Introduction to Machine Learning in Biological Systems (ES60011)

Introduction to Machine Learning

An Overview

- What is Machine Learning?
- Need of Machine Learning
- Types of Machine Learning
- Brief Introduction on Machine Learning Algorithms
- Basic Idea on Machine Learning Steps
- Applications
- Brief Introduction on Biological Systems
- Machine Learning in Biological Systems

Need of Machine Learning

Machine learning is applied in various scenarios, such as:

When human knowledge is unavailable

e.g.: exploring distant planets like Mars

When humans are unable to articulate their skills

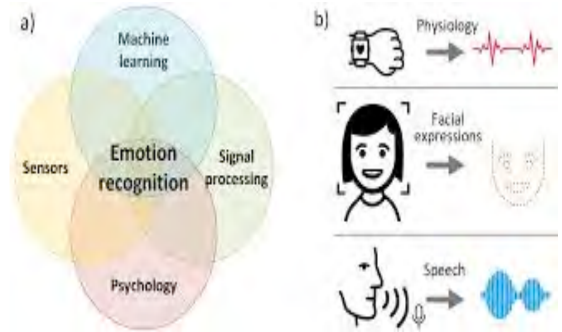
e.g.: recognizing speech patterns

When solutions need to be tailored

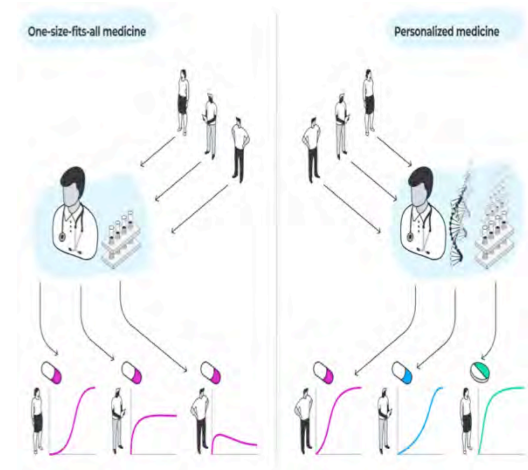
e.g.: designing individualized treatment plans in healthcare

