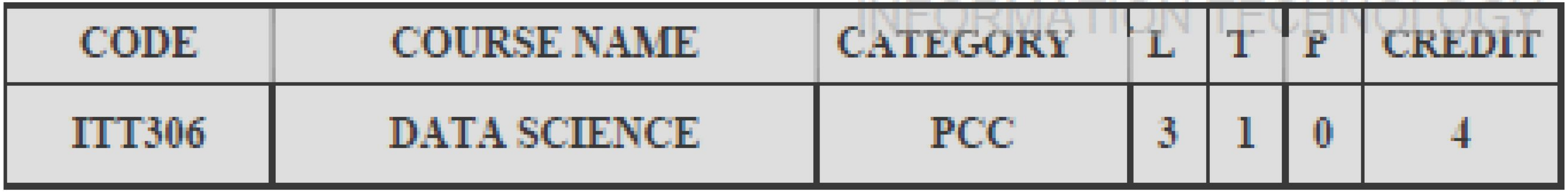
Data Science

ITT306



CODE	COURSE NAME	CATEGORY	L	T	P	CREDIT
ITT306	DATA SCIENCE	PCC	3	1	0	4

- **Preamble**: This course is designed to provide learners with working knowledge of the theoretical background of various aspects of Data Science and enable them to incorporate and apply the principles of statistics and machine learning to solve real-world problems for largescale data analysis.

Prerequisites:

- MAT 101 Linear Algebra and Calculus
- MAT 208 Probability and Statistics and Advanced Graph Theory
- ITT 205 Problem Solving Using Python
- ITT 201 Data Structures
- ITT 206 Database Management Systems

Course Outcomes: After the completion of the course the student will be able to

CO No.	Course Outcome(CO)	Bloom's Category Level
CO 1	Explain the fundamental concepts and various aspects of data science	Level 2: Understand
CO 2	Choose data validationtechniques suitable for statistical analysis andpresent results using data visualization techniques.	Level 2: Understand
CO 3	Identify different statistical learning algorithm for solving a problem	Level 3: Apply
CO 4	Use statistical analysis to characterize and interpret data sets	Level 3: Apply
CO 5	Compare the pros/cons of various models and algorithms used for data analysis and data mining	Level 2: Understand
CO 6	Develop the ability to perform basic data analysis in Python and understand the fundamentals of deep learning.	Level 3: Apply

Module 1: Foundations Data Science, process, and tools

- Introduction to data science
- Properties of data
- Asking interesting questions
- Classification of data science
- Data science process
- Collecting, cleaning and visualizing data
- Languages, and models for data science

