**From Problem to Approach and From Requirements to Collection**

# **Welcome to the Course**

In this module, you will discover what makes data science interesting, learn what a data science methodology is, and why data scientists need a data science methodology. Next, you’ll gain more in-depth knowledge of the first two data science methodology stages: Business Understanding and Analytic Approach. You’ll discover how to identify considerations and steps needed to define the data requirements for decision tree classification during the Data Requirements stage. Next, learn about the processes and techniques data scientists use to assess data content, quality, and initial insights and how data scientists manage data gaps. Round out this week with practical hands-on experience learning how to approach the Business Understanding and the Analytic Approach stage tasks and the Data Requirements and Collection stage tasks for any data science problem.

**Learning Objectives**

* Apply the first four phases of the data science methodology to a case study.
* Compose clearly defined questions that address a business problem.
* Analyze a case study to determine data requirements.
* Apply the data science methodology to a case study.
* Determine data content, data formats, and data sources prior to data collection and data preparation phases.
* Create a decision tree to classify outcomes in a case study.
* Identify appropriate data sources to address a business problem.

## **Course Introduction**

* **What the Course is About**: It introduces you to the "Data Science Methodology" course, which teaches you how data scientists solve problems using a structured approach.
* **What You Will Learn**: You will learn the steps involved in data science, like understanding the problem, collecting data, preparing it, analyzing it, and using it to make decisions.
* **How the Course is Structured**: The course includes videos, readings, hands-on activities, and quizzes to help you practice what you learn.
* **Real-Life Example**: There will be a case study, possibly in healthcare, showing how these methods can improve patient care and decision-making.

**Summary**: The introduction video outlines the course's goal to teach you data science methodologies. By the end, you'll be able to apply these methods to real-world problems, enhancing your skills in data science.

## **Course Syllabus**

**Syllabus**



**Module 1: From Problem to Approach and From Requirements to Collection**

* Video: Course Introduction
* Reading: Helpful Tips for Course Completion
* Reading: Syllabus

**Lesson 1: From Problem to Approach**

* Video: Data Science Methodology Overview
* Video: Business Understanding
* Video: Analytic Approach
* Hands-on Lab: From Problem to Approach
* Reading: Lesson 1 Summary: From Problem to Approach
* Practice Quiz: From Problem to Approach
* Glossary: From Problem to Approach
* Graded Quiz: From Problem to Approach

**Lesson 2: From Requirements to Collection**

* Video: Data Requirements
* Video: Data Collection
* Hands-on Lab: From Requirements to Collection
* Reading: Lesson 2 Summary: From Requirements to Collection
* Practice Quiz: From Requirements to Collection
* Glossary: From Requirements to Collection
* Graded Quiz: From Requirements to Collection

**Module 2: From Understanding to Preparation and from Modeling to Evaluation**

**Lesson 1: From Understanding to Preparation**

* Video: Data Understanding
* Data Preparation - Concepts
* Data Preparation - Case Study
* Hands-on Lab: From Understanding to Preparation
* Reading: Lesson 1 Summary: From Understanding to Preparation
* Practice Quiz: From Understanding to Preparation
* Glossary: From Understanding to Preparation
* Graded Quiz: From Understanding to Preparation

**Lesson 2: From Modeling to Evaluation**

* Video: Modeling - Concepts
* Video: Modeling - Case Study
* Video: Evaluation
* Hands-on Lab: From Modeling to Evaluation
* Reading: Lesson 2 Summary:From Modeling to Evaluation
* Practice Quiz: From Modeling to Evaluation
* Glossary: From Modeling to Evaluation
* Graded Quiz: From Modeling to Evaluation

**Module 3: From Deployment to Feedback**

* Video: Deployment
* Video: Feedback
* Video: Storytelling
* Video: Course Summary
* Reading: Module 3 Summary:From Deployment to Feedback
* Practice Quiz: From Deployment to Feedback
* Glossary: From Deployment to Feedback
* Graded Quiz: From Deployment to Feedback

**Module 4: Final Project and Assessment**

**Final Project**

* Video: Introduction to CRISP-DM
* Reading: Final Assignment Overview
* Peer Review: Final Assignment

**Course Summary and Final Quiz**

* Reading: Review What You Learned
* Graded Quiz: Final Quiz

**Course Wrap Up**

* Reading: Congratulations and Next Steps
* Reading: Thanks from the Course Team
* Reading: IBM Digital Badge

# **Problem to Approach**

methodology:

Overview of Methodology

* **What is Methodology?**
  + A **methodology** is a system of methods used in a specific area of study.
  + It serves as a **guideline** for researchers, helping them make informed decisions during the scientific process.

