Demystifying Machine Learning

Uncloaking the Math in the Black Box

Agenda

What's in an Algorithm?

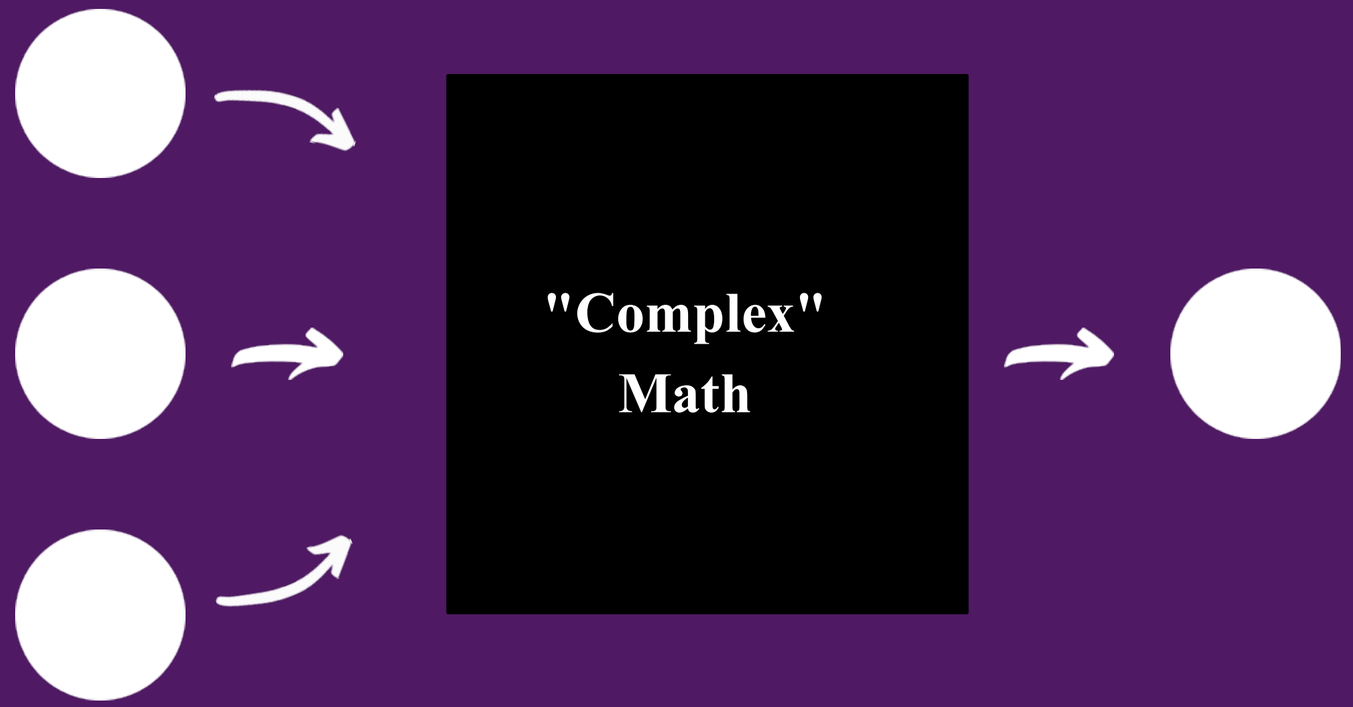
Machine Learning Applications

Math for Machine Learning

Types of ML

WHAT'S IN THE BLACK BOX?

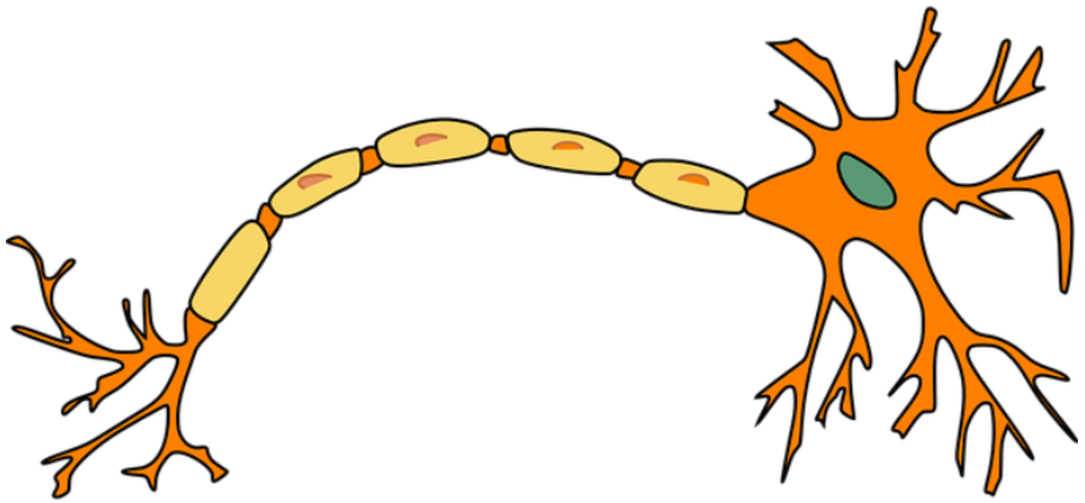
WHAT DOES AN ALGORITHM DO?



Input:
An image
A vector of numbers

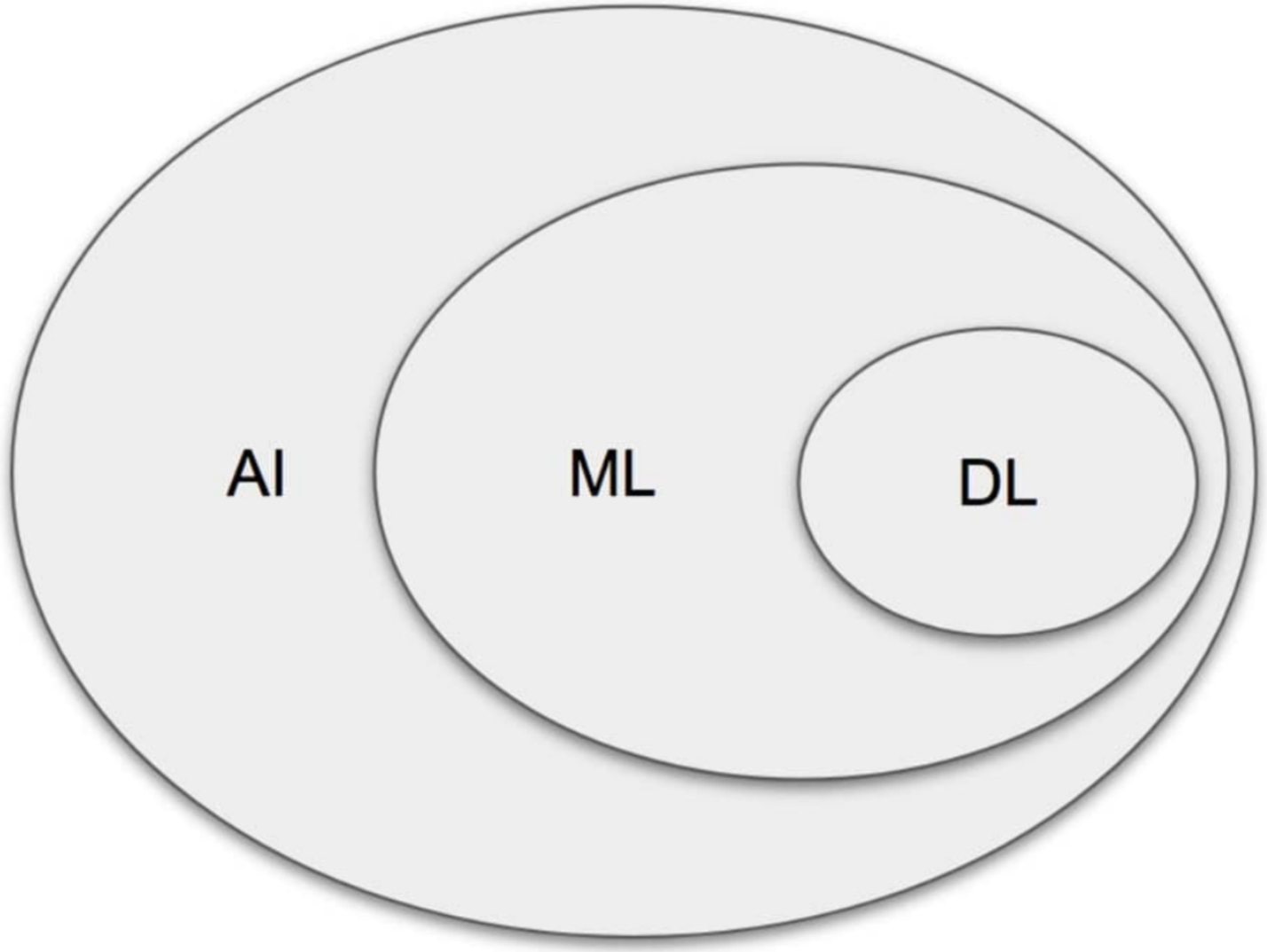
Ouput:
Cat or dog
0 or 1
Chance of rain

ML takes cues from neuroscience



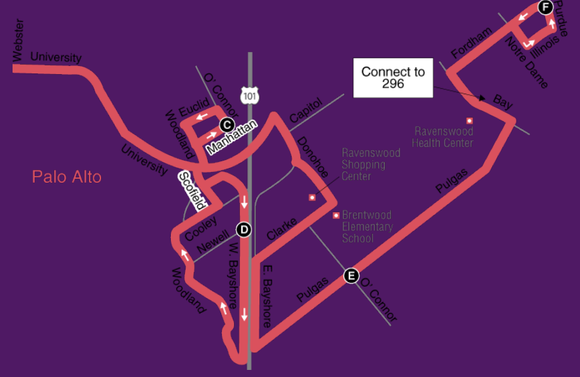
- Training
  - Show a program lots of data
  - Teach it to recognize patterns
  - Check if it learned well

What’s the landscape?



