

# 資料科學 – Pandas教學

---

# 自我介紹

---

- 蔡岳霖
- 學歷
  - 成功大學心理學系(2009-2013)
  - 成功大學心理學研究所(2013-2015)
- 現職
  - 台灣人工智慧學校 AI工程師 / 專任助教

# 資料科學是？

## DATA SCIENTIST



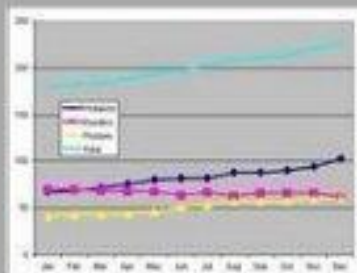
What my friends think I do



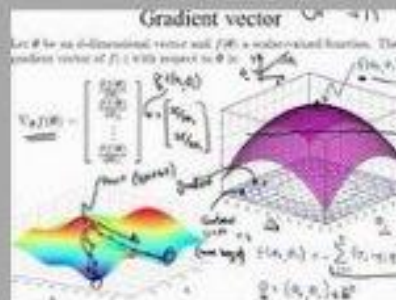
What my mom thinks I do



What society thinks I do



What my boss thinks I do

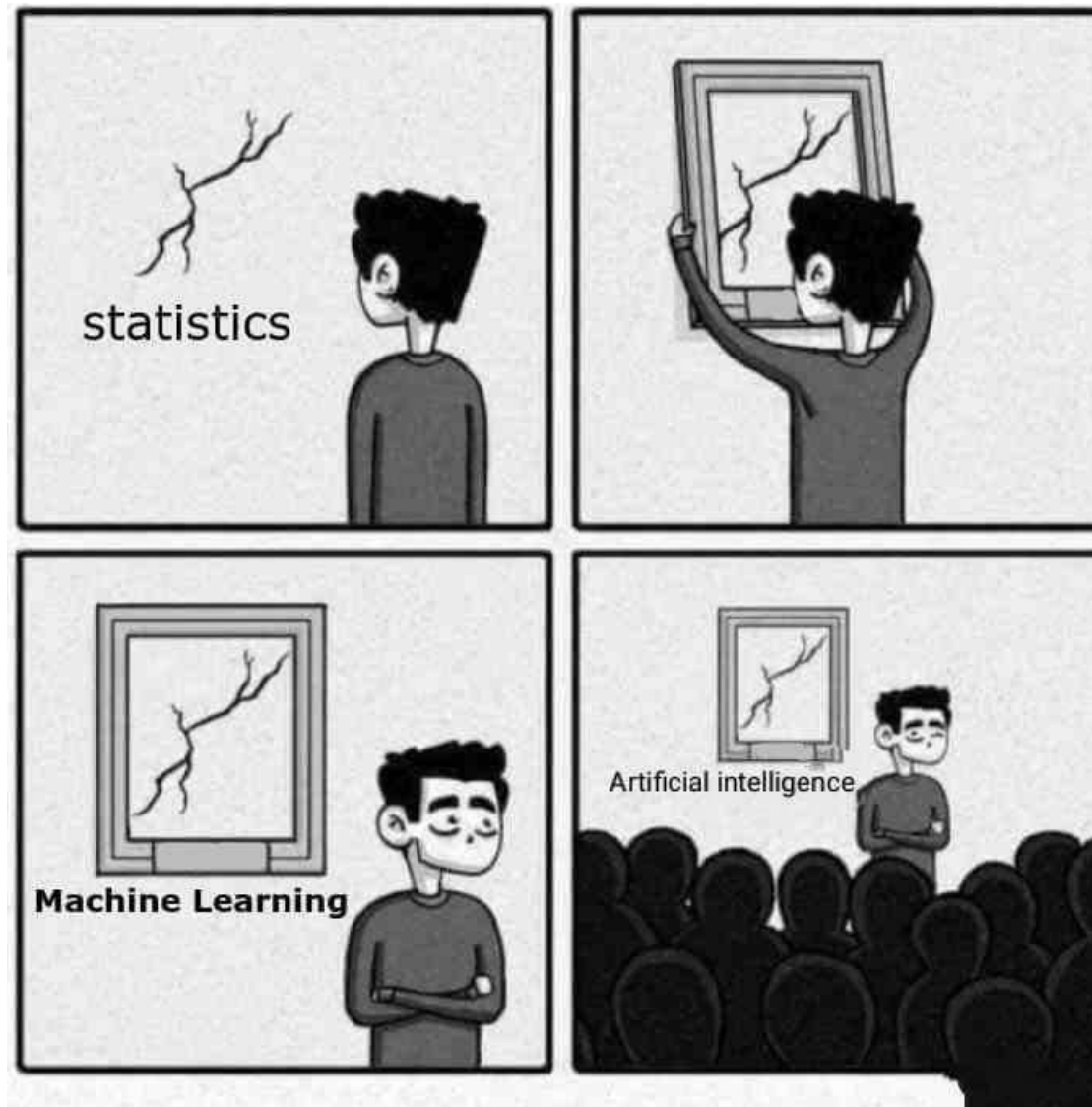


What I think I do

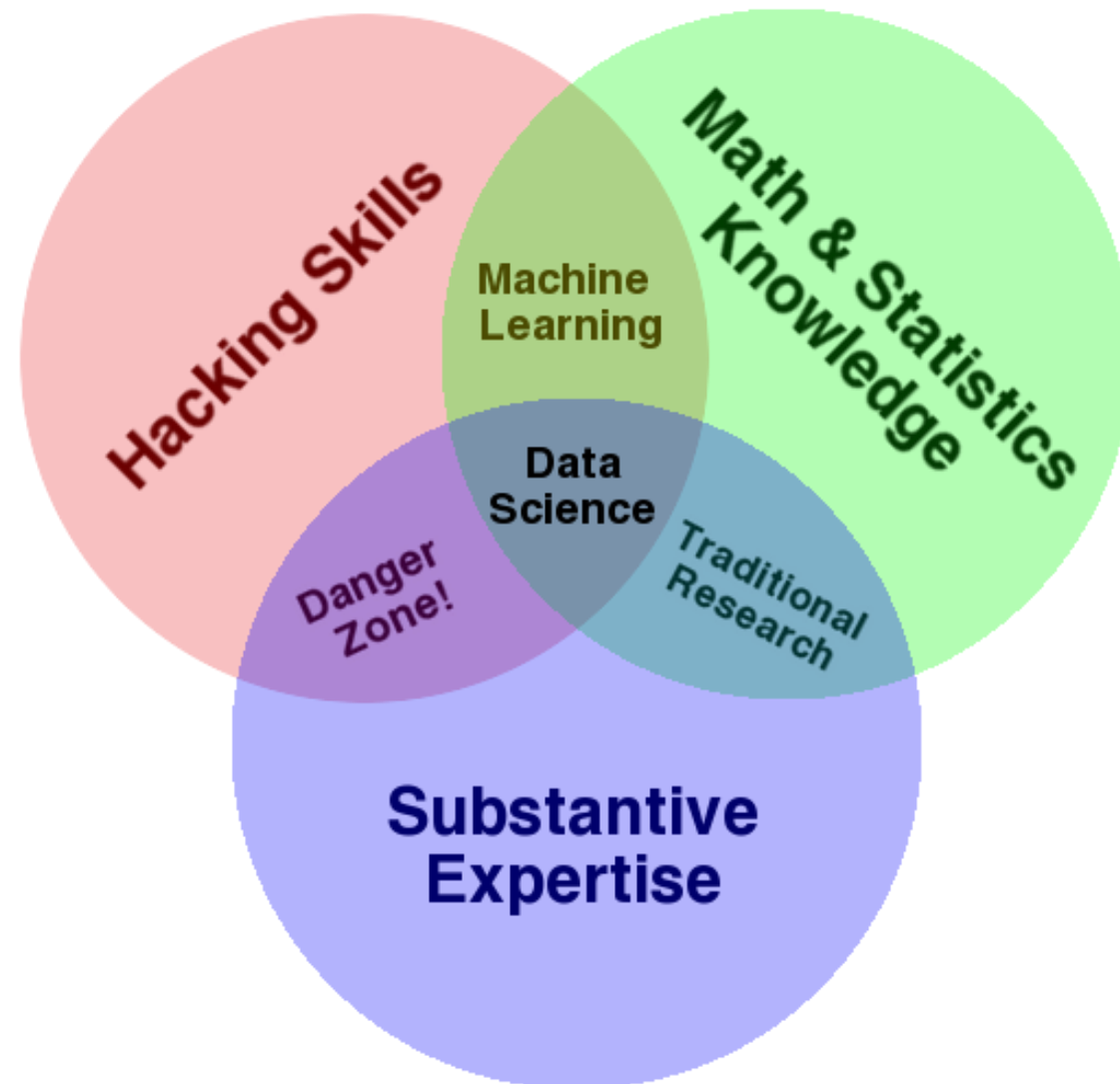


What I actually do

# 資料科學是？



# 資料科學涵蓋領域



## MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

### MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

### PROGRAMMING & DATABASE


- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

### DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

### COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



# 資料科學的流程

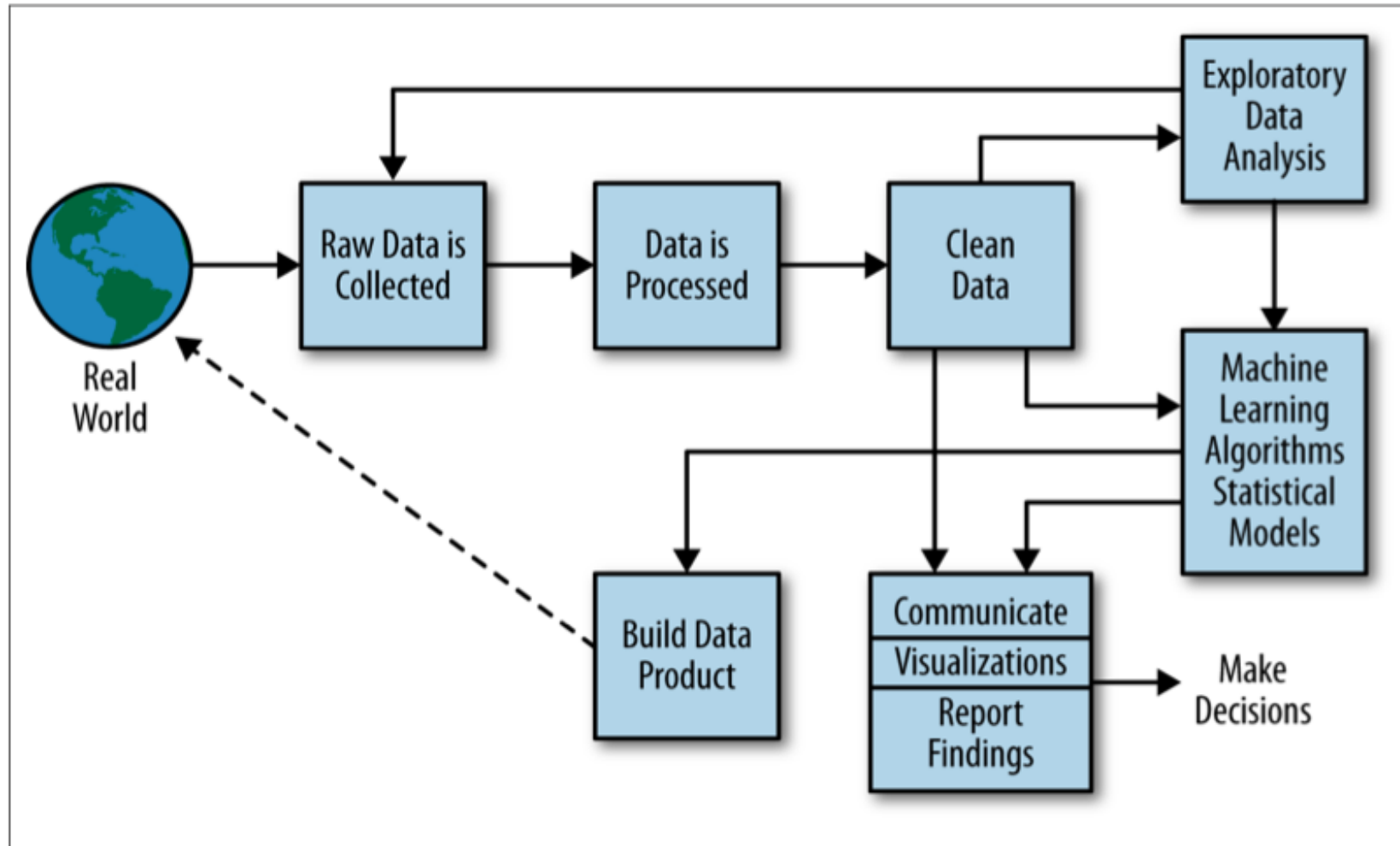


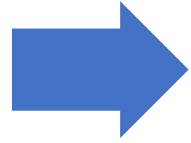
Figure 2-2. The data science process

# 資料科學 with Python

---

## Collect Data

- BeautifulSoup
