



-資管人如何處理大數據與 AI 課題

Yihuang Kang

Data-Driven Discovery (D3) Lab @ NSYSU

Instructor

- Yihuang K. Kang (康藝晃), PhD. 為中山大學資訊管理學系助理教授。曾任美國匹茲堡大學醫學中心資深程式設計師與分析師。專長為統計機器學習、整合分析平台導入、電子化健康醫療數據分析、與遠距醫療系統開發。



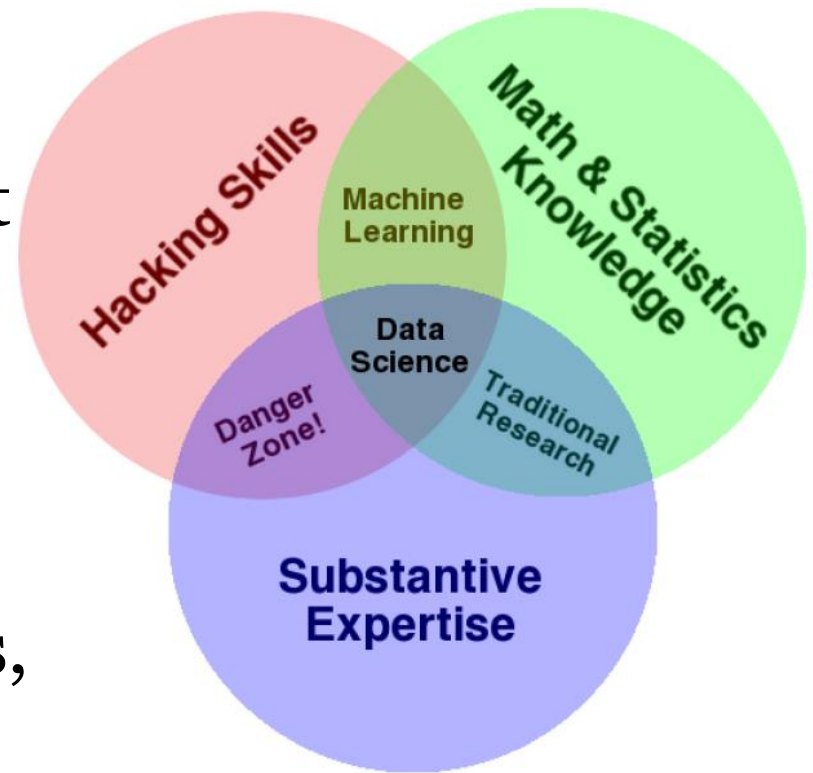
- 近年回台任中山大學資訊管理學系助理教授，協助中山大學建置整合數據分析平台，並結合專長教授數據分析與機器學習相關課程。為中山大學智慧電子商務研究中心與跨領域及數據科學研究中心成員。資策會教育訓練(機器學習)講師。已有多項研究發表於頂尖的資料探勘與健康醫療研究期刊。擁有多個實務資訊系統管理、程式設計、與商業分析專業認證。

Outline

- Brief Introduction to Data Science
- Fundamental of Data Analytics
- Programming vs. Machine Learning
- Towards Interpretable AI
- From Business Intelligence to Unified Analytics
- Build Our Data (Dream) Team!

What is "Data Science"

- “Data science is a multidisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured.”



[Drew Conway's Venn diagram of data science](#)

A Brief History of Data Science

1974

The term “data science” first appeared in Peter Naur’s *Concise Survey of Computer Methods*.

1989

The Knowledge Discovery in Databases, which would mature into the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, organized its first workshop.

1994

Business Week ran the cover story, Database Marketing, revealing the ominous news companies had started gathering large amounts of personal information, with plans to start strange new marketing campaigns

Business

Database Marketing

Jonathan Berry

1994年9月5日 下午12:00 [GMT+8]

It may not be celebrated as a national holiday, but it's a pretty big deal around here. Happy birthday from the Claridge Casino Hotel, Atlantic City.

What time is it now in Israel? What is Mama cooking today? We at AT&T know exactly how you feel and are aware of your need to call and speak with those close to you whenever you wish.

2001

William S. Cleveland laid out plans for training Data Scientists to meet the needs of the future. **He introduced data science as an independent discipline.** He presented an action plan titled, *Data Science: An Action Plan for Expanding the Technical Areas of the field of Statistics*.

2002

Launch of Data Science Journal

2006

Hadoop 0.1.0, an open-source, non-relational database, was released.



2012

Thomas H. Davenport and D.J. Patil published *Data Scientist: The Sexiest Job of the 21st Century*.

2013

IBM shared statistics showing 90% of the data in the world had been created within the last two years.

2015

Using Deep Learning techniques, Google's speech recognition, Google Voice, experienced a dramatic performance jump of 49 percent.

The Rise of Data Science

DATA

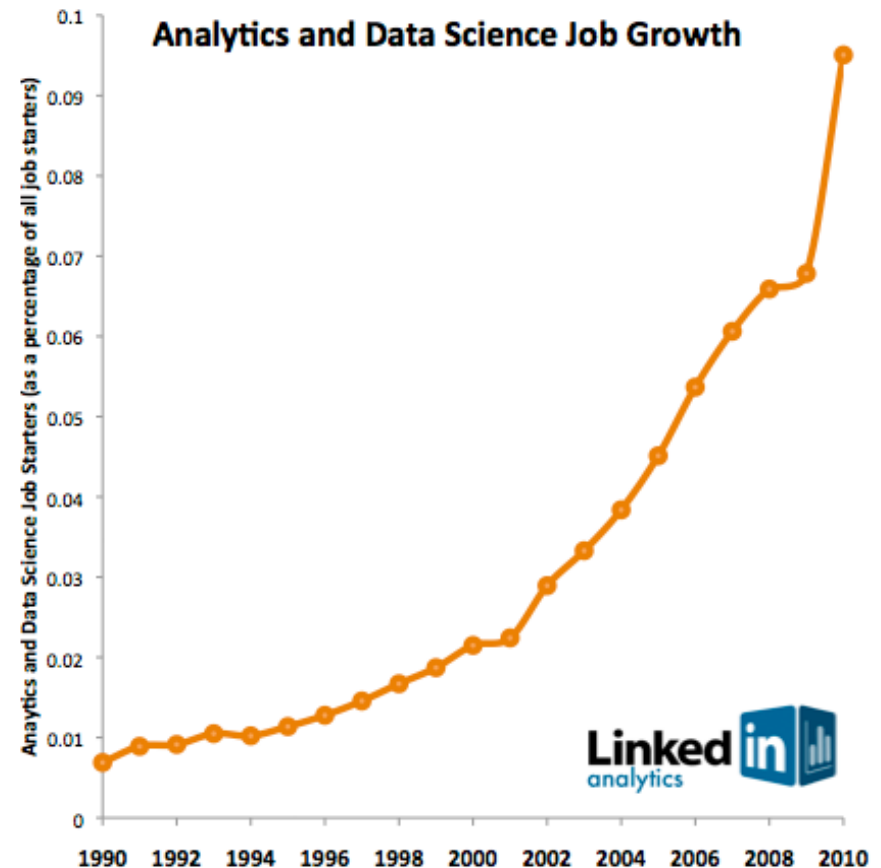
Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

 SUMMARY  SAVE  SHARE  16 COMMENT  TEXT SIZE  PRINT  \$8.95 BUY COPIES

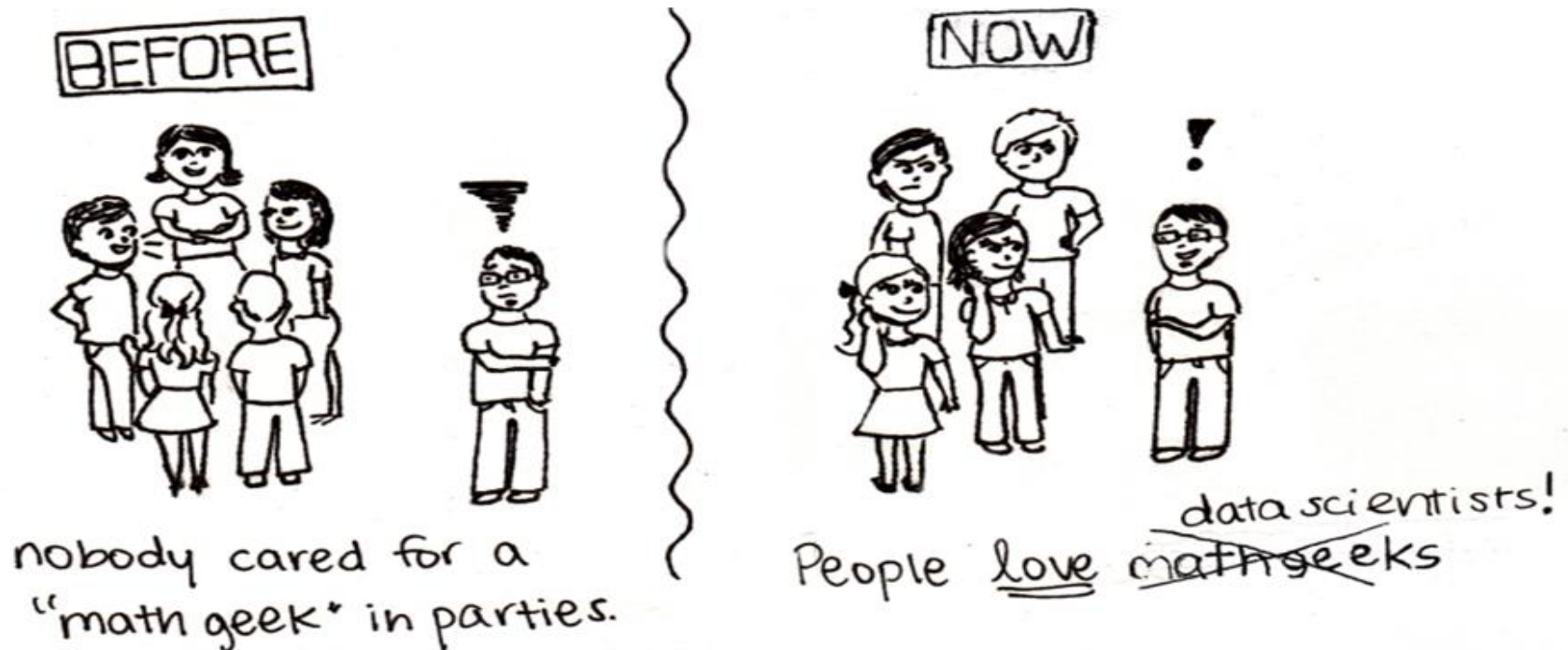
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in



The Rise of "Data Scientists"

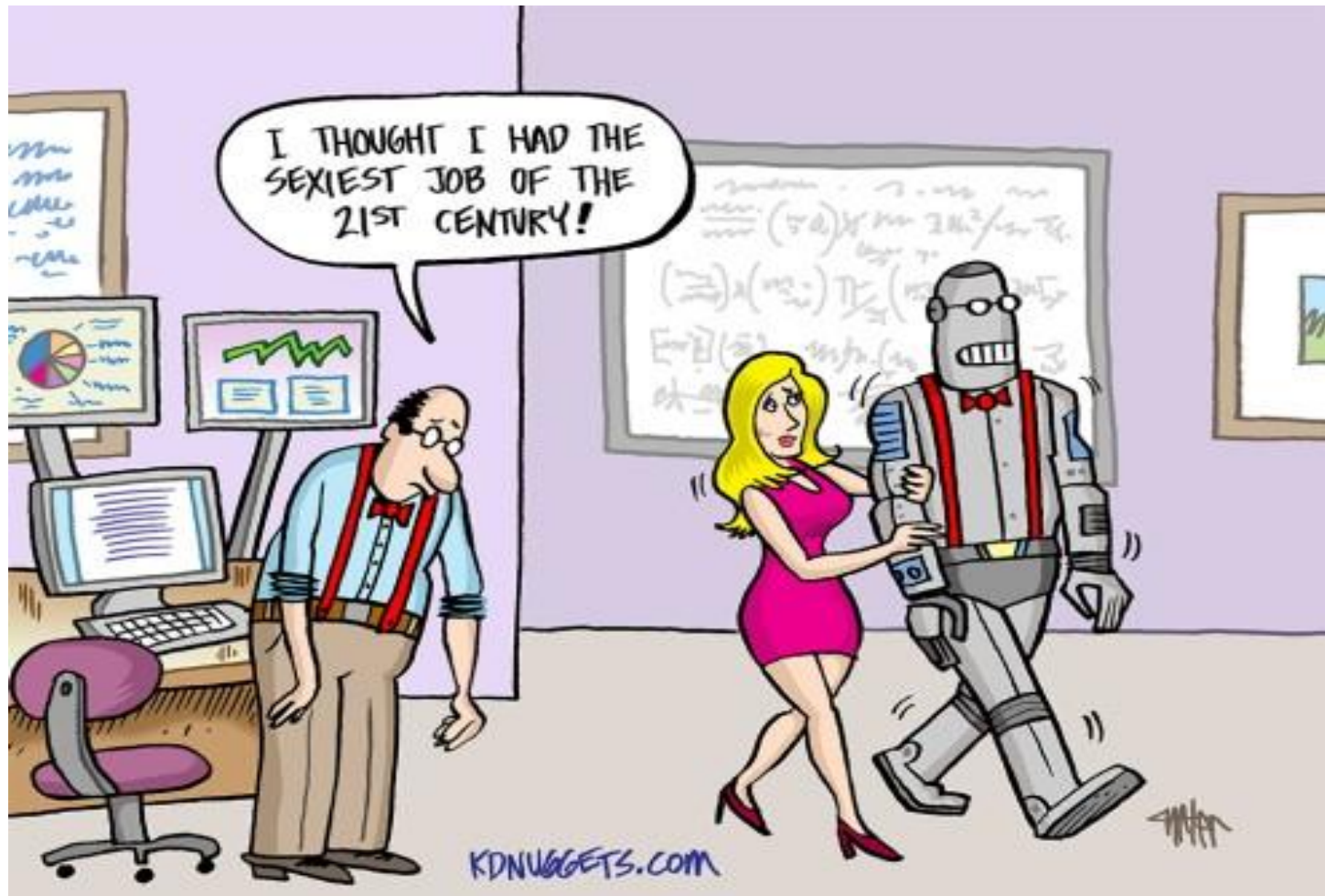
"Data Scientist is The Sexiest Job of the 21st Century"

—T. Davenport & D.J. Patil, *Harvard Business Review*



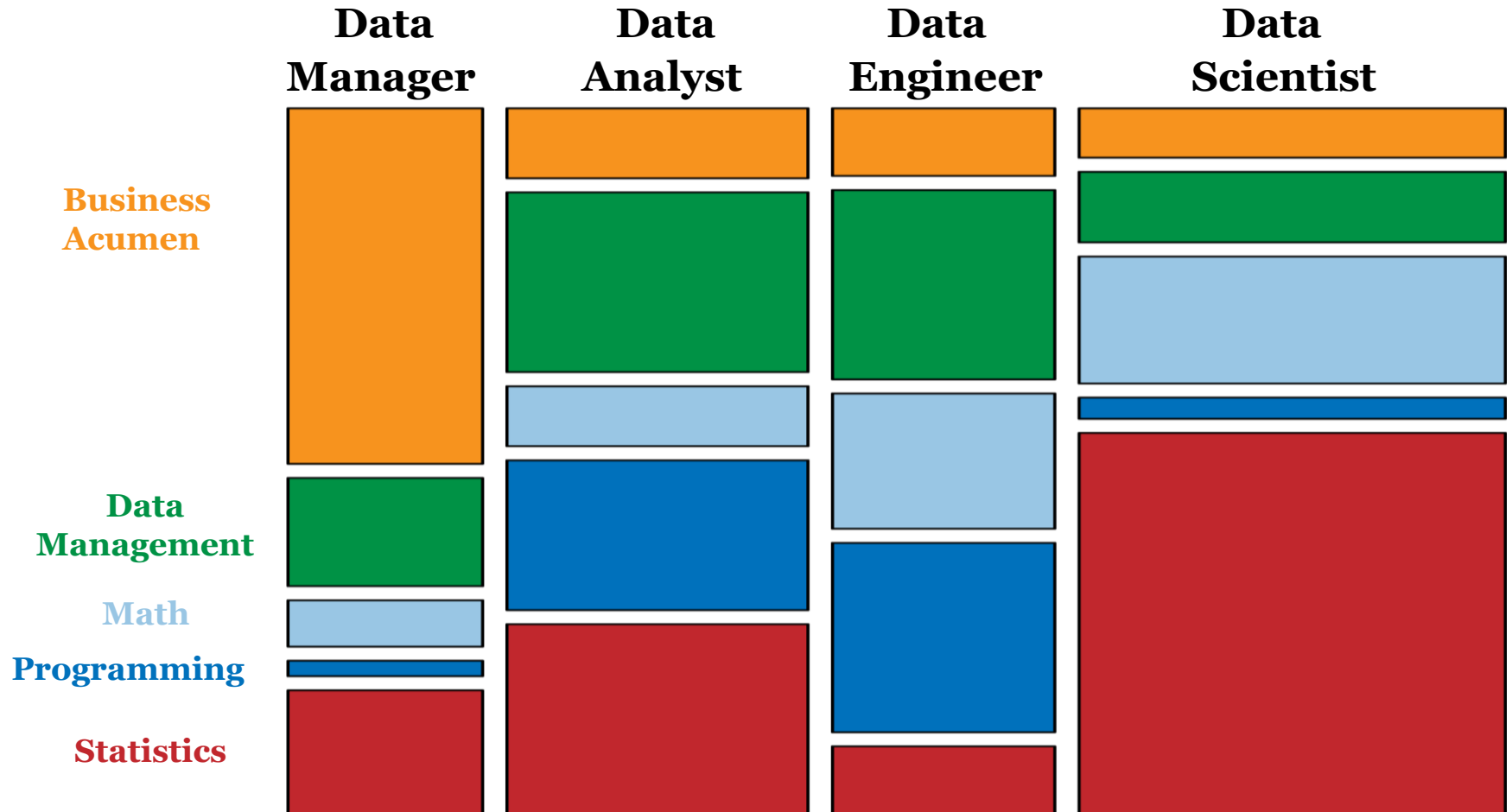
Source: <http://www.techjuice.pk/how-to-become-a-data-scientist-for-free/>

The "Dilemma" of the Data Scientists



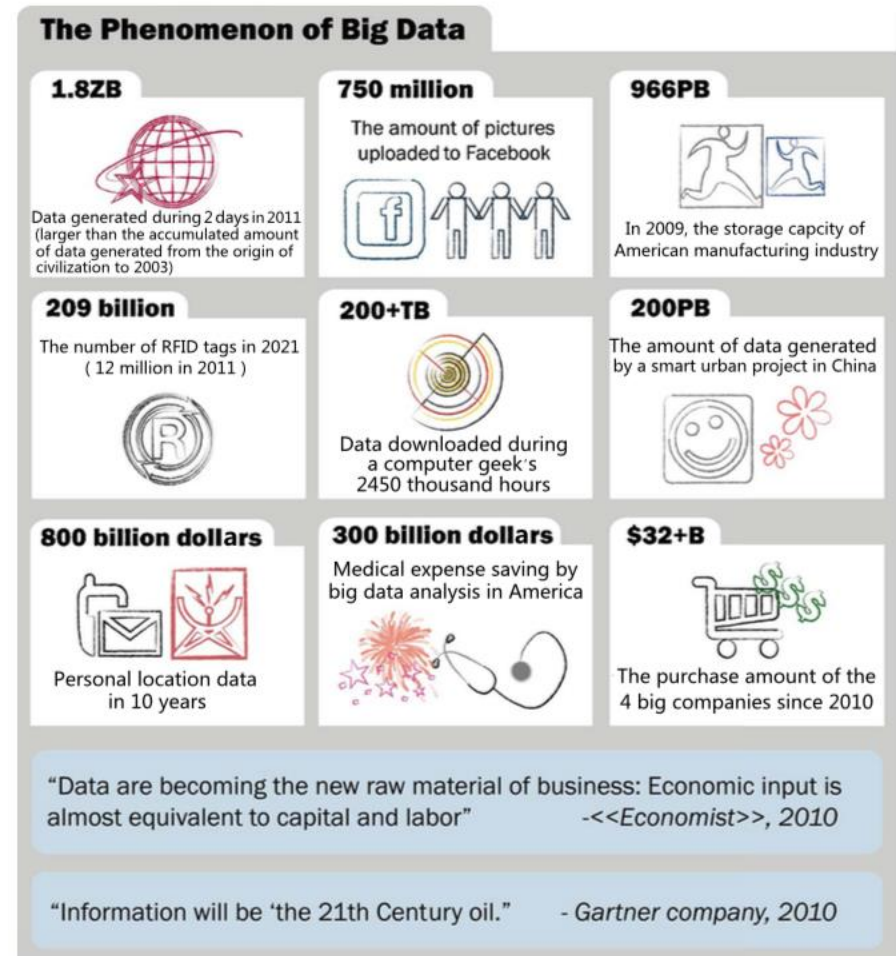
Source: <http://www.kdnuggets.com/2016/08/cartoon-data-scientist-sexiest-job-21st-century.html>

“Data” Careers



Big Data

- According to a report from International Data Corporation (IDC), in 2011, the overall created and copied data volume in the world was 1.8ZB, which increased by nearly 9 times within five years. By 2020, it's estimated 35ZB.
- Challenges: **big data management and analysis**



What is "Big Data"

- Many people have defined "Big Data" with 3Vs, 4Vs, 5Vs..., many Vs!



- My definition is: *"Too much and complicated data to be processed by a single machine with reasonable time or resources"*.

Where does the big data come from?

- **Traditional Data**

- Any digitized contents and/or archives acquired by traditional ways, e.g. survey data, interview records, and documents.

- **Machine Data**

- Sensor data, web logs, any log data from monitoring information systems.

- **Network Data**

- The network of computers (The Internet)
- The network of people (Social Networks)
- The network of things (Internet of Things)

Types of Big Data

- **Structured data**

- Data with clear schema/metadata/data model that describes & defines how the data elements relate to one another. E.g. relational databases, data cubes/warehouses.

- **Semi-Structured data**

- Data with only tag/field definitions but without formal structures of data models to define relations. E.g. data used in information exchanges, such as XML & JSON. Emails/pictures/other files with tags/field definitions.