MSDS Alum Feature

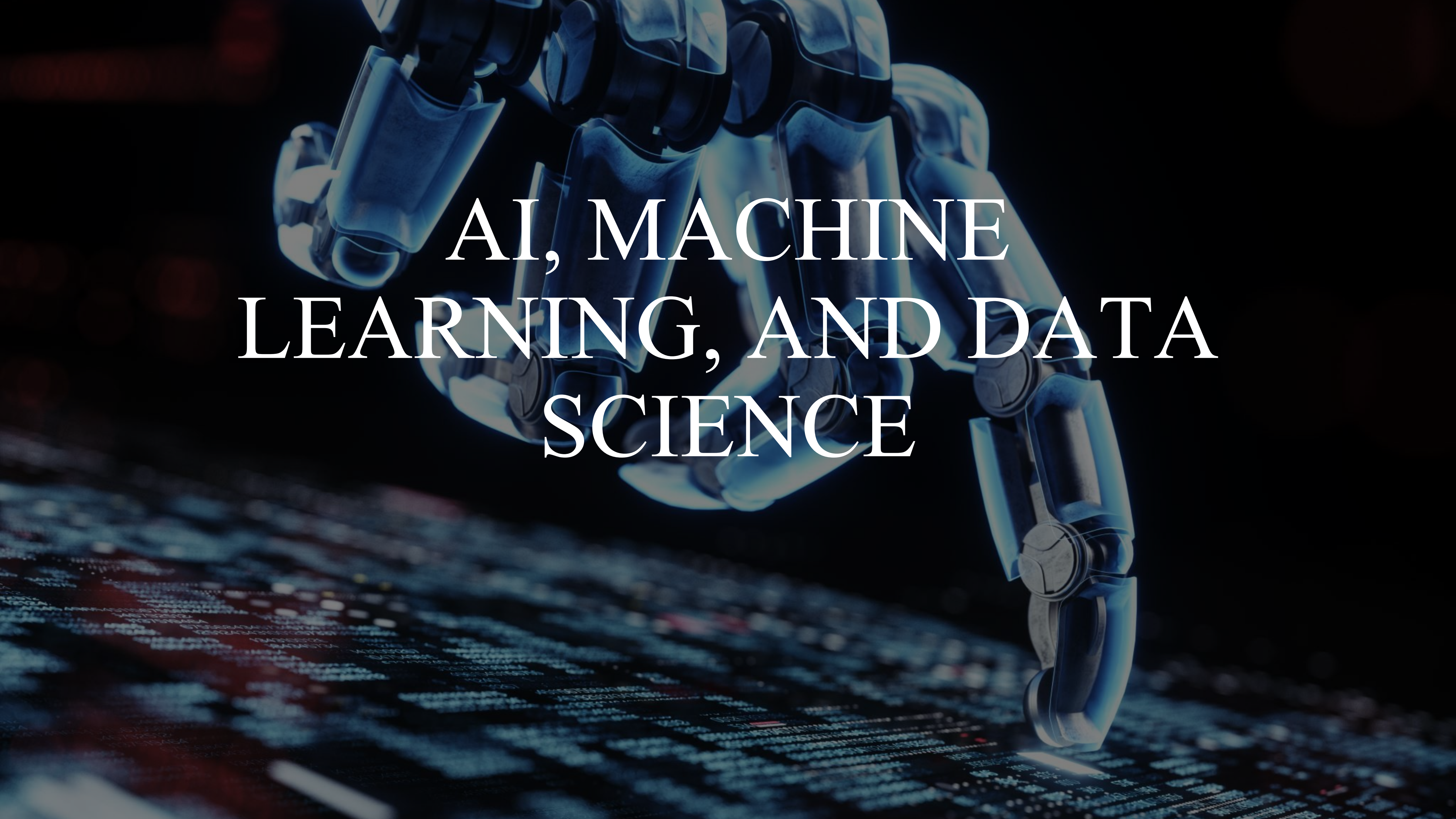
- Sam Garofalo -

Some interesting links

How data transformed the NBA
https://www.youtube.com/watch?v=oUvvfHkXyOA

Using data science for public health
https://www.youtube.com/watch?v=3C7mwm84ndo

How Data is Used in Healthcare
https://www.youtube.com/watch?v=V8TRt-2A0_4



AI, MACHINE LEARNING, AND DATA SCIENCE

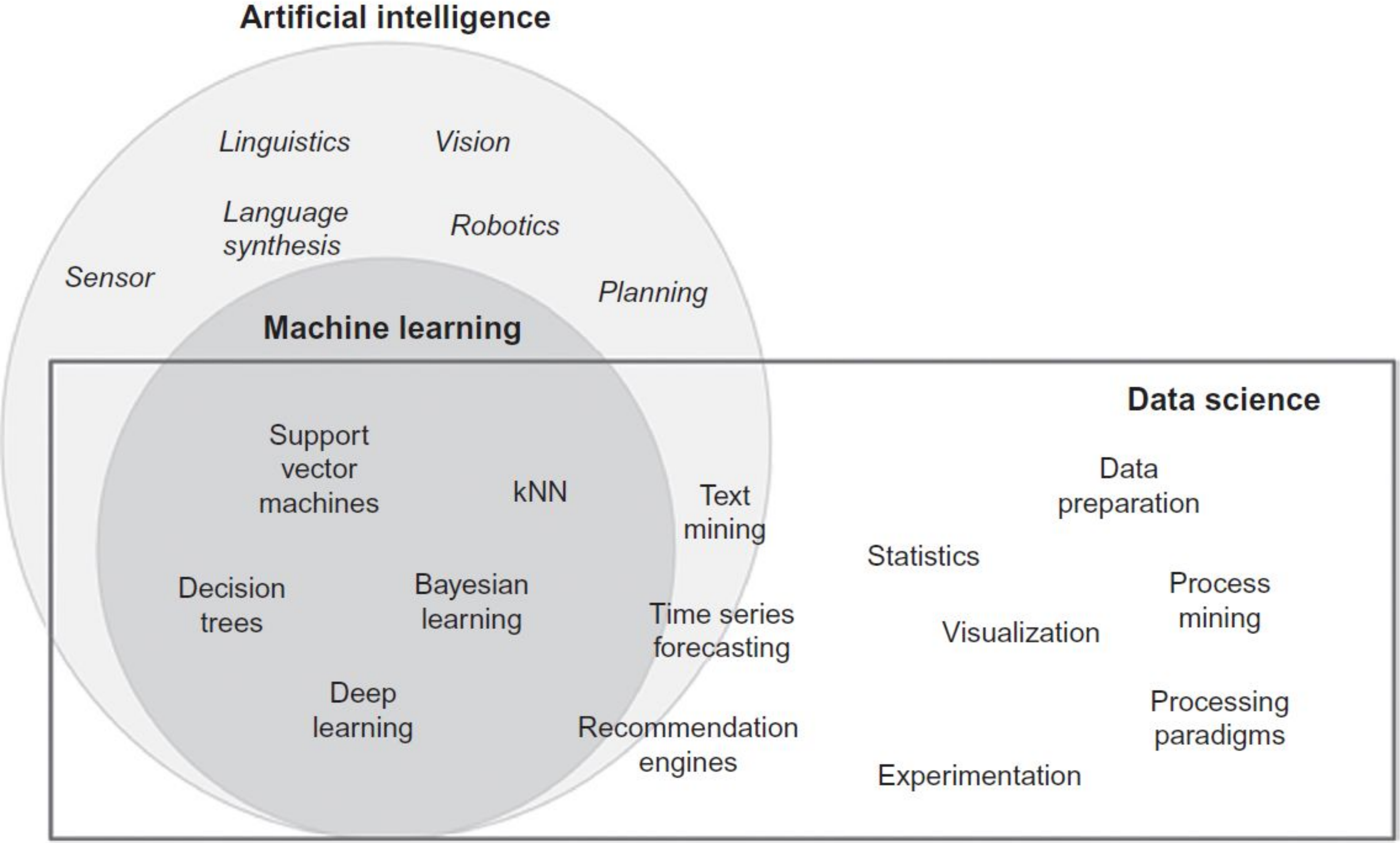


FIGURE 1.1

Artificial intelligence, machine learning, and data science.

AI, MACHINE LEARNING, AND DATA SCIENCE

- Artificial intelligence is about giving machines the capability of mimicking human behavior, particularly cognitive functions.
- 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, is providing machines with the capability of learning from experience.

AI, MACHINE LEARNING, AND DATA SCIENCE

- Experience for machines comes in the form of data.
- Data that is used to teach machines is called **training data**.
- Machine learning algorithms are called **learners**

