Information Systems Group

Hasso Plattner Institut | Universität Potsdam

Duplicate Detection

6.6.2013

Felix Naumann

Overview

- Duplicate detection
- Similarity measures
- Algorithms
- Data sets and evaluation
- Data fusion

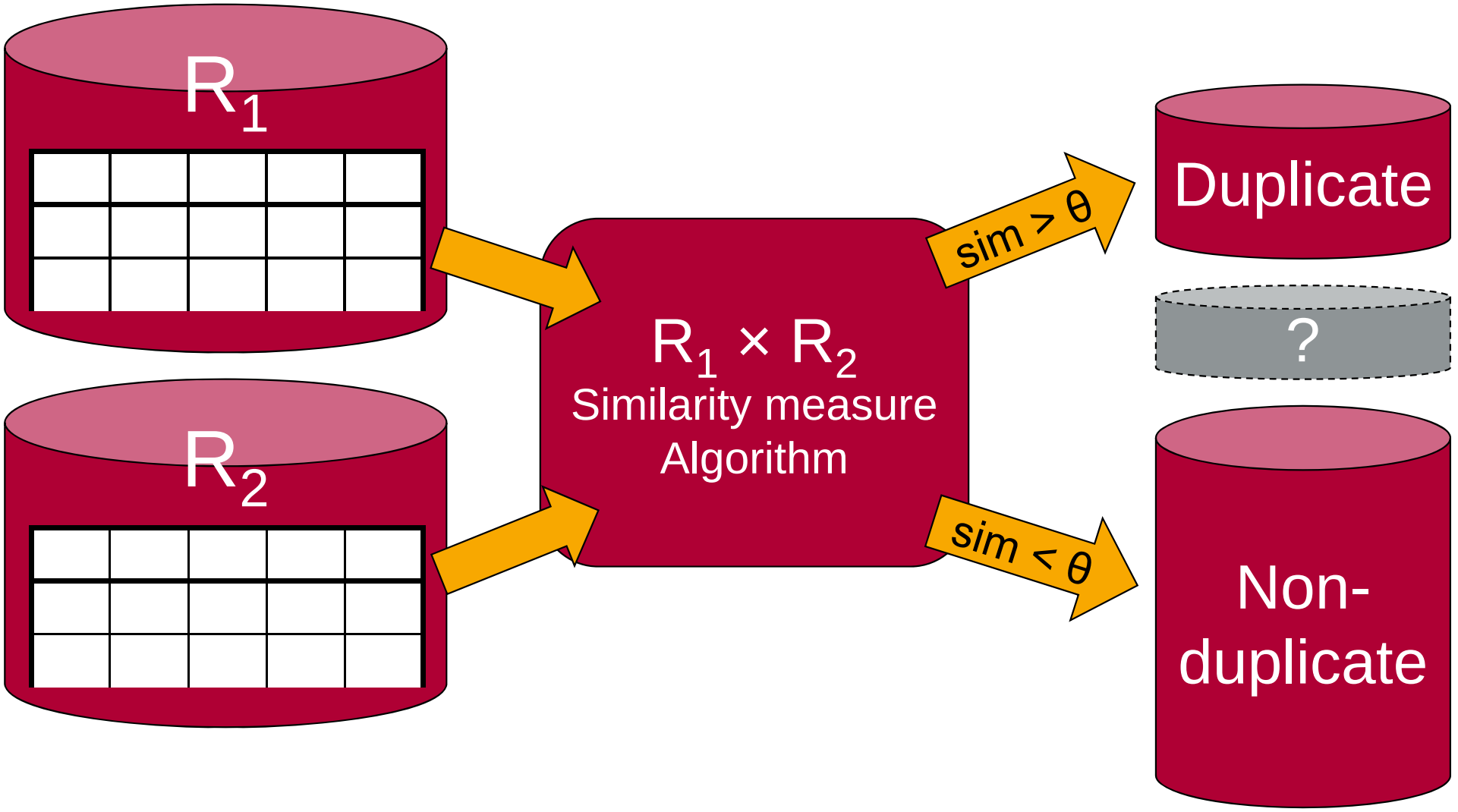
Duplicate Detection

Duplicate detection is the discovery of multiple representations of the same real-world object.

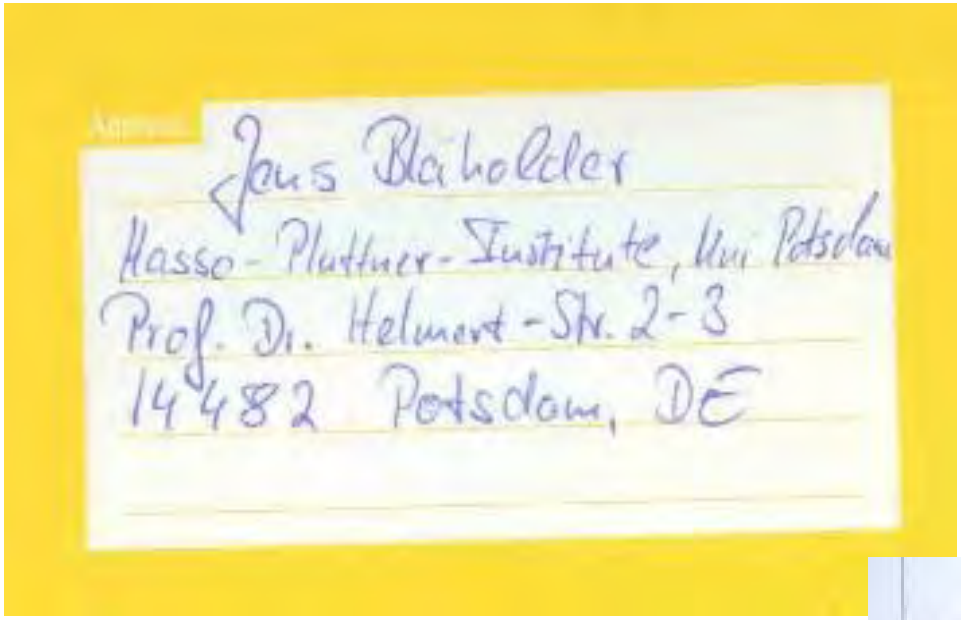
- Problem 1: Representations are not identical.
  - *Fuzzy duplicates*
- Solution: Similarity measures
  - Value- and record-comparisons
  - Domain-dependent or domain-independent

- Problem 2: Data sets are large.
  - Quadratic complexity: Comparison of every pair of records.
- Solution: Algorithms
  - E.g., avoid comparisons by partitioning.

