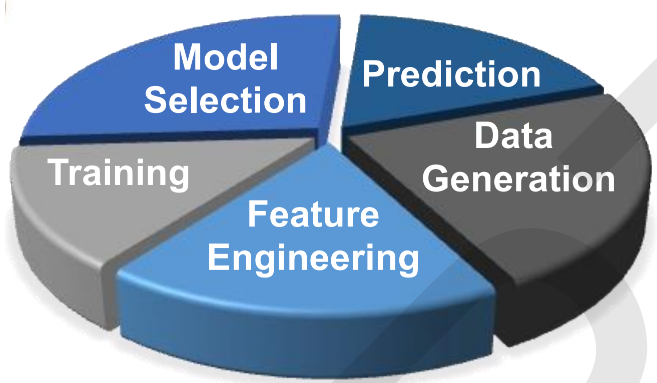
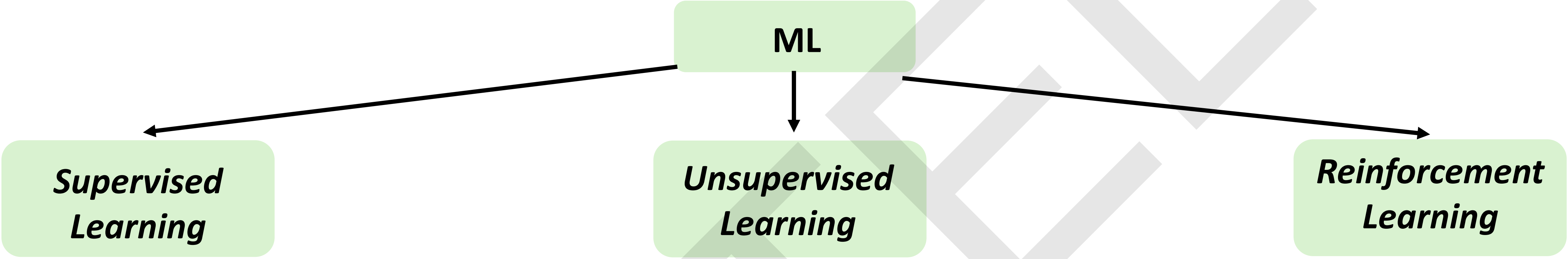
Machine Learning for Core Engineering Disciplines



Prof. Ananth Govind Rajan
Department of Chemical Engineering
Indian Institute of Science, Bengaluru
Website: https://agrgroup.org
Email: ananthgr@iisc.ac.in

NPTEL

Broad categorization of ML algorithms



ML

Supervised Learning

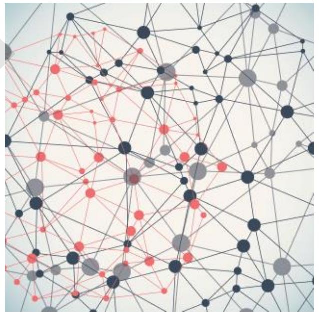
Training happens based on labels assigned to a target variable

Applies to the prediction of both continuous and discrete variables

Unsupervised Learning

Target variables aren’t specified, and labels aren’t required

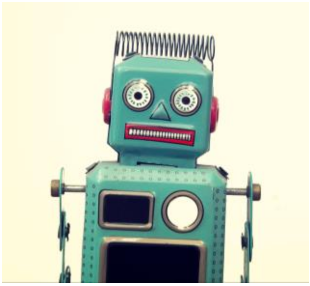
Model “learns” the underlying structure of the data based on some similarity metric



Reinforcement Learning

A virtual agent receives rewards based on favourable outcomes

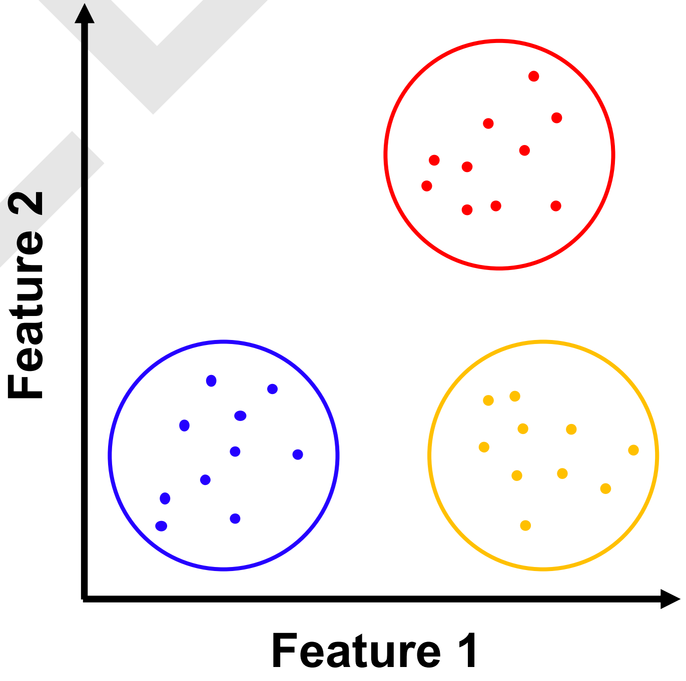
Maximization of the cumulative reward leads to an optimal policy for the problem



