Introduction to Data Analytics and Data-Driven Decision Making

Introduction to Analyzing Data for Business Goals

# What is Data Analytics?

What is data? What is analytics? What is *data analytics* (DA)? *Data* can be characterized as facts, statistics, points in time, characters, numbers, bits, or information. *Analytics* involves making computations, building mathematical models, and using statistics in the process of conducting an analysis. The output of such an analysis can consist of descriptive statistics, predictive statistics, or any of a host of customized measures and models—both qualitative and quantitative. *Data analytics* simply combines the two with a greater focus on obtaining quantitative (numeric) rather than qualitative (descriptive) results.

Essentially, data analytics produce approximations of reality that allows you to draw meaningful insights that can help solve problems. But keep in mind that it’s up to you and your audience to decide what is considered to be “meaningful.” For many professionals including data analysts, business intelligence engineers, data scientists, financial analysts, or marketing analysts, data analytics is their main tool. It helps answer questions such as:

* Why did sales drop during a certain year?
* Why is this particular medicine not working?
* How can I recruit better employees and reduce turnover in my organization?

Analytical results can be produced as easily as dividing two numbers or as complicated as developing a predictive statistical model.

## How DA is Applied Across Industries

Most businesses today search for patterns and correlations in their historical data to support decision making. Below are some examples that highlight past and current applications of analytics across various industries.

### Communications, Media, and Entertainment

Companies such as YouTube, Amazon, and Netflix analyze their users’ browsing habits and patterns, which allows them to create or curate content tailored for specific target audiences.

### Finance

Financial institutions use analytics to identify and prevent fraudulent transactions, and specifically use *predictive* analytics to evaluate a person’s financial behavior and assign them a risk level for credit card approval.

### Healthcare

The healthcare industry has identified quite a few ways to use analytics in the fulfillment of its mission to improve the quality of life for chronically ill patients, provide personalized patient treatment, reduce the rate of hospital-acquired infections, assess and identify treatment risk factors more rapidly, and many more applications.

In addition, large medical centers have used free public health data to create visualizations that can help speed up the identification and analysis of healthcare information and track the spread of diseases.

### Insurance

The insurance industry has numerous opportunities to expand its use of analytics. For instance, insurance carriers can increase the personalization of services and pricing, which allows for additional consumer segmentation. Additionally, there are opportunities for fraud detection and greater industry transparency, in general.

### Transportation

Dataiku DSS (Data Science Studio) applied to freight, sea freight, road freight, and passenger transport uses predictive analytics with sensor data to predict needed maintenance schedules.

## The Four Types of Data Analytics

There are four commonly used types of data analytics that together provide a more comprehensive understanding of available data. They are listed below along with the questions that each type strives to answer and a very brief description.

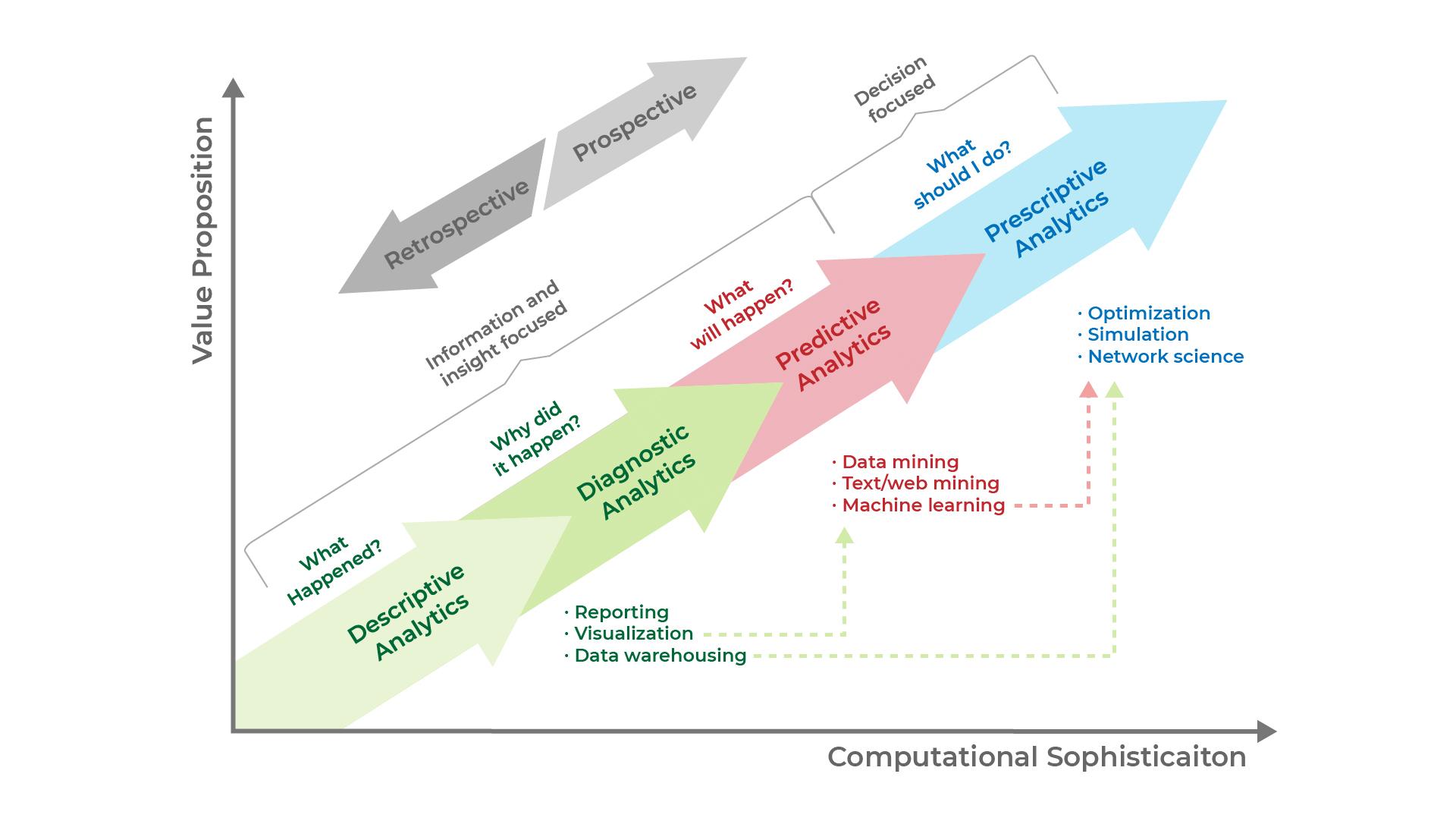
* **Descriptive:** What happened? Helps uncover valuable insight into the data being analyzed
* **Diagnostic:** Why did it happen? Helps understand the relationships and patterns in the data
* **Predictive:** What is likely to happen in the future? Helps forecast the future behavior of people and markets
* **Prescriptive:** What should I do about it? How should we respond to potential future events based on the analysis? Uses optimization and simulation algorithms to provide guidance and understanding on decisions and answers

