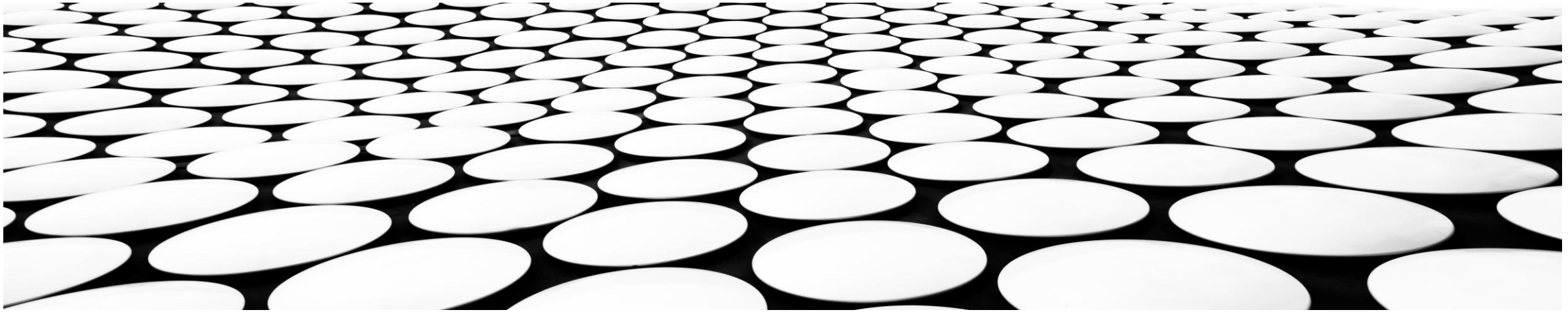
DATA MINING AND PREDICTIVE DATA ANALYSIS

