資料探勘 (Data Mining)

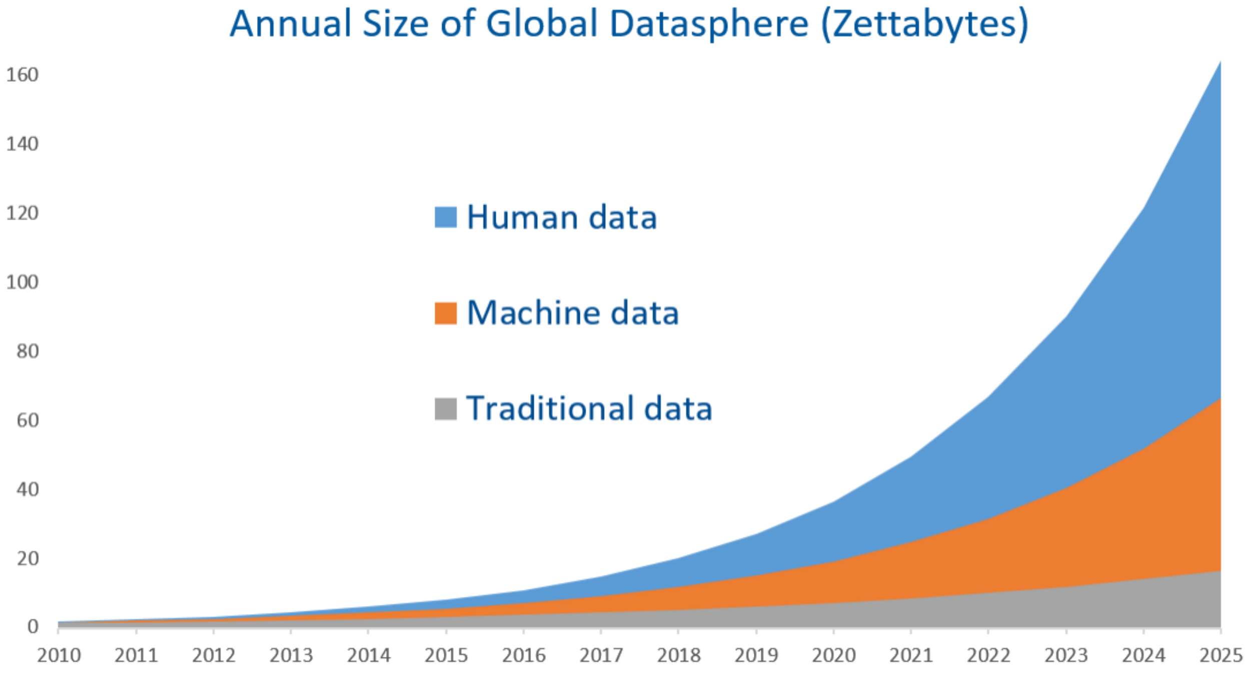
Dr. Tun-Wen Pai

1) Data Mining Introduction
2) Data cleansing and preparation
3) Evaluating Data Mining Techniques
4) Confusion matrix AUC - ROC curve
5) Tools and Platforms for Data Mining

March 8, 2023

- **Data mining** is the art and science of discovering knowledge, insights, and patterns in data. It is the act of extracting useful patterns from an organized collection of data.
- Data mining is a multidisciplinary field. It draws modeling and analytical techniques from **statistics** and **computer science** (artificial intelligence) areas. It also draws the knowledge of decision-making from the field of **business management**.
- The field of data mining emerged in the context of **pattern recognition**.

- The total amount of data in the world is doubling every 18 months (**Moore's law**)
- Gathering and curating data takes time and effort, particularly when it is unstructured or semi-structured.
- Knowledge domain helps select the right streams of data for pursuing new insights. Only the data that suits the nature of the problem being solved should be gathered.

