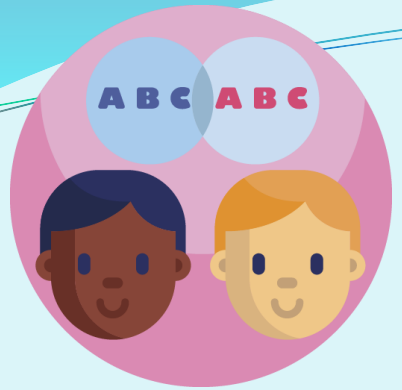


SUPERVISED LEARNING

ONLINE STUDENT TRAINING

FOR

"ARTIFICIAL INTELLIGENCE & MACHINE LEARNING"

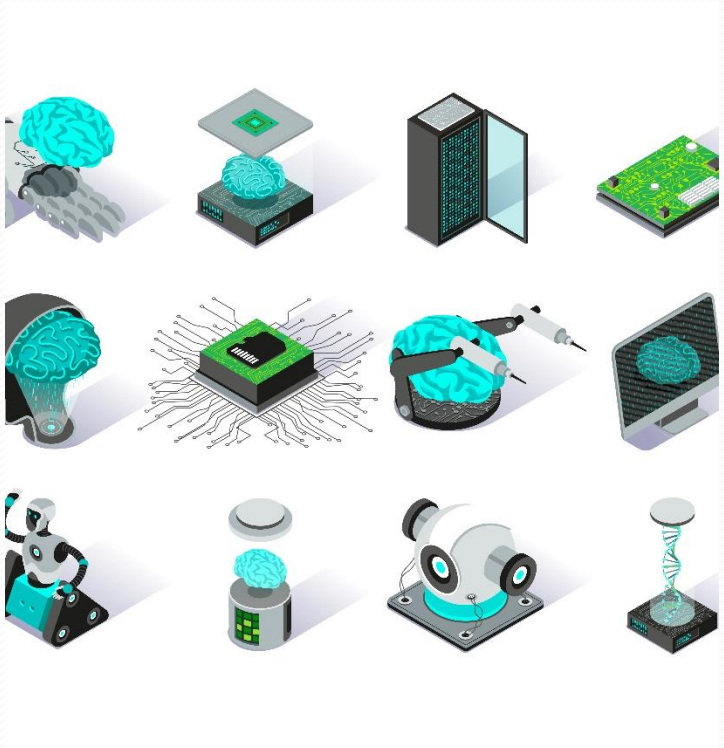
(4TH FEB, 2021 – 17TH MAR, 2021)

FACULTY TRAINER

NAW VARSHA PIPADA

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

ENGINEERING COLLEGE BIKANER



Contents

Supervised Learning

- Supervised machine learning algorithms are designed to learn by example. The name “supervised” learning originates from the idea that training this type of algorithm is like having a teacher supervise the whole process.
- When training a supervised learning algorithm, the training data will consist of inputs paired with the correct outputs.
- During training, the algorithm will search for patterns in the data that correlate with the desired outputs.

