Data Matching Research at the Australian National University

Peter Christen

Research School of Computer Science,

ANU College of Engineering and Computer Science,

The Australian National University

Contact: **peter.christen@anu.edu.au**

http://cs.anu.edu.au/people/Peter.Christen

THE AUSTRALIAN NATIONAL UNIVERSITY

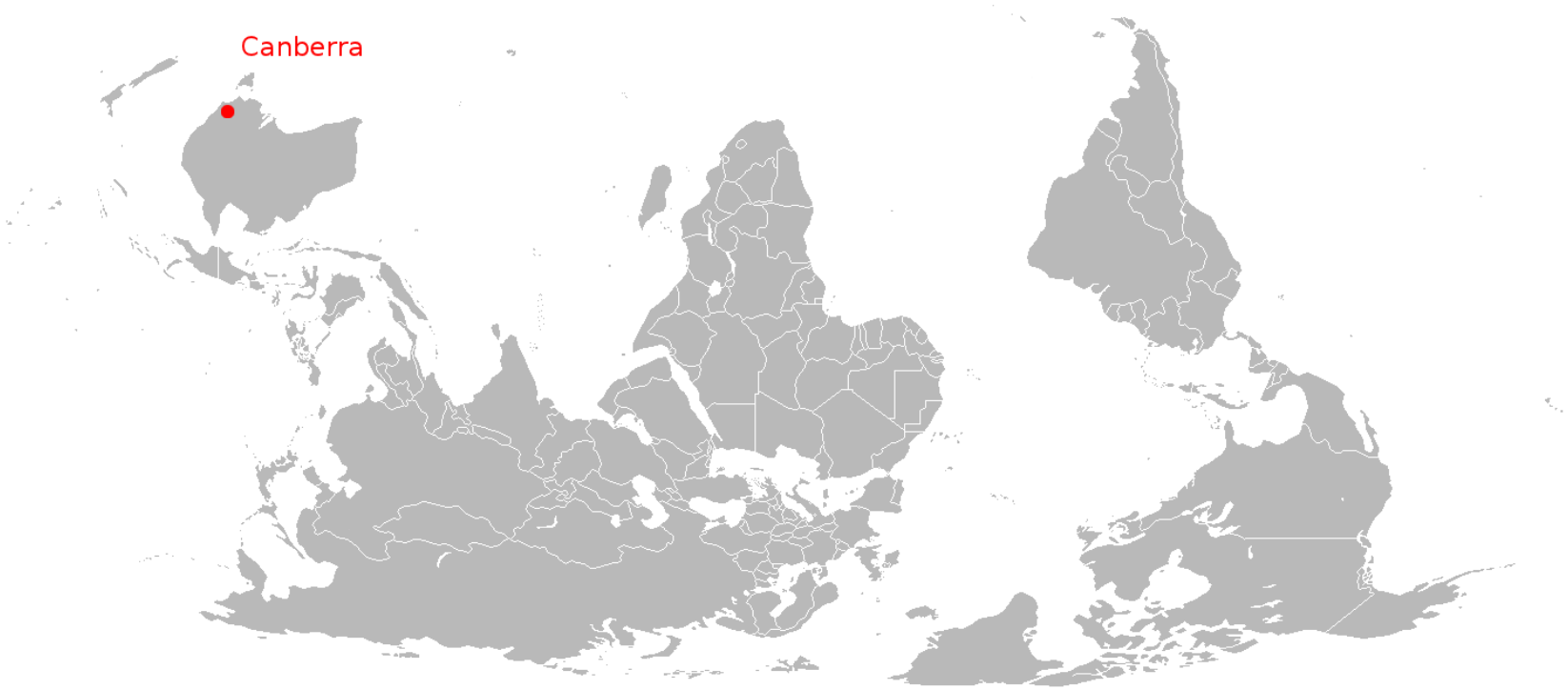
Outline

- Background about me and the ANU
- A short introduction to data matching and its challenges
- Research projects in data matching at the ANU
  - Scalable real-time entity resolution on dynamic databases
  - Scalable privacy-preserving record linkage techniques
  - Efficient matching of historical census data across time
- Conclusions and research directions

Background - Short CV

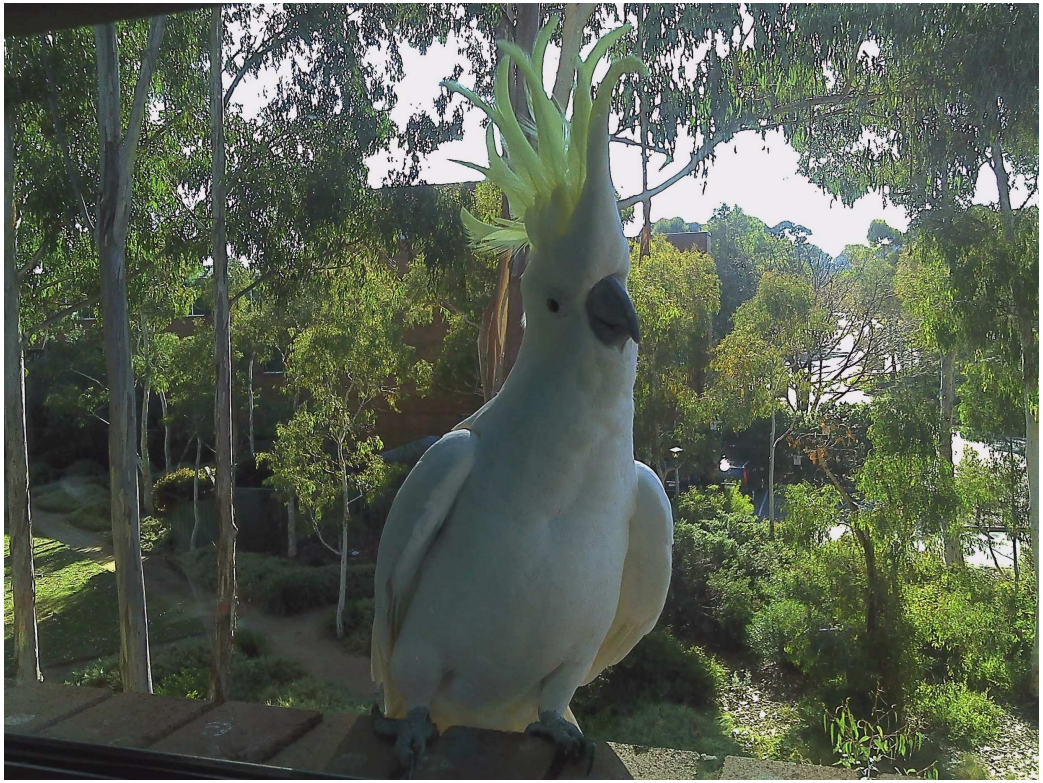
- Born and grew up in Basel, Switzerland
  - Diploma in Computer Science, ETH Zürich in 1995
  - PhD in Parallel Computing, University of Basel in 1999
- Moved to Canberra / ANU in 1999
  - Postdoctoral Researcher (funded by Swiss NSF) from 1999 to 2000
  - Lecturer from 2001 to 2006
  - Senior Lecturer from 2007 to 2012
  - Associate Dean (Higher Degree Research) for Engineering and Computer Science, 2009 to 2011
  - Associate Professor since 2013

Canberra, Australia



THE AUSTRALIAN NATIONAL UNIVERSITY

Research at the ANU (1)



- Around 17,000 students, over 2,000 PhD students (around 100 in computer science)

Research at the ANU (2)

- Over 1,600 academics (around 40 in computer science, including 14 full professors)

