

# **Data Preparation, Training Objectives, Performance Evaluation**

# OUTLINE

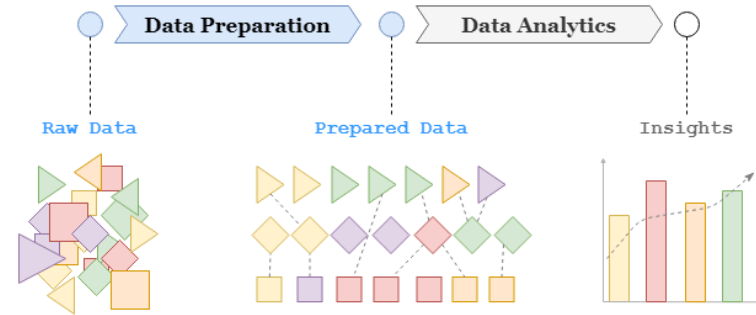
- ❑ Data Preparation
- ❑ Training Objectives
- ❑ Performance Evaluation

# **Data Preparation**

# Data Preparation

## Definition:

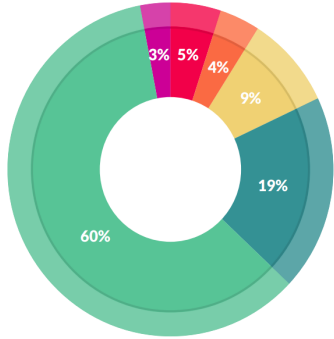
- ❑ Transformation of raw data into an easy-to-understand format.
- ❑ Find relevant data to include in the analysis process so that it can produce information or insights for analysts or business users.
- ❑ Pre-processing steps involve cleaning, transforming, and consolidating data.
- ❑ A process that involves connecting to one or many different data sources, cleaning dirty data, reformatting or restructuring the data, and finally combining this data to use for analysis.



## Other Terms:

- ❑ Data Pre-processing,
- ❑ Data Manipulation,
- ❑ Data Cleansing/ Normalization

# Facts about Data Preparation



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

60-80% of data scientist activities  
(forbes, crowdflower 2016)

- Existing data from many data sources and various formats (structured, semi-structured and unstructured).
- Predictive model quality depends on data quality (Garbage in – Garbage Out).

## Data Preparation Matters

**65%** of organizations said it is **very important to simplify making information available**. The most often required big data preparation activities are:



In the analytic process, the tasks in which organizations spend the most time are reviewing data for quality and consistency (**52%**) and preparing data for analysis (**46%**).

Source:

- <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=6e9aa0e36f63>
- <https://improvado.io/blog/what-is-data-preparation>

# The Importance of Data Preparation

- Data needs to be formatted according to the software used.
- Data needs to be adapted to the data science method used.
- Real-world data tends to be dirty:
  - incomplete: lack of attribute values, lack of certain/important attributes, only contains aggregate data. for example: job = "" (no entry)
  - noisy: has errors or outliers. e.g.: Salary="-10", Age="222"
  - inconsistent: have differences in code and name. e.g: Age = "32" Date of Birth = "03/07/2000"; ratings "1,2,3" -> ratings "A, B, C"
- Alternating columns and rows.
- Multiple variables in the same column.



# Benefits and Challenges of Data Preparation

## **Benefits**

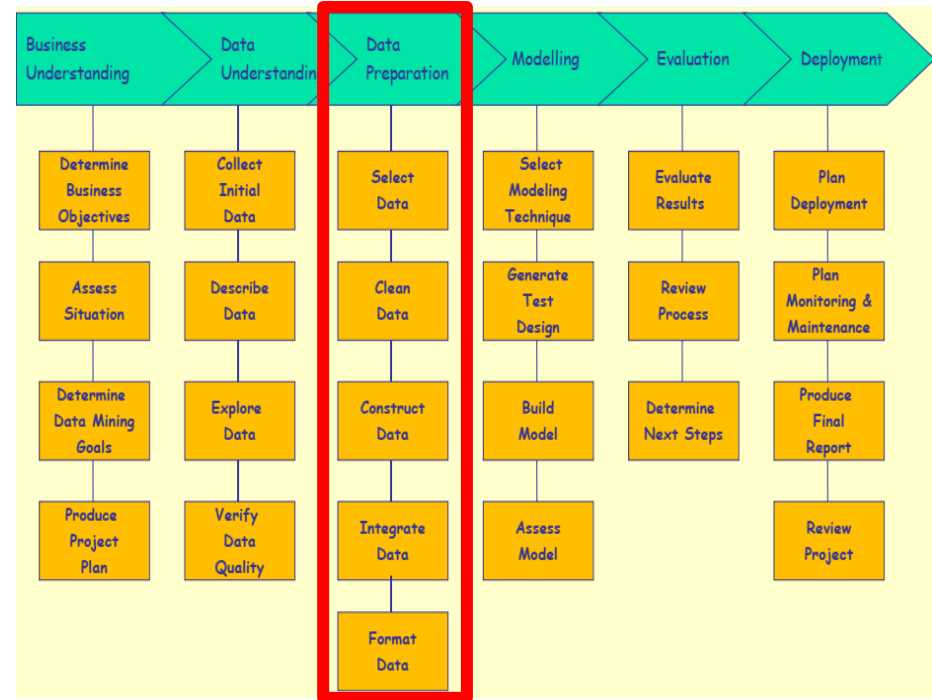
- Data compilation to be efficient and effective (avoid duplication)
- Identify and fix errors
- Easy change globally
- Produce accurate information for decision making
- Business Value and ROI (Return on Investment) will Increase

## **Challenges**

- Take a long time
- Dominant technical portion
- Available data is inaccurate or clear or not directly useable
- Unbalanced data when sampling
- Vulnerable to error

# Data Preparation in CRISP-DM

- ❑ Acronym for: **C**Ross **I**ndustry **S**tandard **P**rocess **D**ata **M**ining
- ❑ A common methodology for data mining, analytics, and data science projects to standardize data mining processes across industries.
- ❑ Used for all levels from beginner to expert.





# Data Preparation Stages

## 1. Select Data

- Determine the dataset to be used.
- Consider data selection.
- Consider using a sampling technique.
- Explain why certain data was included or excluded.
- Collect appropriate additional data (internal or external).

## 2. Clean Data

- Fix, remove or ignore the noise.
- Decide how to deal with special values and their meaning.
- Aggregation level, missing values, etc.
- Clean or manipulate outliers.