CHAPTER-1

AN INTRODUCTION TO DATA MINING AND PREDICTIVE ANALYTICS



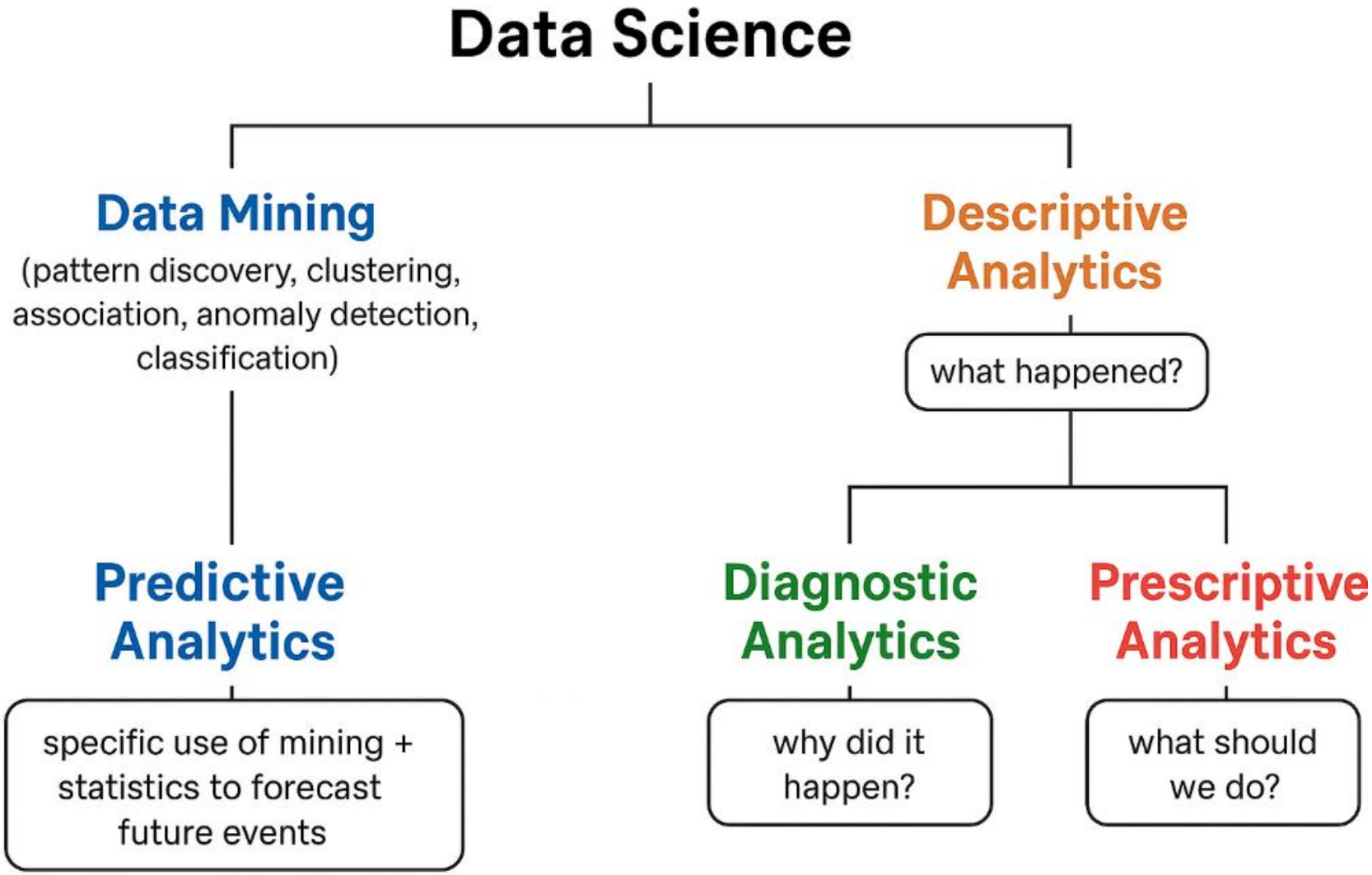
INTRODUCTION

- Data mining
  - Data mining is the process of discovering useful patterns and trends in large data sets.
  - Extracting or “mining” of interesting (non-trivial, implicit, previously unknown and potentially useful) knowledge from large amounts of data.
- Predictive analytics
  - Predictive analytics is the process of extracting information from large data sets in order to make predictions and estimates about future outcomes..
  - The process uses data analysis, machine learning, artificial intelligence, and statistical models to find patterns that might predict future behavior.
- **Data Mining = What’s in the data? (pattern discovery)**
- **Predictive Analytics = What will happen next? (future prediction)**

CORRELATION BETWEEN DATA MINING & PREDICTIVE ANALYTICS

- **Predictive Analytics** is one branch of data mining applications, specifically focused on forecasting.
- Data Mining is often considered a broader field that includes techniques for pattern discovery, clustering, association, anomaly detection, and classification.
- Predictive Analytics is more goal-specific – it uses techniques (many borrowed from data mining & statistics) to predict future outcomes.
- So, you can think of Predictive Analytics as an application area (or specialization) within Data Mining & Machine Learning.

CORRELATION BETWEEN DATA MINING & PREDICTIVE ANALYTICS



CORRELATION BETWEEN DATA MINING & PREDICTIVE ANALYTICS

- **Data Mining**
  - **Meaning**: The process of discovering patterns, correlations, and useful information from large datasets.
  - **Techniques**: Clustering, Association, Anomaly Detection, Classification.
  - **Example**:
    - **Clustering**: Grouping customers into segments based on purchase behavior.
    - **Association**: Market basket analysis – if a customer buys bread, they are likely to buy butter.
    - **Anomaly Detection**: Identifying fraudulent credit card transactions.
    - **Classification**: Predicting whether an email is spam or not spam.

CORRELATION BETWEEN DATA MINING & PREDICTIVE ANALYTICS

- **Predictive Analytics (subset of Data Mining + Statistics)**
  - **Meaning**: Uses past data patterns to predict future outcomes.
  - **Example**:
    - Predicting customer churn (likelihood of a customer leaving a service).
    - Forecasting demand for ride-hailing services based on historical ride data.
    - Predicting disease risk from patient health records.

