Supervised & unsupervised & semi-supervised learning

MohammadReza Ahmadi Teshnizi

Learning

supervised

Supervised learning is a machine learning paradigm where the algorithm learns from a labeled dataset, making predictions or classifications based on input-output pairs. It's used for tasks like image recognition and spam email filtering, where the model learns to generalize patterns from known examples.

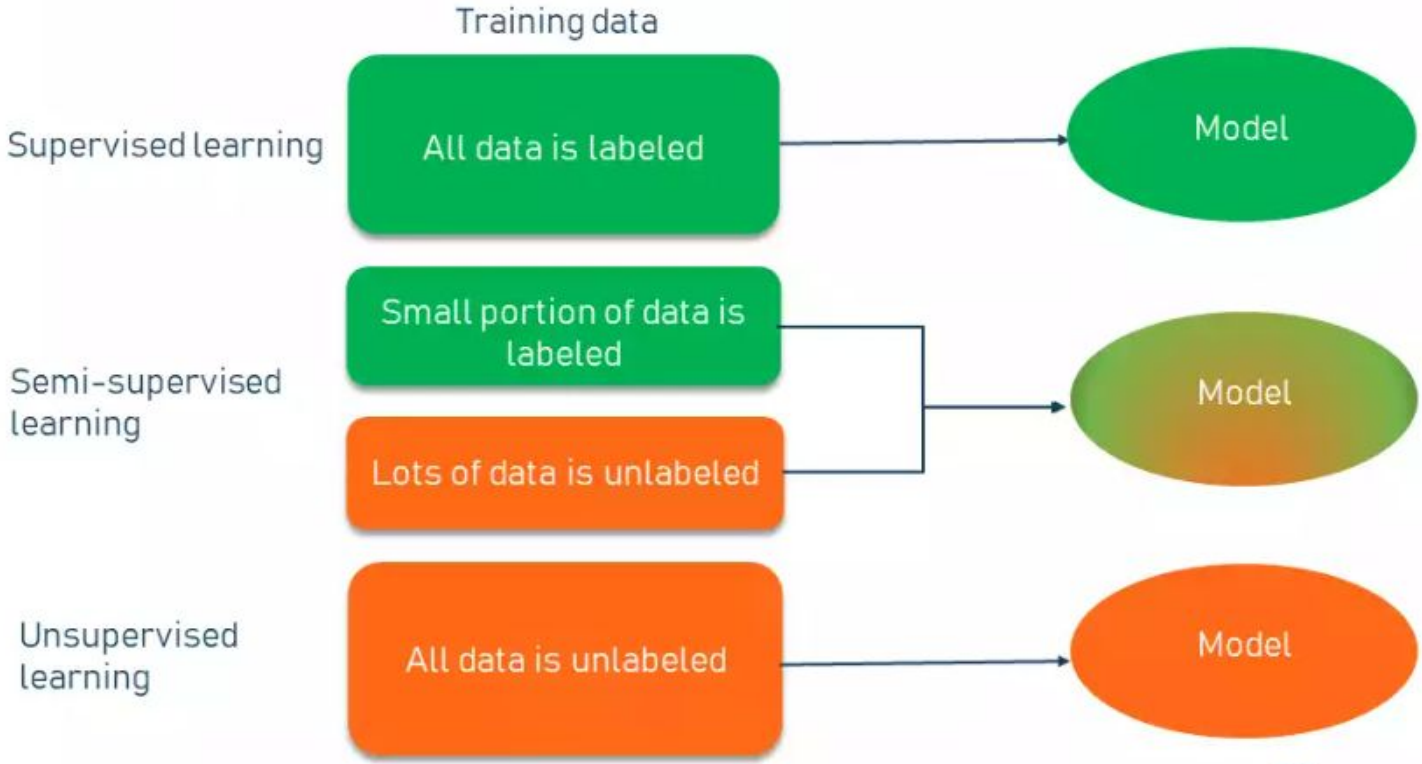
semi-supervised

Semi-supervised learning combines elements of both supervised and unsupervised learning. It uses a partially labeled dataset, leveraging the labeled data for training and the unlabeled data to improve model performance. This approach is useful when obtaining labeled data is costly or time-consuming.

Unsupervised

Unsupervised learning is a machine learning technique where the algorithm learns from unlabeled data to discover patterns or structure within the data itself. It's commonly used for tasks like clustering and dimensionality reduction, helping uncover hidden relationships and insights in the absence of explicit labels.

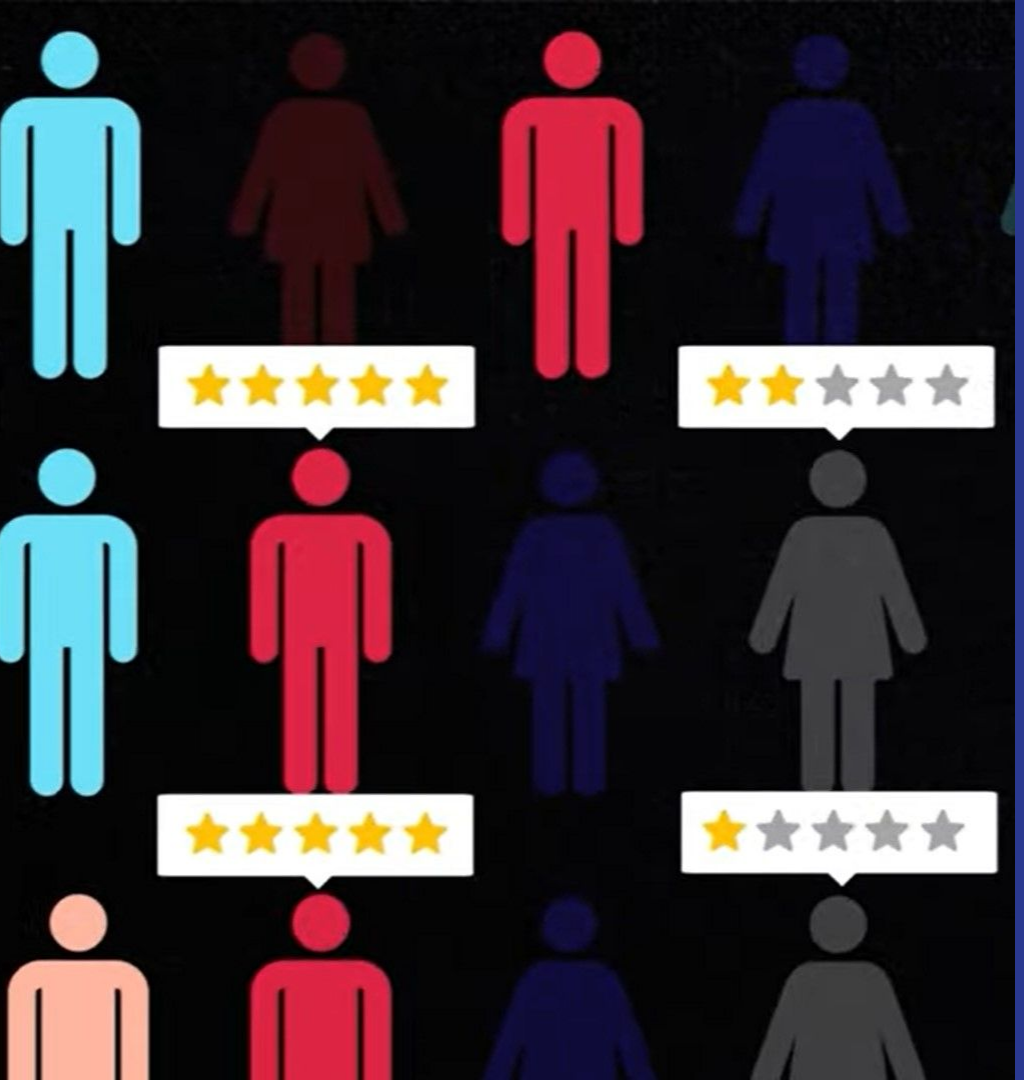
SUPERVISED LEARNING vs SEMI-SUPERVISED LEARNING vs UNSUPERVISED LEARNING



altexsoft

What is labeled data and why is it important?

High-quality data preparation ensures that machine learning models can learn effectively and produce accurate predictions.



Amazon recruitment

How come Amazon's machine learning model turned out to be sexist?

