THE UNIVERSITY OF
MELBOURNE

Workshop 1

COMP20008

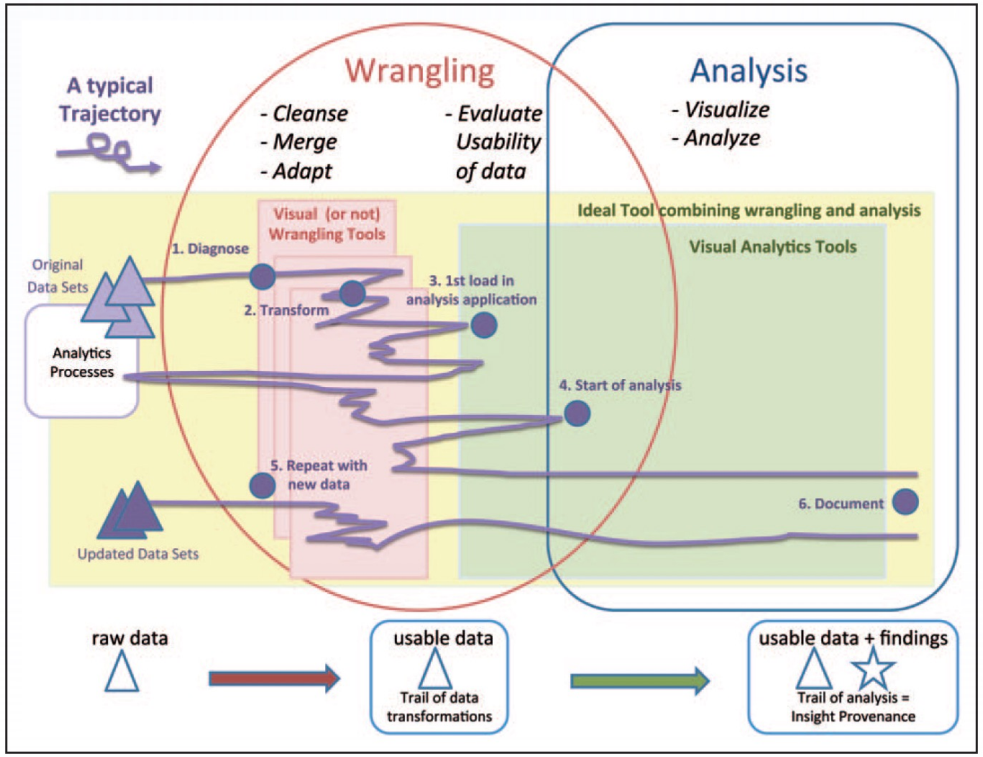
Elements of Data Processing
Zijie Xu

Agenda

- First subject in Data Science
- Jupyter Notebook
- Pandas
    - Series
    - DataFrame
    - Group by

First subject in Data Science

- "Foundations of Data Science"
- Prerequisite to COMP30027 Machine Learning
- Learn about the pipeline of data science



Research directions in Data Wrangling:
visualisations and transformations for crediible data. S. Kandel et al, Information Visualisation 10(4), 2011.

Jupyter Notebook

- Great tool for displaying code projects along with text and visualisations (hence “notebook”)
  - De facto industry standard
  - A server-client app with a backend computing kernel + frontend website for editing

- A notebook consists of
  - Markdown cells: text, figures, HTML etc.
  - Code cells: runnable python code blocks

Pandas

- A software package for data analysis and manipulation in python
  - Open-sourced since 2009
  - Also de facto industry standard (by now)
  - Key benefits: fast, flexible & memory-efficient

```
import pandas as pd
```

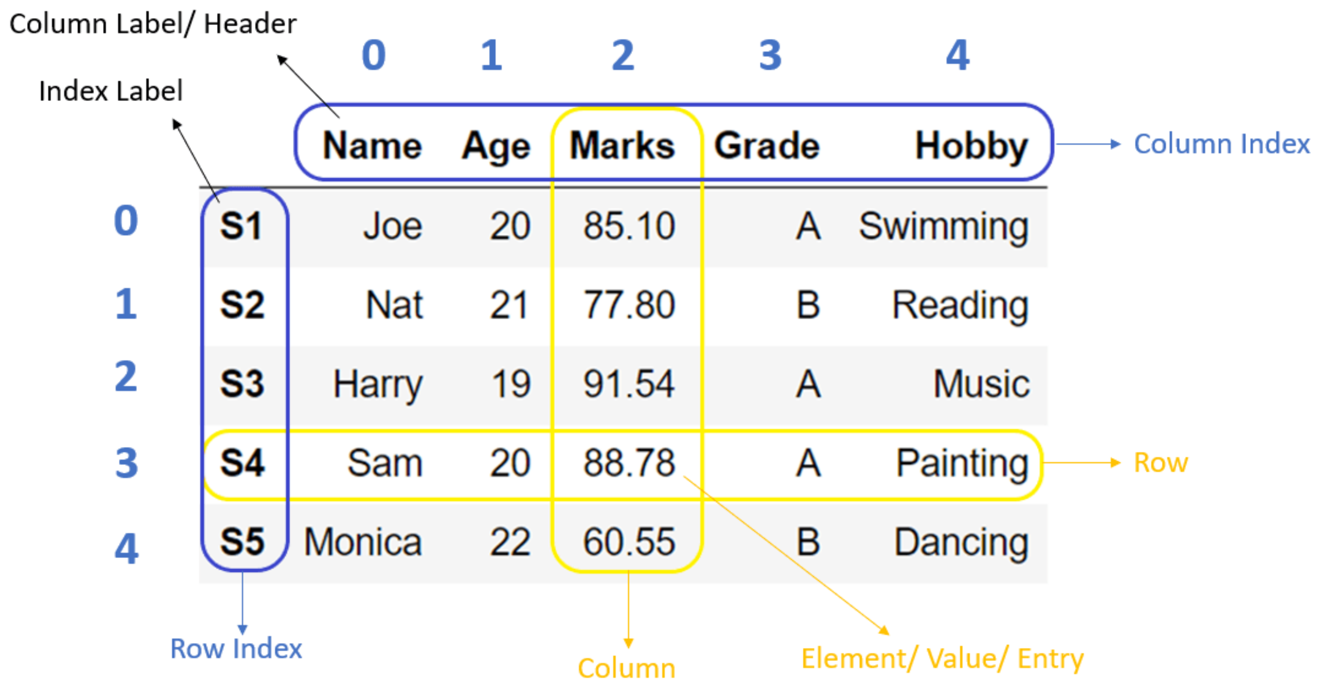
Series and DataFrame

pd.Series

- A column of data

pd.DataFrame

- A spreadsheet
- Multiple columns of data



	Name	Age	Marks	Grade	Hobby
S1	Joe	20	85.10	A	Swimming
S2	Nat	21	77.80	B	Reading
S3	Harry	19	91.54	A	Music
S4	Sam	20	88.78	A	Painting
S5	Monica	22	60.55	B	Dancing

Key skill for COMP20008: Learn to read API documentation

THE UNIVERSITY OF
MELBOURNE

Thank you

More Resources: Canvas