Table 1 shows typical questions for each type of analytical technique for several industries.

| **Field** | **Descriptive Analytics** | **Diagnostic Analytics** | **Predictive Analytics** | **Prescriptive Analytics** |
| --- | --- | --- | --- | --- |
| Healthcare | How many patients went to the hospital/clinic last week? | Why did these patients go to the hospital? | Which patient is likely to need hospitalization? | Extra home treatment to prevent visits |
| Life science | What are the total medicines available by pharmaceuticals? | Why do we not have more medicines or pharmaceuticals? | Which pharmaceutical is likely to monopolize the production of medicines? | Increase the number of pharmaceuticals and medicines |
| Finance | Where was the stock market in the last 6 months? | Why was the stock market that high/low? | What will the level be for the stock market in the next 6 months? | Reduce downside risk by diversifying a portfolio |
| Politics | Who was elected last? | Why were they elected? | Who will be elected next? | Increase voter education to reduce downside societal impact |
| Education | How many students are learning online? | Why are they taking courses online? | How many more courses are needed to support them? | Increase the number of courses |
| Non-profit | How are non-profits independent? | Why are they effective? | How many more non-profits can be created? | Decrease barriers to creating non-profits |
| Human rights | What is the number of refugees in the world? | Why are there so many refugees? | What are the likely future sources of refugees? | Reduce colonialism, imperialism, 'isms and let people live |

*Table 1: Typical questions for conducting analytics*

A pictorial depiction of the progressive nature of the four types of analytics is shown in Figure 1(Delen & Ram, 2018).



*Figure 1: The progressive nature of the four types of analytics*

In an analytics project, you may use all of these types or a combination of only a few of them. For example, if you already know what problem you are trying to solve, you may be able to skip the descriptive analytics step and move forward with diagnostic and prescriptive analytics.

To get started on determining the right mix of analytics for your business or organization, you should start by asking a few questions:

* What data is currently available? Existing data can include customer, marketing, operational, or financial data. Determining what you need to collect is based on your ability to source the data and on what is readily available.
* How deep of a dive into the data will you need to take?
* What data are you currently using and how much additional data do you still need?

The answers to these questions can provide a better understanding of the task at hand and help you decide on a data analytics strategy. You can also determine which resources are available to you and what tools you will want to use to start your project.

## Common Terminology in DA

Below are several common terms and definitions associated with data analytics:

* **Data governance**: The process of establishing how data should be used and ensuring data integrity
* **Outlier**: A value that is significantly different from the general distribution
* **Python:** A computer programming language
* **R:** A computer programming language used often for statistical computing
* **Quantile:** A group of objects classified with similar characteristics
* **Correlation:** The degree to which two variables move together, usually expressed by a *correlation coefficien*t that ranges between -1 and 1
* **Data scientist:** A professional with a high degree of technical skill and knowledge (often computer science, statistics, design skills)
* **Data engineer:** An individual who maintains the systems that data scientists and data analysts use, and sometimes participates in the analysis
* **Algorithm:** Repeatable steps used to solve a problem
* **Unsupervised learning:** A type of machine learning in which outputs are unlabeled
* **Supervised learning:** A type of machine learning in which outputs are labeled

## Basic DA Workflow

The three major categories of tasks in the basic data analytics workflow are presented below:

### Data Preparation

As a first step, pre-processing, preparing, and manipulating raw or partially transformed data is required for all analyses ranging from modeling to visualization. Starting with a basic understanding of the contents of the various columns (i.e., text strings, numbers, etc.), the data needs to be characterized and understood, cleansed of outliers (significant values outside of the distribution), and processed to find errors (missing values, nulls, etc.).

### Analysis and Data Exploration

During this step, the goal is to understand the dataset and the characteristics of the data. The data needs to be loaded, stored, and then explored for completeness to understand how the variables relate to each other. One can use Python, R, or even Excel for data exploration once the data is loaded, and then run basic descriptive statistics (counts, means, modes), etc. One can also create various visualizations such as bar charts or line graphs to further explore the data.

### Presenting and Communicating Analytics Results

The bar charts, line graphs, and other visualizations developed during the analysis and data exploration step comprise a perfect accompaniment for textual reports presented to project stakeholders. As the saying goes, “a picture is worth a thousand words.” Well-organized tables are also useful. Visualization layer tools include Tableau and PowerBi. Visuals are much more “human-friendly” than data dumps as they can show trends and features that would not be apparent from raw data.

# The Value of Data Analytics to Business

Many organizations believe they already possess valuable data but recognize they are not exploiting that data to its full potential. They are unsure of how to use it in an actively beneficial manner. “Actively beneficial” is a key phrase as many data analytics projects fail—not because the analytics were carried out incorrectly, but because the projects were often designed in a way that ultimately did not lead to actionable or beneficial results. This is where *value creation* comes in. Value creation is a process that establishes a clear, direct link between available data and the business decisions that an organization needs to make.

The notion of creating value can refer to a number of different ways that an organization uses its data. These include improving business operations, making strategic decisions, increasing the efficiency with which products and services are delivered, anticipating customer and workforce needs, developing new products, and so forth. While any of these activities can be improved by using available data, many organizations struggle when trying to identify specifically *which* data is most likely to provide a significant strategic business impact.

The discipline of analytics has two overlapping but distinct categories: *data analytics* and *business analytics*. Value creation combines both types of analytics to develop actionable strategies based on an organization’s needs and available data.

*Data analytics* is usually performed by computers and has a focus on identifying patterns that help develop an overall understanding of the data. *Business analytics*, on the other hand, focuses on the application of that understanding and insight to business problems. Although these two types of analytics are distinct, they are symbiotic in the sense that business analytics depends on data analytics to provide insights while data analytics depends on business analytics to have an impact.

