Data Science
COMP5122M

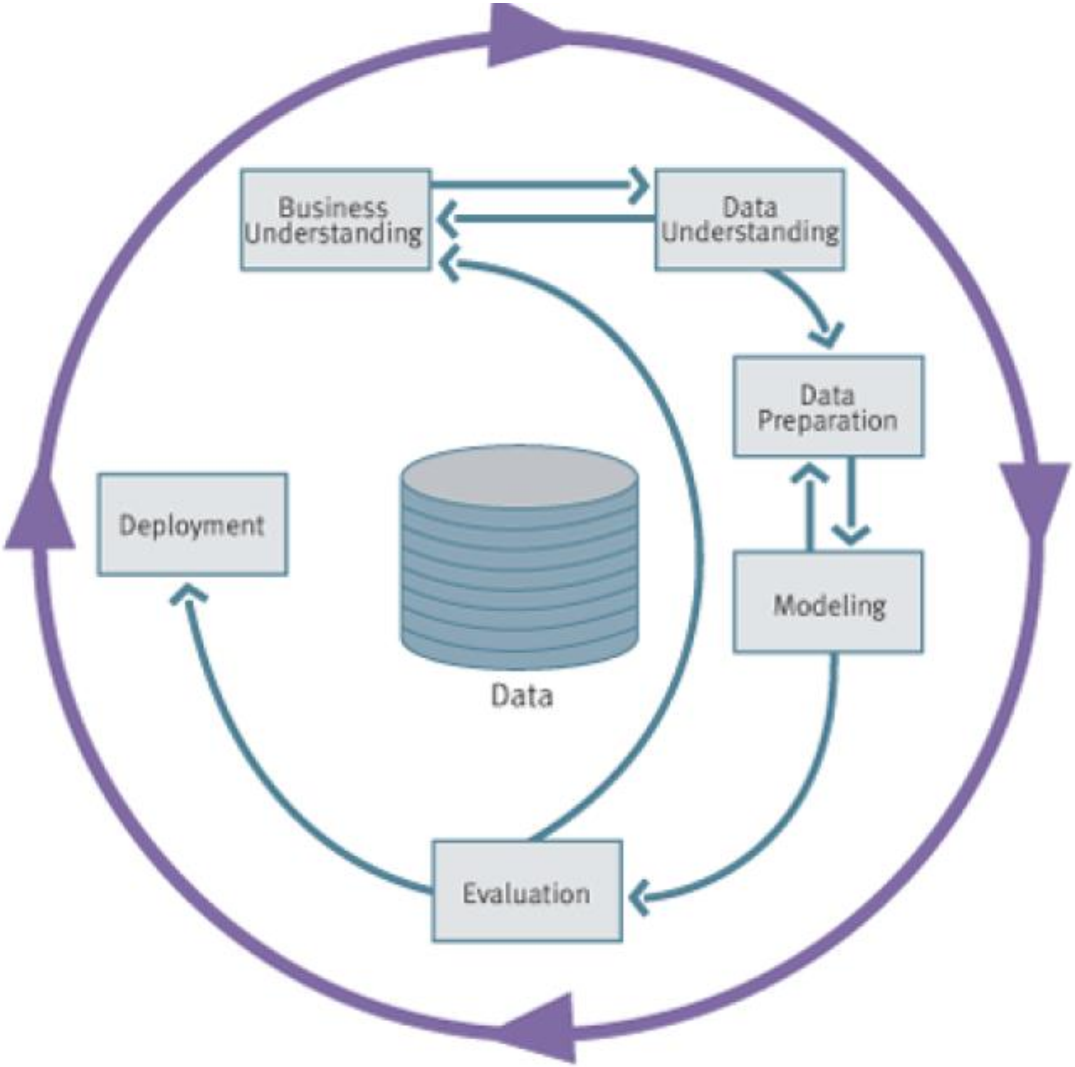
Data understanding

Roy Ruddle

Private study

- See Minerva Announcements for up-to-date info
- None for this lecture

CRISP data mining process



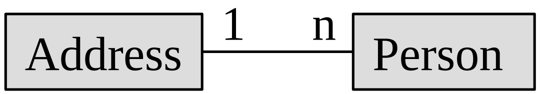
Cross-Industry Standard Process for Data Mining (Shearer, 2000).

What will you learn?

