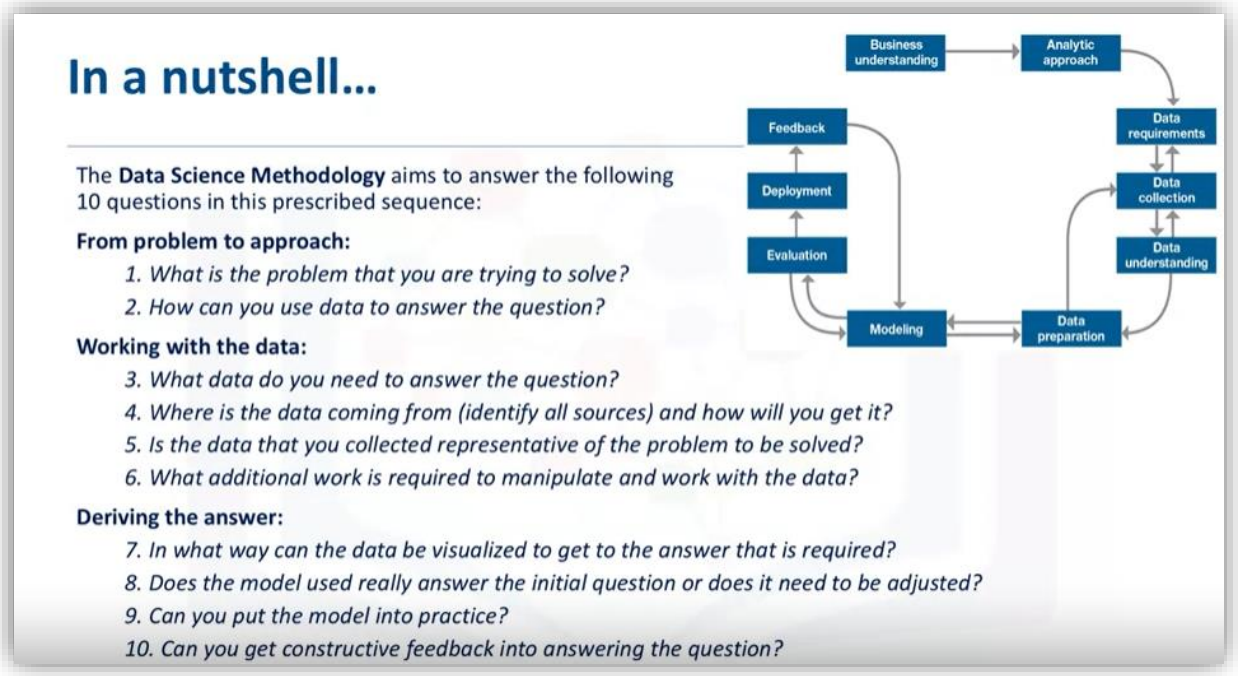
Data Science methodology

A short article from business understanding to solution deployment

Alban STEFF – albansteff@orange.fr – ESILV Paris

How can we divide the **Data Science** process into **small steps** ?



Introduction

The Data Science methodology ensure that we are making the **right decisions** to solve the problem at hand. Jumping too fast to solutions can be a strong mistake, since **Data Science rely a lot on the business understanding,** but also on many **other steps related to the data**, as you can see on the picture above.

Something else is important to see : the **Data Science process is highly iterative.** It means that we need to **go forward** with the process whenever we can, but after getting some feedback or new data or at least a new information, we can also **go backward** and change some parts of the work to get better results.

From problem to approach

The first step of the process is called **business understanding**. During this phase, the stakeholders are **setting the intention of the project**, where it should go. **Communication is the key** at this moment. Each person see the world differently and express ideas based on their objectives. So, to get the best outcome, all aspects of **the problem should be defined**, a perspective should be set and visualize what can be the results after the success of the project can help a lot defining the problem. **It should be written as a question** to make sure at the end of the work that we answered this question.