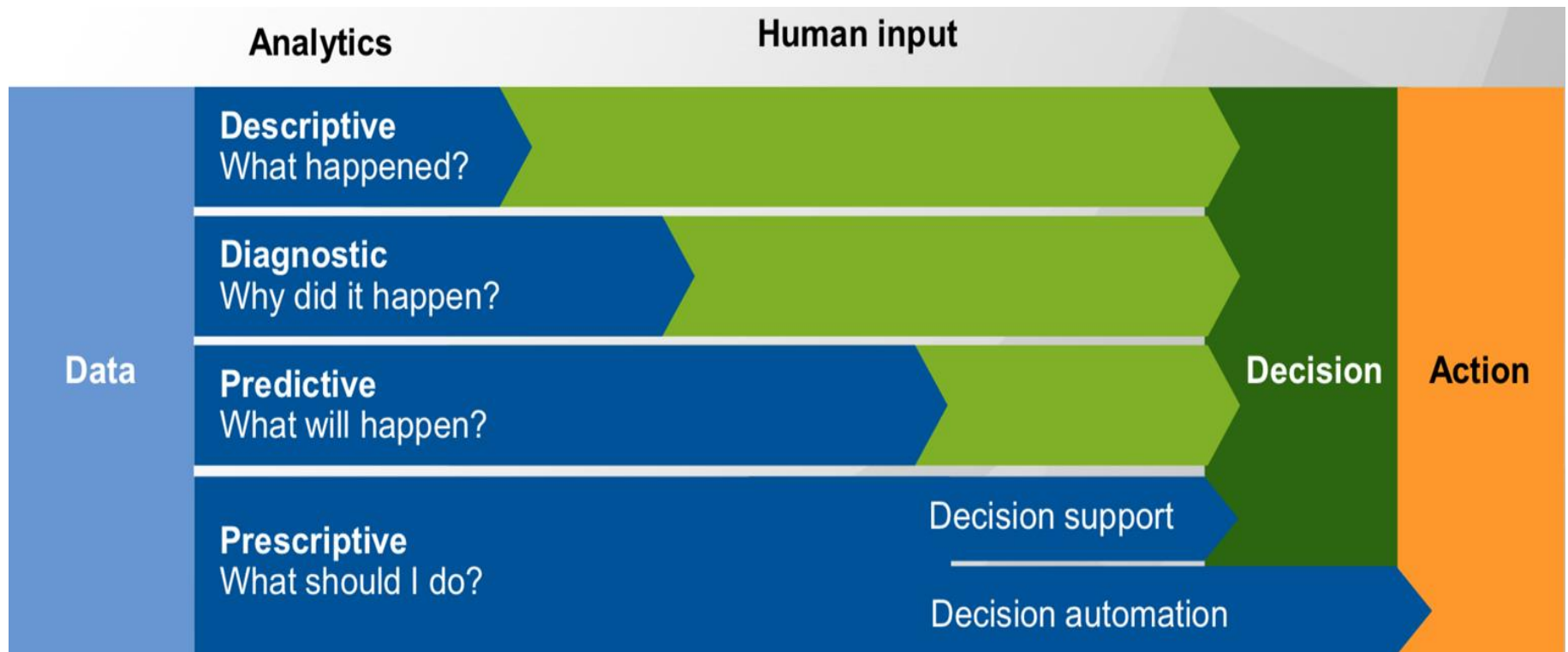
- **Unstructured data**

- Unorganized data without any pre-defined schema. E.g. body of an e-mail message, pictures, audio, and video.

The Value of Data Analytics



The Value of Data Analytics (cont.)



Source: <https://www.gartner.com/en/documents/2594822> , Linden et. Al. "Extend Your Portfolio of Analytics Capabilities", Gartner Research

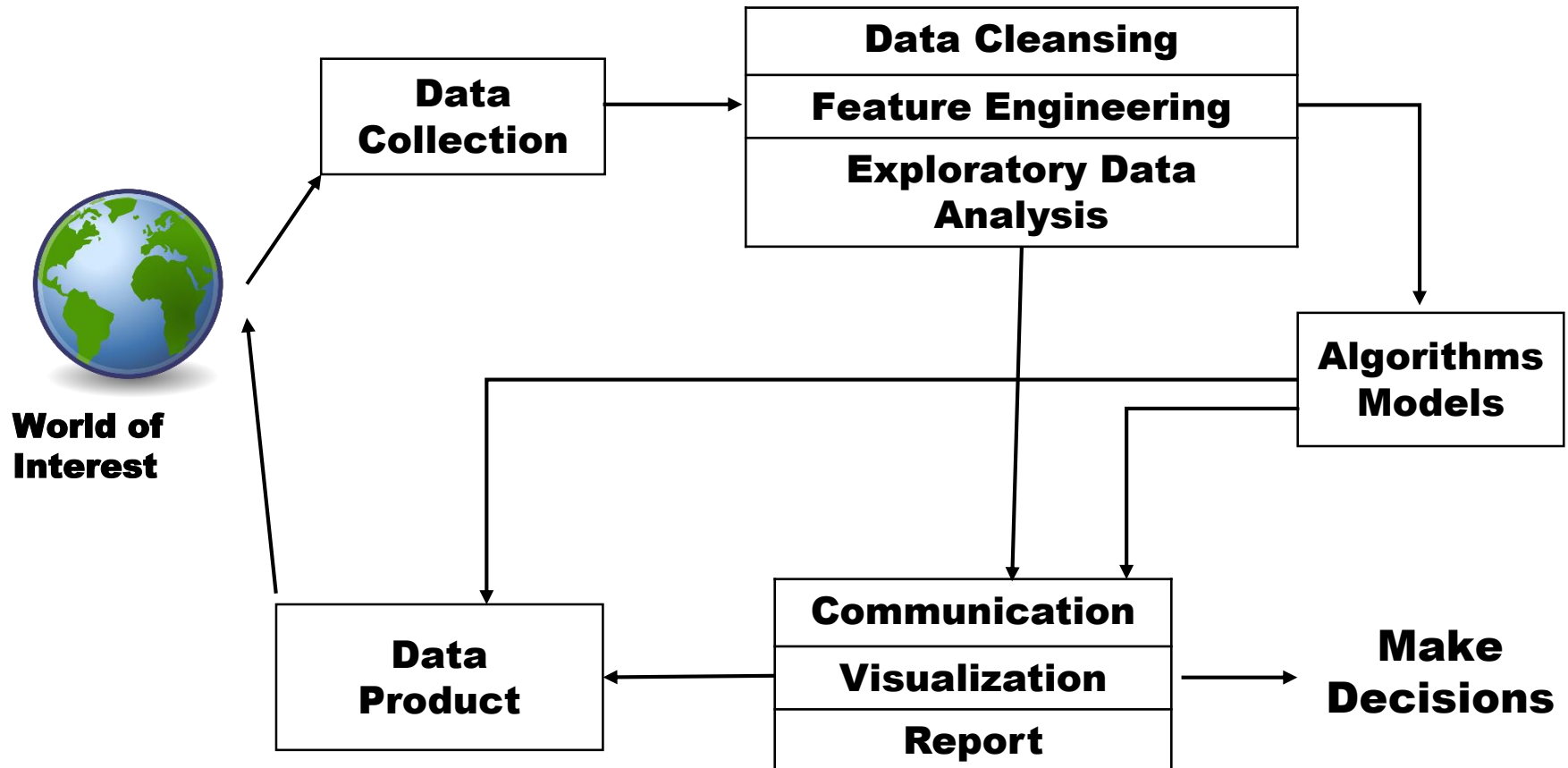
Data is NOT always the cure!

- The "Big Data" does eliminate intuition. However, our interpretations of it have great impact on the results. Let's check out [this article in New York Time](#) . It says “*Let’s put everything in and let the data speak for itself.*” This is a horrible quote and don't let it mislead you.

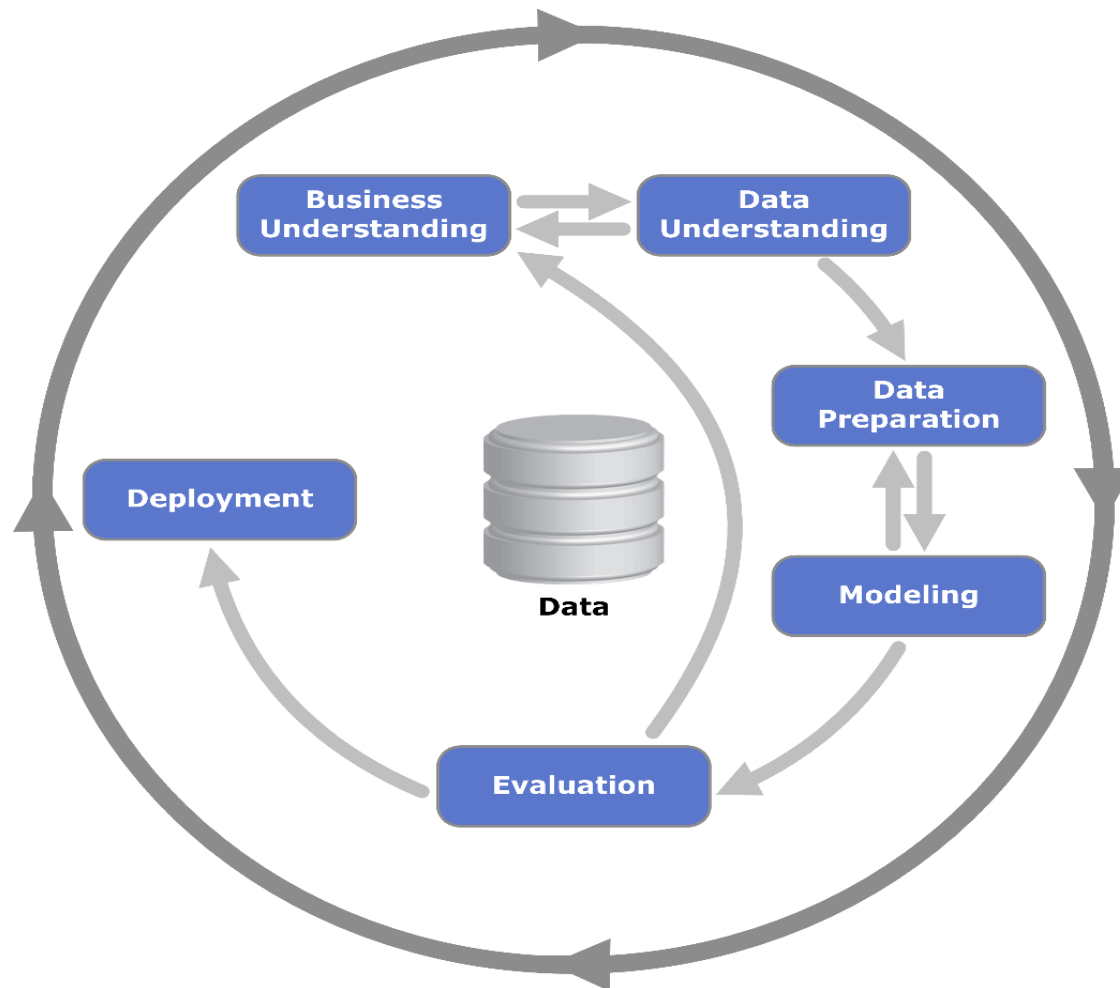
*"...Data is just a quantitative,
pale echo of the events of our society..."*

—[O'Neil, "On Being a Data Skeptic"](#)

The Process of Data Analytics



Cross Industry Standard Process for Data Mining (CRISP-DM)



Group Discussion #1

- 在一個數據分析(Data Analytics)專案中，你認為哪個步驟會花最多的時間，為什麼？



“Data”—The observations of the world

- There are 13,006 data points for total **359** patients who are children between 1 and 6 years old and are hospitalized in Children’s Hospital of Pittsburgh in 2008. The data consists of 1 categorical variable as a patient’s current location (ICU or Floor) and 5 continuous variables as 5 vital signs for a patient from bedside monitoring system (or manually recorded by nurses).

Hidden states?					Observations	
					↓	
Diastolic Blood Pressure (mm Hg)	Systolic Blood Pressure (mm Hg)	Respiratory Rate (bpm)	SpO2 Bedside Monitor (%)	Temperature (C)	Location	Duration (Hour)
64	117	29	100	37.5	ICU	1
65	110	21	99	37.5	ICU	1
65	110	21	99	37.5	ICU	1
65	110	21	98	37.5	ICU	1
66	90	26	96	36.7	Floor	2
67	97	27	98	36.7	Floor	1
67	97	27	96	36.7	Floor	2

Reference
normal ranges:

Diastolic Blood Pressure (mm Hg)	Systolic Blood Pressure (mm Hg)	Respiratory Rate (bpm)	SpO2 Bedside Monitor (%)	Temperature (C)
55 - 75	90 - 100	15 - 30	≥ 95%	36.33 - 37.56

Learning Problem—A Toy Example

Training Set

○	×	○
○	○	×
×	×	○

 = ○

○	×	×
○	○	○
×	×	○

 = ○

×	○	○
○	×	×
×	○	×

 = ×

×	×	×
○	×	○
○	○	×

 = ×

Testing/Unseen Set

×	○	×
○	×	○
×	○	○

 = ?

×	○	×
×	○	○
×	○	○

 = ?

- What models (or patterns, rules, functions, ...) did you learn from the above figure? Is your "model" different from others?

Learning Problem—A Toy Example_(cont.)

Training Set

○	×	○
○	○	×
×	×	○

 = ○

○	×	×
○	○	○
×	×	○

 = ○

×	○	○
○	×	×
×	○	×

 = ×

×	×	×
○	×	○
○	○	×

 = ×

○	×	○
○	○	×
○	×	×

 = ○

○	×	×
×	○	×
×	○	○

 = ×

Testing/Unseen Set

×	○	×
○	×	○
×	○	○

 = ?

×	○	×
×	○	○
×	○	○

 = ?

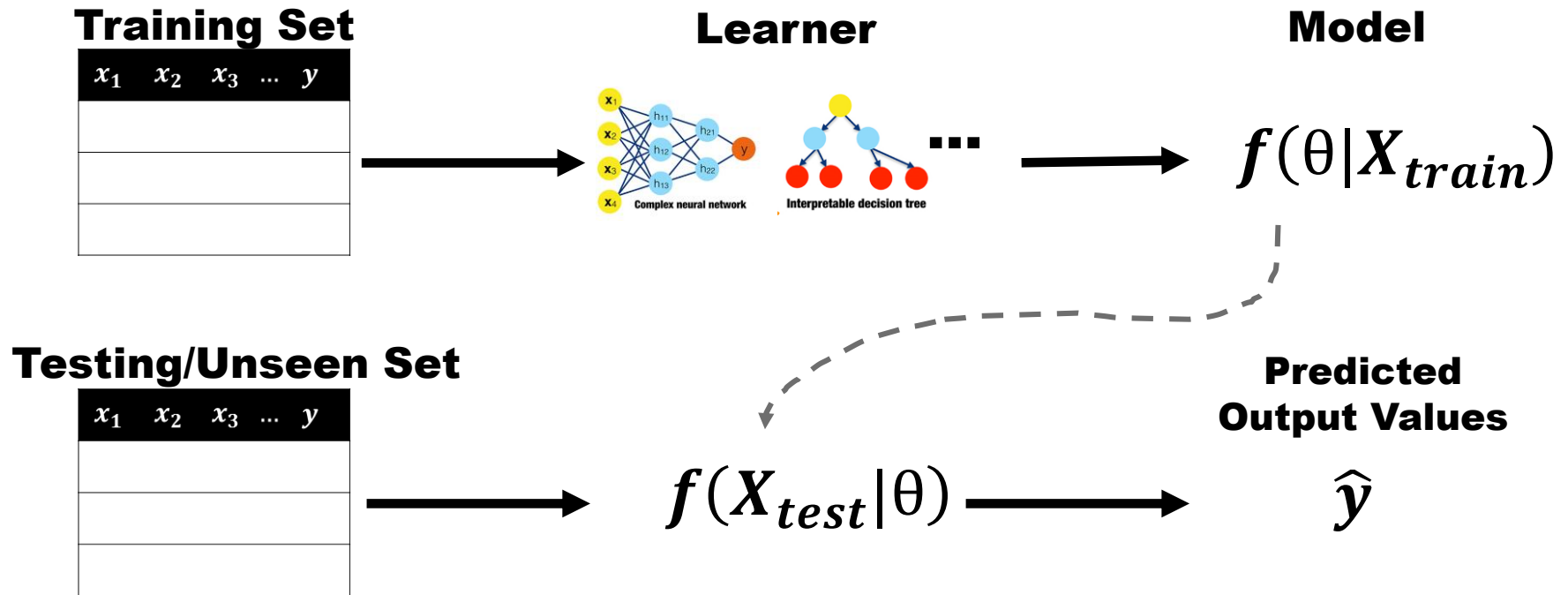
What Is "Learning"?

- We learn to recognize an object, say "cat" (or in Chinese 貓, whatever you call it does not matter) by observing the characteristics and behaviors of a cat instead of learning by its definitions—***we learn from data (observations)***.
- Machine Learning is about ***how to give machines abilities to learn without being definitely programmed***—the abilities to recognize complex patterns, take reasonable actions, and even learn how to learn.

What Is "Learning"? (cont.)

- The **goal** of the learning is to *be able to reason (predict) observations that we haven't seen before*.
- The **quality** of the learning could be determined by *how close our prediction is to the true value*, which is often represented as error functions or measures in statistical machine learning.

What Is "Learning"? (cont.)



Error = Differences between y and \hat{y}

What Is "Learning"? (cont.)

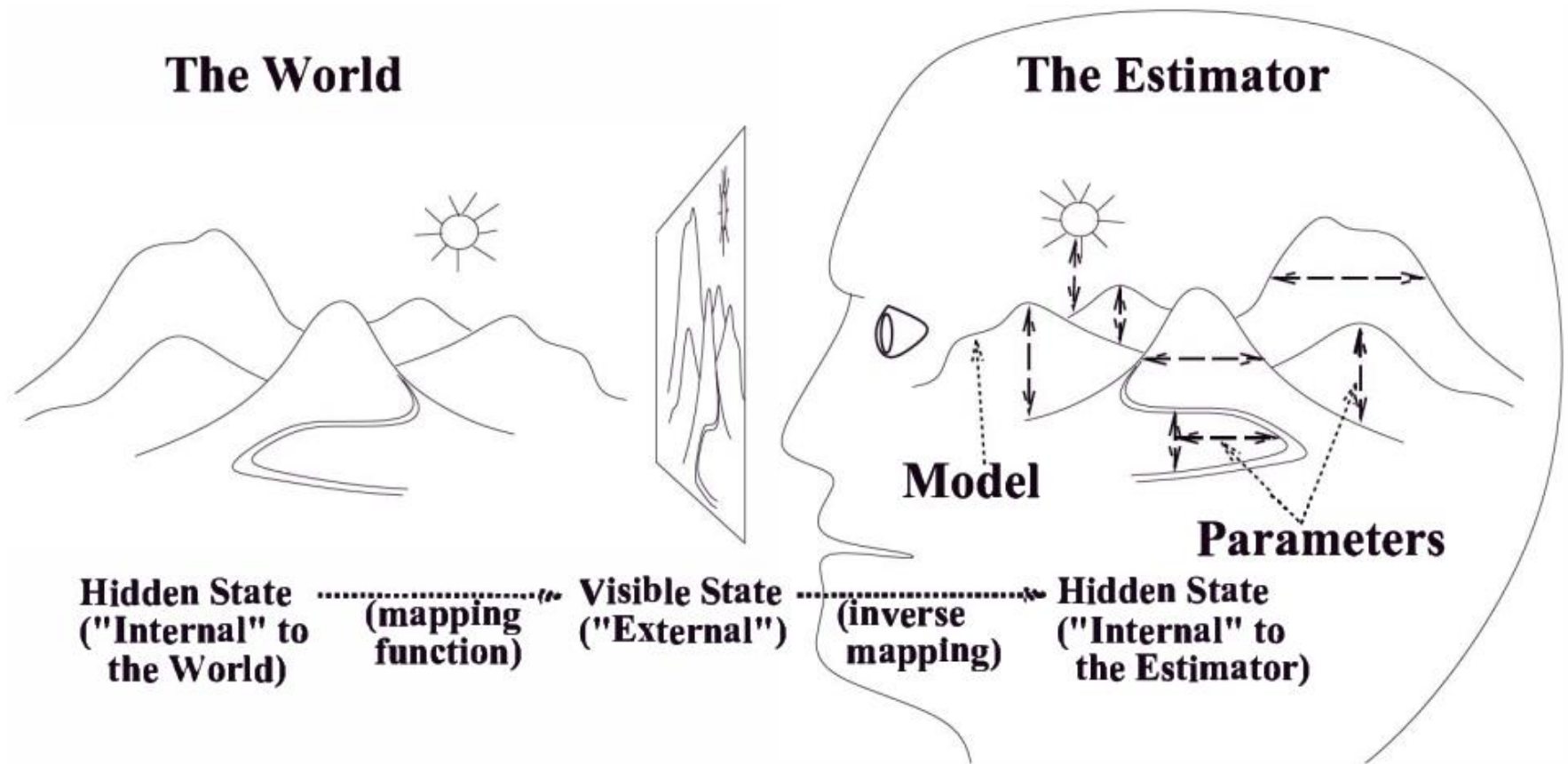
- There exists a complex function $f(\mathbf{x}|\boldsymbol{\theta})$ that minimizes the differences between actual values \mathbf{y} and predicted/estimated values $\hat{\mathbf{y}}$.

$$\mathbf{X} \longrightarrow f(\mathbf{X}|\boldsymbol{\theta}) \longrightarrow \hat{\mathbf{y}}$$

- Learning is about how to find this " $f(\mathbf{x}|\boldsymbol{\theta})$ "!

Model Thinking—

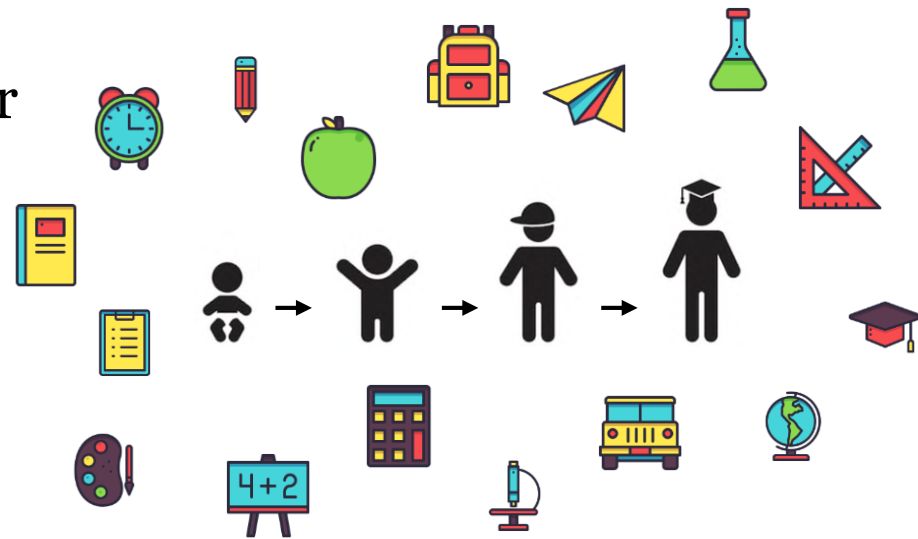
Perception, Reasoning, & Learning of AI



Source: R. P. Rao, "An optimal estimation approach to visual perception and learning," Vision Res., vol. 39, no. 11, pp. 1963–1989, 1999.

