# BC2406 Analytics I
# Visual and Predictive Techniques

Unit 2

## Fundamental Concepts and Principles of Analytics

Based on Chew C. H. (2019) textbook: Analytics, Data Science and AI. Vol 1., Chap 2.

NANYANG TECHNOLOGICAL UNIVERSITY
Nanyang Business School

# Seminar Objectives

- Learn how industry conduct Analytics projects.
- Learn fundamental, critical concepts that distinguish Analytics as a field of study (esp vs Statistics).
- Learn some common misconceptions and avoid them in your analytics work.
- Learn basic use of R and Rscripting



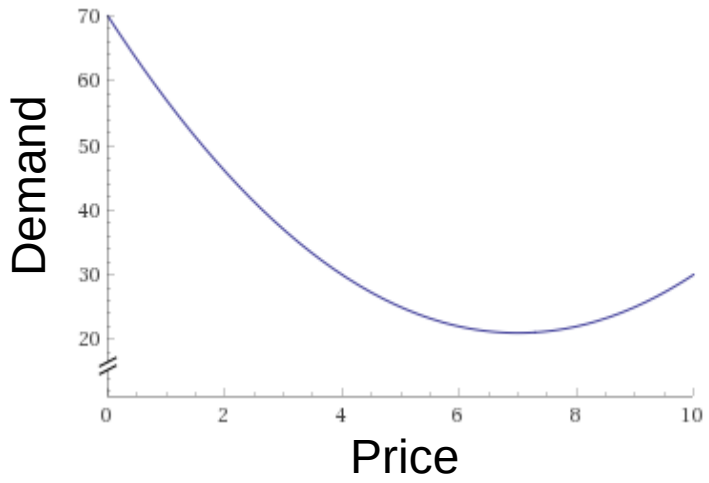
Based on Neumann (2019) textbook: Analytics, Data Science and AI. Vol 1.

# Role of Visualization and Models

**Analytics Model**

$$Demand = Price^2 - 14 * Price + 70$$

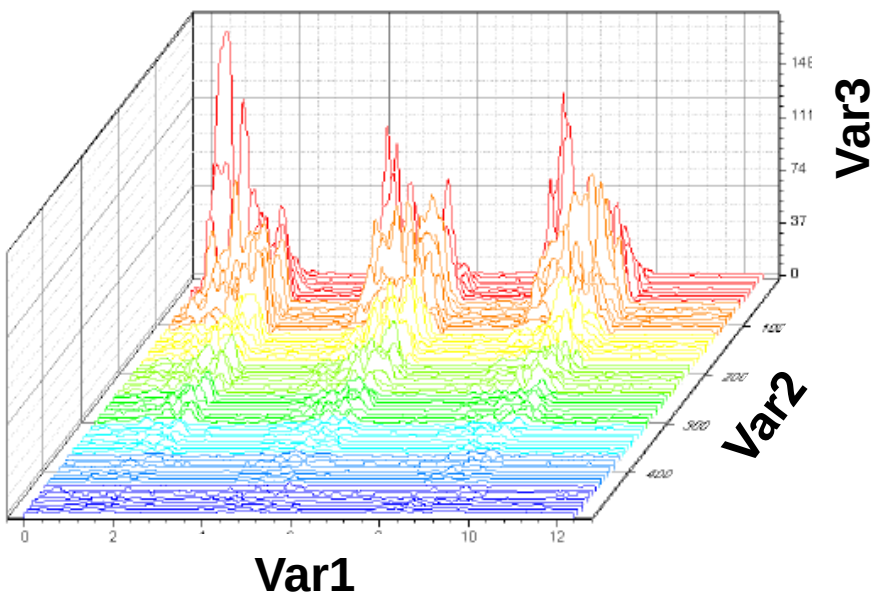
**Visualization**

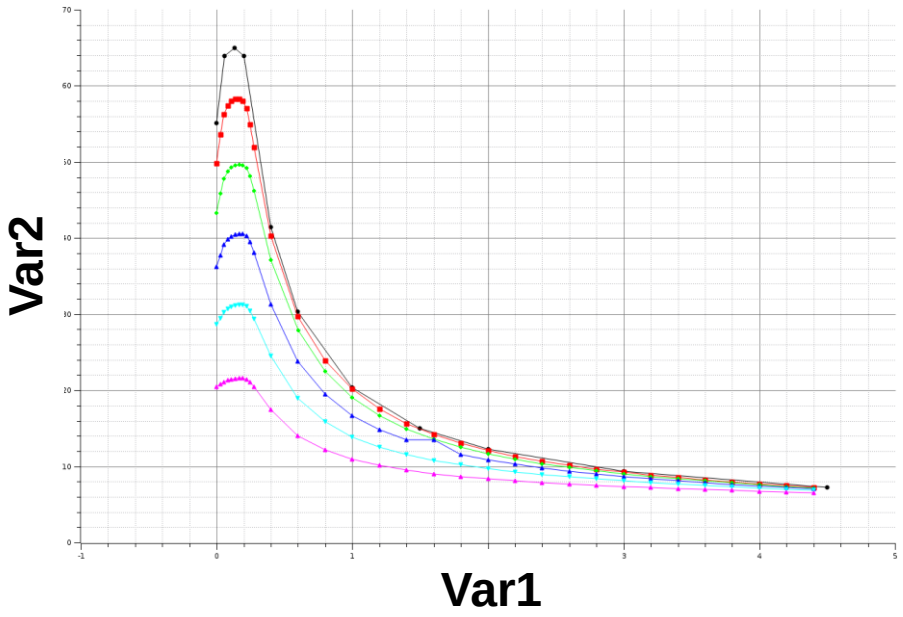



## Visualization

- Used to get easier understanding than analytics models
- Used to communicate/explain the results in an easier way
- Requires minimal training

# Role of Visualization and Models




**Product Sales at Qoo10**

**Affect**

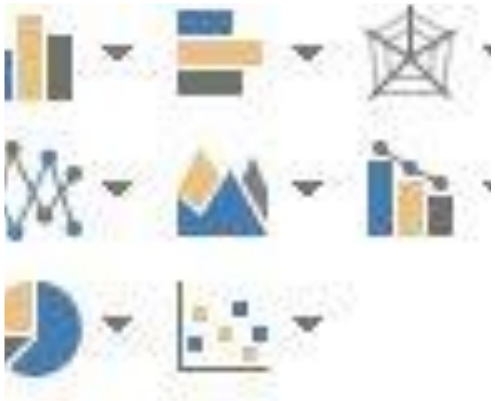
Price, Shipping Fee, Consumer Reviews, Discount Rate, Return Policy, Competitors, Etc.



Based on Neumann (2019) textbook: Analytics, Data Science and AI. Vol 1.