When decisions rely on vast datasets

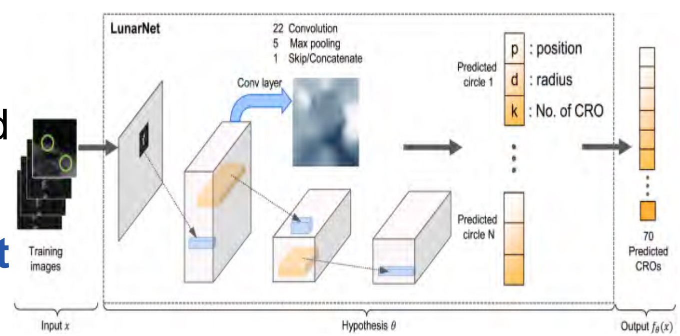
e.g.: analyzing genomic information



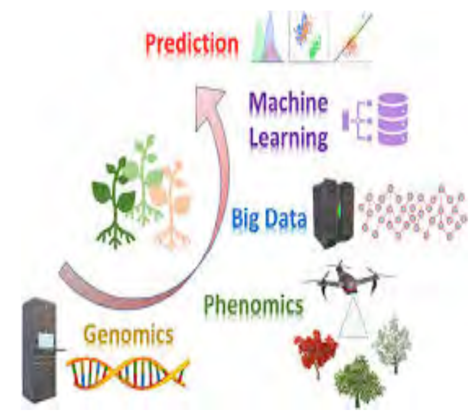
Emotion Pattern Evaluation



Personalised Healthcare



Space Navigation

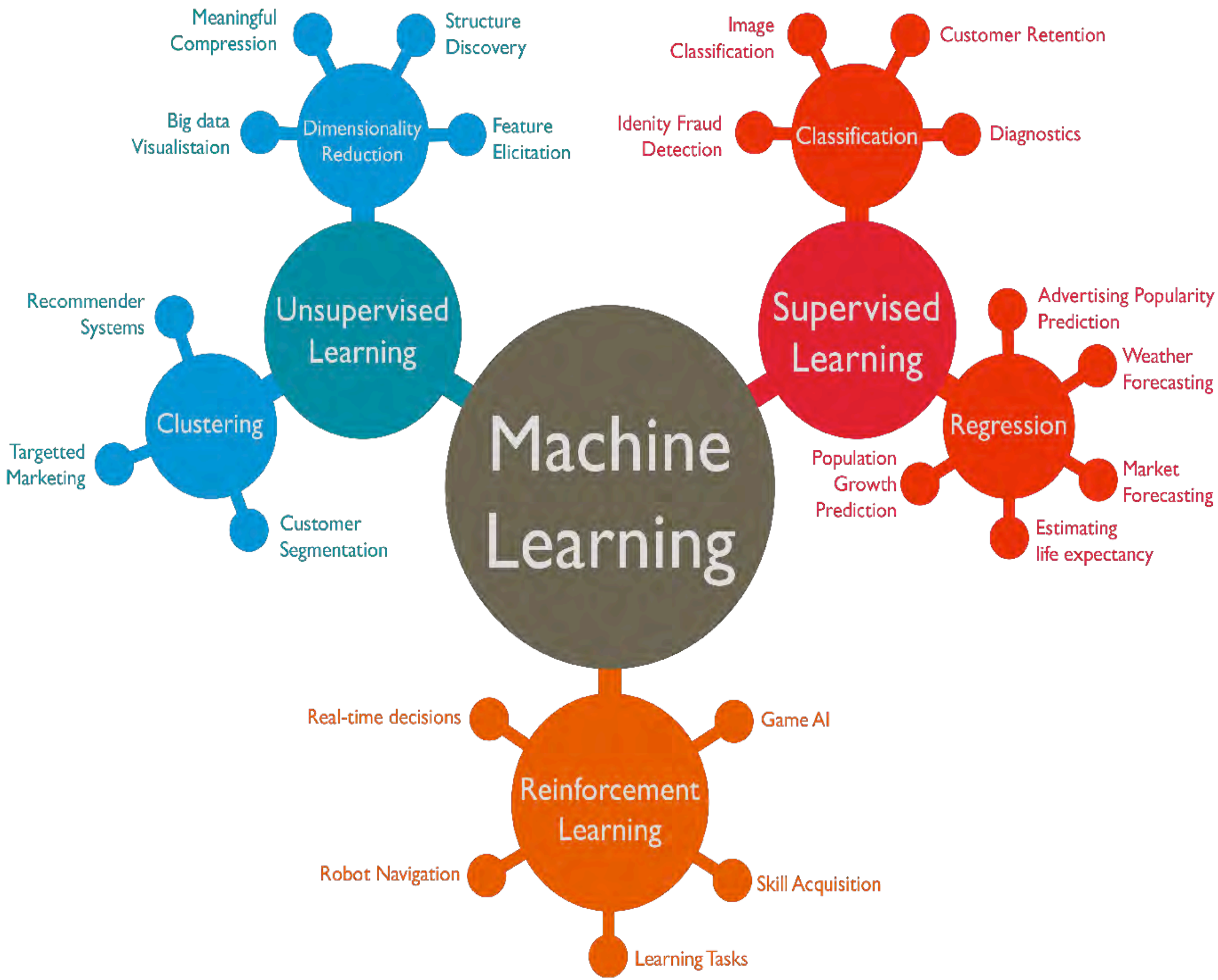


Big Data Handling

What are ML techniques

Machine leaning (ML) techniques enables systems to learn from experience (read data). ML refers to systems ability to acquire and integrate knowledge through large-scale observations and to improve and extend by itself learning new knowledge rather than by being programed with that knowledge (Shapiro 1992)

Types of Machine Learning



Types of Machine Learning

Machine Learning

Supervised ML

Polynomial regression
Random forest (RF)
Linear regression
Logistic regression
Decision trees
K-nearest neighbours
Naive Bayes

Unsupervised ML

Partial least squares
Fuzzy means
Singular value decomposition
K-means clustering
Apriori
Hierarchical clustering
Principal component analysis

Reinforcement Learning

Q-Learning
State-Action-Reward-State-Action (SARSA)
Deep Q Network (DQN)
Deep Deterministic Policy Gradient (DDPG)

Machine Learning: Introduction

Artificial intelligence (AI):

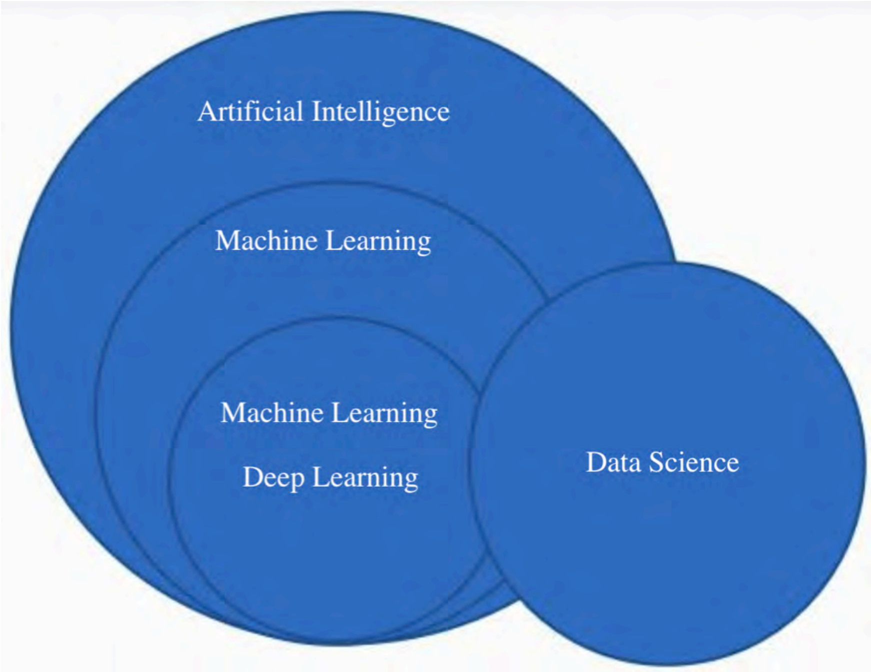
It is computer software that mimics human cognitive abilities in order to perform complex tasks that historically could only be done by humans, such as decision making, data analysis, and language translation.

Machine learning (ML):

Machine learning is a subset of AI in which algorithms are trained on data sets to become machine learning models capable of performing specific tasks.

Deep learning:

Deep learning is a subset of ML, in which artificial neural networks (ANNs) that mimic the human brain are used to perform more complex reasoning tasks without human intervention.



Machine Learning Algorithms

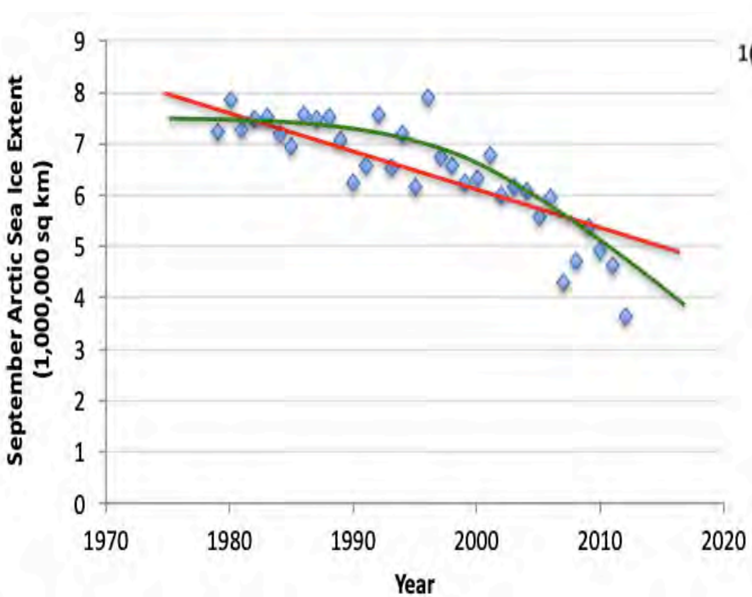
Supervised (inductive) learning

— Given: training data + desired outputs (labels)