A: AI goes rogue

B: Inexperienced data scientists

C: Faulty dataset

D: Alexa gets jealous

15 • $1 MILLION
14 • $500,000
13 • $250,000
12 • $125,000
11 • $64,000
10 • $32,000
9 • $16,000
8 • $8,000
7 • $4,000
6 • $2,000
5 • $1,000
4 • $500
3 • $300
2 • $200
1 • $100

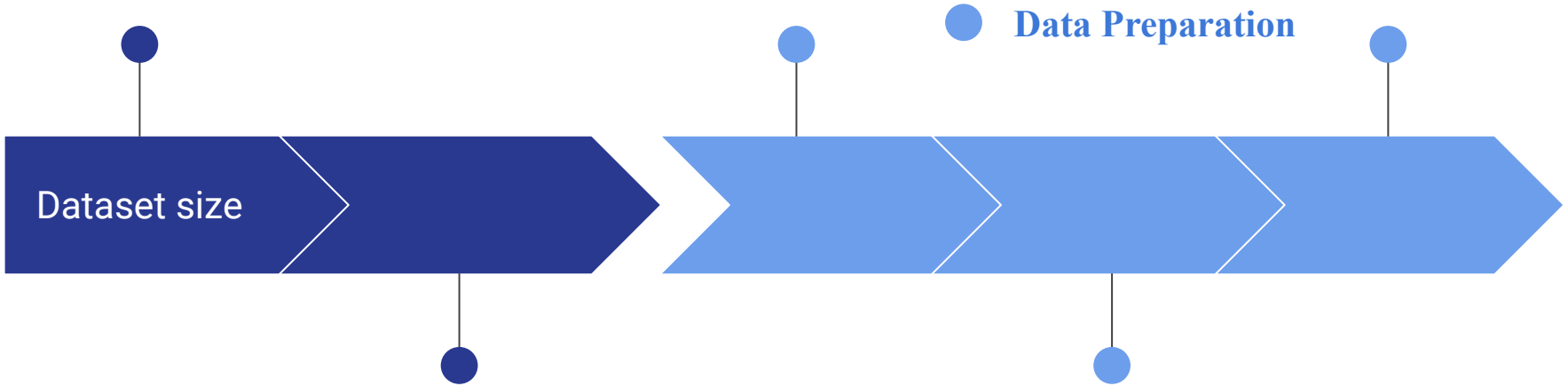
REUTERS 2018

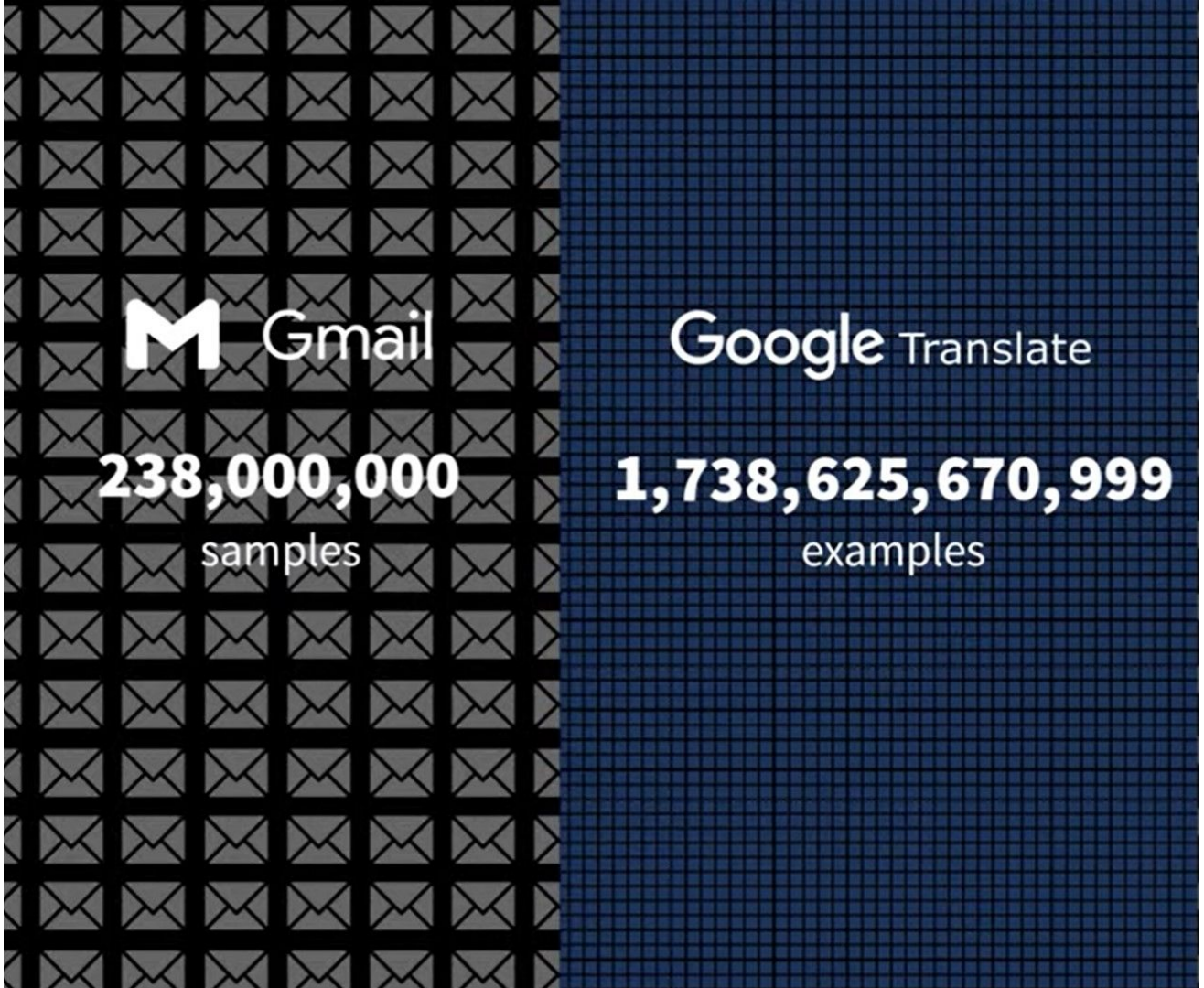
Amazon scraps secret AI recruiting tool that showed bias against women

"...the company realized its new system was not rating candidates for software developer jobs and other technical posts in a gender-neutral way"

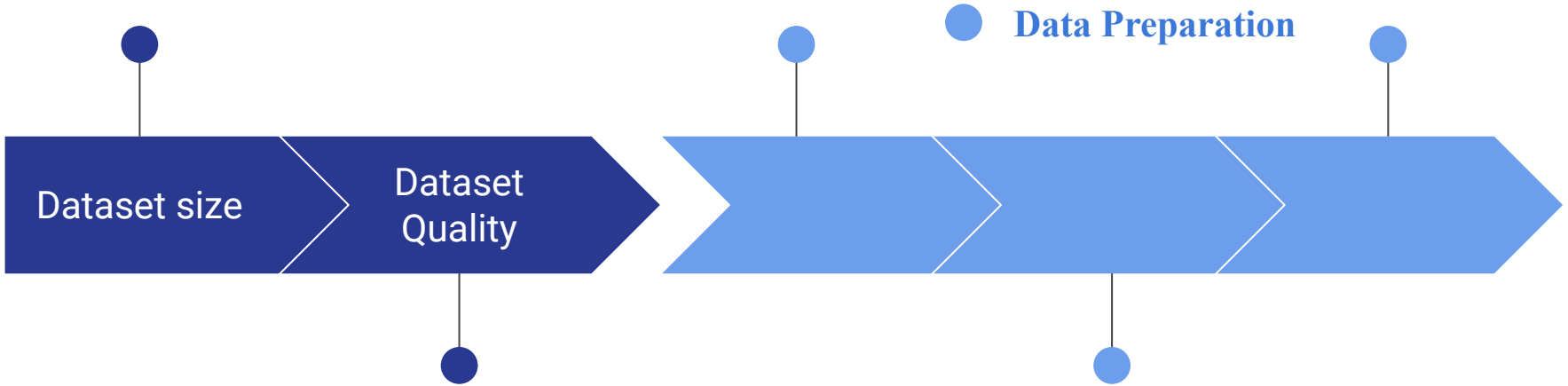
"That is because Amazon's computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry"

As much as you can




Gmail
238,000,000
samples
Google Translate
1,738,625,670,999
examples

As much as you can



High quality data

I-Cheng Yeh

Tamkang University Professor

Gmail

238,000,000

samples

Google Translate

1,738,625,670,999

examples

630

samples

ScienceDirect®

"The proposed framework was applied for the prediction of compressive strength (CS) of femoral trabecular bone in patients suffering from severe osteoarthritis. We reproduce this result on CS data of another porous solid (concrete). When evaluated on independent test samples, the NN achieved accuracy of 98.3%, outperforming an ensemble NN model by 11%."

