

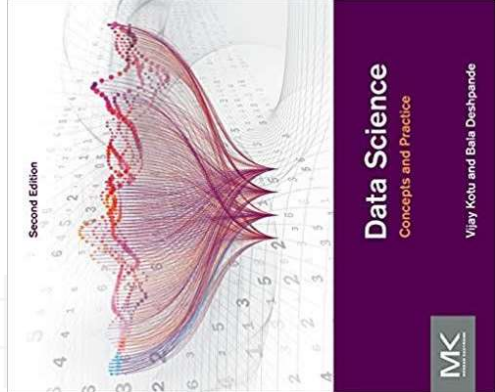
The background features a light gray grid with various numbers (0-9) scattered across it. Overlaid on this grid are numerous thin, black, curved lines that flow from the top left towards the bottom right, creating a sense of motion and data flow.

# Data Science: Concepts and Practice

**Course slides**

# Course Book

# Course Software



## Data Science: Concepts and Practice

Authors : Vijay Kotu & Bala Deshpande  
Publisher : Morgan Kaufmann



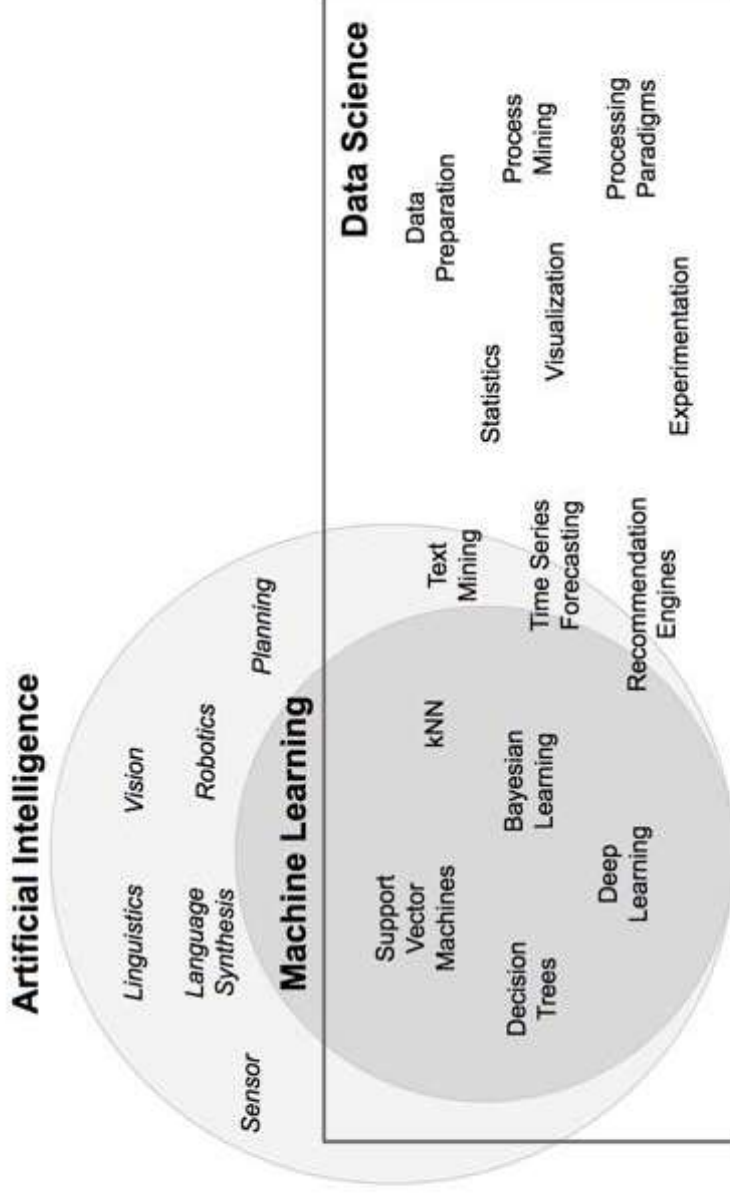
[www.rapidminer.com](http://www.rapidminer.com)

