# *Data, COVID, and LSD*

## *(or how LSD can help combat COVID-19)*

**Peter Christen**

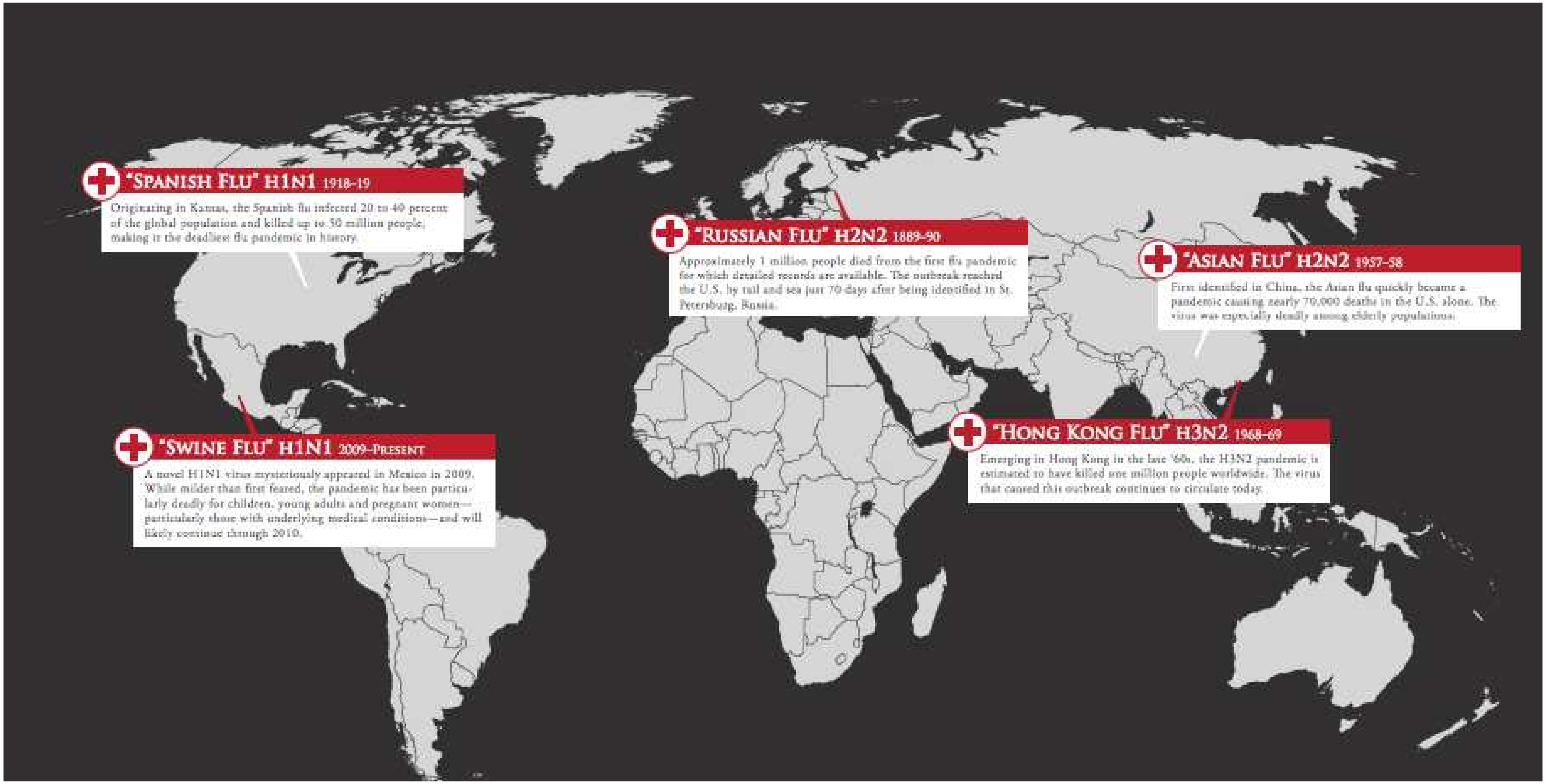
**Leibniz Professor, University of Leipzig**

**Professor, School of Computing,**
**College of Engineering and Computer Science,**
**The Australian National University, Canberra**

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

Homepage: **`http://users.cecs.anu.edu.au/~christen/`**

# Motivating example: A pandemic



"SPANISH FLU" H1N1 1918-19

Originating in Kansas, the Spanish flu infected 20 to 40 percent of the global population and killed up to 50 million people, making it the deadliest flu pandemic in history.

"RUSSIAN FLU" H2N2 1889-90

Approximately 1 million people died from the first flu pandemic for which detailed records are available. The outbreak reached the U.S. by rail and sea just 70 days after being identified in St. Petersburg, Russia.

"ASIAN FLU" H2N2 1957-58

First identified in China, the Asian flu quickly became a pandemic causing nearly 70,000 deaths in the U.S. alone. The virus was especially deadly among elderly populations.