MACHINE LEARNING IN THE WILD

ROUTE OPTIMIZATION



MOVIE RECCOMENDATIONS

VIDEO GAME ENEMIES

PERSONAL ASSISTANTS

WILD MACHINE LEARNING

RECIDIVISM

HR SCREENINGS

POLICING

CREDIT SCORING

Math Foundations

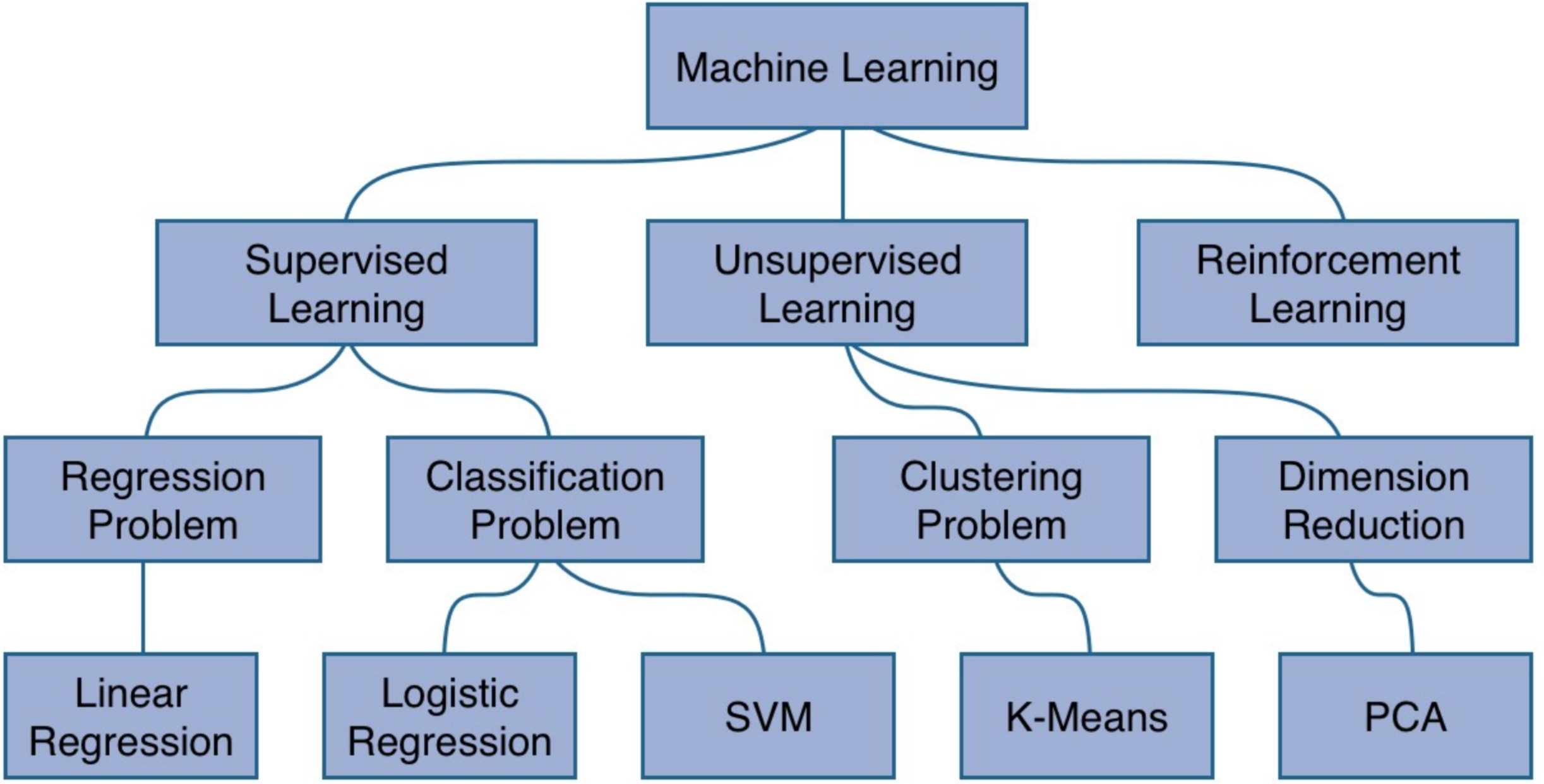
BIGGEST BARRIER TO ENTRY

WHAT THEY SAY YOU NEED

- Linear Algebra
- Calculus
- Statistics
- Discrete Math

WHAT'S USED FREQUENTLY

- SOME Linear Algebra
- Probability
- Statistics


Machine Learning
Supervised Learning
Unsupervised Learning
Reinforcement Learning
Regression Problem
Classification Problem
Clustering Problem
Dimension Reduction
Linear Regression
Logistic Regression
SVM
K-Means
PCA

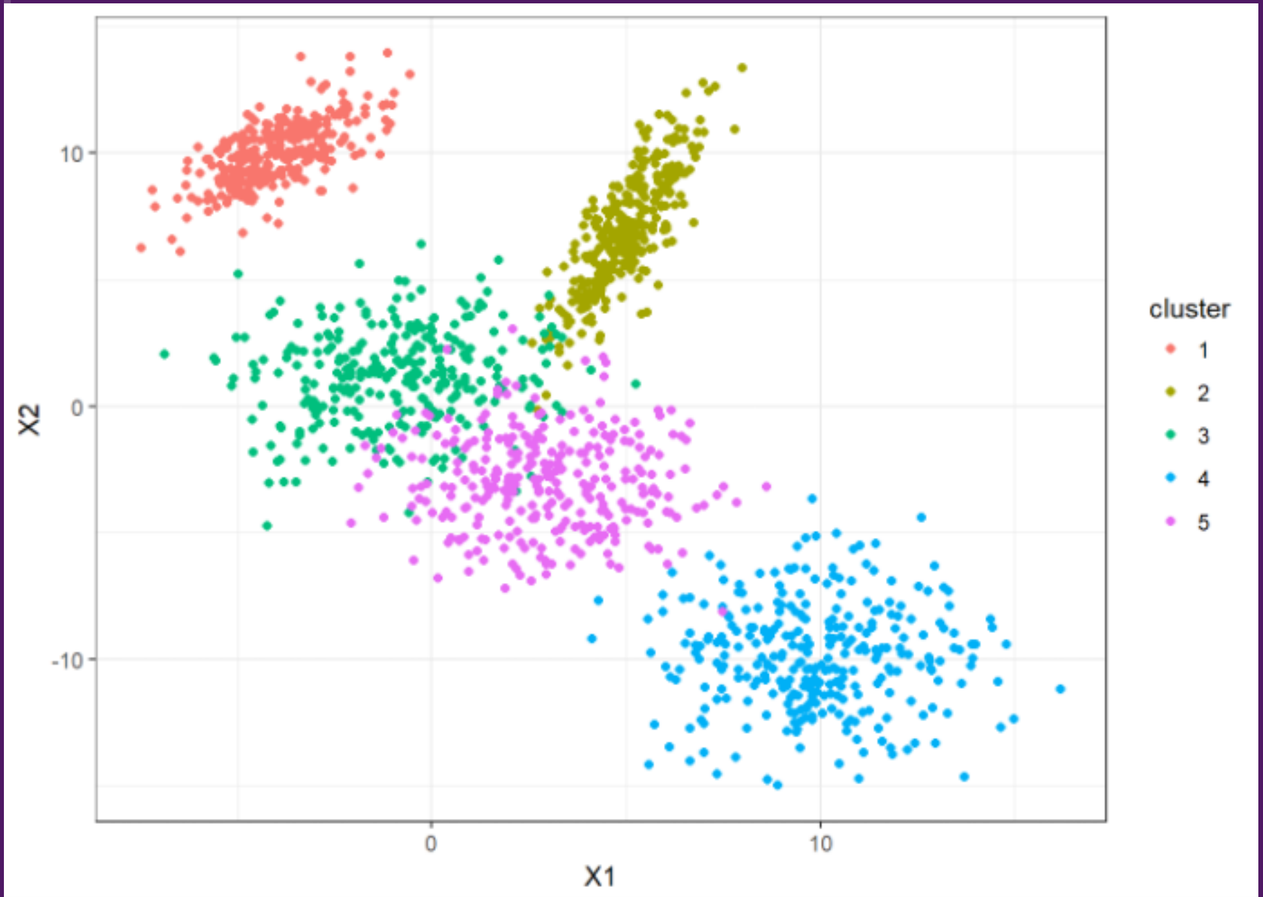
	X	Y	Z	Amplitude	Clip-Present
0	-0.871557	0.868241	9.924039	0.000458	1
1	0.505828	0.870442	9.949194	0.000552	1
2	-0.871557	0.868241	9.924039	0.000358	1
3	-0.871557	0.868241	9.924039	0.000729	1
4	-0.871557	0.868241	9.924039	0.005013	1
5	-0.046967	0.871548	9.961837	0.013645	1
6	0.871557	0.868241	9.924039	0.011765	1
7	0.871557	0.382027	9.954619	0.009366	1
8	0.871557	0.868241	9.924039	0.010479	1
9	0.871557	-0.316847	9.956907	0.013448	1
10	0.603430	-0.301062	9.977236	0.014846	1
11	0.555697	-0.295141	9.980185	0.015184	1
12	0.871557	-0.157112	9.960708	0.017696	1
13	0.871557	0.473205	9.950702	0.016759	1
14	0.698513	0.869429	9.937614	0.012372	1
15	-0.871557	-0.868241	9.924039	0.010986	1

SUPERVISED

LABELED OR NAH?

