KCA University
Nairobi, Kenya

BSD 3101: PRINCIPLES OF DATA SCIENCE

Prepared By:

Linus Aloo

E-mail: linusaloo88@gmail.com

Phone: 0754188380

Recommended Reading

1. Sinan Ozdemir, "Principles of Data Science," Packt Publishing,1st Ed. 2016.
2. Vijay Kotu and Bala Deshpande, "Data Science: Concepts and Practice," Elsevier, Second Edition, 2019.
3. Gareth James, Daniela Witten, Trevor Hastie & Robert Tibshirani, "An Introduction to Statistical Learning (with Applications in R), Springer (1st Ed.), 2014.

Additional Reading

1. Joel Grus, "Data Science from Scratch," O'Reilly Media, 1st Ed., 2015.

CHAPTER ONE: OVERVIEW AND INTRODUCTION

1.0 Course Outline

Pre-requisite: Probability and Statistics, Database and design

Purpose/Aim

- This course aims at introducing students to applications of mathematical theory in data science.

Course Objectives

1. To understand tools and techniques used both to describe and to gain insights into the properties of often large and complex datasets.

2.To demonstrate the use of modern software packages including R for both statistical computation and graphical visualization of statistical properties and results.

CHAPTER ONE: OVERVIEW AND INTRODUCTION

1.0 Course Outline

Course Content

- **Overview and Introduction**: Definitions, career opportunities, Data science tasks and techniques, data science process.
- **Acquiring and Storing data**: Sources and Methods,
- **Data Wrangling and pre-processing**: data collection, cleaning, integration and reduction, data characterization & presentation.
- **Statistical Learning**: Introduction, Estimation, models and prediction accuracy, Notions of supervised and unsupervised learning.
- **Linear Regression**: Simple linear regression, Multiple linear regression, Case studies

1.0 Course Outline

Course Content Continued

- **Classification**: Logistic regression, Linear discriminant analysis, applications
- **Resampling Methods**: Cross-validation, The bootstrap method, Applications
- **Linear Model Selection and Regularization**: Approaches to subset selection Shrinkage methods, Methods for dimension reduction, applications
- **Data visualization**: Techniques and case studies of data visualization

Learning & Teaching Methodologies

Lectures, tutorials and planning exercises.

CHAPTER ONE: OVERVIEW AND INTRODUCTION

1.0 Course Outline

Instructional Materials/Equipment

- Classroom with audio visual aids
- Computer laboratory

Recommended Reading

1. Gareth James, Daniela Witten, Trevor Hastie & Robert Tibshirani, "An Introduction to Statistical Learning (with Applications in R), Springer (1st Ed.), 2014.
2. Sinan Ozdemir, "Principles of Data Science," Packt Publishing,1st Ed. 2016.
3. Vijay Kotu and Bala Deshpande, "Data Science: Concepts and Practice," Elsevier, Second Edition, 2019

Course Assessment

BSD 3101	PRINCIPLES OF DATA SCIENCE	
Course Assessment	Type	Weighting (%)
	Examination	50
	Continuous Assessment	50
	Total	100

Additional Reading

1. Joel Grus, "Data Science from Scratch," O'Reilly Media, 1st Ed., 2015.
2. Cathy O'Neil, Rachel Schutt, "Doing Data Science: Straight Talk from the Frontline," O'Reilly Media

1.2. Introduction

1.2.1. What is Data Science?

- Data science is a collection of techniques used to extract value from data and acquire knowledge.
- It has become an essential tool for any organization that collects, stores, and processes data as part of its operations.
- Data science techniques rely on finding useful patterns, connections, and relationships within data.
- Data science is also commonly referred to as knowledge discovery, machine learning, predictive analytics, and data mining.
- Data science is all about how we take data, use it to acquire knowledge, and then use that knowledge to do the following:
  - Make decisions
  - Understand the past/present
  - Predict the future
  - Create new industries/product

1.2. Introduction

1.2. AI, Machine Learning, and Data Science

Fig. 1.1 shows the relationship between artificial intelligence, machine learning, and data science.

