Data Linkage: Introduction, Recent Advances, and Privacy Issues

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 and DP160101934; and the ARC, Veda, and Funnelback Pty. Ltd., under Linkage Project LP100200079.

Motivation

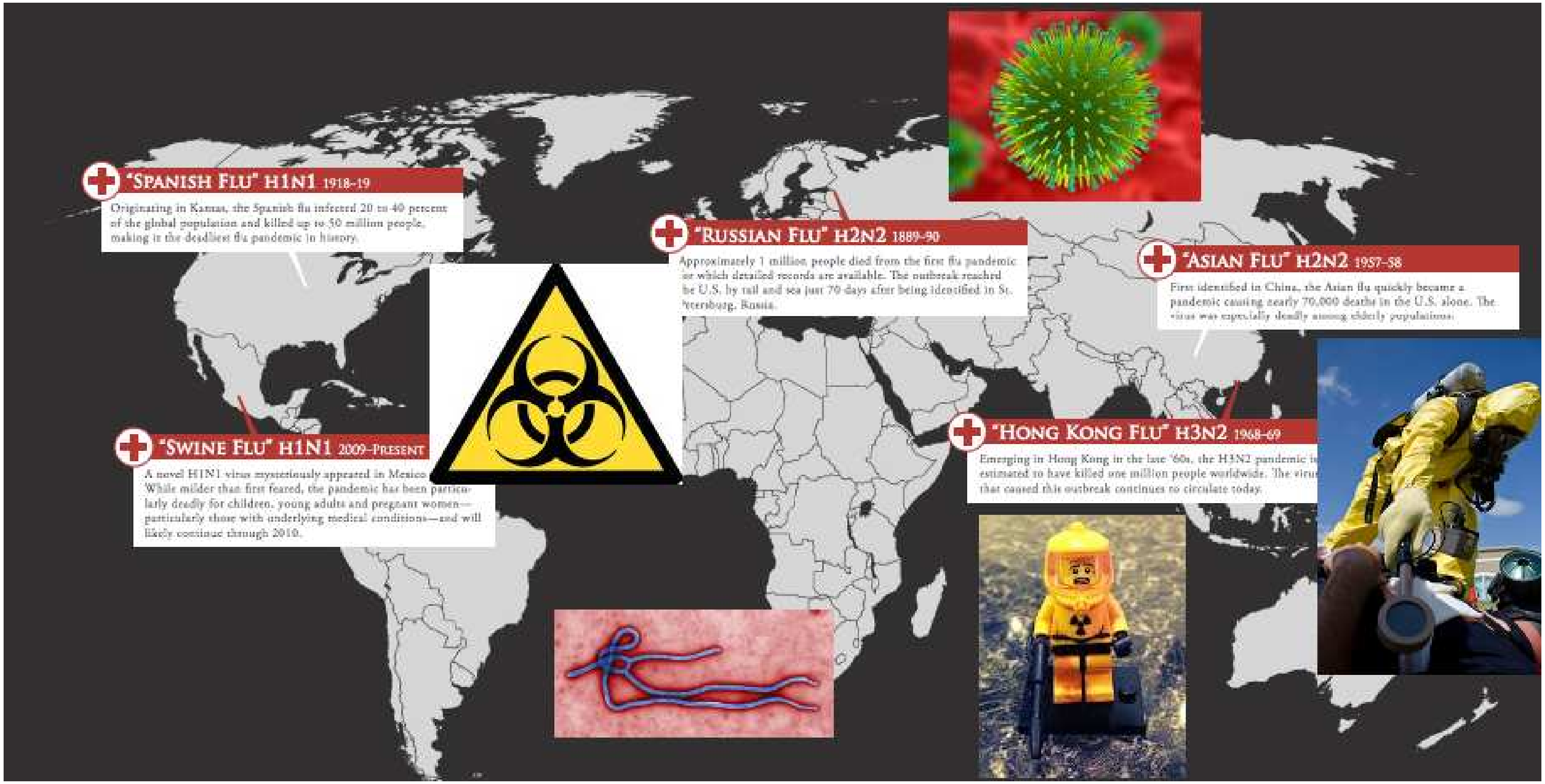
- Large amounts of data are being collected both by organisations in the private and public sectors, as well as by researchers and individuals
- Much of these data are about people, or they are generated by people
  - Financial, shopping, and travel transactions
  - Electronic health records
  - Tax, social security, and census records
  - Location records
  - Emails, tweets, SMSs, Facebook posts, etc.
- Analysing such data can provide huge benefits to businesses, governments and researchers

THE AUSTRALIAN NATIONAL UNIVERSITY

Motivation (continued)

- Often data from different sources need to be integrated and linked
  - To allow data analyses that are impossible on individual databases
  - To improve data quality
  - To enrich data with additional information
- Lack of unique *entity identifiers* means that linking is often 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)



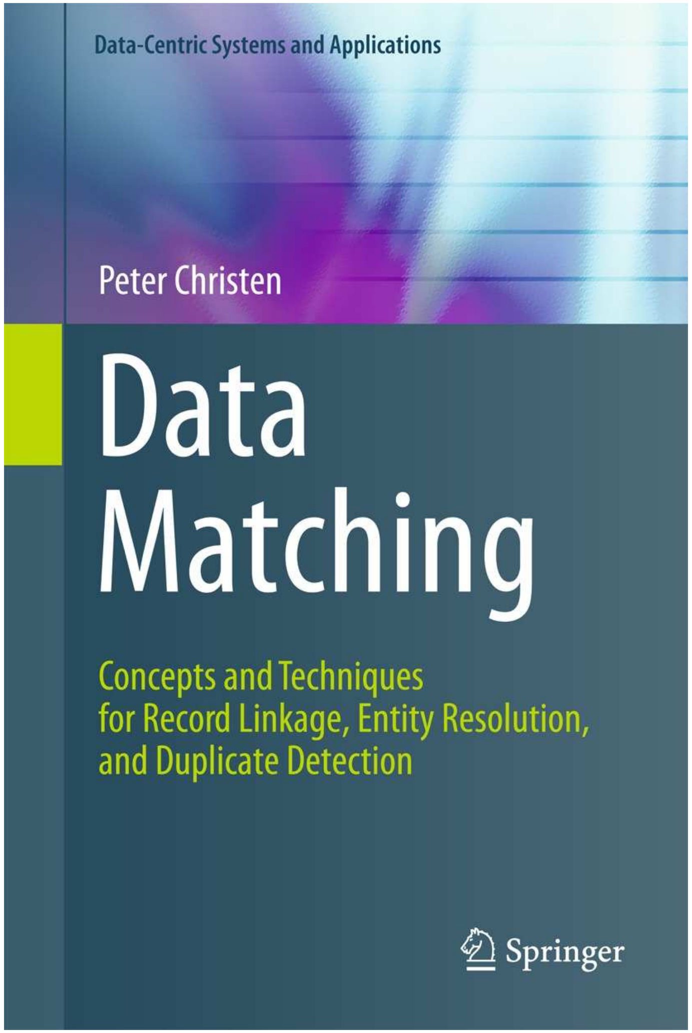
Motivating example: Health surveillance (2)

- 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 **large**, **dynamic**, **complex**, **heterogeneous** and **distributed**, and they require **linking** and **analysis** in near **real time**

Objective of this tutorial

- Provide an understanding of data / record linkage applications, challenges, and techniques
- Understand the data linkage process, and key techniques employed in each step of this process
- Have a basic understanding of advanced techniques for scalable indexing and machine-learning based classification for data linkage
- Appreciate the privacy and confidentiality challenges that data linkage poses
- Have a basic understanding of privacy-preserving record linkage

Content is loosely based on 'Data Matching' (Springer, 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.

Outline

- Part 1: Introduction
  - Applications, history, challenges, and examples
- Part 2: Data linkage process
  - Key techniques used in data linkage
- Part 3: Advanced data linkage techniques
  - Indexing and blocking for scalable data linkage
  - Learning, collective, and graph based techniques
- Part 4: Privacy aspects in data linkage
  - Motivating scenario
  - Privacy-preserving record linkage
- Conclusions and research directions

What is data 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 *record 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*

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

Recent interest in data linkage

- Traditionally, data linkage has been used in statistics (census) and health (epidemiology)
  - First computer based techniques developed in 1960s
- In recent years, increased interest from businesses and governments
  - Massive amounts of data are being collected, and increased computing power and storage capacities
  - 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

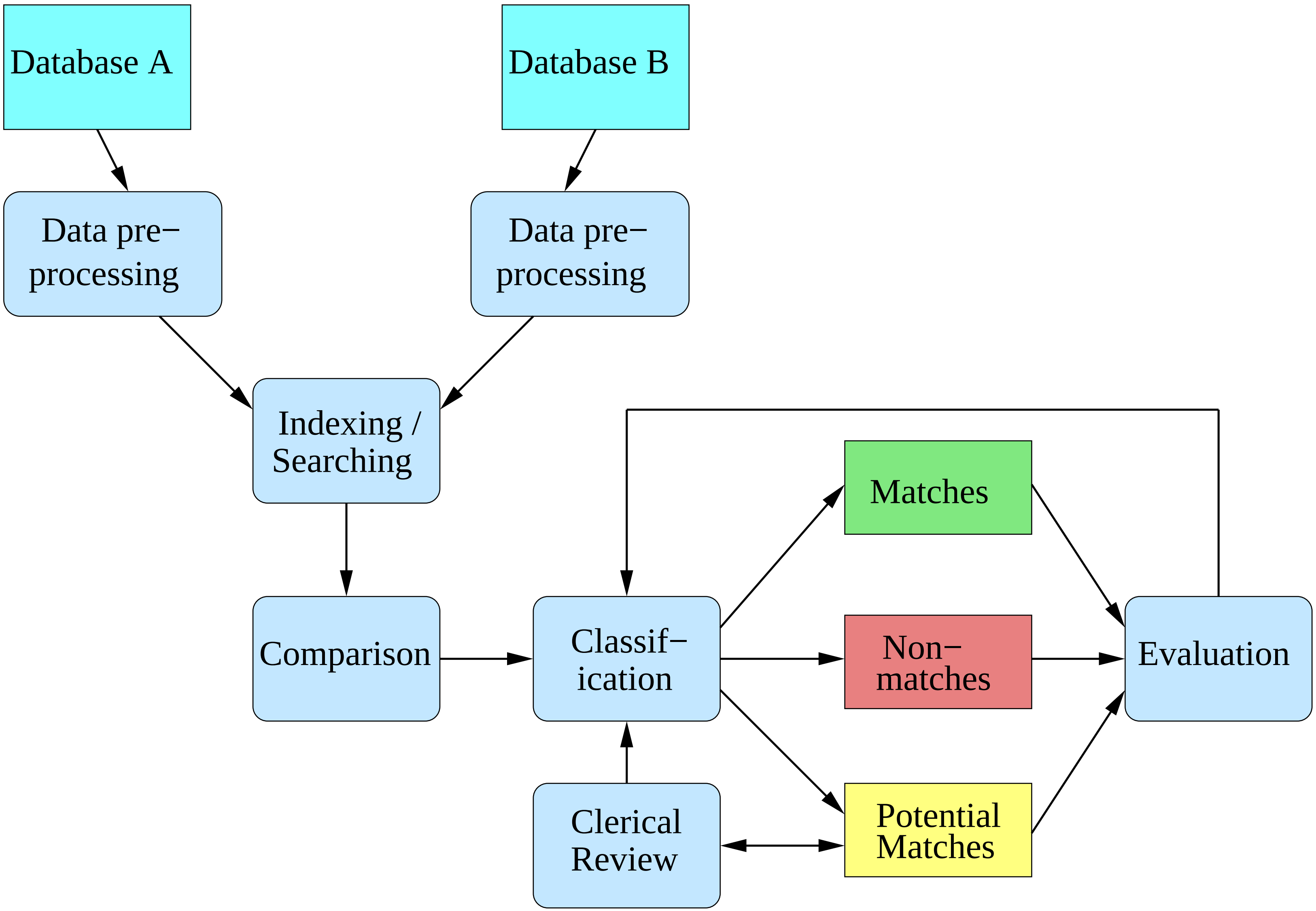
A brief history of data linkage (1)

- Computer assisted data linkage goes back as far as the 1950s (based on ad-hoc heuristic methods)
- Basic ideas of probabilistic linkage were introduced by *Newcombe & Kennedy* (1962)
- Theoretical foundation by *Fellegi & Sunter* (1969)
  - Compare common record attributes (or fields)
  - Compute matching weights based on frequency ratios (global or value specific) and error estimates
  - Sum of the matching weights is used to classify a pair of records as a *match*, *non-match*, or *potential match*
  - Problems: Estimating errors and thresholds, assumption of independence, and *clerical review*

A short history of data linkage (2)

- Strong interest in the last decade from computer science (from many research fields, including data mining, AI, knowledge engineering, information retrieval, information systems, databases, and digital libraries)
- Many different techniques have been developed
- Major focus has been on scalability to large databases, and linkage quality
  - Various indexing/blocking techniques to efficiently and effectively generate candidate record pairs
  - Various machine learning-based classification techniques, both supervised and unsupervised, as well as active learning based

The data linkage process



Data 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)
  - Unsupervised: Clustering, collective, and graph based

Major data 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, are commonly required for linking)

Example 1: Web of Object (WOO)

(based on slides by Hye-Chung Kum, Texas A&M)

