AI Cell's Workshop

ML CLASSIFICATION: SECRETS UNLOCKED

ELGARCH Youssef
TAYI Anouar

CONTENTS

1 What is Classification?

2 Meet the Models!

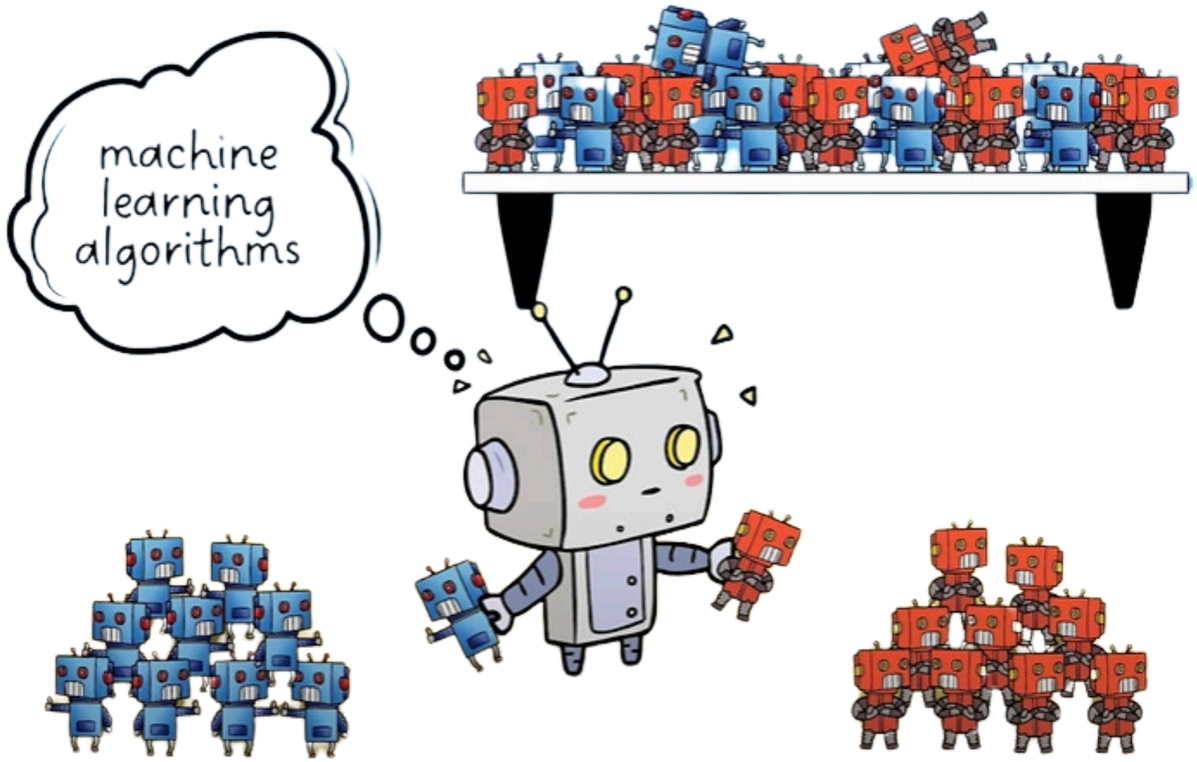
3 Tuning the Models to Perfection!

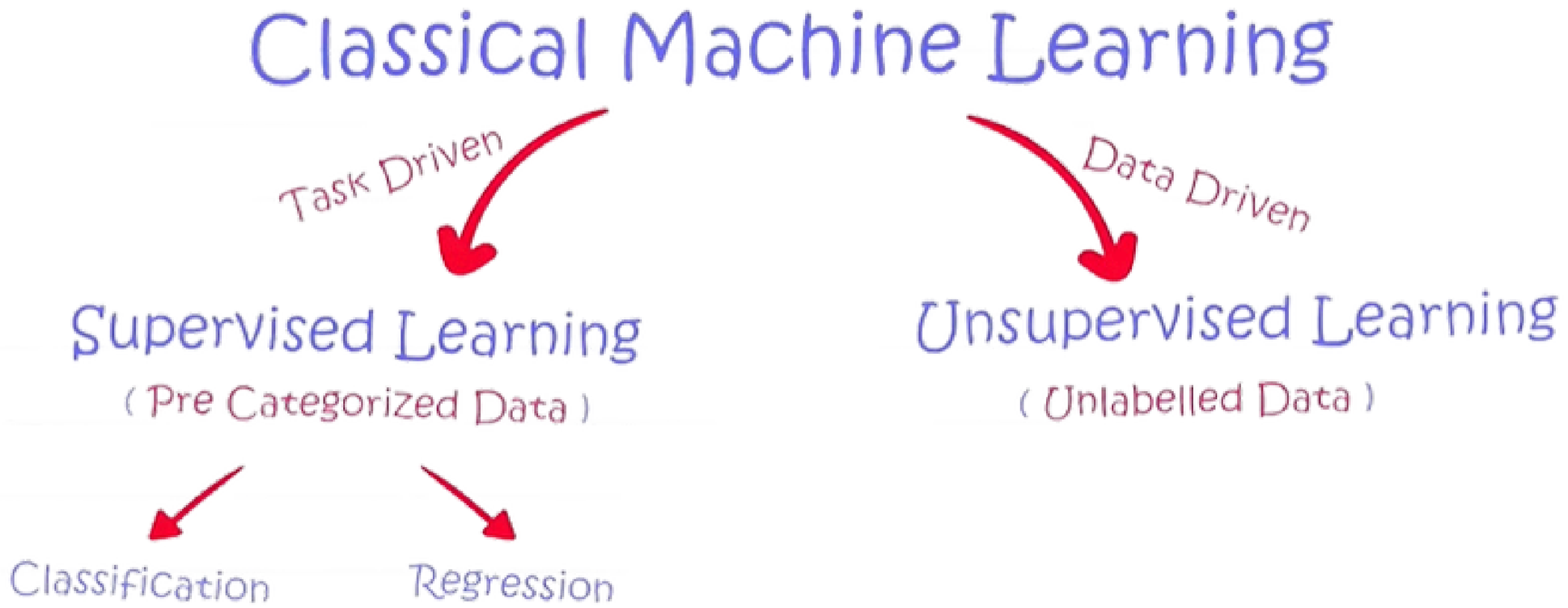
4 How Good Are Our Models?

5 Let’s Build & Test Models!

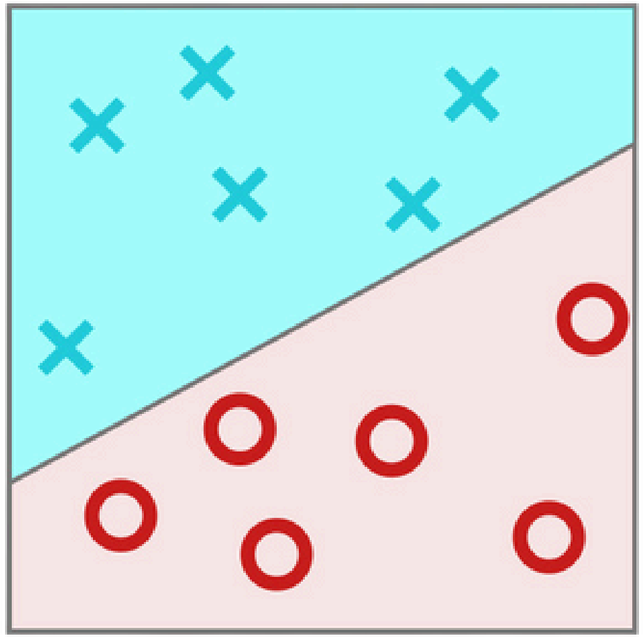
WHAT IS CLASSIFICATION ?

Classification is the process of teaching machines to categorize data into predefined labels or classe

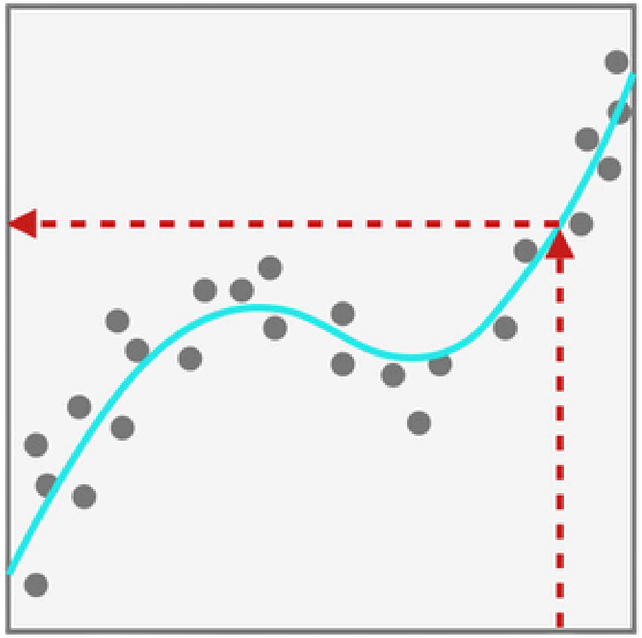



Classical Machine Learning
Task Driven
Data Driven
Supervised Learning
(Pre Categorized Data)
Unsupervised Learning
(Unlabelled Data)
Classification
Regression

REGRESSION VS CLASSIFICATION



Classification

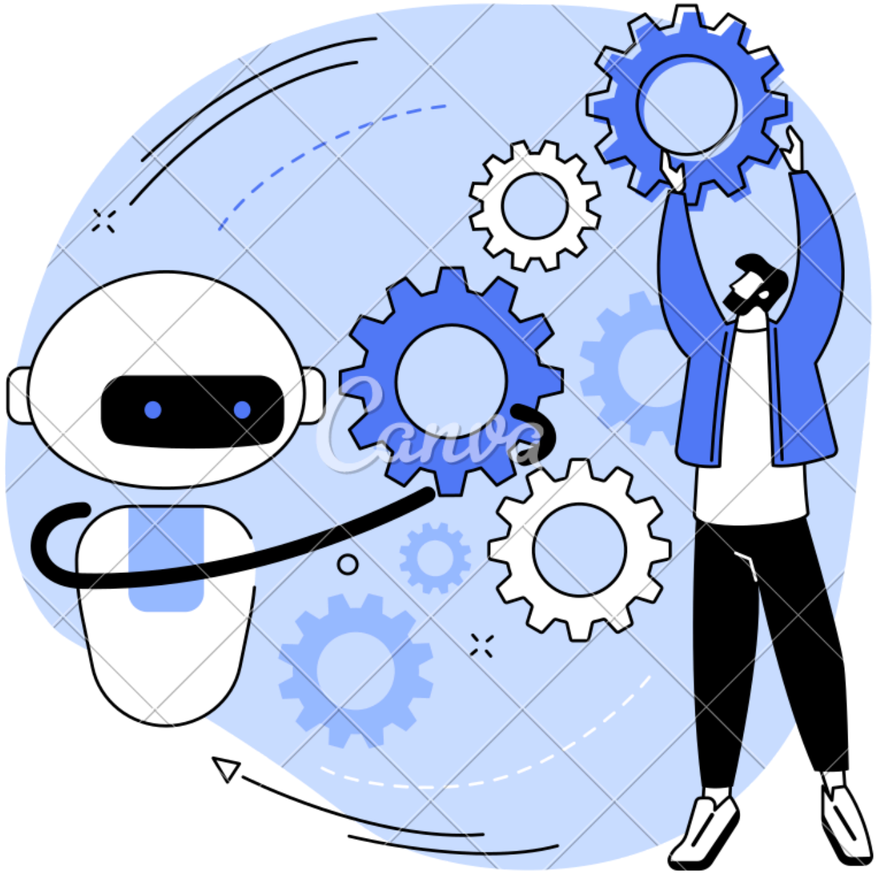


Regression

How to confuse machine learning:

REAL-LIFE EXAMPLES

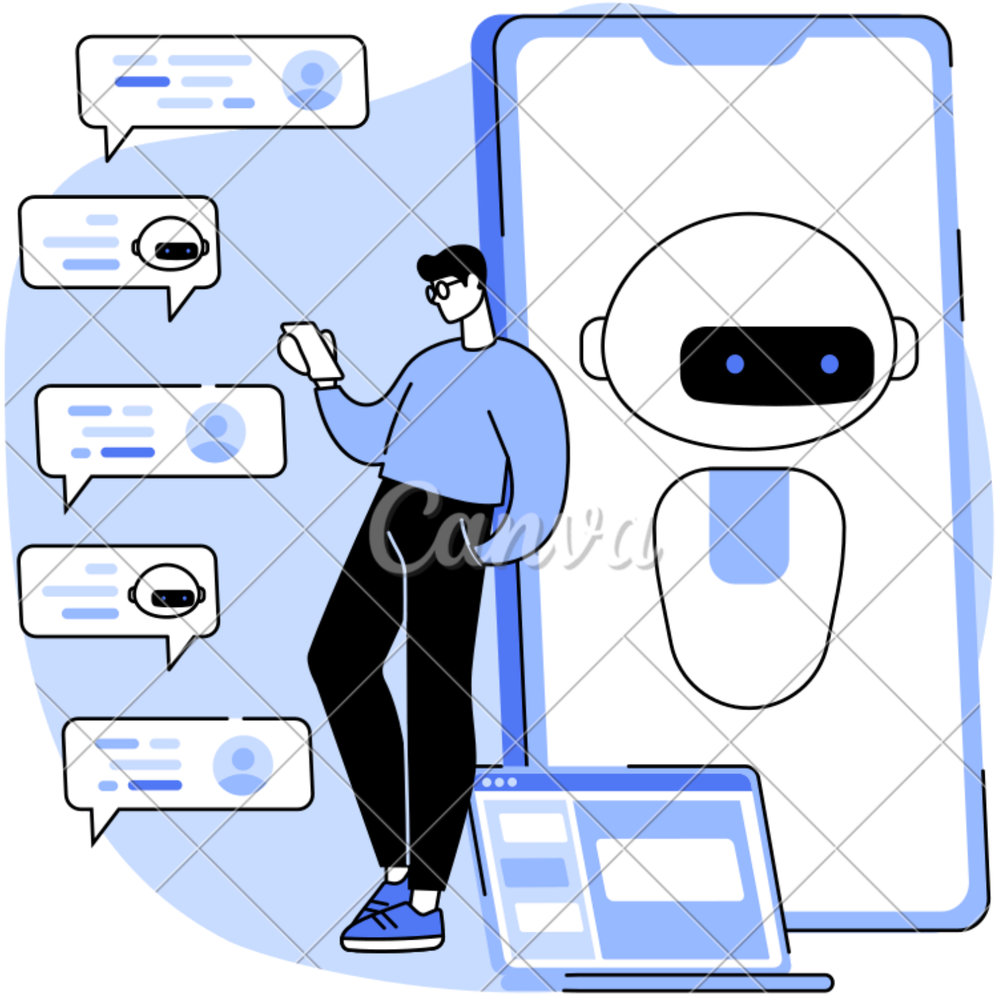
- Spam Emails
- Tumor Detection
- Image Recognition
- Sentiment Analysis



WHERE CLASSIFICATION IS USED

- Healthcare
- Finance
- Social Media:
- E-commerces

CLASSIFICATION MODELS OVERVIEW



START

- Discovering structure
- Finding unusual data points
- Predicting values
- Predicting categories
  - Three or more
  - Two

CLUSTERING

- K-means

ANOMALY DETECTION

- One-class SVM — >100 features, aggressive boundary
- PCA-based anomaly detection — Fast training

REGRESSION

- Ordinal regression — Data in rank ordered categories
- Poisson regression — Predicting event counts
- Fast forest quantile regression — Predicting a distribution
- Linear regression — Fast training, linear model
- Bayesian linear regression — Linear model, small data sets
- Neural network regression — Accuracy, long training time
- Decision forest regression — Accuracy, fast training
- Boosted decision tree regression — Accuracy, fast training

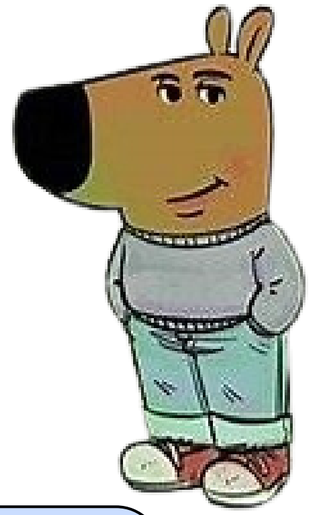
MULTICLASS CLASSIFICATION

- Fast training, linear model — Multiclass logistic regression
- Accuracy, long training times — Multiclass neural network
- Accuracy, fast training — Multiclass decision forest
- Accuracy, small memory footprint — Multiclass decision jungle
- Depends on the two-class classifier, see notes below — One-v-all multiclass

TWO CLASS CLASSIFICATION

- >100 features, linear model — Two-class SVM
- Fast training, linear model — Two-class averaged perceptron
- Fast training, linear model — Two-class logistic regression
- Fast training, linear model — Two-class Bayes point machine
- Accuracy, fast training — Two-class decision forest
- Accuracy, fast training — Two-class boosted decision tree
- Accuracy, small memory footprint — Two-class decision jungle
- >100 features — Two-class locally deep SVM
- Accuracy, long training times — Two-class neural network

MEET THE MODELS!



Decision Trees

"A tree-like structure that splits data based on features to make predictions, offering clear decision-making paths."

Random Forests

"An ensemble of decision trees, combining their strengths to improve accuracy and reduce overfitting."

Support Vector Machines (SVM)

"A powerful model that finds the best boundary to separate different classes in high-dimensional spaces."

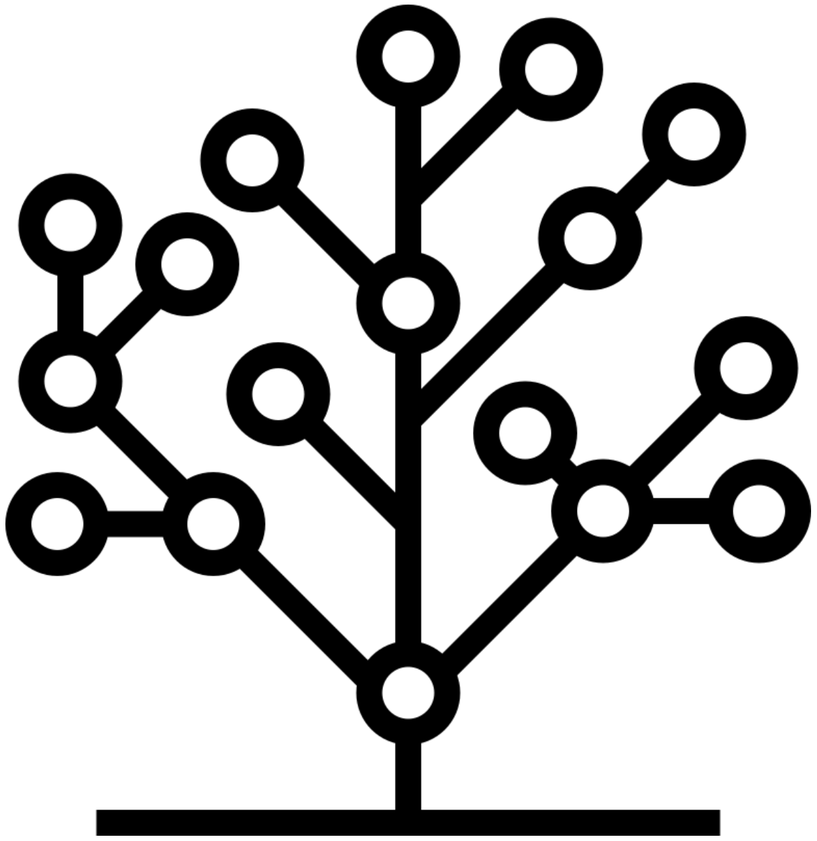
k-Nearest Neighbors (k-NN)

"A simple yet effective model that classifies data based on its closest neighbors in the feature space."