- The (confusing) terminology that people use to describe data
- Three basic data types
  - The importance of considering semantics
  - And to appreciate well-documented data
- Differences between open vs. shared vs. closed data
- Some methods to preserve privacy

Terminology: Data sets

Data set			
Database	**Data table**	**Variable**	**Record**
	Flat file .csv	Column Field Attribute	Row Instance



Person ID	Age	Gender
10000001	27	Female
10000002	24	Male
10000003	3	Female

Terminology: Data types (1)

- **Which terms are**
  - **Synonymous?**
  - **Similar?**
  - **Opposites?**

Alphabetical order			
Boolean	Dimension	Measure	Quantitative
Categorical	Discrete	Nominal	Ratio
Characteristic	Geographic	Numerical	Referential
Connection	Hierarchical	Observation	Spatial
Continuous	Interval	Ordinal	Text
Date	Level of detail	Population	Time
Date/time	Link	Qualitative	

Terminology: Data types (2)

- Referential
  - Context in which measurements were made
  - Three kinds of ‘backdrop’
    - Time
    - Space
    - Population
- Characteristic
  - The measurements

Andrienko & Andrienko. (2006). Exploratory analysis of spatial and temporal data.

Terminology: Data types (3)

- Continuous
  - In principle the variable contains an infinite number of values
    - Some numerical variables (typically decimals, dates & times)
- Discrete
  - Finite number of values
    - Some numerical variables (typically integers)
    - Ordinal variables
    - Categorical variables

Terminology: Data types (4)

- Dimensions and measures
  - Dimension
    - Same as “variable”?
    - Integrally related variables?
      - E.g., X & Y coordinates, but not height & weight
  - Measure
    - Synonymous with observation
    - And characteristic?
  - And Tableau confuse things even more!
    - Dimensions (mostly discrete)
    - Measures (mostly continuous)

Terminology: Basic data types

- **Nominal**
  - **Are only = or ≠ to other values**
- **Ordinal**
  - **Sequence matters; obeys a < relation**
- **Numerical**
  - **Can do arithmetic on them**

Data types		
Categorical	**Ordinal**	**Numerical**
Nominal		Quantitative
Qualitative		Date/time
		Spatial

IEEE Vis 2017 "Test of Time" Award

Card, & Mackinlay. (1997). The structure of the information visualization design space. Proc. IEEE Symposium on Information Visualization.

What data type?

- **Is each variable categorical, ordinal or numerical?**
- **Defined by the values vs. semantics?**

ID	Gender	Date of birth	Height	Ethnicity	Visit number	Day of last visit	Month of last visit	Year of last visit
10035691	1	01/01/2006	140	A	1	3	June	2018
19465810	2	31/01/2005	170	B	3	7	July	2017
23006780	1	01/02/2004	187	B	2	23	May	2016

RTFM

- **Read the *** manual!**
  - **Metadata & documentation**

E.g., https://datamillnorth.org/dataset/off-street-parking-fines

- **Overall description of the dataset**

Every fine issued for vehicles not having a valid parking ticket whilst in car parks.

- **Structure of the data files**

From Quarter 3 2014/15, data on off street fines (car park fines) has been divided in to a fines issued dataset and fines paid dataset.

- **Explanation of each variable**
  - **Name & explanation**
  - **If you’re lucky!**

Issued - Date PCN was issued

- **Google to fill in the gaps**

What is a ‘PCN’?

Data sources

Data sources

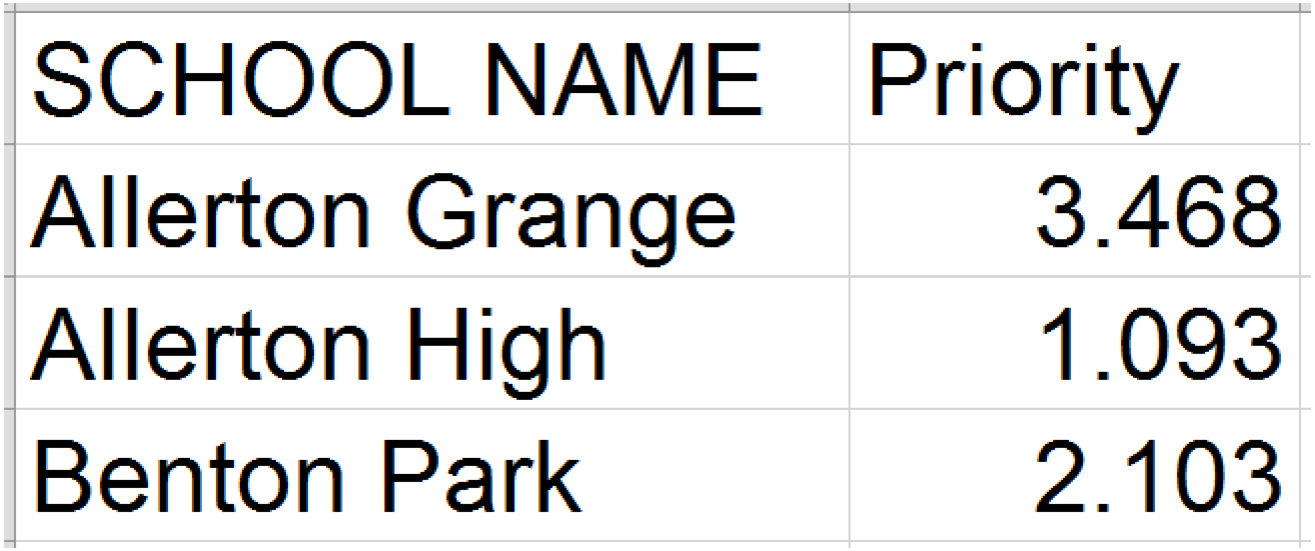
- Publicly available data
  - Open data
  - Other
- Shared data
- Closed data

