

SCC.460 Data Science Fundamentals

Data Gathering, Collection and Cleaning



It all starts with the aim...

- Aim: the general statement of what we want to achieve in the research
 - E.g. to understand what factors are driving obesity in the UK
- Objectives: concrete statements that guide us towards our aim and measure
 - E.g. gather information about obese people's daily habits
- Questions: follow directly from the objectives, the thing we wish to explore
 - Different types of questions for different objectives
 - **Exploratory, relationship, etc.**

From Questions to Gathering Data

- We have a range of data generation methods:
 - Interviews: controlled interaction
 - Questionnaires: pre-defined list of questions
 - Observations: observing what is taking place
 - Documents/reports: artefacts from which we can draw observations
- Two main categories of data:
 - Observational
 - Found

Found vs Observational

	Found Data	Observational Data
<i>Provenance</i>	Unknown / Uncertain	Generation process and conditions are known
<i>Availability</i>	Readily available, easy to acquire	Expensive to produce (i.e. resources)
<i>Familiarity with subject(s)</i>	Little knowledge (demographics, if any)	Awareness of participants
<i>Participant recruitment</i>	Process unknown	Control over recruitment strategy
<i>Pre-processing</i>	Takes time to accumulate and sanitise	Relatively quick
<i>Inference & Revisiting</i>	Limited inference capability Hard to follow-up with	Enables re-correspondence with participants

So why bother with found data?

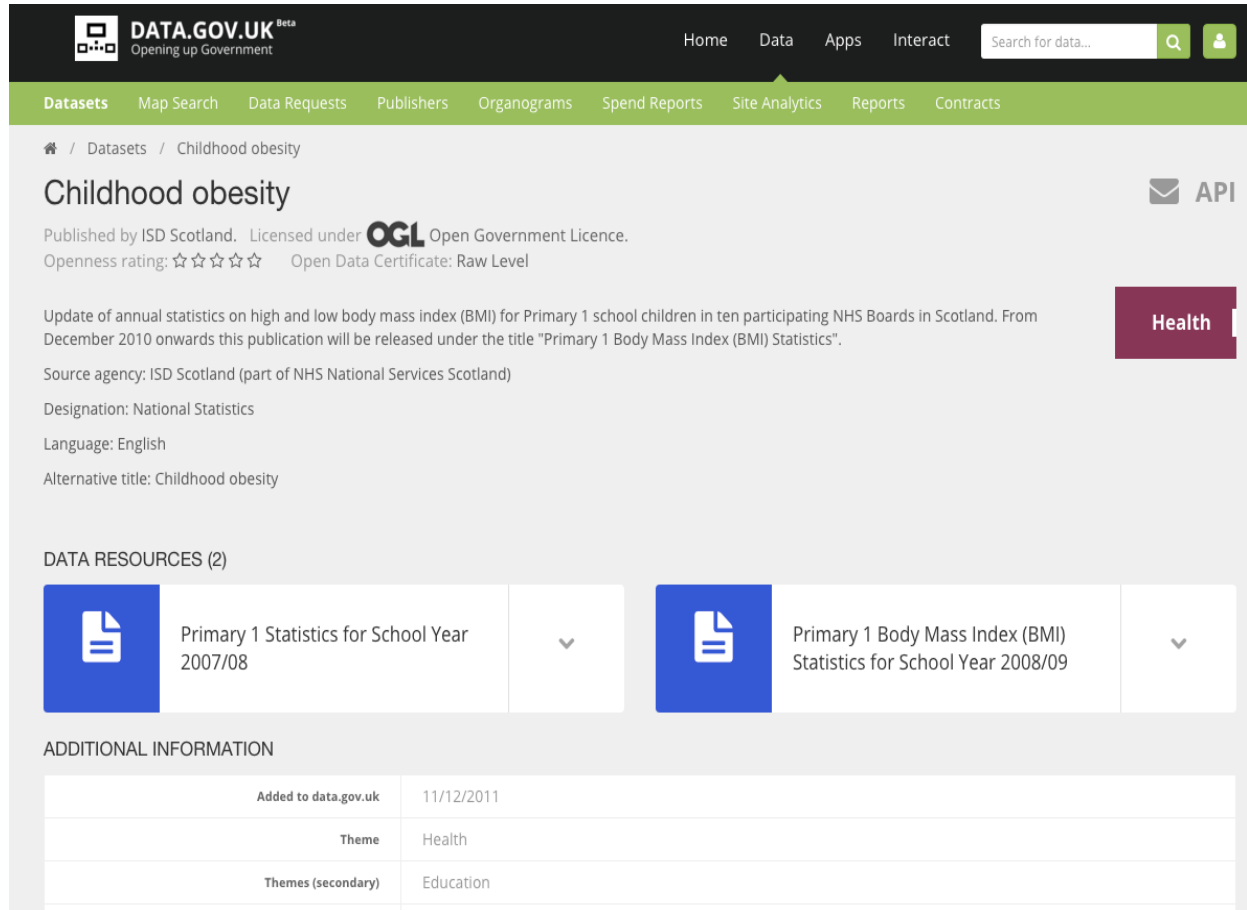
- It is much easier to acquire.
- The overhead of generating observational data can be expensive/prohibitive:
 - Careful design of collection process
 - Designing incentives to encourage participation
 - Recruitment is difficult and often limited
- The Web has provided an ideal resource from which to collect found data.