## 3. Validate Data

- Check/assess data quality
- Check/value data sufficiency level

## 4. Construct Data

- Derived attribute.
- Background knowledge.
- How the missing attributes can be constructed or calculated.

## 5. Integrate Data

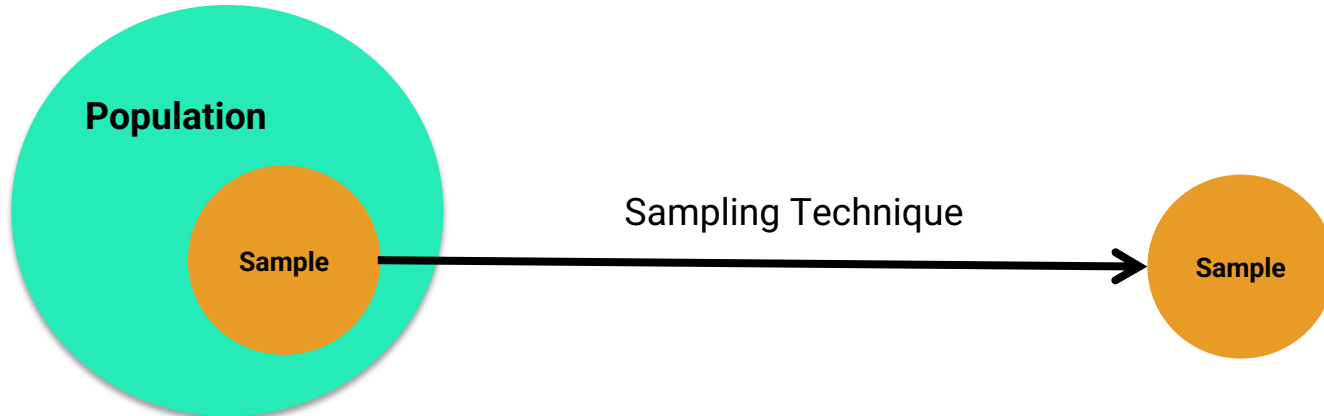
- Integrate sources and store results (new tables and records).

## 6. Format Data

# Data Sampling

Before carrying out the stages in data preparation, first is the selection/determination of objects which can be done by determining:

- **Population** is a group of individuals or subjects in an area and time with certain qualities to be observed/researched.
- **Sample** is part of the population that is used as a research subject as a "representative" of the members of the population.



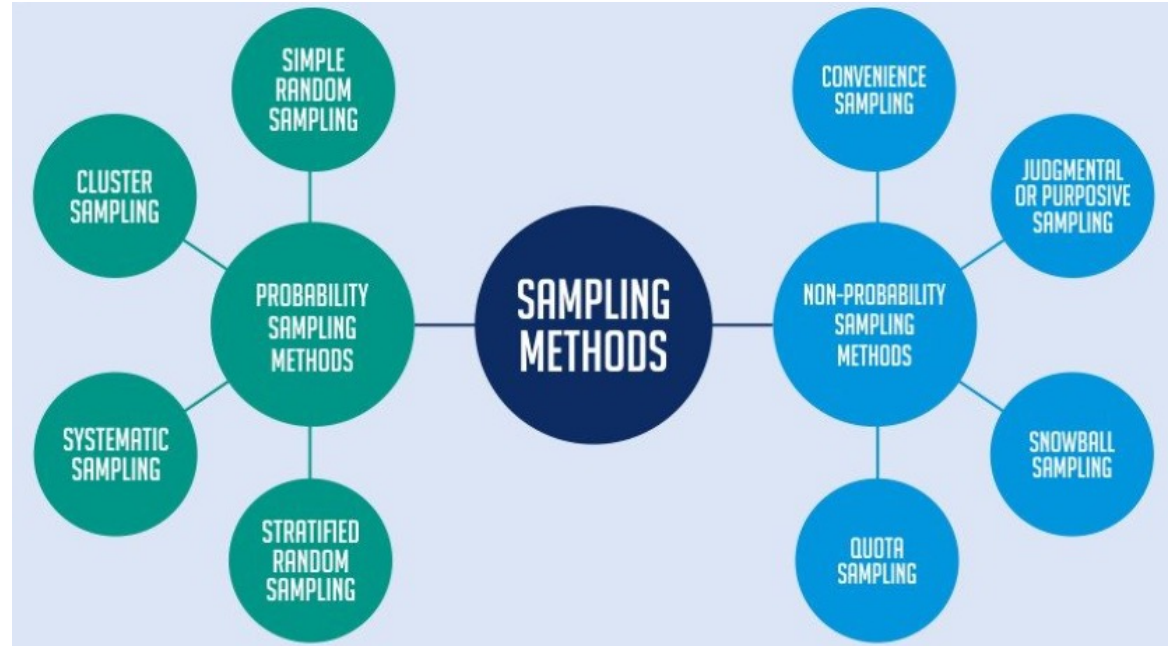
# Sampling Methods

- **Probability Sampling**

- Population known
- Randomization/  
randomness: Yes
- Conclusive
- Result: Unbiased
- Conclusion: Analytics

- **Non-Probability Sampling**

- Population unknown
- Research limitations
- Randomization/  
randomness: No
- Exploratory
- Result: Biased
- Conclusion: Analytics