CORRELATION BETWEEN DATA MINING & PREDICTIVE ANALYTICS

- **Descriptive Analytics (What happened?)**
  - **Meaning**: Summarizes past data to understand trends and patterns.
  - **Example**:
    - Sales dashboard showing monthly revenue trends.
    - A university analyzing past exam results to see pass/fail ratios.
    - Website analytics summarizing total visits, bounce rate, and average session duration.

CORRELATION BETWEEN DATA MINING & PREDICTIVE ANALYTICS

- **Descriptive Analytics (What happened?)**
  - **Meaning**: Summarizes past data to understand trends and patterns.
  - **Example**:
    - Sales dashboard showing monthly revenue trends.
    - A university analyzing past exam results to see pass/fail ratios.
    - Website analytics summarizing total visits, bounce rate, and average session duration.

COMPARISON:

Aspect	Data Mining	Predictive Analytics
Definition	Process of discovering hidden patterns, relationships, or trends in large datasets.	Process of using historical data + statistical models to **predict future outcomes**.
Goal	Knowledge discovery (what is happening in the data).	Forecasting (what will happen in the future).
Focus	**Exploratory** – finding unknown patterns.	**Predictive** – forecasting specific outcomes.
Output	Descriptive patterns, rules, clusters.	Probability scores, forecasts, risk assessments.
Time orientation	Mostly **past and present data patterns**.	Uses past data to **predict the future**.
Techniques Used	Clustering, classification, association rule mining, anomaly detection, regression.	Regression, decision trees, neural networks, time series analysis, ensemble models.

EXAMPLES

- **Example 1: Retail Store**
  - Data Mining: Find that customers who buy bread and butter often also buy milk (association rule).
  - Predictive Analytics: Forecast sales of milk for next month based on past sales, promotions, and seasonality.
- **Example 2: Banking**
  - Data Mining: Discover fraudulent transaction patterns (e.g., unusual location + time + amount).
  - Predictive Analytics: Predict the probability of a customer defaulting on a loan.

