Linking administrative databases

Recent developments and research challenges

Peter Christen

Research School of Computer Science,

The Australian National University

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

THE AUSTRALIAN NATIONAL UNIVERSITY

Outline

- A short introduction to data linkage
- Challenges of linking administrative databases

- Techniques for scalable data linkage
- Automating the linkage process
- Privacy aspects in data linkage
- Evaluating linkage quality and completeness
- Dealing with uncertainty in linked data sets

- Towards an end-to-end linkage framework
- Major research challenges

Two quotes

Each person in the world creates a Book of Life. This Book starts with birth and ends with death. Record linkage is the name of the process of assembling the pages of this Book into a volume.

Halbert L. Dunn, 1946

The biggest payoff will lie in new combinations of designed data and organic data, not in one type alone.

Robert Groves (US Census Bureau), 2011

What is data linkage?

- The process of linking records that represent the same entity in one or more databases

  (patients, customers, businesses, 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)

  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**

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, such as personal details like names, addresses, dates of birth, etc.
  - Use different match weights for individual attributes
- "Computer science" approaches
  - Based on machine learning, data mining, database, or information retrieval techniques
  - Supervised classification: Requires training data (true matches)
  - Unsupervised: Clustering or graph based approaches

General data linkage challenges

- Real world data are dirty

  (typographical errors and variations, missing and out-of-date values, different coding schemes, etc.)

- Scalability
  - Naïve comparison of all record pairs is quadratic
  - Remove likely non-matches as efficiently as possible
- No ground truth data in many linkage applications
  - No record pairs with known true match status, making assessment of linkage quality difficult
- Privacy and confidentiality

  (because personal information, like names and addresses, is commonly required for linking)

Challenges for linking administrative databases

- Size and complexity of administrative databases
  - Possibly hundreds of millions of records
  - Potentially data from many different sources
  - Containing more complex types and more detailed data (free-format text or multimedia)
- Databases are not collected specifically for data linkage projects
  - Attributes required for linking might be missing
  - Databases might be collected at different points in time (challenging as people can change name and address)
- Trustworthiness of (external) data
- Diverse requirements on linked data

Outline

- A short introduction to data linkage
- Challenges of linking administrative databases

- Techniques for scalable data linkage
- Automating the linkage process
- Privacy aspects in data linkage
- Evaluating linkage quality and completeness
- Dealing with uncertainty in linked data sets

- Towards an end-to-end linkage framework
- Major research challenges

Techniques for scalable data linkage

- Number of all record pair comparisons equals the product of the sizes of two databases
- Performance bottleneck in data linkage is usually the detailed comparison of attribute values
  (using approximate (string) comparison functions)
- Aim of indexing / blocking: Cheaply remove record pairs that are obviously not matches
- Traditional blocking only compares record pairs with the same value in a *blocking key*
  - For example, only compare records with the same *postcode*, or the same *surname+birth_year*
  - Common blocking key values will generate large blocks

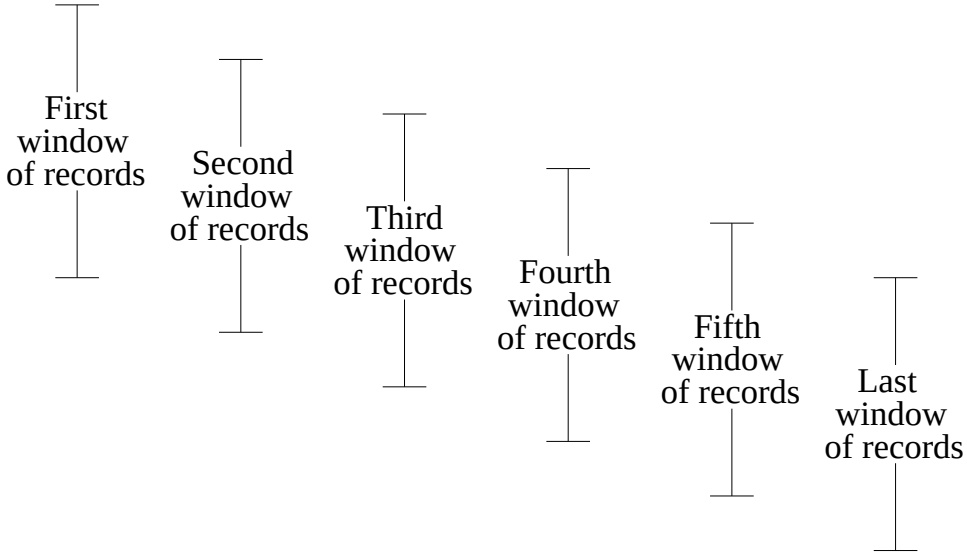
Sorted neighbourhood indexing

(Hernandez and Stolfo, 1995; Draisbach et al., 2012)

