Introduction to Machine Learning

By Jeewaka Perera

Lesson Objectives

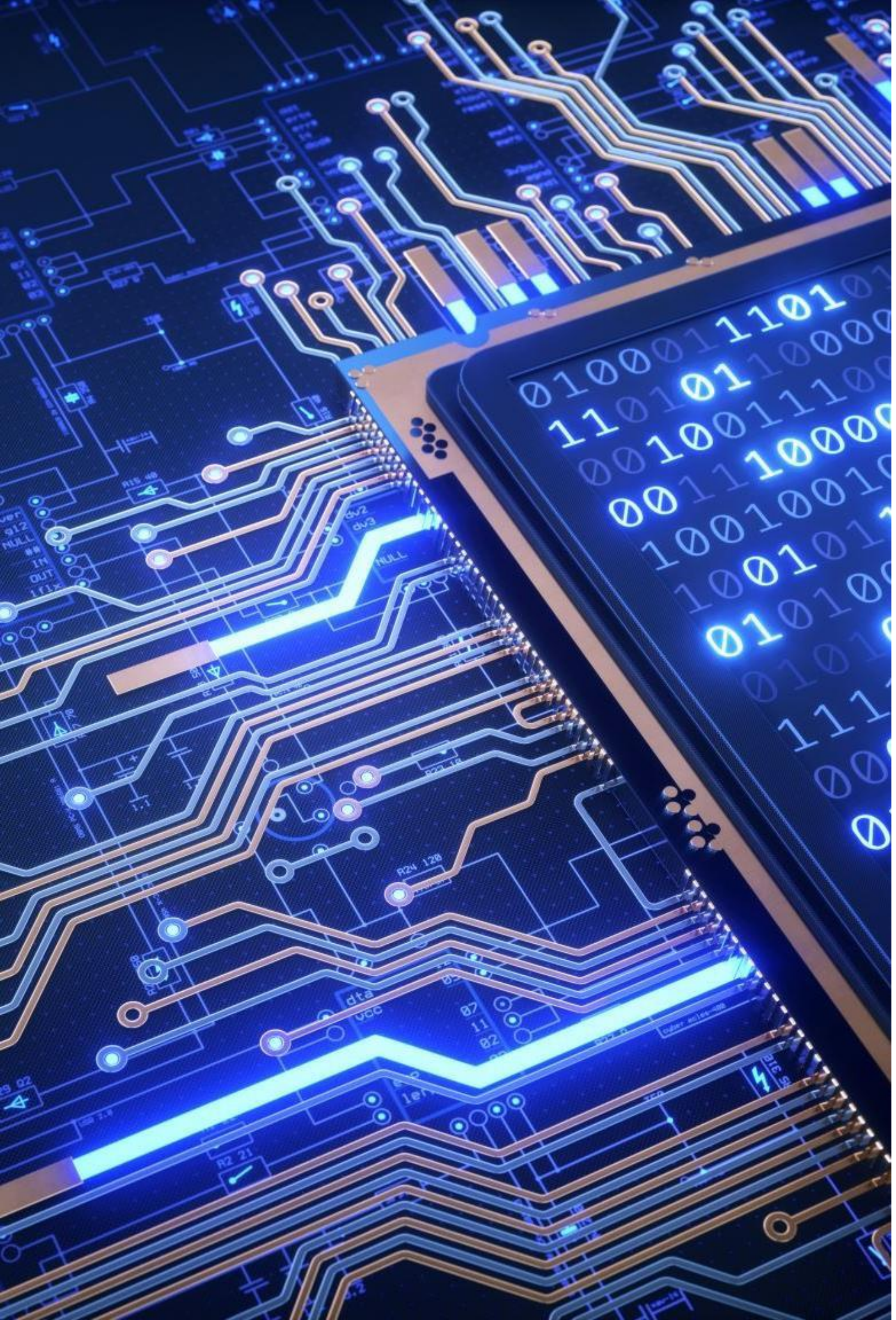
- At the end of the lesson students should be able to explain
  - What is Machine Learning?
  - Why Choose Machine Learning?
  - Different Machine Learning Algorithms and their applications

What is Optimization?

- **Optimization** is the mathematical discipline which is concerned with finding the maxima and minima of functions, possibly subject to constraints.

What is Artificial Intelligence

- " It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable.“ by John McCarthy



What is Machine Learning

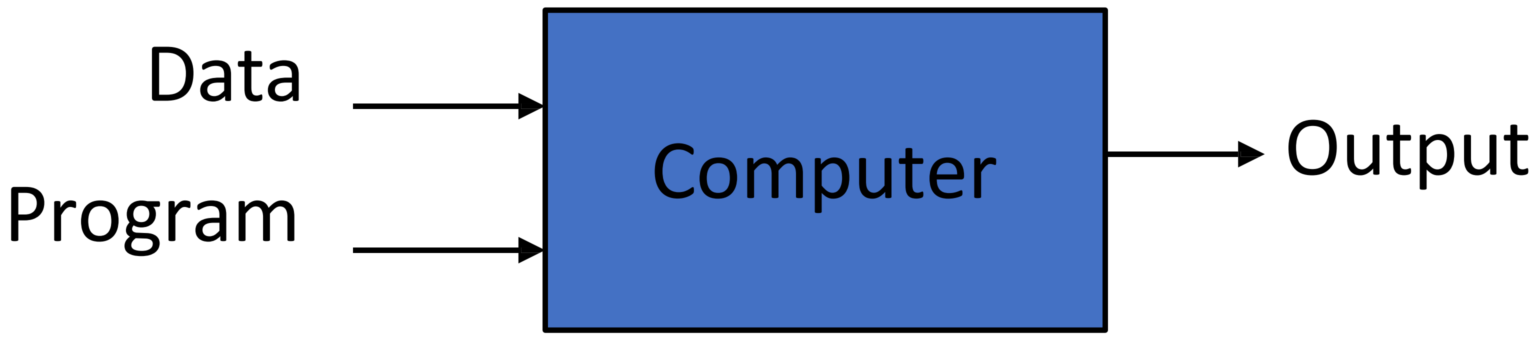
- “field of study that gives computers the ability to learn without explicitly being programmed.” by Arthur Samuel
- Machine learning is a subfield of artificial intelligence, which is broadly defined as the capability of a machine to imitate intelligent human behavior.

