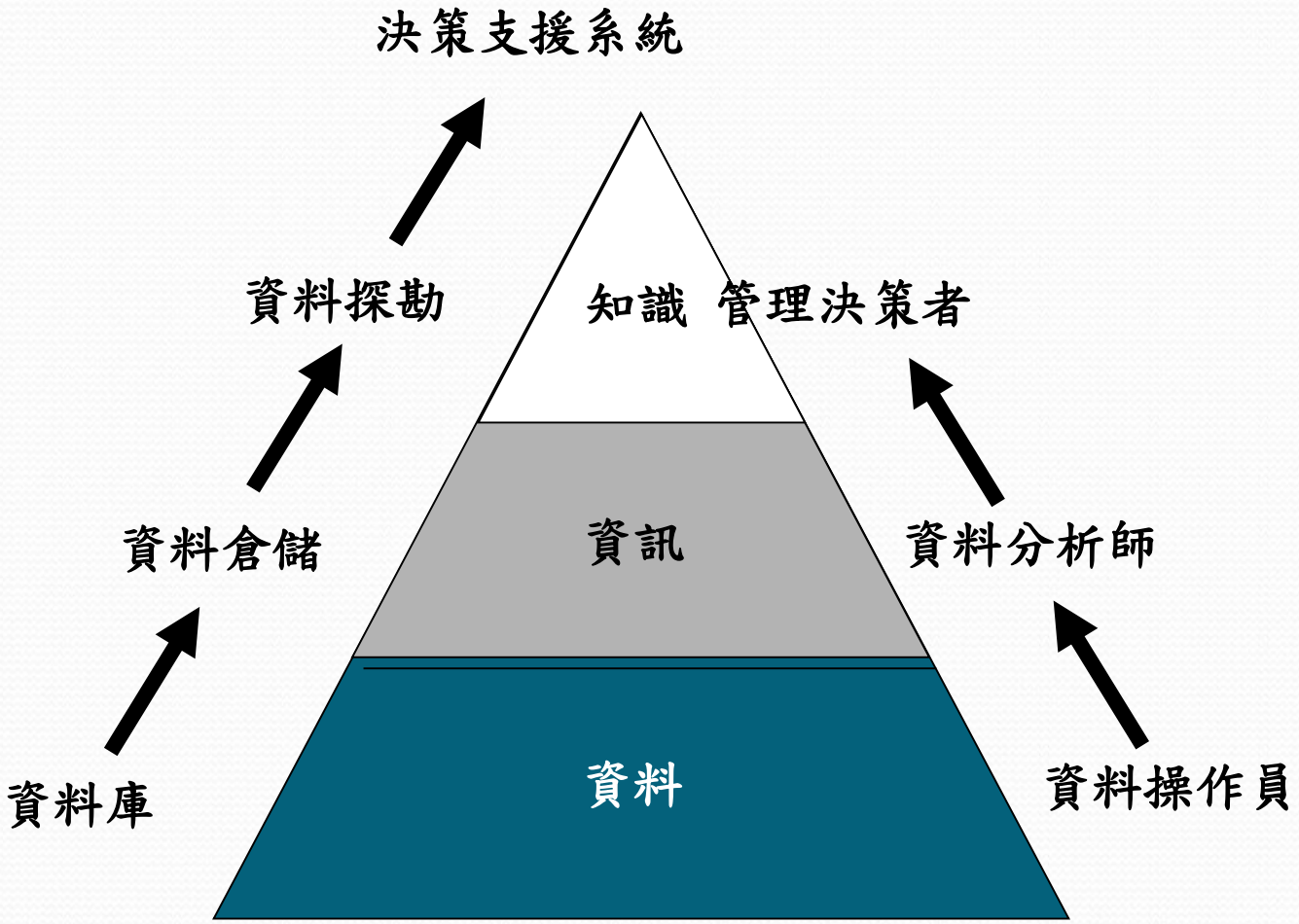
大數據分析之資料前處理與淨化技術

Chih-Fong Tsai
Department of Information Management
National Central University
cftsai@mgt.ncu.edu.tw

