DATA 3421

Data Mining, Management, and Curation

Masoud(Max) Rostami

Department of Science / College of Science
Data Science Program / College of Science
The University of Texas Arlington, Texas

DATA 3421

Data Mining, Management, and Curation

Week 1

Data Mining Pipeline

Dr. Masoud(Max) Rostami

Department of Science / College of Science
Data Science Program / College of Science
The University of Texas Arlington, Texas

Introduce yourself:

1) What is your major?

2) What year?

3) When will you graduate?

4) What is your scientific interest?

5) Any Machine Learning experience?

1) What are your favorite hobbies?

2) What animal inspired you the most?

3) How long is your commute to the university?

Data Mining

Data mining

Process of discovering patterns, correlations, trends, and useful information from large sets of data,

- Using techniques from machine learning, statistics, and database systems.
- The ultimate goal of data mining is to extract valuable information from data and transform it into an understandable structure for further use.

The Process of Data Mining

1. Defining the Problem

The first step in any data mining project is to understand the objectives and requirements.

2. Data Gathering

The second phase covers data collection and exploration. An examination of the data collected will give you an idea of how accurate the fit is to be a base to address your business issue.

3. Data Preparation

The data preparation phase covers tasks such as table, case, and attribute selection. It also includes data cleansing and transformation, duplicate removal, standardizing input titles, and other data checking.

4. Model Building and Evaluation

various modeling techniques are chosen and applied, and parameters are calibrated to the optimum levels. Evaluating again at this point, how the model addresses the business issue is a good idea.

5. Model Deployment

In the final deployment stage, can include the applying the model to any new data, extracting model details, integrating models in applications and more.

Some common data mining techniques and methods:

1. Classification

- Assigns items in a dataset to target categories or classes.
- Decision Trees, Random Forest, Support Vector Machines (SVM), Neural Networks.
- Email spam filtering, customer segmentation, disease diagnosis.

2. Clustering

- Groups a set of objects in a way that objects in the same group (cluster) are more similar to each other than to those in other groups.
- K-Means, Hierarchical Clustering, DBSCAN.
- Market segmentation, social network analysis, astronomical data analysis.

3. Association Rule Learning

- Discover interesting relations between variables in large databases.
- Apriori algorithm, Eclat algorithm, FP-Growth.
- Market basket analysis, cross-selling strategies, catalog design.

4. Regression

- Predicts a numerical value based on inputs.
- Linear Regression, Logistic Regression, Polynomial Regression.
- Real estate pricing, risk assessment in finance, forecasting sales.

5. Anomaly Detection (Outlier Detection)

- Identifies unusual patterns that do not conform to expected behavior.
- Statistical methods, Proximity-based methods, Clustering-based anomaly detection.
- Fraud detection, network security, fault detection.

6. Dimensionality Reduction

- Reduces the number of random variables to consider.
- Principal Component Analysis (PCA), Singular Value Decomposition (SVD), t-Distributed Stochastic Neighbor Embedding (t-SNE).
- Data visualization, noise reduction, feature selection.

7. Ensemble Methods

- Combine predictions from multiple machine learning algorithms to produce better predictive performance.
- Bagging, Boosting, Stacking.
- Improving prediction accuracy in various fields, including competition science and risk management.

8. Neural Networks and Deep Learning

- Capable of learning complex patterns.
- Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Autoencoders.
- Image recognition, speech recognition, natural language processing.

9. Time Series Analysis

- Analyzing time-ordered data points to extract meaningful statistics and other characteristics.
- Autoregression (AR), Moving Average (MA), ARIMA.
- Stock market analysis, economic forecasting, weather prediction.

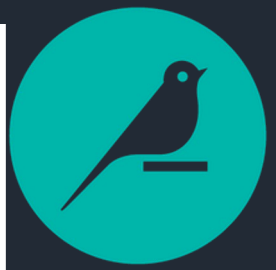
10. Text Mining and Natural Language Processing

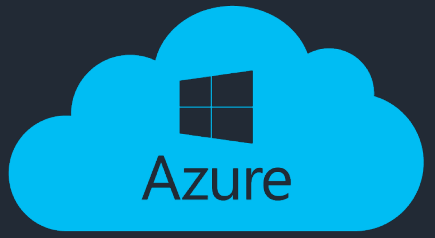
- Extracts useful information and insights from text data.
- Tokenization, Stemming, Lemmatization, TF-IDF, Sentiment Analysis.
- Sentiment analysis, topic modeling, document classification.

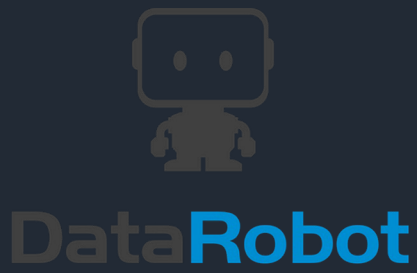
Data mining software and tools

Tools for data mining include:
Alteryx, AWS, Databricks, Dataiku, DataRobot, Google, H2O.ai, IBM, Knime, Microsoft, Oracle, RapidMiner, SAS Institute and Tibco Software, among others.

A variety of free open source technologies can also be used to mine data, including DataMelt, Elki, Orange, Rattle, scikit-learn and Weka.



