Fundamentals of Machine Learning

07 May 2025

A regional partner of the
Digital Research Alliance of Canada

What is ACENET

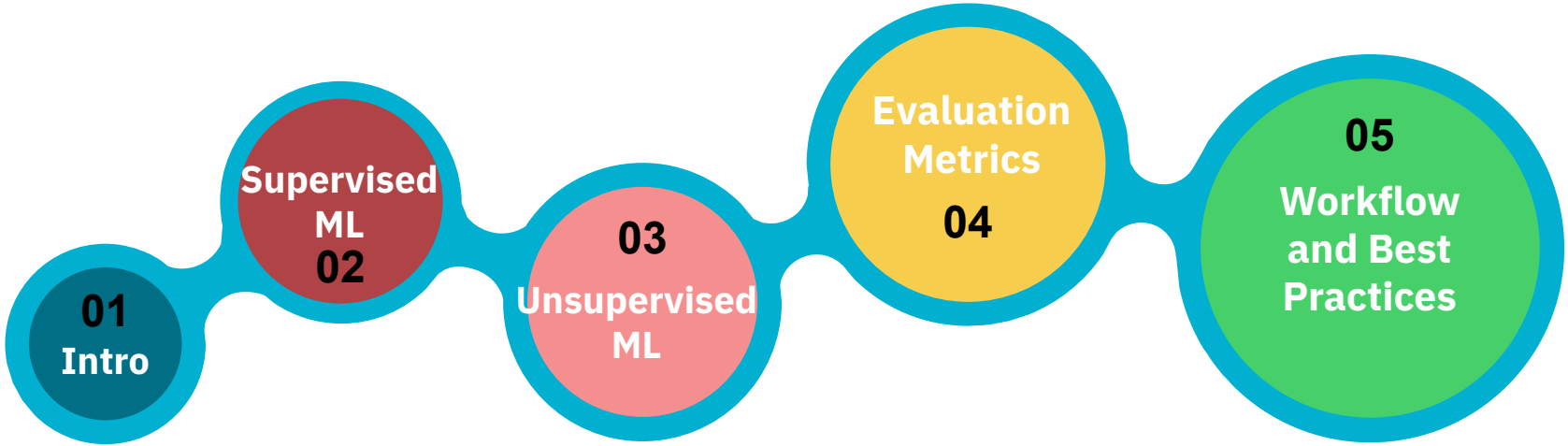
An Atlantic Canadian consortium of universities and community colleges that provides researchers with **access, expertise**, **support** and **training** in digital research resources, including computing and data management.

- Access to thousands of CPUs, GPUs, cloud infrastructure and petabytes of storage when needs outgrow desktop capability.
- Locally-based support across all research disciplines.
- Local training from novice to advanced.
- Regional partner of the Digital Research Alliance of Canada, responsible for digital research infrastructure nationally.

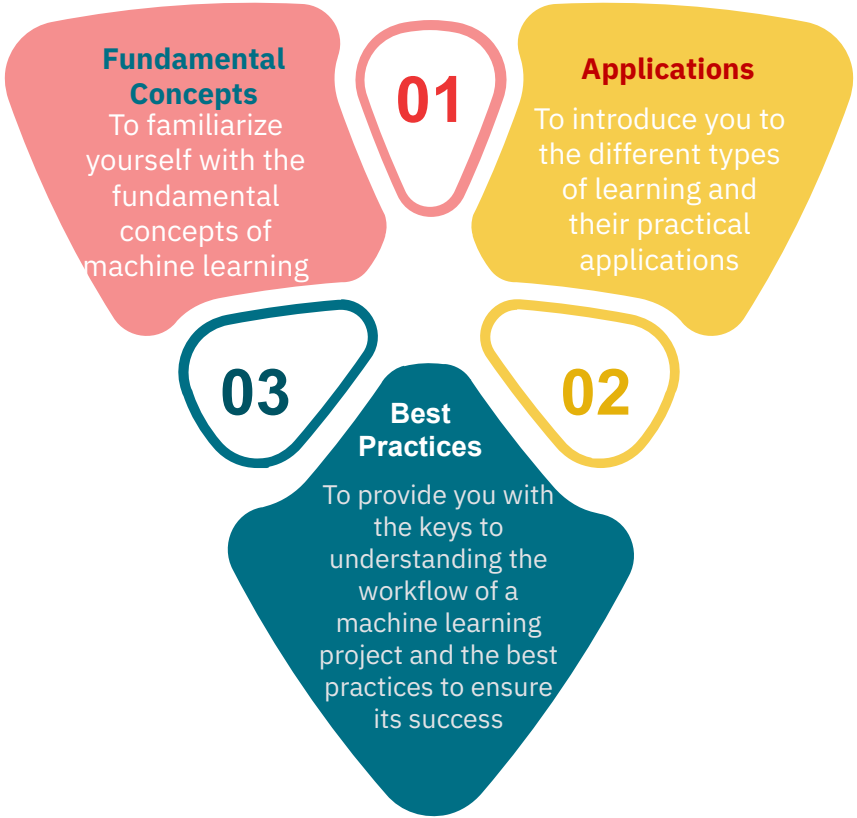
Learning Objectives

- Gain an understanding of the core concept of Machine Learning and its pivotal role in automating tasks.
- Explore and discuss the main types of Machine Learning: supervised and unsupervised learning.
- Define classification and recognize its primary goal within the realm of supervised learning.
- Distinguish between classification tasks, which involve predicting discrete class labels, and regression tasks, which involve predicting continuous numerical values.
- Explore Machine Learning models and their domains of application.

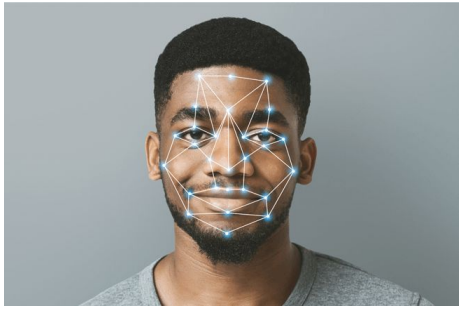
Presentation Outline



Pre-Introduction : More Learning Objectives



Pre-Introduction



01

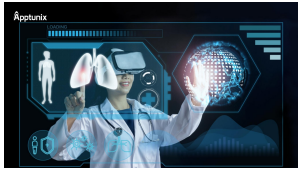
AI

A field in full expansion

02

Application

From facial recognition on our smartphones to self-driving cars



ACENET
accelerate discovery

support@ace-net.ca

What Is ML?

Machine Learning (ML) is a branch of artificial intelligence (AI) and computer science.

Short answer: Applied data science, statistics, and calculus.

ML finds patterns in data, then uses them to predict or model a system's behavior.