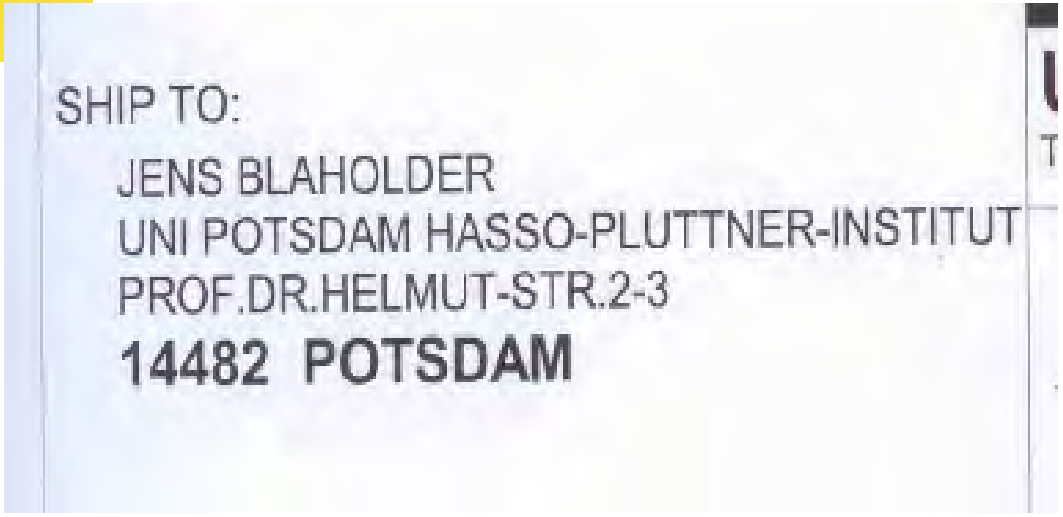
Duplicate Detection



Origins of duplicates

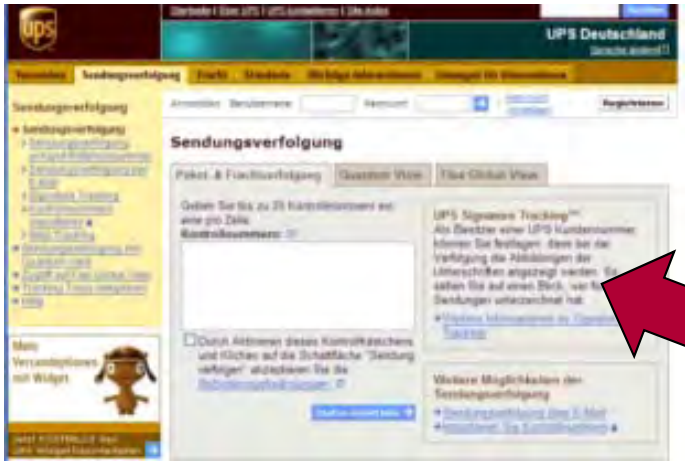


Original

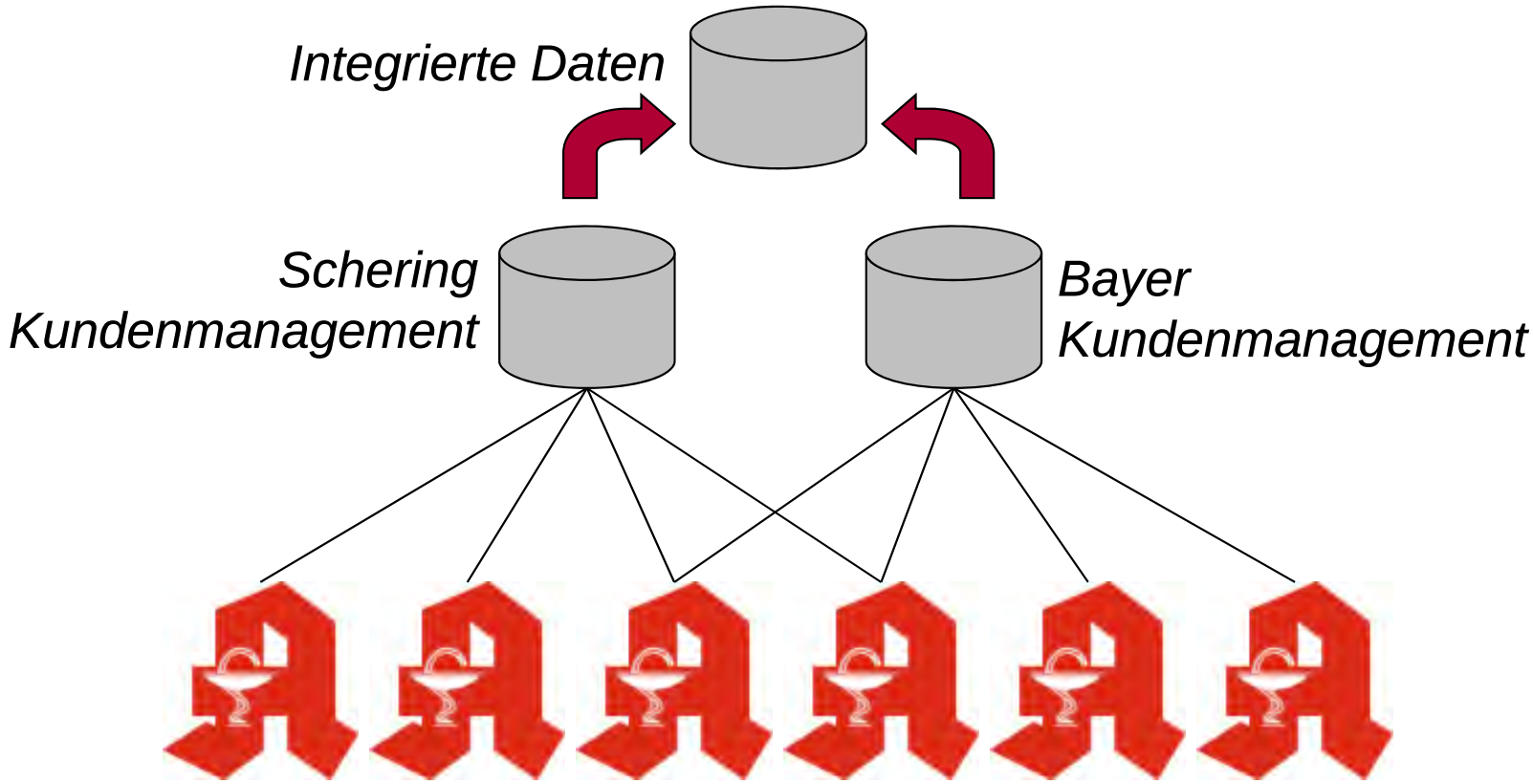


Scanned

Origins of duplicates



Origins of duplicates



Difficult names

488941 britney spears	29 britent spears	9 brinttany spears	5 brney spears	3 britiy sp
40134 brittany spears	29 brittnany spears	9 britanay spears	5 broitney spears	3 britmeny
36315 brittney spears	29 britttany spears	9 britinany spears	5 brotny spears	3 britneeey
24342 britany spears	29 btiney spears	9 britn spears	5 bruteny spears	3 britnehy
7331 britny spears	26 birttney spears	9 britnew spears	5 btiyney spears	3 britnely
6633 briteny spears	26 breitney spears	9 britneyn spears	5 btrittney spears	3 britnesy
2696 britteny spears	26 brinity spears	9 britrney spears	5 gritney spears	3 britnetty
1807 briney spears	26 britenay spears	9 brtiny spears	5 spritney spears	3 britnex s
1635 brittny spears	26 britneyt spears	9 brtittney spears	4 bittny spears	3 britneyxx
1479 brintey spears	26 brittan spears	9 brtny spears	4 bnritney spears	3 britnity
1479 britanny spears	26 brittne spears	9 brytny spears	4 brandy spears	3 britntey
1338 britiny spears	26 btittany spears	9 rbitney spears	4 brbritney spears	3 britnyey
1211 britnet spears	24 beitney spears	8 birtiny spears	4 breatiny spears	3 britterny
1096 britiney spears	24 birteny spears	8 bithney spears	4 breetney spears	3 brittneey
991 britaney spears	24 brightney spears	8 brattany spears	4 bretiney spears	3 brittnney
991 britnay spears	24 brintiny spears	8 breitny spears	4 brfitney spears	3 brittnyey
811 brithney spears	24 britanty spears	8 breteny spears	4 briattany spears	3 brityen s
811 brtiney spears	24 britenny spears	8 brightny spears	4 brieteny spears	3 briytney
664 birtney spears	24 britini spears	8 brintay spears	4 briety spears	3 brltney s
664 brintney spears	24 britnwy spears	8 brinttey spears	4 briitny spears	3 broteny s
664 briteney spears	24 brittni spears	8 briotney spears	4 briittany spears	3 brtaney s
601 bitney spears	24 brittnie spears	8 britanys spears	4 brinie spears	3 brtiiany
601 brinty spears	21 biritney spears	8 britley spears	4 brinteney spears	3 brtinay s
544 brittaney spears	21 birtany spears	8 britneyb spears	4 brintne spears	3 brtinney
544 brittnay spears	21 biteny spears	8 britnrey spears	4 britaby spears	3 brtitany
364 britey spears	21 bratney spears	8 britnty spears	4 britaey spears	3 brtiteny
364 brittiny spears	21 britani spears	8 brittner spears	4 britainey spears	3 brtnet sp
329 brtney spears	21 britanie spears	8 brottany spears	4 britinie spears	3 brytiny s
269 bretney spears	21 briteany spears	7 baritney spears	4 britinney spears	3 btney spe
269 britneys spears	21 brittay spears	7 birntey spears	4 britmney spears	3 drittney
244 britne spears	21 brittinay spears	7 biteney spears	4 britnear spears	3 pretney s
244 brytney spears	21 brtany spears	7 bitiny spears	4 britnel spears	3 rbritney
220 breatney spears	21 brtiany spears	7 breateny spears	4 britneuy spears	2 barittany
220 britiany spears	19 birney spears	7 brianty spears	4 britnewy spears	2 bbbritney
199 britnney spears	19 brirtney spears	7 brintye spears	4 britnmey spears	2 bbitney s
163 britnry spears	19 britnaey spears	7 britianny spears	4 brittaby spears	2 bbritny s

Motivation

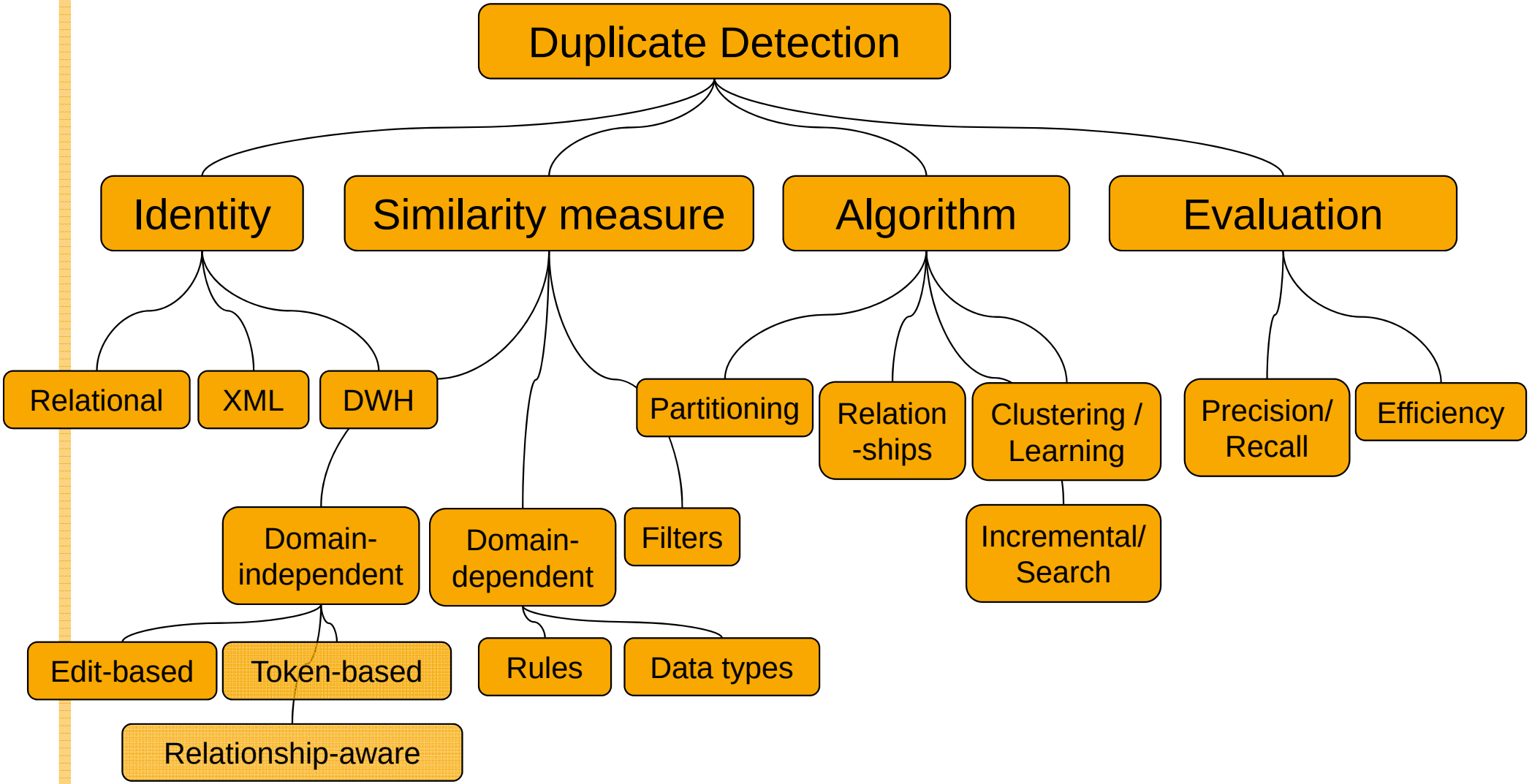
Customer	Revenue
BMW	20.000
BaMoWe	5.000.000
Bayerische Motorenwerke	300.000
...	...

- Possible effects
  - Example: Portfolio Management Offers
  - Credit maximum not detected
  - Too low inventory levels
  - No quantity discount for multiple orders
  - Total revenue of preferred customers unknown
  - Multiple mailings of same catalog to same household
- General problems
  - Additional, unnecessary IT expenses
  - Low customer satisfaction
  - Potentials and dangers not detected
  - Poor quality financial data

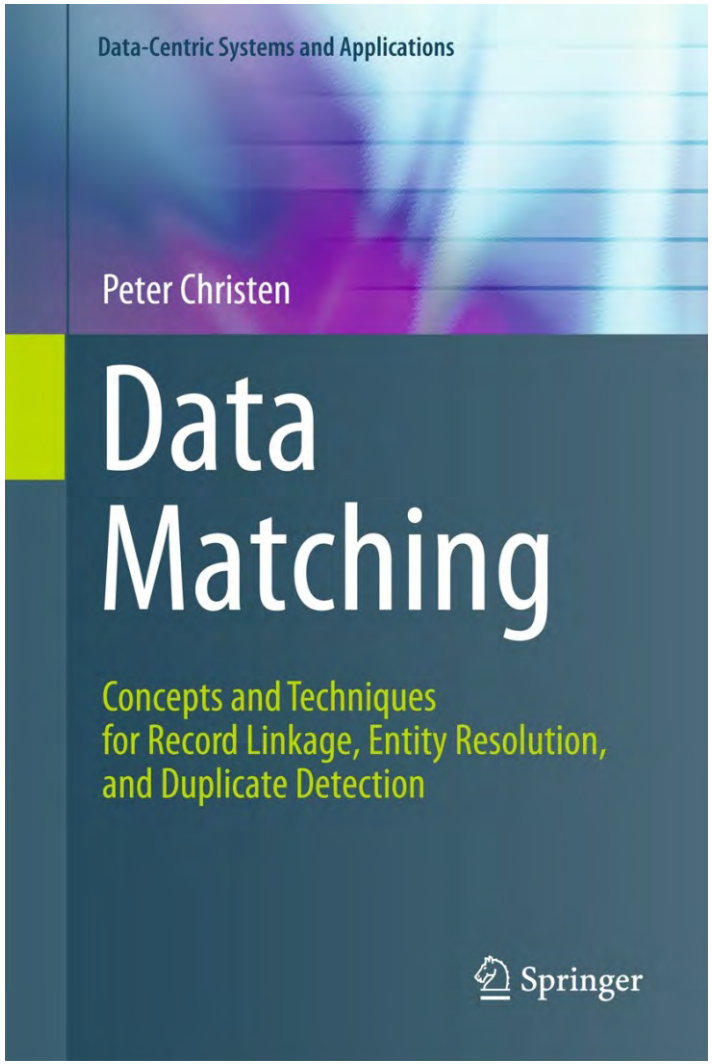
Ironically, “Duplicate Detection” has many Duplicates

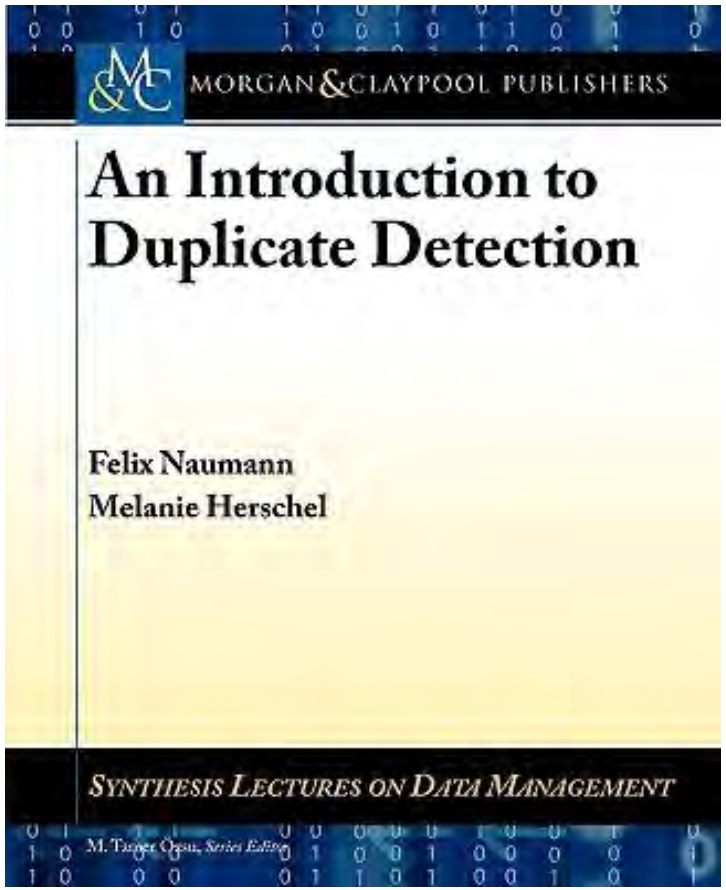
- Household matching
- Doubles
- Duplicate detection
- Mixed and split citation problem
- Record linkage
- Object identification
- Match
- Deduplication
- Fuzzy match
- Object consolidation
- Entity resolution
- Entity clustering
- Approximate match
- Identity uncertainty
- Reference reconciliation
- Merge/purge
- Hardening soft databases
- Householding
- Reference matching

Duplicate Detection – Research



Literature


Data-Centric Systems and Applications
Peter Christen
Data Matching
Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection
Springer


MORGAN & CLAYPOOL PUBLISHERS
An Introduction to Duplicate Detection
Felix Naumann
Melanie Herschel
SYNTHESIS LECTURES ON DATA MANAGEMENT

Overview

- Duplicate detection
- Similarity measures
- Algorithms
- Data sets and evaluation
- Data fusion

Token-based Similarity Measures

- Tokens
  - Words / Terms
  - n-grams
- Jaccard
  - |{common tokens}| / |{all tokens}|
- TFIDF [Cohen et al. 2003]
  - Term frequency: tf
  - Inverse document frequency: idf
  - TFIDF: log (tf+1) x log (idf)
  - Common words have low weight
  - Cosine similarity of term vectors weighted by tfidf
- And many more
  [Koudas Srivastavasa 2005]

Edit-based Similarity Measures

- Jaro [Jaro 1989] / Jaro-Winkler [Winkler 1999]
  - Common letters within ½ string length
  - Transposed letters
- Edit-distance / Levenshtein-distance [Levenshtein 1965]
  - Minimum number of edits from one word to the other
  - Domain-specific costing
  - Dynamic Programming
- Soundex
  - 4-letter code for each word
  - SOUNDEX('Farwick ') = F620
- ...

