# Data Science Methodology

# Contents

# What is Data Science Methodology?

Despite the recent increase in computing power and access to data over the last couple of decades, our ability to use the data within the decision making process is either lost or not maximized as all too often, we don't have a solid understanding of the questions being asked and how to apply the data correctly to the problem at hand.

**Meaning of methodology:**

A system of methods used in a particular area of study or activity.

➡ It's important to consider it because all too often there is a temptation to bypass methodology and jump directly to solutions. Doing so, however, hinders our best intentions in trying to solve a problem.

➡ The data science methodology discussed here is outlined by John

Rollins, a seasoned and senior data scientist currently practising at IBM.

## In a nutshell:

The **Data Science Methodology** aims to answer the following 10 questions in this prescribed sequence

## From problem to approach:

1) What is the problem that you are trying to solve?
2) How can you use data to answer the question?

## Working with the data:

3) What data do you need to answer the question?
4) Where is the data coming form (identify all sources) and how will you get it?
5) Is the data that you collected representative of the problem to be solved?
6) What additional work is requires to manipulate and work with the data?

## Deriving the answer:

7) In what way can the data be visualised to get to the answer that is required?

8) Does the model used really answer the initial question or does it need to be adjusted?

9) Can you put the model into practice?

10) Can you get constructive feedback into answering the questions?

# Introduction to CRISP-DM

## 1) Data Science Methodologies

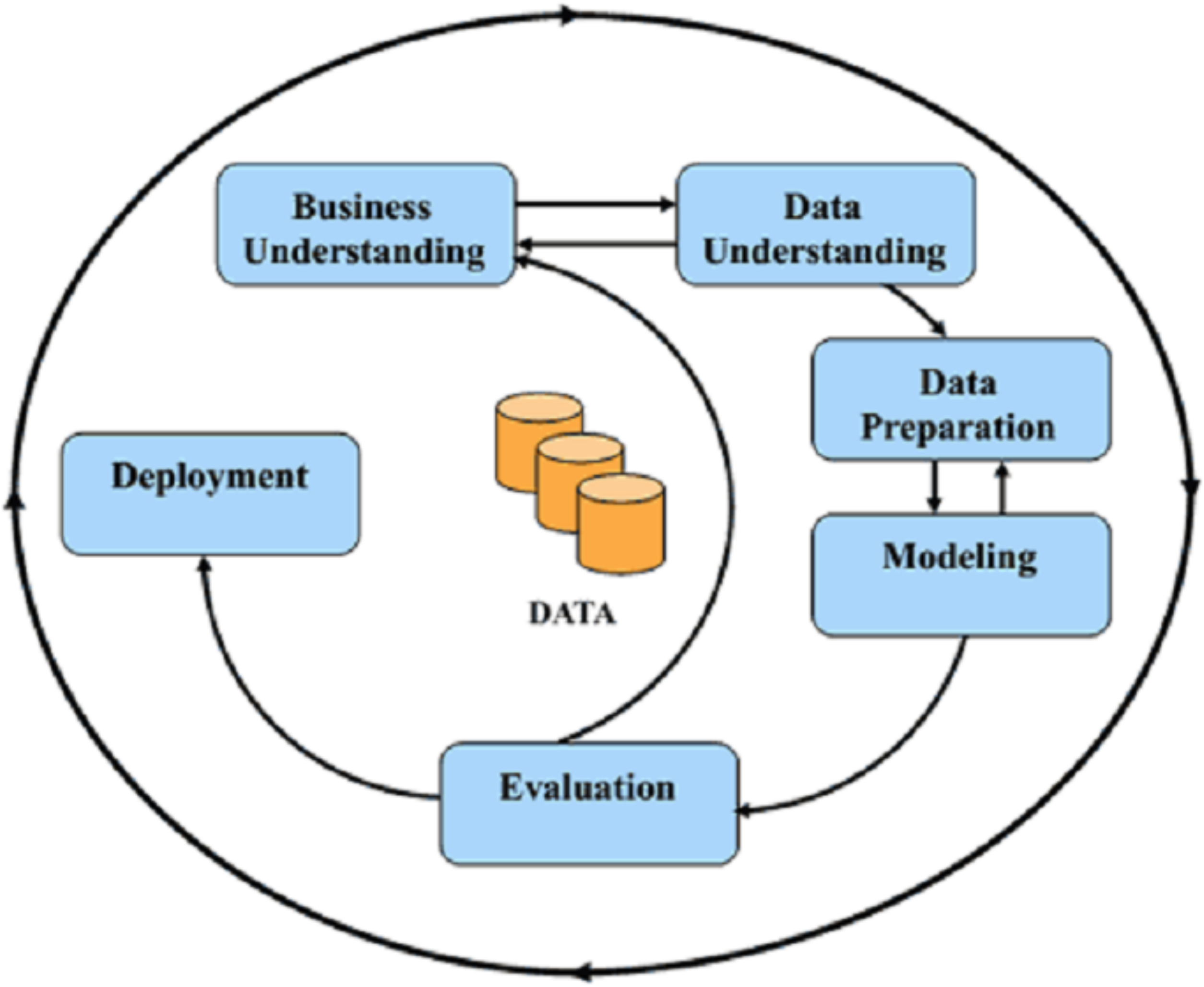
Here we have discussed the data science methodology by John Rollins. However, it is not the only methodology that you will encounter in data science. For example, in data mining, the Cross Industry Process for Data Mining (CRISP-DM) methodology is widely used.

