Advanced record linkage methods and privacy aspects for population reconstruction

Peter Christen

Research School of Computer Science,

ANU College of Engineering and Computer Science,

The Australian National University

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

Work done in collaboration with Zhichun (Sally) Fu (ANU), Dinusha Vatsalan (ANU), Mac Boot (ANU), and Vassilios S. Verykios (Hellenic Open University)

Outline

- A short introduction to record linkage
- Challenges of population reconstruction
- Advanced classification for record linkage
  - Collective classification techniques
  - Group and graph linkage techniques
- Privacy aspects in record linkage
  - Motivating scenario
  - Basic challenges and protocols
  - Privacy issues for population reconstruction
- Conclusions and research directions

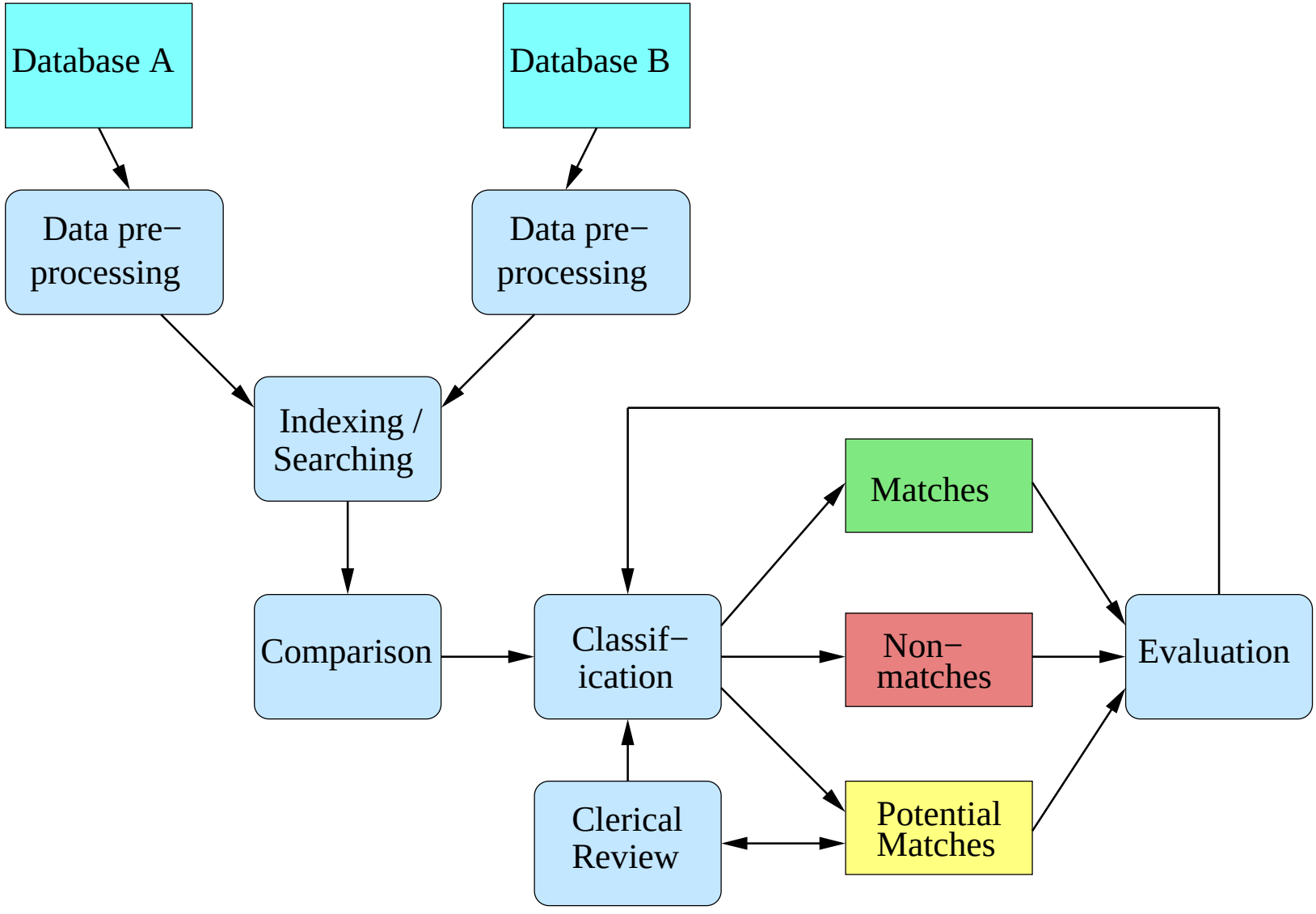
What is record linkage

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

  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 record linkage process



Applications of record linkage

- 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 record linkage
  - Immigration, taxation, social security, census
  - Fraud, crime, and terrorism intelligence
  - Business mailing lists, exchange of customer data
  - Health and **social science research**