"SWINE FLU" H1N1 2009-PRESENT

A novel H1N1 virus mysteriously appeared in Mexico in 2009. While milder than first feared, the pandemic has been particularly deadly for children, young adults and pregnant women—particularly those with underlying medical conditions—and will likely continue through 2010.

"HONG KONG FLU" H3N2 1968-69

Emerging in Hong Kong in the late '60s, the H3N2 pandemic is estimated to have killed one million people worldwide. The virus that caused this outbreak continues to circulate today.

# *Motivating example: Data science*

- Understanding (or even preventing) the outbreak of a pandemic requires identifying 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)
- Such data sets are **large**, **dynamic**, **complex**, **heterogeneous**, and **distributed**
- Privacy and confidentiality concerns arise if such data are stored and linked at a central location

# *Motivating example: LSD*

- To tackle complex issues such as a global pandemic, we need to be able to **integrate** and **link large** and **complex**, and highly **sensitive** and **confidential** databases in near **real time**
- **Linking Sensitive Data** (LSD) is concerned with the development of methods, techniques, algorithms, and processes to achieve this goal
- Besides being a crucial tool in understanding a pandemic, LSD has applications in a variety of different domains (ranging from the health and social sciences to national censuses, crime and fraud detection, and national security)

# *I acknowledge all my LSD colleagues*

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

# *Outline*

- History and application examples for LSD
  - A brief history of linking databases
  - Modern applications of where LSD is required
  - *SNAPS*, the Scotland Family Pedigree Search Tool
  - Challenges for LSD
- Some key technologies used for LSD
  - How to compare names and addresses?
  - How to encrypt and encode sensitive data?
  - How to compare encrypted and encoded data?
  - Privacy-preserving record linkage
- Conclusions and research directions
  - So can LSD help combat COVID-19?

# *Linking data is nothing new...*

‘Linking’ London underground tickets to conduct traffic analysis in 1936.

© Transport for London from the
London Transport Museum collection

# *The book of life (Halbert Dunn, 1946)*

- The idea of creating a book of life for each individual by linking records from birth, marriage, and death certificates, as well as records about individuals from the health and social security systems
- Each such book would start with a birth and end with a death record

- Dunn saw that linked records can provide a wealth of of information that is not available otherwise
- He also realised the challenges of **data quality**, large **volumes of data**, and **sensitive** personal data

# ***Computer-based record linkage***

- Computer assisted record linkage goes back as far as the 1950s (based on ad-hoc heuristic methods)
- Theoretical foundation for *probabilistic record linkage* by statisticians *Fellegi & Sunter* (1969)
  - No unique entity identifiers available (no person numbers or patient identifiers)
  - Compare names, addresses, dates of birth, and so on
  - Assign different importance to different such fields (same name is more important than same gender)
  - Classify a compared record pair as a *match*, a *non-match*, or a *potential match*
- Still the basis of many record linkage systems

# *Enter computer science...*

- Strong interest in the last two decades from computer science (from 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 very large databases and improving linkage quality
  - Blocking techniques to efficiently and effectively generate candidate record pairs
  - Machine learning-based classification techniques
- Development of privacy-preserving record linkage techniques

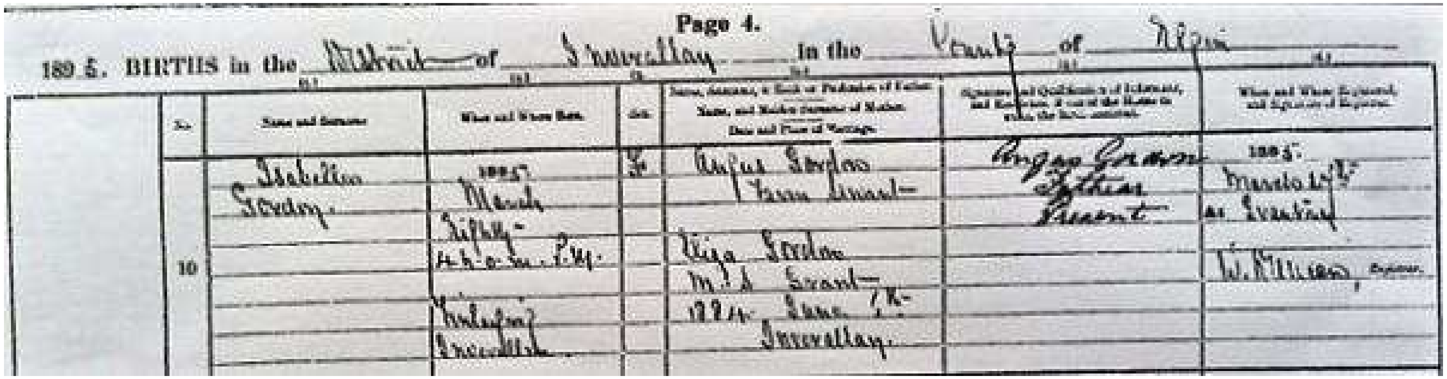
# *Applications of record linkage*

- Remove duplicates in one data set (deduplication)
- Merge new records into a larger master data set
- Create patient or customer oriented statistics
  (for example for longitudinal studies)
- Clean and enrich data for analysis and mining
- Geocode matching (to facilitate spatial data analysis)
- Widespread use of record linkage
  - Health and social science research
  - Immigration, taxation, social security, national censuses
  - Business mailing lists, consumer product matching
  - Crime and fraud detection, and terrorism intelligence

