

資料分析與學習基石

(Fundamental of Data Analytics and Learning)

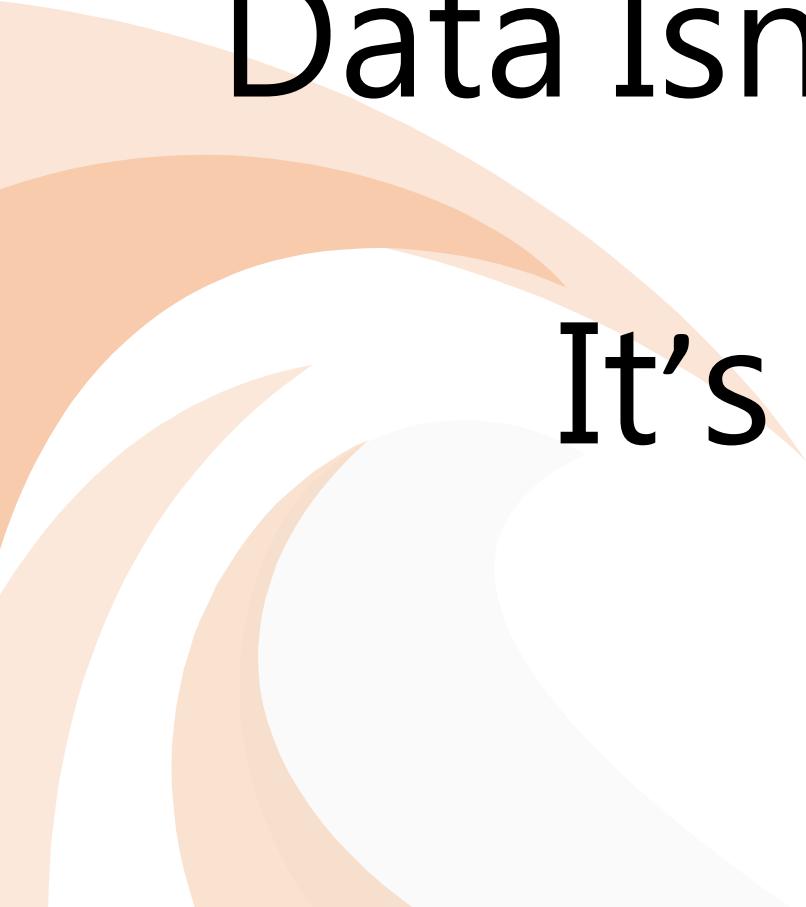
-- Introduction on Data Analytics & Mining

Hung-Yu Kao (高宏宇)
Intelligent Knowledge Management Lab



Master Program of Artificial Intelligence
Institute of Medical Informatics,
Dept. of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan



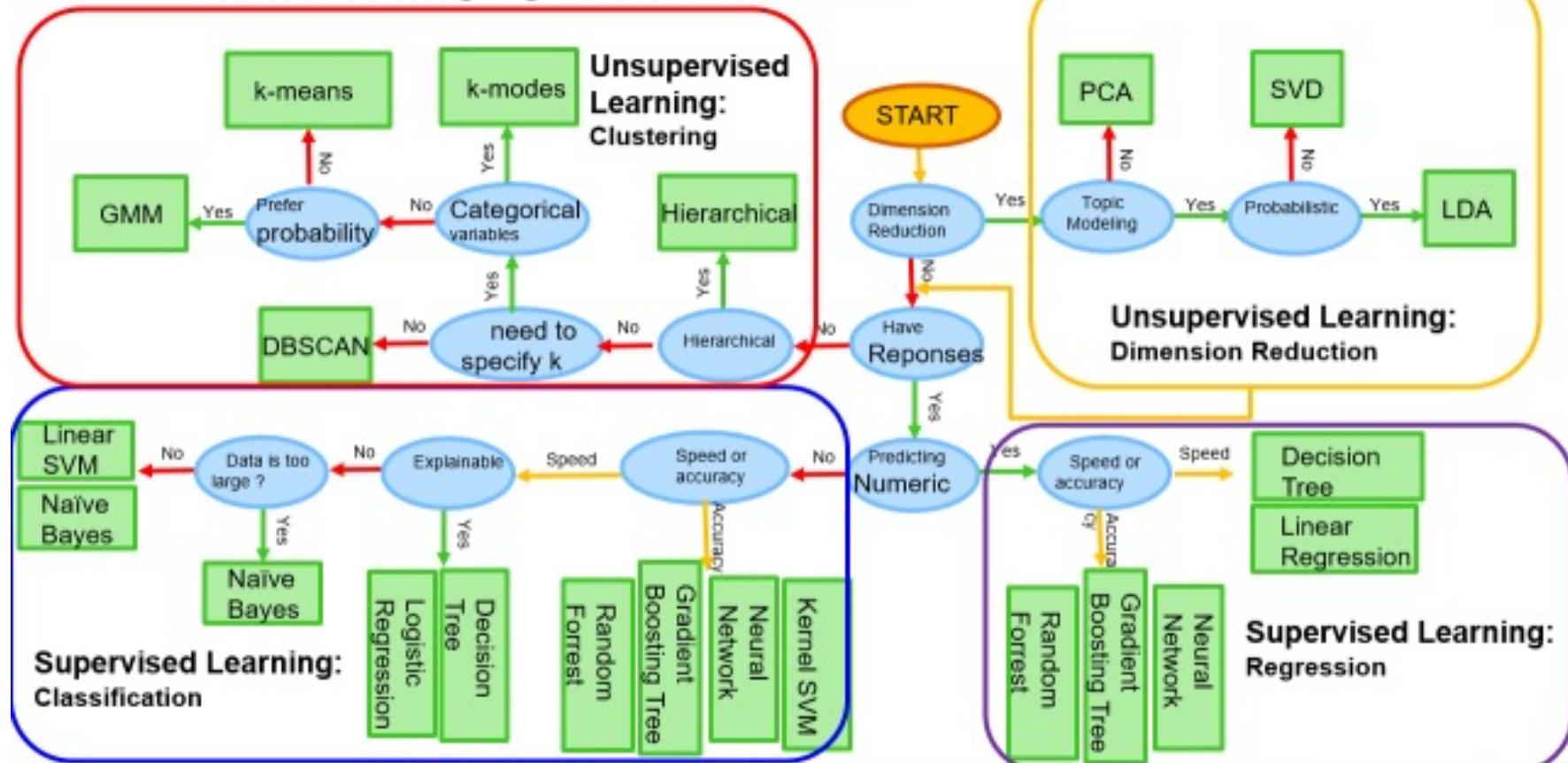


Data Isn't Numbers, It's People.

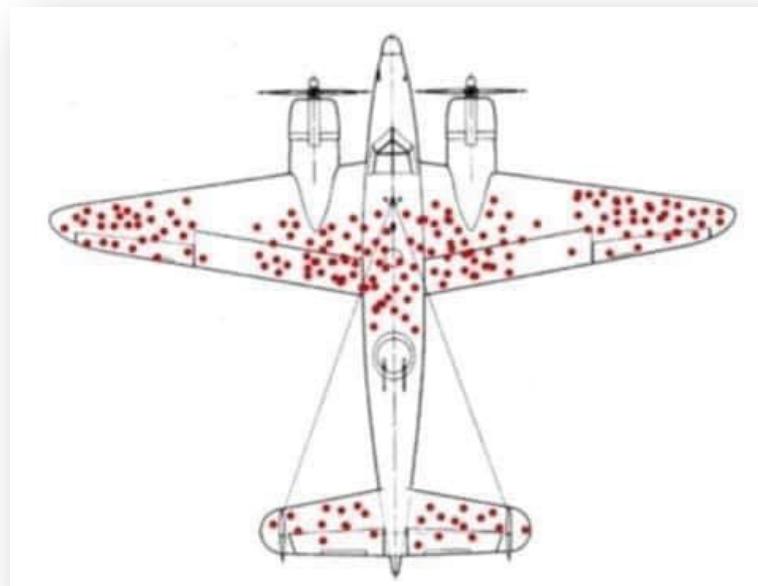
Riley Newman, Airbnb

Top-10 Machine Learning Algorithms for Data Scientist

Machine Learning Algorithms Cheat-sheet



An example



What kind of Data?

Which is useful in your Data?

Which is useful in your Data under some problem definition?

Which is useful in your Data under some **problem definition** and some **evaluation criteria**?

e.g., <https://www.kaggle.com/datasets>

Data-enriched Applications



- Time to take
- Cities to visit?
- Trace pattern?
- Travel purpose?
- Traffic suggestion?
- Congestion report?
- Pokemon hunting path?

Data informed decision making

- Personalization Engine
 - Adobe
 - Salesforce Evergage
 - Dynamic Yield (McD)
 - 得來速menu
 - Sephora推薦系統

The image consists of three vertically stacked screenshots demonstrating data-informed decision making.

Top Screenshot: A photograph of a McDonald's menu board. It displays various items with their prices and calorie counts. The board includes sections for 'Iced Caramel Macchiato', 'Smoothie', 'Iced Latte', and 'Minute Maid Premium Orange Juice'. Below these are promotional offers: 'Add something delicious' featuring a Minute Maid Orange Juice (2.59 | 200 Cal), a Bacon Egg & Cheese Biscuit (3.49 | 450 Cal), and a Sausage Egg McMuffin Meal (5.19 | 630-800 Cal).

Middle Screenshot: A screenshot of a Sephora product recommendation interface. At the top, it says 'Let us help. Have you tried:'. Below are two bullet points:

- Double-checking the spelling of your search term?
- Searching for a similar word or entering a broader term?

Underneath, there is a section titled 'You May Also Like' which shows four products:

- LANEIGE Lip Sleeping Mask** (NEW): Limited edition, \$22.00, ★★★★ 11.2K reviews.
- OLAPLEX No. 3 Hair Perfector** (NEW): Limited edition, \$56.00, ★★★★ 2.9K reviews.
- Sephora Favorites Give Me Some Shine Balm and Gloss Lip Set** (NEW): Limited edition, \$29.00, ★★★★★ 296 reviews.
- OLAPLEX No. 4 Bond Maintenance™ Shampoo**: \$28.00, ★★★★★ 1K reviews.

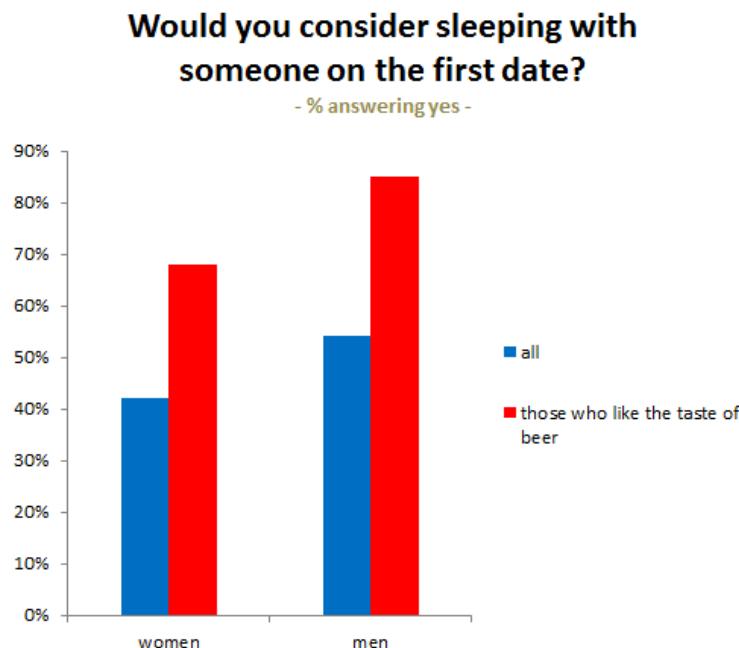
Bottom Screenshot: A screenshot of a Sephora 'Recommended For You' interface. It shows four recommended products:

- exclusive Peace Out Acne Healing Dots**: \$19.00, ★★★★ 2.6K reviews.
- exclusive KVD Vegan Beauty Dagger Tattoo Liner Waterproof Liquid Eyeliner**: \$24.00 | \$12.00 off.
- Too Faced Better Than Sex Mascara**: \$25.00, ★★★★ 16.2K reviews.
- limited edition LANEIGE Lip Glow Balm**: \$17.00, ★★★★★ 2.5K reviews.