Record linkage 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 record linkage applications (true match status not known)
- Privacy and confidentiality
  (because personal information is commonly required for linking)

Types of record linkage techniques

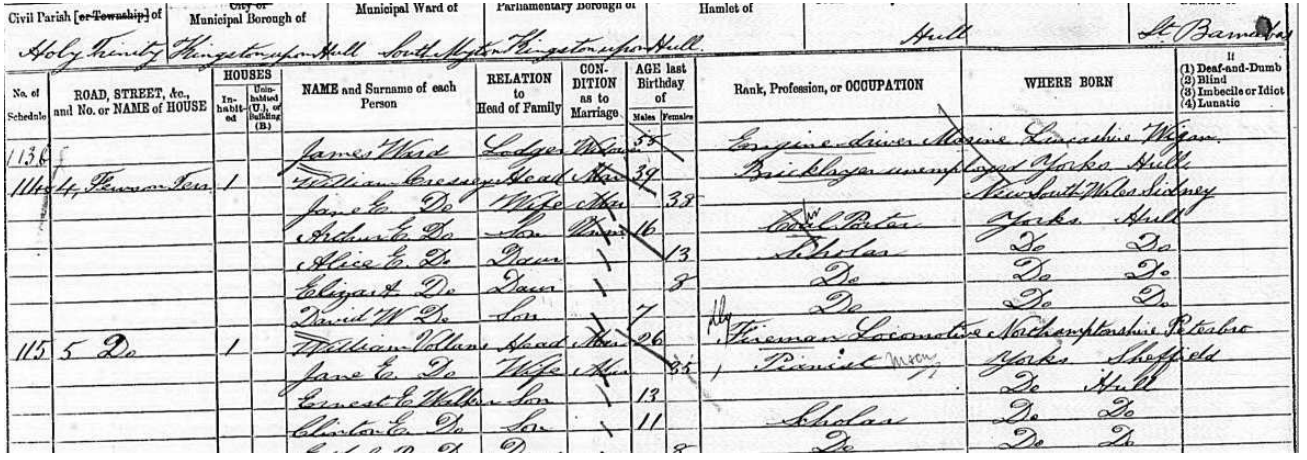
- Deterministic matching
  - Exact matching (if a *unique identifier* of high quality is available: precise, robust, stable over time)
    Examples: *Social security* or *Medicare* numbers
  - Rule-based matching (complex to build and maintain)
- Probabilistic record linkage (*Fellegi and Sunter*, 69)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

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

Challenges for population reconstruction

- Aim is to create "**social genomes**" for individuals by linking large population databases (population informatics, Kum et al. IEEE Computer, 2013)
- Knowing how individuals, families, and house-holds change over time allows for a diverse range of studies (fertility, employment, education, health, etc.)
- Different challenges for historical data compared to contemporary data, but some are common
  - Database sizes (computational aspects)
  - Accurate match classification
  - Coverage of population databases

Challenges for historical (census) data



Civil Parish [or Township] of	City or Municipal Borough of	Municipal Ward of	Parliamentary Borough of	Hamlet of		
Holy Trinity	Kingston upon Hull	South Myton	Kingston upon Hull		Hull	St Barnabas

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
113				James Ward	Lodger	Mar	55		Engine Driver Marine	Lincolnshire Wigan	
114	4 Tennyson Terr	1		William Cressey	Head	Mar	39		Bricklayer unemployed	Yorks Hull	
				Jane E. Do	Wife	Mar		38		New South Wales Sidney	
				Arthur E. Do	Son	Unm	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 Clune	Head	Mar	36		Fireman Locomotive	Northamptonshire Peterboro	
				Jane E. Do	Wife	Mar		35	Pianist	Yorks Sheffield	
				Ernest E. Clune	Son		13			Do Hull	
				Clinton E. Do	Son		11		Scholar	Do Do	
					Do			8	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

