Data Science
COMP5122M

Data Linkage

Roy Ruddle

(with thanks to Anna Palczewska)

Private study

- See Minerva Announcements for up-to-date info
- Private study for this lecture
  - Data linkage technical report https://lida.leeds.ac.uk/wp-content/uploads/2021/01/QuantiCode_technical_report.pdf
  - Watch video about privacy preserving data linkage from Scottish Informatics Programme https://www.youtube.com/watch?v=smnnD9ZXwP0

What will you learn?

- The basic process of data linkage
- How to improve quality & efficiency
- Approximate data linkage methods
- How to preserve privacy
- How to link spatial data

Data linkage

Data linkage is when information from two or more records of independent sources are brought together, when they are perceived to belong to the same individual, family, event or place.

Other names for data linkage:

- **data matching, record linkage**
- **object identification, identifying uncertainty**
- **merge-purge process, entity resolution**

Why data linkage?

- Data source cleaning (removing duplicates) – de-duplication, internal data linkage
- Merge records into larger datasets
- Clean and enrich data for mining and analysis
- Create person-oriented statistics (longitudinal study)
- Geocode matching for spatial analysis of health and geographical information.

Benefits of linking data

- Improved data quality and integrity
- Making better use of available data
- Privacy and consent
- Communication benefits
- Research benefits

History of data linkage

- **1946 Halbert L. Dunn – "Record Linkage" in *American Journal of Public Health***
- **1959 Howard Borden Newcombe, Automatic Linkage of Vital Records, *Science***
- **1969 Ivan Fellegi and Alan Sunter, The Theory of Record Linkage, *Journal of the American Statistical Association***

Linking variables

- **Unique identifiers** → **Problems: May not be consistent across datasets. May sometimes be missing.**

- **Names**
- **Addresses**
- **DoB**
- **Gender**
- **Ethnicity**
- **Time**
- **Geographical location**
- **Picture**
- **Description**
- **…**

→ **a combination of variables**

Data linkage example (two datasets)

ID	NN	Name	DoB	Address	PostCode	GP Practice
23	222-2	David Smith	12/08/1976	10 Lake Road	LS1 1OP	E12345

ID	NN	Name	DoB	Address	PostCode	A&E
01	222-2	David Smith	12 Aug 1076	Flat 10 Lake Road	LS1 1OP	LS123
01		Dave Smith	12/08/1976		LS1 1OP	LS11

Evaluation

- true matches
  - pairs of records correctly classified
- false matches
  - a wrong match (false positive)
- missed matches
  - a missed pair (false negative)

Information retrieval metrics

Precision = true matches / (true matches + false matches)

Recall = true matches / (true matches + missed matches)

Improving quality

Data cleaning and standardisation

ID	NN	Name	DoB	Address	PostCode	GP Practice
23	222-2	David Smith	12/08/1976	10 Lake Road	LS1 1OP	E12345

ID	NN	Name	DoB	Address	PostCode	A&E
01	222-2	David Smith	12 Aug 1076	Flat 10 Lake Road	LS1 1OP	LS123
01		Dave Smith	12/08/1976		LS1 1OP	LS11

Data cleaning and standardisation

- typographical errors (spelling errors, variation of names)
- deferent coding schemes (male/female, M/F)
- missing data
- changing data over time

ID	NN	Name	DoB	Address	PostCode	GP Practice
01	222-2	David Smith	12 Aug 1976	Flat 10 Lake Road Leeds	LS1 1OP	LS123
01	222-2	Dave Smith	12/08/1976	11a Street Lane	LS11 9OL	LS1123

Data cleaning and standardisation

- reformatting values to the common format
- removing punctuation
- phonetic encoding (soundex, methaphone, NYSIIS software)