Supervised Learning

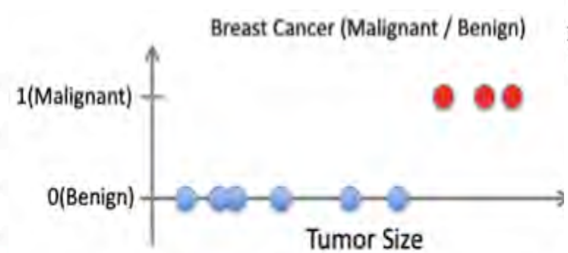
Regression

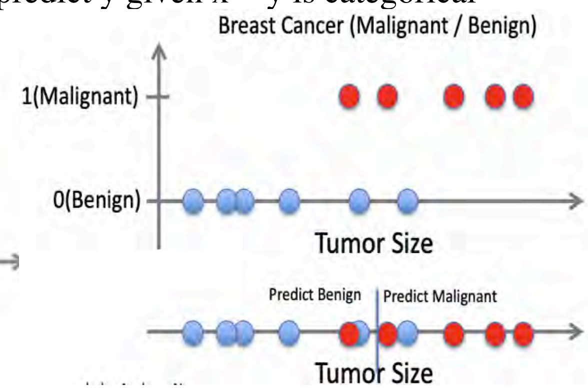
- Given $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$
- Learn a function f(x) to predict y given x – y is real-valued == regression



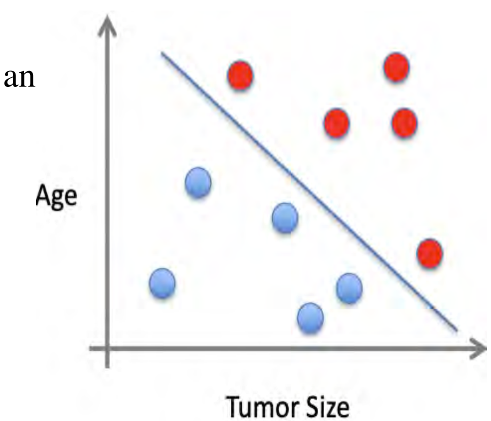
Classification

- Given $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$
- Learn a function f(x) to predict y given x – y is categorical == classification





- x can be multi-dimensional
  – Each dimension corresponds to an



Evaluation metrics - For Classification Problem

- **Analysis:**
- Precision
- Recall
- Accuracy
- F1-score
- Etc…

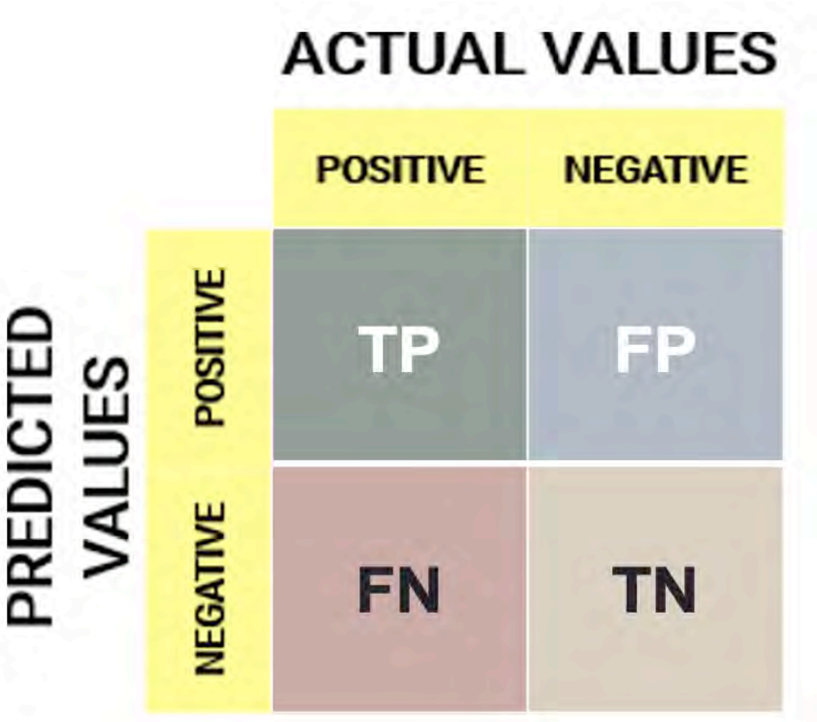
True Positive (TP)
The predicted value matches the actual value. The actual value was positive and the model predicted a positive value.

True Negative (TN)
The predicted value matches the actual value. The actual value was negative and the model predicted a negative value.

False Positive (FP) – Type 1 error
The predicted value was falsely predicted. The actual value was negative but the model predicted a positive value.

False Negative (FN) – Type 2 error
The predicted value was falsely predicted. The actual value was positive but the model predicted a negative value.

