Academic state-of-the-art of data linkage

#datalinkage

Peter Christen

Research School of Computer Science,
The Australian National University,
Canberra, Australia

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

This work was partially supported by a grant from the *Simons Foundation*. The author would also like to thank the *Isaac Newton Institute for Mathematical Sciences*, Cambridge, for support and hospitality during the programme *Data Linkage and Anonymisation* where this presentation was prepared (EPSRC grant EP/K032208/1).

Outline

- A short introduction to data linkage
- Challenges of data linkage
- Techniques for scalable data linkage
- Advanced classification techniques for data linkage
- Privacy aspects in data linkage
- Research directions

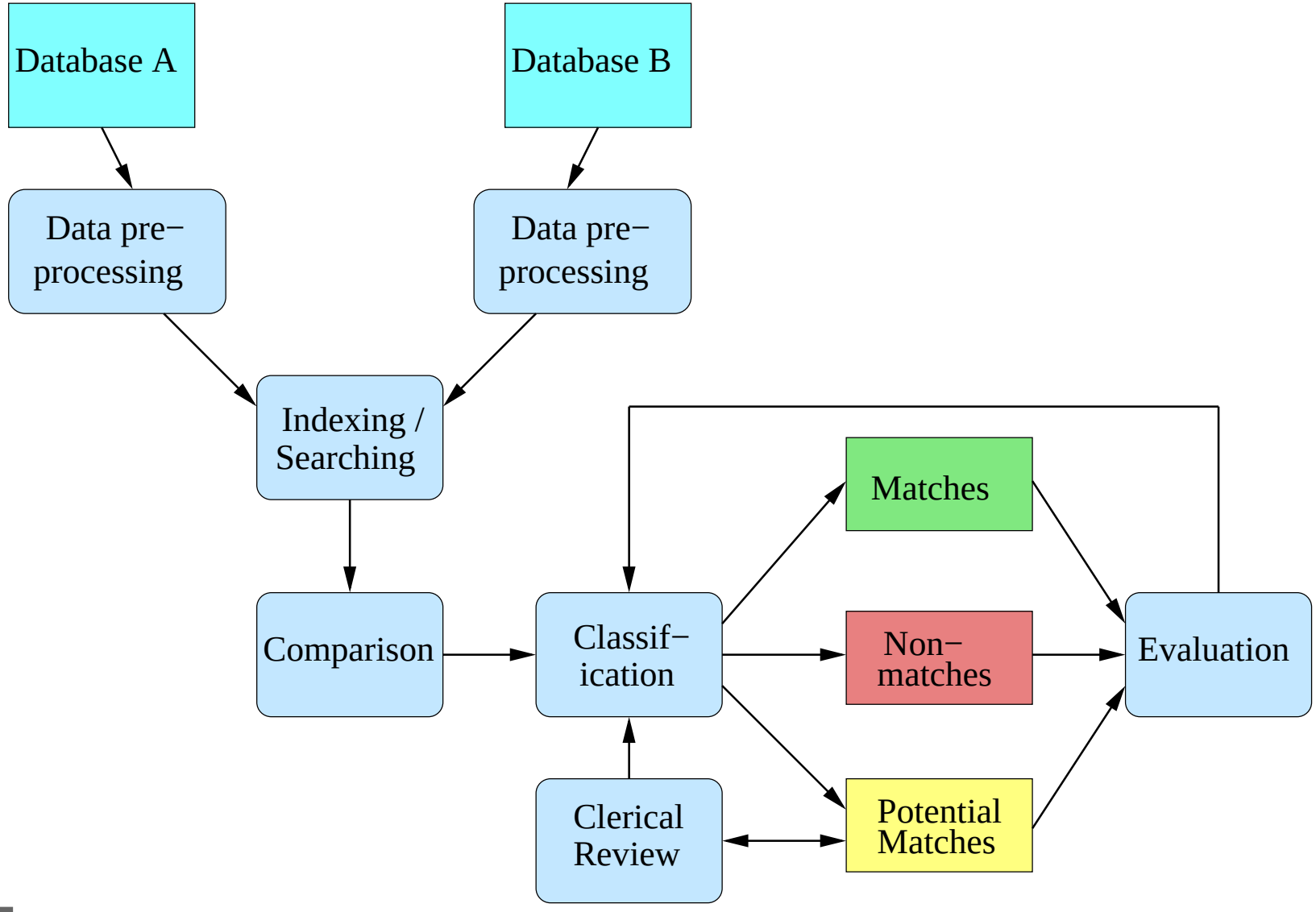
What is data linkage?

