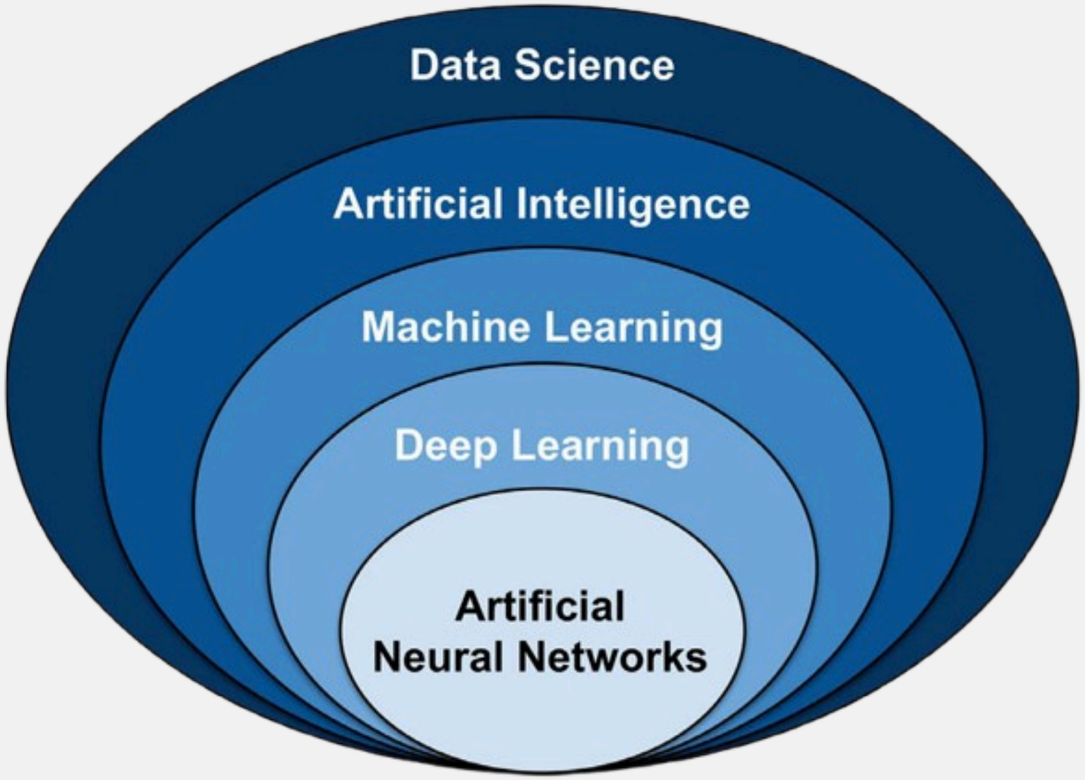
Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
ANG LIJA	ANGEL FORTUNA MAHARANI	ANDREW JIANG	AKUDE TANISHQA SACHIN AKUDE	AVNEET KAUR PANNU	RYAN TAN	RAMKUMAR SAKTHI MAHESHWARI	SREENIVASA NIDHI SREENIVASA
JANELLE HO RUI YEE	JASMINE APRIL AULIA NG	HARNEET KAUR	JOEL YEO	DHIRANA SUNDARAM	SAMMI CHUA YING JIE	XAVIER CHER KAI JUN	THADDEUS NG KAI QI
KIEFER RUSSELL SULIJANTO	JESSICA NOELLYN JAPARA	JINGEN CHEN	LIU YU XUN	LENG JI HERNG	SIVARAMAN LEIKA	SIAH CHIN LOONG	VENSON TOK ZING YUNG
LEE MING XUAN	LIM RONG JUN	LONG XINGXU BENEDICT	MUHAMMAD RAID RAMLEE	MITCH LIAN JUN YU	TAN DE XUN	TAY JING LIN, CHERYL	YAMINI KARLMARX
PAY MEOW LIN ZARA	MUTHURAJAN KOSHIN AMRIT	NADON PANWONG	ROSHAN RAMCHANDANI	NGUYEN THI THUY VAN	UZMA HUSNA BINTE NORDIN	SHAISTA MUSKAN	SOKHENG

Mark Attendance!

Welcome back to DSAcademy!!!

Topic 1

Machine Learning



ML is a **subset of AI** that uses algorithms to learn from data and make predictions without hardcoded rules.

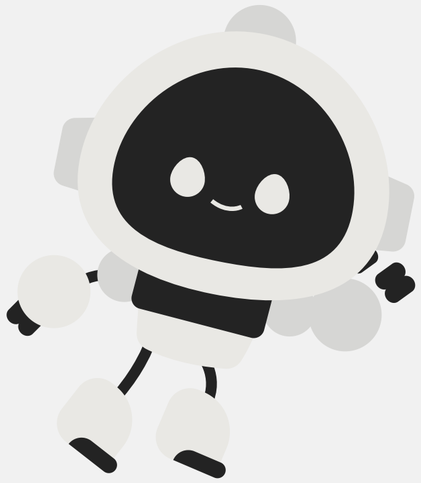
So instead of writing specific instructions, we feed data to the model and let it **learn by itself by detecting patterns**.

“free money” “100% guaranteed”

“limited time” “giveaway” “jackpot”

“you got iPhone” “last chance”

“exclusive deal” “win lucky draw”

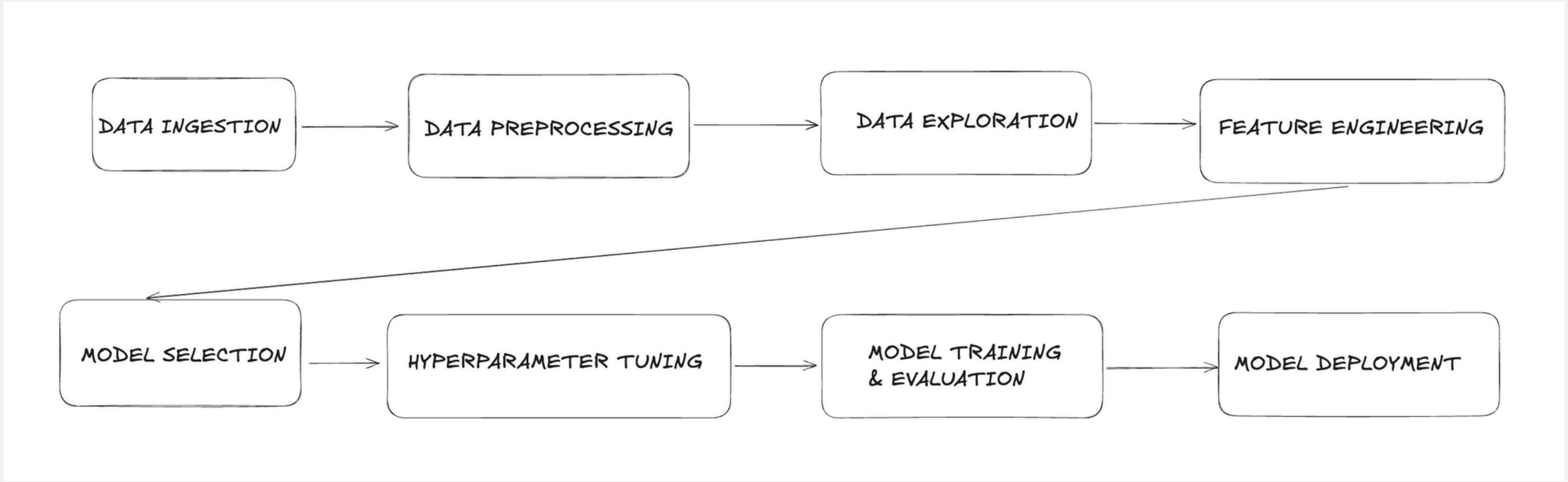


Why is ML important?

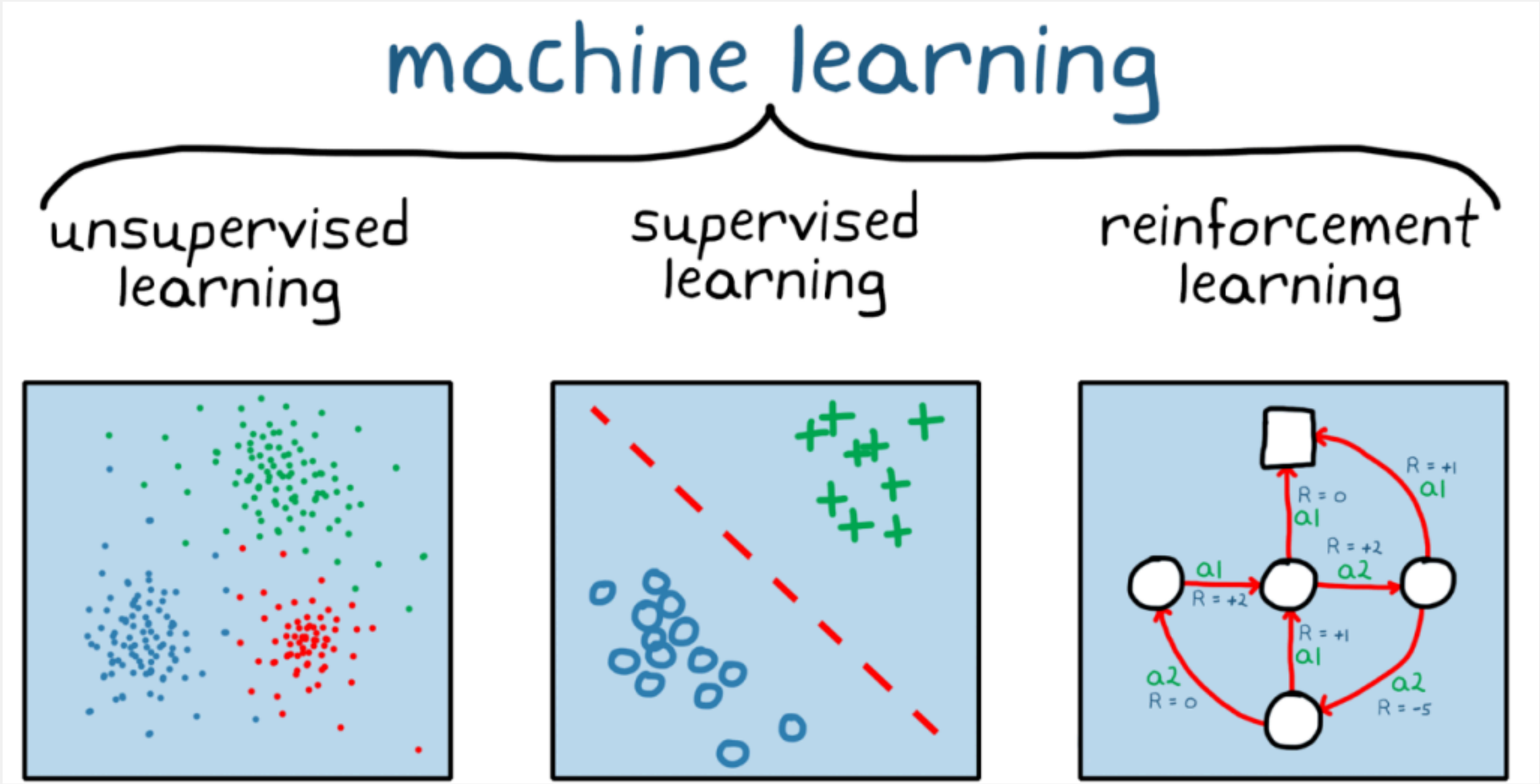
- **Automation**: Automate complex decision-making processes
- **Pattern Recognition**: Find hidden patterns in large datasets
- **Prediction**: Make predictions about future events
- **Personalization**: Customize experiences (like Netflix recommendations)

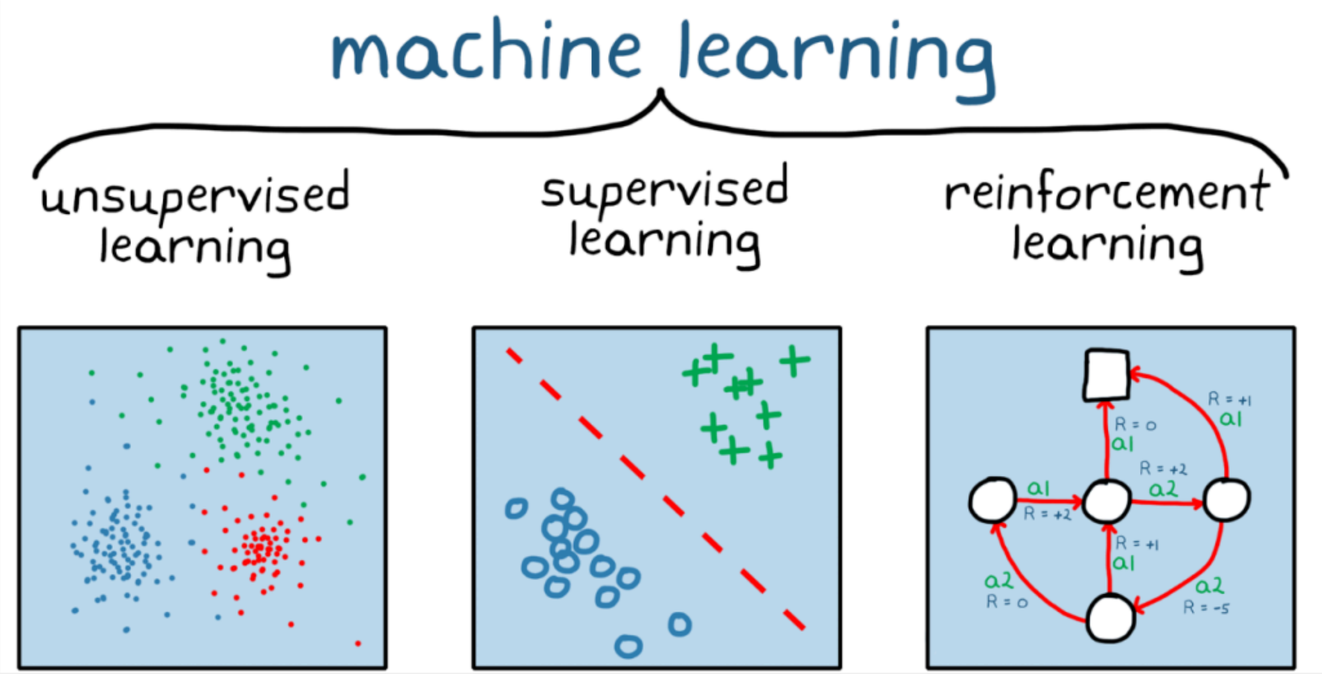
ML Pipeline

Steps to build an ML model:

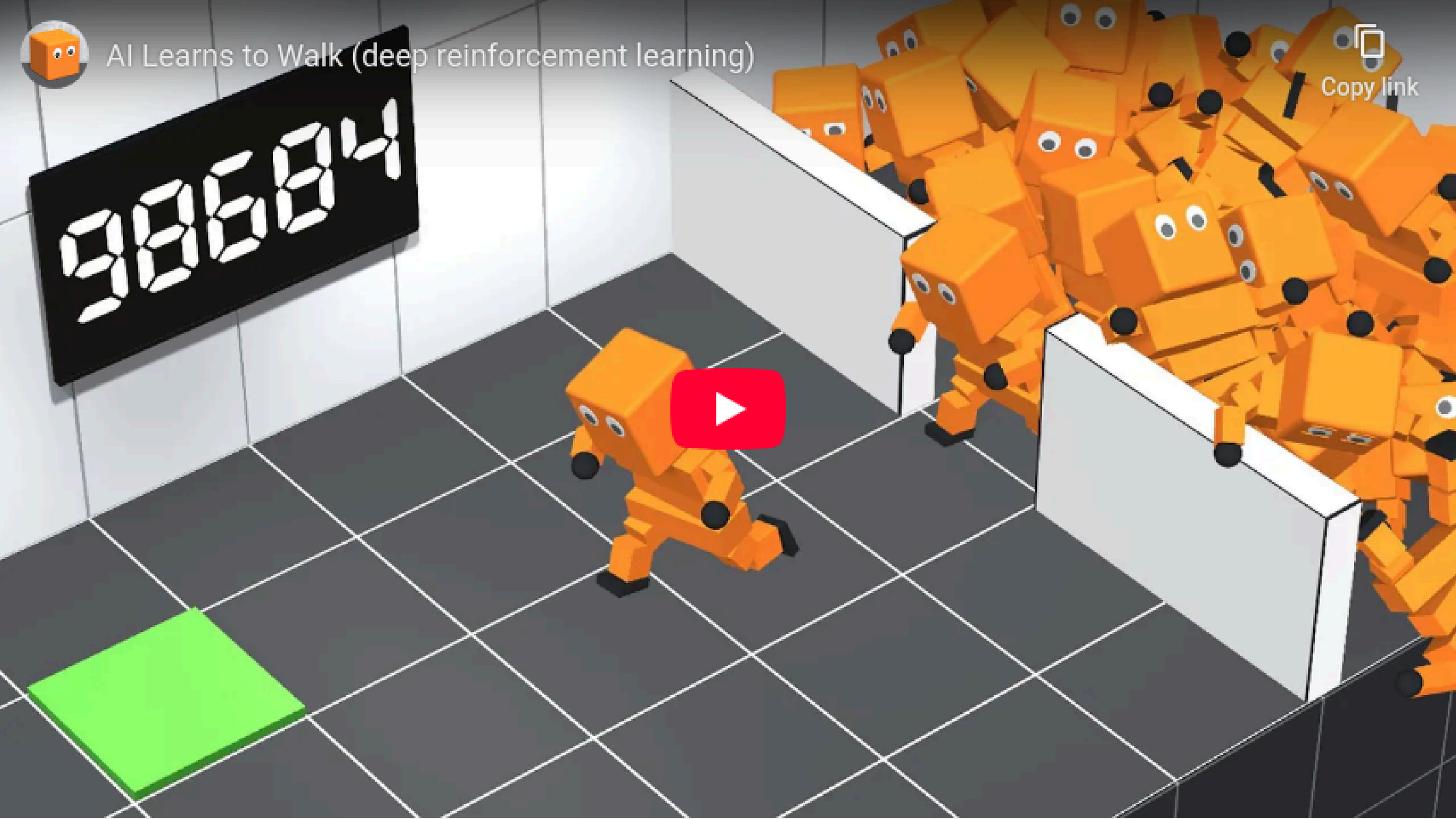


Step	Short Definition
1. Data Ingestion	Load data from files, APIs, or databases
2. Data Preprocessing	Clean and fix messy data
3. Data Exploration	Look at data to understand patterns
4. Feature Engineering	Create better input variables
5. Model Selection	Pick the best algorithm
6. Hyperparameter Tuning	Adjust settings for better performance
7. Model Training and Evaluation	Train model and test accuracy
8. Model Deployment	Put model into real use


machine learning
unsupervised learning
supervised learning
reinforcement learning
R = +1
a1
R = 0
a1
R = +2
a2
a1
R = +2
R = +1
a1
a2
R = 0
a2
R = -5



Definition	Finding patterns in data without labeled examples	Learning with labeled examples (we know the correct answers)	Learning through trial and error with rewards/penalties
Examples	- Customer segmentation - Market basket analysis - Anomaly detection	- Email spam detection (spam/not spam) - House price prediction - Image recognition	- Game playing (chess, Go) - Robot navigation - Trading strategies


AI Learns to Walk (deep reinforcement learning)
98684
Copy link

Let’s Code!

