**Data Science**

**Table of Contents**

1. **Introduction to Data Science**
   * What is science?
   * What is data science?
2. **Key Concepts in Data Science**
   * Data as a commodity
   * Data Science as a process
   * Summary of Data Science
3. **The Importance of Curiosity in Data Science**
   * The 5 Whys Technique
4. **Connection Between Science, Technology, and Data**
5. **Data Analytics vs. Data Science**
   * Differences between Data Analytics and Data Science
   * Diagram of roles
6. **Methodologies in Data Science**
   * What is a methodology?
   * CRISP-DM
   * KDD
   * SEMMA
7. **Steps in Data Science Methodology**
   * Step 1: Business Understanding
   * Step 2: Data Exploration and Preparation
   * Step 3: Data Representation and Transformation
   * Step 4: Data Visualization and Presentation
   * Step 5: Train Data Models
   * Step 6: Deploy Data Models
8. **Applications of Data Science in Various Industries**
   * Healthcare
   * Transportation and Logistics
   * Sports
   * E-commerce
   * Social Platforms
9. **Roles in Data Science Projects**
   * Data Analyst
   * Data Scientist
   * Data Engineer
10. **Key Takeaways and Summary**

**What is science?**   
Science is a system or method reconciling practical ends with scientific laws.

**What is data science?**  
Data science is the understanding of the world through the scientific analysis of digital data.

**Data science** combines the scientific method, math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and even storytelling to uncover and explain the business insights buried in data.

Data science is a **multidisciplinary approach** to extracting **actionable insights** from the large and ever-increasing volumes of data collected and created by today’s businesses.

The goal of data science is to extract value from data in all its forms.

People who work in data science prepare data for analysis and processing, perform **advanced data analysis**, and present the results to reveal patterns and enable stakeholders to draw informed conclusions. People who work in data science also use technologies and tools to build **models** to predict outcomes, discover underlying patterns, and gain insights that can lead to actions that **improve future outcomes**.

Data is a commodity. Businesses need ways to process data to find its value. Data science is a **scientific process** that’s repeatable. That’s not to say data science is mechanical and void of creativity. But data processing—from collecting data sources and data cleansing to machine learning and eventually visualization—includes unique steps that are involved in **transforming raw data into insight**.

**Summary**

Data science can help people find value and gain business insights from data. Data science follows a process, uses tools and technologies, and builds models to discover patterns to improve future outcomes.

**Be curious**

Data scientists who work with data begin by asking questions that start with “**why?**” For instance, why did life move from the primordial oceans to land? Did the moon and the tides have something to do with it?

One of the most important characteristics of a data scientist is to always **be curious**!

**The 5 Whys**

When analyzing data, you’ll find a problem and need to understand why. The [5 Whys(opens in a new tab)](https://en.wikipedia.org/wiki/Five_whys) is a valuable technique in problem solving that is easy to remember. You can determine the root cause of a problem by **asking the question "Why?" five times**.

First ask a "Why?" question, get an answer, and then ask a second "Why?" question. And so on! Each answer will inform your next "Why?" question. The answer to the fifth "Why?" should reveal the root cause of the problem.

Some problems have more than a single root cause. If you want to uncover multiple root causes, then the method must be repeated, asking a different sequence of questions each time.

**Note**: The number 5 is a rule of thumb and the number of "Why?" questions to ask can vary.

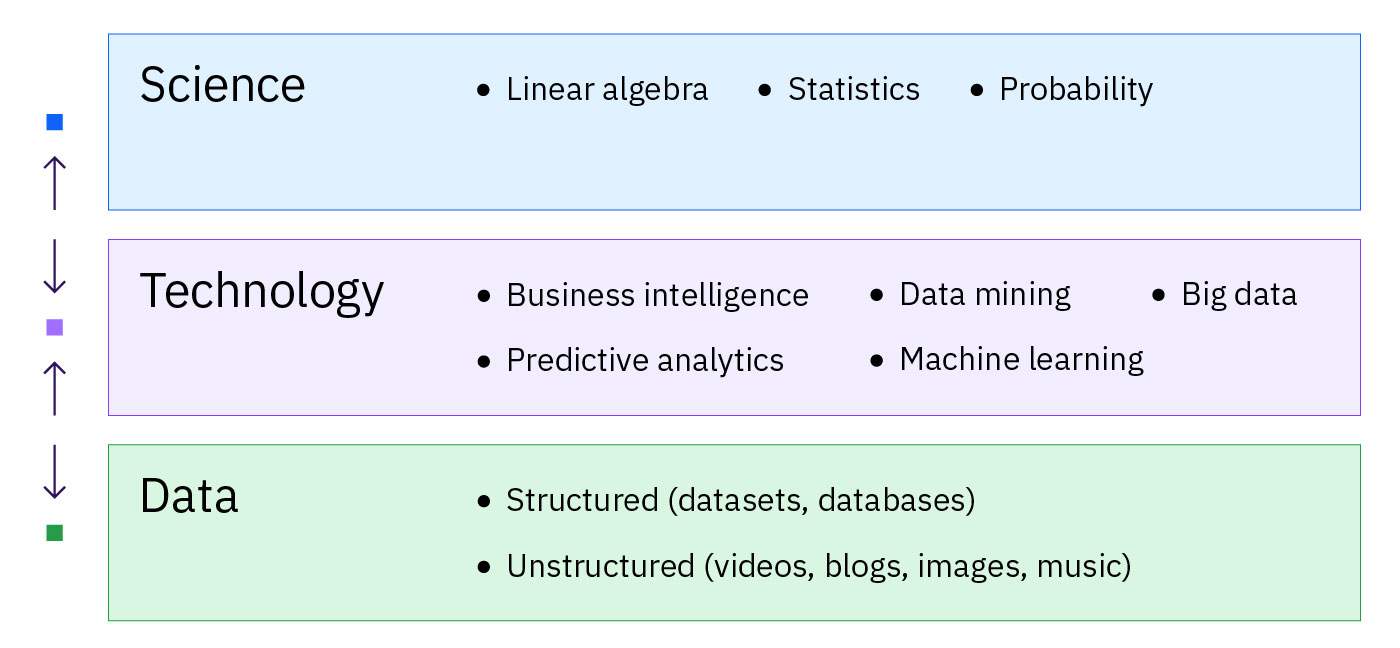
**Note**: When using the 5 Whys technique, avoid assumptions and the temptation to skip to the obvious answer!

# Science, technology, and data are connected

Science, technology, and data are areas that are linked and have given rise to data science. Keep this key concept in mind as you learn about the field of data science.

* A scientific background allows a data scientist to formulate a hypothesis and follow the scientific method.
* Working with technology and tools allows a data scientist to begin classifying data and making predictions.
* A data scientist can gain insights from structured data in relational databases and unstructured data, such as video footage, blogs, and tweets.

Science, technology, and data have been linked since the early days of science in the 1300s and 1400s. Each area continues to evolve.



At this time in the **science plus technology plus data world**, data science is one of the most important fields.

# Data analytics versus data science

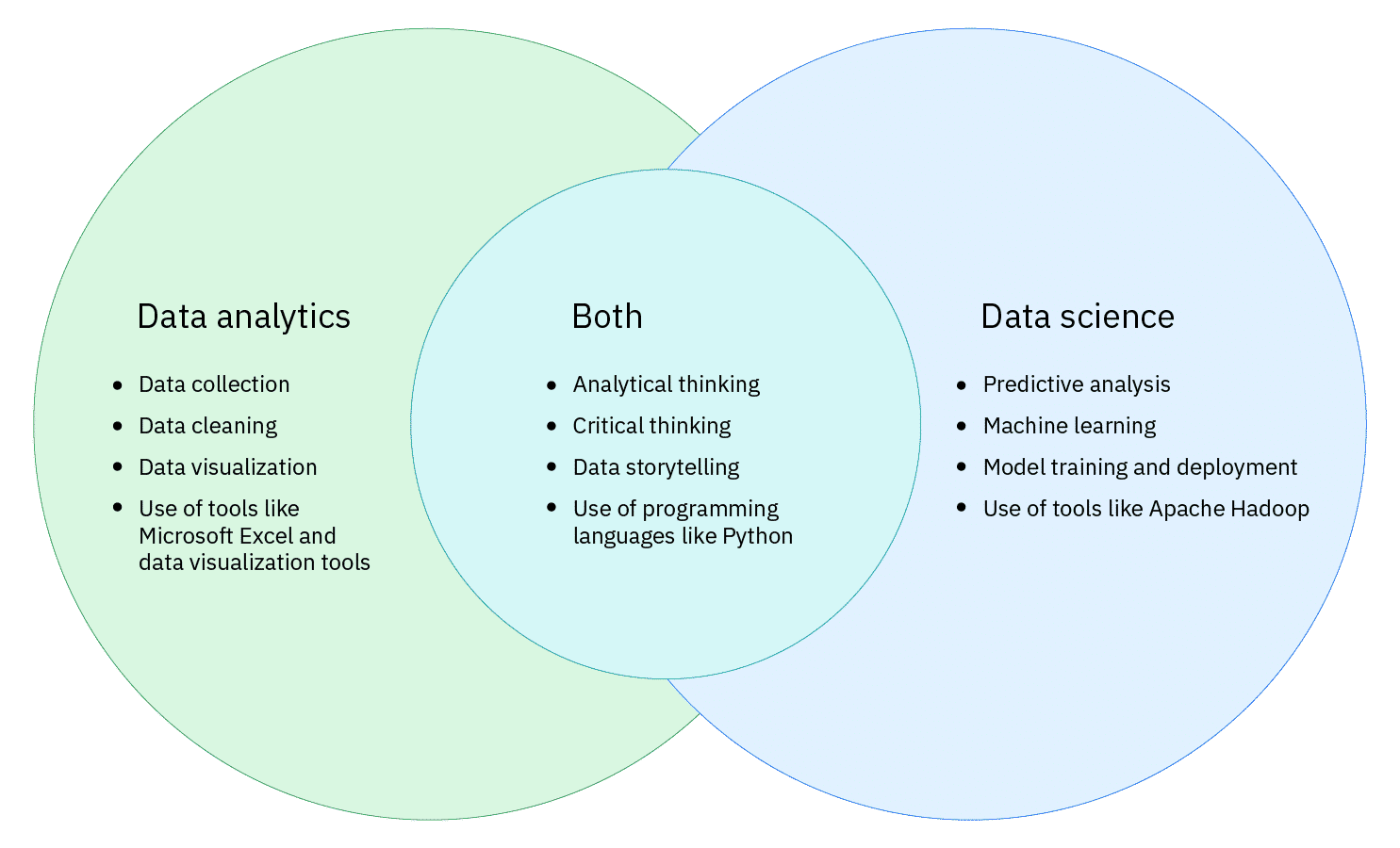
Data analytics and data science are two terms that are often used in the same context. But, it’s important to know they have different definitions.

Both fields work with data, and they share the same goal: to translate data analysis into business intelligence.

The main difference between data analytics and data science is in what data analysts and data scientists **do**with data, meaning the **tactics** used. Here is how they are differentiated:

* **Data analysts** collect and examine large data sets to identify trends, forecasts, and data visualizations to tell a compelling story through actionable insights. These insights help businesses make informed decisions about business needs.
* **Data scientists**design and create new processes for data modeling. They use algorithms, predictive analytics, and statistical analysis. Data scientists have technical skills to arrange unstructured data and build their own methodologies to make predictions based on data trends.

This diagram illustrates differences between data analytics, data science, and what they have in common.



You’ll learn about the role of the data analyst, data scientist, and the teams they work with on data science projects in this course.

**What is a methodology?**

A **methodology** is a general strategy that guides activities within a process. A methodology doesn’t depend on technologies or tools, and it’s not a set of techniques or recipes. Rather, a methodology provides data scientists with a **framework** for how to proceed with whatever methods and processes they will use to **obtain answers or results**.

Scientists have the scientific method. And like scientists, data scientists need a foundational methodology to guide them for solving problems.

It’s good know your history! To begin, here are **three classic and widely adopted data science methodologies**:

* Cross-Industry Standard Process for Data Mining (CRISP-DM)
* Knowledge Discovery in Database (KDD)
* Sample, Explore, Modify, Model, Assess (SEMMA).

