Australian National University

COMP3430 / 8430
Data wrangling

Lecture 12: Overview of record linkage
(Lecturer: Peter Christen)

Lecture outline

- What is record linkage
- Record linkage applications
- A short history of record linkage
- The record linkage process
- Record linkage techniques and challenges

What is record linkage?

- The process of linking records that represent the same entity in one or more databases (patients, customers, businesses, products, publications, etc.)
- Also known as *data linkage*, *data matching*, *entity resolution*, *duplicate detection*, *object identification*, 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 not stable)

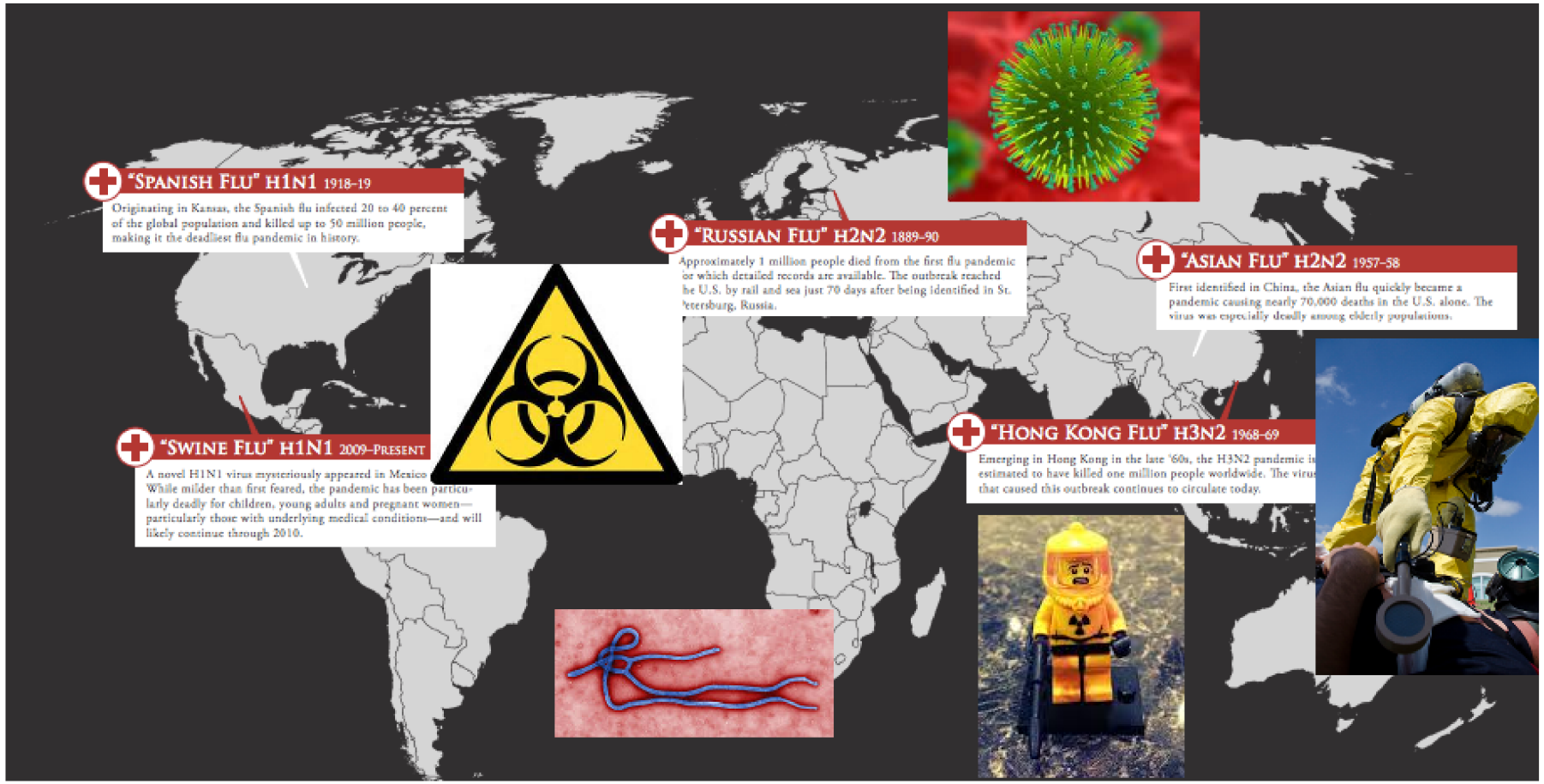
- For example, which of these records represent the same person?

