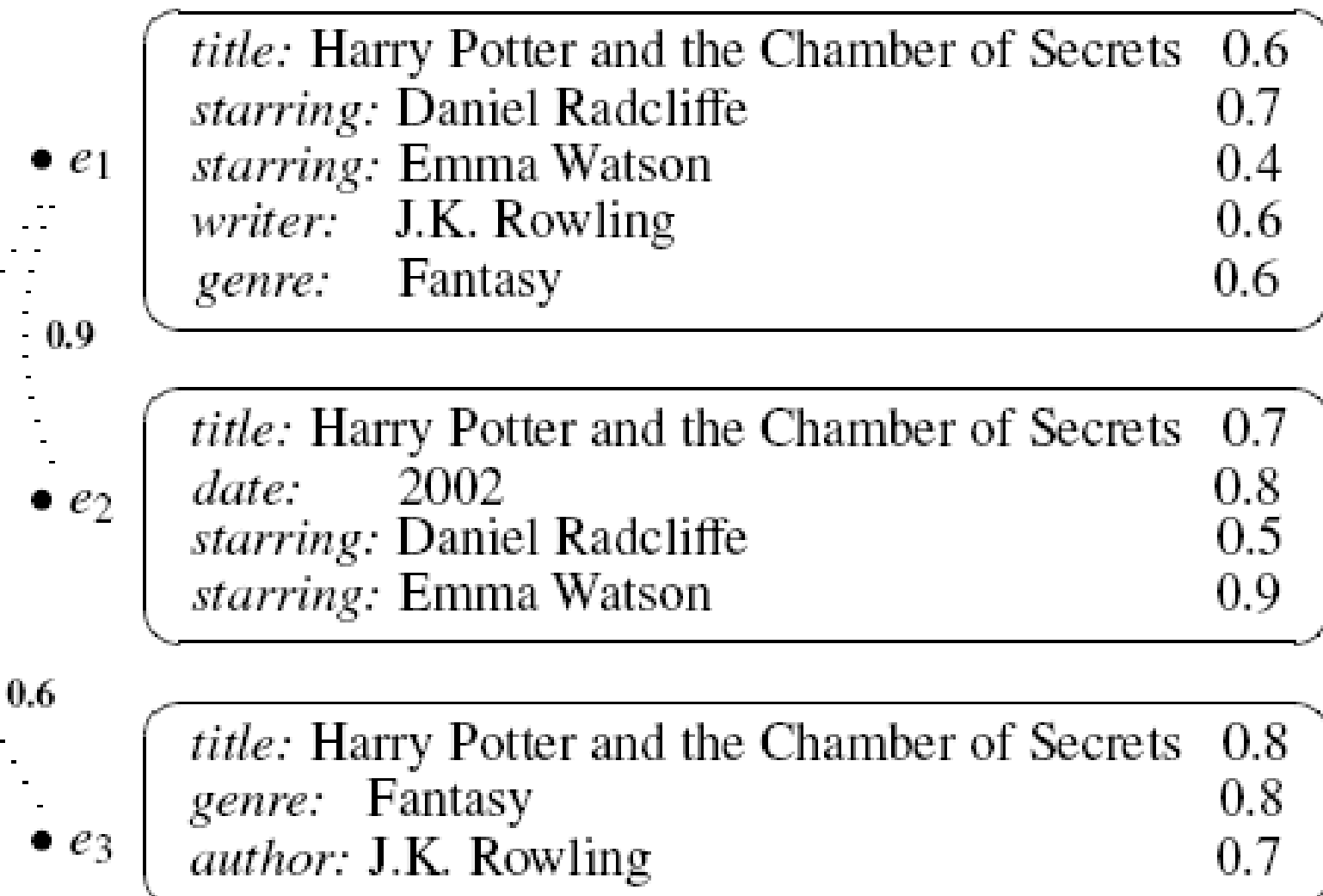
Possible answer with linkages:

- $\text{merge}(e_1, e_2)$
- $\text{merge}(e_1, e_2, e_3)$



Entity-Aware querying over prob. linkages [\[INN10\]](#):

- Linkage prob. represent several possible l -worlds
- Attribute prob. represent several possible worlds
- Efficient query processing:
 - Analyze query conditions
 - Identify the required entity merges
 - Decide useful possible l -worlds
 - Generate possible worlds
 - Compute probability



1. Motivation: Entity Resolution
2. Atomic similarity methods
3. Similarity methods for sets
4. Facilitating inner-relationships
5. Methods in uncertain data
6. **Conclusions**

Discussed methods entity resolution

Four categories of methods

Not presented:

- Blocking mechanisms:
 - Split data into blocks and compare inner-block data
 - Improves efficiency for large-size datasets
 - Examples: [\[WMK+09\]](#), [\[PINF11\]](#)
- Active learning approaches:
 - Use a subset of the data to learn matching rules
 - Apply the rules to remaining data
 - Examples: [\[SB02\]](#), [\[CR01\]](#)
- Similarity Joins [\[GIJ+1\]](#)
- Schema matching
-