Garbage
IN

Garbage
OUT

Example

You have information on the sale of turkey chicken in Canada and you want to predict the sale of turkey chicken during the American holidays!

Historical Data

Sales Prediction

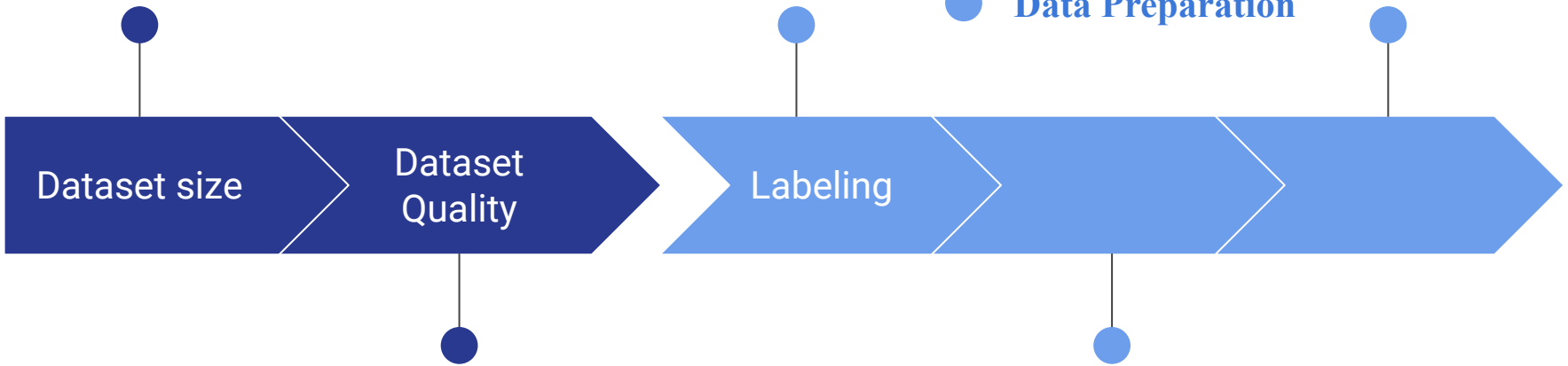
Historical Data

Sales Prediction

As much as you can

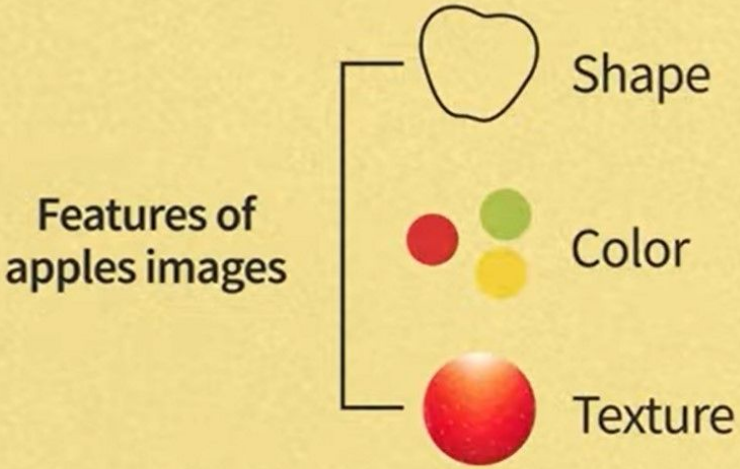
In semi supervised
and supervised

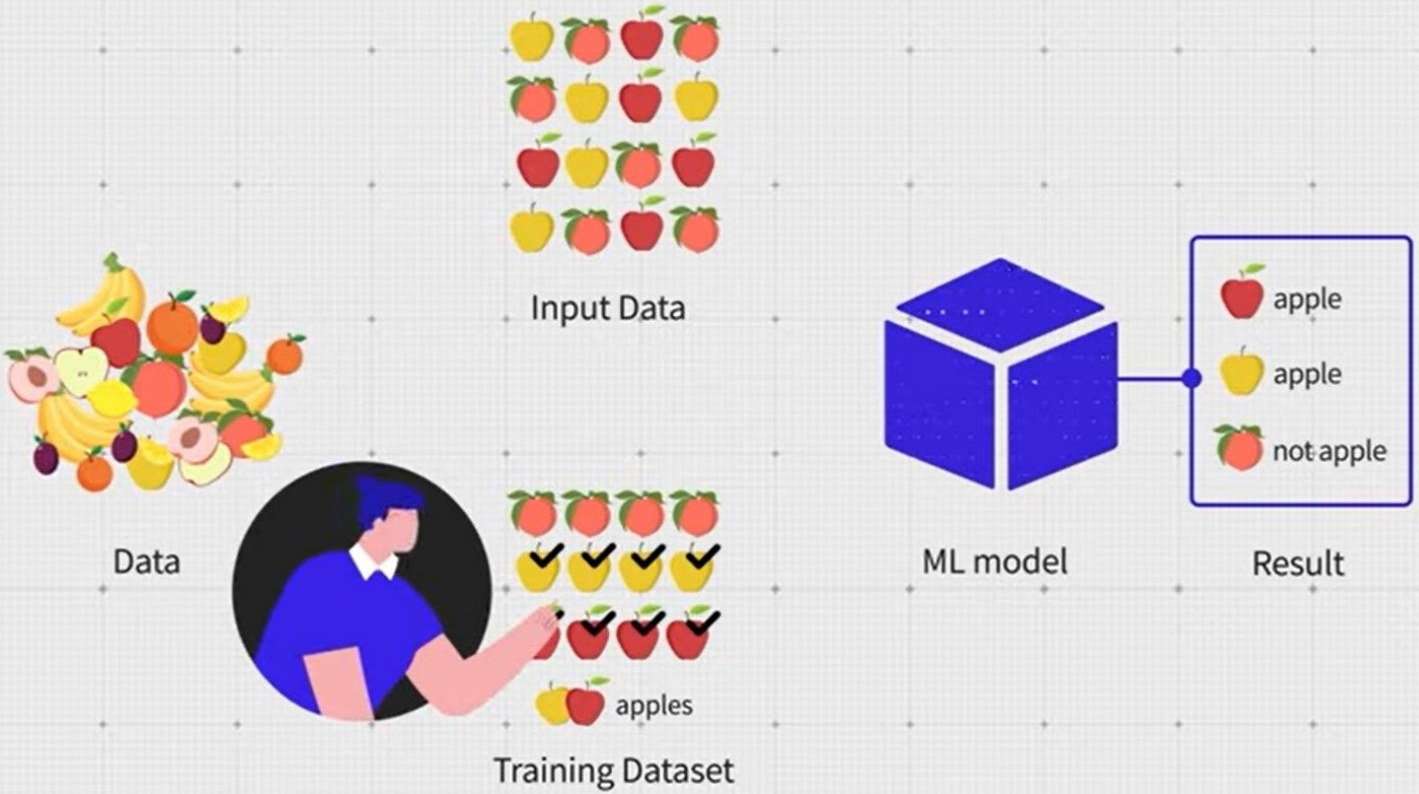
Data Preparation



High quality data

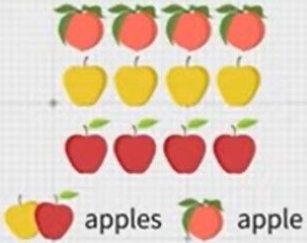

These are apples
APPLE


Features of apples images
Shape
Color
Texture

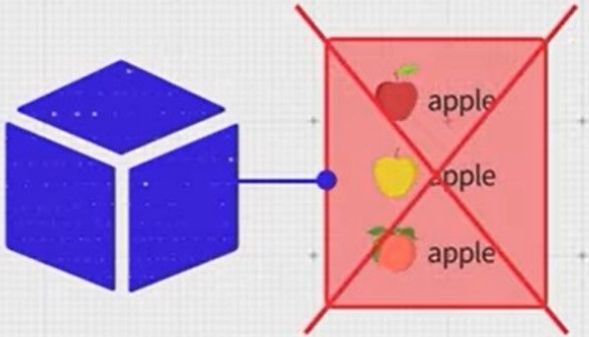

Input Data
Data
apples
Training Dataset
ML model
apple
apple
not apple
Result

Data

Input Data



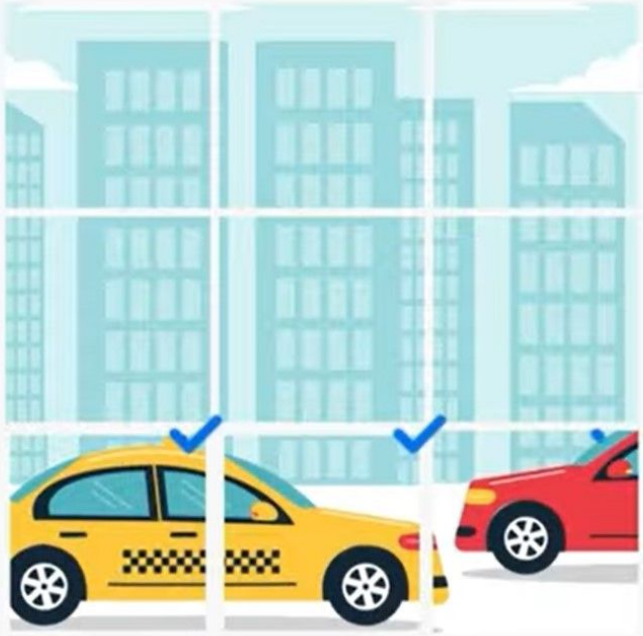
Training Dataset



ML model

Result

Select all squares with
cars

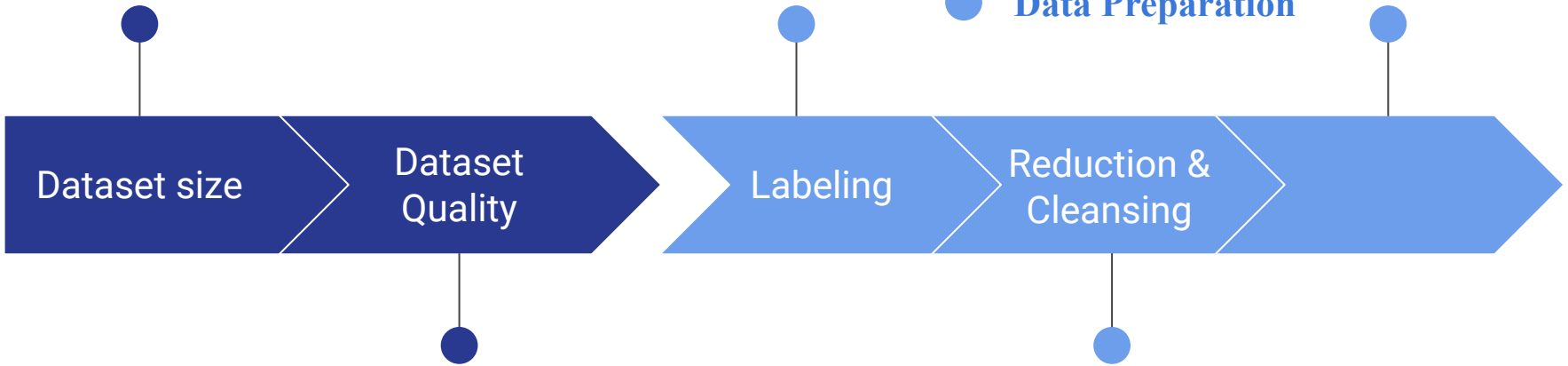


VERIFY

As much as you can

In semi supervised
and supervised

Data Preparation