# Role of Visualization and Models

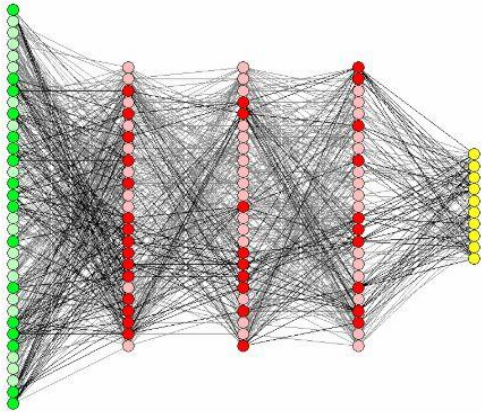
## Data Exploration



Statistics

Graphs

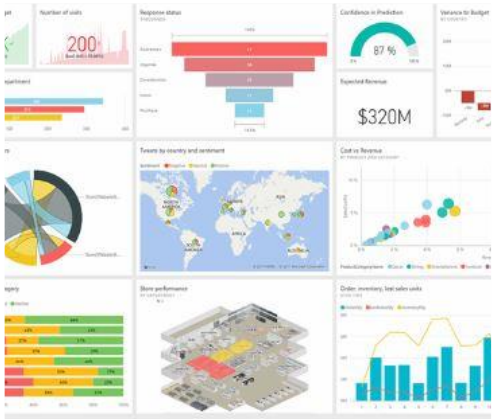
## Model Development



Model Structure

Model Evaluation

## Results Communication




Charts

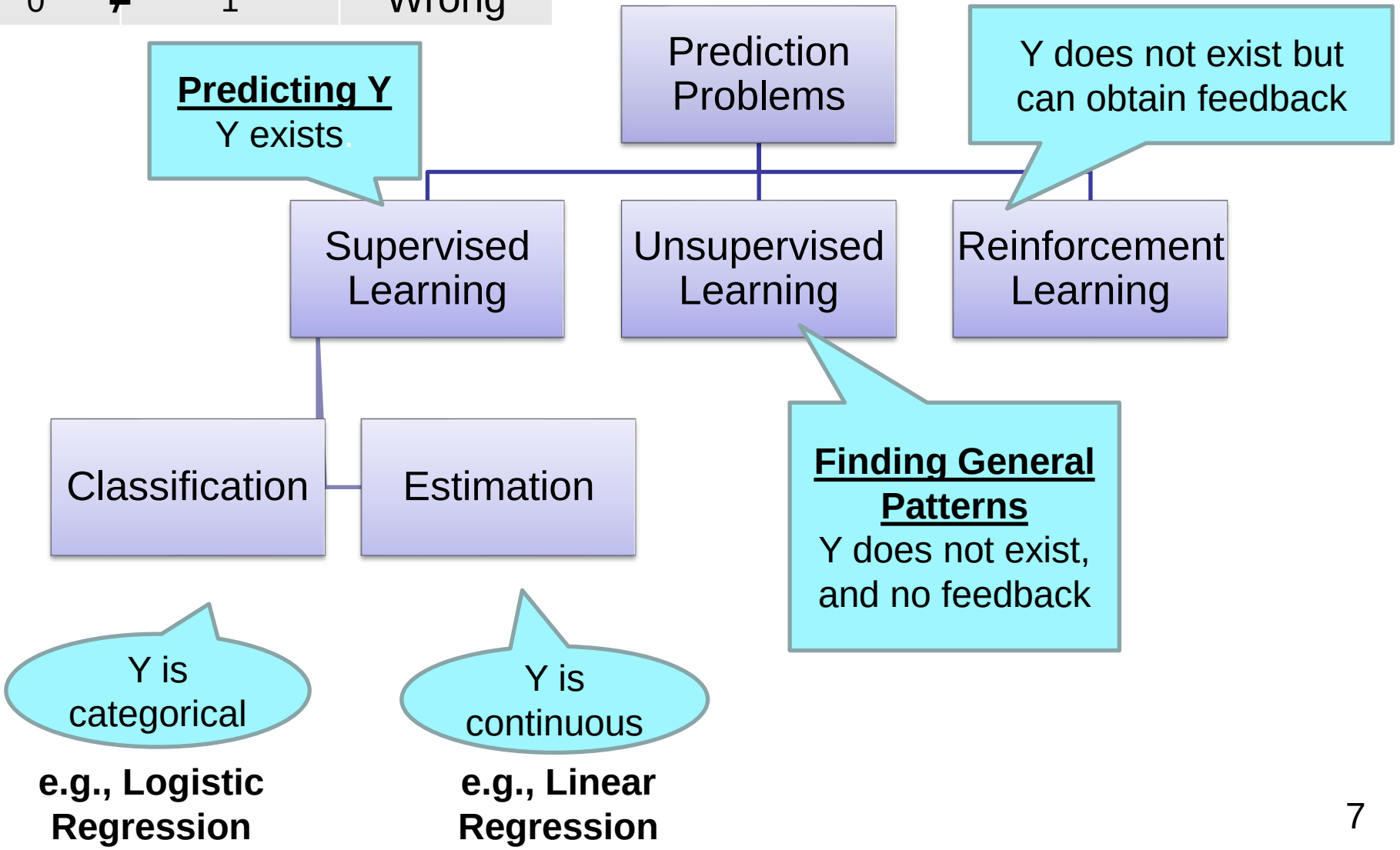
Dashboard



Based on Neumann (2019) textbook: Analytics, Data Science and AI. Vol 1.

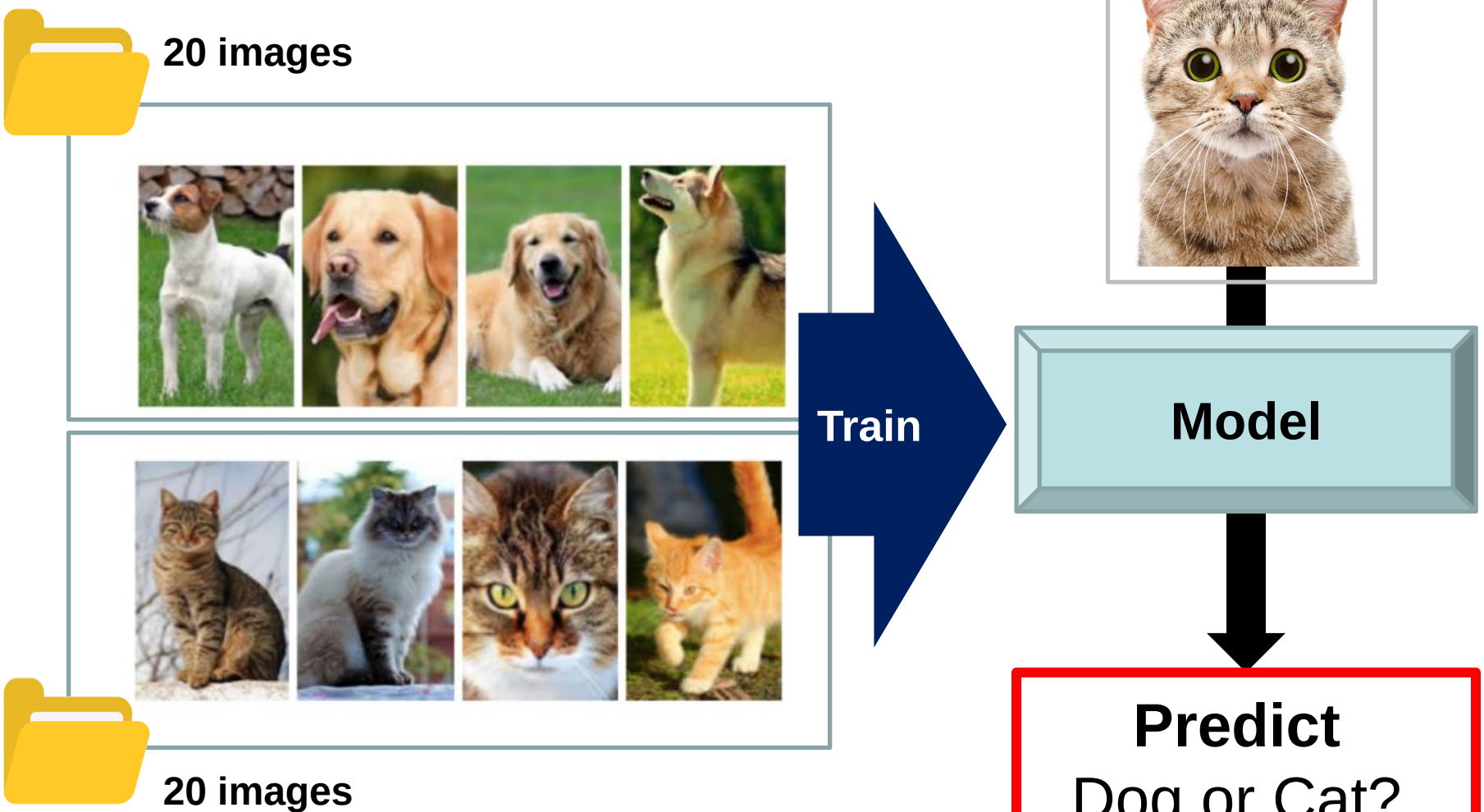
# Classification of Problems in Analytics

| Actual Y | | Predicted Y | Feedback |
|---|---|---|---|
| 1 | = | 1 | Right |
| 0 | ≠ | 1 | Wrong |



