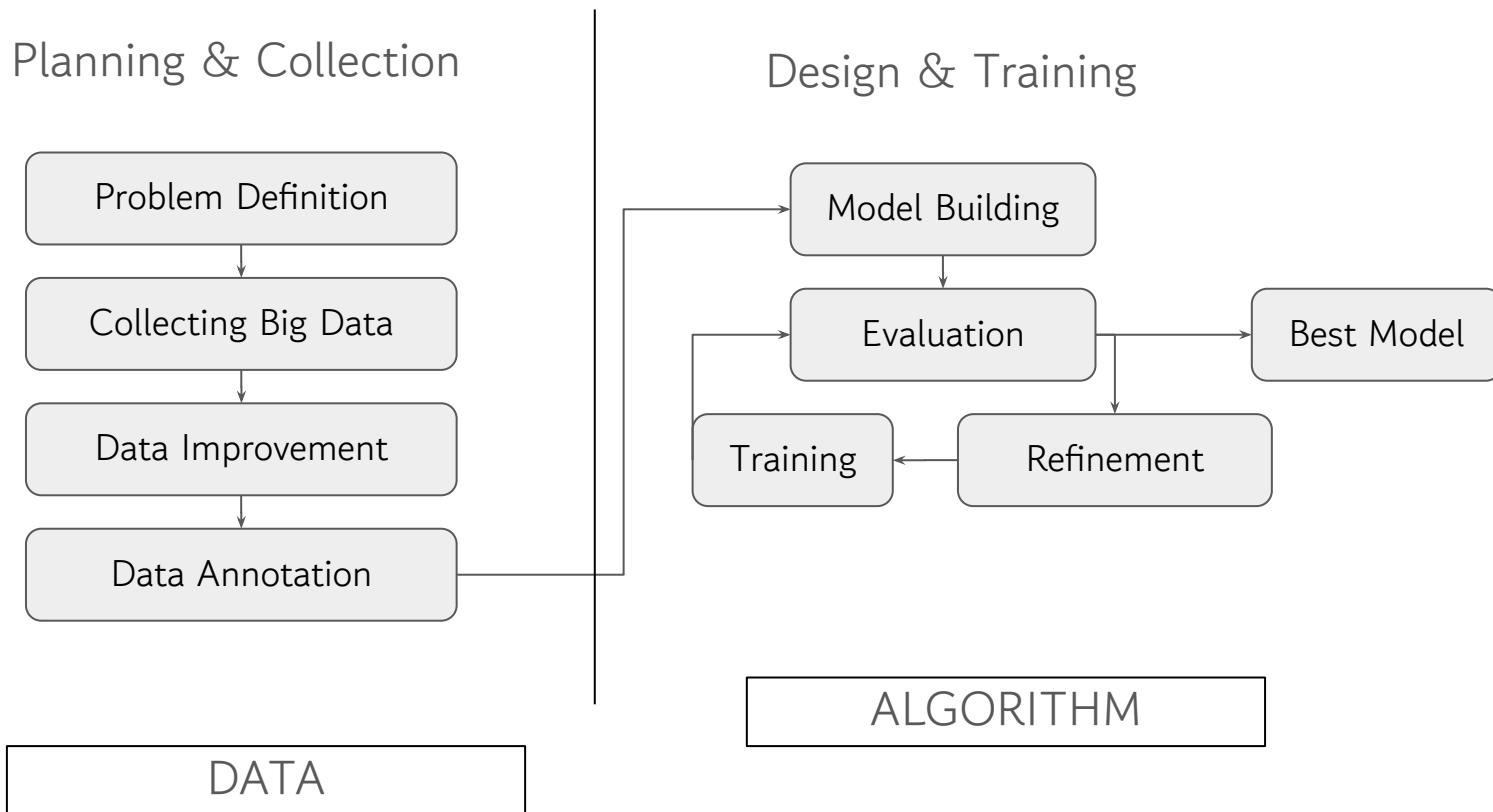




# Machine Learning on DESI Data

by  
John Suárez-Pérez, Ph.D.