# Sampling Methods

## Types of probability sampling

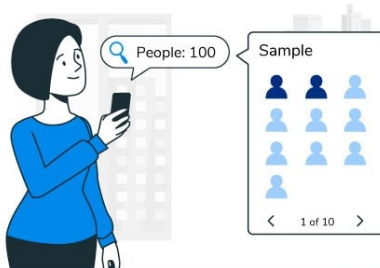
Simple random sampling



Cluster sampling



Systematic sampling



Stratified random sampling



## Types of non-probability sampling

Convenience sampling



Consecutive sampling



Judgmental or Purposive sampling



Quota sampling



Snowball sampling



# Sampling Stages

Step 1

Identify and define Target Population



Step 2

Select Sampling Frame



Step 3

Choose Sampling methods



Step 4

Determine Sample Size

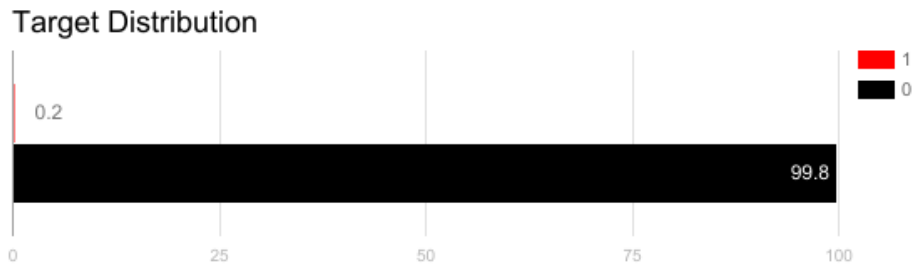


Step 5

Collect the required Data

# Imbalance Dataset

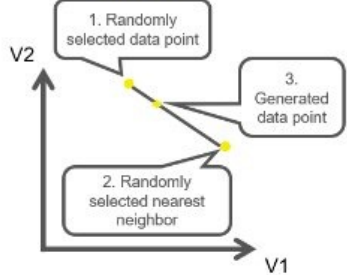
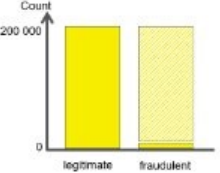
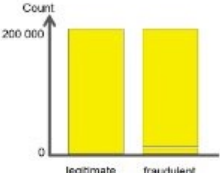
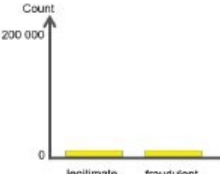
- An **imbalanced dataset** is a condition where the distribution of data classes is unbalanced, and the number of data classes (instances) is one less or more than the number of other data classes.
- In an imbalanced dataset, we usually have classes with few data (rare class) and classes with lots of data (abundant class).
- Examples of cases where dataset imbalance often occurs are credit scoring, fraud, disease data, etc.
- One way to overcome an imbalanced dataset is to do **resampling**.



# Resampling

Resample training data can be done using two methods:

- ❑ **Undersampling:** balancing the dataset by reducing the amount of data in major classes. Performed if the quantity of data is sufficient.
- ❑ **Oversampling:** balancing the dataset by increasing the amount of data in minor classes. Performed if the quantity of data is insufficient.

Resampling method	Description	Target class distribution after resampling
Oversampling (SMOTE)	<p>Generate new synthetic fraudulent transactions until the number of fraudulent transactions is ca. equal to the number of legitimate transactions:</p> <ol style="list-style-type: none"> <li>1. Select one of the fraudulent transactions in the training data randomly</li> <li>2. Select one of its <math>n</math> nearest neighbors in the same fraudulent class randomly</li> <li>3. Select a random point between the existing fraudulent transaction and its nearest neighbor</li> </ol> 	<ul style="list-style-type: none"> <li>• Original data in yellow</li> <li>• New synthetic data in light patterned yellow</li> </ul> 
Oversampling (Bootstrap)	Randomly draw with replacement a sample of fraudulent transactions until the number of fraudulent transactions is ca equal to the number of legitimate transactions	
Undersampling (Bootstrap)	Randomly draw with replacement as many legitimate transactions as there are fraudulent transactions	

# Feature Selection

- After determining the sampling of the data to be taken, then do the **feature selection** for the sampling data.
- Feature selection is a core concept in ML that has a major impact on predictive model performance. Data features that are not/partially relevant and have a negative impact on model performance should be removed.
- The definition of **Feature Selection** is the automatic or manual process of selecting data features that best contribute to the desired predictive or output variable.
- Two types of feature selection:
  - **Unsupervised**: methods that ignore the target variable, such as removing redundant variables.
  - **Supervised**: methods that use the target variable, such as removing irrelevant variables.



# Benefits of Feature Selection

- **Overfitting reduction:** the smaller the redundant data, the fewer noise-based decisions are made.
- **Improve Accuracy:** the smaller the misleading data, the better the accuracy of the model.
- **Reduction of Training Time:** the smaller the data points, the less complex the algorithm is and the faster the algorithm trains.

# Clean Data

Data cleaning is done for **messy data**, such as:

- data with duplicate or irrelevant values
- data with missing values
- data with inconsistent format
- malformed record
- data that has too many outliers

## Data cleaning steps:

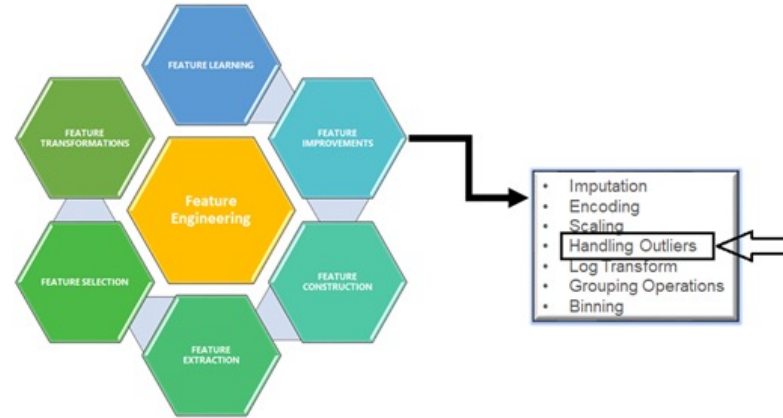
- **Removing unwanted observations:** delete duplicate/redundant/irrelevant values or noise.
- **Missing data handling:** fixing the issue of unknown missing values.
- **Structural error solving:** fixing problems with mislabeled classes, types in names of features, the same attribute with different names, etc. Decide how to deal with special values and their meaning.
- **Outlier management:** unwanted values which are not fit in datasets.

# Validate Data

- **Validation** or **validity** is measuring the difference in scores that reflects the true difference between individuals, groups, or situations regarding the characteristics to be measured, or also actual errors in individuals or the same group from one situation to another.
- **Validation** is a critical step that is often ignored by data scientists.
- **Validation** is done by checking several things including:
  - Data Type (e.g.. integer, float, string)
  - Range Data
  - Uniqueness (e.g. postal code)
  - Consistent expression (e.g. Street, St.)
  - Data Format (e.g. for date “YYYY-MM-DD” VS “DD-MM-YYYY.”)
  - Null/Missing Values
  - Misspelling/Type
  - Invalid Data (gender: M/F: M; Male; F: Female? )
- **Data and model validation techniques:**
  - Accuracy of existing data
  - Data completeness
  - Data consistency
  - Punctuality
  - Trust
  - Value-added
  - Interpretation
  - Ease of Access