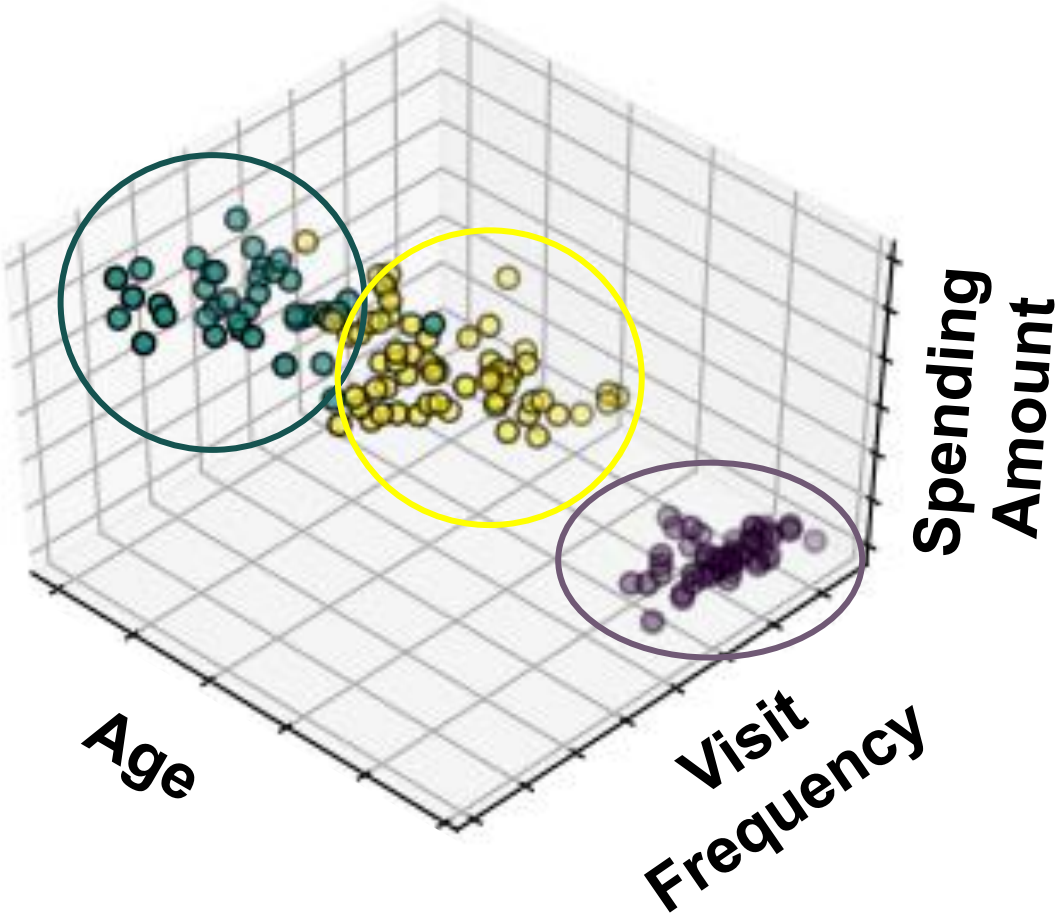
# Classification of Problems in Analytics

## Supervised Learning



# Classification of Problems in Analytics

## 3 Customer Segments




**Unsupervised Learning**

# Classification of Problems in Analytics

## Reinforcement Learning


### Supervised Learning

- Too many situations to be modeled
- Too many datasets are required to train the model

### Reinforcement Learning

- Set basic driving rules and let the model drive a car
- Give penalties (rewards) for bad (good) driving results

# Models* on Explanability Scale


Simple — Complex

Highest Explanability Power (White Box)

Lowest Explanability Power (Black Box)

*: Selected list of models (non-exhaustive) on the Explanability Scale.

Is my model correct? – The Correct Model Fallacy

# PRINCIPLE 1: MANY CORRECT MODELS



Based on Neumann (2019) textbook: Analytics, Data Science and AI. Vol 1.

# Example: Baby Data

| Age | Weight (Y) | Length of Left Arm ($X_1$) | Length of Right Arm ($X_2$) | Length of Left Leg ($X_3$) | Length of Right Leg ($X_4$) |
|---|---|---|---|---|---|
| 2 mth | 8 | 20.1 | 20.2 | 40.1 | 39.9 |
| 3 mth | 10.2 | 30.1 | 30.0 | 52.1 | 52.2 |
| ... | ... | ... | ... | ... | ... |

- Task: Build a Linear Regression Model to predict Weight (Y).
- You can use any input variables $X_i$

Disclaimer: Data is not real and meant for illustration only.

# Example: Baby Data

- $\widehat{Y} = 2 + 0.1X_1 + 0.2X_2$
- $\widehat{Y} = 28 + 2X_1 - 3X_2$
- $\widehat{Y} = -5 + 0.3X_1 + 0.2X_3$
- $\widehat{Y} = 48 + 3X_3 - 4X_4$

- Any of the above model is fine.
- More than one correct model.
- Unlearn previous mathematical education.
- Is my model good enough? How to judge?



Based on Neumann (2019) textbook: Analytics, Data Science and AI. Vol 1.

# Model Predictive Accuracy for Continuous Y

**RMSE**
(Root Mean Square Error) $= \sqrt{\dfrac{\sum_{i=1}^{i=n} (\hat{y}_i - y_i)^2}{n}}$



Based on Neumann (2019) textbook: Analytics, Data Science and AI. Vol 1.

# Model Predictive Accuracy for Categorical Y

## Confusion Matrix

| | | Actual | |
|---|---|---|---|
| | | Not Fraud | Fraud |
| Model Prediction | Not Fraud | 10 | 17 |
| | Fraud | 3 | 20 |



Based on Neumann (2019) textbook: Analytics, Data Science and AI. Vol 1.

# Is my model good enough?

- Based on Predictive Accuracy. Better than target?
- Setting target for predictive accuracy:
  - historical predictive performance in your company
  - best benchmark achieved in recent research papers
  - finding out the desired business impact, and reverse engineer the required target
  - comparing against a standard model (Linear Regression or Logistic Regression)
  - what the boss wants [may or may not be realistic]
- Based on use requirements.
  - A&E Doctors and nurses want a simple model that they could use without statistics background.



Based on Neumann (2019) textbook: Analytics, Data Science and AI. Vol 1.

Zero Prediction Error! The Perfect Model Fallacy.

# PRINCIPLE 2: TRAIN TEST SPLIT

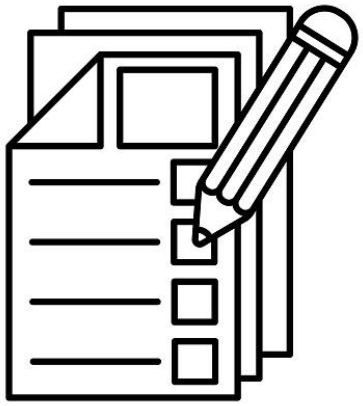


Based on Neumann (2019) textbook: Analytics, Data Science and AI. Vol 1.

