Python Fundamental

Dr. Steve Lai

2021/08

Outline for The Four Time Slots

1. Basic python
   - Python fundamental
   - Data visualization modules
   - Data analysis modules
   - Financial graphics
2. API & Machine learning
   - API demo
   - Machine learning introduction and processing
   - Sci-kit learn
3. Financial model Machine learning Project
   - Financial Models
   - Case Studies
     - House Prices: Advanced Regression Techniques
     - Home Credit Default Risk
4. Python crawling
   - beautifulshop
   - selenium
   - ptt crawling demo

Steve Lai（賴昭榮）

- Founder & CEO of Mathison Intelligence
- 台灣大學資訊工程學系博士

- 中央研究院資訊科技與創新中心助理
- iOS, Android, Python 講師

Quizs

- What is the maximum value of Int in python?
  1. 2^64 -1
  2. 2^32 -1
  3. Infinity

Quizs

- The BIF id() return the identity of an object which the identity is the address of the object in memory. Read the following program

```
1 a = 1
2 id1 = id(a)
3 a = 2
4 id2 = id(a)
5 print(id1 == id2)
```

- What is the output?
  1. An address number
  2. True
  3. true
  4. False
  5. false
  6. Cause an error

Quizs

- What is the output of the following program?

```
1 print = 3
2 print += 3
3 print(2)
```

1. 2
2. 3
3. 6
4. Cause an error

Basic Data Type in Python

- int
- float
- bool
- str

- How is the difficulty so far?
  - (easy) 0 ~ 5 (difficult)

Control Flow

- if-elif-else
- Loop
  - while loop
  - foreach loop

Basic Data Structure

- tuple
- list
- dict
- set

Programming Structure

- Function
- Class

Quizs

- What is benefit of using function?
  - Easy to read
  - Easy to debug
  - Easy to use

Quizs

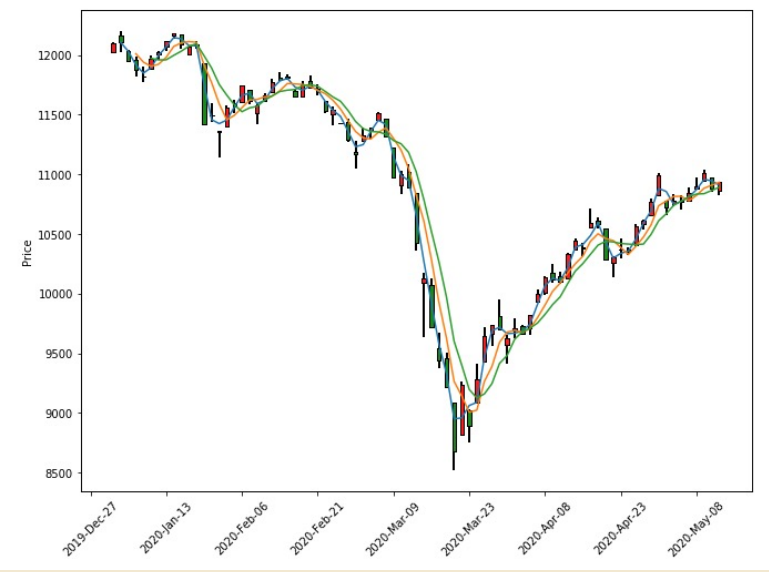
- Which of the following are synonyms?
  1. instance
  2. attributes
  3. method
  4. object
  5. function
  6. variable
  7. class

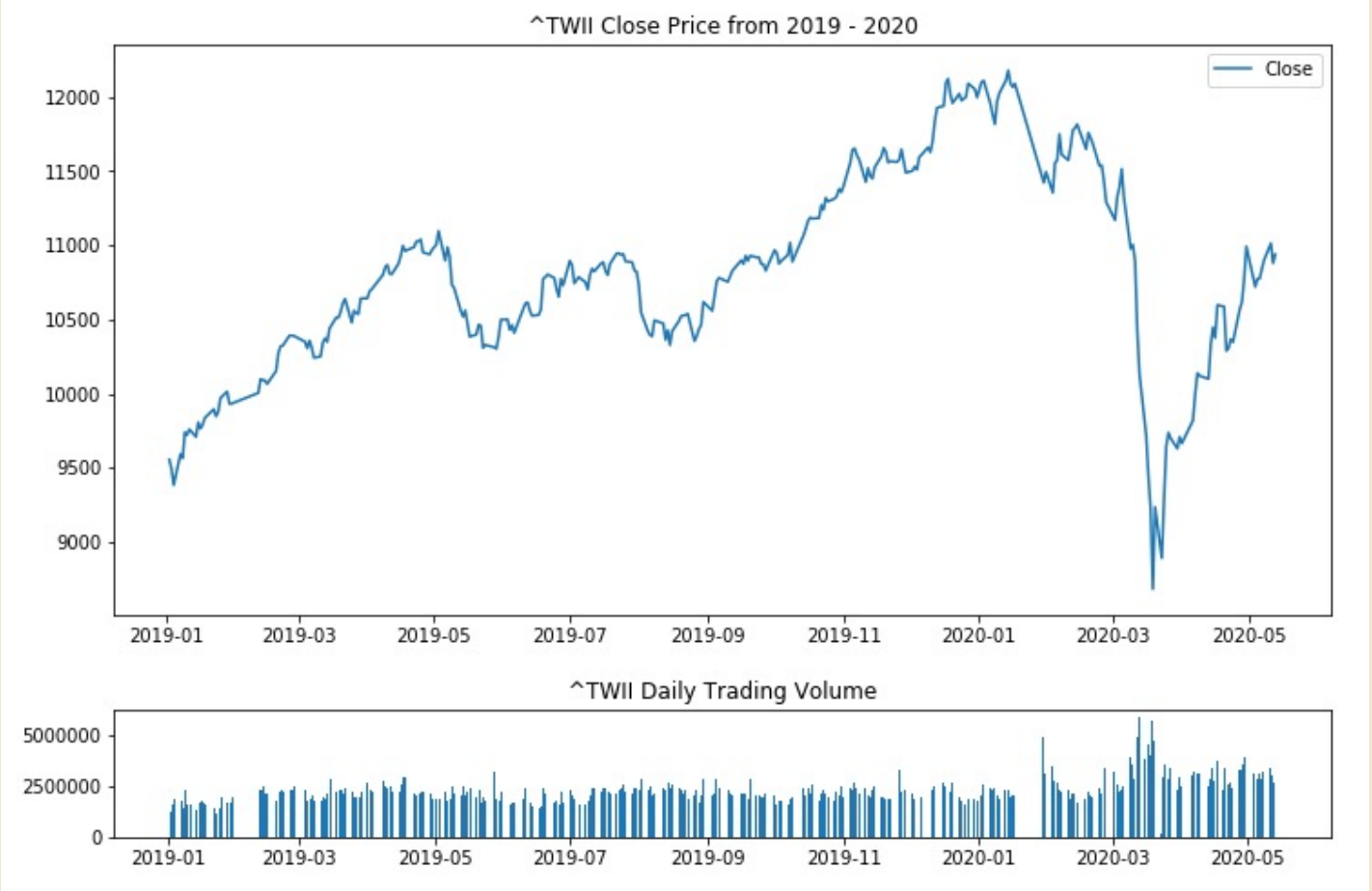
Data Analysis Modules

- Numpy
- Pandas
- Exploratory data analysis

Sample Data Source

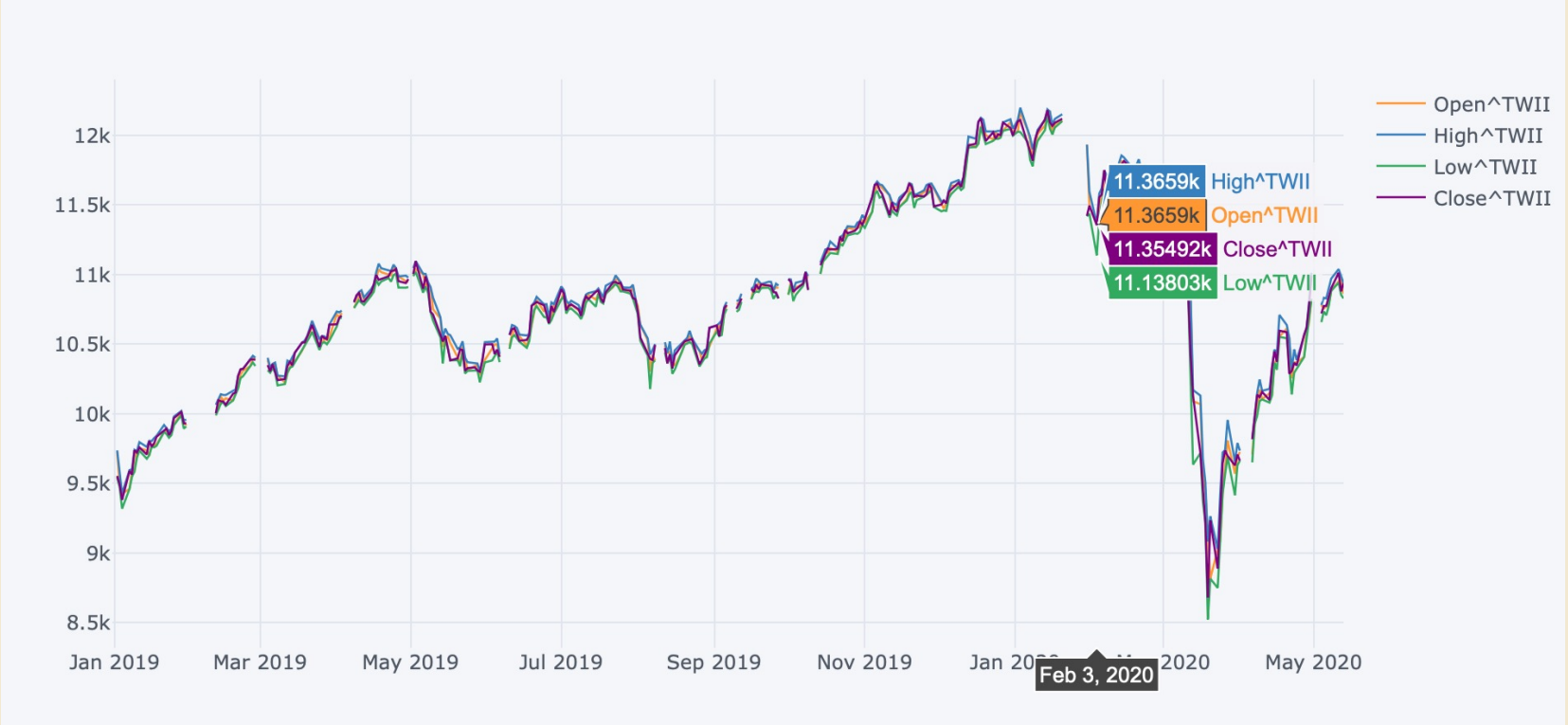
- TWSE
- TAIFEX
- Quandl
  - Need sign up an account
- pandas_datareader
  - yfinance