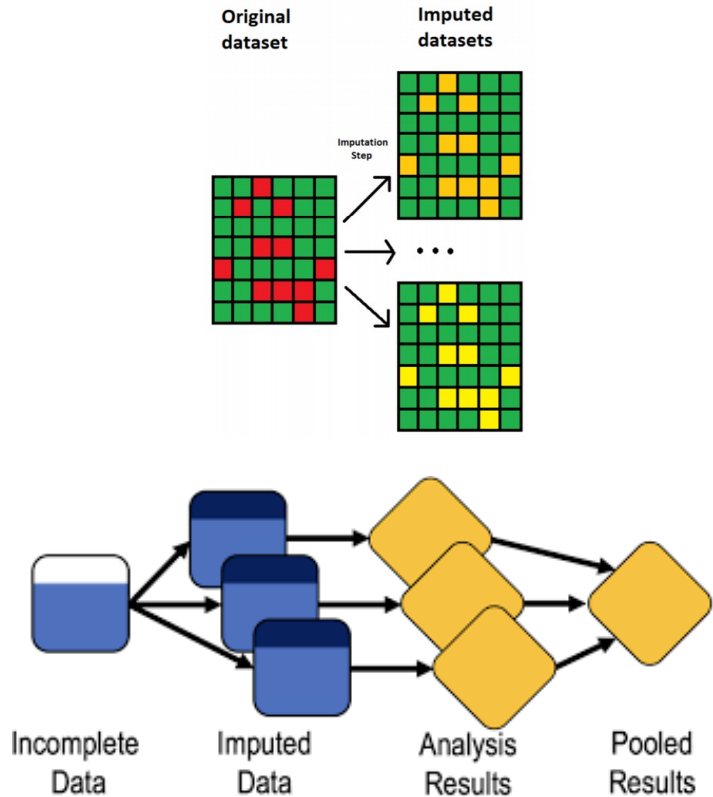
# Data Transformation

- **Feature Representation or Feature Learning:**
  - Techniques that allow the system to work automatically to find the required representation (for feature detection or classification of datasets),
  - Replacing manual feature engineering,
  - Allows machines to learn features and use them to perform specific tasks.
- **Feature Engineering** is the process of turning raw data into features that:
  - Representing the fundamental problem of predictive models,
  - Resulting in better model accuracy on unseen data.



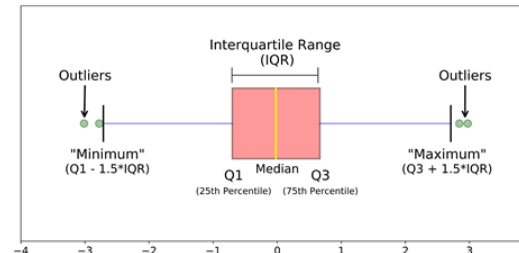
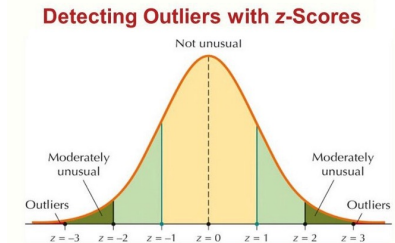
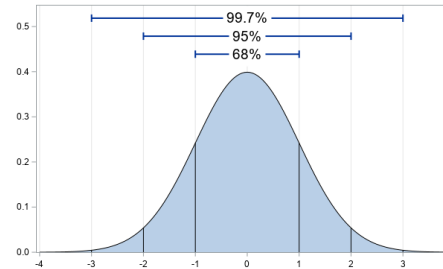
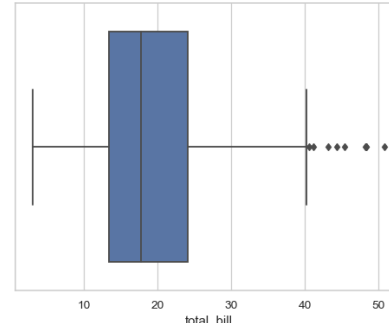
# Imputation

- **Definition:** Replacing missing value/data (missing value; NaN; blank) with a replacement value.
- If the data type is a **Numeric Variables**
  - Imputed mean or median.
  - Imputation of arbitrary values.
  - End of tail value/data imputation.
- If the data type is a **Categorical Variables**
  - Imputation of frequently occurring categories.
  - Add missing categories.
- There is **no perfect way to handle missing values** in a data set. Each strategy performs better for certain data sets and missing data types but can perform much worse on other types of data sets.



# Handling Outliers

- Definition of **Outliers**:
  - Data points that are very different from other data.
  - Observations that deviate from the overall pattern in the sample.
- **Reason**: experimental errors, input errors, instrument errors, intentional (for testing), data processing errors, sampling errors, reasonableness due to anomalies in the data (not errors).
- Outlier detection:
  - **Visualization with Boxplot and Scatterplot**. Most of the data points are located in the middle, but there is one point that is far from the other observations; these could be outliers.
  - **Normal Distribution**  
In a normal distribution, about 99.7% of the data is within three standard deviations of the mean. If any observation is more than three times the standard deviation, it is likely an outlier.
  - **Z-scores**
  - **Inter Quartile Range (IQR)**



Source:  
<https://heartbeat.fritz.ai/hands-on-with-feature-engineering-techniques-dealing-with-outliers-fcc9f57cb63b>  
<https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/>

# Techniques for Handling Outliers

- **Trimming:** remove the outliers from the dataset.
- **Winsorizing:** replace the outliers from the dataset with the percentile values of each upper and lower end/boundary.
- **Imputing:** replace the outliers from the dataset with the mean, median, or any arbitrary values.
- **Discretization/binning:** the process of changing functions, models and continuous variables into discrete (continuous data measured vs. continuous data counted).
- **Censoring**
- **Z-score**
- **Linear Regression Model**