Data → Value

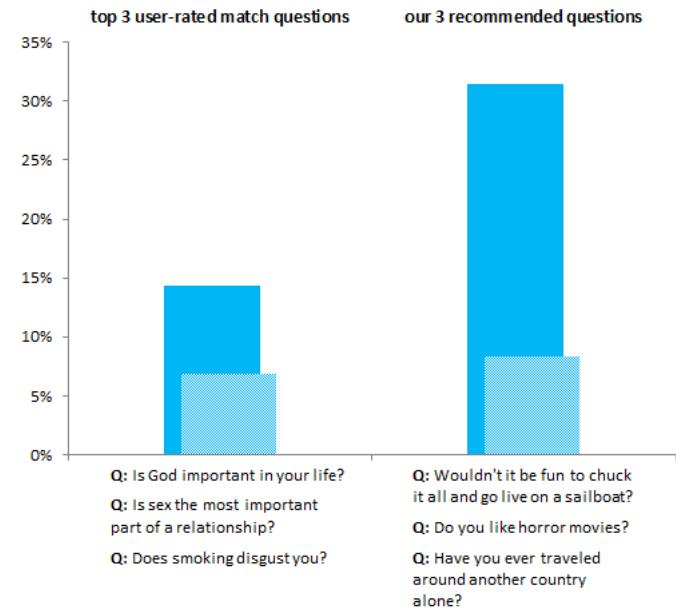


<https://theblog.okcupid.com/the-best-questions-for-a-first-date-dba6adaa9df2>



predictive question	implied odds of first-date sex
Q: In a certain light, wouldn't nuclear war be exciting?	'yes'⇒83%
Q: Assuming you were in the position to do so, would you launch nuclear weapons under any circumstances?	'yes'⇒82%
Q: Could you imagine yourself killing someone?	'yes'⇒82%

■ % of long-term couples who agree on all three questions
■ % agreement expected from pure chance



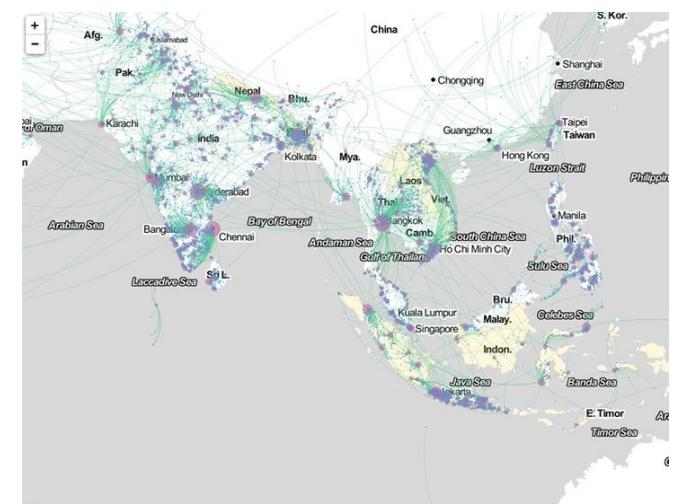
Data → Value

facebook

<https://www.facebook.com/notes/facebook-data-science/coordinated-migration/10151930946453859>

Hometown, current location → migration pattern

Destination city	Country	Number of coordinated migration: n(h)	Urbanization growth between 2000 and 2012 (%)	Hometown countries of the coordinated migrations
Lagos	Nigeria	566	18.6	Nigeria (96%)
Istanbul	Turkey	387	11.7	Turkey (84%), Macedonia (4%), Bulgaria (3%)
Bogota	Colombia	370	4.8	Colombia (98%)
Bangkok	Thailand	322	10.7	Thailand (90%)
Accra	Ghana	315	19.5	Ghana (97%)
Hyderabad	India	307	14.4	India (98%)
Kampala	Uganda	280	32.4	Uganda (93%), Democratic Republic of the Congo (3%)
Lima	Peru	279	6.2	Peru (97%)
Chennai	India	278	14.4	India (98%)
London	Great Britain	270	1.4	Great Britain (94%)



Data empowered Intelligence

HY

請跟上資料學習課程學生介紹你自己



嗨大家好，我是 ChatGPT！我是一個基於人工智慧技術的大型語言模型，由 OpenAI 團隊所訓練。我的主要任務是透過學習大量的文本資料，讓我能夠產生自然語言的回應，並且不斷透過學習來提升我的回應品質與準確性。

作為一個資料學習的學生，你們可能已經了解了機器學習的基本概念和技術，而我所使用的技術正是機器學習中的一個分支 - 自然語言處理。通過學習大量的文本資料，我能夠透過模式識別和統計學的方法來進行文本的自動生成、翻譯、分類、摘要等工作。

當然，我也還有許多可以改進的地方，例如更好的理解上下文、更快速的回應速度等等。但是我會不斷學習和進步，希望能夠在各種語言應用中，為人類帶來更多便利和幫助。

7 leading machine learning use cases (from AWS)

- What makes a good machine learning use case?
 - Solves a **real problem** for your business
 - important enough to get attention, support, and adoption
 - Leverages sources of **untapped data**
 - Increases performance, reduces costs, and/or improves your end-customer experiences
 - Includes technical experts to conduct feasibility assessments and **domain experts** to ensure the solution will be used
 - Can be completed in 6–10 months

<https://www.scribd.com/document/500818131/AWS-7-Leading-Machine-Learning-Use-Cases-eBook-EN>

https://theexperienceofwork.economist.com/pdf/Citrix_The_Experience_of_Work_BriefingPaper.pdf



7 leading machine learning use cases (from AWS)

Use Case 1

Improve employee productivity by quickly and easily finding accurate information

-- NLP, intelligent search, ...

Use Case 2

Make faster decisions by automatically extracting and analyzing data from documents

-- classification, intelligent document processing, ...

-- Thomson Reuters Center



7 leading machine learning use cases (from AWS)

Use Case 3

Add intelligence to your contact center to improve service and reduce cost

Use Case 4

Make personalized recommendations to increase customer engagement

Use Case 5

Analyze media assets to increase value and create new insights



7 leading machine learning use cases (from AWS)

Use Case 6

Forecast key demand metrics faster and more accurately to meet customer demand and reduce waste