- Lxml
- Requests
- Pandas



## Preprocessing and EDA

- Numpy
- Pandas
- Scikit-learn
- Matplotlib
- NLTK



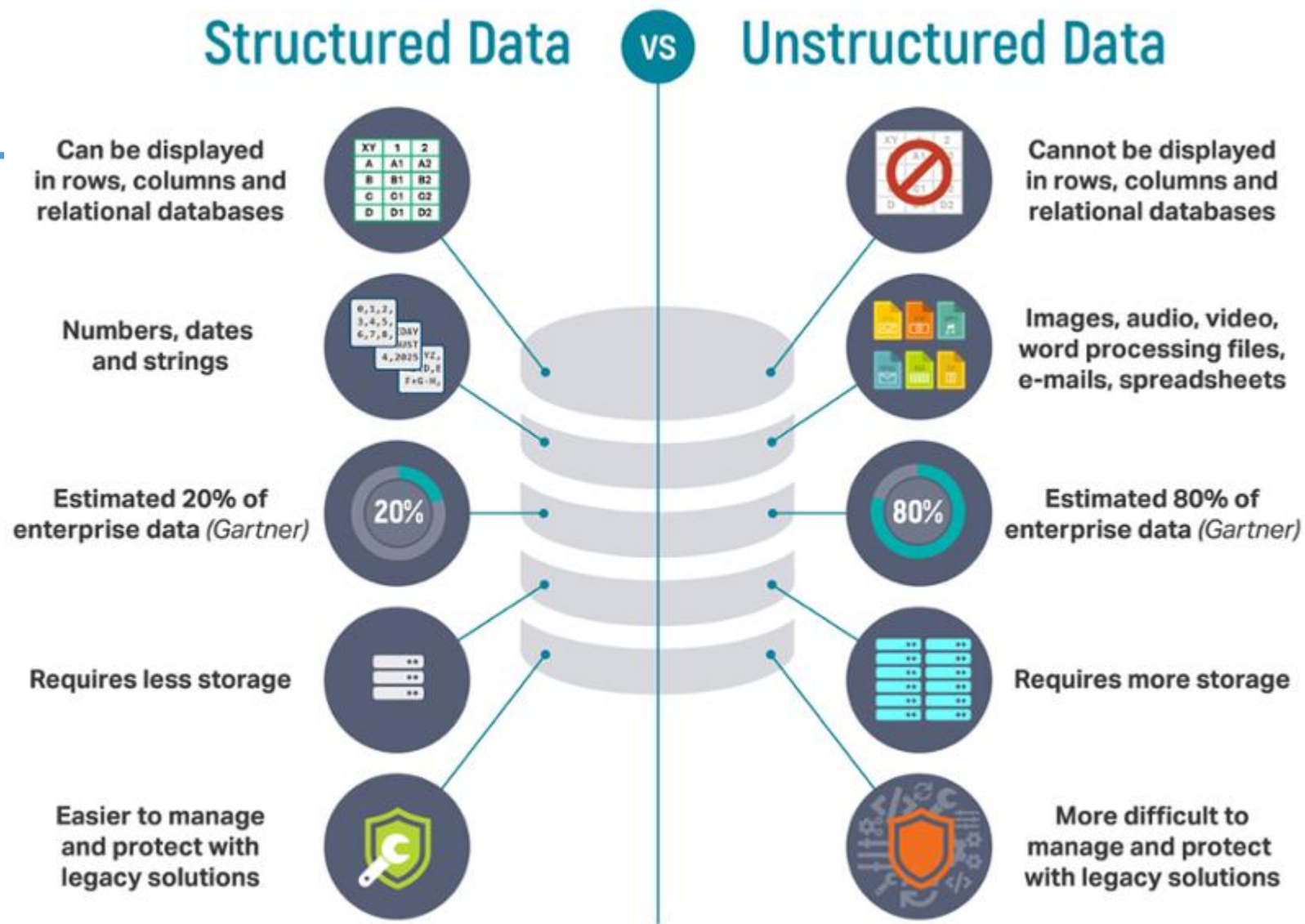
## Analysis and Modeling

- Statsmodels
- Scikit-learn
- Tensorflow
- Keras
- Pytorch

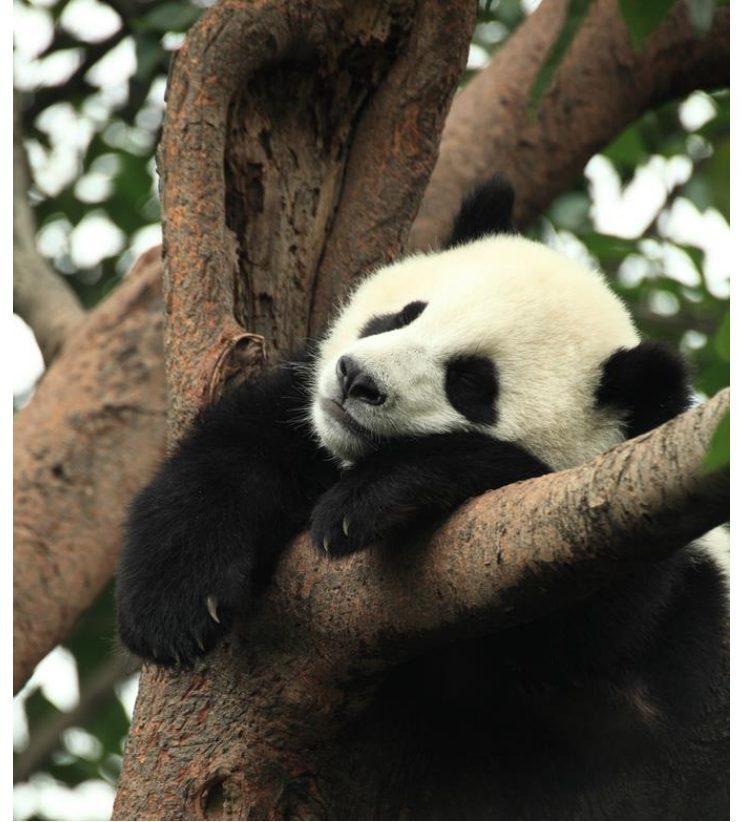


# 資料型態

- 結構化資料
  - 表格
- 非結構化資料
  - 文字
  - 圖片