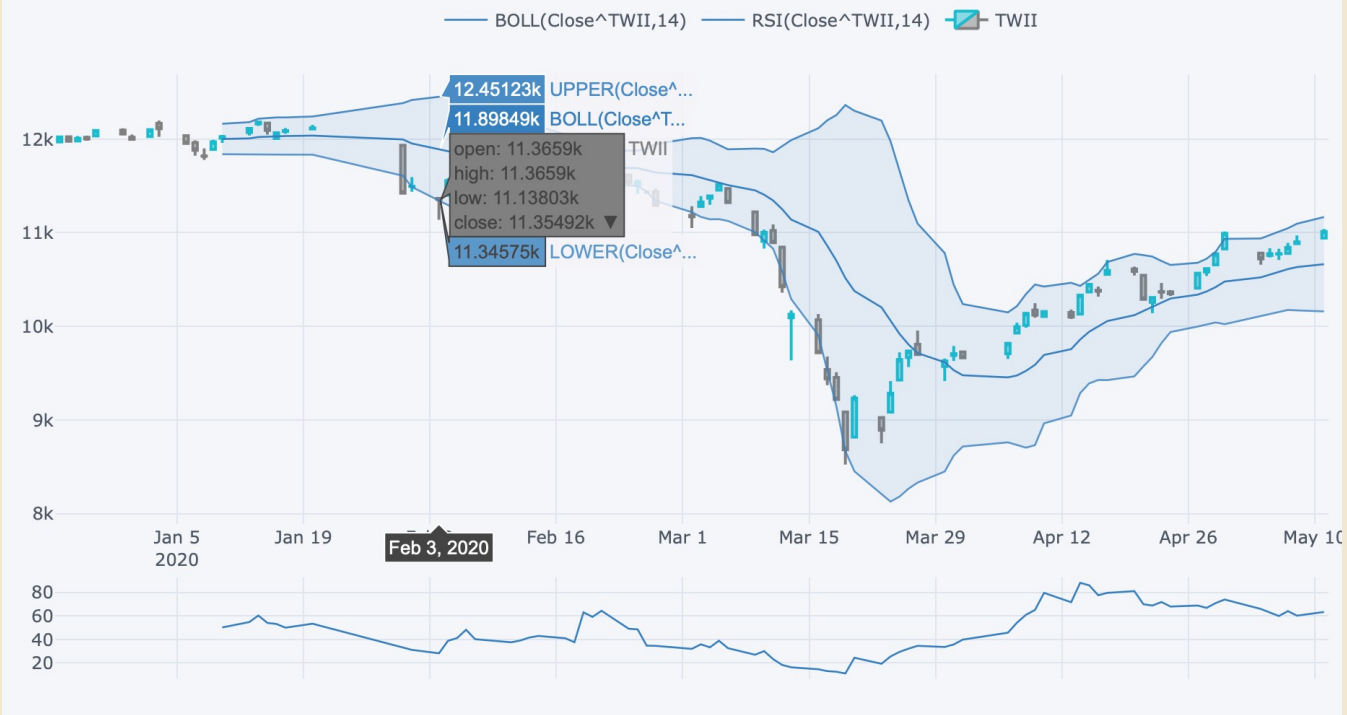
Data Visualization

- Requirement packages
  - Matplotlib
  - Seaborn
  - Plotly
- Graphic Types
  - plot
  - bar
  - histogram
  - scatter
  - pie
  - box
  - polygon



Financial Graphics

- Requirement packages
  - Cufflinks
  - chart_studio
  - mplfinance
- Graphics Types
  - Candlestick chart
  - Bollinger bands
  - Relative strength indicator
  - Moving average convergence divergence



Financial Time Series

- Summary Statistics
  - df.describe()
    - mean, min, max, std, quartile
- Time difference
  - df.diff()
  - df.pct_change()
- Time resample
- Long/Short simple moving averages
  - df.rolling(window=5, min_periods=5).mean()
  - df.rolling(window=30, min_periods=30).mean()
- Exponential moving averages
  - df.ewm(span=5).mean()
  - df.ewm(span=30).mean()
- Correlation Analysis
  - df.pct_change().corr()

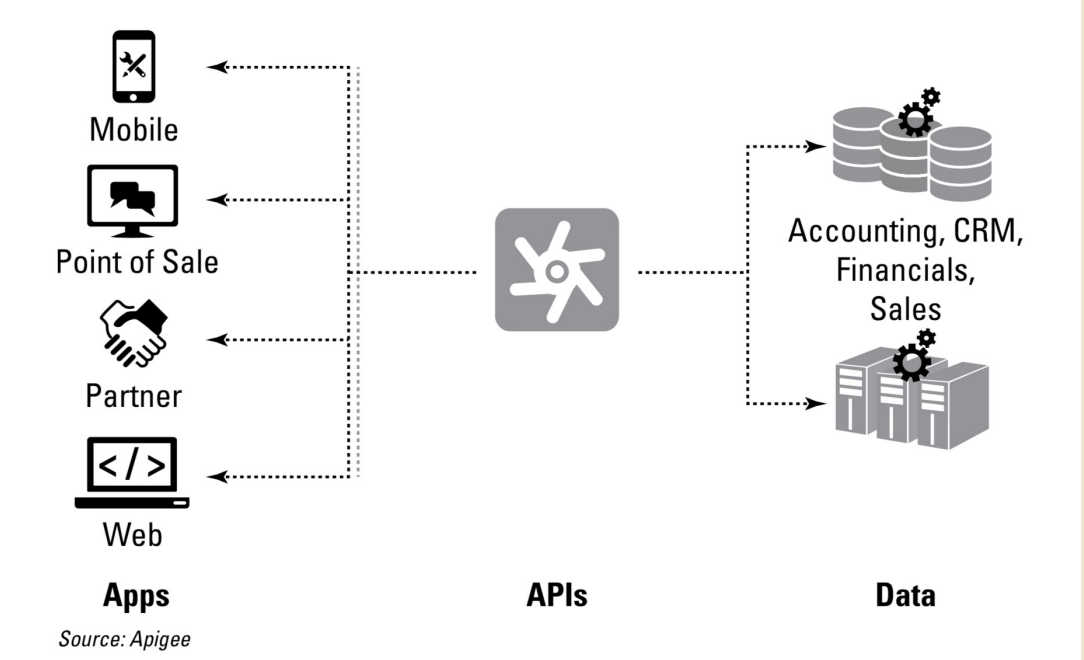
Python with API

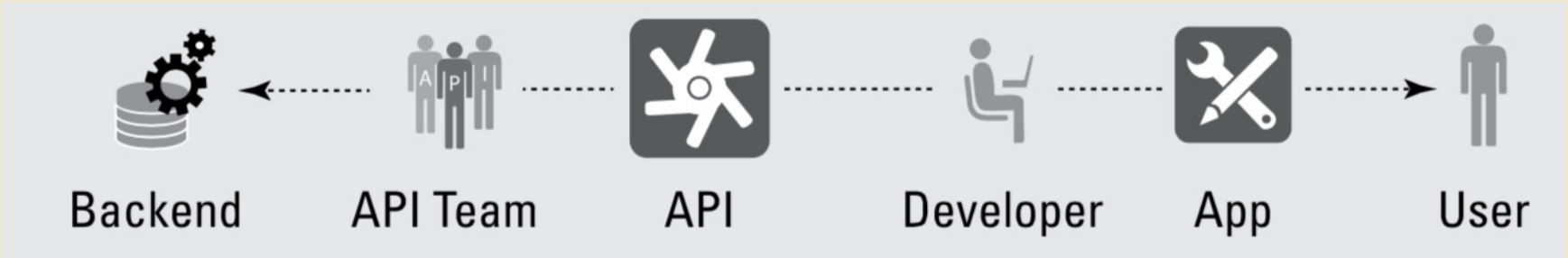
Dr. Steve Lai

2021/08

Connecting to The Appliccation

- API
  - Application Programming Interface
  - Libraries, CSV, JSON, XML, …





JSON

- What is the following similar with JSON Object in Java?
  1. Dictionary
  2. Set
  3. Tuple
  4. List

- What is the following similar with JSON Array in Java?
  1. Array
  2. Set
  3. Map
  4. StringBuilder

What is JSON?

