Privacy-preserving Record Linkage

Peter Christen

Research School of Computer Science,

ANU College of Engineering and Computer Science,

The Australian National University

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

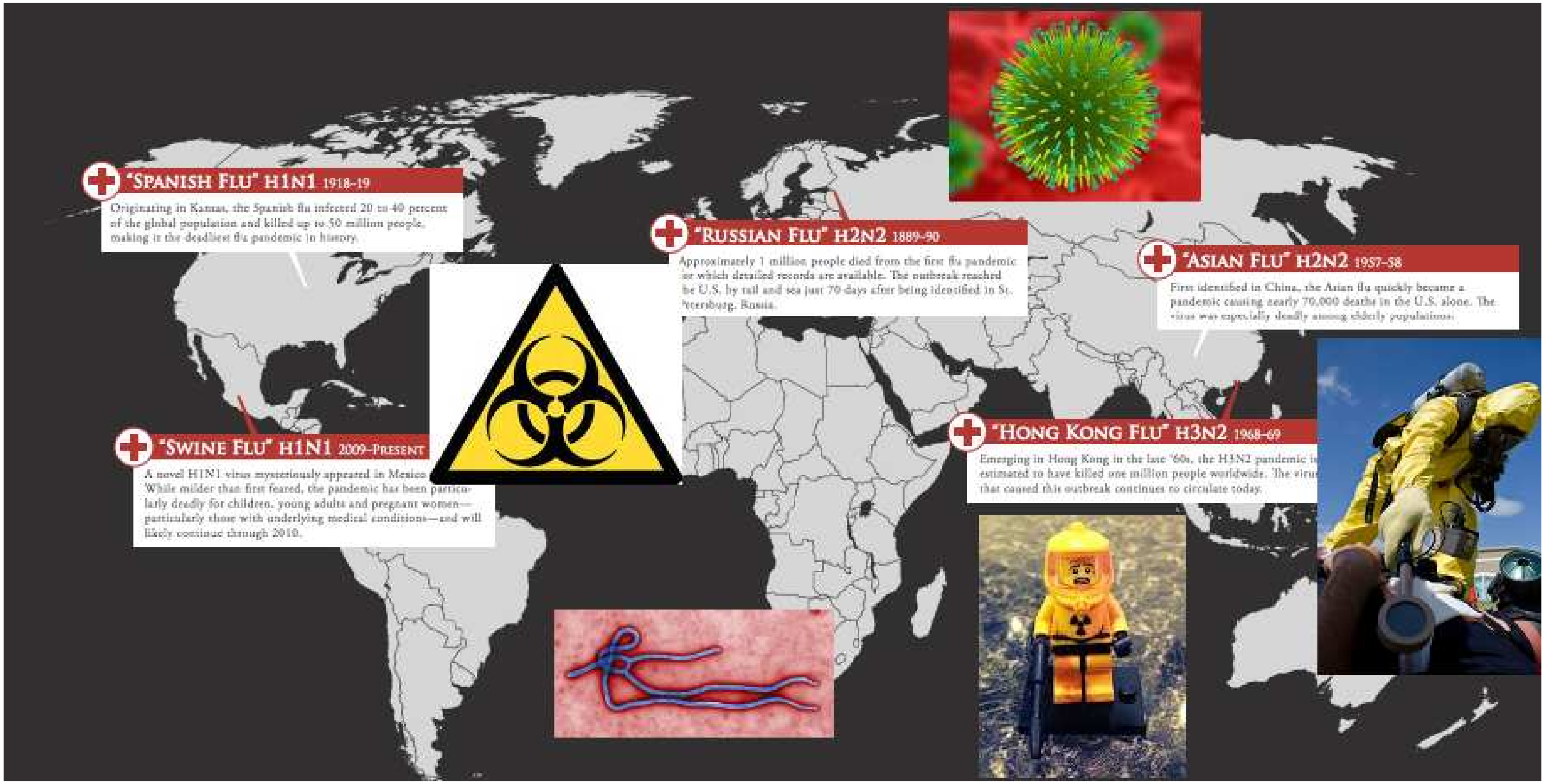
This work was partially funded by the Australian Research Council (ARC)
under Discovery Projects DP130101801.

The author would 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 (https://www.newton.ac.uk/event/dla).

Motivation

- Many Big Data applications require data from different sources to be integrated and linked
  - To allow data analyses that are impossible on individual databases
  - To enrich data with additional information
  - To improve data quality
- Lack of unique *entity identifiers* means that linking often has to be based on personal information
- When databases are linked across organisations, maintaining privacy and confidentiality is vital
- The linking of databases is challenged by **data quality**, **database size**, and **privacy concerns**

Motivating example: Health surveillance (1)



Motivating example: Health surveillance (2)

- Preventing the outbreak of epidemics requires monitoring of occurrences of unusual patterns of symptoms, ideally 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; social network and location data; and possibly even animal health data)
- Privacy and confidentiality concerns arise if such data are stored and linked at a central location
- Such data sets are **sensitive**, **large**, **dynamic**, **heterogeneous** and **distributed**, and they require **linking** and **analysis** in near **real time**

Outline

- A short introduction to record linkage
- Privacy aspects in record linkage
  - Motivating scenarios
  - Basic challenges and protocols
  - Privacy-preserving record linkage
- Conclusions and research directions

What is record linkage?

- The process of linking records that represent the same entity in one or more databases
  (patients, customers, businesses, consumer products, publications, etc.)
- Also known as *data linkage*, *data matching*, *entity resolution*, *duplicate detection*, 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*

