Disclaimer

- This presentation is purely for academic purpose and does not carry any commercial value.
- All images and photos used in this presentation are property of respective image holder(s) and due credit is provided to them. Images are used only for indicative purpose and does not carry any other meaning.
- All information and data in this slide are collected from open domain.

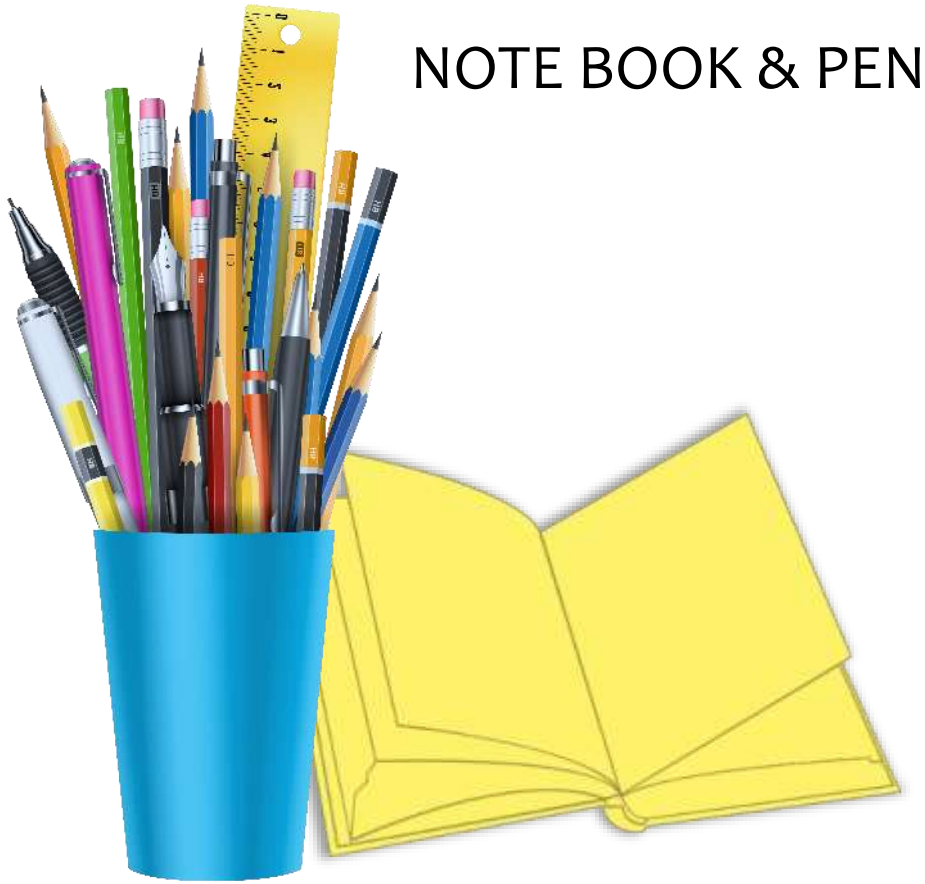
MANISH GODSE, Ph.D.(IIT Bombay)

Request & Instructions

PLEASE OPEN

CALCULATOR

NOTE BOOK & PEN



LAPTOP OR DESKTOP,
IF YOU HAVE.

PLEASE FOLLOW THIS

SILENCE

MUTE MIC

RAISE HAND

NO CHAT

SILENT MODE

MACHINE LEARNING STEPS

BOOKS & REFERENCES

- NO

Table of Contents

ANALYTICS LIFECYCLE

1

ANALYTICS LIFECYCLE

Problem Statement

Capture relevant sample data to solve business problems

DATA DISCOVERY & SAMPLING

DATA STUDY & PREPARATION

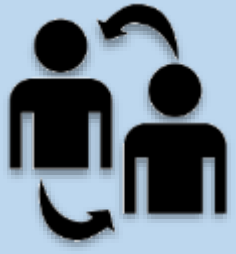
Manipulate data in required format and values

Select useful variables, algorithms to construct analytics model

MODEL DEVELOPMENT

OUTCOME ASSESSMENT

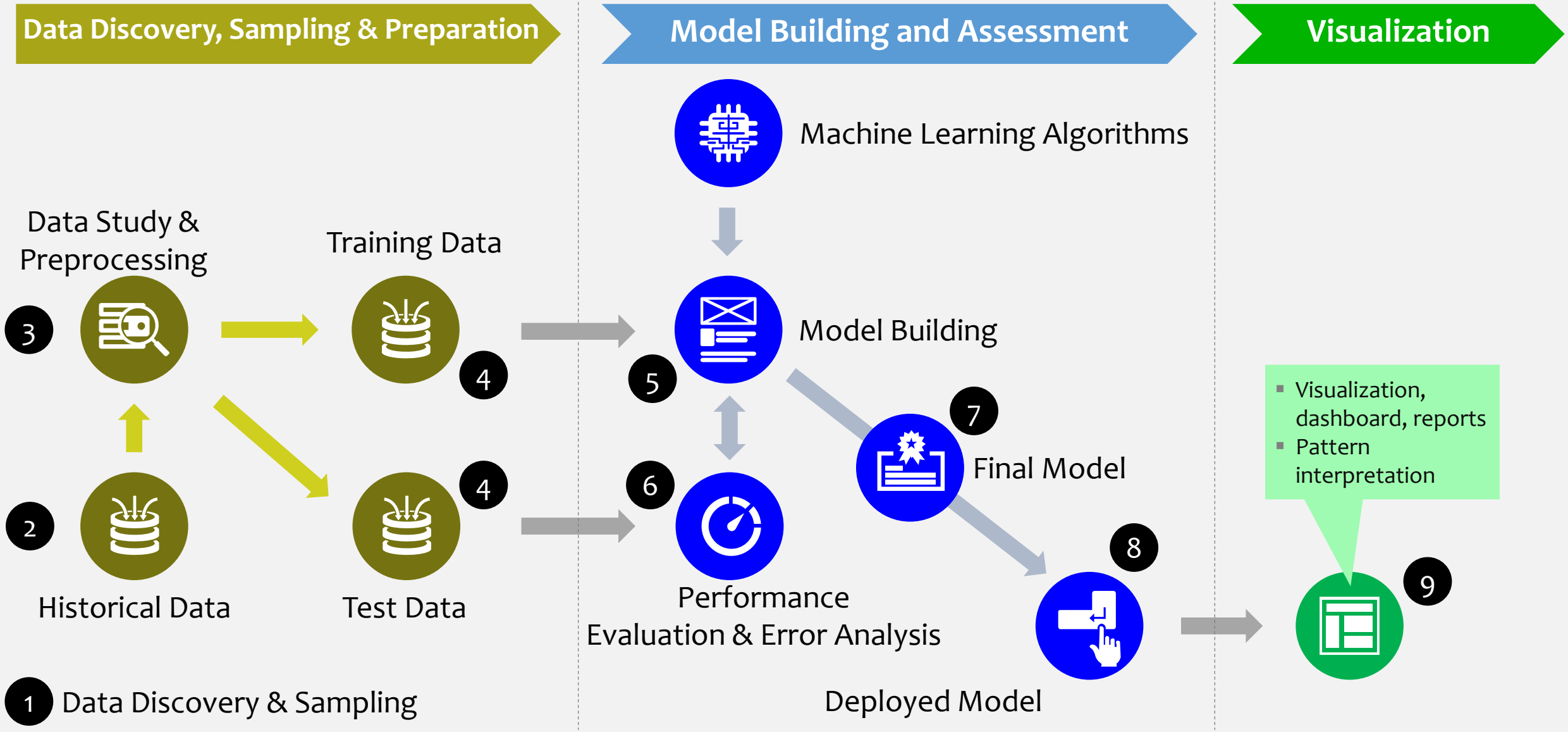
Evaluate reliability of model outcome on appropriate data sets



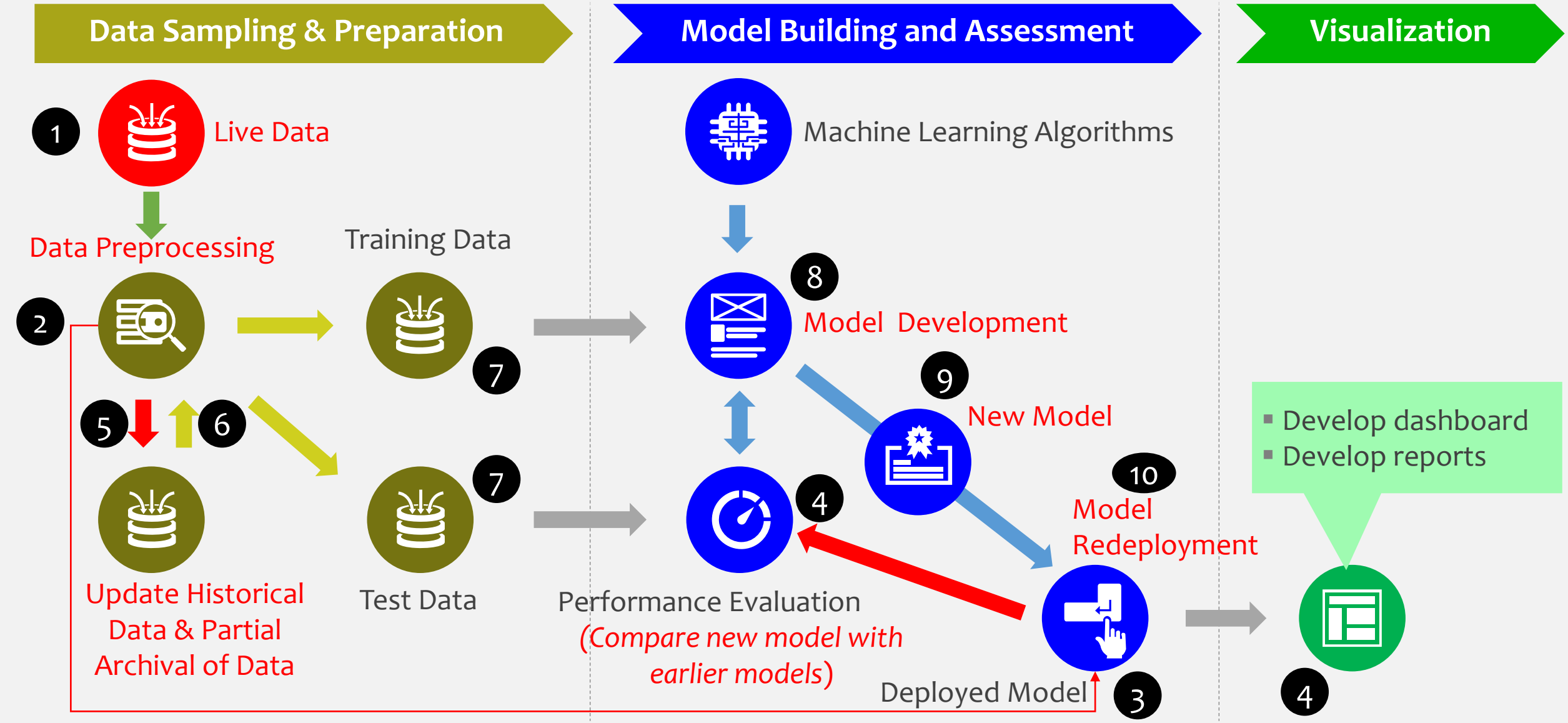
Communicate findings to stakeholders and managers

COMMUNICATION

GENERIC STEPS IN ANALYTICS



AUTOMATED DATA PIPELINE & ANALYTICS



PROBLEM STATEMENT

2

CROSS/UP SELL PROBLEM

You are a sales manager in XYZ company. You wanted to increase sales. One of the channels is cross sell and up sell. You wanted to create a list of potential customers.

How will you do it ?

ASK QUESTIONS?

- What is expected from Analytics? (**GOAL**)
- How to measure output? (**Metric**)
- Is **automation** required?
- Is **dashboard** required?
- What is future **roadmap**?

Problem and Expected Output

DATA DISCOVERY & SAMPLING

3

ASK QUESTIONS?

- Which are data **sources**?
- What is data **type** and **size**?
- Is data **available**?
- Which are the essential **variables**?
- What are data **quality** problems?
- What is expected data **pipeline**?

Data Discovery & Sampling

3.1

POPULATION & SAMPLE

POPULATION & SAMPLE

Data repositories may have millions of observations and over hundreds of variables.

If the volume of data is extremely large (thousands of observations or more), then it is not necessary to use all the data for analysis.

High volume data may have **quality problems** and may have **unnecessary variables** required to solve the business problem in hand.

When dealing with such large data, it is best practice to extract a sample for analysis before developing models.