- Use a sliding window over sorted databases
- Use several passes with different sorting criteria
- Window size can be fixed or adaptive
  (based on similarities between sorting criteria)

For example, database sorted using first and last name:

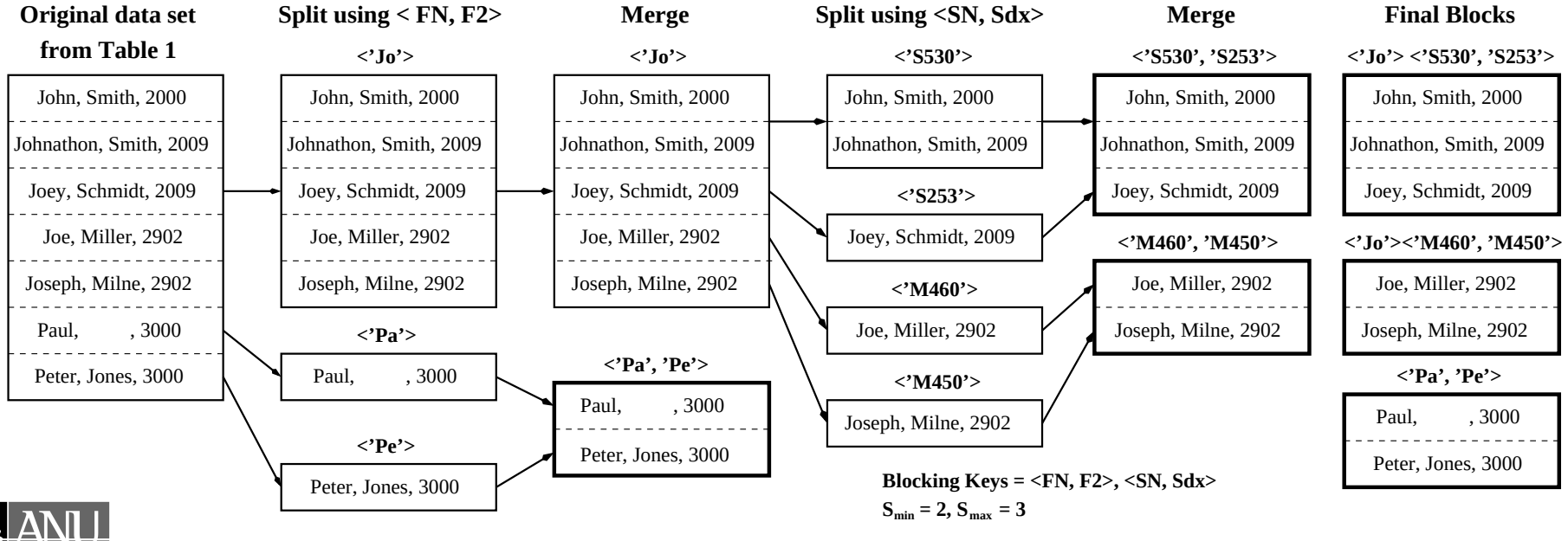
abbybond	r5
paulsmith	r2
pedrosmith	r4
pedrosmith	r9
percysmith	r1
petersmith	r7
petersmith	r10
robinstevens	r3
sallytaylor	r6
sallytaylor	r8



Controlling block sizes

(Fisher et al., 2015)

- Many blocking techniques generate blocks of different sizes (depending upon data characteristics)
  - Having blocks within a certain size range is important for real-time and privacy-preserving record linkage, and with certain machine learning algorithms
- We employ an iterative split-merge approach



Data linkage classification

- Traditional data linkage techniques classify pairs of records individually using similarity thresholds
  (thresholds set manually or based on error estimates)
- Supervised machine learning techniques generally result in much better match quality
  - However, training data in the form of true matches and non-matches are rarely available in practice
  - These have to be manually generated, which is often difficult both in terms of cost and quality
- Two main challenges for generating training data
  1. How to ensure *good* examples are selected
  2. How to *minimise* the user's burden of labelling examples