Human Brain Works Like Machines, But Much More Efficiently

- Human brains learn different tasks in every moment. Human receive multiple **input** and get feedbacks/**output** on whatever they take.
- AlphaGo is the greatest Go player in present. However, it needs to play millions of Go games to achieve the current level. Although AlphaGo is brilliant at playing GO, it can just play Go game. Not like human, machines are hard to learn many tasks.



It is all about “Input” and “Output”

- Learning algorithms (or statistical methods), like human, can take inputs and generate outputs. They can be represented as *predictive modeling* notations—as simple as both sides of an equation. Here we define:

$$[Y_1, Y_2, \dots Y_m; T_1, T_2, \dots T_n] = f([X_1, X_2, \dots X_i; A_1, A_2, \dots A_j])$$

where

Y: continuous output/dependent variable

T: discrete output/dependent variable

X: continuous input/independent variable

A: discrete input/independent variable

A bit about “Levels of Measurement”

- Levels (Scale) of measurement here is classification of information or value of a variable:
 - **Nominal** variable has values that are distinct symbols/labels. No relation, such as ordering or distance, is implied. E.g. Gender (Male, Female).
 - **Ordinal** variable is similar to nominal variable but we can rank order the labels. Note that there is still no “distance” notion among labels. E.g. feel-like temperature (Cool, Warm, Hot).
 - **Interval** variable has values that the order and distance among them make sense. But the distances are fixed and any mathematical operations are not allowed. For example, we normally don't say temperature 10 degree (in C) is “twice colder” than 20 degree.
 - **Ratio** variable allows real numbers that makes any operations, such as ratios and differences, logical. E.g. cost of living in USD.

A bit about “Levels of Measurement”_(cont.)

- In this class, we consider more general terms of the measurement—discrete (categorical) and continuous (numeric) variables.
 - **Discrete variable** is a special data type (also called *factor*) that defines ordered or unordered *levels* with *labels* (possible values of a variable). They include the aforementioned nominal and ordinal variables.
 - **Continuous variable** has real and numeric values that can be used in any aforementioned mathematical operations.

Statistical Inference

- The world is *complex, stochastic, and dynamic*. It keeps producing *data* and will never stop. Now you have *the data*, whatever the format it is, it may give you some *information* about the world of interest.
- The *Statistical Inference* is a process of deducing *properties* of your data assumed to be generated by random processes. By "properties" here, we mean models, estimations, acceptance of hypothesis..., all the information that helps you understand the world of interests.

Different Terms, Same Practices

- Most of terms used in Statistics Modeling and Machine Learning actually mean the same and are often interchangeable.

Statistical Modeling	Machine Learning
Explanatory/Independent Variable	Input, Predictor, Feature
Dependent/Outcome Variable	Output, Target
Model	Network, Graphic, Algorithm
Parameter	Weight
Fitting	Learning
Test data performance	Generalization
Regression/ Classification	Supervised Learning
Density Estimation/Clustering	Unsupervised Learning

So? What's New?

- Often, we measure the world and collect the data with carefully designed experiments, then work for years to learn plausible models from small datasets. The ways to select the key features and to explain the result of analysis rely heavily on domain experts, which sometimes, unfortunately, results in subjective and questionable interpretations of the world.
- In the age of Big Data, however, we start with large datasets with many features, and use statistical learning algorithms to find better models. Instead of manually doing inefficient & costly feature engineering, ***we tend to let machines automatically discover different abstract representations of data that best predict the outcome and describe the characteristics of the data***—a practice of [*Feature Learning*](#).

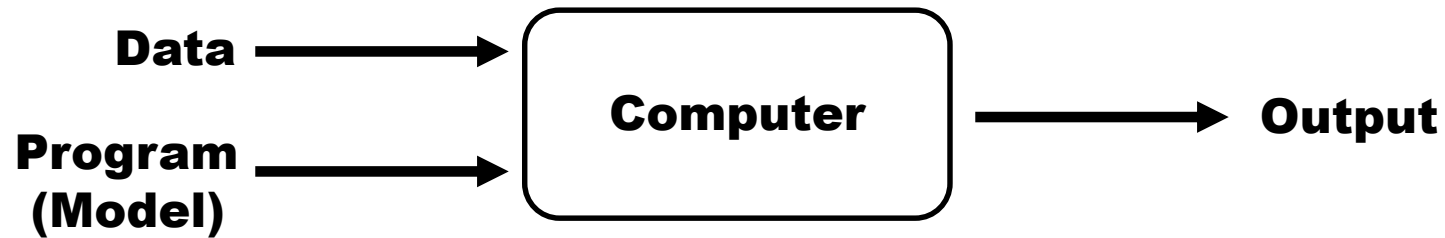


*"...if all you have is a hammer,
everything looks like a nail..."*

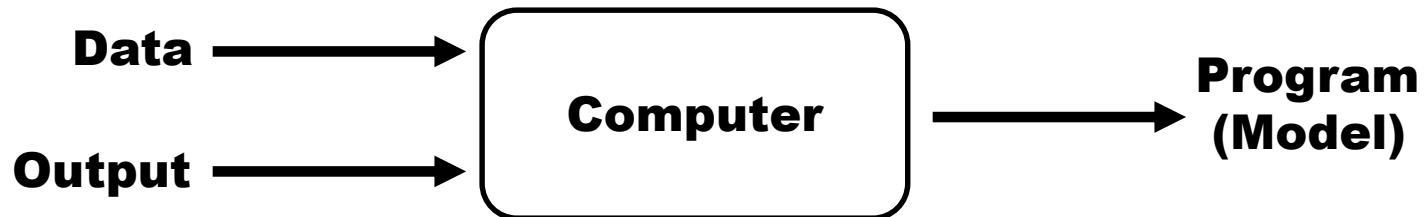
— *Abraham H. Maslow*

Why "Machine Learning"?

- Computer Programming



- Machine Learning



Group Discussion #2

- 從程式設計(Computer Programming)到機器學習(Machine Learning), 是否可以認為是種典範轉移(Paradigm Shift)? 為什麼?

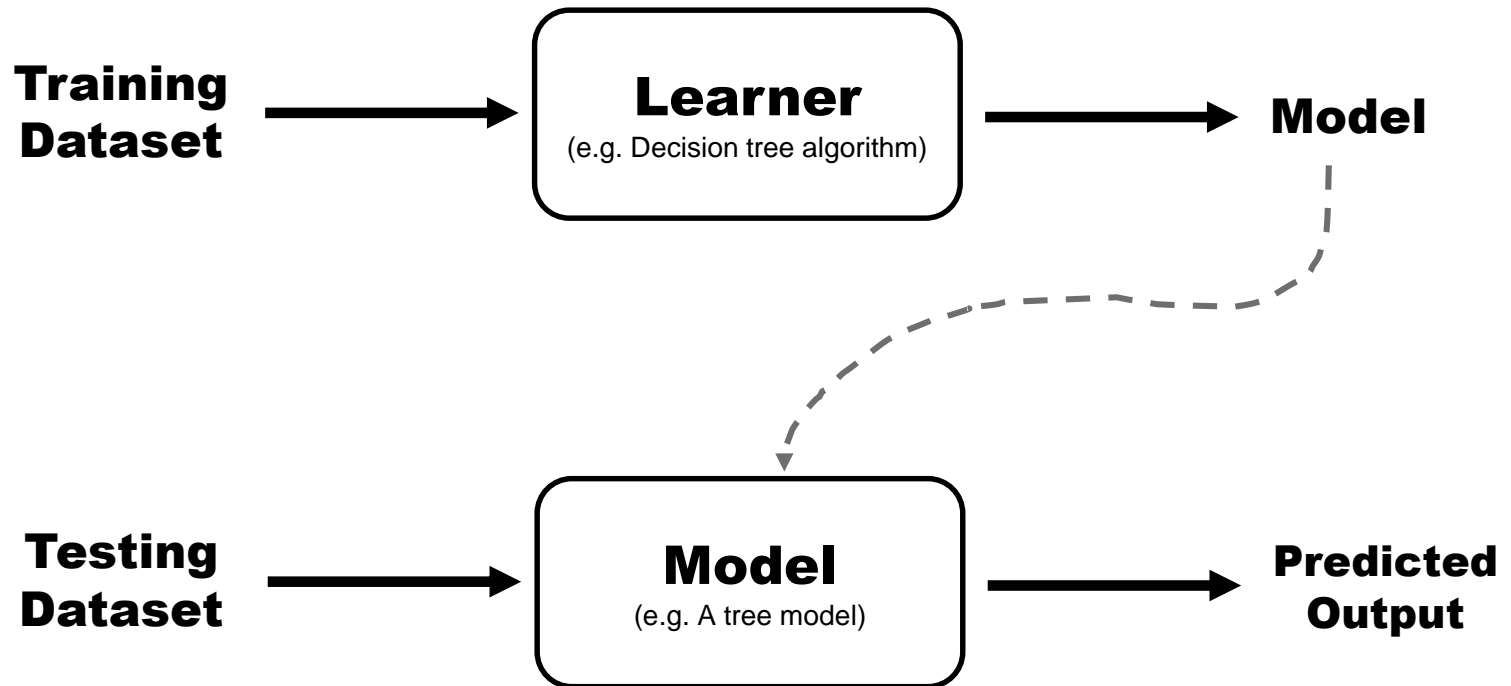


How Does Machine Learn?

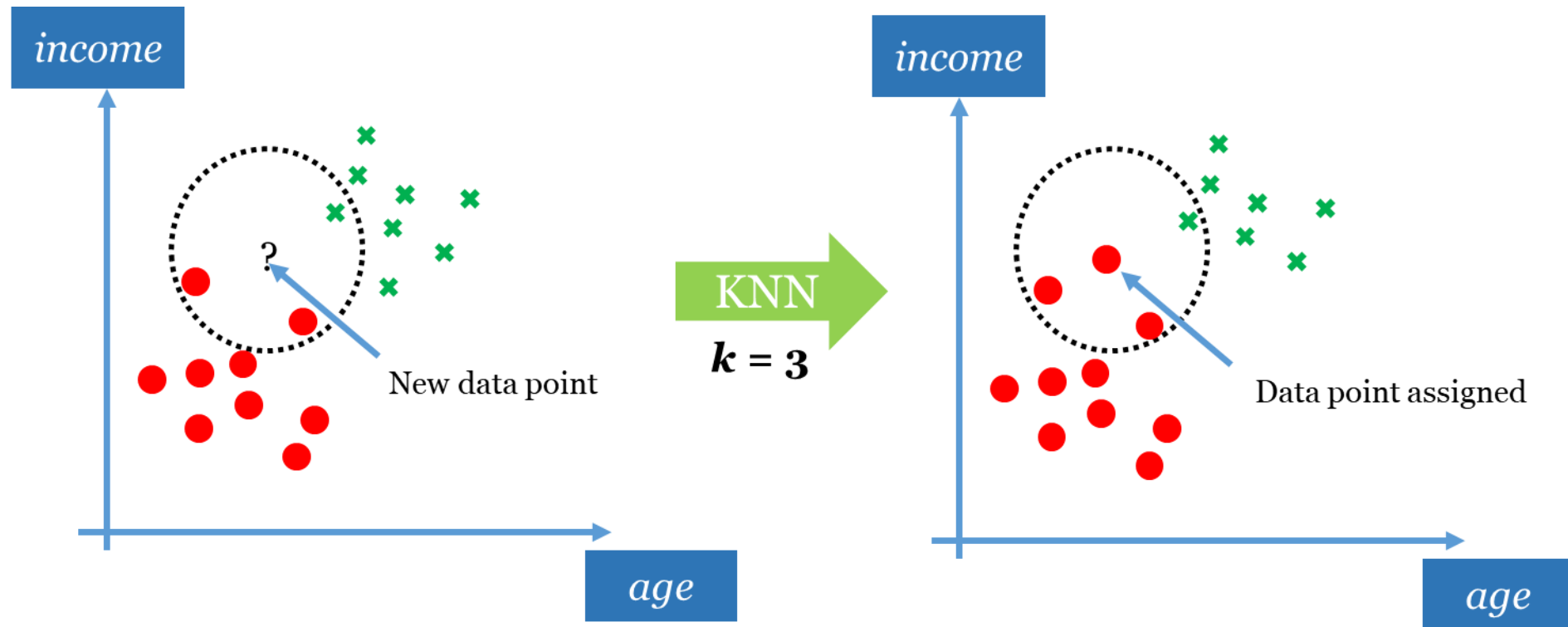
Learning = Representation + Evaluation + Optimization

Representation	Evaluation	Optimization
Instances <i>K</i> -nearest neighbor Support vector machines Hyperplanes Naive Bayes Logistic regression Decision trees Sets of rules Propositional rules Logic programs Neural networks Graphical models Bayesian networks Conditional random fields	Accuracy/Error rate Precision and recall Squared error Likelihood Posterior probability Information gain K-L divergence Cost/Utility Margin	Combinatorial optimization Greedy search Beam search Branch-and-bound Continuous optimization Unconstrained Gradient descent Conjugate gradient Quasi-Newton methods Constrained Linear programming Quadratic programming

Learner vs. Model

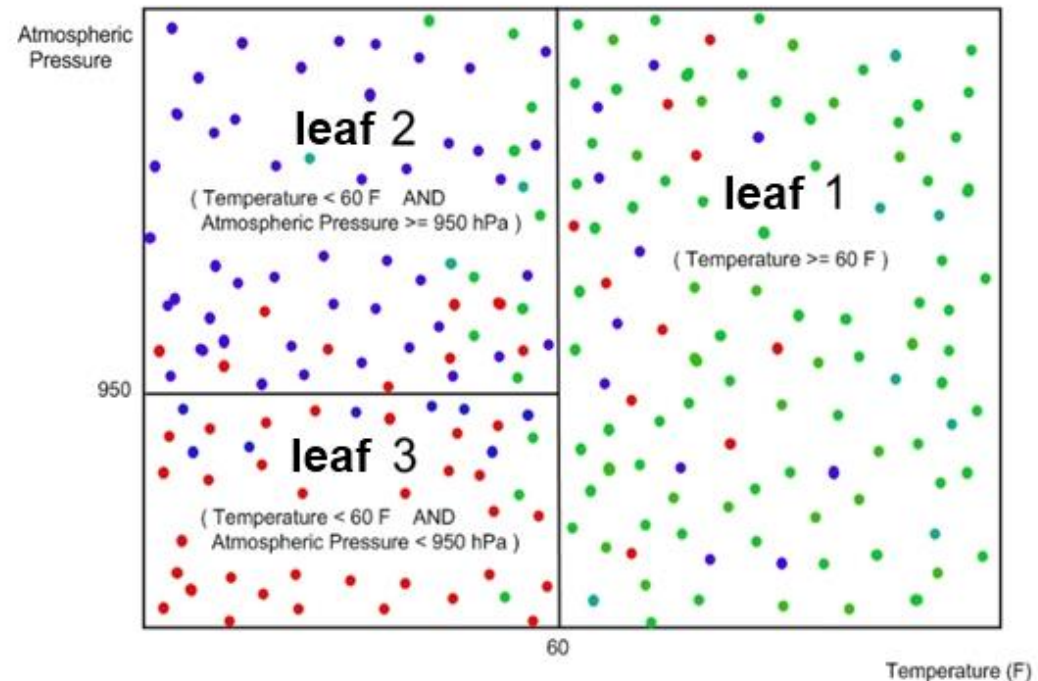
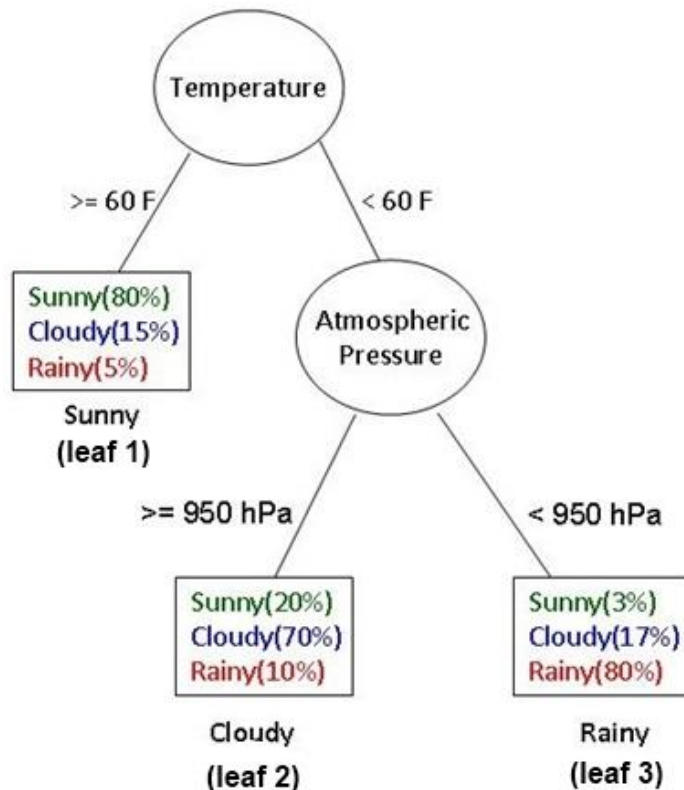


k-Nearest Neighbors—A Simple Learner



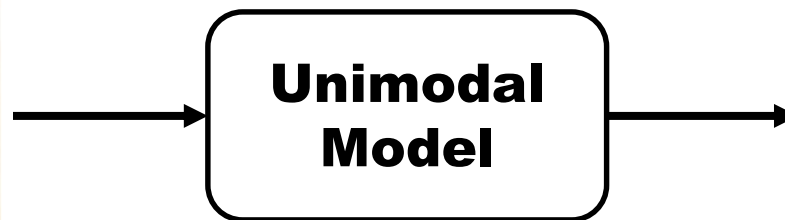
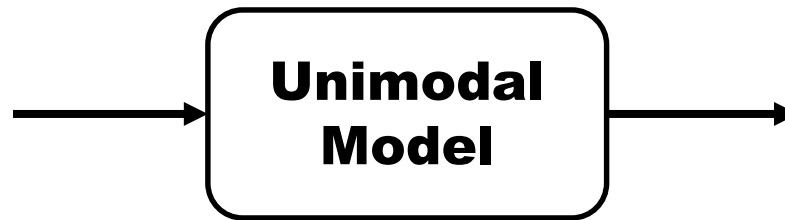
Classification Tree—A Flexible Learner

- The tree divides the predictor space into 3 regions/leaf nodes. These splits increase the purity of each region.



Learning Multi-modal Model

"Sushi"



Ingredients

sushi rice
salmon
avocado
cream cheese
nori

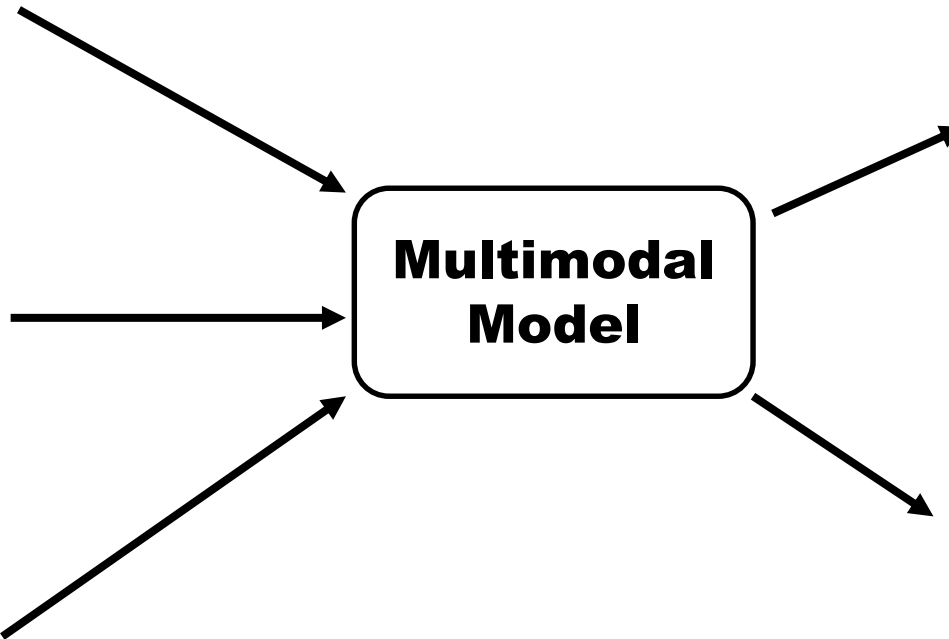
Learning Multi-modal Model_(cont.)

"Sushi"



Ingredients

sushi rice
salmon
avocado
cream cheese
nori







\$15

Instructions

1. Make 2 bowls of sushi rice.
2. Slice the salmon into 24 ultra-thin slices,
- ...

Bigger Data, Better Models?

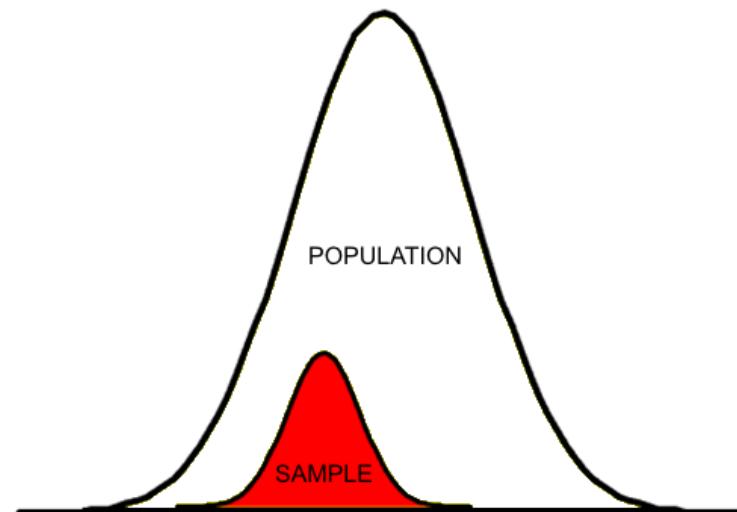
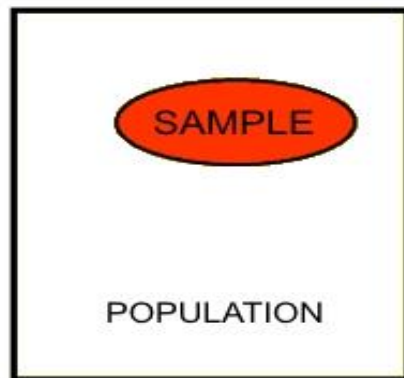
		Number of Variables (p)	
		Small p	Large p
Number of Observations(n)	Small n	 <p>Low accuracy and low precision</p>	 <p>Higher accuracy but low precision</p>
	Large n	 <p>Low accuracy but higher precision</p>	 <p>Higher accuracy and higher precision</p>

Bigger Data, Better Models? (cont.)