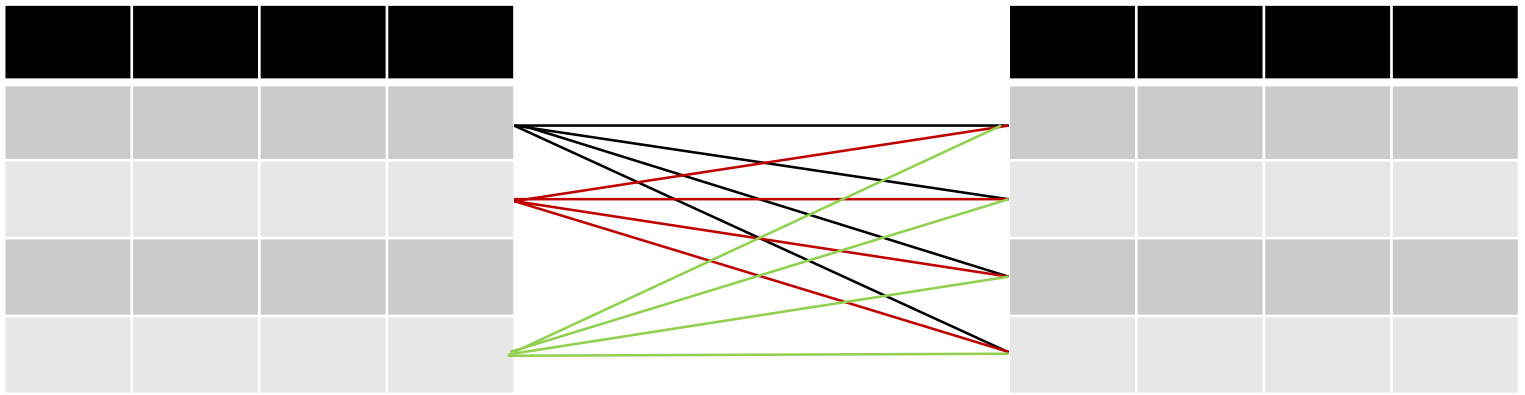
  e.g. Peter, Pete -> p233 Anna, Ana->a566
- name and address standardisation

First Name	Last Name	Number	Street	County	City	Postcode
David	Smith	10	Lake Road	WY	Leeds	LS1 1OP