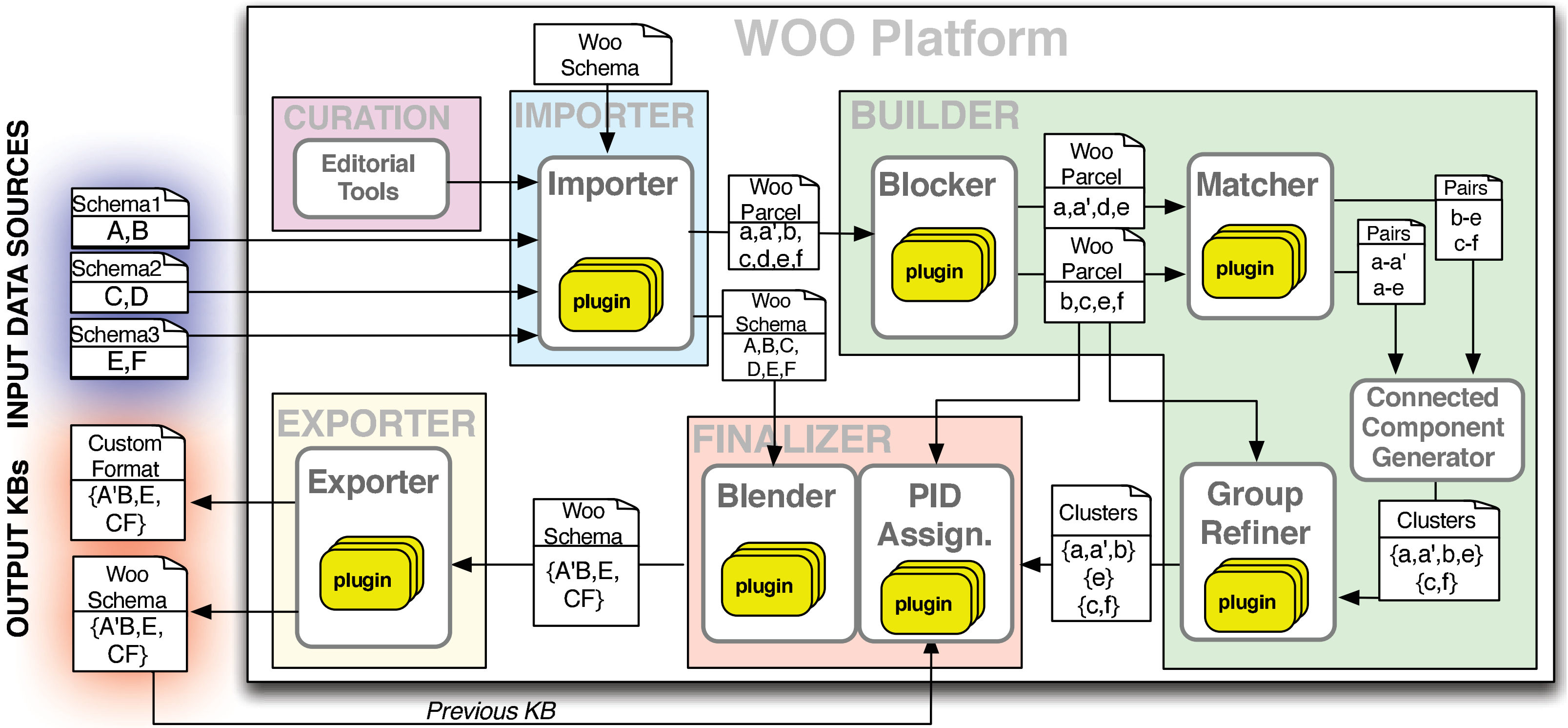
- **Goal**: To enable various products in Yahoo! to synthesise knowledge-bases of entities relevant to their domains (Bellare et al., VLDB, 2013)
- Desiderata:
  - *Coverage*: the fraction of real-world entities
  - *Accuracy*: information must be accurate
  - *Linkage*: the level of connectivity of entities
  - *Identifiability*: one and only one identifier for a real-world entity
  - *Persistence/content continuity*: variants of the same entity across time must be linked
  - *Multi-tenant*: be useful to multiple portals

WOO: Knowledge base synthesis

- **Knowledge base synthesis** is the process of ingestion, disambiguation, and enrichment of entities from a variety of structured and unstructured data sources
  - Sheer scale of the data
    $\Rightarrow$ Hundreds of millions of entities daily
  - Diverse domains
    $\Rightarrow$ From hundreds of data sources
  - Diverse requirements
    $\Rightarrow$ Multiple tenants, such as Locals, Movies, Deals, and Events in (for example) the Yahoo! website

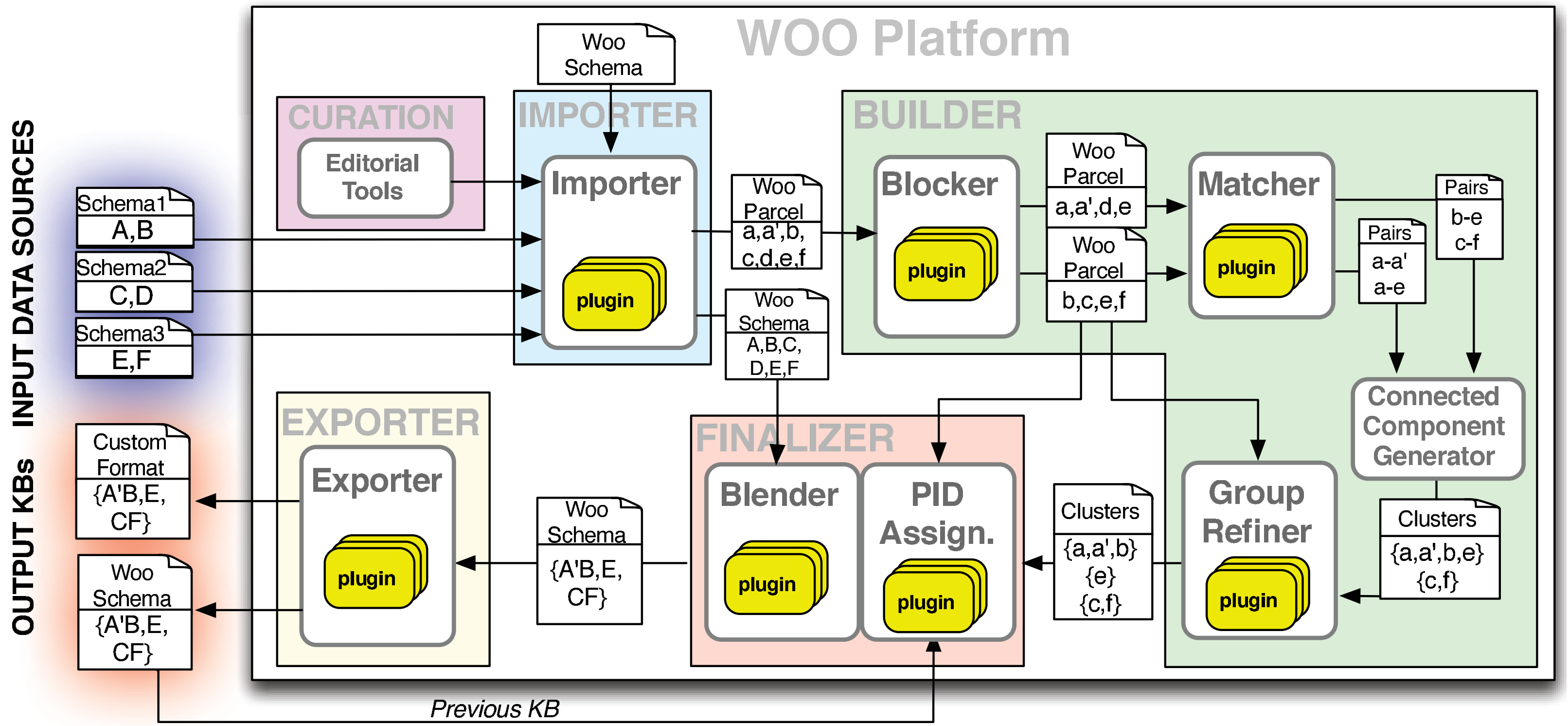
ANU THE AUSTRALIAN NATIONAL UNIVERSITY

The WOO architecture (1)



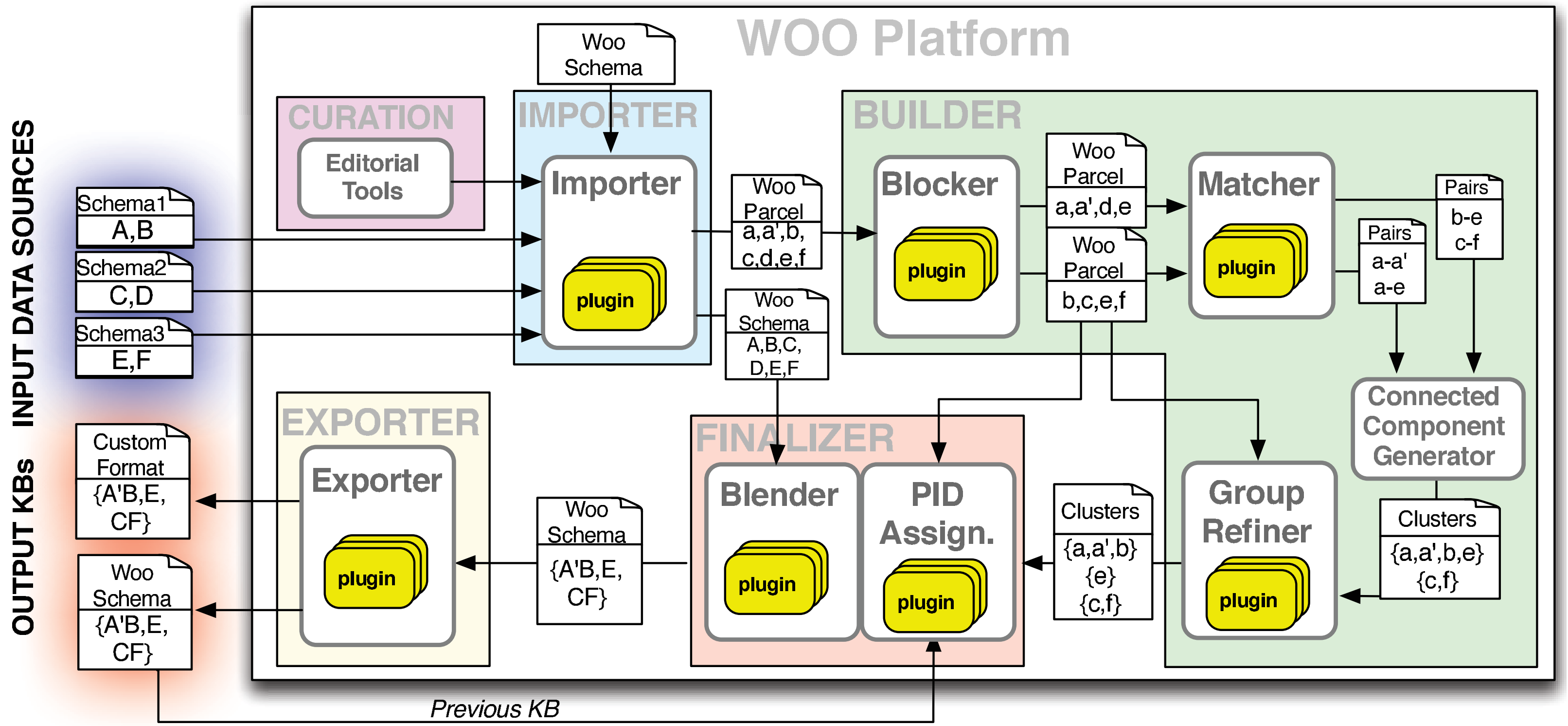
Source: Bellare et al., VLDB, 2013

The WOO architecture (2)



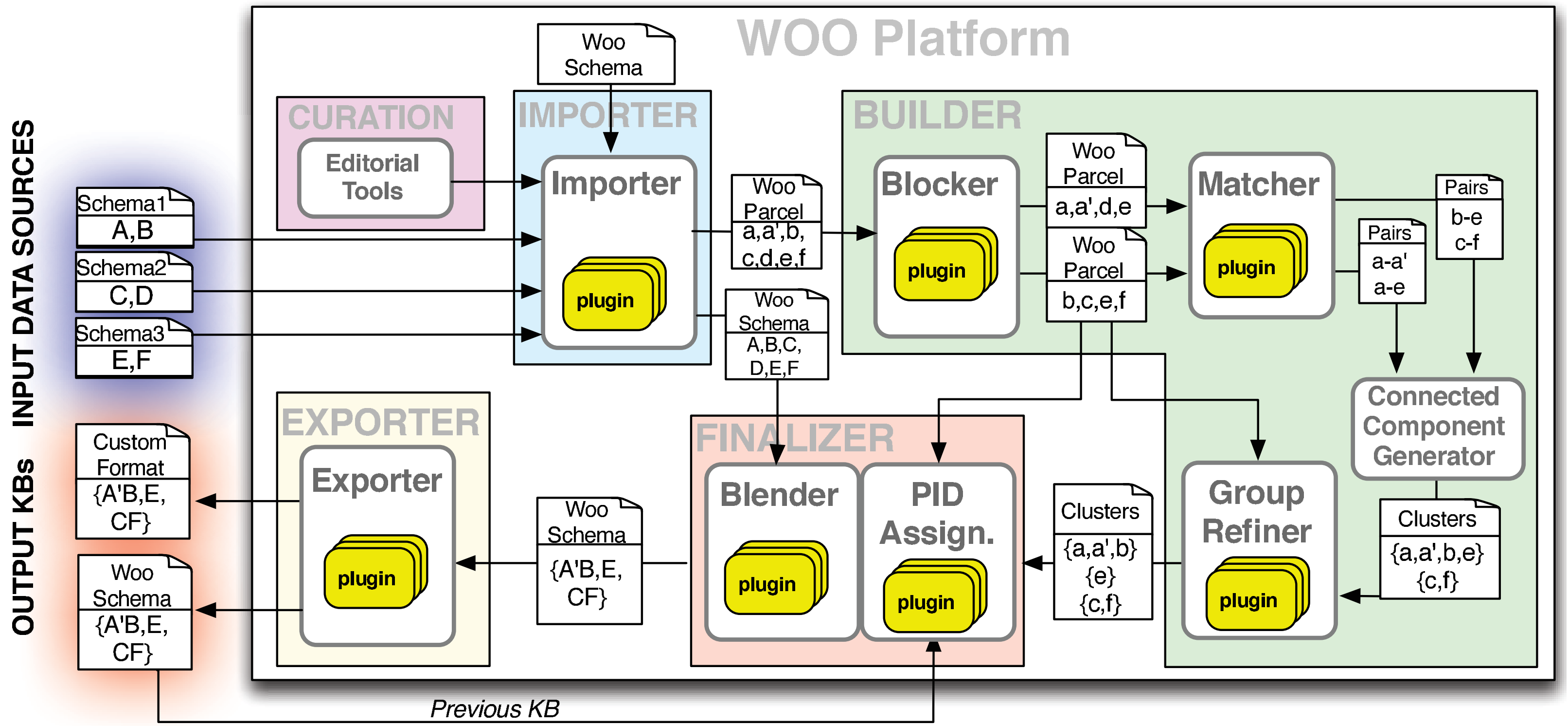
- **Importer** takes a collection of data sources as input (like XML feeds, RDF content, Relational Databases, or other custom formats)

The WOO architecture (3)



