UNIVERSITY OF PORTSMOUTH

Python for Data Analysis

Modeling in Python – KNN

(TB2 - Week 1)

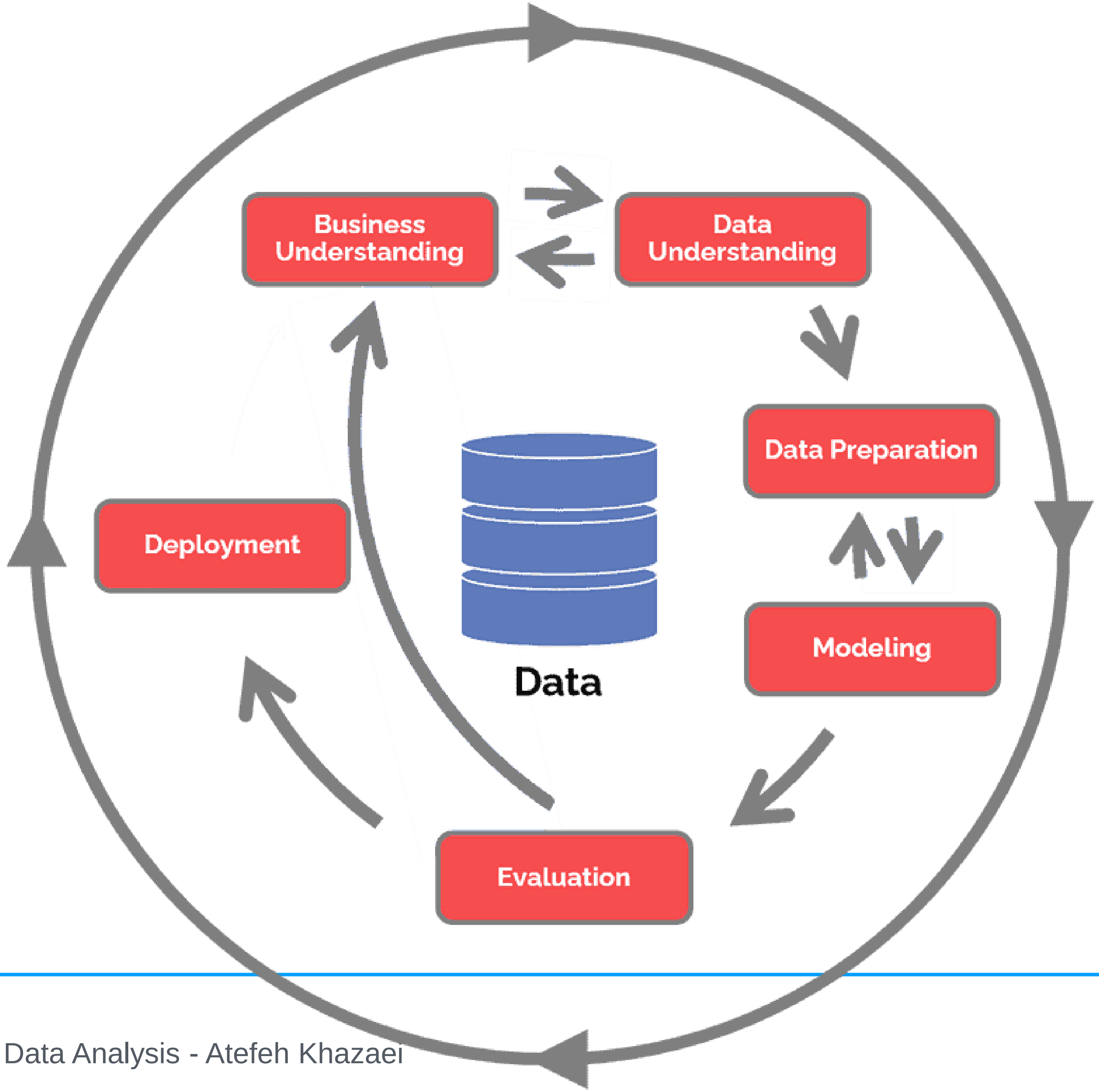
Atefeh Khazaei

atefeh.khazaei@port.ac.uk

What we will learn this week?

