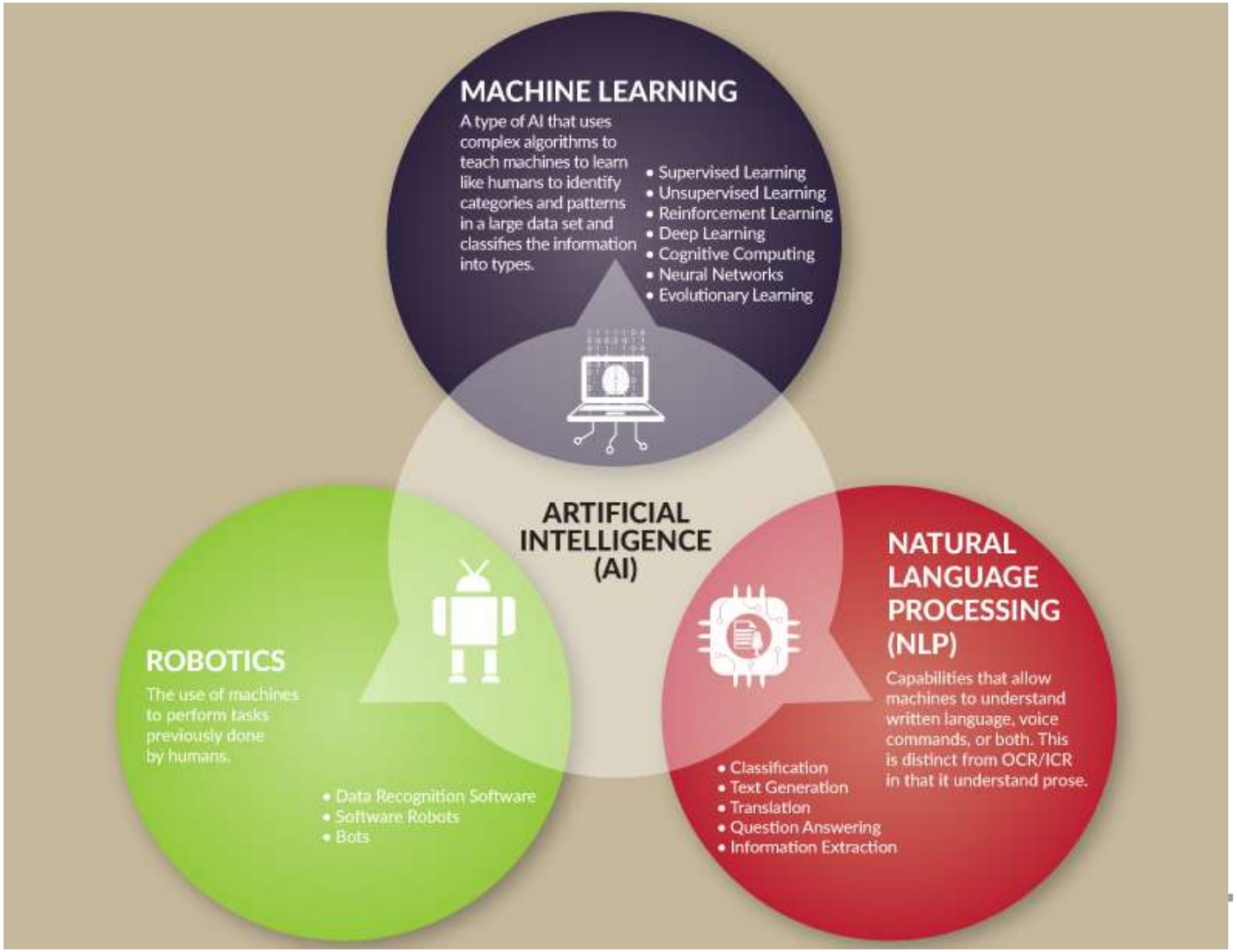
# Machine Learning

By: Dr. Pooja Jain

Indian Institute of Information technology, Nagpur

# Relationship b/w AI and ML.



# Machine Learning Vs Traditional Programming

> Interviewer: What's your biggest strength?
>
> Me: I'm a fast learner.
>
> Interviewer: What's 11 * 11?
>
> Me: 65.
>
> Interviewer: Not even close. It's 121.
>
> Me: It's 121.

