Analytics Framework and Data Preprocessing

Dr. Chia-Yen Lee (李家岩 博士）

Department of Information Management (資訊管理學系)
National Taiwan University (國立台灣大學)

Course Contents

- Data Science & Manufacturing Systems
  - Data, Information, Knowledge, and ML/DS Functions
  - **Analytics Framework and Data Preprocessing**
  - Manufacturing Systems and Factory Dynamics
- Diagnostic and Predictive Analytics
  - Feature Selection and Feature Engineering
  - Regression, Classification, MARS, and Symbolic Regression
  - Tree-based Methods, Random Forest and Boosting
  - SPC, Signal Processing, and PHM
  - Clustering Analysis and Deep Learning
  - Manufacturing Practice
- Prescriptive Analytics
  - Linear Programming and Capacity Planning
  - Metaheuristic Algorithm and Genetic Algorithm
  - Scheduling Optimization and Run-to-Run Control
- Advanced Techniques (if time permits)
  - Concept Drift and Domain Adaptation
  - Transfer Learning, Meta-Learning, Few-shot Learning, Small Samples
- Term-project Presentation (or Exam)

Outline

- An Example of Data Mining
  - Association Rules
- Data Science Framework
  - CRoss-Industry Standard Process for Data Mining (CRISP)
  - Analytics Framework
- Data Preprocessing
  - General Preprocessing
  - Manufacturing dataset

Materials mainly and courteously come from

1. Han, Jiawei, Micheline Kamber, Jian Pei, 2011. Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann (天瓏代理)。王派洲譯，2008，資料探勘：概念與方法，第二版，滄海書局
2. 李家岩，2017，智慧製造與生產線上的資料科學 Data Science in Manufacturing: From Predictive to Prescriptive，臺灣資料科學年會。
3. 簡禎富, 許嘉裕, 2014. 資料挖礦與大數據分析, 前程文化。

Association Rules

(關聯規則)

Materials mainly and courteously come from
Kusiak, A. (2011), Computational Intelligence, Course Lecture Notes, Intelligent Systems Laboratory, The University of Iowa.
Tan, P.-N., M. Steinbach, V. Kumar (2005), Introduction to Data Mining. 1st eds, Addison-Wesley; 1 edition (May 12, 2005).
Tseng, Chi-Yao Tseng (2012), Advanced Algorithms in Computational Biology, Course Lecture Notes, Institute of Information Science, Academia Sinica, Taiwan.

Association Rules (Unsupervised Learning)

關聯規則 : 有哪些itemset高頻出現。製造常常有綁機的規則(這台接續另台)。常一起被購買的商品(無先後)。

Idea

- Frequent pattern: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set
- First proposed by Agrawal, Imielinski, and Swami [AIS93] in the context of frequent itemsets and association rule mining

Motivation

- Finding inherent regularities in data
- What upstream and downstream machines were bound together? (route)
- What products were often purchased together?— Beer and diapers?!
- What are the subsequent purchases after buying a PC?
- What kinds of DNA are sensitive to this new drug?
- Can we automatically classify web documents?

Applications

- Product-machine route, yield analysis, Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis

Transaction dataset

- find the customers’ purchasing behavior pattern

Record	Items（ID）
101	Milk(A), Bread(B), Cookie(C), Juice(D)
102	Bread(B), Cookie (C), Soda(E), Noodles(F)
103	Milk(A), Cookie (C), Fruit(G)
104	Milk(A), Bread(B), Juice(D), Noodles(F), Fruit(G)
105	Cookie(C), Soda (E), Fruit(G)

Basket binary dataset

Record	Milk(A)	Bread(B)	Cookie(C)	Juice(D)	Soda(E)	Noodles(F)	Fruit(G)
101	1	1	1	1	0	0	0
102	0	1	1	0	1	1	0
103	1	0	1	0	0	0	1
104	1	1	0	1	0	1	1
105	0	0	1	0	1	0	1

Similarity of Binary Vector

欄位作為向量Vector

Similarity of Two Binary Vectors

- The similarity value (or correlation coefficient) is between 0 and 1. 1 indicates a total correlation and 0 indicates no correlation.

- If x1 and x2 are two binary vectors with n elements. There are four counters which could be built to estimate the frequency between two elements in the same position.

兩向量都是0的次數
- f_{00}: the counts with x1=0 and x2=0
- f_{01}: the counts with x1=0 and x2=1
- f_{10}: the counts with x1=1 and x2=0
- f_{11}: the counts with x1=1 and x2=1

Example:
x1 = (0 1 0 1 0 1)
x2 = (1 1 1 0 0 0)

we can derive

$$f_{00} = 1$$
$$f_{01} = 2$$
$$f_{10} = 2$$
$$f_{11} = 1$$

- Simple Matching Coefficient, SMC (簡單配對係數)
  - $SMC = \frac{\text{coutns of both equal to 0 or equal to 1}}{\text{number of all elements in the vector}} = \frac{f_{00}+f_{11}}{f_{00}+f_{01}+f_{10}+f_{11}}$

SMC有f00，若商品數兩眾多，則f00會相當多，則SMC近似1。
Jacard則不受資料稀疏性影響。

- Jaccard Coefficient
  - $Jaccard = \frac{\text{coutns of both equal to 1}}{\text{number of all elements in the vector except both equal to 0}}$

$$= \frac{f_{11}}{f_{01}+f_{10}+f_{11}}$$

SMC and Jaccard

- Example:
  - x1 = (1 0 0 0 0 0 0 0 0 0)
  - x2 = (0 0 0 0 0 0 1 0 0 1)
  - we can derive
    - $f_{00} = 7$
    - $f_{01} = 2$
    - $f_{10} = 1$
    - $f_{11} = 0$
  - SMC $= \frac{f_{00}+f_{11}}{f_{00}+f_{01}+f_{10}+f_{11}} = \frac{7+0}{7+2+1+0}$=0.7
  - $Jaccard = \frac{f_{11}}{f_{01}+f_{10}+f_{11}} = \frac{0}{2+1+0}$=0
- Sparse matrix issue

相關性看似低，應用Jaccard修正

Similarity of Binary Vector

- How about the Pearson Correlation Coefficient?

  不推薦使用

- 在零售店中可能的購物籃組合

顧客 #1	啤酒、椒鹽脆餅、洋芋片、阿斯匹靈
顧客 #2	尿布、嬰兒乳液、葡萄柚汁、嬰兒食品、牛奶
顧客 #3	汽水、洋芋片、牛奶
顧客 #4	湯、啤酒、牛奶、冰淇淋
顧客 #5	蘇打、咖啡、牛奶、麵包
顧客 #6	啤酒、洋芋片

Similarity of Binary Vector

- 產品同時被購買的統計表

顧客 #1 ：啤酒、椒鹽脆餅、洋芋片、阿斯匹靈

顧客 #2 ：尿布、嬰兒乳液、葡萄柚汁、嬰兒食品、牛奶

顧客 #3 ：汽水、洋芋片、牛奶

顧客 #4 ：湯、啤酒、牛奶、冰淇淋

顧客 #5 ：蘇打、咖啡、牛奶、麵包

顧客 #6 ：啤酒、洋芋片

	啤酒	洋芋片	牛奶	尿布	汽水
啤酒	3	2	1	0	0
洋芋片	2	3	1	0	1
牛奶	1	1	4	1	1
尿布	0	0	1	1	0
汽水	0	1	1	0	1

Similarity of Binary Vector

- 商品相關係數矩陣
  - 皮爾森(Pearson)相關係數
  - 有偏差的結果

$$\rho_{X,Y} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

	啤酒	洋芋片	牛奶	尿布	汽水
啤酒	1				
洋芋片	0.333333	1			
牛奶	-0.70711	-0.70711	1		
尿布	-0.44721	-0.44721	0.316228	1	
汽水	-0.44721	0.447214	0.316228	-0.2	1

汽水跟洋芋片是低相關

汽水跟尿步是負相關 => 不合理
因為應該是0相關。

Similarity of Binary Vector

- 商品相關性以Jaccard係數表現

	啤酒	洋芋片	牛奶	尿布	汽水
啤酒					
洋芋片	0.500				
牛奶	0.167	0.167			
尿布	0	0	0.250		
汽水	0	0.333	0.250	0	

- 找出購買型態的不同商品間的特性
  - 1. 互補關係，如啤酒和椒鹽脆餅
  - 2. 相似購買週期，如牛奶和水果
  - 3. 反應家中偏好或地理位置的關係

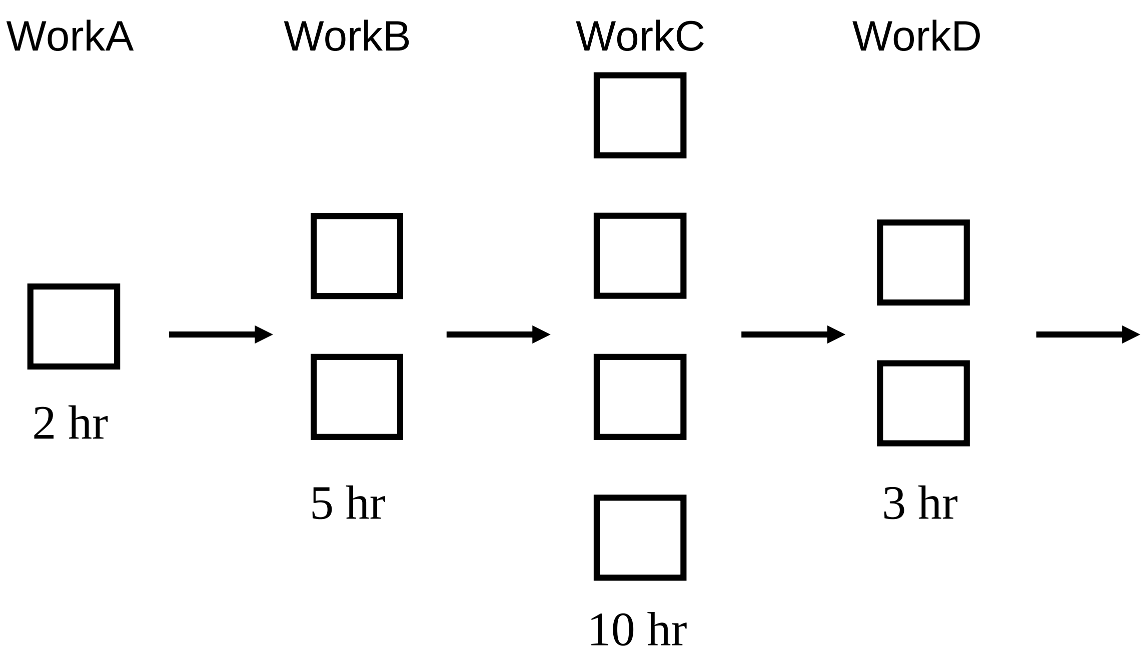
找出 Gold Path 黃金路徑，1加工成功，0加工失敗。

找出第幾站在第幾機台，良率會最高。可以找那些"變數"跟良率高相關

Machine Route and Yield

Record	WorkA_M1	WorkA_M2	WorkB_M1	WorkB_M2	WorkC_M1	WorkC_M2	…	Yield
Lot1	1	0	1	0	1	0		1
Lot2	0	1	1	0	0	1		1
Lot3	1	0	1	0	1	0		0
Lot4	1	0	1	0	1	0		1
Lot5	0	1	0	1	1	0		1

0: no pass
1: process



Machine Parameters and Yield

Record	Var.1	Var.2	Var.3	Var.4	Var.5	Var.6	...	Yield
Lot1	1	0	1	0	1	0		1
Lot2	0	1	1	0	0	1		1
Lot3	1	0	1	0	1	0		0
Lot4	1	0	1	0	1	0		1
Lot5	0	1	0	1	1	0		1

0: low level
1: high level

Material/part/liquid usage and Yield (for each workstation)

Record	WorkA	WorkB	WorkC	WorkD	WorkE	WorkF	...	Yield
Lot1	1	0	1	0	1	0		1
Lot2	0	1	1	0	0	1		1
Lot3	1	0	1	0	1	0		0
Lot4	1	0	1	0	1	0		1
Lot5	0	1	0	1	1	0		1

0: no
1: use

Association Rules

Introduction

- Mining for associations among items in a large database of sales transaction is an important database mining function.
- For example, the information that a customer who purchases a keyboard also tends to buy a mouse at the same time is represented in association rule below: (Transaction data analysis)

  Keyboard ⇒ Mouse [Support = 6%, Confidence = 70%]

  你買鍵盤，則你會買滑鼠 [Support, Confd]

Association Rules

A→B [*support, confidence*]

- Based on the types of values, the association rules can be classified into two categories: Boolean Association Rules and Quantitative Association Rules

Diaper → Beer [0.5%, 75%]
(support, confidence)

- ***Boolean Association Rule***:
  - Keyboard ⇒Mouse [support = 6%, confidence = 70%]
- ***Quantitative Association Rule***:
  - (Age = 26…30) ⇒ (Cars =1, 2) [Support 3%,confidence = 36%]

Rule Evaluation Criteria

- 衡量指標
  - 支持度(support)：衡量關聯規則的顯著性
  - 信賴度(confidence)：衡量關聯規則的正確性
  - 增益(lift)：衡量關聯規則的資訊價值

- 篩選關聯規則
  - 最小支持度 (minimum support)門檻
  - 最小信賴度 (minimum confidence)門檻

Rule Evaluation Criteria

Support 支持度 (how useful is the rule)

- The support of an association pattern is the percentage of task-relevant data tuples for which the pattern is true. (i.e. frequency)
- $Support(A \Rightarrow B) = \dfrac{\#\_tuples\_containing\_both\_A\_and\_B}{total\_\#\_of\_tuples} = P(A \cap B)$

A與B的交集(數量)出現頻率高不高，e.g. 同時買牛奶與麵包。AB可對調。(PS: 不是頻率低就不關注，實務上還是要看商品價值)
影片1:50

交易紀錄	牛奶(A)	麵包(B)	餅乾(C)	柳橙汁(D)	汽水(E)	泡麵(F)	水果(G)
101	1	1	1	1	0	0	0
102	0	1	1	0	1	1	0
103	1	0	1	0	0	0	1
104	1	1	0	1	0	1	1
105	0	0	1	0	1	0	1

$$Support(\text{牛奶} \Rightarrow \text{麵包}) = P(\text{麵包，牛奶}) = \frac{2}{5} = 0.4$$

- 表示關聯規則相對於全部資料須具有一定的普遍性
- Minimum Support Threshold 最小支持度門檻用於控管關聯規則所必須涵蓋的最少資料比率

Rule Evaluation Criteria

Confidence 信賴度 (how true)

- Confidence is defined as the measure of certainty or trustworthiness associated with each discovered pattern.
- The rule X ⇒Y has 90% confidence: means 90% of customers who bought X also bought Y.
- $Confidence(A \Rightarrow B) = \frac{\#\_tuples\_containing\_both\_A\_and\_B}{\#\_tuples\_containing\_A} = \frac{P(A \cap B)}{P(A)} = P(B|A)$

這個規則的信賴程度有多高，是否是有用的，條件機率高低

交易紀錄	牛奶(A)	麵包(B)	餅乾(C)	柳橙汁(D)	汽水(E)	泡麵(F)	水果(G)
101	1	1	1	1	0	0	0
102	0	1	1	0	1	1	0
103	1	0	1	0	0	0	1
104	1	1	0	1	0	1	1
105	0	0	1	0	1	0	1

$$Confidence(\text{牛奶} \Rightarrow \text{麵包}) = P(\text{麵包}|\text{牛奶}) = \frac{2/5}{3/5} = 0.667$$

- 信賴度要達到一定水準時，關聯規則才會具有意義
- Minimum Confidence Threshold最小信賴度門檻主要用於去除信心較低的關聯規則

Rule Evaluation Criteria

Lift 增益值 (how really true)

- 用於比較信賴度與結果項目Y單獨發生時兩者機率間的大小，為衡量該關聯規則之有效性，也就是判定該規則的條件機率是否比原本發生的機率大

$$Lift(X \Rightarrow Y) = \frac{P(Y \mid X)}{P(Y)} = \frac{P(X \cap Y)}{P(X)P(Y)}$$

Lift > 1 時這條規則才有意義 = (買牛奶再加買麵包的機率)/(原本就會買麵包的機率) => (買牛奶再加買麵包的機率) 要大於 (原本就會買麵包的機率)

- 增益值>1，代表此關聯規則的信賴度大於原本結果項目Y發生機率，表示該關聯規則的預測結果比原本表現好
- 增益值<1，表示透過關聯規則的預測結果比原本預測能力差

交易紀錄	牛奶(A)	麵包(B)	餅乾(C)	柳橙汁(D)	汽水(E)	泡麵(F)	水果(G)
101	1	1	1	1	0	0	0
102	0	1	1	0	1	1	0
103	1	0	1	0	0	0	1
104	1	1	0	1	0	1	1
105	0	0	1	0	1	0	1

$$Lift(\text{牛奶} \Rightarrow \text{麵包}) = \frac{P(\text{麵包} \mid \text{牛奶})}{P(\text{麵包})} = \frac{2/3}{3/5} = 1.111$$

Rule Evaluation Criteria

- 顧客於購買牛奶與麵包的同時也會選購餅乾為例：

交易紀錄	牛奶(A)	麵包(B)	餅乾(C)	柳橙汁(D)	汽水(E)	泡麵(F)	水果(G)
101	1	1	1	1	0	0	0
102	0	1	1	0	1	1	0
103	1	0	1	0	0	0	1
104	1	1	0	1	0	1	1
105	0	0	1	0	1	0	1

$$Support(\text{牛奶，麵包} \Rightarrow \text{餅乾}) = P(\text{牛奶，麵包，餅乾}) = 0.2$$

$$Confidence(\text{牛奶，麵包} \Rightarrow \text{餅乾}) = P(\text{餅乾} \mid \text{牛奶，麵包}) = 0.5$$

$$Lift(\text{牛奶，麵包} \Rightarrow \text{餅乾}) = \frac{P(\text{餅乾} \mid \text{牛奶，麵包})}{P(\text{餅乾})} = \frac{0.5}{0.8} = 0.625$$

Itemset

Itemset

- A set of items is referred to as itemset.
- An itemset containing *k* items is called *k*-itemset.
- An itemset can also be seen as a conjunction of items (or a predicate)

Frequent Itemsets

- Suppose min_sup is the minimum support threshold.
- An itemset satisfies minimum support if the occurrence frequency of the itemset is greater than or equal to min_sup.
- If an itemset satisfies minimum support, then it is a frequent itemset.

Strong Rules

- We are often interested in only strong associations, i.e.,
  - ***support*** $\geq$ ***min_sup***
  - ***confidence*** $\geq$ ***min_conf***
- Strong Rules
  - Rules that satisfy both a minimum support threshold and a minimum confidence threshold are called strong.
- Examples:
  - milk $\rightarrow$ bread [5%, 60%]
  - tire and auto_accessories $\rightarrow$ auto_services [2%, 80%].
- Association Rule Mining
  - Find all frequent itemsets
  - Generate strong association rules from the frequent itemsets

Apriori Algorithm

- Apriori algorithm is an influential algorithm for mining *frequent itemsets* for Boolean association rules. Apriori (R. Agrawal and R. Srikant. Fast algorithms for mining association rules. VLDB'94.)

- Uses a ***Level-wise search***, where **k-itemsets** (An itemset that contains k items is a **k-itemse**t) are used to explore (k+1)-itemsets, to mine frequent itemsets from transactional database for Boolean association rules.

- First, the set of frequent 1-itemsets is found. This set is denoted L1. L1 is used to find L2, the set of frequent 2-itemsets, which is used to fine L3, and so on, until no more frequent k-itemsets can be found.

Association Rule Mining Process

Apriori Algorithm

- Derivation of large 1-itemsets L_1: At the first iteration, scan all the transactions and count the number of occurrences for each item.

- Level-wise derivation: At the k^{th} iteration, the candidate set C_k are those whose every (k-1)-item subset is in L_{k-1}. Scan DB and count the # of occurrences for each candidate itemset.

Association rule mining process

- Find all ***frequent itemsets***:
  - Each support S of these frequent itemsets will at least equal to a pre-determined min_sup (An itemset is a subset of items in I, like A)

- Generate ***strong association rules*** from the frequent itemsets:
  - These rules must be the frequent itemsets and must satisfy min_sup and min_conf.

Example

- Example

Transactional data for an *AllElectronics* branch.

TID	*List of item_IDs*
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Example

Transactional data for an *AllElectronics* branch.

