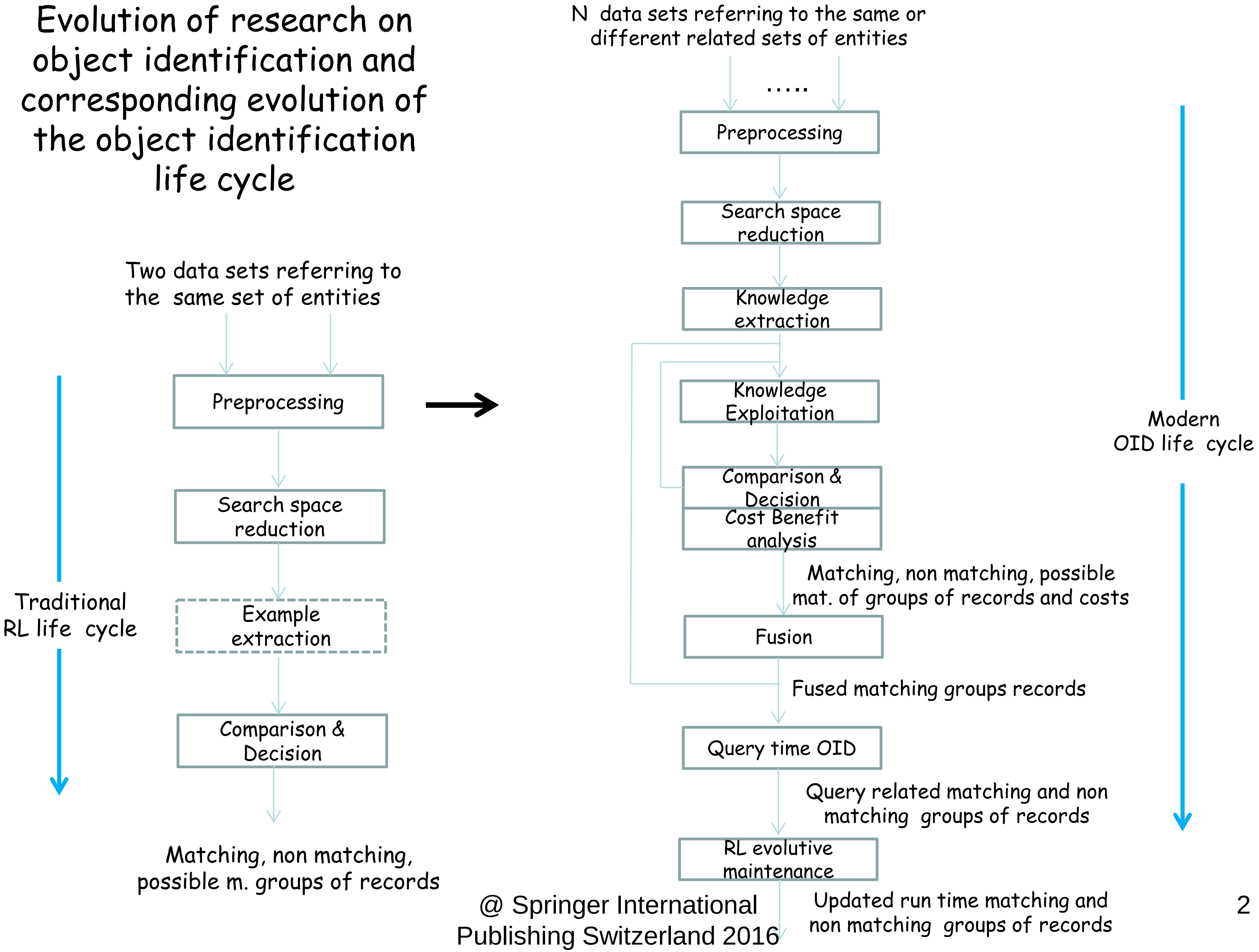
C. Batini & M. Scannapieco
Data and Information Quality Book Figures

Chapter 9: Recent Advances in Object Identification

Evolution of research on object identification and corresponding evolution of the object identification life cycle



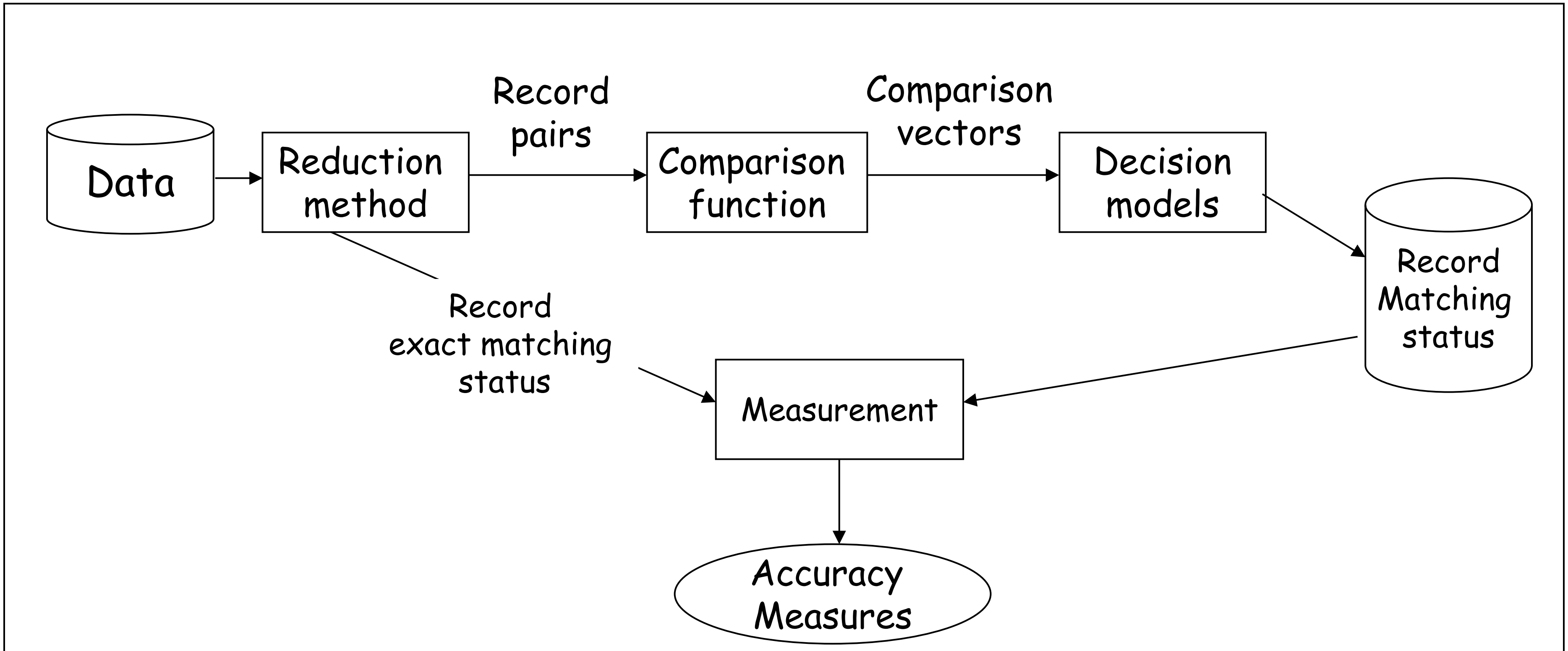
Two data sets referring to the same set of entities
Preprocessing
Search space reduction
Example extraction
Comparison & Decision
Matching, non matching, possible m. groups of records
Traditional RL life cycle
N data sets referring to the same or different related sets of entities
.....
Preprocessing
Search space reduction
Knowledge extraction
Knowledge Exploitation
Comparison & Decision
Cost Benefit analysis
Matching, non matching, possible mat. of groups of records and costs
Fusion
Fused matching groups records
Query time OID
Query related matching and non matching groups of records
RL evolutive maintenance
Updated run time matching and non matching groups of records
Modern OID life cycle

@ Springer International Publishing Switzerland 2016

Comparison of quality measures in the entity space and in the comparison space

Metric	Entity Space	Comparison Space
Precision	72,2%	72,2%
Recall	92,8%	92,8%
F-measure	81,2%	81,2%
Accuracy	94,3%	99,9%
Specificity	94,5%	99,95
False positive rate	5,4%	0.000005%

Architecture of Tailor



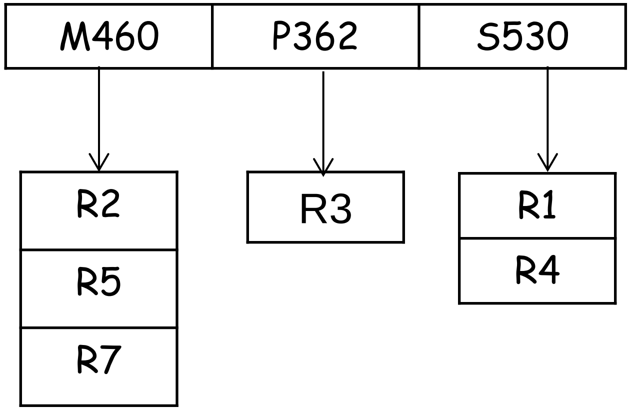
Examples of citation domain string matching from [25]