## 2) What is CRISP-DM?

The CRISP-DM methodology is a process aimed at increasing the use of data mining over a wide variety of business applications and industries.

The intent is to take case specific scenarios and general behaviours to make them domain neutral.

CRISP-DM is comprised of six steps with an entity that has to implement in order to have a reasonable chance of success. The six steps are shown in the following diagram:



## 1) Business Understanding:

This stage is the most important because this is where the intention of the project is outlined.

Foundational Methodology(methodology by John Rollins) and CRISP-DM are aligned here.

It requires communication and clarity. The difficulty here is that stakeholders have different objectives, biases, and modalities of relating information. They don’t all see the same things or in the same manner. Without clear, concise, and complete perspective of what the project goals are resources will be needlessly expended.

## 2) Data Understanding:

Data understanding relies on business understanding. Data is collected at this stage of the process. The understanding of what the business wants and needs will determine what data is collected, from what sources, and by what methods. CRISP-DM combines the stages of Data Requirements, Data Collection, and Data Understanding from the Foundational Methodology outline.

## 3) Data Preparation:

Once the data has been collected, it must be transformed into a useable subset unless it is determined that more data is needed. Once a dataset is chosen, it must then be checked for questionable, missing, or ambiguous cases. Data Preparation is common to CRISP-DM and Foundational Methodology.

## 4) Modelling:

Once prepared for use, the data must be expressed through whatever appropriate models, give meaningful insights, and hopefully new knowledge.

This is the purpose of data mining: to create knowledge information that has meaning and utility. The use of models reveals patterns and structures within the data that provide insight into the features of interest.

Models are selected on a portion of the data and adjustments are made if necessary. Model selection is an art and science. Both Foundational Methodology and CRISP-DM are required for the subsequent stage.

## 5) Evaluation:

The selected model must be tested. This is usually done by having a pre-selected test, set to run the trained model on. This will allow you to see the effectiveness of the model on a set it sees as new. Results from this are used to determine efficacy of the model and foreshadows its role in the next and final stage.

## 6) Deployment

In the deployment step, the model is used on new data outside of the scope of the dataset and by new stakeholders.

The new interactions at this phase might reveal the new variables and needs for the dataset and model. These new challenges could initiate revision of either business needs and actions, or the model and data, or both.
CRISP-DM is a highly flexible and cyclical model.

Flexibility is required at each step along with communication to keep the project on track.

At any of the six stages, it may be necessary to revisit an earlier stage and make changes.

The key point of this process is that it’s cyclical; therefore, even at the finish you are having another business understanding encounter to discuss the viability after deployment. The journey continues.

# Methodology Overview

## From Problem to Approach:

1) Business Understanding
2) Analytic Approach

## From Requirements to Collection

3) Data Requirement
4) Data Collection

## From Understanding to Preparation

5) Data Understanding
6) Data Preparation

## From Modelling to Evaluation

7) Modelling
8) Evaluation

## From Deployment to Feedback

9) Deployment
10) Feedback

# From Problem to Approach

## 1) Business Understanding

Has this ever happened to you? You've been called into a meeting by your boss, who makes you aware of an important task one with a very tight deadline that absolutely has to be met.