Open data

- Usually aggregated to preserve privacy
- UK Open Government Licence (OGL v3)
  - Free to copy, publish, adapt, exploit, etc.
  - You must acknowledge the source, etc.
  - http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/

Aggregated open data example

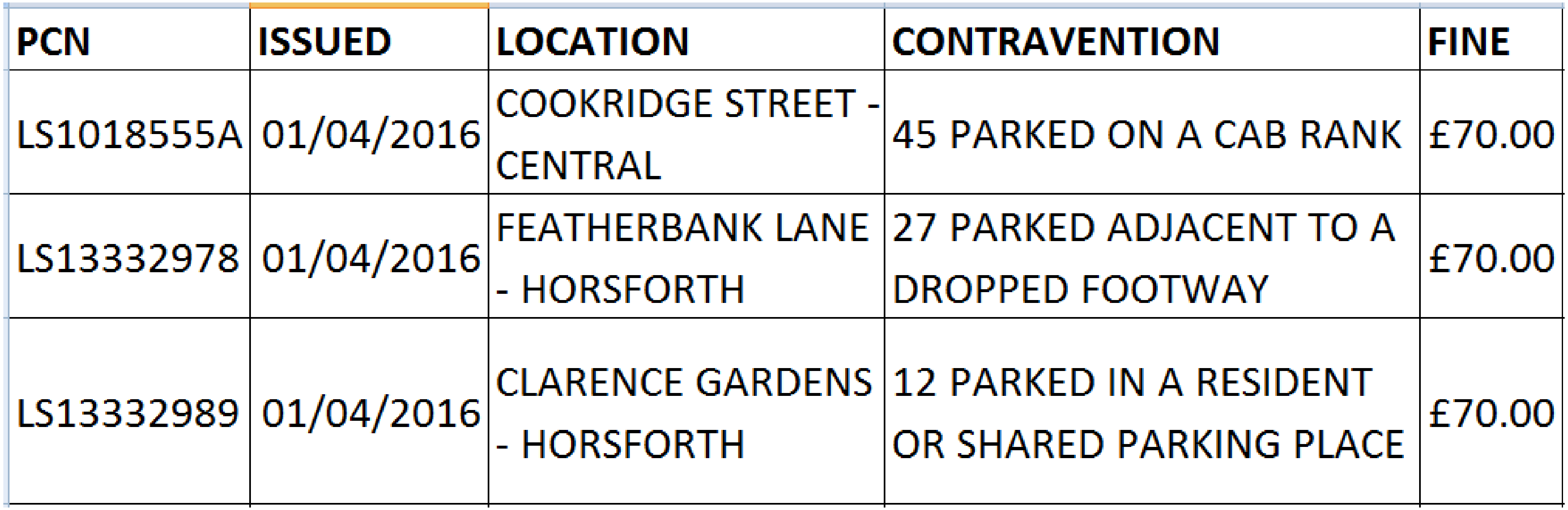
- **Priority cut off distances for secondary schools (https://data.gov.uk/)**



SCHOOL NAME	Priority
Allerton Grange	3.468
Allerton High	1.093
Benton Park	2.103

Non-aggregated open data example

- **Leeds parking fines (http://datamillnorth.org/)**



PCN	ISSUED	LOCATION	CONTRAVENTION	FINE
LS1018555A	01/04/2016	COOKRIDGE STREET - CENTRAL	45 PARKED ON A CAB RANK	£70.00
LS13332978	01/04/2016	FEATHERBANK LANE - HORSFORTH	27 PARKED ADJACENT TO A DROPPED FOOTWAY	£70.00
LS13332989	01/04/2016	CLARENCE GARDENS - HORSFORTH	12 PARKED IN A RESIDENT OR SHARED PARKING PLACE	£70.00