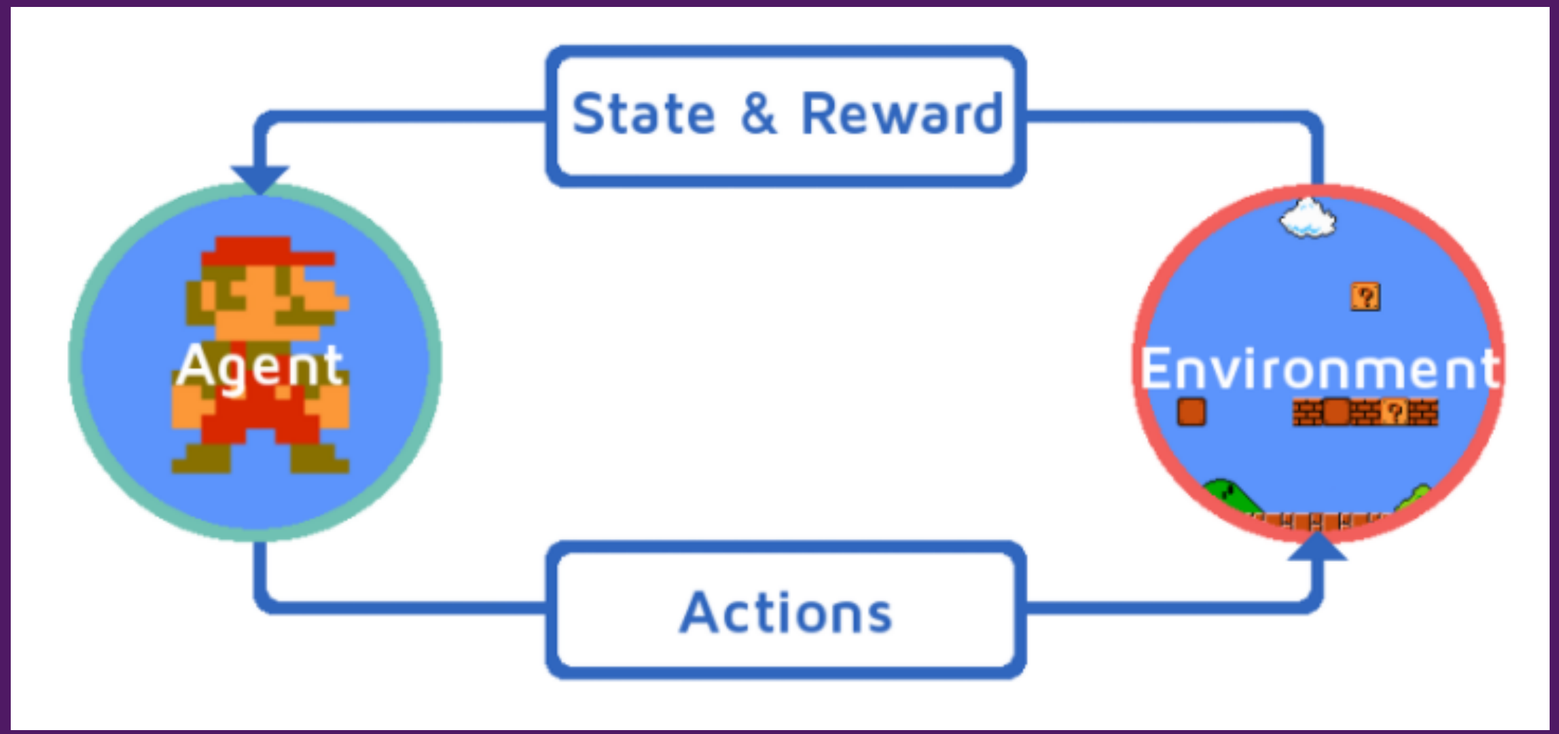
UNSUPERVISED

CREATES SIMILAR GROUPINGS



REINFORCEMENT

WHAT MANY CONSIDER AI



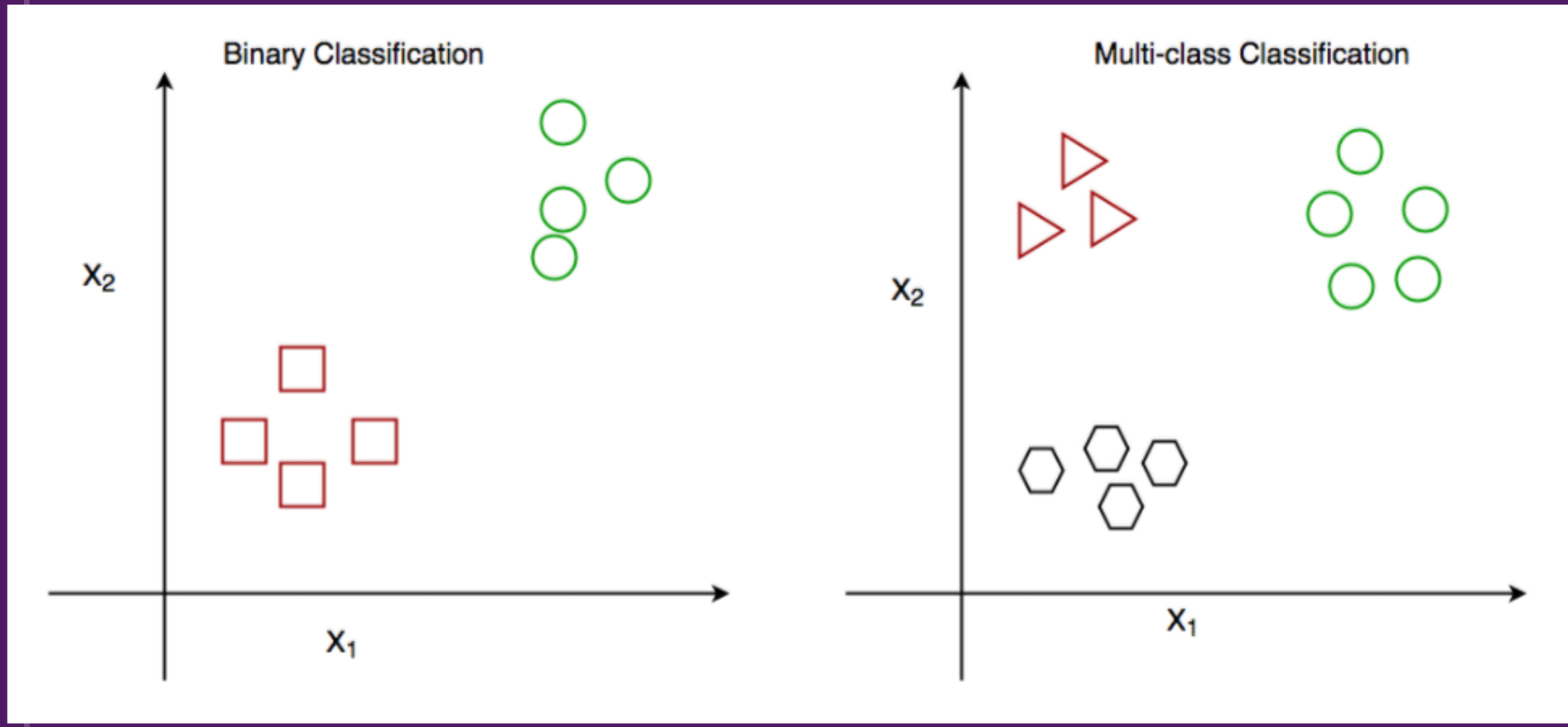
What can machines predict?

REGRESSION

REAL/CONTINUOUS NUMBERS

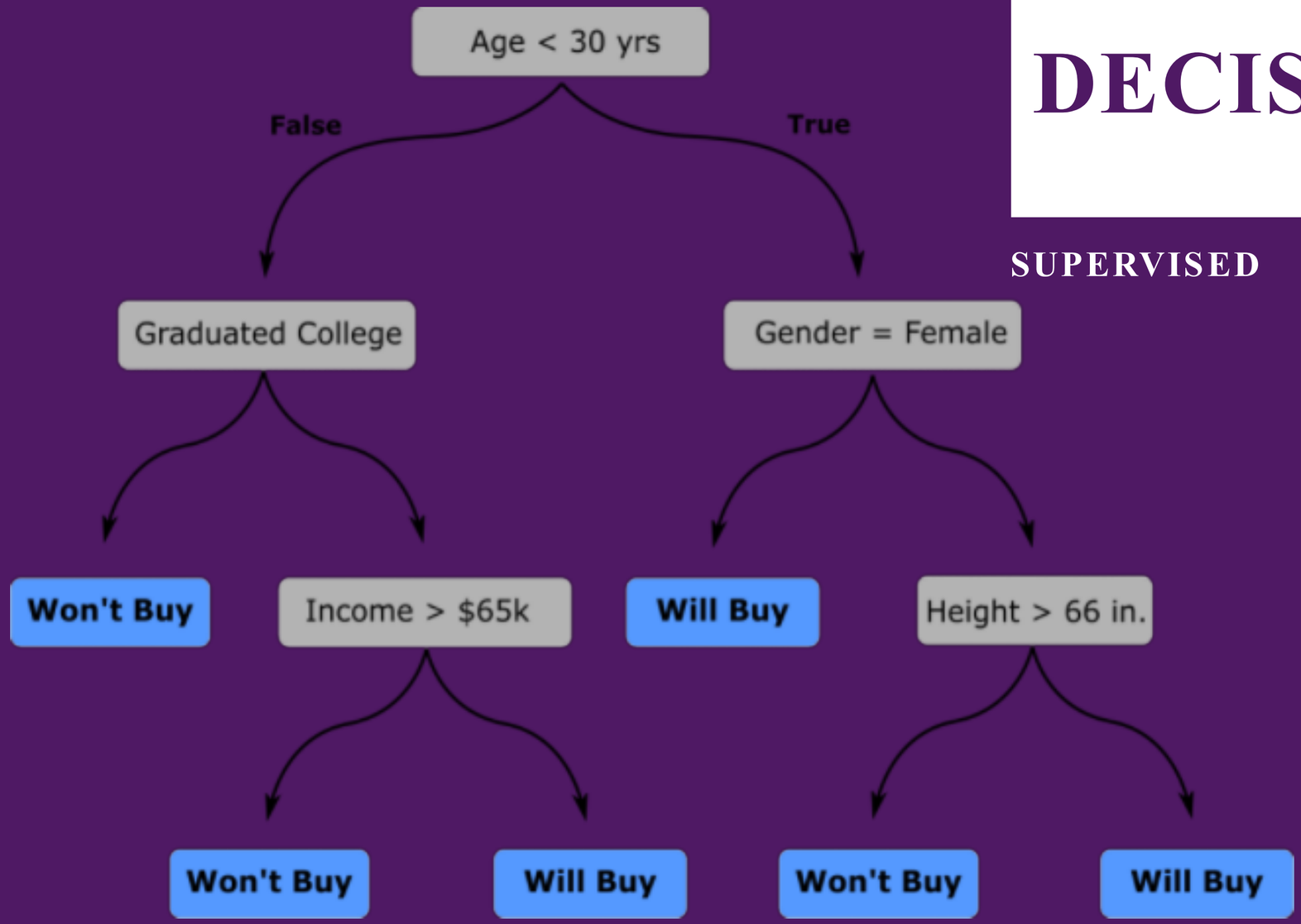
CLASSIFICATION

DISCRETE/CATEGORICAL VARIABLE



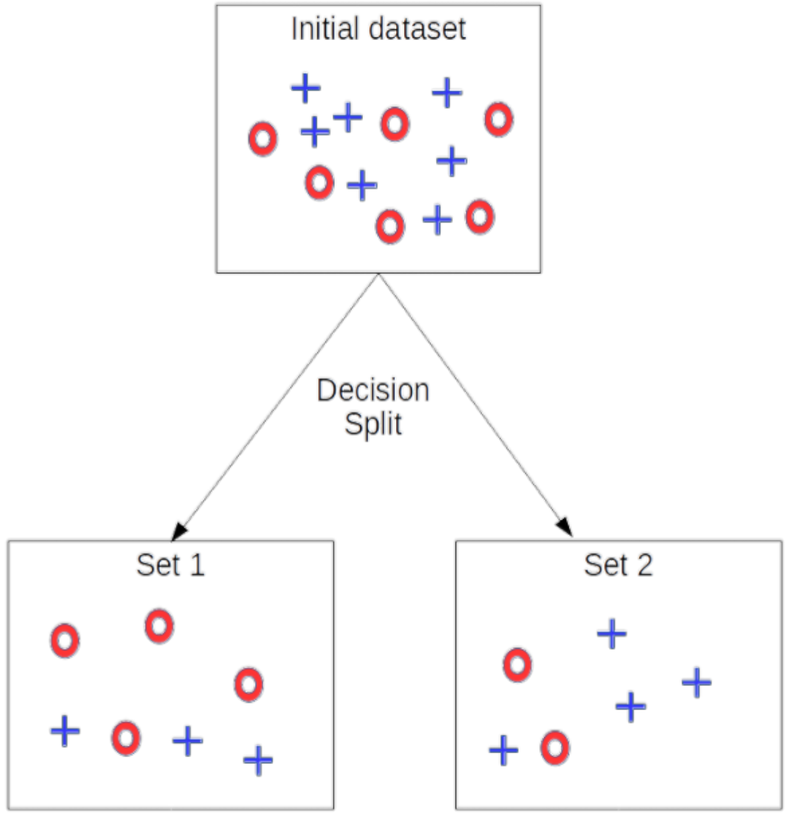
DECISION TREES

SUPERVISED



Entropy

- A measure of the degree of randomness in a variable
- "Good" Decision Trees have homogenous leaf nodes
- The higher the entropy, the harder to draw conclusions



Information Gain

- Used to decide which of the attributes are most relevant
- The purpose is to find the attribute that returns the most information gain
- Expected information gain = decrease in entropy
- The less random the variables, the more information is gained

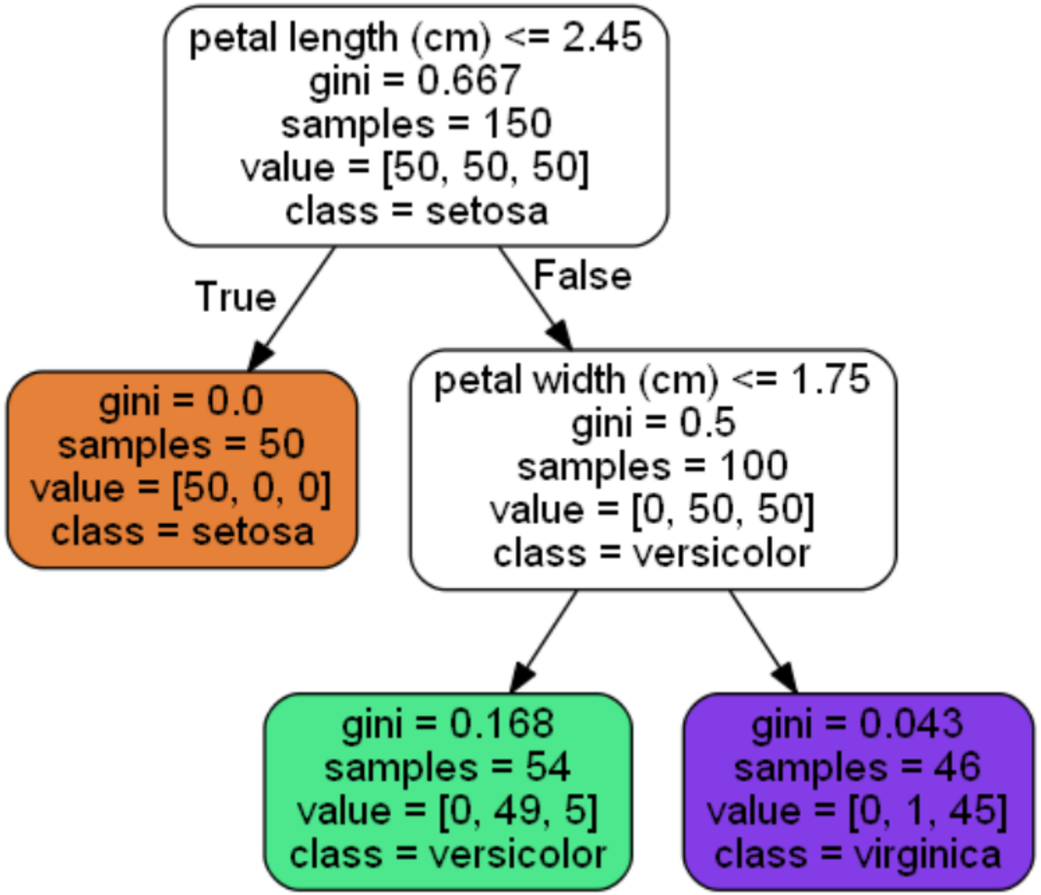
Gini Index

- Measures how impure a node is
- Calculated per node

Gini index is used in CART

(Classification and Regression Trees)

IID3 search algorithm uses entropy and infromation gain



Pre-pruning

Involves setting the tree parameters before building it so it stops early without completely being built.

