Australian
National
University

Lessons from twenty years of working with (administrative) Big Data

Prof Peter Christen
School of Computing, The Australian National University

peter.christen@anu.edu.au

ICTer, Sri Lanka, November 2023

Short background

- Started to work in data mining in 1999, where my first project was to match and analyse Australian Medicare / pharmacy databases
- Collaboration with New South Wales (NSW) Health from 2002 until 2007 to develop record linkage methods, where we linked ten years of records of women giving birth in NSW
- Worked on record linkage with the Australian Taxation Office, the largest Australian credit bureau (real-time matching of customer records to find identity fraud), the Minnesota Population Centre, Fujitsu Japan (synthetic data generation), and many researchers
- Since 2014 working with UK researchers on linking large vital event databases to reconstruct the Scottish population

Many collaborators

Agus Pudjijono · Alan Dearle · Alexandros Karakasidis · Alice Reid · Anika Gross · *Anushka Vidanage* · *Asara Senaratne* · Banda Ramadan · Beata Nowok · Bill Winkler · *Charini Nanayakkara* · *Charith Perera* · Chris Dibben · Christian Borgs · Christine O’Keefe · Christopher Rost · David Hand · David Hawking · Denny · Dimitrios Karapiperis · *Dinusha Vatsalan* · Eilidh Garrett · Erhard Rahm · Felix Naumann · George Papadakis · Graham Kirby · Graham Williams · Grigorios Loukides · Huizhi Liang · James Doidge · Jeffrey Fisher · Karl Goiser · Katie Harron · Kee Siong Ng · Khoi-Nguyen Tran · Kim Lim · Lee Williamson · Lucas Lange · Mac Boot · Maja Schneider · Markus Hegland · Martin Franke · Minkyoung Kim · *Nishadi Kirielle* ·Özgür Akgün · Pouya Omran · Qing Wang · Rainer Schnell · Rohan Baxter · Ross Gayler · Sanjay Chawla · Scott Sanner · Sean Randall · Sirintra Vaiwsri · Solon Pissis · *Sumayya Ziyad* · Themis Palapanas · *Thilina Ranbaduge* · Tim Churches · Timothy de Vries · Tom Dalton · Uwe Draisbach · Vassilios Verykios · Victor Christen · Yichen Hu · Zhichun Fu · Ziad Sehili … *and more* ...

Outline

- An introduction to administrative data, data science, and Big data
- Some example data science studies for the social good
- A bit more about the data science process, data quality, data processing, data linkage (and analysis)
- Misconceptions that can occur when working with (personal) data
- Challenges when working with real-world (personal) data
- Lessons learnt and recommendations

What are administrative data?

Data science

- *“Data science is an interdisciplinary academic field that uses statistics, scientific computing, scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from noisy, structured, and unstructured data.”* [Wikipedia]
- Many data science projects use data about people, as collected by governments (tax payers, travellers, students, unemployed, patients, criminals, and so on) or by businesses (customers)
- Personal data are arguably some of the most valuable data
  (for example, the UK’s National Health Services (NHS) curated data are estimated to be worth as much as GBP 5 Billion (LKR 2,000 Billion) per year)

Administrative data

- In research, explicit data collection via surveys and experiments is increasingly replaced with the use of data collected for other purposes (and often by different organisations, such as governments and businesses)

- Such data are called *administrative data* – they are collected and used for some operational purpose (such as billing customers or tracking the medications given to patients) but not for data science

- Administrative data research is becoming more common in the health and social sciences. It is also known as *Population Data Science* (see journal: https://ijpds.org/)

Big data

- We know Big data to be data characterised by the four or five V's (volume, velocity, variety and veracity, and *value*).

- In the social and health sciences, as well as national statistical institutes, administrative data are also known as Big data (because such data are very much larger than most traditional survey or experimental data sets)

- This fundamental shift from 'small' to possibly very large, dynamic, and uncleaned data sets poses various challenges to researchers in the heath and social sciences – it also leads to misconceptions (we come back to that later)

Examples using administrative data for the social good

Example studies (1)

Deep vein thrombosis and air travel: record linkage study
C. Kelman et al., British Medical Journal, 2003

- Deep vein thrombosis (DVT) occurs when blood clots form that can travel through your body; these clots are life-threatening if they become lodged in your lung. Long travels (flights, bus or train) is one cause of DVT

- 5,408 hospital patients in Western Australia with DVT symptoms were matched with data for arrivals of international flights during the period 1981 and 1999

- Annual risk of DVT was shown to increase by 12% by one long flight per year

- That is why passengers are now encouraged to do exercises during long flights

Example studies (2)

