



Manufacturing Data Science



Foundation of Data Science

(第 3 章 數據科學基礎與模型評估)

Chia-Yen Lee, Ph.D. (李家岩 博士)

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

- 第一章 製造數據科學
- 第二章 製造系統分析與管理
- 第三章 數據科學基礎與模型評估
- 第四章 數據科學分析架構與系統運算決策
- 第五章 數據預處理與製造數據特性
- 第六章 線性分類器
- 第七章 無母數迴歸與分類
- 第八章 決策樹與集成學習
- 第九章 特徵挑選與維度縮減
- 第十章 類神經網路與深度學習
- 第十一章 集群分析
- 第十二章 特徵工程、數據增強與數據平衡
- 第十三章 故障預測與健康管理
- 第十四章 可解釋人工智慧
- 第十五章 概念漂移
- 第十六章 元啟發式演算法
- 第十七章 強化學習

藍：老師課堂講授

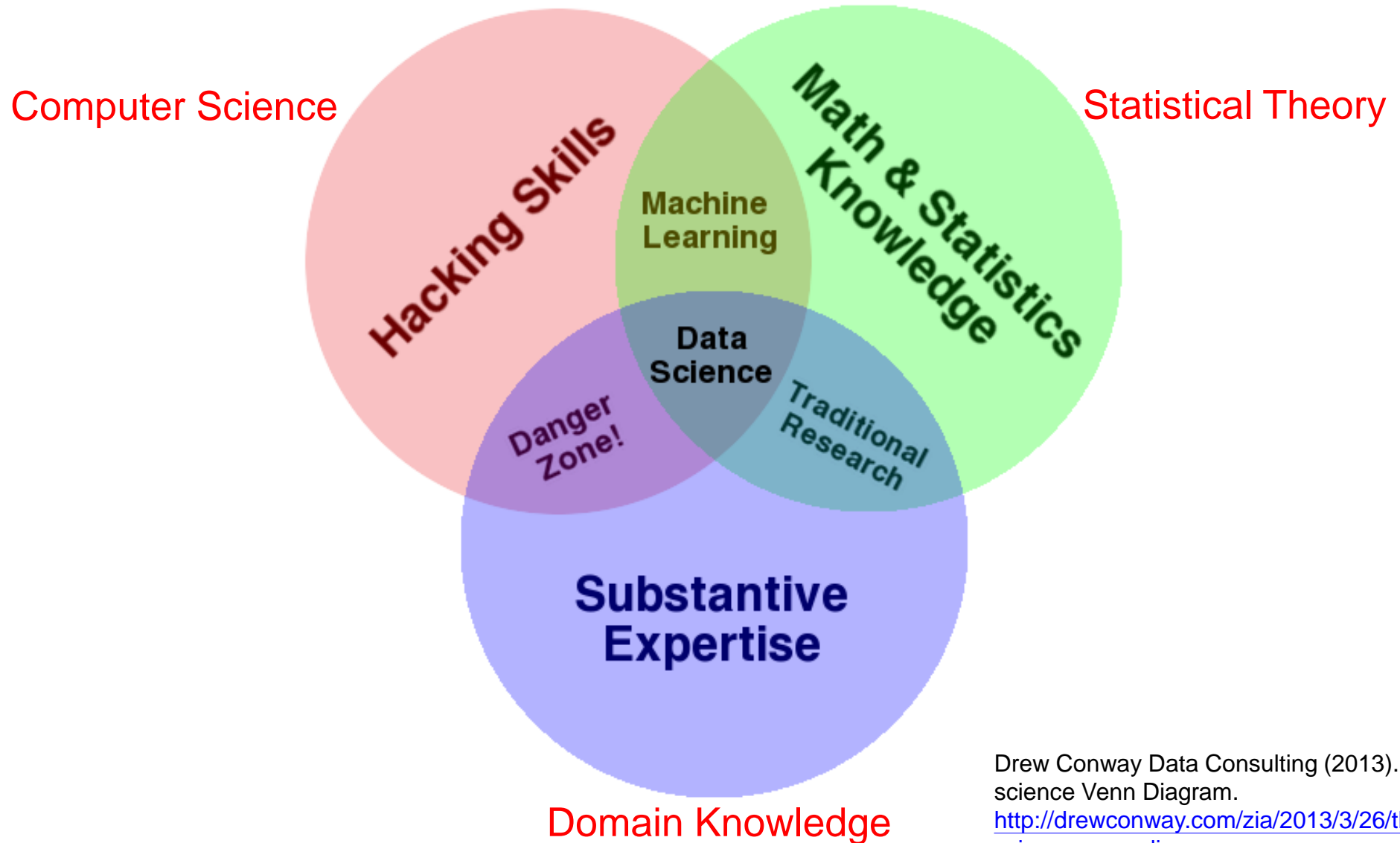
綠：學生自學

- 附錄A 線性迴歸
- 附錄B 支持向量機
- 附錄C 統計製程管制與先進製程控制
- 附錄D 超參數最佳化

- 應用涵蓋

產能規劃、瑕疵檢測、製程監控與診斷、機台保養、需求預測、生產排程、電腦視覺、自動光學檢測、原料價格預測與採購等

Data Scientist: nerd or geek !?



Drew Conway Data Consulting (2013). The data science Venn Diagram.

<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>

□ 人工智慧的歷史

- 1956年夏天於達特茅斯學院（Dartmouth College）所舉行的會議上正式確立了人工智慧研究領域的誕生，美國電腦科學家約翰·麥卡錫（John McCarthy）定義為「The science and engineering of making intelligent.」。

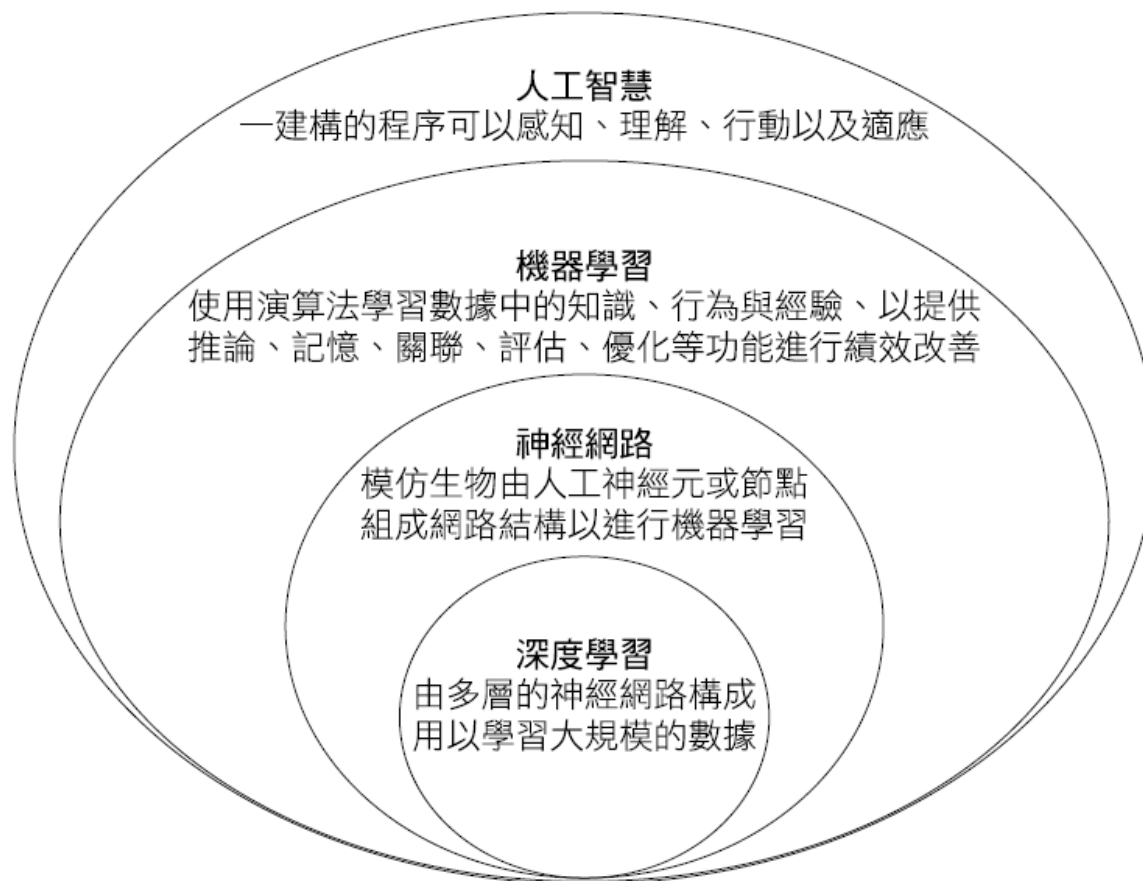


圖 3.1 人工智慧、機器學習、神經網路與深度學習

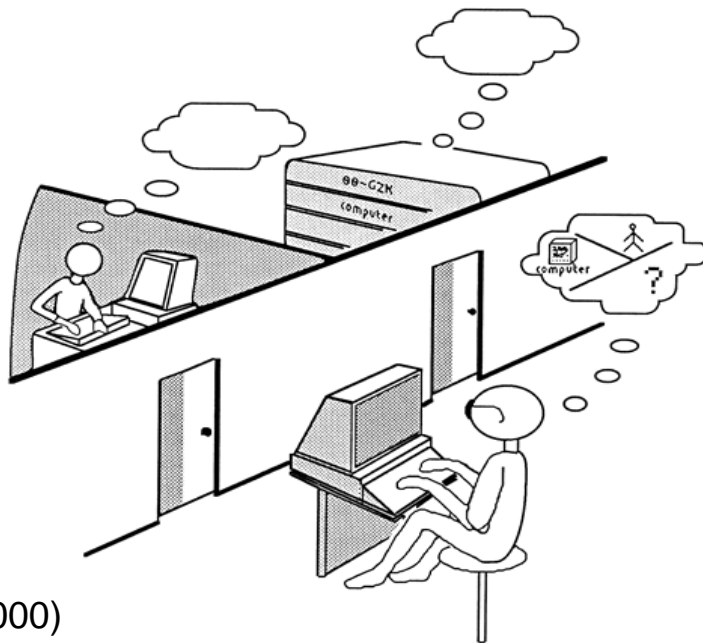
□ One of the first to believe in AI (Turing, 1950)

- A machine can mimic the human brain
- Many people did not want to believe a machine could do the same thing as a human