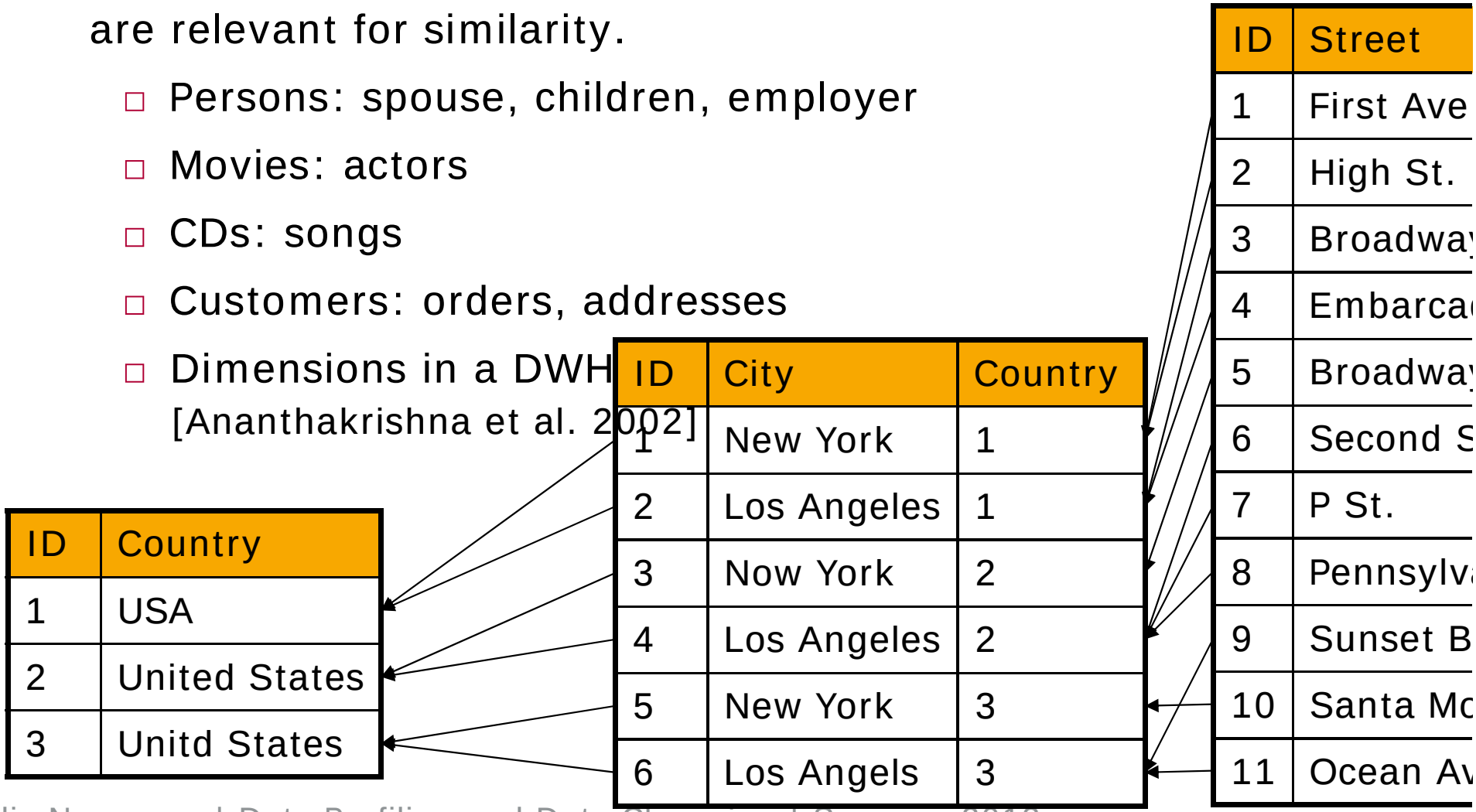
Frass, Fricke, Fahruschi, Feuerhake

Domain-dependent Similarity Measures

- Data Types
  - Special similarity for dates
  - Special similarity for numerical attributes
  - ...
- Rules
  - [Hernandez Stolfo 1998], [Lee et al. 2000]
  - **Given two records, r1 and r2.**
    IF last name of r1 = last name of r2,
    AND first names differ slightly,
    AND address of r1 = address of r2
    THEN r1 is equivalent to r2.

Relationship-aware Similarity Measures

- Idea: Not only values of the records, but values of related records are relevant for similarity.
  - Persons: spouse, children, employer
  - Movies: actors
  - CDs: songs
  - Customers: orders, addresses
  - Dimensions in a DWH [Ananthakrishna et al. 2002]

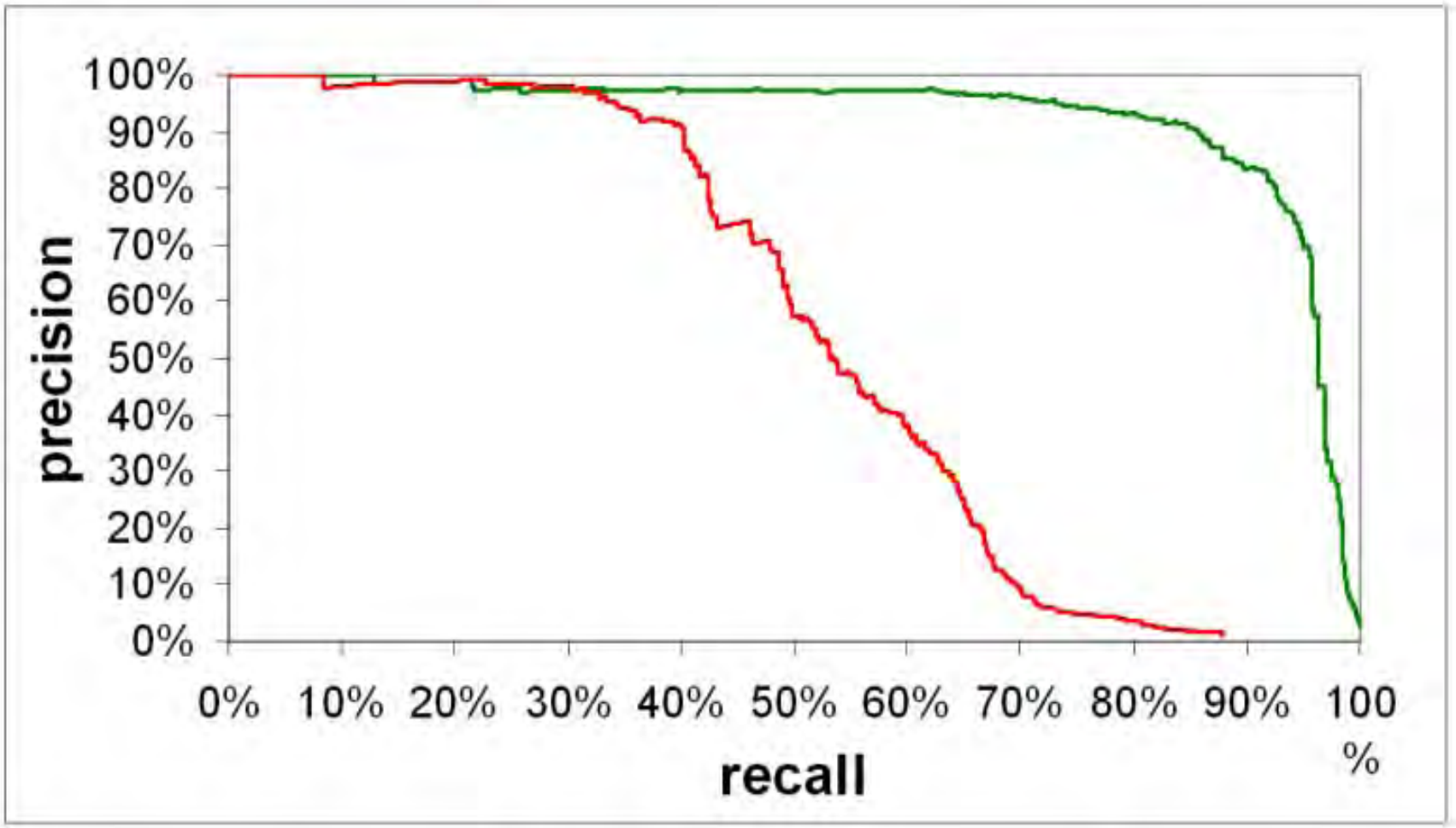


ID	Country
1	USA
2	United States
3	Unitd States

ID	City	Country
1	New York	1
2	Los Angeles	1
3	Now York	2
4	Los Angeles	2
5	New York	3
6	Los Angels	3

ID	Street
1	First Ave
2	High St.
3	Broadway
4	Embarca
5	Broadway
6	Second S
7	P St.
8	Pennsylv
9	Sunset B
10	Santa Mo
11	Ocean Av

Relationship-aware Similarity Measures – Evaluation



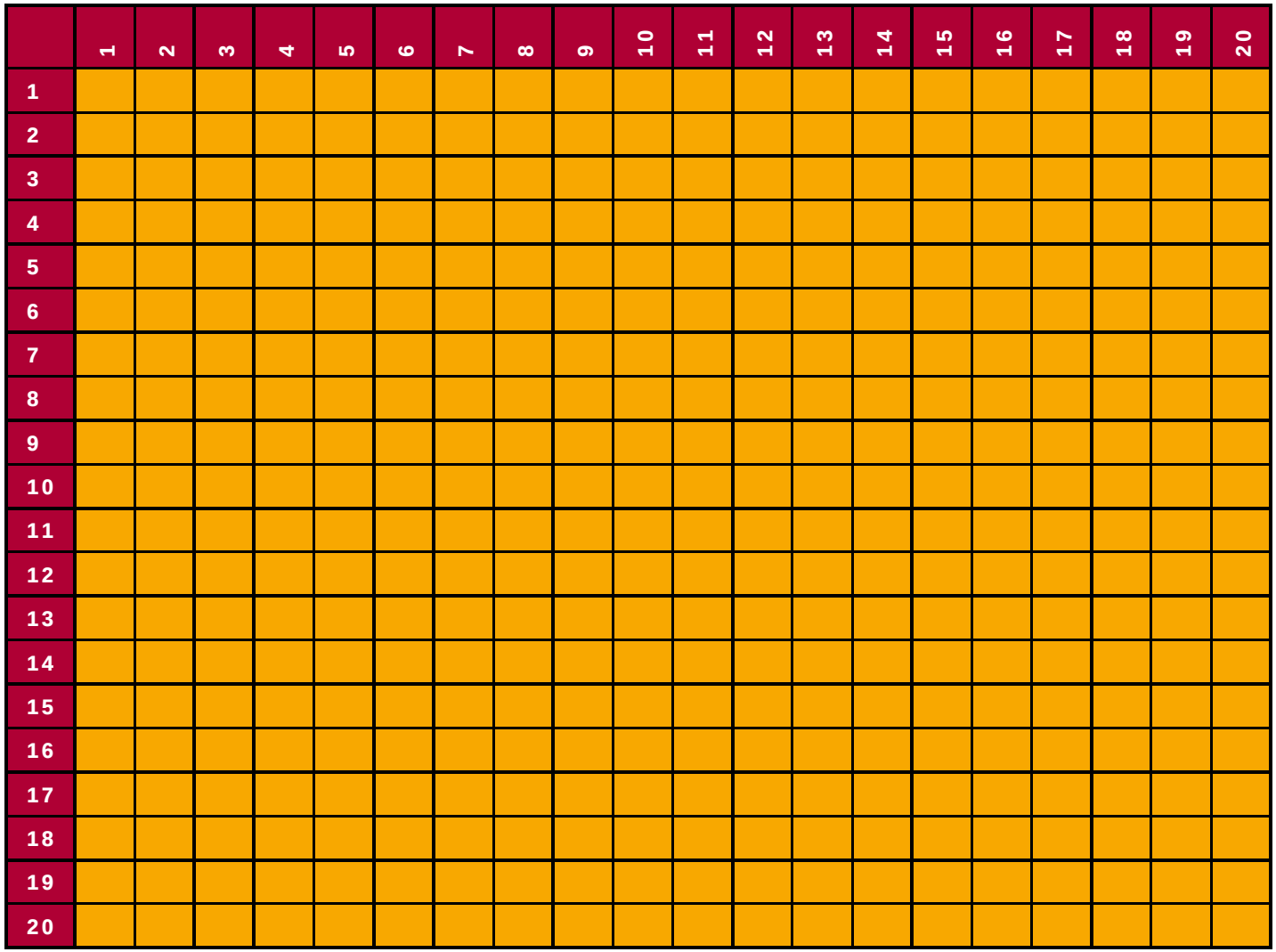
Overview

- Duplicate detection
- Similarity measures
- Algorithms
- Data sets and evaluation
- Data fusion