Learning in a Machine

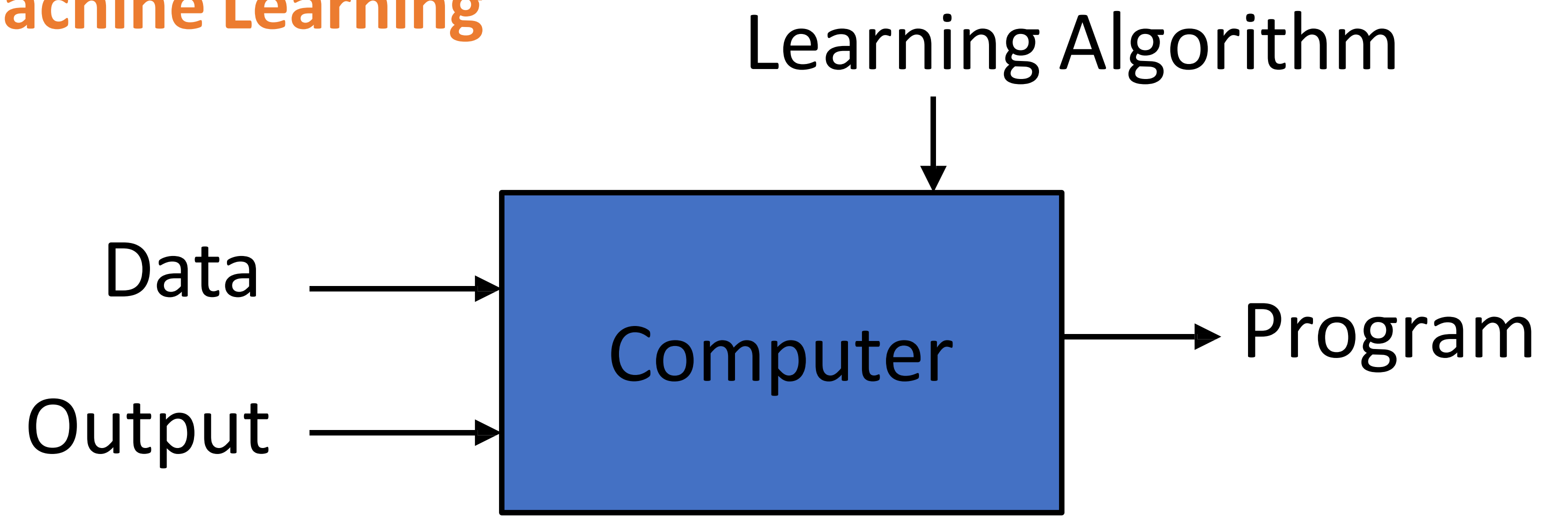
- "A computer program is said to learn from experience (E) with some class of tasks (T) and a performance measure (P) if its performance at tasks in T as measured by P improves with E"

