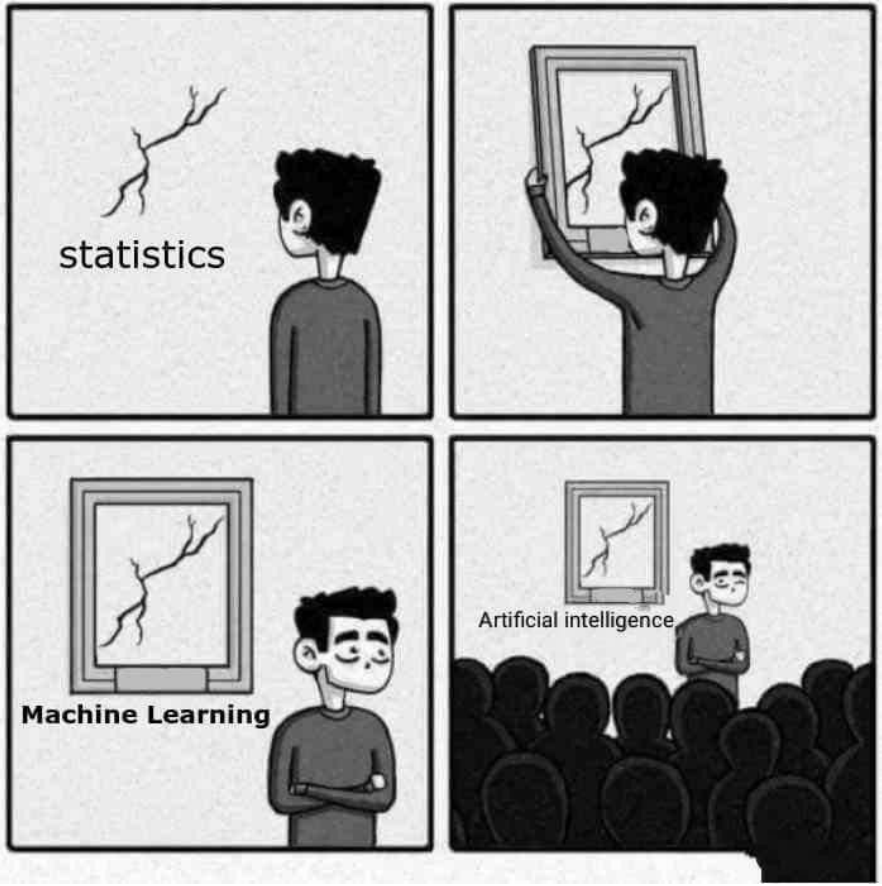
KonvergeAI

CCT 303 AI&CS UNIT 1 Foundations for ML

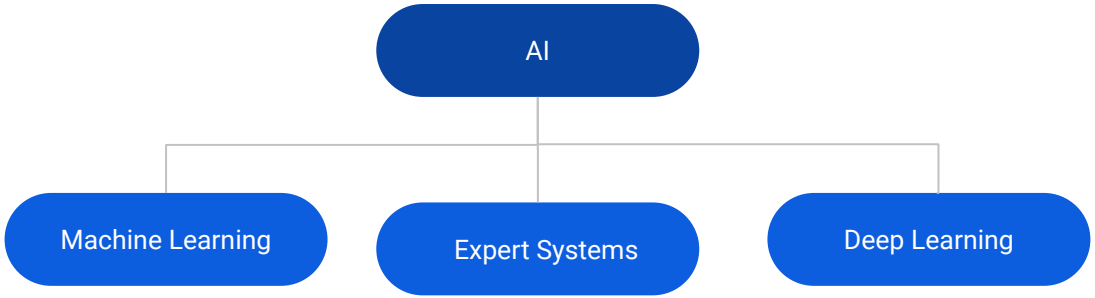
ML Techniques Overview

AI Landscape



https://www.instagram.com/sandserifcomics/

AI Landscape...



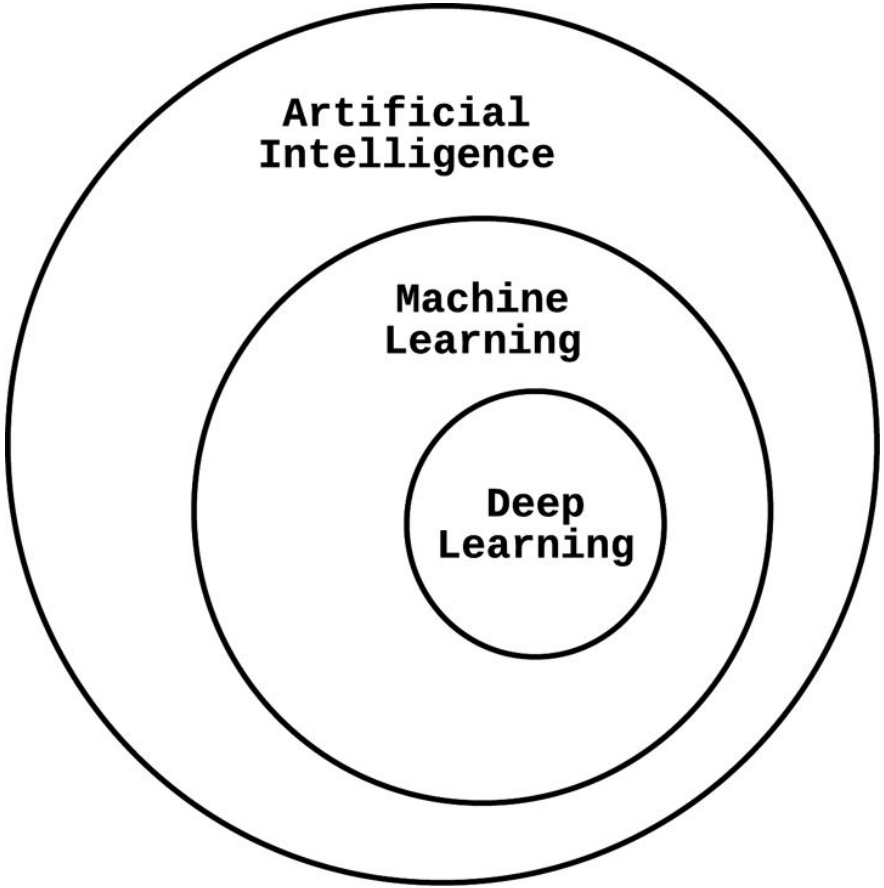
Stats + Algorithms

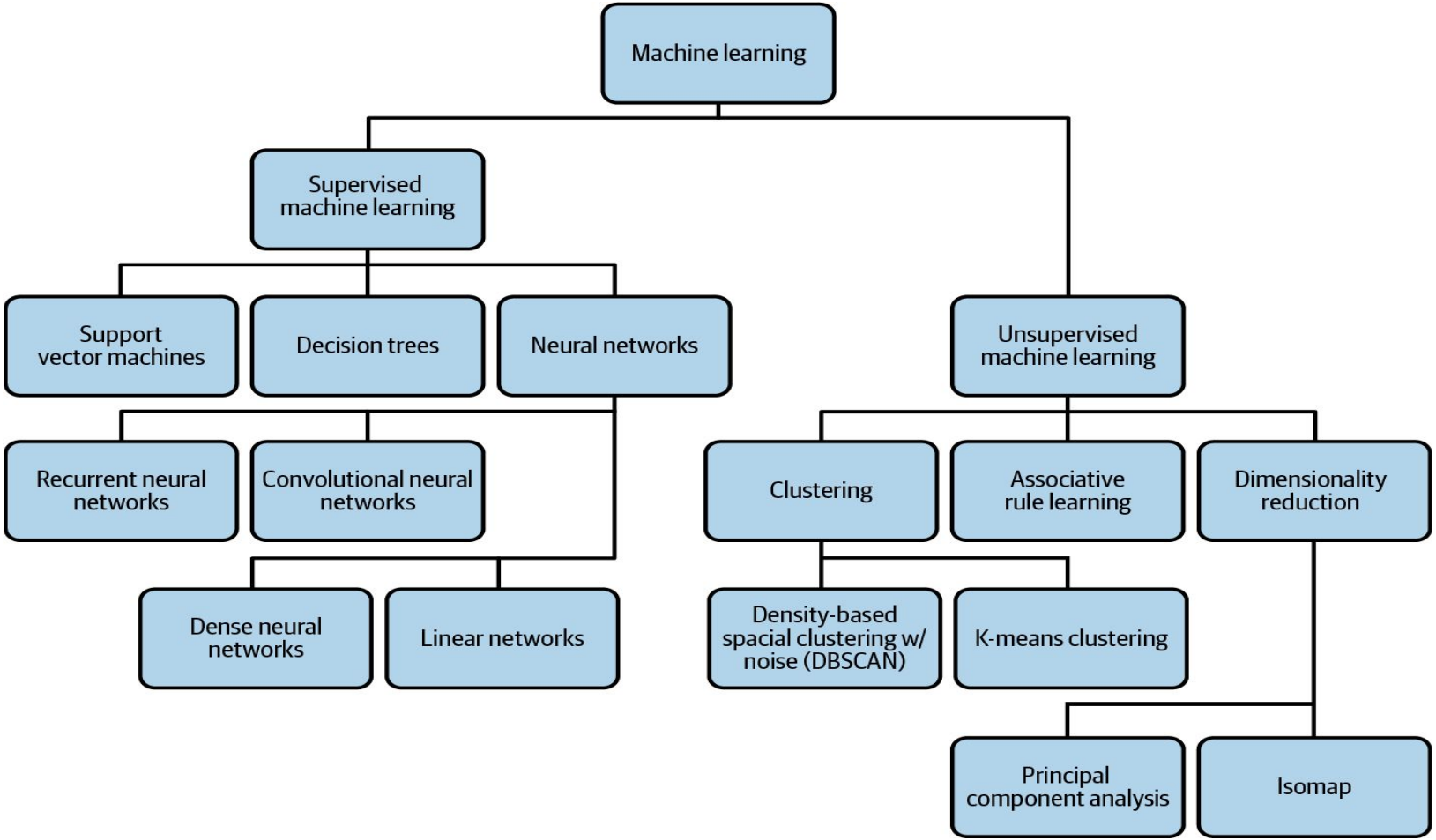
Rule based systems to solve specific problems