# Feature Scaling

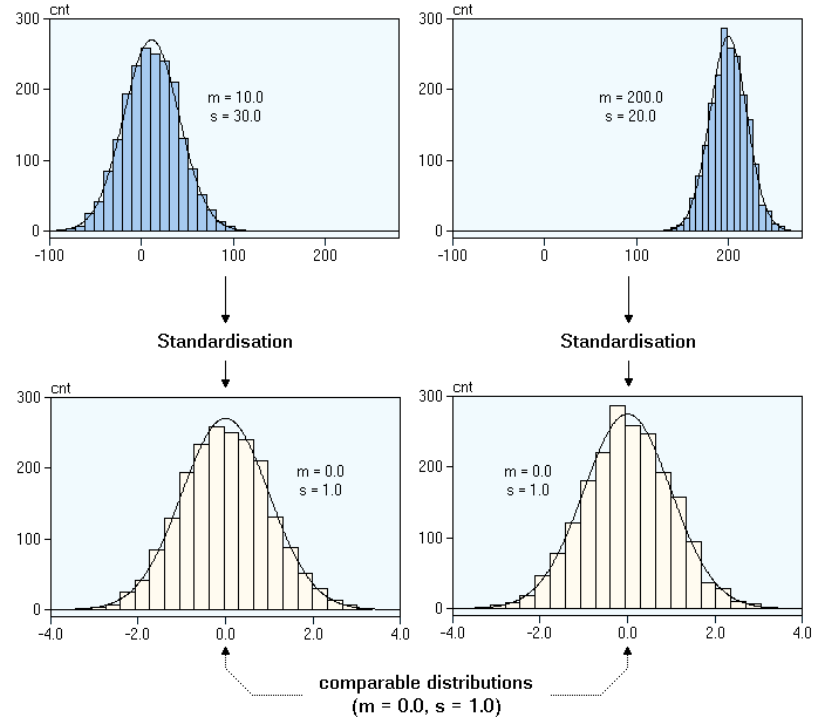
- Numerical data (usually) have different ranges and characteristics. Therefore, to compare different data attributes, it is necessary to do **feature scaling**.
- **Feature Scaling** is a way to make numerical data in datasets have the same range of values (scale). Thus, there is no longer one data variable that dominates other data variables.

Standardisation	Normalisation
$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation}(x)}$	$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$



# Feature Scaling: Standardization

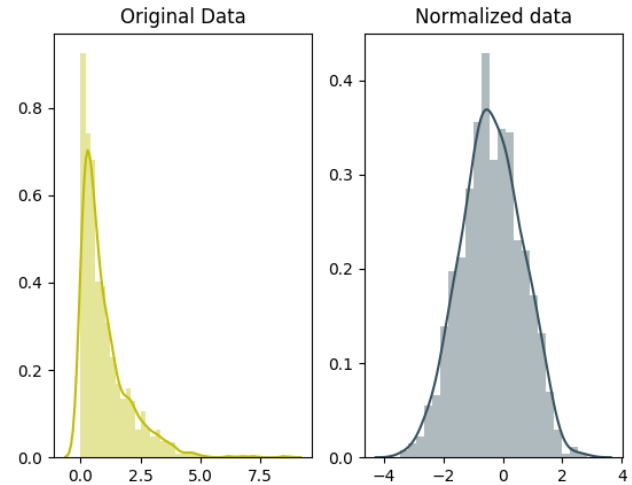
- **Goal:** focuses on turning raw data into usable information prior to analysis.
- **Definition:** A technique that scales data so that it has to mean = 0 and standard deviation = 1



# Feature Scaling: Normalization (Min-Max Scaling)

- **Definition:** A scaling technique in which values are shifted and scaled so that values range between 0 and 1 (range  $[0,1]$  ).
- $X_{max}$  and  $X_{min}$  are the maximum and minimum values of the feature, respectively.
  - When the value  $X$  is the minimum value in a column, the numerator is 0, and therefore  $X'$  is 0.
  - Otherwise, when  $X$  is the maximum value in a column, the numerator is the same as the denominator, so  $X'$  is 1.
  - If the value of  $X$  is between the minimum and maximum values, then the value of  $X'$  is between 0 and 1.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$



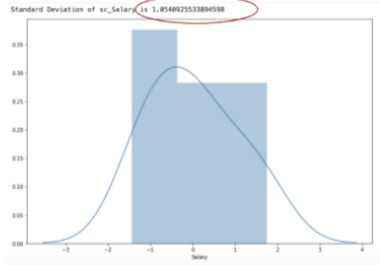
# Feature Scaling Example

*After Feature scaling.*

Column: Salary

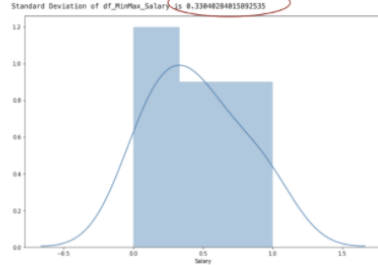
Standard Deviation (Salary):  
Max-Min Normalization (0.33) < Standardisation (1.05)

Standardisation



*Normal distribution and Standard Deviation of Salary.*

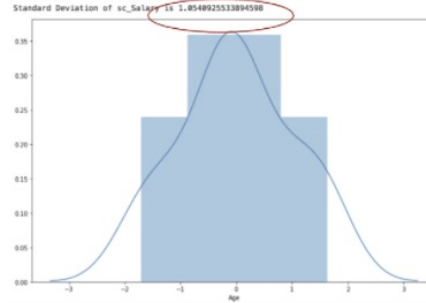
Max-Min Normalisation



Column: Age

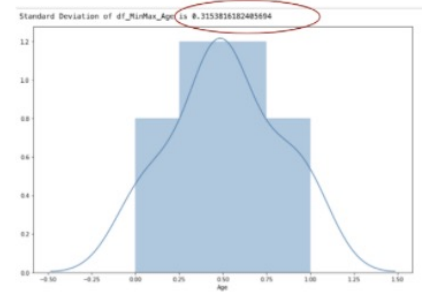
Standard Deviation (Age):  
Max-Min Normalization (0.315) < Standardisation (1.05)