Importance of Methodology in Data Science

* Data science combines **statistical analysis**, **technological expertise**, and **domain knowledge** to extract insights from large datasets.
* Despite advancements in computing power and data accessibility, challenges remain in understanding the questions posed and applying data effectively.
* A structured methodology helps address these challenges, ensuring a systematic approach to problem-solving.

John Rollins's Contributions

* The course is based on the methodology outlined by **John Rollins**, an experienced IBM Senior Data Scientist.
* His insights emphasize the importance of following a methodology for successful outcomes in data science.

10 Stages of Data Science Methodology

1. **Business Understanding**: Define the business problem and objectives.
2. **Analytic Approach**: Determine the analytical methods to be used.
3. **Data Requirements**: Identify what data is needed to address the problem.
4. **Data Collection**: Gather the necessary data from various sources.
5. **Data Understanding**: Analyze the data to understand its structure and quality.
6. **Data Preparation**: Clean and prepare the data for analysis.
7. **Modeling**: Develop models to analyze the data.
8. **Evaluation**: Assess the model's performance and its ability to answer the business question.
9. **Deployment**: Implement the model in a real-world setting.
10. **Feedback**: Gather feedback to refine the model and improve future analyses.

The Role of Questions

* **Questions are Central**: Asking the right questions is crucial for success in data science.
* Each stage of the methodology is driven by specific questions that guide the process.

Key Questions Aligned with the Stages

1. **Defining the Problem**:
   * What is the problem you are trying to solve?
   * How can you use data to answer this question?
2. **Organizing Data**:
   * What data do you need to answer the question?
   * Where is the data coming from, and how will you receive it?
   * Does the data collected accurately represent the problem?
   * What additional work is required to manipulate and work with the data?
3. **Validating the Approach**:
   * When applying data visualizations, do you see answers that address the business problem?
   * Does the data model answer the initial business question, or does it need adjustments?
   * Can you implement the model in practice?
   * Can you obtain constructive feedback from the data and stakeholders to address the business question?

Conclusion

* The video emphasizes that a well-defined data science methodology is essential for guiding data scientists in solving complex problems effectively.
* It highlights the importance of structured approaches, data collection forms, measurement strategies, and comparisons of data analysis methods relative to different research goals.

## **Business Understanding**

"Business Understanding" in Data Science Methodology:

Simplified Explanation:

* **Business Understanding** is the first step in data science. It means clearly defining the problem you need to solve before jumping into data analysis.
* Imagine your boss gives you a task with a tight deadline. You need to ask questions to fully understand what is required. This helps you know what data to use later.
* A clear question guides your analysis. For example, if a business owner asks how to reduce costs, you need to clarify if the goal is to improve efficiency or increase profits.
* After understanding the goal, you break it down into smaller objectives. This helps in organizing discussions and planning how to solve the problem.
* In a case study, a healthcare insurance provider wanted to know how to best use a limited budget for quality care. They worked with data scientists to define their goals and objectives before collecting data.
* They found that many patients were being readmitted to rehab centers, especially those with heart issues. The data scientists proposed a workshop to understand the problem better and set clear business requirements for their analysis.

Summary:

* **Business Understanding** is crucial in data science as it sets the foundation for effective analysis.
* Clearly defining the problem and understanding the goals helps in determining the right data to use.
* Engaging stakeholders and breaking down objectives leads to better planning and problem-solving.
* A case study illustrated how a healthcare provider used this approach to address patient readmissions and optimize budget allocation.

## **Analytic Approach**

the **analytic approach** in data science methodology, focusing on how to select the right analytic approach based on the question being asked. Here’s a simplified explanation and summary:

Simplified Explanation:

1. **Understanding the Question**: Before choosing an analytic method, it's important to clarify the question being asked. This helps in selecting the most suitable approach.
2. **Types of Analytic Approaches**:
   * **Predictive Models**: Used when you want to determine the likelihood of an action.
   * **Descriptive Approaches**: Useful for showing relationships or patterns in data.
   * **Statistical Analysis**: Applied when you need to count or classify responses (like yes/no).
   * **Machine Learning**: Helps in identifying trends and relationships in data without explicit programming.
3. **Case Study**: The video presents a case study using a **decision tree classification model** to predict patient readmission risks. This model helps clinicians understand which conditions lead to high-risk scores for patients.

Summary:

* The video emphasizes the importance of selecting the right analytic approach based on the specific question in data science.
* It covers various analytic methods, including predictive, descriptive, and classification models.
* A case study illustrates the use of a decision tree model to assess patient readmission risks, making it easier for non-data scientists to understand and apply.