ML driven by Neural Networks

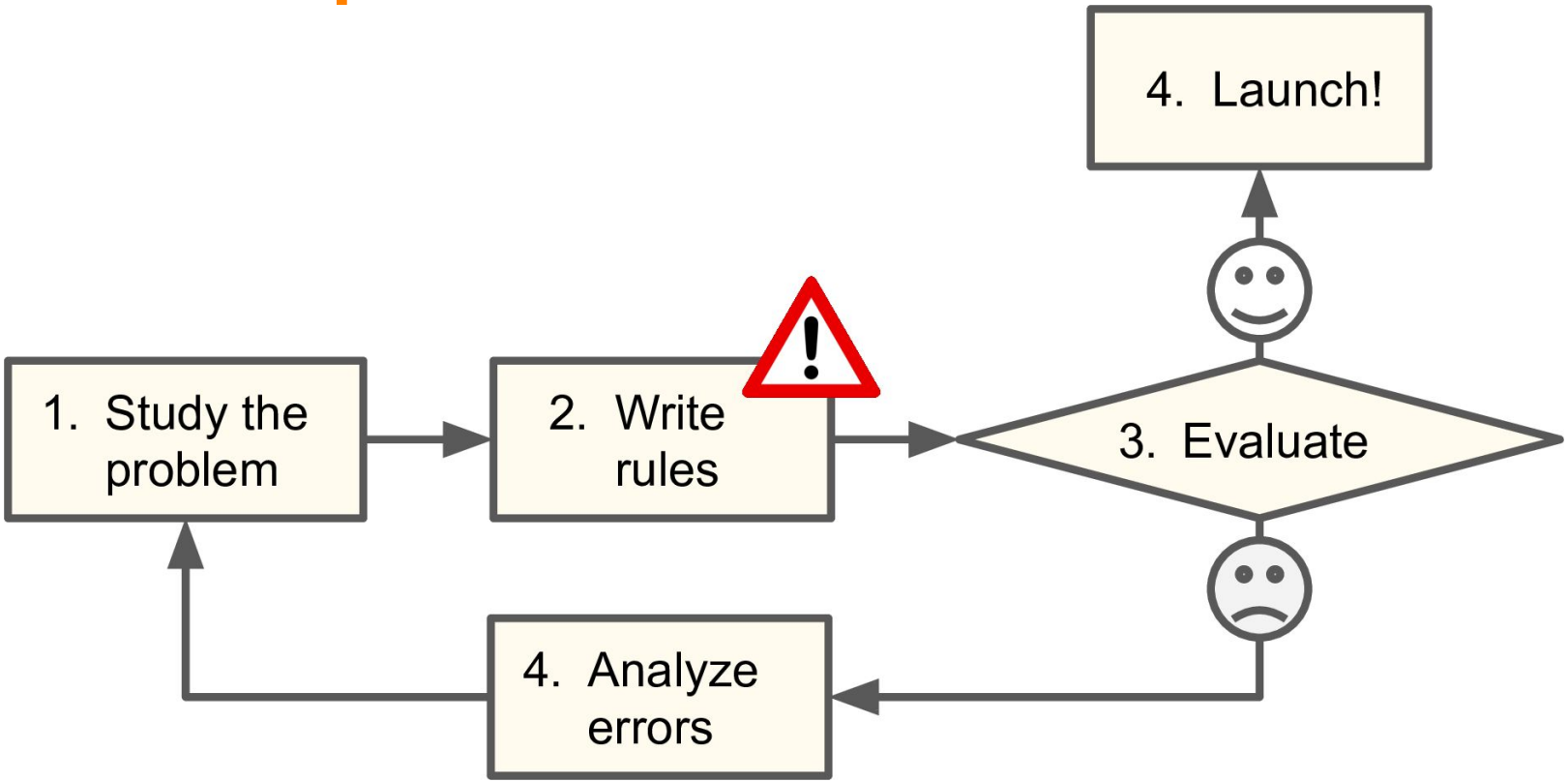
What are the factors driving the development of AI ?

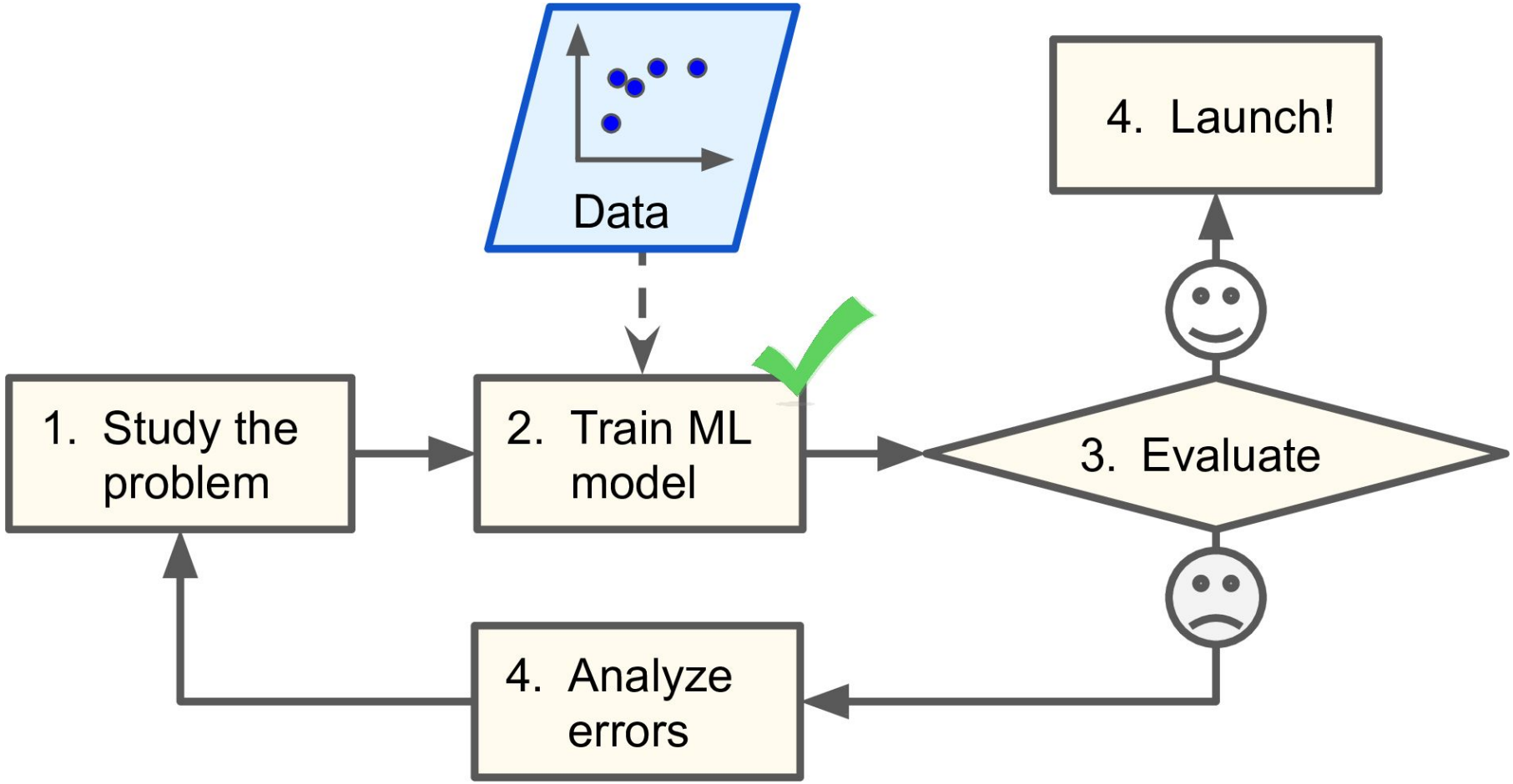
Another way of looking

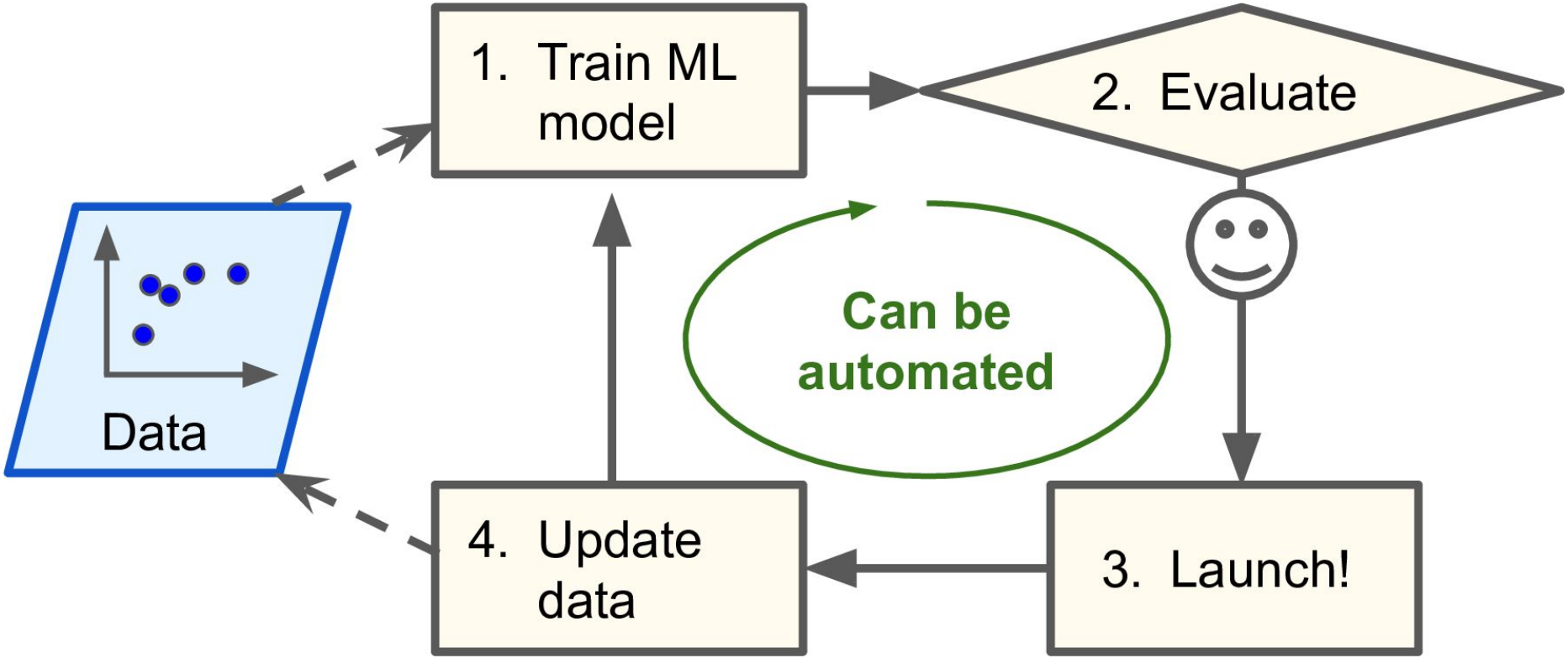



Machine learning
Supervised machine learning
Support vector machines
Decision trees
Neural networks
Recurrent neural networks
Convolutional neural networks
Dense neural networks
Linear networks
Unsupervised machine learning
Clustering
Associative rule learning
Dimensionality reduction
Density-based spacial clustering w/ noise (DBSCAN)
K-means clustering
Principal component analysis
Isomap