# Zero Prediction Error

# = Excellent ?

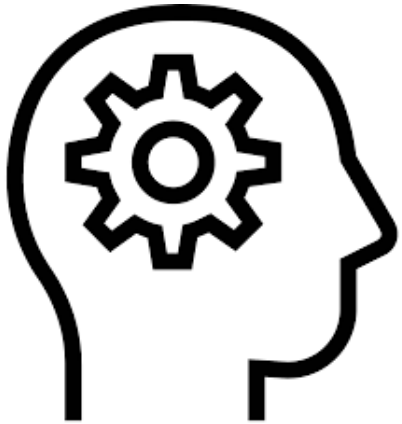
**Previous Exam Questions (Data)**



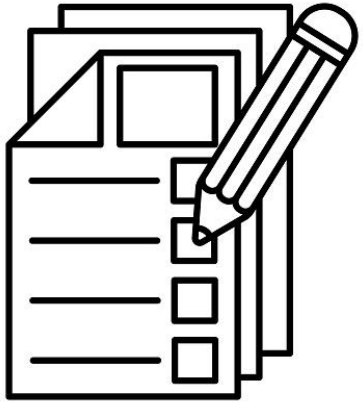
**Train**

**Test**

**Knowledge (Model)**

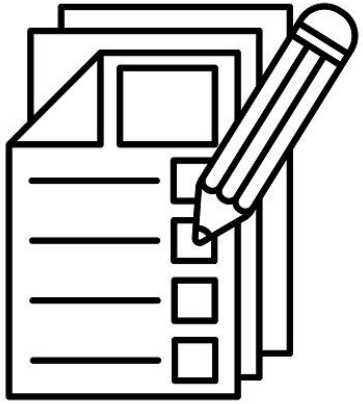


**Train Set of Previous Exam Questions (Data)**



mutually exclusive

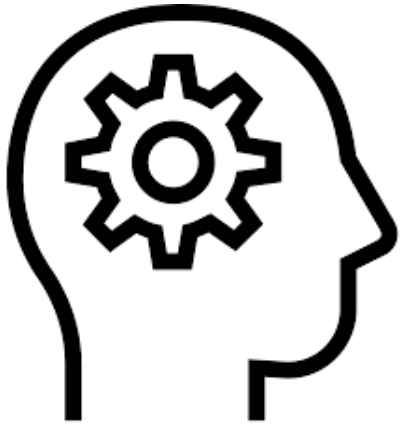
**Test Set of Previous Exam Questions (Data)**



Train

Test

**Knowledge (Model)**



Nanyan

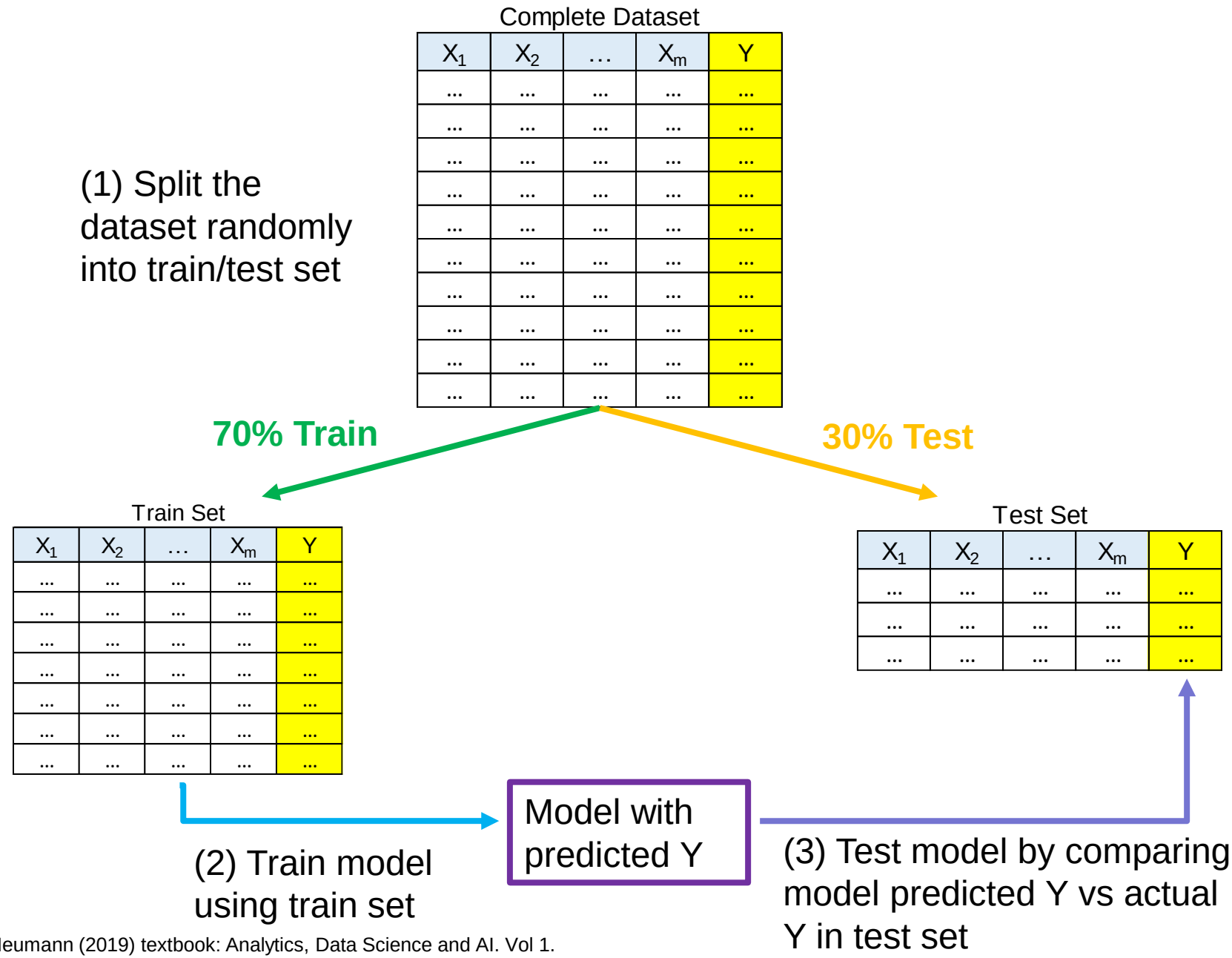
# Zero Prediction Error?

- **Too good to be true**
- **Typically obtained based on trainset**
  - Implications of Polynomial Interpolation Theorem
  - Neural Network infamous for
  - Biased and Over-optimistic
- Do not be misled
  - Predict Historical Data, or
  - Predict Future Data
- Model Predictive Accuracy
  - How to decide if model is good enough to be used, if not based on historical data?
  - What would be a fair, unbiased estimate of error?