ML vs Traditional programming

Traditional Programming



Machine Learning



Types of Artificial Intelligence Algorithms

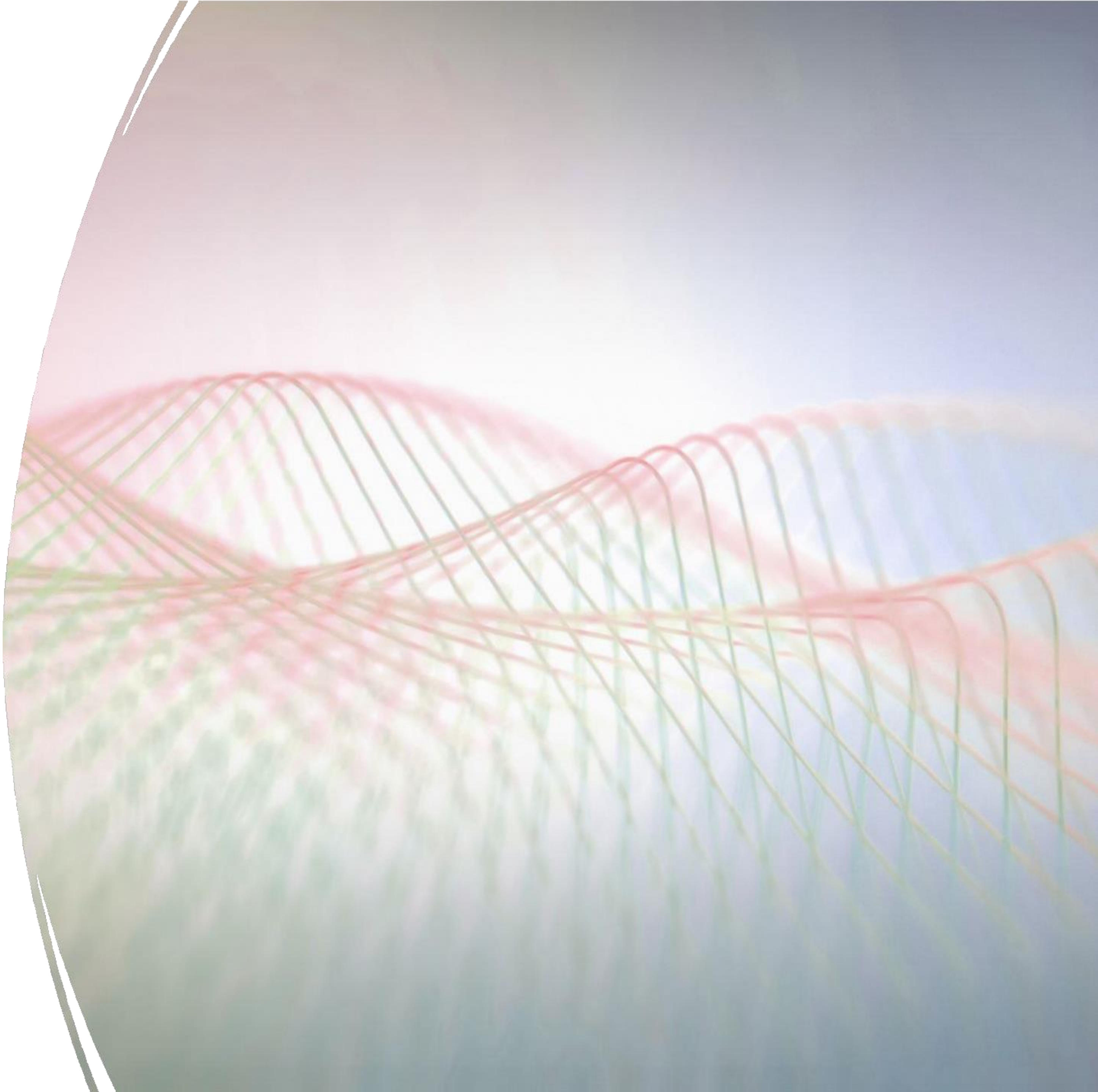
Rule-based expert

Systems

Search Algorithms

Evolutionary Algorithms and Swam Intelligence Algorithms

Machine Learning Algorithms

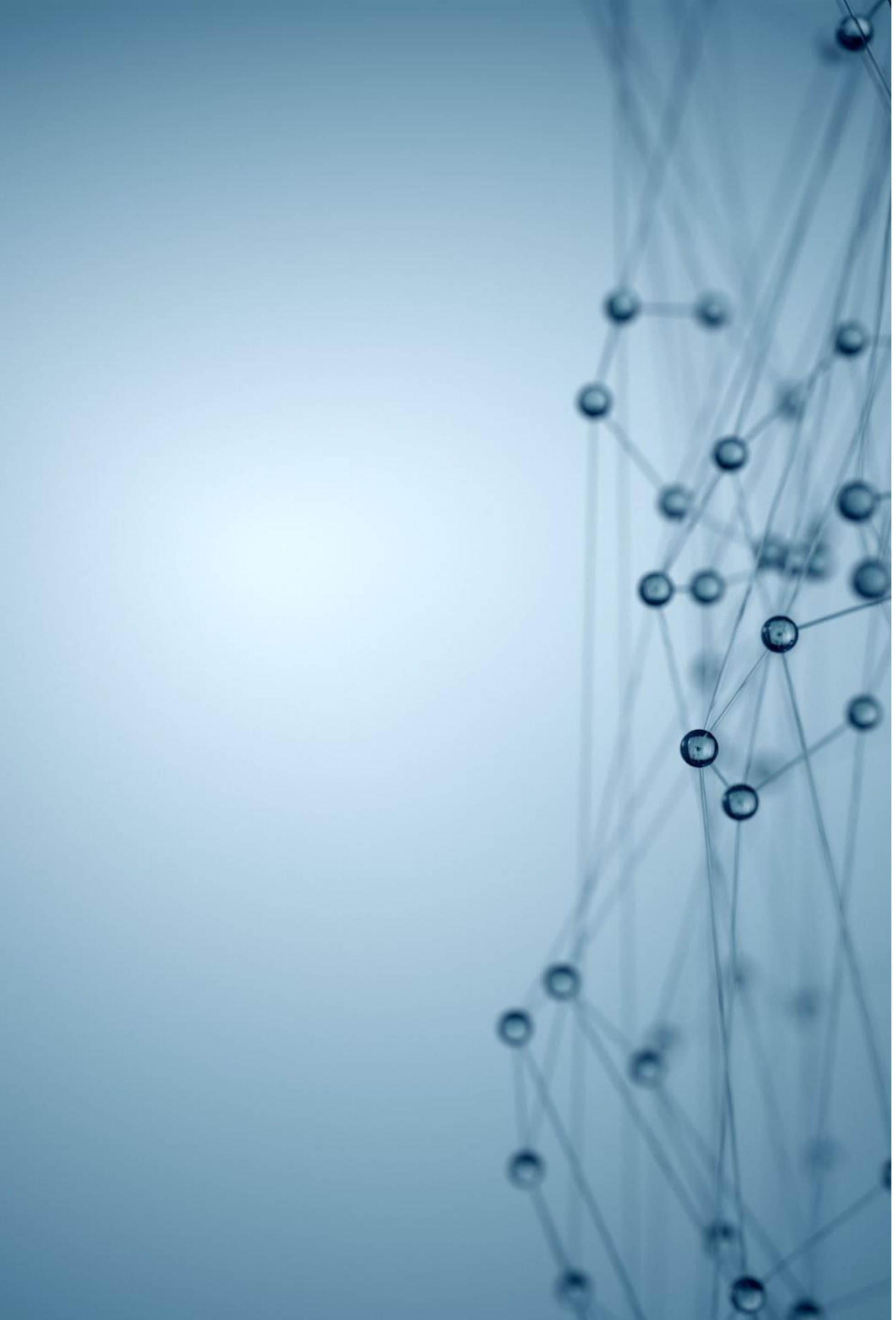


Rule Based Systems

- Expert Systems
- [PROSPECTOR]
- [MYCIN]
- Based on pre-defined Rules
- Rules defined based on domain knowledge
- Designed to mimic the decision-making process of human experts

Search Algorithms

- Breadth-First Search
- Depth First Search
- Iterative Deepening Search
- Uniform Cost Search
- Dijkstra's algorithm
- A* Search



Evolutionary Algorithms

- Evolutionary Algorithms are a family of nature-inspired optimization algorithms that mimic biological evolution—natural selection, mutation, recombination, and survival of the fittest.
  - Genetic Algorithms
  - Particle Swarm Optimization
  - Cultural Algorithms
  - HCA KCA
  - SI Algorithms(ACO, FA)

Steps in a generic Evolutionary Algorithm

Initialize a population of potential solutions randomly.

→

Evaluate each solution using a fitness function.

→

Select the best individuals.

→

Recombine (crossover) and **mutate** them to create new individuals.

→

Repeat until a stopping condition is met.

Applications of EA

SCHEDULING AIRLINE CREWS

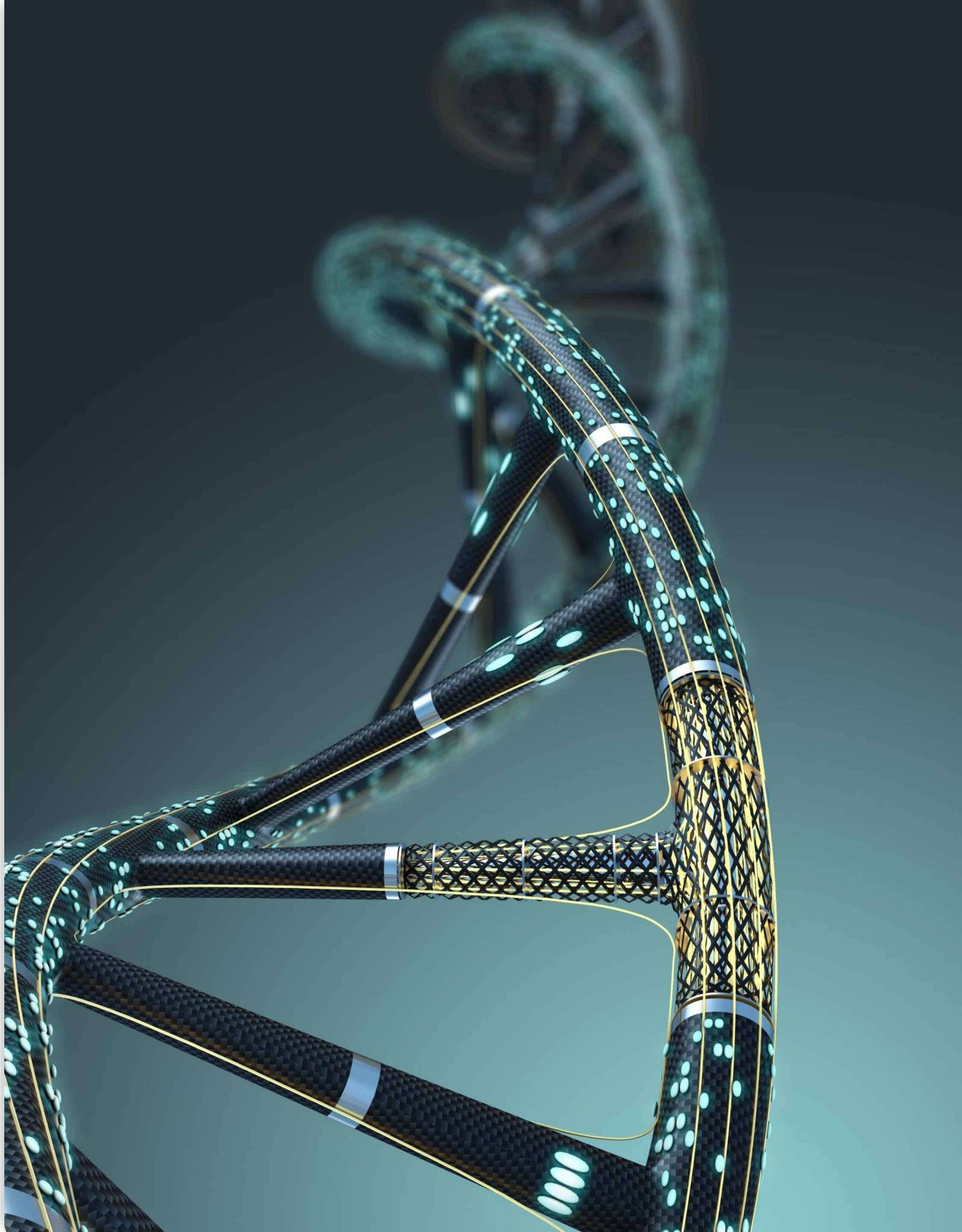
TUNING HYPERPARAMETERS IN ML

DESIGNING ANTENNAS (NASA EXAMPLE!)