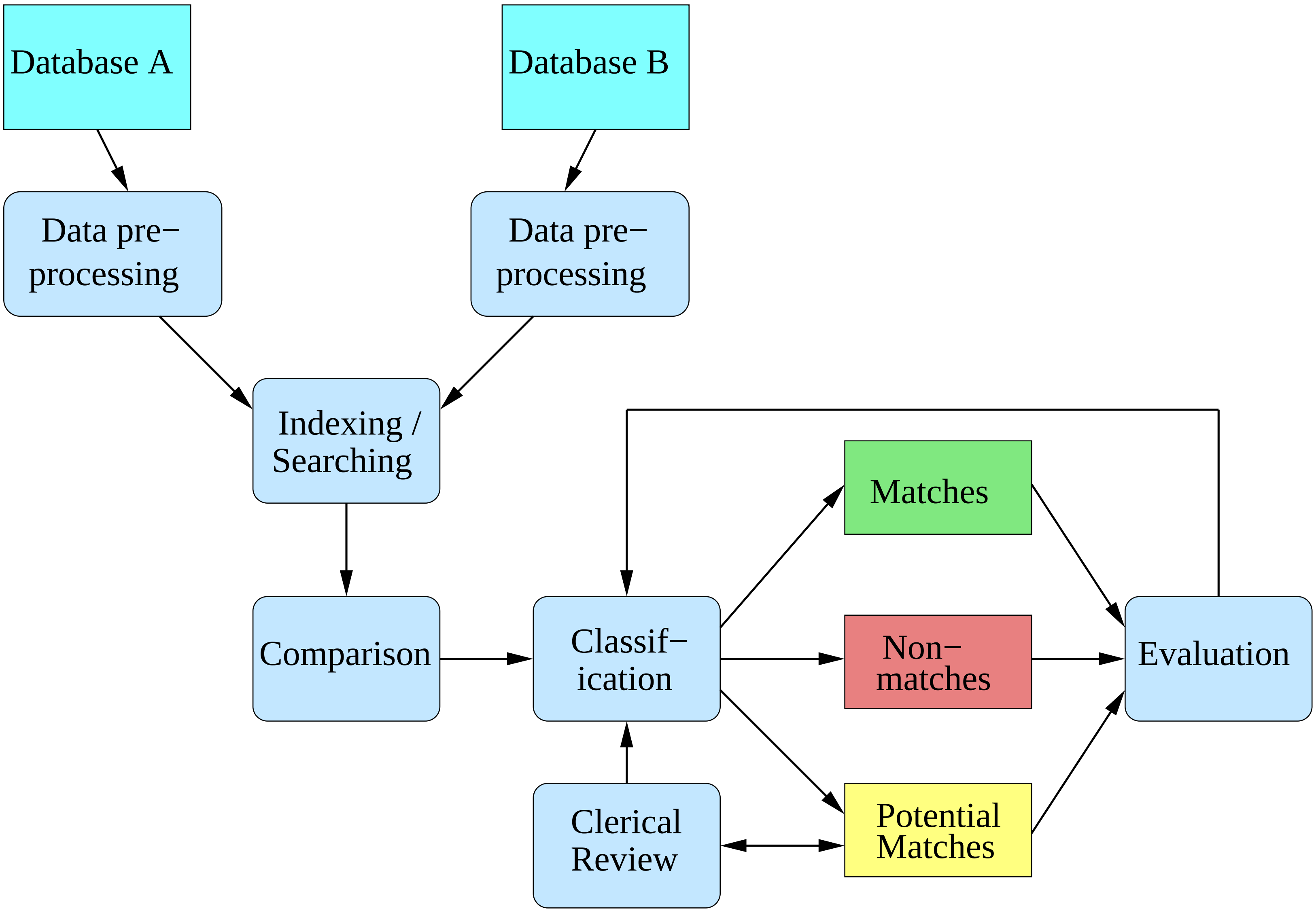
THE AUSTRALIAN NATIONAL UNIVERSITY

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

ANU
THE AUSTRALIAN NATIONAL UNIVERSITY

The record linkage process



Record linkage techniques

- Deterministic matching
  - Rule-based matching (complex to build and maintain)
- Probabilistic record linkage (*Fellegi and Sunter*, 1969)
  - 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
  - Supervised classification: Requires training data (true matches and true non-matches)
  - Unsupervised: Clustering, collective, and graph based

Major record linkage challenges

- No unique entity identifiers available
- 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)

Privacy aspects in record linkage

Privacy aspects in record linkage

- Objective: *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 organisation involved in the linking, or any external organisation.*
- Main challenges
  - Allow for approximate linking of values
  - Being able to asses 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 AUSTRALIAN NATIONAL UNIVERSITY

Privacy and record linkage: Motivating scenario 1

- 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 (however, personal details are needed to conduct the actual linkage)

ANU
THE AUSTRALIAN NATIONAL UNIVERSITY

Privacy and record linkage: Motivating scenario 2

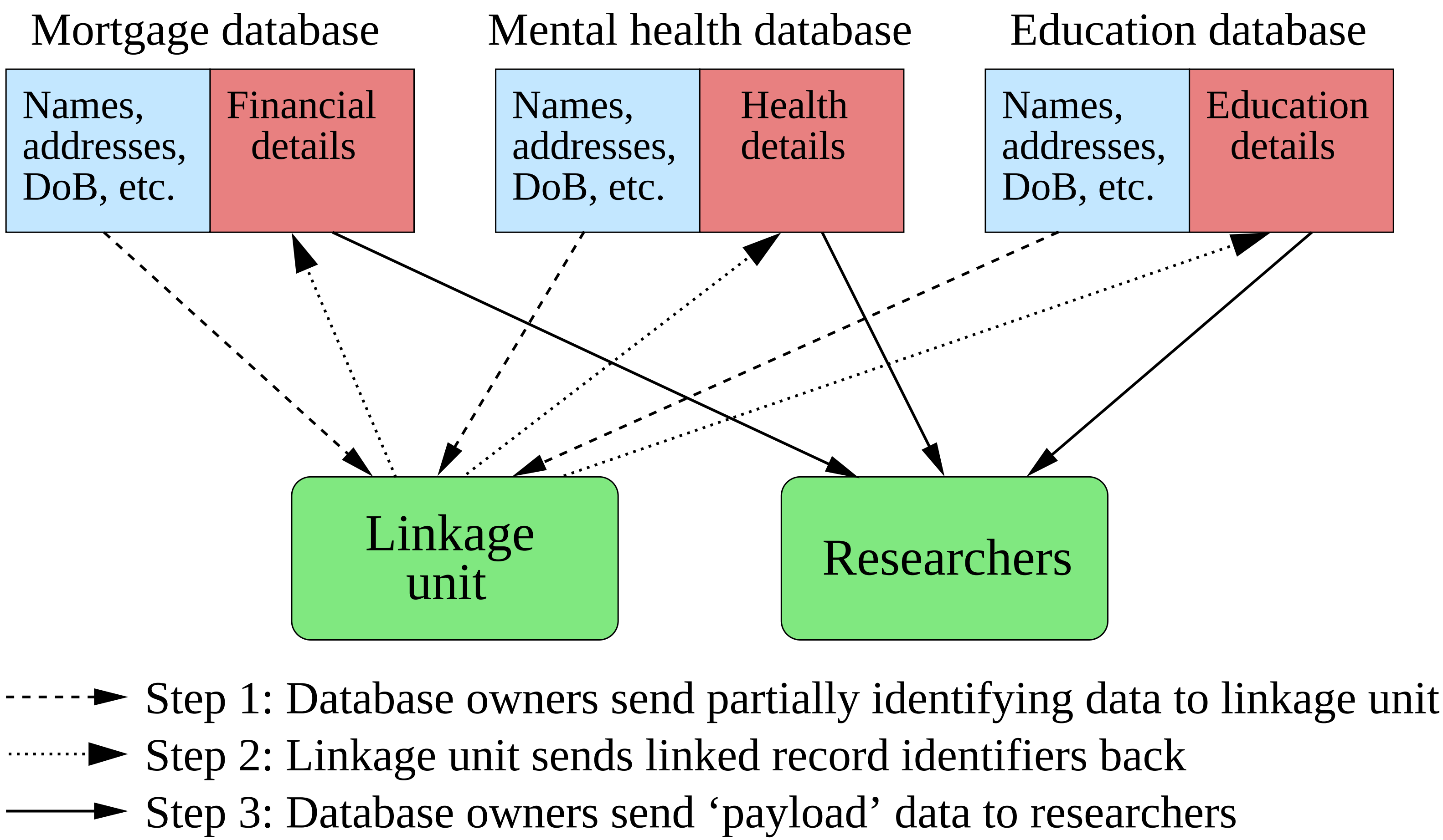
- A national crime investigation unit is tasked with fighting against crimes that are of national significance (organised crime or money laundering)
- This unit will likely manage various national databases which draw from different sources (law enforcement and tax agencies, Internet service providers, and financial institutions)
- These data are highly sensitive; and storage, analysis and sharing must be tightly regulated (collecting such data in one place makes them vulnerable to outsider attacks and internal adversaries)
- Ideally, only linked records (such as those of suspicious individuals) are available to the unit

Current best practice approach used in the 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 never sees payload data
    - Researchers do not see personal details
    - All communication is encrypted

THE AUSTRALIAN NATIONAL UNIVERSITY

Current best practice approach used in the health domain (2)