Standardisation



*Normal distribution and Standard Deviation of Age.*

Max-Min Normalisation

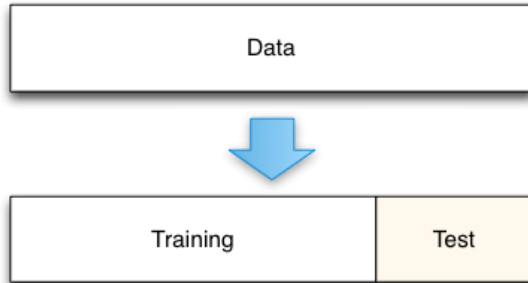


# **TRAINING OBJECTIVES**

# Model Selection and Assessment

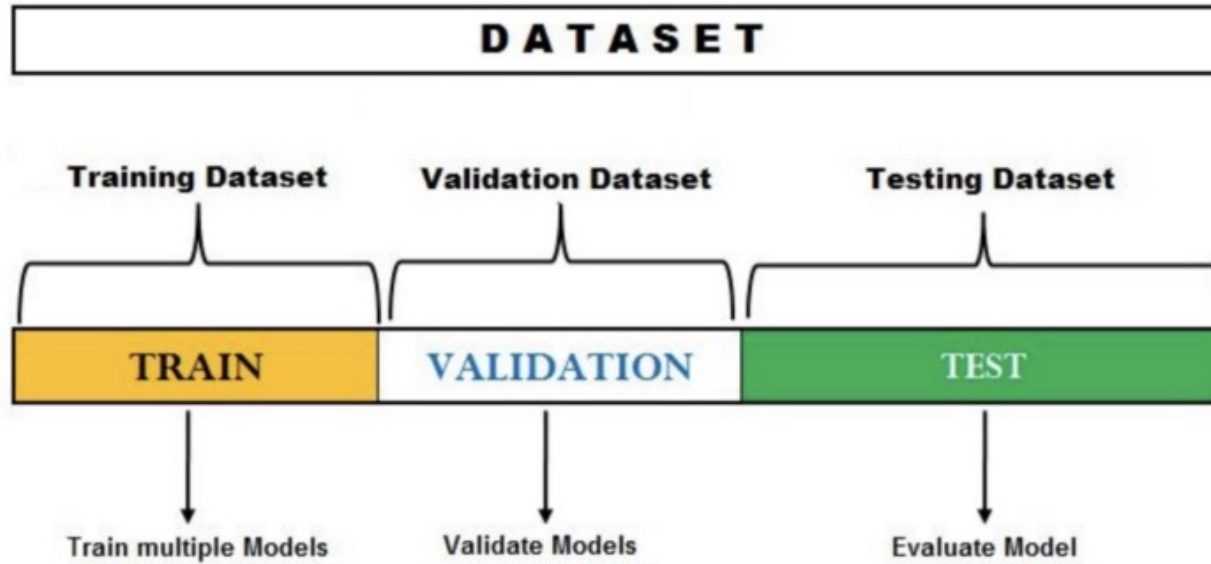
- ❑ **Model Selection:** Estimating performances of different models to choose the best one (produces the minimum of the test error)
- ❑ **Model Assessment:** Having chosen a model, estimate the prediction error on new data.

# Splitting the Data



X_train				y_train	Training Data 70%
Panjang Sepal	Lebar Sepal	Panjang Petal	Lebar Petal	Kelas	
5.1	3.5	1.4	0.2	Iris Setosa	
6.3	3.3	6	2.5	Iris Virginica	
7	3	4.6	1.4	Iris Versicolour	
...	...	...	...	...	
...	...	...	...	...	
...	...	...	...	...	
5.8	3.3	6	2.4	Iris Virginica	
6.8	3.1	4.5	1.5	Iris Versicolour	
X_test				y_test	Testing Data 30%
4.9	3	1.4	0.2	Iris Setosa	
...	...	...	...	...	
6.8	3.2	4.4	1.6	Iris Versicolour	

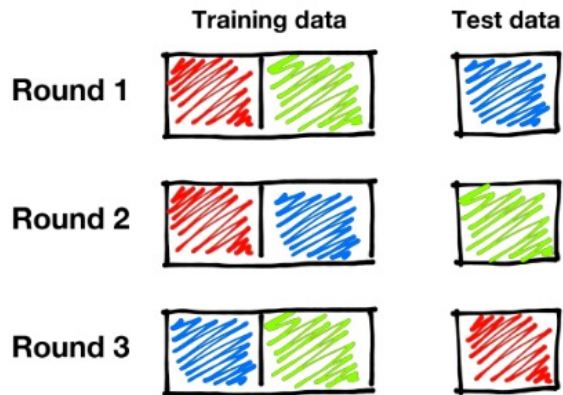
# Splitting the Data



**Figure: Data splitting**  
(source: <https://medium.com>)

# Cross Validation

Original data, divided into k parts



© Machine Learning @ Berkeley

**Figure: Cross Validation**

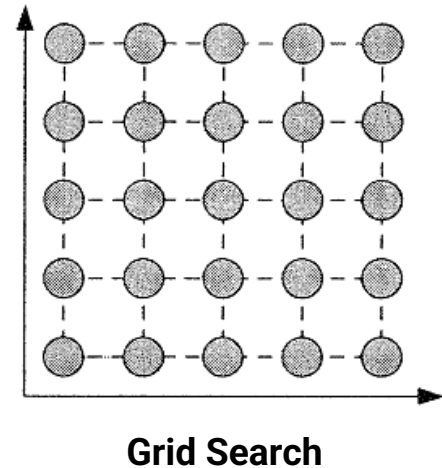
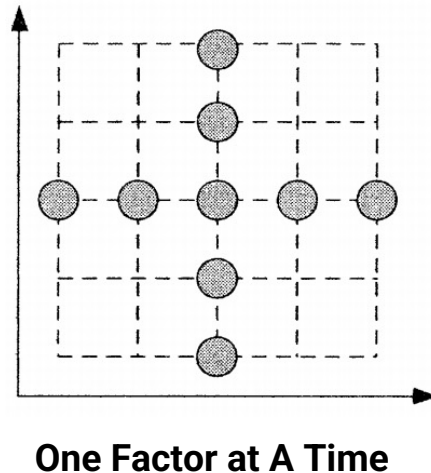
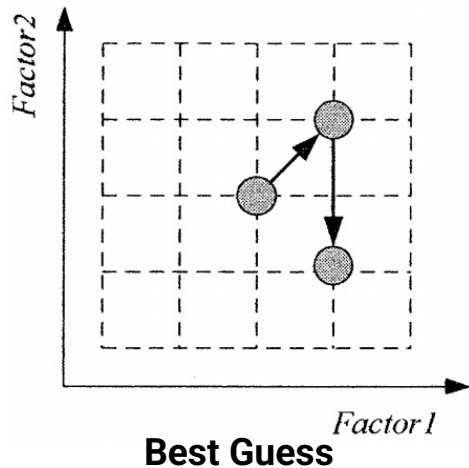
(source: <https://ml.berkeley.edu/blog/2017/07/13/tutorial-4/>)

- ❑ **Cross-validation (CV)** is one of the techniques used to **test the effectiveness of ML models**, it is also a re-sampling procedure used to evaluate a model if we have **limited data**.
- ❑ **Goal:** test the model's ability to predict new data that was not used in estimating it, in order to **handle problems** like overfitting and selection bias, and to give an insight into an independent dataset (i.e., **how the model will generalize** an unknown dataset, for instance from a real problem).



# Training Objectives

- Each machine learning model/method has certain parameters.
- Experiments were carried out with several variations of the parameters.
- Parameters that produce the best performance model will be used next.
- Several parameter search strategies to produce the best model.