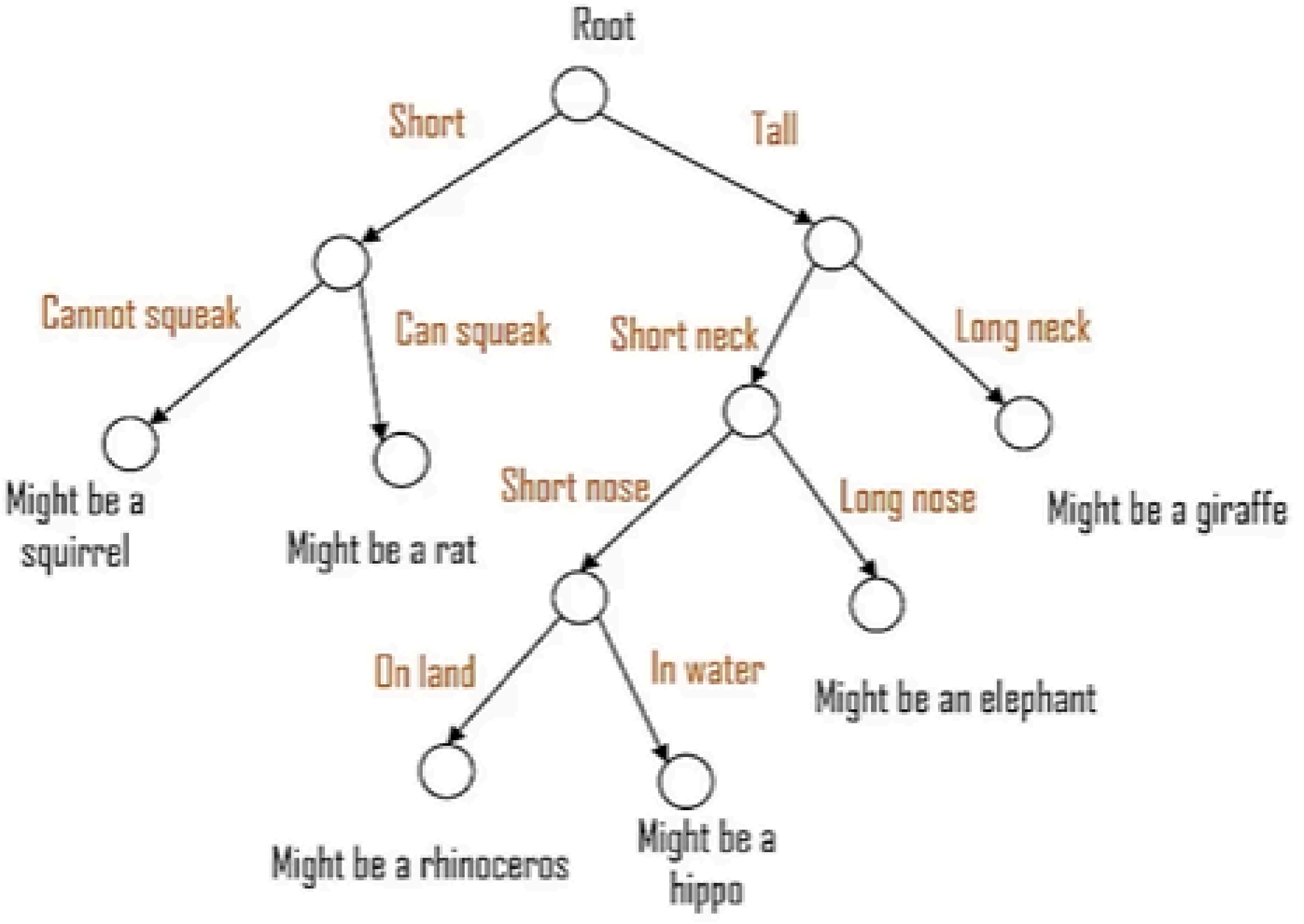
LM = Language Model

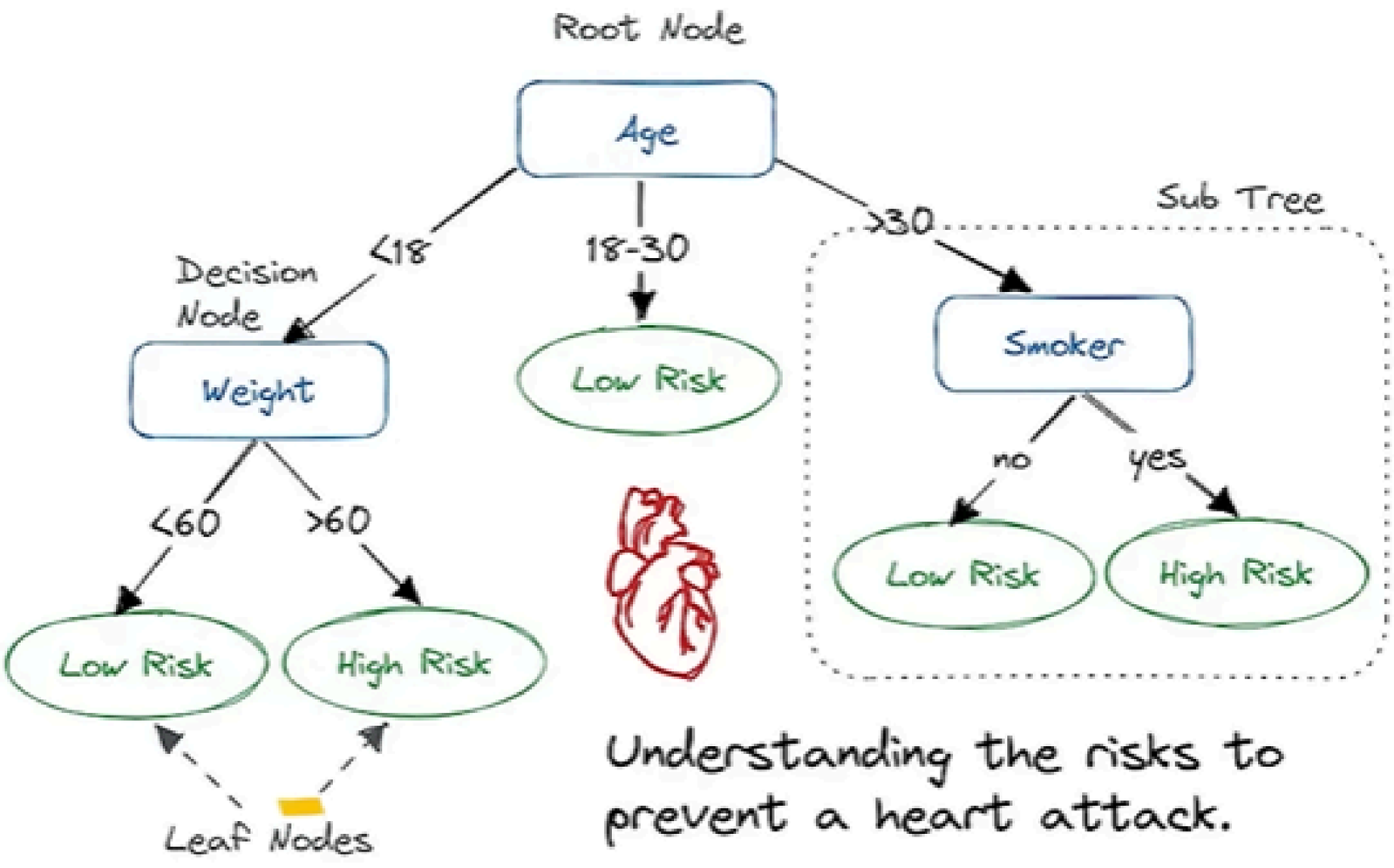
DECISION TREES

Decision Trees: Nature’s Decision Maker!