## **Analytic Approach Based on the Question Type**

**Analytic Approach Based on the Question Type**

When choosing an analytic approach for a problem, the type of question you’re trying to answer greatly influences the methodology. Here are five common types of questions and corresponding analytic approaches:

**1. Descriptive Questions: “What is the current status?”**

**Approach: Descriptive Analytics**

**Question:** "What is the current status of our sales?"

**Techniques:**

* Data aggregation: Combining data from various sources into a unified view.
* Data mining: Extracting useful information from large datasets.
* Data visualization: Using visual tools to present data in an easily understandable format.

**Examples:**

* Summarizing sales data
* Creating dashboards
* Generating reports

**2. Diagnostic Questions: “Why did it happen?”**

**Approach: Diagnostic Analytics**

**Question:** "Why did our sales decline in the last quarter?"

**Techniques:**

* Drill-down: Exploring detailed data to find underlying causes.
* Data discovery: Identifying patterns and relationships in data.
* Correlation analysis: Assessing the relationship between different variables.

**Examples:**

* Identifying root causes of sales decline
* Analyzing customer complaints
* Understanding failure points in a process

**3. Predictive Questions: “What is likely to happen?”**

**Approach: Predictive Analytics**

**Question:** "What is our sales forecast for the next year?"

**Techniques:**

* Regression analysis: Predicting outcomes based on relationships between variables.
* Time series forecasting: Predicting future values based on past trends.
* Machine learning models: Using algorithms to predict future outcomes based on historical data.

**Examples:**

* Forecasting sales
* Predicting customer churn
* Estimating future demand

**4. Prescriptive Questions: “What should we do?”**

**Approach: Prescriptive Analytics**

**Question:** "What should we do to increase website traffic?"

**Techniques:**

* Optimization models: Finding the best solution from a set of alternatives.
* Simulation: Modeling scenarios to predict outcomes.
* Decision analysis: Evaluating and comparing different decisions.

**Examples:**

* Recommending inventory levels
* Optimizing marketing campaigns
* Determining pricing strategies

**5. Classification Questions: “Which category does this belong to?”**

**Approach: Classification (Supervised Learning)**

**Question:** "Which category does this data point belong to?"

**Techniques:**

* Logistic regression: Predicting the probability of a categorical outcome.
* Decision trees: Splitting data into branches to classify it.
* Support vector machines: Finding the best boundary to separate categories.
* Neural networks: Using interconnected nodes to classify data.

**Examples:**

* Email spam detection
* Image classification
* Disease diagnosis

Understanding these different types of questions and the corresponding analytic approaches can help you unlock your data's true potential.

## Activity

**Business Understanding: Asking Questions**

**Business Goal**

**The company's e-commerce business goal is to optimize its pricing strategy to maximize revenue and profitability. By leveraging data science, the company aims to identify patterns in historical sales data, pricing changes, and customer behavior to make informed decisions on pricing and promotional strategies.**

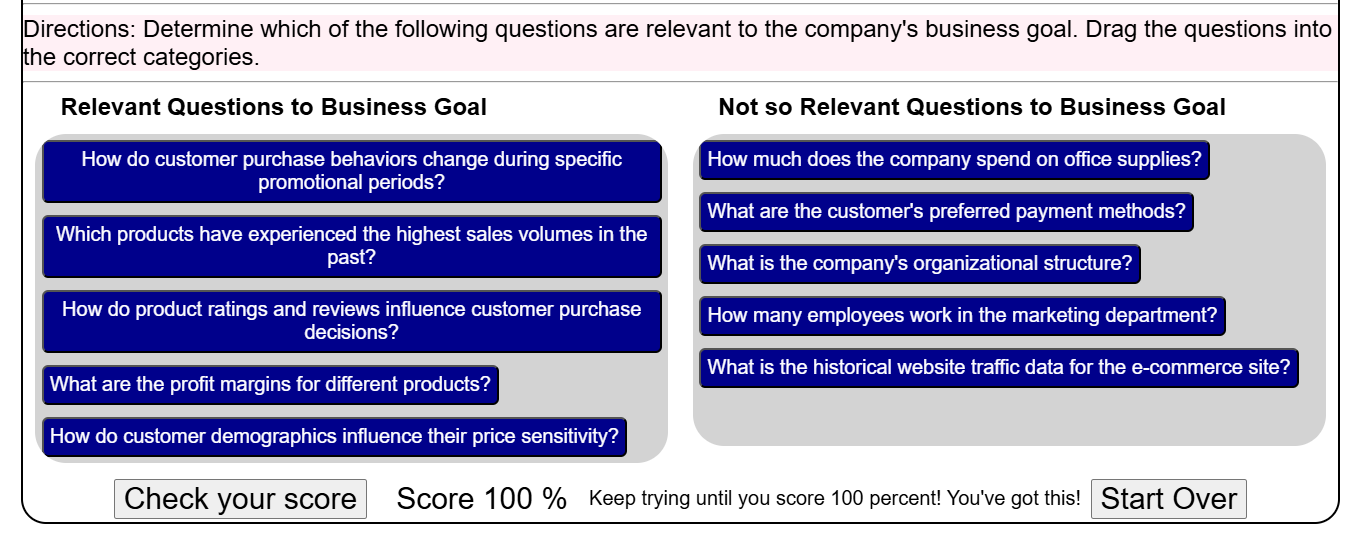
Directions: Determine which of the following questions are relevant to the company's business goal. Drag the questions into the correct categories.