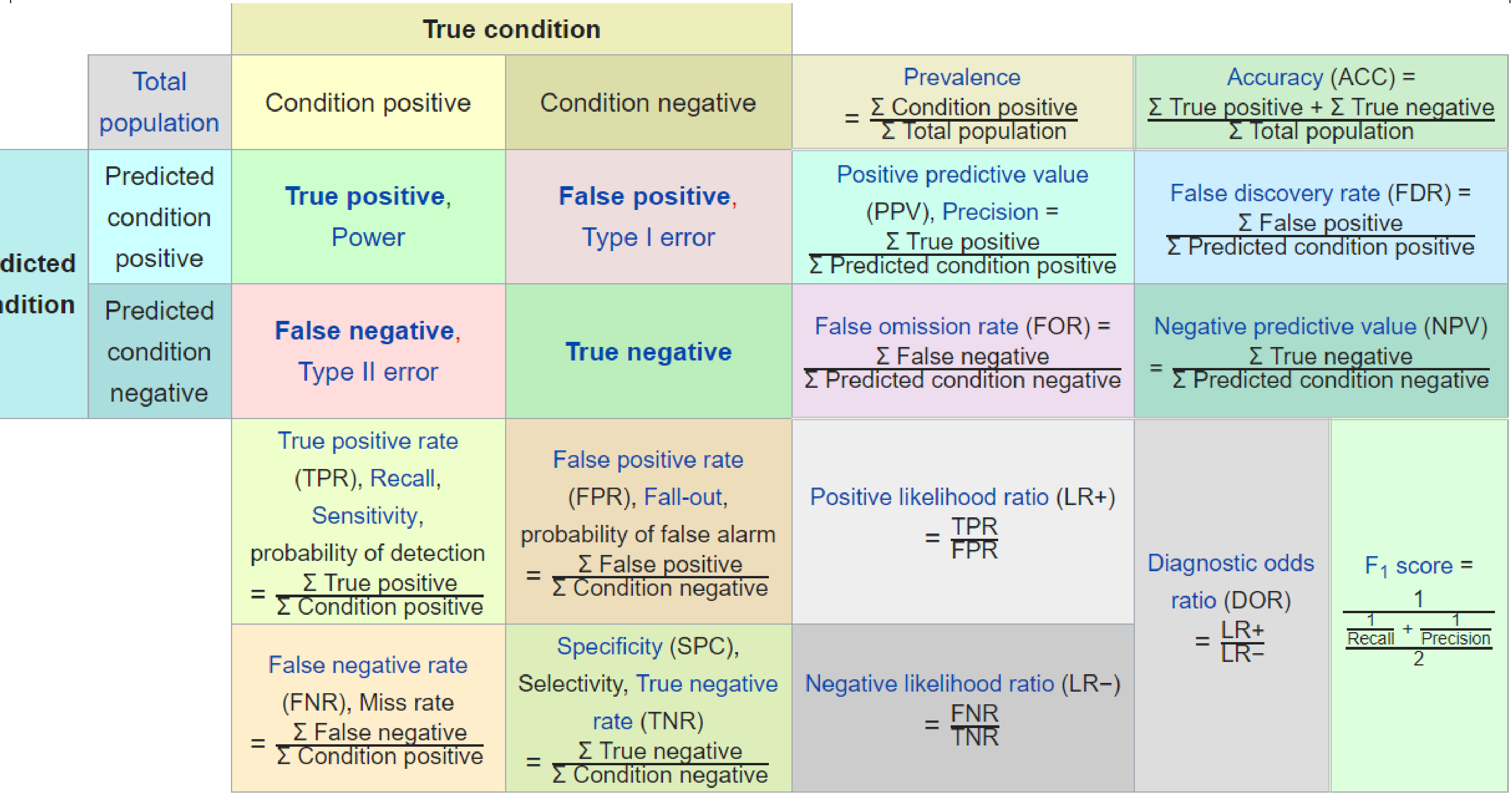


Source: IDC, insideBIGDATA

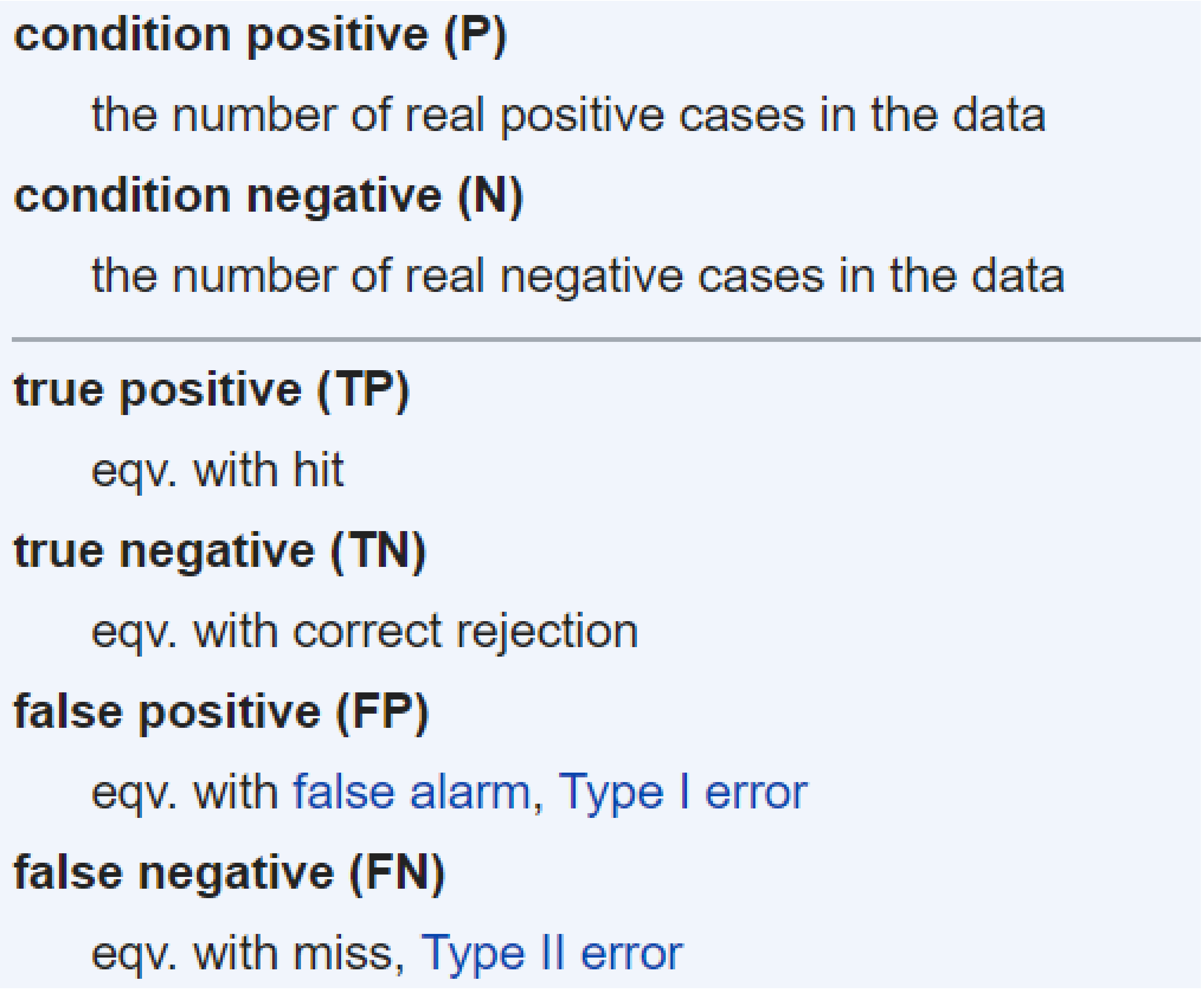
Data cleansing and preparation

- The **quality** of data is critical to the success and value of the data mining project. Otherwise, the situation will be of the kind of **garbage in and garbage out** (GIGO).
- Data almost certainly needs to be **cleaned** and **transformed** before it can be used for data mining. removing duplicate data, filling missing values, reigning in the effects of outliers, transforming fields (comparable data elements), binning continuous variables, increasing information density. …. before it can be ready for analysis.
- Labor-intensive or semi-automated activity that can take up to 60-80% of the time needed for a data mining project.

Evaluating Data Mining Results
Confusion matrix (error matrix)



		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
dicted dition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score = $\frac{1}{\frac{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}{2}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		



condition positive (P)

the number of real positive cases in the data

condition negative (N)

the number of real negative cases in the data

---

true positive (TP)

eqv. with hit

true negative (TN)

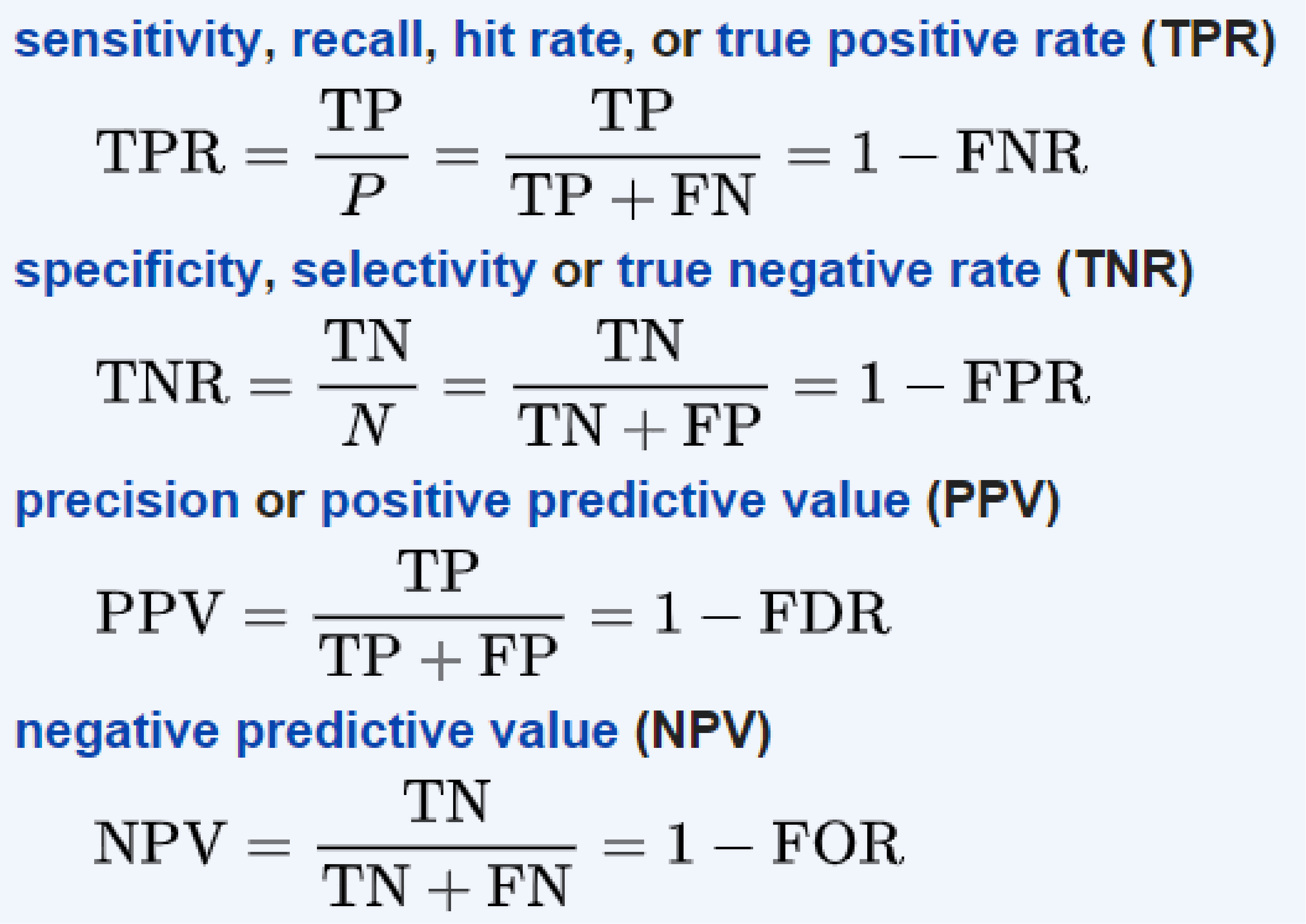
eqv. with correct rejection

false positive (FP)

eqv. with false alarm, Type I error

false negative (FN)

eqv. with miss, Type II error



sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$

specificity, selectivity or true negative rate (TNR)