具備 Analysis(分析)、Learning(學習)、Adaptation(調適)、Reasoning(推理)、或 Decision(決策)等能力

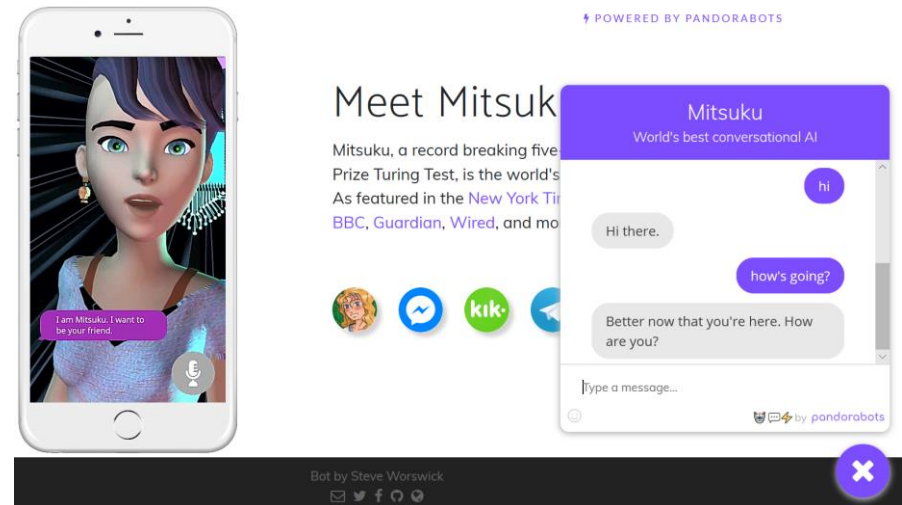
□ Turing Test

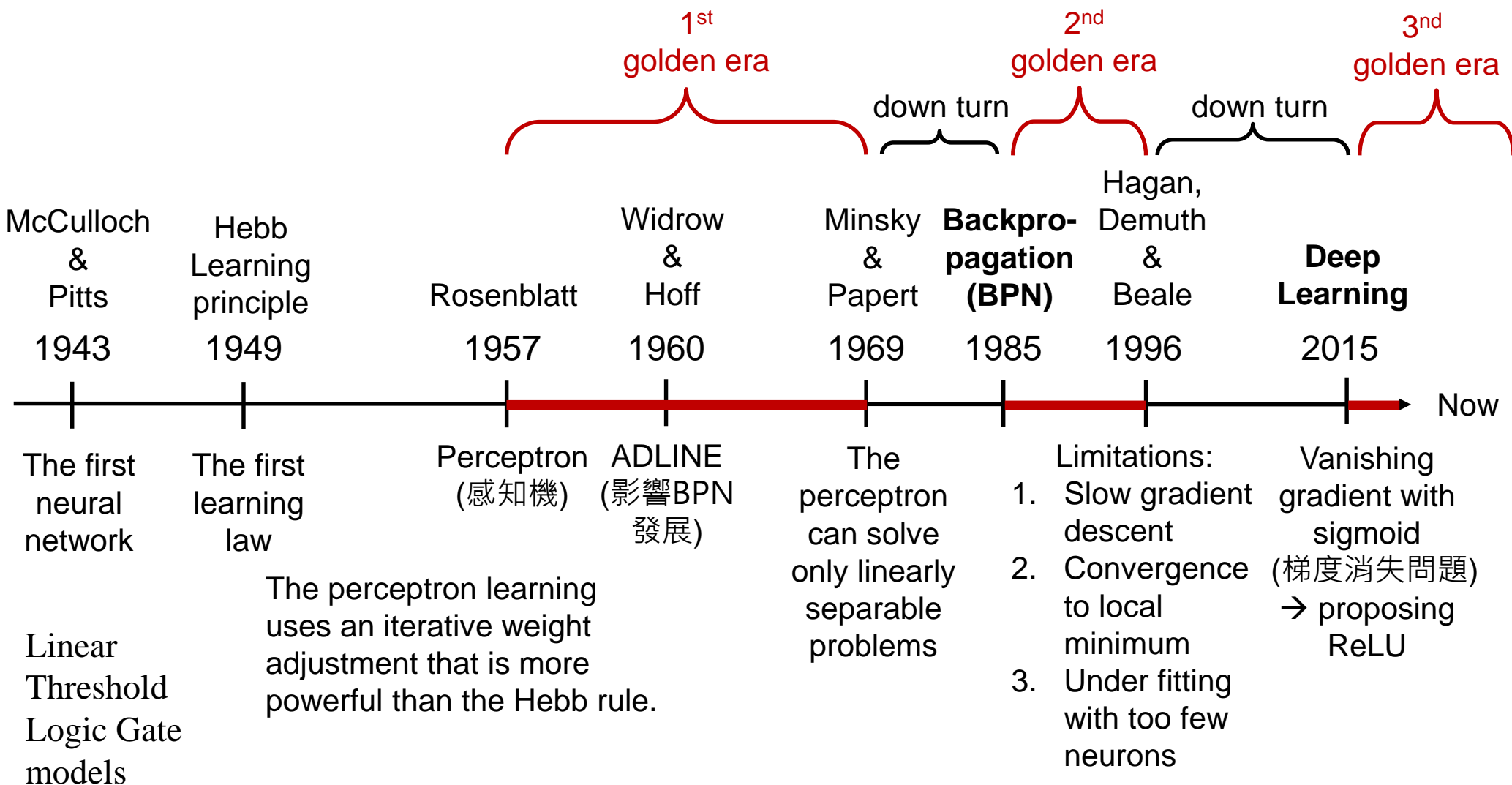
- A person asked questions on a keyboard to a person and a machine, if they could not tell the difference after some time the machine was considered intelligent (BBC News, 1999)



Copeland (2000)

Chatbot: <https://www.pandorabots.com/mitsuku>

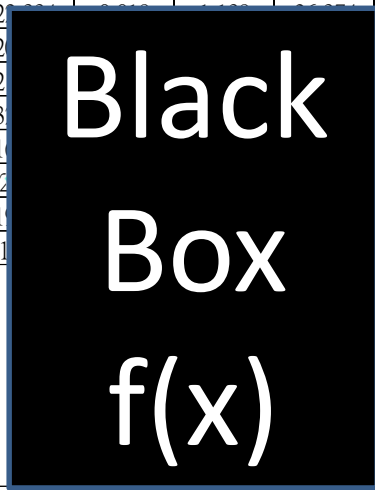




□ Generally, the Prediction of ML/DS...

- Supervised Learning, Unsupervised Learning, Reinforcement Learning

	Var_1	Var_2	Var_3	Var_4	Var_5	Var_6	Var_7	Var_8	Var_9	Var_10	Var_11	Var_12	Var_13	Var_14	Var_15	Var_16	Var_17	C lass
O bs1	-5349	19.8	20.2	0	2.713	0.324	24.069	0.011	2.092	24.301	0.024	0.031	0.002	0.01	6.653	24.478	295.528	1
O bs2	-5597	11.682	28.318	0	2.713	0.319	22.302	0.013	3.949	23.668	0.029	0.032	0	0.01	5.942	23.592	90.394	1
O bs3	-5440.5	22.435	17.566	0	2.713	0.35	21.646	0.016	0.685	23.346	0.023	0.013	0	0.01	3.073	19.719	101.866	1
O bs4	-5614.25	17.163	22.837	0	2.713	0.29	17.521	0.016	1.066	27.508	0.021	0.032	0.002	0.01	2.635	27.749	224.542	1
O bs5	-5534.5	25.457	14.543	0	3.198	0.35	22.798	0.018	1.619	29.305	0.042	0.024	0.001	0.012	3.533	18.54	162.479	1
O bs6	-2649.25	20.551	19.449	0	2.914	0.324	20.481	0.022	1.411	21.722	0.03	0.022	0.001	0.013	2.526	14.507	0	1
O bs7	-5554.25	18.683	21.317	0	3.198	0.38	33.628	0.023	1.641	19.756	0.046	0.034	0	0.012	3.513	24.921	147.607	1
O bs8	-6566	25.443	14.557	0	2.951	0.403	16.265	0.016	1.947	21.162	0.029	0.04	0	0.023	5.361	21.856	0	1
O bs9	-5613.75	17.637	22.363	0	2.914	0.247	20.281	0.031	1.183	16.635	0.039	0.022	0	0.013	7.098	23.817	913.882	1
O bs10	-6546.5	24.351	15.649	0	2.951	0.403	15.373	0.015	4.217	30.47	0.023	0.033	0.006	0.023	4.796	32.937	183.158	1
O bs11	-5652	12.063	27.937	0	2.951	0.27	22.971	0.022	4.47	25.754	0.026	0.029	0	0.023	3.714	24.679	526.739	1
O bs12	-6656.75	9.523	30.477	0	2.914	0.287	20.734	0.02	1.612	13.433	0.026	0.033	0	0.013	5.745	16.663	219.026	1
O bs13	-5681.5	22.925	17.076	0	3.198	0.26	20.734	0.026	0.02	0.001	0.012	2.813	12.986	163.435	1			
O bs14	-5537.5	9.334	30.666	0	2.951	0.27	20.734	0.029	0.031	0.002	0.023	4.219	24.275	150.967	1			
O bs15	-5349.25	12.539	27.461	0	2.692	0.342	20.734	0.025	0.016	0	0.016	3.356	23.889	168.404	1			
O bs16	-5495.75	12.431	27.569	0	2.642	0.324	30.47	0.026	0.026	0	0.011	7.639	25.813	208.284	0			
O bs17	-6518.25	28.636	11.365	0	2.692	0.289	10.005	0.036	0.044	0.005	0.016	3.703	32.604	329.604	1			
O bs18	-4556	8.693	31.307	0	2.914	0.27	20.734	0.029	0.031	0.001	0.013	3.831	38.056	149.578	1			
O bs19	-5432.5	-5.306	25.306	0	2.914	0.37	10.005	0.04	0.043	0	0.016	3.584	33.733	977.346	1			
O bs20	-5677.75	-8.749	28.749	0	2.914	0.268	10.005	0.012	0.034	0	0.016	2.849	23.108	771.212	1			



O bs101	-5384.25	2.583	37.417	0	2.828	0.378	33.6	0.012	2.439	19.77	0.018	0.059	0	0.039	4.417	31.293	226.108	?
---------	----------	-------	--------	---	-------	-------	------	-------	-------	-------	-------	-------	---	-------	-------	--------	---------	---

M. McCann, Y. Li, L. Maguire, A. Johnson, Causality Challenge: Benchmarking relevant signal components for effective monitoring and process control, Journal of Machine Learning Research: Workshop and Conference Proceedings, 6 (2008) 277–288.

□ 監督式學習(Supervised Learning)

- 有明確的學習目標，也就是數據中的目標值（target value）〔或稱標籤（label）、反應變數（response variable）、相依變數（dependent variable）〕，一般以符號 Y 表示。

- Data pair (X, Y)

- Y is label

- continuous \rightarrow prediction
- categorical \rightarrow classification

- 舉例：產品良率預測

- 產品數據的良率精確至0到100%的連續數值時，可使用迴歸模型預測
- 當產品數據為良品與不良品的離散二元變數時，則可使用二元分類模型預測
- 「產品良率」預測的價值
 - 部分節省品質檢測的成本（例如虛擬量測Virtual Metrology）
 - 從模型中期望建構「因果關係」（causal relationship），進行重要「特徵挑選」（feature selection），探討特徵對目標值的影響，協助「根本原因分析」（root-cause analysis）進而調整特徵改善良率

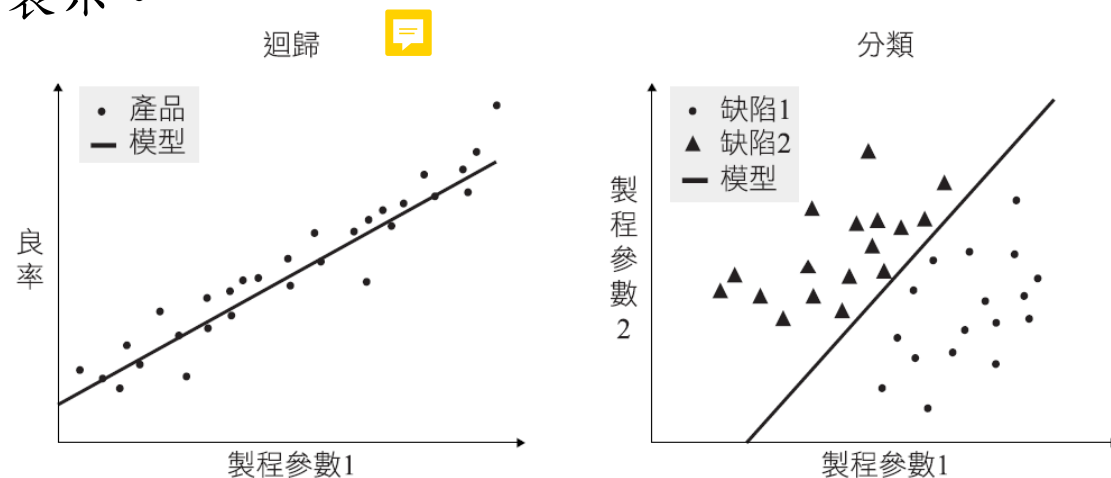


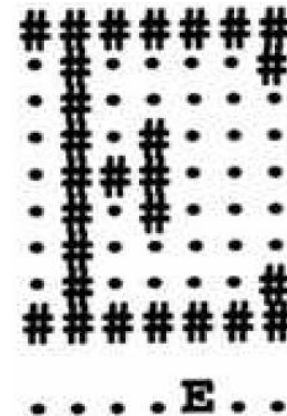
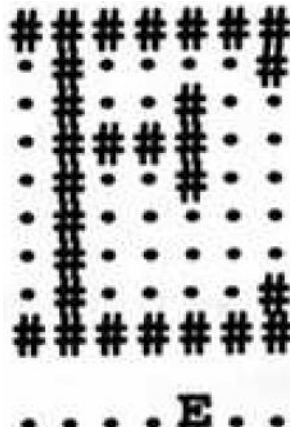
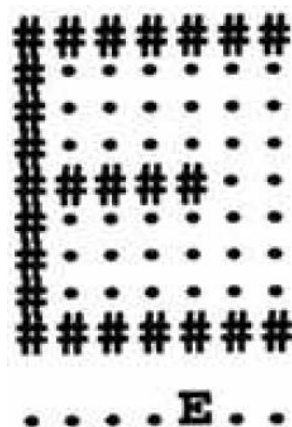
圖 3.4 迴歸與分類