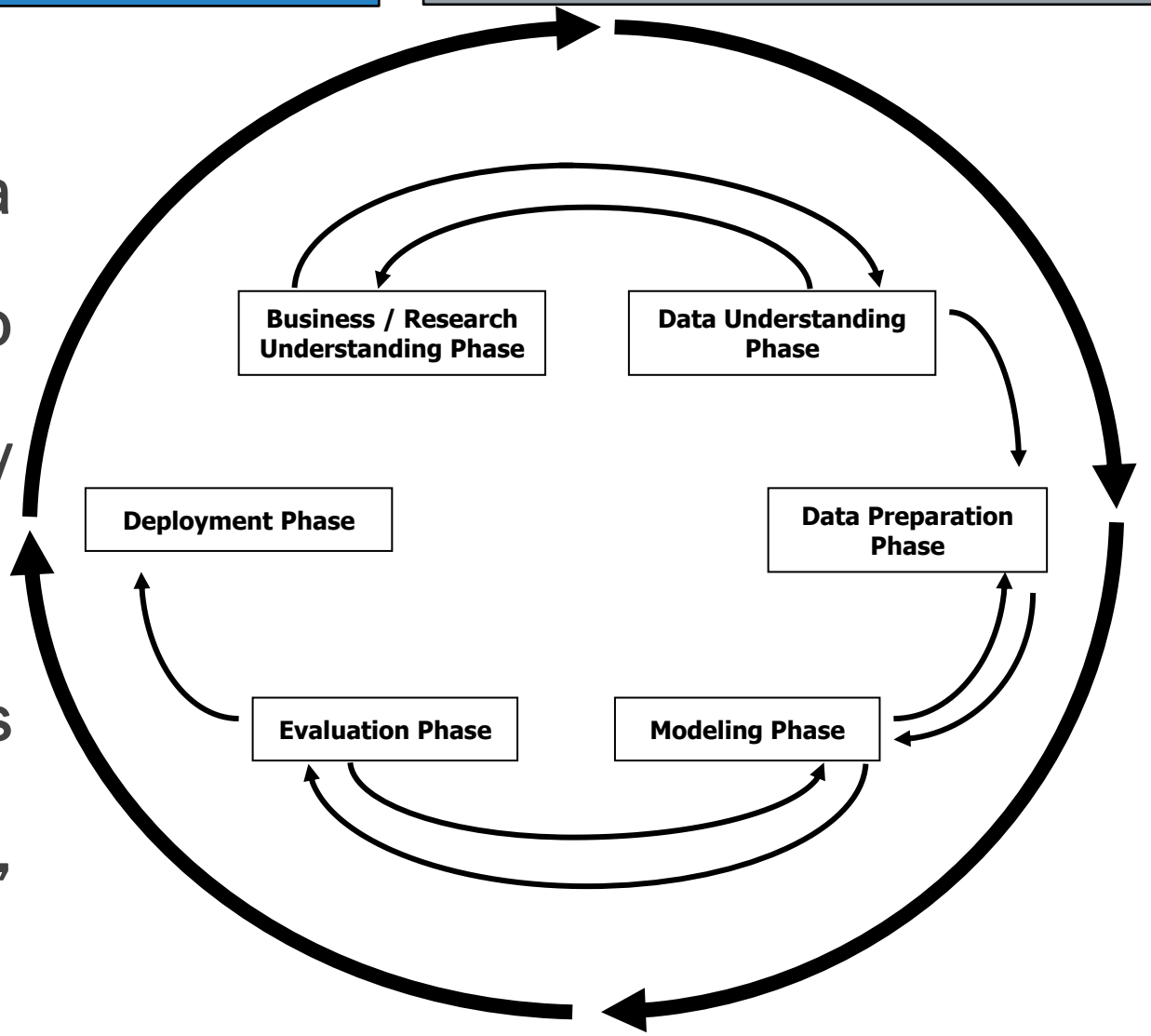
THE NEED FOR HUMAN DIRECTION OF DATA MINING

- **Automation is no substitute for human oversight in Data Mining**.
  - Humans need to be actively involved at every phase of the data mining process.
  - Task of data mining should be integrated into human process of problem solving.
- The very power of the readily available data mining algorithms embedded in the black box software makes their misuse proportionally more dangerous
  - Understanding of the data and the statistical and mathematical model structures underlying the software is required