Id	Left	Right
1	Katayama,T., 2A hierarchical and functional software process description and its enaction", Proc. 11th ICSE, IEEE, 1989, pp.343-352	T. Katayama, “A hierarchical and functional software process description and its enaction," In: Proceedings of the Eleventh Int. Conf. On Soft. Eng. Pages: 343{352, IEEE Computer Society Press, Pittsburgh, PA, Jan 1989.
2	Knuth, D., The art of Computer Programming, Vol. III, Addison-Wesley, (1973).	8. D. Knuth, The art of Computer Programming, Volume 3: Sorting and Searching, Addison-Wesley, Reading, MA, 1973.
3	[ESWARAN76] Eswaran, K. P., J. N. Gray, R. A. Lorie, I. L. Traiger, \The notions of consistency and predicate locks in a database system", Communications of the ACM, Vol. 19, No. 11, November, 76	[14] K. P. Eswaran, J. N. Gray, R. A. Lorie, and I. L. Traiger, \The notions of consistency and predicate locks in a database system," Commun. Assoc. Comput. Mach., Vol. 19, No. 11, Nov. 1976

Example of traditional blocking (here and in the following of the section examples are inspired to [139])

Identifier	Surname	BK (Soundex encoding)
R1	Smith	S530
R2	Miller	M460
R3	Peters	P362
R4	Smyth	S530
R5	Millar	M460
R6	Miller	M460

a. Records table with BKVs


M460 P362 S530

R2 R5 R7

R3

R1 R4

b. Inverted index data structure

Example of traditional sorted neighborhood

Window position	BK (Surname)	Identifier
1	Millar	R6
2	Miller	R2
3	Miller	R8
4	Myler	R4
5	Peters	R3
6	Smith	R1
7	Smyth	R5
8	Smyth	R7

a. Records table with BKVs and window positions

Window range	Candidate record pairs
1-3	(R6,R2), (R6,R8), (R2,R8)
2-4	(R2,R8), (R2,R4), (R8,R4)
3-5	(R8,R4), (R8,R3), (R4,R3)
4-6	(R4,R3), (R4,R1), (R3,R1)
5-7	(R3,R1), (R3,R5), (R1,R5)
6-8	(R1,R5), (R1,R7), (R5,R7)

b. Record pairs in windows

Example of sorted neighborhood based on inverted index

Window position	BK (Surname)	Identifier
1	Millar	R6
2	Miller	R2, R8
3	Myler	R4
4	Peters	R3
5	Smith	R1
6	Smyth	R5,R7

a. Records table with inverted index

Window range	Candidate record pairs
1-3	(R6,R2), (R6,R8), (R6,R4), (R2,R8), (R2,R4), (R8,R4)
2-4	(R2,R8), (R2,R4), (R8,R4), (R8,R4),(R8,R3),(R4,R3)
3-5	(R4,R3), (R4,R1), (R3,R1)
4-6	(R3,R1), (R3,R5), (R3,R7), (R1,R5), (R1,R7), (R5,R7)

b. Record pairs in windows

Example of suffix array based blocking

Identifier	BK (Given Name)	Suffixes
R1	Catherine	Catherine, atherine, therine, herine, erine, rine
R2	Katherina	Katherina, atherina, therina, herina, erina, rina
R3	Catherina	Catherina, atherina, therina, herina, erina, rina
R4	Catrina	Catrina, atrina, trina, rina
R5	Katrina	Katrina, atrina, trina, rina

a. Records table with BK and suffixes

Suffix	Identifier
atherina	R2,R3
atherine	R1
atrina	R4,R5
catherina	R3
catherine	R1
catrina	R4
erina	R2,R3
erine	R1
herina	R2,R3
herine	R1
katherina	R2
katrina	R5
rina	R2,R3,R4,R5
rine	R1
therina	R2,R3
therine	R1
trina	R4,R5

b. Sorted suffix-array

Examples of blocking predicates from [76]

Domain	Blocking Predicate
Census data	Same first three chars in Last Name
Product normalization	Common token in Manufacturer
Citations	Publication Year same or off-by-one

Blocking key values for a sample record from [76]

Author	Year	Title	Venue	Other
Freund, Y.	(1995)	Boosting a weak learning algorithm by majority	Information and computation	(121(2), 256-285

a. Sample record

Predicate	Author	Title	Venue	Year	Other
Contain common token	(freund, y)	(boosting, a, weak, learning, algorithm, by, majority)	(information, computation)	(1995)	(121,2,256,285)
Exact match	(“freund y”)	(“Boosting a weak learning algorithm by majority”)	(“information and computation)	(“1995”)	(“121 2 256 285”)
Same 1st three Chars	(fre)	(boo)	(inf)	(199)	(121)
Contain same or off-by-one integer	-	-	-		(120_121, 121_122, 1_2, 2_3,,255_256, 256_257, 284_285,285_286)

b. Blocking predicates and key sets produced by their indexing functions for the record

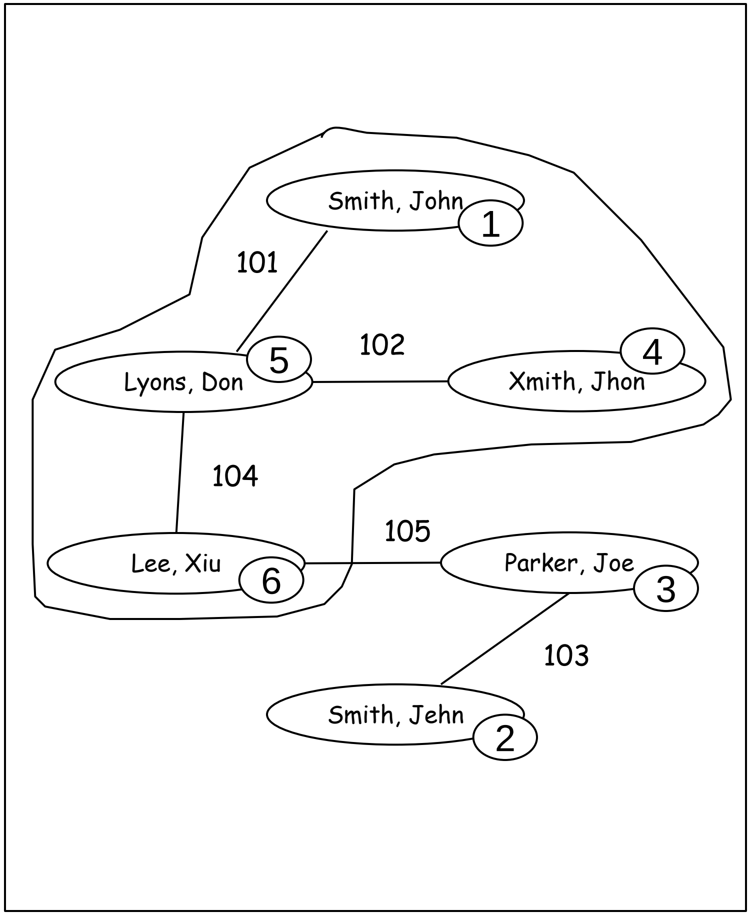
Example of semantic blocking from [473]

1	Smith, John	
2	Smith, Jehn	
3	Parker, Joe	
4	Xmith, Jhon	
5	Lyons, Don	
6	Lee, Xiu	

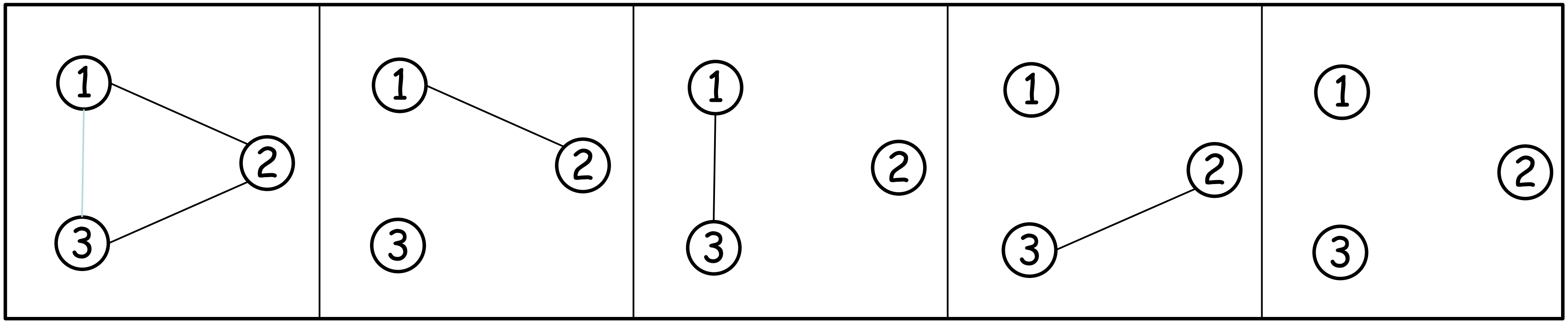
101	Title1	.....
102	Title2	
103	Title3	
104	Title4	
105	Title5	

....	1	101	....
	5	101	
	4	102	
	5	102	
	2	103	
	3	103	
	5	104	
	6	104	
	3	105	
	6	105	