AI怎麼訓練的呢？

案例：手寫識別 (Letter Recognition)

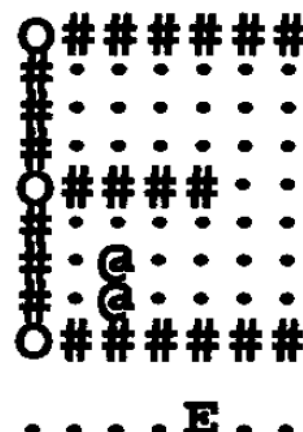
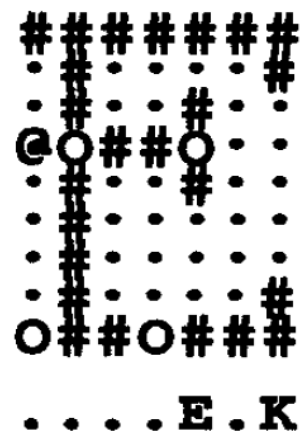
□ Handwritten Alphabet/Digit Recognition (Fausett, 1994)

- A perceptron to classify letters from different fonts: one output class
- 7 x 9 = 63 grid of pixels



#: 1
.: 0

- With noise (@) and missing data (O)



If noise exists, use bipolar better!

: 1
• : -1
Noise: 0

Handwritten Alphabet/Digit Recognition

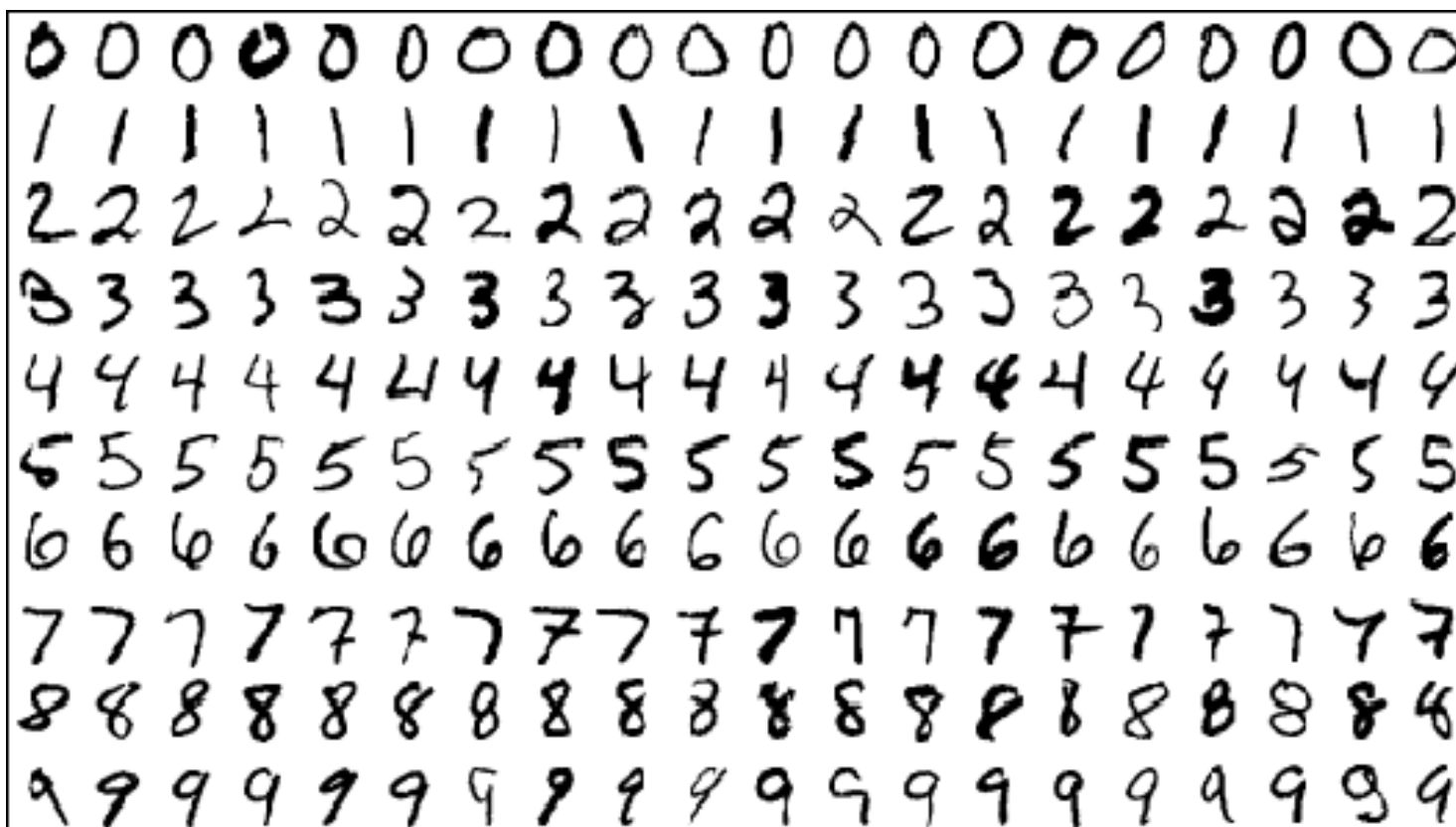
- 16 x 16 = 256 grid of pixels

Pixel															
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64
65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96
97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112
113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128
129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144
145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160
161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176
177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192
193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208
209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224
225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240
241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256

1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0
0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0
0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0
0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
0	0	1	1	1	0	0	0	0	0	0	0	0	1	1	1
0	0	1	1	1	1	0	0	0	1	1	1	1	1	1	0
0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0

