



THE UNIVERSITY OF
MELBOURNE

Workshop 1

COMP20008

Elements of Data Processing
Zijie Xu



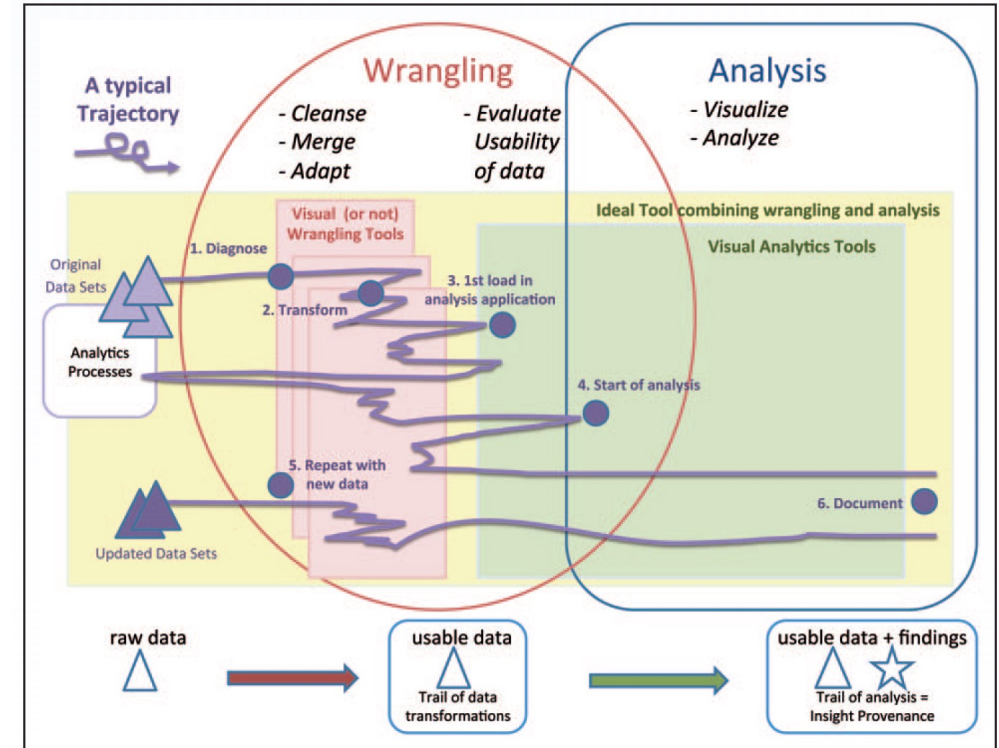


Agenda

- 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 credible 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
```

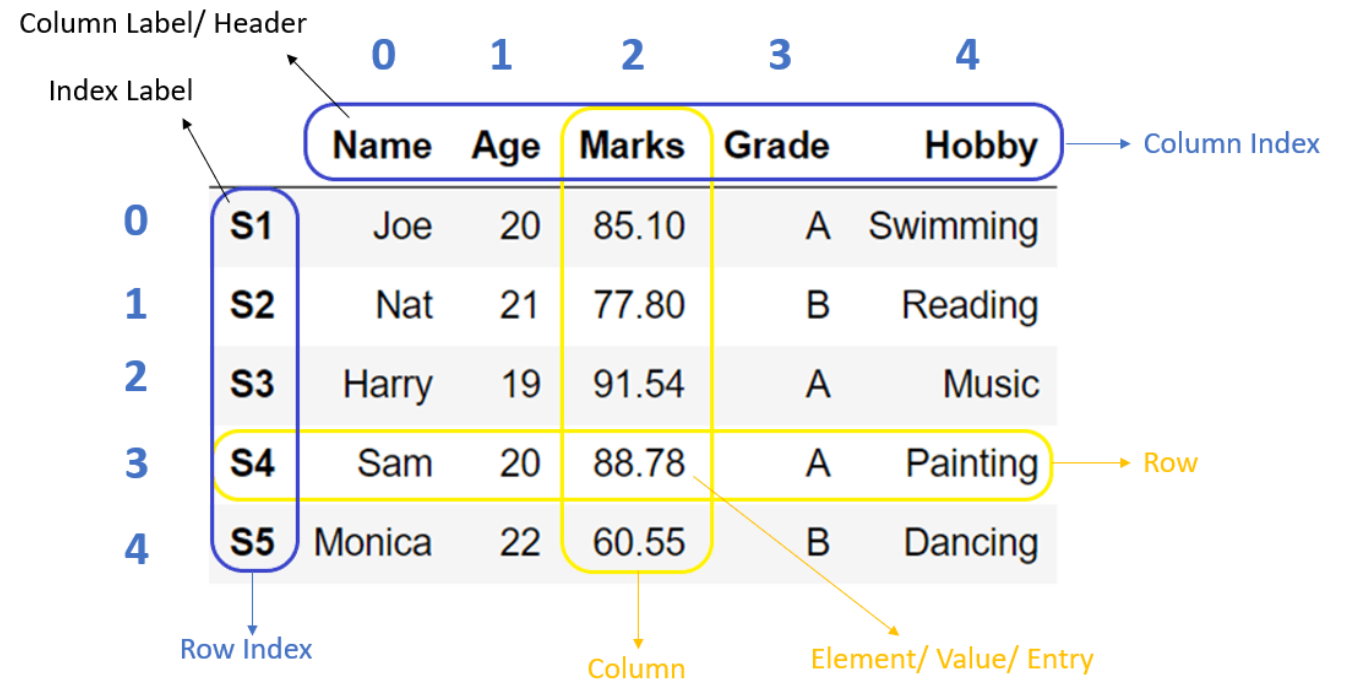
Series and DataFrame

pd.Series

- A column of data

pd.DataFrame

- A spreadsheet
- Multiple columns of data



The diagram illustrates a pandas DataFrame with the following structure:

	Column Label/ Header	0	1	2	3	4
Index Label		Name	Age	Marks	Grade	Hobby
0	S1	Joe	20	85.10	A	Swimming
1	S2	Nat	21	77.80	B	Reading
2	S3	Harry	19	91.54	A	Music
3	S4	Sam	20	88.78	A	Painting
4	S5	Monica	22	60.55	B	Dancing

Annotations in the diagram:

- Column Index:** Points to the header row (Name, Age, Marks, Grade, Hobby).
- Row Index:** Points to the index column (S1, S2, S3, S4, S5).
- Column:** Points to the 'Marks' column.
- Row:** Points to the 'S4' row.
- Element/ Value/ Entry:** Points to the value '88.78' at the intersection of row 'S4' and column 'Marks'.

Key skill for COMP20008: Learn to read API documentation



THE UNIVERSITY OF
MELBOURNE

Thank you

More Resources: Canvas