การอบรม Machine Learning for Data Analytics
ณ ห้องประชุม 501 ชั้น 5 สมาคมประกันวินาศภัยไทย
(30 สิงหาคม 2562)

ผศ.ดร.สุรณพีร์ ภูมิวุฒิสาร

Instructor

- **ผศ.ดร.สุรณพีร์ ภูมิวุฒิสาร**
  (ผู้อำนวยการบัณฑิตศึกษา คณะวิทยาการและเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีมหานคร)
- Education
  - วศบ. วิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย
  - M.Sc., Information Science, University of Pittsburgh, USA
  - Ph.D Computer Science and Engineering, University of New South Wales, Australia
- Teaching Experience
  - Machine Learning
  - Mathematics for Research
- Contact
  - suronape@mut.ac.th

MSIT MUT

AGENDA

Lecture (9.00 – 10.30)
- Machine Learning Overview

Lecture (10.45 – 12.00)
- Applied Analytics through Case Studies

Lecture (13.00 – 16.00)
- Demo – Insurance Case
- Model Assessment

MSIT MUT

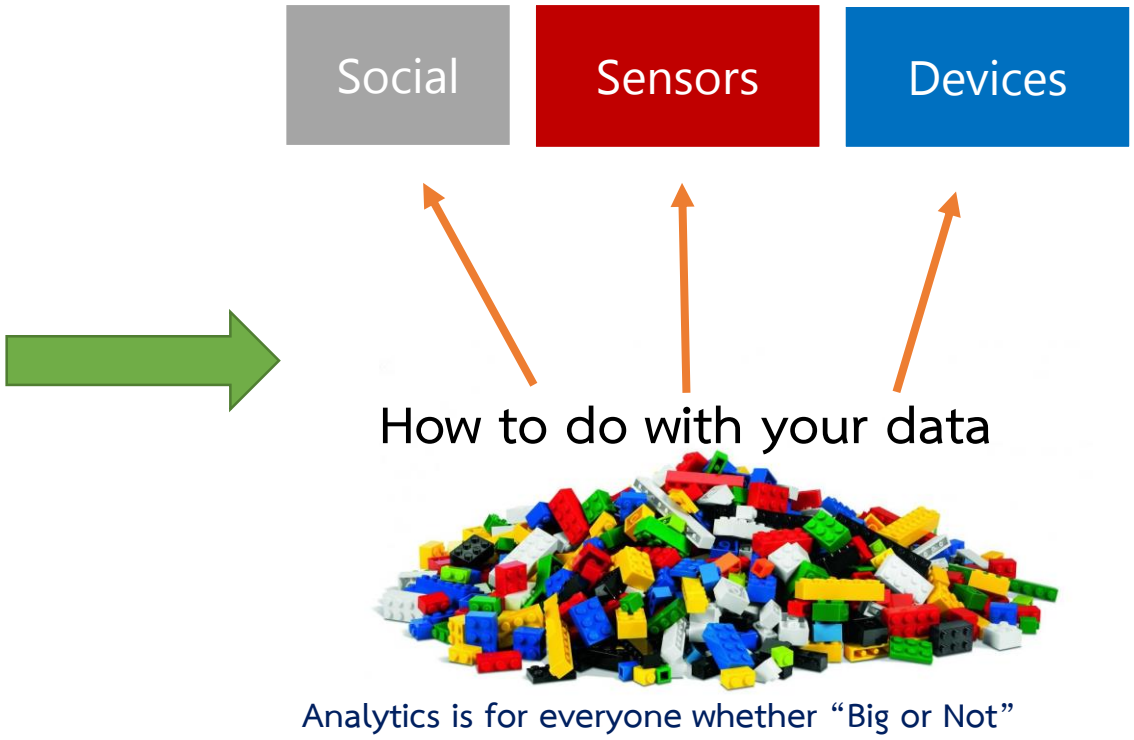
Big Data Analytics

MSIT MUT

- We are entering the era of big data
  - Big Data Analytics create "Business Value" in today's world (better decision)

Picture: Internet

Recent Big Success Story



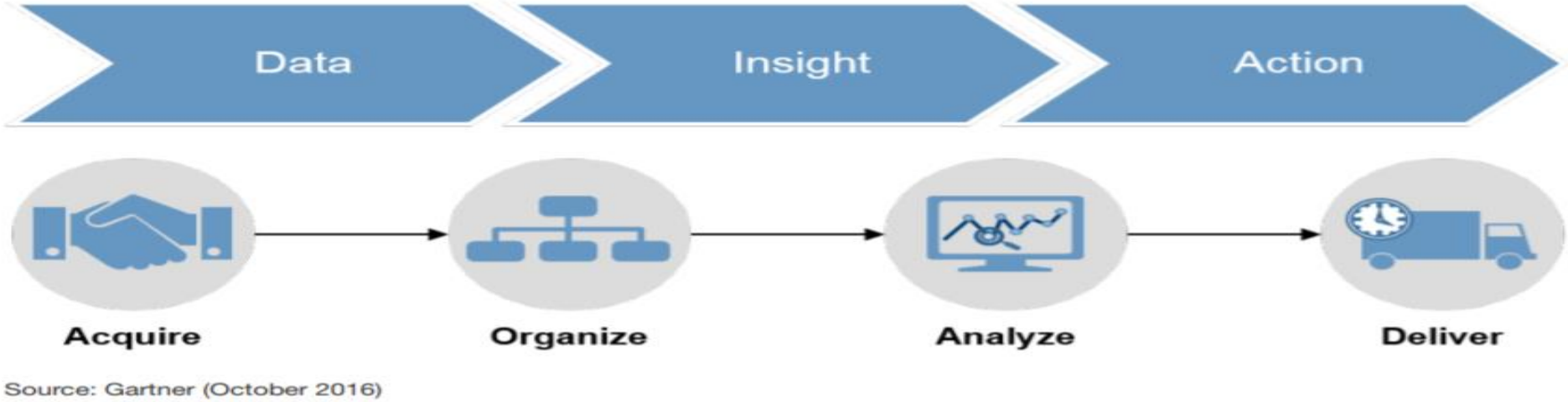
Analytics is for everyone whether "Big or Not"

Picture: https://www.promptcloud.com/blog/datafication-era-of-big-data

Big Data Analytics

MSIT MUT

- The process of discovering actionable insights through the analysis from the data”

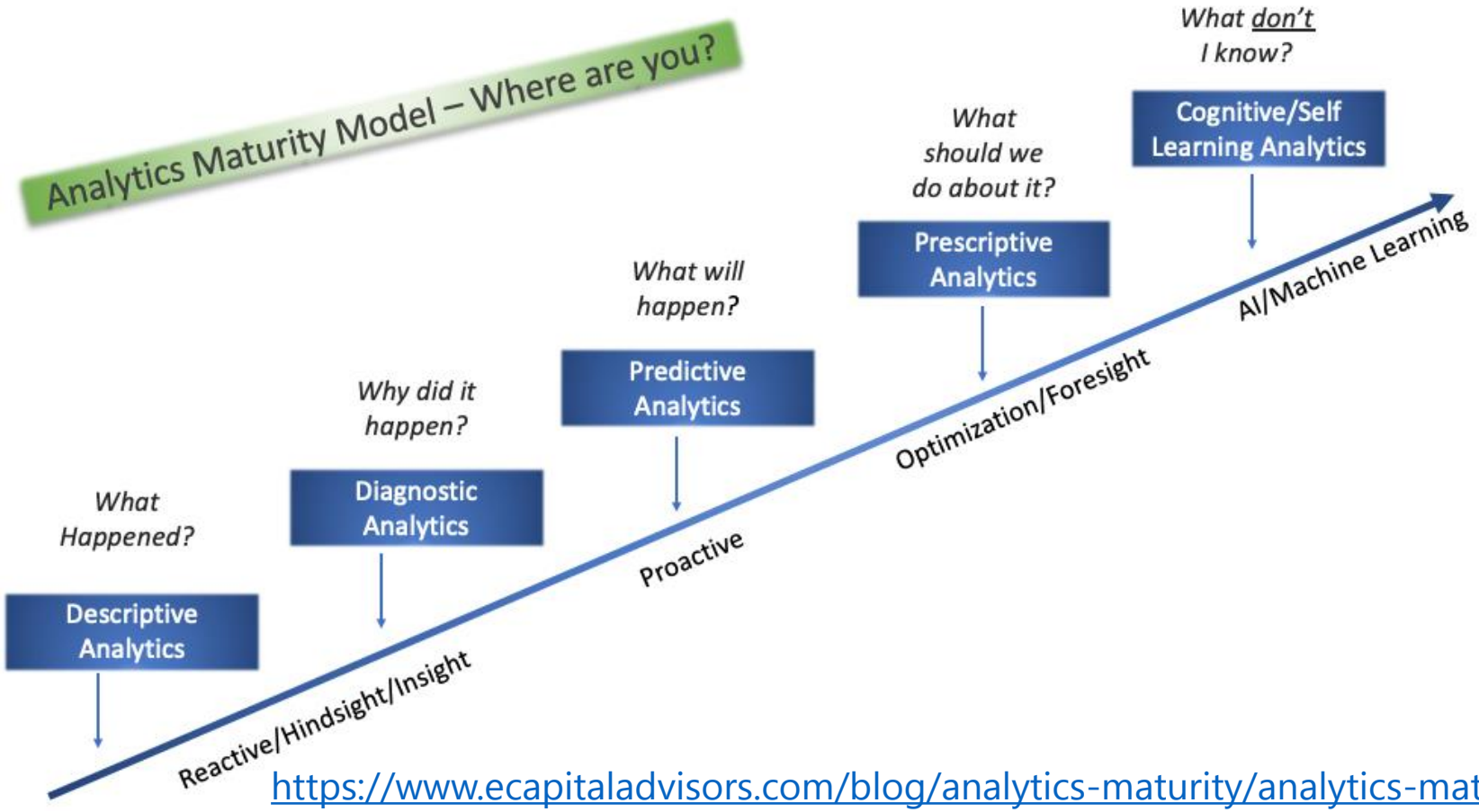


Source: Gartner (October 2016)

- Example – NETFLIX
  - What is data? ข้อมูลการรับชมภาพยนตร์ของสมาชิก
  - What is insight? ภาพยนตร์เรื่องไหน ประเภทใด ที่เป็นที่สนใจที่ผู้ชม
  - What is action? ระบบแนะนำโปรแกรมหนัง/การสร้างภาพยนตร์

Types of Big Data Analytics

MSIT MUT



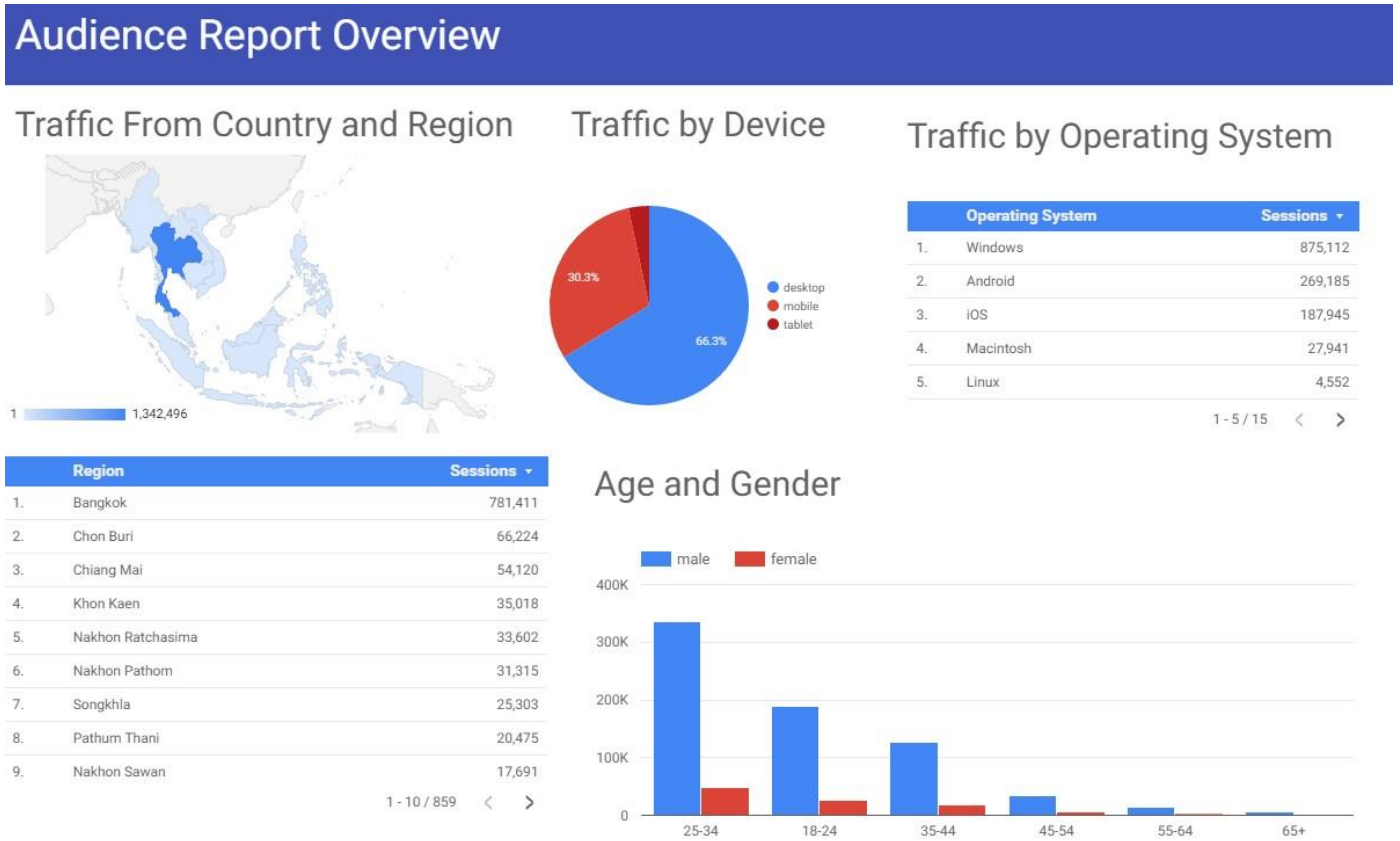
https://www.ecapitaladvisors.com/blog/analytics-maturity/analytics-maturity/

Descriptive Analytics

MSIT MUT

- "What happened?"
- "Describe" or "Summarize" raw data and make it interpretable by humans.
- Example
  - Google Analytics

Google Analytics



Picture: https://www.itopclass.com/blog/google-analytics-คืออะไร

Diagnostic Analytics

MSIT MUT

- Why did it happen?