		Number of Variables (p)	
		Small p	Large p
Number of Observations(n)	Small n	<ul style="list-style-type: none"> ✗ Our models and inferences remain questionable. ✗ Our understanding of the world is limited by a few variables. 	<ul style="list-style-type: none"> ✓ Improve our ability to make inferences, as we can "control for" more variables and/or investigate more outcomes. ✗ How to select variables most relevant to the outcome? Are all variables required to make exogeneity plausible have been included?
	Large n	<ul style="list-style-type: none"> ✓ Increase the precision of estimates and the power of hypothesis tests. Many statistical learning techniques and/or estimation methods that require more observations could be used. ✗ We still cannot see the big picture of the world 	<ul style="list-style-type: none"> ✓ Creates numerous possibilities for descriptions, explorations, and inferences of the world of interest. ✗ More chance to get "coincident connections" among variables due to randomness. ✗ Harder to identify key predictors or factors.

Populations and Samples of Big Data

- "Big Data" doesn't mean the concepts of sampling is no longer important—simply, we would just have more information to understand the world.
- DO NOT believe those hype of Big Data that implies **N=ALL**. We would never observe everything (again, we're not the God!).



Populations and Samples of Big Data(cont.)

- Note that *biases are still there*, as we are making *assumptions* about our data (e.g. the underlying process that generated the data).
- Different interpretations of contexts may result in creating different models. Back in the days before the Big Data, we often look at the data, make assumptions, and then create parametric models for complex mechanism devised by the nature. Unfortunately, again, we would never find the perfect model, and conclusions drawn from our models remain questionable.
- Now, we have *more information* in the age of Big Data. When you analyze your data, use your imagination. Be careful and open-minded. Embrace different statistical learning techniques. And don't be afraid to make mistakes!

P-values, the ~~hunger~~ hacking! games



Source: <http://www.peaya.com/peaya.php?comicsid=1013>

peaya.com

hacking! *P-values, the ~~hunger~~ games*(cont.)

- A [recent article](#) about the p-values reports that "*...half of all published psychology papers that use null-hypothesis significance testing contained at least one p-value that was inconsistent with its test statistic and degrees of freedom. One in eight papers contained a grossly inconsistent p-value that may have affected the statistical conclusion...*".
- The p-values has been misused for years. The most notorious issue is that it has become a criterion for whether work is publishable. Some researchers have been hunting for "p-values < 0.05" so that they can claim their findings are "statistically significant". It has even worsen in the age of Big Data, as we can make the p-values as small as we want by increasing the sample size and/or by using some statistical tests sensitive to trivially small effects.
- Although the p-values should be considered outdated, it does not mean that we should skip p-values or shouldn't have large samples. Instead, when we interpret our findings, we must understand that the significance depends on the effect size and test sensitivity, and the "statistically significant" may not be meaningful. Also remember to check out [ASA Statement on Statistical Significance and P-values](#).

ASA Statement on Statistical Significance and P-values

- 1) "P-values can indicate how incompatible the data are with a specified statistical model."
- 2) "P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone."
- 3) "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold. "
- 4) "Proper inference requires full reporting and transparency."
- 5) "A p-value, or statistical significance, does not measure the size of an effect or the importance of a result."
- 6) "By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis."

P-values in the age of Big Data

- So, if the p-values is "outdated", how to evaluate our models? My answer to it is that it simply means using p-values cautiously. Very often, we get almost everything statistically significant when learning models from massive datasets. Instead of just using p-values to select and evaluate models, we should also **consider the accuracy of prediction** (e.g. accuracy for classification problems and mean squared error for regression problems), variable importance (different learning algorithms have different importance evaluations), and model interpretability.
- Researchers have also proposed other possible solutions. For example, we can simplify models by finding sparse representations of the input data; learn simple [*surrogate models*](#) (e.g. linear models) with the outcomes from complex models; impose some regularizations/constraints on model learning may also help; or, just use those "big data-ready" learning algorithms, such as [*Elastic Net*](#) instead of typical generalized linear models. Again, **getting "p-value < 0.05" simply means it's worth of a second look. It is just the beginning of your research, not the goal of it.**

P-values in the age of Big Data (cont.)

—Hypothesis Testing of Big Data-高雄捷運票價效益分析

- 2011年至2015年高雄捷運交易資料
- 月票期間(100/1/1~104/3/8)與無月票期間(104/3/8~104/12/31)
- 旅客有無購買月票(同張票卡曾設定月票者)，如何影響其搭乘行為？

Station ID	Station	月票期間日運量
15	高雄車站	14678
10	左營	13962
12	巨蛋	12966
18	三多商圈	11634
17	中央公園	9646
16	美麗島	7088
24	小港	6608
14	後驛	5189
20	凱旋	4870
13	凹子底	4731

P-values in the age of Big Data (cont.)

—Hypothesis Testing of Big Data-高雄捷運票價效益分析

- 不購買月票仍繼續搭乘捷運者，與購買月票期間相比：

1. 普卡30日搭乘金額超過月票售價(999元)250元以上？
2. 學生卡30日搭乘金額超過月票售價(799元)250元以上？

- 問題

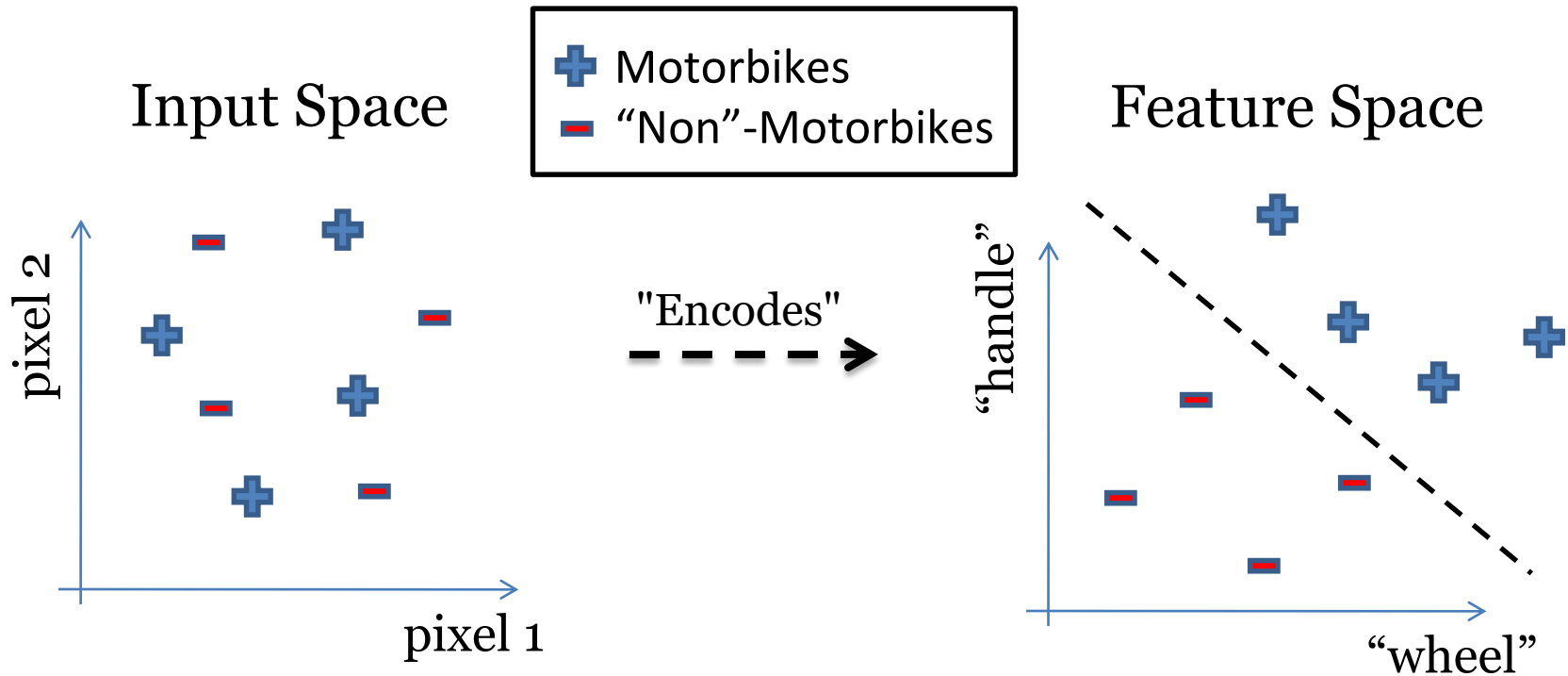
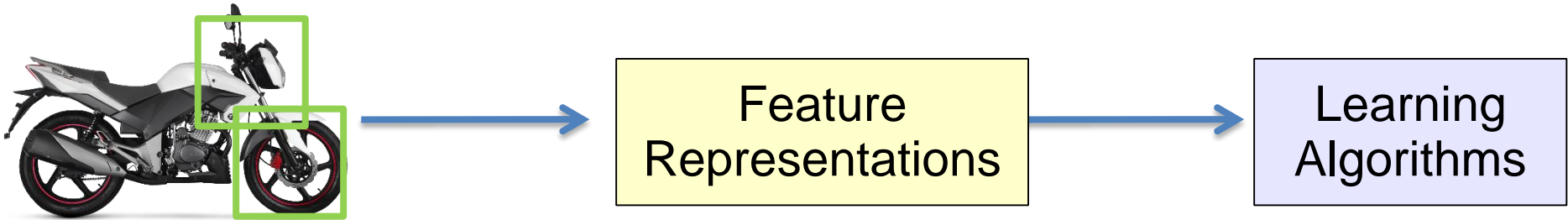
- 大樣本(big n) 造成 p-values 微小差異就變的很小？

在2014年曾設定為月票，而且在無月票期間仍繼續使用的票卡，於2014 年共有8,771,444筆搭乘紀錄，而在無月票期間共有4,621,623筆搭乘紀錄。

- 30日搭乘金額檢定後不符合常態分佈，使用無母數Wilcoxon Signed Rank Test的單尾檢定？

以重抽樣(resampling)的方法抽相對較小的樣本多次，來估計 p-value可能的值，或是以FDR or q-value 來取代 p-value 來推論。

Deeper Features, Better Models?

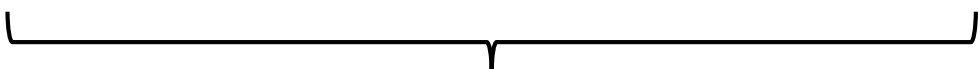


Feature Engineering Is (Still) The Key

What we think we're doing...

$$X \longrightarrow f(X) \longrightarrow \hat{y}$$

What we're actually doing...

$$X \longrightarrow f^1(X_1) \longrightarrow f^2(X_2) \longrightarrow \cdots \longrightarrow f^n(X_n) \longrightarrow \hat{y}$$


Manually—*Feature Engineering*

Automatically—*Feature Learning*

Real-world Feature Engineering

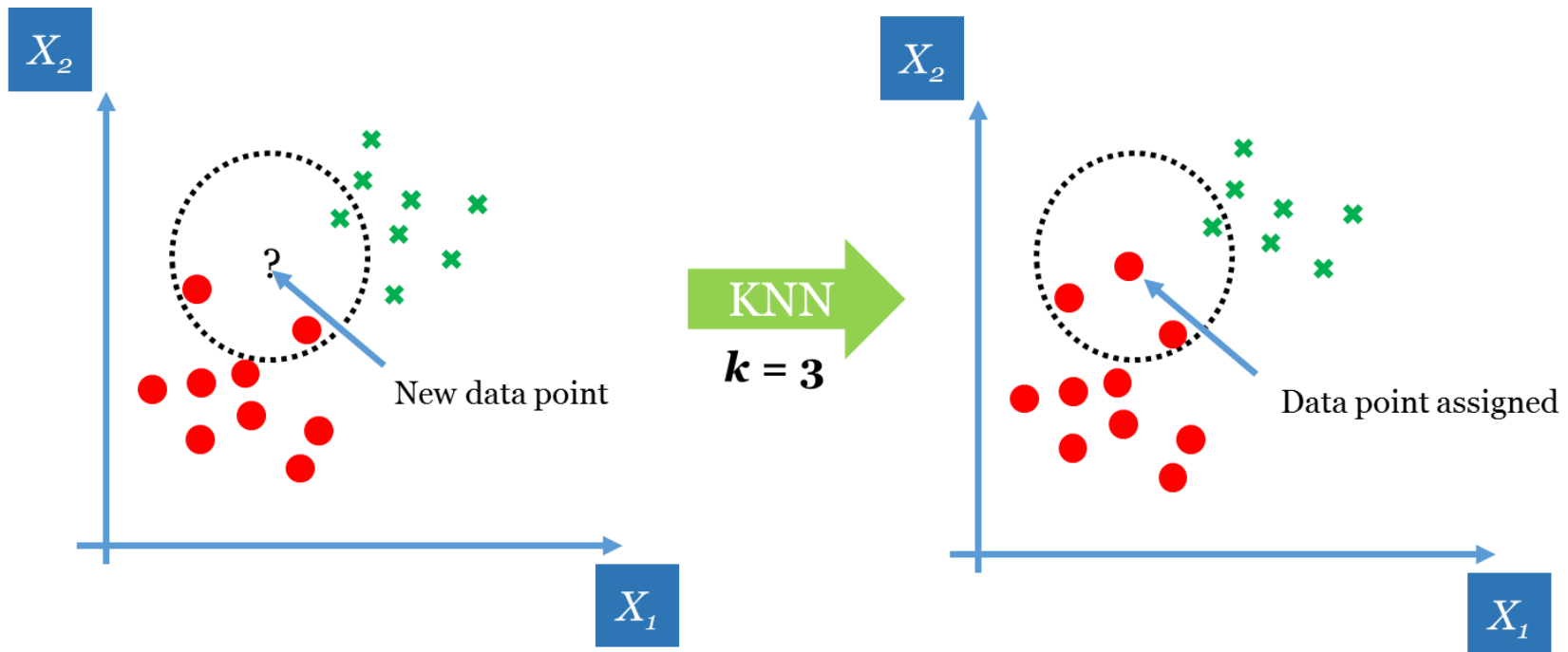
—Scaling & Transformation

- Some statistical tests and learning algorithms may be "fooled" by characteristics of features. For instance, **those instance-based techniques**, such as *k-Nearest Neighbors* (KNN) and *Principal Component Analysis* (PCA), **are sensitive to the ranges of continuous features**. They tend to give more weights to those features that exhibit larger variance and thus perform poorly.
- A quick, but not the best, solution is to make sure features are on the same (expected) scales, or just use scale-invariant techniques (e.g. tree-based models). Surely, most data analysis software allow us to create such data transformation/pre-processing "plans", which document how features would be re-scaled or transformed. Those plan objects can later be applied to training and testing (unseen) datasets to make sure all datasets are transformed into expected scales, ranges, or domains.

Real-world Feature Engineering (cont.)

—Scaling & Transformation

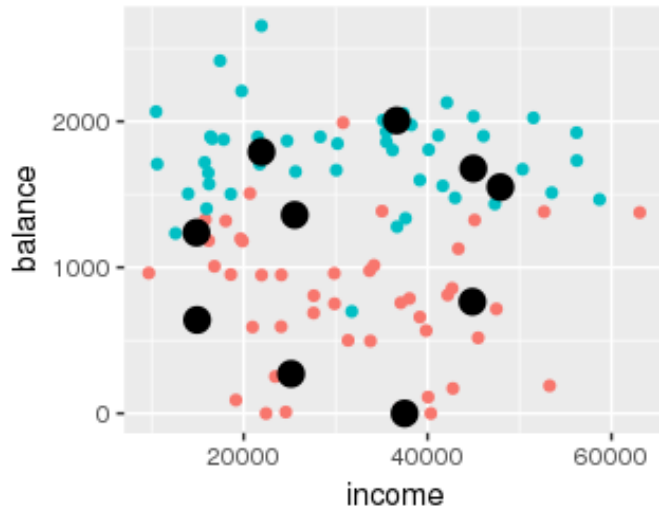
- Consider using KNN to solve classification problems. As the "nearest neighbors" is usually defined based on Euclidean distance, features used in computing the distances (X_1 and X_2 in below example) must be on the same scale.



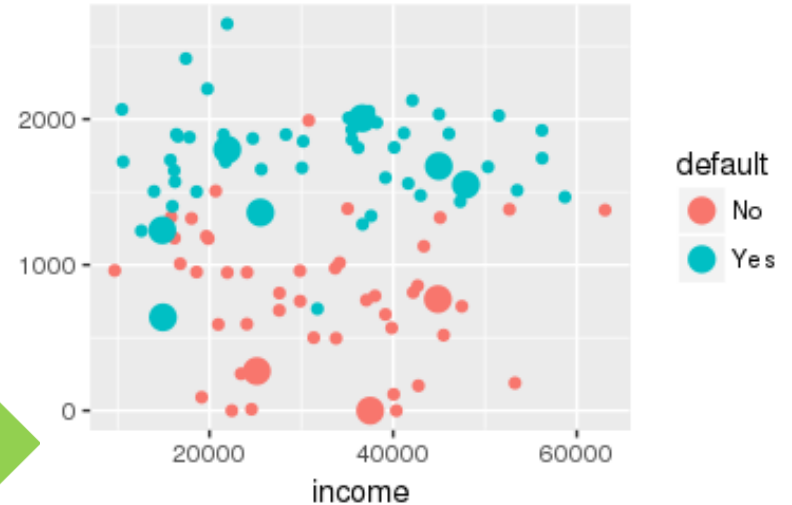
Real-world Feature Engineering (cont.)

—Scaling & Transformation

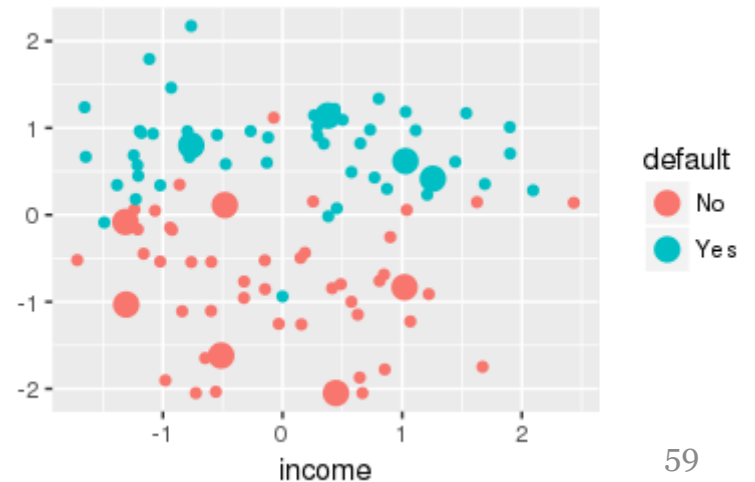
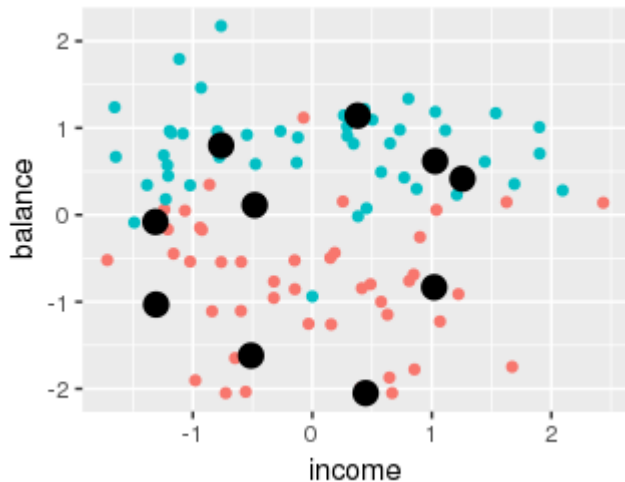
Unscaled



KNN
 $k = 3$



Scaled



Group Discussion #3

- 數據的特徵工程(Feature Engineering) 目的為何?
它是必要的嗎?



The Two Cultures of Modeling

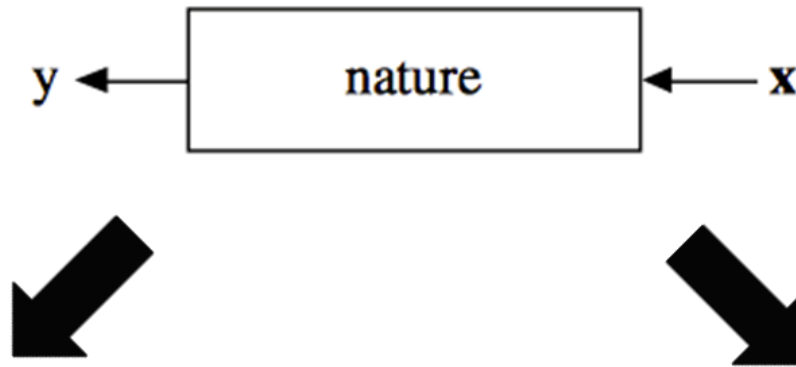


"...the focus in the statistical community on data models has led to irrelevant theory and questionable scientific conclusions..."

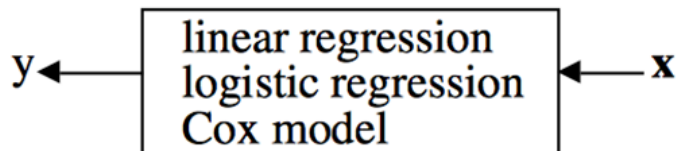
— [Leo Breiman](#)

- Remember to check out the article "[Statistical modeling: The two cultures](#)".

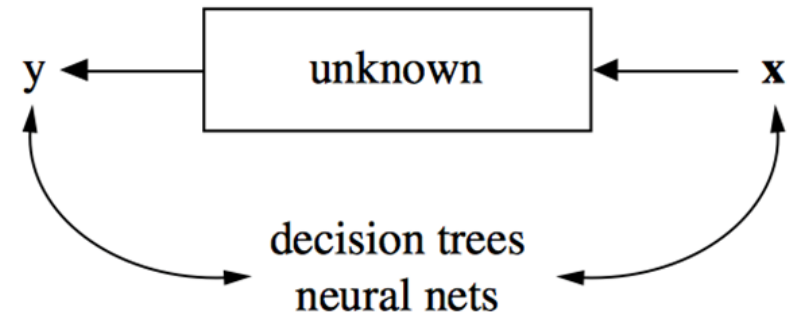
The Two Cultures of Modeling (cont.)