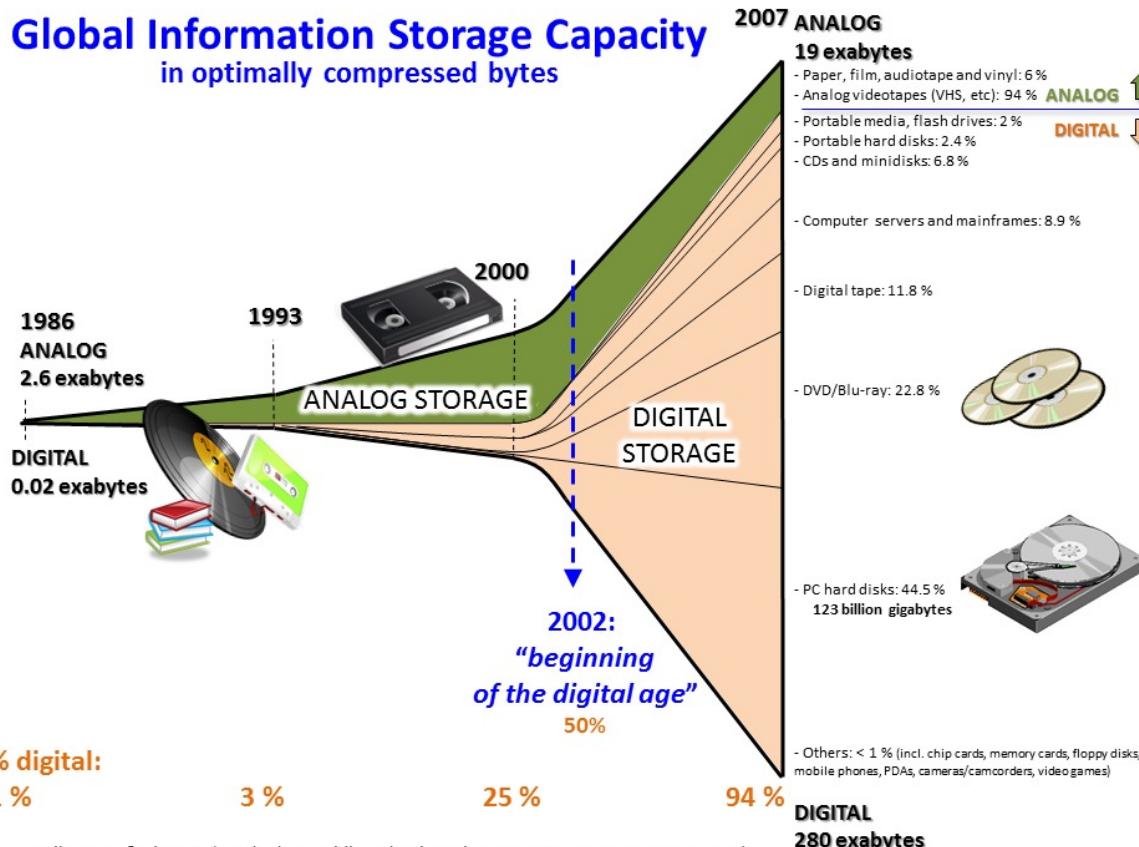
Use Case 7

Make it easy to identify potential fraudulent online activities

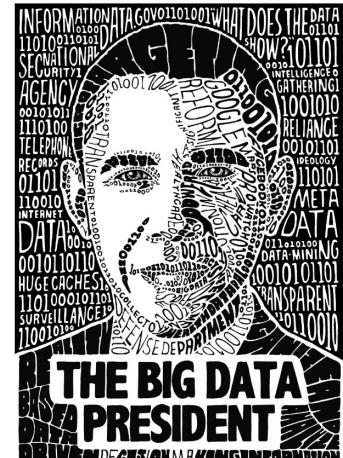
We are data rich

Driving force – Digital Storage

Global Information Storage Capacity
in optimally compressed bytes

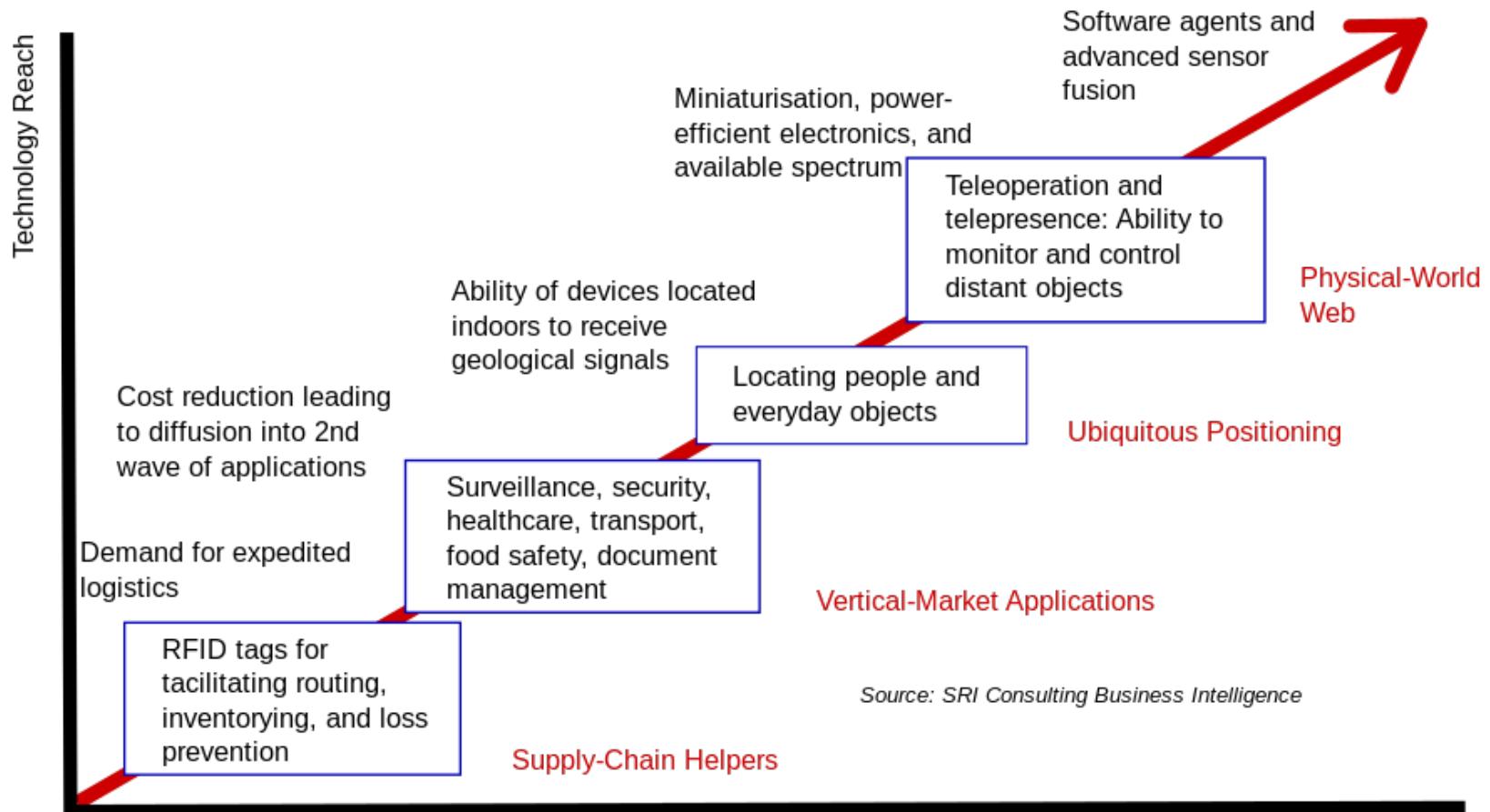


Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60–65. <http://www.martinhilbert.net/WorldInfoCapacity.html>



(Sarah A. King for The Washington Post)

More data in the coming years Everything is on-line



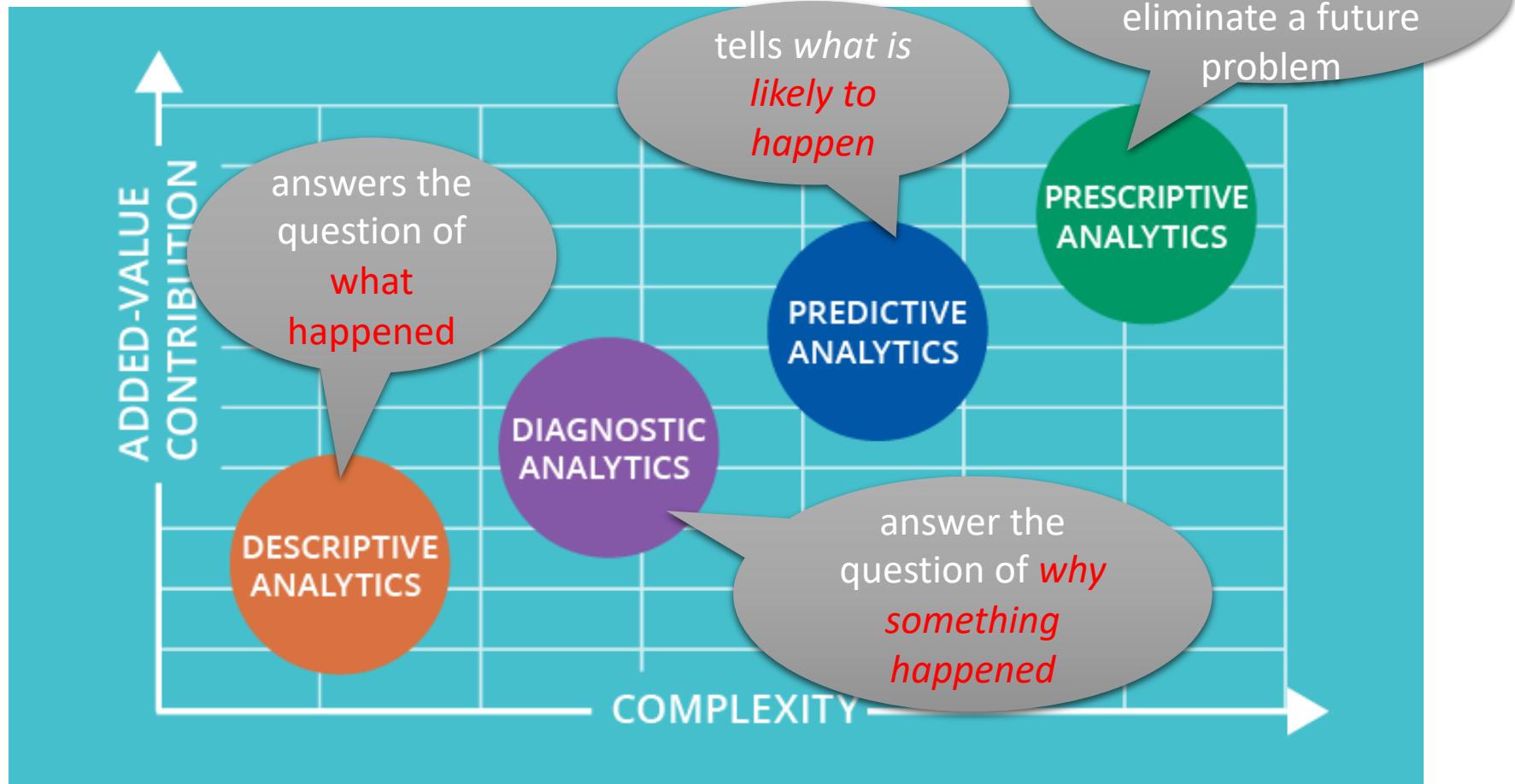
Data Analytics (DA)

- Data analytics is the process of examining data sets in order to **draw conclusions** about the **information** they contain.
 - Refer to different applications: **Business Intelligence (BI)**, **Online Analytical Process (OLAP)**
- **Boosting business performance**
 - Help business increase revenues
 - Improve operational efficiency
 - Optimize marketing campaigns and customer service efforts
 - Response more quickly to emerging market trend
 - gain a competitive edge over rivals

More Advanced types of Data Analytics

- **Data Mining**
 - involves sorting through large data sets to identify trends, patterns and relationships
- **Predictive Analytics**
 - seeks to predict customer behavior, equipment failures and other future events
- **Machine Learning**
 - an artificial intelligence technique that uses automated algorithms to churn through data sets more quickly than data scientists can do via conventional analytical modeling.
- **Big data analytics** applies data mining, predictive analytics and machine learning tools to sets of big data that often contain unstructured and semi-structured data.
 - **Text mining** provides a means of analyzing documents, emails and other text-based content.

Types of data analytics

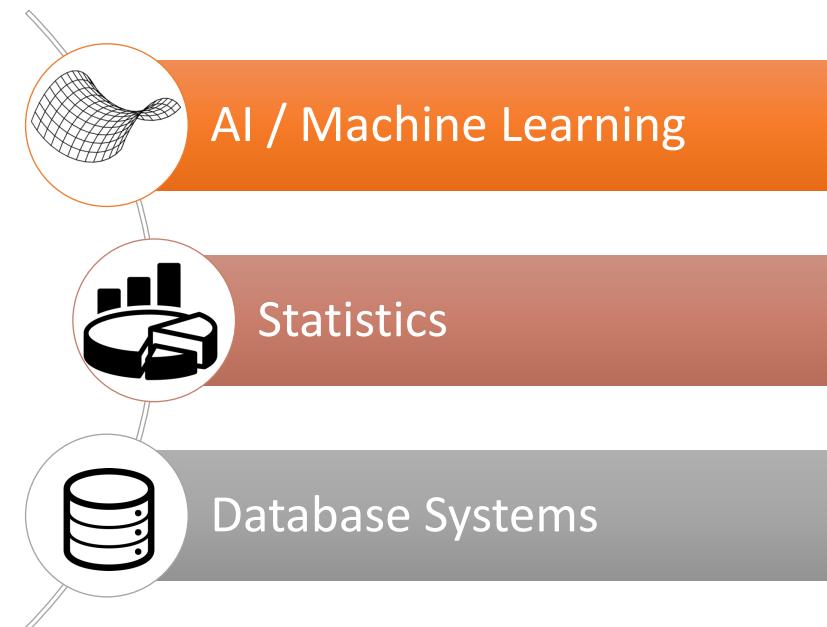


Data Mining

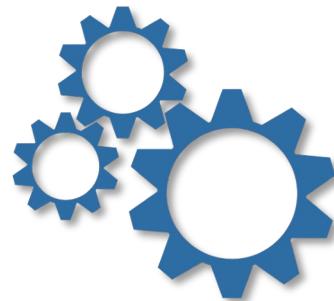
- Interdisciplinary subfield of Computer Science & Statistics
- the computational process of discovering patterns in large data sets involving of artificial intelligence, machine learning, statistics, and database systems.
- the automatic or semi-automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining).
- Data mining: Knowledge discovery in databases

Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- **Traditional Techniques** may be unsuitable due to
 - Enormity of data
 - High dimensionality of data
 - Heterogeneous, distributed nature of data



Artificial Intelligence



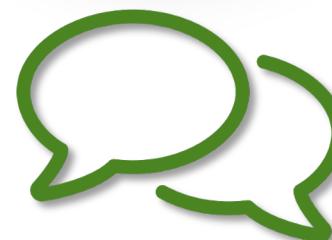
Reasoning



Knowledge



Learning



Communication



Planning

Reason & Planning

1. 農夫要帶著狗、雞、包心菜過河。
2. 小船不夠大，因此農夫每次只能帶一樣東西過河。
3. 當農夫在的時候，狗、雞、包心菜都不會有事情。
4. 當農夫不在時，狗會吃雞，雞會吃包心菜。

問題：農夫要如何將三樣東西平安的帶過河呢？

解答：

農夫先帶雞過河，再回頭帶包心菜過河，留下包心菜並將雞帶回，留下雞將狗帶過河，再回頭帶雞過河。



Machine Learning

- "*Gives computers the ability to learn without being explicitly programmed*" -- Arthur Samuel, 1959.



