National Workshop on Bioinformatics: AI in Healthcare

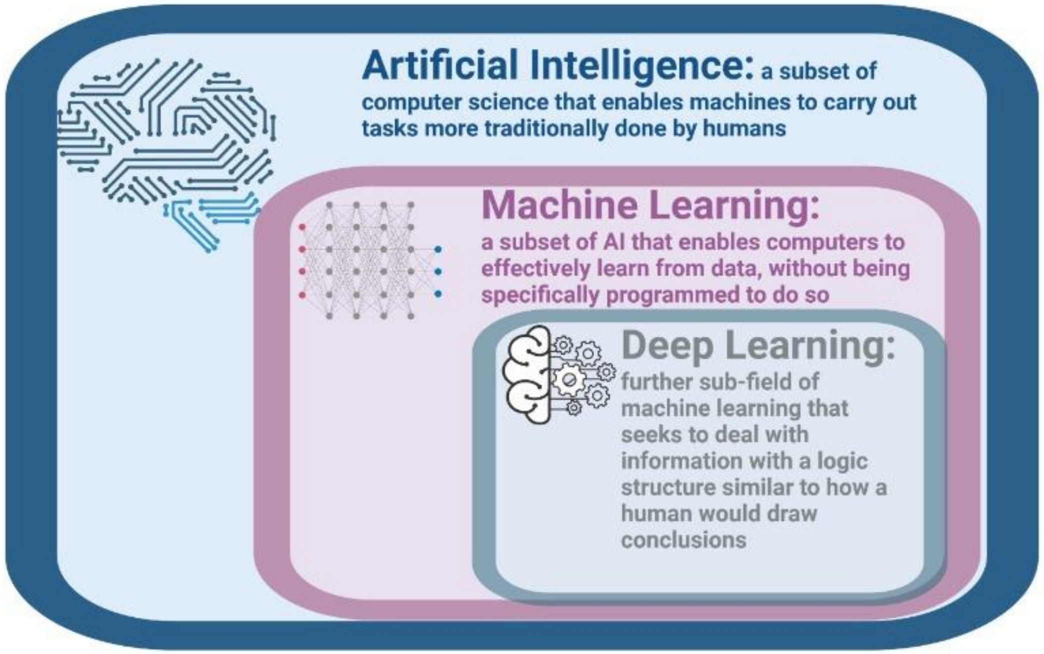
Hands-on session

Machine Learning Algorithm Applications using WEKA

16th January 2024
Bose Institute, Kolkata

What is Artificial Intelligence?

- AI is a **subset of computer science** that enables machines to carry out tasks traditionally done by humans.
- It is an over-arching term that includes two subsets: machine learning and deep learning.

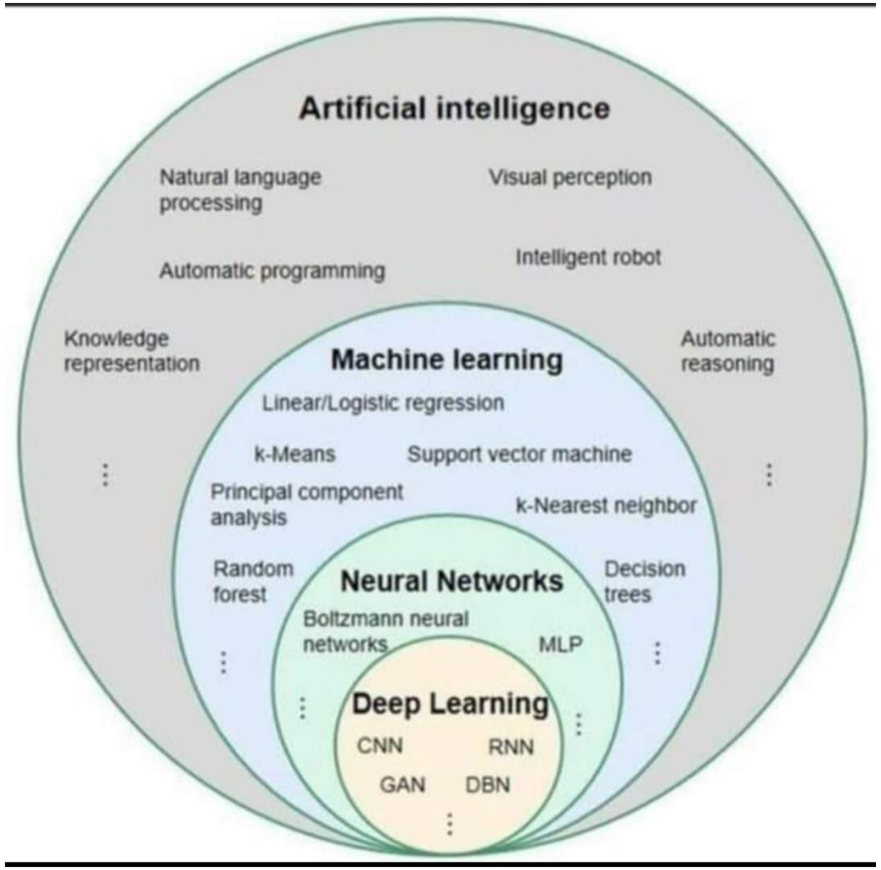


Artificial Intelligence: a subset of computer science that enables machines to carry out tasks more traditionally done by humans

Machine Learning: a subset of AI that enables computers to effectively learn from data, without being specifically programmed to do so

Deep Learning: further sub-field of machine learning that seeks to deal with information with a logic structure similar to how a human would draw conclusions

Figure 1: The three key terms in AI and how they are related



Machine Learning

- Machine learning was defined in 90's by ***Arthur Samuel*** as, "***it is a field of study that gives the ability to the computer for self-learn without being explicitly programmed***", which means imbuing knowledge to machines without hard-coding it.

- In machine learning, **algorithms are 'trained' to build a model based on sample data**, enabling them to make subsequent predictions or decisions.

- Examples of Machine Learning are k-nearest neighbor, Naïve Bayes, Support Vector Machine (SVM)

Machine Learning vs Artificial Intelligence

- **What is the difference between machine learning and artificial intelligence?**

AI solves tasks that require human intelligence while ML is a subset of **artificial intelligence that solves specific tasks by learning from data and making predictions.**

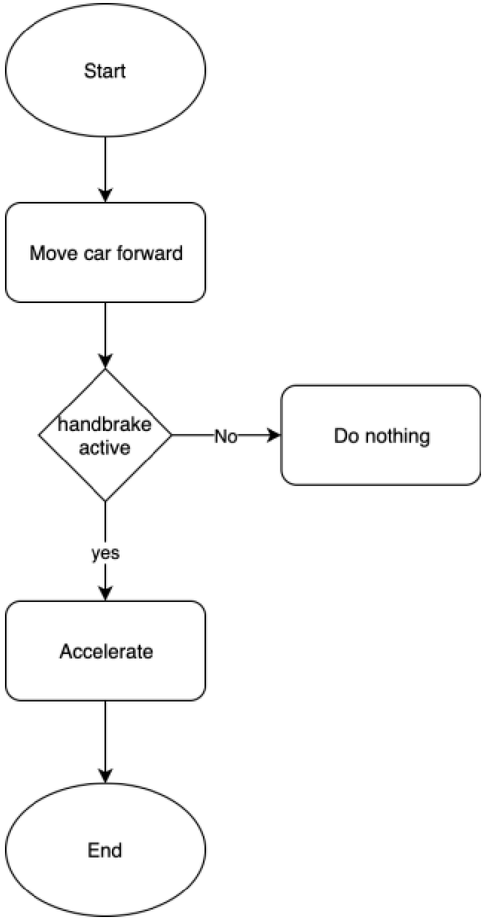
This means that all machine learning is AI, but not all AI is machine learning.

