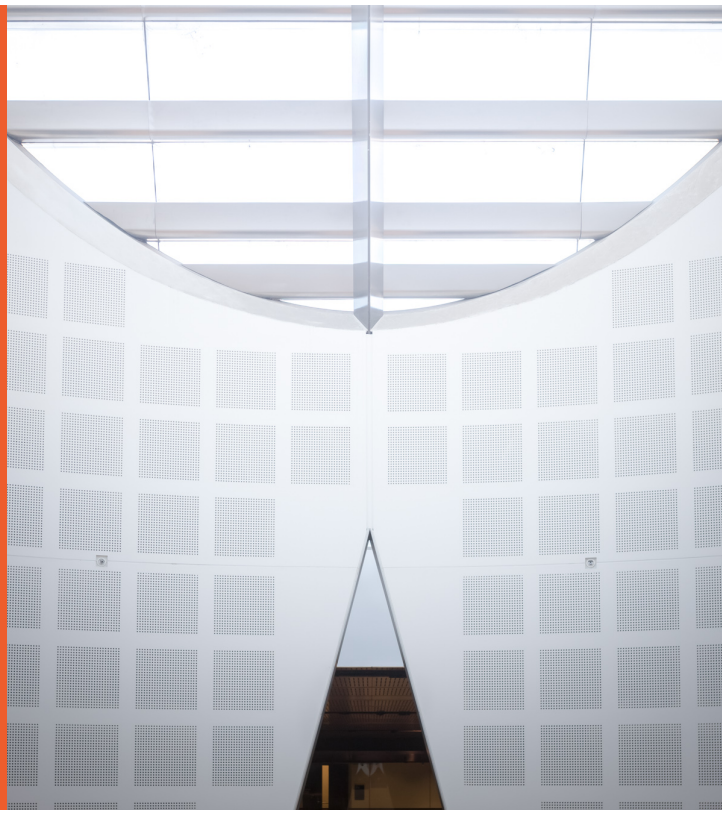
DATA2001 – Data Science, Big Data and Data Diversity

Presented by
A/Prof Uwe Roehm
School of Computer Science

THE UNIVERSITY OF SYDNEY

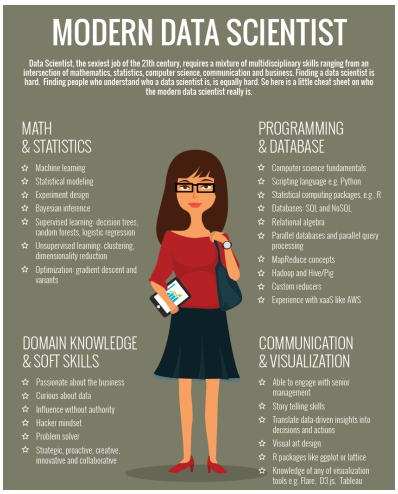


What is a Data Scientist?

THE UNIVERSITY OF SYDNEY

Data Scientists build intelligent systems to derive knowledge from data.

Data Science skills



http://www.marketingdistillery.com/2014/11/29/is-data-science-a-buzzword-modern-data-scientist-defined/

Data scientists help organisations:

- understand their data,
- ask meaningful questions,
- derive transformative insights,
- lead empirically grounded decision making.

Example: Reducing costs through route optimisation

http://www.bloomberg.com/news/articles/2013-10-30/ups-uses-big-data-to-make-routes-more-efficient-save-gas

- Use customer, vehicle and delivery data
- 1 mile less per day for every driver saves $50M per year in fuel, maintenance and time
- Less idling, e.g., by avoiding left turns, saved 6 million liters of fuel in 2012