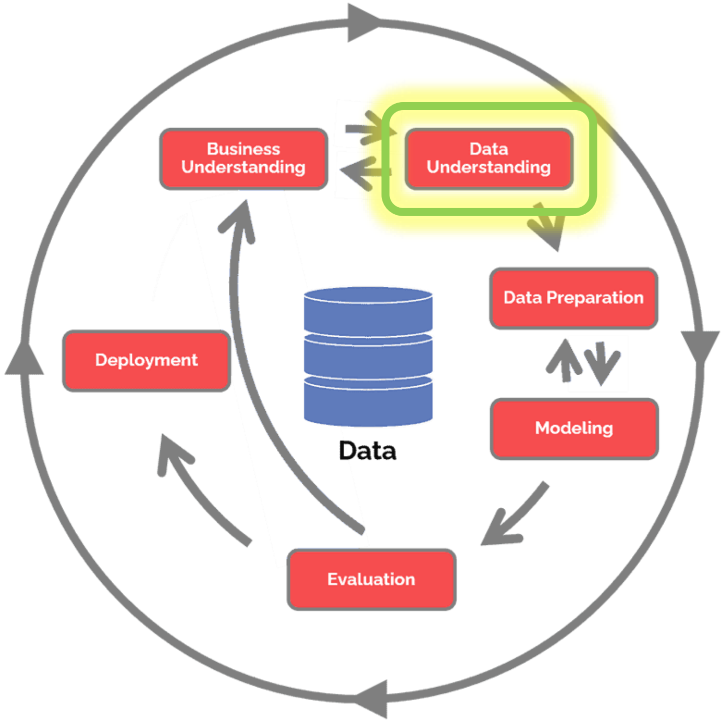
- ❑ Reviewing the Previous Teaching Block
- ❑ Modelling Algorithms
  - ❑ K-Nearest Neighbour

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

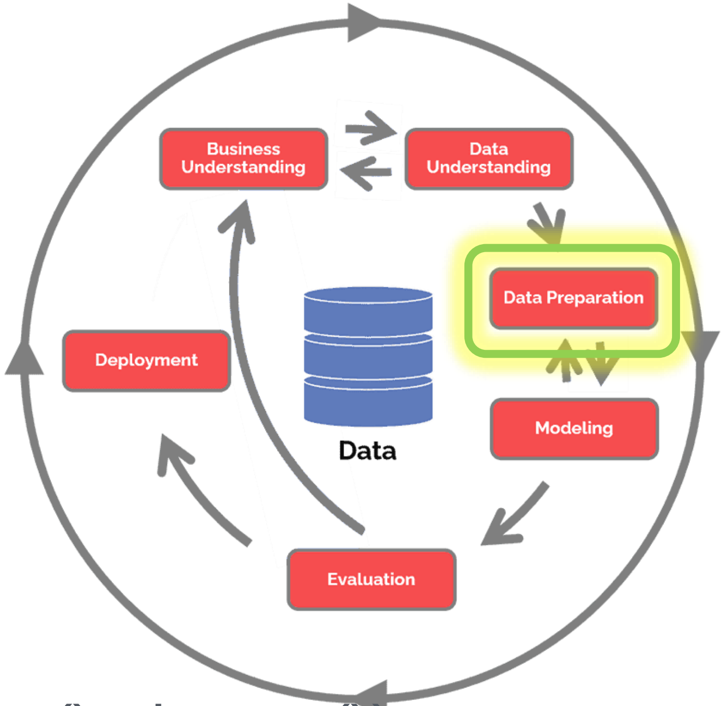
CRISP-DM
Data Understanding

- ❑ Descriptive functions
  - ❑ type(), head(), tail(), info(), describe()
- ❑ Plotting and Visualisation
  - ❑ Line, Bar, Histogram, Density, Scatter, Box Plots