□ Handwritten Alphabet/Digit Recognition

- Data Collection- MNIST dataset

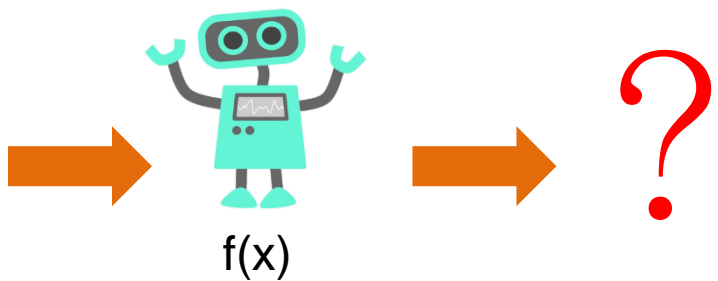


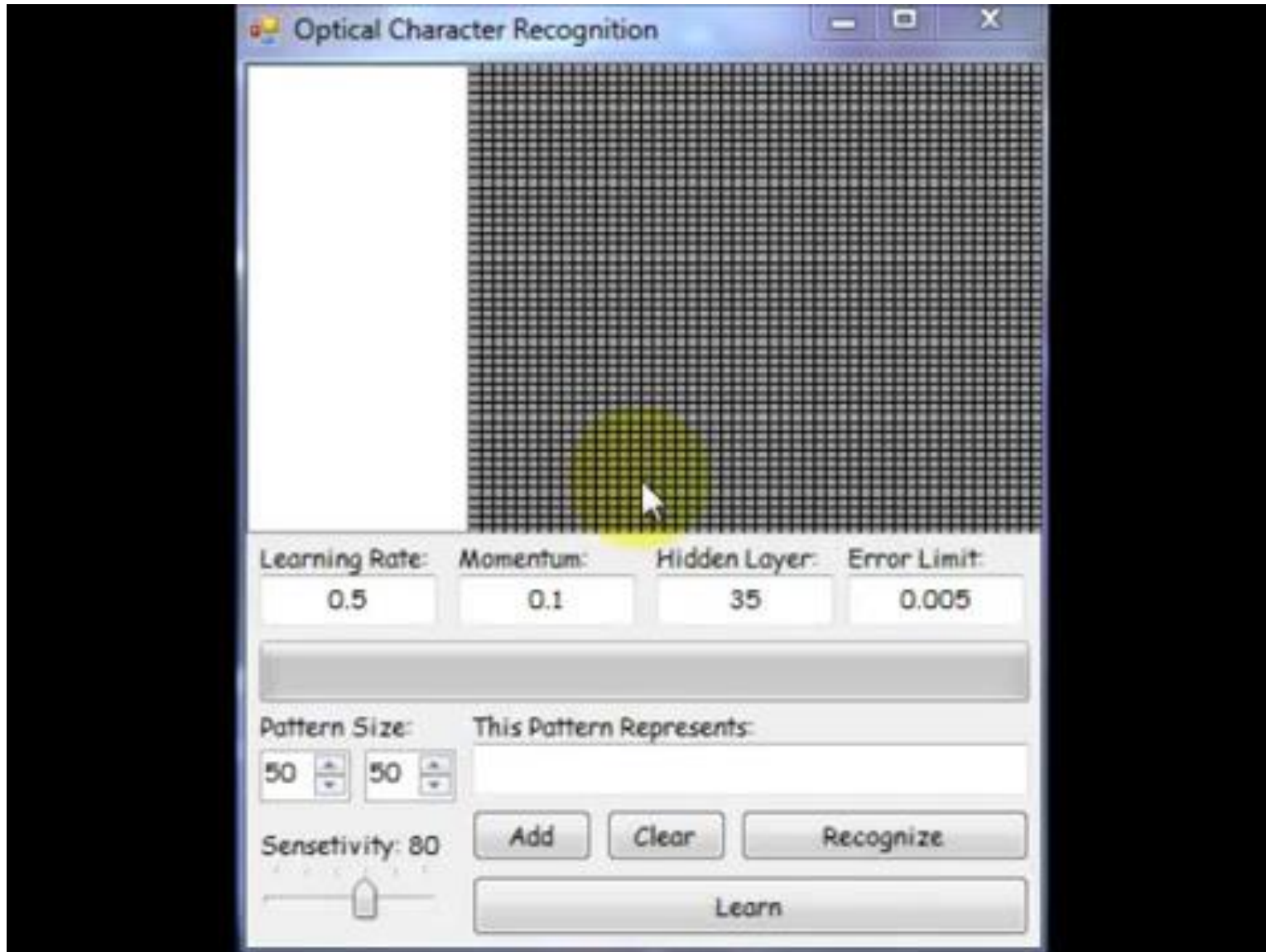
<http://yann.lecun.com/exdb/mnist/>

	A	B	C	D	E	F	G	H	I	J	K	IW	IX	IY	IZ	JA	JB	JC	JD	JE
1	No.	Pixel001	Pixel002	Pixel003	Pixel004	Pixel005	Pixel006	Pixel007	Pixel008	Pixel009	Pixel010	...	Pixel256	Target0	Target1	Target2	Target3	Target4	Target5	Target6
2	1	0	0	0	0	0	0	1	1	1	1	...	0	1	0	0	0	0	0	0
3	2	0	0	0	0	0	1	1	1	1	1	...	0	1	0	0	0	0	0	0
4	3	0	0	0	0	0	0	0	0	0	0	1	...	0	1	0	0	0	0	0
5	4	0	0	0	0	0	0	1	1	1	1	1	...	0	1	0	0	0	0	0
6	5	0	0	0	0	0	0	0	0	0	0	1	...	0	1	0	0	0	0	0
7	6	0	0	0	0	1	1	1	1	1	1	1	...	0	1	0	0	0	0	0
8	7	0	0	0	0	0	1	1	1	1	1	1	...	0	1	0	0	0	0	0
9	8	0	0	0	0	1	1	1	1	1	1	1	...	0	1	0	0	0	0	0
10	9	0	0	0	0	0	1	1	1	1	1	1	...	0	1	0	0	0	0	0
11	10	0	0	0	1	1	1	1	1	1	1	1	...	0	1	0	0	0	0	0
12	11	0	0	0	0	1	1	1	1	1	1	1	...	0	1	0	0	0	0	0
13	12	0	0	0	0	0	0	0	0	0	0	1	...	0	1	0	0	0	0	0
14	13	0	0	0	0	0	1	1	1	1	0	...	0	1	0	0	0	0	0	0
15	14	0	0	1	1	1	1	1	1	1	1	1	...	0	1	0	0	0	0	0
16	15	0	0	0	0	0	0	1	1	1	1	1	...	0	1	0	0	0	0	0
17	16	0	0	0	0	1	1	1	1	1	1	1	...	0	1	0	0	0	0	0
18	17	0	0	0	0	0	0	1	1	1	1	1	...	0	1	0	0	0	0	0
19	18	0	0	0	0	0	0	1	1	1	1	1	...	0	1	0	0	0	0	0
20	19	0	0	0	1	1	1	1	1	1	1	1	...	0	1	0	0	0	0	0
21	20	0	0	0	0	0	0	0	0	1	1			0	1	0	0	0	0	0



[0, 0, 0, 0, 0, 1, 1, 0, ..., 1, 0, 0, 0]





□ 非監督式學習(Unsupervised Learning)

- 相對於監督式學習是數據中不包含標籤，僅有特徵 X ，也就是沒有明確的學習目標。以化繁為簡的思維幫助決策的執行。
- 應用類型：
 - 「**分群**」(clustering)：站在觀測值（或特徵）的角度，將特徵相似的樣本（或特徵）分為同一個集群
 - 「**維度縮減**」(dimension reduction)：站在特徵的角度，在盡可能保留較多數數據資訊量的前提下，將多數特徵濃縮以少數特徵表示。

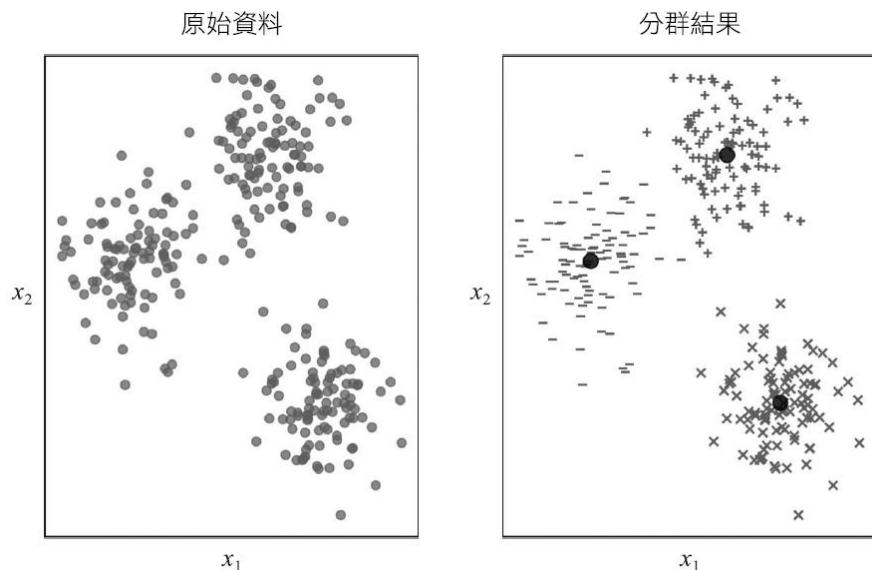


圖 3.5 分群示意圖

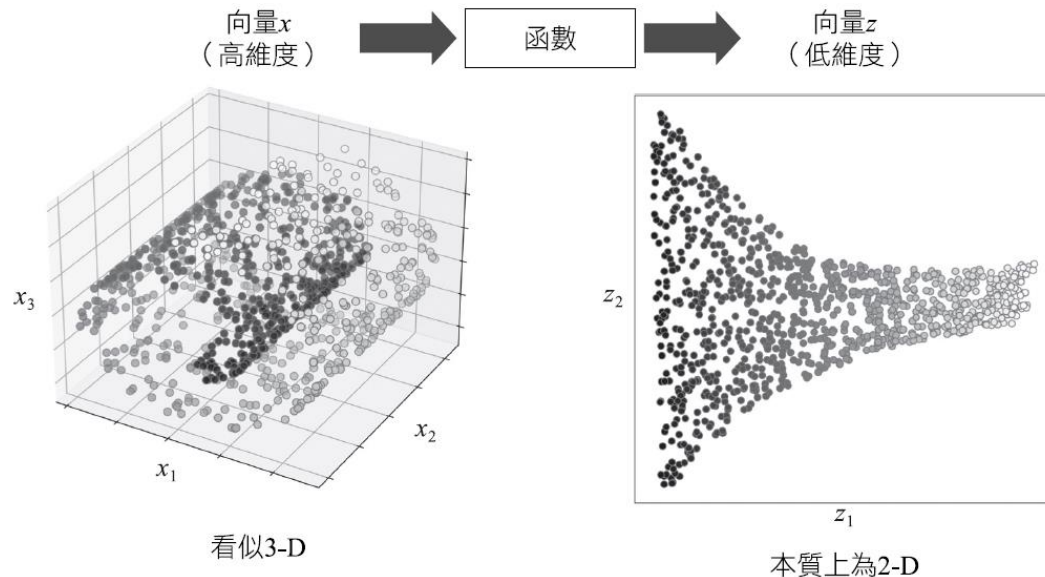
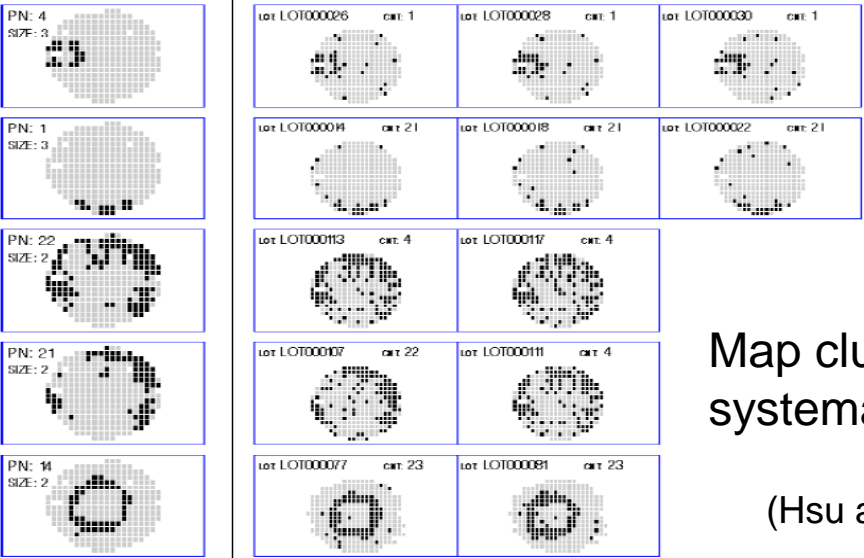
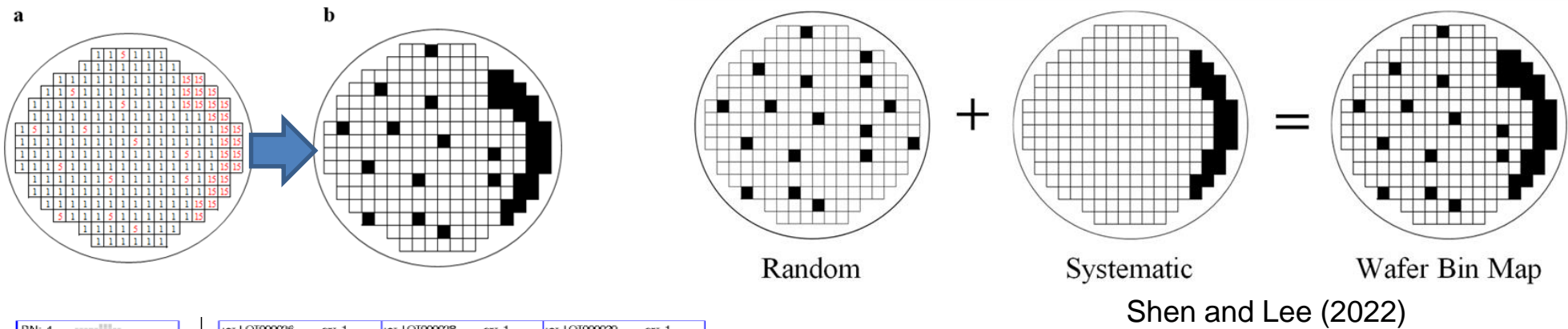


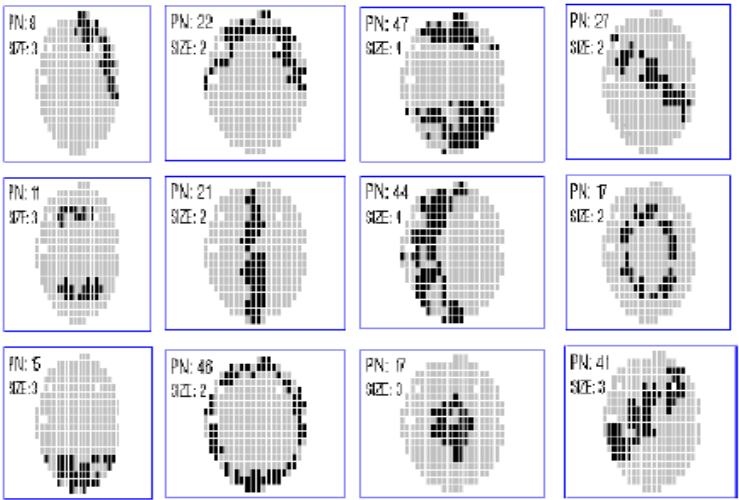
圖 3.6 維度縮減示意圖

非監督式學習

- 分群舉例：「瑕疵檢測」 (defect inspection)
 - 期望能將缺陷的「樣型」 (pattern) 分為多個集群，從而回溯剖析製程中可能發生的問題。

