

DATA2001 – Data Science, Big Data and Data Diversity

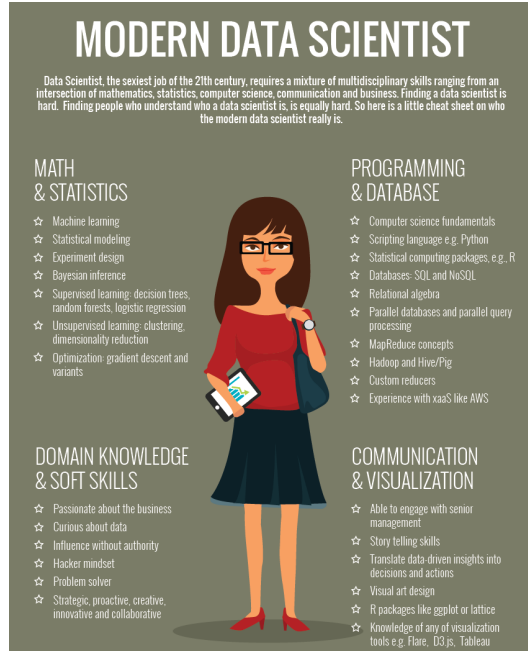
Dr Ali Anaissi

School of Computer Science



What is a Data Scientist?

Data Science skills



Data scientists help organisations:

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

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

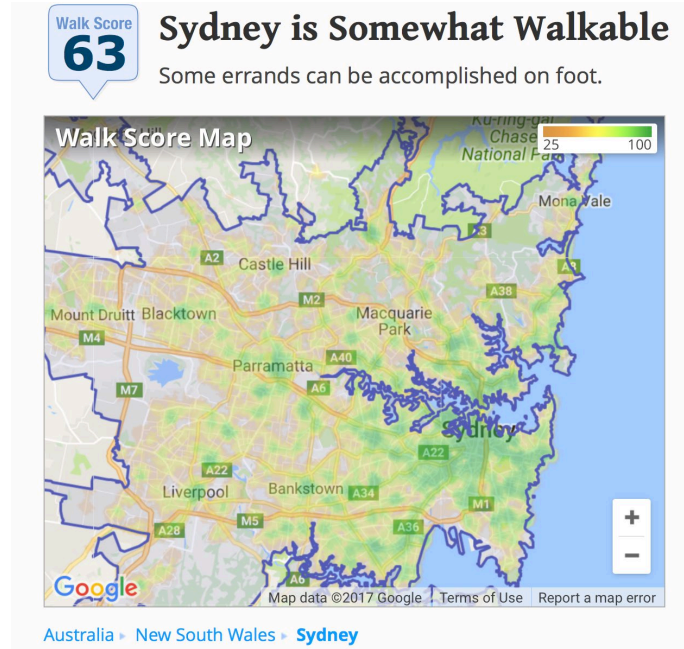
Example: Reducing costs through route optimisation



