FIT5145 Introduction to Data Science

Module 4

Data Resources, Processes, Standards and Tools

2019 Lecture 7

Monash University

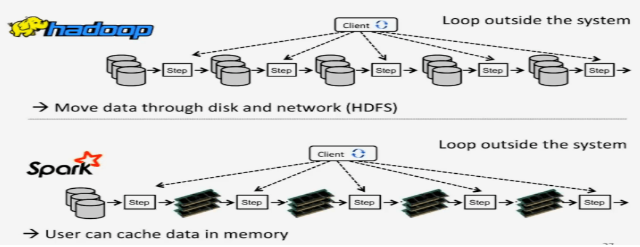
Discussion: Hadoop and Spark

Hadoop

- an inexpensive and open source platform for parallelising processing
- not suited to streaming (suitable for offline processing)

Spark

- include Map-Reduce capabilities
- provides real-time, in-memory processing
- much faster than Hadoop



→ Move data through disk and network (HDFS)

→ User can cache data in memory

FLUX Question

What is MapReduce?

A. A way to make maps smaller.

B. Owned by Apache.

C. A multi stage process to break up then analyse data.

D. No longer used, Google has found an alternative.

Unit Schedule: Modules

Module	Week	Content
1.	1	overview and look at projects
	2	(job) roles, and the impact
2.	3	data business models
	4	application areas and case studies
3.	5	characterising data and "big" data
	6	data sources and case studies
4.	**7**	**resources and standards**
	8	**resources case studies**
5.	9	data analysis theory
	10	data analysis process
6.	11	issues in data management
	12	GUEST SPEAKER & EXAM INFO

Learning Outcomes (Week 7)

By the end of this week you should be able to:

- Locate the new data sources
- Identify the clever and creative use of existing multiple data sources
- Identify issues and complexities in data sources
- Explain what data wrangling is
- Identify appropriate set of wrangling tasks in a give dataset
- Apply Python for data wrangling

Introduction to Resources (ePub section 4.1)

introduction to issues

- using data
  - want new data sources or clever and creative use of existing multiple data sources
- open data
  - organisations provide machine readable to support data science
- wrangling
  - manipulating data to make it directly usable for analysis

Introduction to Resources
Using data

access to new data sources or clever and creative use of existing multiple data sources are hallmarks of innovative data science

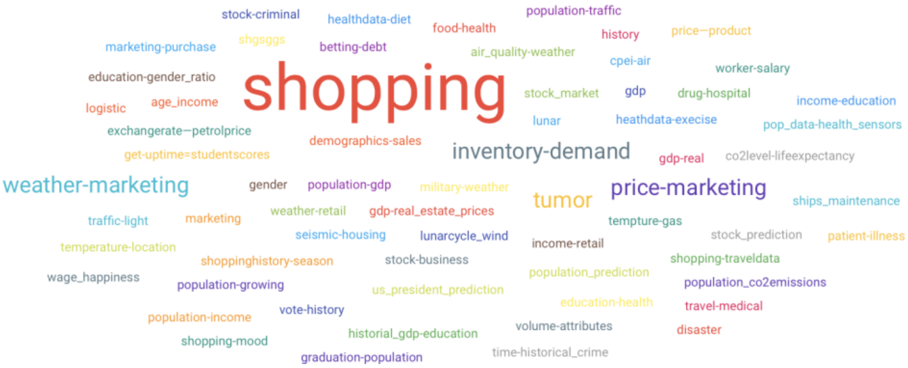
Where to find and how to use data sources

Task: forecasting traffic: blockages, clearing, surprising situations, alternate routes

- Critical data:
  - GPS data on traffic flow
  - Maps
  - incidents and events
  - weather
- Challenge:
  - collect different sources of data

FLUX Question

Give an example where two very different data sets needed to be combined in order to make a data science project work.



Three Examples of Using Data

We'll now look at three examples of public data and using data?