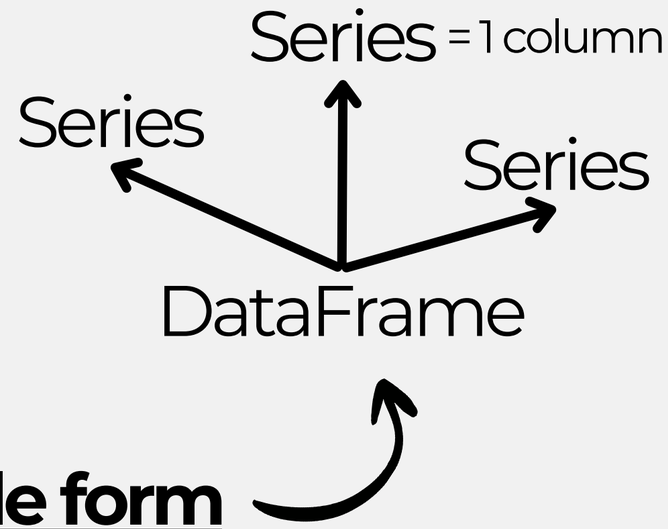
YouTube Trending Video Analysis

Goal: Walk through the full pipeline of a mini data science project using real-world YouTube trending data.

BUT! Before that...

Recap for Python Libraries:

pandas



= python library that **keeps your data in <u>table form</u>** (like excel/spreadsheet) in memory, so you can do:

- sort
- filter
- group
- join
- etc

matplotlib

= python library for **data visualization**

data

chart

What we'll do later:

Raw Data

→ pandas →

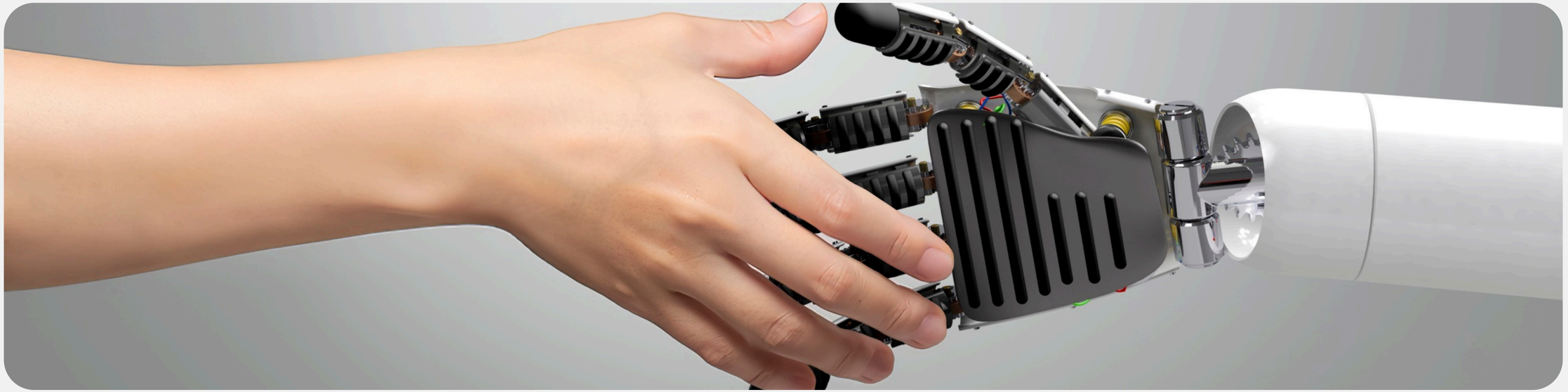
DataFrame

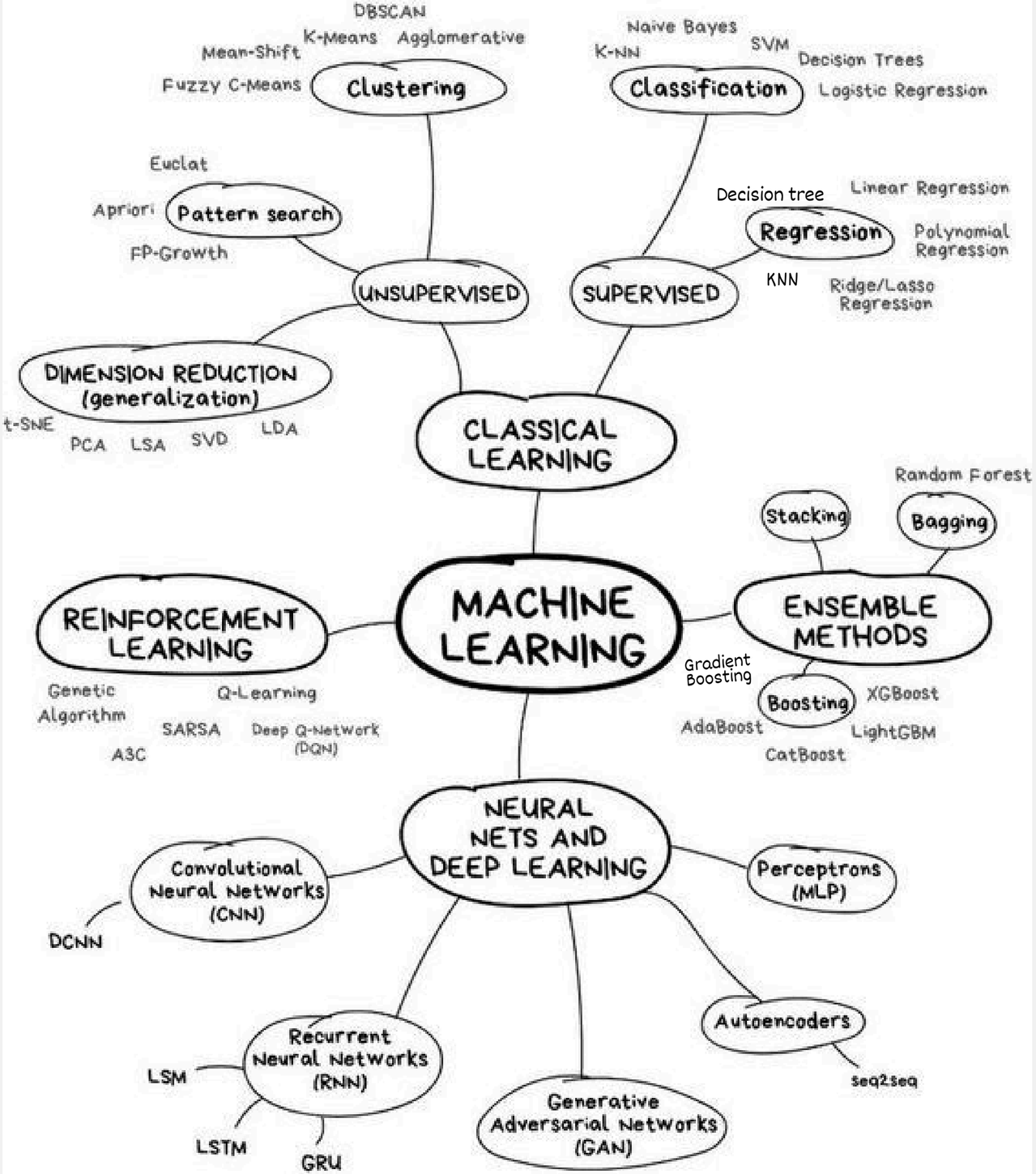
Visualize

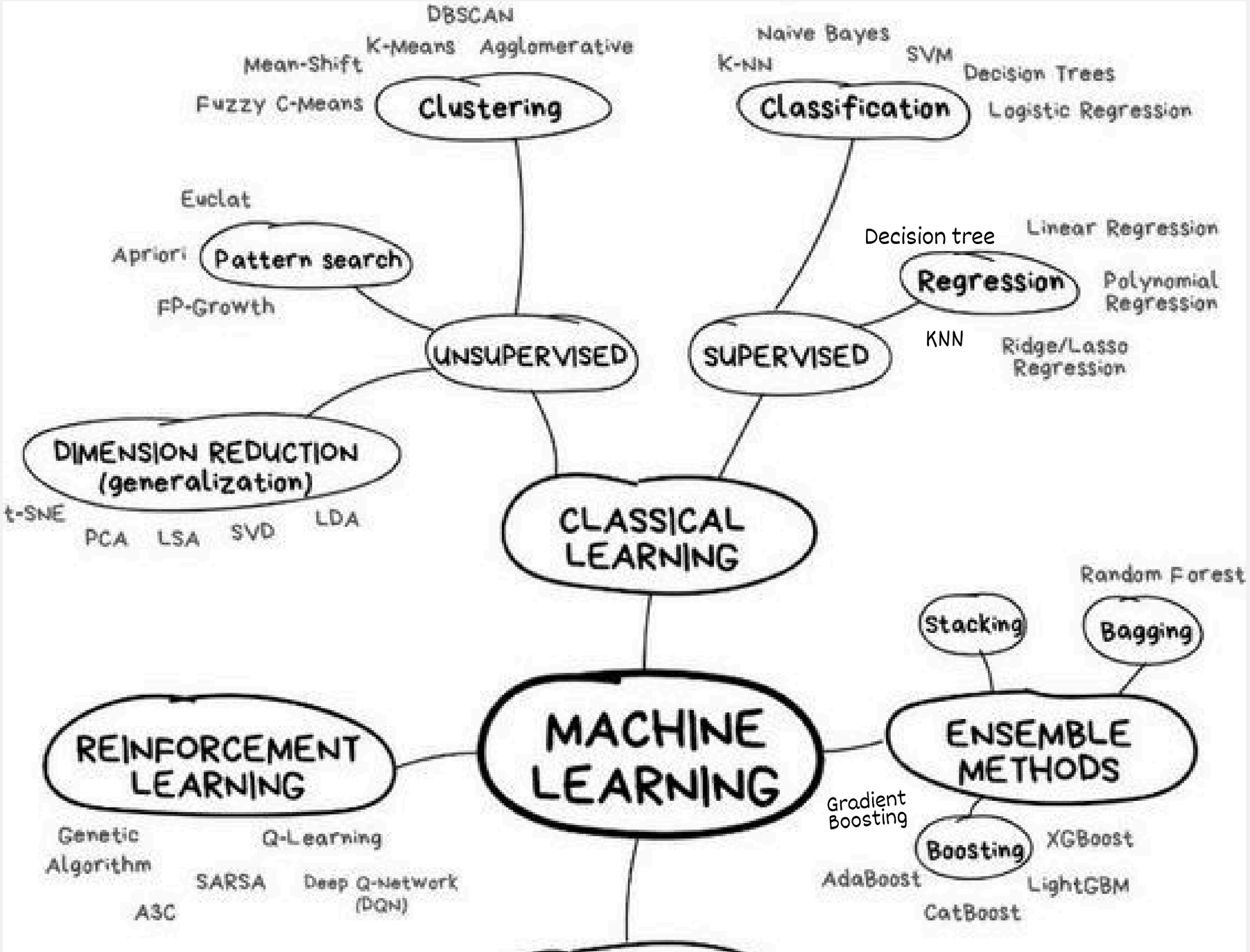

Learn SQL in 5 minutes! (Can you?)
Copy link
learning
SQL
in 5 minutes
?
1SELECT COUNT(*) FROM
2(SELECT COUNT(*) FROM
3(SELECT food_type, COUNT(*)
4FROM zoo
5LEFT JOIN food
6 ON food.animal = zoo.animal
7GROUP BY food_type
Watch on YouTube

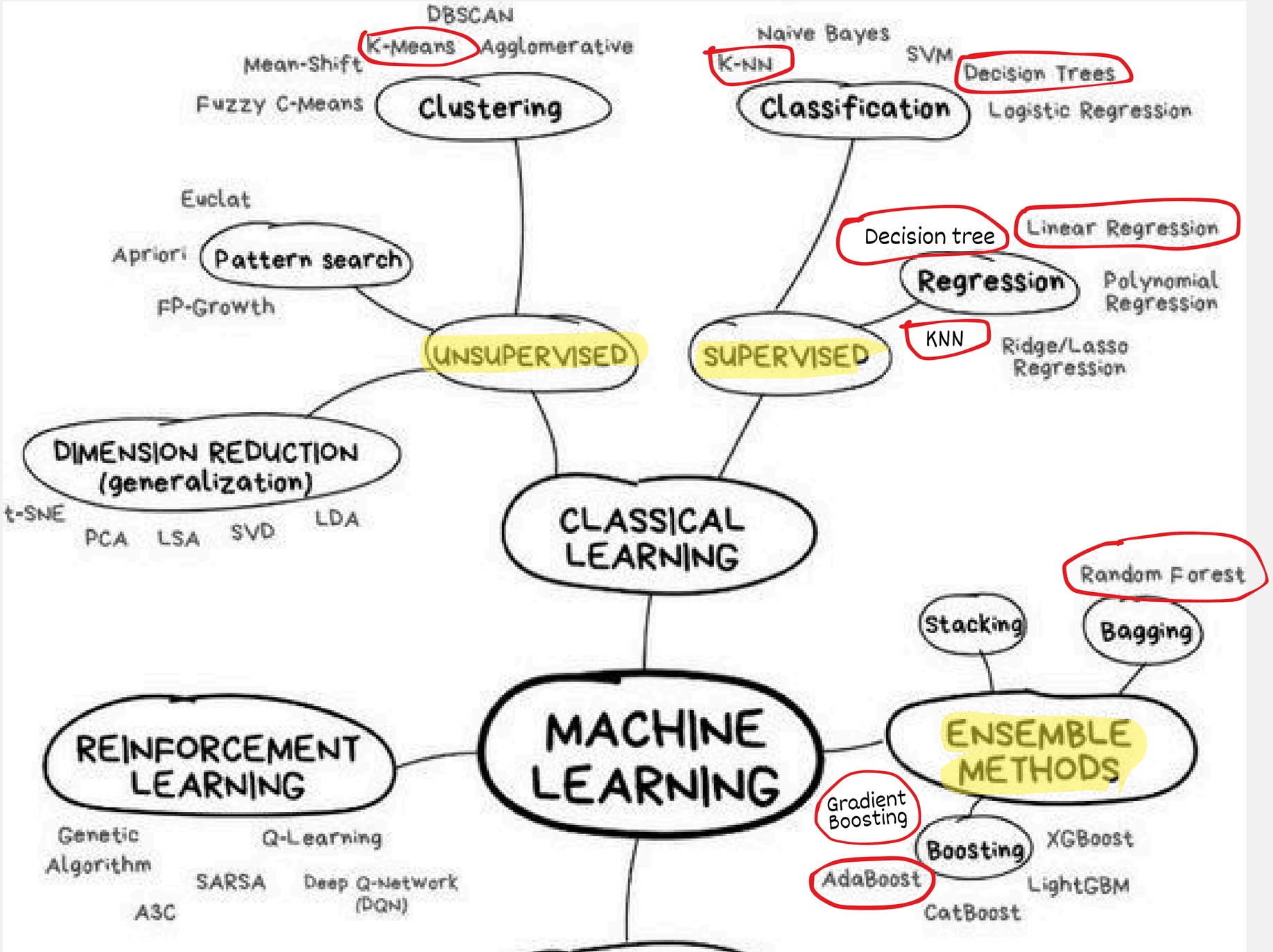
Common Algorithm

= a set of **mathematical rules** and steps that a model use to learn pattern from data




MACHINE LEARNING
CLASSICAL LEARNING
UNSUPERVISED
Clustering
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Pattern search
Euclat
Apriori
FP-Growth
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
SUPERVISED
Classification
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Regression
Decision tree
Linear Regression
Polynomial Regression
KNN
Ridge/Lasso Regression
REINFORCEMENT LEARNING
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C
ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
Gradient Boosting
XGBoost
AdaBoost
LightGBM
CatBoost
NEURAL NETS AND DEEP LEARNING
Convolutional Neural Networks (CNN)
DCNN
Perceptrons (MLP)
Recurrent Neural Networks (RNN)
LSM
LSTM
GRU
Generative Adversarial Networks (GAN)
Autoencoders
seq2seq


DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Clustering
Naive Bayes
SVM
K-NN
Decision Trees
Classification
Logistic Regression
Euclat
Apriori
Pattern search
FP-Growth
Decision tree
Linear Regression
Regression
Polynomial Regression
KNN
Ridge/Lasso Regression
UNSUPERVISED
SUPERVISED
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
CLASSICAL LEARNING
Random Forest
Stacking
Bagging
REINFORCEMENT LEARNING
MACHINE LEARNING
ENSEMBLE METHODS
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C
Gradient Boosting
Boosting
XGBoost
AdaBoost
LightGBM
CatBoost


MACHINE LEARNING
CLASSICAL LEARNING
UNSUPERVISED
SUPERVISED
Clustering
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Classification
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Pattern search
Euclat
Apriori
FP-Growth
Regression
Decision tree
Linear Regression
Polynomial Regression
KNN
Ridge/Lasso Regression
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
Gradient Boosting
XGBoost
AdaBoost
LightGBM
CatBoost
REINFORCEMENT LEARNING
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C

What is Supervised Learning

(Recap)

- Train with **labelled data**
  - the dataset has the correct answers.
- Learn to predict outcomes (numbers or categories).