TID	*List of item_IDs*
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

1-Itemsets

Sup_cunt = 出現次數，min_sup 通通 = 2

Scan *D* for count of each candidate →

C_1

Itemset	Sup. count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Compare candidate support count with minimum support count →

L_1

Itemset	Sup. count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

2-Itemsets

Generate C_2 candidates from L_1 →

C5選2的組合

C_2

Itemset
{I1, I2}
{I1, I3}
{I1, I4}
{I1, I5}
{I2, I3}
{I2, I4}
{I2, I5}
{I3, I4}
{I3, I5}
{I4, I5}

Scan *D* for count of each candidate →

C_2

Itemset	Sup. count
{I1, I2}	4
{I1, I3}	4
{I1, I4}	1
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2
{I3, I4}	0
{I3, I5}	1
{I4, I5}	0

Compare candidate support count with minimum support count →

L_2

Itemset	Sup. count
{I1, I2}	4
{I1, I3}	4
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2

Min Support = 2

Frequent 3-Itemsets

觀察出，{1,2,3}是同時會被買的商品

Generate C_3 candidates from L_2 →

C_3

Itemset
{I1, I2, I3}
{I1, I2, I5}

Scan *D* for count of each candidate →

C_3

Itemset	Sup. count
{I1, I2, I3}	2
{I1, I2, I5}	2

Compare candidate support count with minimum support count →

L_3

Itemset	Sup. count
{I1, I2, I3}	2
{I1, I2, I5}	2

Example

Rule Generation

由{1,2,5}這個itemset，可以衍生出6條排列組合

Generating association rules. Let's try an example based on the transactional data for *AllElectronics* shown in Table 5.1. Suppose the data contain the frequent itemset $l = \{I1, I2, I5\}$. What are the association rules that can be generated from l? The nonempty subsets of l are $\{I1, I2\}$, $\{I1, I5\}$, $\{I2, I5\}$, $\{I1\}$, $\{I2\}$, and $\{I5\}$. The resulting association rules are as shown below, each listed with its confidence:

$$
\begin{array}{ll}
I1 \wedge I2 \Rightarrow I5, & confidence = 2/4 = 50\% \\
I1 \wedge I5 \Rightarrow I2, & confidence = 2/2 = 100\% \\
I2 \wedge I5 \Rightarrow I1, & confidence = 2/2 = 100\% \\
I1 \Rightarrow I2 \wedge I5, & confidence = 2/6 = 33\% \\
I2 \Rightarrow I1 \wedge I5, & confidence = 2/7 = 29\% \\
I5 \Rightarrow I1 \wedge I2, & confidence = 2/2 = 100\%
\end{array}
$$

買2，也會買1,5的機率是100%

If the minimum confidence threshold is, say, 70%, then only the second, third, and last rules above are output, because these are the only ones generated that are strong. Note that, unlike conventional classification rules, association rules can contain more than one conjunct in the right-hand side of the rule. ■

Apriori Algorithm

Pseudocode

Algorithm: Apriori. Find frequent itemsets using an iterative level-wise approach based on candidate generation.

Input:

- D, a database of transactions;
- min_sup, the minimum support count threshold.

Output: L, frequent itemsets in D.

Method:

L_1 = find_frequent_1-itemsets(D);
for ($k = 2; L_{k-1} \neq \phi; k{+}{+}$) {
    C_k = apriori_gen(L_{k-1});
    for each transaction $t \in D$ { // scan D for counts
        C_t = subset(C_k, t); // get the subsets of t that are candidates
        for each candidate $c \in C_t$
            c.count++;
    }
    $L_k = \{c \in C_k | c.count \geq min\_sup\}$
}
return $L = \cup_k L_k$;

procedure apriori_gen(L_{k-1}:frequent $(k-1)$-itemsets)
for each itemset $l_1 \in L_{k-1}$
    for each itemset $l_2 \in L_{k-1}$
        if $(l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge ... \wedge (l_1[k-2] = l_2[k-2]) \wedge (l_1[k-1] < l_2[k-1])$ then {
            $c = l_1 \bowtie l_2$; // join step: generate candidates
            if has_infrequent_subset(c, L_{k-1}) then
                delete c; // prune step: remove unfruitful candidate
            else add c to C_k;
        }
return C_k;

procedure has_infrequent_subset(c: candidate k-itemset;
        L_{k-1}: frequent $(k-1)$-itemsets); // use prior knowledge
for each $(k-1)$-subset s of c
    if $s \notin L_{k-1}$ then
        return TRUE;
return FALSE;

Apriori 演算法

會找出所有itemse, so, 計算量相當高

- 在大量的資料集中，利用項目集來建立關聯規則，並計算每一個候選項目出現的數目，依據所設定的支持度來衡量候選項目是否可建立顯著的關聯規則

- 採用水平方向進行項目集的搜尋；透過 *k*項目集之組合去探索 *k*+1項目集，提升發現高頻項目集的效率

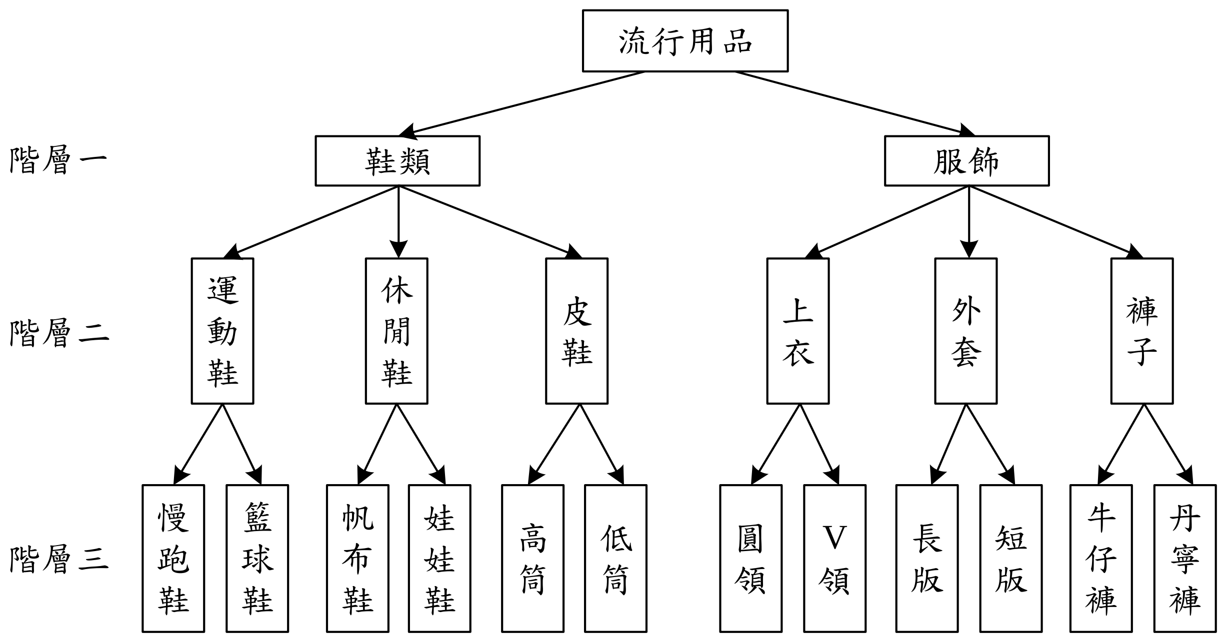
- 由單一項目集(1-itemset)開始，反覆產生候選項目集與蒐集項目集之步驟，直到找出所有高頻項目集為止

- 應用類似遞移律的概念，稱為反單調性：若某候選項目集為高頻，則其所有的子集合必定是高頻項目集

重點: 子集合!!! (ex; {i1,i2,i3}的子集合，也都是高頻，出現次數滿足min_sup)

以所涵蓋的抽象層級為基礎

- 單一層級關聯規則(single-level association rule)
  - 規則屬性或項目全為同一層級
  - 如購買牛奶⇒購買麵包，可從中得到較具體與精確的資訊
- 多階層關聯規則(multilevel association rule)
  - 同時包含較低階層和較高階層的項目集集合的多階層資料
  - 先建立概念層級樹(concept hierarchy tree)



多層關聯規則方法：一致支持度

- 一致支持度: 在各層次間使用相同的支持度
  - **優點**： 如果一項目集合的父項目集合不滿足最小支持度，那其本身也不會滿足最小支持度。
  - **缺點**：底層項較不容易成為頻繁集合，如果支持度
    - 太高 ⇒ 丟失去底層關聯規則
    - 太低 ⇒ 產生較多無趣的高層關聯規則

層1
min_sup = 5%

上層的發生頻率一定高!
所以應該要遞減min_sup_cnt

牛奶
[support = 10%]

層 2
min_sup = 5%

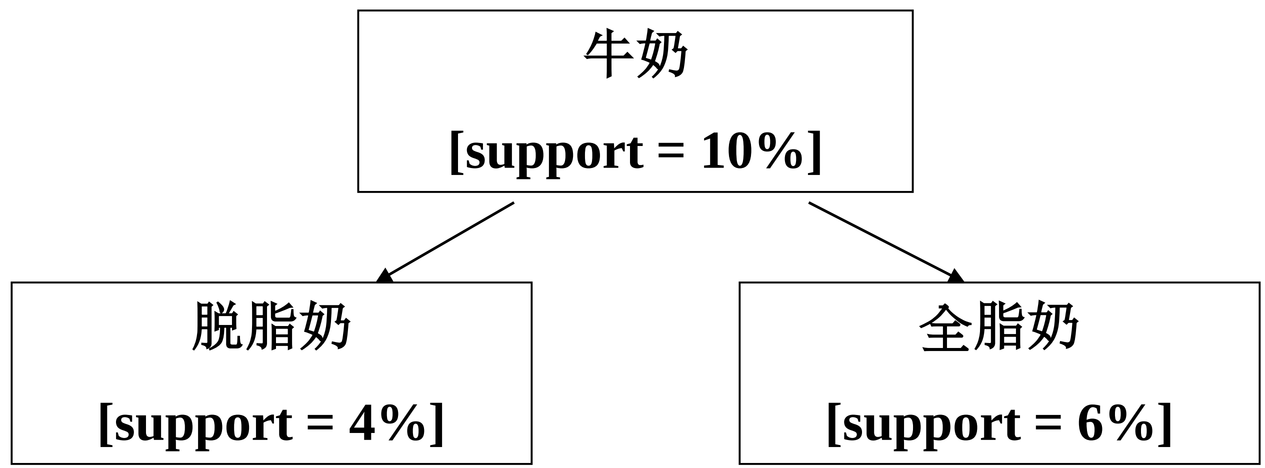
脫脂奶
[support = 4%]

全脂奶
[support = 6%]

- 遞減支持度: 隨著層次降低支持度遞減
  - 搜尋策略：
    - 層與層互相獨立
    - 用單項目跨階層過濾
    - 用k-項目集合進形階層交叉過濾

層1
min_sup = 5%

層 2
min_sup = 3%



檢查冗餘的多層關聯規則

- 由於項目之間的“祖先”關係，有些規則可能是冗餘的。
- 例子
  - 牛奶 ⇒ 白麵包 [support = 8%, confidence = 70%]
  - 全脂奶 ⇒ 白麵包 [support = 2%, confidence = 72%]
- 稱第一個規則是第二個規則的“祖先”
- 根據規則的祖先版本計算，若它的支持度和信賴度接近我們“預期”的值，則此規則是冗餘的(redundant)。
- 若規則一有70%信賴度與8%支持度，且約有1/4牛奶其銷售為全脂，則規則二之度量值接近“預期”的值(sup. 8%*1/4, 大約70% conf.)，其並非有趣的規則。

Association Rules Visualization

POLab

SECOM dataset

做關聯規則 => 資料要是二元的，比平均高=1，比平均低=0。
是非監督式學習，所以label也可以拿進當當rule的參數。

- Data source: https://archive.ics.uci.edu/ml/datasets/SECOM
- 1567 lots with 590 variables, 1 label for inspection results (PASS, FAIL)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	VS	VT	VU	VV	VW
1		SVID_1	SVID_2	SVID_3	SVID_4	SVID_5	SVID_6	SVID_7	SVID_8	SVID_9	SVID_10	SVID_11	SVID_12	SVID_13	SVID_14	SVID_15	...	SVID_590	Label	Time	
2	Lot_1	3030.93	2564	2187.733	1411.127	1.3602	100	97.6133	0.1242	1.5005	0.0162	-0.0034	0.9455	202.4396	0	7.9558	...	NaN	-1	19/07/2008 11:55:0	
3	Lot_2	3095.78	2465.14	2230.422	1463.661	0.8294	100	102.3433	0.1247	1.4966	-0.0005	-0.0148	0.9627	200.547	0	10.1548	...	208.2045	-1	19/07/2008 12:32:0	
4	Lot_3	2932.61	2559.94	2186.411	1698.017	1.5102	100	95.4878	0.1241	1.4436	0.0041	0.0013	0.9615	202.0179	0	9.5157	...	82.8602	1	19/07/2008 13:17:0	
5	Lot_4	2988.72	2479.9	2199.033	909.7926	1.3204	100	104.2367	0.1217	1.4882	-0.0124	-0.0033	0.9629	201.8482	0	9.6052	...	73.8432	-1	19/07/2008 14:43:0	
6	Lot_5	3032.24	2502.87	2233.367	1326.52	1.5334	100	100.3967	0.1235	1.5031	-0.0031	-0.0072	0.9569	201.9424	0	10.5661	...	73.8432	-1	19/07/2008 15:22:0	
7	Lot_6	2946.25	2432.84	2233.367	1326.52	1.5334	100	100.3967	0.1235	1.5287	0.0167	0.0055	0.9699	200.472	0	8.6617	...	44.0077	-1	19/07/2008 17:53:0	
8	Lot_7	3030.27	2430.12	2230.422	1463.661	0.8294	100	102.3433	0.1247	1.5816	-0.027	0.0105	0.9591	202.0901	0	9.035	...	44.0077	-1	19/07/2008 19:44:0	
9	Lot_8	3058.88	2690.15	2248.9	1004.469	0.7884	100	106.24	0.1185	1.5153	0.0157	0.0007	0.9481	202.417	0	13.6872	...	95.031	-1	19/07/2008 19:45:0	
10	Lot_9	2967.68	2600.47	2248.9	1004.469	0.7884	100	106.24	0.1185	1.5358	0.0111	-0.0066	0.9494	202.4544	0	12.6837	...	111.6525	-1	19/07/2008 20:24:0	
11	Lot_10	3016.11	2428.37	2248.9	1004.469	0.7884	100	106.24	0.1185	1.5381	0.0159	0.0049	0.944	202.5999	0	12.4278	...	90.2294	-1	19/07/2008 21:35:0	
12	Lot_11	2994.05	2548.21	2195.122	1046.147	1.3204	100	103.34	0.1223	1.5144	-0.019	0.0013	0.9433	201.7125	0	11.8566	...	57.8122	1	19/07/2008 21:57:0	
13	Lot_12	2928.84	2479.4	2196.211	1605.758	0.9959	100	97.9156	0.1257	1.469	0.017	-0.0154	0.9445	202.1264	0	9.1084	...	75.5077	1	19/07/2008 22:52:0	
14	Lot_13	2920.07	2507.4	2195.122	1046.147	1.3204	100	103.34	0.1223	1.531	-0.0259	0.0216	0.9595	202.1269	0	8.4828	...	52.2039	-1	20/07/2008 03:35:0	
15	Lot_14	3051.44	2529.27	2184.433	877.6266	1.4668	100	107.8711	0.124	1.5236	-0.0209	-0.0031	0.9441	226.0086	0	9.7686	...	52.2039	-1	21/07/2008 08:21:0	
16	Lot_15	2963.97	2629.48	2224.622	947.7739	1.2924	100	104.8489	0.1197	1.4474	0.0144	-0.0119	0.9582	195.3787	0	9.7561	...	142.908	1	21/07/2008 11:53:0	
17	Lot_16	2988.31	2546.26	2224.622	947.7739	1.2924	100	104.8489	0.1197	1.5465	0.025	-0.0024	0.9616	192.9787	0	12.4364	...	100.2745	-1	22/07/2008 00:03:0	
18	Lot_17	3028.02	2560.87	2270.256	1258.456	1.395	100	104.8078	0.1207	1.4368	0.015	-0.0037	0.9623	195.1742	0	12.1805	...	82.0989	-1	22/07/2008 02:59:0	
19	Lot_18	3032.73	2517.79	2270.256	1258.456	1.395	100	104.8078	0.1207	1.5537	0.022	-0.0027	0.9613	195.3425	0	10.0002	...	82.0989	-1	22/07/2008 08:41:0	
20	Lot_19	3040.34	2501.16	2207.389	962.5317	1.2043	100	104.0311	0.121	1.5481	-0.0367	0.0014	0.9634	196.2746	0	8.4061	...	82.0989	-1	22/07/2008 11:47:0	
21	Lot_20	2988.3	2519.05	2208.856	1157.722	1.5509	100	107.8022	0.1233	1.5362	-0.0259	-0.0179	0.9614	197.1793	0	13.3419	...	47.1586	-1	22/07/2008 14:00:0	
22	Lot_21	2987.32	2528.81	NaN	NaN	NaN	NaN	NaN	0.1195	1.6343	-0.0263	0.0116	0.9587	200.8256	0	11.9224	...	47.1586	-1	22/07/2008 15:30:0	
23	Lot_22	NaN	2481.85	2207.389	962.5317	1.2043	100	104.0311	0.121	1.5559	0.0002	-0.0044	0.9617	196.6315	0	13.6262	...	34.4153	-1	23/07/2008 05:15:0	
24	Lot_23	3002.27	2497.45	2207.389	962.5317	1.2043	100	104.0311	0.121	1.5465	0.0195	-0.0114	0.9491	199.6394	0	15.4346	...	114.5979	-1	23/07/2008 19:22:0	
25	Lot_24	2884.74	2514.54	2160.367	899.9488	1.4022	100	105.4978	0.124	1.5585	-0.0317	-0.0138	0.9638	196.1842	0	11.6229	...	216.8869	1	25/07/2008 15:23:0	
26	Lot_25	3010.41	2632.8	2203.9	1116.413	1.2639	100	102.2733	0.1199	1.4227	0.0194	0.0073	0.9765	199.0177	0	4.704	...	125.06	-1	27/07/2008 04:18:0	
27	Lot_26	2979.74	2446.56	2257.167	1437.957	1.4918	100	106.34	0.1203	1.5136	0.0018	0.0058	0.958	205.9919	0	10.6611	...	125.06	-1	27/07/2008 09:37:0	
28	Lot_27	3067.35	2456.33	2257.167	1437.957	1.4918	100	106.34	0.1203	1.486	-0.0019	-0.0056	0.9587	206.4033	0	14.8136	...	125.06	-1	27/07/2008 11:10:0	
29	Lot_28	2988.99	2607.63	2223.033	1533.993	1.3548	100	109.7067	0.1211	1.5582	-0.0101	0.0204	0.9572	199.6076	0	5.7692		216.9552	-1	27/07/2008 15:46:0	
30	Lot_29	2972.78	2431.57	2190.489	1059.439	0.8614	100	102.1178	0.1216	1.5438	0.0065	0.0032	0.9589	206.4436	0	11.3083		127.5067	-1	27/07/2008 16:06:0	
31	Lot_30	2981.85	2529.11	2180.378	1208.741	1.2998	100	100.2789	0.1209	1.42	-0.0016	0.0138	0.962	198.7199	0	6.8715		146.8715	-1	27/07/2008 16:49:0	

McCann & Johnston (2008), https://archive.ics.uci.edu/ml/datasets/SECOM
Kao, H., Hsieh, Y., Chen, C., & Lee, J. (2017). Quality prediction modeling for multistage manufacturing based on classification and association rule mining. MATEC Web of Conferences 123, 00029.