Validation of a machine learning model to predict childhood lead poisoning E. Potash et al., JAMA Network, 2020

- Lead-based paint in older housing can result in elevated blood lead levels which can cause irreversible neurological damage in children (such as learning difficulties)

- A machine learning model was developed that used 2.5 Million blood tests, 70,000 public health investigations about lead, 2 Million building permits and violations; as well as age, size and condition of housing, and sociodemographic data from the US Census

- Compared to simpler earlier models, this more sophisticated approach correctly identified the children at highest risk of lead poisoning about twice the rate

Australian National University

Example studies (3)

- **Early life PM2.5 exposure, childhood cognitive ability and mortality between age 11 and 86: A record-linkage life-course study from Scotland** G. Baranyi et al., Environ. Res., 2023

- An analysis of records of almost 3,000 people born in 1936 showed that those exposed to significant air pollution in early childhood are up to 5% more likely to die early than those raised in areas with better air quality

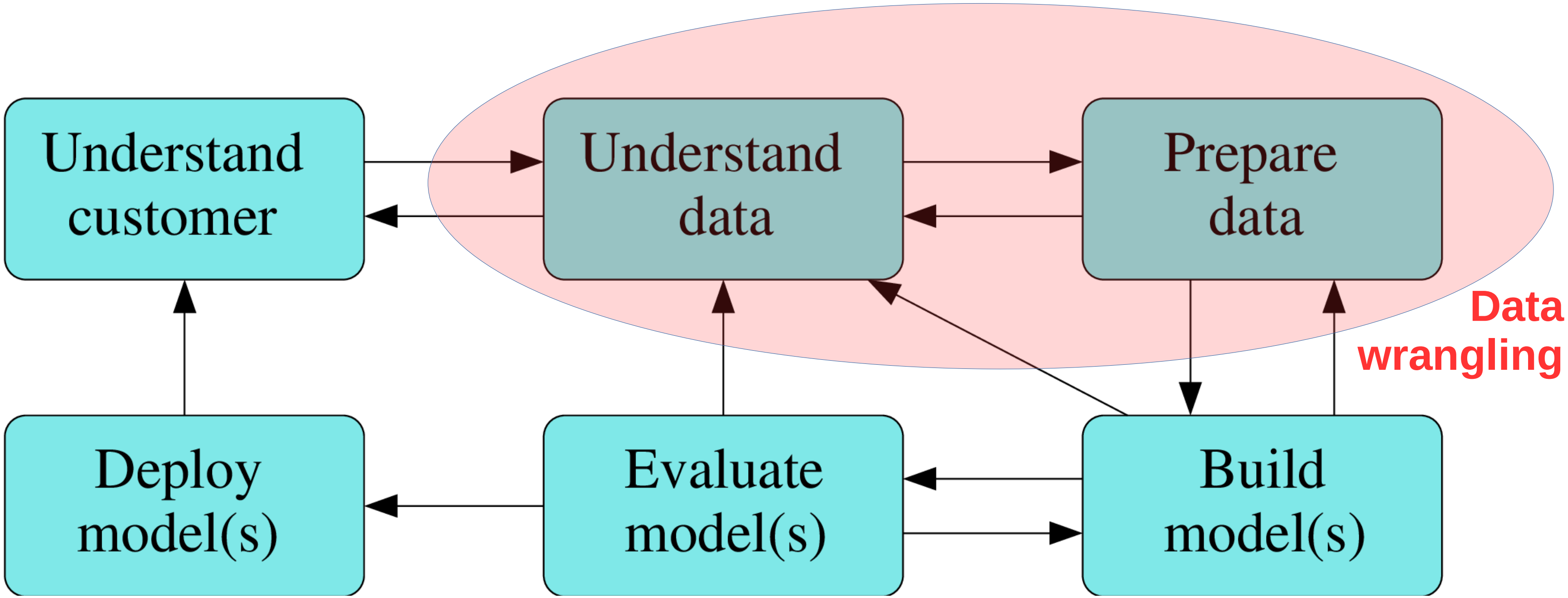
- Historical air pollution data was estimated using atmospheric chemistry models and matched to each participant’s home address in 1939

“*We are lucky, in Scotland, to have an increasing number of studies following people from childhood to old age. This is helping us to better understand what type of environments we need now to support healthy ageing in the future.*” – Chris Dibben, University of Edinburgh

Australian
National
University

A short overview of the data science process, data quality, data capturing, processing, and linkage