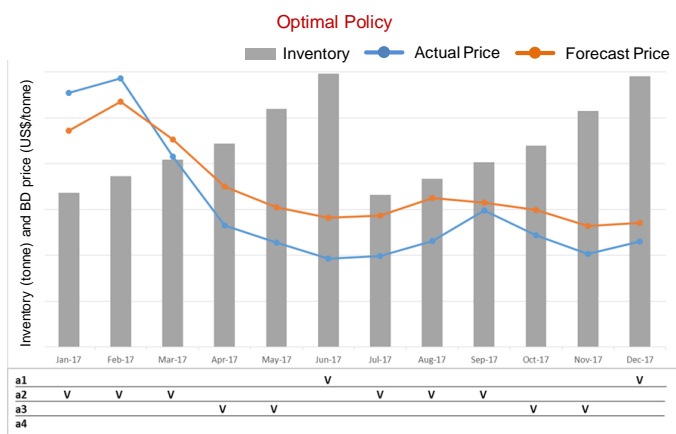
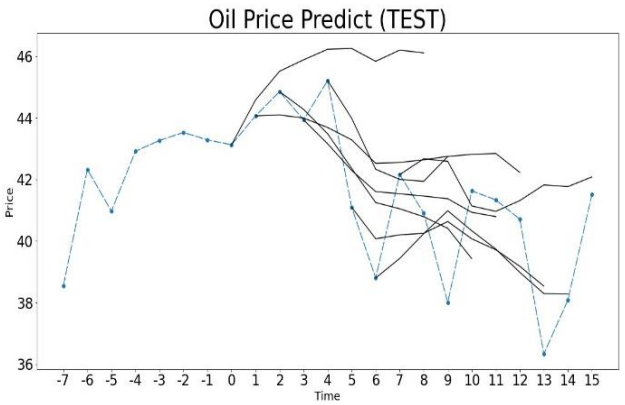


(Hsu and Chien, 2007)



強化學習

- 「強化學習」所做的並非是預測，而是建構一個連續決策的「代理人」(agent)，因此要最佳化「決策行動」(action)過程有賴於代理人與「環境」互動後所得到的「報酬」(reward)以及「狀態」(state)轉移的數據，而最佳化連續決策期望在長期行動下有最大的累積報酬。
- 常見的應用：自動駕駛汽車、AlphaGo圍棋以及機器人和工業自動化等。
- 設備保養策略
- 原料價格預測與採購決策



Lee, C.-Y., Chou, B.-J., and Huang, C.-F. 2022. Data science and reinforcement learning for price forecasting and raw material procurement in petrochemical industry. Advanced Engineering Informatics, 51, 101443.

□ 預測的關鍵是什麼？

□ 預測的本質

- 因果關係建構
- 重要因子的數據收集

□ 模型的本質

- 模型複雜度與預測力
 - 「模型配適」 (model fitting) 的目的在於刻畫與描述特徵與目標值之間的關係
 - 「模型複雜度」 (model complexity)
 - 「偏誤與變異」
- 維度的詛咒

□ 一般性、特殊性與模型複雜度

- 一般性（**generalization**）：具備較低的模型複雜度（例如簡單線性迴歸模型）。而在現實中，由於特徵與目標值之間的關係通常是複雜且非線性的，因此複雜度低的模型在預測準確度表現上較差
 - 對於數據有更「**全局**」（global）的理解
 - 對噪音較不敏感
 - 通常有較高的「**解釋力**」（interpretability）
- 特殊性（**specification**）：具備較高的模型複雜度（例如多項式的非線性模型）。然而特殊性過高的模型不經意地將噪音考慮進模型內，使得模型配適產生較大的變異。
 - 對於數據有更「**局部**」（local）的描述
 - 更精確地描繪特徵與目標值之間非線性的關係
 - 模型所具備的解釋力也相對較低
- 唯有在「一般性與特殊性」之間取得權衡（也就是合適的模型複雜度）才是合適且預測力強的模型

● 過度配適 (overfitting) 與欠缺配適 (underfitting)

- 說明了一般性與特殊性之間平衡不佳的情況。
- 「欠缺配適(underfit)」：一個過於「一般性」的模型
- 「過度配適(overfit)」：一個過於「特殊性」的模型。

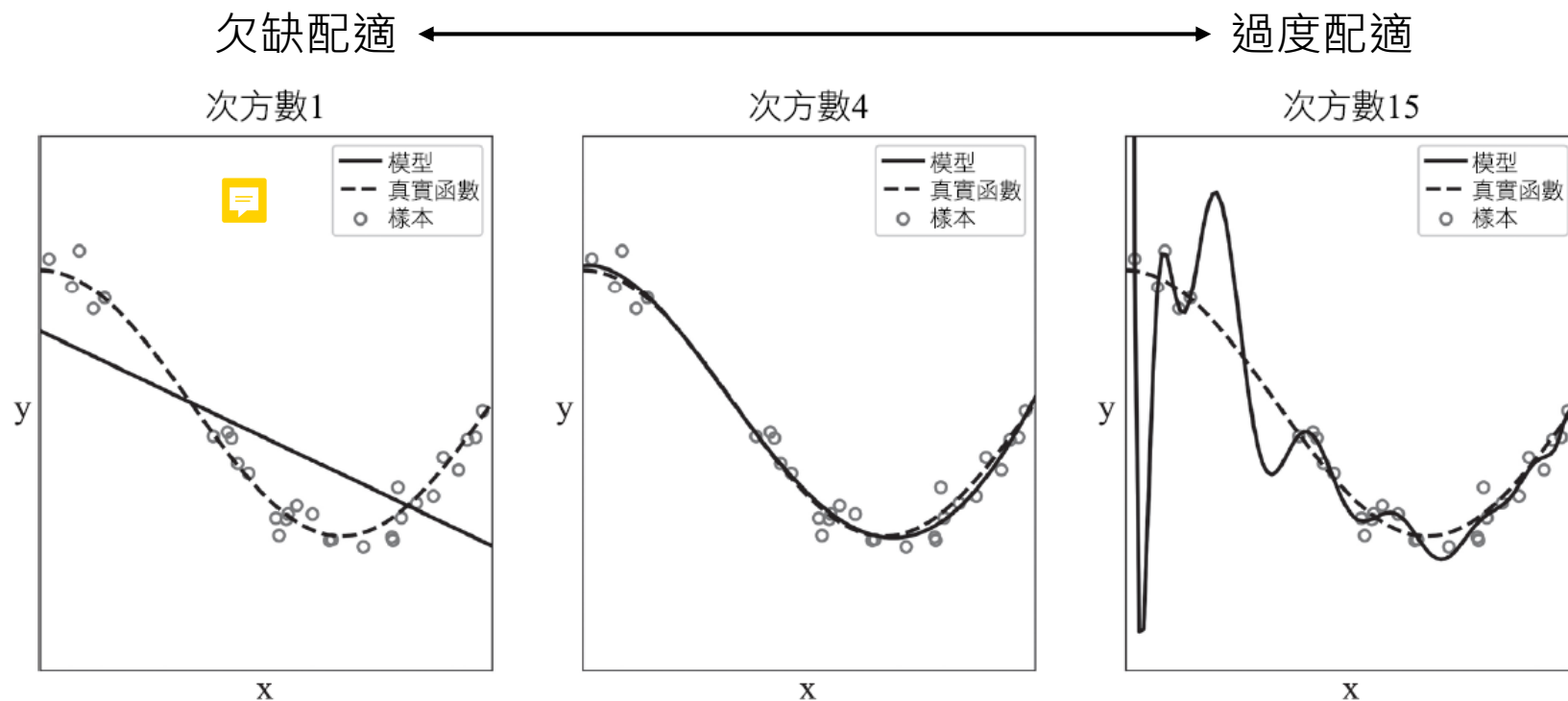


圖 3.8 過度配適與欠缺配適

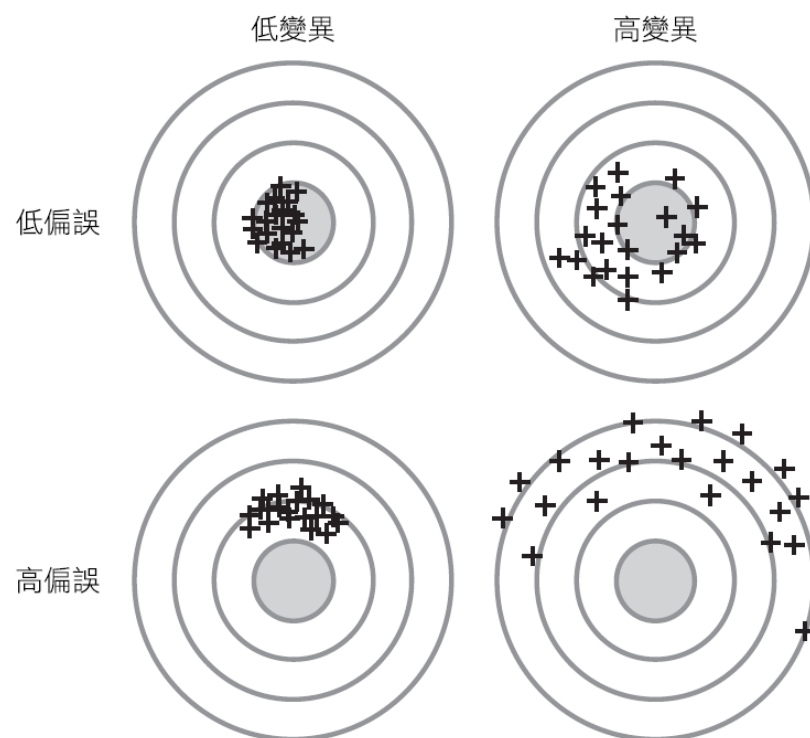
□ 模型複雜度與預測力

- 偏誤與變異的分解—如何量化模型的欠缺配適與過度配適？
- 分解期望預測誤差（expected prediction error）
 - 例如使用平方誤差損失（squared-error loss）拆解成「偏誤」（bias）與「變異」（variance）
 - 不可避免誤差（irreducible error）：不可避免的隨機噪音，理論上即是配適最佳模型的最小誤差
 - 偏誤：平均預測值與真實值的相差
 - 變異：預測值的變異

$$\begin{aligned}\text{Error}(x_0) &= E \left[\left(Y - \hat{f}(x_0) \right)^2 \middle| X = x_0 \right] \\ &= \sigma_\varepsilon^2 + [E[\hat{f}(x_0)] - f(x_0)]^2 + E \left[\hat{f}(x_0) - E[\hat{f}(x_0)] \right]^2 \\ &= \sigma_\varepsilon^2 + \text{Bias}^2 \left(\hat{f}(x_0) \right) + \text{Var} \left(\hat{f}(x_0) \right) \\ &= \text{不可避免誤差 (Irreducible Error)} + \text{偏誤 (Bias)}^2 + \text{變異 (Variance)}\end{aligned}$$

□ 偏誤與變異

- 在最小化期望預測誤差的目標下，我們期望能同時減低偏誤與變異，使得預測模型越集中且越靠近真實函數（靶心）越好
- 偏誤與變異的分解：與模型配適之間的關係
 - 偏誤：模型的「配適度」（偏誤越小配適度越好）
 - 變異：模型的「穩定度」（變異越小穩定度越好）
 - 我們所期望找出的「模型複雜度」，實際上就是模型「偏誤與變異」的權衡



模型的本質

● 偏誤與變異的權衡

- 在模型訓練時，模型複雜度將由低逐漸增高（例如集成學習與神經網路等），而模型配適的狀態將由欠缺配適到合適的模型，再到過度配適。
- 最小的期望預測誤差將落於「偏誤與變異」之間的權衡點。

