Week 2 Live Session
Data Science Tools 2

NEBA NFONSANG

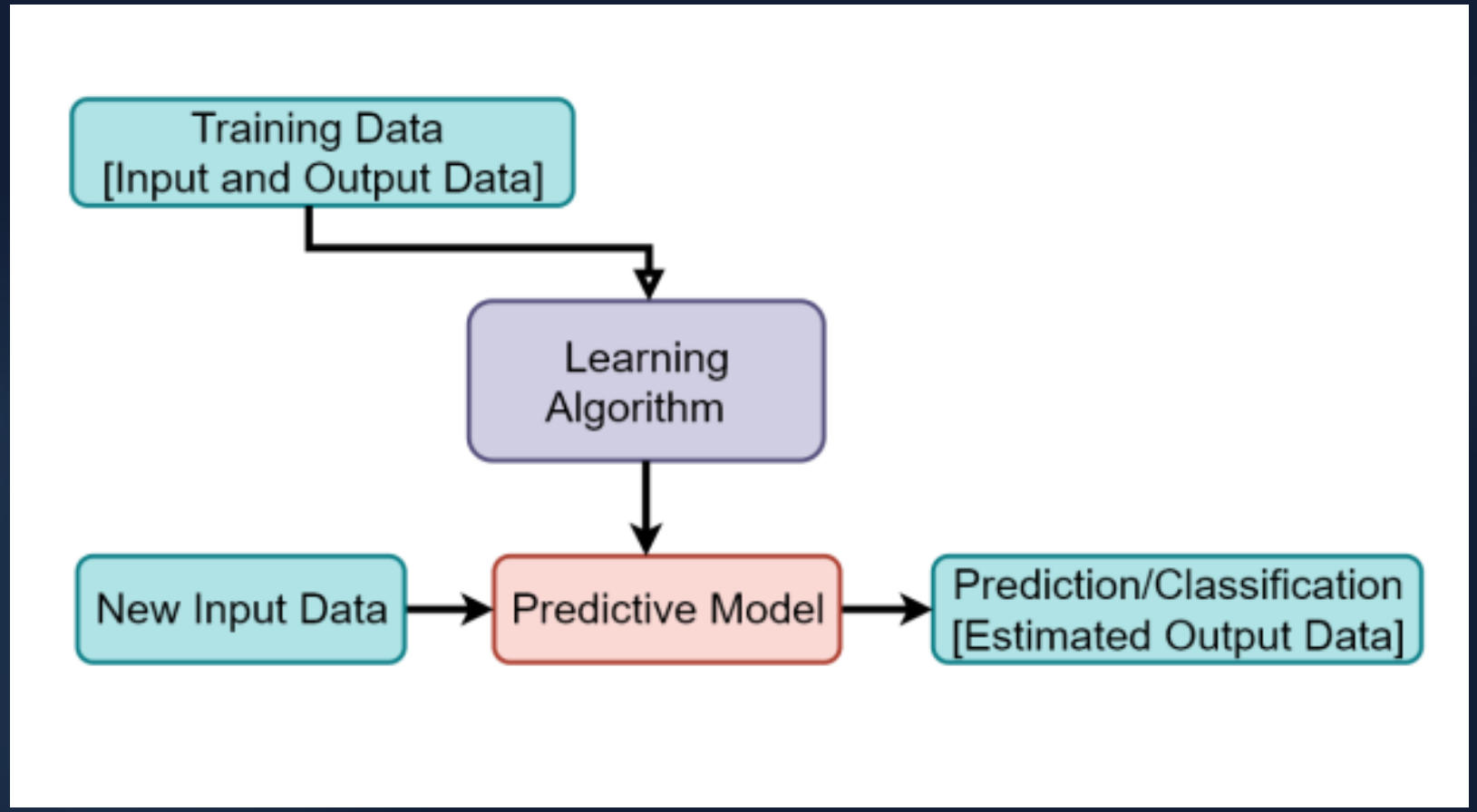
Data Representation

- What is data science?
- What are the components of a data set?
- What is the difference between categorical and numerical data?
- How does a training example in supervised learning look like compared to a training example in unsupervised learning?
- What advantage do we have in using algorithms to solve problems compared to solving problems analytically?