PORTFOLIO OPTIMIZATION

Genetic Algorithms

- A Genetic Algorithm (GA) is a search and optimization method inspired by how living things evolve over time through natural selection.
- We represent potential solutions as chromosomes.
- Better solutions (those with higher fitness) are selected
- New solutions are created through crossover and mutation
- Over generations, solutions get better!



Swarm Intelligence Algorithms

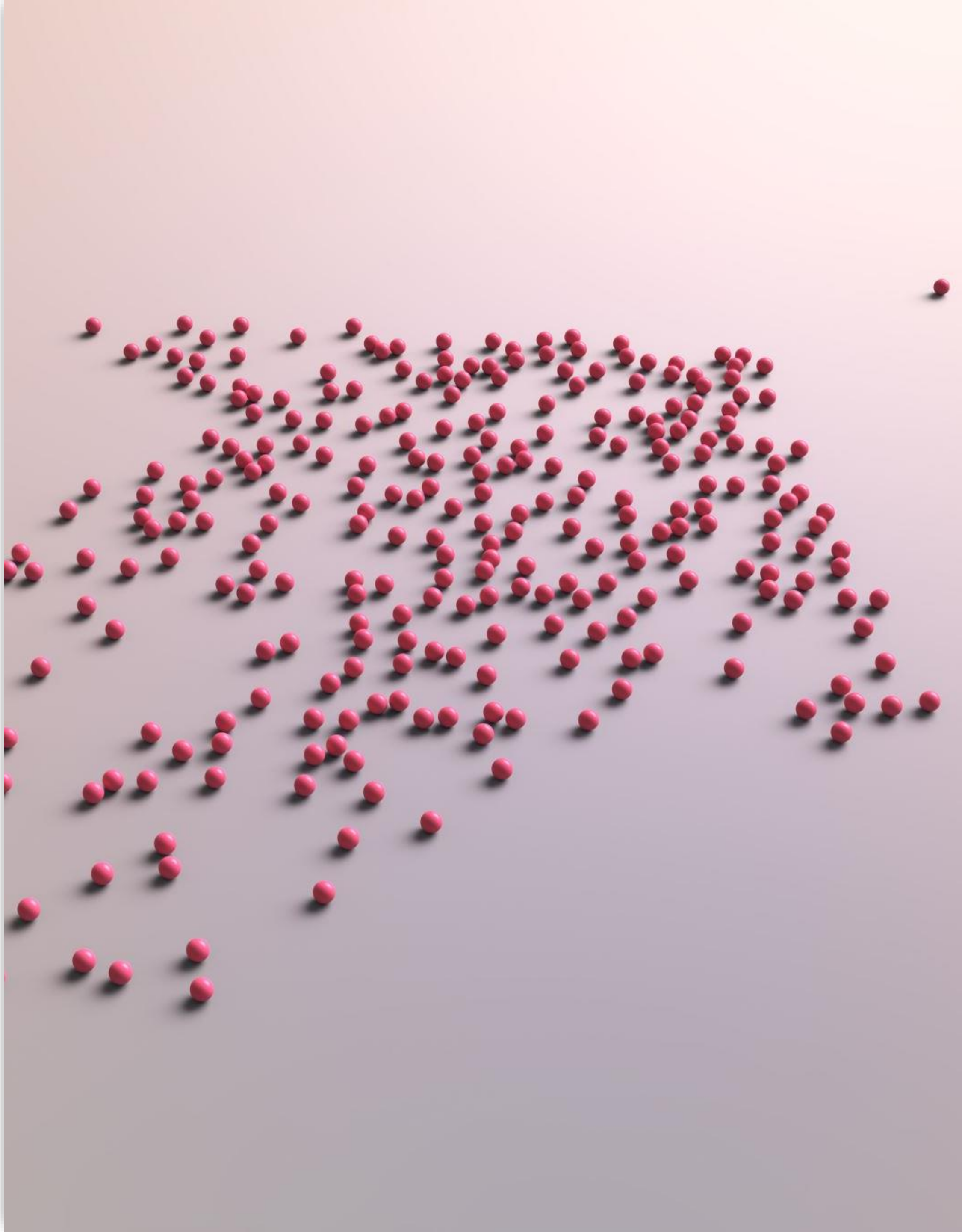
Inspired by: Collective behavior of decentralized, self-organized systems (e.g., birds, ants, fish)

Ant Colony Optimization (ACO)

- Inspired by ants finding shortest paths using **pheromone trails**.
- Good for **discrete path-based problems** like the **Traveling Salesman Problem (TSP)**.

Firefly Algorithm (FA)

- Fireflies are attracted to brighter (better) solutions.
- Uses light intensity and distance for movement.



Types of Machine Learning Algorithms

SUPERVISED LEARNING

UNSUPERVISED LEARNING

SEMI-SUPERVISED LEARNING

REINFORCEMENT LEARNING

Supervised Learning Algorithms

- Learned under supervision.
  - Supervision of what?
  - Humans?
- Supervised by the Labeled data
  - Require labeled data. (Inputs, output)
  - This is the most difficult part of supervised learning.

Types of Supervised Learning Models

- Regression
  - Predicting a Linear value
  - Linear Regression
  - SVR
  - DT

- Classification
- Predicting a class/label
  - Logistic Regression
  - SVM
  - DT
  - NB

Artificial Neural Network Models such as MLP, CNN, RNNs are Considered as supervised Learning Models

Supervised learning examples

A Bank may have borrower details (age, income, gender, etc.) of the past **(features)**

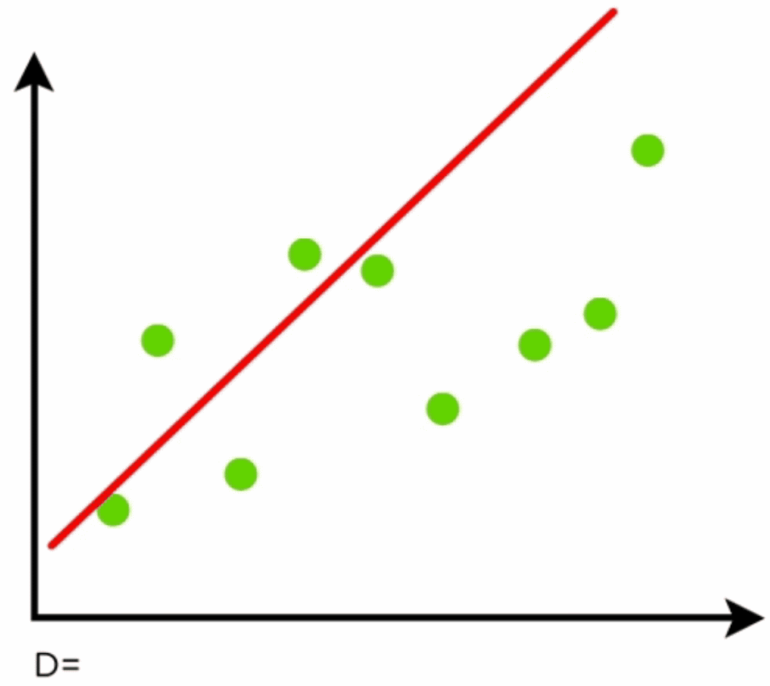
Also it may have details of the borrowers who defaulted in the past **(labels)**