- JSON
  - JavaScript Object Notation
- JSON Object
  - Key-Value pairs
- JSON Array
  - Sequential data with order
- import json
  - load, loads
  - dump, dumps

CSV

- import csv

APIs

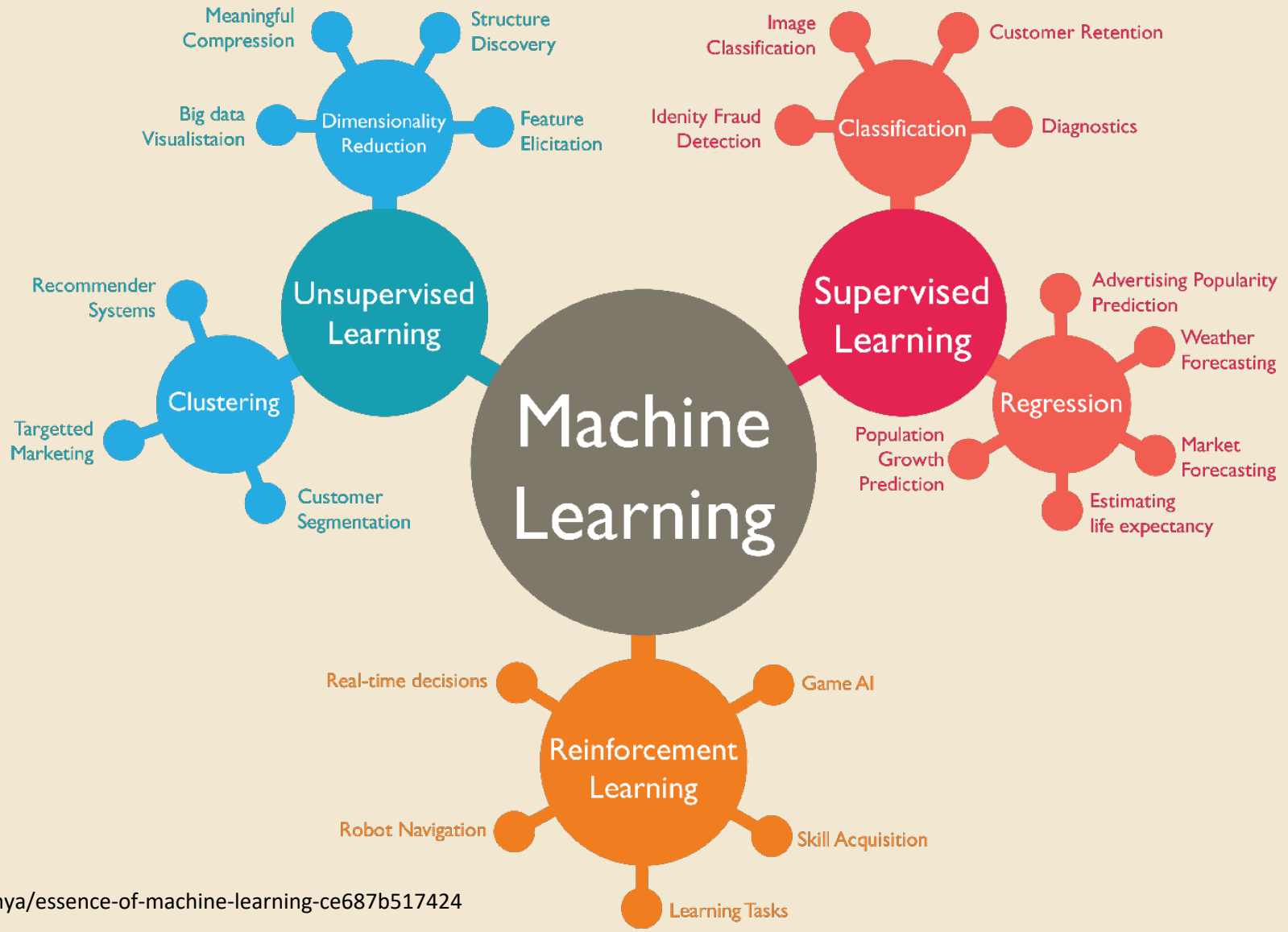
- https://api.github.com/search/repositories?q=language:python&sort=stars
- https://www.travel.taipei/open-api/zh-tw/Attractions/All?categoryIds=12&page=1