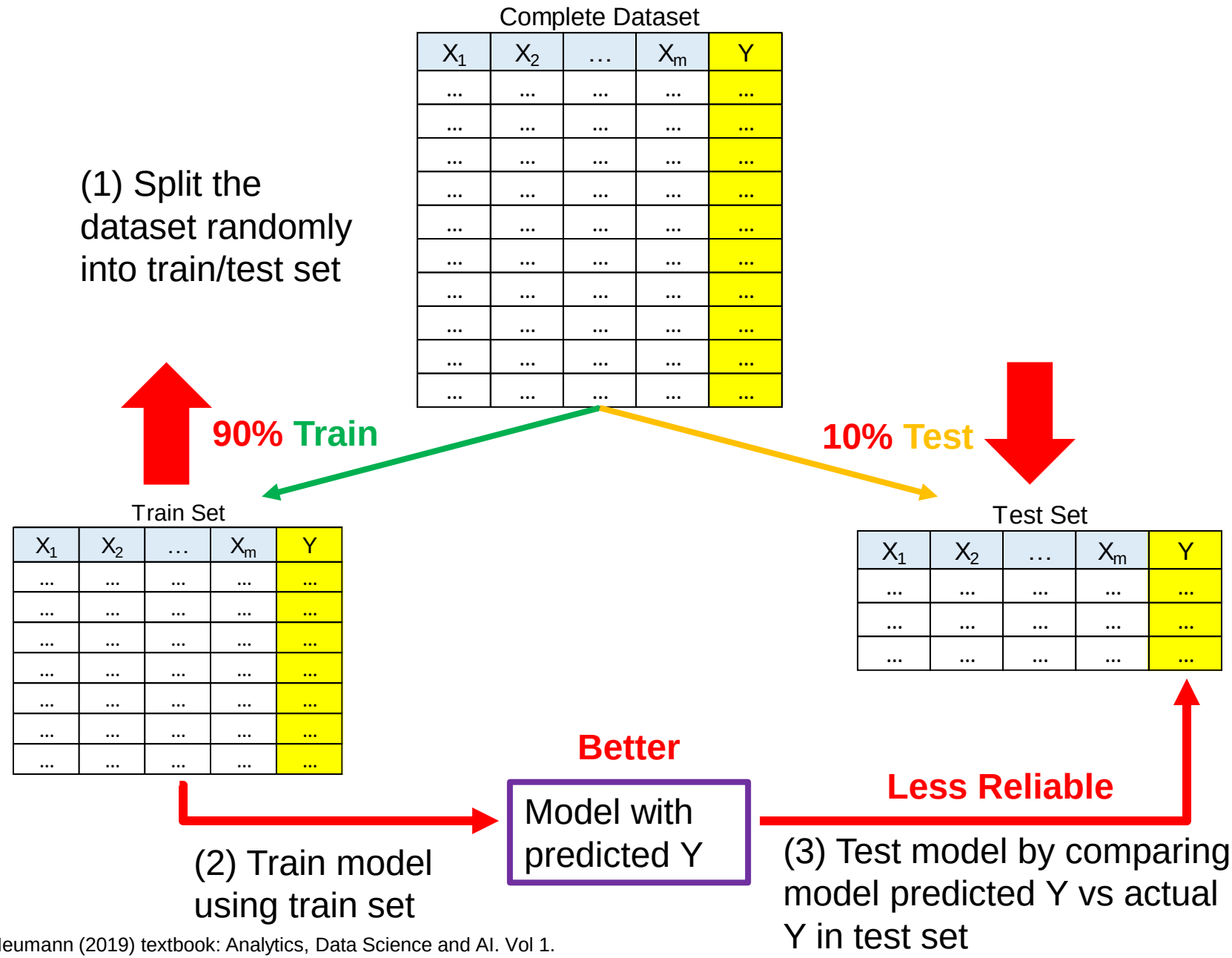
Based on Neumann (2019) textbook: Analytics, Data Science and AI. Vol 1.

# Train – Test Split Procedure (Basic)



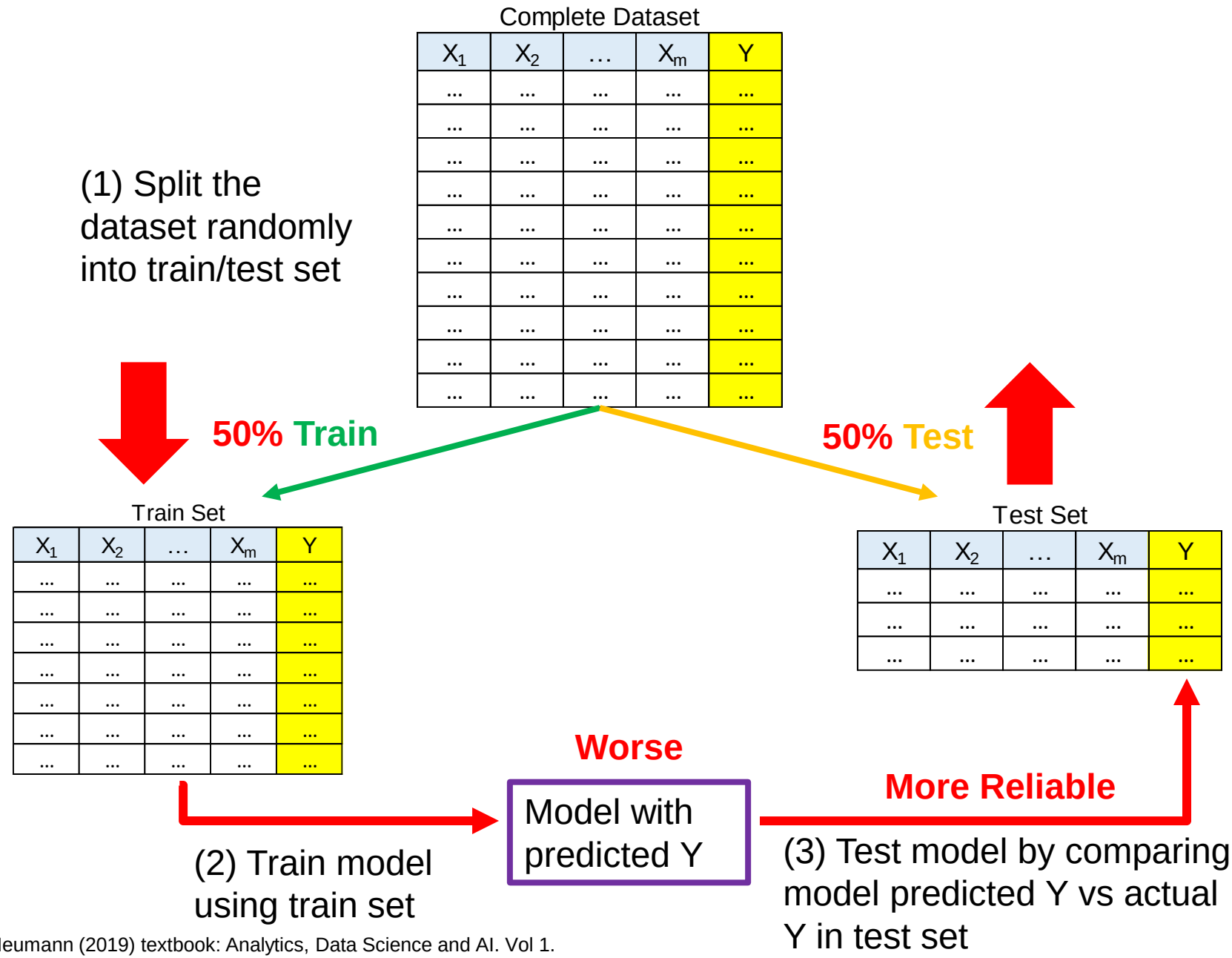

Based on Neumann (2019) textbook: Analytics, Data Science and AI. Vol 1.

# Train – Test Split Procedure (Basic)




Based on Neumann (2019) textbook: Analytics, Data Science and AI. Vol 1.

# Train – Test Split Procedure (Basic)




Based on Neumann (2019) textbook: Analytics, Data Science and AI. Vol 1.

# Train – Test Split Procedure (Basic)

Complete Dataset

| $X_1$ | $X_2$ | … | $X_m$ | Y |
|---|---|---|---|---|
| … | … | … | … | 0 |
| … | … | … | … | 0 |
| … | … | … | … | 1 |
| … | … | … | … | 0 |
| … | … | … | … | 0 |
| … | … | … | … | 1 |
| … | … | … | … | 0 |
| … | … | … | … | 0 |
| … | … | … | … | 0 |
| … | … | … | … | 0 |

(1) Split the dataset randomly into train/test set

**If Y is Categorical & Rare Event**
(e.g., Cancer)

**70% Train**

**30% Test**

Train Set

| $X_1$ | $X_2$ | … | $X_m$ | Y |
|---|---|---|---|---|
| … | … | … | … | 0 |
| … | … | … | … | 1 |
| … | … | … | … | 0 |
| … | … | … | … | 1 |
| … | … | … | … | 0 |
| … | … | … | … | 0 |
| … | … | … | … | 0 |

**All Y=1 Cases**

Test Set

| $X_1$ | $X_2$ | … | $X_m$ | Y |
|---|---|---|---|---|
| … | … | … | … | 0 |
| … | … | … | … | 0 |
| … | … | … | … | 0 |