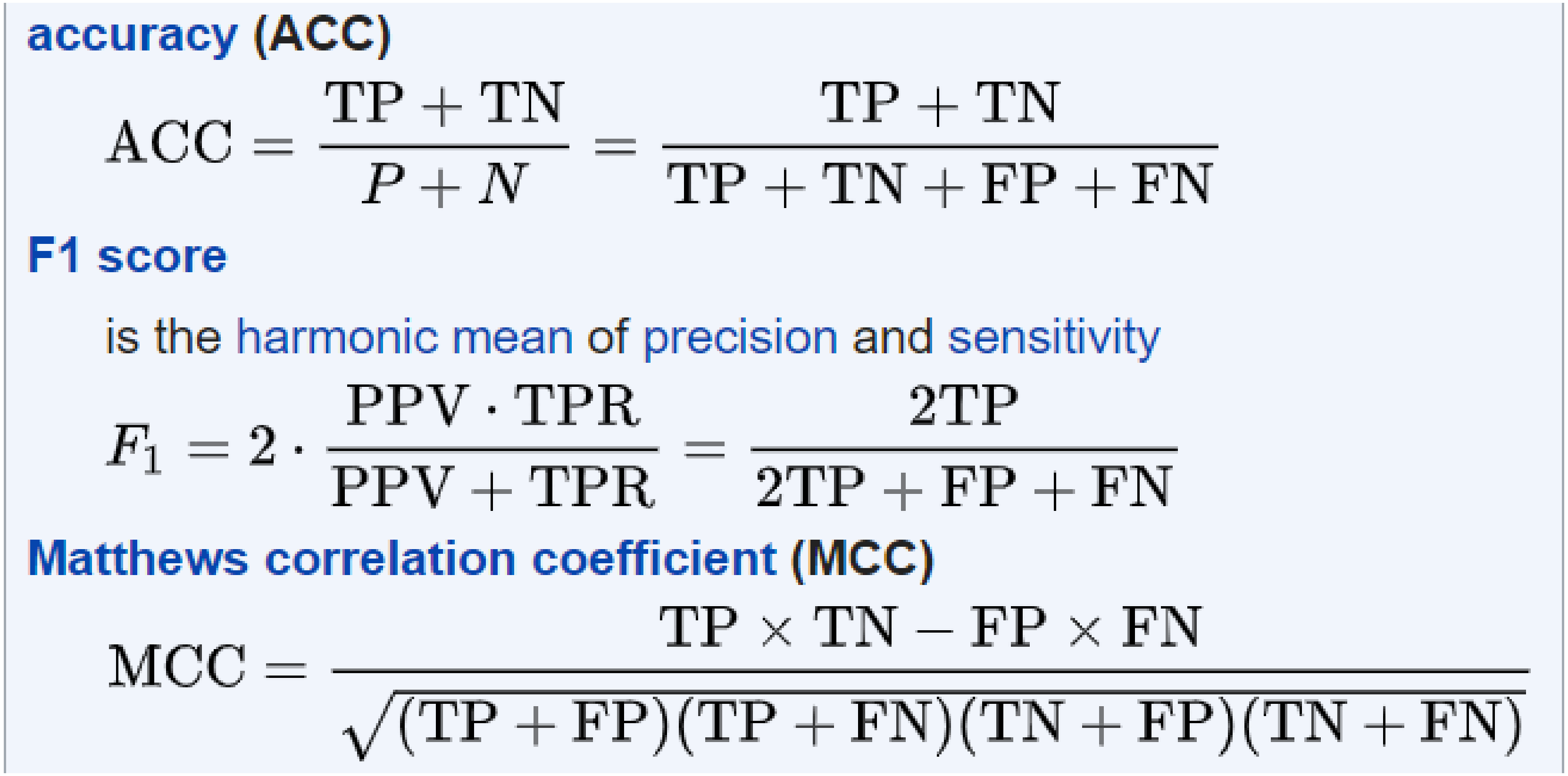
$$\mathrm{TNR} = \frac{\mathrm{TN}}{N} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} = 1 - \mathrm{FPR}$$

precision or positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} = 1 - \mathrm{FDR}$$

negative predictive value (NPV)

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} = 1 - \mathrm{FOR}$$



accuracy (ACC)

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{P+N} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

F1 score

is the harmonic mean of precision and sensitivity

$$F_1 = 2\cdot\frac{\mathrm{PPV}\cdot\mathrm{TPR}}{\mathrm{PPV}+\mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$$

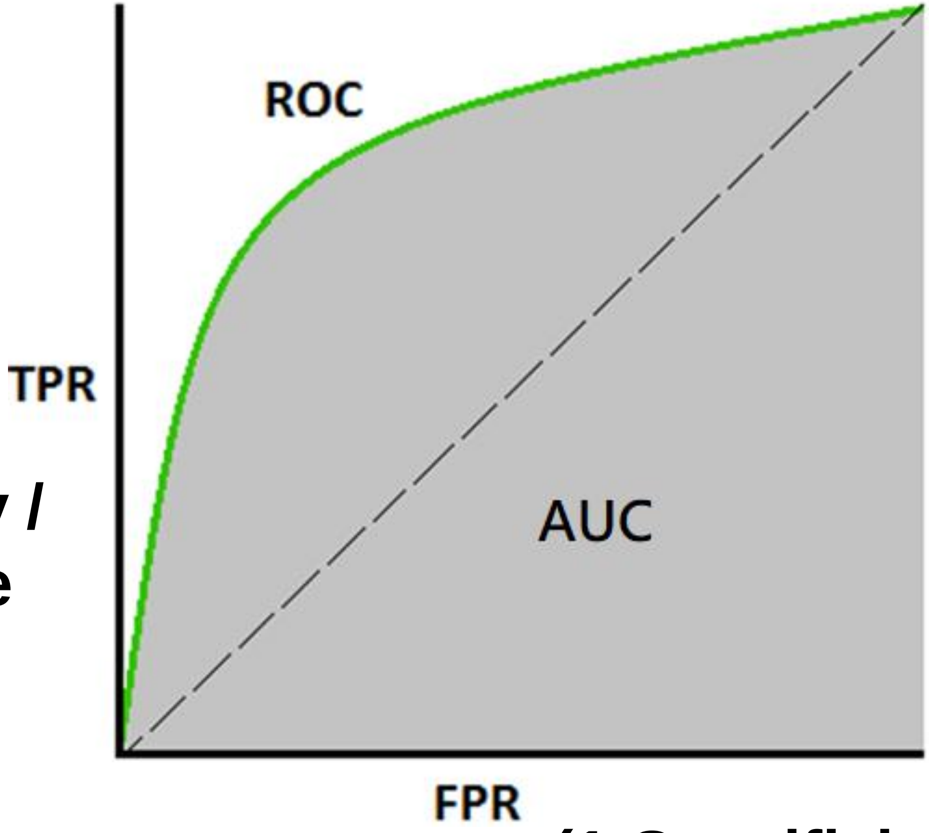
Matthews correlation coefficient (MCC)

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

The highest value can be 100%. In practice, predictive models with more than 70% accuracy can be considered usable in business domains, depending upon the nature of the business.