**Relevant Questions to Business Goal**

**Not so Relevant Questions to Business Goal**

How do customer purchase behaviors change during specific promotional periods?Which products have experienced the highest sales volumes in the past?How do product ratings and reviews influence customer purchase decisions?What are the profit margins for different products?How do customer demographics influence their price sensitivity?

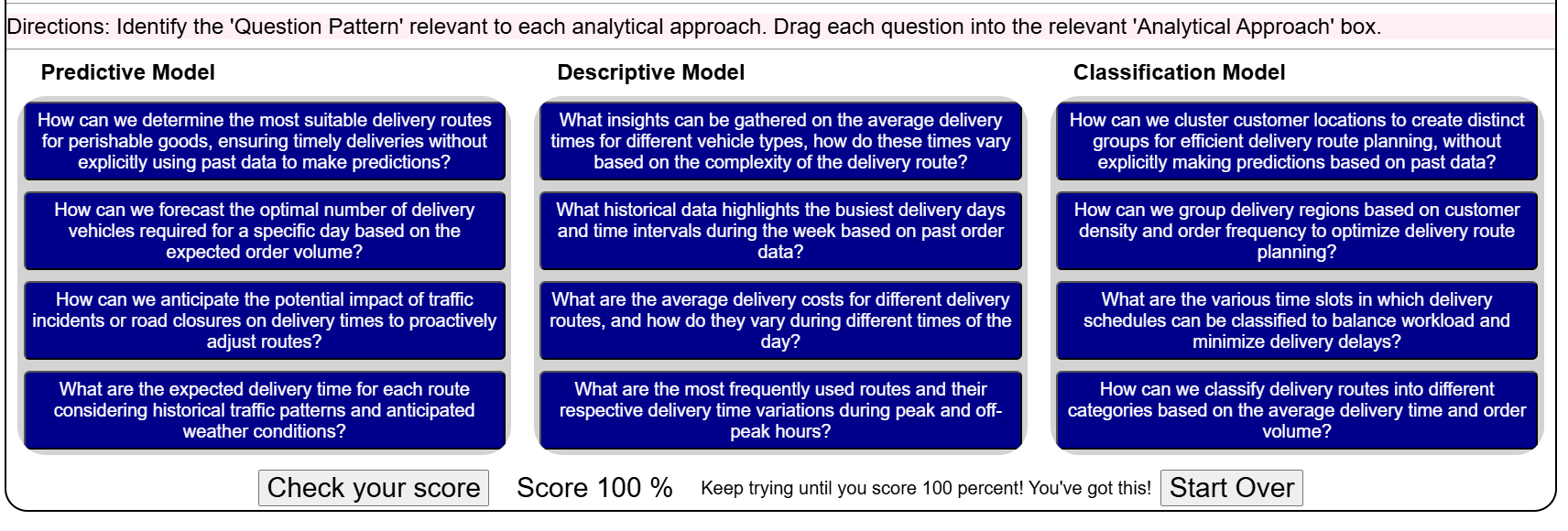
How much does the company spend on office supplies?What are the customer's preferred payment methods?What is the company's organizational structure?How many employees work in the marketing department?What is the historical website traffic data for the e-commerce site?