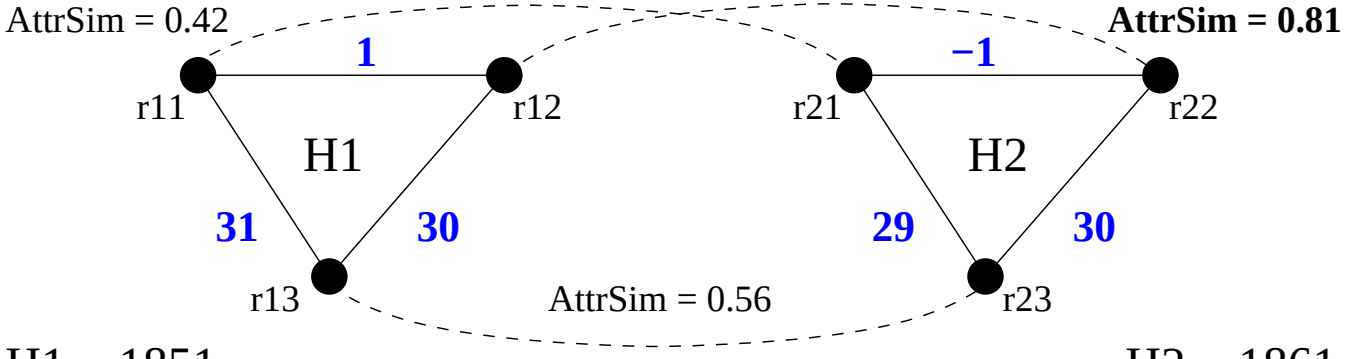
Advanced classification: Active learning and group linkage

- 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)
  - Iteratively train and improve a classification model using manually labelled data
- Group linkage
  - First conduct pair-wise linking of individual records
  - Then calculate group similarities based on pair-wise record similarities to identify new record pairs that are part of a group

Graph-based group linkage

(Fu et al., 2014)

- 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 contain information that does not change over time (like age differences)



H1 − 1851

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

H2 − 1861

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

Privacy aspects in record linkage
An example scenario

- A national crime investigation unit is tasked with identifying crimes that are of national significance (organised crime or money laundering)
- This unit will likely manage various national databases collected 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
- Ideally, only linked records should be available to the unit (such as those of suspicious individuals)

Privacy-preserving 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 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, cryptanalysis, etc.)
  - Have techniques that are scalable to linking large databases across multiple parties

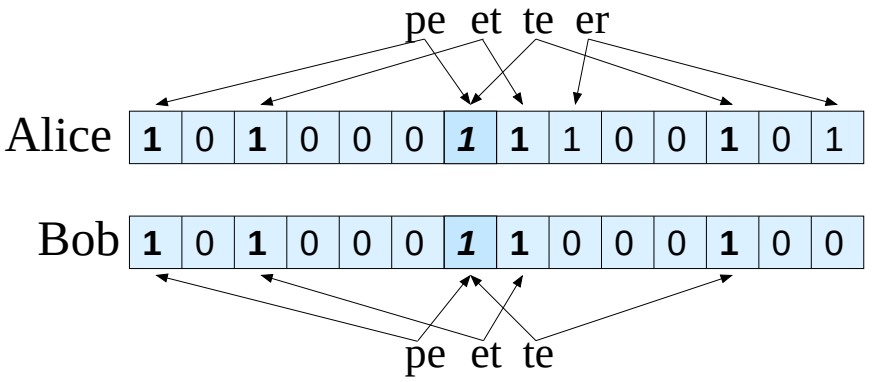
Hash-encoding for PPRL

- 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

Bloom filter encoding for PPRL

(Schnell et al., 2009)