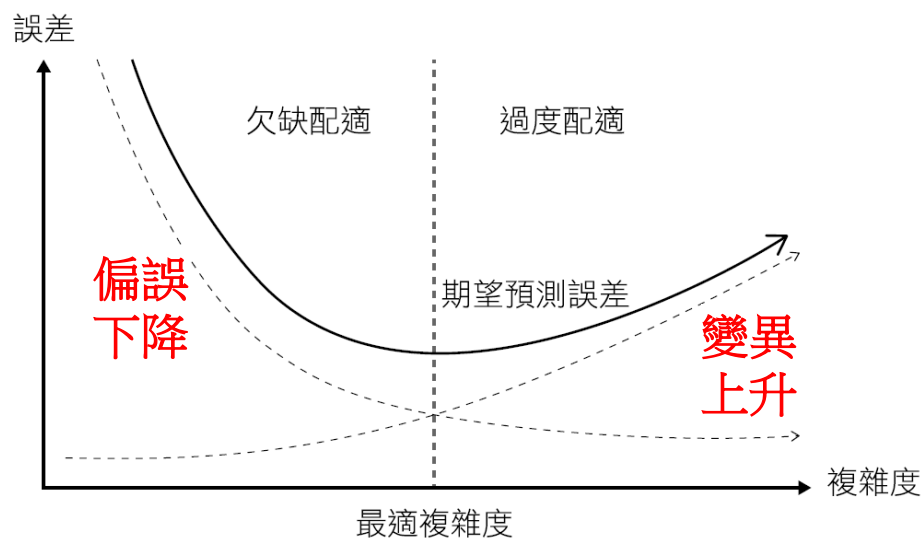
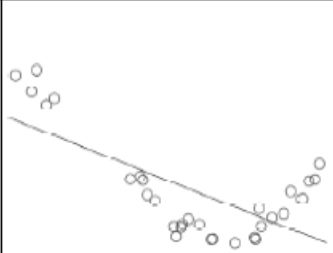
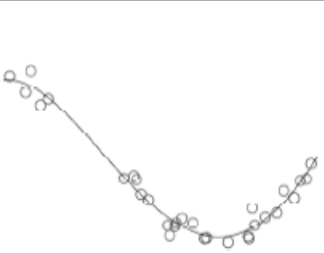
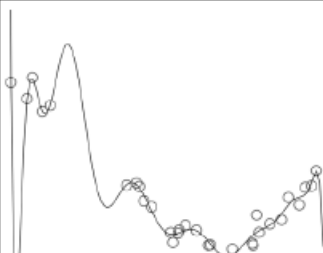
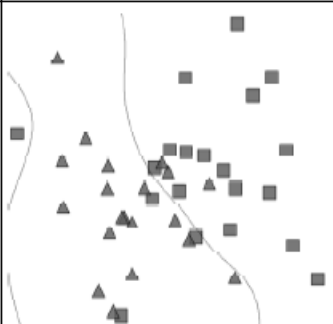


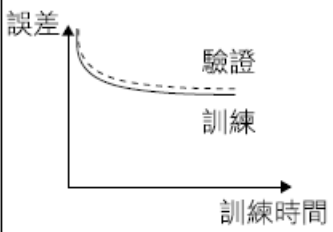
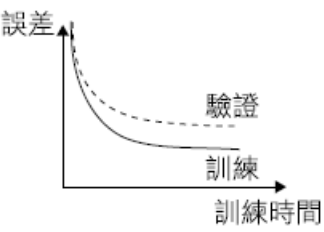
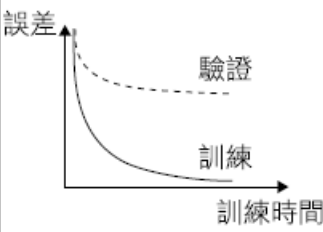


圖 3.10 模型複雜度由低至高的學習過程

	欠缺配適	合適配適	過度配適
特性	<ul style="list-style-type: none"> - 高訓練誤差 - 訓練誤差接近測試誤差 - 高偏誤 	<ul style="list-style-type: none"> - 訓練誤差略低於測試誤差 	<ul style="list-style-type: none"> - 低訓練誤差 - 訓練誤差甚低於測試誤差 - 高變異
迴歸			
分類			
深度學習			
建議	<ul style="list-style-type: none"> - 加強模型複雜度 - 增加更多特徵 - 增加訓練時間 		<ul style="list-style-type: none"> - 正規化 - 增加資料量

模型的本質

▣ 維度的詛咒 (curse of dimensionality)

- 定義：隨著函數（或模型）所需要估計未知參數的增加，樣本所需要的數量會呈現指數型的成長
- 若樣本數用 n 、變數個數用 p 表示，當 $p \gg n$ ，也就是維度過大（或所需要估計的未知參數過多）但樣本數不足的情況下，此時機器學習常發生：
 - 花很長時間學習但演算法不易收斂
 - 收斂時出現多重解 (multiple solutions) 或過度配適
- 維度的詛咒造成數據在幾何空間呈現稀疏性 (sparsity)，此時傳統的距離量測的有效性變差，使迴歸或分類器不易收斂或多重解的可能

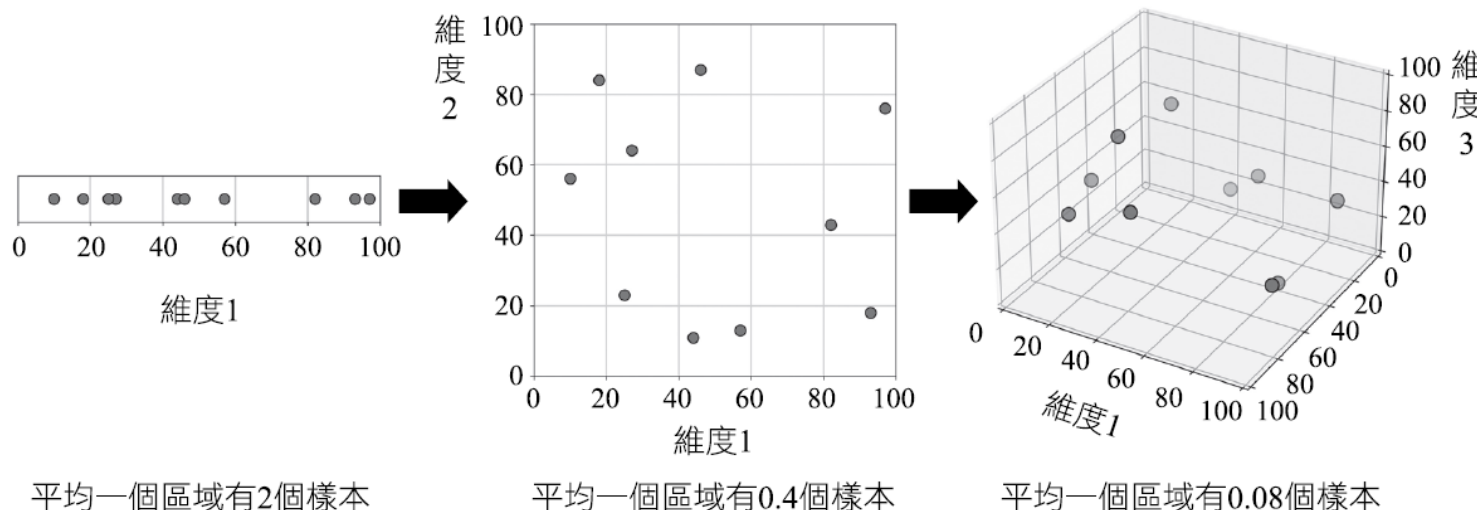
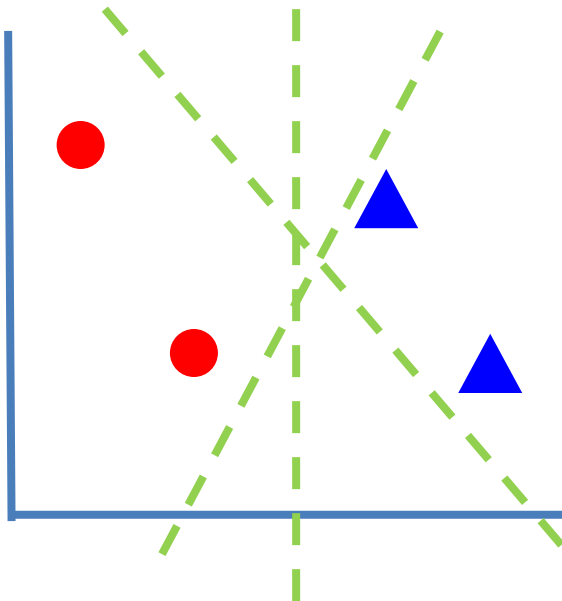


圖 3.11 數據維度與模型的預測的準確度

□ 1-D classification



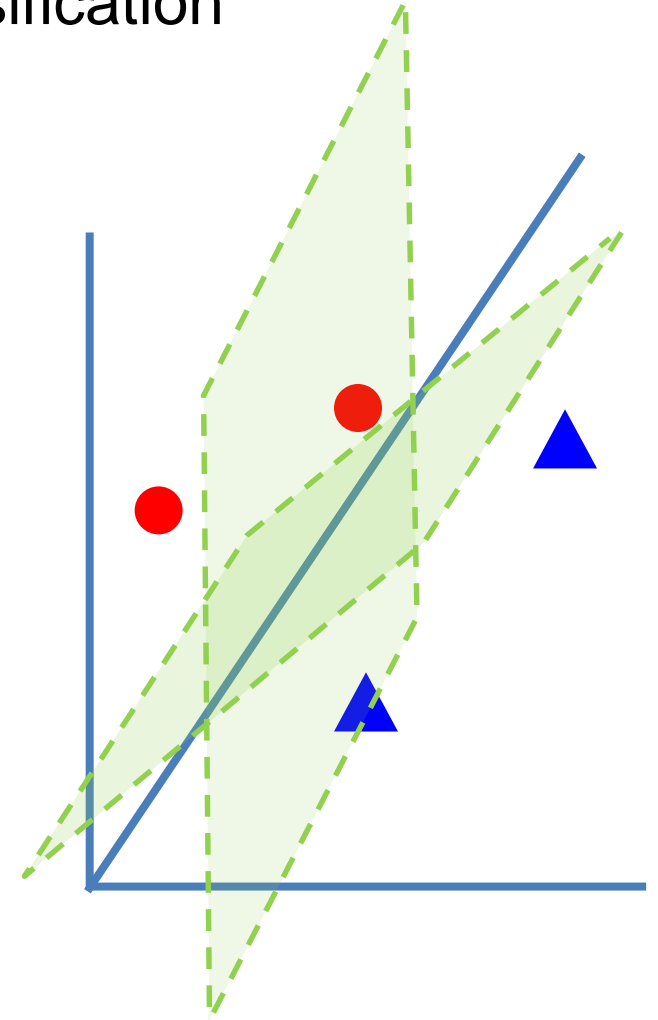
□ 2-D classification



Slopes are different.
For linear classifier,
which one is correct?

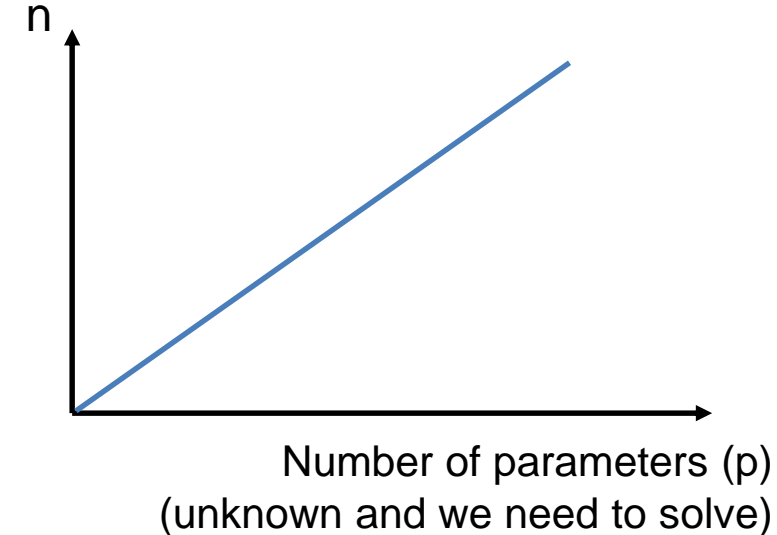
For nonlinear classifier,
you have more choices.