A sample is **representative** from the entire **population** of data.

SAMPLE SIZE

There are no definite rules to decide the size of the sample. However it should be large enough to contain **significant information** to solve the problem, yet small enough to be **manipulated quickly**. It is also important that sample should cover **all cases** required in learning of algorithms. The algorithms typically are more effective for use of more data in **training of algorithms**.

When obtaining a representative sample, it is also important not to **carelessly discard variables** from consideration. It is generally best to include as many **variables** as possible in the sample. After **exploring the data** with descriptive statistics and visualization, the analyst can **eliminate variables** that are not of interest.

THUMB RULE FOR SAMPLE SIZE

- There are **no definite rules** for the size of the sample.
- For prediction tasks, a rule of thumb is to have sample equal or more than **ten times of variables**.
- For classification tasks, a rule of thumb is to have at least $\mathbf{6 \times m \times q}$ observations, where *m* is **number of outcome** categories and *q* is the **number of variables**.
- When we are interested in predicting a rare event, such as click-through on an advertisement posted on a Web site, it is recommended that the training set oversample the number of observations corresponding to the rare events to provide the machine learning algorithm sufficient data to learn about the rare events.

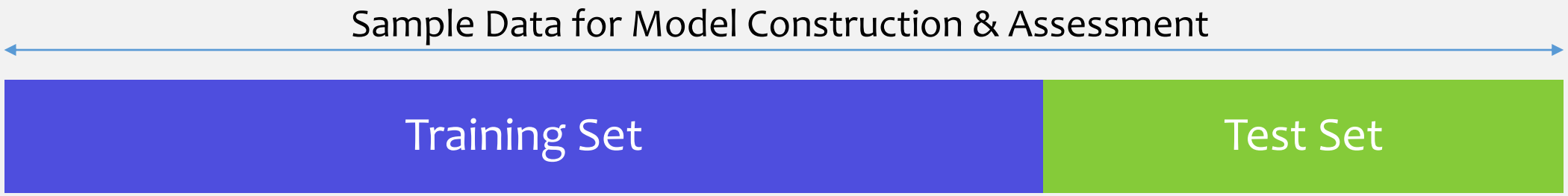
3.2

TRAINING, TEST & VALIDATION DATA

TRAINING & TEST DATA

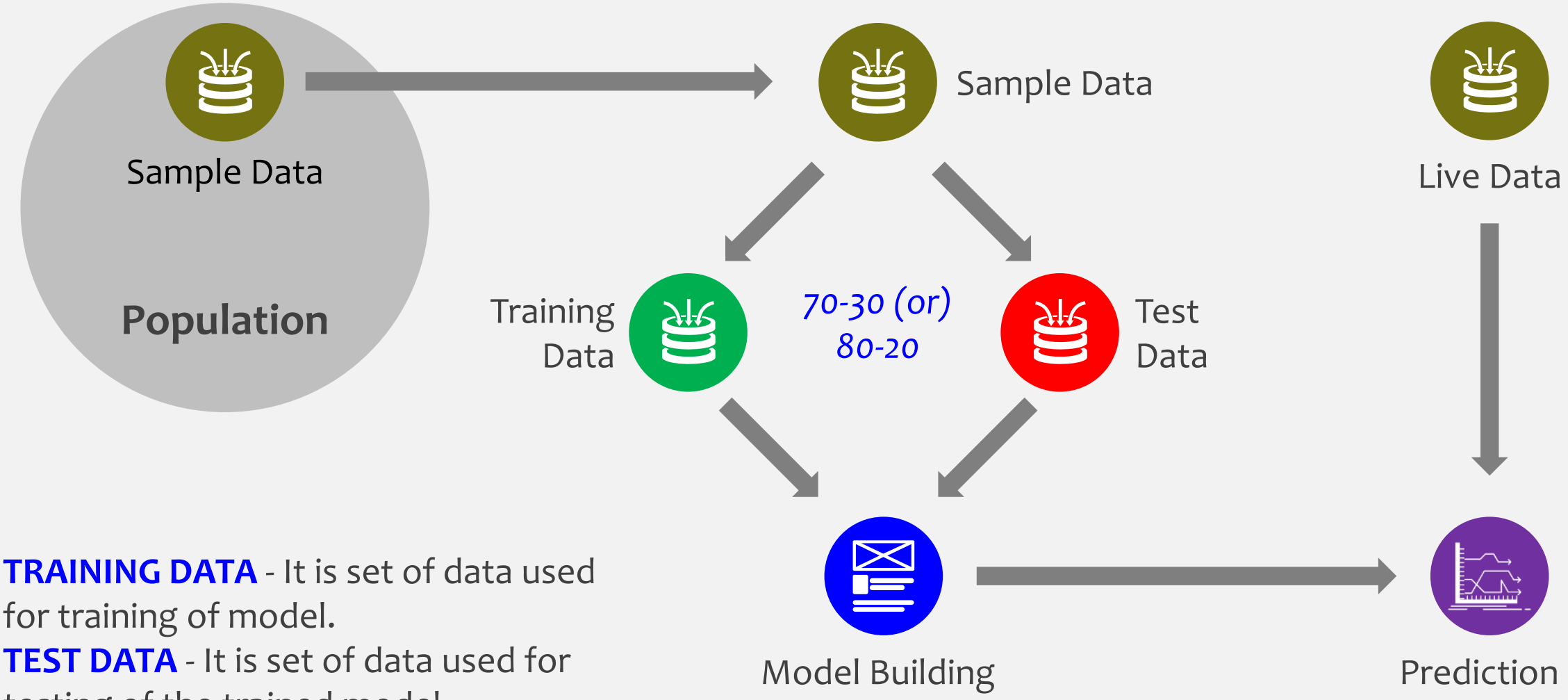
The **TRAINING SET** consists of the data used to build the analytics models. For example, a training set can be used to estimate the slope and coefficients in a multiple regression.

The final model should be evaluated with the **TEST SET** in order to estimate model effectiveness when applied to data that have not been used to build or select the model.



Reference - Brian Ripley, page 354, Pattern Recognition and Neural Networks, 1996

TRAINING & TEST DATA



- **TRAINING DATA** - It is set of data used for training of model.
- **TEST DATA** - It is set of data used for testing of the trained model.

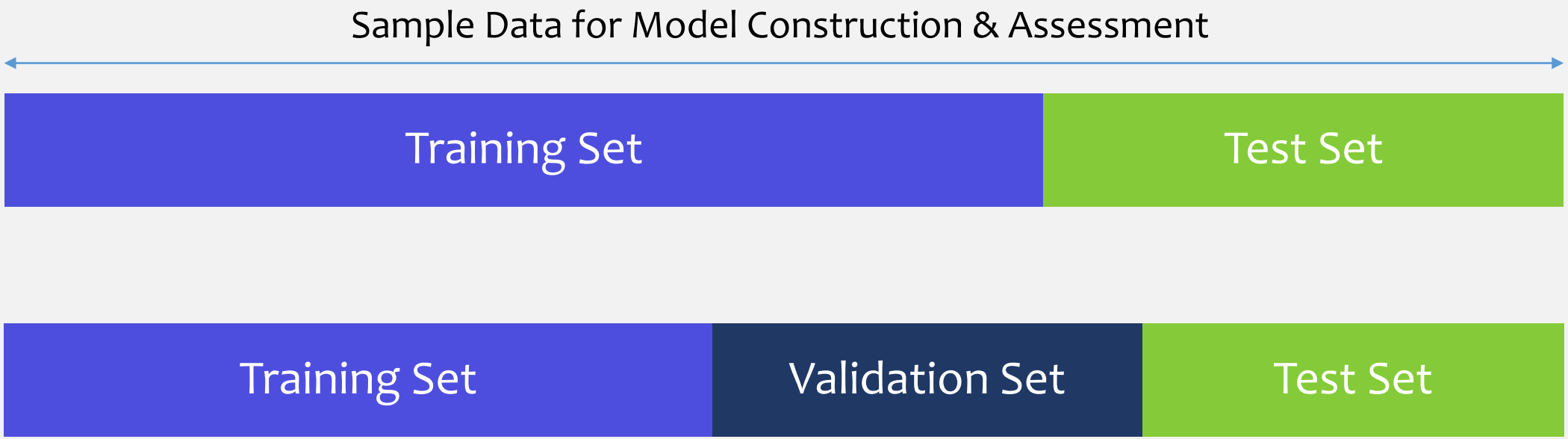
VALIDATION & TEST DATA

The training data is used to construct a model. This model should be evaluated on samples that are not used to build or fine-tune the model, so that model provide an unbiased sense of effectiveness.

The validation set approach involves randomly dividing the available set of observations into two parts, a training set and a validation set or hold-out set. The model is constructed on the training set, and the fitted model is used to predict the responses for the observations in the validation set. The resulting validation error helps in fine-tuning the model.

A **VALIDATION SET** is a sample of data held back from training a model. The validation data is used in tuning hyper-parameters of model. Both **TEST** and **VALIDATION** dataset are used to qualify performance of trained model.

VALIDATION & TEST DATA



Validation data may or may not be used. There is no standard way to split the sample data, but can be done as 70:20:10.

Reference - Brian Ripley, page 354, Pattern Recognition and Neural Networks, 1996

3.3

RESAMPLING

RESAMPLING

- Resampling is the method of drawing **repeated samples** from the original data samples.
- Resampling involves the selection of **randomized cases** with replacement from the original data sample in such a manner that each number of the sample drawn has a number of cases that are similar to the original data sample.
- Resampling generates a **unique sampling distribution** on the basis of the actual data.
- Resampling approaches can be computationally expensive, because it involves fitting the same statistical method **multiple times** using different subsets of the training data.
- The most commonly used resampling methods, **cross-validation** and the **bootstrap**.

DATA PREPARATION

4

DATA PREPARATION

Data preparation is the process of **CLEANING** and **TRANSFORMING** raw data so that it cab be used in analytics.

Data preparation tasks are as below.

- DATA CLEANING - Identify and correct errors in the data.
- FEATURE SELECTION - Identify input variables which are most relevant to the task.
- DATA TRANSFORMATION - Change the scale or distribution of variables.
- FEATURE ENGINEERING - Derive new variables from available data.
- DIMENSIONALITY REDUCTION - Creating compact projections of the data.

METHODS, ALGORITHMS & MODELS

5

5.1

METHODS & ALGORITHMS

ASK QUESTIONS?

- Which machine learning method required for given data and problem statement?
- Which is suitable algorithm?
- What type of visualization is required?
- Which are the model evaluations metrics?

Methods & Algorithms

MACHINE LEARNING METHODS

01 REGRESSION

02 CLASSIFICATION

03 CLUSTERING

04 ASSOCIATION MINING

05 ENSEMBLE METHODS

and many more +

ALGORITHM v/s MODEL

Machine learning involves the use of machine learning algorithms and models

ALGORITHM

Machine learning algorithm is a procedure that is run on training data to create a machine learning model.

MODEL

A model is the output of machine learning algorithm run on training data.

A model is the outcome of learning by machine learning algorithm.

A machine learning model is a file that is trained to do some task may be predictions.

MACHINE LEARNING ALGORITHMS

Machine learning algorithm is a procedure that is run on training data to create a machine learning model.