Confusion matrix



Evaluation metrics

Accuracy= **(TP+TN)/(TP+TN+FP+FN)**

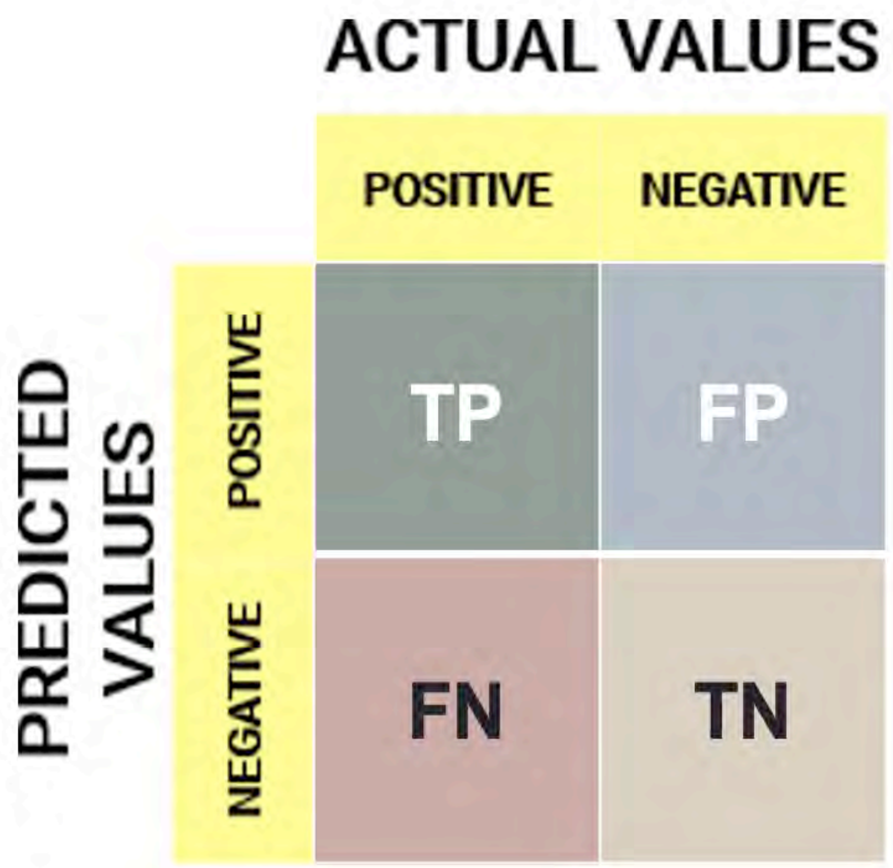
Precision P = **TP/(TP+FP)**

Recall/Sensitivity R= **TP/(TP+FN)**

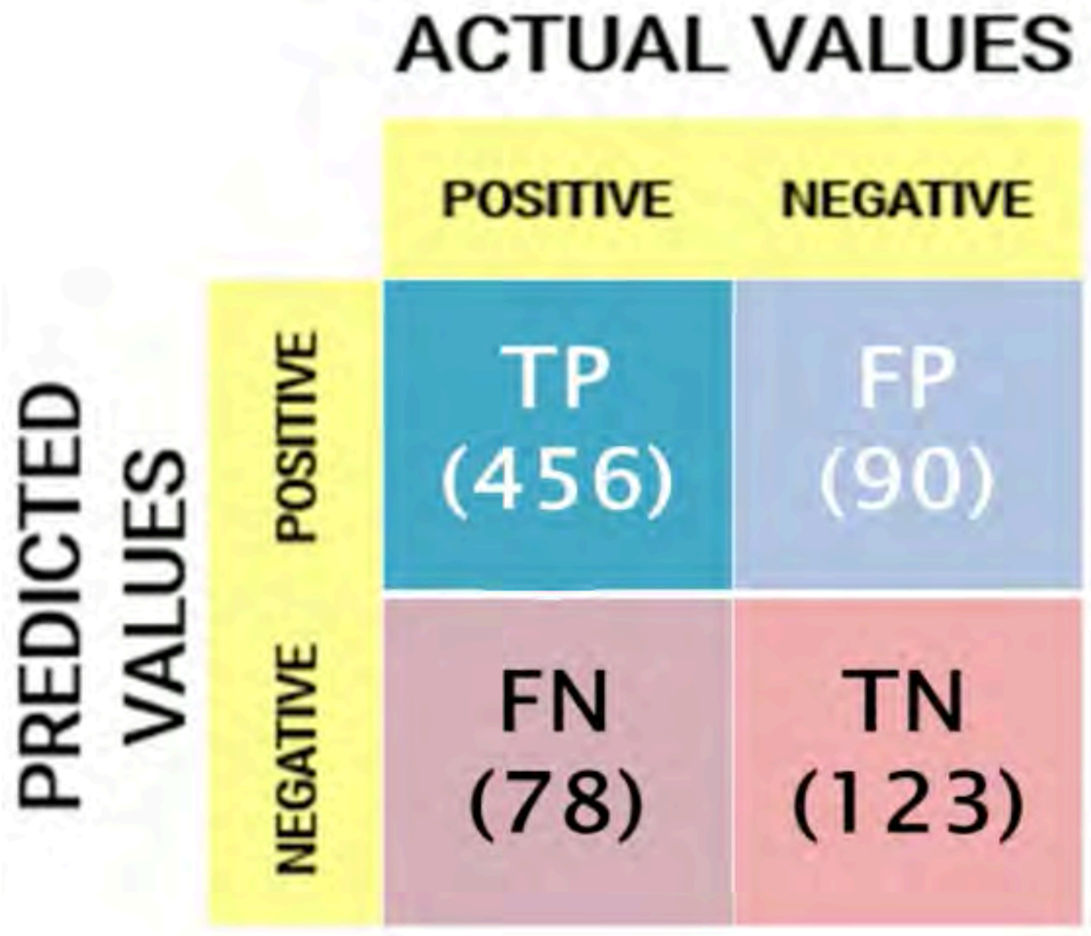
Specificity S = **TN/(TN+FP)**

F-measure **F=2*(P*R)/(P+R)**
[harmonic mean of precision and recall]

Confusion matrix



Evaluation metrics



Accuracy= **(TP+TN)/(TP+TN+FP+FN)**

= 77.51 %

Numericals

1. Suppose a regression model follows y=mx+c, if the dataset (X,y) with X=[1,2,3,4,5] and y = [2,4,6,8,10], find the parameters m (slope) and c (intercept) for this model.

