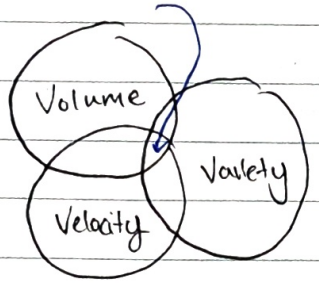
Wk 1 - lecture

Data Science skills

- Math & Science
- Programming & statistics
- Domain knowledge
  - ex) ↓ costs through route optimisation → greater impact considering more vehicles, time
  - ↳ get knowledge from experts, domain changes by research topics.
- communication

Big Data



Volume: what size — never ending with manual processing ½

Velocity: how fast data added ~~changing~~ — how many ID vids twitter twits

Variety: how various the structure is — review

biggest challenge

csv, json, text, image, emails, etc

structured | semi- | unstructured

≈ 80% of data from enterprises

From certain volume, manual processing is NOT AN OPTION.

ex) Panama paper leak (TBs of data)

Data-Driven Journalism

data → filter → visualize → story

Data can be

- too big
- too fast
- needs to be combined from diverse sources
  - ↳ scripting &

Why Python?

R: for research

Python: for industry projects

- interpreted x compiling
- dynamically typed ← data types may vary (errors occur in java)
- readable (:) for devs
- deployable ← server-side language
- productivity ← facilitates rapid & interactive prototyping

Most common IDE

- Jupyter Notebook ← download in Anaconda & prefered.
  - or google collab
  - or ~~Kaggle~~ uni server
  - ✓ live code share
  - ✓ code by 'cells'
- VS code
- R studio
- etc

Importing packages.

```
import csv (as ___)
from csv import DictReader
```

) b easy importing !? 👍

Open-Source libraries

- scipy : scientific & technical computing - scipy.org
  - numpy : arrays & matrices
  - matplotlib : visualization
- scikit-learn : machine learning
- nltk : natural language processing
- pandas : R-like dataframe manipulation