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

Record linkage applications

- Remove duplicates in one data set (deduplication)
- Merge new records into a larger master data set
- Create patient or customer oriented statistics (for example for longitudinal studies)
- Clean and enrich data for analysis and mining
- Geocode matching (with reference address data)
- Widespread use of record linkage:
  - Immigration, taxation, social security, census
  - Fraud and crime detection, and national security
  - Business mailing lists, exchange of customer data
  - Health and social science research

Example application: Health outbreak (1)

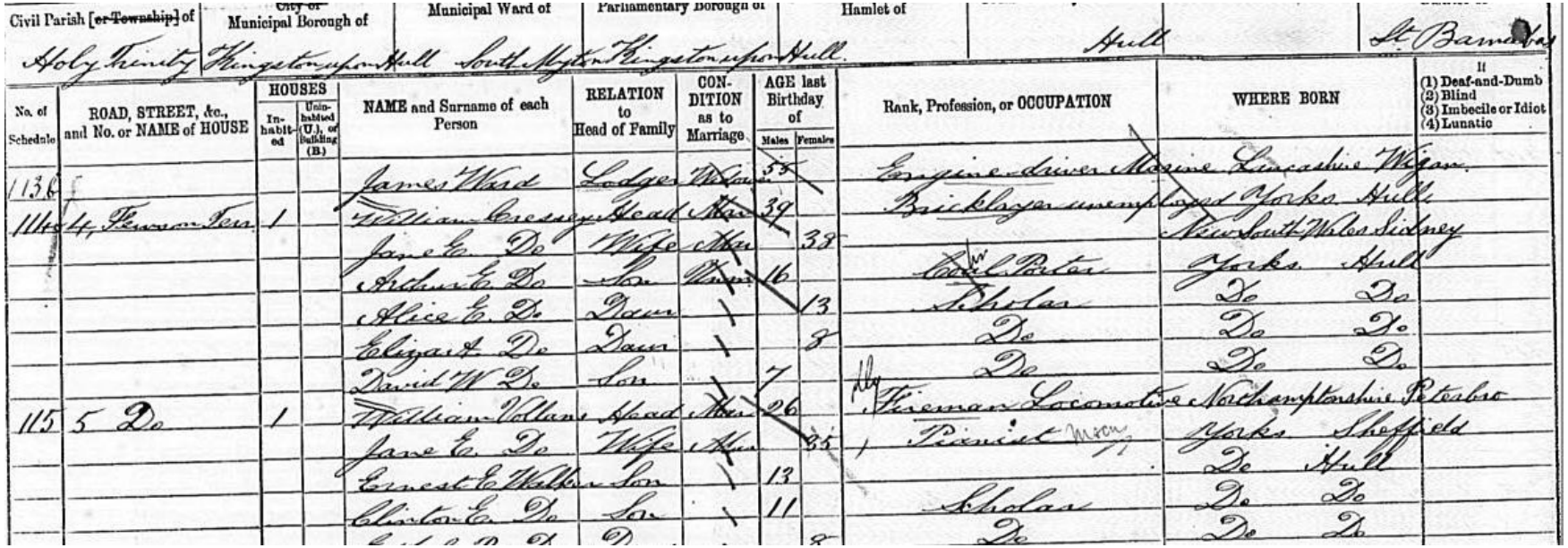


Example application: Health outbreak (2)

- Preventing the outbreak of epidemics requires monitoring of occurrences of unusual patterns of symptoms
- Data from many different sources will need to be collected and linked (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

Example application: Census data (1)

- Linking (historical) census data over time can help unlock a wealth of information about a society (such as changes in family structure, health, education, fertility, immigration, etc.)