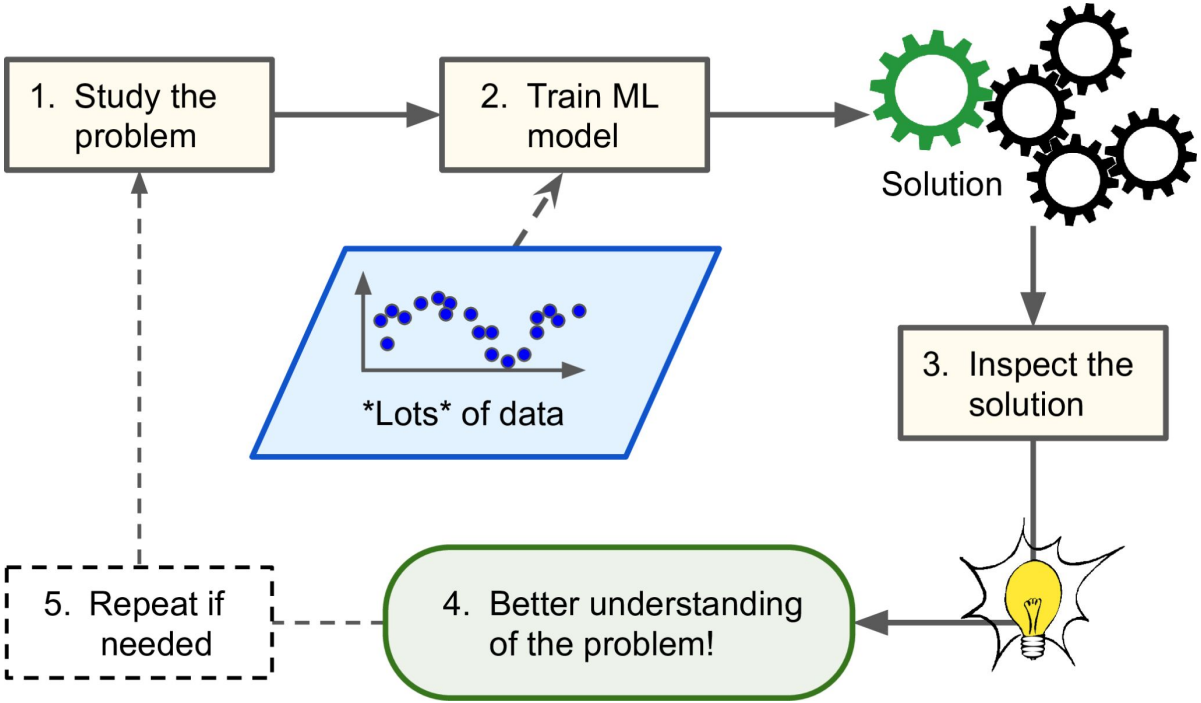
Why do we need ML?
The Email Spam Problem




Data
4. Launch!
1. Study the problem
2. Train ML model
3. Evaluate
4. Analyze errors


1. Train ML model
2. Evaluate
Data
Can be automated
4. Update data
3. Launch!

Help Humans Learn



For instance, once a spam filter has been trained on enough spam, it can easily be inspected to reveal the list of words and combinations of words that it believes are the best predictors of spam. Sometimes this will reveal unsuspected correlations or new trends, and thereby lead to a better understanding of the problem.

Where to use ML?

- Problems for which existing solutions require a lot of fine-tuning or long lists of rules: one Machine Learning model can often simplify code and perform better than the traditional approach.
- Complex problems for which using a traditional approach yields no good solution: the best Machine Learning techniques can perhaps find a solution.
- Fluctuating environments: a Machine Learning system can easily be retrained on new data, always keeping it up to date.
- Getting insights about complex problems and large amounts of data.

Some Typical Applications

Analyzing images of products on a production line to automatically classify them

Detecting tumors in brain scans

Forecasting your company's revenue next year, based on many performance metrics

Segmenting clients based on their purchases so that you can design a different marketing strategy for each segment

Recommending a product that a client may be interested in, based on past purchases

Questions to be kept in mind when designing an ML System?

What question(s) am I trying to answer? Do I think the data collected can answer that question?

What is the best way to phrase my question(s) as a machine learning problem?

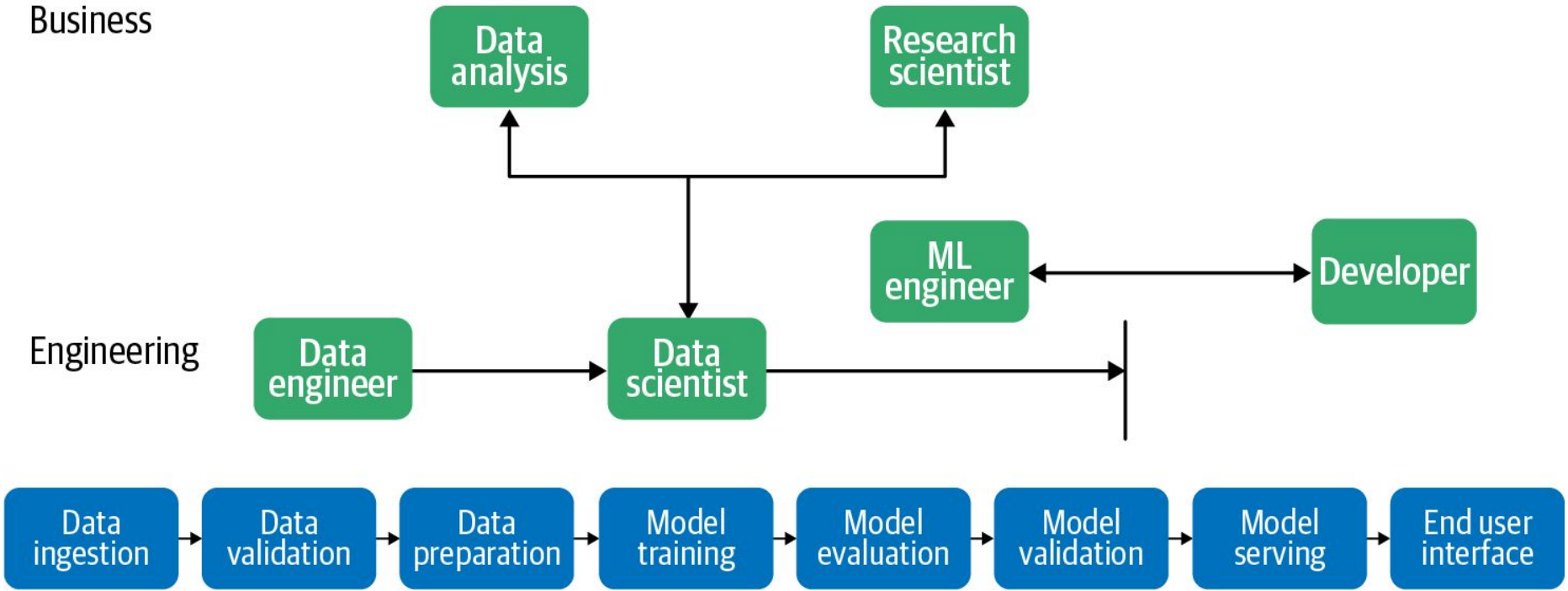
Have I collected enough data to represent the problem I want to solve?

What features of the data did I extract, and will these enable the right predictions?

How will I measure success in my application?

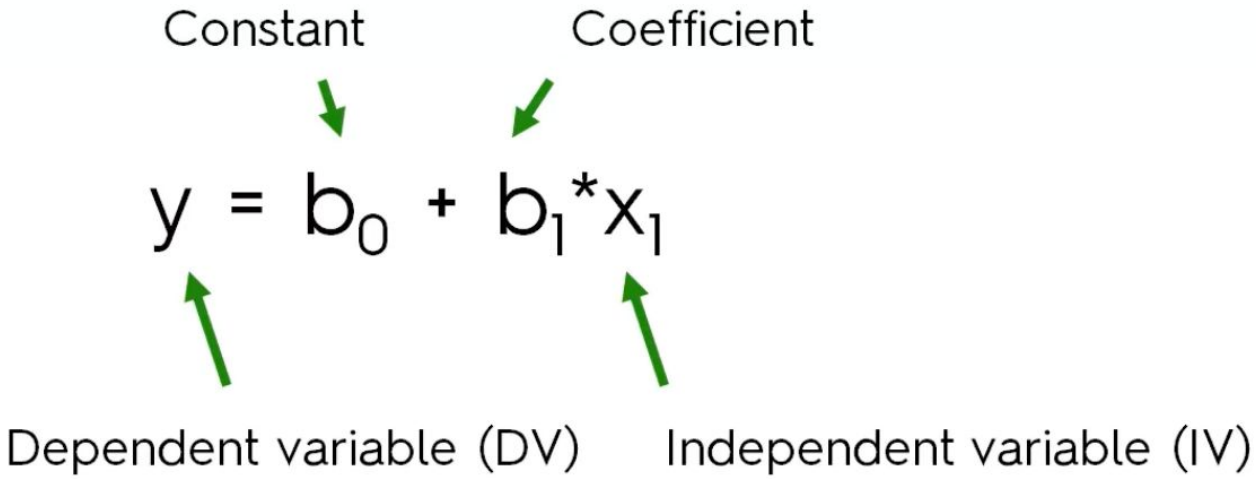
How will the machine learning solution interact with other parts of my research or business product?