Record Pairs as Matrix

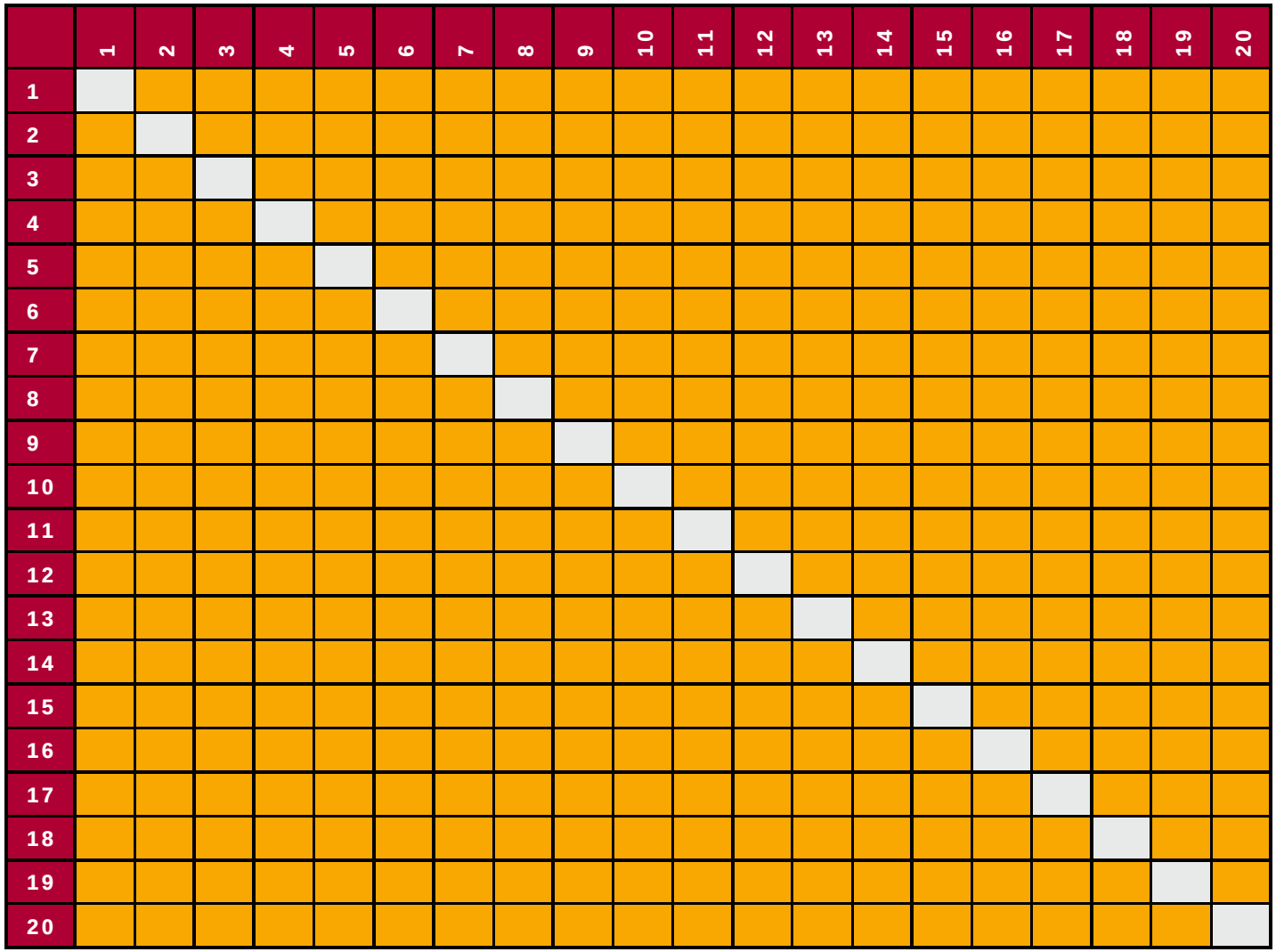
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2																				
3																				
4																				
5																				
6																				
7																				
8																				
9																				
10																				
11																				
12																				
13																				
14																				
15																				
16																				
17																				
18																				
19																				
20																				

Number of comparisons: All pairs



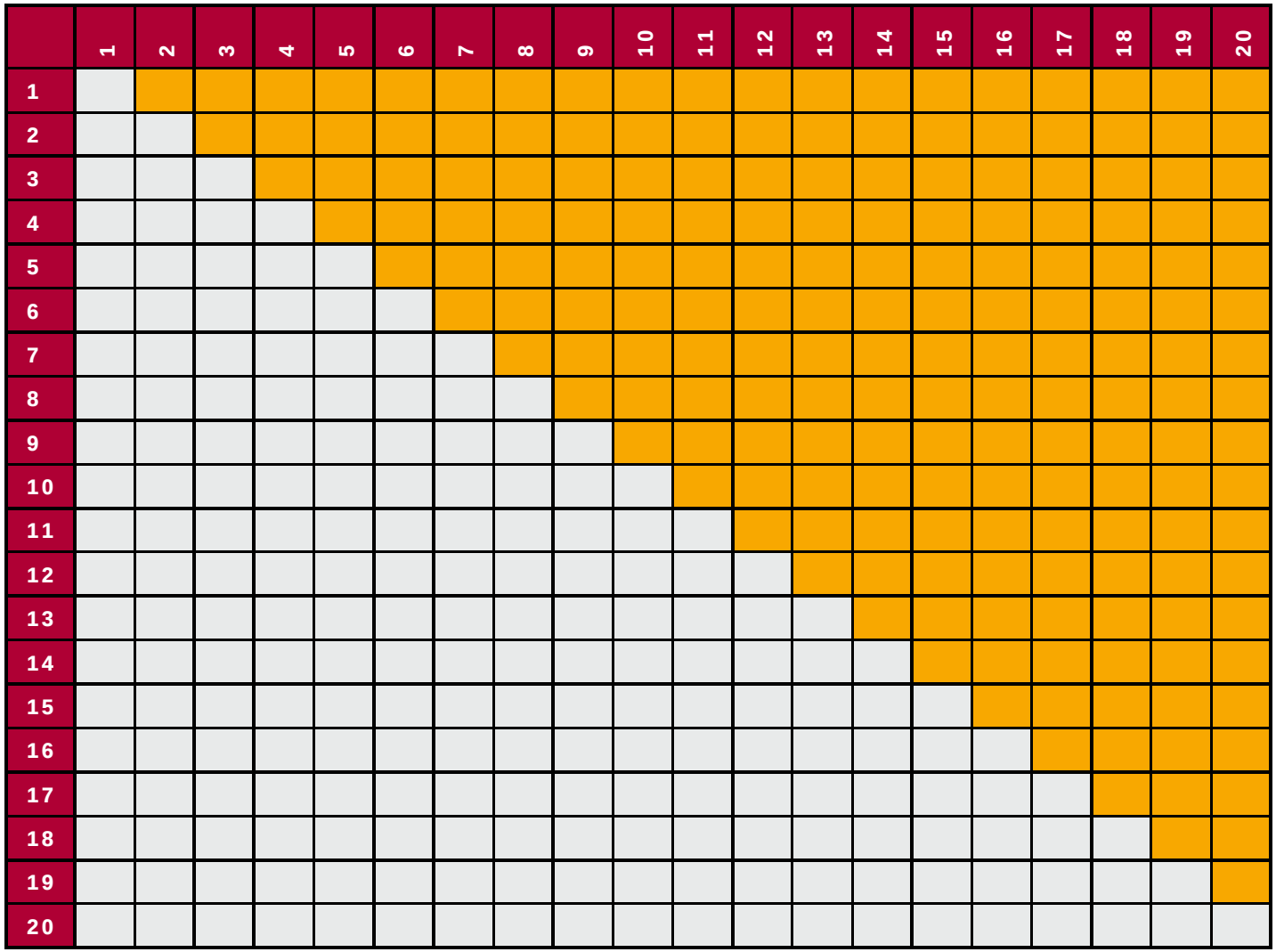
400 comparisons

Reflexivity of Similarity



380 comparisons

Symmetry of Similarity



190 comparisons

Complexity

- Problem: Too many comparisons!
  - 10.000 customers
    => 49.995.000 comparisons
    - $(n^2 - n) / 2$
    - Each comparison is already expensive.
- Idea: Avoid comparisons…
  - … by filtering out individual records.
  - … by partitioning the records and comparing only within a partition.

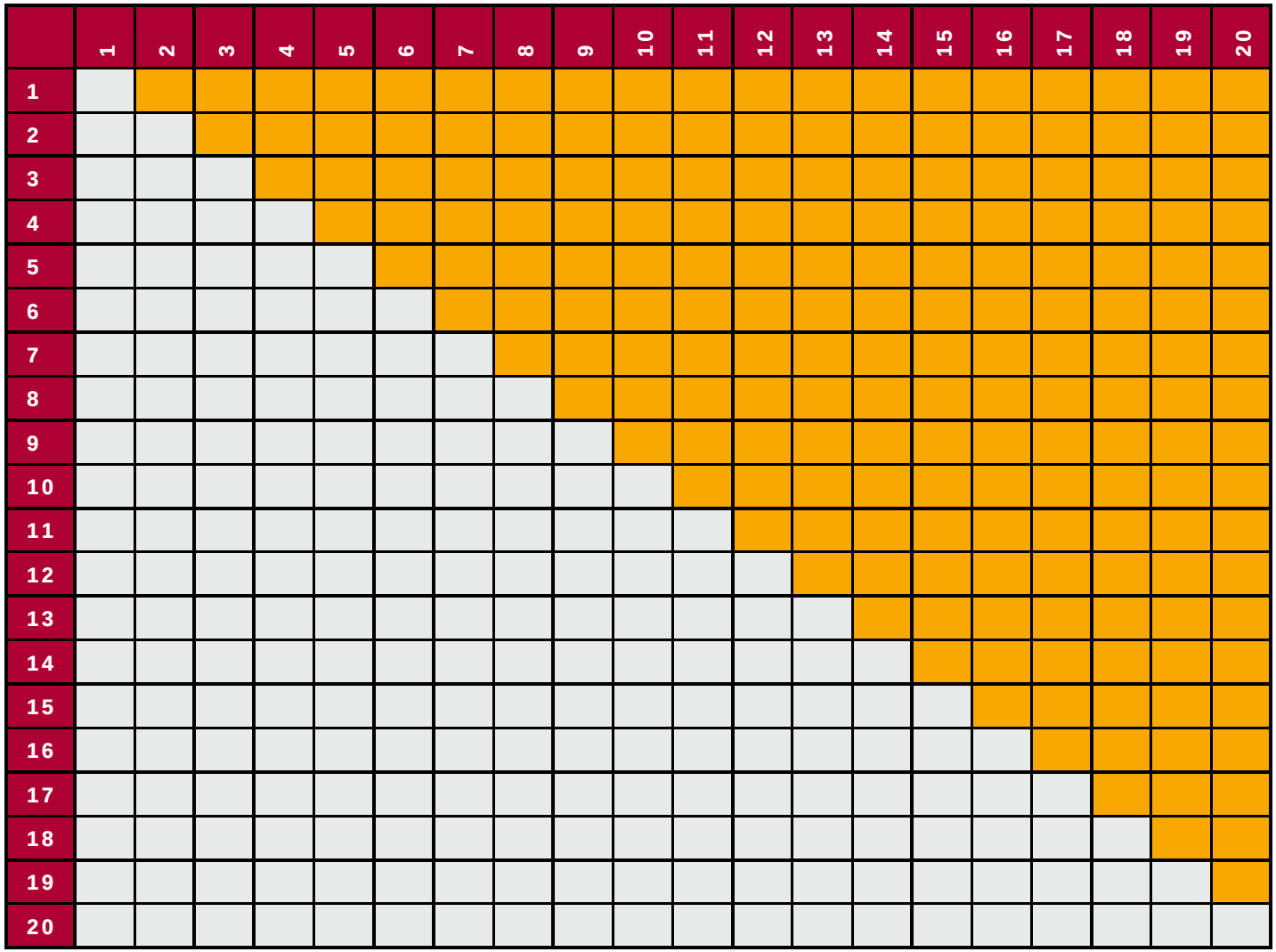
Partitioning / Blocking

- Partition the records (horizontally) and compare pairs of records only within a partition.
  - Partitioning by first two zip-digits
    - Ca. 100 partitions in Germany
    - Ca. 100 customers per partition
    - => 495.000 comparisons
  - Partition by first letter of surname
  - ...

Source: wikipedia.de

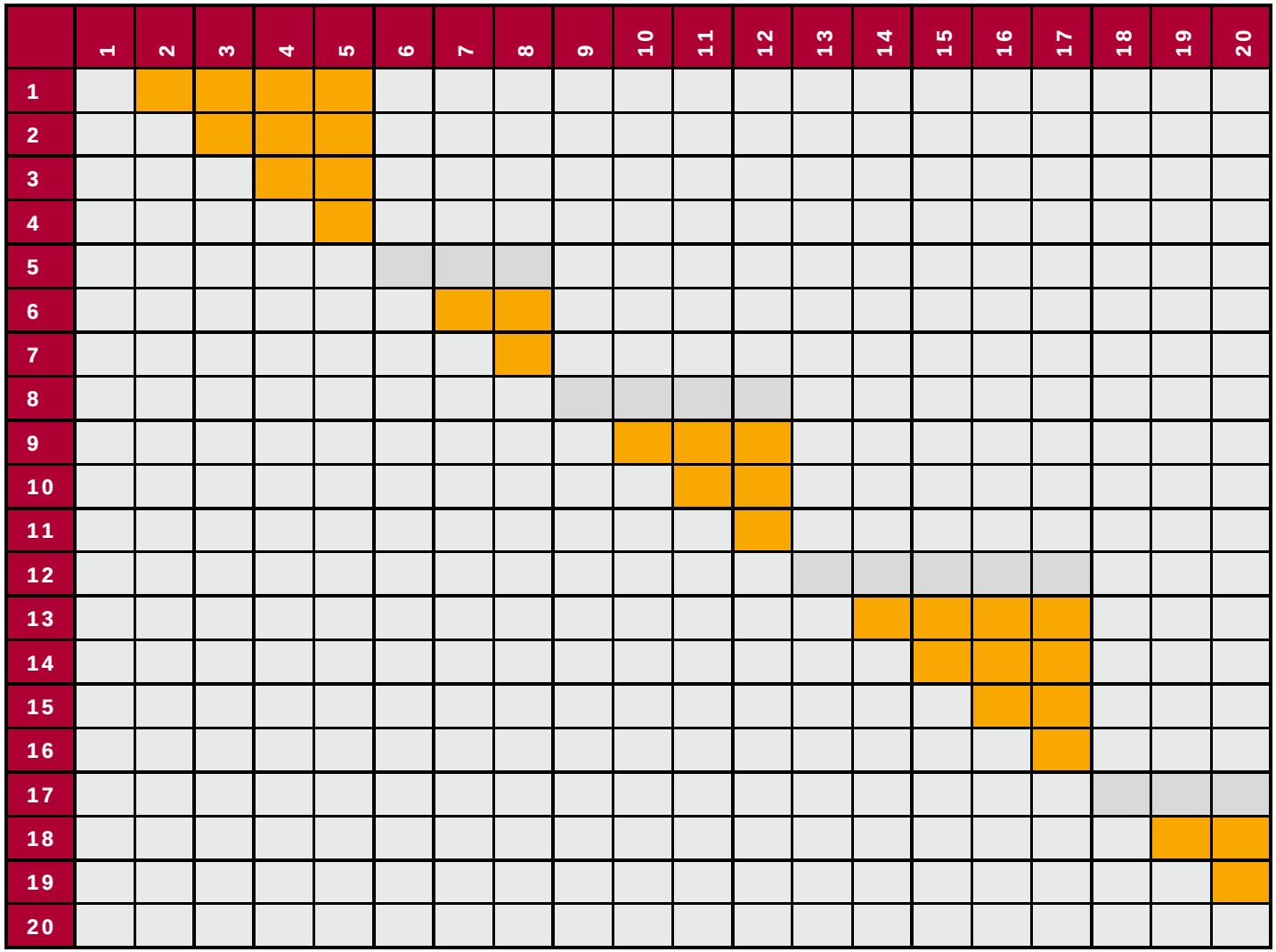
- Idea: Partition multiple times by different criteria.
  - Then apply transitive closure on discovered duplicates.