CRISP-DM: CROSS INDUSTRY STANDARD PROCESS

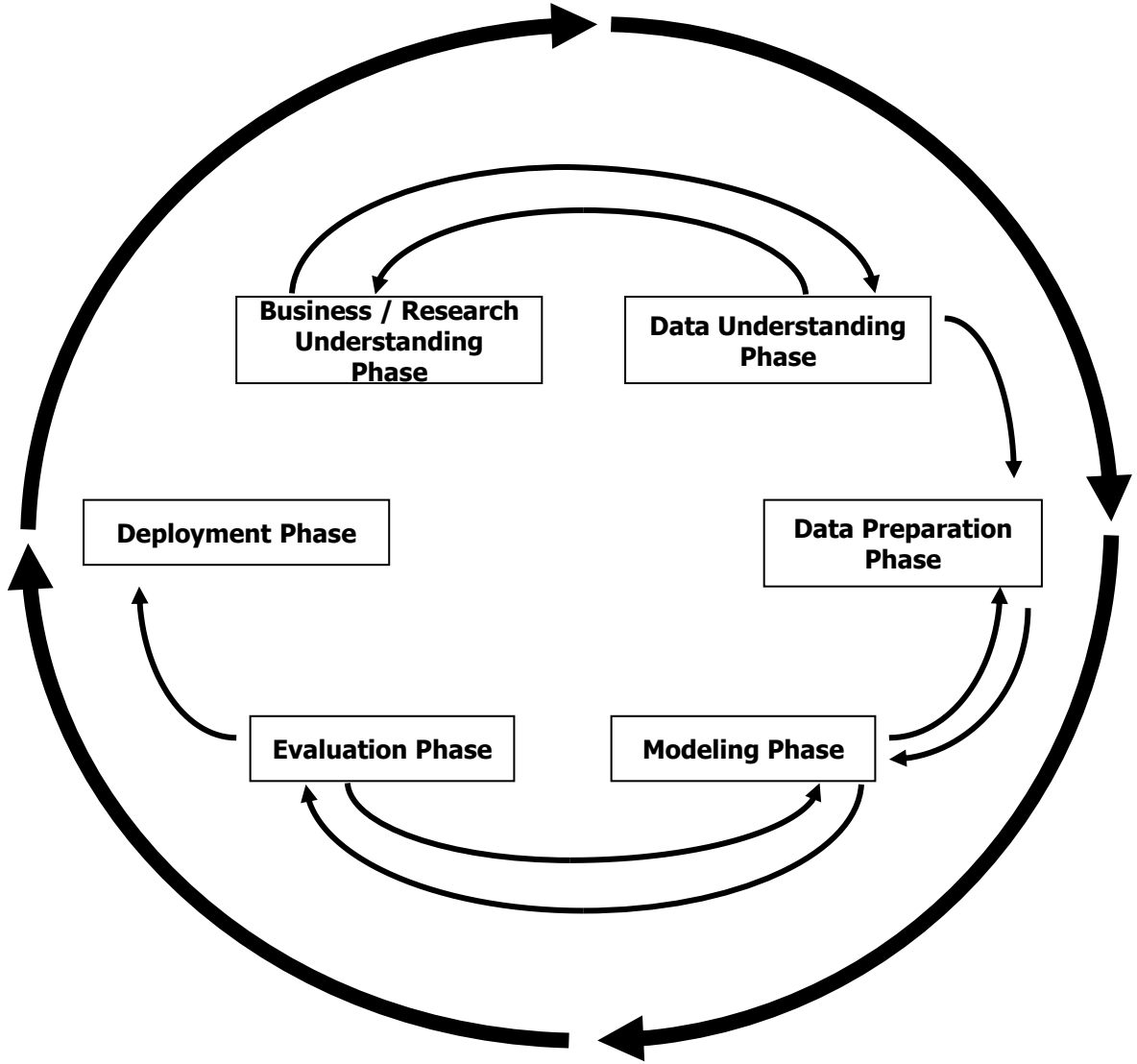
CRISP-DM:

- A widely used framework that provides a structured and systematic approach to guide data mining projects efficiently and effectively
- Outlines a six-phase process that helps organizations effectively plan, execute, and evaluate data mining initiatives.



CRISP-DM: CROSS INDUSTRY STANDARD PROCESS

- Iterative CRIP-DM process shown in outer circle
- Most significant dependencies between phases shown
- Next phase depends on results from preceding phase
- Returning to earlier phase possible before moving forward



CRISP-DM: THE SIX PHASES

❑CRISP-DM: The Six Phases

1. Business/Research Understanding Phase

- Define project requirements and objectives
- Translate objectives into data mining problem definition
- Prepare preliminary strategy to meet objectives

2. Data Understanding Phase

- Collect data
- Perform exploratory data analysis (EDA)
- Assess data quality
- Optionally, select interesting subsets

CRISP-DM: THE SIX PHASES

3. Data Preparation Phase

- Prepares for modeling in subsequent phases
- Select cases and variables appropriate for analysis
- Cleanse and prepare data so it is ready for modeling tools
- Perform transformation of certain variables, if needed

4. Modeling Phase

- Select and apply one or more modeling techniques
- building and assessing models, and assessing model quality
- Calibrate model settings to optimize results
- If necessary, additional data preparation may be required for supporting a particular technique

CRISP-DM: THE SIX PHASES

5. Evaluation Phase

- Evaluate one or more models for effectiveness
- Determine whether defined objectives achieved
- Establish whether some important facet of the problem has not been sufficiently accounted for
- Make decision regarding data mining results before deploying to field

6. Deployment Phase

- Make use of models created
- This final phase involves planning the deployment, planning the monitoring and maintenance, producing the final report, and reviewing the project
- In businesses, customer often carries out deployment based on your model

CRISP-DM: THE SIX PHASES

- CRISP-DM helps organizations ensure that their data mining efforts are well-defined, managed, and aligned with their business objectives, ultimately leading to more valuable and reliable results.
- CRISP-DM enhances collaboration by aligning technical tasks with business objectives.
- CRISP-DM framework ensures that all stakeholders–data scientists, business analysts, and leadership–are on the same page throughout the project.