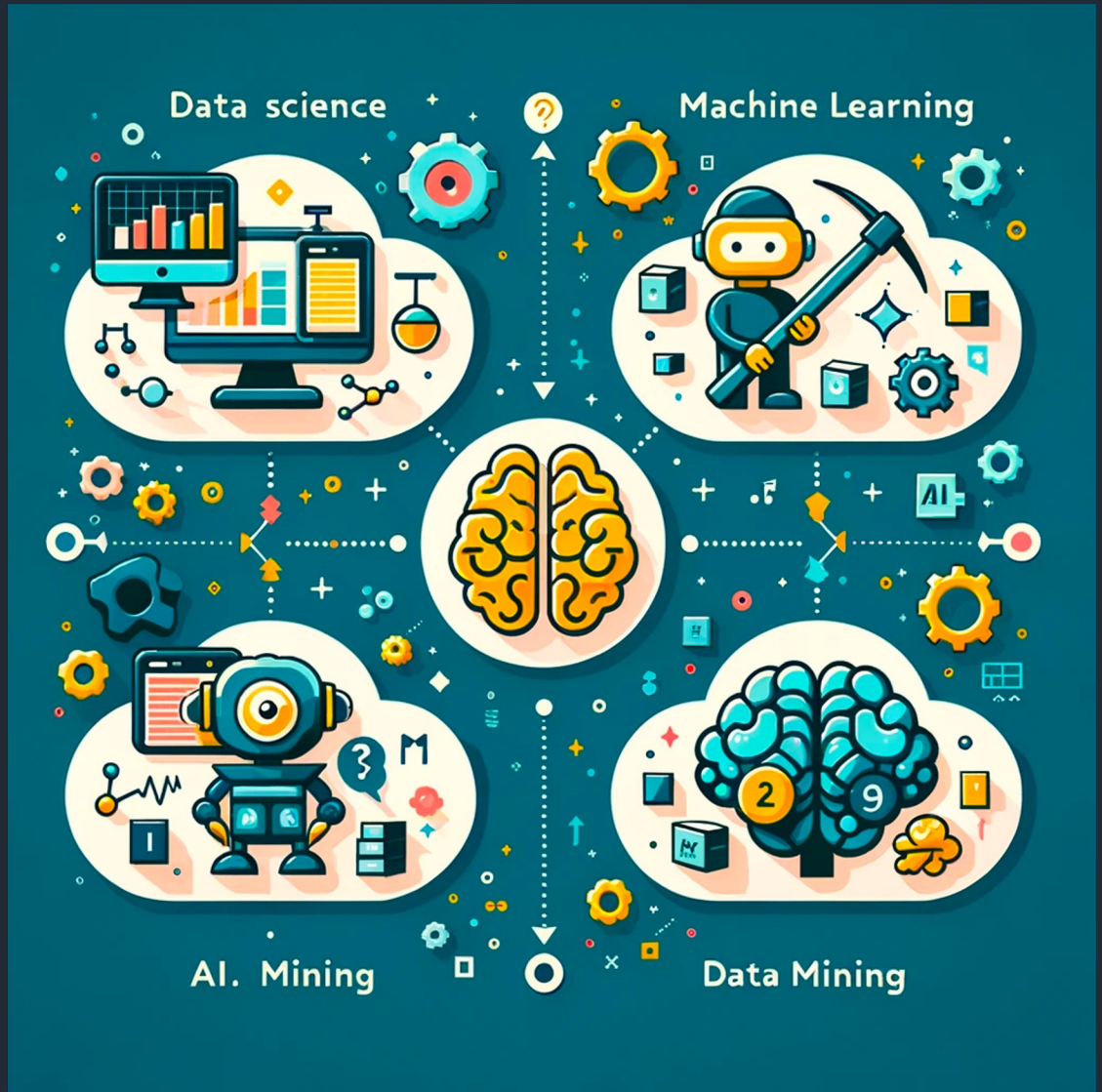




Challenges in Data Mining:

- **Data Quality:** The effectiveness of data mining is heavily dependent on the quality of the data. Poor data quality can lead to inaccurate conclusions.
- **Privacy and Security:** Ensuring the privacy and security of data is a major concern, especially with the mining of sensitive personal or financial information.
- **Complex and High-Dimensional Data:** Handling and processing large, complex datasets can be computationally challenging.
- **Ethical Concerns:** Ethical issues can arise over the use of personal data in data mining, especially without consent.

What is the difference between Data Science, Machine Learning, AI and Data mining?



- **Data Science vs. Machine Learning:** Data Science is a broader field that includes data preprocessing, analysis, and visualization, as well as the use of algorithms (which includes machine learning). Machine Learning is specifically about learning from data to make predictions or decisions.
- **Data Mining vs. Data Science:** Data Mining is focused on finding patterns in data. It's a part of Data Science, which also includes many other aspects like data cleansing, preparation, and analysis.

Fallacies of Data Mining

Fallacy 1: The data mining process is autonomous, requiring little or no human oversight.