**Business Goal**

**A transportation company aims to optimize its delivery routes and schedules to minimize costs and improve delivery efficiency. The company wants to use data science to identify the most optimal routes and delivery time windows based on historical delivery data and external factors such as traffic and weather conditions.**

**Various questions are targeted by data scientist to achieve this business goal**



## **Lesson Summary**

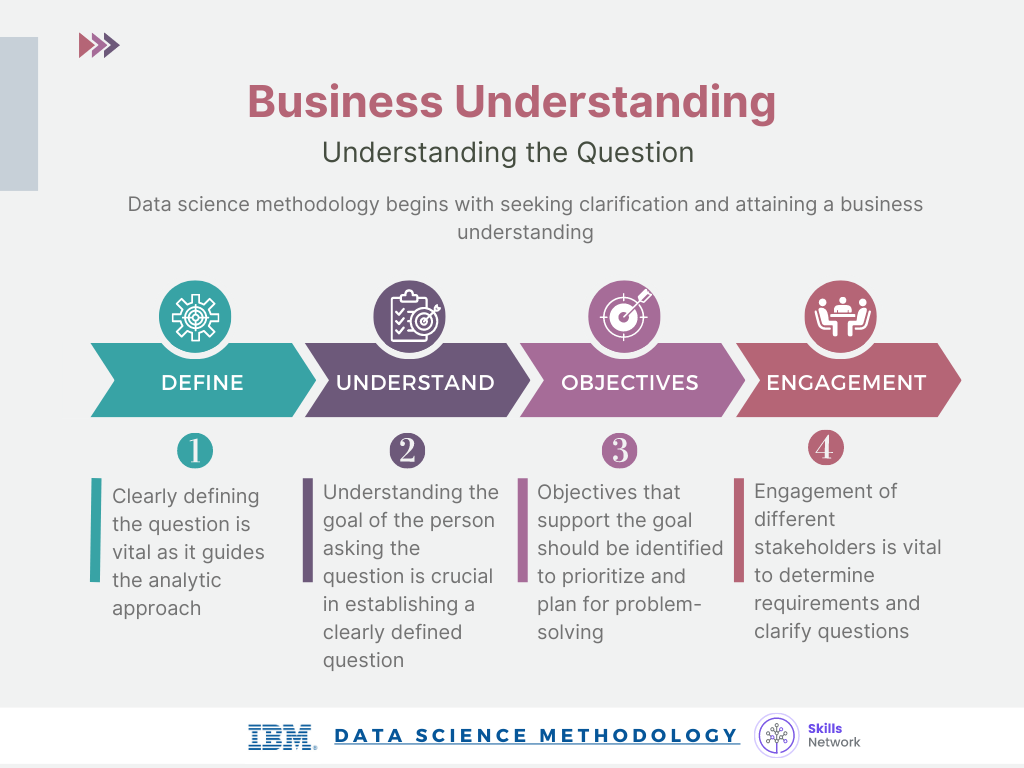
**Lesson summary**

**Module 1 Lesson 1: From Problem to Approach**

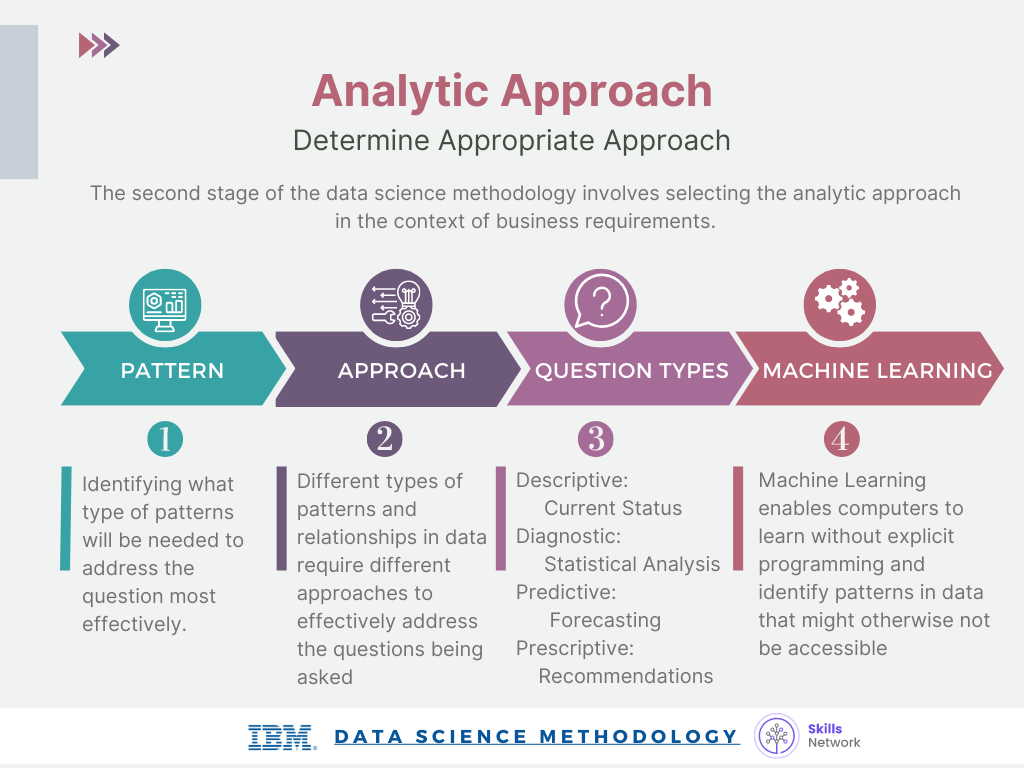


**Congratulations! You have completed this lesson. At this point in the course, you know:**

* Foundational methodology, a cyclical, iterative data science methodology developed by John Rollins, consists of 10 stages, starting with Business Understanding and ending with Feedback.
* The primary goal of the Business Understanding stage is to understand the business problem and determine the data needed to answer the core business question.



* During the Analytic Approach stage, you can choose from descriptive, diagnostic, predictive, and prescriptive analytic approaches, whether to use machine learning with clustering associations.



* Decision tree classification is a predictive analytics approach that's easy for non-data scientists to implement

## **Glossary**

**Glossary: From Problem to Approach**

Welcome! This alphabetized glossary contains many of the terms you'll find within this lesson. These terms are important for you to recognize when working in the industry, when participating in user groups, and when participating in other certificate programs.

|  |  |
| --- | --- |
| **Term** | **Definition** |
| **Analytic Approach** | The process of selecting the appropriate method or path to address a specific data science question or problem. |
| **Analytics** | The systematic analysis of data using statistical, mathematical, and computational techniques to uncover insights, patterns, and trends. |
| **Business Understanding** | The initial phase of data science methodology involves seeking clarification and understanding the goals, objectives, and requirements of a given task or problem. |
| **Clustering Association** | An approach used to learn about human behavior and identify patterns and associations in data. |
| **Cohort** | A group of individuals who share a common characteristic or experience is studied or analyzed as a unit. |
| **Cohort study** | An observational study where a group of individuals with a specific characteristic or exposure is followed over time to determine the incidence of outcomes or the relationship between exposures and outcomes. |
| **Congestive Heart Failure (CHF)** | A chronic condition in which the heart cannot pump enough blood to meet the body's needs, resulting in fluid buildup and symptoms such as shortness of breath and fatigue. |
| **CRISP-DM** | Cross-Industry Standard Process for Data Mining is a widely used methodology for data mining and analytics projects encompassing six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. |
| **Data analysis** | The process of inspecting, cleaning, transforming, and modeling data to discover useful information, draw conclusions, and support decision-making. |
| **Data cleansing** | The process of identifying and correcting or removing errors, inconsistencies, or inaccuracies in a dataset to improve its quality and reliability |
| **Data science** | An interdisciplinary field that combines scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. |
| **Data science methodology** | A structured approach to solving business problems using data analysis and data-driven insights. |
| **Data scientist** | A professional using scientific methods, algorithms, and tools to analyze data, extract insights, and develop models or solutions to complex business problems. |
| **Data scientists** | Professionals with data science and analytics expertise who apply their skills to solve business problems. |
| **Data-Driven Insights** | Insights derived from analyzing and interpreting data to inform decision-making |
| **Decision tree** | A supervised machine learning algorithm that uses a tree-like structure of decisions and their possible consequences to make predictions or classify instances. |