NPTEL

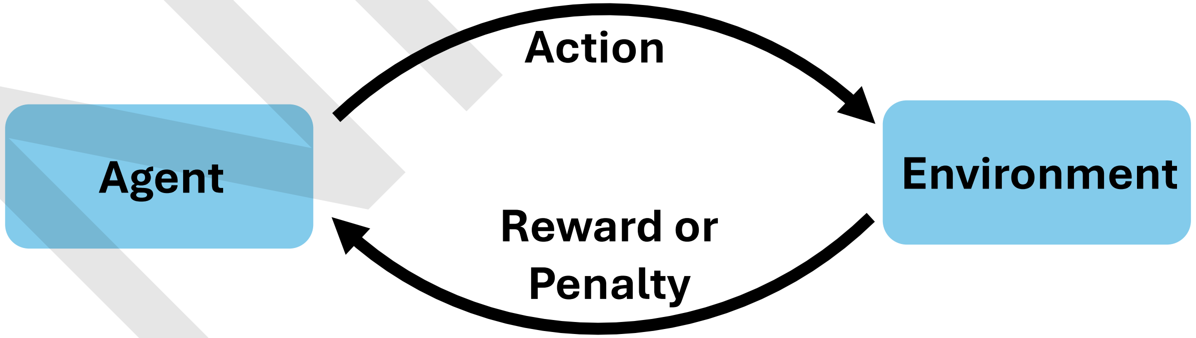
Broad categorization of ML algorithms

- **Supervised learning**
  - Seeks to map each input feature vector to an output value
  - Training based on specified values (labels) of the target variable



- **Unsupervised learning**
  - Seeks to learn underlying structure and inherent patterns of the data in the feature space
  - Dimensionality reduction falls into this class of ML

- **Reinforcement learning (RL)**



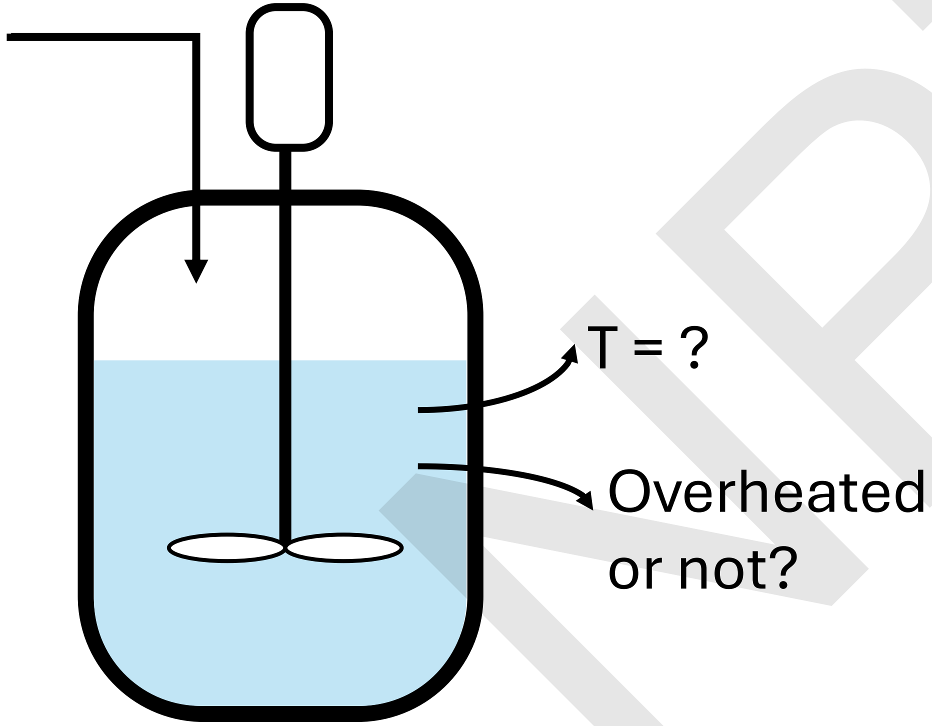
NPTEL

Types of supervised learning

Regression

Used to predict a continuous variable

Example: Given plant data, what is the temperature of the reactor?



Classification

Used to classify data into a fixed number of categories

Can be "binary" or "multiclass"

Example: Given plant data, has the reactor overheated or not?

Given a microscopy image:

(i) Is the material brittle?