**No Y=1 Case**

Model with predicted Y

(2) Train model using train set

(3) Test model by comparing model predicted Y vs actual Y in test set

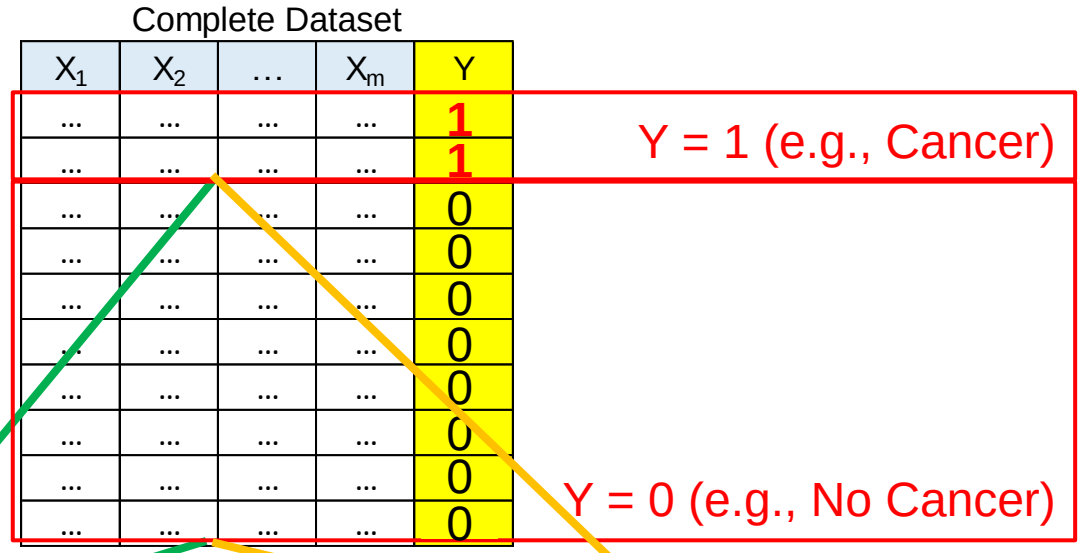
NANYANG TECHNOLOGICAL UNIVERSITY
Nanyang Business School

Based on Neumann (2019) textbook: Analytics, Data Science and AI. Vol 1.

# Train – Test Split Procedure (Enhanced)

(1) Stratify on categorical Y

(2) Split the dataset randomly, within each strata, into train/test set




(3) Train model using train set

Model with predicted Y

(4) Test model by comparing model predicted Y vs actual Y in test set

Based on Neumann (2019) textbook: Analytics, Data Science and AI. Vol 1.

# Train Test Split

- Split your historical dataset into two pieces
  - Stratify on y. [Why?]
  - Typically, but not always, 70% Training Set
  - Typically, but not always, 30% Test Set
- Develop your model on the Training Set
- Test your model on the Test Set
  - Obtain unbiased estimate of model prediction error.

How to select best model within the model type?

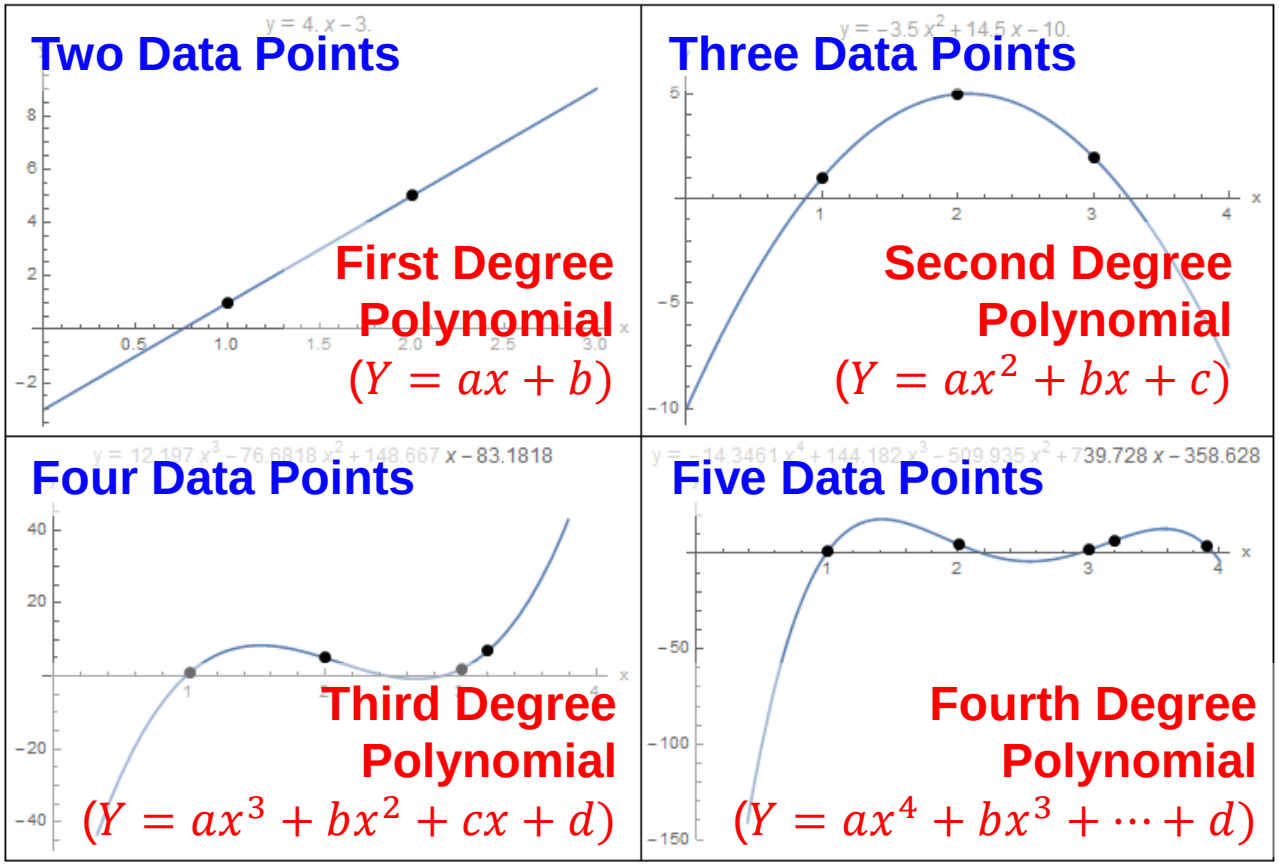
# PRINCIPLE 3: COMPLEXITY ADJUSTED MODEL PERFORMANCE



Based on Neumann (2019) textbook: Analytics, Data Science and AI. Vol 1.

# Implication of Model Complexity

- Idea from Polynomial Interpolation Theorem
- Hit zero error on trainset by using very complex model



