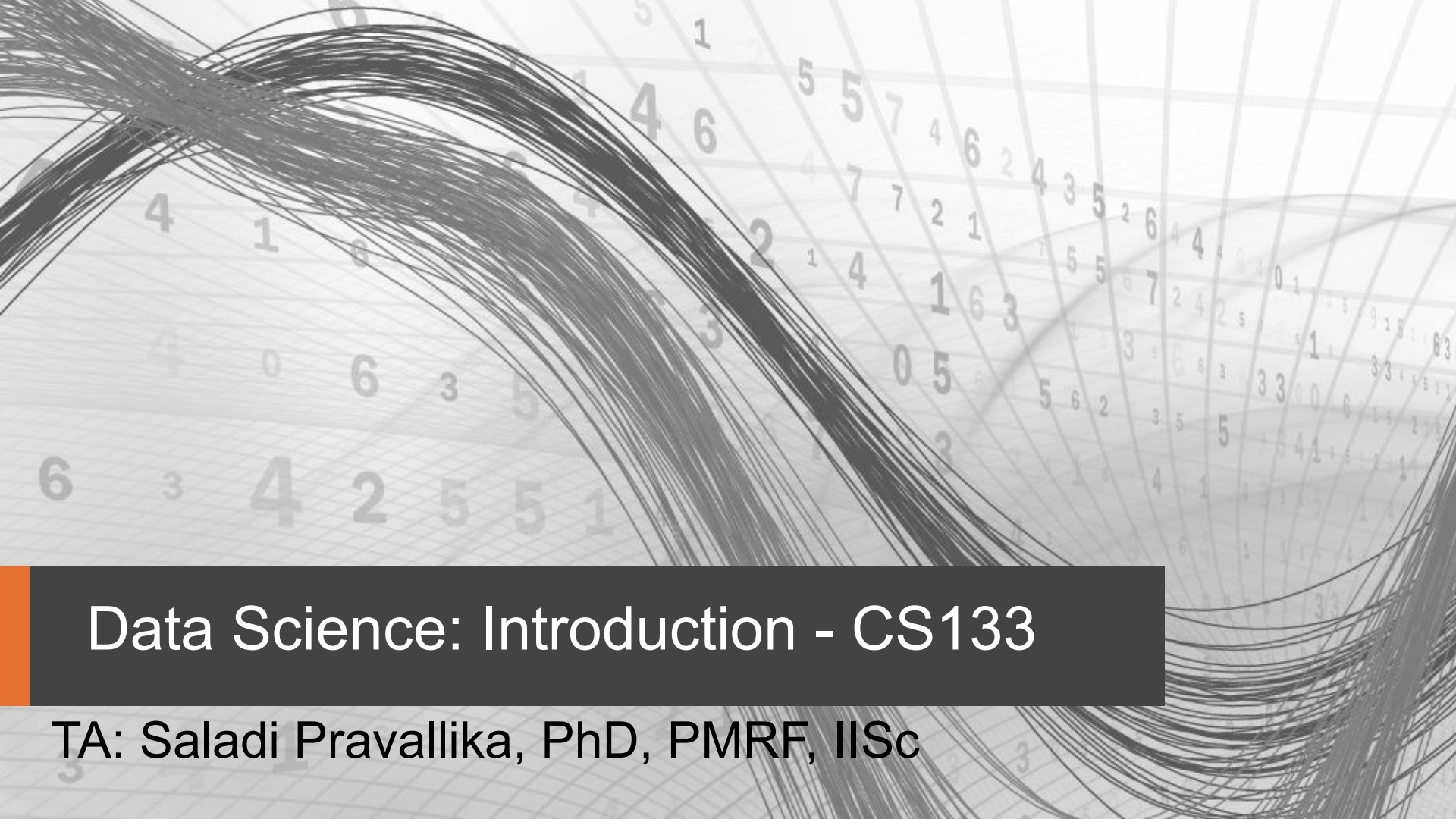