- Underline why something happened in the past and help locate the root cause of the problem

- Example
  - factors influence medical expenses for insurance company

Which factors influence medical expenses charged per year?

```
> str(insurance)
'data.frame':   1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8 25.8 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2$
 $ expenses: num  16885 1726 4449 21984 3867 ...
```

Predictive Analytics

MSIT MUT

- "What will happen?"

- Make predictions about the future

- Example
  - Southwest Airlines

What might happen in the future

For example, Southwest Airlines analyses sensor data on their planes in order to identify patterns that indicate a potential malfunction, thus allowing the airlines to the necessary repairs before its schedule.

Picture: https://www.edureka.co/blog/big-data-analytics/

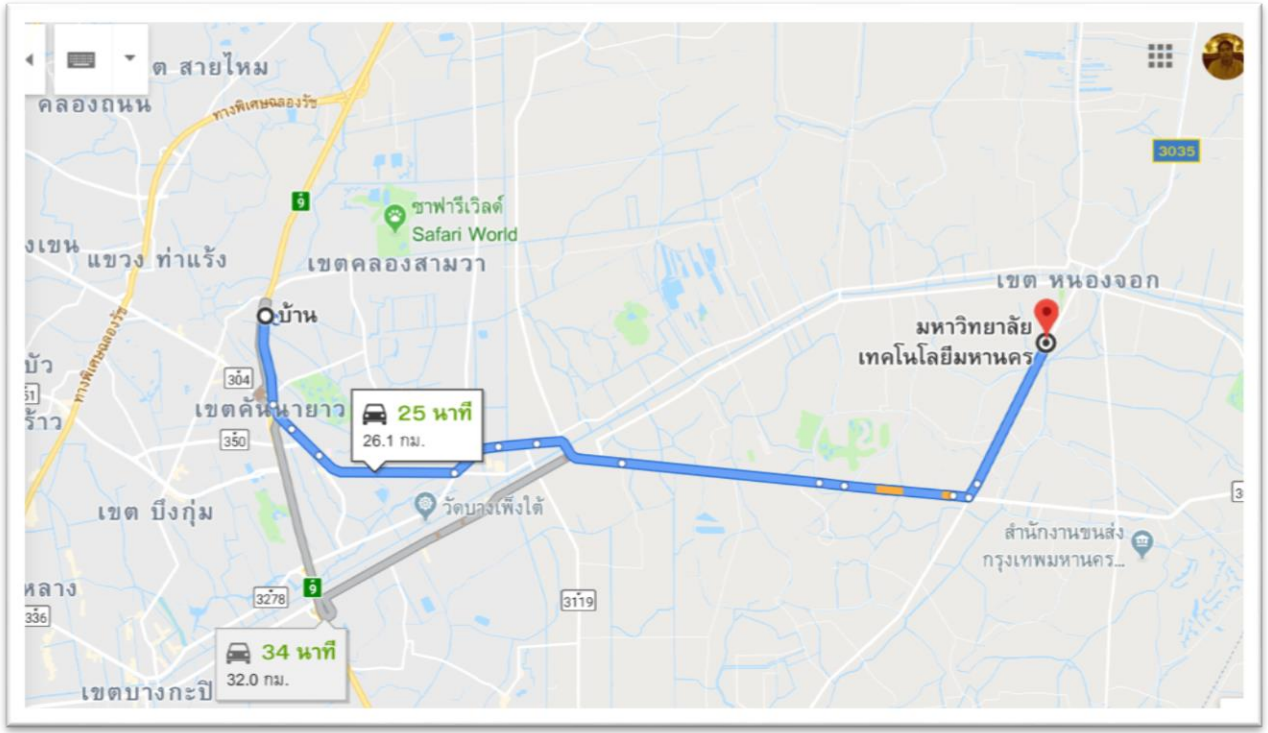
Prescriptive Analytics

MSIT MUT

- "What should we do about this?"

- Uses optimization and simulation algorithms to advice on possible outcomes

- Example
  - Google Maps – best root



Google Maps

Cognition/ Self Learning Analytics

MSIT MUT

- What don’t I know?

- Cognition is all about thinking, understanding, learning and remembering. Cognitive computing is all about creating analytics technology that mimics the human brain's ability to perform these functions.

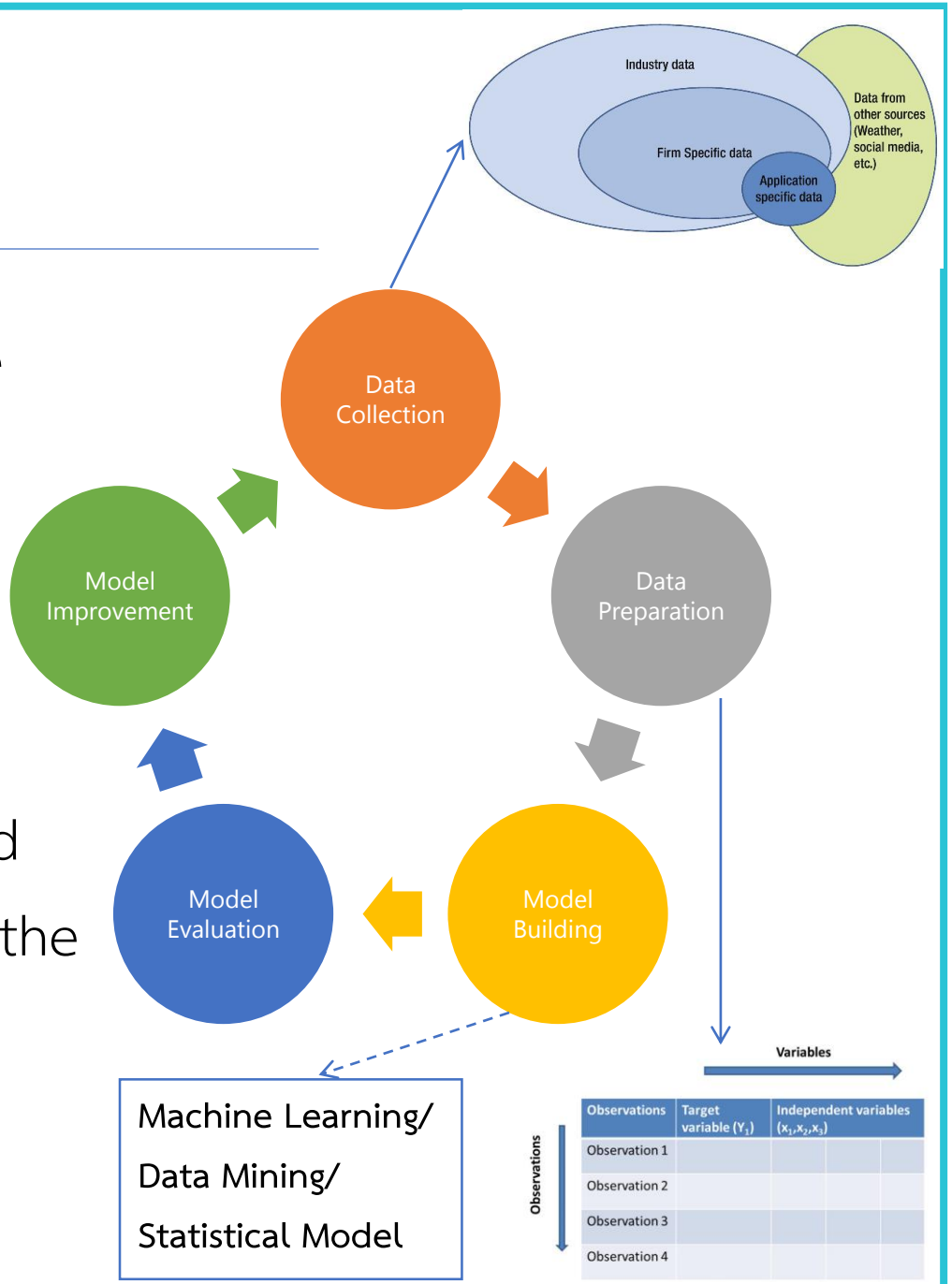
- Example
  - Google ‘s self-driving car

Picture: adapt from https://www.edureka.co/blog/big-data-analytics/

Data Analytics Process

- **Data collection**: gather data an algorithm will use
- **Data preparation**: prepare the data
- **Model Building**: chose a learning algorithm to represent

  the data in the form of a model.
- **Model evaluation**: evaluate the algorithm learned
- **Model improvement**: improve a performance of the model

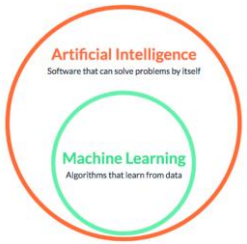


Source: Applied Analytics through case studies

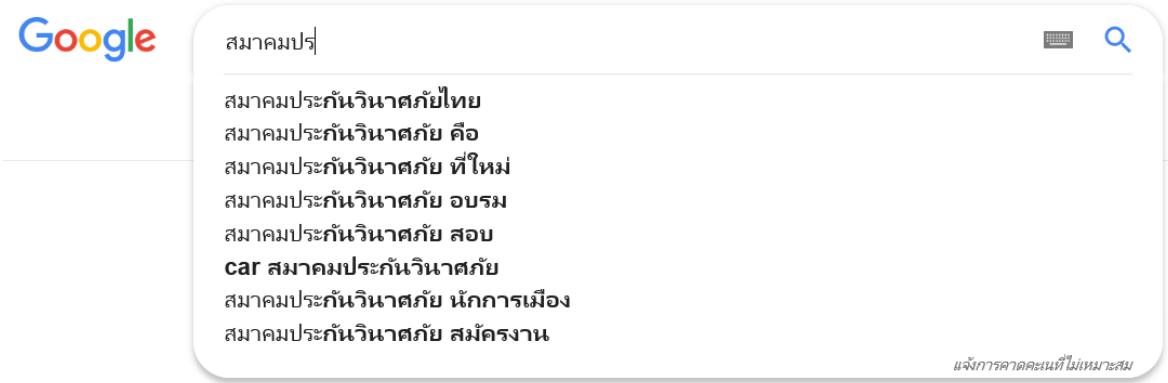
Machine Learning

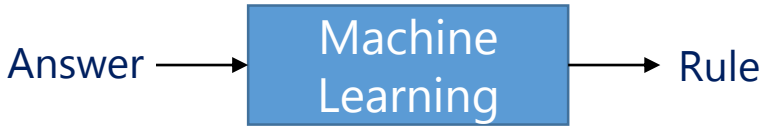
MSIT MUT

- What is Machine Learning?



"the field of study that gives computers the ability to learn without being explicitly programmed" Arthur Samuel (1959)

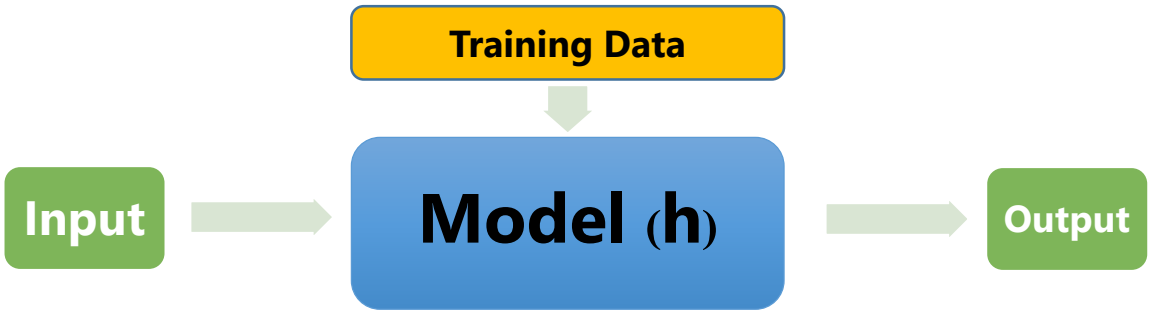




How does the Machine Learn?

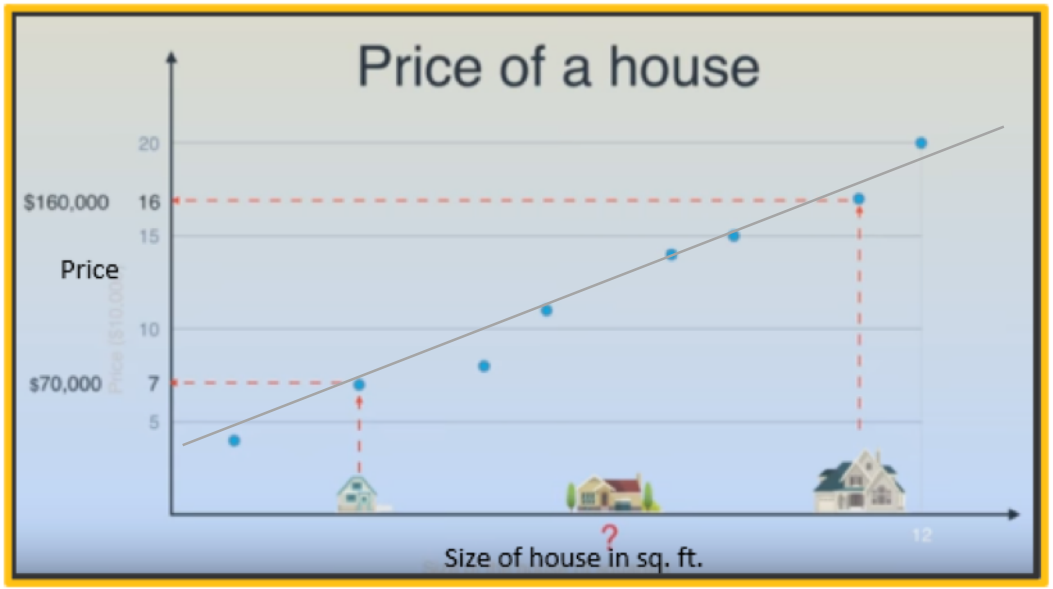
MSIT MUT

- How Machine Learns?