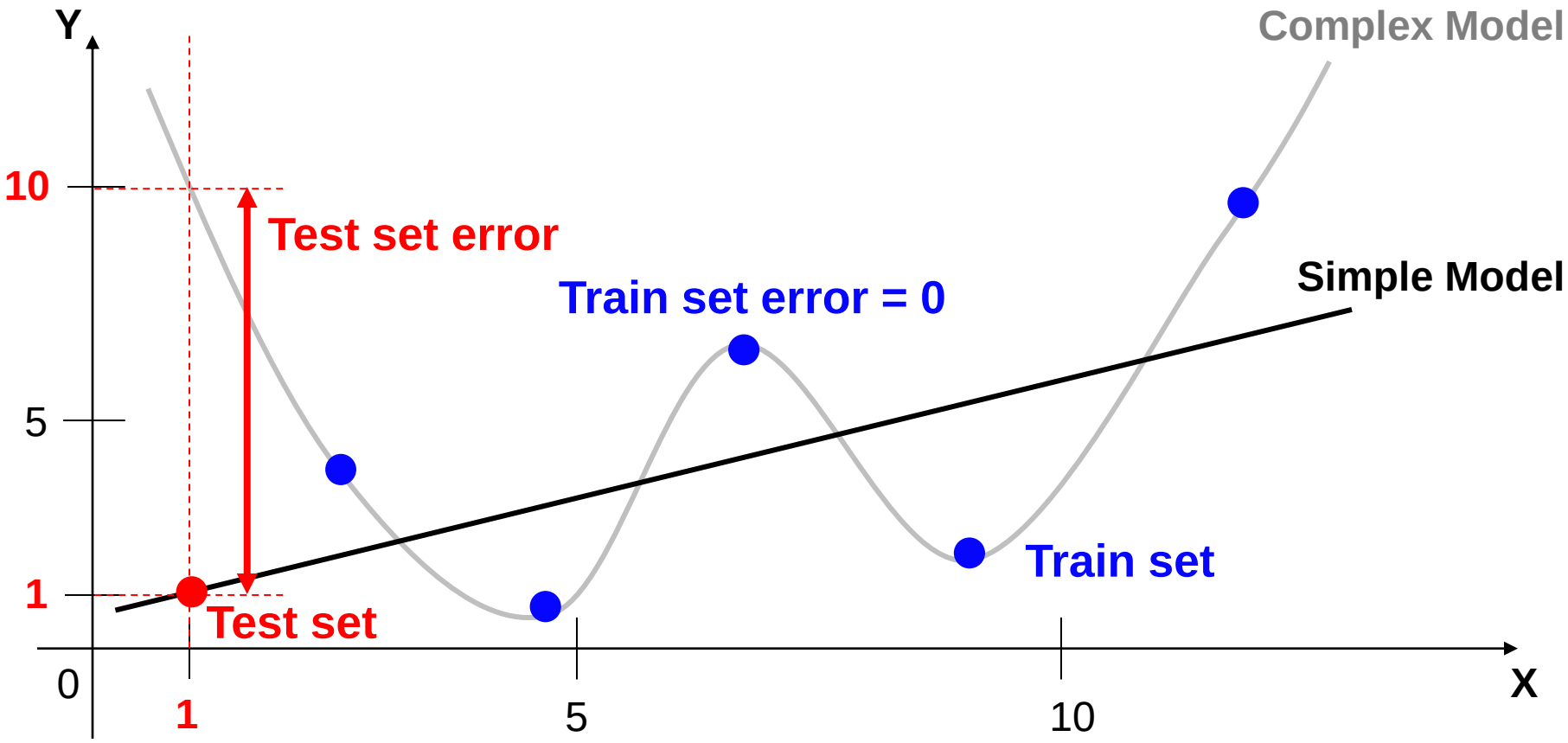

**N Data Points**

**N-1 th Degree Polynomial**

$(Y = a_{n-1}x^{n-1} + \cdots + a_0)$

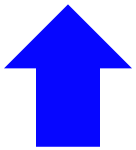
# Implication of Model Complexity

- Idea from Polynomial Interpolation Theorem
- Hit zero error on trainset by using very complex model



# What is Model Complexity?

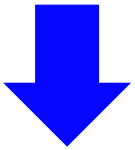
| Model Type | Complexity Measure |
|---|---|
| Linear Regression | Number of Xs |
| Logistic Regression | Number of Xs |
| Classification & Regression Tree | Number of Terminal Nodes or Number of Splits |
| Neural Network | Number of Hidden Layers and Hidden Nodes |
| MARS | Number of Xs |



**Model Complexity**
**= # Input Variables**

**:**

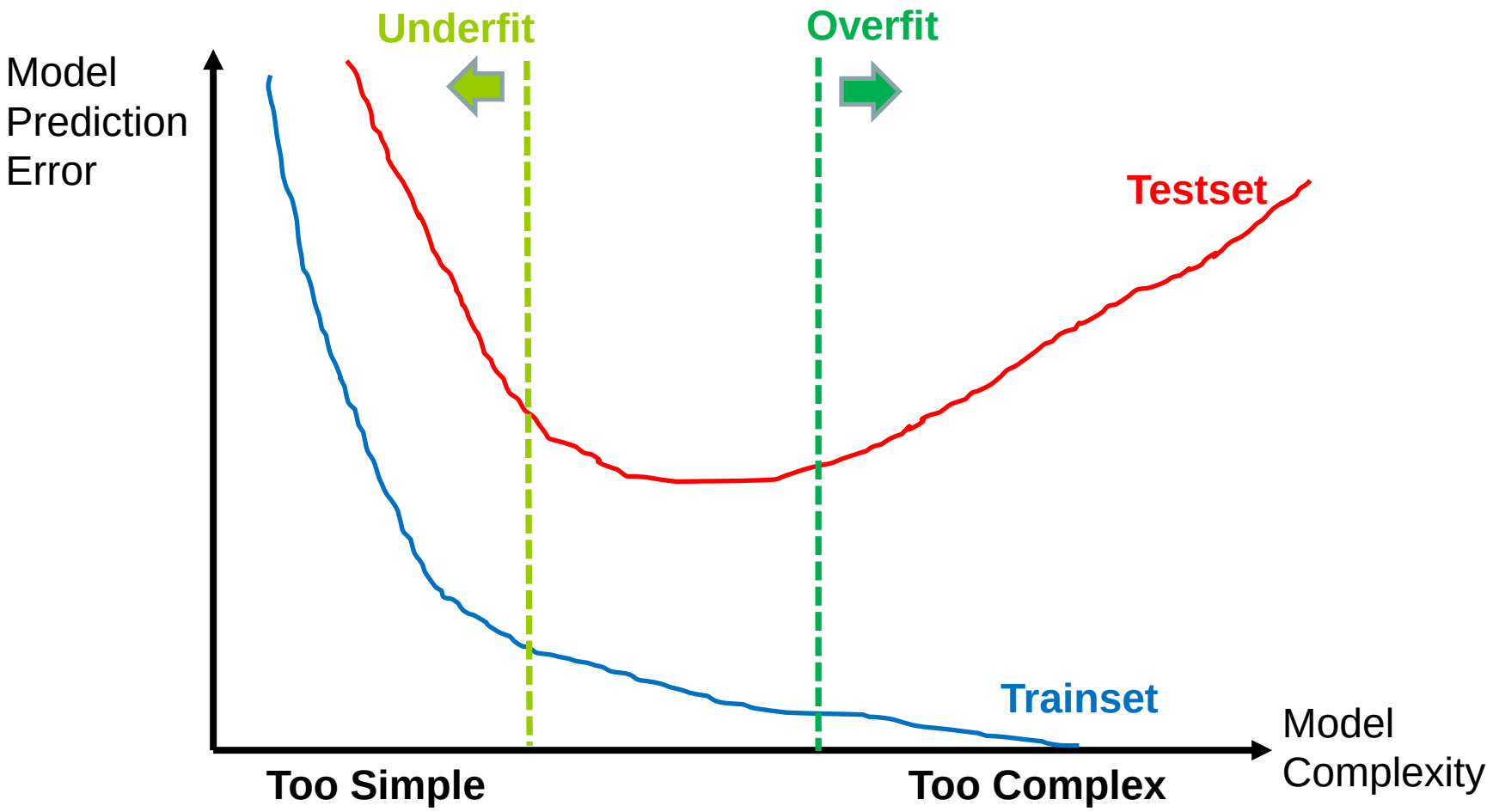
**Train Set Error**



**:**

**Test Set Error**

# Overfitting Risk




Beyond a certain model complexity, testset error increases while trainset error continue decreasing.

Based on Neumann (2019) textbook: Analytics, Data Science and AI. Vol 1.

# Complexity-Adjusted Model Performance

- Find a way to balance both model prediction error and model complexity.
- Classification and Regression Tree (CART) does this automatically with an “alpha” parameter
  - More details in Unit 8 Decision Trees.