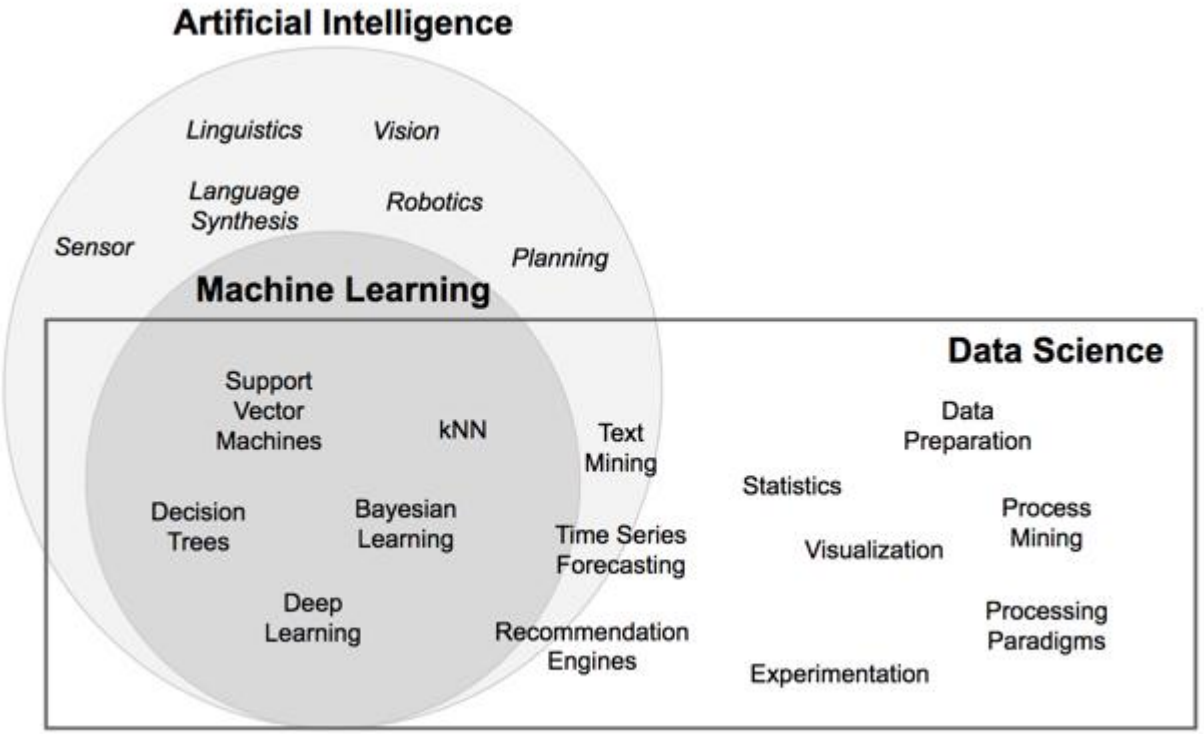