High quality data

Relevant data

Customer	Year of birth	Age	Country	Single room	Double bedroom	Twin room	Suite
Customer A	1990	31	US	2	-	-	-
Customer B	1963	58	US	-	3	-	1
Customer C	1969	52	US	-	-	6	-
Customer D	2001	20	US	2	-	-	3

D ~ 0

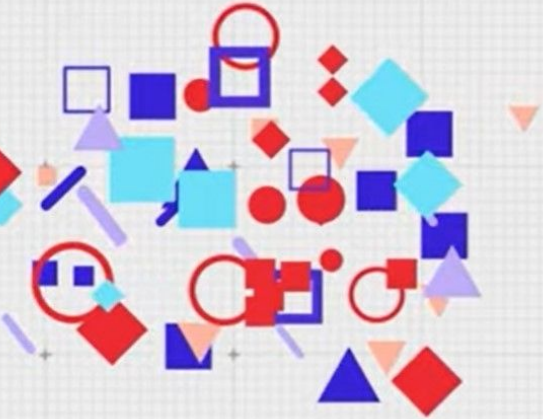
Duplicate

Customer	Year of birth	Age	Single room	Double bedroom	Twin room	Suite
Customer A	1990	31	2	-	-	-
Customer B	1963	58	-	3	-	1
Customer C	1969	52	-	-	6	-
Customer D	2001	20	2	-	-	3

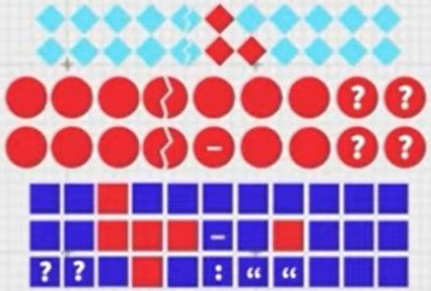
Customer	Age	Country	Booking	Date-Time
Customer A	56	US	Single	2021-03-24 T10:46:07Z
Customer B	22	Canada	Single	2021-03-24 T12:14:23Z
Customer C	37	US	Double	2021-03-22 T10:00:14Z
Customer D	48	Brazil	Twin	2021-03-18 T11:03:45Z

Customer	Age	Country	Booking	Date	Time
Customer A	56	US	Single	2021-03-24	T10:46:07Z
Customer B	22	Canada	Single	2021-03-24	T12:14:23Z
Customer C	37	US	Double	2021-03-22	T10:00:14Z
Customer D	48	Brazil	Twin	2021-03-18	T11:03:45Z