# *Application example: SNAPS*

- The *Digitising Scotland* project has transcribed 26 million birth, death, and marriage certificates from 1855 to 1973
- We now aim to reconstruct the full Scottish population to provide a unique data set that will facilitate a multitude of research studies
  - Social mobility, education, fertility, and family changes
  - Mortality during the 1918 flu pandemic
- The *Genetics Genealogy Team* of *Public Health Scotland* investigates heritability and genetic patterns of diseases for patients with familial cancer or other inherited genetic conditions

# Challenges of (historical) data



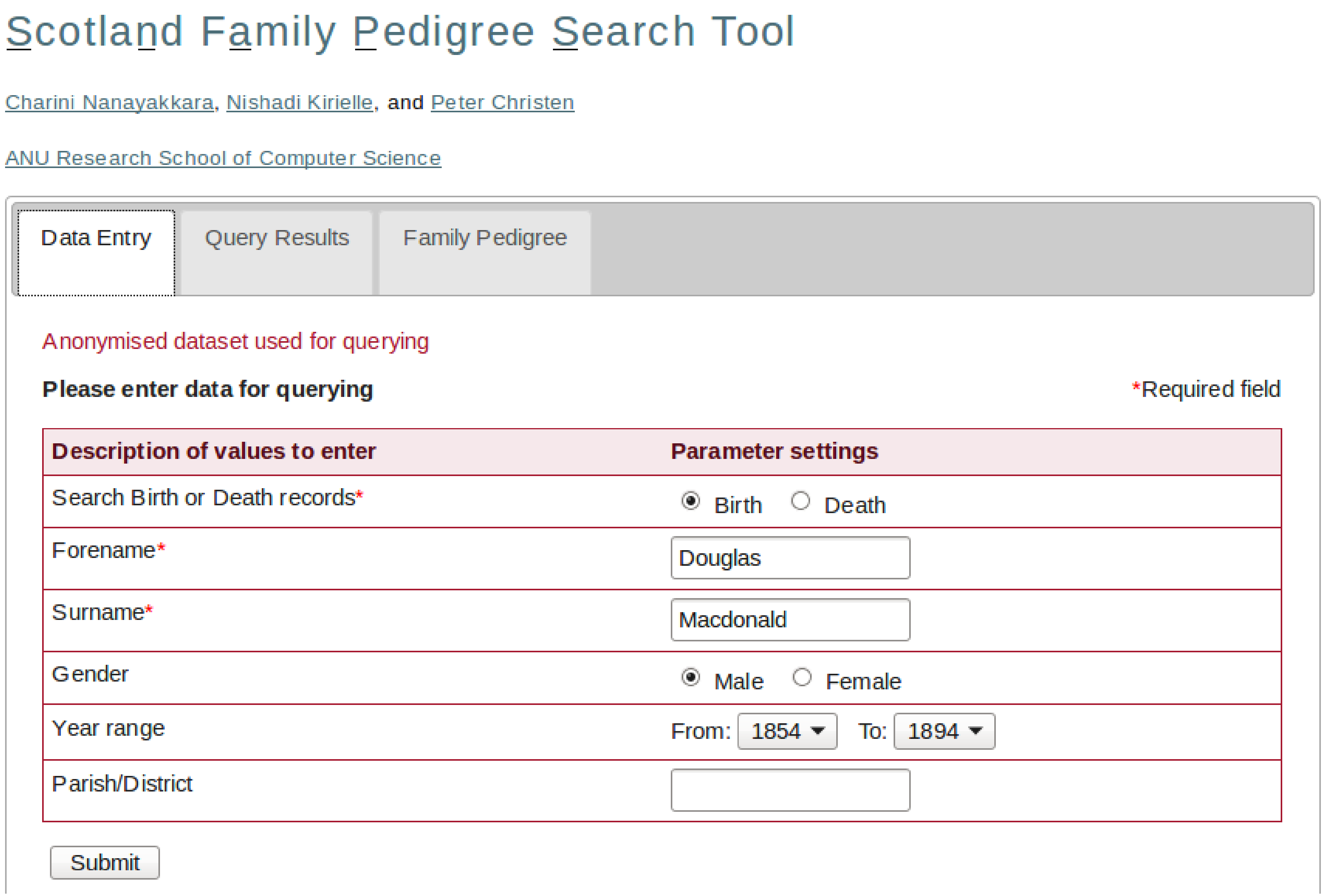
Page 4.