Performance measurement

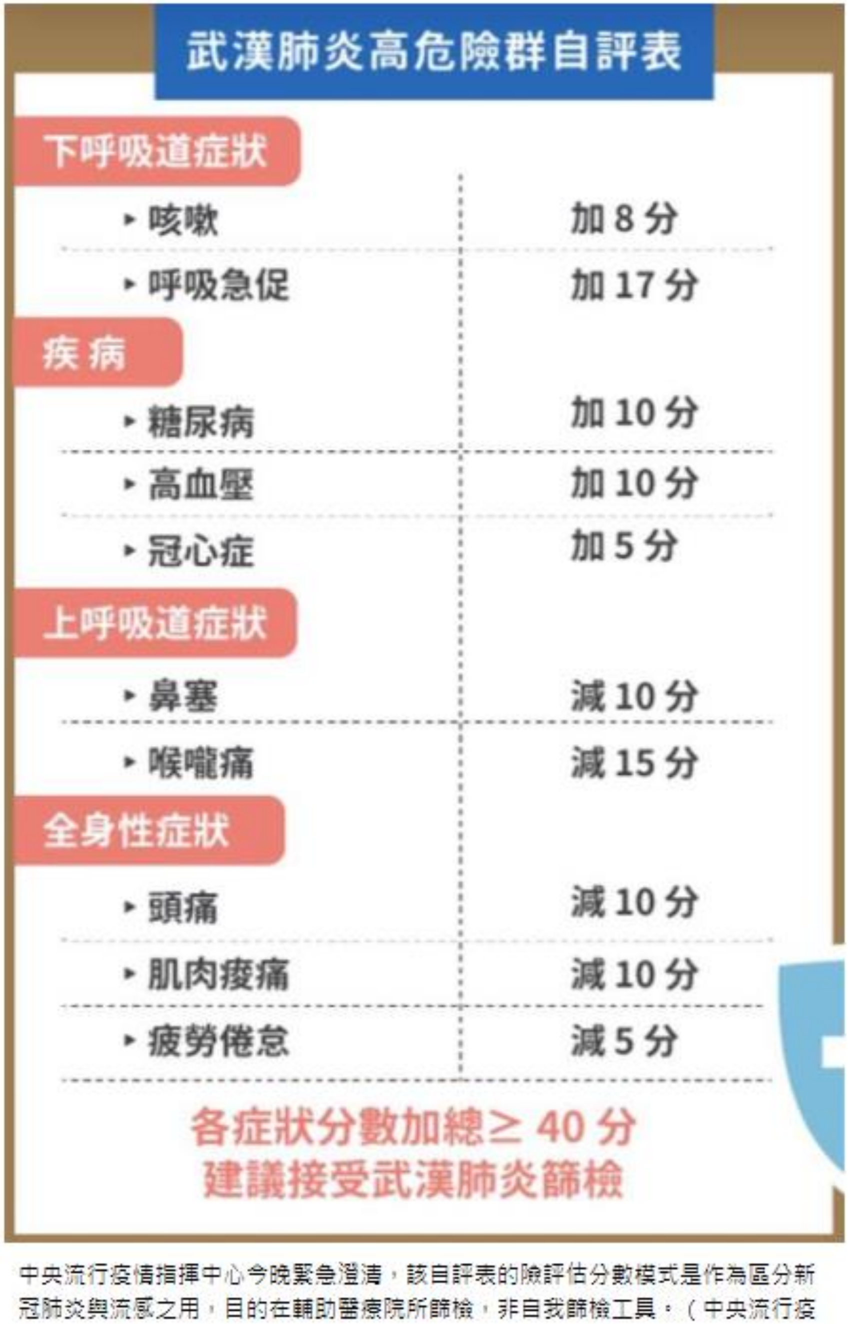
- For classification problem, measurement based on an AUC - ROC Curve, AUC - ROC curve is a performance measurement for classification problem at various thresholds settings.
- AUC (**Area Under The Curve**)
- ROC (**Receiver Operating Characteristics**)
- AUROC (**Area Under the Receiver Operating Characteristics**)
- Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.



Sensitivity / Recall rate

(1-Specificity)