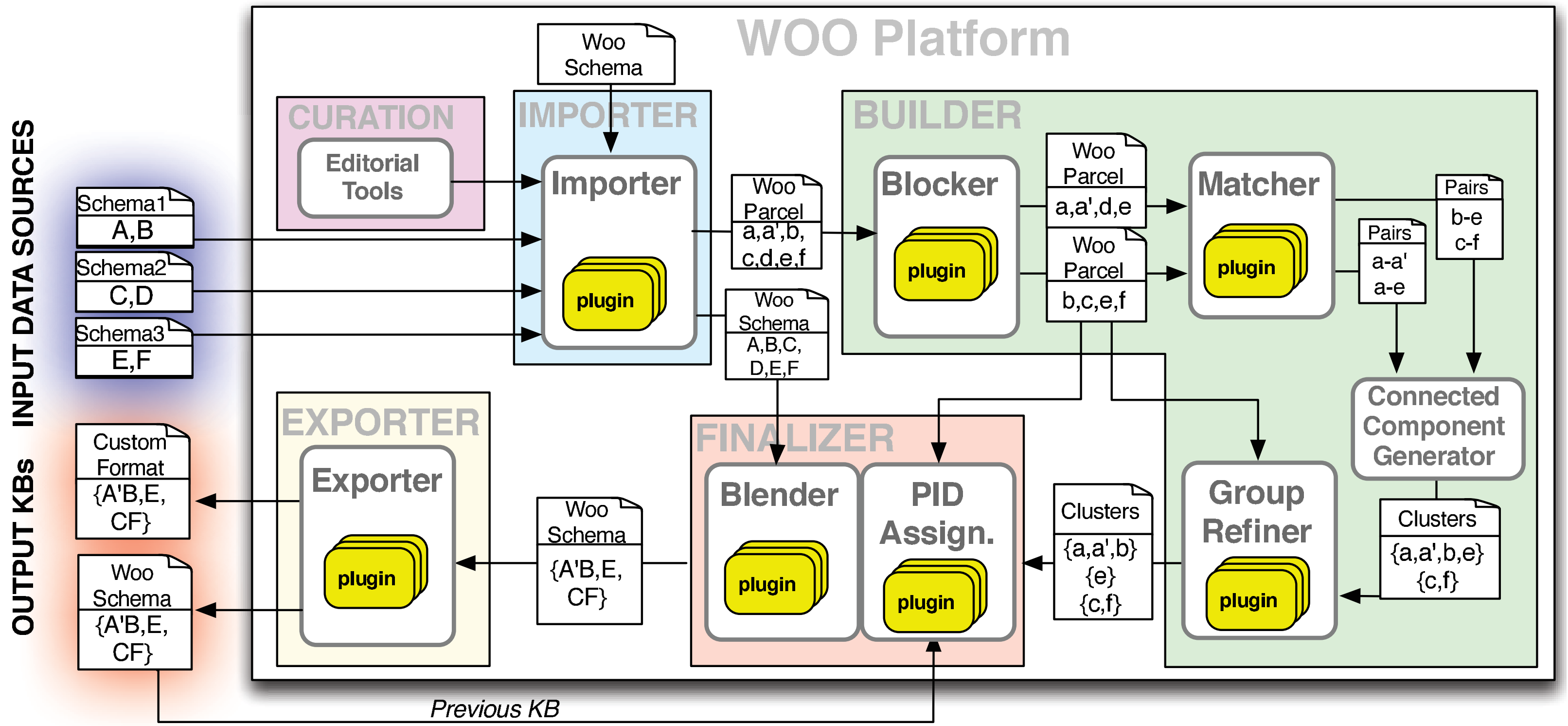
- Each data source is converted into a common format called *the WOO schema*
- The *WOO Parcel*, containing only the attributes needed for matching, is pushed to the **Builder**

The WOO architecture (4)



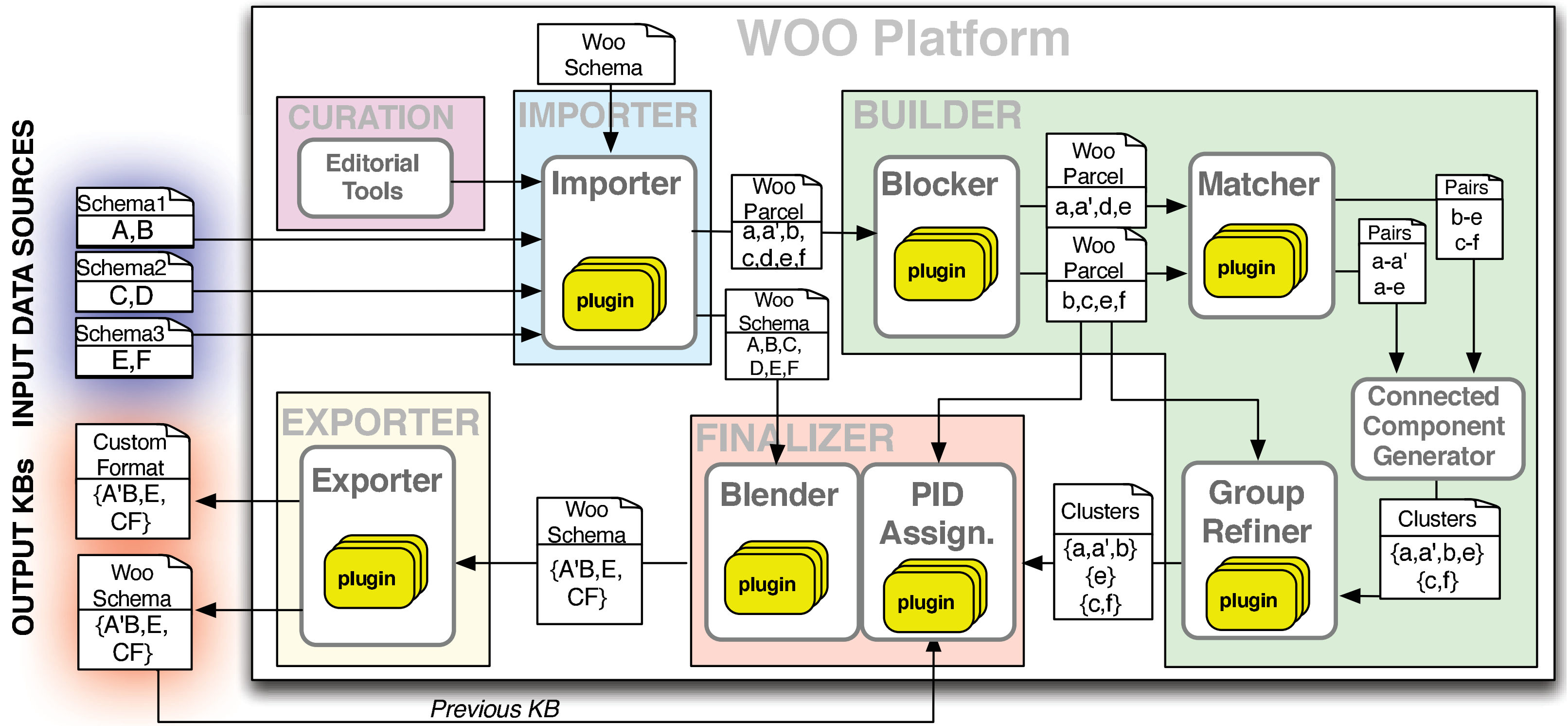
- **Builder** performs the entity deduplication and produces a clustering decision, including (1) *blocker*, (2) *matcher*, (3) *connected component generator*, and (4) *group refiner*

The WOO architecture (5)



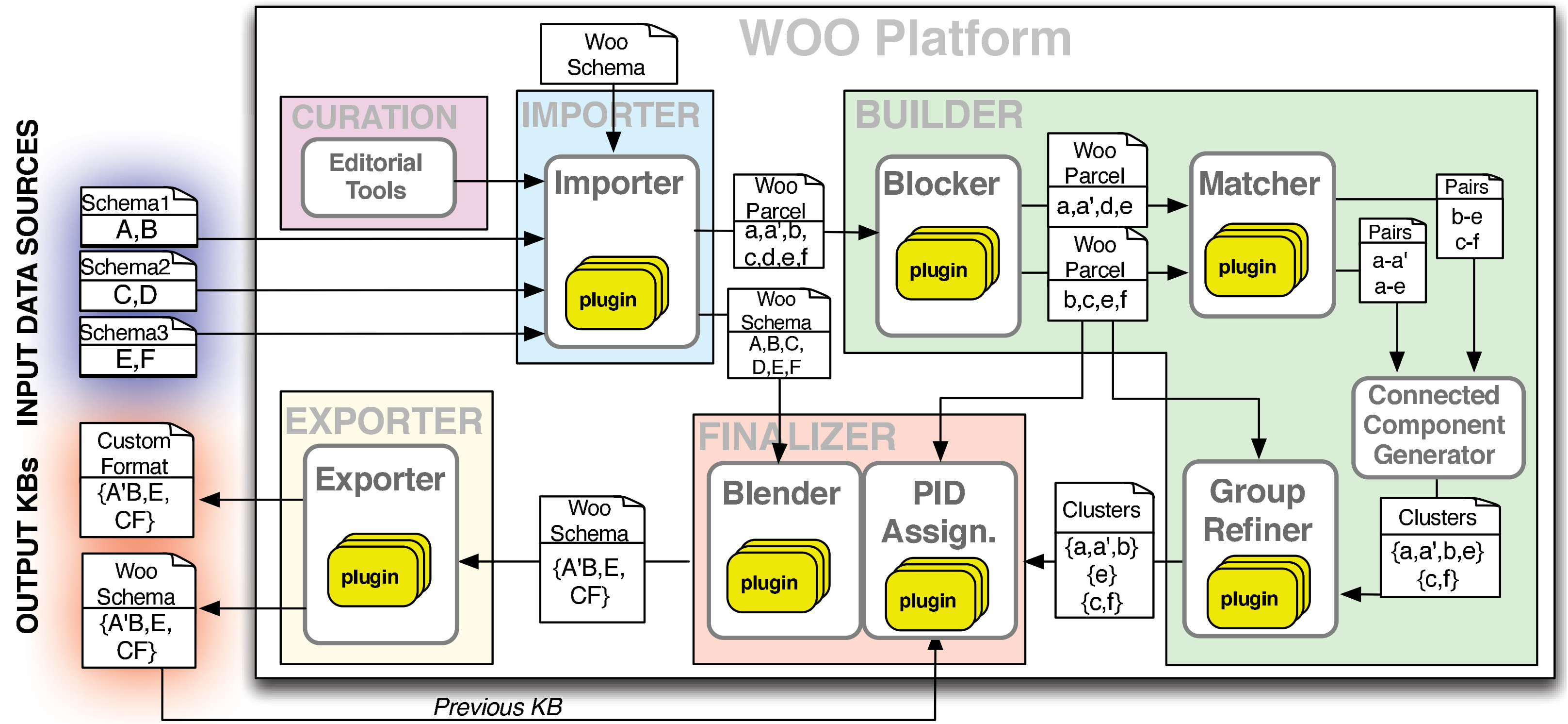
- **Finaliser** is responsible for handling the persistence of object identifiers and the blending of the attributes of the (potentially many) entities that are being merged

The WOO architecture (6)



- **Exporter** generates a fully integrated and de-duplicated knowledge-base, both in a format consistent with the WOO schema and in any custom format

The WOO architecture (7)



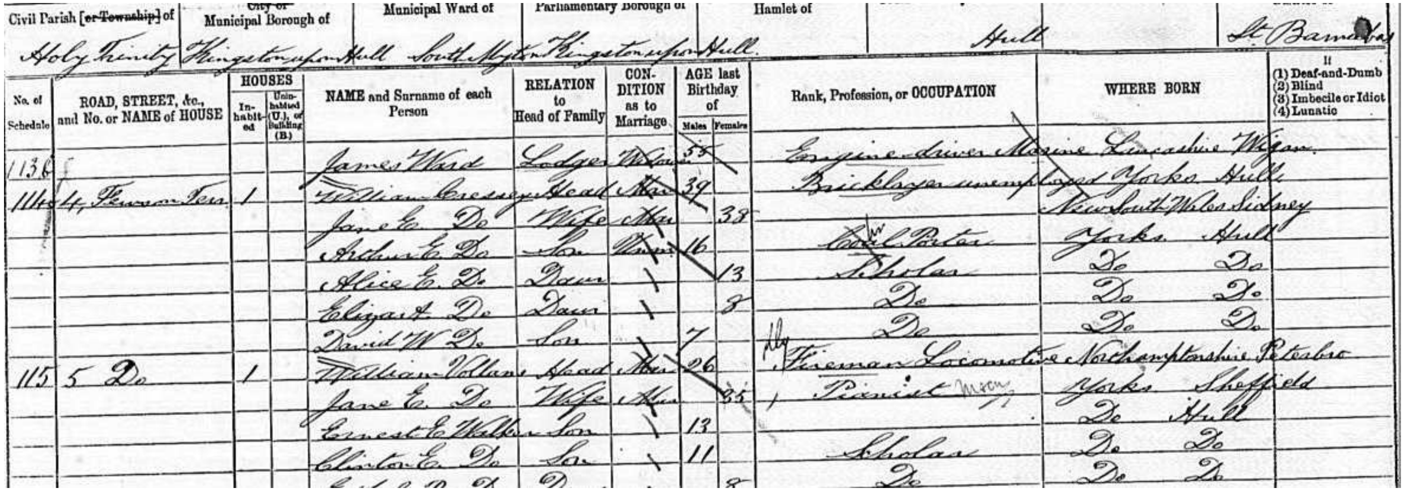
- **Curation** enables domain experts to influence the system behaviour through a set of GUIs, such as: forcing or disallowing certain matches between entities, or by editing attribute values

THE AUSTRALIAN NATIONAL UNIVERSITY

Example 2: Linking 'big' social science data

- Increasing use of large databases in social science research
- Often the aim is to create '*social genomes*' for individuals by linking population databases (*Population Informatics*, Kum et al. IEEE Computer, 2013)
- Knowing how individuals and families change over time allows for a diverse range of studies (fertility, employment, education, health, crime, etc.)
- Different challenges for historical data compared to contemporary data, but some are common
  - Database sizes (computational aspects)
  - Accurate match classification (data quality)

Challenges for historical 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 In-habited	Unin-habited (U), or Building (B)	NAME and Surname of each Person	RELATION to Head of Family	CON-DITION as to Marriage	AGE last Birthday of Males	Females	Rank, Profession, or OCCUPATION	WHERE BORN	If (1) Deaf-and-Dumb (2) Blind (3) Imbecile or Idiot (4) Lunatic
136				James Ward	Lodger	Widr	55		Engine driver Marine	Lancashire Wigan	
114	4, Flemon 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	Unm	16		Coal Porter	Yorks Hull	
				Alice E. Do	Daur			13	Scholar	Do Do	
				Elizabeth Do	Daur			8	Do	Do Do	
				David W. Do	Son		7		Do	Do Do	
115	5 Do	1		William Vollans	Head	Mar	26		Fireman Locomotive	Northamptonshire Peterboro	
				Jane E. Do	Wife	Mar		35	Pianist	Yorks Sheffield	
				Ernest E. Walker	Son		13			Do Hull	
				Clinton E. Do	Son		11		Scholar	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