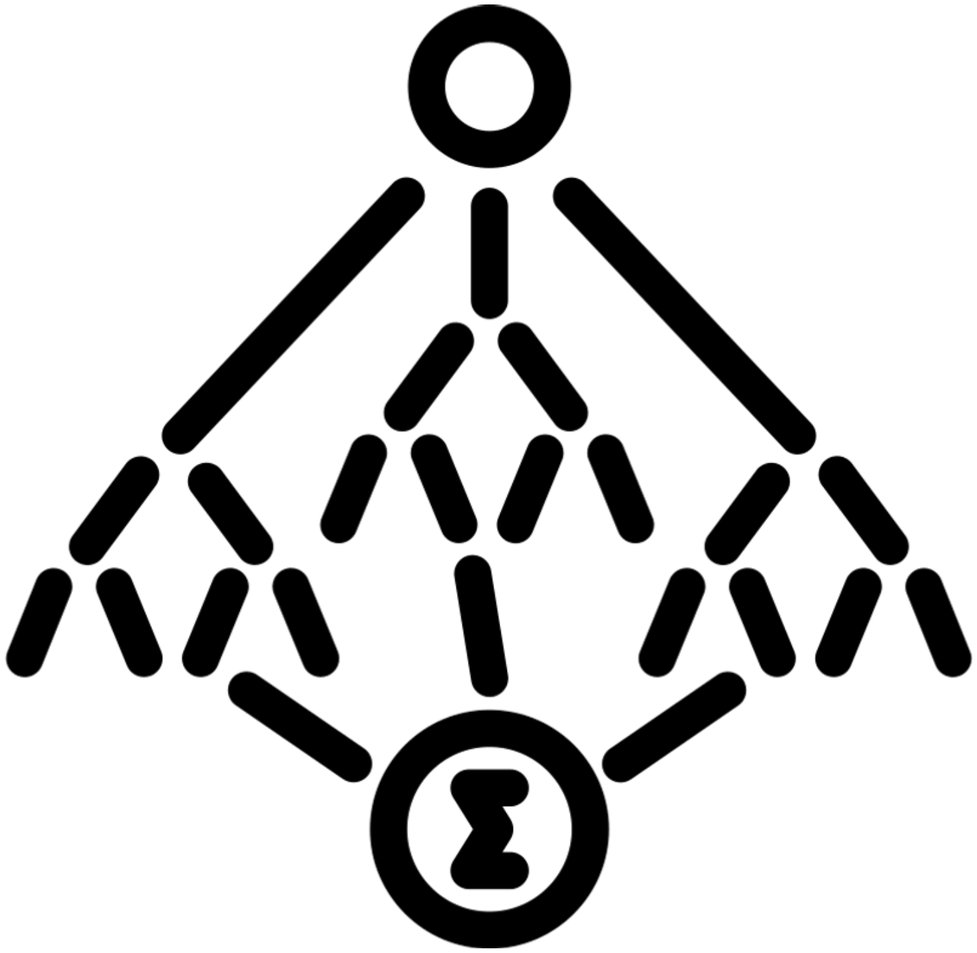
DECISION TREE ANIMAL CLASSIFIER

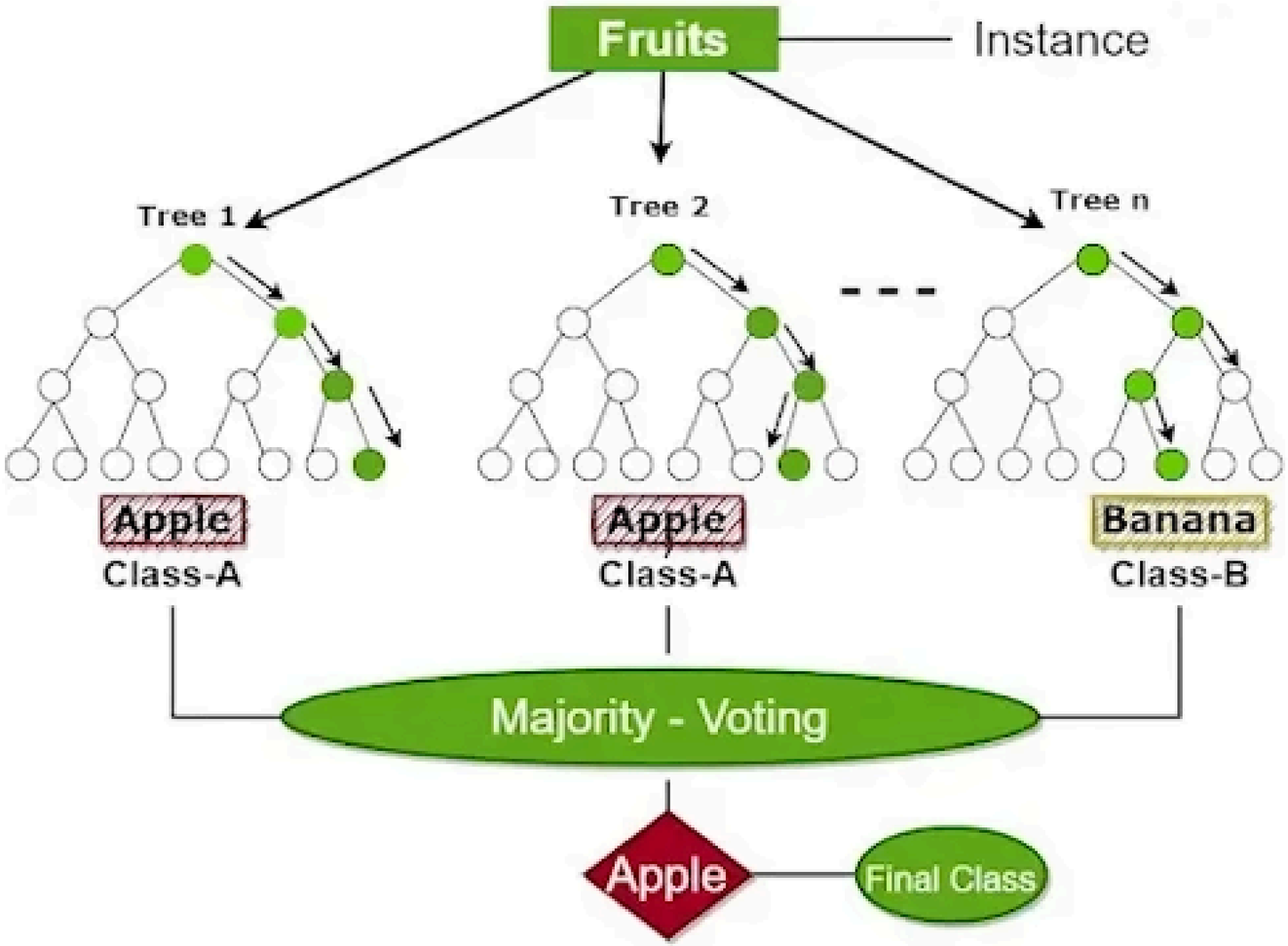


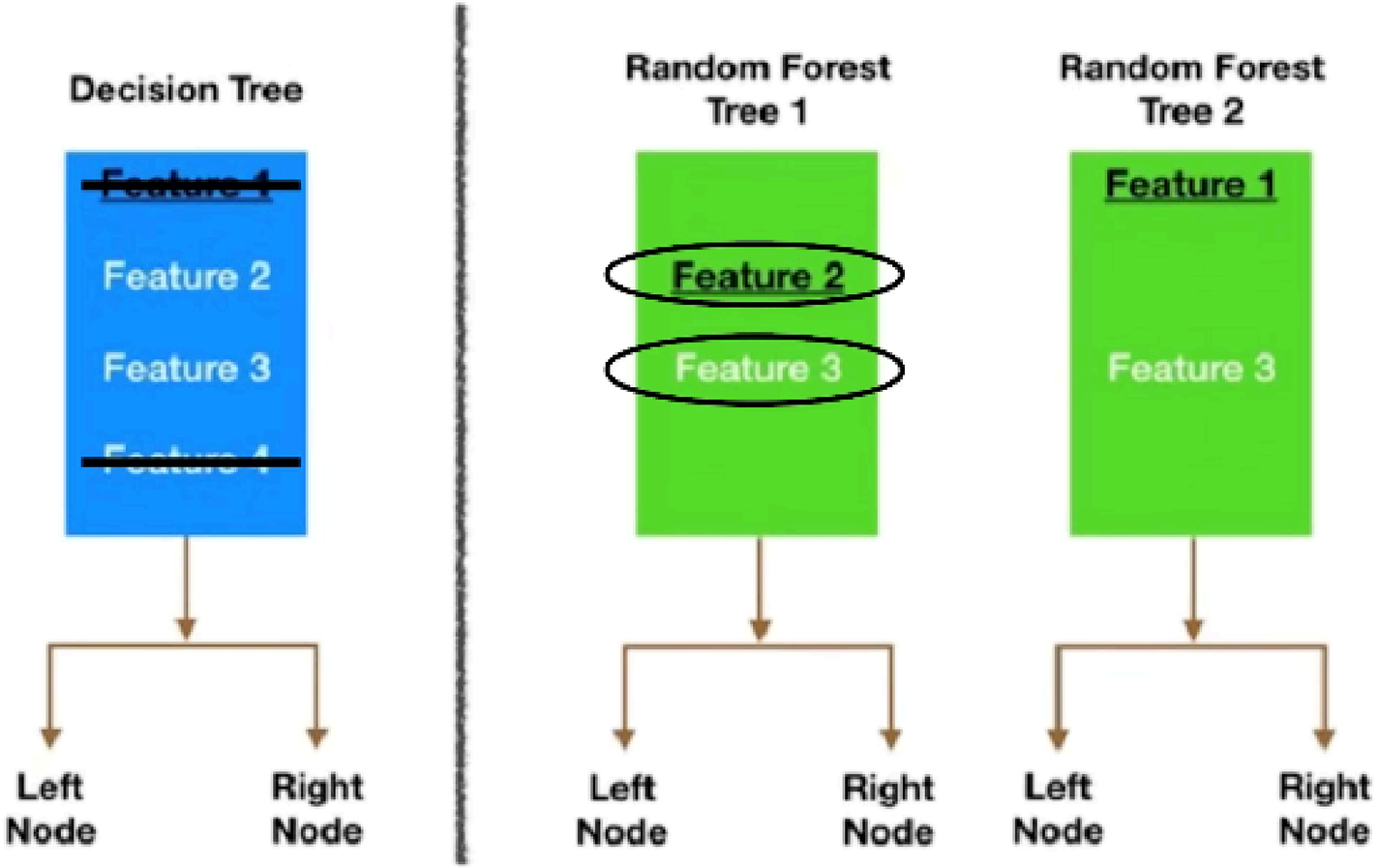


Understanding the risks to prevent a heart attack.

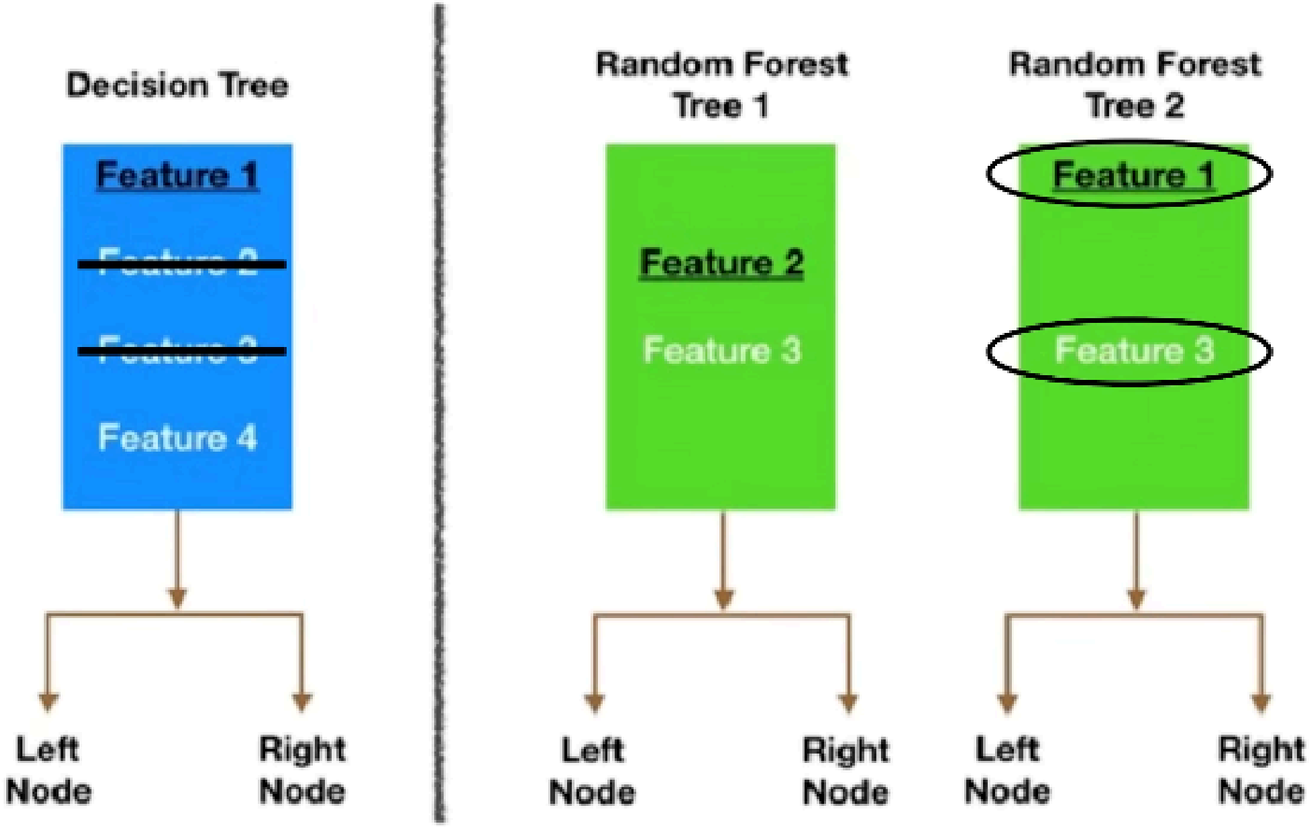
RANDOM FORESTS



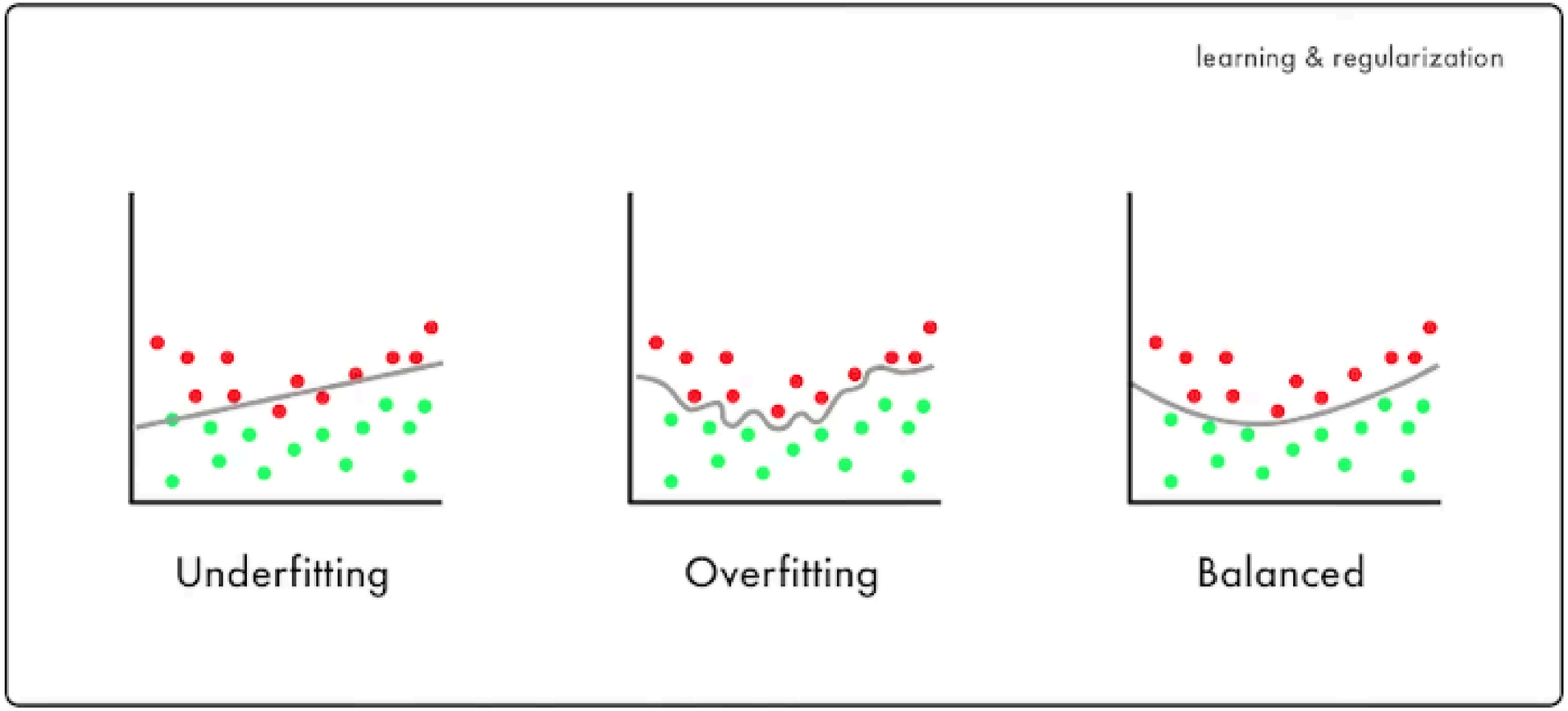

Fruits
Instance
Tree 1
Tree 2
Tree n
Apple
Class-A
Apple
Class-A
Banana
Class-B
Majority - Voting
Apple
Final Class