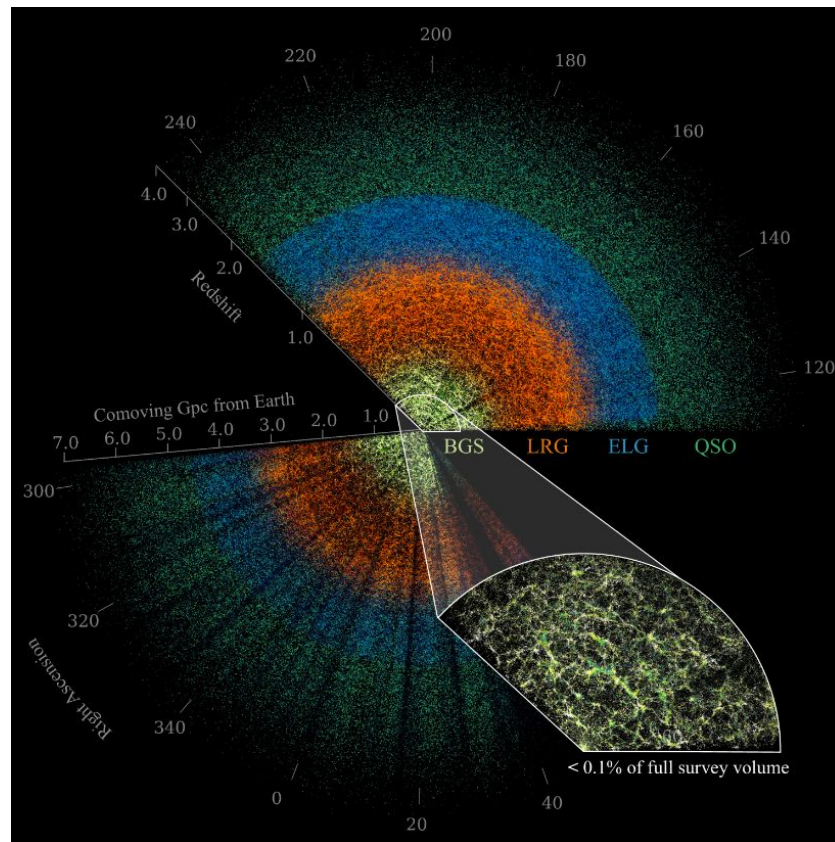




# DESI Spectra Classification



- ❑ 50 million spectra measured (as of Dec 2024) (36.3 M galaxies & quasars. 13.7 M stars.)
- ❑ Data Release 1 (DR1): 18 million unique objects.
- ❑ Data Release 2 (DR2, March 2025): includes the first 3 years of observations.
- ❑ Backup program: ~7 M Milky Way stars observed (1.2 M in DR1).





### DESI Data

Home

Data Releases

Overview

EDR

**DR1 (latest)**

Data Access

Data Organization

Data Model [external]

Software

Tutorials

Glossary

Value-added Catalogs

Papers

Data License and  
Acknowledgments

Other Resources

Contact Us

## Data Release 1 (DR1)

### Overview

DESI Data Release 1 (DR1) includes spectra for more than 18 million unique targets from [Main Survey](#) observations taken between May 2021 and June 2022. In addition, DR1 includes all the data taken as part of DESI [Survey Validation](#) which was originally released as part of the [Early Data Release](#) but reprocessed with the same reduction pipeline as the Main Survey data.

The DR1 data are released under the [Creative Commons Attribution 4.0 International License](#) (CC BY 4.0). Use of DESI data requires including the citation and acknowledgment text given on the [Data License and Acknowledgments](#) page.

**Data URL:** <https://data.desi.lbl.gov/public/dr1>

**European mirror URL:** <https://webdav-hdfs.pic.es/data/public/DESI/DR1>

**Paper:** [DESI Collaboration et al. \(2025\), Data Release 1 of the Dark Energy Spectroscopic Instrument](#)

**Cosmology Results:** [DESI Key Project Papers using DR1](#)

### Table of contents

Overview

Coverage area

Summary statistics

Redshift distributions

Large-Scale Structure Catalogs

Value-added Catalogs

General VACs

Milky Way Survey (MWS)