To highlight the similarities and differences between data analytics and business analytics, let’s examine the variations of skills needed for each (Table 2).

| **Skills for Data Analytics** | **Overlap Skills** | **Skills for Business Analytics** |
| --- | --- | --- |
| * Data profiling * Data preparation * Statistical analysis * Data mining * Analytic modeling | * Basic understanding of business * Basic data management understanding * Basic data visualization skills | * Business subject matter expertise * Understanding of business strategy and tactics * Skills for analyzing cause and effect * Business management (decision-making ability based on analysis) |
|
|
|
|

*Table 2: Skills required for data and business analytics*

How can you use data to create business value? In this context, the goal is to bring value to owners, customers, and employees from a business perspective. To identify specifically *what* has value, it’s necessary to look at the impact it has on the business, such as increasing the return on investment, reducing costs, or reducing downside risks. You establish some goals and then work backward to identify the actions you must take to reach them, and then determine the data analyses that would help you choose the best actions. This means you have to develop an understanding of your organization by talking to its owners, customers, and employees.

For example, suppose you work at a hospital and your goal is to improve patient outcomes by having them leave the hospital sooner and in a healthier condition. This would generate more profit for the hospital’s owners and perhaps motivate the employees to be more efficient (assuming they receive a portion of the increased profits). Of course, the level of care needed for healthier outcomes might be more costly, which would reduce the potential for profit. Value creation must take the full process into account.

All variables (cost, risk, return) matter in value creation. You have a limited amount of data available so you must both identify the data that is most relevant for your organization and optimize the use of that data. You must decide what specifically constitutes “value” for your organization in terms of its goals (e.g., increasing profit, reducing staff turnover, improving the product line, etc.), and then look at whether the actions taken as a result of your data analytics actually helped move your organization toward those goals. Did profitability increase? Did staff turnover decrease? Did the quality of your product improve? Of course, you want to quantify the results in terms of dollars or percentages, or evaluate growth using a time-based metric.

Naturally, input variables and their associated values depend on the industry with which they are associated. If you work on a farm and want to increase milk production (because that’s your most profitable product), you can use computer modeling to estimate the impact on production that increasing the amount of feed given to the cows or changing the milking schedule will have.

Here is a simplified step-by-step process that summarizes value creation:

* Identify a variable of interest and determine how it relates to improving value (e.g., reducing downside risk, increasing returns, or reducing costs).
* Engage (“socialize”) with relevant people to develop a full understanding of the organizational goals and values (i.e., people such as audiences, clients, owners, or employees)
* Take appropriate action based on the results of your data analyses to execute, capture, or create value.

# Standard Processes for Data Analytics

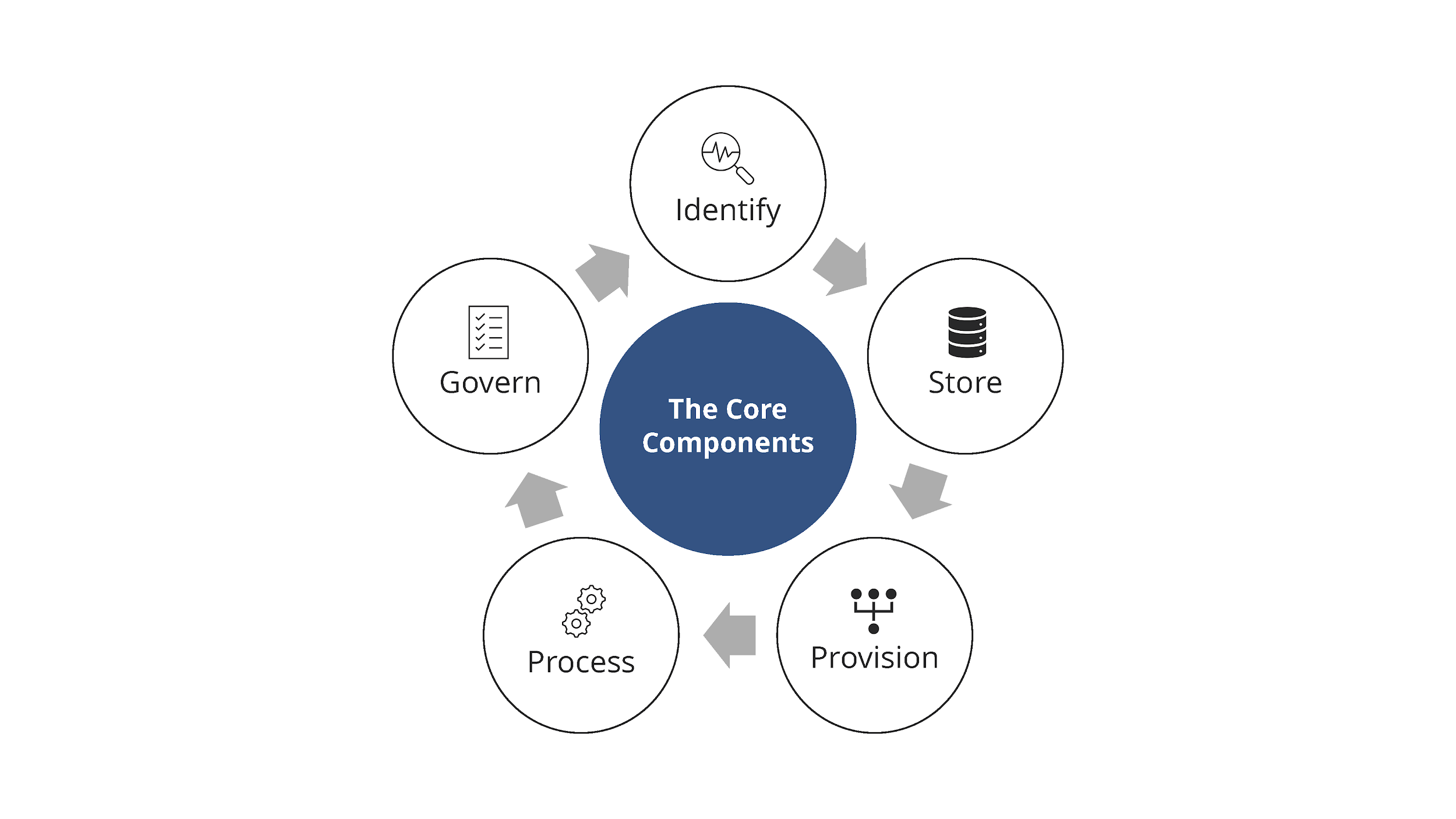
It is important to gather data with intent and focus. You want high-quality data (i.e., data having few errors) that can be used to evaluate, solve, and answer important questions, and that covers a sufficient time span to help answer your questions. Remember, data is purposeful only if it can be used for making computations that lead to actionable conclusions.