- nick name and abbreviation lookups

Improving efficiency

Blocking method



1 million records

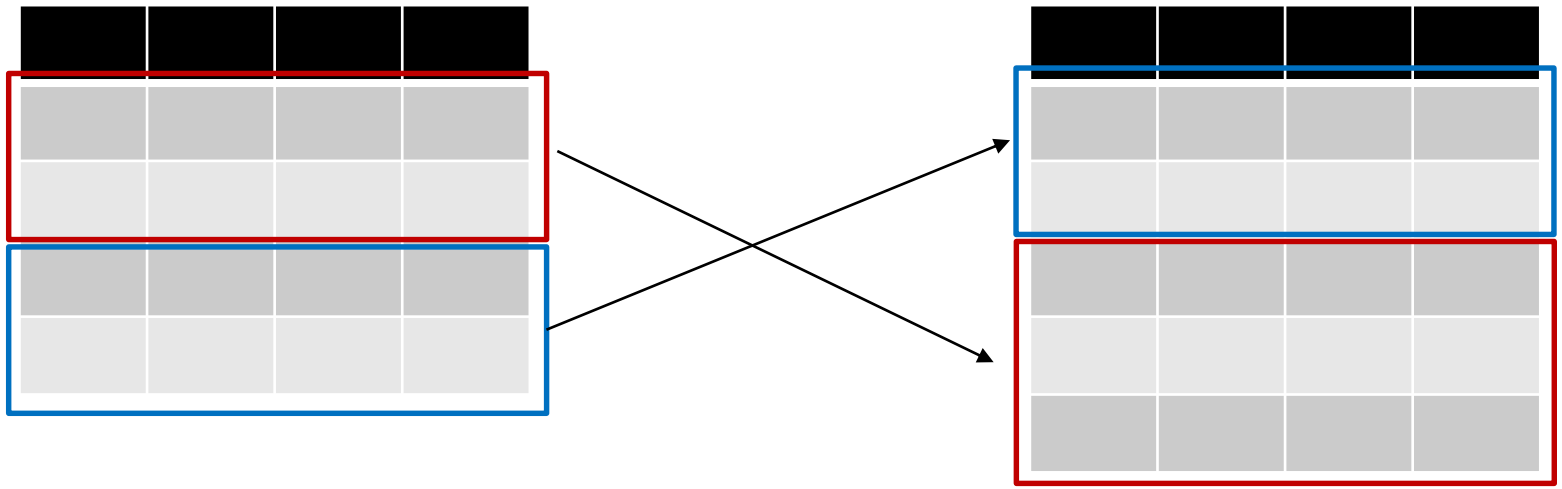
5 million records

5×10^{12} (5 trillion record pairs)

Assume: 1 comparison takes 1ms

- **57870.3 days**
- **1902.5 months**
- **158.54 years**

Blocking methods



- reduce the large amount of comparison
- remove the candidate record pairs which are not matches
- compare record pairs that have the same value (blocking key) for blocking variable

Blocking methods

- **Traditional blocking**

Blocking variable

ID	NN	Name	DoB	Address	PostCode	GP Practice
01	222-2	David Smith	12 Aug 1976	Flat 10 Lake Road	LS1 1OP	LS123
01	222-2	Dave Smith	12/08/1976	11a Street Lane	LS1 1OP	LS1123

Blocking key

- **Sorted neighbourhood approach**
- **Q-gram blocking**
- **Canopy clusters**

Evaluation

- all record pairs
- candidate record pairs (generated by blocking)

Reduction ratio = 1 – (candidate record pairs / all record pairs)

Approximate data linkage

See also similarity measures (Exploratory analysis lectures)

Methods

- **Deterministic linkage**
  - Exactly match on specified common fields
  - Easiest, quickest linkage strategy
  - Results in errors due to non-matches
- **Probabilistic linkage**
  - Statistically estimate likelihood that two records describe the same individual\entity, even if they disagree on some fields
  - Computationally complicated
  - Fewer non-matches
- **Artificial intelligence approaches**

Deterministic linkage

ID	NN	Name	DoB	Address	PostCode	GP Practice
23	222-2	David Smith	12/08/1976	10 Lake Road	LS1 1OP	E12345

ID	NN	Name	DoB	Address	PostCode	A&E
01	222-2	David Smith	12 Aug 1076	Flat 10 Lake Road	LS1 1OP	LS123
01		Dave Smith	12/08/1976		LS1 1OP	LS11

1. If NN agrees then match

2. If not NN agrees and (any two from {Name, DoB, Address} agrees then match

Deterministic linkage tools

- sort-merge algorithms in Excel, R, Python, and other programming languages
- sql select with joins
  - https://www.youtube.com/watch?v=HyZtBGXLN00

Probabilistic linkage

ID	NN	Name	DoB	Address	PostCode	GP Practice
23	222-2	David Smith	12/08/1976	10 Lake Road	LS1 1OP	E12345

w_1 w_2 w_3 w_4

ID	NN	Name	DoB	Address	PostCode	A&E
01	222-2	David Smith	12 Aug 1076	Flat 10 Lake Road	LS1 1OP	LS123
01		Dave Smith	12/08/1976		LS1 1OP	LS11

$$w_i = \frac{m_i}{u_i}$$

Probability that a common variable agrees on a matched pair.