Types of Outcome

(for Supervised Learning)

Regression

(Predict Number)

Remaining lease,
Floor area,
Location ,
etc

HDB resale price

Classification

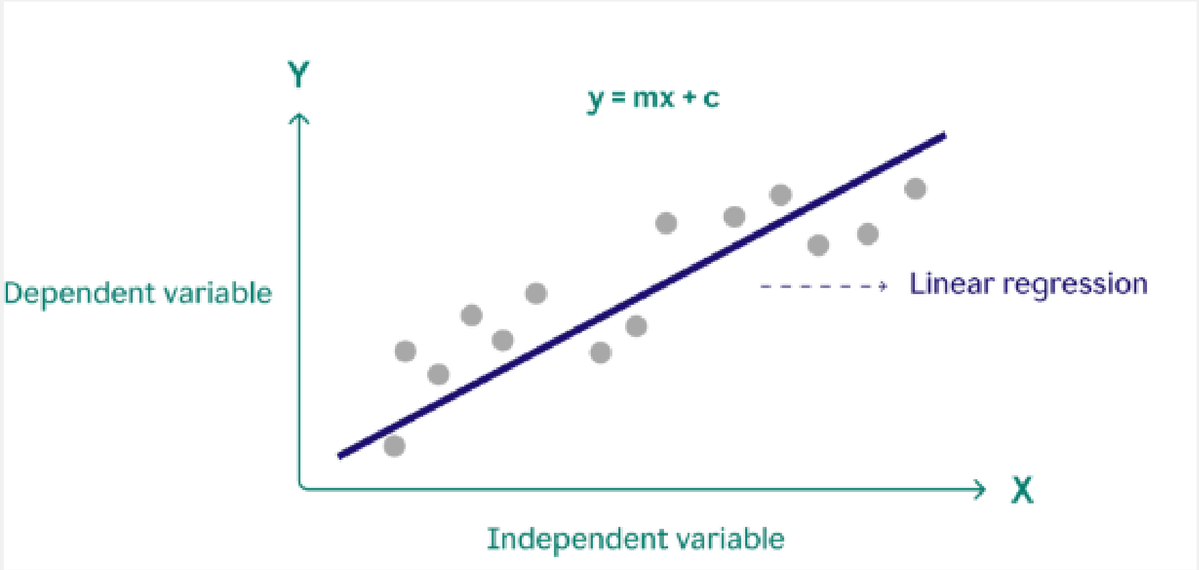
(Predict Category)

Hours studied,
Attendance rate,
Assignment score,
etc

Pass
or
Fail

Linear Regression

(for Supervised Learning)



Strengths	Weaknesses
Simple and easy to interpret	Struggles with non-linear patterns
Works well when the relationship is linear	Sensitive to outliers

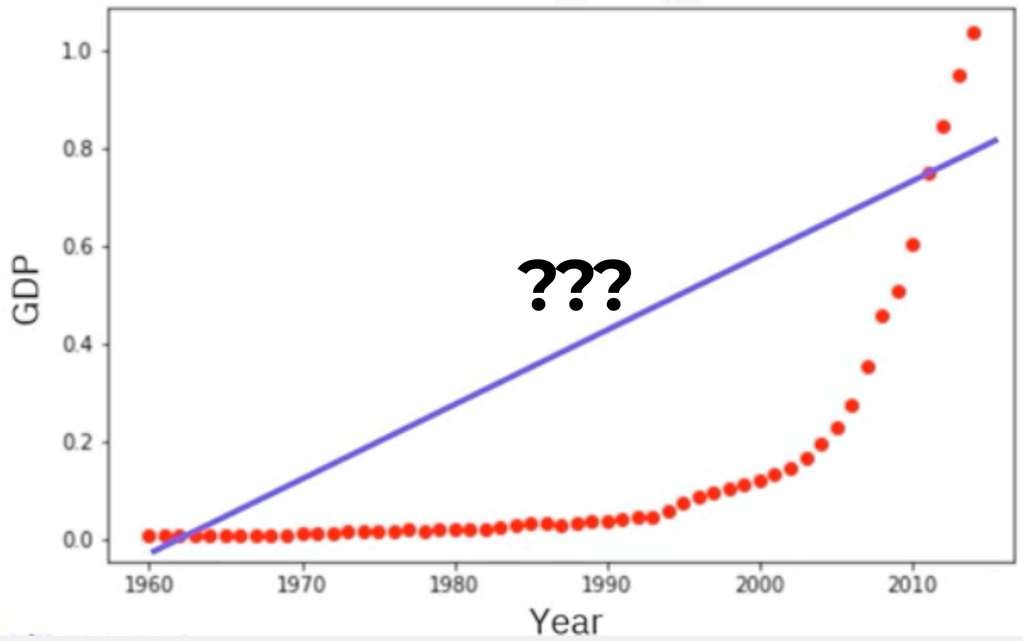
Definition: A model that predicts a continuous value based on the relationship between input features and the target.

Use Case: Predicting prices, trends (e.g., house price prediction).

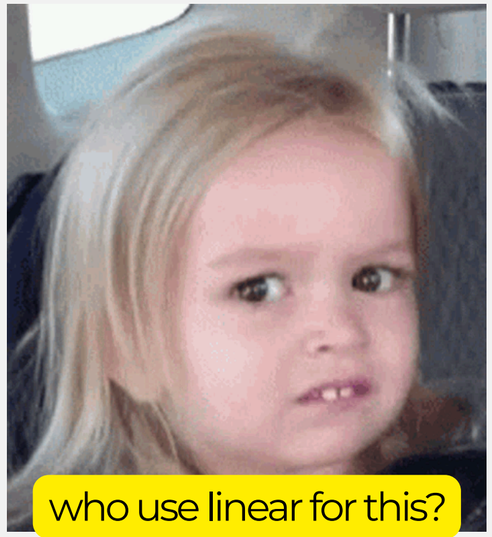
IRL, its not as simple as linear..

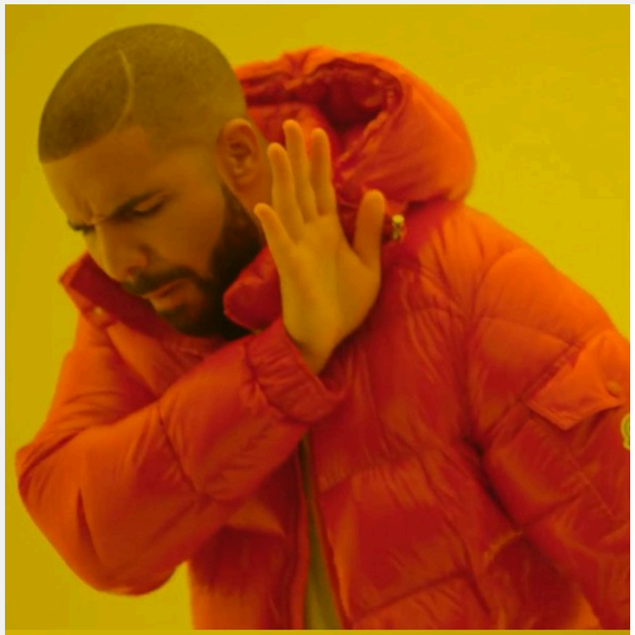
Compound growth example

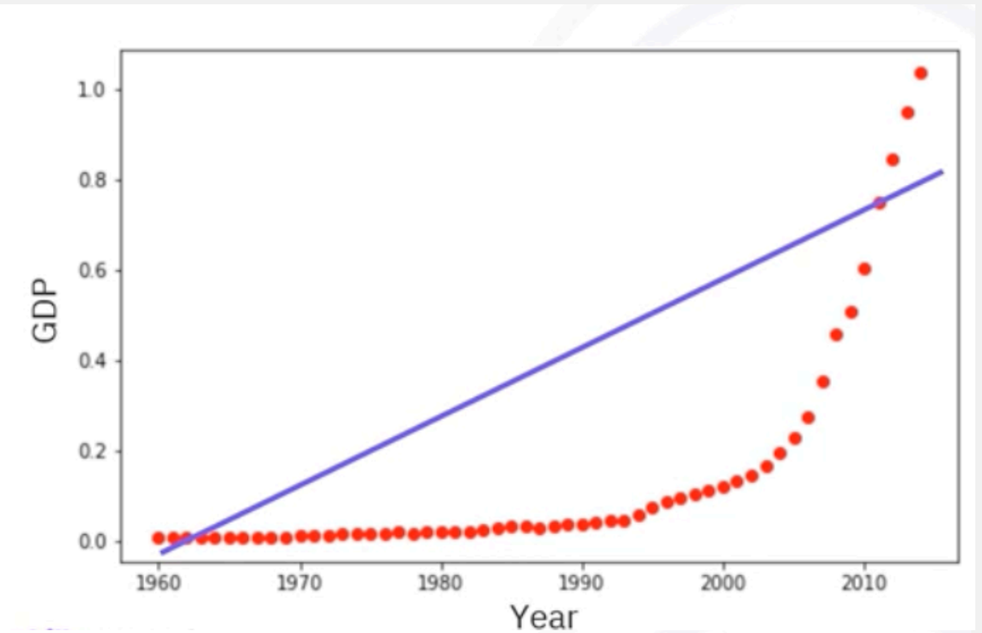
- Scatterplot displays strong dependence of GDP on time, but relationship is nonlinear

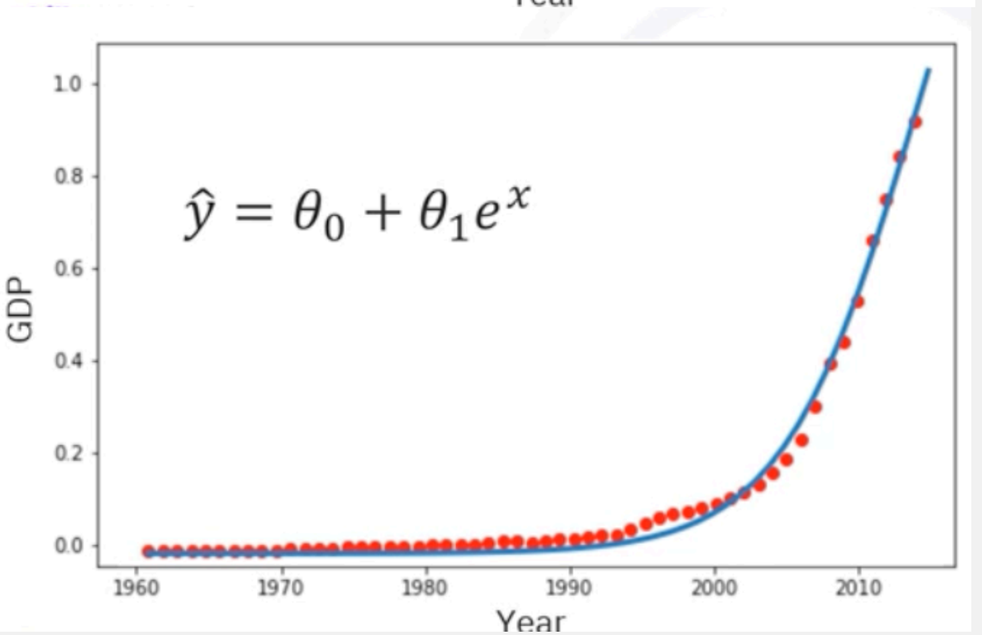


	Year	GDP (B USD)
0	1960	59.2
1	1961	49.6
2	1962	46.7
3	1963	50.1
4	1964	59.1
5	1965	69.8
6	1966	75.9
7	1967	72.1
...	.....	.....






GDP
Year
1.0
0.8
0.6
0.4
0.2
0.0
1960
1970
1980
1990
2000
2010


$\hat{y} = \theta_0 + \theta_1 e^x$
GDP
Year
1.0
0.8
0.6
0.4
0.2
0.0
1960
1970
1980
1990
2000
2010

Non-Linear Regression

(an extension of Linear Regression)

Polynomial Regression

- Quadratic – 2nd order

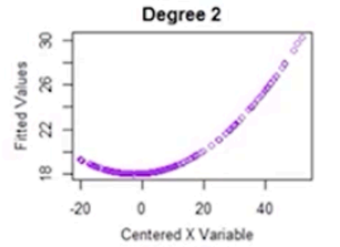
$$\hat{Y} = b_0 + b_1 x_1 + b_2 (x_1)^2$$

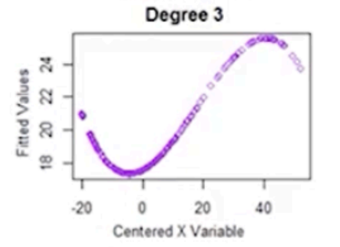
- Cubic – 3rd order

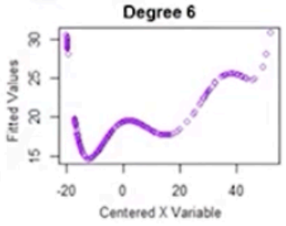
$$\hat{Y} = b_0 + b_1 x_1 + b_2 (x_1)^2 + b_3 (x_1)^3$$

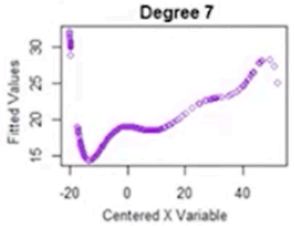
- Higher order

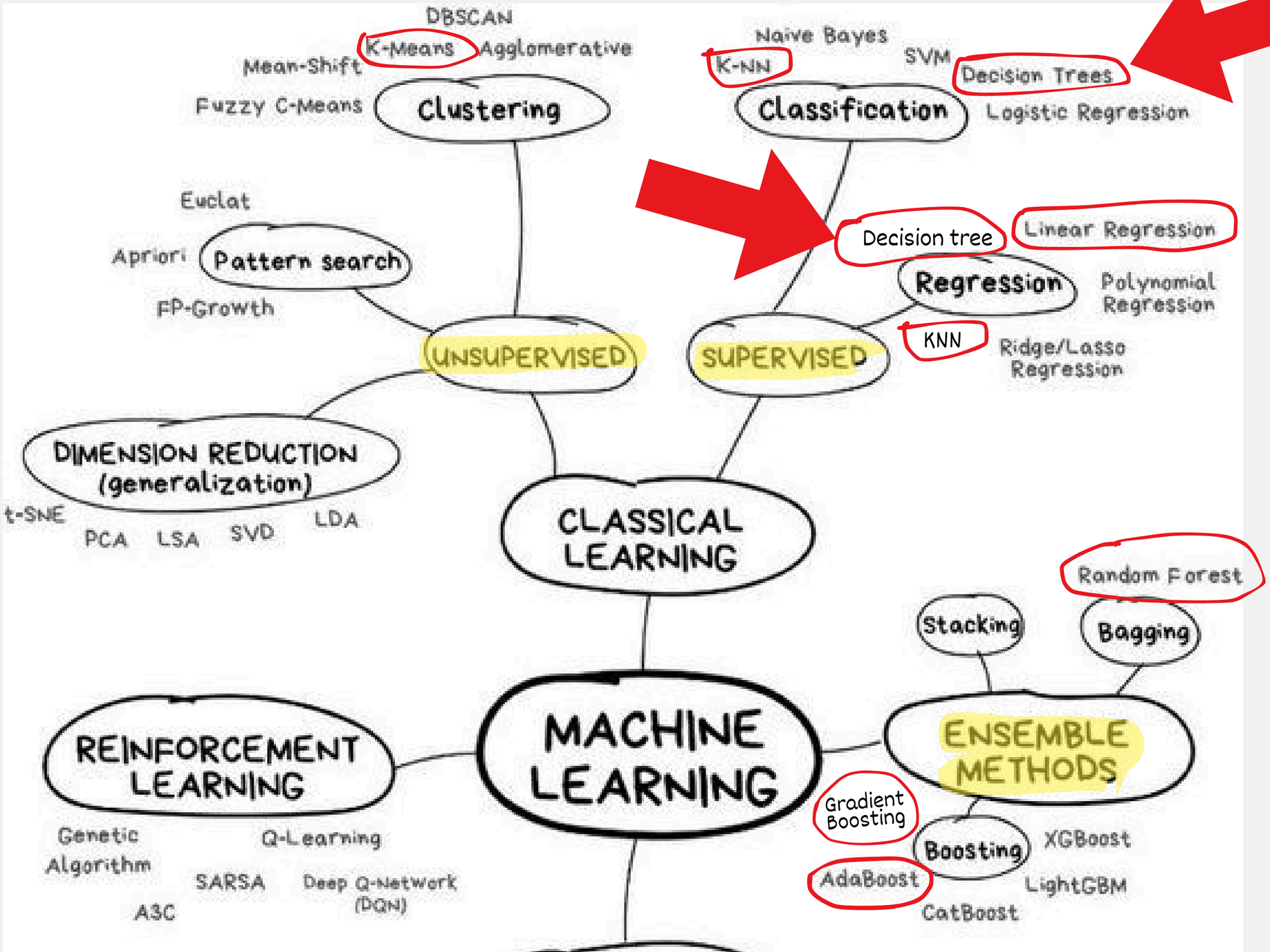
$$\hat{Y} = b_0 + b_1 x_1 + b_2 (x_1)^2 + b_3 (x_1)^3 + ..$$





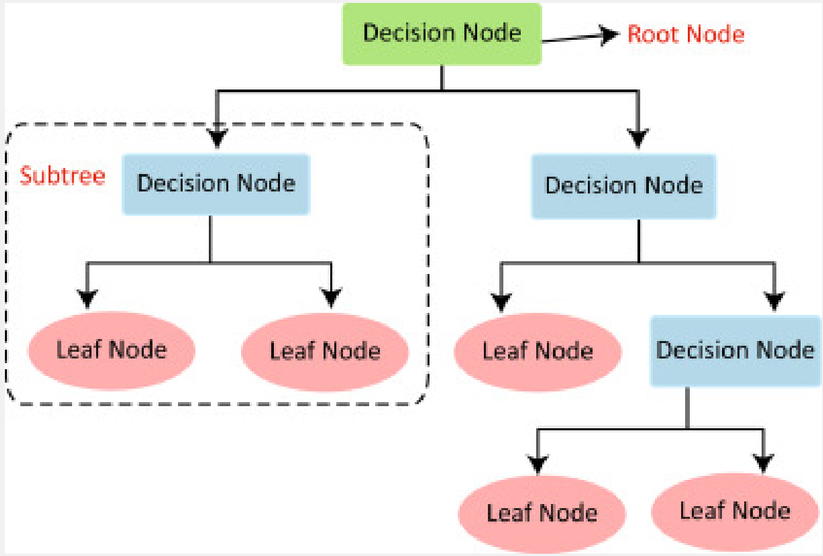





DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Clustering
Naive Bayes
K-NN
SVM
Decision Trees
Classification
Logistic Regression
Euclat
Apriori
Pattern search
FP-Growth
UNSUPERVISED
Decision tree
Linear Regression
Regression
Polynomial Regression
SUPERVISED
KNN
Ridge/Lasso Regression
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
CLASSICAL LEARNING
Random Forest
Stacking
Bagging
MACHINE LEARNING
REINFORCEMENT LEARNING
ENSEMBLE METHODS
Gradient Boosting
Boosting
XGBoost
AdaBoost
LightGBM
CatBoost
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C