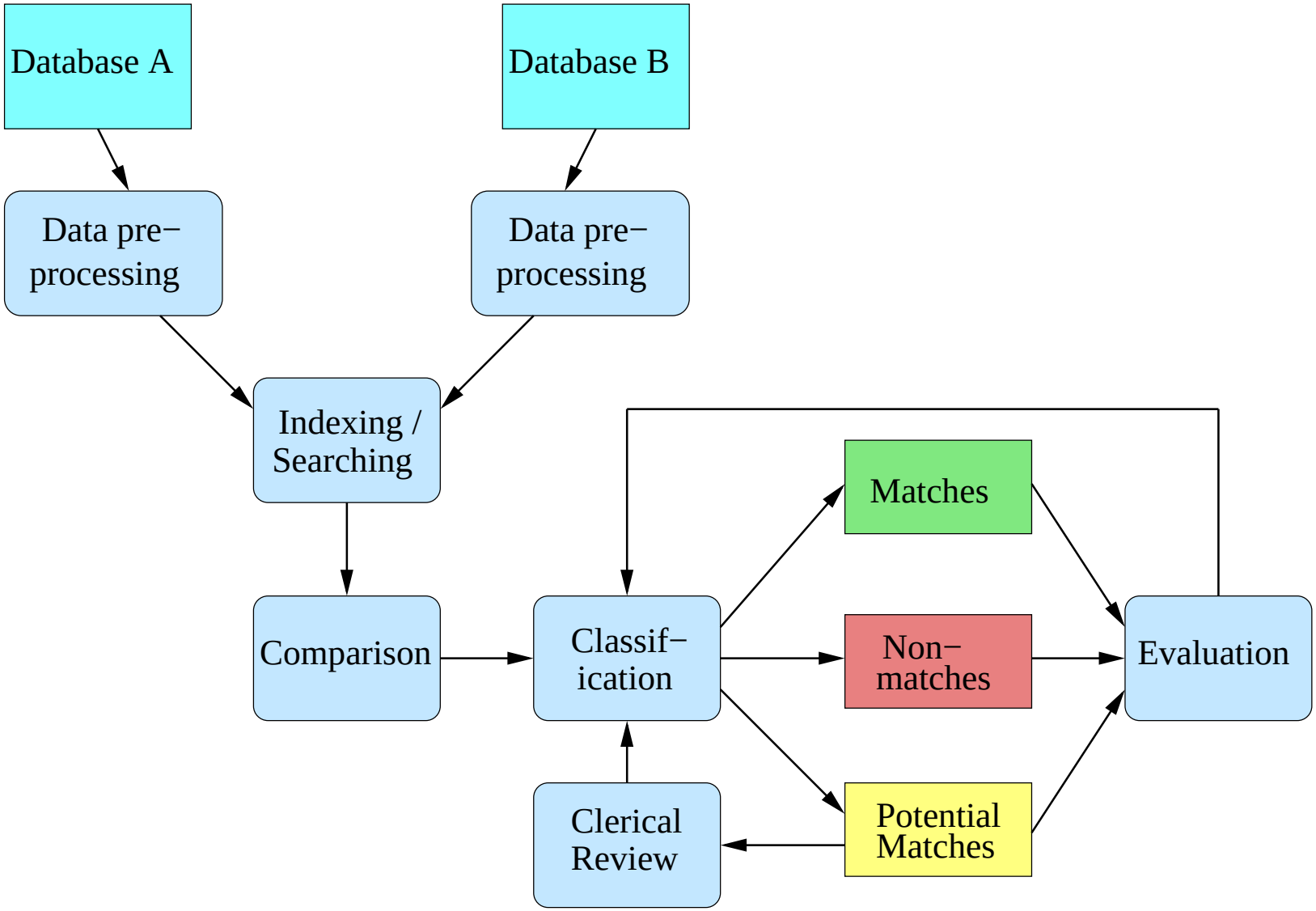
What is data matching?

- The process of matching records that represent the same entity in one or more databases
  (patient, customer, business name, etc.)
- Also known as *record linkage*, *entity resolution*, *object identification*, *duplicate detection*, *identity uncertainty*, *merge-purge*, etc.
- Major challenge is that unique entity identifiers are often not available in the databases to be matched (or if available, they are not consistent)

  E.g., which of these records represent the same person?

Dr Smith, Peter	*42 Miller Street 2602 O'Connor*
Pete Smith	*42 Miller St 2600 Canberra A.C.T.*
P. Smithers	*24 Mill Rd 2600 Canberra ACT*

The data matching process



Applications of data matching

- Remove duplicates in one data set (deduplication)
- Merge new records into a larger master data set
- Create patient or customer oriented statistics
  (for example for longitudinal studies)
- Clean and enrich data for analysis and mining
- Geocode matching (with reference address data)
- Widespread use of data matching
  - Immigration, taxation, social security, census
  - Fraud, crime, and terrorism intelligence
  - Business mailing lists, exchange of customer data
  - Health and social science research

Data matching challenges

- No unique entity identifiers are available
  (use approximate (string) comparison functions)
- Real world data are dirty
  (typographical errors and variations, missing and out-of-date values, different coding schemes, etc.)
- Scalability to very large databases
  (naïve comparison of all record pairs is quadratic; some form of blocking, indexing or filtering is needed)
- No training data in many data matching applications (true match status not known)
- Privacy and confidentiality
  (because personal information is commonly required for matching)

Types of data matching techniques

- Deterministic matching
  - Exact matching (if a *unique identifier* of high quality is available: precise, robust, stable over time)
    Examples: *Social security* or *Medicare* numbers
  - Rule-based matching (complex to build and maintain)
- Probabilistic record linkage (*Fellegi and Sunter*, 69)
  - Use available attributes for matching (often personal information, like names, addresses, dates of birth, etc.)
  - Calculate matching weights for attributes
- 'Computer science' approaches

(based on machine learning, data mining, database, or information retrieval techniques)

Advanced classification techniques

- View record pair classification as a *multi-dimensional binary classification* problem

  (use attribute similarities to classify record pairs as *matches* or *non-matches*)

- Many machine learning techniques can be used
  - Supervised: Decision trees, SVMs, neural networks, learnable string comparisons, active learning, etc.
  - Un-supervised: Various clustering algorithms
- Recently, *collective* classification techniques have been investigated (build graph of database and conduct overall classification, rather than each record pair independently)

Scalable real-time entity resolution on dynamic databases

Scalable real-time entity resolution on dynamic databases

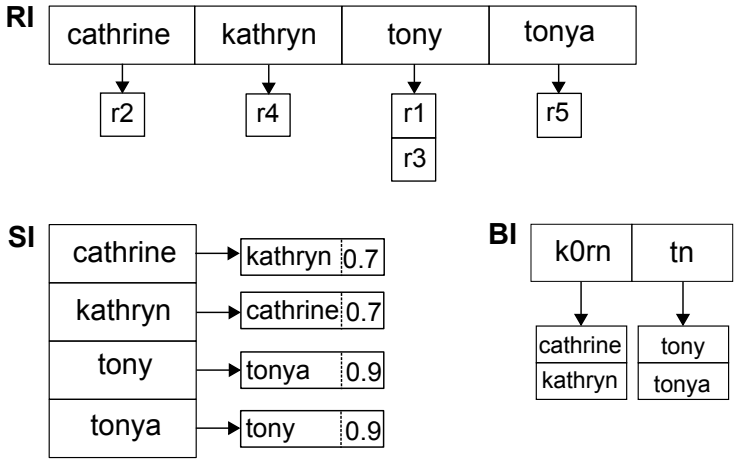
- A Linkage Project funded by the Australian Research Council, Veda (credit bureau), and Funnelback (web and enterprise search)
- Collaborators:
  - Dr Huizhi (Elly) Liang (Post-doc, ANU)
  - Ms Banda Ramadan (PhD student, ANU)
  - Assoc Prof Peter Strazdins (ANU)
  - Dr Ross Gayler (Veda)
  - Prof David Hawking (Funnelback and ANU)

Motivation and objectives

- Credit bureau requires matching in real-time of query records to a large database of entity records (credit enquiries)
- Improve indexing to retrieve candidate records faster, therefore have more time for advanced classification (currently proprietary rules-based)
- Objectives are to develop:
  - Novel indexing techniques that allow for real-time matching of query records on dynamic databases
  - Techniques that consider temporal data aspects
  - Improved techniques for real-time classification of query records (to match with database records)

Dynamic similarity-aware indexing (1)

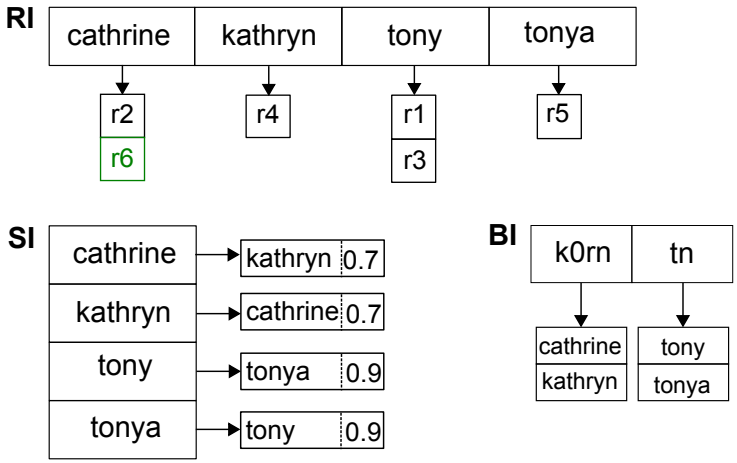
RecID	Given-name	Double-Metaphone
r1	tony	tn
r2	cathrine	k0rn
r3	tony	tn
r4	kathryn	k0rn
r5	tonya	tn