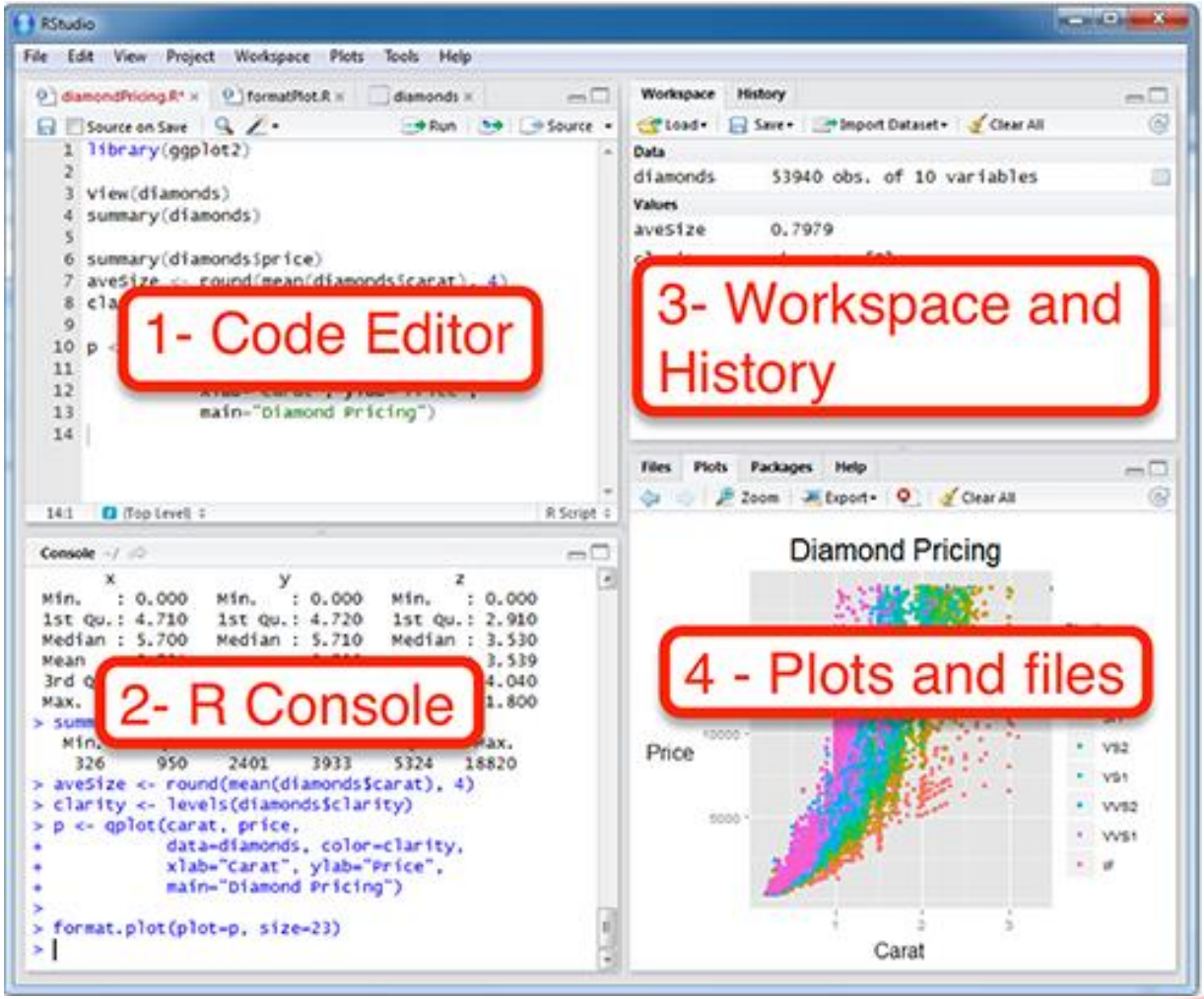
Learning R

# BASICS

# Install R and RStudio

- Install R first
- Then install RStudio
- Before week 2 class.
- Bring your laptop to class
  - To conserve battery, switch off wi-fi when there is no need for internet access.

# 4 Panels work area in RStudio




NANYANG TECHNOLOGICAL UNIVERSITY

Nanyang Business School

# Operators: R as a calculator

```
1 / 200 * 30
#> [1] 0.15
(59 + 73 + 2) / 3
#> [1] 44.7
sin(pi / 2)
#> [1] 1
```

| Operator | Description |
|---|---|
| + | addition |
| - | subtraction |
| * | multiplication |
| / | division |
| ^ or ** | exponentiation |
| x %% y | modulus (x mod y) 5%%2 is 1 |
| x %/% y | integer division 5%/%2 is 2 |

| Operator | Description |
|---|---|
| < | less than |
| <= | less than or equal to |
| > | greater than |
| >= | greater than or equal to |
| == | exactly equal to |
| != | not equal to |
| !x | Not x |
| x \| y | x OR y |
| x & y | x AND y |
| isTRUE(x) | test if X is TRUE |

# The assignment Operator: <-

You can create new objects with `<-` :

```
x <- 3 * 4
```

All R statements where you create objects, **assignment** statements, have the same form:

```
object_name <- value
```

You can inspect an object by typing its name:

```
x
#> [1] 12
```



RStudio's keyboard shortcut: Alt and - (the minus sign)

# Object Naming Convention in R

Object names must start with a letter, and can only contain letters, numbers, `_` and `.`. You want your object names to be descriptive, so you'll need a convention for multiple words. I recommend **snake_case** where you separate lowercase words with `_`.

```
i_use_snake_case
otherPeopleUseCamelCase
some.people.use.periods
```

# R is spelling and case sensitive

Make yet another assignment:

```
r_rocks <- 2 ^ 3
```

Let's try to inspect it:

```
r_rock
#> Error: object 'r_rock' not found
R_rocks
#> Error: object 'R_rocks' not found
```

# R functions

R has a large collection of built-in functions that are called like this:

```
function_name(arg1 = val1, arg2 = val2, ...)
```

Type se and hit TAB. A popup shows you possible completions. Specify seq() by typing more (a “q”) to disambiguate, or by using ↑/↓ arrows to select. Notice the floating tooltip that pops up, reminding you of the function’s arguments and purpose. If you want more help, press F1 to get all the details in the help tab in the lower right pane.

Press TAB once more when you’ve selected the function you want. RStudio will add matching opening (() and closing ()) parentheses for you. Type the arguments 1, 10 and hit return.

```
seq(1, 10)
#>  [1]  1  2  3  4  5  6  7  8  9 10
```

# Text String

Type this code and notice you get similar assistance with the paired quotation marks:

```
x <- "hello world"
```

Quotation marks and parentheses must always come in a pair. RStudio does its best to help you, but it's still possible to mess up and end up with a mismatch. If this happens, R will show you the continuation character "+":

```
> x <- "hello
+
```

The `+` tells you that R is waiting for more input; it doesn't think you're done yet. Usually that means you've forgotten either a `"` or a `)`. Either add the missing pair, or press ESCAPE to abort the expression and try again.

# Create Vectors with c function

A vector is a set of values that are **all of the same type.**

```
a <- c(1,2,5.3,6,-2,4) # numeric vector
b <- c("one","two","three") # character vector
c <- c(TRUE,TRUE,TRUE,FALSE,TRUE,FALSE) #logical vector
```

Hint: Use the class function to check on the type of object.

# Resources for Learning R

- BC3407 R and Python for BA
- Textbooks
- Blogs
- Online Tutorials (Quick R):
  - Quick R: https://www.statmethods.net/

# Summary

- Visualization and Models have important roles.
- Fundamental Analytics concepts.
  - More than one correct model.
  - The importance of Trainset vs Testset.
  - Beware of low/zero prediction error on trainset.
- R Basics
  - Operators, Functions, Basic Plot.
  - Import Dataset into R.
  - Create new Data and Dataset within R.
  - Export processed Dataset out of R.
  - Saving and running Rscripts.
  - Errors are natural in programming. Just resolve them.