Learning problems

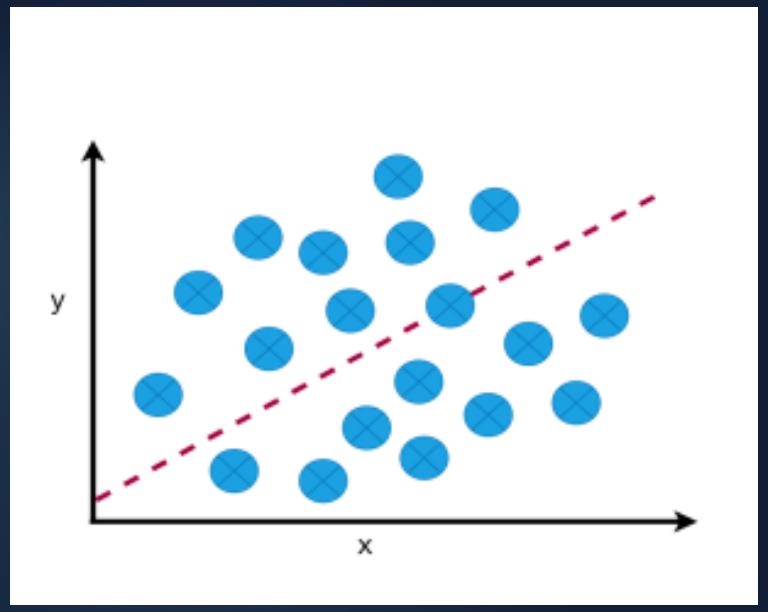
- What are the different types of learning problems in data science or machine learning?