- **How is AI linked to machine learning?**

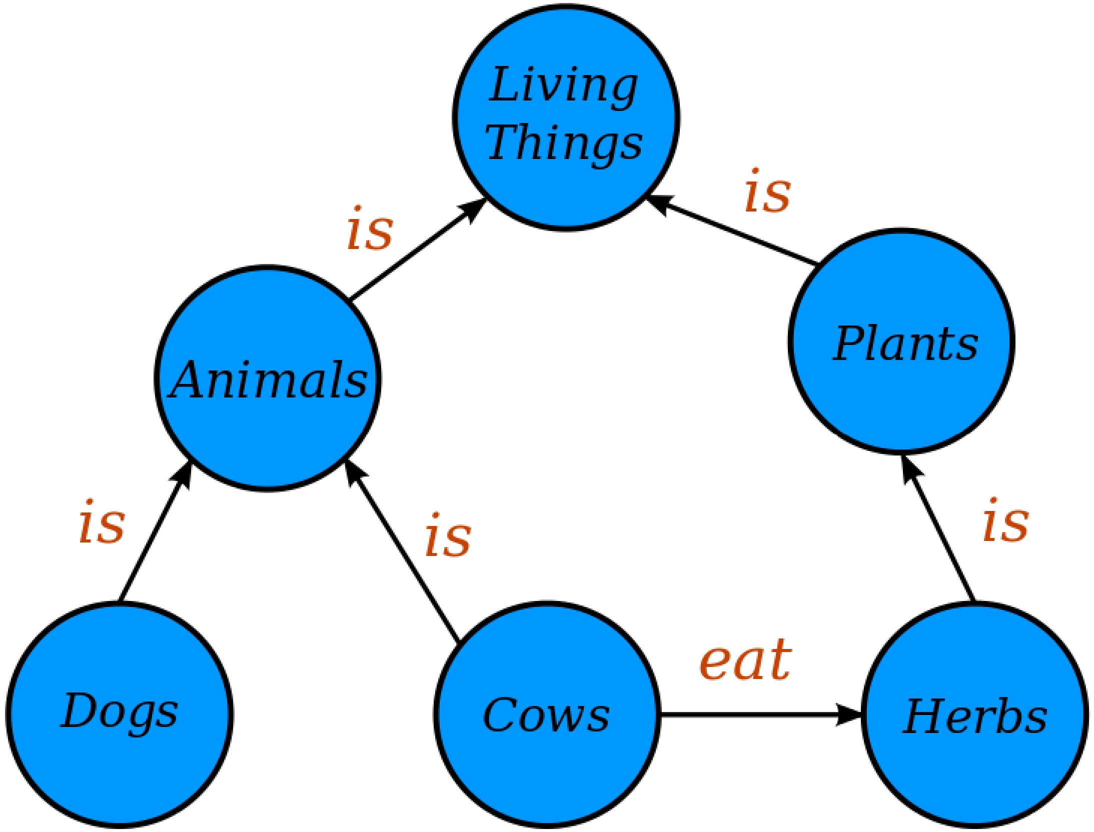
Machine learning is an application of AI. It's the process of using mathematical models of data to help a computer learn without direct instruction. This enables a computer system to continue learning and improving on its own, based on experience.

Examples of AI that are not machine learning?

For example, symbolic logic – rules engines, expert systems and knowledge graphs – could all be described as AI, and none of them are machine learning.



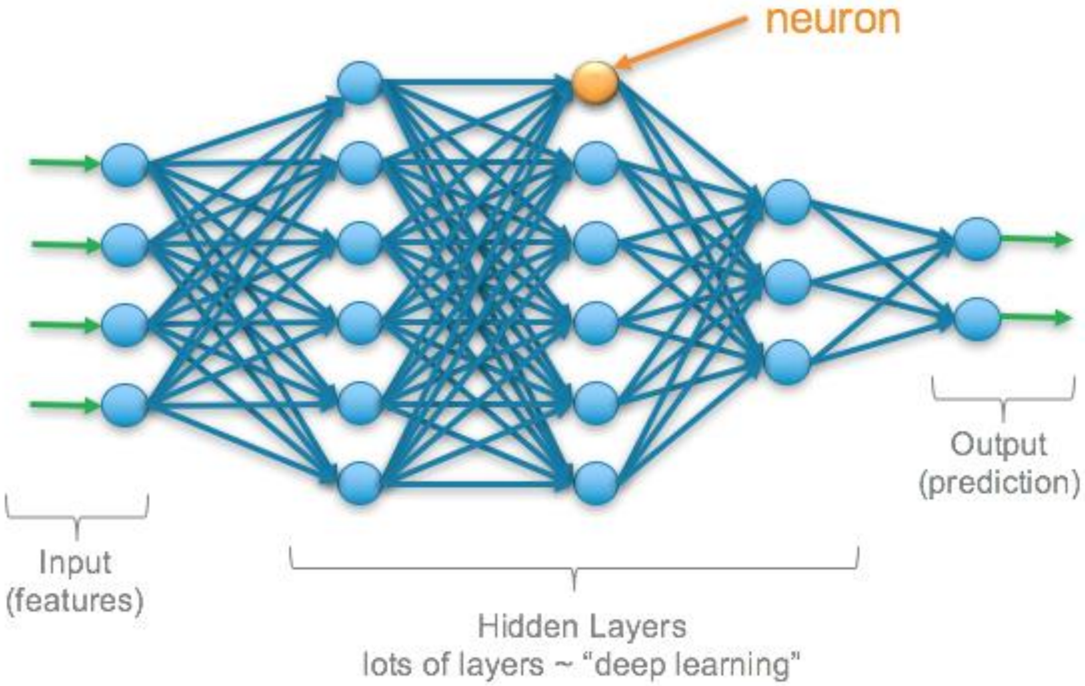
symbolic logic – rules engines



Knowledge graph

Deep Learning

- Deep learning is a further subset of machine learning that seeks to mimic the way humans handle information and draw conclusions using a 'neural network'.
- Deep learning methods build on work done on artificial neurons developed as an idea back in the 1940s to model real biological brains.



- Examples are Feed forward Neural Network (FNN), Recurrent Neural Network.

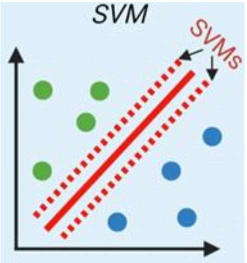
Types of Machine Learning

Supervised machine learning

fitting of a model to data (or a subset of data) that have been **labelled** — where there exists some ground truth property

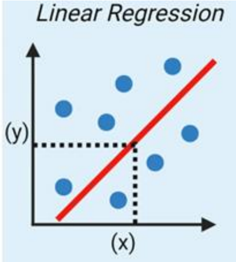
Classification

sorting of observations into pre-determined discrete categories



Regression

predict numeric outcomes from one or more variables

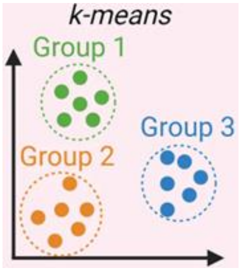


Unsupervised machine learning

able to identify patterns in **unlabelled data**, without the need to provide the system with the ground truth information

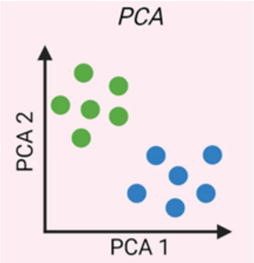
Clustering

determine the similarity between observations

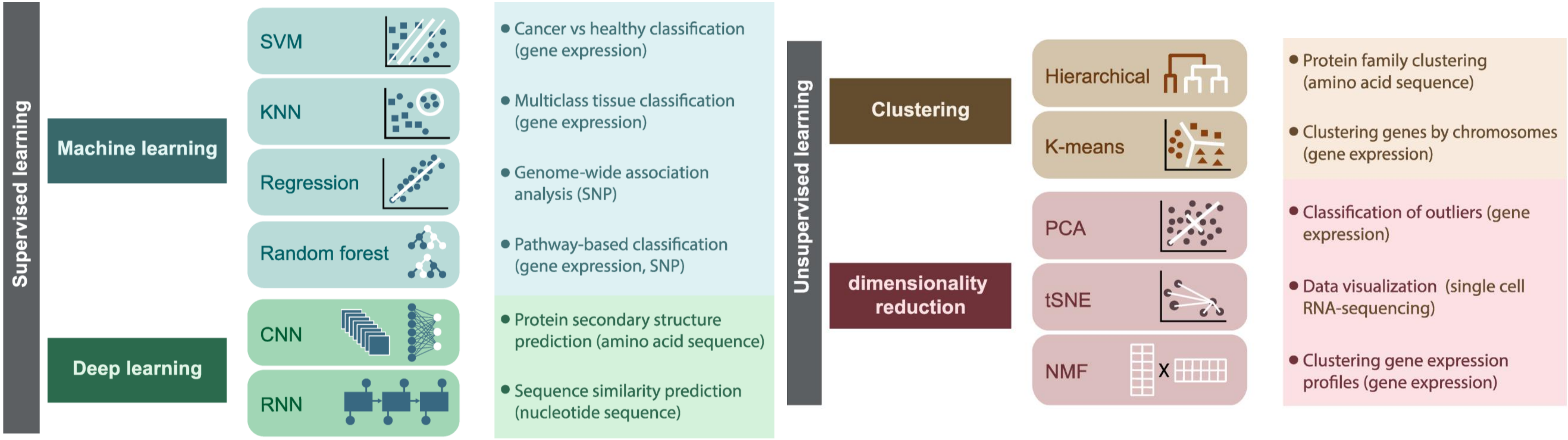


Dimensionality reduction

reduce the number of variables under assessment



Machine Learning Application in Bioinformatics



Supervised learning

Machine learning

SVM

KNN

Regression

Random forest

- Cancer vs healthy classification (gene expression)
- Multiclass tissue classification (gene expression)
- Genome-wide association analysis (SNP)
- Pathway-based classification (gene expression, SNP)

