徐唯哲
Paul Wei-Che HSU

機器學習與多體學資料分析

Machine learning and integrative analysis of multi-omics data

分子與基因醫學研究所
專案技術助研究員

2023/9/15

課程相關資料

https://drive.google.com/drive/folders/1JwKLgLGLvscr9fVsAKc_AacJyYgkEfZ-?usp=sharing

群聚（Crowding）

引發的災難

蝗災

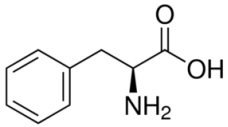
蝗災的起源: CYP305M2

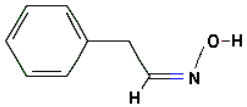
蚱蜢

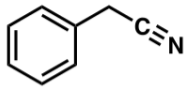
Crowding

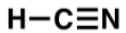
Active CYP305M2

蝗蟲









L-Phenylalanine (苯丙胺酸) $\xrightarrow{\text{CYP305M2}}$ Z-Phenylacetaldoxime (苯乙醛肟) $\xrightarrow{\text{CYP71}}$ Phenylacetonitrile, PAN (苯乙腈) $\longrightarrow$ HCN (氰化氫)

細胞色素P450　CYP450 (Cytochrome P450)

CYP enzymes cover about **80%** of oxidative metabolism (Zhao et al., 2021)

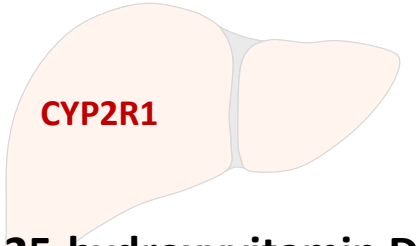
70-80% of clinical drugs are metabolized by CYP enzymes (Zanger and Schwab., 2013)

藥物學名	適應症	相關基因
Clomipramine	憂鬱症	CYP2D6
Clopidogrel	中風，心肌梗塞，周邊動脈血管疾病	CYP2C19
Clozapine	精神分裂症，	CYP2D6
Codeine	鎮咳、鎮痛	CYP2D6
Daclatasvir	慢性 C 型肝炎	IFNL3
Darifenacin	膀胱過動症	CYP2D6
Dasabuvir, Ombitasvir, Paritaprevir, and Ritonavir	慢性 C 型肝炎	IFNL3
Desipramine	憂鬱症	CYP2D6
Desvenlafaxine	台灣未上市	CYP2D6
Deutetrabenazine	遲發性運動障礙	CYP2D6
Dexlansoprazole	糜爛性食道炎	CYP2C19
Dextromethorphan & Quinidine	台灣未上市	CYP2D6
Diazepam	焦慮	CYP2C19
Donepezil	阿滋海默症	CYP2D6
Doxepin	憂鬱症，焦慮症，睡眠障礙，搔癢症	CYP2C19
Doxepin	憂鬱症，焦慮症，睡眠障礙，搔癢症	CYP2D6
Dronabinol	癌症化療止吐，刺激食慾	CYP2C9
Drospirenone & Ethinyl Estradiol	避孕	CYP2C19
Duloxetine	憂鬱症，糖尿病週邊神經痛	CYP2D6
Elagolix	子宮內膜異位症	SLCO1B1

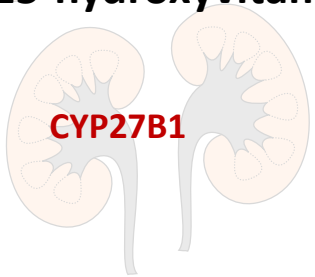
藥物學名	適應症	相關基因
Lesinurad	痛風	CYP2C9
Lofexidine	毒品戒斷	CYP2D6
Lusutrombopag	血小板減少症	F5
Meclizine	暈車，暈船，暈機	CYP2D6
Mercaptopurine	急性白血病，慢性骨髓性白血病	NUDT15
Mercaptopurine	急性白血病，慢性骨髓性白血病	TPMT
Metoclopramide	嘔吐，胃食道逆流，糖尿病引起之胃腸異常	CYP2D6
Metoprolol	高血壓，心絞痛，心衰竭，心肌梗塞，心律不整	CYP2D6
Mirabegron	膀胱過動症	CYP2D6
Modafinil	猝睡症	CYP2D6
Nebivolol	高血壓	CYP2D6
Nefazodone	憂鬱症	CYP2D6
Nortriptyline	憂鬱症	CYP2D6
Ombitasvir, Paritaprevir, and Ritonavir	慢性 C 型肝炎	IFNL3
Omeprazole	胃潰瘍，十二指腸潰瘍，胃食道逆流	CYP2C19
Ondansetron	癌症化療止吐	CYP2D6
Ospemifene	女性停經後骨質疏鬆	CYP2C9
Palonosetron	癌症化療止吐	CYP2D6
Pantoprazole	胃潰瘍，十二指腸潰瘍，胃食道逆流	CYP2C19
Paroxetine	憂鬱症，恐慌症，焦慮症，強迫症	CYP2D6
Peginterferon Alfa-2b	慢性 C 型肝炎	IFNL3
Perphenazine	鎮靜	CYP2D6

Vitamin D_2

Vitamin D_3



25-hydroxyvitamin D

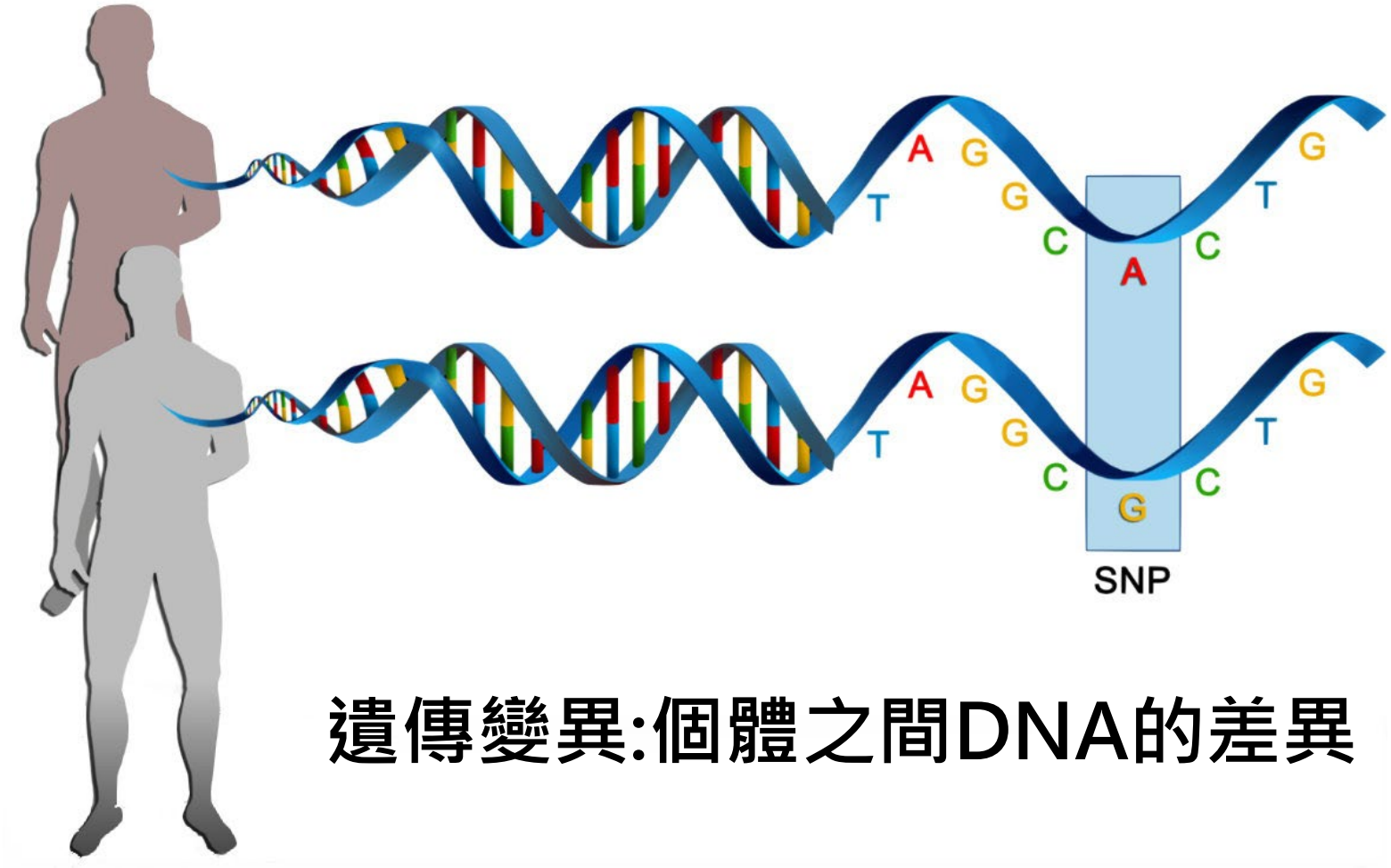


l, 25-dihydroxyvitamin D

1,25(OH)2D

為什麼每個人獲得的維他命D的能力不一樣?

Genetic Variants



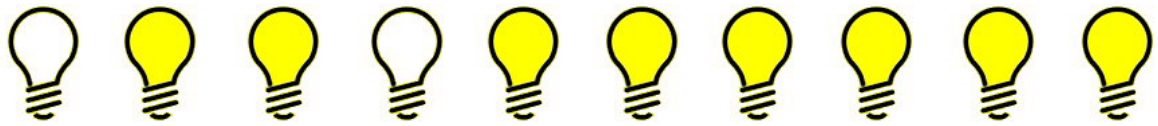
遺傳變異:個體之間DNA的差異

Phenotype Variability

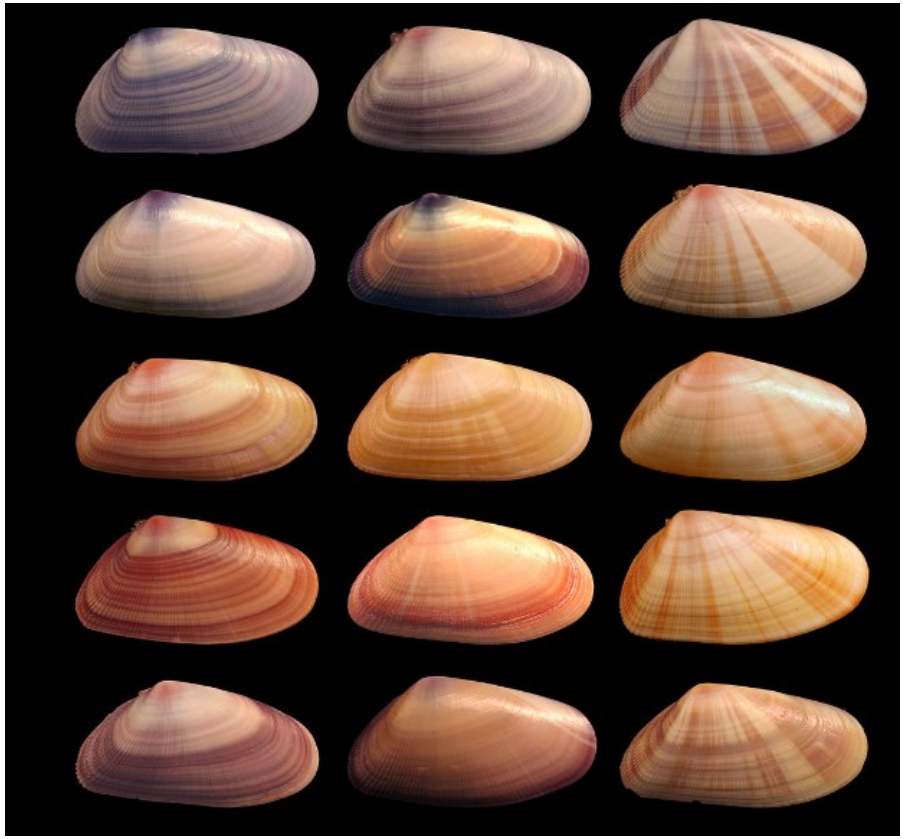
表型多變性

- **外顯率** (Penetrance)

相同基因突變下，造成phenotype表現的比率



80%



(https://www.wikiwand.com/en/Phenotype)

- **表現度** (Expressivity)

相同基因突變下，phenotype表現的程度

CYP450, CYP (細胞色素P450)

PharmVar
Pharmacogene Variation Consortium

https://www.pharmvar.org/

23 Pharmacogenes

21 CYP genes

CYP1 family

CYP1A1
CYP1A2
CYP1B1

CYP2 family

CYP2A6
CYP2A13
CYP2B6
CYP2C8
CYP2C9
CYP2C19
CYP2D6
CYP2E1
CYP2F1
CYP2J2
CYP2R1
CYP2S1
CYP2W1

CYP3 family

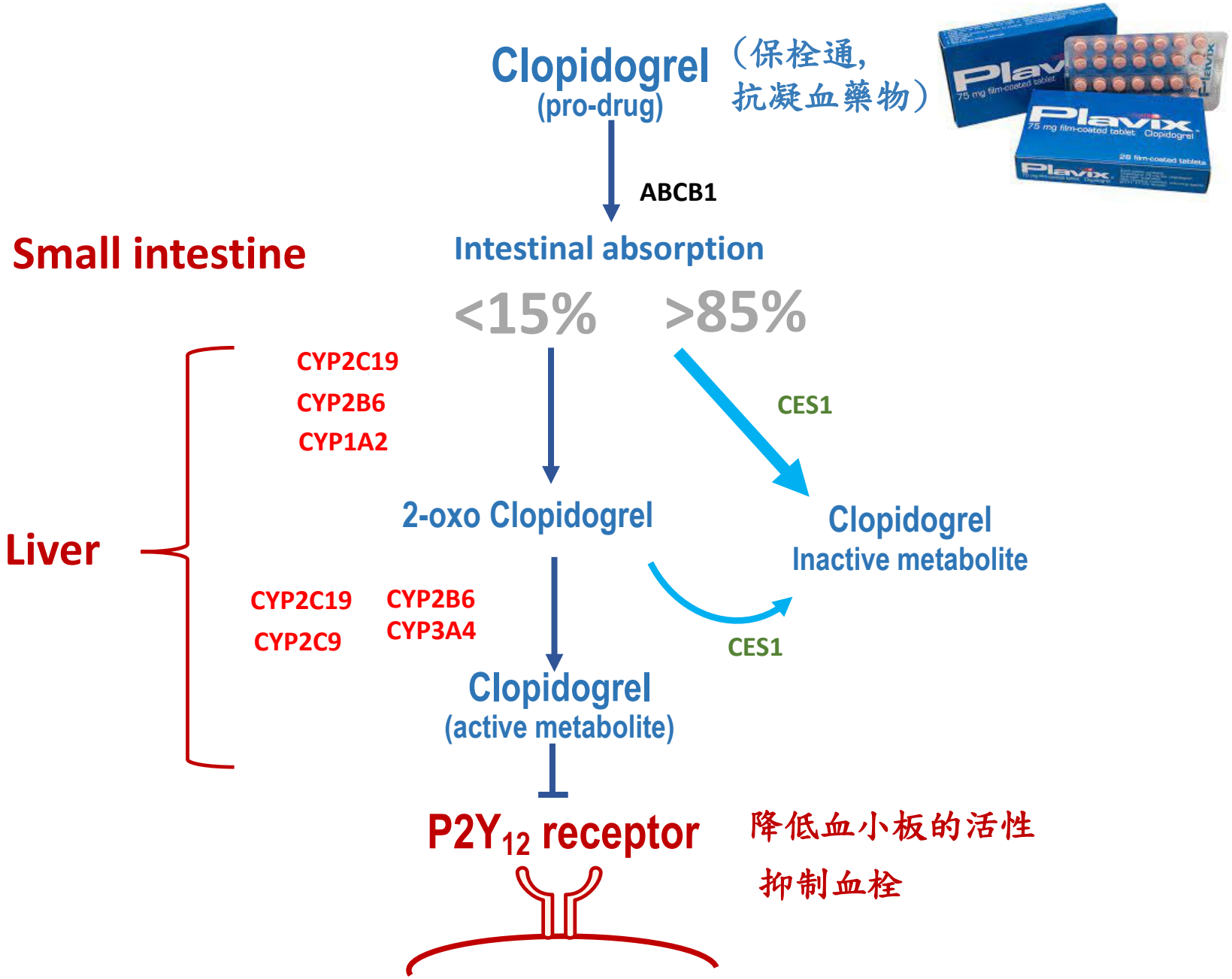
CYP3A4
CYP3A5
CYP3A7
CYP3A43

CYP4 family

CYP4F2

Non-Cytochrome P450

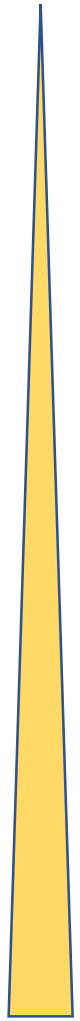
DPYD
NUDT15


Clopidogrel
(pro-drug)
（保栓通，
抗凝血藥物）
ABCB1
Small intestine
Intestinal absorption
<15%
>85%
CYP2C19
CYP2B6
CYP1A2
CES1
2-oxo Clopidogrel
Clopidogrel
Inactive metabolite
Liver
CYP2C19
CYP2C9
CYP2B6
CYP3A4
CES1
Clopidogrel
(active metabolite)
P2Y12 receptor
降低血小板的活性
抑制血栓

Platelet reactivity unit (PRU)

血小板高反應性

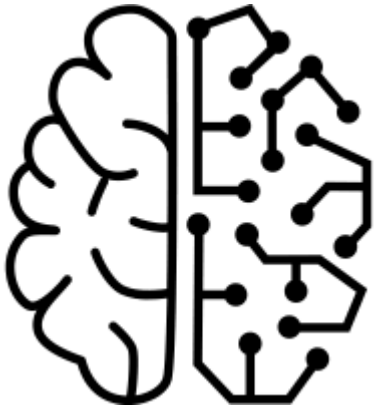
PRU
低
（抗凝血效果佳）



高
（抗凝血效果差）

sample_id	PRU	CYP2C19	CYP2B6	CYP2C9	CYP3A4	CYP3A5
1	1	*1B/*1B	*1/*4	*1/*1	*1/*1	*3/*3
2	6	*1B/*1B	*1/*1	*1/*1	*1/*1	*1/*1
3	6	*2/*2	*1/*6	*1/*1	*1/*1	*1/*3
4	6	*1B/*2	*1/*6	*1/*1	*1/*1	*3/*3G
5	102	*1B/*1B	*1/*6	*1/*1	*1/*1	*3/*3
6	106	*1B/*1B	*1/*2	*1/*1	*1/*5	*1/*3
7	119	*1B/*1B	*1/*1	*1/*1	*1/*1	*3/*3
8	121	*1B/*2	*1/*1	*1/*1	*1/*18	*1/*3
9	122	*1B/*1B	*1/*1	*1/*1	*1/*1	*3/*3
10	131	*1B/*2	*1/*6	*1/*1	*1/*1	*1/*3
11	137	*1B/*3	*6/*6	*1/*1	*1/*1	*3/*3
12	142	*1B/*3	*1/*1	*1/*1	*1/*1	*3/*3
13	143	*1B/*2	*1/*1	*1/*1	*1/*1	*1/*3
14	147	*1B/*1B	*1/*1	*1/*1	*1/*1	*3/*3
15	154	*1B/*2	*1/*6	*1/*1	*1/*1	*3/*3
16	158	*1B/*2	*1/*4	*1/*1	*1/*1	*1/*1
17	160	*1B/*1B	*1/*4	*1/*1	*1/*1	*1/*1
18	171	*1B/*2	*1/*4	*1/*1	*1/*1	*3/*3
19	175	*1B/*2	*2/*26	*1/*1	*1/*1	*3/*3
20	179	*1B/*2	*1/*6	*1/*1	*1/*1	*3/*3
21	188	*1B/*1B	*1/*1	*1/*1	*1/*1	*3/*3
22	197	*2/*2	*1/*1	*1/*1	*1/*1	*1/*3
23	197	*2/*2	*2/*6	*1/*1	*1/*1	*3/*3
24	198	*1B/*1B	*1/*4	*1/*1	*1/*1	*3/*3
25	202	*1/*2	*1/*1	*3/*13	*1/*1	*1/*3
26	204	*1/*1B	*1/*6	*1/*3	*1/*1	*1/*3
27	208	*1B/*1B	*1/*4	*1/*1	*1/*1	*1/*1
28	211	*1/*1+rs4	*1/*4	*1/*3	*1/*1	*1/*3
29	213	*1B/*2	*1/*1	*1/*1	*1/*1	*1/*3
30	215	*1/*1B	*1/*1	*1/*1	*1/*1	*3/*3
31	218	*1B/*2	*1/*6	*1/*27	*1/*1	*1/*3
32	225	*1B/*3	*1/*4	*1/*1	*1/*1	*1/*3
35	240	*1B/*2	*1/*1	*1/*1	*1/*1	*1/*3
36	242	*1B/*2	*1/*1	*1/*1	*1/*1	*1/*3
37	249	*1B/*3	*4/*6	*1/*1	*1/*1	*3/*3
38	257	*3/*3	*1/*6	*1/*1	*1/*1	*3/*3
39	310	*1B/*1B	*4/*6	*1/*1	*1/*18	*1/*1
40	325	*1B/*3	*1/*4	*1/*1	*1/*1	*3/*3
41	356	*1B/*2	*1/*6	*1/*1	*1/*1	*1/*3
42	467	*2/*2	*1/*6	*1/*1	*1/*1	*1/*3

Pharmacogenes



Machine learning

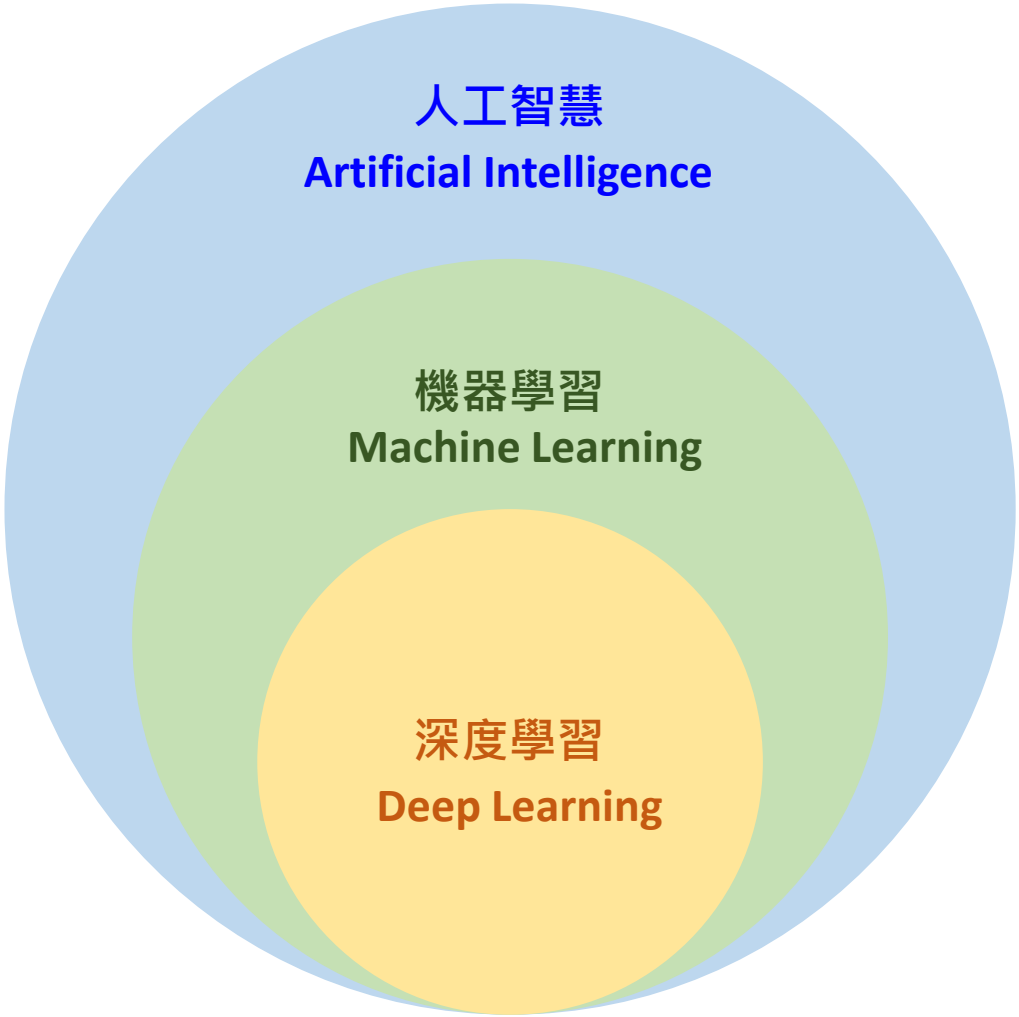
從基因型態預測藥效

Machine learning

Machine Learning:
"The field of study that gives computers the ability to learn without being explicitly programmed."