Challenges for present-day data

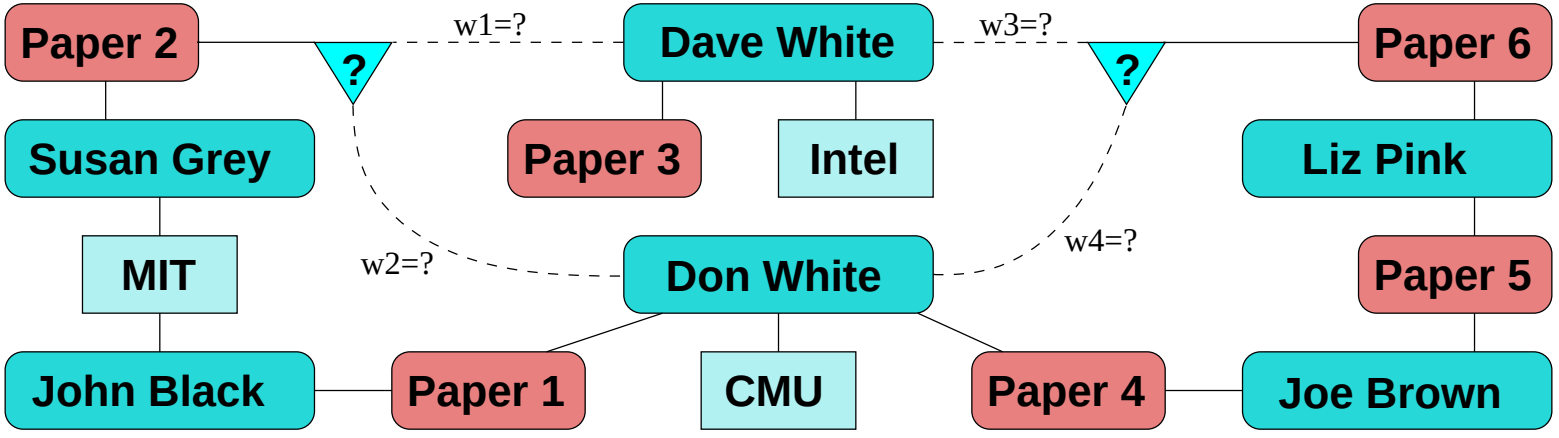
- These data are about living people, and so privacy is of concern when data are linked between organisations
- Linked data allow analysis not possible on individual databases (potentially revealing highly sensitive information)
- Modern databases contain more details and more complex types of data (free-format text or multimedia)
- Data are available from different sources (governments, businesses, social network sites, the Web)
- Major questions: Which data are suitable?
  Which can we get access to?

Advanced classification 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: Requires training data (record pairs with known true match status)
  - Un-supervised: Clustering
- Recently, *collective* classification techniques have been investigated (build graph of database and conduct overall classification, and also take **relational similarities** into account)

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)

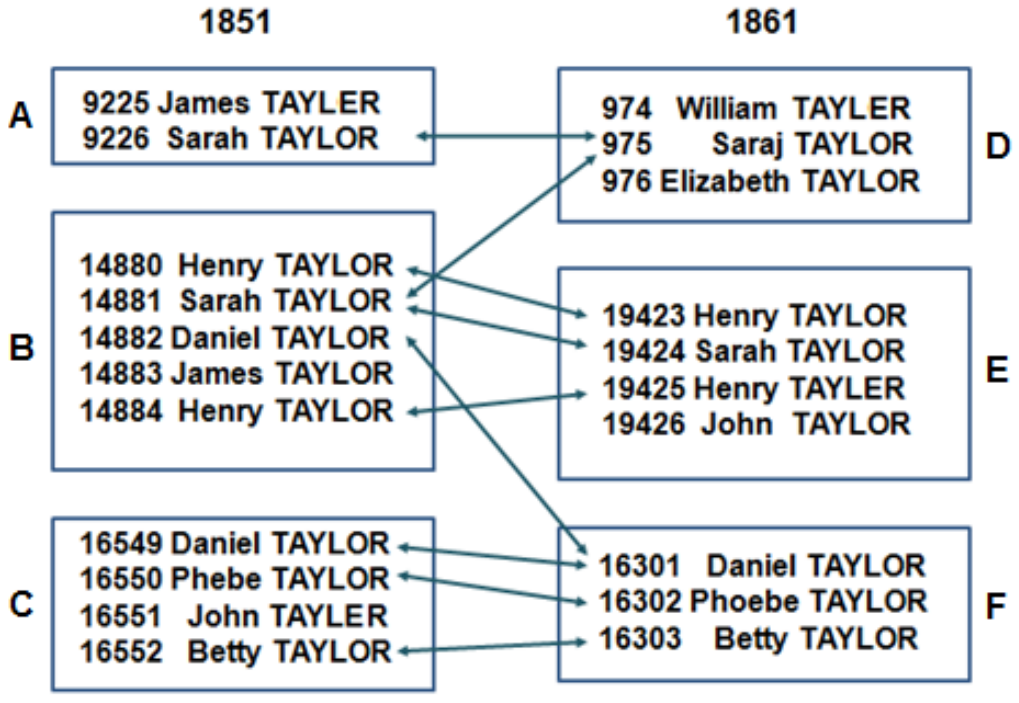
(P1, John Black / Don White)
(P2, Sue Grey / **D. White**)
(P3, Dave White)
(P4, Don White / Joe Brown)
(P5, Joe Brown / Liz Pink)
(P6, Liz Pink / **D. White**)