→



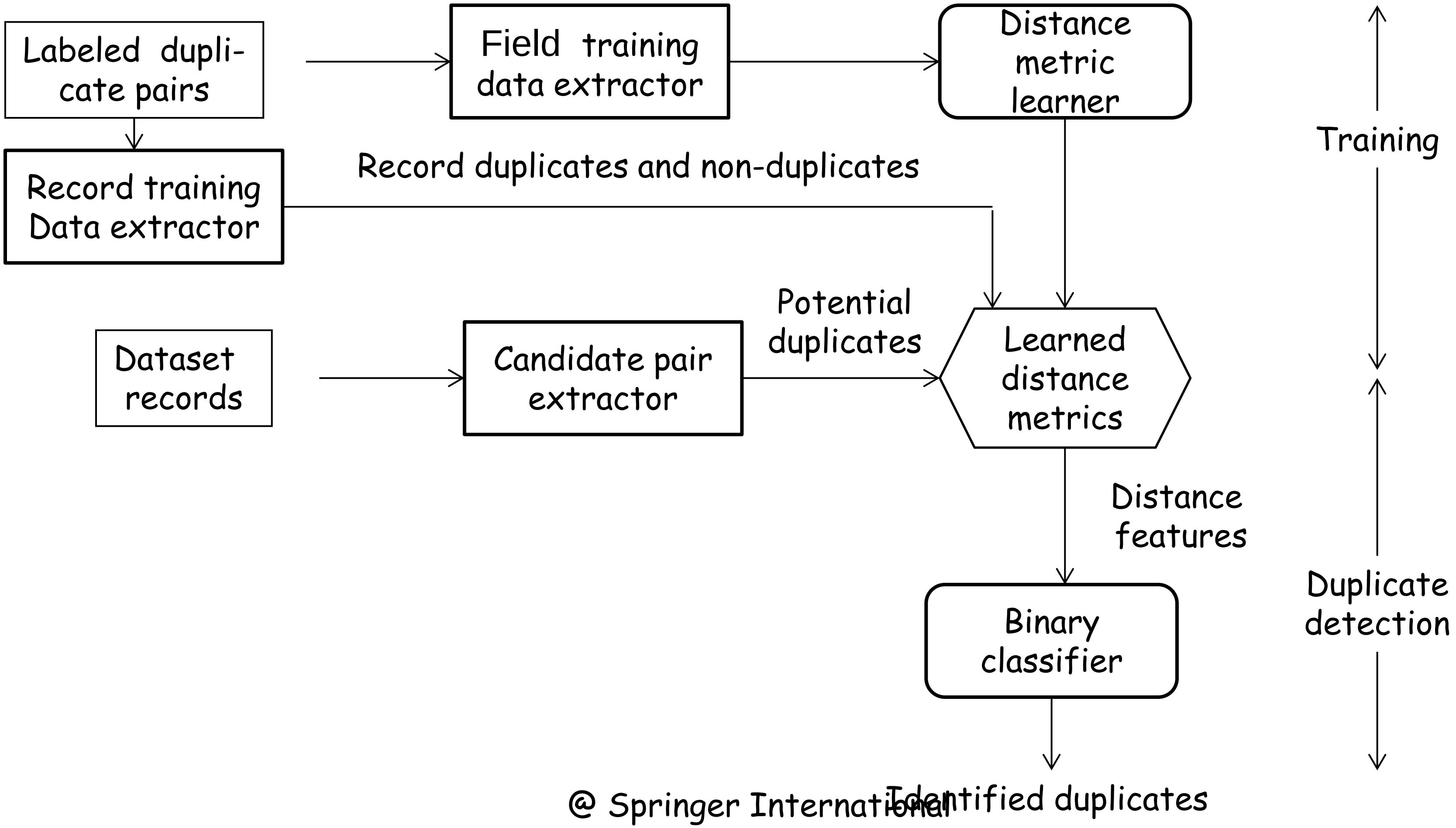
Possible paths of agreement for three data sets in [536]



Examples of features in [145]

Name of Feature	**Description**
SubstringMatch	true iff one of the two strings is a substring of the other
PrefixMatch	true iff one of the two strings is a prefix of the other
StrongNumberMatch	true iff the two strings contain the same number
Edit distance	usual meaning
Jaccard distance	usual meaning

Phases of knowledge extraction and exploitation in [75]



@ Springer International Publishing Switzerland 2016

Example of weight vectors from [138]

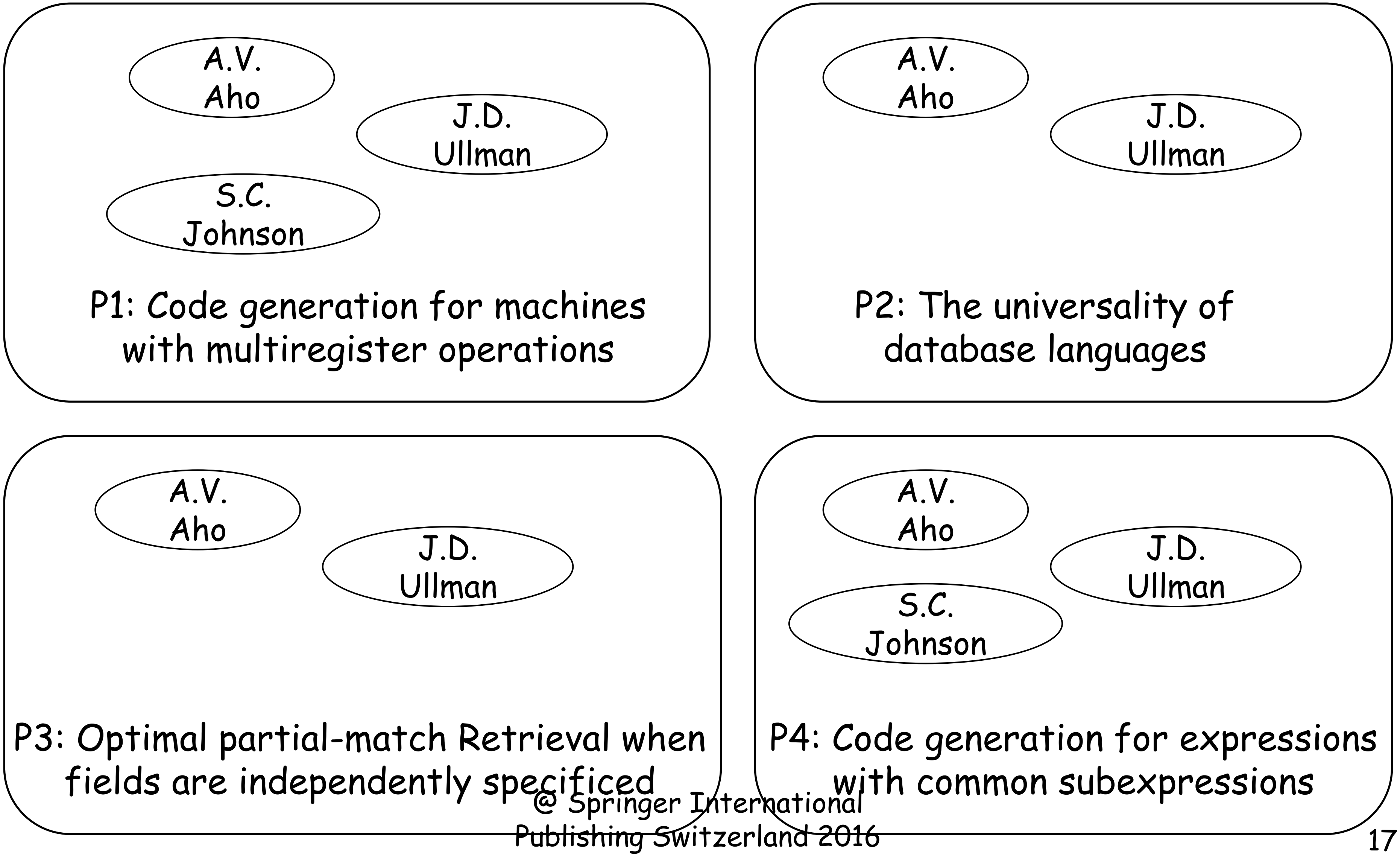
Record	Name		Address		
R1	Christine	Smith	42	Main	Street
R2	Christina	Smith	42	Main	St.
R3	Bob	O'Brian	11	Smith	Rd
R4	Robert	Bryee	12	Smythe	Road

a. Four record examples

WV(R1,R2): [0.9, 1.0, 1.0, 1.0, 0.9]
WV(R1,R3): [0.0, 0.0, 0.0, 0.0, 0.0]
WV(R1,R4): [0.0, 0.0, 0.5, 0.0, 0.0]
WV(R2,R3): [0.0, 0.0, 0.0, 0.0, 0.0]
WV(R2,R4): [0.0, 0.0, 0.5, 0.0, 0.0]
WV(R3,R4): [0.7, 0.3, 0.5, 0.7, 0.9]

b. Corresponding wieght vectors

An example Author/Paper resolution problem from [66]. Each box represents a paper reference (in this case unique) and each oval represents an author reference



A.V. Aho

J.D. Ullman

S.C. Johnson

P1: Code generation for machines with multiregister operations

A.V. Aho

J.D. Ullman

P2: The universality of database languages

A.V. Aho

J.D. Ullman

P3: Optimal partial-match Retrieval when fields are independently specificed

A.V. Aho

J.D. Ullman

S.C. Johnson

P4: Code generation for expressions with common subexpressions

Example of exploitation of context information in [179]

```
Person  (name, email, *coAuthor, *emailContact)
Article ( title, year, pages, *authoredBy, *publishedIn)
Conference (name, year, location)
Journal (name, year, volume, number)
```

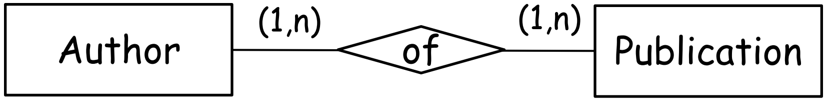
Related records and corresponding Entity Relationship schema as adapted from [353]

