Introduction to Data Science

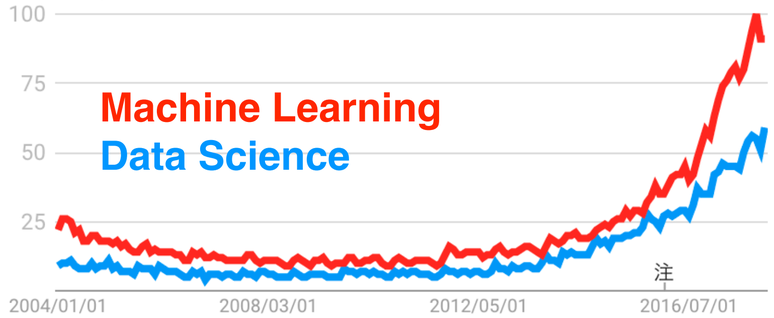
Part.1 Background
Part.2 Basic statistics
Part.3 Regression

Takuma Kawahara
Jan. 25th 2018

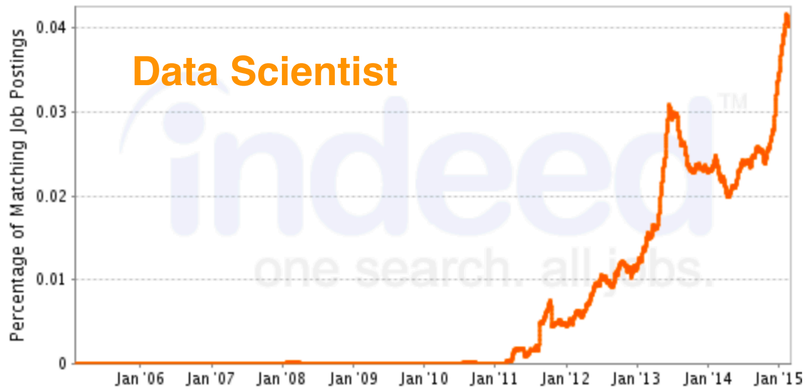
Part.1: Background

Data scientist

Google Trends



Job Trends from indeed.com



https://www.quora.com/What-is-the-job-outlook-for-data-scientists-analysts-in-2015-and-beyond

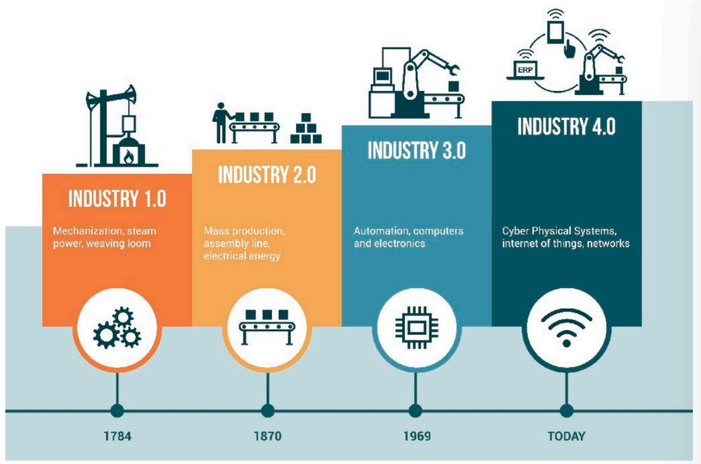
Media

Data Scientist: The Sexiest Job of the 21st Century

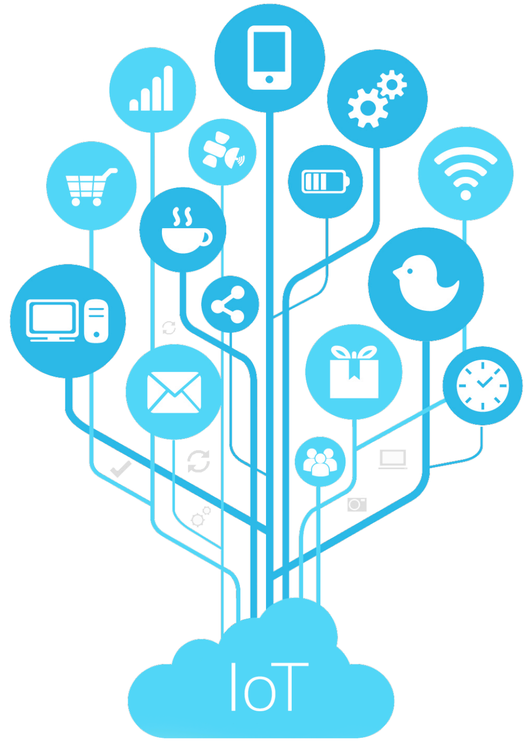
by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century



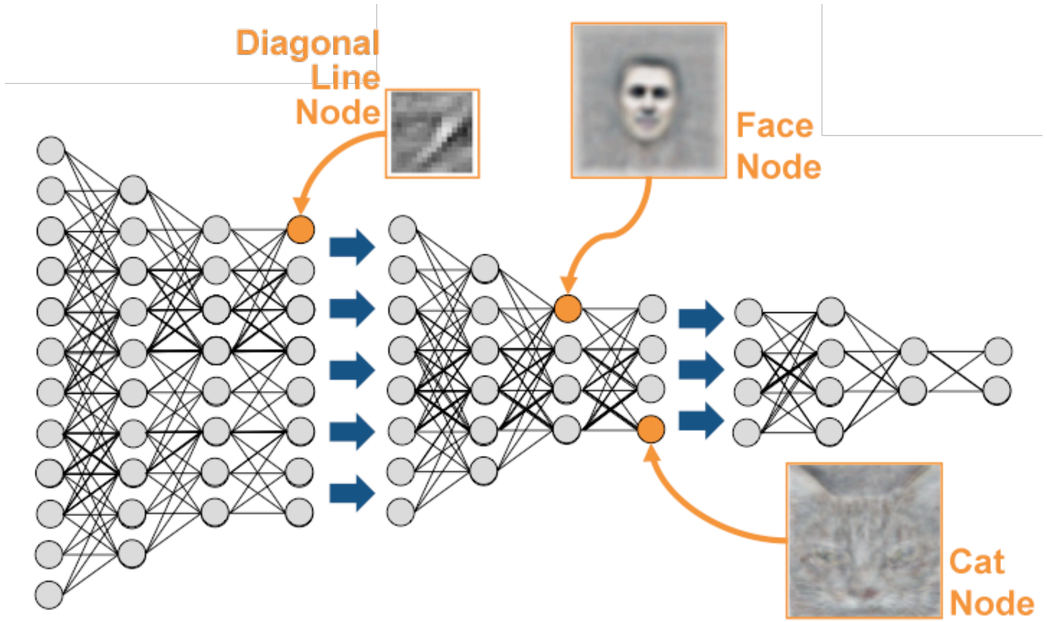
http://www.aberdeenessentials.com/opspro-essentials/industry-4-0-industrial-iot-manufacturing-sneak-peek/