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

Classification challenges

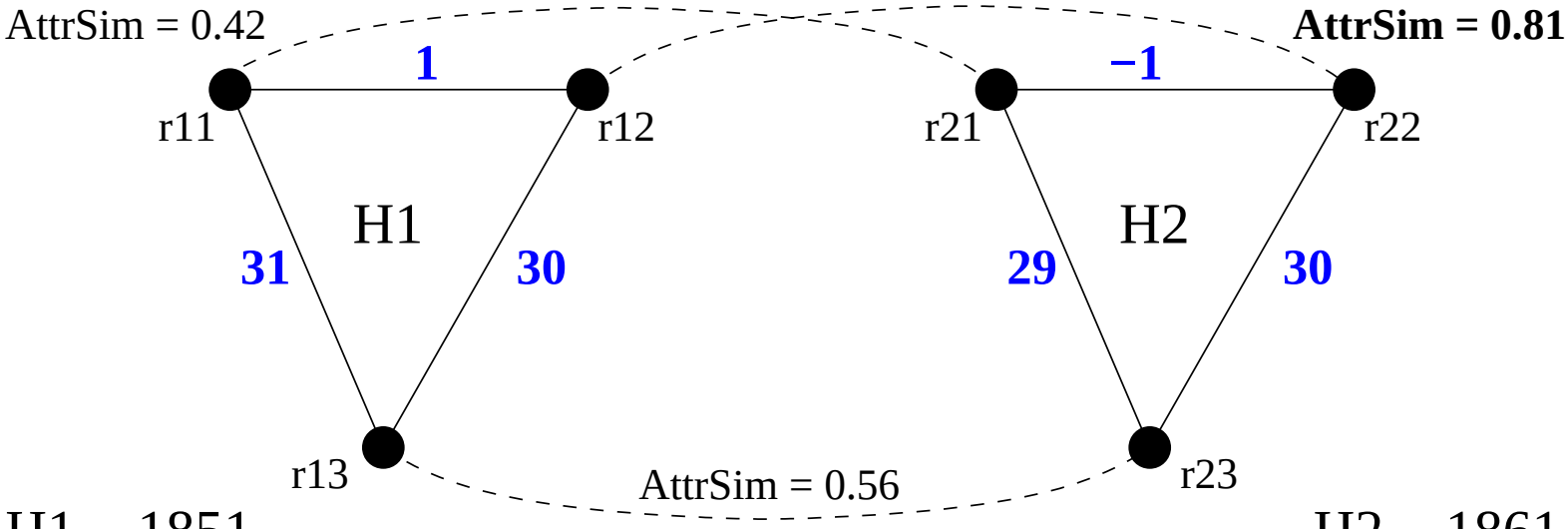
- In many cases there are no training data available
  (no data set with known true match status)
  - Possible to use results of earlier record linkage projects? Or from manual *clerical review* process?
  - How confident can we be about correct manual classification of *potential* matches?
- No large test data collections available
  (unlike in information retrieval or machine learning)
- Many record linkage researchers use synthetic or bibliographic data
  (which have very different characteristics to personal data)

Group matching using household information (Fu et al. 2011, 2012)



- Conduct pair-wise linking of individual records
- Calculate household similarities using Jaccard or weighted similarities (based on pair-wise links)
- 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 (Fu et al. 2014)



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)

Privacy aspects in record linkage

Privacy aspects in record linkage

- Objective is to link data across organisations such that besides the linked records (the ones classified to refer to the same entities) no information about the sensitive source data can be learned by any party involved in the linking, or any external party.
- Main challenges
  - Allow for approximate linking of values
  - Have techniques that are not vulnerable to any kind of attack (frequency, dictionary, crypt-analysis, etc.)
  - Have techniques that are scalable to linking large databases