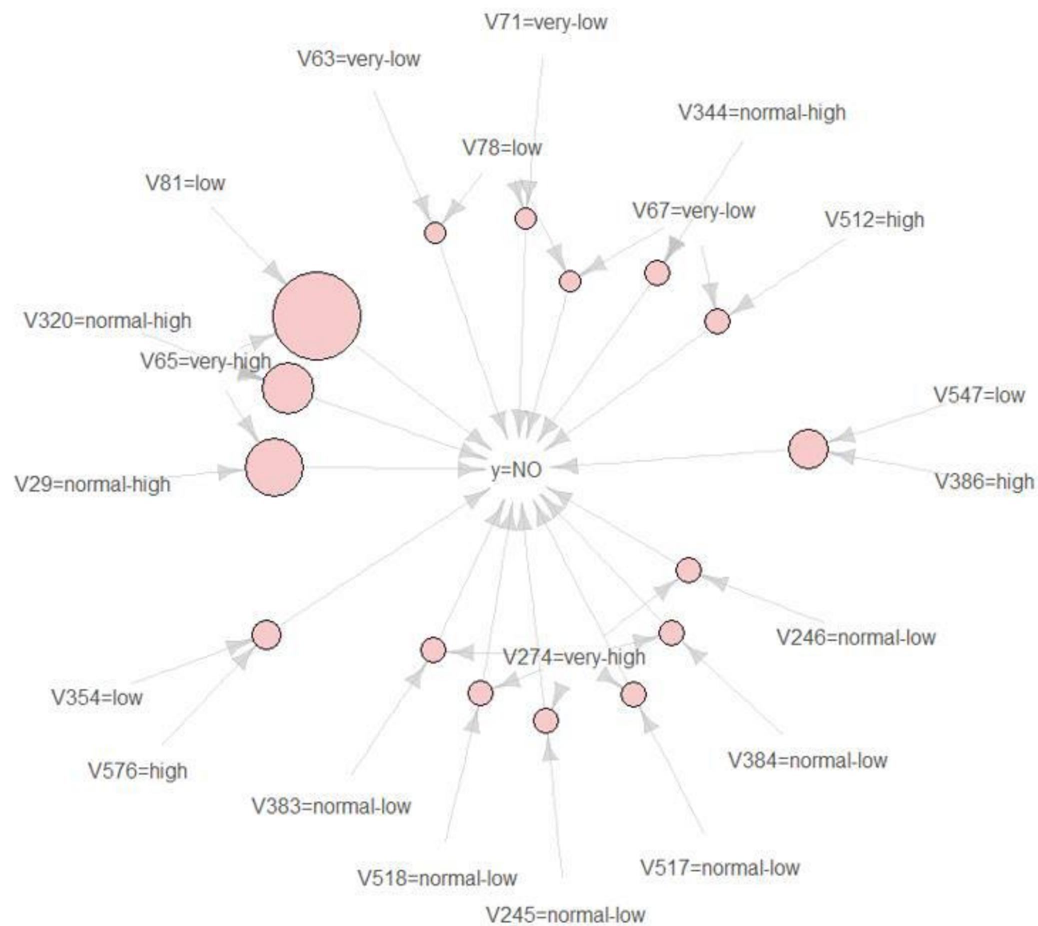
Association Rules Visualization

SECOM dataset

- Data source: https://archive.ics.uci.edu/ml/datasets/SECOM
- 1567 lots with 590 variables, 1 label for inspection results (PASS, FAIL)

Rule	Support	Confidence
{V65=very-high, V81=low}	0.0283	1
{V29=normal-high, V65=very- high}	0.0247	1
{V65=very-high, V320=normal-high}	0.0238	1
{V386=high,V547=low}	0.0224	1

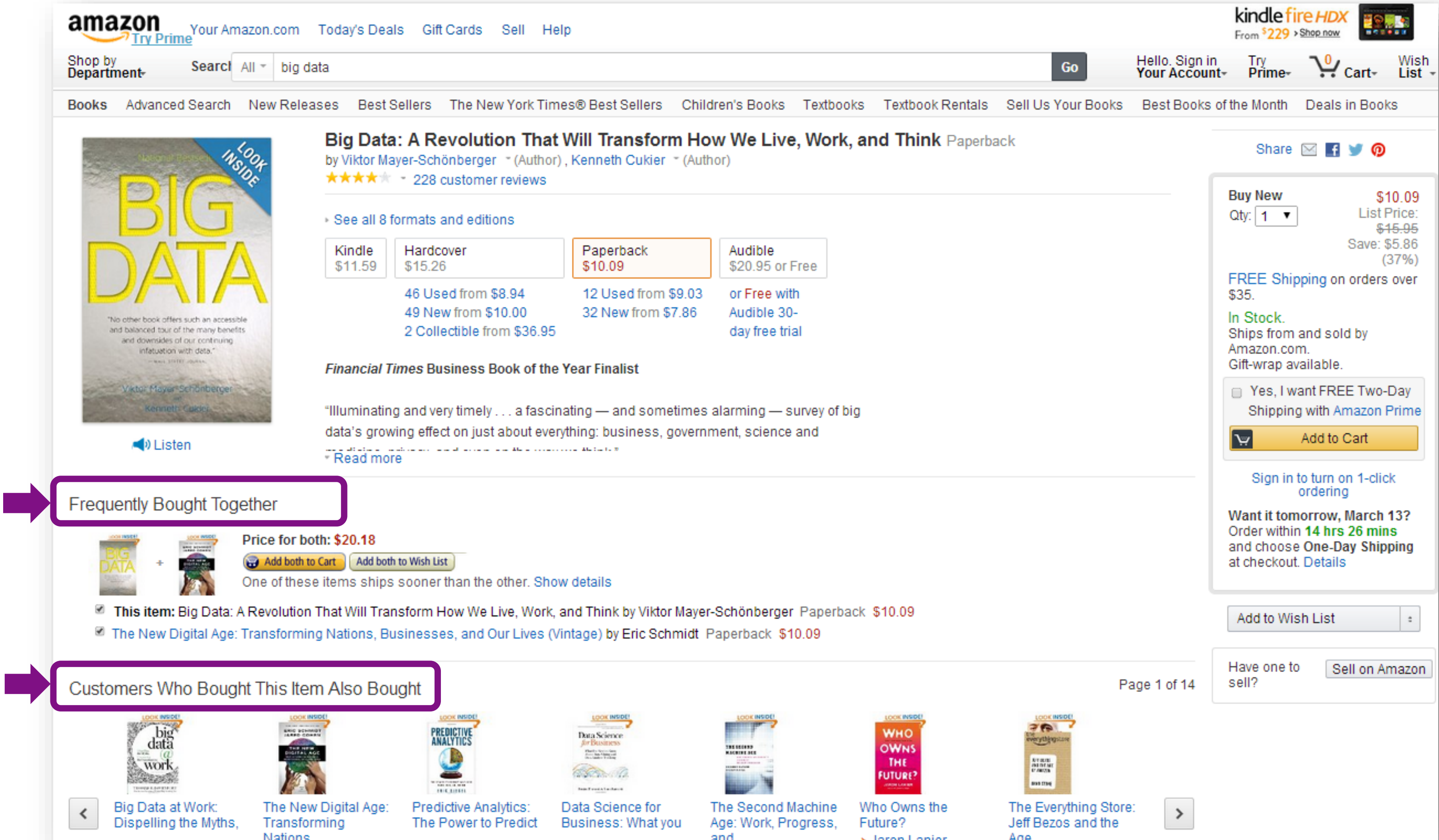
ID	Cause
368	{V157=high,V456=low},{V157=high,V184=low},{V157=high,V320=low},{V356=normal-high,V386=high},{V2=low,V430=high
370	{V355=low,V586=high},{V317=low,V568=low},{V217=low,V586=high},{V355=low,V584=high}
390	{V124=low,V561=very-high}



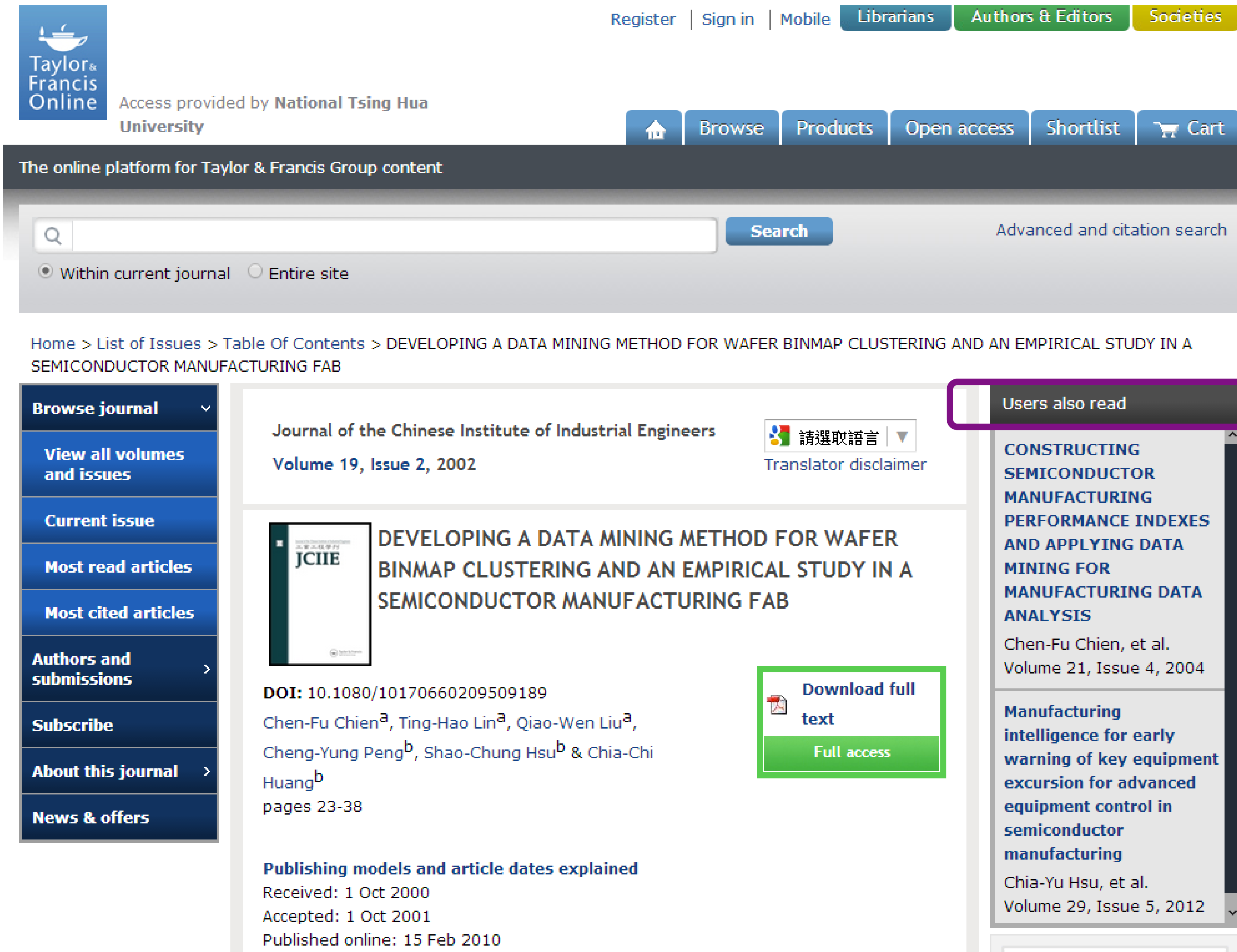
Kao, H., Hsieh, Y., Chen, C., & Lee, J. (2017). Quality prediction modeling for multistage manufacturing based on classification and association rule mining. MATEC Web of Conferences 123, 00029.

應用實例：Amazon

可以應用於推薦系統，but"缺點"需要被關心。



應用實例：相關期刊論文推薦



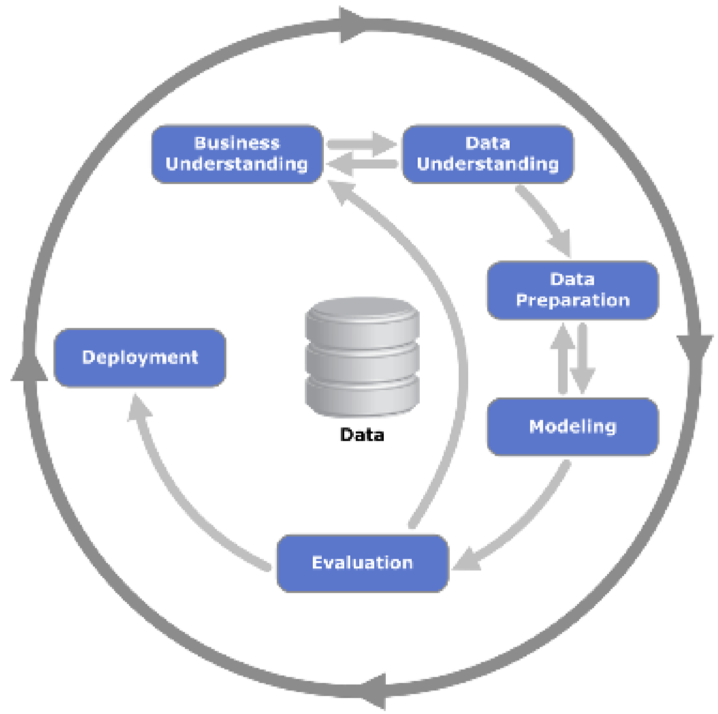
總結

- 顧客的消費行為會隨著時間而改變，所以需不斷地重新挖礦以更新資料庫並週期性地執行關聯規則運算，以萃取出最新的關聯規則來洞悉顧客消費型態
- 在產生關聯規則的程序中，會有許多重複或不重要的關聯規則，如何制定合適之支持度、信賴度與增益值門檻亦為議題
- 關聯規則的優點
  - 計算模式簡單易懂、能產生簡單明瞭的結論
    - 應用關連規則，以「如果買了，則也會購買」為模式
  - 能運用在非監督式資料採礦上
  - 適用不同形式的原始資料
- 關聯規則的缺點
  - 當商品數量增加，必須進行的運算會成幾何級數增加
  - 對於資料的個別資訊不甚重視
  - 難以決定適當的商品組合 means: 得出的規則，會難以常理解釋?
  - 容易剔除罕見商品

Data Science Framework

CRISP-DM 標準流程

- CRoss-Industry Standard Process for Data Mining (CRISP)
  - Proposed by Integral Solutions Ltd (ISL), Teradata, Daimler AG, NCR Corporation and OHRA (1996)
  - Business understanding (商業理解)
  - Data understanding (資料理解)
  - Data preparation (資料預備)
  - Modeling (塑模)
  - Evaluation (評估)
  - Deployment (佈署)



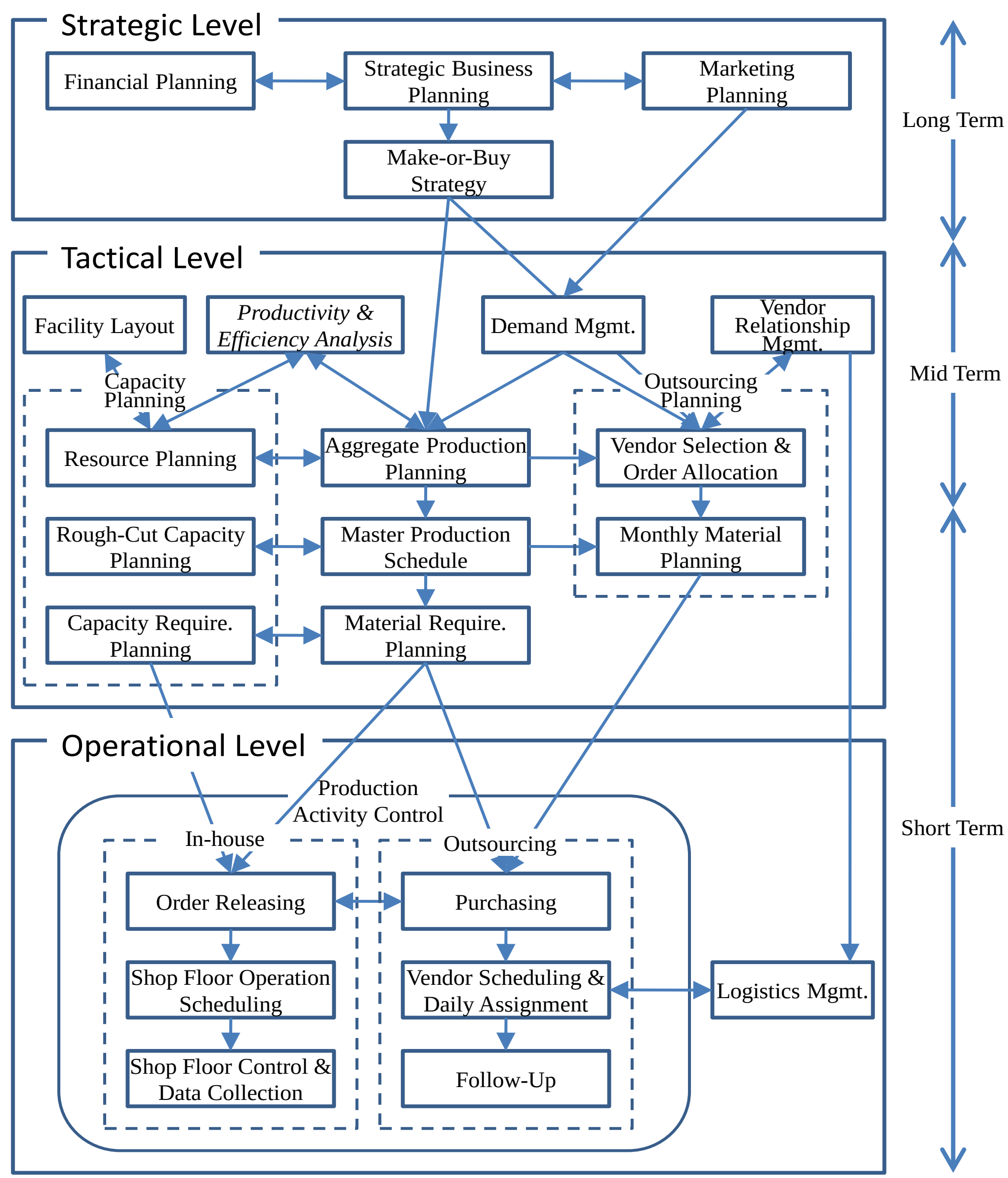
- Notes
  - Data is the focus (i.e. center)
  - The sequence of the phases is not strict and moving back and forth between different phases as it is always required.
  - The outer circle symbolizes the cyclic nature of data mining itself.

https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

資料探勘標準流程：商業理解

- 重點：瞭解探勘的方向與目標
  - 與企業的戰略、戰術、現場作業的環節息息相關。
- 問題在哪？什麼是問題？
- 客戶導向
  - 釐清客戶目標後，再規劃如何進行這樣的知識需求；規劃如何收集資料、分析資料及呈現目標的報告形式，也包含預算的規劃。

客戶導向



Lee and Johnson (2013)

資料探勘標準流程：資料理解

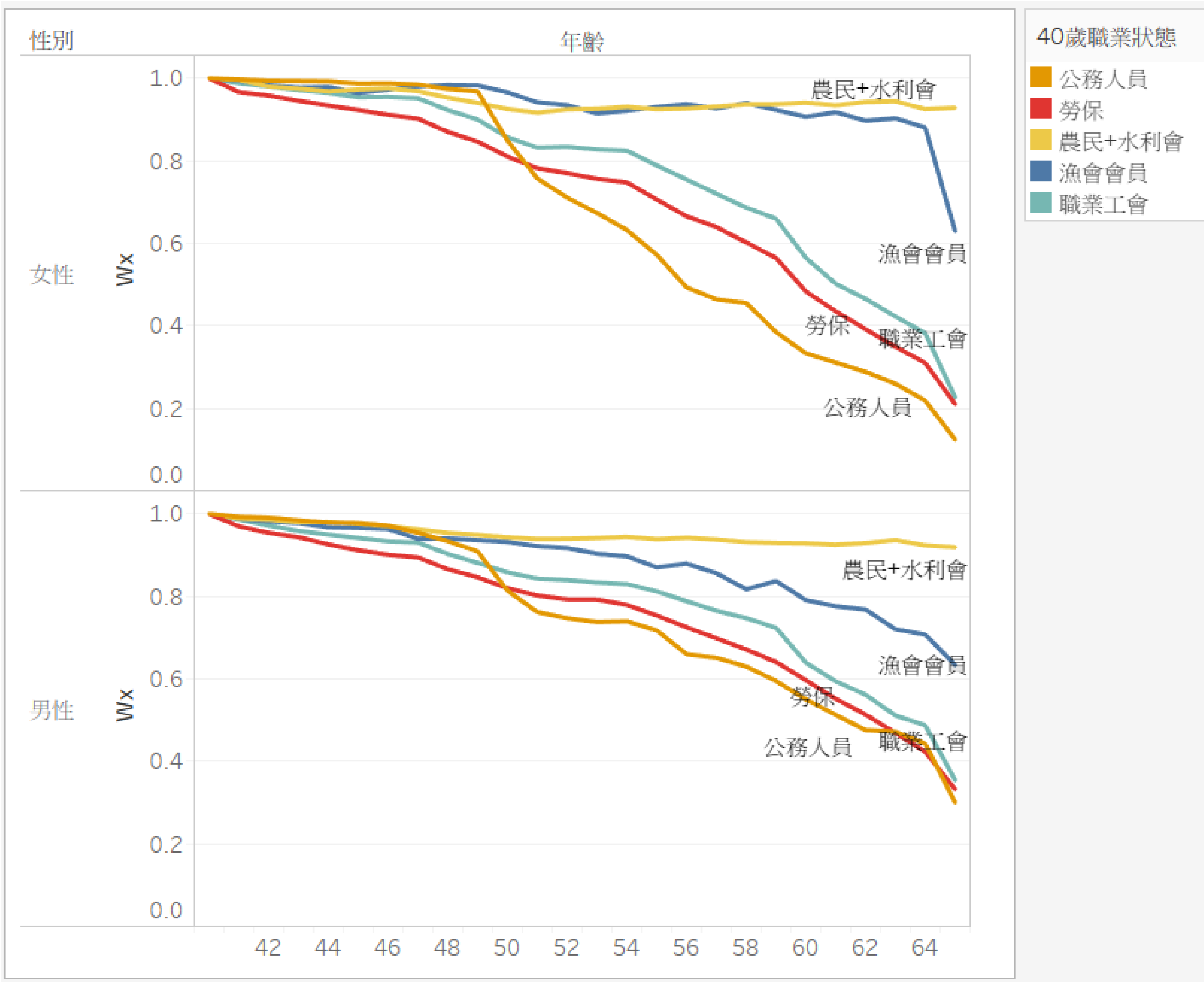
- 資料理解就是選取資料、資料異質、資料品質
  - 從資料庫與檔案中濾出相關的資料
    - 建立簡潔、明確的探勘工作內容描述，以利資料的收集
    - 選擇適當的獨立相關變數；變數間的獨立性，可降低變數內含資訊的重疊性
- 資料檢視：敘述統計與視覺化

資料品質

  - 資料數量
    - 檢視量化資料的三個維度：樣本個數(n)、變數/特徵個數(p)、異質性
    - 樣本個數太少會影響結果的解釋程度；當個數太多時，則統計上的顯著不見得有實質意義 (WHY?)
      - 資料的”多元性”、"異質性" 及"代表性“
      - 一個母體平均數的單尾檢定時，當樣本數佔母體總數的比例R=80%時，顯著水準0.05 要調整為0.0001. (馬瀰嘉，2019)
      - 樣本大到接近母體，不需要檢定，直接敘述統計視覺化
    - 變數個數太多會造成維度過高，使得分析時間過長→ 維度詛咒
  - 資料品質
    - 檢視資料的集中趨勢(平均數、中位數、眾數等)以及變異程度
    - 以不同圖表來檢視資料遺漏、資料雜訊、離群值等

馬瀰嘉(2019)，巨量資料分析下如何調整顯著水準，智慧科技與應用統計學報，16(2)，19-36.