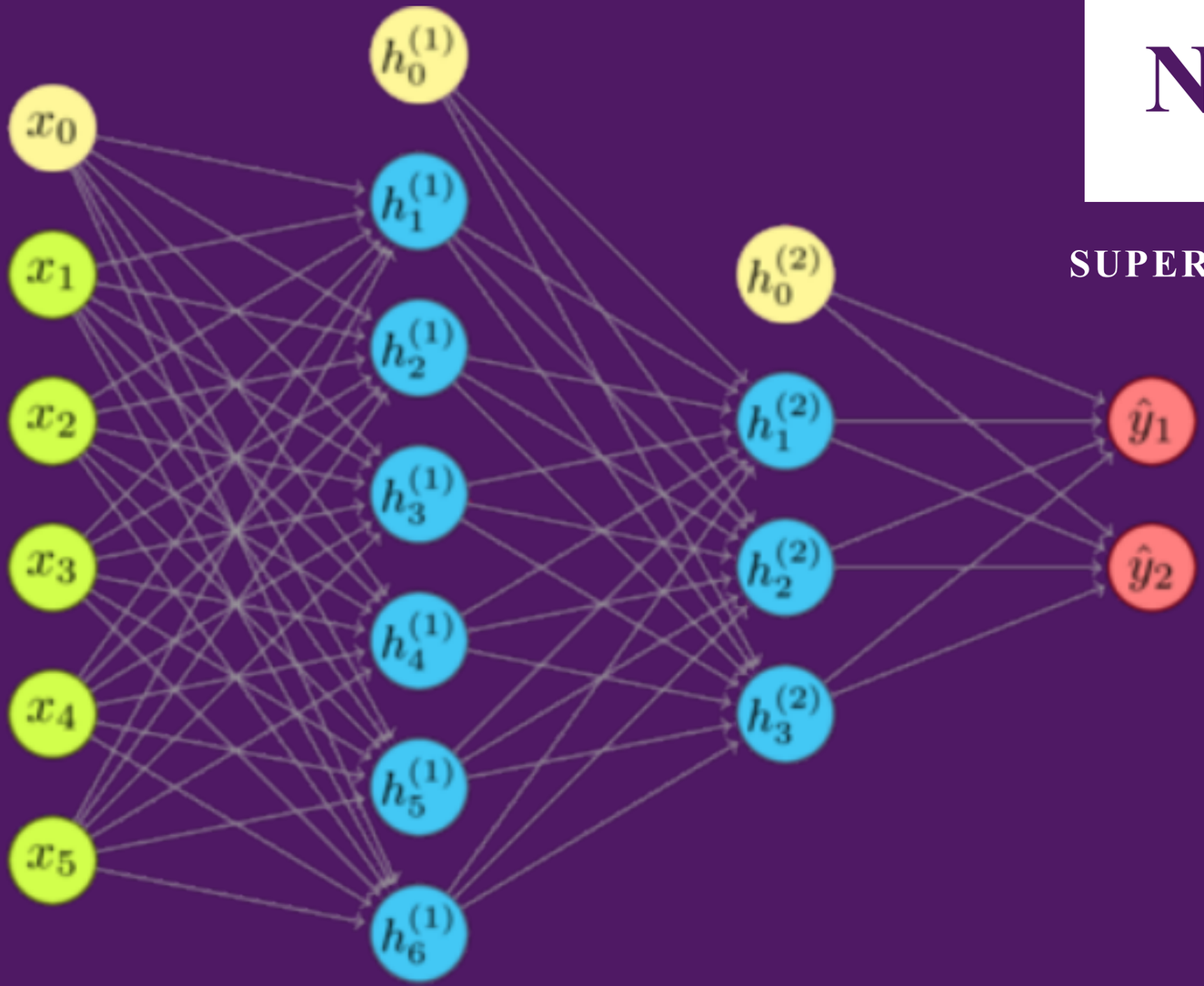
Variables to tune:

- Set max tree depth
- Set max terminal nodes
- Set max number of features
- Set max samples for a node split
  - controls the size of terminal nodes

Post-pruning

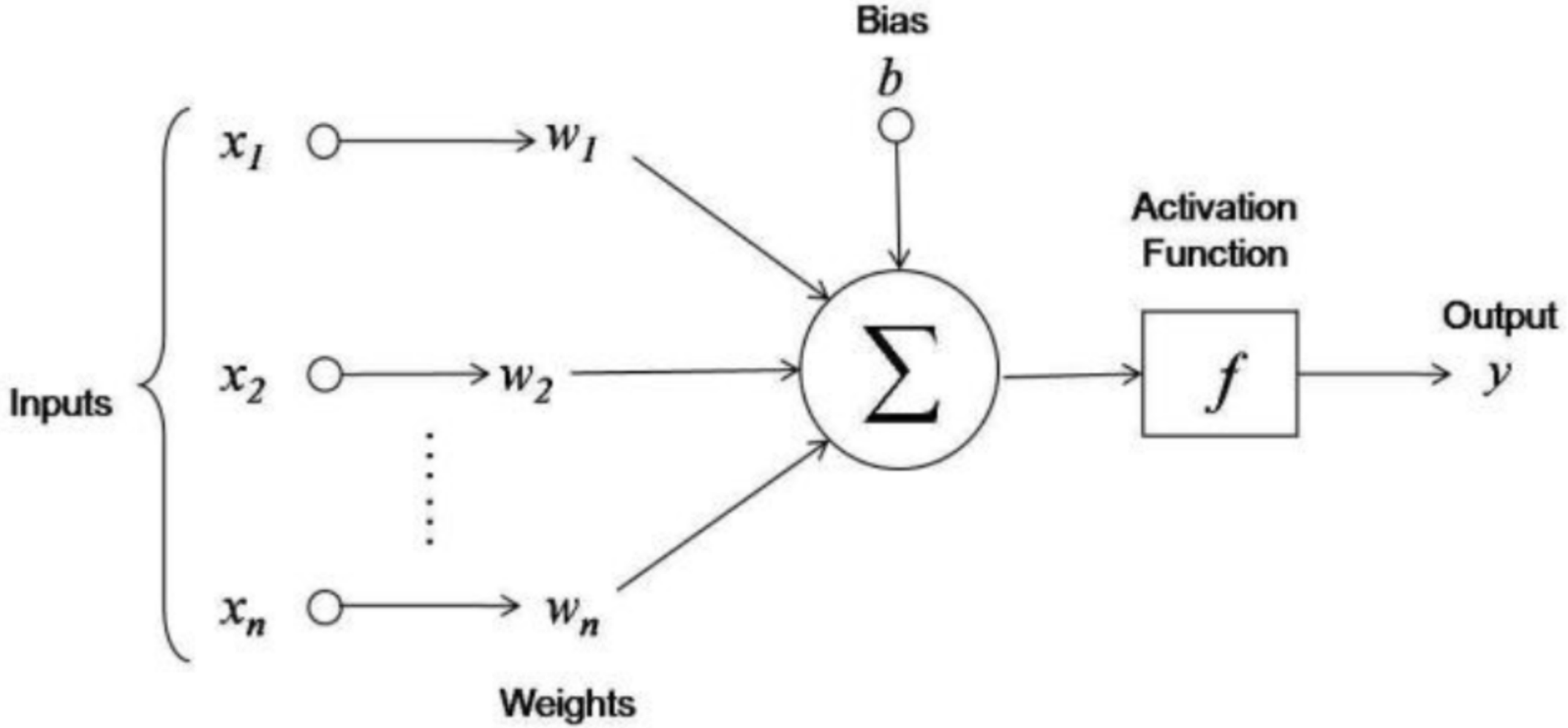
- Validate the performance of the model on a test
- Cut back splits that seem to overfit the noise in the training set
- Removes a branch from a fully grown tree