| **Decision Tree Classification Model** | A model that uses a tree-like structure to classify data based on conditions and thresholds provides predicted outcomes and associated probabilities. |
| **Decision Tree Classifier** | A classification model that uses a decision tree to determine outcomes based on specific conditions and thresholds. |
| **Decision-Tree Model** | A model used to review scenarios and identify relationships in data, such as the reasons for patient readmissions |
| **Descriptive approach** | An approach used to show relationships and identify clusters of similar activities based on events and preferences |
| **Descriptive modeling** | Modeling technique that focuses on describing and summarizing data, often through statistical analysis and visualization, without making predictions or inferences |
| **Domain knowledge** | Expertise and understanding of a specific subject area or field, including its concepts, principles, and relevant data |
| **Goals and objectives** | The sought-after outcomes and specific objectives that support the overall goal of the task or problem. |
| **Iteration** | A single cycle or repetition of a process often involves refining or modifying a solution based on feedback or new information. |
| **Iterative process** | A process that involves repeating a series of steps or actions to refine and improve a solution or analysis. Each iteration builds upon the previous one. |
| **Leaf** | The final nodes of a decision tree where data is categorized into specific outcomes. |
| **Machine Learning** | A field of study that enables computers to learn from data without being explicitly programmed, identifying hidden relationships and trends. |
| **Mean** | The average value of a set of numbers is calculated by summing all the values and dividing by the total number of values. |
| **Median** | When arranged in ascending or descending order, the middle value in a set of numbers divides the data into two equal halves. |
| **Model (Conceptual model)** | A simplified representation or abstraction of a real-world system or phenomenon used to understand, analyze, or predict its behavior. |
| **Model building** | The process of developing predictive models to gain insights and make informed decisions based on data analysis. |
| **Pairwise comparison (correlation)** | A statistical technique that measures the strength and direction of the linear relationship between two variables by calculating a correlation coefficient. |
| **Pattern** | A recurring or noticeable arrangement or sequence in data can provide insights or be used for prediction or classification. |
| **Predictive model** | A model used to determine probabilities of an action or outcome based on historical data. |
| **Predictors** | Variables or features in a model that are used to predict or explain the outcome variable or target variable. |
| **Prioritization** | The process of organizing objectives and tasks based on their importance and impact on the overall goal. |
| **Problem solving** | The process of addressing challenges and finding solutions to achieve desired outcomes. |
| **Stakeholders** | Individuals or groups with a vested interest in the data science model's outcome and its practical application, such as solution owners, marketing, application developers, and IT administration. |
| **Standard deviation** | A measure of the dispersion or variability of a set of values from their mean; It provides information about the spread or distribution of the data. |
| **Statistical analysis** | Stand deviations are applied to problems that require counts, such as yes/no answers or classification tasks. |
| **Statistics** | The collection, analysis, interpretation, presentation, and organization of data to understand patterns, relationships, and variability in the data. |
| **Structured data (data model)** | Data organized and formatted according to a predefined schema or model and is typically stored in databases or spreadsheets. |
| **Text analysis data mining** | The process of extracting useful information or knowledge from unstructured textual data through techniques such as natural language processing, text mining, and sentiment analysis. |
| **Threshold value** | The specific value used to split data into groups or categories in a decision tree. |