- The process of linking records that represent the same entity in one or more databases
  (patients, customers, businesses, publications, etc.)
- Major challenge is that unique *entity identifiers* are not available in the databases to be linked
- Various applications of data linkage
  - Remove duplicates in one data set (deduplication)
  - Merge new records into a larger master data set
  - Clean and enrich data for analysis and mining
  - Geocode matching (with reference addresses)
  - Create longitudinal data set

Recent interest in data linkage

- Traditionally, data linkage has been used in national statistics (census) and health research
- In recent years, increased interest from businesses and governments
  - Massive amounts of data are being collected, and computing power and storage capacities are increasing
  - Often data from different sources need to be integrated
  - Need for data sharing between organisations
  - Data mining (analysis) of large data collections
  - E-Commerce and Web services (comparison shopping)
  - Spatial data analysis and online map applications

The data linkage process



Data linkage techniques

- Deterministic matching
  - Exact matching (if a *unique identifier* of high quality is available: precise, robust, stable over time)
  - Rule based matching (complex to build and maintain)
- Probabilistic record linkage (*Fellegi and Sunter*, 1969)
  - Use available attributes for linking (often personal information) and calculate match weights
- "Computer science" approaches
  - Based on machine learning, data mining, database, or information retrieval techniques
  - Supervised classification: Requires training (truth) data
  - Unsupervised: Clustering, collective, and graph based

Major data linkage challenges

- Real world data are dirty
  (typographical errors and variations, missing and out-of-date values, different coding schemes, etc.)
- Scalability
  - Naïve comparison of all record pairs is quadratic
  - Remove likely non-matches as efficiently as possible
- No training data in many linkage applications
  - No record pairs with known true match status
- Privacy and confidentiality
  (because personal information, like names and addresses, is commonly required for linking)

Challenges for linking Big Data

- Size (*volume*) and complexity (*variety*) of data
  - Possibly hundreds of millions of records about entities
  - From many different sources (internal and external)
  - Contain more complex data types
- Dynamic nature of Big Data (*velocity*)
  - Streams of data (unpredictable rate and volume)
  - (Near) real-time linking and analysis are required
- Trustworthiness of (external) data (*veracity*)
- Diverse requirements on linked data
- Privacy and confidentiality

Outline

- A short introduction to data linkage
- Challenges of data linkage
- Techniques for scalable data linkage
- Advanced classification techniques for data linkage
- Privacy aspects in data linkage
- Research directions

Techniques for scalable data linkage

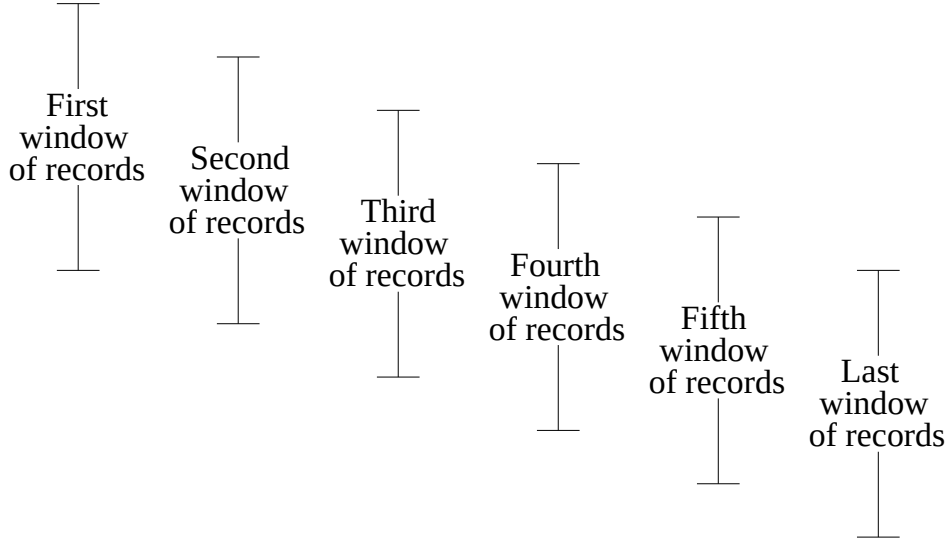
- The number of all record pair comparisons equals the product of the sizes of the two databases
  - But the number of true matches is often less than the number of records in the smaller database
- Performance bottleneck in data linkage is usually the detailed comparison of attribute values
- Aim of indexing / blocking: Cheaply remove record pairs that are obviously not matches
- Traditional blocking only compares record pairs with the same value in a *blocking key*
  (e.g., only compare records with the same *postcode*)

Advanced indexing approaches (1)

- Sorted neighbourhood approach
  - Sliding window over sorted databases
  - Use several passes with different sorting criteria
  - Window size can be fixed or adaptive (based on similarities between records)

For example, database(s) sorted using first and last name:

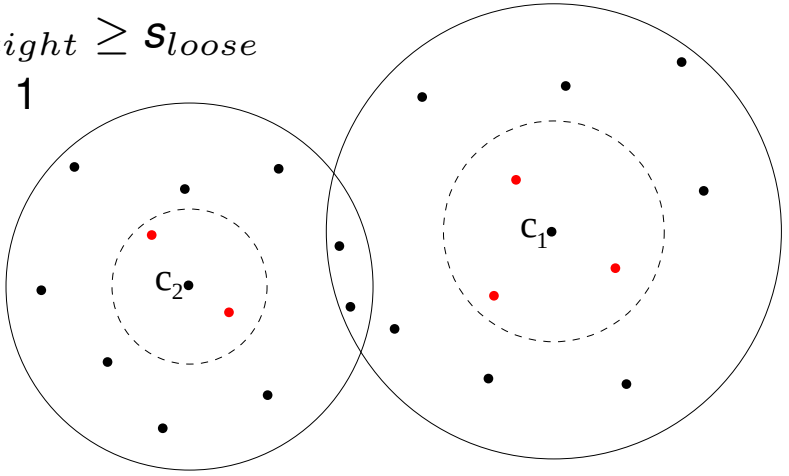
abbybond	r5
paulsmith	r2
pedrosmith	r4
pedrosmith	r9
percysmith	r1
petersmith	r7
petersmith	r10
robinstevens	r3
sallytaylor	r6
sallytaylor	r8



Advanced indexing approaches (2)

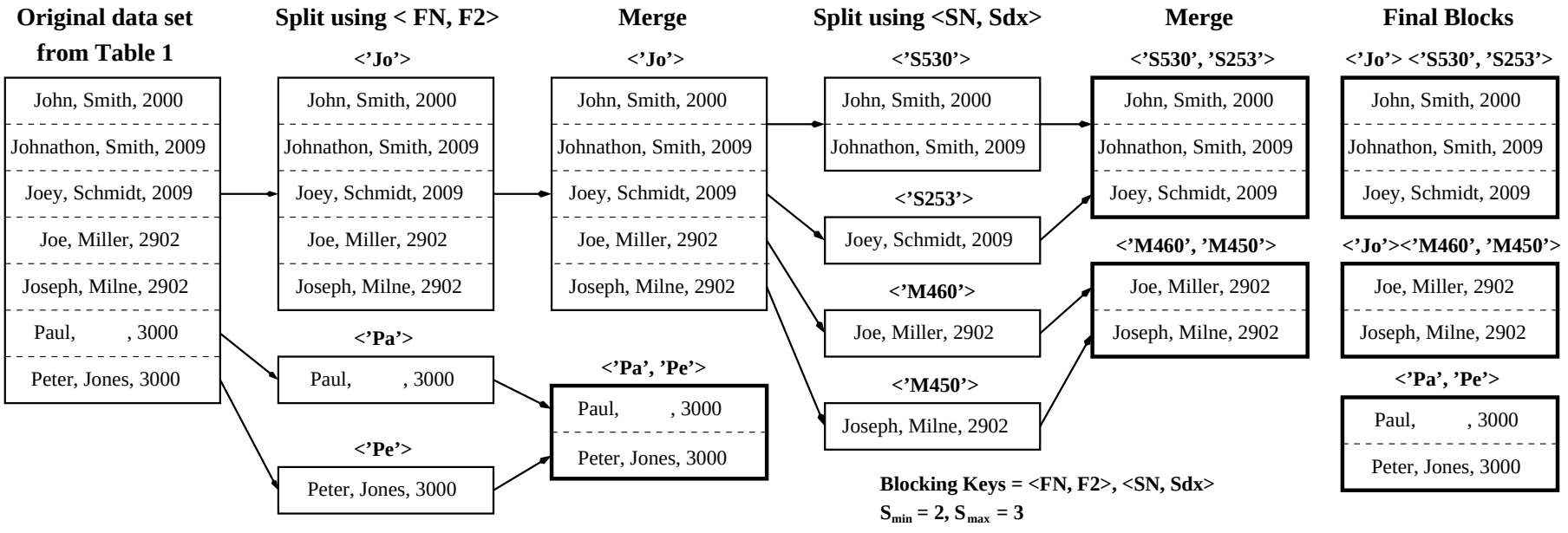
- *Canopy* clustering
  - Based on a fast similarity measure such as Jaccard (set intersection based on q-grams: *john* $\rightarrow$ *jo*, *oh*, *hn*)
  - Records will be inserted into several clusters / blocks
  - Algorithm steps:

    1) Randomly select a record in data set *D* as cluster centroid c_i, $i = 1, 2, \ldots$
    2) Insert all records that have a similarity of at least s_{loose} with c_i into cluster C_i
    3) Remove all records $r_j \in C_i$ (including c_i) that have a similarity of at least s_{tight} with c_i from *D*, with $s_{tight} \geq s_{loose}$
    4) If data set *D* not empty go back to step 1



Controlling block sizes

- Important for real-time and privacy-preserving linkage, and with certain machine learning algorithms (that have a quadratic or higher complexity)
- We have developed an iterative split-merge clustering approach (Fisher et al., ACM SIGKDD, 2015)



Outline

- A short introduction to data linkage
- Challenges of data linkage
- Techniques for scalable data linkage
- Advanced classification techniques for data linkage
- Privacy aspects in data linkage
- Research directions

Advanced classification techniques

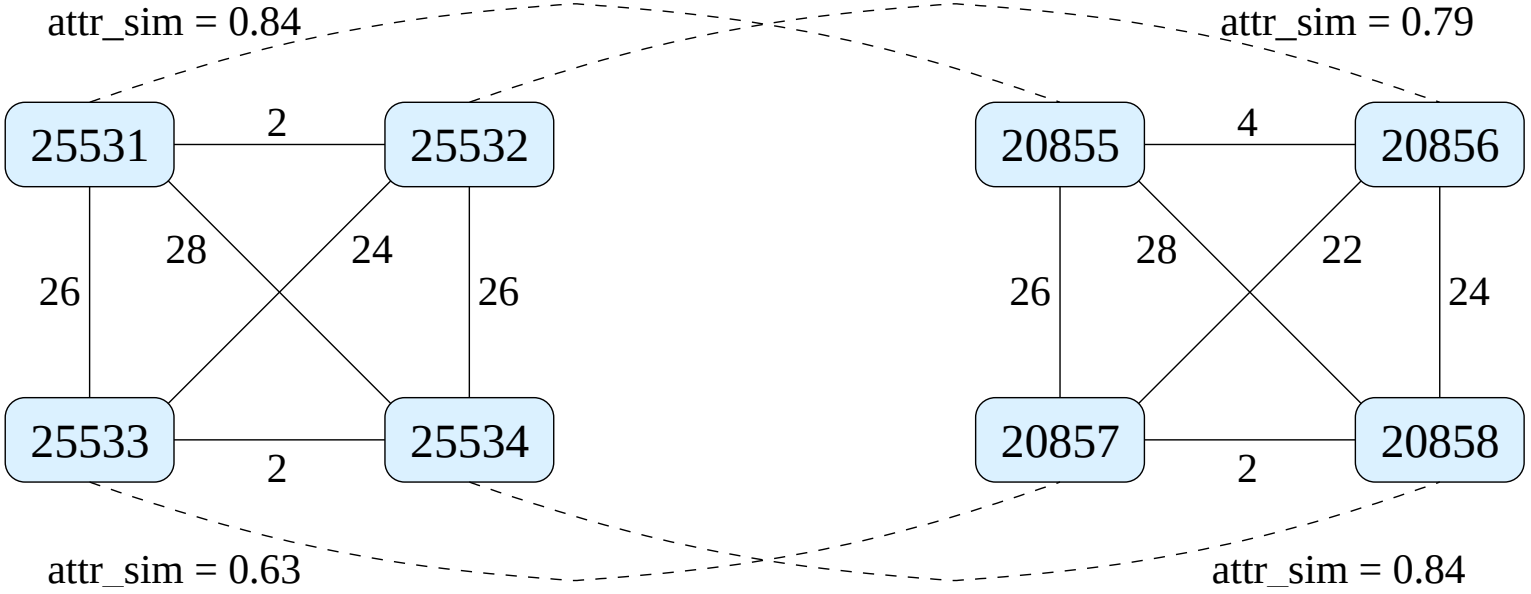
- View record pair classification as a *multi-dimensional binary classification* problem
  - Use all attribute similarities to classify record pairs
  - Only classify into *matches* and *non-matches*
- Many machine learning techniques can be used
  - Supervised: Requires training data (record pairs with known true match and non-match status)
  - Different supervised techniques have been used: *Decision trees*, *support vector machines*, *neural networks*, *learnable string comparisons*, etc.
  - Active and semi-supervised learning
  - Unsupervised: *Clustering*

Classification challenges

- In many cases there are no training data available
  - Possible to use results of earlier matching projects? Or from manual *clerical review* process?
  - How confident can we be about correct manual classification of *potential matches*?
- Often there is no *gold standard* available
  (no data sets with known true match status)
- No large test data set collections available
  (like in information retrieval or machine learning)
  - Due to privacy and confidentiality concerns
  - Therefore much research (in computer science) has been using bibliographic data