Records sorted by ZIP



190 comparisons

Blocking by ZIP



32 comparisons

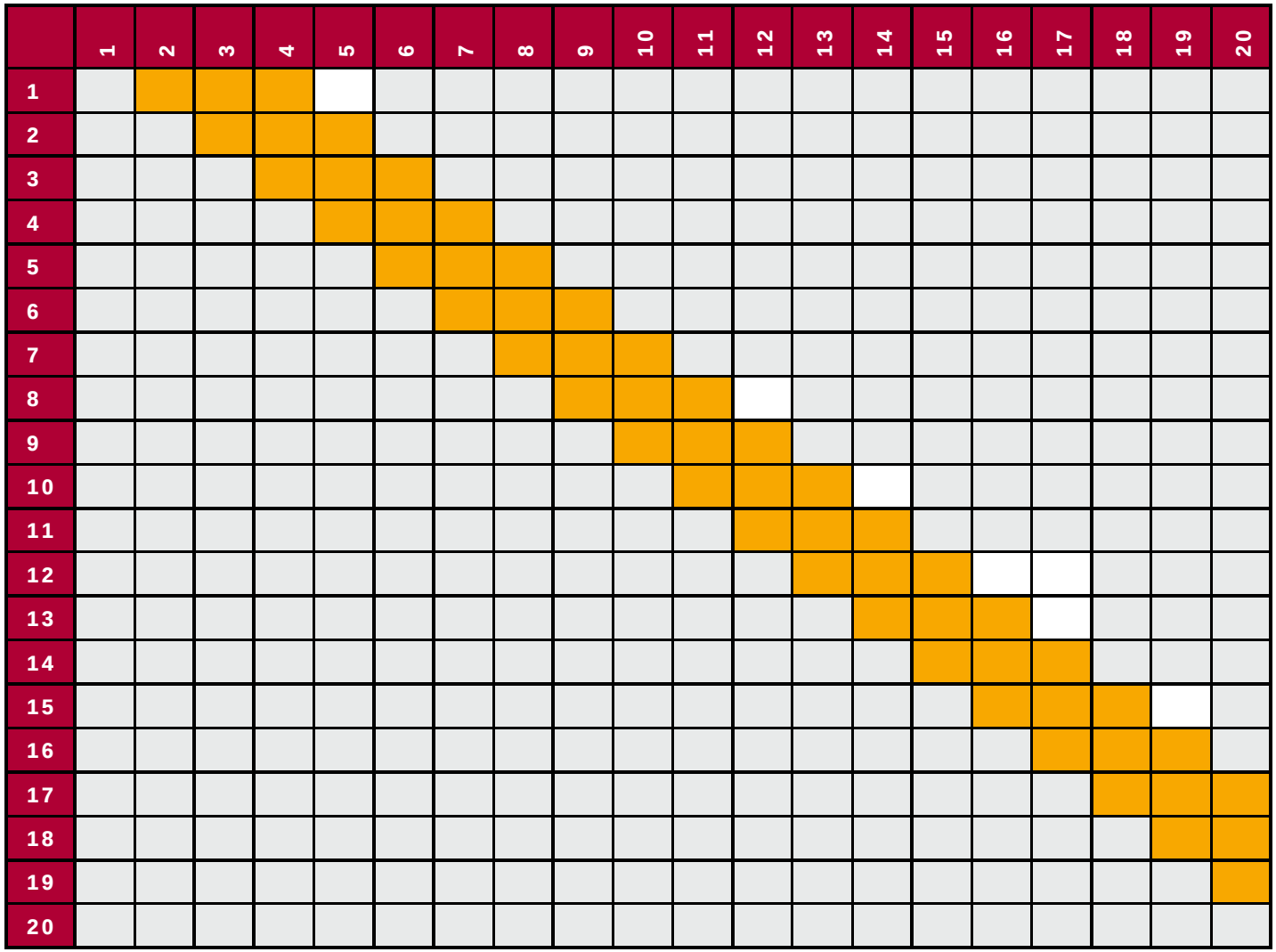
Sorted Neighborhood
[Hernandez Stolfo 1998]

- Idea
  - Sort tuples so that similar tuples are close to each other.
  - Only compare tuples within a small neighborhood (window).

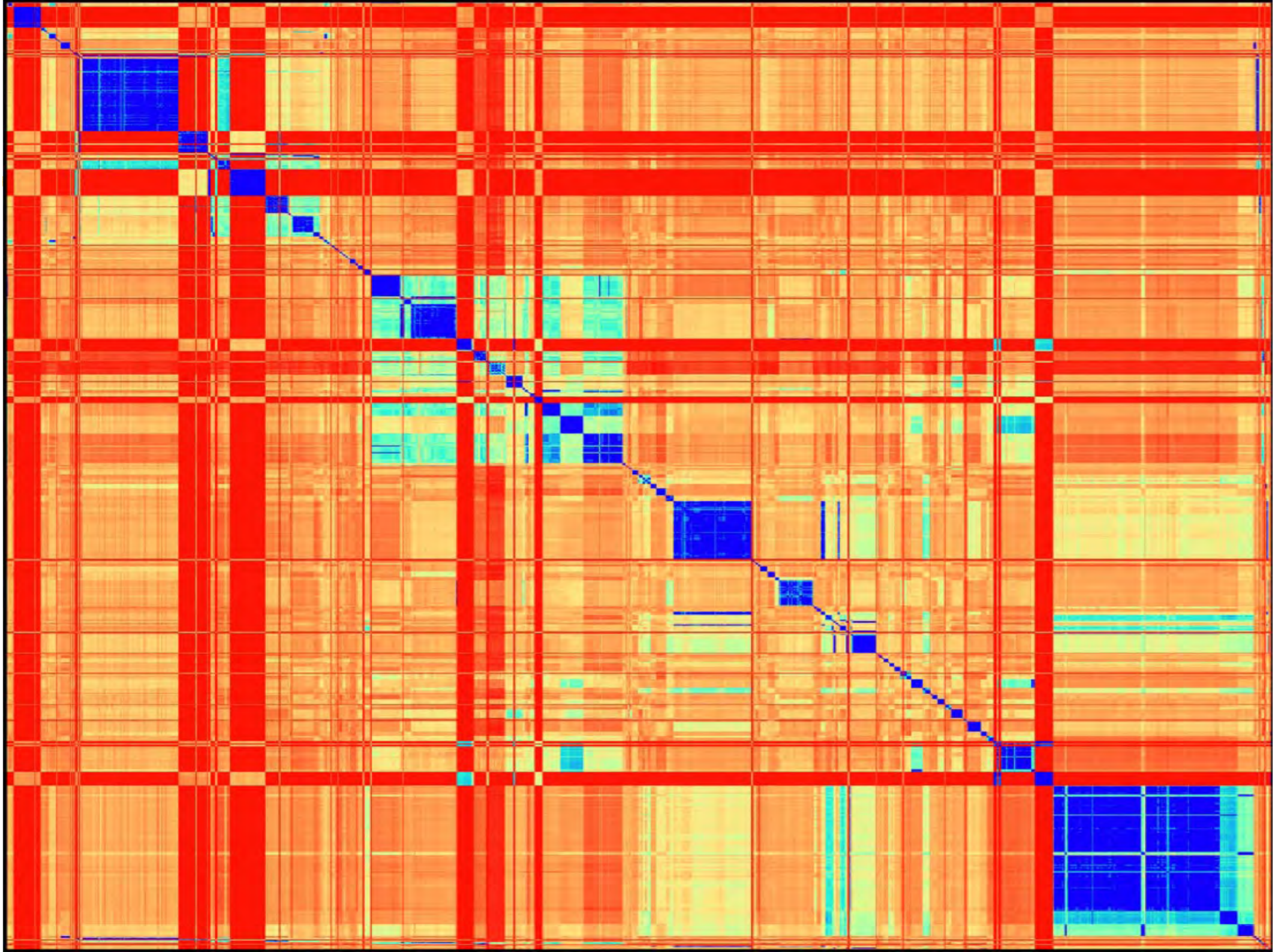
1. Generate key
   - E.g.: SSN+“first 3 letters of name“ + ...
2. Sort by key
   - Similar tuples end up close to each other.
3. Slide window over sorted tuples
   - Compare all pairs of tuples within window.

- Problems
  - Choice of key
  - Choice of window size
- Complexity: At least 3 passes over data
  - Sorting!

SNM by ZIP (window size 4)



54 comparisons



Overview

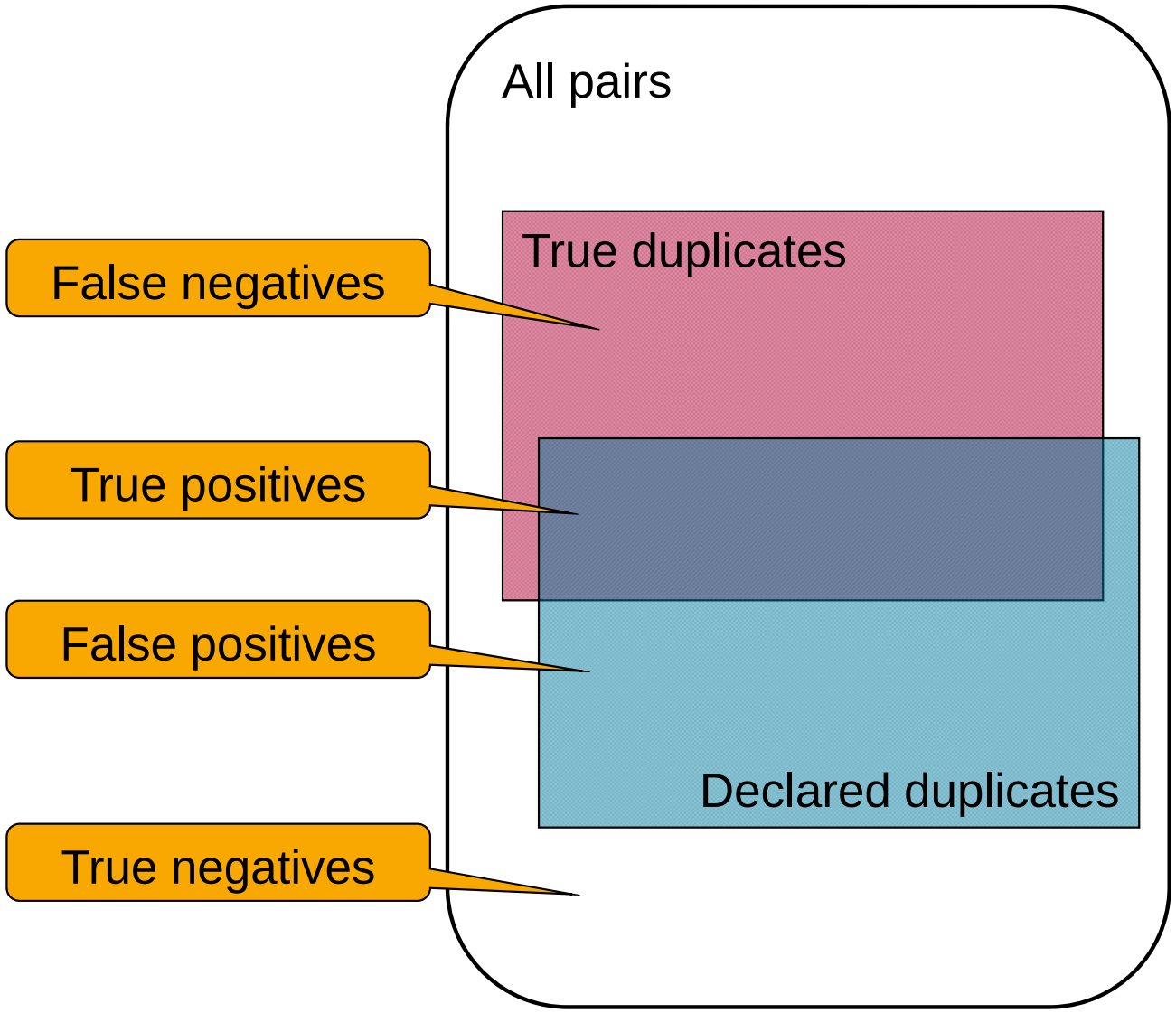
- Duplicate detection
- Similarity measures
- Algorithms
- Data sets and evaluation
- Data fusion

Precision & Recall

- True positives (TP): Correctly declared duplicates
- False positives (FP): Incorrectly declared duplicates
- True negatives (TN): Correctly avoided pairs
- False negatives (FN): Missed duplicates

- Precision = TP / (TP + FP)
  - = TP / declared dups
  - Proportion of found matches that are correct
- Recall = TP / (TP + FN)
  - = TP / all dups
  - Proportion of correct matches that are found