**Traditional Programming**


**Machine Learning**

# Machine Learning

**Definition:**
Machine Learning is the study of algorithms

- that improve their performance P
- at some task T
- with experience E.

A well-defined learning task is given by <P, T, E>.
If a computer program improves with experience then we can say that it has learned.

# Machine Learning

Improve on task T, w.r.t performance metric P, based on experience E.

T: Playing checkers

P: Percentage of games won against an arbitrary opponent

E: Playing practice games against itself

T: Recognizing hand-written words

P: Percentage of words correctly classified

E: Database of human-labeled images of handwritten words

T: Driving on four-lane highways using vision sensors

P: Average distance traveled before a human-judged error

E: A sequence of images and steering commands recorded while observing a human driver.

# Question

Suppose your email program watches which emails you do or do not mark as spam, and based upon on that learns how to better filter spam. What is the task T in this setting?

a) Classifying emails as spam or not spam

b) Watching you label emails as spam or not

c) The number (or fraction) of emails correctly classified as spam or not spam

d) None of the above- this is not a ML problem

# Question

Suppose your email program watches which emails you do or do not mark as spam, and based upon on that learns how to better filter spam. What is the task T in this setting?

a) Classifying emails as spam or not spam

b) Watching you label emails as spam or not

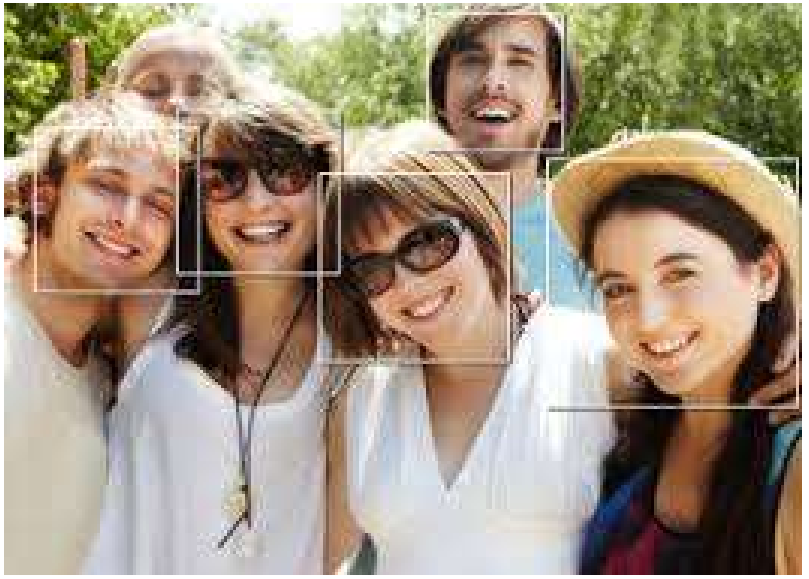
c) The number (or fraction) of emails correctly classified as spam or not spam