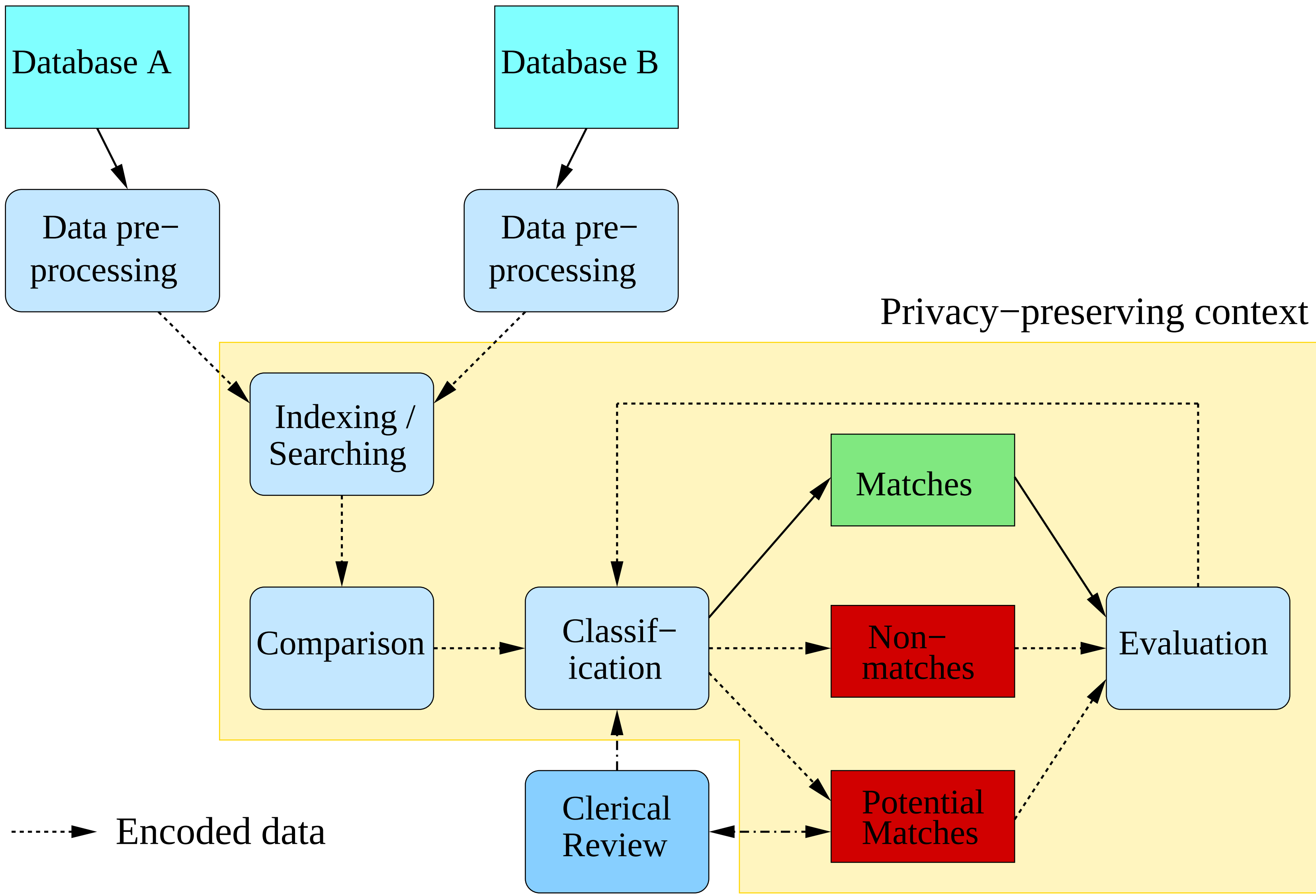


Details given in: Chris Kelman, John Bass, and D’Arcy Holman: *Research use of Linked Health Data – A Best Practice Protocol*, Aust NZ Journal of Public Health, vol. 26, 2002.

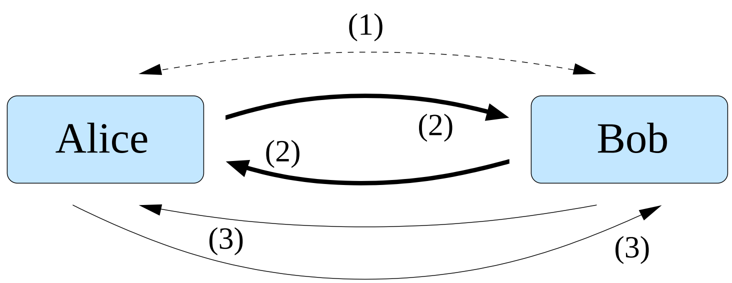
Current best practice approach used in the health domain (3)

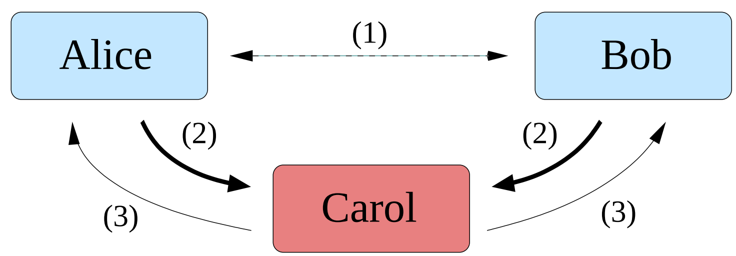
- Problem with this approach is that the linkage unit needs access to personal details
  (metadata might also reveal sensitive information)
- Collusion between parties, and internal and external attacks, make these data vulnerable
- *Privacy-preserving record linkage* (PPRL) aims to overcome these drawbacks
  - 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, databases, etc.)

The PPRL process



Basic PPRL protocols





- Two basic types of protocols
  - Two-party: Only the two database owners who wish to link their data
  - Three-party: Use a (trusted) third party (linkage unit) to conduct the linkage (this party will never see any unencoded values, but collusion is possible)
- Multi-party protocols: Linking records from more than two databases (with or without a linkage unit)