Pre-Introduction : AI vs ML

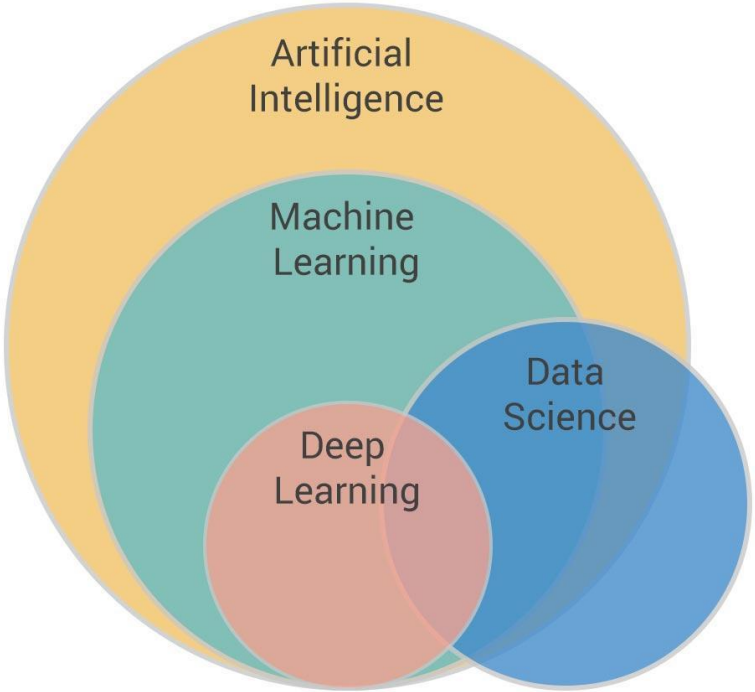
	Artificial Intelligence (AI)	Machine Learning (ML)
Definition	A broader field aimed at creating systems capable of simulating human intelligence	A subset of AI that focuses on learning automatically from data
Objective	Automate intelligent tasks, simulate cognitive processes	Use algorithms to enable systems to learn from data without being explicitly programmed
Approche	Uses various techniques (heuristics, rule-based engines, search algorithms, etc.)	Utilizes statistical models and algorithms to detect patterns in data
Examples of Techniques	Bayesian networks, expert systems, genetic algorithms	Neural networks, decision trees, regressions, SVM (Support Vector Machines)
Interaction with Data	Can function with or without data (e.g., manually defined rules)	Learns solely from data
Examples of Applications	Recommendation systems, speech recognition, autonomous robots	Image classification, spam detection, facial recognition

*We will focus today on **Machine Learning***

Pre-Introduction : AI vs ML

	Artificial Intelligence (AI)	Machine Learning (ML)
Definition	A broader field aimed at creating systems capable of simulating human intelligence	A subset of AI that focuses on learning automatically from data
Objective	Automate intelligent tasks, simulate cognitive processes	Use algorithms to enable systems to learn from data without being explicitly programmed
Approche	Uses various techniques (heuristics, rule-based engines, search algorithms, etc.)	Utilizes statistical models and algorithms to detect patterns in data
Examples of Techniques	Bayesian networks, expert systems, genetic algorithms	Neural networks, decision trees, regressions, SVM (Support Vector Machines)
Interaction with Data	Can function with or without data (e.g., manually defined rules)	Learns solely from data
Examples of Applications	Recommendation systems, speech recognition, autonomous robots	Image classification, spam detection, facial recognition

Pre-Introduction : AI vs ML



*We will focus today on **Machine Learning***

What Machine Learning is **Not?**

- Magic,
  - ML cannot transform poor-quality data into highly accurate models as if by magic.
- An answer for every problem,
  - ML is for specific types of problems, not a magic bullet paid for with technical debt.
  - Though not strictly accurate, ML can learn what a very dedicated person could learn from studying data.
- Omniscience,
  - ML models are limited to the information they have been trained on and can't make decision or generate insights beyond their training data or designed capabilities.

What Machine Learning is **Not?**

- Terribly complicated to get started with,
  - There are dozens of simple models that can be used to gain insight on data.
  - https://scikit-learn.org
- Sentience and Emotion,
  - ML algos do not have emotions, consciousness, or self-awareness, they operate purely through mathematical and statistical methods.
- A replacement for human experts,
  - ML should assist human experts; it is not a replacement.
  - Machines cannot understand nuances, only data.

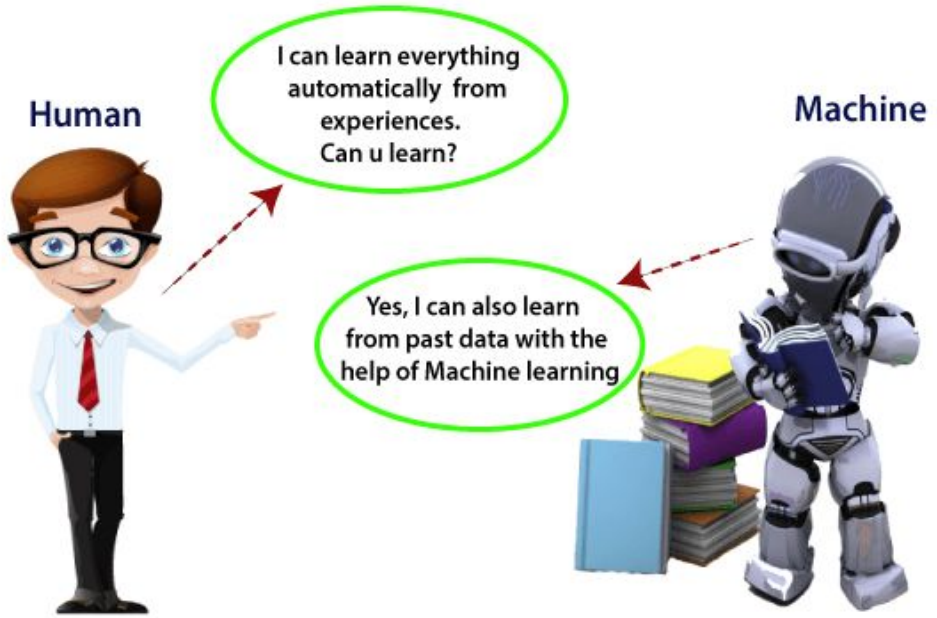
What Machine Learning is **Not?**

- Responsible waiver,
  - You are still responsible for the outcomes of your algorithms.
  - Companies are responsible for their platforms, even if AI makes decisions.
  - There are real-world financial consequences to algorithmic decisions.

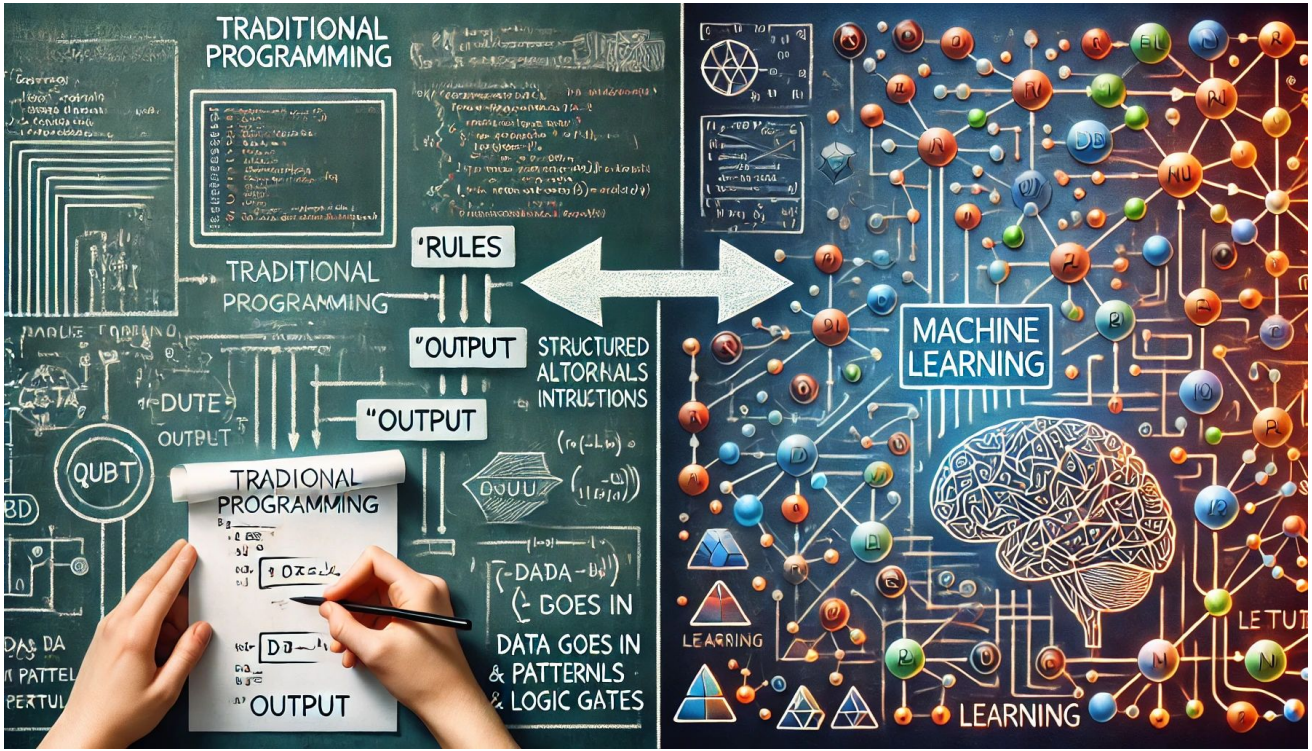
1. Introduction to ML

What is Machine Learning?