Decision tree

(For regression and classification)



Strengths	Weaknesses
Easy to understand and visualize	Prone to overfitting
Can handle both numerical and categorical data	Small changes in data can lead to a completely different tree

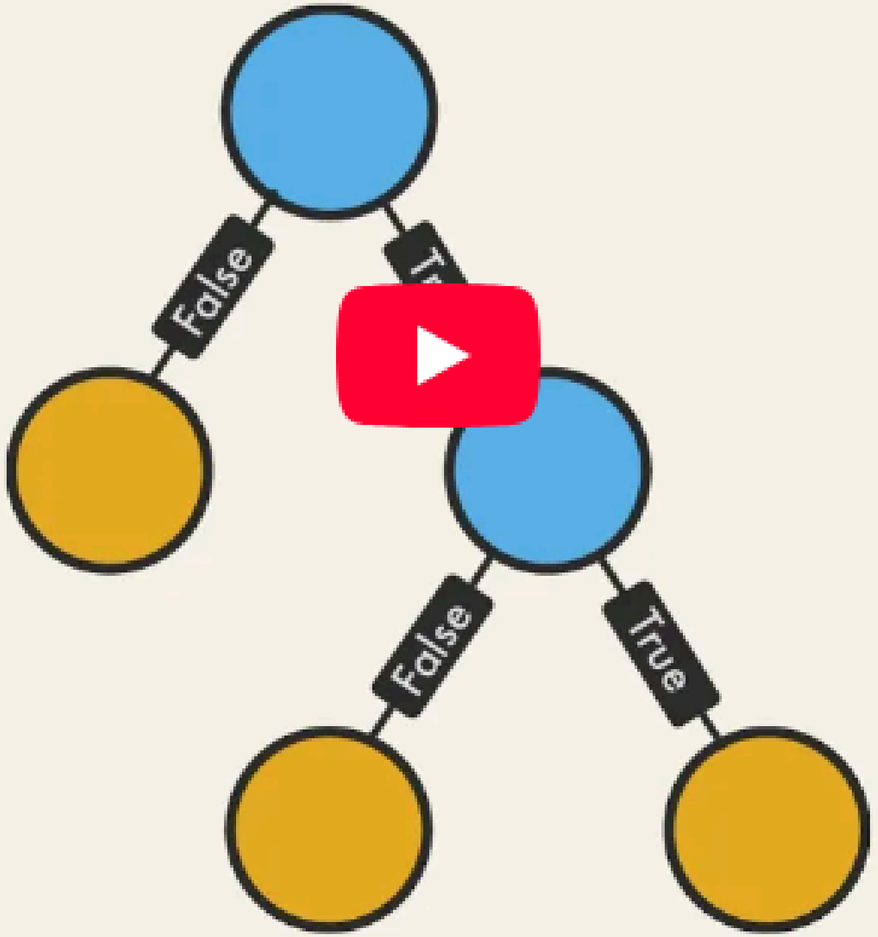
Definition: A tree-based model that splits the data based on feature values to make decisions.

Use Case: Classification problems like determining whether an email is spam or not.

Decision Tree: Important things to know

Copy link

Decision Tree



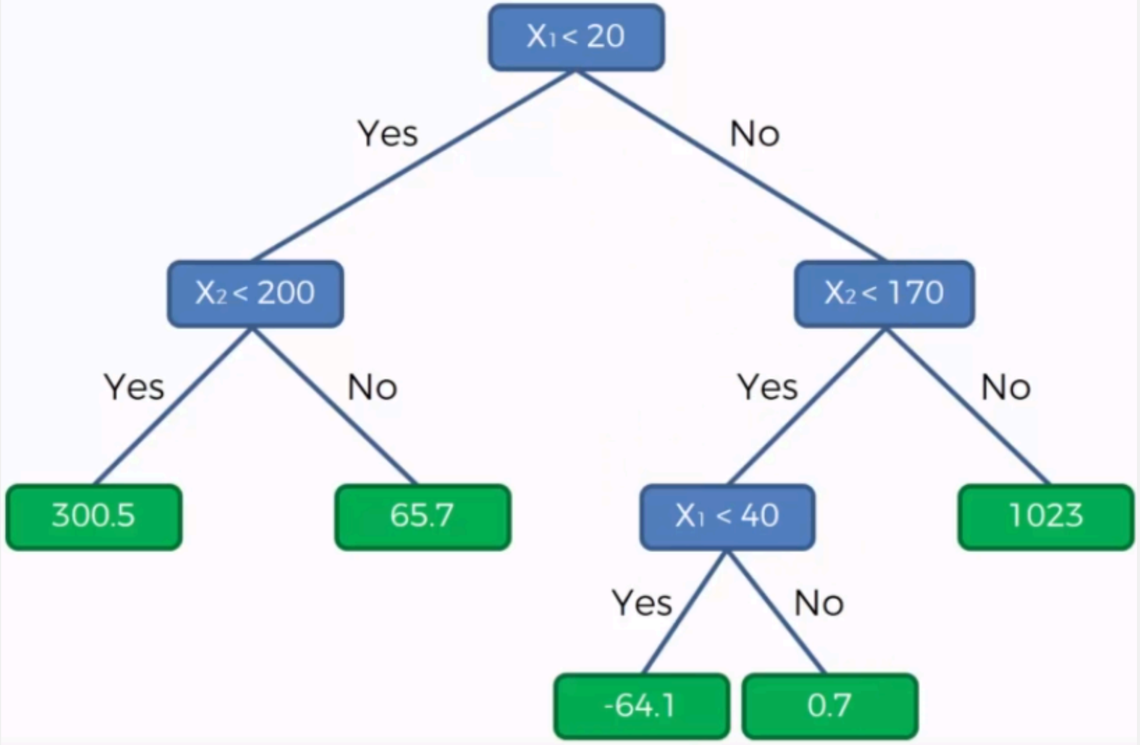
Watch on YouTube

Types of Decision tree

(for Supervised Learning)

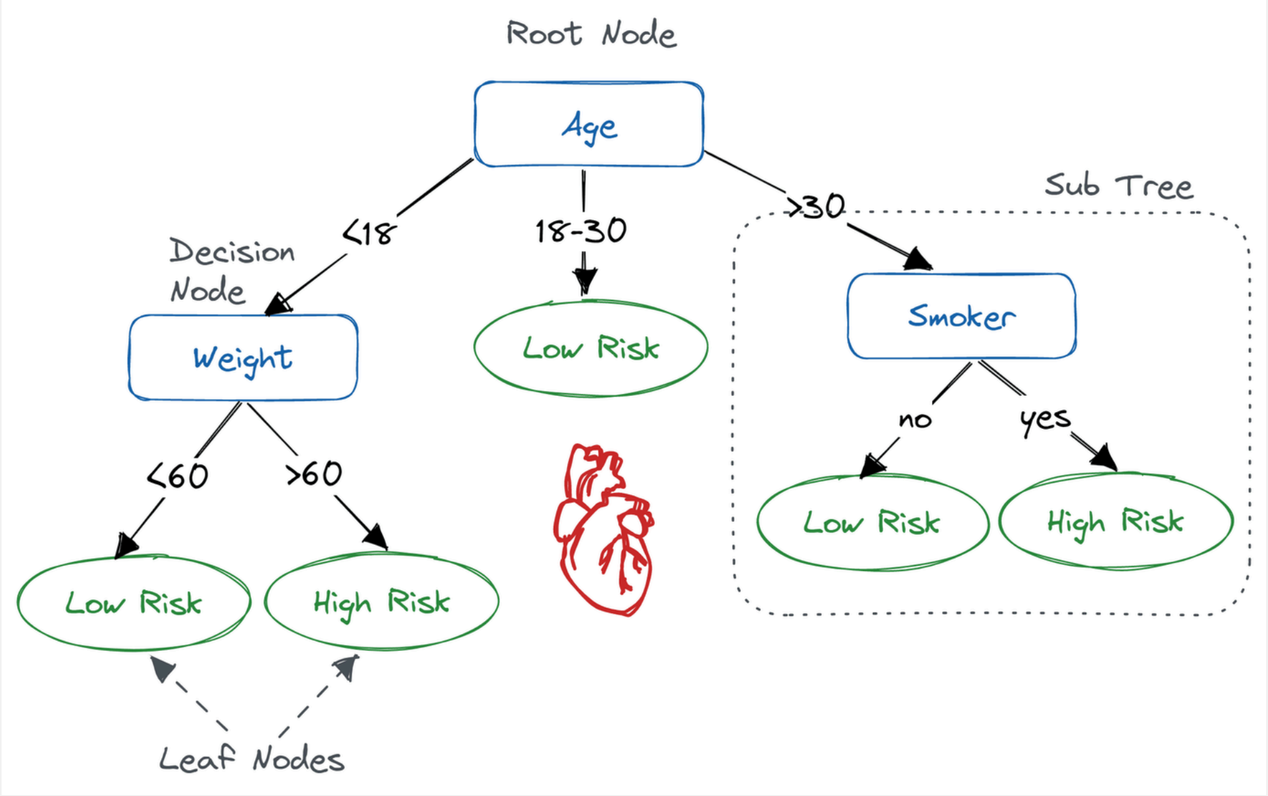
Regressor

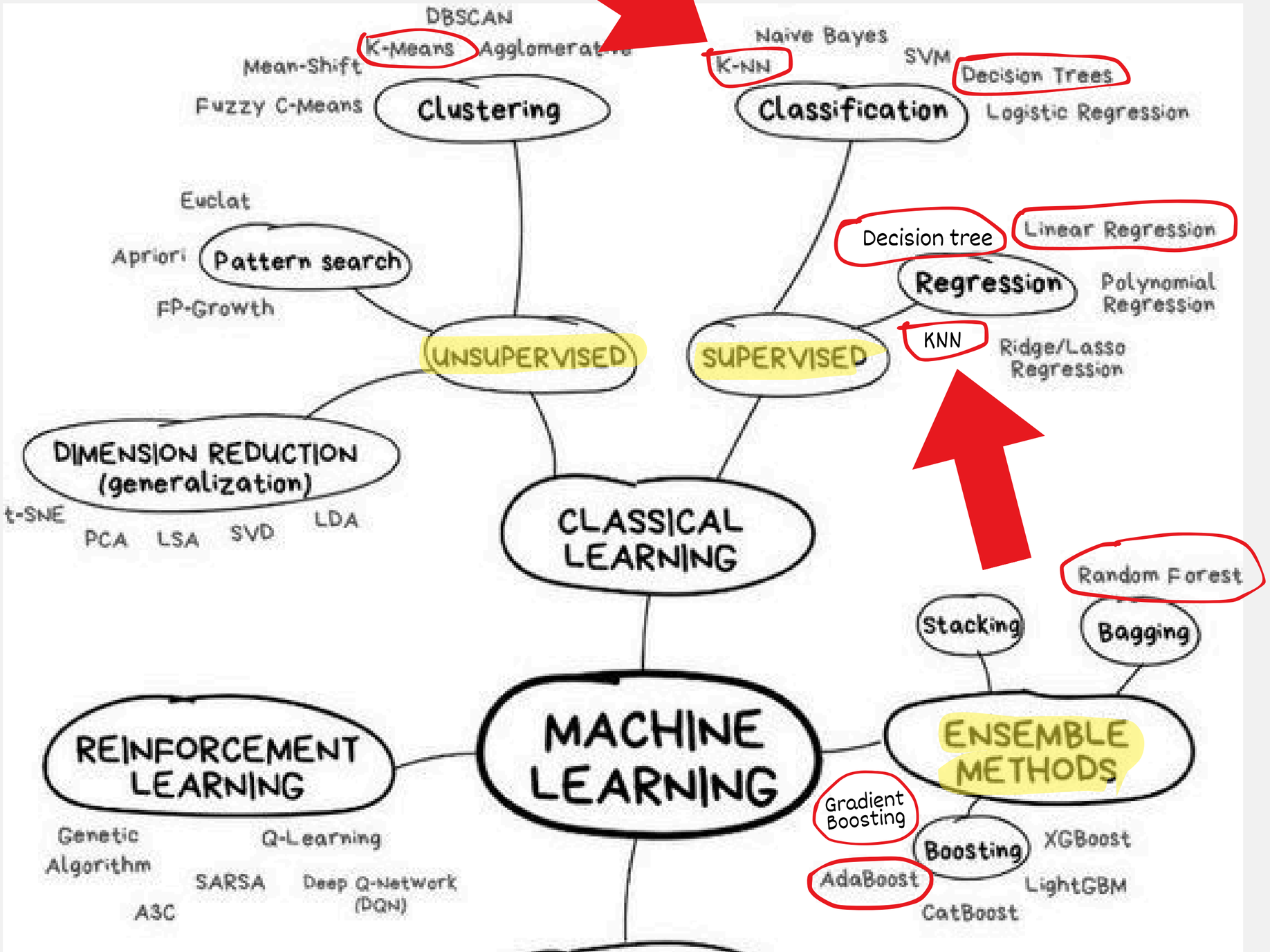
(Predict Number)



Classifier

(Predict Category)




DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Clustering
Naive Bayes
K-NN
SVM
Decision Trees
Classification
Logistic Regression
Euclat
Apriori
Pattern search
FP-Growth
UNSUPERVISED
SUPERVISED
Decision tree
Linear Regression
Regression
Polynomial Regression
KNN
Ridge/Lasso Regression
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
CLASSICAL LEARNING
Random Forest
Stacking
Bagging
MACHINE LEARNING
REINFORCEMENT LEARNING
ENSEMBLE METHODS
Gradient Boosting
Boosting
XGBoost
AdaBoost
LightGBM
CatBoost
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C

K-Nearest Neighbours (K-NN)

(for Supervised Learning)

- It predicts based on the labels of its nearest neighbors in the training data.

Strengths	Weaknesses
No training phase (just stores data)	Sensitive to irrelevant features or different scales
Works for both classification and regression.	Slow for large datasets (must compare to all points)

Use Case: “Users who liked this movie also liked…”

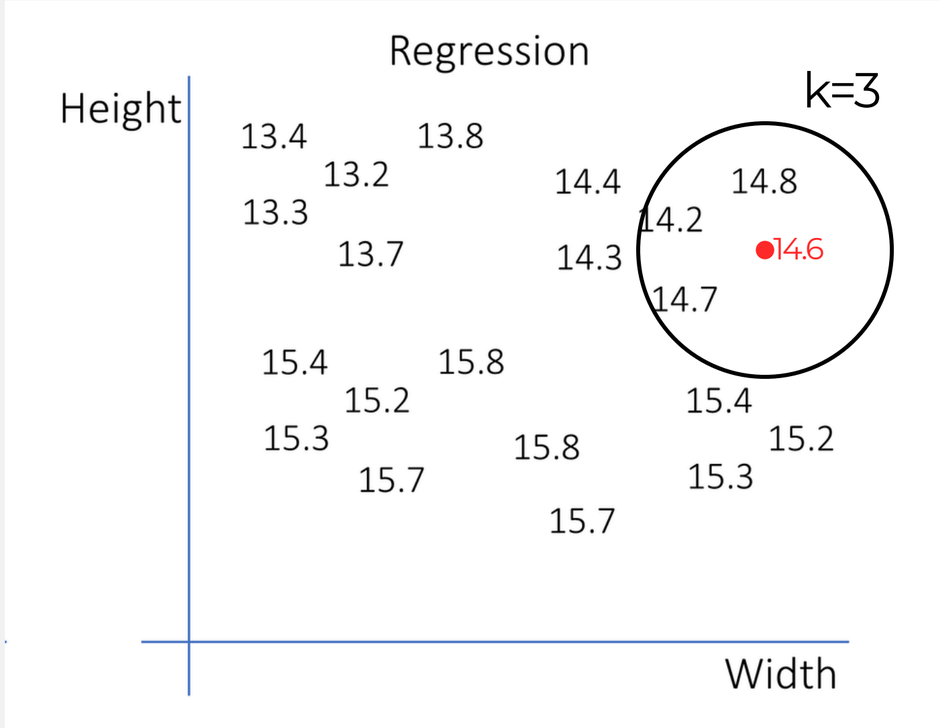
- KNN finds users with similar viewing histories 🎬

Types of KNN

(for Supervised Learning)