- Housing Price Prediction

Picture : internet

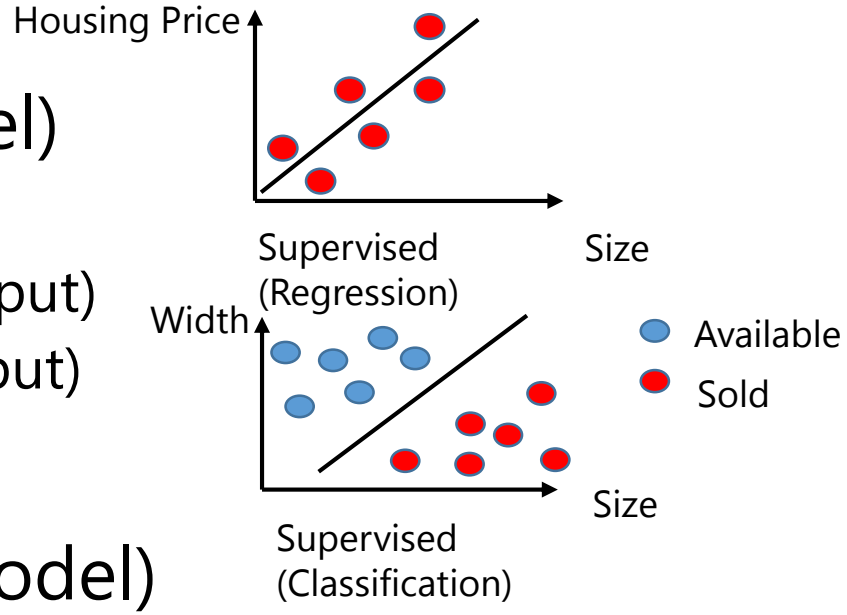


Choose algorithm to build the model

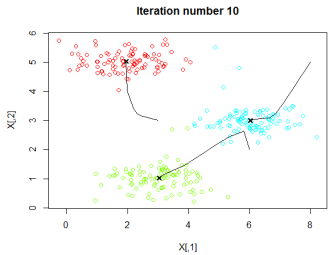
Picture : internet

Types of Machine Learning

- Supervised learning (h = predictive model)
  - *"correct answers" given*
    - Regression problems (predict continuous output)
    - Classification problems (predict discrete output)



- Unsupervised learning (h = descriptive model)
  - *" no correct answers" given*



Unsupervised - Market Segmentation (example)

Classification or Regression problems?

- **You're running a company, and you want to develop learning algorithms to address each of two problems**
  - **Problem 1**: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.
  - **Problem 2**: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.

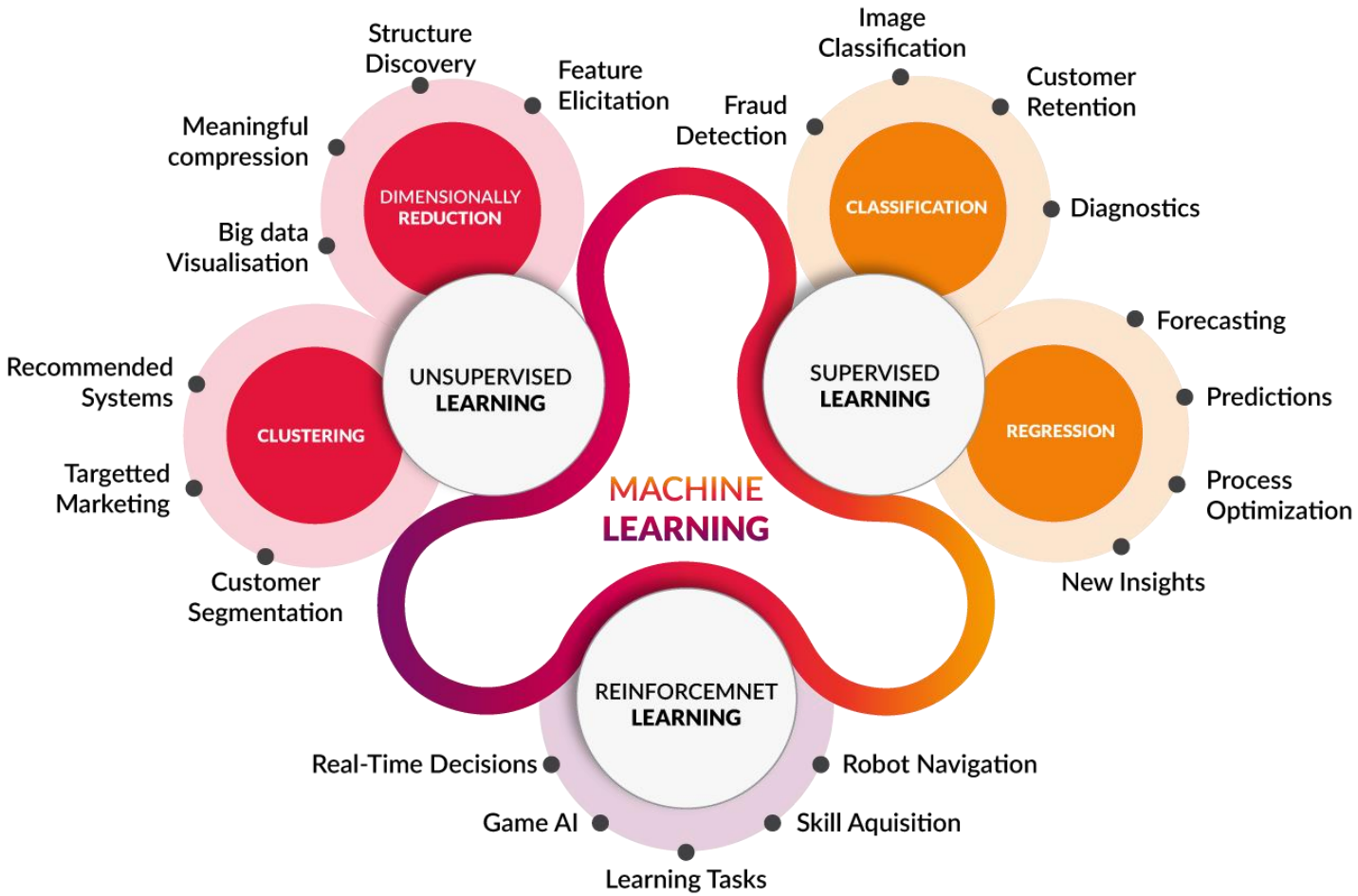
Recent Success Stories

Source: Google

- **Examples**

Application	Input	Output
Real Estate	Size, #Bedrooms, Location	Price of the House
Spam Classification	Keywords	Spam/not Spam
Self-Driving Car	Image, Radar Info.	Position of other Cars
Face Recognition	Image	Yes/No
Speech Recognition	Audio	Text
Game Application	User Profile, Ad Info.	Click/ not Click

Types of Machine Learning (con’t)



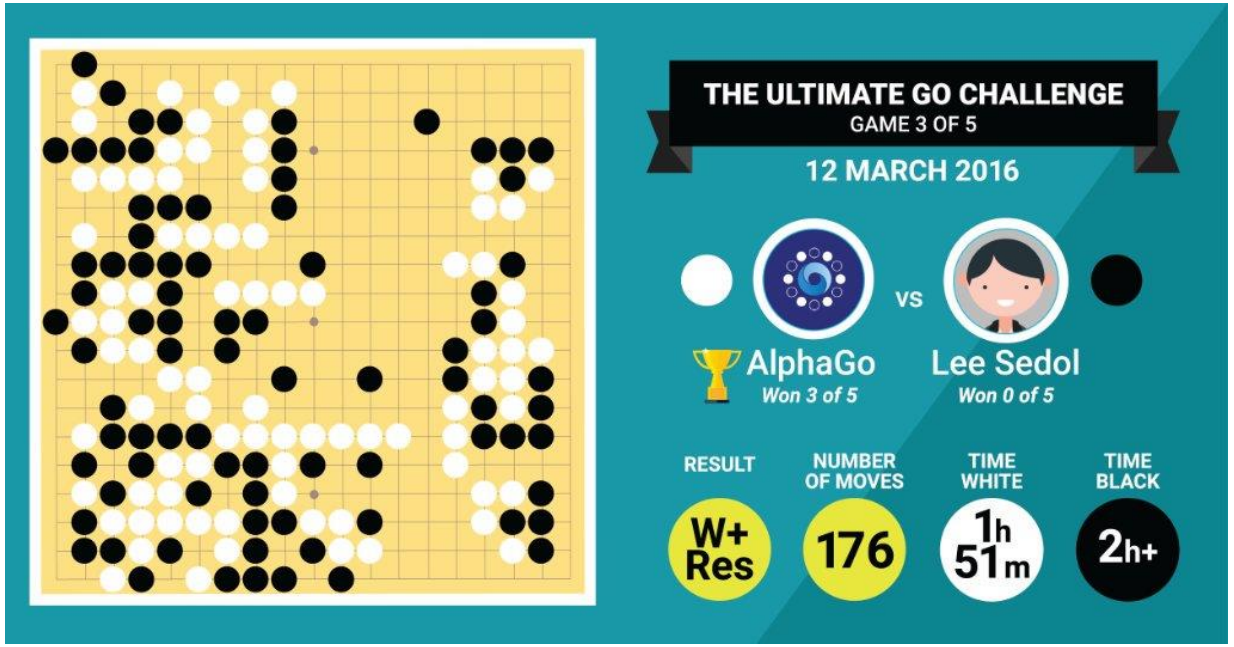
MSIT MUT

MSIT MUT

The Limits of Machine Learning

MSIT MUT

- **Machine learning, at this time, is not in any way a substitute for a human brain.**
  - little flexibility to extrapolate outside of the strict parameters it learned and knows no common sense



Source: Google Twitter

Common Notation

x = input variable/features $x = \{x_1, x_2, \ldots, x_n\}$

Numeric – year, price, mileage

categorical (or Nominal) – model, color, transmission

y = output variable/target feature $y = \{y_1, y_2, \ldots, y_n\}$

Training examples = $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), .. , (x^{(m)}, y^{(m)})\}$

m = number of training examples

Size	House Price
100	1,000,000
200	2,000,000
300	3,000,000

features

year	model	price	mileage	color	transmission
2011	SEL	21992	7413	Yellow	AUTO
2011	SEL	20995	10926	Gray	AUTO
2011	SEL	19995	7351	Silver	AUTO
2011	SEL	17809	11613	Gray	AUTO
2012	SE	17500	8367	White	MANUAL
2010	SEL	17495	25125	Silver	AUTO
2011	SEL	17000	27393	Blue	AUTO
2010	SEL	16995	21026	Silver	AUTO
2011	SES	16995	32655	Silver	AUTO

examples

AGENDA

Lecture (9.00 – 10.30)
- Machine Learning Overview

Lecture (10.45 – 12.00)
- Applied Analytics through Case Studies

Lecture (13.00 – 16.00)
- Demo – Insurance Case
- Model Assessment

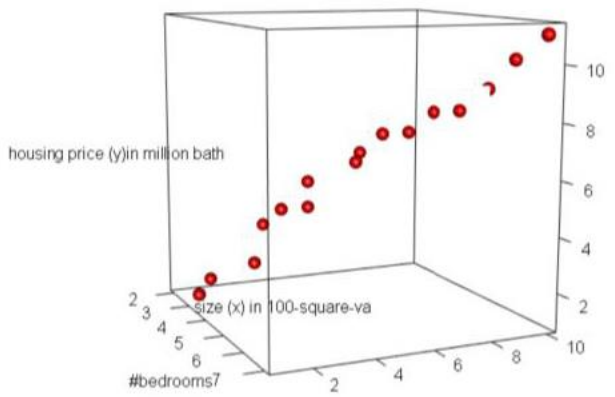
MSIT MUT

Case Studies – Supervised (Regression)

MSIT MUT

- Understanding Regression
  - specifying the relationship between one or more numeric independent variables (the predictors) and a single numeric dependent variable (the value to be predicted)

- More example
  - Predict housing price based on width and #bedroom



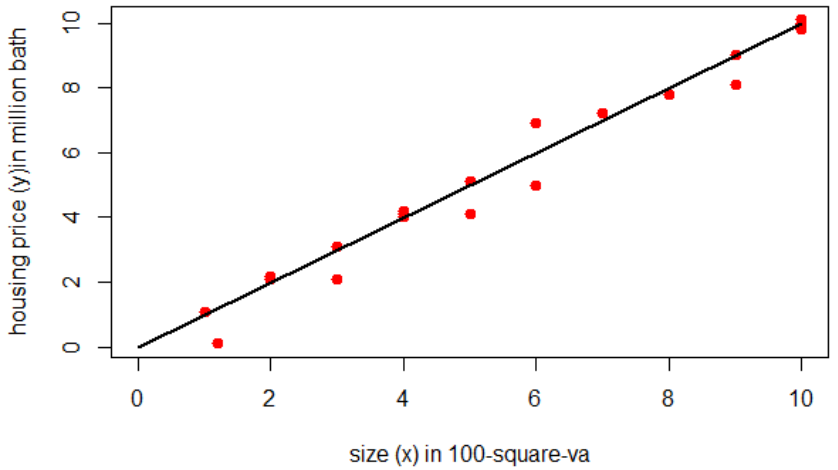
Housing Price Prediction

Linear Regression

MSIT MUT

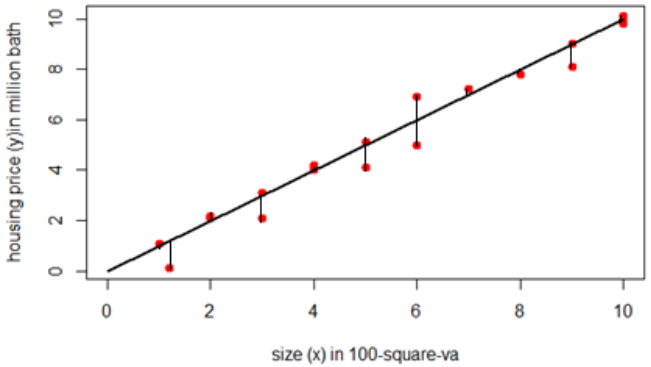
- Hypothesis Function

$$h_\theta(x) = \theta_0 + \theta_1 x$$



- Cost Function
  - Measure the accuracy of the hypothesis function

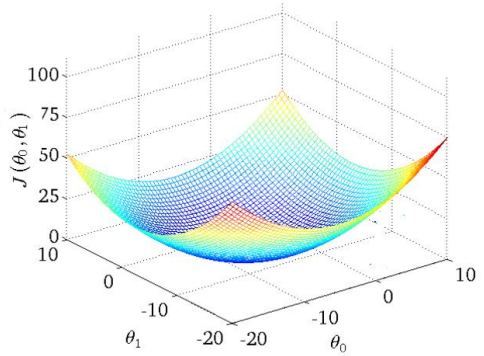
$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$



Linear Regression (2)

MSIT MUT

- Choose θ_0, θ_1 to minimize the cost function



Source: Andrew Ng Coursera

- Parameter Learning - Gradient Descent
  - Start with some parameters, for example θ_0, θ_1
  - Keep changing θ_0, θ_1 until we end up at a minimum

- Gradient Descent Algorithm

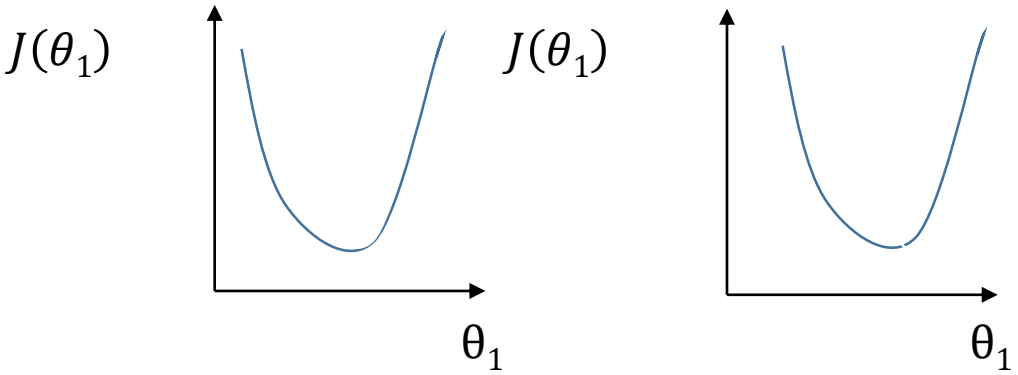
  Repeat until convergence{

  $$\theta_i := \theta_i - \alpha \frac{\partial J(\theta_i)}{\partial \theta_i}$$

  }

  $\alpha = learning\ rate$

suppose $\theta_0 = 0$



Example – Regression Problem (How much/How many?)