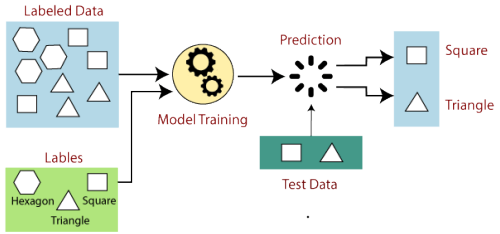
Arthur Samuel popularized the term "machine learning" in 1959

人工智慧，機器學習與深度學習

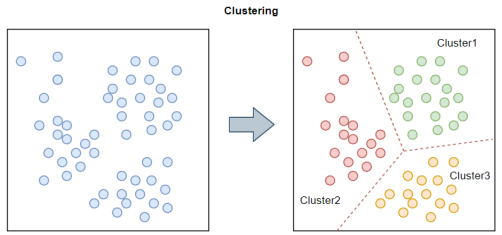


Machine learning

Supervised Learning 監督學習

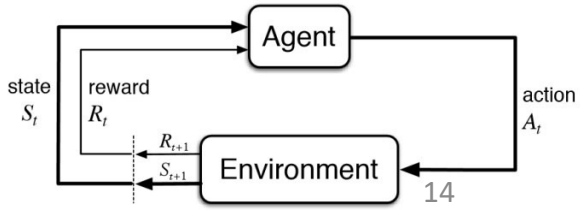


Unsupervised Learning 無監督學習

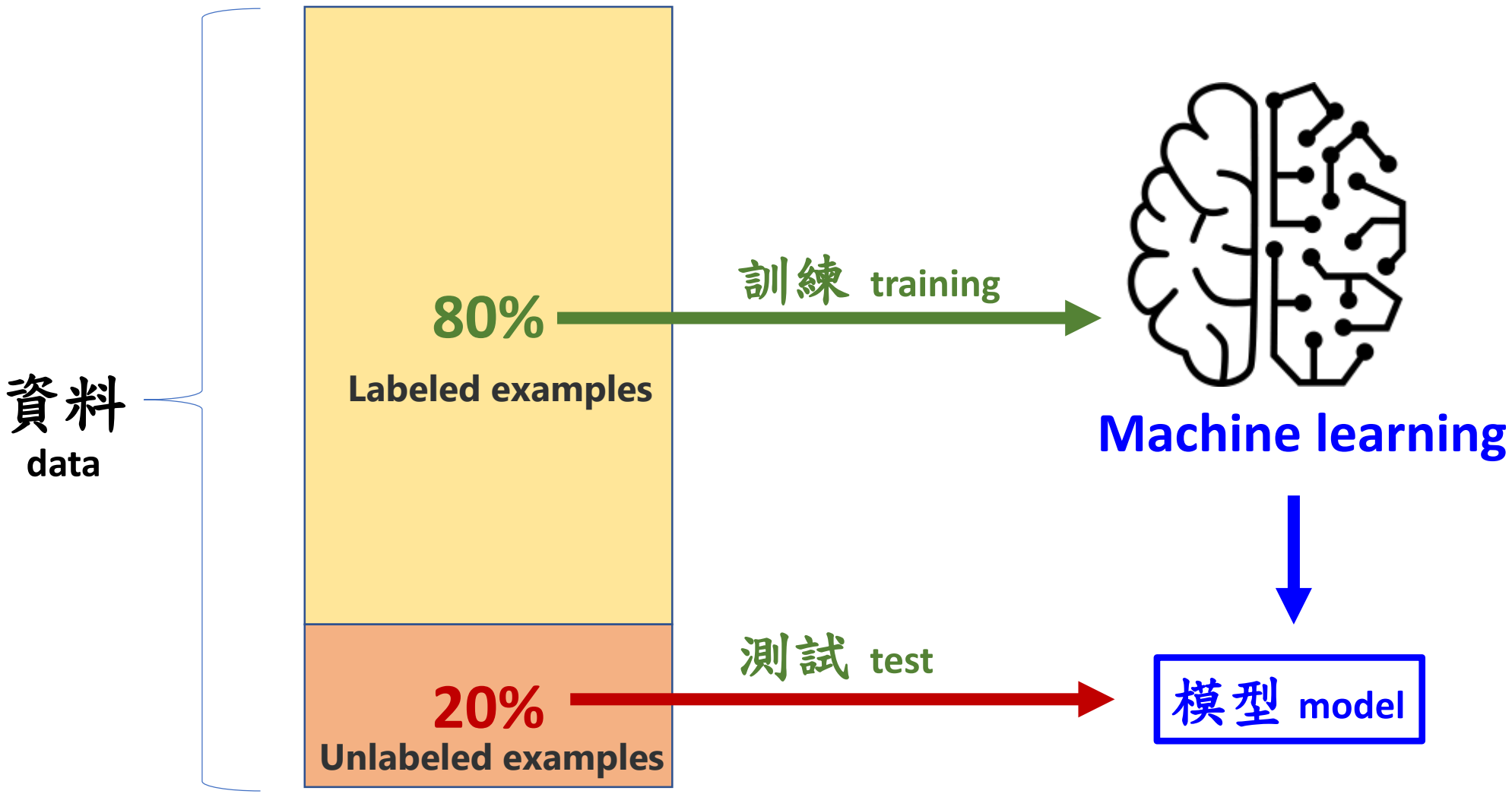


Semi-supervised learning 半監督學習

Reinforcement Learning 強化學習



Supervised Learning 監督學習



Supervised Learning 監督學習

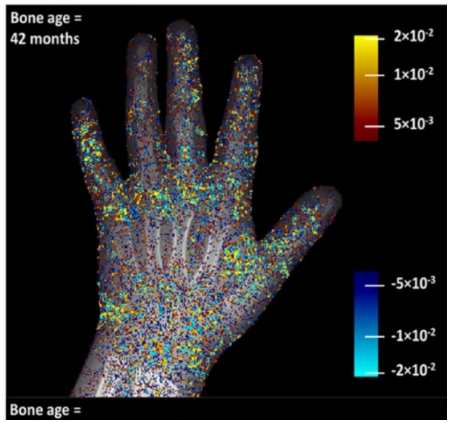
Regression(回歸)

骨齡預測

房價評估

降雨機率

predicting a quantity



Classification(分類)

良性腫瘤/惡性腫瘤

圖片： 食物/狗

垃圾郵件/非垃圾郵件

predicting a label

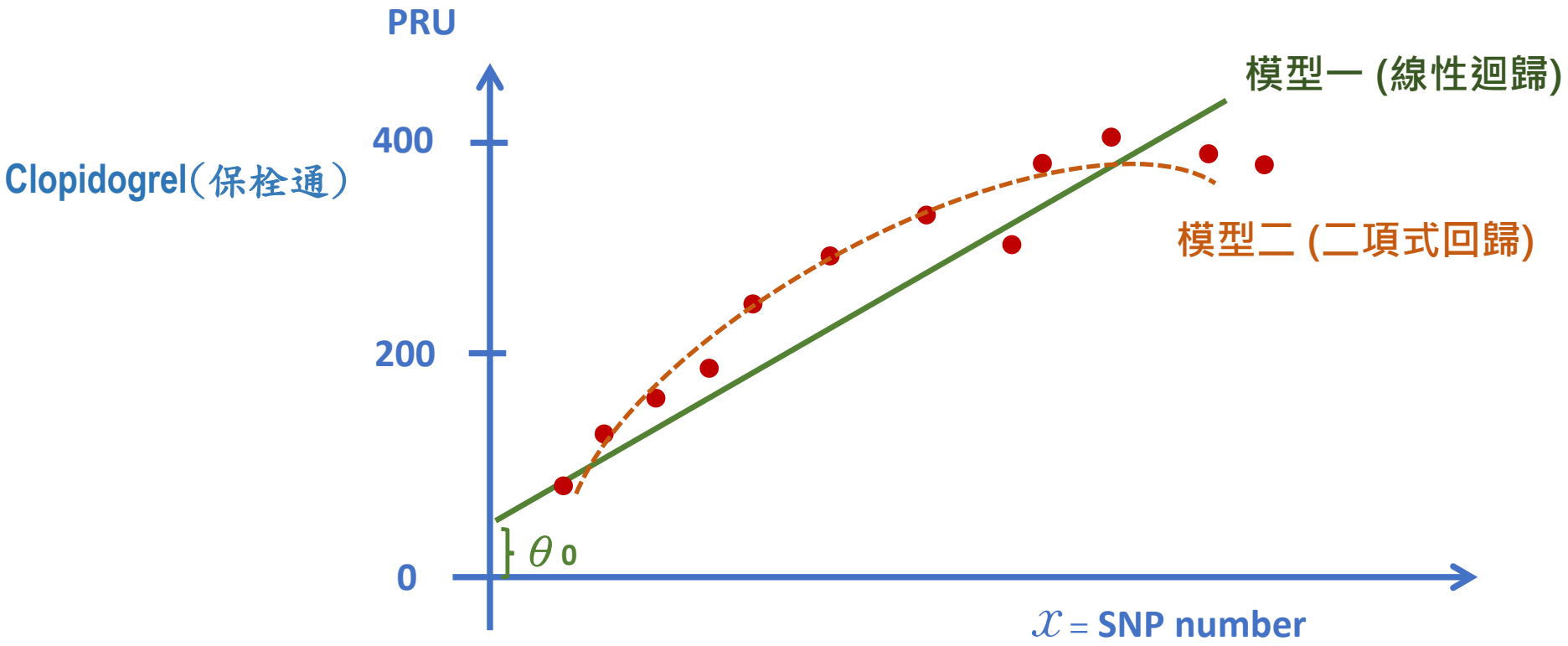
(a)

(b)

(by Karen Zack)

從受試者的特徵值去預測PRU值

Regression(回歸)



X 特徵值(features)

x_0 = 藥效0/1

x_1 = SNP number

x_2 = age of subject

x_3 = sex of subject

x_4 = is CYP2C19*17 (0/1)

⋮

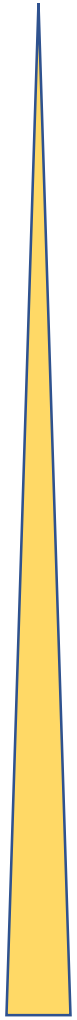
x_{100}

$$h_\theta(x) = \theta_0 + \theta_1 x$$ 模型一 (線性迴歸)

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$ 模型二 (二項式回歸)

Classification(分類)

PRU



高
（抗凝血效果差）

sample_id	PRU	CYP2C19	CYP2B6	CYP2C9	CYP3A4	CYP3A5
1	1	*1B/*1B	*1/*4	*1/*1	*1/*1	*3/*3
2	6	*1B/*1B	*1/*1	*1/*1	*1/*1	*1/*1
3	6	*2/*2	*1/*6	*1/*1	*1/*1	*1/*3
4	6	*1B/*2	*1/*6	*1/*1	*1/*1	*3/*3G
5	102	*1B/*1B	*1/*6	*1/*1	*1/*1	*3/*3
6	106	*1B/*1B	*1/*2	*1/*1	*1/*5	*1/*3
7	119	*1B/*1B	*1/*1	*1/*1	*1/*1	*3/*3
8	121	*1B/*2	*1/*1	*1/*1	*1/*18	*1/*3
9	122	*1B/*1B	*1/*1	*1/*1	*1/*1	*3/*3
10	131	*1B/*2	*1/*6	*1/*1	*1/*1	*1/*3
11	137	*1B/*3	*6/*6	*1/*1	*1/*1	*3/*3
12	142	*1B/*3	*1/*1	*1/*1	*1/*1	*3/*3
13	143	*1B/*2	*1/*1	*1/*1	*1/*1	*1/*3
14	147	*1B/*1B	*1/*1	*1/*1	*1/*1	*3/*3
15	154	*1B/*2	*1/*6	*1/*1	*1/*1	*3/*3
16	158	*1B/*2	*1/*4	*1/*1	*1/*1	*1/*1
17	160	*1B/*1B	*1/*4	*1/*1	*1/*1	*1/*1
18	171	*1B/*2	*1/*4	*1/*1	*1/*1	*3/*3
19	175	*1B/*2	*2/*26	*1/*1	*1/*1	*3/*3
20	179	*1B/*2	*1/*6	*1/*1	*1/*1	*3/*3
21	188	*1B/*1B	*1/*1	*1/*1	*1/*1	*3/*3
22	197	*2/*2	*1/*1	*1/*1	*1/*1	*1/*3
23	197	*2/*2	*2/*6	*1/*1	*1/*1	*3/*3
24	198	*1B/*1B	*1/*4	*1/*1	*1/*1	*3/*3
25	202	*1/*2	*1/*1	*3/*13	*1/*1	*1/*3
26	204	*1/*1B	*1/*6	*1/*3	*1/*1	*1/*3
27	208	*1B/*1B	*1/*4	*1/*1	*1/*1	*1/*1
28	211	*1/*1+rs4	*1/*4	*1/*3	*1/*1	*1/*3
29	213	*1B/*2	*1/*1	*1/*1	*1/*1	*1/*3
30	215	*1/*1B	*1/*1	*1/*1	*1/*1	*3/*3
31	218	*1B/*2	*1/*6	*1/*27	*1/*1	*1/*3
32	225	*1B/*3	*1/*4	*1/*1	*1/*1	*1/*3
35	240	*1B/*2	*1/*1	*1/*1	*1/*1	*1/*3
36	242	*1B/*2	*1/*1	*1/*1	*1/*1	*1/*3
37	249	*1B/*3	*4/*6	*1/*1	*1/*1	*3/*3
38	257	*3/*3	*1/*6	*1/*1	*1/*1	*3/*3
39	310	*1B/*1B	*4/*6	*1/*1	*1/*18	*1/*1
40	325	*1B/*3	*1/*4	*1/*1	*1/*1	*3/*3
41	356	*1B/*2	*1/*6	*1/*1	*1/*1	*1/*3
42	467	*2/*2	*1/*6	*1/*1	*1/*1	*1/*3

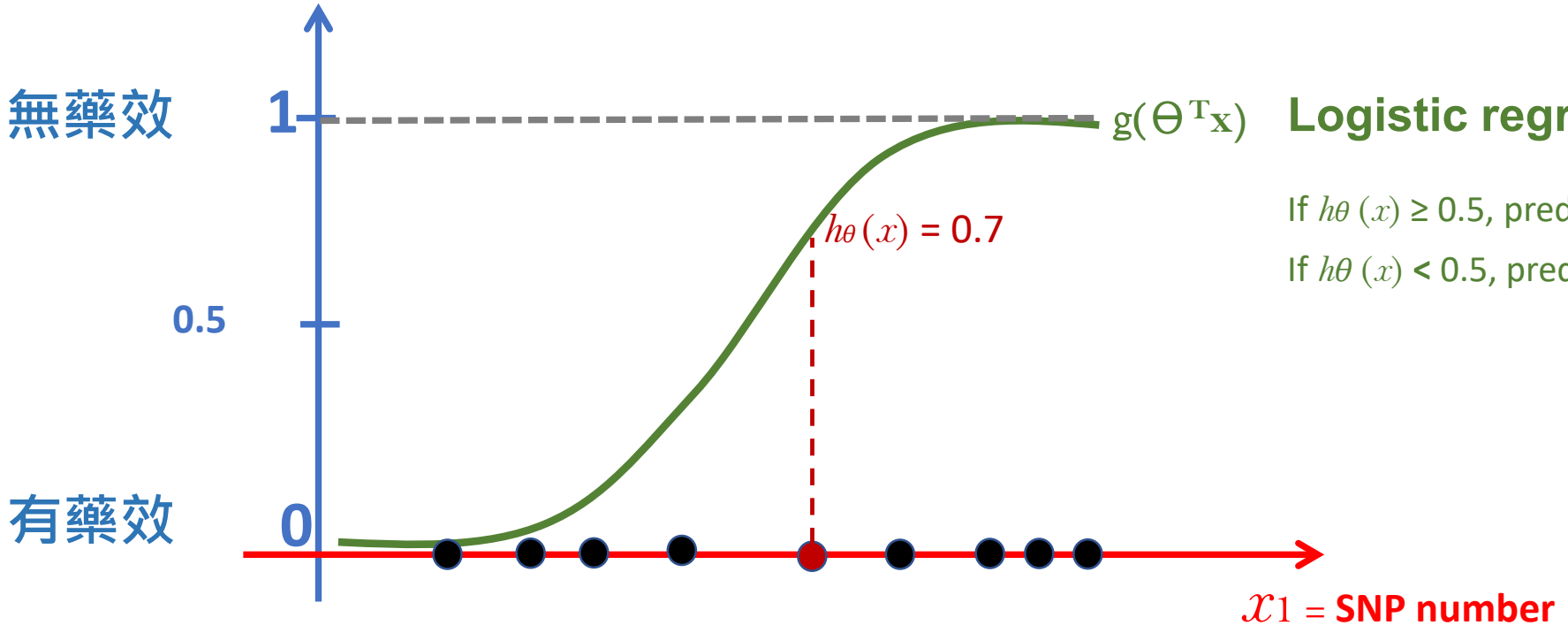
Pharmacogenes
(特徵值)

PRU=200

○ 有藥效
(PRU低)

× 無藥效
(PRU高)

從受試者的特徵值去預測Clopidogrel(保栓通)的藥效



Logistic regression

If $h\theta(x) \geq 0.5$, predict "$y = 1$" 無藥效

If $h\theta(x) < 0.5$, predict "$y = 0$" 有藥效

$h\theta(x)$ = estimated probability that y = 1 on input x

$$\text{If } x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{SNP number} \end{bmatrix}$$

$h\theta(x) = 0.7$

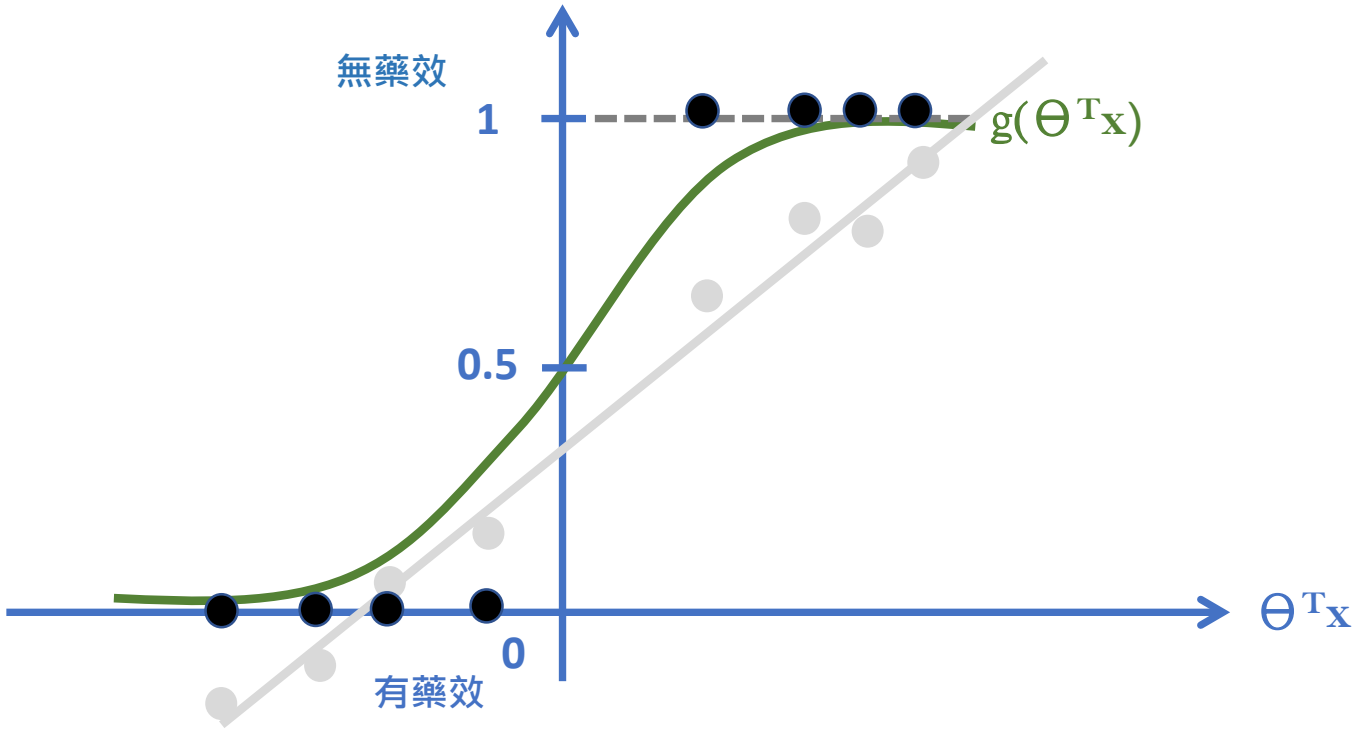
該名患者服藥後70%無效

X 特徵值(features)

x_0 = 藥效0/1

x_1 = SNP number

從受試者的特徵值去預測Clopidogrel(保栓通)的藥效



X 特徵值(features)

x_0 = 藥效0/1

x_1 = SNP number

x_2 = age of subject

x_3 = sex of subject

x_4 = is CYP2C19*17 (0/1)

⋮

x_{100}

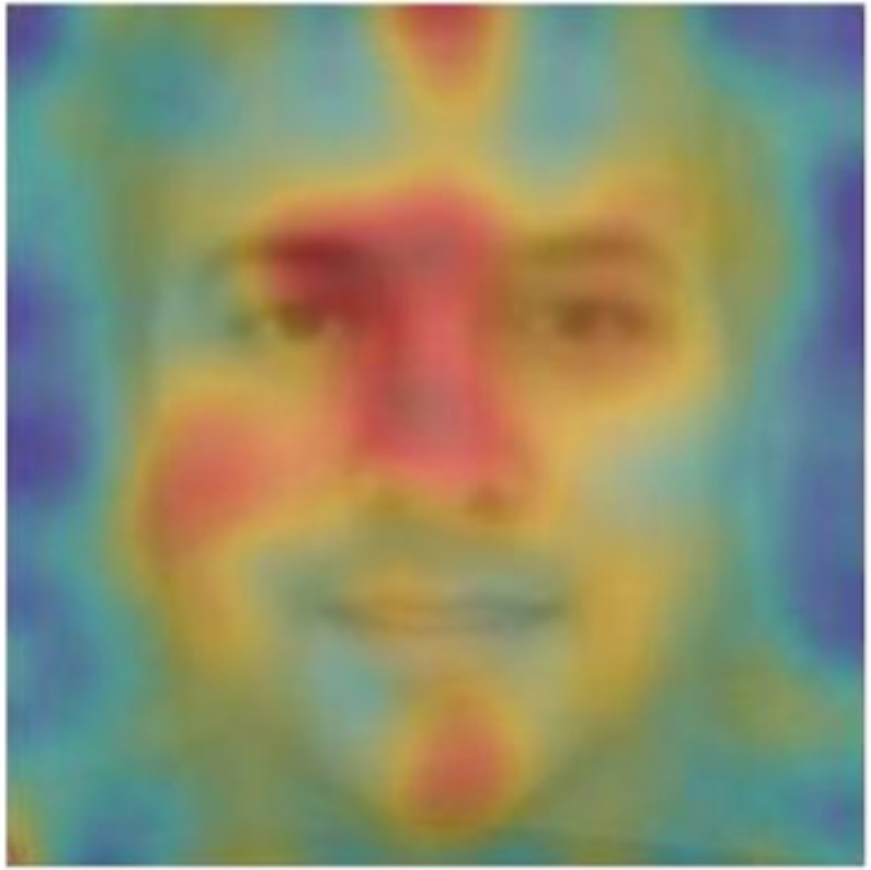
Logistic regression

$$h_\theta(x) = g(\Theta^T x)$$

$$= \frac{1}{1+e^{-\theta^T x}}$$

Sigmoid function

找出重要的特徵值(features) $\boldsymbol{x}$



Male

Female

(Michal Kosinski and Yilun Wang, 2020)

如何診斷代謝症候群?