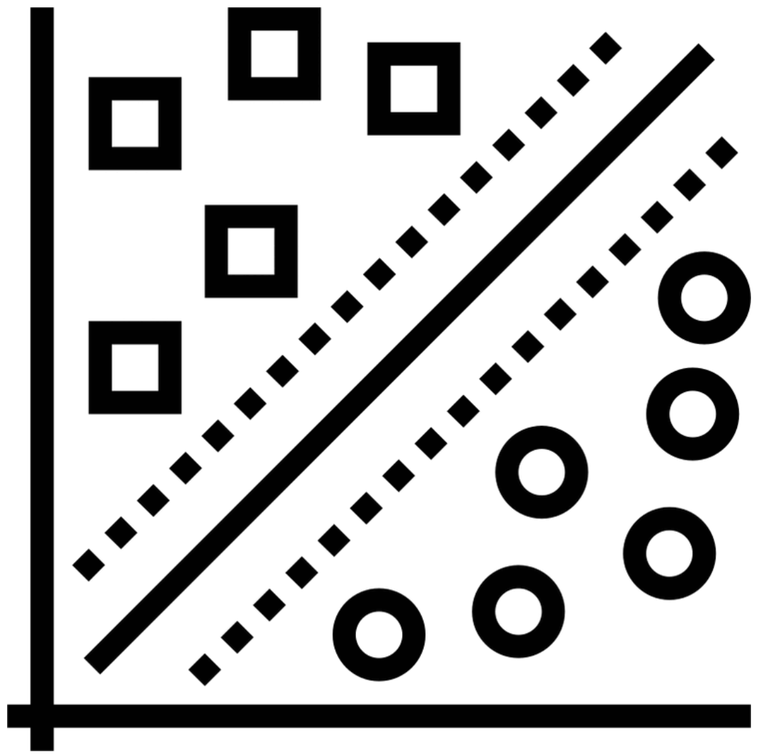
Node splitting in a random forest model is based on a random subset of features for each tree.

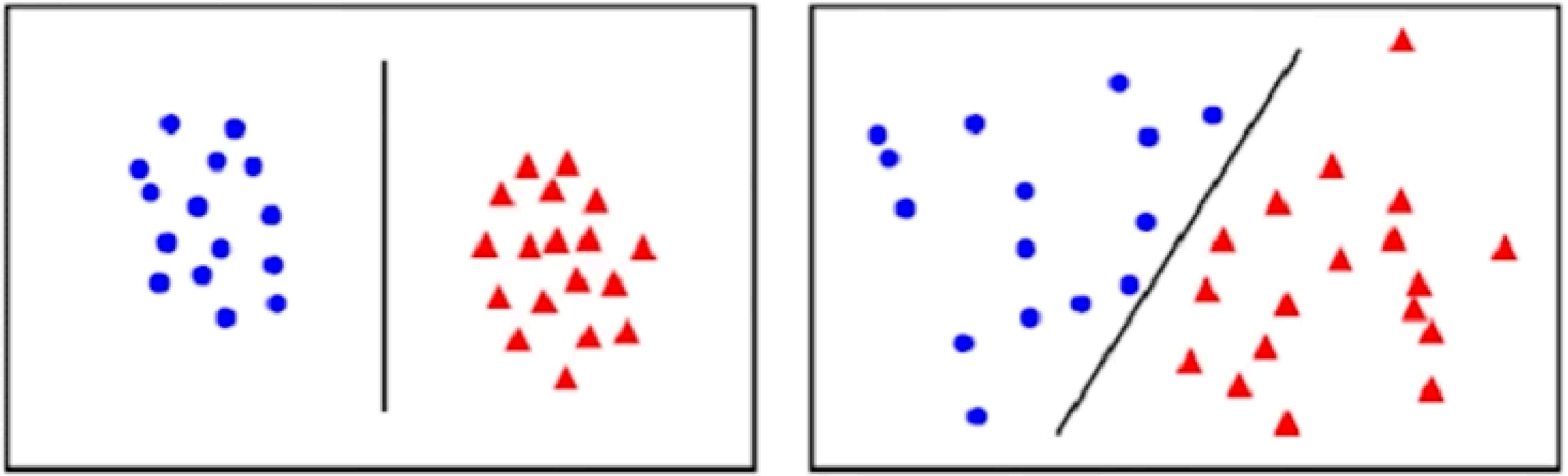


Node splitting in a random forest model is based on a random subset of features for each tree.

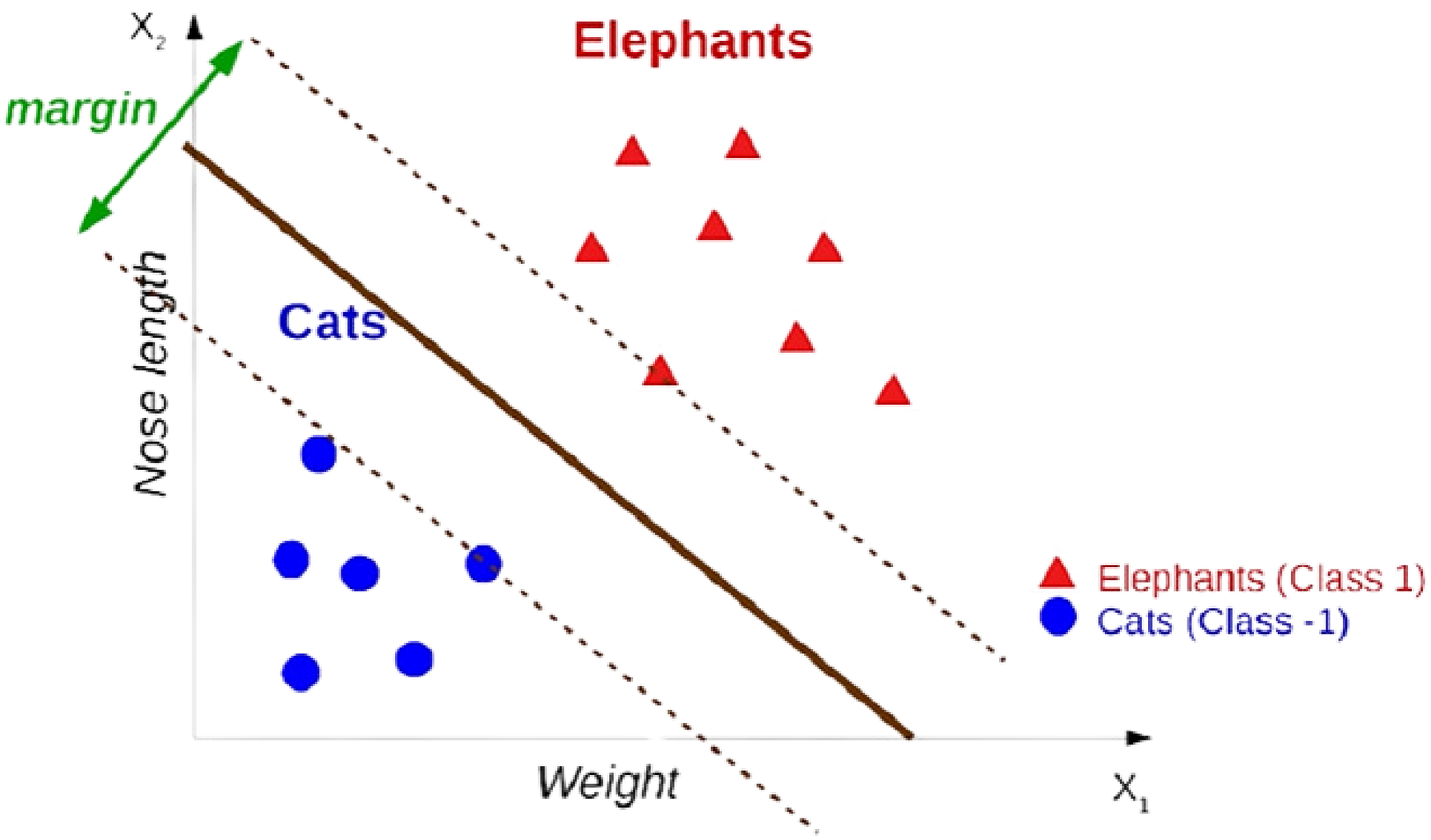

learning & regularization
Underfitting
Overfitting
Balanced

SUPPORT VECTOR MACHINES (SVM)



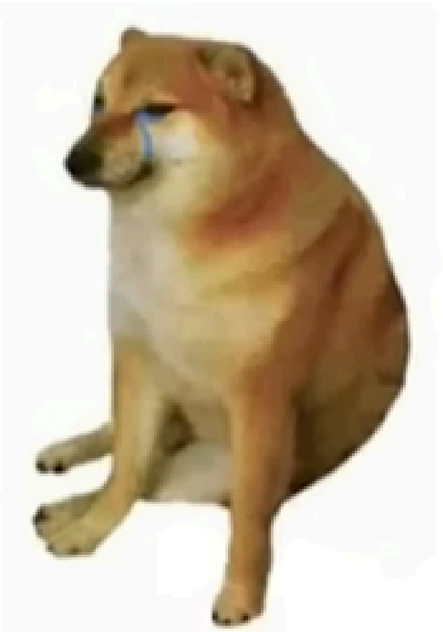


DECISION BOUNDARY


X_2
margin
Elephants
Cats
Nose length
Weight
X_1
Elephants (Class 1)
Cats (Class -1)