Learning from experience

Teaching,
Reading/Observati
on,
Memorizing/Under
standing



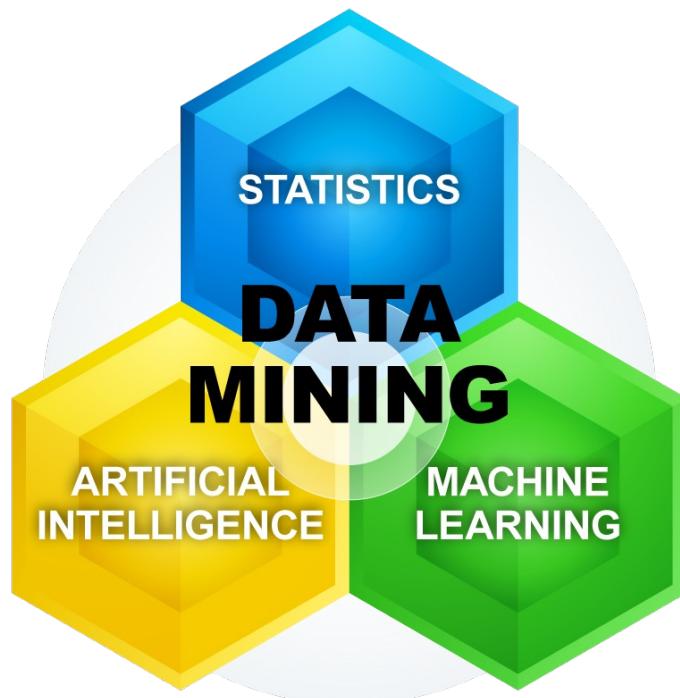
Learning from ~~experience~~

Data

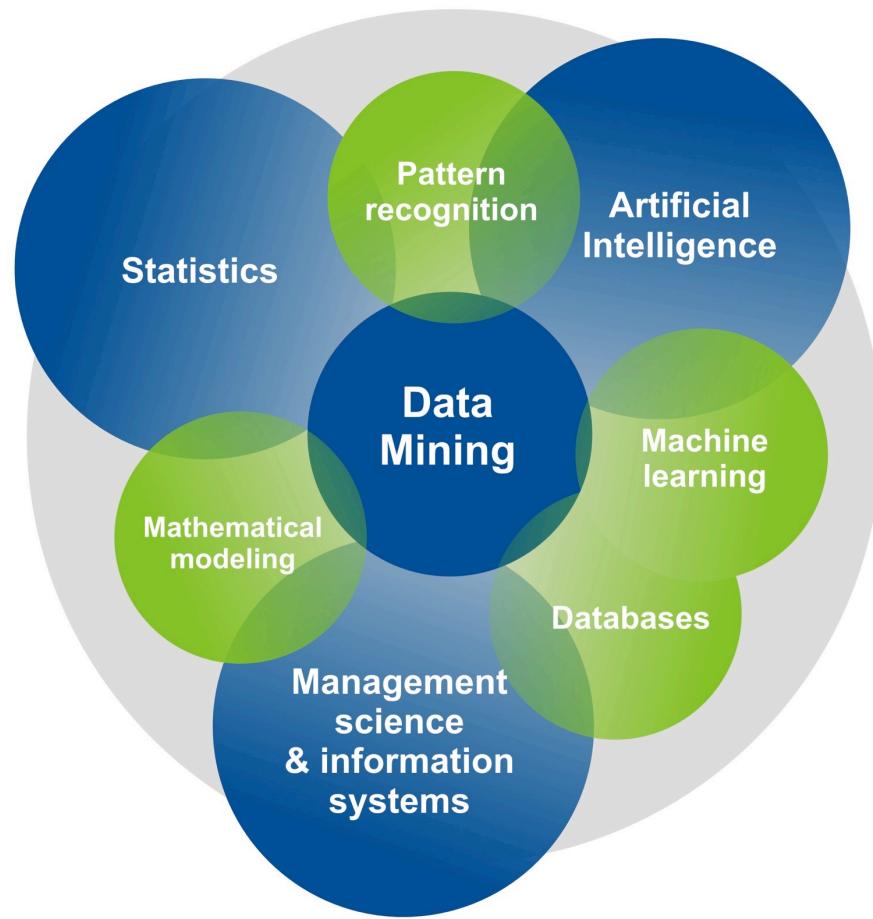


Designed instructions

What's their difference?

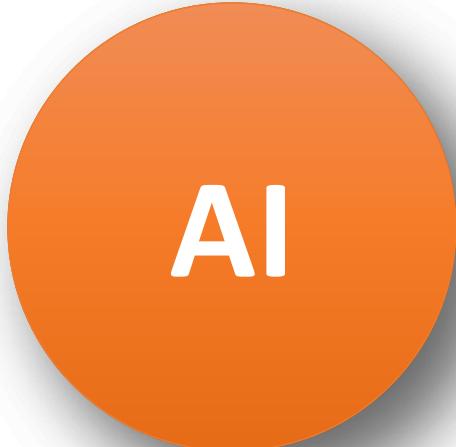


from sas.com



from <http://frontender.com/blog/enablers/data-mining/>

What's their difference?

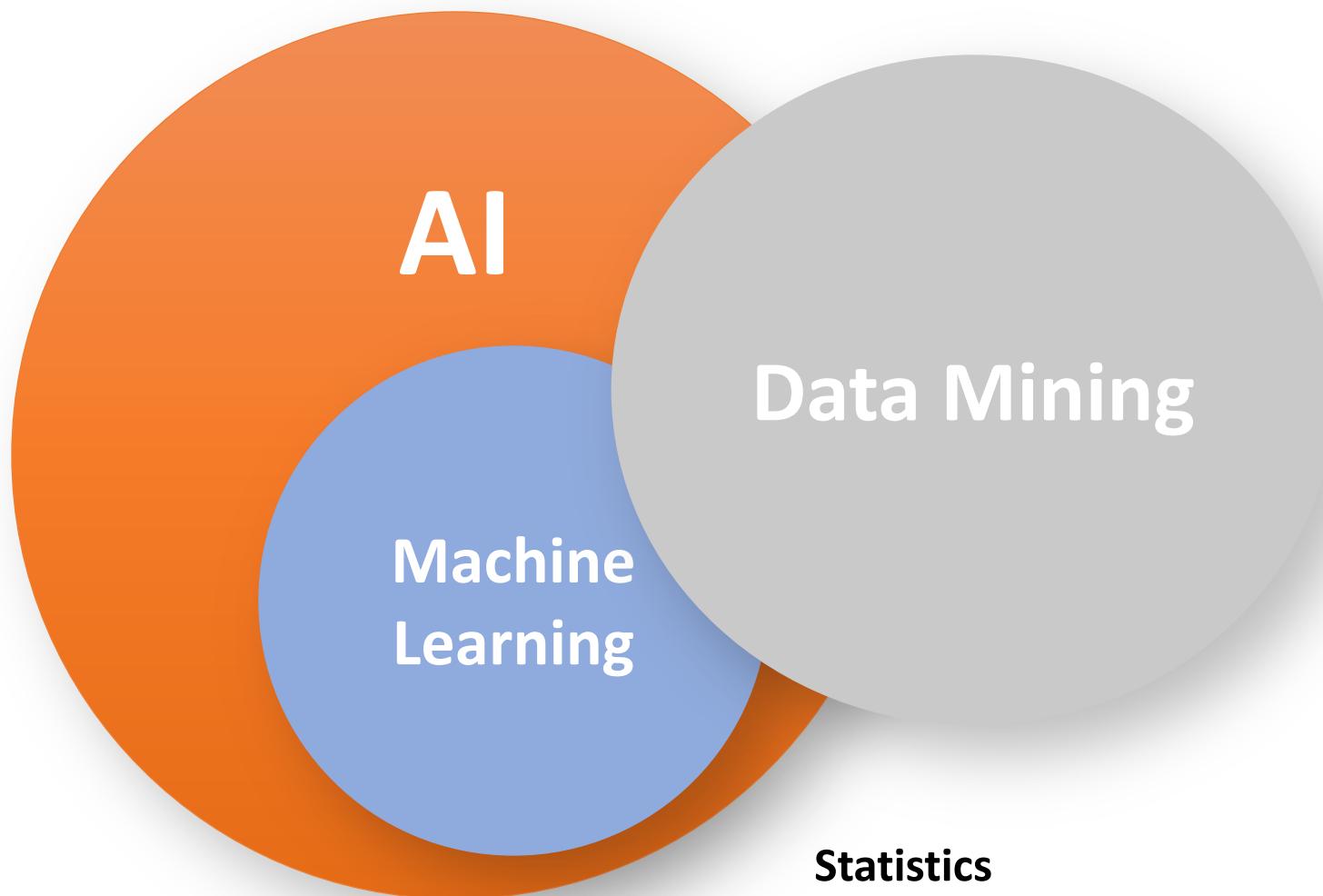


AI



Statistics

What's their difference?



Machine learning

- Machine learning develops algorithms for making **predictions** from **data**
- Machine learning is no Voodoo
 - The word “predictions” can be misleading



Machine Learning

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T , as measured by P , improves with experience E .

-- Tom M. Mitchell, 1997



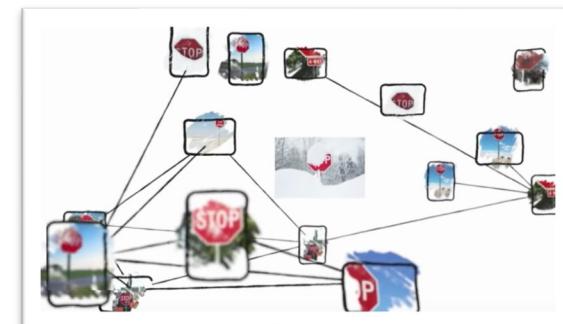
Machine Learning

- Model + Parameter + Learner

Model: makes the predictions and identifications



Parameter: Signals and factors that models need to make its decision



Learner: A system that adjusts the parameters by looking at differences between predictions and the actual outcomes

Machine Learning

- Closely related to (overlaps with) Computational Statistics, Probability, Optimization, and Data Mining

- Type of Tasks

- Supervised Learning

- 監督式學習

- Classification 分類

- Unsupervised Learning

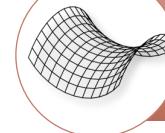
- 非監督式學習, 分群

- Reinforcement Learning

- 強化學習



Computational Statistics

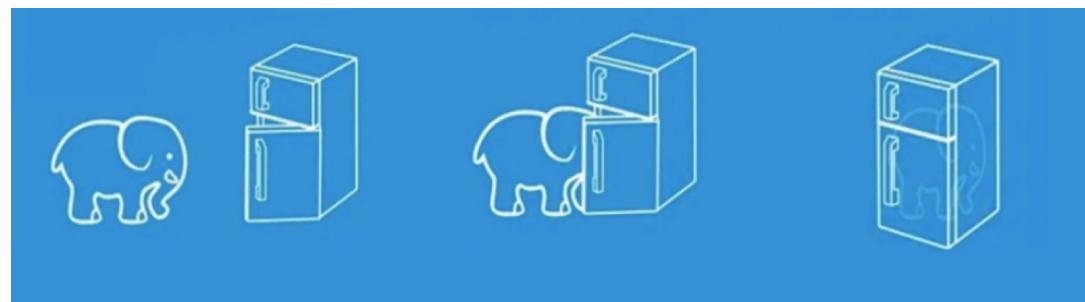


Optimization



Data mining

Memorize / Learn / Understand



Data Mining/Analytics Potential Applications

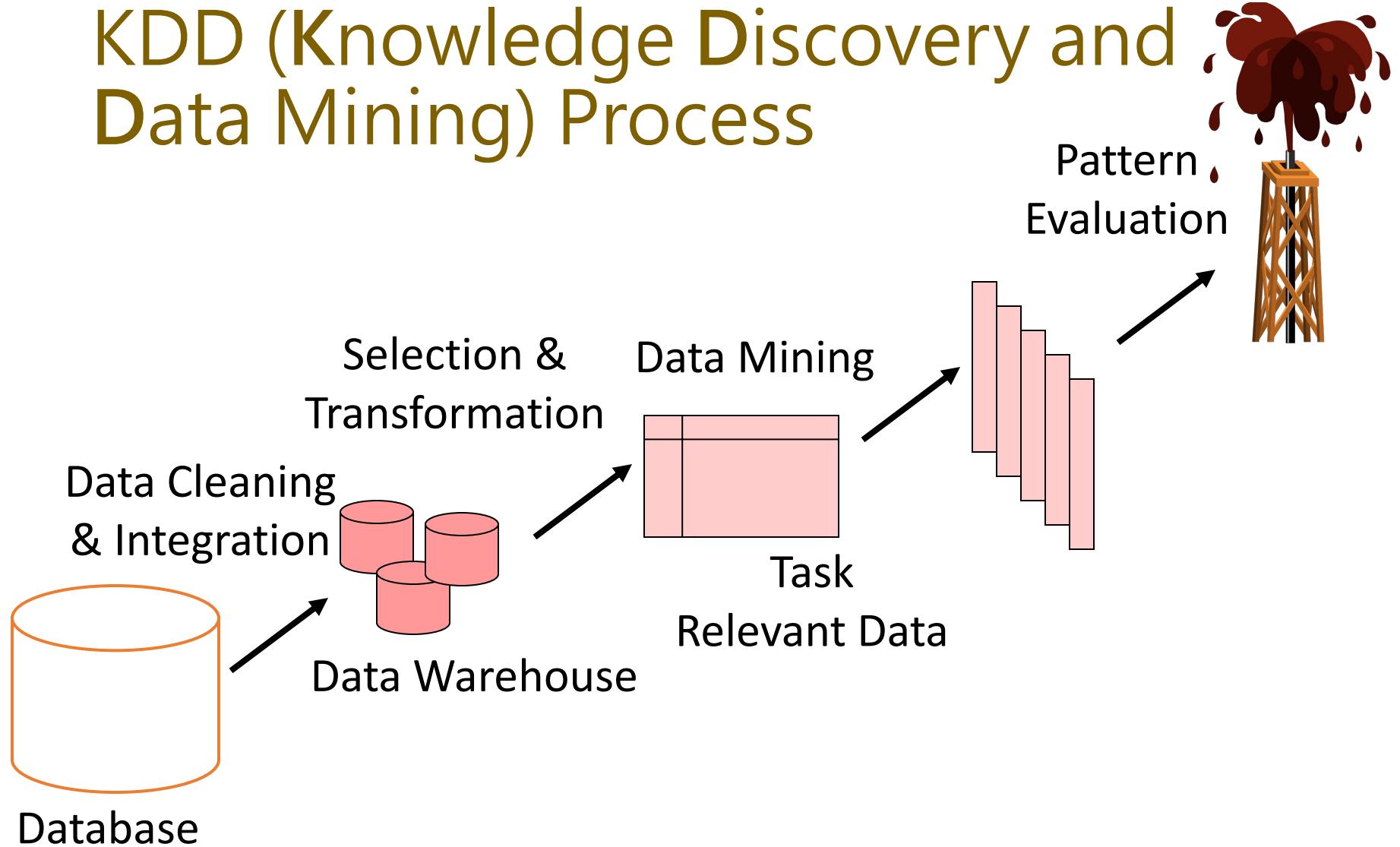
- Data analysis and decision support
 - Market analysis and management
 - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
 - Risk analysis and management
 - Forecasting, customer retention, competitive analysis
 - Fraud detection and detection of unusual patterns (outliers)
 - “Whoscall”