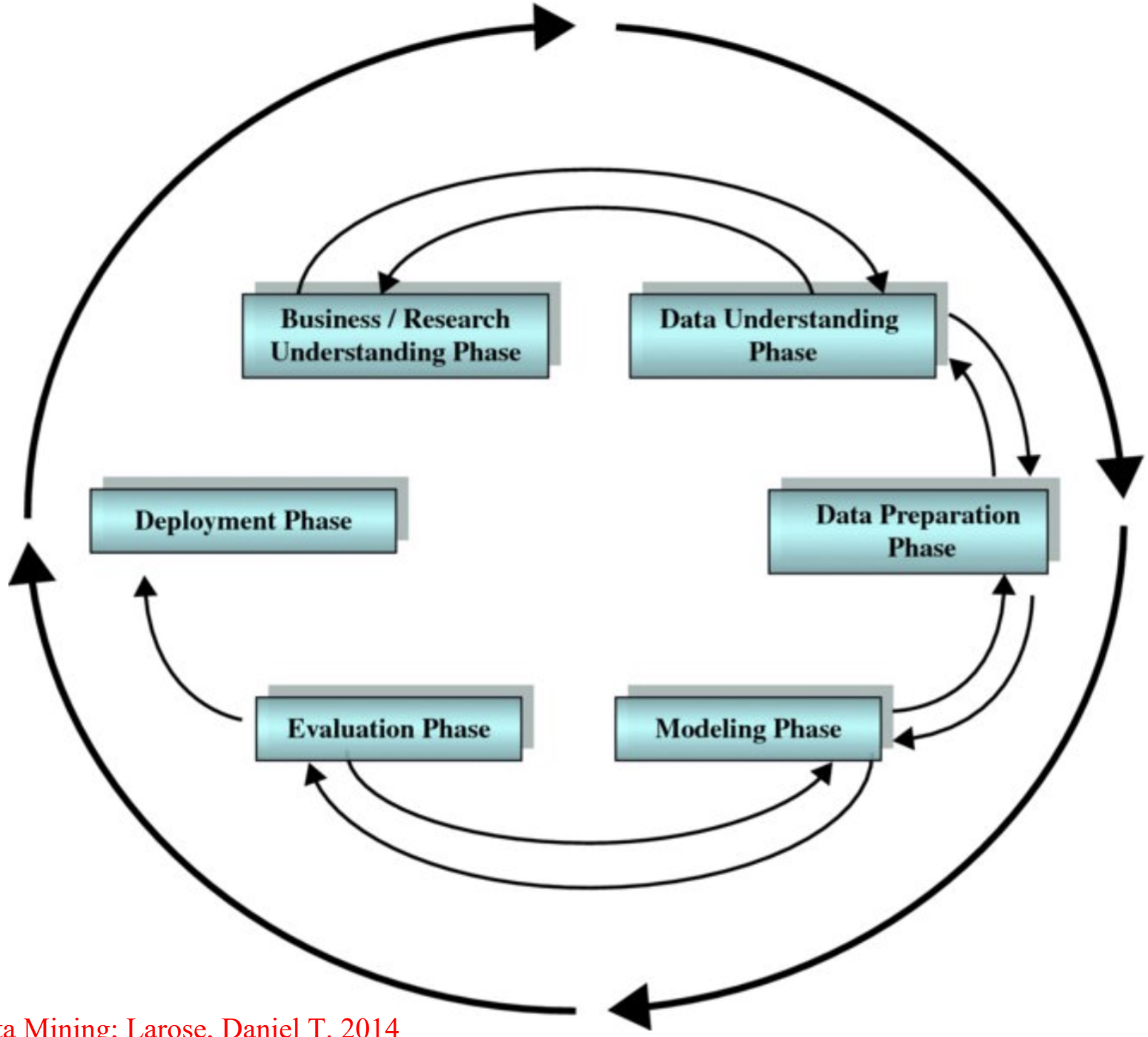
Fallacy 2: Data mining software packages are intuitive and easy to use.

Fallacy 3: Data mining will identify the causes of our business or research problems.

Fallacy 4: Data mining will automatically clean up our messy database.

Fallacy 5: Data mining always provides positive results.

Data mining project has a life cycle consisting of six phases:

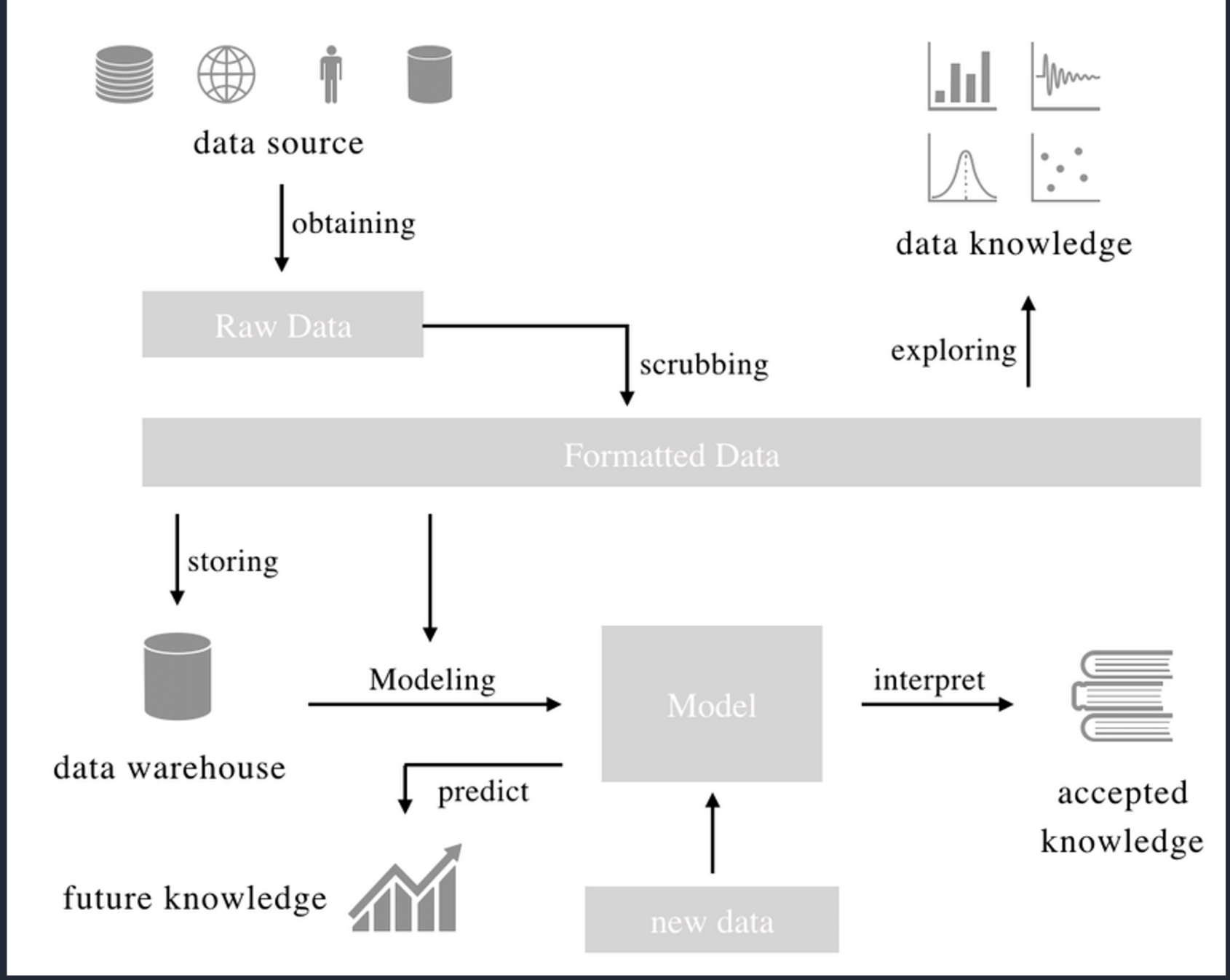


Discovering Knowledge in Data; An introduction to data Mining; Larose, Daniel T. 2014