Deep learning

CNN

RNN

- Protein secondary structure prediction (amino acid sequence)
- Sequence similarity prediction (nucleotide sequence)

Unsupervised learning

Clustering

Hierarchical

K-means

- Protein family clustering (amino acid sequence)
- Clustering genes by chromosomes (gene expression)

dimensionality reduction

PCA

tSNE

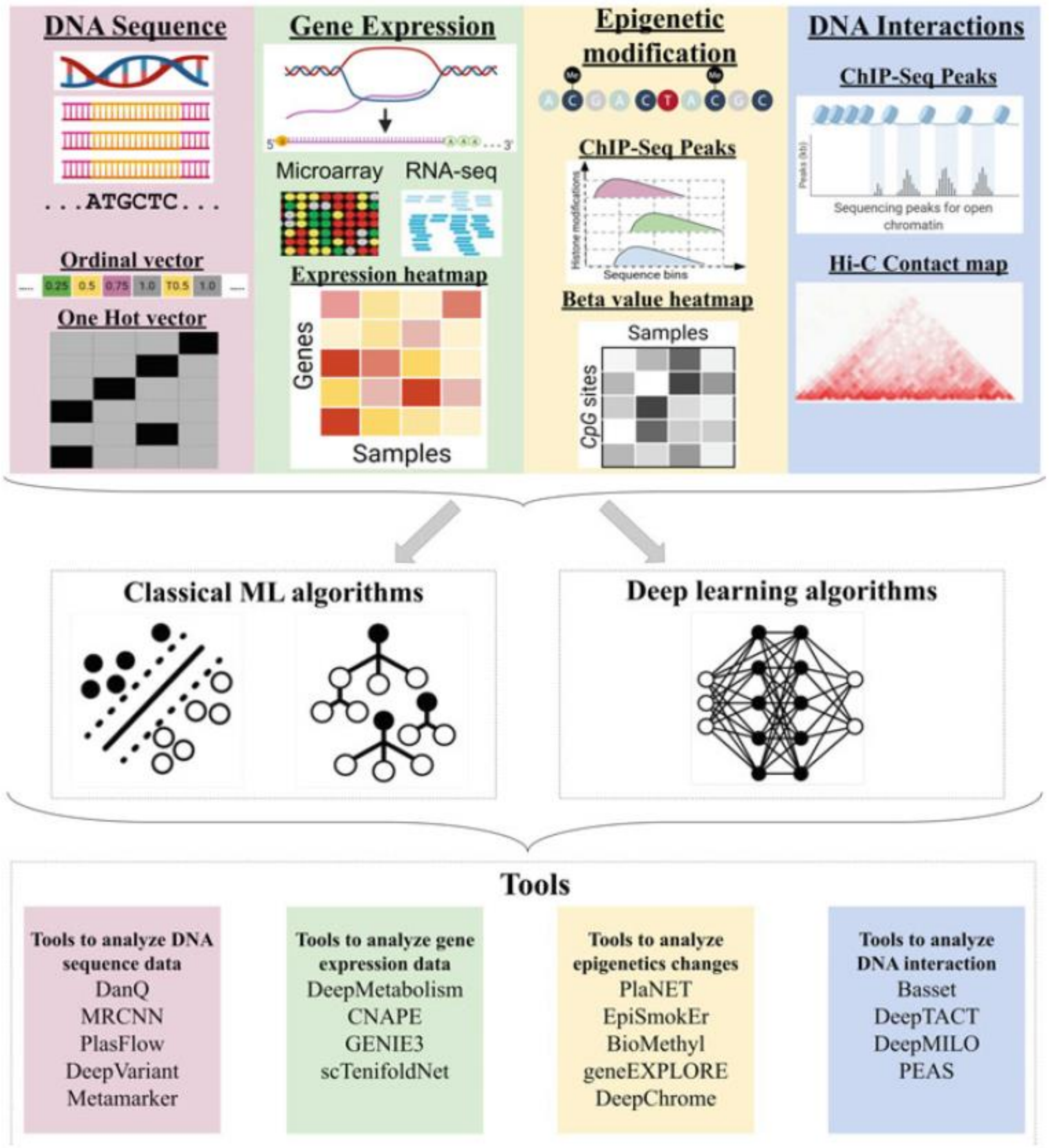
NMF

X

- Classification of outliers (gene expression)
- Data visualization (single cell RNA-sequencing)
- Clustering gene expression profiles (gene expression)

Auslander N. et al. *Int. J. Mol. Sci.* **2021**

Role of ML in Omics data analysis



Tools to analyze DNA sequence data
DanQ
MRCNN
PlasFlow
DeepVariant
Metamarker

Tools to analyze gene expression data
DeepMetabolism
CNAPE
GENIE3
scTenifoldNet

Tools to analyze epigenetics changes
PlaNET
EpiSmokEr
BioMethyl
geneEXPLORE
DeepChrome

Tools to analyze DNA interaction
Basset
DeepTACT
DeepMILO
PEAS

Bhattacharjee S. et al. **2022**

Tools and packages

Weka

Classification and Regression Training (caret)

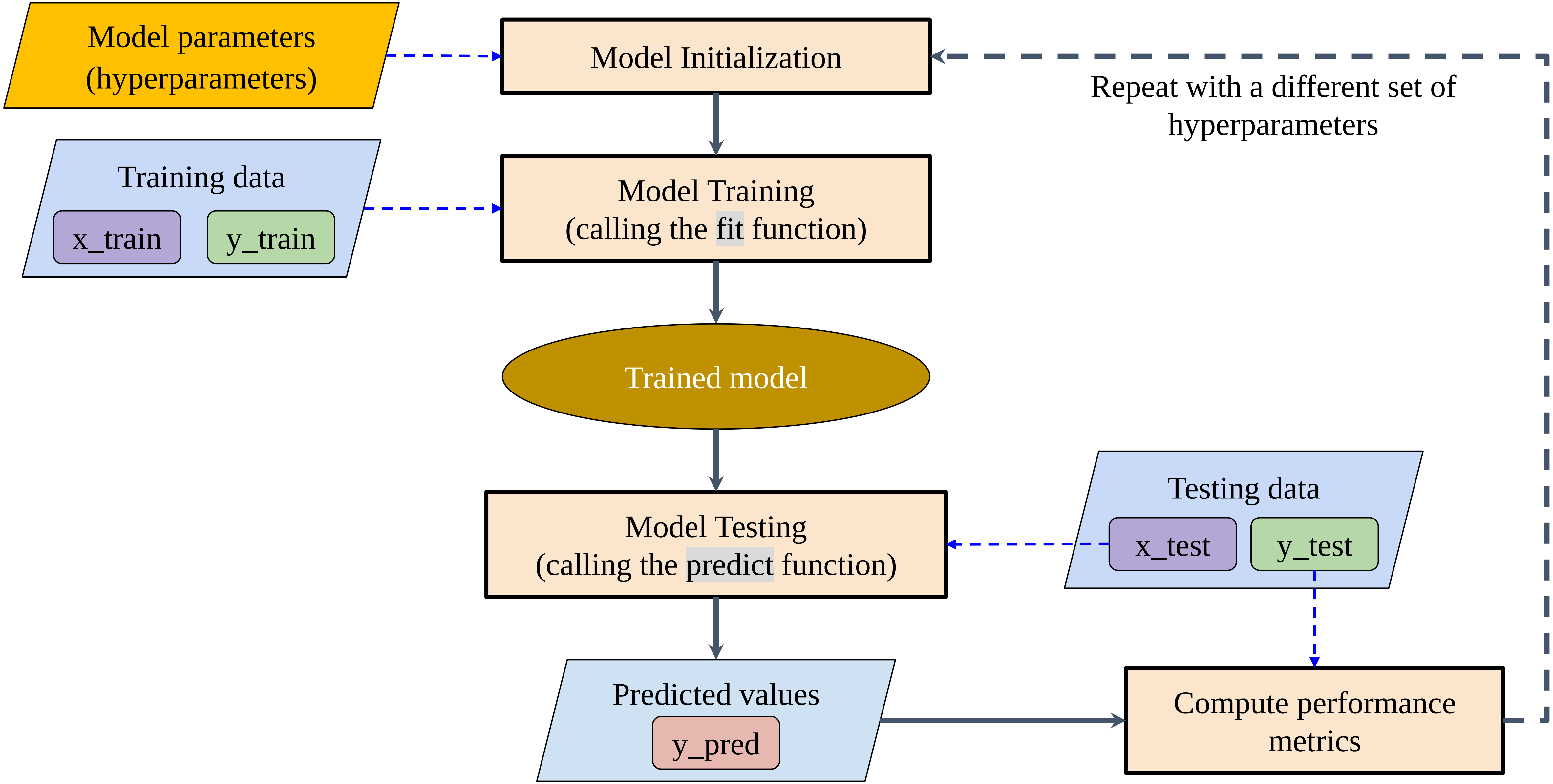
- Package in R
- http://topepo.github.io/caret/index.html