d) None of the above- this is not a ML problem

# When Do We Use Machine Learning ?

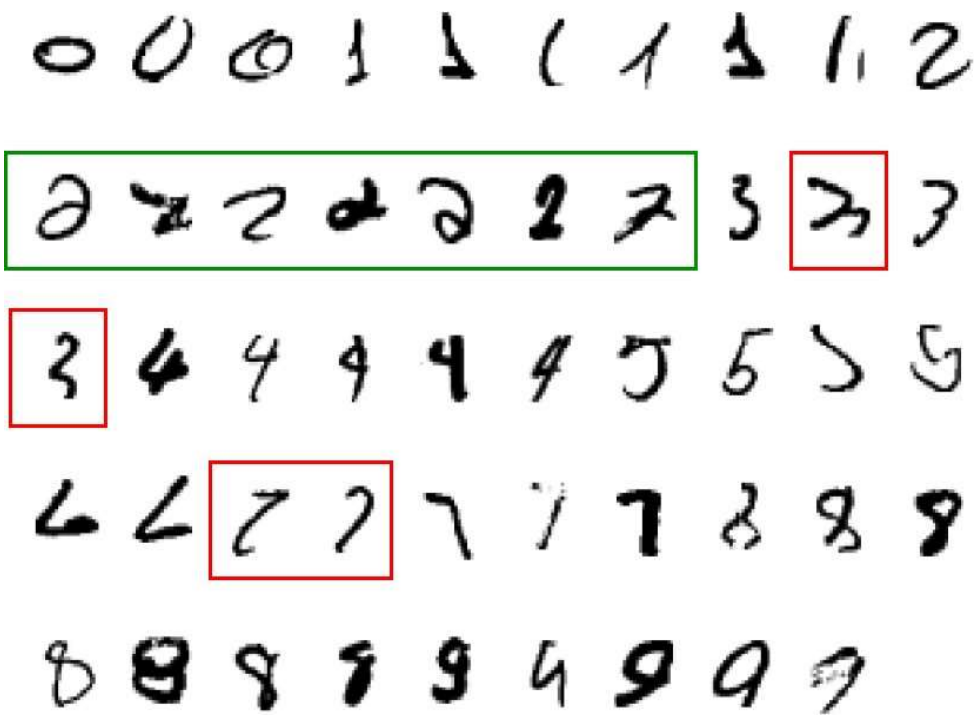
**ML is used when:**

- Humans can't explain their expertise -
  Face Recognition
  - Facebook pre-tags a user using facial recognition software to match newly uploaded photos with photos that have been tagged elsewhere.
  - Recognize Faces in the Image
- Image Classification
  - Process in computer vision that can classify an image according to its visual content.



# When Do We Use Machine Learning

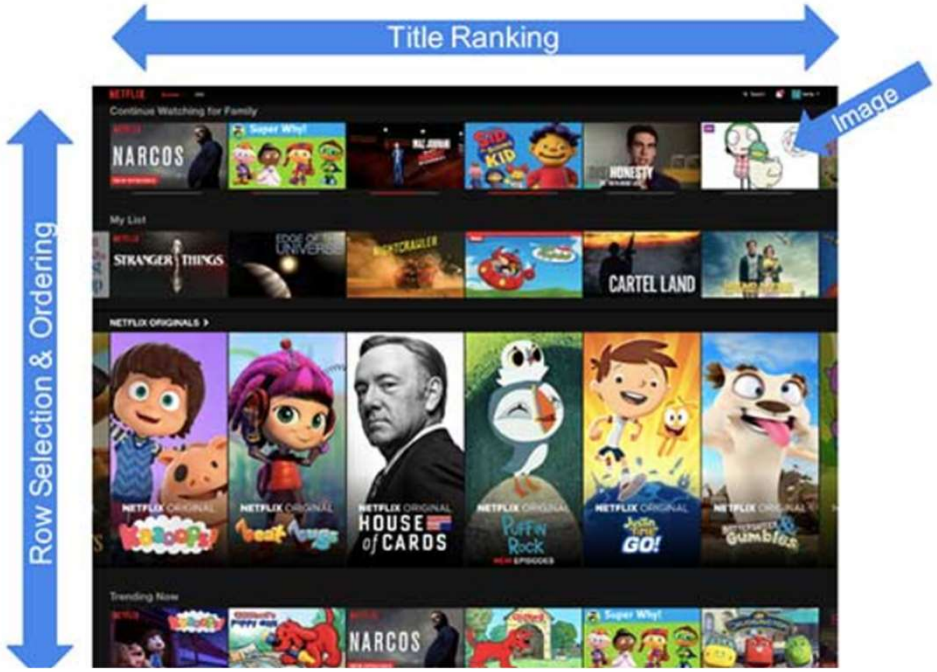
- No well defined algorithm to say what is “2”



# When Do We Use Machine Learning ?

**ML is used when:**

- Models must be customized - <u>Recommendation Systems</u>
  - Netflix developed machine learning models to predict movie ratings, user preferences and recommend movies to users.
  - Recommendation systems are based on collaborative filtering models.



# When Do We Use Machine Learning

**ML is used when:**

- Human expertise does not exist (Autonomous Robots)

- Models are based on huge amounts of data – Decision Making
  - Forecast sales
  - Predict downfalls in the stock market
  - Identify risks and anomalies, etc.