Probability that a common variable agrees on an unmatched pair.

$$w_t = \sum_{i}^{k} w_i$$

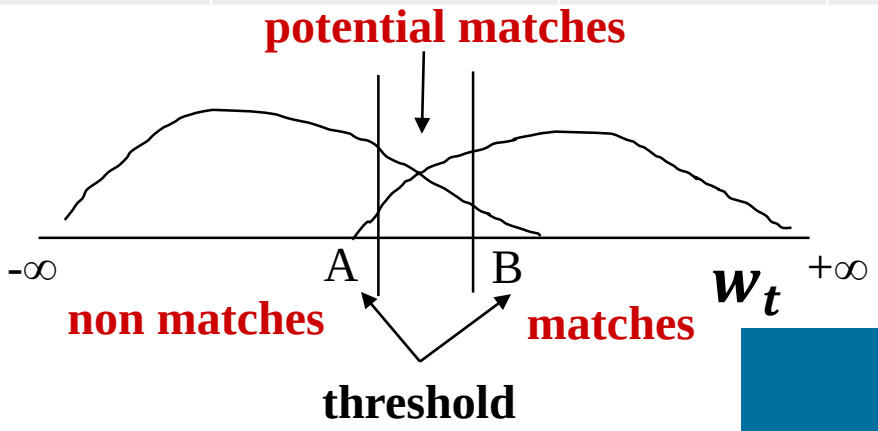
Probabilistic linkage

ID	NN	Name	DoB	Address	PostCode	GP Practice
23	222-2	David Smith	12/08/1976	10 Lake Road	LS1 1OP	E12345

w_1 w_2 w_3 w_4

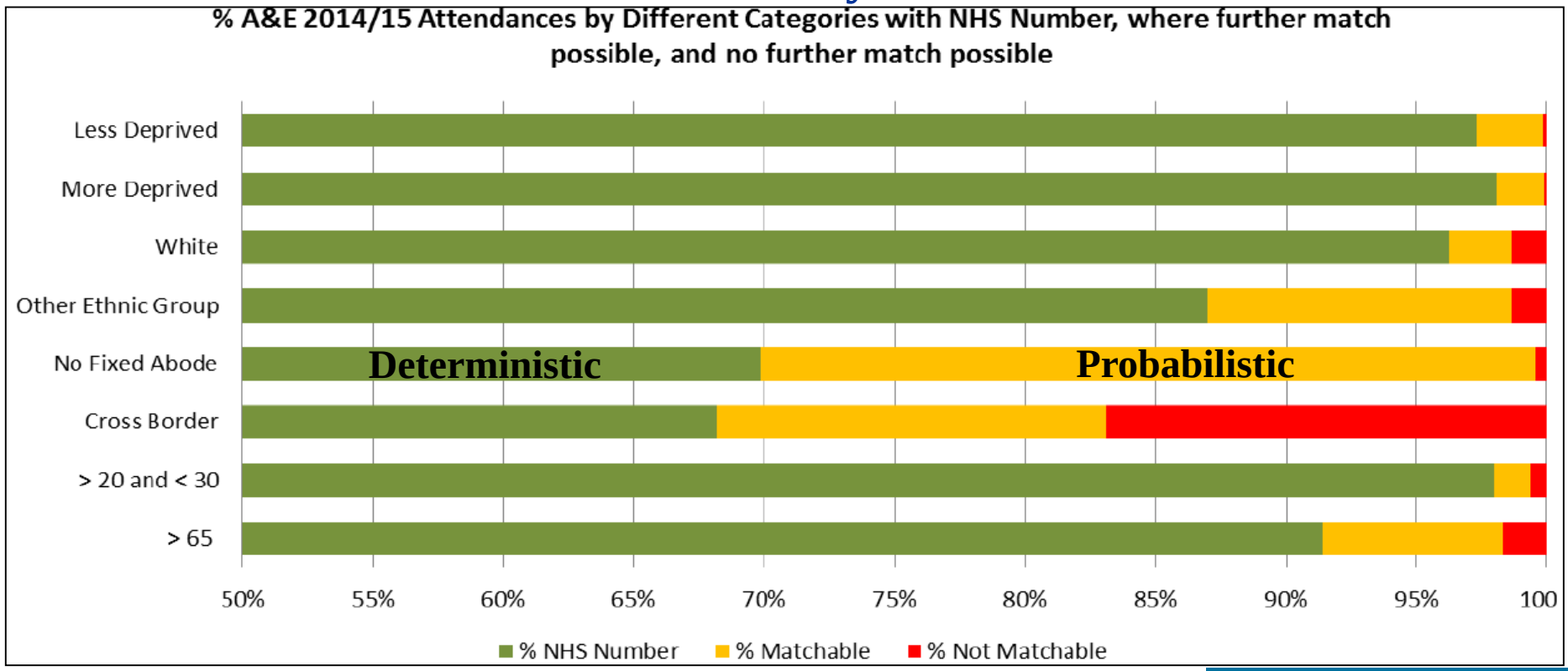
ID	NN	Name	DoB	Address	PostCode	A&E
01	222-2	David Smith	12 Aug 1076	Flat 10 Lake Road	LS1 1OP	LS123
01		Dave Smith	12/08/1976		LS1 1OP	LS11