(A1: “Dave White”; “Intel”)
(A2: “Don White”; “CMU”)
(A3: “Susan Grey”; “MIT”)
(A4: “John Black”; “MIT”)
(A5: “Joe Brown”; unknown)
(A6; “Liz Pink”; unknown)

a. Authors records

(P1: “Databases….”; “John Black”; “Don White”)
(P2: “Multimedia……”; “Sue Gray”; “D. White”)
(P3: “Title3….”; “Dave White”)
(P4: “Title4…”; “Don White”; “Joe Brown”)
(P5: “Title5…”; “Joe Brown”; “Liz Pink”)
(P6; “Title6…”; “Liz Pink”; “D. White”)

b. Publications records

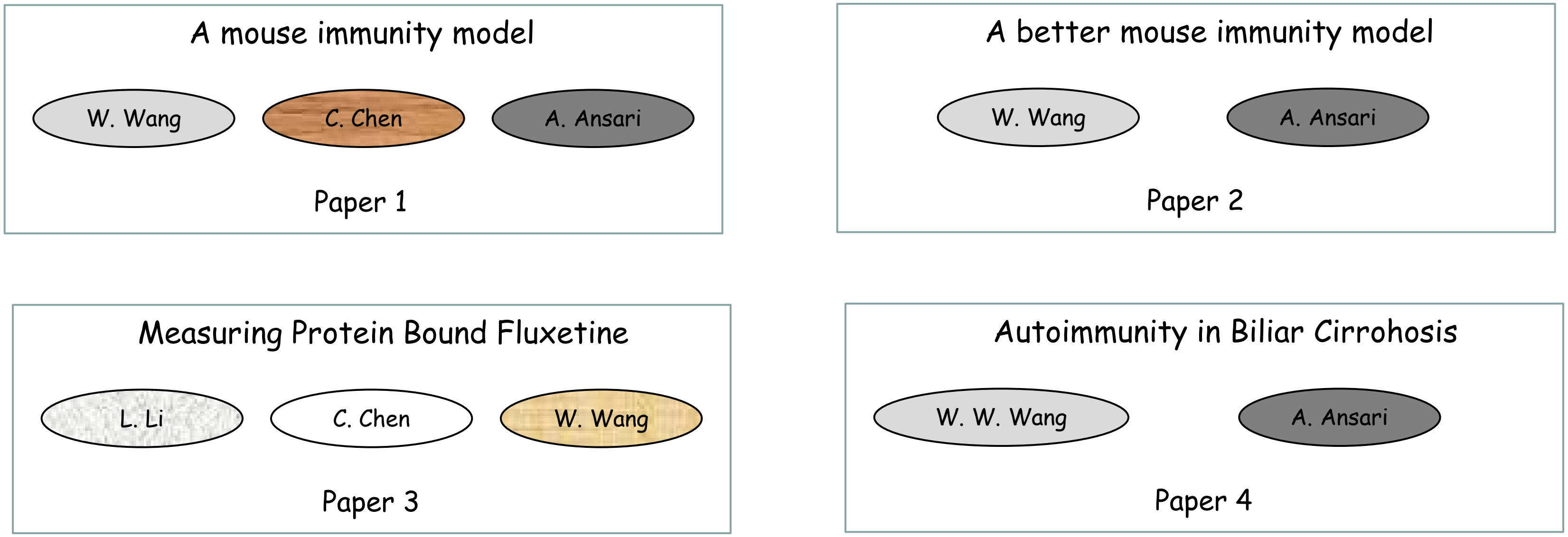


c. Corresponding Entity Relationship schema

Bibliographic example from [68]

(1) W. Wang , C. Chen, A. Ansari – A mouse immunity model
(2) W. Wang, A. Ansari – A better mouse immunity model
(3) L. Li, C. Chen, W. Wang – Measuring protein-bound fluxetine
(4) W.W. Wang, A. Ansari – Autoimmunity in biliar cirrhosis

a. A set of four papers



A mouse immunity model

W. Wang
C. Chen
A. Ansari

Paper 1

A better mouse immunity model

W. Wang
A. Ansari

Paper 2

Measuring Protein Bound Fluxetine

L. Li
C. Chen
W. Wang

Paper 3

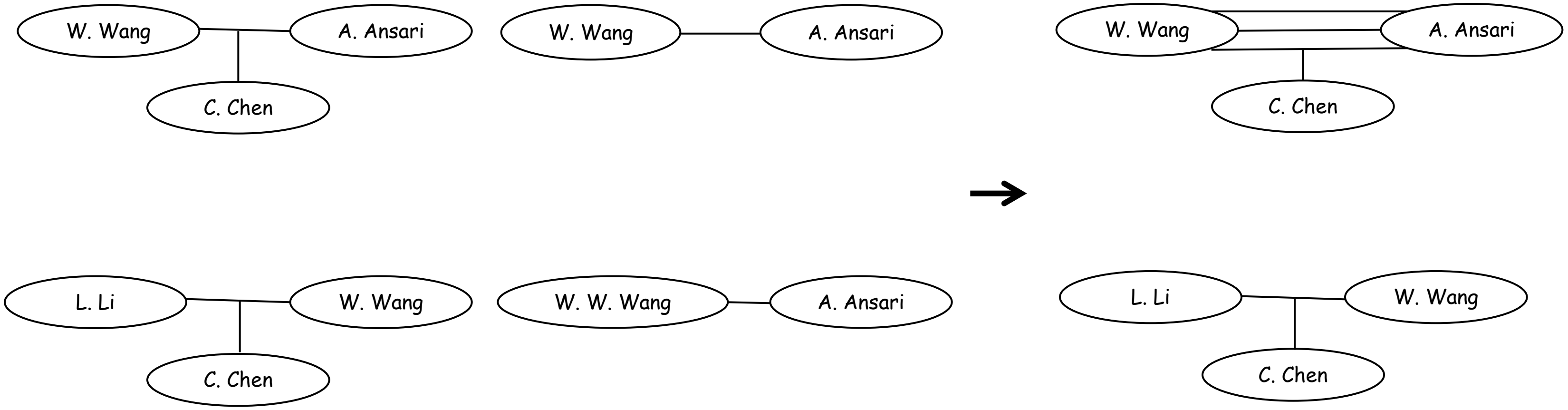
Autoimmunity in Biliar Cirrohosis

W. W. Wang
A. Ansari

Paper 4

b. References to the same author are identically shaded

Reference graph and entity graph for the author resolution example in [68]



a. Reference graph

b. Entity graph

Motivating example in [159]

PublID	Author	Title	Venue	VenueID	Year
0	X.Li	Predicting the stock market	KDD	10	2010
1	X.Li	Predicting the stock market	Int’l Conference on Knowledge Discovery	20	2010
2	J.Smith	Semi-Definite Programming for Link Prediction	KDD	30	2011
3	J.Smith	Semi-Definife Programing for Link Prediction	Conference on Knowledge Discovery	40	2011

Example of aggregate constraint in [121]

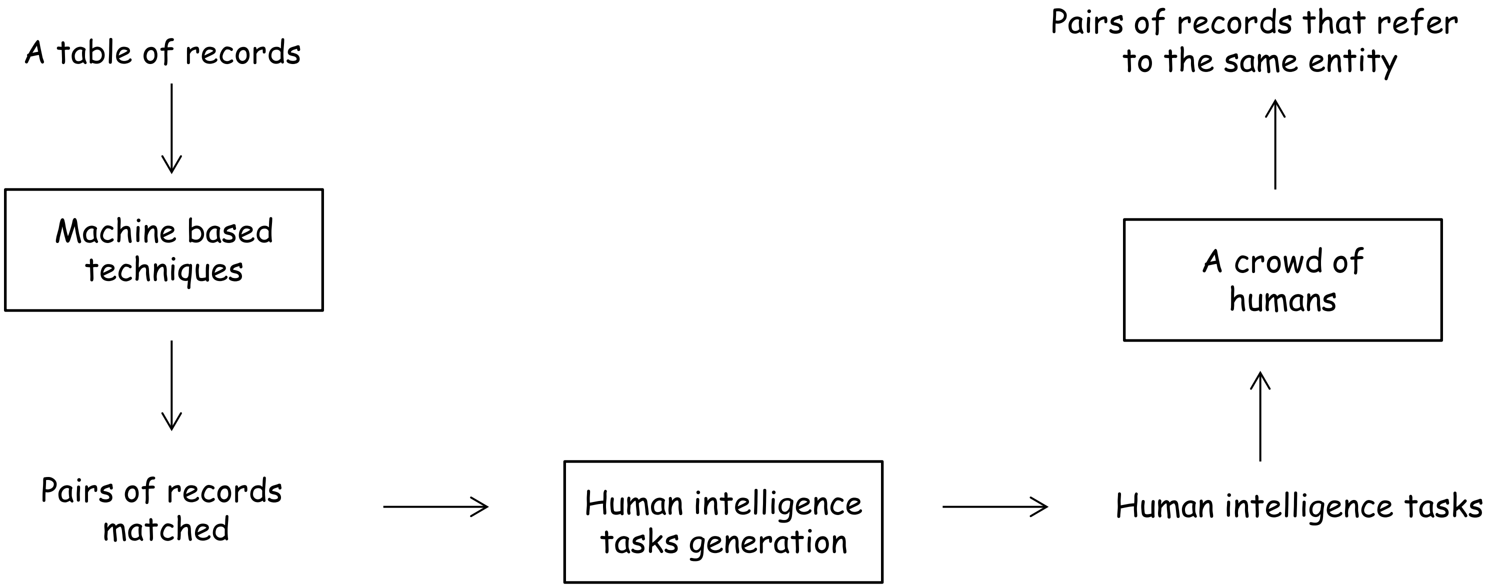
Member	Fees stored	Fees derived
John Doe	100	130
J. Doe	40	10
…………	…..	…..