Machine Learning

Dr. Steve Lai

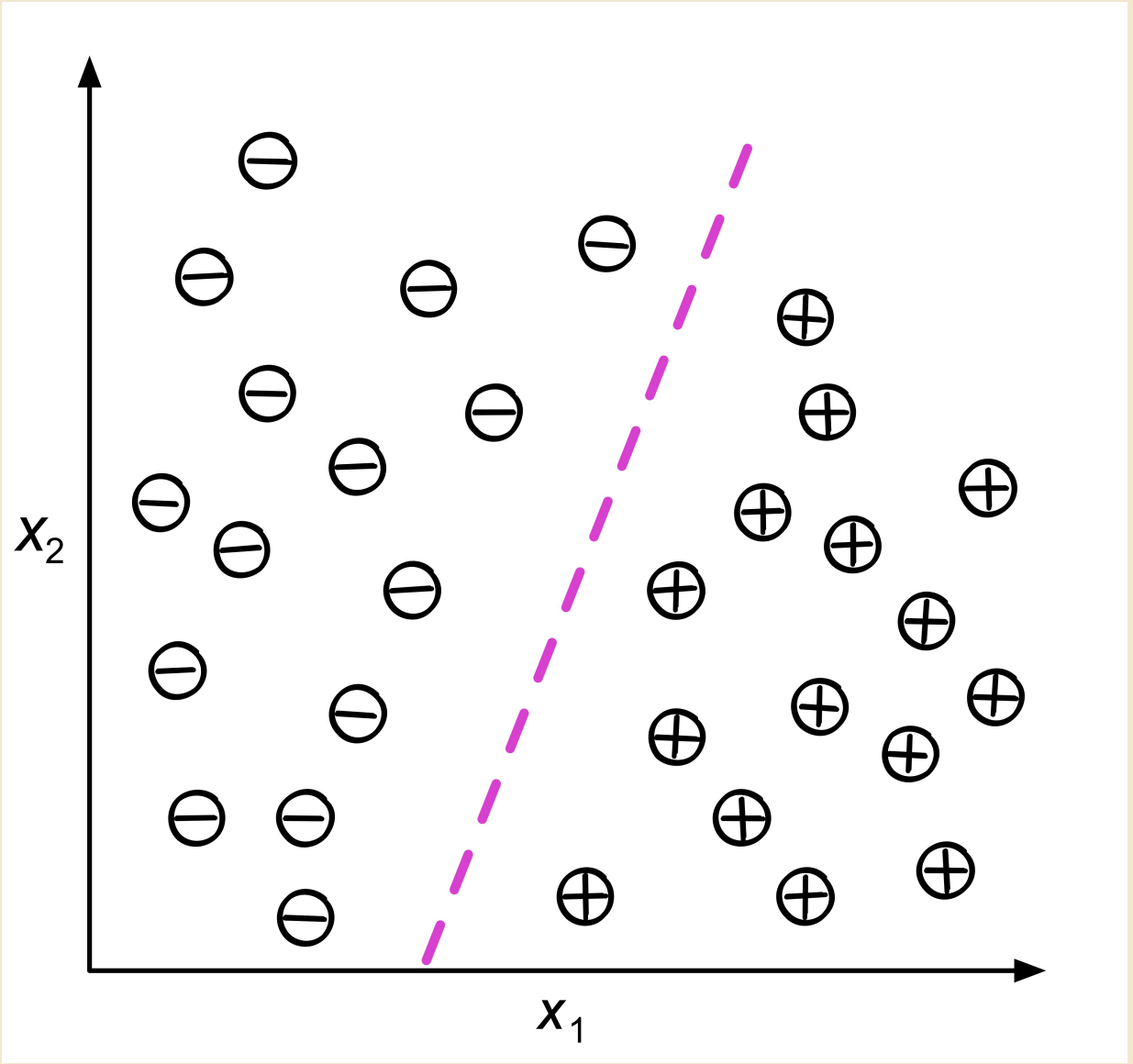
2021/08

Machine Learning

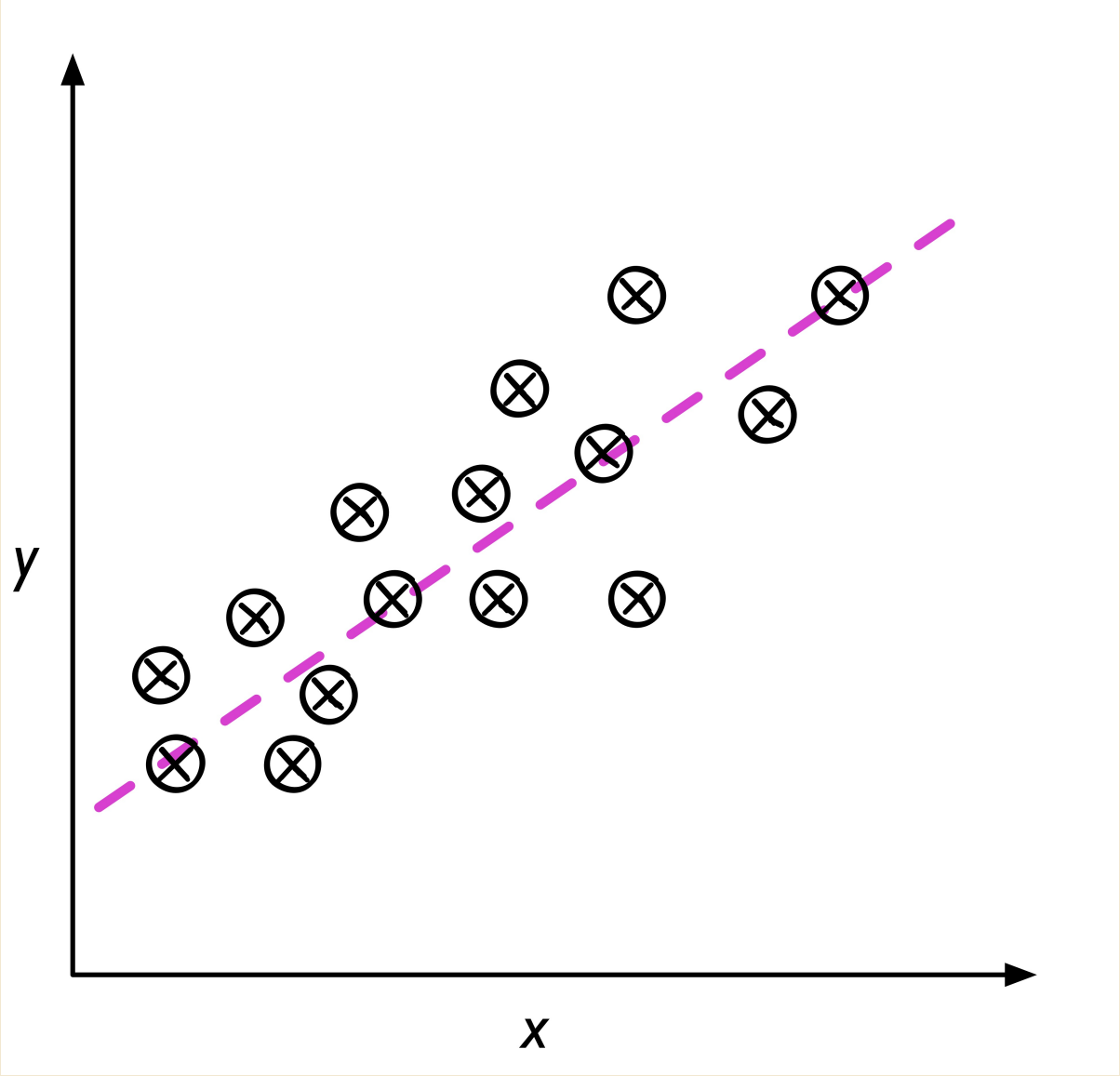


https://medium.com/analytics-vidhya/essence-of-machine-learning-ce687b517424

Classification

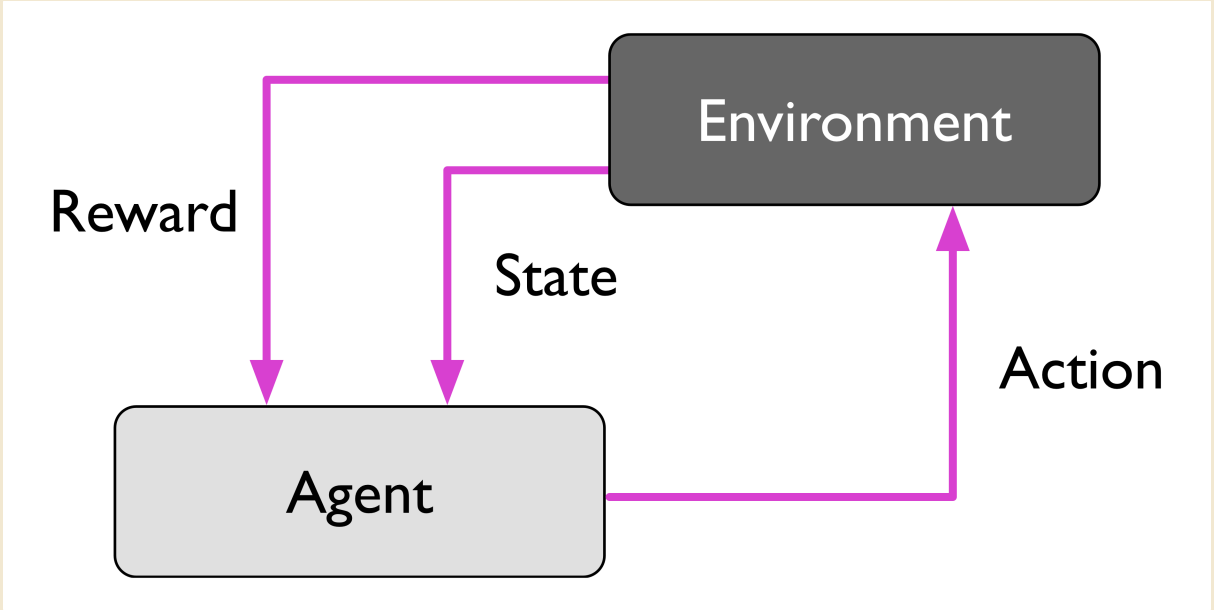


Regression



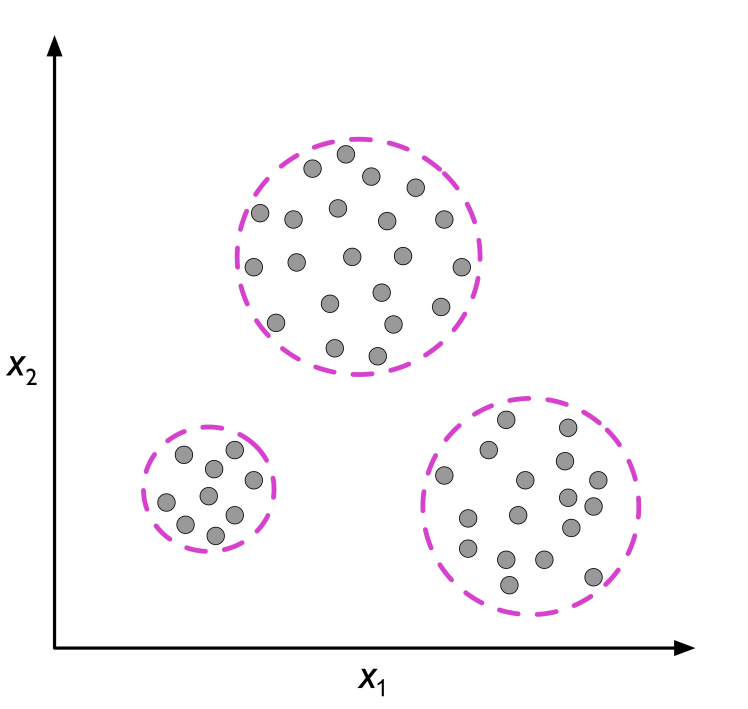
Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2, 3rd Edition

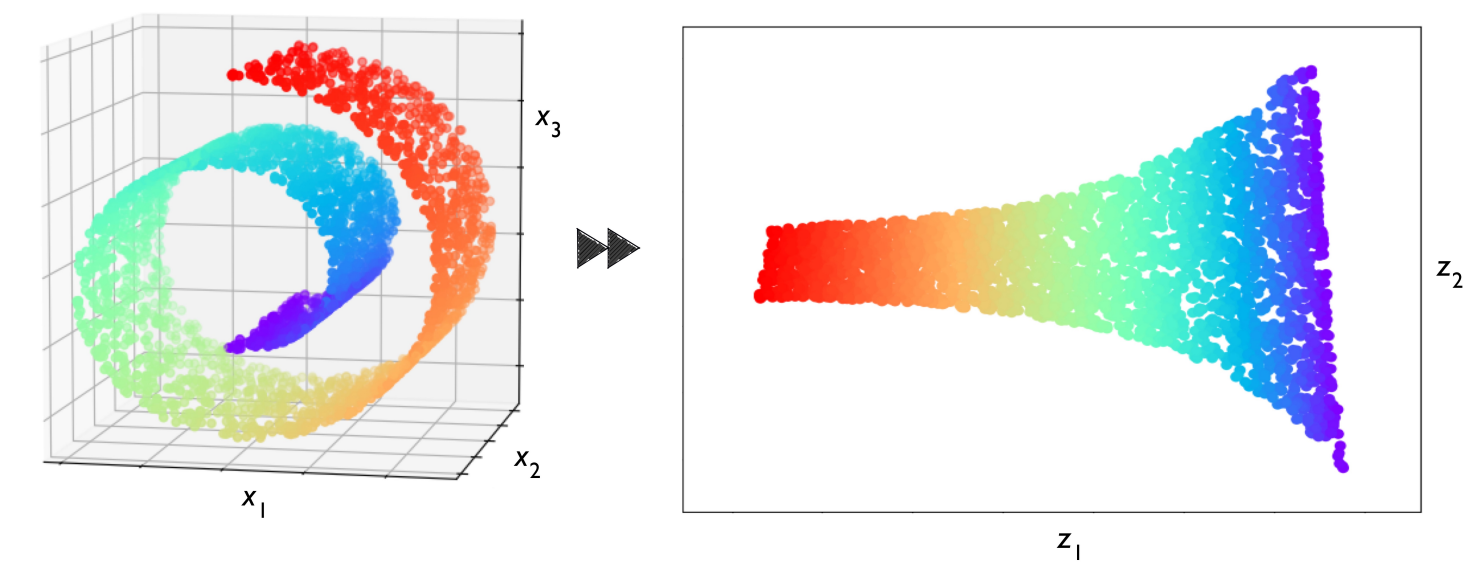
Reinforcement Learning



Unsupervised Learning

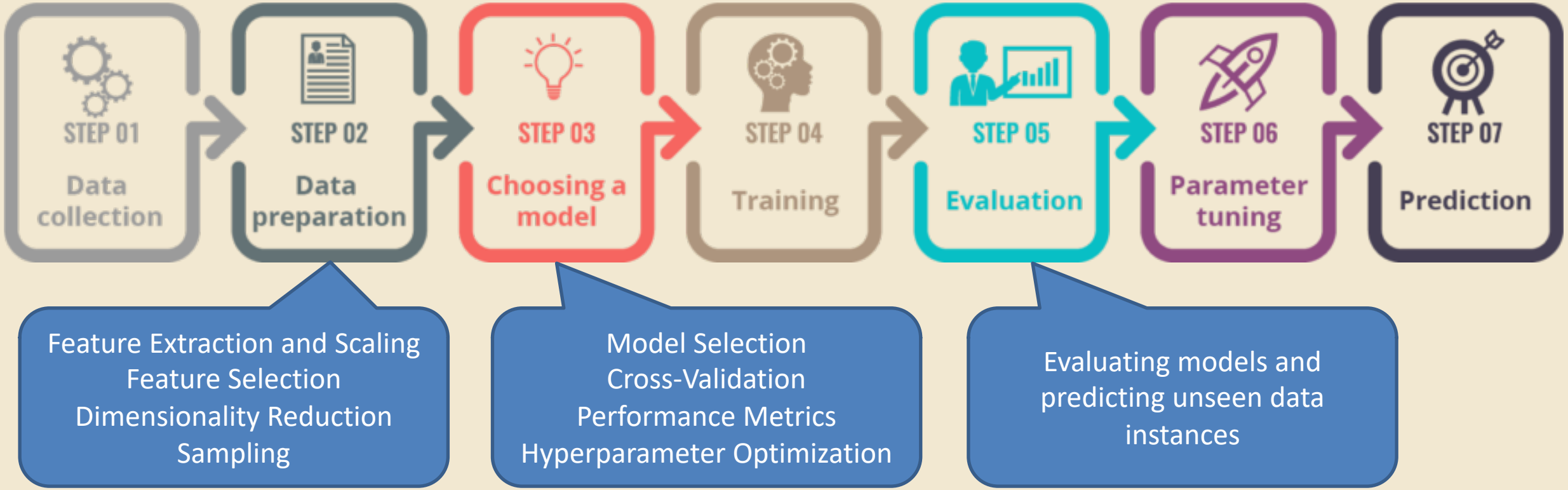
- Finding subgroups
  - Clustering
- Dimensionality reduction





Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2, 3rd Edition

7 Steps to Machine Learning

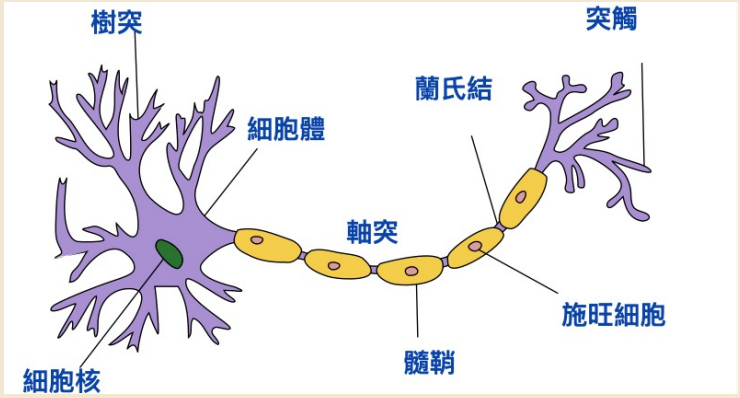


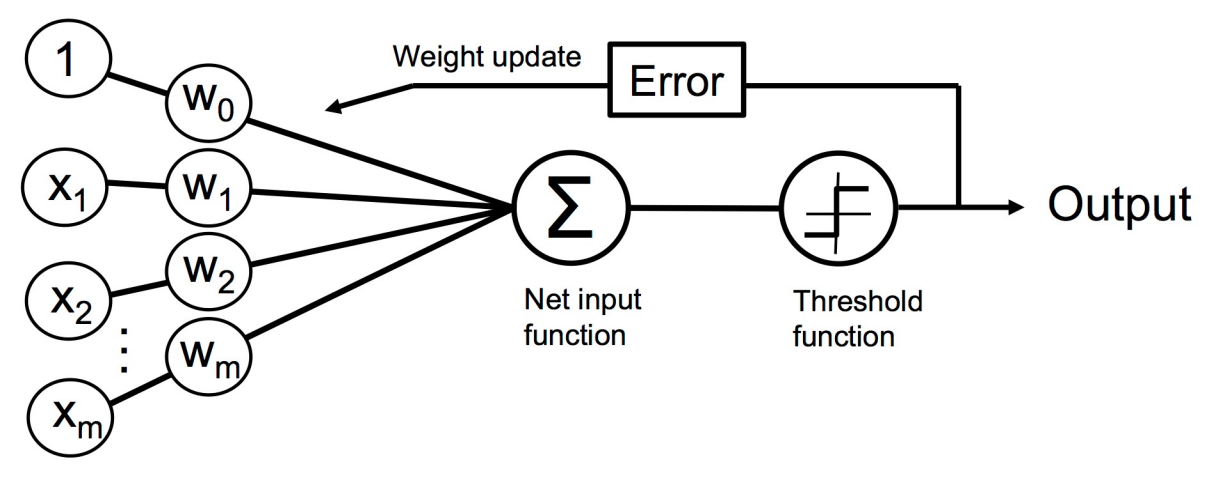
Google Cloud Platform
https://www.youtube.com/watch?v=nKW8Ndu7Mjw
The 7 steps of machine learning (AI Adventures)

https://medium.com/dataseries/7-steps-to-machine-learning-how-to-prepare-for-an-automated-future-78c7918cb35d

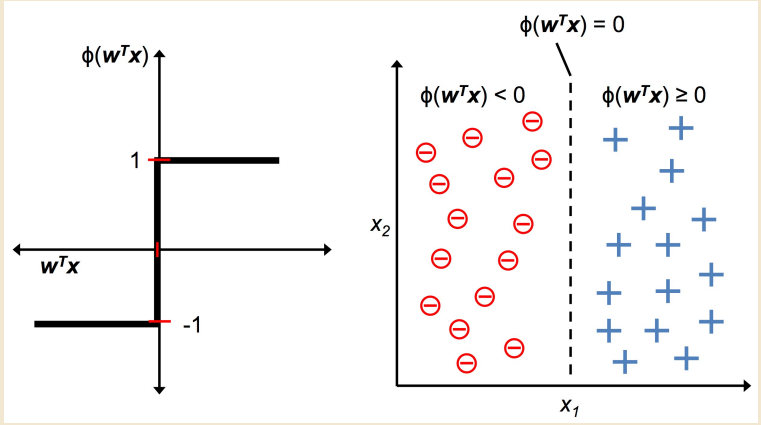
Learning from Data

- Biological neurons



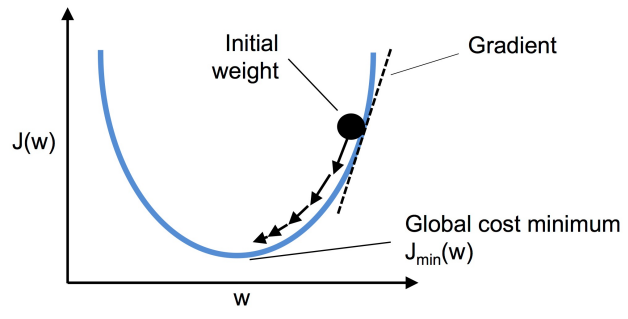


- Artificial neurons: Perceptron
  - $w = \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix}, x = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}, z = w_1x_1 + w_2x_2 + \cdots + w_mx_m, \phi(z) = \begin{cases} 1, & if\ z \geq \theta, \\ -1, & otherwise \end{cases}$
  - Update the weights simultaneoous
    - $w_j := w_j + \Delta w_j$
    - $\Delta w_j = \eta(y^{(i)} - \hat{y}^{(i)})x_j^{(i)}$
  - w_j: the weights of feature j
  - $x_j^{(i)}$: the j_{th} feature of data i
  - $y^{(i)}$: the target of data i

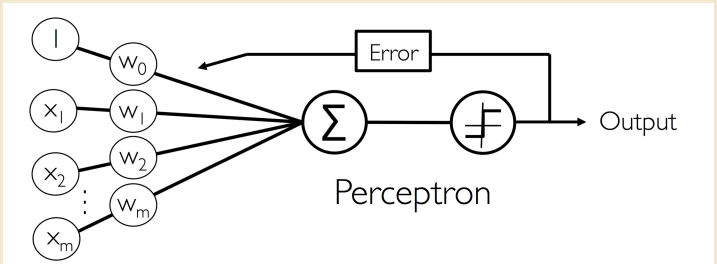


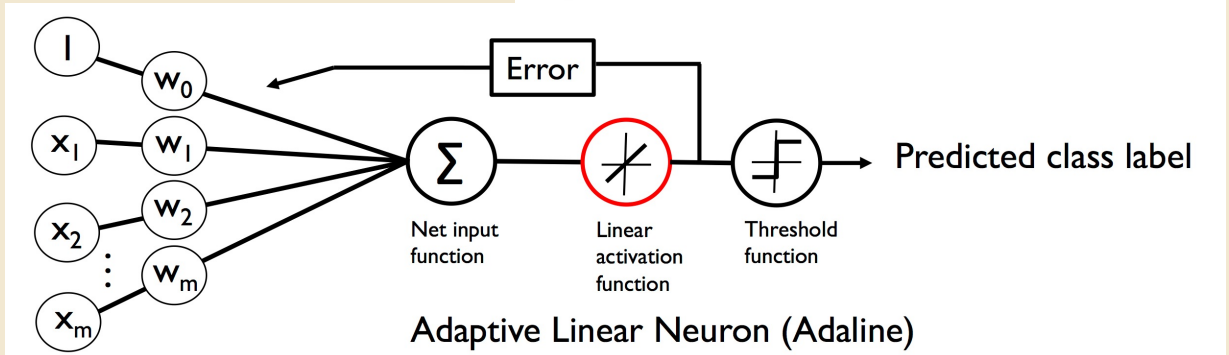
Different Activation Function

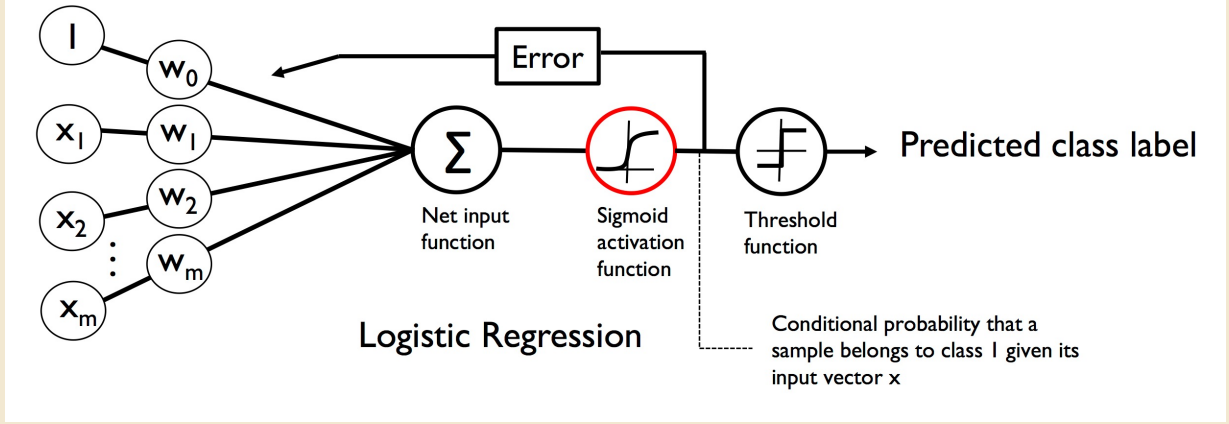
- ADAptive LInear NEuron (Adaline)
  - Perceptron may divergence
  - Add a linear function after net input
  - Gradient descent to get the optimal