Machine Learning (ML) is a field of artificial intelligence that enables computers to learn from data without being explicitly programmed.



Comparison with Traditional Programming



Comparison with Traditional Programming

Comparison with Traditional Programming

Traditional Programming

- **Explicit Programming:** Specific instructions (code) are written for every scenario.
- **Rule-Based:** Follows predefined rules and logic to achieve outcomes.
- **Input Data, Output Result:** Processes input data according to fixed rules, producing a set output.
- **Predictable Behavior:** Always produces the same output given the same input.

Machine Learning

- **Learns from Data:** Discovers patterns and relationships from data without explicit instructions.
- **Driven by Algorithms:** Analyzes data to identify trends and make predictions or decisions.
- **Data as Input and Teacher:** Continuously learns and improves performance using data.
- **Adaptive and Evolving:** Refines its model with new data for more accurate predictions.

Applications of Machine Learning in 2025

E-Commerce

Widely used in the field of e-commerce, as it helps the organization establish strong engagement between the user and the business.

Education

It helps faculty and students by recommending courses, analyzing certain data, and making decisions related to the student, etc.

Robotics

ML enables robots to make real-time decisions and increase productivity.

Healthcare

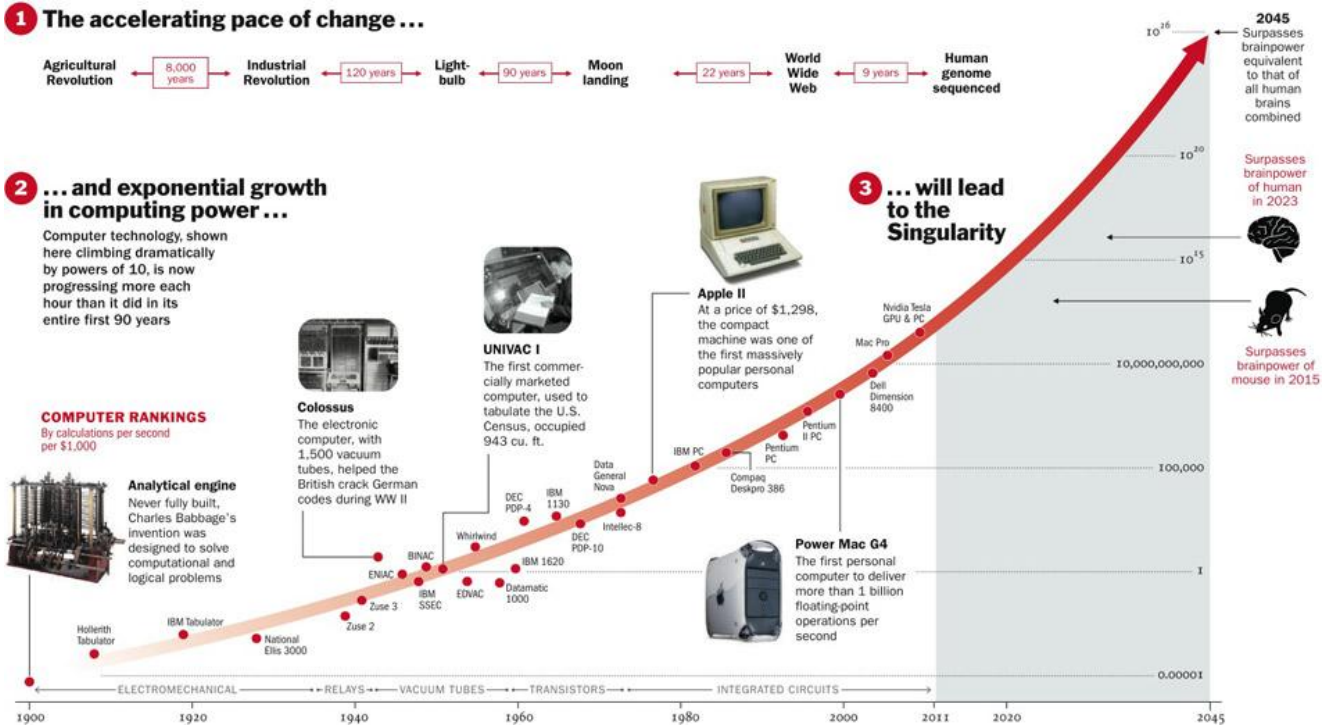
Different ML algorithms are used to build accurate machines capable of detecting minor diseases within the human body.

Agriculture

It is used to detect various parameters such as the amount of water and moisture, the level of nutrient deficiencies, etc., in the soil.

Why is Machine Learning Important Today?