- 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

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

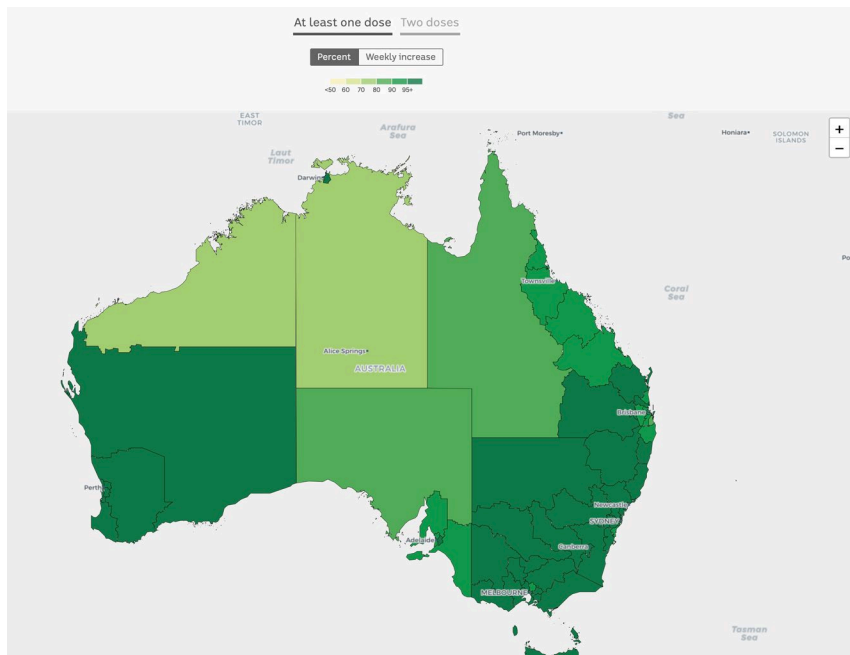
Example: Urban & Transport Planning



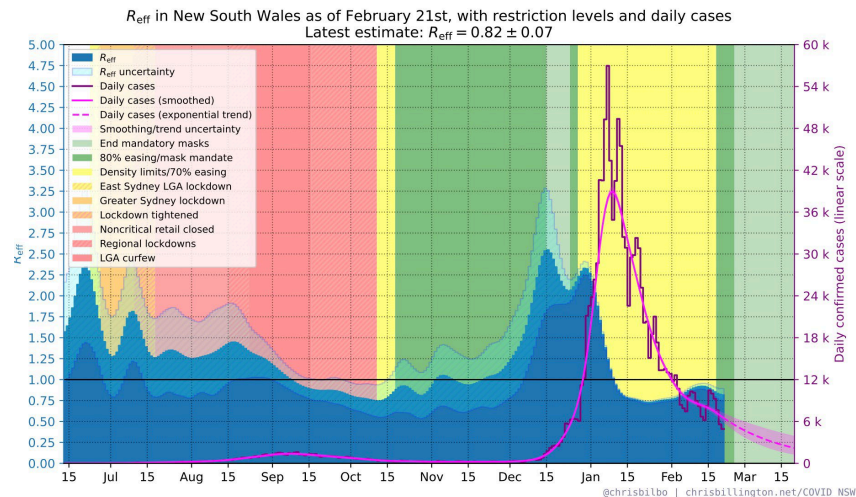
- Integrated various datasets: population, road and public transport network, services, restaurants, amenities etc.
- Summarising *Walkability Score*
- overlaid on map visualisation
- Prediction of impact of new developments
- API for use in 3rd party apps, eg. supporting real estate agents

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

Example: Public Health – data-driven decision making



<https://www.abc.net.au/news/2021-03-02/charting-australias-covid-vaccine-rollout/13197518>