Advanced classification: Graph-based linkage

- Based on structure between groups of records
  (for example linking households from different censuses)
  - One graph per household, finds best matching graphs using both record attribute and structural similarities
  - Edge attributes are information that does not change over time (like age differences)

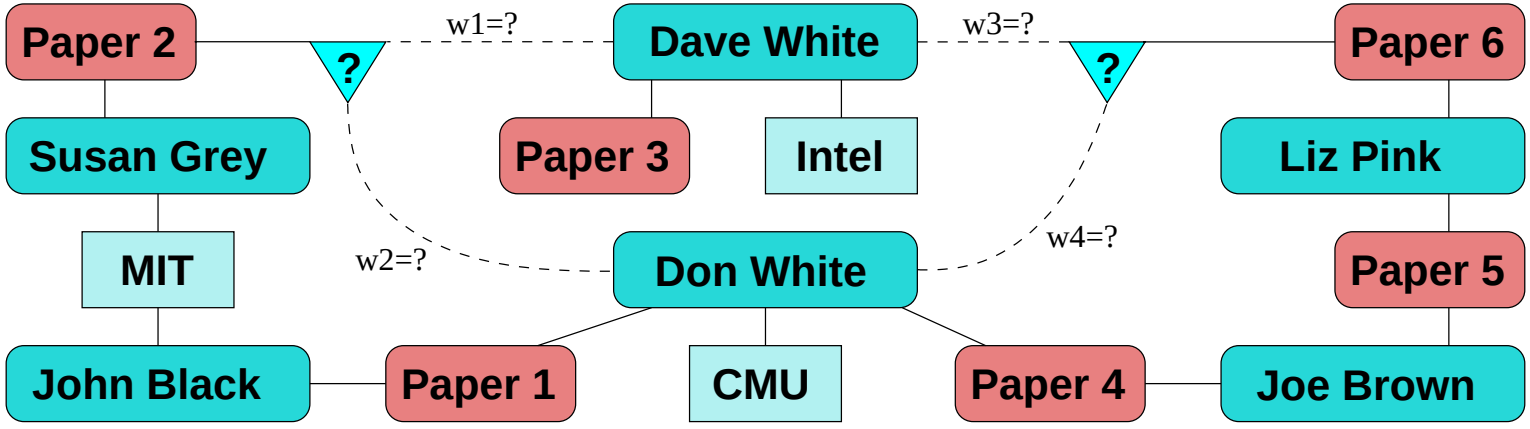


Advanced classification: Active learning and group linkage

- **Active learning**
  - Semi-supervised by human-machine interaction
  - Overcomes the problem of supervised learning that requires training data
  - Selects a sample of record pairs to be manually classified (budget constraints)
  - Trains and improves a classification model using manually labelled data
- **Group linkage** (like families, households, publications)
  - First conduct pair-wise linking of individual records
  - Then calculate group similarities using Jaccard or weighted similarities (based on pair-wise similarities)

Advanced classification: Collective entity resolution

- Considers *relational similarities* not just attribute similarities



(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, 2006

Outline

- A short introduction to data linkage
- Challenges of data linkage
- Techniques for scalable data linkage
- Advanced classification techniques for data linkage
- Privacy aspects in data linkage
- Research directions