Based on the above, can train a classifier to learn the patterns of borrowers who are likely to default on their payments

Linear Regression

- **Model Explanation:**
  Draws a straight line that best fits the data.
- **Key Concept:**
  Learns the relationship as a **linear equation**:
  $y = mx + c$.
- **What is learnt through training:**
  Finds the best slope (m) and intercept (c) to minimize error.
- **Example Use Cases:**
  - Predicting house prices
  - Salary estimation
  - Sales forecasting
- **Limitations:**
  - Only works well when the relationship is **linear**
  - Not suitable for complex, non-linear patterns
  - Sensitive to **outliers**

Linear regression



"Predictor":

Evaluate line:

$$r = \theta_0 + \theta_1 x_1$$

return r

Types of Unsupervised Learning Algorithms

- Clustering Algorithms
  - K Means
  - DBSCAN
- Dimensionality Reduction Algorithms
  - PCA
  - MDS (Multidimensional Scaling)
  - LDA (Linear Discriminant Analysis)

- Graph Based Models can be considered as Unsupervised Learning

Unsupervised learning examples

A Supermarket may store each buyer's

basket content details **(features)**

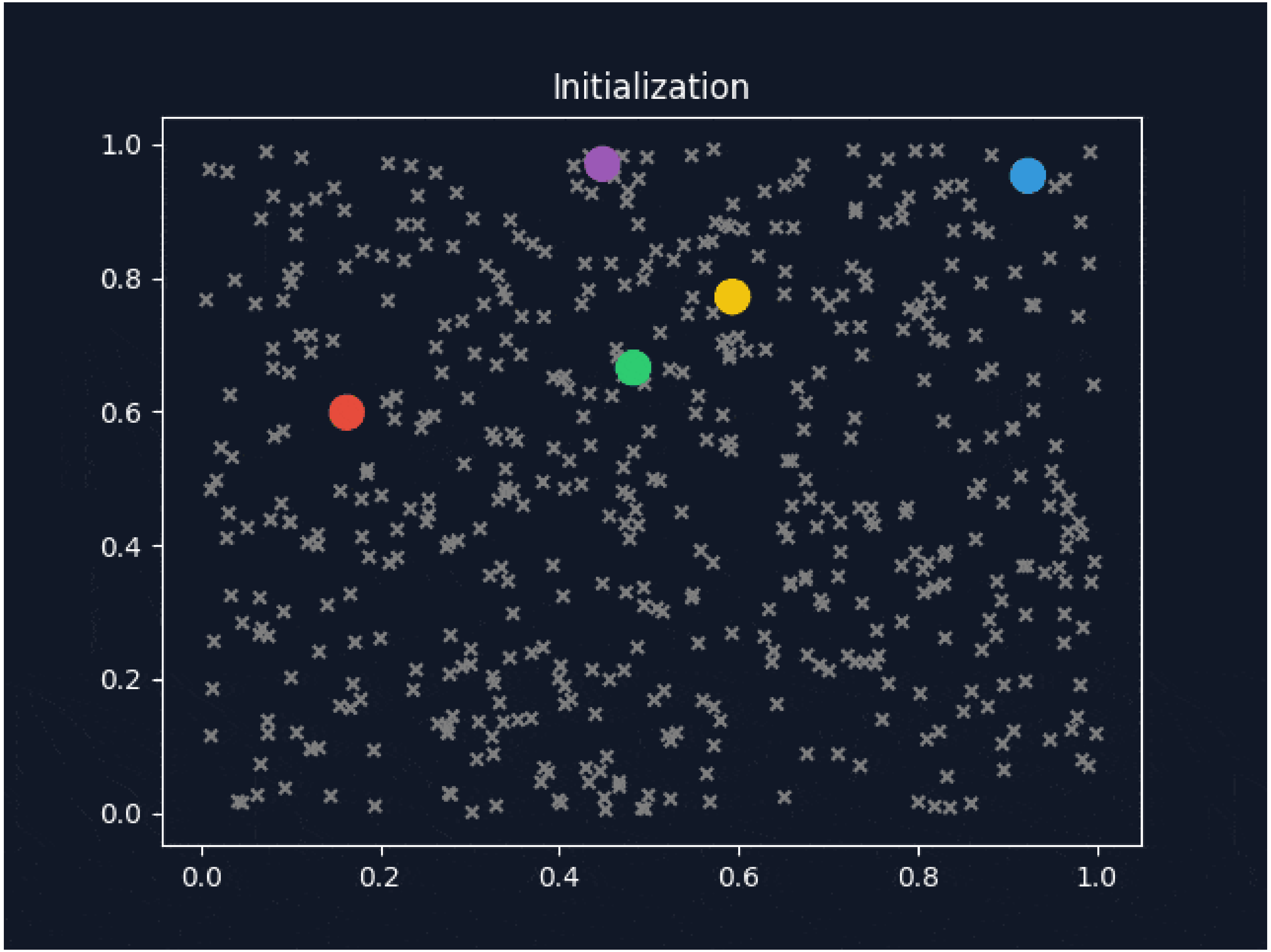
There are **NO** grouping **(labels)**

Need to group the buyers based on their buying patterns in order to best use the shelf space (recommendation)