- Linear Regression
- Logistic Regression
- Classification and Regression Trees
- Naive Bayes
- K-Nearest Neighbors (KNN)
- Support Vector Machines (SVM)
- Random Forest
- Boosting
- AdaBoost
- Apriori
- K-means
- Linear Discriminant Analysis
- Principal Component Analysis

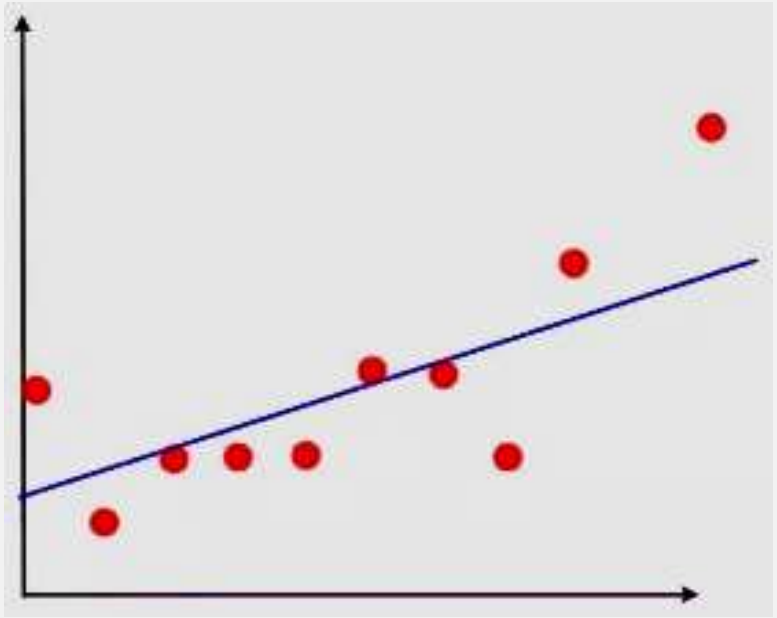
5.2

ML MODELS

MODELS (1/3)

Simple Model

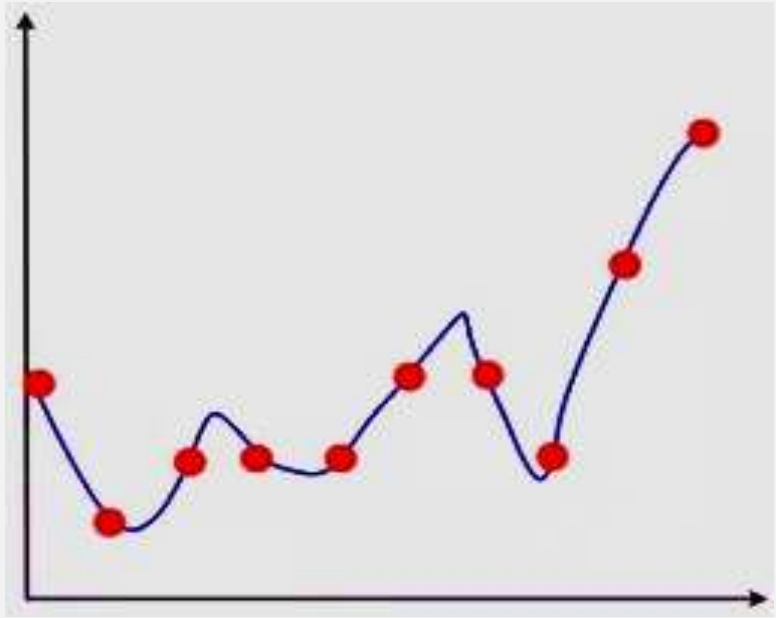
$y = mx + c + \varepsilon$



Model is not passing through all dataset.

Complex Model

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0,$$

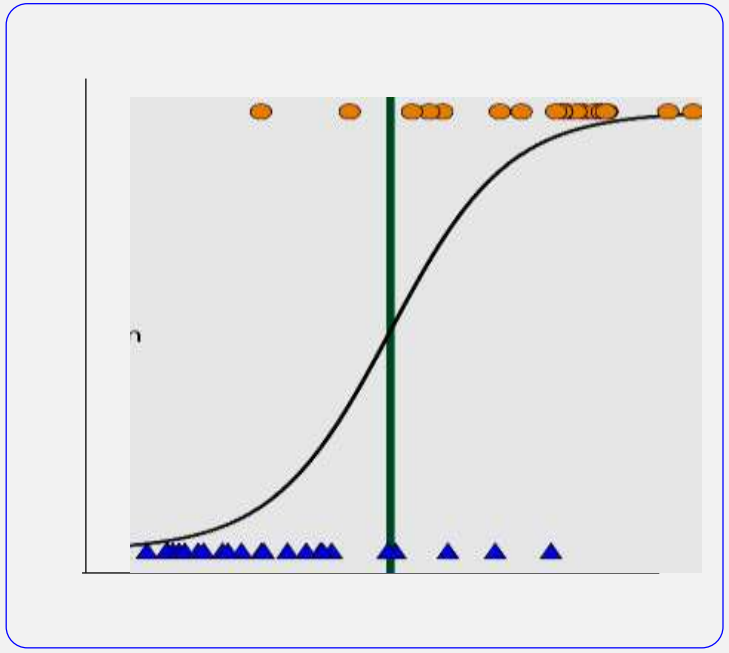


Model is passing through all dataset.

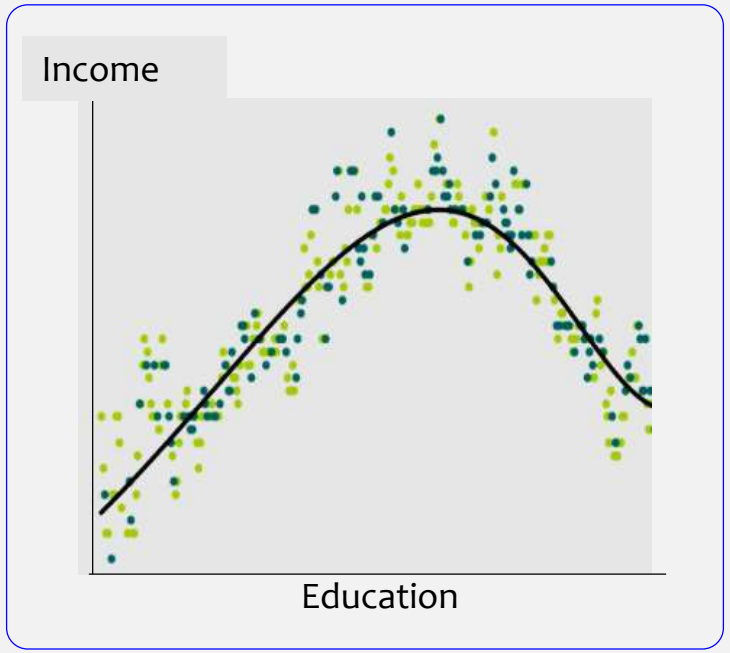
MODELS (2/3)

Linear Regression

Logistic Regression

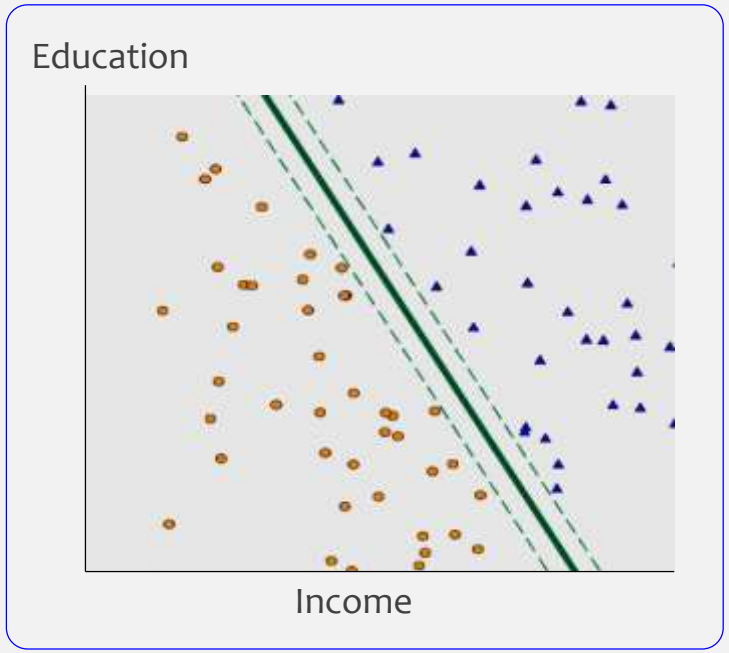


Polynomial Regression

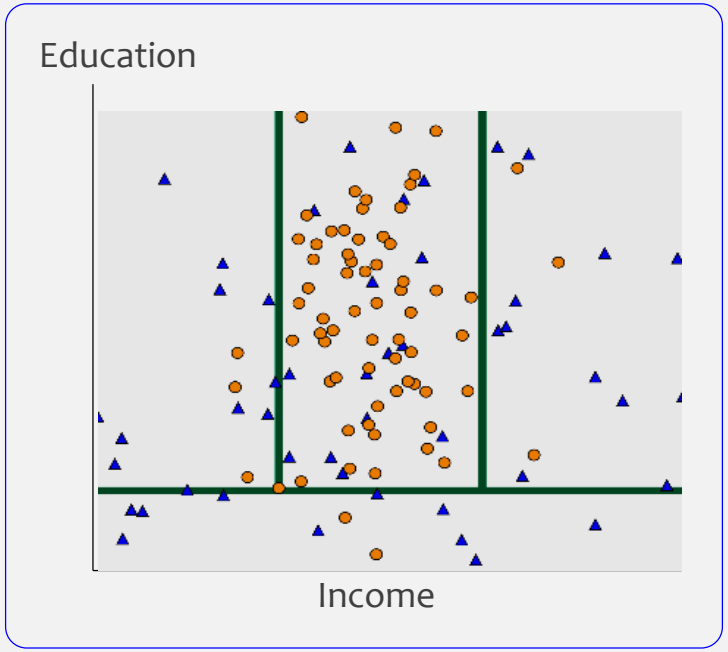


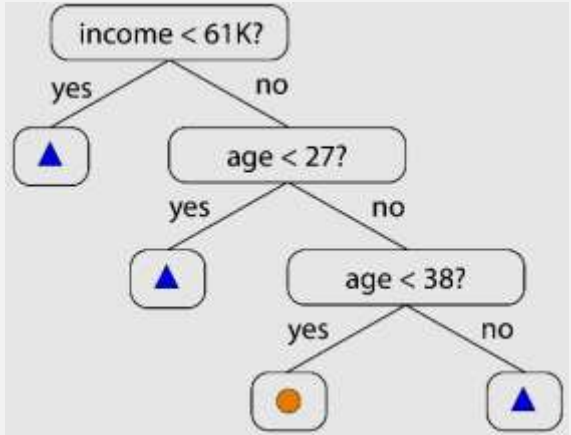
MODELS (3/3)

Classification

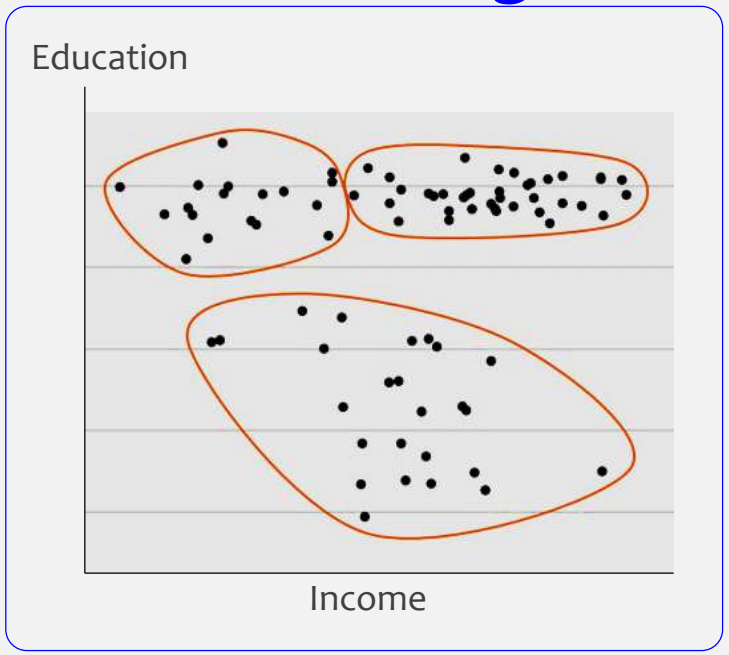


Decision Tree





Clustering



TERMINOLOGY – STAT AND ML

Statistics and machine learning uses different terminology for same words.

Statistics	Machine Learning
Variable	Feature
Dependent variable	Label
Transformation	Feature creation

5.2.1 PARAMETRIC AND NONPARAMETRIC METHODS

PARAMETRIC AND NONPARAMETRIC METHODS (1/3)

PARAMETRIC

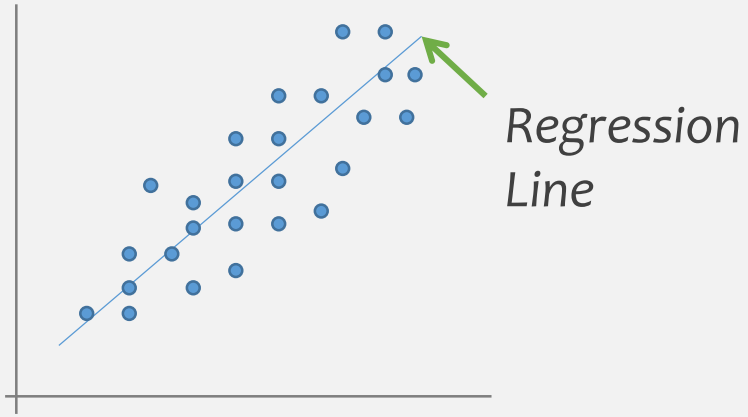
- Algorithms that simplify the function to a known form are called parametric machine learning algorithms.
- Parametric methods can be parametrized by a finite number of parameters
- Assume functional form or shape for estimate (f) such as linear regression. $f(X) = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta pXp$

NONPARAMETRIC

- Algorithms that don't make strong assumptions about the form of mapping function are called nonparametric machine learning algorithms.
- Nonparametric methods cannot be parametrized by a fixed number of parameters.
- Non-parametric methods don't assume functional form of " f ".