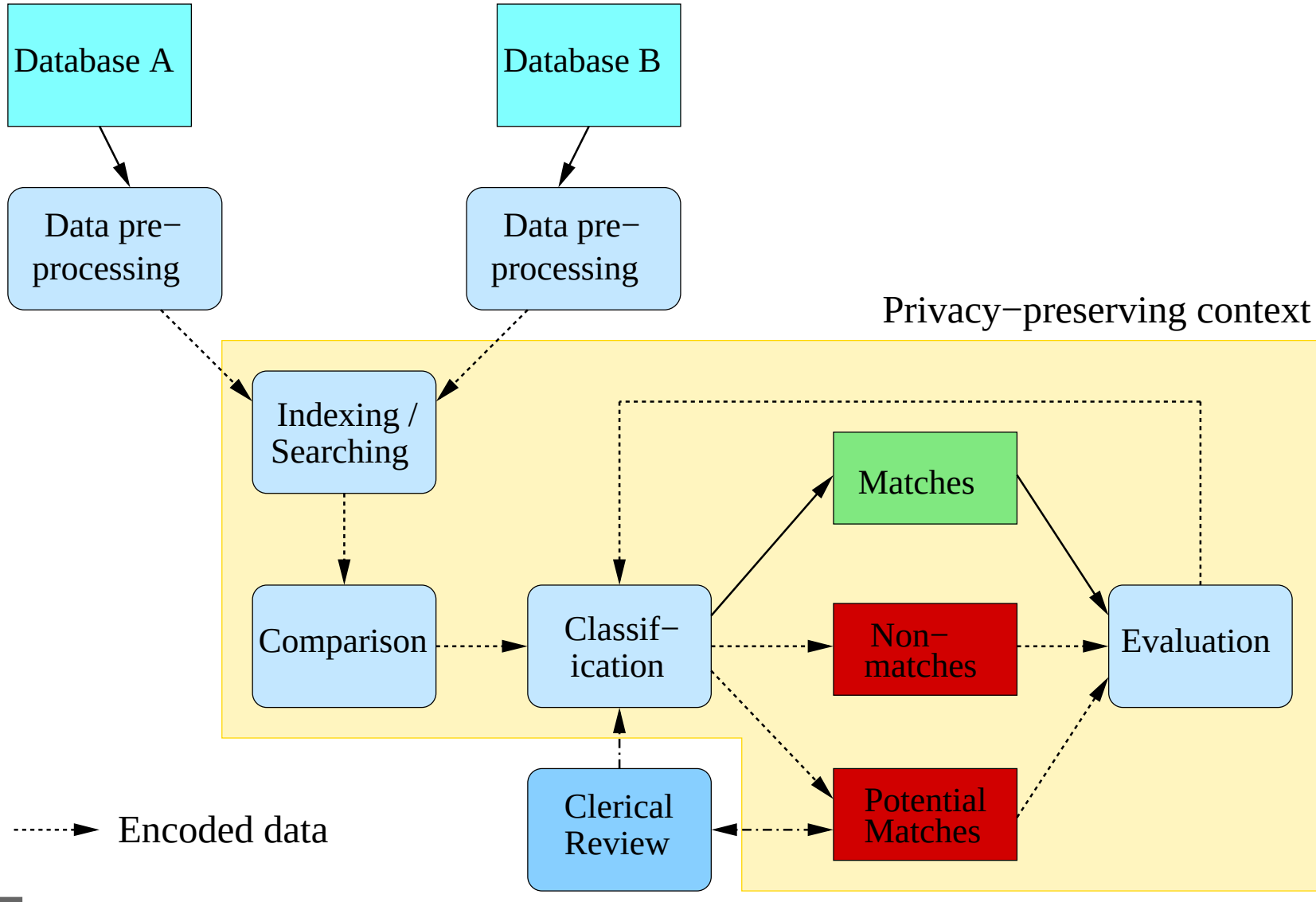
Privacy and data linkage: A motivating scenario

- A demographer aims to investigate how mortgage stress is affecting people with regard to their mental and physical health
- She will need data from financial institutions, government agencies, and private sector health providers (banks; social security, health, and education agencies; GPs and specialists; and health insurances)
- 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 (however personal details are needed to conduct the actual linkage)

Privacy-preserving record linkage

- *The objective of PPRL is to link databases across organisations such that besides certain attributes of matches (record pairs classified to refer to the same entity) no information about sensitive data can be learned by any party involved in the linkage, or any external party.*
- PPRL has many challenges
  - Allow for approximate linking of values
  - Being able to assess linkage quality and completeness
  - 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 across multiple parties

The PPRL process

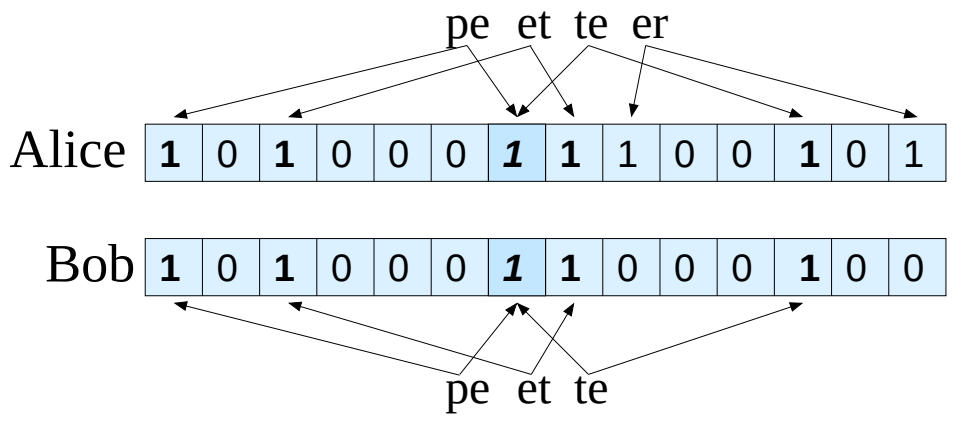


Hash-encoding for PPRL

- 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 between two input values results in completely different hash-codes
  - For example:
    SHA('peter') $\rightarrow$ '4R#x+Y4i9!e@t4o]'
    SHA('pete') $\rightarrow$ 'Z5%o-(7Tq1@?7iE/'
  - Only exact matching is possible