## Developing a Data Strategy

In developing a data strategy, you must define the specific business needs you want to address, create a well-defined problem statement, identify your goals (the problems you are trying to solve), and determine how analytics can help you achieve those goals (solve those problems).

A data strategy can be described as *a set of choices and decisions that when put together allow you to chart a high-level course of action toward achieving high-level goals.* This data strategy includes a business plan for using the information you obtain to reach specific business goals such as gaining a competitive advantage. An effective data strategy requires you to understand the fundamental data needs inherent in your business strategy.

The [SAS Institute](https://www.sas.com/en_us/home.html) defines data strategy in terms of five core components: *identify, store, provision, process*, and *govern*. These core components work together as building blocks to comprehensively support data management across an organization (Figure 2).



*Figure 2: The core components supporting data management across and organization*

## Defining the Data Needs

By choosing the right data and its sources, you can develop business-relevant analytics that can be put to use building models that predict and optimize business outcomes. With the right data, you can:

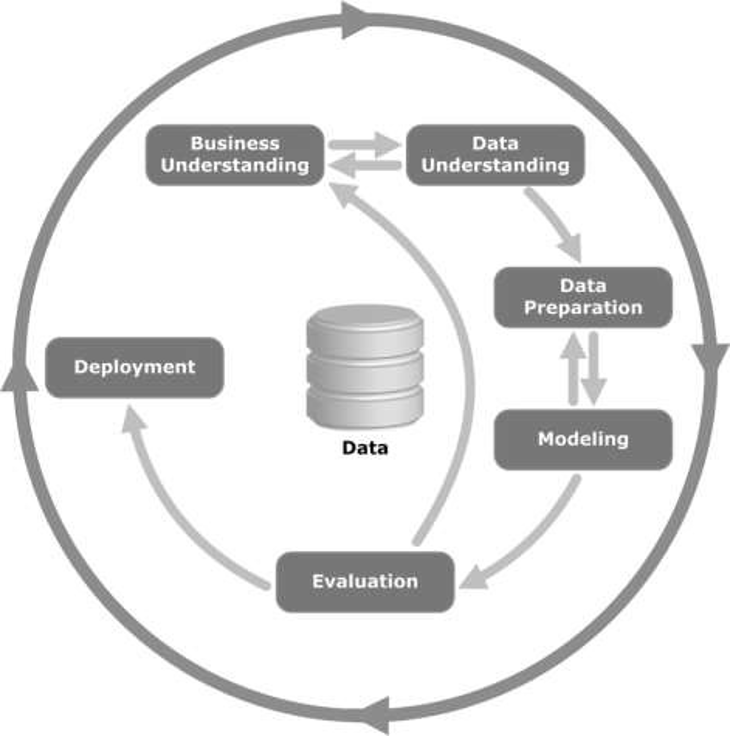
* Use the data in your business
* Make data-driven, evidence-based business decisions
* Understand what your customers think and feel
* Uncover the latest trends relevant to your business
* Deliver smarter products and services
* Improve internal operations
* Stay compliant with data protection regulations

The most popular data strategy methodologies in use are the *C*r*oss-Industry Standard Process for Data Mining* (CRISP-DM) and the *Sample, Explore, Modify, Model, and Assess* (SEMMA) approach.

### CRISP-DM (Cross Industry Standard Process for Data Mining) Methodology

The CRISP-DM methodology consists of a cyclical sequence of phases commonly used by data mining experts for traditional business intelligence (BI) data mining and data strategy development. Figure 3 (below) summarizes this cycle, which is dynamic, features bidirectional movement between the six phases, and entails multiple iterations. In the diagram:

* The Iterative CRISP-DM process is shown in the outer circle.
* The most significant dependencies between phases are shown.
* Subsequent phases depend on results from the preceding phases.
* Returning to earlier phases is possible before moving forward.