Precision & Recall



$$\text{Precision} = \frac{\text{True positives}}{\text{Declared duplicates}}$$

$$\text{Recall} = \frac{\text{True positives}}{\text{True duplicates}}$$

$$\text{F-Measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

F Measure

- *Harmonic mean* of recall and precision

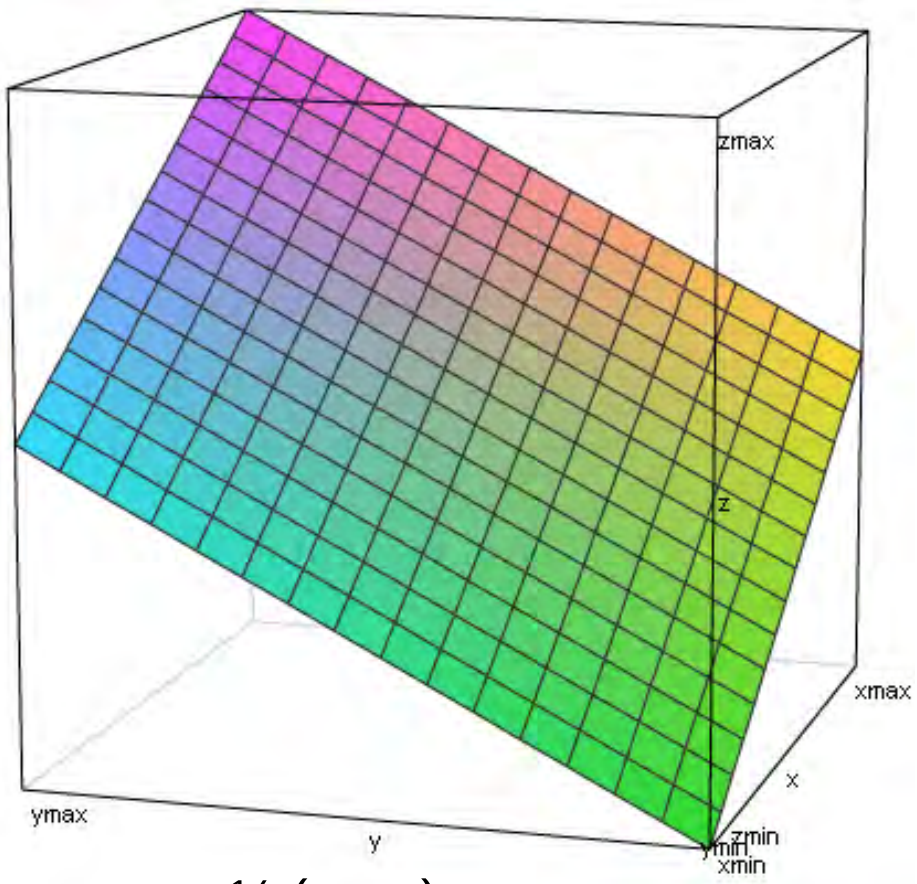
$$F = \frac{1}{\frac{1}{2}\left(\frac{1}{R} + \frac{1}{P}\right)} = \frac{2RP}{R+P}$$

  - Harmonic mean emphasizes the importance of small values, whereas arithmetic mean is affected more by outliers that are unusually large.
- More general form: Weighted harmonic mean

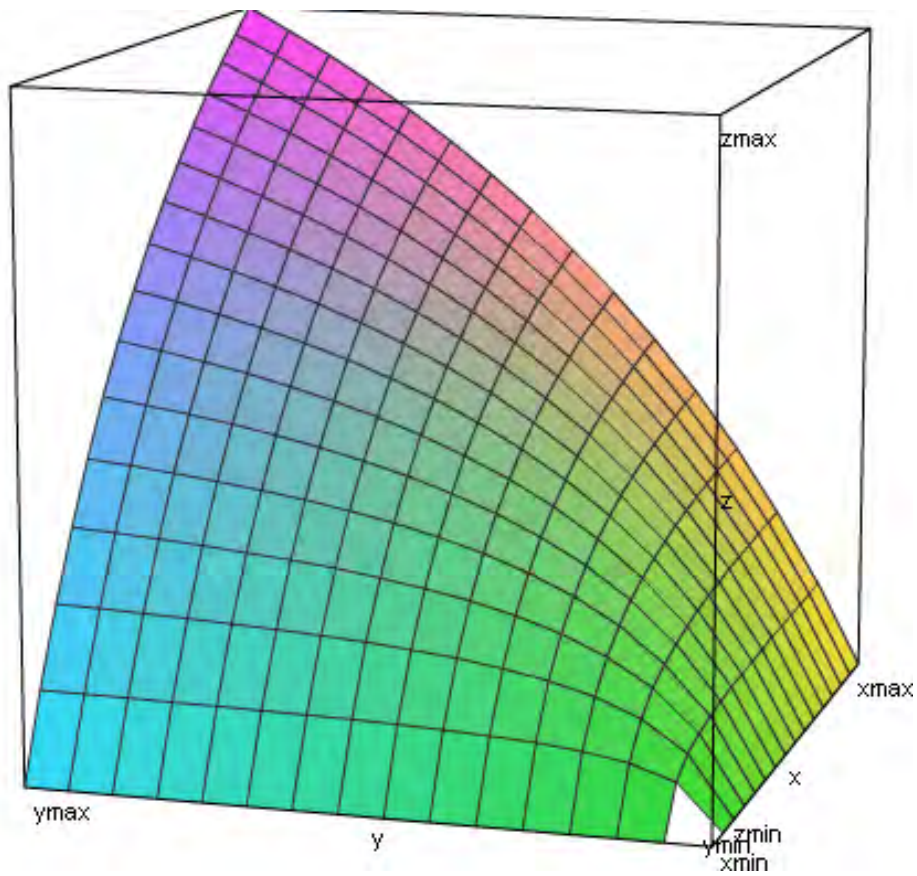
$$F_\alpha = \frac{RP}{\alpha R + (1-\alpha)P}$$

  - Thus, harmonic mean is $F_{1/2}$

Arithmetic mean („Average“) vs. Harmonic mean („F-Measure“)