- Logistic regression
  - Conditional probabilities
  - Sigmoid function



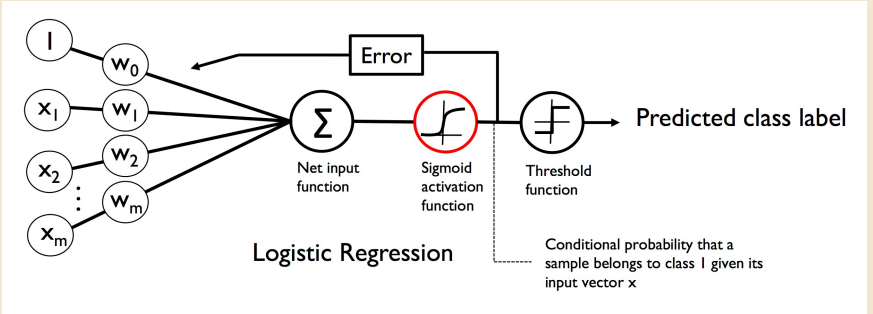


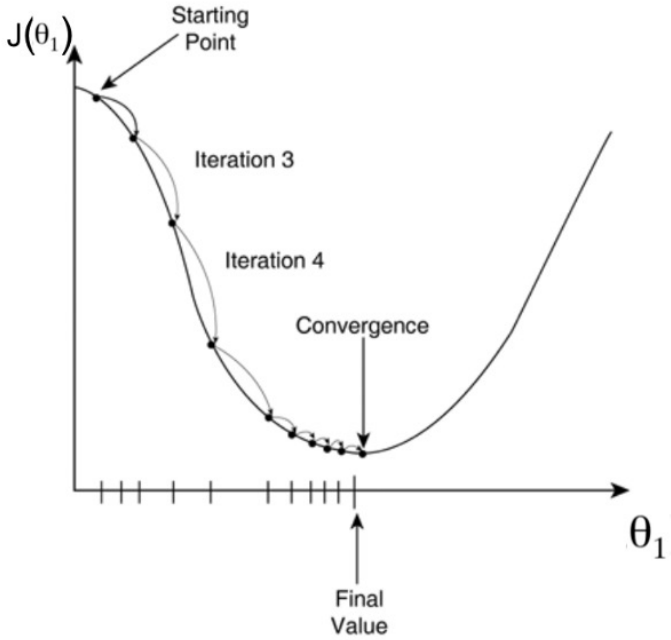


Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2, 3rd Edition

Logistic Regression

- Linear classification model
- Logistic sigmoid function
  - $\phi(z) = \frac{1}{1+e^{-z}}, z = \boldsymbol{w}^T\boldsymbol{x} = w_0x_0 + w_1x_1 + w_2x_2 + \cdots + w_mx_m$
- Loss (Cost) function
  - $\hat{y} = \begin{cases} 1, & if\,\phi(z) \geq 0.5 \\ 0, & otherwise \end{cases}$
  - $J(\boldsymbol{w}) = \sum_i \frac{1}{2}\left(\phi\left(z^{(i)}\right) - y^{(i)}\right)^2$





Cost Function – "One Half Mean Squared Error":

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

Objective:

$$\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$$

Derivatives:

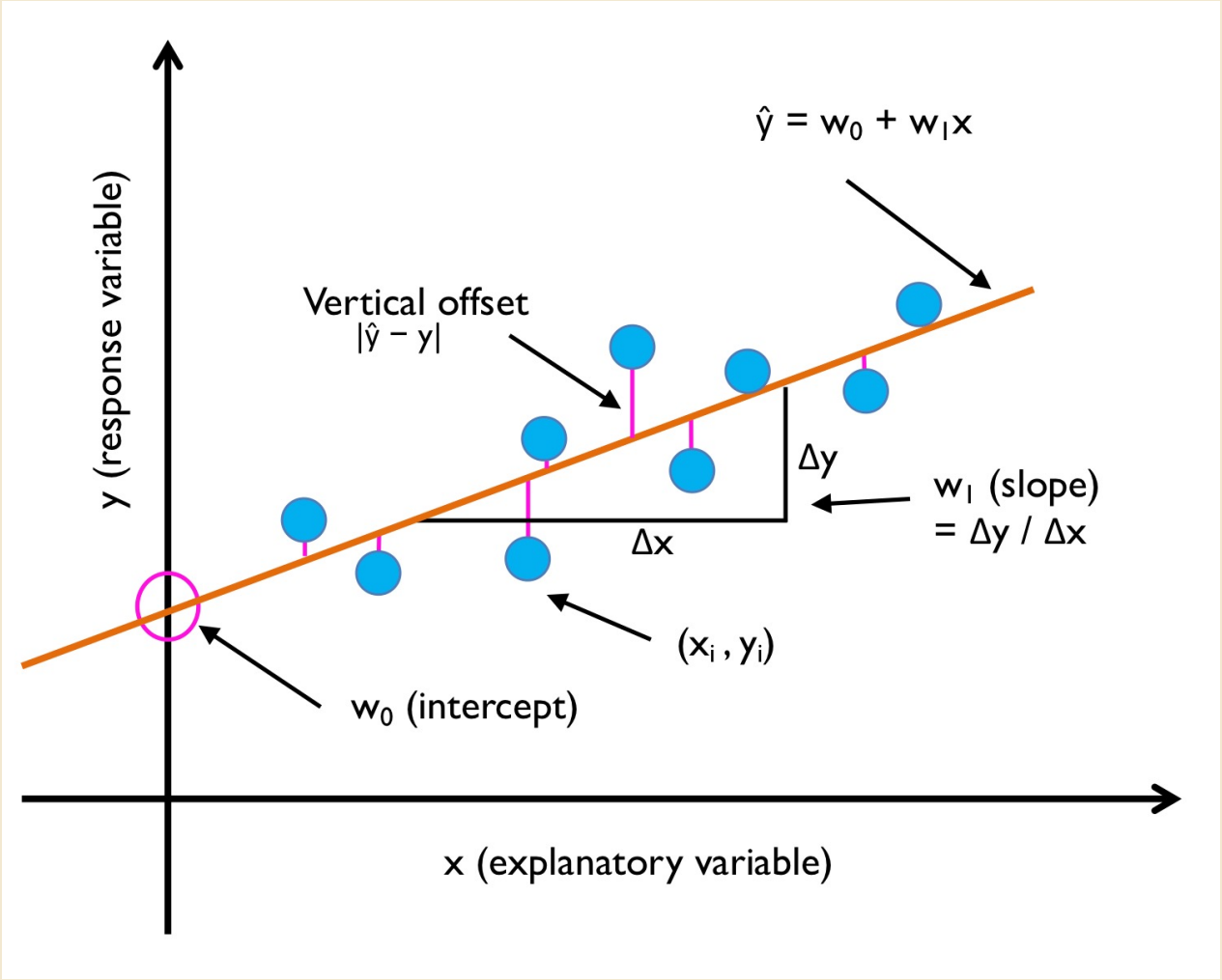
$$\frac{\partial}{\partial\theta_0}J(\theta_0, \theta_1) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)$$

$$\frac{\partial}{\partial\theta_1}J(\theta_0, \theta_1) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)\cdot x^{(i)}$$

- Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2, 3rd Edition
- https://www.kdnuggets.com/2020/05/5-concepts-gradient-descent-cost-function.html