TensorFlow

PyTorch

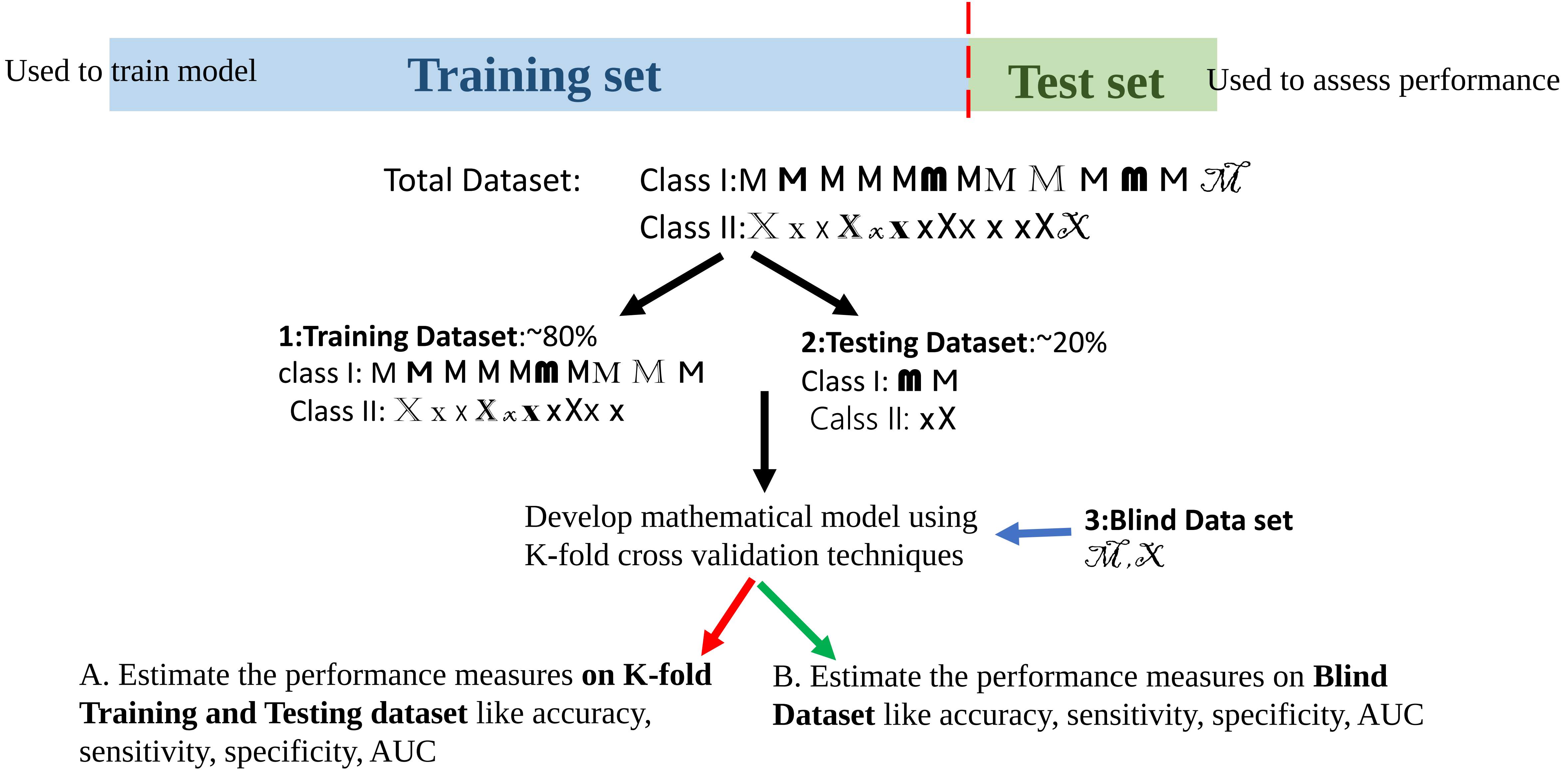
❑ machine learning library in Python
❑ https://scikit-learn.org/stable/

Workflow for training and testing machine learning models



Splitting the dataset

- Split the labelled dataset into two parts (often 80% training and 20% testing).



Used to train model

Training set

Test set

Used to assess performance

Total Dataset: Class I:M M M M MM MM M M M M M

Class II:X x x X x x xXx x xXX

1:Training Dataset:~80%

class I: M M M M MM MM M M

Class II: X x x X x x xXx x

2:Testing Dataset:~20%

Class I: M M

Calss II: xX

Develop mathematical model using K-fold cross validation techniques

3:Blind Data set

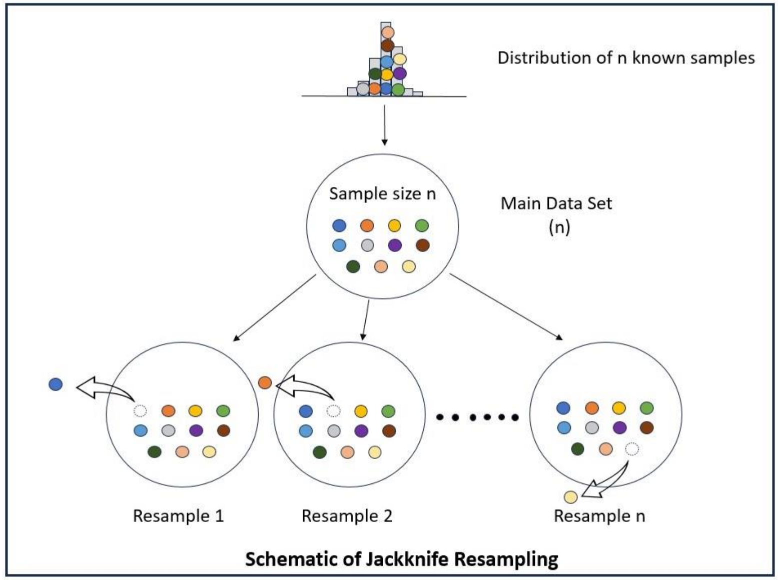
M, X

A. Estimate the performance measures **on K-fold Training and Testing dataset** like accuracy, sensitivity, specificity, AUC

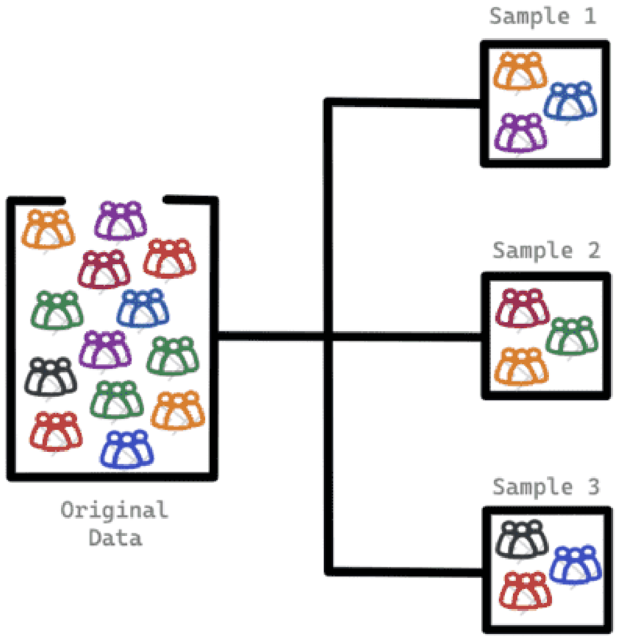
B. Estimate the performance measures on **Blind Dataset** like accuracy, sensitivity, specificity, AUC