Data Mining Pipelines

This is your dataset, how you start?

Temperature (°C)	*Humidity (%]*	*Rainfall (mm)*
16.49	97.53	0.8
16.88	-70	0.6
16.58	98.73	1
15.24	97.88	-0.9
15.84	99.71	0.3
250	97.47	
15.07	-300	0.1
16.19	89.07	-15



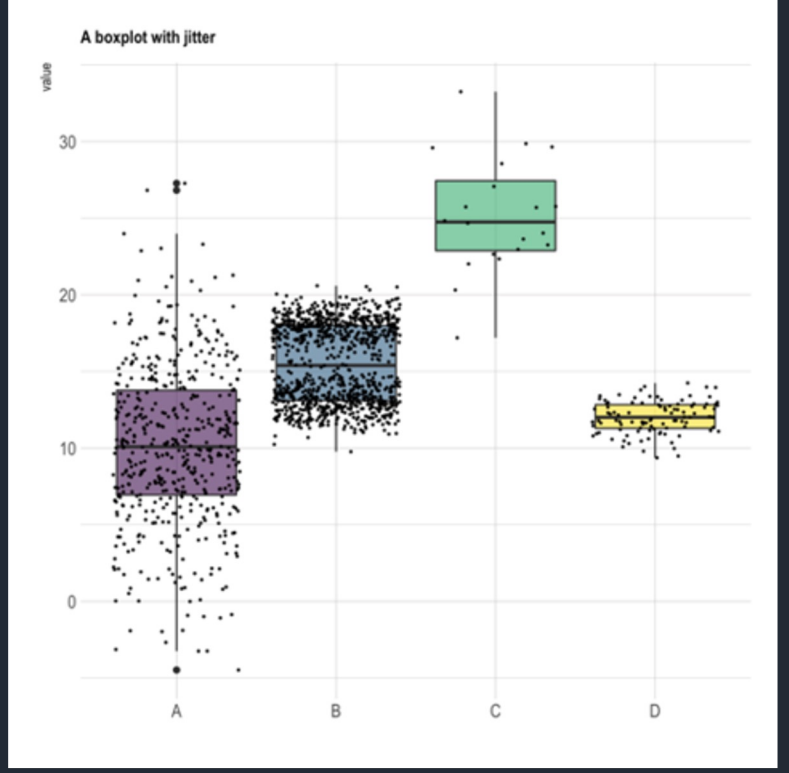
https://www.researchgate.net/figure/An-illustration-of-data-mining-pipeline_fig1_333915926

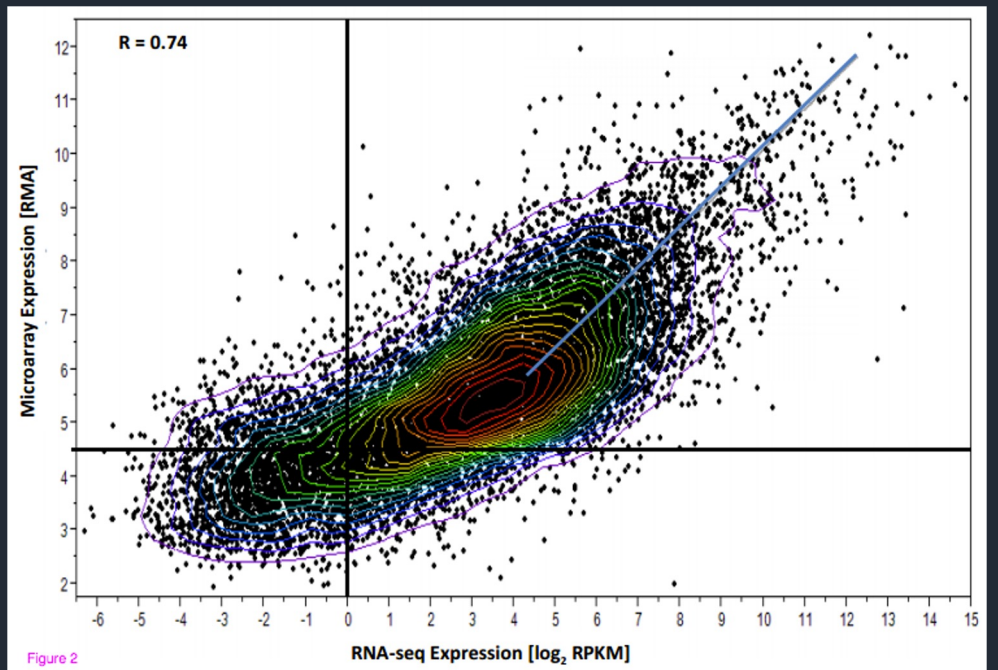
Data Collection

2. Define the Project/Business Questions

Data understanding

- What types of data?
- What do they look like?
- Statistics & visualization
- Similarity vs. dissimilarity
- General patterns vs. anomalies