RI: Record index, BI: Block index, SI: Similarity index

Dynamic similarity-aware indexing (2)

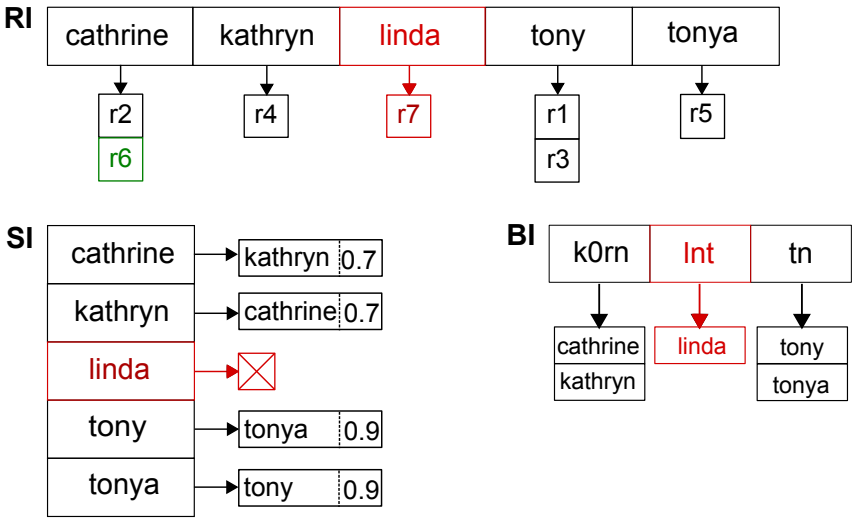
RecID	Given-name	Double-Metaphone
r1	tony	tn
r2	cathrine	k0rn
r3	tony	tn
r4	kathryn	k0rn
r5	tonya	tn
r6	cathrine	k0rn



RI: Record index, BI: Block index, SI: Similarity index

Dynamic similarity-aware indexing (3)

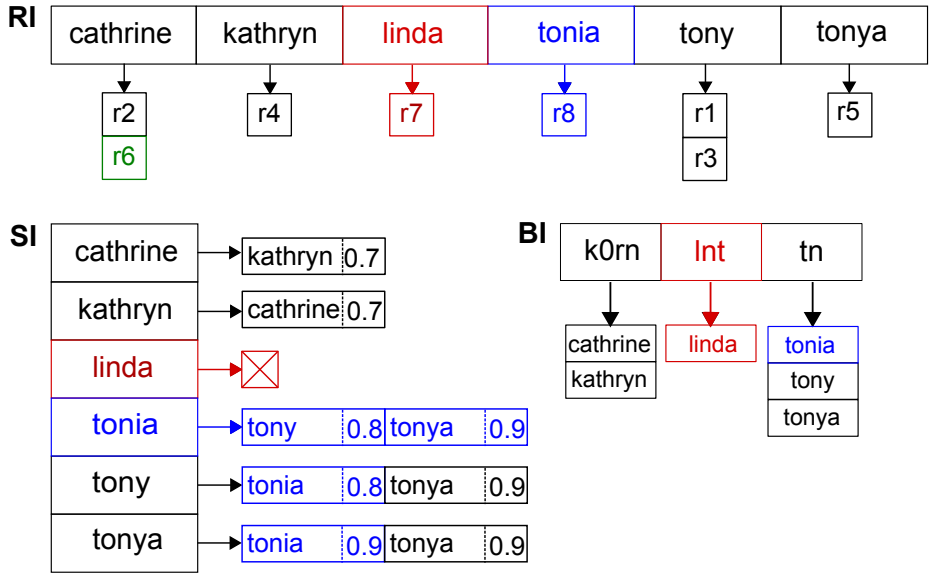
RecID	Given-name	Double-Metaphone
r1	tony	tn
r2	cathrine	k0rn
r3	tony	tn
r4	kathryn	k0rn
r5	tonya	tn
r6	cathrine	k0rn
r7	linda	lnt



RI: Record index, BI: Block index, SI: Similarity index

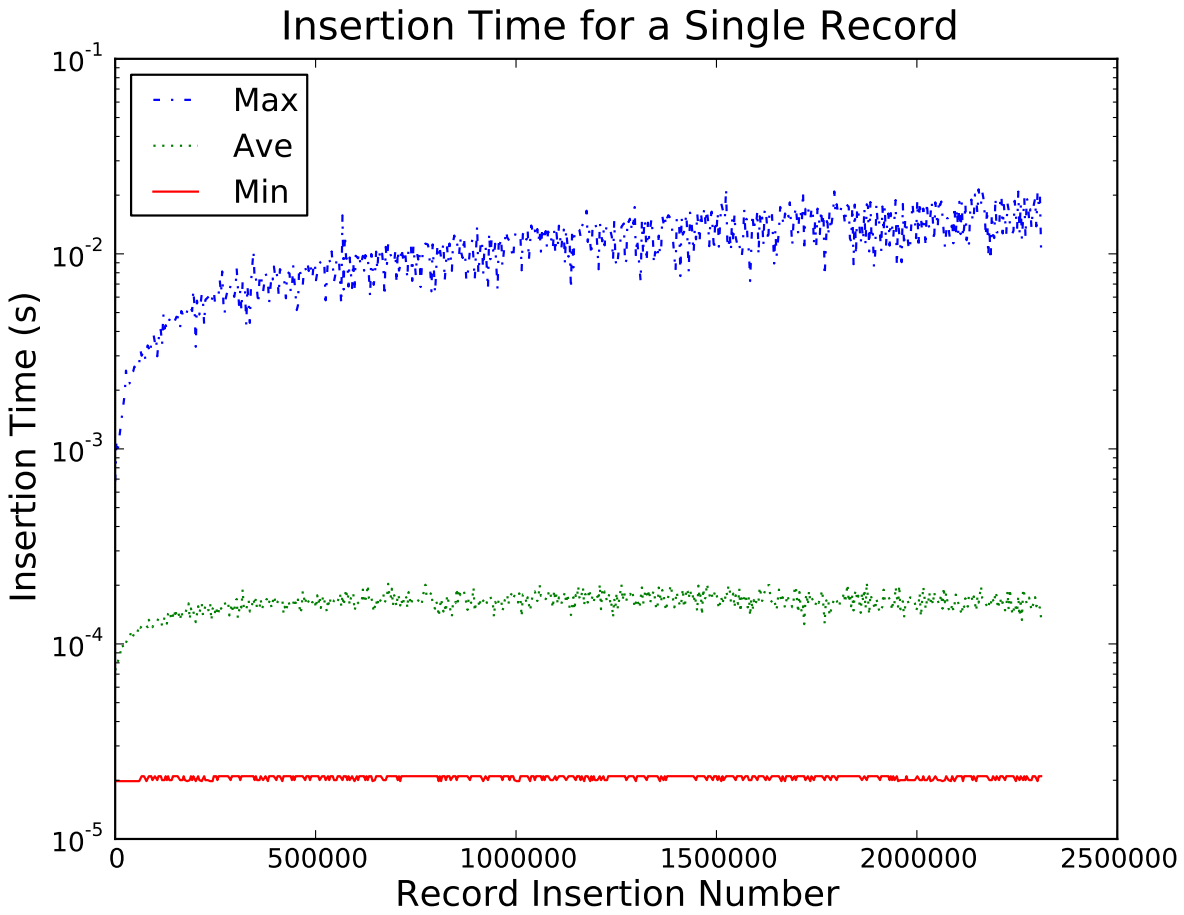
Dynamic similarity-aware indexing (4)

RecID	Given-name	Double-Metaphone
r1	tony	tn
r2	cathrine	k0rn
r3	tony	tn
r4	kathryn	k0rn
r5	tonya	tn
r6	cathrine	k0rn
r7	linda	lnt
r8	tonia	tn