K- Means Clustering

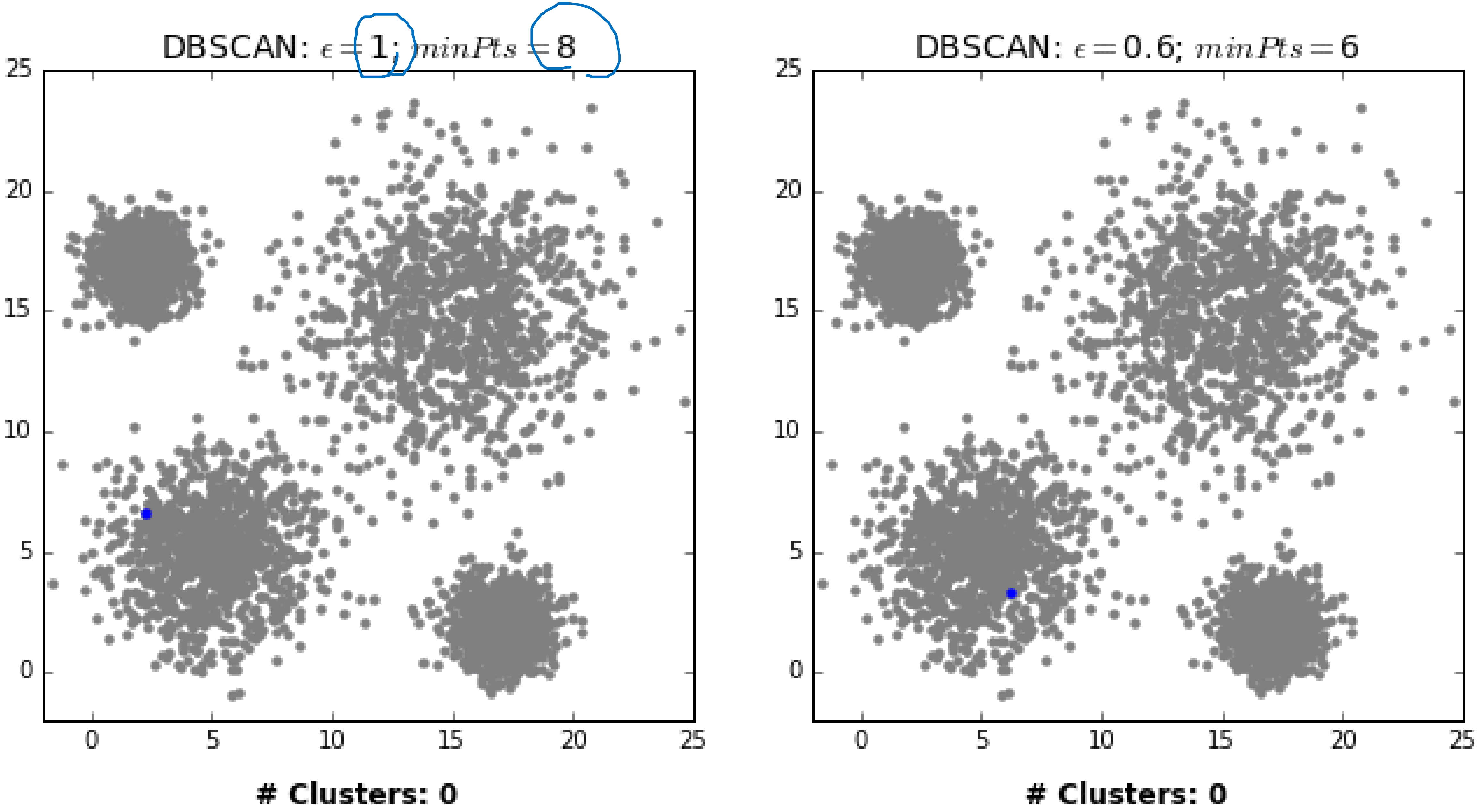
- **Model Explanation:**
  - Groups data into **K clusters** based on similarity.
- **Key Concept:**
  - Finds **cluster centers** and assigns each point to the nearest one.
- **What is learnt through training:**
  - Learns the **position of cluster centers**.
- **Example Use Cases:**
  - Customer segmentation
  - Grouping articles by topic
  - Organizing images
- **Limitations:**
  - You must choose **K (number of clusters)** beforehand
  - Assumes **spherical-shaped clusters**
  - Struggles with **uneven or noisy data**

K-means clustering example

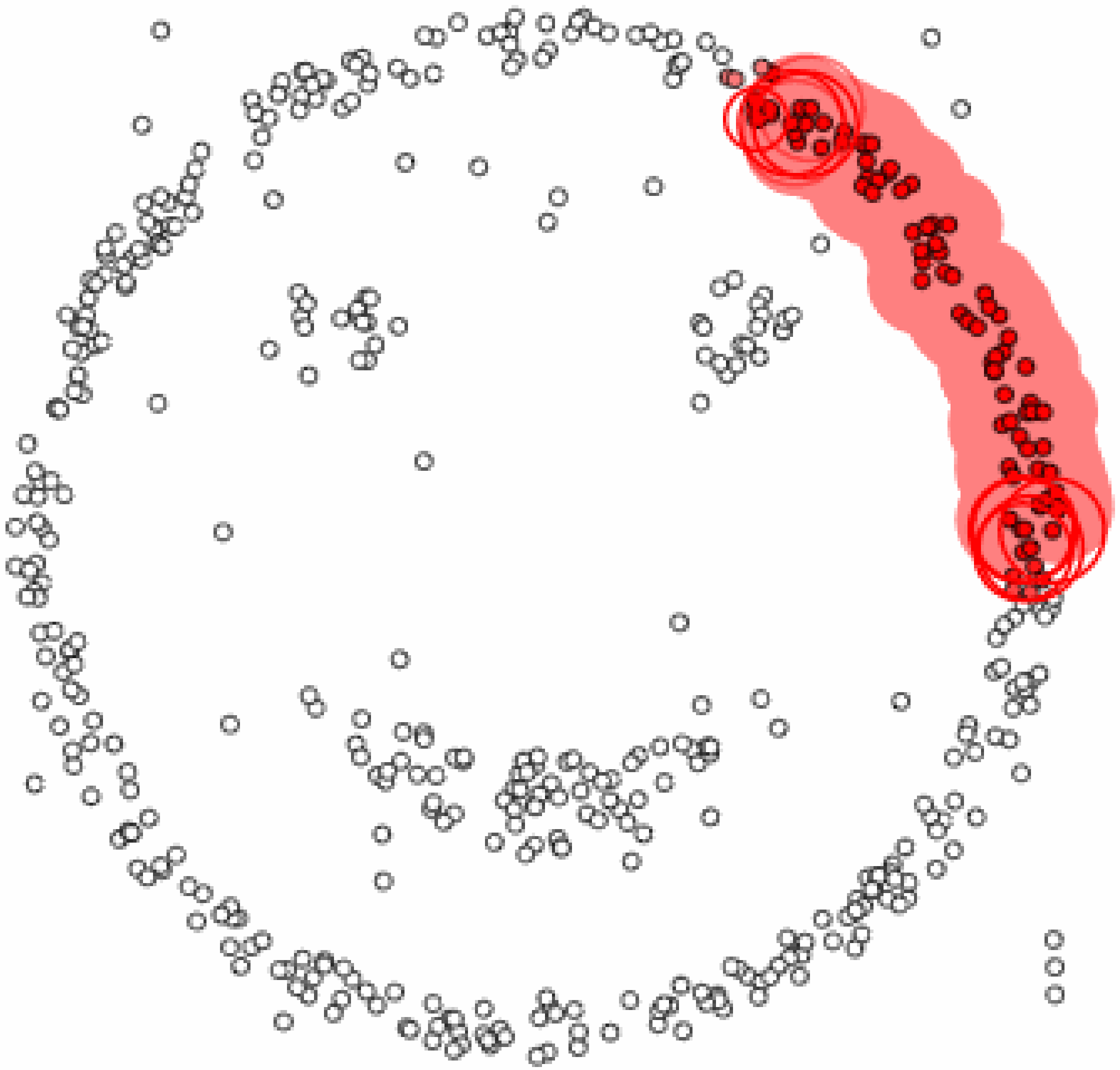


DBSCAN

- **Model Explanation:**
  - Groups together dense areas; labels sparse points as outliers.
- **Key Concept:**
  - Clusters are formed based on **density**, not shape or number.
- **What is learnt through training:**
  - Learns which points belong to **dense clusters** and which are **noise**.
- **Example Use Cases:**
  - Fraud detection
  - Identifying event hotspots
  - Anomaly detection in GPS or sensor data
- **Limitations:**
  - Struggles with **varying densities**
  - Requires tuning of **eps** and **min points**
  - Not ideal for **high-dimensional data**



