C. Batini & M. Scannapieco
Data and Information Quality Book Figures

Chapter 8: Object Identification

How three agencies represent the same business

Agency	Identifier	Name	Type of activity	Address	City
Agency 1	CNCBTB765SDV	Meat production of John Ngombo	Retail of bovine and ovine meats	35 Niagara Street	New York
Agency 2	0111232223	John Ngombo canned meat production	Grocer's shop, beverages	9 Rome Street	Albany
Agency 3	CND8TB76SSDV	Meat production in New York state of John Ngombo	Butcher	4, Garibaldi Square	Long Island

Examples of the matching objects of the three data typologies

R(FirstName, LastName, Region, State)

Patrick	Metzisi	MM	Kenia

Patrick	Metzisi	Masai Mara	KE

(a) Two tuples

R1(FirstName, LastName, Region) R2(Region, State) R3(State, Continent)

Patrick	Metzisi	MM	→ MM	Kenia	→ Kenia	Africa
Patrick	Mezisi	Masai Mara	→ Masai Mara	KE	→ KE	Africa

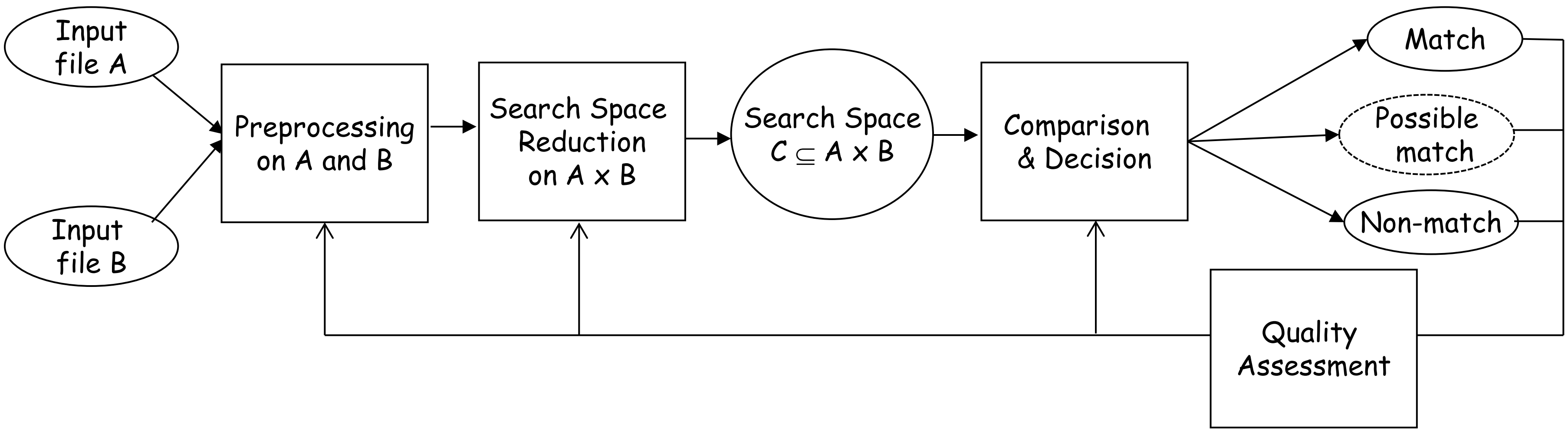
(b) Two hierarchical groups of tuples

```
<country>
  <name> Kenia </name>
  <cities> Nairobi, Mombasa, Malindi
  </cities>
  <lakes>
  <name> Lake Victoria </name>
  </lakes>
</country>
```

```
<country>
  Kenia
  <city> Nairobi  </city>
  <city> Mombasa  </city>
  <lakes>
    <lake> Lake Victoria </lake>
  </lakes>
</country>
```