http://codesfiction.com/what-is-iot/

Artificial Intelligence

Topics of AI

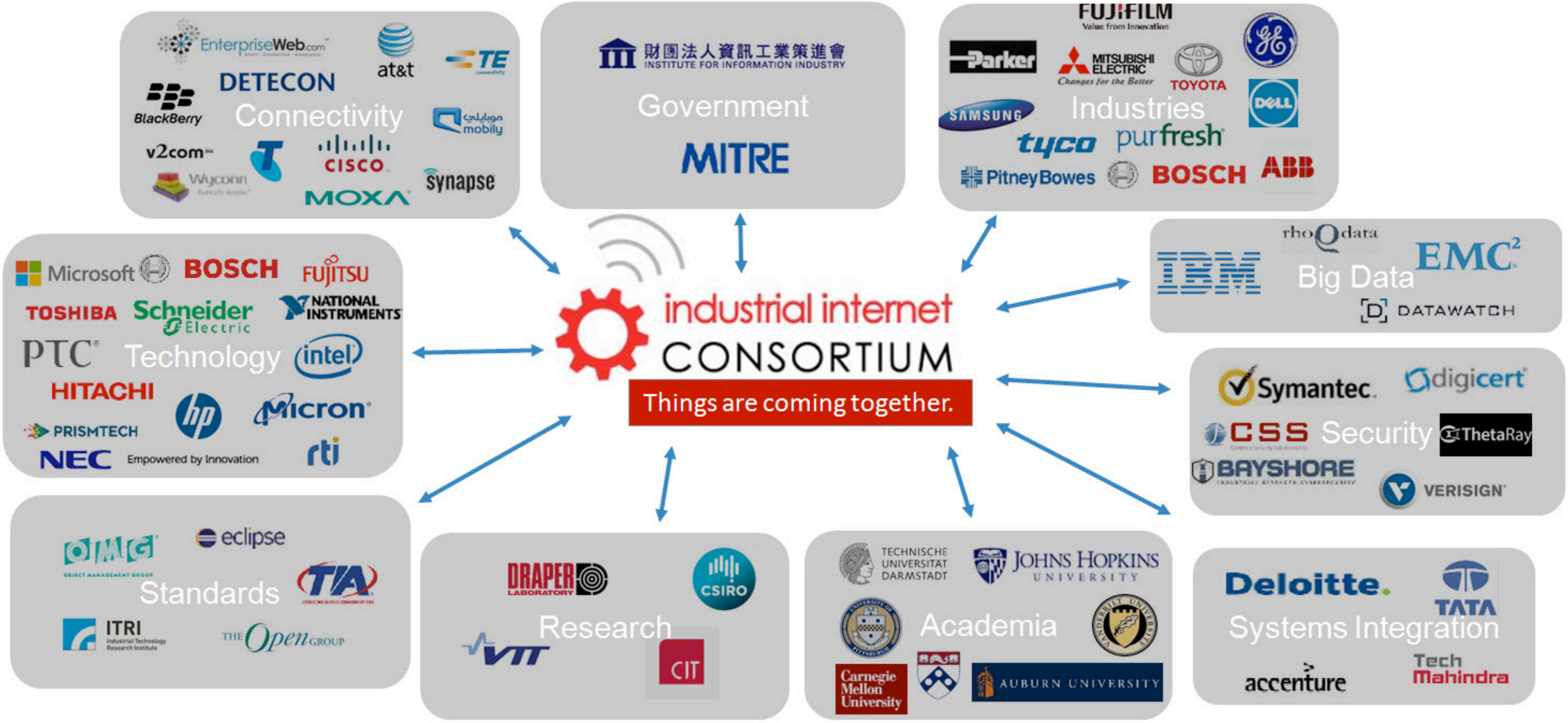


https://theanalyticsstore.ie/deep-learning/

https://deepmind.com/research/alphago/

https://cs.stanford.edu/people/esteva/nature/

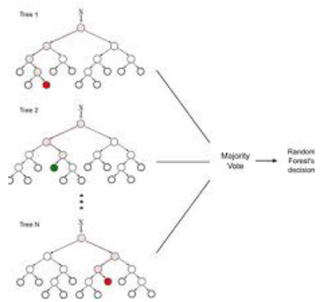
Industrial internet

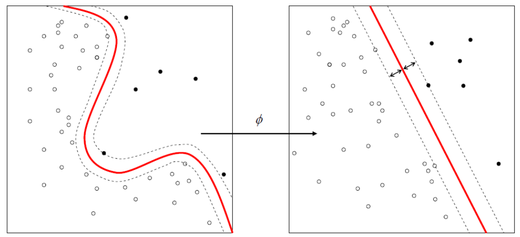


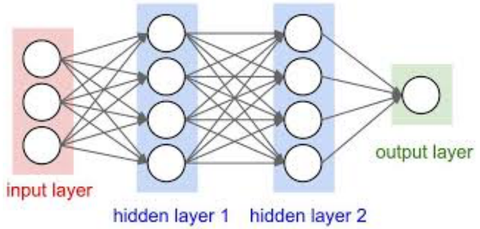
http://www.infiniteinformationtechnology.com/sap-se-company-and-the-industrial-internet-consortium

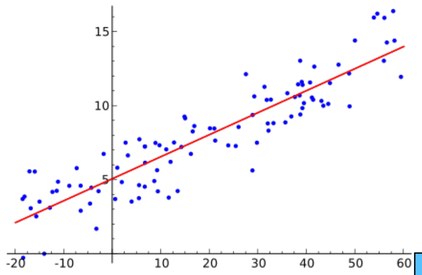
AI, MachineLearning & Regression

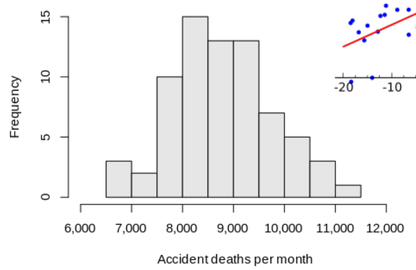
Regression is first step to Machine Learning and AI.











complex

simple

Deep Learning

Support vector machine

Random forest

PLS regression, Lasso, Ridge, Elastic net

Generalized linear model

Multiple linear regression, Principal component analysis

Simple linear regression

Fundamental statistics (Mean, Standard deviation, Histogram, Scatter plot)

Part.2: Basic statistics

Fundamental statistics

Data set x_i $x = \{1, 3, 3, 6, 7, 8, 9\}$

Mean

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = 5.29$$

Median

The value separating the higher half of a data sample, a population, or a probability distribution, from the lower half.

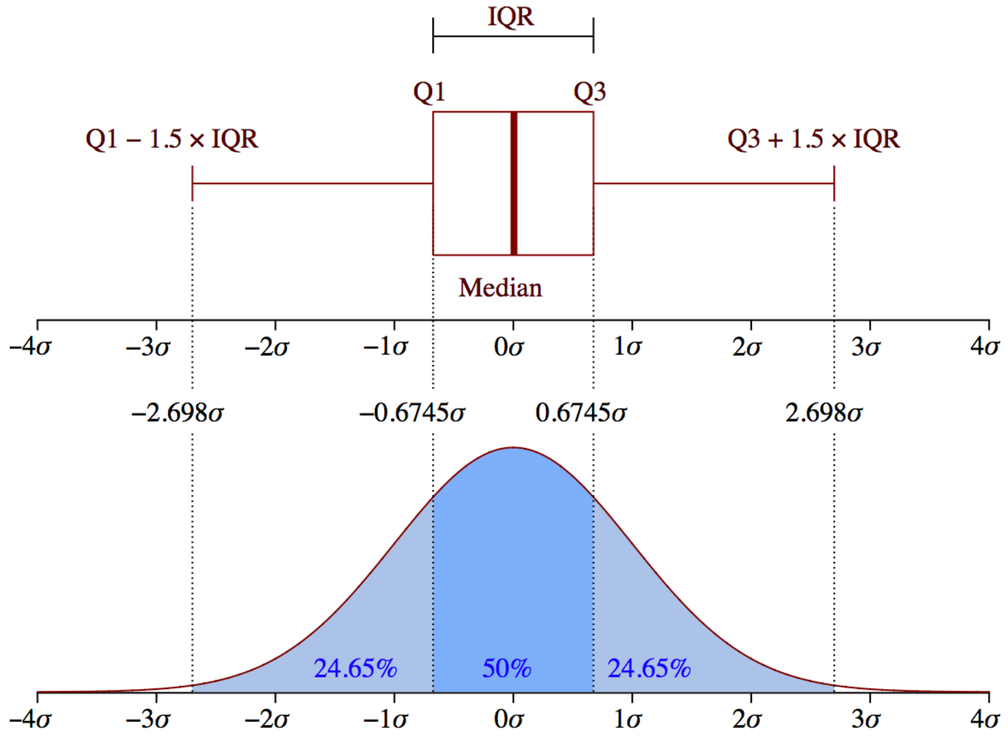
$m = 6$

Standard deviation:σ

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \qquad \sigma = 2.76$$

Variance

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad \sigma^2 = 7.63$$



https://en.wikipedia.org/wiki/Probability_density_function

First quartile $Q1 = 3$

Third quartile $Q3 = 7.5$

Types of variable

There are four types of variable as below. We need to pay attention to type of variable.