The Data Modeling



Algorithmic Modeling



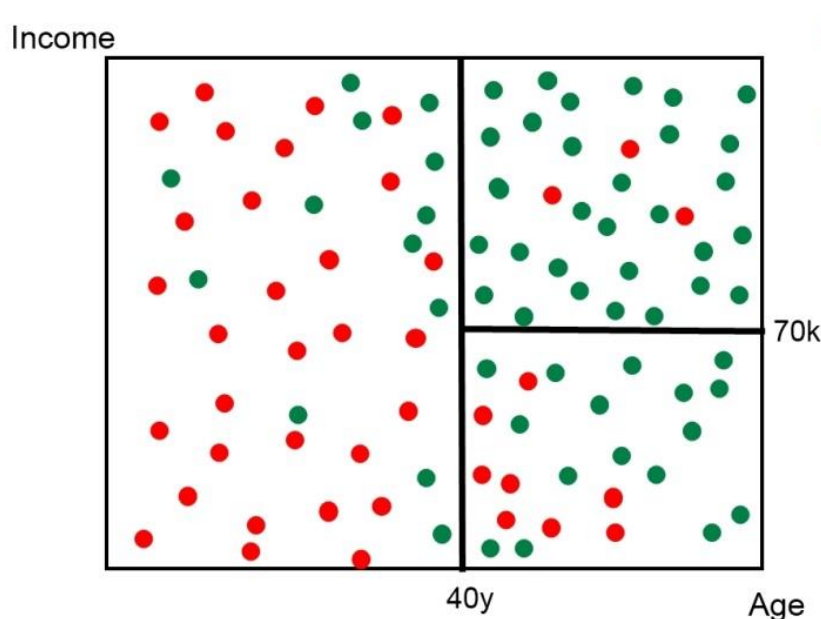
The Two Cultures of Modeling (cont.)

- Breiman and some statisticians believe that it is questionable to characterize the nature's functions from the inputs to outputs with simple and unique models that make assumptions of data generation processes. They suggest that ***we should instead develop models that account for observed data well and generalize well to unseen test datasets.***
- Now, such *Statistical Machine Learning* techniques are widely used and are proved successful in many different fields. However, some of these models & algorithms, such as Artificial Neural Networks and other ensemble learning techniques, are often complex and harder to interpret—just like the complexity of our worlds of interests. They are sometimes called "black-box" models.

The Two Cultures of Modeling (cont.)

- There does have data scientists who are trying to [open the black box](#) (e.g. Rule Induction & Extraction), which is beyond the scope of this class. Still, how to choose modeling practices (cultures) depends on the goal of your data analytics projects. Again, a good analytics practice is to always focus on the goal and "see" the data from different angles (i.e. to identify better representations of features). ***Different representations of the data may capture hidden patterns that could boost the model accuracy and interpretability.***
- In many real-world data applications, you might find that simple models are easy-to-understand but come with lower accuracy and rough findings, whereas complex models are hard-to-interpret but have relatively higher predictive power. Note that, however, ***there is no necessary connection between model accuracy and model interpretability/complexity.***

"To Explain?" or "To Predict?"

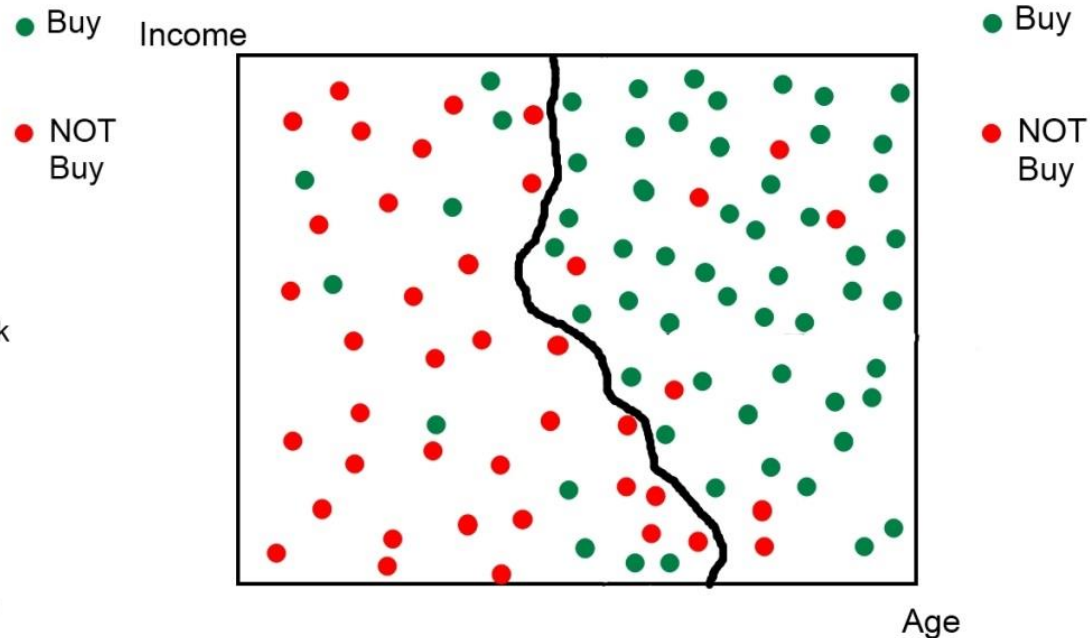


Misclassification Rate = 15% ??

Easy to understand the model
by those "rules":

- 1: IF Age >= 40y AND Income >= 70k
THEN Buy = True
- 2: IF Age >= 40y AND Income < 70k
THEN Buy = True
- 3: IF Age < 40y
THEN Buy = False

Simple!!



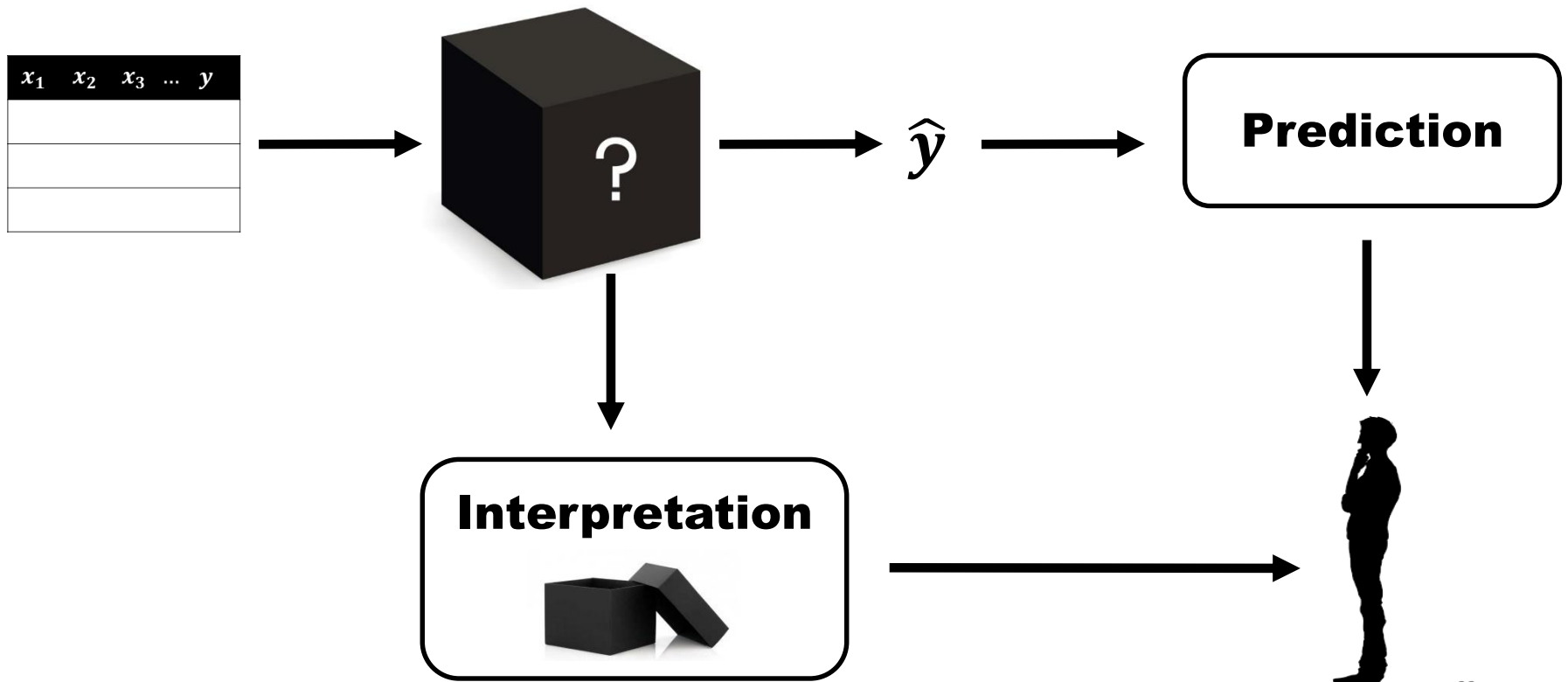
Misclassification Rate = 10% ?? (usually lower)

Rules ? What is that?

Obscure!!
Rule "extraction"?!

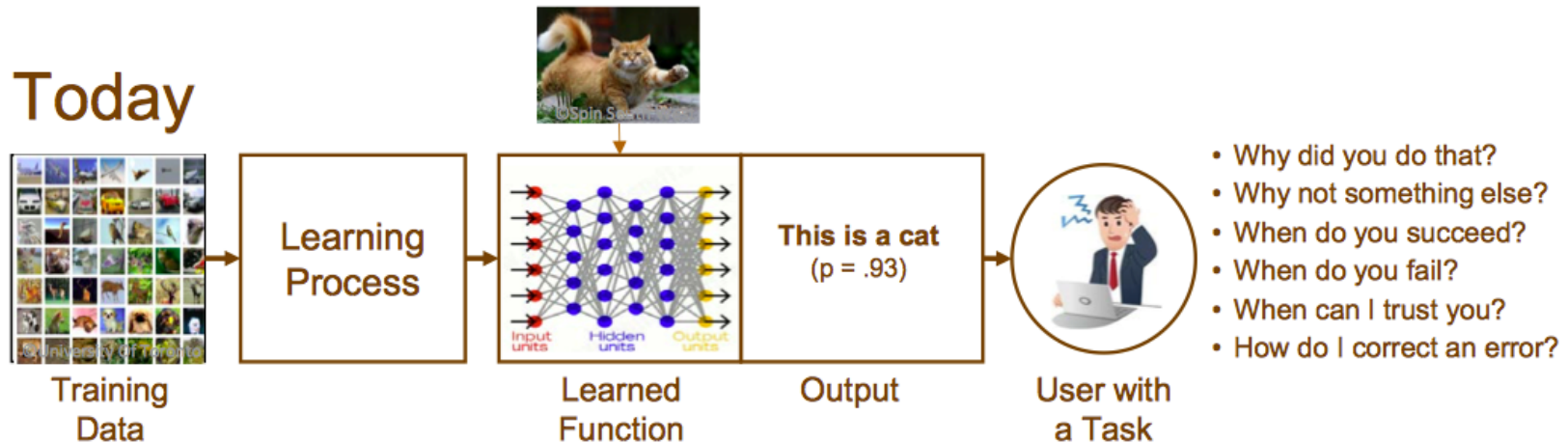
Machine Learning Interpretability

- The *accountability/interpretability* is the ability to explain or present in understandable terms to human.

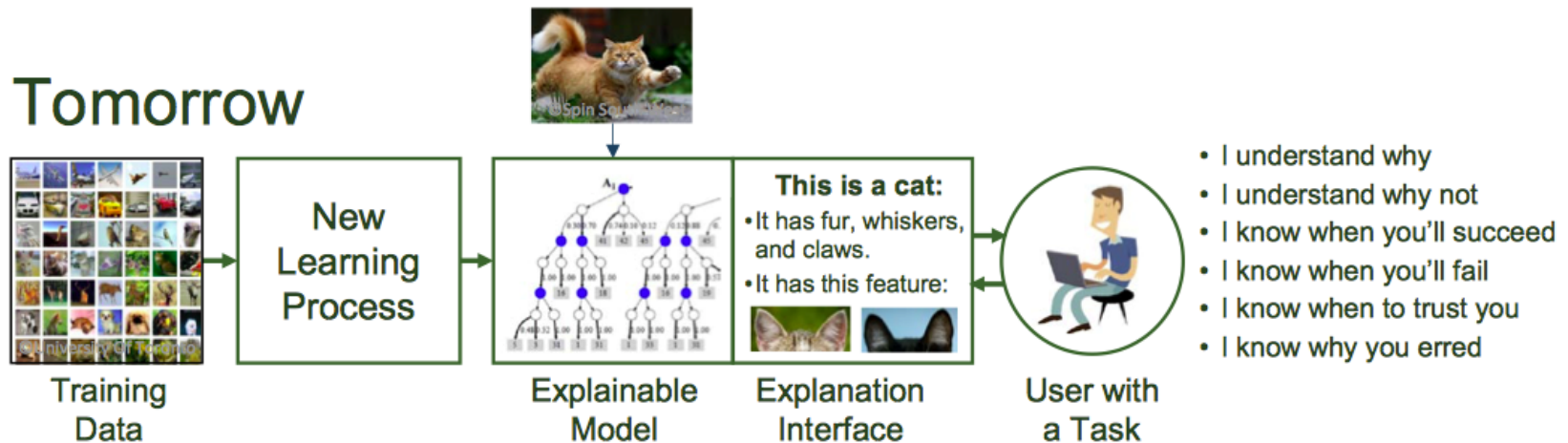


Towards Interpretable Machine Learning

Today



Tomorrow



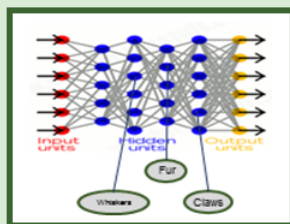
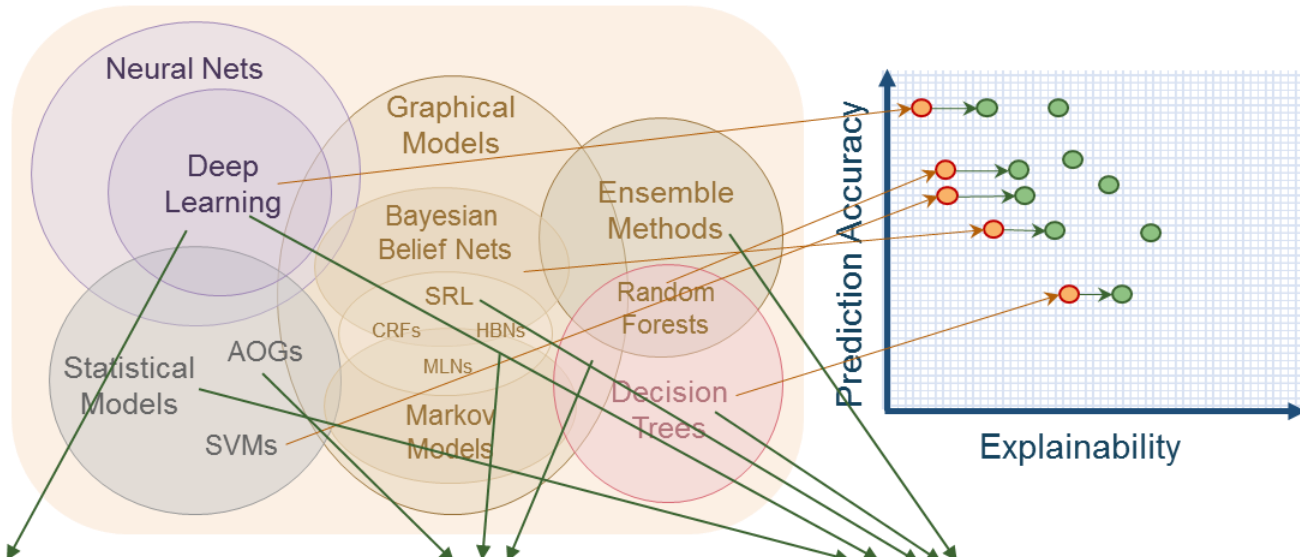
Towards Interpretable Machine Learning

—Performance vs. Explainability

New Approach

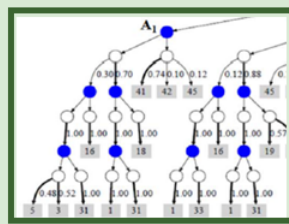
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



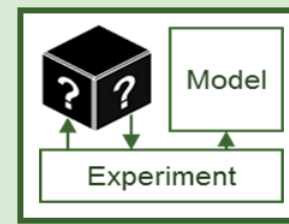
Deep Explanation

Modified deep learning techniques to learn explainable features



Interpretable Models

Techniques to learn more structured, interpretable, causal models

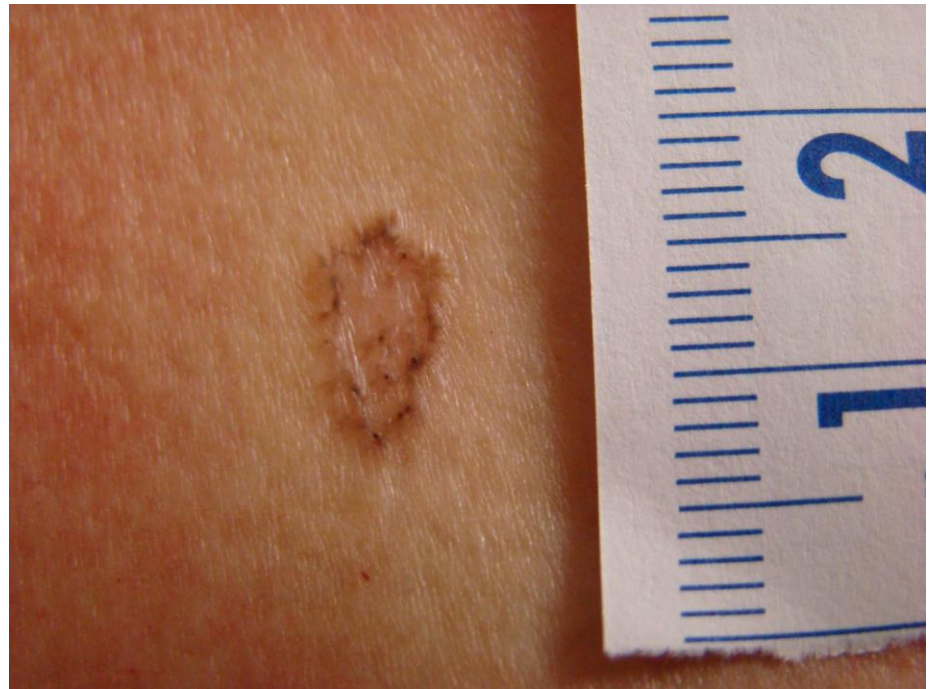


Model Induction

Techniques to infer an explainable model from any model as a black box

Towards Interpretable Machine Learning —A Big Problem

Skin cancer?

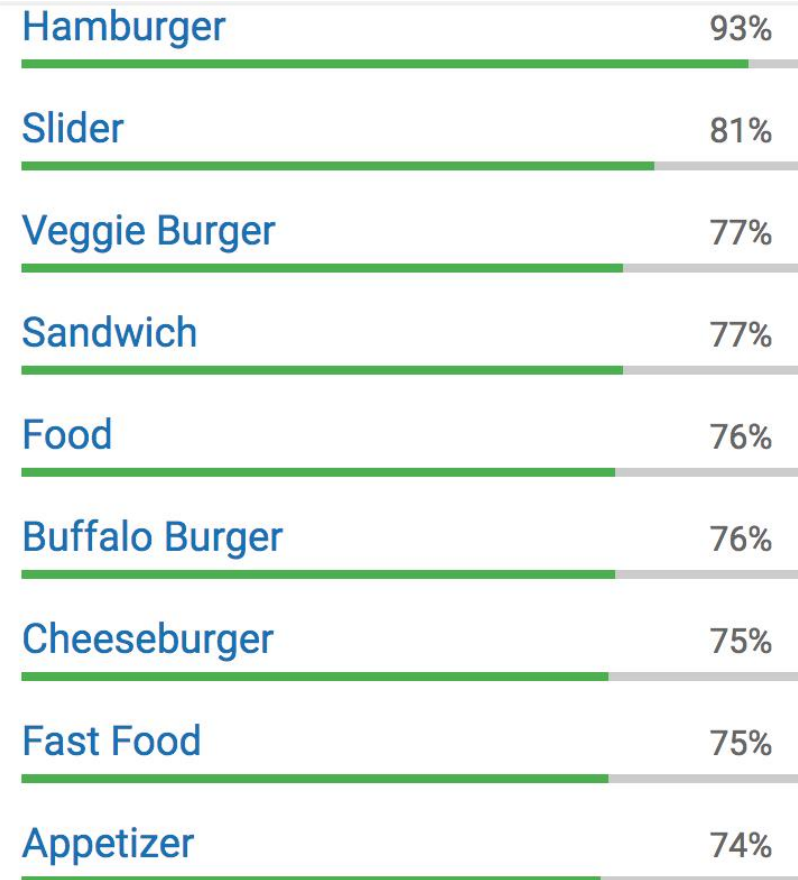


Source: <https://visualsonline.cancer.gov/details.cfm?imageid=9288>, "Malignant Melanoma", National Cancer Institute

Towards Interpretable Machine Learning (cont.)