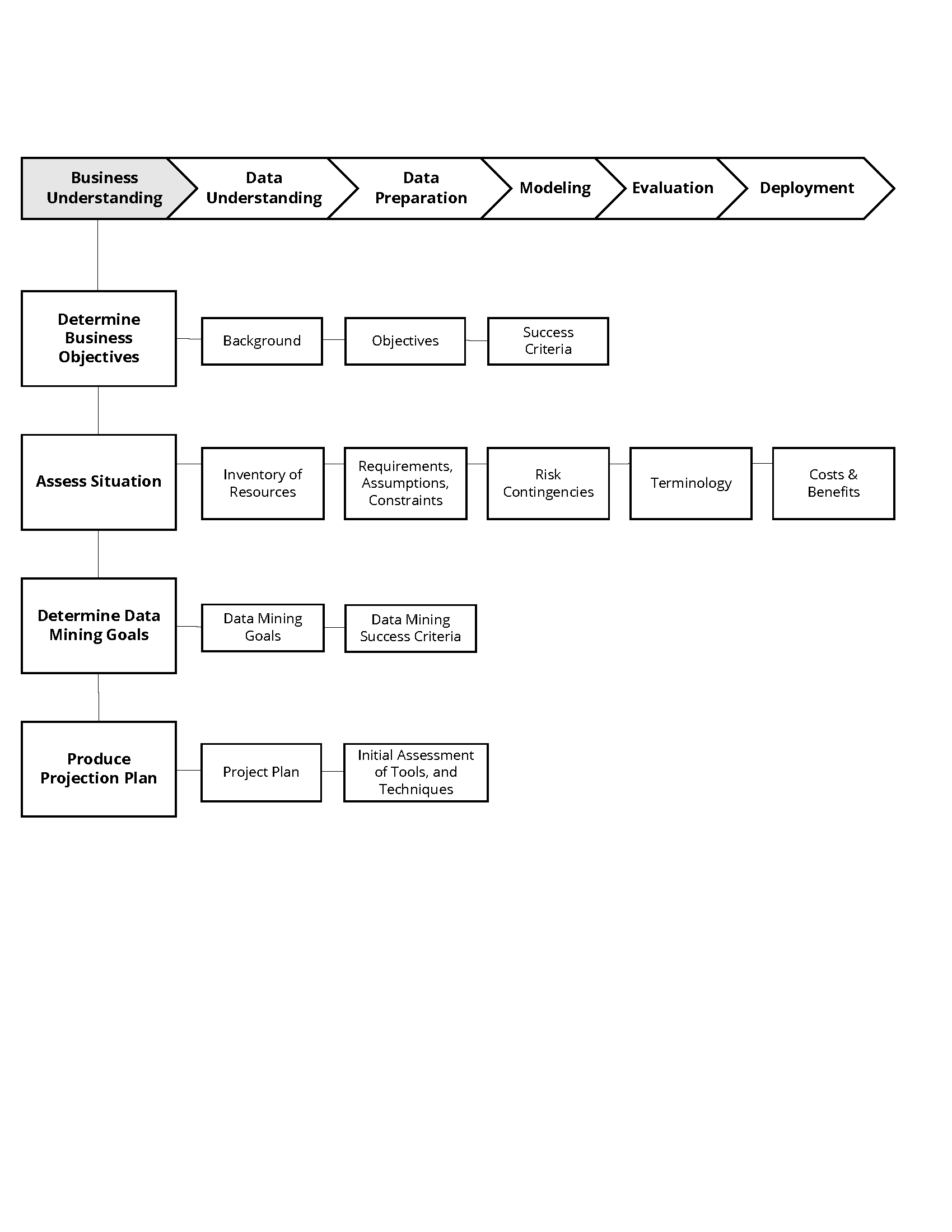
*Image source:* [*https://commons.wikimedia.org/wiki/File:CRISP-DM\_Process\_Diagram.png*](https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png)

*Figure 3: The CRISP-DM process*

Specific activities for each phase are highlighted below:

### Business/Research Understanding Phase (Figure 4)

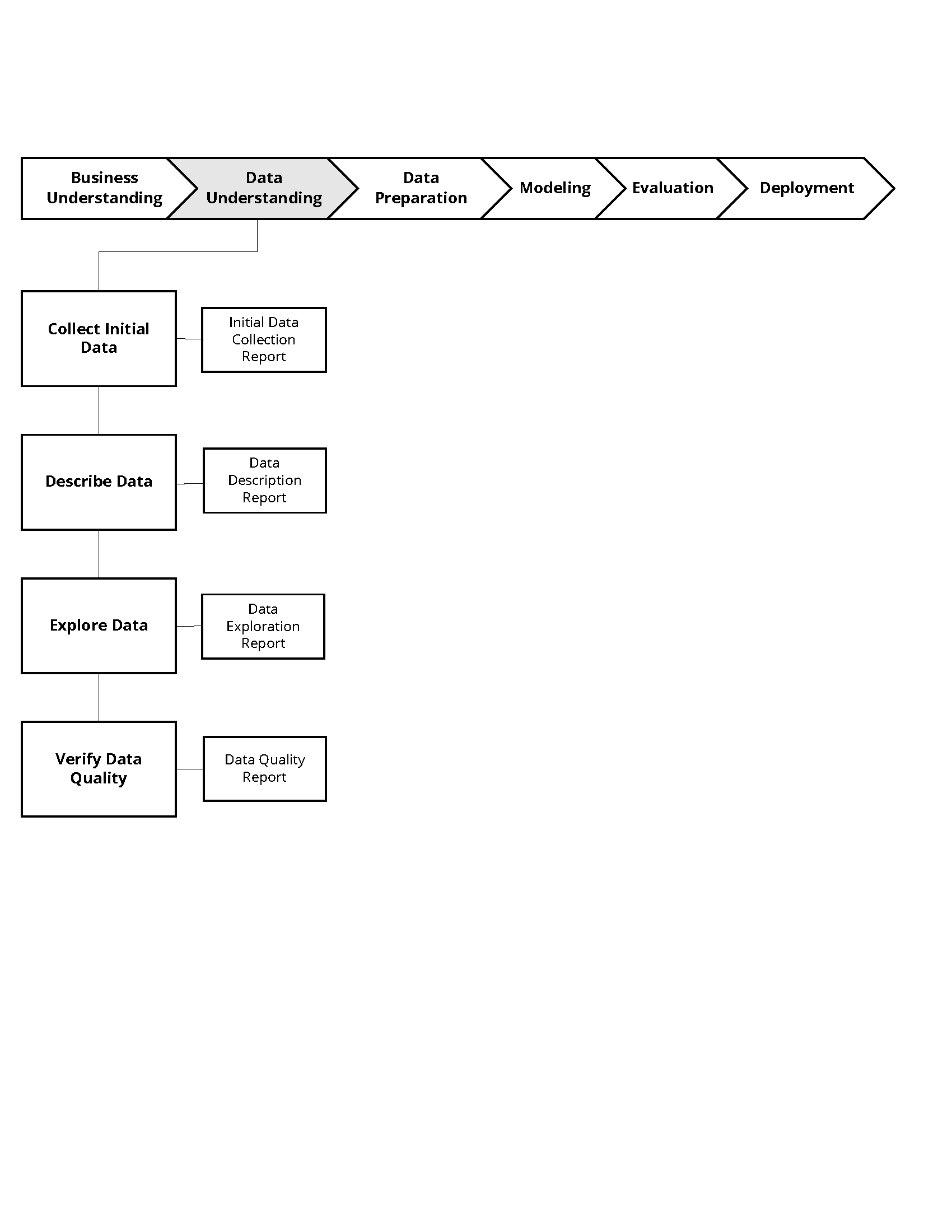
* Define project requirements and objectives
* Translate objectives into a data mining problem definition
* Prepare a preliminary strategy to meet the objectives