Bloom filter encoding

(Schnell et al., BMC Med Inform Decis Mak, 2009)



'peter': x_1=*7*, 'pete': x_2=*5*, *c*=*5*, therefore sim_{Dice} = 2×5/(7+5)= 10/12 = 0.83

- Bloom filters are bit vectors initially set to *0*-bits
- Use *k* hash functions to hash-map a set of elements by setting corresponding *k* bit positions to *1*
- A set of *q*-grams (from strings) are hash-mapped to allow approximate matching
- Dice similarity of two Bloom filters b_1 and b_2 is: $sim_{Dice}(b_1, b_2) = \frac{2 \times c}{(x_1 + x_2)}$, with: $c = |b_1 \cap b_2|$, $x_i = |b_i|$

Outline

- A short introduction to data linkage
- Challenges of data linkage
- Techniques for scalable data linkage
- Advanced classification techniques for data linkage
- Privacy aspects in data linkage
- Research directions

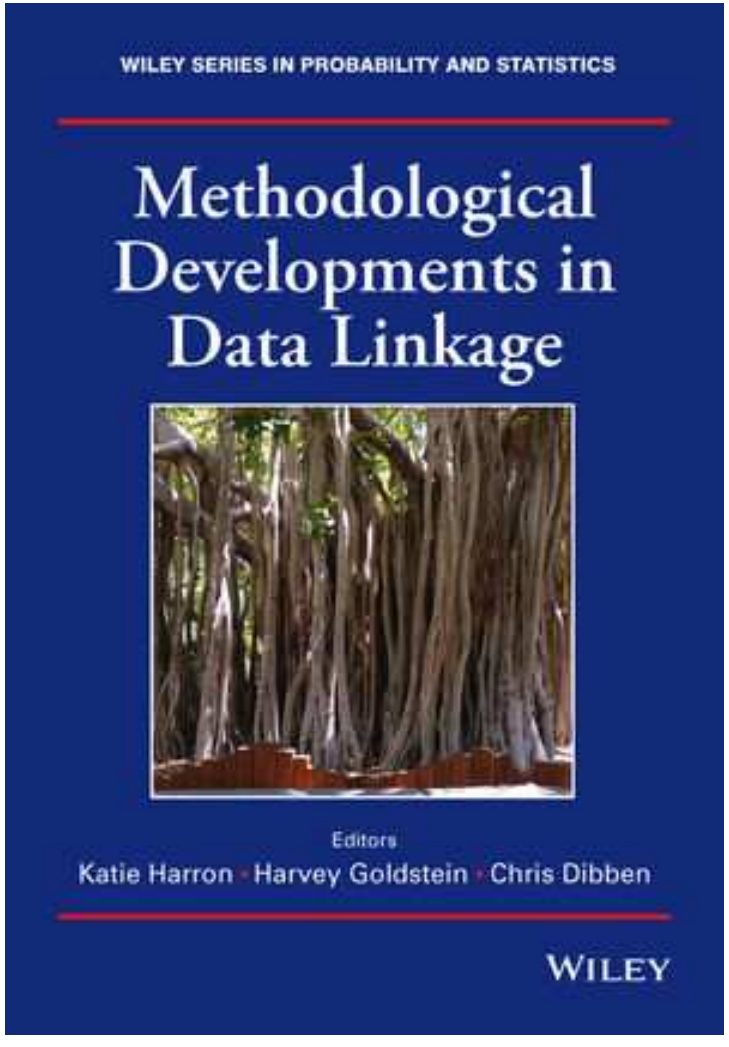
Research directions (1)