- [AFM06] Periklis Andritsos, Ariel Fuxman, and Renée J. Miller. Clean answers over dirty databases: A probabilistic approach. In ICDE, 2006.
- [BG04a] Indrajit Bhattacharya and Lise Getoor. Deduplication and group detection using links. In LinkKDD, 2004.
- [BG04b] Indrajit Bhattacharya and Lise Getoor. Iterative record linkage for cleaning and integration. In DMKD, pages 11–18, 2004.
- [BMC+03] Mikhail Bilenko, Raymond J. Mooney, William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5):16–23, 2003.
- [CR01] W. Cohen and J. Richman. Learning to match and cluster entity names. In MF/IR Workshop co-located with SIGIR, 2001.
- [CRF03] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. A Comparison of String Distance Metrics for Name-Matching Tasks. In IIWeb co-located with IJCAI, pages 73–78, 2003.
- [DH05] AnHai Doan and Alon Y. Halevy. Semantic integration research in the database community: A brief survey. *AI Magazine*, 26(1):83–94, 2005.
- [DHM05] Xin Dong, Alon Halevy, and Jayant Madhavan. Reference Reconciliation in Complex Information Spaces. In SIGMOD, pages 85–96, 2005.
- [DHY07] Xin Luna Dong, Alon Y. Halevy, and Cong Yu. Data integration with uncertainty. In VLDB, pages 687–698, 2007.
- [DLLH03] AnHai Doan, Ying Lu, Yoonkyong Lee, and Jiawei Han. Object matching for information integration: A profiler-based approach. In IIWeb co-located with IJCAI, pages 53–58, 2003.
- [DS07a] Nilesh N. Dalvi and Dan Suciu. Management of probabilistic data: foundations and challenges. In PODS, pages 1–12, 2007.
- [EIV07] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1–16, 2007.
- [GD05] Lise Getoor and Christopher P. Diehl. Link mining: a survey. *SIGKDD Explorations*, 7(2):3–12, 2005.

Bibliography (II)

- [GIJ+01] Luis Gravano, Panagiotis G. Ipeirotis, H. V. Jagadish, Nick Koudas, S. Muthukrishnan, and Divesh Srivastava. Approximate string joins in a database (almost) for free. In VLDB, pages 491–500, 2001.
- [GM03] Ramanathan V. Guha and Rob McCool. TAP: a SemanticWeb Platform. *Computer Networks*, 42(5):557–577, 2003.
- [HS95] Mauricio A. Hernández and Salvatore J. Stolfo. The merge/purge problem for large databases. In SIGMOD Conference, pages 127–138, 1995.
- [HS98] Mauricio A. Hernández and Salvatore J. Stolfo. Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Min. Knowl. Discov.*, 2(1):9–37, 1998.
- [INN08] Ekaterini Ioannou, Claudia Niederée, and Wolfgang Nejdl. Probabilistic entity linkage for heterogeneous information spaces. In CAiSE, pages 556–570, 2008.
- [INNV10] Ekaterini Ioannou, Wolfgang Nejdl, Claudia Niederée, and Yannis Velegrakis. On-the-fly entity-aware query processing in the presence of linkage. *PVLDB*, 3(1):429–438, 2010.
- [Jar89] Matthew A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *American Statistical Association*, 84, 1989.
- [KM06] Dmitri V. Kalashnikov and Sharad Mehrotra. Domain-independent data cleaning via analysis of entity-relationship graph. *ACM TODS*, 31(2):716–767, 2006.
- [KMC05] Dmitri V. Kalashnikov, Sharad Mehrotra, and Zhaoqi Chen. Exploiting relationships for domain-independent data cleaning. In SIAM SDM, 2005.
- [KSS06] Nick Koudas, Sunita Sarawagi, and Divesh Srivastava. Record linkage: similarity measures and algorithms. In SIGMOD Conference, pages 802–803, 2006.
- [Lev66] V. I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, vol. 10, no. 8, pages 707-710, 1966.
- [MPC+10] Enrico Minack, Raluca Paiu, Stefania Costache, Gianluca Demartini, Julien Gaugaz, Ekaterini Ioannou, Paul-Alexandru Chirita, and Wolfgang Nejdl. Leveraging personal metadata for desktop search: The beagle++ system. *Journal of Web Semantics*, 8(1):37–54, 2010.

Bibliography (III)

- [Nav01] Gonzalo Navarro. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88, 2001.
- [OKLS07] Byung-Won On, Nick Koudas, Dongwon Lee, Divesh Srivastava. Group Linkage. In *ICDE*, pages 496-505, 2007.
- [OS99] Aris M. Ouksel and Amit P. Sheth. Semantic interoperability in global information systems: A brief introduction to the research area and the special section. *SIGMOD Record*, 28(1):5–12, 1999.
- [PD04] Parag and P. Domingos. Multi-relational record linkage. In *MRDM Workshop co-located with KDD*, pages 31–48, 2004.
- [PINF11] George Papadakis, Ekaterini Ioannou, Claudia Niederée, and Peter Fankhauser. Efficient entity resolution for large heterogeneous information spaces. In *WSDM*, 2011.
- [RDS07] Christopher Re, Nilesh N. Dalvi, and Dan Suciu. Efficient top-k query evaluation on probabilistic data. In *ICDE*, pages 886–895, 2007.
- [RVMB09] Flavio Rizzolo, Yannis Velegrakis, John Mylopoulos, Siarhei Bykau: Modeling Concept Evolution: A Historical Perspective. In *ER*, pages 331-345, 2009.
- [SB02] Sunita Sarawagi and Anuradha Bhamidipaty. Interactive deduplication using active learning. In *KDD*, pages 269–278, 2002.
- [TKM02] Sheila Tejada, Craig A. Knoblock, and Steven Minton. Learning domain-independent string transformation weights for high accuracy object identification. In *KDD*, pages 350–359, 2002.
- [Win99] William Winkler. The state of record linkage and current research problems, 1999.
- [Win06] William Winkler. Overview of Record Linkage and Current Research Directions. Bureau of the Census, 2006.
- [WMK+09] Steven Euijong Whang, David Menestrina, Georgia Koutrika, Martin Theobald, and Hector Garcia-Molina. Entity resolution with iterative blocking. In *SIGMOD Conference*, pages 219–232, 2009.