		Available Calc.	Example
Quantitateve	Continuous	>, < , +, -, x, ÷	10m/s, 2.7g, 3.22mm
	Discrete	>, < ,+, -	12deg.C, Jan-12th
Categorical	Ordinal	>, <	Like/Dislike, High/Mid/Low
	Nominal	-	Dog, Cat, Man, Woman

Correlation coefficient

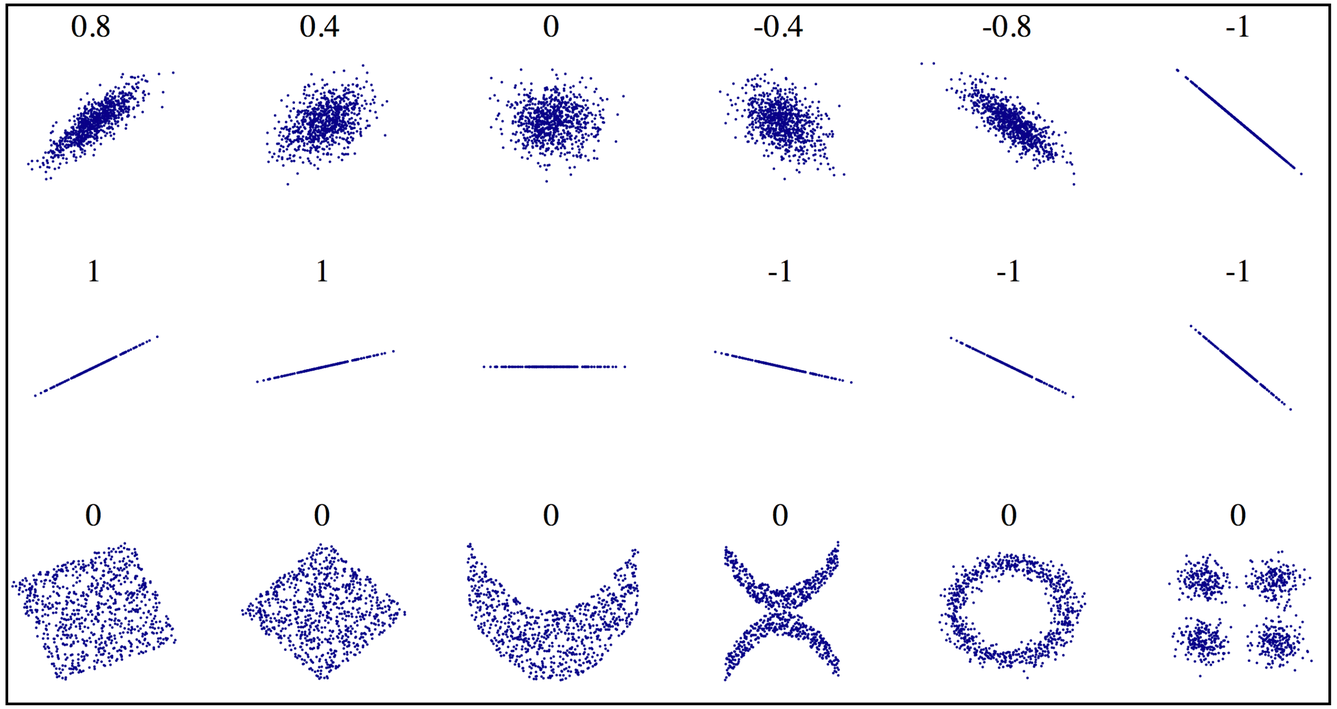
We can quantify the correlation between two variables using Correlation coefficient.

Correlation coefficient:r

In statistics, the Pearson correlation coefficient is a measure of the linear correlation between two variables *X* and *Y*.[1]

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2)(\sum_{i=1}^{n}(y_i - \bar{y})^2)}}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

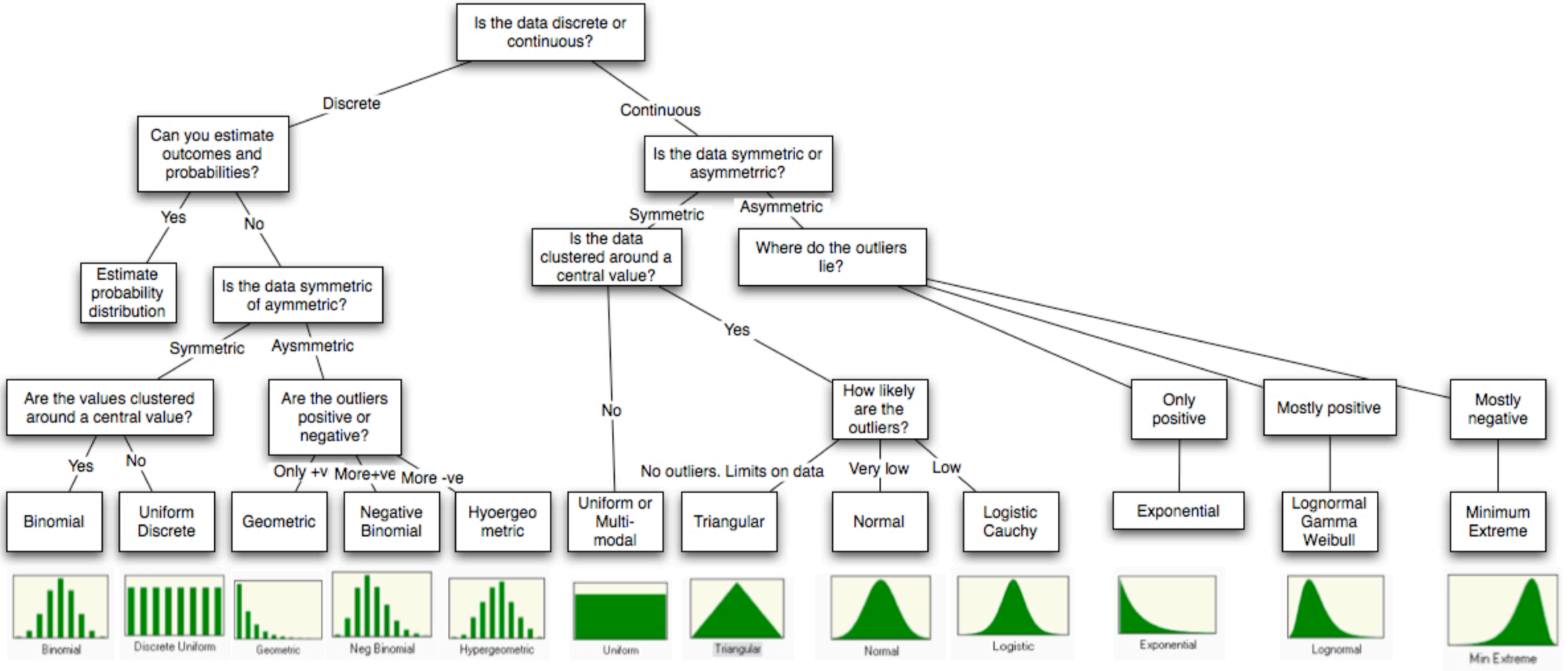


It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

[1]https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Statical Distribution

There are varies statical distribution based on type of data set.