大數據分析構成要素

- 大數據 (或巨量資料)
  - 不同領域問題有不同的定義
- 分析技術 (或資料探勘技術)
  - 統計方法 (資料庫/資料倉儲系統)
  - 機器學習方法

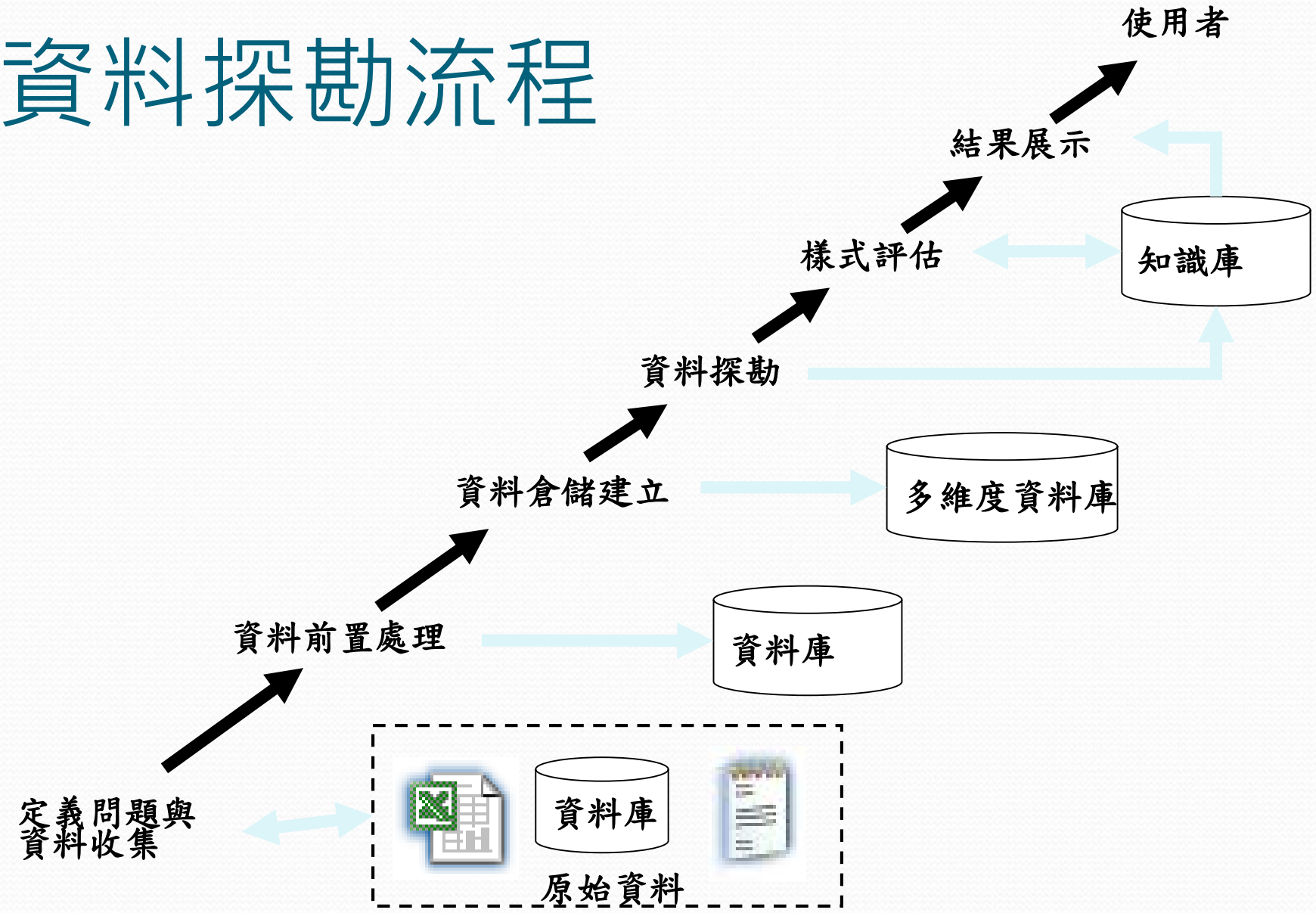
資料探勘



資料探勘 vs 資料倉儲系統

線上分析處理	資料探勘
多少人曾購買筆記型電腦？	哪些顧客可能會購買筆記型電腦？
上個月有多少顧客沒有進入網站瀏覽商品？	哪些顧客較有可能在未來三個月內不上站瀏覽商品？
顧客的平均單月消費總金額是多少？	哪些顧客下個月的消費有可能會超過一萬元？
哪些顧客訂單超過三天未付款？	哪些顧客較有可能延遲付款？
電子報的點閱率多少？	電子報行銷方式對那些會員較有效？
去年的銷售業績統計報表	明年預期之銷售業績額度。

資料探勘流程


使用者
結果展示
樣式評估
知識庫
資料探勘
資料倉儲建立
多維度資料庫
資料前置處理
資料庫
定義問題與
資料收集
資料庫
原始資料

資料前處理 (資料清理)

- 常見的資料正確性問題

檢查內容	說明
屬性的有效值或有效範圍	例如：性別屬性的值不是男性就是女性；生日的月份應該介於 1 和 12 之間。
數值的唯一性	例如：身分證字號或是顧客編號不可有重複。
參考完整性 (referential integrity)	例如：存在於訂單資料表中的會員編號必須同時存在於會員資料表中。
資料的合理性驗證	例如：從會員的生日計算出該會員的年齡只有 10 歲，但是該會員所填寫的學歷卻是博士，顯然不合理。

資料前處理 (資料清理)

- 常見的資料完整性問題

檢查內容	說明
是否缺少探勘所需的屬性	例如：當我們想要探勘顧客年齡與購買商品種類的關係時，卻發現資料庫中並未包含年齡這個屬性。
是否只包含統計整合過的資訊，而缺少詳細的單筆資料	例如：當我們想要分析某網站的瀏覽率以了解一天當中哪一個時段最多人拜訪這個網站時，卻發現該網站每天只有記錄一筆當天的總瀏覽人次，而缺少每個小時的瀏覽人次資料。

資料前處理 (資料淨化與清理)

- 其他基本前處理