Adversary models

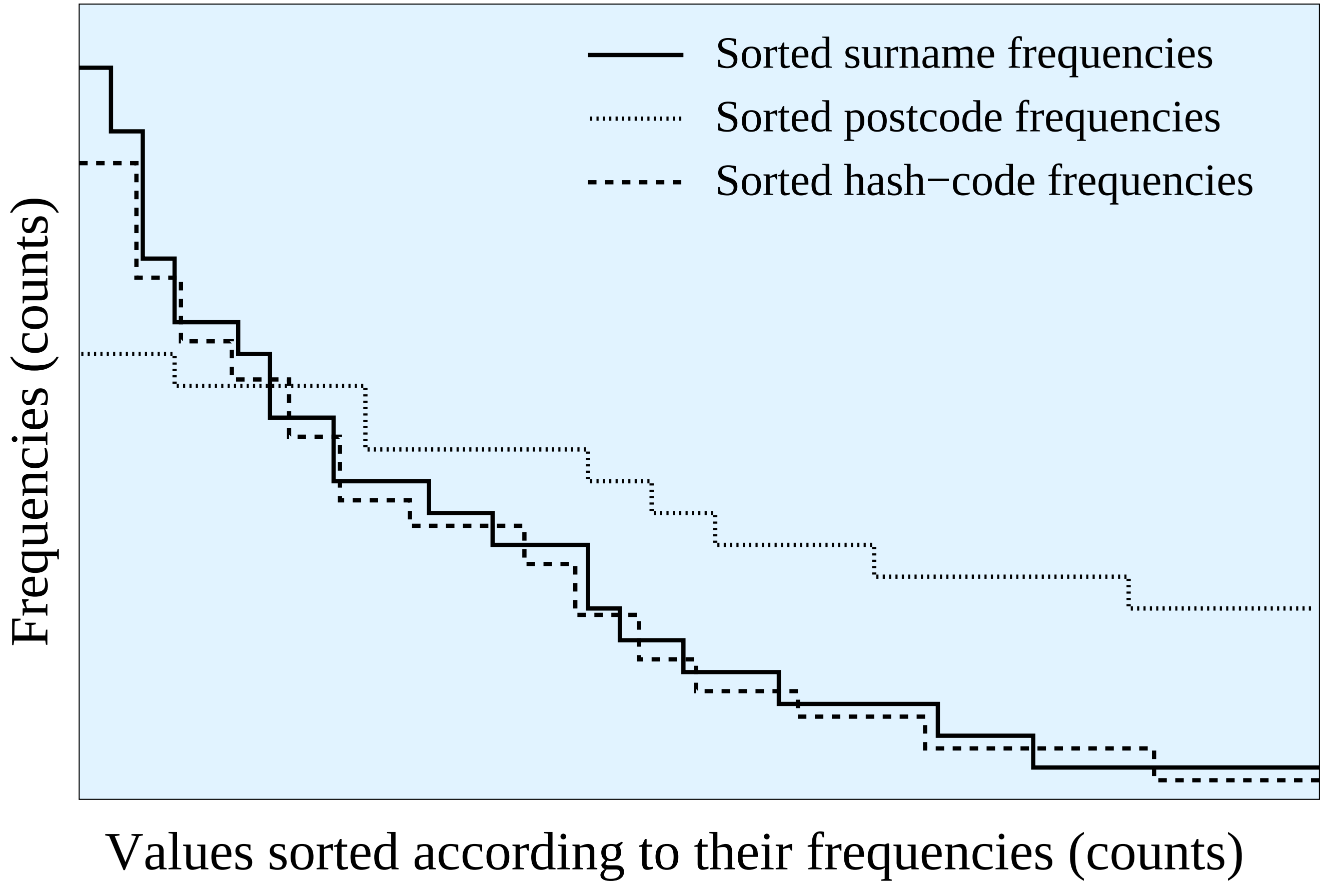
- *Honest-but-curious* (HBC) model assumes that parties follow the protocol while being curious to find about another party's data
  - HBC model does not prevent collusion
  - Most existing PPRL protocols assume HBC model
- *Malicious* model assumes that parties behave arbitrarily (do not follow the protocol)
  - Protocols under this model often have high complexity
- *Accountable computing and covert model*
  - Allow for proofs if a party has followed the protocol or the misbehaviour can be detected with high probability
  - Lower complexity than malicious and more secure than HBC

Attack methods

- *Dictionary* attacks
  An adversary encodes a list of known values using existing encoding functions until a matching encoded value is identified (a keyed encoding approach, like HMAC, can help prevent this attack)
- *Frequency* attacks
  Frequency distribution of encoded values is matched with the distribution of known values
- *Cryptanalysis* attack
  A special category of frequency attack applicable to Bloom filter based encoding
- *Collusion*
  A set of parties (in three- or multi-party protocols) collude with the aim to learn about another party's data

ANU
THE AUSTRALIAN NATIONAL UNIVERSITY

Frequency attack example



- If frequency distribution of hash-encoded values closely matches the distribution of values in a (public) database, then ‘re-identification’ of values might be possible

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, embedding space, generalisation, noise addition, differential privacy, or secure multi-party computation (SMC)

PPRL techniques: 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:

    'peter' $\rightarrow$ '101010...100101' or '4R#x+Y4i9!e@t4o]'
    'pete' $\rightarrow$ '011101...011010' or 'Z5%o-(7Tq1@?7iE/'
  - Only exact matching is possible

THE AUSTRALIAN NATIONAL UNIVERSITY

PPRL techniques: Reference values and embedding

- Reference values
  - Values extracted from a publicly available source in the same domain (e.g. telephone directory) or randomly generated values
  - Calculate similarities between private values using the similarities of each private value with the reference value (triangular inequality)
- Embedding space
  - Embeds records into multi-dimensional space while preserving the distances
  - Difficult to determine the dimension of space and select suitable pivots

PPRL techniques: Noise and differential privacy

- Noise addition
  - Extra (fake) records to perturb data
  - Overcomes frequency attack (improves privacy) at the cost of more comparisons and loss in linkage quality (due to false matches)
- Differential privacy
  - Alternative to noise addition
  - The probability of holding any property on the perturbed database is approximately the same whether or not an individual value is present in the database
  - Magnitude of noise depends on privacy parameter and sensitivity of data

THE AUSTRALIAN NATIONAL UNIVERSITY

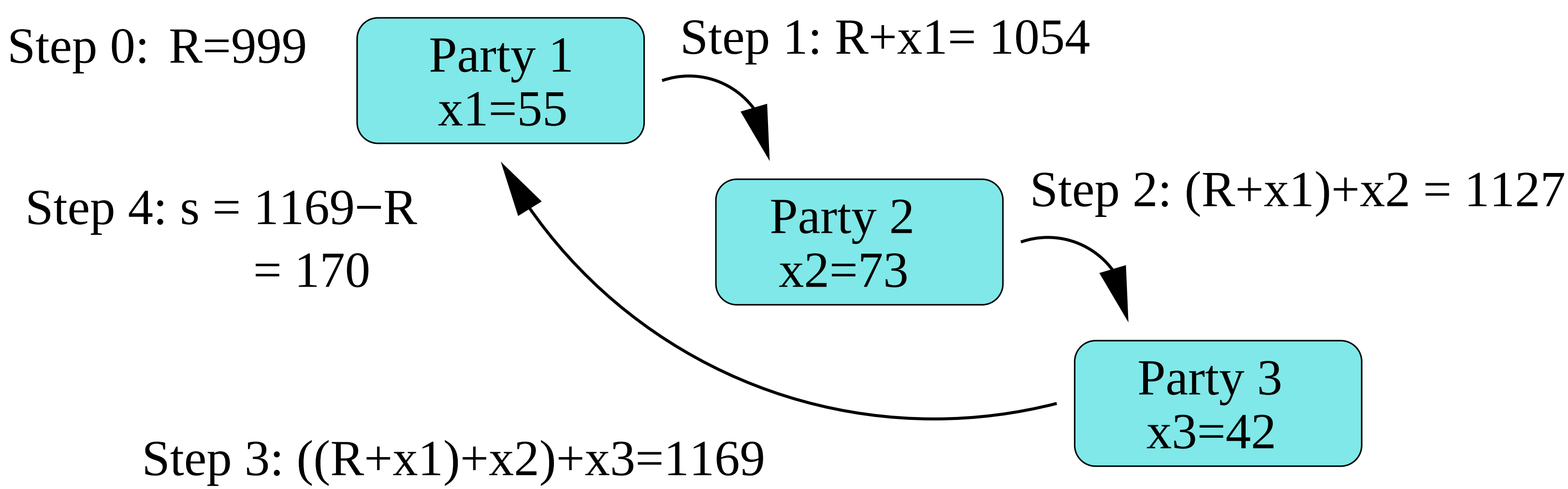
PPRL techniques: Encryption and generalisation

- Value generalisation
  - Generalises the records to overcome frequency attacks
  - For example k-anonymity: ensure every combination of attribute values is shared by at least k records
  - Other techniques are value generalisation hierarchies, top-down specialisation, and binning
- Encryption schemes (SMC)
  - Commutative and homomorphic encryption are used
  - Secure scalar product, secure set intersection, and secure set union are the most commonly used SMC techniques
  - However, many are computationally expensive

THE AUSTRALIAN NATIONAL UNIVERSITY

PPRL techniques: 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, Foundations of Computer Science, 1982]
- Simple example: Secure summation $s = \sum_i x_i$.