You both go back and forth to ensure that all aspects of the task have been considered and the meeting ends with both of you confident that things are on track.

Later that afternoon, however, after you've spent some time examining the various issues at play, you realize that you need to ask several additional questions in order to truly accomplish the task.

Unfortunately, the boss won't be available again until tomorrow morning.

Now, with the tight deadline still ringing in your ears, you start feeling a sense of uneasiness.

So, what do you do? Do you risk moving forward or do you stop and seek clarification.

Data science methodology begins with spending the time to seek clarification, to attain what can be referred to as a business understanding.

Having this understanding is placed at the beginning of the methodology because getting clarity around the problem to be solved, allows you to determine which data will be used to answer the core question.

Rollins suggests that having a clearly defined question is vital because it ultimately directs the analytic approach that will be needed to address the question.

All too often, much effort is put into answering what people **THINK** is the question, and while the methods used to address that question might be sound, they don't help to solve the actual problem.

Establishing a clearly defined question starts with understanding the **GOAL** of the person who is asking the question.

**Example:**

For example, if a business owner asks: "How can we reduce the costs of performing an activity?" We need to understand, is the goal to improve the efficiency of the activity? Or is it to increase the businesses profitability? Once the goal is clarified, the next piece of the puzzle is to figure out the objectives that are in support of the goal. By breaking down the objectives, structured discussions can take place where priorities can be identified in a way that can lead to organizing and planning on how to tackle the problem. Depending on the problem, different stakeholders will need to be engaged in the discussion to help determine requirements and clarify questions.

## 2. Analytic Approach

Selecting the right analytic approach depends on the question being asked. The approach involves seeking clarification from the person who is asking the question, so as to be able to pick the most appropriate path or approach.

Once the problem to be addressed is defined, the appropriate analytic approach for the problem is selected in the context of the business requirements. This is the second stage of the data science methodology.

Once a strong understanding of the question is established, the analytic approach can be selected. This means identifying what type of patterns will be needed to address the question most effectively.

➡ If the question is to determine probabilities of an action, then a predictive model might be used.

➡ If the question is to show relationships, a descriptive approach maybe be required. This would be one that would look at clusters of similar activities based on events and preferences.

➡ Statistical analysis applies to problems that require counts. For example if the question requires a yes/ no answer, then a classification approach to predicting a response would be suitable.

➡ Machine Learning is a field of study that gives computers the ability to learn without being explicitly programmed. Machine Learning can be used to identify relationships and trends in data that might otherwise not be accessible or identified. In the case where the question is to learn about human behaviour, then an appropriate response would be to use Clustering Association approaches. So now, let's look at the case study related to applying Analytic Approach.

## Summary of From Problem to Approach:

- ➡ The need to understand and prioritize the business goal.
- ➡ The way stakeholder support influences a project.
- ➡ The importance of selecting the right model.
- ➡ When to use a predictive, descriptive, or classification model.

# From Requirements to Collection

## 3) Data Requirement

If your goal is to make a spaghetti dinner but you don't have the right ingredients to make this dish, then your success will be compromised.

Think of this section of the data science methodology as cooking with data. Each step is critical in making the meal.

So, if the problem that needs to be resolved is the recipe, so to speak, and data is an ingredient, then the data scientist needs to identify: which ingredients are required? how to source or the collect them? how to understand or work with them? and how to prepare the data to meet the desired outcome?

Building on the understanding of the problem at hand, and then using the analytical approach selected, the Data Scientist is ready to get started.

This includes identifying the necessary data content, formats and sources for initial data collection.

## 4) Data Collection

After the initial data collection is performed, an assessment by the data scientist takes place to determine whether or not they have what they need.

As is the case when shopping for ingredients to make a meal, some ingredients might be out of season and more difficult to obtain or cost more than initially thought.

In this phase the data requirements are revised and decisions are made as to whether or not the collection requires more or less data.

Once the data ingredients are collected, then in the data collection stage, the data scientist will have a good understanding of what they will be working with.