1. NYC data
2. traffic prediction
3. predictive analytics for banks

New York City Data

Under Mayor Bloomberg, NYC embarked on a program to make the city's data accessible:

- *"How data and open government are transforming NYC"* in *Radar.O'Reilly*:
  - "In God We Trust," tweeted New York City Mayor Mike Bloomberg this month. "Everyone else, bring data."
  - applications of the data provided:
    - "real-time updates on your phone based on where the buses are located using very low-cost technologies"
    - applying predictive analytics to building code violations and housing data to try to understand where potential fire risks might exist
- *Bloomberg signs NYC 'Open Data Policy'* into law, plans web portal for 2018," in *Engadget*
- *NYC Open Data portal*
- Melbourne has a similar portal: *City of Melbourne's open data platform*

NYC Data, cont.

“How we found the worst place to park in New York City” is examples, and a discussion of the complexities of getting data out of NYC:

Map of road speed by day+time: GPS data for NYC cabs gives; data obtained via FOIL request, then made public by recipient

Danger spots for cycles: *NYPD crash data* obtained by daily download of PDF files followed by (non-trivial) extraction
NB. they now have Excel data to ease the work!

Dirty waterways: *fecal coliform measurements on waterways* from Department of Environmental Protection’s website; extracted from Excel sheets per site; each in a different format

Faulty road markings: parking tickets for fire-hydrants by location from *NYC Open Data portal* need to normalize the addresses supplied

Traffic Prediction

Back in 2008, *Microsoft Introduced a Tool for Avoiding Traffic Jams*

The system was called Clearflow:

- it aimed to forecast traffic: blockages, clearing, surprising situations, etc.
- and to suggest alternate routes
- critical data use to build the application included:
  - GPS data on traffic flow
  - maps
  - incidents and events
  - weather

Predictive Analytics for Banks

See *this video* of a seminar on
"Predictive Analytics with Fine-grained Behavior Data"

- by Foster Provost (Professor at NYU and author of *this book*) presented at Stata+Hadoop in 2013
- describes customer prediction problem for banking products

He discusses about whether bigger data is "always" better. So is big data better?

- His answer is that it's not always (much) better.
- But that big data can certainly be better if the data is *richer and more fine-grained*.

Lessons Learnt from the examples

What lessons have we learnt from these “data” examples?

- NYC data
  - data requires work to clean up,
  - be creative about sources
- traffic prediction
  - combine many sources
  - you might have to generate some of your own
- predictive analytics for banks
  - fine-grained data really helps, but is harder to use

Introduction to Resources
Open data

organisations provide machine readable to support data science

Start with the video *The year open data went worldwide* a TED talk by Prof. Sir Tim Berners-Lee (video, 6 mins)

Democratization of Data

From *“the New Data Republic: Not Quite a Democracy”* in MIT Sloan Review 2015

- from Hal Varian (at Google): “information that once was available to only a select few ... available to everyone”
- from Robert Duffner (at Salesforce): “finally puts crucial business information in the hands of those who need it”
- government and IT departments building data and infrastructure to allow sharing
  - e.g. USA Open Gov Initiative
- analytic tools, (desktop and web-based), available to analyse it
- but people need the right skills!
  - open data is all good and well, but people need to be able to use it too!

Open Data Recommendations

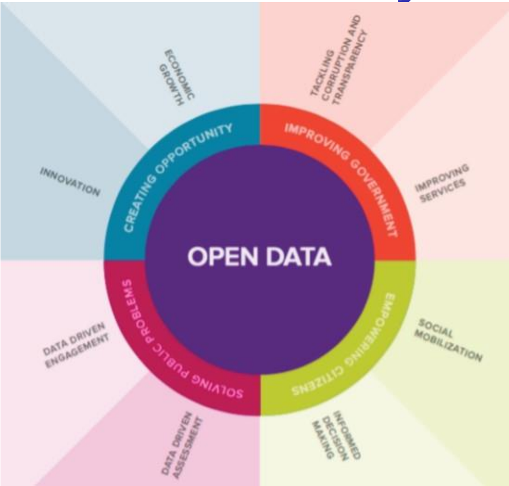
The reports:

- *“Open data:* Unlocking innovation and performance with liquid information” by MGI, and
- *“Science as an open enterprise”* by the Royal Society (UK)

claim that:

- open data provides new opportunities for business, new products and services, and can raise productivity
- open data supports public understanding and citizen engagement
- scientists need to better publicise their data (with help from universities, *etc.*)
- industry sectors should work with regulators and coordinate industry collaboration
- collaboration across sectors in both public and private settings, *e.g.*, disaster response, education

Open Data Taxonomy of Impact



see *“the-global impact of open data”* for a large catalogue of examples

Open Data: Impact for Weather Data

- National Oceanic and Atmospheric Administration (NOAA) in the USA started creating open data portal in early 2000’s
- now supports many industries and organisations
  - The Weather Channel and others
  - energy utilisation prediction for utilities
  - support for emergency response, coastal shipping, storm surge, tornado warnings, agricultural forecasting
  - support for environmental monitoring
  - weather derivatives financial industry

  N.B. several of these use predictive modelling!
- both governmental (local, state, international) support and commercial support

What’s Wrong with Open Data Sites

The Scientific American report:

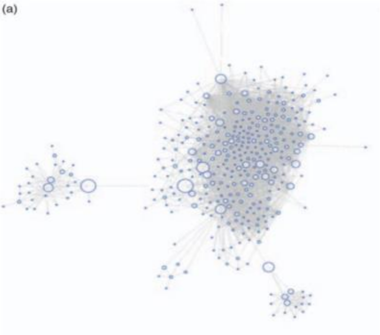
- *“What’s Wrong* with Open-Data Sites–and How We Can Fix Them”

discusses:

- its hard to make sense of the huge amount of government data
  - Data.GOV has 230k datasets, and Data.GOV.AU has 30k
- authors developed *Data USA*
- merge multiple datasets and transform them into stories
- the stories show people what data is available

Open Data ...

- A common format for open data is “Linked Open Data (LOD)”
- Remember graph database
  - Triples: subject, verb and object
  - DBPedia page for “Arnold Schwarzenegger



Linked Open Data

LOD project started by inventor of the Web,
Prof. Sir Tim Berners-Lee, OM, KBE, FRS, FREng, FRSA, DFBCS.

Aim of Linked Open Data (LOD) is to make data accessible, machine readable and **self-describing**.

- objects given a URI (like a URL)

  e.g. NYT or *Eighth Avenue in Manhattan*
- relationships between two objects represented as a triple:

  (subject, verb, object)
- relation itself is another URI
- data has an open license for use
- see this *tutorial on LOD* by Tom Heath

Open Data - Summary

- Publicly available
  - government and IT departments building data and infrastructure to allow sharing
  - e.g., Data.GOV has 230k datasets, and Data.GOV.AU has 30k
- Machine readable
- But..
  - it is not always usable
  - people need the right skills

FLUX Question

Graph database is commonly used to store…?

A. Structured data
B. Open data
C. Linked open data
D. None of the above options

Introduction to Resources
Wrangling

manipulating data to make it directly usable for analysis

What is Data Wrangling?

Process of transforming “raw” data into data that can be analyzed to generate valid actionable results and insights

Why Wrangling?

- Working with raw data is challenging!
  - Data comes in all shapes and sizes
  - Different files have different formatting
  - Mistakes in data entries

We need techniques to cleanse and prepare data

Wrangling Examples

Examples of wrangling tasks:

- extract the core news text, title, and date from a webpage:
  Apple’s iPhone loses top spot to Android in Australia
  - some news sites support “Reader View” which deletes a lot of the additional adverts/fluff/indices, but no all
- extract the text plus details from a PDF file:
  “Data Wrangling: The Challenging Journey ...”
- extract all article titles from an XML file:
  PUBMED results xml
- digitize the text from a scanned image:
  scanned letter
- extract all the sentences referring to particular individual in an article:
  a news article about Hillary Clinton

FLUX Question

One example of data wrangling is extract dates from text and converting them to a digitized date format.

Which of the following text can be a challenge in converting them to a digitized date format.

A. next Tuesday
B. January 3 next year
C. 3rd Friday in the month
D. 03/12/18
E. All of the text

Wrangling Examples (cont.)