□ 3-D classification



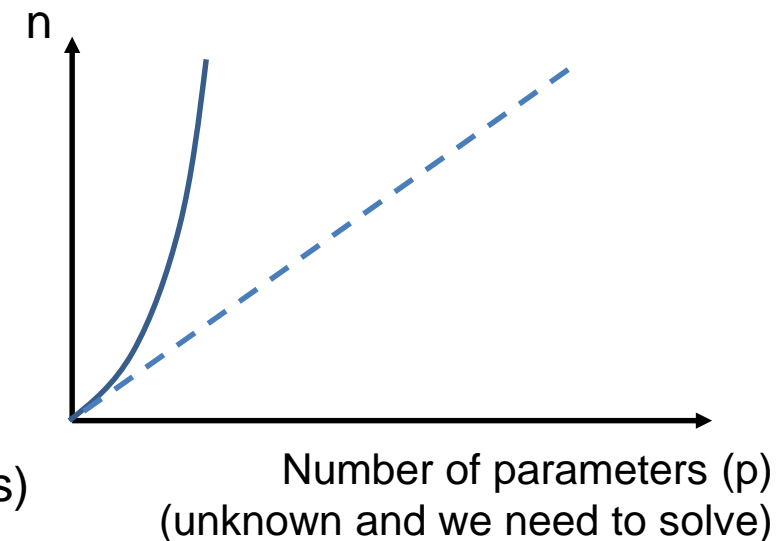
□ Curse of Dimensionality

- system of equations
- $\begin{cases} 2x + y = 12 \\ x + 3y = 11 \end{cases}$
- 2 parameters (p) need 2 equations (n)



- ML/DS 
 - # of obs. (n) required exponentially grows...

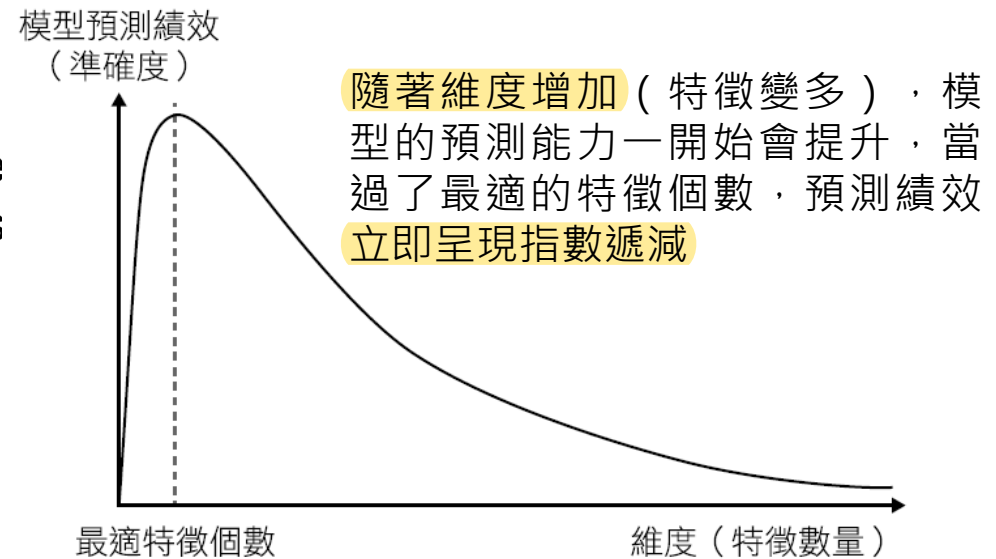
- Reason:
 - Nonlinear (eg. quadratic shows + and - roots)
 - Combination (eg. x, y, xy)
 - NN/DL system (eg. fully-connected arc/weights)



□ Problem

- In practice, the **curve of learning** performance w.r.t. the feature dimension looks like this

For a **fixed sample size n** , there is an optimal number of features to use...



- Address “Curse of Dimensionality”
 - The number of observations required exponentially grows to estimate the function or model parameters.
- Volume of Dataset (nodes N , weights W , # of samples I)
 - Widrow’s rule of thumb (Widrow, 1987): $I \geq 10W$
 - Vapnik-Chervonenkis (**VC**) Dimension: $I \geq O\left(\frac{W}{\varepsilon} \log \frac{N}{\varepsilon}\right)$

□ 評估模型的主要方法以及重要的評估指標

- 先針對「監督式學習」。
- 分割數據
 - 訓練集 (training dataset)：作為訓練、配適模型所使用的數據，一般會佔原有數據多數的比例。
 - 驗證集 (validation dataset)：作為調整模型超參數 (hyperparameter) 與複雜度的數據 (此議題又被稱為「超參數最佳化」)
 - 測試集 (testing dataset)：作為模型預測結果評估的數據，此數據僅作為評估模型用，不會參與模型訓練與配適的任何一個環節。
 - 分割比例會基於樣本數與變數個數而有所差異，常用為70:15:15或8:1:1
- 評估模型
 - 樣本外誤差 (extra-sample error)
 - 交叉驗證 (cross validation)
 - 包外誤差 (out-of-bag)
 - 樣本內誤差 (in-sample error)
 - Cp統計量
 - 赤池資訊量準則 (Akaike information criterion, AIC)
 - 貝氏資訊量準則 (Bayesian information criterion, BIC)

□ 訓練、驗證與測試集

- 期望能最小化的「期望預測誤差」實際上就是「測試誤差」（testing error） $\text{Error}_{\text{Test}}$ ，是在給定的訓練集下所計算出的期望誤差

- $\text{Error}_{\text{Test}}(\mathcal{T}) = \mathbb{E} \left[L(Y, \hat{f}(X)) \mid \mathcal{T} \right]$

- 其中，特徵 X 與目標值 Y 是由它們來自母體的聯合分配隨機抽出的樣本， L 為損失函數（loss function）， τ 為給定的訓練集

- 在模型訓練時，實際上是最小化「訓練誤差」（training error） $\text{Error}_{\text{Train}}$ ，是直接由訓練集所計算出的平均損失

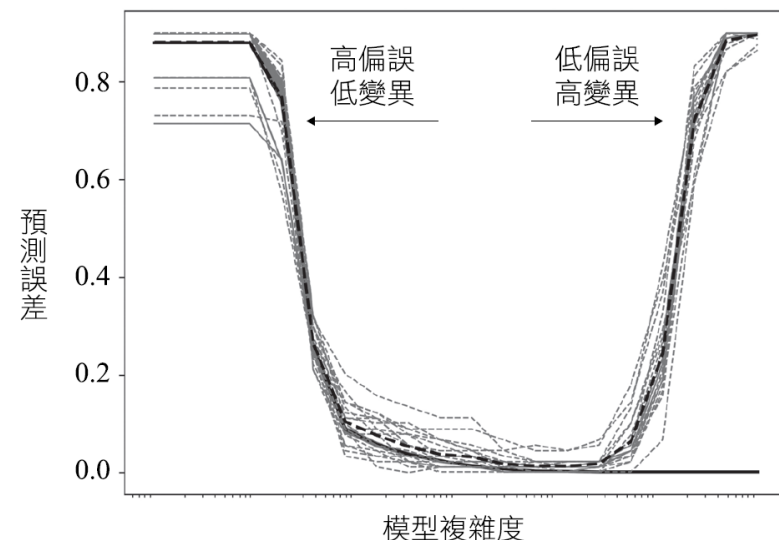
- $\text{Error}_{\text{Train}} = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}(x_i))$