  - 資料格式轉換
  - 資料正規化 (欄位數值:0~1之間)
  - 資料離散化
- 進階處理
  - 特徵選取 (feature selection/dimensionality reduction)
  - 案例選取 (instance selection/outlier detection)
  - 遺漏值填補 (missing value imputation)
- 有高品質的資料，才有高品質的探勘結果

探勘技術

- 預測
  - 分類 (classification)
  - 推估 (estimation/regression)
- 分群 (clustering analysis)
- 關聯法則 (購物籃分析; association rule mining)

預測技術 (監督式學習技術)

- 統計
  - 邏輯/線性迴歸(分類/推估)
  - 區別分析
  - 單純貝式分類法, 等等
- 機器學習
  - 決策樹
  - 類神經網路
  - 支援向量機
  - k-最鄰近分類法, 等等

資料精簡

- 資料精簡在資料探勘過程中所扮演的角色:

  - 主要應用在資料的前置處理階段

  - 從資料集合中挑選、過濾出具代表性與解釋能力較高的資料，進而減少整個資料探勘的時間和成本，甚至可以增進探勘的結果

- 資料精簡:

  - 特徵選取又稱做維度精簡

  -案例選取

資料精簡之優點

- 提高知識的應用性與準確性，降低無效、錯誤資料之影響
- 挑選少量且具代表性的資料將大幅縮減資料探勘所需的時間
- 使資料探勘方法的可用性提高
- 助於高價值知識的取得與提升知識可讀性
- 降低儲存的成本

資料精簡所包含之觀點

- 資料精簡：資料維度精簡、資料記錄精簡與資料數值精簡(資料離散化)
  - 會員資料集合

資料維度

會員編號	平均月收入(千)	教育程度	年齡	會員等級
1	21	高中	30	低
2	24	大學	29	高
3	33	國中	28	高
4	20	國中	32	低
5	42	高中	31	低
6	38	大學	35	高
7	37	高中	36	高