(c) Two XML records

Relevant steps of object identification techniques



Example of string comparison

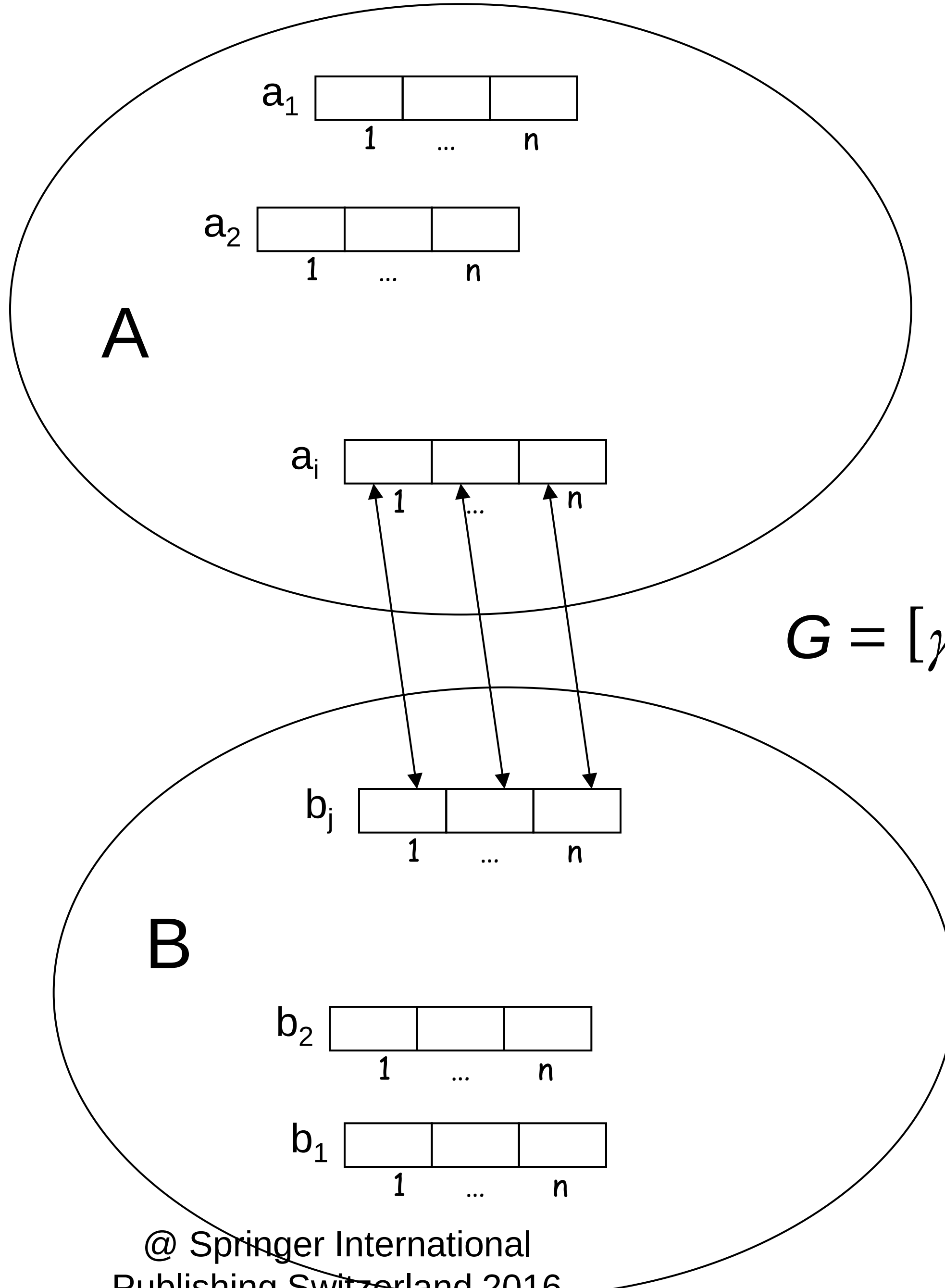
...	...	...
ATT	...	...
IBM Corporation	...	...
...	...	...

...	...	...
ATT Corporation	...	...
IBM Corporation	...	...
...	...	...