Knowledge Discovery from Databases

- Nontrivial process of extraction of
 - valid (with some degree of certainty)
 - novel (implicit, previously unknown)
 - potential useful
 - ultimately understandable
 - patterns from large collection of data
- Pattern
 - expression in languages describing subset of data
 - model (structure) applicable to subset of data



KDD (Knowledge Discovery and Data Mining) Process



Classification of Data Mining Techniques

- (input) What kinds of databases to work on
- (output) What kind of knowledge to be mined
- (process) What kind of techniques to be utilized



Databases to Work on

- Relational
- Transactional
- Spatial
- Time series data
- Multimedia
- Unstructured text
- Graph

Knowledge to Be Mined

- Association rules
 - $\text{Buy(bread)} \wedge \text{Buy(milk)} \Rightarrow \text{Buy(butter)}$
 - $\text{Age(20\text{--}29)} \wedge \text{Income(20\text{--}30k)} \Rightarrow \text{Buy(CD player)}$
- Classification
- Clustering
- Time series data analysis
- Semantics

Data Mining Tasks

- Prediction Methods
 - Use some variables to **predict unknown** or future values of other variables.
- Description Methods
 - Find human-interpretable patterns that **describe** the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

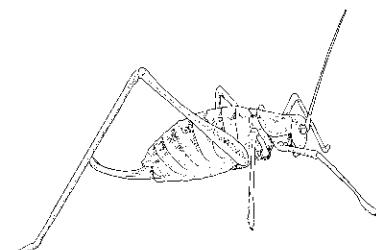


Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive] detect outliers

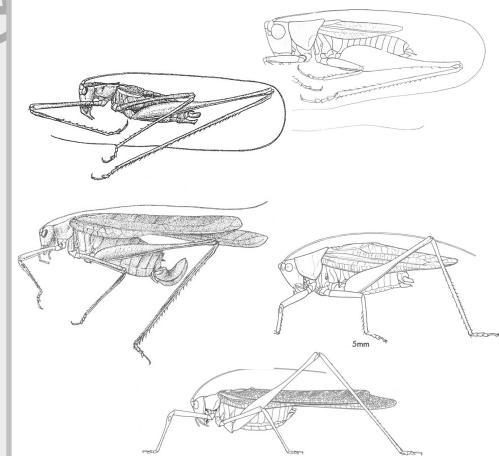
Classification example

- Given a collection of annotated data. In this case 5 instances **Katydid**s and five of **Grasshoppers**, decide what type of insect the unlabeled example is.

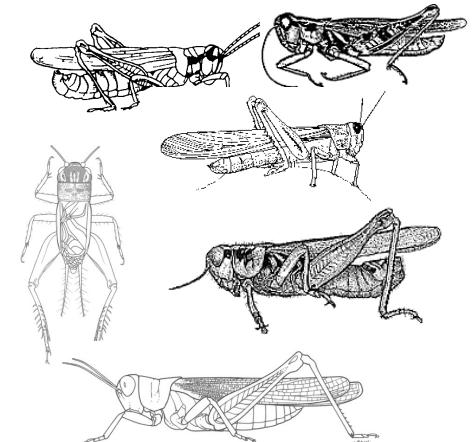


Katydid or Grasshopper?

Katydids



Grasshoppers

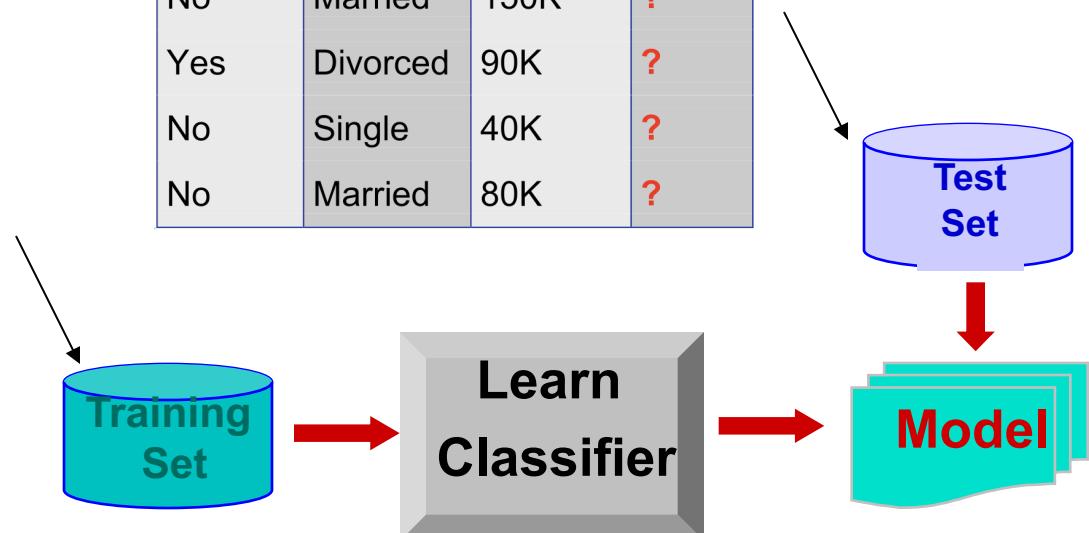


Classification Example

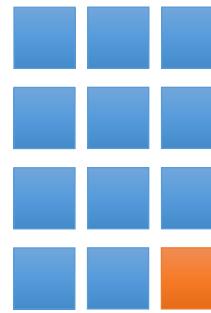
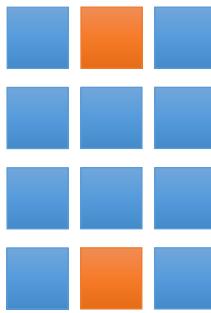
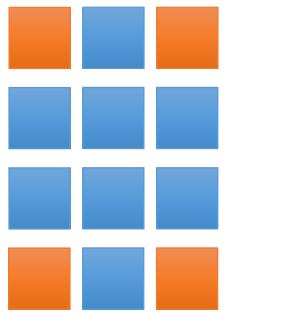
categorical
categorical
continuous
class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

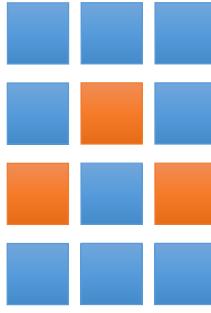
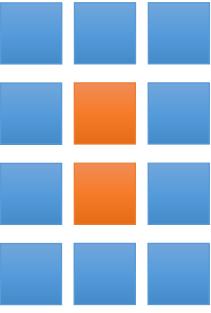
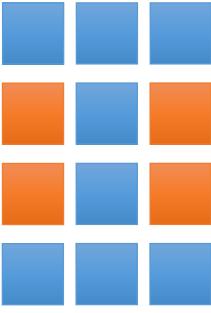
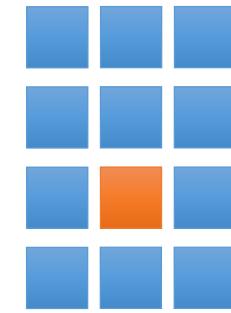
Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification Example



Class A

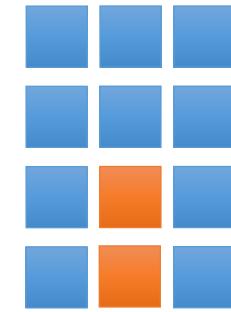


Class B

都是一個橘點->上面的會分到A
把橘點當成不同的意義會有不同的解釋（分類）
做出的AI要有可解釋性

kNN 新的直近來，就去看他跟哪一組比較像就去那一組

How about this?

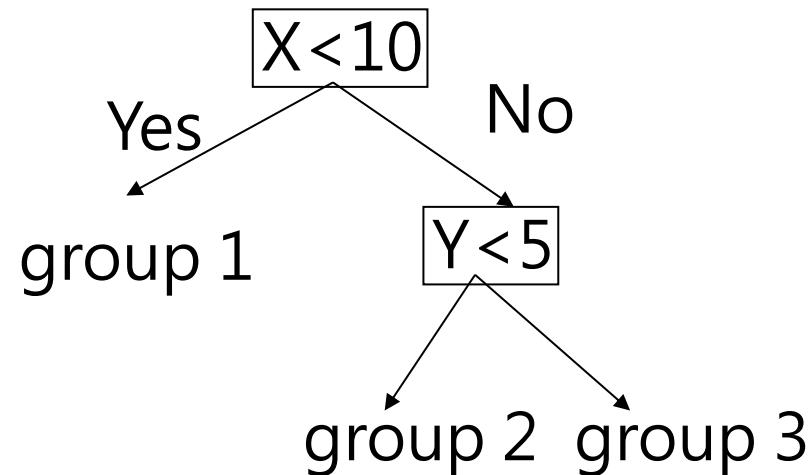


How about this?



Classification

- Supervised classification
- Organizes data into given classes based on attribute values
- Machine learning



Classification: Definition

- Given a collection of records (*training set*)
 - Each record contains a set of *attributes*, one of the attributes is the *class*. class : Yes, NO ; 多類別
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
 - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

交叉驗證，把一部分資料的結果遮起來，train完後測試準不準

Start from this X

overfitting 太完整解釋現有的東西可能造成預測失誤
可能現在的資料是outlyar, 可能資料的一小部分

訓練資料
(已發生)

上帝 $f(X) = Y$ 不知道 f 是什麼

機器學習：
給定 X, Y , 利用計算,
找到— g , minimize $|g(X) - f(X)|$

測試資料
(新資料)

Y

Y
N
Y
Y
Y
N
N
N
N
Y
Y
Y
Y
N

? Y 0.7
? N 0.2
? N 0.3
? Y 0.6
? Y 0.9



Classification: Application 1

- Direct Marketing
 - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
 - Approach:
 - Use the data for a similar product introduced before.
 - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
 - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
 - Type of business, where they stay, how much they earn, etc.
 - Use this information as input attributes to learn a classifier model.