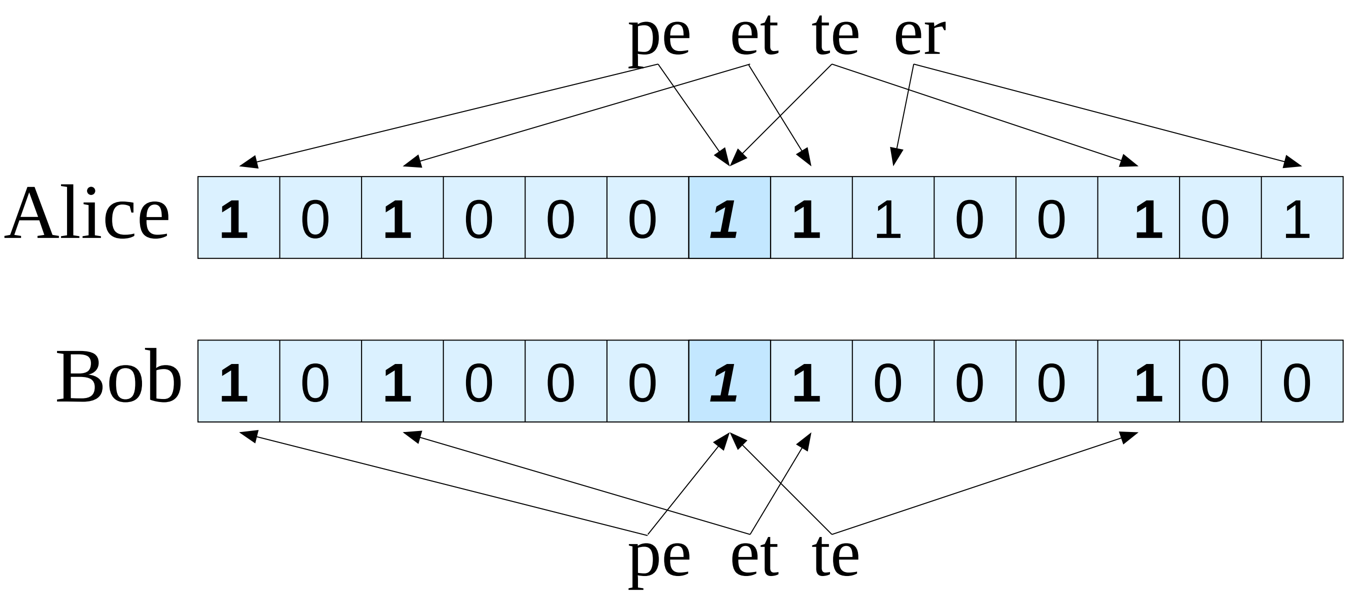


PPRL techniques: Bloom filter encoding (1)

- Proposed by Schnell et al. (Biomed Central, 2009)
- A Bloom filter is a bit-array, where a bit is set to 1 if a hash function $H_k(x)$ maps an element x of a set into this bit (elements in our case are q-grams)
  - $0 \leq H_k(x) < l$, with l the number of bits in Bloom filter
  - Many hash functions can be used (Schnell: $k = 30$)
  - Number of bits can be large (Schnell: $l = 1000$ bits)
- Basic idea: Map q-grams into Bloom filters using hash functions only known to database owners, send Bloom filters to a third party which calculates Dice coefficient (number of *1*-bits in Bloom filters)

THE AUSTRALIAN NATIONAL UNIVERSITY

PPRL techniques: Bloom filter encoding (2)



- *1*-bits for string ‘peter’: 7, *1*-bits for ‘pete’: 5, common *1*-bits: 5, therefore $Dice\_sim = 2 \times 5/(7+5)$= 10/12 = 0.83
- Collisions will effect the calculated similarity values
- Number of hash functions and length of Bloom filter need to be carefully chosen

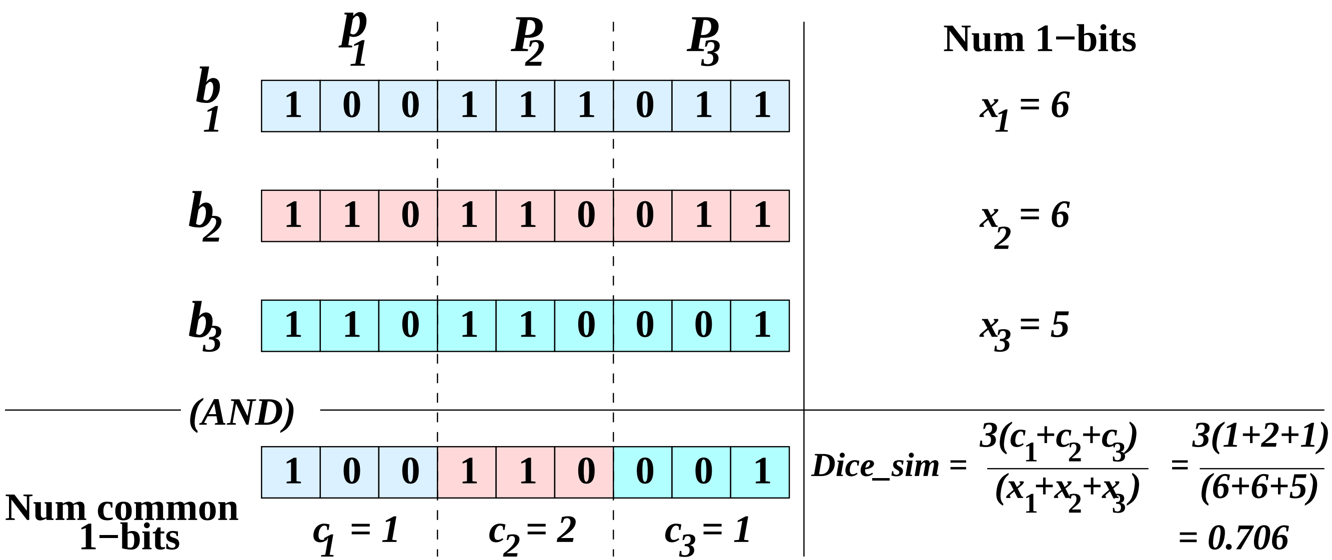
Multi-Party PPRL (1)

- Privacy-preserving linking of multiple databases (more than two sources)
- Example applications:
  - Health outbreak systems require data to be integrated across human health data, travel data, drug data, and animal health data
  - National security applications need to integrate data from law enforcement agencies, Internet service providers, businesses, and financial institutions
- Additional challenges:
  - Exponential complexity with number of sources
  - Increased privacy risk of collusion

Multi-party Bloom filter based PPRL

(Vatsalan and Christen, CIKM, 2014)

- Distribute similarity calculation across all parties:
  - Bloom filters are split into segments such that each party processes a segment to calculate the number of common *1*-bits in its segment
  - Secure summation is applied across parties to sum the number of common *1*-bits (c_i) and total *1*-bits (x_i) in their Bloom filters to calculate the similarity



Conclusions and research directions

To make sure everybody is awake.. :-)

Conclusions

- The linking of databases is challenged by **data quality**, **database size**, and **privacy concerns**
- When databases are linked across organisations, maintaining privacy and confidentiality is vital
- A variety of PPRL techniques has been developed in the past two decades
  - They allow approximate matching
  - Are scalable to medium–large databases
  - Work on static databases
- More research is needed to make PPRL practical for Big Data applications

Research directions (1)