Extragalactic Science

Quasar Science

Lya Forest

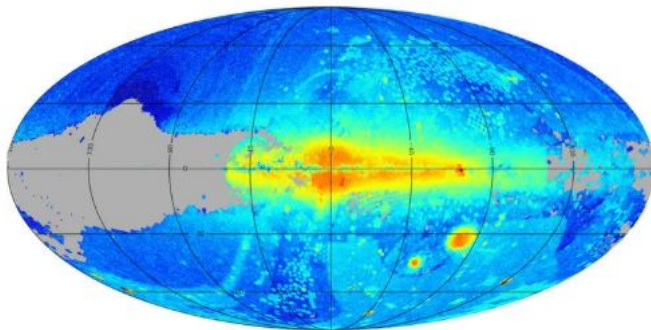
Known Issues

Software Package Versions



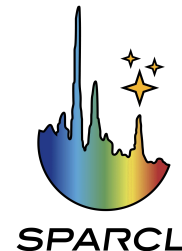
## Astro Data Lab Science Platform

---



The Astro Data Lab Science Platform enables efficient exploration and analysis of the large datasets now being generated by instruments on NOIRLab and various other wide-field telescopes.

SPARCL: Spectra Analysis  
& Retrievable Catalog Lab

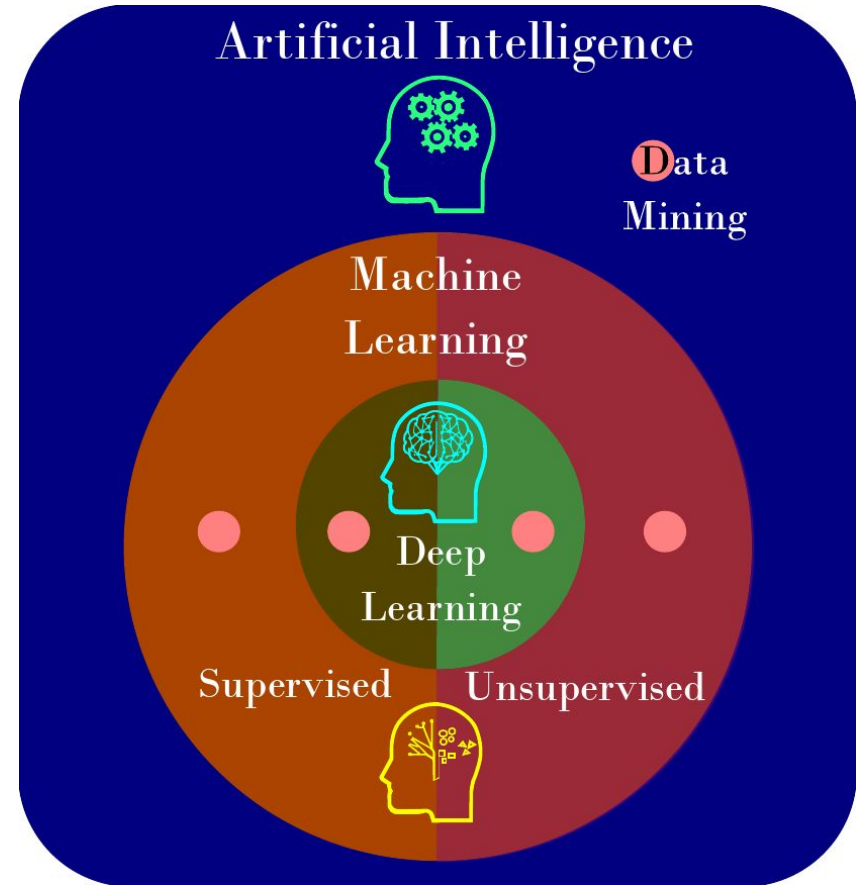


- ❑ Online service for discovery and retrieval of 1D optical/infrared spectra.
- ❑ Currently hosts data from SDSS and BOSS spectrographs.
- ❑ Designed to support DESI spectra, to be included with the first public data release.

# What is AI?



- ❑ Simulating of human intelligence in machines (visual perception, speech recognition, decision-making, language processing).
- ❑ Algorithms and models that can process large amounts of data, trained with techniques as Supervised or Unsupervised learning.
- ❑ Main subfields of AI are Machine & Deep learning. Data Mining cut across both.





## Supervised

- Used for classification or regression tasks.
- Requires to use labels to make predictions.

### ML Algorithms

- \* Support Vector Machines
- \* K-nearest neighbors
- \* Decision Trees
- \* Random Forest.