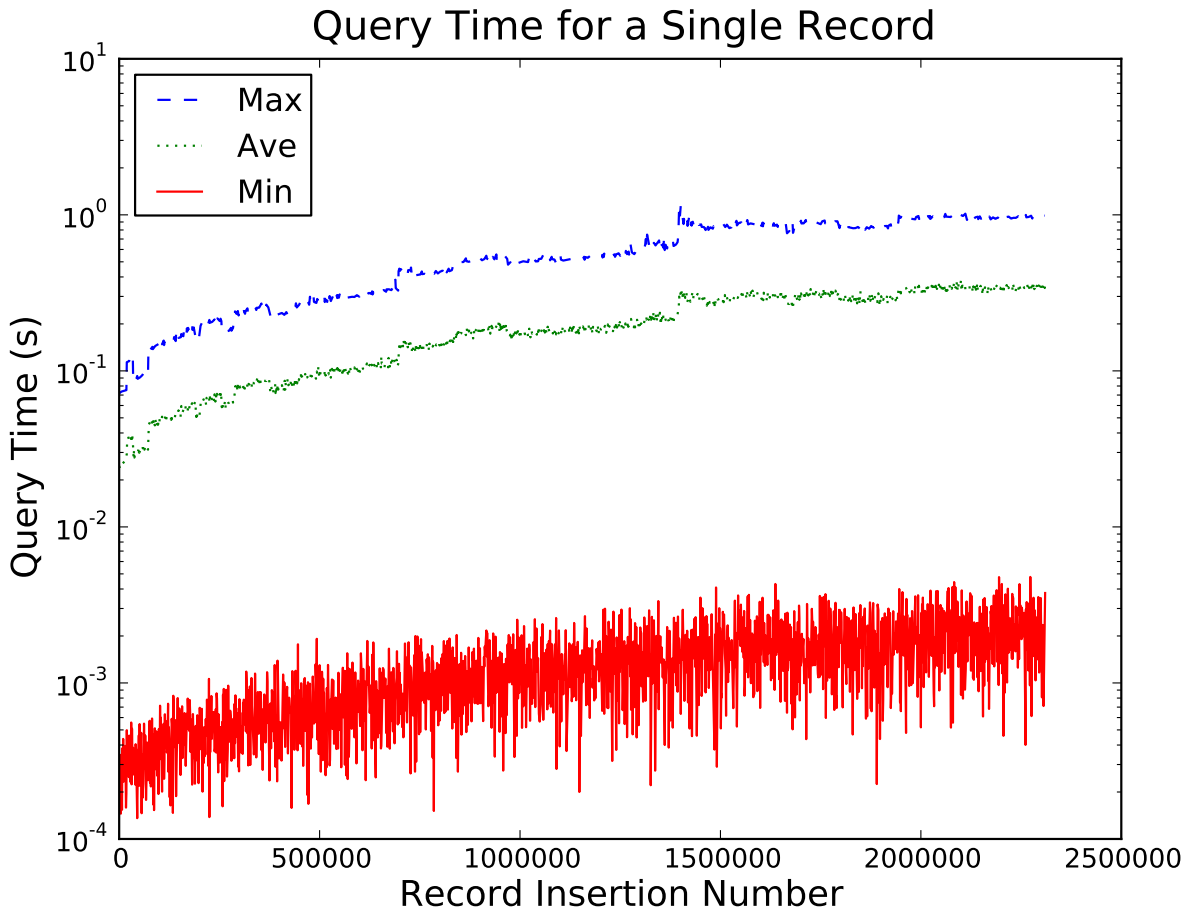
RI: Record index, BI: Block index, SI: Similarity index

Dynamic similarity-aware indexing (5)



On North Carolina voter database (around 2.4 million records)

Dynamic similarity-aware indexing (6)



Scalable privacy-preserving record linkage (PPRL)

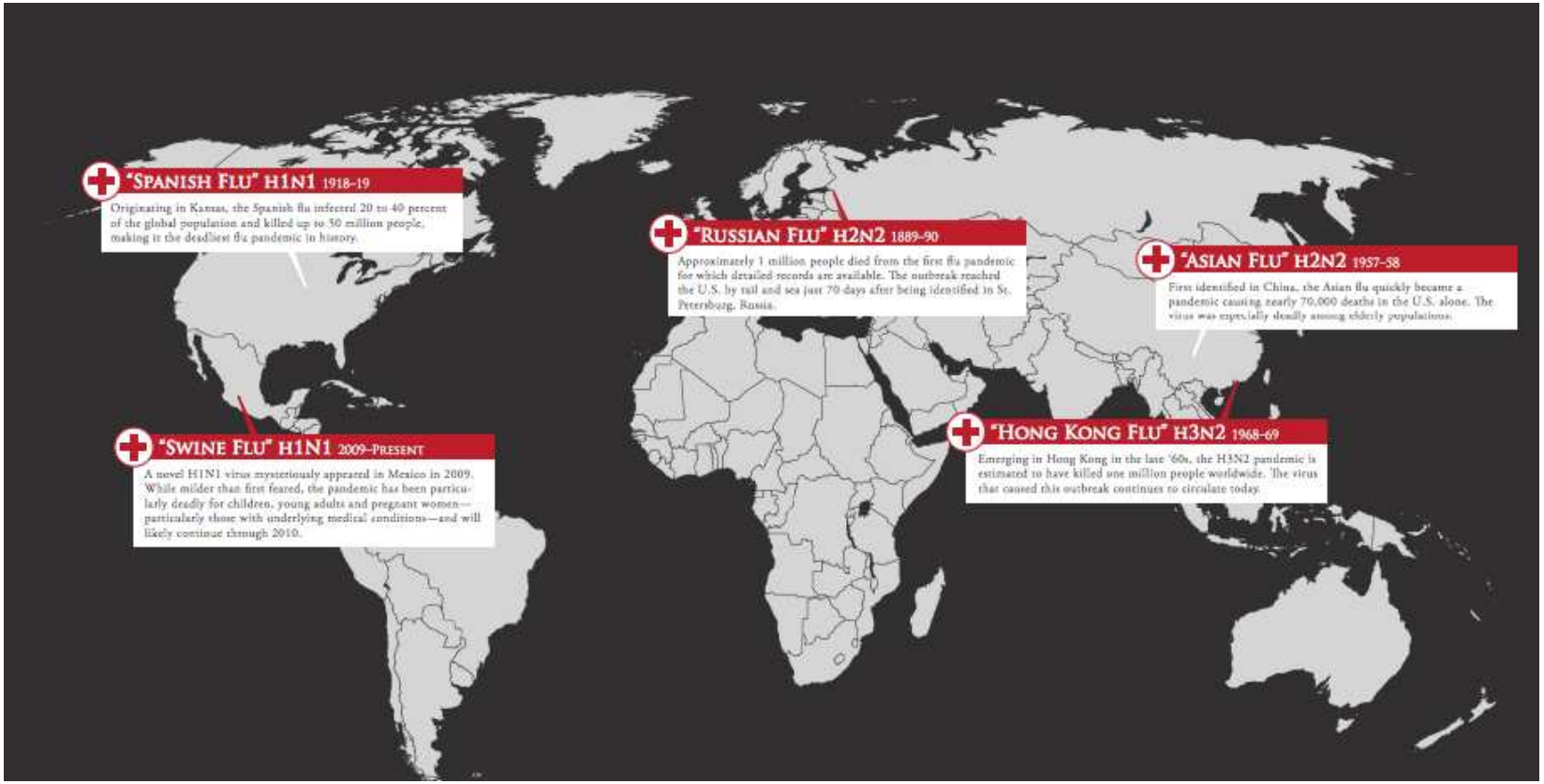
Scalable privacy-preserving record linkage

- A Discovery Project funded by the Australian Research Council
- Collaborators:
  - Ms Dinusha Vatsalan (PhD student, ANU)
  - Assoc Prof Vassilios Verykios (Hellenic Open University)
  - Mr Thilina Ranbaduge (PhD student, starting 2014)

Motivation and objectives

- Privacy concerns in many applications where data are matched between organisations
- Matched data can allow analysis not possible on individual databases
  (potentially revealing highly sensitive information)
- Objectives are to develop:
  - Scalable techniques to facilitate PPRL
  - Techniques that allow PPRL on multiple databases
  - Improved classification techniques for PPRL
  - Methods to assess matching quality and completeness in a privacy-preserving framework