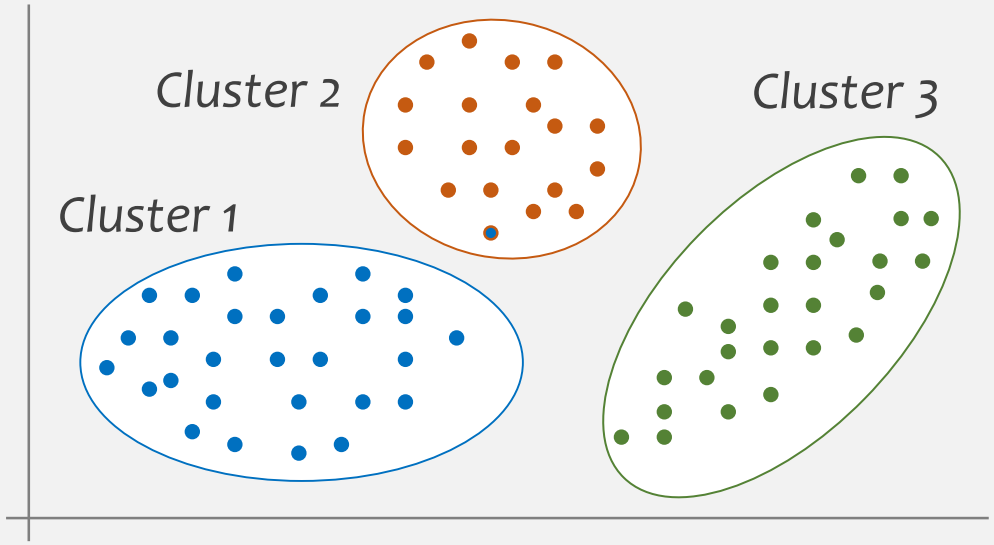
PARAMETRIC AND NONPARAMETRIC METHODS (2/3)

PARAMETRIC METHOD



Example *- Find height, when weight is given.*

NON-PARAMETRIC METHOD



Example *- Discover groups of customers for targeted marketing program.*

PARAMETRIC AND NONPARAMETRIC METHODS (3/3)

PARAMETRIC

Advantage

- **Simple Form**: Problem is simplified because functional form is assumed.
- **Easy Learning**: Learning is easier from small set of data.
- **Less Data**: Parametric method can work well for smaller data set.

Disadvantage

- **Incorrect Form**: Model selection may not be incorrect because true form of " f " is unknown.
- **Poor Fit**: Flexible model may be selected to fit many different possible functional forms.
- **Overfitting**: Flexible model selection may make model complex, which may lead to over-fitting the data, hence more noise in model.

NONPARAMETRIC

Advantage

- **Flexibility**: Capable of fitting large number of functional forms
- **No Prior Knowledge** is required to fit functional form.
- **Accuracy**: As there is no assumption on functional form, it results in the better accuracy of model.

Disadvantage

- **Complex Form**: Problem is not simplified as there is no assumption.
- **Large Data:** Very large number of observations are required to accurate estimate for " f ".

5.2.2 GENERALIZATION, OVERFITTING & UNDER-FITTING

GENERALIZATION (1/2)

Generalization is a term used to describe a model's ability to react to new data. OR After training of model on training dataset, how accurately model predict values on new dataset.

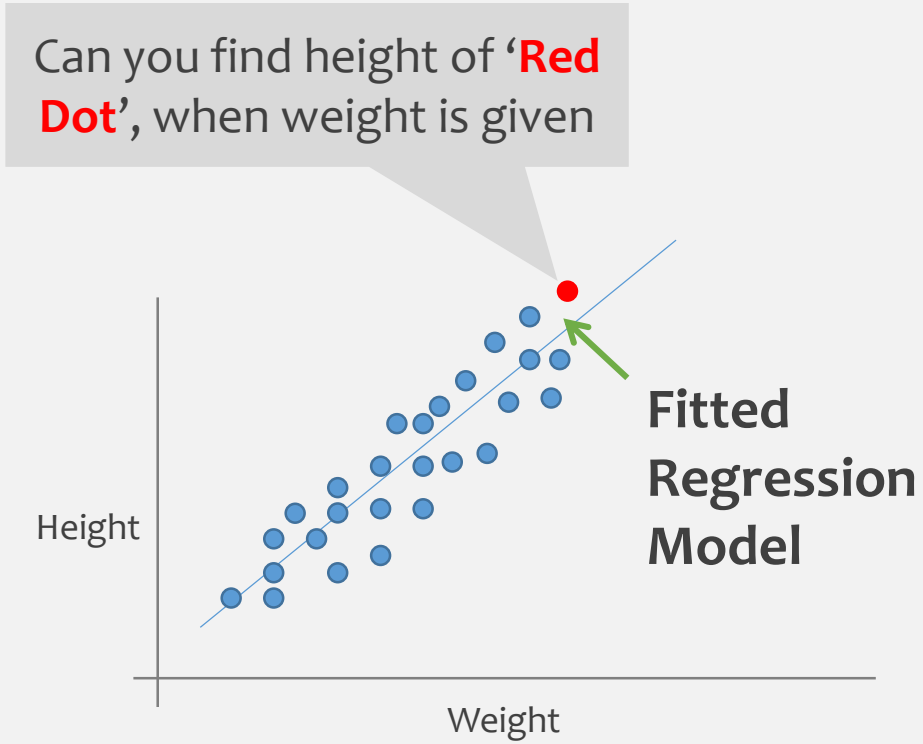
The goal of a good machine learning model is to generalize well from the training data to any data from the problem domain. This allows us to make predictions in the future on data the model has never seen.

If model is making inaccurate predictions when given new data, then the model useless; even though it is able to make accurate predictions for the training data.

Reference : *https://wp.wwu.edu/machinelearning/2017/01/22/generalization-and-overfitting/*

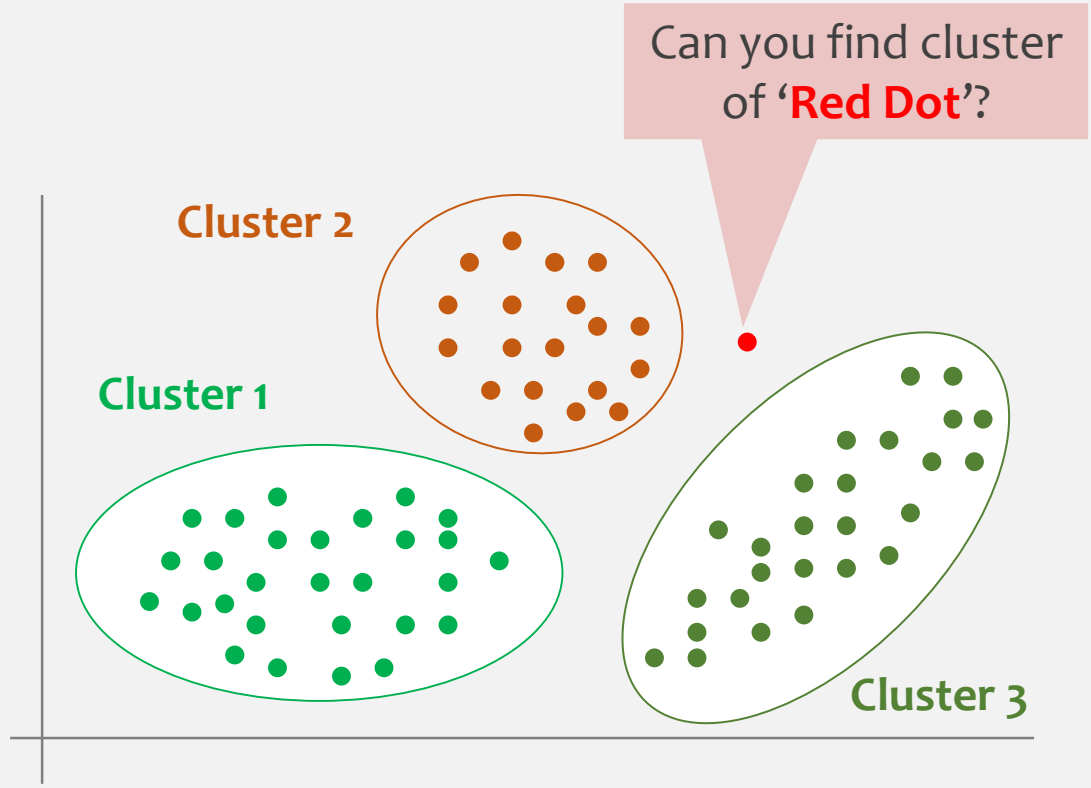
GENERALIZATION (2/2)

Generalization ☑



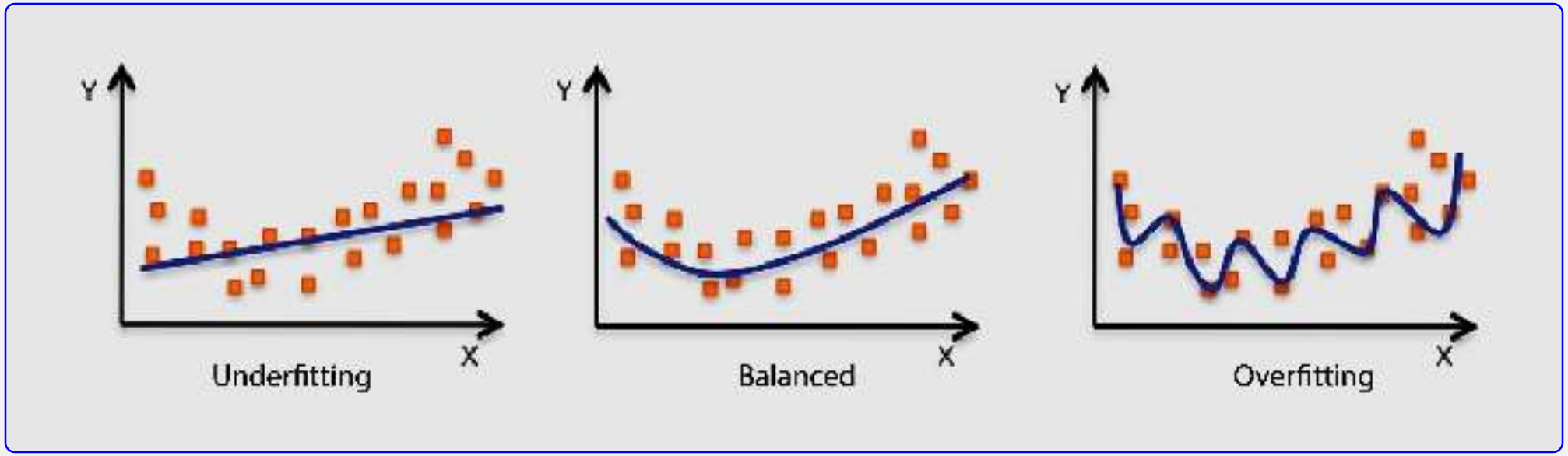
Find height, when weight is given.

Generalization ☒



Discover groups of customers, which can be used for targeted marketing program.

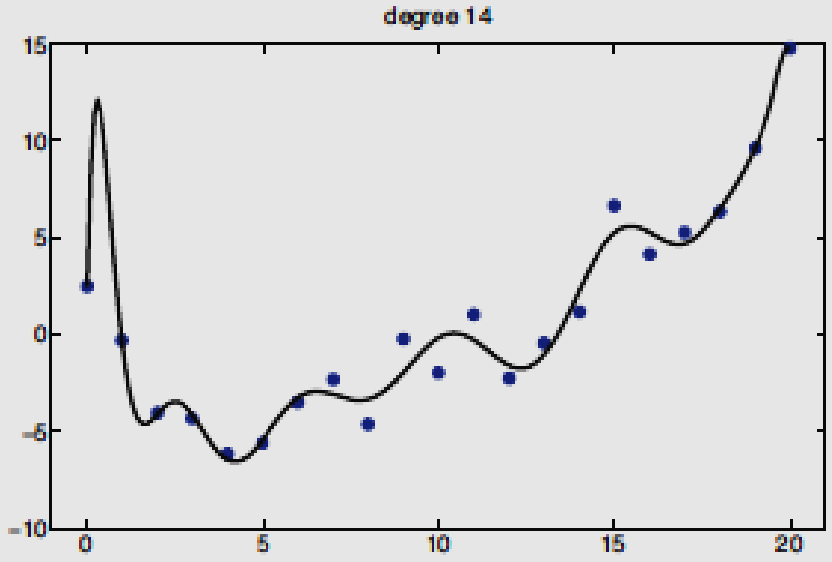
Overfitting and **Underfitting** are the two biggest causes for poor **performance** of machine learning algorithms.