台灣
目前的評估標準

衛生福利部
國民健康署

2007.1.18公告實施

ATP-III

代謝症候群的診斷標準

符合五項指標任三項者

如何判斷指數的嚴重性?

Taiwan Biobank

Taiwan Biobank

n=217,196

1 Scores Based

建立各項指標風險程度的評分

根據不同性別年齡/性別，標準化各項指標 (Z-score)

2 Weighting

分析各項指標的權重

統計模型

Pearson Correlation

Regression

LASSO

Confirmatory factor analysis (CFA)

從台灣人體資料庫來看不同性別與年齡，各項指數的重要性也不同

男性

年齡	20-39	40-54	55-64	≥65
TG	0.593812	0.564152	0.54498	0.529211
HDL	0.32498	0.323892	0.286285	0.265515
腰圍	0.894987	0.891831	0.902409	0.938052
血糖	0.778805	0.683606	0.641712	0.63396
血壓	0.423434	0.335106	0.261107	0.215299

重要特徵排名
feature scores analysis

20-39	40-54	55-64	≥65
1. 腰圍	1. 腰圍	1. 腰圍	1. 腰圍
2. 血糖	2. 血糖	2. 血糖	2. 血糖
3. TG	3. TG	3. TG	3. TG
4. 血壓	4. 血壓	4. HDL	4. HDL
5. HDL	5. HDL	5. 血壓	5. 血壓

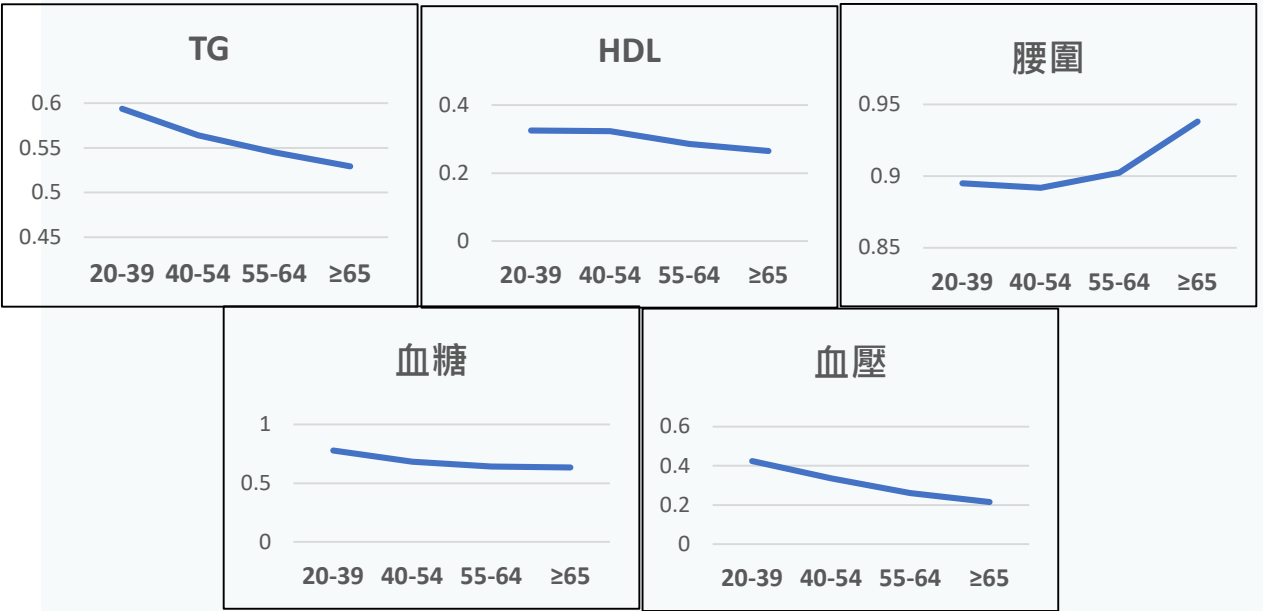
女性

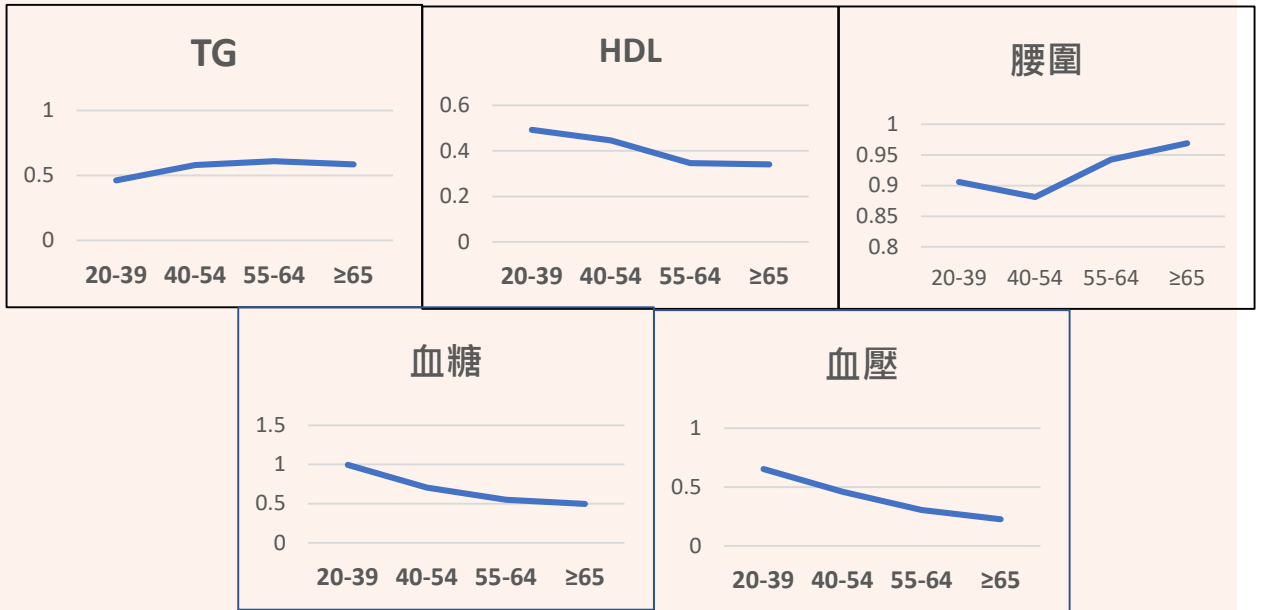
年齡	20-39	40-54	55-64	≥65
TG	0.460873	0.580948	0.608649	0.584282
HDL	0.493218	0.446388	0.346302	0.341003
腰圍	0.905682	0.881332	0.942321	0.969141
血糖	0.995502	0.706625	0.54936	0.497753
血壓	0.651754	0.458852	0.306021	0.224916

重要特徵排名
feature scores analysis

20-39	40-54	55-64	≥65
1.血糖	1. 腰圍	1. 腰圍	1. 腰圍
2.腰圍	2. 血糖	2. TG	2. TG
3.血壓	3. TG	3. 血糖	3. 血糖
4. HDL	4. 血壓	4. HDL	4. HDL
5. TG	5. HDL	5. 血壓	5. 血壓

隨著年齡的增加，腰圍的重要性逐漸上升。



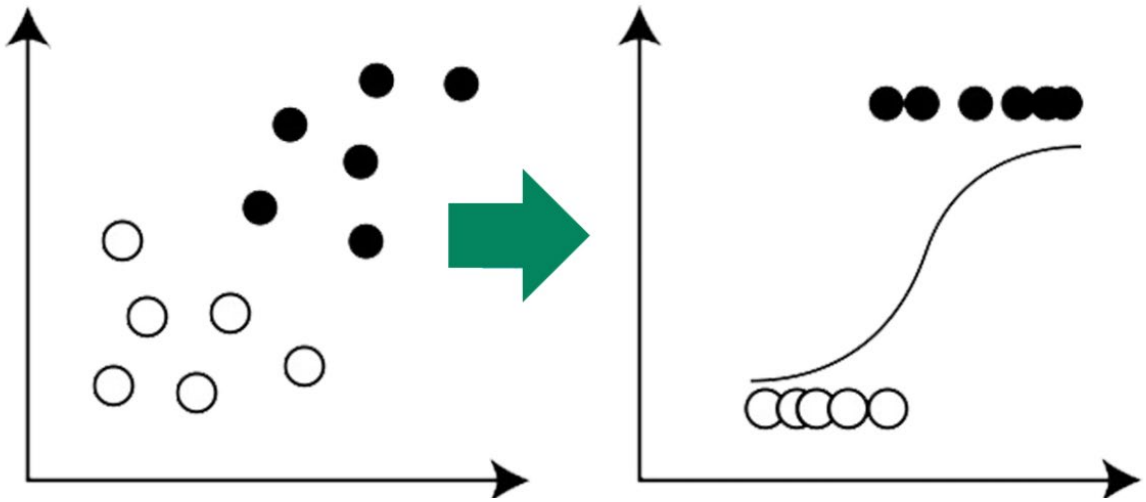


監督學習常用的演算法

如何從細胞外觀判定惡性與良性腫瘤

羅吉斯迴歸 (Logistic Regression)

分類



將細胞特徵綜合評分，畫出惡性與良性腫瘤的分界

單純貝氏(Naïve Bayes)　　分類

Likelihood　　Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

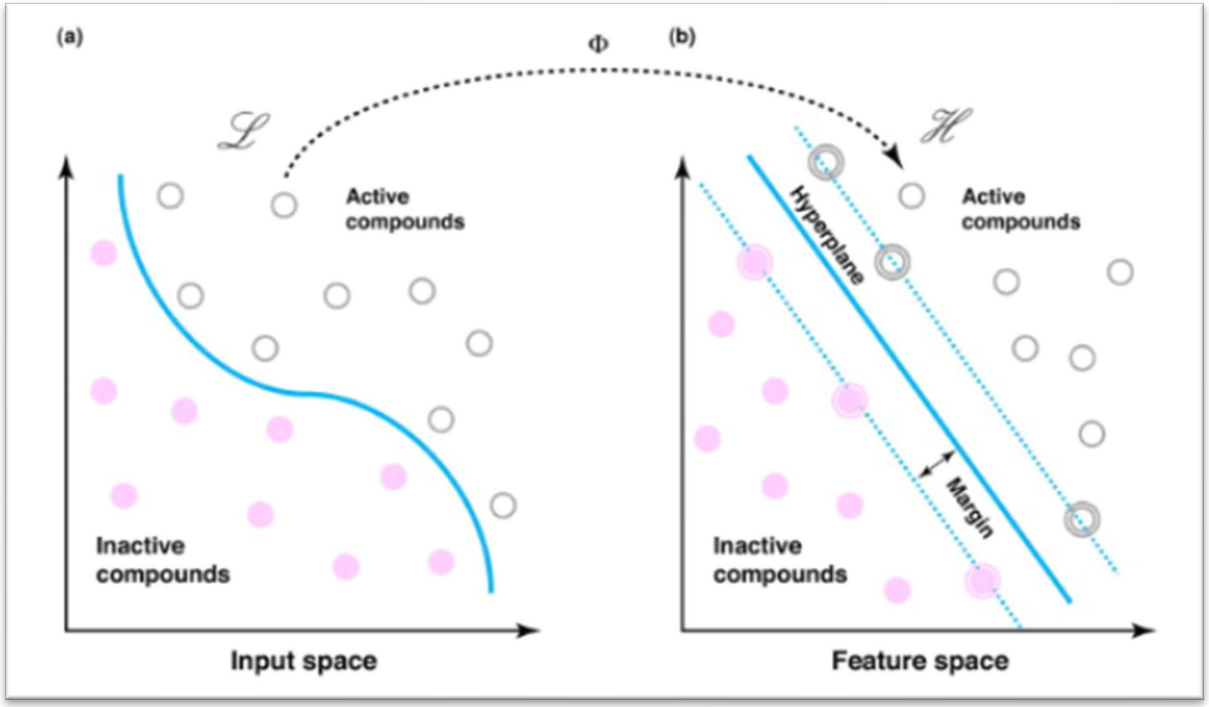
Posterior Probability　　Predictor Prior Probability

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

根據經驗，腫瘤越大，惡性的機率越大

支援向量機

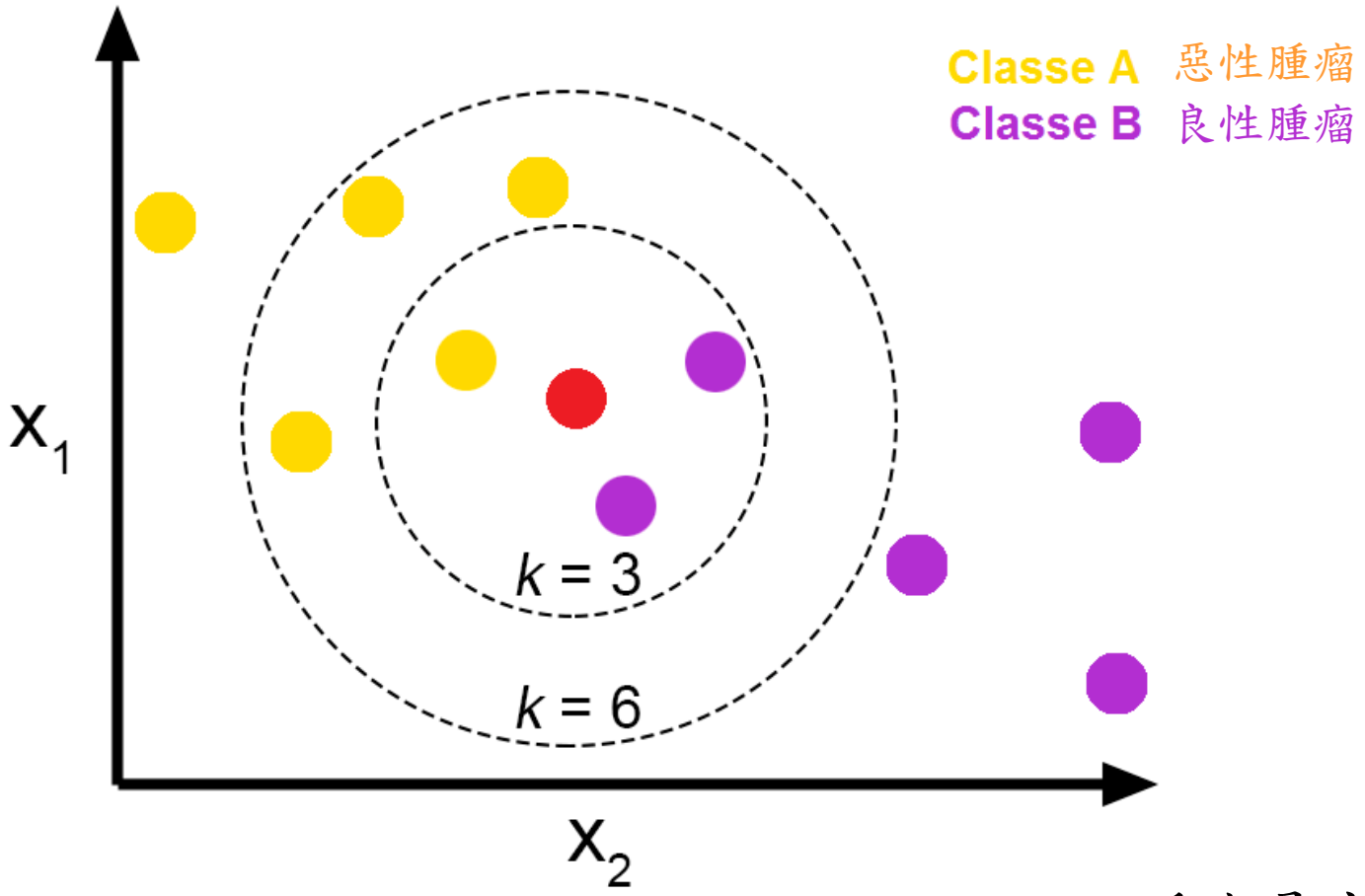
分類(SVM)和迴歸(SVR)



把二維變成三維，找出超平面分開惡性與良性腫瘤

K最近鄰居 (KNN, K-Nearest Neighbors)

分類和迴歸



腫瘤是良性或惡性，由周圍的細胞表決

決策樹 (Decision Tree)

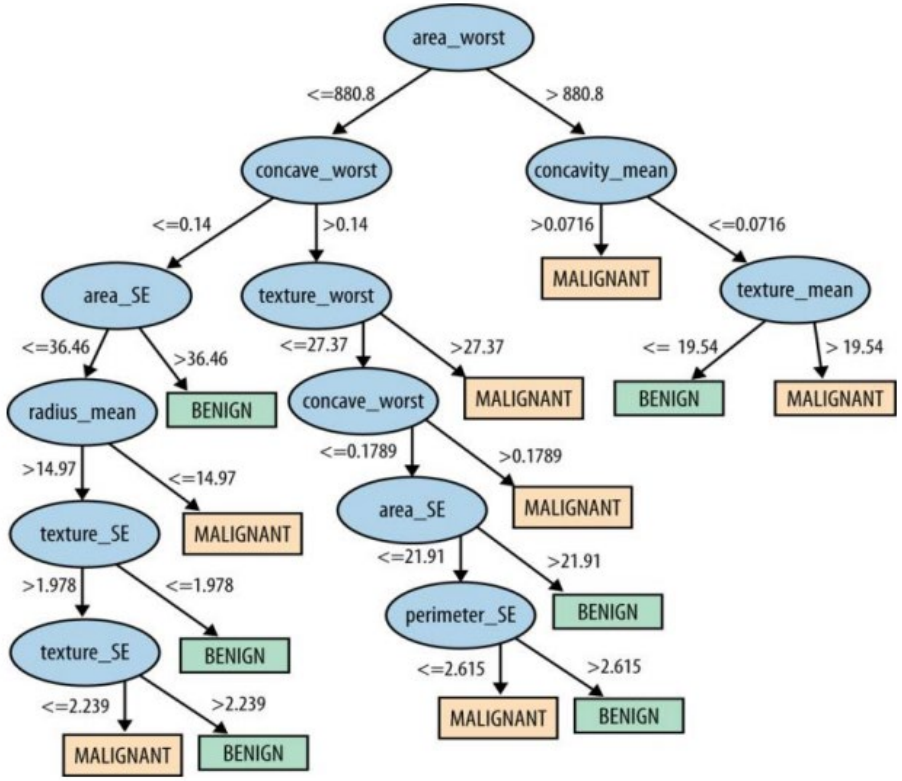
分類和迴歸

隨機森林 (Random Forest)

輕量級的高效梯度提升樹(lightGBM)

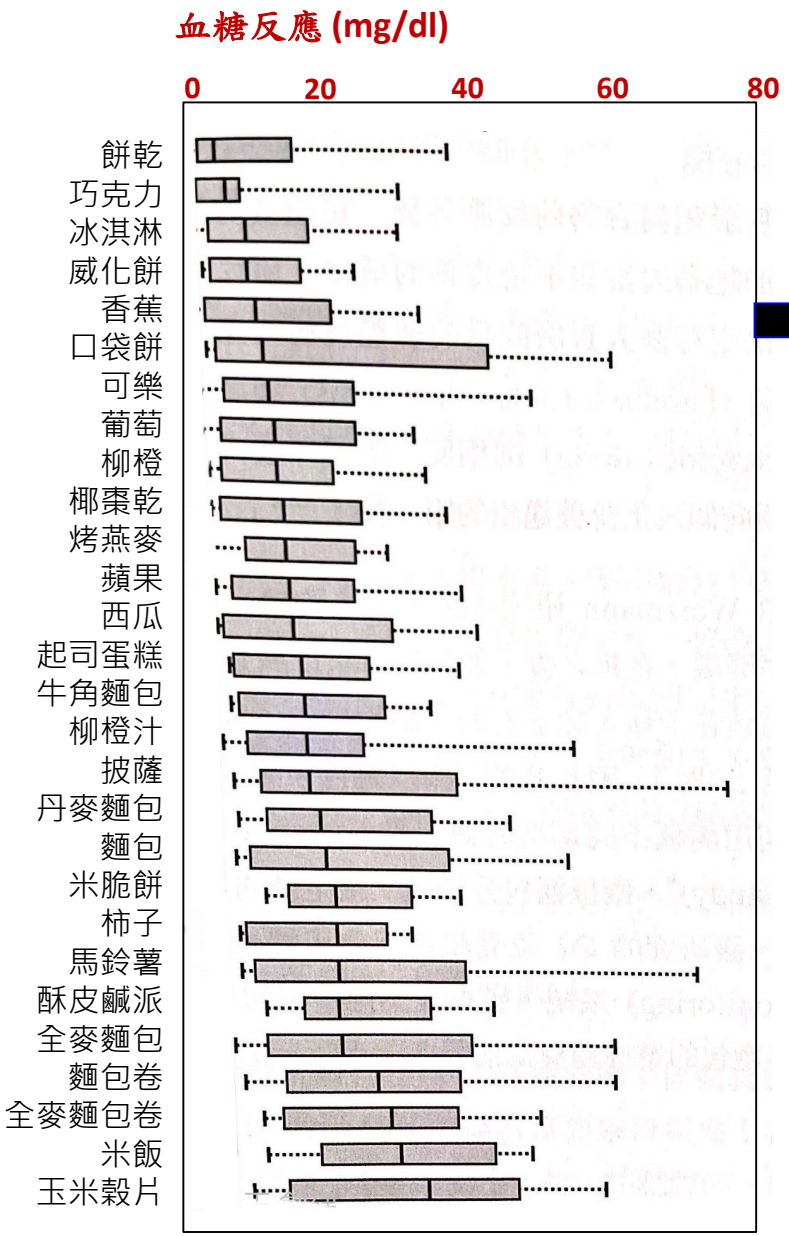
極端梯度提升樹 (XGBoost)