Linear Regression

- Predict the continues targets

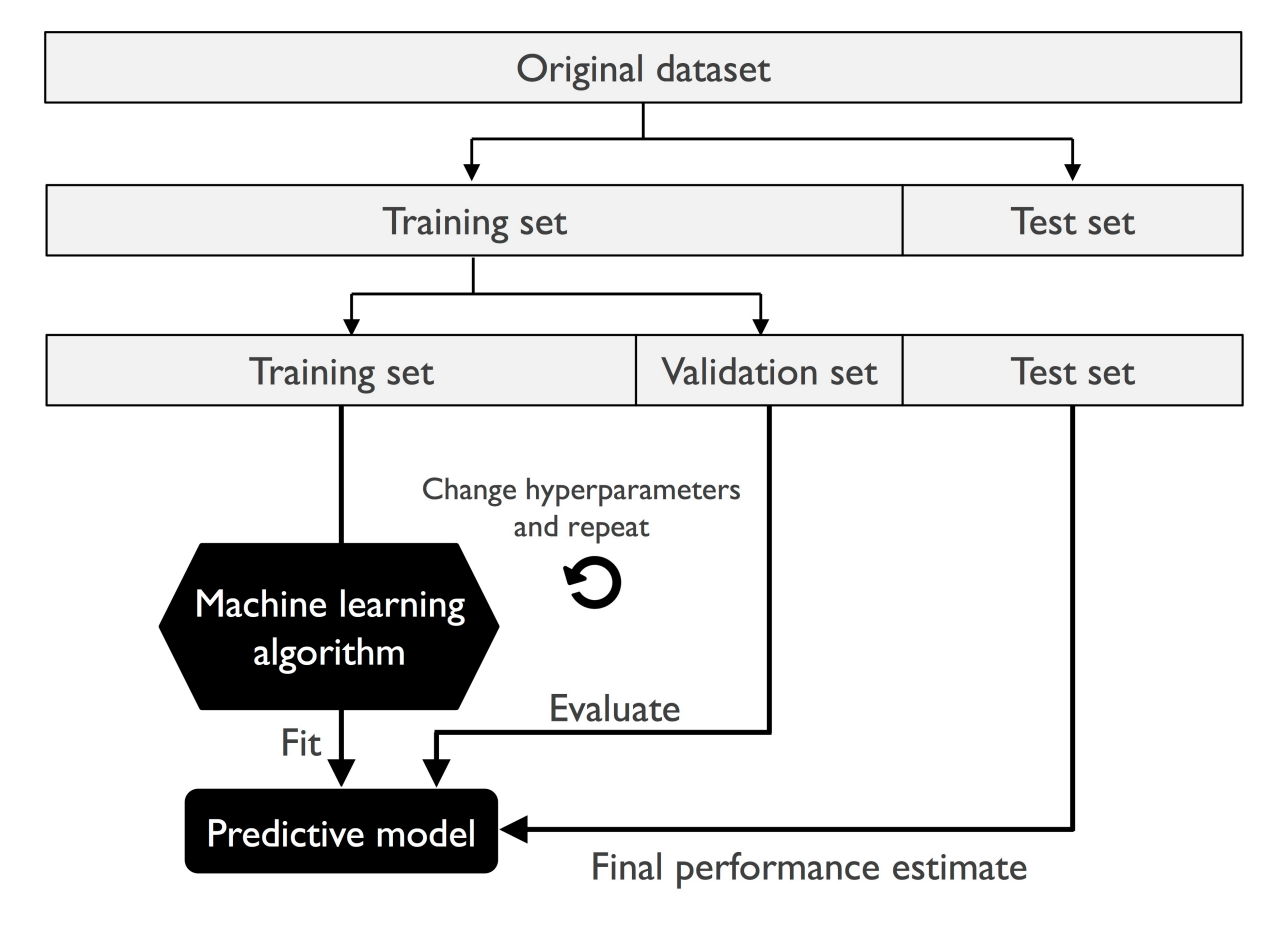


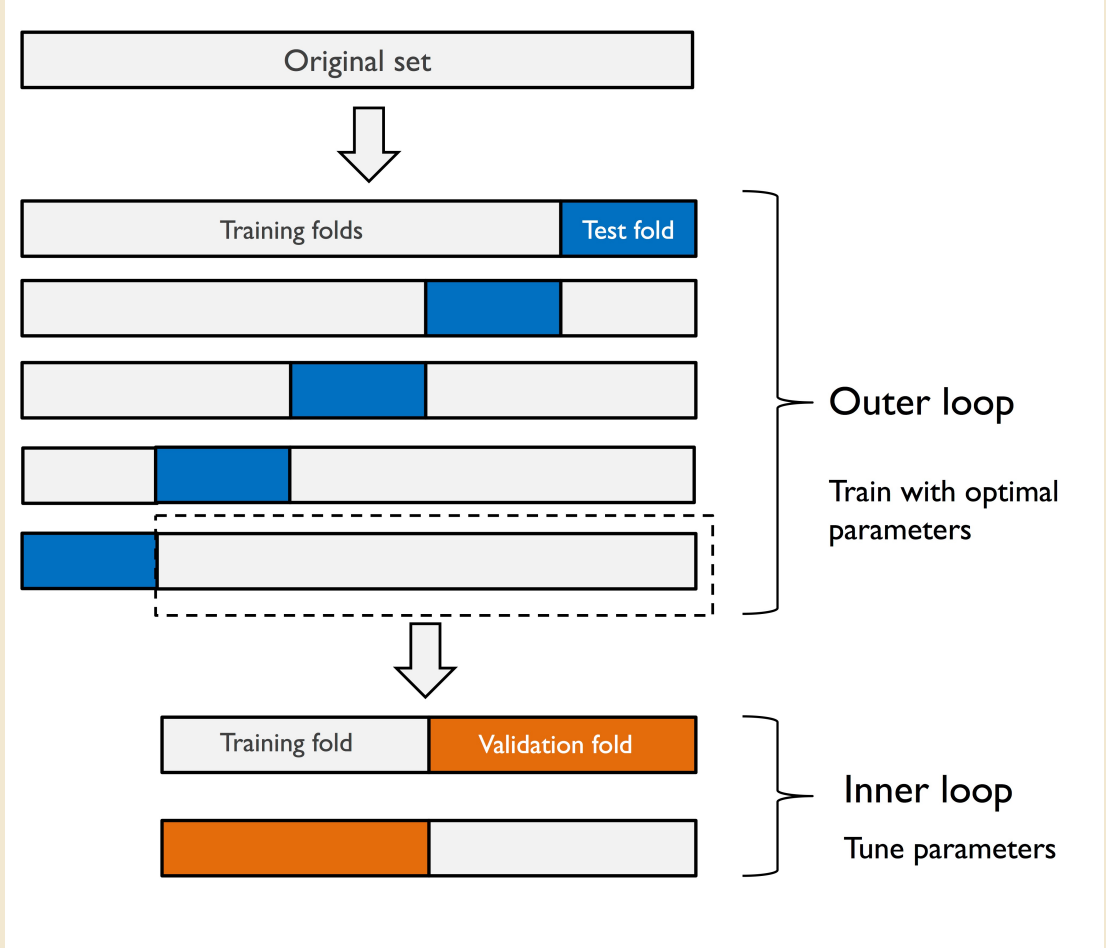
Data Preprocessing

- Dealing with missing data
  - Identifying
  - Eliminating, examples or features
  - Imputing
- Handling categorical data
  - Ordinal, label encoding
  - Nominal, one-hot encoding
- Separate training and test datasets
- Scaling
  - Min-max scaling
  - Standard scaling

Model Selection and Hyperparameter Tuning

- Cross-validation





Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2, 3rd Edition

Evaluation Metrics – Confusion Matrix

		Predicted condition		Sources: [13][14][15][16][17][18][19][20] view · talk · edit	
	Total population $= P + N$	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP),** hit	**False negative (FN),** type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP),** type I error, false alarm, overestimation	**True negative (TN),** correct rejection	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN} = 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI) $= \frac{TP}{TP + FN + FP}$

https://en.wikipedia.org/wiki/Confusion_matrix

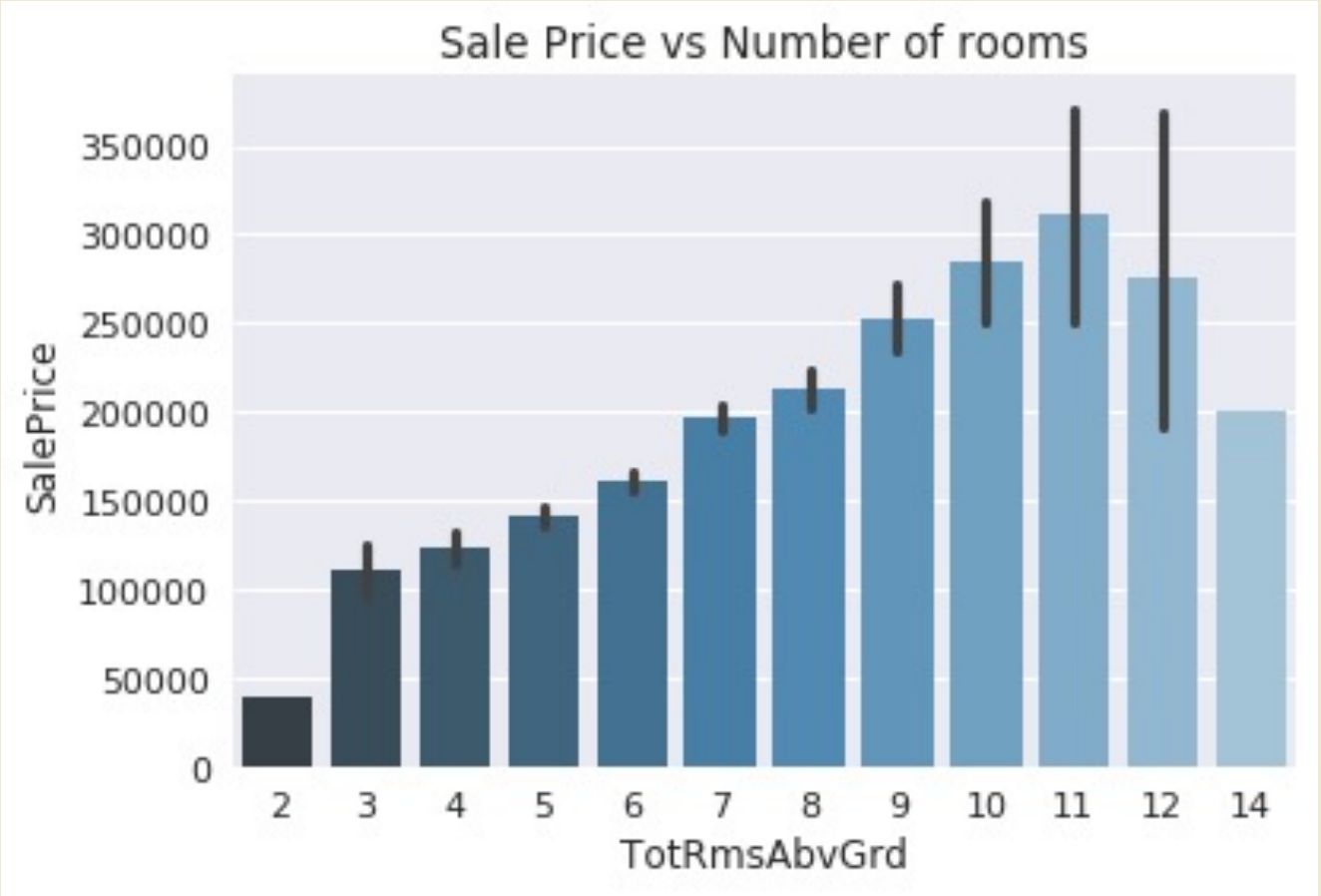
Case Study: Minimizing Risks for Loan Investments

- Higher credit scores (more trustworthy and less risky) get lower interest rates for their loans
  - lower credit scores (less trustworthy and more risky) get higher rates.
- However, the loans with higher interest rates are more attractive because they provide higher return on investment (ROI)
  - They pose risks of being not returned at all.
- The machine learning model that could predict which of the high interest loans are more likely to be returned, would bring added value by minimizing the associated risks.