40歲後各行業工作率 (2010-2013)



詹晨偉 (2019)

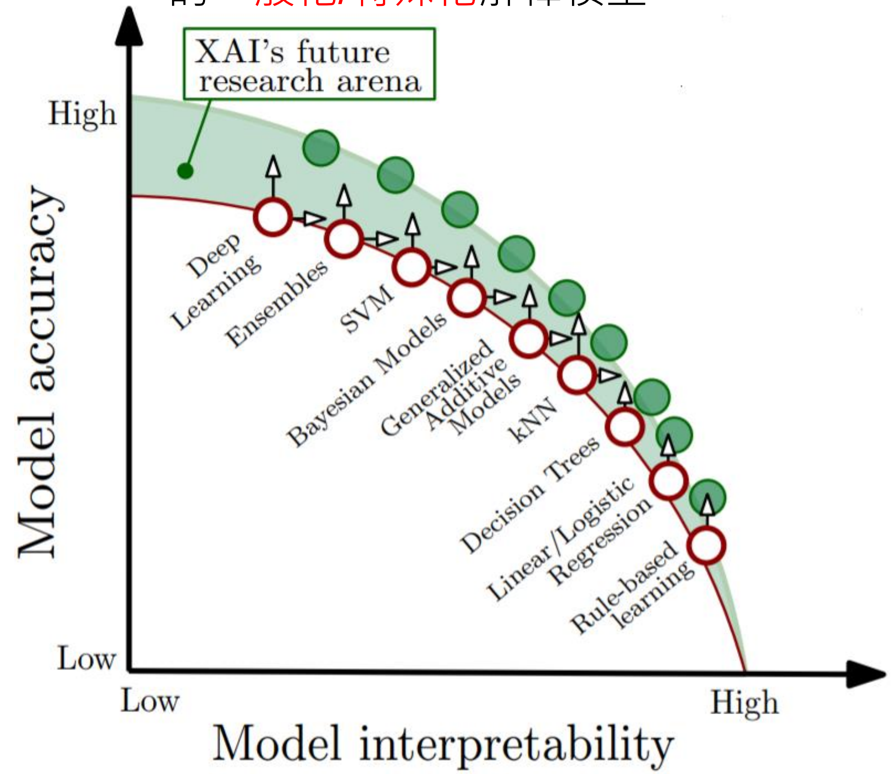
Supervised Learning Models

- Regression, Logistic regression, Partial Least Squares (PLS), Multivariate Adaptive Regression Splines (MARS), SVM/SVR, Decision Tree, Random Forest, LightGBM, Deep learning (DNN, CNN, RNN, graphCNN), GAN…
- Time Series: ARIMA, SARIMAX, LSTM, convolutional LSTM, GRU, …
- Accuracy Enhancement: Ensemble, Stacking, Attention, Transformer…

Unsupervised Learning Models

- Clustering: Hierarchical(Ward's Method), non-hierarchical(k-means, k-medoids), density-based (DBSCAN), Spectral Clustering, SOM, ART, EM-GMM…
- Dimension Reduction: PCA, ICA, t-SNE…

數學建模用以產生適合各種不同情境的一般化/特殊化解釋模型



Arrieta et al. (2019)

Issues

- Curse of Dimensionality, Overfitting vs. Underfitting, Accuracy vs. Interpretation, Data Imbalance

Arrieta et al., "Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI," https://arxiv.org/pdf/1910.10045.pdf.

資料探勘標準流程：評估

- Prediction
  - P-value, F statistic
  - AIC, BIC, WIC
  - R-squared, MSE, RMSE, MAE, MAPE, sMAPE, RMSLE, cross entropy
- Classification
  - Accuracy, Sensitivity(recall), Specificity, Precision, Miss rate (miss, type II error), fall-out (false alarm, type I error), F1-score, MAP
  - DCG(discounted cumulative gain), NDCG, P@k (precision at k)
- Cross Validation
  - Data split: training dataset (eg. 80%) and testing dataset (eg. 20%)
  - K-fold CV, Random-Sampling CV (Holdout), stratified sampling CV, Leave-one-out CV, Time Series Nested CV, Group CV…

https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval)

1. 適合用於萃取If-Then Rules/Features
2. 適合用於Ensemble的弱分類器

TABLE 10.1. *Some characteristics of different learning methods. Key:* ▲ *= good,* ◆ *=fair, and* ▼ *=poor.*

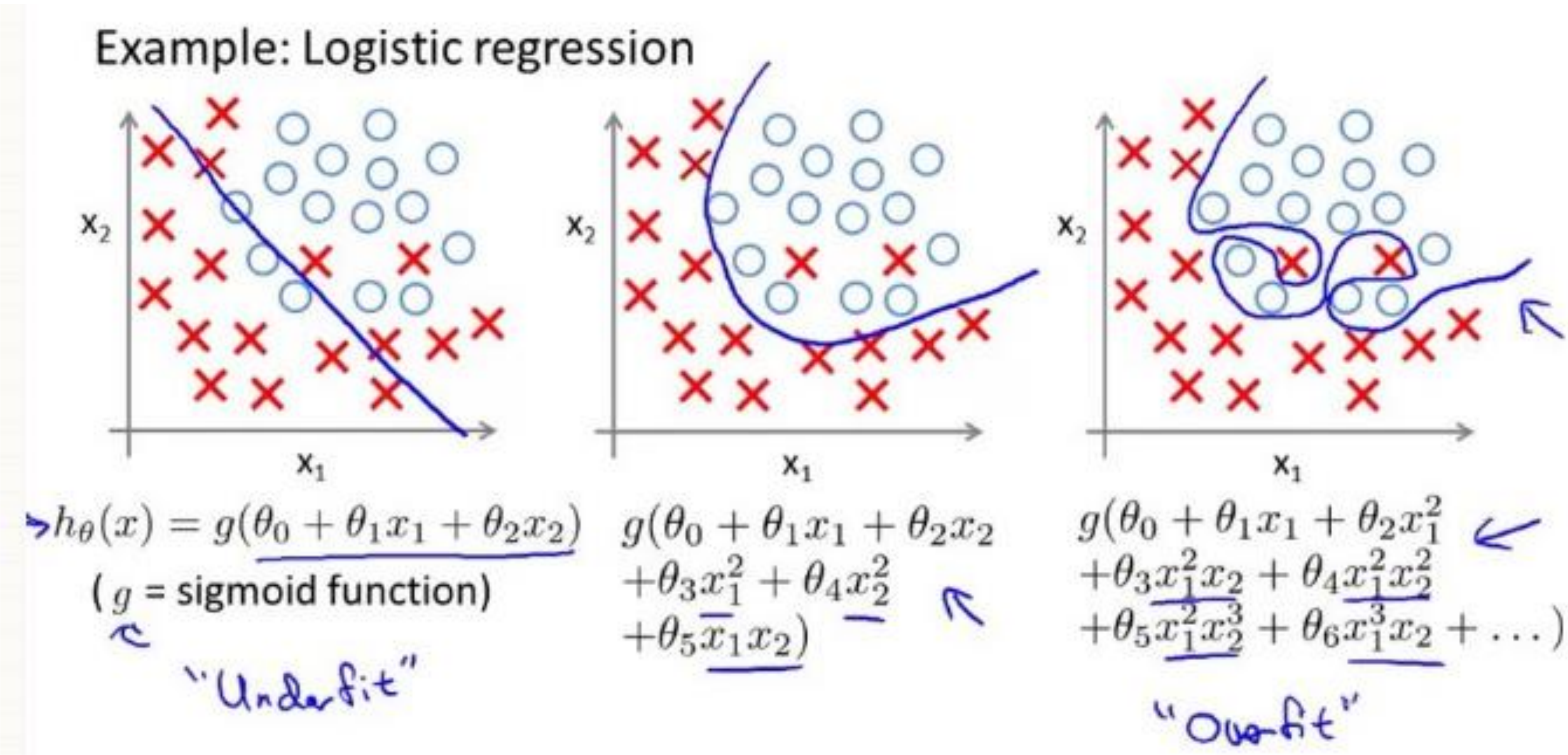
(revised from Hastie et al., 2009)

Characteristic	Neural Nets	SVM	Trees	MARS	k-NN, Kernels	ARIMA	GBM
Natural handling of data of "mixed" type	▼	▼	▲	▲	▼	▼	▲
Handling of missing values	▼	▼	▲	▲	▲	▼	▲
Robustness to outliers in input space	▼	▼	▲	▼	▲	▼	▲
Insensitive to monotone transformations of inputs	▼	▼	▲	▼	▼	▼	▲
Computational scalability (large N)	▼	▼	▲	▲	▼	◆	▲
Ability to deal with irrelevant inputs	▼	▼	▲	▲	▼	▼	▲
Ability to extract linear combinations of features	▲	▲	▼	▼	◆	▼	◆
Interpretability	▼	▼	◆	▲	▼	▲	◆
Predictive power	▲	▲	▼	◆	▲	◆	▲

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2 ed.: Springer.

資料探勘標準流程：評估

- Underfit vs. Overfit



Andrew Ng's, https://zh-tw.coursera.org/learn/machine-learning#

資料探勘標準流程：佈署

- This phase focuses on Decision and Action
- Explain and interpret the results of data science to extract the potential knowledge in the dataset
- Identify the value through the visualization tool
  - Does the ML/DS model meet business objectives/KPIs?
  - Does model make sense? Is model actionable and feasible?
  - How to evaluate the decision risk after taking action?
- New knowledge should enhance the core competence for business competition
- Model and knowledge must be monitored & updated over time
  - domain adaptation and concept drift

Analytics Framework

Data Description
- Data Source & Goal
- Summary & Visualization
- Load Packages
- Problem Clarification

Data Preprocessing
- Redundant & Duplicate
- Missing Value Imputation
- Transform & Rescale
- Data Imbalance

Feature Selection
- Filter Methods
- Wrapper Methods
- Embedded Methods
- Dimension Reduction
- Physical Causality

Modeling & Validation
- Regression-based
- Tree-based
- Boosting
- Deep Learning
- Cross Validation
- Evaluation

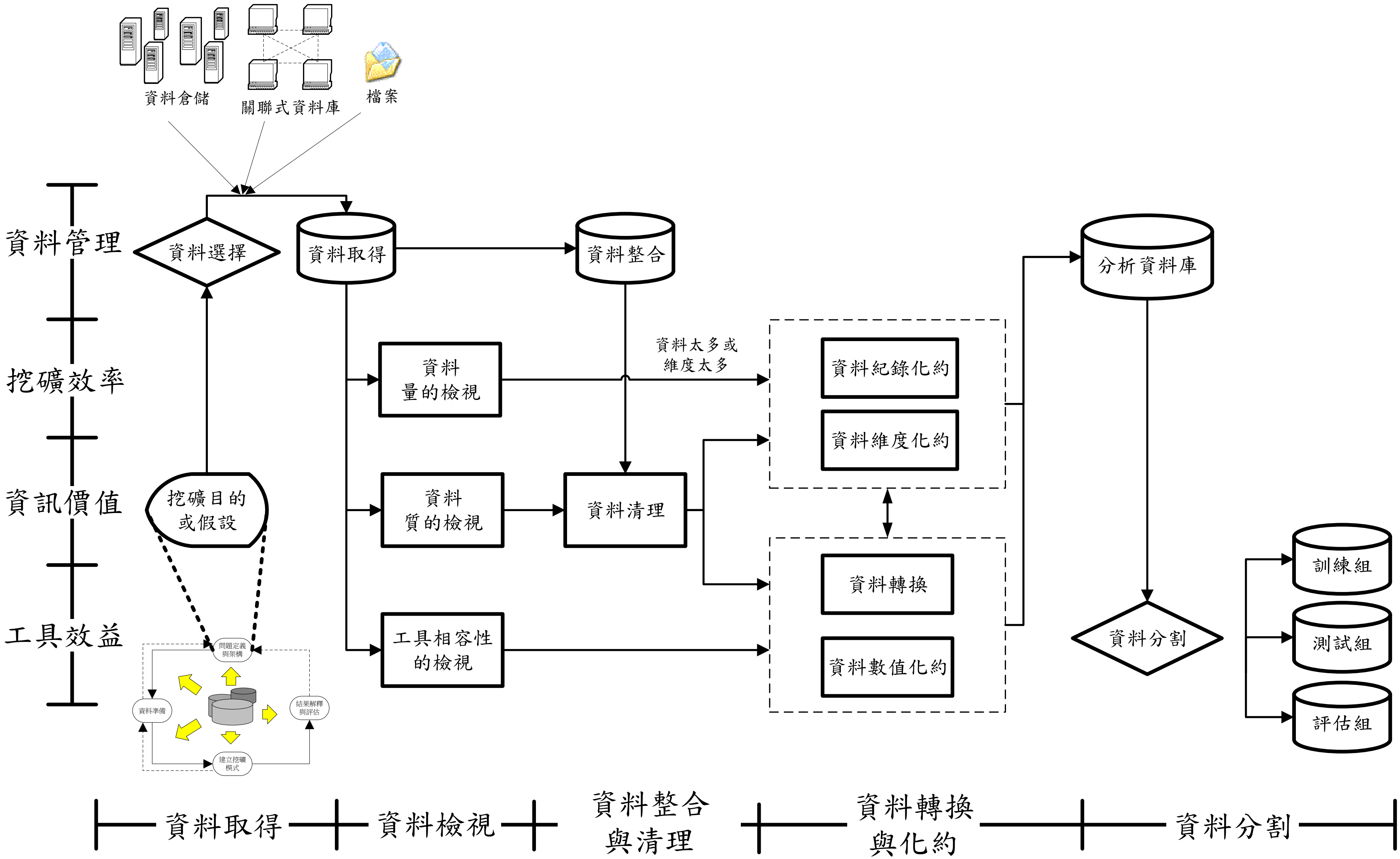
Visualization & Conclusion
- Visualization
- Interpretation
- Criticism
- Risk Assessment
- Value of Information
- Decision

以視覺化起，也以視覺化終

數據預處理

Data Preprocessing

Data Preprocessing Framework



簡禎富 and 許嘉裕 (2014)

Outline

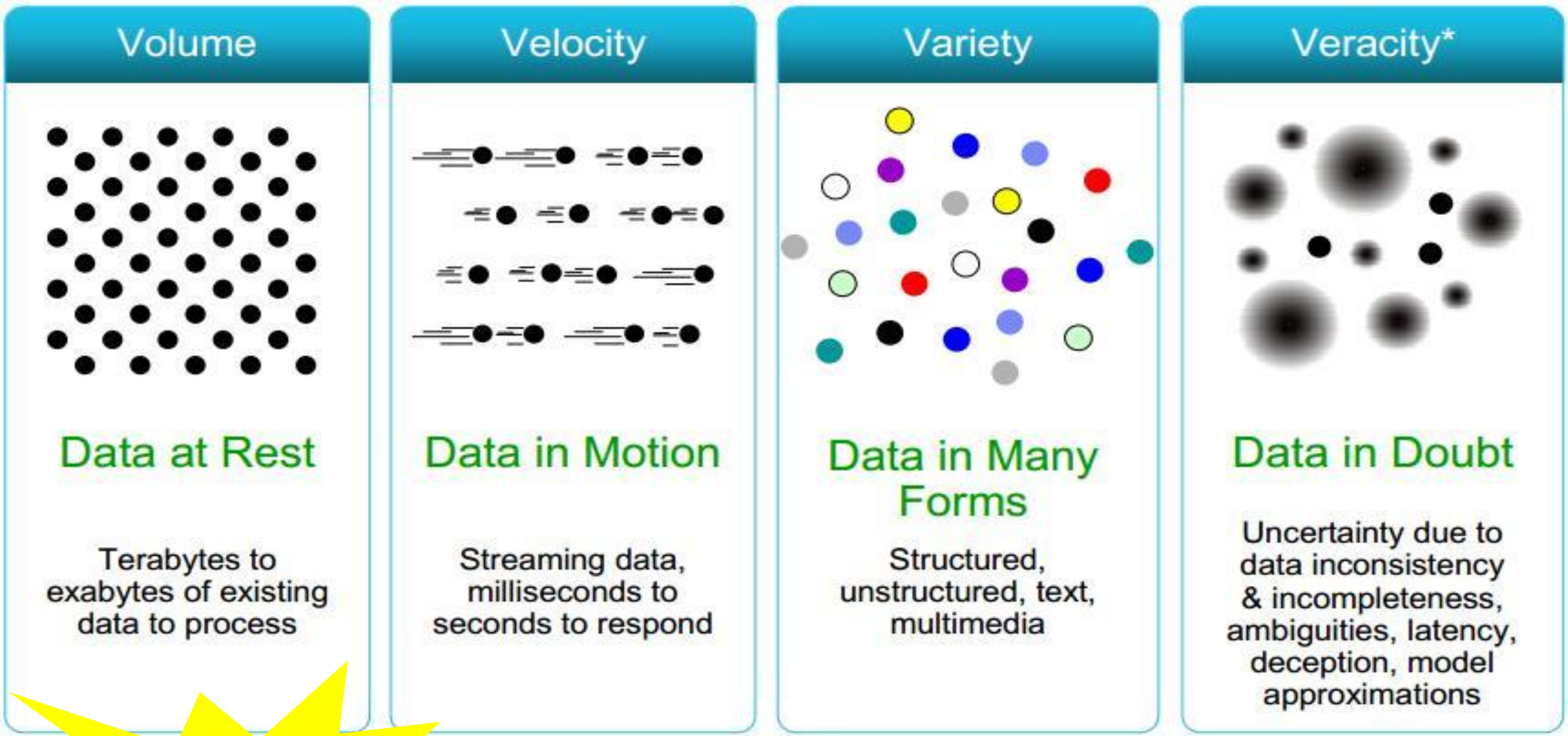
- Data Preprocessing Framework
  - Data Quality
  - Data Type and Scale
  - Data Integration and Cleaning
  - Data Transformation
  - Data Similarity and Dissimilarity
  - Data reduction and Partition
- Empirical Study: Data Science in Manufacturing
- 智慧製造與生產線上的資料科學
  - http://polab.im.ntu.edu.tw/Talk/Data_Science_in_Manufacturing.pdf

Materials mainly comes from

1. 李家岩，2017，智慧製造與生產線上的資料科學 Data Science in Manufacturing: From Predictive to Prescriptive，臺灣資料科學年會。
2. 簡禎富, 許嘉裕, 2014. 資料挖礦與大數據分析, 前程文化。

Big Data 4V

資料量龐大　資料變動速度快　資料多樣性　資料真實性



Value!

圖片來源：http://www.datasciencecentral.com/profiles/blogs/data-veracity

Why Data Preprocessing?