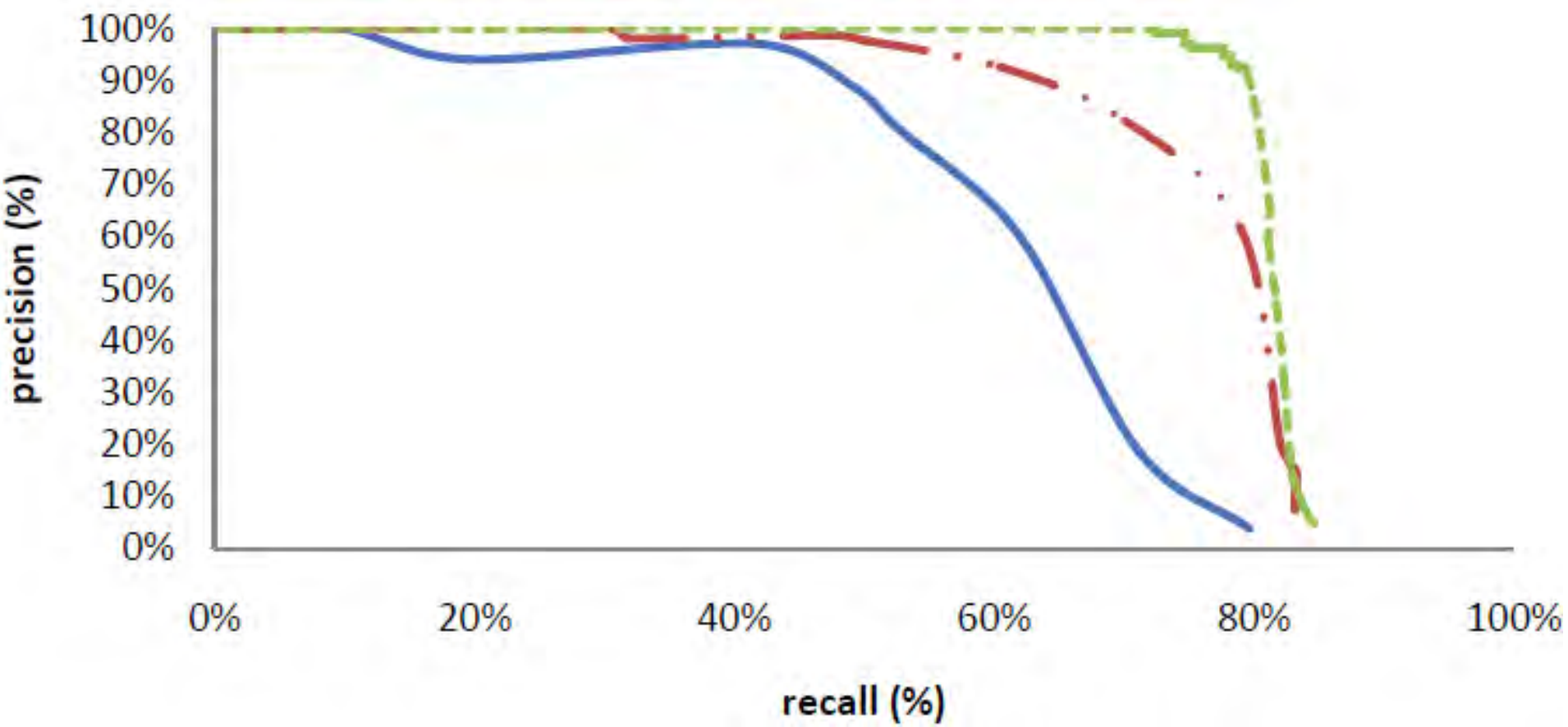


$z = ½ (x + y)$

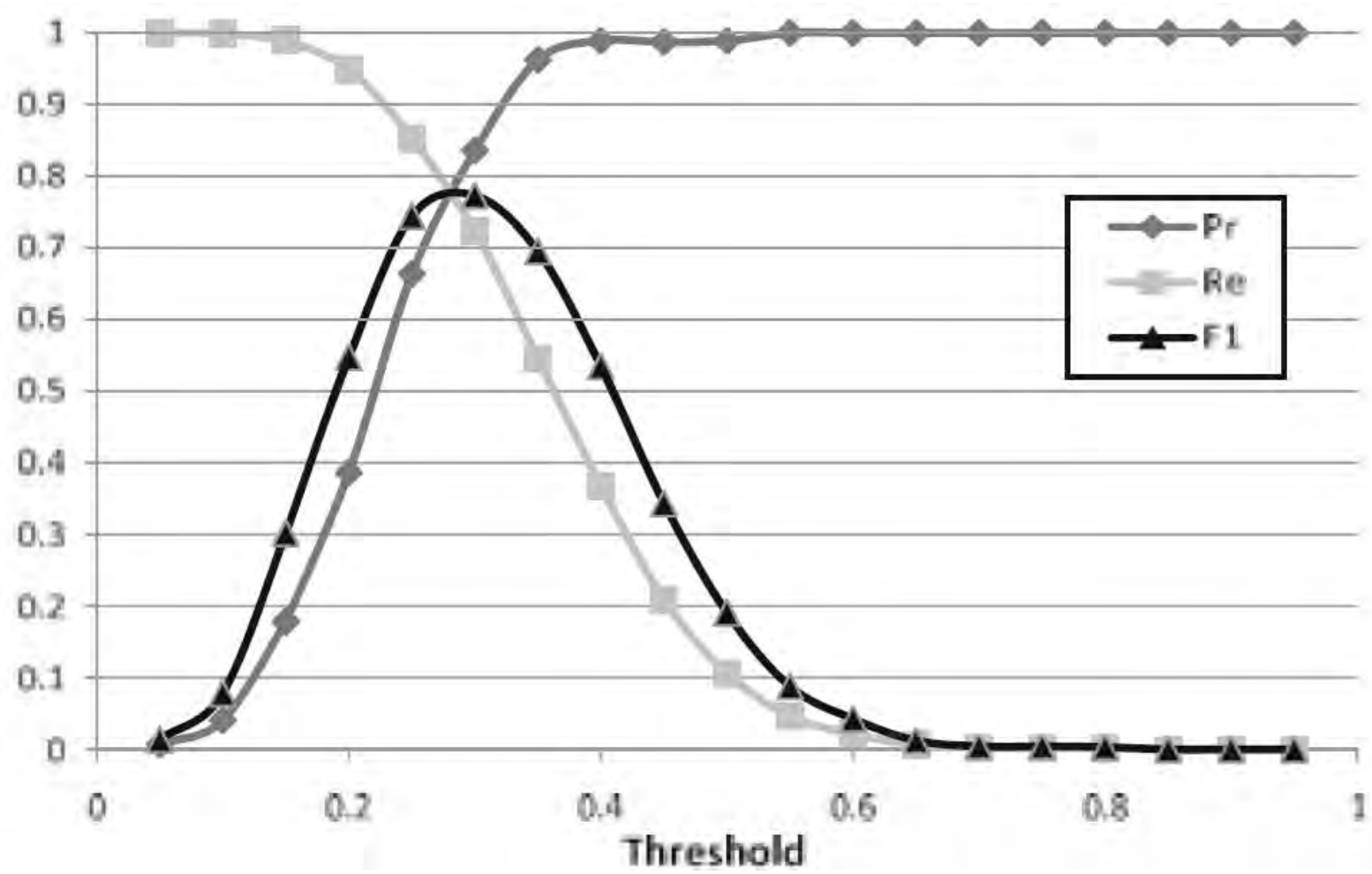


$z = 2 (x \cdot y) / (x + y)$

Recall-precision-diagram

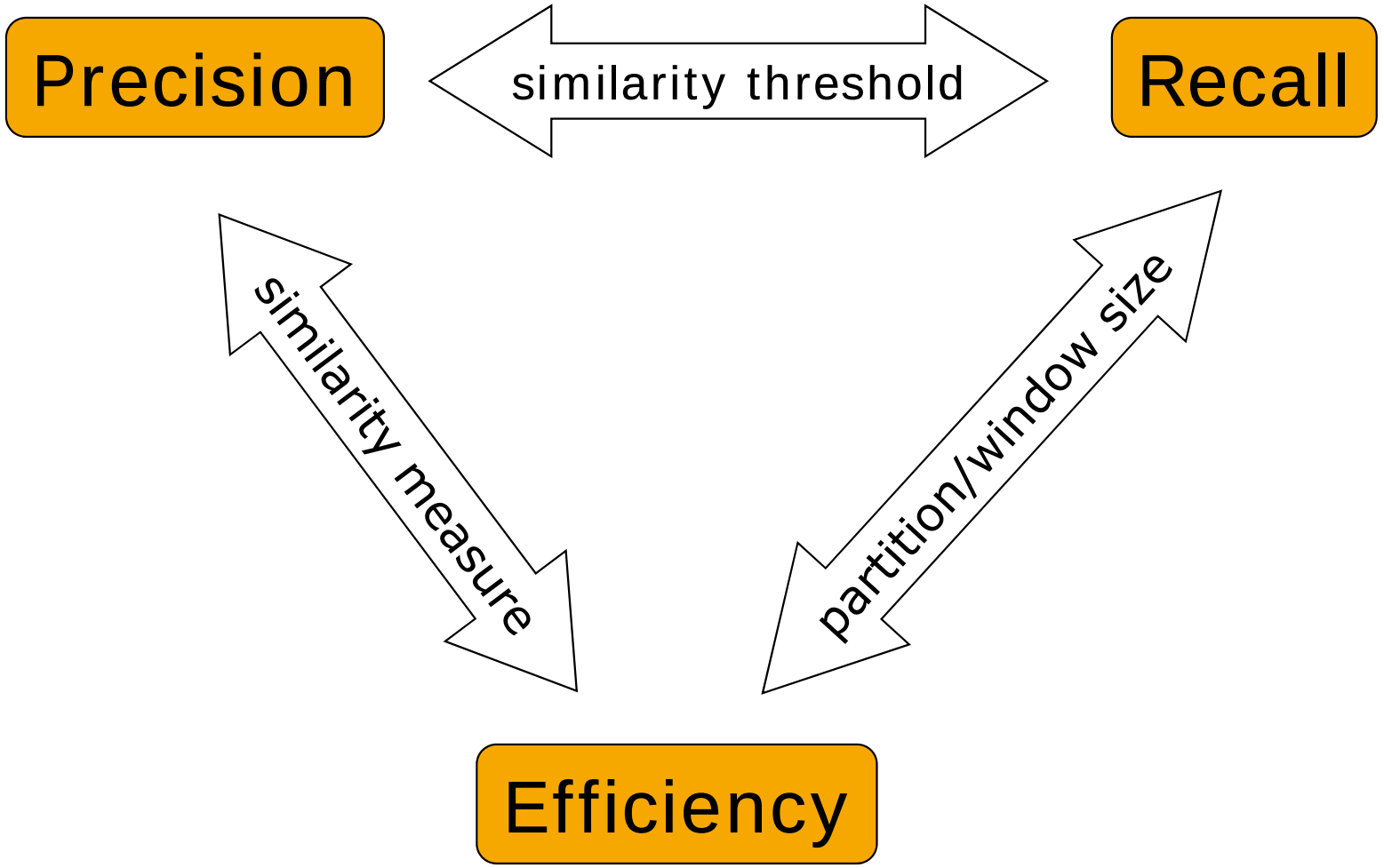


F1 Graph



From Creating probabilistic databases from duplicated data
Oktie Hassanzadeh · Renée J. Miller (VLDBJ)

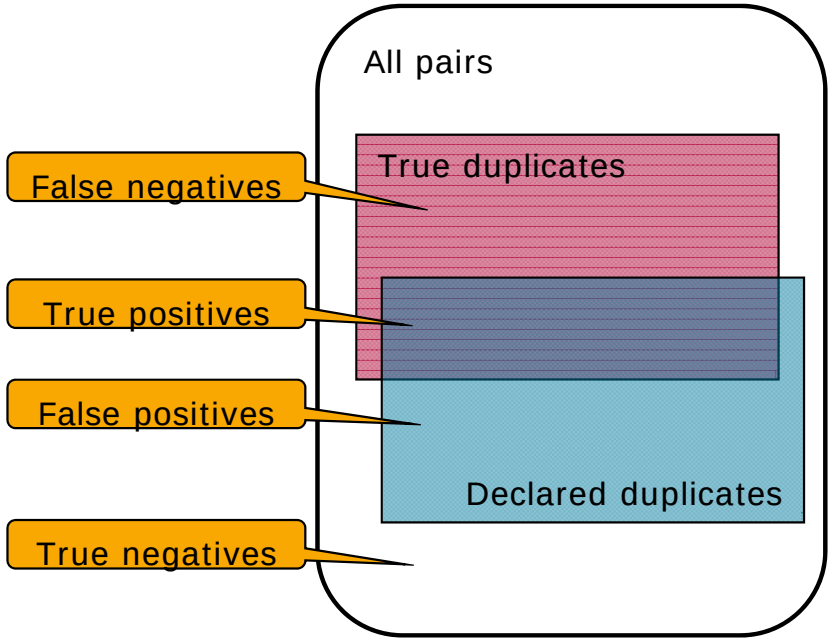
Evaluating Duplicate Detection



Other effectivness measures

- Accuracy
  - (TP + TN) / (TP + FP + TN + FN)
  - Used for balanced classes
  - For duplicate detection, TN usually dominates overall result

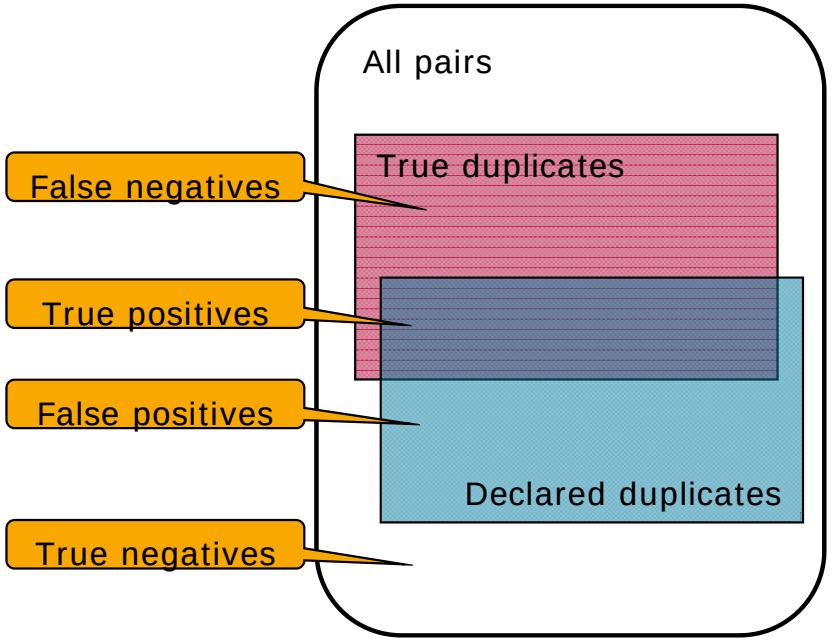
- Specificity
  - TN / (TN + FP)
  - = TN / true non-matches
  - Again: TN dominates overall result



Complexity measures

- Problem: Measure not only quality of similarity measure, but also that of algorithm
  - Solution 1: Runtime measurements
    - But: Different hardware, difficult repeatability
  - Solution 2: Measure how well/poor algorithms filter candidates

- Reduction ratio
  - 1 – ((TP+TN)/(FN+TP+FP+TN)) = 1 – accuracy
- Pairs completeness
  - TP / (FN + TP) = recall
- Pairs quality
  - TP / (TP + TN)



Data sets to evaluate deduplication

- Requirements
  - Real-world application
  - Interestingly large
  - Interestingly dirty
  - Gold standard
  - Publicly available
  - (Relational)
  - (Understandable)
- Hardly available
  - Privacy issues
  - Security issues
  - Embarrasment
  - Data fiefdoms

Small datasets with gold standard

CORA

- 1878 bibliographic references in XML format
- http://www.hpi.uni-potsdam.de/naumann/projekte/repeatability/datasets/cora_dataset.html

DBLP

- 50,000 bibliographic references in XML format
- http://www.hpi.uni-potsdam.de/naumann/projekte/repeatability/datasets/dblp_dataset.html

Restaurants

- 864 restaurants with 112 duplicates
- http://www.cs.utexas.edu/users/ml/riddle/data.html

Whirl datasets

- 11 smaller datasets with a single string attribute
- http://www.cs.purdue.edu/commugrate/data/whirl/match/

Large datasets without gold standard

Places

- 1.4 million POIs from Facebook, Gowalla, Foursquare
- https://www.hpi.uni-potsdam.de/naumann/trac/LuSim/wiki/quellen

WheelMap

- 120,000 places/things in Germany

FreeDB

- 1.9 million CDs, dirty, some duplicate clusters quite large
- original: http://www.freedb.org/en/download__database.10.html
- derived: http://www.hpi.uni-potsdam.de/naumann/projekte/repeatability/datasets/cd_datasets.html

CITESEERX

- 1.3 million publications in CSV
- http://asterix.ics.uci.edu/data/csx.raw.txt.gz

Data generation

- For lack of gold standard: create one
- Data base
  - Real-world data sets (without or without enough duplicates)
  - Real-world values (from dictionaries)
  - Synthetic strings
- Data corruption: Duplicate and modify some percentage of tuples
  - Duplication: Cluster sizes?
  - Data values
    - Insert/remove/transpose/change certain letters
    - Delete values
    - Swap values (within tuple, from dictionary, across tuples)
- General suspicion: Similarity measure and candidate selction is geared towards known types of errors.

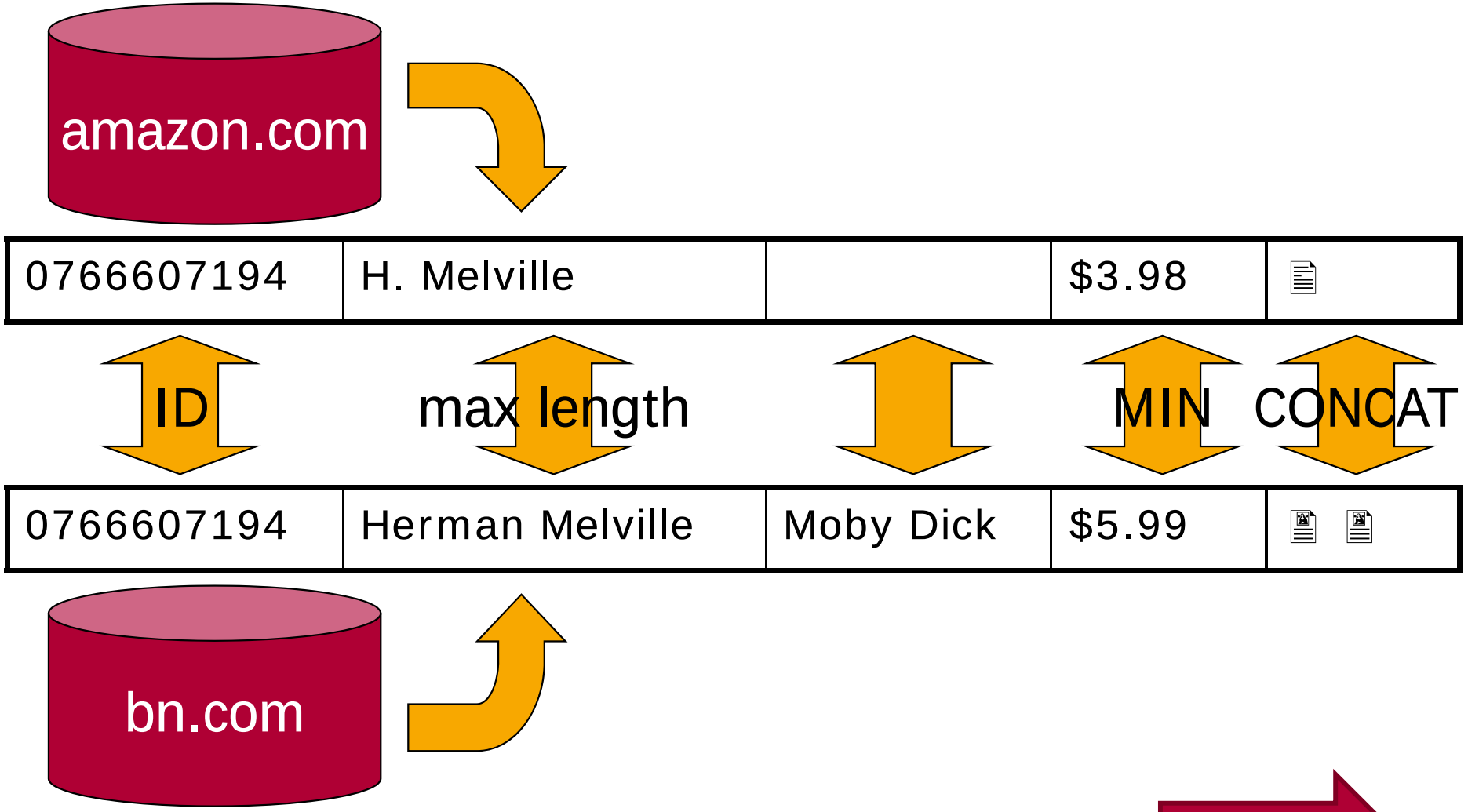
Data generators

- UIS Database Generator
  - Generates a list of randomly perturbed names and US mailing addresses.
  - Written by Mauricio Hernández.
  - http://www.cs.utexas.edu/users/ml/riddle/data/dbgen.tar.gz
- FEBRL-Generator
  - Part of a cleansing suite
  - Dictionaries with frequencies
  - http://sourceforge.net/projects/febrl/
- Dirty XML Generator
  - http://www.hpi.uni-potsdam.de/naumann/projekte/completed_projects/dirtyxml.html