### DL Algorithms

- \* Multi-Layer Perceptron
- \* Convolutional Neural Network
- \* Recurrent Neural Network
- \* Transformers

## Unsupervised

- Used for clustering or dimensionality reduction tasks.
- Doesn't require to use labels. Used to find patterns.

### DL Algorithms

- \* Autoencoders

### ML Algorithms

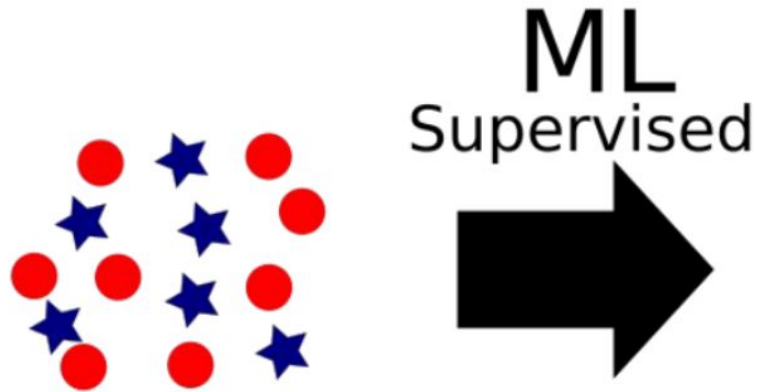
#### Clustering:

- \* K-means clustering
- \* DBScan
- \* Gaussian Mixture Models

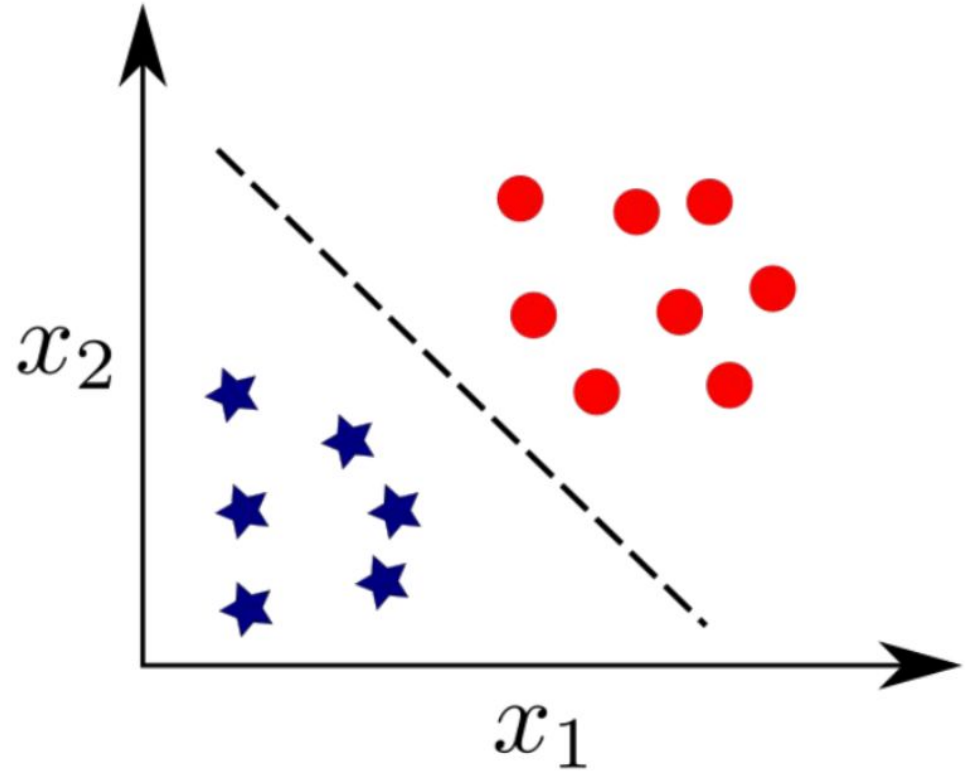
#### Dimensionality Reduction:

- \* PCA
- \* Isometric Map
- \* T-SNE
- \* UMAP

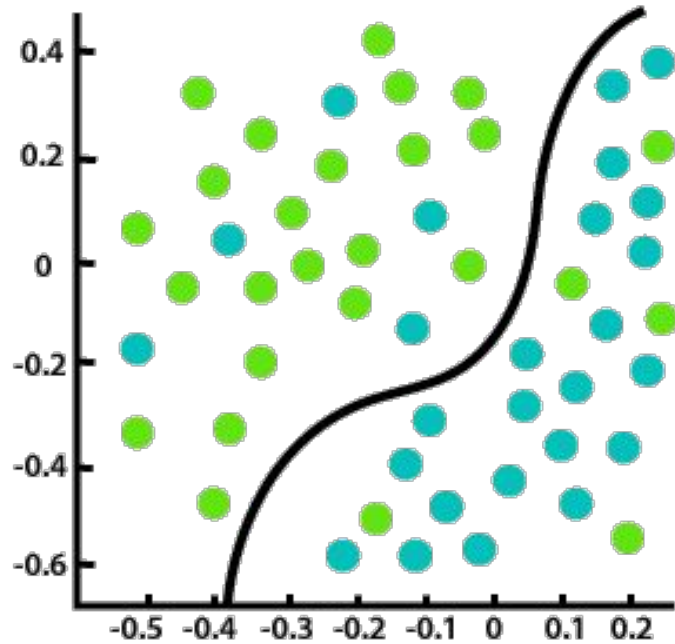




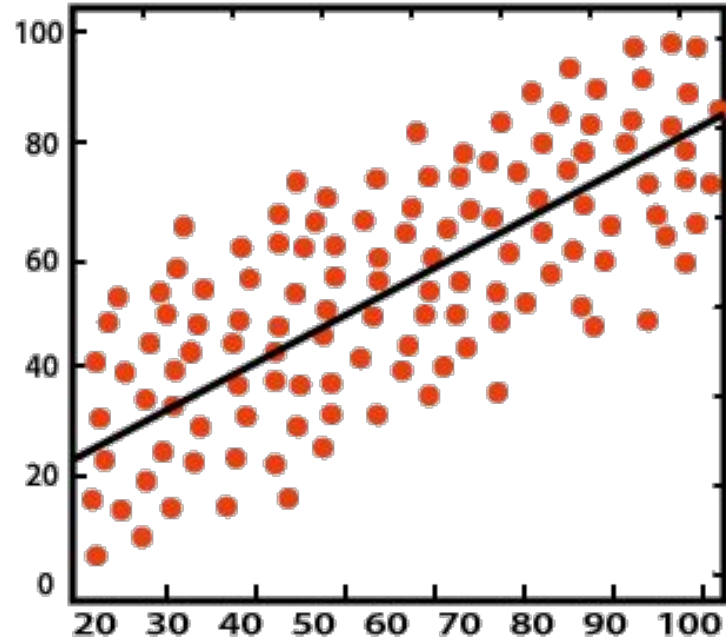
- ❑ Features: (color, number of corners, ...)
- ❑ target: (class: circle, star)  
(mass,...)



Classification



Regression

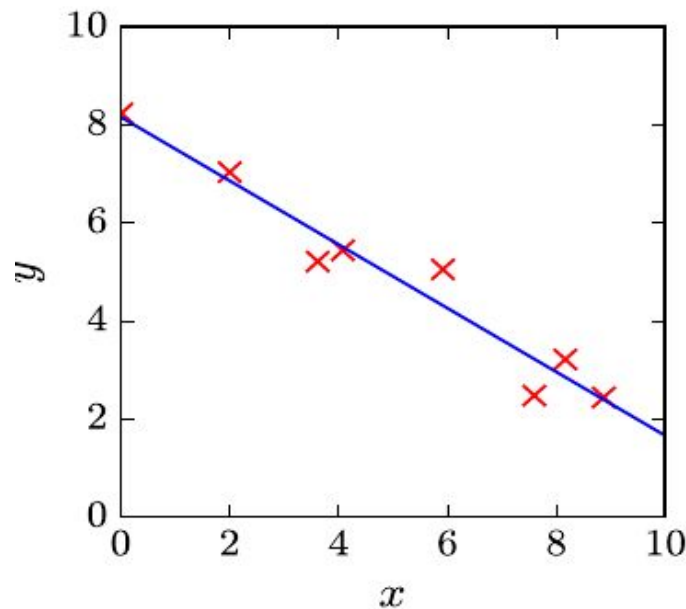




$$f(x, \vec{\beta}) = \beta_0 + \beta_1 x$$

$$MSE : L(x, y, \vec{\beta}) = \frac{1}{N} \sum_{i=1}^N (f(x, \vec{\beta}) - y_i)^2$$

$$\beta_{best} = \beta \text{ when } \left( \frac{\partial L(x, y, \vec{\beta})}{\partial \vec{\beta}} \right) \bigg|_{x, y} = 0$$