Validation using resampling techniques

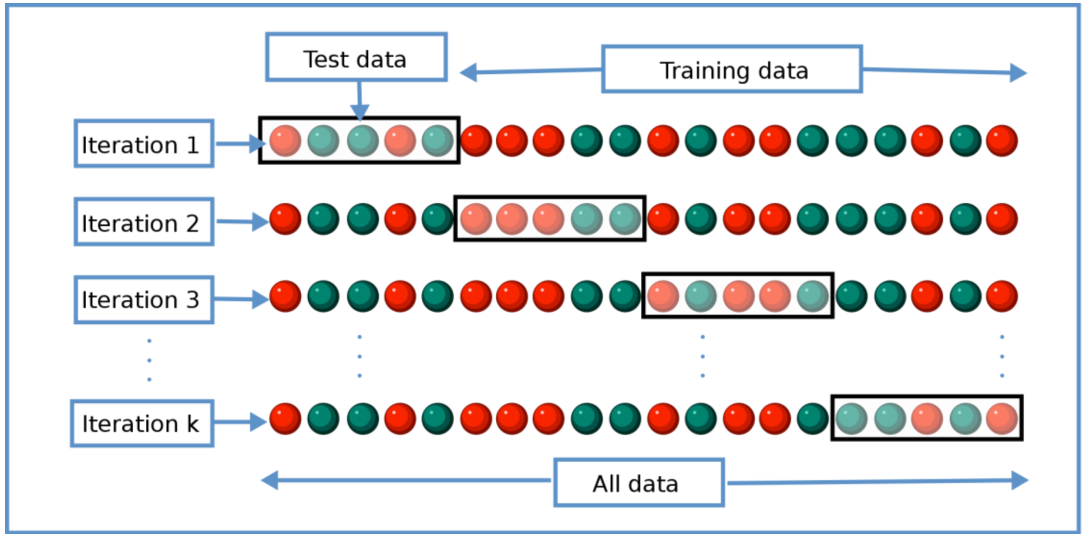
- ✓ Jackknife
- ✓ Bootstrap
- ✓ k-fold cross-validation (k = 5, 10)



Schematic of Jackknife Resampling



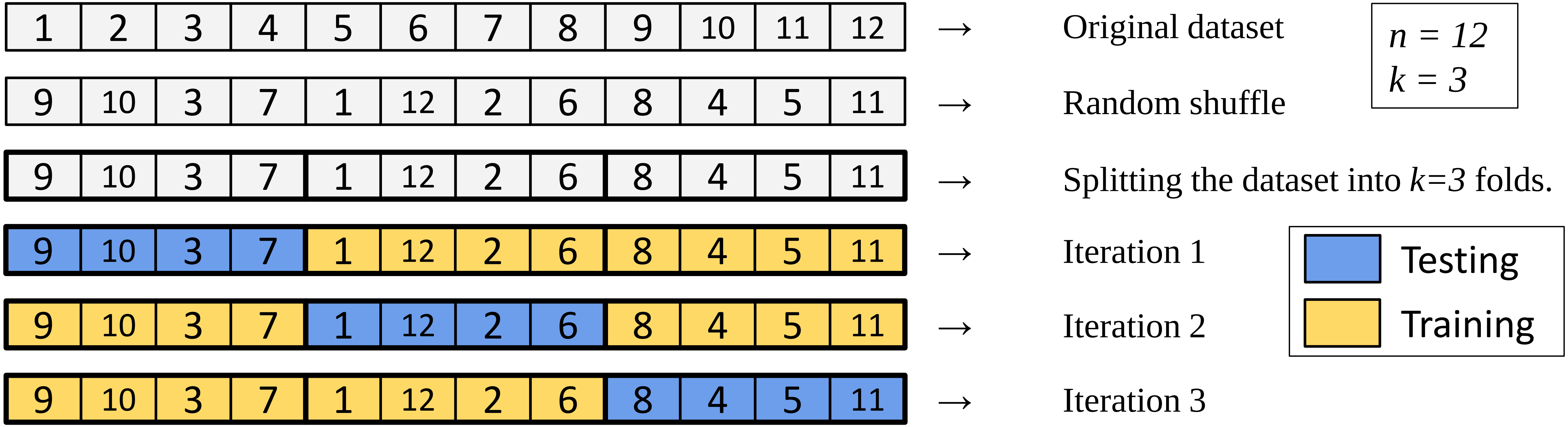
Bootstrapping



k-fold cross-validation

k-fold cross validation

- It is a cross validation technique in which the samples are randomly partitioned into *k* equal-sized and disjoint subsets, called folds.
- In each iteration, a single fold is used as the validation data for testing the model.
- The remaining ***k* − *1* folds are used as training data**



1	2	3	4	5	6	7	8	9	10	11	12	→	Original dataset
9	10	3	7	1	12	2	6	8	4	5	11	→	Random shuffle
9	10	3	7	1	12	2	6	8	4	5	11	→	Splitting the dataset into *k*=*3* folds.
🟦 9	🟦 10	🟦 3	🟦 7	🟨 1	🟨 12	🟨 2	🟨 6	🟨 8	🟨 4	🟨 5	🟨 11	→	Iteration 1
🟨 9	🟨 10	🟨 3	🟨 7	🟦 1	🟦 12	🟦 2	🟦 6	🟨 8	🟨 4	🟨 5	🟨 11	→	Iteration 2
🟨 9	🟨 10	🟨 3	🟨 7	🟨 1	🟨 12	🟨 2	🟨 6	🟦 8	🟦 4	🟦 5	🟦 11	→	Iteration 3

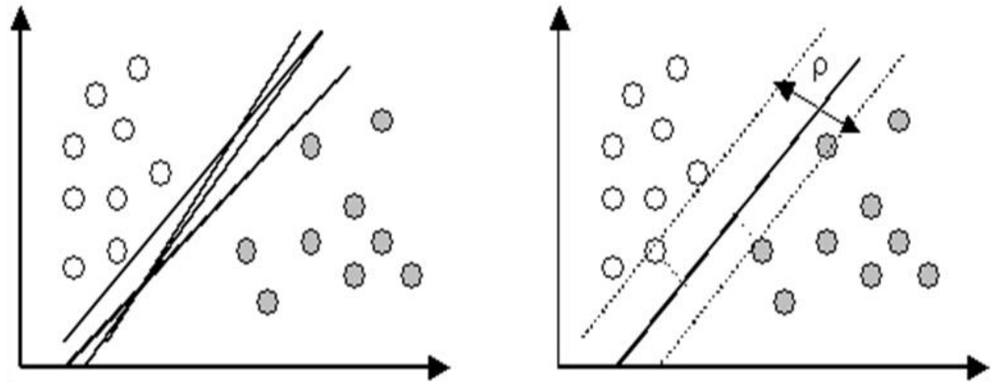
$n = 12$
$k = 3$

🟦 Testing
🟨 Training

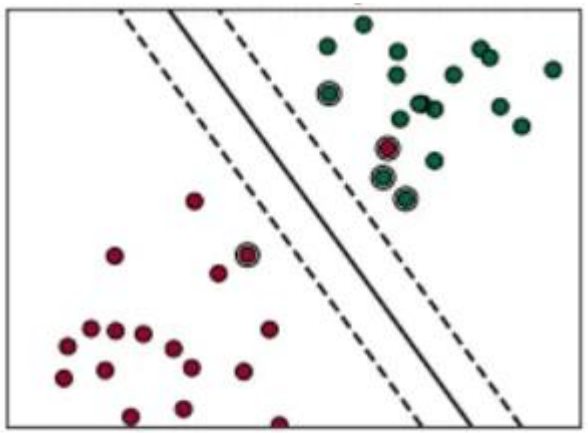
- A stratified strategy can also be used to perform the splits.
  - For **classification tasks**, each partition contains roughly the same proportions of the class labels.
  - For **regression tasks**, the mean target value is approximately equal in all the partitions.

Support Vector Machines

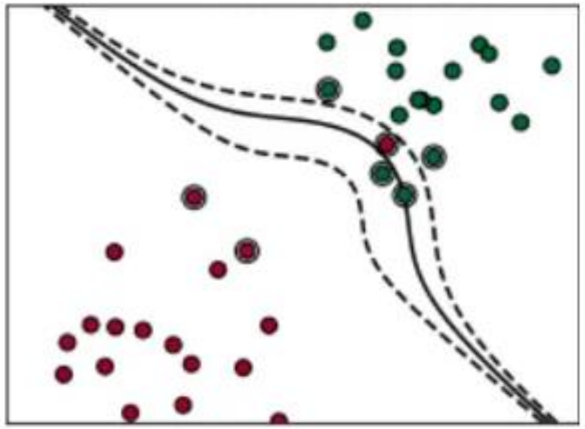
- ✓ used for classification and regression
- ✓ based on the labelled data (training data) the algorithm tries to find the **optimal hyperplane** which can be used to classify new data points.



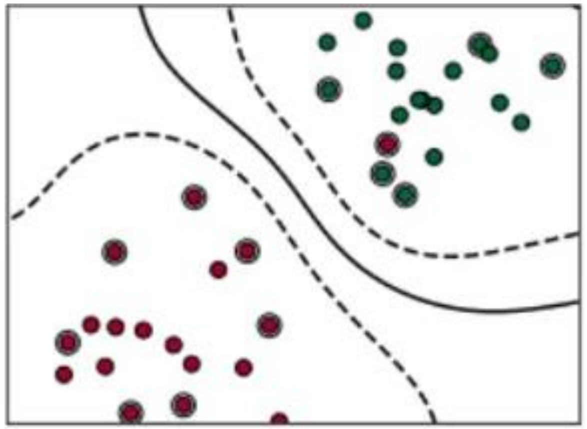
The hyperplane that separates the positive (open circle) and negative examples (closed circle). Examples closest to the hyperplane are support vectors, margin ρ of the separator is the distance between support vectors.