Different Roles in an Organization's ML Model Development Process

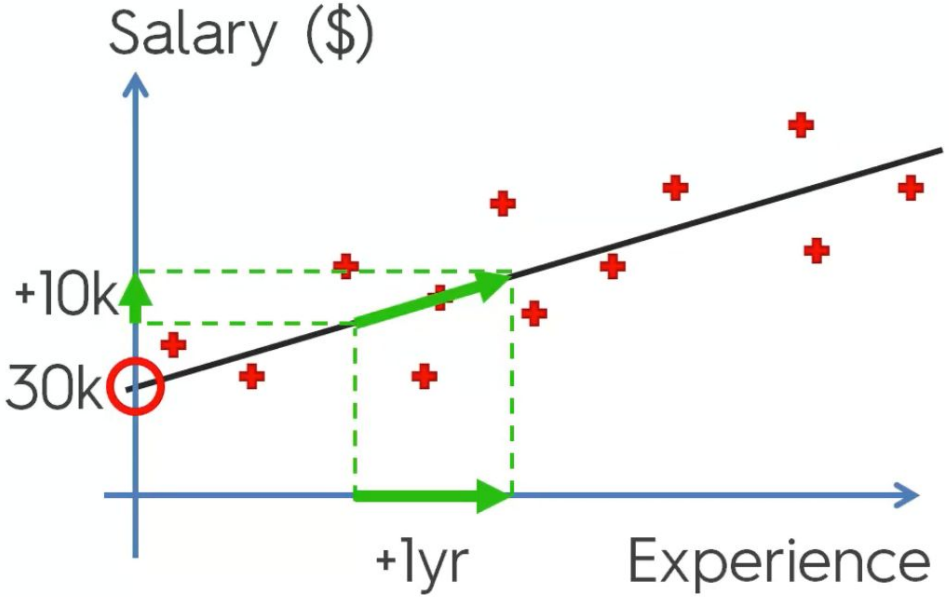


Linear and Logistic Regression

Simple Linear Regression



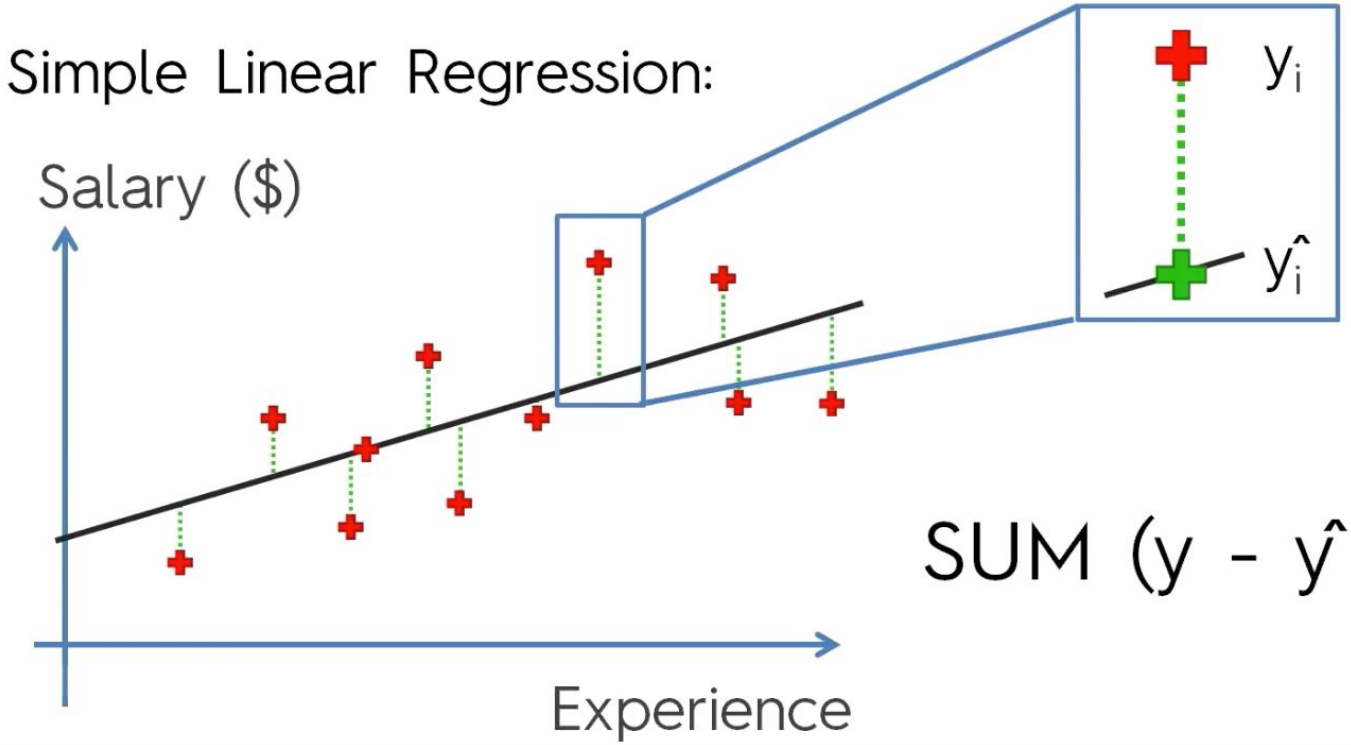
Simple Linear Regression:



$$y = b_0 + b_1 * x$$

Salary = b_0 + b_1 *Experience

Ordinary Least Squares



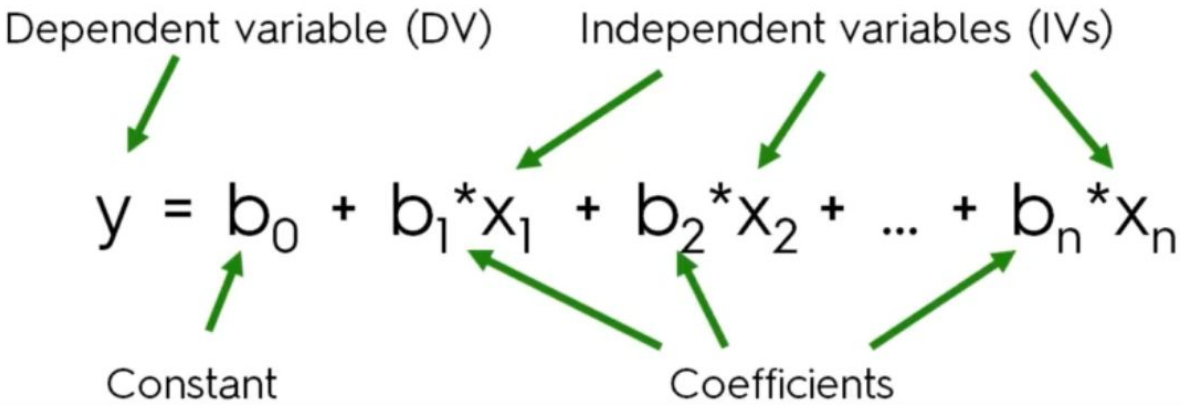
$$SUM\ (y - \hat{y})^2 \rightarrow min$$

Salary Dataset write a program for simple linear regression in python

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Multiple Linear Regression



Multiple Linear Regression

Problem Statement :-

Data of 50 companies.
R&D Spends/Administration/Marketing Spend/State (Independent Variables)
Profit (Dependent Variable)

A VC Fund is interested in investing in these companies. But has questions like :-
Where companies perform better? Are these companies are those who spend more money on R&D Spend or on Marketing Spend ?
Help VC Fund to build a model

Dummy Variable/Categorical Variable/One hot encoding

Profit	R&D Spend	Admin	Marketing	State	Dummy Variables: New York	Dummy Variables: California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

$$y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 + b_4*D_1$$

Dummy Variable Trap

Not Truly Independent Variables
Multicollinearity = High correlation between 2 or more independent variables

Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70			ia	0	1
191,050.39	153,441.51			ia	0	1
182,901.99	144,372.41			rk	1	0
166,187.94	142,107.34			ia	0	1

Dummy Variables

$D_2 = 1 - D_1$

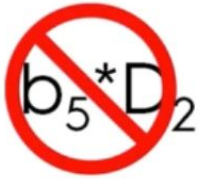
$$y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 + b_4*D_1 + b_5*D_2$$

Dummy Variable Trap

Dummy Variables

Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

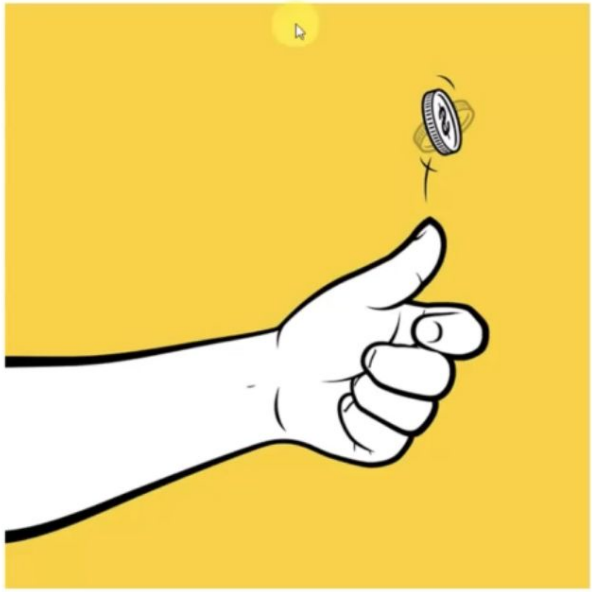
$$y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 + b_4*D_1 + b_5*D_2$$



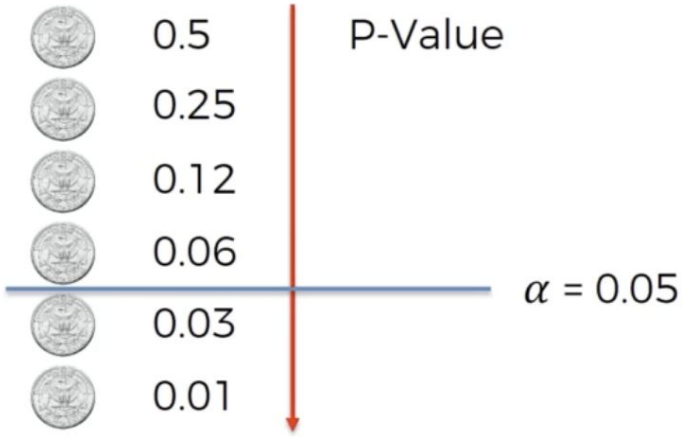
Always omit one dummy variable

P Value = How likely it is to get a particular result when the null hypothesis is assumed to be true.