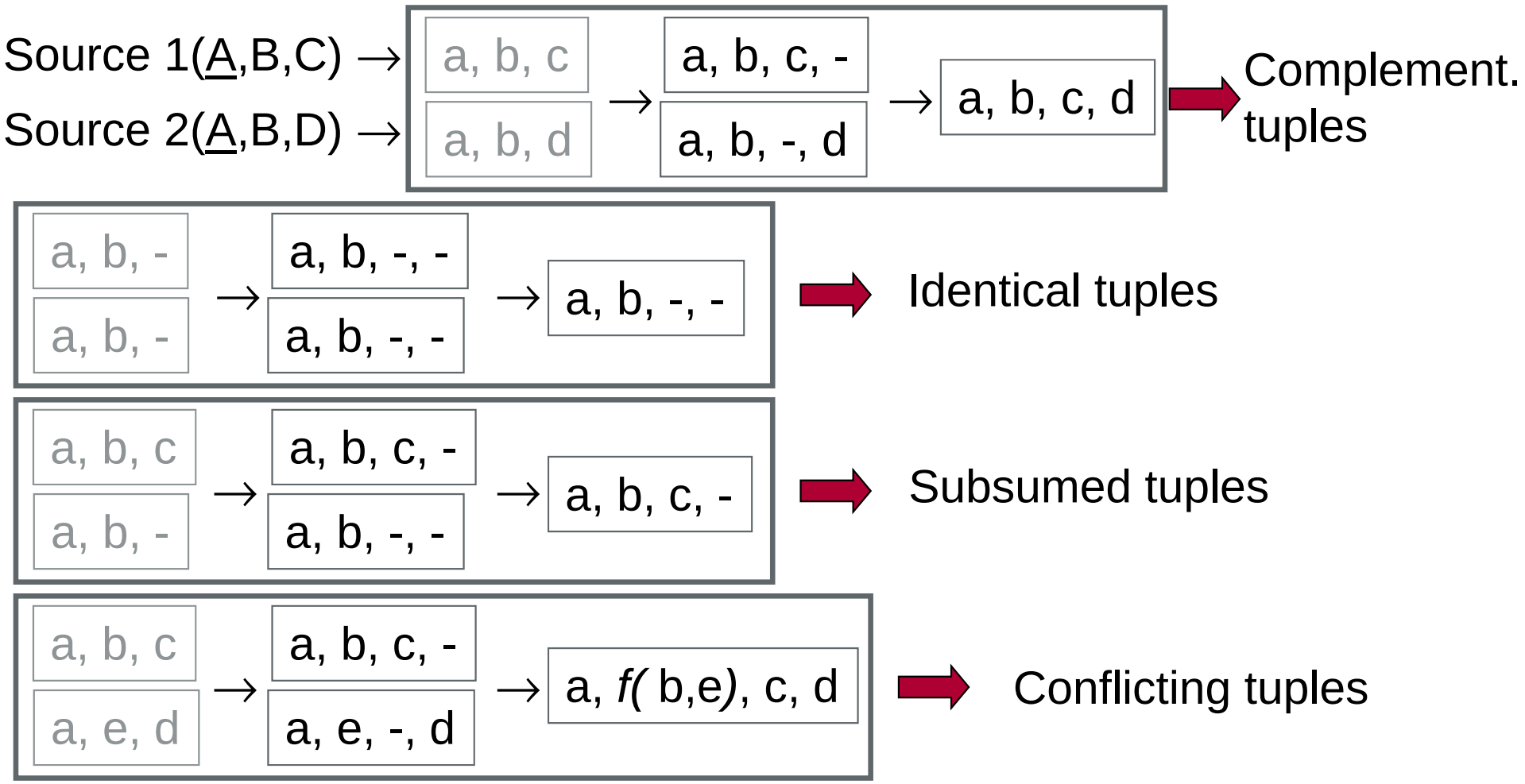
Overview

- Duplicate detection
- Similarity measures
- Algorithms
- Data sets and evaluation
- Data fusion

Data Fusion


amazon.com
0766607194
H. Melville
$3.98
ID
max length
MIN
CONCAT
0766607194
Herman Melville
Moby Dick
$5.99
bn.com
Its late

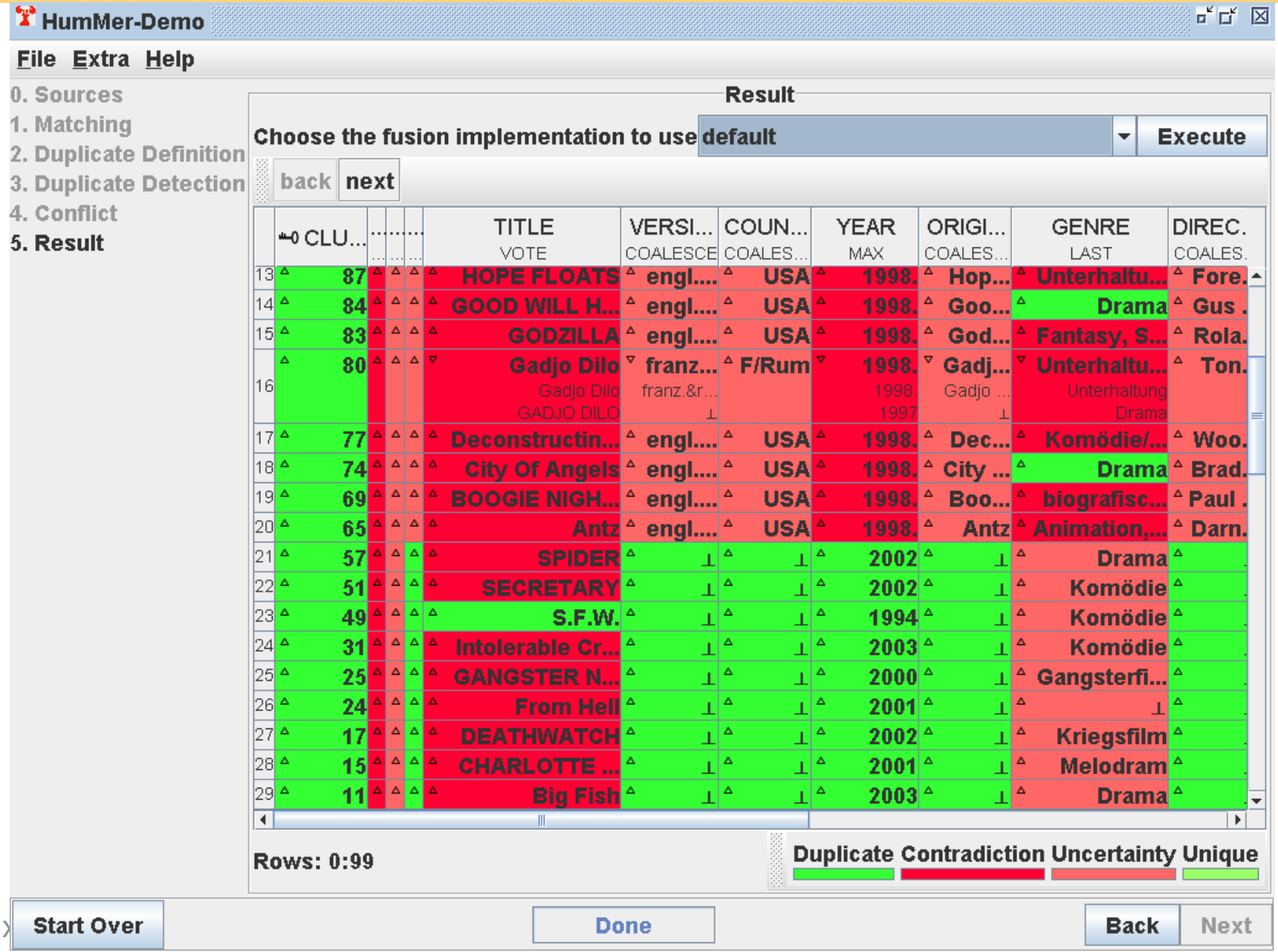
“Proper” Data Fusion



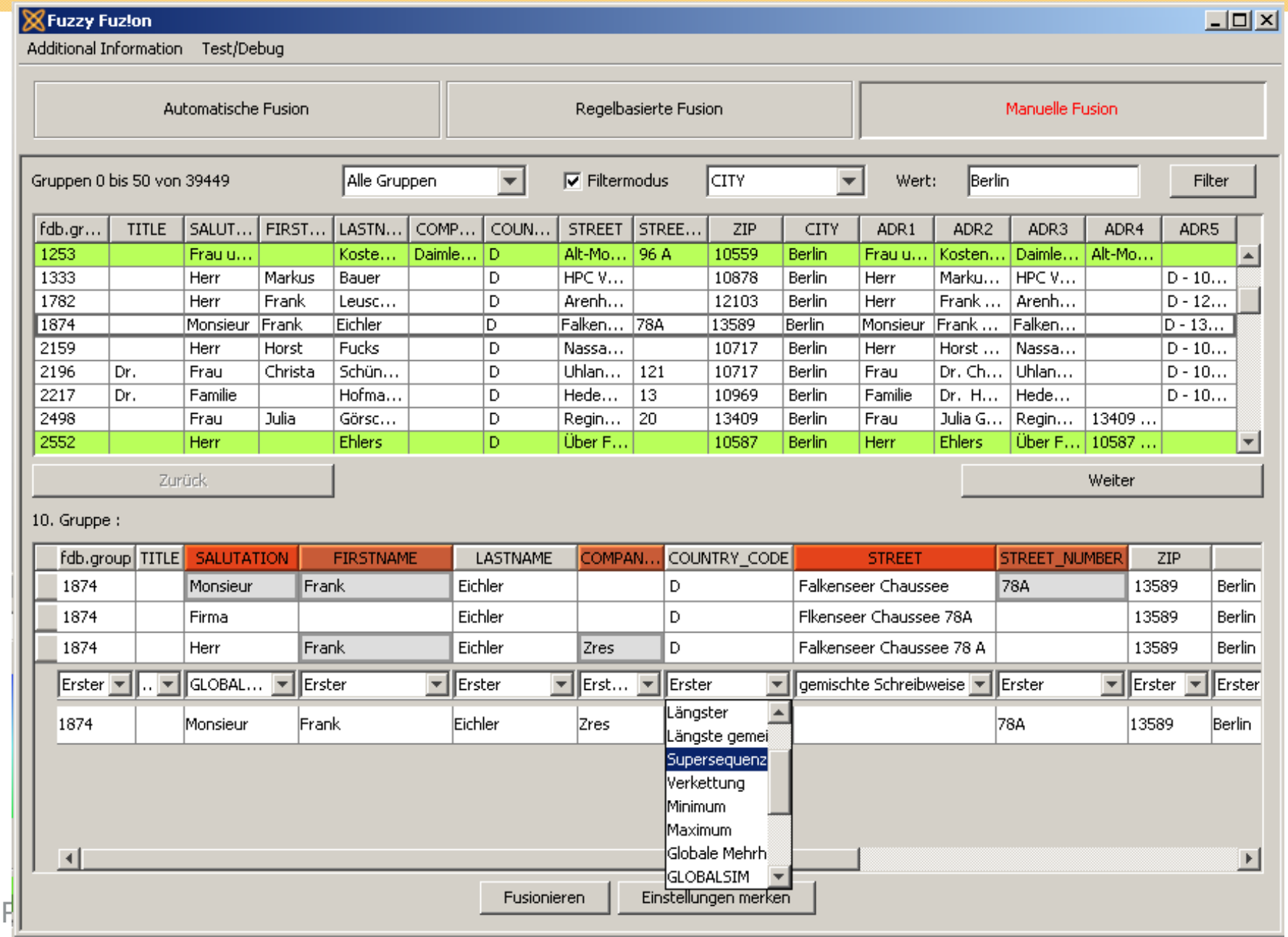
Conflict Resolution Functions

Min, Max, Sum, Count, Avg, StdDev	Standard aggregation
Random	Random choice
First, Last	Choose first/last value; depends on order
Longest, Shortest	Choose longest/shortest value
Choose(source)	Choose value froma particular source
ChooseDepending(col, val)	Choose depending on val in other column col
Vote	Majority decision
Coalesce	Choose first non-null value
Group, Concat	Group or concatenate all values
MostRecent	Choose most recent (up-to-date) value
MostAbstract, MostSpecific	Use a taxonomy / ontology
….	….

Visualization of Integrated Data



Tool-based Data Fusion



References

- [Ananthakrishna et al. 2002] R. Ananthakrishna, S. Chaudhuri, V. Ganti: Eliminating Fuzzy Duplicates in Data Warehouses, *Proc. of the 28th VLDB Conference*, Hong Kong, China, pages 586-597, 2002.
- [Dey et al. 2002] D. Dey, S. Sarkar, P. De: A Distance-based Approach to Entity Reconciliation in Heterogeneous Databases, *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 14(3): 567-582, 2002.
- [Gravano et al. 2001] L. Gravano, P. Ipeirotis, H. Jagadish, N. Koudas, S. Muthukrishnan, D. Srivastava: Approximate String Joins in a Database (Almost) for Free, *Proc. of the 27th VLDB Conference*, Roma, Italy, pages 491-500, 2001.
- [Hernandez Stolfo 1995] M. Hernandez, S. Stolfo: The Merge/Purge Problem for Large Databases, *Proc. ACM SIGMOD Conference 1995*, San Jose, USA, pages 127-138, 1995.
- [Hernandez Stolfo 1998] M. Hernandez, S. Stolfo: Real-world Data is Dirty: Data Cleansing and the Merge/Purge, *Journal of Data Mining and Knowledge Discovery*, 2(1):9-37, 1998.
- [Jaro 1989] M. Jaro: Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida, *Journal of the American Statistical Association* 84(406):414-420, 1989.
- [Low et al. 2001] W. Low, M. Lee, T. Ling: A Knowledge-based Approach for Duplicate Elimination in Data Cleaning, *Information Systems* 26(8):585-606, 2001.
- [Monge 2000] A. Monge: Matching Algorithms within a Duplicate Detection System, *IEEE Data Engineering Bulletin*, 23(4):14-20, 2000.
- [Navarro 2001] G. Navarro: A Guided Tour of Approximate String Matching, *ACM Computing Surveys* 31(1):31-88, 2001.
- [Weis Naumann 2005] Melanie Weis, Felix Naumann: DogmatiX Tracks down Duplicates in XML. In Proc. of the ACM Conference on Management of Data (SIGMOD) 2005.
- [GL94] Outerjoins as Disjunctions, Cesar A. Galindo-Legaria, SIGMOD 1994 conference
- [Cod79] E. F. Codd: Extending the Database Relational Model to Capture More Meaning. TODS 4(4): 397-434 (1979)
- [Ull89] Jeffrey D. Ullman: Principles of Database and Knowledge-Base Systems, Volume II. Computer Science Press 1989.