梯度提升樹 (GBDT)

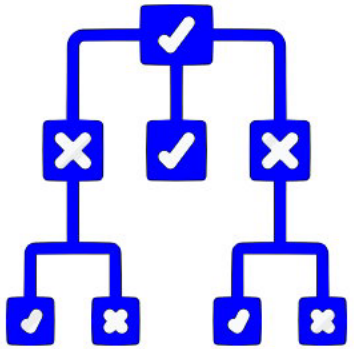


先看腫瘤大小，再看形狀，然後看硬度

由決策樹 (Decision Tree)來預測個人對特定食物的血糖反應



決策樹 Decision Tree



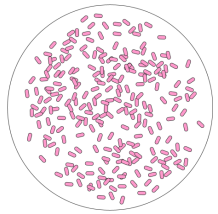
137特徵值

血糖預測模型

腸道菌才是關鍵

狄氏副擬桿菌

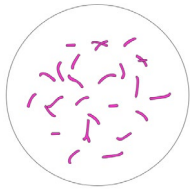
Parabacteroides distasonis



血糖反應

多氏擬桿菌

Bacteroides dorei



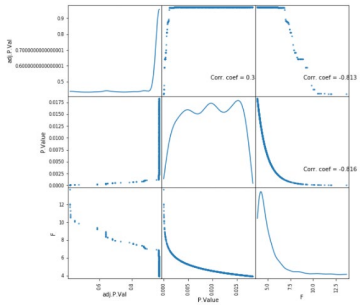
血糖反應

(E. Segal & E. Elinav, 2017)

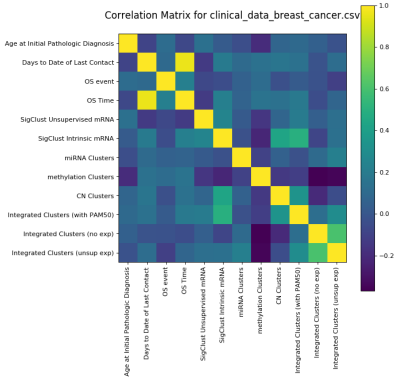
如何入門?

kaggle 全球最大的資料科學社群

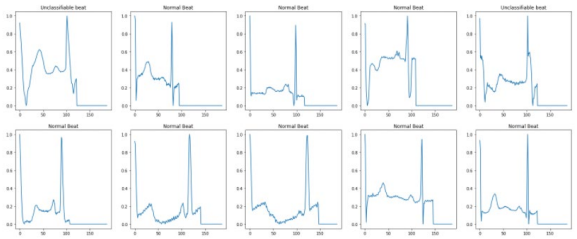
Inside Kaggle you'll find all the code & data you need to do your data science work. Use over 50,000 public datasets and 400,000 public notebooks to conquer any analysis in no time.



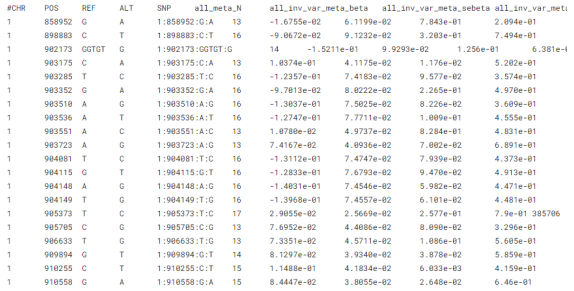
Alzheimer Microarray Analysis



Breast Cancer Analysis



ECG Heartbeat Categorization

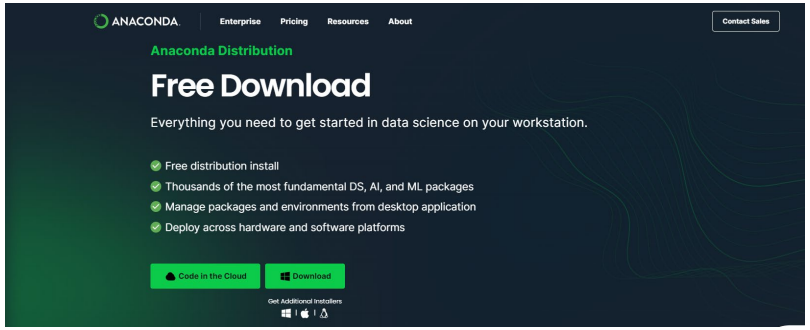


#CHR	POS	REF	ALT	SNP	all_meta_N	all_inv_var_meta_beta	all_inv_var_meta_sebeta	all_inv_var_meta
1	858952	G	A	1:858952:G:A	13	-1.6755e-02	6.1199e-02	7.843e-01	2.094e-01
1	898883	C	T	1:898883:C:T	16	-9.0672e-02	9.1232e-02	3.203e-01	7.494e-01
1	902173	GGTGT	G	1:902173:GGTGT:G	14	-1.5211e-01	9.9293e-02	1.256e-01	6.381e-0
1	903175	C	A	1:903175:C:A	13	1.0374e-01	4.1175e-02	1.176e-02	5.202e-01
1	903285	T	C	1:903285:T:C	16	-1.2357e-01	7.4183e-02	9.577e-02	3.574e-01
1	903352	G	A	1:903352:G:A	16	-9.7013e-02	8.0222e-02	2.265e-01	4.970e-01
1	903510	A	G	1:903510:A:G	16	-1.3037e-01	7.5025e-02	8.226e-02	3.609e-01
1	903536	A	T	1:903536:A:T	16	-1.2747e-01	7.7711e-02	1.009e-01	4.555e-01
1	903551	A	C	1:903551:A:C	13	1.0780e-02	4.9737e-02	8.284e-01	4.831e-01
1	903723	A	G	1:903723:A:G	13	7.4167e-02	4.0936e-02	7.002e-02	6.891e-01
1	904081	T	C	1:904081:T:C	16	-1.3112e-01	7.4747e-02	7.939e-02	4.373e-01
1	904115	G	T	1:904115:G:T	16	-1.2833e-01	7.6793e-02	9.470e-02	4.913e-01
1	904148	A	G	1:904148:A:G	16	-1.4031e-01	7.4546e-02	5.982e-02	4.471e-01
1	904149	T	G	1:904149:T:G	16	-1.3968e-01	7.4557e-02	6.101e-02	4.481e-01
1	905373	T	C	1:905373:T:C	17	2.9055e-02	2.5669e-02	2.577e-01	7.9e-01 385706
1	905705	C	G	1:905705:C:G	13	7.6952e-02	4.4086e-02	8.090e-02	3.296e-01
1	906633	T	G	1:906633:T:G	13	7.3351e-02	4.5711e-02	1.086e-01	5.605e-01
1	909894	G	T	1:909894:G:T	14	8.1297e-02	3.9340e-02	3.878e-02	5.859e-01
1	910255	C	T	1:910255:C:T	15	1.1488e-01	4.1834e-02	6.033e-03	4.159e-01
1	910558	G	A	1:910558:G:A	15	8.4447e-02	3.8055e-02	2.648e-02	6.46e-01

COVID19-hg GWAS round 5 release (GRCh38)

快速入門

上課資料



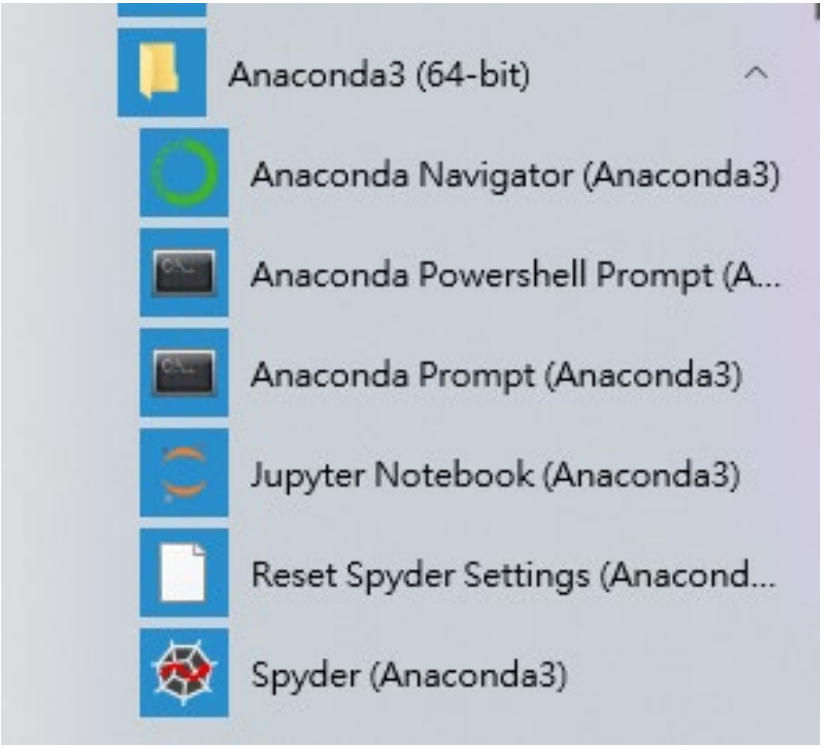
1. 下載 **ANACONDA**: https://www.anaconda.com

2. 安裝

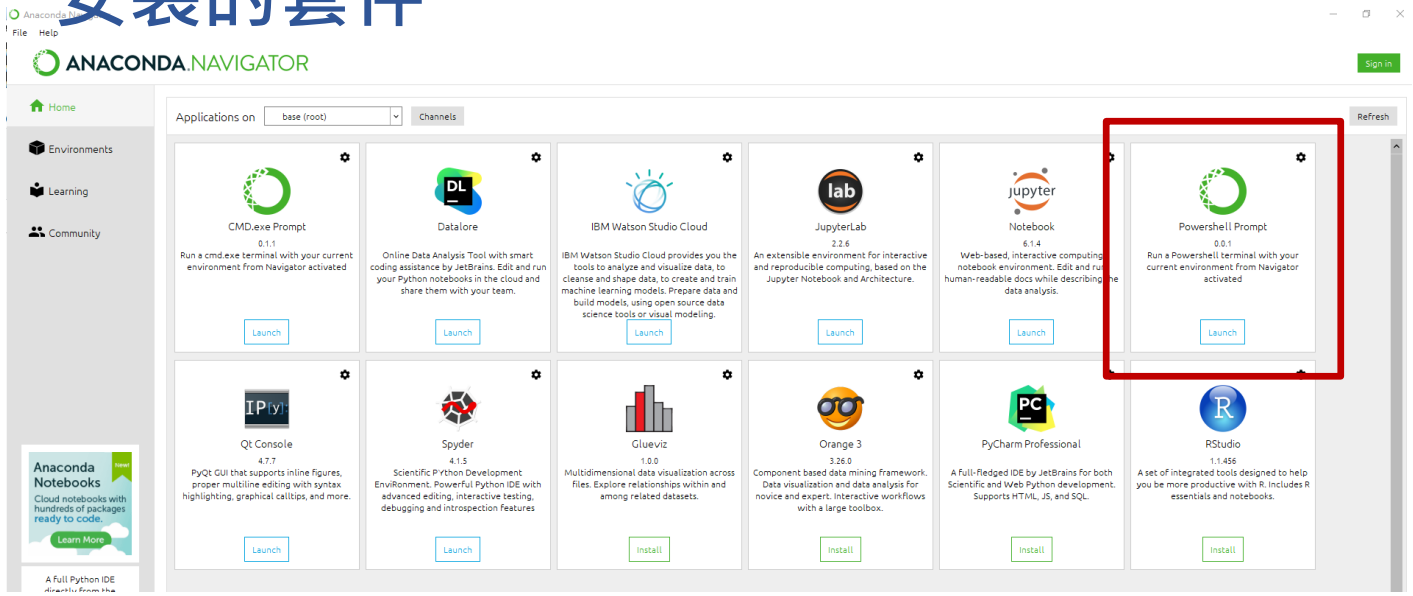
3. 打開**ANACONDA**

4. 準備工作: 安裝套件 ← 只需要做一次

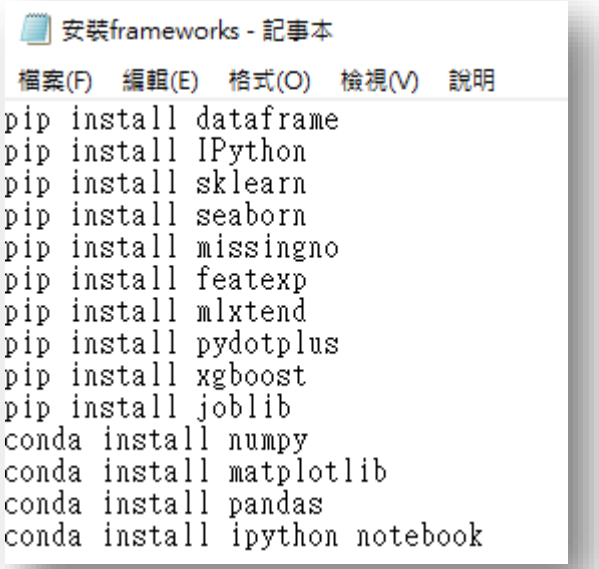
5. 跑程式 (Jupyter)

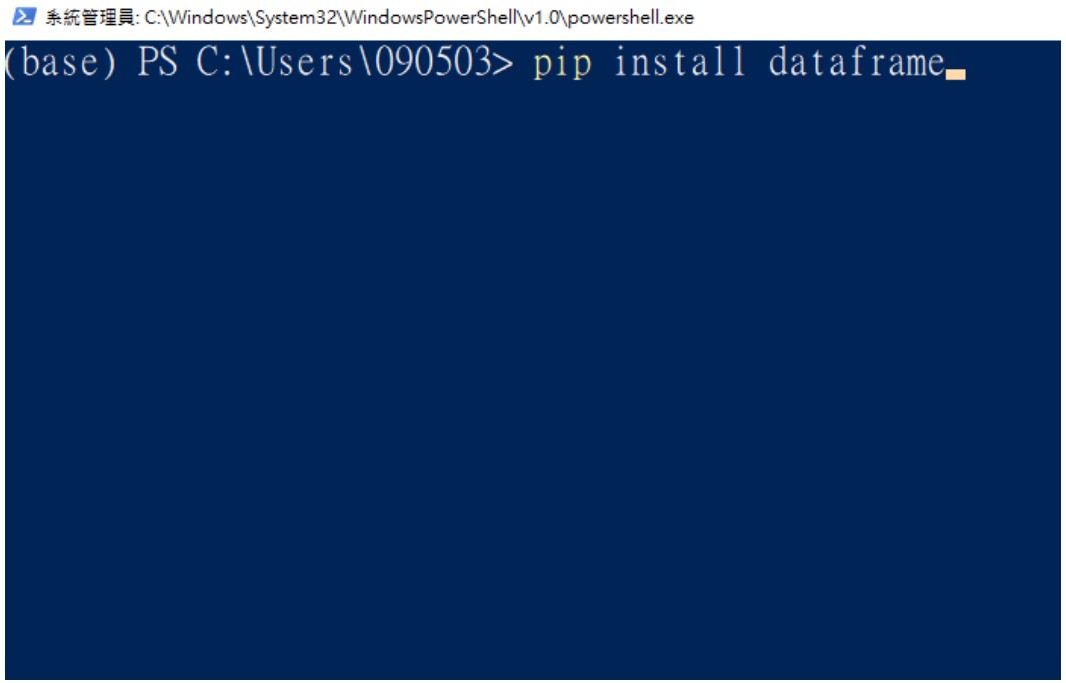


安裝的套件



打開”安裝frameworks.txt”，一行一行輸入需要安裝的套件

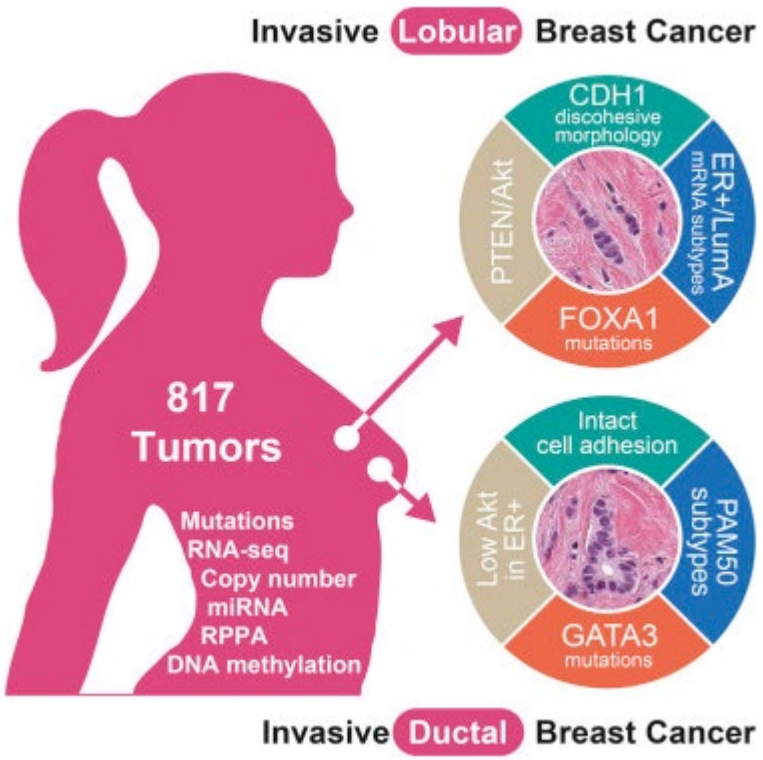




實戰演練

檔案: https://www.kaggle.com/datasets/samdemharter/brca-multiomics-tcga

paper: https://www.cell.com/cell/fulltext/S0092-8674(15)01195-2



Multi-Omics
- Copy Number Variations (860)
- Somatic Mutations (249)
- Gene Expression (604)
- Protein Expression (223)
- Total: 1936 features

Total: 1936 features

n=705 (611名存活，94名死亡)