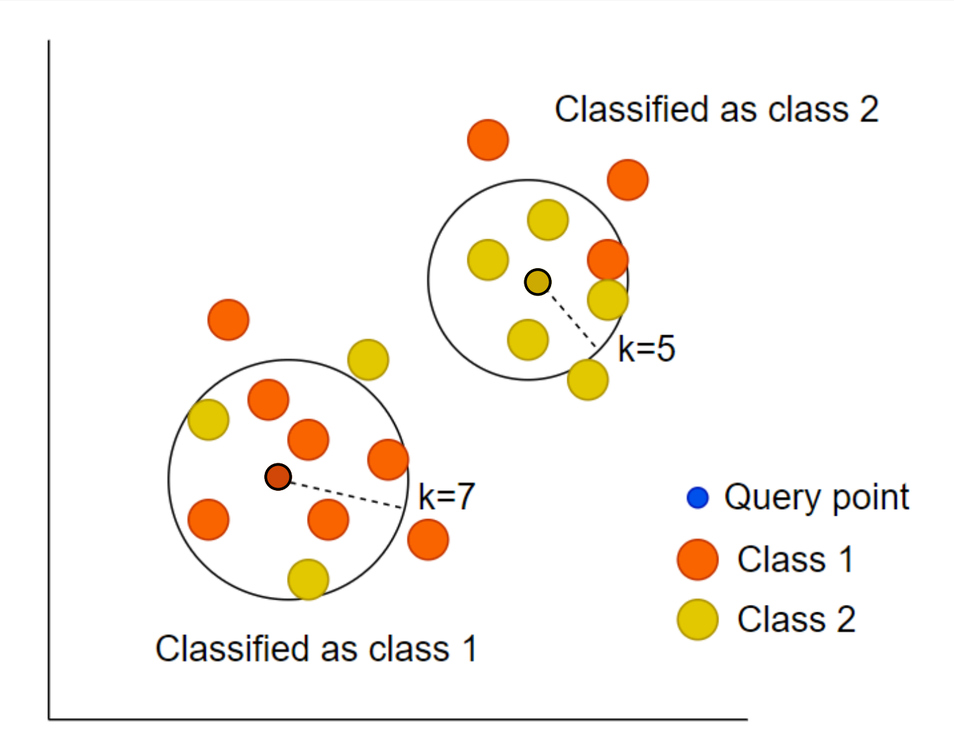
Regressor

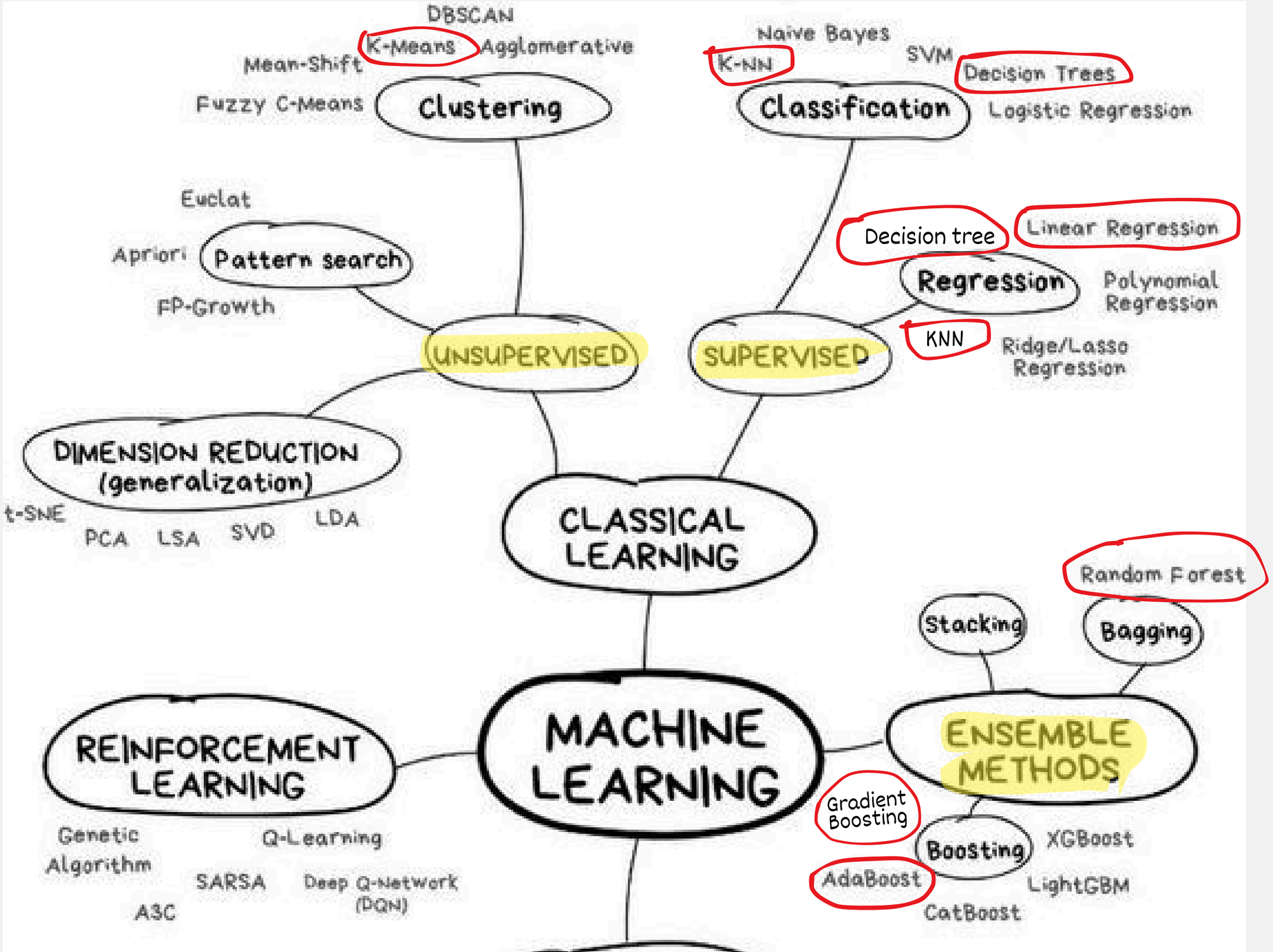
average (or median) of neighbors' values



Classifier

by majority vote of neighbors' labels



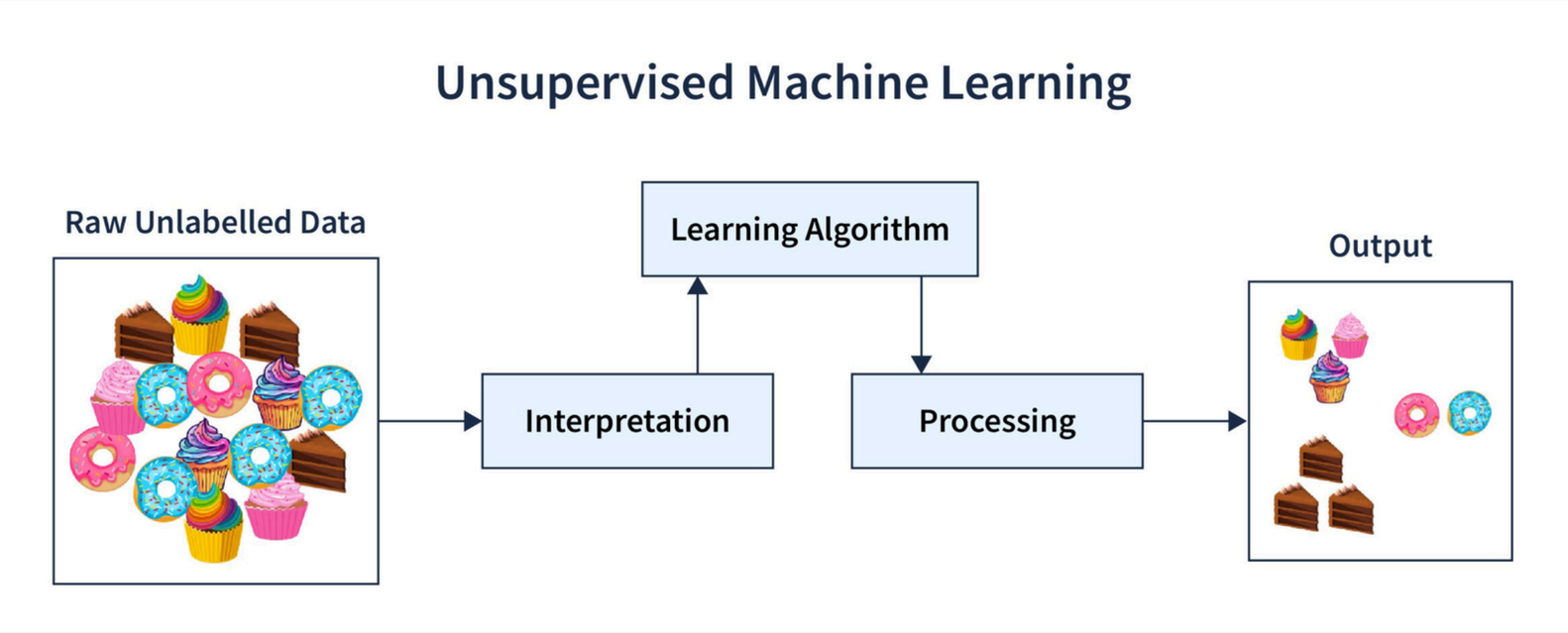

MACHINE LEARNING
CLASSICAL LEARNING
UNSUPERVISED
SUPERVISED
Clustering
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Classification
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Pattern search
Euclat
Apriori
FP-Growth
Regression
Decision tree
Linear Regression
Polynomial Regression
KNN
Ridge/Lasso Regression
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
Gradient Boosting
XGBoost
AdaBoost
LightGBM
CatBoost
REINFORCEMENT LEARNING
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C

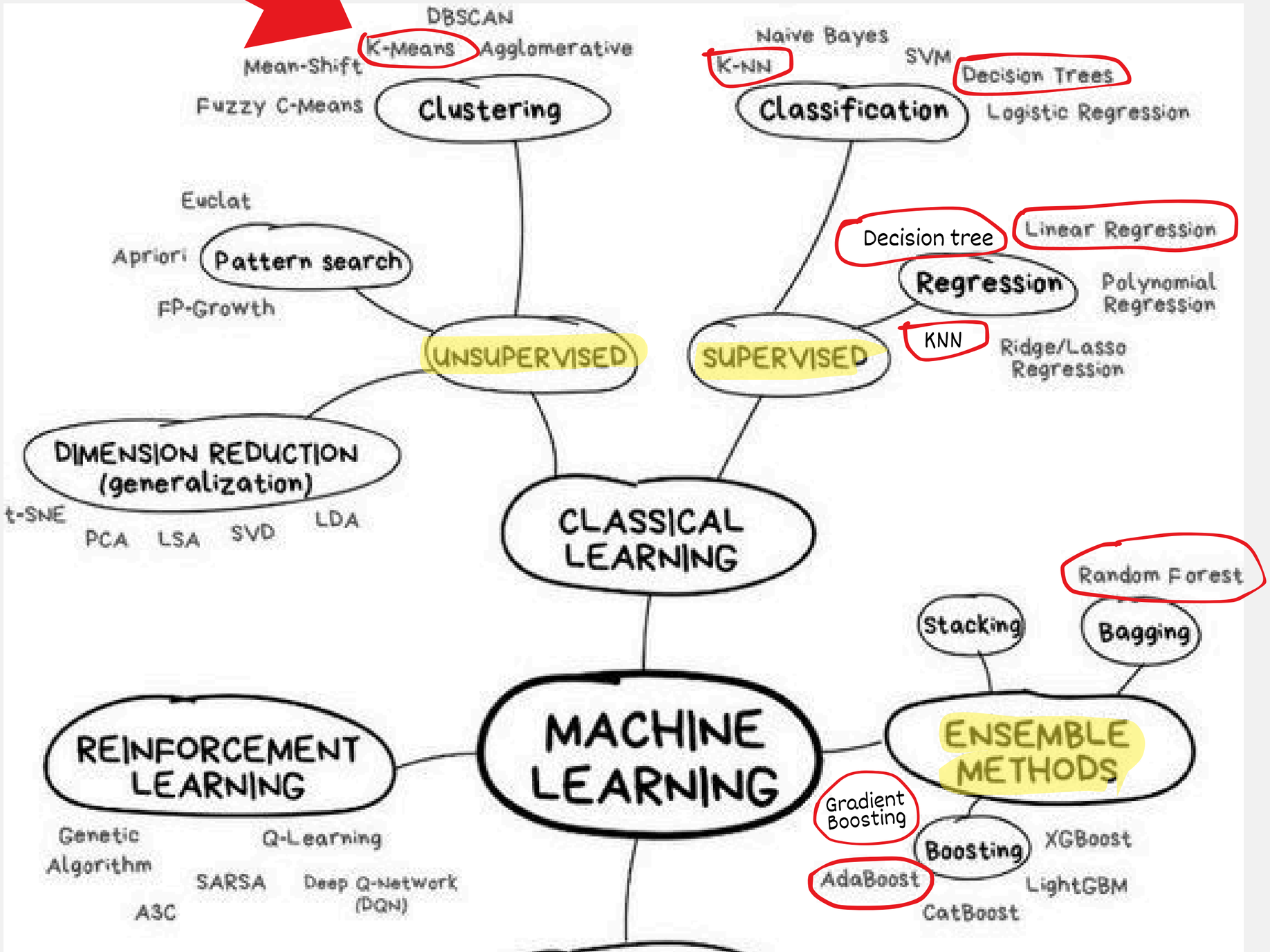
What is Unsupervised Learning

(Recap)

- You train with unlabelled data
  - There are no given answers.

- Goal: find hidden patterns or group similar data.




MACHINE LEARNING
CLASSICAL LEARNING
UNSUPERVISED
SUPERVISED
Clustering
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Classification
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Pattern search
Euclat
Apriori
FP-Growth
Regression
Decision tree
Linear Regression
Polynomial Regression
KNN
Ridge/Lasso Regression
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
REINFORCEMENT LEARNING
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C
ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
Gradient Boosting
XGBoost
AdaBoost
LightGBM
CatBoost

Classification

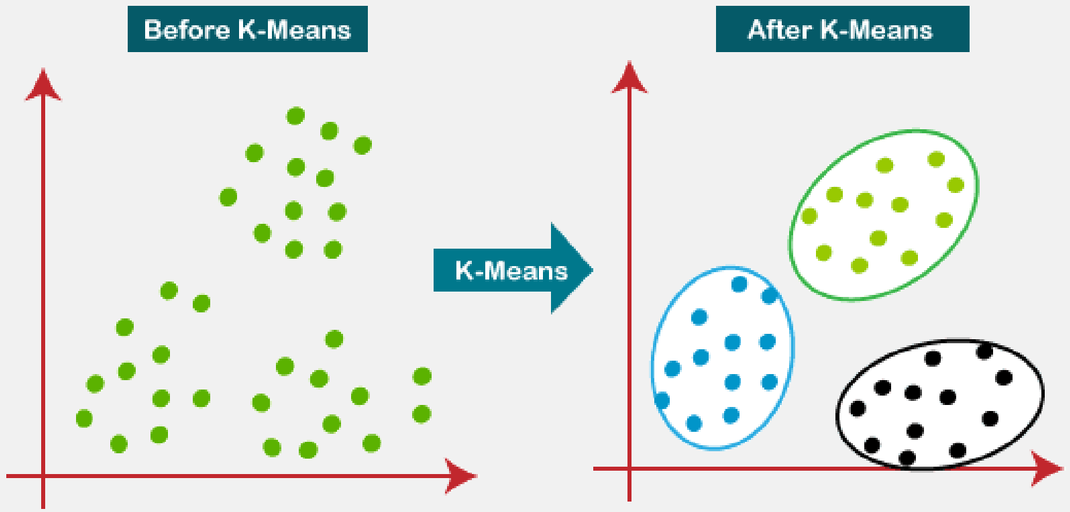
- You already know what categories exist
  - You train your model to recognize them.

Clustering

- You don’t know the categories
  - You just want to group similar items.

K-Means Clustering

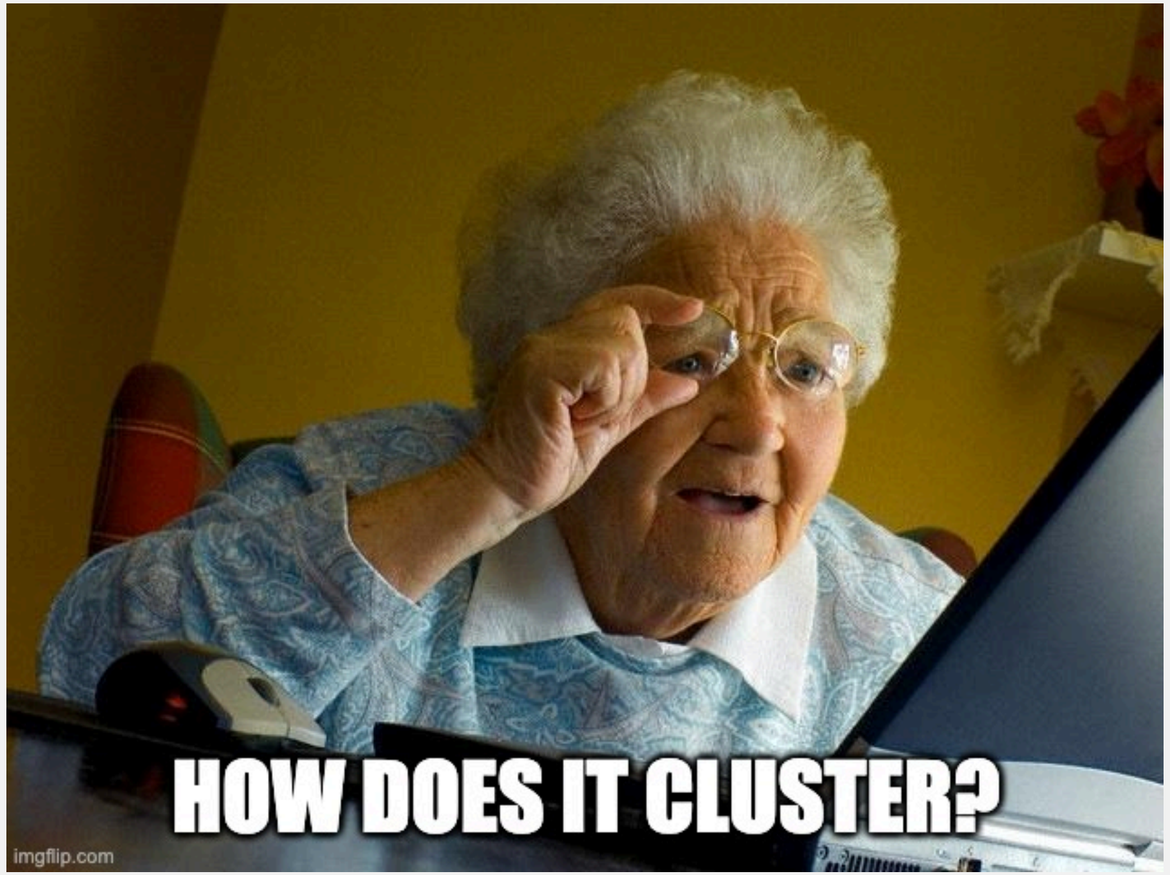
(for Unsupervised learning)