- Predict medical expense each person charged to the insurance plan for the year

กลุ่มตัวอย่าง 1,338 ตัวอย่าง

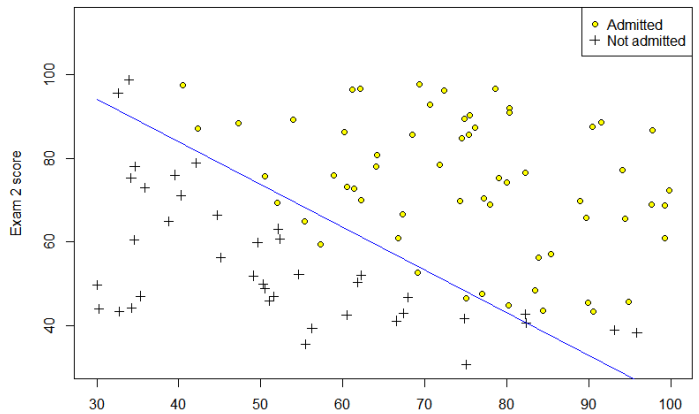
ตัวแปรต้น 6 ตัว ตัวแปรตามที่เราสนใจจะทำนาย 1 ตัว (expenses)

```
> str(insurance)
'data.frame':   1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8 25.8 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2$
 $ expenses: num  16885 1726 4449 21984 3867 ...
```

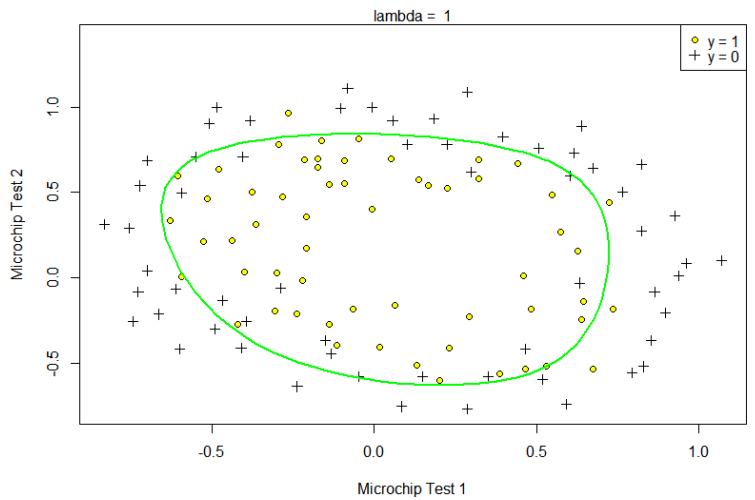
- Prediction Model
  - expenses = 139.0053 - 32.6181*age + 3.7307*age2 + 678.6017 * children + 119.7715*BMI - 496.7690*sexmale - 997.9355*BMI30 + 13404.5952*smokeryes + 19810.1534*bmi30:smokeryes

Case Studies – Supervised (Classification)

- Predict whether a student gets admitted or rejected into a university



- Predict whether a microship pass a quality test

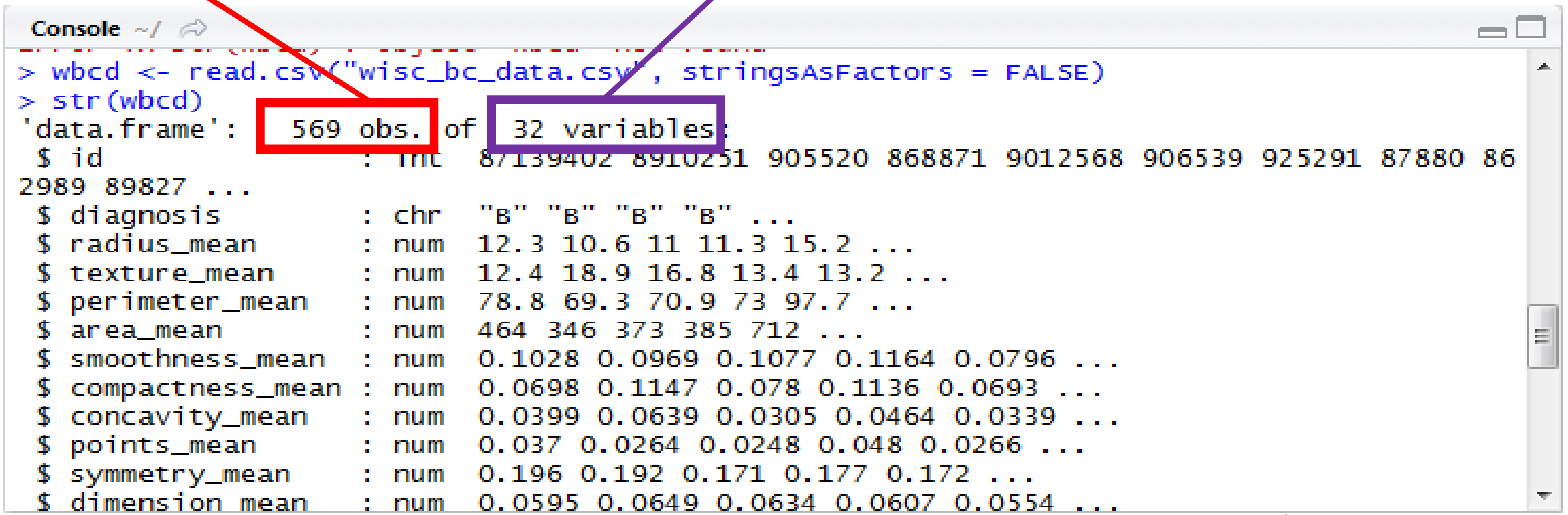


Example – Classification Problem (A or B or..)

- Diagnosing breast cancer by determining whether the mass is likely to be malignant or benign

ตัวแปรต้น 31 ตัว ตัวแปรตามที่เราสนใจจะทำนาย 1 ตัว (diagnosis)

กลุ่มตัวอย่าง 569 ตัวอย่าง

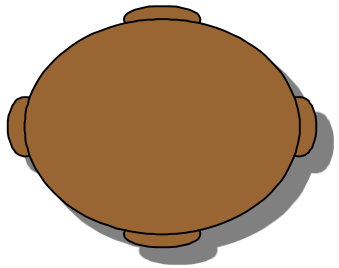


(Credit จาก demographic statistics from the US Census Bureau)

K-Nearest Neighbor Classification (KNN)

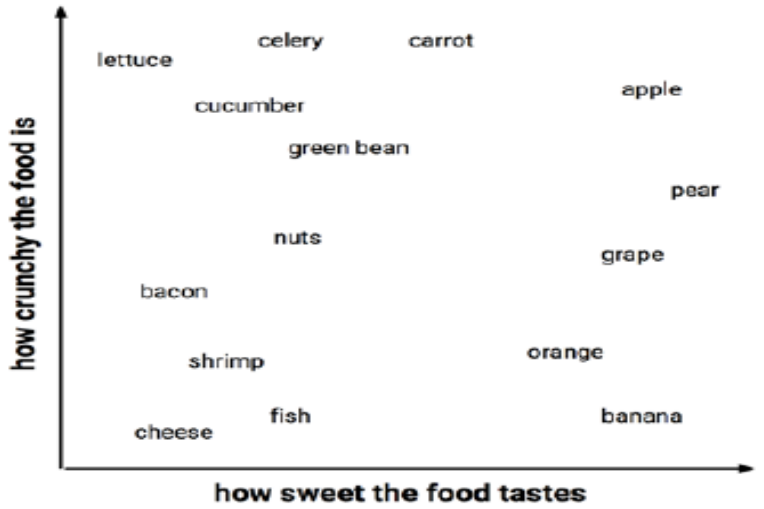
MSIT MUT

- Uses an example's k-nearest neighbors to classify unlabeled examples.
- Blind Tasting Experience Example



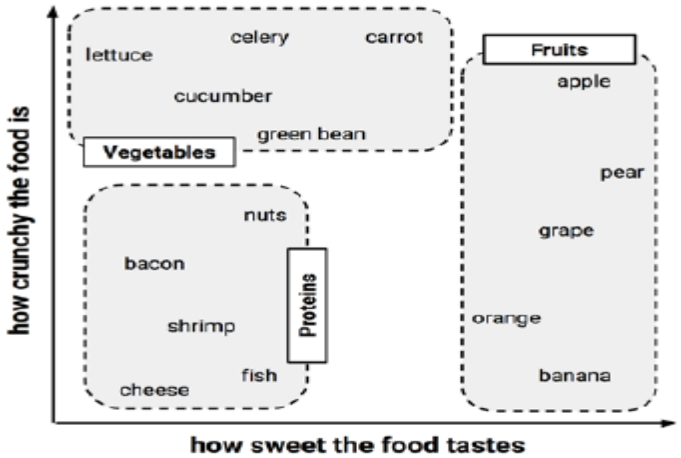
Ingredient	Sweetness	Crunchiness	Food type
apple	10	9	fruit
bacon	1	4	protein
banana	10	1	fruit
carrot	7	10	vegetable
celery	3	10	vegetable
cheese	1	1	protein

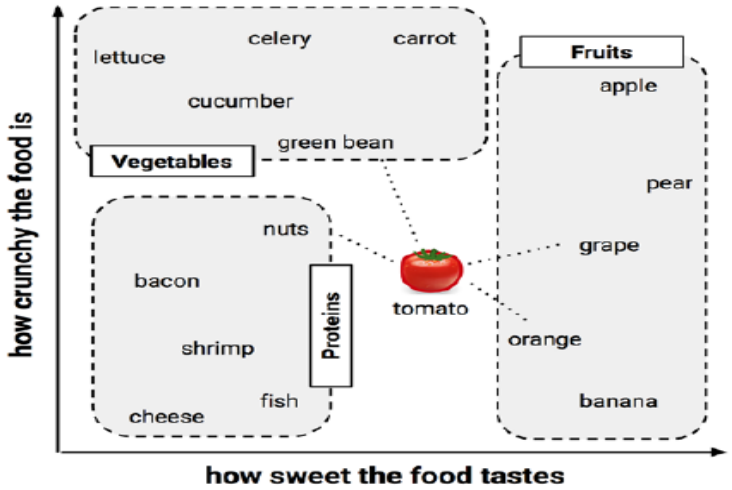
Source: Machine Learning with R



K-Nearest Neighbor Classification (KNN) (2)

- Is tomato a fruit or vegetable?





- Let ‘s calculate the distance between the tomato (sweetness = 6, crunchiness = 4), and the green bean (sweetness = 3, crunchiness = 7).

$$\text{dist}(\text{tomato, green bean}) = \sqrt{(6-3)^2 + (4-7)^2} = 4.2$$

Ingredient	Sweetness	Crunchiness	Food type	Distance to the tomato
grape	8	5	fruit	*sqrt((6 - 8)^2 + (4 - 5)^2) = 2.2*
green bean	3	7	vegetable	*sqrt((6 - 3)^2 + (4 - 7)^2) = 4.2*
nuts	3	6	protein	*sqrt((6 - 3)^2 + (4 - 6)^2) = 3.6*
orange	7	3	fruit	*sqrt((6 - 7)^2 + (4 - 3)^2) = 1.4*

Source: Machine Learning with R

K-NN and ML Algorithms

- Classification algorithms using nearest neighbor are considered lazy learning algorithms
  - Heavily reliance on the training instances rather than an abstracted model (also known as instance-based learning)
  - no parameters are learned about the data (non-parametric learning methods)

Source: Machine Learning with R Second Edition, Brett Lantz

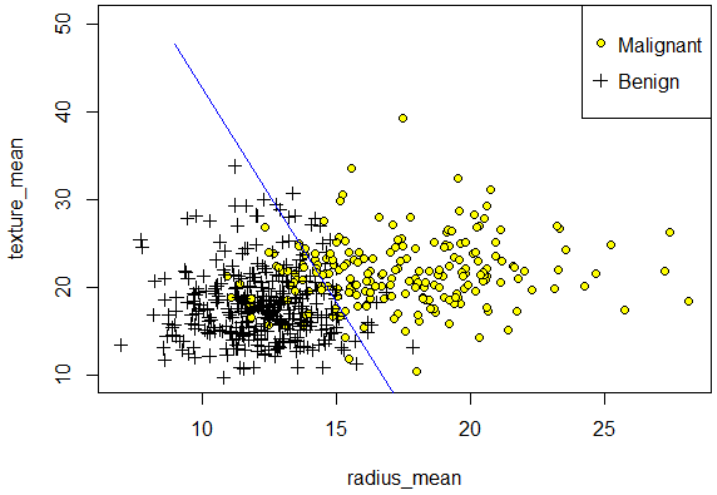
```
Console ~/
> wbcd <- read.csv("wisc_bc_data.csv", stringsAsFactors = FALSE)
> str(wbcd)
'data.frame':   569 obs. of  32 variables:
 $ id                : int  87139402 8910251 905520 868871 9012568 906539 925291 87880 86
2989 89827 ...
 $ diagnosis         : chr  "B" "B" "B" "B" ...
 $ radius_mean       : num  12.3 10.6 11 11.3 15.2 ...
 $ texture_mean      : num  12.4 18.9 16.8 13.4 13.2 ...
 $ perimeter_mean    : num  78.8 69.3 70.9 73 97.7 ...
 $ area_mean         : num  464 346 373 385 712 ...
 $ smoothness_mean   : num  0.1028 0.0969 0.1077 0.1164 0.0796 ...
 $ compactness_mean : num  0.0698 0.1147 0.078 0.1136 0.0693 ...
 $ concavity_mean    : num  0.0399 0.0639 0.0305 0.0464 0.0339 ...
 $ points_mean       : num  0.037 0.0264 0.0248 0.048 0.0266 ...
 $ symmetry_mean     : num  0.196 0.192 0.171 0.177 0.172 ...
 $ dimension_mean    : num  0.0595 0.0649 0.0634 0.0607 0.0554 ...
```