Example: Urban & Transport Planning, Public Health

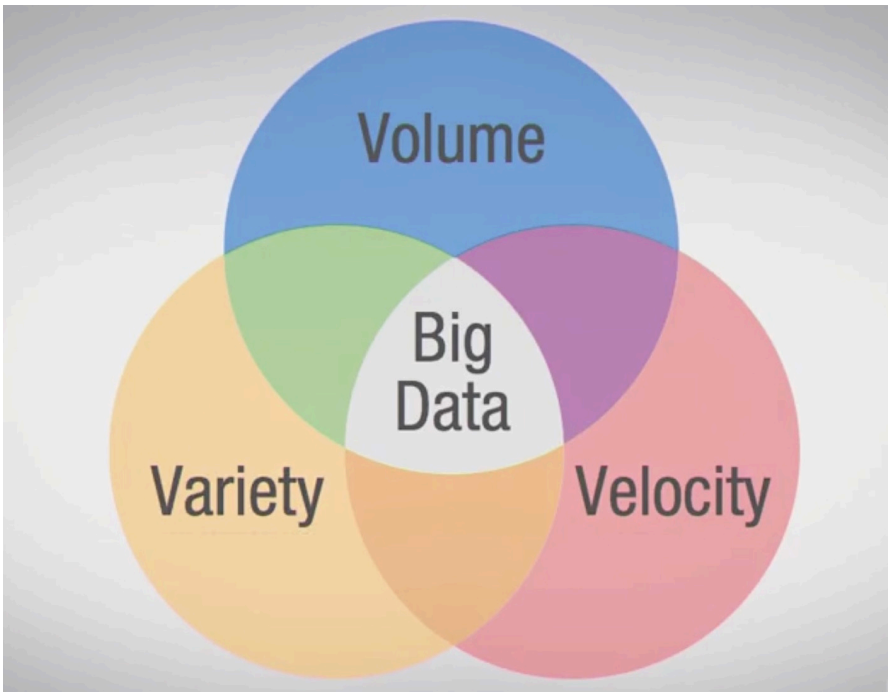
http://www.walkscore.com/research/

- Integration of data about road and public transport network with data about population, services, restaurants, amenities etc.
- Summarising *Walkability Score overlayed* on map visualisation
- Prediction of impact of new developments
- API for use in 3rd party apps, eg. supporting real estate agents

Big Data

the three Vs:

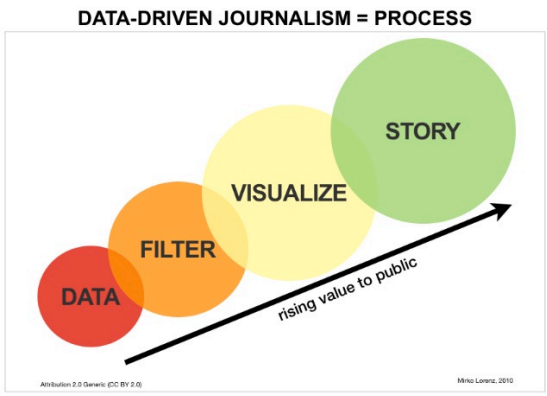
[cf. article by Doug Laney, 2001]



[Barton Poulson "Techniques and Concepts of Big Data", 2014]

Big Data: Data Volume

- Data Analysis problems growing too much
  - There is a certain size (volume) from which onwards manual processing is not an option anymore
- Example: Data-Driven Journalism
  - **April 2016: Panama Papers Leak**
  - leaked documents from Mossack Fonseca, a Panamanian law firm that sells anonymous offshore companies around the world
  - 2.6TB in 11.5 million documents, 214,000 companies
  - http://panamapapers.sueddeutsche.de/en/



[Source: Wikipedia]

Big Data: Variety

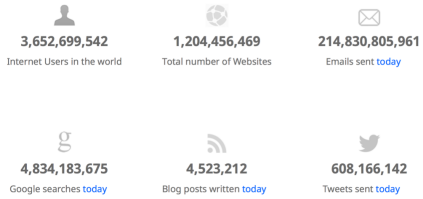
- Structured Data, such as CSV or RDBMS
- Semi-structured Data, such as JSON or XML
- Unstructured Data, ie. text, e-mails, images, video
  - an estimated 80% of enterprise data is unstructured

- study by Forester Research: **variety biggest challenge in Big Data**

Big Data: Velocity

- conventional scientific research:
  - months to gather data from 100s cases, weeks to analyze the data and years to publish.
  - Example: Iris flower data set by Edgar Anderson and Ronal Fisher from 1936

- on the other end of the scale: Twitter
  - average 6000 tweets/sec, 500 million per day or 200 billion per year
  - Cf. life Twitter Usage Statistics http://www.internetlivestats.com/twitter-statistics/



Sources of Big Data / More Vs

- Human-generated Big Data
  - E.g. photos, posts, likes etc
- Machine-generated data
  - Communication logs, Internet-of-Things, etc.

- More Vs of Big Data:
  - Validity (data quality), Variability (data consistency), Veracity (data accuracy / trustworthiness), Value…

DATA2001: Data Scientists need Big Data skills

- Data is either
  - too large (volume),
  - too fast (velocity), or
  - needs to be combined from diverse sources (variety)

  for processing with scripts or on single server.

- DATA2001: Focus on key technologies for large-scale data science
  - Data Models
  - Processing Abstractions
  - Data Scalability, Big Data

Exploratory Data Analysis with Python

THE UNIVERSITY OF SYDNEY

Why Python?

