Lecture-1

Relationship Between DS, ML and AI:

- **Data Science** provides the data and insights.
- **Machine Learning** develops models to make predictions.
- **AI** uses ML and other techniques to create intelligent systems.

Artificial Intelligent:

 AI enables machines to mimic human behavior and cognitive functions.
 Examples: Facial recognition, self-driving cars, sorting mail.
 AI techniques:

- Linguistics
- Natural language processing
- Decision science
- Vision
- Robotics
- Planning

Machine Learning:

 Machine Learning (ML) is a sub-field of AI, allowing machines to learn from experience.
 Experience for machines comes in the form of data.
 ML algorithms (learners) use training data (input and output) to create models that predict outputs from inputs.
 Example: Social media platforms and review sites use ML to automatically moderate and remove abusive content.

Machine learning turns the traditional programing model upside down

.

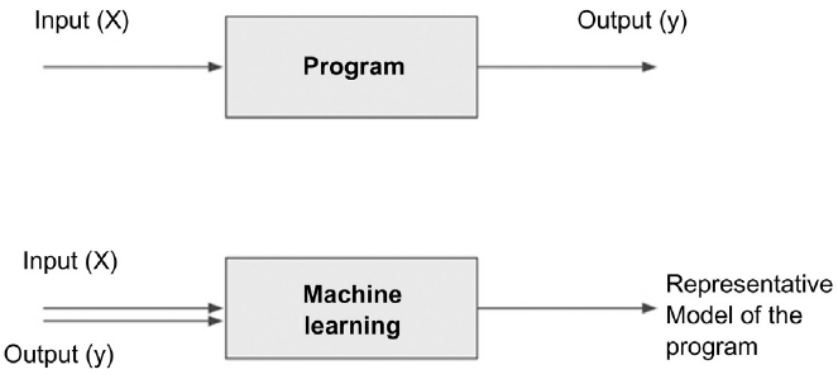


FIGURE 1.2
Traditional program and machine learning.

Data Science:

- **Data Science** applies AI, ML, statistics, and visualization to solve business problems.
- It extracts value from data and is often called data mining.
- Relies heavily on machine learning for predictive and analytical tasks.
- **Examples**:
  - Movie recommendation engines
  - Fraud detection for credit card transactions
  - Predicting customer churn
  - Forecasting revenue for the next quarter

What is Data Science?

 Data Science begins with data, ranging from simple arrays to complex datasets with millions of observations.
 It uses specialized computational methods to uncover meaningful patterns and insights.
 Closely associated fields:

- Database systems
- Data engineering
- Visualization
- Data analysis
- Experimentation
- Business intelligence (BI)

Data science is a collection of techniques used to extract value from data.

Key features and Motivations:

Extracting Meaningful Patterns:

Knowledge discovery is finding useful patterns in data for decision-making. Data science helps generalize these patterns to uncover new insights that can be acted on.

Building Representative Models:

In statistics, a model represents relationships between variables in a dataset, showing how some variables relate to others. Modeling involves creating an abstraction from data. For example, a model can predict loan interest rates based on credit score, income, and loan amount.

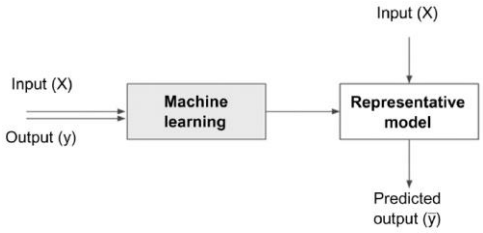


FIGURE 1.3
Data science models.

Data science involves building a model that fits observational data to predict outputs and understand relationships between variables. For instance, it can predict loan interest rates based on credit score, income, and loan amount, while also revealing how each factor influences the rate. A model is useful for both prediction and explanation.

Combination of Statistics, Machine Learning, and Computing:

Data science finds valuable insights in large datasets using statistics, machine learning, and database techniques. Success depends on understanding both the data and the business, known as subject expertise. It's done in steps, learning more about data patterns along the way.

Learning Algorithms:

 Data Science discovers hidden patterns in data through automatic, iterative methods.

 Uses advanced algorithms, setting it apart from traditional data analysis.