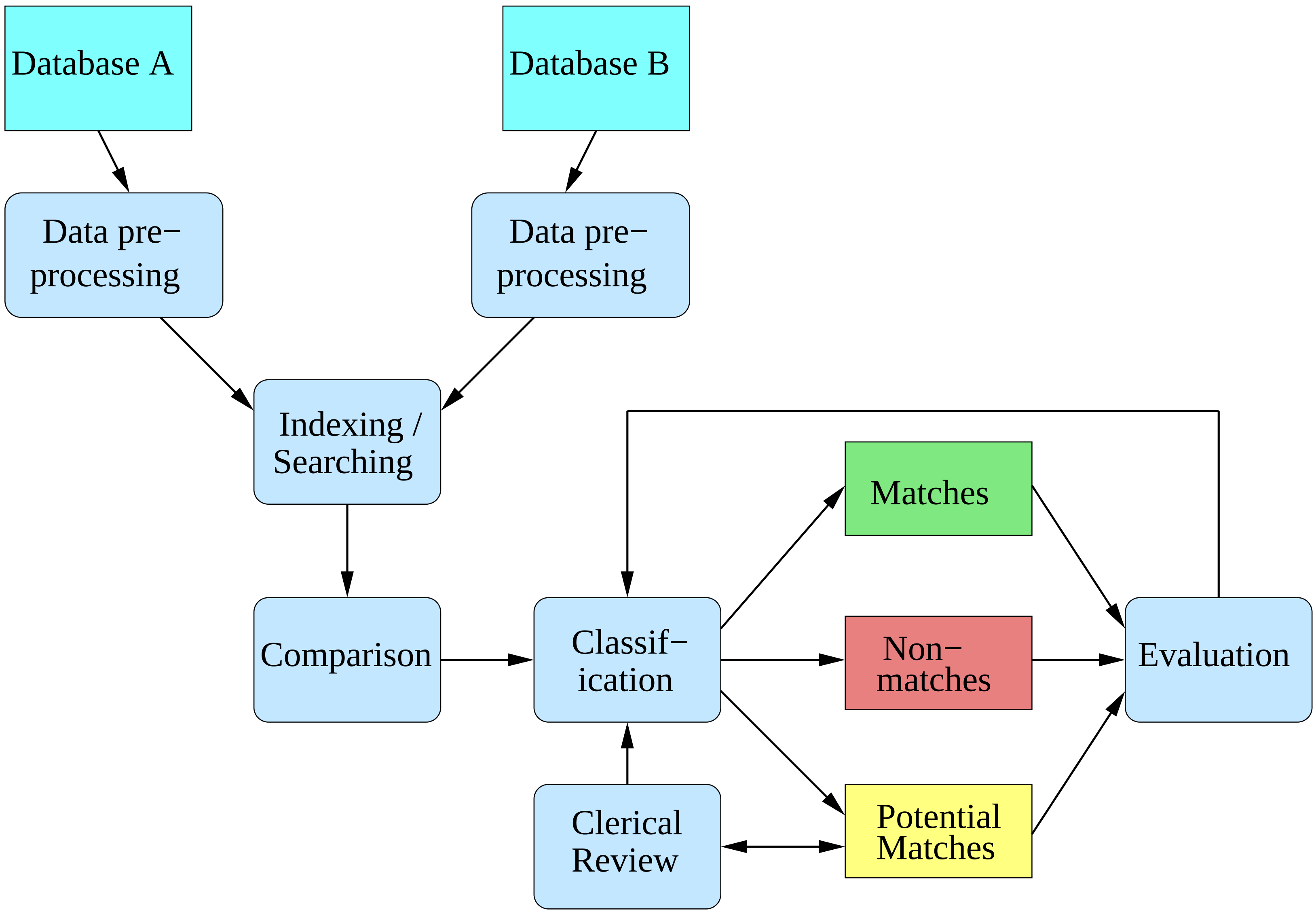
Challenges for present-day data

- These data are about living people, and so privacy is of major concern when data are linked between organisations
  - Linked data allow analyses 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?

Outline

- Part 1: Introduction
  - Applications, history, challenges, and examples
- Part 2: Data linkage process
  - Key techniques used in data linkage
- Part 3: Advanced data linkage techniques
  - Indexing and blocking for scalable data linkage
  - Learning, collective, and graph based techniques
- Part 4: Privacy aspects in data linkage
  - Motivating scenario
  - Privacy-preserving record linkage
- Conclusions and research directions

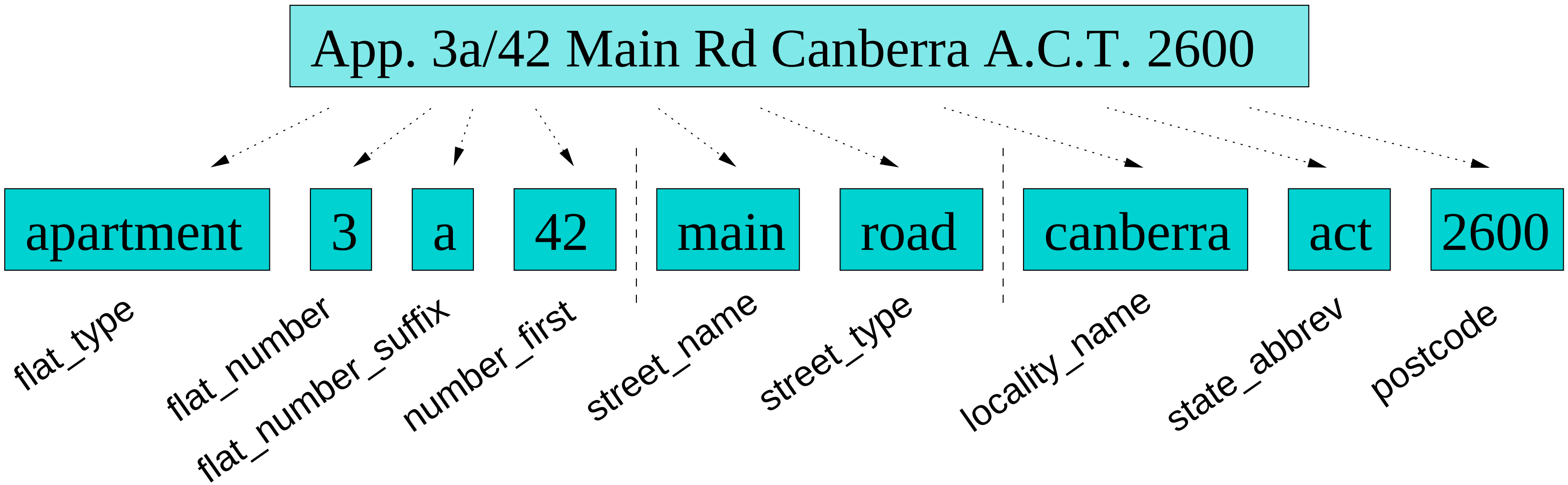
The data linkage process



Why cleaning and standardisation?

- Real world data are often *dirty*
  - Typographical and other errors
  - Different coding schemes
  - Missing values
  - Data changing over time
- Name and addresses are especially prone to data entry errors
  - Scanned, hand-written, over telephone, hand-typed
  - Same person often provides her/his details differently
  - Different correct spelling variations for proper names (e.g. *'Gail'* and *'Gayle'*, or *'Dixon'* and *'Dickson'*)

Example: Address standardisation



1. Clean input
   - Remove unwanted characters and words
   - Expand abbreviations and correct misspellings
2. Segment address into well defined *output fields*
3. Verify if address (or parts of it) exists in reality

Standardisation approaches

- Rules based
  - Manually developed parsing and transformation rules
  - Time consuming and complex to develop and maintain
- Probabilistic methods
  - Based for example on *hidden Markov models* (HMMs)
  - More flexible and robust with regard to new unseen data
  - Drawback: Training data needed for most methods
    (for example, sets of correctly standardised addresses)

HMMs are widely used in natural language processing and speech recognition, as well as for text segmentation and information extraction.

Standardisation steps

- Cleaning
  - Based on look-up tables and correction lists
  - Remove unwanted characters and words
  - Correct various misspellings and abbreviations
- Tagging
  - Split input into a list of *tokens* (words, characters, numbers, and separators)
  - Assign one or more *tags* to each token using look-up tables and/or features
- Segmenting
  - Use for example a trained HMM to assign list elements into *output fields*

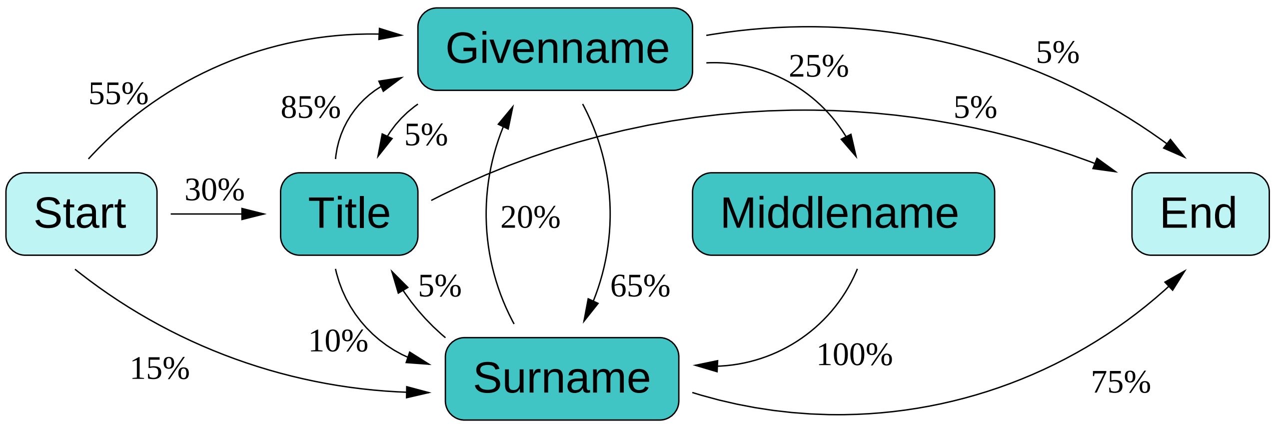
Data tagging example

- Tags provide information about the category / type of a token, such as:
  - `TI` Name title words ('ms', 'mr', 'dr', etc.)
  - `GM` Male given names ('thomas', 'paul', etc.)
  - `SN` Surnames ('smith', 'miller', 'thomas', etc.)
  - `N4` Four-digit numbers ('2602', '3000', etc.)
- Specific tags for names, addresses, and other domains (some overlapping, like street names)
- Example tagging:
  - Uncleaned input string: *'Doc. Thomas Paul MILLER'*
  - Cleaned string: *'dr thomas paul miller'*
  - Token and tag lists:

```
['dr',  'thomas',  'paul',  'miller']
['TI',  'GM/SN',   'GM',    'SN'    ]
```

THE AUSTRALIAN NATIONAL UNIVERSITY

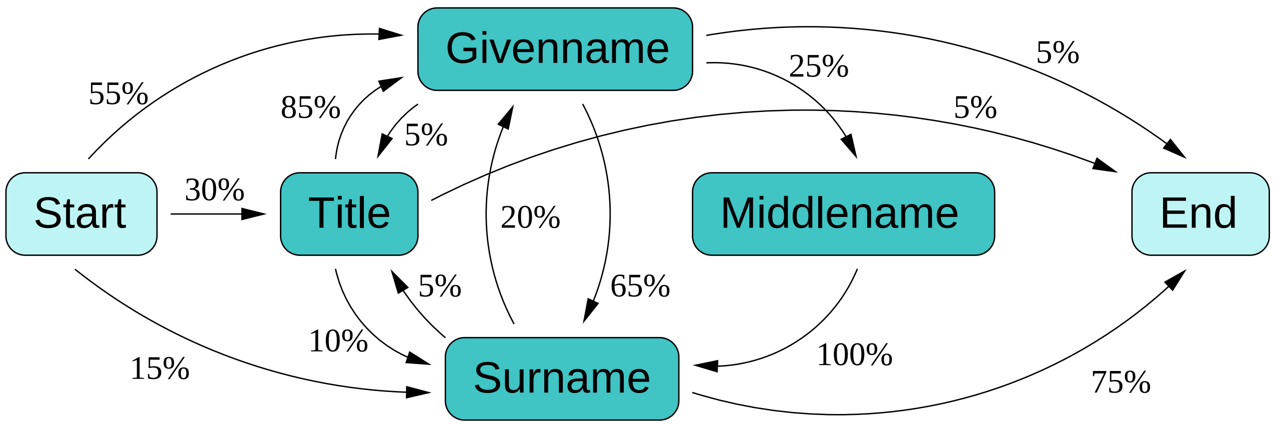
Hidden Markov model (HMM)



- A HMM is a *probabilistic* finite state machine
  - Made of a set of *states* and *transition probabilities* between these states
  - In each state an *observation* symbol is emitted with a certain probability distribution
  - For data segmentation, the observation symbols are *tags* and the states correspond to the *output fields*

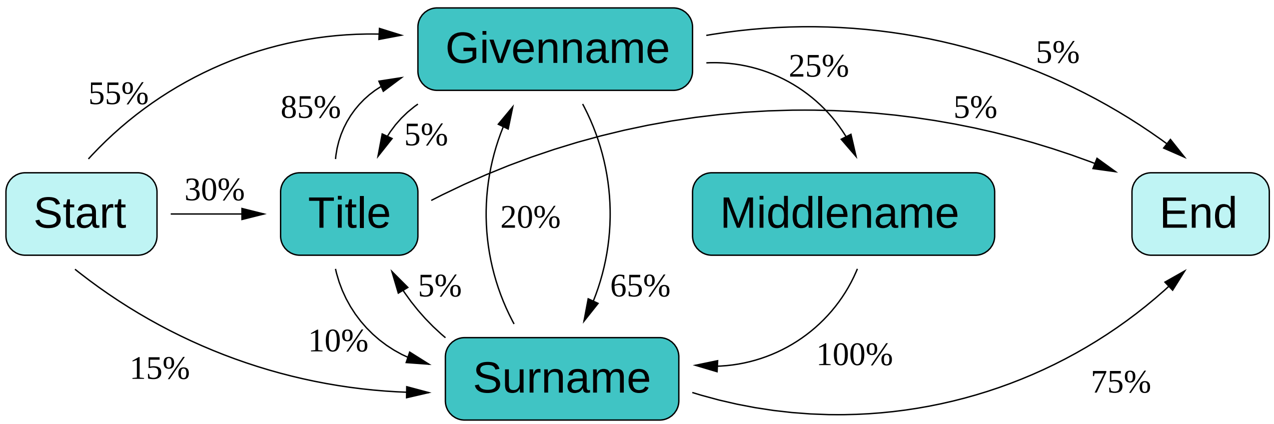
THE AUSTRALIAN NATIONAL UNIVERSITY

HMM probability matrices



	State					
Observation	*Start*	Title	Givenname	Middlename	Surname	*End*
`TI`	–	96%	1%	1%	1%	–
`GM`	–	1%	34%	33%	15%	–
`GF`	–	1%	36%	27%	14%	–
`SN`	–	1%	9%	14%	45%	–
`UN`	–	1%	20%	25%	25%	–

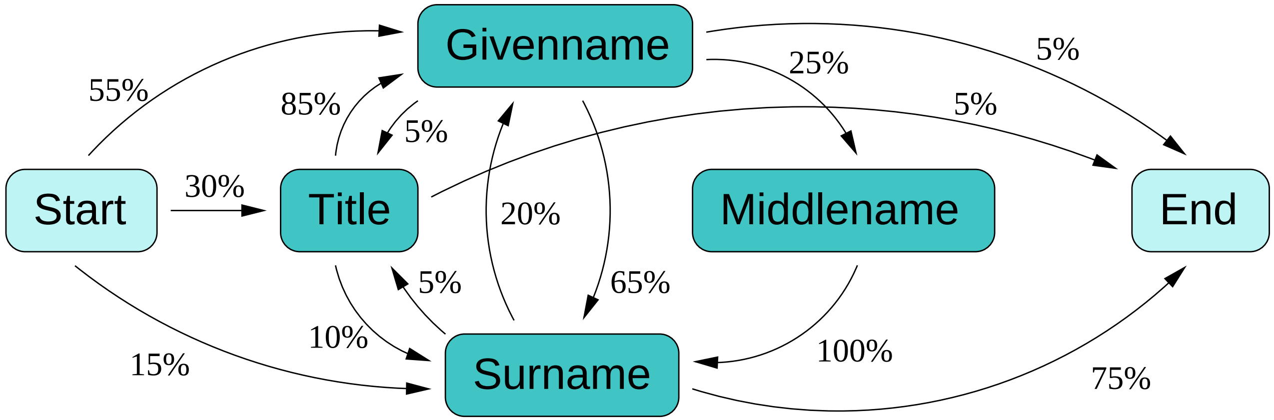
HMM data segmentation



- For an observation sequence we are interested in the most probable path through a given HMM
  (in our case an observation sequence is a list of *tags*)
- The *Viterbi* algorithm is used for this task
  (a dynamic programming approach)
- *Smoothing* is applied to account for unseen data
  (assign small probabilities for unseen observation symbols)