The Web as a Data Source

- Large Dataset Dumps: DBPedia, data.gov.uk
- Application Programming Interfaces (APIs): Facebook, Twitter, YouTube
- Spidering and Scraping:
 - Spidering: identifying URLs in a web page, accessing them to find more URLs
 - Scraping: extracting content from a web page's HTML
 - Various tools: import.io, scrapy (Python), jaunt (Java)
 - Ethics of scraping:
 - Always observe the robots.txt file on the site you are going to scrape
 - Scraping is ethically dubious, sites prefer you to access their content via APIs in order to be able to enforce authentication and rate limiting

Found Data: data.gov.uk

- UK government produces open data describing a range of topics:
 - Transport
 - Education
 - Economics
 - Ecology
- Often data is uploaded but with limited provenance
 - No information about collection process
 - No information about parties who compiled the data



The screenshot shows the data.gov.uk website interface. At the top is a dark navigation bar with the 'DATA.GOV.UK Beta' logo and links for Home, Data, Apps, and Interact. Below this is a green secondary navigation bar with links for Datasets, Map Search, Data Requests, Publishers, Organograms, Spend Reports, Site Analytics, Reports, and Contracts. The main content area is for the 'Childhood obesity' dataset, published by ISD Scotland under the OGL (Open Government Licence). It includes an Openness rating of four stars and an Open Data Certificate of Raw Level. The description states it's an update of annual statistics on BMI for Primary 1 school children in Scotland. Source agency is ISD Scotland (part of NHS National Services Scotland). Designation is National Statistics, Language is English, and Alternative title is Childhood obesity. There is an API icon and a 'Health' category tag. Under 'DATA RESOURCES (2)', two datasets are listed: 'Primary 1 Statistics for School Year 2007/08' and 'Primary 1 Body Mass Index (BMI) Statistics for School Year 2008/09'. At the bottom, an 'ADDITIONAL INFORMATION' table provides metadata.

ADDITIONAL INFORMATION	
Added to data.gov.uk	11/12/2011
Theme	Health
Themes (secondary)	Education

Found Data: Other sources

- <http://aws.amazon.com/datasets>
- <http://books.google.com/ngrams>
- <http://archive.ics.uci.edu/ml> **
- <http://crawdad.org> **
- <http://www.re3data.org>
- <http://reddit.com/r/datasets>
- <http://factual.com>
- ...and many more

Cleaning and Pre-processing



Cleaning and Pre-processing

- Data is very rarely perfect and ready to use.
- Often it contains errors (caused by humans or machines):
 - Missing data
 - Outliers
 - Duplicates
 - Wrongly labeled
- In natural language: Misspellings, Grammatical errors, Odd abbreviations, etc.
- Handling this manually is expensive and potentially infeasible.
- Coding up different data 'preparation' stages (Java, Python, R, etc.).