First scenario

Member	Fees stored	Fees derived
John Doe	100	100
J. Doe	40	10
…………	…..	…..

Second scenario

Example of hybrid human-machine workflow proposed in [642]



Example proposed in [284]

a. Data sources

Source	Name	Phone	Address
S1	Microsofe Corp.	xxx-1255	1 Microsoft Way
	Microsofe Corp.	xxx-9400	1 Microsoft Way
	Macrosoft Inc.	xxx-0500	2 Sylvan W.
S2	Microsoft Corp.	xxx-1255	1 Microsoft Way
	Microsofe Corp.	xxx-9400	1 Microsoft Way
	Macrosoft Inc.	xxx-0500	2 Sylvan Way
S3	Microsoft Corp.	xxx-1255	1 Microsoft Way
	Microsoft Corp.	xxx-9400	1 Microsoft Way
	Macrosoft Inc.	xxx-0500	2 Sylvan Way
S4	Microsoft Corp.	xxx-1255	1 Microsoft Way
	Microsoft Corp.	xxx-9400	2 Sylvan Way
	Macrosoft Inc.	xxx-0500	1 Microsoft Way
S5	Microsoft Corp.	xxx-1255	1 Microsoft Way
	Microsoft Corp.	xxx-9400	1 Microsoft Way
	Macrosoft Inc.	xxx-0500	2 Sylvan Way
S6	Microsoft Corp.	xxx-2255	1 Microsoft Way
	Macrosoft Inc.	xxx-0500	2 Sylvan Way
S7	MS Corp.	xxx-1255	1 Microsoft Way
	Macrosoft Inc.	xxx-0500	2 Sylvan Way
S8	MS Corp.	xxx-1255	1 Microsoft Way
	Macrosoft Inc.	xxx-0500	2 Sylvan Way
S9	Macrosoft Inc.	xxx-0500	2 Sylvan Way
S10	MS Corp.	xxx-0500	2 Sylvan Way

b. Real-world entities

Name	Phone	Address
Microsofe Corp., Microsofe Corp, MS Corp.	xxx-1255 Xxx-9400	1 Microsoft Way
Microsoft Inc.	xxx-0500	2 Sylvan Way, 2 Sylvan W.

Example from [54]

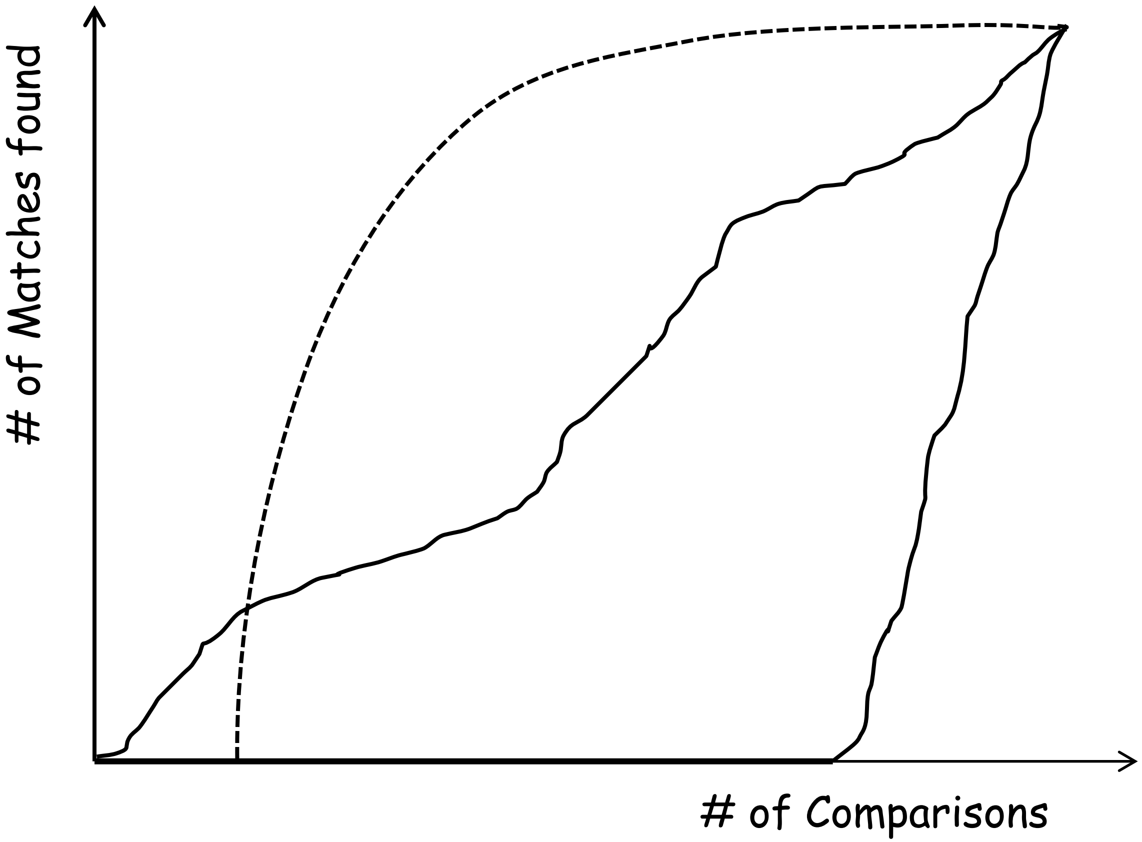
	Name	Phone	E-mail
r1	JohnDoe	235-2635	jdoe@yahoo
r2	J.Doe	234-4358	
r3	JohnD.	234-4358	jdoe@yahoo

a. An instance of persons representing persons

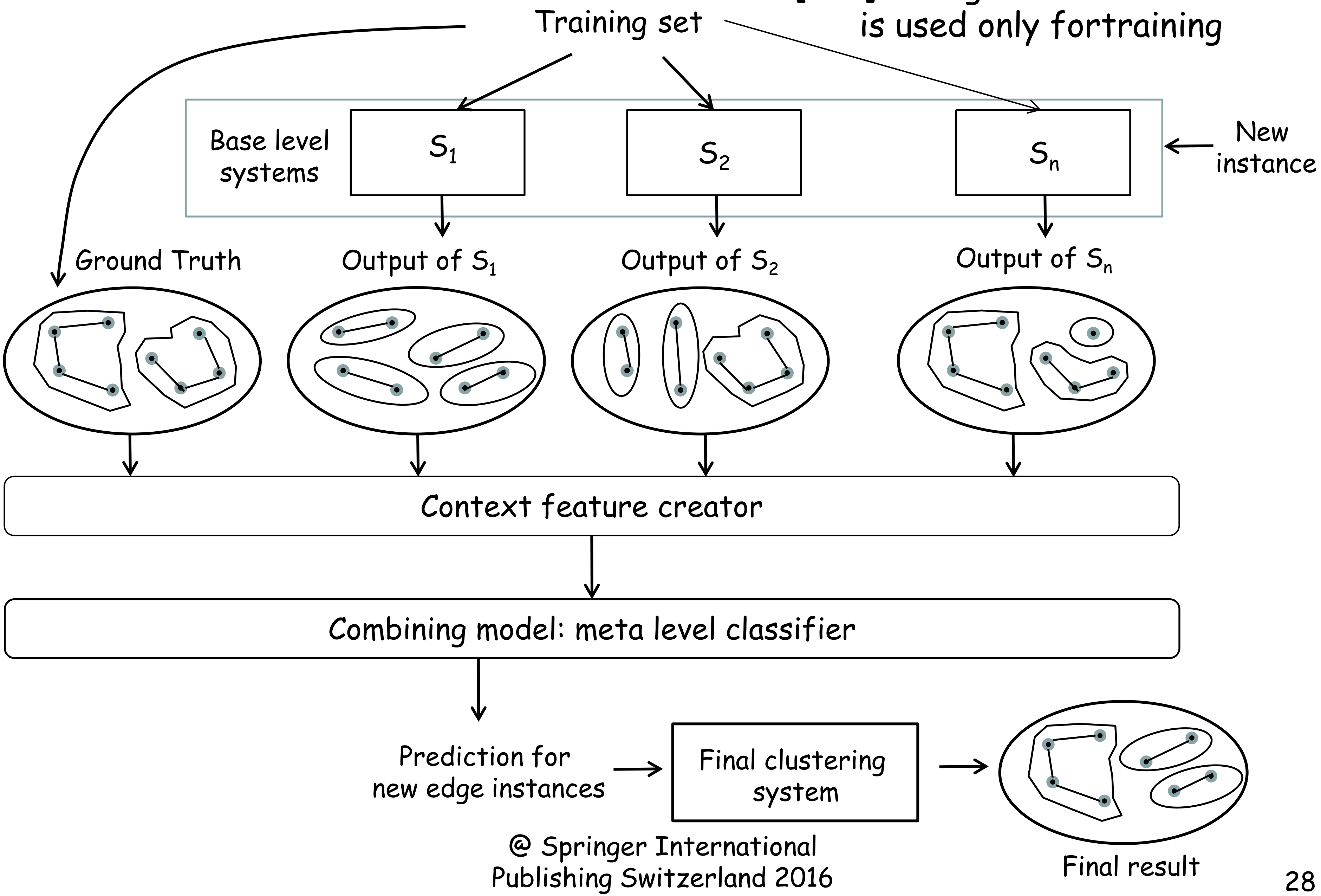
r4	John Doe	234-4358 235-2635	jdoe@yahoo

b. A new record generated by merging

Pay-as-you-go approach in [664]