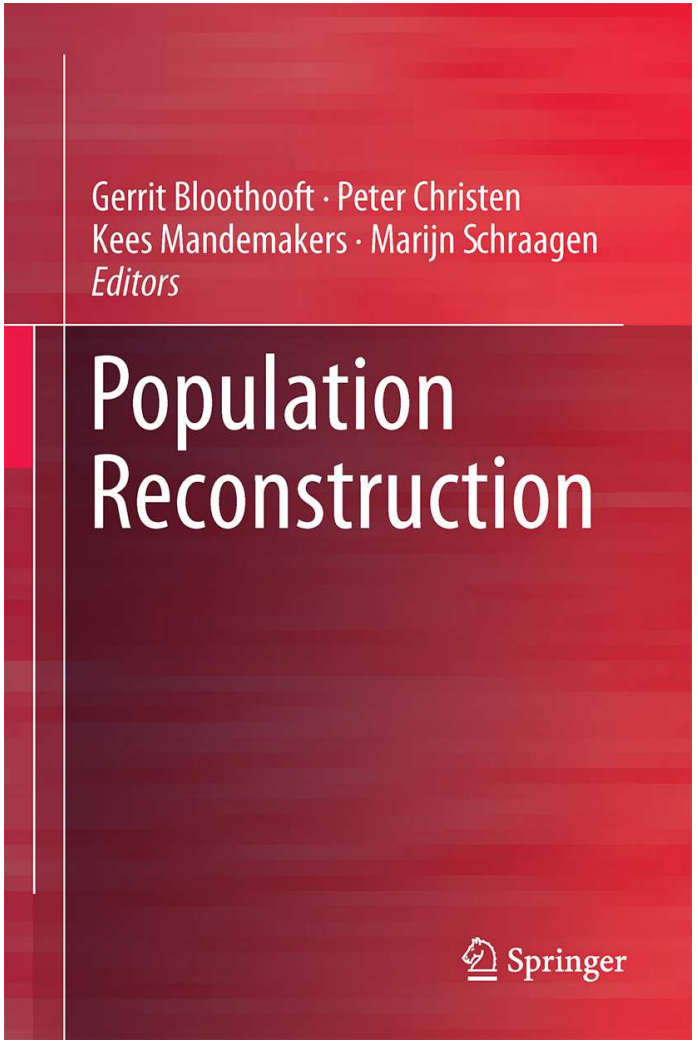
- Linkage techniques for massive-scale Big data collections (parallel, distributed, cloud based)
- Linking data from many sources
  (pair-wise linking does not scale to many databases)
- Linking dynamic data and linking data in real-time
  (dynamic indexing techniques and classification models)
- No training data in most applications
  - Active learning approaches
  - Visualisation for improved manual clerical review
- Frameworks for data linkage that allow comparative experimental studies

Research directions (2)

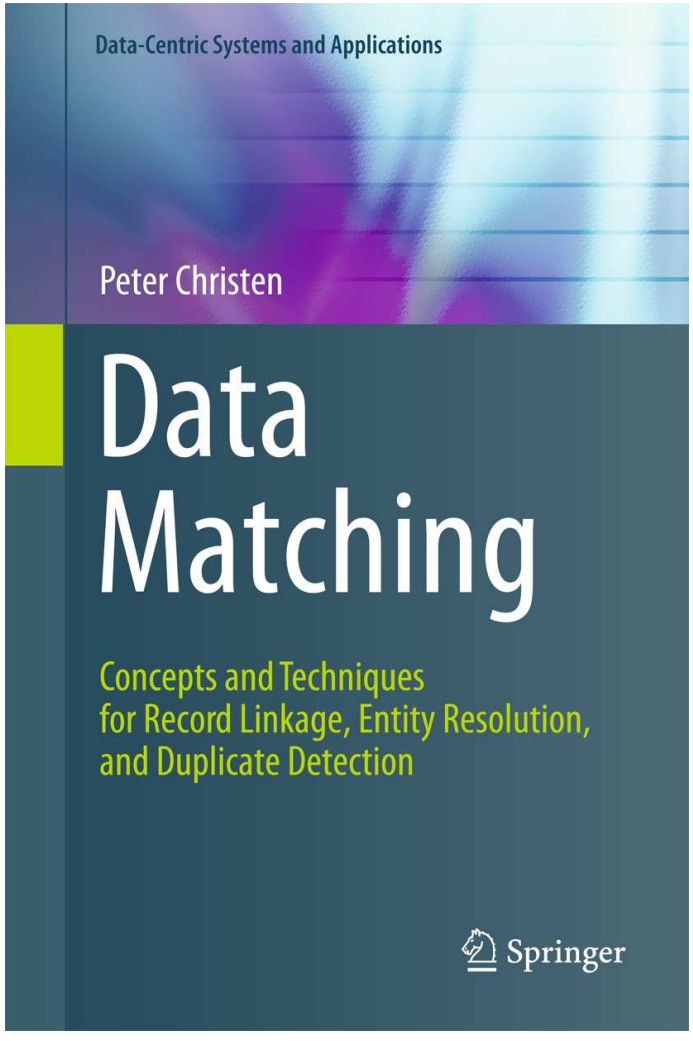
- Publicly available test data collections
  - Challenging (impossible?) to have true match status
  - Challenging as most data are proprietary or sensitive
- Making PPRL practical
  - Improved classification (not only simple thresholds)
  - Assessing linkage quality (access to actual record values not possible as this reveals sensitive information)
  - PPRL on multiple databases (preventing collusion between (sub-groups of) parties becomes more difficult)
- Pragmatic challenge: Collaborations across multiple (research) disciplines

Advertisement: Recent books





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.