MSIT MUT

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|
```

Total Observations in Table: 100

	wbcd_test_pred		
wbcd_test_labels	Benign	Malignant	Row Total
Benign	61	0	61
	1.000	0.000	0.610
	0.968	0.000	
	0.610	0.000	
Malignant	2	37	39
	0.051	0.949	0.390
	0.032	1.000	
	0.020	0.370	
Column Total	63	37	100
	0.630	0.370	



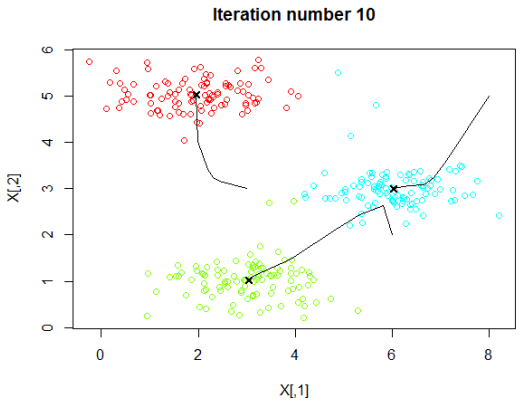
Case Studies - Unsupervised Learning

MSIT MUT

- Discover hidden patterns or a structures of data.
  - *"no correct answers"* - no target to learn
  - no single feature is more important than any other

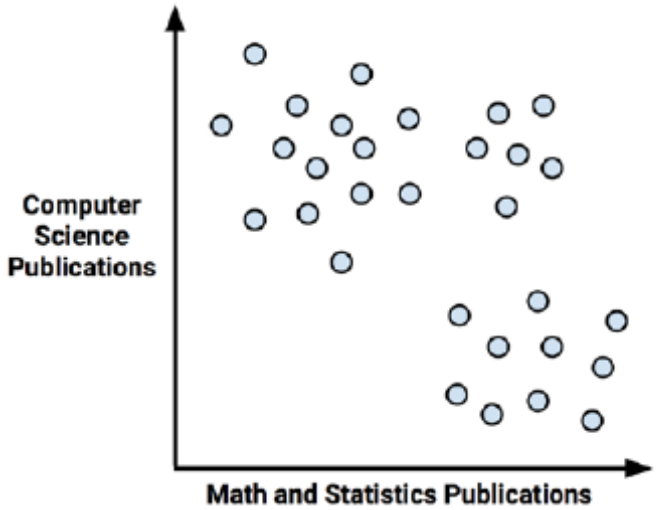
"no correct answers"

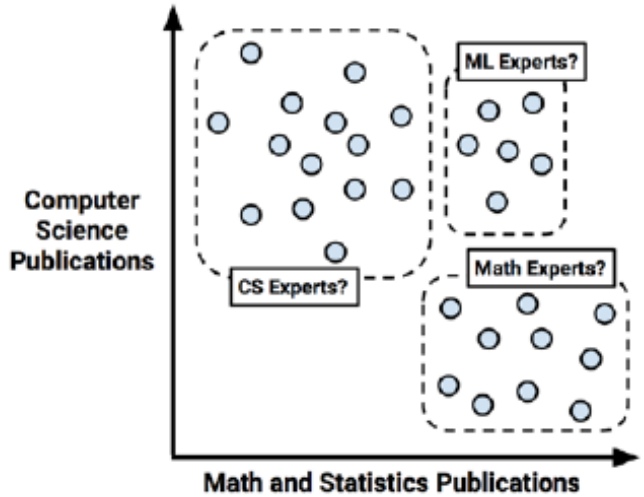
- Example
  - Teen market segments
    - เป็นการจัดกลุ่มวัยรุ่นที่ใช้งาน Social Media ตามลักษณะที่คล้ายๆกัน เช่น อายุ เพศ จำนวนเพื่อนความชอบส่วนตัว



Organizing a Data Science Conference

- To facilitate professional networking and collaboration, you planned to seat people in groups according to one of three research specialties: computer and/or database science, math and statistics, and machine learning.

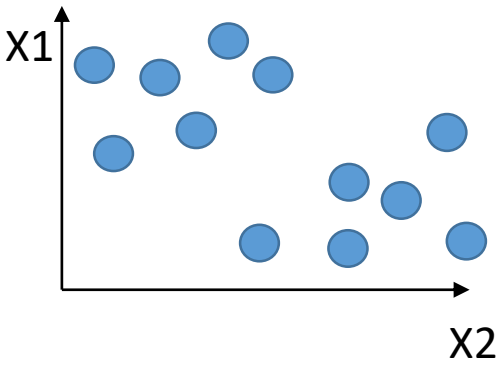




Source: Machine Learning with R

Clustering

- Clustering is an unsupervised machine learning task that automatically divides the data into **clusters,** or groups of similar items.



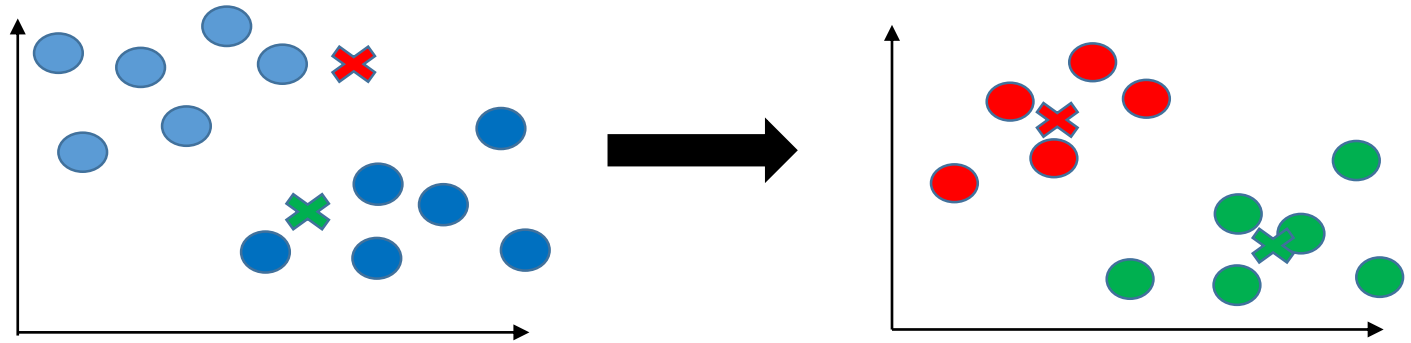
- Clustering is guided by the principle that items inside a cluster should be very similar to each other

- Clustering is used for knowledge discovery rather than prediction
  - It provides an insight into the natural groupings found within data.

