Introduction to Data Science Module 4

Data Resources, Processes, Standards and Tools

FIT1043 2018 Lecture 6

Monash University

Discussion: R Language

- Powerful language for visualising and building predictive models of data

- Very easy to use with lots of inbuilt functionality.

- Great for exploratory data analysis

- Not as scalable as programming languages: Java, Python, C++

MARS Question

Which of the following statements about SQL and NoSQL databases is TRUE?

A. SQL is suitable to store unstructured data
B. Both SQL and NoSQL are suitable when data changes rapidly
C. None of the above options

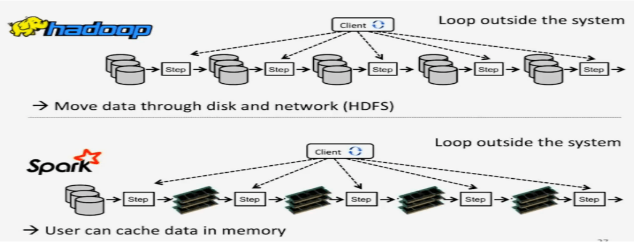
Discussion: SQL and NoSQL

- Use SQL database when:
  - Data is structured
  - Data is unchanging

- Use NoSQL database when:
  - Data has little to no structured data
  - Data changes rapidly

Discussion: Hadoop and Spark

- Hadoop
  - not suited to streaming (suitable for offline processing)
- Spark
  - include Map-Reduce capabilities
  - provides real-time, in-memory processing
  - much faster than Hadoop



Discussion: Unix Shell

Useful for managing and manipulating **large files without ever loading them fully into memory**

- using pipes allow us to process files as a stream
- allows us to deal with files that are too big for

  applications and/or don't fit into memory

Shell contains many useful commands, like

- **less** to view large files
- **grep** to search large files
- **awk** to process them one line at a time

Unit Schedule: Modules

Module	Week	Content
1.	1 2	Overview and look at projects (Job) roles, and the impact
2.	3	Data business models / application areas
3.	4 5	Characterising data and "big" data Data sources and case studies
4.	6 7	**Resources and standards** **Resources case studies**
5.	8 9 10	Data analysis theory Regression and decision trees Data analysis process
6.	11 12	Issues in data management GUEST SPEAKER & EXAM INFO

Introduction to Resources (ePub section 4.1)

- Where to find and how to use data sources
  - Examples
- Open data
  - Machine readable and publicly available
- Data Wrangling
  - Data manipulation and preparation for data analysis

Introduction to Resources: Finding and using data

access to new data sources or clever and creative use of existing multiple data sources are very important in a data science project

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

Introduction to Resources:
Open data

organizations provide machine readable to support data science

Open Data

- 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

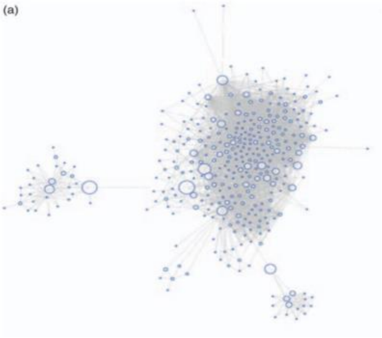
MARS Question

Graph database is commonly used to store…?

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

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



Introduction to Resources: Data Wrangling

manipulating data to make it directly usable for analysis

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

What is Data Wrangling?

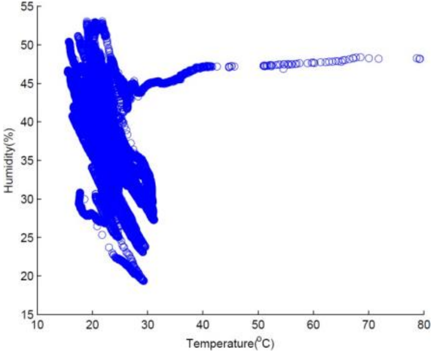
Process of transforming “raw” data into data that can be analyzed to generate valid actionable results and insights

Data Wrangling

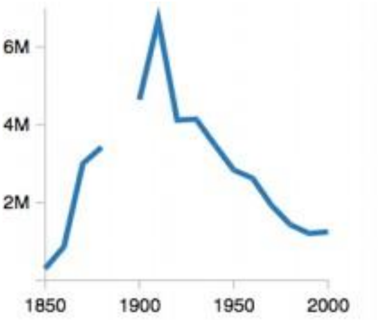
- Data pre-processing
- Data preparation
- Data cleansing
- Data transformation
- etc

Data Wrangling- Visualisation

Outliers data



Missing



MARS 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

Homework!

Watch a TED talk by Prof. Sir Tim Berners-Lee about Open Data:

“The year open data went worldwide”

Unit Schedule: Next Week

Module	Week	Content
1.	1 2	Overview and look at projects (Job) roles, and the impact
2.	3	Data business models / application areas
3.	4 5	Characterising data and "big" data Data sources and case studies
4.	6 7	**Resources and standards** **Resources case studies**
5.	8 9 10	Data analysis theory Regression and decision trees Data analysis process
6.	11 12	Issues in data management GUEST SPEAKER & EXAM INFO