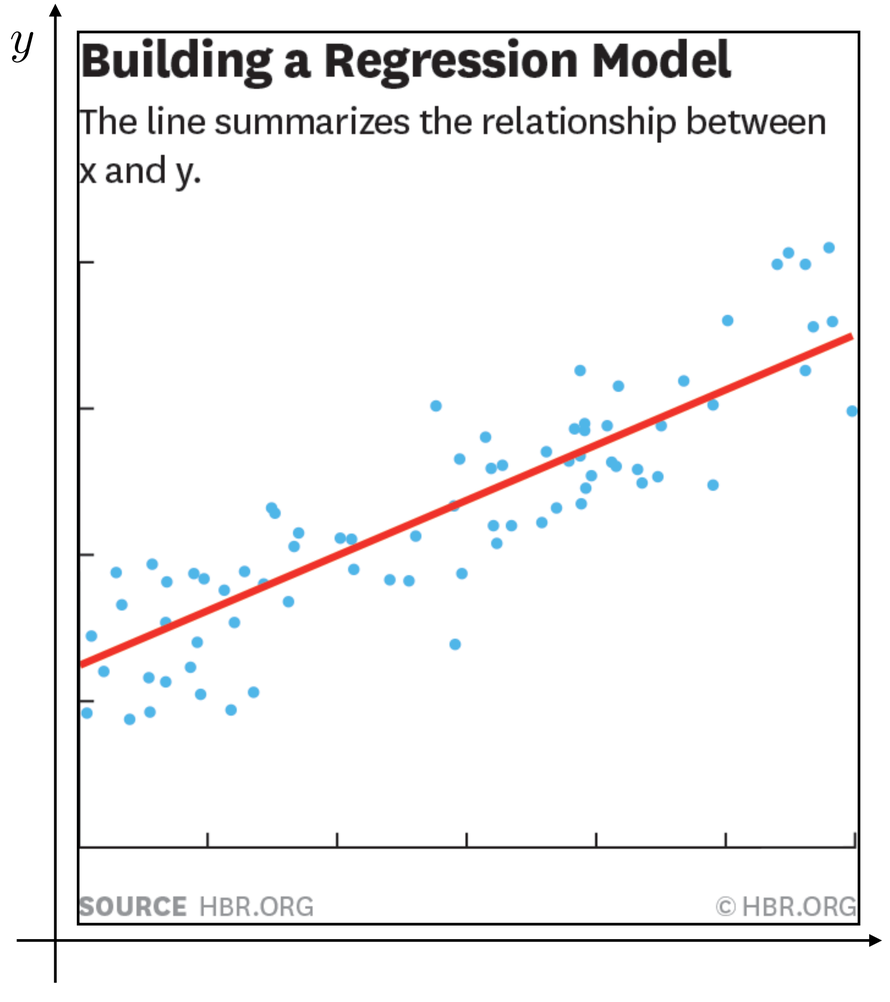


http://people.stern.nyu.edu/adamodar/pdfiles/papers/probabilistic.pdf

Part.3: Regression

Regression

Build model (train model) to explain actual data with small error.



https://hbr.org/2015/11/a-refresher-on-regression-analysis

$$y = f(\beta_i, x_i)$$

y : The dependent variable

x_i : The independent variable

β_i : Regression coefficient (also called parameter, weight)

Typical regression models

Linear regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n$$

Logistic regression

$$logit(y) = \ln \frac{y}{1-y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n$$

Poisson loglinear regression

$$\ln(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n$$

Data analysis process

Typical process of data analysis is as shown here.

#1. Prepare and clean data set

Set the dependent variable and collect many independent variables.

#2. Perform exploratory data analysis

Histgram, Scatter plot, Scatterplot matrix, Boxplot

#3. Data transformation

Data normalization. One-hot-encoding (dummy variable),
Split original data to “Training data”, (“Validation data”) & “Test data”

#4. Build and Train the model

Least squares, Maximum likelihood method, Variable selection,
Coefficient of determination (R^2), Stepwise

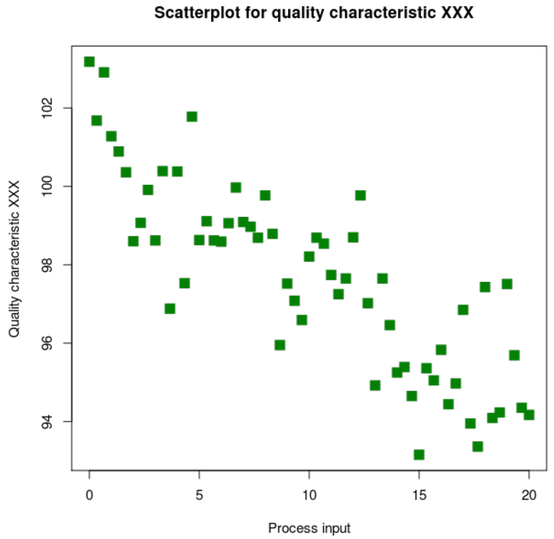
#5. Validate the model

RMSE, K-fold cross validation, Leave-one-out cross validation

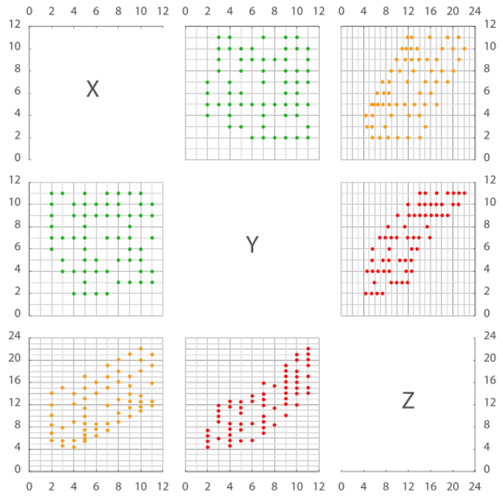
#2. Perform exploratory data analysis (EDA)

First step of data analysis is EDA. In this step, we summarize data with fundamental statistics (max, min, etc) and/or some visual methods (scatter plot, histogram, etc.).

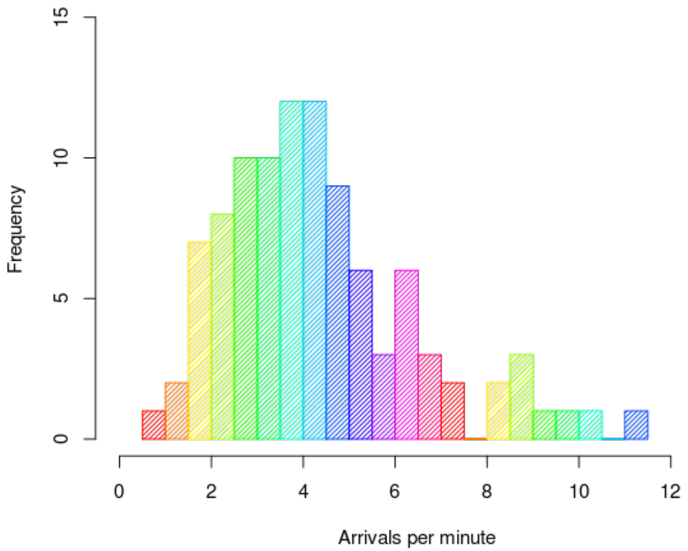
Scatter plot[1]



Scatter plot matrix[1]



Histogram[2]