  - 音訊







# Pandas

# Pandas套件相關資源

---

- [從 pandas 開始 Python 與資料科學之旅](#)
- [Pandas tutorials](#)
- [Pandas exercises](#)

# Pandas功能

---

- 讀取檔案
- 觀察資料
- 改變資料型態
- 遺漏值偵測與處理
- 挑選欄位
- 篩選資料
- 刪除/新增欄位
- 描述統計量
- 排序資料
- 群組
- 合併資料
- Apply
- 交叉列聯表
- 繪圖
- 寫出檔案



# Pandas

補充資料

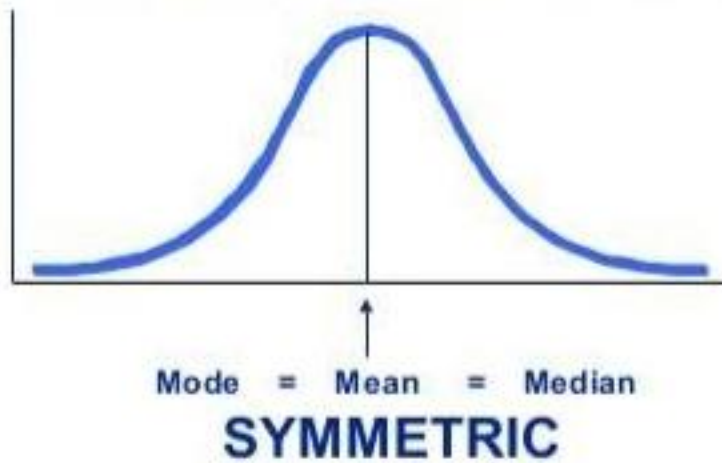
# 欄位型態

參考資料：[Wikipedia](#)