Then, we have to move to the **analytic approach phase**. Based on the type of question we set in the previous phase, we can pick the right approach :

- **Descriptive :** to show relationships between the variables and understand what happened.
- **Diagnostic (statistical analysis) :** determine the reasons why something happened.
- **Predictive (forecasting) :** determine what will happen next.
- **Prescriptive :** determine how to make something happen.

Each approach can lead to different information in the end, so picking one that is answering the problem is a must. Also, we can mention that for a yes/no answer, a **classification model** is often a good choice to predict the response, like a **decision tree** for instance. And **machine learning** can be used to **identify relationships** in the data that might not be identified without using it.

From data requirements to data collection

After picking the right analytic approach, we have to **define the data requirements** we need to get our result. Data can be identified by its **content, format and sources.** By the way, data can be found by doing **web scraping** when we don't have access to a good database.

Having this all set, **data collection** can start. Techniques should be used to **assess the quality and the content of the data.** For instance, **statistics and visualizations** are great tools to get in touch with the data and get an initial insight about its content. If data is missing, it can be collected later in the process if it's not necessary data. Otherwise, the Data Scientist should proceed with more data collection.

From data understanding to data preparation

Now, you can ask yourself the question : is the data collected **representative of the problem to be solved ?** To answer this question, **descriptive statistics** are a great help. Calculating the **mean, median, min, max and standard deviation** is showing the variables behaviour. Then, **correlation** between variables shows how related they are to each other. Using a **heatmap** is a good way to visualize the **correlation matrix**. Otherwise, a **scatter plot** is useful to see how two variables are related. To check the distribution of the variables, we can plot **histograms.**

Based on these new information, actions should be taken to **prepare the data**. Dealing with **gaps** and **duplicates** can be the first step. Then, we can do more specific actions like **correcting** some values, **standardize** the data format and maybe **merge** data if needed. **Feature engineering** is also part of the preparation process. By creating features within the data, we ensure that the model we create in the next step will have better results toward the initial problem.

From modeling to evaluation

Data modelling focuses on developing a **model** that fits the data. This **model** can be **statistically** driven or **machine learning** driven. We can divide the data in a **training and a test set**. Usually, up to 80% of the total data is used for the training set and up to 20% for the test set. Then, we can use different **algorithms** to check which one works best by testing the model with the test set and compare the prediction and the reality.

After creating the model, we need to **evaluate** it to determine if it needs a **calibration**. For a classification problem, when something is **misclassified**, we can adjust the cost of the **relative weight** of misclassifying each outcome. The default value is set to 1-1 but we can change it to 9-1 or 4-1 and so on to see which model is best. These errors are called **type I error** for a **false-positive** and **type II error** for a **false-negative**. In a yes/no problem, we call the yes accuracy as **sensitivity** and the no accuracy as **specificity**. Using a **ROC curve** is often a good choice to evaluate a classification model. The **optimal model** is at the maximum separation. Evaluating the model can be divided in two parts :

- **Diagnostic measures :** ensure that the model is working as intended.
- **Statistical significance :** ensure that the data is properly interpreted by the model.

A great tool to determine if the model predictions are good is to create a **confusion matrix**, also called **error matrix**. It shows the accuracy of the model by displaying the values predicted for each class.

From deployment to feedback

Once the model is ready to be **implemented**, the key is to get the stakeholders involved by getting them familiar with the tool created by the Data Scientist. Then, depending on the purpose of the model, the solution will be deployed in a **test environment** or in a **limited group** of users.

The last part of the Data Science process is the **feedback**. This is an important step, since it tells you how **effective** is your solution when it is applied to the real world. The plan for the feedback step can be :

- **Review process :** define how to know if the model has an impact on the problem.
- **Tracking :** record the impact of the model on the problem.
- **Measurement :** measure the efficiency of solving the problem at hand.

The real value of the model will depend on incorporating the feedback until the problem is completely solved. As we saw in the data requirements step, we can also tell if we finally need more data to get better results. If everything looks fine, the model is **implemented on a larger scale** and its effects will be reviewed one year later. Based on this year extra knowledge, the model can be **refined** again if needed.

Feedbacks tell you where to go