Object identification techniques

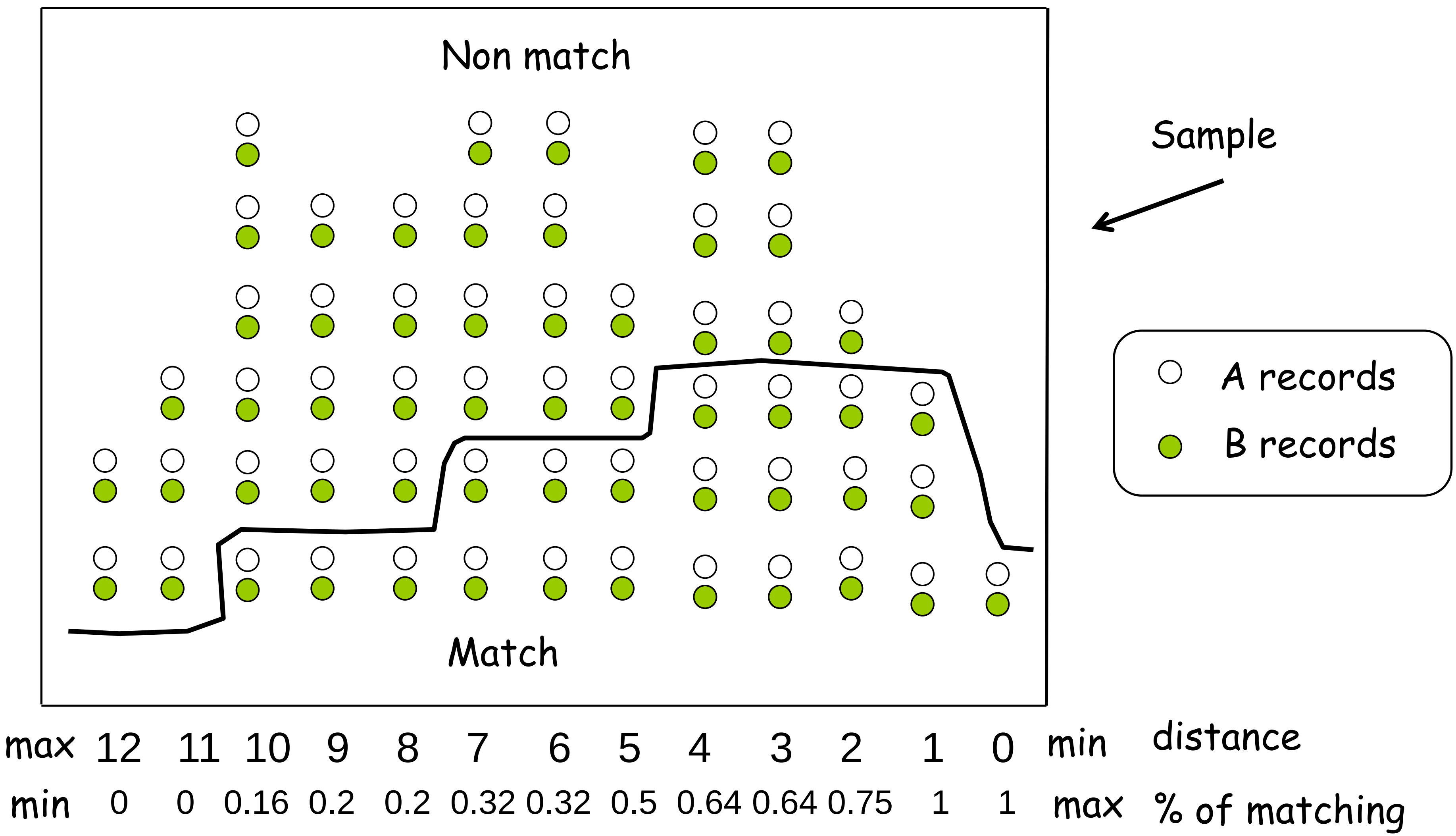
Name	Technical Area	Type of data
Fellegi and Sunter and extensions	probabilistic	Two files
Cost-based	probabilistic	Two files
Sorted Neighborhood and variants	empirical	Two files
Delphi	empirical	Two relational hierarchies
DogmatiX	empirical	Two XML documents
Intelliclean	knowledge-based	Two files
Atlas	knowledge-based	Two files

