What is Data Science ?

A little article about how I see this job

Alban STEFF – albansteff@orange.fr – ESILV

What is **Data Science** ?

Data Science is a process to understand the world using data. It's an amazing process to deliver insights and to uncover a **new path** that was hidden between the million lines of data that you collected.

A **Data Scientist** is providing **answers** to unsolved problems and should present it in a compelling way. **Data Visualization** is a tool to make data express its secrets in a human understandable language and scale.

The best noun to define a Data Scientist is "**storyteller**". Why ? When you want to explain what patterns or what information you found in your data, you have to **tell a story**. It makes your presentation **convincing** and easy to follow for everyone.

The key qualities of a **Data Scientist** are **curiosity** and a strong **argumentation ability**. I believe that curiosity is essential because it tells you where to start whith your data and allows you to test ideas that may seem to be unrelevant hypothesis. But only **Data Analysis** can tell. When you find interesting patterns in your data, communicating them is your role. A **Data Scientist** is not just doing analysis : he's a real **presenter**. Arguing on what you found is the best way to convince your audience that your findings are relevant. They should understand how you think and how you discovered these patterns. Yes, a **Data Scientist** is a good **storyteller** !

As a **Data Scientist**, your role is to make **recommendations** for the project stakeholders, based on the field of application of the project. Having great knowledge about the project subject is a must since your mission is to **generate insights** from the data you have. Not being able to understand what is the project may lead your **Data Analysis** on a wrong path and reveal useless patterns.

Talking about **Data Analysis**, a **Data Scientist** must also have good **computational skills**. Working with computers is necessary to do strong **Data Analysis** with other useful tools like **statistics, programming, databases, cloud, artifical intelligence, machine learning, deep learning** and so on. Data can be structured or unstructured and you have to know how to deal with both. You can find **gaps**, missing values or systematically missing values. Making relevant decisions by doing some **Data Cleaning** on your **Collected Data** can help your findings become closer to the reality.

A **Data Scientist** can work with data but also with **Big Data** and classic computational power is not convenient to deal with this amount of data. Luckily, **Cloud Platforms** allows to deploy a **Database** in a **Cluster** on high performance machines that are not yours. That's amazing when you work with **Big Data**. It's also making the **teamwork** more efficient than ever by giving access to the same data to your coworkers at the same time.

The skillset of a **Data Scientist** can be used for many **real-world purposes** such as healthcare systems, fraud detection, natural disasters predictions, recommendation systems and so on. It's up to each **Data Scientist** to focus on a field he cares about, because that's through the motivation to find solutions in a specific domain that he will really give useful information to the relevant stakeholders.

What are the **ten main components** of a **report** that would be delivered at the end of a **Data Science** project ?

Since a **Data Scientist** has to **share** his results to other workers or people, writing a **good report** is important to make sure your ideas are **understandable**. To do so, you can focus on these components :

- **Informative cover page :** title, authors (name, affiliation, contact), publisher, date of publication and why not a good design in the background.
- **Table of contents :** always use a table of contents as a map for your readers to see where they are going. It gives an idea of what is in the document.
- **Executive summary :** there is nothing stronger than a little summary of your ideas and arguments for the reader who wants to get into the subject faster.
- **Introductory :** set up the problem and explain the subject.
- **Literature review :** sometimes, the subject of the report can be contested. Using online references is useful to show what is other people’s opinion and which knowledge might be missing.
- **Methodology :** here you have to explain your research methods and where you found the data you used for analysis.
- **Results :** present your findings (descriptive statistics, visualization, hypothesis test, regression models or categorical analysis, data mining results). If your readers are not from a scientifical background, using mostly visualizations to summarize your results can be better.
- **Discussion :** explain your main arguments and communicate your main idea, that will be supported by the presentation you just did.
- **Conclusion :** promote your findings and identify future applications of them.
- **Acknowledgements, references, appendices**