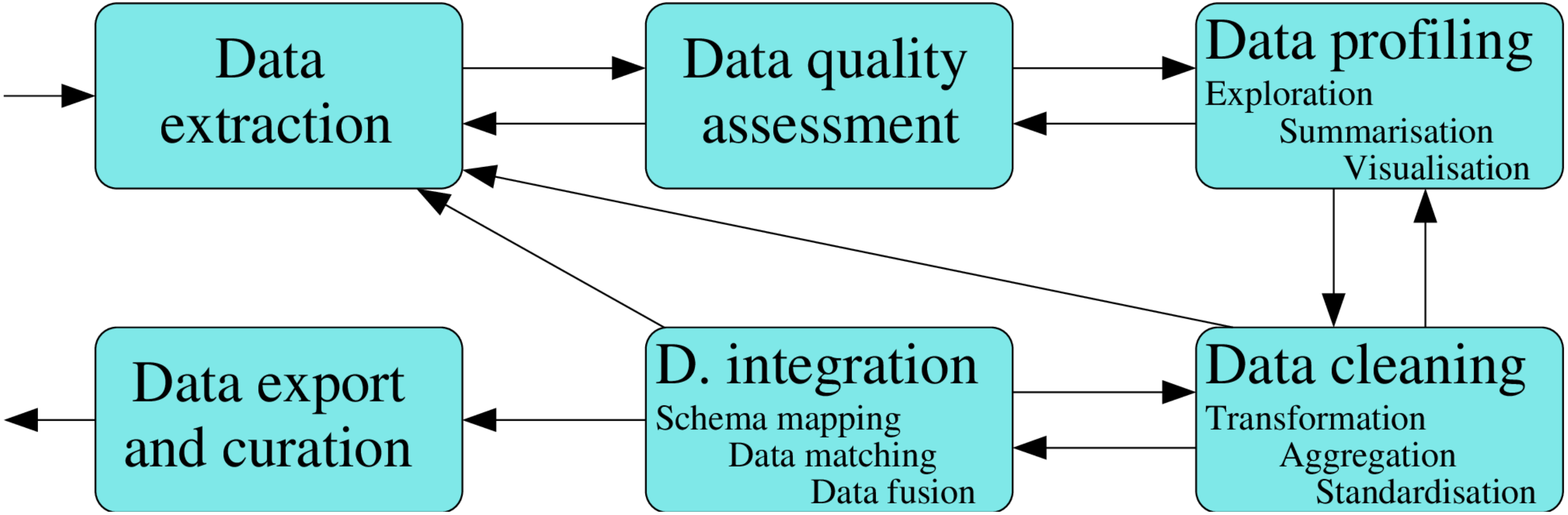
The data mining / science process



Typically up to 90% of time and effort are spent in the first three steps

(based on the *CRoss Industry Standard Process for Data Mining* (CRISP-DM), 1996)

The data wrangling process



Many of the challenges and misconceptions in the data science process occur within these steps