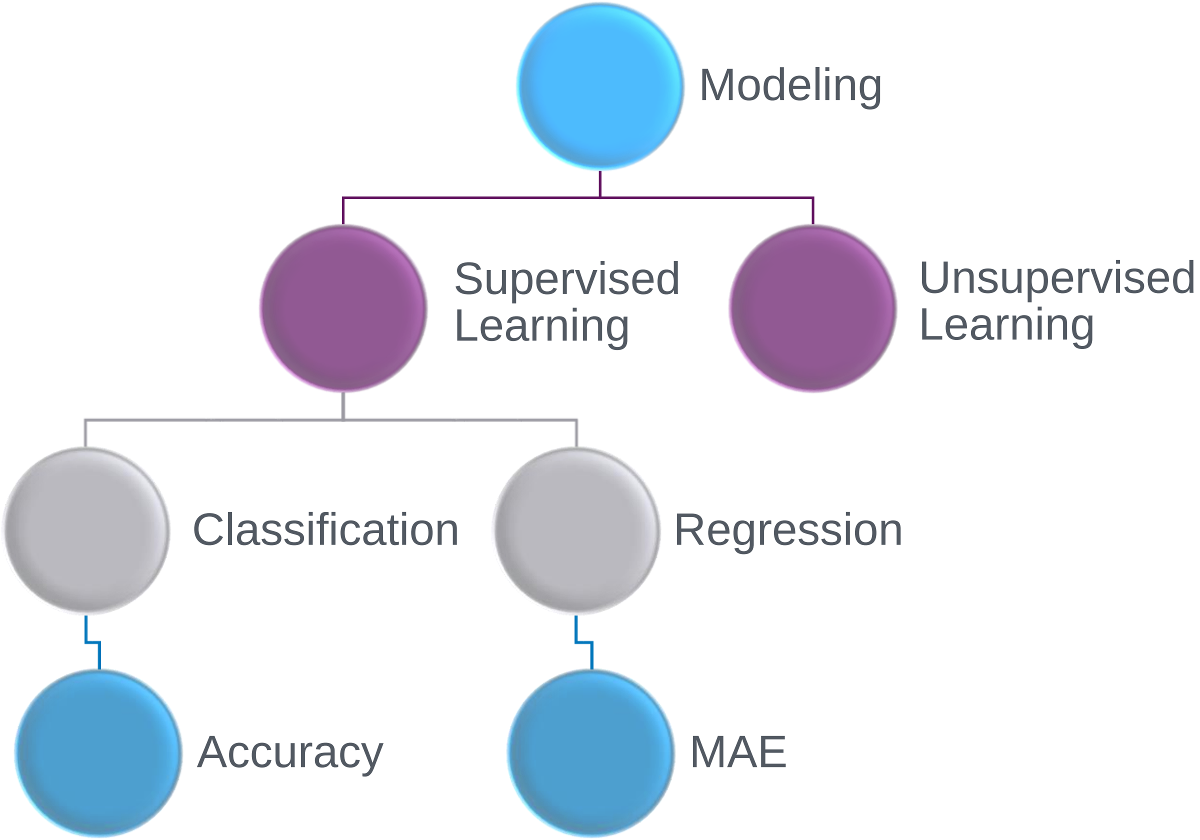
CRISP-DM

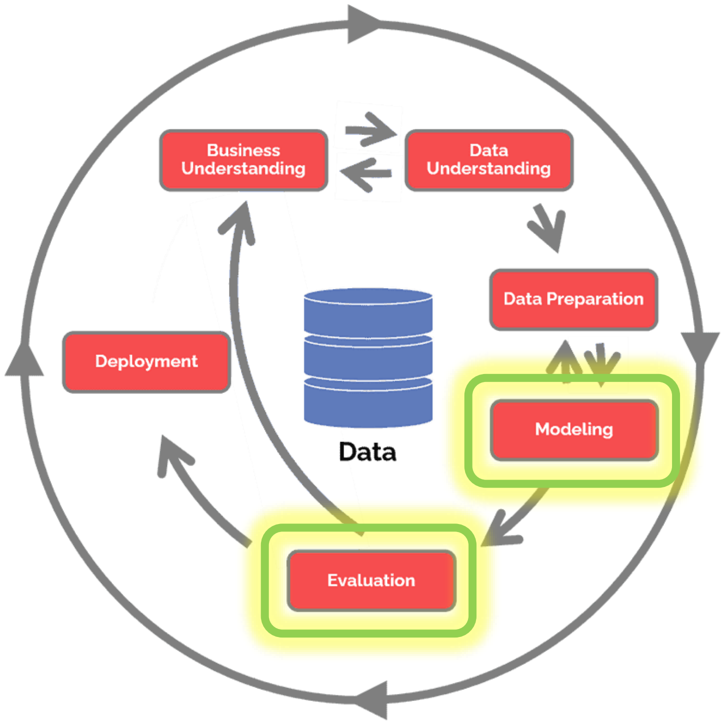
Data Preparation



- ❑ Data Aggregation and Group By
  - ❑ groupby(), agg()
- ❑ Data Cleaning
  - ❑ Handling missing values (isnull(), notnull(), fillna(), dropna())
  - ❑ Duplicate samples (duplicated(), drop_duplicated())
  - ❑ Data transformation (replace(), get_dummies())
- ❑ Normalization
  - ❑ min_max, z-score