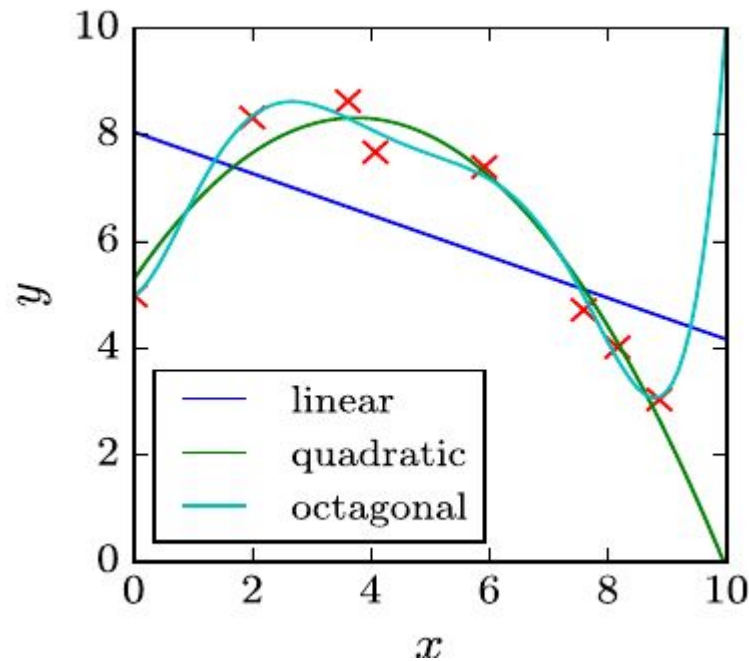


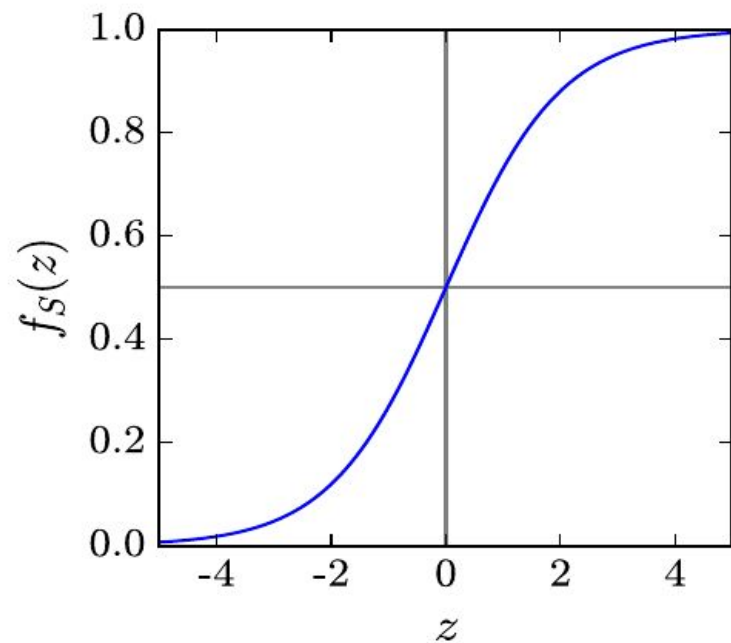


$$f(x, \vec{\beta}) = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$MSE : L(x, y, \vec{\beta}) = \frac{1}{N} \sum_{i=1}^N (f(x, \vec{\beta}) - y_i)^2$$

$$\beta_{best} = \beta \text{ when } \left( \frac{\partial L(x, y, \vec{\beta})}{\partial \vec{\beta}} \right) \bigg|_{x, y} = 0$$