Categories

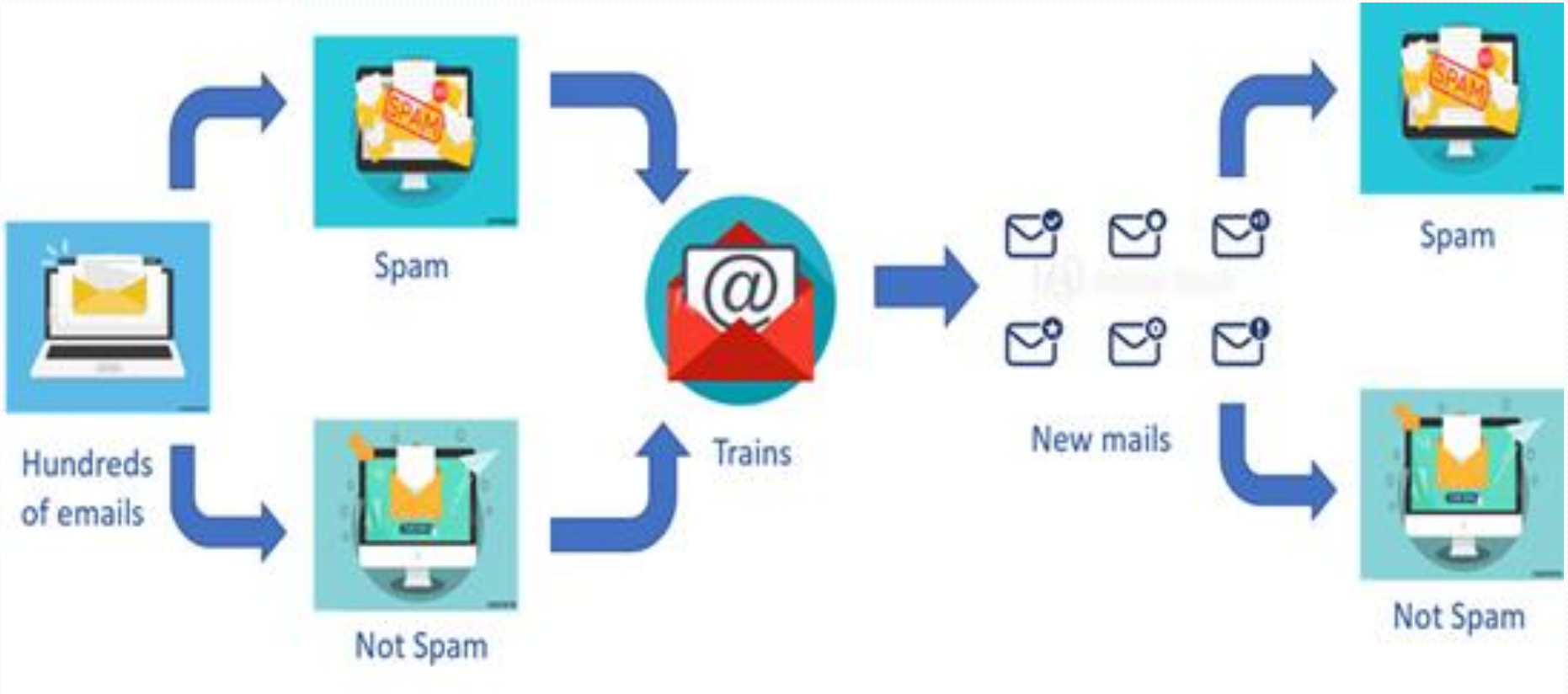
Classification

- It is used for problems where the output variable can be categorized, such as “Yes” or “No”, or “Pass” or “Fail.”
- Classification Models are used to predict the category of the data.
- Examples - spam detection, sentiment analysis, scorecard prediction of exams, etc.

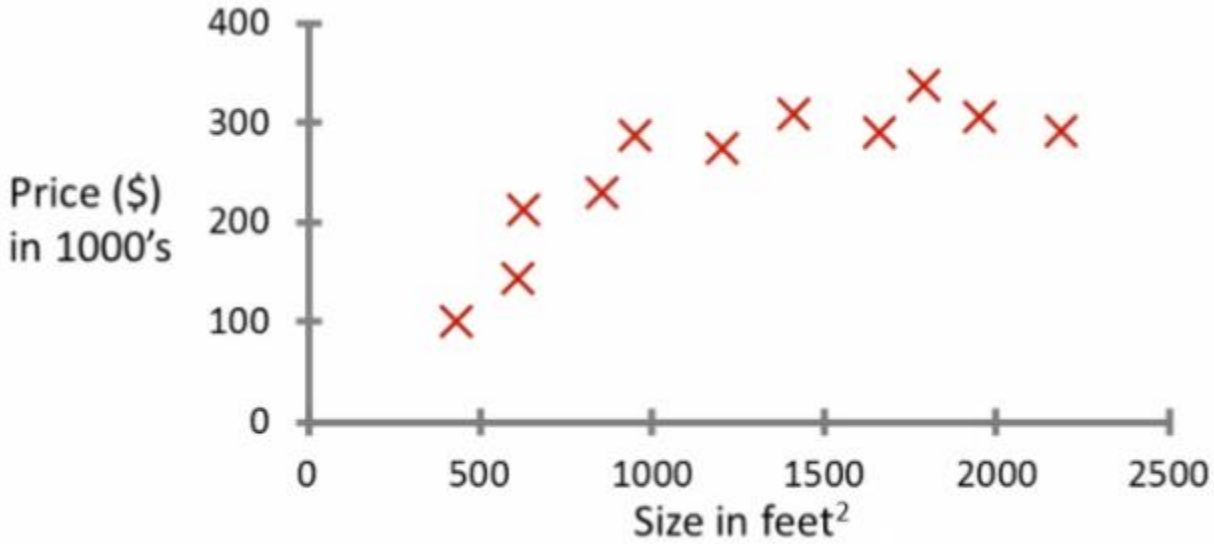
Regression

- Regression models are used for problems where the output variable is a real value such as a unique number, dollars, salary, weight or pressure, for example
- Some of the more familiar regression algorithms include linear regression, logistic regression, polynomial regression, and ridge regression