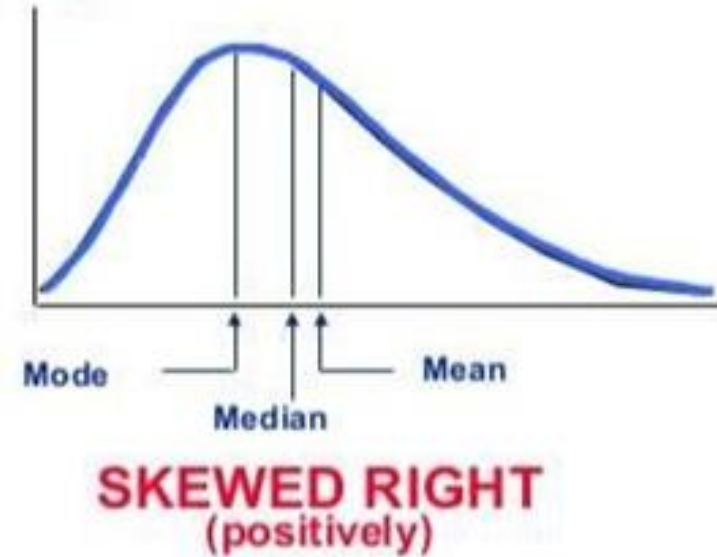
名稱	範例	特性	較常對應的pandas資料類型
名目尺度(Nominal)	國家	數學運算	'Object' 'Category'
次序尺度(Ordinal)	名次	排序	
等距尺度(Interval)	攝氏溫度	加減	'float' 'int'
等比尺度(Ratio)	重量	乘除	