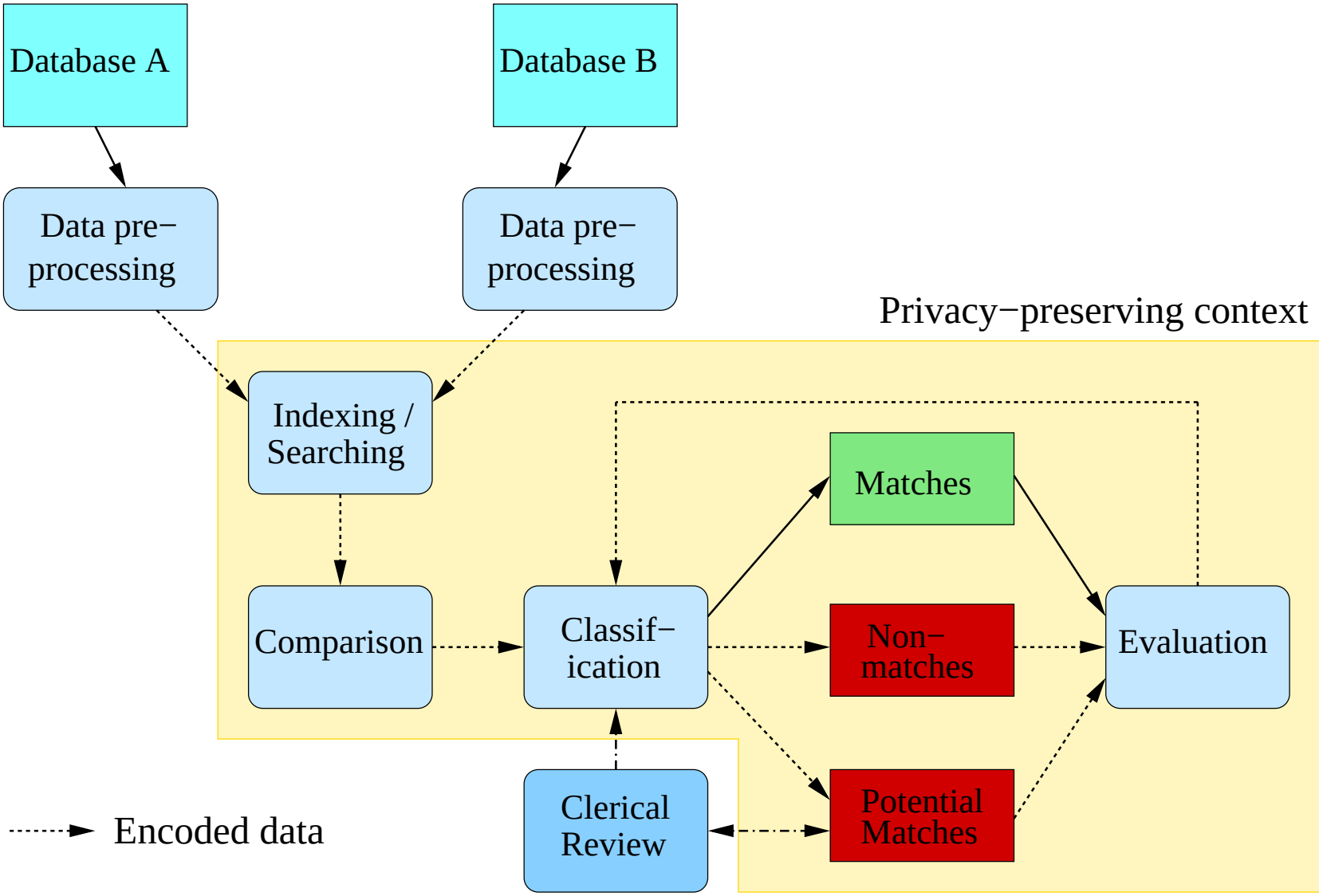
Privacy and data matching: An example scenario (1)



Privacy and data matching: An example scenario (2)

- Preventing the outbreak of epidemics requires monitoring of occurrences of unusual patterns in symptoms (in real time!)
- Data from many different sources will need to be collected (including travel and immigration records; doctors, emergency and hospital admissions; drug purchases in pharmacies; animal health data; etc.)
- Privacy concerns arise if such data are stored and matched at a central location
- Matched sensitive patient data and confidential data from healthcare organisations must be kept secure, while still allowing analysis

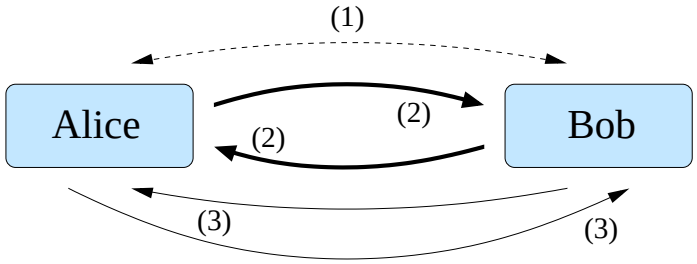
The PPRL process

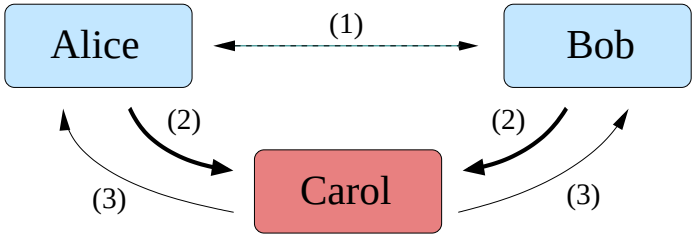


PPRL challenges and basic protocols

- Main challenges in PPRL
  - Allow for approximate matching of values (given real world data are often ‘dirty’)
  - Have techniques that are not vulnerable to any kind of attack, and are scalable to matching large databases
- Two basic types of protocols
  - Two-party protocol: Only the two database owners who wish to link their data
  - Three-party protocols: Use a (trusted) third party (linkage unit) to conduct the linkage (this party will never see any unencoded values, but collusion is possible)

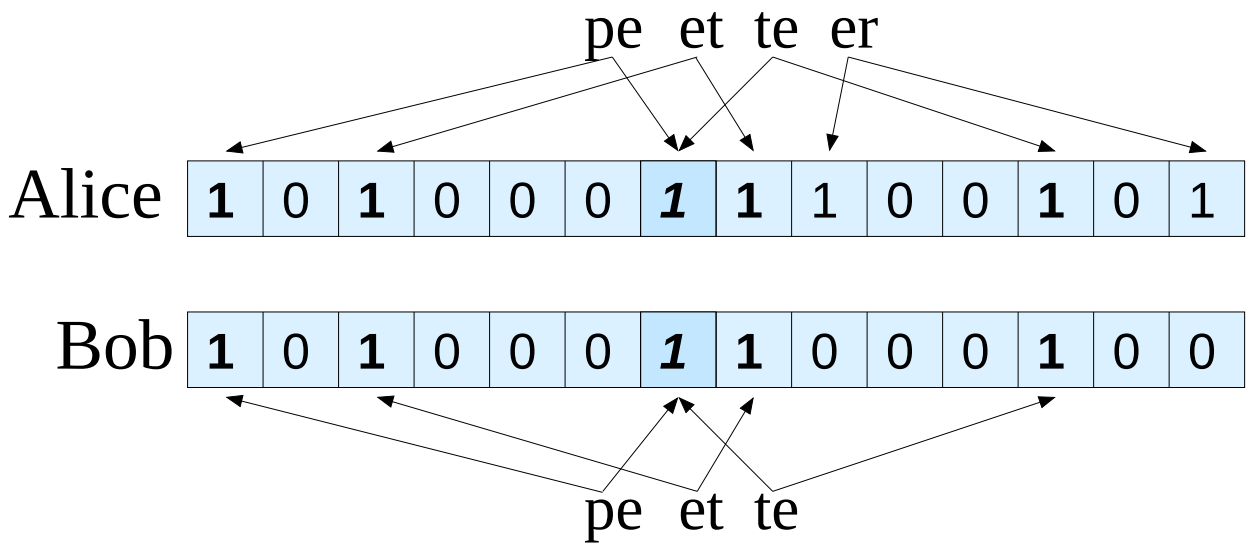
Basic protocol steps





- Generally, three main communication steps
  1. Exchange of which attributes to use in a linkage, pre-processing methods, encoding functions, parameters, secret keys, etc.
  2. Exchange of the *somehow* encoded database records
  3. Exchange of records (or selected attribute values, or identifiers only) of records classified as matches