**Free Download**

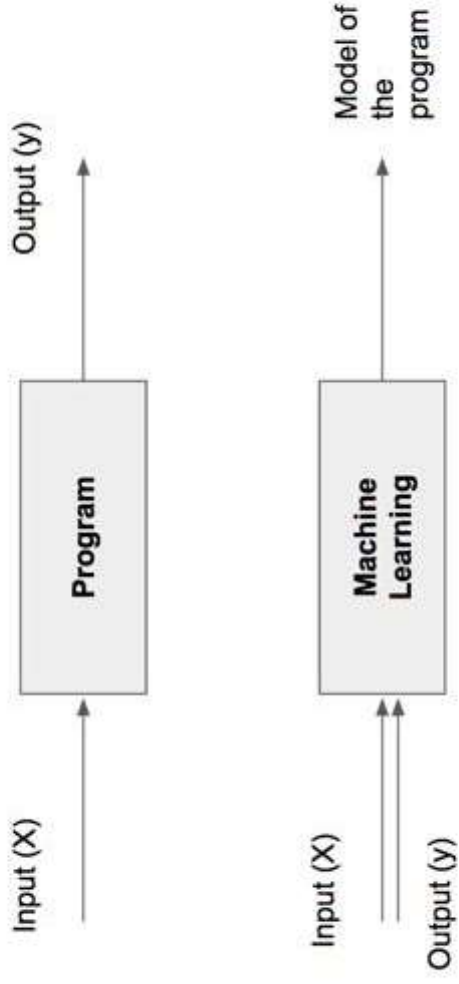
The background of the slide is a complex, abstract composition. It features a grid of thin, light gray lines that intersect to form a pattern of small squares. Overlaid on this grid are numerous numbers, including 0, 1, 2, 3, 4, 5, 6, 7, and 8, in various sizes and orientations. Some numbers are bold and black, while others are lighter and more faded. A series of thick, dark, curved lines sweep across the lower half of the image, creating a sense of motion and depth. A solid black horizontal bar is positioned in the upper right, containing the text '1. Introduction' in white. The bottom left corner of the slide has a small orange square.