THE AUSTRALIAN NATIONAL UNIVERSITY

HMM segmentation example



- Input word and tag list

  `[‘dr’, ‘thomas’, ‘paul’, ‘miller’]`

  `[‘TI’, ‘GM/SN’, ‘GM’, ‘SN’ ]`

- Two example paths through HMM:

  Start → Title (TI) → Givenname (GM) → Middlename (GM) →
  Surname (SN) → *End*

  Start → Title (TI) → Surname (SN) → Givenname (GM) →
  Surname (SN) → *End*

HMM training

- Both transition and observation probabilities need to be trained (maximum likelihood estimates (MLE) are derived by accumulating frequency counts for transitions and observations)
- Training data consists of records, each being a sequence of `tag:hmm_state` pairs

  Examples training records:

  – `GM:Givenname`, `SN:Surname` ('peter', 'miller')
  – `UN:Givenname`, `SN:Surname` ('zikia', 'smith')
  – `TI:Title`, `GM:Givenname`, `GF:Surname` ('mr', 'john', 'kelly')
- Only a few person days are needed to get a HMM that results in an accurate standardisation (instead of weeks or even month to develop rules)

Blocking / indexing / filtering

- Number of record pair comparisons equals the product of the sizes of the two data sets
  (matching two data sets containing 1 and 5 million records will result in $1{,}000{,}000 \times 5{,}000{,}000$ record pairs)
- Performance bottleneck in a data linkage system is usually the (expensive) detailed comparison of field values between record pairs
  (such as approximate string comparison functions)
- Blocking / indexing / filtering techniques are used to reduce the large amount of comparisons
- Aim of blocking: Cheaply remove candidate record pairs which are obviously not matches

ANU
THE AUSTRALIAN NATIONAL UNIVERSITY

Traditional blocking

- Traditional blocking works by only comparing record pairs that have the same value for a *blocking variable* (for example, only compare records that have the same *postcode* value)
- Problems with traditional blocking
  - An erroneous value in a blocking variable results in a record being inserted into the wrong block (several *passes* with different blocking variables can solve this)
  - Values of blocking variable should have uniform frequencies (as the most frequent values determine the size of the largest blocks)

    Example: Frequency of *'Smith'* in NSW: *25,425*

    Frequency of *'Dijkstra'* in NSW: *4*

Phonetic encoding

- Bringing together spellings variations of the same name for improved blocking
- Techniques such as *Soundex*, *NYSIIS*, or *Double-Metaphone*
- Examples:

Name	Soundex	NYSIIS	Double-Metaphone
stephen	s315	staf	stfn
steve	s310	staf	stf
gail	g400	gal	kl
gayle	g400	gal	kl
christine	c623	chra	krst
christina	c623	chra	krst
kristina	k623	cras	krst

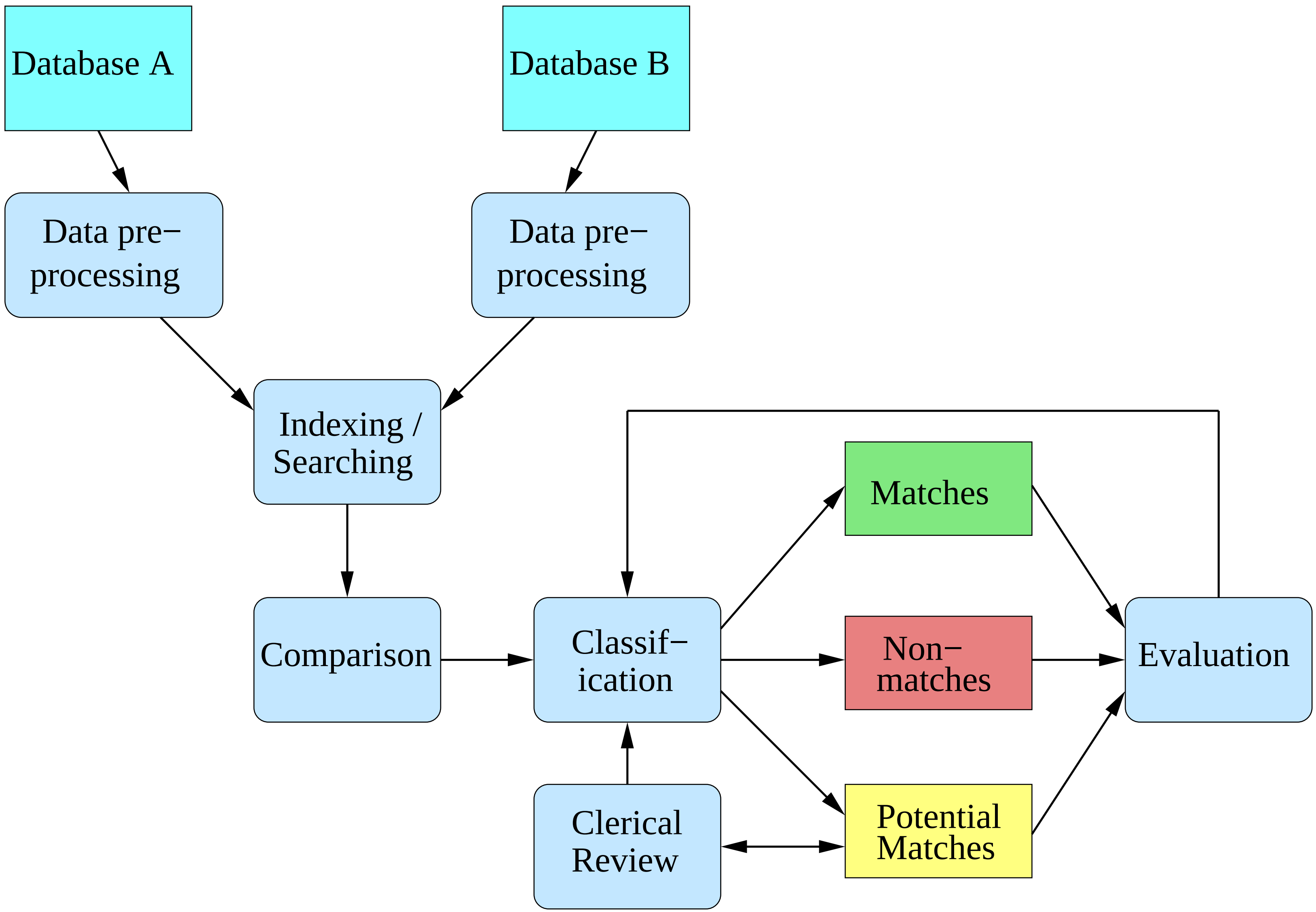
THE AUSTRALIAN NATIONAL UNIVERSITY

Soundex algorithm

- Keep first letter of a string (name), and remove all following occurrences of *a, e, i, o, u, y, h, w*
- Replace all consonants from position 2 onwards with digits using these rules:

  b, f, p, v $\rightarrow$ 1
  c, g, j, k, q, s, x, z $\rightarrow$ 2
  d, t $\rightarrow$ 3
  l $\rightarrow$ 4
  m, n $\rightarrow$ 5
  r $\rightarrow$ 6
- Only keep unique adjacent digits
- If length of code is less than 4 add zeros, if longer truncate at length 4

The data linkage process



Approximate string comparison

- Aim: Calculate a normalised similarity between two strings ($0 \leq sim_{approx} \leq 1$)
  - $sim_{approx} = 1 \rightarrow$ Same ('peter', 'peter')
  - $sim_{approx} = 0 \rightarrow$ Totally different ('peter', 'chris')
  - $0 < sim_{approx} < 1 \rightarrow$ Somewhat similar ('peter', 'pedro')
- Many different techniques available, some generic, others specific for certain types of strings
  - Edit-distance based (number of character edits)
  - Set-based (Jaccard, Dice, and Overlap coefficients)
  - Jaro-Winkler (specific for personal names)
  - Monge-Elkan and Soft-TFIDF (specific for strings that contain several words)

Q-gram based string comparisons

- Convert a string into q-grams (sub-strings of length q)
  - For example, for q = 2: 'peter' $\rightarrow$ ['pe','et','te','er']
- Find q-grams that occur in two strings, for example using the Dice coefficient:

  $sim_{Dice} = 2 \times c_c \,/\, (c_1 + c_2)$

  where c_c is number of common q-grams, and c_1, c_2 the number of q-grams in string s_1 and s_2
- With $s_1 =$ 'peter' and $s_2 =$ 'pete': $c_1 = 4$, $c_2 = 3$, and $c_c = 3$ ('pe','et','te'):

  sim_{Dice}('peter', 'pete') $= 2\times3/(4+3)$= 6/7 = 0.86

THE AUSTRALIAN NATIONAL UNIVERSITY

Edit-distance based string comparisons

- The number of character edits needed to convert one string into another (insert, delete, substitute)
- Can be calculated using a dynamic programming algorithm (of quadratic complexity in length of strings)
- Convert distance into a similarity as:

  sim_{ED} = 1 - $dist_{ED}$ / max(l_1, l_2)

  where l_1 = length of string s_1 and l_2 = length of s_2
- With s_1 = 'peter' and s_2 = 'pete': $l_1 = 5$, $l_2 = 4$, $dist_{ED} = 1$ (delete 'r'): sim_{ED} = 1 - 1/5 = 4/5 = 0.8
- Variations consider transposition of two adjacent characters, allow for gaps, or different edit costs (learned from training data)

THE AUSTRALIAN NATIONAL UNIVERSITY

Edit distance calculation example

- Matrix *D* shows number of edits between sub-strings (for example, 'ga' and 'gayle' -> 3 inserts)

D		g	a	y	l	e
	0	1	2	3	4	5
g	1	**0**	1	2	3	4
a	2	1	**0**	1	2	3
i	3	2	1	**1**	2	3
l	4	3	2	2	**1**	**2**

- If $s_1[i] = s_2[j]$, then $D[i,j] = D[i-1, j-1]$
- If $s_1[i] \neq s_2[j]$, then $D[i,j] =$

$$min \begin{cases} D[i-1,j]+1 & \text{del} \\ D[i,j-1]+1 & \text{ins} \\ D[i-1,j-1]+1 & \text{subst} \end{cases}$$

- Edit path: 'gail' $\rightarrow$ substitute 'i' with 'y' $\rightarrow$ insert 'e' $\rightarrow$ 'gayle' (final edit distance *dist*$_{ED}$('gail','gayle') $= 2$)

THE AUSTRALIAN NATIONAL UNIVERSITY

Probabilistic record linkage

- Basic ideas of probabilistic linkage were introduced by *Newcombe & Kennedy, 1962*
- Theoretical foundation by *Fellegi & Sunter, 1969*
  - Compare common record attributes (or fields) using approximate (string) comparison functions
  - Calculate matching weights based on frequency ratios (global or value specific ratios) and error estimates
  - Sum of the matching weights is used to classify a pair of records as a *match*, *non-match*, or *potential match*
  - Problems: Estimating errors, find optimal thresholds, assumption of independence, and manual *clerical review*

Fellegi and Sunter classification (1)

- For each compared record pair a vector of *matching weights* is calculated

Record A:	`['dr', 'thomas', 'paul', 'miller']`
Record B:	`['mr', 'john', '', 'miller']`
Matching weights:	`[0.2, -3.2, 0.0, 2.4 ]`

- A ratio *R* is calculated for each compared record pair *r = (a,b)* in the product space $\mathbf{A} \times \mathbf{B}$:

$$R = P(\gamma \in \Gamma \mid r \in M) / P(\gamma \in \Gamma \mid r \in U),$$

where M and U are the sets of true matches and true non-matches, and γ is an agreement pattern in the comparison space Γ, with:

$$\mathbf{A} \times \mathbf{B} = \{(a,b) : a \in \mathbf{A}, b \in \mathbf{B}\} \text{ for files } \mathbf{A} \text{ and } \mathbf{B}$$

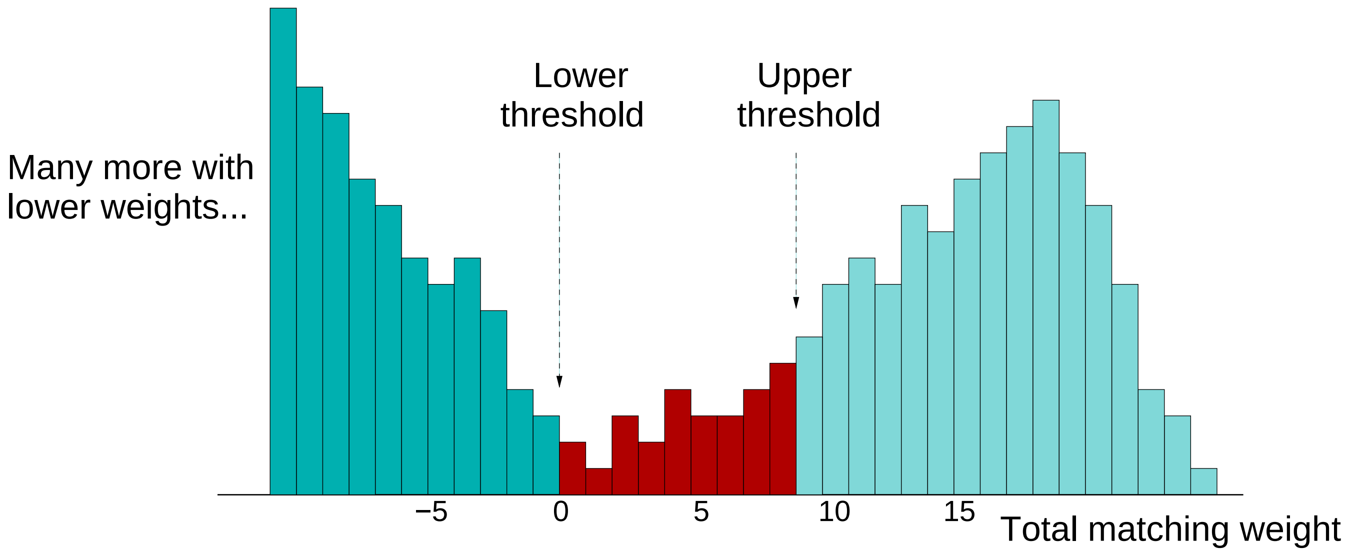
$$M = \{(a,b) : a = b,\ a \in \mathbf{A}, b \in \mathbf{B}\}$$

$$U = \{(a,b) : a \neq b,\ a \in \mathbf{A}, b \in \mathbf{B}\}$$

Fellegi and Sunter classification (2)

- Fellegi and Sunter proposed the following decision rule:

$$
\begin{array}{rcll}
R \geq t_u & \Rightarrow & r \rightarrow & \text{Match} \\
t_l < R < t_u & \Rightarrow & r \rightarrow & \text{Potential Match} \\
R \leq t_l & \Rightarrow & r \rightarrow & \text{Non-Match}
\end{array}
$$



Fellegi and Sunter classification (3)