(ii) What is the hardness of the material?

NPTEL

Broad types of unsupervised learning

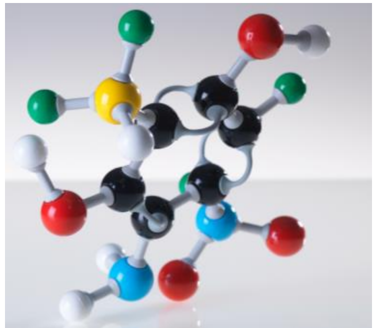
- **Clustering**
  - Groups datapoints to uncover the similarities and differences between them
  - **Examples:** k-means clustering, density-based clustering

- **Dimensionality reduction**
  - Reduces the number of features in the dataset without losing important relationships
  - **Examples:** principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), uniform manifold approximation and projection (UMAP)

- **Generative algorithms**
  - Learn the distribution underlying the data and thus generate new samples which are similar to the input data
  - **Examples:** generative adversarial networks (GANs), variational autoencoders (VAEs), restricted Boltzmann machines (RBMs)

NPTEL

Examples of supervised learning



→ Is the molecule soluble in water?

→ What is the voltage provided by a battery?

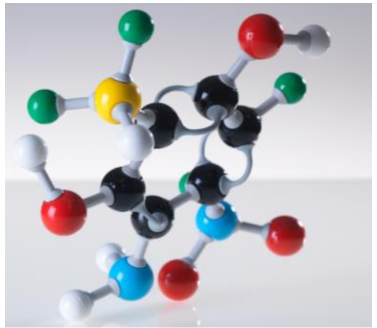
→ Is the person suffering from chest congestion or not?

In each case, you would have to train the computer with some labeled data

Various algorithms: linear/nonlinear regression, decision trees, random forests, support vector machines, neural networks & deep learning, ...

NPTEL

Examples of supervised learning



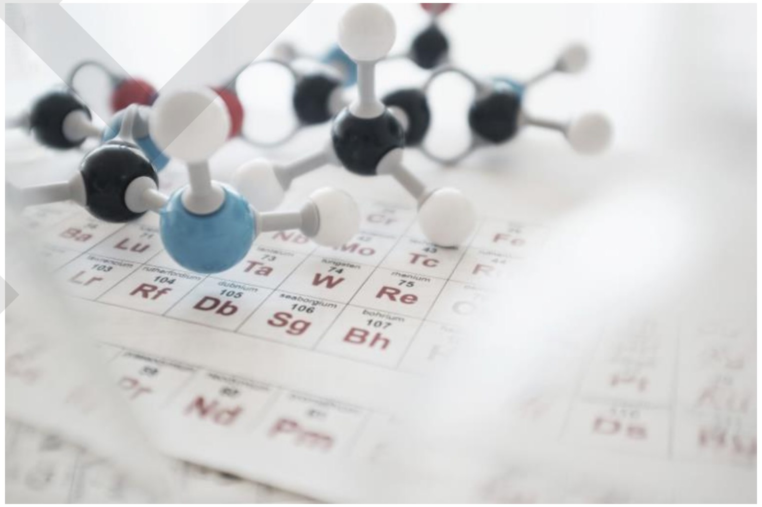
Example	Type of Problem	Input Features	Target Label
Molecule soluble in water?	Binary Classification	Molecular descriptors/ fingerprints	Soluble / not soluble
Voltage provided by battery?	Regression	Battery type, temperature, chemistry, age, load conditions	Voltage (continuous value)
Person suffering from chest congestion or not?	Binary Classification	Chest X-ray and/or symptoms	Yes / no

NPTEL

Examples of unsupervised learning

Categorizing elements to find similarities between them based on their features

Generating molecules similar to those in the training dataset



In each case, you would like the computer to figure out the structure of the data itself based on the underlying features

Various algorithms: k-means clustering, PCA, GANs, VAEs, etc.

NPTEL

Supervised ML as a function approximator

$$y = \boldsymbol{f}(x_1, x_2, \dots, x_n; \beta_1, \beta_2, \dots, \beta_p; \alpha_1, \alpha_2, \dots, \alpha_h)$$

Target variable **Features** **Parameters** **Hyperparameters**

Target variable: variable of interest that one desires to predict using a supervised ML model; it could be a continuous or a discrete variable