The framework presented in [127]. The ground truth cluster is used only fortraining



Training set

Base level systems

S_1

S_2

S_n

New instance

Ground Truth

Output of S_1

Output of S_2

Output of S_n

Context feature creator

Combining model: meta level classifier

Prediction for new edge instances

Final clustering system

Final result

@ Springer International Publishing Switzerland 2016

Example from [17]

P_id	P_title	Cited	Venue	Authors	Year
P1	Towards efficient entity resolution	65	Very Large Data Bases	Alon Halevy	2000
P7	Towards efficient ER	45	VLDB	Alon Halevy	2000
P2	Entity Resolution on dynamic data	25	ACM SIGMOD	Alon Halevy, Jane Doe	2005
P3	ER on dynaminc data	20	Proc of ACM SIGMOD Conf	A.Y. Halevy, J. Doe	2005
P4	Entity Resolution for dynamic data	15	SIGMOD Conf.	A. Halevy, Jane D.	2005
P5	Entity Resolution for Census data	10	ICDE Conf.	Alon Halevy	2002
P6	ER on census data	5	Proc of ICDE Conf	Alon Y. Halevy	2002

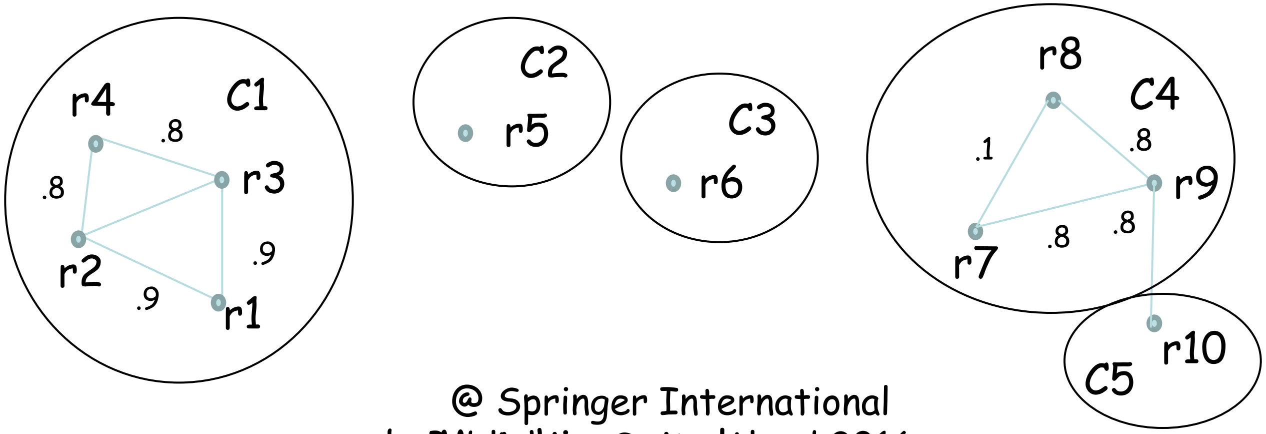
Relation R after being clustered using an entity resolution algorithm

Cluster	P_id	P_title	Cited	Venue	Authors	Year
C1	P1, P7	Towards efficient entity resolution	110	Very Large Data Bases	Alon Halevy	2000
C2	P2, P3, P4	Entity Resolution on dynamic data	60	Proc of ACM SIGMOD Conf	Alon Halevy, Jane Doe	2005
C3	P5, P6	Entity Resolution for Census data	15	ICDE Conf. Proc of ICDE Conf	Alon Halevy	2002

Original business listings and object identification results in [278]

	BizId	Id	Name	Street address	City	Phone
D0	B1	r1	Starbucks	123 MISSION ST STE ST1	SAN FRANCISCO	4155431510
	B1	r2	Starbucks	123 MISSION ST	SAN FRANCISCO	4155431510
	B1	r3	Starbucks	123 Mission St	SAN FRANCISCO	4155431510
	B2	r4	Starbucks Coffee	340 MISSION ST	SAN FRANCISCO	*4155431510*
	B3	r5	Starbucks Coffee	333 MARKET ST	SAN FRANCISCO	415534786
	B3	r6	Starbucks	MARKET ST	San Francisco	
	B4	r7	Starbucks Coffee	52 California St	San Francisco	4153988630
	B4	r8	Starbucks Coffee	52 CALIFORNIA ST	SAN FRANCISCO	4153988630
	B5	r9	Starbucks Coffee	295 California St	SAN FRANCISCO	415986234
	B5	r10	Starbucks	295 California ST	SF	

a. Original business listings



b. Matching results

New updates in [278]

	BizId	Id	name	Street address	city	phone
D1	B6	r11	Starbucks Coffee	201 Spear Street	San Francisco	4159745077
D2	B3 B3	r12 r13	Starbucks Coffee Starbucks	MARKET STREET 333 MARKET ST	San Francisco San Francisco	4155434786 4155434786
D3	B1 B1	r14 r15	Starbucks Starbucks	123 MISSION ST STE ST1	SAN FRANCISCO San Francisco	4155431510 4155431510
D4	B5 B4	r16 r17	Starbucks Starbucks	295 CALIFORNIA ST 52 California St	SAN FRANCISCO SF	4155431510 4153988630

Records to match and evolving rules in [663]

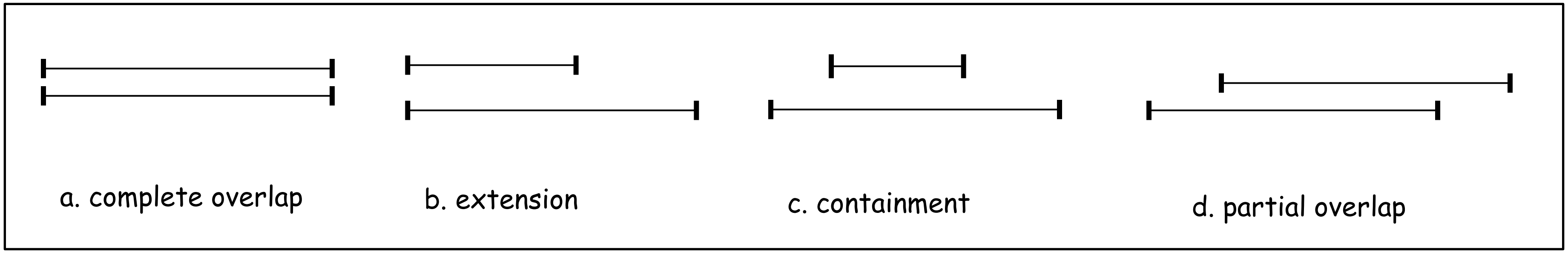
Record	Name	Zip	Phone
r1	John	54321	123-4567
r2	John	54321	987-6543
r3	John	11111	987-6543
r4	Bob	null	121-1212

a. Records to match

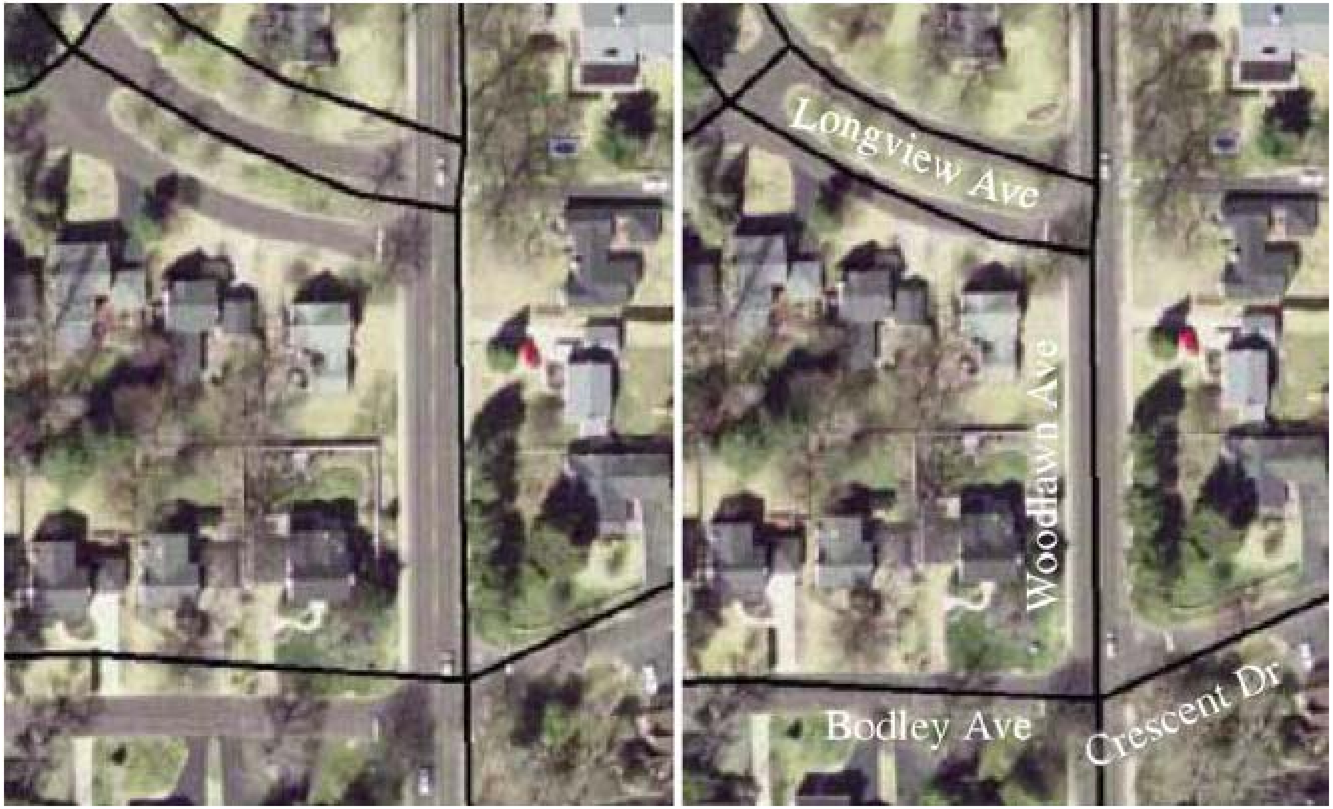
Comparison Rule	Definition
B1	p_{name}
B2	P_{name} AND P_{zip}
B3	P_{name} AND P_{phone}