# **PERFORMANCE EVALUATION FOR SUPERVISED LEARNING (CLASSIFICATION)**

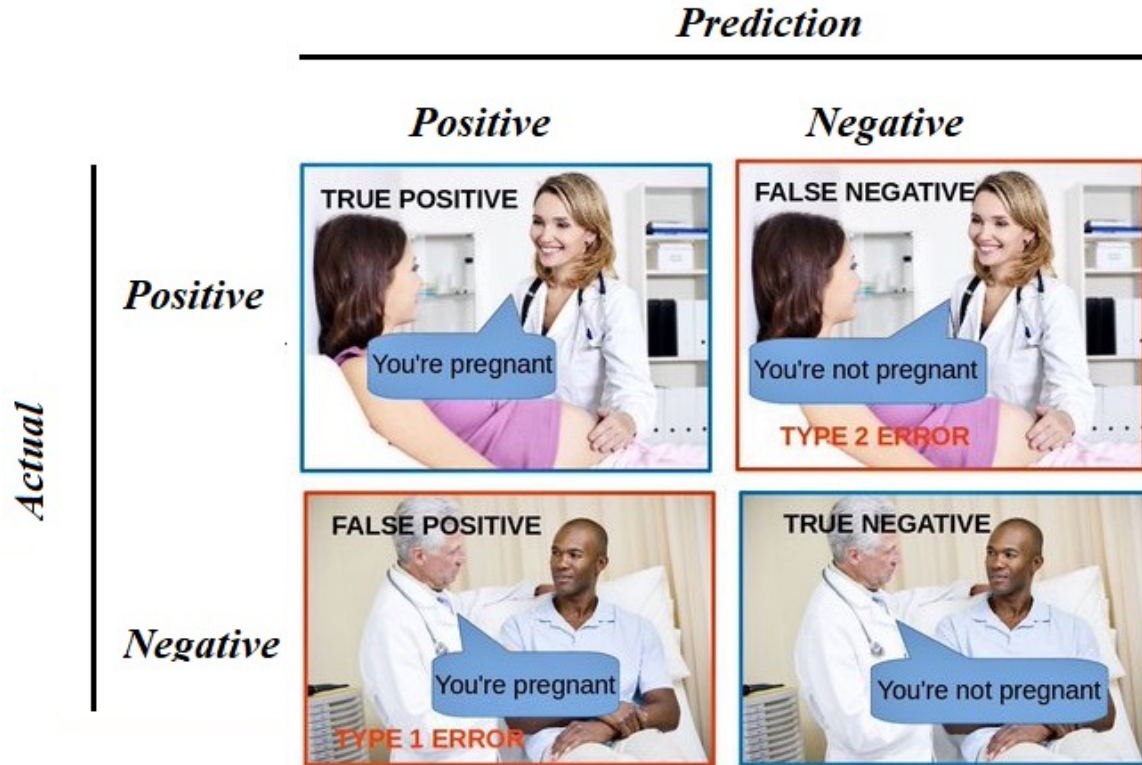
# Binary Classification

- In **binary classification**, there are only two classes: 0 or 1, valid or invalid, true or false, positive or negative, good or not good, beautiful or not beautiful, recommended or not, passed or not, spam or not spam, etc.
- The most common form: one class is declared as a positive class (became the focus of classification), and one other class is declared as a negative class.
- We can use a **confusion matrix** to evaluate binary classification.

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

- **True Positive (TP)**: the actual value is positive and is predicted to be positive
- **False Positive (FP)**: the actual value is negative but is predicted to be positive
- **True Negative (TN)**: the actual value is negative and is predicted to be negative
- **False Negative (FN)**: the actual value is positive but is predicted to be negative.

# Illustration of Confusion Matrix for Binary Classification



# False Positive vs False Negative

- Prediction errors are found in **False Positive** and **False Negative**.
- Each problem has a different parameter pressure point. Model evaluation is highly dependent on cases and data distribution.
- The FP and FN parameters are error parameters that need to be reduced in value so that the resulting errors are not fatal to humans. Parameters that are **more crucial** between the two will be studied based on the case.
- **Early detection of a tsunami** (tsunami: + )
  - FN: It is predicted that there was no tsunami, but in fact, there was a tsunami.
  - FP: It is predicted that there will be a tsunami, but it turns out that there is no tsunami.
  - In this case, an unpredictable tsunami (FN) is more dangerous and very detrimental than FP.
- **Spam detection used as a spam filter** (spam: +)
  - FN: Spam Emails enter the inbox
  - FP: A normal email goes to the spam folder
  - In this case, generally, spam in the inbox (FN) is not detrimental compared to important emails in the spam folder (FP) which can result in failed work/projects/schools, etc.

# Evaluation Metrics for Binary Classification

		Nilai Prediksi		
		Positive	Negative	
Nilai Aktual	Positive	True Positive (TP)	False Negative (FN)	Recall, Sensitivity, True Positive Rate $\frac{TP}{TP + FN}$
	Negative	False Positive (FP)	True Negative (TN)	Specificity, True Negative Rate $\frac{TN}{FP + TN}$ False Positive Rate $\frac{FP}{FP + TN}$
		Precision $\frac{TP}{TP + FP}$	Negative Predictive Value $\frac{TN}{TN + FN}$	Accuracy $\frac{TP + TN}{TP + TN + FP + FN}$

**Precision:** % of selected items that are correct

**Recall:** % of correct items that are selected

## A Combined Measure: F-Score

- A combined measure that assesses the Precision-Recall tradeoff is the **F measure** (weighted harmonic mean):

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- If precision and recall are equally important, people usually use a balanced F1 score (i.e., with  $\beta = 1$ , that is,  $\alpha = \frac{1}{2}$ ):

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

## More Than Two Classes: Sets of Binary Classifiers

- Dealing with **any-of** or **multivalue** classification
  - A document can belong to 0, 1, or >1 classes.
- For each class  $c \in C$ 
  - Build a classifier  $\gamma_c$  to distinguish  $c$  from all other classes  $c' \in C$
- Given test doc  $d$ ,
  - Evaluate it for membership in each class using each  $\gamma_c$
  - $d$  belongs to **any** class for which  $\gamma_c$  returns true



## More Than Two Classes: Sets of Binary Classifiers

- One-of or multinomial classification
  - Classes are mutually exclusive: each document in exactly one class
- For each class  $c \in C$ 
  - Build a classifier  $\gamma_c$  to distinguish  $c$  from all other classes  $c' \in C$
- Given test doc  $d$ ,
  - Evaluate it for membership in each class using each  $\gamma_c$
  - $d$  belongs to the one class with maximum score

## Micro- vs. Macro-Averaging

If we have more than one class, how do we combine multiple performance measures into one quantity?