2. If a binary classification model has 90 true positives, 30 false positives, 20 false negatives, and 160 true negatives, calculate the accuracy, precision, recall, and F1-score.

Machine Learning Algorithms

Unsupervised learning

– Given: training data (without desired outputs)

Unsupervised Learning

Clustering

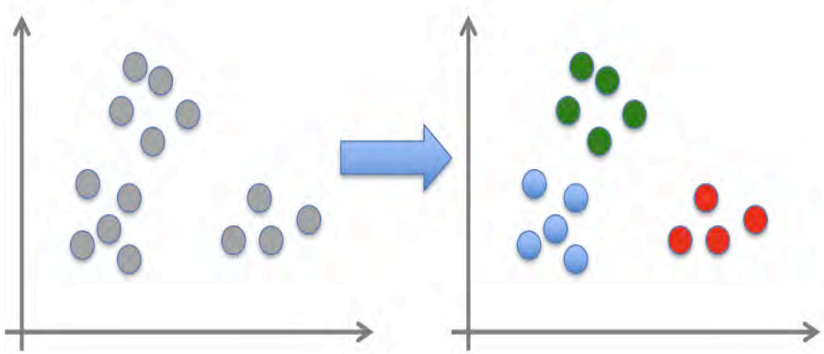
Given $x_1, x_2, ..., x_n$ (without labels)

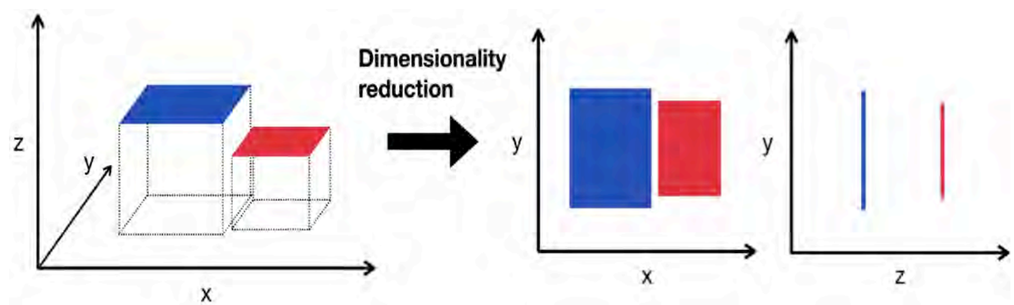
Output hidden structure behind the x's

Dimensionality Reduction

Given $x_1, x_2, ..., x_n$ (without labels)

Reduces the number of variables to get the exact information





Numericals

Given the following points: (1, 1), (1, 4), (4, 1), and (4, 4), if we want to cluster them into 2 clusters, what are the initial centroids and the final centroids after one iteration of K-means clustering?

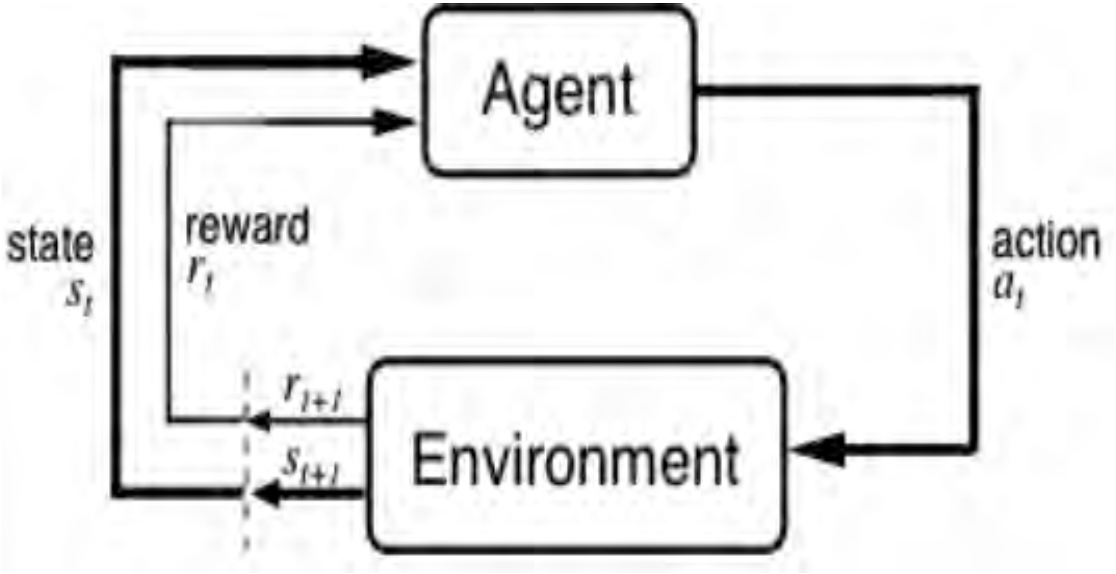
Machine Learning Algorithms

Reinforcement Learning

Discovers data through a process of trial and error and then decides what action results in higher rewards.

Major components: the agent/learner/decision-maker, the environment, and the actions.