 Tasks in Data Science:

- **Classification**: Categorizes data (e.g., decision trees, neural networks).
- **Association Analysis**: Finds relationships between items.
- **Clustering**: Groups similar data points (e.g., k-means clustering).
- **Regression**: Predicts continuous outcomes.

 Common Algorithms: Decision trees, neural networks, k-nearest neighbors, k-means clustering.

Associated Fields:

Data science relies on several key associated fields:

- **Descriptive Statistics**: Summarizes data with measures like mean and standard deviation.
- **Exploratory Visualization**: Uses charts and graphs to explore data patterns.
- **Dimensional Slicing**: Breaks data into subsets for deeper analysis.
- **Hypothesis Testing**: Tests assumptions or predictions about data.
- **Data Engineering**: Manages and prepares data for analysis.
- **Business Intelligence**: Extracts actionable insights to support decisions.

Case for Data Science:

Data science helps manage large, complex data to find patterns. It's used for:

1. **Volume** – Large amounts of data.
2. **Dimensions** – Many features in data.

3. **Complex Questions** – Solving tough problems.

Data Science Classification:

Data science problems can be divided into:

1. **Supervised Learning**
2. **Unsupervised Learning**

Common tasks include: Classification, Regression, Association Analysis, Clustering, Anomaly Detection, Recommendation Engines, Feature Selection, Time Series Forecasting, Deep Learning, Text Mining

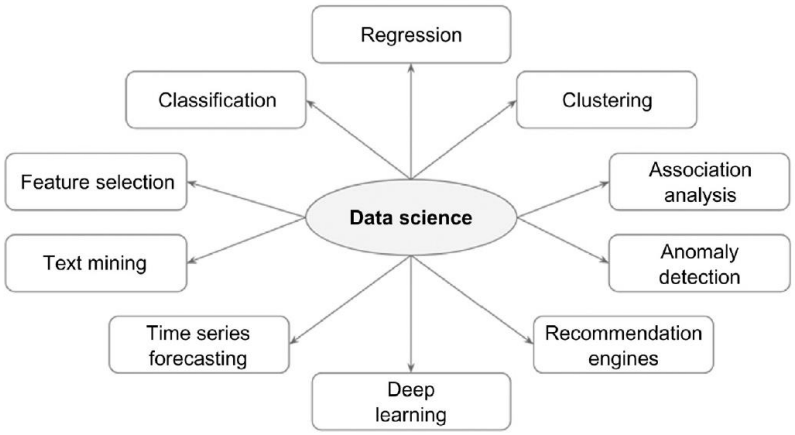


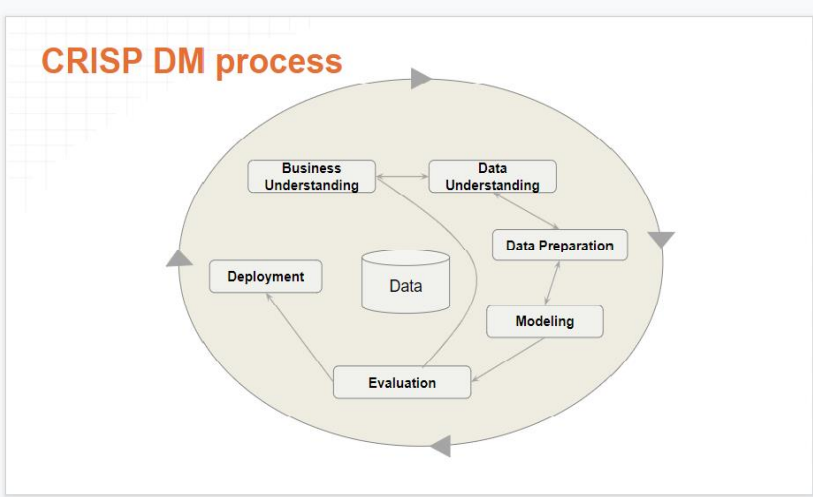
FIGURE 1.4
Data science tasks.

Applications of DS:

- Internet Search
- Transport
- Finance
- HealthCare
- Image Recognition
- Airline Routing Planning
- Data Science in Gaming

Lecture-2(Data Science Process)

CRISP Data Mining Framework:



The CRoss Industry Standard Process for Data Mining (CRISP-DM) is a process model that serves as the base for a data science process. It has six sequential phases:

Business understanding – What does the business need?