*Figure 4: The business understanding phase*

### Data Understanding Phase (Figure 5)

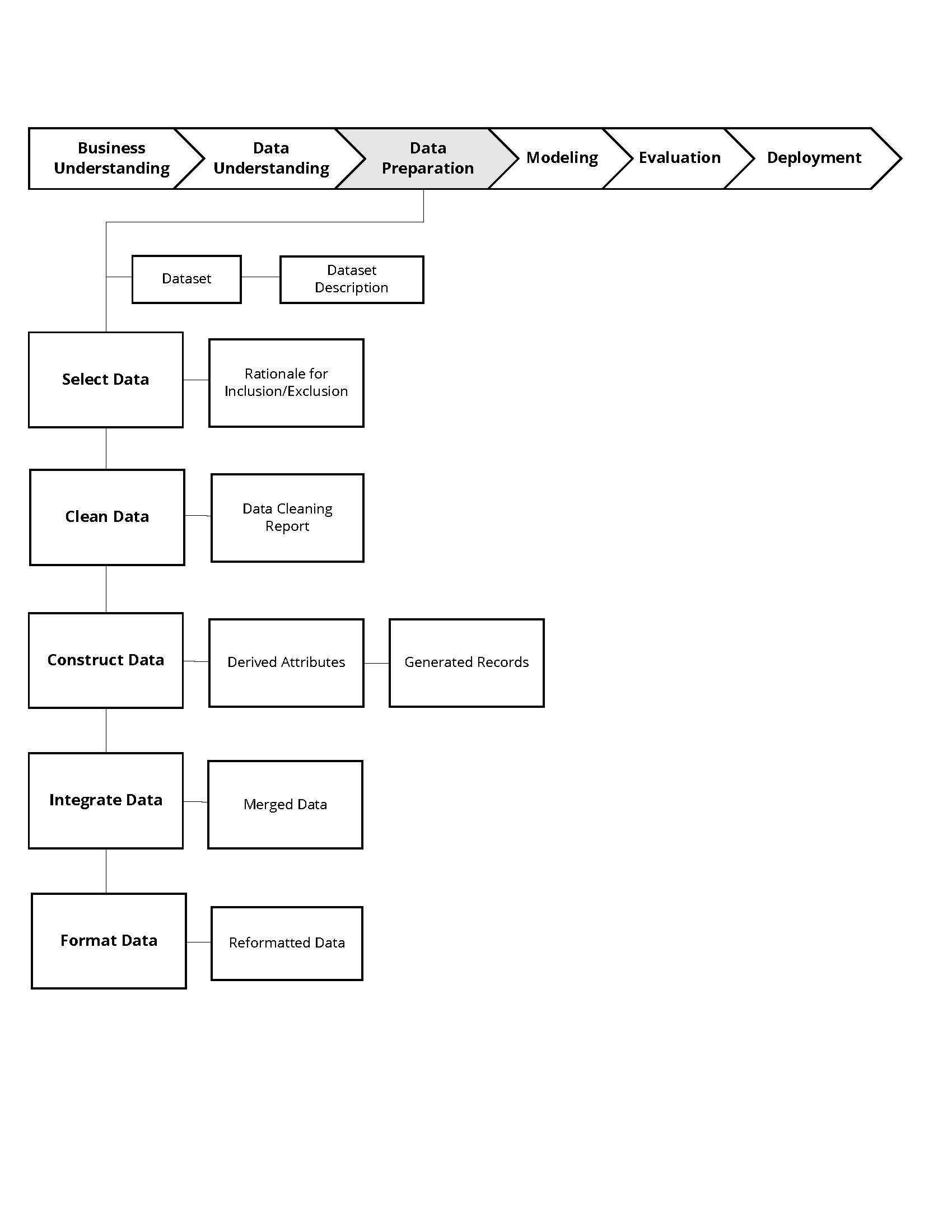
* Collect data
* Perform exploratory data analysis (EDA)
* Assess data quality
* Select interesting subsets (optional)



*Figure 5: The data understanding phase*

### Data Preparation Phase (Figure 6)

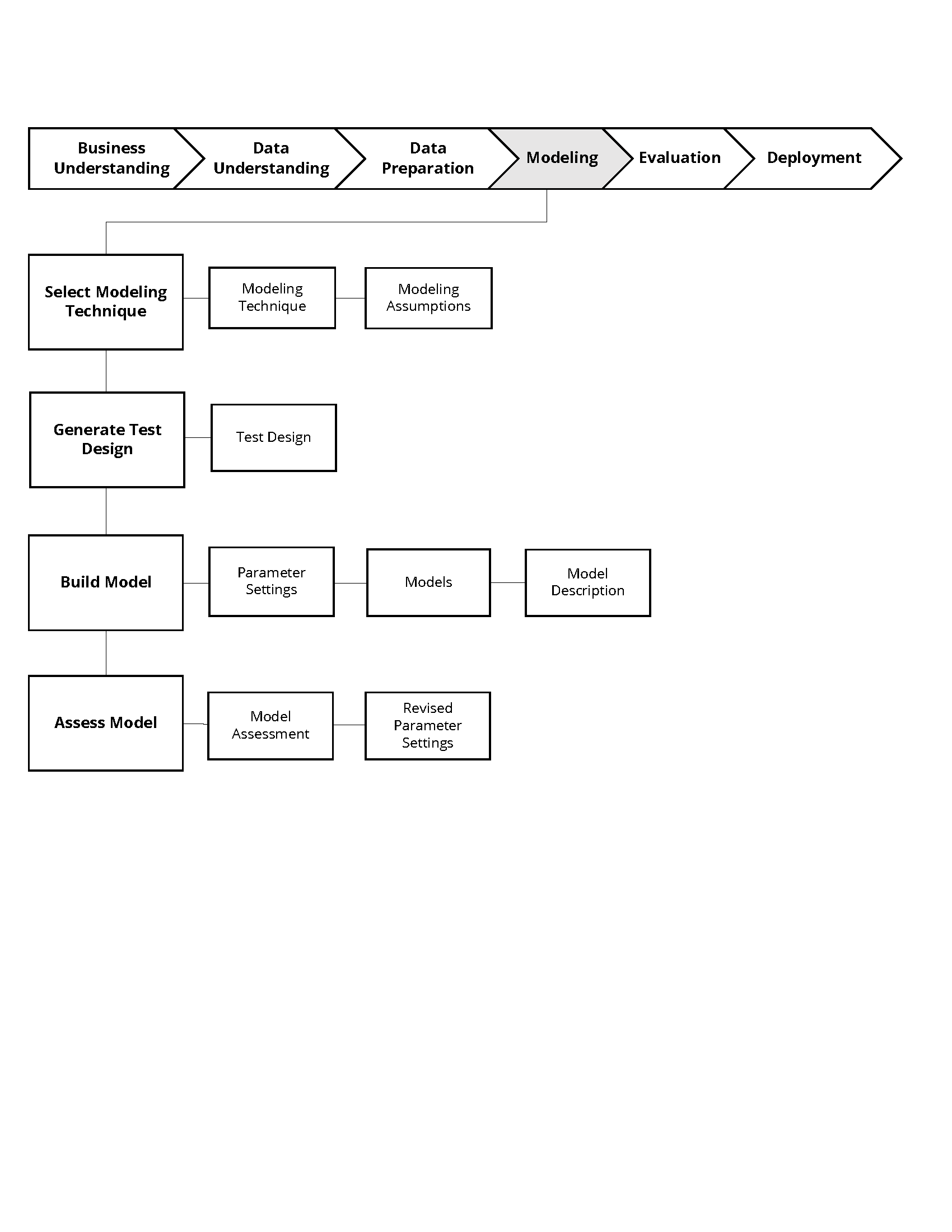
* Prepare for modeling in subsequent phases
* Select cases and variables appropriate for analysis
* Cleanse and prepare data so it is ready for modeling tools
* Perform transformation of certain variables, if needed