data.csv

1936 features

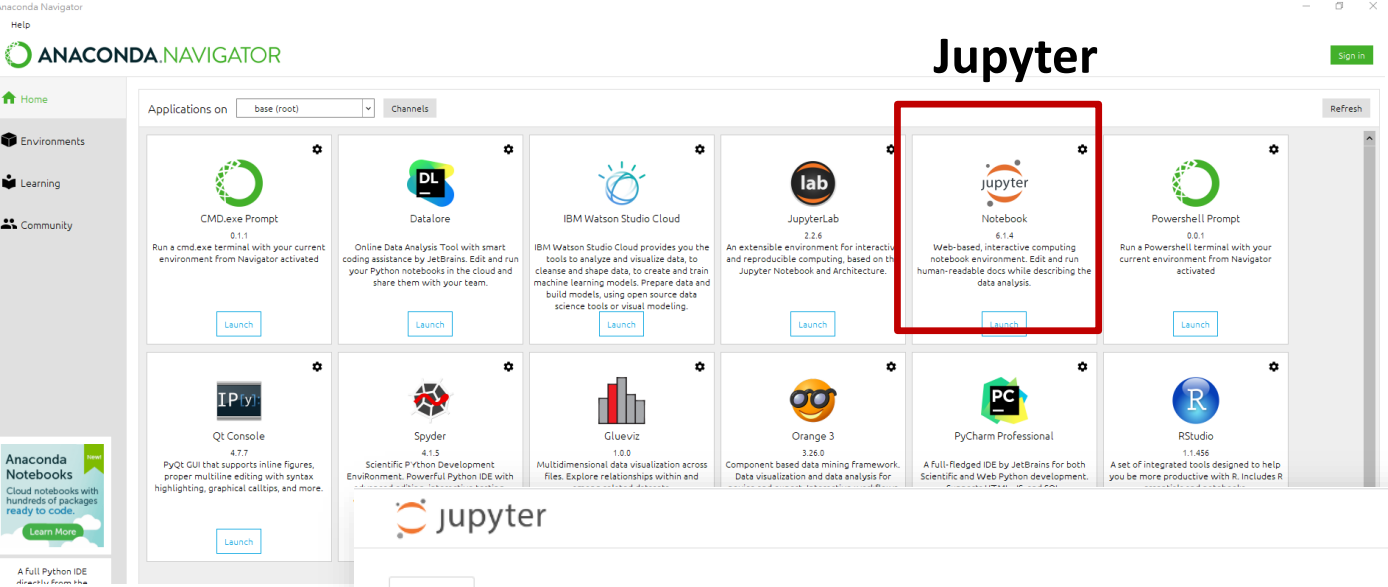
survival: 0
death: 1

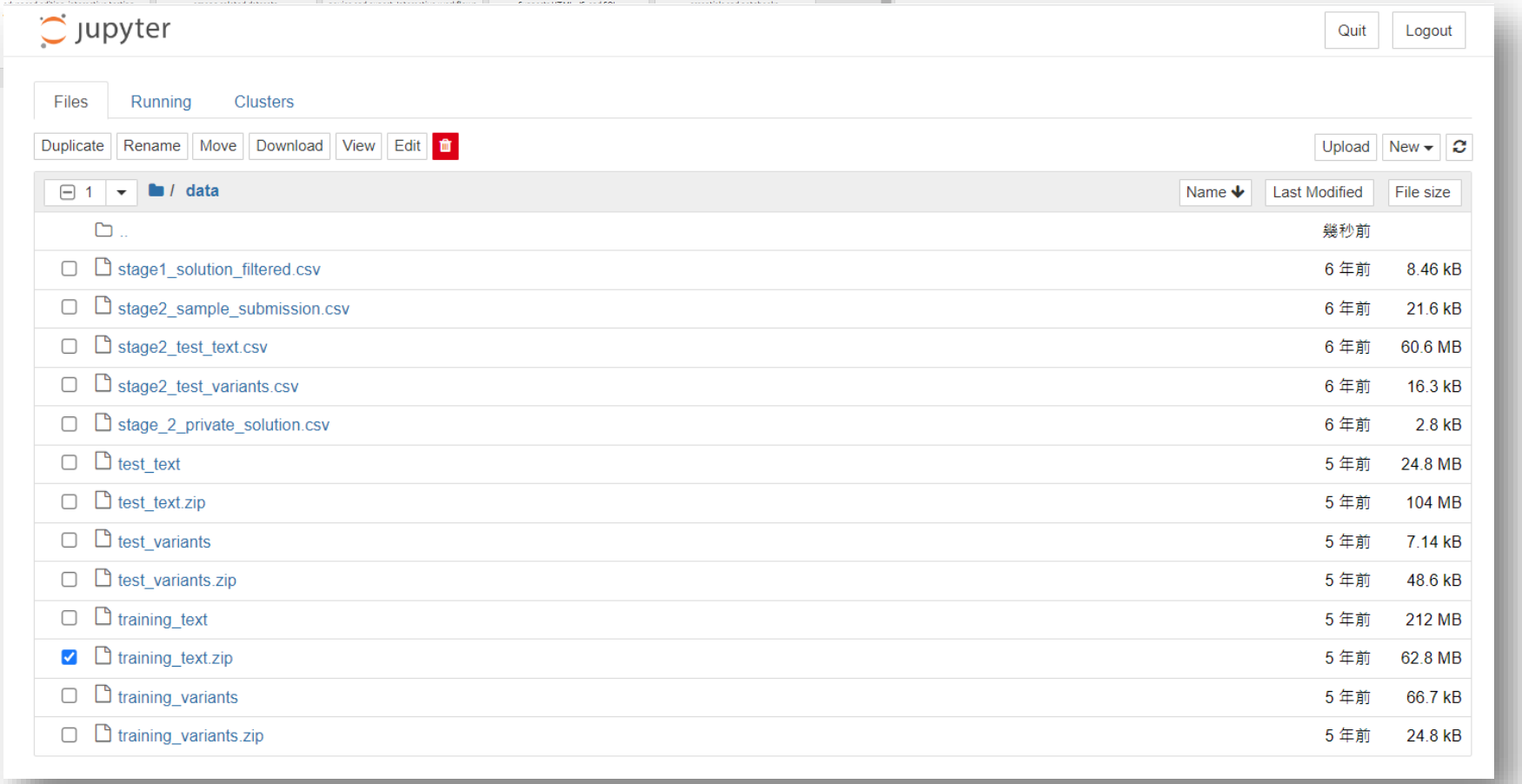
705 patients

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	BVL	BVM
1	rs_CLEC3	rs_CPB1	rs_SCGB2	rs_SCGB1	rs_TFF1	rs_MUCL1	rs_GSTM1	rs_PIP	rs_ADIPO	rs_ADH1B	rs_S100A7	rs_HMGCS	rs_CYP2B	rs_ANKRD	rs_PRAME	lpp_p90RS	vital.status
2	0.892818	6.580103	14.12367	10.6065	13.18924	6.649466	10.52033	10.33849	10.24838	10.22997	0	7.904609	8.754801	9.681176	9.363258	-0.20726	0
3	0	3.691311	17.11609	15.51723	9.867616	9.691667	8.179522	7.911723	1.289598	1.818891	0.444879	1.289598	5.698288	7.885523	1.818891	0.26753	0
4	3.74815	4.375255	9.658123	5.326983	12.10954	11.64431	10.51733	5.114925	11.97535	11.91144	4.126056	9.383379	8.630913	9.61031	4.733674	-0.19852	0
5	0	18.23552	18.53548	14.53358	14.07899	8.91376	10.55746	13.30443	8.205059	9.211476	0	4.347992	7.836265	9.732374	9.069201	-0.0469	0
6	0	4.583724	15.71186	12.80452	8.881669	8.430028	12.96461	6.806517	4.294341	5.385714	0.841893	6.11157	7.701901	10.7592	1.370053	0.037473	0
7	0.602837	1.026658	9.248794	5.095713	0	6.085525	7.773748	11.08078	6.916635	7.935318	8.671148	5.846375	4.665484	0	11.48724	-0.04847	1
8	5.654312	2.842074	8.231955	7.062477	7.285706	8.490481	0	6.823051	5.907011	6.793263	6.287818	2.762306	2.390172	4.593306	8.209576	0.67704	0
9	2.370945	0.430499	0	2.635987	0.761711	2.788999	5.45192	2.552943	7.388579	6.464041	0.430499	1.625925	1.453491	0	11.82103	-0.10164	0
10	0	0.603597	2.21313	1.027933	1.847075	0	5.949234	1.622087	1.027933	0	0	1.847075	1.847075	0	8.118521	-0.10292	0
11	0	8.187287	5.250765	3.058923	10.29397	0	0.493032	0	9.300252	9.766249	0	0.859811	10.32087	10.69306	1.94557	-0.40892	1
12	0	1.878333	5.078392	2.741467	14.30011	14.61199	2.58938	10.27948	10.80134	10.55123	1.878333	8.667065	14.1371	10.77706	1.418352	-0.34292	0
13	0	8.152976	11.38638	8.111426	9.813765	6.457277	1.594644	7.26423	11.55646	11.6801	0	8.013315	11.8963	9.243775	1.594644	0.090501	0
14	12.08081	1.078542	4.287783	1.311852	9.451286	4.367665	9.611614	8.252329	11.80453	11.73297	0.454808	0	10.24658	2.344005	3.725883	-0.09927	0
15	0.54745	3.648235	0.54745	0.54745	0	7.824536	6.152814	0.54745	5.614657	4.474436	6.440336	0	0.54745	0	1.253747	-0.09538	1
16	10.22	10.13828	14.25146	10.51583	8.020775	12.02788	0.740021	10.50855	8.096916	9.961419	1.880098	8.743535	8.933501	9.745235	0	-0.419	0
17	0	5.03562	6.944828	6.301805	13.35564	5.467071	12.55429	17.15364	5.082013	5.478677	2.193141	7.069034	16.35991	11.34807	0	-0.17159	0
18	1.225028	7.223793	15.18962	12.26103	11.92687	13.07427	2.942646	12.181	8.253096	8.42015	2.764792	8.853269	5.837464	7.359244	4.269115	0.429118	0
19	0	12.18525	10.23754	8.880304	9.218761	9.478781	9.841524	15.88566	0	0	5.641419	3.537768	10.56868	9.668254	9.091519	0.241038	1
20	5.207627	6.388463	7.419122	4.816287	3.759348	4.83754	11.86855	3.714103	4.367986	4.899475	3.408087	4.338481	8.238421	0	2.11919	-0.38334	0
21	0	0.818605	1.337597	0.818605	8.2703	11.07211	8.985609	1.87688	11.8184	10.92404	0	1.101314	8.133799	8.391719	10.60697	0.024336	1
22	4.376457	4.162299	11.51267	10.10907	7.922094	11.97035	10.18525	12.85716	4.041988	7.48328	1.587701	7.839681	6.939862	9.395491	1.443182	-0.38335	0
23	5.272363	5.58896	5.424925	2.927915	12.06274	9.128513	12.65949	9.305378	2.814612	1.590626	12.65178	3.082754	2.408984	7.026013	9.361294	-0.04657	0
24	12.6757	1.38648	10.08193	4.788534	7.36861	14.90865	1.935497	9.742503	3.181389	0	11.00965	10.61633	1.774671	0	4.458566	0.016026	0
25	0	0	0.497638	0.497638	0	0	11.38833	0	0	0	11.44249	0.497638	0.497638	0	9.782566	-0.50348	0
⋮																	
706	1.687374	0.530171	15.89707	12.01882	13.90623	14.52932	11.95344	8.899572	9.476798	9.451396	3.649374	9.933222	10.60729	9.804425	1.687374	-0.37637	0

…

Jupyter





開啟brca.ipynb

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

Prediction of Breast Cancer Severity with Multi-Omics Data

導入必要的套件

In [1]:
```
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split, StratifiedKFold, cross_val_score
from sklearn.metrics import roc_curve, roc_auc_score
import matplotlib.pyplot as plt
from joblib import dump
```

步驟1: 載入數據

In [2]:
```
data_path = 'data.csv'
df = pd.read_csv(data_path)
```

步驟2: 將資料分為特徵和目標變量

In [3]:
```
X = df.drop("vital.status", axis=1)
y = df["vital.status"]
```

步驟3: 根據目標變量進行層次化分割，將資料分為訓練集和測試集

In [4]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, stratify=y, train_size=0.85, random_state=40)
```

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

步驟4: 訓練隨機森林分類器

In [5]:
```
rf = RandomForestClassifier(random_state=30)
rf.fit(X_train, y_train)
```

Out[5]:
```
RandomForestClassifier
RandomForestClassifier(random_state=30)
```

步驟5: 在測試集上進行預測

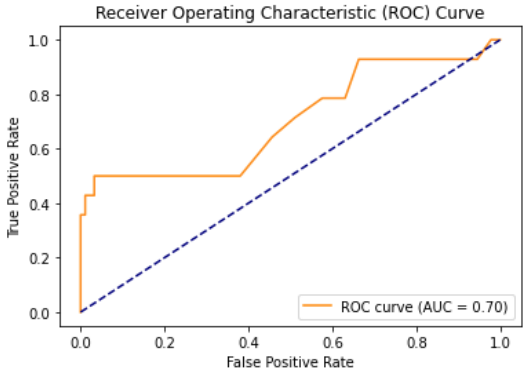
In [6]:
```
y_pred_prob = rf.predict_proba(X_test)[:, 1]
```

步驟6: 計算 ROC 曲線和 AUC（曲線下面積）

In [7]:
```
fpr, tpr, _ = roc_curve(y_test, y_pred_prob)
roc_auc = roc_auc_score(y_test, y_pred_prob)
```

步驟7: 繪製 ROC 曲線

In [8]:
```
plt.figure()
plt.plot(fpr, tpr, color='darkorange', label=f'ROC curve (AUC = {roc_auc:.2f})')
plt.plot([0, 1], [0, 1], color='navy', linestyle='--')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.legend(loc="lower right")
plt.show()
```



步驟8: 使用 K-Fold 交叉驗證進行模型驗證

In [9]:
```
cv = StratifiedKFold(n_splits=10)
scores = cross_val_score(RandomForestClassifier(random_state=42), X, y, cv=cv, scoring="accuracy")
print(f"Accuracy: {scores.mean():.4f} (+/- {scores.std() * 2:.4f})")
```

```
Accuracy: 0.8809 (+/- 0.0563)
```

步驟9: 重要特徵排序

In [10]:
```
importance = pd.Series(data=rf.feature_importances_, index=X_train.columns)
importance = importance.sort_values(ascending=False)
print(importance)
```