Techniques such as descriptive statistics and visualization can be applied to the data set, to assess the content, quality, and initial insights about the data.

Gaps in data will be identified and plans to either fill or make substitutions will have to be made.

In essence, the ingredients are now sitting on the cutting board.

It is alright to defer decisions about unavailable data, and attempt to acquire it at a later stage.

## Summary of From Requirements to Collection:

➡ The significance of defining the data requirements for your model.

➡ Why the content, format, and representation of your data matter.

- ➡ The importance of identifying the correct sources of data for your project.

- ➡ How to handle unavailable and redundant data.

- ➡ To anticipate the needs of future stages in the process.

# From Understanding to Preparation

## 5) Data Understanding

Data understanding encompasses all activities related to constructing the data set.

Essentially, the data understanding section of the data science methodology answers the question: Is the data that you collected representative of the problem to be solved?

## 6) Data Preparation

The data preparation stage of the methodology answers the question: What are the ways in which data is prepared?

In a sense, data preparation is similar to washing freshly picked vegetables in so far as unwanted elements, such as dirt or imperfections, are removed.

Together with data collection and data understanding, data preparation is the most time-consuming phase of a data science project, typically taking seventy percent and even up to even ninety percent of the overall project time.

Automating some of the data collection and preparation processes in the database, can reduce this time to as little as 50 percent. This time savings translates into increased time for data scientists to focus on creating models.

To continue with our cooking metaphor, we know that the process of chopping onions to a finer state will allow for its flavours to spread through a sauce more easily than that would be the case if we were to drop the whole onion into the sauce pot. Similarly, transforming data in the data preparation phase is the process of getting the data into a state where it may be easier to work with.

Specifically, the data preparation stage of the methodology answers the question: What are the ways in which data is prepared? To work effectively with the data, it must be prepared in a way that addresses missing or invalid values and removes duplicates, toward ensuring that everything is properly formatted.

## Feature Engineering:

➡ Feature engineering is also part of data preparation.

➡ It is the process of using domain knowledge of the data to create features that make the machine learning algorithms work.

➡ A feature is a characteristic that might help when solving a problem.

➡ Features within the data are important to predictive models and will influence the results you want to achieve.

➡ Feature engineering is critical when machine learning tools are being applied to analyze the data.

When working with text, text analysis steps for coding the data are required to be able to manipulate the data. The data scientist needs to know what they're looking for within their dataset to address the question. The text analysis is critical to ensure that the proper groupings are set, and that the programming is not overlooking what is hidden within.

The data preparation phase sets the stage for the next steps in addressing the question. While this phase may take a while to do, if done right the results will support the project. If this is skipped over, then the outcome will not be up to par and may have you back at the drawing board. It is vital to take your time in this area, and use the tools available to automate common steps to accelerate data preparation. Make sure to pay attention to the detail in this area. After all, it takes just one bad ingredient to ruin a fine meal.

## Summary of From Understanding to Preparation:

➡ The importance of descriptive statistics.

➡ How to manage missing, invalid, or misleading data.

➡ The need to clean data and sometimes transform it.

➡ The consequences of bad data for the model.

➡ Data understanding is iterative; you learn more about your data the more you study it.

# From Modelling to Evaluation

## 7) Modelling

Modelling is the stage in the data science methodology where the data scientist has the chance to sample the sauce and determine if it's bang on or in need of more seasoning!

Data Modelling focuses on developing models that are either descriptive or predictive.

An example of a descriptive model might examine things like: if a person did this, then they're likely to prefer that.

A predictive model tries to yield yes/no, or stop/go type outcomes. These models are based on the analytic approach that was taken, either statistically driven or machine learning driven.

The data scientist will use a **training set** for predictive modelling.

A training set is a set of historical data in which the outcomes are already known. The training set acts like a gauge to determine if the model needs to be calibrated.

In this stage, the data scientist will play around with different algorithms to ensure that the variables in play are actually required.

The success of data compilation, preparation and modelling, depends on the understanding of the problem at hand, and the appropriate analytical approach being taken.

The data supports the answering of the question, and like the quality of the ingredients in cooking, sets the stage for the outcome.

Constant refinement, adjustments and tweaking are necessary within each step to ensure the outcome is one that is solid.