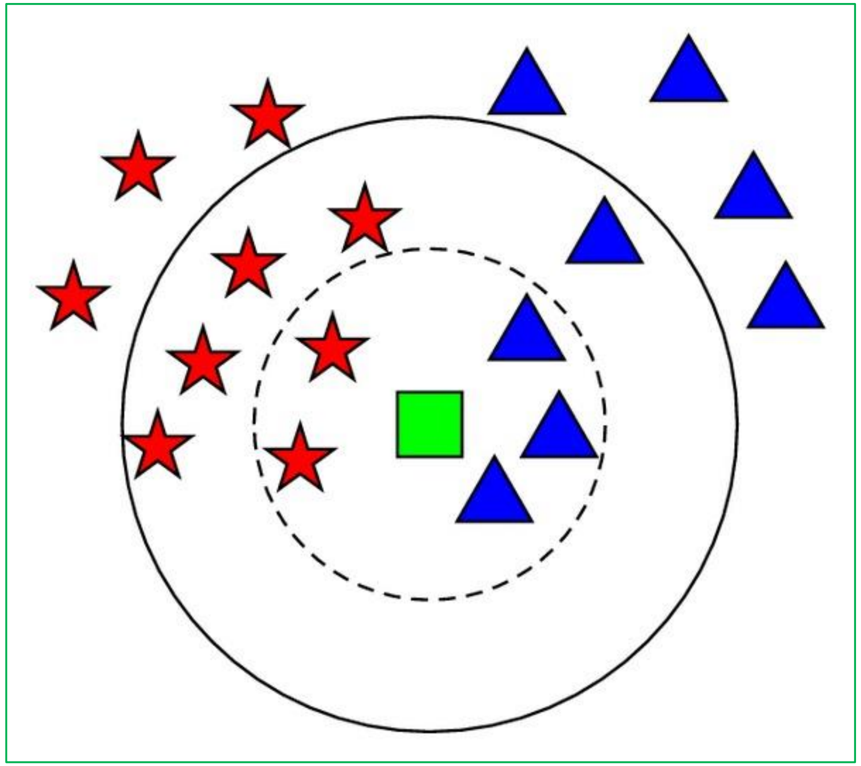
CRISP-DM
Modeling & Evaluation (Intro in TB1)

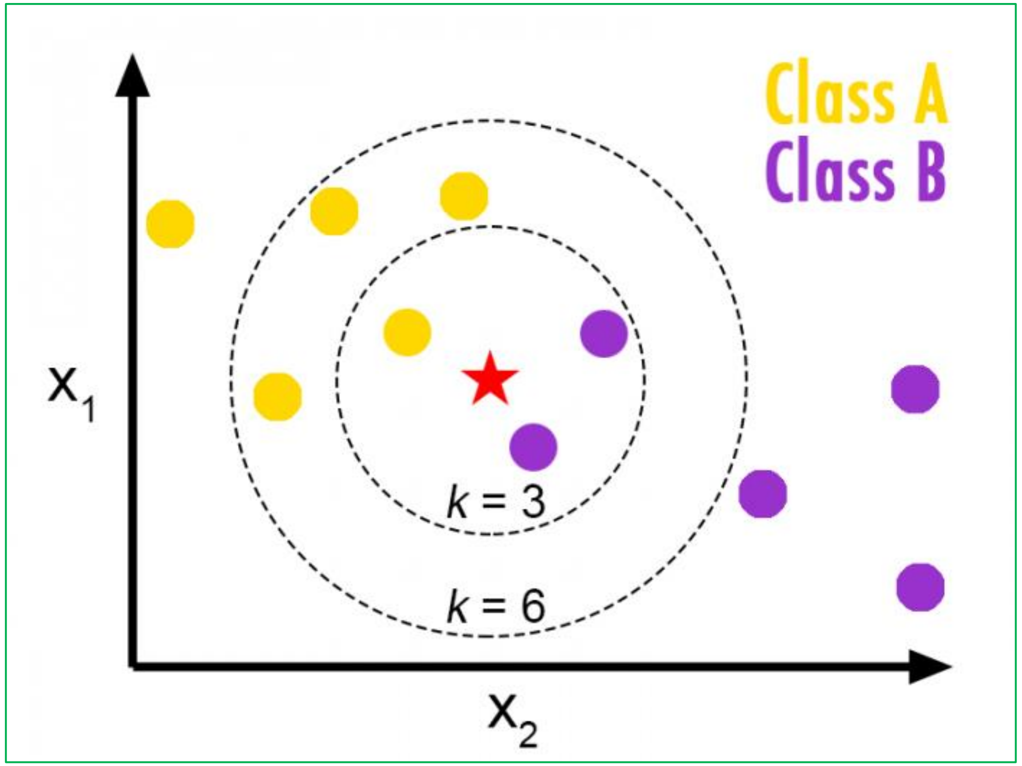




- □ Related functions:
  - □ train_test_split()
  - □ fit()
  - □ predict()