Data Preprocessing

- **Potential issues with data**

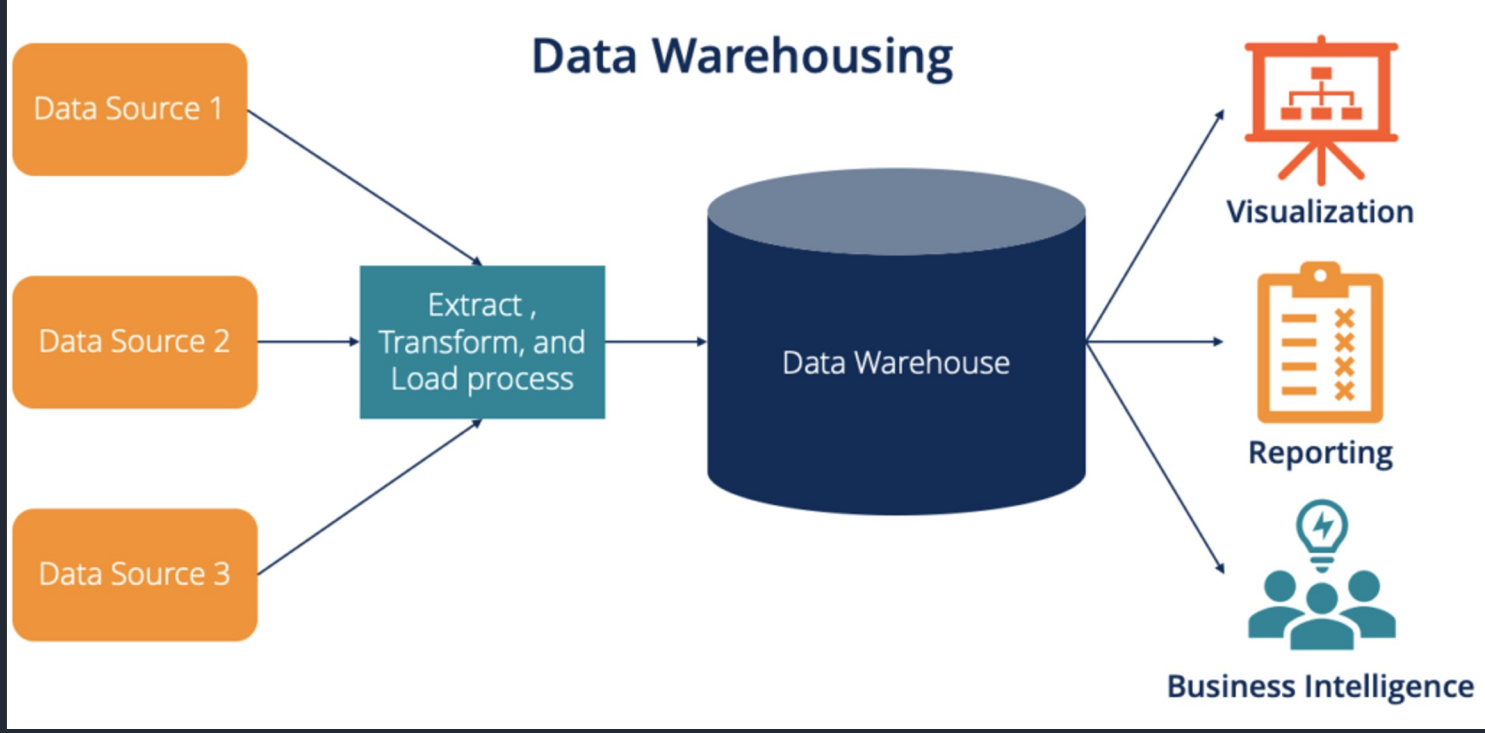
E.g., missing data, errors, inconsistency