- **Interpreted**: direct execution without compilation
- **Dynamically-typed**: don't have to declare a static type
- **Readable**: easy-to-understand syntax
- **Deployable**: easy to incorporate in applications
  - The Dropbox desktop client is written entirely in Python (>40 million users)
- **Productivity**: facilitates rapid, interactive prototyping

Python Import System

- Modules built by third party can be employed easily
- gives access to classes and functions within the module
  - Example: **csv:** comma-separated file format support

```
import csv
for row in csv.reader( ['one,apple,green', 'two,tomato,red'] ):
    print(row[1])
```

  - alternative usages to introduce shortcuts or import only certain functions:

```
import csv as X
from csv import DictReader
```


Python has excellent open-source data libraries

- **scipy**: libraries for scientific and technical computing
  - **numpy**: support for large multidimensional arrays and matrices
  - **matplotlib**: port of matlab plotting functionality
- **scikit-learn**: machine learning library
- **nltk**: natural language toolkit

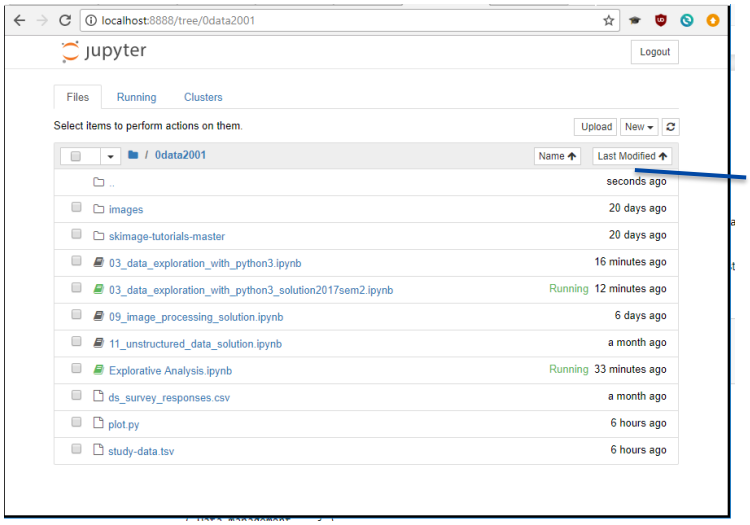
- **pandas**: R-like data frame and associated manipulations

Installing Python and Jupyter using Anaconda

- We make some Jupyter servers available
  - but those can be slow
- You can also install Python and Jupyter privately using the Anaconda Distribution, which includes Python, the Jupyter Notebook, and other commonly used packages for scientific computing and data science.

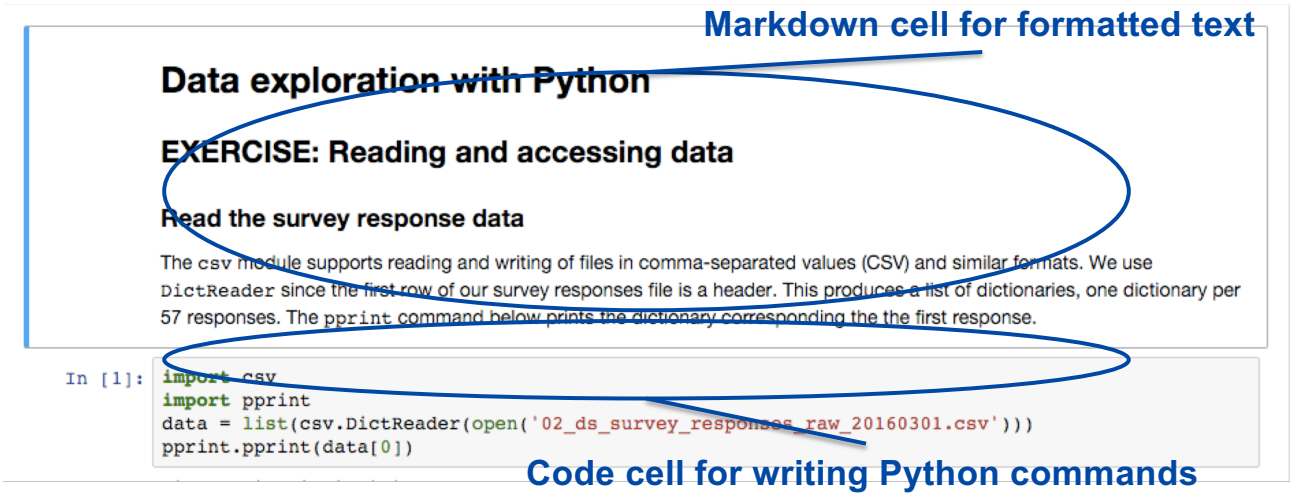
Jupyter Notebooks support interactive Data Science

- IPython interactive command shell offers:
  - Introspection
  - Tab completion
  - Command history
- Jupyter runs in a browser and supports:
  - Sharing and documenting of live code
  - Data cleaning, visualisation, machine learning, …
  - Jupyter's gallery of interesting notebooks: https://github.com/ipython/ipython/wiki/A-gallery-of-interesting-IPython-Notebooks
- SIT provides a Jupyter server which runs Python 3



1. Click here for file open dialogue
2. Click upload next to file name

Jupyter notebook cells



Jupyter menu bar