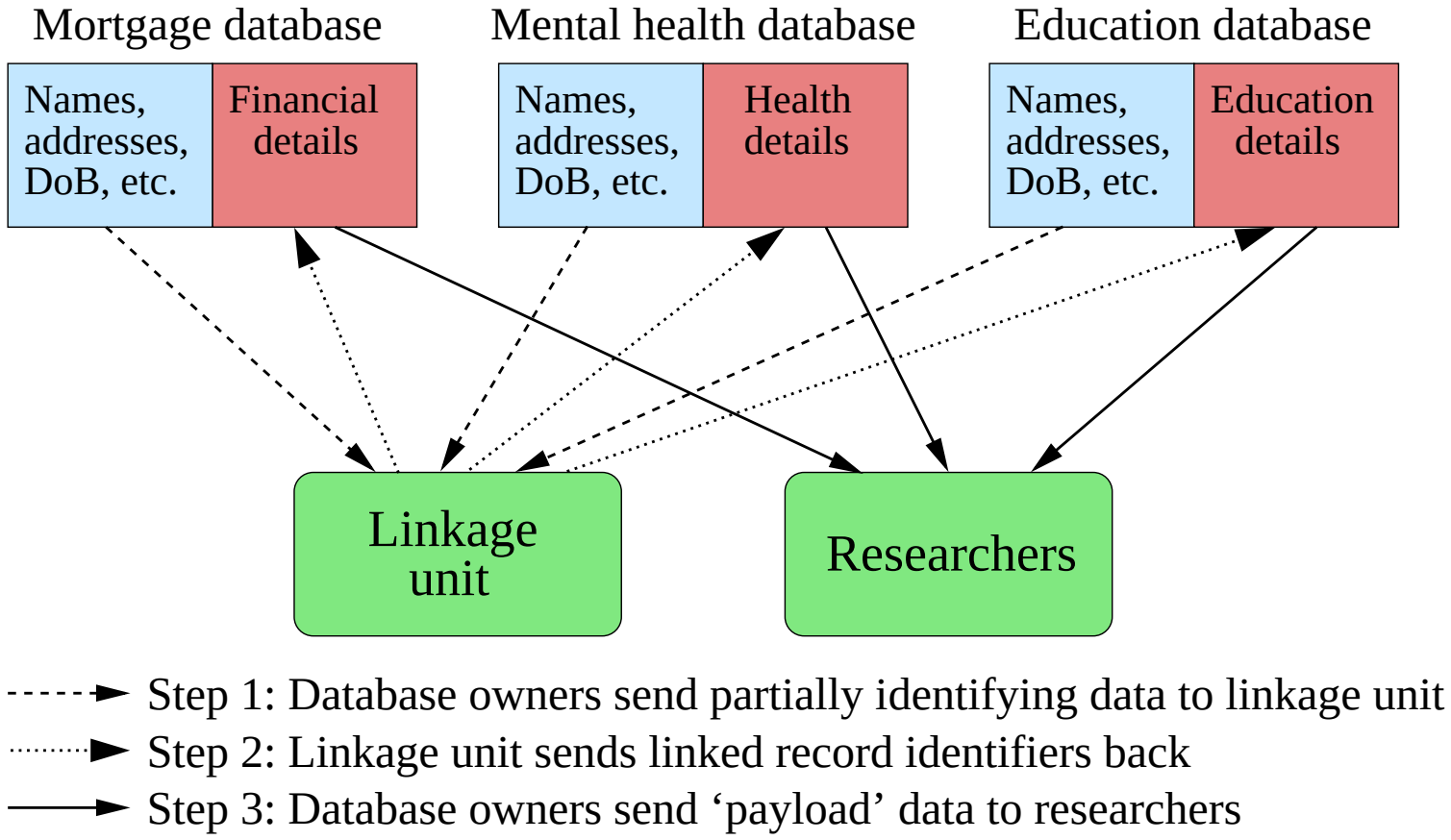
Privacy and record linkage: An example scenario

- A demographer who aims to investigate how mortgage stress is affecting different people with regard to their mental and physical health
- She will need data from financial institutions, government agencies (social security, health, and education), and private sector providers (such as health insurers)
- It is unlikely she will get access to all these databases (for commercial or legal reasons)
- She only requires access to some attributes of the records that are linked, but not the actual identities of the linked individuals (but personal details are needed to conduct the actual linkage)

Current best practice approach used in health domain (1)

- Linking of health data is common in public health (epidemiological) research
- Data are sourced from hospitals, doctors, health insurers, police, governments, etc
- Only identifying data are given to a trusted linkage unit, together with an encrypted identifier
- Once linked, encrypted identifiers are given back to the sources, which ‘attach’ payload data to identifiers and send them to researchers
- Linkage unit does never see payload data
- Researchers do not see personal details
- All communication is encrypted

Current best practice approach used in health domain (2)