DBSCAN: $\epsilon = 1; minPts = 8$
Clusters: 0
DBSCAN: $\epsilon = 0.6; minPts = 6$
Clusters: 0



epsilon = 1.00
minPoints = 4

Restart

Pause

Semi-supervised learning

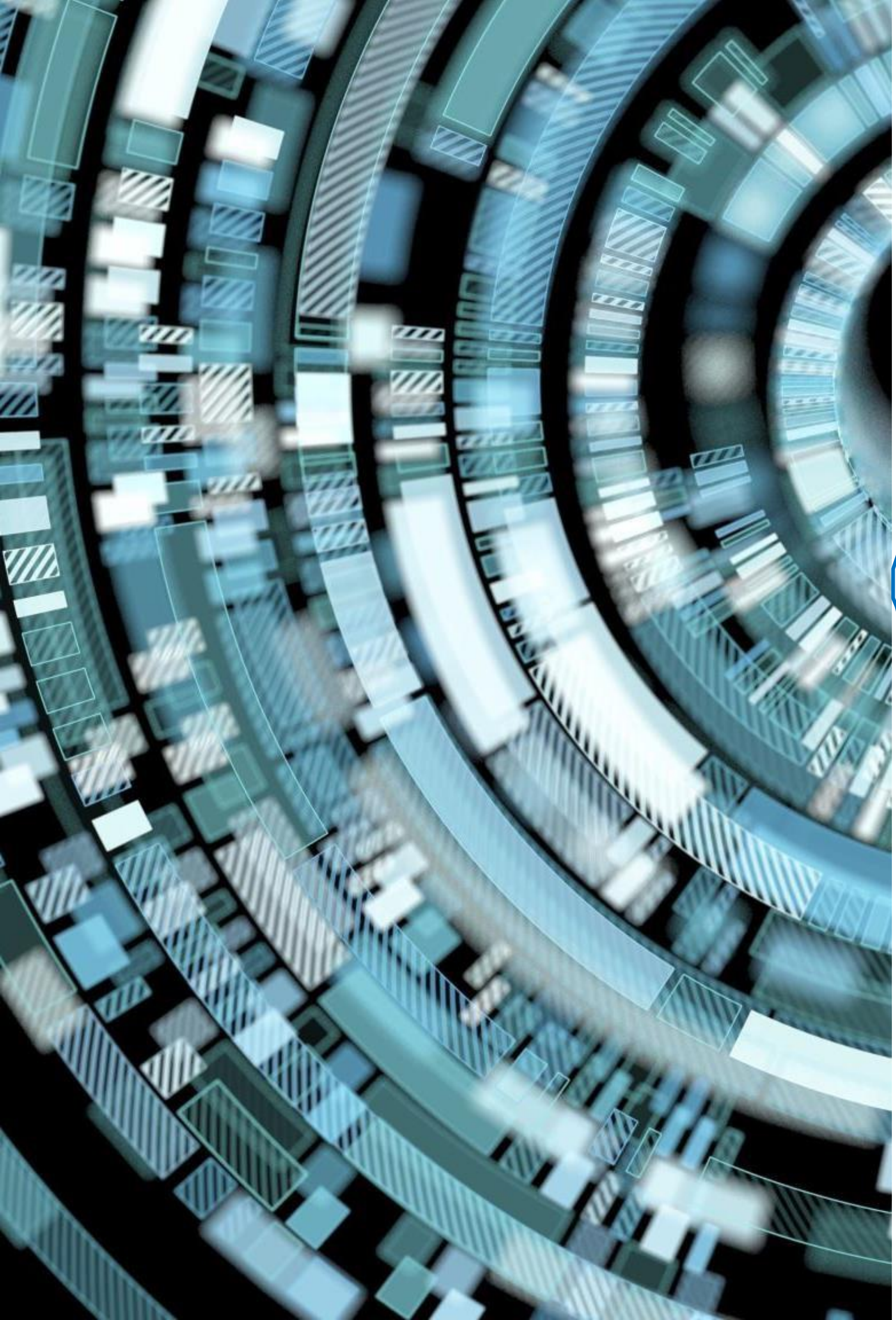
Labeled data is expensive/difficult to get

Unlabeled data is cheap/easier to get

The idea is to use smaller amount of labelled data with larger amount of unlabeled data to creating the training/testing datasets

Algorithms - Self Training, Generative models

- Semi-Supervised Support Vector Machines, etc.