Cleaning: Missing Values

- In the age of Big Data?!
- What can you do about it:
 - Ignore it
 - Many dependent processes / software will break
 - Go back and collect it
 - Approximate it: Use statistical methods with domain knowledge

Cleaning: Missing Values

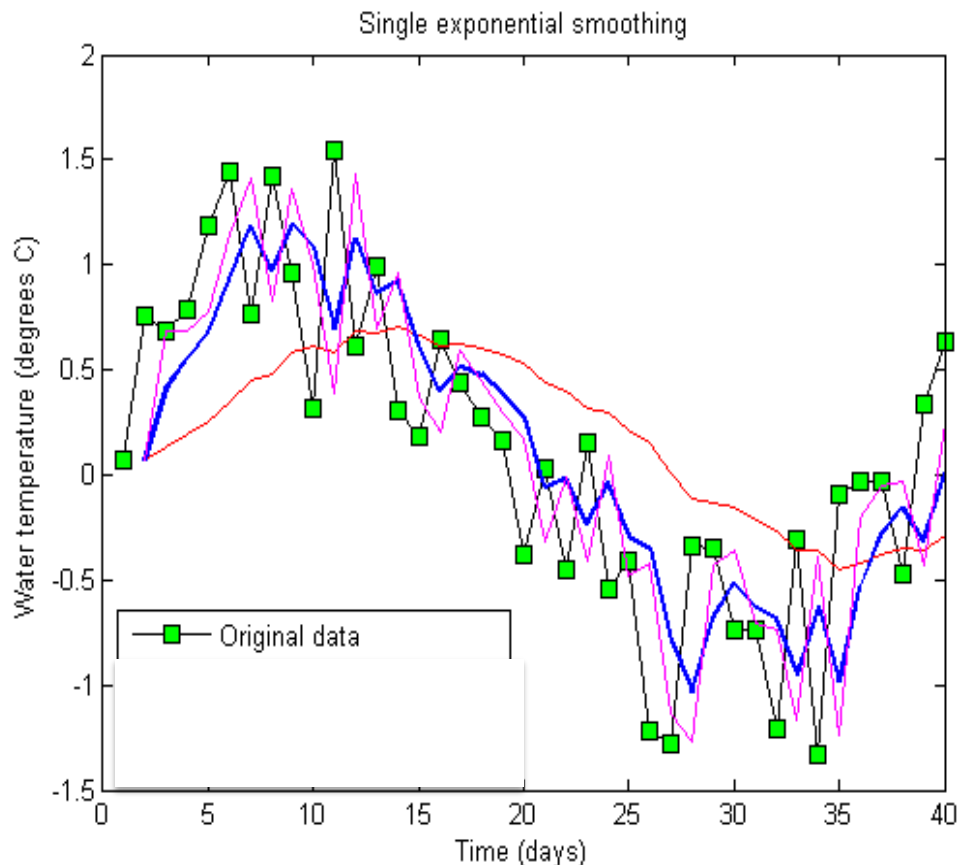
- Statistical methods:
 - Replace with mean
 - Linear interpolation, quadratic regression
- Do you know which data are “missing”?
- Missing data could tell you much more than its unknown value:
 - Experiment design / assumption is faulty
 - Instruments used are not suitable
 - Is the missing data the golden nugget?

Cleaning: Smoothing

- Data may contain variation (i.e. in a given feature's values) over time
 - E.g. time series of GDP for the past 250 years
- Such variation can be 'jumpy' and prone to spurious fluctuations
- Smoothing techniques 'level-out' such jumpiness:
 - Moving average with window size n
 - Exponential Smoothing: controlled by parameter α
 - High $\alpha \Rightarrow$ more prioritisation to recent feature values
 - Low $\alpha \Rightarrow$ less preference for recent feature values

Cleaning: Smoothing

- Choose the setting of alpha that minimises the MSE.
- R and Java both provide libraries to smooth time-series data using simple moving averages and exponential smoothing.



Cleaning: Outliers

- “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.” Douglas Hawkins, 1980
- Detection (and removal) is important as outliers can affect analyses drawn from the data; e.g. calculation of mean, standard deviation, etc.
- What do outliers signify?
 - Data capturing error (revisit missing values causes and possible indications)
 - Deliberate imprecision from the provider of the data (cf SA in GPS before 2000)
 - Anomaly in the real observed data
 - Rare (once in a blue moon) abnormality
 - Tip of the iceberg of an unanticipated trend

Cleaning: Outliers

- Proximity-based methods: The locality of an outlier deviates significantly (sparsely populated) from that of most of the others in the data set.
- Three types:
 - Distance-based approaches
 - Neighborhood does not have enough other points (using distance metrics)
 - Density-based approaches
 - Point density is dissimilar to that of its neighbours
 - Cluster-based approaches
 - Membership of and distance to clusters indicate potential outliers

Cleaning: Outliers

- Visual methods: Identifying potential outliers through visual representation.
- Easy to do with 1 dimensional data (=> best as exploratory graphical device).
- Gets increasingly trickier with each additional dimension.

• Scatter plots

• Boxplots