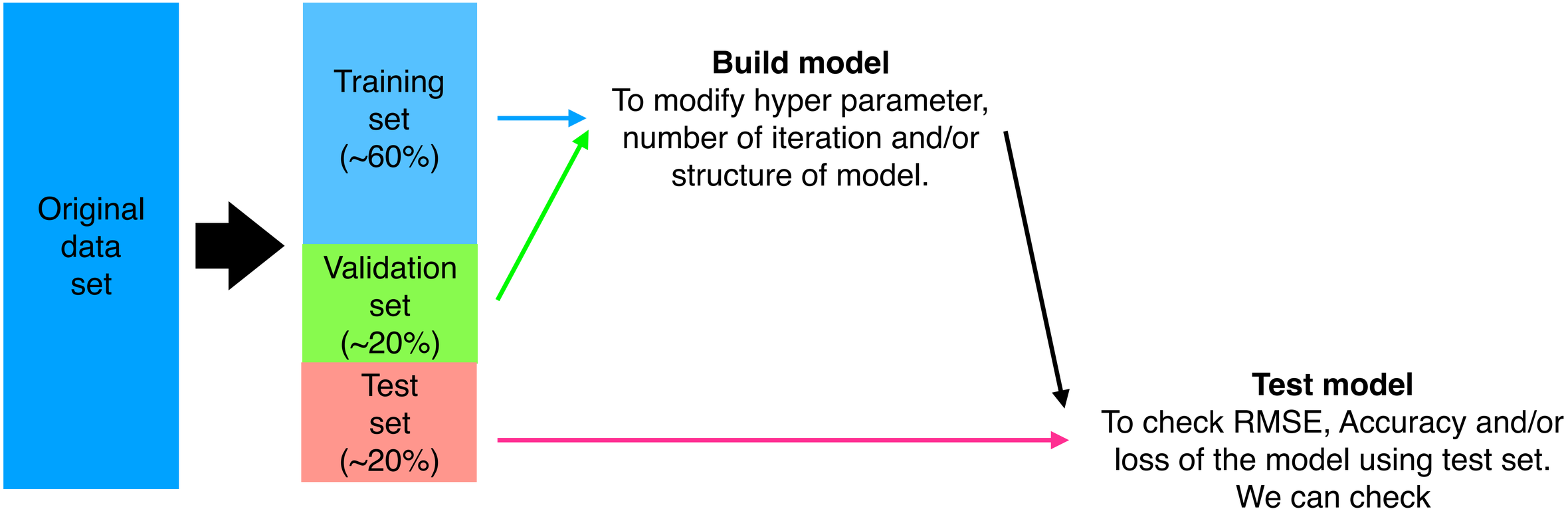
[1] https://en.wikipedia.org/wiki/Scatter_plot

[2] https://en.wikipedia.org/wiki/Histogram

#3. Data transformation

We use the “Training data” and “Validation data” to Build/Train the model and use the “Test data” to check performance of the model for *Unknown* data.

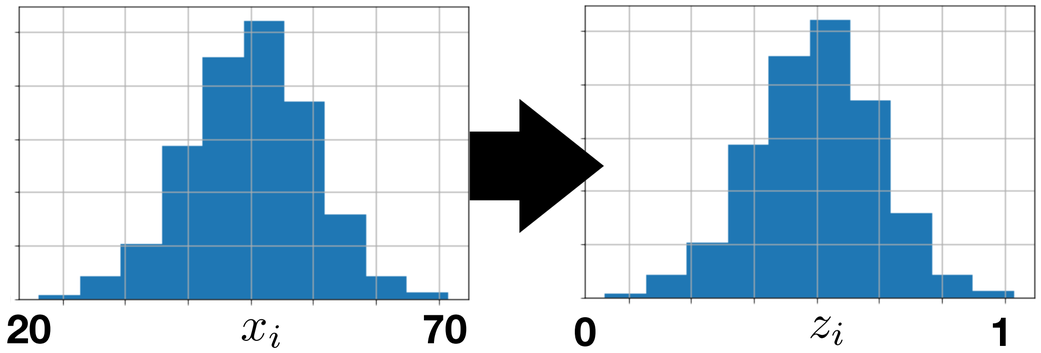
Split the Data set to Training, Validation and Test set.



#3. Data transformation

Data normalization is needed to compare coefficients of variables which have different scale.
One-hot-encoding is needed to use categorical variables for regression.

Data normalization



$$z_i = \frac{x_i - min(x)}{max(x) - min(x)}$$

One-hot-encoding (dummy variable)

Material
A
B
A
D
C
B
A

x_1	x_2	x_3
1	0	0
0	1	0
1	0	0
0	0	0
0	0	1
0	1	0
1	0	0

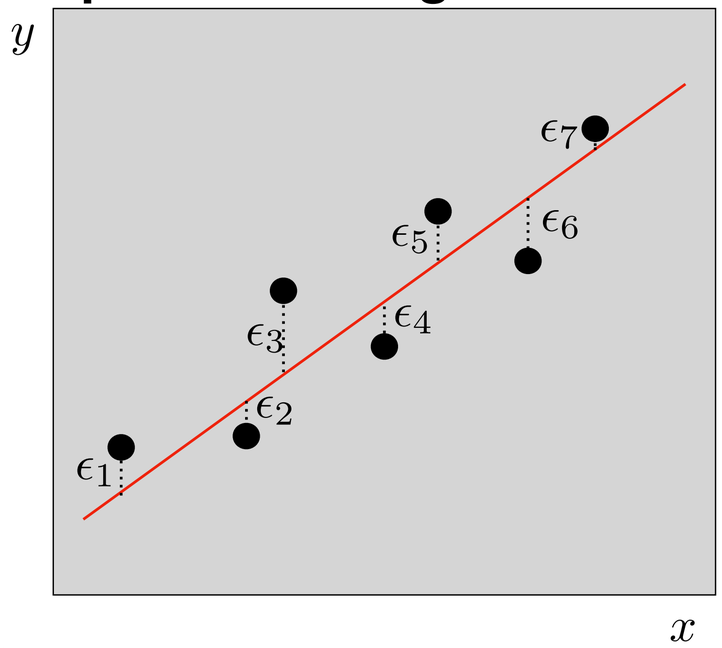
*x_1,x_2,x_3: Dummy variables

Change categorical variables to 0 or 1 number with “One-hot-encoding“

#4. Build and Train the model

We can get regression coefficient which can minimize error of model by “Least square method”.

Simple Linear Regression



Data set

$$x = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$$
$$y = \{y_1, y_2, y_3, y_4, y_5, y_6, y_7\}$$

Model

$$\hat{y} = \beta_0 + \beta_1 x$$

#1. Calculate error of model

$$\epsilon_i = (y_i - \hat{y}_i)^2 \longrightarrow \epsilon = \sum_{i=1}^{n} \epsilon_i$$

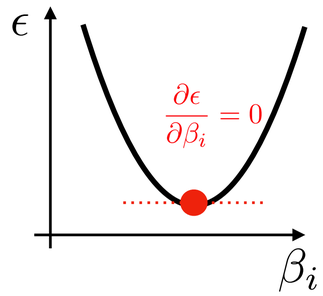
$\epsilon_i = |y_i - \hat{y}_i|$ is mathematically difficult to calculate to find minimum value.