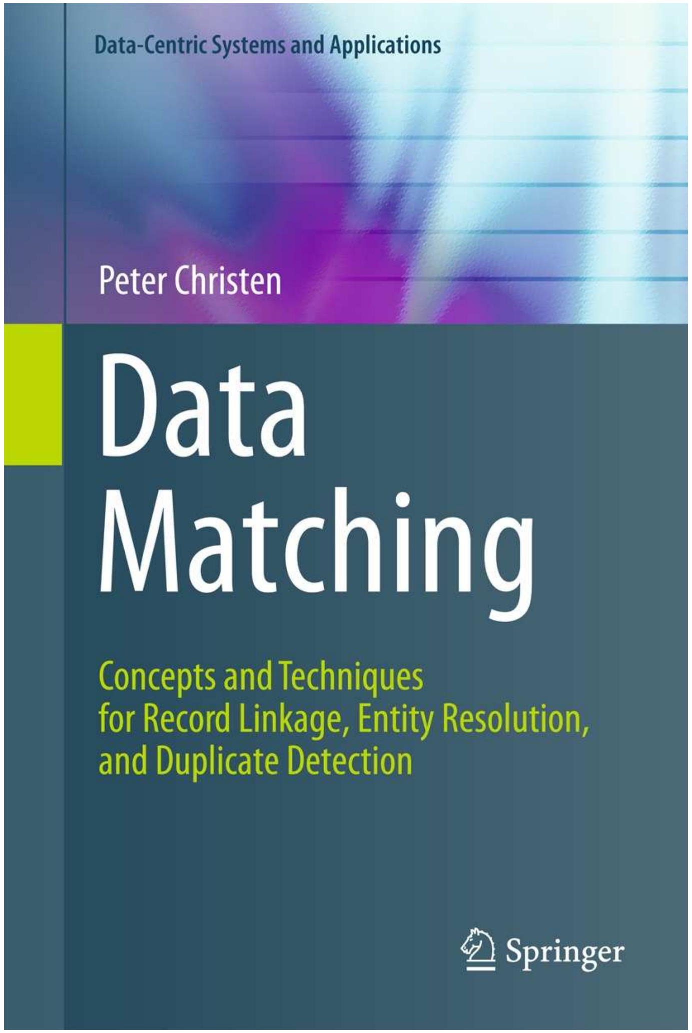
- Improved classification for PPRL
  - Mostly simple threshold-based classification is used
  - No investigation into advanced methods, such as collective / relational techniques
  - Supervised classification is difficult (no training data in many situations)
- Assessing linkage quality and completeness
  - How to assess linkage quality (precision and recall)?

    – How many classified matches are true matches?
    – How many true matches have we found?
  - Access to actual record values is not possible (as this would reveal sensitive information)

Research directions (2)

- A framework for PPRL is needed
  - To facilitate comparative experimental evaluation of PPRL techniques
  - Needs to allow researchers to plug-in their techniques
  - Benchmark data sets are required (biggest challenge, as such data are sensitive!)
- PPRL on multiple databases
  - Most work so far is limited to linking two databases (in practice, often records from several organisations need to be linked)
  - Pair-wise linking does not scale up
  - Preventing collusion between (sub-groups of) parties becomes more difficult

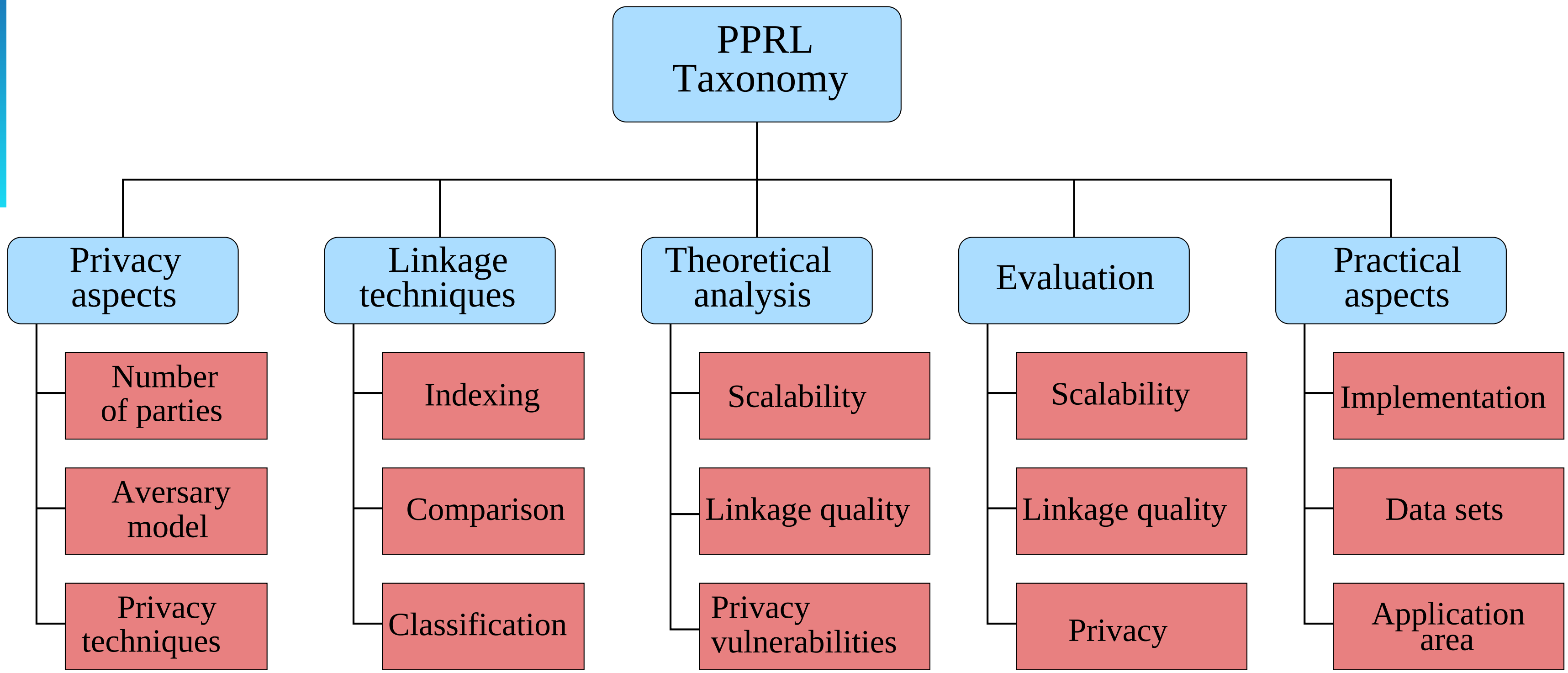
Advertisement: Book 'Data Matching' (2012)



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.

A taxonomy for PPRL



A taxonomy of privacy-preserving record linkage techniques
Dinusha Vatsalan, Peter Christen, and Vassilios Verykios
Elsevier Information Systems, 38(6), September 2013

References (1)

- Agrawal R, Evfimievski A, and Srikant R: *Information sharing across private databases.* ACM SIGMOD, San Diego, 2005.
- Al-Lawati A, Lee D and McDaniel P: *Blocking-aware private record linkage.* IQIS, Baltimore, 2005.
- Atallah MJ, Kerschbaum F and Du W: *Secure and private sequence comparisons.* WPES, Washington DC, pp. 39–44, 2003.
- Bachteler T, Schnell R, and Reiher J: *An empirical comparison of approaches to approximate string matching in private record linkage.* Statistics Canada Symposium, 2010.
- Barone D, Maurino A, Stella F, and Batini C: *A privacy-preserving framework for accuracy and completeness quality assessment.* Emerging Paradigms in Informatics, Systems and Communication, 2009.
- Bellare K, Curino C, Machanavajihala A, Mika P, Rahurkar M, and Sane A: *Woo: A scalable and multi-tenant platform for continuous knowledge base synthesis*. VLDB Endowment, 6(11), pp. 1114–1125, 2013.