1885. BIRTHS in the District of Inverallan in the County of Elgin

| | No. | Name and Surname | When and Where Born | Sex | Name, Surname, & Rank or Profession of Father.<br>Name, and Maiden Surname of Mother.<br>Date and Place of Marriage. | Signature and Qualification of Informant, and Residence, if out of the House in which the Birth occurred. | When and Where Registered, and Signature of Registrar. |
|---|---|---|---|---|---|---|---|
| | 10 | Isabella<br>Gordon. | 1885<br>March<br>Fifth<br>4h 0m P.M.<br>[illegible]<br>[illegible] | F | Angus Gordon<br>Farm Servant<br>Eliza Gordon<br>M.S. Grant<br>1884 June 7th<br>Inverallan. | Angus Gordon<br>Father<br>Present | 1885.<br>March 16th<br>at Inverallan<br>W. Williams, Registrar. |

- 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, optical character recognition (OCR), and transcription errors

# SNAPS: Querying a person



## Scotland Family Pedigree Search Tool

Charini Nanayakkara, Nishadi Kirielle, and Peter Christen

ANU Research School of Computer Science

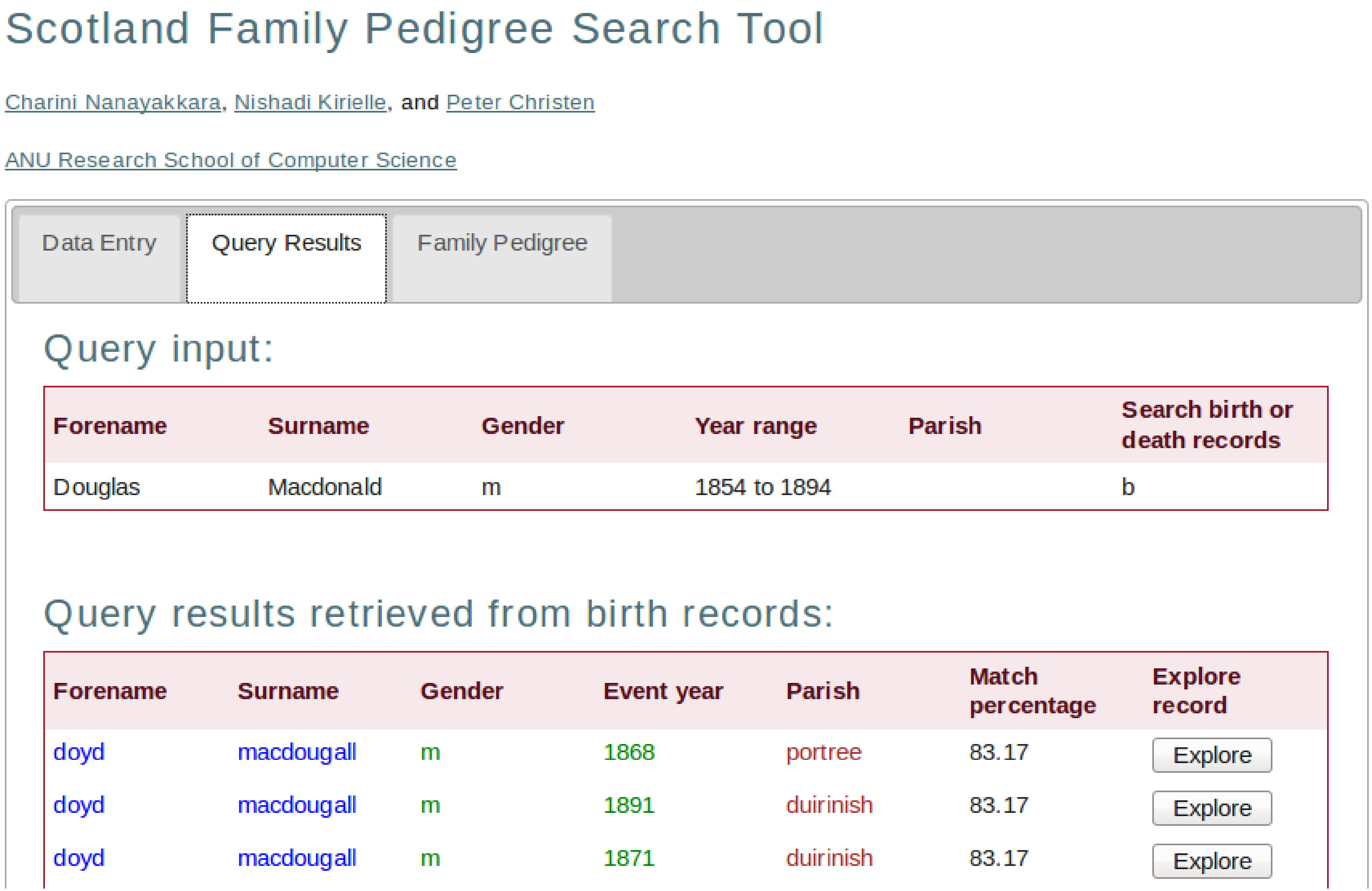
Anonymised dataset used for querying

**Please enter data for querying**

*Required field

| Description of values to enter | Parameter settings |
|---|---|
| Search Birth or Death records* | ● Birth ○ Death |
| Forename* | Douglas |
| Surname* | Macdonald |
| Gender | ● Male ○ Female |
| Year range | From: 1854 To: 1894 |
| Parish/District | |

# *SNAPS: Query results*



## Scotland Family Pedigree Search Tool

Charini Nanayakkara, Nishadi Kirielle, and Peter Christen

ANU Research School of Computer Science

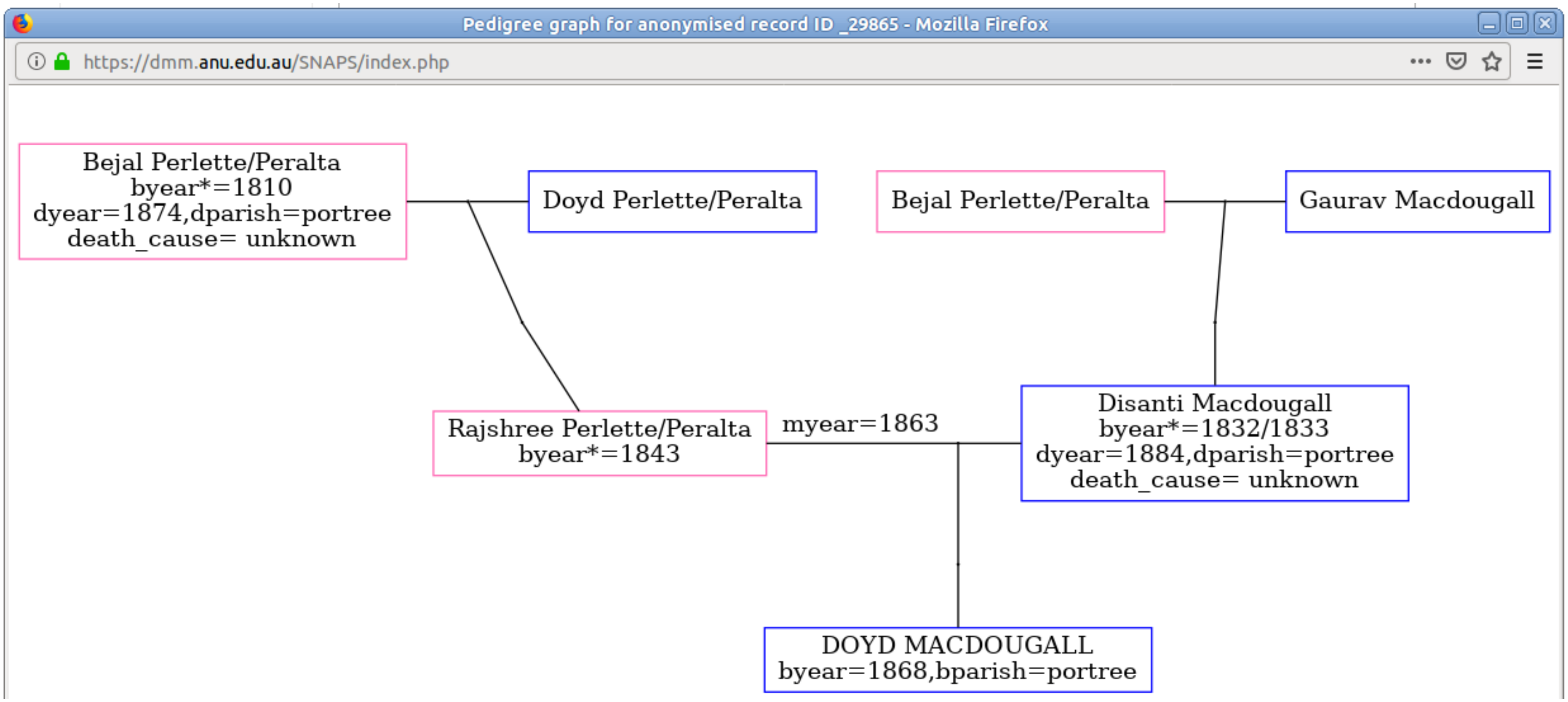
### Query input:

| Forename | Surname | Gender | Year range | Parish | Search birth or death records |
|---|---|---|---|---|---|
| Douglas | Macdonald | m | 1854 to 1894 | | b |