Classification



Regression


Price ($)
in 1000's
400
300
200
100
0
0
500
1000
1500
2000
2500
Size in feet²

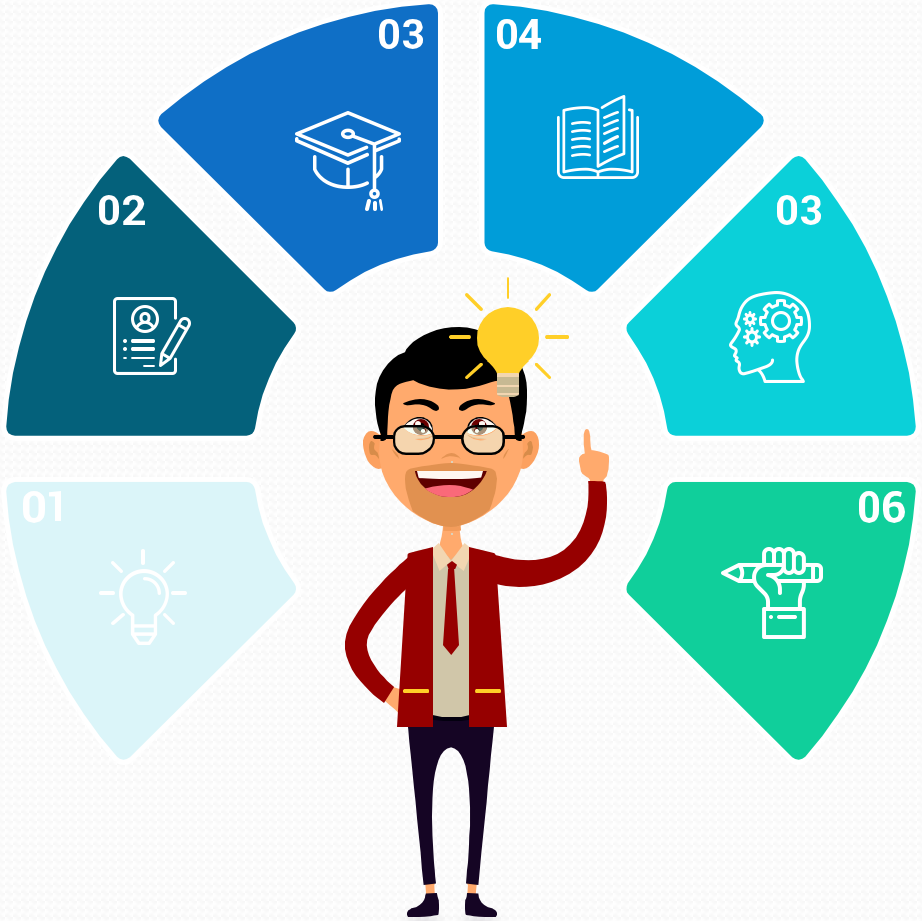
Machine Learning Process

Some Basic Concepts

Features & Target

Training Data

Testing Data



Bias-Variance Tradeoff

Model Evaluation Techniques

Model Evaluation Metrics

Terminology : Features & Its Types

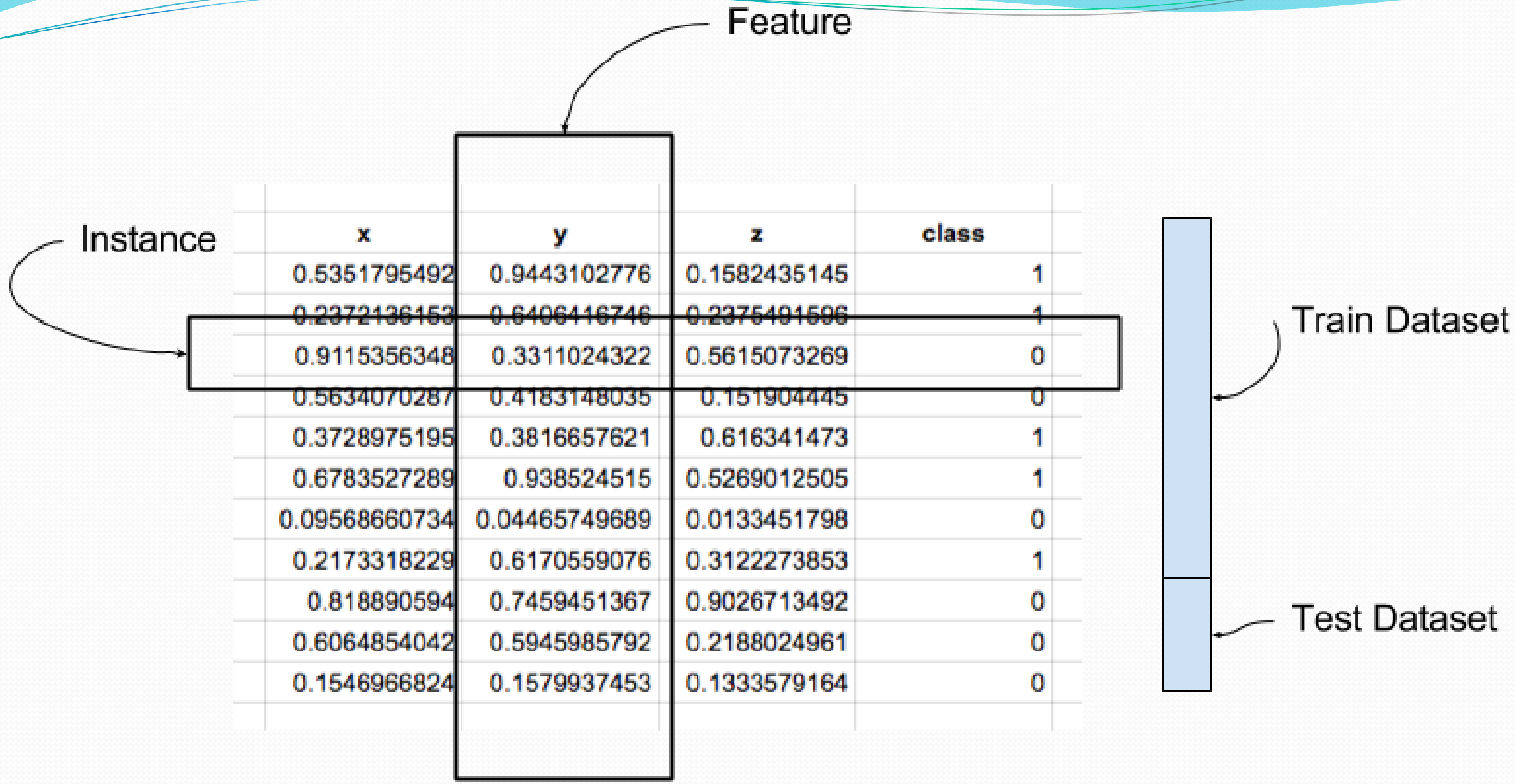
- Features are individual independent variables that act like a input to the system.
- Features are individual observations when analyzed into a set of quantifiable properties. It can be interpreted as column in dataset.
- Features are building blocks of datasets as the quality of the features in your dataset has major impact on the quality of the insights you will get while using the dataset for machine learning.
- Types
  - Categorical (e.g., Blood Group)
  - Ordinal (e.g., Shirt Size)
  - Integer (e.g., No. of words in text)
  - Real Values (e.g., Height)