Data Science: Introduction - CS133

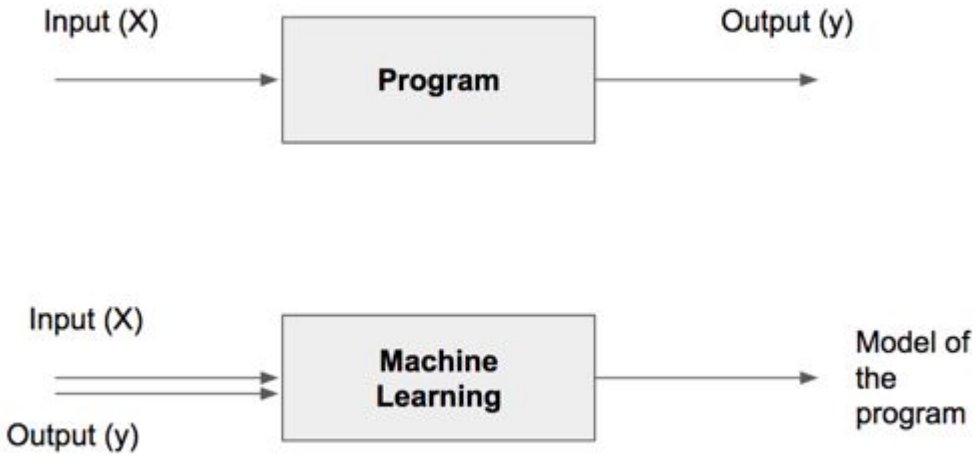
TA: Saladi Pravallika, PhD, PMRF, IISc

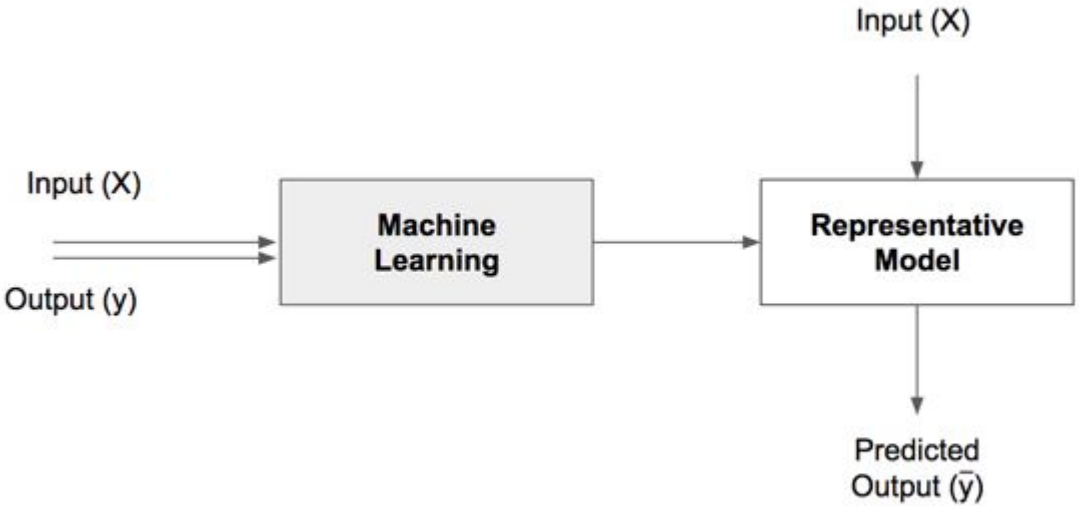
1. Introduction

What is Data Science

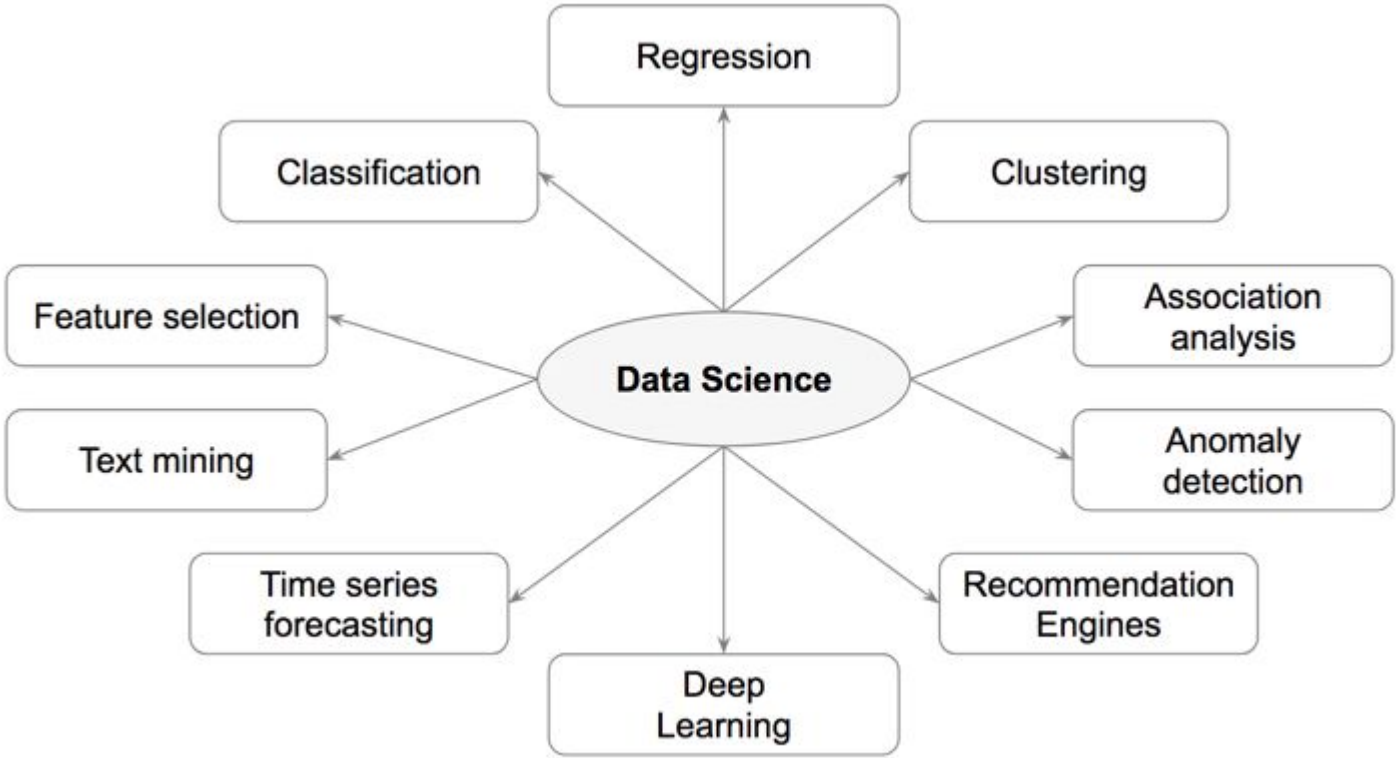


Models


Input (X)
Program
Output (y)
Input (X)
Output (y)
Machine Learning
Model of the program


Input (X)
Input (X)
Output (y)
Machine Learning
Representative Model
Predicted Output (ȳ)

Types of Data Science



Tasks	Description	Algorithms	Examples
Classification	Predict if a data point belongs to one of predefined classes. The prediction will be based on learning from known data set.	Decision Trees, Neural networks, Bayesian models, Induction rules, K nearest neighbors	classifying votes among 3 parties.
Regression	Predict the numeric target label of a data point. The prediction will be based on learning from known data set.	Linear regression, Logistic regression	Predicting unemployment rate for next year. Estimating insurance premium.
Anomaly detection	Predict if a data point is an outlier compared to other data points in the data set.	Distance based, Density based, LOF	Fraud transaction detection in credit cards. Network intrusion detection.
Time series	Predict if the value of the target variable for future time frame based on history values.	Exponential smoothing, ARIMA, regression	Sales forecasting, production forecasting, virtually any growth phenomenon that needs to be extrapolated
Clustering	Identify natural clusters within the data set based on inherit properties within the data set.	K means, density based clustering - DBSCAN	Finding customer segments in a company based on transaction, web and customer call data.
Association analysis	Identify relationships within an itemset based on transaction data.	FP Growth, Apriori	Find cross selling opportunities for a retailor based on transaction purchase history.