Civil Parish [~~or Township~~] of: Holy Trinity | Municipal Borough of: Kingston upon Hull | Municipal Ward of: South Myton | Parliamentary Borough of: Kingston upon Hull | Hamlet of: Hull | St Barnabas

No. of Schedule	ROAD, STREET, &c., and No. or NAME of HOUSE	HOUSES Inhabited	HOUSES Uninhabited (U.), or Building (B.)	NAME and Surname of each Person	RELATION to Head of Family	CONDITION as to Marriage	AGE last Birthday of Males	AGE last Birthday of Females	Rank, Profession, or OCCUPATION	WHERE BORN	If (1) Deaf-and-Dumb (2) Blind (3) Imbecile or Idiot (4) Lunatic
113				James Ward	Lodger	~~Wid~~	55		Engine driver Marine	Lincolnshire Wigan	
114	4 Pearson Terr	1		William Creasey	Head	~~Mar~~	39		Bricklayer unemployed	Lincs Yorks Hull	
				Jane E. Do	Wife	~~Mar~~		38		New South Wales Sidney	
				Arthur E. Do	Son	~~Wid~~	16		Cabinet Painter	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 Adams	Head	~~Mar~~	36		Fireman Locomotive	Northamptonshire Peterboro	
				Jane E. Do	Wife	~~Mar~~		35	Pianist Mom	Yorks Sheffield	
				Ernest E. Walker	Son		13			Do Hull	
				Clinton E. Do	Son		11		Scholar	Do Do	
									Do	Do Do	

Example application: Census data (2)

- Many challenges to linking such historical data:
  - 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 (optical character recognition), and transcription errors
  - Lost pages (fire, eaten by mice, etc.)
- For modern data, privacy and confidentiality are major challenges
  - Because these data are about living people, and so privacy is of major concern when data are linked between organisations

Increased interest in record linkage (1)

ATO Data Matching – What is it and will it affect me?

You may have heard that ATO data matching technology is now being used in Australia. The ATO launched the system to shine a light on people who have undeclare returns, opening a new opportunity to catch people out and collect may understate their income.

What is ATO Data Matching?

The ATO data matching might best be described as a two-step proc

1. The ATO collects information about taxpayers. This includes the ATO pulling-in data from…
your employers,

Centrelink's controversial data matching program to target pensioners and disabled, Labor calls for suspension

By political reporter Henry Belot
Updated 17 Jan 2017, 7:01pm

The Federal Government will expand Centrelink's automatic debt recovery program later this year to focus on aged pensioners and disability support payments.

Parliamentary Budget Office charts reveal the Government plans to use a similar data-matching program to save nearly $1.5 billion over four years.

Increased interest in record linkage (2)

- Traditionally, record linkage has been used in statistics (census) and health (epidemiology)
  - First computer based techniques developed in 1960s
- In recent years, much 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 record 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