The Fellegi and Sunter record linkage formulation



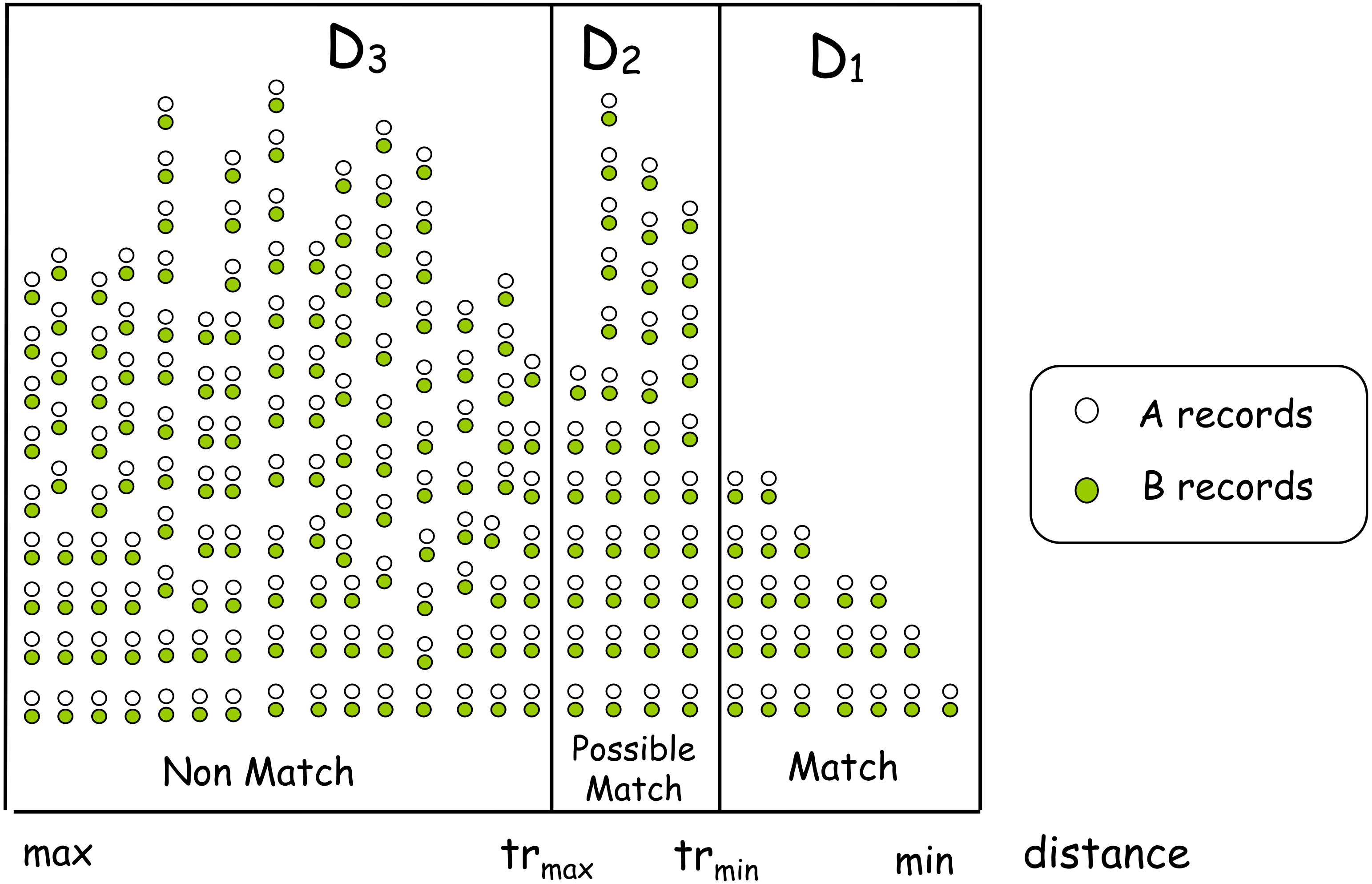
$$G = [\gamma_1^{ij}, \ldots, \gamma_n^{ij}]$$

Example distribution of match and not match in the sample as a function of distance among pairs