*Figure 6: The data preparation phase*

### Modeling Phase (Figure 7)

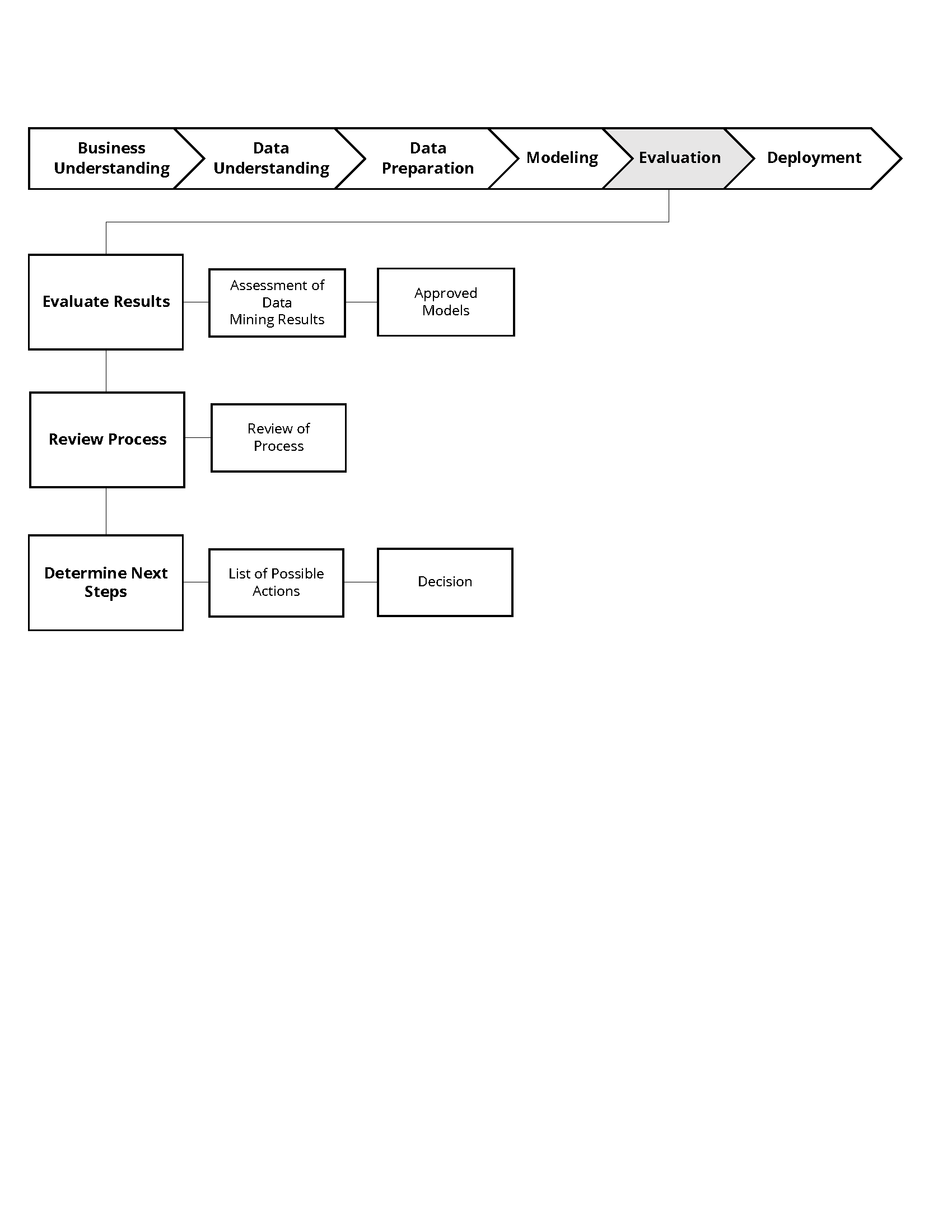
* Select and apply one or more modeling techniques
* Calibrate model settings to optimize results
* Prepare additional data if required to support a particular technique