Bloom filter based PPRL

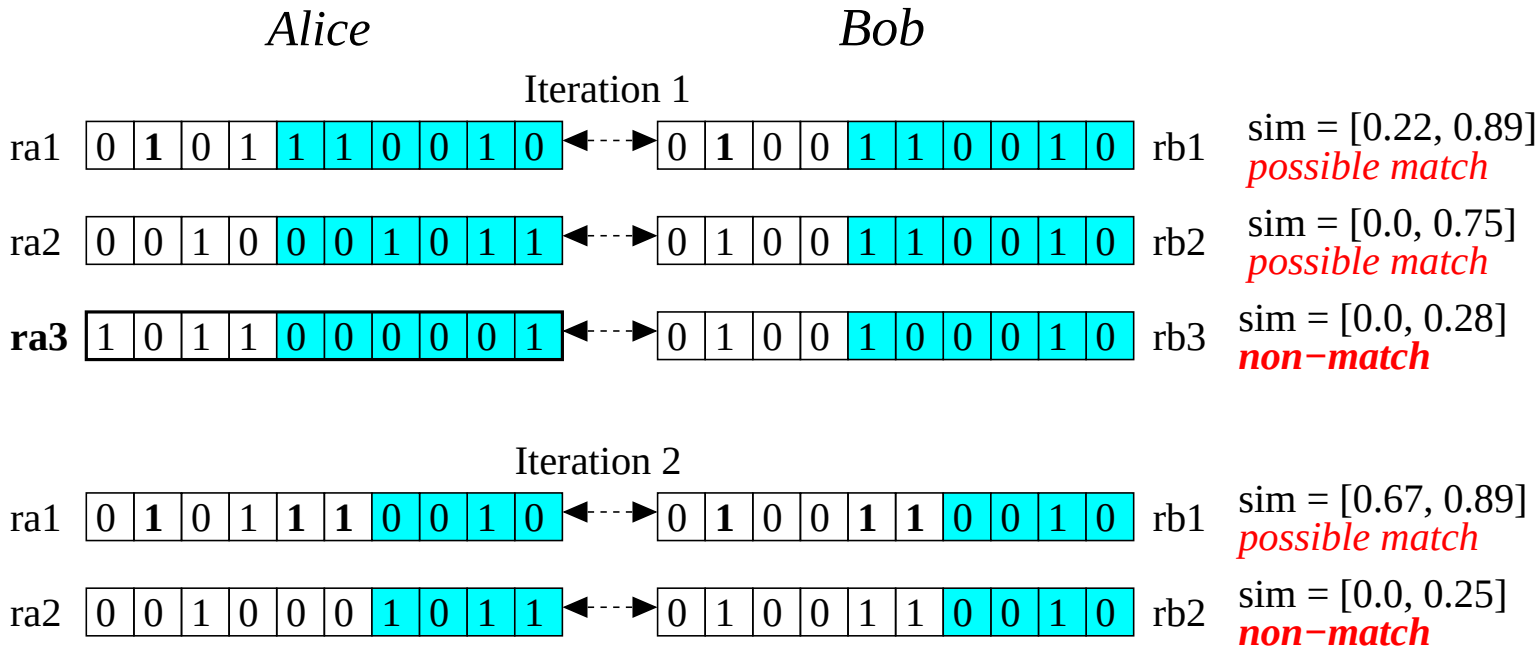


- Proposed by Schnell et al. (Biomed Central, 2009)
- Idea: Map q-grams into Bloom filters using hash functions only known to database owners, send Bloom filters to linkage unit to calculate Dice similarity
- *1*-bits for string ‘peter’: 7, *1*-bits for ‘pete’: 5, common *1*-bits: 5, therefore $sim_{Dice} = 2 \times 5/(7+5) = 10/12 = 0.83$

Two-party Bloom filter protocol (1)

- Iteratively exchange certain bits from the Bloom filters between database owners
- Calculate the minimum Dice similarity from the bits exchanged, and classify record pairs as matches, non-matches, and possible matches
- Pairs classified as possible matches are taken to the next iteration (where more bits are exchanged)
  - The number of bits revealed in each iteration is calculated such that the risk of revealing more bits for non-matches is minimised
  - Minimum similarity of possible matches increases as more bits are revealed

Two-party Bloom filter protocol for PPRL (2)



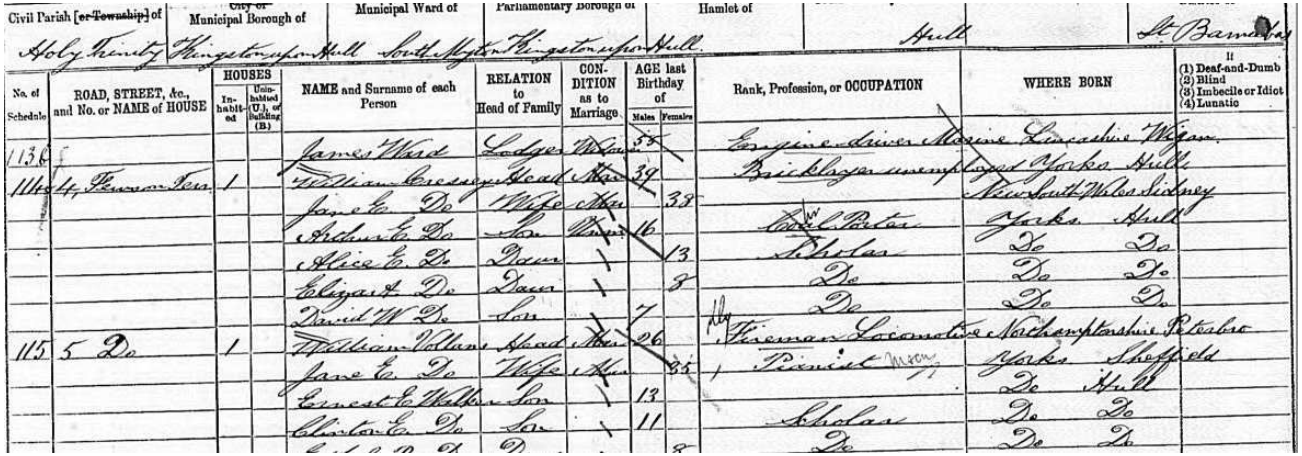
- Each party knows how many 1-bits are set in total in a Bloom filter received from the other party
- In iteration 1, for example, there is one unrevealed 1-bit in *ra3*, and so the maximum possible Dice similarity with *rb3* is: $max(sim(ra3, rb3)) = 2 \times 1/(4+3) = 2/7 = 0.28$

Efficient matching of historical census data across time

Project 3: Efficient matching of historical census data across time

- Collaborators:
  - Ms Zhichun (Sally) Fu (PhD student, ANU)
  - Assoc Prof Mac Boot (Australian Demographic and Social Research Institute, ANU)
- Motivation
  - Shift in the social sciences from small-scale studies to using population databases
  - New field of ‘population informatics’ to analyse the ‘social genome’
  - Develop techniques to compile family trees over time from large data collections (population reconstruction)

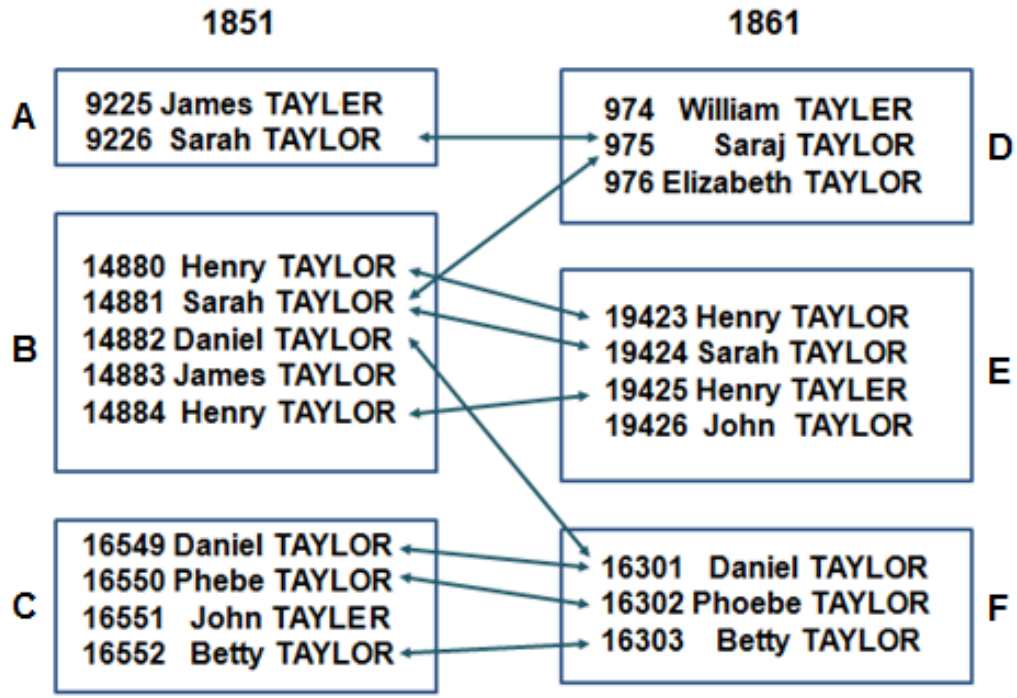
Challenges with historical (census) data