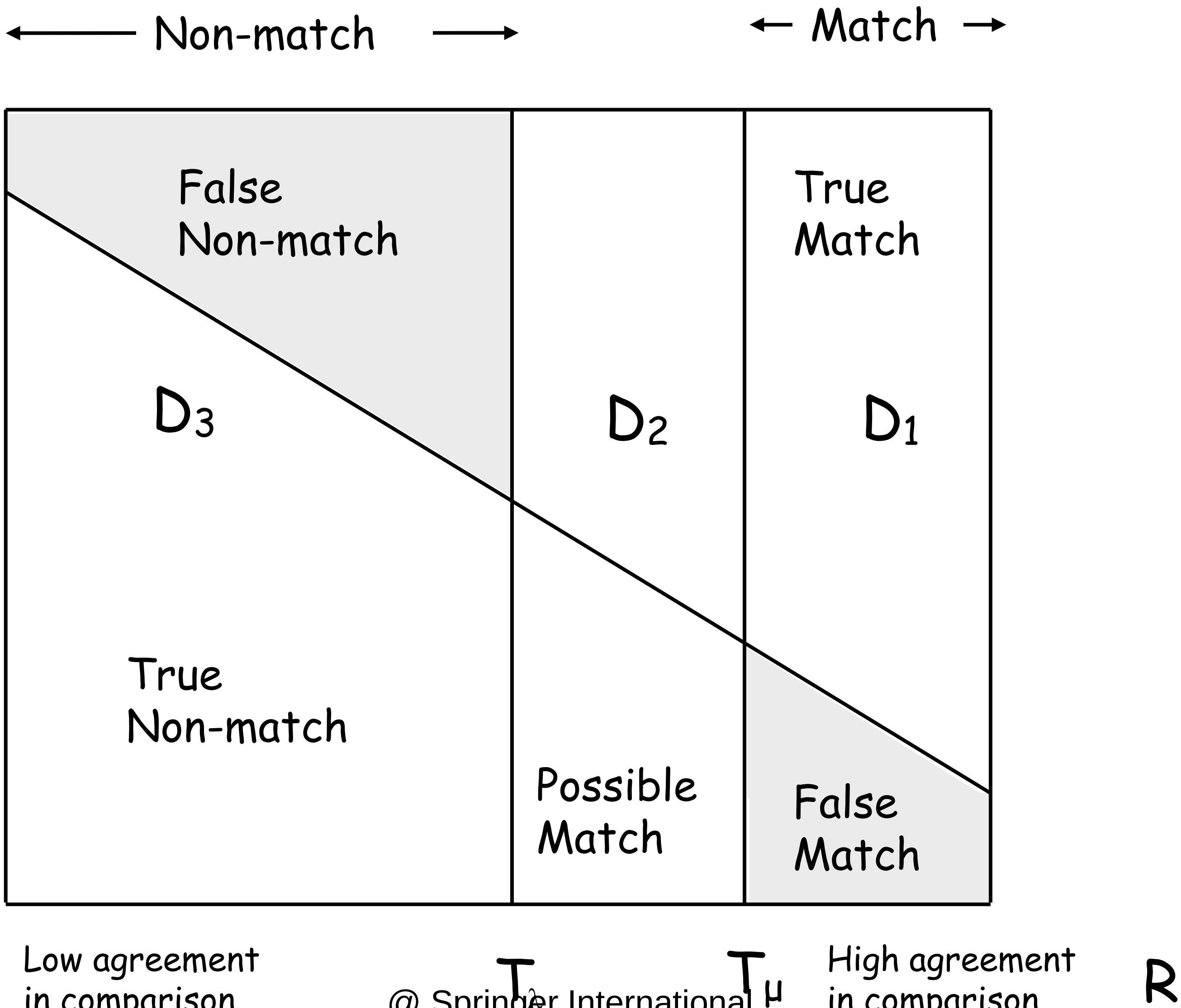
Vertical regions contain pairs of records ordered according to decreasing values of distance

Distribution of matching and unmatching applied to the universe U



Vertical regions contain pairs of records ordered according to decreasing values of distance