$$w_t = \sum_{i}^{k} w_i$$



Impact of data quality on linkage

- **False/missed matches often not randomly distributed**
  - **Leads to bias in data analysis**



Source: NHS England (2016).

Preserving privacy

Privacy-preserving record linkage

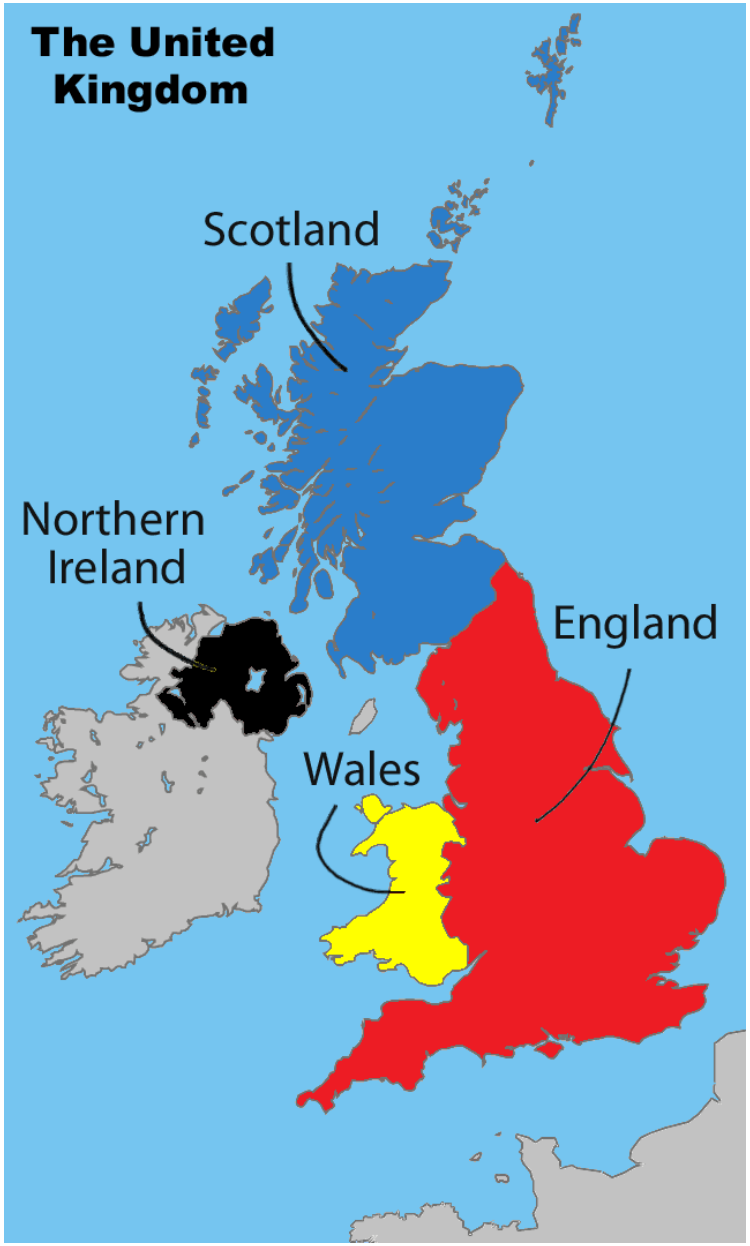
- secure way to link record of data from two or more organizations (e.g. governmental agencies and health institution)
  - E.g., Scottish Informatics Programme https://www.youtube.com/watch?v=smnnD9ZXwP0