The Exponential Growth of Computing Power

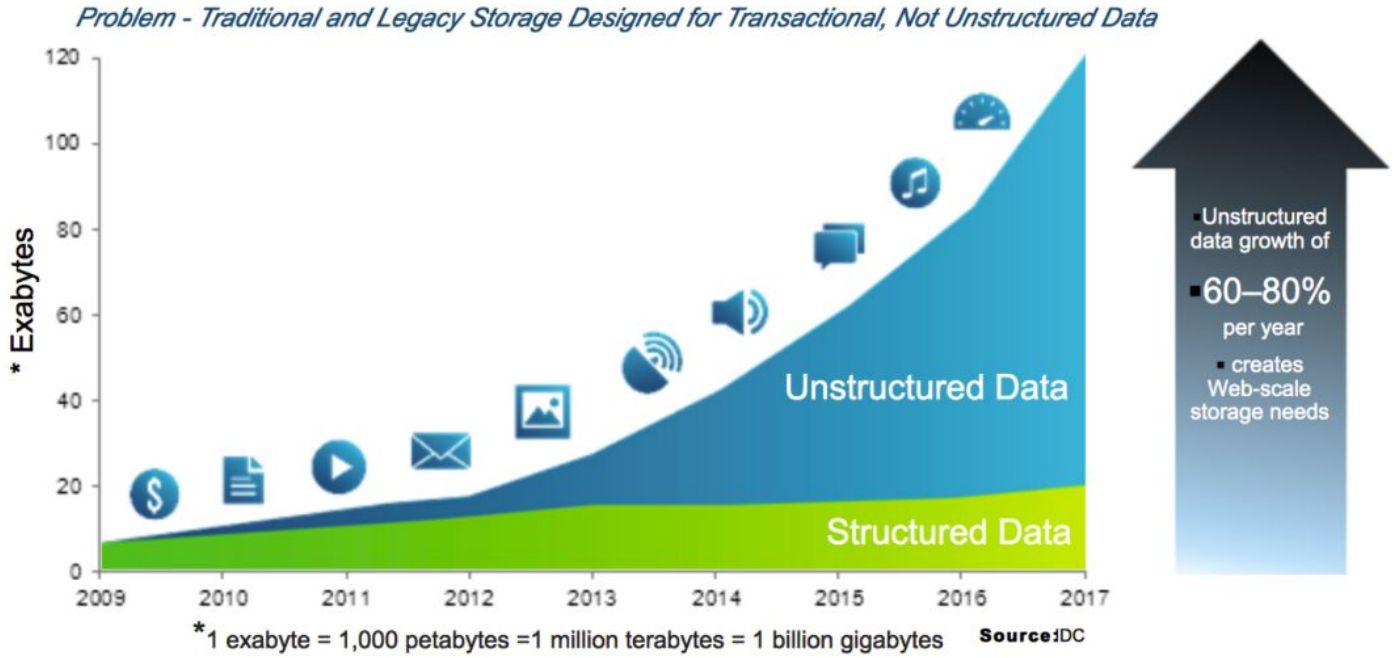


ACENET
accelerate discovery

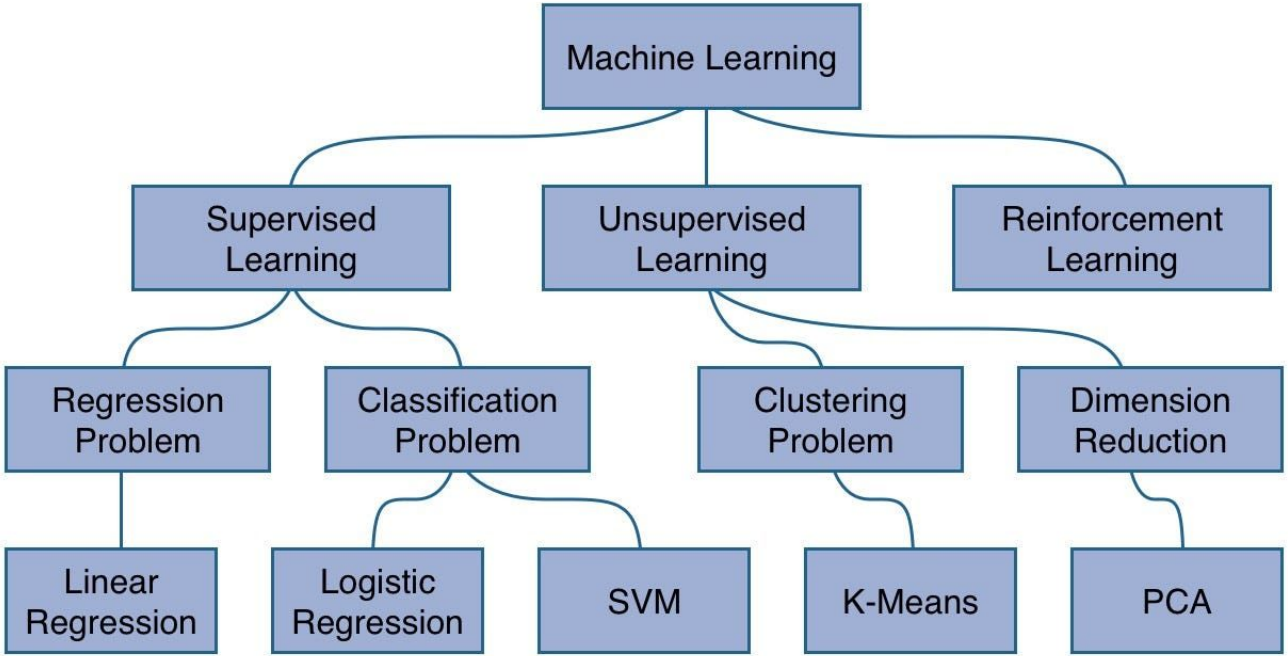
support@ace-net.ca

Why is Machine Learning Important Today?

Exponential growth of data



Types of Machine Learning



2. Supervised Learning

Learning Through Labels

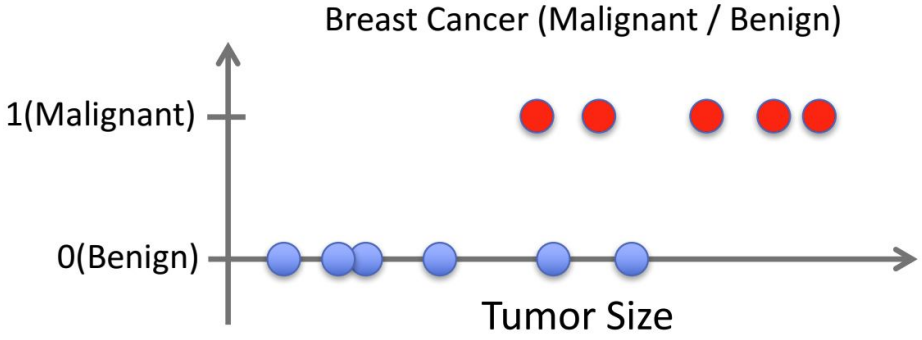
Supervised Learning - Techniques

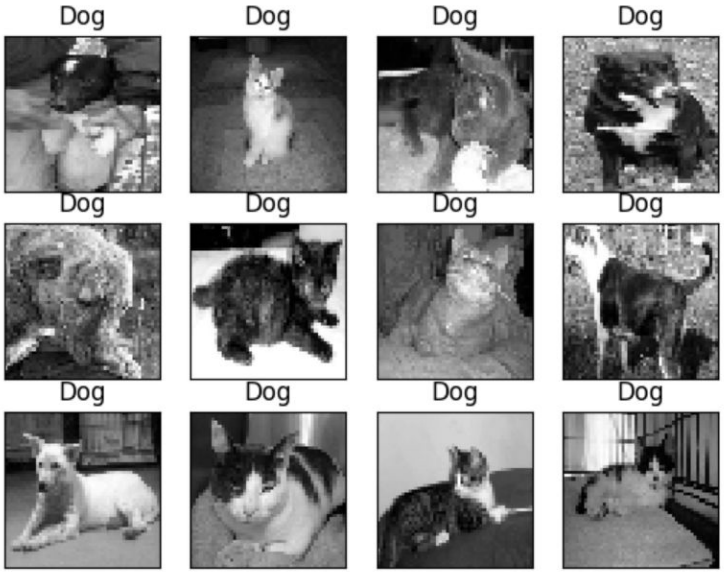
Classification – Placing objects into categories.

Regression – Predictions on a spectrum.

Supervised Learning

Supervised learning is a type of machine learning where the model is trained on labeled examples, meaning examples for which the desired output is known.





Supervised Learning

Supervised learning is a type of machine learning where the model is trained on labeled examples, meaning examples for which the desired output is known.

Regression

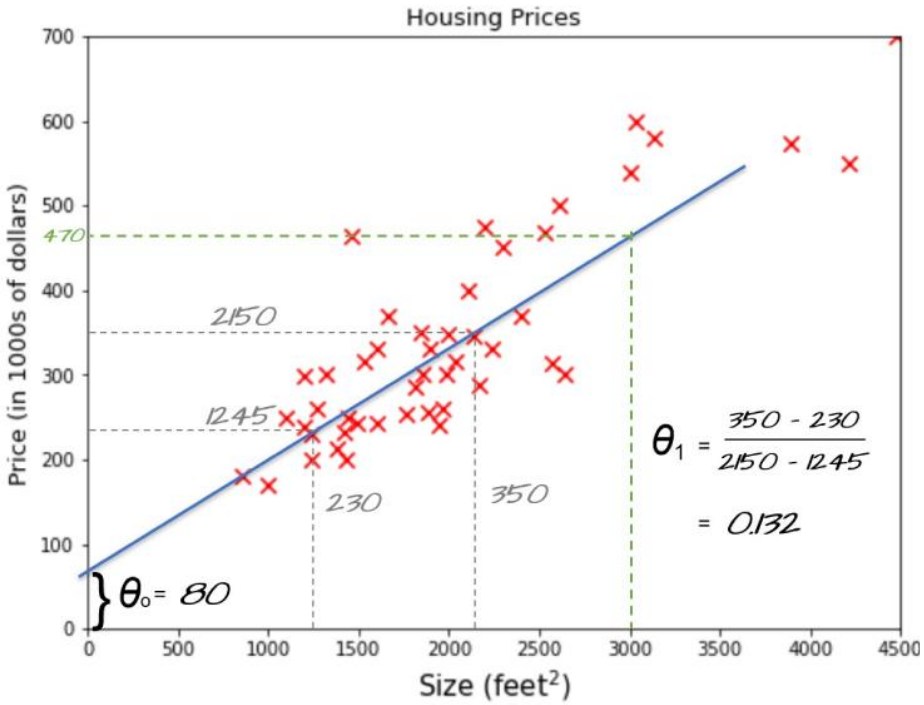
Supervised Learning Methods

Linear Regression

The model learns a linear function (a line in the case of two dimensions) that allows it to predict the output from the inputs.