Supervised Learning

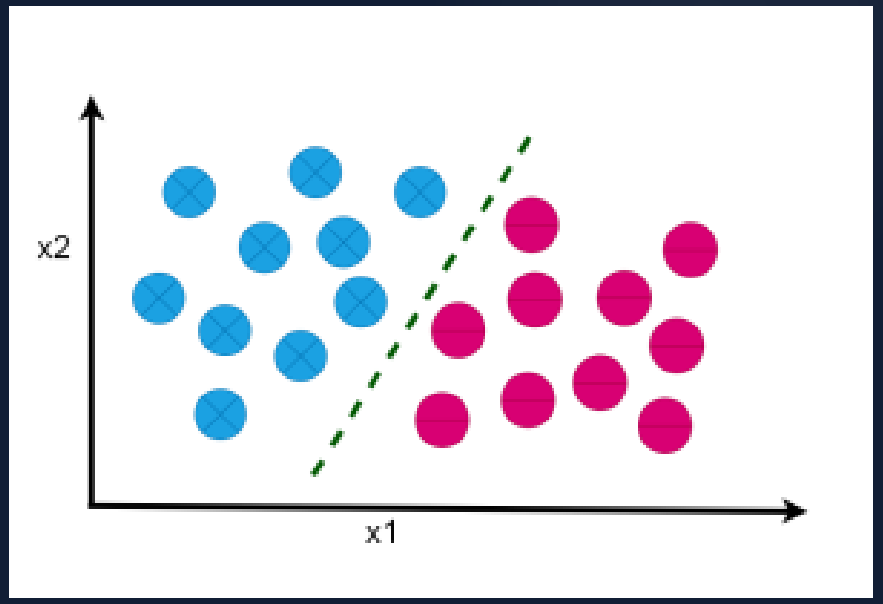


Supervised learning

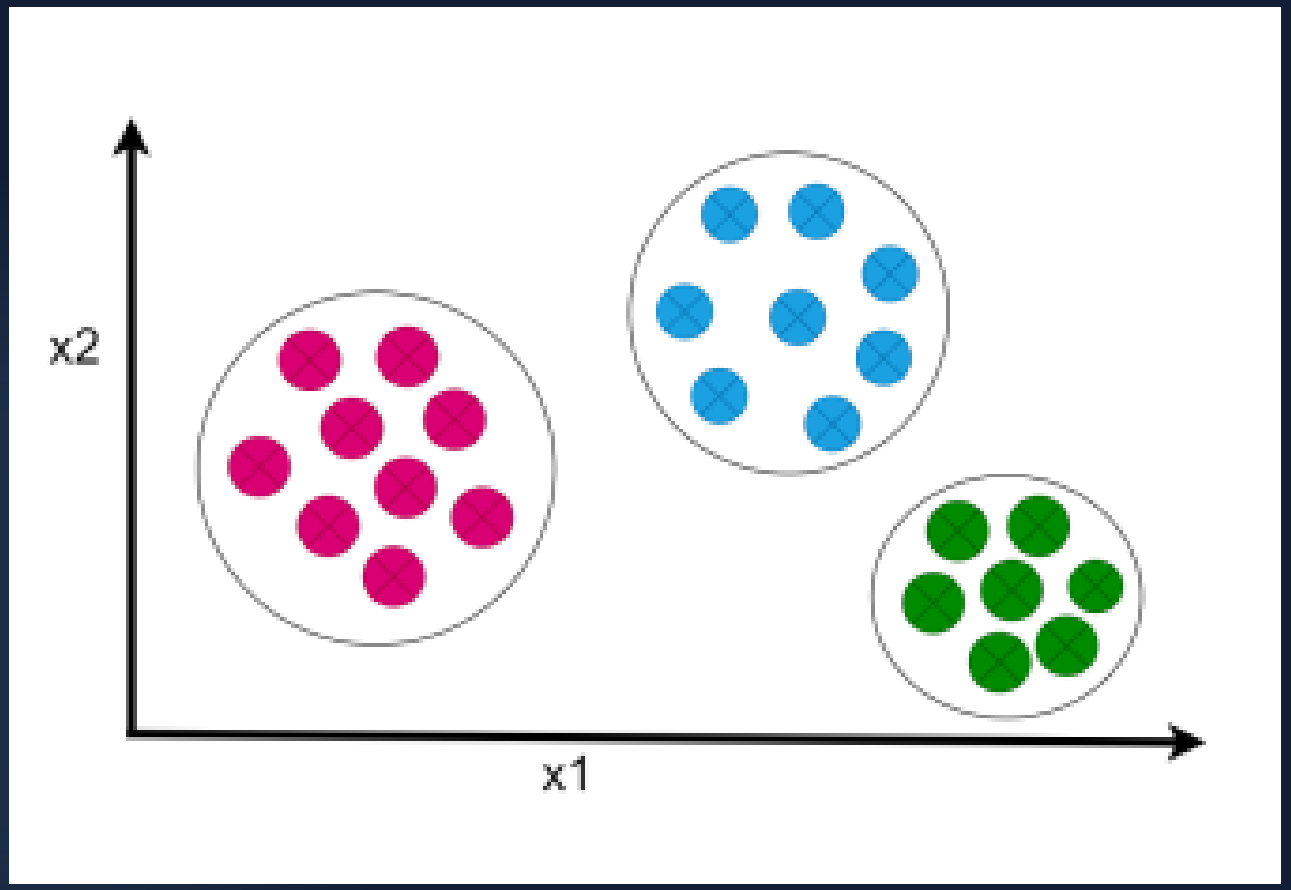
- What type of supervised learning is this?