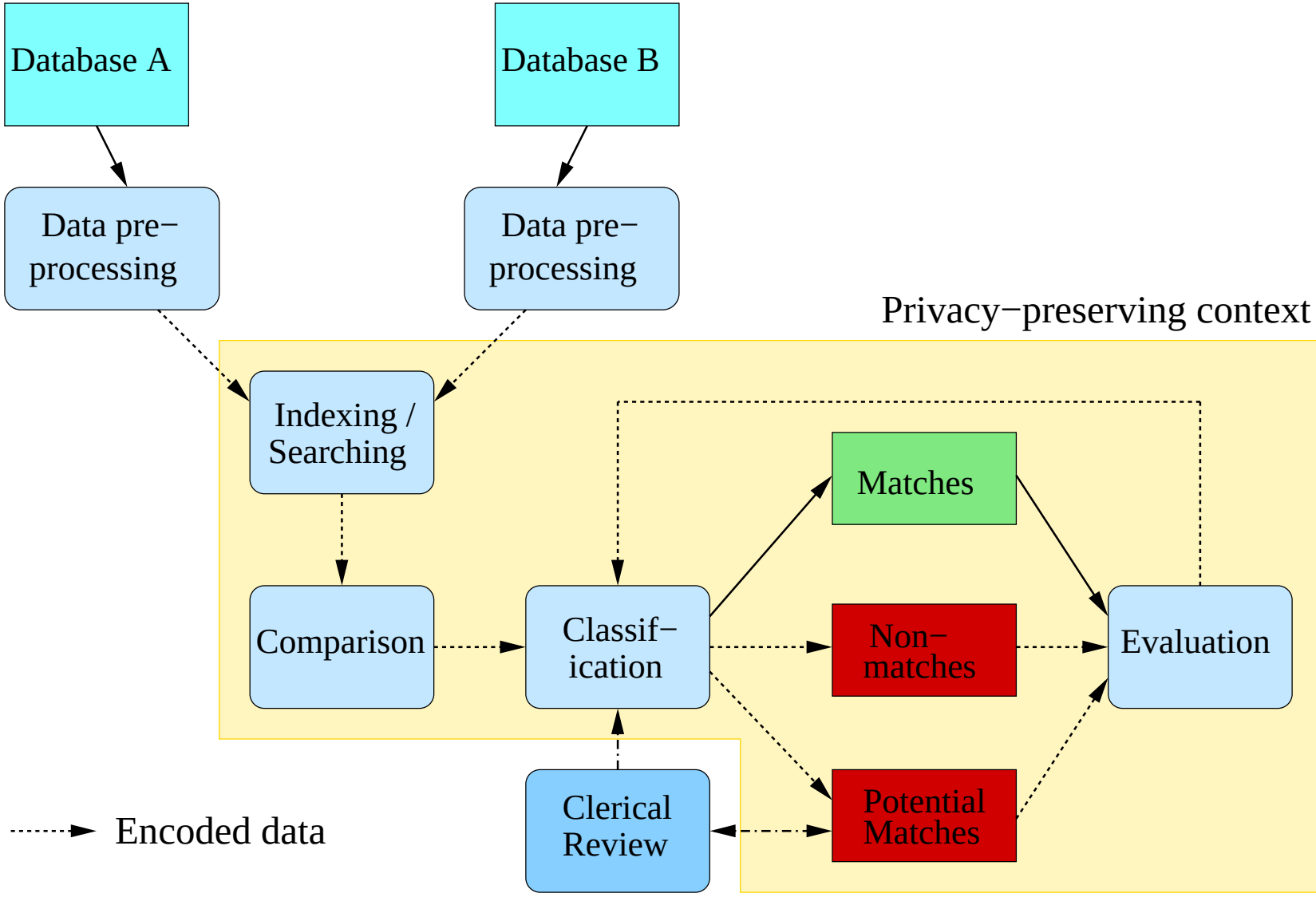


Details given in: Chris Kelman, John Bass, and D'Arcy Holman: *Research use of Linked Health Data – A Best Practice Protocol*, Aust NZ Journal of Public Health, vol. 26, 2002.

Current best practice approach used in health domain (3)

- Problem with this approach is that the linkage unit needs access to personal details
  (metadata might also reveal sensitive information)
- Collusion between parties, and internal and external attacks, make these data vulnerable
- Privacy-preserving record linkage (PPRL) aims to overcome these drawbacks
  - No unencoded data ever leave a data source
  - Only details about matched records are revealed
  - Provable security against different attacks
- PPRL is challenging (employs techniques from cryptography, machine learning, databases, etc.)

The PPRL process



Hash-encoding for PPRL

- A basic building block of many PPRL protocols
- Idea: Use a one-way hash function (like SHA) to encode values, then compare hash-codes
  - Having only access to hash-codes will make it nearly impossible to learn their original input values
  - But dictionary and frequency attacks are possible
- Single character difference in input values results in completely different hash codes
  - For example:
    'peter' $\rightarrow$ '101010...100101' or '4R#x+Y4i9!e@t4o]'
    'pete' $\rightarrow$ '011101...011010' or 'Z5%o-(7Tq1@?7iE/'
  - Only exact matching is possible

Advanced PPRL techniques

- First generation (mid 1990s): exact matching only using simple hash encoding
- Second generation (early 2000s): approximate matching but not scalable (PP versions of edit distance and other string comparison functions)
- Third generation (mid 2000s): take scalability into account (often a compromise between PP and scalability, some information leakage accepted)
- Different approaches have been developed for PPRL, so far no clear best technique