```
rs_KIAA0408    0.010571
rs_MMRN1       0.010038
rs_ADIPOQ      0.005953
rs_SLC19A3     0.005876
rs_KCNIP2      0.005841
                 ...
cn_AMZ1        0.000000
cn_FRMD1       0.000000
cn_SYT5        0.000000
cn_SIRPG       0.000000
cn_SLC1A2      0.000000
Length: 1936, dtype: float64
```

Step 10: 輸出訓練好的模型

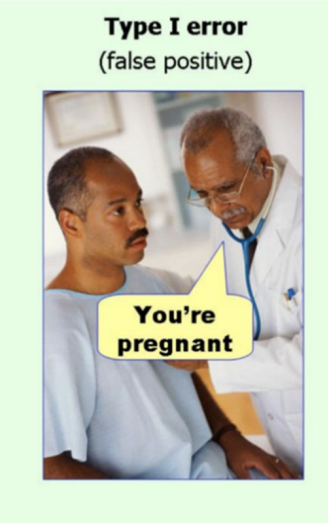
In [11]:
```
dump(rf, 'random_forest_model.joblib')
```

Out[11]:
```
['random_forest_model.joblib']
```

效能評估

confusion matrix 混淆矩陣

實際 \ 預測	Positive	Negative
Positive	True Positive (TP)	False Negative (FN) 偽陰性
Negative	False Positive (FP) 偽陽性	True Negative (TN)



偽陽性率，沒病被診斷有病的比率

false positive rate (FPR) $\frac{FP}{N} = \frac{FP}{FP + TN}$

偽陰性率，有病被診斷沒病的比率

false negative rate (FNR) $\frac{FN}{P} = \frac{FN}{FN + TP}$

Recall/Sensitivity $$\frac{TP}{P} = \frac{TP}{TP + FN}$$ 所有得病者被正確診斷出得病的比率

Specificity $$\frac{TN}{N} = \frac{TN}{TN + FP}$$ 所有非得病者被正確診斷非得病的比率

Precision $$\frac{TP}{TP + FP}$$ 診斷出得病者中，真正得病的比率

False Discovery Rate (FDR) $$\frac{FP}{FP + TP} = 1 - \text{Precision}$$

Accuracy $$\frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN}$$ 所有受試者被正確診斷為有病和沒病的比率

Harmonic Precision-Recall Mean (F1 Score)

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

診斷出得病者中，真正得病的比率

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

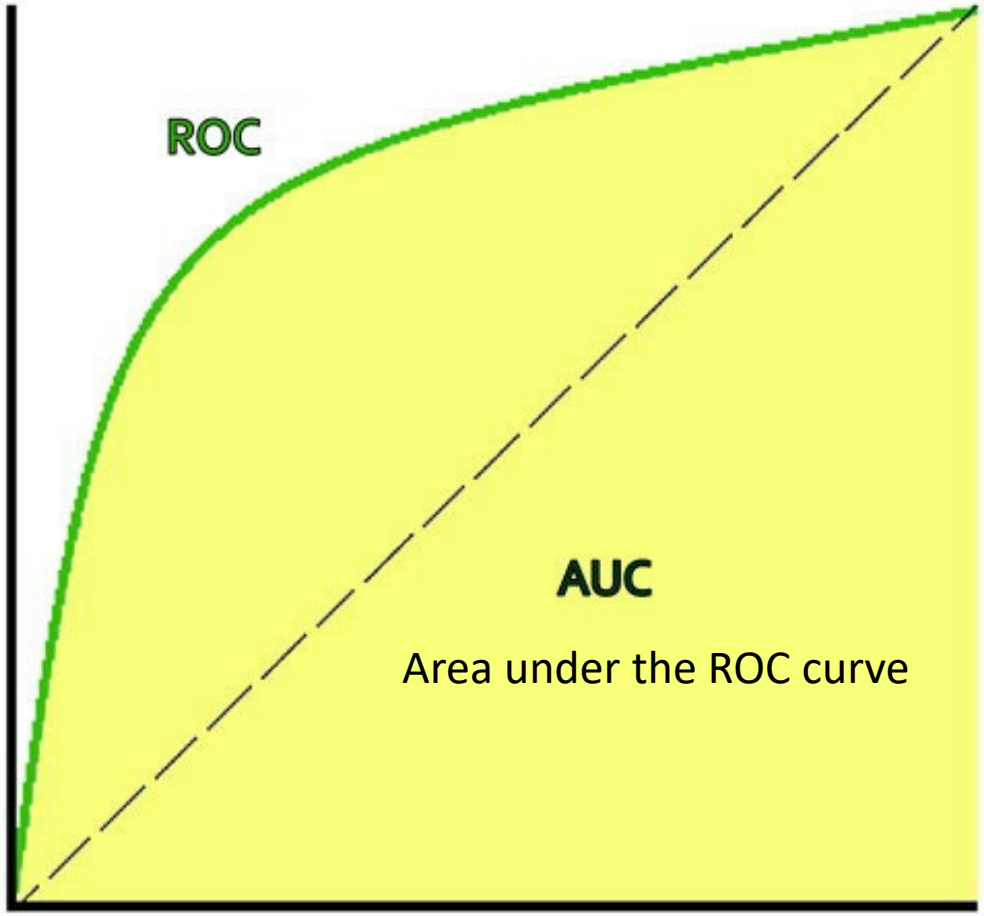
所有得病者被正確診斷出得病的比率

$$F_1 = 2 * \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Receiver operating characteristic curve, ROC

TPR

真陽性率(true positive rate)
=
敏感度(sensitivity)
=
召回率 (recall)



FPR

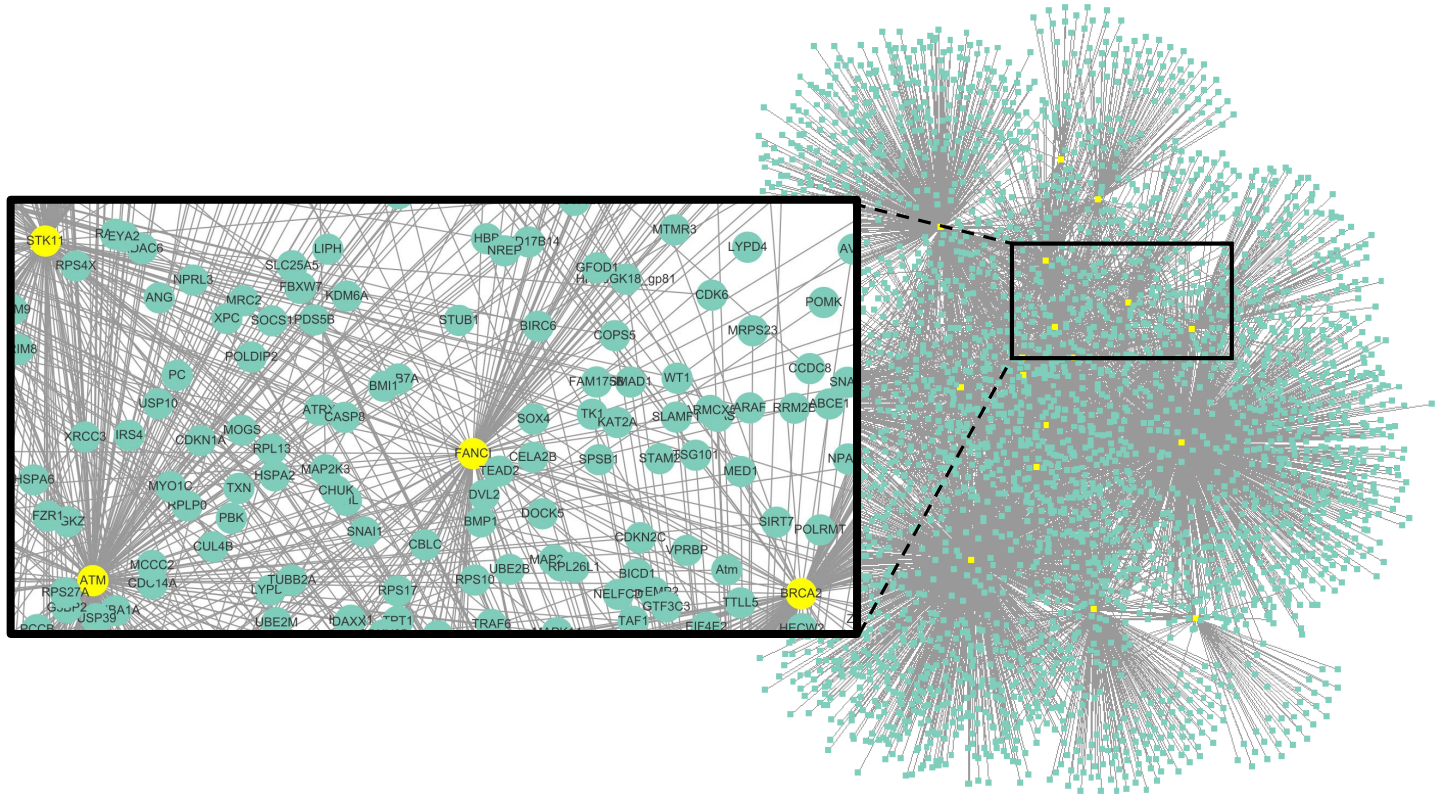
研究罕病用**Precision**

偽陽性率(false-positive rate) = 1 - Specificity

與遺傳性乳癌相關的基因

14%遺傳性乳癌患者找不到相關基因

ATM	ATM serine/threonine kinase
BARD1	BRCA1 associated RING domain 1
BRCA1	BRCA1 DNA repair associated
BRCA2	BRCA2 DNA repair associated
BRIP1	BRCA1 interacting helicase 1
CDH1	cadherin 1
CHEK2	checkpoint kinase 2
FANCA	FA complementation group A
FANCI	FA complementation group I
FANCL	FA complementation group L
NBN	nibrin
NF1	neurofibromin 1
PALB2	partner and localizer of BRCA2
PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
PMS2	PMS1 homolog 2, mismatch repair system component
PPM1D	protein phosphatase, Mg2+/Mn2+ dependent 1D
PTEN	phosphatase and tensin homolog
RAD51C	RAD51 paralog C
STK11	serine/threonine kinase 11
TP53	tumor protein p53

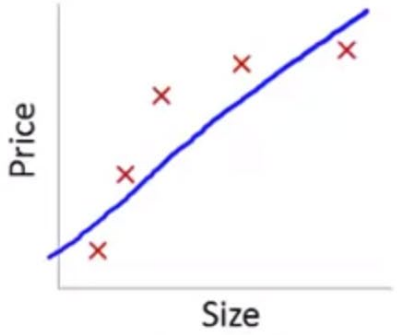


僅有10%乳癌患者有以上基因突變

5%乳癌患者有BRCA1/2突變

欠擬合(Underfitting) &過擬合(overfitting)

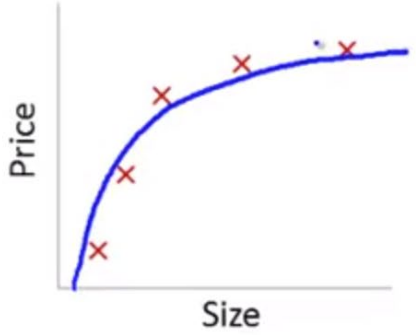
欠擬合



$\theta_0 + \theta_1 x$

High bias
(underfit)

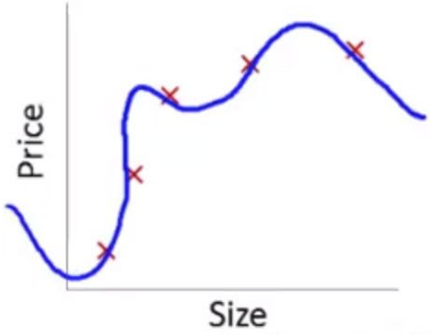
完美



$\theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"

過擬合



$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

High variance
(overfit)

特徵數/參數太少

單純以遺傳性乳癌相關基因
建立乳癌預測模型

樣本數太少

特徵數/參數太多

(Source from Andrew Ng)

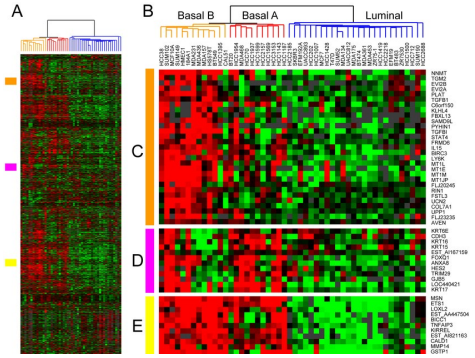
過擬合

從人的睡姿來預測床的形狀

Unsupervised Learning 無監督學習

Clustering (群集)

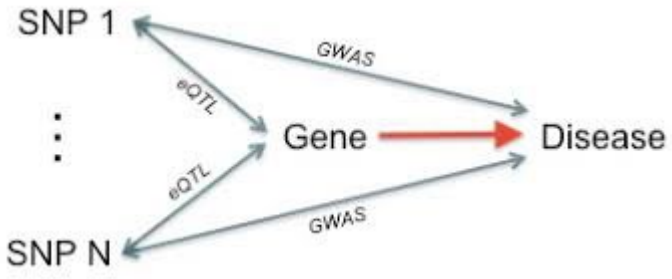
Microarray 基因表現分群



(Kao, et al., 2009)

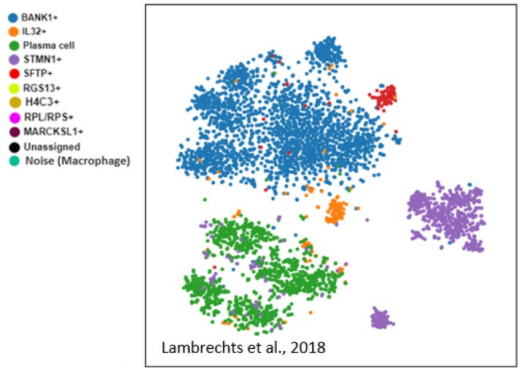
Association (關聯)

GWAS (全基因組關聯分析)



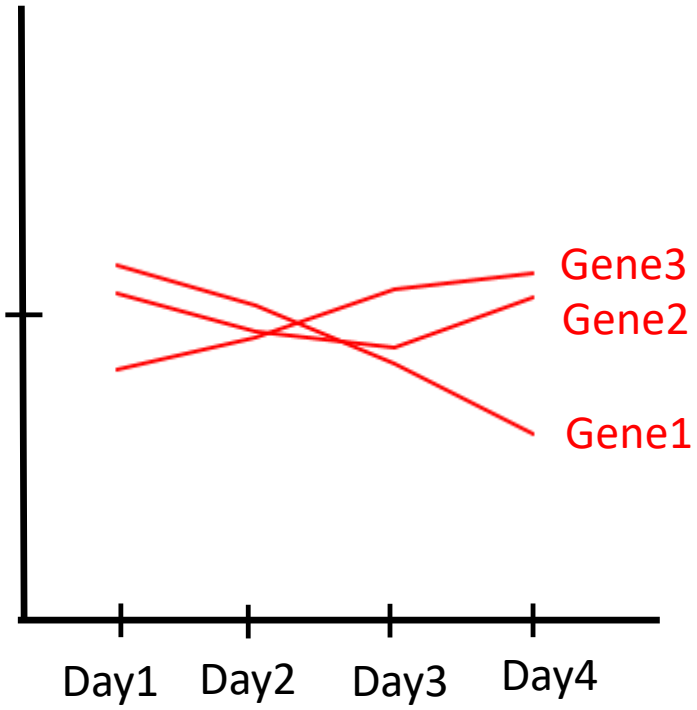
Dimensionality Reduce (降維)

Single-cell analysis



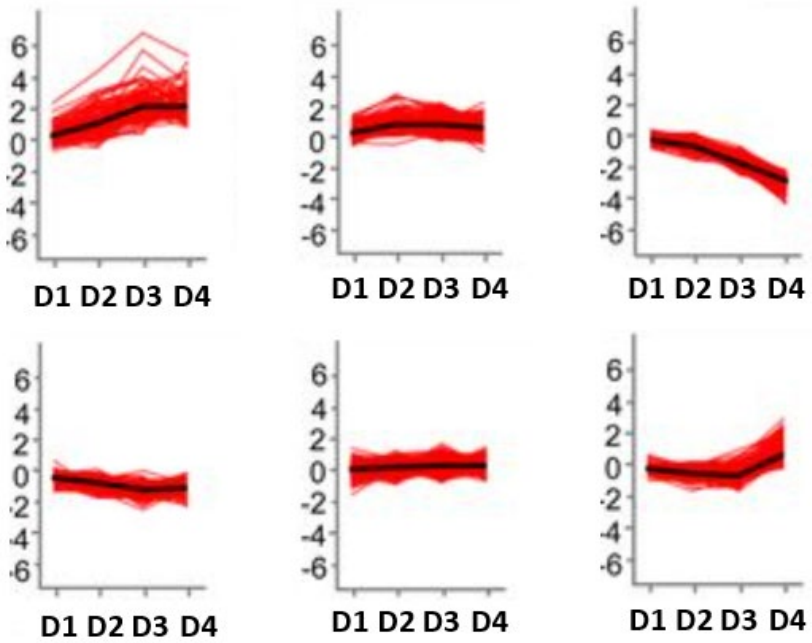
施打疫苗後基因產生的變化　Clustering (群集)

Microarray/RNA-seq 基因表現分群



k-Means Clustering

k=6

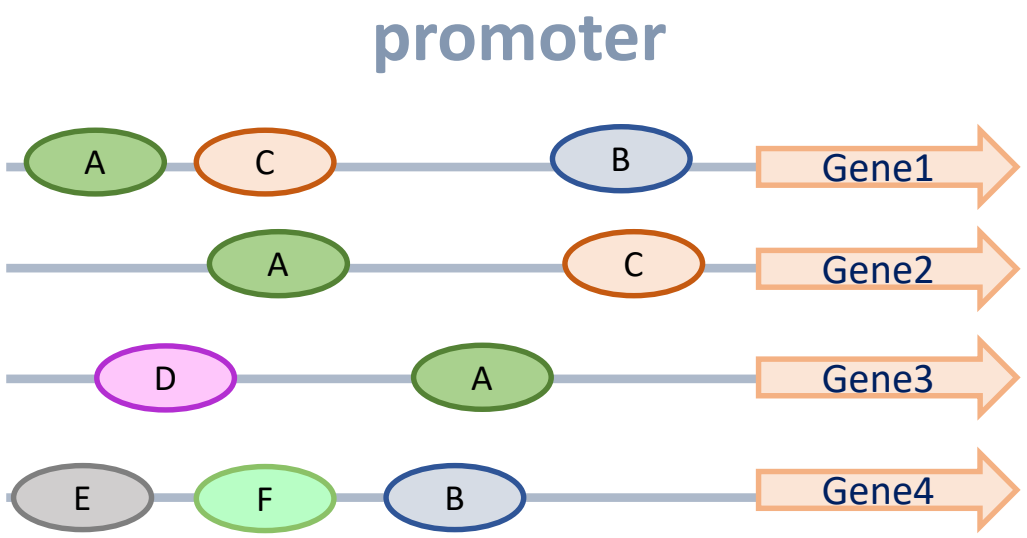


Gene Functional Analysis

Day0: Control

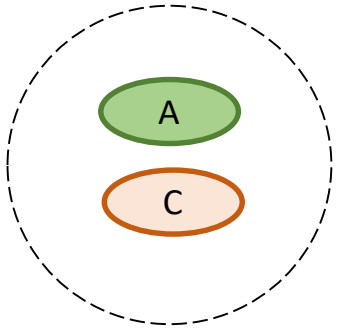
找出轉錄因子(Transcription factors)的組合

Association (關聯)



Apriori algorithm

Transcription factors (TF)



Gene	TF
Gene1	A,B,C
Gene2	A,C
Gene3	A,D
Gene4	B,E,F

計算頻率 →

Frequent Items	Support
{A}	0.75
{B}	0.5
{C}	0.5
{A,C}	0.5
{E}	0.25
{F}	0.25
{A,B,C}	0.25
{A,D}	0.25
{B,E,F}	0.25

Minimum support **0.5**
Minimum confidence **0.5**

For site combination $A \Rightarrow C$:

Support = support({A, C}) = 0.5

Confidence = support({A, C})/support({A}) = 0.67

Confidence = support({A, C})/support({C}) = 1

Whiteld *et al. Genome Biology* 2012, **13**:R50
http://genomebiology.com/2012/13/9/R50

Genome Biology

RESEARCH Open Access

Functional analysis of transcription factor binding sites in human promoters

Troy W Whiteld[1], Jie Wang[1], Patrick J Collins[2], E Christopher Partridge[3], Shelley Force Aldred[2], Nathan D Trinklein[2], Richard M Myers[3] and Zhiping Weng[1*]

Genome Biology 2012

Transcription factor binding resulted in transcriptional repression in more than a third of functional sites.

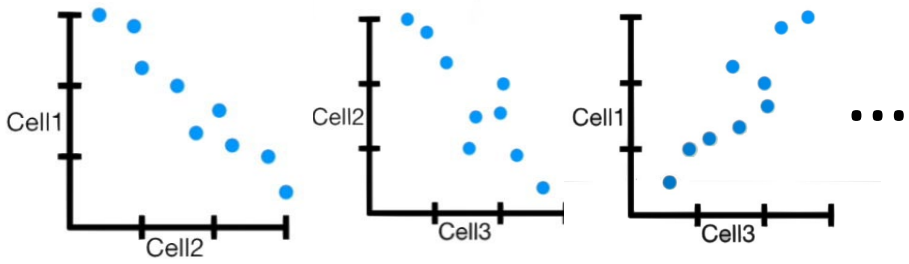
The activating TF binding sites are significantly closer than repressing TF binding sites to the TSS

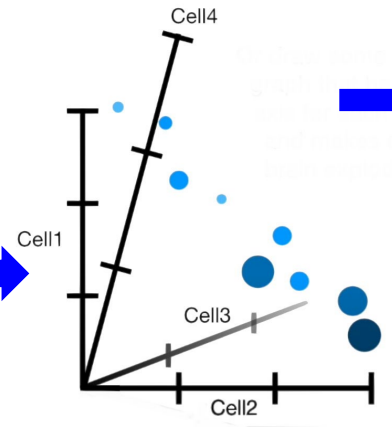
Positive elements on human promoters between 40 and 350 bp away from the TSS, as well as the presence of negative elements from 350 to 1,000 bp upstream of the TSS (Cooper et al., *Genome Res*, 2006).

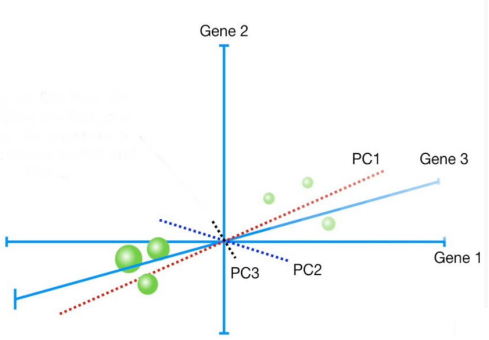
利用基因表現值將細胞分類

Dimensionality Reduce (降維)

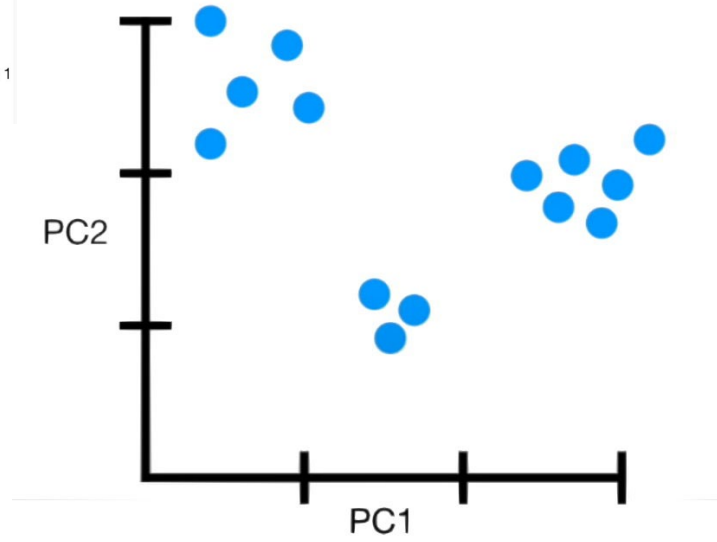
	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	2.2	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2	0.3	...
Gene5	1.3	1.6	1.6	0	...
Gene6	1.5	2	2.1	3	...
Gene7	1.1	2.2	1.2	2.8	...
Gene8	1	2.7	0.9	0.3	...
Gene9	0.4	3	0.6	0.1	





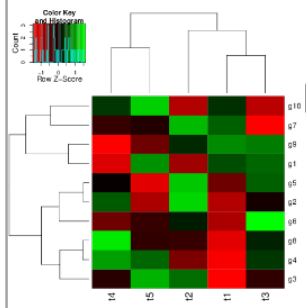


PCA (主成分分析)



(Source from StatQuest)

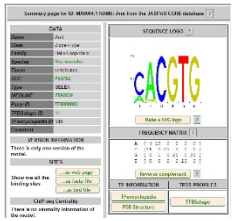
機器學習早
已融入到生物資訊
各個分析領域



Clustering, K-means
+
Functional Annotation

DAVID

g:Profiler **GSEA**



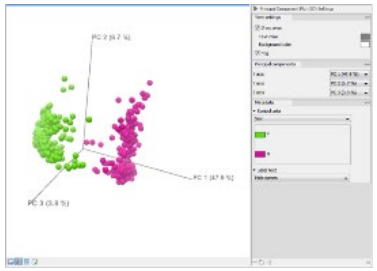
Transcription factor

Position specific scoring matrix (PSSM)

MEME

JASPAR

TRANSFAC



PCA (gene expression)

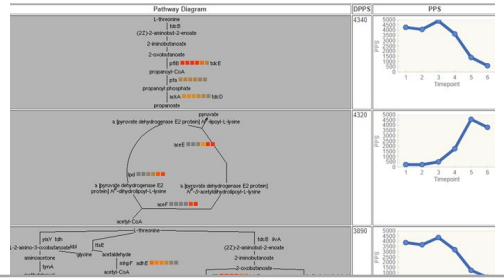
PCoA, NMDS (microbiota)

QIIME 2 (microbiota)

Protein structure:

AlphaFold

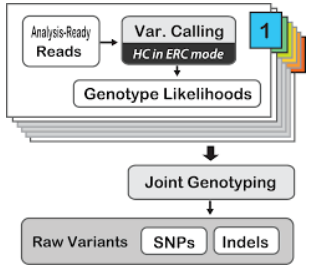
I-TASSER



Gene expression regulation pathway:

BioCyc

IPA



SNP and InDel caller:

GATK

Illumina DRAGEN (國網中心)

GWAS/PRS:

plink

PRSice-2 (R程式)

深度學習 Deep Learning

Radiomics 放射組學

生物醫學影像分析流程

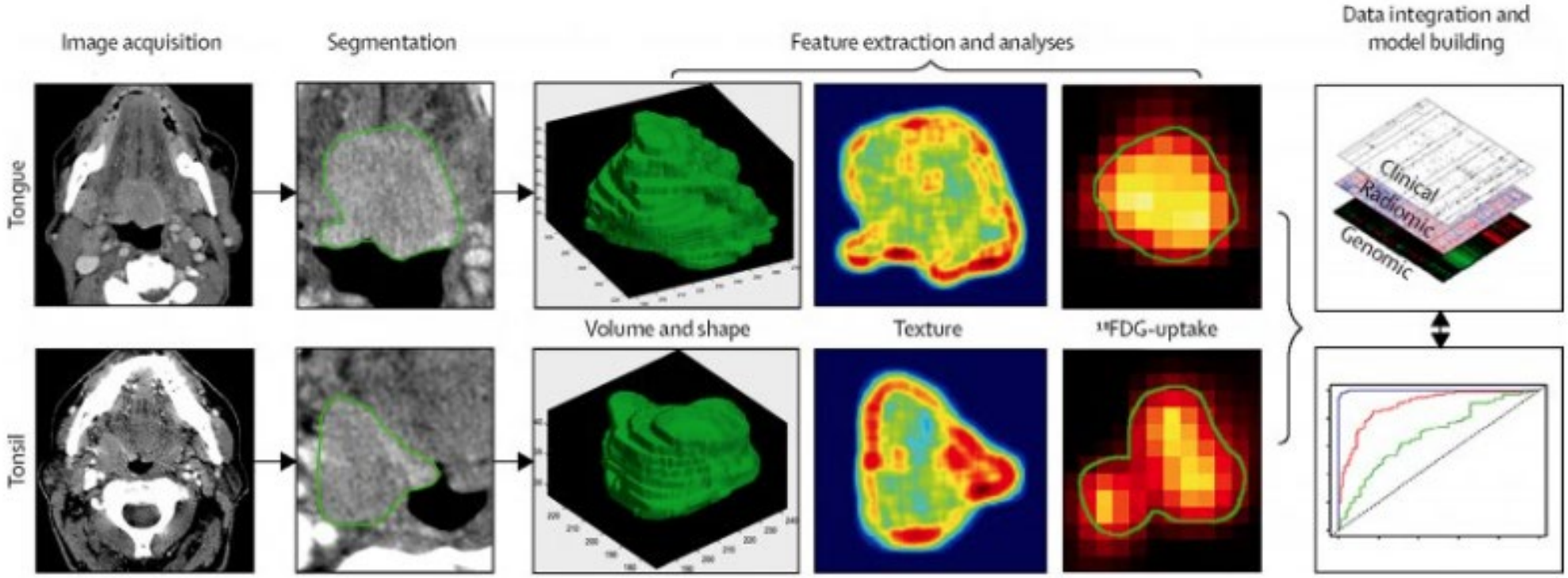
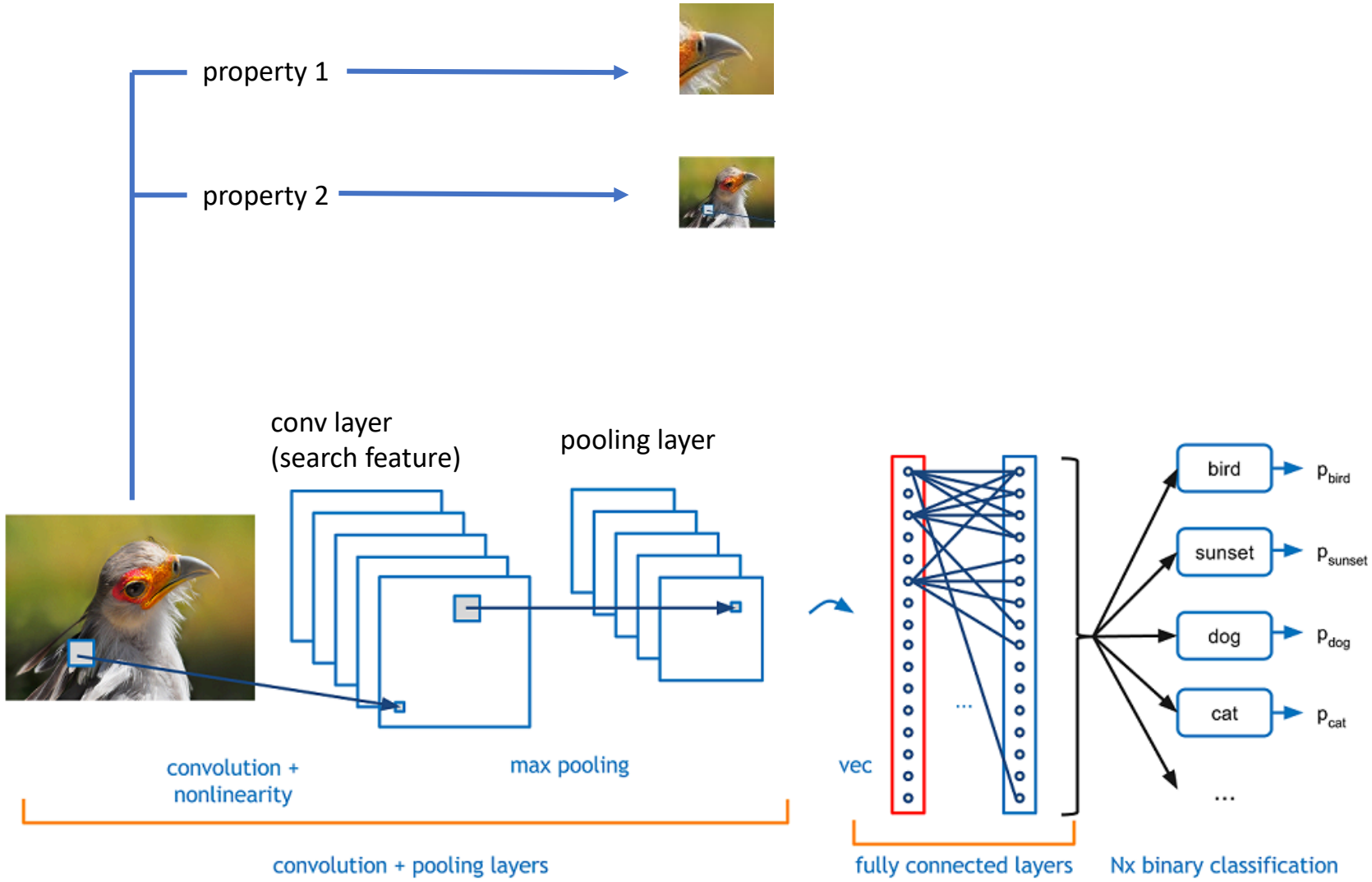


Illustration of the radiomics workflow

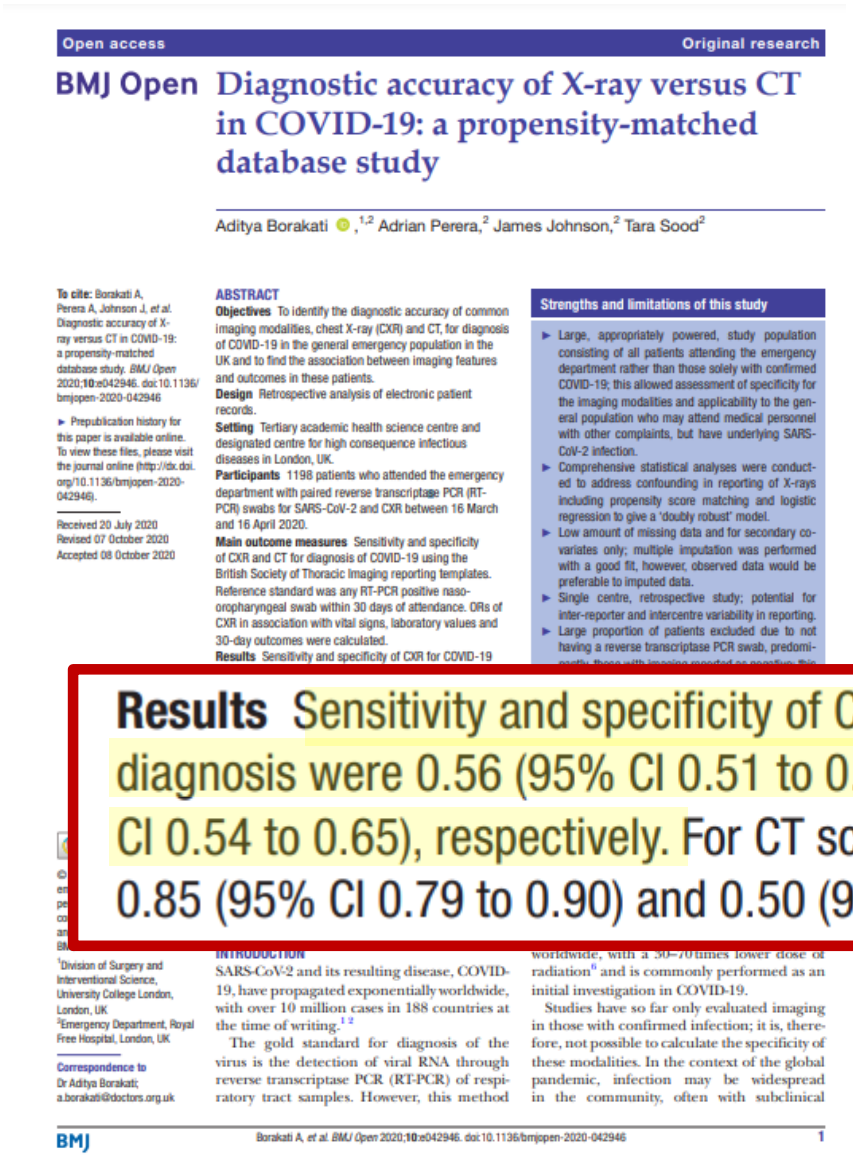
Philippe Lambin 2019

Convolutional Neural Networks



https://adeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convolutional-Neural-Networks/

靠人工判讀X光片無法正確診斷是否確診新冠肺炎



Open access Original research

BMJ Open

Diagnostic accuracy of X-ray versus CT in COVID-19: a propensity-matched database study

Aditya Borakati,[1,2] Adrian Perera,[2] James Johnson,[2] Tara Sood[2]

To cite: Borakati A, Perera A, Johnson J, *et al*. Diagnostic accuracy of X-ray versus CT in COVID-19: a propensity-matched database study. *BMJ Open* 2020;**10**:e042946. doi:10.1136/bmjopen-2020-042946

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-042946).

Received 20 July 2020
Revised 07 October 2020
Accepted 08 October 2020

ABSTRACT

Objectives To identify the diagnostic accuracy of common imaging modalities, chest X-ray (CXR) and CT, for diagnosis of COVID-19 in the general emergency population in the UK and to find the association between imaging features and outcomes in these patients.

Design Retrospective analysis of electronic patient records.

Setting Tertiary academic health science centre and designated centre for high consequence infectious diseases in London, UK.

Participants 1198 patients who attended the emergency department with paired reverse transcriptase PCR (RT-PCR) swabs for SARS-CoV-2 and CXR between 16 March and 16 April 2020.

Main outcome measures Sensitivity and specificity of CXR and CT for diagnosis of COVID-19 using the British Society of Thoracic Imaging reporting templates. Reference standard was any RT-PCR positive naso-oropharyngeal swab within 30 days of attendance. ORs of CXR in association with vital signs, laboratory values and 30-day outcomes were calculated.

Results Sensitivity and specificity of CXR for COVID-19

Strengths and limitations of this study

- Large, appropriately powered, study population consisting of all patients attending the emergency department rather than those solely with confirmed COVID-19; this allowed assessment of specificity for the imaging modalities and applicability to the general population who may attend medical personnel with other complaints, but have underlying SARS-CoV-2 infection.
- Comprehensive statistical analyses were conducted to address confounding in reporting of X-rays including propensity score matching and logistic regression to give a 'doubly robust' model.
- Low amount of missing data and for secondary covariates only; multiple imputation was performed with a good fit, however, observed data would be preferable to imputed data.
- Single centre, retrospective study; potential for inter-reporter and intercentre variability in reporting.
- Large proportion of patients excluded due to not having a reverse transcriptase PCR swab, predominantly those with imaging reported as negative, this

Results Sensitivity and specificity of CXR for COVID-19 diagnosis were 0.56 (95% CI 0.51 to 0.60) and 0.60 (95% CI 0.54 to 0.65), respectively. For CT scans, these were 0.85 (95% CI 0.79 to 0.90) and 0.50 (95% CI 0.41 to 0.60),

[1]Division of Surgery and Interventional Science, University College London, London, UK
[2]Emergency Department, Royal Free Hospital, London, UK

Correspondence to
Dr Aditya Borakati;
a.borakati@doctors.org.uk

INTRODUCTION

SARS-CoV-2 and its resulting disease, COVID-19, have propagated exponentially worldwide, with over 10 million cases in 188 countries at the time of writing.[1 2]

The gold standard for diagnosis of the virus is the detection of viral RNA through reverse transcriptase PCR (RT-PCR) of respiratory tract samples. However, this method

worldwide, with a 30–70 times lower dose of radiation[6] and is commonly performed as an initial investigation in COVID-19.

Studies have so far only evaluated imaging in those with confirmed infection; it is, therefore, not possible to calculate the specificity of these modalities. In the context of the global pandemic, infection may be widespread in the community, often with subclinical

只從X光片，確診者被診斷出確診的比率

Sensitivity = 0.56

只從X光片，非確診者被診斷非確診的比率

Specificity = 0.60

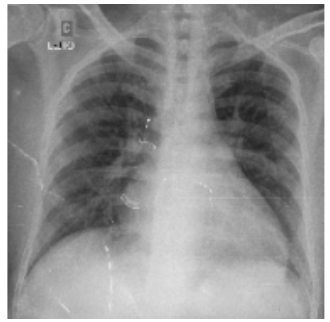
影像分析:胸腔X光判讀感染肺炎

正常(n =1,349)

非COVID-19病毒性肺炎(n=1,345)

細菌性肺炎(n=2,538)

COVID-19肺炎(n=3616)



資料來源: Kang Zhang et. al., 2018, *Cell*

Chowdhury et. al., 2020. *IEEE*

Mendeley Data

kaggle

開發框架:

- TensorFlow-21.02-tf2-py3

硬體設定:

- a100.2g.20gb

Convolutional neural networks(CNN)

開發框架:

- TensorFlow-21.02-tf2-py3

硬體設定:

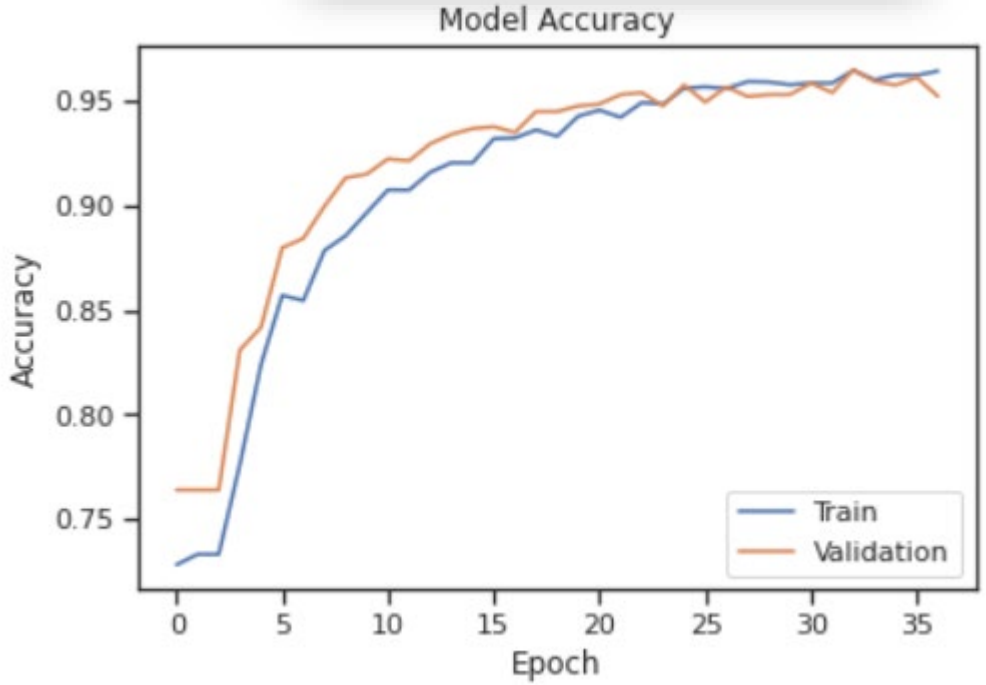
- a100.2g.20gb

Conv2D
MaxPooling2
Dropout
Conv2D
MaxPooling2
Dropout
Conv2D
Flatten
Dense
Dropout
Dense

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_55 (Conv2D)           (None, 68, 68, 128)       3584
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 34, 34, 128)       0
_________________________________________________________________
dropout_3 (Dropout)          (None, 34, 34, 128)       0
_________________________________________________________________
conv2d_56 (Conv2D)           (None, 32, 32, 64)        73792
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 16, 16, 64)        0
_________________________________________________________________
dropout_4 (Dropout)          (None, 16, 16, 64)        0
_________________________________________________________________
conv2d_57 (Conv2D)           (None, 14, 14, 64)        36928
_________________________________________________________________
flatten_1 (Flatten)          (None, 12544)             0
_________________________________________________________________
dense_2 (Dense)              (None, 16)                200720
_________________________________________________________________
dropout_5 (Dropout)          (None, 16)                0
_________________________________________________________________
dense_3 (Dense)              (None, 2)                 34
=================================================================
Total params: 315,058
Trainable params: 315,058
Non-trainable params: 0
_________________________________________________________________
CPU times: user 39.5 ms, sys: 4.12 ms, total: 43.6 ms
Wall time: 42.6 ms
```