b. Evolving from rule B1 to rule B2

Possible relationships between polylines



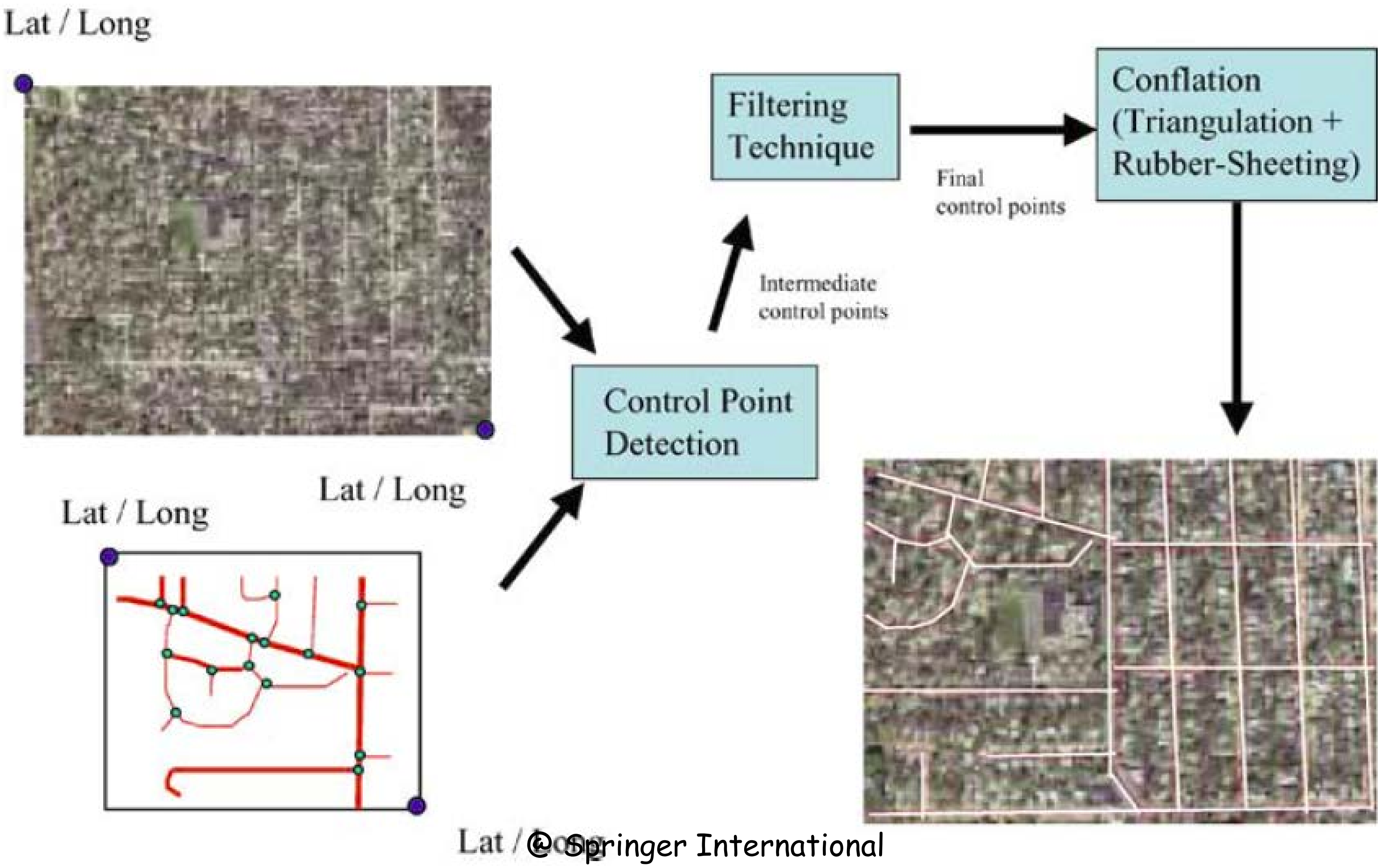
Matching between road vector map and orthoimagery,from [123] @Springer 2006



a. map and image not aligned

b. map and image aligned

The approach presented in [123] @Springer 2006



Control Point Detection

Intermediate control points

Filtering Technique

Final control points

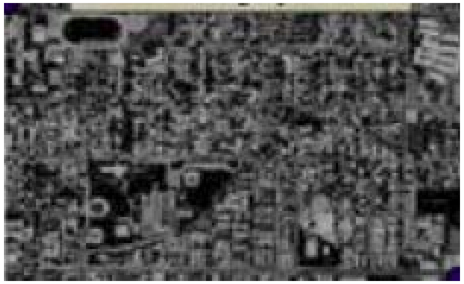
Conflation (Triangulation + Rubber-Sheeting)