Terminology : Training Data

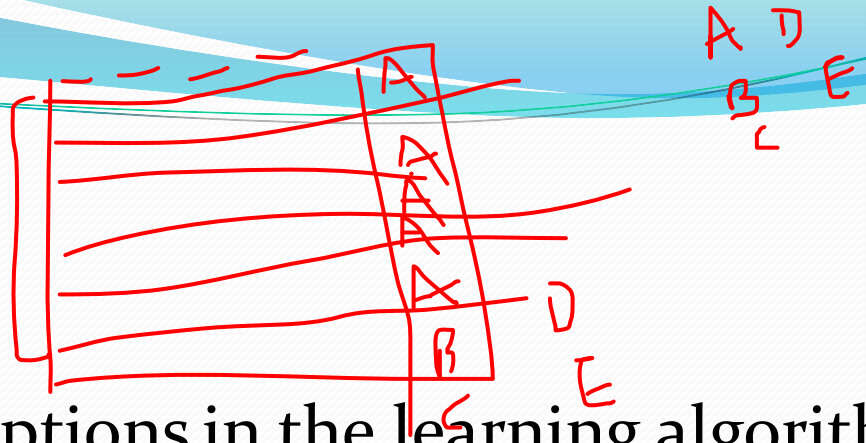
- Target
  - The **Target** is the information the machine learns to predict. In mathematical formulas, the target is usually called y or y_i for a single instance.
- Training Data
  - The observations in the training set form the experience that the algorithm uses to learn.
  - In supervised learning problems, each observation consists of an observed output variable and one or more observed input variables.

Terminology : Test Data

- Test Data
  - The test set is a set of observations used to evaluate the performance of the model using some performance metric.
  - It is important that no observations from the training set are included in the test set.
  - If the test set does contain examples from the training set, it will be difficult to assess whether the algorithm has learned to generalize from the training set or has simply memorized it.


Feature
Instance
x
y
z
class
0.5351795492
0.9443102776
0.1582435145
1
0.2372136153
0.6406416746
0.2375491596
1
0.9115356348
0.3311024322
0.5615073269
0
0.5634070287
0.4183148035
0.151904445
0
0.3728975195
0.3816657621
0.616341473
1
0.6783527289
0.938524515
0.5269012505
1
0.09568660734
0.04465749689
0.0133451798
0
0.2173318229
0.6170559076
0.3122273853
1
0.818890594
0.7459451367
0.9026713492
0
0.6064854042
0.5945985792
0.2188024961
0
0.1546966824
0.1579937453
0.1333579164
0
Train Dataset
Test Dataset