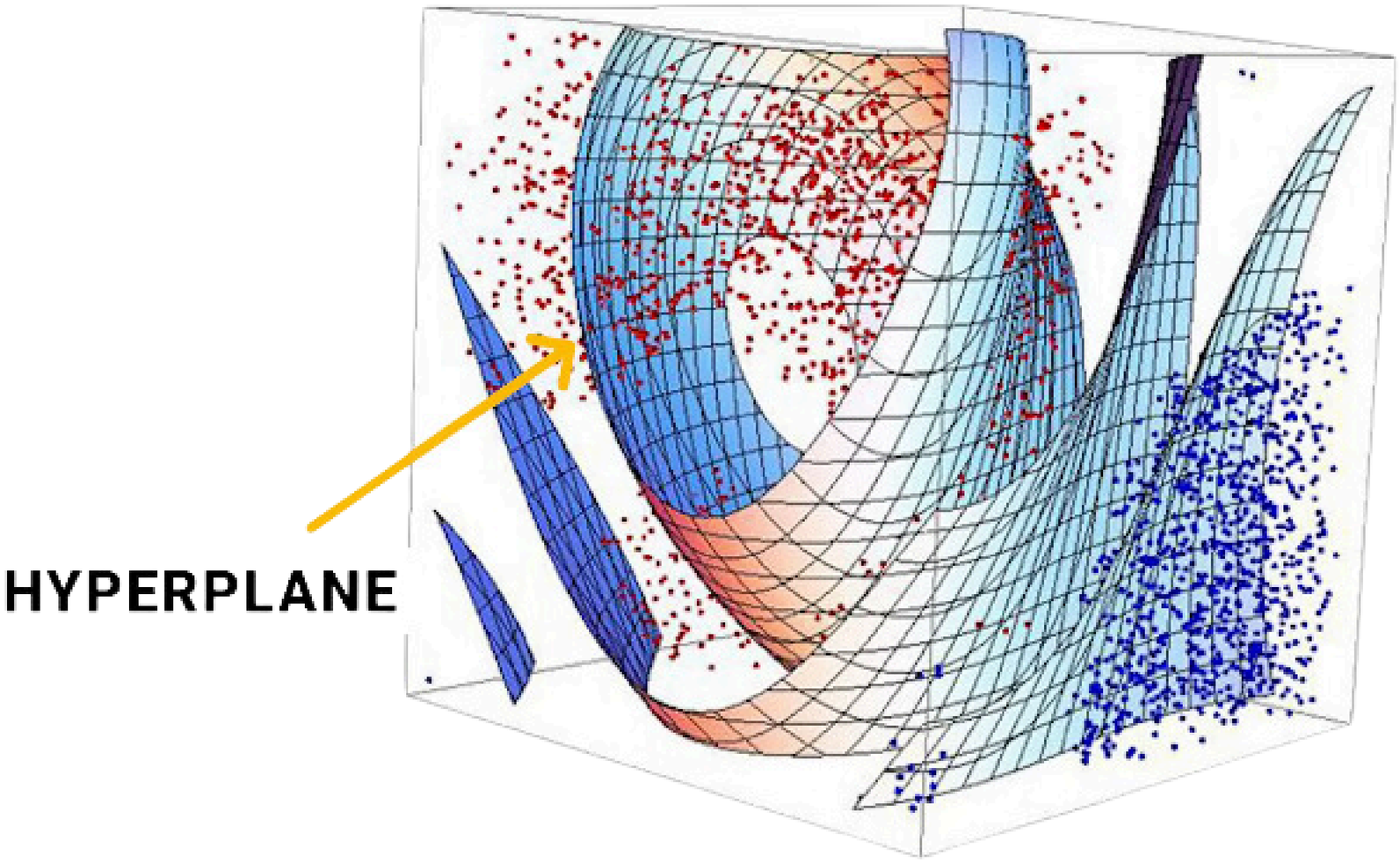
8MB of RAM then

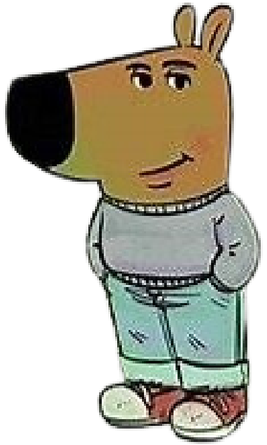
8GB of RAM now



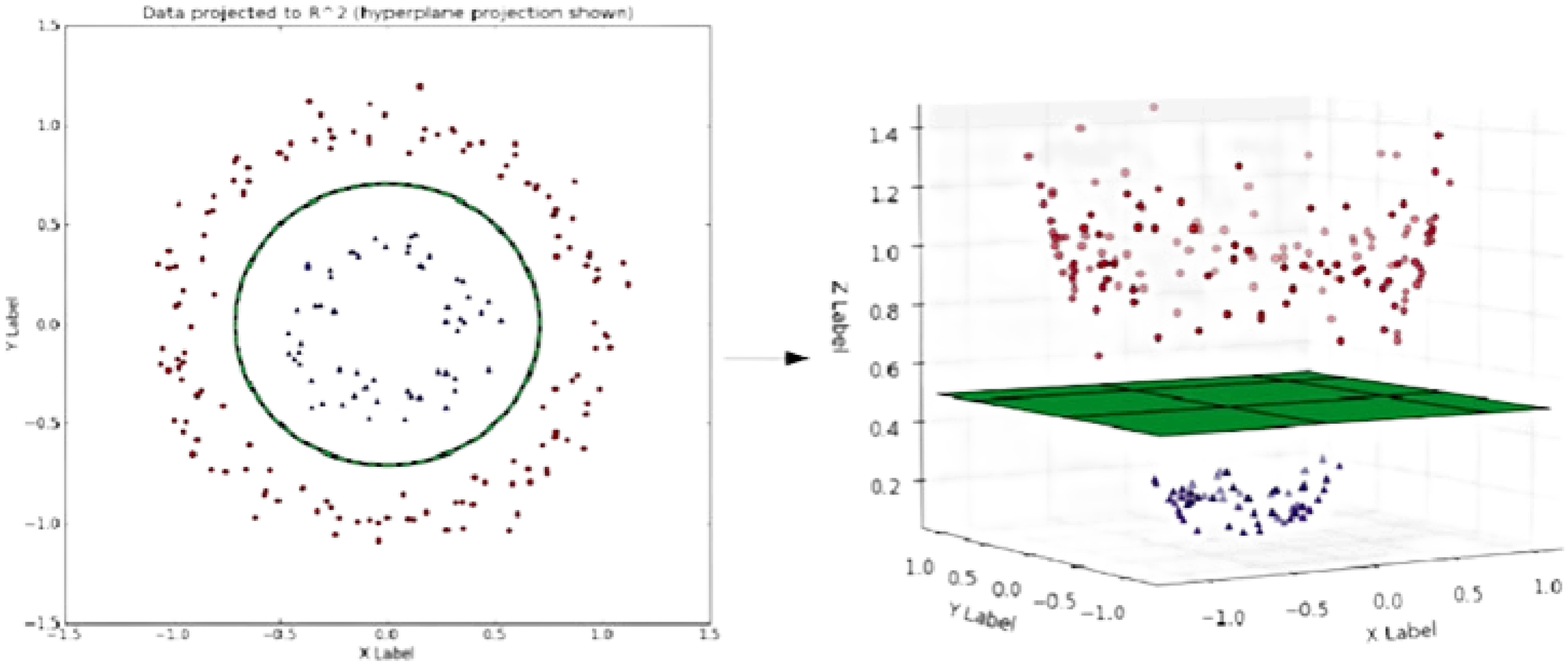
New chrome tab scary

BENEFITS OF SVM





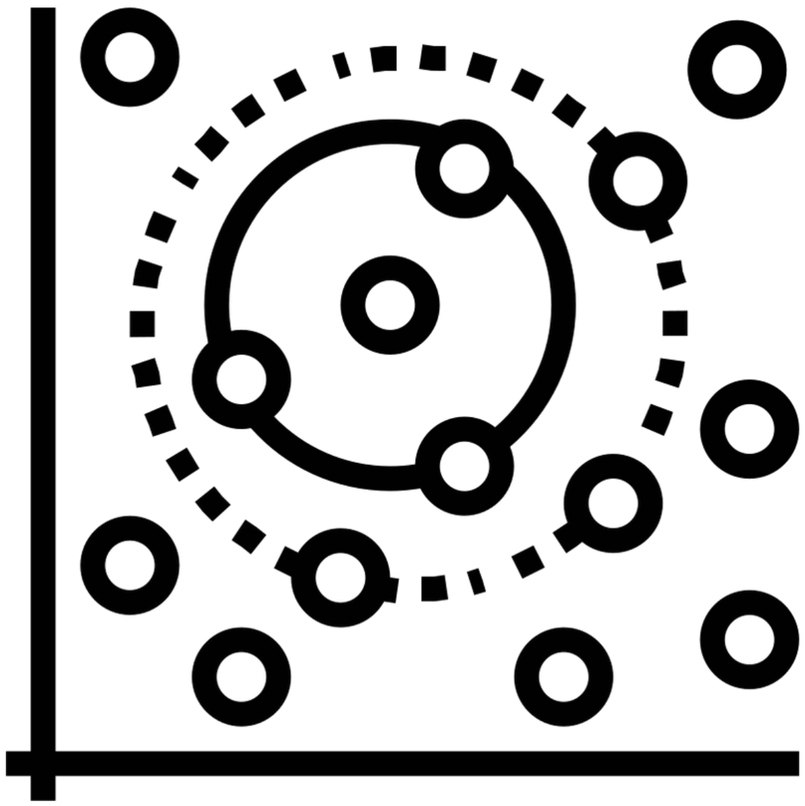
HIGH DIMENSIONAL SPACES (MANY FEATURES)



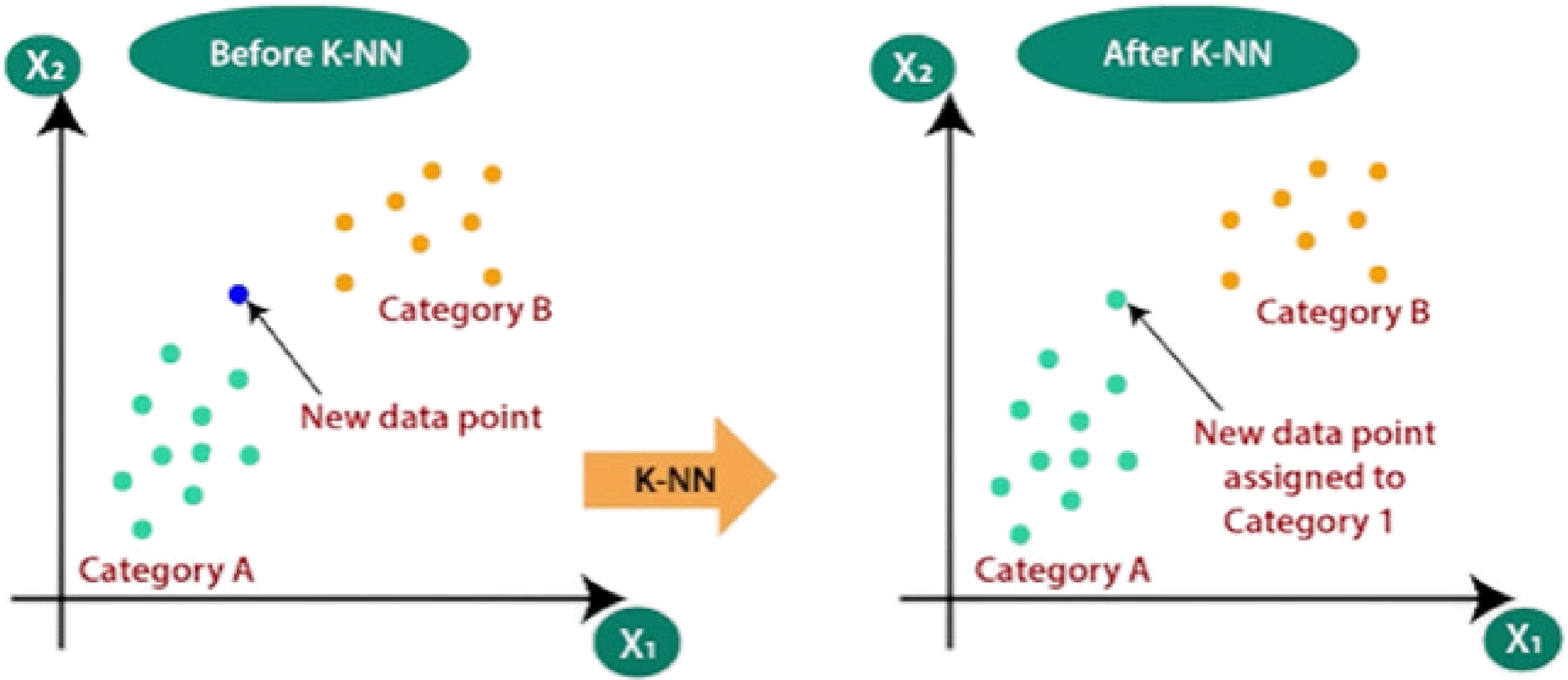
KERNEL FUNCTIONS

No	Kernel function	Formula	Optimization parameter
1	Dot -product	$K(x_n, x_i) = (x_n, x_i)$	C
2	RBF	$K(x_n, x_i) = \exp(-\gamma \parallel x_n - x_i \parallel^2 + C$	C and γ
3	Sigmoid	$K(x_n, x_i) = \tanh(\gamma(x_n, x_i) + r)$	C, γ, and r
4	Poly-nomial	$K(x_n, x_i) = (\gamma(x_n, x_i) + r)^d$	C, γ, r, d