Formula: $y = w_1x_1 + w_2x_2 + \ldots + w_nx_n + b$

- **Y** is the predicted output,
- $x_1, x_2, \ldots, x_n$ are the input features,
- $w_1, w_2, \ldots, w_n$ are the corresponding weights (parameters) of the model, tilts the slope of the line
- **b** is the bias term, which allows the line to be shifted up or down.



Linear Regression - Advantages and Disadvantages

Advantages:

- Simple to understand and implement.
- Easy to interpret: weights show how much each feature influences the prediction.
- Computationally efficient

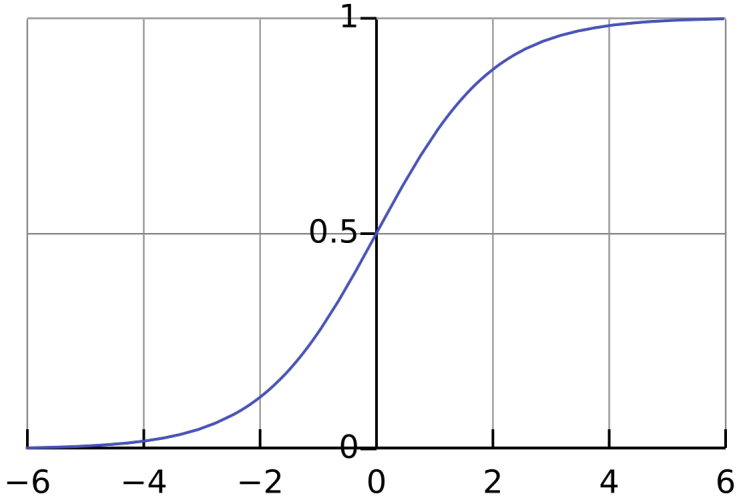
Disadvantages:

- Cannot model non-linear relationships between variables.
- Sensitive to outliers (extreme values).

Logistic Regression

Logistic regression uses a sigmoid function to predict the probability of belonging to a class (0 or 1). The sigmoid function converts any real-valued input into a value between 0 and 1, making it ideal for binary classification tasks.

- **Sigmoïde Function**: σ(z) = 1 / (1 + exp(-z))
  - Z is the linear combination of inputs and weights, similar to linear regression.
  - The sigmoid function transforms z into a value between 0 and 1, representing probability.



$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Logistic Regression - Advantages and Disadvantages

Advantages:

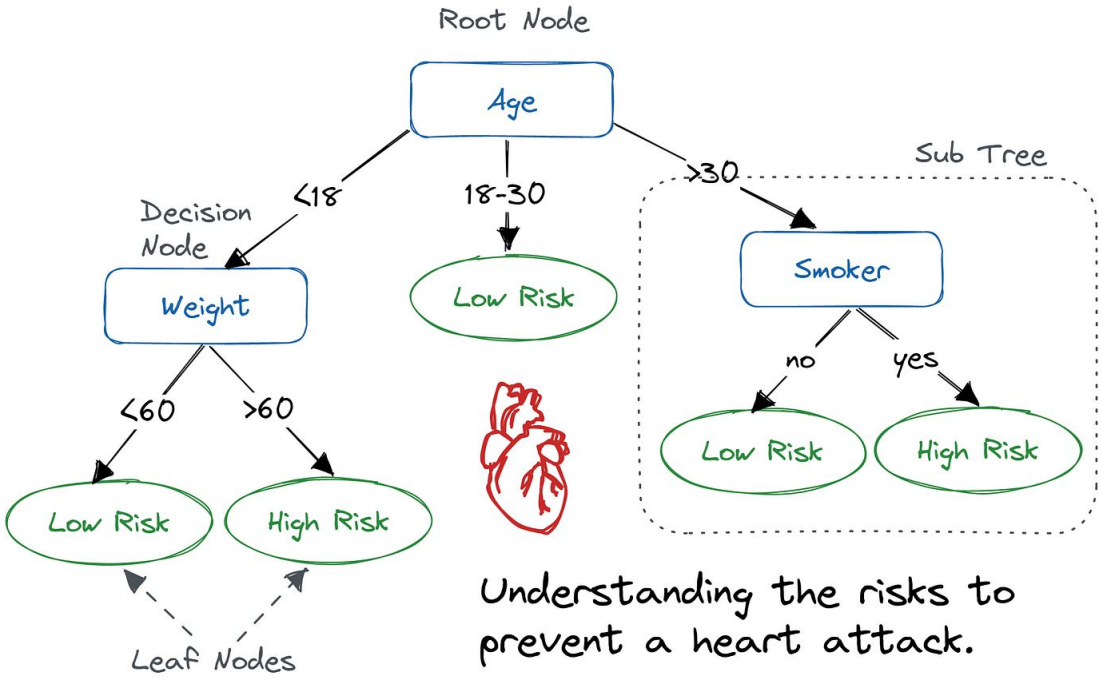
- Simple and easy to interpret.
- Provides probabilistic outputs (confidence scores).
- Robust to moderate outliers.

Disadvantages:

- Assumes a linear relationship (struggles with non-linear patterns).
- Sensitive to imbalanced data.
- Requires feature scaling for accurate predictions.

Decision Trees

The model learns a tree structure that allows it to make decisions based on the values of the attributes.



Decision Trees - Advantages and Disadvantages

Advantages:

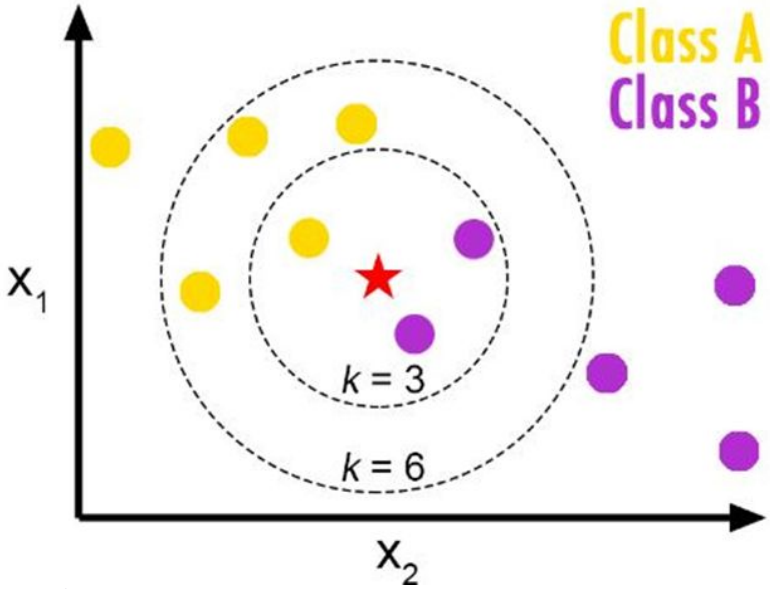
- Easy to understand and interpret: the decision process can be visualized.
- Can handle both numerical and categorical data.
- Robust to outliers.

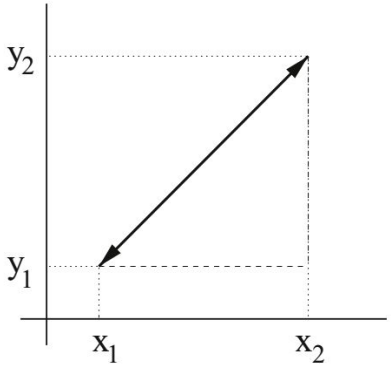
Disadvantages:

- Can overfit the training data if the tree is too deep.
- Sensitive to small variations in the data.

k-Nearest Neighbors (k-NN)

The model classifies a new example based on the classes of its k nearest neighbors in the feature space.