NEURAL NETWORKS

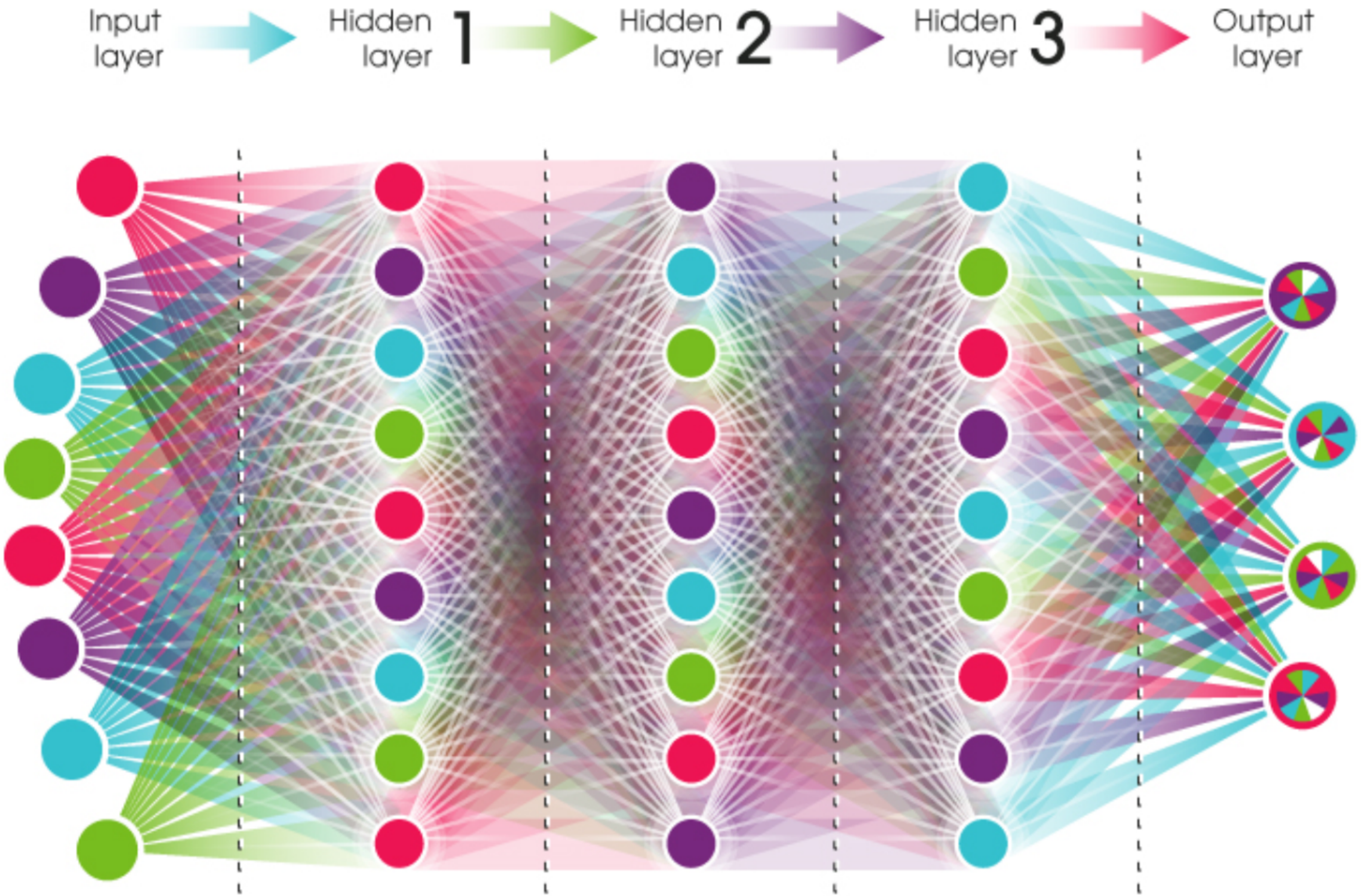


SUPERVISED

Neuron Structure

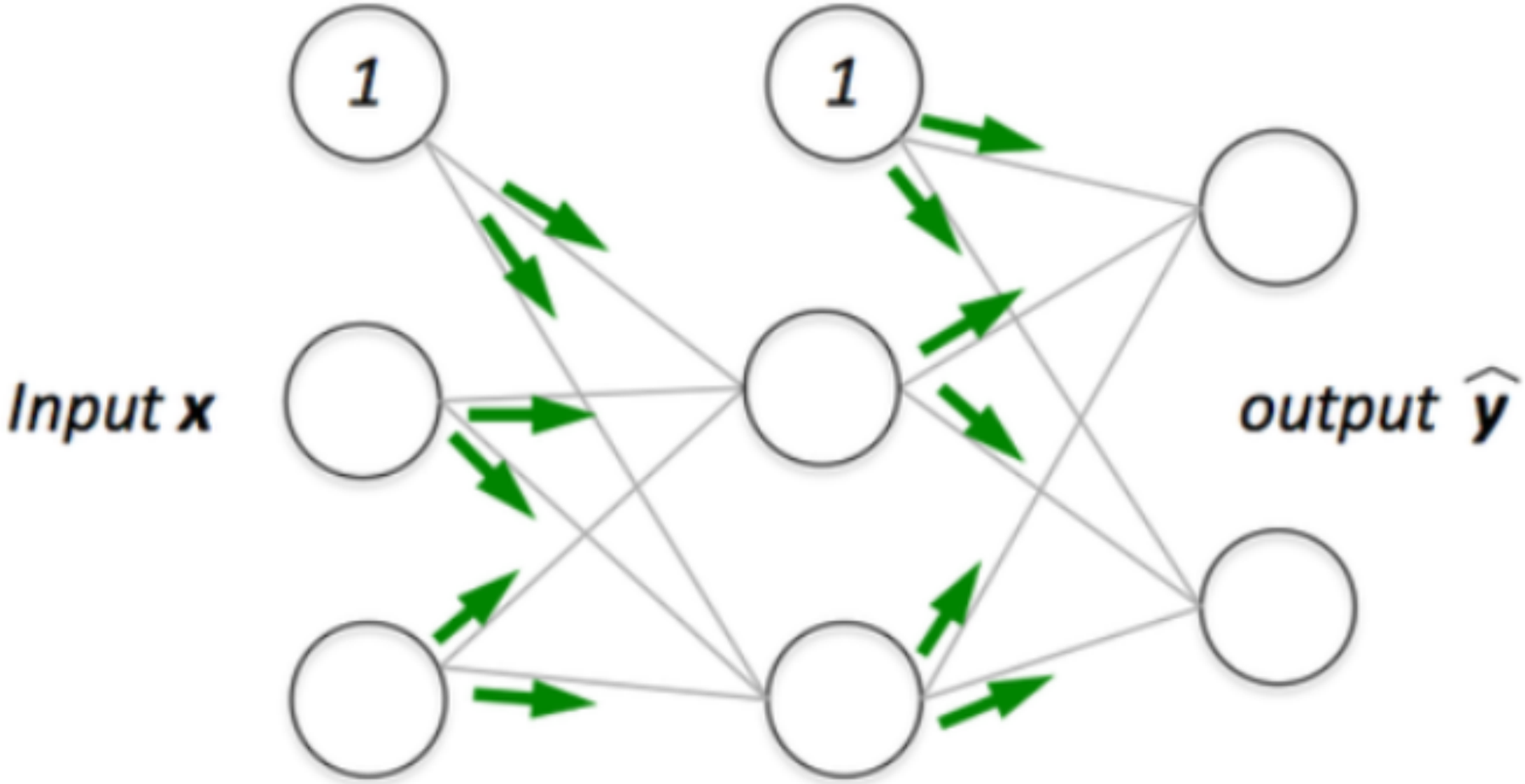


DEEP NEURAL NETWORK

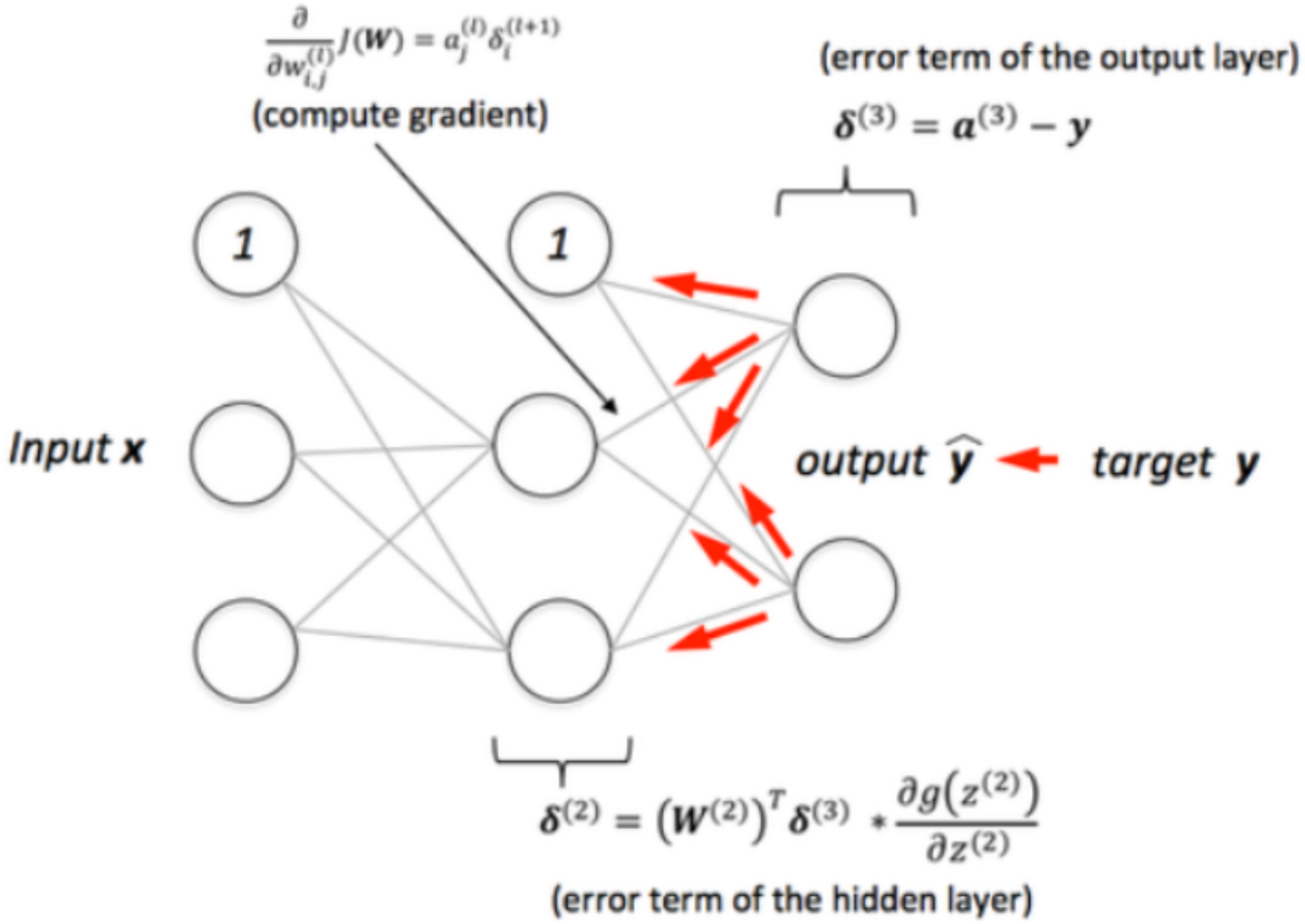


neuralnetworksanddeeplearning.com - Michael Nielsen, Yoshua Bengio, Ian Goodfellow, and Aaron Courville, 2016.