# **From Requirements to Collection**

## **Data Requirements**

Simplified Explanation:

* **Cooking Analogy**: Just like making a spaghetti dinner requires specific ingredients, data science requires specific data to solve problems.
* **Identifying Data Needs**: Before collecting data, data scientists must determine what data is needed, how to collect it, and how to prepare it for analysis.
* **Case Study Example**: A case study is presented where a data scientist needs to analyze patients with congestive heart failure. They set criteria for selecting patients to ensure the data is relevant and accurate.
  + **Criteria**:
    - Patients must be admitted as in-patients.
    - They should have a primary diagnosis of congestive heart failure for a year.
    - Continuous enrollment for at least six months before admission is required.
* **Data Preparation**: The data must be organized into a specific format, with one record per patient, including all relevant medical history.

Summary:

* The video emphasizes the importance of defining **data requirements** before data collection in data science.
* It uses a case study to illustrate how to select appropriate data based on specific criteria to ensure accurate analysis.
* Proper data preparation is crucial for effective data analysis and modeling.

## **Data Collection**

Simplified Explanation:

1. **Data Collection**: After deciding what data is needed, data scientists gather the required information, similar to shopping for ingredients to cook a meal.
2. **Assessing Data Needs**: They check if they have all the necessary data or if they need to adjust their requirements based on what's available.
3. **Identifying Gaps**: If some data is missing, they plan how to fill those gaps later, possibly after analyzing initial results.
4. **Collaboration**: Database Administrators (DBAs) and programmers work together to collect and organize the data, ensuring it’s ready for analysis.
5. **Automation**: They may also discuss ways to automate data collection processes to make it easier and faster in the future.

Summary:

* The video explains the importance of the data collection stage in data science.
* It highlights the need to assess data requirements, identify gaps, and collaborate with DBAs and programmers.
* The goal is to ensure that the collected data is ready for the next stage of analysis.