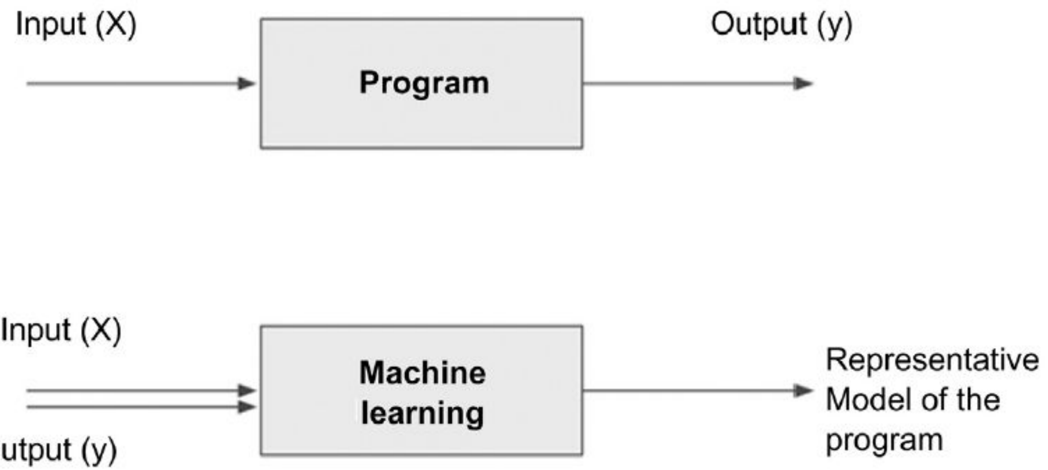
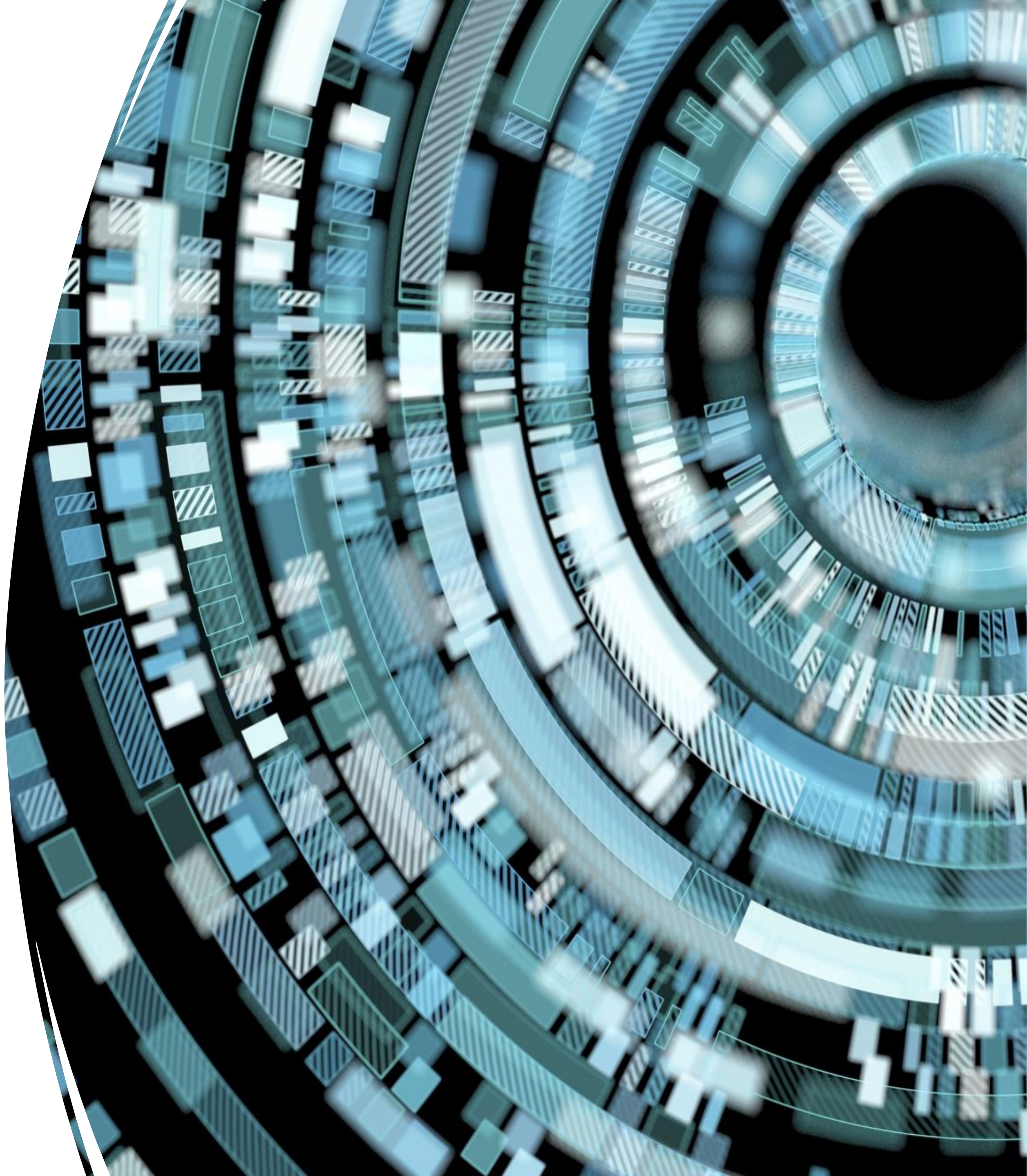