$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

$$d_E(\mathbf{x}, \mathbf{y}) = \sqrt{\Sigma_{i=1}^{n}(x_i - y_i)^2}$$

k-Nearest Neighbors - Advantages and Disadvantages

Advantages:

- Simple to understand and implement.
- No need for explicit training: learning happens during classification.

Disadvantages:

- Computationally expensive: distances need to be calculated for all training examples.
- Sensitive to data dimensionality: performance decreases with many attributes.
- Affected by the choice of distance metric and the value of **k**.

Conclusion: Diversity of Supervised Models

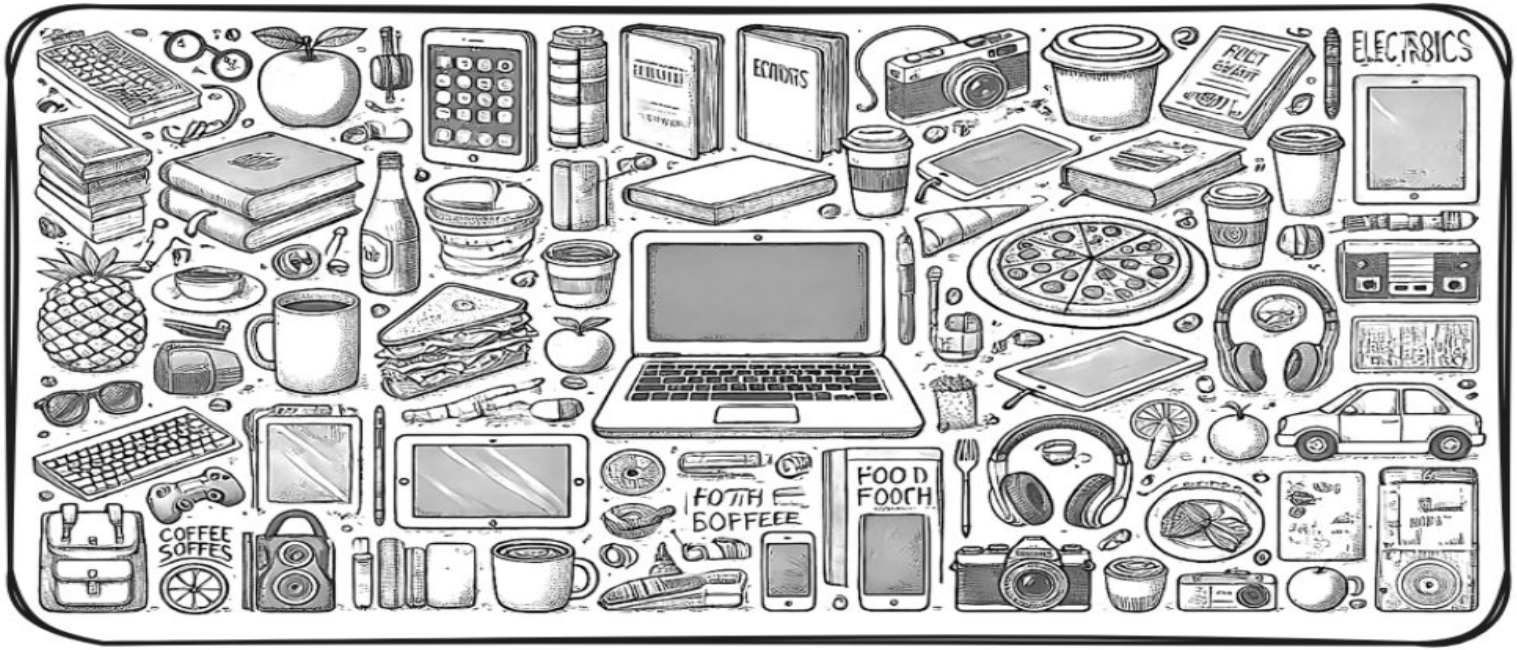
- We have reviewed the different algorithms, along with their strengths and limitations.
- The choice of the best model depends on the problem, the data, and the goals of the application.
- No single model is perfect for all tasks.

But what can be done when the data is not labeled?

3. Unsupervised Learning

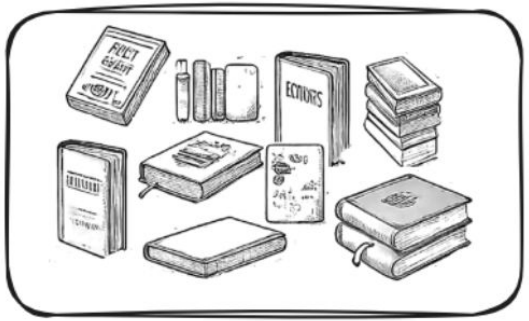
Unsupervised Learning: Discovering hidden structures in data

Unsupervised learning is a form of machine learning where the model is trained on unlabeled data.

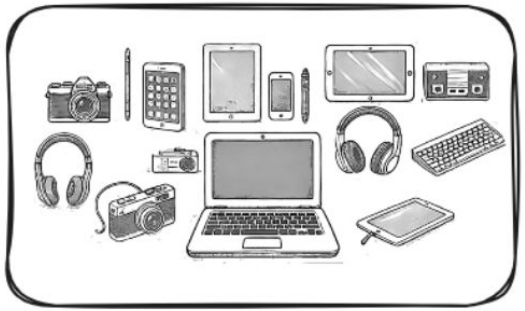


Unsupervised Learning: Discovering hidden structures in data

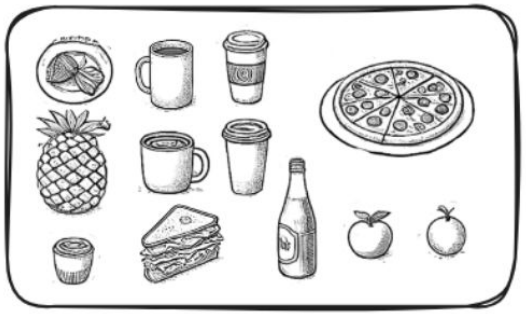
Unsupervised learning is a form of machine learning where the model is trained on unlabeled data.