13808 samples

全部運行完畢時間: 約4min 30s



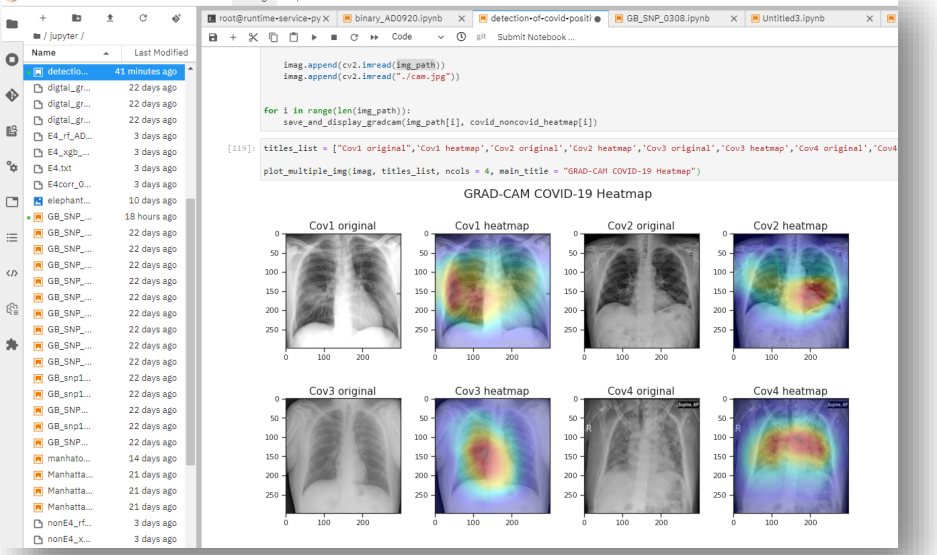
只從X光片，確診者被診斷出確診的比率

Sensitivity = 0.93

只從X光片，非確診者被診斷非確診的比率

Specificity = 0.96

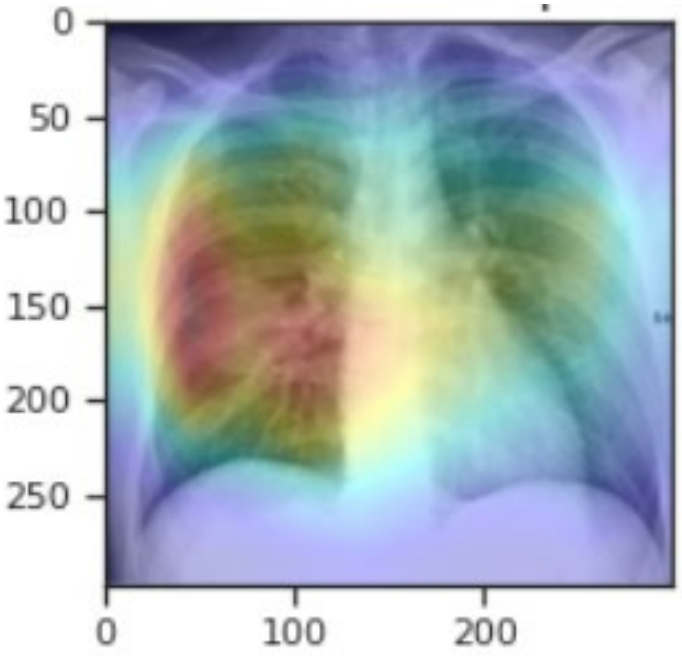
找出AI判讀X光片為確診的依據



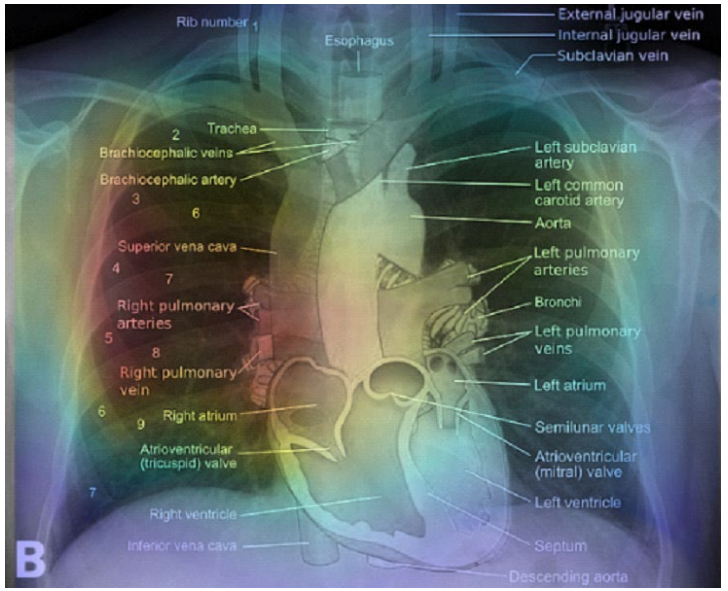
X-ray

Heatmap

Grad - CAM Algorithm



加上註釋



偏紅色為AI重要的參考區域

神經網路與深度學習

Neural networks and deep learning

1

Python or Julia

Program language

2

TensorFlow or PyTorch

Software library for machine learning

3

Keras

Deep learning API

https://drive.google.com/drive/folders/1ITEhTJqNivPWu5h7rZsGmOeQwzlFZTT?usp=sharing

神經網路在基因體學的應用

2019

SpliceAI

nature REVIEWS GENETICS

Research Highlight | Published: 25 January 2019

RNA

Learning the language of splicing

Dorothy Clyde

Nature Reviews Genetics **20**, 132–133 (2019) | Download Citation

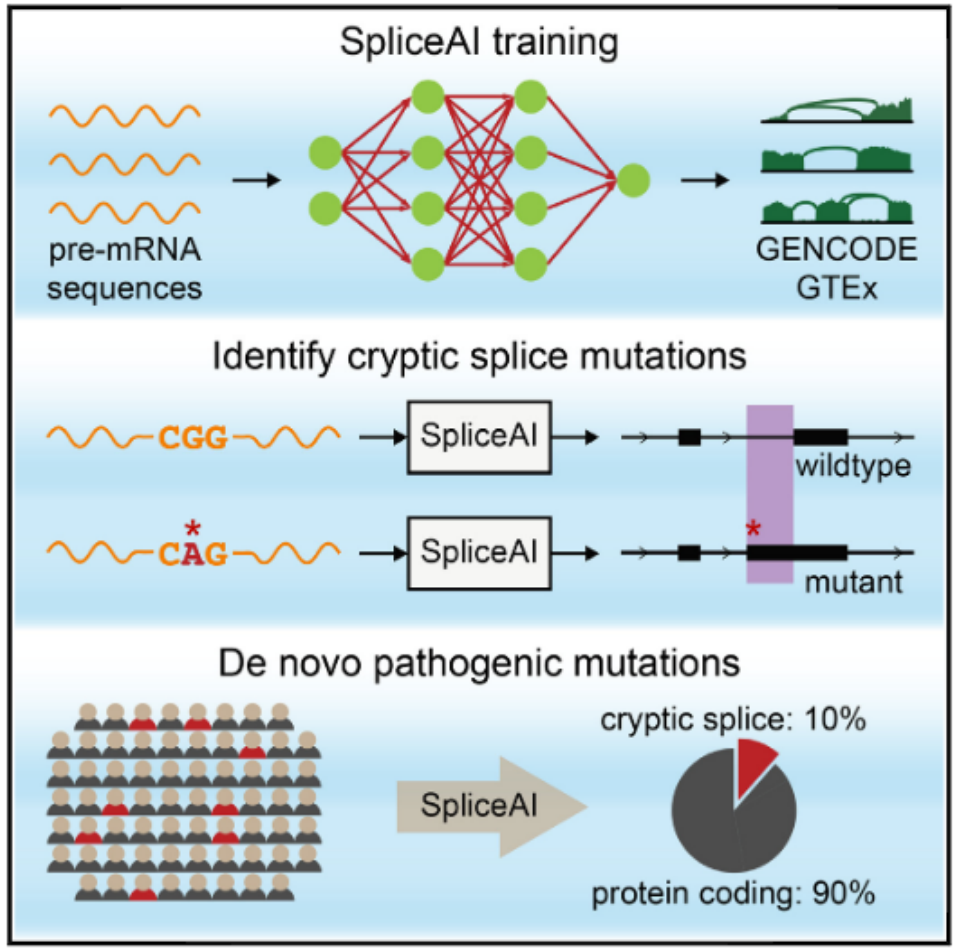
Precise splicing of pre-mRNAs is essential for proper gene function, and defective splicing can lead to disease. However, our understanding of the sequence features that underlie the accuracy of this process remains incomplete, which makes it difficult to identify genetic variation that could disrupt it. Now, a paper in *Cell* describes SpliceAI, a deep neural network that accurately predicts not only splice sites in previously unseen pre-mRNA sequence but also the effects of sequence variants on splicing patterns.

Article

Cell

Predicting Splicing from Primary Sequence with Deep Learning

Graphical Abstract



Authors

Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F. McRae, ..., Serafim Batzoglou, Stephan J. Sanders, Kyle Kai-How Farh

Correspondence

kfarh@illumina.com

In Brief

A deep neural network precisely models mRNA splicing from a genomic sequence and accurately predicts noncoding cryptic splice mutations in patients with rare genetic diseases.

Modeling

Input: pre-mRNA sequences

CATGCATCTCGA**G**CATGCGTACACT

40 nt	40 nt
200 nt	200 nt
1000 nt	1000 nt
5000 nt	5000 nt

Convolutional Neural Networks (CNNS)

Residual neural network (ResNet)

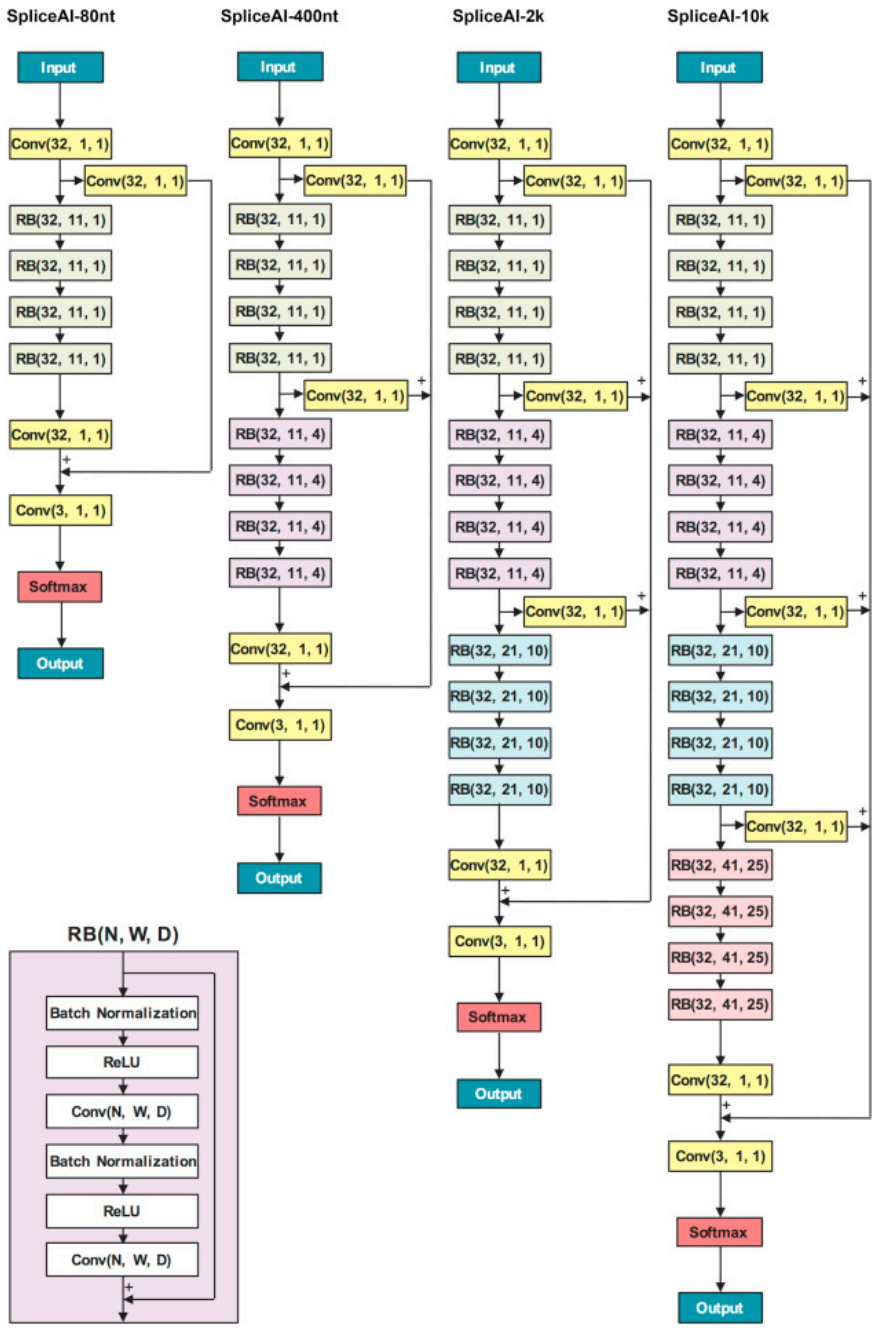
Output: probability

splice acceptor Loss/Gain
splice donor Loss/Gain
neither

variant: chr6-2317763-T-A

Acceptor Loss

pre-mRNA position: 106 bp



(Jaganathan et al., Cell, 2019)

Training and Testing

- **hg19/GRCh37 Genome assembly**
- **GENCODE Version 24lift37**

bin	name	chrom	strand	txStart	txEnd	cdsStart	cdsEnd	exonCount	exonStarts	exonEnds	score	name2	cdsStartStat	cdsEndStat	exonFrames
585	ENST00000473358.1	chr1	+	29553	31097	29553	29553	3	29553,30563,30975,	30039,30667,31097,	0	RP11-34P13.3	none	none	-1,-1,-1,
585	ENST00000469289.1	chr1	+	30266	31109	30266	30266	2	30266,30975,	30667,31109,	0	RP11-34P13.3	none	none	-1,-1,
585	ENST00000417324.1	chr1	-	34553	36081	34553	34553	3	34553,35276,35720,	35174,35481,36081,	0	FAM138A	none	none	-1,-1,-1,
585	ENST00000461467.1	chr1	-	35244	36073	35244	35244	2	35244,35720,	35481,36073,	0	FAM138A	none	none	-1,-1,
585	ENST00000335137.3	chr1	+	69090	70008	69090	70008	1	69090,	70008,	0	OR4F5	cmpl	cmpl	0,
585	ENST00000466430.5	chr1	-	89294	120932	89294	89294	4	89294,92090,112699,120774,	91629,92240,112804,120932,	0	RP11-34P13.7	none	none	-1,-1,-1,-1,
585	ENST00000495576.1	chr1	-	89550	91105	89550	89550	2	89550,90286,	90050,91105,	0	RP11-34P13.8	none	none	-1,-1,
585	ENST00000477740.5	chr1	-	92229	129217	92229	92229	4	92229,112699,120720,129054,	92240,112804,120932,129217,	0	RP11-34P13.7	none	none	-1,-1,-1,-1,
585	ENST00000471248.1	chr1	-	110952	129173	110952	110952	3	110952,112699,129054,	111357,112804,129173,	0	RP11-34P13.7	none	none	-1,-1,-1,
73	ENST00000610542.1	chr1	-	120724	133723	120724	120724	4	120724,120873,129054,133373,	120869,120932,129223,133723,	0	RP11-34P13.7	none	none	-1,-1,-1,-1,

20,287 protein-coding gene

removed genes without any splice junctions

~90% → training

~10% → testing 1

chromosomes 2, 4, 6, 8, 10-22, X, and Y
(13,384 genes, 130,796 donor-acceptor pairs)

chromosomes 1, 3, 5, 7, and 9 without any paralogs
(1,652 genes, 14,289 donor-acceptor pairs)

GTEx Portal
The Genotype-Tissue Expression project

95% accuracy !