*Figure 7: The modeling phase*

### Evaluation Phase (Figure 8)

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

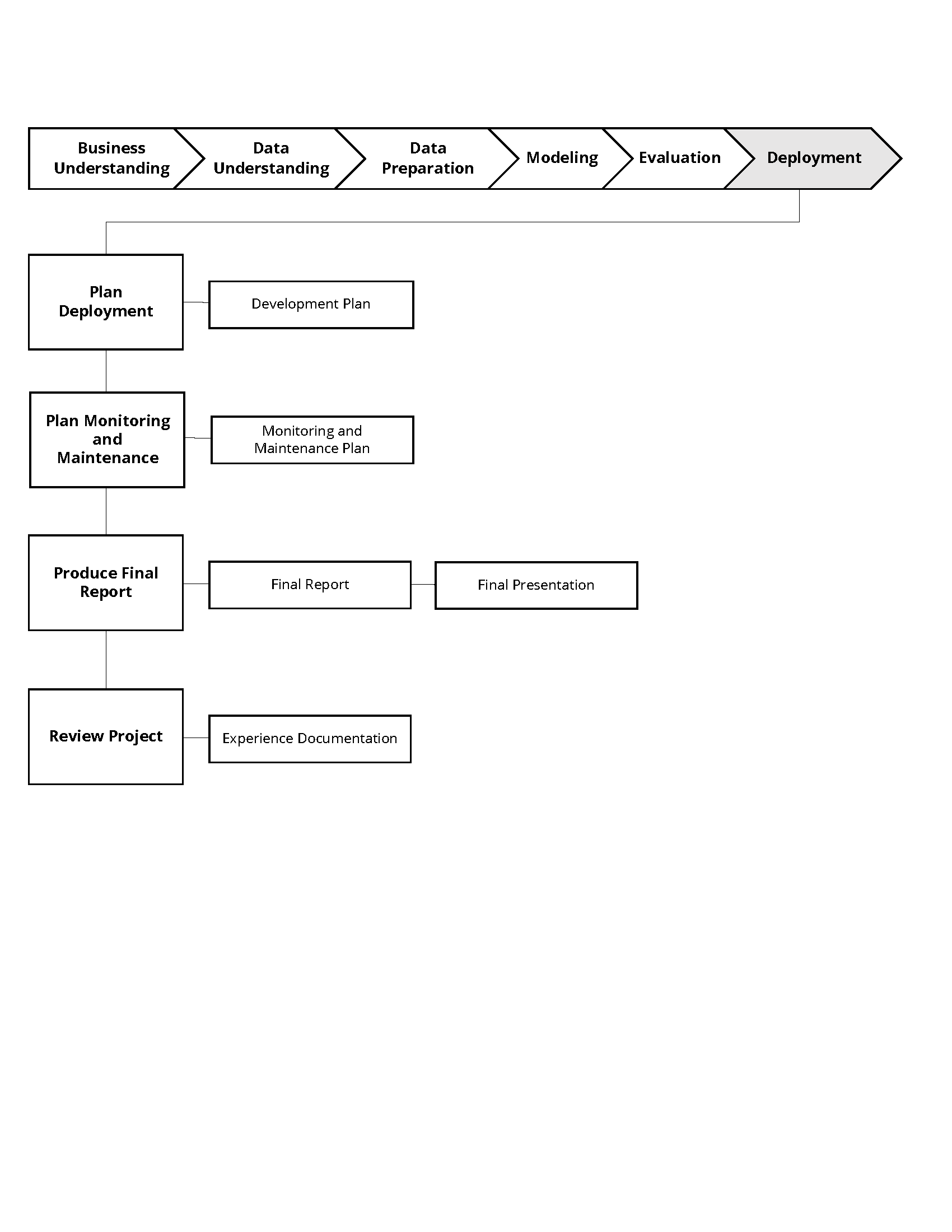


*Figure 8: The evaluation phase*

### Deployment Phase (Figure 9)

* Make use of models created
* Generate the report (simple deployment example)
* Implement a parallel data mining effort in another department (complex deployment example)

Note: In some businesses, the customer often carries out the deployment based on your model.



*Figure 9: The deployment phase*

### SEMMA Methodology (Sample, Explore, Modify, Model, and Assess)

SEMMA is another data strategy developed by the SAS Institute. It guides users through tools in the SAS Enterprise Miner application for data mining problems. SEMMA is an acronym that stands for *sample, explore, modify, model*, and *assess* (Figure 10).

* **Sample** the data by creating one or more data tables. The samples should be large enough to contain significant information, yet small enough to process.
* **Explore** the data by searching for anticipated relationships, unanticipated trends, and anomalies in order to gain understanding and ideas.
* **Modify** the data by creating, selecting, and transforming the variables so they focus on the model selection process.
* **Model** the data by using analytical tools to search for a combination of data elements that reliably predict a desired outcome.
* **Assess** the data by evaluating the usefulness and reliability of the findings from the data mining process.



*Figure 10: The SEMMA methodology*

## Text Mining

Sourcing or creating value from text-based data (whether structured or unstructured) is very relevant these days. Textual data is often produced and consumed by humans, and contains various kinds of information such as news, corporate documents, email content and more. It can be used in contextual advertising, customer care service, content enrichment, and more.

To work with textual data, you begin with a question you’re trying to answer or a problem you’re trying to solve. Next, you look at the text itself to see if it has the potential to help you answer that question or solve that problem. You may need to store your data using a NoSQL structure (such as MongoDB, OrientDB, or CouchDB), after which you can use Python, R, or similar tool to perform natural language processing (NLP) to interpret the data. NLP can identify parts of speech, process grammar, engage in syntactic parsing, etc. As an example, you might feed in the transcript of a phone call and determine various characteristics about the speakers such as their emotional states. You can then present your findings to your audience, client, or customer.

# Ethics and Data Privacy

Ethics is of paramount importance in business. While striving to achieve the goals of business owners and their employees is of course important, it is not so important that ethics can be ignored. Our global society has produced numerous examples of unethical practices including child labor, the sale of tobacco to children, and even governments conducting war in the name of questionable objectives. Nevertheless, achieving business objectives and being ethical are not necessarily mutually exclusive. We do have the ability and perhaps even the responsibility to make decisions about how we use our skills that take into consideration the ethical impact of our work.

In terms of data privacy, you are most likely transferring information about your own online behavior at this very moment, albeit anonymously. For instance, your email application may be reading your messages (without identifying you by name, but rather identifying you categorically). So be careful about your online behavior. Encrypt your messages or, better yet, respect the power lurking behind data and seek to keep private information private.

# References:

* Chapman, P., Clinton, J., Kerber, R., Khabaza, T., et al. (2000). [CRISP-DM 1.0.](https://the-modeling-agency.com/crisp-dm.pdf)
* Delen, D. & Ram, S. (2018). [Research Challenges and Opportunities in Business Analytics.](https://www.tandfonline.com/doi/full/10.1080/2573234X.2018.1507324)
* Kempe, S. (2017). [DI&A Webinar: Descriptive, Prescriptive, and Predictive Analytics.](https://acadgild.com/blog/different-types-of-data-analytics)
* SAS Institute Inc. (2020). [Introduction to SEMMA.](https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=14.3&locale=en)
* SAS Institute Inc. (2018). [The 5 Essential Components of a Data Strategy.](https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/5-essential-components-of-data-strategy-108109.pdf)