Figures adopted from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

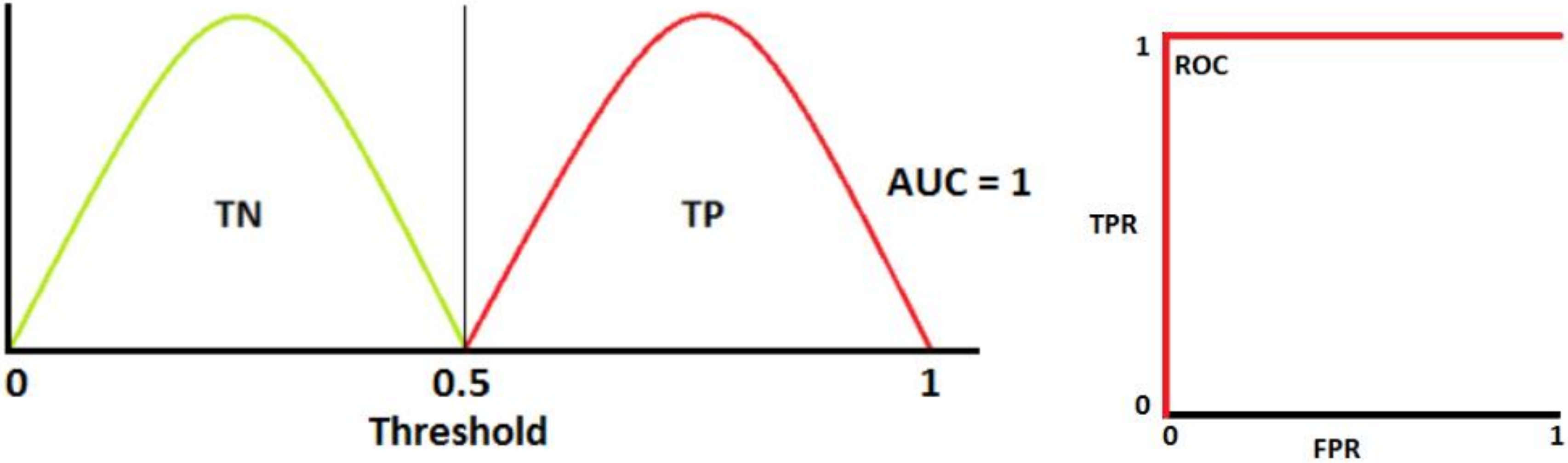


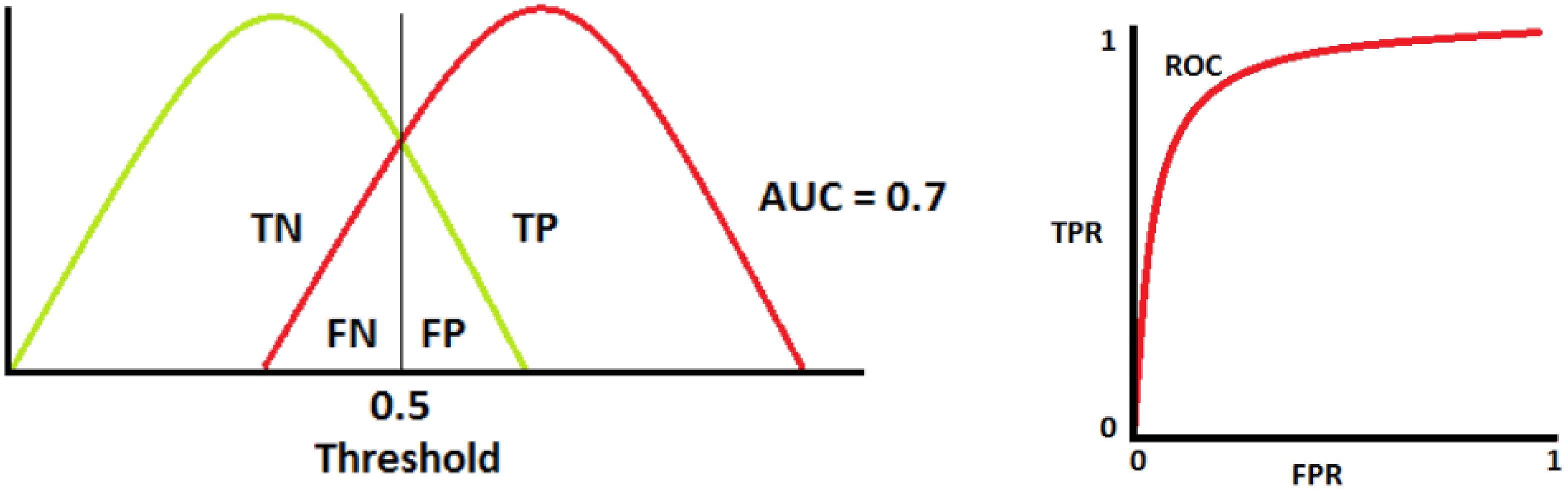
武漢肺炎高危險群自評表

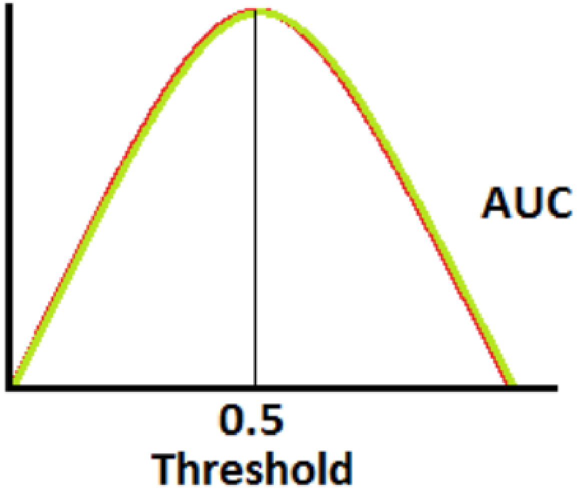
下呼吸道症狀	
▸ 咳嗽	加 8 分
▸ 呼吸急促	加 17 分
疾 病	
▸ 糖尿病	加 10 分
▸ 高血壓	加 10 分
▸ 冠心症	加 5 分
上呼吸道症狀	
▸ 鼻塞	減 10 分
▸ 喉嚨痛	減 15 分
全身性症狀	
▸ 頭痛	減 10 分
▸ 肌肉痠痛	減 10 分
▸ 疲勞倦怠	減 5 分

各症狀分數加總≥ 40 分
建議接受武漢肺炎篩檢

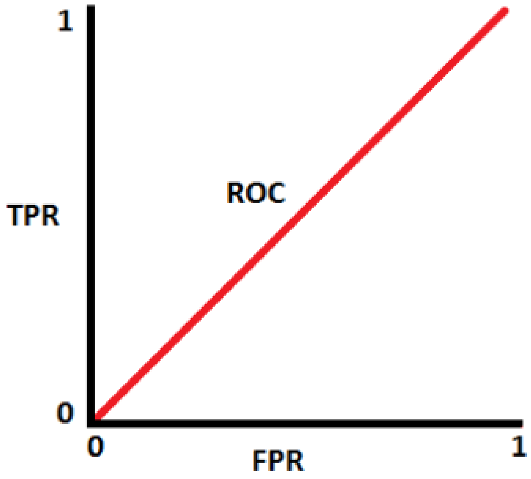
中央流行疫情指揮中心今晚緊急澄清，該自評表的險評估分數模式是作為區分新冠肺炎與流感之用，目的在輔助醫療院所篩檢，非自我篩檢工具。（中央流行疫