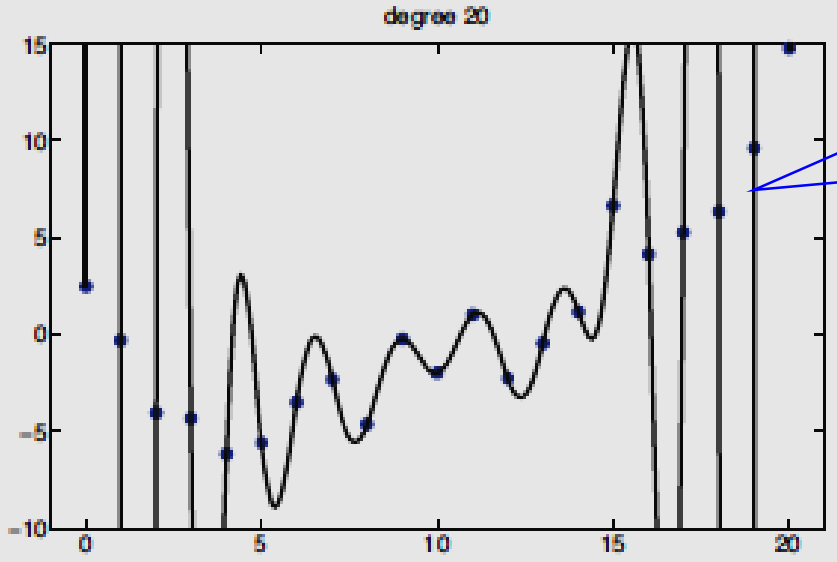


Reference :
https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/
https://wp.wwu.edu/machinelearning/2017/01/22/generalization-and-overfitting/
http://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

OVERFITTING

- When model is more complex or highly flexible, then it can lead to a phenomenon known as overfitting the data.
- Overfitting means model follow the errors, or noise, too closely.
- Modelling every minor variation in the input, is more likely using every to be noise in data than true signal.



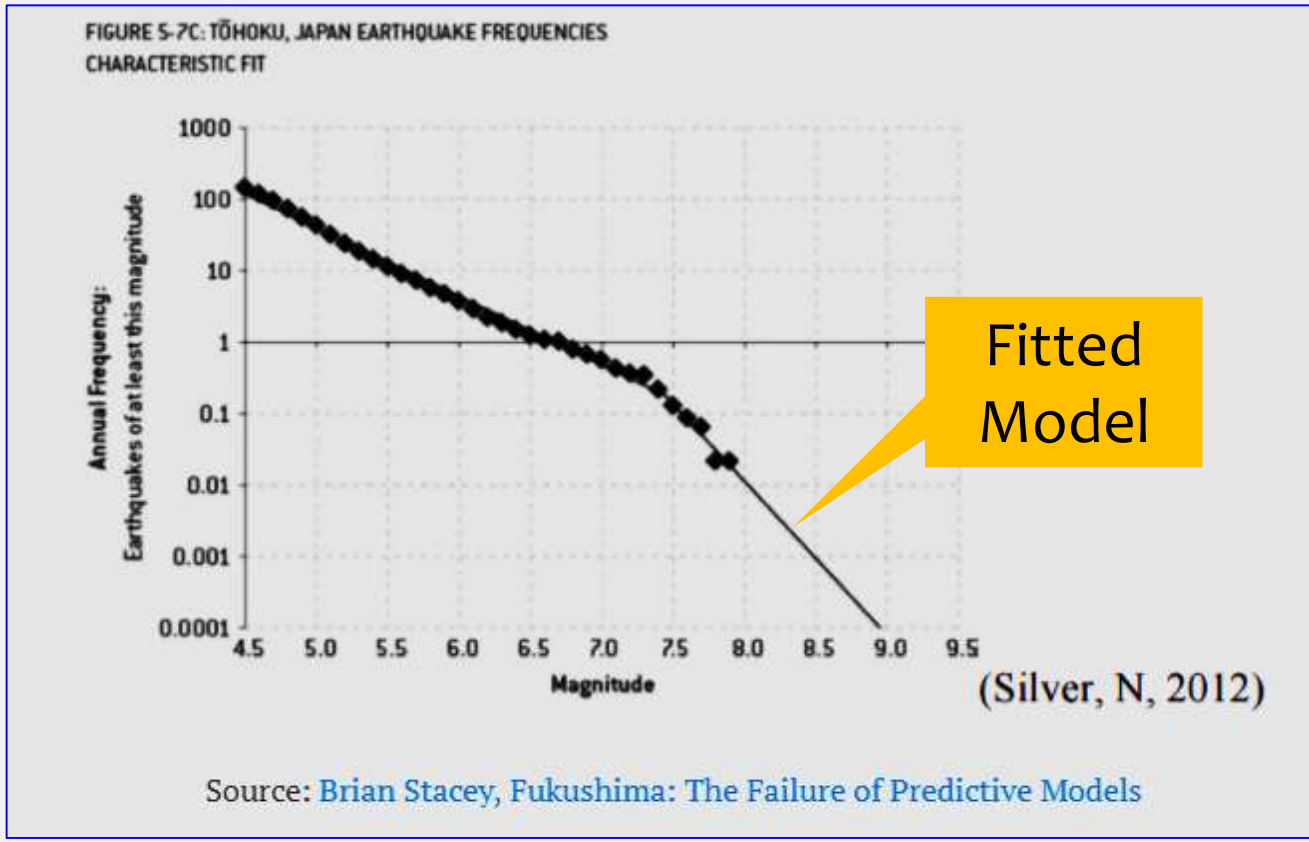


Overfitted Model

Reference : *Kevin P. Murphy, "Machine Learning - A Probabilistic Perspective"*

OVERFITTING - FUKUSHIMA POWER PLANT DISASTER (1/3)

- The Fukushima power plant disaster is a example of overfitting. Engineers designed power plant based on - how often earthquakes would occur.
- Earthquake data from the past 400 years was used to train a regression model.
- The diamonds represent actual data while the thin line shows the fitted regression.

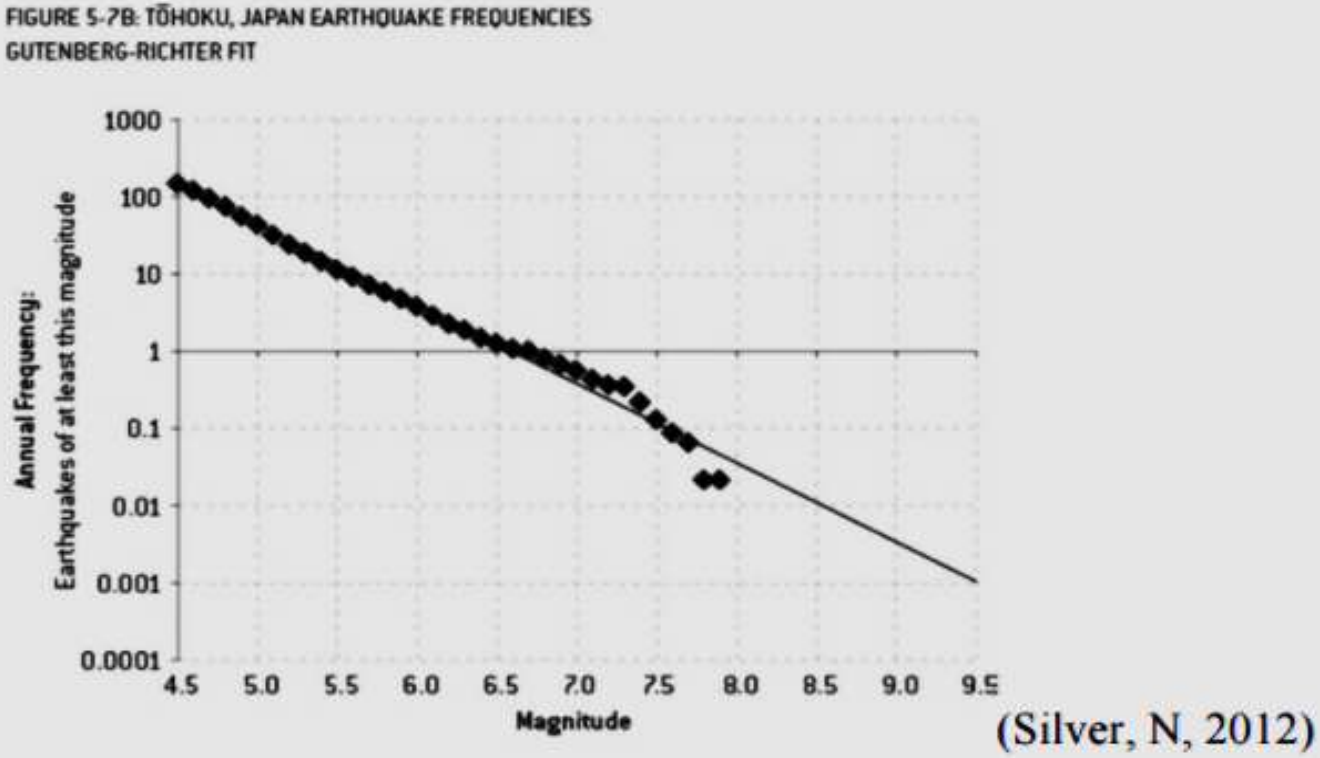


Source: Brian Stacey, Fukushima: The Failure of Predictive Models

Model hugs the data points very closely. In fact, model makes a kink at around a magnitude of 7.3 – decidedly not linear.

Reference: *https://ml.berkeley.edu/blog/2017/07/13/tutorial-4/*

OVERFITTING - FUKUSHIMA POWER PLANT DISASTER (2/3)

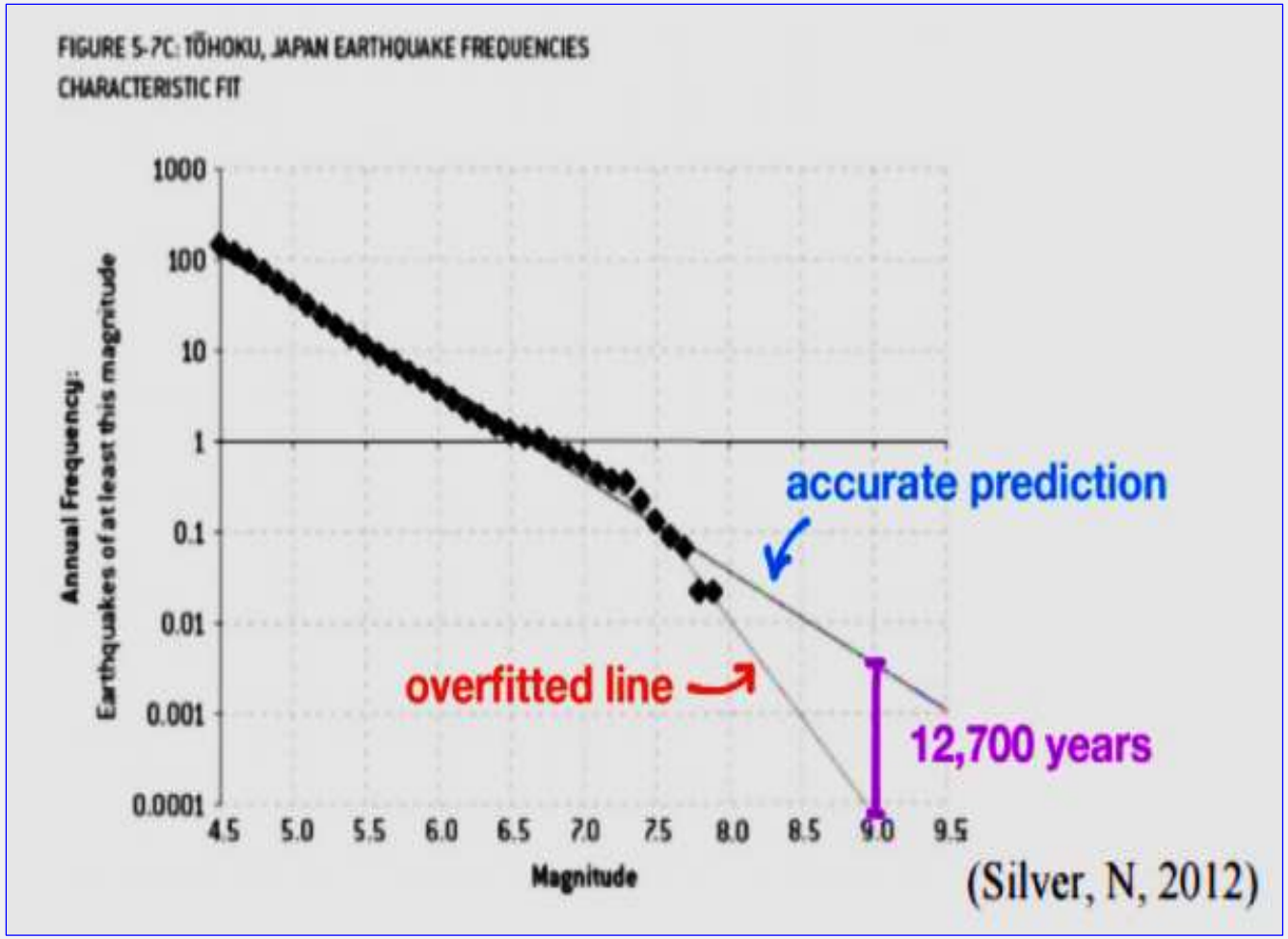


Notice there's no kink this time, so the line isn't as steeply sloped on the right.