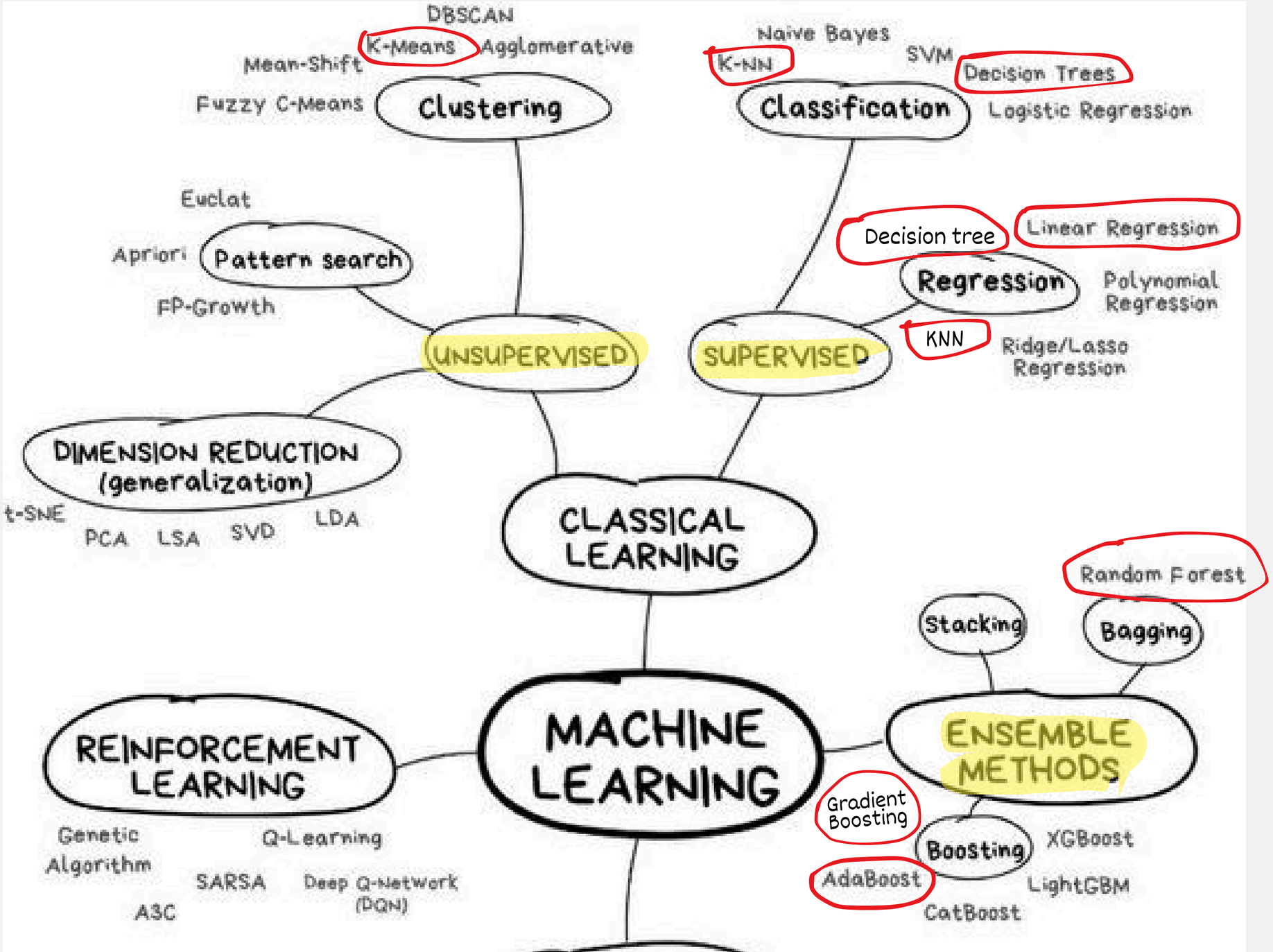
Strengths	Weaknesses
Fast and efficient for large datasets	Requires choosing K manually
Easy to understand and implement	Doesn't work well with non-spherical clusters or different cluster sizes

Definition: An unsupervised algorithm that groups data into K clusters based on similarity.

Use Case: Customer segmentation, grouping similar users or products.

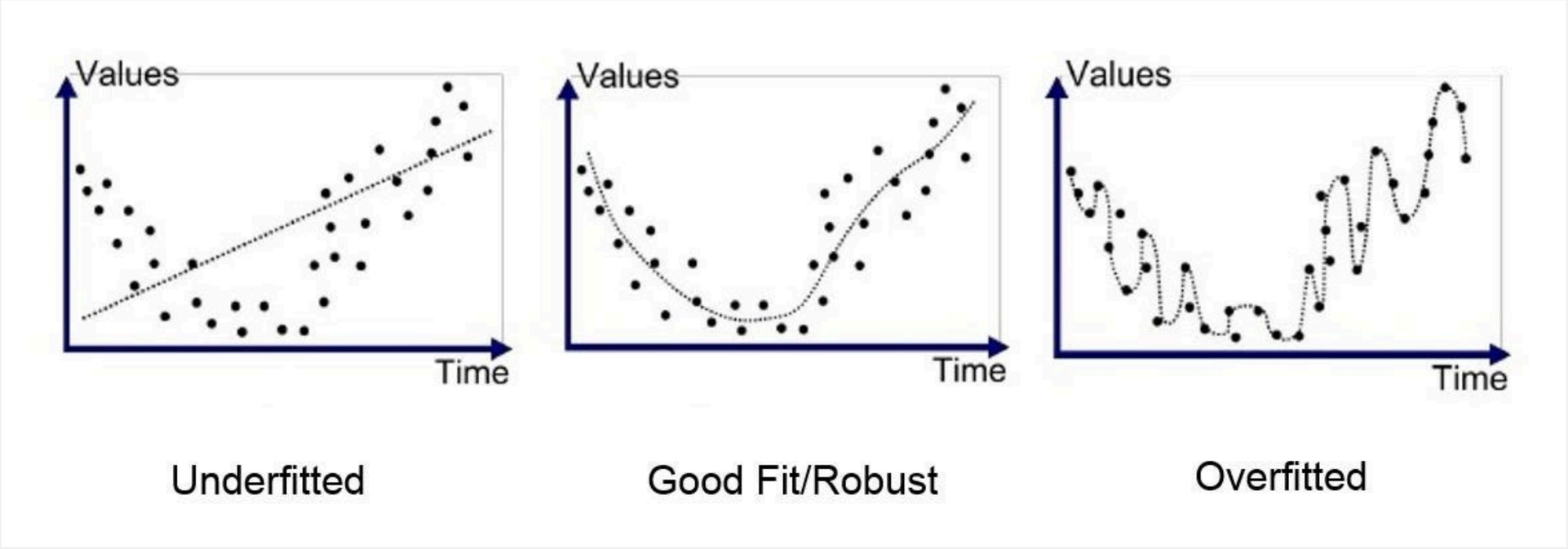


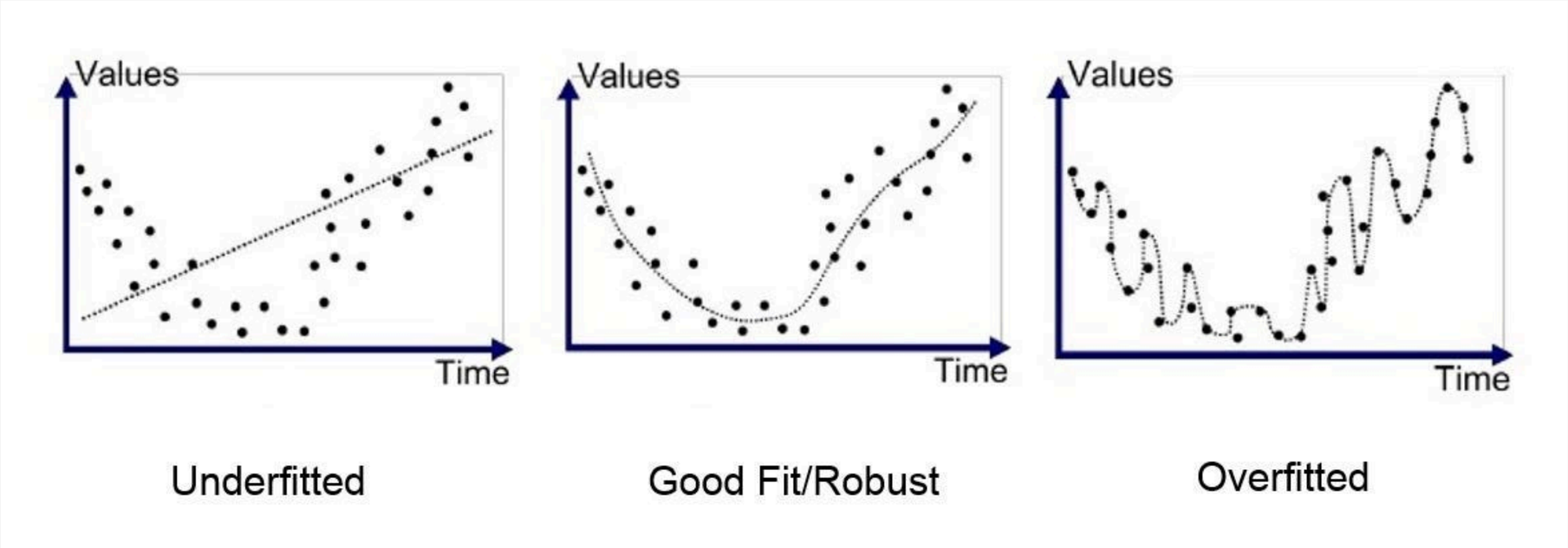
https://www.naftaliharris.com/blog/visualizing-k-means-clustering/


MACHINE LEARNING
CLASSICAL LEARNING
UNSUPERVISED
SUPERVISED
Clustering
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Classification
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Pattern search
Euclat
Apriori
FP-Growth
Regression
Decision tree
Linear Regression
Polynomial Regression
KNN
Ridge/Lasso Regression
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
Gradient Boosting
XGBoost
AdaBoost
LightGBM
CatBoost
REINFORCEMENT LEARNING
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C

Common Problem

Underfitting vs Overfitting





What it is	Too basic to capture patterns	Captures true underlying patterns	Memorizes training data, not general patterns
Example	Straight line for curved house prices	Student understands and applies concepts	Student memorizes pyp answers but fails on exam
Signs	Poor performance on both train/test	Good on both train/test	Great on training, bad on new data
Symptoms	High bias, low variance	Balanced bias and variance	Low bias, high variance
Solution	Use more complex model or features	Balance model complexity and data quality	Use simpler model, more data, or regularization

How to Determine?

Case 1:
Training accuracy: 60%
Test accuracy: 58%

low training accuracy
low test accuracy

= Underfitting

Case 2:
Training accuracy: 98%
Test accuracy: 65%

high training accuracy
low test accuracy

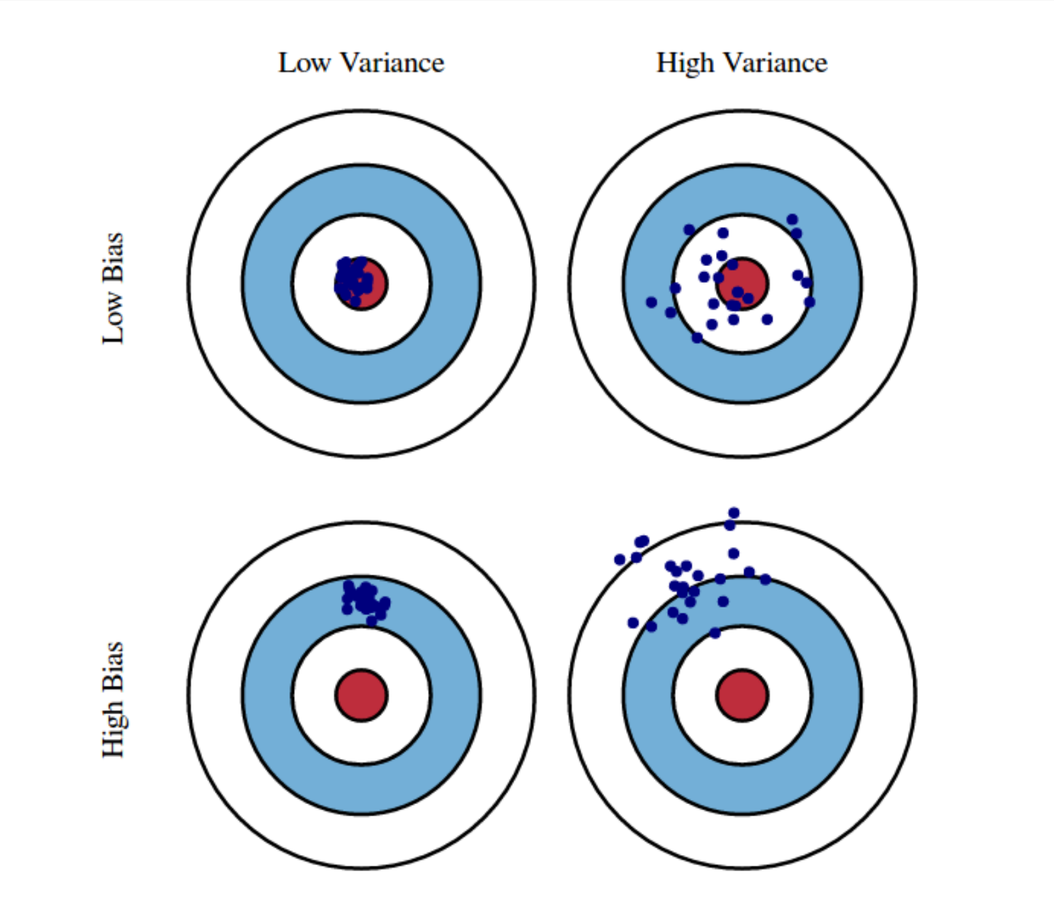
= Overfitting

Case 3:
Training accuracy: 90%
Test accuracy: 90%

high training accuracy
high test accuracy

= Good fit ✅

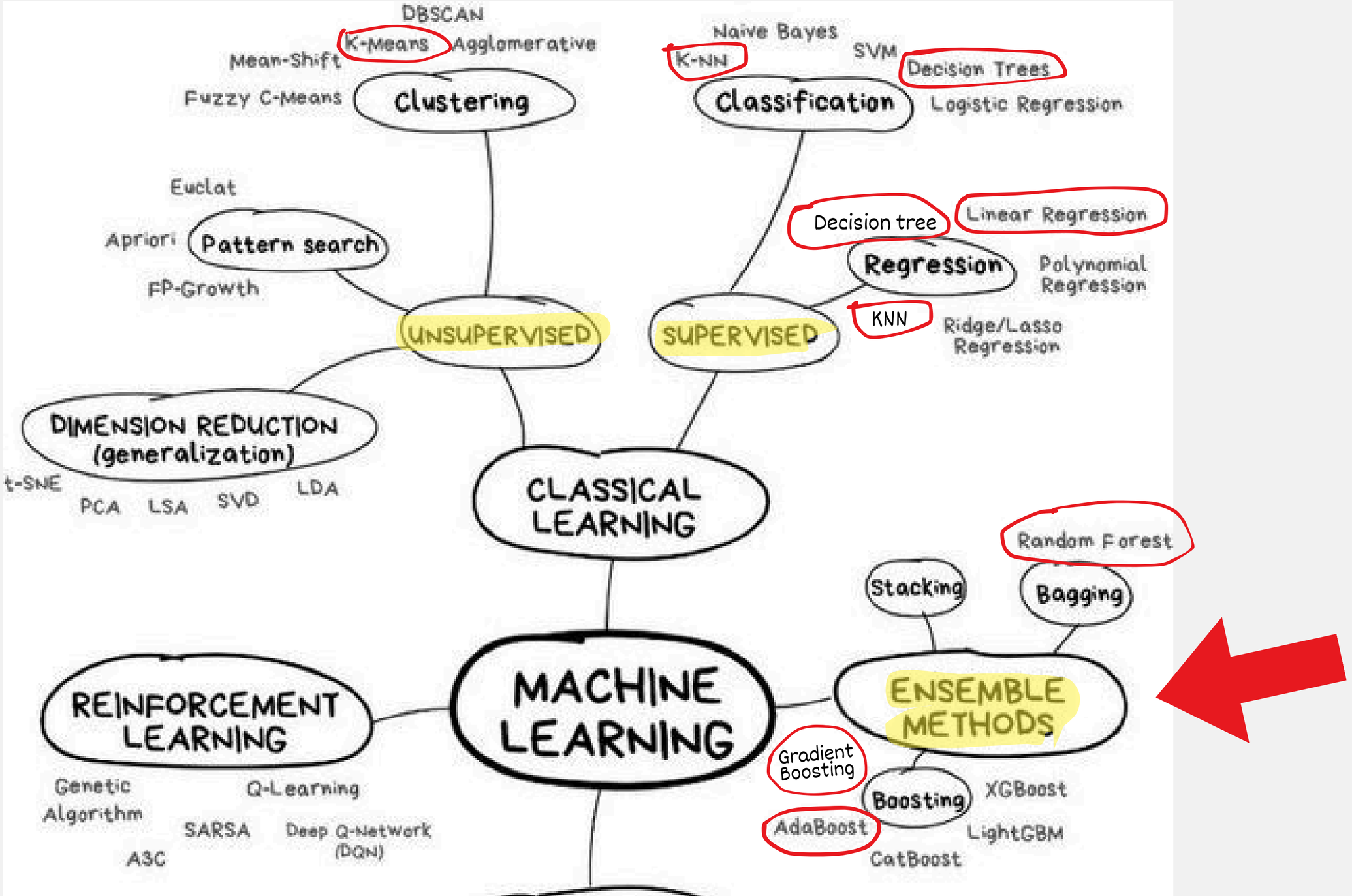
Bias-Variance Tradeoff



Bias: Error due to overly simple assumptions (underfitting)

Variance: Error due to too much complexity (overfitting)

Goal: Find a balance where both are low


MACHINE LEARNING
CLASSICAL LEARNING
UNSUPERVISED
SUPERVISED
REINFORCEMENT LEARNING
ENSEMBLE METHODS
Clustering
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Classification
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Pattern search
Euclat
Apriori
FP-Growth
Regression
Decision tree
Linear Regression
Polynomial Regression
KNN
Ridge/Lasso Regression
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
Stacking
Bagging
Random Forest
Boosting
Gradient Boosting
XGBoost
AdaBoost
LightGBM
CatBoost
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C

Ensemble Methods

(Combine models to improve results)

⚙️ How It Works

- You train multiple models then combine their outputs by:
  - Averaging (for regression)
  - Voting (for classification)