Reinforcement Learning

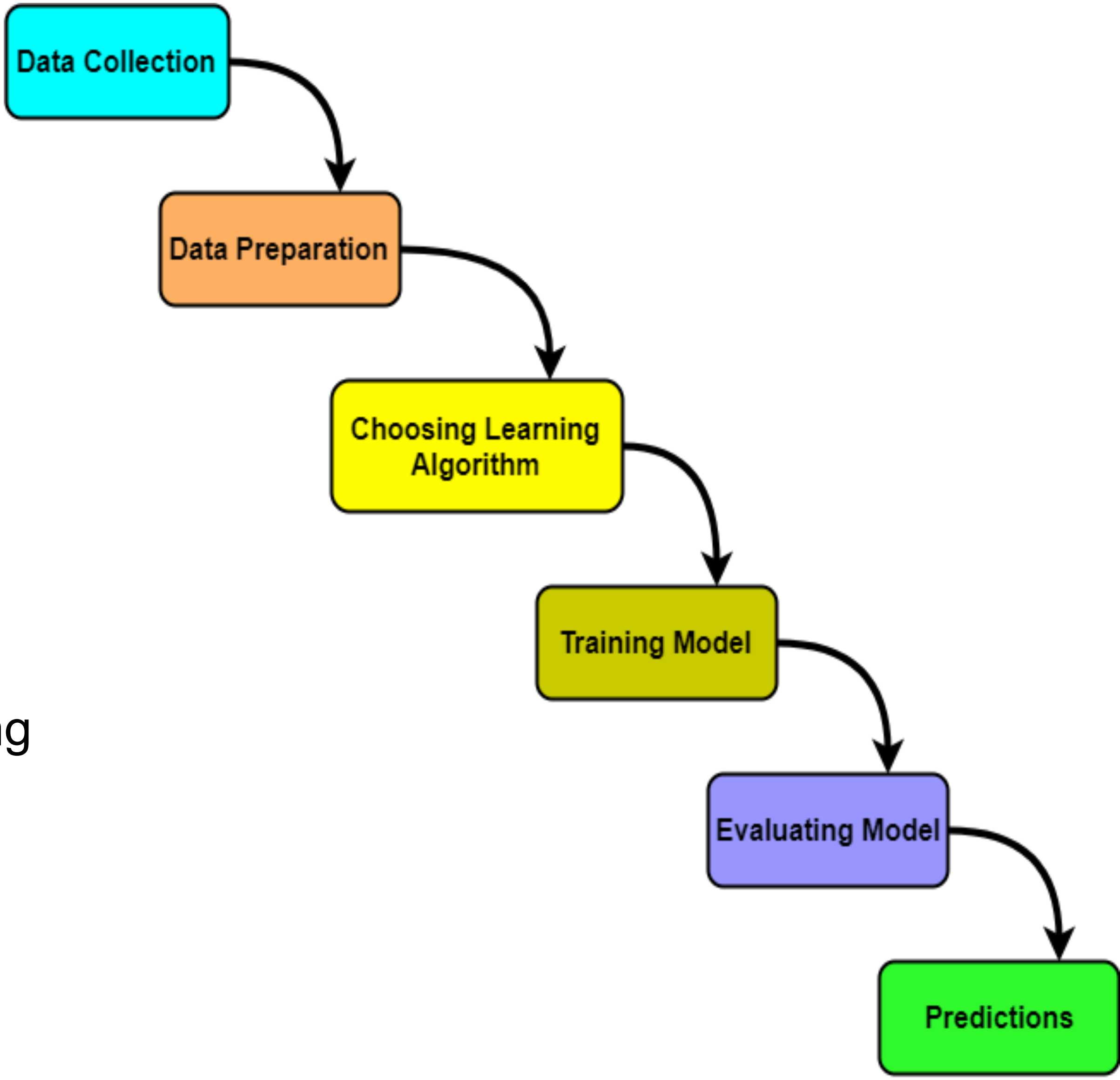


Given a sequence of states and actions with (delayed) rewards, output a policy

Policy is a mapping from states ⟶ actions that tells you what to do in a given state

Machine Learning Steps

- Data Collection
- Data Cleaning and Preprocessing
- Exploratory Data Analysis (EDA)
- Feature Engineering
- Model Selection
- Model Training
- Model Evaluation
- Hyperparameter Tuning
- Model Deployment
- Monitoring and Maintenance
- Documentation and Reporting



Machine Learning Workflow

Applications

Supervised ML

- Image Classification
- Spam Detection
- Sentiment Analysis
- Predictive Maintenance
- Medical Diagnosis
- Fraud Detection
- Speech Recognition
- Recommendation Systems
- Customer Churn Prediction
- Stock Price Prediction
- Handwriting Recognition
- Credit Scoring

Unsupervised ML

- Customer Segmentation
- Anomaly Detection
- Market Basket Analysis
- Document Clustering
- Dimensionality Reduction
- Image Compression
- Gene Sequence Analysis
- Topic Modeling
- Social Network Analysis
- Recommender Systems
- Data Visualization
- Feature Learning

Reinforcement Learning

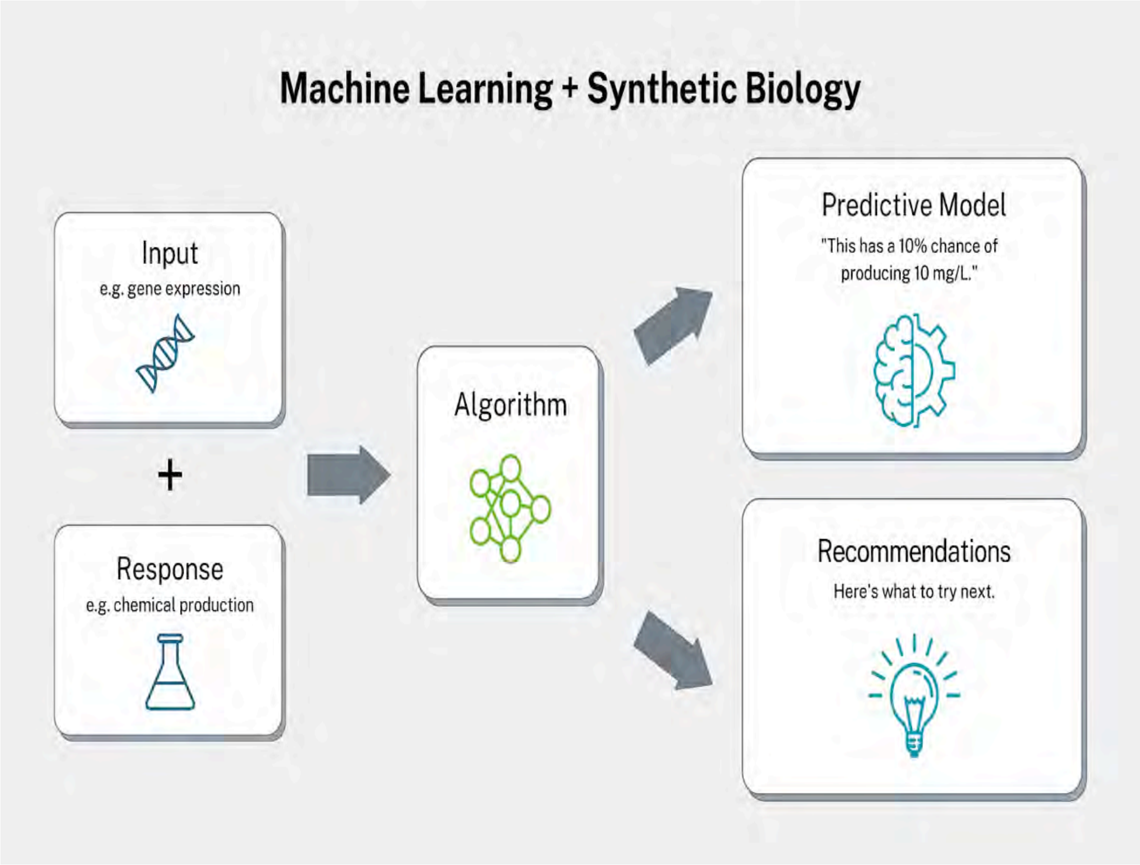
- Robotics
- Game Playing (e.g., Chess, Go, and Video Games)
- Autonomous Vehicles
- Industrial Automation
- Personalized Recommendations
- Financial Trading
- Healthcare Treatment Planning
- Natural Language Processing
- Smart Grid Management
- Supply Chain Optimization
- Advertising Bidding
- Dynamic Pricing