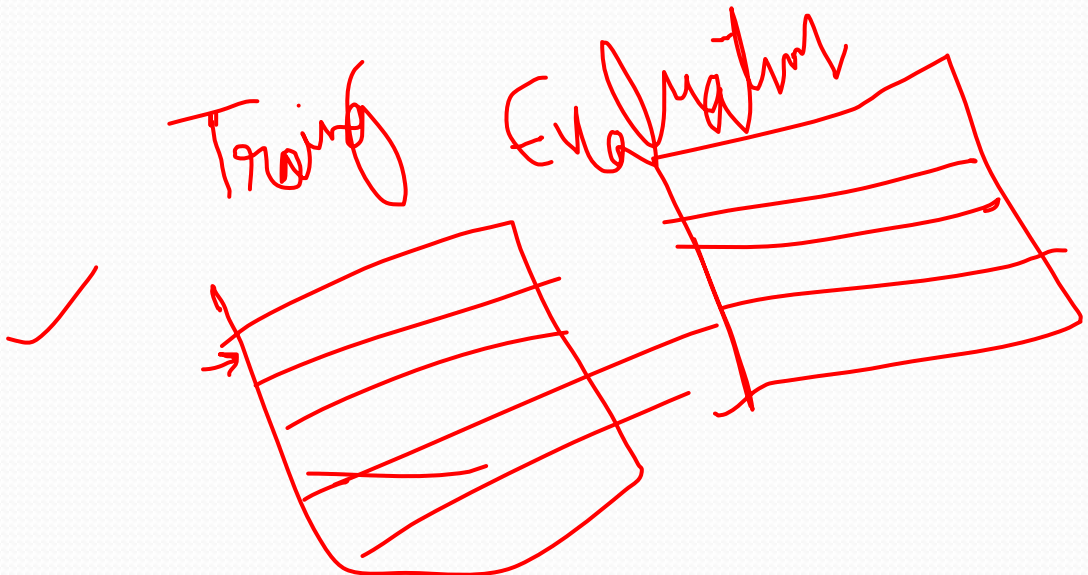
Terminology: Bias



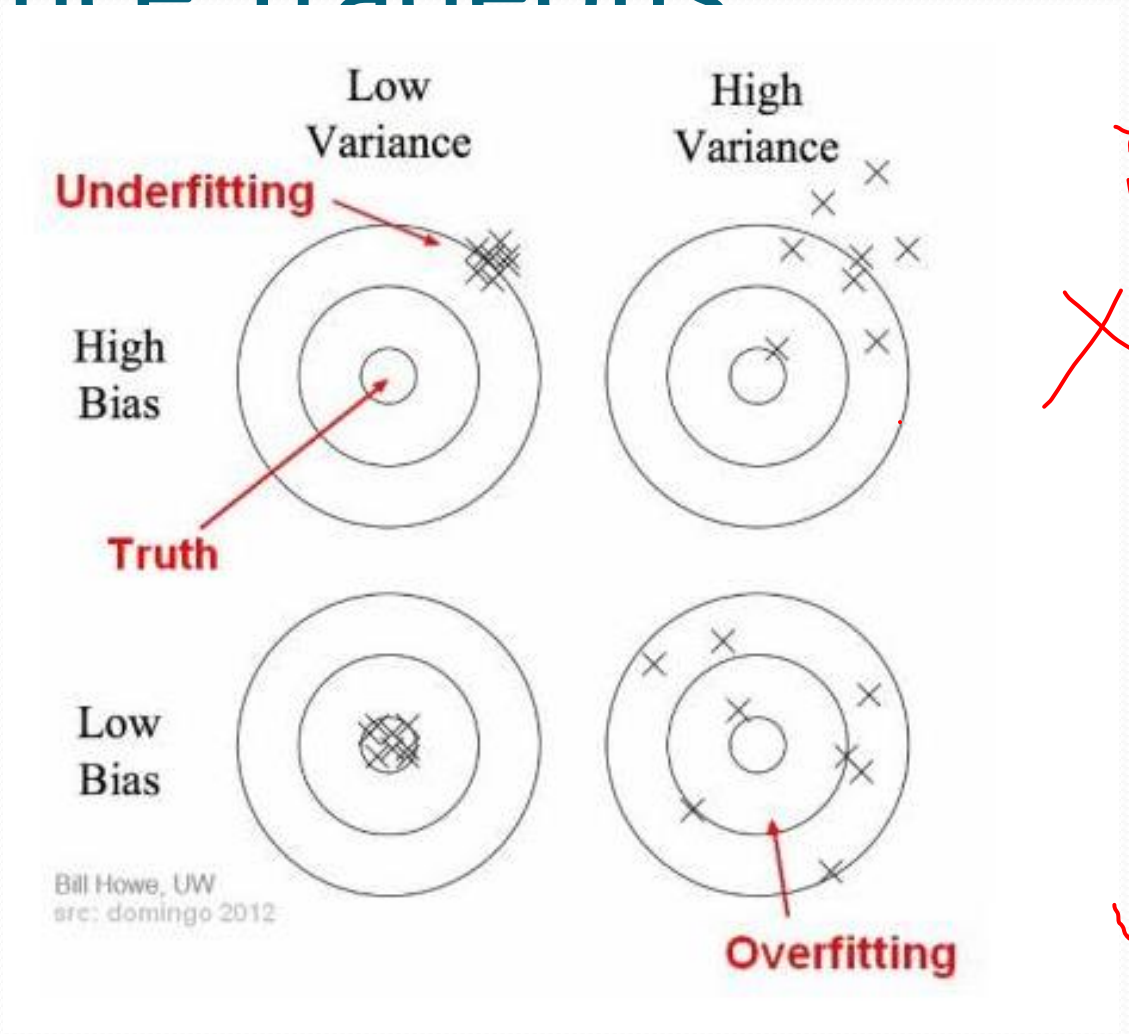
- Bias is an error from erroneous assumptions in the learning algorithm.
- High bias can cause an algorithm to miss the relevant relations between features and target outputs
- Bias is the algorithm's tendency to consistently learn the wrong thing by not taking into account all the information in the data
- Example
  - If you have a simple model, you might conclude that every "Alex" are amazing people. This presents a High Bias and Low Variance problem. Your dataset is 'biased' towards people with the name Alex. Thus, most predictions will be similar, since you believe people with 'Alex' act a certain way.