Data quality dimensions

Six core dimensions

Completeness: No missing data

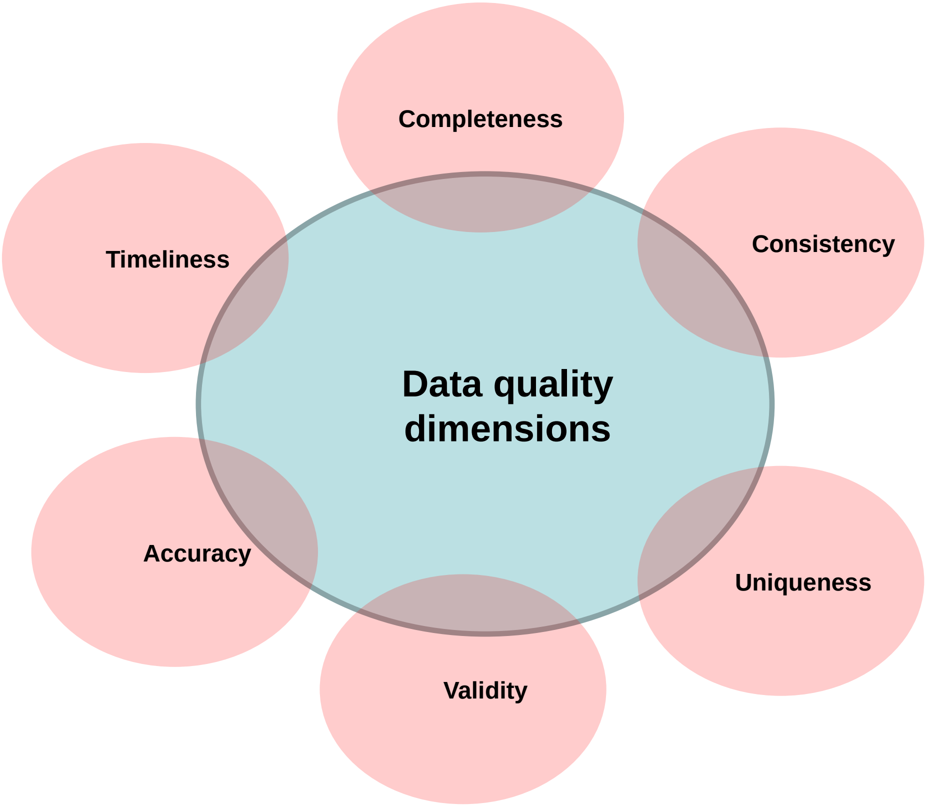
Consistency: Across different sources

Uniqueness: Single view of data

Validity: Meet constraints and rules

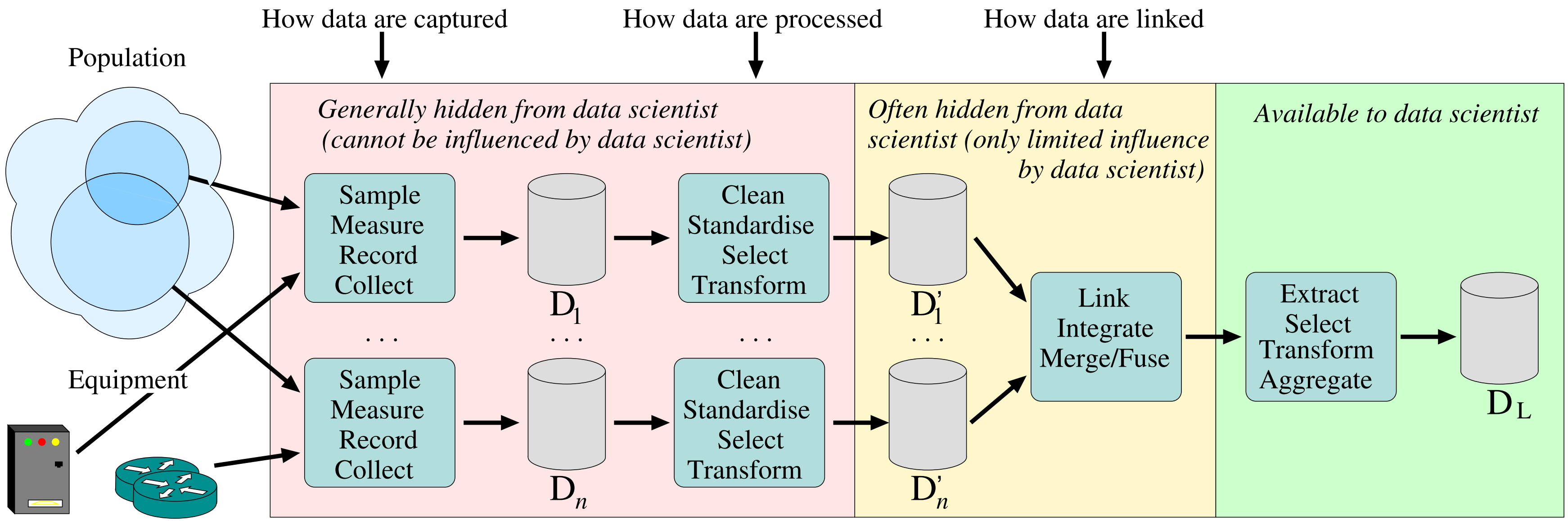
Accuracy: Correct and reflect reality

Timeliness: No out-of-date values



The six primary dimensions for data quality assessment, *DAMA UK Working Group*, 2013

How data gets onto a data scientist's computer


How data are captured
How data are processed
How data are linked
Population
Equipment
Generally hidden from data scientist (cannot be influenced by data scientist)
Often hidden from data scientist (only limited influence by data scientist)
Available to data scientist
Sample Measure Record Collect
. . .
Sample Measure Record Collect
D_1
. . .
D_n
Clean Standardise Select Transform
. . .
Clean Standardise Select Transform
D'_1
. . .
D'_n
Link Integrate Merge/Fuse
Extract Select Transform Aggregate
D_L

(adapted from: *Thirty-three myths and misconceptions about population data: from data capture and processing to linkage. P. Christen and R. Schnell, IJPDS, 2023)*

Big data is not (yet) the new oil: Misconceptions that can occur when working with (personal) data