**CRISP-DM**, **KDD**, and **SEMMA**:

* Use data mining methods
* Are best suited for structured data
* Are useful for using descriptive and predictive analytics
* Share some common activities, such as data gathering, data transformation, data modeling, and model evaluation

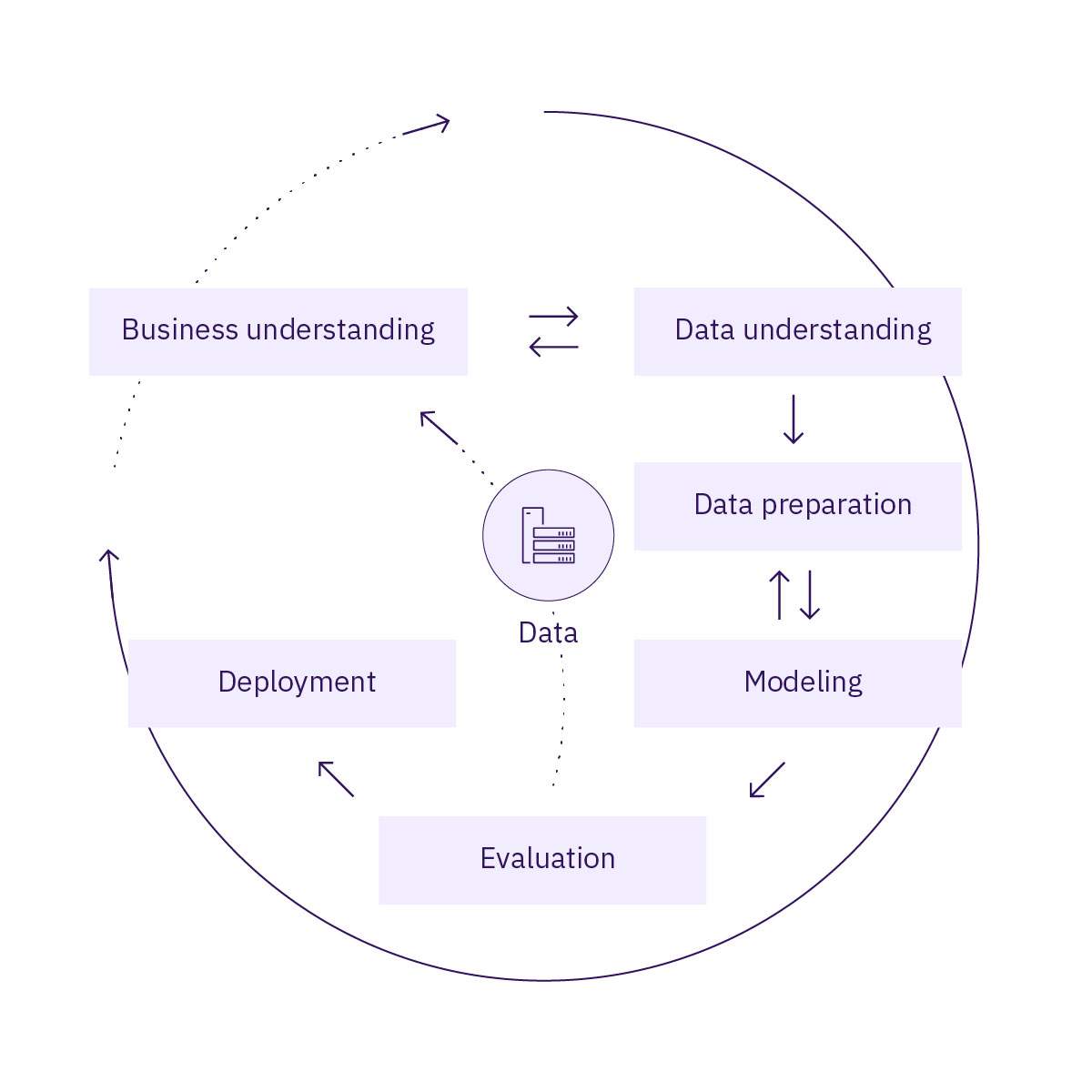
**Note**: These methodologies are not useful on projects that work with unstructured data, such as images and text.

You will learn about these widely adopted data science methodologies in upcoming lessons.

# Example 1: CRISP-DM

CRISP-DM stands for**Cross-Industry Standard Process for Data Mining.**

Founded by the [European Strategic Program on Research in Information Technology(opens in a new tab)](https://en.wikipedia.org/wiki/European_Strategic_Programme_on_Research_in_Information_Technology_(ESPRIT)) initiative, CRISP-DM is a proven way to guide data mining efforts. Any industry can use this methodology to help structure a data science project. CRISP-DM is a flexible and comprehensive data science approach.



CRISP-DM consists of six phases with arrows indicating the most important and frequent dependencies between phases:

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

The sequence of the phases is not strict. CRISP-DM is **iterative**, meaning that the phases can be repeated to incrementally improve the result. The results of some stages might require the project cycle to go back to earlier stages.

What’s unique to the CRISP-DM methodology is that it begins with **Business understanding**. This phase focuses on understanding the project objectives and requirements from a business perspective, and defining the data problem to solve.

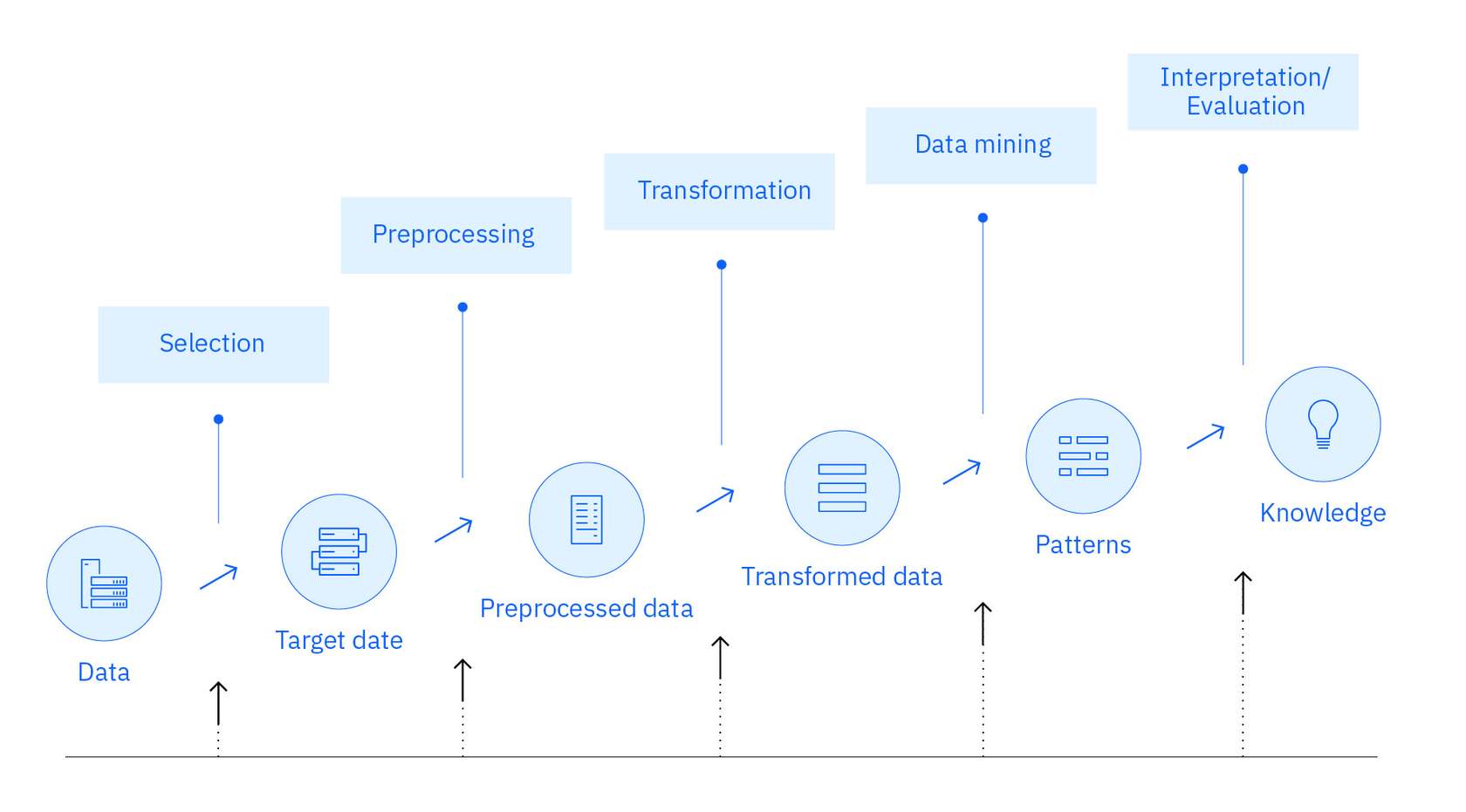
CRISP-DM is among the most popular approaches for data science projects.

**Example 2: KDD**

KDD stands for **Knowledge Discovery in Database.**

KDD represents the overall process of collecting data and methodically refining it. KDD typically has five steps:

1. Selection
2. Preprocessing
3. Transformation
4. Data Mining
5. Interpretation/Evaluation



The KDD methodology can help businesses stay current with customer needs and behaviors and predict future purchasing trends to stay competitive. But, the process doesn’t address many of the modern realities of data science projects, such as the setup of big data architecture, considerations of ethics, or the various roles in a data science team.

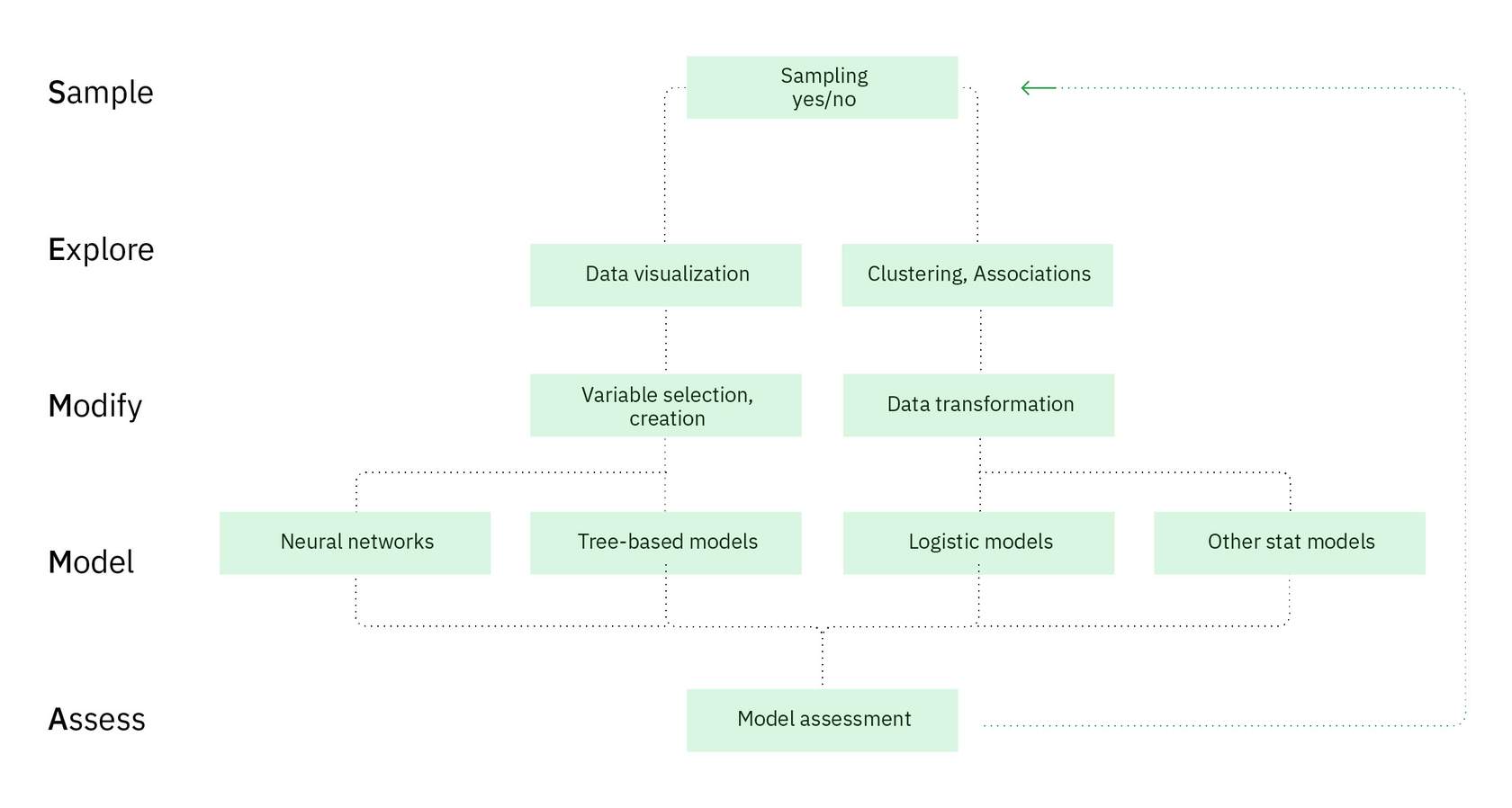
KDD is **iterative**, meaning new data can be integrated and transformed to get different and more appropriate results. The knowledge acquired can be cycled back into the process, enhancing its effectiveness.

**Source:** [KDD and Data Mining(opens in a new tab)](https://www.datascience-pm.com/kdd-and-data-mining/), Data Science Process Alliance, by Nick Hotz, July 2021

**Example 3: SEMMA**

SEMMA stands for its five steps:

1. **S**ample
2. **E**xplore
3. **M**odify
4. **M**odel
5. **A**ssess



SEMMA is a data science methodology that helps convert data into knowledge. SEMMA can help solve a range of business problems, such as fraud identification, customer retention and turnover, database marketing, customer loyalty, market segmentation, and risk analysis.

The [SAS Institute(opens in a new tab)](https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm) developed SEMMA as a process of data mining. SEMMA primarily focuses on the modeling tasks of data mining projects.

SEMMA is also an **iterative**process, in which answering one set of questions often leads to more interesting and more specific questions.

**Source:** [What is SEMMA?(opens in a new tab)](https://www.datascience-pm.com/semma/), Data Science Process Alliance, by Nick Hotz, May 2021

**Summary:**

The three methodologies are all **iterative**! This means that the phases or steps can be repeated. Knowledge acquired can be cycled back into the process to gain more or different insights.