- ***Artificial intelligence*** is about giving machines the capability of mimicking human behavior, particularly cognitive functions. Examples would be: **facial recognition, automated driving, sorting mail based on postal code**.
- There are quite a range of techniques that fall under artificial intelligence: ***linguistics, natural language processing, decision science, bias, vision, robotics, planning***, etc.
- ***Machine learning*** can either be considered a sub-field or one of the tools of artificial intelligence that is providing machines with the capability of learning from experience.
- Experience for machines comes in the form of data.
- Data that is used to teach machines is called ***training data***.

1.2. Introduction

1.2. AI, Machine Learning, And Data Science

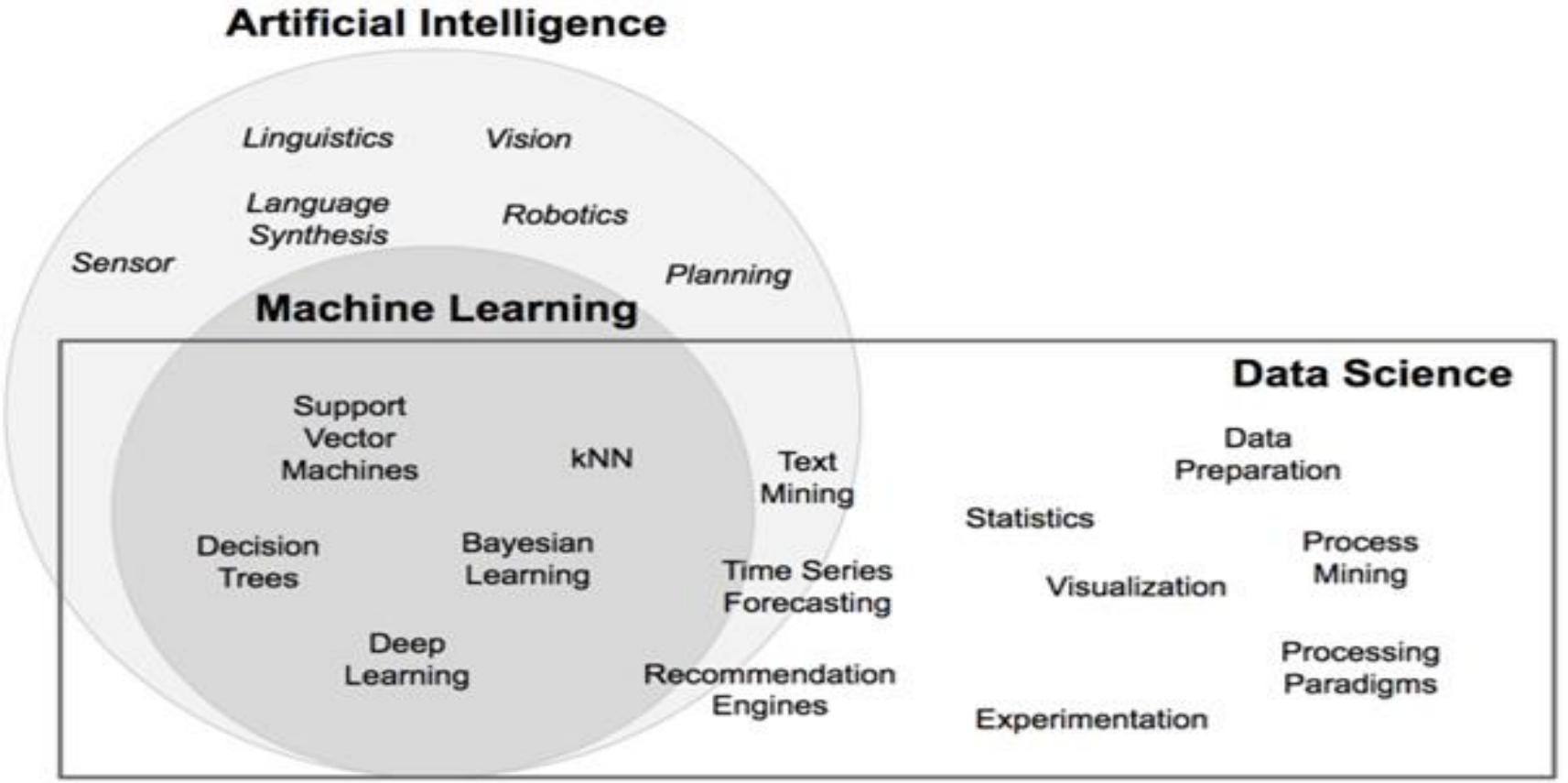

Fig. 1.1 The relationship between artificial intelligence, machine learning, and data science

1.3. Models

1.3.1 Model of the program

- Machine learning turns the traditional programing model upside down (Fig. 1.2).