Machine Learning in Biological Systems

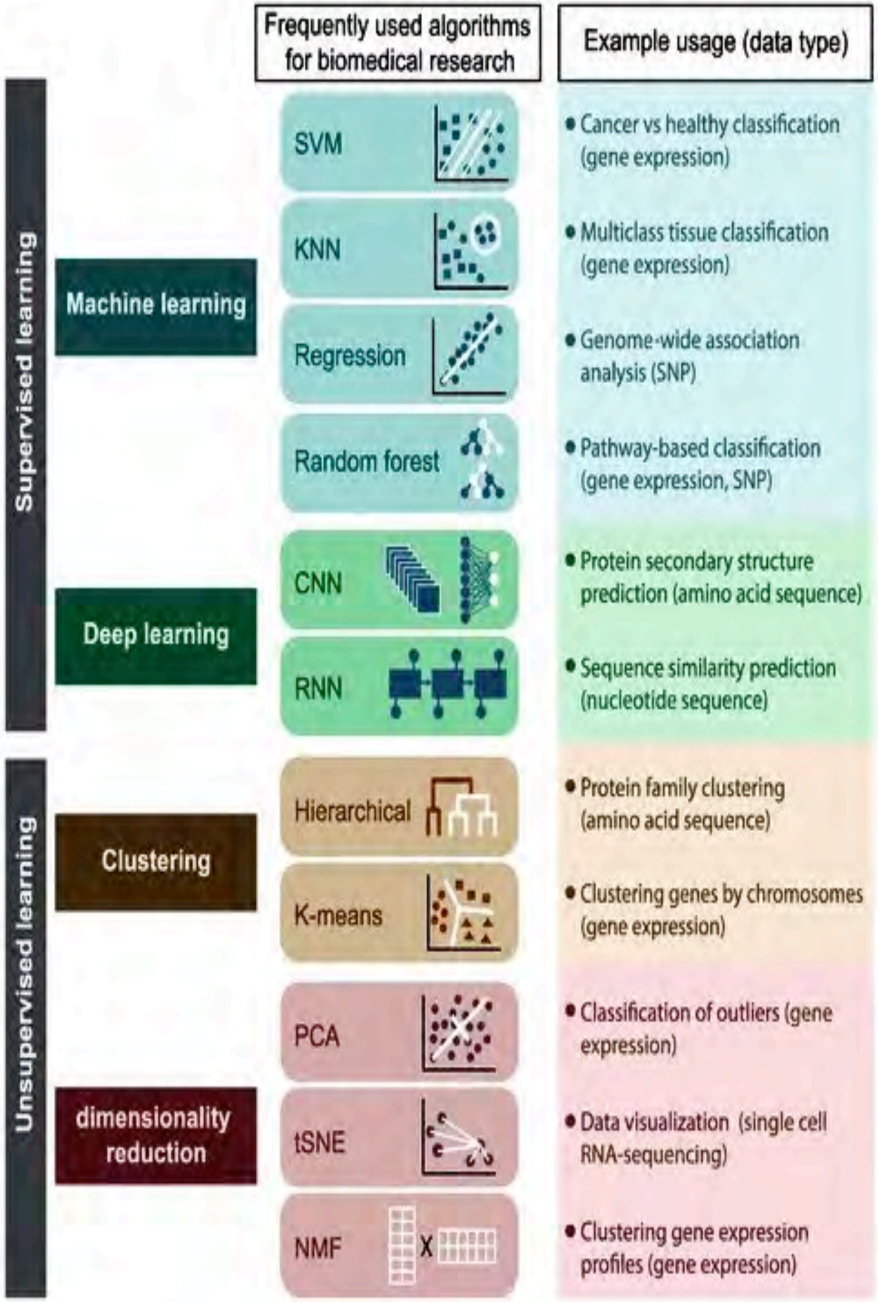
Ecology and Environmental Biology

Species Identification: ML models are used to identify species from images, sounds, and other data, aiding in biodiversity studies.

Ecosystem Modeling: ML helps in modeling ecological systems and predicting the impact of environmental changes on ecosystems.