Data understanding – What data do we have / need? Is it clean?

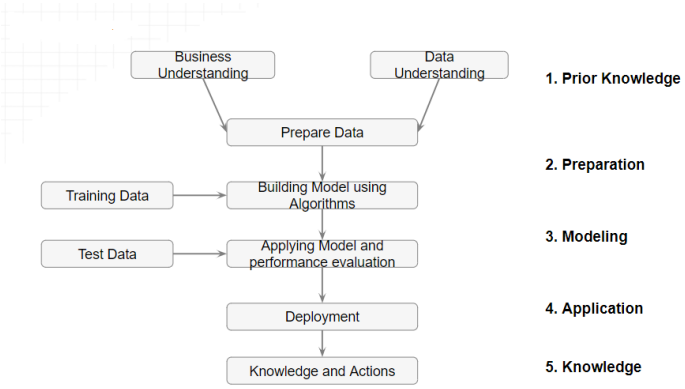
Data preparation – How do we organize the data for modeling?

Modeling – What modeling techniques should we apply?

Evaluation – Which model best meets the business objectives?

Deployment – How do stakeholders access the results?

Data Science Process:



Prior Knowledge:

1. **Objective of the Problem**
   The process starts with a clear business goal or question. This guides dataset selection and algorithm choice.
   Example: Can we predict the interest rate for a new borrower based on past credit scores?
2. **Subject Area of the Problem**
   Understanding the business context is essential to identify relevant patterns in the data.
   Example: Knowing how the lending business works is crucial for predicting interest rates.
3. **Data**
   Understanding how data is collected and used helps create a dataset that answers the business question.
   Example: A dataset with credit scores and interest rates can be used to predict future rates.

Data Types:

Two Types of Data:

1. **Labelled Data**
   Data with a designated attribute (label) that we use to predict its value for new, unseen instances.
   Example: A dataset with credit scores and interest rates, where the interest rate is the label, we want to predict.

Outlook	Temp (°F)	Humidity (%)	Windy	Class
sunny	75	70	true	play
sunny	80	90	true	don't play
sunny	85	85	false	don't play
sunny	72	95	false	don't play
sunny	69	70	false	play
overcast	72	90	true	play
overcast	83	78	false	play
overcast	64	65	true	play
overcast	81	75	false	play
rain	71	80	true	don't play
rain	65	70	true	don't play
rain	75	80	false	play
rain	68	80	false	play
rain	70	96	false	play

Outlook=sunny	Temp=79	Humidity=88	Windy=false	Class=?

2. **Unlabeled Data**
   Data without any designated attribute. The goal is to extract useful information or patterns from the data itself.
   Example: A dataset of customer features without knowing the target variable, like interest rate or loan approval.

Age	Gender	Income	Profession	Tenure	City
35	M	60,000	IT	12	KRK
23	F	90,000	Sales	3	WAW
18	M	12,000	Student	1	KRK
42	F	128,000	Doctor	13	KRK
34	M	63,000	Manager	8	WAW
56	M	82,000	Teacher	30	WAW

Learning Methods:

Supervised Learning
Data mining with labelled data is called supervised learning.

1. **Classification**
   When the target attribute is categorical, the task is classification. It predicts a label, such as categorizing patients into risk groups or grading student projects.
   Techniques: Nearest Neighbour Matching, Classification Rules, Classification Trees.
2. **Numerical Prediction (Regression)**
   When the target attribute is numerical, the task is regression. It predicts a numerical value, such as company profits or stock prices.
   Technique: Neural Networks.

Unsupervised Learning
Data mining with unlabelled data is called unsupervised learning.

1. **Association Rules**
   Finds relationships among variables, often in the form of rules.
   Technique: APRIORI, Market Basket Analysis.
2. **Clustering**
   Groups similar data points together into clusters.
   Techniques: K-Means Clustering, Agglomerative Hierarchical Clustering.

A dataset (example set) is a collection of data with a defined structure. Table 2.1 shows a dataset. It has a well-defined structure with 10 rows and 3 columns along with the column headers. This structure is also sometimes referred to as a "data frame".

A data point (record, object or example) is a single instance in the dataset. Each row in Table 2.1 is a data point. Each instance contains the same structure as the dataset.