資料記錄

資料數值

特徵選取/維度精簡

- 步驟與流程:

  (1) 給定一個資料集 D (M個維度)

  (2) 透過特徵選取演算法分析出 N個重要的維度 (N<M)

  (3) 使用處理後的資料集 D' (N個維度) 進行後續的資料探勘流程

- 維度越高的資料經過特徵選取後的影響越高

特徵選取/維度精簡

- 特徵選取演算法可分為三種技術類型:
  - filter (過濾器) (非監督式)
  - wrapper (包裝器) (監督式)
  - embedded (嵌入式) (監督式)

Filter Based Feature Selection

- 統計方法為主, 例如:
  - principal component analysis
  - information gain
  - stepwise regression
  - t-test
  - factor analysis, etc.

Filter Based Feature Selection

- 主要目的是分析每個特徵屬性的重要程度
- 分析的結果可將所有特徵進行重要性排名
- 優點: 執行時間快速
- 缺點: 表現不一定最好
- 如何篩選?
- 基本上保留80%重要的特徵

Filter Based Feature Selection

- 範例: Wine dataset (分類葡萄酒品種; 13個維度)
- UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php)

	Non-feature selection	PCA (80%)
CART	0.932	0.989 (+0.057)
MLP	0.978	0.994 (+0.016)
SVM	0.983	0.933 (-0.05)

Wrapper Based Feature Selection

- 機器學習方法為主, 例如:

  - genetic algorithms (GA)

  - particle swarm optimization (PSO)

Wrapper Based Feature Selection

- 主要目的是透過適應函數 (fitness function) 以監督式學習的方式選取能獲得最佳正確率的資料子集(即特定的欄位)
- 優點: 透過一定的迭代次數的學習篩選結果會比較好
- 缺點: 執行時間過長
- 影響篩選結果的因素: 演算法本身的參數, 例如適應函數與迭代次數

Wrapper Based Feature Selection

- 範例: Wine dataset

	Non-feature selection	PCA (80%)	GA
CART	0.932	0.989 (+0.057)	0.989 (+0.057)
MLP	0.978	0.994 (+0.016)	0.994 (+0.016)
SVM	0.983	0.933 (-0.05)	0.994 (+0.011)

Embedded Based Feature Selection

- 常見的方法: C4.5/CART 決策樹與隨機森林 (Random Forest)
- 主要目的是篩選特徵與建立預測模型兩個階段同時完成
- 優缺點: 包含上述兩種類型的技術

Embedded Based Feature Selection

- 範例: 糖尿病 dataset (182個特徵欄位)
- 以 SVM 分類器為例

Non-feature selection	PCA	GA	C4.5
0.643	0.643	0.683 (+0.04)	0.741 (+0.098)

- C4.5 只選了10個特徵

特徵選取/維度精簡 - 小結

- 資料收集完成後, 進行探勘探勘前可以先執行此階段進行分析
- 有執行特徵選取一定比未執行特徵選取好嗎?
- 哪種技術最好?
- 不同特性或領域的資料會影響此階段的分析結果: 特徵數量, 資料數量, 分類數量, 欄位值 (連續型, 離散型, 混合型數值)

案例選取

- 目的:隨著資料表中的資料記錄愈來愈多，整個資料探勘所需的時間將跟著拉長，同時儲存資料的空間也變大
- 當資料集合中存在無關、偏差的資料記錄時(離群值;outlier)，將資料記錄作適當的精簡，將能獲得更準確有效的知識

案例選取

- 步驟:

  - 給定一個資料集 *D* 由訓練資料集 *T* 與測試資料集 *U* 組成

  - 執行案例選取將 *T* 篩選出重要且具代表性的子集 *S*

  - 使用 *S* 訓練並建立分類器 Model1

  - 使用 *U* 測試 Model1

- Baseline: 使用 *T* 訓練並建立分類器 Model2 並使用 *U* 測試 Model2

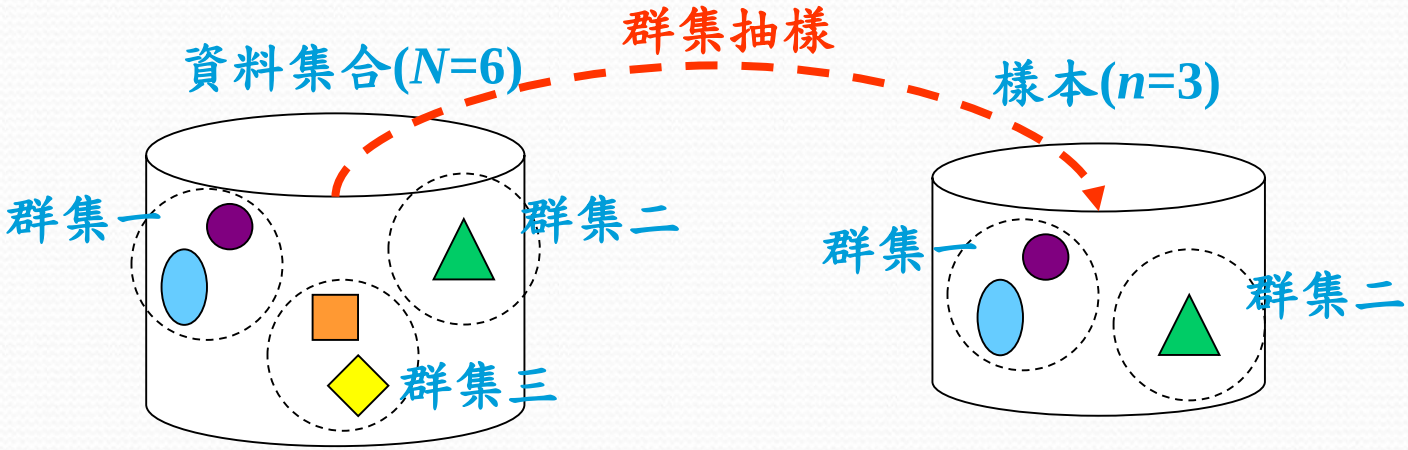
- 當資料量越大時, Model1 比 Model2 的差異會越顯著 (前提是需要搭配合適的案例選取演算法)

案例選取

- 可分為統計抽樣方法與監督式學習方法
- 抽樣方法:
  - 隨機抽樣 (容易導致偏差的結果)
  - 分群抽樣

群集抽樣(cluster sampling)

- 步驟一：利用群集分析技術，將整個資料集合區分成數個群集，使得每個群集中的資料記錄相似度很高，不同群集間的資料記錄相似度很低
- 步驟二：透過定義好的標準將這些群集中選取(1)某些群集或(2)每群的某些樣本當做結果



案例選取-監督式學習方法

- Edited Nearest Neighbor (ENN): 最早期具代表性的演算法之一
- 首先將 *S* 指定與 *T* 相同, 當某個樣本 *i* 經由 k 最近鄰居法則(k-nearest neighbor; k-NN) (k 通常設為3)進行分類後, 若 k-NN 能將 *i* 正確分類至其所屬類別, 則 *i* 將保留於 *S* 中, 否則 *i* 將從 *S* 中刪除

案例選取-監督式學習方法

- ENN 的延伸

  - Instance-Based Learning (IB1, IB2, IB3)

  - Decremental Reduction Optimization Procedure (DROP; DROP1, DROP2, DROP3)

  多加入了一些每個樣本需要被保留或刪除的特殊指標, 但是其中一個流程都會依照 *k*-NN 分類法則進行確認

  - 特性: 不錯的精簡率與執行時間效能快速

- Genetic Algorithms

  - 特性: 精簡率更高但非常耗時

案例選取

- 範例：KDD Cup 2008 Breast Cancer dataset (102294個樣本; 117個特徵)

	IB3	DROP3	GA
CART	79.32%	99.38%	99.31%
K-NN	77.84%	99.44%	99.35%
SVM	87.83%	99.44%	99.36%

	IB3	DROP3	GA
Reduction rate	42.02%	10.85%	56.96%
Processing time (min.)	12.13	455.29	1839.43

案例選取 - 小結

- Over selection 過度篩選: 精簡率增加可能導致正確率下降
- 有執行案例選取一定比未執行特徵選取好嗎?
- 哪種案例選取演算法最好?