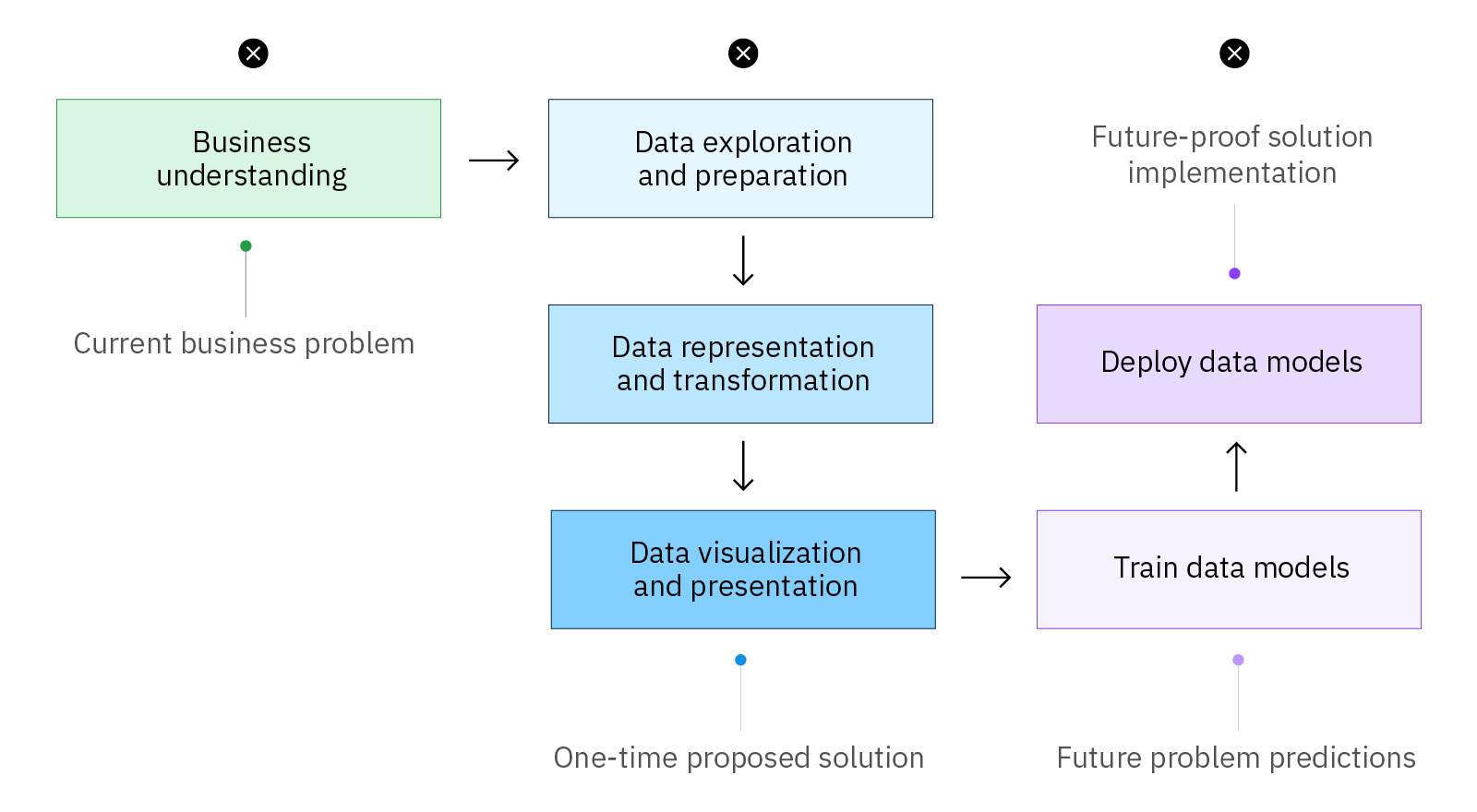
In conclusion, it’s good to know the CRISP-DM, KDD, and SEMMA methodologies for reference and background. Organizations and businesses typically determine their own methodology for a data science project. You’ll learn more about this is an upcoming lesson.

# Methodology overview

In previous lessons, three well-known data science methodologies were presented. To review, they are Cross-Industry Standard Process for Data Mining (CRISP-DM), Knowledge Discovery in Database (KDD), and Sample, Explore, Modify, Model, Assess (SEMMA). In this lesson, you will explore a data science methodology in depth.

Here are the steps in the data science methodology you’ll learn about:

1. Business understanding
2. Data exploration and preparation
3. Data representation and transformation
4. Data visualization and presentation
5. Train data models
6. Deploy data models

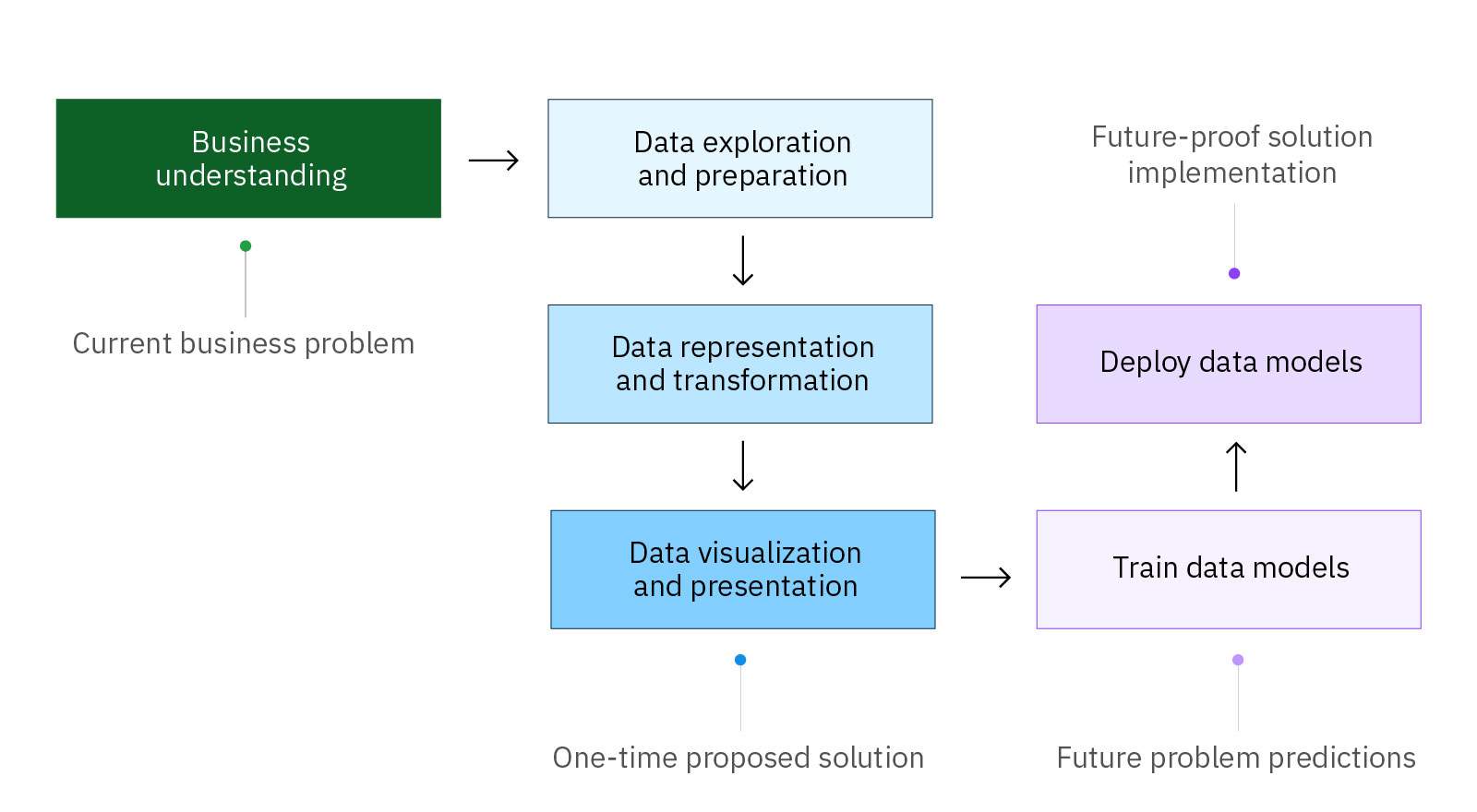


The methodology steps you use might vary. But you can follow the steps to learn about typical tasks that a company might implement for their data science projects. This methodology is independent of technologies or tools and provides a framework for processes that data scientists can use to obtain answers and results. The methodology is also scientific, so the key is that it is repeatable.

**Step 1: Business understanding**

Every project, regardless of how big it is, starts with **business understanding**. Before any data exploration can be done, the team must understand the problem that needs to be solved.

The **business sponsor**plays a critical role. Business sponsors are in a leadership position. They initiate the project because they have a “pain point”, so they bring the business problem to the data science project team  and then support the project.



The data science project team examines the business problem. They can do this using **design thinking**. Design thinking is a problem solving methodology that focuses on the user, having empathy for the user, and determining the best user experience.

As part of this method, the data science project team might host a design thinking workshop and apply techniques to:

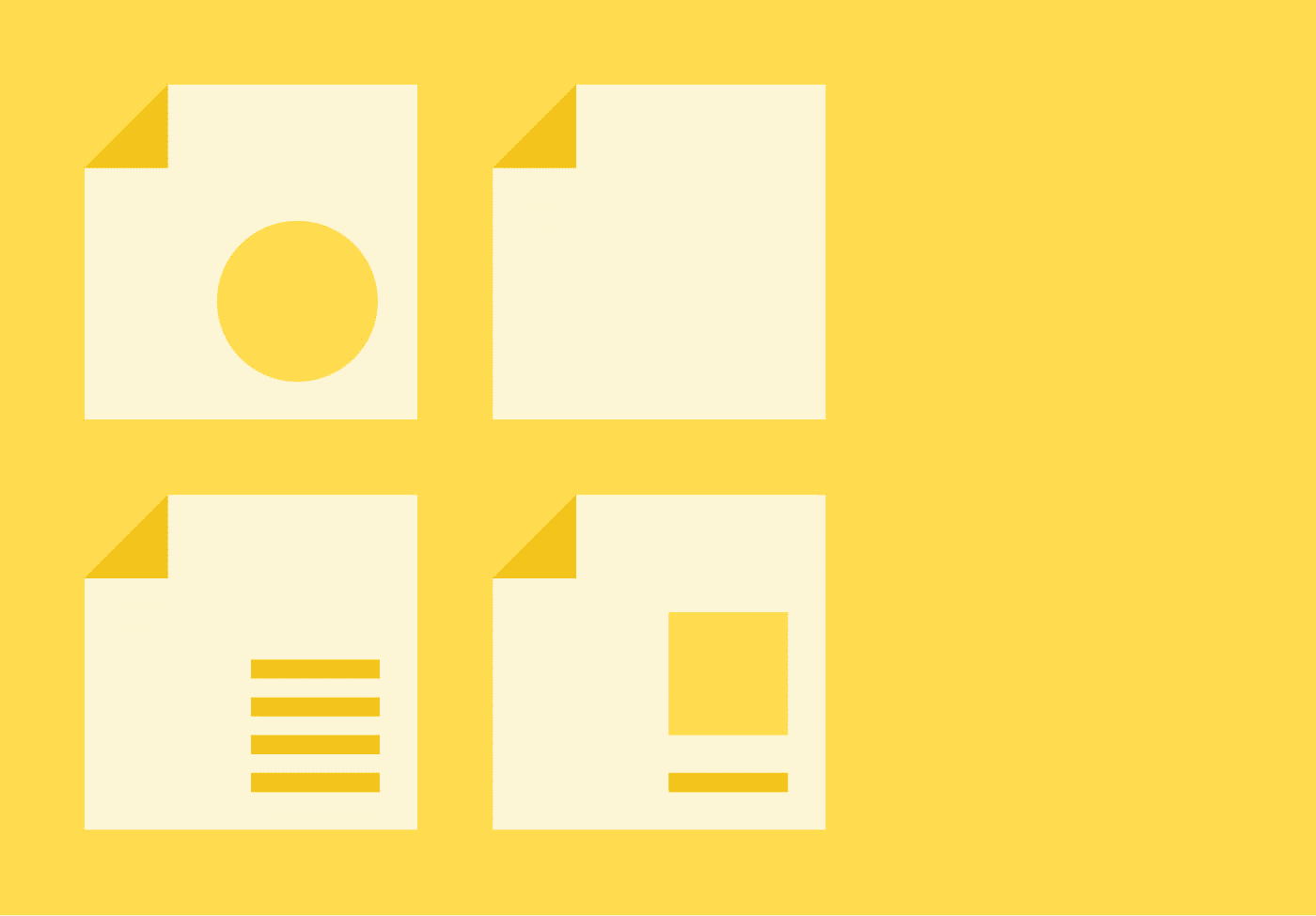
* Define the problem
* Determine the project objectives
* Develop personas or fictional characters that represent typical end users
* Document solution requirements from a business perspective

Defining the problem helps lay the foundation for a potential resolution to the business problem. The business sponsor should also be involved throughout the project to provide expertise, review data findings, and ensure the project stays on track.

Once the business problem is clearly stated, a **data scientist** on the team defines the analytic approach to solving the problem. This involves expressing the problem in the context of statistical and machine learning techniques. For example, if the goal is to predict a customer’s response in terms of a “yes” or a “no”, then the analytic approach could be defined as building, testing, and implementing something called a logistical regression model. Data scientists are experts who have many technologies and methods in their toolbox!

**Project scenario: Business understanding**

A fictitious business, named GAXR, develops software for companies that design video games. GAXR has over 1000 employees. Suddenly, GAXR is facing a human resources (HR) issue. The HR director doesn’t know why so many people are leaving the company. There haven’t been any company layoffs, salaries are competitive, and the amenities at work are great, including a free onsite gym!



The **business problem** is that there has been higher-than-expected employee attrition over the past six months. Employee attrition is when employees depart a company for any reason, such as resignation, termination, or retirement. In the case of GAXR, the employee attrition is due to voluntarily resignations or people quitting.