- Assuming conditional independence between attributes allows to calculate individual attribute-wise probabilities

  $$m_i = P([a_i = b_i, a \in \mathbf{A}, b \in \mathbf{B}] \mid r \in M) \text{ and}$$
  $$u_i \;= P([a_i \neq b_i, a \in \mathbf{A}, b \in \mathbf{B}] \mid r \in U),$$

  where a_i and b_i are the values of attribute *i* being compared

- Based on these *m*- and *u*-probabilities, we calculate a *matching weight* w_i for attribute i as:

$$w_i = \begin{cases} log_2(\frac{m_i}{u_i}) & \text{if } \; a_i = b_i \; \text{ (agreement weight)} \\ log_2(\frac{(1-m_i)}{(1-u_i)}) & \text{if } \; a_i \neq b_i \; \text{ (disagreement weight)} \end{cases}$$

THE AUSTRALIAN NATIONAL UNIVERSITY

Weight calculation: Month of birth

- Assume two data sets with a *3%* error in field *month of birth*
- Probability that two matched records (representing the same person) have the same month value is *97%* (m_i)
- Probability that two matched records do not have the same month value is *3%* ($1\text{-}m_i$)
- Probability that two (randomly picked) un-matched records have the same month value is *1/12 = 8.3%* (u_i)
- Probability that two un-matched records do not have the same month value is *11/12 = 91.7%* ($1\text{-}u_i$)
- Agreement weight $log_2(m_i / u_i)$: $log_2(0.97 / 0.083) = 3.54$

  Disagreement weight $log_2(1\text{-}m_i) / (1\text{-}u_i)$: $log_2(0.03 / 0.917) = -4.92$

Data linkage evaluation (1)

- At the end we need to evaluate how good the results of a data linkage project are
- Main measures for linkage complexity
  - **Reduction ratio**: How many candidate record pairs were generated by blocking, compared to all pairs?

$$rr = 1 - \left(\frac{number\ of\ candidate\ pairs}{number\ of\ all\ record\ pairs} \right)$$

  - **Pairs completeness**: How many true matches were generated by blocking, divided by all true matches?

$$pc = \frac{number\ of\ true\ matching\ candidate\ pairs}{number\ of\ all\ true\ matching\ pairs}$$

Data linkage evaluation (2)

- To measure linkage quality, we need true matches (gold standard, ground truth data)
  - Two types of errors:
    - – A missed true match (false non-match, *false negative*)
    - – A wrong match (false match, *false positive*)
- Data linkage is often a very *imbalanced* problem
  - Most records pairs (even after blocking) are true non-matches
- Calculating *accuracy* is not meaningful (percentage of false matches and false non-matches)
  - Classifying all record pairs as non-matches can give very high accuracy

THE AUSTRALIAN NATIONAL UNIVERSITY

Data linkage evaluation (3)

- Commonly used measures are similar to information retrieval (Web search)
  - **Precision**: How many true matches are in the set of classified matches?

$$prec = \frac{number\ of\ true\ matching\ pairs}{number\ of\ classified\ matching\ pairs} = \frac{tp}{tp + fp}$$

  - **Recall**: How many true matches did we find from all known true matches?

$$reca = \frac{number\ of\ true\ matching\ pairs}{number\ of\ all\ true\ matching\ pairs} = \frac{tp}{tp + fn}$$

  - Number of true non-matches (*tn*) are not used for precision and recall

THE AUSTRALIAN NATIONAL UNIVERSITY

Outline

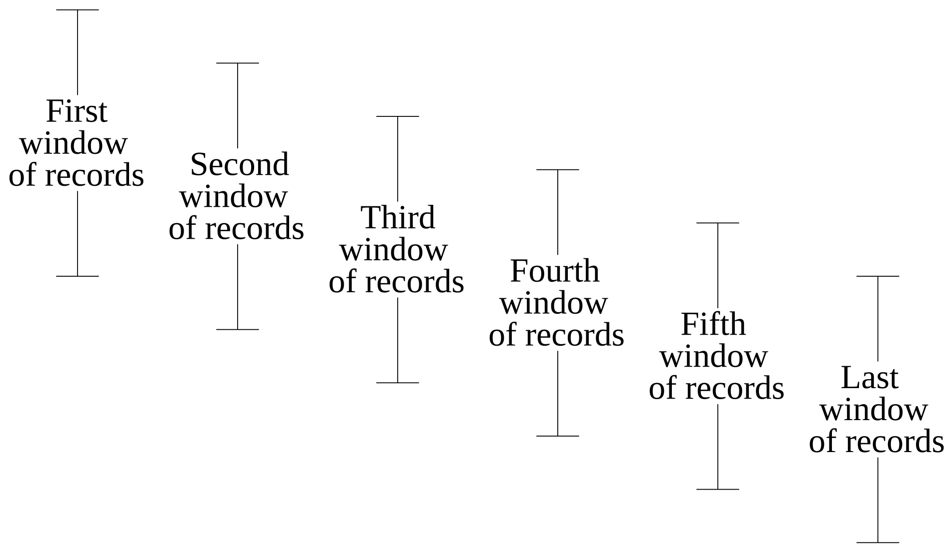
- Part 1: Introduction
  - Applications, history, challenges, and examples
- Part 2: Data linkage process
  - Key techniques used in data linkage
- Part 3: Advanced data linkage techniques
  - Indexing and blocking for scalable data linkage
  - Learning, collective, and graph based techniques
- Part 4: Privacy aspects in data linkage
  - Motivating scenario
  - Privacy-preserving record linkage
- Conclusions and research directions

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 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)

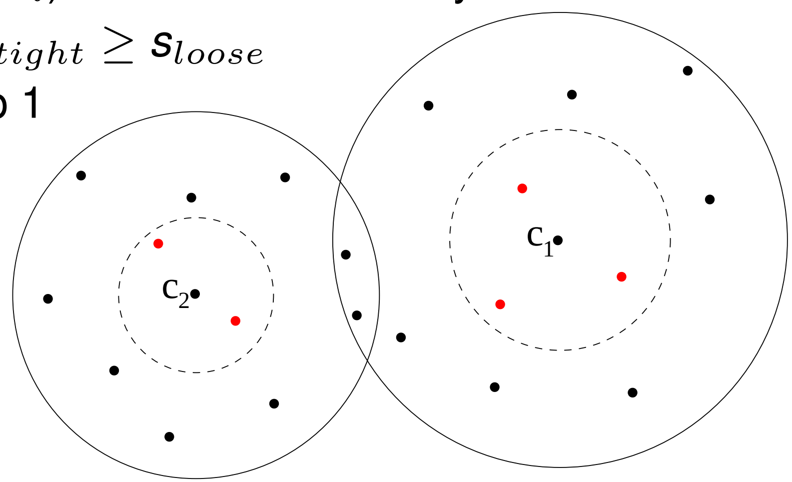
- *Canopy* clustering
  - Based on a computationally ‘cheap’ similarity measure such as Jaccard (set intersection based on q-grams)
  - Records will be inserted into several clusters / blocks
  - Algorithm steps:

    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



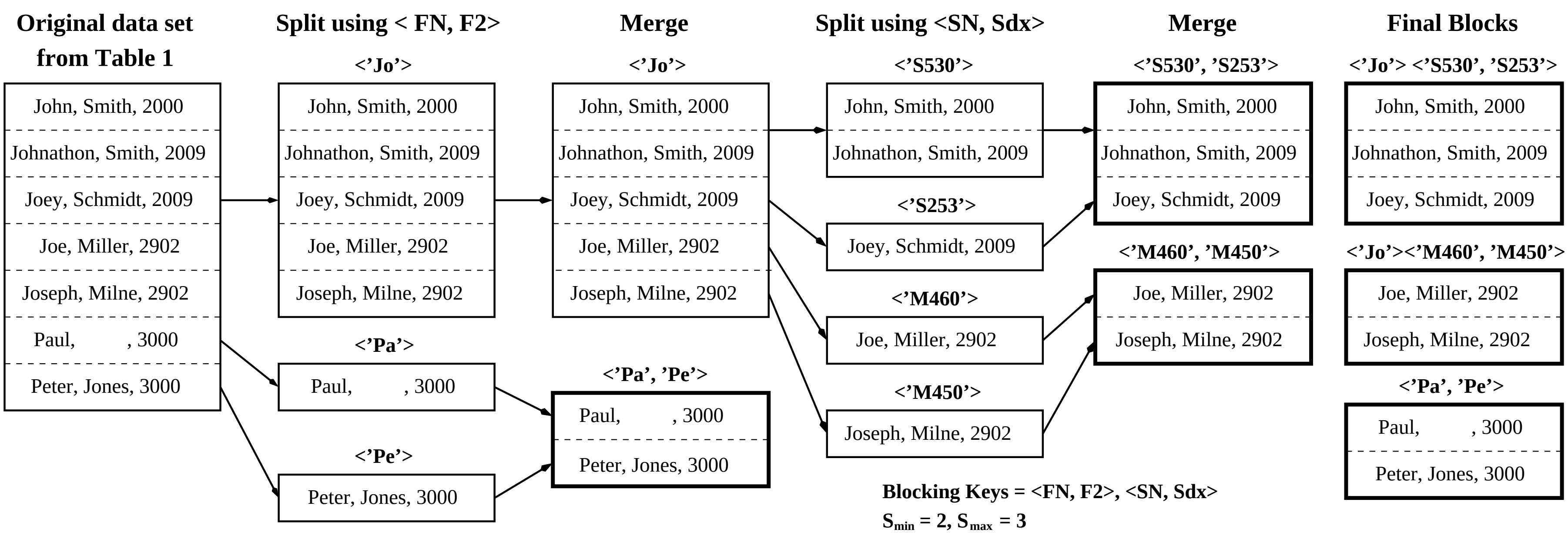
Advanced indexing approaches (3)

- *Q*-gram based blocking (e.g. *2-grams / bigrams*)
  - Convert values into *q*-gram lists, then generate sub-lists

    'peter' $\rightarrow$ *['pe','et','te','er'],* ***['pe','et','te']****, ['pe','et','er'], ..*
    'pete' $\rightarrow$ ***['pe','et','te']****, ['pe','et'], ['pe','te'], ['et','te'], ...*
  - Records with the same sub-list value are inserted into the same block
  - Each record will be inserted into several blocks
  - Works well for'dirty' data but has high computational costs
- Mapping-based blocking
  - Map strings into a multi-dimensional space such that distances between strings are preserved

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



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

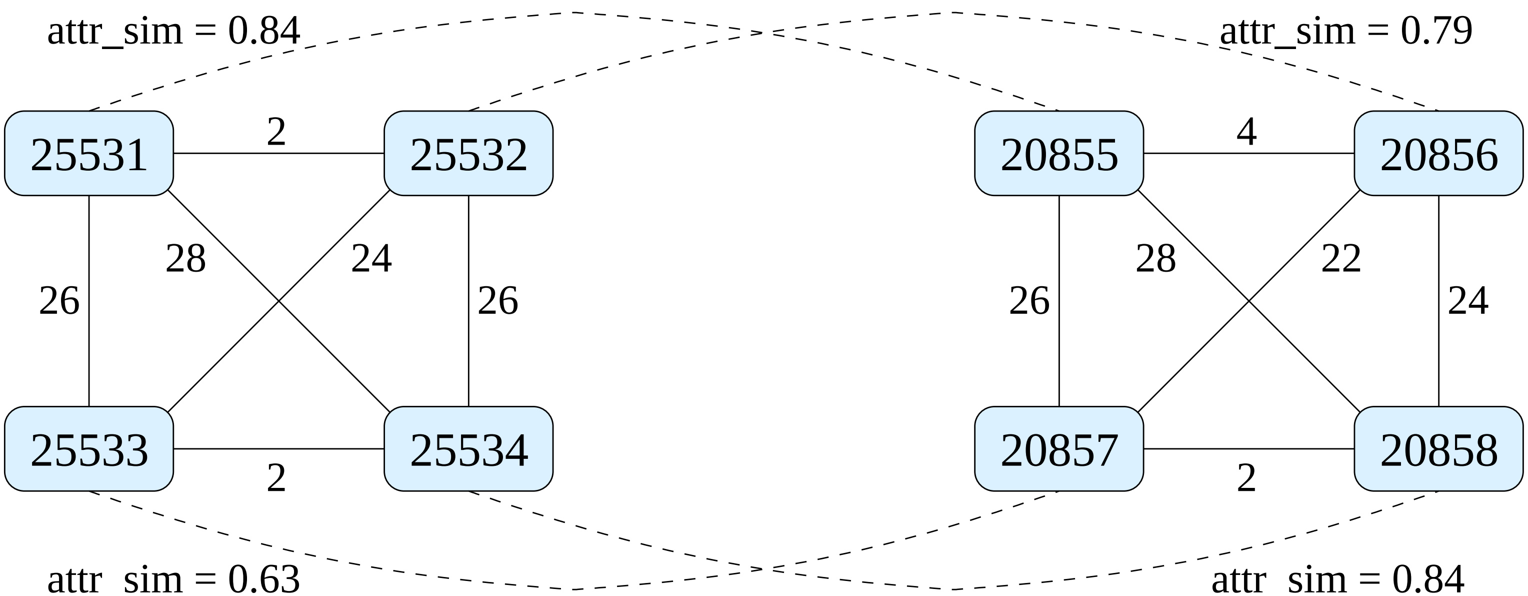
- In many cases there are no training data available
  - Possible to use results of earlier matching projects? Or from manual *clerical review* process?
  - How confident can we be about correct manual classification of *potential matches*?
- Often there is no *gold standard* available (no data sets with known true match status)
- No large test data set collection 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: 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**
  - First conduct pair-wise linking of individual records
  - Then calculate group similarities using Jaccard or weighted similarities (based on pair-wise similarities)

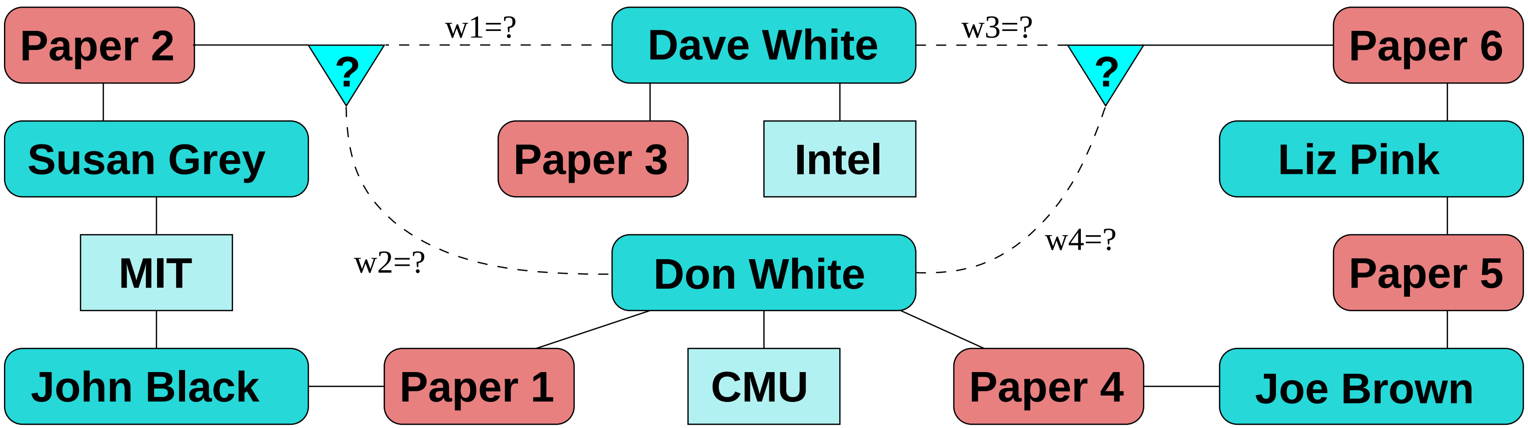
Advanced classification: Graph-based linkage