1599px-Umami_Burger_hamburger.jpg



Source: https://en.wikipedia.org/wiki/File:Umami_Burger_hamburger.jpg (adapted), by Jun Seita, licensed under CC BY 2.0

Interpretable AI – An Example



Fried Chicken

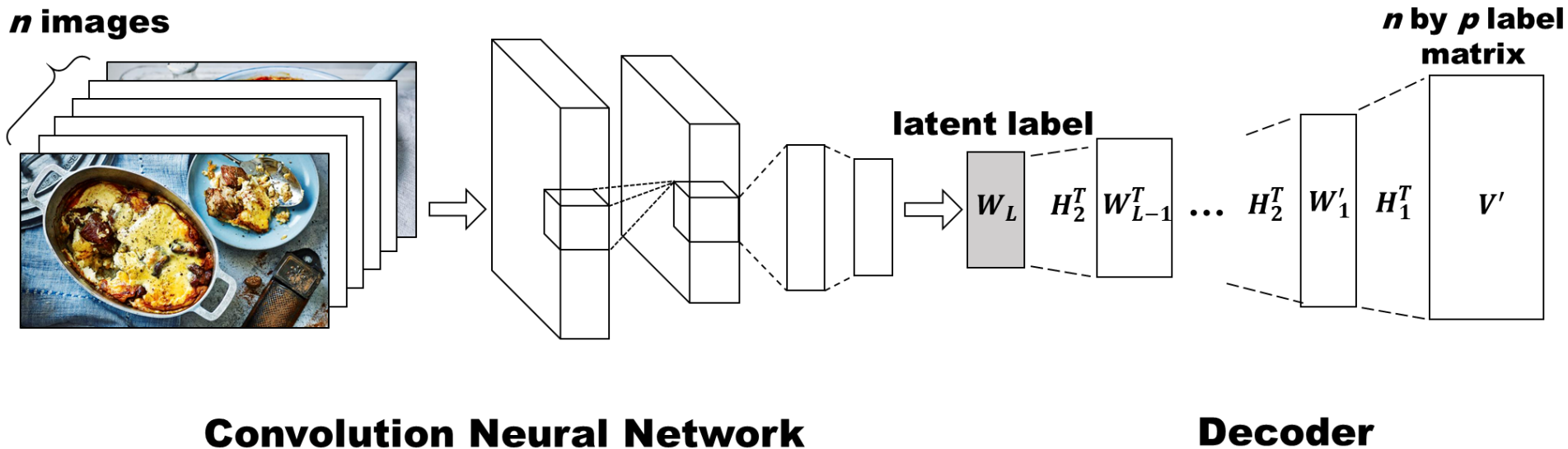
Actual Ingredient

chicken
buttermilk
plain flour
cornflour
oregano
chilli powder
sage
basil
marjoram
white pepper
salt
paprika
smoked paprika
onion
garlic
oil
milk
coleslaw

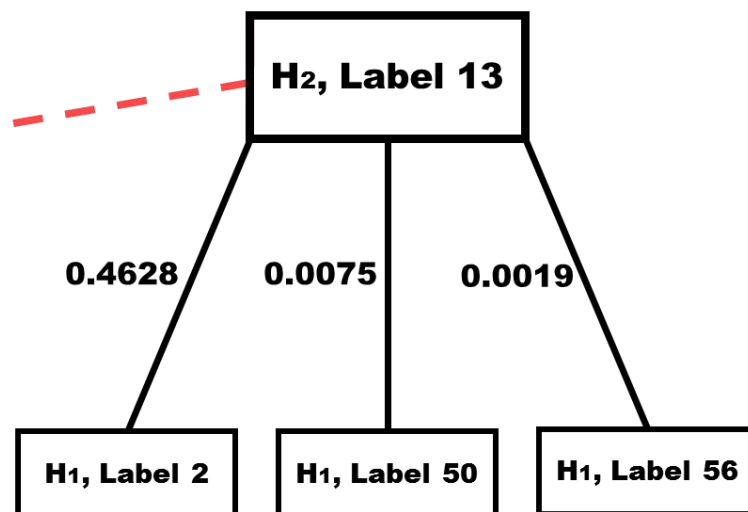
Predicted Ingredient

plain flour
buttermilk
oil
black pepper
egg
salt
onion
garlic
milk
lemon
olive oil
sugar
white wine
nut
stock
caster sugar
cream
vinegar
chicken
baking powder

Interpretable AI – An Example_(cont.)



Interpretable AI – An Example_(cont.)



H1, Label 2 : garlic, garlic bread, chicken, stock, silverside, lamb shank, onion, chicken stock, boar, field mushroom

H1, Label 50 : white wine, vinegar, red wine vinegar, mustard, white bread, dijon mustard, anchovy essence, white cabbage, turkey mince, egg white, white pepper

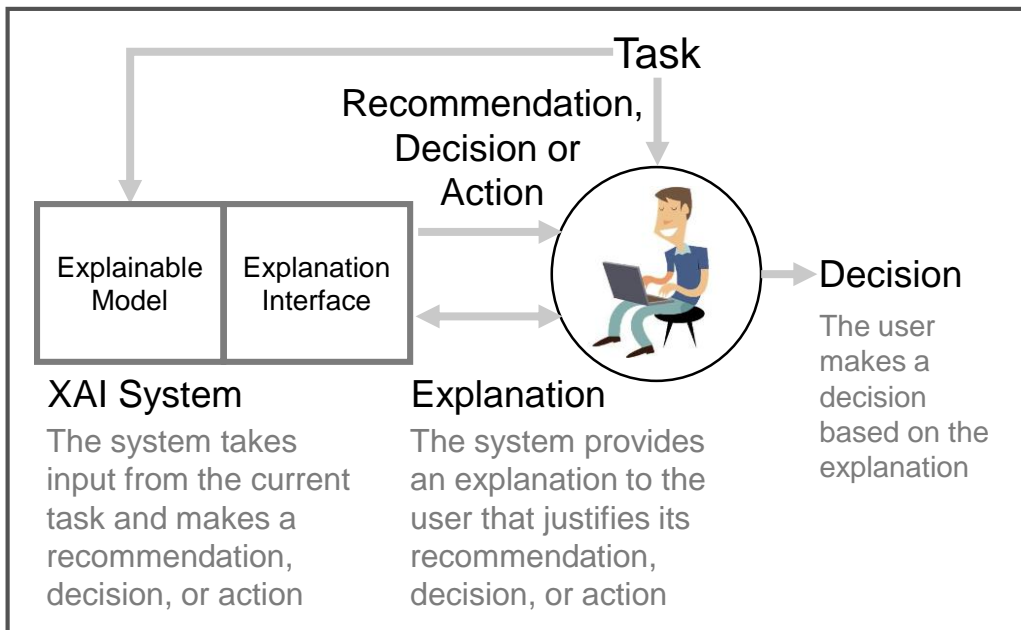
H1, Label 56 : white wine, vinegar, charlotte potato, bread sauce, double gloucester, Dover sole, smoked trout, anchovy essence, kohlrabi, pink fir apples

Measuring Interpretability

- Interpretability is NOT formally defined but is closely related to the development of more ***Fair, Accountable, Transparent, and Ethical (FATE)*** AI. However, there does have researchers who argue that it is impossible for AI to explain everything it does, as not all the intelligence are interpretable.
- The other researchers argue that the interpretability is "a means to engender trust", "a concept that algorithms can provide explanations", or "the ability to explain or to present to understandable terms to a human". For simplicity, here we define model interpretability as ***how algorithms arrive at a decision.***

Measuring Interpretability (cont.)

XAI Explanation Framework



Source: <https://www.darpa.mil/program/explainable-artificial-intelligence>,
"Explainable Artificial Intelligence (XAI)"

Interpretability Measurement

User Satisfaction

- Clarity of the explanation (user rating)
- Utility of the explanation (user rating)

Mental Model

- Understanding individual decisions
- Understanding the overall model
- Strength/weakness assessment
- 'What will it do' prediction
- 'How do I intervene' prediction

Task Performance

- Does the explanation improve the user's decision, task performance?
- Artificial decision tasks introduced to diagnose the user's understanding

Trust Assessment

- Appropriate future use and trust

Correctability (Prediction Accuracy)

- Identifying errors
- Correcting errors, Continuous training

EU General Data Protection Regulation (GDPR) & A Right to Explanation

- In the regulation of algorithms, particularly artificial intelligence and its subfield of machine learning, ***a right to explanation*** is a right to be given an explanation for an output of the algorithm.
- Such rights primarily refer to individual rights to be given an explanation for decisions that significantly affect an individual, particularly legally or financially.

Group Discussion #4

- 請舉例說明需要”可解釋的 AI 模型”的情境



The pursuit of “good model”

Q: So, what is good model? **A:** ○ Lower testing error
○ Better interpretation
○ Smaller & faster in applications

Model #1

IF gender = “male”
THEN buy = “yes”

Model #2

IF gender = “male” **AND**
income > 22k
THEN buy = “yes”

Model #3

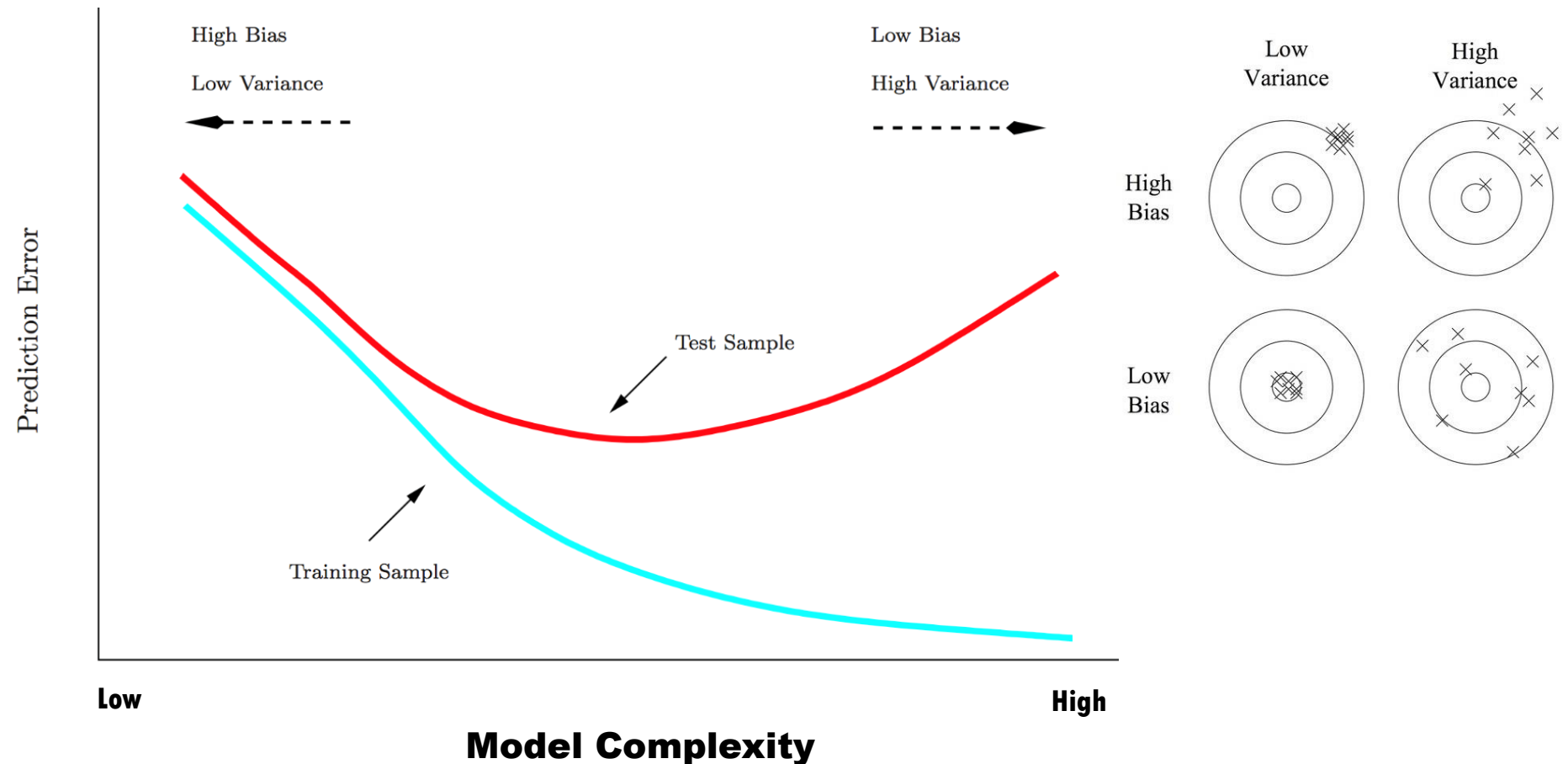
IF (gender = “male” **AND**
income > 22k) **OR**
(hasRichDad = “yes”)
THEN buy = “yes”

Low

High

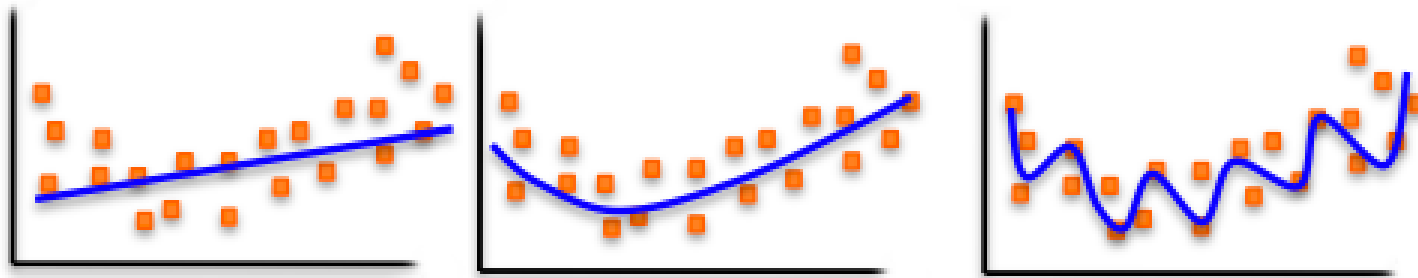
Model Complexity/Capacity

Model Complexity and Bias-Variance Tradeoff

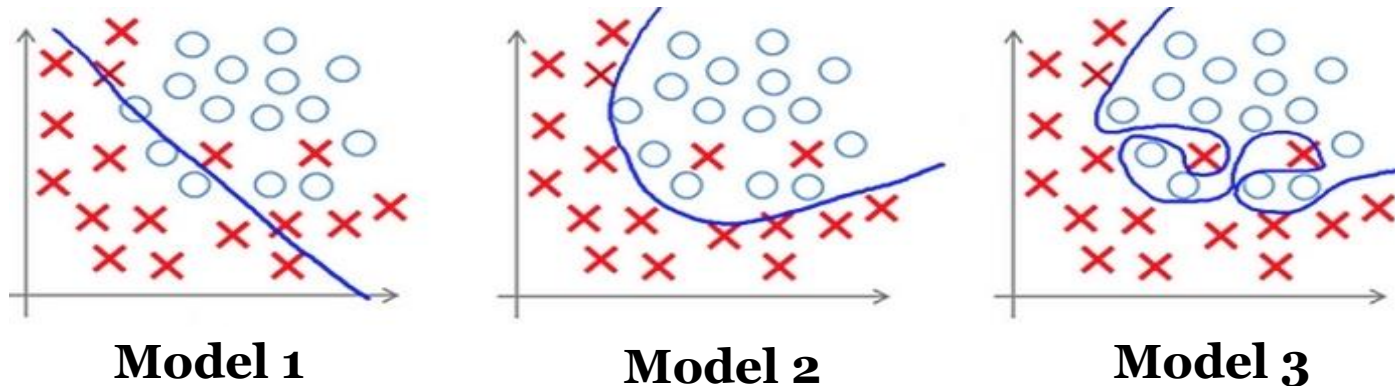


Underfitting and Overfitting

Regression:



Classification:



Learning Ensemble of Models

- To identify and learn the complex function $f(X)$ manually is very difficult. Researchers have found that we learn and combine many models—ensemble of models, either **deeper** or **diverser**, may help.

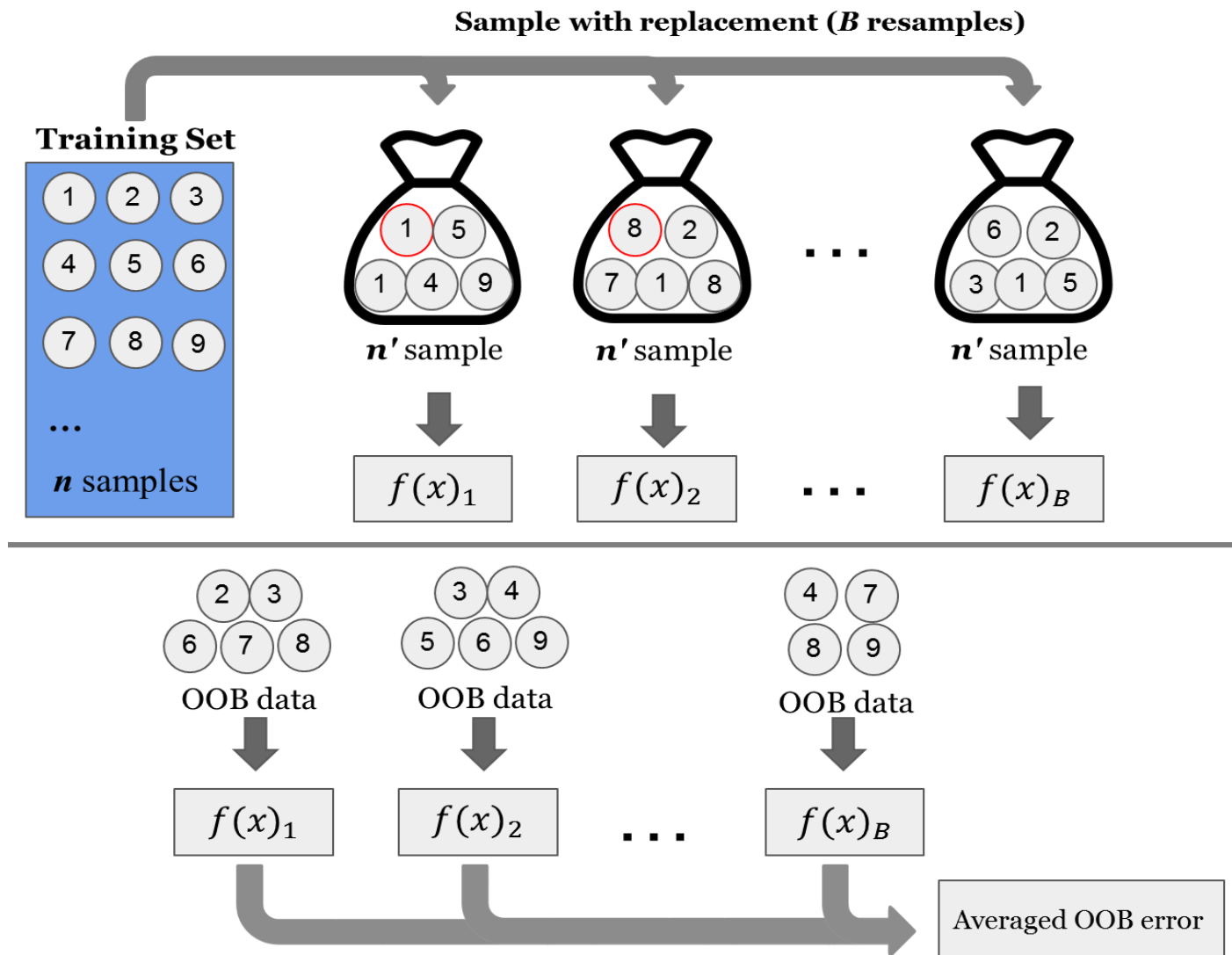
$$X \longrightarrow f^A(f^1(X), f^2(X), \dots, f^n(X),) \longrightarrow \hat{y}$$

$$X \longrightarrow f^4(f^3(f^2(f^1(x)))) \longrightarrow \hat{y}$$

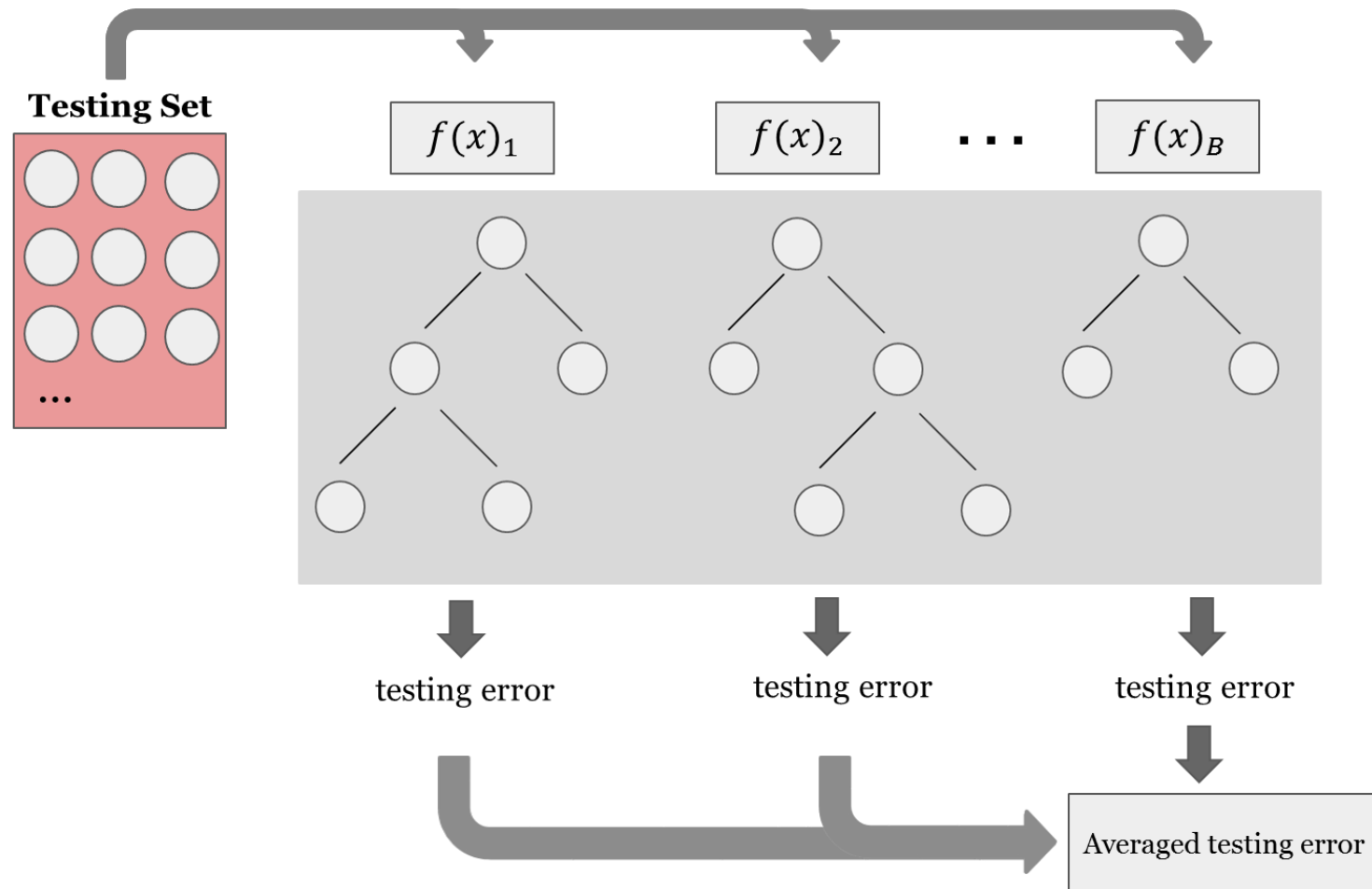
$$X \longrightarrow f^A(f^{11}(f^1(x)), (f^{22}(f^2(x)))) \longrightarrow \hat{y}$$

⋮

Bootstrap Aggregating—An Example



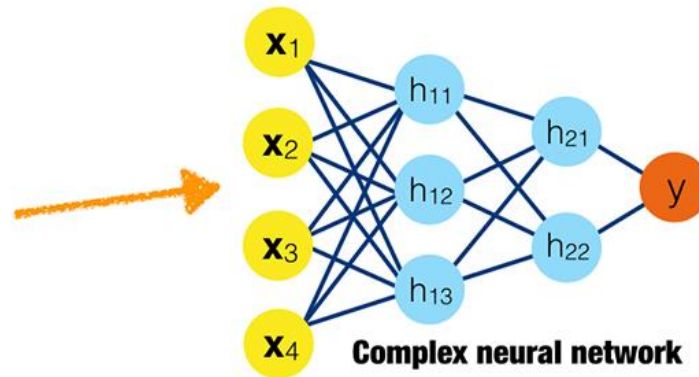
Bootstrap Aggregating—An Example(cont.)