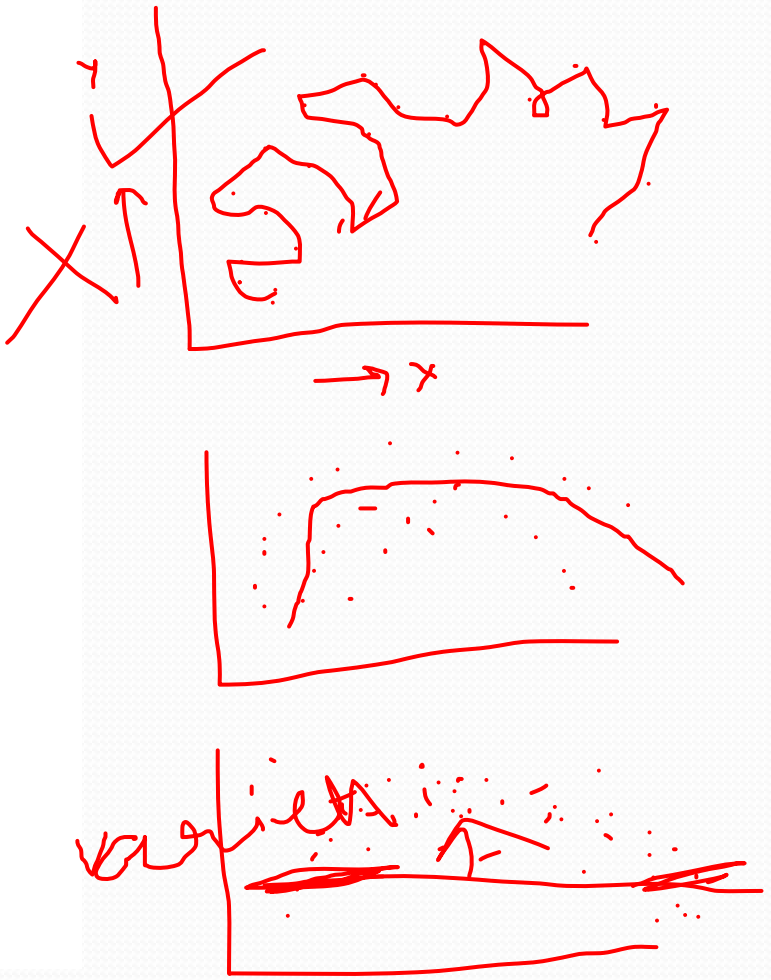
Terminology: Variance

- Variance is the variability of model prediction for a given data point or a value which tells us spread of our data.
- Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before.
- As a result, such models perform very well on training data but has high error rates on test data.



Bias-Variance Tradeoffs





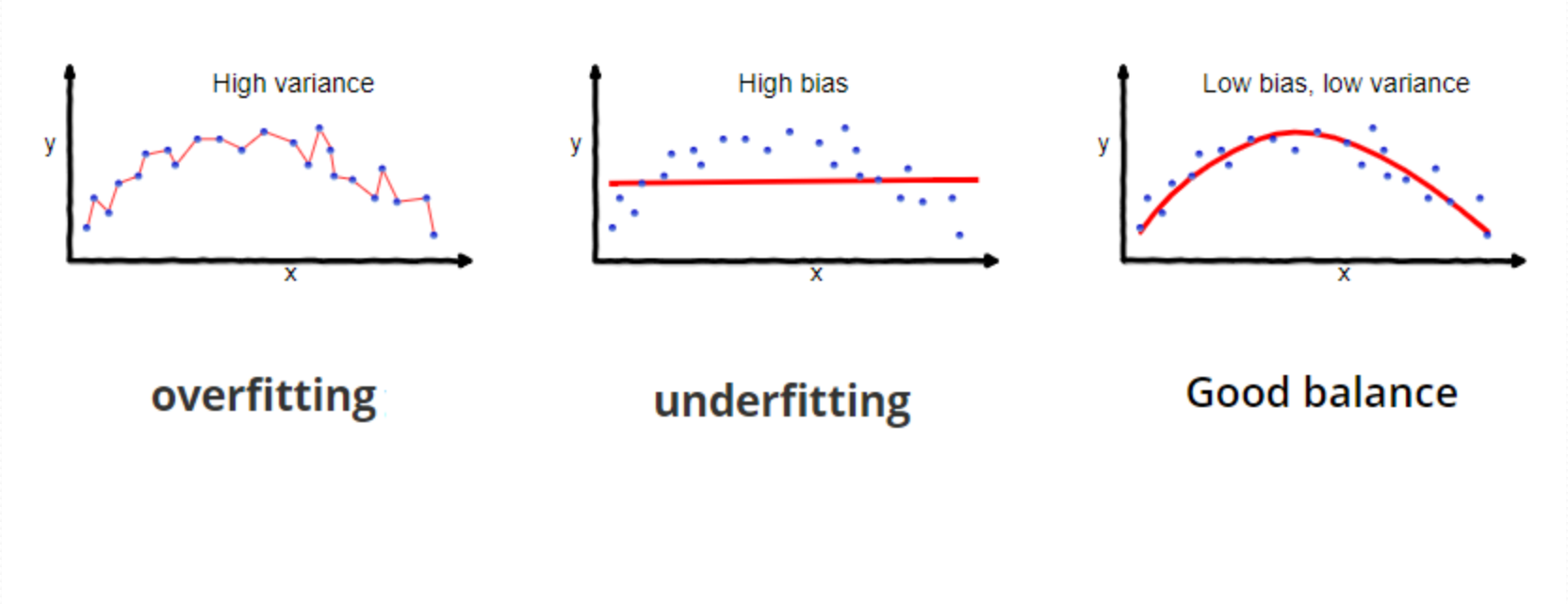
Overfitting

- Overfitting refers to a model that models the training data too well.
- Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data.
- This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model.
- The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize.