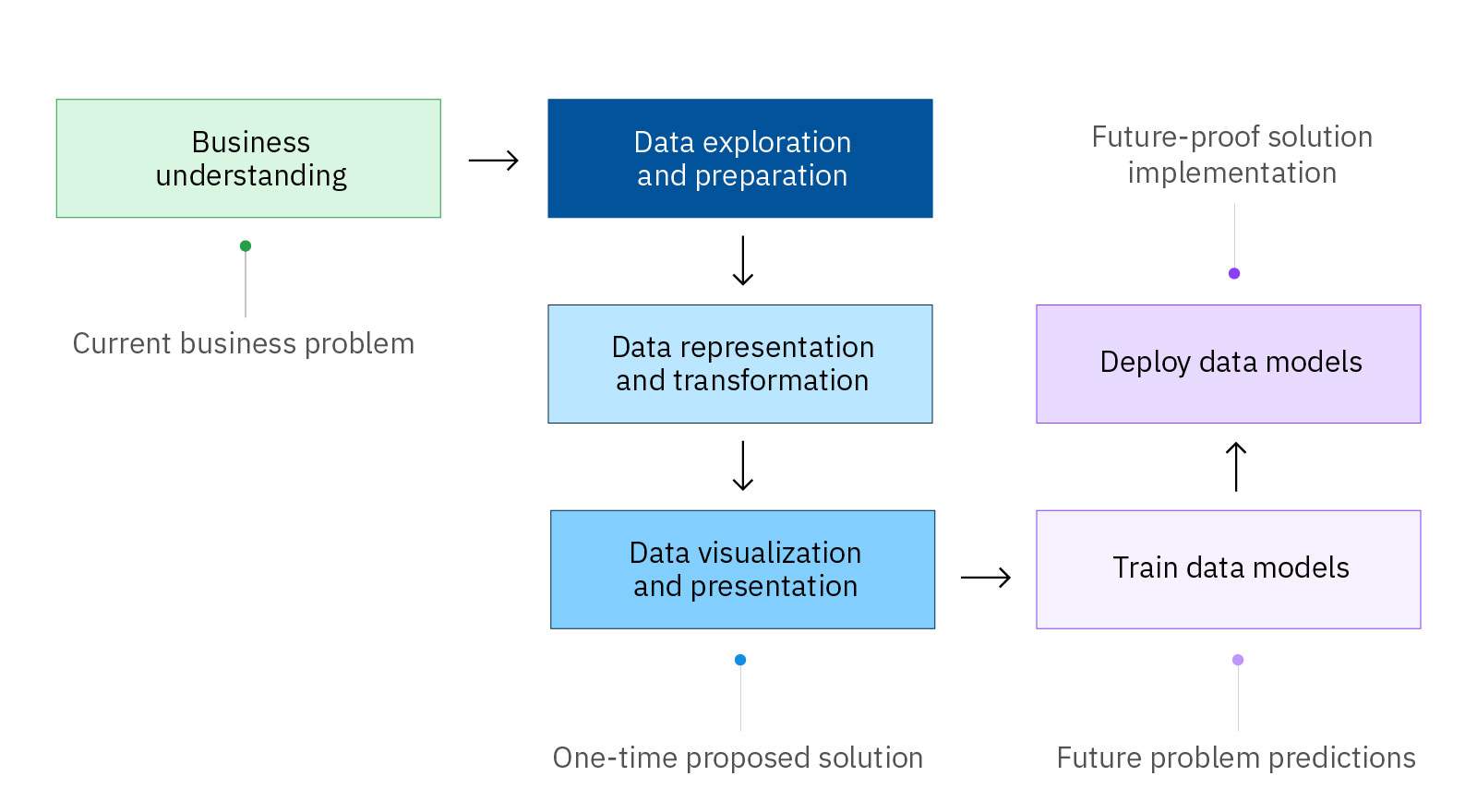
The HR director, Marilyn Shah, is the **business sponsor**. She gathers a team and hosts a one-day design thinking workshop so the team can understand the business problem and consider how to address the problem.

The **data science project**begins! The team begins to brainstorm and hypothesize why people are leaving the GAXR. The team **data scientist**, Scott Hill, participates. He partners with his colleagues to define the best analytical approach and determine which company tools the team will use.

**Step 2: Data exploration and preparation**

Data scientists identify and collect data from existing and often new sources in a business. This could be structured and unstructured data that is relevant to the problem. They might retrieve data from sources such as:

* Static files, such as spreadsheets
* Databases
* The internet



Step 2: Data exploration and preparation

If a data scientist encounters any issues or gaps in data collection, then the data scientist might need to revise the data requirements and collect more data.

Once data is in a format a data scientist needs to work with it, **data exploration** can begin. The **initial**exploration of a data set is important because it helps data scientists find patterns and relationships and discover initial insights from the data. Here are a few questions a data scientist might think about during the initial exploration of data:

* Which data characteristics seem promising for further analysis?
* Has exploring revealed new characteristics about the data?
* Has exploring changed the initial hypothesis?

Next, the data scientist next **prepares the data**. Data preparation is very important and the most time-consuming step in a data science project. It involves constructing the data set that will be used in the modeling step. Data preparation also includes cleaning the data, combining data from multiple sources, and making sure the data doesn’t have any gaps. Additionally, data preparation includes cleaning or “wrangling” the data so it’s ready to transform.

Data scientists can’t assume that data is ready to use, even if it’s structured data. Real-world data usually needs some work because it might be:

* Incomplete or have incorrect values
* Corrupted with broken lines or have fields in the wrong place
* Too random
* Irrelevant
* An outlier, which is a value that lies far away from other values and will skew the data
* A missing value in some fields

Manually verifying large volumes of stored data can be challenging for data scientists. So, data scientists use automated processes and tools to prepare data quickly and accurately.

**Project scenario: Data exploration and preparation**

Scott Hill, the team data scientist, has a meeting with Lena, the GAXR database administrator, to get her guidance on the types of data that the human resources group collects for its employees. Scott reviews and determines the employee data set he wants, such as:

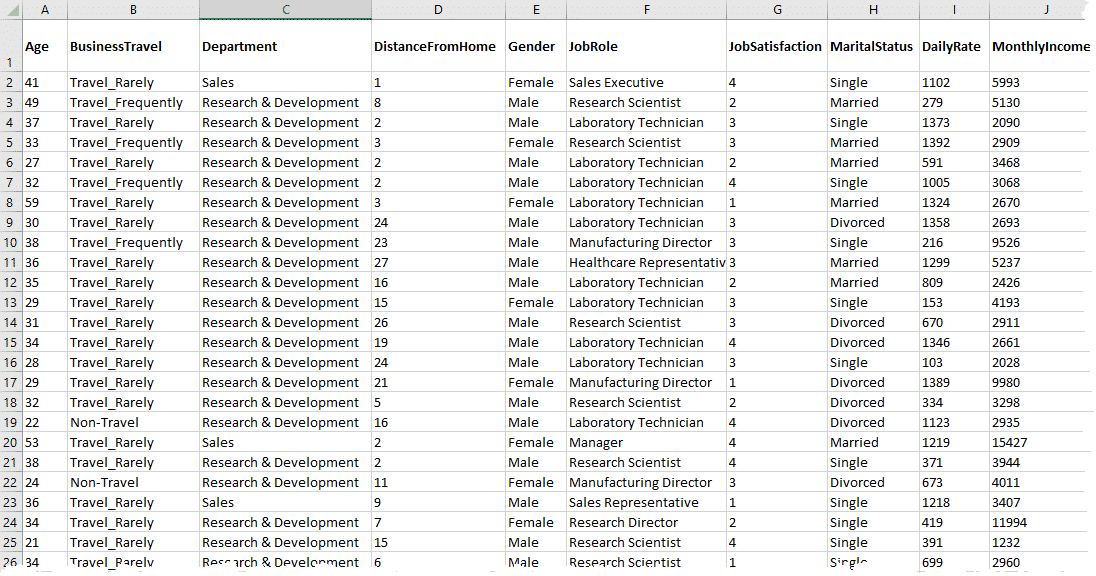
* First name
* Last name
* Age
* Department
* Distance from home (to work in kilometers)
* Hourly rate
* Daily rate
* Monthly income
* Marital status



Scott uses **Structured Query Language (SQL)** to move the data he needs for the project into a **CSV file**.

* SQL is the most common language for extracting, organizing, and managing data in a relational database to then perform various operations on the data.
* A comma separated values (CSV) fileallows data to be saved in a tabular format.

Here is the CSV file that Scott is cleaning up:



Each line of the file is a data record for one GAXR employee.

Next, Scott cleans the data! For instance, he removes the **Employee Count** field since all GAXR employees have the value of 1 (because each line is for one (1) employee). This is unnecessary data. Scott also updates the**Hourly Rate** field and all other monetary fields to remove decimal places and round up to the nearest US dollar.

Scott also removes **personal information (PI)**since it’s not pertinent to his data science project. This includes the **First Name** and **Last Name** fields.

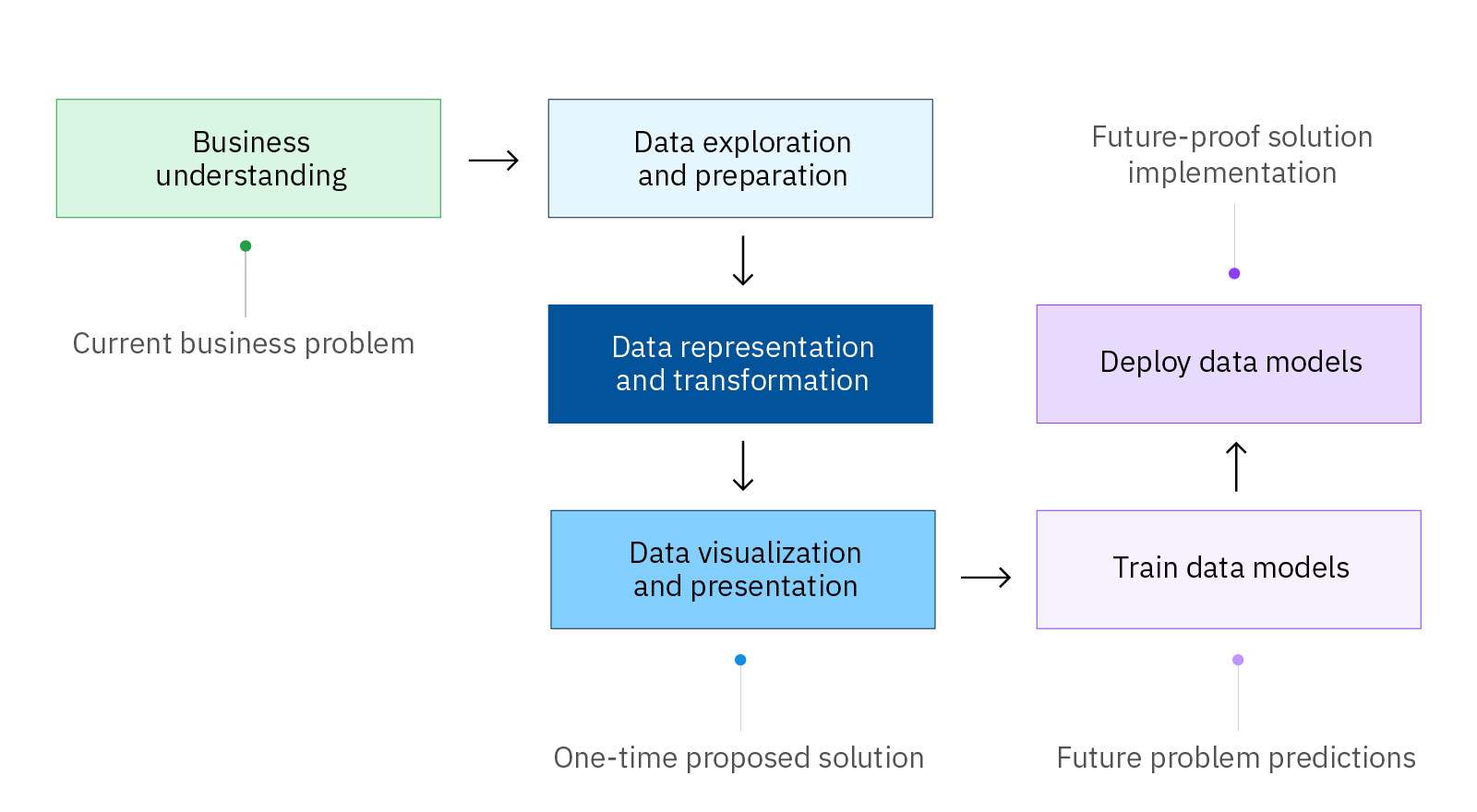
In order to run data comparisons in the tool, Scott also removes columns that are not helpful in making predictions, such as the **Telephone Number**.

Additionally, Scott ensures there aren’t any gaps and that he has enough data. In all, he has 1470 employee records to use for his analysis. He revisits and confirms his hypothesis: Employees are resigning from GAXR because they need a higher salary in today’s economy.

**Step 3: Data representation and transformation**

The data representation and transformation step in the data science methodology is about:

* Understanding the data
* Assessing data quality
* Discovering initial insights about the data



Think of data scientists as detectives investigating a case. Data scientists use many techniques, tools, and types of analysis to better **represent and transform data** to detect insights. This lesson presents some of these techniques, but remember that there are many more.