  (for example based on Bloom filters, phonetic encodings, generalisation, randomly added values, or secure multi-party computation)

Challenges and research directions

To make sure everybody is awake.. :-)

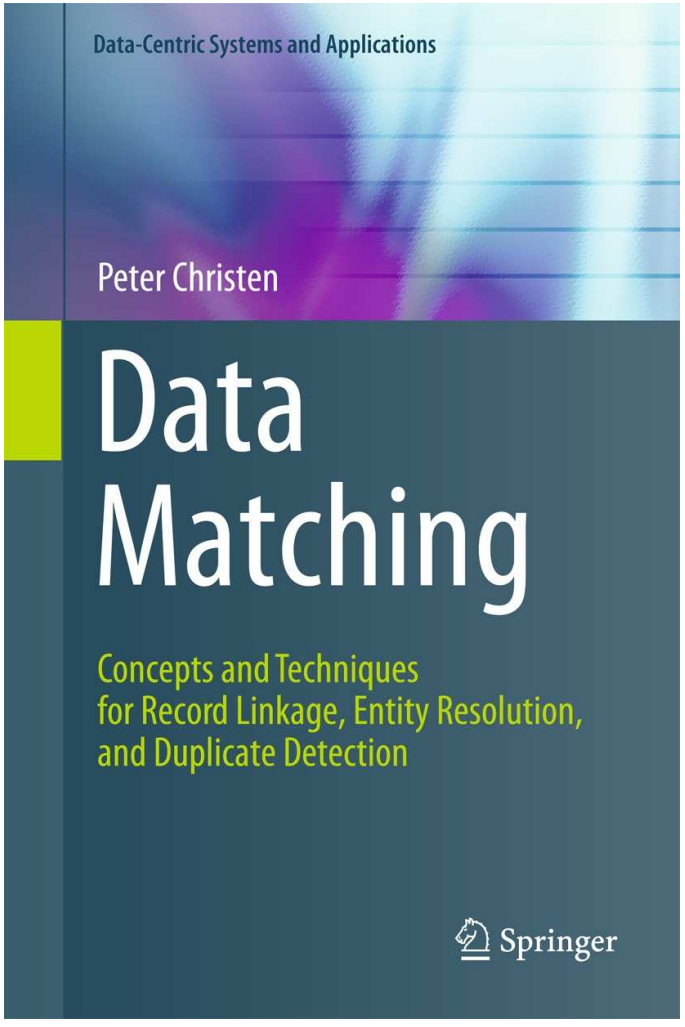
Challenges and research directions (1)

- For historical data, the main challenge is data quality (develop (semi-)automatic data cleaning and standardisation techniques)
- How to employ collective classification techniques for data with personal information?
- No training data in most applications
  - Employ active learning approaches
  - Visualisation for improved manual clerical review
- Linking data from many sources (significant challenge in PPRL, due to issue of collusion)
- Frameworks for record linkage that allow comparative experimental studies

Challenges and research directions (2)

- Collections of test data sets which can be used by researchers
  - Challenging (impossible?) to have true match status
  - Challenging as most data are either proprietary or sensitive
- Develop practical PPRL techniques
  - Standard measures for privacy
  - Improved advanced classification techniques for PPRL
  - Methods to assess accuracy and completeness
- Pragmatic challenge: Collaborations across multiple research disciplines

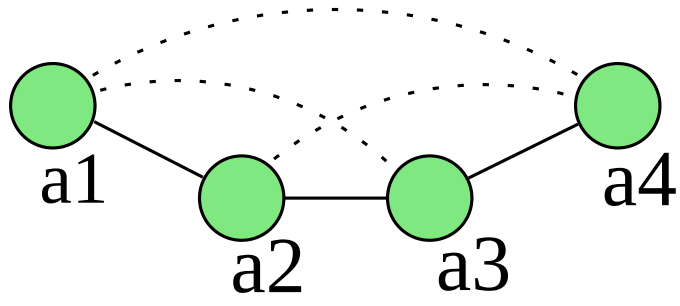
Advertisement: Book 'Data Matching'



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.