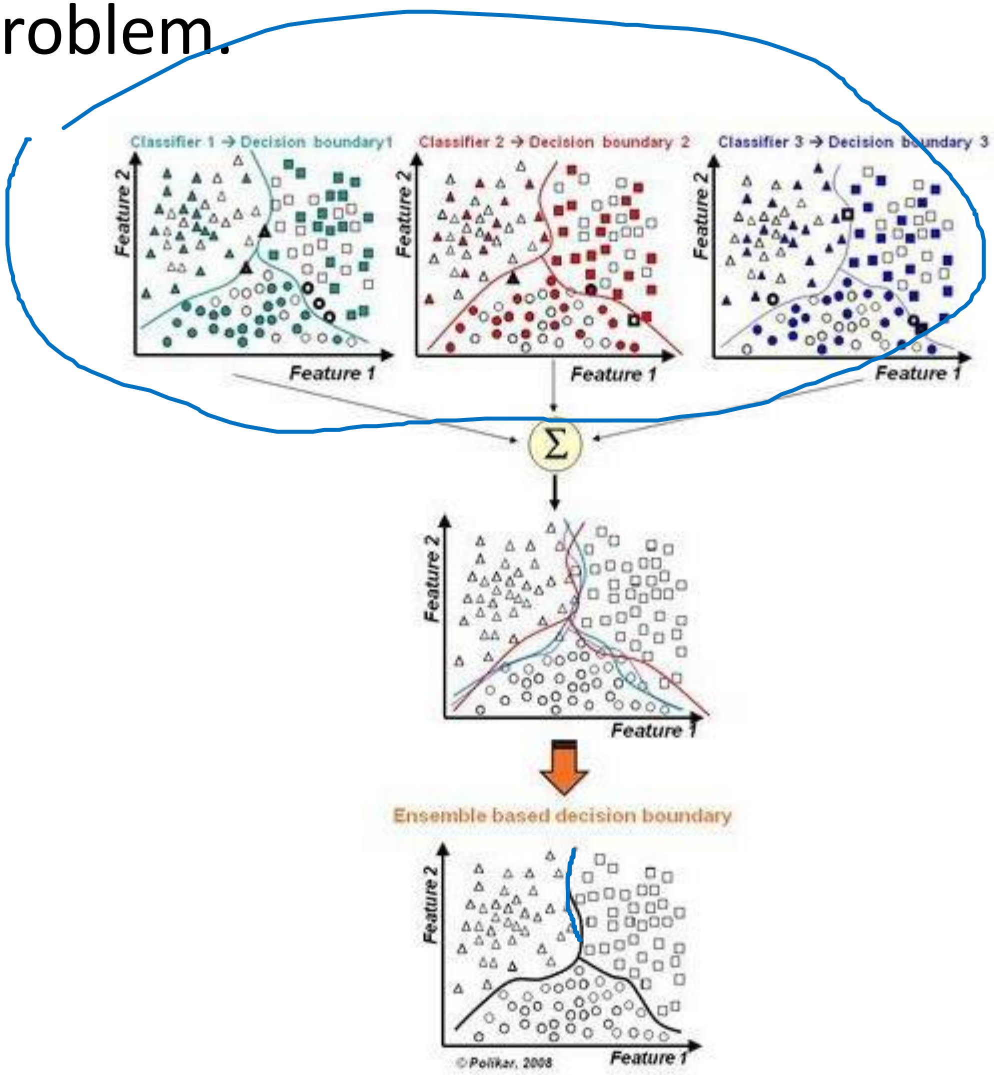


Semi-Supervised Learning Algorithms

- Generative Adversarial Networks
- Auto-encoders
- Variational Auto-encoders

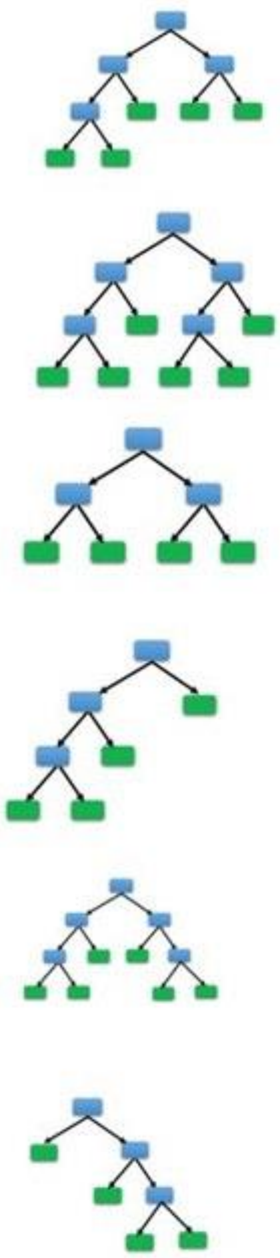
Ensemble Learning

- Often, multiple classifiers need to be combined to solve a real-world problem.



Random Forest

- **Model Explanation:**
  - Combines predictions from many decision trees.
- **Key Concept:**
  - Uses **voting** or **averaging** to make decisions.
- **What is learnt through training:**
  - Learns different rules from subsets of data to make **robust predictions**.
- **Example Use Cases:**
  - Spam detection
  - Credit risk analysis
  - Disease classification
- **Limitations:**
  - Can be **slow** with large datasets
  - Less interpretable than a single decision tree
  - Needs tuning (like number of trees)



Random Forest in Action!!!

Reinforcement Learning

- **Model Explanation:**
  - Learns by interacting with an environment and receiving feedback (rewards or penalties).
- **Key Concept:**
  - Uses **trial-and-error** and **reward maximization** to improve decision-making over time.
- **What is learnt through training:**
  - Learns an **optimal policy** or **action strategy** that maximizes long-term rewards.
- **Example Use Cases:**
  - Game playing (e.g., AlphaGo)
  - Robotics (e.g., navigation or motor control)
  - Dynamic pricing or recommendation systems
- **Limitations:**
  - Requires **many interactions** with the environment (sample inefficiency)
  - Can be **unstable** or hard to converge
  - Needs **careful reward design** to avoid unintended behaviors

Reinforcement learning examples

A group of robots have been deployed in an unknown territory

The objective is for them to collaboratively find the navigation path to reach a particular destination/goal

Can use reinforcement learning where achieving the goal/getting closer to the goal gives a positive reward. Negative reward otherwise

Can share the information among robots (multi-agent system)

Comparing Machine Learning Models

Not all models are created equal — and neither are the ways we evaluate them.

Key Questions to Ask:

Is it a **classification**, **regression**, or **clustering** task?	Do we care more about **correctness**, **fairness**, or **interpretability**?	What are the **costs of wrong predictions**?

Common Evaluation Matrices

Task Type	Metrics Used
Classification	Accuracy, Precision, Recall, F1 Score, ROC-AUC
Regression	MSE, MAE, RMSE, R^2 Score
Clustering	Silhouette Score, Davies-Bouldin Index, Inertia
Ranking/Recommendation	MAP, NDCG, Hit Rate

Classification Matrices Explained

Metric	Use When...	Notes
Accuracy	Classes are balanced, all errors matter	Can be misleading with imbalance
Precision	False positives are costly (e.g., spam)	TP / (TP + FP)
Recall	False negatives are costly (e.g., cancer)	TP / (TP + FN)
F1 Score	Balance between precision and recall	Harmonic mean
ROC-AUC	Need to evaluate ranking ability	Works for probabilistic models

Regression Model

Metric	Use When...	Notes
MSE	Large errors are very bad	Penalizes large errors more
MAE	Equal penalty for all errors	More robust to outliers
RMSE	Like MSE but in original units	Square root of MSE
R^2 Score	Want to explain variability in output	1 = perfect, 0 = no

Clustering Algorithms

Metric	Use When...	Notes
Silhouette Score	Want to measure how distinct clusters are	1 = well-clustered, -1 = wrong
Davies-Bouldin Index	Lower is better (compact & separated)	Good for comparing k-values
Inertia (within-cluster SSE)	Used in K-Means	Lower is better, but not scaled

Practical Evaluation Factors

Factor	**Why It Matters**
Interpretability	Do we understand how/why it makes predictions?
Training Time	Important for real-time or big data
Fairness	Does it treat all groups equally?
Generalization	Does it perform well on new data?
Explainability	Can we explain decisions to stakeholders?

Things to consider in Selecting a ML Algorithm

- If there’s an algorithmic way instead of ML, use it!!! (ML is messy)
- Refer the literature!!!
- Try different ML algorithms (no single algorithm is the best)
- Check the dataset against the usage/strength of each algorithm (e.g. RNNs, ARIMA is good in time-series predictions)
- Be mindful of ‘external factors’ (e.g. seasonal effects, RL if you don’t have data, Clustering if you have unlabeled data, etc.)
- Test your algorithm(s) with test data and select the best performing one for production (include the test results in your thesis/publications)
- No algorithm will be perfect! (There will be an error. The objective is to keep the error at an acceptable rate)

Popular Frameworks/Tools

- Scikit-learn - Python (Anaconda Python Distribution)
- R (R studio)
- Matlab/Octave (can export DLLs)
- Weka (Java based)
- Java OpenNLP/Python NLTK (Natural language processing + ML)
- Apache Spark (part of the Apache Hadoop platform)
- Google Tensorflow (Python library for Deep neural networks)
- Apache Keras (Python library of neural networks)
- Theano (Python library for Multicore processing of DNNs)
- Amazon AWS Services/Microsoft Azure ML (Cloud based ML)

Commonly used python libraries

✓	NumPy	Matrix algebra
	Pandas	Data Frames, Series
	Matplotlib	Visualization

Summary

- AI is a vast discipline with many varying branches.
- AI attempts to give machine the ability to mimic human decision making/learning capabilities

Thank You

Jeewaka Perera

Jeewaka.p@sliit.lk