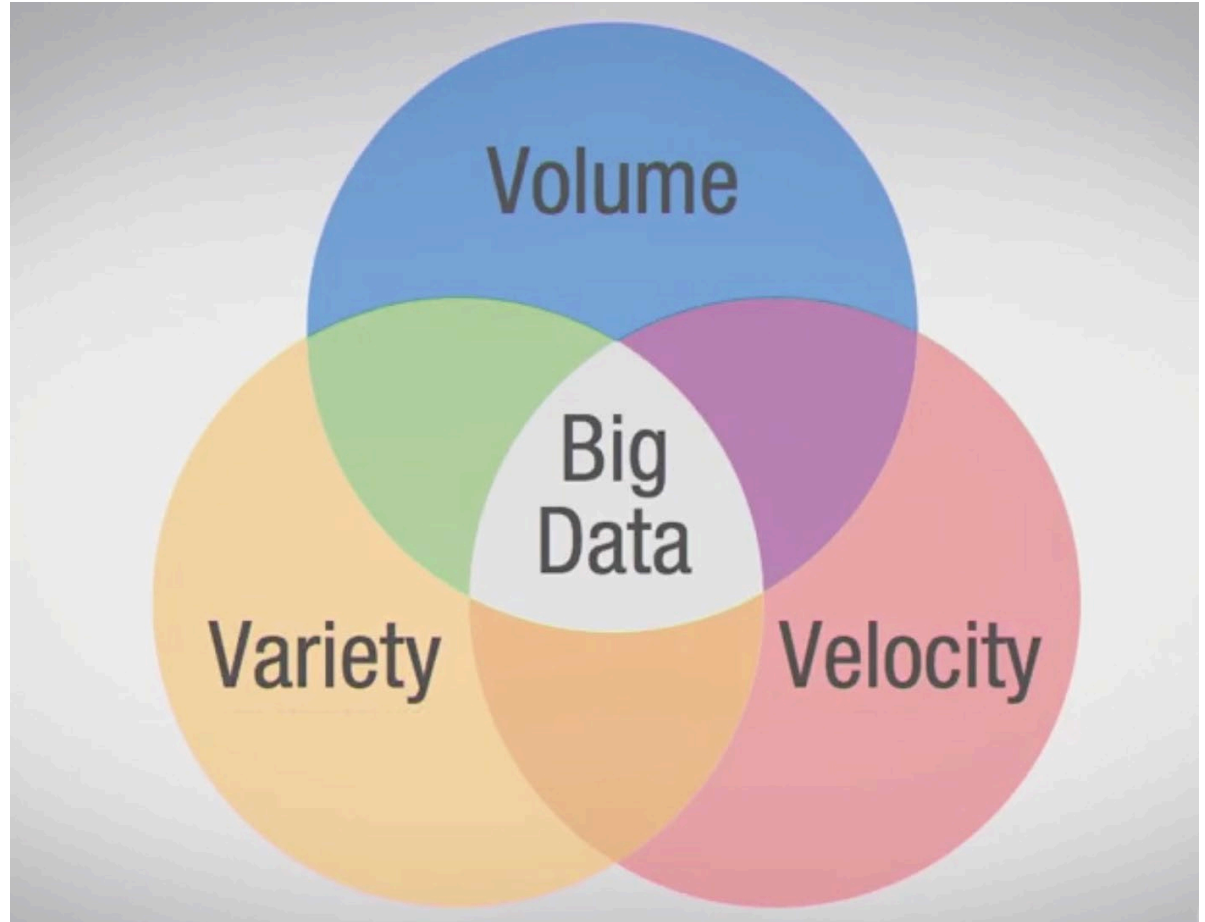


https://chrisbillington.net/COVID_NSW.html

Big Data

the three Vs:

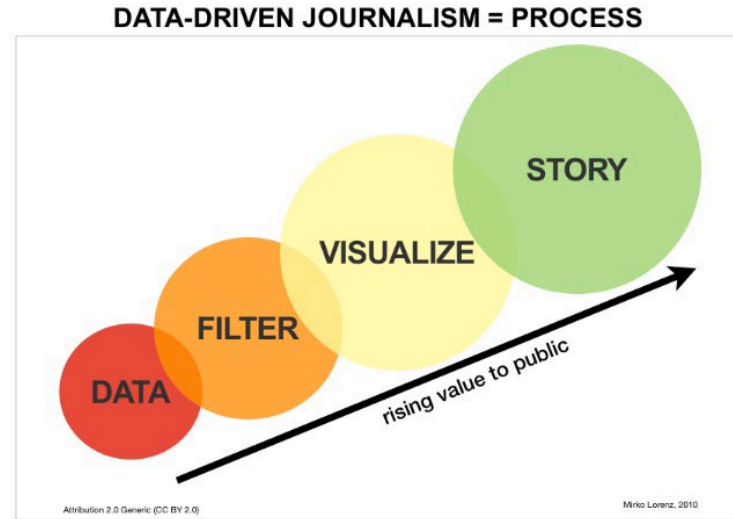
[cf. article by
Doug Laney, 2001]



[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/>



3,652,699,542
Internet Users in the world


1,204,456,469
Total number of Websites


214,830,805,961
Emails sent [today](#)