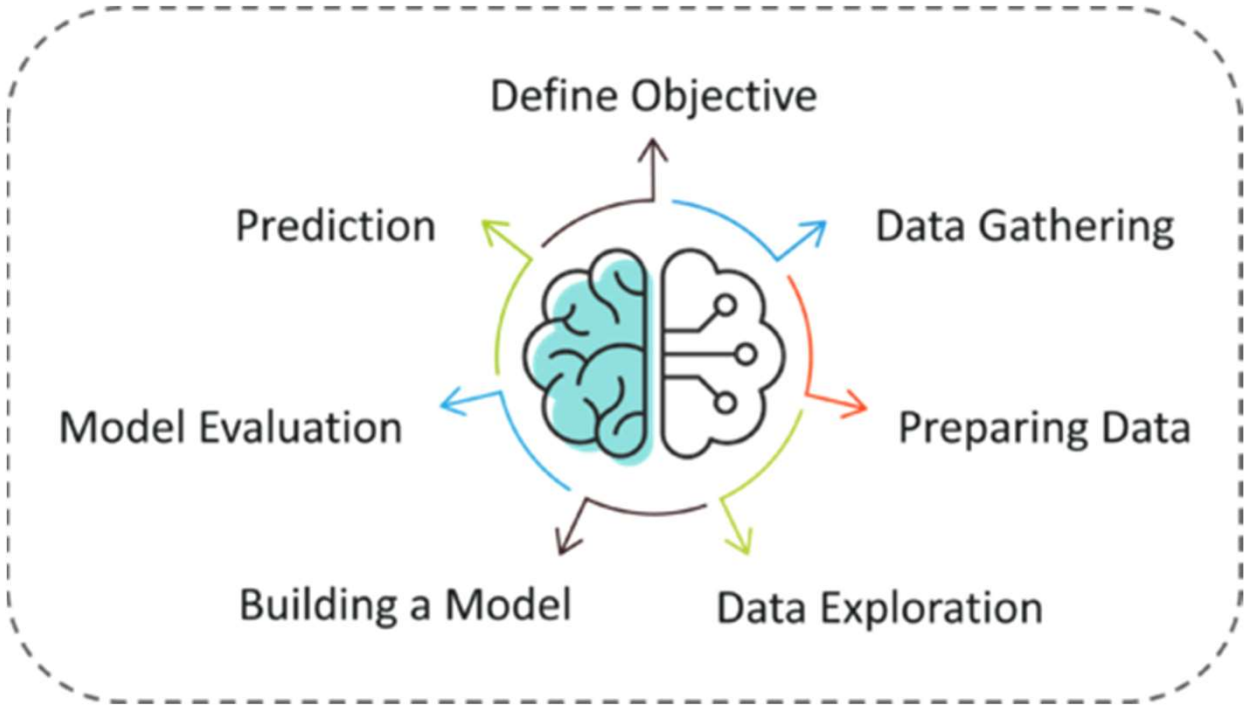
**Learning isn't always useful:**

- There is no need to "learn" to calculate matrix inverse!

# Types of Learning

- Supervised (inductive) learning
  - Given: training data + desired outputs (labels)
- Unsupervised learning
  - Given: training data without ground truth labels
- Semi-supervised learning
  - Given: training data + a few desired outputs
- Reinforcement learning
  - Rewards from sequence of actions

# Solving Machine Learning Problem



# Step 1: Define the problem statement

1. Understand what exactly needs to be predicted/estimated
    1. Class of the Image
    2. Whether Email is spam or not.
    3. Cancer Prediction

2. What kind of data can be used to solve this problem ?
    1. Class of the Image : Set of images of each class
    2. Email : Text of emails with labels
    3. Cancer Prediction : Blood Reports

3. Type of approach you must follow to get to the solution.
    1. Neural Networks
    2. Linear Regression/SVM

# Step 2 : Data Gathering & Preparation

1. Is the data available?
2. How can I get the data?
    I. Publicly Available Datasets such as MNIST, CIFAR-10.
    II. Obtain Datasets from Kaggle.
3. Remove Inconsistencies
    1. Missing Values
    2. Redundant Variables
    3. Duplicate Values

# Step 3 :Exploratory Data Analysis

1. Diving deep into data and finding all the hidden data mysteries.
2. Data Exploration involves understanding the patterns and trends in the data.
3. All the useful insights are drawn and correlations between the variables are understood.

# Step 4 : Training a machine learning model


Interviewer: What's your biggest strength?

Me: I'm an expert in machine learning.

Interviewer: What's 9 + 10?

Me: Its 3.

Interviewer: Not even close. It's 19.

Me: It's 16.

Interviewer: Wrong. Its still 19.

Me: It's 18.

Interviewer: No, it's 19.

Me: it's 19.