Linear



Polynomial



Radial Basis Function

Tutorial 1: Classification using WEKA

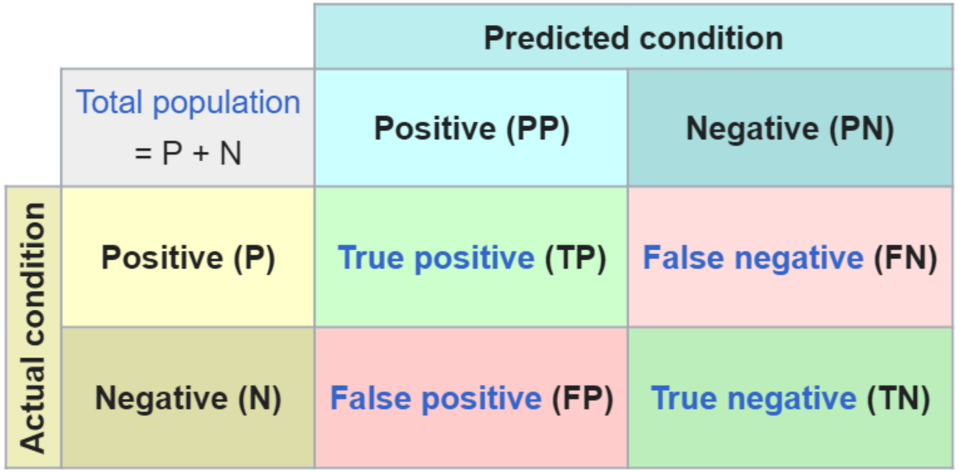
Drug resistance class prediction using Support Vector Machine

Dataset has two labels – **S (susceptible) and M (Multidrug resistant) –** Categorical datatype

Features – Alternate allele ratio of genomic mutations in *Mycobacterium tuberculosis*

Model performance

Confusion Matrix for Binary Classification



		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Type II error

Type I error

Metrics	Formula
Accuracy	(TP+TN)/(TP+TN+FP+FN)
Precision	TP/(TP+FP)
Recall/Sensitivity/True Positive Rate (**TPR**)	TP/(TP+FN)
False Positive Rate (FPR)	FP/(FP+TN)
Specificity (**1-FPR**)	TN/(TN+FP)
F1 score	2TP/(2TP+FP+FN)

$$Sensitivity = \frac{TP}{(TP+FN)} * 100$$

how many observations of **positive** class are predicted as **positive**

$$Specificity = \frac{TN}{(TN+FP)} * 100$$

how many observations of **negative** class are predicted as **negative**

$$Accuracy = \frac{TP+TN}{(TP+FP+TN+FN)} * 100$$

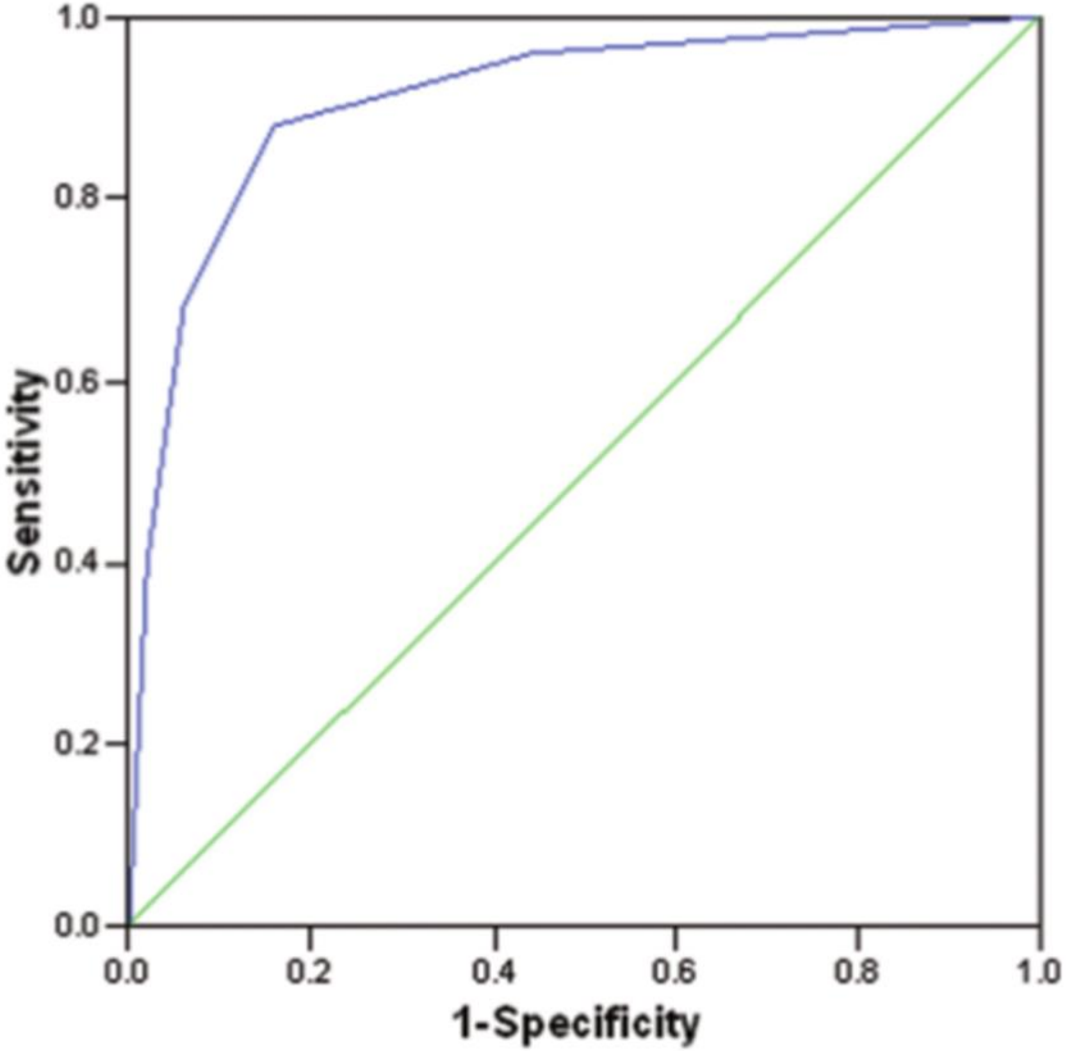
how often the classifier makes the **correct** prediction

$$F_1\,score = \frac{2TP}{2TP+FP+FN}$$

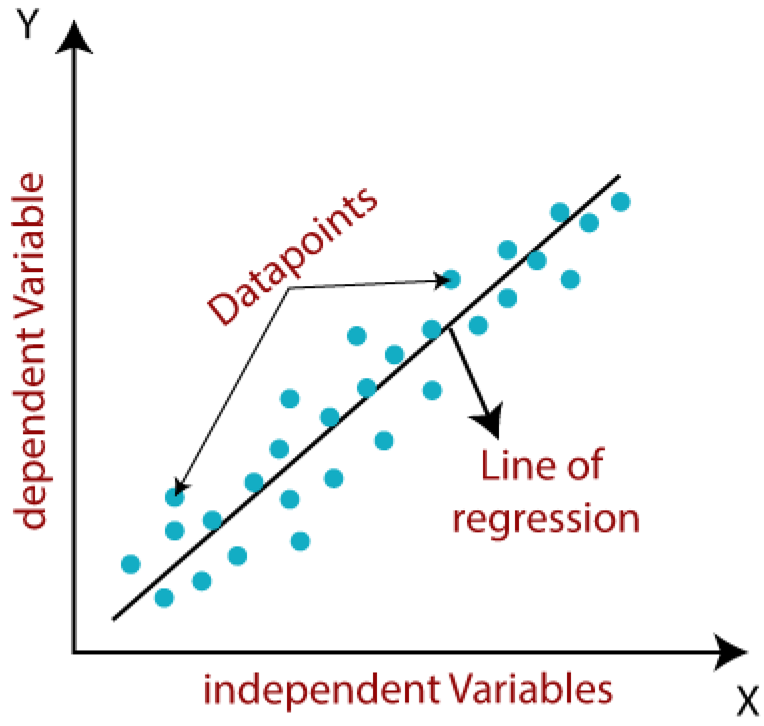
harmonic mean of precision and recall

Receiver Operating Characteristic curve, or ROC curve

- ROC curve is the plot of the Sensitivity (TPR) against the FPR (1- Specificity) at each threshold.
- The area under the ROC curve (AUC) is a measure of how well a binary classifier can distinguish between classes.