Statistical Significance



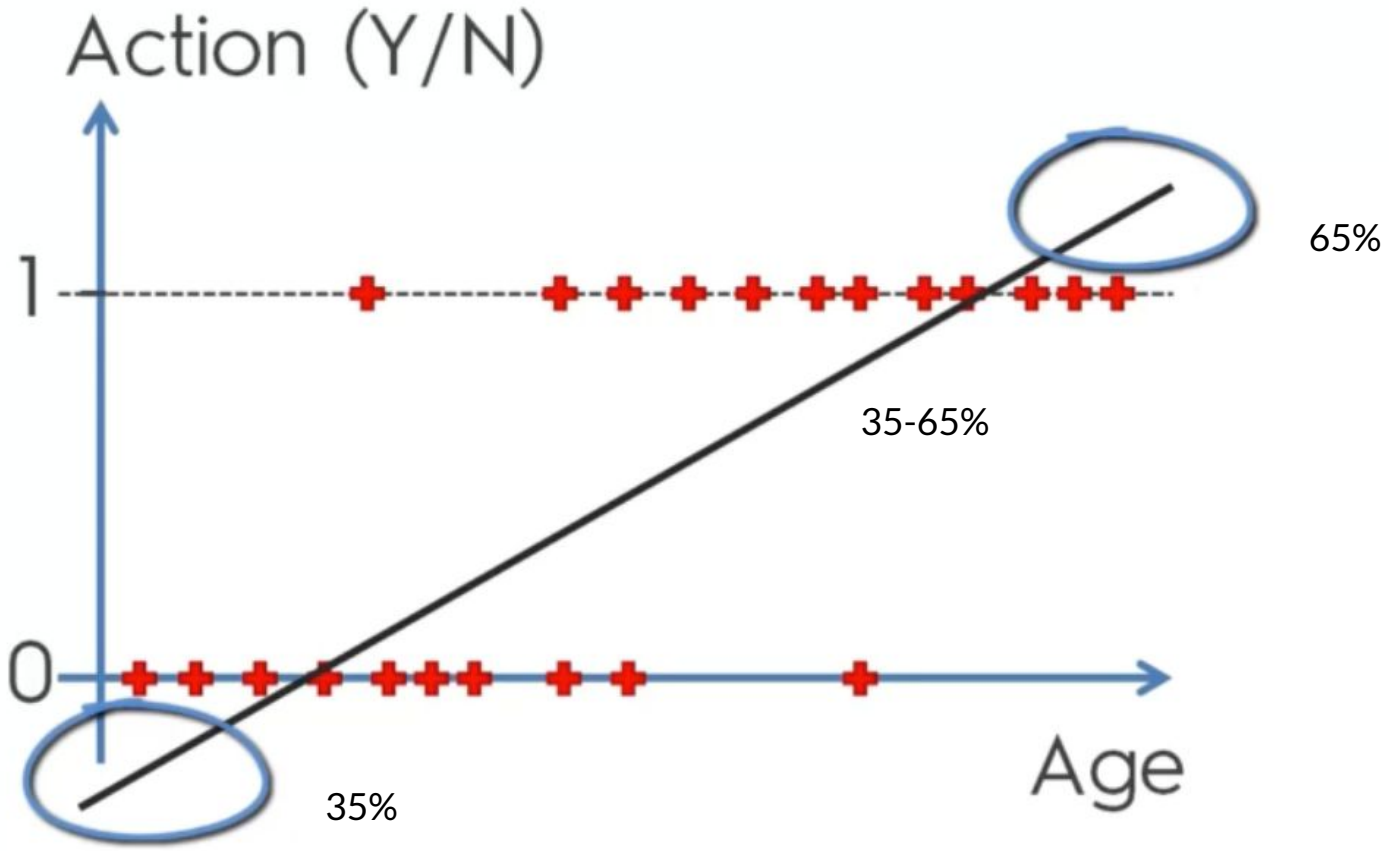
H_0: This is a fair coin
H_1: This is not a fair coin

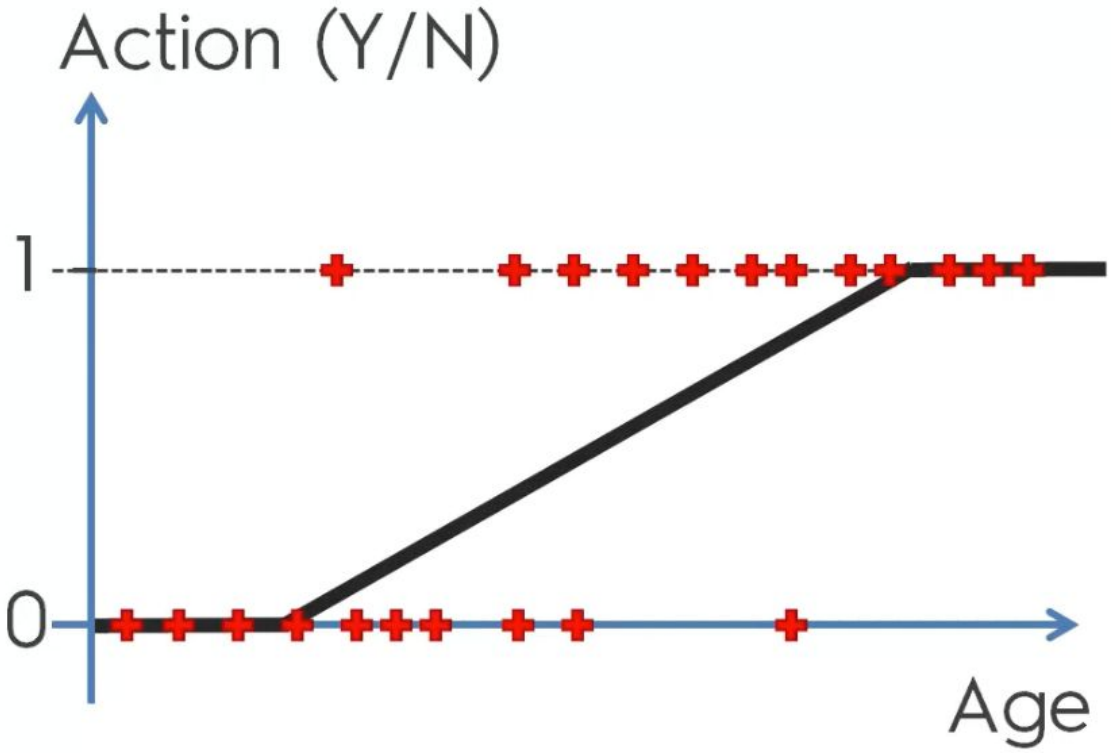


Try Building a Model

Logistic Regression

Unlike regression where you predict a continuous number, you use classification to predict a category. There is a wide variety of classification applications from medicine to marketing