Spatial linkage

Geographical location

- **Direct georeference (GPS, surveys)**
  - **Point on a map defined coordinates, line, or polygon (boundaries)**
- **Indirect georeferenced**
  - **Postal addresses, postal codes and place names**
  - **Do not include explicit coordinates**

The UK



https://www.youtube.com/watch?v=rNu8XDBSn10

UK geographies

- **Census geography**
- **Postal geography**
- **Health geography**
- **Electoral geography**
- **Administrative geography**
- **Other**
  - Local Education Authority
  - Build-up areas
  - National Parks
  - Police Force Areas
  - Fire and Rescue Authorities

Census geography

Geography	Population		Household	
	Min	Max	Min	Max
Output Area (OA)	100	625	40	250
Lower SOA	1000	3000	400	1200
Middle SOA	5000	15000	2000	6000

- OAs are the lowest geographical level at which census estimates are provided
- OAs are built from clusters of adjacent unit postcodes
- OAs are subject to change due to the changes in the population, postcode and local authorities areas

Postal geography

- Geographic data (e.g., post codes; LS2 9JT)
  - LS (the area)
  - 2 (the district)
  - 9 (the sector)
  - JT (the unit; ≈ addresses)
- E.g.
  - https://en.wikipedia.org/wiki/LS_postcode_area

Lookup tables

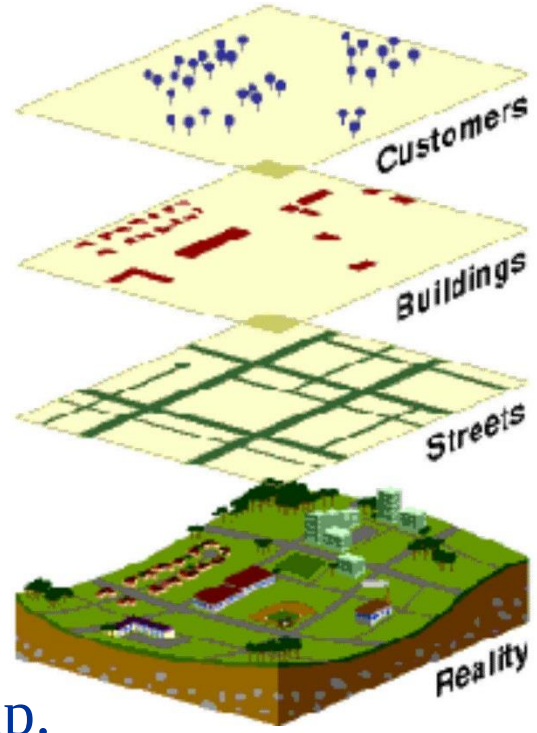
- ONS Code History Database (CHD)
- Postcode lookup files: ONS Postcode Directory, NHS Postcode Directory, and https://data.gov.uk/
- Lookup tables between geographies

Methods:

1. Exact-fit – when one geography falls within boundary of other geography
2. Best-fit – when one geography boundaries straddles the boundary of other geography (based on population weighted centroid or mean grid reference of all the addresses)

Geographical Information Systems GIS

- are designed to capture, store, manipulate, analyse, manage, and present all types of spatial or geographical data
- enables people to more easily see, analyse, and understand patterns and relationships
- maps create overlays from which we can extract the features of one data set that fall within the spatial extent of another dataset
- are used for geocoding (e.g. linking an address to a physical location on the earth) GIS calculates geographic coordinates before an address can be displayed on a map.



Mapping tools

- **GeoConvert, MapInfo, QGIS, ArcGIS**