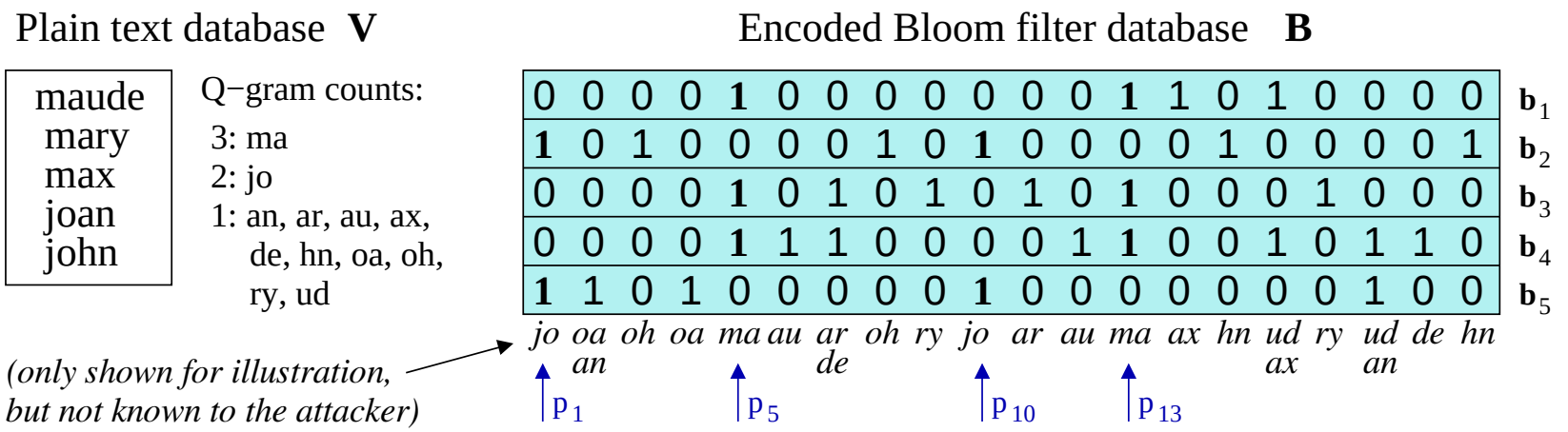


'peter': x_1=*7*, 'pete': x_1=*5*, *c*=*5*, therefore sim_{Dice} = 2×5/(7+5)= 10/12 = 0.83

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

Attacks on Bloom filter based PPRL

(Christen et al., 2017 and 2018)



- Several recent cryptanalysis attacks exploit the patterns and frequencies within and between Bloom filters
- Our most recent attack is successful even when each Bloom filter in an encoded database is unique
- *Hardening* techniques have been proposed, which modify Bloom filters to make them more resistant to attacks (at the costs of reduced linkage quality)

Data linkage evaluation

- Assuming ground truth data is available
  - **Precision**: How many true matches are in the set of classified matches? (aka *positive predictive value*)

$$P = \frac{number\ of\ true\ matching\ pairs}{number\ of\ classified\ matching\ pairs} = \frac{TP}{TP + FP}$$

  - **Recall**: How many true matches did we find from all known true matches? (aka *sensitivity*)

$$R = \frac{number\ of\ true\ matching\ pairs}{number\ of\ all\ true\ matching\ pairs} = \frac{TP}{TP + FN}$$

  - Often combined into the **F-measure** (or *F-score* or *F1*)

$$F1 = \frac{2 \times P \times R}{P + R}$$

The F-measure can be misleading!

(Hand and Christen, 2017)

- Traditionally the F-measure is seen as the harmonic mean of precision and recall
- We have shown that the F-measure is also a weighted arithmetic mean where recall is given weight *p* and precision weight *1-p*, with:

$$p = \frac{TP + FN}{FN + FP + 2 \times TP}$$

- The problem is that *p* depends upon the number of classified matches and non-matches

The measure being used to evaluate classification performance therefore depends on the thing being evaluated!

Dealing with uncertainty in linked data sets

- A majority of studies based on linked data sets ignore potential bias and have untested assumptions about the data
  - Information about the linkage process is rarely passed to analysts and considered in downstream processes
  - Only summary linkage quality measures are considered
  - Researchers using linked data sets often assume that linkage has not introduced any bias
- The choices made in every linkage step influence the quality and completeness of a linked data set

Limited work on dealing with uncertainty and bias

- A recent study evaluated linkage error and resulting bias in hospital records (Harron et al, 2014)
  - Two linkage methods: Highest pair-wise weight versus an imputation based approach (Goldstein et al., 2012)
  - Substantial bias when linkage error differed by hospitals, especially with lower match rates
- The Minnesota Population Center (IPUMS) aims to achieve high precision (Ruggles et al., 2017)
  - The bias of having missed matches can be measured and corrected
  - False matches can introduce systematic bias into many studies and is hard to detect

Outline

- A short introduction to data linkage
- Challenges of linking administrative databases

- Techniques for scalable data linkage
- Automating the linkage process
- Privacy aspects in data linkage
- Evaluating linkage quality and completeness
- Dealing with uncertainty in linked data sets

- Towards an end-to-end linkage framework
- Major research challenges

Towards an end-to-end data linkage framework

- Data linkage is only one step in a larger process

  (of data generation / collection, processing, storage, linkage, analysis, and dissemination)

- Data linkage is often done in isolation from the other steps
  - Often by different people / groups from the researchers who will make use of the linked data sets
  - Data linkage is commonly seen as a ‘black box’

Researchers who use linked data need to be knowledgeable about linkage methodologies and techniques, their potential limitations, and any potential bias they might introduce.