Linear Regression



❖ **method for understanding the relationship between independent variables or features and a dependent variable or outcome.**

$$y = mx + b$$

Where:
•y is the dependent variable (target),
•x is the independent variable (feature),
•m is the slope of the line (also called the weight or coefficient),
•b is the y-intercept.

Performance metrics

✓ **R-squared** or coefficient of determination is a statistical measure of how close the data are to the fitted regression line.

✓ **Other measures – MSE, MAE, Adjusted R-squared, and RMSE.**

Tutorial 2: Regression using WEKA

Age prediction using Linear Regression

Dataset has 1 label – **Age –** Numerical datatype
Features – Few marker gene expressions

Computing regression metrics

- **Mean Absolute Error (MAE)**: It is the mean of the absolute errors between the actual and predicted values of the target variable.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|actual_i - predicted_i\right|$$

- **Mean Squared Error (MSE)**: It is the mean of the square of errors between the actual and predicted values of the target variable.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(actual_i - predicted_i\right)^2$$

- **Coefficient of determination (*R*^{2}-score)**: It is a statistical measure of how well the regression predictions approximate the real data points.
  - It is the proportion of variance of the model's errors with the total variance (of the data).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(actual_i - predicted_i\right)^2}{\sum_{i=1}^{n}\left(actual_i - \overline{actual}\right)^2}$$

THANK YOU

Let us move to the hands-on session ………

Naïve Bayes Algorithm

- based on **Bayes' Theorem**
- Bayes' Theorem describes the **probability of an event**, based on a **prior knowledg**e of conditions that might be related to that event.
- assumes that the features we use to predict the target are **independent** and do not affect each other.

Given a features vector X=(x1,x2,…,xn) and a class variable y, Bayes Theorem states that:

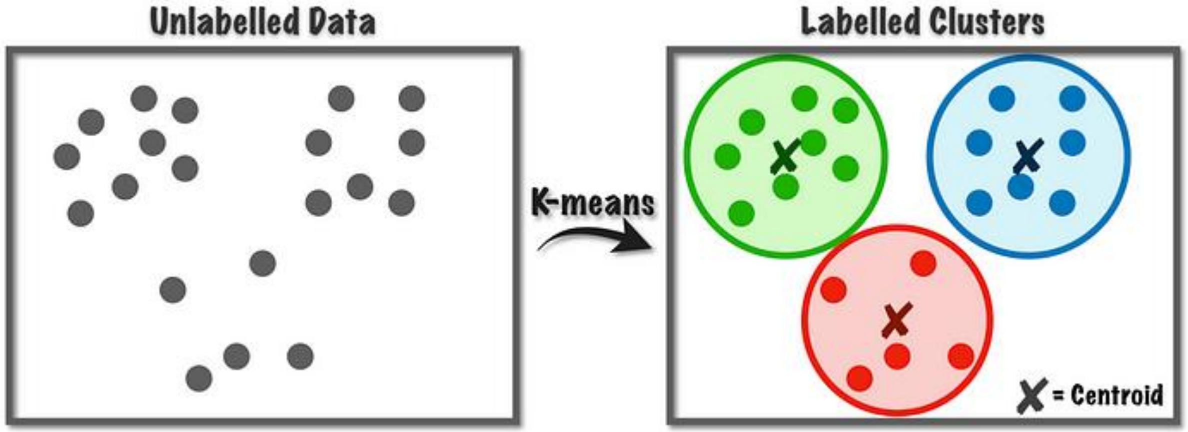
$$P(y|X) = \frac{P(X|y) * P(y)}{P(X)}$$

We're interested in calculating the posterior probability P(y | X) from the likelihood P(X | y) and prior probabilities P(y),P(X).

Clustering

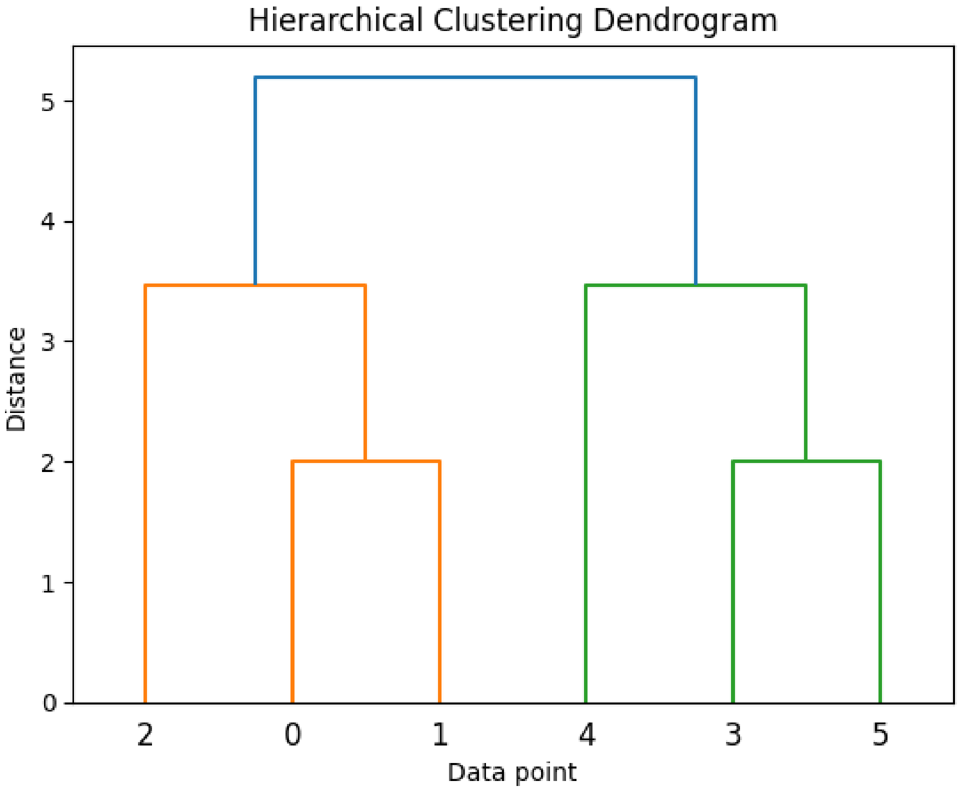
k-means clustering

- iteratively groups a collection of data points into a fixed number of clusters (k) according to their similarity.
- The algorithm aims to reduce the distance between each data point and its corresponding cluster center, also called the centroid.



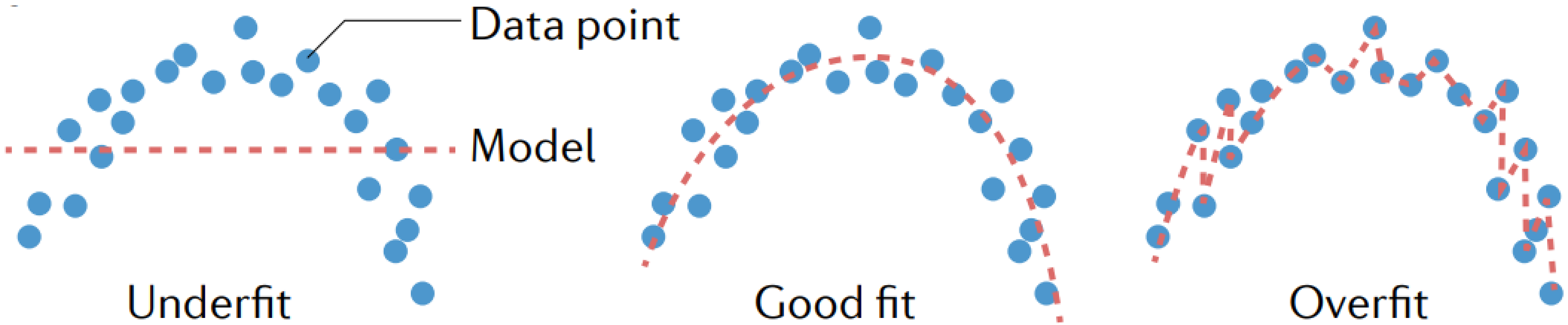
Hierarchical clustering

- aims at finding similarity between instances—quantified by a distance metric—to group them into segments called clusters
- the result of clustering is visualized as a **dendrogram**