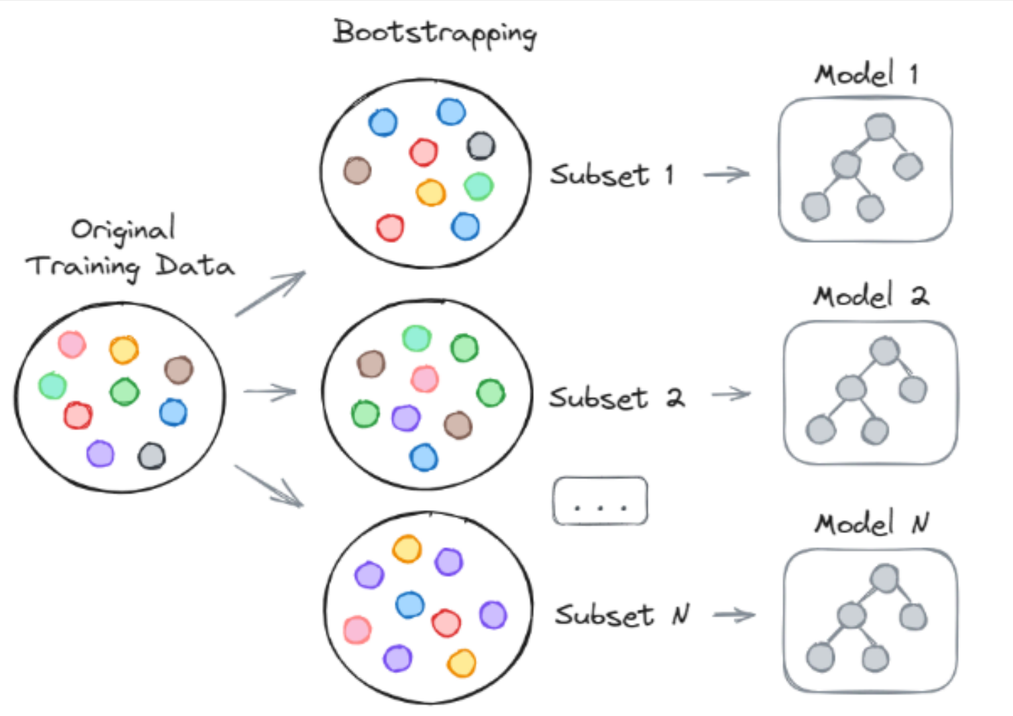
Types of Ensemble Methods

(Mainly used for tree)

Bagging

Train on random subsets of the data in parallel

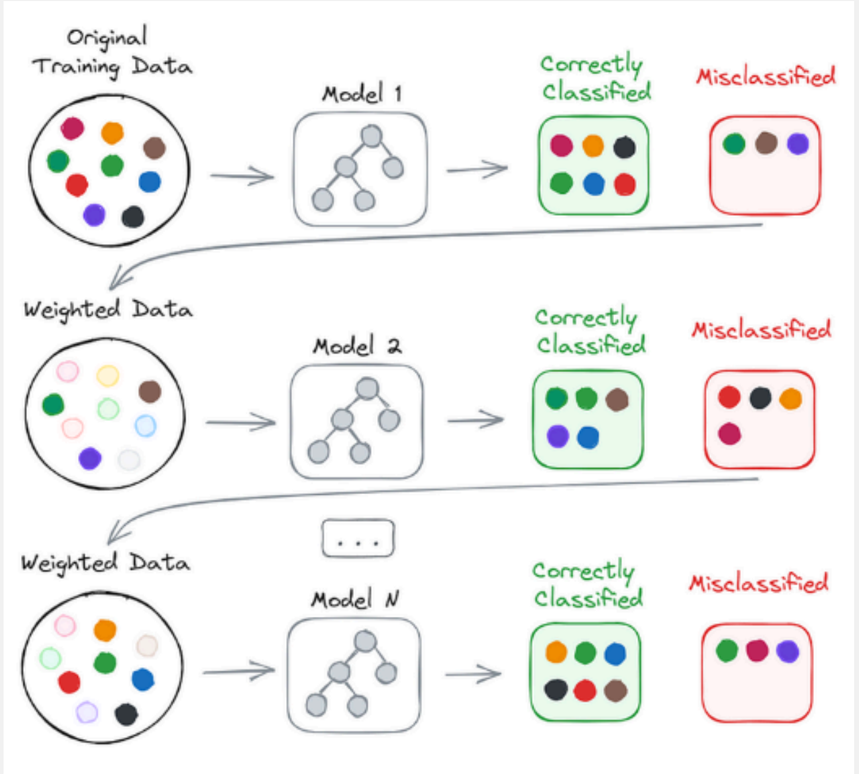
(Reduce variance)



Boosting

Train models sequentially to fix previous mistakes

(Reduce bias)



Bagging & Boosting

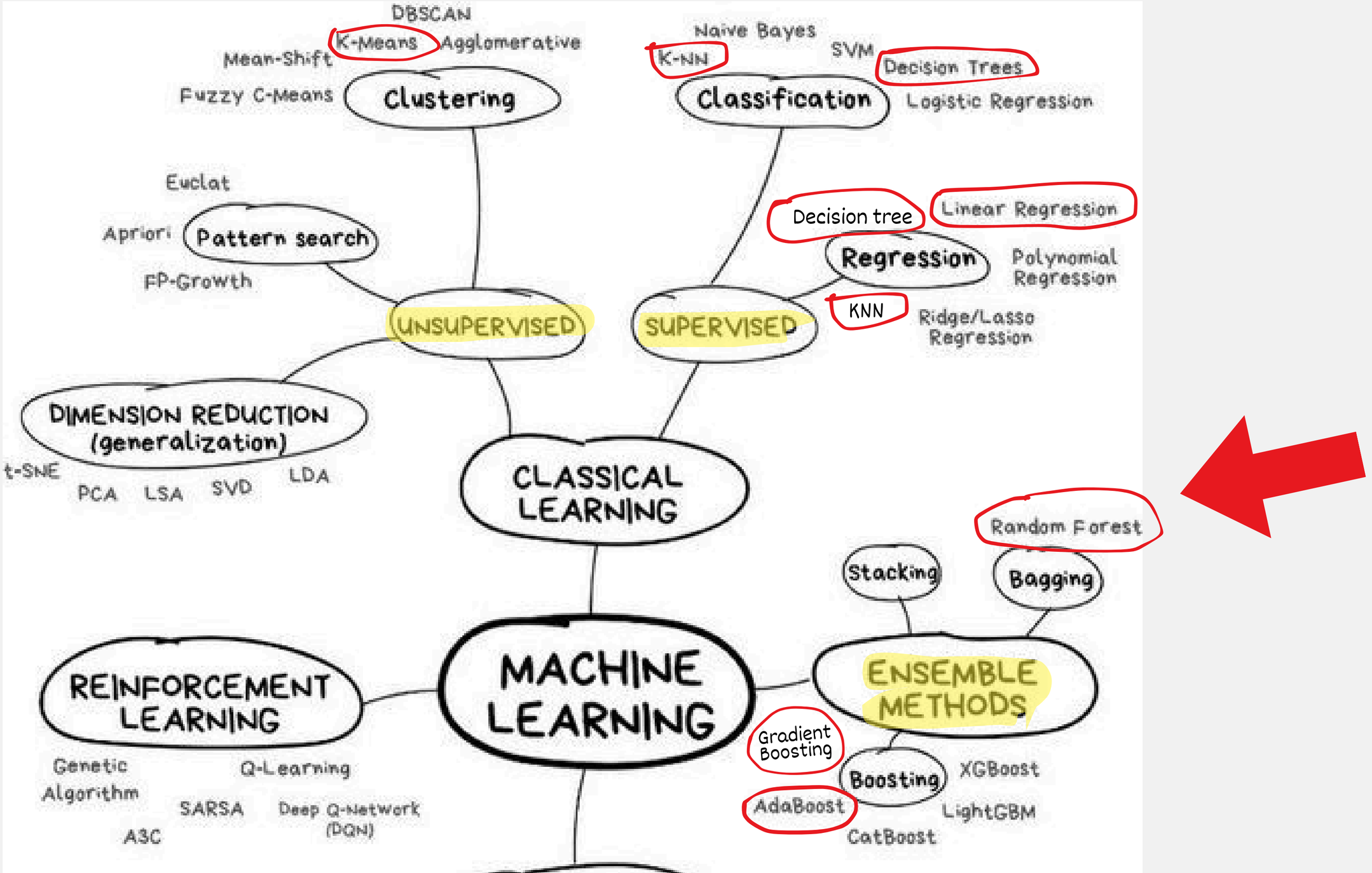
✅ Pros

- Higher accuracy 📈
- Works well on complex data
- Reduces overfitting (esp. bagging)

❌ Cons

- More computationally expensive 🧮
- Harder to interpret
- Risk of overfitting if not tuned properly

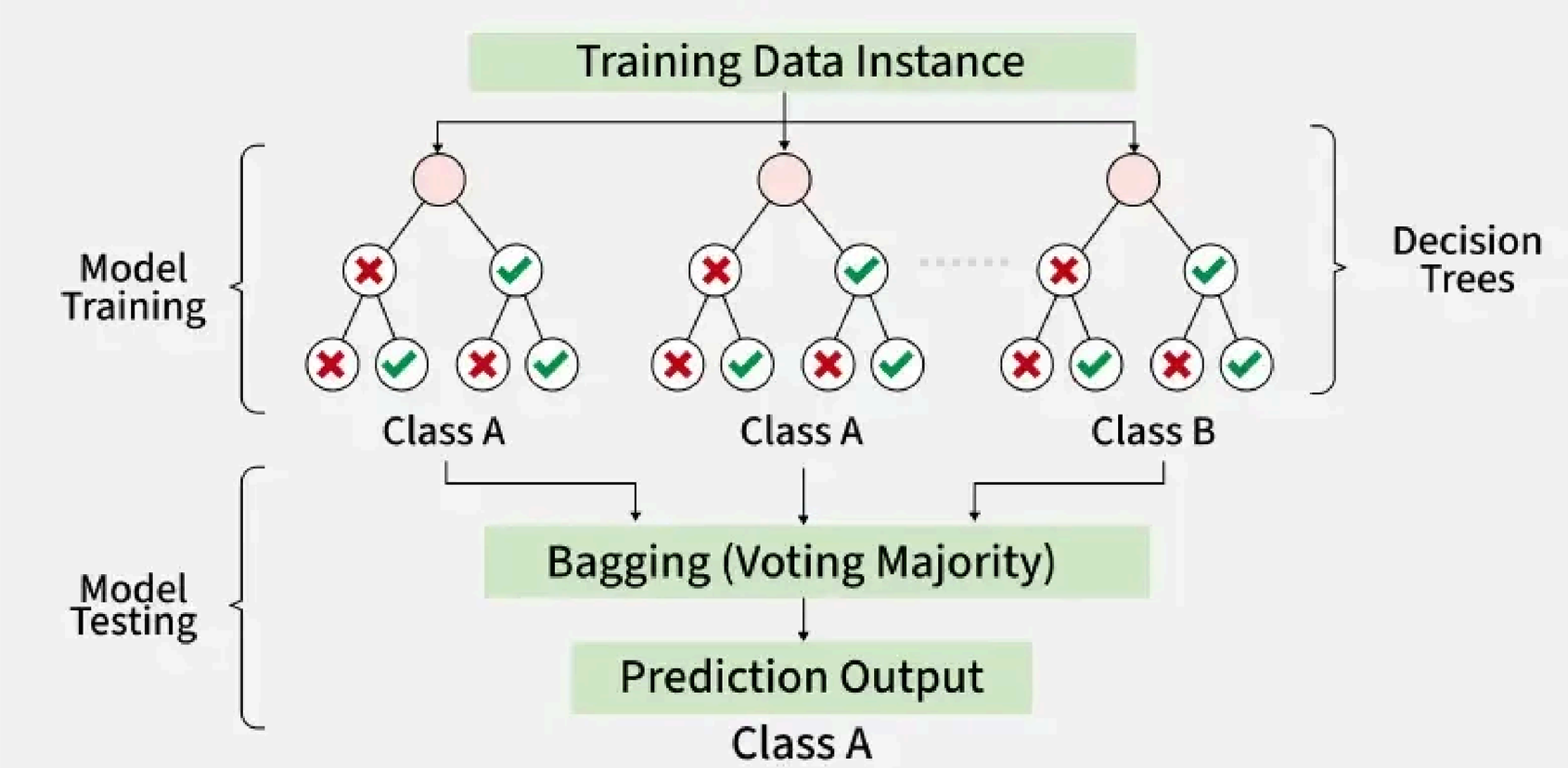
WE COVER EACH OTHER'S WEAKNESSES


DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Clustering
Naive Bayes
K-NN
SVM
Decision Trees
Classification
Logistic Regression
Euclat
Apriori
Pattern search
FP-Growth
Decision tree
Linear Regression
Regression
Polynomial Regression
KNN
Ridge/Lasso Regression
UNSUPERVISED
SUPERVISED
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
CLASSICAL LEARNING
Random Forest
Stacking
Bagging
ENSEMBLE METHODS
MACHINE LEARNING
REINFORCEMENT LEARNING
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C
Gradient Boosting
Boosting
XGBoost
AdaBoost
LightGBM
CatBoost

Random Forest

(Bagging ensemble algorithms)

- A bunch of decision trees voting together



Random Forest

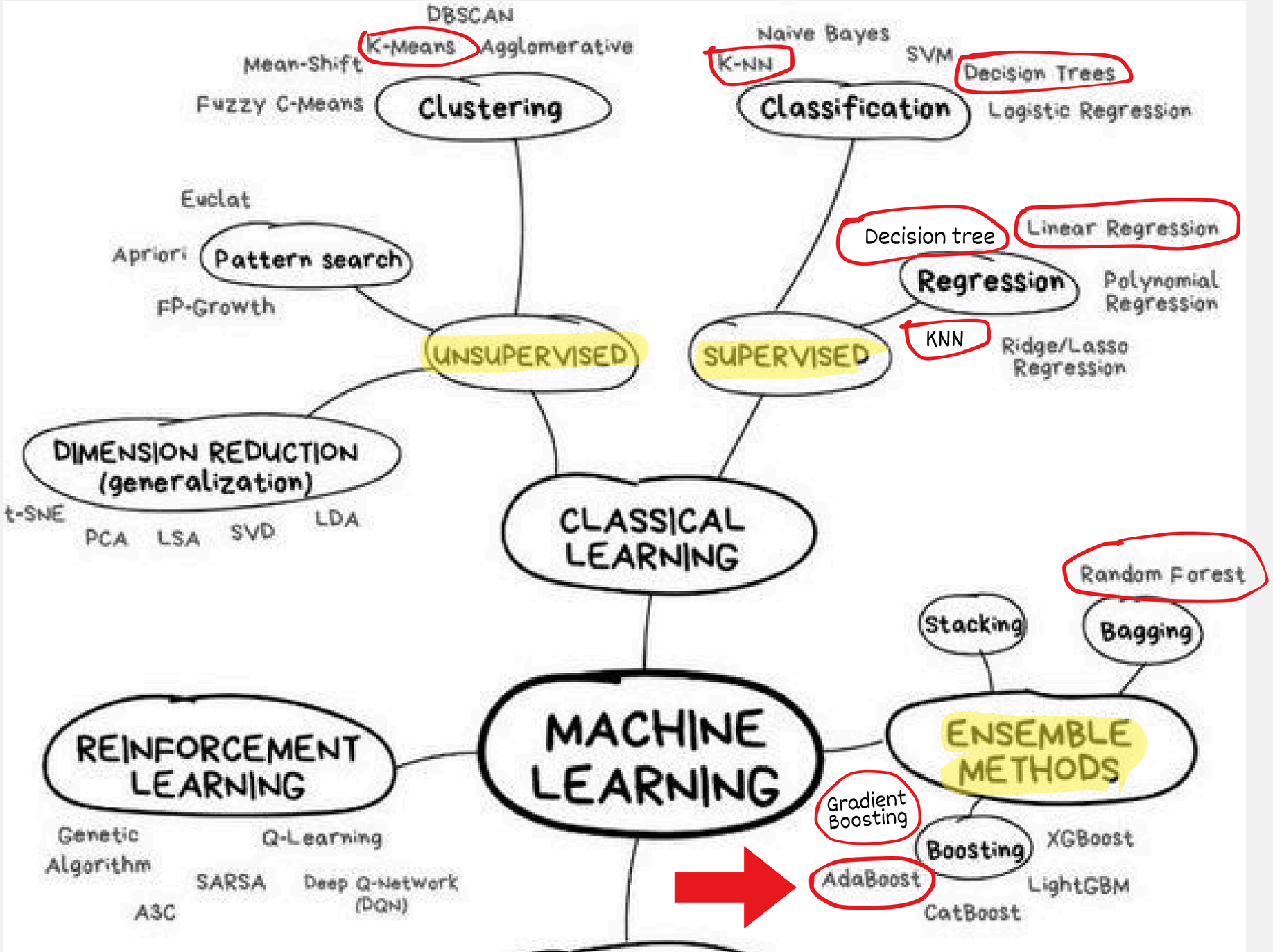
✅ **Pros**

- Handles both regression & classification
- Great accuracy out of the box
- Reduces overfitting (thanks to averaging)
- Handles missing values & noisy data well

❌ **Cons**

- Slower to train with many trees
- Hard to interpret (black-box feel)
- Can be memory-heavy

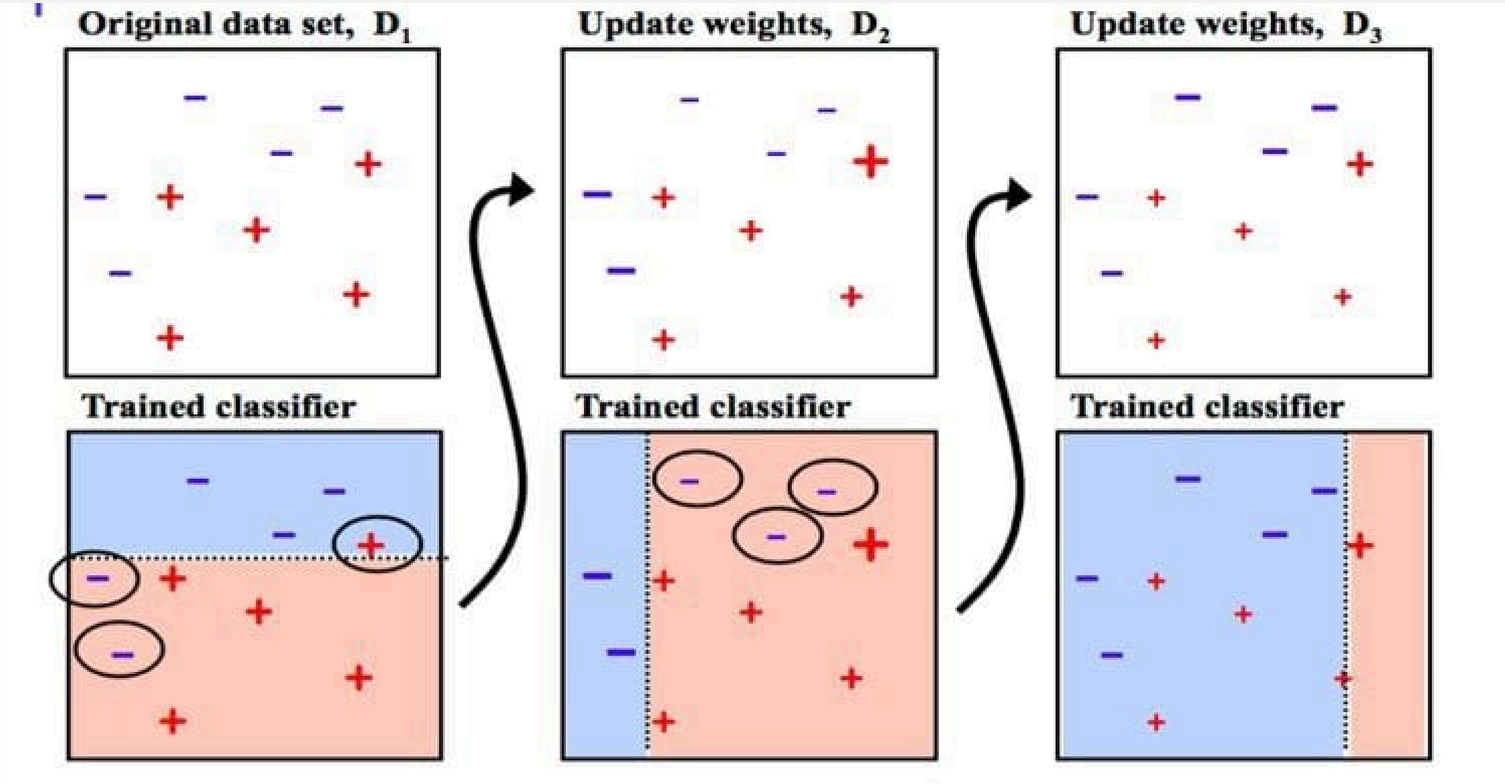
ONE TREE MIGHT BE WRONG, BUT TOGETHER THEY'RE USUALLY RIGHT.