Interviewer: You're hired

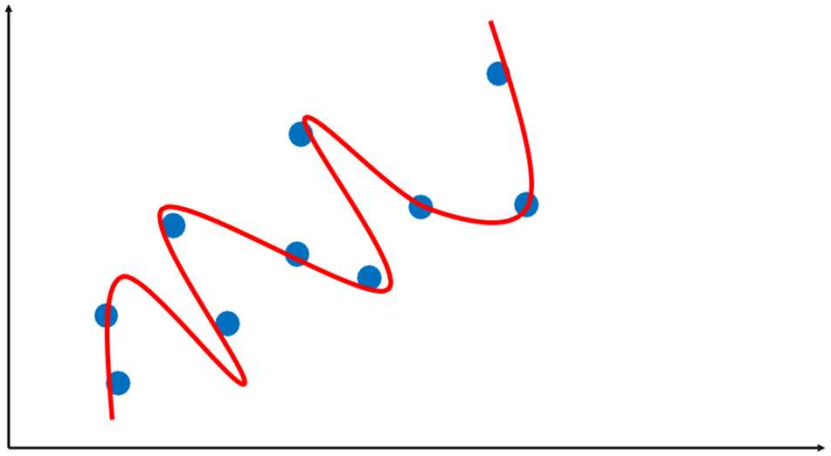
# Step 4 : Training a machine learning model

1. All the insights and patterns derived during Data Exploration are used to define the hypothesis function/model.
2. This stage always begins by splitting the data set into two parts
    1. Training data ( 80% ) and
    2. Testing data ( 20% ).
3. The training data will be used to train the model while testing data will be used to validate the results of the machine learning model.

# Step 5 :Model Evaluation and Optimization

1. After building a model by using the training data set, it is finally time to put the model to a test.
2. The testing data is used to check the efficiency of the model and how accurately it can predict the outcome. Improvements in the model are done at this stage.
3. Methods like parameter tuning and cross-validation can be used to improve the performance of the model. Confusion Matrix also gives insights to the trained model.

# Step 5 :Model Evaluation and Optimization

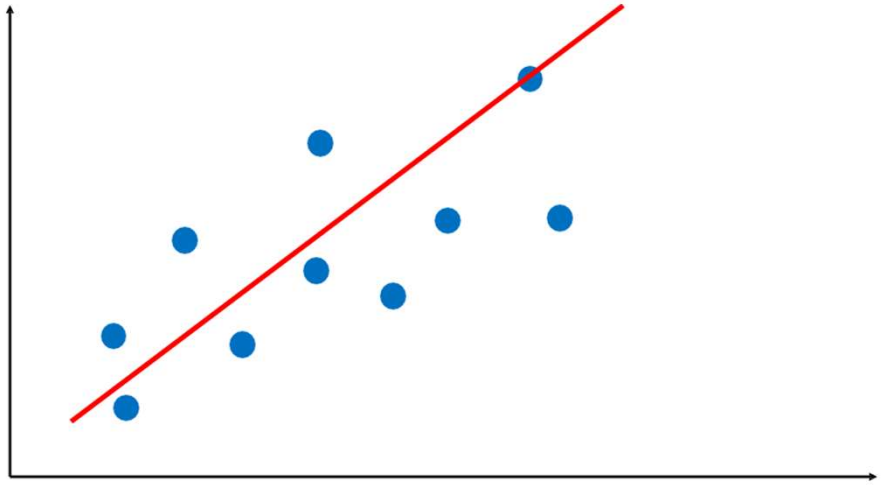


Overfitting of Model

( Learned the Training Data but not able to generalize. )