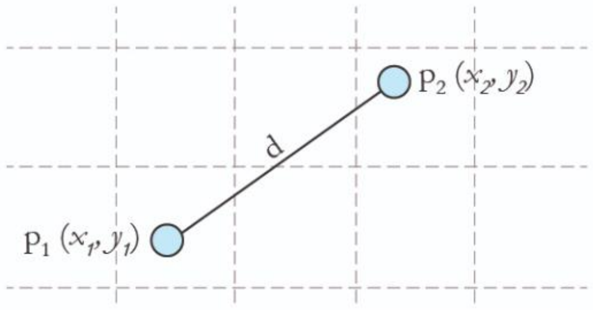
K-Nearest Neighbour (KNN)





K-Nearest Neighbour (KNN) (cont.)

- Among the simplest of all data mining algorithms
- Requires 3 things:
  - Feature space (training data)
  - Distance metric
    - Euclidean distance
  - The value of k
- Applicable to both classification and regression problems.



Euclidean distance (d) $= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

p, q = two points in Euclidean n-space

q_i, p_i = Euclidean vectors, starting from the origin of the space (initial point)

n = n-space

K-Nearest Neighbour (KNN) (cont.)

- Combining the labels of k nearest neighbours:
  - Take the majority vote (average) of labels among the neighbours
  - Weighting: $w = 1/d \quad or \quad 1/d^2$
- Example: Assume these are the three nearest neighbours

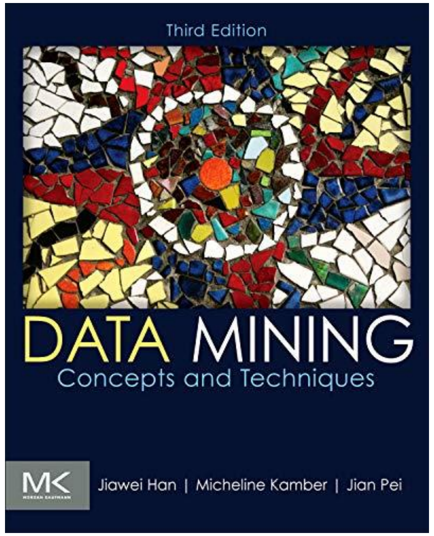
Value	5	8	9
Distance	3	2	5

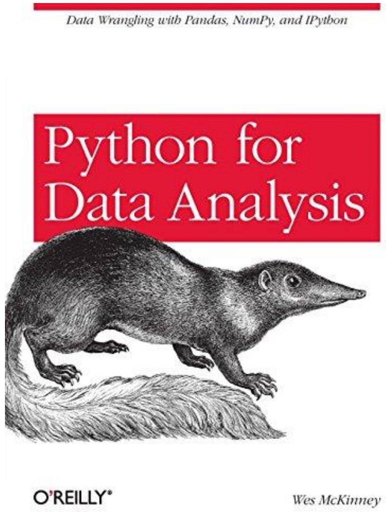
$$\text{prediction} = \frac{\frac{1}{3}\times 5 + \frac{1}{2}\times 8 + \frac{1}{5}\times 9}{\frac{1}{3} + \frac{1}{2} + \frac{1}{5}}$$

- Choosing the value of *k*:
  - If *k* is too small, sensitive to noisy points
  - If *k* is too large, neighbourhood may include irrelevant points
  - Choose an odd value for *k*, to eliminate ties

References & More Resources

- References:
  - McKinney, Wes. *Python for data analysis: Data wrangling with Pandas, NumPy, and Ipython*, O'Reilly Media, Inc., 2012.
  - Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.





Practical Session

- ❑ Please download TB2_Week01_Sample-Model.ipynb file, and run it to learn new points.
- ❑ Revise the Titanic Case Study (Last session of TB1) to remind the material of the previous TB. Build some KNN models for Titanic with several different numbers of neighbors and compare them together.