FIGURE 1.2
Traditional program and machine learning.

AI, MACHINE LEARNING, AND DATA SCIENCE

- Data science is the business application of machine learning, artificial intelligence, and other quantitative fields like statistics, visualization, and mathematics.
- It is an interdisciplinary field that extracts value from data.
- Datamining



%
$\frac{\partial^2 u}{\partial t^2}$
$+1=0$
∞
$F-E+V=2$
$J+\frac{\partial D}{\partial t}$
$\nabla \cdot B=0$
$E=mc^2$
‰

What is Data Science?

- It starts with **data**
- Data can range from a simple array of a few numeric observations to a complex matrix of millions of observations with thousands of variables.
- Data science utilizes certain specialized computational *methods* to discover meaningful and useful structures within a dataset.

Data Science - key features and motivations

1. Extracting Meaningful Patterns

- **Knowledge discovery** in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns or relationships within a dataset to make important decisions
- One of the key aspects of data science is the process of **generalization** of patterns from a dataset.
- Data science is also a process with defined steps, each with a set of tasks

2. Building Representative Models

- Modeling is a process in which a representative abstraction is built from the observed dataset.
- This model serves two purposes:

1. it predicts the output based on the new and unseen set of input variables
2. the model can be used to understand the relationship between the output variable and all the input variables.

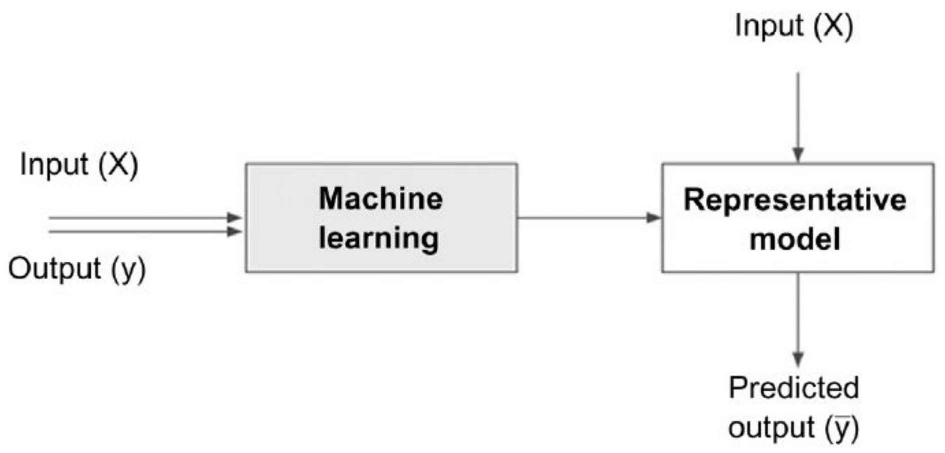


FIGURE 1.3
Data science models.

3. Combination of Statistics, Machine Learning, and Computing

- In the pursuit of extracting useful and relevant information from large datasets, data science borrows computational techniques from the disciplines of statistics, machine learning, experimentation, and database theories.
- One of the key ingredients of successful data science is substantial prior knowledge about the data and the business processes that generate the data, known as *subject matter expertise*.