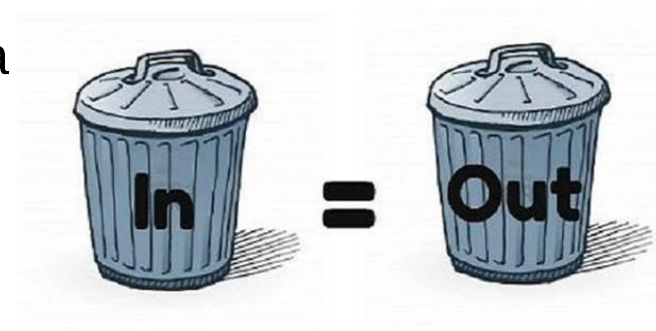
- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data

    職業=“ ”
  - noisy: containing errors or outliers

    薪資=“-10”
  - inconsistent: containing discrepancies (不一致) in codes or names

    過去分類 “1,2,3”, 現在分類 “A, B, C”

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration
  - Garbage-in Garbage-out



Data Description

Data Summary

```
##   Class              Time                             Feature1        Feature2
##  pass:1463   Min.   :2008-07-19 11:55:00   Min.   :2743   Min.   :2159
##  fail: 104   1st Qu.:2008-08-22 00:55:30   1st Qu.:2966   1st Qu.:2452
##              Median :2008-09-11 08:06:00   Median :3011   Median :2499
##              Mean   :2008-09-09 18:37:39   Mean   :3014   Mean   :2496
##              3rd Qu.:2008-09-29 11:33:00   3rd Qu.:3057   3rd Qu.:2539
##              Max.   :2008-10-17 06:07:00   Max.   :3356   Max.   :2846
##                                            NA's   :6      NA's   :7
##     Feature3        Feature4        Feature5            Feature6
##  Min.   :2061   Min.   :   0   Min.   :   0.6815   Min.   :100
##  1st Qu.:2181   1st Qu.:1082   1st Qu.:   1.0177   1st Qu.:100
##  Median :2201   Median :1285   Median :   1.3168   Median :100
##  Mean   :2201   Mean   :1396   Mean   :   4.1970   Mean   :100
##  3rd Qu.:2218   3rd Qu.:1591   3rd Qu.:   1.5257   3rd Qu.:100
##  Max.   :2315   Max.   :3715   Max.   :1114.5366   Max.   :100
##  NA's   :14     NA's   :14     NA's   :14          NA's   :14
```

Hung (2018). https://rpubs.com/jeff_datascience/Semiconductor_Manufacturing

Data Preprocessing

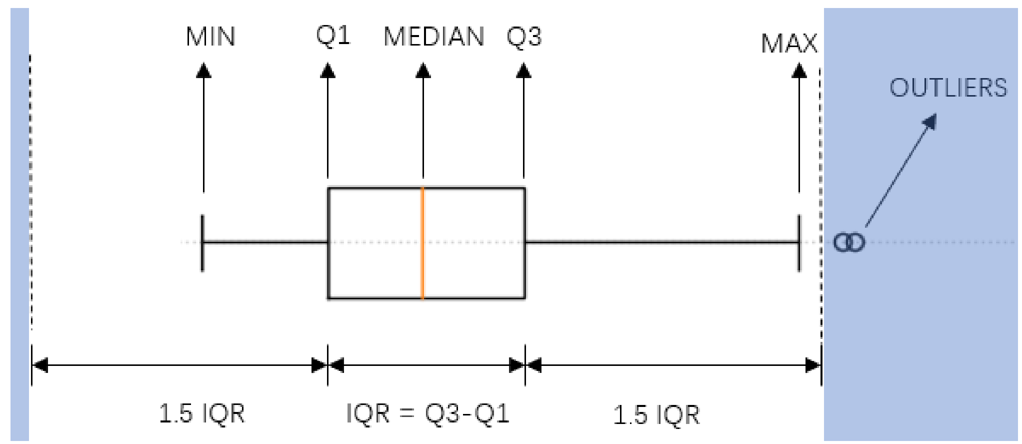
- Variable and Observation Summary
  - n and p: # of observations, # of variables
  - Statistics: mean, median, standard deviation, min, max, 1st quarter, 3rd quarter, etc.
  - Missing: # of NA

- Preliminary Variable Removal
  - Identical column
    - with identical values in this column
  - Duplicate column
    - the same data value but with different types
  - Redundant column
    - the value in this column can be derived trivially from another column… with high correlation
  - Highly-correlated variables (?)
    - Regression-based → multicollinearity
    - Machine learning-based → just keep it! (?)
  - Counter, material code, time (if which you don't care), …

Data Quality Investigation

- Single Column
- Multiple Columns
- Validity and Reliability (信度與效度)
- Metadata and Entity-Relation Model

- Single Column
  - Box plot (outlier)



- Note: all outlier should be removed? (it depends)
- Outlier could also provide information for improving the noise or sensitivity.

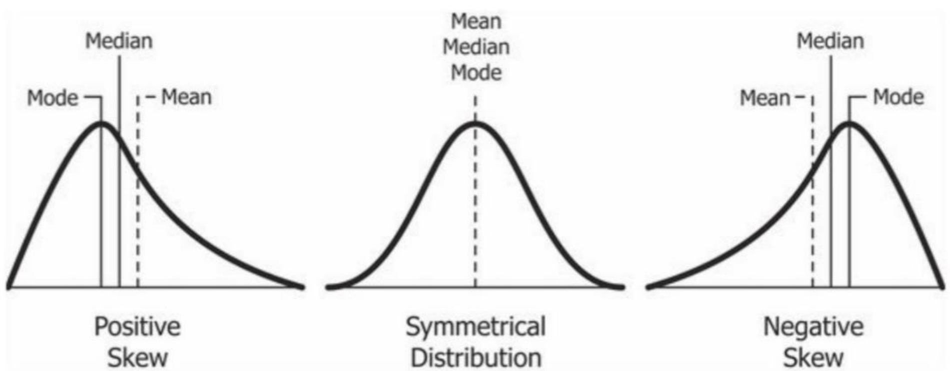
Outlier Detection

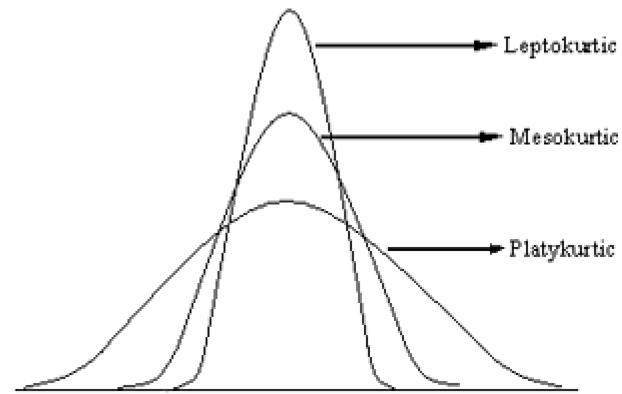
- Univariate
  - box plot, violin plot
- Multivariate (for observation)
  - Scatter plot, Correlation coefficient (kind of clustering)
  - Cook's distance
    - https://en.wikipedia.org/wiki/Cook%27s_distance
  - Clustering
    - DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
    - Gaussian mixture models (GMM) with expectation-maximization (EM) algorithm/ Bayesian Gaussian Mixture
- Outlier Treatment
  - Keep (due to having information content) or Remove
  - Transformation: log transformation, binning
  - Treat as missing value imputation (eg. KNN or impute “Other”)
  - Adding new column for marking outlier

https://www.mathworks.com/help/stats/cooks-distance.html
https://towardsdatascience.com/best-clustering-algorithms-for-anomaly-detection-d5b7412537c8

Data Quality Investigation

- Univariate Statistic
  - Statistics: mean(median, mode), variance, skewness, kurtosis
  - benchmarking with Engineer's intuition



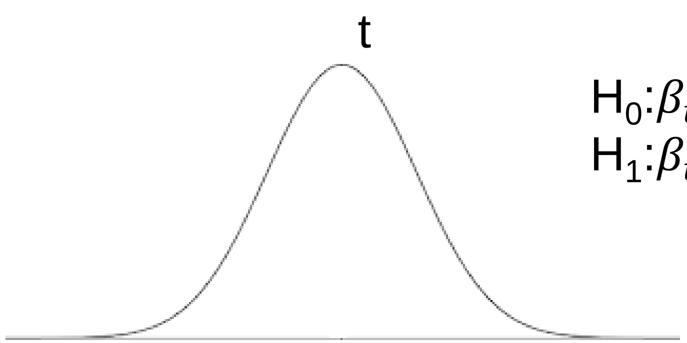


- Univariate Interpretation
  - Linear Regression (with response variable Y).
  - $y = \beta_0 + \beta_i x_i + \varepsilon$
  - If the estimate of the β_i violates the engineering experience (sign changed)
    - Eg. Etching time (x) and thickness (y) should be with negative β_i .
  - Then the variable may have quality issue.

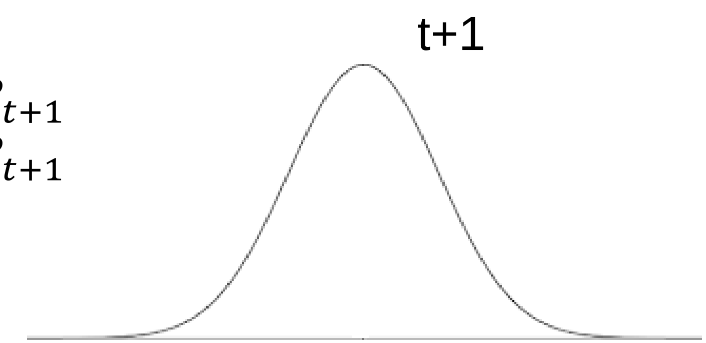
Data Preprocessing

Data Quality Investigation

- Multiple Column
  - Correlation or Covariance between two variables
    - Eg. Height and weight should be positive correlation.
    - Eg. Grades of MATH and SCIENCE should be positive correlation.
    - Eg. Lithography: Exposure latitude (EL) versus depth of focus (DOF) should be negative.
- Validity and Reliability (信度與效度)
  - 敘述性統計大部分做的是checking within distribution
  - 也可根據不同時間點對同一母體收集資料，做between distribution checking



$$H_0{:}\beta_t = \beta_{t+1}$$
$$H_1{:}\beta_t \neq \beta_{t+1}$$

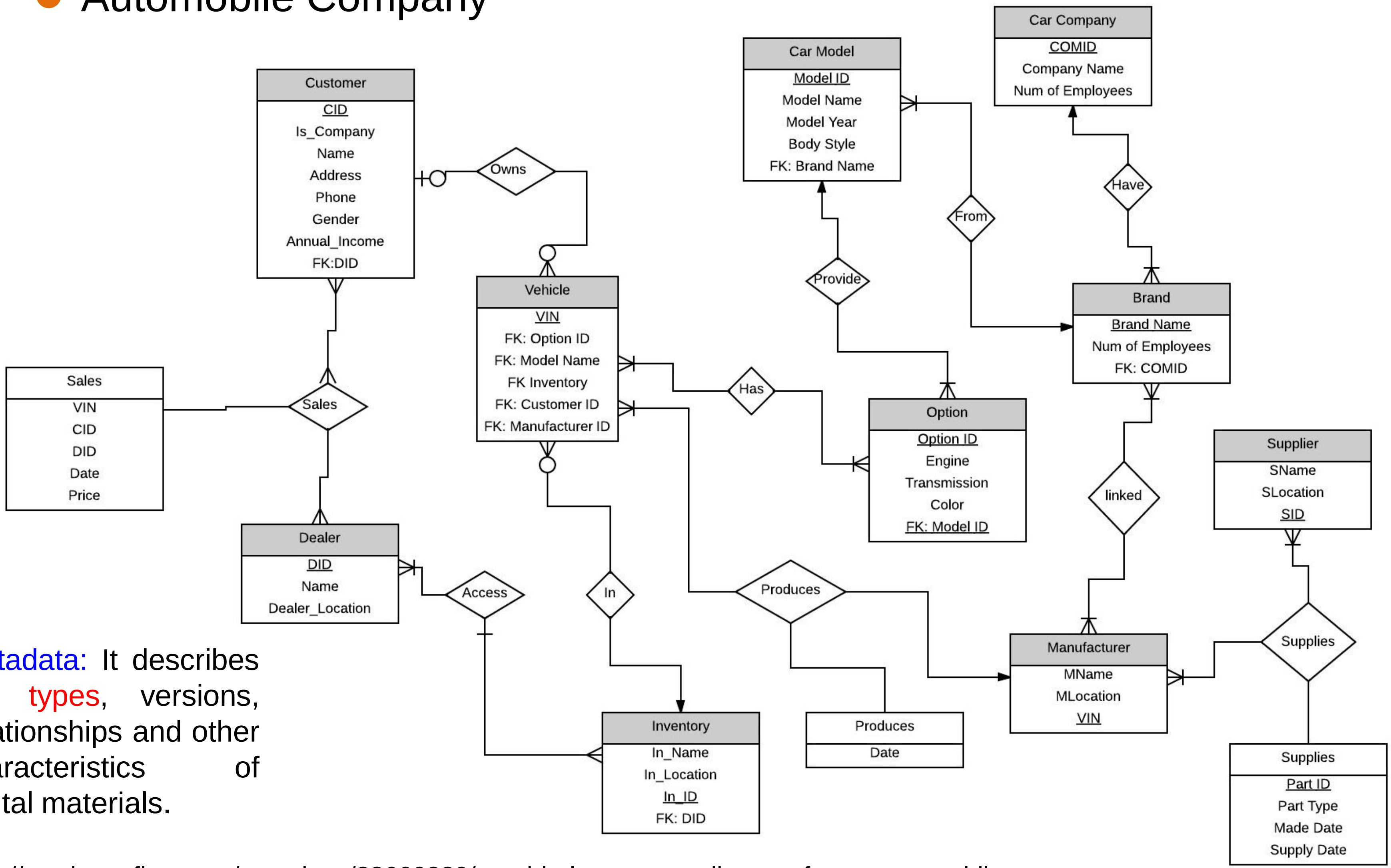


$$y = \beta_0 + \beta_t x + \varepsilon$$

$$y = \beta_0 + \beta_{t+1} x + \varepsilon$$

Data Quality Investigation- Entity-Relationship (ER) Model

- Automobile Company



Metadata: It describes the types, versions, relationships and other characteristics of digital materials.

https://stackoverflow.com/questions/23660839/need-help-on-an-er-diagram-for-an-automobile-company

資料尺度 (Scale)

衡量的層次	內容說明
名目尺度 (nominal Scale) (=,≠)	衡量的數字僅是作為代碼，數字大小不具任何意義，也不能做數學運算。範例：員工編號、眼睛顏色、郵遞區號、機台編號、貨批編號等
類別尺度 (categorical Scale)	衡量的數字僅是用來表示歸屬的類別，因此類別尺度的資料可以重複。範例：先對縣市編碼，歸類成北、中、南、東地區。
順序尺度 (ordinal Scale) (<,>)	衡量的數字表示方案之間的大小順序關係。範例：成績排名、金屬硬度
區間尺度 (interval Scale) (+,−)	衡量的數字可有意義地描述並比較數字之間的差距大小。無固定原點，也可以調整分隔的間距大小。範例：日期、華氏或攝氏溫度、機台的溫度、量測的參數、學業成績
比率尺度 (ratio Scale) (×,÷)	衡量的數字可做比率倍數的比較。有固定原點 範例：溫度、電子現金、化學藥劑使用量、重量
絕對尺度 (absolute scale)	所衡量的數字具有絕對的意義，無法再做其他有意義的轉換。範例：機率、自然數

(簡禎富、許嘉裕，2014)

資料探勘標準流程：資料預處理

問題	原因	步驟
不正確的資料	資料的值超出合理範圍	
不一致的資料	不同來源資料整合後所出現的分歧 數值不一致、資料內容不一致、欄位不一致	
重複的資料 (Duplication)	重複記錄的欄位或數值 (data type: single, double…) (同樣的資料卻不同的寫法，“做了36顆”, “打出36粒”, “生產36個”，”左上角區塊有產生defects”, “defects發現於左上方 區域”)	**資料整合**
冗餘的資料 (Redundant)	出現相同意義的資料或欄位 具有相同意義或彼此間存有已知數學關係的欄位，此變數的屬性或意義可由另一變數推導而得 (有些冗餘資料可以經由相關分析偵測到) eg. 地址vs.地區	
遺漏值	量測設備或人為因素所造成的資料遺漏	
雜訊	資料本身的誤差或資料輸入的偏差	**資料清理**
離群值	資料本身的特性、不當量測或資料輸入錯誤	
資料尺度不適	資料格式不符合挖礦工具的假設 將不同尺度或單位的資料轉換成有一致的數值尺度，或類別資料與連續數值資料間轉換。	**資料轉換 (正規化)**
資料太多	資料或維度過高	**特徵篩選**

(簡禎富、許嘉裕，2014)

遺漏值填補是...補「資料」?

	English	Math
Student_A	80	76
Student_B	80	91
Student_C	80	83
Student_D	80	62
Student_E	80	?
Avg.	80	

Max: 91
Min: 62
Avg: 78

迷思: 補遺漏值?

- 填補遺漏值一般會造成部分失真或偏差
- 使用者應根據"資料特性"以及"分析目的"，來決定填補遺漏值的方法，以避免忽略原本應有的資訊
- 方法(補值一定要找關係!!!!)
  - 忽略變數值 (“N/A” 與 “0” 是不一樣的!!)
  - 移除觀測值 (remove the tuple)：當依變數Y遺漏時
  - 人工填寫遺失值
  - 使用一個全域常數填充遺漏值 (eg. N/A)
  - 使用屬性平均值
  - 使用與給定變數值屬於同一類別的所有樣本之平均值
  - 模型: 簡單/多元線性迴歸、類神經網路、最鄰近估計法K-Nearest Neighbor (KNN)、Random Forest…”MICE”..

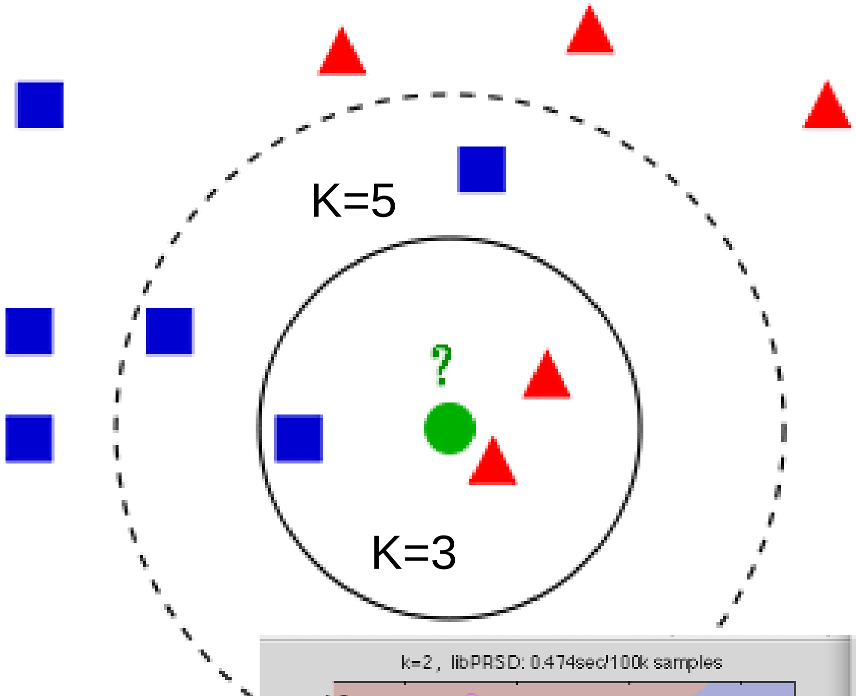
最鄰近估計法 (Nearest Neighbor Estimator)

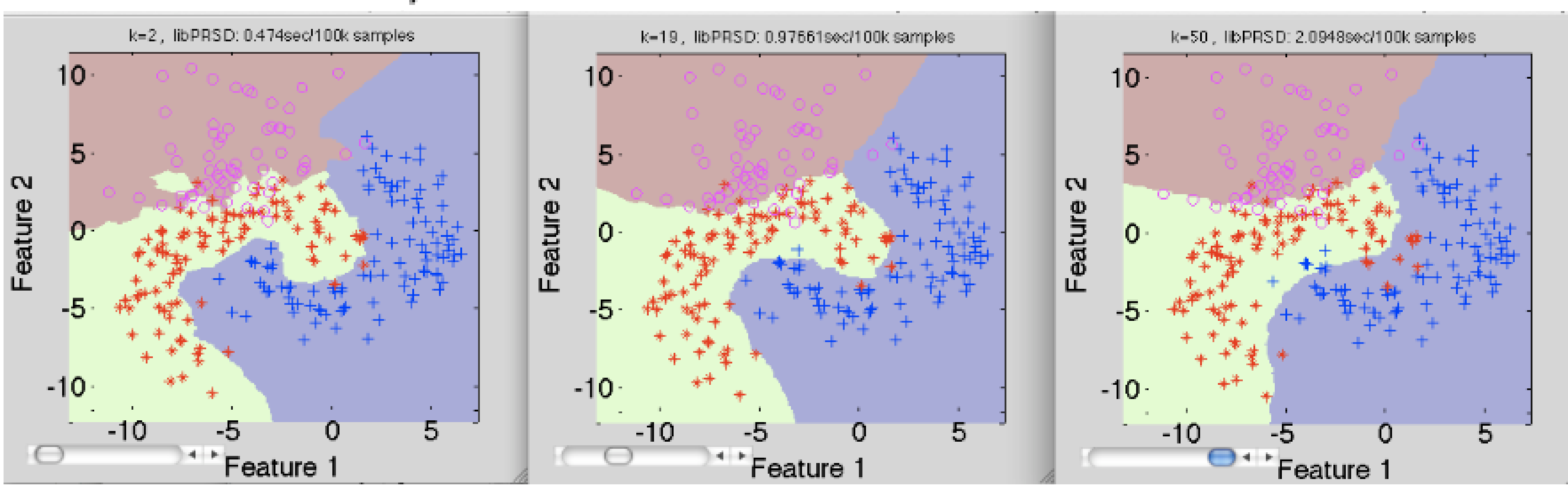
- 補值一定要找關係!!!!
  - 利用其他變數與遺漏值之間的關係來估計遺漏值
  - 補值：可利用其他變數與遺漏值之間的關係來估計遺漏值
  - 例如，若「收入水準」變數發生遺漏值，或許可能用「房子坪數」這變數來做預測

- 假設在現有的資料庫中發現某一顧客其購買反應的態度為一遺漏值

顧客	性別	年齡	薪水	購買反應
A	女	27	$19,000	No
B	男	51	$64,000	Yes
C	男	52	$105,000	Yes
D	女	33	$55,000	Yes
E	男	45	$45,000	No
F	女	45	$100,000	?

最鄰近估計法K-Nearest Neighbor (KNN)





Wikipedia, https://zh.wikipedia.org/wiki/%E6%9C%80%E8%BF%91%E9%84%B0%E5%B1%85%E6%B3%95

perClass, 2017. kb16: Visualize the effect of a change of parameters in a trained classifier. http://perclass.com/doc/kb/16.html

迷思: 補遺漏值?