**Descriptive statistics**

To understand data, a data scientist can use a mathematical approach, such as descriptive statistics. **Descriptive statistics** quantitatively summarizes a data set. It can answer the question, “What is happening?” Data scientists can build a table to describe a large, complex data set and make quick observations about:

* Number (N): What is the total number of observations?
* Mean: What is the average of a set of two or more numbers?
* Median: What is the middle number or “center” in a sorted list of numbers?
* Mode: What is the most observed value in a data set?
* Minimum: What is the minimum extreme of a data set?
* Maximum: What is the maximum extreme of a data set?
* Standard deviation: How spread out is the data in relation to the mean?

For instance, the following table indicates the mean household income in US dollars for a company’s customers, based on their education level. The data is fictitious and for illustrative purposes.

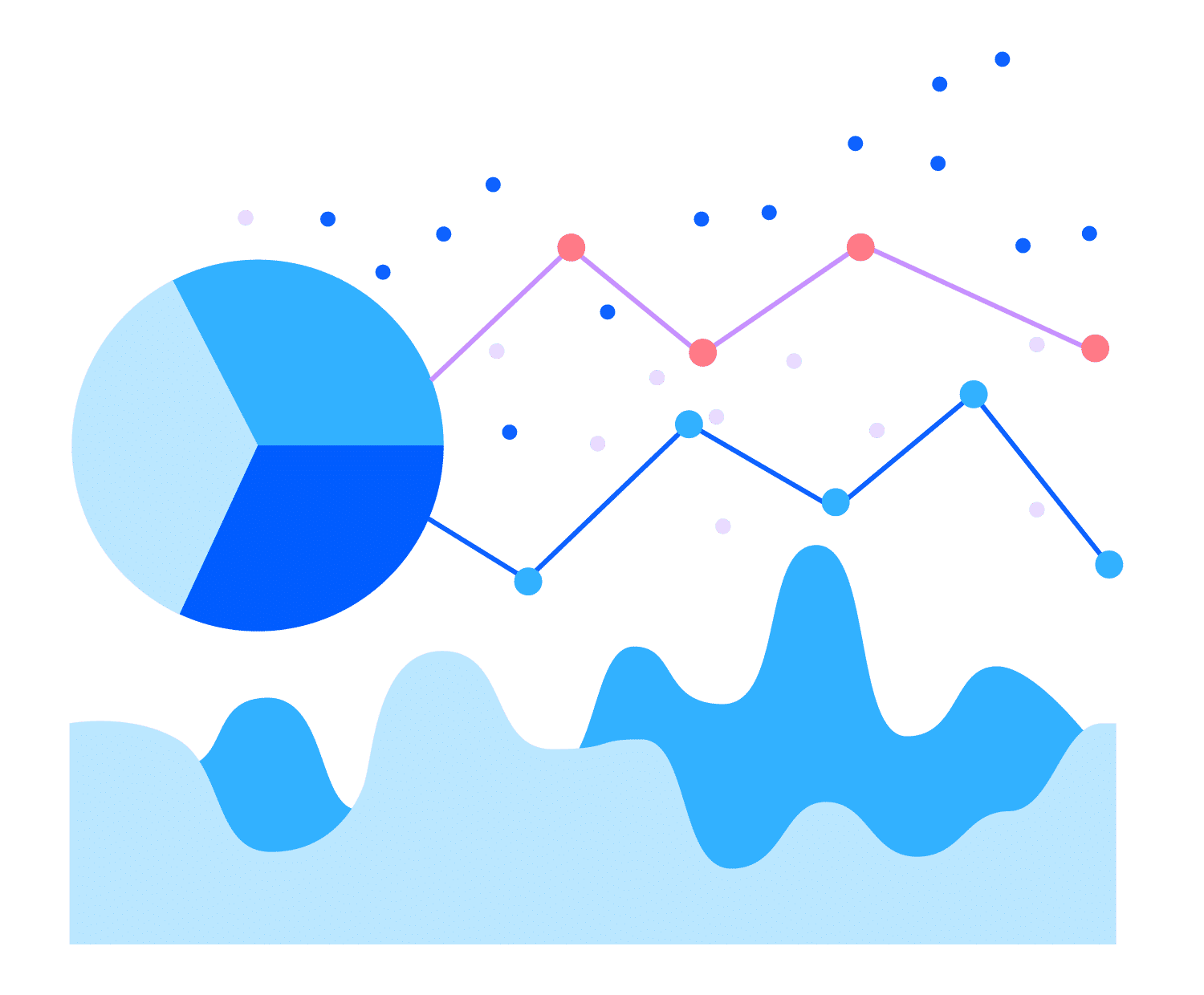
|  | **Did not complete high school** | **High school degree** | **Some college** | **College degree** |
| --- | --- | --- | --- | --- |
| **Mean** | 51.48 | 52 | 56.90 | 70.94 |
| **Standard deviation** | 51.855 | 56.370 | 53.836 | 67.940 |
| **N** | 246 | 527 | 333 | 310 |
| **Median** | 36.00 | 35.00 | 39.00 | 49.00 |
| **Minimum** | 15 | 12 | 13 | 15 |
| **Maximum** | 497 | 533 | 403 | 512 |

**Summary**

Descriptive statistics can begin to reveal insights. For example, the table data suggests that the household income will be higher as the education level increases. Descriptive statistics can quickly provide a data scientist with an **overall sense**of the data being analyzed.

**Exploratory visualizations**

Sometimes, statistics can be misleading when on their own, so it's important to have other techniques. Data scientists also use **exploratory visualizations**. Exploratory visualizations help make complex data more accessible and revealing. Data scientists use initial visualizations, like charts, graphs, and maps, to uncover distributions, find patterns, and understand trends.



Descriptive statistics, visualization techniques, and many other techniques help data scientists understand data and **assess its quality**. Data science teams must validate the quality of the data they use as input for predictive modeling because poor quality data will lead to poor model performance later in the process.

**Transforming data with machine learning**

Data representation is followed by**data transformation**. Regardless of the format of the source data, a data scientist structures and organizes data in a format that supports the most efficient machine-learning model possible. **Machine learning** is a branch of artificial intelligence (AI) and computer science that focuses on using data and algorithms to imitate the way that humans learn, gradually improving its accuracy. Machine-learning models understand only numbers, not text or images, so data scientists might need to transform unstructured data into a “0” or “1”.

For instance, a data scientist can break up text into words, phrases, symbols, or other meaningful elements called **tokens**. Then, a list of tokens becomes input for further processing into numbers. This technique is called **tokenization**, which is one of many data transformation techniques.

Data scientists also use data transformation tools and languages, like [Python(opens in a new tab)](https://www.python.org/), to conduct more in-depth analysis and create more complex visualizations. They perform programmatical activities on the data formatting, including depicting which models may be the best fit for the data.

To make analysis easier, data scientists often need to standardize data into model-friendly formats so that it’s “tidy data” rather than “messy data”.

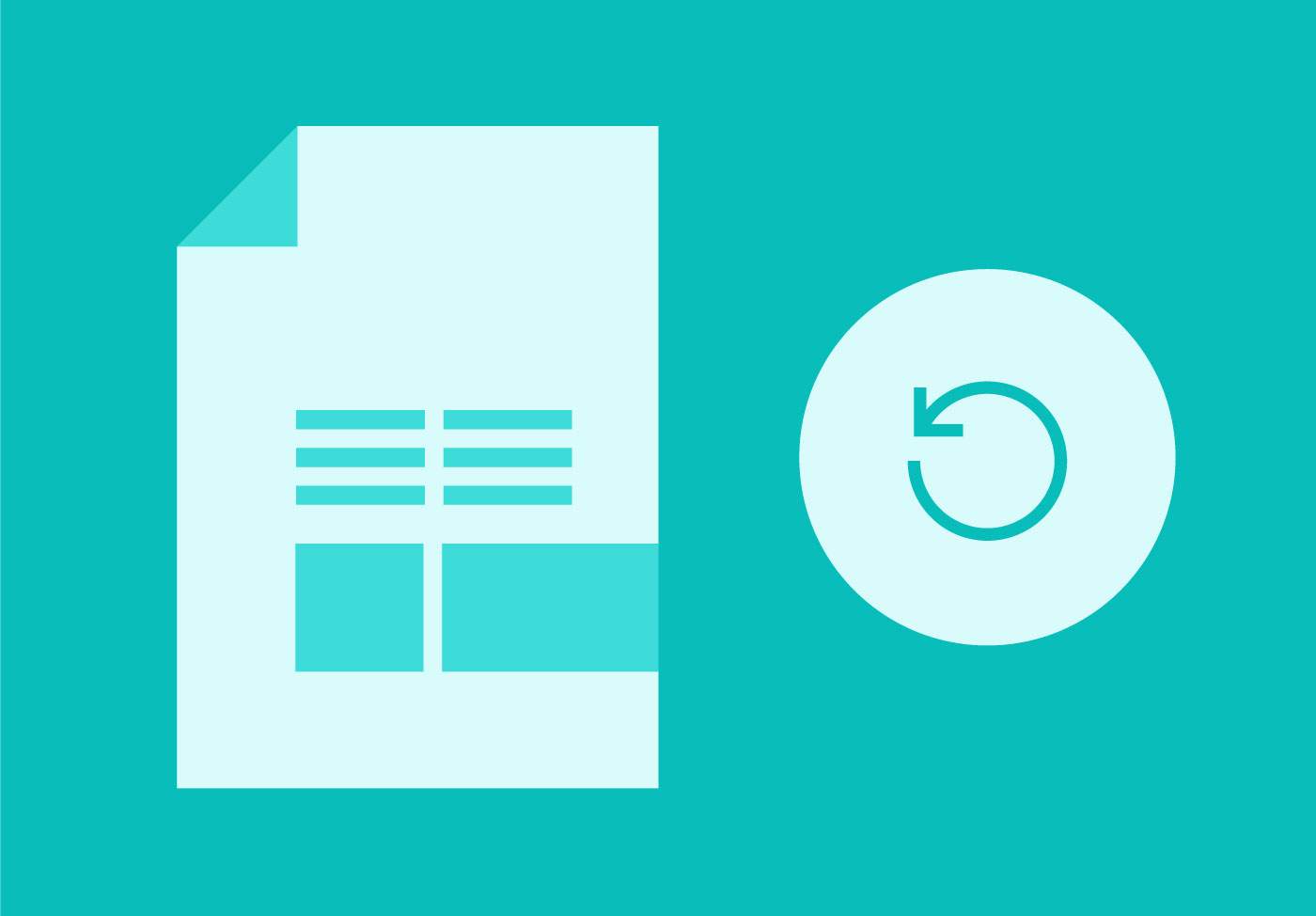
Data scientists might need to refine their analytics approach they selected in the Business understanding step.

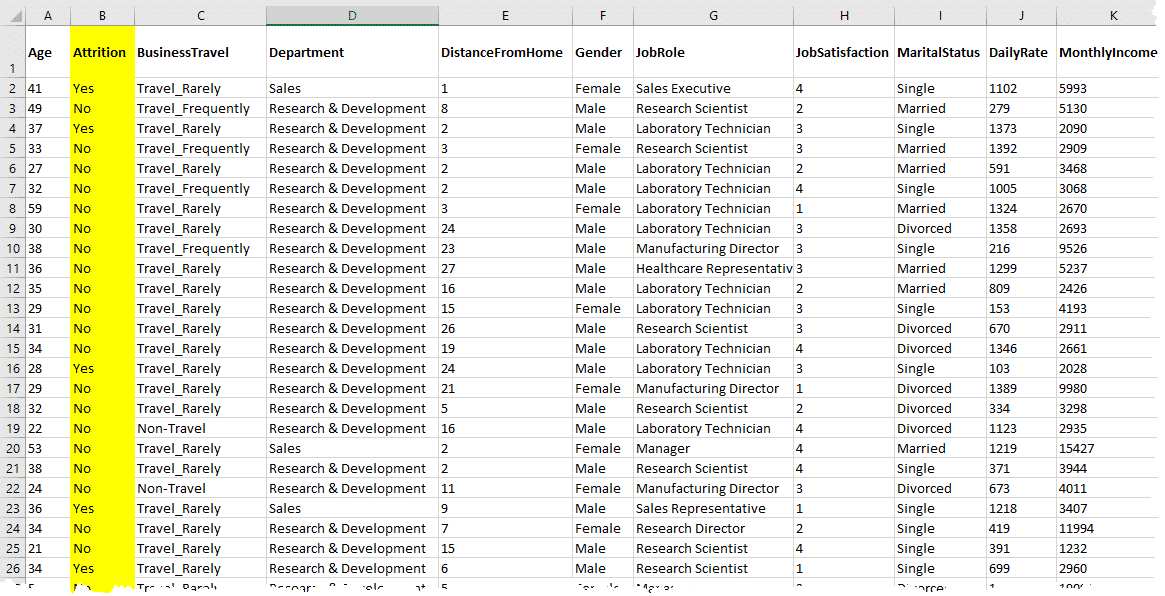
**Project scenario: Data representation and transformation**

Scott Hill, the team data scientist, continues with the data science project to find out **why employees are leaving GAXR**.

Scott realizes he needs to access another data source to better understand all of the data. He needs data on why a particular employee left the company. Scott uses SQL to merge more data into his CSV file. He adds a new column for **Attrition**. This is an important step!

Take a moment to study column B—the new, highlighted **Attrition** column—in the following image of the GAXR CSV file.





The Attrition column contains what’s called **dichotomous data**, meaning each cell has only two possible values: Yes or No. These values can help Scott quickly answer the question, “Did the employee leave GAXR or not?”

Scott makes sure the data is in the right format for the model the data science project team plans to use.

Scott also uses a data visualization tool to begin creating different charts and graphs. The charts and graphs can provide Scott with initial insights he can use to draw conclusions about the data.