遺漏值填補

- 在實務上，資料收集後會發現某些資料的一或多個欄位值是缺失的，此資料集稱為不完整資料集

	A	B	C	D	E
1	5	1	1	1	2
2	5	4	4	5	7
3	3	1	1	1	2
4	6	8	8	1	3
5	4	1	1	3	2
6	8	10	10	8	7
7	1	1	1	1	2
8	2	1	2	1	2
9	2	1	1	1	2
10	4	2	1	1	2

完整資料(complete data)

	A	B	C	D	E
1	5	1	1	1	2
2	5	4	4	NaN	7
3	3	1	1	NaN	2
4	NaN	8	8	1	3
5	4	1	1	3	2
6	NaN	NaN	10	NaN	7
7	1	1	1	1	NaN
8	2	1	2	1	2
9	2	1	1	1	2
10	4	NaN	NaN	1	2

不完整資料(incomplete data)

遺漏值填補

- 發生的原因: 機器和人為因素
- 機器因素: 資料儲存的失敗、儲存器損壞、機器故障導致某段時間資料未能儲存
- 人為因素: 人的資料輸入失誤、資料庫系統之設計局限或有意隱瞞或無意造成的資料缺失，例如問卷調查

遺漏值填補

- 方法一: 直接刪除發法 (case deletion/listwise deletion)
- 適用在缺失值比例(或遺漏率)較小的資料集，例如10%
- 不適合用在資料數量有限而且遺漏資料過多的資料集

遺漏值填補

- 遺漏值填補: 統計與機器學習方法
- 統計方法:
  - mean/mode (平均數/眾數)
  - regression, etc.
- 監督式學習方法
  - KNN
  - MLP
  - Decision trees: C4.5/CART/Random forests
  - SVM/SVR, etc.

遺漏值填補

- 範例:

	Name	Sex	Age	Height	Weight
1	Alfred	M	14	69	112.5
2	Alice		13	56.5	84
3	Barbara	F	13	65.3	98
4	Carol	F	14	.	102.5
5	Henry	M	14	63.5	102.5
6	James	M	12	57.3	83
7	Jane	F	12	.	84.5
8	Janet	F	15	62.5	112.5
9	Jeffrey	M	13	62.5	.
10	John	M	12	59	99.5
11	Joyce	F	11	51.3	50.5
12	Judy	F	14	64.3	90
13	Louise	F	12	56.3	77
14		F	15	66.5	112
15	Philip	M	16	72	150
16	Robert	M	12	64.8	128