Action (Y/N)
1
0
Age
65%
35-65%
35%

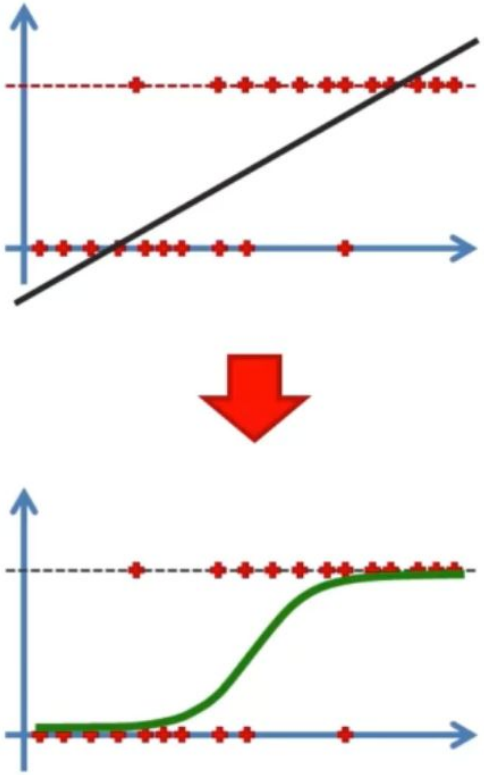

Action (Y/N)
1
0
Age

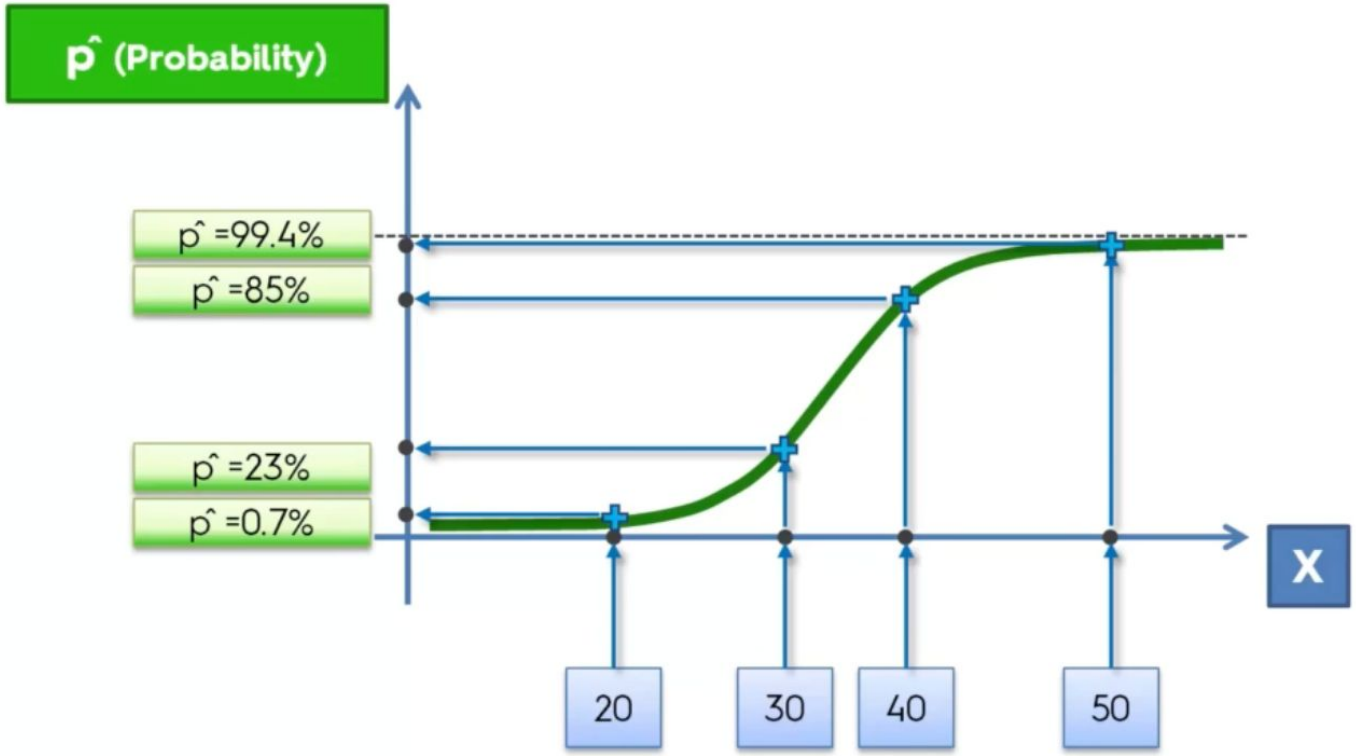
$$y = b_0 + b_1 * x$$

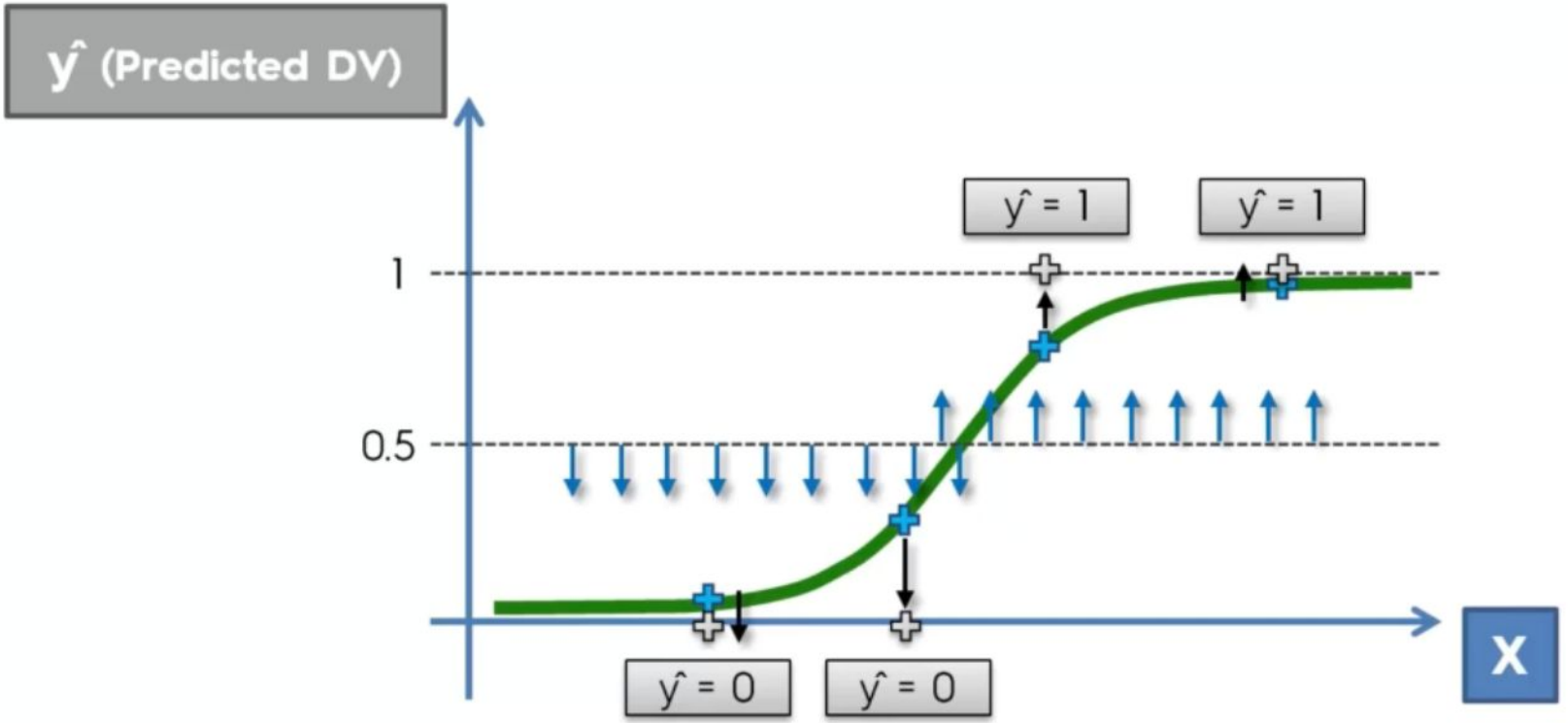
Sigmoid Function

$$p = \frac{1}{1 + e^{-y}}$$

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 * x$$

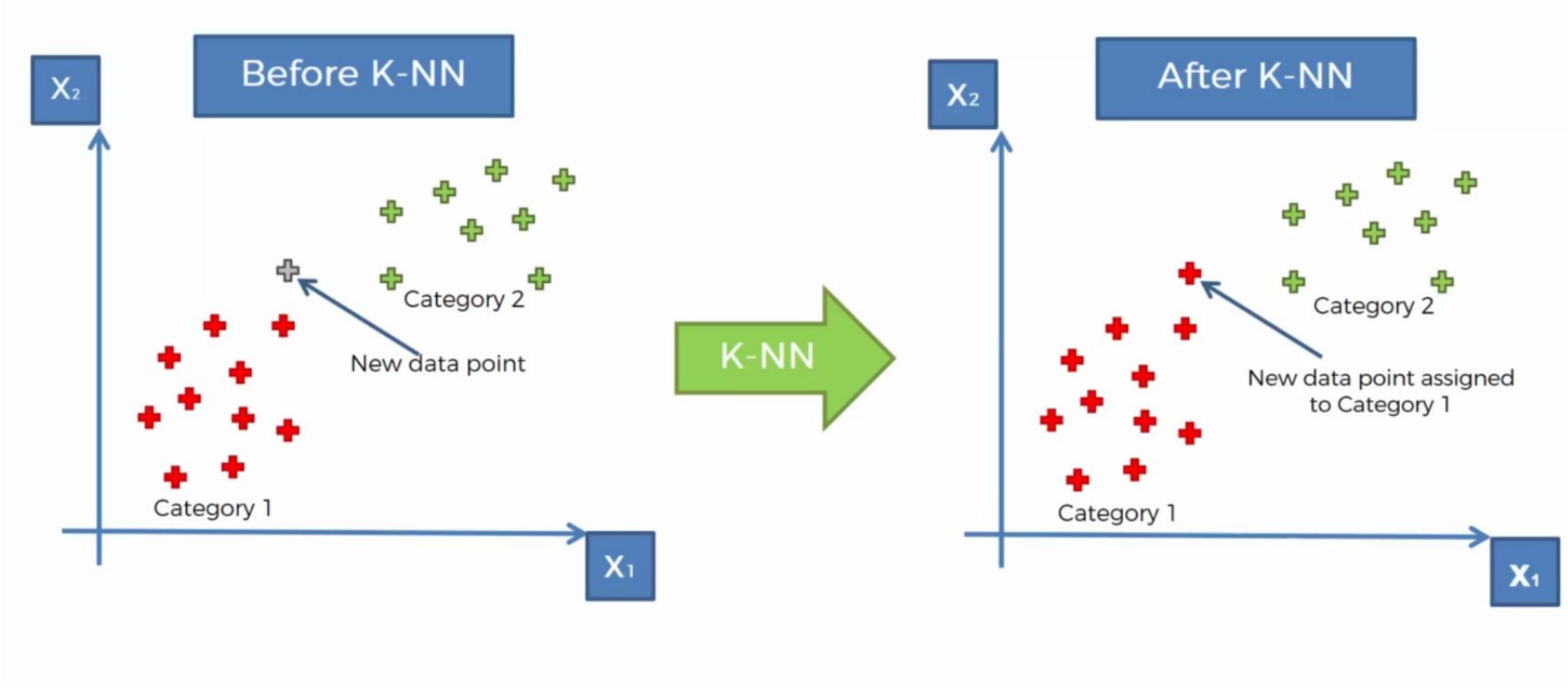



p̂ (Probability)
p̂ =99.4%
p̂ =85%
p̂ =23%
p̂ =0.7%
X
20
30
40
50


ŷ (Predicted DV)
1
0.5
ŷ = 1
ŷ = 1
ŷ = 0
ŷ = 0
X

Try Building a Model

K Nearest Neighbor


Before K-NN
X2
X1
Category 2
New data point
Category 1
K-NN
After K-NN
X2
X1
Category 2
New data point assigned
to Category 1
Category 1