## **Summary**

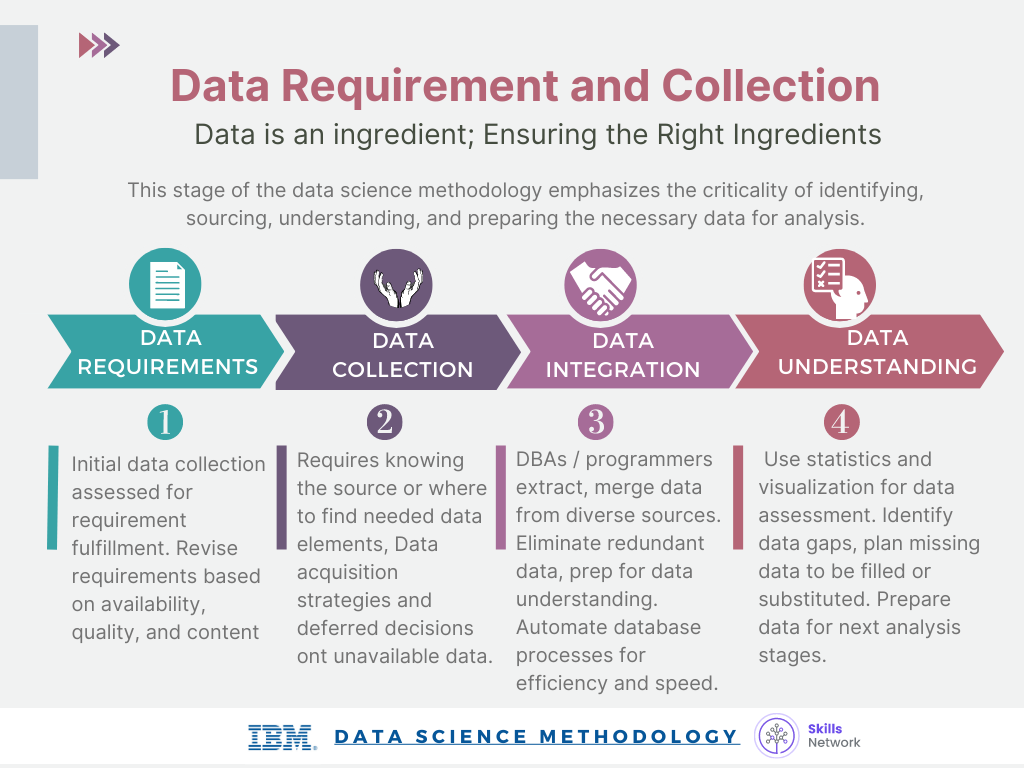
**Lesson summary**

**Module 1 Lesson 2: From Requirements to Collection**



**Congratulations! You have completed this lesson. At this point in the course, you know:**

* Data Requirements stage tasks include identifying the correct and necessary data content, data formats, and data sources for the specific analytical approach.
* During the Data Collection stage, expert data scientists meticulously revise data requirements and make critical decisions regarding the quantity and quality of data.



* Data scientists apply descriptive statistics and visualization techniques to thoroughly assess the content, quality, and initial insights gained from the collected data, identify gaps, and determine if new data is needed or to substitute existing data.

## **Glossary**

**Glossary: From Requirements to Collection**

Welcome! This alphabetized glossary contains many of the terms you'll find within this lesson. These terms are important for you to recognize when working in the industry, when participating in user groups, and when participating in other certificate programs.

|  |  |
| --- | --- |
| **Term** | **Definition** |
| **Analytics team** | A group of professionals, including data scientists and analysts, responsible for performing data analysis and modeling. |
| **Data collection** | The process of gathering data from various sources, including demographic, clinical, coverage, and pharmaceutical information. |
| **Data integration** | The merging of data from multiple sources to remove redundancy and prepare it for further analysis. |
| **Data Preparation** | The process of organizing and formatting data to meet the requirements of the modeling technique. |
| **Data Requirements** | The identification and definition of the necessary data elements, formats, and sources required for analysis. |
| **Data Understanding** | A stage where data scientists discuss various ways to manage data effectively, including automating certain processes in the database. |
| **DBAs (Database Administrators)** | The professionals who are responsible for managing and extracting data from databases. |
| **Decision tree classification** | A modeling technique that uses a tree-like structure to classify data based on specific conditions and variables. |
| **Demographic information** | Information about patient characteristics, such as age, gender, and location. |
| **Descriptive statistics** | Techniques used to analyze and summarize data, providing initial insights and identifying gaps in data. |
| **Intermediate results** | Partial results obtained from predictive modeling can influence decisions on acquiring additional data. |
| **Patient cohort** | A group of patients with specific criteria selected for analysis in a study or model. |
| **Predictive modeling** | The building of models to predict future outcomes based on historical data. |
| **Training set** | A subset of data used to train or fit a machine learning model; consists of input data and corresponding known or labeled output values. |
| **Unavailable data** | Data elements are not currently accessible or integrated into the data sources. |
| **Univariate** | Modeling analysis focused on a single variable or feature at a time, considering its characteristics and relationship to other variables independently. |
| **Unstructured data** | Data that does not have a predefined structure or format, typically text images, audio, or video, requires special techniques to extract meaning or insights. |
| **Visualization** | The process of representing data visually to gain insights into its content and quality. |