Prototype

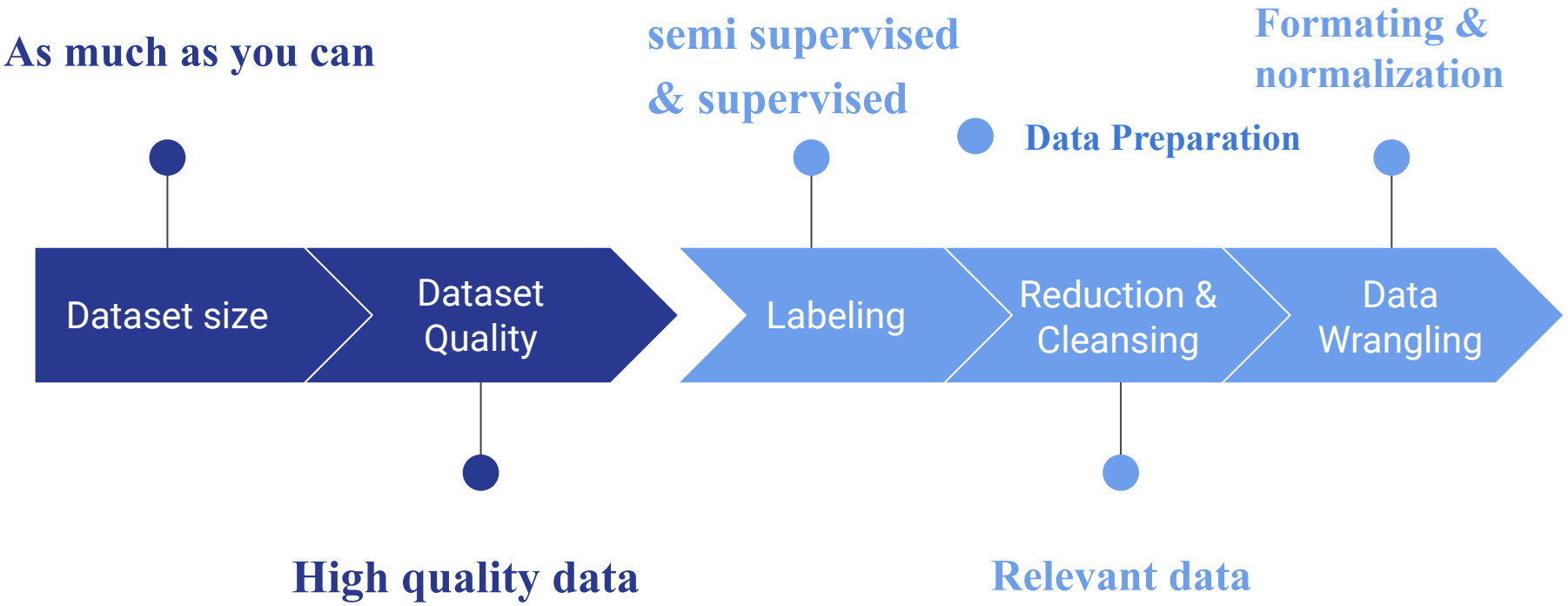


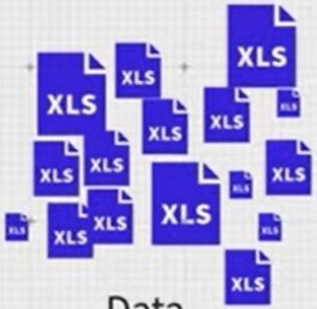
Data



Training Dataset

ML model


As much as you can
semi supervised
& supervised
Formating &
normalization
Data Preparation
Dataset size
Dataset
Quality
Labeling
Reduction &
Cleansing
Data
Wrangling
High quality data
Relevant data



Data

Training Dataset

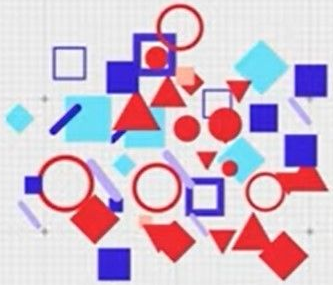
ML model

Historical Data

Historical Data

Historical Data ⟶ Normalized Data

Min-Max Normalization

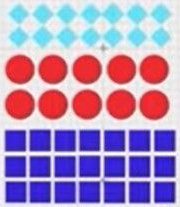


Data

Data Preparation

80%

of project time



Training Dataset

ML model

Supervised learning

slow (it requires human experts to manually label training examples one by one)

costly (a model should be trained on the large volumes of hand-labeled data to provide accurate predictions).

Unsupervised learning

has a limited area of applications (mostly for clustering purposes)

provides less accurate results.

Semi-Supervised learning

Unlike unsupervised learning, SSL works for a variety of problems from classification and regression to clustering and association.

Unlike supervised learning, the method uses small amounts of labeled data and also large amounts of unlabeled data, which reduces expenses on manual annotation and cuts data preparation time.

Unsupervised Learning vs. Supervised Learning

Unsupervised learning operates without labeled data.

Supervised learning relies on labeled data for training.

Key difference: Supervised has explicit target output, unsupervised does not.

Clustering

Clustering is a fundamental unsupervised learning technique.

It groups similar data points together based on certain criteria.

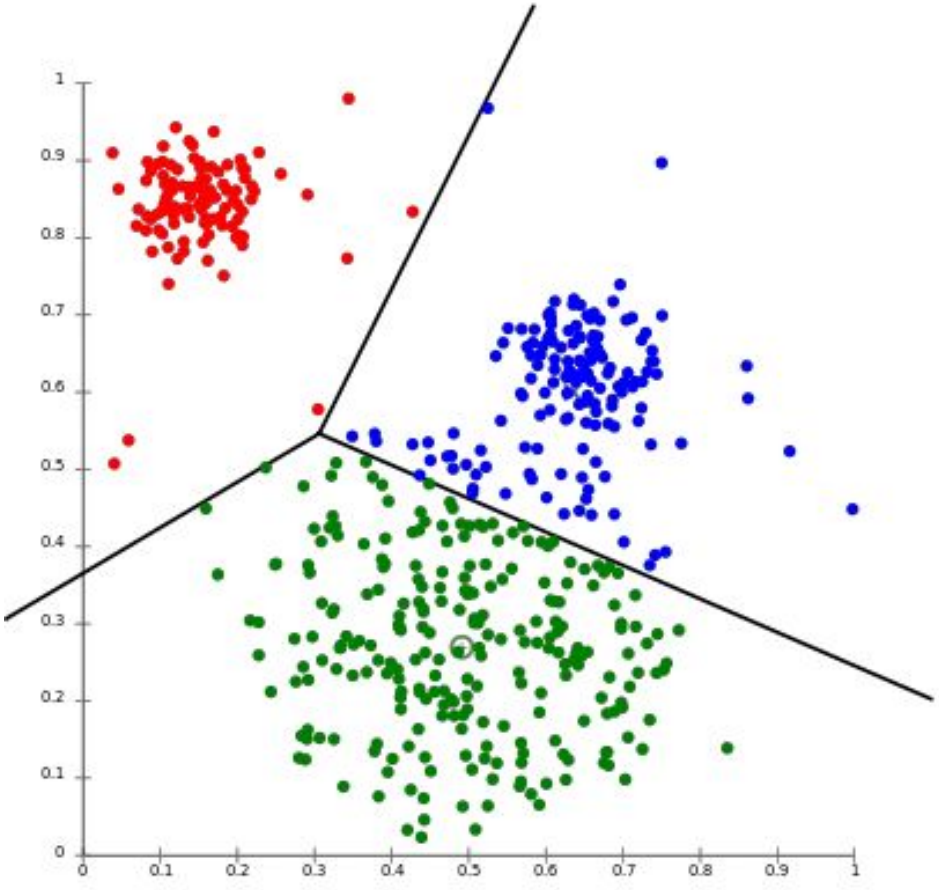
Common clustering algorithms include K-Means, Hierarchical, and DBSCAN.

K-Means Clustering

K-Means is a popular clustering algorithm.

Steps:

1. Choose the number of clusters (K).
2. Initialize cluster centroids.
3. Assign data points to the nearest centroid.
4. Update centroids.
5. Repeat steps 3 and 4 until convergence