More wrangling tasks:

- integrate data sources:
  company has customer records in 4 different databases in different formats; you want a single standardised set of customer names and addresses
- geocoding:
  convert addresses in your customer database into geographic latitude and longitude
- convert free text dates to standard format:
  e.g. map: “next Tuesday” → “2nd January 15”,
  other date examples: “January 3 next year” , “3rd Friday in the month” “03/31/15”, “31/03/15”

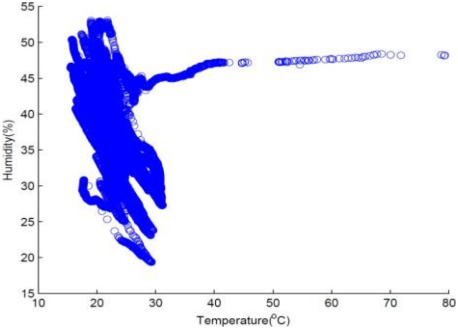
Wrangling Examples (cont.)

More wrangling tasks:

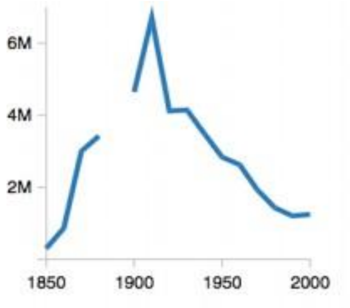
- recognise missing values and deal with them, by e.g.
  - removing the row or column,
  - replace with a special “unknown” value,
  - replace with an average value,
  - or doing nothing
- deal with outliers or “illegal” values,
  e.g. remove extremely large values that are likely due to sensor noise
- discretise the data into a set of values
  discretisation is necessary if the predictive model being learnt cannot handle continuous data

Data wrangling- visualisation

- Outliers



- Missing data



FLUX Question

How to deal with missing data?

A. Removing the row or column
B. Replace with a special “unknown” value
C. Replace with an average value

Standards and Issues
(ePub section 4.5)

more on standards and issues

- some standards
  - some standards for semi-structured data, data science process and predictive models
- open data and open source software
  - critical infrastructure and tools
- APIs and SaaS
  - think Web 3.0

Example Standards

Examples of standards

- Metadata standards
  - such as *Dublin Core*, examples at *A Gentle Introduction to Metadata*
- XML formats for sharing models,
  - e.g. PMML (see below)
- Standards for describing the data mining/science process,
  - such as *CRISP-DM*
- Standard vocabularies for use in Medicine, e.g.
  - health codes: disease and health problem codings *ICD-10*
  - systematized nomenclature of medicine, clinical terms, *SNoMed-CT*

Standards support cooperation, reuse, *etc.*

What other sorts of things might you have standards for?

Model Language

PMML ::= Predictive Model Markup Language

PMML provides a standard language for describing a (predictive) model that can be passed between analytic software (e.g. from R to SAS).

- *PMML: An Open Standard for Sharing Models*
- A list of products working with PMML is the *PMML Powered page* on DMG site.

Data Science Process

We’ve seen many data science processes and lifecycles:

- e.g. our own “standard Data Science value chain”
- *CRISP-DM* discussed previously, is a standardised data science process
- statisticians sometimes use the term exploratory data analysis for part of the process

Semi-Structured Data

Semi-structured data is data that is presented in XML or JSON:

- see some examples for *here*
- Note YAML (Yet Another Markup Language), which is just an indentation (easier to read) version of JSON
- standard libraries for reading/writing/manipulating semi-structured data exist in Python, Perl, Java
- don't need to know all the details of XML (and related Schema languages)

  many good online tutorials, *e.g. W3schools.com*
- their use in systems leads to the open world assumption about data, where we may download relevant data on the fly from APIs *etc.*

Unit Schedule: Next Week

Module	Week	Content
1.	1	overview and look at projects
	2	(job) roles, and the impact
2.	3	data business models
	4	application areas and case studies
3.	5	characterising data and "big" data
	6	data sources and case studies
4.	7	resources and standards
	8	**resources case studies**
5.	9	data analysis theory
	10	data analysis process
6.	11	issues in data management
	12	GUEST SPEAKER & EXAM INFO