- A program, a set of instructions to a computer, transforms input signals into output signals using predetermined rules and relationships.
- Machine learning algorithms, also called "learners", take both the known input and output (training data) to figure out a model for the program, which converts input to output.

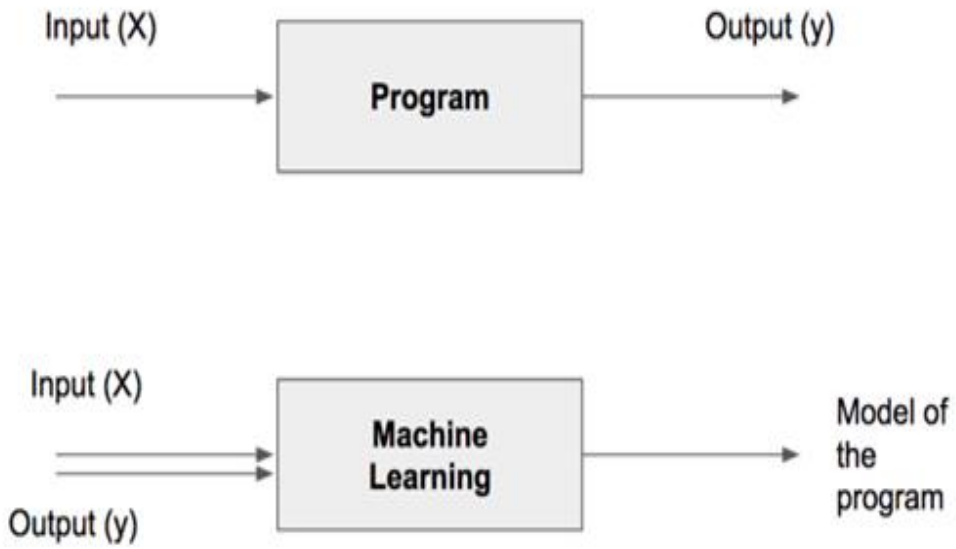

Fig. 1.2

- For example, many organizations like social media platforms, review sites, or forums are required to moderate posts and remove abusive content.

1.3.2. Building Representative Models

- In statistics, a model is the representation of a relationship between variables in a dataset.
- It describes how one or more variables in the data are related to other variables.
- Modeling is a process in which a representative abstraction is built from the observed dataset.
- For example, based on credit score, income level, and requested loan amount, a model can be developed to determine the interest rate of a loan**.**
- Fig. 1.3 shows the process of generating a model.