High variance

Low bias



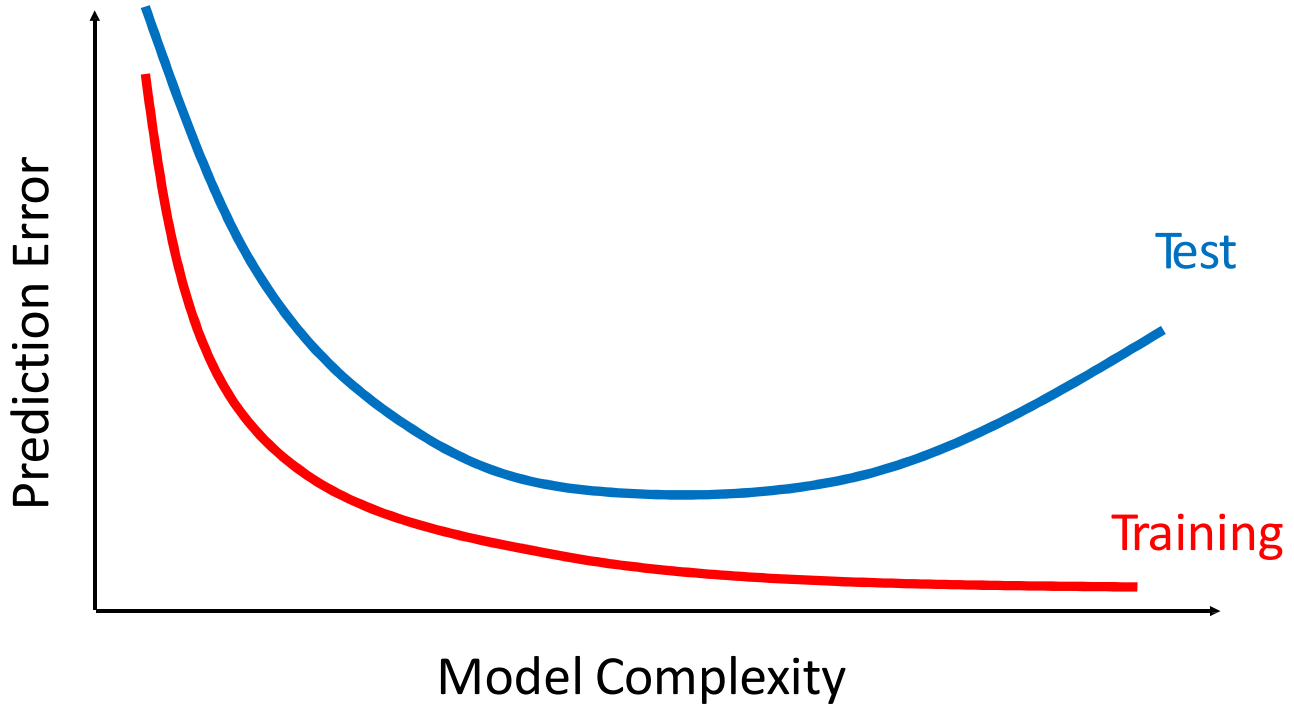
Underfitting of Model

( Model not able to learn the data properly )