遺漏值填補

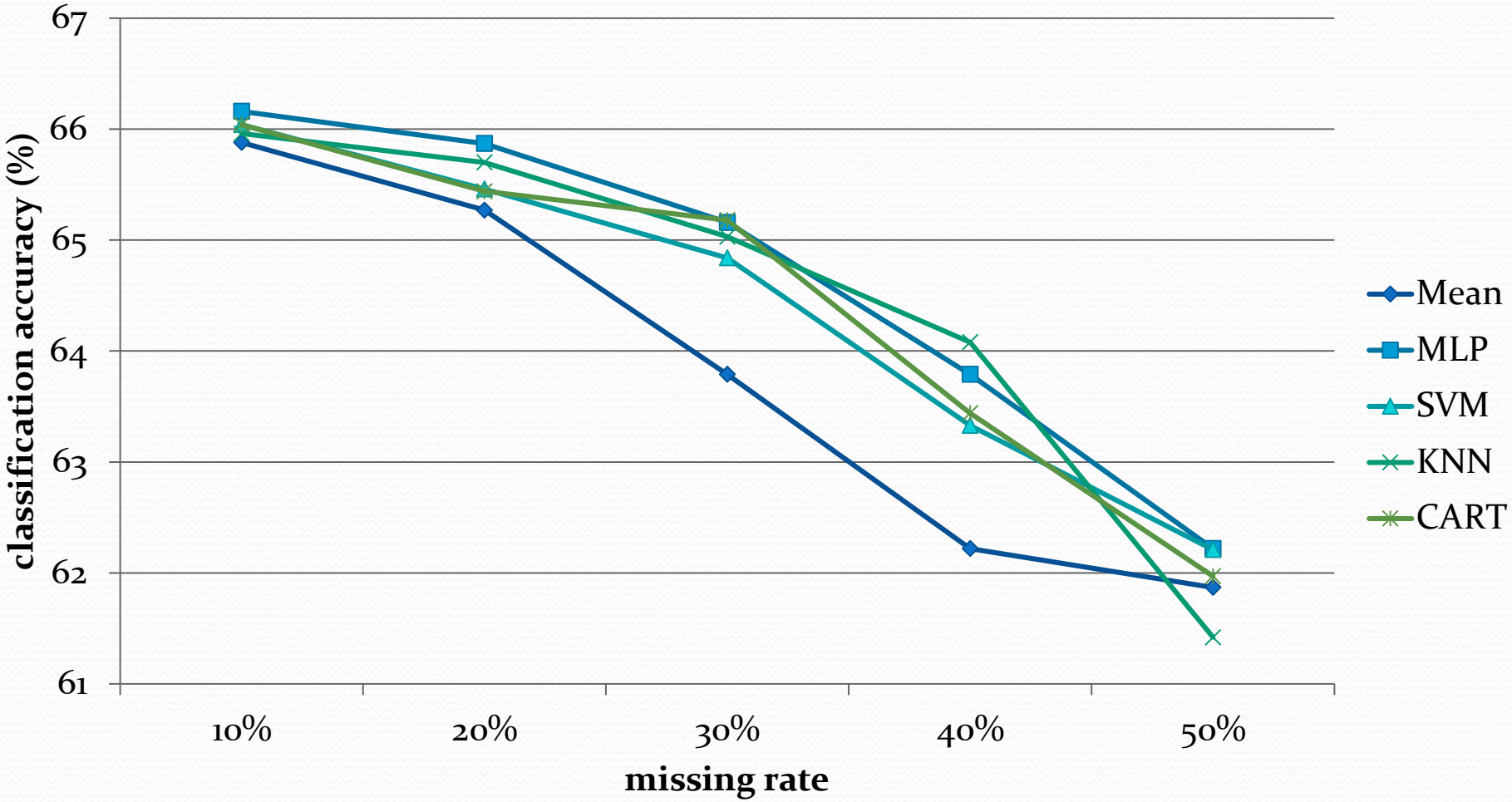
- 第2筆資料的Sex特徵是遺漏值，假設要填補Sex這個特徵值
- 必須將Sex視為被預測的特徵值(依變數)，Name、Age、Height、Weight這四個特徵值為訓練資料之特徵值(自變數)
- 第4、7、9、14筆資料皆有遺漏值，所以不可以當作訓練資料
- 因此除了上述資料(2為測試資料, 4, 7, 9, 14)，其他筆資料皆為訓練資料用以訓練預測模型
- 訓練完畢後即可將第2筆資料輸入預測模型，而其輸出值即可取代Sex特徵的遺漏值