Organisational challenges to data linkage *(Harron et al., 2017)*

- Organisations move away from individual ‘ad-hoc’ linkage of two static data sets
- Linkage is increasingly done in an ongoing fashion, of many data sets
- Different consumers of linked data sets have different requirements
  - Some require linked data of high precision (low false match rate), others high recall (low missed match rate)
- Privacy concerns might limit the use of certain data sets for some linkage projects
- Recent initiatives to provide guidelines (Gilbert et al., 2017 – GUILD)

Major research challenges (1)

- Linkage techniques for massive-scale Big data collections (parallel, distributed, cloud based)
- Linking data across many organisations
  (computational challenges, as well as the issue of collusion between parties in PPRL)
- Linking dynamic data and linking data in real-time
  (dynamic indexing techniques and classification models)
- How to further automate the linkage process
  (automatic / semi-automatic data cleaning techniques)
- How to make use of additional information in data
  (not only attribute / field similarities like addresses and names, but also relationships, social network data, etc.)

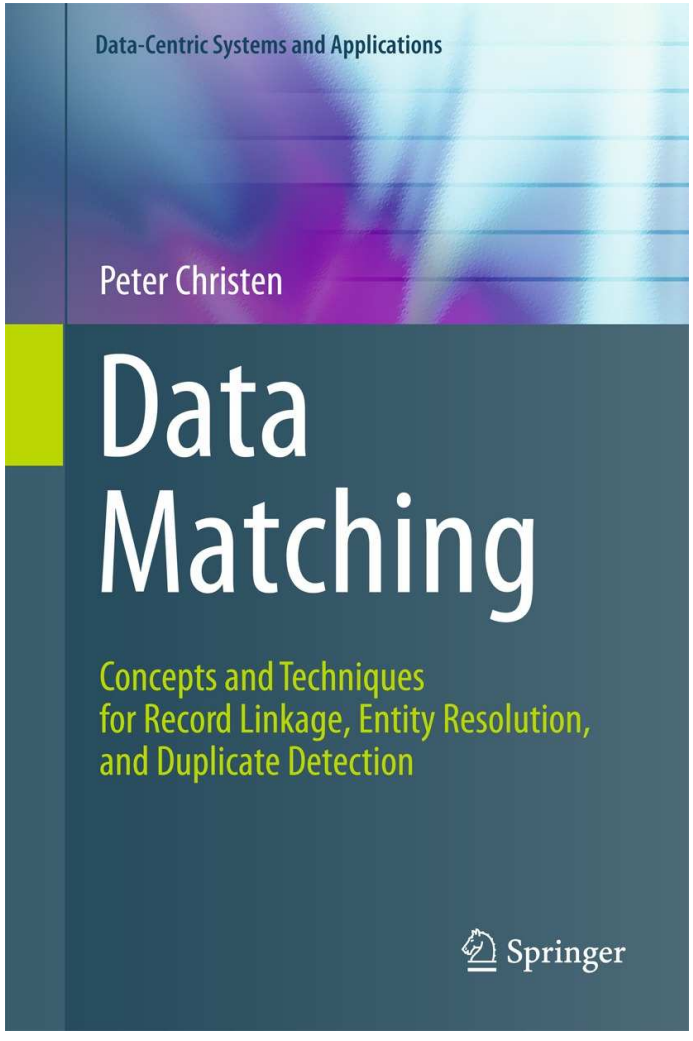
Major research challenges (2)

- No training data in most applications
  - Develop improved active learning approaches
  - Visualisation for improved manual clerical review
- Frameworks and toolboxes for data linkage to allow comparative experimental studies
- Publicly available test data collections
  - Challenging (impossible?) to have true match status
  - Challenging as most data are either proprietary or sensitive
- Pragmatic challenge: Collaborations across multiple research disciplines

Major research challenges – PPRL

- 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)

Advertisement: Book 'Data Matching'



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

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