Books



Devices



Food

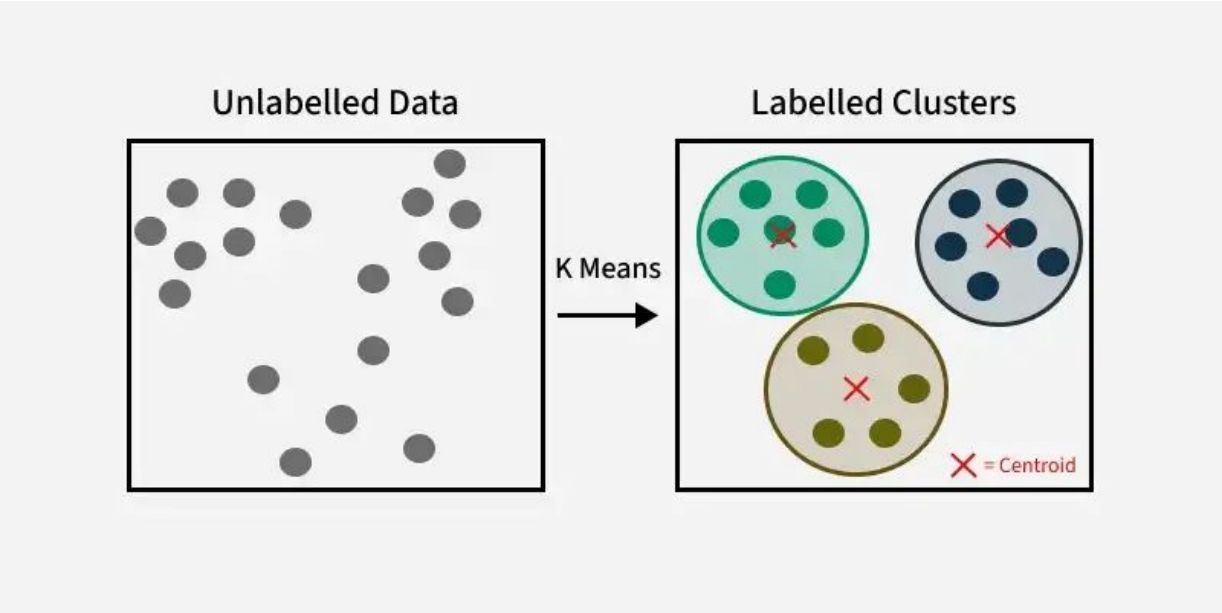
Unsupervised Learning: Key Points

- Unsupervised learning is used for **clustering** and **dimensionality reduction**.
- Interpreting the results is often more complex than in supervised learning.
- No need for labeled data, making it useful when labeled data is unavailable.
- Useful for data exploration, revealing hidden patterns and relationships.

Unsupervised Learning Methods

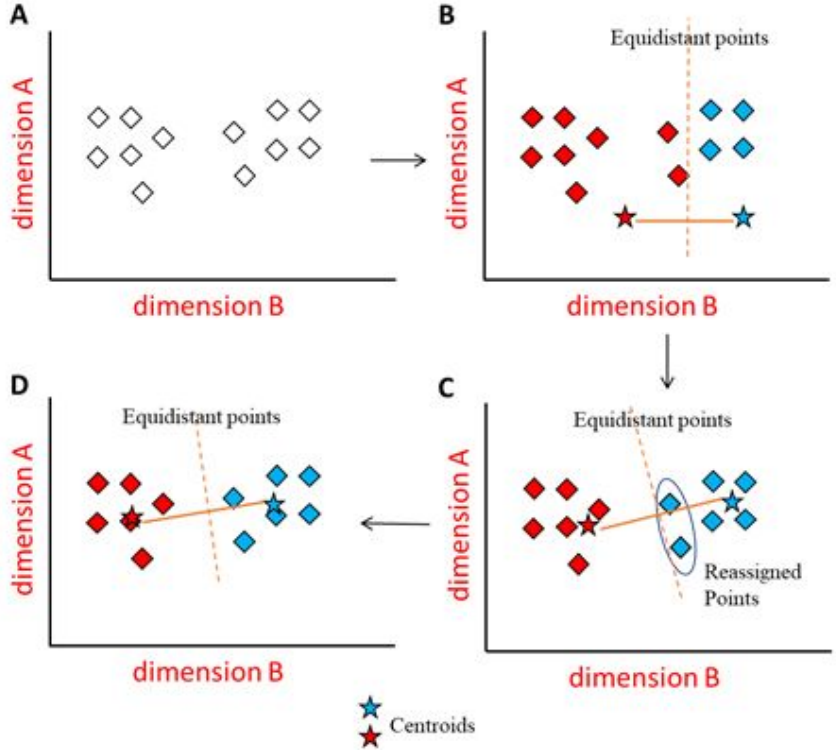
k-means Clustering

The algorithm seeks to partition the data into k clusters by minimizing the distance between the examples and the center (centroid) of their cluster.



k-means Clustering

The algorithm seeks to partition the data into k clusters by minimizing the distance between the examples and the center (centroid) of their cluster.



k-means Clustering - Advantages and Disadvantages

Advantages:

- Simple and easy to implement.
- Efficient for large datasets.
- Scales well with data size.

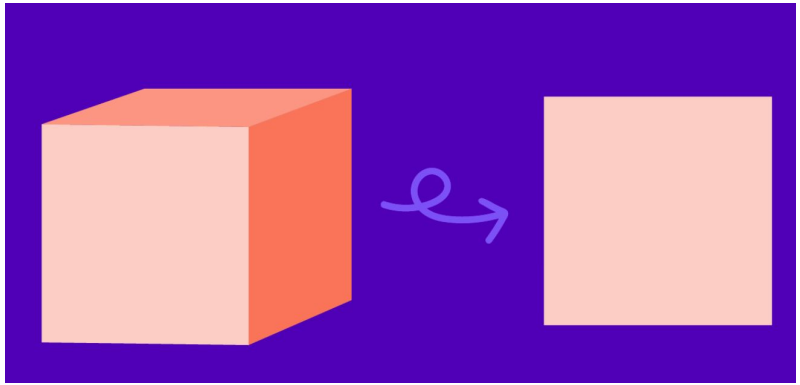
Disadvantages:

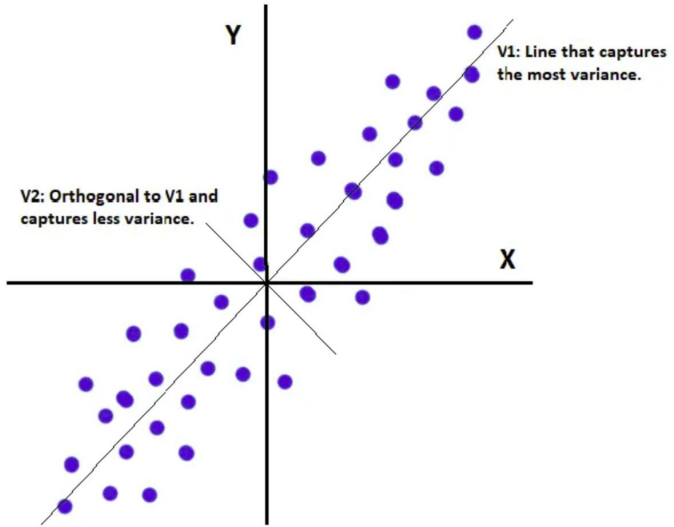
- Sensitive to initial centroids.
- Struggles with non-spherical clusters.
- Requires predefining the number of clusters (k).

Principal Component Analysis (PCA): Dimensionality reduction for visualization and analysis.

PCA is a statistical technique that transforms a set of correlated variables into a set of uncorrelated variables (principal components).

- **Principle:** PCA identifies the directions of greatest variance in the data and projects the data onto these directions, allowing for dimensionality reduction while preserving essential information.





Principal Component Analysis (PCA): Advantages and Disadvantages

Advantages:

- Allows for the visualization of multidimensional data.
- Reduces noise and redundancy in the data.
- Improves the performance of certain machine learning algorithms.

Disadvantages:

- Can be difficult to interpret: Principal components do not always correspond to real-world concepts.
- Sensitive to outliers.

Principal Component Analysis : Applications

1. **Image Compression:** Reducing the dimensionality of image data while preserving essential information, which is crucial for image storage and transmission.

2. **Bioinformatics:** Analyzing multidimensional gene expression data to identify patterns and reduce noise.

3. **Facial Recognition:** Extracting essential facial features for recognition tasks.

4. **Recommendation Systems:** Reducing the dimensionality of user-item interaction data for efficient recommendation algorithms.

5. **Finance:** Analyzing financial data to identify trends and underlying patterns.

Applications of Unsupervised Learning

- **Data Visualization:** Analyze and explore complex data by projecting it onto a reduced-dimensionality space.
- **Customer Segmentation:** Group customers with similar behaviors for targeted marketing campaigns.
- **Anomaly Detection:** Identify suspicious behavior in financial transactions or security systems.
- **Data Compression:** Reduce the size of data for storage and transmission.
- **Data Preprocessing:** Improve the performance of supervised learning algorithms by reducing noise and redundancy.

Conclusion: Unsupervised learning reveals the secrets of unlabeled data.