### Query results retrieved from birth records:

| Forename | Surname | Gender | Event year | Parish | Match percentage | Explore record |
|---|---|---|---|---|---|---|
| doyd | macdougall | m | 1868 | portree | 83.17 | Explore |
| doyd | macdougall | m | 1891 | duirinish | 83.17 | Explore |
| doyd | macdougall | m | 1871 | duirinish | 83.17 | Explore |

# *SNAPS: A generated family tree*


Pedigree graph for anonymised record ID _29865 - Mozilla Firefox
https://dmm.anu.edu.au/SNAPS/index.php
Bejal Perlette/Peralta
byear*=1810
dyear=1874,dparish=portree
death_cause= unknown
Doyd Perlette/Peralta
Bejal Perlette/Peralta
Gaurav Macdougall
Rajshree Perlette/Peralta
byear*=1843
myear=1863
Disanti Macdougall
byear*=1832/1833
dyear=1884,dparish=portree
death_cause= unknown
DOYD MACDOUGALL
byear=1868,bparish=portree

- Based on real historical Scottish data from the Isle of Skye, however with names and dates anonymised
- See demo at: **`https://dmm.anu.edu.au/SNAPS/`**

# *Major challenges when linking data*

- No unique entity identifiers are available
- Real world data are dirty

  (typographical errors and variations, missing and outdated values, and various other data quality issues)
- Scalability to linking large databases
  - Naive comparison of all record pairs does not scale
- No ground truth data (gold standard) in many linkage applications
  - No record pairs with known true match status
- Privacy and confidentiality

  (because personal information, such as names and addresses, are commonly required for linking)

# *Outline*

- History and application examples for LSD
  - A brief history of linking databases
  - Modern applications of where LSD is required
  - *SNAPS*, the Scotland Family Pedigree Search Tool
  - Challenges for LSD
- Some key technologies used for LSD
  - How to compare names and addresses?
  - How to encrypt and encode sensitive data?
  - How to compare encrypted and encoded data?
  - Privacy-preserving record linkage
- Conclusions and research directions
  - So can LSD help combat COVID-19?

# *Comparing names and addresses*

- A key requirement to achieve high linkage quality
- Personal data is prone to errors and variations
  - Scanned, hand-written, over telephone, hand-typed
  - Different correct spelling variations for proper names (*Christopher*, *Kristopher*, *Christoffer*, *Christophir*, *Christoph*, *Kristoffe*, *Christophe*, and many more..)
  - Nicknames (*Tash* for *Natasha* or *Tosh* for *Macintosh*)
  - Fake values (my phone number is often *04 1234 5678*)
- Changes occur over time (names can change due to marriage and addresses when people move)
- All these mean exact comparison of names and addresses will not give good linkage results

# Approximate name comparison

- Aim: Compare two names (or addresses) and calculate a numerical similarity between 0 and 1
  - Comparing a name with itself gives a similarity of 1 (compare *Peter* with *Peter*)
  - Comparing completely different names gives a similarity of 0  (compare *Peter* with *David*)
  - Comparing somewhat similar names gives a similarity between 0 and 1  (compare *Peter* with *Petros*)
- Many different techniques have been developed, some specific to names, others for more general text  (comparing text is a fundamental aspect in many applications, such as Web search and spell checkers)

# *Q-gram based name comparison*

- Convert a name into *q-grams* (segments of length *q*)
  - For example, for $q = 2$: *peter* → [**pe**, **et**, te, er]<br>*petros* → [**pe**, **et**, tr, ro, os]
- Find how many q-grams are common between two names (for our example, two: [pe, et])
- Calculate a similarity, for example using the Sørensen-Dice coefficient (developed by botanists in the 1940s to calculate the similarity between plant communities)