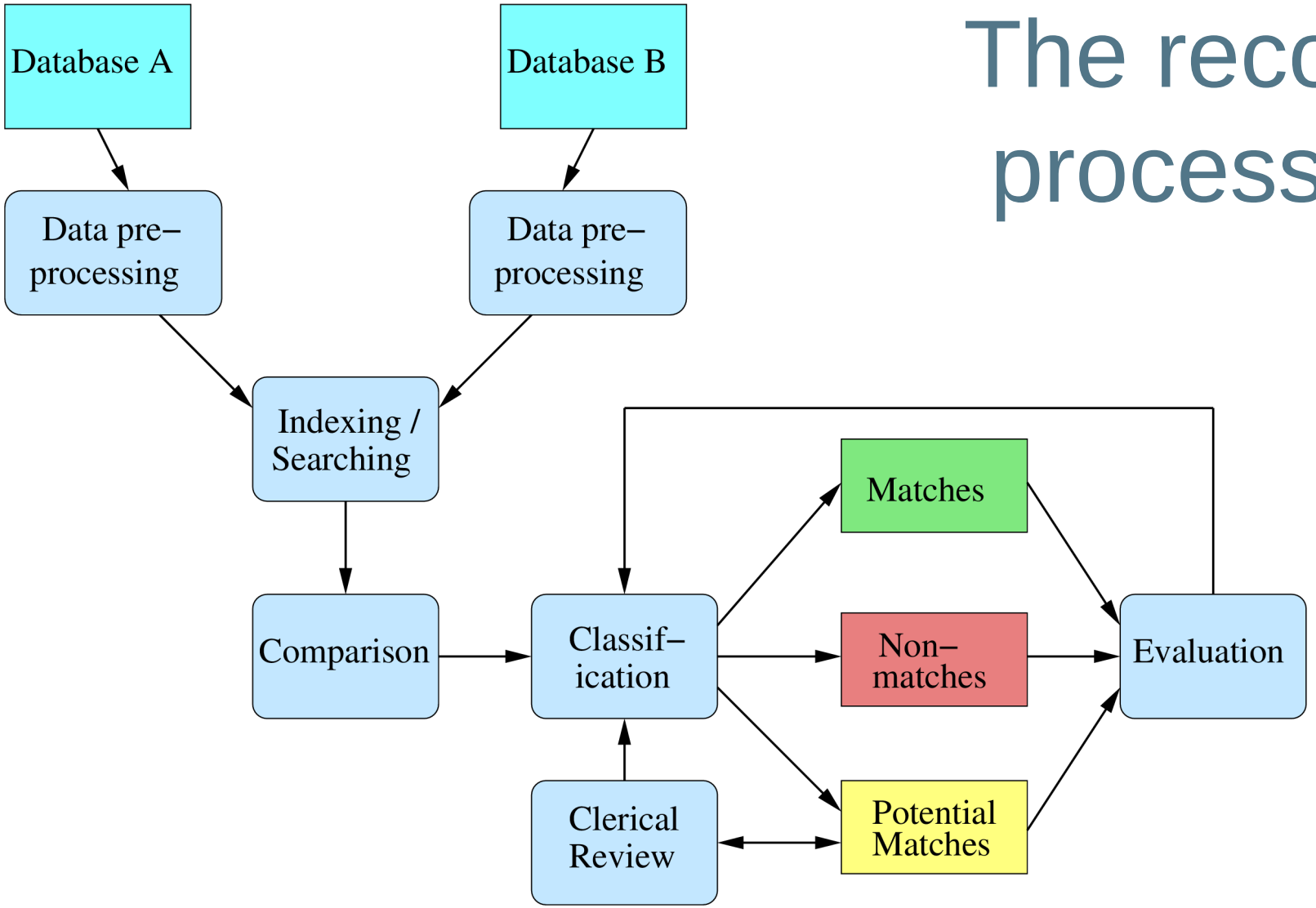
A brief history of record linkage (2)

- 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 **record pair classification** techniques, both supervised and unsupervised, as well as active learning based

The record linkage process



Record linkage techniques

- Deterministic matching
  - Rule-based matching (complex to build and maintain)
- Probabilistic record linkage (Fellegi and Sunter, 1969)
  - Use available attributes for linking (often personal information, like names, addresses, dates of birth, etc.)
  - Calculate match weights for attributes
- “Computer science” approaches
  - Based on machine learning, data mining, database, or information retrieval techniques
  - Supervised classification: Requires training data (true matches)
  - Unsupervised: Clustering, collective, and graph based

Record linkage challenges

- No unique entity identifiers available
- Real world data are dirty (typographical errors and variations, missing and out-of-date values, different coding schemes, etc.)
- Scalability
  - Naïve comparison of all record pairs has a quadratic complexity
  - 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)