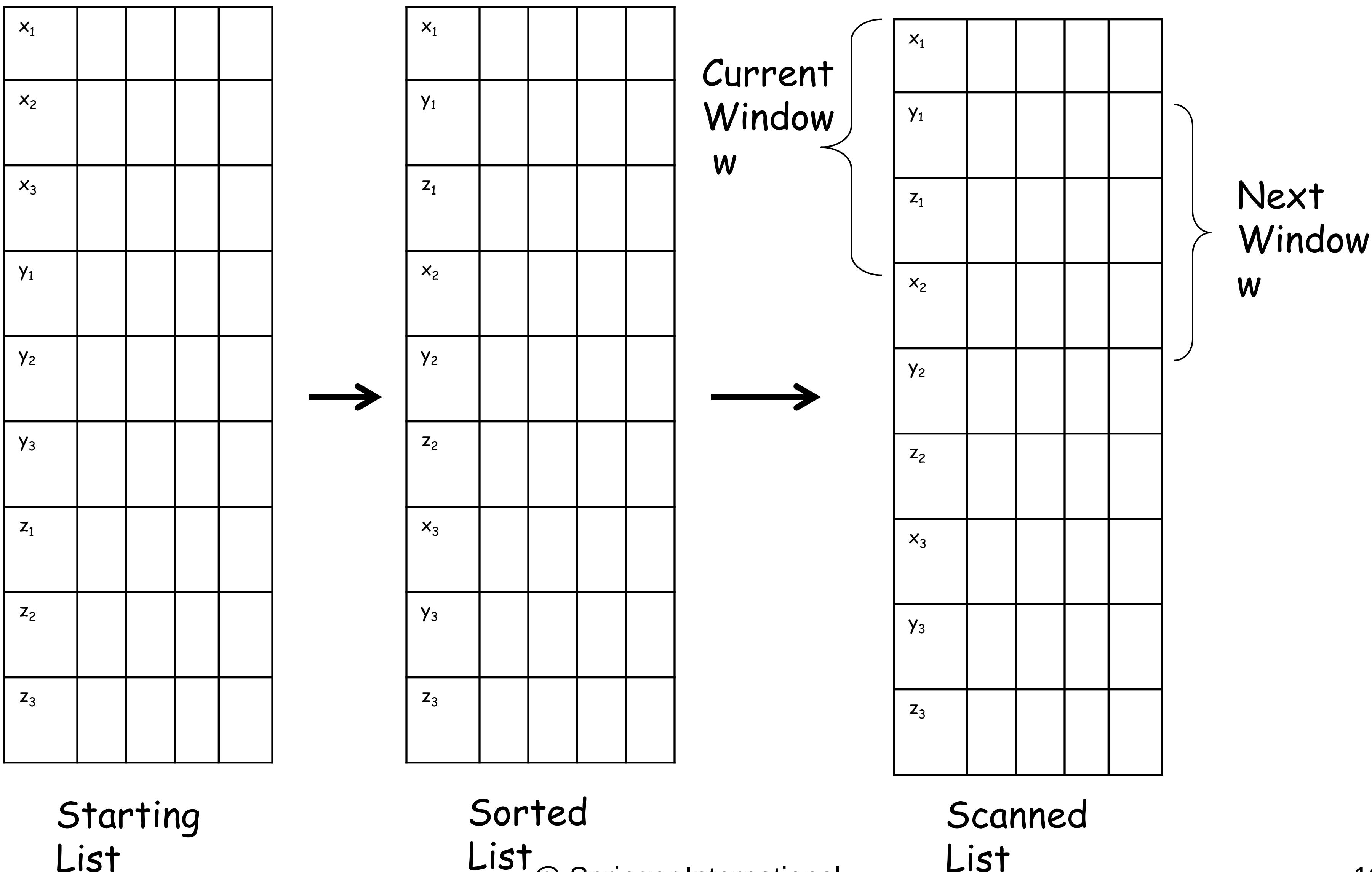
The regions of the Fellegi and Sunter decision model [281]



Costs corresponding to various decisions

Cost	Decision	Actual Matching
C_{10}	D_1	M
C_{11}	D_1	U
C_{20}	D_2	M
C_{21}	D_2	U
C_{30}	D_3	M
C_{31}	D_3	U