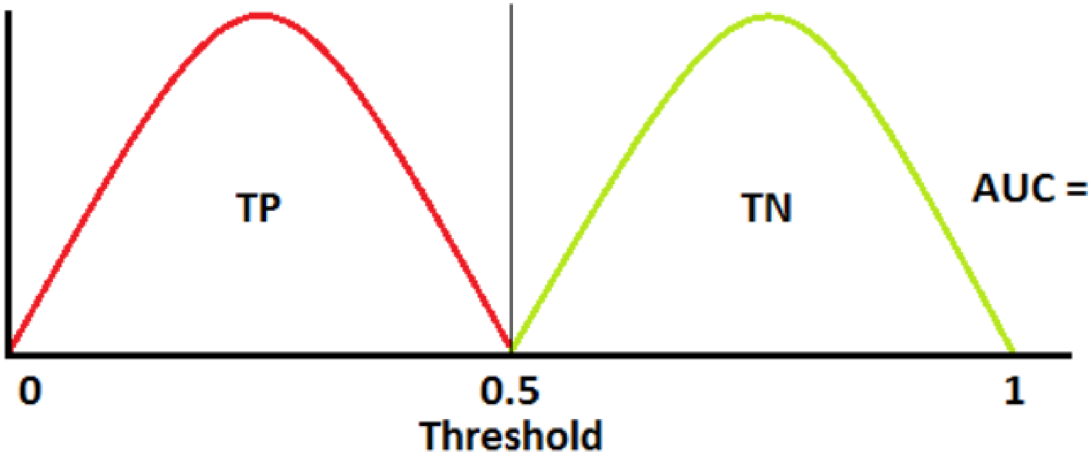

TN
TP
AUC = 1
0
0.5
1
Threshold
1
ROC
TPR
0
0
FPR
1


TN
TP
FN
FP
AUC = 0.7
0.5
Threshold
1
ROC
TPR
0
0
FPR
1

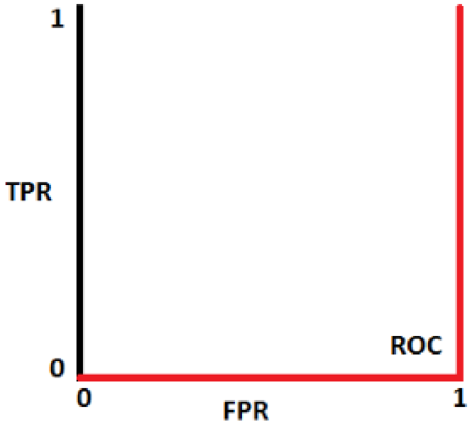

0.5
Threshold

AUC = 0.5

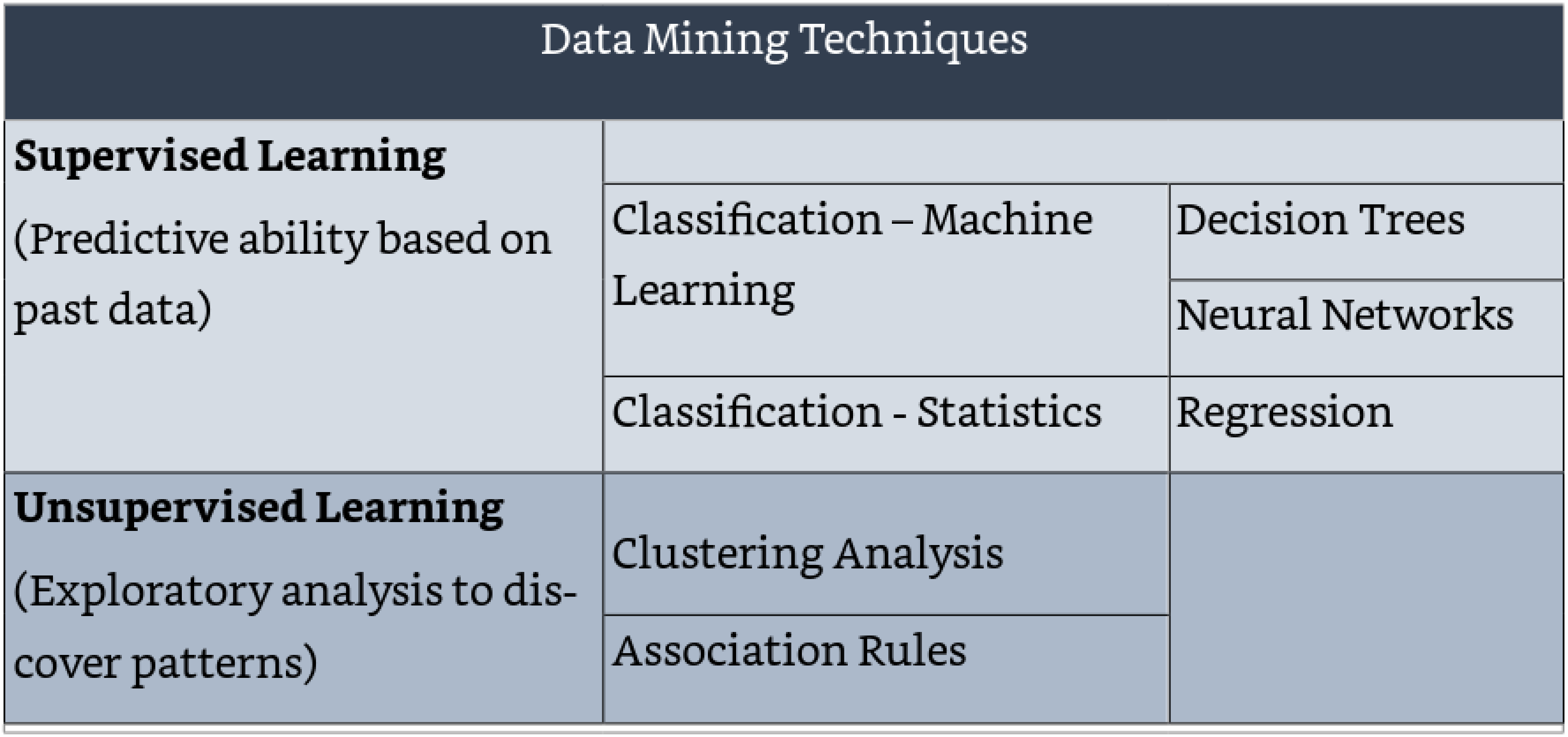

1
TPR
ROC
0
0
FPR
1


TP
TN
0
0.5
1
Threshold

AUC = 0


1
TPR
ROC
0
0
FPR
1

Data mining Techniques



Data Mining Techniques		
Supervised Learning (Predictive ability based on past data)		
	Classification – Machine Learning	Decision Trees
		Neural Networks
	Classification - Statistics	Regression
Unsupervised Learning (Exploratory analysis to discover patterns)	Clustering Analysis	
	Association Rules	

Types of Learning in ML

- Learning Problems
  1. Supervised Learning
  2. Unsupervised Learning
  3. Reinforcement Learning
- Hybrid Learning Problems
  4. Semi-Supervised Learning
  5. Self-Supervised Learning
  6. Multi-Instance Learning
- Statistical Inference
  7. Inductive Learning
  8. Deductive Inference
  9. Transductive Learning
- Learning Techniques
  10. Multi-Task Learning
  11. Active Learning
  12. Online Learning
  13. Transfer Learning
  14. Ensemble Learning
  15. Federated Learning

https://machinelearningmastery.com/types-of-learning-in-machine-learning/

1. Supervised Learning

- There are two main types of supervised learning problems: they are classification that involves predicting a class label and regression that involves predicting a numerical value.

Classification: Supervised learning problem that involves predicting a class label.

Regression: Supervised learning problem that involves predicting a numerical label.

- Both classification and regression problems may have one or more input variables and input variables may be any data type, such as numerical or categorical.

- An example of a classification problem would be the MNIST handwritten digits dataset where the inputs are images of handwritten digits (pixel data) and the output is a class label for what digit the image represents (numbers 0 to 9).
- An example of a regression problem would be the Boston house prices dataset where the inputs are variables that describe a neighborhood and the output is a house price in dollars.

- Some machine learning algorithms are described as "supervised" machine learning algorithms as they are designed for supervised machine learning problems. Popular examples include: decision trees, support vector machines, neural net and many more.

2. Unsupervised Learning

- Unsupervised learning describes a class of problems that involves using a model to describe or extract relationships in data.
- Compared to supervised learning, unsupervised learning operates upon only the input data without outputs or target variables. As such, unsupervised learning does not have a teacher correcting the model, as in the case of supervised learning.

- There are many types of unsupervised learning, although there are two main problems that are often encountered by a practitioner: they are clustering that involves finding groups in the data and density estimation that involves summarizing the distribution of data.
- Clustering: Unsupervised learning problem that involves finding groups in data.
- Density Estimation: Unsupervised learning problem that involves summarizing the distribution of data.

- An example of a clustering algorithm is k-Means where k refers to the number of clusters to discover in the data. An example of a density estimation algorithm is Kernel Density Estimation that involves using small groups of closely related data samples to estimate the distribution for new points in the problem space.

2. Unsupervised Learning(cont.)

- Clustering and density estimation may be performed to learn about the patterns in the data.

- Additional unsupervised methods may also be used, such as visualization that involves graphing or plotting data in different ways and projection methods that involves reducing the dimensionality of the data.

- Visualization: Unsupervised learning problem that involves creating plots of data.
- Projection: Unsupervised learning problem that involves creating lower-dimensional representations of data.
- An example of a visualization technique would be a scatter plot matrix that creates one scatter plot of each pair of variables in the dataset. An example of a projection method would be Principal Component Analysis that involves summarizing a dataset in terms of eigenvalues and eigenvectors, with linear dependencies removed.

3. Reinforcement Learning

- Reinforcement learning describes a class of problems where an agent operates in an environment and must learn to operate using feedback.
- Reinforcement learning is learning what to do — how to map situations to actions—so as to maximize a numerical reward signal. The learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them.
- Some machine learning algorithms do not just experience a fixed dataset. For example, reinforcement learning algorithms interact with an environment, so there is a feedback loop between the learning system and its experiences.
- An example of a reinforcement problem is playing a game where the agent has the goal of getting a high score and can make moves in the game and received feedback in terms of punishments or rewards.
- Impressive recent results include the use of reinforcement in Google’ s AlphaGo in out-performing the world’ s top professuional human Go player.
- Some popular examples of reinforcement learning algorithms include Q-learning, temporal-difference learning, and deep reinforcement learning.