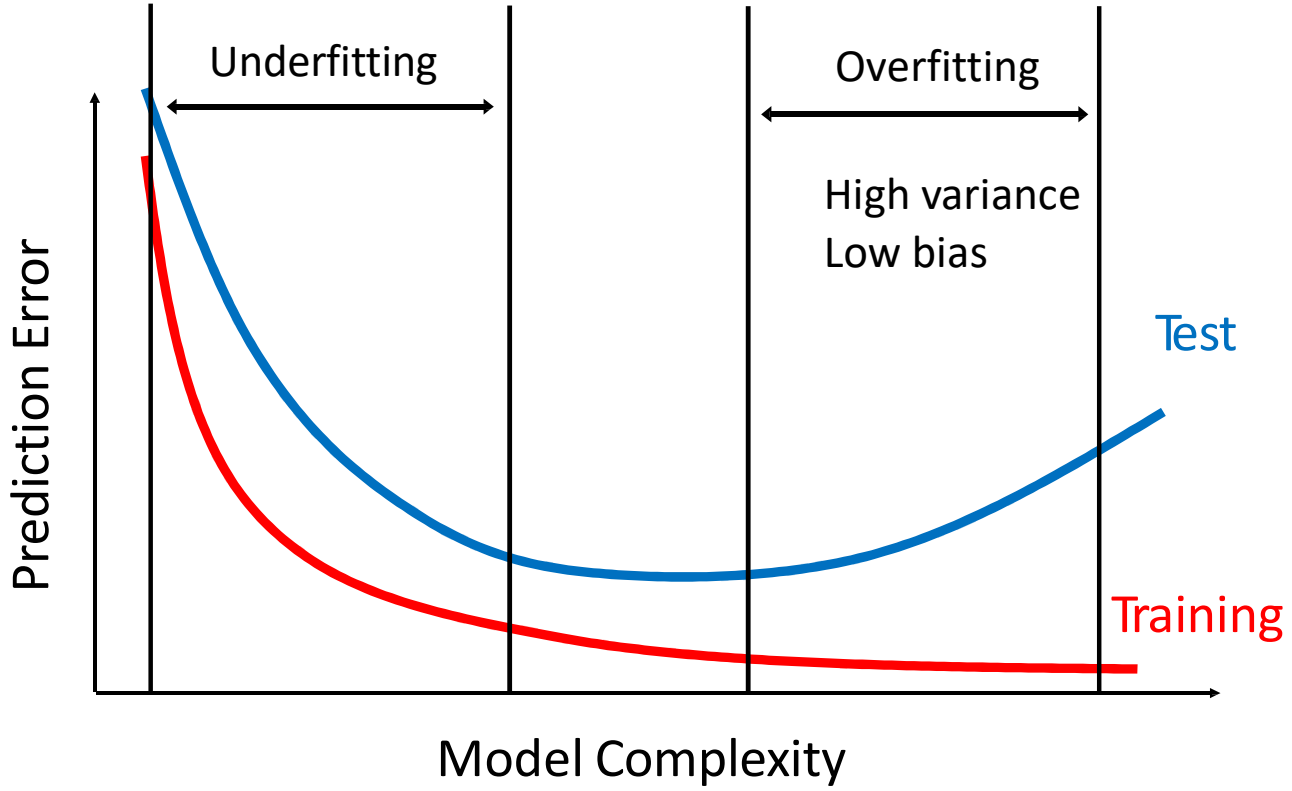
Low variance

High bias

# Model Complexity Vs Performance



# Model Complexity Vs Performance



# Confusion Matrix

- Given a dataset of P positive instances and N negative instances:

| actual class \ predicted class | Yes | No |
|---|---|---|
| Yes | TP | FN |
| No | FP | TN |

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}$$

- For a classifier which identifies positive cases (e.g. tumor detection)

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

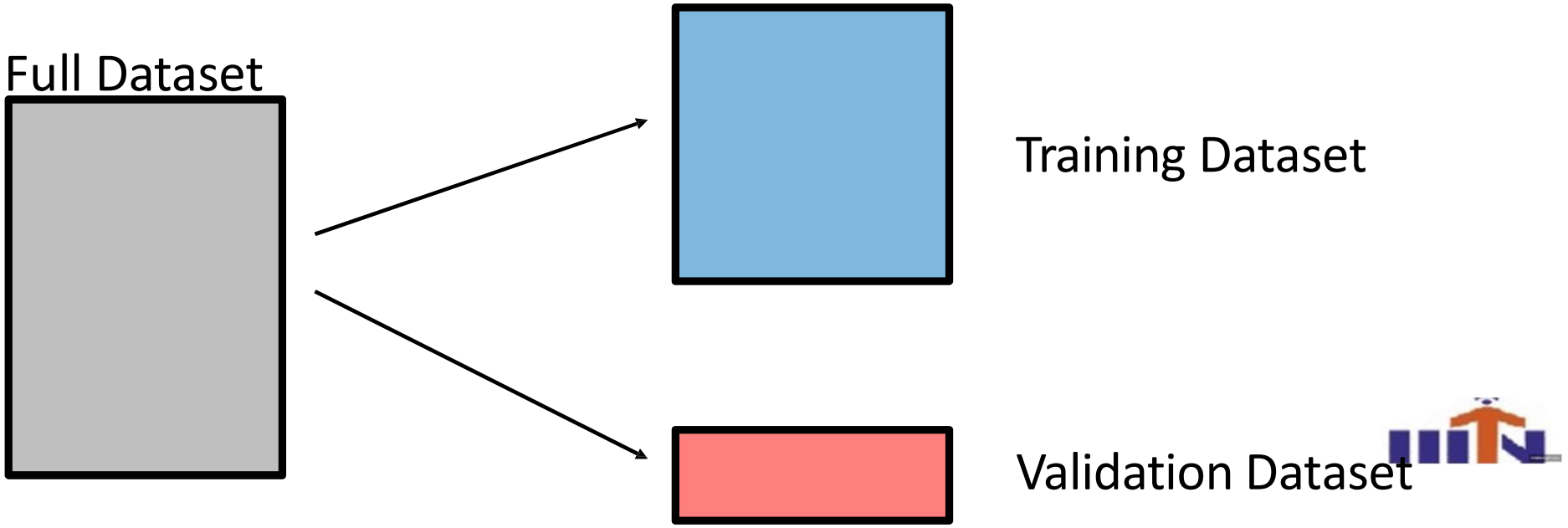
probability that a randomly selected result is relevant

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

probability that a randomly selected relevant object is retrieved

# Validation

- **Idea:** Create multiple models with different “hyperparameters”.
- Train each model on the “training data”.
- Test each model’s accuracy on the validation data, and report the best model.