K-means Clustering

MSIT MUT

- *K*-means clustering partition training examples into a pre-specified number of clusters.

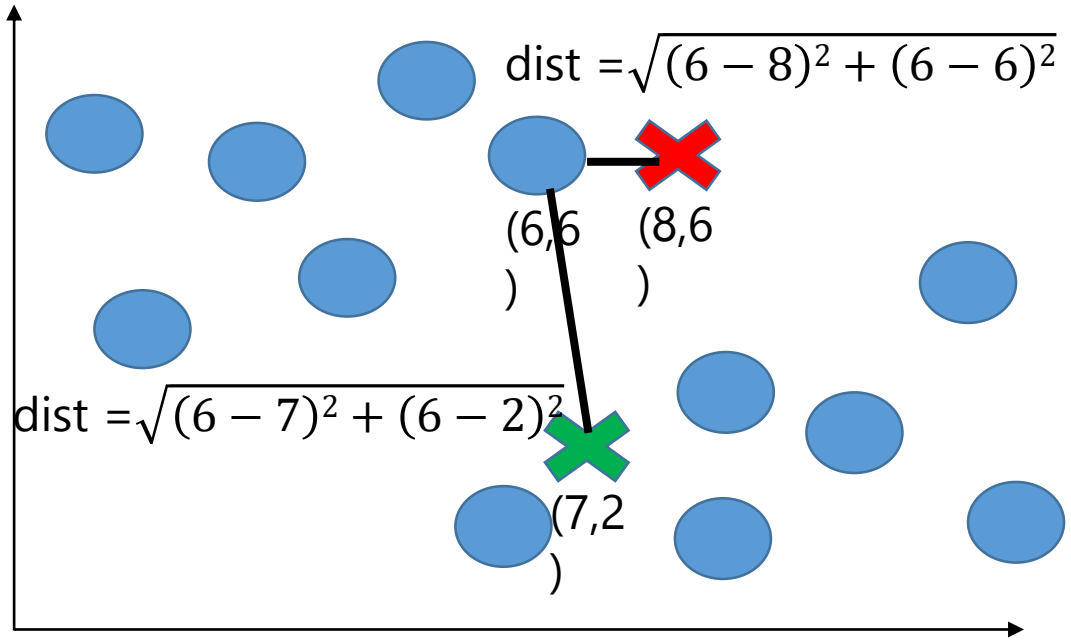


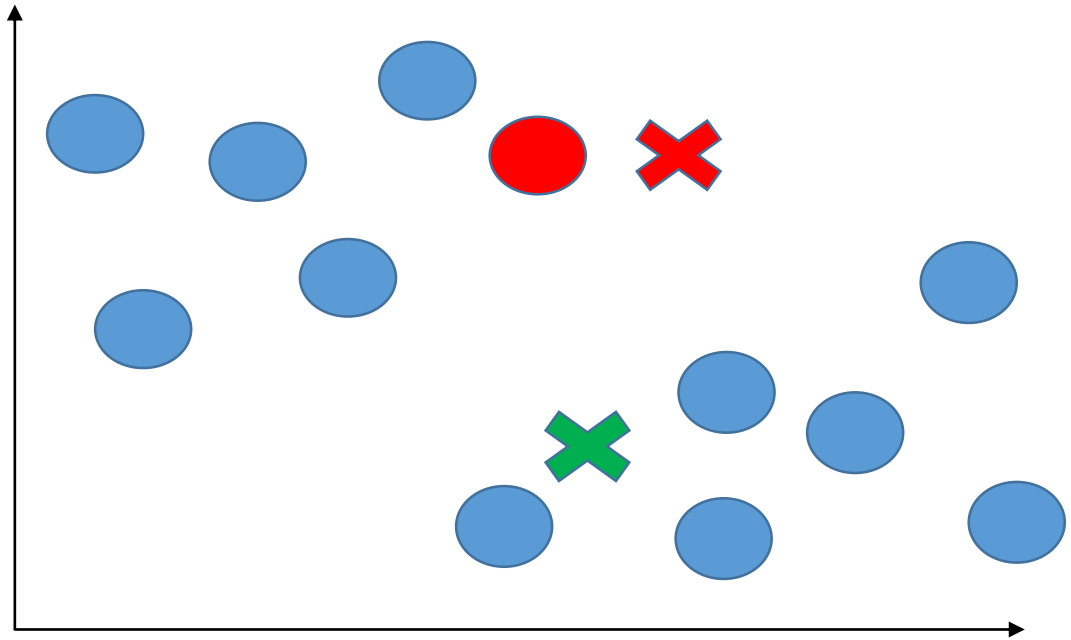
- K-means is consists of 2 phases
  - Cluster assignment
    - assign each example to one of the *K* clusters
  - Updates cluster centroids
    - adjusting centroids of clusters

* The process of updating and assigning occurs several times until changes no longer improve the cluster fit.

Iteration 1

- Cluster assignment

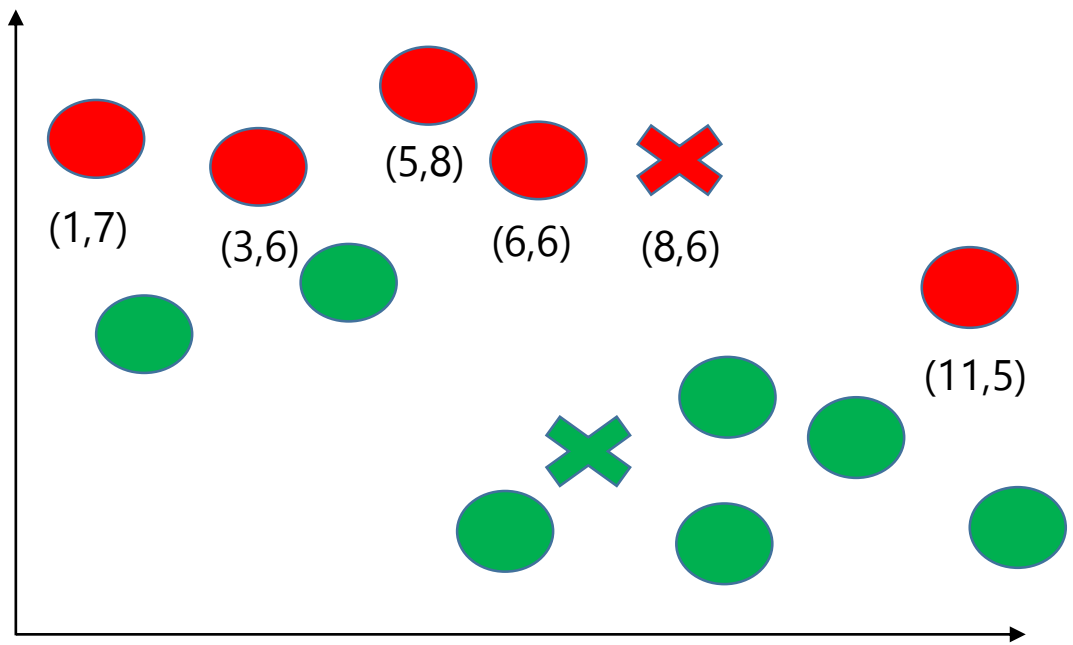


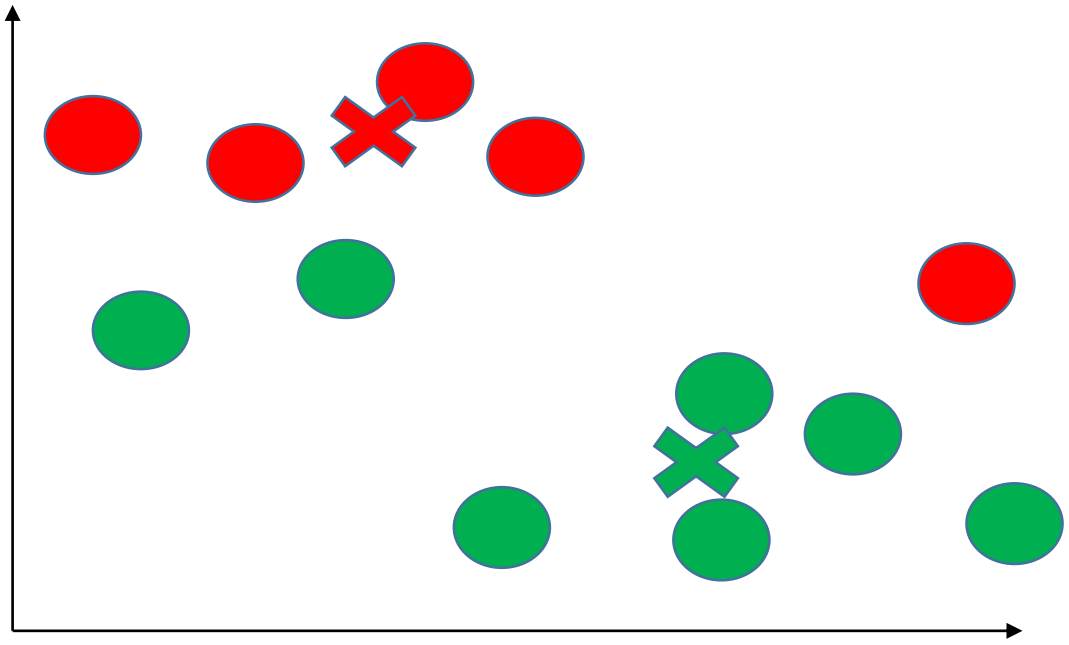


$$\text{Euclidean distance } d(p,q) = d(q,p) = \sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-pn)^2}$$

Iteration 1

- Update cluster centroids Centroid (red) $= (\frac{(1+3+5+6+11)}{5}, \frac{(7+6+8+6+5)}{5})$

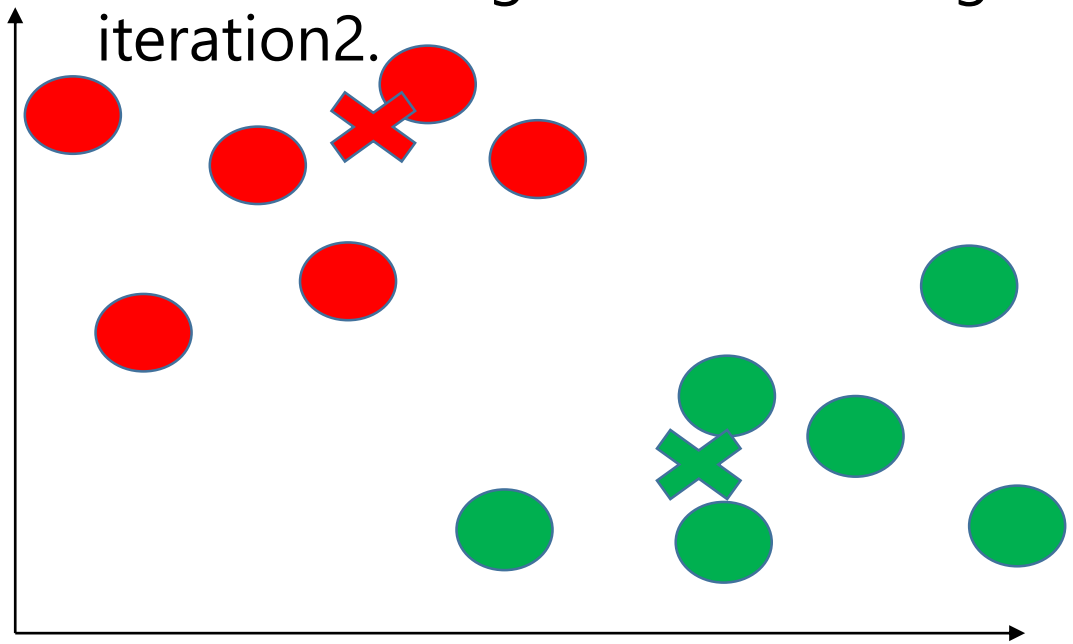




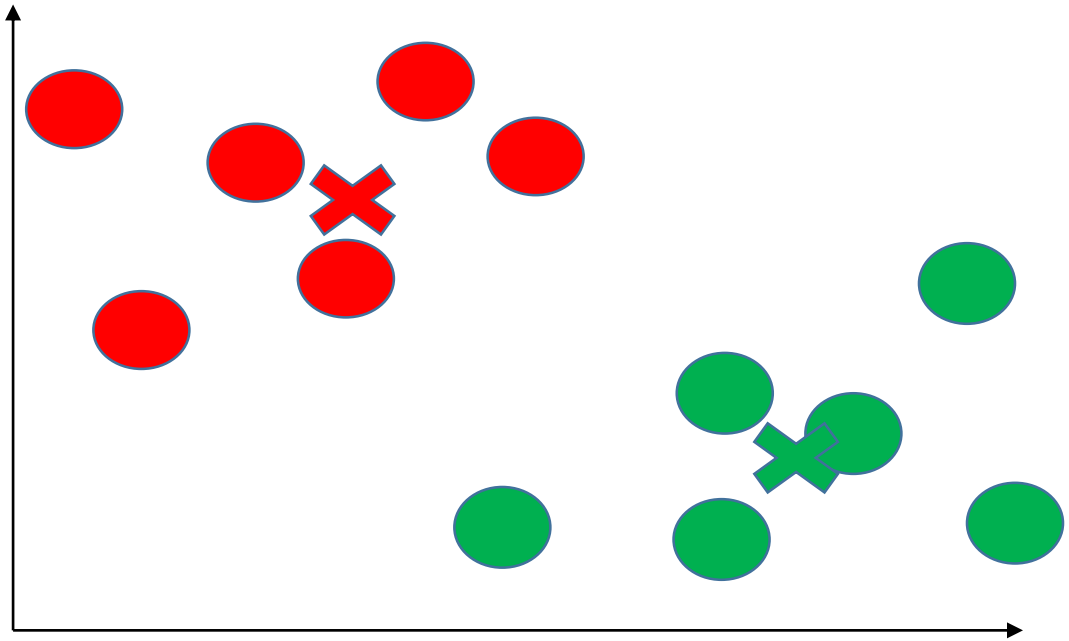
Iteration 2

MSIT MUT

- The same process of iteration1
  - Since no change in cluster assignment occurs, K-means stops at iteration2.



Cluster assignment



Update cluster centroids

AGENDA

Lecture (9.00 – 10.30)
- Machine Learning Overview

Lecture (10.45 – 12.00)
- Applied Analytics through Case Studies

Lecture (13.00 – 16.00)
- Demo – Insurance Case
- Model Assessment

MSIT MUT

Demo

MSIT MUT

- Insurance Case Study
  - Which factors influence medical expenses charged per year?
  - How many expenses should the insurance company should charge individual customers per year?

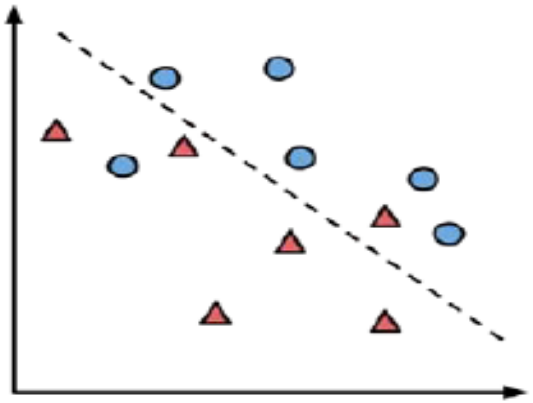