- **Preparing data for the mining process**

Data cleaning, integration, transformation, reduction

- **No good data, no good data mining!**

Data Warehousing



Data warehouse

- vs. operational data

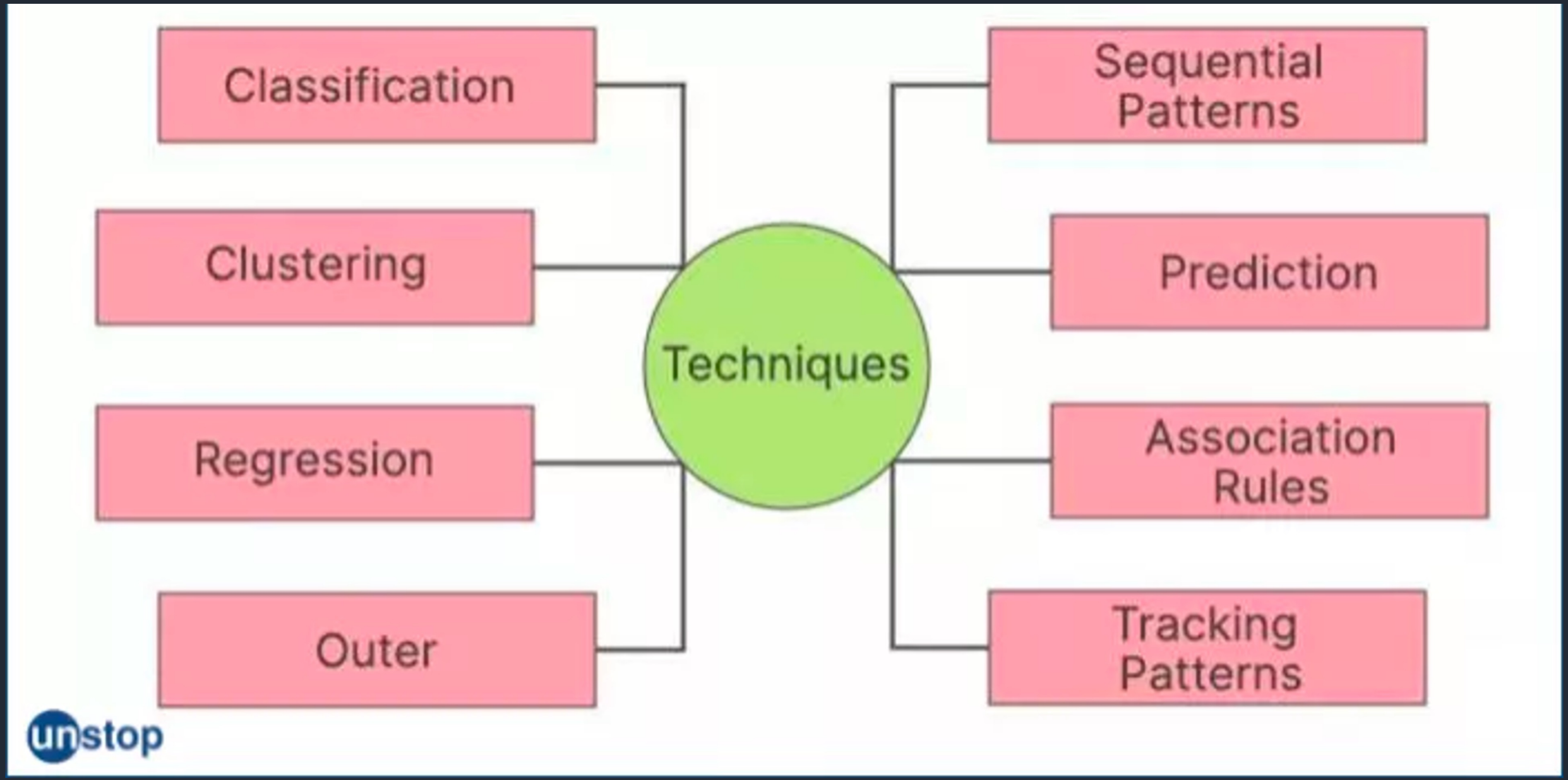
Data cube & OLAP

- Multi-dimensional data management

Data warehouse architecture

A data warehouse is **a central repository of information**

Data Modeling



unstop

Pattern evaluation

Finding Interesting Patterns from Data:

- Identifying patterns, trends, or relationships within a dataset.
- Classification, clustering, regression, and association rule mining.

Evaluation Metrics:

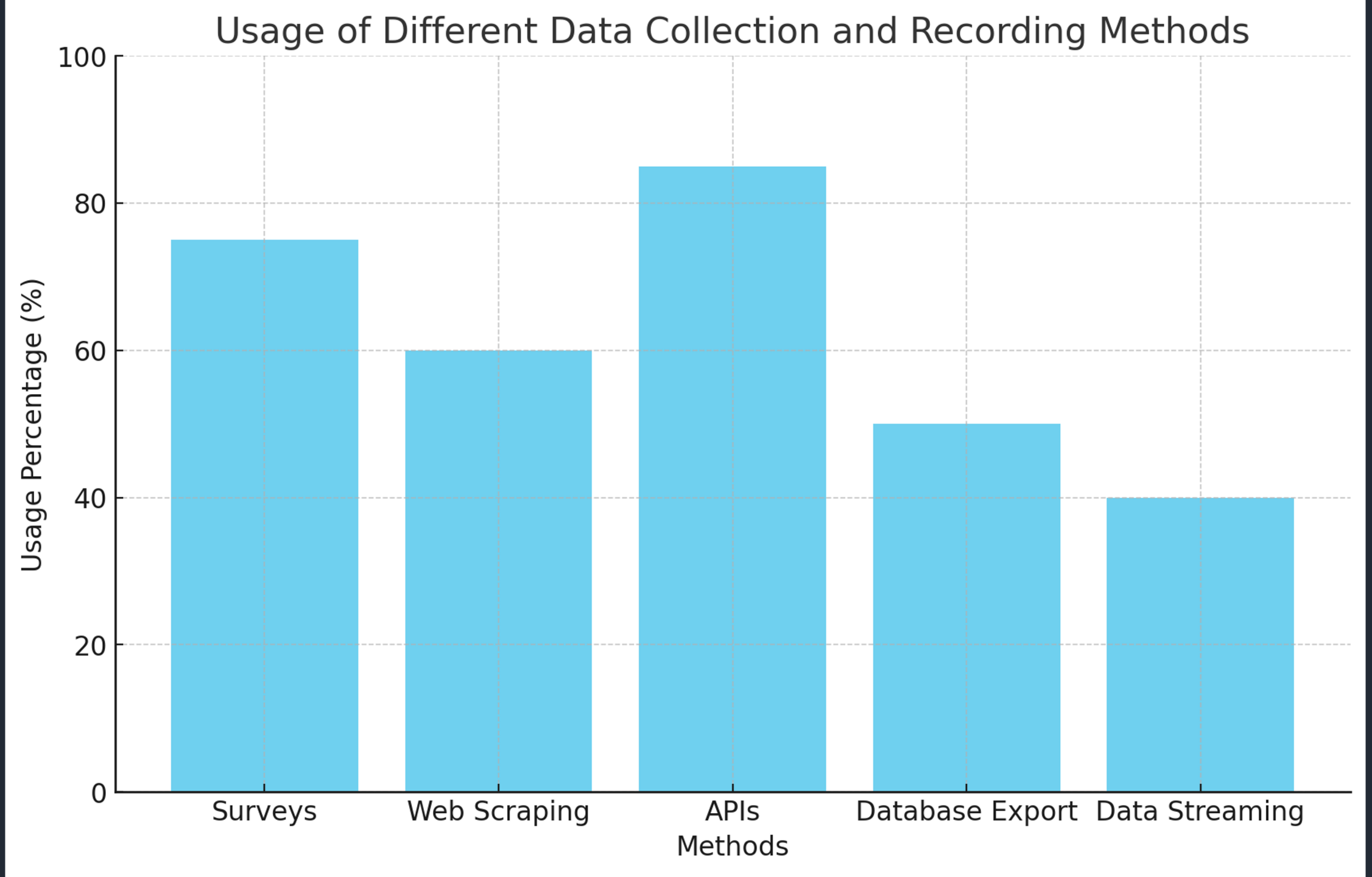
- Assessing the performance of the patterns or models discovered.
- Accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC).