From [Berry & Linoff] Data Mining Techniques, 1997



Classification: Application 2

- Fraud Detection
 - Goal: Predict **fraudulent** cases in credit card transactions.
 - Approach:
 - Use credit card transactions and the information on its account-holder as attributes.
 - When does a customer buy, what does he buy, how often he pays on time, etc
 - Label past transactions as fraud or fair transactions. This forms the class attribute.
 - Learn a model for the class of the transactions.
 - Use this model to detect fraud by observing credit card transactions on an account.

ex: whosecall 行為分析

困難的不是分析行為，最困難的是取得資料

公司半開放資料競賽，但不想洩漏資料，用一個mapping轉換資料要維持資料的特性與分佈，較困難

就算只公布pattern、去識別化，利用現有的資訊可能也能猜某人的資訊
ex: 蔡英文



Classification: Application 3

kkbox 想抓到覺得不錯，但有覺得有點貴的客戶，給一個discount他就會續訂
要預測客戶的行為

- Customer Attrition/Churn:
 - Goal: To predict whether a customer is likely to be lost to a competitor.
 - Approach:
 - Use detailed record of transactions with each of the past and present customers, to find attributes.
 - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
 - Label the customers as loyal or disloyal.
 - Find a model for loyalty.



From [Berry & Linoff] Data Mining Techniques, 1997

WSDM Cup 2018 <https://wsdm-cup-2018.kkbox.events/>



Classification: Application 4

- Sky Survey Cataloging
 - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
 - 3000 images with $23,040 \times 23,040$ pixels per image.
 - Approach:
 - Segment the image.
 - Measure image attributes (features) - 40 of them per object.
 - Model the class based on these features.
 - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

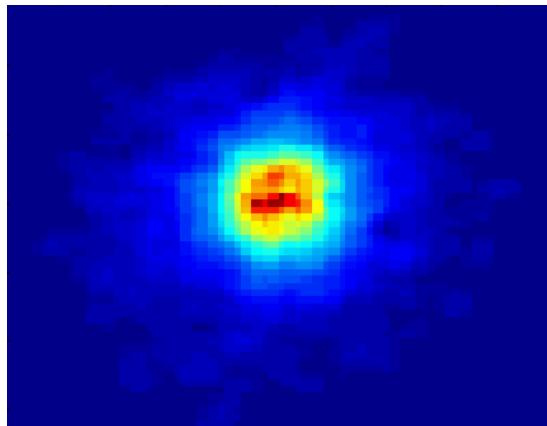
From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996



Classifying Galaxies

Courtesy: <http://aps.umn.edu>

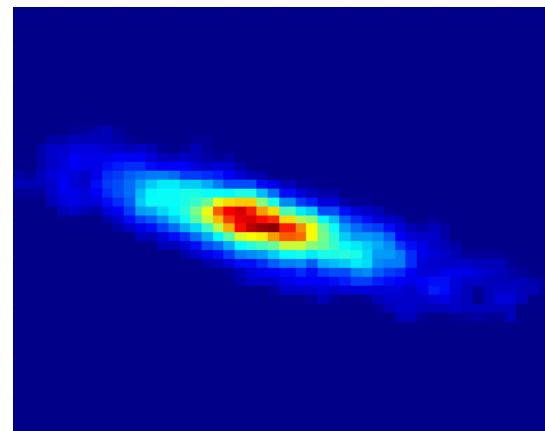
Early



Class:

- Stages of Formation

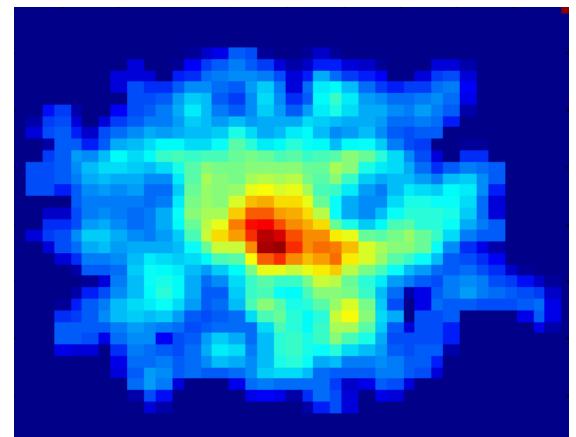
Intermediate



Attributes:

- Image features,
- Characteristics of light waves received, etc.

Late



Data Size:

- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

Clustering Definition

- Given a set of data points, each having a set of attributes, and **a similarity measure** among them, find clusters such that
 - Data points in one cluster are **more similar** to one another.
 - Data points in separate clusters are **less similar** to one another.
- Similarity Measures:
 - Euclidean Distance if attributes are continuous.
 - Other Problem-specific Measures.
- Unsupervised classification

分析哪些使用者是淺在客戶
先整理資料（分群）資料會有一些屬性
unsupervised classification: 沒有標注資料
ex:學生要分五群，分巨大大學部or研究生、性別、學運...目前只能分出三群->大群的再細分

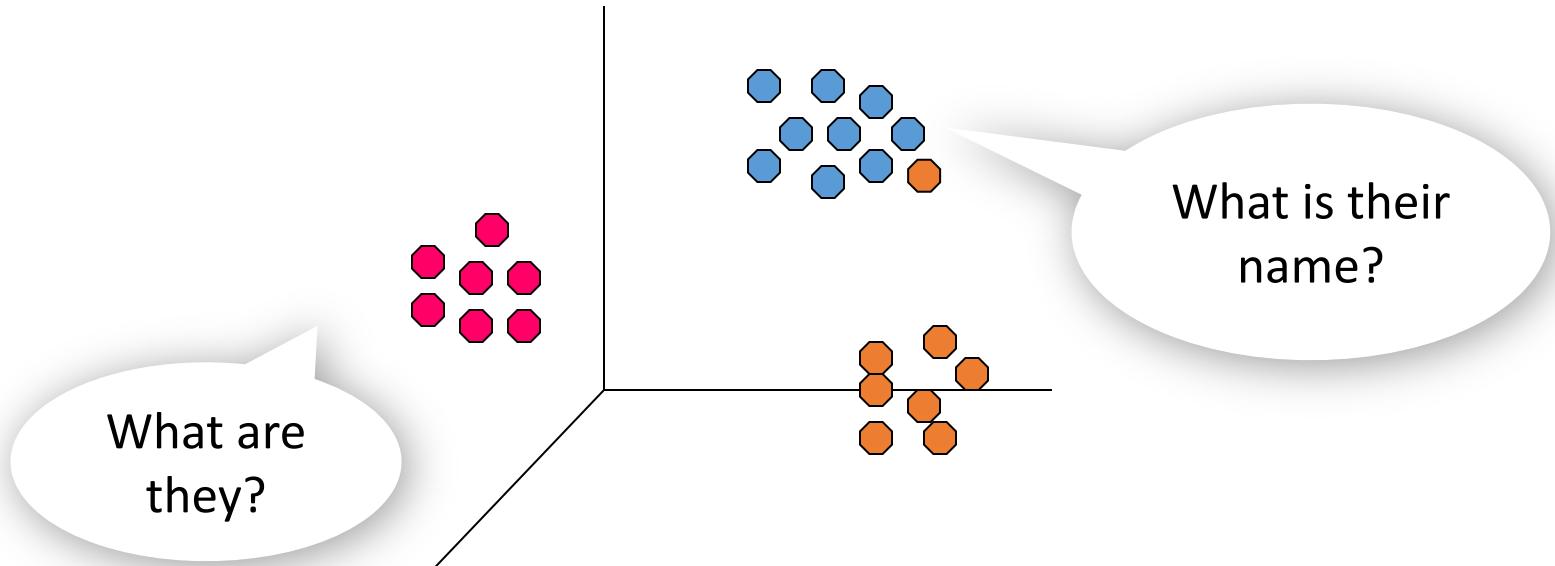


Illustrating Clustering

clustering: 像的擺在一起，不像的區分

Intracluster distances
are minimized

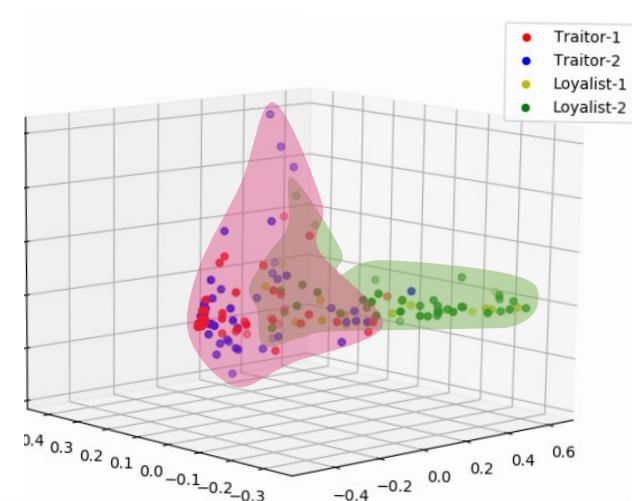
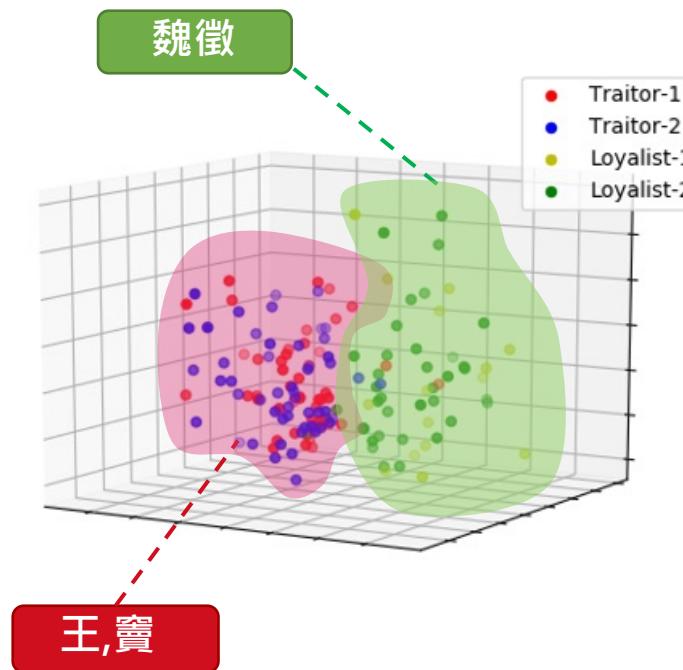
Intercluster distances
are maximized



| Euclidean Distance Based Clustering in 3-D space.

Word embedding clustering

- Use **Doc2Vec** to train paragraph vector
- We take **two types of volumes** which described for 魏徵 and 王竇



LDA doc-topic vector result

Clustering: Application 1

- Market Segmentation:
 - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
 - Approach:
 - Collect different attributes of customers based on their geographical and lifestyle related information.
 - Find clusters of similar customers.
 - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

- Document Clustering:
 - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
 - Approach: To identify **frequently occurring terms** in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
 - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Illustrating Document Clustering

- Clustering Points: 3204 Articles of Los Angeles Times.
- Similarity Measure: How many words are common in these documents (after some word filtering).

<i>Category</i>	<i>Total Articles</i>	<i>Correctly Placed</i>
<i>Financial</i>	555	364
<i>Foreign</i>	341	260
<i>National</i>	273	36
<i>Metro</i>	943	746
<i>Sports</i>	738	573
<i>Entertainment</i>	354	278

How about 記者快抄
<http://news.ptt.cc/>

Also, 島民衛星
<https://islander.cc/latest/>

監控台灣的新聞媒體



Clustering of S&P 500 Stock Data

分析股票也可以用分群
ex: 這幾家公司的上下游是一樣的，走勢會很像

- Observe Stock Movements every day.
- Clustering points: Stock-{UP/DOWN}
- Similarity Measure: Two points are more similar if the events described by them frequently happen together on the same day.
 - We used association rules to quantify a similarity measure.