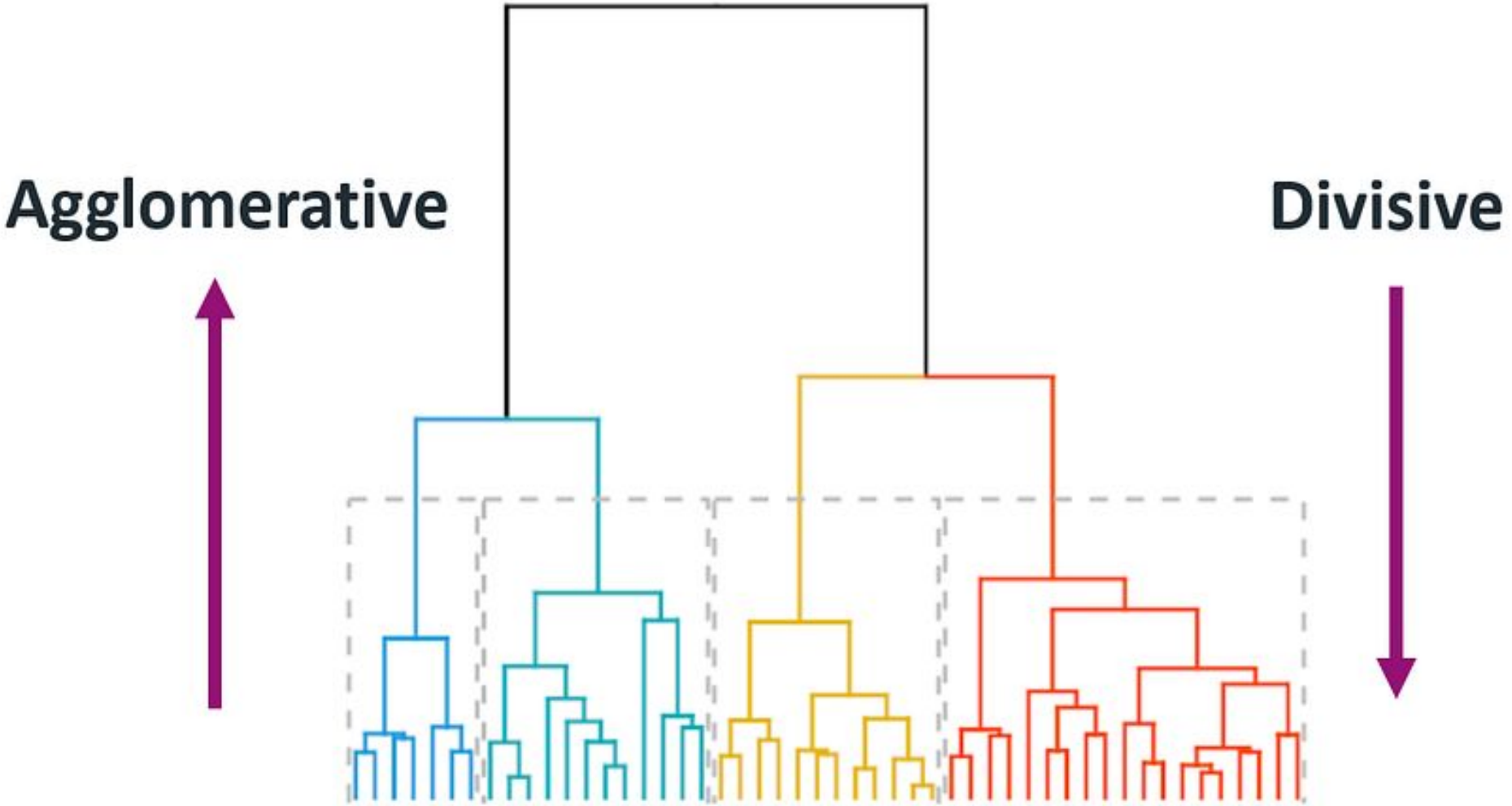

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Hierarchical Clustering

Hierarchical clustering builds a tree-like structure of clusters.

Agglomerative: Starts with individual data points as clusters and merges them.

Divisive: Starts with all data points in one cluster and recursively splits them.


Agglomerative
Divisive

AGGLOMERATIVE OR BOTTOM-UP HIERARCHICAL CLUSTERING

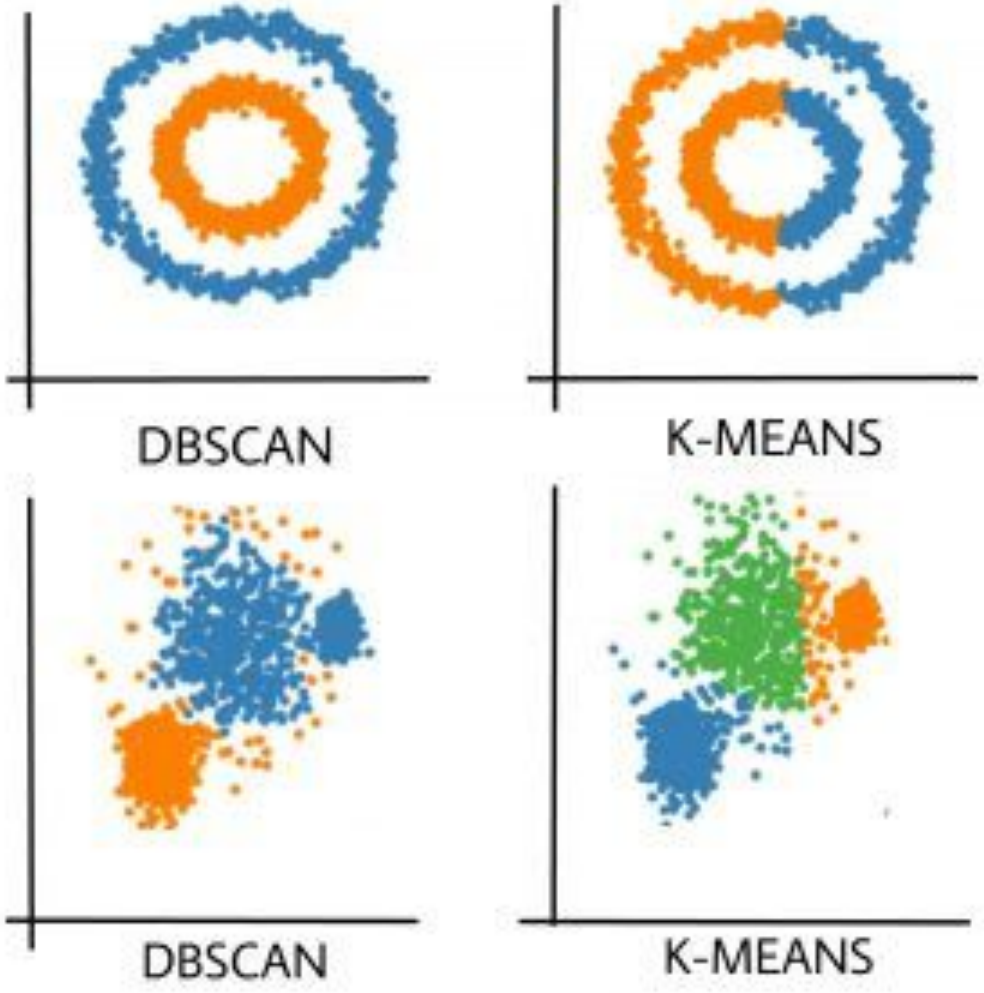
altexsoft
software r&d engineering

Density-Based Clustering (DBSCAN)

DBSCAN clusters data based on density.

Core points have a minimum number of data points within a specified radius.

Border points are within the radius of a core point but do not meet the density requirement.


DBSCAN
K-MEANS
DBSCAN
K-MEANS

Other Clustering Algorithms

Dimensionality Reduction:

- Dimensionality reduction reduces the number of features while preserving data patterns.
- Common techniques: PCA (Principal Component Analysis) and t-SNE (t-distributed Stochastic Neighbor Embedding).

Principal Component Analysis (PCA)

- PCA identifies and eliminates correlated features.
- Steps:
- Calculate the covariance matrix.
- Find eigenvectors and eigenvalues.
- Select top eigenvectors as principal components.
- Example: Reducing a dataset from 3D to 2D.

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- t-SNE is used for visualization.
- It preserves data relationships in a lower-dimensional space.
- Example: Visualizing high-dimensional data clusters.

Use Cases of Unsupervised Learning

Customer Segmentation: Grouping customers based on behavior.

Anomaly Detection: Detecting unusual patterns in data.

Image Compression: Reducing image file size without significant quality loss.

Challenges and Limitations

Choosing the right number of clusters is challenging.

Domain knowledge may be necessary to interpret results.

Scalability can be an issue with large datasets.

Semi-Supervised vs. Supervised Learning

In Supervised Learning, models are trained solely on labeled data.

Semi-Supervised Learning combines the use of limited labeled data with a significant amount of unlabeled data.

SSL aims to improve model performance by making the most of available data resources.

Why Semi-Supervised Learning?

Labeled data can be scarce, expensive, or time-consuming to obtain.

Unlabeled data often goes unused, leading to a loss of valuable information.

SSL bridges the gap between the data efficiency of supervised learning and the data abundance of unsupervised learning.

SSL Techniques

Self-Training: Iteratively labeling unlabeled data with high-confidence model predictions.

Multi-View Learning: Combining information from multiple data perspectives.

Co-Training: Using multiple weak models to label unlabeled data.