Model Selection:

- Choosing the right algorithm or model for the specific problem and dataset.

Deep Dive into Key Topics

1.Data Collection and Recording


Usage of Different Data Collection and Recording Methods
Usage Percentage (%)
100
80
60
40
20
0
Surveys
Web Scraping
APIs
Database Export
Data Streaming
Methods

2. Define the Project/Business Questions

Six Types of Questions

- Descriptive
- Exploratory
- Inferential
- Predictive
- Causal
- Mechanistic

Characteristics of Each Question Type

1. Descriptive:

Summarizes characteristics of a dataset. Focuses on existing data without consideration of external factors.

1. **What is the average age of users on a particular social media platform?**
2. **How many sales transactions were completed in each month of the last year?**
3. **What is the average age of our customers?**
4. **What is the most common product purchased in our store?**

- **Why It's Important:** Provide a *foundational understanding of the data* and *set the stage for deeper analysis*. They help in *summarizing the data* and i*dentifying patterns* or *trends that may require further exploration*.

2. Exploratory:

Looking to find patterns, relationships, or structures within the data that were not previously known.

1. **Is there a relationship between the number of hours studied and scores on final exams among university students?**
2. **Do higher temperatures correlate with increased ice cream sales?**
3. **What groups of customers tend to buy similar products?"**

Why It's Important: These questions are vital for *discovering insights* that may not be immediately obvious. They help in shaping the direction of further analysis and research.

3. Inferential:

Inferential questions seek to make generalizations or inferences about a population based on a sample of data. These questions are rooted in statistics and involve estimating population parameters.

1. **What is the average income of our entire customer base based on a sample survey?**

1. **Is there a significant difference in spending between male and female customers?**

Why It's Important: Making it possible to draw conclusions about larger populations from smaller samples.

4. Predictive:

Focuses on predicting outcomes based on features in the dataset.

Emphasizes correlation rather than causation.

1. **Can we predict the future stock price of a company based on its past quarterly earnings?**
2. **Based on current health metrics, what is the likelihood that a patient will develop diabetes in the next year?**
3. **Which customers are likely to churn in the next month?**

Why It's Important: By predicting future outcomes, businesses can strategize accordingly to maximize opportunities or mitigate risks.

5. Causal:

Investigates how changes in one variable affect another. Often requires controlled experiments or careful analysis of observational data.

1. **Does implementing a new employee training program lead to increased productivity in the workplace?**
2. **What is the effect of reducing class size on student performance in primary schools?**

Why It's Important: Understanding causality is key to making informed decisions and implementing changes that will have the desired effect.

6. Mechanistic:

Mechanistic questions go deeper than causal questions by exploring the underlying mechanisms or processes that lead to a particular outcome. They seek to explain how and why something happens.

1. **How does changing the pH level in soil affect the growth rate of a specific plant species?**
2. **What impact does altering the wing design have on the fuel efficiency of an airplane?**
3. **How does customer satisfaction lead to increased sales?**

Why It's Important: Mechanistic questions provide a deeper understanding of the phenomena, which can lead to more targeted and effective interventions or innovations.

Characteristics of a Good Question

- Saves time, money, and frustration.
- Sets the direction for the entire process.
- Define audience: collaborators, scientific community, funders, etc.
- Ensure relevance to their interests and needs.
- Not Already Answered Conduct literature review. Consult experts for novelty assurance.
- Ability to explain expected outcomes.
- Ensure feasibility of data collection and analysis.
- Simplifies data collection and analysis.

Importance of Identifying Question Types

- Each type of question is associated with a specific analytical approach, data requirements, and interpretation of results.
- Misidentification can lead to inappropriate methodologies, misleading conclusions, and ultimately, unsuccessful research outcomes.

Importance of correctly identifying Questions, and the potential consequences of misidentification:

1. Descriptive Questions

Example: What are the average sales figures for our product across different regions?

Effect of Misidentification:

If a descriptive question is misidentified as an exploratory or inferential one, you might incorrectly apply statistical tests or attempt to infer relationships that don’t exist, leading to confusion or incorrect conclusions about the basic data characteristics.

Predictive Questions

- **Example:** What will be the next quarter's sales figures based on current trends?

Effect of Misidentification:

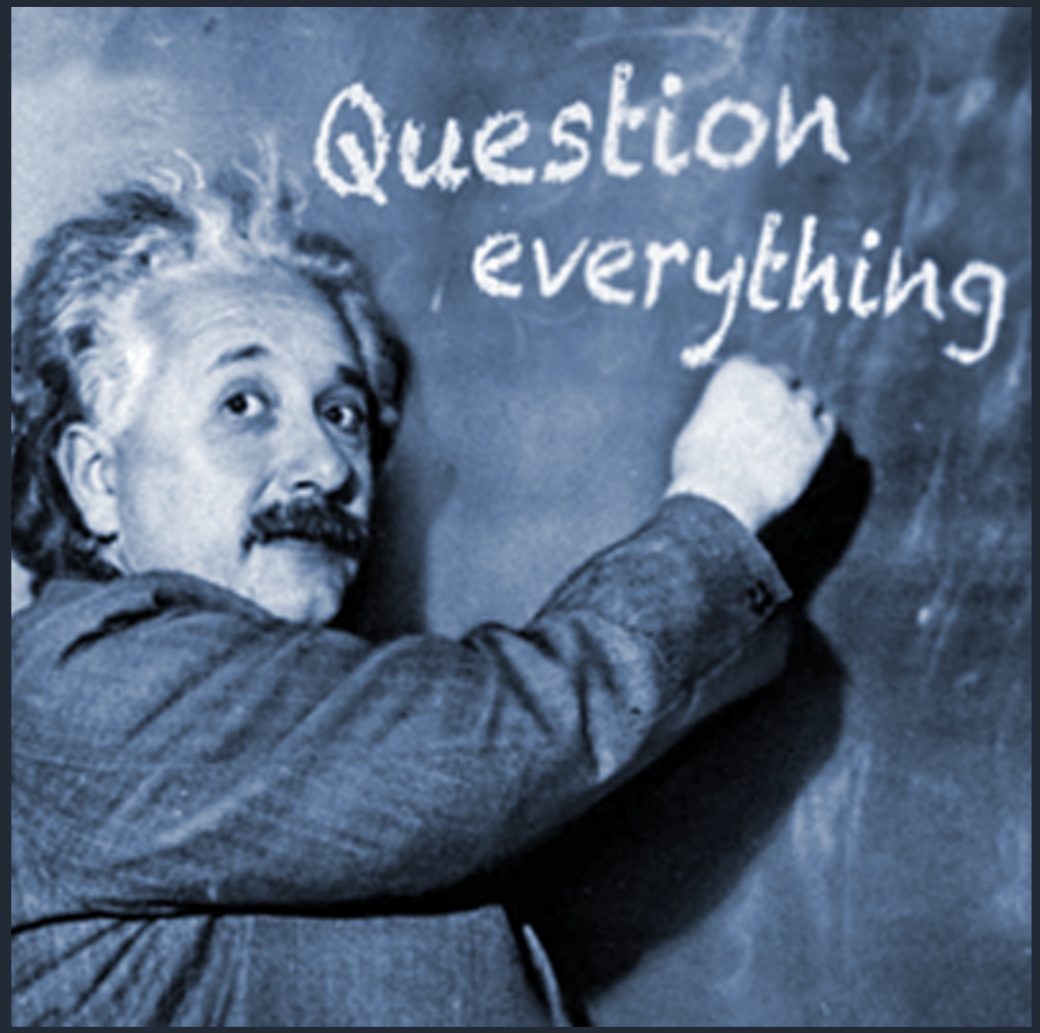
- Misidentifying a predictive question as descriptive or causal could result in failing to develop or apply appropriate predictive models, leading to inaccurate forecasts and poor decision-making.

Causal Questions

- **Example:** Does increasing our advertising budget lead to higher sales?

Effect of Misidentification:

If a causal question is misidentified as inferential or predictive, you might wrongly infer causality from correlation, leading to incorrect conclusions and possibly wasting resources on ineffective strategies.



Challenges


CHALLENGE ACCEPTED

Question 1

A company wants to understand the average age of its customers.

What type of question is this?

- A) Descriptive
- B) Exploratory
- C) Inferential
- D) Predictive
- E) Causal
- F) Mechanistic

Question 2

A researcher is investigating if there is a relationship between smoking and lung cancer.

What type of question is this?

- A) Descriptive
- B) Exploratory
- C) Inferential
- D) Predictive
- E) Causal
- F) Mechanistic

Question 3

A data analyst is trying to determine if the amount of money spent on advertising is associated with an increase in sales.

What type of question is this?

- A) Descriptive
- B) Exploratory
- C) Inferential
- D) Predictive
- E) Causal

Question 4

A data scientist is examining a large dataset to find patterns and relationships that were not previously known.

What type of question is this?

- A) Descriptive
- B) Exploratory
- C) Inferential
- D) Predictive
- E) Causal
- F) Mechanistic

Question 5

A healthcare provider wants to know the proportion of patients who recovered from an illness within a year.

What type of question is this?

- A) Descriptive
- B) Exploratory
- C) Inferential
- D) Predictive
- E) Causal
- F) Mechanistic

Question 6

An engineer is studying how different levels of stress affect the mechanical properties of a new material.

What type of question is this?

- A) Descriptive
- B) Exploratory
- C) Inferential
- D) Predictive
- E) Causal

Knowing the types of questions in data science can significantly enhance your data science project in several ways:

1. Clear Objectives and Scope

2. Appropriate Method Selection

Different types of questions require different analytical methods and tools. For example:

- **Descriptive** questions might use summary statistics and visualization.
- **Inferential** questions might use hypothesis testing and confidence intervals.
- **Predictive** questions might require machine learning models.
- **Causal** questions might involve controlled experiments or advanced statistical techniques like regression analysis.
- **Exploratory** questions might use clustering or association rule learning.
- **Mechanistic** questions might require detailed scientific models and simulations.

3. Data Collection Strategy

Knowing the type of question informs your data collection strategy. For instance, causal questions may require experimental or longitudinal data, while descriptive questions may only need cross-sectional data.

4. Efficient Resource Allocation

For example, predictive modeling might require significant computational power, while exploratory analysis might need extensive data preparation and cleaning.

5. Targeted Communication

Stakeholders may be more interested in predictions, causal relationships, or simply understanding current trends, depending on their goals.

6. Improved Validation and Reliability

For instance, predictive models need to be validated on unseen data, while causal inferences may require robustness checks and sensitivity analyses.

Practical Example

Imagine you are working on a project to improve customer satisfaction for an online retailer. Knowing the question types can guide your approach:

- **Descriptive:** What is the current customer satisfaction score?
- **Exploratory:** What are the common complaints or issues raised by customers?
- **Inferential:** Do customers in different regions have significantly different satisfaction levels?
- **Predictive:** Can we predict which customers are likely to be dissatisfied based on their purchase history?
- **Causal:** Does faster shipping lead to higher customer satisfaction?
- **Mechanistic:** How does the customer service response time affect customer satisfaction?