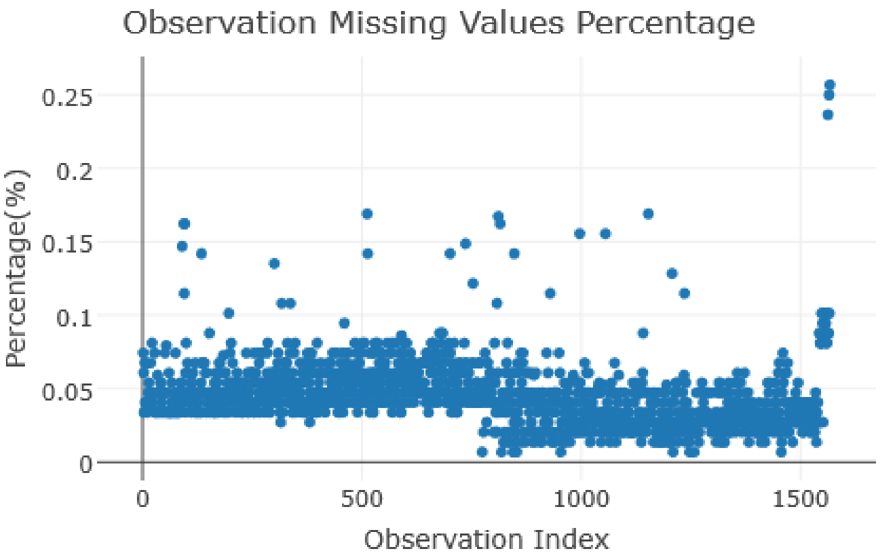
不偏估計量 vs.變異程度

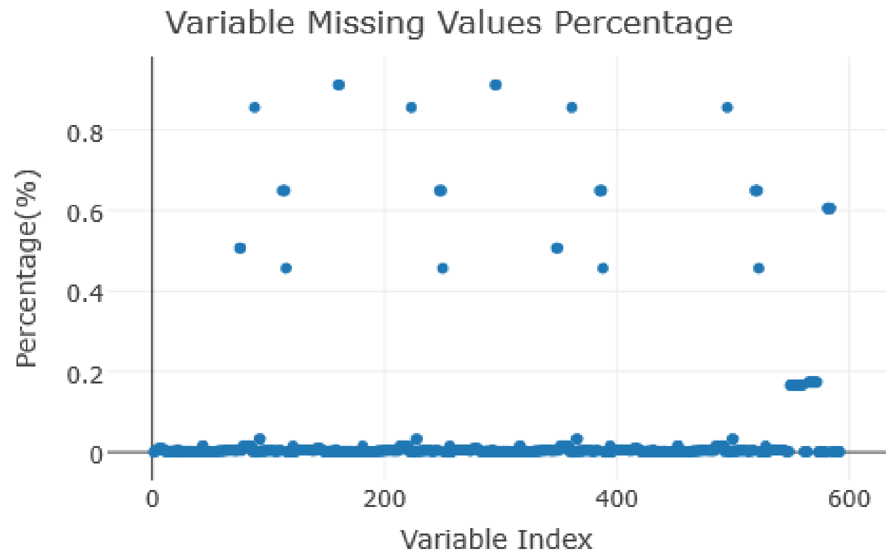
觀測值	原始資料值	第 11 筆遺漏	利用平均數估計	利用標準差估計
1	0.0886	0.0886		
2	0.0684	0.0684		
3	0.3515	0.3515		
4	0.9874	0.9874		
5	0.4713	0.4713		
6	0.6115	0.6115		
7	0.2573	0.2573		
8	0.2914	0.2914		
9	0.1662	0.1662		
10	0.44	0.44		
11	0.6939	?		
平均值	0.4023	0.3731		
標準差	0.2785	0.2753		
誤差值				

(簡禎富、許嘉裕，2014)

Data Preprocessing

Missing Value Imputation





- Remove the column or row with too many missing values (>40%)
- Remove the row without label/target value/Y (or unsupervised learning)
- Impute: Mean and Median (by the same category), Mode, "Others", "N/A", "NaN"
- Impute by Model: K nearest neighbor (KNN), Multivariate Imputation by Chained Equations (MICE), Inverse Distance Weighting (IDW), etc.
- Ignore: some algorithms can handle the missing values, eg. LightGBM and XGBoost (ignore is different from imputing "NA" or "0").

Jeff Hung (2018), Data Analysis in Semiconductor Manufacturing. https://rpubs.com/jeff_datascience/Semiconductor_Manufacturing

Duplicate Entries Detection

- Duplicate/Redundant Entries Detection
  - means that the observations having the same value of features show different target value. (同樣一組x有不同的y 怎辦?)
  - Two rows with the same feature values but their labels are different.

- Treatment
  - Remove the “old/out-of-date” one (from time aspect)
  - Cause: some observation having outlier y (有些觀測值是outlier在y)
    - For binary class in y, in the same x we can find the majority class in y and remove the minority class. (找到這組x在y上表現較多majority的反應，把另一反應較少的刪除)
  - Need one more additional/new variable to discriminate these two
    - Eg. Furnace with the same lot QC, adding Furnace_location (low, median, high) or wafer_location (1 to 25)
    - Use feature engineering to generate new feature
  - Cause: smaller noise
    - Consider the stochastic model with noise (eg. regression) when prediction/classifier modelling (eg. noise caused by particles)
  - Cause: larger noise
    - Denoise: use moving average to smooth the noise

Data Transformation and Scaling

- Feature Scaling (轉換原始資料或重新編碼以提升資料價值)
  - Smoothing: eliminate noise from data (eg. moving average)
  - Aggregation: summarization, data cube construction
    - 例：以日銷售量，計算月和年銷售量。
  - Generalization: concept hierarchy climbing
    - 以高層級概念代替低層級。例 “street”以 “city”或 “country”代替
  - 資料數值的轉換
    - 正規化 (Normalization): min-max

$$X_{nom} = \frac{X - X_{min}}{X_{max} - X_{min}} \in [0\,,1]$$

      - MinMaxScaler transforms to a specific range (often [0, 1]). Note that zero values will probably be transformed to non-zero values.
    - 標準化 (Standardization): Z-score

$$Z = \frac{X - \bar{X}}{s}$$

  - 資料類別的轉換
    - 離散型資料轉成連續型資料
      - 例如：學生成績的等第為A對應至92分，若為B+，則應該對應至88分
    - 連續型資料轉成離散型資料
  - Tree-based method and a typical linear regression doesn’t need scaling because of not sensitive.

Data Transformation and Scaling

連續型資料轉成離散型資料

- Binarization 二值化
  - If x is larger than a threshold, then transform x to 1; otherwise 0.
    - For categorical color, if you don't care what color is, then you can use 1 representing color; 0 for no color (transparent or white).
- Binning and Bucketization 裝箱 (for discretization 離散化)
  - Bucketizer transforms a column of continuous features to a column of feature buckets, where the buckets are specified by users.
    - Eg. age: [0-19] for 1; [20-39] for 2; [40-59] for 3; [60 above] for 4.
    - For categorical variable, you may use Group By or Select Count(), and then transform x to "Other" if the frequency less than a threshold.
- Question: age: [0-19] for 1; [20-39] for 2; [40-59] for 3; [60 above] for 4.
  - Is this a good method with Likert scale? Misleading? (WHY?)

類別!! 不可+-x÷

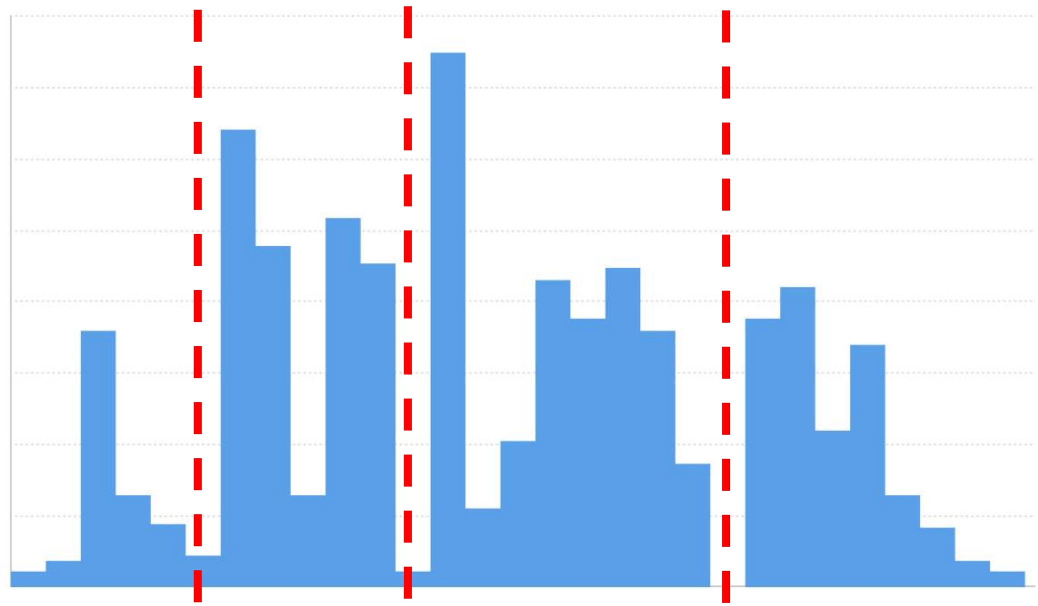
Half-day	1	$100
Day	2	$200
Half-week	3	$500
Week	4	$1000
Half-month	5	$2000
Month	6	$4000

為啥"月"是"半天"的40倍?需解釋!

Data Transformation and Scaling

- 連續型資料轉成離散型資料
  - 離散化(discretization): 將連續資料分配到數個小區間
  - You may use distribution for binning, eg. quantization or quantile binning.

- Draw distribution and then binning



https://antv-2018.alipay.com/zh-cn/vis/chart/histogram.html

Data Transformation and Scaling

- 離散型資料轉成連續型資料
- Integer/Label Encoding
  - Mapping the category to number. Ordinal can be mapped to 0, 1, 2, 3.
  - Frequency of category: most frequent is 0, and then 1, 2, 3 in turn. (**StringIndexer**) (eg. quantile binning and then stringindexer)
  - Note: if the categorical variable shows no ordinal scale (eg. nominal), then the transformed numbers (0, 1, 2, 3) are not comparable.
  - You may consider dummy/binary variable transformation (i.e. **One-hot Encoding (OHE)**)

Pet
Cat
Dog
Turtle
Fish
Cat

Cat	Dog	Turtle	Fish
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	0	0	0

https://towardsdatascience.com/stop-one-hot-encoding-your-categorical-variables-bbb0fba89809

Feature Transformation

- One-hot Encoding (OHE)
  - If one categorical variable is with L levels, then we can generate M binary variables (i.e. a binary vector with L elements).
  - You may use dummy coding to generate only L-1 variables.
  - Suggestion
    - Ordinal scale → Integer encoding (eg. age, morning/afternoon/night)
    - Nominal scale → OHE (eg. color, city, machine ID, recipe, occupation)
  - Disadvantage: curse of dimensionality; break the relationship in the group
  - Treatment
    - Grouping: product → product group; tool → tool group
    - Specify time interval for data collection
    - Delete the level which only appears once, because of no reproducibility
  - Note: if you use logistic regression, OHE is good. Otherwise you may use feature hashing or bin counting. In addition, if you use gradient boosting tree (GBDT), you can directly use categorical or nominal value.

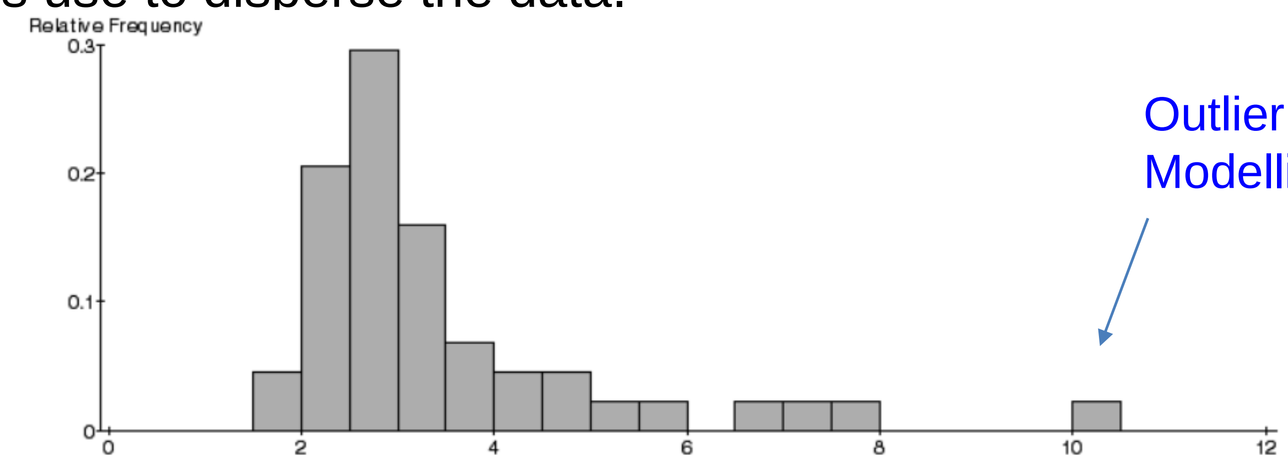
Data Transformation and Scaling

Rounding

- Replacing a number with a different number at the specified number of digits that is approximately equal to the original.
- Even you can round(value * m), round(log(value)), or round the value as categorical feature.

Log/SquareRoot/Exp Transformation (Log for long tail)

- log(x) slowly increase over x, that is, log can compress a large number and expand a small number. Eg. Log([100,1000],10)=[2,3]. Also, we may use log(1 + x) or log(x / (1 - x)). Similarly to square root or cube root.
- Exp is use to disperse the data.



Outlier → remove?
Modelling outlier?

Similarity vs. Dissimilarity (相似度與不相似度)

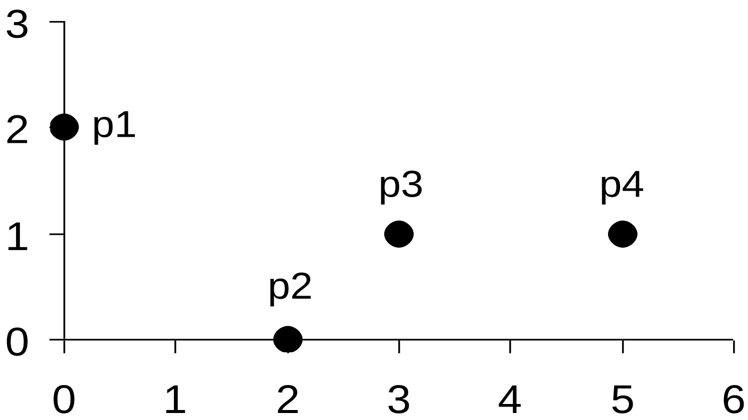
- 比較觀測值間(observation)或屬性間(attribute)的相似與不相似
  - Redundant Data Identification; Clustering Analysis
  - 相似度(其值基本上會介於0－1之間)
    - 相似度表示物件間相同的程度 。物件之間的相似度愈高，其物件愈相像
  - 不相似度
    - 不相似度表示兩個物件間差異的程度
    - 不相似度和距離其實是同義字，距離愈大，不相似度愈高
  - 計算方法
    - 兩個物件o_i與o_j 各有一個屬性，d(o_i,o_j) 與 s(o_i,o_j) 表示不相似度及相似度

屬性型態	不相似度	相似度
名目	$d = \begin{cases} 0, 若 o_i = o_j \\ 1, 若 o_i \neq o_j \end{cases}$	$s = \begin{cases} 1, 若 o_i = o_j \\ 0, 若 o_i \neq o_j \end{cases}$
順序	$d = \lvert o_i - o_j \rvert$/(n-1) (數值映射至0~n-1之間)	$s = 1 - d$
區間或比例	$d = \lvert o_i - o_j \rvert$	$s = -d;\ s = \frac{1}{1+d};\ s = e^{-d};$ $s = 1 - \frac{d - d\_min}{d_{max} - d\_min}$

歐幾里德距離

其中n是指維度個數，而x_k 及y_k 分別表示x與y的第k個屬性

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$



四個二維樣本點

Point	x 軸	y 軸
p1	0	2
p2	2	0
p3	3	1
p4	5	1

x 與 y 座標軸上的四個點

	p1	p2	p3	p4
p1	0.0	2.8	3.2	5.1
p2	2.8	0.0	1.4	3.2
p3	3.2	1.4	0.0	2.0
p4	5.1	3.2	2.0	0.0

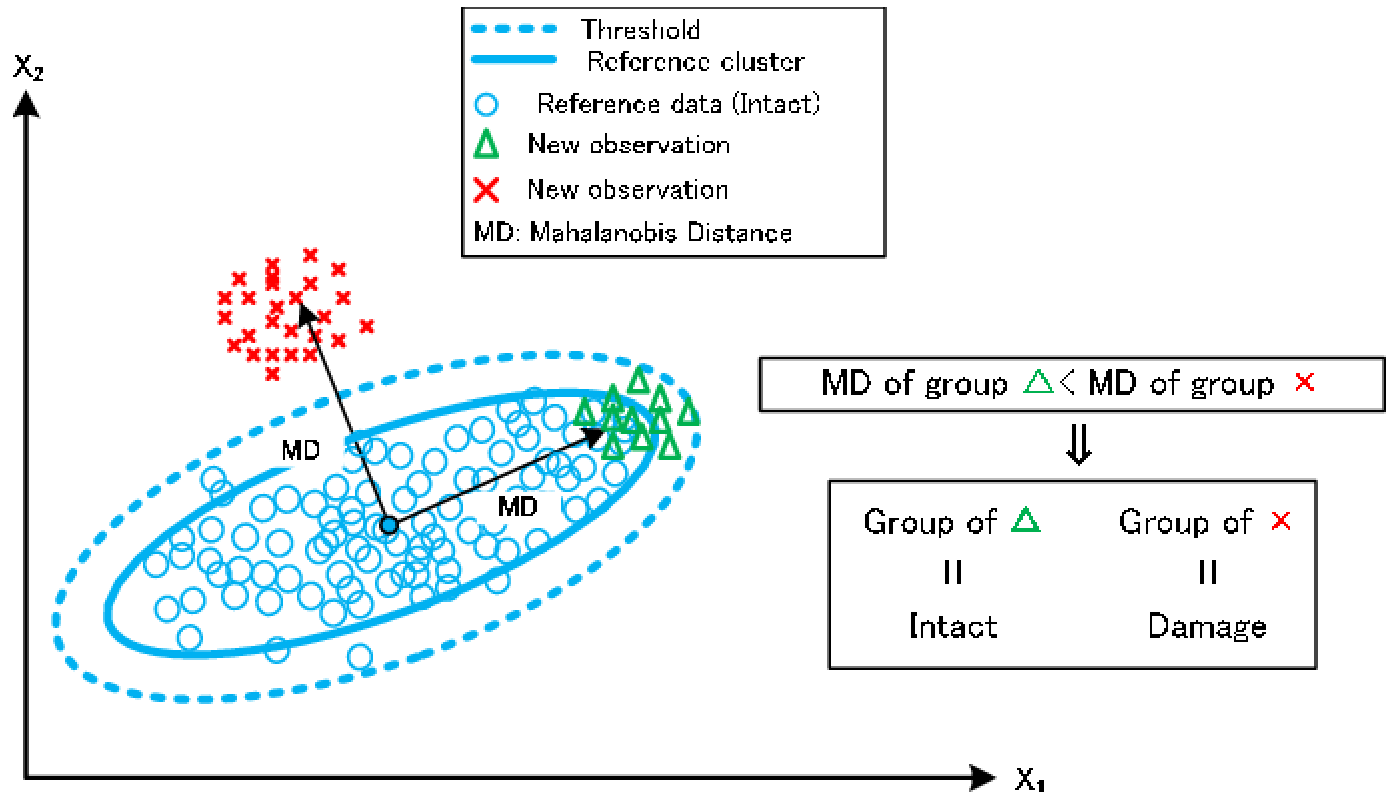
距離矩陣

馬氏(Mahalanobis)距離

- 可用來處理屬性間具有相關性的問題：

$$\text{mahalanobis}(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{y})^T$$

其中$\boldsymbol{\Sigma}^{-1}$是共變異矩陣的反矩陣



https://www.researchgate.net/figure/Concept-of-Mahalanobis-distance-MD_fig7_275701517

Feature Selection

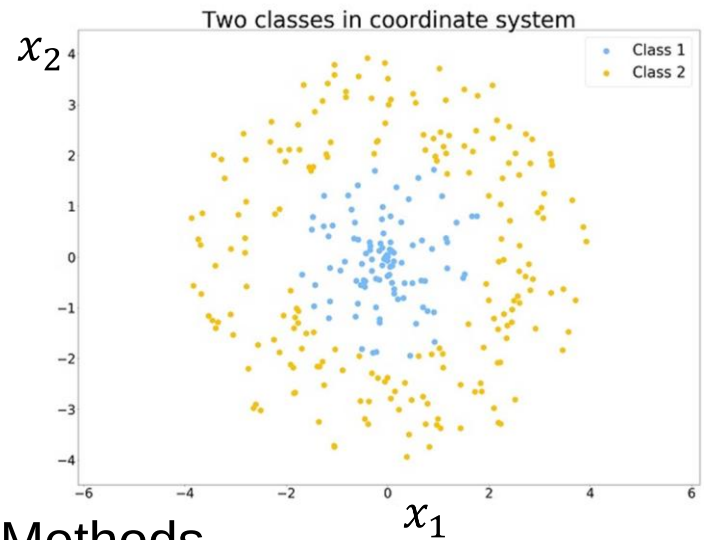
- 資料本身的價值因資料解析度(resolution)不同而有所差別，可經由資料匯總提升資料代表的意義
  - 資料表中描述資料集合所用的特徵或屬性稱為資料維度(dimension)