Model Compression for Smaller and Faster models

—Surrogate Models

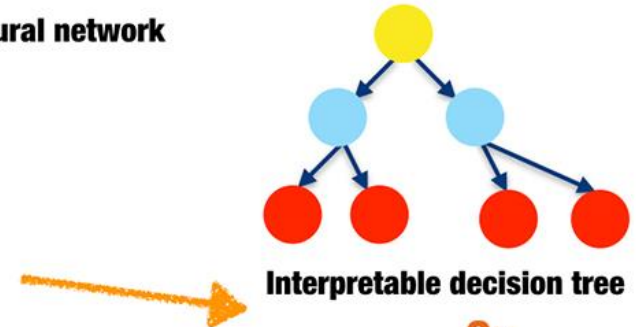
BAD	CUSTOMER_DTI	LOAN_PURPOSE	CHANNEL
0	0.18	MORT	7
1	0.42	HELOC	10
0	0.11	MORT	10
0	0.21	MORT	1



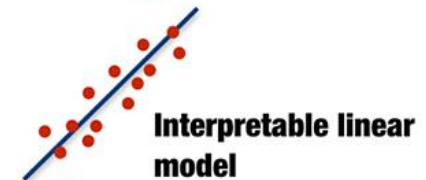
1. Train a complex machine learning model

BAD	PREDICTED_BAD	CUSTOMER_DTI	LOAN_PURPOSE	CHANNEL
0	0.47	0.18	MORT	7
1	0.82	0.42	HELOC	10
0	0.18	0.11	MORT	10
0	0.12	0.21	MORT	1

2. Train an interpretable model on the original inputs and the predicted target values of the complex model

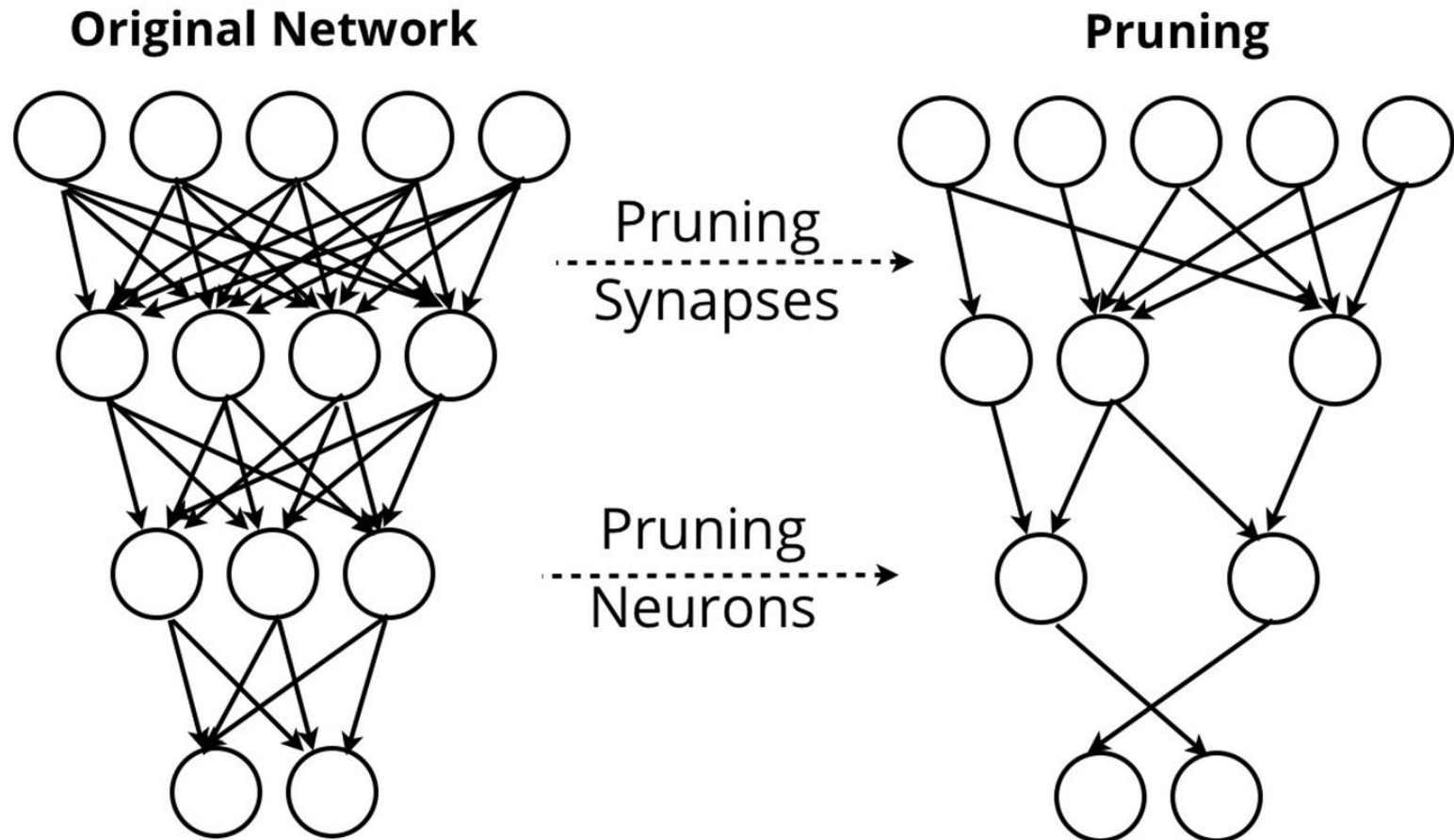


Or



Model Compression for Smaller and Faster models

—Neural Network Pruning





*"Essentially, all models are wrong,
but some are useful."*

— *George E. P. Box*

Group Discussion #5

- 對你來說，模型（model）是什麼？如何定義模型的好與壞？



Trends in Data Analytics

- ✓ The flood of "data lake"
- ✓ The rise of out-of-core learning algorithms
- ✓ The dawn of *fast scalable data applications*
- ✓ The use of *in-memory*, *in-database*, and *heterogeneous* computing
- ✓ The pursuit of interpretable analytics and explainable AI



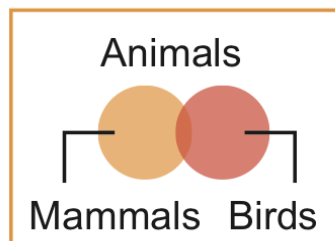
"AI is the New Electricity."
— Andrew Ng (吳恩達)

Getting started with AI

- Build a data pipelining infrastructure (Unified Data Analytics platform)
- Acquire data strategically with purposes, and always start small, quick, & “dirty”
- Promote pervasive automations, iteratively and incrementally

Towards “real” AI—The ML 5 Tribes

Symbolists



Use symbols, rules, and logic to represent knowledge and draw logical inference

Favored algorithm

Rules and decision trees

Bayesians

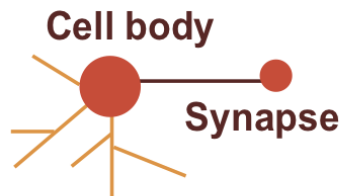


Assess the likelihood of occurrence for probabilistic inference

Favored algorithm

Naive Bayes or Markov

Connectionists



Recognize and generalize patterns dynamically with matrices of probabilistic, weighted neurons

Favored algorithm

Neural networks

Evolutionaries



Generate variations and then assess the fitness of each for a given purpose

Favored algorithm

Genetic programs

Analogizers



Optimize a function in light of constraints (“going as high as you can while staying on the road”)

Favored algorithm

Support vectors

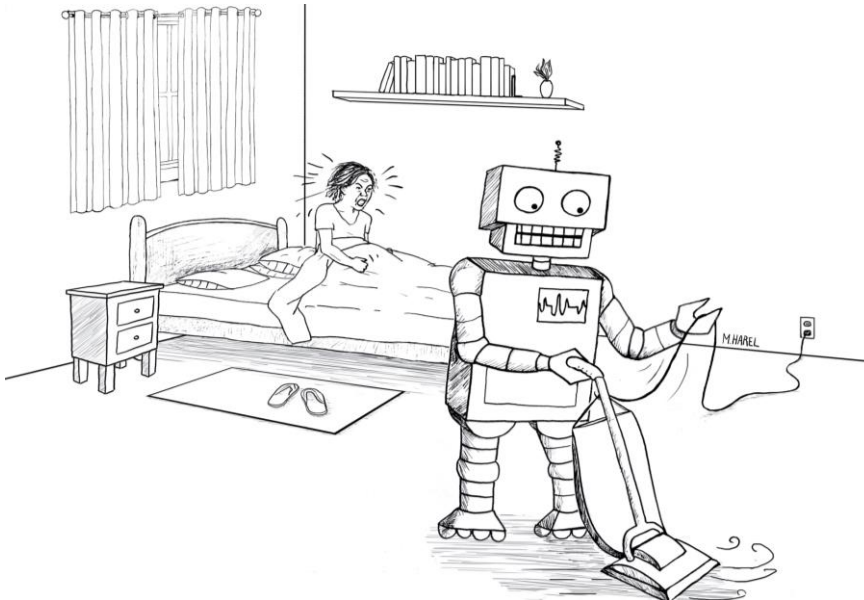
Towards “real” AI—Cognition of AI

- Perception (sensing, understanding...)

**Theoretically solved by
Connectionists?!**

- Reasoning (logic, probability, ...)

**Yes! we are counting on you,
Symbolists and Bayesians!**



Group Discussion #6

- 現今所謂的 AI 應用是否為大眾所認知的 AI? 如果不是, 有哪些問題還沒解決, 為什麼?



Better Infrastructures, Better Data and Better Models

- *“More data beats a cleverer algorithm.”*

P. Domingos

- *“More data beats clever algorithms, but better data beats more data.”*

P. Norvig

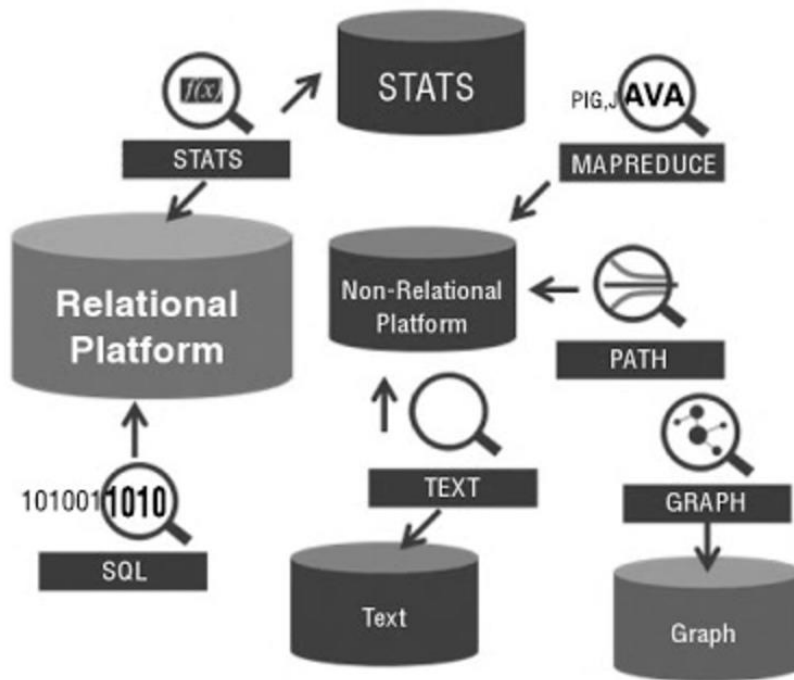
I’ll say ***“No better data infrastructures, no better data and cleverer algorithms”!***

The Age of Business Intelligence (BI)

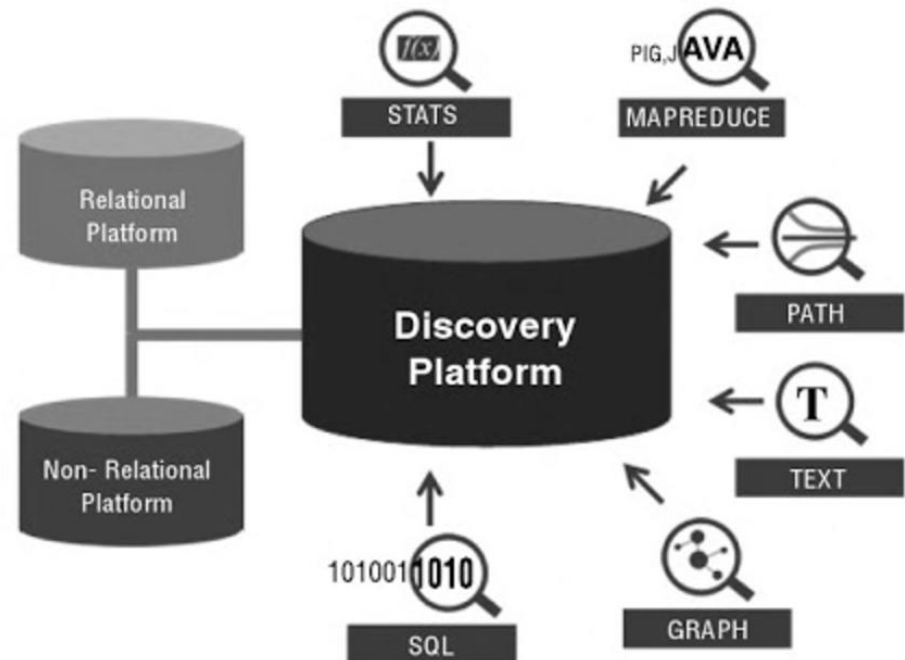


Towards Unified Analytics Platform

Traditional (Messy BI) Data Warehouse?



Discovery (Unified Analytics) Data warehouse + Data Lake?



From Business Intelligence to Unified Analytics

- **Collaboration**—Teams working in silos
- **Unification**—Data pipeline becomes more complex
- **Flexibility**—Lack of an agile AI ecosystem
- **Scalability**—When the infrastructure gets in the way

Teams working in silos



Great for Data, but not AI



Data Engineers



PYTORCH



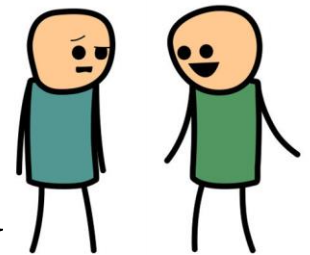
Great for AI, but not Data



Data Scientists

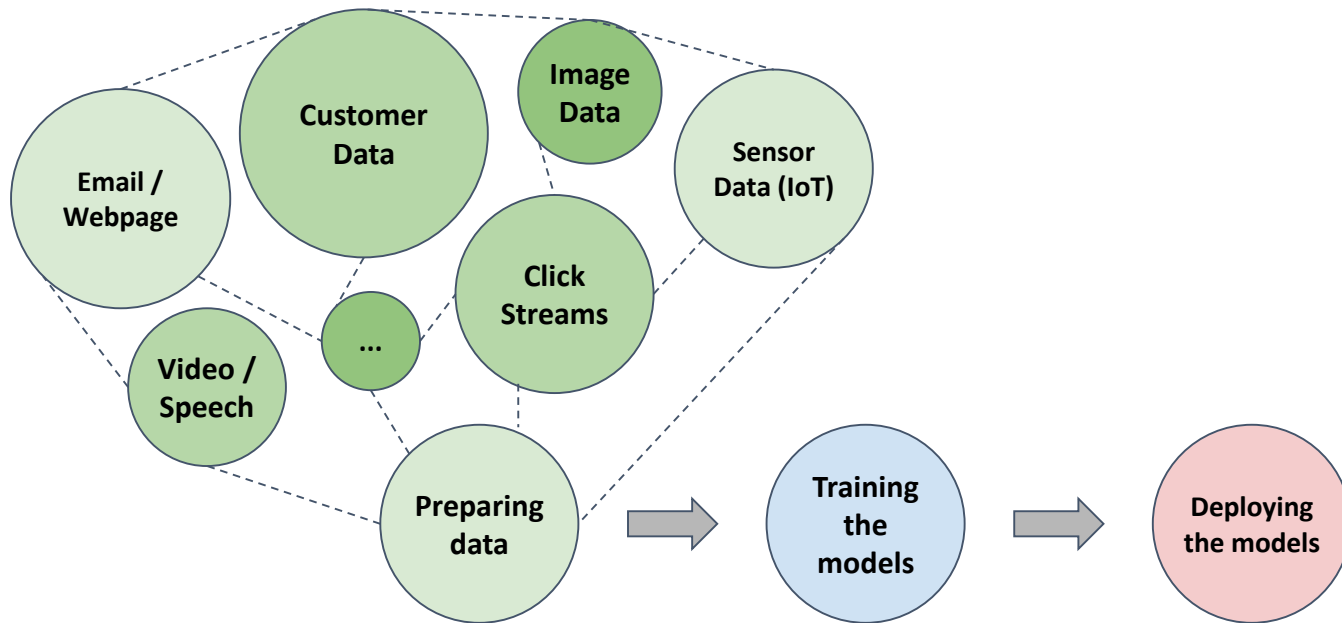
Teams working in silos (cont.)

- To avoid a decrease in AI productivity due to “siloed” teams, make sure to be aware of the following indicators:
 - Lack of integrated analytics environment for data engineers to iteratively create and provide high quality datasets to data scientists.
 - Lack of collaboration capabilities limiting knowledge sharing among data scientists, and the ability to iteratively explore data, train, and fine-tune models as a team.
 - Complex procedures when deploying models into production leading to multiple hands-off between data scientists, data engineers, and developers, slowing down processes and increasing risks to introduce errors



Data pipeline becomes complex

- Preparing data for analytics is a major bottleneck. Make sure to invest time in defining an integrated and flexible data strategy to avoid.



Lack of an agile AI ecosystem

- Efficiently setting up and maintaining proper machine learning environments are difficult due to too many existing ML & AI frameworks.

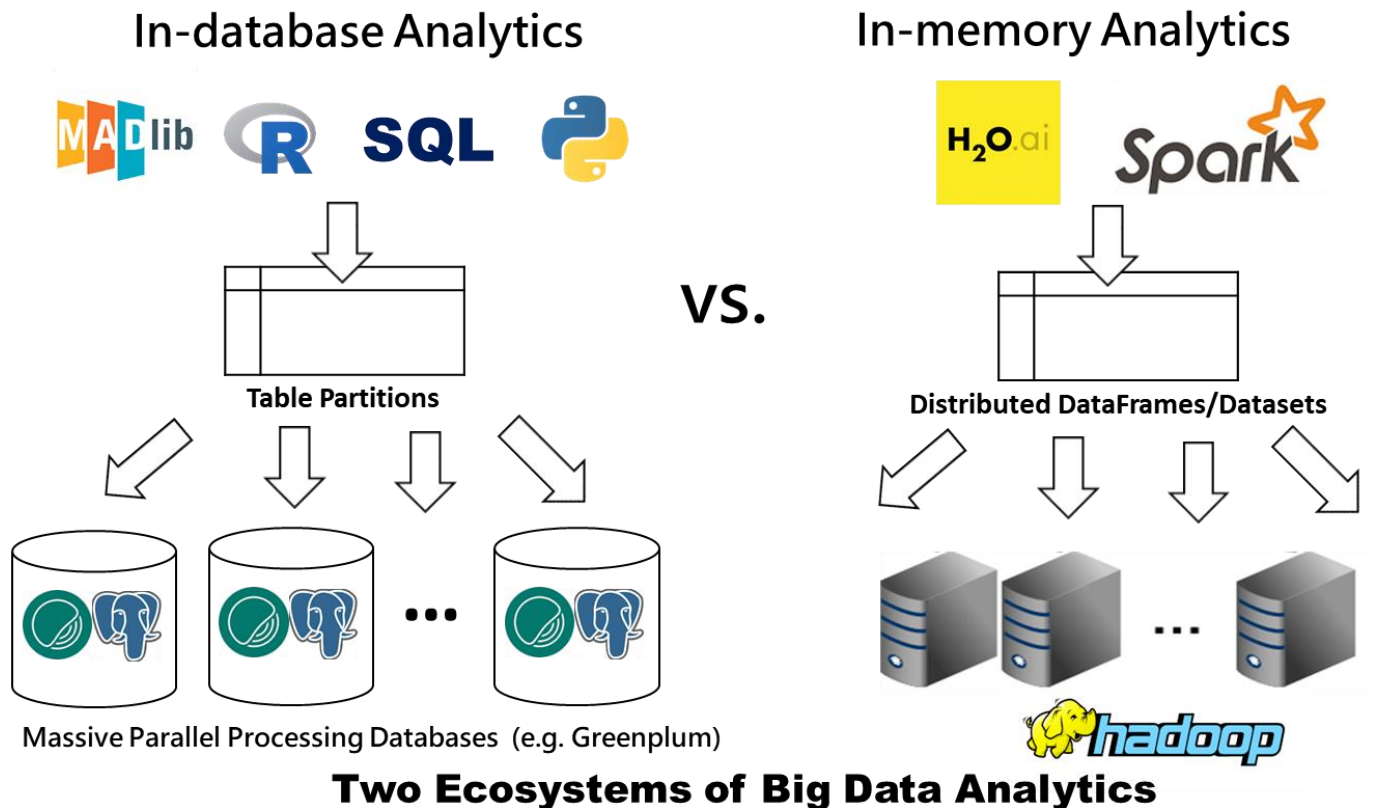
Machine Learning	Deep Learning	Supporting Libraries	Serving and Monitoring
Scikit-learn, Spark, MLlib, H2O, Mlpack, Mahout, ...	Tensorflow, Keras, Caffe, PyTorch, BigDL, SparkDL, ...	Python, R, Numpy, Scipy, Anaconda, ...	MLeap, TF serving, Azure ML, Cassandra, TensorBoard, Redis, ...