Process Basics

Data Science Process

Data Exploration

Model Evaluation

Core Algorithms

Classification

Decision Trees

Rule Induction

k-Nearest Neighbors

Naïve Bayesian

Artificial Neural Networks

Support Vector Machines

Ensemble Learners

Regression

Linear Regression

Logistic Regression

Association Analysis

Apriori

FP-Growth

Clustering

k-Means

DBSCAN

Self-Organizing Maps

Common Applications

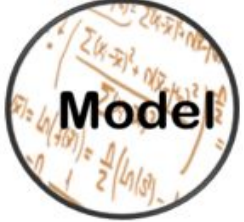
Text Mining

Time Series Forecasting

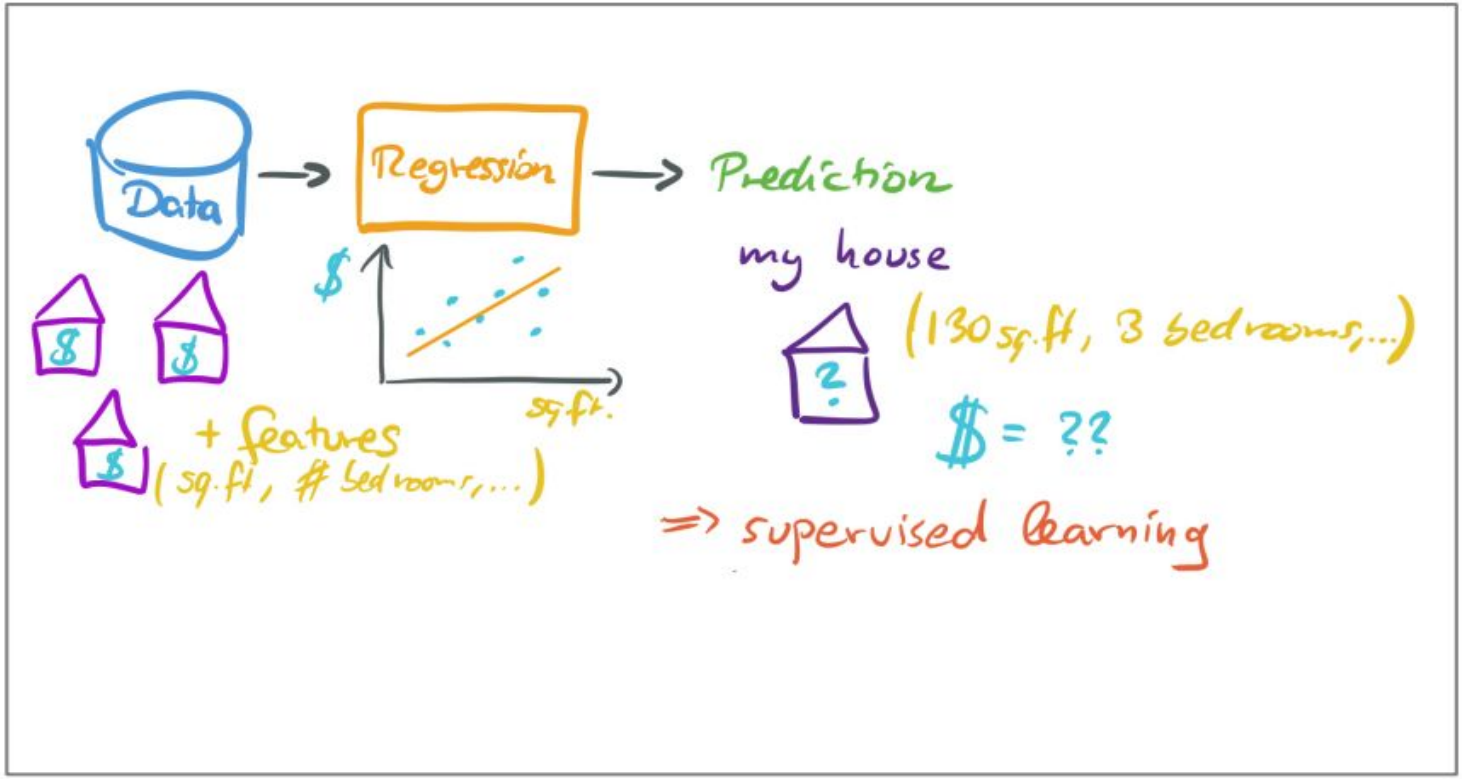
Anomaly Detection

Feature Selection

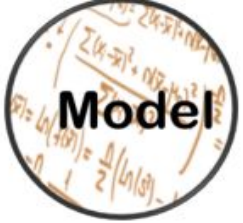
LEARNING FROM DATA



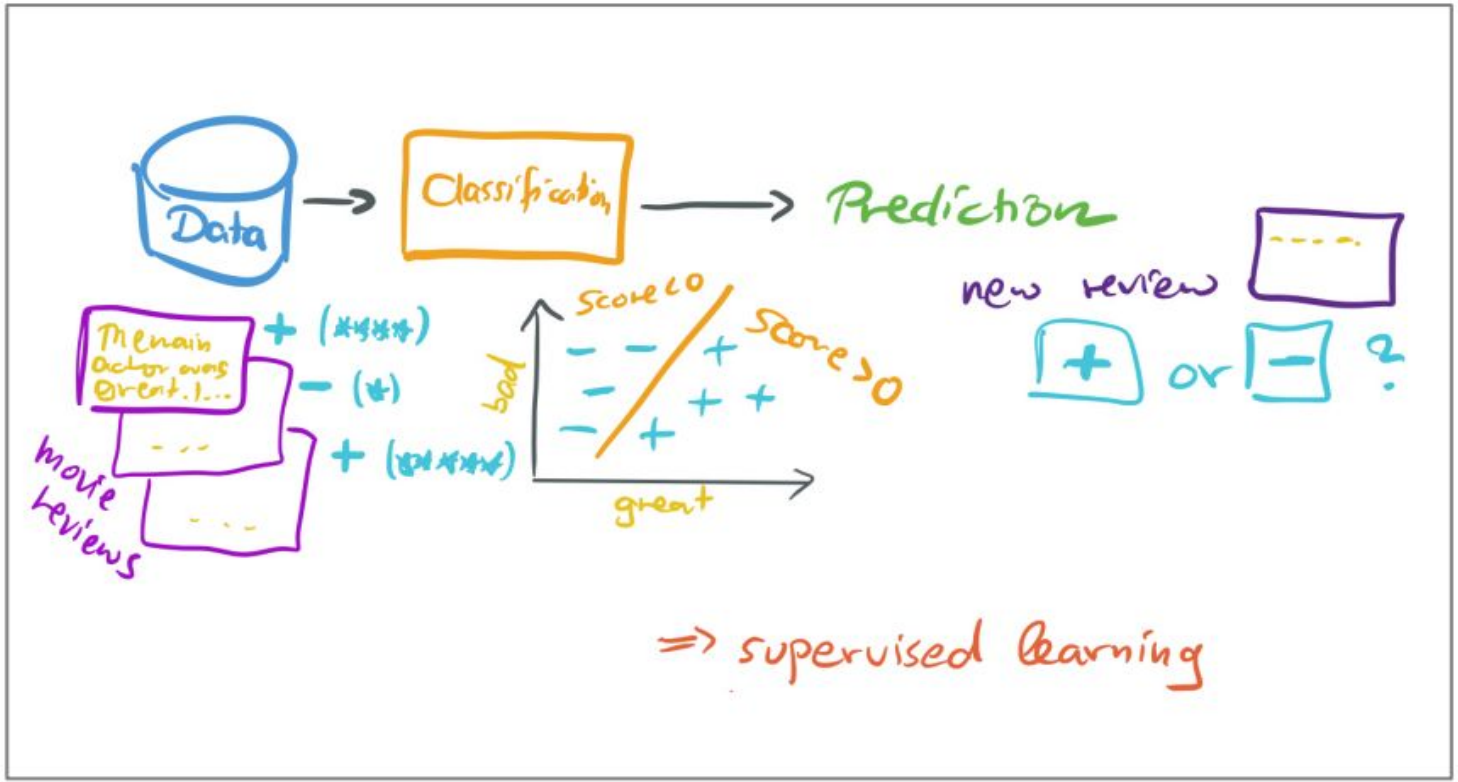
- Regression



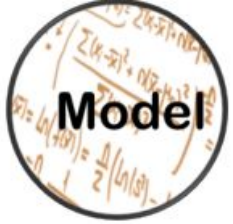
LEARNING FROM DATA



- Classification



LEARNING FROM DATA



- Clustering