Other publicly available data

- E.g. from
  - Office for National Statistics
    - https://www.ons.gov.uk/
  - NHS Digital
    - http://digital.nhs.uk/searchcatalogue
- Usage terms sometimes unclear
  - It's your responsibility to check

Shared data

- Data that is shared only with named people or organisations, or groups who meet certain criteria, e.g.,
  - Environmental datasets provided under licence for a specific purpose (e.g. teaching)
  - Hospital episode statistics (NHS Digital)
- Promotes re-use
- Sometimes record-level but anonymised

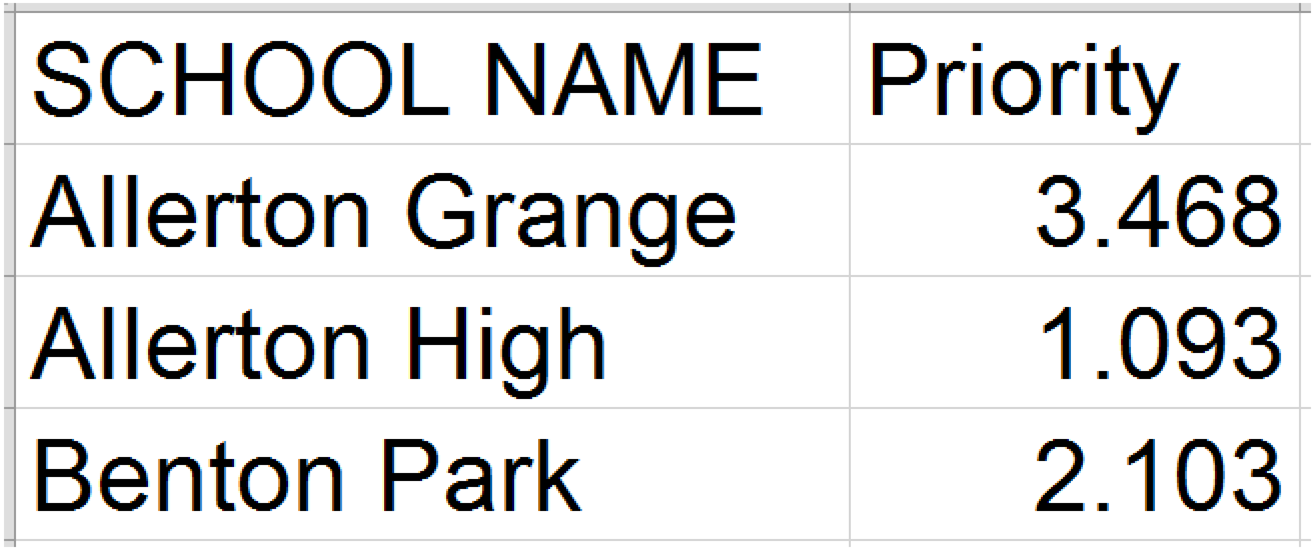
Closed data

- Data that can only be accessed by its subject, owner or organisations that have been granted permission (see ethics lectures)
- E.g.,
  - Your electronic health record
    - Record-level and identified
  - Product sales
    - Commercially confidential

Preserving privacy

Aggregation

- Create a new record from a set of others
  - Use home postcode to calculate each pupil’s distance from school
  - Publish the maximum
    - Preserves pupils’ privacy



SCHOOL NAME	Priority
Allerton Grange	3.468
Allerton High	1.093
Benton Park	2.103

A variable’s level of detail (LOD)

- Reduce the precision (e.g., age)
- Coarser level of detail (e.g., spatial region)
- Group categories, e.g., “other”

LOD: Reduce precision

- **E.g., Numbers of Patients Registered at a GP Practice**
  - **Publicly available data from NHS Digital**
  - **Why grouped for 95+?**

PRACTICE_CODE	MALE_0_1	MALE_1_2	**Etc.**	**MALE_95+**
A81001	14	21		0
A81002	105	95		9
A81003	14	15		0

LOD: Spatial region

- Geographic data (e.g., post codes; LS2 9JT)
  - Outcode
    - LS (the area)
    - 2 (the district)
  - Incode
    - 9 (the sector)
    - JT (the unit; ≈ addresses)
- See
  - Tableau practicals
    - GP surgeries in Yorkshire.xlsx
  - Data Linkage lecture
  - https://en.wikipedia.org/wiki/LS_postcode_area

Final things to consider

- Data provider
  - Reputation
  - Data quality (even from reputable provider)
- Lawful usage
  - Licensing terms for each dataset
  - Ethics & information governance (see later lectures)