Hybrid Learning Problems:
4. Semi-Supervised Learning

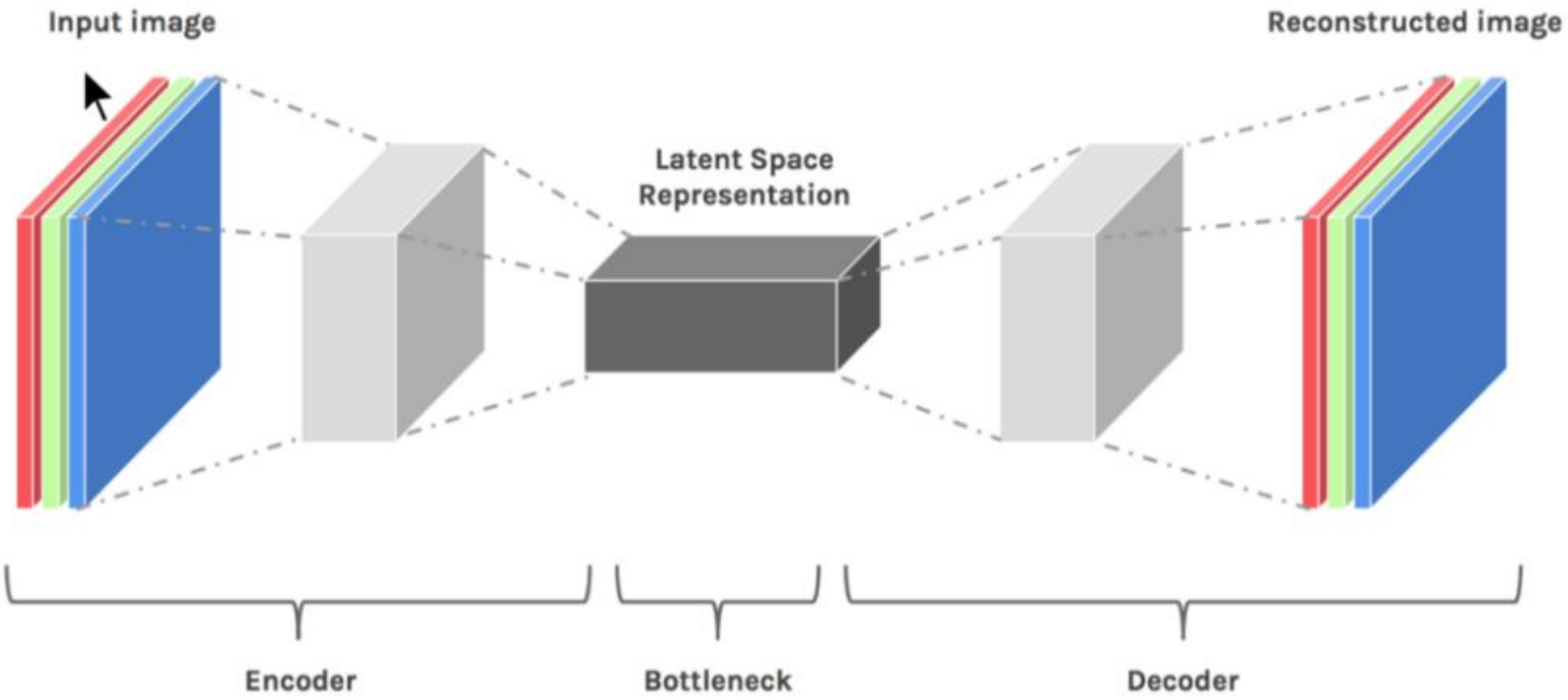
- Semi-supervised learning is supervised learning where the training data contains very few labeled examples and a large number of unlabeled examples. The goal of a semi-supervised learning model is to make effective use of all of the available data, not just the labelled data like in supervised learning.
- Making effective use of unlabelled data may require the use of or inspiration from unsupervised methods such as clustering and density estimation. Once groups or patterns are discovered, supervised methods or ideas from supervised learning may be used to label the unlabeled examples or apply labels to unlabeled representations later used for prediction.
- It is common for many real-world supervised learning problems to be examples of semi-supervised learning problems given the expense or computational cost for labeling examples. For example, classifying photographs requires a dataset of photographs that have already been labeled by human operators.
- Many problems from the fields of computer vision (image data), natural language processing (text data), and automatic speech recognition (audio data) fall into this category and cannot be easily addressed using standard supervised learning methods.

5. Self-Supervised Learning

- The self-supervised learning framework requires only unlabeled data in order to formulate a pretext learning task (委託學習任務) such as predicting context or image rotation, for which a target objective can be computed without supervision.
- A common example of self-supervised learning is computer vision where a corpus of unlabeled images is available and can be used to train a supervised model, such as making images grayscale and having a model predict a color representation (colorization) or removing blocks of the image and have a model predict the missing parts (inpainting).
- A general example of self-supervised learning algorithms are **autoencoder**s. These are a type of neural network that is used to create a compact or compressed representation of an input sample (code, latent vector, feature vector) . They achieve this via a model that has an encoder and a decoder element separated by a bottleneck that represents the internal compact representation of the input.
- SSL trained model is considered as a pretrained model, and applies linear evaluation approaches to evaluate and fine-tune pretrained models

- An autoencoder is a neural network that is trained to attempt to copy its input to its output. Internally, it has a hidden layer h that describes a code used to represent the input. Traditionally, autoencoders were used for dimensionality reduction or feature learning.

Autoencoder



source : Julien Despois

- Another example of self-supervised learning is generative adversarial networks, or GANs (生成對抗網路). These are generative models that are most commonly used for creating synthetic photographs using only a collection of unlabeled examples from the target domain.
- GAN models are trained indirectly via a separate discriminator model that classifies examples of photos from the domain as real or fake (generated), the result of which is fed back to update the GAN model and encourage it to generate more realistic photos on the next iteration.
- The generator network directly produces samples. Its adversary, the discriminator network, attempts to distinguish between samples drawn from the training data and samples drawn from the generator. The discriminator emits a probability value given by d(x; θ (d)), indicating the probability that x is a real training example rather than a fake sample drawn from the model.
- 生成對抗網路通過讓兩個神經網路相互博弈的方式進行學習。由一個生成網絡與一個判別網絡組成。生成網絡從潛在空間（latent space）中隨機取樣作為輸入，其輸出結果需要盡量模仿訓練集中的真實樣本。判別網絡的輸入則為真實樣本或生成網絡的輸出，其目的是將生成網絡的輸出從真實樣本中盡可能分辨出來。而生成網絡則要盡可能地欺騙判別網絡。兩個網絡相互對抗、不斷調整參數，最終目的是使判別網絡無法判斷生成網絡的輸出結果是否真實。

6. Multi-Instance Learning

- Multi-instance learning (多示例學習) is a supervised learning problem where individual examples are unlabeled; instead, bags or groups of samples are labeled.
- In multi-instance learning, an entire collection of examples is labeled as containing or not containing an example of a class, but the individual members of the collection are not labeled.
- Instances are in “bags” rather than sets because a given instance may be present one or more times, e.g. duplicates.
- Modeling involves using knowledge that one or some of the instances in a bag are associated with a target label, and to predict the label for new bags in the future given their composition of multiple unlabeled examples.
- In supervised multi-instance learning, a class label is associated with each bag, and the goal of learning is to determine how the class can be inferred from the instances that make up the bag.
- Simple methods, such as assigning class labels to individual instances and using standard supervised learning algorithms, often work as a good first step.
- 多示例學習的標記目標不是一個樣本，而是一個數據包（bag）。當一個bag的標記爲負時，這個bag裏面所有樣本的標記都是負的。當一個bag的標記爲正時，這個bag裏面至少有一個樣本的標記爲正。目標是學習得到一個分類器，使得對新輸入的樣本，可以給出它的正負標記。這樣的一類問題就是多示例問題。

Statistical Inference
7. Inductive Learning(歸納學習)