- 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: Collective entity resolution

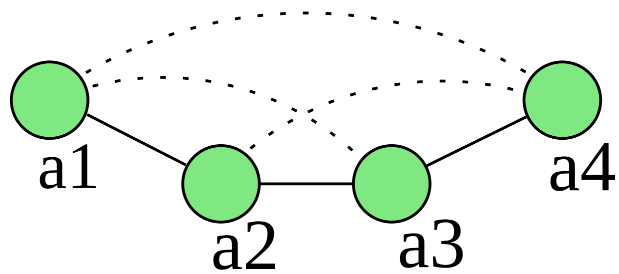
- Considers *relational similarities* not just attribute similarities



(A1, Dave White, Intel)	(P1, John Black / Don White)
(A2, Don White, CMU)	(P2, Sue Grey / **D. White**)
(A3, Susan Grey, MIT)	(P3, Dave White)
(A4, John Black, MIT)	(P4, Don White / Joe Brown)
(A5, Joe Brown, unknown)	(P5, Joe Brown / Liz Pink)
(A6, Liz Pink, unknown)	(P6, Liz Pink / **D. White**)

Adapted from: [Kalashnikov and Mehrotra, ACM TODS, 2006]

Managing transitive closure

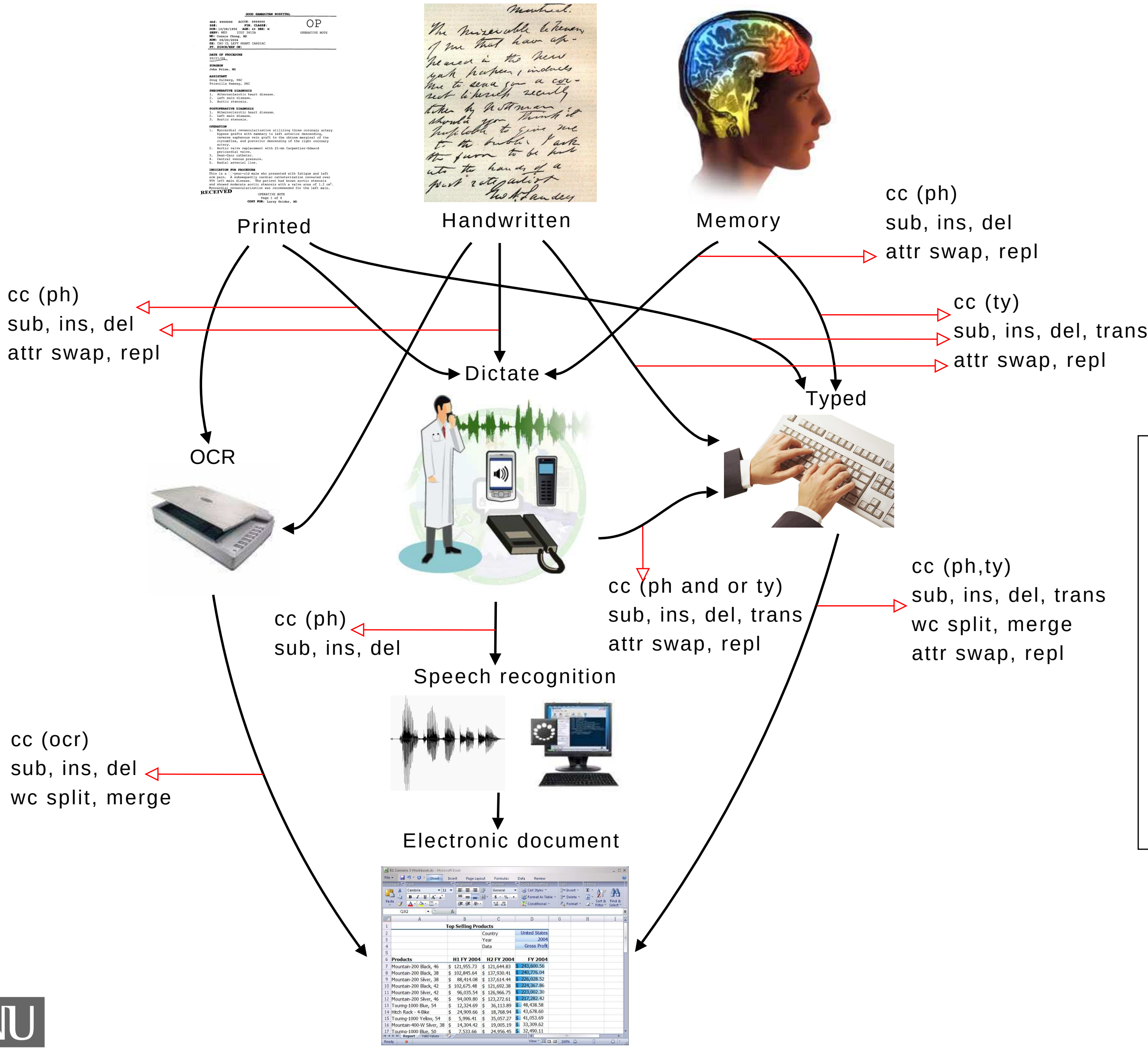


- If record *a1* is classified as matching with record *a2*, and record *a2* as matching with record *a3*, then records *a1* and *a3* must also be matching
- Possibility of chains of linked records occurring
- Various algorithms have been developed to find optimal solutions (special clustering algorithms)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

Generating and using synthetic data

- Privacy issues prohibit publication of real personal information
- De-identified or encrypted data cannot be used for data linkage research
  (as real name and address values are required)
- Several advantages of synthetic data
  - Volume and characteristics can be controlled (errors and variations in records, number of duplicates, etc.)
  - It is known which records are duplicates of each other, and so matching quality can be calculated
  - Data and the data generator program can be published (allowing others to repeat experiments)

Modelling of variations and errors



Abbreviations:
cc : character change
wc : word change
subs : substitution
ins : insertion
del : deletion
trans : transpose
repl : replace
ty : typographic
ph : phonetic
attr : attribute

Example of generated data

rec_id	age	given_name	surname	street	suburb
rec-1-org	***33***	***Madison***	Solomon	Tazewell ***Circuit***	***Beechboro***
rec-1-dup-0	33	Madisoi	Solomon	Tazewell Circ	Beech Boro
rec-1-dup-1		Madison	Solomon	Tazewell Crct	Bechboro
rec-2-org	39	***Desirae***	***Contreras***	Maltby Street	***Burrawang***
rec-2-dup-0	39	Desirae	Kontreras	Maltby Street	Burawang
rec-2-dup-1	39	Desire	Contreras	Maltby Street	Buahrawang
rec-3-org	***81***	***Madisyn***	Sergeant	***Howitt*** Street	***Nangiloc***
rec-3-dup-0	87	Madisvn	Sergeant	Hovvitt Street	Nanqiloc

- rec-1: typing/abbreviations; rec-2: phonetic; rec-3: OCR
- Generated using the *Febrl* and *GeCo* data generators (see: `https://dmm.anu.edu.au/geco/`)

THE AUSTRALIAN NATIONAL UNIVERSITY

Outline

- Part 1: Introduction
  - Applications, history, challenges, and examples
- Part 2: Data linkage process
  - Key techniques used in data linkage
- Part 3: Advanced data linkage techniques
  - Indexing and blocking for scalable data linkage
  - Learning, collective, and graph based techniques
- Part 4: Privacy aspects in data linkage
  - Motivating scenario
  - Privacy-preserving record linkage
- Conclusions and research directions

Privacy aspects in data 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 party involved in the linking, or any external party.*
- 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

ANU THE AUSTRALIAN NATIONAL UNIVERSITY

Privacy and data linkage: A motivating scenario

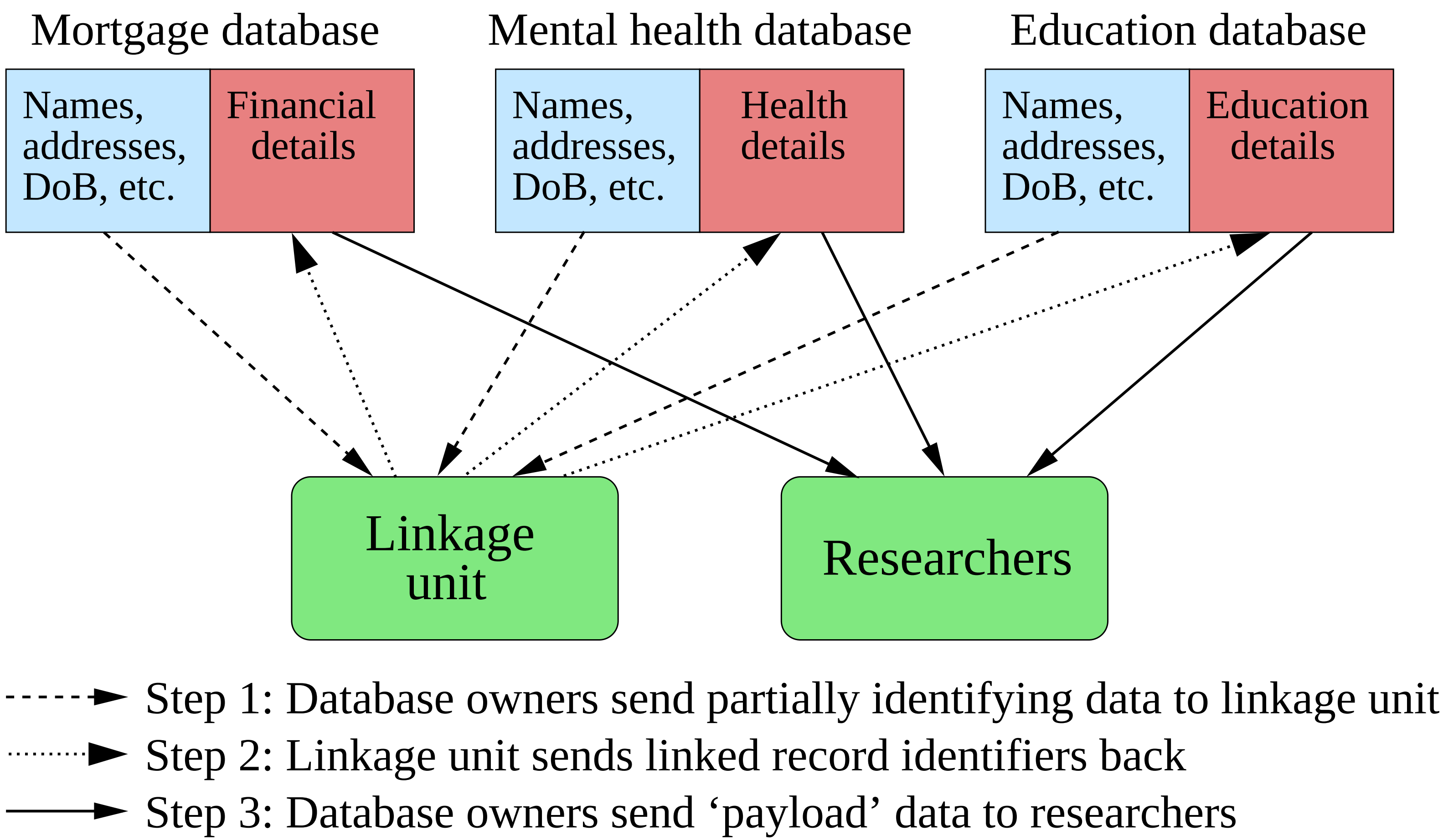
- A demographer who aims to investigate how mortgage stress is affecting different people with regard to their mental and physical health
- She will need data from financial institutions, government agencies (social security, health, and education), and private sector providers (such as health insurers)
- 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)

THE AUSTRALIAN NATIONAL UNIVERSITY

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 does never see payload data
- Researchers do not see personal details
- All communication is encrypted

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

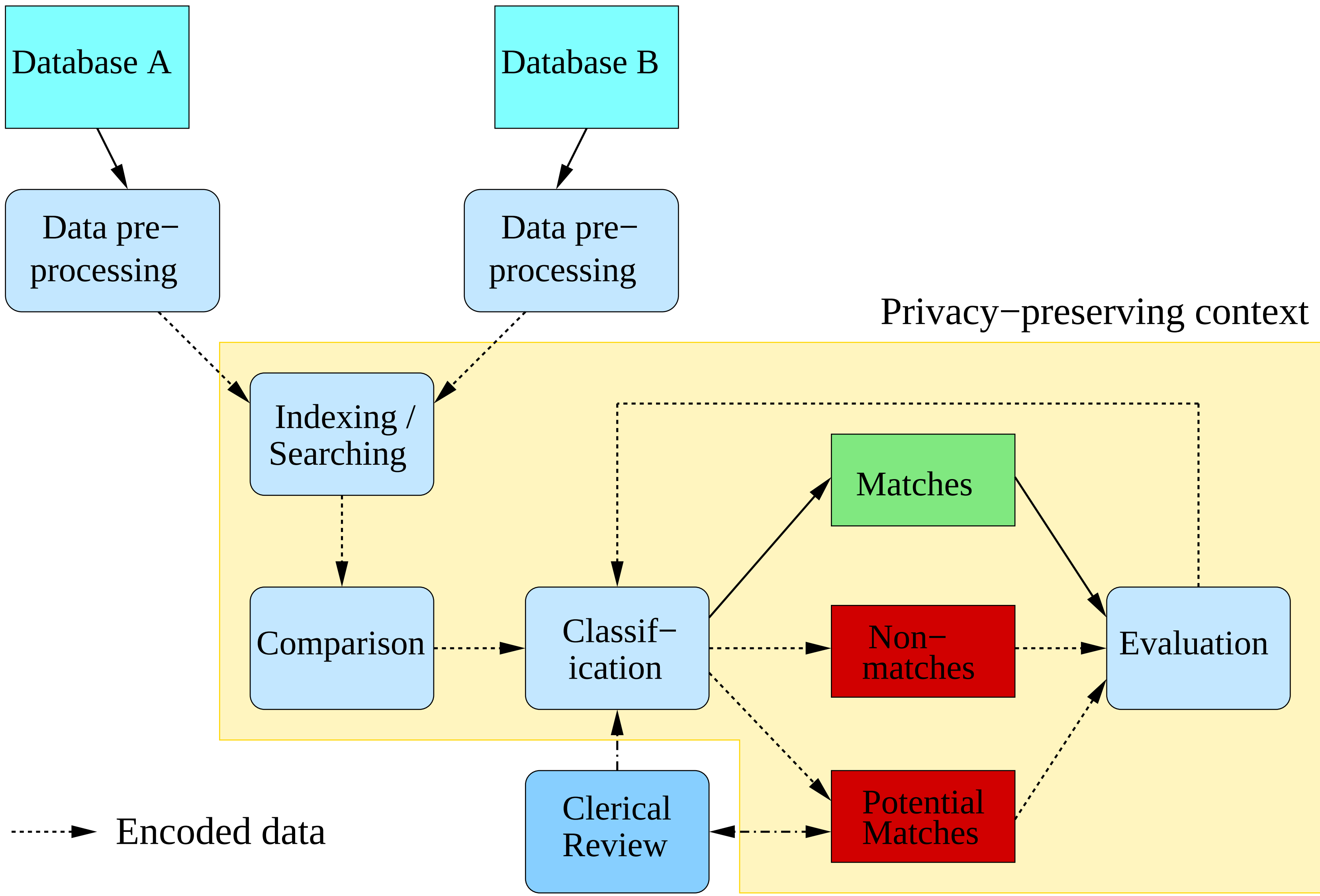


Details given in: Chris Kelman, John Bass, and D'Arcy Holman: *Research use of Linked Health Data – A Best Practice Protocol*, Aust NZ Journal of Public Health, vol. 26, 2002.