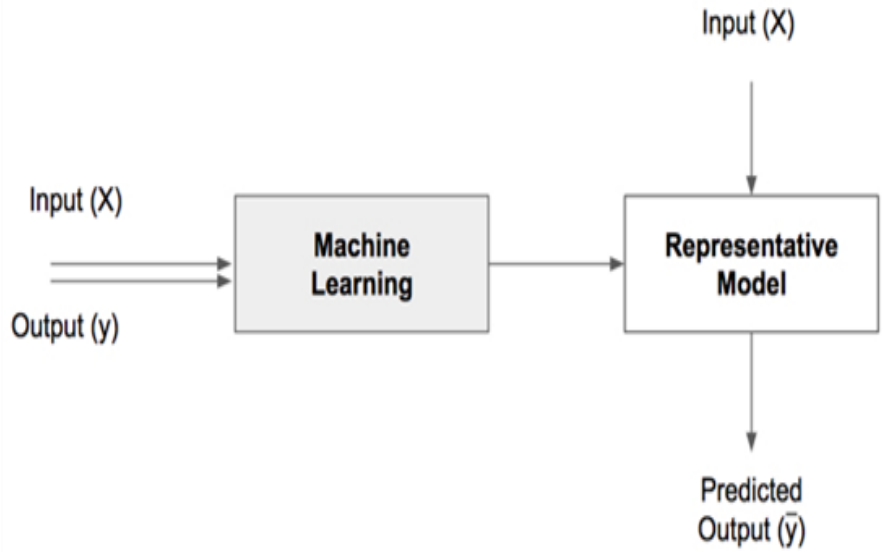

FIGURE 1.3 Data science models.

1.3.2. Building Representative Models

- Once the representative model is created, it can be used to predict the value of the interest rate, based on all the input variables.
- Data science is the process of building a representative model that fits the observational data.
- This model serves two purposes: on the one hand, it ***predicts the output*** (interest rate) based on the new and unseen set of input variables (credit score, income level, and loan amount), and on the other hand, the model can be used to understand the **relationship between the output variable and all the input variables**.

1.3.2. Building Representative Models

- **Probabilistic model**: This refers to using probability to find a relationship between elements that includes a degree of randomness.
- **Statistical model**: This refers to taking advantage of statistical theorems to formalize relationships between data elements in a (usually) simple mathematical formula.

Study Questions:

1. Explain the difference between organized data and unorganized data as used in data science.

Ans.

• **Organized data**: This refers to data that is sorted into a row/column structure, where every row represents a single *observation* and the columns represent the *characteristics* of that observation.

• **Unorganized data**: This is the type of data that is in the free form, usually text or raw audio/signals that must be parsed further to become organized.

Study Questions Continued

2. Distinguish between **Exploratory data analysis** (**EDA**) and **Data mining**

Ans.

- **Exploratory data analysis (EDA)** refers to preparing data in order to standardize results and gain quick insights. EDA is concerned with data visualization and preparation. This is where we turn unorganized data into organized data and also clean up missing/incorrect data points.
- **Data mining** is the process of finding relationships between elements of data. Data mining is the part of data science where we try to find relationships between variables.

1.3.3. The data science Venn diagram

- Understanding data science begins with three basic areas:
  - **Math/statistics**: This is the use of equations and formulas to perform analysis
  - **Computer programming**: This is the ability to use code to create outcomes on the computer
  - **Domain knowledge**: This refers to understanding the problem domain (medicine, finance, social science, and so on).
- The Venn diagram shown in Fig.1.4 provides a visual representation of how the three areas of data science intersect.