MACHINE LEARNING
CLASSICAL LEARNING
UNSUPERVISED
SUPERVISED
Clustering
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Classification
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Pattern search
Euclat
Apriori
FP-Growth
Regression
Decision tree
Linear Regression
Polynomial Regression
KNN
Ridge/Lasso Regression
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
REINFORCEMENT LEARNING
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C
ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
Gradient Boosting
XGBoost
AdaBoost
LightGBM
CatBoost

AdaBoost

(Boosting ensemble algorithms)

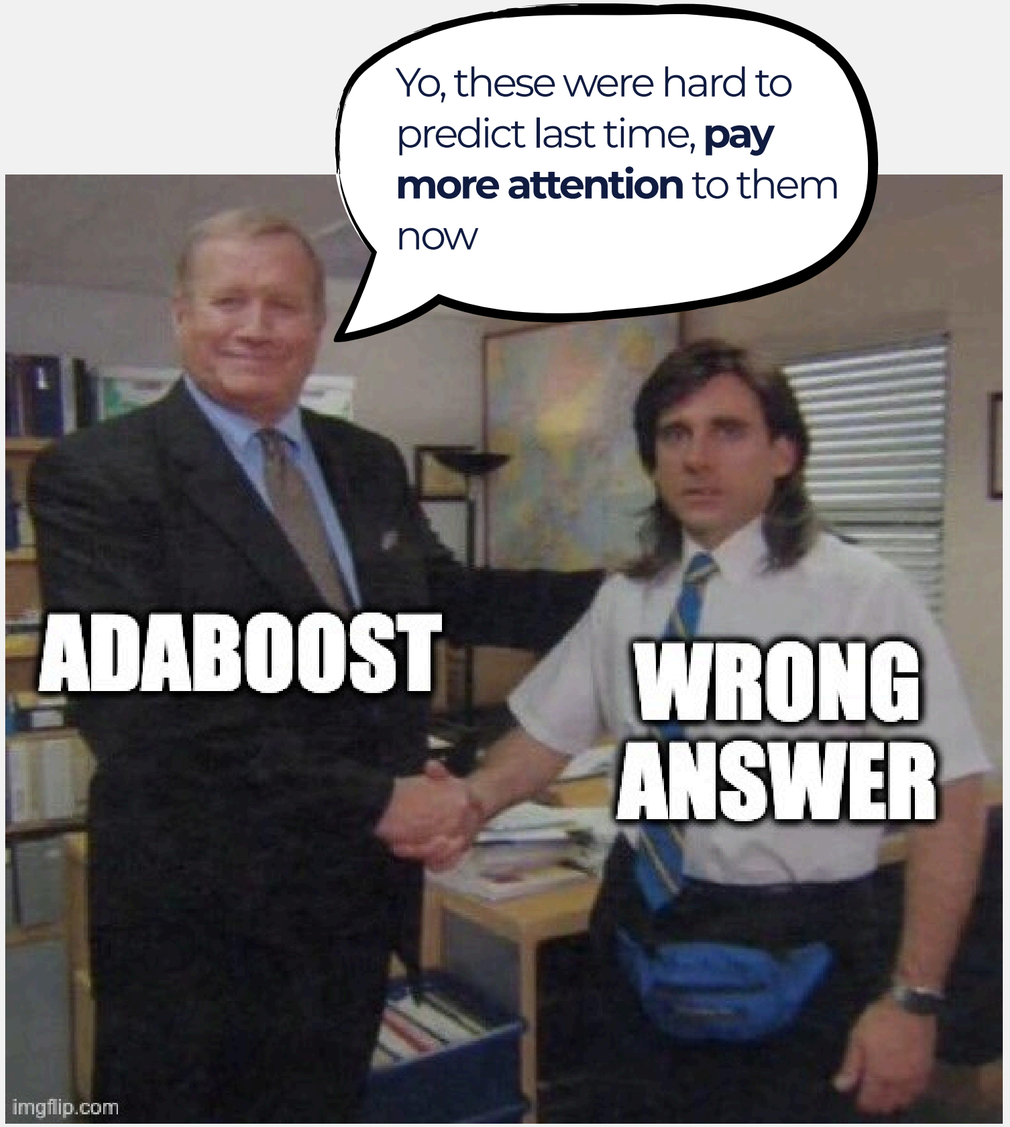
- Fix wrong classifications by increasing the weight on wrongly predicted samples

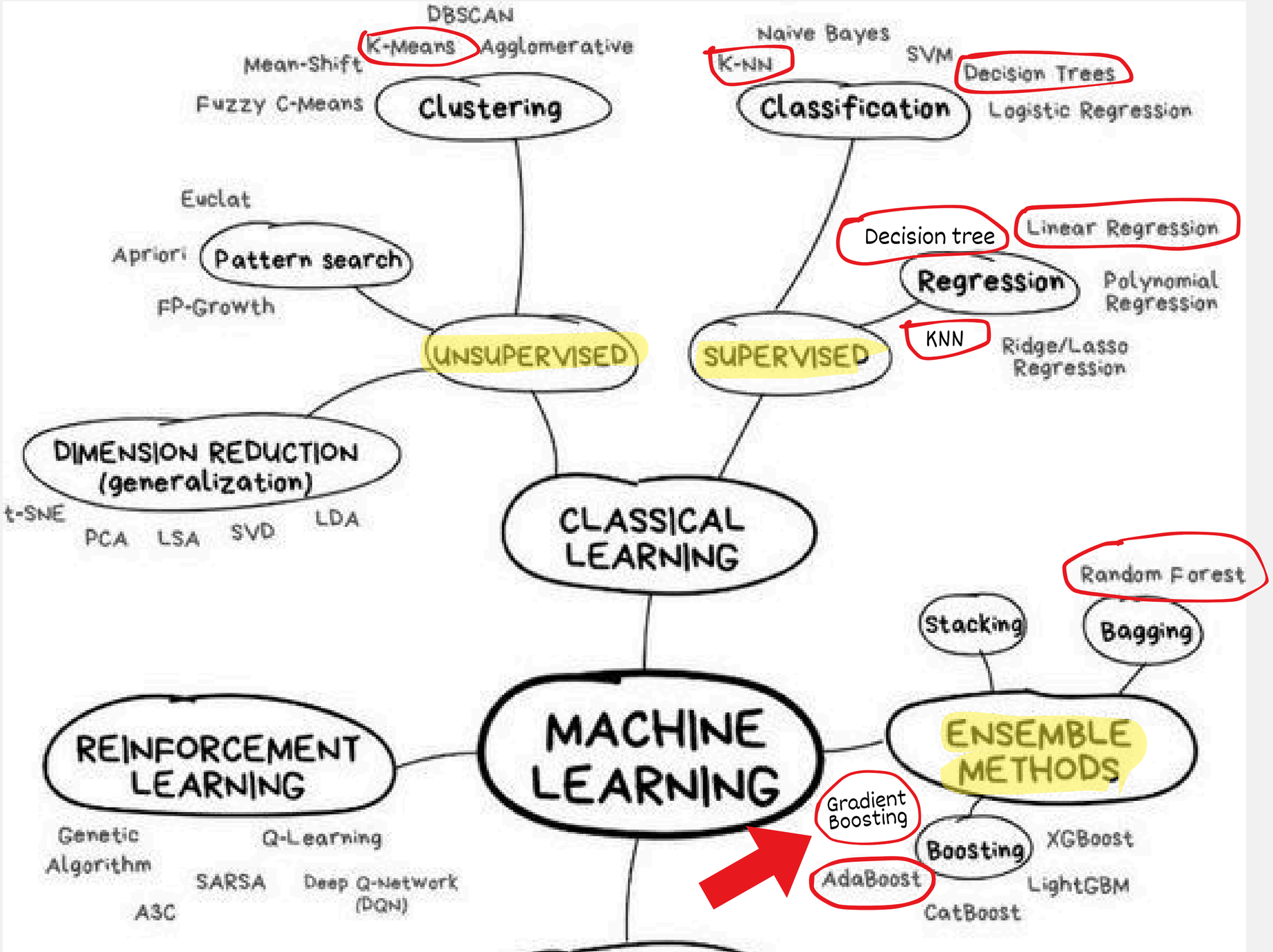


AdaBoost

(Boosting ensemble algorithms)

- Teacher re-testing students
  - each time, spending more time on the students who failed before

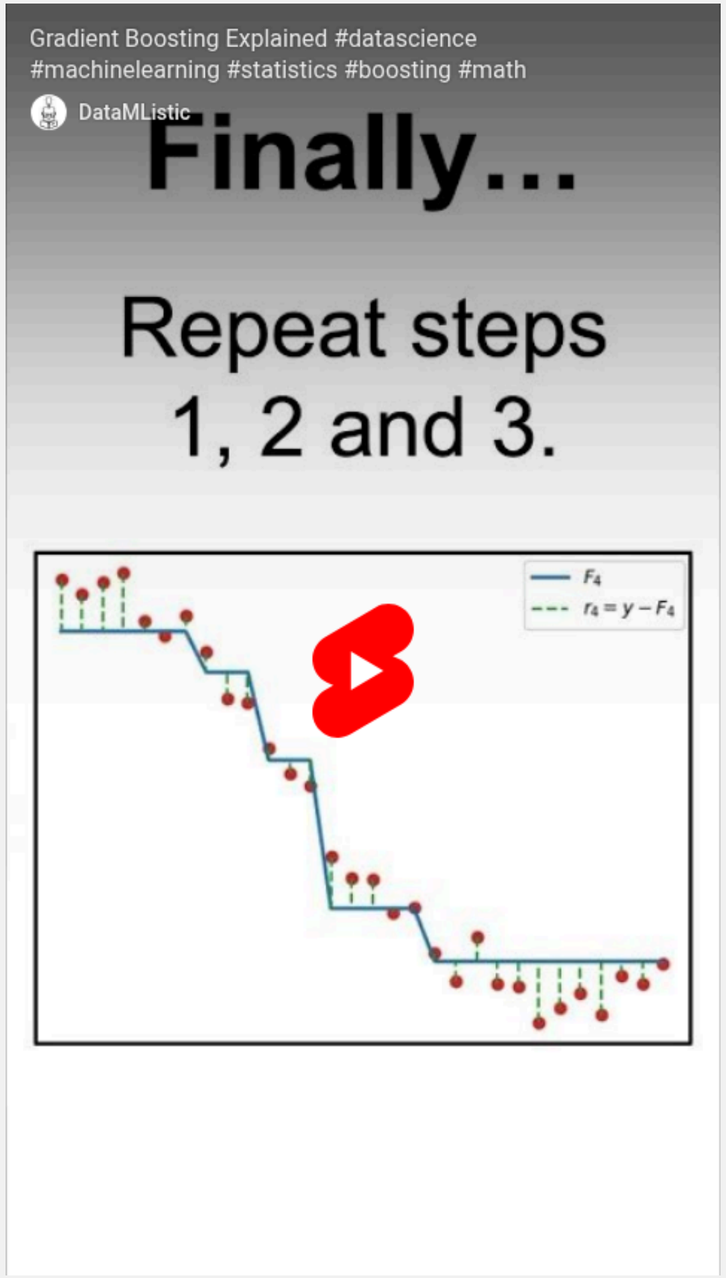



MACHINE LEARNING
CLASSICAL LEARNING
UNSUPERVISED
SUPERVISED
Clustering
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Classification
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Pattern search
Euclat
Apriori
FP-Growth
Regression
Decision tree
Linear Regression
Polynomial Regression
KNN
Ridge/Lasso Regression
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
XGBoost
LightGBM
CatBoost
AdaBoost
Gradient Boosting
REINFORCEMENT LEARNING
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C

Gradient Boosting

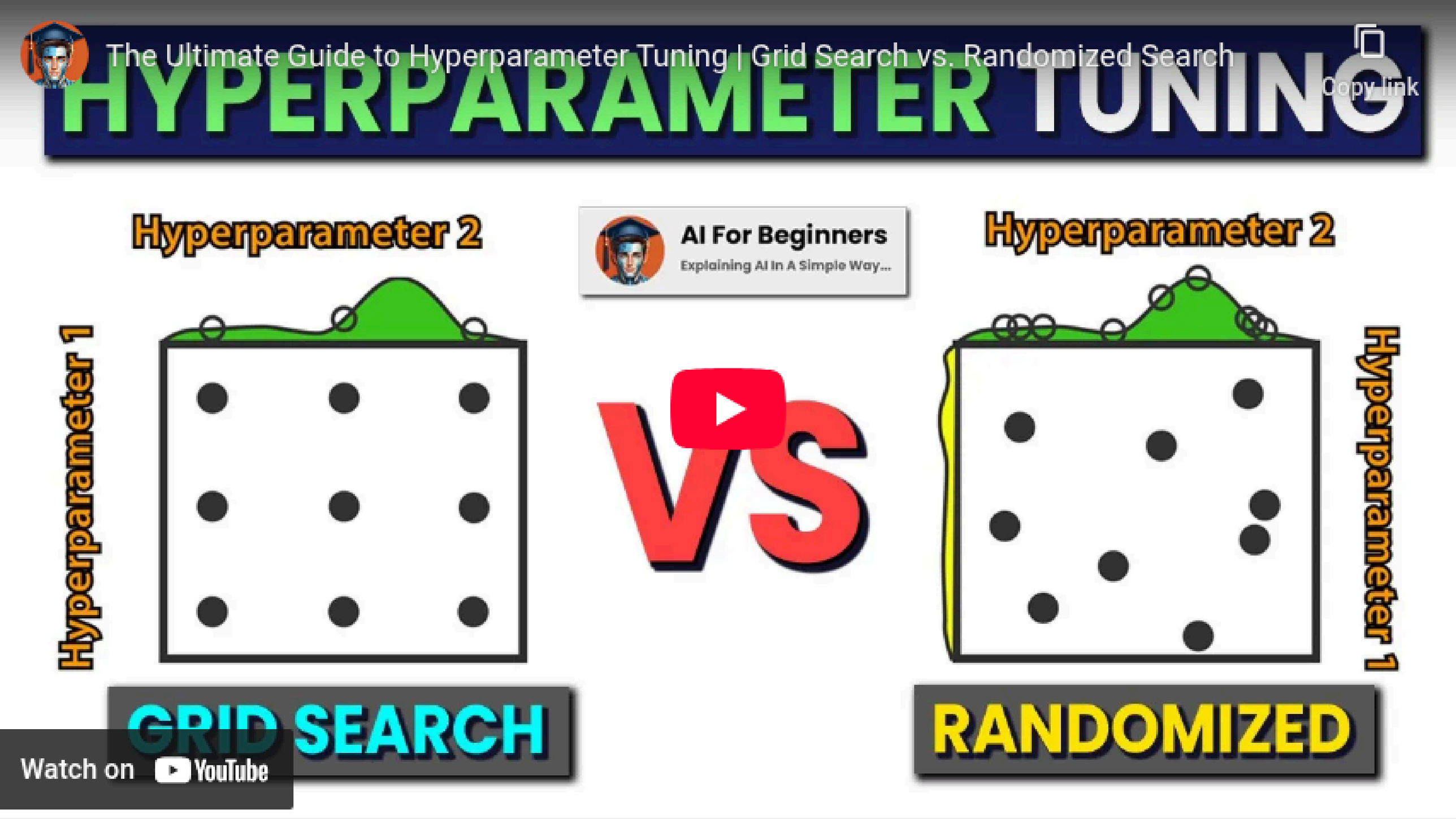
(Boosting ensemble algorithms)

- “learn from mistakes” strategy
- Each tree corrects the previous one
- Slower than Random Forest, but usually more accurate when tuned well



LEARNING
THE
THEORY FOR ML

CODE
THE THING


The Ultimate Guide to Hyperparameter Tuning | Grid Search vs. Randomized Search
Copy link
HYPERPARAMETER TUNING
Hyperparameter 2
Hyperparameter 1
AI For Beginners
Explaining AI In A Simple Way...
VS
Hyperparameter 2
Hyperparameter 1
GRID SEARCH
RANDOMIZED
Watch on YouTube