4. Learning Algorithms

- Data science can also be defined as a process of discovering previously unknown patterns in data using *automatic iterative methods*.
- Data science is classified into tasks such as classification, association analysis, clustering, and regression.
- Each data science task uses specific learning algorithms like decision trees, neural networks, k-nearest neighbors (k-NN), and k-means clustering, among others.

5. Associated Fields

Descriptive statistics

Exploratory visualization

Dimensional slicing

Hypothesis testing

Data engineering

Business intelligence

CASE FOR DATA SCIENCE

- Massive accumulation of data
- A paradigm is needed to manage the massive volume of data, explore the inter-relationships of thousands of variables, and deploy machine learning algorithms to deduce optimal insights from datasets.
- Data science is one such paradigm that can handle large volumes with multiple attributes and deploy complex algorithms to search for patterns from data.

Each key motivation for using data science techniques are

1. Volume
2. Dimensions
3. Complex Questions

DATA SCIENCE CLASSIFICATION

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.
- ✔ Unsupervised or undirected data science uncovers hidden patterns in unlabeled data.

DATA SCIENCE CLASSIFICATION

- 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

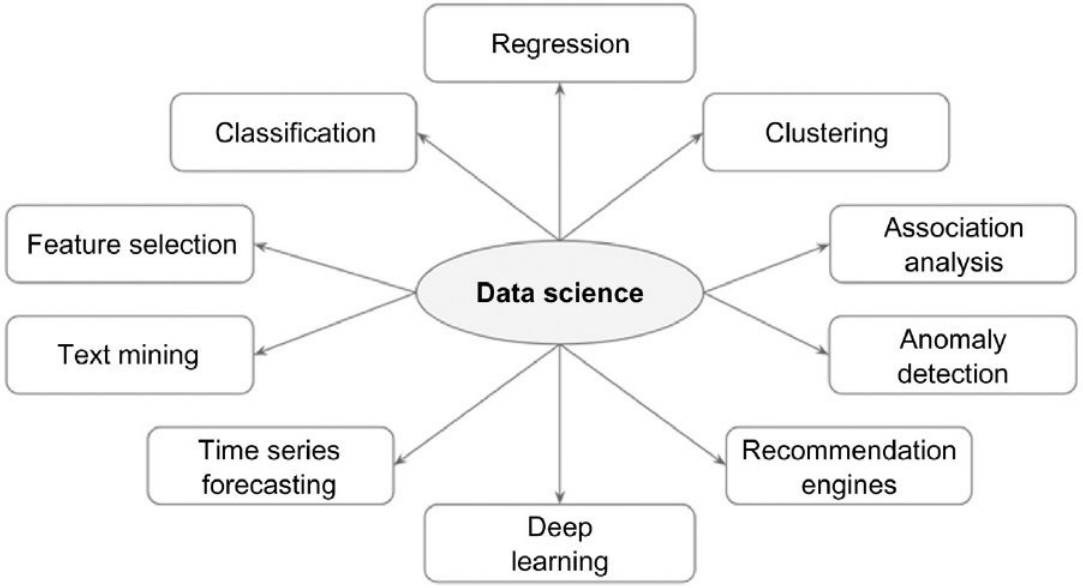


FIGURE 1.4
Data science tasks.

- **Classification** and **regression** techniques predict a target variable based on input variables
- **Deep learning** is a more sophisticated artificial neural network that is increasingly used for classification and regression problems.
- **Clustering** is the process of identifying the natural groupings in a dataset.
- **Market basket analysis** or **association analysis** used in cross selling
- **Recommendation engines** are the systems that recommend items to the users based on individual user preference.
- **Anomaly or outlier detection** identifies the data points that are significantly different from other data points in a dataset.
- **Text mining** is a data science application where the input data is text, which can be in the form of documents, messages, emails, or web pages.
- **Feature selection** is a process in which attributes in a dataset are reduced to a few attributes that really matter.

king5.com