Phases of the SNM method



Three hierarchical relations

Person

PId	First name	Last Name	RegId
1	Patrick	Mezisi	1
2	Amanda	Rosci	2
3	George	Oado	3
4	John	Mumasia	4
5	Vusi	Oymo	7
6	Luyo	Msgula	5
7	Frial	Keyse	8
8	Wania	Nagu	6
9	Paul	Kohe	7

Administrative Region

RegId	RegionName	CtryId
1	MM	1
2	MM	2
3	Masai Mara	1
4	Eastern Cape	3
5	Free State	3
6	FS	4
7	HHohho	5
8	Lumombo	6

Country

CtryId	CountryName
1	KE
2	Kenia
3	SOA
4	South Africa
5	SWA
6	Swaziland

The Delphi algorithm

1. Process first the top most relation
2. Group relations below the top most relation into clusters of tuples
3. Prune each cluster according to properties of distance functions eliminating tuples that cannot be duplicates.
4. Compare pairs of tuples within each group according to two comparison functions and corresponding thresholds
    - ✓Textual similarity between two tuples
    - ✓Co-occurrence similarity between the children sets of the tuples
5. Decide for duplicates comparing a suitable combination of the two measures against a given threshold or a set of thresholds.
6. Dynamically update thresholds
7. Move one level down in the hierarchy

Bridging file example

A	A&B	B
$Tax_{1,1}$	$Name_1$, $Surname_1$, $Address_1$	$SocialService_{2,1}$
$Tax_{1,2}$	$Name_2$, $Surname_2$, $Address_2$	$SocialService_{2,2}$
…	…	…
…	…	…
$Tax_{1,n}$	$Name_n$, $Surname_n$, $Address_n$	$SocialService_{2,n}$

A small portion of the registry of US citizens

Record #	First Name	Last Name	State	Area	Age	Salary
1	Ann	Albright	Arizona	SW	65	70.000
2	Ann	Allbrit	Florida	SE	25	15.000
3	Ann	Alson	Louisiana	SE	72	70.000
4	Annie	Olbrght	Washington	NW	65	70.000
5	Georg	Allison	Vermont	NE	71	66.000
6	Annie	Albight	Vermont	NE	25	15.000
7	Annie	Allson	Florida	SE	72	70.000
8	George	Alson	Florida	SE	71	66.000

An example of the duplicate identification rule in Intelliclean

Define rule Restaurant_Rule
Input tuples: R1, R2
IF (R1.telephone = R2.telephone)
AND (ANY_SUBSTRING (R1.ID, R2.ID) = TRUE)
AND (FIELDSIMILARITY (R1.address = R2.address) > 0.8)
THEN
DUPLICATES (R1,R2) CERTAINTY = 0.8

The complete Intelliclean strategy

1. Preprocessing
Perform data type checks and format standardization
2. Processing
  2.1 The compared records are fed into an expert system engine together with a set of rules of the form IF <condition> THEN <action>.
  2.2 Check iteratively within a sliding window first Duplicate Identification rules and then Merge Purge rules using a basic production system to see which ones should fire based on the facts in the database, looping back to the first rule when it has finished.
  2.3 Perform transitive closure under uncertainty using an improved version of the multi-pass Sorted Neighborhood searching method
3. Human verification and validation stage
Human intervention to manipulate the duplicate record groups for which merge/purge rules are not defined

Examples of transformations

1. Soundex converts an item into a Soundex code. Items that sound similar have the same code
2. Abbreviation replaces an item with corresponding abbreviation (e.g., third → 3rd)
3. Equality compares two items to determine if each item contains the same characters in the same order
4. Initial computes if one item is equal to the first character of the other.
5. Prefix computes if one item is equal to a continuous subset of the other starting at the first character
6. Suffix computes if one item is equal to a continuous subset of the other starting at the last character
7. Abbreviation computes if one item is equal to a subset of the other (e.g., Blvd, Boulevard)
8. Acronym computes if all characters of one item string are initial letters of all items from the other string