An attribute (feature, input, dimension, variable, or predictor) is a single property of the dataset. Each column in Table 2.1 is an attribute. Attributes can be numeric, categorical, date-time, text, or Boolean data types. In this example, both the credit score and the interest rate are numeric attributes.

Table 2.1 Data Set

Borrower ID	Credit Score	Interest Rate
01	500	7.31%
02	600	6.70%
03	700	5.95%
04	700	6.40%
05	800	5.40%
06	800	5.70%
07	750	5.90%
08	550	7.00%
09	650	6.50%
10	825	5.70%

- A label (class label, output, prediction, target, or response) is the special attribute to be predicted based on all the input attributes. In Table 2.1, the interest rate is the output variable.

Table 2.2 New Data With Unknown Interest Rate

Borrower ID	Credit Score	Interest Rate
11	625	?

Data Preparation:

Data Exploration:

 Data Preparation: The first step is to explore and understand the dataset thoroughly.

 Exploratory Data Analysis (EDA): A process of examining data to find patterns, trends, and outliers.

 Methods in EDA:

- **Descriptive Statistics**: Summarizing data with measures like mean, median, standard deviation.
- **Visualization**: Using charts and graphs to better understand the data's distribution and relationships.

Data quality:

 Data Quality: The measure of how well a dataset meets its intended purpose.It's based on:

- **Data Correctness**
- **Data Freshness**
- **Data Completeness**

• **Data Correctness**: Refers to how accurately data represents real-world facts. Correct data reflects truth and is free from errors.

- **Importance**: Incorrect data can lead to poor decisions and unreliable results, affecting credibility. For example, incorrect patient data in healthcare can lead to wrong treatment.
- **Challenges**: Issues like collection noise, outdated data, and incorrect schema can affect correctness.

• **Data Freshness**: Refers to how up-to-date and timely the data is, ensuring it accurately reflects the current state of an entity.

- **Example**: A customer's address must be updated in the database whenever it changes to maintain accuracy.

• **Data Completeness**: Ensures all necessary values are present in the dataset.

- **Example**: An employee dataset must include key details like name, ID, and position for each entry. Missing data can lead to incomplete and inaccurate insights.

Missing values:

Missing values in datasets can occur due to irrelevant attributes, equipment malfunctions, changes in data forms, or unavailable information (e.g., medical data).

The **k-NN algorithm** handles missing values well by ignoring missing attributes during distance calculations.
In contrast, **neural networks** require data preparation (e.g., imputing missing values) to perform well.

Methods to Handle Missing Values

1. Discard Instances:
This strategy involves removing all instances with at least one missing value.

- **Advantage:** It avoids introducing errors or biases from imputation.
- **Disadvantage:** It may lead to a loss of valuable data, reducing the reliability of the results.

2. Replace by Most Frequent/Average Value:
This strategy estimates missing values by using existing data.

- For **categorical attributes**, replace missing values with the most frequent value.
- For **continuous attributes**, replace missing values with the average value.

It's a simple and effective method but may introduce bias if the missing data isn't random.

Noisy values:

A valid data point that is incorrectly recorded.

- Example: 69.72 entered as 6.972.
- Example: "brown" recorded as "blue."

Noisy values can lead to inaccuracies in data analysis and models.

Invalid values:

- 69.7X for 6.972 or bbrown for brown
- An invalid value can easily be detected and either corrected or rejected

Data types and Conversion:

Attribute Data Types:

1. **Continuous Numeric**:
   - Example: Interest rate (e.g., 5.2%).
2. **Integer Numeric**:
   - Example: Credit score (e.g., 750).
3. **Categorical**:
   - Example: Credit score as categories (e.g., poor, good, excellent).

Different data science algorithms have specific requirements or restrictions on the data types they can handle effectively. For instance:

- Some algorithms may require numeric data.
- Others, like decision trees, can handle both numeric and categorical data.

Transformation:

k-NN Algorithm:

- Requires **numeric** and **normalized** input attributes.
- Normalization prevents attributes with larger values from dominating the distance calculation.

Normalization Process:

Scales attribute values to a range of 0 to 1.

Normalized value= $\frac{a-min}{max-min}$

min: Lowest value, **max**: Highest value of the attribute.