Rapid adoption and inflated expectations

- Due to the perceived advantages of administrative data, the number of projects adopting existing databases for research and decision making is increasing

- The use of buzzwords like Big data, AI, and machine learning, in the context of administrative data seems to suggest for non-technical users and decision makers that any kind of question can be answered when analysing population databases

- Neither data quality issues (how data was captured and processed), nor the techniques used to link data, are clear to decision makers and researchers who are used to working with smaller data sets

Data quality and misconceptions of population data

- There is much work on general data quality, but only little specific to population data
- The misconceptions we consider here are usually underestimated by non-specialists, leading to inflated expectations
- Such over-expectations might cause costly mis-management in areas such as public health or in government decision making
- Failing administrative data projects might even result in the loss of trust in governments and science by the public
- Hence, misconceptions on population data need to be avoided

Types of misconceptions (1)

- We have (so far) identified over 30 misconceptions across the three stages (capturing, processing, linking), here are some examples:

- **Data capturing**:
  - A population database contains all individuals in a population
  - The population covered in a database is well defined
  - Each individual in a population is represented by a single record
  - Records in a population database always refer to real people
  - Certain personal details do not change over time
  - Errors in personal data are not intentional
  - Missing data have no meaning
  - Data values are in their correct attributes / fields
  - Automatically collected data are always correct, complete, and valid
  - Population data provide the same answers as survey data
  - Hardware and software used to capture data are error free

Types of misconceptions (2)

- **Data processing**:
  - Data processing can be fully automated
  - Data processing is always correct
  - Aggregated data are sufficient for research
  - Metadata are correct, complete, and up-to-date
  - Synthetically generated data fully reflect reality
  - Software used to process data is bug-free
- **Data linking**:
  - A linked data set corresponds to an actual population
  - Linked databases represent the conditions of individuals / entities at the same time
  - A linked data set contains no duplicates
  - A linked data set is unbiased
  - Attribute / field values in linked records are correct
  - Modern record linkage techniques can handle databases of any size.

Challenges when working with real-world (personal) data

(or why becoming a theoretical physicist might make your life as a researcher easier)

Australian National University

Challenge 1: Limited control over provenance of data

- Data scientists often have no or very limited control about the provenance of the data they are working with, and any processing done on these data
- They likely will also have only limited metadata that are needed to fully understand the characteristics and quality of their data
- When both the collection and processing of data are outside the control of a data scientist, conducting proper research can become challenging

“In science, three things matter: the data, the methods used to collect the data (which give them their probative value), and the logic connecting the data and methods to conclusions.” – Andrew Brown et al., Proceedings of the National Academy of Sciences, 2018

Challenge 2: Spurious results

- Most machine learning and AI methods will produce results on any input data (for example, k-means will generate *k* clusters)
- This does not mean these results are meaningful
- Administrative data provide different results from survey data (the former are about what people are and what they do, the latter about their attitudes, beliefs, expectations, or intentions)

“The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.” – John Tukey, The American Statistician, 1986

“One of the lessons that we have learnt from data mining practice over the past 20 years is that most of the unusual structures in large data sets arise from data errors, rather than anything of intrinsic interest.” – David Hand, Statistics in Society, 2018

Challenge 3: Accessing personal data is often difficult

- Personal data can be highly sensitive and are covered by privacy regulations (such as the EU and UK GDPR, or the US HIPAA)
- Accessing such data generally requires extensive information governance (access and confidentiality agreements, secure and trusted computing environments, data anonymisation, and so on)
- These can be months if not years to set up, often delaying data access and therefore research
- Frameworks such as the Five Safes are aimed to support access to sensitive data (safe data, safe projects, safe people, safe settings, safe outputs)
- Increasingly, synthetic data are used for training, education, and software development and testing

Australian National University

Lessons learnt and recommendations

Lessons learnt and recommendations

- Obtaining access to real-world (personal) data always takes much longer than anticipated
- Real-world data are dirty, so never trust any data you are given (do thorough data exploration, profiling, and data quality assessment)
- Aim to obtain detailed metadata (and try to talk to the data owner about how data were captured and processed)
- Data sets have limitations in what questions can be answered by them
- Always carefully assess and question any results from your machine learning or AI algorithms (results of 100% accuracy are likely incorrect)
- Use a suitable performance evaluation measure (accuracy, F-score, etc.)
- But don’t give up (and don’t become a theoretical physicist), there is much (good) data science can contribute to the challenging world we live in

Lessons from twenty years of working with (administrative) Big Data

Prof Peter Christen
School of Computing, The Australian National University

peter.christen@anu.edu.au

ICTer, Sri Lanka, November 2023