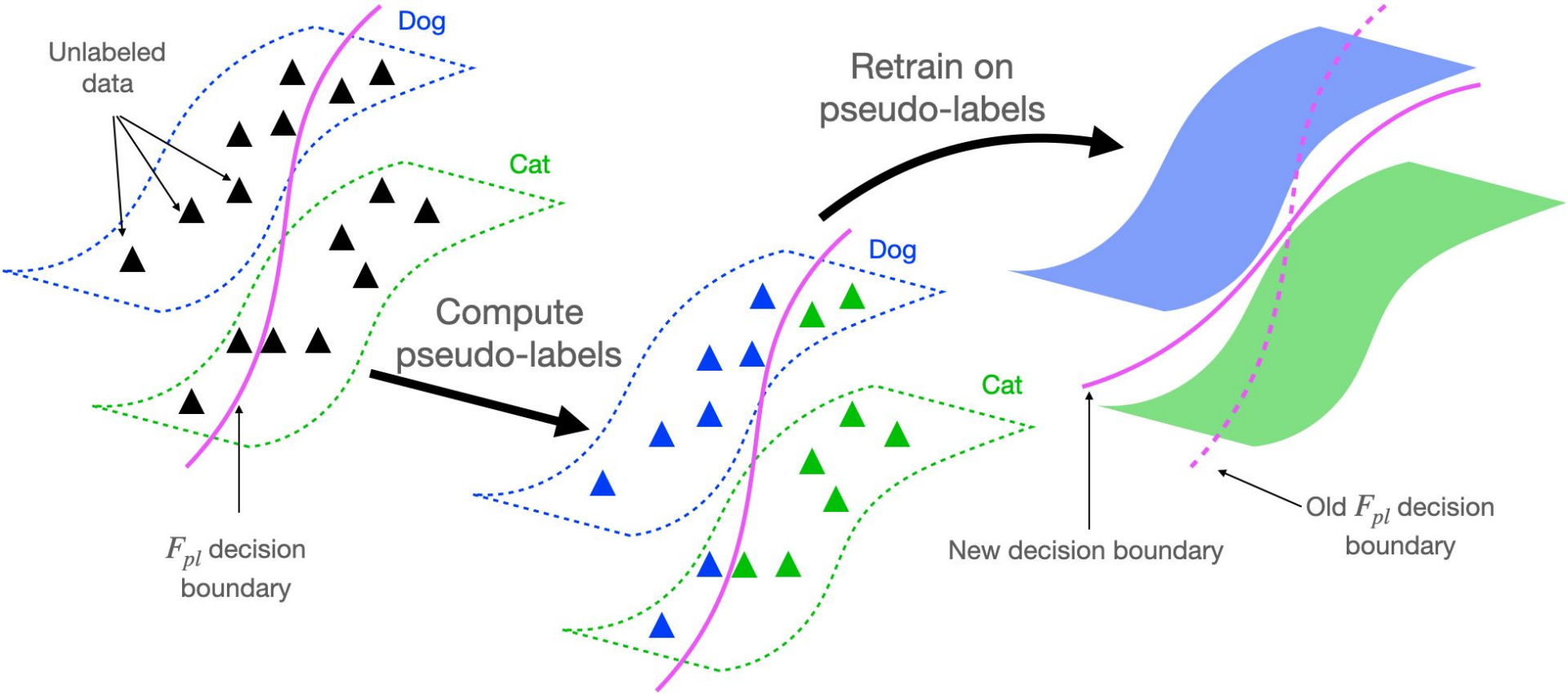
Label Propagation: Spreading labels through a graph or network structure.

Self-Training

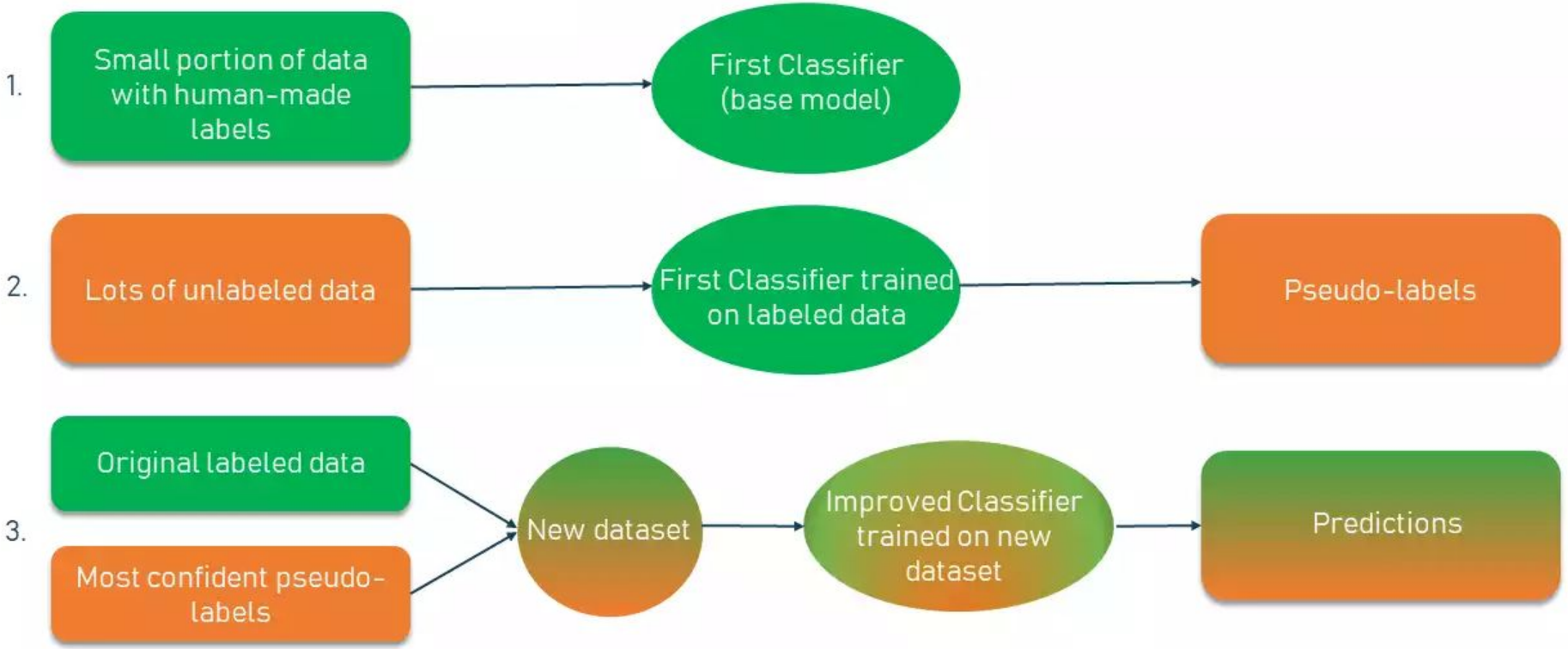
Self-Training is a popular SSL technique.

It starts with a small labeled dataset and expands it by labeling additional unlabeled data points.

The model iteratively labels unlabeled data points with high-confidence predictions.


Unlabeled data
Dog
Cat
F_{pl} decision boundary
Compute pseudo-labels
Dog
Cat
Retrain on pseudo-labels
New decision boundary
Old F_{pl} decision boundary