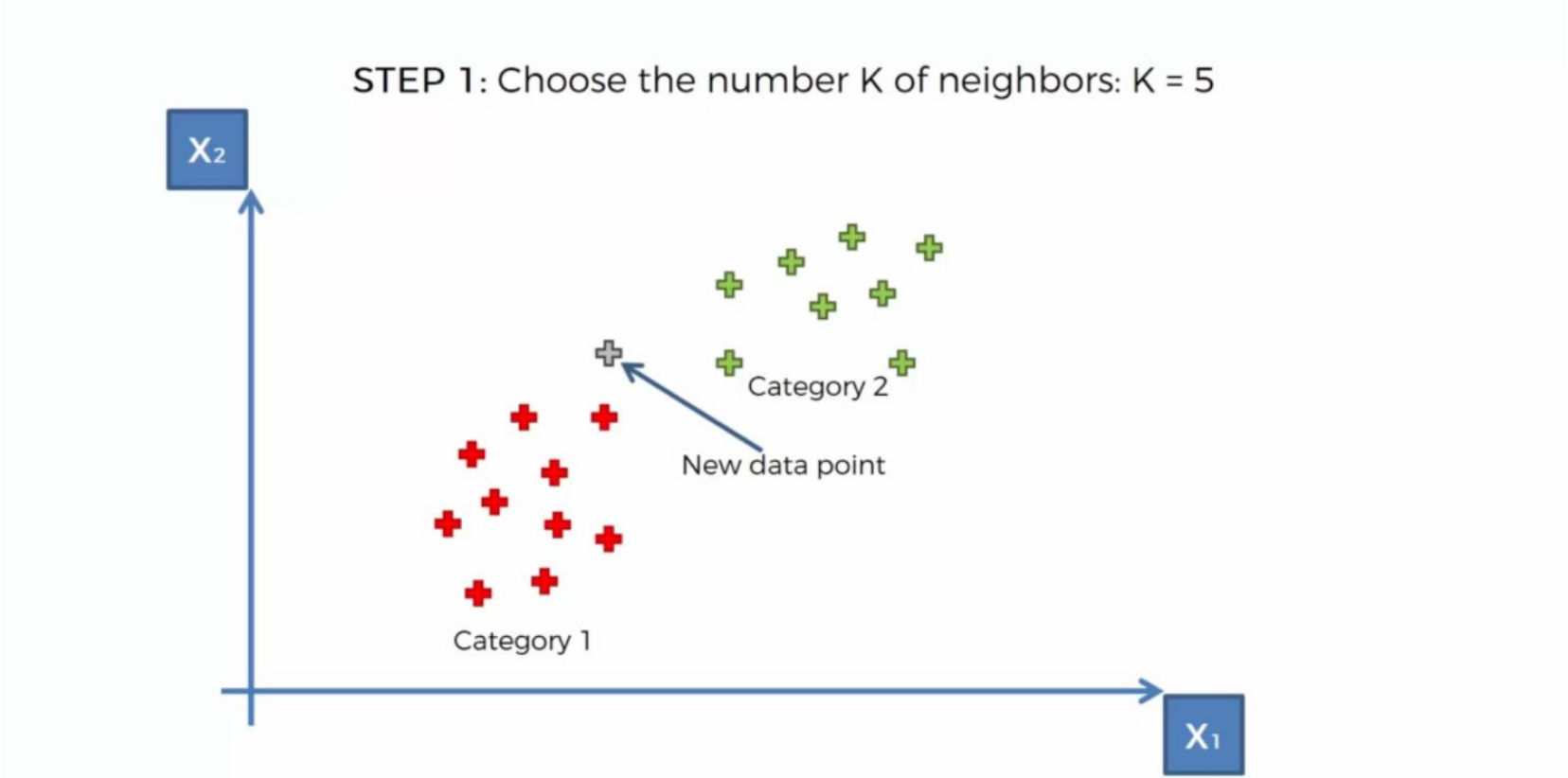
STEP 1: Choose the number K of neighbors

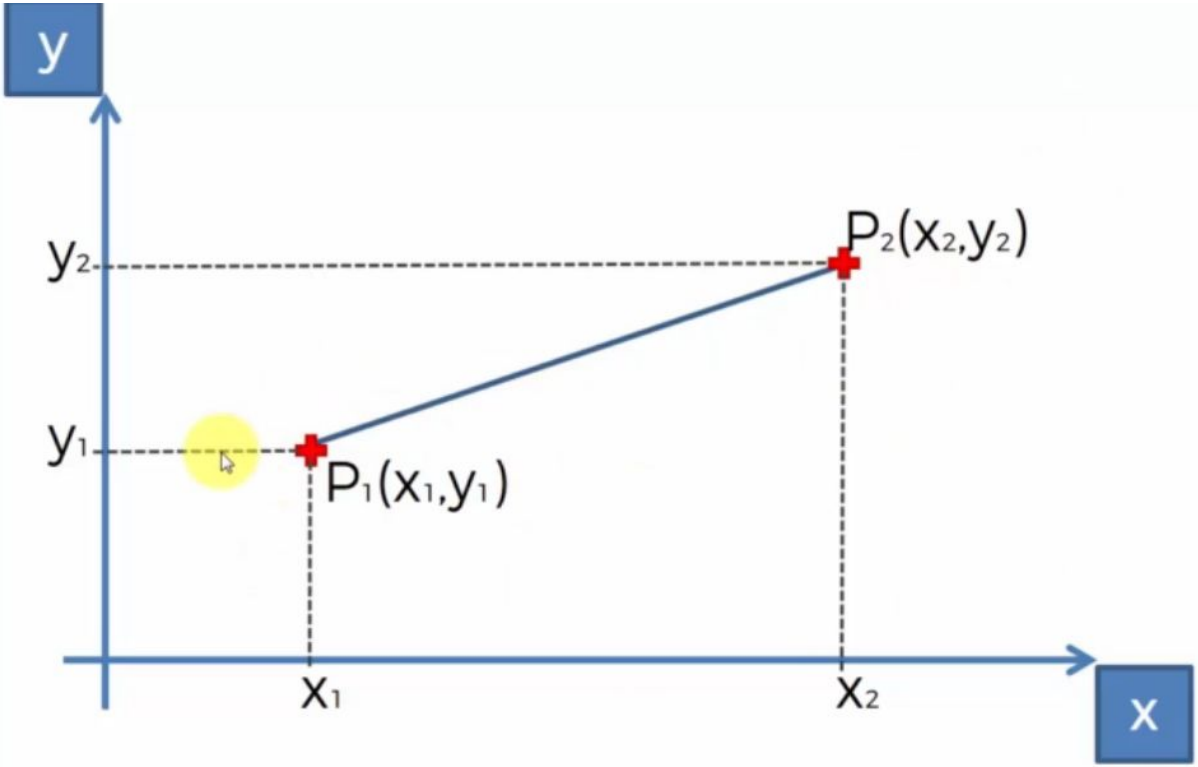
STEP 2: Take the K nearest neighbors of the new data point, according to the Euclidean distance

STEP 3: Among these K neighbors, count the number of data points in each category

STEP 4: Assign the new data point to the category where you counted the most neighbors

Your Model is Ready

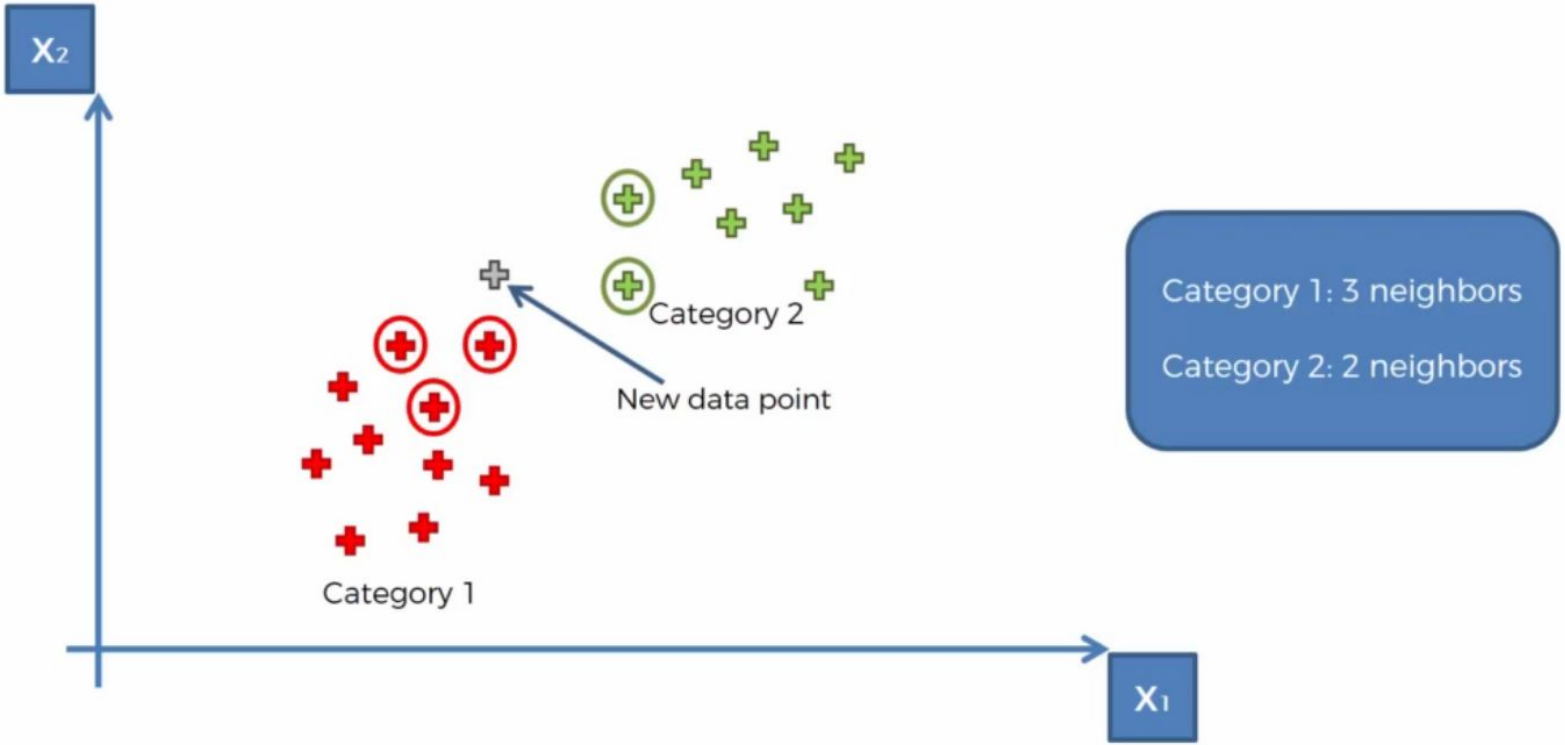

STEP 1: Choose the number K of neighbors: K = 5
X2
Category 2
New data point
Category 1
X1