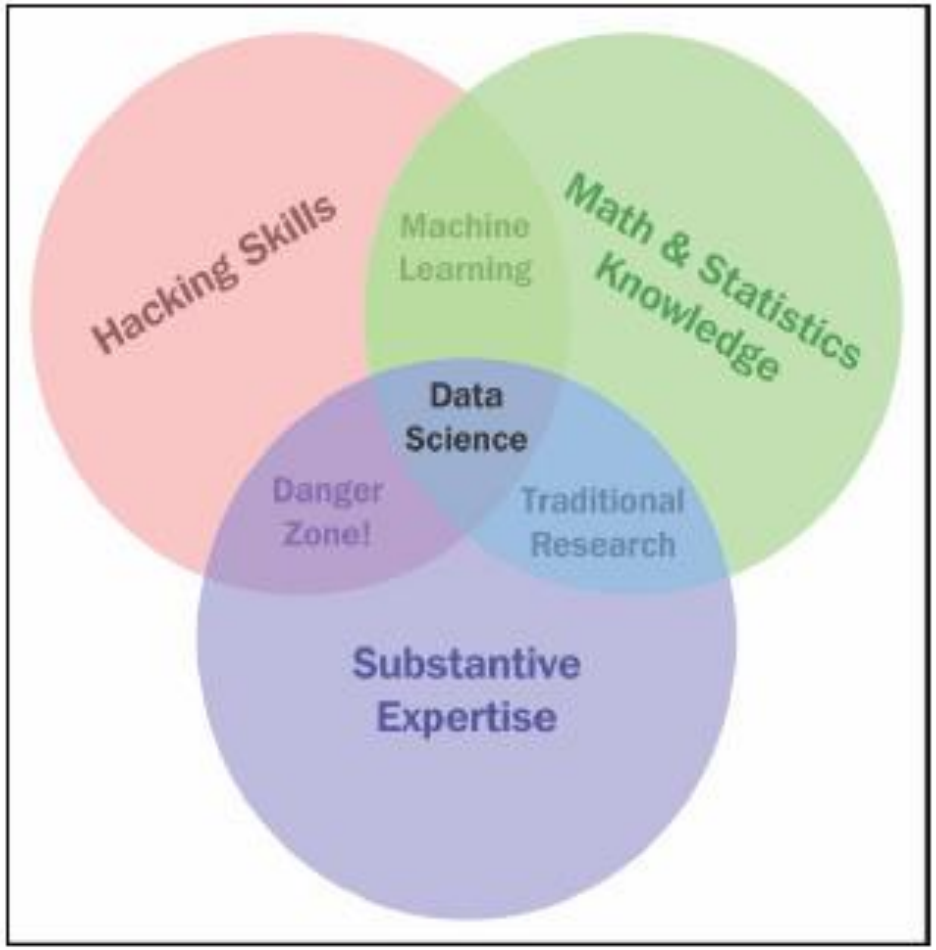

Fig.1.4. The Venn diagram of data science

1.4 Data Science Algorithms

- An algorithm is a logical step-by-step procedure for solving a problem.
- In data science, it is the blueprint for how a particular data problem is solved.
- Many of the learning algorithms are recursive, where a set of steps are repeated many times until a limiting condition is met.
- Some algorithms also contain a random variable as an input and are aptly called randomized algorithms.
- A classification task can be solved using many different learning algorithms such as decision trees, artificial neural networks, k-NN, and even some regression algorithms.
- The choice of which algorithm to use depends on the ***type of dataset, objective, structure of the data, presence of outliers, available computational power, number of records, number of attributes***, and so on.

1.4 Data Science Algorithms

- Data science algorithms can be implemented by custom-developed computer programs in almost any computer language.
- This obviously is a time consuming task. In order to focus the appropriate amount of time on data and algorithms, data science tools or statistical programing tools, like R, RapidMiner, Python, SAS Enterprise Miner, etc., which can implement these algorithms with ease, can be leveraged.
- These data science tools offer a library of algorithms as functions, which can be interfaced through programming code or configured through graphical user interfaces.
- Table 1.1 provides a summary of data science tasks with commonly used algorithmic techniques and example cases.

1.5. Types of Data Science

- Data science problems can be broadly categorized into **supervised or unsupervised learning models**.
- Supervised or directed data science tries to infer a function or relationship based on labeled training data and uses this function to map new unlabeled data.
- Supervised techniques predict the value of the output variables based on a set of input variables.
- Unsupervised or undirected data science uncovers hidden patterns in unlabeled data.
- In unsupervised data science, there are no output variables to predict.
- The objective of this class of data science techniques is to find patterns in data based on the relationship between data points themselves.

1.5. Types of Data Science

Table 1.1. Key features of each category of data science task.

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.

1.5. Types of Data Science

Table 1.1. Key features of each category of data science task Continued

Tasks	Description	Algorithms	Examples
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.

1.5. Types of Data Science

- Data science problems can also be classified into tasks such as: classification, regression, association analysis, clustering, anomaly detection, recommendation engines, feature selection, time series forecasting, deep learning, and text mining (Fig. 1.5).