- The main techniques of unsupervised learning and their applications.
- The importance of unsupervised learning for knowledge discovery and data exploration.

But what happens when the relationships become more complex?

4. Evaluation Metrics

Evaluation Metrics: How to Measure Performance?

In machine learning, evaluation is crucial for understanding a model's performance and choosing the best model for a given application.

		Labels returned by the classifier	
		pos	neg
True labels:	pos	N_{TP}	N_{FN}
	neg	N_{FP}	N_{TN}

Evaluation of Supervised Models : Error & Accuracy

The error rate of a classifier, E, is the frequency of errors made by the classifier on a given set of examples.

		Labels returned by the classifier	
		pos	neg
True labels:	pos	N_{TP}	N_{FN}
	neg	N_{FP}	N_{TN}

$$E = \frac{N_{FP} + N_{FN}}{N_{FP} + N_{FN} + N_{TP} + N_{TN}}$$

Sometimes, the engineer prefers to work with the opposite quantity, classification accuracy, Acc: the frequency of correct classifications made by the classifier on a given set of examples.

$$Acc = \frac{N_{TP} + N_{TN}}{N_{FP} + N_{FN} + N_{TP} + N_{TN}}$$

Evaluation of Supervised Models: Recall & Precision

Precision represents the percentage of true positives (TP) among all the examples that the classifier has labeled as positive.

$$Pr = \frac{N_{TP}}{N_{TP} + N_{FP}}$$

In other words, precision is the probability that the classifier is correct when it labels an example as positive.

Recall represents the probability that a positive example is correctly recognized as such by the classifier.

$$Re = \frac{N_{TP}}{N_{TP} + N_{FN}}$$

		Labels returned by the classifier	
		pos	neg
True labels:	pos	N_{TP}	N_{FN}
	neg	N_{FP}	N_{TN}

Note that both Recall and Precision differ only by the denominator. This makes sense.

While Precision is the frequency of true positives among all the examples labeled as positive by the classifier, Recall is the frequency of those same true positives among all the positive examples in the dataset.

Evaluation of Supervised Models: Example

Recall represents the probability that a positive example is correctly recognized as such by the classifier.

		Labels returned by the classifier	
		pos	neg
True labels:	pos	20	50
	neg	30	900

We want to obtain Precision, Recall, and Accuracy:

$$\text{precision} = \frac{20}{50} = 0.40 \quad \text{recall} = \frac{20}{70} = 0.29 \quad \text{accuracy} = \frac{920}{1000} = 0.92$$

Evaluation of Supervised Models: Example

$$\text{precision} = \frac{20}{50} = 0.40 \qquad \text{recall} = \frac{20}{70} = 0.29 \qquad \text{accuracy} = \frac{920}{1000} = 0.92$$

		Labels returned by the classifier	
		pos	neg
True labels:	pos	30	70
	neg	20	880

- The example we just reviewed illustrates the behavior of the two metrics in a simple domain with an imbalanced representation of two classes, pos and neg.
- The induced classifier, although it shows an impressive classification accuracy, suffers from low precision and low recall.
- More specifically, a precision of 0.40 means that out of the 50 examples labeled as positive by the classifier, only 20 are actually positive, while the remaining 30 are false positives.
- As for recall, the situation is even worse: out of the 70 positive examples in the test set, only 20 are correctly identified as such by the classifier.

5: Practical Workflow in Machine Learning and Best Practices

Best Practices - Before you Start a Machine Learning project

- **Define the Problem Clearly:** Understand the problem you are trying to solve and how success will be measured.
- **Collect and Explore Data:** Ensure you have the right data and spend time understanding its structure, quality, and gaps.
- **Set Clear Goals:** Establish what you want to achieve, like accuracy targets or specific outputs.
- **Choose the Right Tools:** Select the appropriate frameworks, libraries, and tools for your project.

Best Practices - While You Work on the ML Project

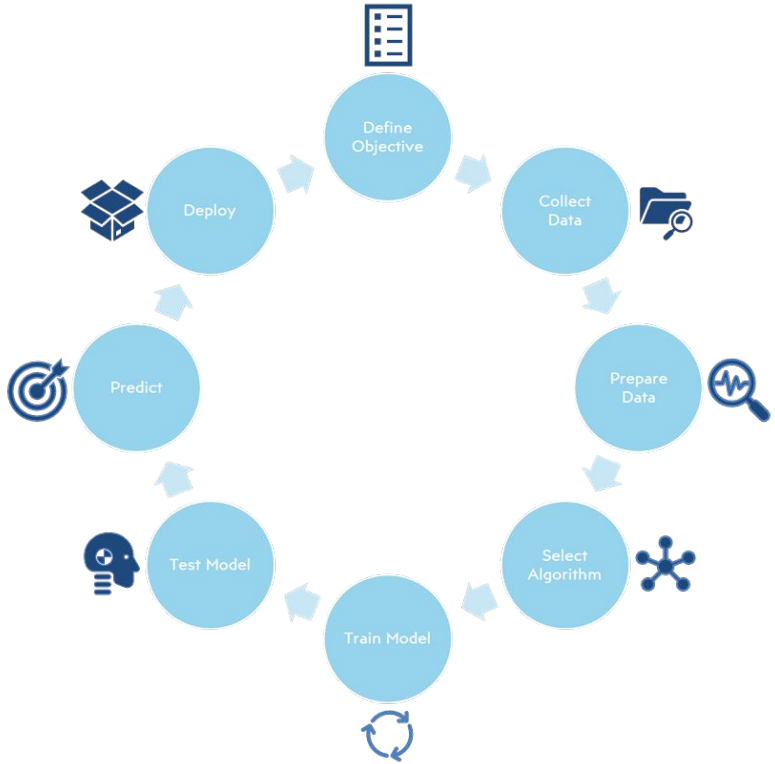
- **Clean and Preprocess Data:** Handle missing data, remove outliers, and normalize your data for consistency.
- **Split Data into Training and Test Sets:** Use one set to train the model and another to evaluate its performance.
- **Experiment with Different Models:** Try different algorithms to see which performs best for your data.
- **Use Cross-Validation:** Regularly check model performance using cross-validation to avoid overfitting.
- **Document Your Process:** Keep notes on the steps you take, including the models tried and their performance metrics.

Best Practices - After You Finish the ML Project

- **Evaluate the Model Thoroughly**: Use relevant metrics (accuracy, precision, recall) and analyze model behavior.
- **Deploy the Model**: Implement the model into production in a way that's scalable and easy to use.
- **Monitor the Model in Production**: Continuously track how the model performs with new data and make updates as needed.
- **Ensure Model Interpretability**: Make sure the predictions are explainable, especially in sensitive applications like healthcare or finance.
- **Plan for Retraining**: Set a schedule to retrain the model as data evolves or new information becomes available.

Practical Workflow in Machine Learning

- **Data Collection:** Gather relevant data from various sources.
- **Data Preprocessing:** Clean, transform, and normalize the data to prepare it for modeling.
- **Model Selection:** Choose the appropriate algorithm based on the problem type and data.
- **Model Training:** Train the model using the prepared data.
- **Model Evaluation:** Assess the model's performance using metrics like accuracy, precision, recall.
- **Model Deployment:** Deploy the model into a production environment for use.



Google Colab

ACENET
accelerate discovery

support@ace-net.ca

Google Colab

- Google Colab, short for Google Colaboratory, is a platform offered free of charge by Google that lets you write and run python code in your browser.
- Colab is a hosted Jupyter Notebook service that requires no setup to use and provides free access to computing resources, including GPUs and TPUs.
- Google Colab is great if you need to work across multiple devices as it syncs seamlessly across devices.

Google Colab

- To start, go to: https://colab.google

Responsible machine learning is crucial to ensuring the ethical and sustainable development of AI.

Thank you!

ACENET
accelerate discovery

support@ace-net.ca

Questions ?

ace-net.ca

info@ace-net.ca

support@ace-net.ca