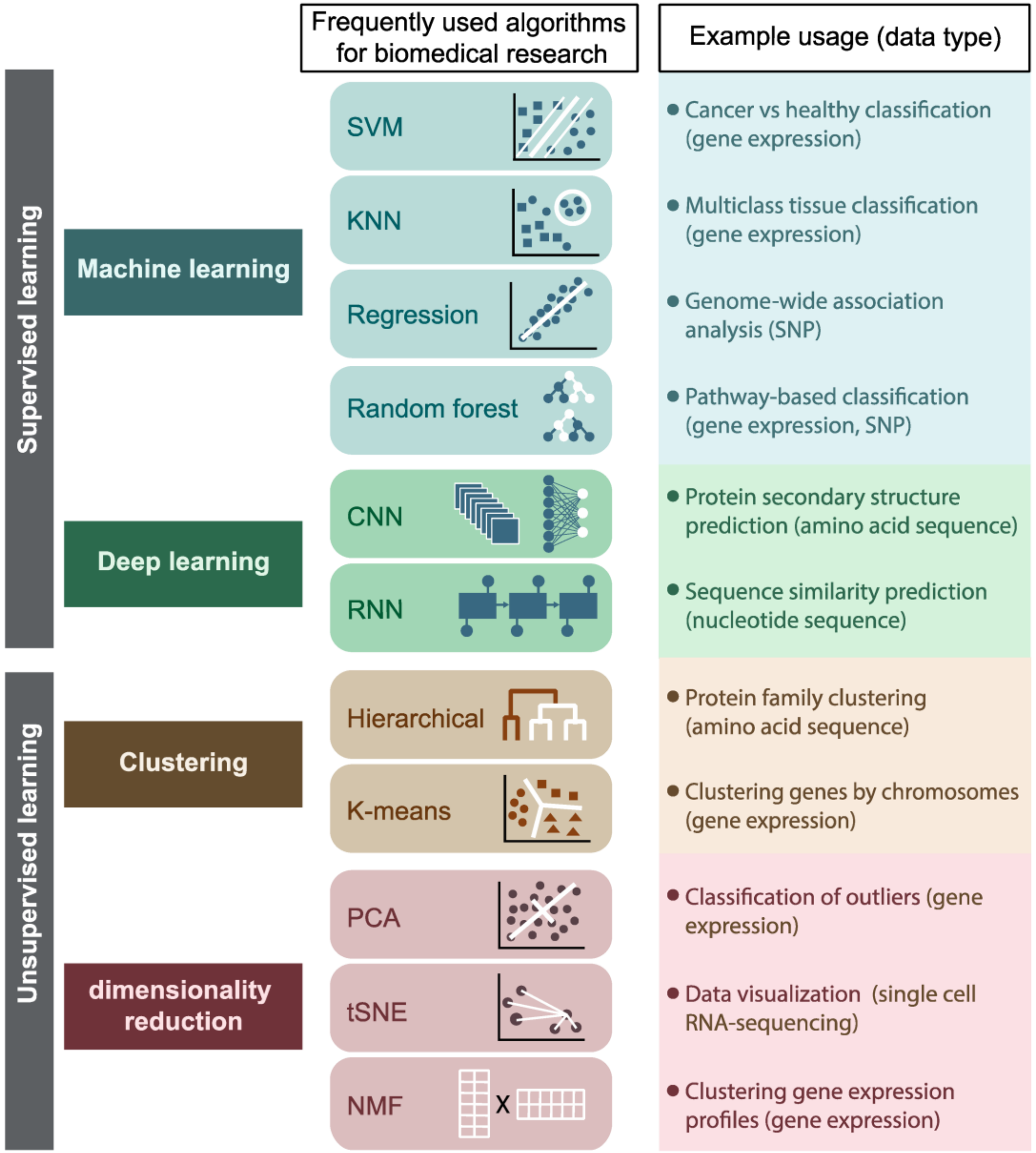
Challenges

❖ **Models that are either overfitted or underfitted will produce poor predictions on data not in the training set**

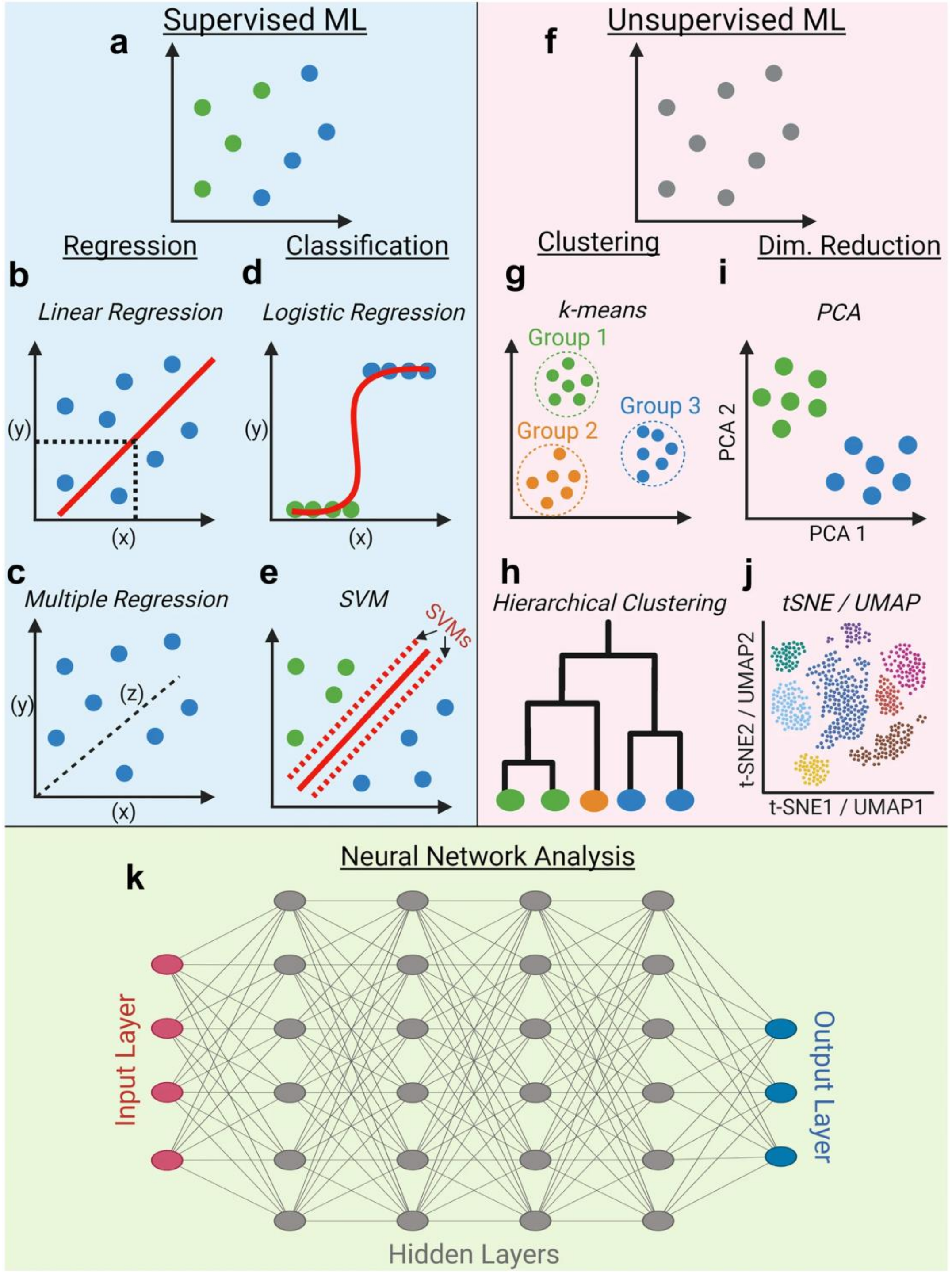


- Failing to learn the underlying relationship between the variables
- using a model **without sufficient complexity**

- learning the noise in the training data
- using a model **with too many parameters**



		Frequently used algorithms for biomedical research	Example usage (data type)
Supervised learning	Machine learning	SVM	Cancer vs healthy classification (gene expression)
		KNN	Multiclass tissue classification (gene expression)
		Regression	Genome-wide association analysis (SNP)
		Random forest	Pathway-based classification (gene expression, SNP)
	Deep learning	CNN	Protein secondary structure prediction (amino acid sequence)
		RNN	Sequence similarity prediction (nucleotide sequence)
Unsupervised learning	Clustering	Hierarchical	Protein family clustering (amino acid sequence)
		K-means	Clustering genes by chromosomes (gene expression)
	dimensionality reduction	PCA	Classification of outliers (gene expression)
		tSNE	Data visualization (single cell RNA-sequencing)
		NMF	Clustering gene expression profiles (gene expression)


Supervised ML
a
Regression
Classification
b
Linear Regression
(y)
(x)
d
Logistic Regression
(y)
(x)
c
Multiple Regression
(y)
(z)
(x)
e
SVM
SVMs
Unsupervised ML
f
Clustering
Dim. Reduction
g
k-means
Group 1
Group 2
Group 3
i
PCA
PCA 2
PCA 1
h
Hierarchical Clustering
j
tSNE / UMAP
t-SNE2 / UMAP2
t-SNE1 / UMAP1
Neural Network Analysis
k
Input Layer
Hidden Layers
Output Layer