Mileage (miles)	Number of doors	Age (years)	Number of owners
18,457	2	12	8
26,292	4	3	1

Outliers:

- **Definition**: Data points that are significantly different from the rest of the dataset.
- **Causes**:
  - Inconsistent data entry.
  - Erroneous observations.
- **Impact**: Outliers can skew the data distribution and affect the performance of data models.

Applications:

- Outlier detection is crucial in applications like:
  - **Fake email detection**.
  - **Fraud detection**.
  - **Intrusion detection**.

Feature selection:

- Reducing the number of attributes in a dataset without sacrificing model performance.
- Datasets can have many attributes, especially in text mining (e.g., each word as an attribute).

Challenges:

- Not all attributes are useful or relevant.
- Highly correlated attributes (e.g., income and taxes) can add redundancy.
- Too many attributes increase model complexity and can degrade performance due to the **curse of dimensionality**.

Data Sampling

- **Definition**: Selecting a subset of records from the original dataset to represent the whole dataset for analysis or modeling.
- **Purpose**:
  - Reduces the amount of data to be processed.
  - Speeds up model building.
  - Samples are expected to have similar properties (e.g., mean) as the full dataset.

Benefits:

- Often sufficient for gaining insights and building predictive models.
- The error introduced by sampling is generally small and outweighed by the benefits, especially in large datasets.

Modeling:

Splitting Training and Test Datasets:

- **Training Dataset**: Used to build the model, containing known attributes and target values.
- **Test/Validation Dataset**: Used to evaluate the model's performance and check its validity.

Process:

- The original dataset is split into two parts:
  - **Two-thirds** for training.
  - **One-third** for testing.

This ensures the model is trained on one set of data and validated on a separate set to avoid overfitting and assess its generalization ability.

Training Set and Test Set

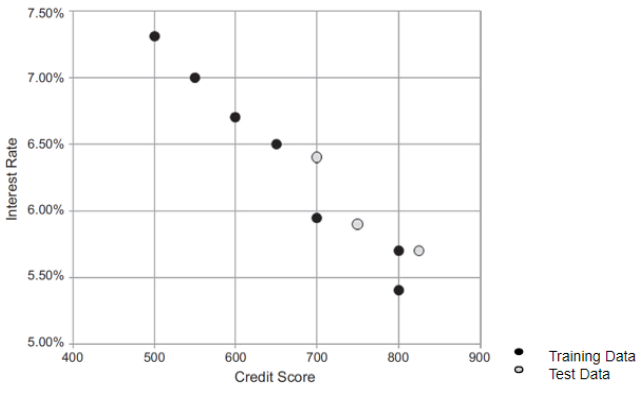
Table 2.3 Training Data Set

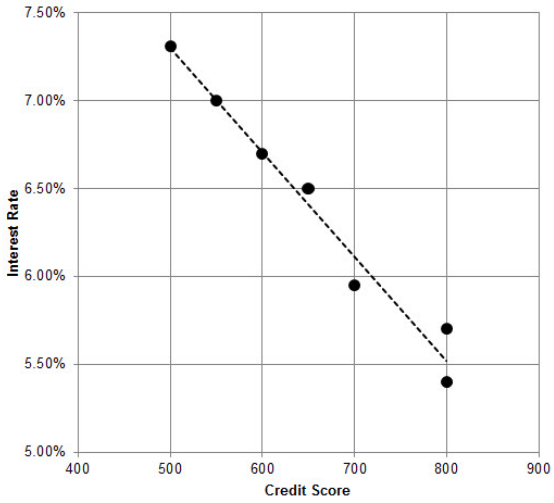
Borrower	Credit Score (X)	Interest Rate (Y)
01	500	7.31%
02	600	6.70%
03	700	5.95%
05	800	5.40%
06	800	5.70%
08	550	7.00%
09	650	6.50%

Table 2.4 Test Data Set

Borrower	Credit Score (X)	Interest Rate (Y)
04	700	6.40%
07	750	5.90%
10	825	5.70%

Modeling





Evaluation of test dataset

Table 2.5 Evaluation of Test Data Set

Borrower	Credit Score (X)	Interest Rate (Y)	Model Predicted (Y)	Model Error
04	700	6.40%	6.11%	-0.29%
07	750	5.90%	5.81%	-0.09%
10	825	5.70%	5.37%	-0.33%