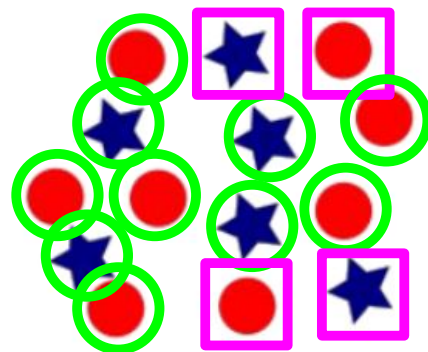
$$f_s(z) = \frac{1}{1+e^{-z}}$$

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$$

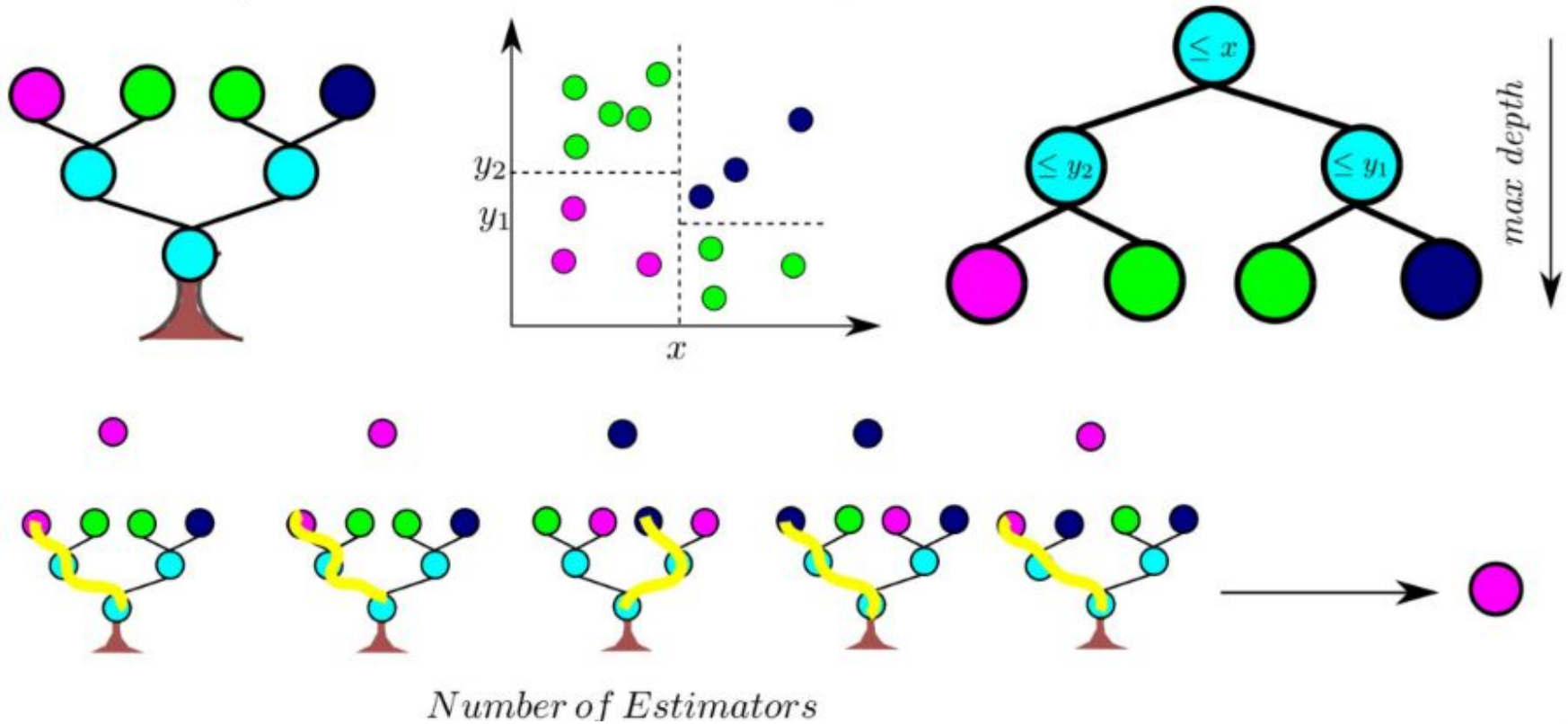
- ❑ Must **train** the model with a fraction of the data (**train data**) and **evaluate** its performance with the remaining fraction (**test data**).
- ❑ To **avoid overfitting** and to **assess the quality** of the fit/prediction.

 *train data*  
(70%)

 *test data*  
(30%)



# Decision Trees & Random Forest



# Multilayer Perceptron