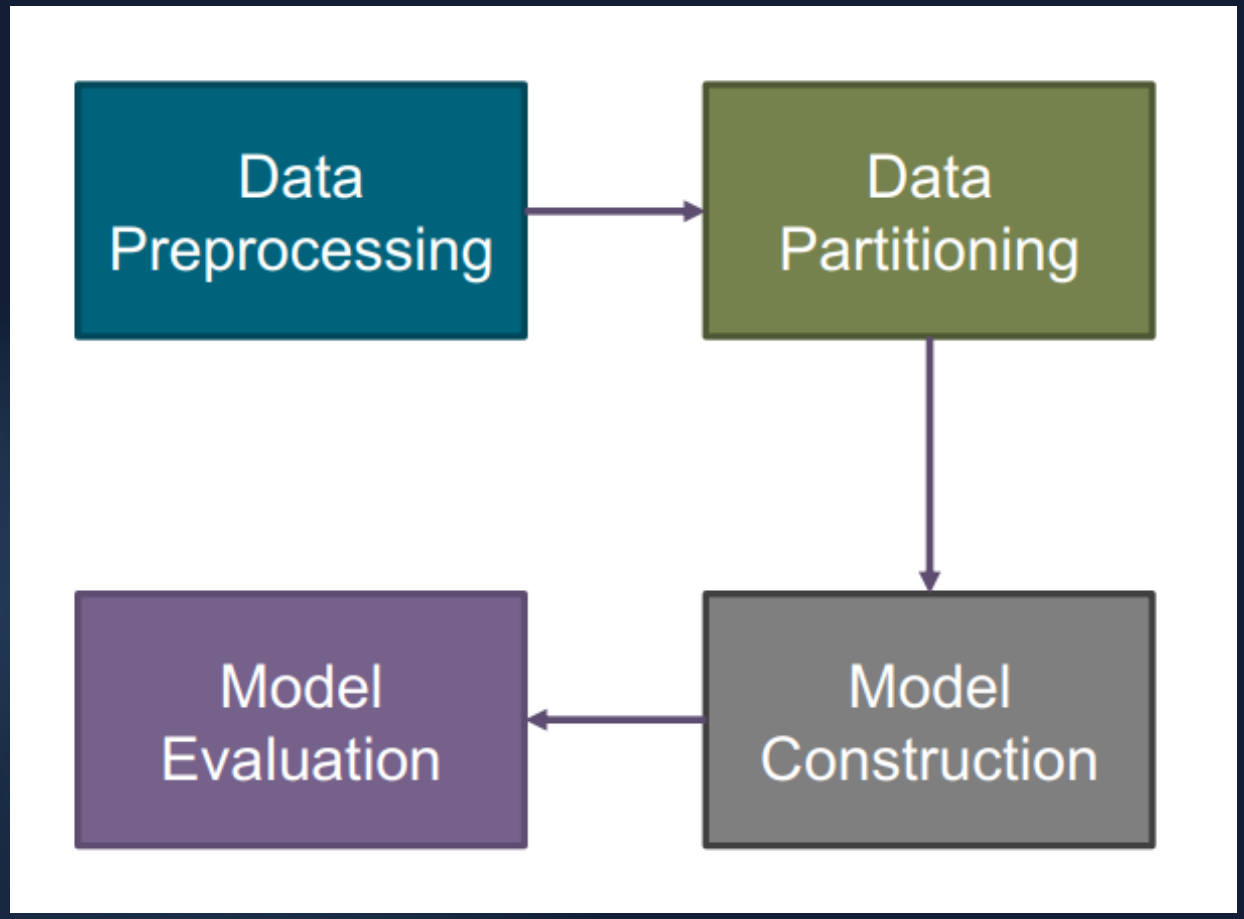
- What type of supervised learning is this?



Unsupervised Learning



Data Science Process



Data Preprocessing

Preprocessing Task	Activities
Data cleaning	Remove noisy data, fix inconsistent data, handle missing data and outliers.
Data transformation	Scale the data to fall within an appropriate range such as 0 to 1 (normalization), transform the data to an appropriate format, etc.
Feature extraction	Data reduction, reduce data size by eliminating redundant or meaningless features
Data exploration	Descriptive statistics and visualization

It used to be that a data scientist spends 80 percent of the time cleaning and preparing the data.

Now, it seems to be that, a data scientist spends 80% of the time worrying about the data.

Dealing with missing data

- How are these missing data mechanism different?
- How do you deal with missing data?
- When do we drop cases?
- When should we impute with mean, median or mode?
- When should we impute using regression?

- **MCAR:** Missing Completely at Radom
- **MAR:** Missing at Random
- **NMAR**: Not Missing at Random

Outliers

- How do outliers affect the results of data analysis?
- What methods can we use to check for outliers?
- How should we handle outliers?

Handling Outliers

- The interquartile method
  - Filter values below Q1 –1.5*IQR and values above Q3 + 1.5*IQR as outliers.
- The mean and standard deviation method
  - Filter values that are 3 standard deviations away from the mean as outliers

Data transformation – Group activity

- What are different ways of transforming data?
- When should we transform our data?
- Why should we transform our data?
- What is the difference between dummy variables and indicator variables?
- What is the difference between feature extraction and feature selection? Give examples?
- What is the curse of dimensionality? How do too many features affect your model? How would you handle high dimensional data?

Tidy Data

- Tidy data is rectangular in shape, that is have columns, rows, and cells.
- Each column should hold a single variable.
- Each row should hold a single observation
- Each cell should hold a single value.

Tidy Data Problems Requiring Data Preparation

1. Column headers are values, not variable names.
2. Multiple variables are stored in one column.
3. Variables are stored in both rows and columns.
4. Multiple types of observational units are in the same table.
5. A single observational unit is in multiple tables.

Read this paper: Tidy Data by Hadley Wickham
https://vita.had.co.nz/papers/tidy-data.pdf