Study Question: State and explain FOUR career opportunities in data science.

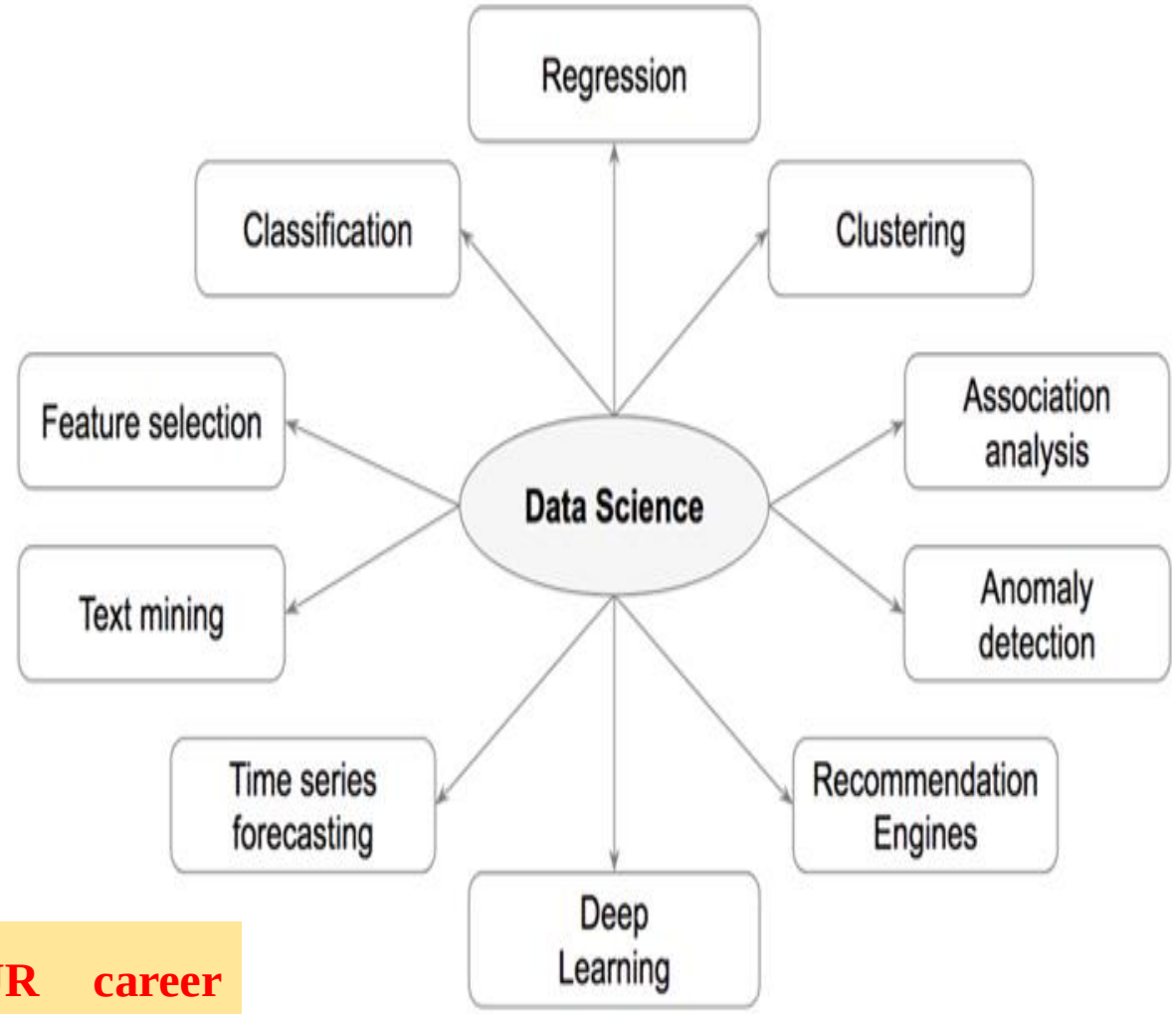

Figure 1.5 Data science tasks.