4,834,183,675
Google searches [today](#)


4,523,212
Blog posts written [today](#)


608,166,142
Tweets sent [today](#)

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

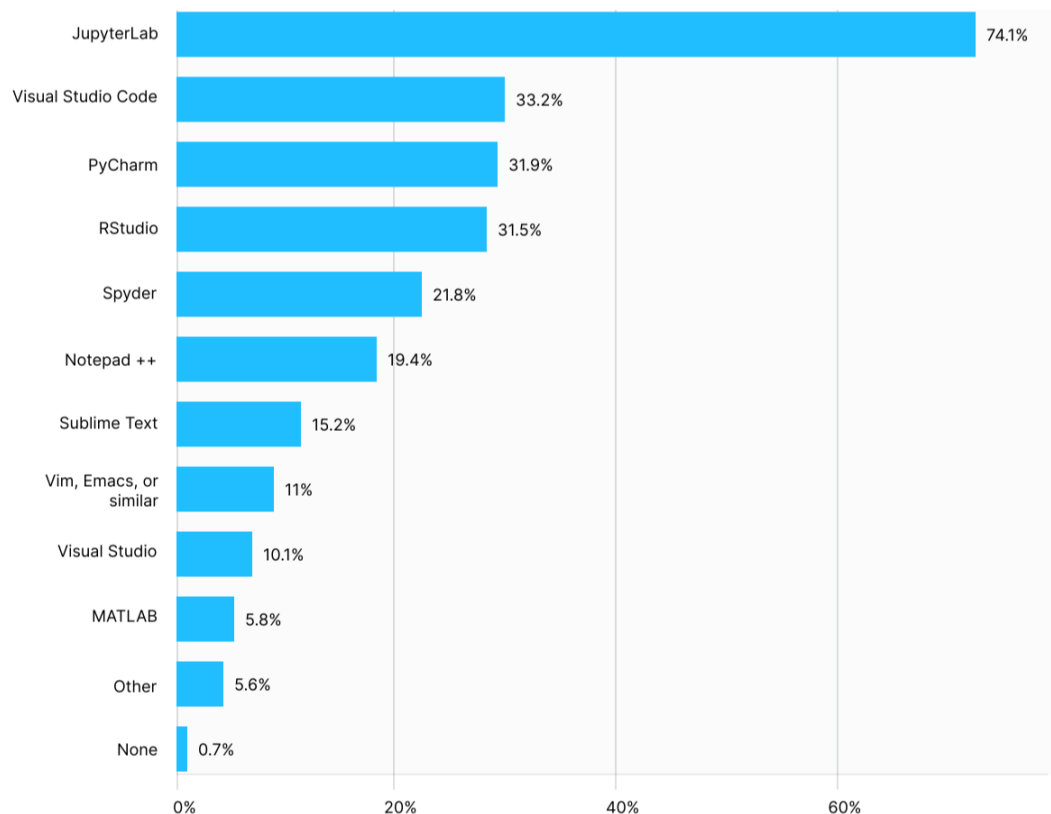
Exploratory Data Analysis with Python

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

Kaggle "State of Data Science" Survey 2020

POPULAR IDE USAGE



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
```

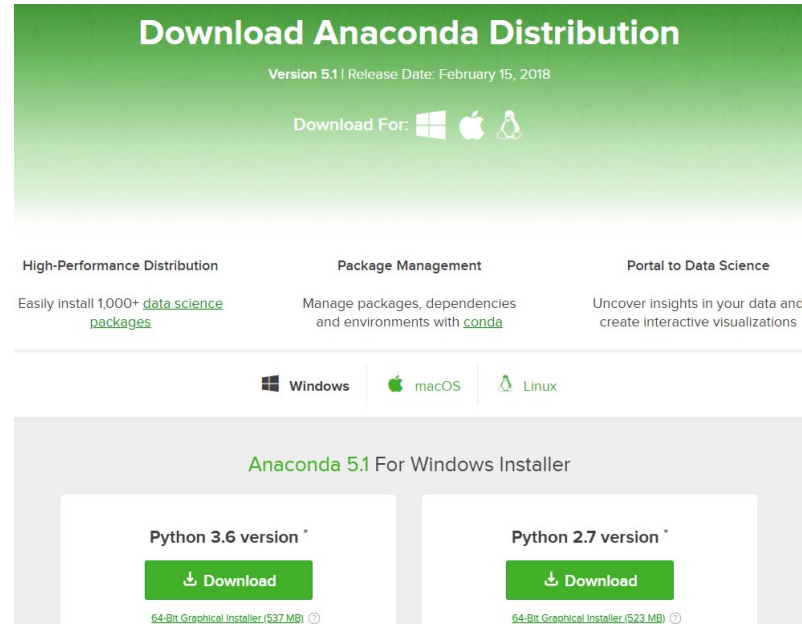
Python has excellent open-source data libraries

- **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



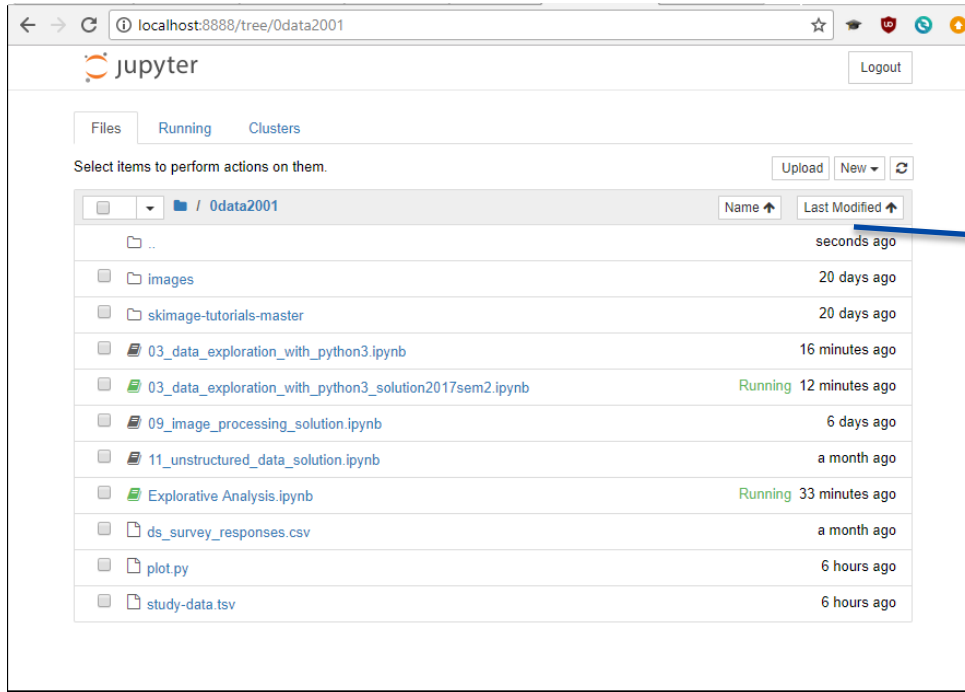
Installing Python and Jupyter using Anaconda

- 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.
- Alternatively, try Google Colab (colab.research.google.com)
 - But might give problems in using our university database servers later in the semester...



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

Markdown cell for formatted text

Data exploration with Python

EXERCISE: Reading and accessing data

Read the survey response data

The `csv` module supports reading and writing of files in comma-separated values (CSV) and similar formats. We use `DictReader` since the first row of our survey responses file is a header. This produces a list of dictionaries, one dictionary per 57 responses. The `pprint` command below prints the dictionary corresponding to the first response.

```
In [1]: import csv
import pprint
data = list(csv.DictReader(open('02_ds_survey_responses_raw_20160301.csv')))
pprint.pprint(data[0])
```

Code cell for writing Python commands

Jupyter menu bar