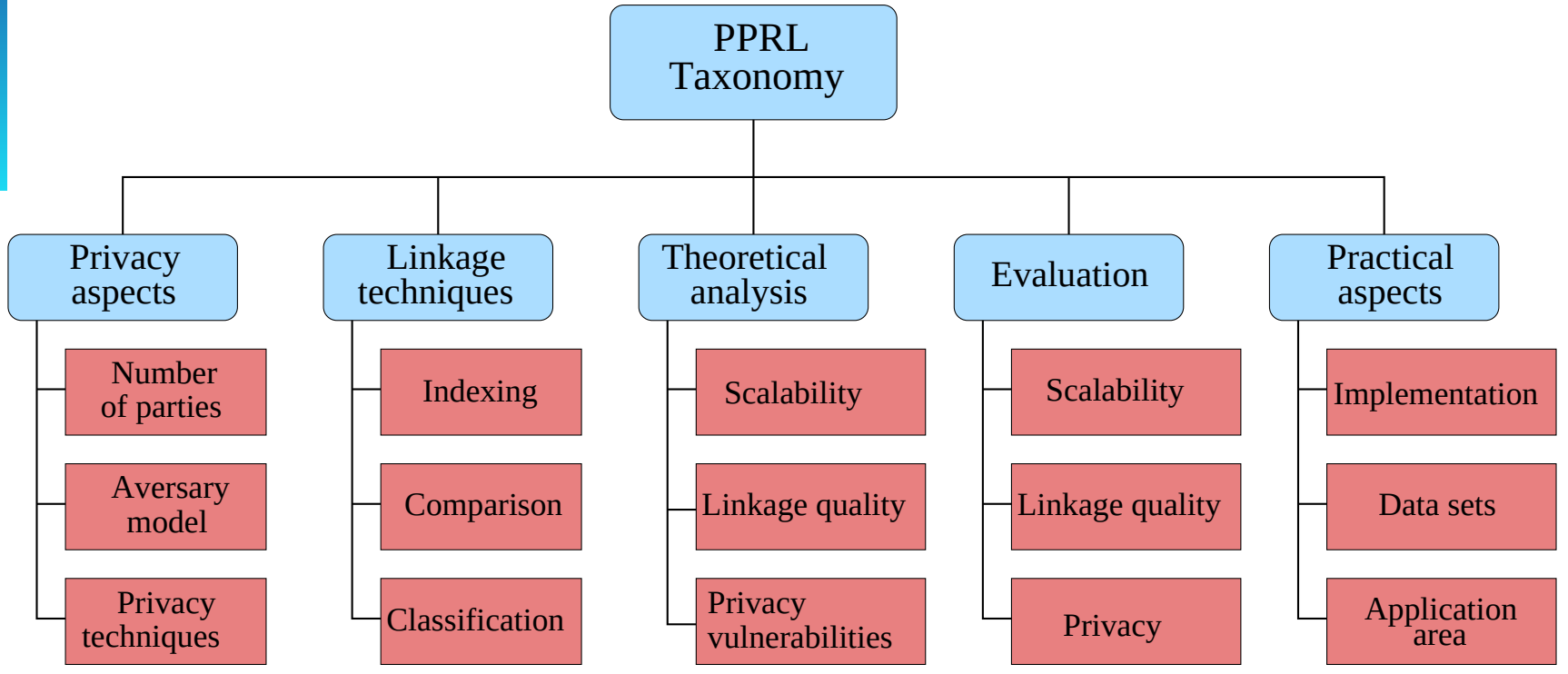
William E. Winkler, U.S. Bureau of the Census.

Managing transitive closure

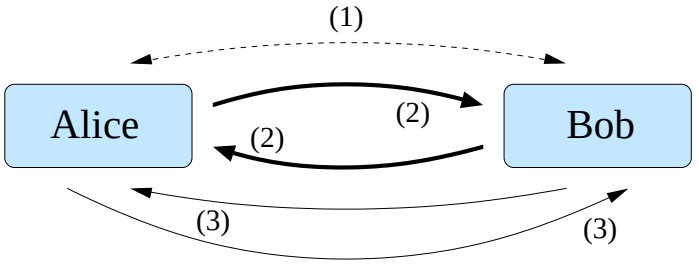


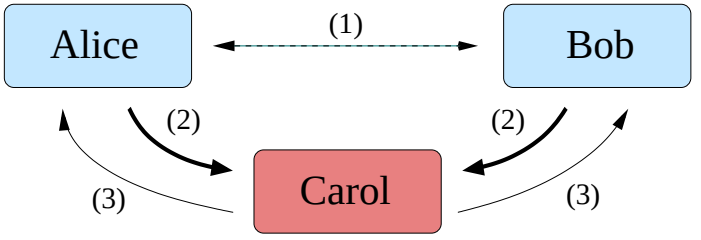
- If record *a1* is classified as matching with record *a2*, and record *a2* as matching with record *a3*, then records *a1* and *a3* must also be matching.
- Possibility of record chains occurring
- Various algorithms have been developed to find optimal solutions (special clustering algorithms)
- Collective classification and clustering approaches deal with this problem by default

A taxonomy for PPRL



Basic PPRL protocols





- 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)

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 $s = \sum_i x_i$.