FALLACIES OF DATA MINING

	Fallacy	Reality
1	• Set of tools can be turned loose on data repositories • Finds answers to all business problems	• No automatic data mining tools solve problems • Rather, data mining is process (CRISP-DM) • Integrates into overall business objectives
2	• Data mining process is autonomous • Requires little oversight	• Requires significant intervention during every phase • After model deployment, new models require updates • Continuous evaluative measures monitored by analysts
3	• Data mining quickly pays for itself	• Return rates vary • Depending on startup, personnel, data preparation costs, etc.
4	• Data mining software easy to use	• Ease of use varies across projects • Analysts must combine subject matter knowledge with specific problem domain

FALLACIES OF DATA MINING

	Fallacy	Reality
5	• Data mining identifies causes of business problems	• Knowledge discovery process uncovers patterns of behavior • Humans interpret results and identify causes
6	• Data mining automatically cleans data in databases	• Data mining often uses data from legacy systems • Data possibly not examined or used in years • Organizations starting data mining efforts confronted with huge data preprocessing task
7	• Data mining always provides positive results.	• There is no guarantee of positive results • But used properly, data mining can provide actionable and highly profitable results.

WHAT TASKS CAN DATA MINING ACCOMPLISH?

- **Six Common Data Mining Tasks:**

1. Description
2. Estimation
3. Prediction
4. Classification
5. Clustering
6. Association

WHAT TASKS CAN DATA MINING ACCOMPLISH?

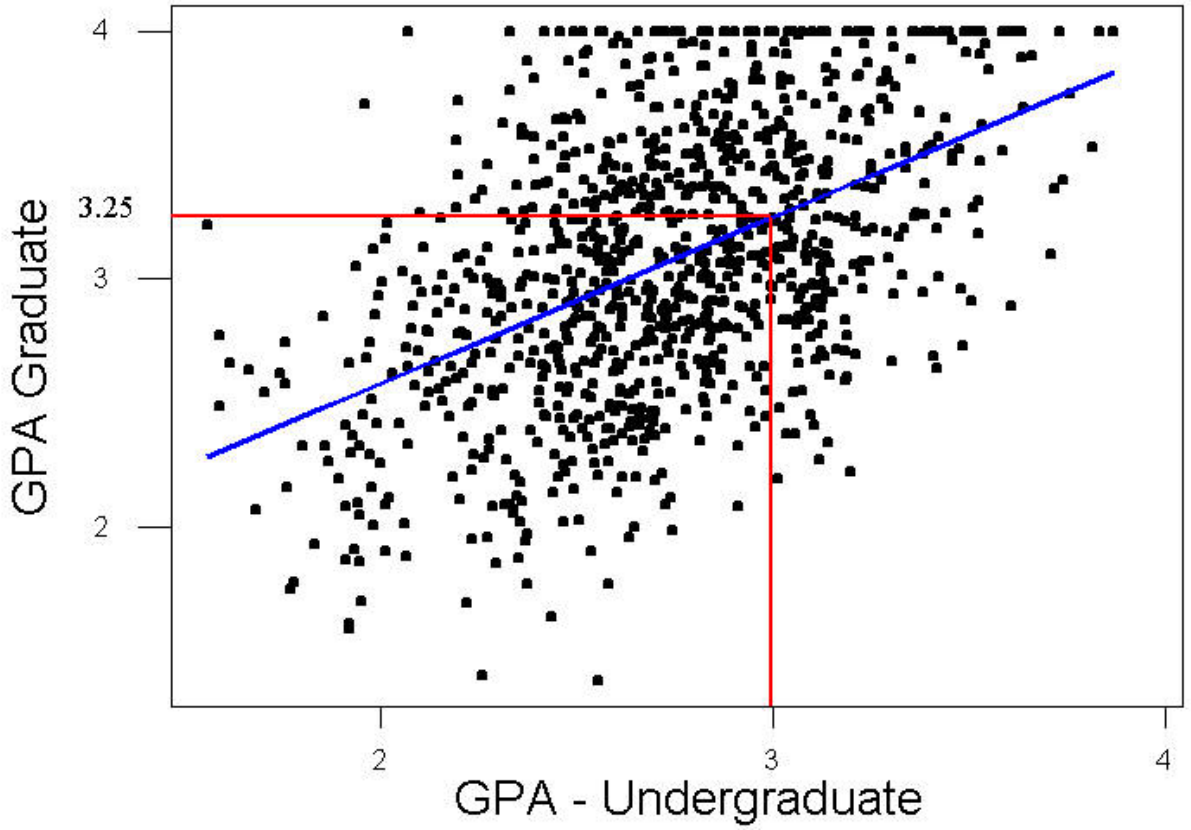
- **Description**
  - Describes the general properties, characteristics, patterns and general behavior of the data. it focuses on summarizing and interpreting what is contained in the dataset.
  - Goal: To provide insights and understanding of data without necessarily making predictions.
  - Techniques Used:
    - Descriptive statistics (mean, median, standard deviation, frequency distribution).
    - Data visualization (histograms, pie charts, heat maps).
    - Data summarization (OLAP, dashboards).
  - Example: Analyzing hospital records to find the average patient age, common diseases, and seasonal variations in admissions.
  - Example: A retail chain analyzes transaction data to find the average purchase amount, most common product categories, and sales trends across seasons.