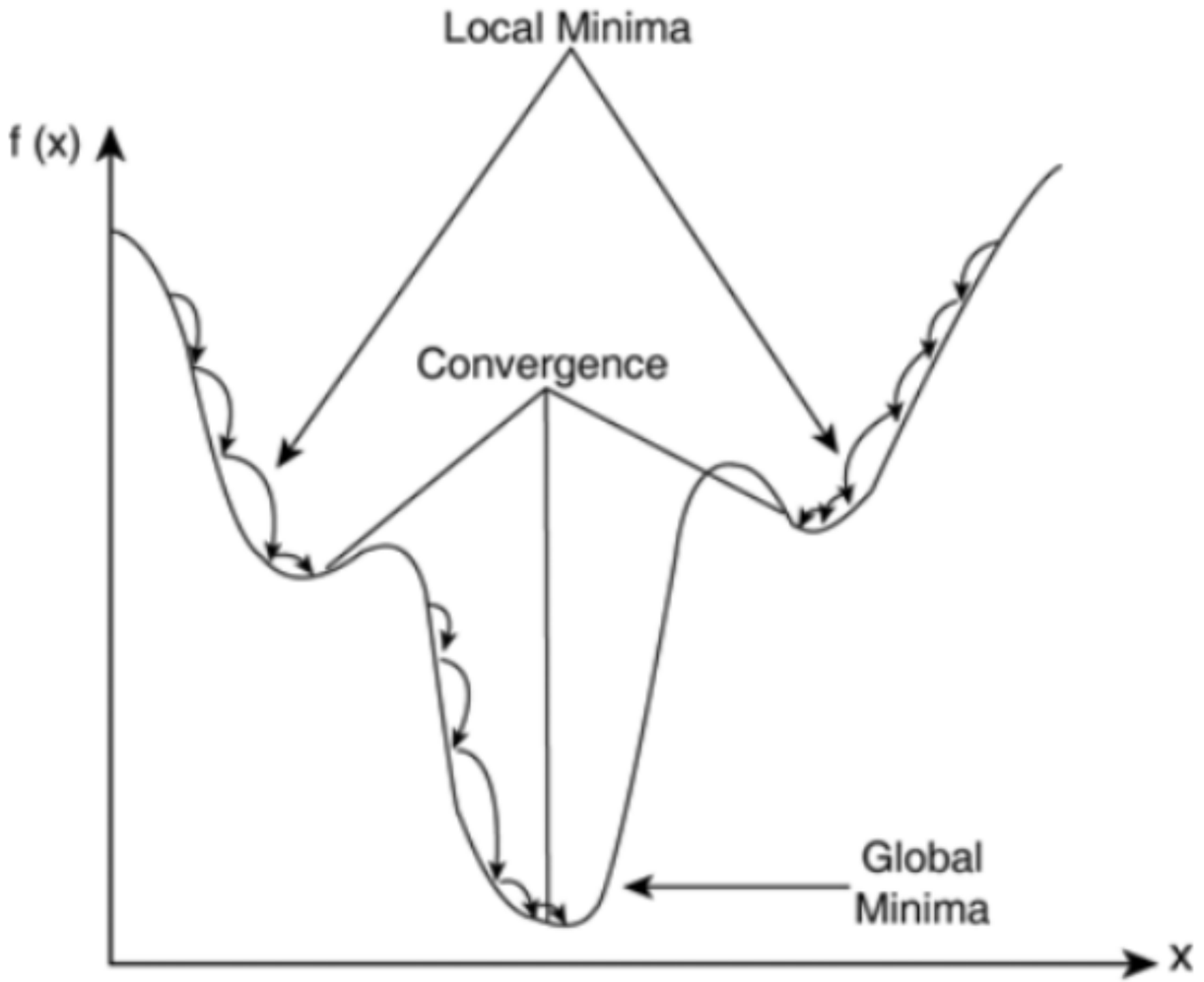
Forward Propogation



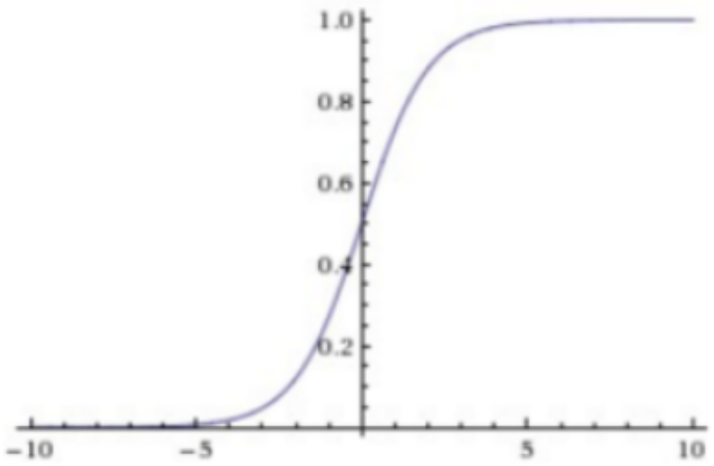
Back Propogation



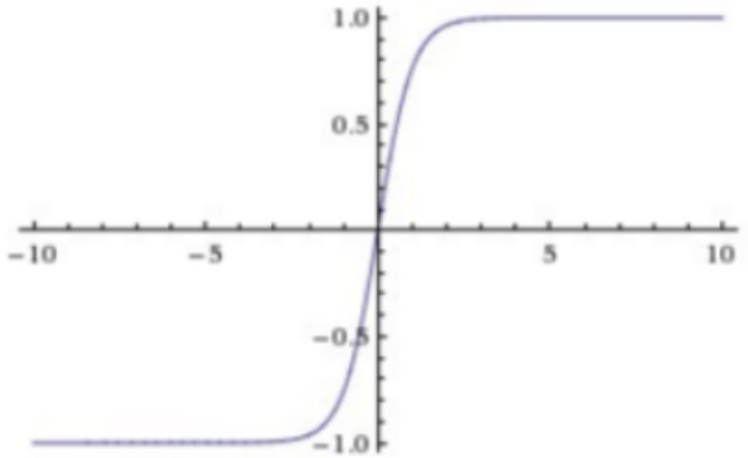
Learning Rate



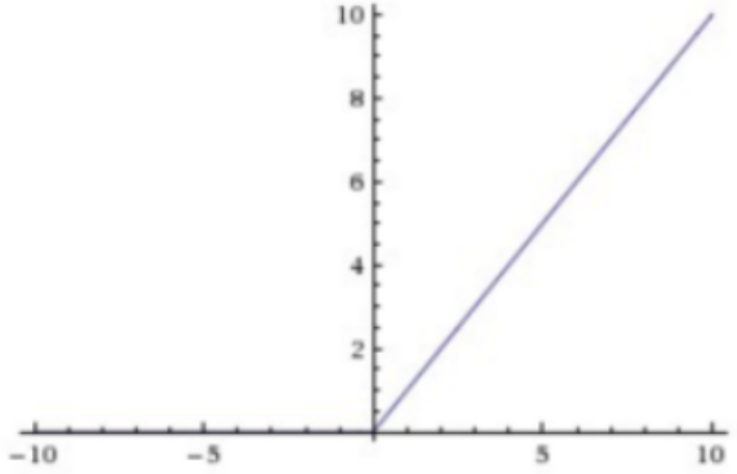
Activation Functions



Sigmoid

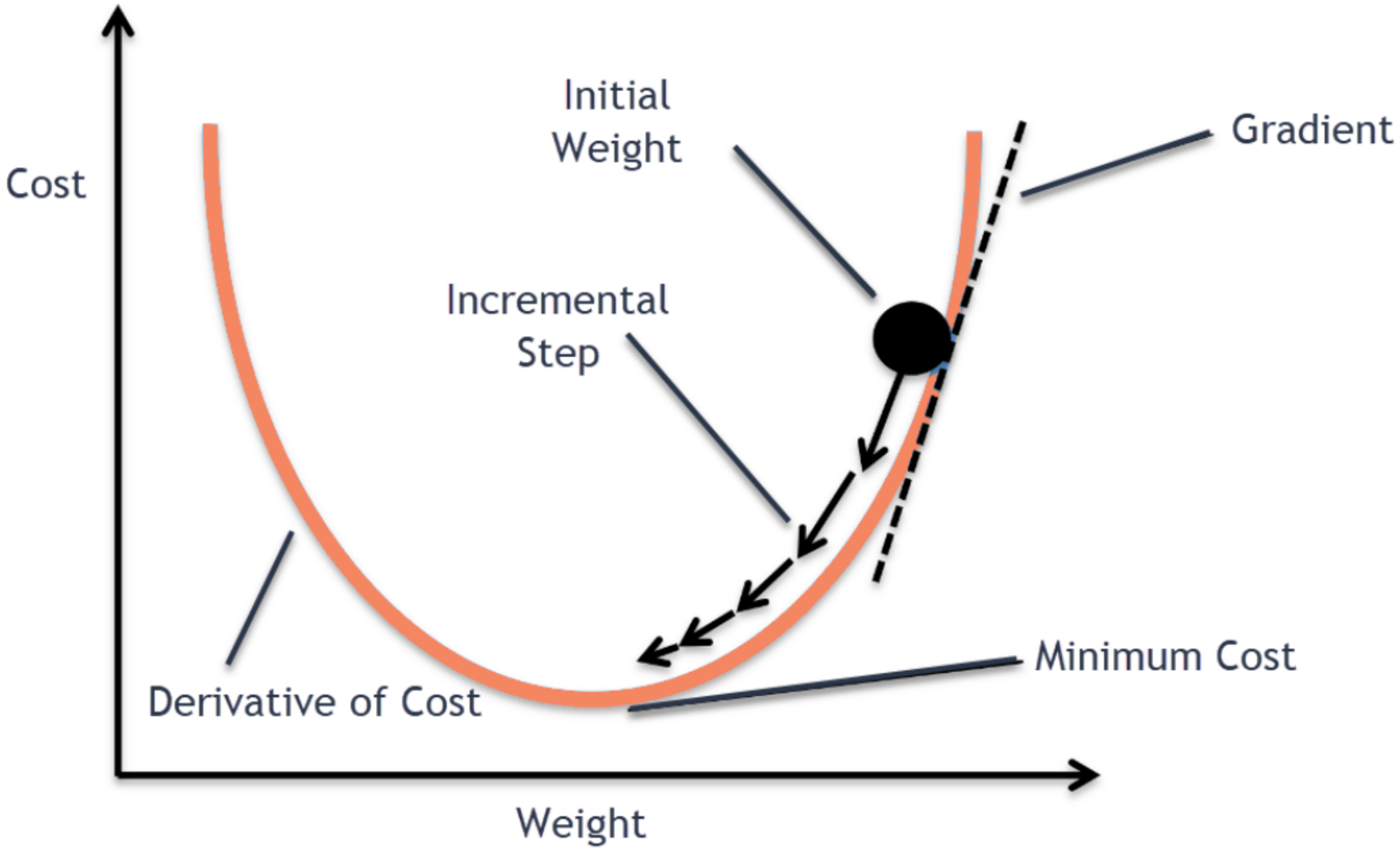


tanh



ReLU

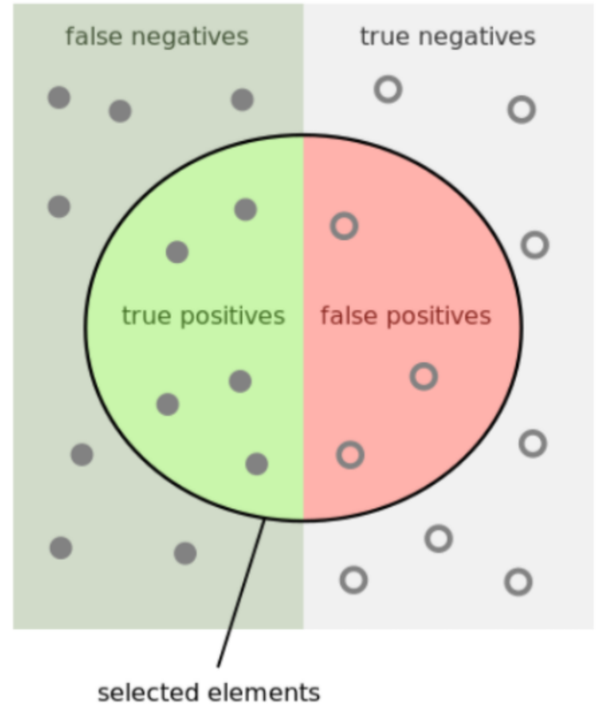
Gradient Descent

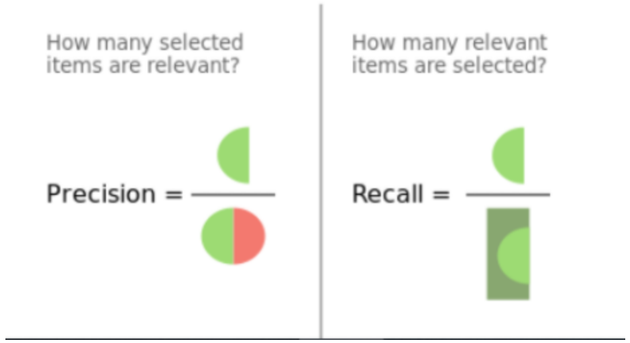


What Makes A Successful Model?

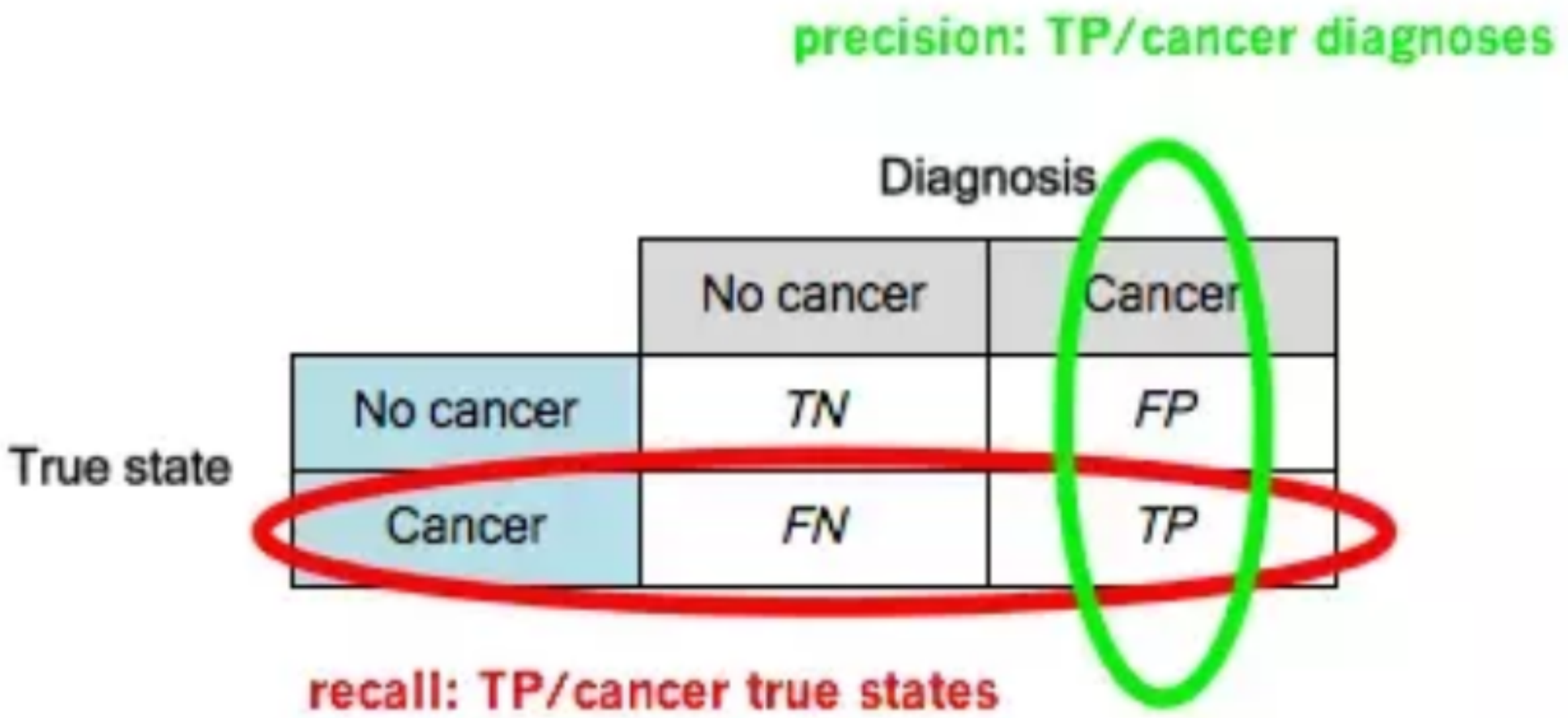
Evaluating Models

- **Classification Accuracy**
- **Confusion matrix**
- Logarithmic Loss
- Area under curve (AUC)
- F-Measure