Lack of an agile AI ecosystem (cont.)

- Organizations are using on average 7 different tools within their AI technology stack. This explosion of options is actually a good thing because data scientists should have the choice to use the right framework to solve the right problems. A disjointed ecosystem lacks the ability to secure sufficient capacity and relevant data feeds for model training.
- Data scientists should be able to choose their favorite languages to visualize data and train models. This is the only way enterprises can truly solve the talent gap, which makes data scientists productive in their existing skills.
- Researchers and practitioners have noticed this problem, and started promoting ML & AI model/framework interoperability and interchangeability. Check out [PMML](#) and [ONNX](#) for more information.

Data infrastructure gets in the way

- Many organizations tend to have the data infrastructures unable to scale with growing data volumes for large teams of data scientists.



Towards Unified Analytics

—Four Key Parts Review

- Unifying all your siloed datasets.

Access your “big data” in one (virtual) place!

- Unifying data pipelining.

Foster a collaborative environment for data scientists and data engineers!

- Unifying data engineering and AI technologies

Simplify model deployment and management!

- Unifying infrastructure supporting DevOps.

Integrate software development (Dev) and operation (Ops) to improve monitoring of all steps of software construction!

Unified Analytics - Unlocking the Potential of AI

Modern data-driven enterprises need to run fast, iterative experiments to test and refine learnings supported by a strong collaboration between data science, engineering, and the business.

Massive Data

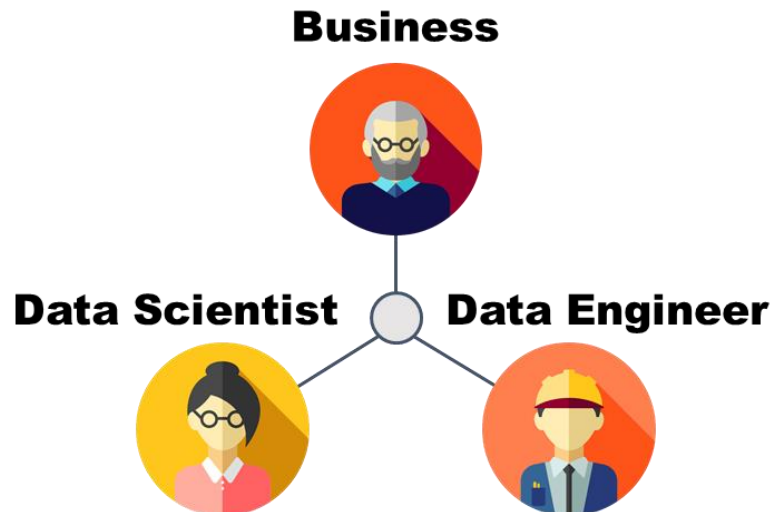
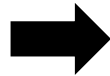
Customer Data

Video/Speech

Click Streams

Sensor Data(IoT)

...



Opportunity

Fraud Detection

Genome Sequencing

Recommendation System

Predictive Maintenance

...



Towards Unified Analytics Platform

—NSYSU 's experience



NSYSU Unified Analytics Platform

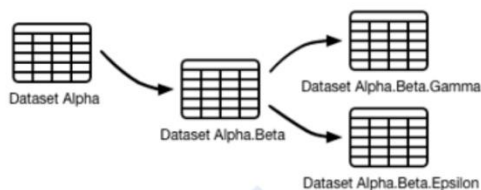
國立中山大學 管理學術研究中心

大數據運算平台 + 商業數據分析技術

研究資源

技術諮詢

分析方法



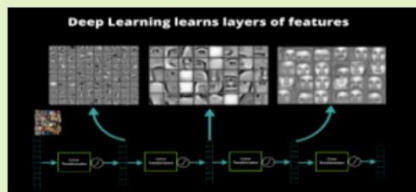
商業應用

平行運算叢集



轉換收集

深度學習主機



開源軟體

線上課程

技術社群

開放性資料

<http://cm.nsysu.edu.tw/~msrc/wp/>

II. 系統技術規格：

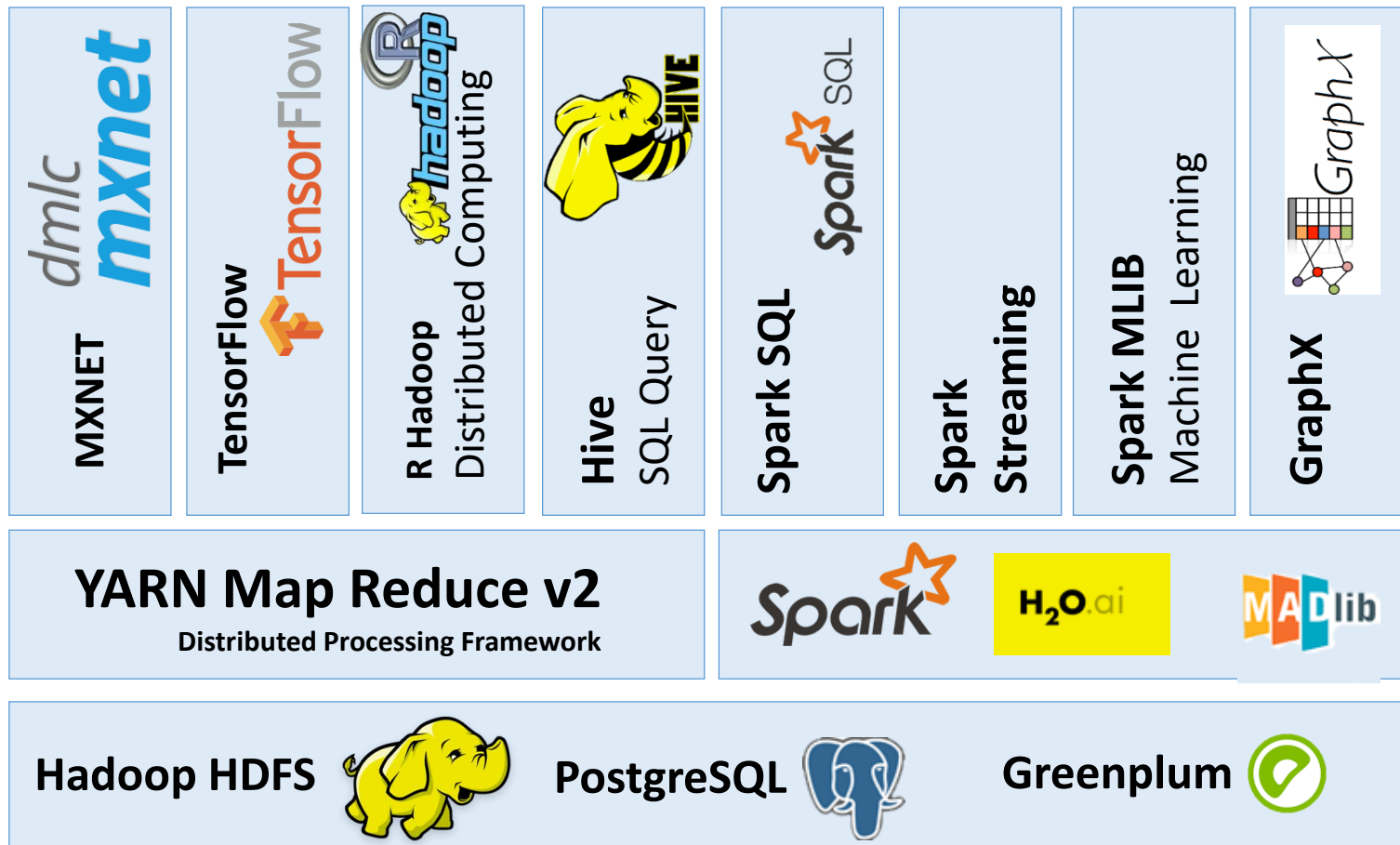
A. CPU 平行運算叢集：

- 1 x Master Node: 32 cores, 100G
- 17 x Worker Nodes: 8 cores, 61G
- R 3.4.1, RStudio 1.0.153
- Apache HADOOP 2.7
- Apache Spark 2.1.0
- Apache Zeppelin 0.7.2
 - R 3.4.1
 - Python 2.7.5
 - scala 2.11.8

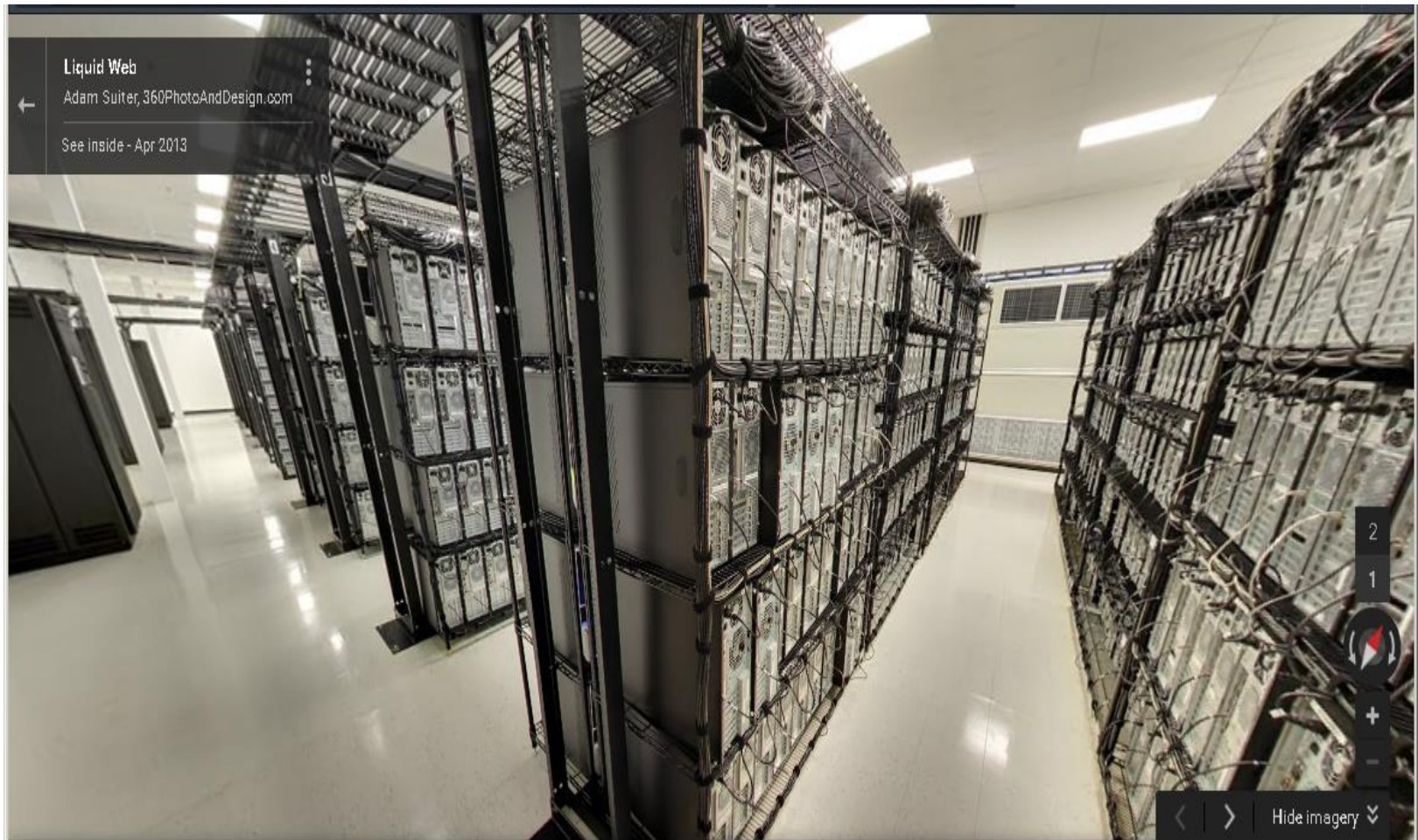
B. GPU 深度學習主機：

- 1 x Server: 48 cores, 280G
- 1 x GPU Card: Tesla P100-PCIE-16GB
- R 3.4.1, RStudio 1.0.153
- Python 2.7.5
- MxNet 0.11.0
- Tensorflow-GPU 1.3.0
- Keras 1.2.2 + MxNet 0.11.0
- Keras 2.0.8 + Tensorflow 1.3.0

NSYSU Unified Analytics Software Stack



Future of the platform...



NSYSU Unified Analytics Platform

The screenshot displays the RStudio environment. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Tools, and Help. The top right shows the user 'yihuang' and a 'Sign Out' link. The top status bar indicates 'Project: (None)'. The left sidebar contains the Environment, History, and Files panels. The Environment panel shows the Global Environment with a search bar. The Data panel shows a data frame 'big_num_4...' with 400,000 observations and 1,000 variables. The Functions panel shows a function 'crosstab_...' with arguments 'dfs_data', 'x', and 'y'. The bottom left sidebar contains the Plots, Packages, Help, and Viewer panels. The Packages panel shows a table of installed and available packages.

Name	Description	V...
acepack	ace() and avas() for selecting regression transformations	1.3-3.3
arules	Mining Association Rules and Frequent Itemsets	1.1-9
assertthat	Easy pre and post assertions.	0.1
BH	Boost C++ Header Files	1.58.1
bit	A class for vectors of 1-bit booleans	1.1-12
bit64	A S3 Class for Vectors	0.9-

The main editor window shows an R script 'Untitled1*' with the following code:

```
1 Sys.setenv(HADOOP_CMD="/home/hadoop/hadoop/bin/hadoop")
2 Sys.setenv(JAVA_HOME="/usr/java/latest")
3 Sys.setenv(HADOOP_STREAMING="/home/hadoop/hadoop/share/hadoop/tools/lib/hadoop
4 Sys.setenv(HADOOP_HOME="/home/hadoop/hadoop")
5 #Sys.setenv(HADOOP_OPTS="$HADOOP_OPTS -Djava.library.path=/home/hadoop/hadoop/
6
7 library(rhdfs)
8 library(rmr2)
9
10
```

The console window shows the output of the script, including Hadoop job progress and completion status:

```
15/09/15 11:46:01 INFO mapreduce.Job: map 30% reduce 0%
15/09/15 11:46:04 INFO mapreduce.Job: map 42% reduce 0%
15/09/15 11:46:07 INFO mapreduce.Job: map 47% reduce 0%
15/09/15 11:46:10 INFO mapreduce.Job: map 53% reduce 0%
15/09/15 11:46:13 INFO mapreduce.Job: map 59% reduce 0%
15/09/15 11:46:15 INFO mapreduce.Job: map 76% reduce 0%
15/09/15 11:46:16 INFO mapreduce.Job: map 79% reduce 0%
15/09/15 11:46:19 INFO mapreduce.Job: map 81% reduce 0%
15/09/15 11:46:22 INFO mapreduce.Job: map 100% reduce 0%
15/09/15 11:46:26 INFO mapreduce.Job: map 100% reduce 67%
15/09/15 11:46:29 INFO mapreduce.Job: map 100% reduce 100%
15/09/15 11:46:29 INFO mapreduce.Job: Job job_1442057141596_0007 completed successfully
```

NSYSU Unified Analytics Platform(cont.)



NSYSU Unified Analytics Platform(cont.)

Unified Analytics

» Data Management

» Upload Data

» Visualize Data

» Import to Cluster

Cluster Management

Monitor

Logout

Import to Cluster Service

HDFS

Please Select a File on Server

Select

CSV Preview

☐ None

☒ Head

☐ All

Header

☒ Yes

☐ No

Delimiter

☒ Comma

☐ Semicolon

☐ Tab

☐ User Defined

Quote

☐ None

☒ Double Quote

☐ Single Quote

Import

Data Preview

Show 10 entries

Search:

	sepal_length	sepal_width	petal_length	petal_width	species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

Showing 1 to 10 of 10 entries

Previous1Next

Import log

NSYSU Unified Analytics Platform(cont.)

Unified Analytics

Data Management

Cluster Management

Monitor

» Process Status

» Cluster Status

» GPU Status

» Spark Status

» Service Status

Logout

Update every:
30 seconds

Overview

Service Name	Status	ON / OFF
Hadoop	✓	<div>Strat</div> <div>Stop</div>
Spark	✓	<div>Start</div> <div>Stop</div>
H2O	✓	<div>Start</div> <div>Stop</div>
Greenplum	✓	<div>Start</div> <div>Stop</div>
RStudio(.5)	✓	<div>Start</div> <div>Stop</div>
Jupyterhub(.5)	✓	<div>Start</div> <div>Stop</div>
RStudio(.100)	✓	<div>Start</div> <div>Stop</div>
Jupyterhub(.100)	✓	<div>Start</div> <div>Stop</div>

Output of the Service:

Group Discussion #7

- 商業智慧(Business Intelligence)與整合分析(Unified Analytics) 兩者是否不同? 為什麼?



Build our Data/AI Dream Team

- Understanding data science process

**Understand
business problems**

**Decide
analytic
approaches**

**Collect
data**

**Engineer &
unsterstand
data**

**Model &
analyze
data**

Evaluation

Decision

Build our Data/AI Dream Team (cont.)

- Roles in a Data Team



Team Leader/Coordinator

Connection between the team and the business
Set up regular forums
(not necessarily technically skilled)



Data Engineer/Data Architect

Infrastructure (storage, management & cleaning)
Data collection



Data Analyst

Reporting
Visualization
Research projects



Business Intelligence Professional

Know the organization's key objectives
Communicate
Decision making



Data Scientist

Machine learning, statistics, big data
Predictive analytics
Can come up with new algorithms



Other

Based on the needs of the business
Multidisciplinary

Understand business problems



- Know the problems in the organization that you want to solve and decide the objectives or KPIs.
- Set the expectations, scope of the project & responsibilities
- The coordinator, possibly sourced from the business, must ensure the needs of the business to be communicated well to the data team.

Decide analytic approaches



- Determining which analytic/machine learning approach that can best address the business problems is the data scientist's job.
- A data scientist who knows **SQL, Python or R** is ideal.

Collect data



- This process includes data collection and storage. The data engineer is in charge of constructing and maintaining the infrastructure that improves the quality and efficiency of data.
- Solid knowledge of database skills & engineering skills

Engineer & understand data



- Pre-process and clean the data (might take 80% of the data scientists' time).
- Report on a regular basis (hourly, daily, weekly, monthly)
- Visualization

Model & analyze data



- The data scientist uses advanced analytical techniques to create models and reports the results.
- Big data applications (SQL, Python or R)

Evaluation



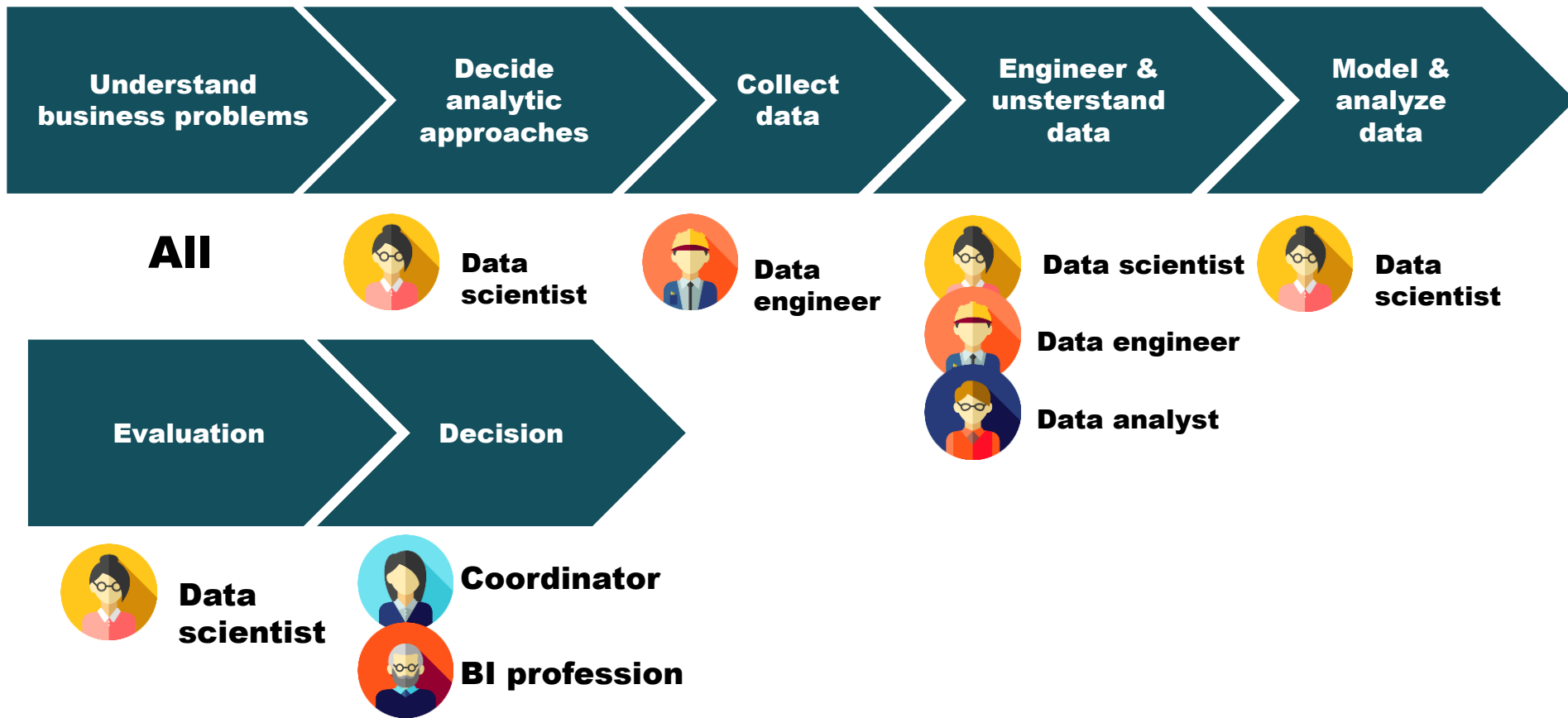
- Evaluate the model in terms of performance, accuracy, and interpretability. To improve the model, go to the previous step to modify the model

Decision



- With an understanding of the organizational strategy and decision-making, the process involves the coordinator and BI professions to address the business problems by the reports.

Roles in the data science process



Build our Data/AI Dream Team (cont.)

- Again, set up the basic data pipelining infrastructure.
An Unified Analytics Platform is a good start!
 - ✓ Databases
 - ✓ Software
 - ✓ Servers
 - ❑ ...
- Tips and tricks:
 - ✓ Communications across the data team
 - ✓ Hire one person at a time
 - ✓ Data analyst or data scientist?
 - ✓ Transfer/complementary skills
 - ❑ Anything else you can come up with!?



Group Discussion #8

- 建立我們的 data analytics team 需要注意哪些事項？你的角色比較可能會是什麼？



Thank you so much!

Questions?