- 資料蒐集階段應盡可能地蒐集所有可記錄的變數或資料(或Feature Engineering)，再經由資料化約，得到與原始資料具有相同資訊但卻較精簡的資料集

- 其效益為：
  - 提升資料品質
  - 縮短資料挖礦時間
  - 提升資料價值、知識價值的取得與增加可讀性
  - 降低資料儲存成本
  - 避免維度的詛咒

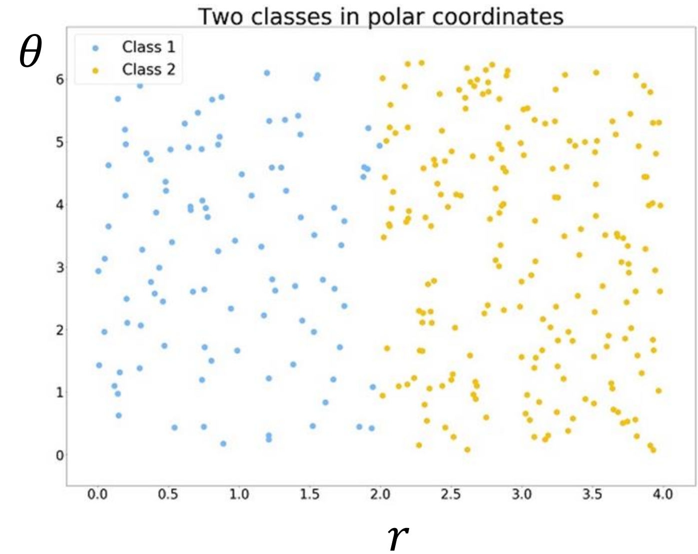
Feature Engineering

- is to manipulate the new feature manually from the original features.



$$r = \sqrt{x_1^2 + x_2^2}$$

$$\theta = \arctan \frac{x_2}{x_1}$$



https://www.kdnuggets.com/2018/12/feature-engineering-explained.html

Methods

- Temporal Features
  - "hour" can be binning as "morning/afternoon/night", or different shifts(早/晚班)
  - "day" can be binning as "weekday", "weekend"
- Image Features: Colorful to Gray Sale/Black-or-white, Rotation, Convolution, Pooling, etc.
- Text Features: Chopping, stemming, lemmatization, Word2Vec/GloVe/Doc2Vec, TF*IDF
- Spatial Features
  - Location in space, such as GPS-coordinates, cities, countries, addresses
  - Latitude and longitude can build the "median_distance_within_2_miles"

Feature Selection

- Variable Selection
  - Attribute subset selection
  - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
  - reduce # of patterns in the patterns, easier to understand
  - Heuristic methods are suggested due to exponential # of choices
    - stepwise regression, decision-tree induction

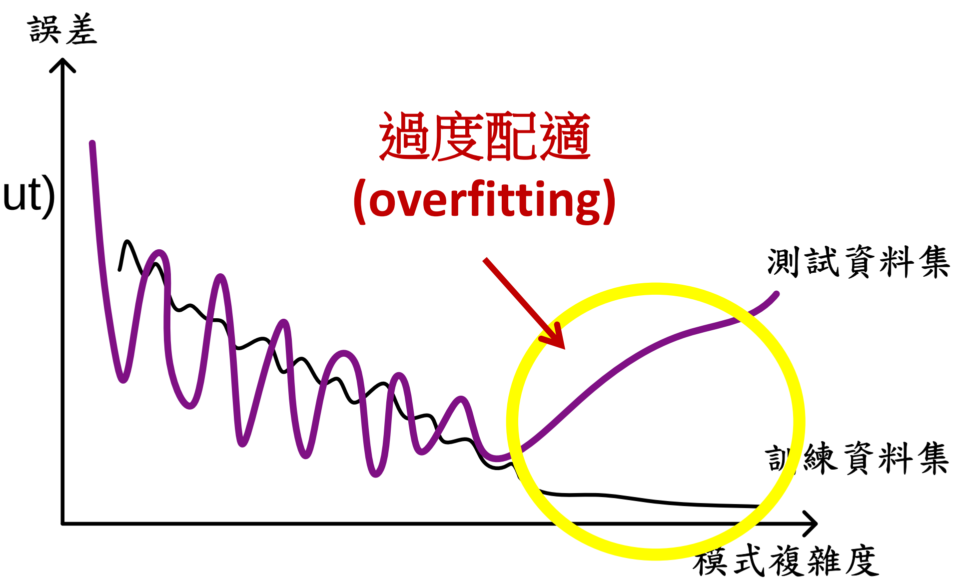
- Dimension Reduction (also called feature extraction/ variable transformation)
  - Extract the features from the original dataset (eg. linear combination)
  - Methods: Principal Component Analysis (PCA), Independent Component Analysis (ICA)

Data Partition

Data Partition

- Training, Validation, & Testing Datasets
  - Split the dataset for the model construction and validation
  - Training dataset: model training via cross validation
    - Validation dataset could be used for tuning hyper-parameters in the model
  - Test dataset: the model has never seen in training
  - eg. training data versus testing data = 8:2

- Cross Validation (CV)
  - K-fold CV
  - Time-series nested CV
  - Random-Sampling CV (Holdout)
  - Stratified sampling CV
  - Leave-one-out CV
  - Group CV…

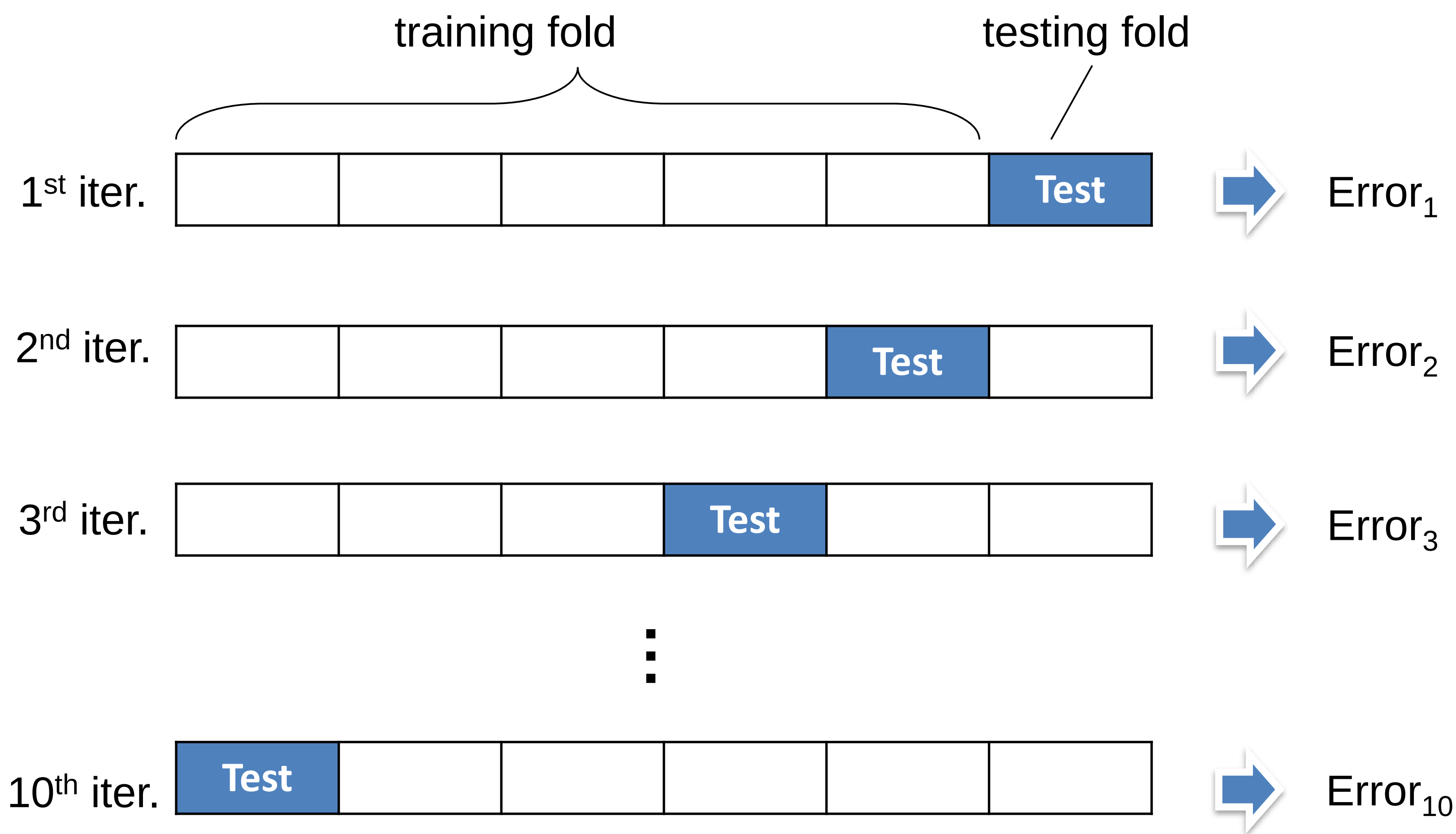


Hung (2018). https://rpubs.com/jeff_datascience/Semiconductor_Manufacturing

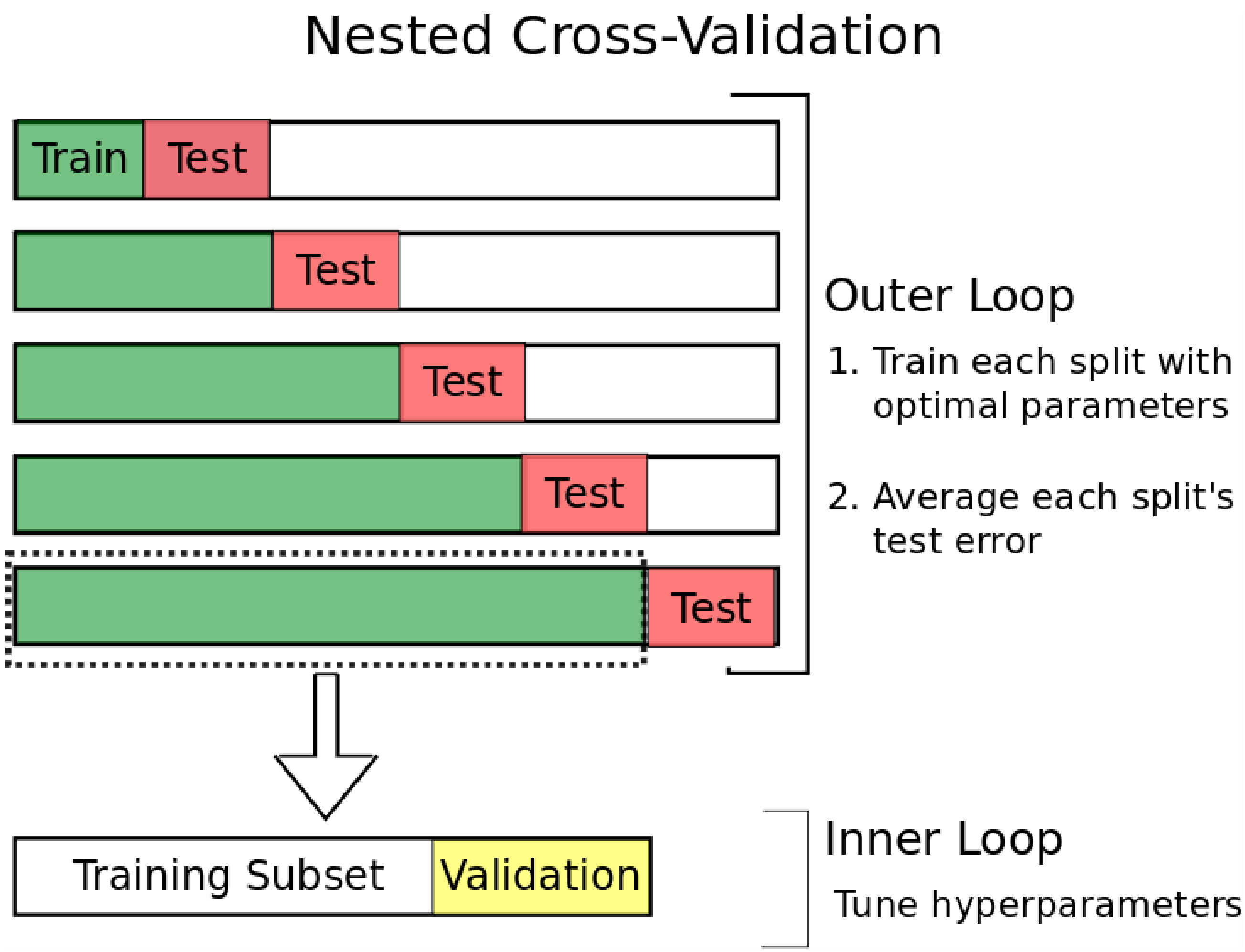
Data Partition

- K-fold Cross Validation
  - eg. 10-fold cross validation

$$\text{Minimize Error}= \frac{1}{10}\sum_{i=1}^{10} \text{Error}_i$$



- Time-series Nested Cross Validation



https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9

製造資料科學要做到…

看到資料，就能對應到現場的特性與問題

製造現場與其資料特性

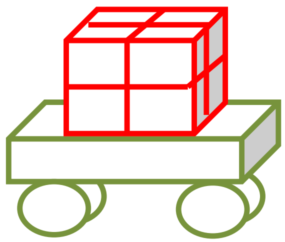
Characteristics	Data & Management Issues
Batch size (生產批量)	Lot ID decomposition, lot tracing, merge/split
Parallel machine (平行機台)	Missing value, high dimension, multicollinearity
Golden machine (黃金機台)	Utilization, class imbalance → Inference bias
Recipe and parts (處方與零件)	Nominal or categorical variable → too many levels → too many dummy variables → high dimension
Sampling testing (抽樣檢測)	Missing value, multi-response, metrology delay
Engineering or R&D lot (工程與實驗貨)	Outlier, machine contamination, setup capacity loss, small dataset
Maintenance (維修保養)	When? how (大保養 or 小保養)? capacity loss, reliability, typing error, text, choosing "others"
Changeover (換線、換模)	Sequence-dependent setup time, capacity loss
Bottleneck shift(瓶頸站轉移)	Different treatment, WIP transfer, product-mix
Queue time limit(等候時間限制)	Defects, WIP
Data imbalance (資料不平衡)	Inference bias

Inventory = Lead Time + Uncertainty

Lee and Chien (2020)

Parallel Machine

- Not identical (有機差) → Tool Matching



WS_A_Mach_1

WS_A_Mach_2

Data Preparation

- Missing Value

Lot ID	WS_A_ Mach_1_ Temp	WS_A_ Mach_2_ Temp
Lot001	820	N/A
Lot002	820	N/A
Lot003	N/A	840
Lot004	N/A	840

Lot ID	WS_A_ Temp	WS_A_ Mach_Type
Lot001	820	1
Lot002	820	1
Lot003	840	2
Lot004	840	2

Recipe/ Parts- Nominal (名目) or Categorical (類別) Variable

- Transfer to dummy variable (啞變數, 虛擬變數)

Lot ID	WS_A_Mach_1_Parts
Lot001	PartsA
Lot002	PartsB
Lot003	PartsA
Lot004	PartsC
Lot005	PartsD
Lot006	PartsE
Lot007	PartsB
Lot008	PartsA
Lot009	PartsC
Lot010	PartsE

Lot ID	WS_A_Mach_1_PartsA	WS_A_Mach_1_PartsB	WS_A_Mach_1_PartsC	WS_A_Mach_1_PartsD
Lot001	1	0	0	0
Lot002	0	1	0	0
Lot003	1	0	0	0
Lot004	0	0	1	0
Lot005	0	0	0	1
Lot006	0	0	0	0
Lot007	0	1	0	0
Lot008	1	0	0	0
Lot009	0	0	1	0
Lot010	0	0	0	0

- Given N levels, the method will generate N-1 dummy variables.

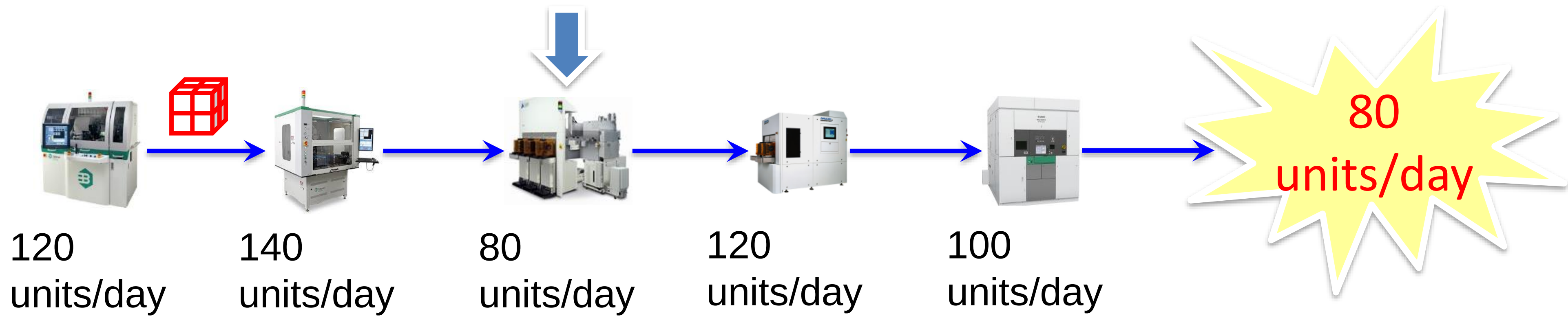
- 某類別變數level過多 (Recipe or Parts數目過多)
  - 轉成Dummy Variables會產生許多新變數
    - Issue: Curse of Dimensionality (維度的詛咒)
  - 建議方式
    - 將部分level整合 (grouping)
      - eg. 產品 → 產品族
      - eg. tool → tool group
    - 選取特定時間區間的資料進行分析
      - 降低該變數level的數目
    - 將某類別中只有出現一次觀測值的level刪除
      - 沒有再現性!