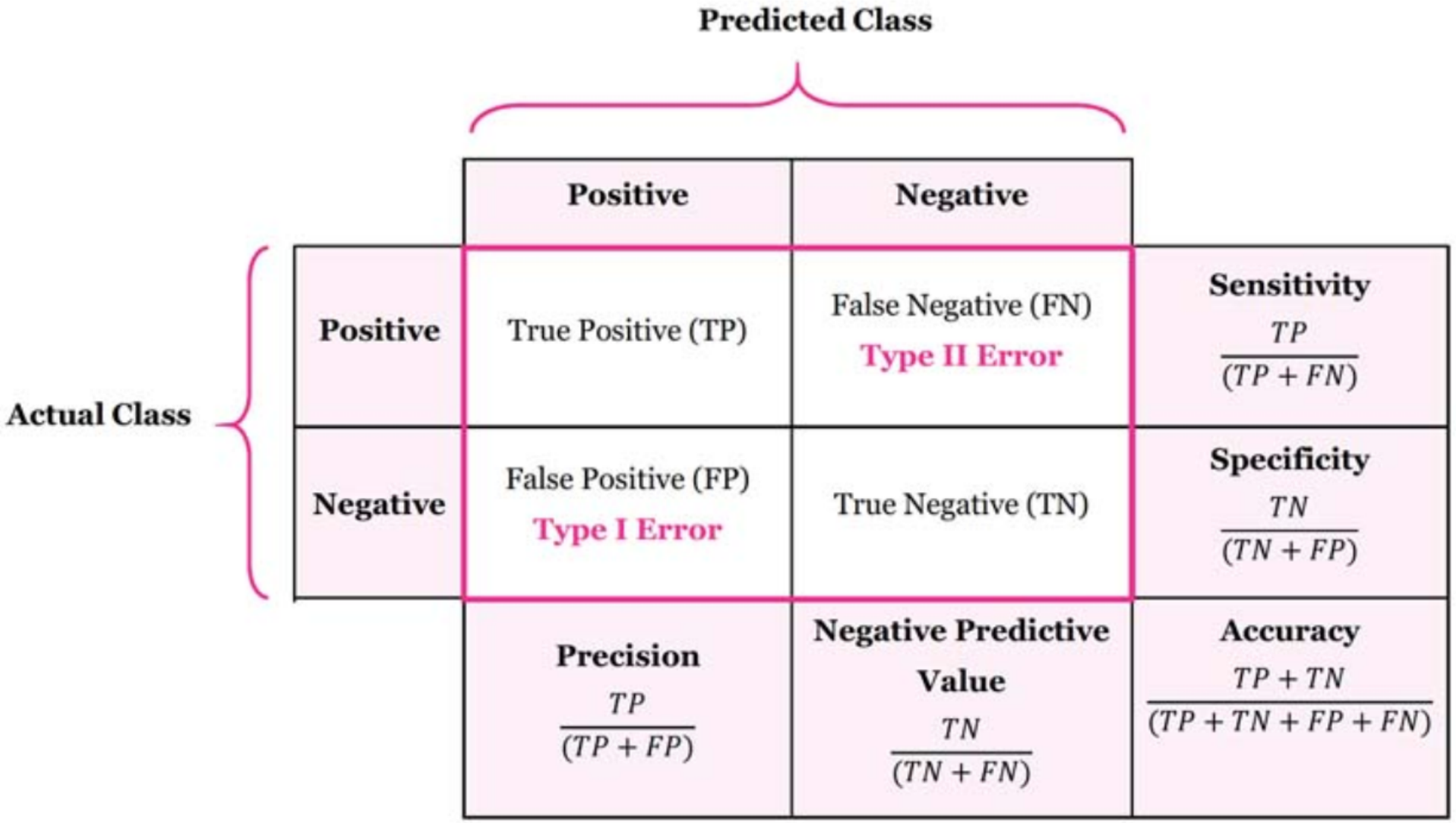




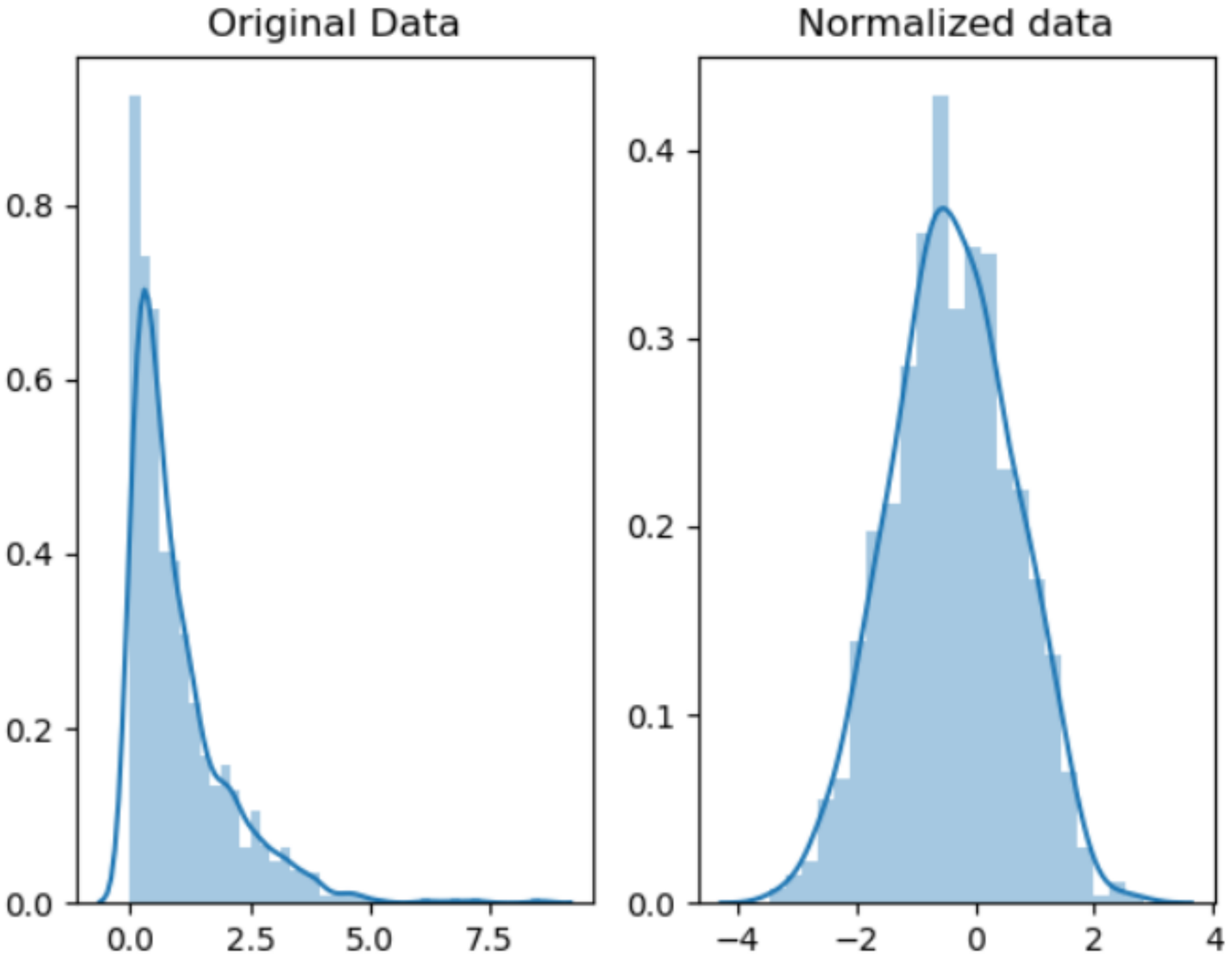
Classification Accuracy



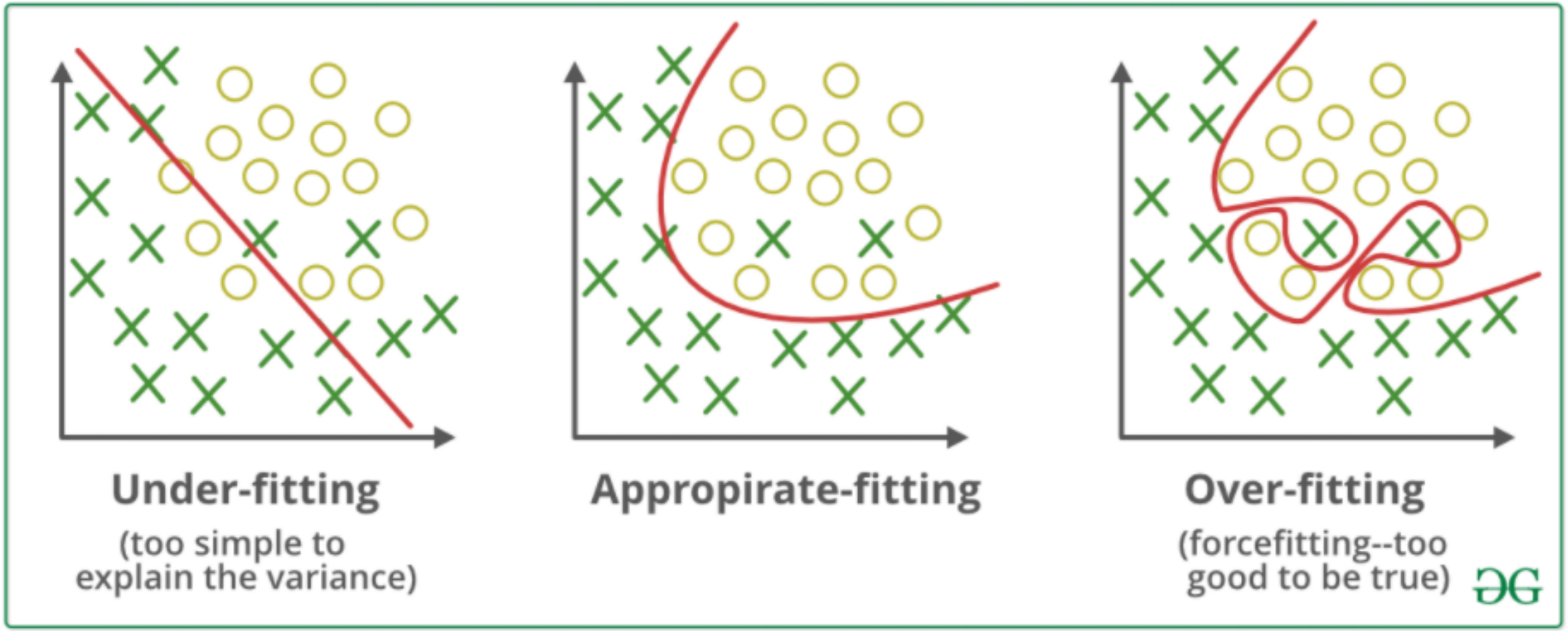
Confusion Matrices



Normalization



Overfitting



Regularizarion

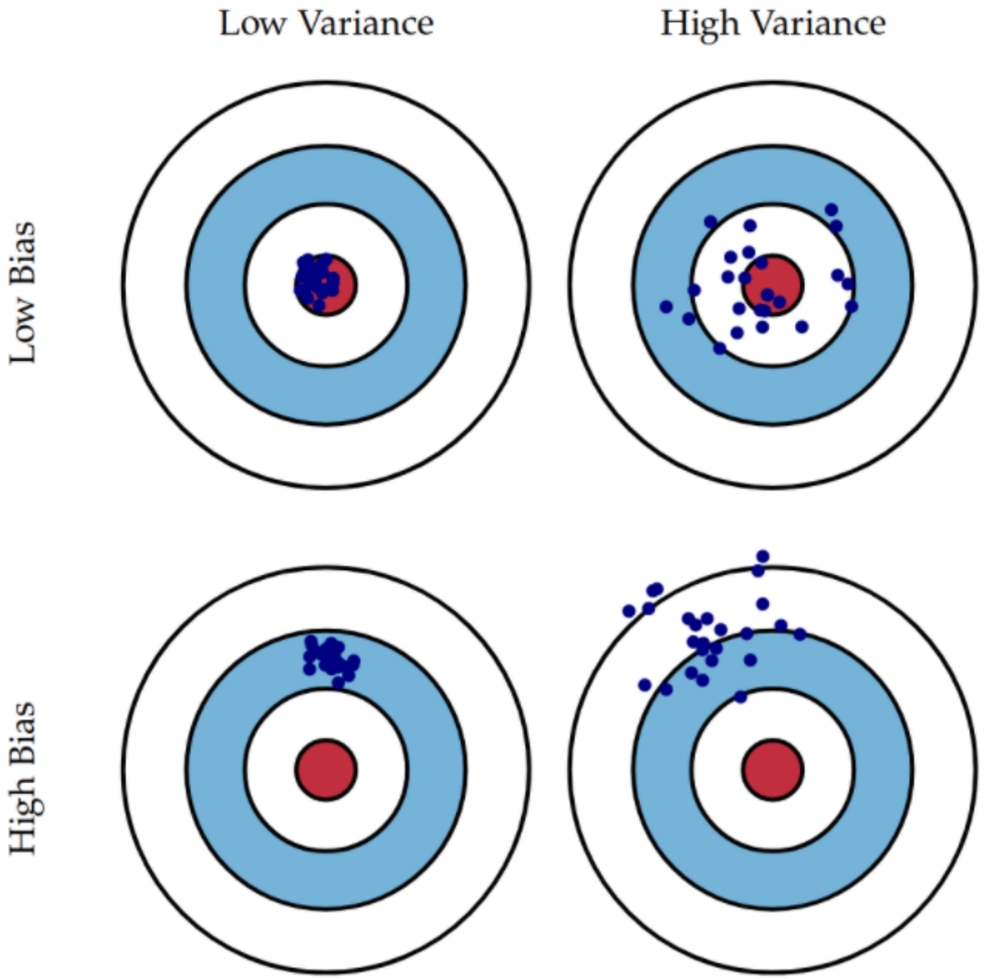
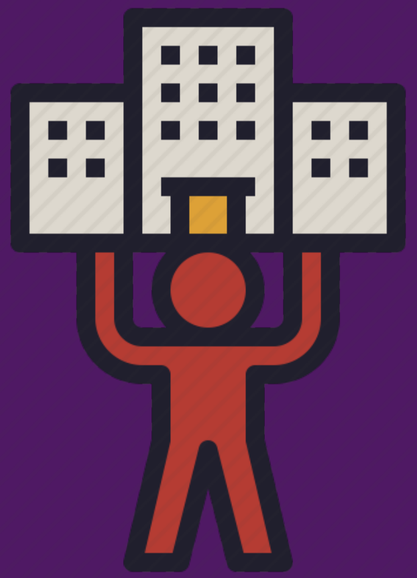


Fig. 1 Graphical illustration of bias and variance.