Underfitting

- Underfitting refers to a model that can neither model the training data nor generalize to new data.
- An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data.
- Underfitting is often not discussed as it is easy to detect given a good performance metric.

Overfitting Vs. Underfitting



Model Evaluation

- How the model generalizes on unseen data ?
- Whether it actually works and, consequently, if we can trust its predictions.
- Could the model be merely memorizing the data it is fed with, and therefore unable to make good predictions on future samples, or samples that it hasn't seen before?
- Model evaluation aims to estimate the generalization accuracy of a model on future (unseen/out-of-sample) data.

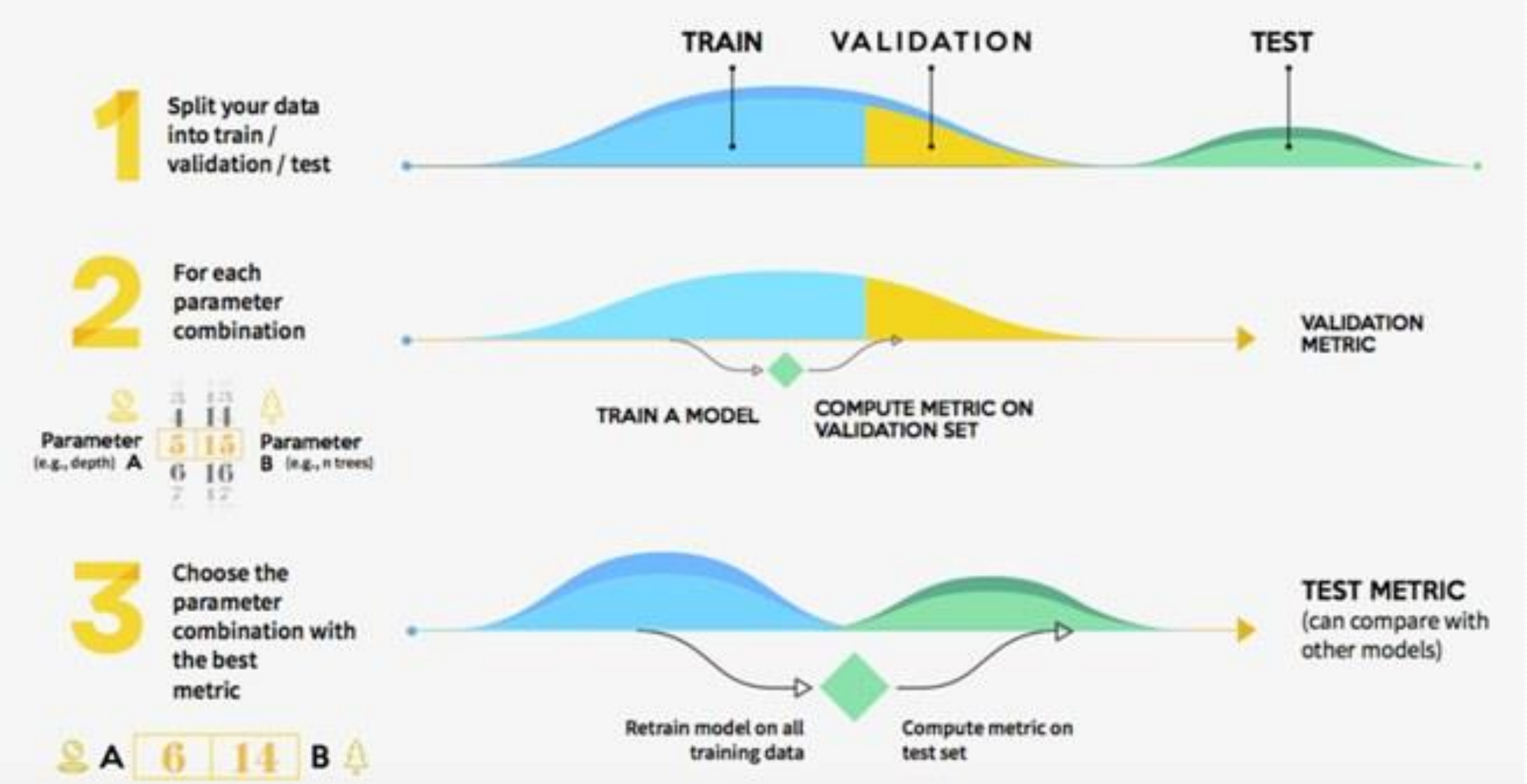
Model Evaluation Techniques

- Most common methods
  - Holdout
  - Cross-validation
- Both methods use a test set (i.e data not seen by the model) to evaluate model performance.
- It's not recommended to use the data we used to build the model to evaluate it.
- This is because our model will simply remember the whole training set, and will therefore always predict the correct label for any point in the training set. That means overfitting.