If the engineers had used the correct linear model, their results would have looked something like this.

Reference : *https://ml.berkeley.edu/blog/2017/07/13/tutorial-4/*

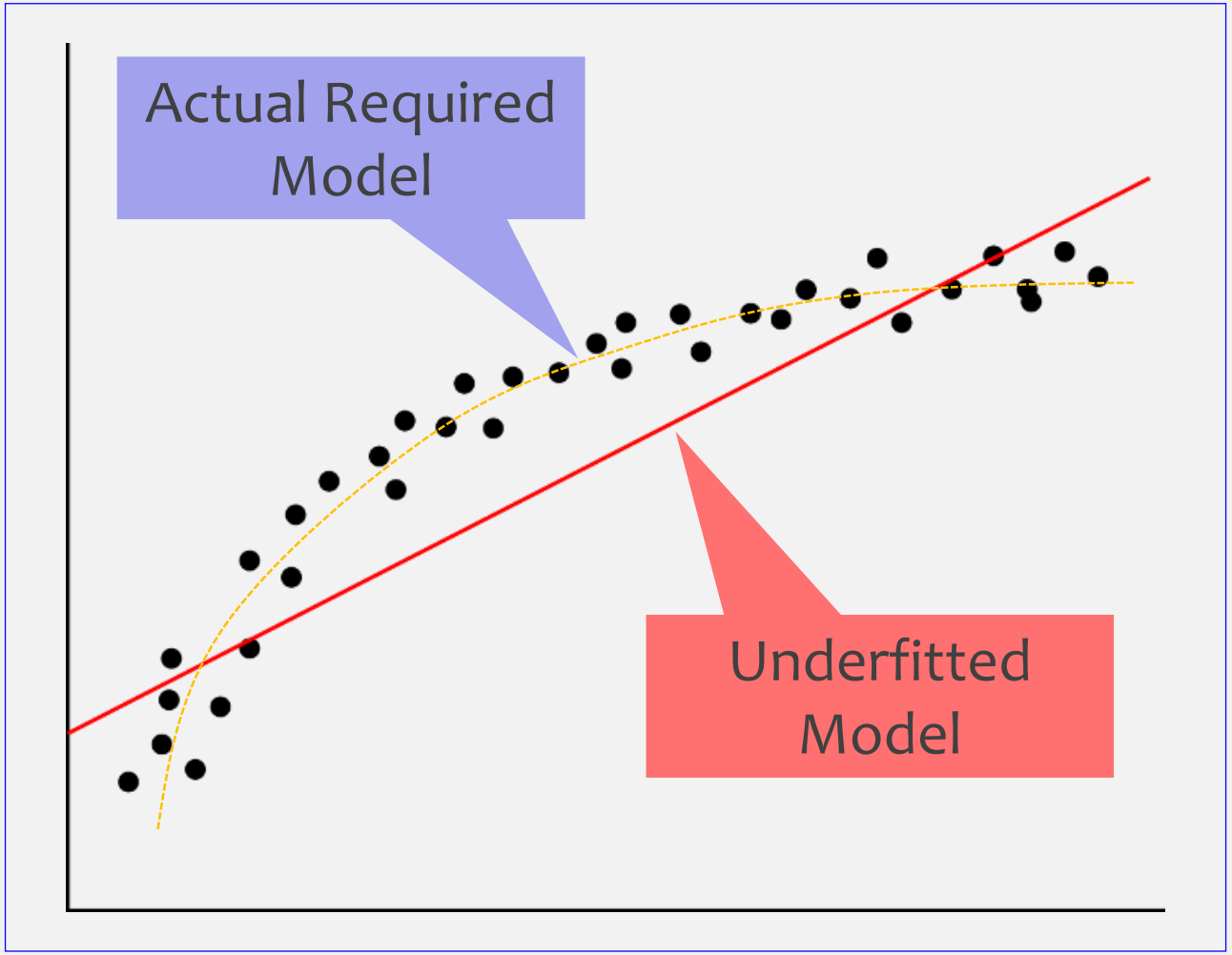
OVERFITTING - FUKUSHIMA POWER PLANT DISASTER (3/3)



- The overfitted model predicted earthquake of magnitude 9 about every 13000 years while the correct model predicted earthquake of at least magnitude 9 just about every 300 years.
- Because of this, the Fukushima Nuclear Power Plant was built to withstand an earthquake of magnitude 8.6.
- In 2011 earthquake that devastated the plant was of magnitude 9.

Reference: *https://ml.berkeley.edu/blog/2017/07/13/tutorial-4/*

UNDERFITTING



- Underfitting happens when a model has not been trained enough on the data.
- Underfitted model fails to capture relevant information.
- In the case of Underfitting, it makes the model just as useless and it is not capable of making accurate predictions, even with the training data.

Reference: *https://wp.wwu.edu/machinelearning/2017/01/22/generalization-and-overfitting/*
https://ml.berkeley.edu/blog/2017/07/13/tutorial-4/

ERROR IN OVERFITTING AND UNDERFITTING

	ERROR	
	Overfitting	Underfitting
TRAINING of MODEL	LOW	HIGH
TESTING of MODEL	HIGH	HIGH

HOW TO AVOID OVERFITTING?

There are several methods to avoid overfitting. Few methods are listed as below.

CROSS VALIDATION

Data is divided into training data and test data. Model is built on training data and then validated on test data.

REGULARIZATION

It adds a penalty on the different parameters of the model to reduce the freedom of the model. Hence, the model will be less likely to fit the noise of the training data and will improve the generalization abilities of the model.

EARLY STOPPING

Stop training process at early stage, thus reducing number of iterations.

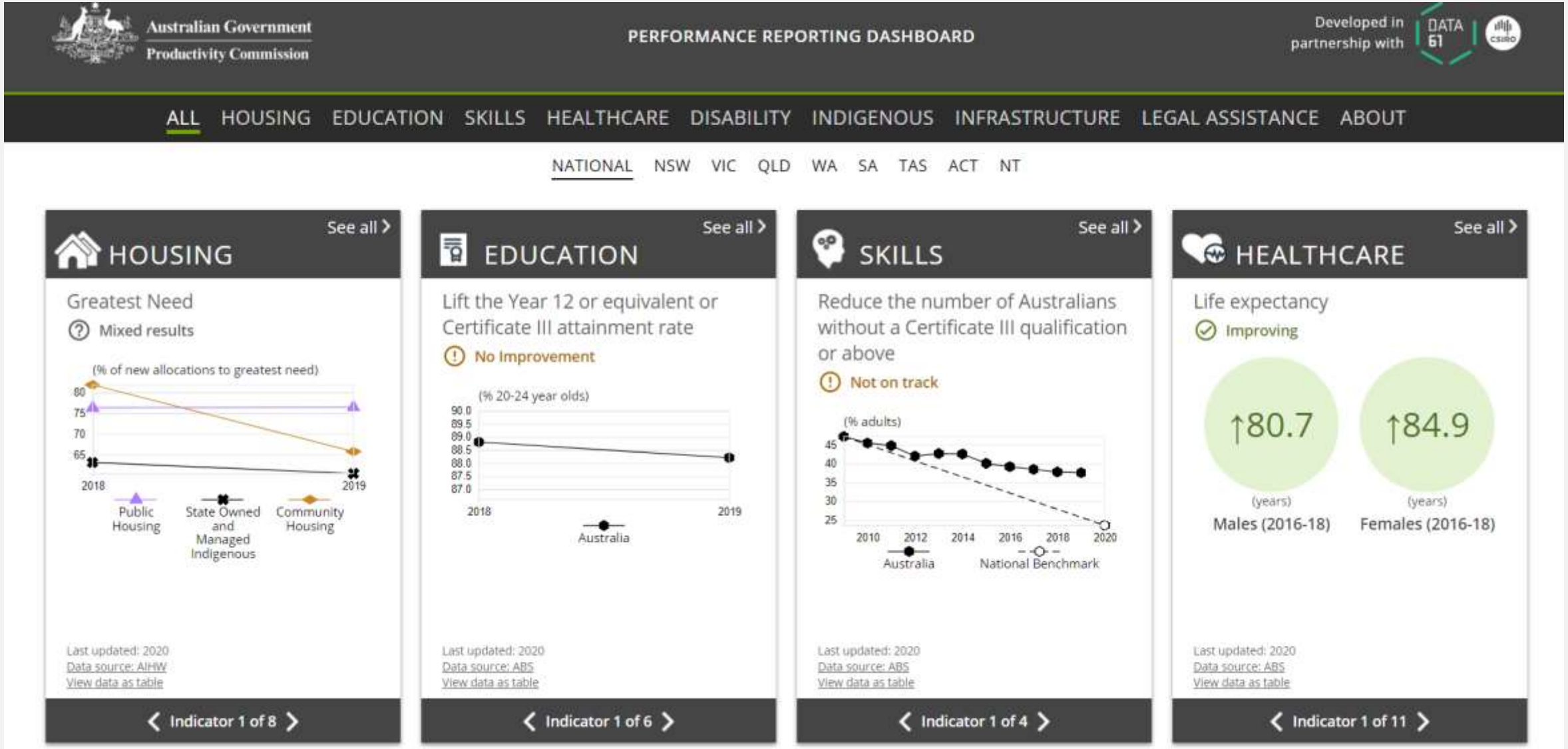
FEATURES

New features can be added or existing features can be eliminated.