#2. Minimize error by changing regression coefficient

Get coefficient by “Least square method”.

$$\frac{\partial \epsilon}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

$$\frac{\partial \epsilon}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$



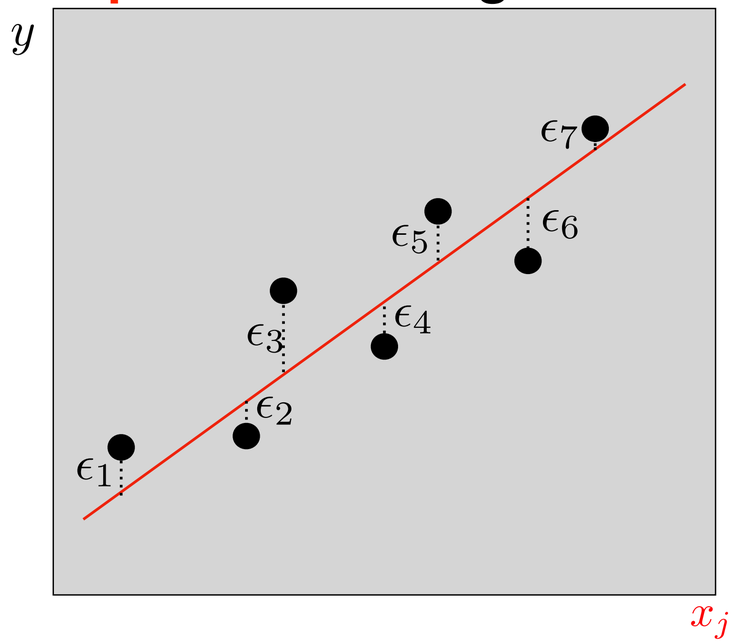
#3. Get regression coefficient

$$\beta_0 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \beta_1 = \bar{y} - \beta_0 \bar{x}$$

#4. Build and Train the model

We can get regression coefficient which can minimize error of model by “Least square method”.

Multipule Linear Regression



Data set

$$x_j = \{x_{1,j}, x_{2,j}, x_{3,j}, x_{4,j}, x_{5,j}, x_{6,j}, x_{7,j}\}$$

$$y = \{y_1, y_2, y_3, y_4, y_5, y_6, y_7\}$$

Model

$$\hat{y} = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$$

#1. Calculate error of model

$$\epsilon_i = (y_i - \hat{y}_i)^2 \longrightarrow \epsilon = \sum_{i=1}^{n} \epsilon_i$$

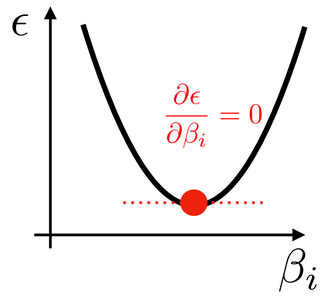
$\epsilon_i = |y_i - \hat{y}_i|$ is mathematically difficult to calculate to find minimum value.

#2. Minimize error by changing regression coefficient

Get coefficient by “Least square method”.

$$\frac{\partial \epsilon}{\partial \beta_m} = \frac{\partial}{\partial \beta_m}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{i,j})^2\right) = 0$$

$$-2\sum_{i=1}^{n} x_{i,m}(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{i,j}) = 0$$



#3. Get regression coefficient

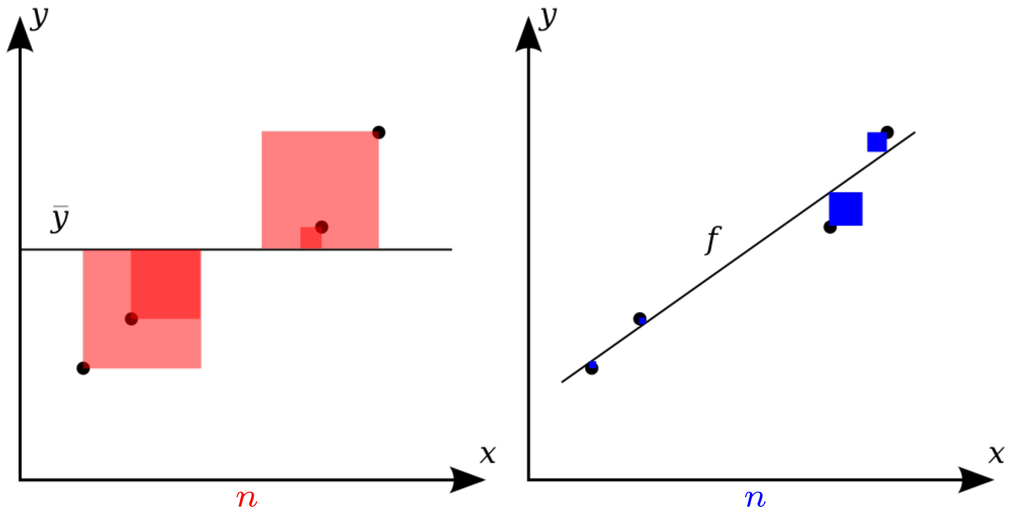
From #2, calculate k+1 equation and get coefficient.
This calculation can solve with the help of vectors and matrices.
Please find more detail in text (e.g. [1,2]).

$$\beta = (X^T X)^{-1} X^T y$$

[1]http://mezeylab.cb.bscb.cornell.edu/labmembers/documents/supplement%205%20-%20multiple%20regression.pdf
[2] http://fs1.law.keio.ac.jp/~aso/ecnm/pp/reg2.pdf

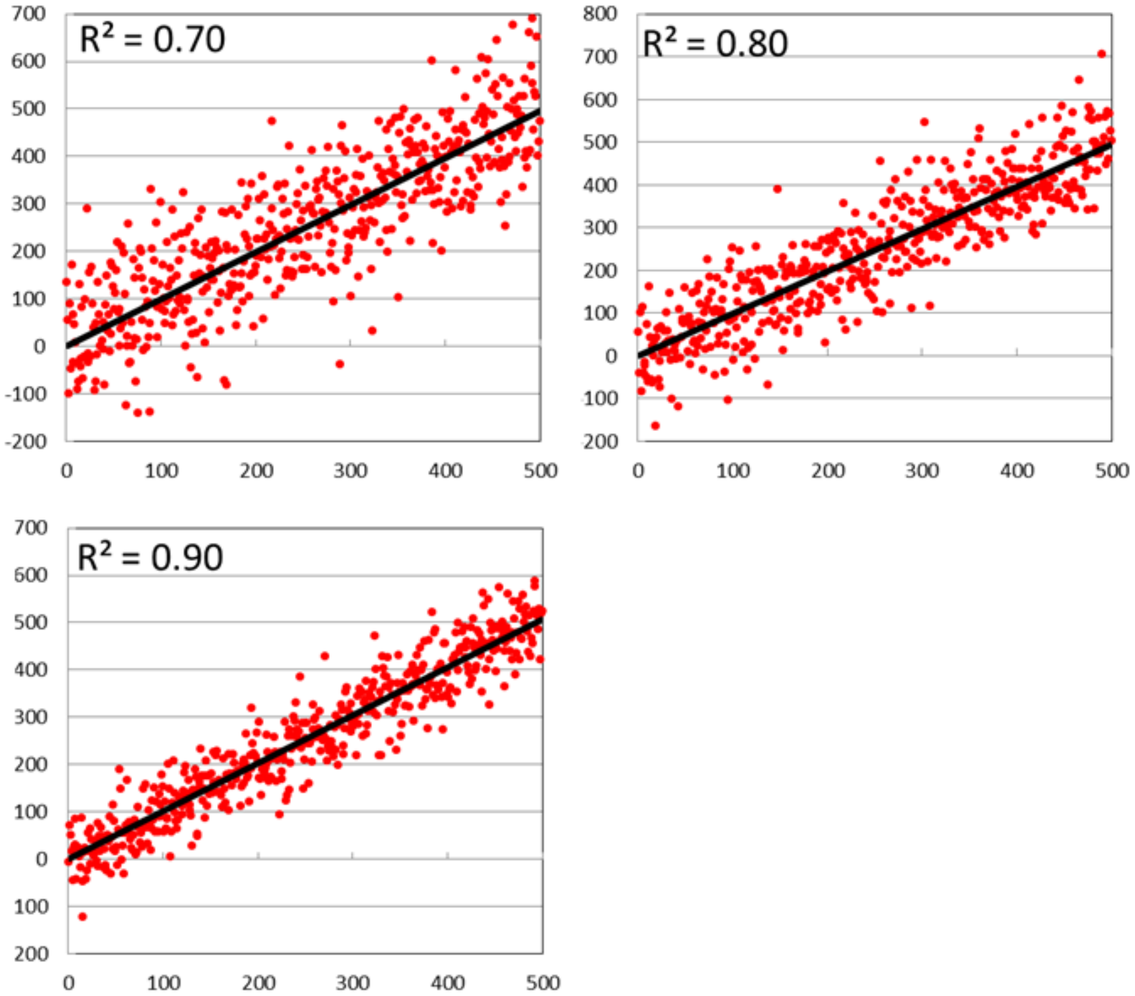
#4. Build and Train the model

We quantify how well the model can explain the data with Coefficient of determination (R^2 ,R squared).