- Inductive learning involves using evidence to determine the outcome.
- 歸納推理(inductive reasoning)是由具體至廣泛的推理方式，
- Inductive reasoning refers to using specific cases to determine general outcomes, e.g. specific to general.
- Most machine learning models learn using a type of inductive inference (歸納推理) or inductive reasoning where general rules (the model) are learned from specific historical examples (the data).
- Fitting a machine learning model is a process of induction. The model is a generalization of the specific examples in the training dataset.
- A model or hypothesis is made about the problem using the training data, and it is believed to hold over new unseen data later when the model is used.
- 從多個個例歸納出普遍性，再演繹到個例，例如大陸法案判決方式，先對過往的判例歸納總結出法律條文，再應用到實際案例進行判決。但是從有限的實際樣本，企圖歸納出普遍真理，傾向形而上，往往會不由自主地成為教條。若從嚴謹實驗方法去運用歸納推理當然沒有問題，但是在現實生活，可能會胡亂運用歸納推理，造成錯誤推論 (ex.看見幾個美女都和有錢人在一起就斷定全世界所有美女都是貪錢的、看見幾個中東人是恐怖份子就斷定了所有中東人都是恐怖份子)
- 歸納推理中的一個經典方法是Bayes決策，通過求解$P(Y|X)=P(X|Y)P(Y)/P(X)$得到從樣本X到類別Y的概率分佈$P(Y|X)$，進而使用$P(Y|X)$預測測試樣本的類別。

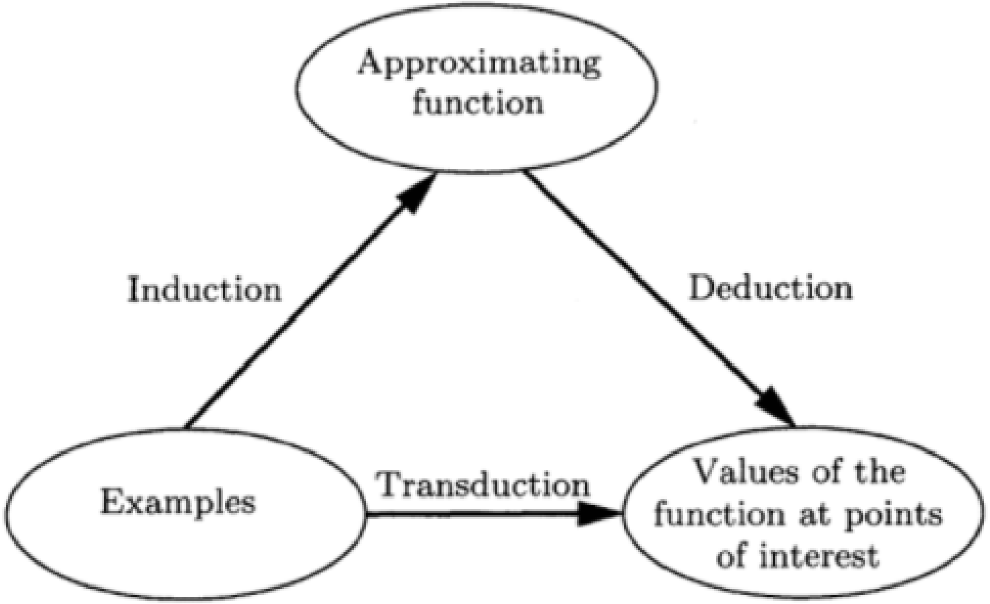
8. Deductive Learning (演繹學習)

- Deduction or deductive inference refers to using general rules to determine specific outcomes.
- Deductive reasoning是由廣泛至具體的推理方式，在推理的過程中，會由一系列較為廣泛的前提(premises)去嘗試推論出較具體的結論。
- Deduction is the reverse of induction. If induction is going from the specific to the general, deduction is going from the general to the specific.
- Deduction is a top-down type of reasoning that seeks for all premises to be met before determining the conclusion, whereas induction is a bottom-up type of reasoning that uses available data as evidence for an outcome.
- In the context of machine learning, once we use induction to fit a model on a training dataset, the model can be used to make predictions. The use of the model is a type of deduction or deductive inference.
- 北科學生英文不好; **有些**英文不好的人程式設計能力也不好; 所以北科學生的程式設計能力不好。這個例子的推理結論是不正確的因為不是所有英文不好的人程式設計能力也不好。
- 預測結果是否正確是非常取決於前提(premises)是否正確。但是這個前提的形成偏偏會受著許多前期資料偏誤所影響。

9. Transductive Learning (轉導學習)

- Transduction or transductive learning is used in the field of statistical learning theory to refer to predicting specific examples given specific examples from a domain.
- 統計學習中，轉導推理（Transductive Inference）是一種通過觀察特定的訓練樣本，進而預測特定的測試樣本的方法。
- It is different from induction that involves learning general rules from specific examples, e.g. specific to specific.
- Induction, deriving the function from the given data. Deduction, deriving the values of the given function for points of interest. Transduction, deriving the values of the unknown function for points of interest from the given data.
- Unlike induction, no generalization is required; instead, specific examples are used directly. This may, in fact, be a simpler problem than induction to solve.
- The model of estimating the value of a function at a given point of interest describes a new concept of inference: moving from the particular to the particular. We call this type of inference transductive inference. Note that this concept of inference appears when one would like to get the best result from a restricted amount of information.
- A classical example of a transductive algorithm is the k-Nearest Neighbors algorithm that does not model the training data, but instead uses it directly each time a prediction is required.
- 轉導推理能利用無標註的測試樣本的資訊發現聚簇，進而更有效地分類。而這正是隻使用訓練樣本推導模型的歸納推理所無法做到的。

- **Induction**: Learning a general model from specific examples. (specific cases to determine general)
- **Deduction**: Using a model to make predictions. (general rules to determine specific outcomes)
- **Transduction**: Using specific examples to make predictions. (specific to specific)



Relationship Between Induction, Deduction, and Transduction
Taken from The Nature of Statistical Learning Theory.

Types of Learning Techniques

- 10. Multi-Task Learning

Multi-task learning is a type of supervised learning that involves fitting a model on one dataset that addresses multiple related problems.

- 11. Active Learning

Active learning is a technique where the model is able to query a human user operator during the learning process in order to resolve ambiguity during the learning process.

- 12. Online Learning

Online learning involves using the data available and updating the model directly before a prediction is required or after the last observation was made.

Types of Learning Techniques

- 13. Transfer Learning

Transfer learning is a type of learning where a model is first trained on one task, then some or all of the model is used as the starting point for a related task.

- 14. Ensemble Learning

Ensemble learning is an approach where two or more modes are fit on the same data and the predictions from each model are combined.

- 15. Federated Learning

Federated learning (also known as collaborative learning) is a machine learning technique that trains an algorithm across multiple decentralized edge devices or servers holding local data samples, without exchanging them. This approach stands in contrast to traditional centralized machine learning techniques where all the local datasets are uploaded to one server, as well as to more classical decentralized approaches which often assume that local data samples are identically distributed.

Tools and Platforms for Data Mining

- MS Excel is a relatively simple and easy data mining tool. It can get quite versatile once Analyst Pack and some other add-on products are installed on it.
- scikit-learn provides machine learning techniques in Python, Prviding simple and efficient tools for data mining and data analysis, accessible to everybody, and reusable in various contexts. It is built on NumPy, SciPy, and matplotlib. (Open source, commercially usable - BSD license)
- Weka is an open-source GUI based tool that offers a large number of data ining algorithms.
- R is an extensive and extensible, versatile open-source statistical programming language with 600+ libraries and 120,000 functions. It is very popular with startup companies, and increasingly so in the large organizations.
- IBM's SPSS Modeler is an industry-leading data mining platform. It offers a powerful set of tools and algorithms for most popular data mining capabilities. It has colorful GUI format with drag-and-drop capabilities. It can accept data in multiple formats including reading Excel files directly.
- ERP systems include some data analytic capabilities, too. SAP has its Business Objects (BO) software. BO is considered one of the leading BI suites in the industry, and is often used by organizations that use SAP.
- Rapid Miner/ Orange/ KNIME / Sisense /SSDT (SQL Server Data Tools) /Apache Mahout /Oracle Data Mining/ Rattle/ DataMelt/IBM Cognos /SAS Data Mining/Teradata//Board/Dundas BI/MS Power BI ……..