No. of Schedule	ROAD, STREET, &c., and No. or NAME of HOUSE	HOUSES Inhabited	HOUSES Uninhabited (U), or Building (B)	NAME and Surname of each Person	RELATION to Head of Family	CONDITION as to Marriage	AGE last Birthday of Males	AGE last Birthday of Females	Rank, Profession, or OCCUPATION	WHERE BORN	If (1) Deaf-and-Dumb (2) Blind (3) Imbecile or Idiot (4) Lunatic
136				James Ward	Lodger	Wid	54		Engine driver Marine	Lincolnshire Wigan	
114	4 Ferguson Ter	1		William Cressey	Head	Mar	39		Bricklayer unemployed	Yorks Hull	
				Jane E. Do	Wife	Mar		38		New South Wales Sidney	
				Arthur E. Do	Son		16		Cabt Porter	Yorks Hull	
				Alice E. Do	Dau		13		Scholar	Do Do	
				Elizabeth Do	Dau		8		Do	Do Do	
				David W. Do	Son		7		Do	Do Do	
115	5 Do	1		William Williams	Head	Mar	36		Fireman Locomotive	Northamptonshire Peterboro	
				Jane E. Do	Wife	Mar		35	Pianist	Yorks Sheffield	
				Ernest E. Williams	Son		13			Do Hull	
				Clinton E. Do	Son		11		Scholar	Do Do	
									Do	Do Do	

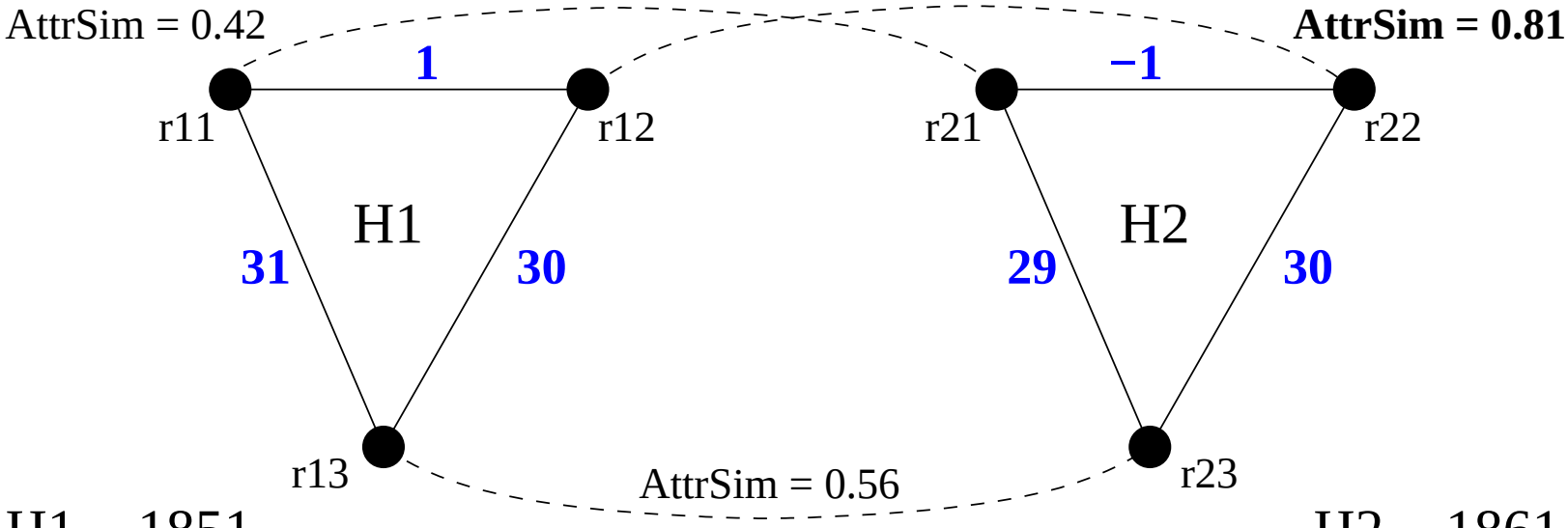
- Low literacy (recording errors and unknown exact values), no address or occupation standards
- Large percentage of a population had one of just a few common names (‘John’ or ‘Mary’)
- Households and families change over time
- Immigration and emigration, birth and death
- Scanning, OCR, and transcription errors

Group matching using household information



- Conduct pair-wise matching of individual records
- Calculate household similarities using Jaccard or weighted similarities (based on pair-wise matching)
- Promising results on UK Census data from 1851 to 1901 (Rawtenstall, with around 17,000 to 31,000 records)

Graph-matching based on household structure



H1 − 1851

ID	Address	SN	FN	Age
r11	goodshaw	smith	john	32
r12	goodshaw	smith	mary	31
r13	goodshaw	smith	anton	1

H2 − 1861

ID	Address	SN	FN	Age
r21	goodshaw	smith	jack	39
r22	goodshaw	smith	marie	40
r23	goodshaw	smith	toni	10

- One graph per household, find best matching graphs using both record attribute and structural similarities
- Edge attributes are information that does not change over time (like age differences)

To make sure everybody is awake..

Conclusions and research directions

- We address various challenges in data matching: real-time matching and dynamic data; temporal aspects; privacy; and population reconstruction
- Challenges in data matching
  - Improved classification for matching personal data
  - Matching data from many sources
  - Use of cloud computing platforms for large-scale data matching
  - Frameworks for data matching that allow comparative experimental studies, and test data collections
  - Develop practical PPRL techniques (assessing accuracy and completeness)

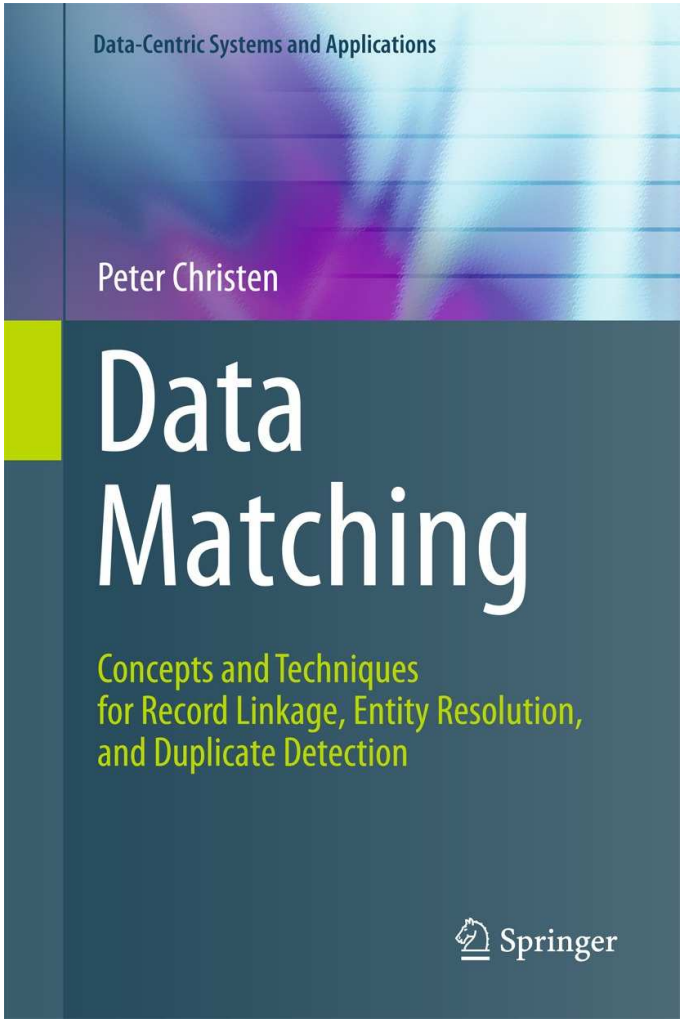
Recent publications (1)

- Christen P and Gayler R: *Adaptive temporal entity resolution on dynamic databases.* PAKDD, Gold Coast, Australia, Springer LNCS vol. 7819, 2013.
- Christen P and Vatsalan D: *Flexible and extensible generation and corruption of personal data.* ACM CIKM, San Francisco, 2013.
- Christen P: *Advanced record linkage methods and privacy aspects for population reconstruction.* Workshop on Population Reconstruction, Amsterdam, 2014.
- Fisher J, Wang Q, Wong P and Christen P: *Data cleaning and matching of institutions in bibliographic databases.* AusDM, Canberra, CRPIT vol. 146, 2013.
- Fu Z, Zhou J, Christen P and Boot M: *Multiple instance learning for group record linkage.* PAKDD, Kuala Lumpur, Malaysia, Springer LNCS vol. 7301, 2012.
- Fu Z, Boot M, Christen P and Zhou J: *Automatic record linkage of individuals and households in historical census data.* International Journal of Humanities and Arts Computing, 2014.
- Fu Z, Christen P and Zhou J: *A graph matching method for historical census household linkage.* PAKDD, Tainan, Taiwan, 2014.