	<i>Discovered Clusters</i>	<i>Industry Group</i>
1	Applied-Matl-DOW N,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Ho me-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP



Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
 - Produce **dependency rules** which will predict **occurrence** of an item based on occurrences of other items.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

$\{\text{Milk}\} \rightarrow \{\text{Coke}\}$

$\{\text{Diaper, Milk}\} \rightarrow \{\text{Beer}\}$

AB

$A \rightarrow B$

$P(B | A) = P(A \cap B) / P(A)$

$B \rightarrow A$

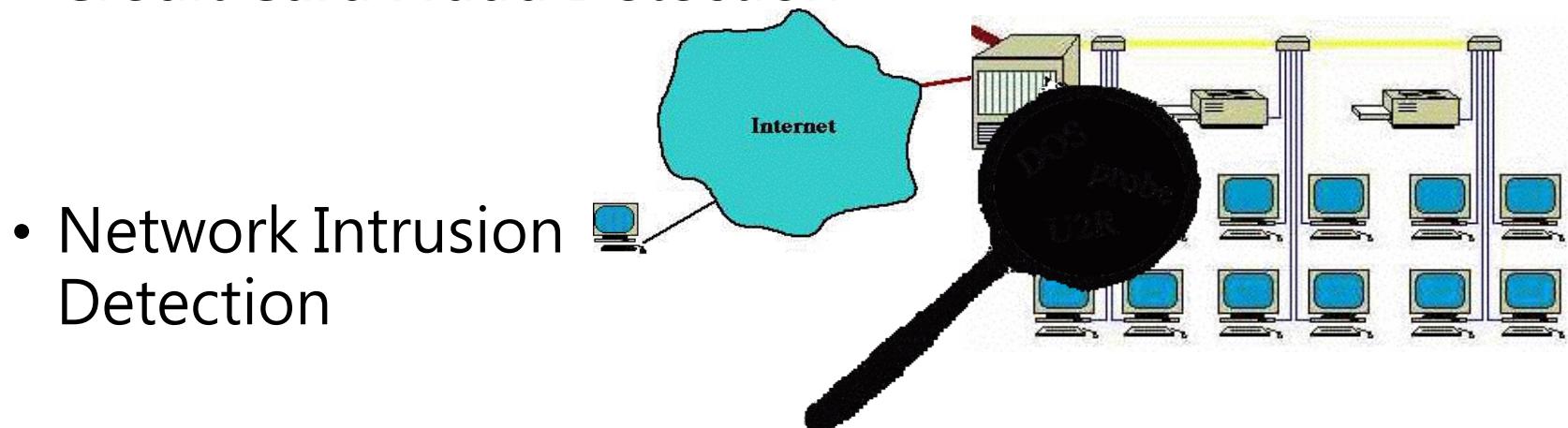
$P(A | B) = P(B \cap A) / P(B)$

Regression

- Predict a value of a given continuous valued variable based on **the values of other variables**, assuming a *linear* or *nonlinear* model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
 - Predicting sales amounts of new product based on advertising expenditure.
 - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
 - Time series prediction of stock market indices.

Deviation/Anomaly Detection

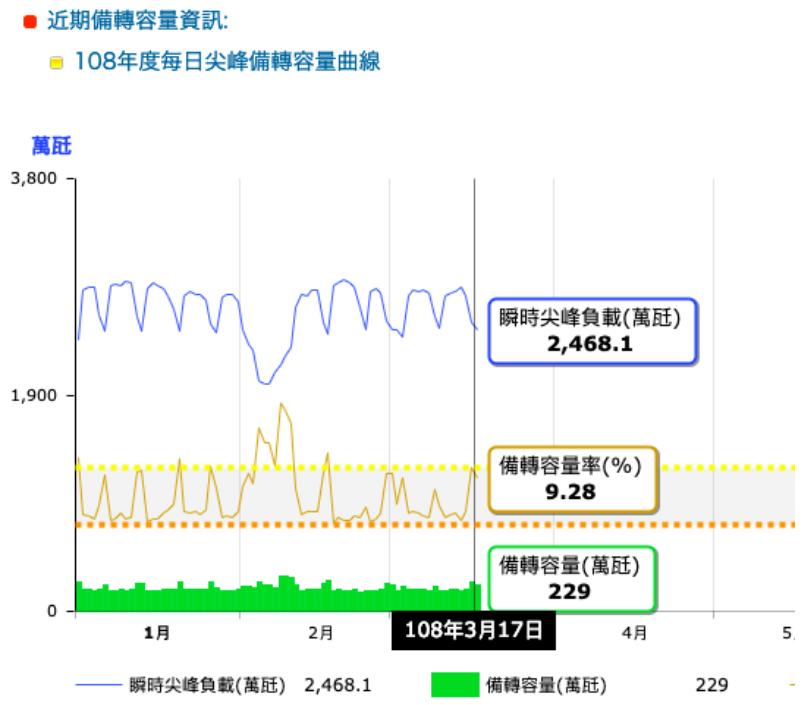
- Detect significant deviations from normal behavior
- Applications:
 - Credit Card Fraud Detection



Typical network traffic at University level may reach over 100 million connections per day

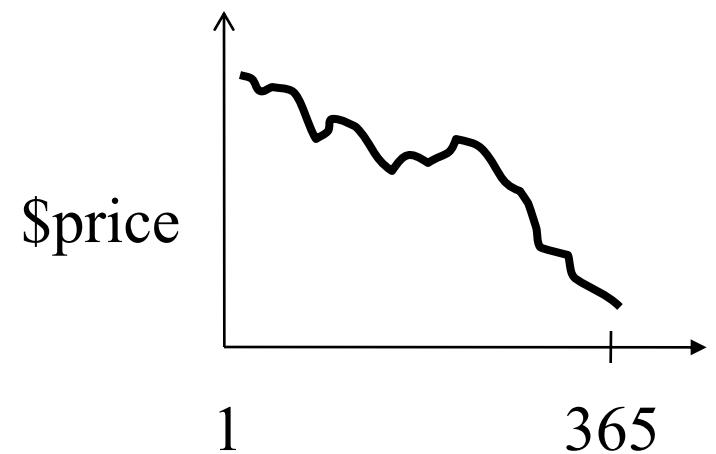
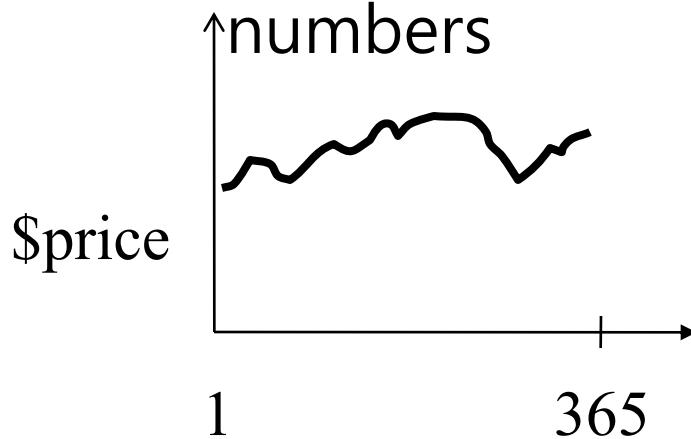
Time Series Analysis

- Trends analysis
- Regression
- Sequential patterns
- Similar sequences



Time Series Database

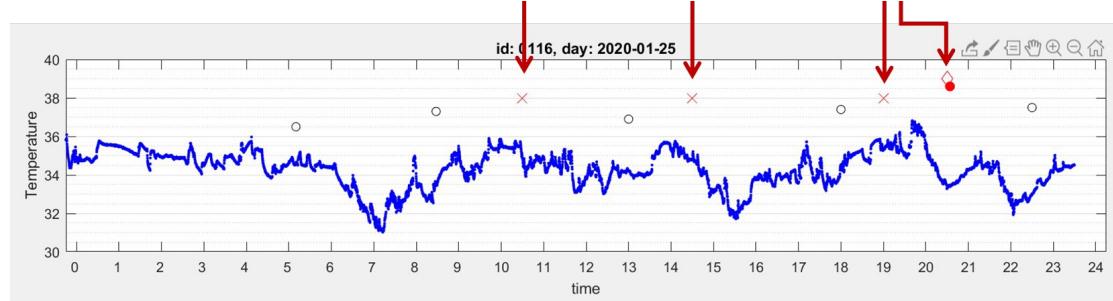
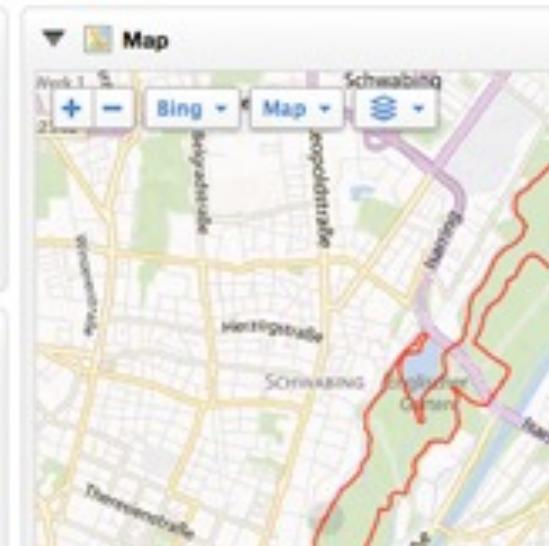
- Time series
 - Financial, marketing & production: stock price, sales number
 - Scientific: weather data, geological, astrophysics
- Time series DB
 - Databases with many time series of real numbers



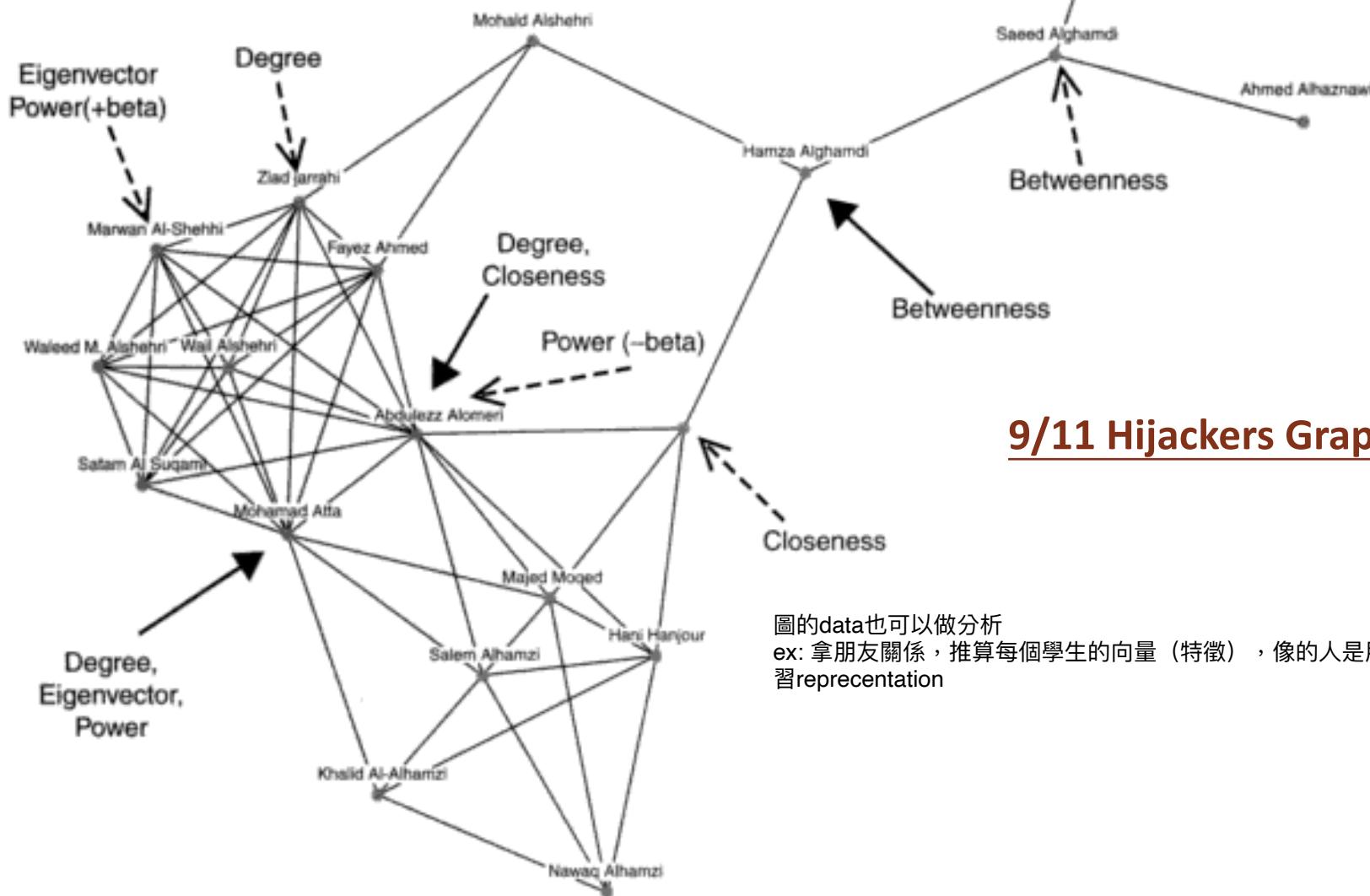
Time Series Database

- Query in time series DB
 - Searching for similar patterns
 - Whole matching
 - Subsequence matching
 - Examples
 - Identify companies with similar pattern of growth
 - Determine products with similar selling patterns
 - Discover stocks with similar movement in stock prices
 - Find if a musical score is similar to one of the copyrighted scores

Time series data in IoT Devices



Graph Mining



9/11 Hijackers Graph

圖的data也可以做分析
ex: 拿朋友關係，推算每個學生的向量（特徵），像的人是朋友->表示的學習representation

Reference from "The Text Mining Handbook", Ronen Feldman, James Sanger, P257.