```
> str(insurance)
'data.frame':   1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8 25.8 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2$
 $ expenses: num  16885 1726 4449 21984 3867 ...
```

MSIT MUT

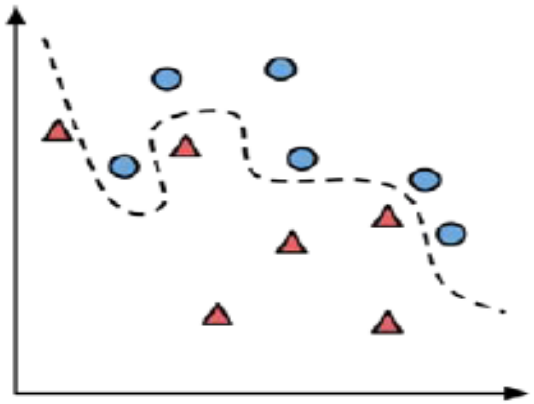
Evaluating Model Performance

MSIT MUT

- Evaluate how well a learned hypothesis can be expected to perform on new training examples
  - A hypothesis may have a low error for the training examples but still be inaccurate to predict output
  - A hypothesis may not even have a low error for the training examples and therefore be inaccurate to predict output



Underfit



Overfit

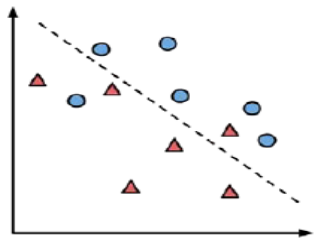
Source: Picture from Machine Learning with R Second Edition, Brett Lantz)

MSIT MUT

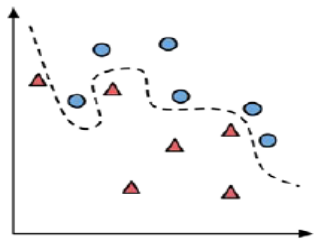
Bias versus Variance

MSIT MUT

- Every model has bias and variance error components
  - **Bias** component comes from erroneous assumptions of chosen learning algorithms (**underfitting problems**)
  - **Variance** component comes from sensitivity to change in a model (**overfitting problems**)



Underfit



Overfit

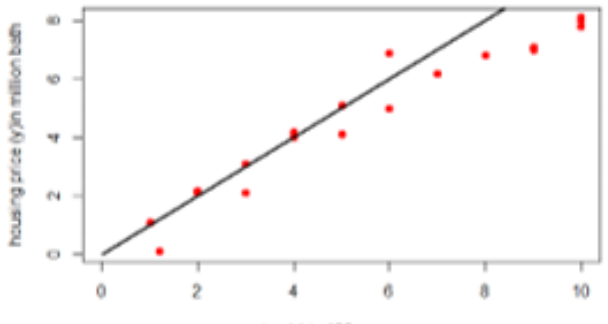
- **Bias and Variance Trade-off**
  - Bias and Variance are inversely related to each other
    - Balancing between both components is an ideal solution (**low bias and low variance**)

MSIT MUT

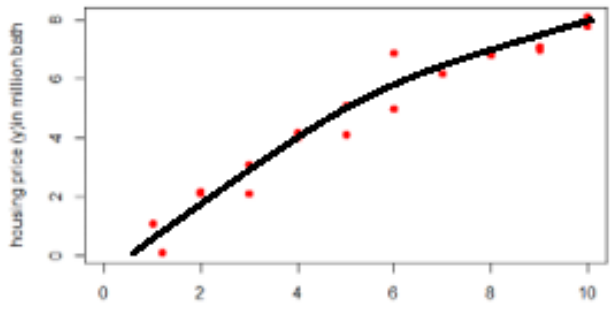
Example

MSIT MUT

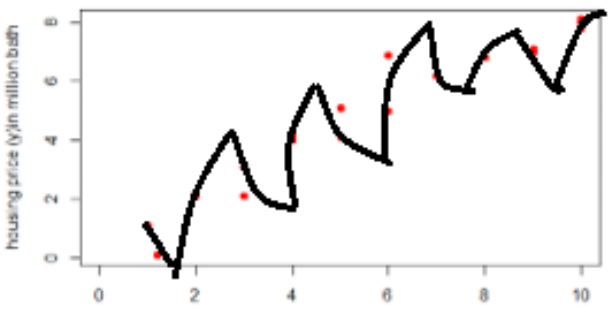
- **Regression Problems**



$$h_\theta(x) = \theta_0 + \theta_1 x$$

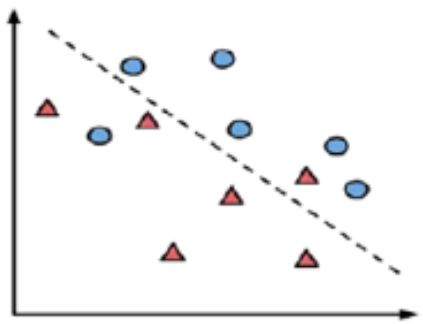


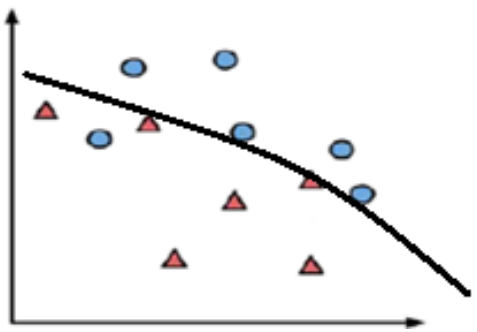
$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$

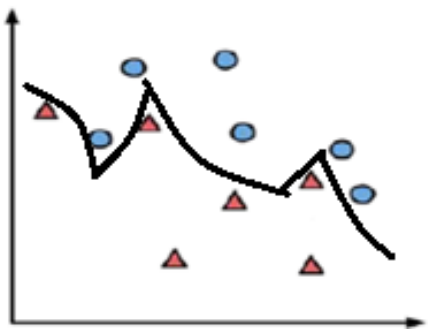


$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

- **Classification Problems**

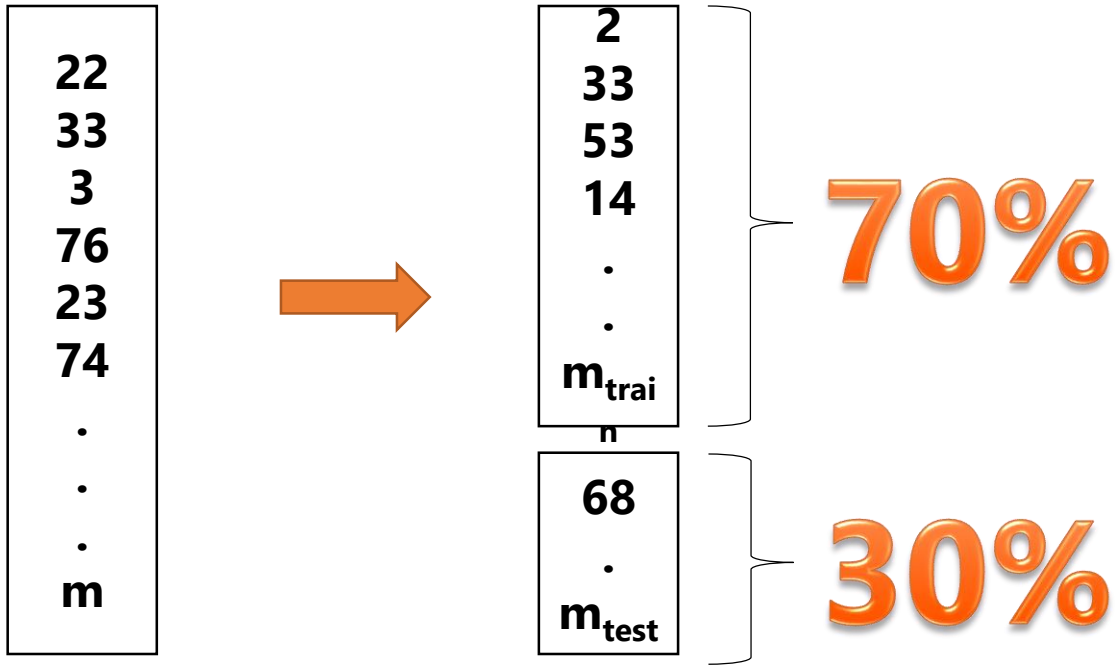






Evaluating a Model

- Splitting training examples into 2 parts – train and test datasets
  - randomly 70-30 or 80-20 into train and test dataset



Splitting Training Examples into 2 Parts

- The hypothesis is fit on the training set, and its performance is evaluated on the test set:
  - Learn θ and minimize $J_{train}(\theta)$ using a training set
  - Compute test error $J_{test}(\theta)$
    - For linear regression

$$J_{test}(\theta) = \frac{1}{m_{test}} \sum_{i=1}^{m_{test}} (h_\theta(x_{test}^{(i)}) - y_{test}^{(i)})^2$$

    - For KNN
      - simple approximately misclassification error

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{Err}_i \qquad \mathrm{Err}_i = I(y_i \neq \hat{y}_i)$$

MSIT MUT

Problems with Splitting Data into 2 parts

- Consider the case of underfitting
- Model Selection
  - select the appropriate level of flexibility for a model

$$h_\theta(x) = \theta_0 + \theta_1 x$$

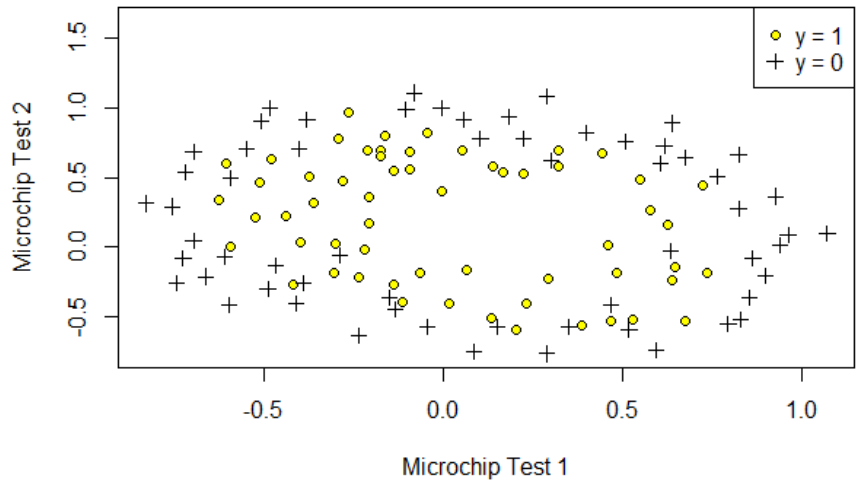
$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

..

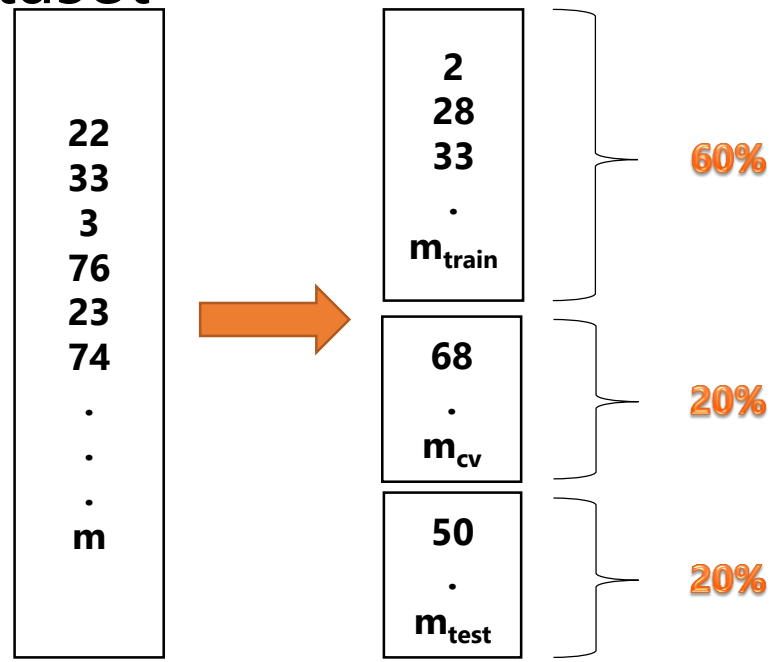
$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + ... + \theta_{10} x^{10}$$



How does the model generalize?

Evaluating a Model

- Splitting training examples into 3 parts – train, cross validation and test datasets (Machine Learning World – not for statistics!)
  - randomly 60-20-20 or 50-25-25 into train, cross validation and test dataset



Splitting Training Examples into 3 parts

MSIT MUT

- The hypothesis is fit on the training set, and its performance is evaluated on the test set:
  - Optimize parameters θ using training set for each polynomial degree
    - Goal: min $J(\theta)$ = $$J_{train}(\theta) = \frac{1}{m_{train}} \sum_{i=1}^{m_{train}} (h_\theta(x_{train}^{(i)}) - y_{train}^{(i)})^2$$
  - Find polynomial degree d with the least error using cross validation set

$$J_{cv}(\theta) = \frac{1}{m_{cv}} \sum_{i=1}^{m_{cv}} (h_\theta(x_{cv}^{(i)}) - y_{cv}^{(i)})^2$$

  - Estimate the generalization error using test set wit $J_{test}(\theta^{(i)})$, $\theta(i)$= theta from polynomial with lowest error)

$$J_{test}(\theta) = \frac{1}{m_{test}} \sum_{i=1}^{m_{test}} (h_\theta(x_{test}^{(i)}) - y_{test}^{(i)})^2$$

MSIT MUT

Model Improvement

- Advanced strategies may be needed to improve a model's performance

- Some Strategies
  - Collect more examples
  - Try smaller set of features
  - Try adding relevant features
  - Try higher degree of features
  - Increasing regularization parameter
  - decreasing regularization parameter

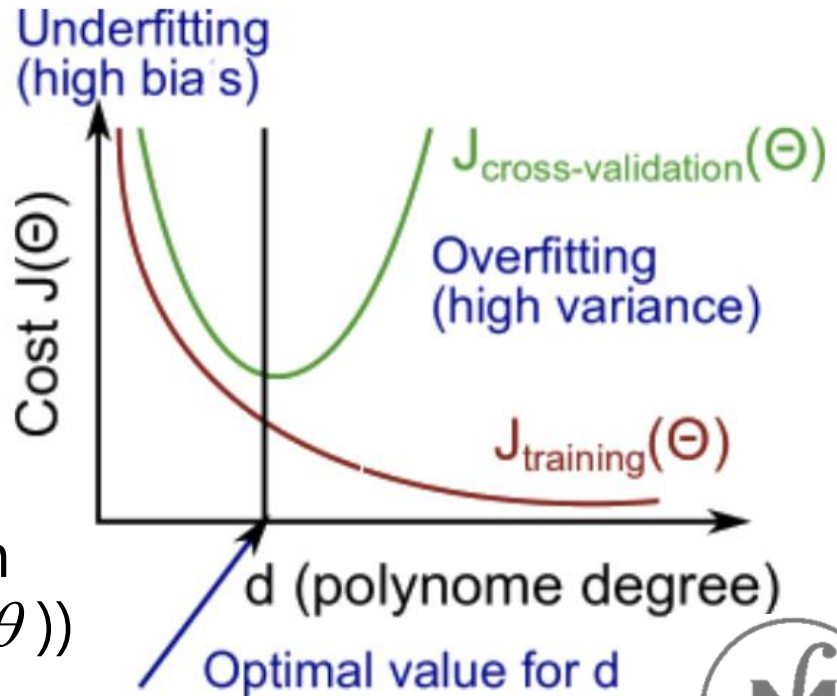
-> ***Goal*** *: perform* a test to gain insight what is/isn't work with a learning algorithms so as to correctly improve the model's performance