ANU
THE AUSTRALIAN NATIONAL UNIVERSITY

References (2)

- Bhattacharya, I and Getoor, L: *Collective entity resolution in relational data.* ACM TKDD, 2007.
- Blakely T, Woodward A and Salmond C: *Anonymous linkage of New Zealand mortality and census data.* ANZ Journal of Public Health, 24(1), 2000.
- Bloom, BH: *Space/time trade-offs in hash coding with allowable errors.* Communications of the ACM, 1970.
- Bonomi L, Xiong Li, Chen R, and Fung B: *Frequent grams based embedding for privacy preserving record linkage.* ACM Information and knowledge management, 2012.
- Bouzelat H, Quantin C, and Dusserre L: *Extraction and anonymity protocol of medical file.* AMIA Fall Symposium, 1996.
- Chaytor R, Brown E and Wareham T: *Privacy advisors for personal information management.* SIGIR workshop on Personal Information Management, Seattle, pp. 28–31, 2006.
- Christen P: *Privacy-preserving data linkage and geocoding: Current approaches and research directions.* PADM held at IEEE ICDM, Hong Kong, 2006.

THE AUSTRALIAN NATIONAL UNIVERSITY

References (3)

- Christen P: *Geocode Matching and Privacy Preservation.* ACM PinKDD, 2009.
- Christen, P: *A survey of indexing techniques for scalable record linkage and deduplication.* IEEE TKDE, 2012.
- Christen, P: *Data matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection.* Springer, 2012.
- Christen, P: *Preparation of a real voter data set for record linkage and duplicate detection research.* Technical Report, The Australian National University, 2014.
- Christen P and Churches T: *Secure health data linkage and geocoding: Current approaches and research directions.* ehPASS, Brisbane, 2006.
- Christen, P and Goiser, K: *Quality and complexity measures for data linkage and deduplication.* In *Quality Measures in Data Mining.* Springer Studies in Computational Intelligence, vol. 43, 2007.
- Christen P, Vatsalan D, and Verykios V: *Challenges for privacy preservation in data integration.* In *Journal of Data and Information Quality.* ACM, vol. 5, 2014.

References (4)

- Churches T: *A proposed architecture and method of operation for improving the protection of privacy and confidentiality in disease registers.* BMC Medical Research Methodology, 3(1), 2003.
- Churches T and Christen P: *Some methods for blindfolded record linkage.* BMC Medical Informatics and Decision Making, 4(9), 2004.
- Clifton C, Kantarcioglu M, Vaidya J, Lin X, and Zhu MY: *Tools for privacy preserving distributed data mining.* ACM SIGKDD Explorations, 2002.
- Clifton C, Kantarcioglu M, Doan A, Schadow G, Vaidya J, Elmagarmid AK and Suciu D: *Privacy-preserving data integration and sharing.* SIGMOD workshop on Research Issues in Data Mining and Knowledge Discovery, Paris, 2004.
- Dinur I and Nissim K: *Revealing information while preserving privacy.* ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, California, 2003.
- Du W, Atallah MJ, and Kerschbaum F: *Protocols for secure remote database access with approximate matching.* ACM Workshop on Security and Privacy in E-Commerce, 2000.

References (5)

- Dusserre L, Quantin C and Bouzelat H: *A one way public key cryptosystem for the linkage of nominal files in epidemiological studies.* Medinfo, 8:644-7, 1995.
- Durham, EA: *A framework for accurate, efficient private record linkage.* PhD Thesis, Vanderbilt University, 2012.
- Durham, EA, Toth C, Kuzu, M. Kantarcioglu M, and Malin B: *Composite Bloom for secure record linkage.* IEEE Transactions on Knowledge and Data Engineering, 2013.
- Durham, EA, Xue Y, Kantarcioglu M, and Malin B: *Private medical record linkage with approximate matching.* AMIA Annual Symposium, 2010.
- Durham, EA, Xue Y, Kantarcioglu M, and Malin B: *Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage.* Information Fusion, 2012.
- Dwork, C: *Differential privacy.* International Colloquium on Automata, Languages and Programming, 2006.
- Elmagarmid AK, Ipeirotis PG and Verykios VS: *Duplicate record detection: A survey.* IEEE TKDE 19(1), pp. 1–16, 2007.

References (6)

- Fienberg SE: *Privacy and confidentiality in an e-Commerce World: Data mining, data warehousing, matching and disclosure limitation.* Statistical Science, IMS Institute of Mathematical Statistics, 21(2), pp. 143–154, 2006.
- Hall R and Fienberg SE: *Privacy-preserving record linkage.* Privacy in Statistical Databases, Springer LNCS 6344, 2010.
- Herzog TN, Scheuren F, and Winkler WE: *Data quality and record linkage techniques.* Springer, 2007.
- Holman et al.: *A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system.* CSIRO Australian Health Review, 32(4), pp. 766–777, 2008.
- Ibrahim A, Jin H, Yassin AA, and Zou D: *Approximate Keyword-based Search over Encrypted Cloud Data.* IEEE ICEBE, pp. 238–245, 2012.
- Inan A, Kantarcioglu M, Bertino E and Scannapieco M: *A hybrid approach to private record linkage.* IEEE ICDE, Cancun, Mexico, pp. 496–505, 2008.
- Inan A, Kantarcioglu M, Ghinita G, and Bertino E: *Private record matching using differential privacy.* EDBT, 2010.

ANU
THE AUSTRALIAN NATIONAL UNIVERSITY

References (7)

- Kang H, Getoor L, Shneiderman B, Bilgic M, and Licamele L: *Interactive entity resolution in relational data: A visual analytic tool and its evaluation.* IEEE Transactions on Visualization and Computer Graphics, 14(5), pp. 999–1014, 2008.
- Kantarcioglu M, Jiang W, and Malin B: *A privacy-preserving framework for integrating person-specific databases.* Privacy in Statistical Databases, 2008.
- Kantarcioglu M, Inan A, Jiang W and Malin B: *Formal anonymity models for efficient privacy-preserving joins.* Data and Knowledge Engineering, 2009.
- Karakasidis A and Verykios VS: *Privacy preserving record linkage using phonetic codes.* IEEE Balkan Conference in Informatics, 2009.
- Karakasidis A and Verykios VS: *Advances in privacy preserving record linkage.* E-activity and Innovative Technology, Advances in Applied Intelligence Technologies Book Series, IGI Global, 2010.
- Karakasidis A and Verykios VS: *Secure blocking+secure matching = Secure record linkage.* Journal of Computing Science and Engineering, 2011.

THE AUSTRALIAN NATIONAL UNIVERSITY

References (8)

- Karakasidis A, Verykios VS, and Christen P: *Fake injection strategies for private phonetic matching.* International Workshop on Data Privacy Management, 2011.
- Karakasidis A and Verykios VS: *Reference table based k-anonymous private blocking.* Symposium on Applied Computing, 2012.
- Karakasidis A and Verykios VS: *A sorted neighborhood Approach to multidimensional privacy preserving blocking.* IEEE ICDM workshop, 2012.
- Karapiperis D and Verykios VS: *A distributed framework for scaling Up LSH-based computations in privacy preserving record linkage.* Balkan Conference in Informatics, 2013.
- Kelman CW, Bass AJ and Holman CDJ: *Research use of linked health data – A best practice protocol.* ANZ Journal of Public Health, 26(3), pp. 251–255, 2002.
- Kum, HC, Duncan DF and Stewart CJ: *Supporting self-evaluation in local government via Knowledge Discovery and Data mining.* Government Information Quarterly, 26(2), pp. 295-304, 2009.