# Step 6 :Predictions

1. Once the model is evaluated and improved, it is finally used to make predictions.
2. The final output can be:
    1. Categorical variable (e.g. True or False)
    2. It can be a Continuous Quantity (e.g. the predicted value of a stock).

# Classification Metrics

$$\text{accuracy} = \frac{\#\text{correct predictions}}{\#\text{test instances}}$$

$$\text{error} = 1 - \text{accuracy} = \frac{\#\text{incorrect predictions}}{\#\text{test instances}}$$

# ML in a Nutshell

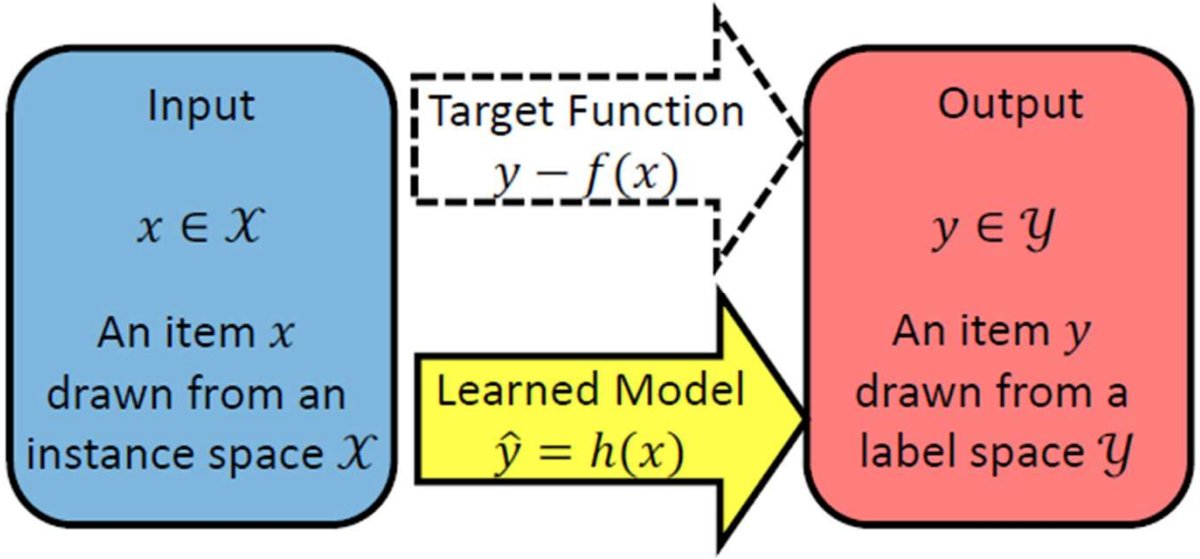
Every ML technique has three components:

- Representation/Hypothesis
- Evaluation/Loss
- Optimization

# Hypothesis Space for Various ML Techniques

We need to choose what kind of model we want to learn.

- Numerical functions
  - Linear regression
  - Neural networks
  - Support vector machines
  - Bayesian Model
  - Naïve Bayes
- Symbolic functions
  - Decision trees
  - Rules in first-order predicate logic



# Evaluation: Loss Function

- How do we measure how "good" a hypothesis is on the training data?
- Typically done by introducing a **loss function**

$$\mathrm{l} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_{\&}$$

- E.g. for regression, squared loss:

$$l(h_\theta(x), y) = (h_\theta(x) - y)^2$$

- Other Loss Evaluation Functions

  I. Accuracy, F1-score
  II. Mean Squared Error (mse)
  III. Posterior ProbabilityCross - Entropy
  IV. K-L Divergence

# Search/Optimization Algorithms

- Convex Optimization

  Gradient Descent in Regression

- Combinatorial Optimization

  Grid Search in SVM for hyperparameter tuning

- Constrained Optimizaion

  Sequential Minimal Organisation to solve SVM/