遺漏值填補

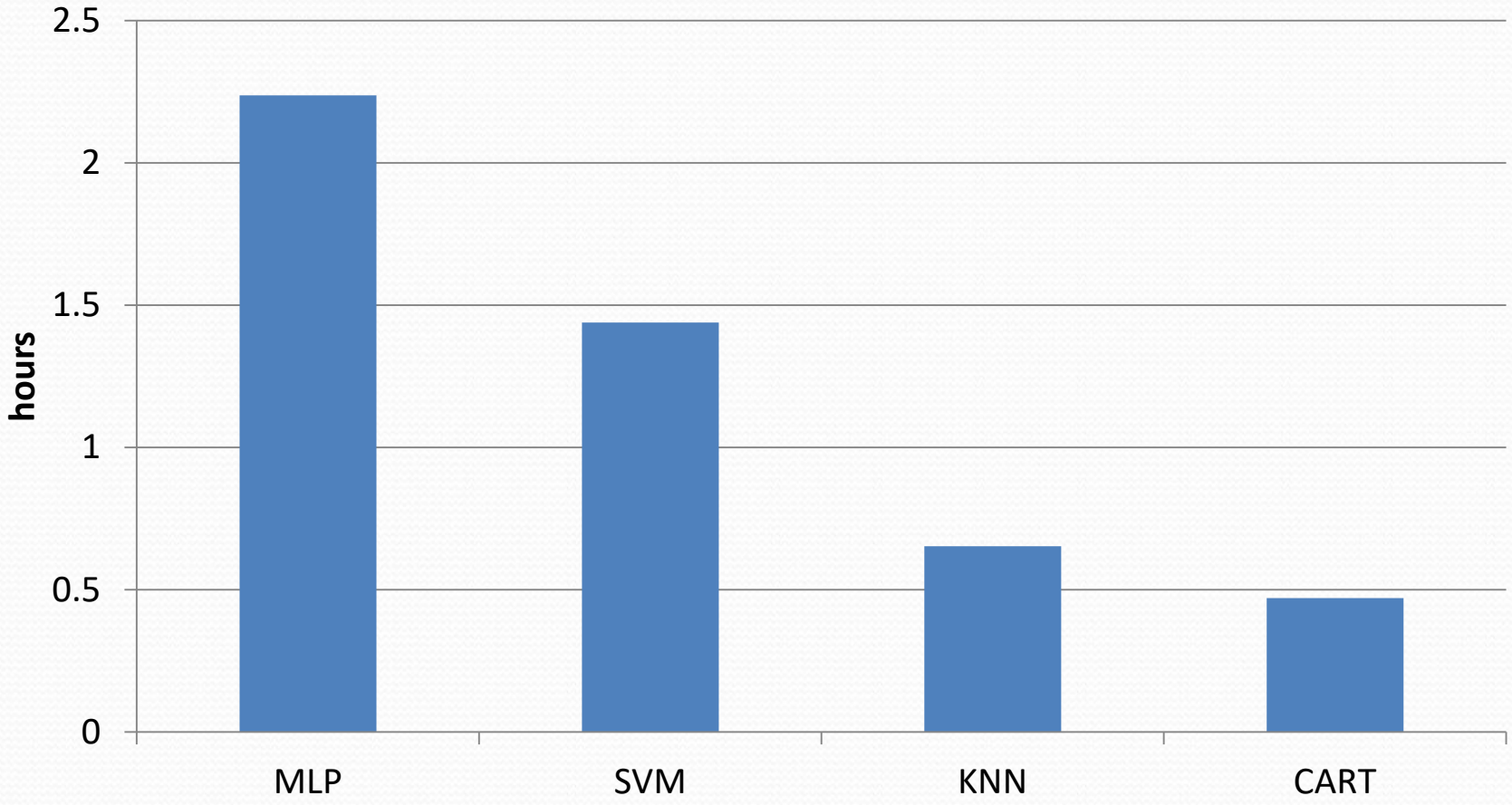
- 範例: 5個醫療資料集, 20%遺漏率, SVM 分類器正確率 (%)

	Mean	MLP	SVM	KNN	CART
Blood	74.56	74.99	74.72	74.56	74.51
Breast_cancer	76.62	77.04	75.77	76.61	75.2
Ecoli	69.07	71.27	70.72	71.38	71.27
Pima	64.88	64.94	64.94	64.94	64.94
Yeast	41.24	41.1	41.16	41.03	41.29
Avg.	65.27	**65.87**	65.46	65.7	65.44

遺漏值填補


67
66
65
64
63
62
61
classification accuracy (%)
10%
20%
30%
40%
50%
missing rate
Mean
MLP
SVM
KNN
CART

遺漏值填補



遺漏值填補 - 小結

- 監督式學習演算法和統計方法何種較佳?
- 資料集特性: 連續型/離散型/混合型數值, 資料維度, 資料數量, 分類數量以及遺漏率等會影響補值法的結果

Q&A