$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad SS_{reg} = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

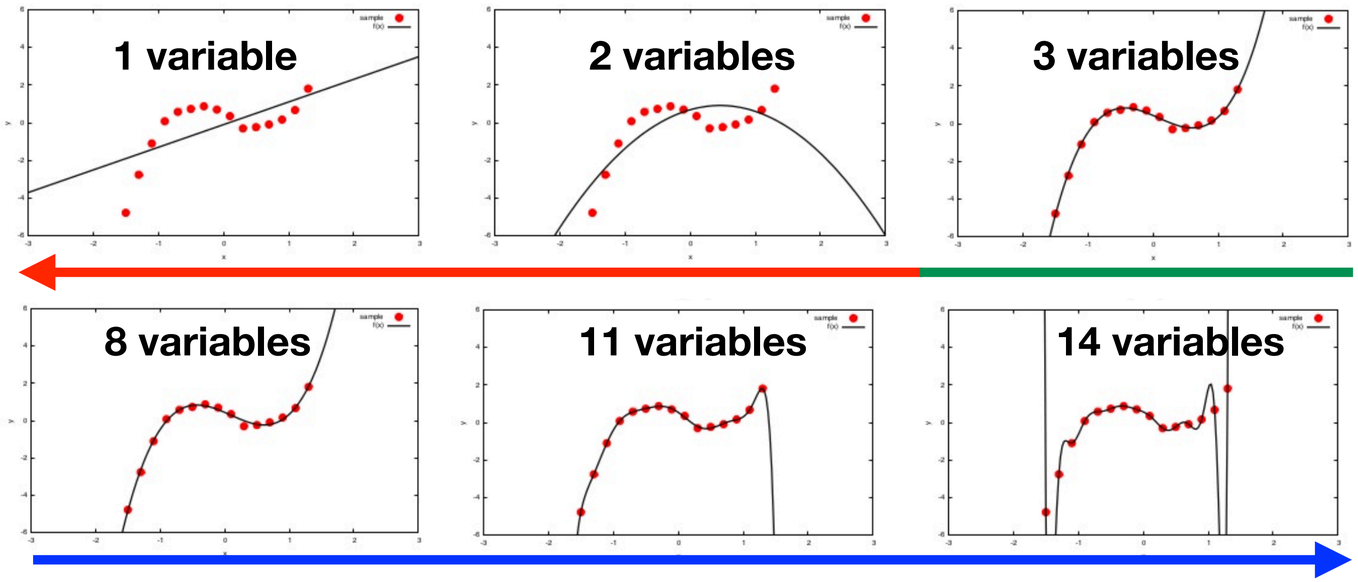
$$R^2 \equiv 1 - \frac{SS_{reg}}{SS_{tot}}$$



https://en.wikipedia.org/wiki/Coefficient_of_determination

#4. Build and Train the model

We need to optimize number of variable. With fewer variables, it is difficult to get high accuracy for both training set and test set. With too many variables, we expect to get higher accuracy for training set, but get lower accuracy for test set.



Simple model
High bias
Low variance

Complex model
Low bias
High variance

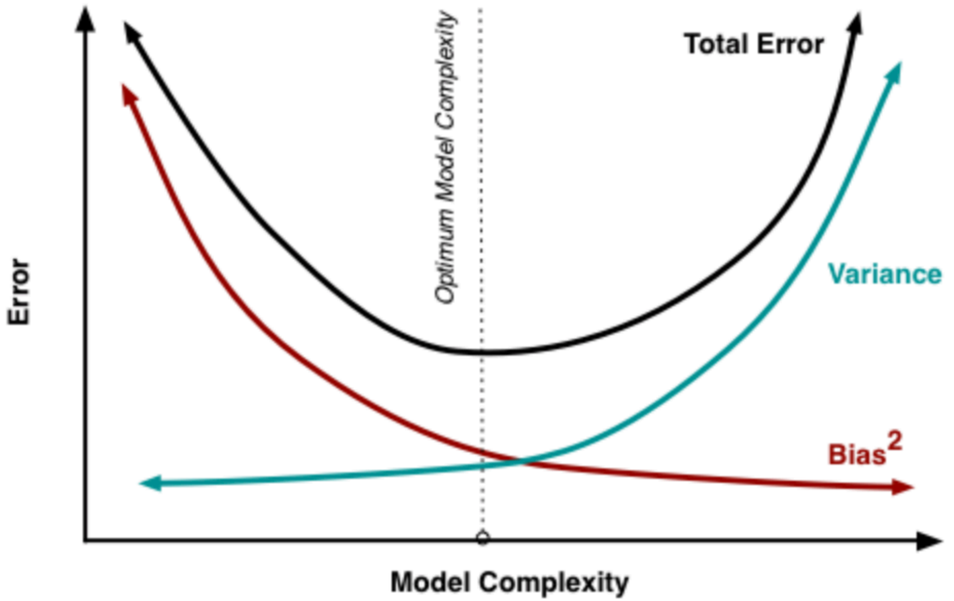


Fig. 6 Bias and variance contributing to total error.

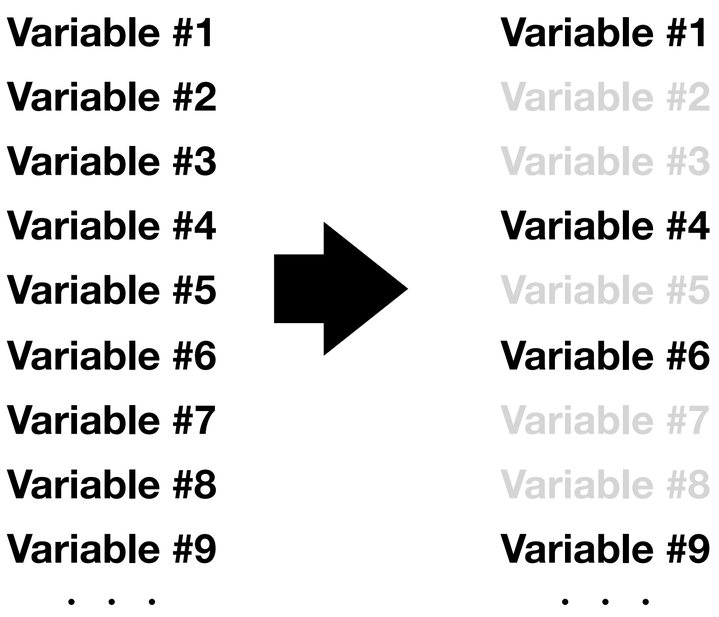
https://www.slideshare.net/ShinyaShimizu/ss-11623505

http://scott.fortmann-roe.com/docs/BiasVariance.html

#4. Build and Train the model

To get optimal number of variable, we select variable with “Stepwise Backward elimination” or “Stepwise Forward selection”.

Variable selection



Stepwise Backward elimination

A variable selection procedure in which all variables are entered into the equation and then sequentially removed.The procedure stops when there are no variables in the equation that satisfy the removal criteria.

Stepwise Forward selection

A stepwise variable selection procedure in which variables are sequentially entered into the model. The procedure stops when there are no variables that meet the entry criterion.