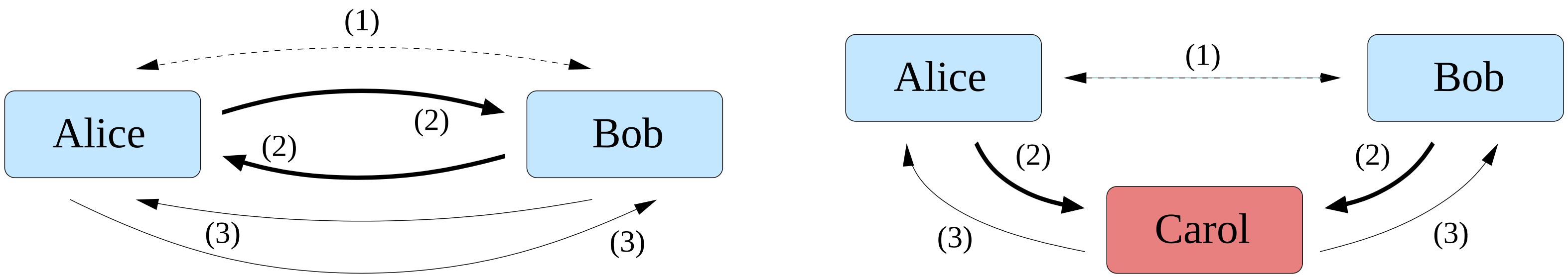
Current best practice approach used in the health domain (3)

- 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)

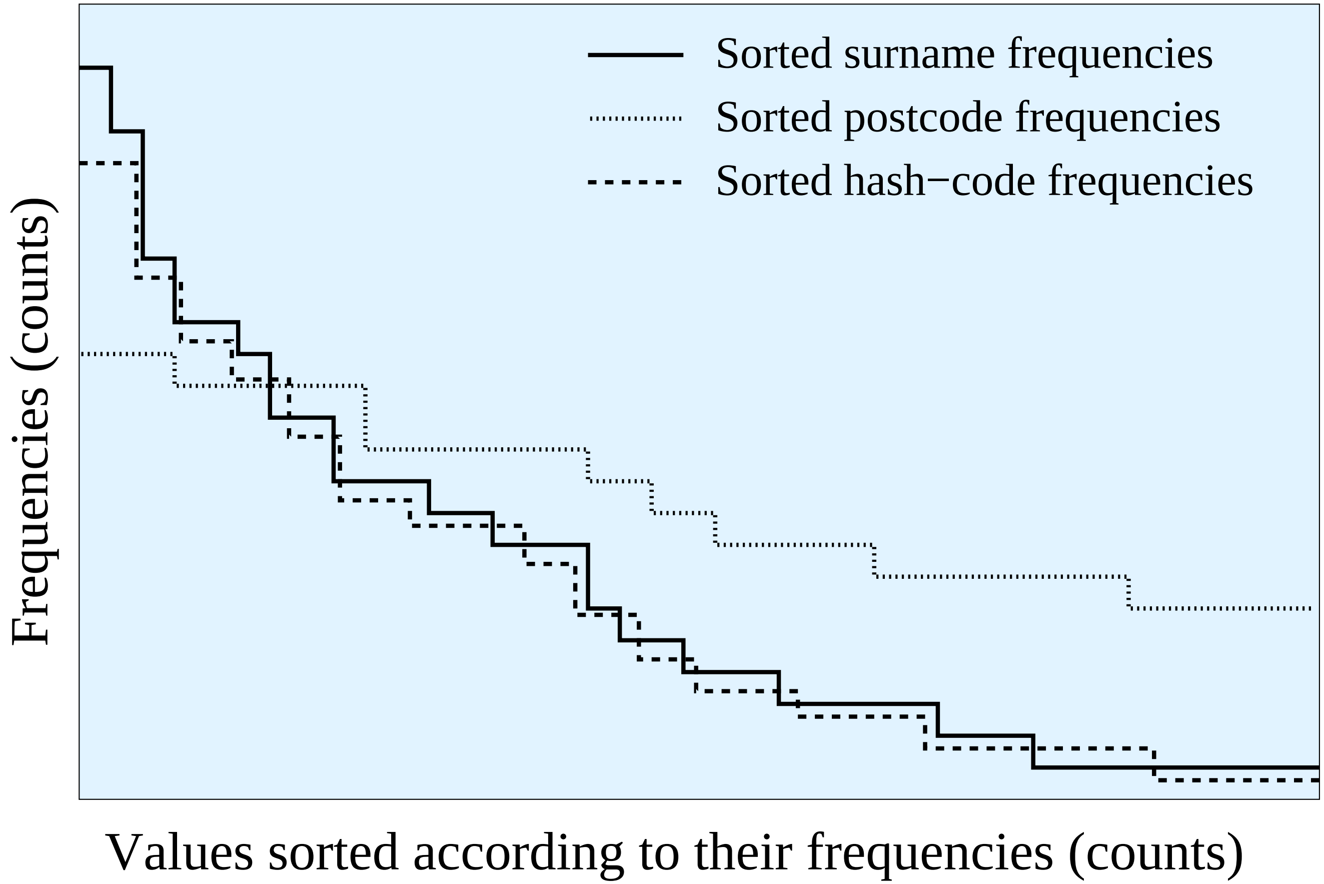
Adversary models

- *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

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)

ANU THE AUSTRALIAN NATIONAL UNIVERSITY

Hash-encoding for PPRL

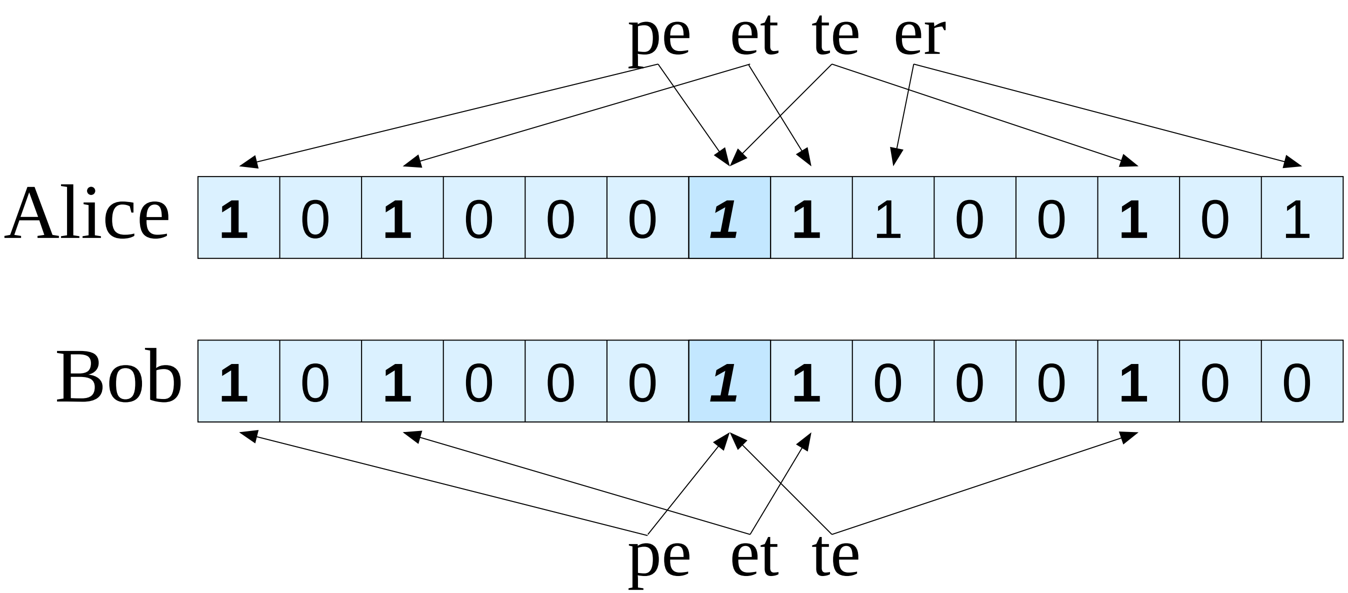
- A basic building block of many PPRL protocols
- Idea: Use a one-way hash function (like SHA) to encode values, then compare hash-codes
  - Having only access to hash-codes will make it nearly impossible to learn their original input values
  - But dictionary and frequency attacks are possible
- Single character difference between two input values results in completely different hash codes
  - For example:

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

Bloom filter based PPRL (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)

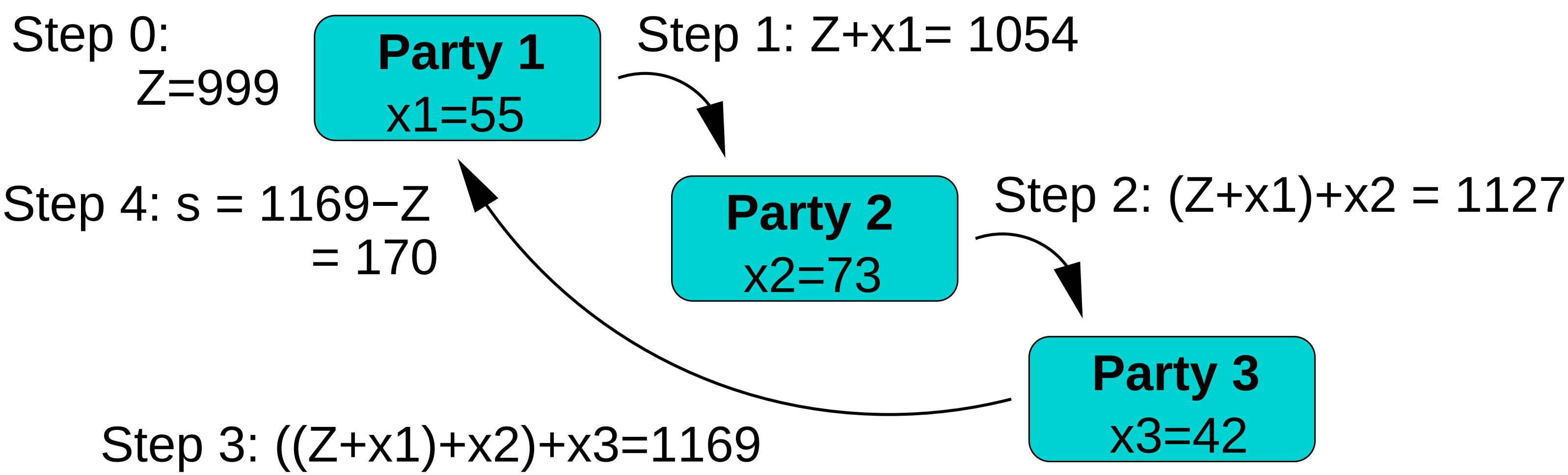
Bloom filter based PPRL (2)



- *1*-bits for string ‘peter’: 7, *1*-bits for ‘pete’: 5, common *1*-bits: 5, therefore $sim_{Dice} = 2\times5/(7+5)$= 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

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
- Simple example: Secure summation $s = \sum_i x_i$.



Outline

- Part 1: Introduction
  - Applications, history, challenges, and examples
- Part 2: Data linkage process
  - Key techniques used in data linkage
- Part 3: Advanced data linkage techniques
  - Indexing and blocking for scalable data linkage
  - Learning, collective, and graph based techniques
- Part 4: Privacy aspects in data linkage
  - Motivating scenario
  - Privacy-preserving record linkage
- Conclusions and research directions

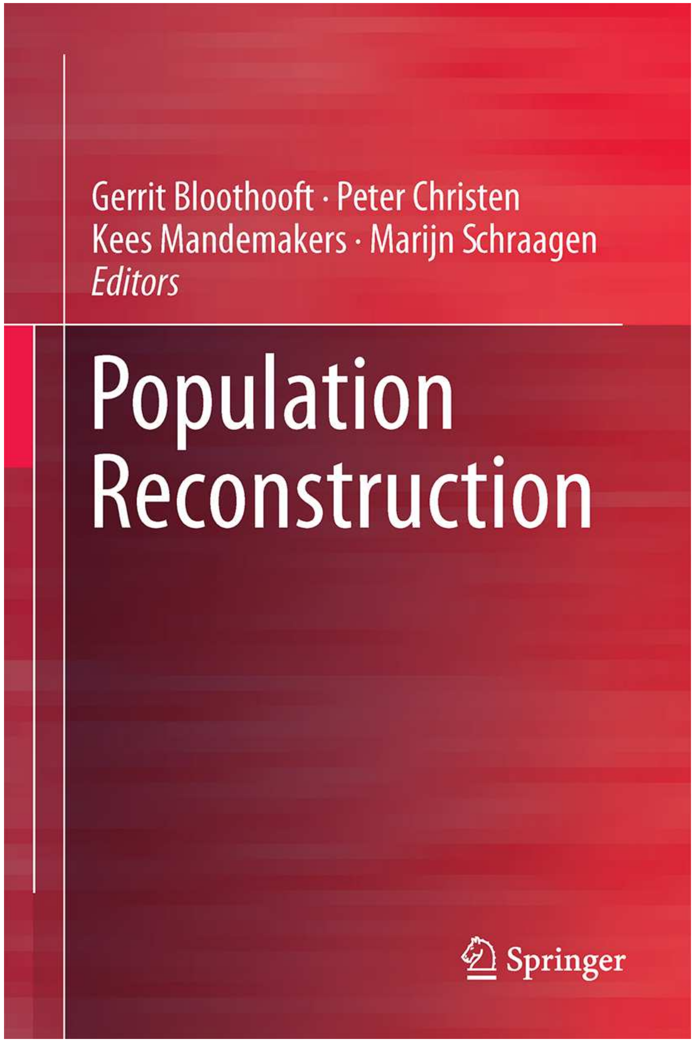
Conclusions and research directions (1)

- For historical data, a major 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 available in many 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 data linkage that allow comparative experimental studies

Conclusions and research directions (2)

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

Advertisement: Book 'Population Reconstruction' (August 2015)



The book details the possibilities and limitations of information technology with respect to reasoning for population reconstruction.

Follows the three main processing phases from handwritten registers to a reconstructed digitized population.

Combines research from historians, social scientists, linguists, and computer scientists.

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.

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.

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.

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.

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.

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.

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