Model Evaluation Techniques : Holdout

- The purpose is to test a model on different data than it was trained on and hence to get an unbiased estimate of learning performance.
- In this method, the dataset is *randomly* divided into three subsets:
  - **Training set** is a subset of the dataset used to build predictive models.
  - **Validation set** is a subset of the dataset used to assess the performance of the model built in the training phase. It provides a test platform for fine-tuning a model's parameters and selecting the best performing model. Not all modeling algorithms need a validation set.
  - **Test set**, or unseen data, is a subset of the dataset used to assess the likely future performance of a model. If a model fits to the training set much better than it fits the test set, overfitting is probably the cause.

Hold-Out Method



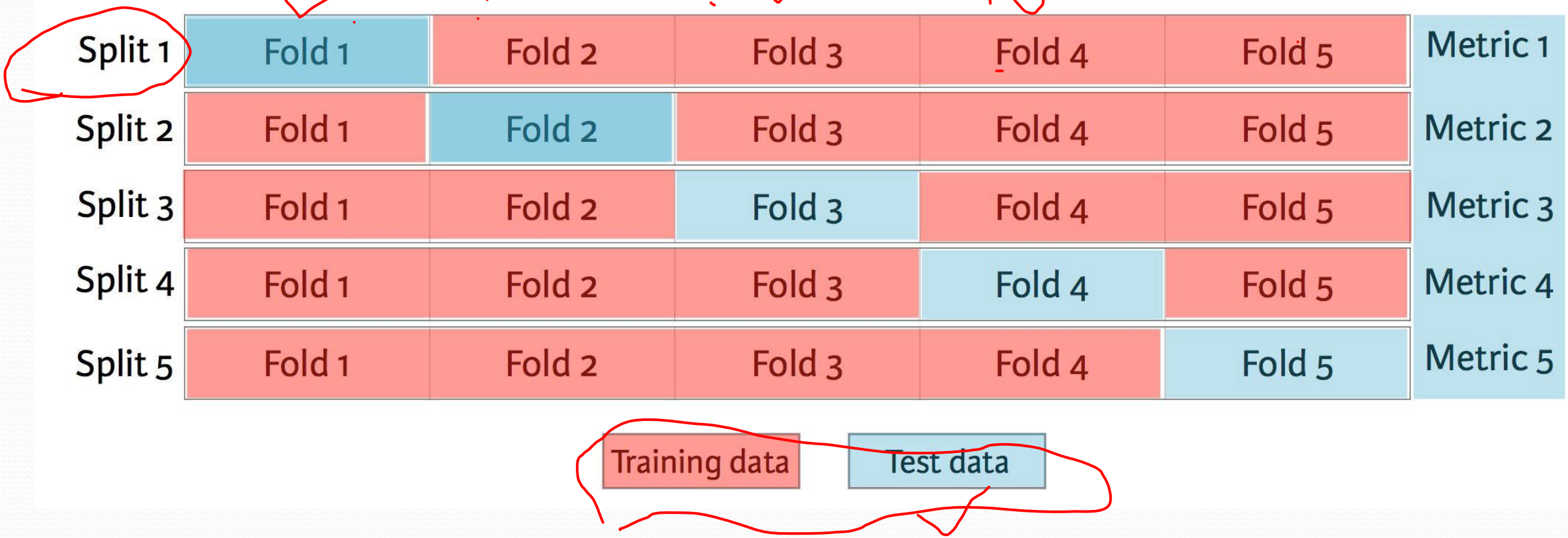
Model Evaluation Techniques : Cross-Validation

- The most common cross-validation technique is **k-fold cross-validation**, where the original dataset is partitioned into k equal size subsamples, called folds.
- The k is a user-specified number, usually with 5 or 10 as its preferred value.
- This is repeated k times, such that each time, one of the k subsets is used as the test set/validation set and the other k-1 subsets are put together to form a training set.
- The error estimation is averaged over all k trials to get the total effectiveness of our model.

Model Evaluation Techniques : Cross-Validation

- As can be seen, every data point gets to be in a test set exactly once and gets to be in a training set k-1 times.
- This significantly reduces bias, as we're using most of the data for fitting, and it also significantly reduces variance, as most of the data is also being used in the test set.
- Interchanging the training and test sets also adds to the effectiveness of this method.
- Cross-validation techniques can also be used to compare the performance of different machine learning models on the same data set and can also be helpful in selecting the values for a model's parameters that maximize the accuracy of the model—also known as parameter tuning.

Model Evaluation Techniques : Cross-Validation



Model Evaluation Metrics

- Model evaluation metrics are required to quantify model performance.
- The choice of evaluation metrics depends on a given machine learning task (such as classification, regression, ranking, clustering, topic modeling, among others).

Model Evaluation Metrics

- Classification Accuracy
- Confusion matrix
- Logarithmic Loss
- Area under curve (AUC)
- F-Measure
- Mean Absolute Error (or MAE)
- Root Mean Squared Error (RMSE)

Thank You

For your precious time