# 集中趨勢指標與偏態

眾數 = 中位數 = 平均數

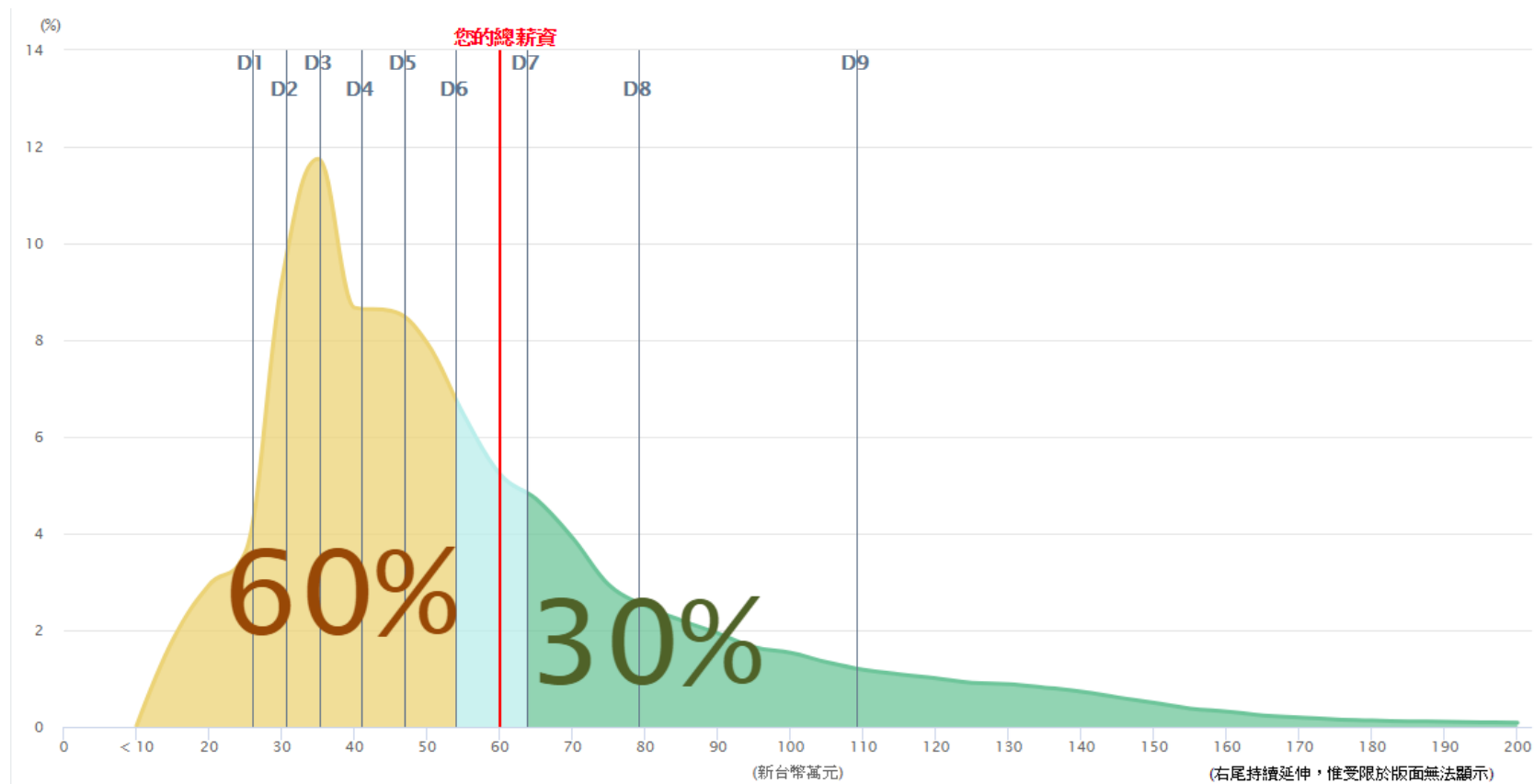


眾數 < 中位數 < 平均數





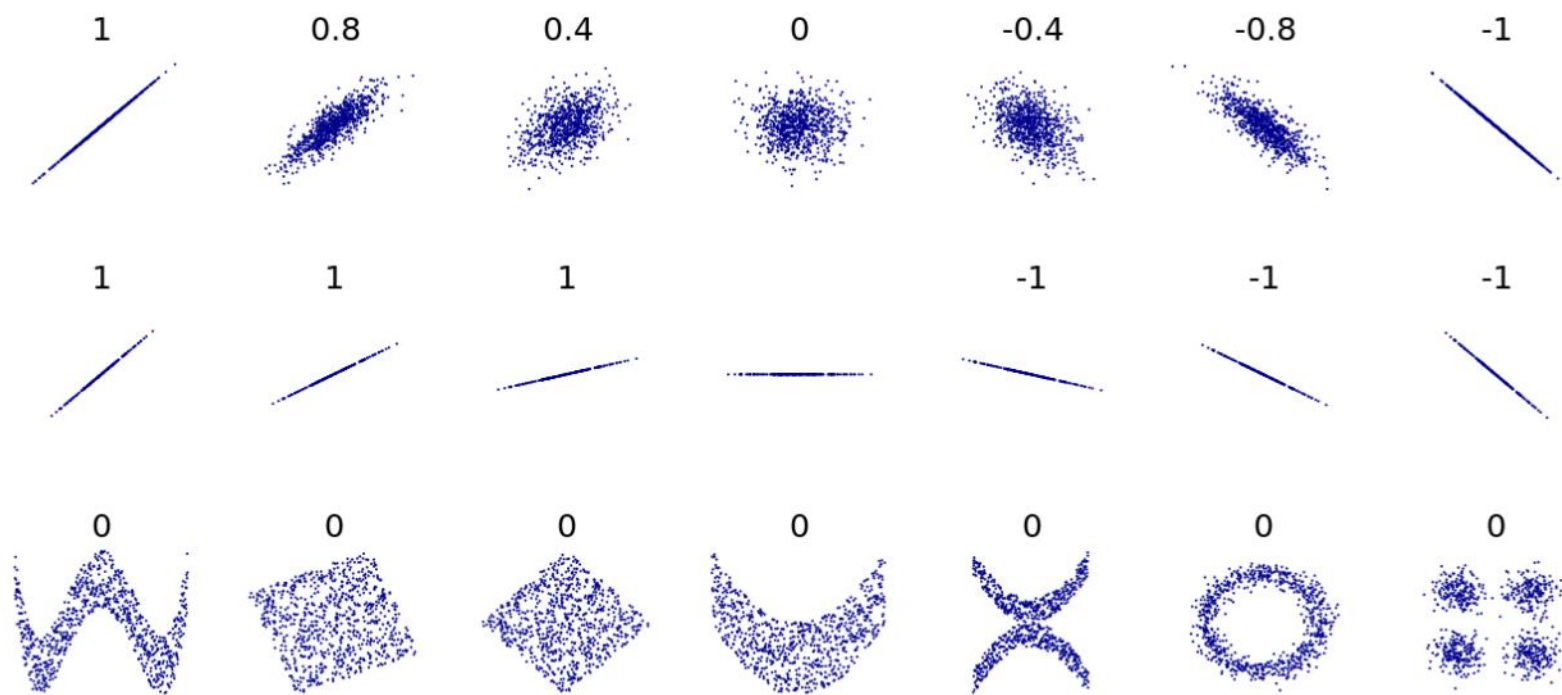
# 集中趨勢指標與偏態



行政院主計總處：[106年每年每月總薪資平均為49,989元](#)

行政院主計總處：[薪資平台](#)

# 相關係數



[Source](#)