Selection criterion

Akaike Information Criterion (AIC) $AIC \equiv -2\ln(L) + 2k$

Bayes Information Criterion (BIC) $BIC \equiv -2\ln(L) + k\ln(n)$

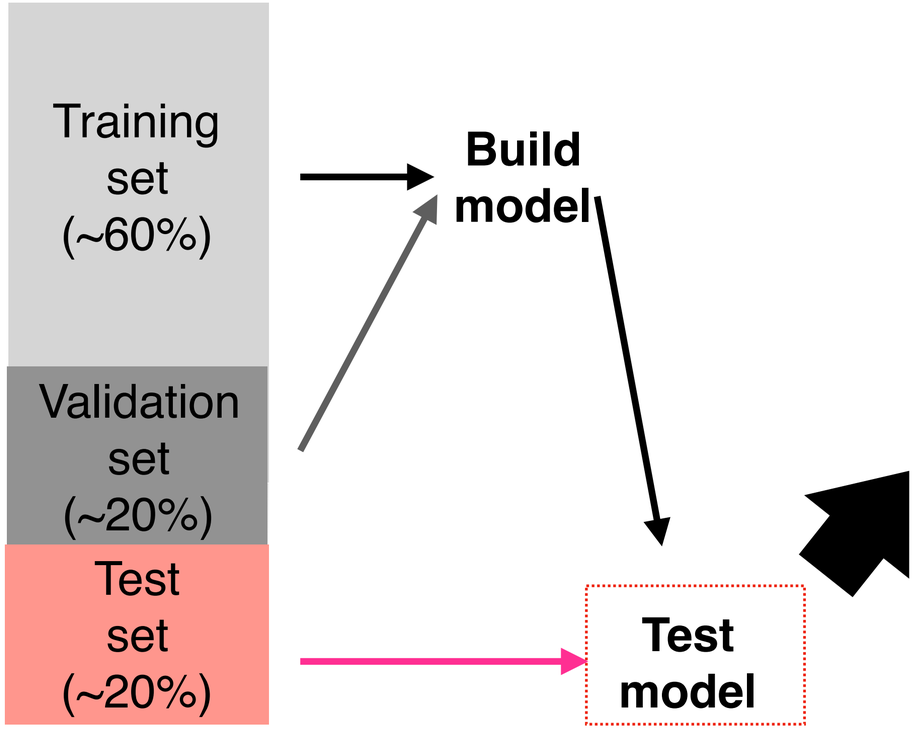
Adjusted R^2, R'^2 $R'^2 \equiv 1 - \frac{n-k-1}{n-1}R^2$

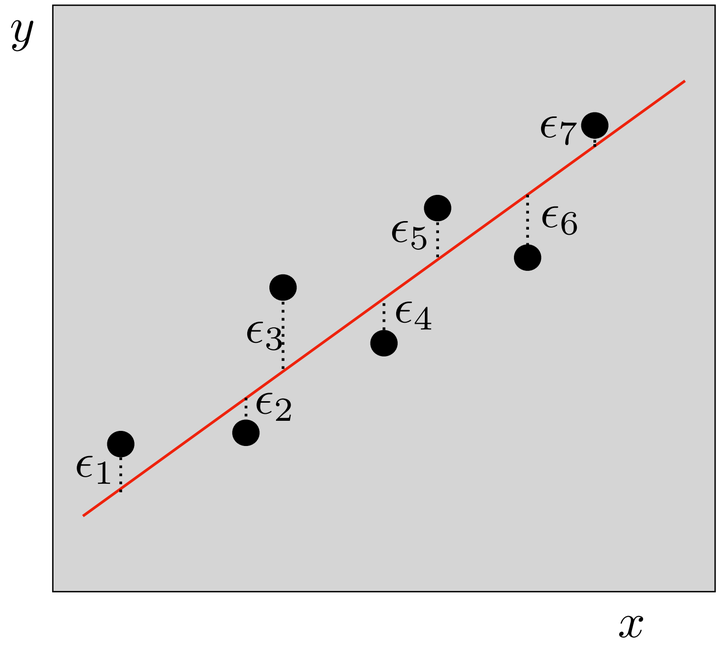
*L: likehood, k:number of variable, n:number of data.

https://www.ibm.com/support/knowledgecenter/en/SSLVMB_23.0.0/spss/base/linear_regression_methods.html

#5. Validate the model

We can validate the trained model with RMSE for test set.





Model

$$\hat{y} = \beta_0 + \beta_1 x$$

Test set

$$x = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$$
$$y = \{y_1, y_2, y_3, y_4, y_5, y_6, y_7\}$$

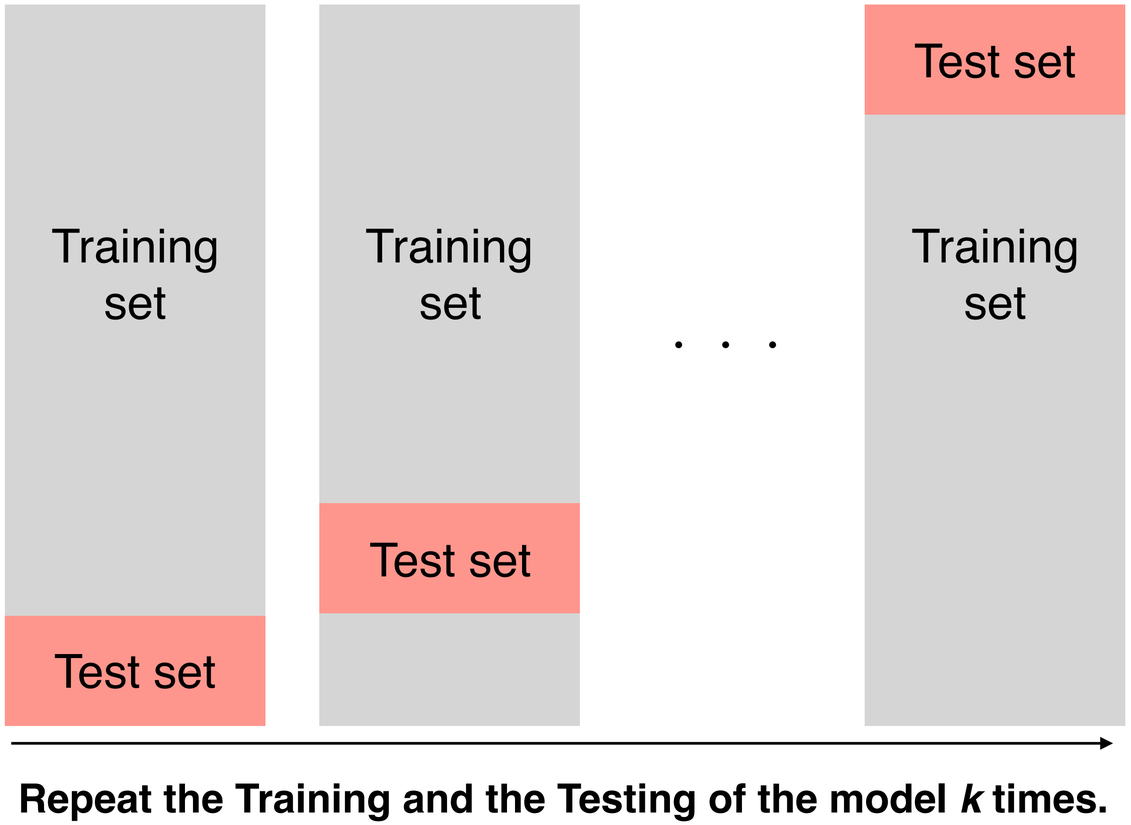
RMSE (Root Mean Squared Error)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

#5. Validate the model

We can validate the trained model with RMSE for test set.

k-fold cross-validation, Leave-one-out cross-validation



Repeat the Training and the Testing of the model *k* times.

Number of original data set: N

Number of test data set: n

k-fold cross-validation

$$n = \frac{N}{k}$$

Leave-one-out cross-validation

$$n = 1$$