5.3

VISUALIZATION FOR MODEL OUTPUT

DASHBOARD

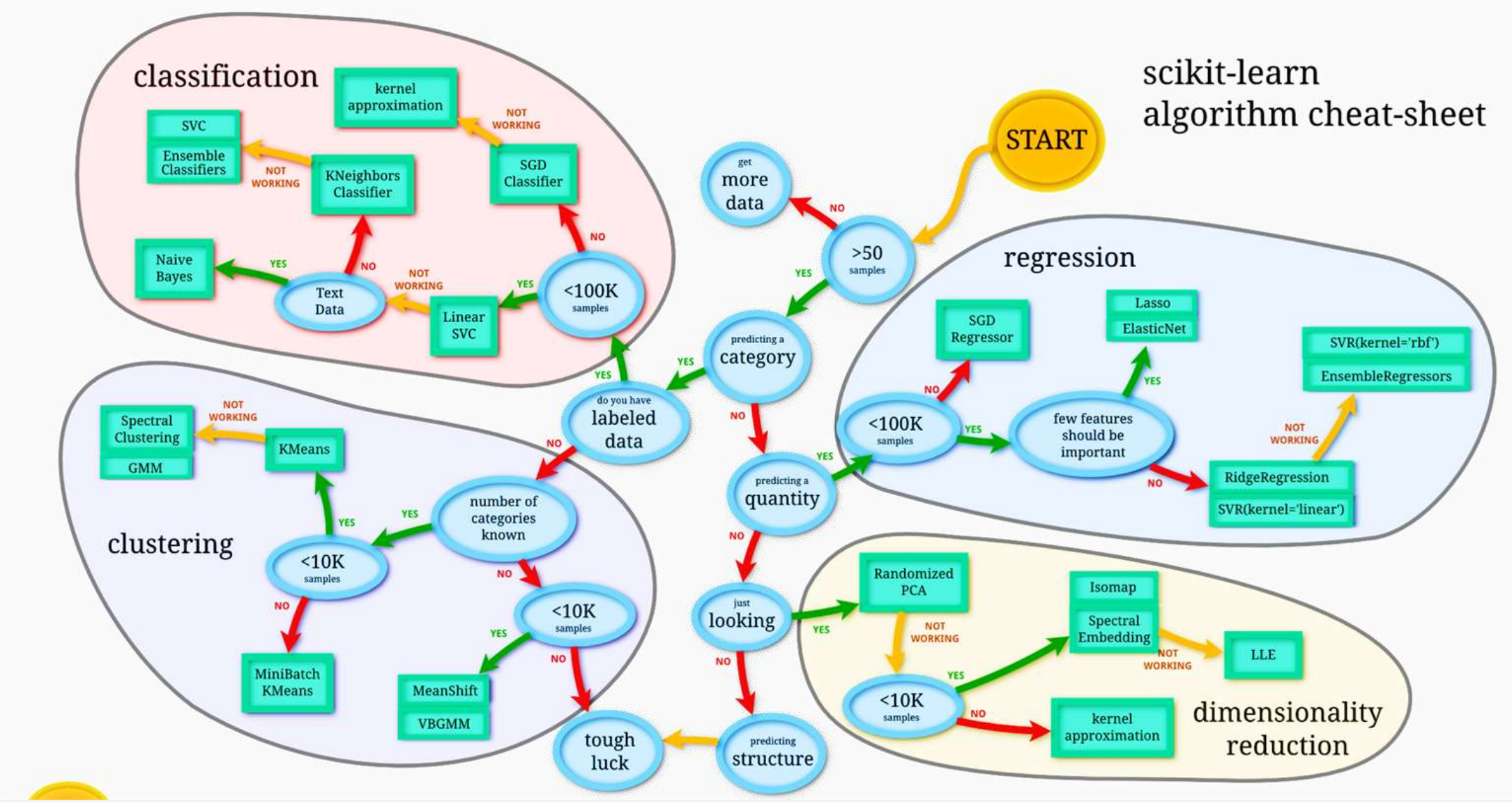


https://performancedashboard.d61.io/aus

SELECT METHODS & ALGORITHMS

6

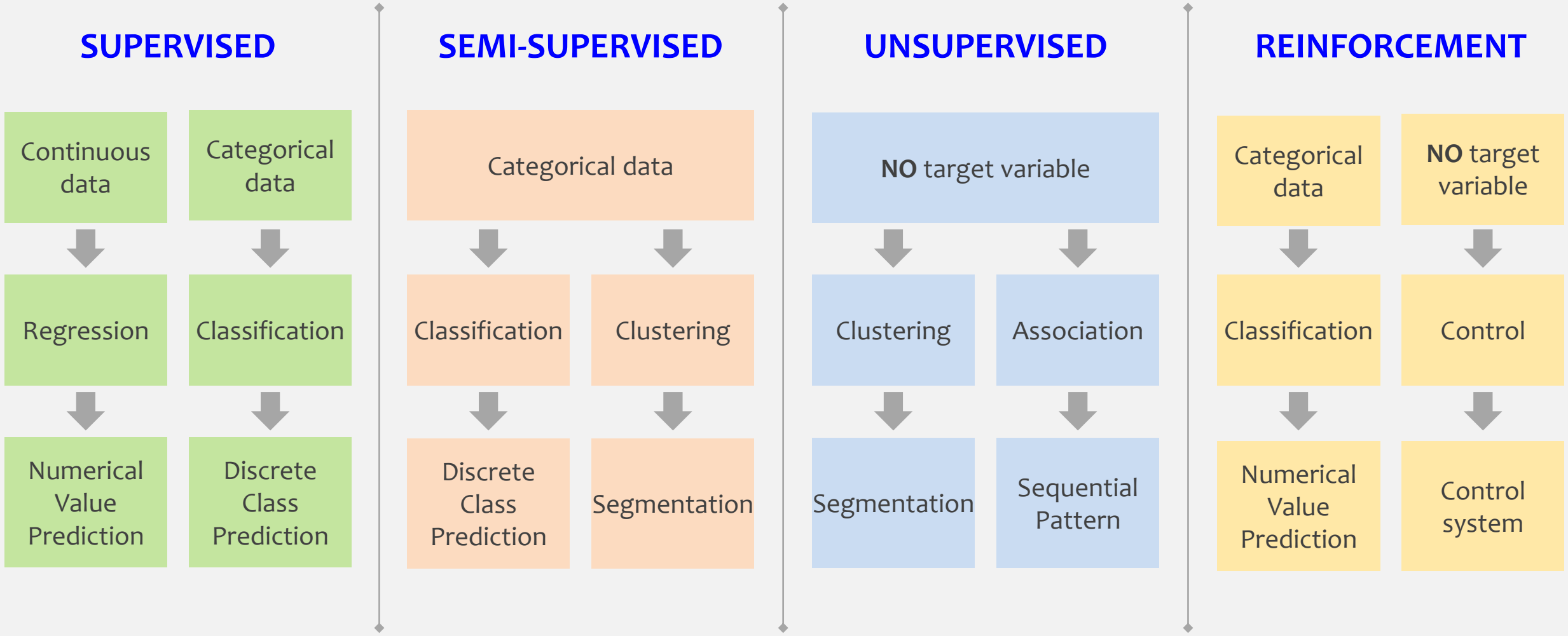
CHEAT-SHEET TO SELECT METHOD & ALGORITHM



MACHINE LEARNING STYLES

Learning Styles	SUPERVISED	SEMI-SUPERVISED	UNSUPERVISED	REINFORCEMENT
Input Data	All data labelled	Partially labelled & rest unlabeled	All data unlabeled	Data may be labelled or unlabeled
Target Variable	Continuous or categorical variable	Categorical variable	No target variable	Categorical or No target variable
Model Type	Predictive	Predictive and Descriptive	Descriptive *(Summary, patterns, rules)*	Continuous iterative learning

SELECT MACHINE LEARNING STYLE



FACTORS TO CHOOSE AN ALGORITHM (1/5)

Parameters affecting decision to choose the right algorithms are as below.

1. Number of features
2. Linearity of data
3. Size of training data
4. Training time
5. Interpretation and flexibility of a model
6. Memory requirement

FACTORS TO CHOOSE AN ALGORITHM (2/5)

1. NUMBER OF FEATURES

- A feature is a variable used in machine learning.
- A large number of features requires more time to train the model. Hence feature selection is important step in model development.
- Support vector machine (SVM) is suitable when features are high in numbers. Hence, SVM is suitable in applications like information retrieval, text classification and image classification. SVM can be used for both classification and regression tasks.

2. LINEARITY OF DATA

- If the data can be linearly separable or if it can be represented using a linear model, then algorithms like SVM, linear regression or logistic regression can be used. Otherwise, deep neural networks or ensemble models can be used.

FACTORS TO CHOOSE AN ALGORITHM (3/5)

3. SIZE OF TRAINING DATA

- If the training data has a less number of observations and a high number of features, then select algorithm with high bias and low variance like Linear regression, Naïve Bayes, or Linear SVM.
- If the training data is appropriately large and the number of observations is higher as compared to the number of features, then select algorithm with low bias and high variance algorithms like KNN, Decision trees, or kernel SVM.

4. SIZE OF TRAINING DATA

- More training time means more accuracy of a model. Also large training data may require more time.
- Naïve Bayes and Linear regression require less time. While SVM may require more time to solve the same problem with same training data. Neural networks and random forests are highly time consuming.

FACTORS TO CHOOSE AN ALGORITHM (4/5)

5. INTERPRETATION AND FLEXIBILITY OF A MODEL

- A highly interpretable algorithms can easily indicate association of predictor with the response, whereas flexible models give higher accuracy but low interpretability. Hence there is a trade-off between accuracy and interpretability. Neural networks are highly accurate but are black box in interpretation.
- Restrictive algorithms generates small range of shapes of the mapping function. Linear regression is a restrictive approach as it generates linear functions like lines.
- Flexible algorithms generates a wider range of possible shapes of the mapping function. For example, KNN with k=1 is highly flexible as it will consider every input data point to generate the mapping output function. The below picture displays the trade-off between flexible and restrictive algorithms.
- For better inference, restrictive models are preferred. For better accuracy, flexible models are better. There is a trade-off between flexibility and interpretability. As flexibility increases, interpretability decreases.

FACTORS TO CHOOSE AN ALGORITHM (5/5)

6. MEMORY REQUIREMENT

Memory requirement depends upon the data size and models. Neural networks or support vector machines requires more memory as compared to regression models.

MODEL EVALUATION

7

MODEL EVALUATION TASKS

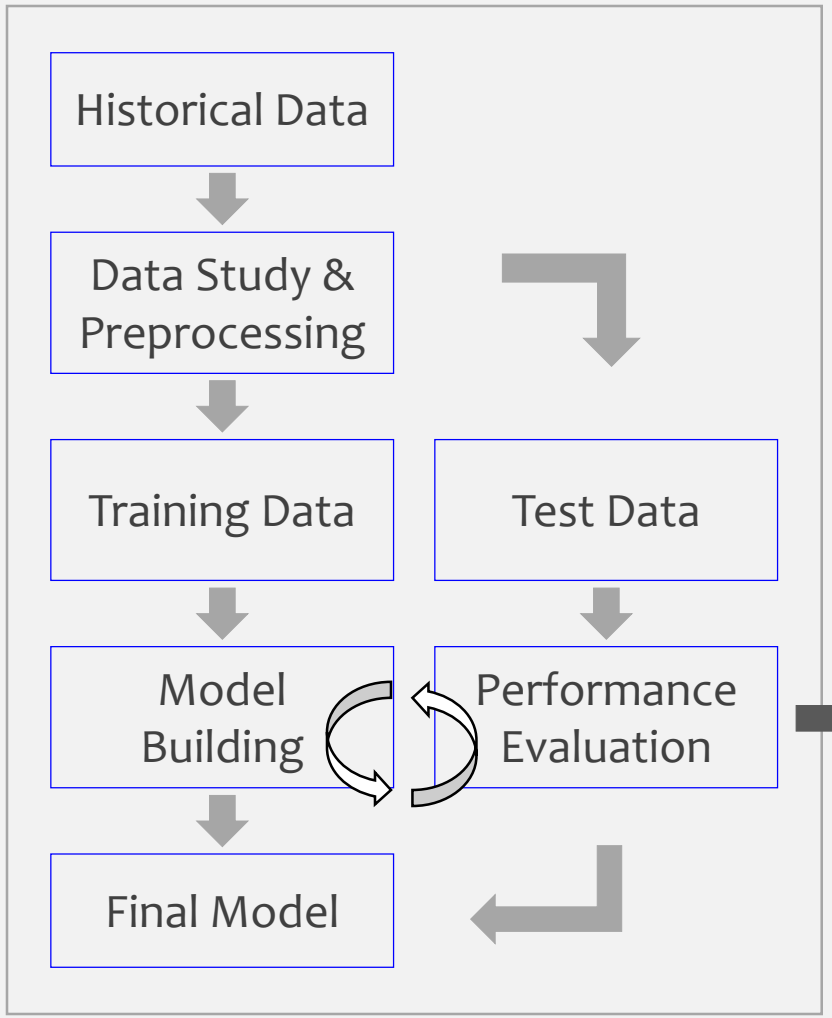
- Select a **metric** for evaluating model.
- Decide a model evaluation steps.
- Evaluate the **first model** using test data and evaluation metric.
- Tune algorithm **hyper-parameters**, if model is not within acceptable metric.
- Combine models into **ensembles** to enhance accuracy.

A hyperparameter is a parameter that is set before the learning process starts. These parameters are tunable and can be changed to improve accuracy of model. These parameters affect the model training.

Examples of hyperparameters in machine learning are -

- Learning Rate of neural network
- Number of branches in a decision tree
- Number of clusters in a clustering algorithm

QUALITY OF FIT



How to measure quality of fit?