- 特徵 x_i 與目標值 y_i 為訓練集的真實數據。

- 在不同的模型複雜度下

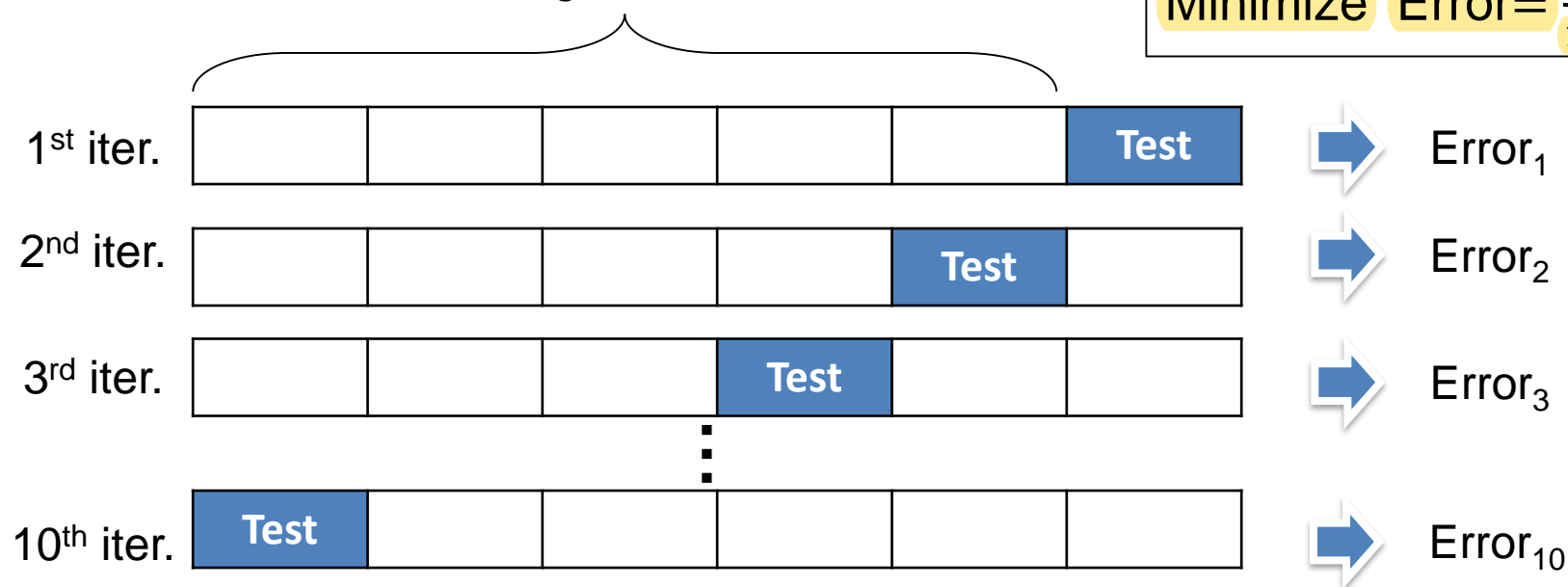
- 「訓練誤差」是在固定的數據下試圖最小化損失函數，因而其誤差只會不斷地遞減。

- 在模型複雜度低時，有較高的偏誤與較低的變異，產生「欠缺配適」；而在模型複雜度高時，有較低的偏誤與較高的變異，產生「過度配適」



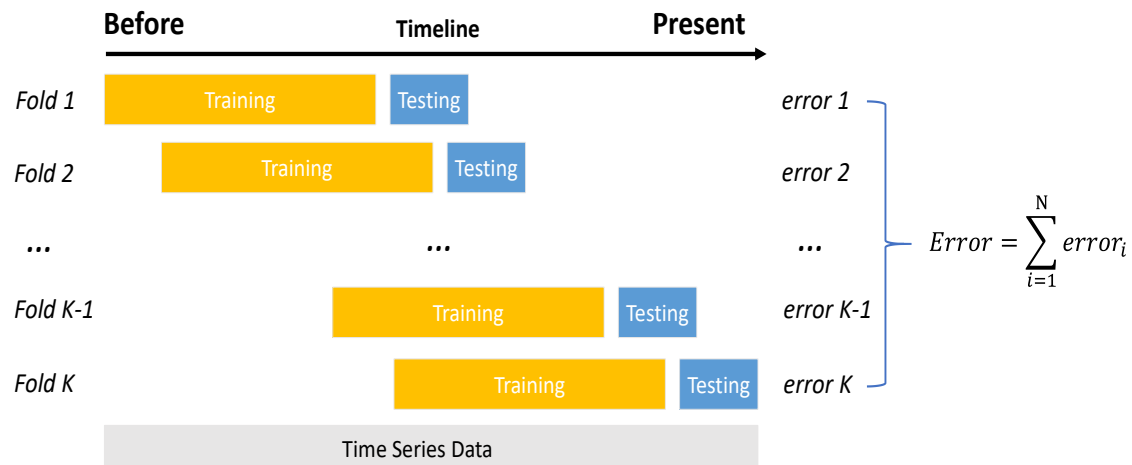
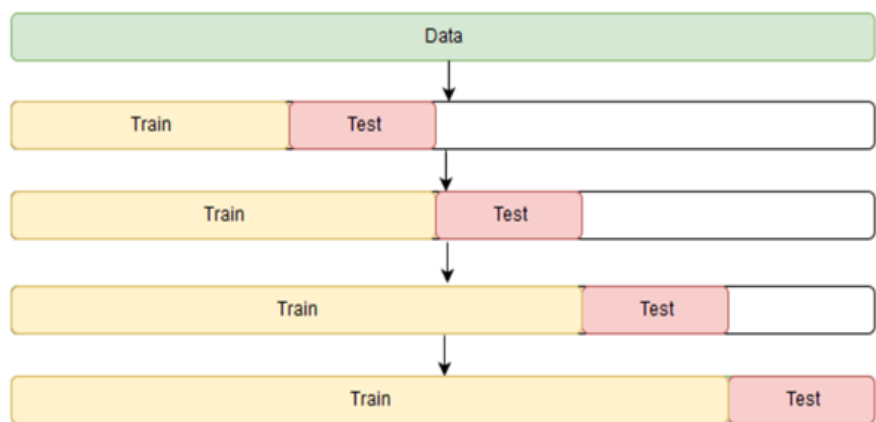
模型評估與挑選

□ K-Fold Cross Validation




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

□ Time-Series Nested CV/ Rolling Time Window



□ 樣本外誤差 (extra-sample error)

- 直接估計「樣本外誤差」，也就是「測試誤差」。
- K 折交叉驗證 (K-fold cross-validation, CV)
 - 將數據切分成多個等份，輪流作為訓練集與驗證集，計算出K個「預測誤差」，最後再將這K個誤差取平均近似「期望預測誤差」
 - $CV(\hat{f}) = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}^{-\kappa(i)}(x_i))$
 - 一般我們會設定「K折交叉驗證」中的折數為5至10折。
 - 當 $K = N$ 時，就是「留一交叉驗證」(leave-one-out validation, LOOCV)
- 重複隨機抽樣的驗證 (repeated random sub-sampling validation；又稱蒙地卡羅交叉驗證 Monte Carlo CV)
- 時間序列巢狀交互驗證 (time series nested CV)：根據時間序列的數據可使用
- 群K折交叉驗證 (Group K-fold CV)：當數據非滿足獨立同分布 (i.i.d.) 假設且樣本有相依性時可用
- Stratified sampling CV 


□ 樣本外誤差 (extra-sample error)

● 包外誤差 (out-of-bag) :



- 在某些「集成學習」(ensemble learning) 方法中 (例如隨機森林)，我們可能透過「拔靴法」(bootstrap) 的方式對樣本進行重複抽樣 (resampling) 以產生多個子數據集 (sub-dataset)，再針對每個子數據集各別訓練模型，最後再由多個模型分別預測後加以彙總。
- 由於重複抽樣會造成某些樣本取用而有些樣本未使用，有取用到的樣本會形成訓練集拿來訓練模型，未取用到的樣本便可用以做驗證集，來評估模型訓練的績效。
- 那些未取用的樣本叫做「包外」(out-of-bag, OOB) 數據，而利用包外數據來計算的誤差，稱為「包外誤差」(out-of-bag error)。

□ 樣本內誤差

- 若我們想直接以「訓練集」估計預測誤差，可以計算「樣本內誤差」
 - $\text{Error}_{\text{in}}(\mathcal{T}) = \frac{1}{N} \sum_{i=1}^N \mathbb{E} \left[L(Y, \hat{f}(x_i)) \mid \mathcal{T} \right]$
 - 其中特徵 x_i 為訓練集真實數據，隨機變數 Y 為在固定特徵下目標值潛在的可能
- 若將「樣本內誤差」與前述提及的「訓練誤差」 $\text{Error}_{\text{Train}}$ 相比，「訓練誤差」存在的問題在於其未考慮樣本的隨機性，過於樂觀地低估實際的誤差，因此我們可以將低估的「樂觀誤差」 Error_{op} 定義
 - $\text{Error}_{\text{op}} \equiv \text{Error}_{\text{in}} - \text{Error}_{\text{Train}}$
 - 而實際上，這個「樂觀誤差」的期望值可藉由計算實際目標值 y 與預測目標值 \hat{y} 的共變異數得到
 - $\text{op} = \mathbb{E}[\text{Error}_{\text{op}}] = \frac{2}{N} \sum_{i=1}^N \text{Cov}(\hat{y}_i, y_i)$ 
- 也就是說，對於誤差的樂觀程度，取決於預測值 \hat{y} 究竟受實際值 y 影響的大小，當影響過大時（與訓練集過於接近），將發生「過度配適」的情形。在量化「期望樂觀誤差」後，便能推導「樣本內誤差」期望值的估計
 - $\mathbb{E}[\text{Error}_{\text{in}}] = \mathbb{E}[\text{Error}_{\text{Train}}] + \underbrace{\frac{2}{N} \sum_{i=1}^N \text{Cov}(\hat{y}_i, y_i)}_{\text{模型複雜度}}$

□ 樣本內誤差

- 樣本內誤差估計是由訓練誤差與模型複雜度(後者為懲罰項)的加總

- $\widehat{\text{Error}}_{\text{in}} = \text{Error}_{\text{Train}} + \widehat{\text{op}}$

- 當「模型複雜度」越高時，對於誤差估計的懲罰越大。

- Cp統計量

- $C_p = \text{Error}_{\text{Train}} + 2 \cdot \frac{d}{N} \hat{\sigma}_\varepsilon^2$



- 其中， $\text{Error}_{\text{Train}}$ 是以平均平方誤差作為損失函數所計算出的訓練誤差， d 為特徵數， $\hat{\sigma}_\varepsilon^2$ 為殘差的變異數。此式以特徵數 d 與樣本數 N 來衡量「模型複雜度」與「預測不確定性」作為懲罰項：一方面，當使用越多的特徵時，則代表模型複雜度越高；另一方面，當有越多樣本時，會降低模型預測不確定性。

- 「赤池資訊量準則」(AIC)與「貝氏資訊量準則」(BIC)

- 在相同的樣本數下，BIC所受的懲罰會比AIC所受的來得大，總結來說，在估計「樣本內誤差」時，BIC相較於AIC更加保守，因而經最小化後所挑選的模型較為「一般化」，相反地，樂觀的AIC所挑選的模型則較為「特殊化」，產生「過度配適」的可能也較高。

- 實務上，「樣本內誤差」可能受於模型的限制(「概似函數」與「模型複雜度」不存在或難以評估)。因此，絕大多數均採用「交叉驗證」作為模型評估方法。

❑ 損失函數 (loss function) 與評估指標 (evaluation metric) 差異

- 兩者都是用來估計模型誤差的函數或指標
- 模型訓練時使用損失函數，模型訓練後使用評估指標
- 損失函數
 - 訓練與配適模型時最小化的目標函數，作為調整模型複雜度與權衡偏誤與變異
 - 其函數特性常需要滿足一定的要求，例如具備「微分特性」、「連續性」等性質。通常只能比較相對性，無法解釋絕對的誤差大小
- 評估指標
 - 用以衡量訓練後的模型表現，評估最終結果，需具備「公平性」與「比較性」

❑ Regression

- 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, DICE, MAP
- DCG(discounted cumulative gain), NDCG, P@k (precision at k)

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

□ 分類結果- Confusion Matrix

- 以測試組資料的結果來客觀評估較佳的分類模型
- 假設是二元分類模型，也就是只有兩種類別的模型，例如：
 - （良品／不良品）（陽性／陰性）（有病／沒病）（發生／不發生）等。

		預測類別	
		類別1 (不良品)	類別2 (良品)
實際類別	類別1 (不良品)	TP (true positive)	FN (false negative) (Type II error) (miss)
	類別2 (良品)	FP (false positive) (Type I error) (false alarm)	TN (true negative)

□ 根據分類結果，可計算出**正確率** $Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$

模型評估與挑選

□ 靈敏度 (Sensitivity, Recall)

- 為實際類別1(不良品)當中，被正確預測的比率

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

- Miss (漏報) Rate = 1 - Sensitivity

□ 特異度 (Specificity)

- 為實際類別2(良品)當中，被正確預測的比率

$$\text{Specificity} = \frac{TN}{TN+FP}$$

- False Alarm (假警報) Rate = 1 - Specificity

□ 精確度 (Precision)

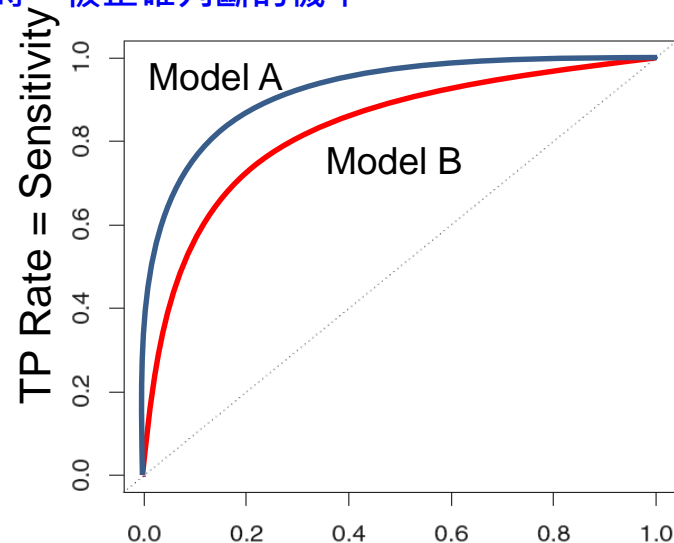
- 為預測類別1(不良品)當中，被正確預測的比率

$$\text{Precision} = \frac{TP}{TP+FP}$$

		預測	
		類別1	類別2
實際	類別1	TP	FN
	類別2	FP	TN

□ ROC曲線 (Receiver Operating Characteristic curve)

描述當資料屬於類別1(不良品)時，被正確判斷的機率



FP Rate = 1 - Specificity

當資料不屬於類別1(不良品)時，被誤判的機率 (Type I error)

一般來說，**右下方的面積愈大**，該模型的分類效果愈佳！

▣ 精度/良率預測



● Model performance (128個觀測值)

Model A		預測	
		Bad	Good
實際	Bad	61	7
	Good	29	31

Model B		預測	
		Bad	Good
實際	Bad	47	21
	Good	7	53

	Testing			
	Accuracy	TP Rate	TN Rate	AUC
Model A	71.9%	89.7%	51.7%	70.2%
Model B	78.1%	69.1%	88.3%	78.9%

AUC: Area under the Curve of ROC

□ 定義人工智慧與機器學習的關聯

- 歷史與範疇

□ 為了瞭解模型的本質

- 模型的複雜度與預測能力
- 「過度配適」與「欠缺配適」
- 偏誤與變異的權衡
- 維度的詛咒

□ 模型評估與挑選

- 樣本外誤差
- 樣本內誤差
- 評估指標

□ 留意「數據不平衡」議題的影響才能確保模型的預測準確度

Thanks for your attention



NTU Dept. of Information Management
name: 李家岩 (FB: Chia-Yen Lee)
phone: 886-2-33661206
email: chiayenlee@ntu.edu.tw
web: <https://polab.im.ntu.edu.tw/>