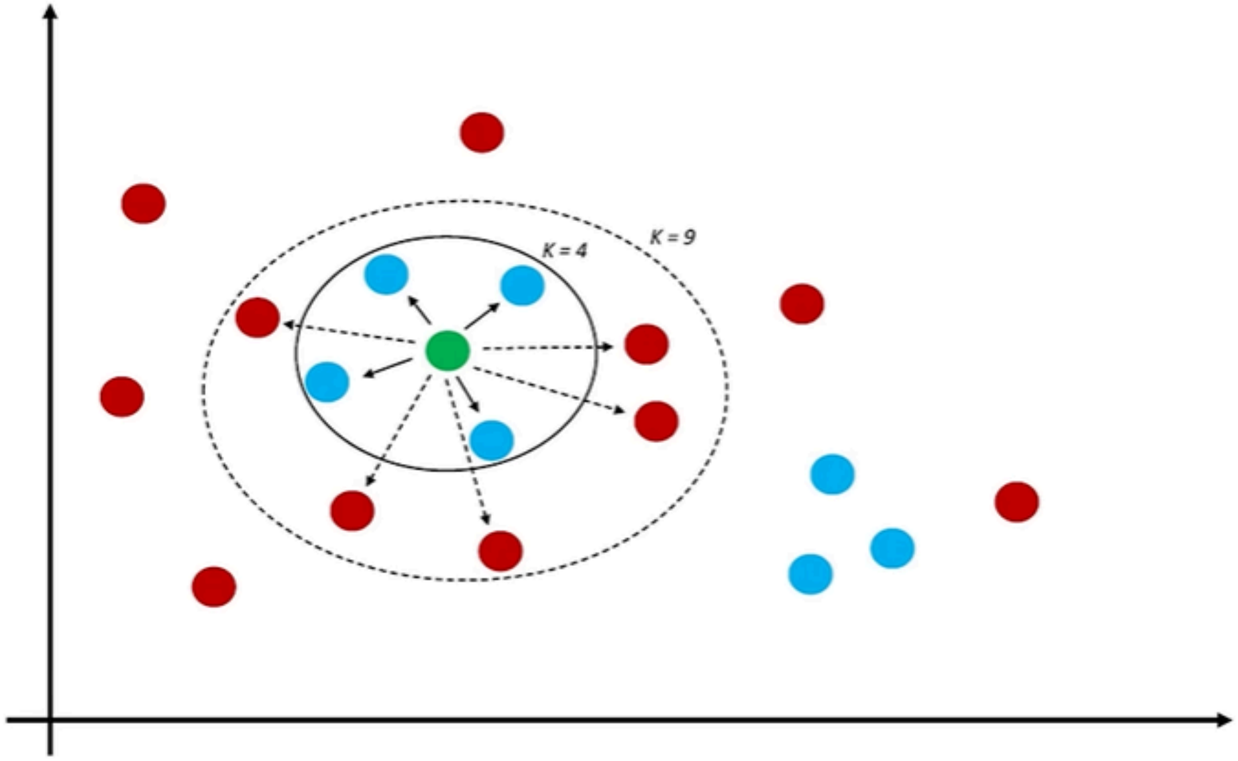
K-NEAREST NEIGHBORS (K-NN)

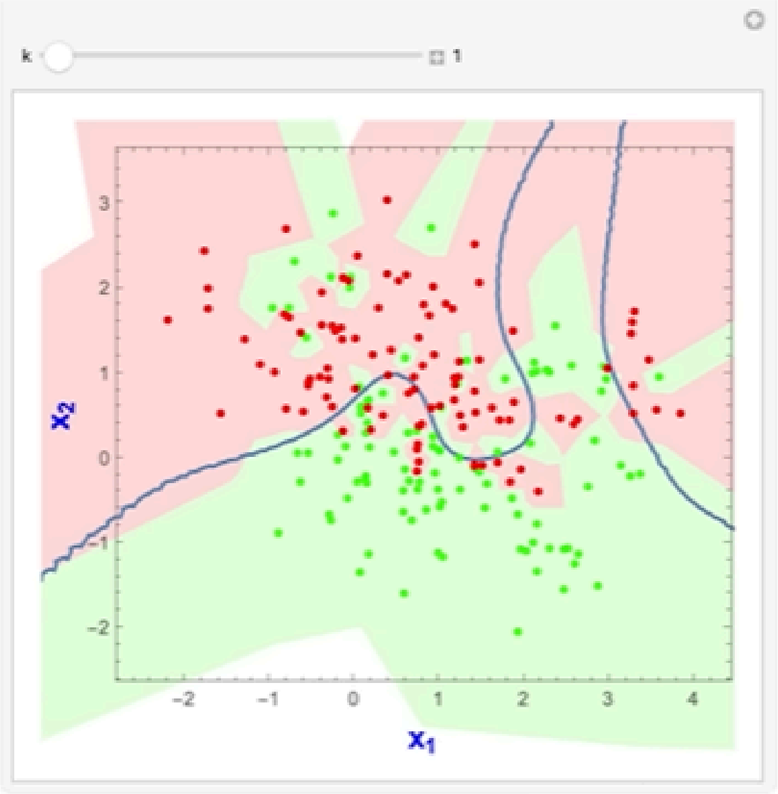


NON-PARAMETRIC MODEL



MAIN CONCEPT

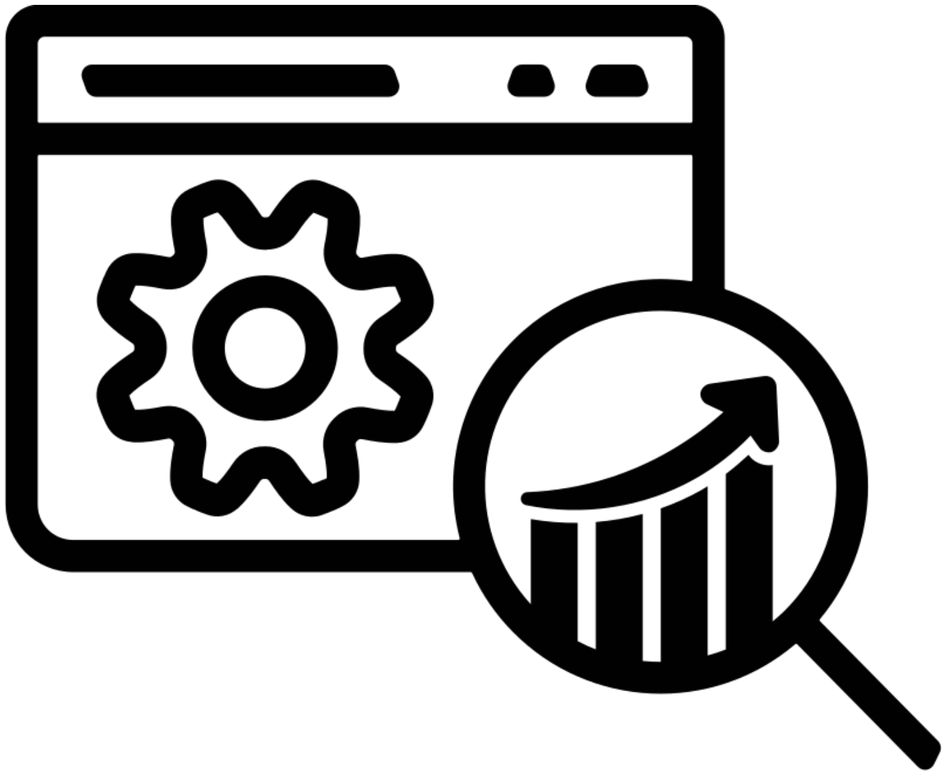




COMPLICATED NON-LINEAR DECISION BOUNDARY

HYPERPARAMETER TUNING

Hyperparameters are the adjustable settings that control how a model learns, like knobs on a machine that can be turned to optimize its performance.