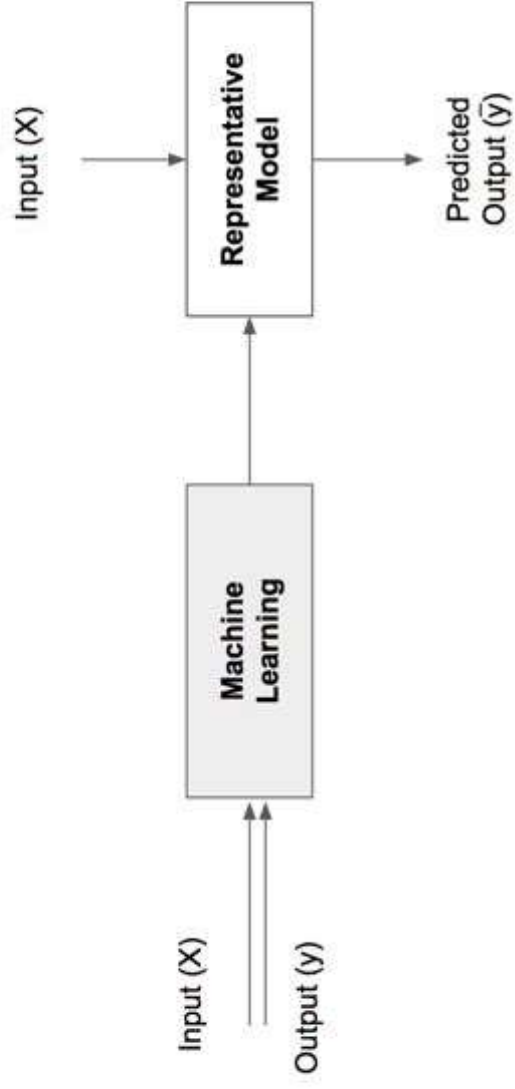
# 1. Introduction

# What is Data Science

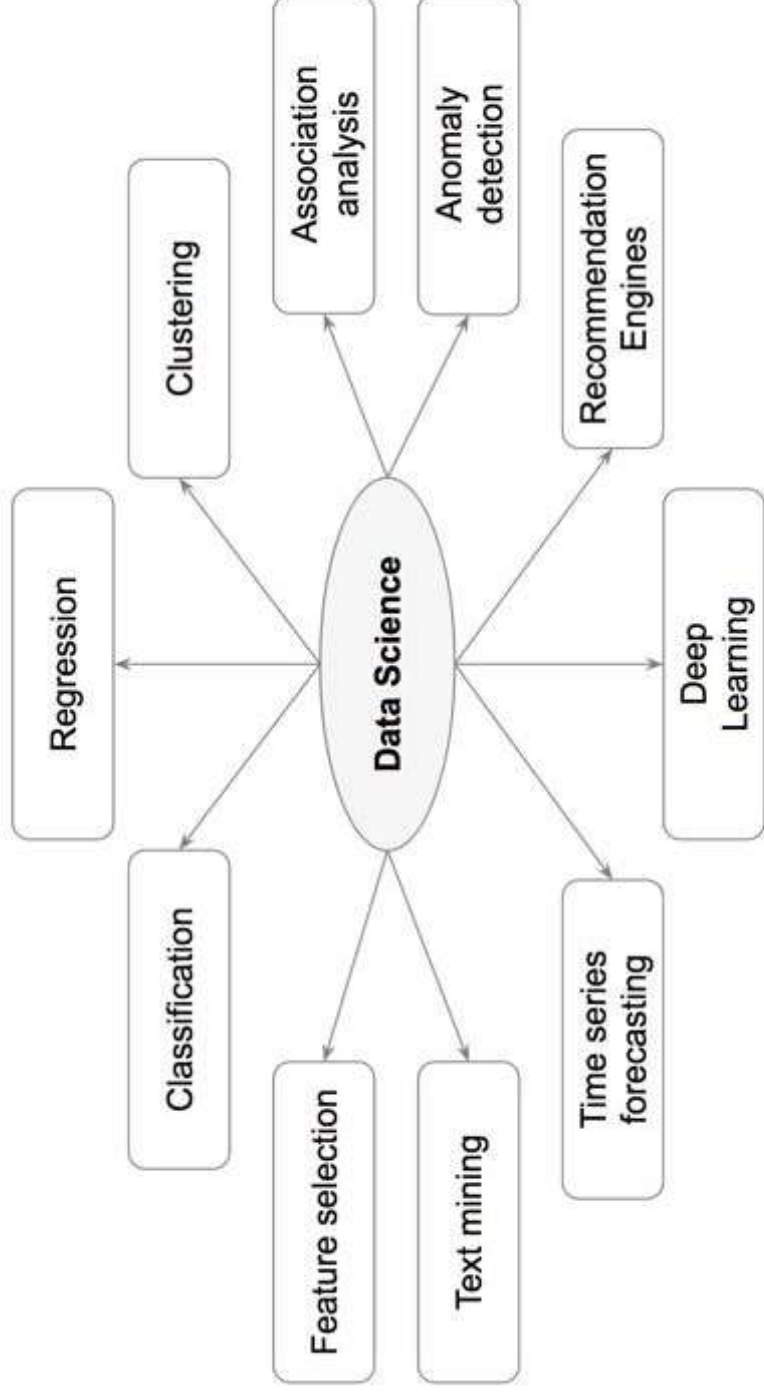


# Models





# 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	Assigning voters into known buckets by political parties eg: soccer moms. Bucketing new customers into one of known customer groups.
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 inherent 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 retailer based on transaction purchase history.



# Course outline

## 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

## Process Basics

**Data Science**

**Process**

**Data Exploration**

**Model Evaluation**

## Common Applications

**Text Mining**

**Time Series Forecasting**

**Anomaly Detection**

**Feature Selection**