SEMI-SUPERVISED SELF-TRAINING METHOD

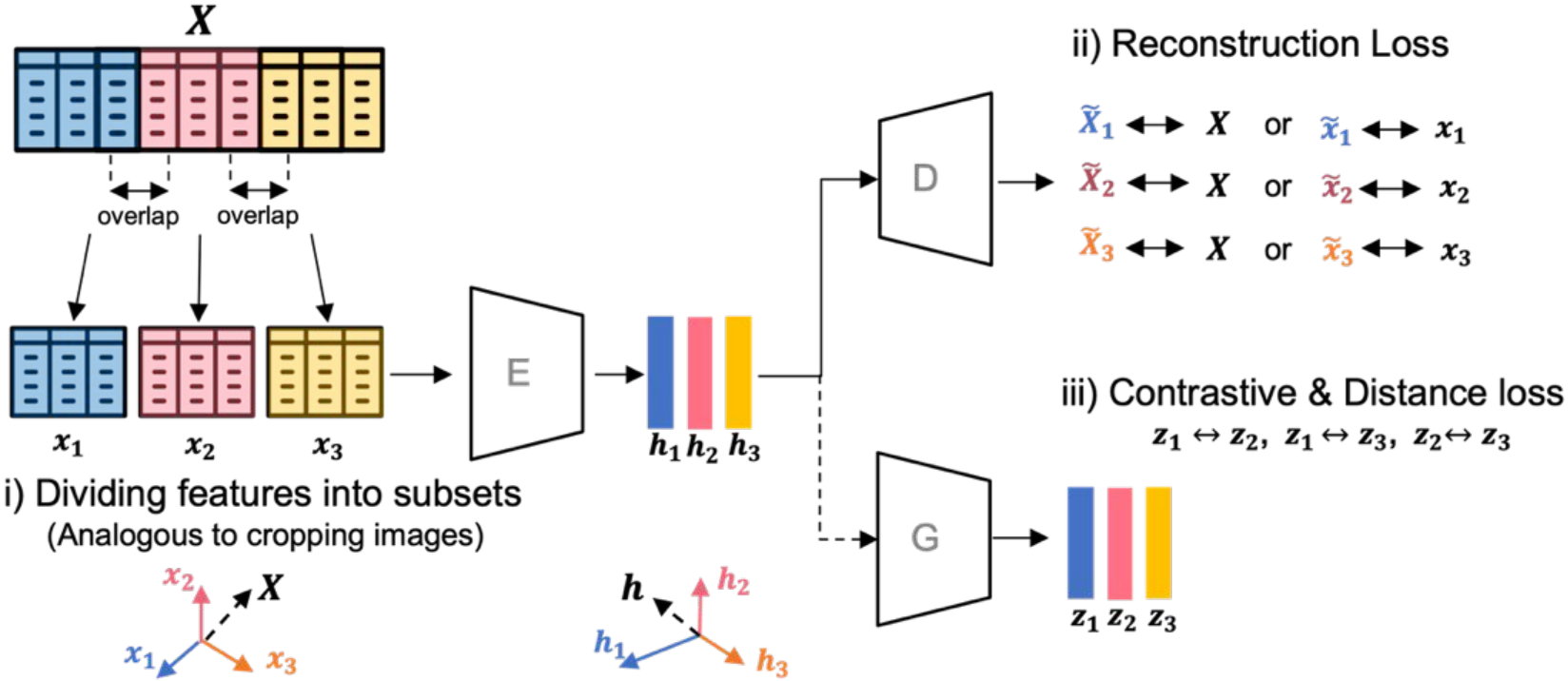


altexsoft

Multi-View Learning

Multi-View Learning combines information from multiple data sources or perspectives.

It enhances SSL by considering different angles of the data, improving generalization.

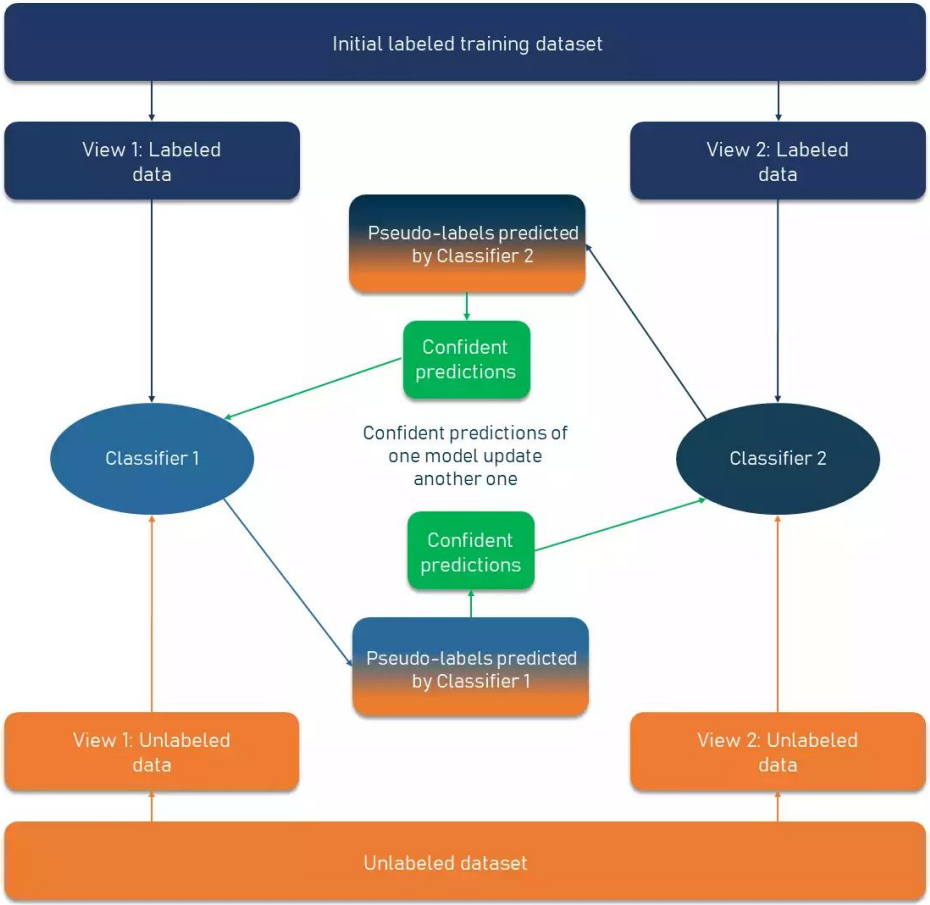

X
overlap
overlap
x_1
x_2
x_3
E
h_1 h_2 h_3
D
G
ii) Reconstruction Loss
$\tilde{X}_1 \leftrightarrow X$ or $\tilde{x}_1 \leftrightarrow x_1$
$\tilde{X}_2 \leftrightarrow X$ or $\tilde{x}_2 \leftrightarrow x_2$
$\tilde{X}_3 \leftrightarrow X$ or $\tilde{x}_3 \leftrightarrow x_3$
iii) Contrastive & Distance loss
$z_1 \leftrightarrow z_2$, $z_1 \leftrightarrow z_3$, $z_2 \leftrightarrow z_3$
z_1 z_2 z_3
i) Dividing features into subsets
(Analogous to cropping images)
x_2
X
x_1
x_3
h
h_2
h_1
h_3

Co-Training

Co-Training involves using multiple weak models, each trained on a different subset of features or views of the data.

Unlabeled data points are labeled based on the consensus of the models.

SEMI-SUPERVISED CO-TRAINING METHOD



altexsoft

Label Propagation

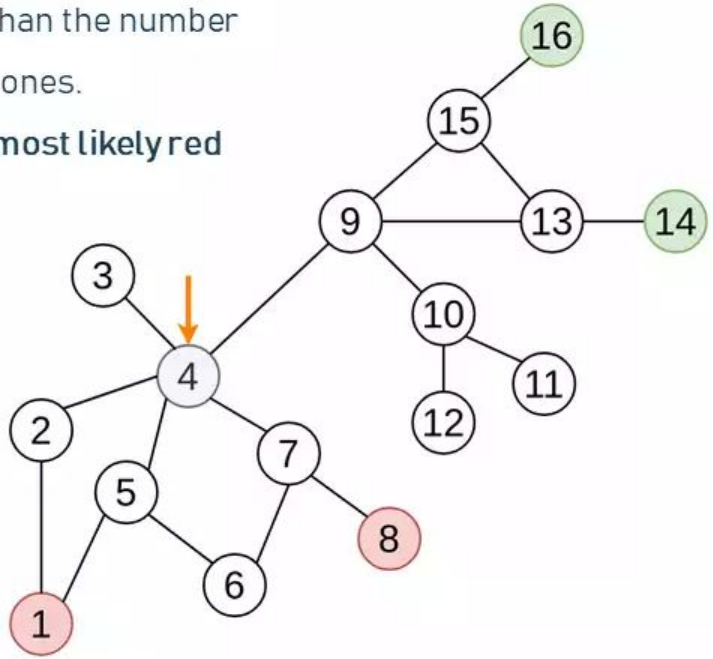
Label Propagation is a technique that spreads labels through a graph or network.

It assigns labels to unlabeled data points based on the labels of their neighbors in the graph.

GRAPH-BASED LABEL PROPAGATION

The number of walks to red nodes is greater than the number of walks to green ones.

Assumption: **4 is most likely red**



Walks that end up in green nodes

1. 4 → 9 → 15 → 16
2. 4 → 9 → 13 → 14
3. 4 → 9 → 13 → 15 → 16
4. 4 → 9 → 15 → 13 → 14

Walks that end up in red nodes

1. 4 → 7 → 8
2. 4 → 7 → 6 → 5 → 1
3. 4 → 5 → 1
4. 4 → 5 → 6 → 7 → 8
5. 4 → 2 → 1

altexsoft

Challenges in Semi-Supervised Learning

The assumption that unlabeled data follows the same distribution as labeled data.

The risk of propagating incorrect labels.

Balancing the utilization of labeled and unlabeled data effectively.

Active Learning

Active Learning is a complementary approach to SSL.

It selects the most informative data points for labeling, reducing the need for extensive labeled data.


Append
Add newly labeled examples to training data
Train
Train model on labeled training data
Active Learning Loop
Annotate
Human experts annotate selected examples
Query
Use acquisition function to select examples from unlabeled data

Conclusion

Semi-Supervised Learning offers a powerful approach to leverage both labeled and unlabeled data.

It is especially valuable in scenarios where obtaining labeled data is challenging.

Consider exploring SSL techniques for various machine learning tasks.

Thank you!