$$sim = 2 \cdot 2 \, / \, (4 + 5) = 4 \, / \, 9 = 0.44$$

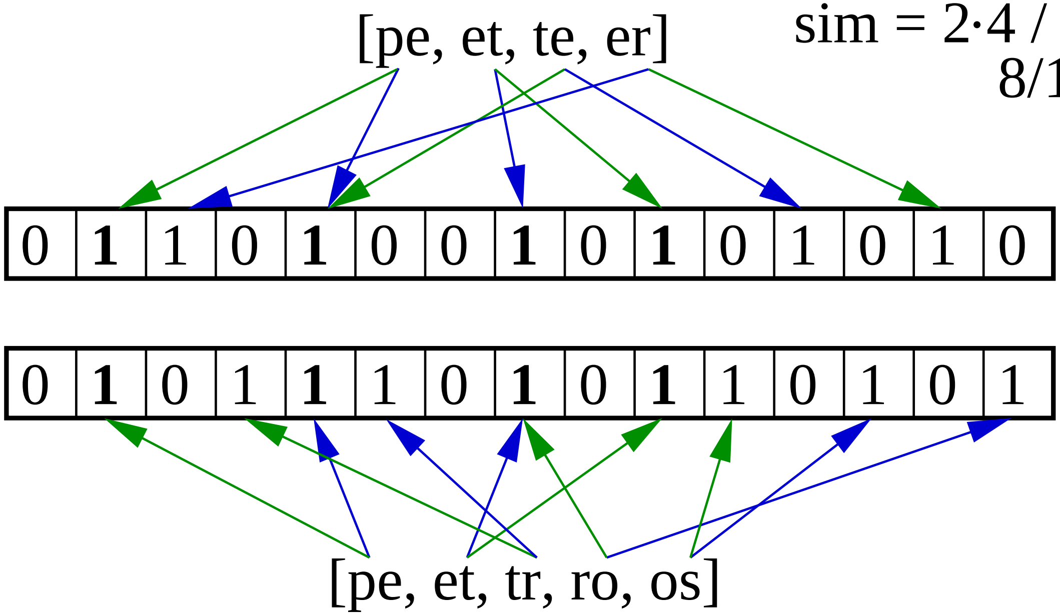
- The more q-grams two names have in common the more similar they are

# *Encoding sensitive data*

- We cannot share sensitive data (such as personal information) between organisations
- We need to encode and/or encrypt sensitive data
- We employ techniques such as those used for secure Internet communication (like online banking)
- One key technique is *secure one-way hashing*
  - A function that converts an input into a hash code
  - If we only have the code then it is almost impossible to obtain the input
  - For example: *peter* → *4R#x+Y4i9!e@t4o]W*
    *petros* → *Z5%o-(7Tq1g?7iE/#*
- But this only allows for exact matching!

# *Bloom filter based encoding*

- Bloom filters were developed in 1970; their use for LSD was proposed by Rainer Schnell in 2009
- We *map* q-grams into bit arrays (0s and 1s) where the number of common 1-bits approximates the similarity



[pe, et, te, er]

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | **1** | 1 | 0 | **1** | 0 | 0 | **1** | 0 | **1** | 0 | 1 | 0 | 1 | 0 |
| 0 | **1** | 0 | 1 | **1** | 1 | 0 | **1** | 0 | **1** | 1 | 0 | 1 | 0 | 1 |

[pe, et, tr, ro, os]

sim = 2·4 / (7+9) = 8/16 = 0.5

- Basic Bloom filters can be susceptible to attacks aimed at re-identifying sensitive values

# Privacy-preserving record linkage

- We aim to link records in databases across organisations without revealing any sensitive data
- We require techniques that:
  - Allow for approximate matching and high linkage quality
  - Are provably secure (cannot be attacked) and do not allow the re-identification of encoded sensitive values
  - Are scalable to linking very large databases
- An active area of research since the mid 1990s
  - Contributions from computer science, statistics, as well as the health and social sciences
  - Besides the health domain, there is increasing interest by governments (such as for national censuses and digital vaccination passports)

# *Outline*

- History and application examples for LSD
  - A brief history of linking databases
  - Modern applications of where LSD is required
  - *SNAPS*, the Scotland Family Pedigree Search Tool
  - Challenges for LSD
- Some key technologies used for LSD
  - How to compare names and addresses?
  - How to encrypt and encode sensitive data?
  - How to compare encrypted and encoded data?
  - Privacy-preserving record linkage
- Conclusions and research directions
  - So can LSD help combat COVID-19?

# *Conclusions and research directions*

- The technical building blocks for LSD exist, and we can now link large sensitive databases in privacy-preserving ways
- There are various open questions and challenges
  - How do we securely link new types of data, such as images, biometrics, or genetic data?
  - How do we evaluate a linkage if only encoded or encrypted values are available?
  - How do we prove our techniques are secure and cannot be attacked? Who are the adversaries?
  - How do we measure and formalise privacy?
  - How to do real-time linking of very large and dynamic databases?