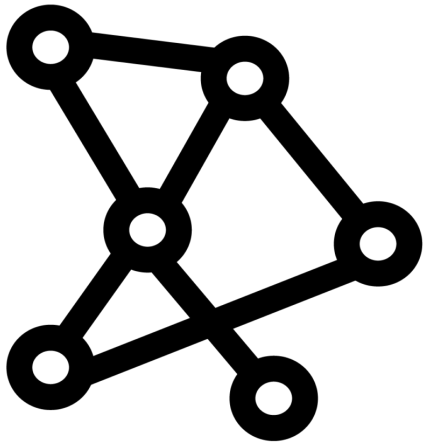
正確觀念

不預設立場

多面向觀察

＋

正確方法



邏輯正確

結果符合生物學常識

＝

有效的精準醫療

肥胖是一種疾病，是心血管疾病與糖尿病元凶?

Dr. Jaime Guevara-Aguirre　Dr. Valter Longo

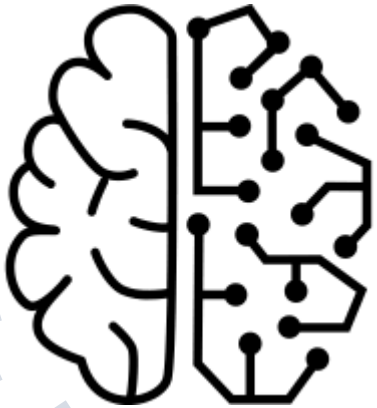
Dr. Jaime Guevara-Aguirre　Dr. Arlan Rosenbloom

	厄瓜多小人（西班牙後裔）	一般厄瓜多人
肥胖 Obese	20%	12%
糖尿病 Diabetes	~0%	5%
癌症 Cancer	~0%	22%

Little People of Ecuador: Laron Syndrome 生長激素無法刺激肝臟分泌胰島素生長因子IGF1

機器學習

精準醫療



基因體學

代謝體學

腸道菌體學

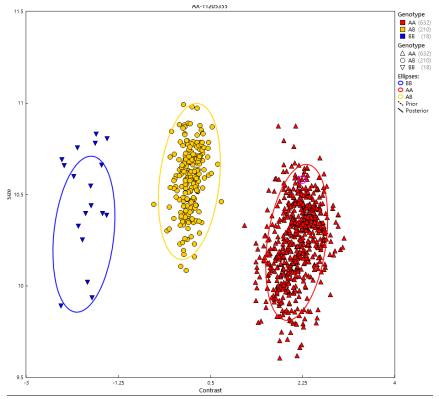
性別

年齡

人種

族群（肥胖族群， 抽菸族群）

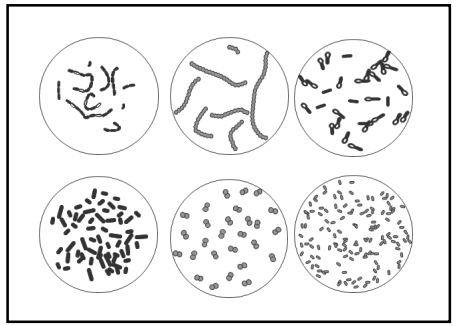
Genotyping



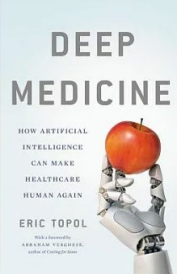
Metabolites data

Fat No	C0	C2	C3	C4	C5	C5:1-DC	C7-DC	C8	C10	C12	C14:1	C16	C18	C18:1	C18:2
C0200400	-0.28348	0.174167	-0.12124	0.075949	0.244419	-0.502500341	0.026967	0.269461	0.106087	0.96384	0.254183	0.747937	1.082462	-0.11301	0.181447
C0200402	0.297066	-0.10645	0.118549	0.812914	0.096573	-1.087462841	0.594008	-0.01294	-0.21853	-0.35443	-0.05394	-0.04474	0.258312	0.009648	0.422455
C0200407	-0.23575	-0.89711	-0.50371	-0.33498	-0.54663	-1.087462841	0.026967	0.18133	-0.05914	-0.39142	-0.6389	-0.94596	-1.28011	-0.84997	0.086588
C0200427	-0.10524	0.446385	-0.50371	-0.614451	-0.53051	-0.502500341	0.557482	0.146841	0.184656	0.439821	0.786405	0.195396	0.477322	0.244804	0.220266
C0200431	-0.04417	-0.70447	0.218754	0.032133	0.290515	-0.917537784	-0.68353	0.927355	0.621199	-0.46839	-0.26039	-0.31918	-0.3505	-0.27096	0.508612
C0200435	0.523247	0.331154	0.544424	0.308872	0.73252	-0.502500341	1.040264	-0.06589	0.056806	0.083256	0.17433	0.204982	-0.21299	0.4468	0.400087
C0200436	-0.12487	0.342307	-0.49344	-0.57525	-0.49881	-1.087462841	0.401363	0.012824	-0.02926	0.096676	0.381563	0.306427	0.304855	0.167727	0.365871
C0200440	0.17073	0.390688	0.361358	-0.16506	0.317482	-0.280107919	0.442005	-0.29166	-0.14622	0.014214	0.401741	0.010971	-0.3505	-0.07088	-0.06075
C0200448	0.173947	-0.78193	0.171136	-0.08706	-0.152	-1.502500341	-0.77372	-0.91254	-0.86197	-0.98579	-0.74502	-0.55769	-0.38702	-0.68203	-0.09212
C0200455	0.17073	0.085923	0.209355	0.277046	0.253757	-0.502500341	0.154723	0.311586	0.401942	0.083256	0.329848	0	0	-0.11301	0.058299
C0200465	-0.01092	0.550344	0.057003	-0.01963	-0.29445	-0.502500341	0.27208	0.572023	0.264	-0.24882	0.026232	-0.3055	-0.5025	-0.21254	0.220266
C0200468	-0.3102	-0.61546	-0.37561	0.153951	-0.70044	-0.628031223	-0.40599	-0.50357	-0.76033	-0.82532	-0.5014	-0.18793	-0.11784	-0.42164	0.154975
C0200474	0.23374	-0.05538	1.044724	0.025763	0.043721	-1.087462841	0.026967	-0.10692	-0.11451	-0.21527	0.401741	0.085518	0.304855	0.019629	0.029444
C0200479	-0.37453	-0.05956	-0.49344	-0.61451	-0.73697	-1.087462841	0.316474	-0.14202	-0.03519	-0.07325	0.254183	-0.55769	-0.62803	0.019629	-0.53382
C0200483	0.258193	0.147166	0.392144	-0.30256	-0.03368	-0.387023123	0.731511	0.854423	0.791542	0.149143	-0.13883	0.375968	-0.21299	0.132104	0.245577
C0200484	-0.08588	0.265715	-0.31107	-0.23217	-0.57942	0	0.421827	0.567685	0.901085	0.471072	0.498602	-0.31918	0.055495	-0.13454	-0.22487
C0200485	-0.3605	-1.14598	-0.69048	-0.48153	-0.43741	-0.502500341	-0.17333	-1.1635	-1.15265	-1.0435	-0.76818	-0.8635	-0.71823	-1	-0.92749
C0200487	0.377396	0.643454	0.54193	0.462972	-0.18949	-0.502500341	0.27208	-0.07943	0.179549	-0.07325	0.126633	0	0.13493	0.454032	0.128008
C0200492	-0.342	-0.22609	-0.1173	-0.67547	-0.51457	-0.917537784	-0.30166	1.280108	1.432315	0.319068	0.026232	-0.26524	-0.14886	-0.11301	-0.66918
C0200494	0.186743	-0.1594	0.138494	-0.05295	0.606989	-0.087462841	0.421827	-0.06589	-0.14622	0	-0.09576	-0.05614	0.210218	-0.1564	-0.55552
C0200496	-0.04417	-0.0223	-0.62177	0.171368	-0.20219	-0.502500341	-0.40599	-1.09311	-0.99422	-0.75132	-0.44626	-0.33298	-0.31487	-0.43494	-0.87199

Gut microbiota data



參考書籍



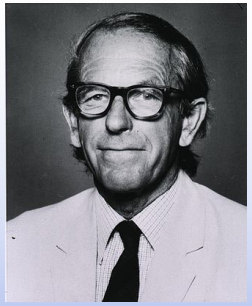
Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, Topol, Eric, 2019

精準醫學：早期預防癌症，破解基因迷思對症下藥，曾嶔元，2020

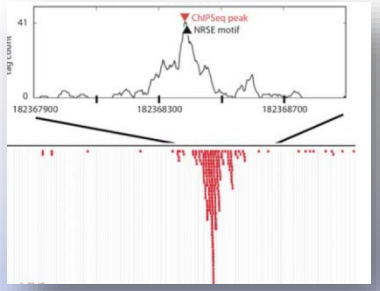
最後彩蛋

1. 人工智慧與定序都是老技術新發展，並開創熱潮

Sequencing



Frederick Sanger
1975
Dideoxynucleotide termination

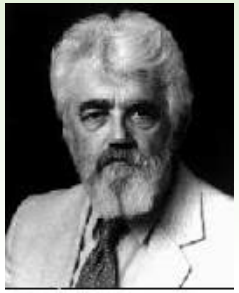


(Johnson et al, Science 2007)

2007 first real application of Next Generation Sequencing

NGS

AI



John McCarthy
1956
coined the term "artificial intelligence" (AI)

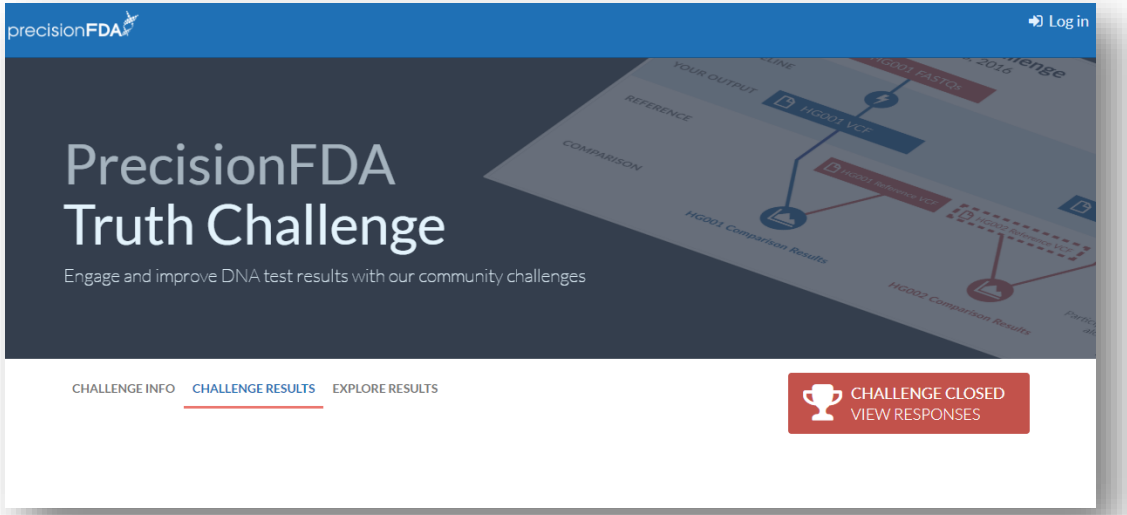
Geoffrey Hinton
2006
Deep learning

Deep Learning

2. Deep learning與NGS第一次結合是2016年，Google贏得SNP variant (from DNA-seq) 預測大賽

Google
Google ❤ Open Source

DeepVariant

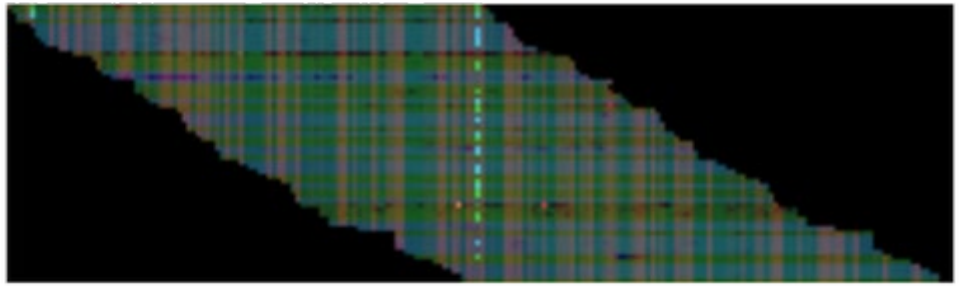


precisionFDA

DeepVariant

(Poplin et al., Nature Biotechnology, 2018)

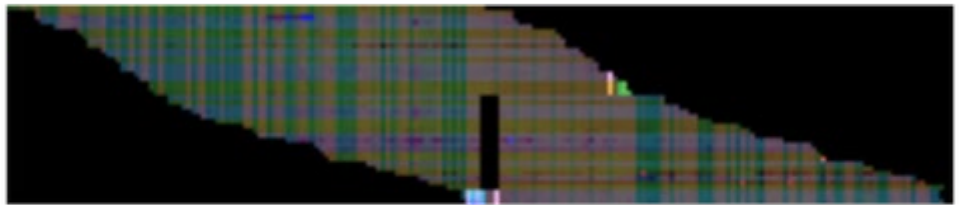
A



SNP

true SNP on one chromosome pair

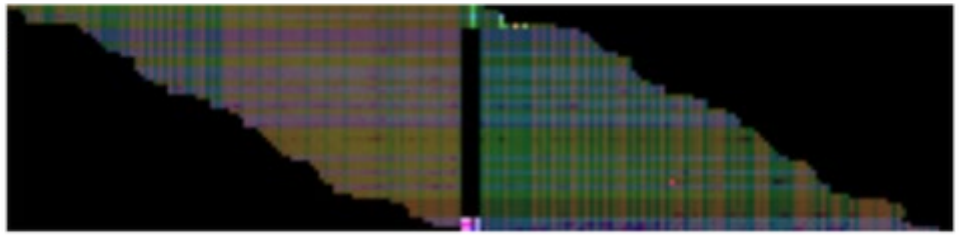
B



Hetero Deletion

a deletion on one chromosome

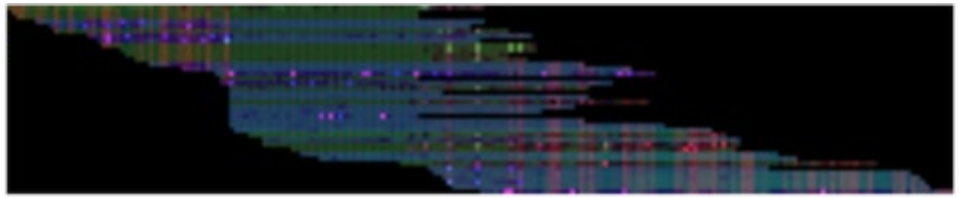
C



Homo Deletion

a deletion on both chromosomes

D

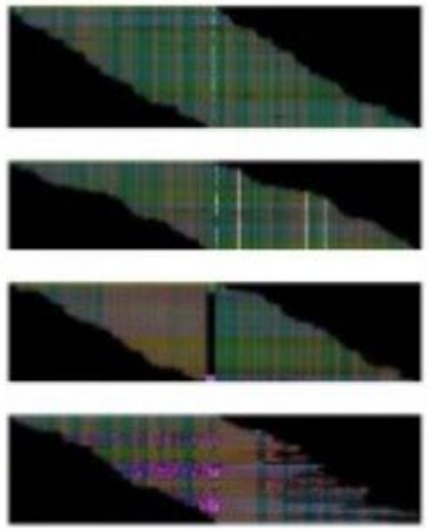


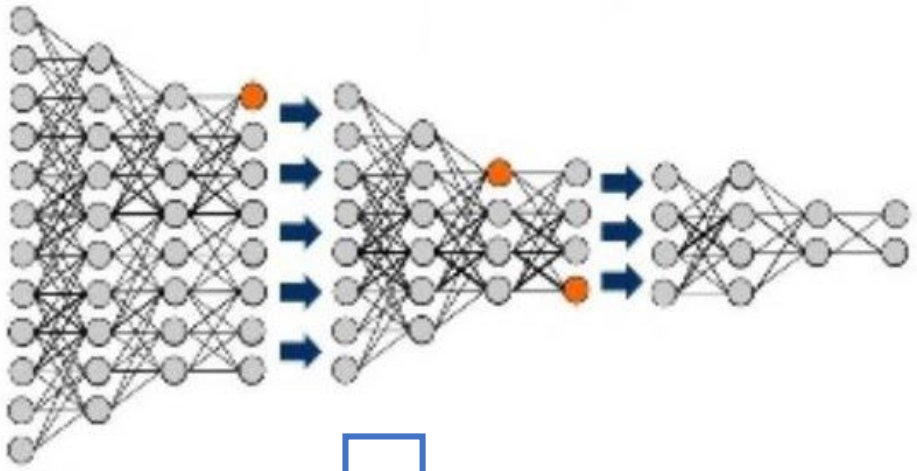
False Positive

a false variant caused by errors

Red = {ACGT} Green={quality score} Blue={read strand}

Convolutional neural network (CNN)





Genotype

Homozygous reference: 0.01

Homozygous variant: **0.95**

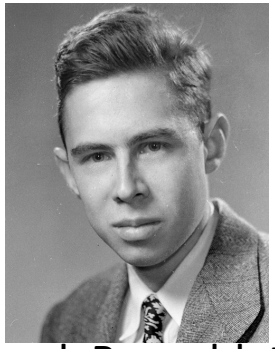
Heterozygous: 0.04

99.9% accuracy

Homozygous variant call

(Poplin et al., Nature Biotechnology, 2018)

3. 1958年，Frank Rosenblatt當初為了解開大腦的奧秘，模仿神經元設計了感知器，是人工智慧的先驅，也是AI神經網絡最小單位



Frank Rosenblatt
(American psychologist)

Vol. VI, No. 2, Summer 1958

research trends

CORNELL AERONAUTICAL LABORATORY, INC., BUFFALO 21, NEW YORK

The Design of an

Intelligent AUTOMATON

by FRANK ROSENBLATT

Introducing the perceptron — A machine which senses, recognizes, remembers, and responds like the human mind.

STORIES about the creation of machines having human qualities have long been a fascinating province in the realm of science fiction. Yet we are now about to witness the birth of such a machine — a machine capable of perceiving, recognizing, and identifying its surroundings without any human training or control.

Development of that machine has stemmed from a search for an understanding of the physical mechanisms which underlie human experience and intelligence. The question of the nature of these processes is at least as ancient as any other question in western science and philosophy, and, indeed, ranks as one of the greatest scientific challenges of our time.

Our understanding of this problem has gone perhaps as far as had the development of physics before Newton. We have some excellent descriptions of the phenomena to be explained, a number of interesting hypotheses, and a little detailed knowledge about events in the nervous system. But we lack agreement on any integrated set of principles by which the functioning of the nervous system can be understood.

We believe now that this ancient problem is about to yield to our theoretical investigation for three reasons:

First, in recent years our knowledge of the functioning of individual cells in the central nervous system has vastly increased.

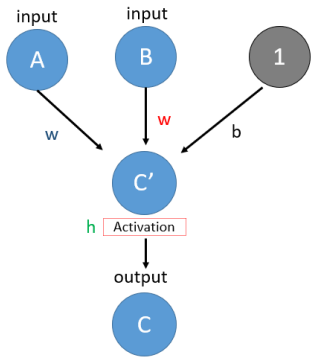
Second, large numbers of engineers and mathematicians are, for the first time, undertaking serious study of the mathematical basis for thinking, perception, and the handling of information by the central nervous system, thus providing the hope that these problems may be within our intellectual grasp.

Third, recent developments in probability theory and in the mathematics of random processes provide new tools for the study of events in the nervous system, where only the gross statistical organization is known and the precise cell-by-cell "wiring diagram" may never be obtained.

Receives Navy Support

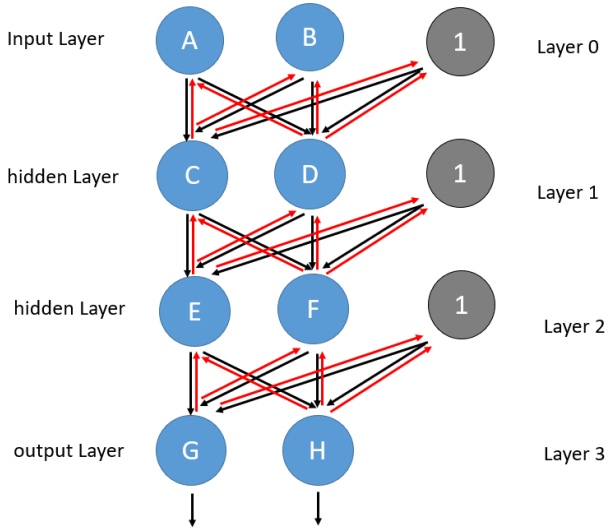
In July, 1957, Project PARA (Perceiving and Recognizing Automaton), an internal research program which had been in progress for over a year at Cornell Aeronautical Laboratory, received the support of the Office of Naval Research. The program had been concerned primarily with the application of probability theory to

感知器 Perceptron

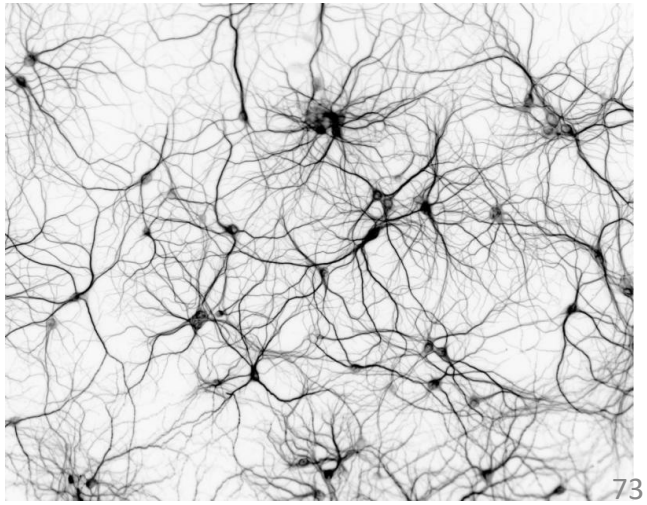


神經元 Neuron

AI神經網絡 Neural Network



神經網路 Nerve Net



Published: 01 January 1966

Transfer of Conditioned Responses from Trained Rats to Untrained Rats by Means of a Brain Extract

FRANK ROSENBLATT, JOHN T. FARROW & WILLIAM F. HERBLIN

Nature **209**, 46–48 (1966) | Cite this article

Thanks for your attention~

My Email: paul@nhri.edu.tw