Recent publications (2)

- Li S, Liang H and Ramadan B: *Two stage similarity-aware indexing for large-scale real-time entity resolution.* AusDM, Canberra, CRPIT vol. 146, 2013.
- Liang H, Wang Y, Christen P and Gayler R: *Noise-tolerant approximate blocking for dynamic real-time entity resolution.* PAKDD, Tainan, Taiwan, 2014.
- Ramadan B, Christen P, Liang H, Gayler R, and Hawking D: *Dynamic similarity-aware inverted indexing for real-time entity resolution.* PAKDD Workshops (DMApps), Gold Coast, Australia, Springer LNCS vol. 7867, 2013.
- Tran KN, Vatsalan D and Christen P: *GeCo: an online personal data generator and corruptor.* ACM CIKM, San Francisco, 2013.
- Vatsalan D and Christen P: *Sorted nearest neighborhood clustering for efficient private blocking.* PAKDD, Gold Coast, Australia, Springer LNCS vol. 7819, 2013.
- Vatsalan D, Christen P and Verykios VS: *A taxonomy of privacy-preserving record linkage techniques.* Journal of Information Systems, 2013.
- Vatsalan D, Christen P and Verykios VS: *Efficient two-party private-blocking based on sorted nearest neighborhood clustering.* ACM CIKM, San Francisco, 2013.

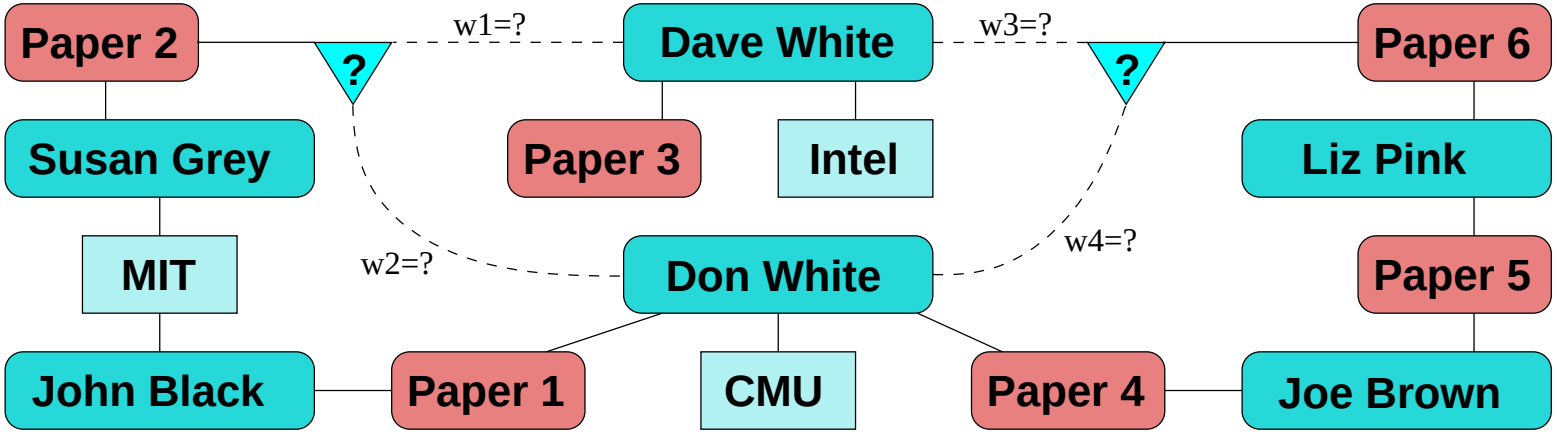
Advertisement: Book 'Data Matching'



The book is very well organized and exceptionally well written. Because of the depth, amount, and quality of the material that is covered, I would expect this book to be one of the standard references in future years.

William E. Winkler, U.S. Bureau of the Census.

Collective classification example



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

(P1, John Black / Don White)
(P2, Sue Grey / **D. White**)
(P3, Dave White)
(P4, Don White / Joe Brown)
(P5, Joe Brown / Liz Pink)
(P6, Liz Pink / **D. White**)

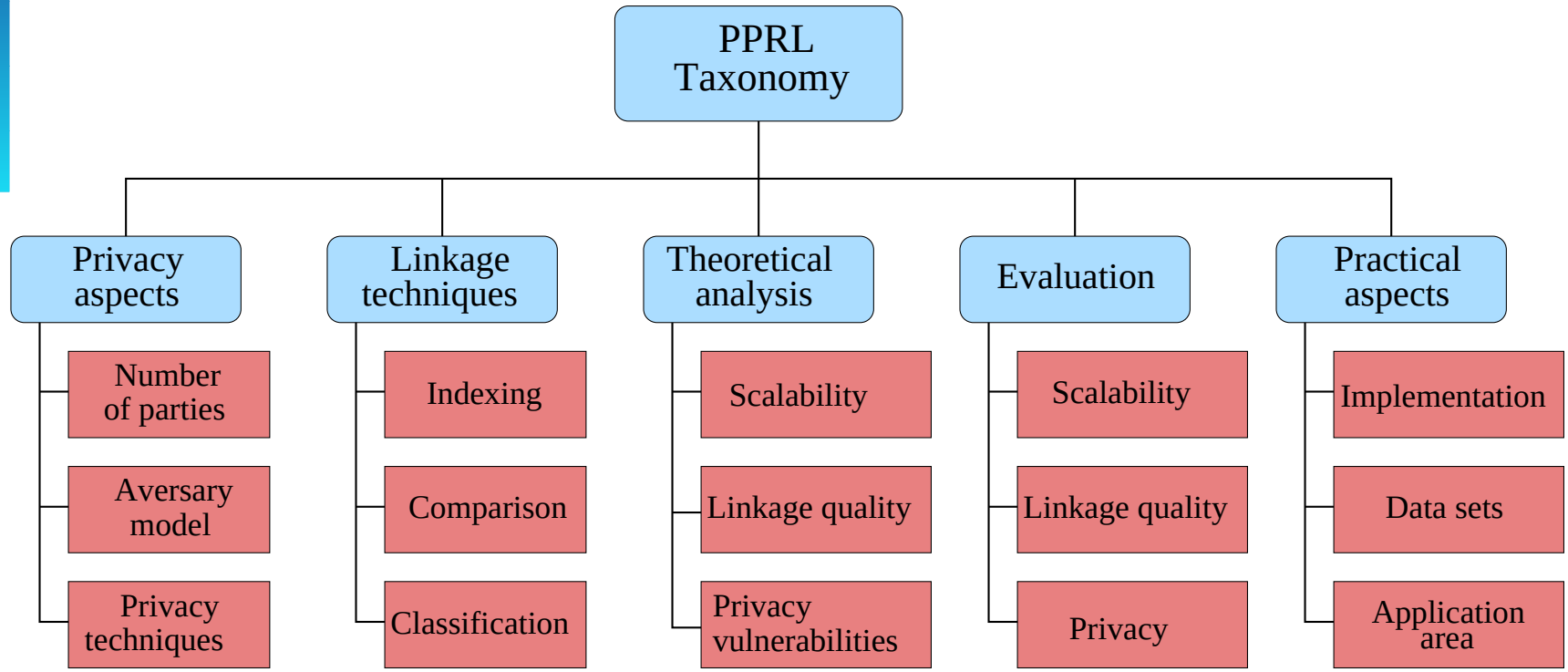
Adapted from Kalashnikov and Mehrotra, ACM TODS, 31(2), 2006

A definition of PPRL

- Assume $O_1 \cdots O_d$ are the d owners of their respective databases $D_1 \cdots D_d$
- They wish to determine which of their records $r_1^i \in D_1$, $r_2^j \in D_2$, $\cdots$, and $r_d^k \in D_d$, match according to a decision model $C(r_1^i, r_2^j, \cdots, r_d^k)$ that classifies pairs (or groups) of records into one of the two classes M of matches, and U of non-matches
- $O_1 \cdots O_d$ do not wish to reveal their actual records $r_1^i \cdots r_d^k$ with any other party

  (they are, however, prepared to disclose to each other, or to an external party, the outcomes of the matching process — certain attribute values of record pairs in class M — to allow further analysis)

A taxonomy for PPRL



Secure multi-party computation

- Compute a function across several parties, such that no party learns the information from the other parties, but all receive the final results
  [Yao 1982; Goldreich 1998/2002]
- Simple example: Secure summation $s = \sum_i x_i$.