The approach and example presented in [124] @Springer 2008

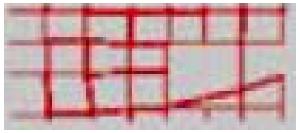
Inputs

Map with unknown coordinates

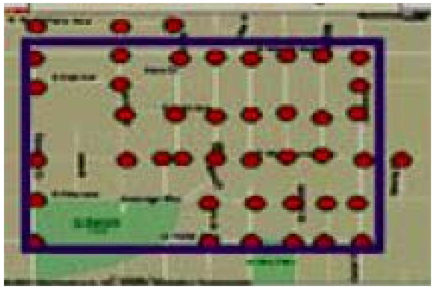
Geo-referenced imaginery



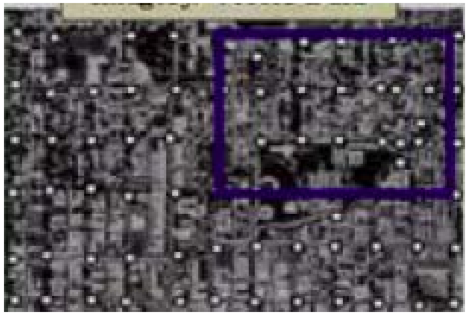
Vector data



1. Detect Intersection Points on the Map

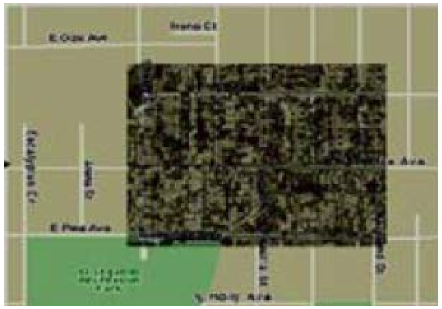


2. Vector-Imaginery Conflation

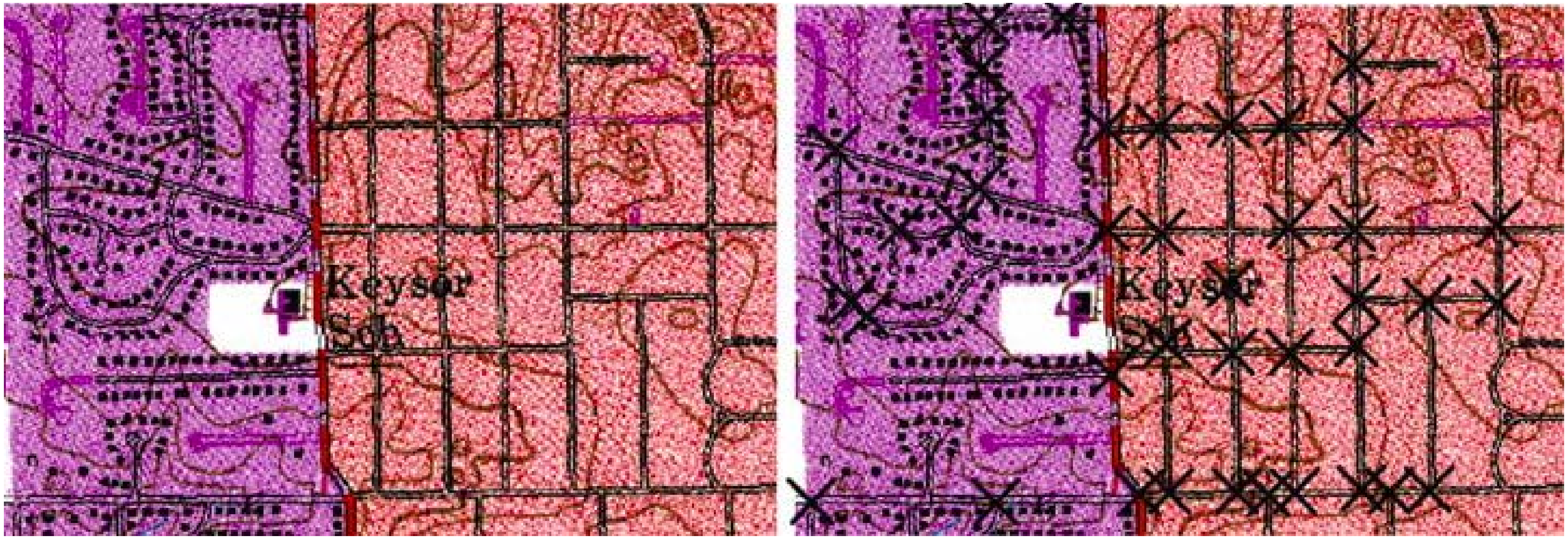


3. Point Pattern Matching & Map-Imaginery Conflation

Output



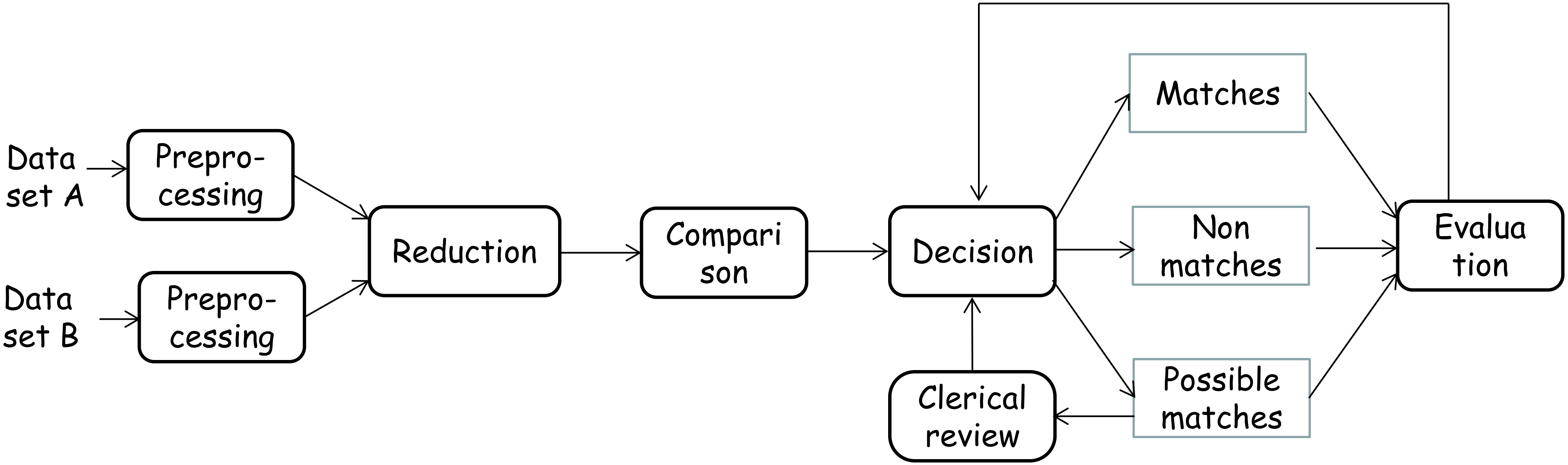
Intersection points automatically detected on a map in [124]



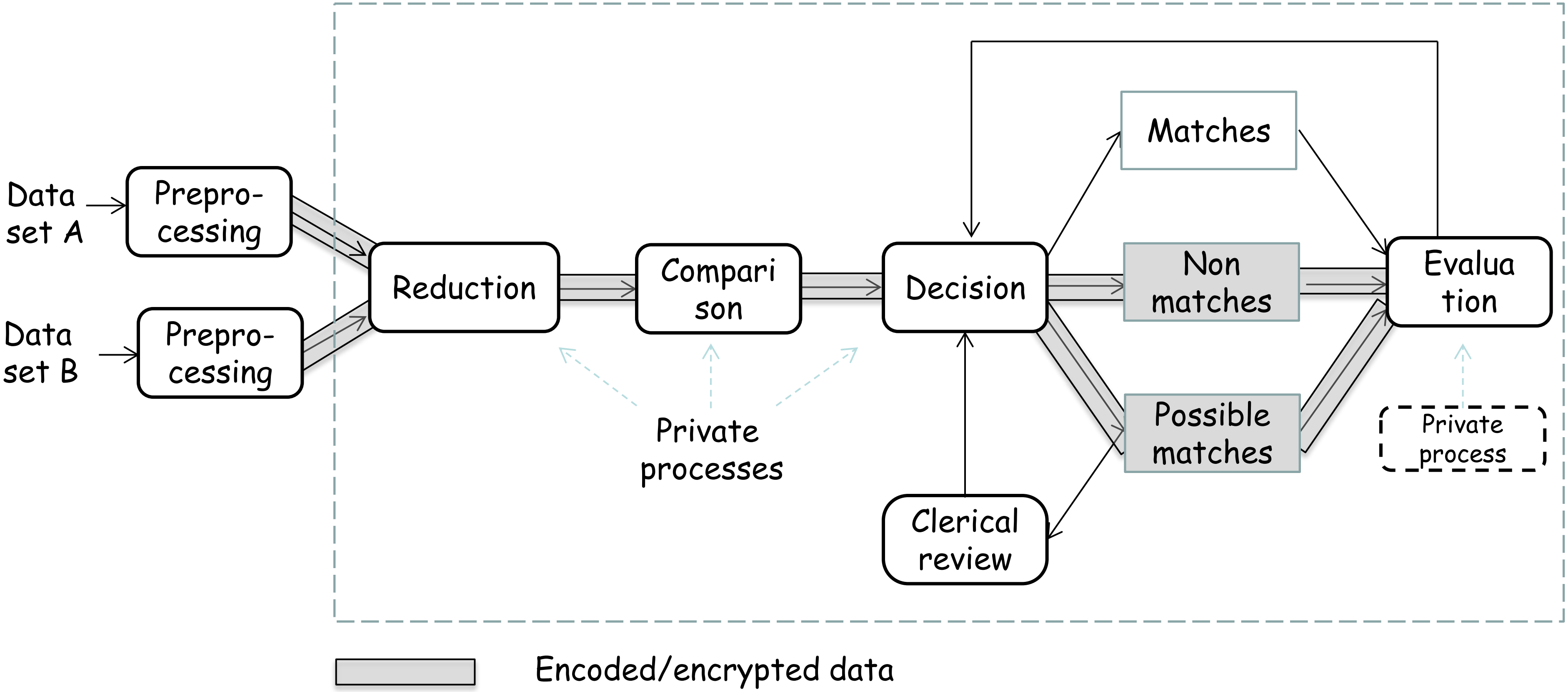
Countries and languages investigated in [518]

Country	Languages
China	Standard Chinese (Mandarin), Cantonese, Shangainese, Fozhou, Hokkinen-Taiwanese, Xiang, Gan, Hakka dialects, and others
France	French, regional dialects
Germany	German
Italy	Italian, German, French, Slovene
Japan	Japanese
Mexico	Spanish, indigenous languages (Mayan, Nauhatl, and others)
Saudi Arabia	Arabic
Spain	Castilian Spanish, Catalan, Galician, Basque
Taiwan	Mandarin Chinese, Taiwanese, Hakka dialects
United Kingdom	English, Scots, Scottish Gaelic, Welsh, Irish, Cornish
Yemen	Arabic

Classical object identification process

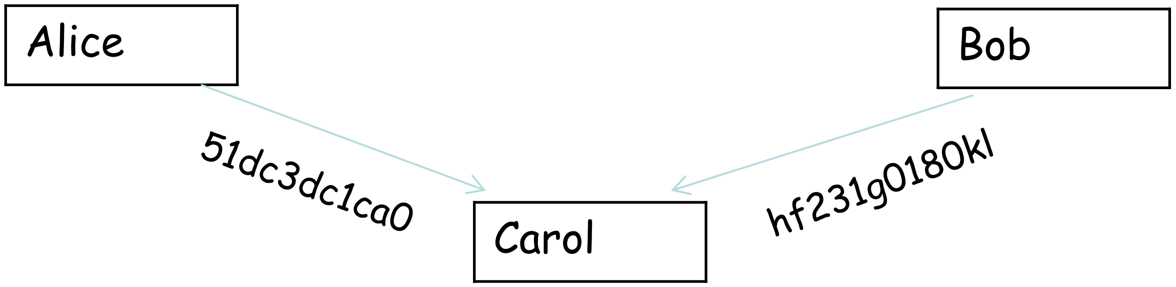


Privacy preserving object identification (inspired to [623])



Secure hash encoding

First Name	Surname	Compound string	Hash string
peter	christen	peterchristen	51dc3dc1ca0
pete	christen	petechristen	h231g0180kl



Alice

Bob

51dc3dc1ca0

hf231g0180kl

Carol

k-anonymized tuples as used in [323]

Alice

Age	Zip Code
25	20133
50	12205
70	12209
30	40100

→

Bob

Age	Zip Code
(20-40)	20***
(40-60)	122**
(60-80)	12***
(20-40)	40***