ACCURACY	Measure correctness of predictions
CONFUSION MATRIX	Number of correct and incorrect classes
PER-CLASS ACCURACY	Average accuracy for each class
ROC CURVE	Shows sensitivity of classifier

And many more … …

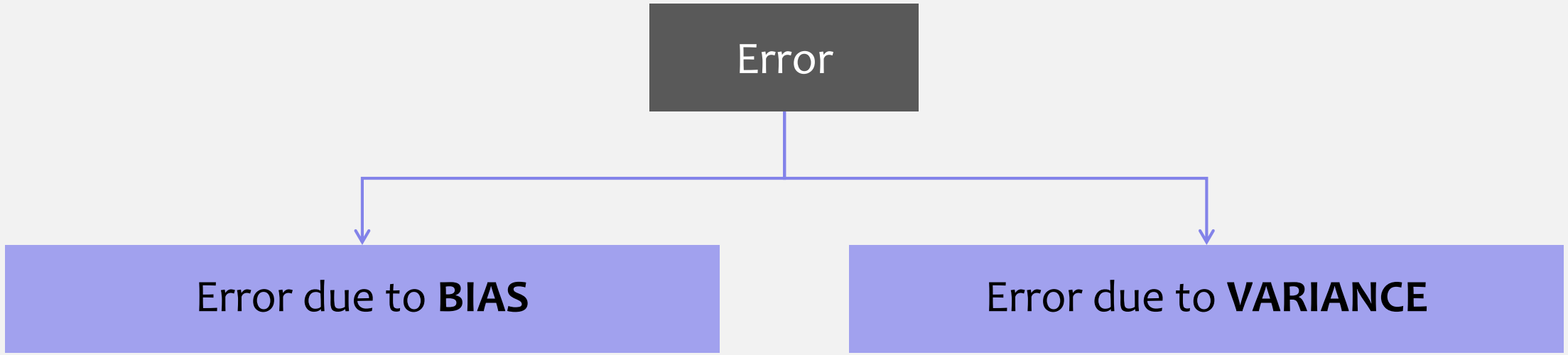
It is necessary to **quantify** the **qualify of fit** to compare which the predicted response value for a given observation is close to the true response value for that observation.

DISTRIBUTION DRIFT

Distribution of Drift - Practically distribution of data changes over time. This is called distribution of drift.

- Historical data is used for **Model Development and Evaluation**.
- **Model Deployment**: Model will be deployed for use on live data and predictions will be done.
- **Distribution of Data**: It is assumed that live data will have same distribution as Historical data.
- To find out distribution drift, the **performance of model** is continuously checked on live data.
- When performance of model starts **degrading**, then distribution might have drifted.
- If distribution drift occurs then model has to be **retrained**.

ERROR IN MODEL

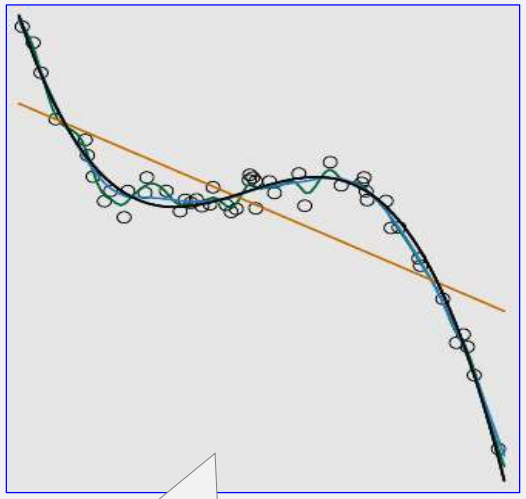


Bias refers to the error that is introduced by approximating a real-life problem.

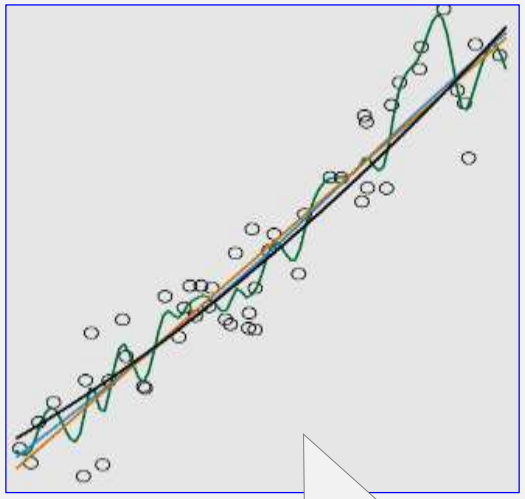
Variance refers to the amount by which estimate would change if we estimated it using a different training data set.

BIAS IN MODEL

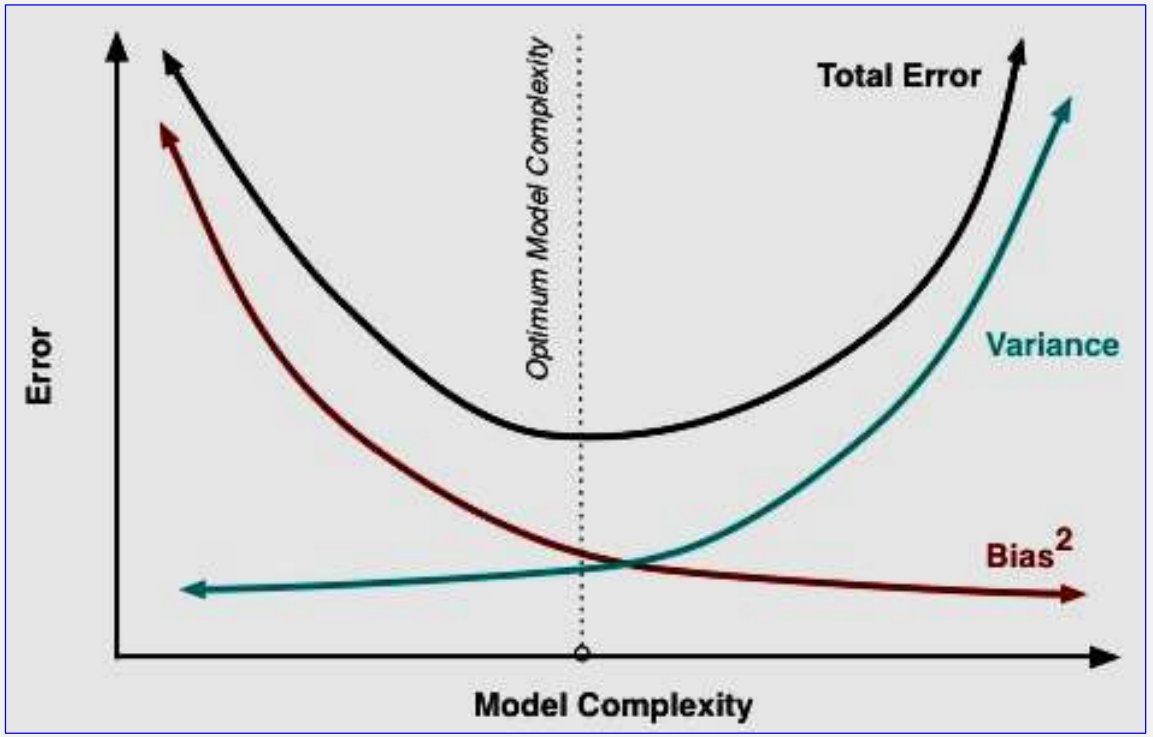
- Real life problems are mostly complicated, hence machine learning models are simplified to fit in given data.
- This simplification of problem introduces error in model.
- This error is referred as **BIAS**.



linear regression results in **HIGH BIAS**

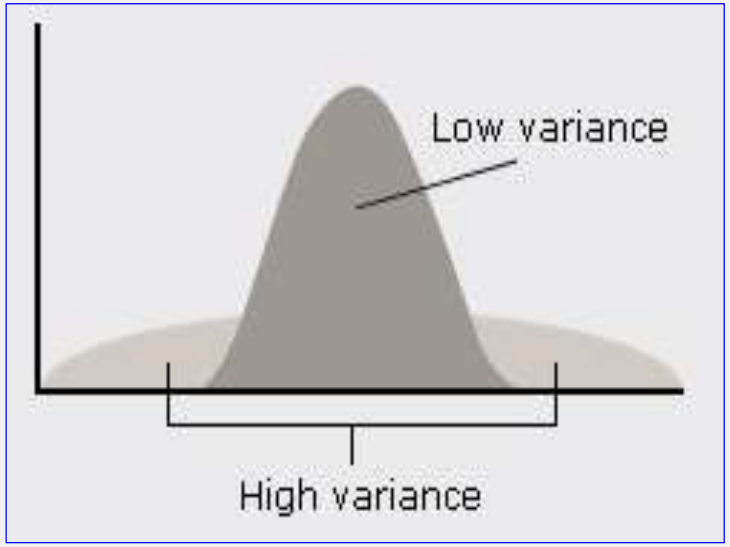


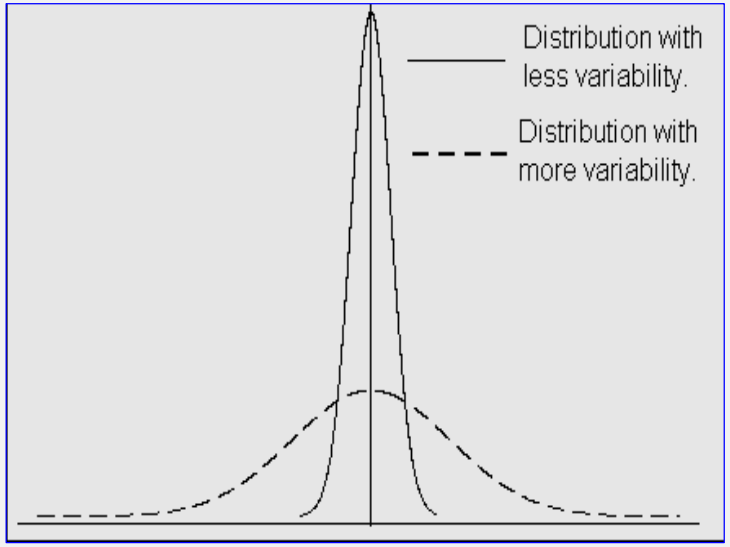
linear regression results in **LOW BIAS**



VARIANCE

- Variance measures **volatility in data**.
- Variance refers to the amount by which **estimate of function** would change if we estimated it using a different training data set. Since the training data are used to fit the statistical learning method, different training data sets will result in a different function. But ideally the estimate for function should not vary too much between training sets. However, if a method has high variance then **small changes** in the training data can result in large changes in estimate of function.
- In general, more flexible statistical methods have higher variance.

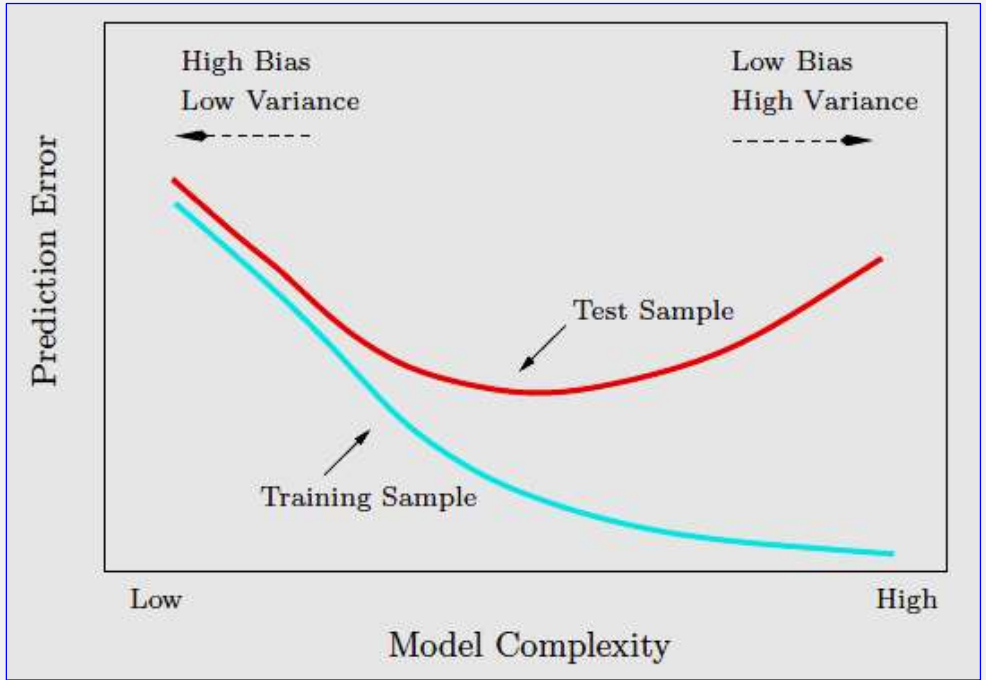




BIAS-VARIANCE TRADE OFF

- Real life problems are mostly complicated, hence **simplified.**
- The simplification of problem introduces **error.**
- This error is referred as **bias.**
- Use more **flexible** methods, the **variance** will increase and the **bias** will decrease.
- Use more flexible methods, the variance will **increase** and the bias will **decrease.**

Error	Overfitting	Underfitting
Model Complexity	HIGH	LOW
Bias	LOW	HIGH
Variance	HIGH	LOW



QUESTION AND ANSWERS