DATA IS VALUABLE

WHO COLLECTS IT

WHO OWNS IT



IF IT'S PROTECTED

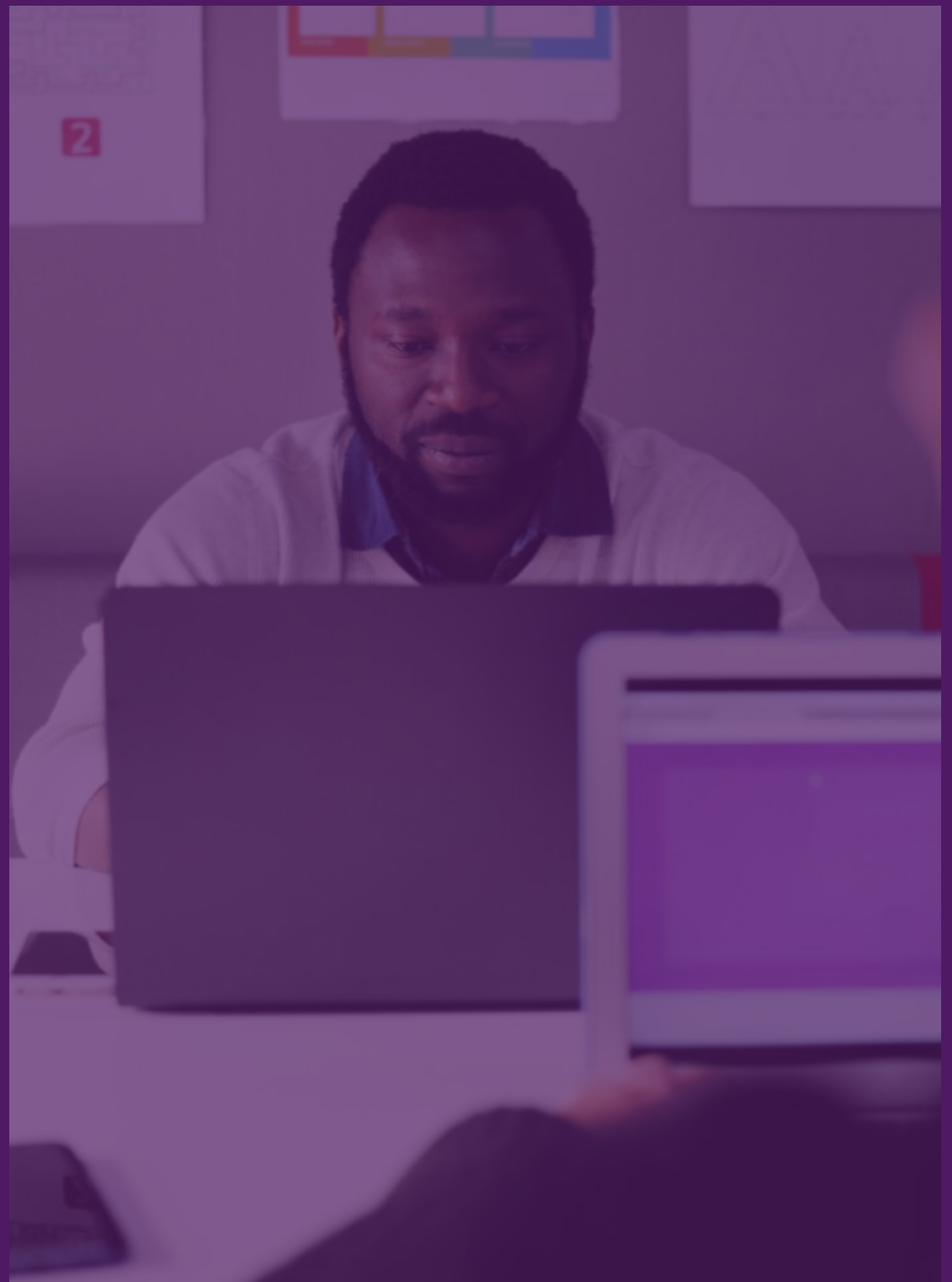
WHO'S USING IT

Critical Questions

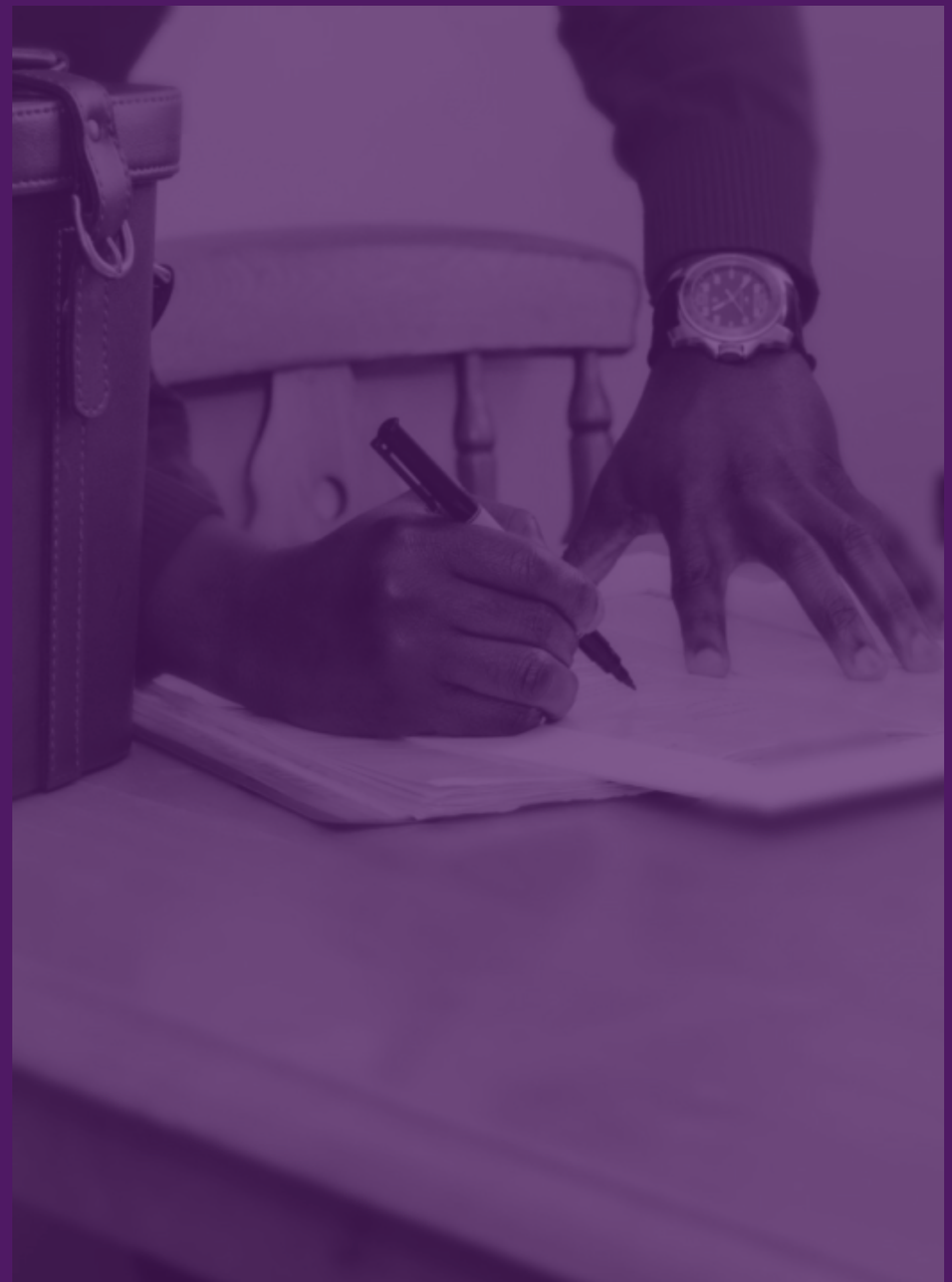
What does the prediction data look like?

If I don’t have the data, can I get it?

When will models make predictions?



What do we do about protected classes?



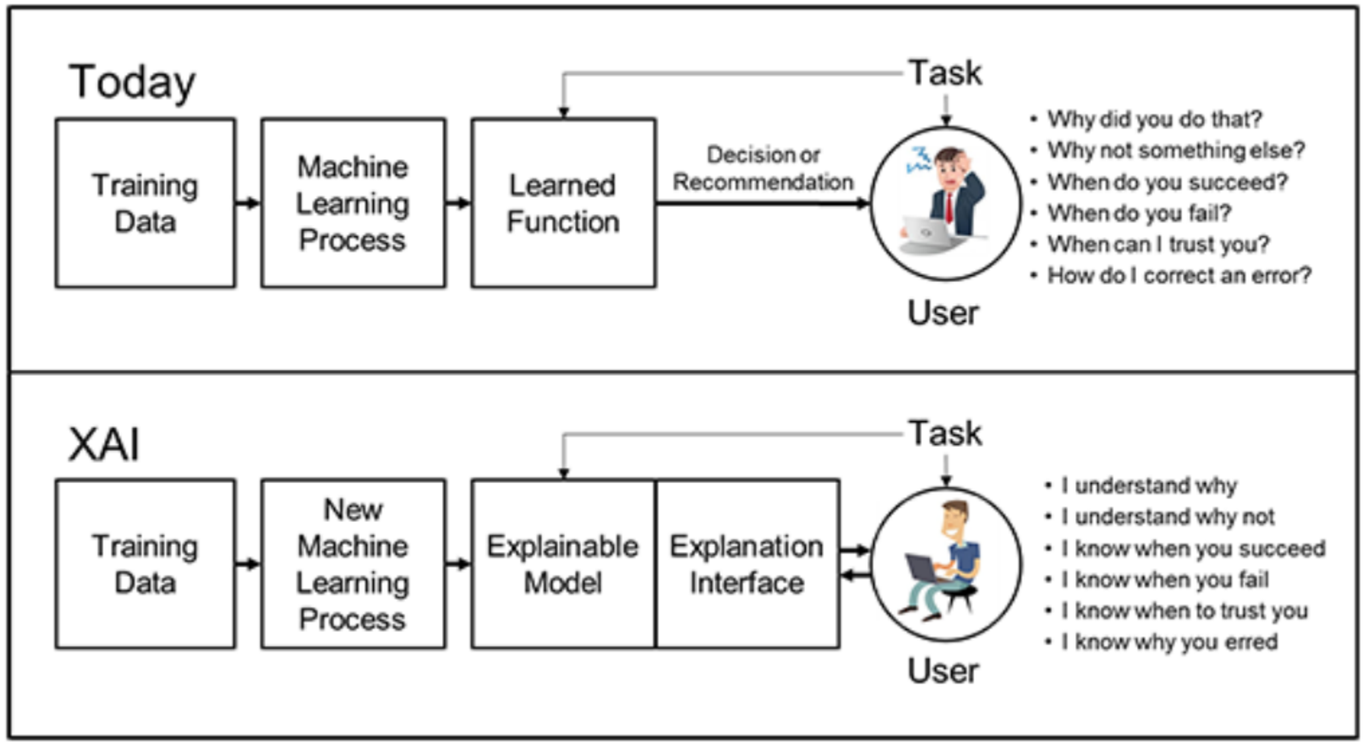
How often will this model get feedback?

PICKING A MODEL

NO FREE LUNCH

- What type of data are my outputs?
- What am I trying to predict
- What is wrong with the data?
- What data cleaning did I do?
  - Will this be a problem when running the model in the real world?

The Future: XAI



Thank you!

@DataSciBae

@DataSciBae

/ayodeleodubela