Two relations

Relation1

LastName	Address	City	Region	Telephone
Ngyo	Mombsa Boulevard	Mutu	MM	350-15865

Relation2

LastName	Address	Region	Telephone
Ngoy	Mombasa Blvd.	Masai Mara	350-750123

Notation on matching decision cases

M	Actual match w.r.t. real world
U	Actual non match w.r.t. real world
FP	Declared match while actual non match
FN	Declared non-match while actual match
TP	Declared match while actual match
TN	Declared non match while actual non match

Comparison of decision methods

Technique	Input	Output	Objective	Human interaction	Selection/Construction of a representative for the matching records
Fellegi&Sunter	γ vector of comparison functions Estimation of $T\mu$ and $T\lambda$ m- and u-probabilities	For each record pair, decision on match, non-match, possible match with given error rates	Low error rates (false match and false non-match) Minimization of possible matches	Clerical Review of possible matches	No
Cost Based	Matrix of costs of decision rules m- and u- probabilities	For each record pair, decision on match, non-match, possible match with given error rates	Minimization of cost of errors (false match and false non-match)	Clerical Review of possible matches Matrix of costs of decision rules	No
SNM	Declarative rules encoding domain knowledge (for tuple level decision) Comparison functions (for attribute value decision) Threshold (for attribute value decision)	For each record pair, decision on match or non-match	Precision/Recall tradeoff	Choice of the matching key Threshold Specification Decision Rules	No (only for incremental SNM)
Priority-Queue	Smith Waterman comparison function Threshold (for tuple value decision)	For each record pair, decision on match or non-match	Precision/Recall tradeoff	Threshold Specification	No
Delphi	Textual Comparison Function Co-occurrence metric Set of thresholds (dynamically updated)	For each record pair, decision on match or non-match	Precision/Recall tradeoff	None	No
DogMatix	XML Threshold similarity (object level)	For each XML element pair, decision on match or non-match	Precision/Recall tradeoff	Selection of candidates Threshold Specification	No
IntelliClean	Duplicate Identification Rules (for tuple decision) Merge Purge Rules (for tuple decision) Set of thresholds (for attribute comparison and for tuple merging)	For each record pair, decision on match or non-match Merged Result for matching records	Precision/Recall tradeoff User controlled confidentiality for merging	Duplicate Identification,Merge/Purge Rules Specification Threshold Specification Human verification for merging duplicates when rules are not specified	Yes
Atlas	Learnt Decision rules Set of domain independent transformations Thresholds	For each record pair, decision on match or non-match	Precision/Recall tradeoff	Mapping rule learning	No

Metrics used by to evaluate object identification by empirical techniques and related results

Technique	Metrics	Synttc/ Real Dta	Data Dimensions	Results
SNM	Precision False Positive Percentage	Synthetic	1..000.000 records (120Mb)	Precision 50%-70% on independent pass Precision close to 90% with transitive closure False Positive Percentage not significant (0.05 - 0.2%)
	Precision False positive Percentage False negative Percentage	Real	128.438 records (13,6 Mb)	Not significant False Negatives Percentage Not significant False Positive Percentage
Priority-Queue	Precision Efficiency (Number of comparisons)	Synthetic	From around 300.000 records to around 480,000 records	Precision similar to SNM Efficiency : 5 times less than SNM
	Efficiency (Number of comparisons)	Real	255. 000 records	Precision not provided as for real data difficult to identify actual duplicate s Efficiency - Number of reduced comparisons similar to the one for the synthetic data set
Delphi	False Positive Percentage False Negative Percentage	Real	270.000 records	False Positive Percentage less than 25% False Negative Percentage around 20%
DogMatix	Precision Recall	Real	Experiment1:1000 records Experiment2:10000 records	For similarity measure: Experiment 1: Precision 70-100% Experiment 1: Recall: 2%-35% Experiment 2: Precision 60-100%
IntelliClean	Precision	Real	Experiment1: 856 records Experiment2: 22.122 records	Experiment 1: Precision 80% Experiment 1: Less than 8% Recall Experiment 2 :Precision: 100% Experiment 2 :Recall:100%
Atlas	Precision (accuracy)	Real	Experiment1: 1.000 records Experiment2: 10.000 records	Experiment 1: Precision 100% Experiment 2: Precision 99%