ANU
THE AUSTRALIAN NATIONAL UNIVERSITY

References (9)

- Kum HC and Ahalt S: *Privacy by design: understanding data access models for secondary data.* AMIA Joint Summits on Translation Science and Clinical Research Informatics, 2013.
- Kum HC, Krishnamurthy A, Machanavajjhala A, and Ahalt S: *Social genome: Putting big data to work for population informatics.* IEEE Computer, 2014.
- Kum HC, Ahalt S, and Pathak D.: *Privacy-Preserving Data Integration Using Decoupled Data.* Security and Privacy in Social Networks, Springer, pp. 225-253, 2013.
- Kum HC, Krishnamurthy A, Pathak D, Reiter M, and Ahalt S: *Secure Decoupled Linkage (SDLink) system for building a social genome.* IEEE International Conference on BigData, 2013.
- Kum HC, Krishnamurthy A, Machanavajjhala A, Reiter MK, and Ahalt S: *Privacy preserving interactive record linkage (PPIRL).* Journal of the American Medical Informatics Association, 21(2), pp. 212–220, 2014.

THE AUSTRALIAN NATIONAL UNIVERSITY

References (10)

- Kuzu M, Kantarcioglu M, Durham EA and Malin B: *A constraint satisfaction cryptanalysis of Bloom filters in private record linkage.* Privacy Enhancing Technologies, 2011.
- Kuzu M, Kantarcioglu M, Inan A, Bertino E, Durham EA and Malin B: *Efficient privacy-aware record integration.* ACM Extending Database Technology, 2013.
- Kuzu M, Kantarcioglu M, Durham EA, Toth C, and Malin B: *A practical approach to achieve private medical record linkage in light of public resources.* Journal of the American Medical Informatics Association, vol. 20, pp. 285–292, 2013.
- Lai PK, Yiu SM, Chow KP, Chong CF, and Hui LC: *An efficient Bloom filter based solution for multiparty private matching.* International Conference on Security and Management, 2006.
- Li Y, Tygar JD and Hellerstein JM: *Private matching.* Computer Security in the 21st Century, Lee DT, Shieh SP and Tygar JD (editors), Springer, 2005.
- Li F, Chen Y, Luo B, Lee D, and Liu P: *Privacy preserving group linkage.* Scientific and Statistical Database Management, 2011.

THE AUSTRALIAN NATIONAL UNIVERSITY

References (11)

- Lyons R et al.: *The SAIL databank: linking multiple health and social care datasets.* BMC Medical Informatics and Decision Making, 9(1), 2009.
- Malin B, Airoldi E, Edoho-Eket S and Li Y: *Configurable security protocols for multi-party data analysis with malicious participants.* IEEE ICDE, Tokyo, pp. 533–544, 2005.
- Malin B and Sweeney L: *A secure protocol to distribute unlinkable health data.* American Medical Informatics Association, Washington DC, pp. 485–489, 2005.
- Mohammed N, Fung BC and Debbabi M: *Anonymity meets game theory: secure data integration with malicious participants.* VLDB Journal, 2011.
- Murugesan M, Jiang W, Clifton C, Si L and Vaidya J: *Efficient privacy-preserving similar document detection.* VLDB Journal, 2010.
- Naumann F and Herschel M: *An introduction to duplicate detection.* Synthesis Lectures on Data Management, Morgan and Claypool Publishers, 2010.
- Navarro-Arribas G and Torra V: *Information fusion in data privacy: A survey.* Information fusion, 2012.

References (12)

- O'Keefe CM, Yung M, Gu L and Baxter R: *Privacy-preserving data linkage protocols.* WPES, Washington DC, pp. 94–102, 2004.
- Pang C, Gu L, Hansen D and Maeder A: *Privacy-preserving fuzzy matching using a public reference table.* Intelligent Patient Management, 2009.
- Quantin C, Bouzelat H and Dusserre L: *Irreversible encryption method by generation of polynomials.* Medical Informatics and The Internet in Medicine, Informa Healthcare, 21(2), pp. 113–121, 1996.
- Quantin C, Bouzelat H, Allaert FAA, Benhamiche AM, Faivre J and Dusserre L: *How to ensure data quality of an epidemiological follow-up: Quality assessment of an anonymous record linkage procedure.* International Journal of Medical Informatics, 49, pp. 117–122, 1998.
- Quantin C, Bouzelat H, Allaert FAA, Benhamiche AM, Faivre J and Dusserre L: *Automatic record hash coding and linkage for epidemiological follow-up data confidentiality.* Methods of Information in Medicine, Schattauer, 37(3), pp. 271–277, 1998.

References (13)

- Ravikumar P, Cohen WW and Fienberg SE: *A secure protocol for computing string distance metrics.* PSDM held at IEEE ICDM, Brighton, UK, 2004.
- Scannapieco M, Figotin I, Bertino E and Elmagarmid AK: *Privacy preserving schema and data matching.* ACM SIGMOD, 2007.
- Schadow G, Grannis SJ and McDonald CJ: *Discussion paper: Privacy-preserving distributed queries for a clinical case research network.* CRPIT'14: Proceedings of the IEEE international Conference on Privacy, Security and Data Mining, Maebashi City, Japan, pp. 55–65, 2002.
- Schnell R, Bachteler T and Reiher J: *Privacy-preserving record linkage using Bloom filters.* BMC Medical Informatics and Decision Making, 9(1), 2009.
- Schnell R, Bachteler T and Reiher J: *A novel error-tolerant anonymous linking code.* German record linkage center working paper series, 2011.
- Schnell R: *Privacy-preserving record linkage and privacy-preserving blocking for large files with cryptographic keys using multibit trees.* ASA JSM Proceedings, Alexandria, VA, 2013.

References (14)

- Sweeney L: *Privacy-enhanced linking.* ACM SIGKDD Explorations, 7(2), 2005.
- Trepetin S: *Privacy-preserving string comparisons in record linkage systems: a review.* Information Security Journal: A Global Perspective, 2008.
- Vatsalan D, Christen P and Verykios VS: *An efficient two-party protocol for approximate matching in private record linkage.* AusDM, CRPIT, 2011.
- Vatsalan D and Christen P: *An iterative two-party protocol for scalable privacy-preserving record linkage.* AusDM, CRPIT, vol. 134, 2012.
- 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.* CIKM, 2013.
- Vaidya J and Clifton C: *Secure set intersection cardinality with application to association rule mining.* Journal of Computer Security, 2005.

References (15)

- Verykios VS, Karakasidis A and Mitrogiannis VK: *Privacy preserving record linkage approaches.* International Journal of Data Mining, Modelling and Management, 2009.
- Weber SC, Lowe H, Das A and Ferris T: *A simple heuristic for blindfolded record linkage.* Journal of the American Medical Informatics Association, 2012.
- Weitzner D.J et al.: *Information accountability.* ACM Communications, 51(6), pp. 82–87, 2008.
- Winkler WE: *Overview of record linkage and current research directions.* RR 2006/02, US Census Bureau, 2006.
- Yakout M, Atallah MJ and Elmagarmid AK: *Efficient private record linkage.* IEEE ICDE, 2009.
- Yao, AC: *How to generate and exchange secrets.* Annual Symposium on Foundations of Computer Science, 1986.
- Zhang Q and Hansen D: *Approximate processing for medical record linking and multidatabase analysis.* International Journal of Healthcare Information Systems and Informatics, 2(4), pp. 59–72, 2007.

ANU
THE AUSTRALIAN NATIONAL UNIVERSITY