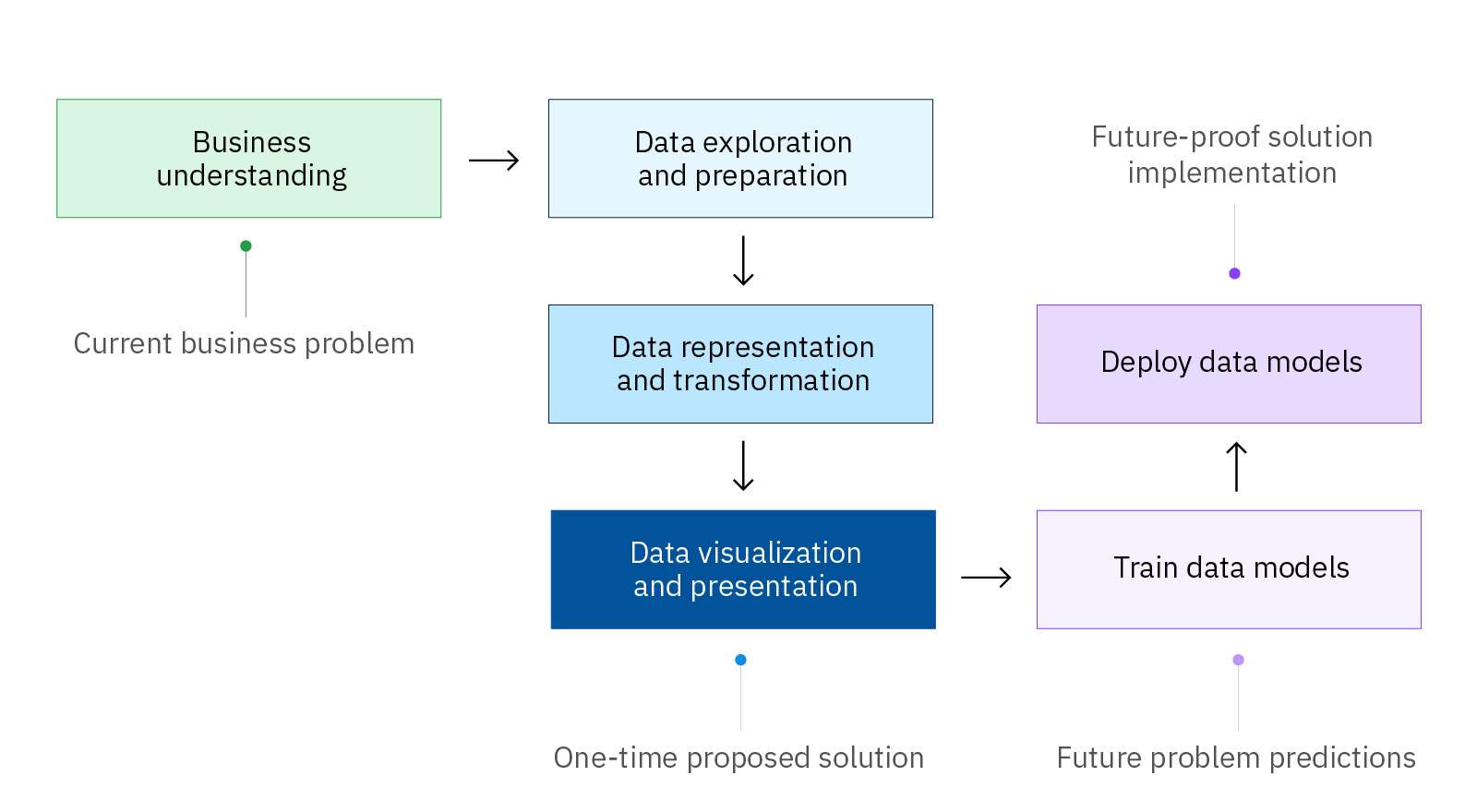
**Important note:** Column B, for **Attrition**, will be training data to help build a supervised learning model. The model will help determine the cause of employee attrition that the company can also consider in the *future*.

**Step 4: Data visualization and presentation**

**Data visualization** is the culmination of a data science team’s efforts to view the insights that their data transformation efforts have produced.

Visualizations help data scientists test their hypotheses and check assumptions.

The data needs to **tell a story** and address the business problem or question that a project is trying to answer!



Step 4: Data visualization and presentation

The data science team uses software to help visualize and conclude final insights from the data. The goal is to have a visualization that’s**effective**, **attractive**, and **impactive**.

**Types of data visualization**

One type of visualization can display data in a better way compared to another type of visualization. Here are some of these types of visualizations.

It has been said that “**a** **picture is worth a thousand words**.” This is particularly true when creating a data visualization. A single chart can answer quite a few questions. These graphical representations can help clarify and simplify complicated issues.

**Considerations for data presentation**

The final **data presentation** must be meaningful, compelling, and, most importantly, easy to interpret for the business sponsor. A team's presentation design must go beyond just showing results and the look of the data visualization or visualizations. The team must consider the:

* **Purpose**: What problem are you trying to address and why will data visualization help to solve it?
* **Audience**: Who is viewing the presentation and how can it be valuable to them?
* **Data**: Is the data represented in the best way and will the visualization need to be updated in the future?
* **Context**: Where will the visualization reside (for example, in software, on a website, or in a business report)?

Remember, the data visualization and project presentation must answer the current business problem and share a proposed solution with the business sponsor.

**Explore more!**

Check out these resources:

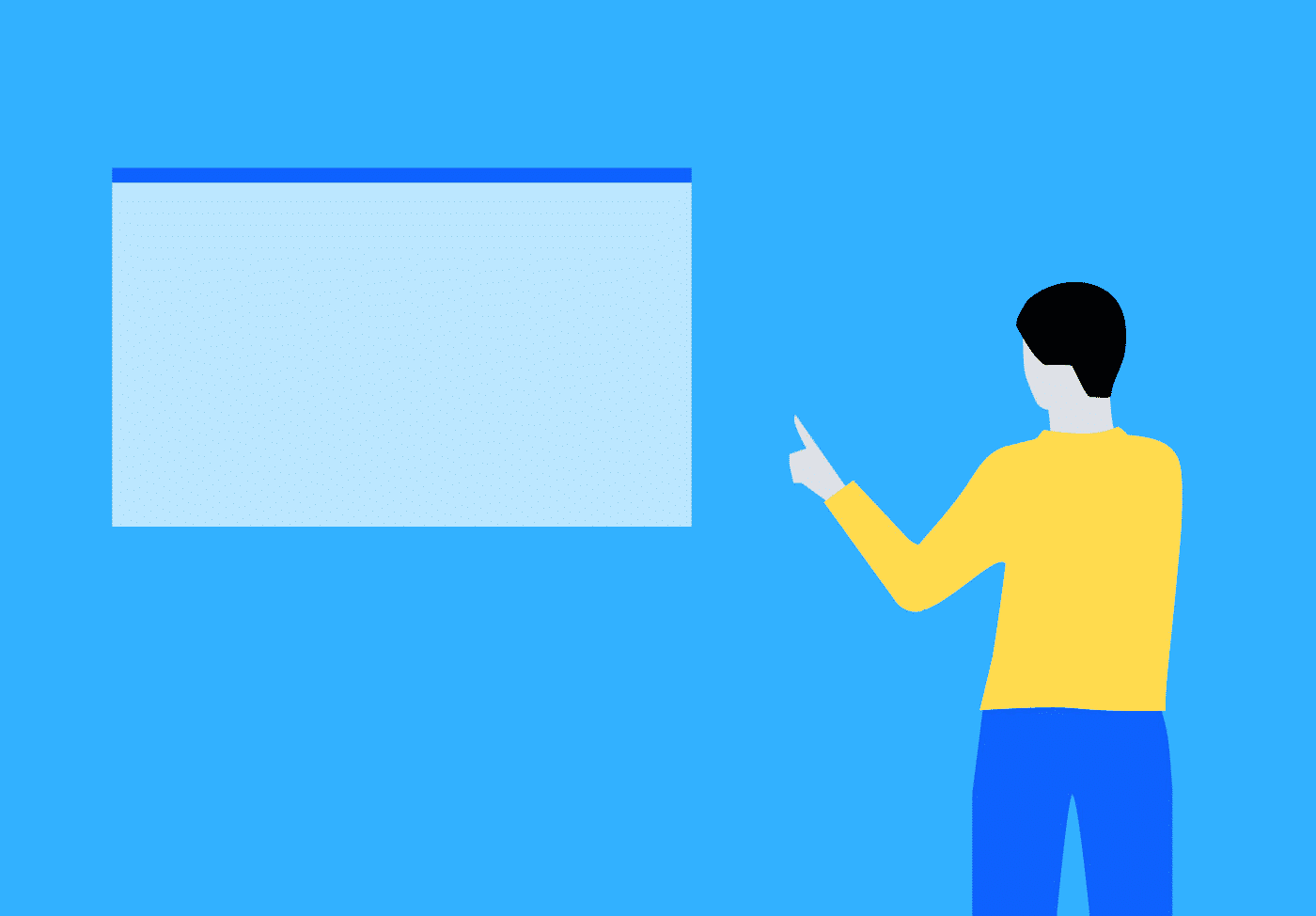
* [From Data to Viz(opens in a new tab)](https://www.data-to-viz.com/) – an online tool to learn more about charts and help determine the most appropriate graph for data
* [How to Choose the Right Data Visualization(opens in a new tab)](https://chartio.com/learn/charts/how-to-choose-data-visualization/) – a data tutorial from Chartio (Atlassian) by Mike Yi and Mel Restori
* [Viz of the Day(opens in a new tab)](https://public.tableau.com/app/discover/viz-of-the-day) – a site from Tableau Public that features visualizations shared by people around the world

**Project scenario: Data visualization and presentation**

Scott Hill, the team data scientist, is ready to find insights and discuss a solution to GAXR’s business problem!

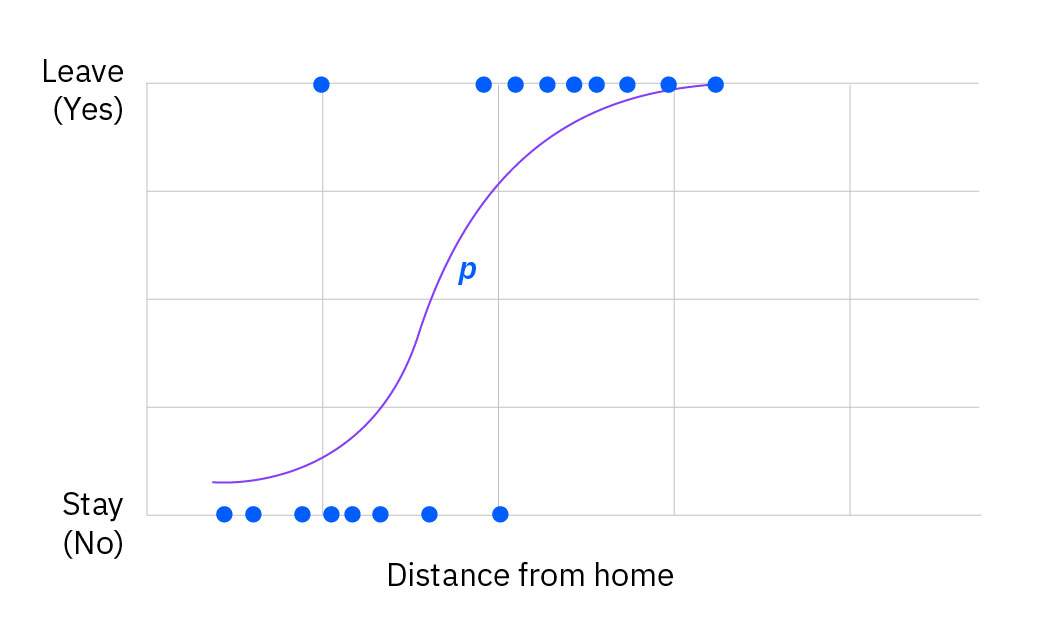
First, Scott uses a data visualization tool to refine and study the data. Scott finds the best chart to represent the data. The chart helps him make a conclusion.

Scott concludes that **employees who live far away from GAXR do not seem to stay at the company long-term**. The **Distance From Home** field, on its own, didn’t show this, but the visualization does!



The best data visualization of this data is a **scatter plot chart** **with an S-curve**. This visualization shows the relationship between two variables:

* How one variable impacts another
* How the value of both variables changes because of this impact



Scott's scatter plot chart is called an S-curve because you can see the curve in an “S” shape.

* The dots on the chart represent employees in the data set.
* The chart shows that employees stay and leave the company. This is **dichotomous data**, meaning “Yes, they leave” or “No, they stay.”
* As the distance from home increases, more dots (or employees) are clustered.
* The “p” near the s-shaped curve represents the probability that employees will leave, the greater the distance from home.

The **conclusion** is that employees who live a certain distance away from the office demonstrate a greater probability of leaving the company, compared to employees who live closer to the office.

Scott prepares a **presentation**for the team and Marilyn, the business sponsor from HR. In his presentation Scott covers these topics:

* The business problem
* The data sources used
* The steps taken to prepare and transform the data
* The data visualization to show the insight
* The answer to the business problem

Scott saves time at the end of the presentation of the team to brainstorm about a **proposed solution** HR might consider. The team identifies a couple of potential action items to address the business problem.

Ultimately, the team's proposed solution is to offer employees the option to work from home two days of the work week. Most likely, Marilyn Shah, the HR director and business sponsor, will share the team’s proposed solution with the leadership.

Scott saves his data and plans to revisit it again in another six months.