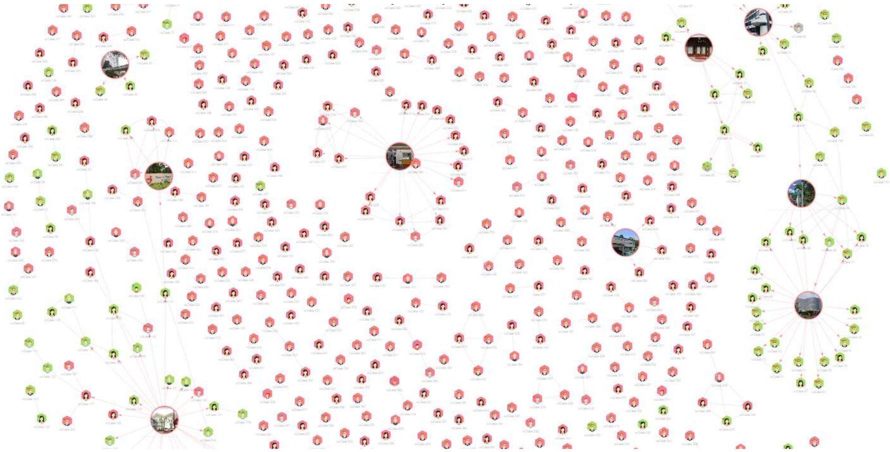
# *So can LSD help combat COVID-19?*

- Techniques developed for LSD are clearly crucial to facilitate advanced analysis of databases related to COVID-19
  - Such databases are held by different organisations (both public and private)
  - Only by linking them can we collect and combine the information we need to understand a pandemic
- The techniques exist, however most are still at the stage of research prototypes
- A lack of understanding of and trust in advanced techniques, and concerns about violating privacy regulations hinder the application of LSD techniques

# *Yes – LSD can help combat COVID-19*

- For ‘traditional’ linkage of sensitive databases from diverse sources for health research
- To collect encoded data in anonymised ways for digital vaccination passports
- To support (semi-) automatic contact tracing

Example from Singapore:

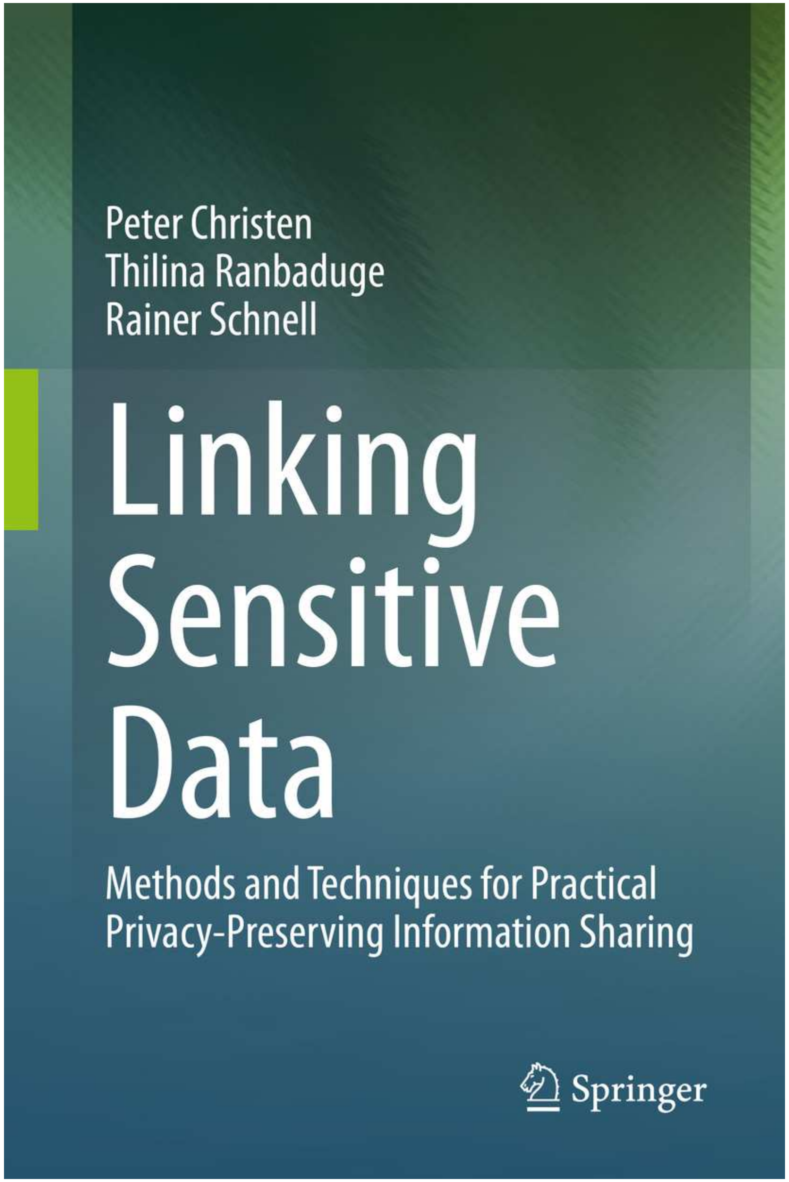


*The more we know and the quicker we know it, the better the chance to minimise the damages.*

Source: `https://op.europa.eu/en/web/eudatathon/covid-19-linked-data`

# ‘Linking Sensitive Data’ book

## (Springer, 2020)



Peter Christen
Thilina Ranbaduge
Rainer Schnell

Linking Sensitive Data

Methods and Techniques for Practical Privacy-Preserving Information Sharing

Springer

*The Book describes how linkage methods work and how to evaluate their performance. It covers all the major concepts and methods and also discusses practical matters such as computational efficiency, which are critical if the methods are to be used in practice – and it does all this in a highly accessible way!*

Prof David J. Hand OBE,
Imperial College, London