- https://www.kaggle.com/wordsforthewise/lending-club

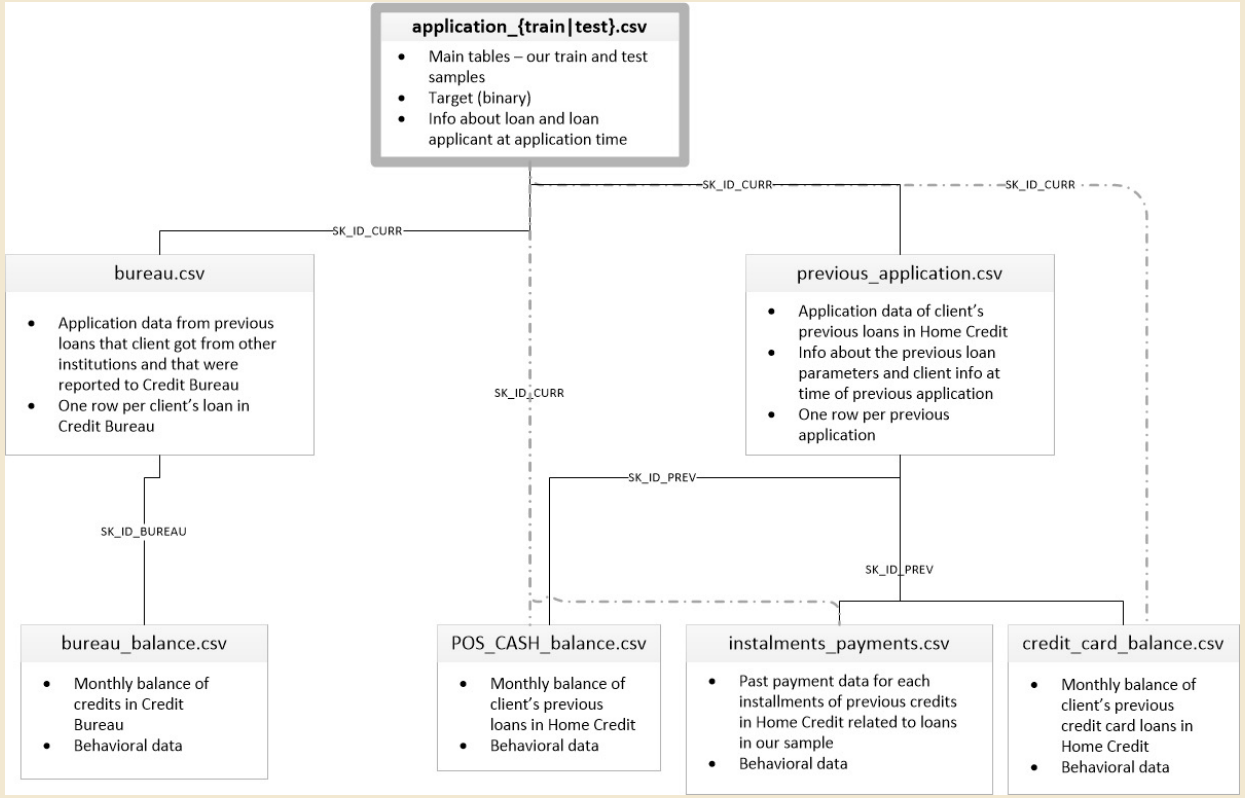
Case Study: House Prices

- Kaggle
  - House Prices: Advanced Regression Techniques
  - https://www.kaggle.com/c/house-prices-advanced-regression-techniques



Case Study: Home Credit Default Risk

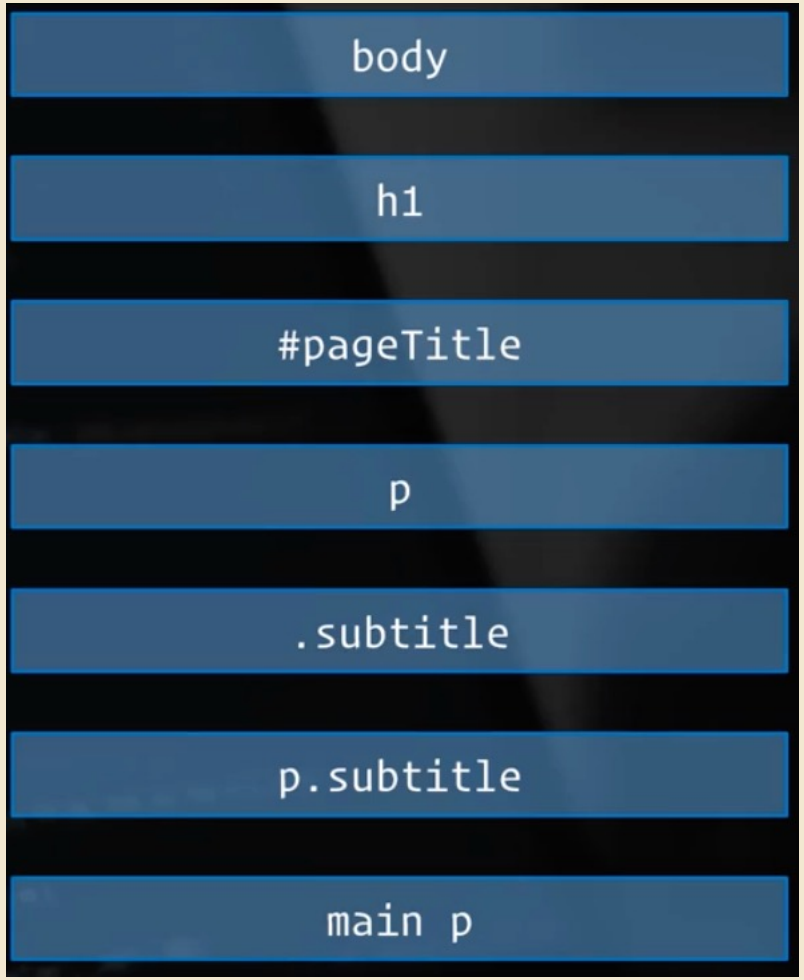
- Kaggle
  - Home Credit Default Risk
  - https://www.kaggle.com/c/home-credit-default-risk



The Structure of HTML Code

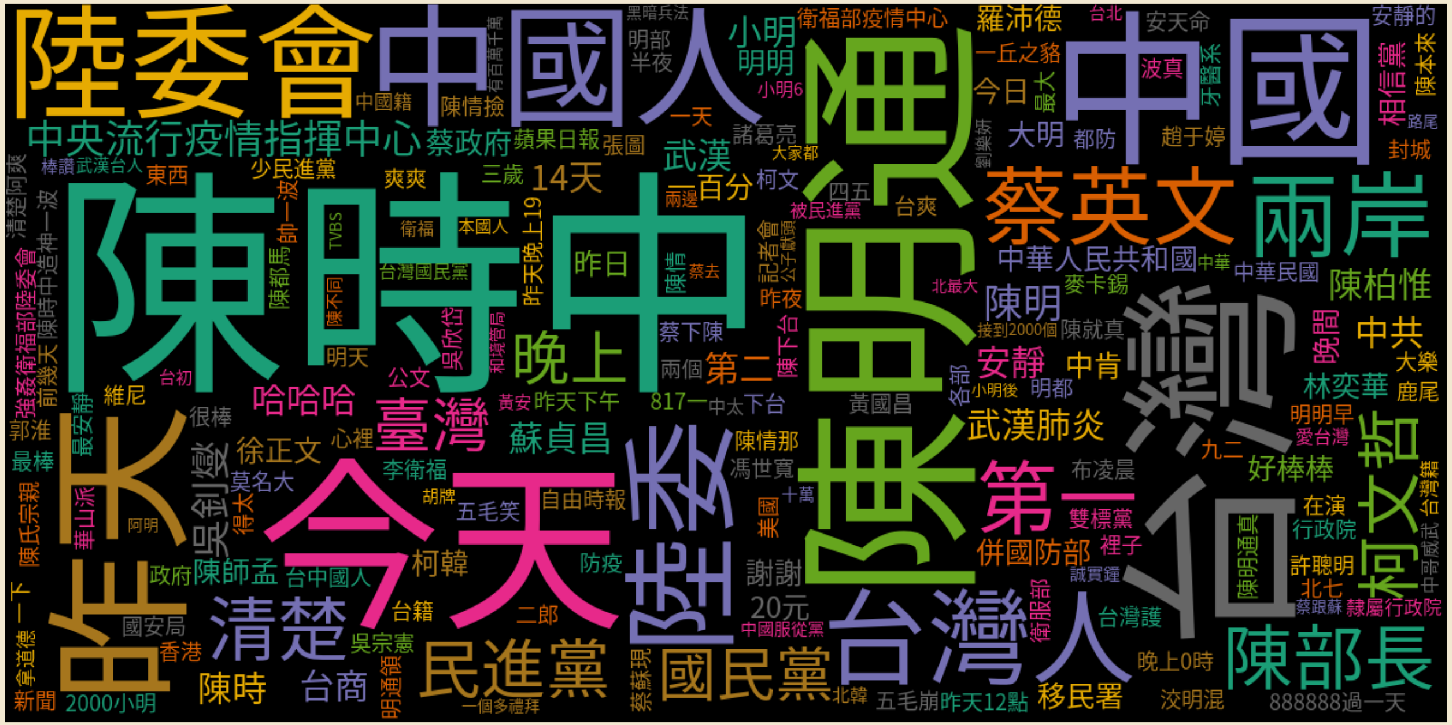
-

```
<body>
    <h1 class="header" id="pageTitle">Hello World!</h1>
    <p class="subtitle">A subtitle.</p>
    <main>
        <p>This is the main text.</p>
    </main>
</body>
```



PTT Demo

- PTT crawling
- NLP



Code Reference

- https://github.com/lzrong0203/fin_ios_python/blob/master/Python20210823.pdf
- https://github.com/lzrong0203/fin_ios_python