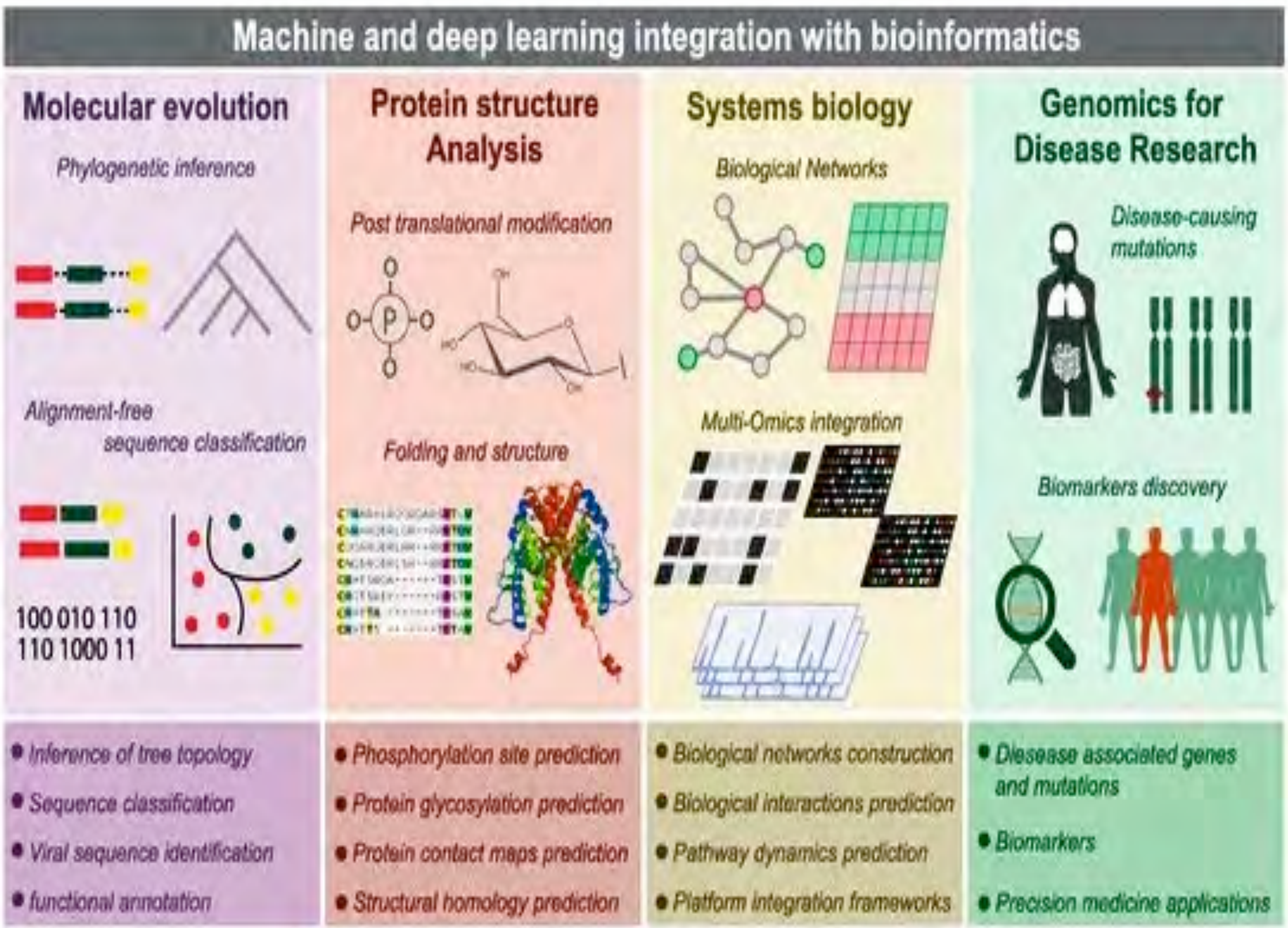


Machine Learning in Biological Systems



Ref.: Auslander, N., Gussow, A. B., & Koonin, E. V. (2021). Incorporating machine learning into established bioinformatics frameworks. *International journal of molecular sciences*, *22*(6), 2903.

Machine Learning in Biological Systems



Ref.: Auslander, N., Gussow, A. B., & Koonin, E. V. (2021). Incorporating machine learning into established bioinformatics frameworks. *International journal of molecular sciences*, *22*(6), 2903.

Machine Learning in Biological Systems

- Machine learning is increasingly being utilized to understand and model biological systems due to its ability to handle and analyze large, complex datasets.
- Genomics:
- Identifying patterns in DNA sequences,
- Predicting gene expression,
- Understanding genetic variations associated with diseases.
- Proteomics:
- Analyzing protein structures and functions,
- Predicting protein interactions,
- Identifying biomarkers for diseases.
- Drug Discovery:
- Predicting the efficacy and toxicity of new compounds,
- Identifying potential drug targets,
- Optimizing drug design.

Machine Learning in Biological Systems

- Medical Imaging:
- Enhancing the analysis of medical images (e.g., MRI, CT scans) for disease diagnosis and treatment planning.
- Systems Biology:
- Modeling complex biological networks to understand cellular processes, metabolic pathways, and disease mechanisms.
- Personalized Medicine:
- Tailoring medical treatments to individual patients based on their genetic and phenotypic information.
- Epidemiology:
- Predicting the spread of infectious diseases,
- understanding risk factors,
- planning public health interventions.
- Neuroscience:
- Analyzing neural activity data to understand brain function, cognitive processes, and neurological disorders.
- Machine learning models can uncover hidden patterns and relationships within biological data, leading to new insights and advancements in biological research and healthcare.

Challenges and Future Directions

- Data Integration: Combining data from different sources and types (e.g., genomic, proteomic, clinical) remains a challenge.
- Interpretability: Understanding how ML models make decisions is crucial for their acceptance in critical fields like medicine.
- Scalability: Handling large-scale biological data efficiently requires scalable ML methods.
- Ethical Considerations: Issues such as data privacy and the ethical implications of ML-driven decisions need careful consideration.
- Overfitting
- Underfitting