Marilyn has an answer and a potential solution to the business problem that’s been occurring during the past six months at GAXR. But, she’s thinking ahead, too. **What about the future?** She asks the data scientist and team to continue to the next steps in the methodology. Marilyn believes it will be valuable to train and deploy a model to **predict** future problems about employee attrition.

A data science project might conclude at this point in the methodology, with a **one-time proposed solution**.

Or, the project might continue to build a model for future problem predictions. With clean data in place and a good understanding of it, data scientists can build a model.

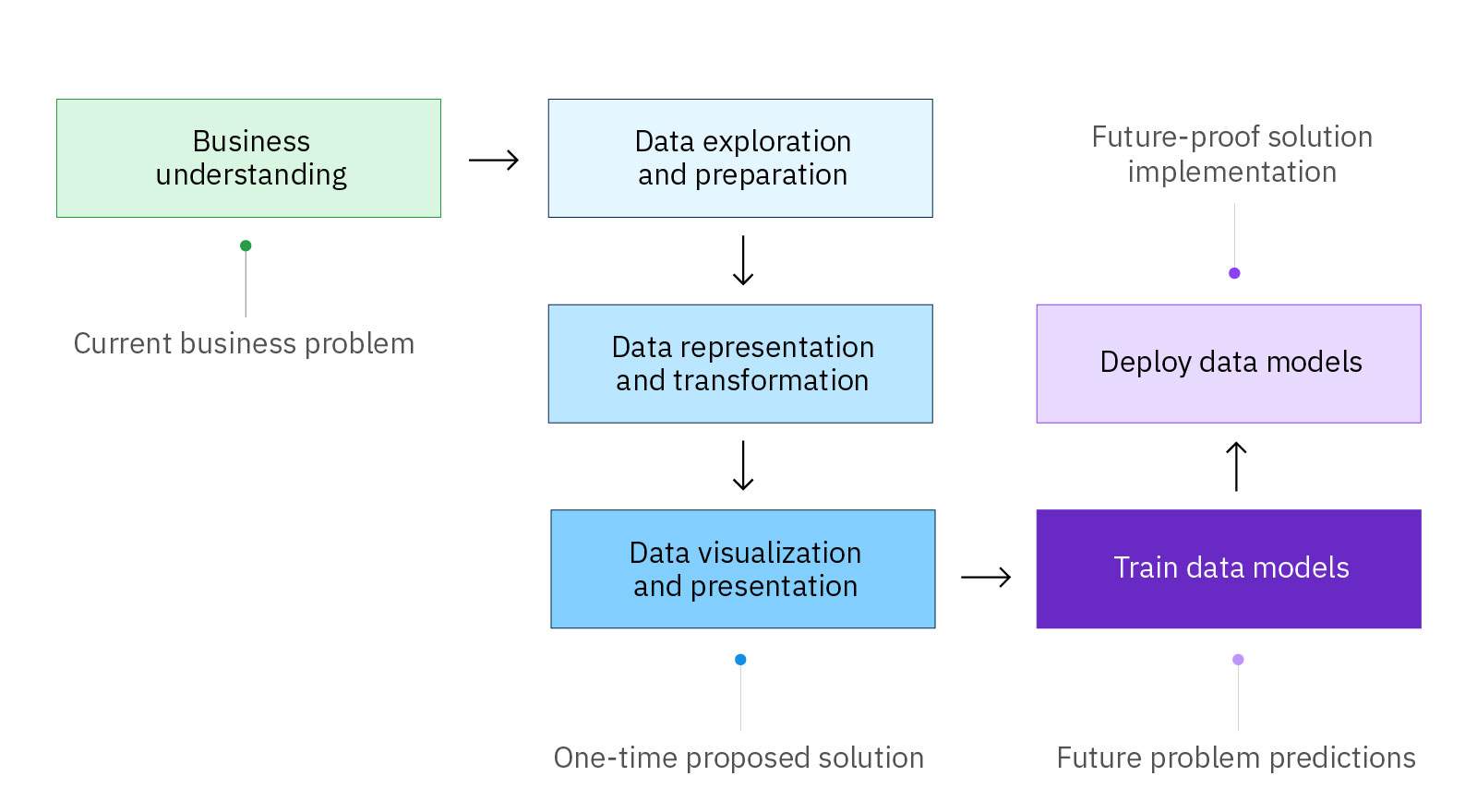
**Step 5: Train data models**

What is meant by a **model**?

* **What is a model?** A data model identifies the data, data attributes, and relationships or associations with other data. A data model provides a generalized view of data that represents the real business scenario and data.
* **Why build a model?** A data scientist can develop a more systematic approach to address an identified business problem by building a model. The main goal of building a model is to make better predictions for the business and gain a better understanding of the system being modeled.

Remember, data scientists use the scientific method, so they experiment! They can select from many different types of models to make predictions about future outcomes.

Data scientists must train a model. How? They use **machine learning**. Essentially, machine learning is teaching a computer to solve problems. Machine learning allows a machine to learn from data without programming it with rules. The machine can learn from the data it’s given. A machine learning algorithm "ingests" data so it can improve its accuracy.



Step 5: Train data models

Here are three methods for machine learning, based on the algorithms used and the results that are required.

***Supervised Learning***

In supervised learning, a machine ingests many questions and their answers—essentially a set of pre-structured information. The information might, for example, be drawings and photos of animals, some of which are dogs and are labeled “dog”. The machine attempts to  identify patterns so that when it sees a new photo of a dog and is asked, “What is this?”, it can respond, “dog”, with high accuracy.

Supervised learning trains machines on data to build general rules that can be applied to future problems. The better the training set of data, the better the output.

**Unsupervised Learning**

In unsupervised learning, a machine ingests an enormous amount of information, asked a question, then allowed to determine how to answer the question by itself. For example, a machine might receive many photos and articles about dogs. The machine ingests and classifies the information within all of the photos and articles. When the machine is shown a new photo of a dog, the machine is intended to be able to identify it as a dog, with reasonable accuracy.

Unsupervised learning trains machines on a huge volume of unlabeled or unstructured data.

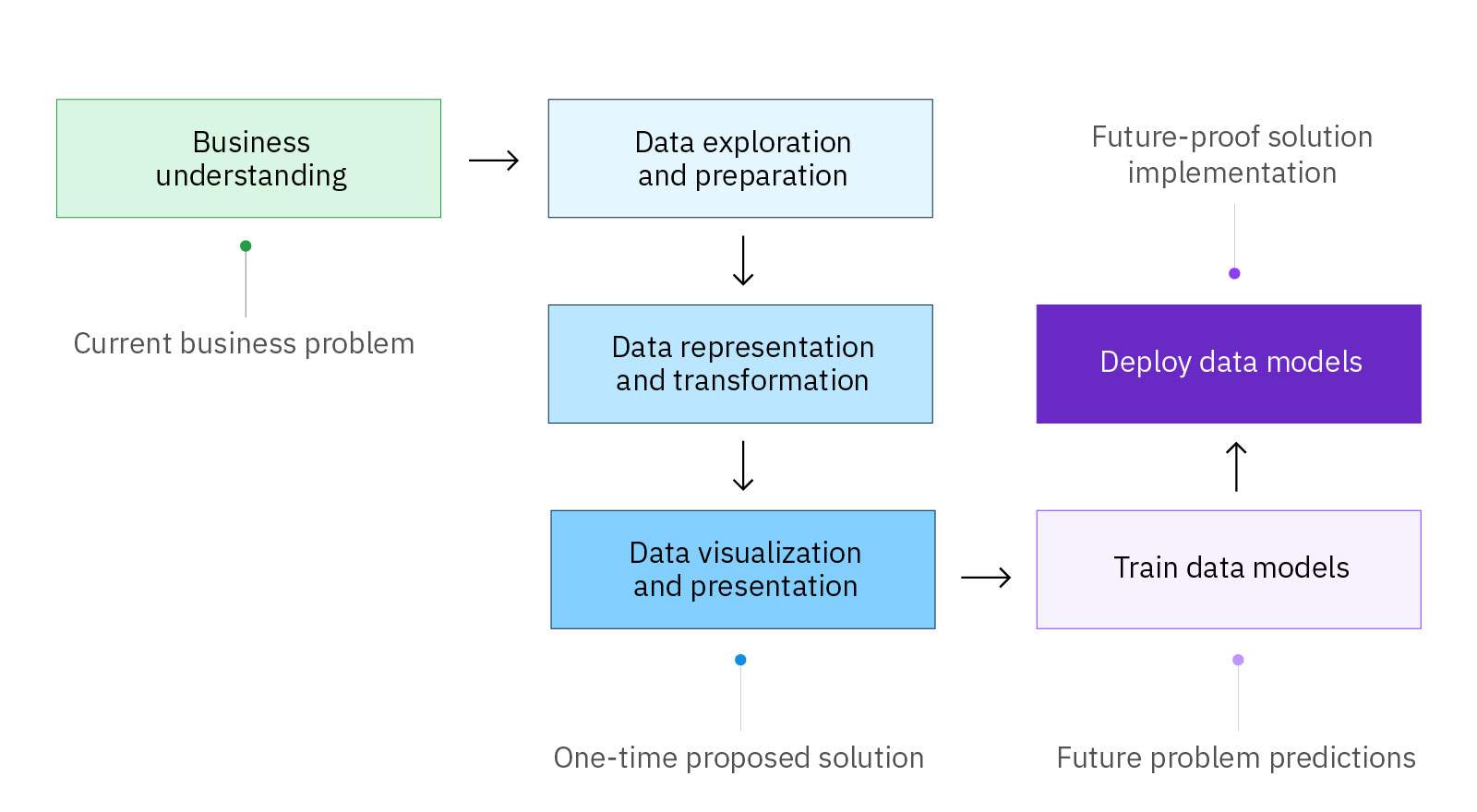
**Reinforcement Learning**

Humans and machines can learn through reinforcement learning. Reinforcement learning is a feedback-based, machine-learning technique. Through reinforcement learning, a machine determines how to behave in an environment by performing and observing the results of its actions. For each “good” action, the machine receives positive feedback (a reward). For each “bad” action, the machine receives negative feedback (a penalty). As a result, the machine learns automatically, through its experience and feedback.

Reinforcement learning doesn’t involve a specific goal. Rather, it involves learning from trial and error or “learn as you go”. Reinforcement learning is widely used in self-driven cars, drones, and other robotics applications.

# Step 6: Deploy data models

Deploying a model is the step in which the machine learning model is integrated into a business’ production environment. Data scientists perform this step using a company’s chosen tool set and software. Once a model is saved and deployed, the model can be used to continue making**better predictions for future solutions**. The model operates according to a schedule, and it must be maintained by a data scientist.



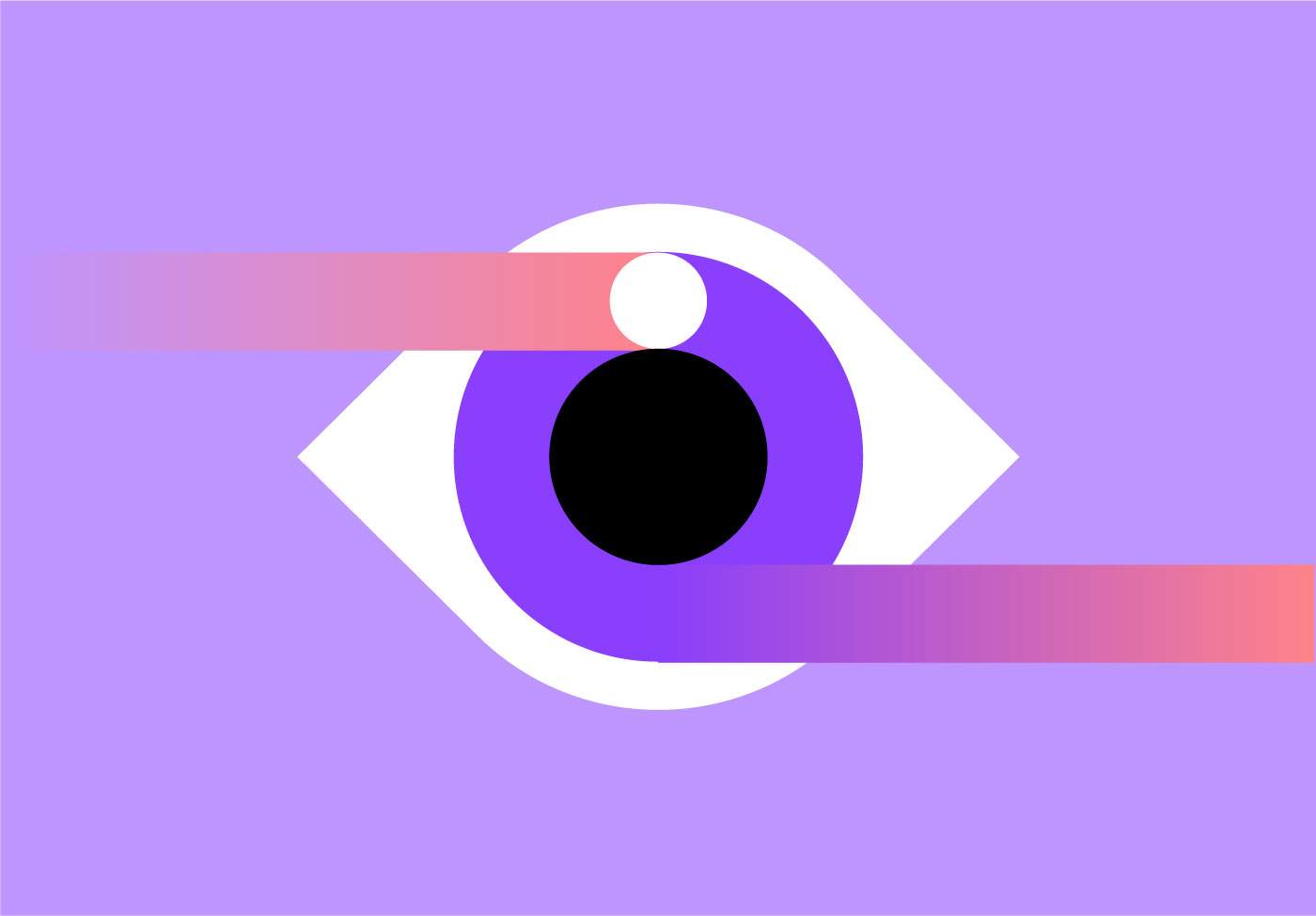
Step 6: Deploy data models

**Project scenario: Deploy data models**

Scott Hill, the team data scientist, deploys the model into production for GAXR using the IBM AutoAI technology. The deployment results in a machine learning file type. Future employee data sets can now be used to help mitigate the risk of employees leaving the company due to a long commute time.