- ❑ **Macroaveraging:** Compute performance for each class, then average.
- ❑ **Microaveraging:** Collect decisions for all classes, compute contingency table, evaluate.

Class 1

	Truth: yes	Truth: no
Classifier: yes	10	10
Classifier: no	10	970

Class 2

	Truth: yes	Truth: no
Classifier: yes	90	10
Classifier: no	10	890

Micro Average Table

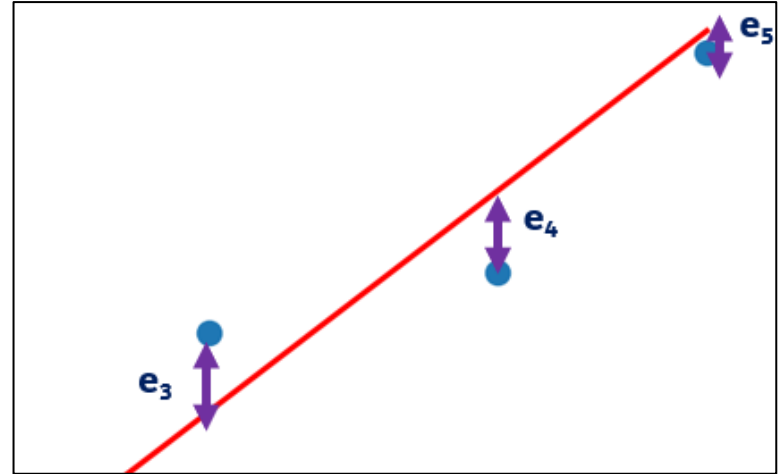
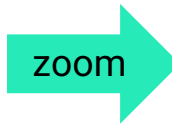
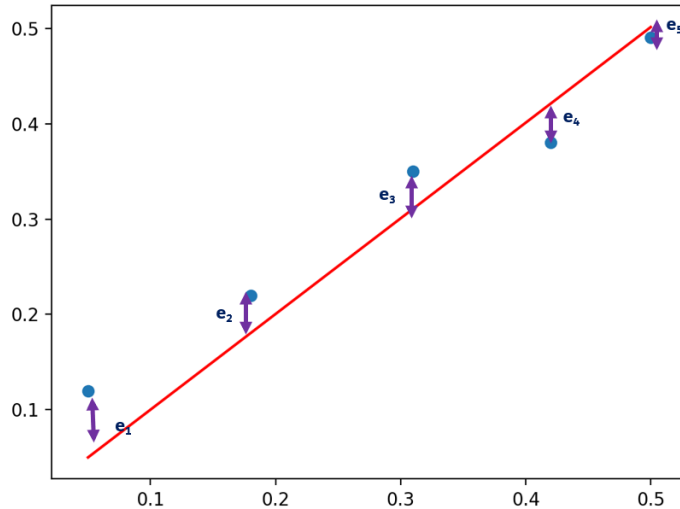
	Truth: yes	Truth: no
Classifier: yes	100	20
Classifier: no	20	1860

- Macroaveraged precision:  $(0.5 + 0.9)/2 = 0.7$
- Microaveraged precision:  $100/120 = 0.83$
- Microaveraged score is dominated by score on common classes

# **PERFORMANCE EVALUATION FOR SUPERVISED LEARNING (REGRESSION)**

# Performance Evaluation for Regression

- The model predicts a **continuous value** (a real number), not a discrete value (in the form of a class/label).
- Example: house price prediction, maximum temperature, earthquake strength, the stock price.
- **Error** is the difference between the actual value and the predicted (real) value.



# Evaluation Metrics for Regression

- ❑ Mean Absolute Error (MAE)
- ❑ Relative Absolute Error (RAE)
- ❑ Mean Squared Error (MSE)
- ❑ Relative Squared Error (RSE)
- ❑ Root Mean Squared Error (RMSE)
- ❑ Mean Absolute Percentage Error (MAPE)
- ❑ Mean Percentage Error (MPE)
- ❑ R-squared

Acroynm	Full Name	Residual Operation?	Robust To Outliers?
MAE	Mean Absolute Error	Absolute Value	Yes
MSE	Mean Squared Error	Square	No
RMSE	Root Mean Squared Error	Square	No
MAPE	Mean Absolute Percentage Error	Absolute Value	Yes
MPE	Mean Percentage Error	N/A	Yes

# Mean Squared Error (MSE)

$$MSE = \frac{1}{n} \sum \left( \underbrace{y - \hat{y}}_{\substack{\text{The square of the difference} \\ \text{between actual and} \\ \text{predicted}}} \right)^2$$

# Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Predicted_i - Actual_i)^2}{N}}$$

- It is preferred more in some cases because the errors are first squared before averaging which poses a high penalty on large errors.
- This implies that RMSE is useful when large errors are undesired.

# Mean Absolute Error (MAE)

The diagram illustrates the Mean Absolute Error (MAE) formula with the following components and annotations:

- Divide by the total number of data points:** A blue line points to the  $\frac{1}{n}$  term, which is enclosed in a blue box.
- Sum of:** A blue line points to the summation symbol  $\Sigma$ .
- Actual output value:** A green line points to the  $y$  term inside a green box.
- Predicted output value:** An orange line points to the  $\hat{y}$  term inside an orange box.
- The absolute value of the residual:** A bracket under the  $|y - \hat{y}|$  term is labeled with this text.

$$MAE = \frac{1}{n} \sum \left| y - \hat{y} \right|$$

- The MAE is more robust to outliers and does not penalize the errors as extremely as MSE.
- MAE is a linear score which means all the individual differences are weighted equally.
- It is not suitable for applications where you want to pay more attention to the outliers.



# R-Squared

$$R^2 = 1 - \frac{\text{MSE}(\text{model})}{\text{MSE}(\text{baseline})}$$

- The coefficient of Determination or  $R^2$  is another metric used for evaluating the performance of a regression model.
- The metric helps us to compare our current model with a constant baseline and tells us how our model is better.
- The constant baseline is chosen by taking the mean of the data and drawing a line at the mean.
- $R^2$  is a scale-free score that implies it doesn't matter whether the values are too large or too small, the  $R^2$  will always be less than or equal to 1.

# THANK YOU