In John Rollin’s descriptive Data Science Methodology, the framework is geared to do 3 things:

1) Understand the question at hand.
2) Select an analytic approach or method to solve the problem,
3) Obtain, understand, prepare, and model the data.

The end goal is to move the data scientist to a point where a data model can be built to answer the question.

With dinner just about to be served and a hungry guest at the table, the key question is: Have I made enough to eat? Well, let's hope so. In this stage of the methodology, model evaluation, deployment, and feedback loops ensure that the answer is near and relevant. This relevance is critical to the data science field overall, as it ís a fairly new field of study, and we are interested in the possibilities it has to offer. The more people that benefit from the outcomes of this practice, the further the field will develop.

# 8) Evaluation

A model evaluation goes hand-in-hand with model building as such, the modelling and evaluation stages are done iteratively.

Model evaluation is performed during model development and before the model is deployed.

Evaluation allows the quality of the model to be assessed but it's also an opportunity to see if it meets the initial request.

**Evaluation answers the question:**
Does the model used really answer the initial question or does it need to be adjusted?

**Model evaluation can have two main phases:**

1) The first is the diagnostic measures phase, which is used to ensure the model is working as intended.

➡ If the model is a predictive model, a decision tree can be used to evaluate if the answer the model can output, is aligned to the initial design. It can be used to see where there are areas that require adjustments.

➡ If the model is a descriptive model, one in which relationships are being assessed, then a testing set with known outcomes can be applied, and the model can be refined as needed.

2) The second phase of evaluation that may be used is statistical significance testing. This type of evaluation can be applied to the model to ensure that the data is being properly handled and interpreted within the model. This is designed to avoid unnecessary second guessing when the answer is revealed.

## Summary of From Modelling to Evaluation:

➡ The difference between descriptive and predictive models.

- ➡ The role of training sets and test sets.
- ➡ The importance of asking if the question has been answered.
- ➡ Why diagnostic measures tools are needed.
- ➡ The purpose of statistical significance tests.
- ➡ That modelling and evaluation are iterative processes.

# From Deployment to Feedback

## 9) Deployment

While a data science model will provide an answer, the key to making the answer relevant and useful to address the initial question, involves getting the stakeholders familiar with the tool produced.

In a business scenario, stakeholders have different specialties that will help make this happen, such as the solution owner, marketing, application developers, and IT administration.

Once the model is evaluated and the data scientist is confident it will work, it is deployed and put to the ultimate test. Depending on the purpose of the model, it may be rolled out to a limited group of users or in a test environment, to build up confidence in applying the outcome for use across the board.

## 10) Feedback

Once in play, feedback from the users will help to refine the model and assess it for performance and impact.

The value of the model will be dependent on successfully incorporating feedback and making adjustments for as long as the solution is required.

Throughout the Data Science Methodology, each step sets the stage for the next. Making the methodology cyclical, ensures refinement at each stage in the game. The feedback process is rooted in the notion that, the more you know, the more that you'll want to know. That's the way John Rollins sees it and hopefully you do too.

Once the model is evaluated and the data scientist is confident it'll work, it is deployed and put to the ultimate test: actual, real-time use in the field.

## Summary to From Deployment to Feedback:

➡ The importance of stakeholder input.

➡ To consider the scale of deployment.

➡ The importance of incorporating feedback to refine the model.

➡ The refined model must be redeployed.

➡ This process should be repeated as often as necessary.

# Summary of Data Science Methodology

We’ve learned how to think like a data scientist, including taking the steps involved in tackling a data science problem and applying them to interesting, real-world examples.

These steps have included: forming a concrete business or research problem, collecting and analyzing data, building a model, and understanding the feedback after model deployment.

You've also learned how to model the data by using the appropriate analytic approach, based on the data requirements and the problem that you were trying to solve Once the approach was selected, you learned the steps involved in evaluating and deploying the model, getting feedback on it, and using that feedback constructively so as to improve the model.

Remember that the stages of this methodology are iterative! This means that the model can always be improved for as long as the solution is needed, regardless of whether the improvements come from constructive feedback, or from examining newly available data sources.

Your success within the data science field depends on your ability to apply the right tools, at the right time, in the right order, to the address the right problem. And that is the way John Rollins sees it.