WHAT TASKS CAN DATA MINING ACCOMPLISH?

- **Estimation**
  - Estimates the value of an **unknown (**numeric and continuous**)** variable based on input data.
  - Purpose: To approximate quantities that are not yet known (**numerical prediction**)
  - Common Techniques:
    - Regression analysis (linear and nonlinear regression).
    - k-Nearest Neighbors (k-NN), Neural networks, Decision trees (for regression tasks).
  - Example: Estimate a patient's systolic blood pressure, based on patient's age, gender, body-mass index, and sodium levels
    - Use training data to develop model that estimates blood pressure based on predictor variables
    - Apply model to new cases, to obtain estimated blood pressure

WHAT TASKS CAN DATA MINING ACCOMPLISH?

- Example: **estimating** the GPA of a graduate student, based on that student's undergraduate GPA.
- Figure 1.2 shows scatter plot of graduate GPA against undergraduate GPA (1000 students)
- Linear regression finds line best approximating relationship between two variables



WHAT TASKS CAN DATA MINING ACCOMPLISH?

- **Prediction**
  - Forecasting unknown future outcomes or values based on patterns learned from historical and present data.
  - Similar to classification and estimation, except results lie in the future
  - Unlike estimation, prediction may involve both numeric and categorical variables.
  - Common Techniques:
    - Time series forecasting (ARIMA, Prophet, LSTM models).
    - Regression models.
    - Machine learning algorithms (Random Forest, Gradient Boosting, k-NN, NN)
  - Example: Predicting whether a patient will develop diabetes within 5 years based on lifestyle and medical history
  - Predict price of stock 3 months into future, based on past performance.

WHAT TASKS CAN DATA MINING ACCOMPLISH?

Classification

- Classification assigns records into predefined categories (classes) based on input attributes.
- Similar to Estimation task, except target variable is categorical.
- It is a supervised learning technique -- the training data already contains known class labels.
- **Goal:** To learn a model from labeled data and use it to classify new, unseen data record.
- Example: Classify the Income Bracket of an individual as Low, Middle or High based their Age, Gender and Occupation.
  - for example, 63-year-old female professor -> high
- Common Techniques:

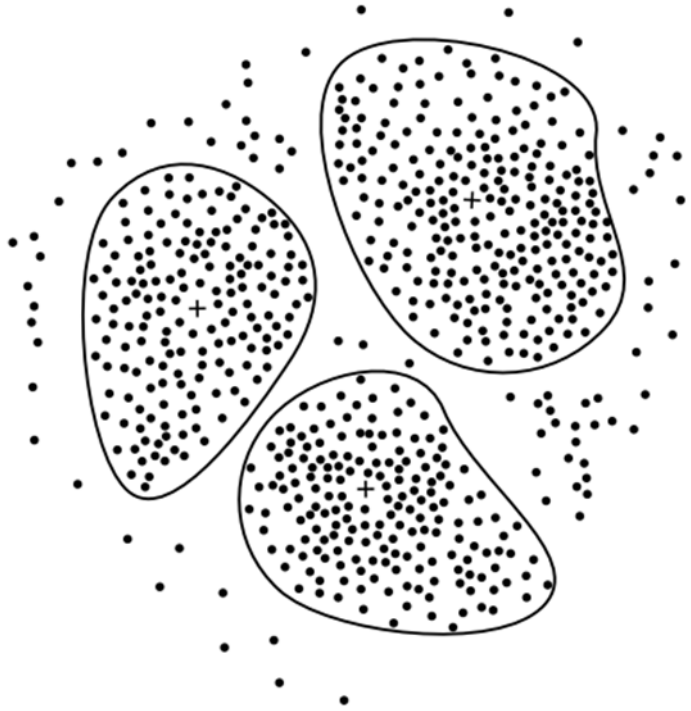
Subject	Age	Gender	Occupation	Income Bracket
001	47	F	Software Engineer	High
002	28	M	Marketing Consultant	Middle
003	35	M	Unemployed	Low
…	…	…	…	…

WHAT TASKS CAN DATA MINING ACCOMPLISH?

- Clustering
  - Grouping data objects into clusters - such that objects in the same cluster are more similar to each other than to those in other clusters.
  - Cluster – a collection of records similar to one another, and dissimilar to records in other clusters
  - • Purpose: To discover hidden groupings in data without predefined labels.
  - • Techniques: K-means clustering, hierarchical clustering, DBSCAN.
  - Clustering is unsupervised method– Target variable not specified
    - –-Clustering does not try to classify/estimate/predict target variable

WHAT TASKS CAN DATA MINING ACCOMPLISH?

- Example: Segmenting customers into groups based on purchasing behavior.
- Clustering is often used as a preliminary step in a data mining process, with the resulting clusters being used as further inputs into a different technique downstream
  - Dimensionality Reduction
  - Pattern Recognition
  - Market Analysis
  - Spatial Data Analysis
  - Image Processing



WHAT TASKS CAN DATA MINING ACCOMPLISH?

- **Association**
  - The task of association seeks to uncover rules for quantifying the relationship between two or more attributes
  - Discovering interesting relationships or associations among variables in datasets.
  - Purpose: To identify patterns of co-occurrence.
  - Techniques: Association rule mining, Apriori algorithm, FP-Growth.
  - Example: Market basket analysis — “Customers who buy bread are also likely to buy butter.”

WHAT TASKS CAN DATA MINING ACCOMPLISH?

- Association Rules- Quantify relationships between two or more attributes in the form of rules as:
  - IF antecedent THEN consequent
- Rules measured using support and confidence
- Example: A particular Digital Store might find that:
  - **Thursday night 200 of 1,000 customers bought 'Computers', and of those buying 'Computers', 50 purchased 'Antivirus s/w'.**
  - **Association Rule: "IF buy Computers, THEN buy Antivirus s/w"**
  - **Support = 200/1,000 = 5%, and confidence = 50/200 = 25%**

WHAT TASKS CAN DATA MINING ACCOMPLISH?

Task	Data Type	Purpose	Example
Description	All	Summarize characteristics	Average customer income
Estimation	Numeric	Approximate value	Estimate house price
Prediction	Future/Categorical	Forecast outcome	Predict loan default
Classification	Categorical	Assign to predefined classes	Spam detection
Clustering	Mixed	Group by similarity	Customer segmentation
Association	Transactional	Find relationships	Market basket analysis