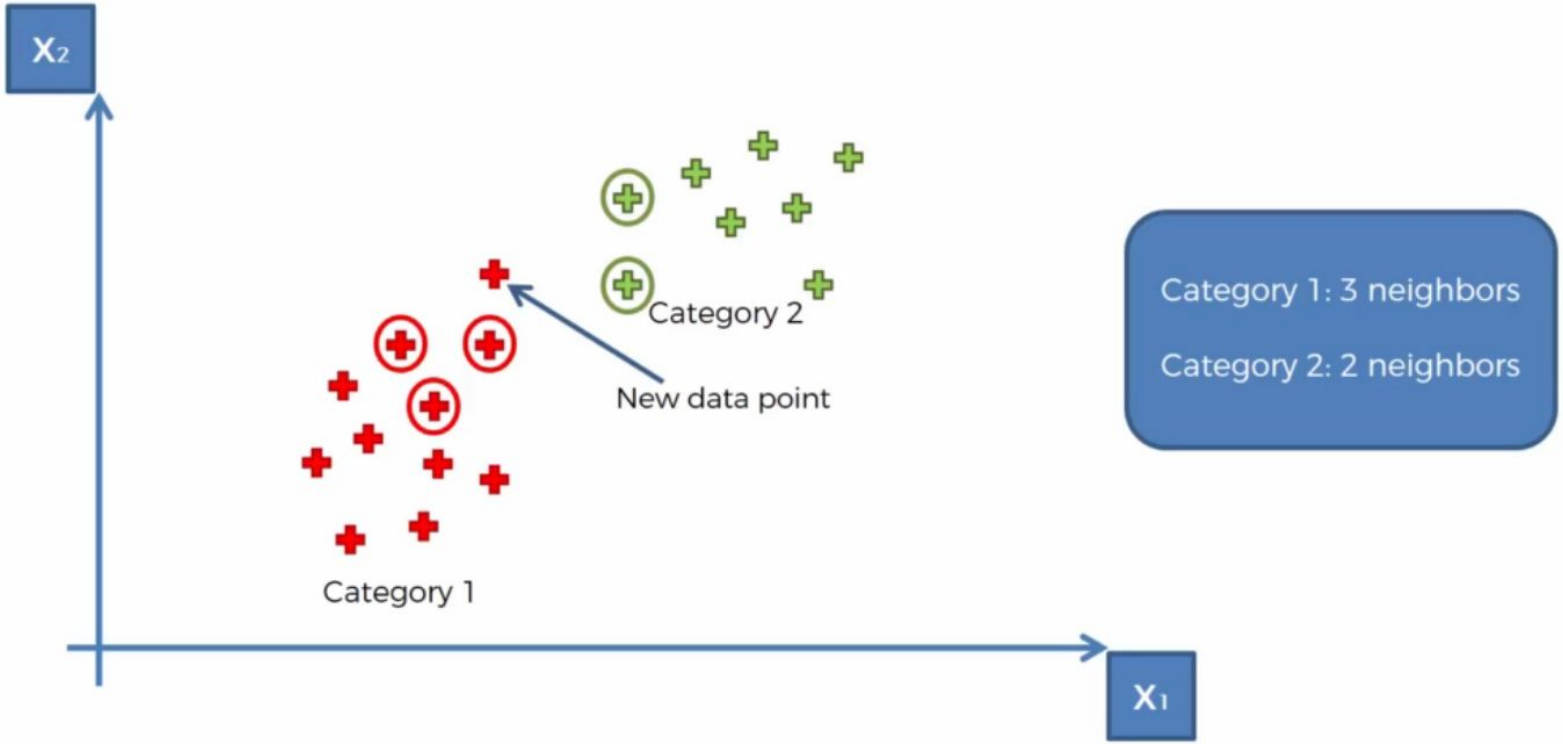
$$\text{Euclidean Distance between } P_1 \text{ and } P_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

STEP 2: Take the K = 5 nearest neighbors of the new data point, according to the Euclidean distance

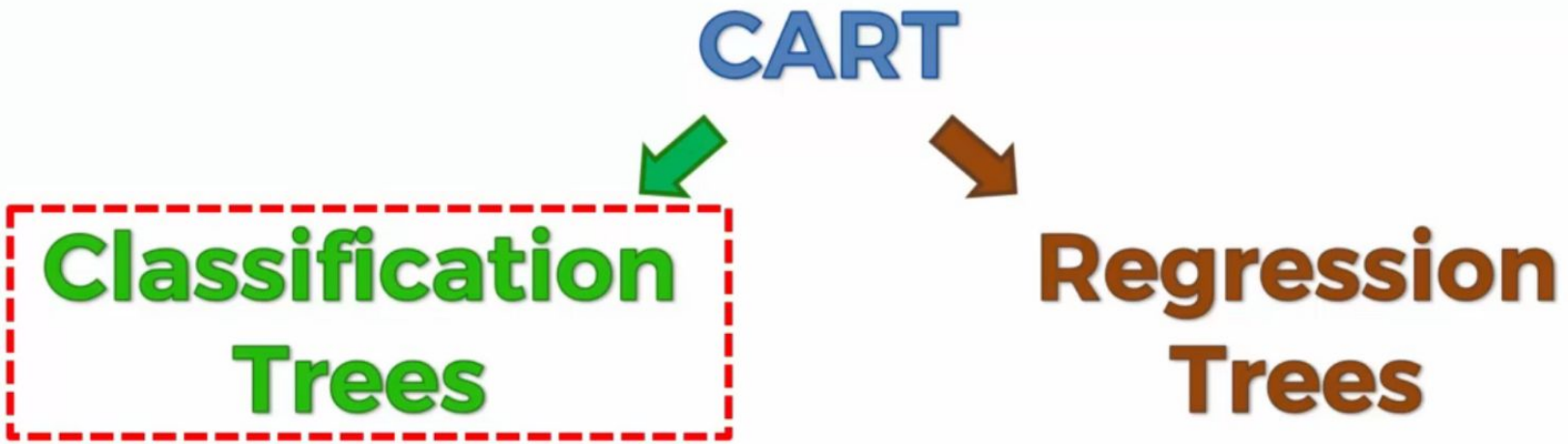
STEP 3: Among these K neighbors, count the number of data points in each category



STEP 4: Assign the new data point to the category where you counted the most neighbors



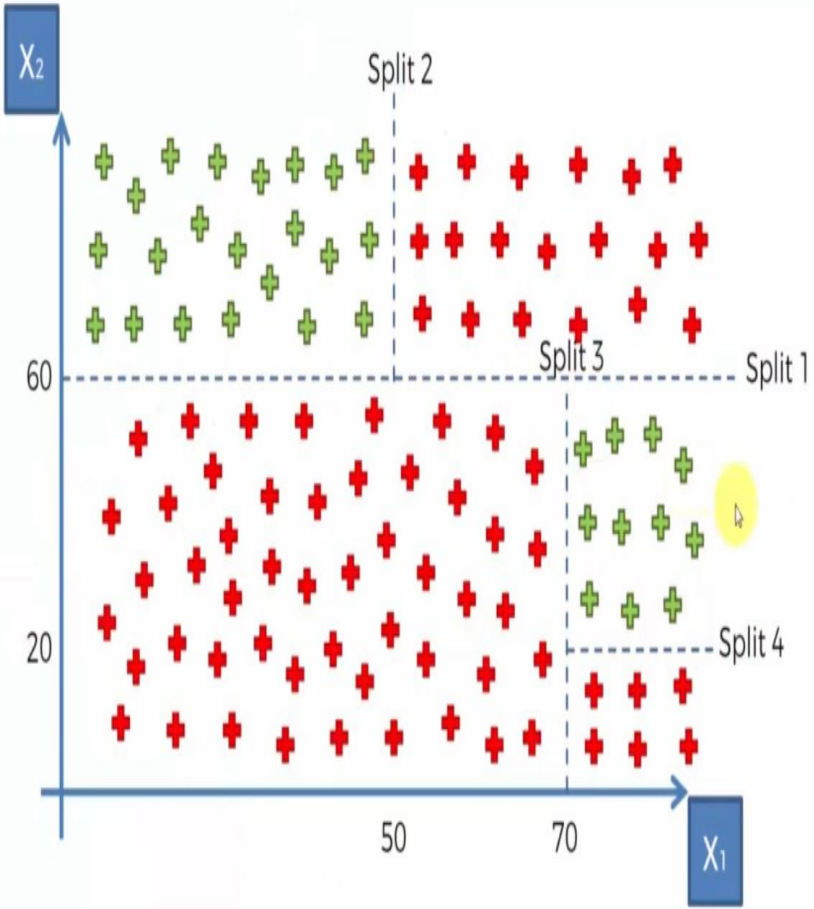
Decision Tree

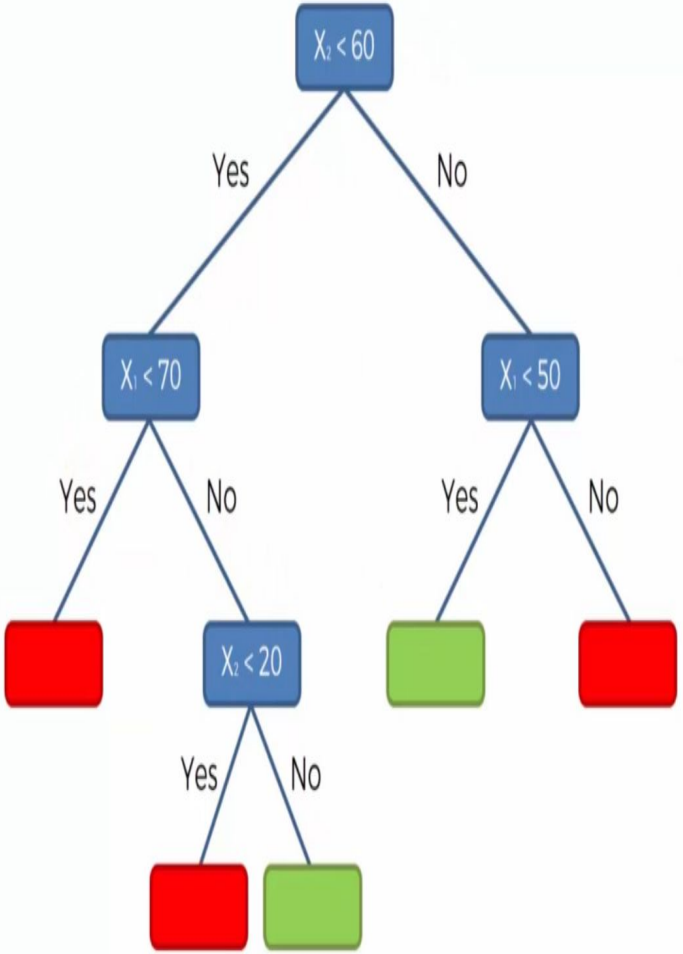


Help In Classification in categories like M/F etc.

Help In Prediction of a value such as a salary etc.

Splits are reducing entropy, entropy is measure of uncertainty or disorder (Intuitive Understanding)





- Old Method
- Reborn with upgrades
- Random Forest
- Gradient Boosting
- etc.

Random Forest

Ensemble Learning

Put together Multiple Machine Learning Algorithms to create one bigger Machine Learning Algorithm

Random Forest basically combines lot of Decision Tree Methods

STEP 1: Pick at random K data points from the Training set.

STEP 2: Build the Decision Tree associated to these K data points.

STEP 3: Choose the number Ntree of trees you want to build and repeat STEPS 1 & 2

STEP 4: For a new data point, make each one of your Ntree trees predict the category to which the data points belongs, and assign the new data point to the category that wins the majority vote.

Power of Trees,

Taking Majority Vote

https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/BodyPartRecognition.pdf

Random Forest used in Microsoft Kinect to understand where body parts are