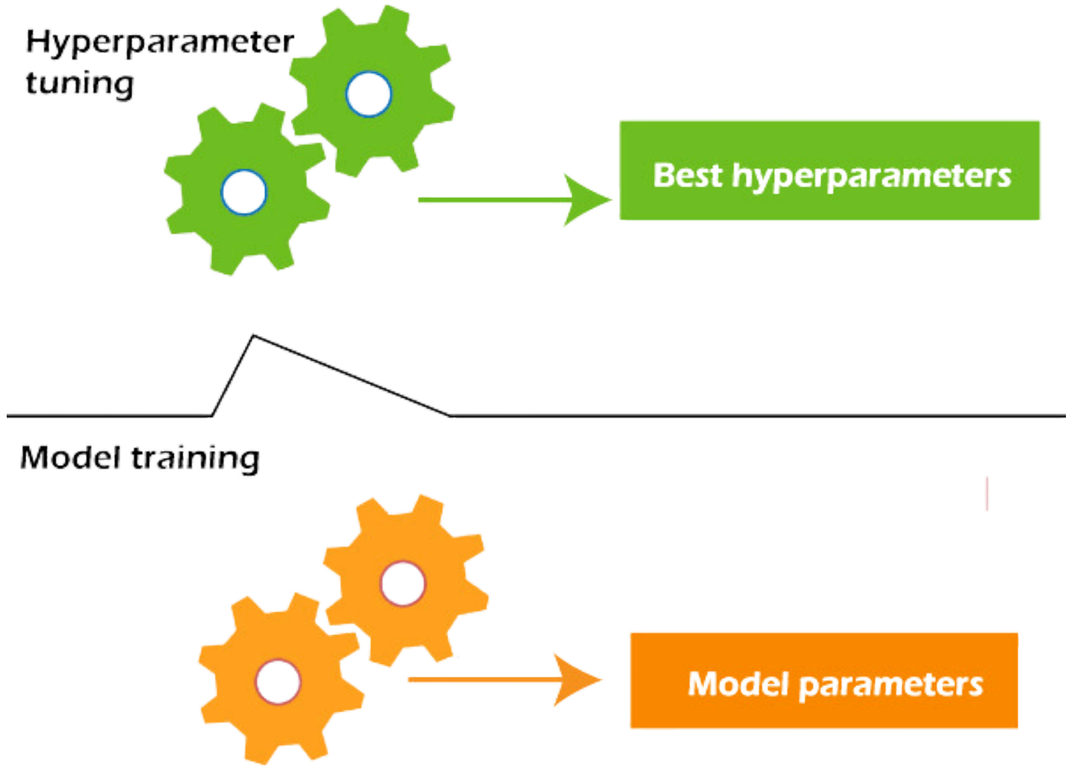


PARAMETER

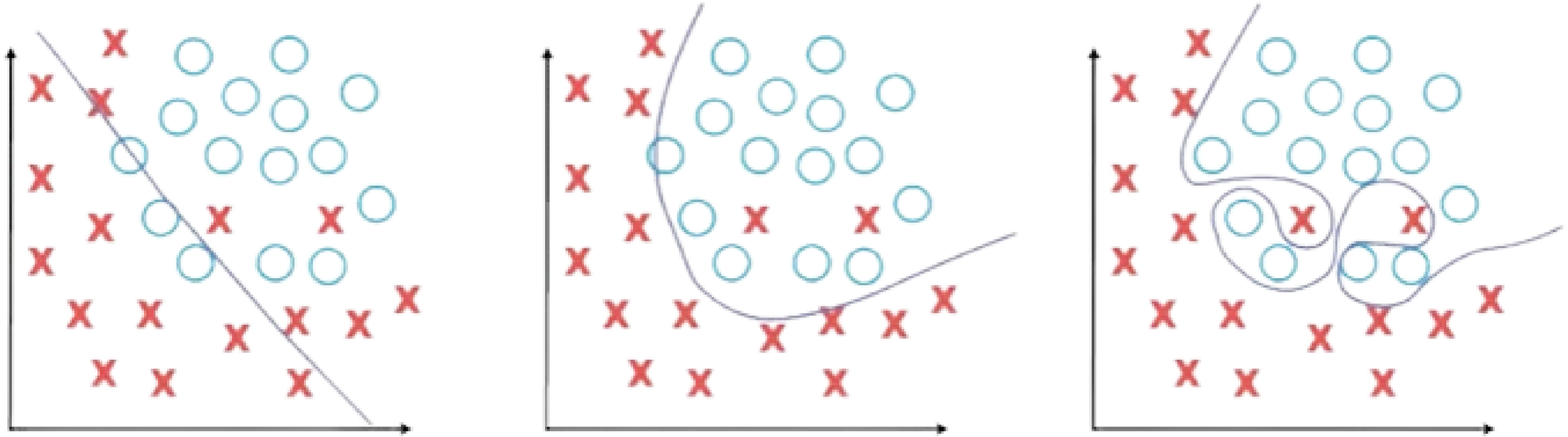
A model parameter is a configuration variable that is internal to the model and whose value can be estimated from the given data.

HYPERPARAMETER

A model hyperparameter is a configuration that is external to the model and whose value cannot be estimated from data.



LET'S TAKE EXAMPLE OF KNNS



Under-Fitting

(too simple to explain the variance)

K=1000

Appropriate-Fitting

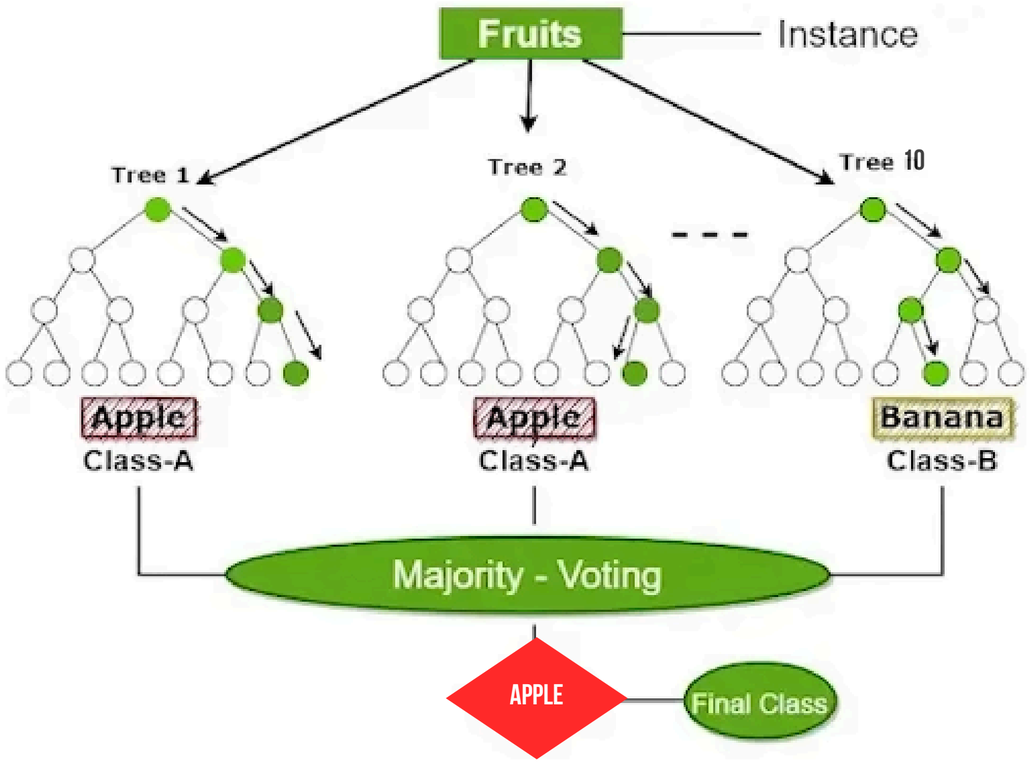
K=5

Over-Fitting

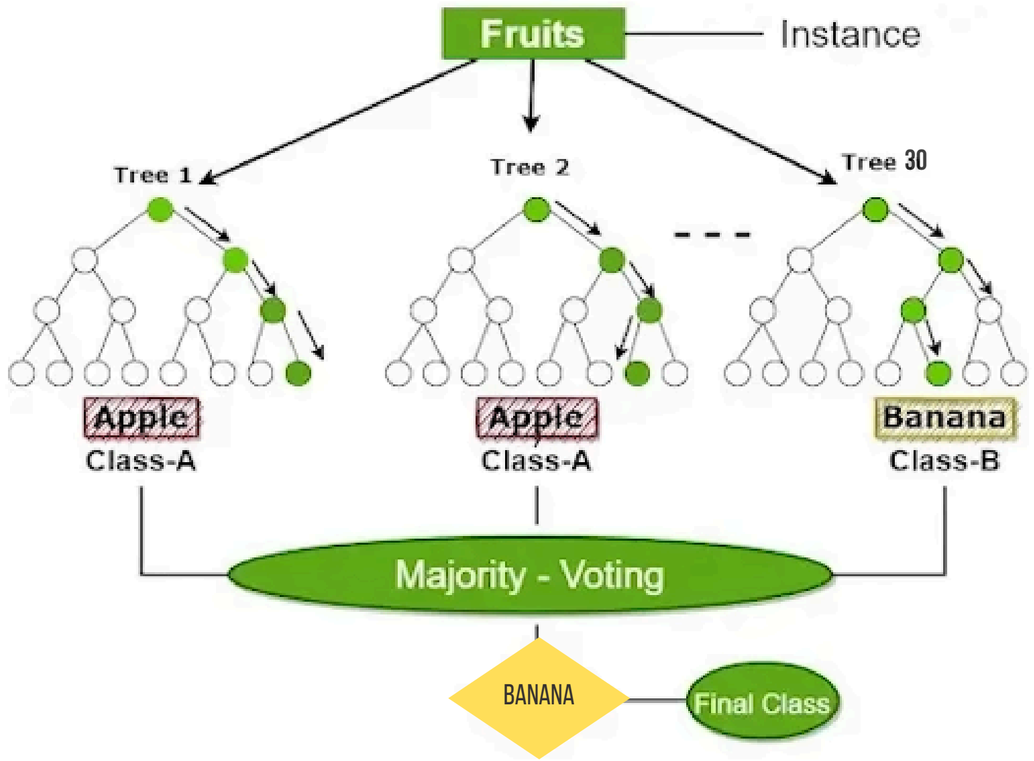
(force-fitting – too good to be true)

K=1

LET'S TAKE EXAMPLE OF RANDOM FORESTS



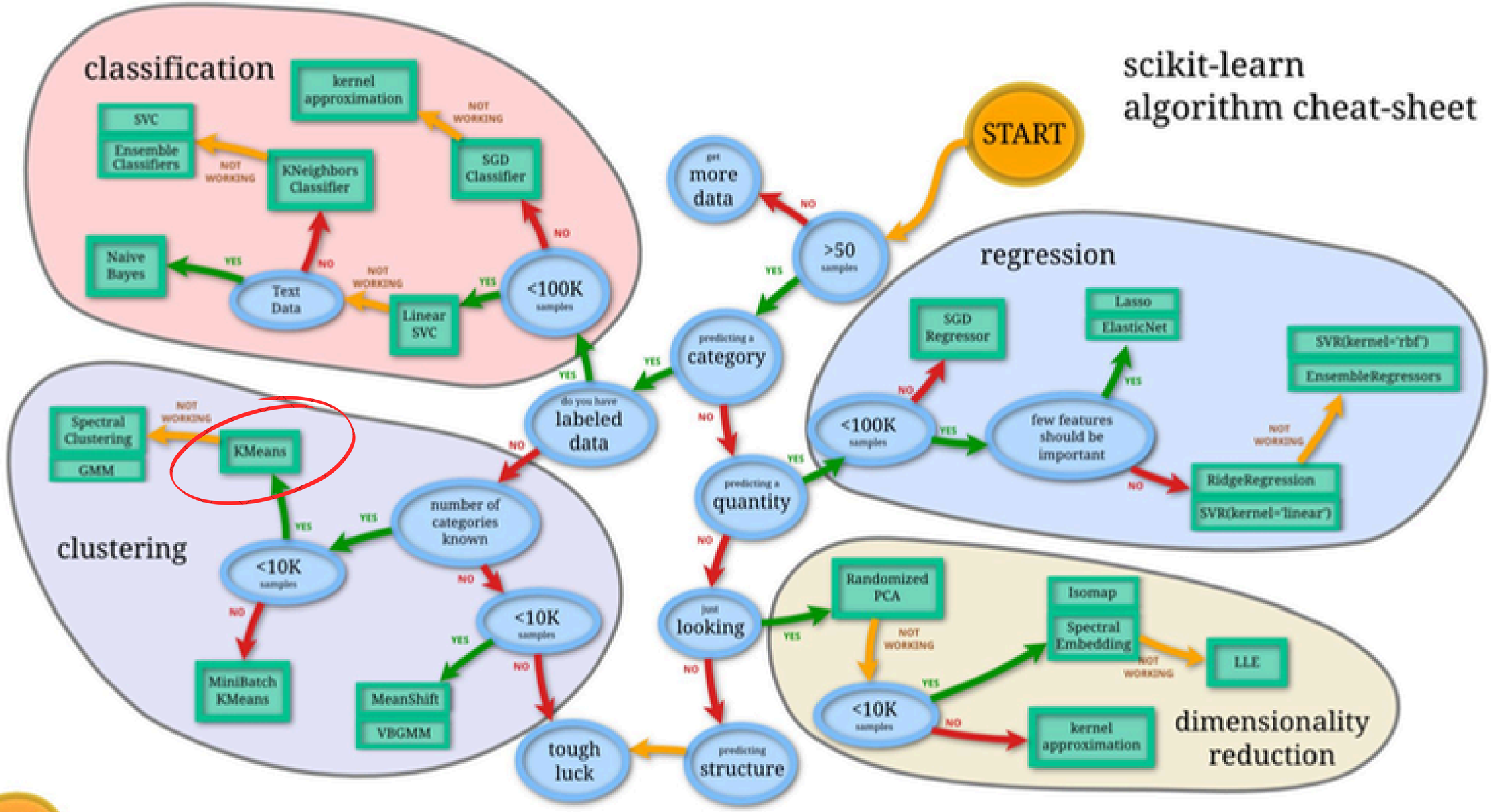
number of trees = 10



number of trees = 30

Metric	Grid Search	Random Search
Comprehensiveness	High	Moderate
Computational Cost	High	Lower
Scalability	Limited	Higher

scikit-learn algorithm cheat-sheet



HANDS ON LABS

THANK YOU BEYONCE !