Features: independent variables that characterize each datapoint in the dataset

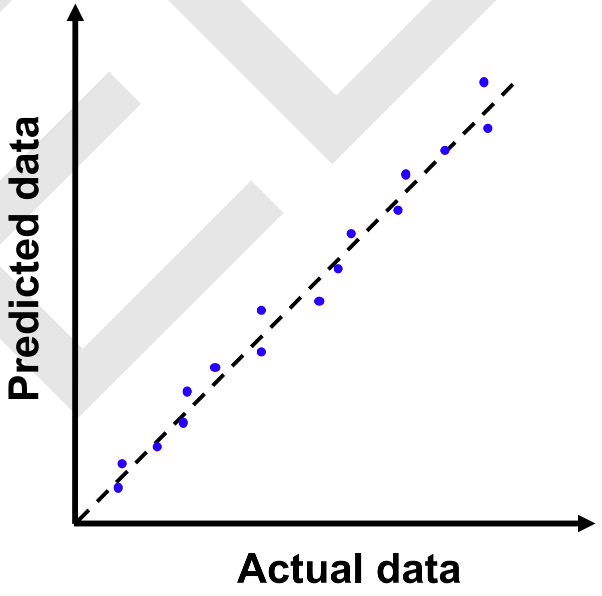
Parameters: learnable coefficients (weights) that the model learns during model training using multivariate optimization

Hyperparameters: model configuration variables specified before training which determine the model architecture and training process

NPTEL

Training and test sets in ML models

Features	Data
$(x_{11}, x_{12}, \ldots, x_{1n})$	y_1
$(x_{21}, x_{22}, \ldots, x_{2n})$	y_2
$\vdots$	$\vdots$
$(x_{d1}, x_{d2}, \ldots, x_{dn})$	y_d

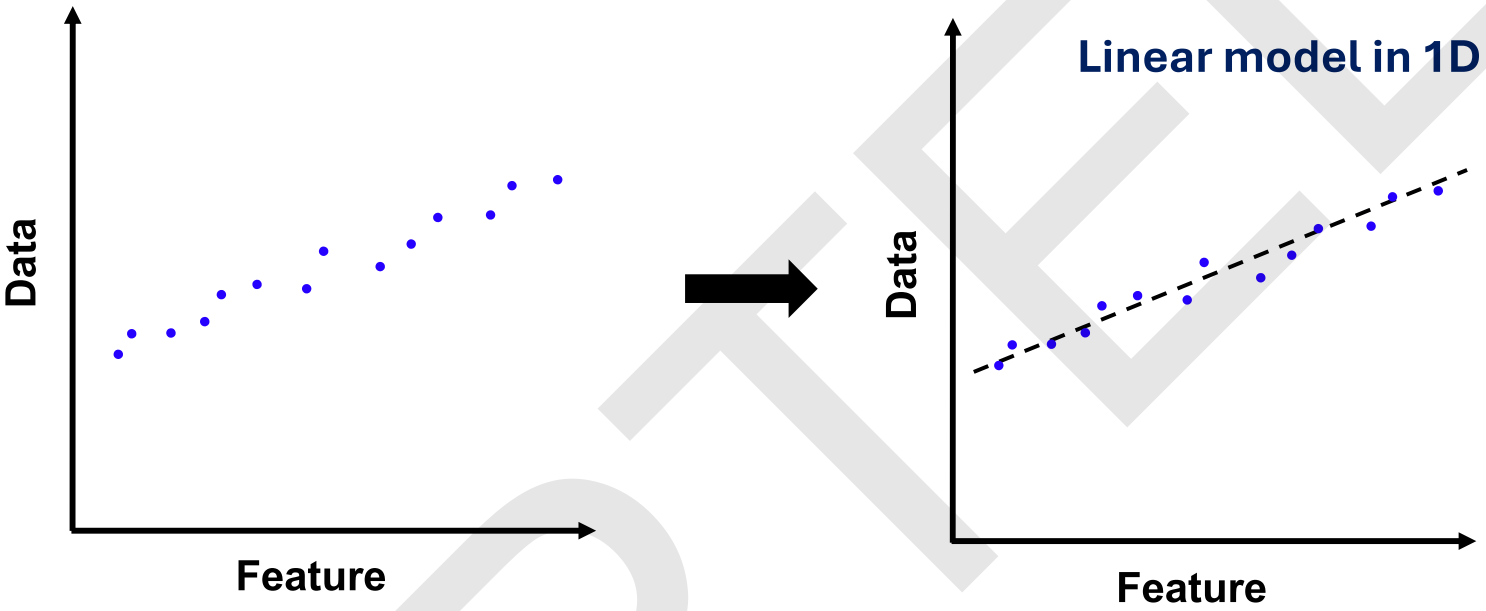


Training set: collection of data points, i.e., the set of feature and corresponding target variable values used to train the ML model

Test set: collection of unseen datapoints used to independently verify the performance of the final model after training is completed

NPTEL

How do we enable ML models to learn?



Loss function: metric used to quantify the performance of an ML model in terms of the errors in the predicted values of the target variable versus its actual values

Cross validation: a method to determine the generalizability of the model across various realizations of training data and unseen (validation) data; allows one to rationally choose hyperparameters

NPTEL