Performance Measurement

- Efficiency
 - Time
 - Accuracy, Precision, Recall, Purity, P@k, ROC, ...
- Effectiveness (interestingness)
 - Objective measures; based on statistics & structures of patterns
 - e.g. support, confidence
 - Subjective: based on user's beliefs in data
 - e.g. unexpectedness, novelty

Interestingness

- A pattern is interesting if it is
 - Easily understood by humans
 - Valid on new or test data with some degree of certainty
 - Potentially useful
 - Validates some hypothesis that a user seeks to confirm



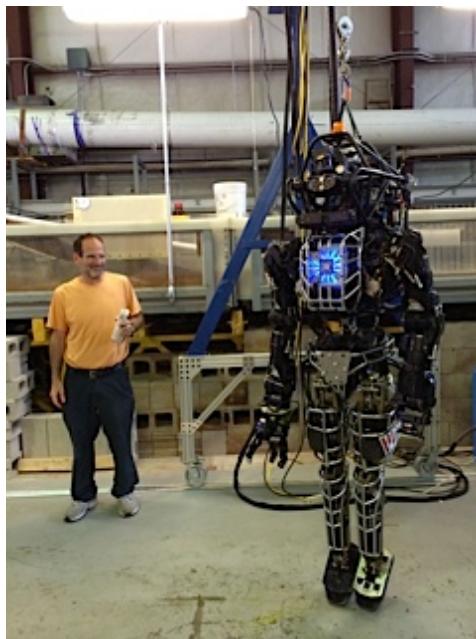
Challenges of Data Analytics / Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

Features & Challenges of KDD

- Big data
- Feature engineering
- Deep learning

A chatbot developed by Microsoft (Tay.ai) has gone rogue on Twitter, swearing and making racist remarks and inflammatory political statements.



Dream the future world:

More data about human behavior and knowledge of human
Google Now, Siri and Cortana show one example of mining and understanding your behavior
Ultimate goal: We are building brain for robots

[Robot 1](#)
[Robot 2](#)

模型訓練師筆記

Basic Data-driven solution

Machine learning Approach

Describe your data
/ features computer
can understand

需要對問題設計特徵讓電腦容易理解：需要一個專家

Domain Specific,
require domain
experts

Learning algorithm

如何建模與訓練

Optimizing the
weights on features

如何解釋，而不是fitting data

ISSUE 1: 問題定義

先有資料

- 資料關聯與特徵
- 更多資料?
- Value finding (problem design)
- Domain knowledge
- 資料收集與前處理

先有問題

- label/feature relation

資料與問題同時

ISSUE 2: 資料意義

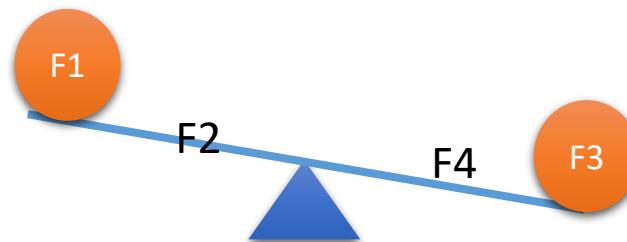
出乎意料的好做，只要喂夠大量的資料，但是電腦並不知道加法的邏輯
只是了解資料的特性而已：ex: F3, F6夠大的時候L2會多一

F1	F2	F3	F4	F5	F6	L1	L2	L3
1	4	5	3	5	2	4	9	7
2	3	1	5	2	2	7	5	3
4	3	5	3	2	2	7	5	7
3	2	2	1	7	8	5	0	0
4	0	3	2	9	9	7	0	2
1	3	2	7	1	0	8	4	2
6	2	1	1	4	5	7	6	6

ISSUE 2: 資料意義

在做時四個資料是獨立的，會做得不好

F1	F2	F3	F4	L
10	5	10	4	1
5	10	8	7	0
6	7	9	4	1
3	15	12	4	0
3	9	8	4	0



ISSUE 3: Coverage and Diversity in your Feature Space

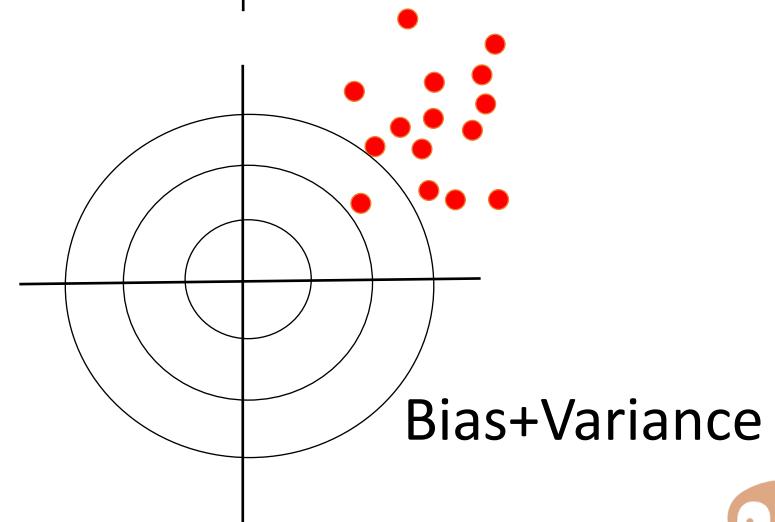
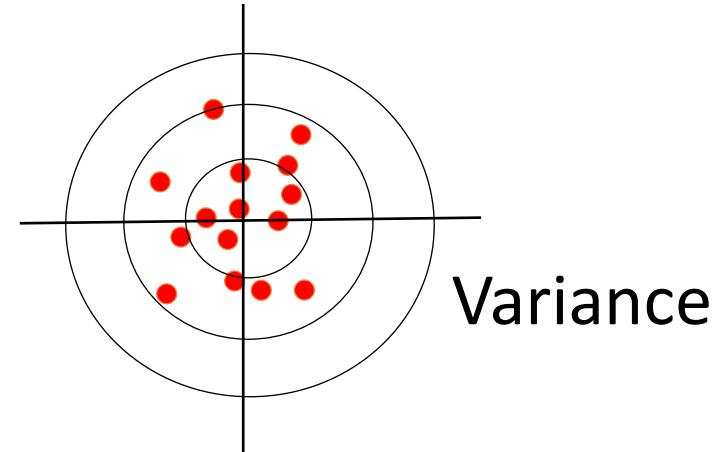
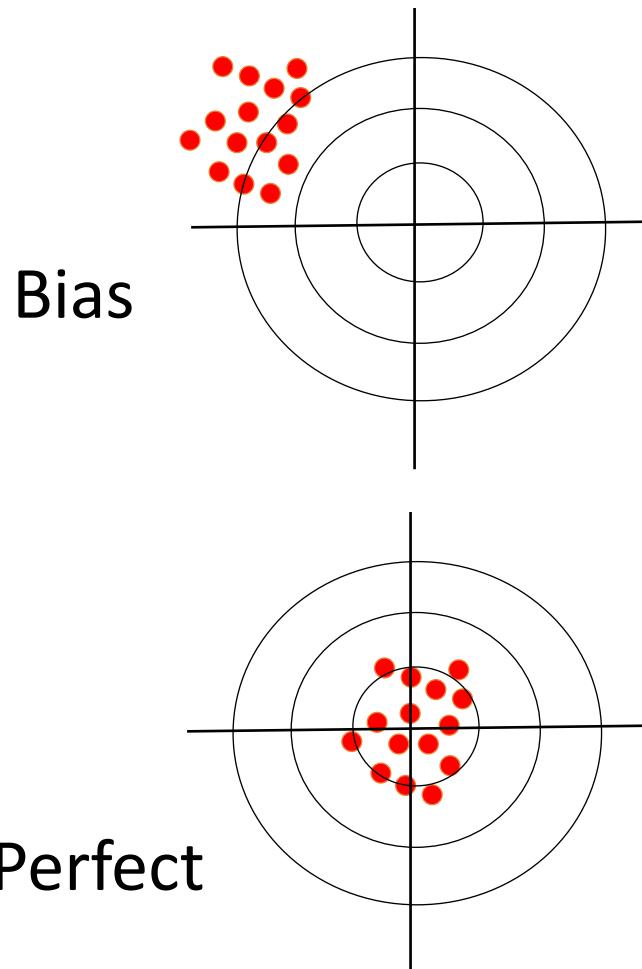
- full combinations of feature values?
- Numerical range
- Normalization / scaling

條參數train data反而不困難，有固定的pattern



ISSUE 4: Bias & Variance

希望模型train越好，會用越複雜的東西，fit非常好，但結果卻不好
bias 越小 variance 越大



Bias-Variance Trade-Off

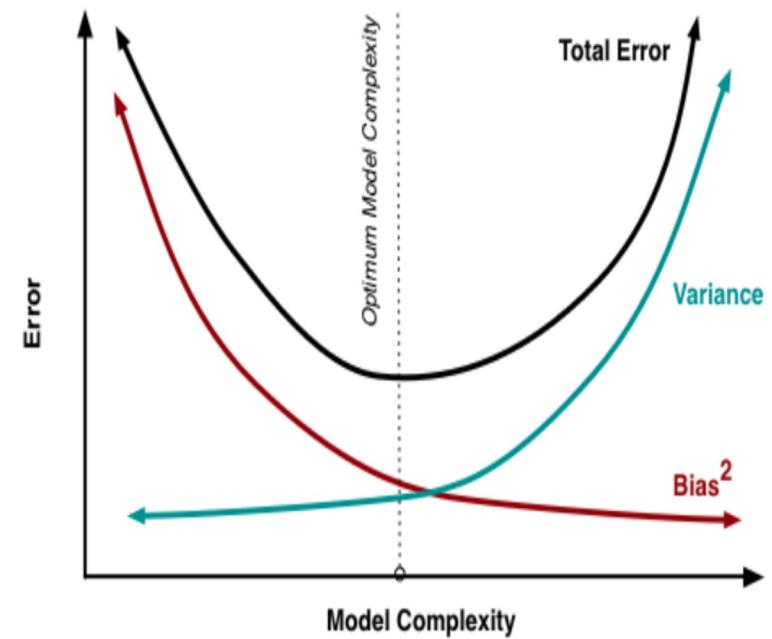
As model complexity increases we see the characteristic U-shape of total error

We reduce bias to an optimum where increasing variance starts to dominate

Finding the right balance is a key machine learning skill

High bias, low variance = underfitting

Low bias, high variance = overfitting



ISSUE 5: Can't fully trust model

ML model is just an optimization process to **fit** your data in a **simple** representation.