-> **Diagnose Bias or Variance**

Bias or Variance Problem?

MSIT MUT

- Suppose our learning hypothesis is bad predictions: bias or variance problem?
  - High bias (underfitting):
    - both $J_{train}(\theta)$ and $J_{cv}(\theta)$ will be high
  - High variance (overfitting):
    - $J_{train}(\theta)$ will be low and $J_{cv}(\theta)$ will be high (muchmore than $J_{train}(\theta)$)



Source: Figure by Andrew Ng

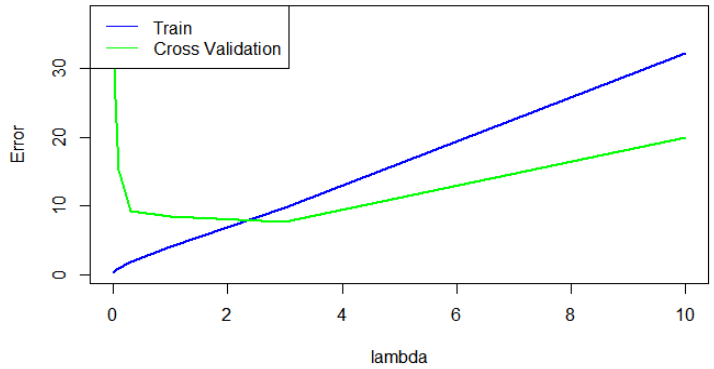
MSIT MUT

Overcome Overfitting

- Reduce a number of features
  - Select only relevant features
  - Decrease degree of polynomial

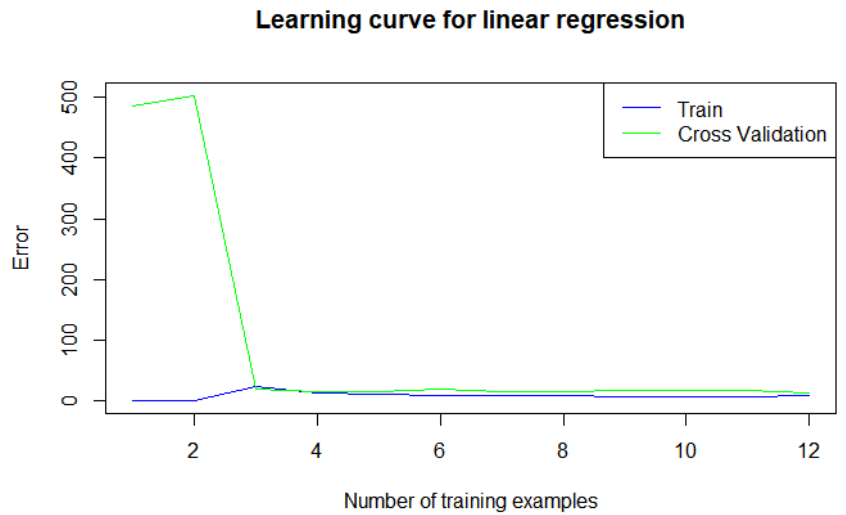
- Regularization
  - Keep all features (each contributes a bit to predict output) but *constrains* or *regularizes* the parameters to small values (many cases towards zero but not zero).
  - Regularized Linear Regression
    - Cost Function

$$J(\theta) = \frac{1}{m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\theta_j{}^2\right]$$



การตรวจสอบ Bias vs Variance ด้วย Learning Curve

- Learning Curve เป็นกราฟที่เราใช้สำหรับตรวจสอบ Model ว่ามีลักษณะ High Bias หรือ High Variance โดยจะทำการ plot cost ที่เกิดขึ้นใน train dataset (error_train) และ cross validation dataset (error_cv) ตามจำนวน training example ที่เพิ่มขึ้น

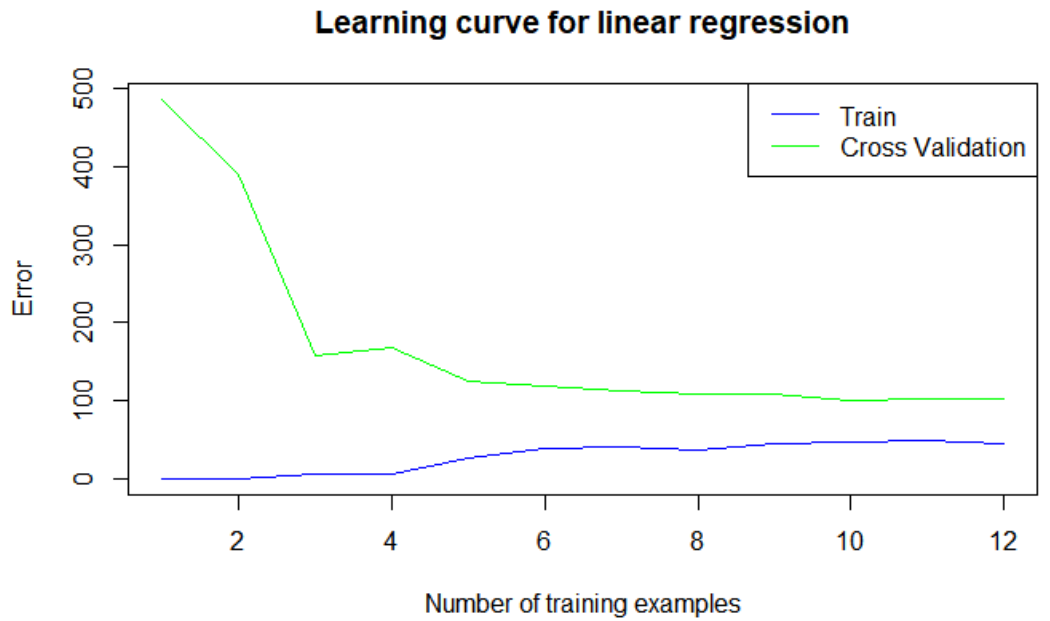


Model ที่ดีที่ Balance เรื่อง Bias (low) และ Variance (low)

MSIT MUT

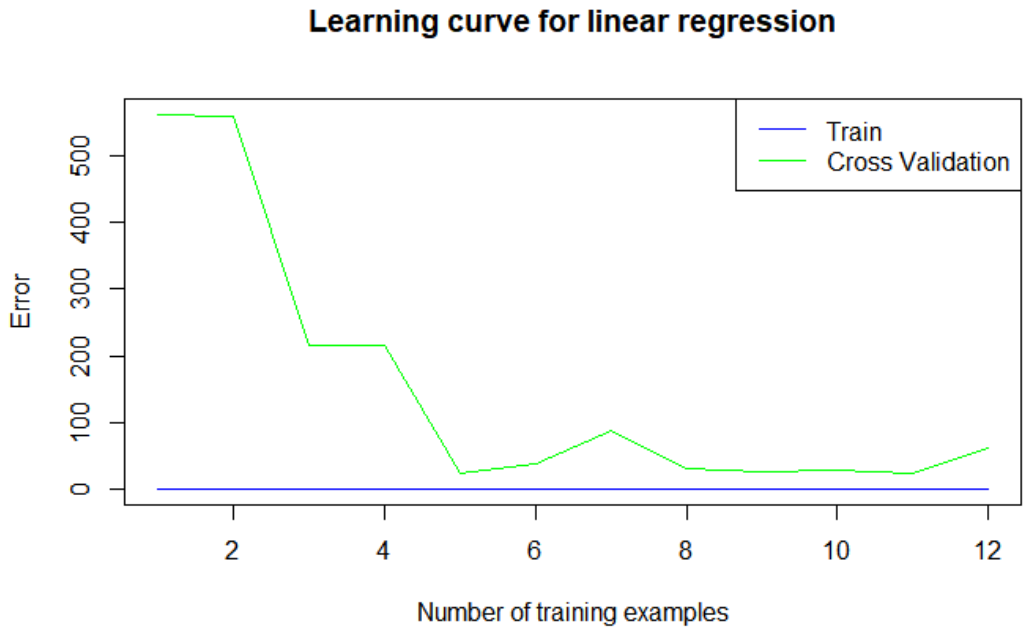
Learning Curve – High Bias

- ถ้าเส้น error train และ cross validation มีค่าสูงทั้งคู่ Model จะมีลักษณะ High Bias เพราะ จาก error ที่เกิดขึ้น แสดงว่า Model ไม่สามารถอธิบายข้อมูลทั้งที่นำมาทดสอบและข้อมูลกลุ่มตัวอย่างใหม่ได้



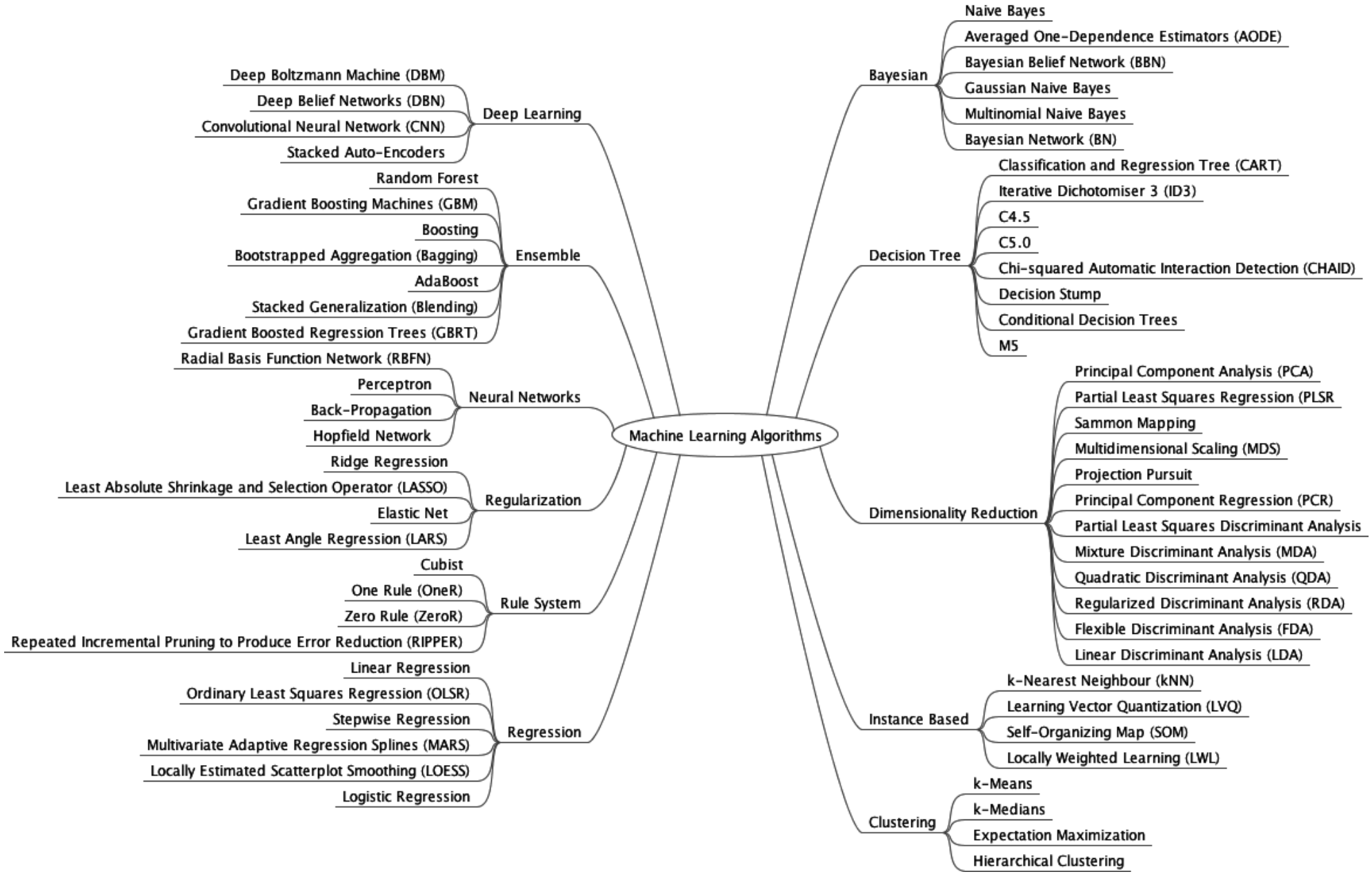
Learning Curve – High Variance

- ถ้าเส้น error จาก train มีค่าต่ำ และ error จาก cross validation มีค่าสูง Model จะมีลักษณะ High Variance เพราะ จาก error ที่เกิดขึ้น แสดงว่า Model จำรูปแบบ training examples มากเกินไป ทำให้อธิบายข้อมูลกลุ่มตัวอย่างใหม่ไม่ดี



วิธีการในการ Improve Model

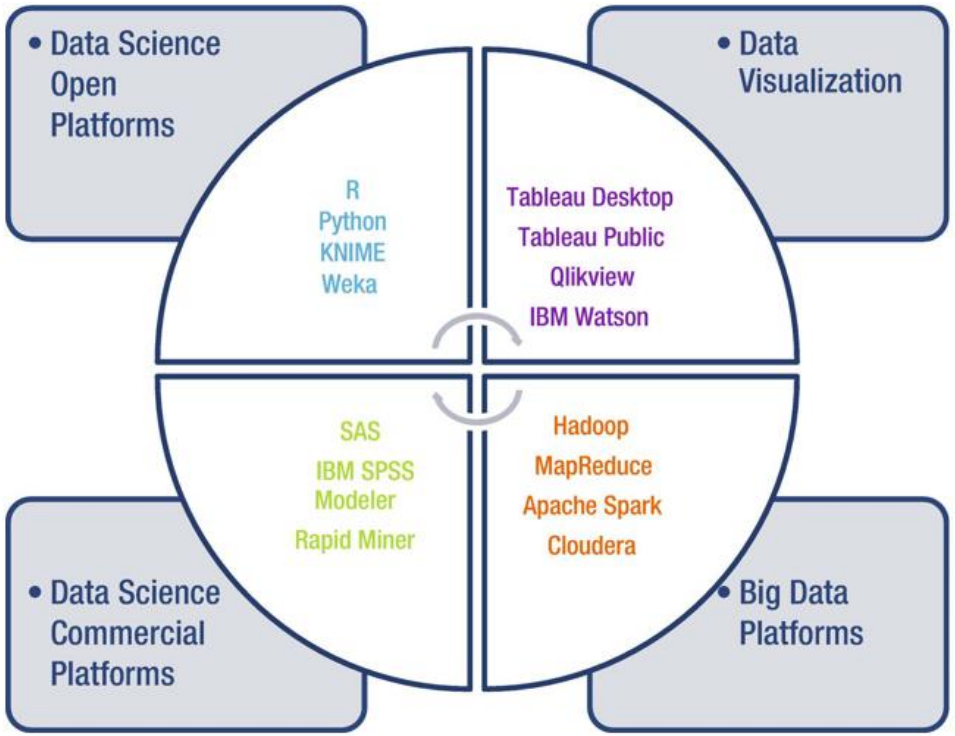
	Fix High Bias	Fix High Variance
การเพิ่มจำนวน training examples		★
การเพิ่มจำนวน input features	★	
การเพิ่ม degress of polynomial	★	
การลดจำนวน input features		★
การเพิ่มค่า regularized parameters		★
การลดค่า regularized parameters	★	


Machine Learning Algorithms
Deep Learning
Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders
Ensemble
Random Forest
Gradient Boosting Machines (GBM)
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (Blending)
Gradient Boosted Regression Trees (GBRT)
Neural Networks
Radial Basis Function Network (RBFN)
Perceptron
Back-Propagation
Hopfield Network
Regularization
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least Angle Regression (LARS)
Rule System
Cubist
One Rule (OneR)
Zero Rule (ZeroR)
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
Regression
Linear Regression
Ordinary Least Squares Regression (OLSR)
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Logistic Regression
Bayesian
Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Gaussian Naive Bayes
Multinomial Naive Bayes
Bayesian Network (BN)
Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
C5.0
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
Conditional Decision Trees
M5
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLSR
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Principal Component Regression (PCR)
Partial Least Squares Discriminant Analysis
Mixture Discriminant Analysis (MDA)
Quadratic Discriminant Analysis (QDA)
Regularized Discriminant Analysis (RDA)
Flexible Discriminant Analysis (FDA)
Linear Discriminant Analysis (LDA)
Instance Based
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)
Clustering
k-Means
k-Medians
Expectation Maximization
Hierarchical Clustering

MSIT MUT

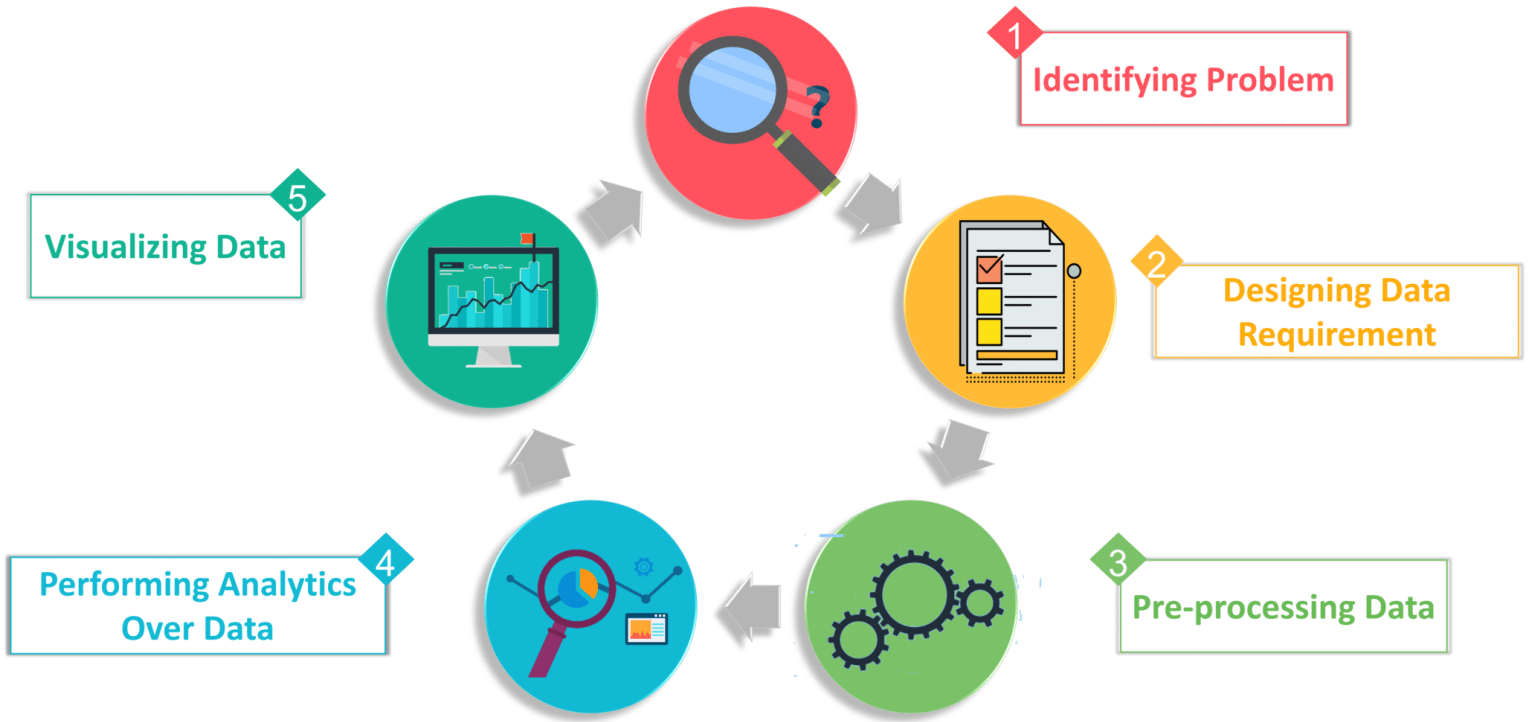
Programming for Data Analytics and Demo

- R Programming Language
  - A collection of R functions that can be shared among users is called a package
  - View a list at Comprehensive R Archive Network (CRAN) http://cran.r-project.org/index.html
- Python Programming Language
- Others programming platforms such as SAS



Source: Applied Analytics through case studies

How to Start a Project?



- Challenges
  - Successful businesses will be defined by their ability to collect and curate the right data and apply analytics in order to make insights actionable at the point of decision and make better decisions.

Top 10 Skills ที่ Data Scientist ต้องรู้

MSIT MUT

- Source: http://www.msit.mut.ac.th

Thank you

- Question?

MSIT MUT