實驗貨工程貨

- 主要是研發產品、或機台測試校正用，資料上有時會呈現Outlier。
- 若針對一般性產品資料分析，需要在分析前先濾掉或移除。
- 一般而言，有特殊的LotID，在收集資料時可先過濾掉，或在預處理中進行。若無給定特殊LotID，那需要在資料中觀察，例如使用特殊機台、特殊recipe，該產品只經過某些特定製程等。

特性 \ 階段	實驗貨工程貨	一般正常貨
資料量	剛起步，較少 (n<<p)	較多 (大量生產) (n>>p)
資料數值	實驗設計，參數較分散	很多參數已成為定值
成本	需要反覆試驗，較高	大量生產，較低
良率	較低	較高且穩定
分析方法	最佳化方法、無母數、實驗設計/田口方法、LASSO、SVM、Forward Stepwise...	有母數、GLM、Random Forests, Boosting, Deep Learning…

□ 瓶頸 (bottleneck)



- 一般來說，瓶頸機台常是利用率高且週期時間長的機台→ WIP堆積多

楊大和(2016)

□ 內部瓶頸

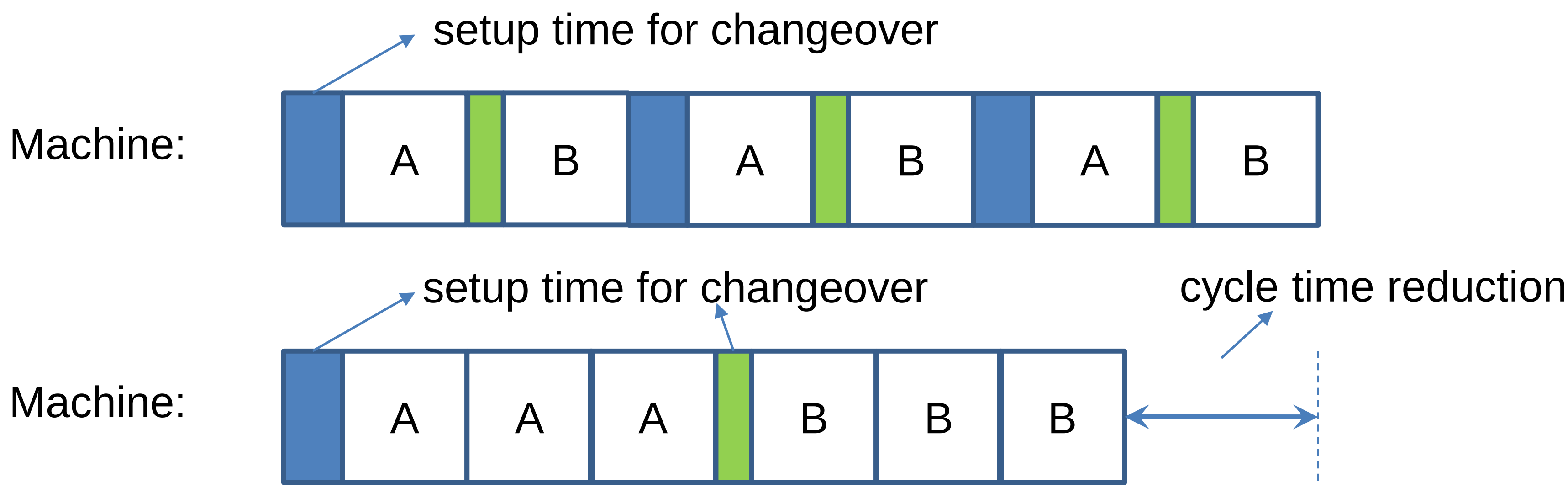
- 特定機台或工作站的產能限制
- 薪資水準或工作環境無法吸引到優秀員工
- 搬運/運輸/物流形成為生產的瓶頸
- 現場管理團隊能力/生產規劃團隊的排程/規劃
- 管理階層對於系統產能不正確假設/認知

□ 外部瓶頸

- 原物料的供應
- 特定區域的人力供給（勞工和幹部）
- 公司產品的品牌知名度
- 公司產品的配銷通路

換線、換模

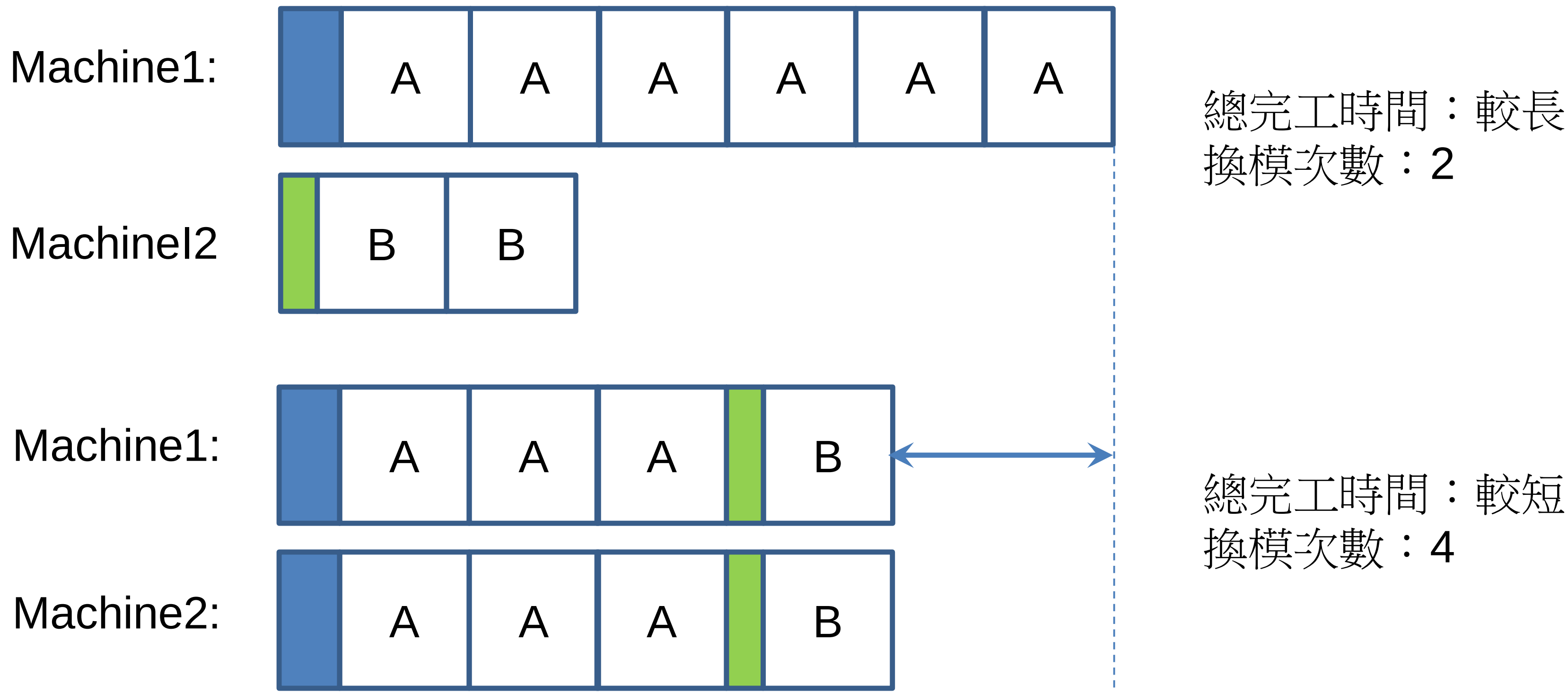
- 當產線要生產不同的產品時，會針對機台進行換線/模的動作。
- 同樣產品類型的儘可能排在一起，以減少換模(線)次數



- 以大批量生產的方式分攤換模(線)的時間
- 大批量生產方式會增加無訂單的庫存，因此必須預估數量，設定"經濟生產批量(Economic Lot Sizing)"來因應

總完工時間(makespan) vs. 換模次數??

- 原則上換模次數愈少，makespan下降，然而…當有大單時…
- 需留意多個平行機台(parallel machine)進行排程時的取捨(tradeoff)



- 資料會反應這件事情!
- 不單只是排程、良率也會因換模狀況而有所改變..

等候時間限制 (Queue Time Limit)

- 由於半導體製程晶圓表面上為化學物質，若長期曝露於一般空氣中，會造成氧化反應而導致缺陷(defect)產生。
- 為了避免在製品於生產線上等待過久而造成製程缺陷，會根據製程與產品特性，在特定製程完成加工之後，規定等候時間的限制（Queue Time Limit）以維持產品良率。為了延遲此限制，多於FOUP中填充惰性氣體。
- 等候時間若發生在批次工作站(例如爐管製程，furnace)問題可能更嚴重。對於到達此工作站之晶圓，除了需要等前一批次加工完，還需要另外等候集批(Form Batch)，換言之，需要等待多個批量後(有相同的recipe製程條件)，該工作站才進行作業。此加工型態會造成產品的等候時間過長，甚至超出等候時間限制，而造成不良品產生。
- 通常可計算Qtime當作獨立變數(x)來對良率(y)進行建模，以瞭解Qtime長短如何影響良率的情況。
  - estimated by the difference between check-out of A and check-in of B

資料合併

- 表單串接 – 注意必須為相同的欄位名稱, i.e., KEY
- Key通常為Lot ID, Machine ID等

Event-based record

Time	SVID 1	SVID 2	.....
2/11 00:06:29			
2/11 00:10:41			
2/11 03:41:09			
. . .			
2/11 23:11:57			

Periodic-based record

Time	SVID 101	SVID102	.....
2/11 00:00:00			
2/11 01:00:00			
2/11 02:00:00			
. . .			
2/11 23:00:00			

兩種不同類型的資料紀錄，該如何合併串接呢? Which one could be “Main Table”?

Data Merge

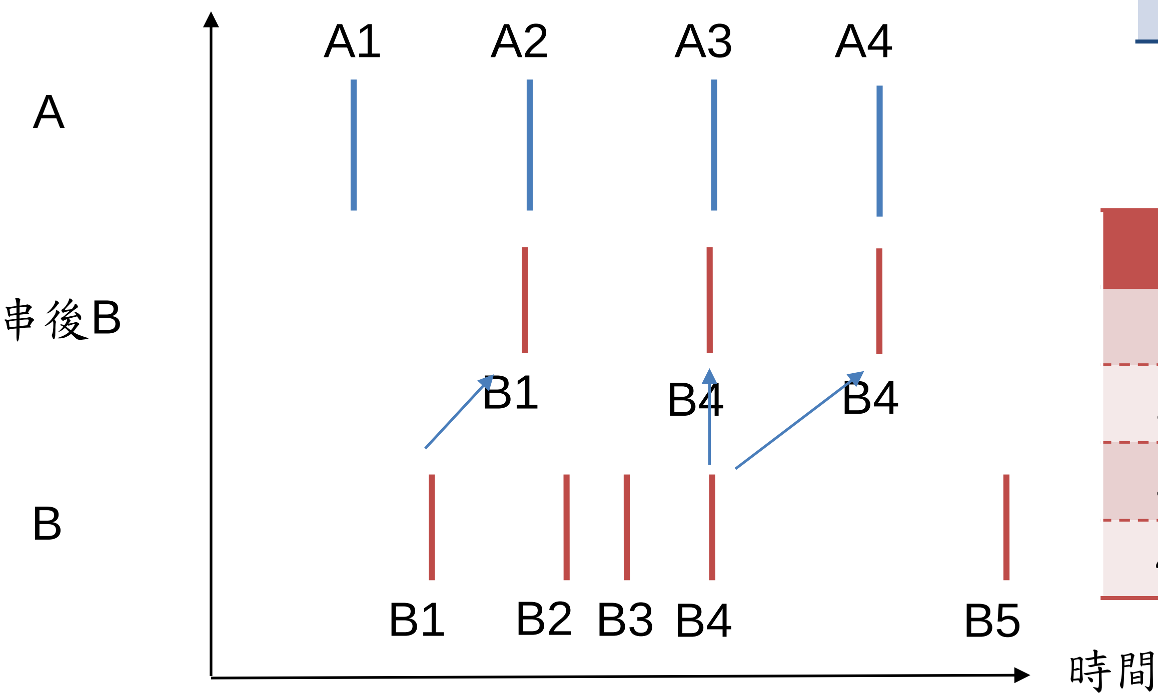
比較表	以 Event 為基準做串聯	以 Periodic 為基準做串聯
記錄方式	有"事件"才記錄。 例如機台換模、停機、人為調機等	固定"週期"記錄。 例如1小時一次
串接前 表單特性	資料筆數通常較少且稀疏	資料筆數通常較完整
串接後優點	資料較完整 (串接後可能遺漏值較少)	可觀察週期性變化
串接後缺點	可能有某"長"時間區段無資料	資料可能有部分缺失 (串Event會造成大量遺漏值)
建議串接 方法	Rolling Forward Nearest time	Rolling Forward Rolling Backward Nearest time
目的或 使用時機	**Troubleshooting**	**Monitoring**

Lee and Dong (2019)

資料合併- Rolling forward (過去歷史資料當中離現在最近的填進來)

	Date	A
1	2016-01-01	A1
2	2016-04-01	A2
3	2016-07-01	A3
4	2016-10-01	A4

	Date	B
1	2016-02-20	B1
2	2016-05-01	B2
3	2016-06-15	B3
4	2016-07-01	B4
5	2016-12-31	B5

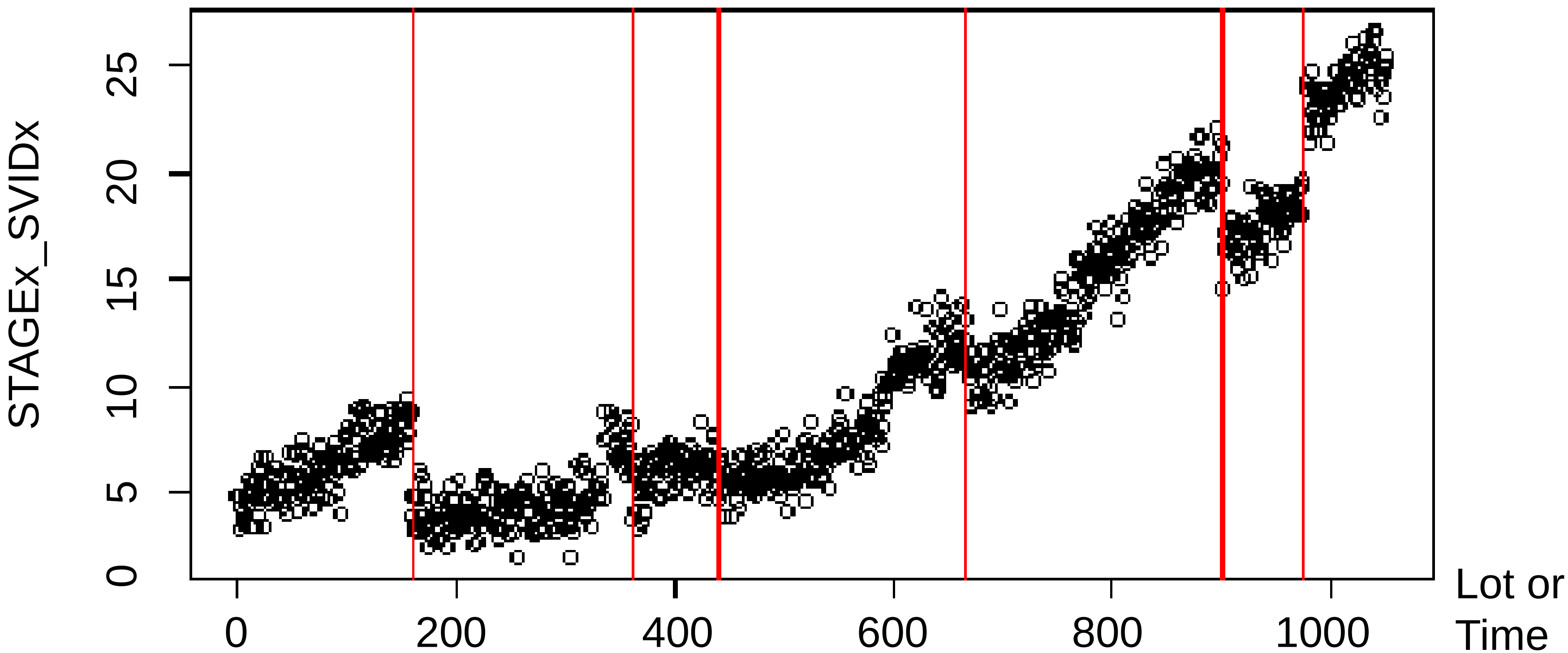


	Date	A	B
1	2016-01-01	A1	
2	2016-04-01	A2	B1
3	2016-07-01	A3	B4
4	2016-10-01	A4	B4

維修保養

- 定期保養 (年保、季保、月保、機台生產10,000產品…)
- 若有收集保養資料，可直接找出保養時間點。若無，可藉由推估
  - 機台up與down的時間 (Overall Equipment Effectiveness, OEE)
  - 產品的queue time
  - 機台參數(eg. status variable identification, SVID)的監控

再與相關部門(例如設備)進行確認。



維修保養troubleshooting

- 機台故障表單 + MES (含Recipe跟使用的零組件材料) + 良率
- 針對某一機台ID，用"時間"進行資料的合併串接

Time	Down code	Repair		Recipe	Part	Material		Yield
	機台故障表單				MES			良率
2017-05-07 14:05:28	**Run**	**No或NA**		Recip18	Part01	Mater05		94.3%
2017-05-07 16:12:14	**Run**	**No**		Recip18	Part01	Mater05		93.1%
2017-05-07 17:41:30	Down04	Part19 (換零件)	…	**Recip18**	**Part19**	**Mater05**	…	**82.5% 或內插**
2017-05-07 19:22:43	**Run**	**No**		Recip18	Part19	Mater05		82.5%
2017-05-07 20:18:17	**Run**	**No**		Recip02	Part19	Mater10		76.7%

$$\frac{93.1\% \times [(17:41:30) - (16:12:14)] + 82.5\% \times [(19:22:43) - (17:41:30)]}{(19:22:43) - (16:12:14)}$$

Data/Class Imbalance原因

- # of qualified product extremely dominates the # of defective product
- 資料不平衡大多發生於類別型態的資料上(一般泛指兩類)，若以連續分佈的資料來說，資料不平衡代表資料可能集中在某些區段，而這些區段也可以稱作"群/類別"。
- 資料不平衡的情況可能出現在獨立變數或是相依變數。

資料多不平衡才算不平衡?

- For the two classes (0 and 1), rule of thumb…
  - 10% vs. 90%? 5% vs 95%? or 1% vs. 99%?
  - It depends… on your industry applications.
- From a theoretical viewpoint, it occurs if it skews the model training for prediction…
- 也就是說，如果你訓練的模型準確率"異常地高"
  - Overfitting? Class Imbalance?

Lot ID	X1	...	X100	Inspection
Lot01				PASS
Lot02				PASS
Lot03				PASS
Lot04				PASS
Lot05				PASS
Lot06				PASS
Lot07				**FAIL**
Lot08				PASS
Lot09				PASS
Lot10				PASS
Lot11				PASS
Lot12				PASS

- 預測Inspection的結果
  - 由於只有1筆FAIL
  - 預測模型全部都猜PASS
    - 不需要分析變數X1~X100
    - 準確度可達 11/12 = 91.7%

Class/Data Imbalance

Class Imbalance Solutions

- Random sampling deals with the issue.
- Undersampling: samples a subset of the majority class.
- The main deficiency is that many majority class examples are ignored.
- Thus, we sample several subsets from the majority class (resampling).
- Others: oversampling, cost-sensitive, SMOTE, ensemble-based…

Example

- For Y label, 良品 vs. 不良品 = 1000 : 50
- Samples 50 良品 at a time for model training
- # of replications: 20 times
- Rank the variables by the “voting”
- Hint: 1:1 can be properly extended to 5:1

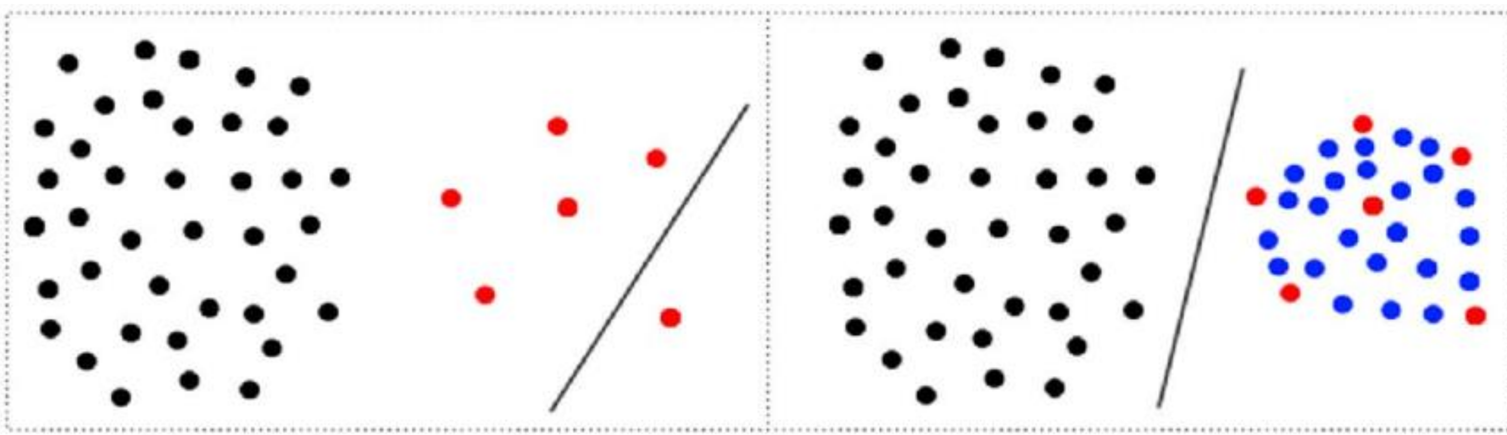
SVID	Voting by Undersampling
SVID_003	19
SVID_101	18
SVID_021	18
SVID_040	18
SVID_002	17
SVID_128	17
SVID_062	17
SVID_077	17
⋮	⋮

Pros and Cons

- Improve running time and storage problem
- Neglect potential useful information

Class Imbalance Solutions

- Data Imbalance
  - Synthetic Data Generation
    - The Synthetic Minority Over-sampling TEchnique (SMOTE)
    - It uses bootstrapping and k-nearest neighbors to generate artificial data.
  - Pass:Fail: 1463:104 → Pass:Fail: 1463:607



•majority class samples • minority class samples • synthetic sample

Dang et al. (2013).A novel over-sampling method and its application to miRNA prediction. Journal of Biomedical Science and Engineering, 6 (2A), 236-248.
Hung (2018). https://rpubs.com/jeff_datascience/Semiconductor_Manufacturing

Data Preparation Issues

Data Source	Scale	Issues
Production data (MES)	Categorical/continuous /time	High dimension, multicollinearity, class imbalance, missing value
Equipment data	Categorical/continuous	High dimension, too many categorical levels, time series, missing value
Parts/Supplier data	Categorical	Too many categorical levels
Transportation data	Categorical/continuous	too many categorical levels, time series, missing value
Maintenance/ Repair Data	Binary/categorical/continuous	Typing error, text, missing value, Choosing "others" or "NA"
Testing/Inspection Metrology data	Binary/continuous/ figure	Sampling data, time series, multi-response, metrology delay

Revised from Chen (2015)

Summary

- Association Rules
  - Decision-oriented system: process, resources, function
- Data Science Framework
  - data description, data preprocessing, feature selection, modelling & validation, visualization & conclusion
  - Prediction is only the process, decision-making is the purpose
- Data Preprocessing
  - 受到資料來源的不同，資料挖礦分析時需處理的資料型態也不盡相同，適當的瞭解蒐集的資料特性將有助資料挖礦模式的選擇
  - 有意義的資料呈現已成為資料挖礦與巨量資料分析的重點，視覺化的工具將可提供資料挖礦分析者更多元的整合資訊
  - 資料準備為資料挖礦的重要步驟，所需耗費的時間可能遠高於其他步驟