Now that the model is in place, Scott must consider its future maintenance. He asks himself the following questions:

* “How will I evaluate the model’s performance in production?”
* “How frequently do I plan on re-training or running my model?”
* “Will the format of the production input data differ drastically from the model training data?”
* “Will it come in batches or as a stream?”
* “Do I need to run the model offline?”



# Data scientists have changed almost every industry

Data science has revolutionized the way data is perceived. There are many data science applications in healthcare, banking, e-commerce, manufacturing, and more. Big data companies like Amazon, Google, and Facebook use data science concepts for business insights and decisions for their organizations.

Data science projects are happening around the world, from predicting the side effects of medications to optimizing routes during traffic rush hours. Data science is used in a variety of industries and all kinds of organizations—some might even surprise you.

Data science can:

* Identify and predict disease and personalize recommendations in **healthcare**
* Optimize shipping routes in real time for **transportation**
* Accurately evaluate athletes’ performance in **sports**
* Prevent tax evasion and predict incarceration rates for **governments**
* Automate digital ad placement in**e-commerce**
* Improve online experiences for **gaming**
* Create algorithms to pinpoint compatible partners for **social media**[(opens in a new tab)](https://builtin.com/data-science/data-science-applications-examples)

**Source**: [22 Data Science Applications and Examples(opens in a new tab)](https://builtin.com/data-science/data-science-applications-examples), Built In, by Mae Rice and updated by Jessica Powers, June 2022

Explore examples of how companies are applying data science to improve and solve global problems.

**Healthcare**

one of the biggest opportunities for data science is in healthcare. Healthcare is an enormous industry, prime for disruption and new insights. Think about the huge amounts of electronic medical records, patient populations, and medical imaging.

Each person will generate 1 million gigabytes (GB) of health-related data in their lifetime, which is equivalent to about 300 million books. In that data is the secret to health and well-being.

Check out these two examples of data science in healthcare:

* Google developed a tool, [LYNA(opens in a new tab)](https://ai.googleblog.com/2018/10/applying-deep-learning-to-metastatic.html) (short for Lymph Node Assistant) to identify breast cancer tumors that metastasize to nearby lymph nodes. This can be difficult for the human eye to see, especially when the new cancer growth is small. In one trial, LYNA accurately identified metastatic cancer 99 percent of the time using its machine-learning algorithm. More testing is required, however, before doctors can use it in hospitals.
* [Oncora's software(opens in a new tab)](https://www.oncora.ai/) uses machine learning to create personalized recommendations for current cancer patients based on data from past ones. Healthcare facilities using the company’s platform include [UT Health San Antonio and Scripps Health(opens in a new tab)](https://www.prnewswire.com/news-releases/scripps-health-and-ut-health-san-antonio-purchase-oncoras-software-to-improve-physician-documentation-and-enhance-personalized-cancer-care-301523973.html). Their radiology team collaborated with Oncora data scientists to mine 15 years’ worth of data on diagnoses, treatment plans, outcomes, and side effects from more than 50,000 cancer records. Based on this data, Oncora’s algorithm learned to suggest personalized chemotherapy and radiation regimens.

**Source**: [22 Data Science Applications and Examples(opens in a new tab)](https://builtin.com/data-science/data-science-applications-examples), Built In, by Mae Rice and updated by Jessica Powers, June 2022



**Transportation and logistics**

More cities, transit organizations, and departments of transportation are using data science to solve problems and prioritize investments.

Check out these two examples of data science in transportation:

* [StreetLight(opens in a new tab)](https://www.streetlightdata.com/) uses data science to model traffic patterns for cars, bikes, and pedestrians on North American streets. Based on a monthly influx of trillions of data points from smartphones, in-vehicle navigation devices and more, Streetlight’s traffic maps stay up-to-date. The company’s maps inform various city planning enterprises, including commuter transit design.
* The United Parcel Service, [UPS(opens in a new tab)](https://www.ups.com/us/en/Home.page), uses data science to optimize package transport from drop-off to delivery. The company’s integrated navigation system, ORION, helps drivers choose fuel-efficient routes. [ORION has saved(opens in a new tab)](https://about.ups.com/ae/en/newsroom/press-releases/innovation-driven/ups-to-enhance-orion-with-continuous-delivery-route-optimization.html) UPS approximately 100 million miles and 10 million gallons of fuel per year with the use of advanced algorithms, AI, and machine learning.

**Source:**[22 Data Science Applications and Examples(opens in a new tab)](https://builtin.com/data-science/data-science-applications-examples), Built In, by Mae Rice and updated by Jessica Powers, June 2022



**Sports**

The sports industry also benefits from data science, which can be used to calculate statistics across different sports.

Check out these two examples of data science in sports:

* [WHOOP(opens in a new tab)](https://www.whoop.com/) makes wearable devices that track athletes’ physical data, like resting heart rate, sleep cycle, and respiratory rate. The goal is to help athletes understand when to push their training and when to rest, and to make sure they’re taking the necessary steps to get the most out of their body.
* [Trace(opens in a new tab)](https://traceup.com/) provides soccer coaches with recording gear and an AI system that analyzes game film. Players wear a tracking device, called a Tracer, while its specially designed camera records the game. The AI bot then takes that footage and stitches together all of the most important moments in a game—from shots on goal to defensive lapses and more. This technology allows coaches and players to have more detailed insights from game film.

**Source**: [22 Data Science Applications and Examples(opens in a new tab)](https://builtin.com/data-science/data-science-applications-examples), Built In, by Mae Rice and updated by Jessica Powers, June 2022



**e-Cemmerce**

e-Commerce has exploded in popularity since the internet has become a primary mode of communication for many people. Companies use data science to track everything involved in the customer journey, from marketing to purchases to consumer trends.

Does this sound familiar? For example, if you use social media and search for a product online, you might see ads for that same product **everywhere** immediately afterwards across your social media, right?! This is an example of how e-commerce companies are using data science for advertising.

Check out this example of data science in e-commerce:

* [Instagram(opens in a new tab)](https://www.instagram.com/) uses data science to target its sponsored posts, which sell everything from trendy sneakers to influencers posting sponsored ads. The company’s data scientists pull data about users, including age and education. From there, the team crafts algorithms that convert users’ likes and comments, their use of other apps, and their web history into predictions about the products they might buy.

**Source:** [22 Data Science Applications and Examples(opens in a new tab)](https://builtin.com/data-science/data-science-applications-examples), Built In, by Mae Rice and updated by Jessica Powers, June 2022



**Social platforms**

The rise of social networks has completely altered how people socialize. Now that many relationships begin online, data about people impacts who they get to know next.

Check out this example of data science shaping human connection:

* When singles match on [Tinder(opens in a new tab)](https://builtin.com/company/tinder), they can thank the company’s data scientists. A carefully crafted algorithm works behind the scenes, boosting the probability of matches. It prioritizes matches between active users, users near each other, and users who seem like each other’s “types” based on their swiping history.

**Source**: [22 Data Science Applications and Example(opens in a new tab)](https://builtin.com/data-science/data-science-applications-examples)s, Built In, by Mae Rice and updated by Jessica Powers, June 2022



**Note**:

There’s really no limit to the number or kind of organizations that can potentially benefit from the opportunities data science is creating. If you can think of it, it probably can be done!

In conclusion, data science is being applied across industries in many ways:

* Data science has changed the way pharmaceutical drugs are discovered and developed for people.
* Retailers can develop new products that respond to the needs of customers.
* Companies can detect and prevent cybersecurity threats.
* Organizations can work on solving global challenges like climate change, poverty, inequity, and terrorism.

Data tells businesses a lot and provides a potentially limitless horizon for research and technology.

**The data science project team**

Data science isn’t just about data sources and methodologies—it’s also about people. In practice, several people work on a team. It's not just one data scientist! Analysis results are only as good as the team that is responsible for collecting, analyzing, and interpreting the data.

A data science project team works on projects with the goal of extracting valuable insights from data to answer business problems.

In this lesson, you’ll learn about and compare the unique roles of the:

* Data analyst
* Data scientist
* Data engineer

**Joseph,Data Analyst**

**What does Joseph do?**

Joseph is in the office day-to-day, working with structured data in databases and using statistics.

* Joseph’s primary responsibility is to collect, organize, clean, and analyze mountains of data. He uses standardized methods and tools from his company to identify trends, find patterns, and make forecasts.
* Joseph knows the industry and uses his knowledge of the business.
* Joseph communicates findings and depicts insights to the business sponsors using data visualizations and presentations to help them make decisions and drive action.

**What are helpful characteristics and skills to have?**

* Methodical
* Critical thinker
* Skilled in data transformation tools to clean data and visualization tools to show insights
* Great communicator with presentation skills

**zhanna,Data Scientist**

**What does Zhanna do?**

Zhanna is in the office and involved from the beginning to the end of a project.

* Given the business problem, Zhanna develops a hypothesis to research and find hidden patterns.
* Zhanna works with data from many sources. Depending on the project, she might travel or go into the field to collect data and measurements. The data can be unstructured or structured, if it has already been “wrangled”.
* Zhanna conducts experiments to build custom models using the company’s data science methodology and tools.
* Zhanna uses techniques such as machine learning to build and train models to predict future outcomes.
* Overall, Zhanna transforms data into knowledge to produce actionable insights that can be used to improve future outcomes.

**What are helpful characteristics and skills to have?**

* Always curious and wondering “why?”
* Research and scientific mindset
* Problem solver
* Math, statistics, and machine learning
* Skilled in data analytics tools

**Ronnie,Data engineer**

**What does Ronnie do?**

Ronnie manages the data infrastructure for the company. His primary responsibility is to set up systems and processes for the data analysts and data scientists to use and rely on when working with data.

* Ronnie understands the flow of data and transforms large volumes of raw data into usable “pipelines” for projects.
* Ronnie is focused on tools and uses advanced programming techniques.
* Ronnie works on the steps to test and deploy machine-learning models into production for the company.

**What are helpful characteristics and skills to have?**

* Tech savvy
* Math, statistics, machine learning
* Skilled in programming
* Familiar with infrastructure architecture (IA)

**Interesting notes**

* Everyone on the team is responsible for **understanding the business problem** so they can work towards proposing a solution.
* Organizations use **different job titles**for the roles that work with data. Depending on the size of an organization, a person might have a combined role and responsibilities.
* There are other roles which aren’t covered here. For instance, you may find someone who is a **data journalist**. A data journalist maintains a client-focused mindset and their primary responsibility is to communicate results. Although this is not a technical role, a data journalist is a highly skilled communicator who turns facts into a compelling story with insights and data visualizations.

**You’ve come so far!**

And learned so much! You learned about the field of data science, including widely adopted methodologies, an example project as it moves through the steps of a data science methodology, the application of data science in our world, and the role data scientists and their fellow colleagues.

Now that you’ve completed this course, you should be able to:

* Define data science
* Recognize the importance of being curious to solve problems with data
* Differentiate between the fields of data analytics and data science
* Identify three widely adopted data science methodologies
* Explore a data project scenario and identify key tasks as it moves through a methodology
* Recognize industries and applications of data science that help solve global problems and discover innovations
* Compare the roles and characteristics of a data analyst, data scientist, and data engineer

**Key points to remember**

* 1

Data science combines the scientific method, math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and even storytelling to uncover and explain the business insights buried in data.

* 2

Data scientists use methodologies that include processes or activities to perform to obtain results. The methodologies are scientific, so the key is that they are repeatable.

* 3

Three widely used data science methodologies are: Cross-Industry Standard Process for Data Mining (CRISP-DM), Knowledge Discovery in Database (KDD), and Sample Explore, Modify, Model, and Assess (SEMMA).

* 4

Key activities that take place during a data science project are:

* 1. Business understanding
  2. Data exploration and preparation
  3. Data representation and transformation
  4. Data visualization and presentation
  5. Train data models
  6. Deploy data models
* 5

Those who work in data science use technologies and tools to build models. The models are used to predict outcomes or discover underlying patterns. The intent is to gain insights that lead to actions that improve future outcomes.

* 6

Data scientists use machine learning to train models. Machine learning is essentially teaching a computer to solve problems. It allows a machine to learn from data without programming it with rules. The machine can learn from the data it’s given.

* 7

Data visualizations should be effective, attractive, and impactive.

* 8

One of the most important characteristics of a data scientist is to always be curious!

* 9

Data science projects are happening in our world, across social platforms and industries like healthcare, transportation, sports, e-commerce, and much more.

* 10

Data science is a team effort! Data analysts, data scientists, and data engineers collaborate to solve business problems.