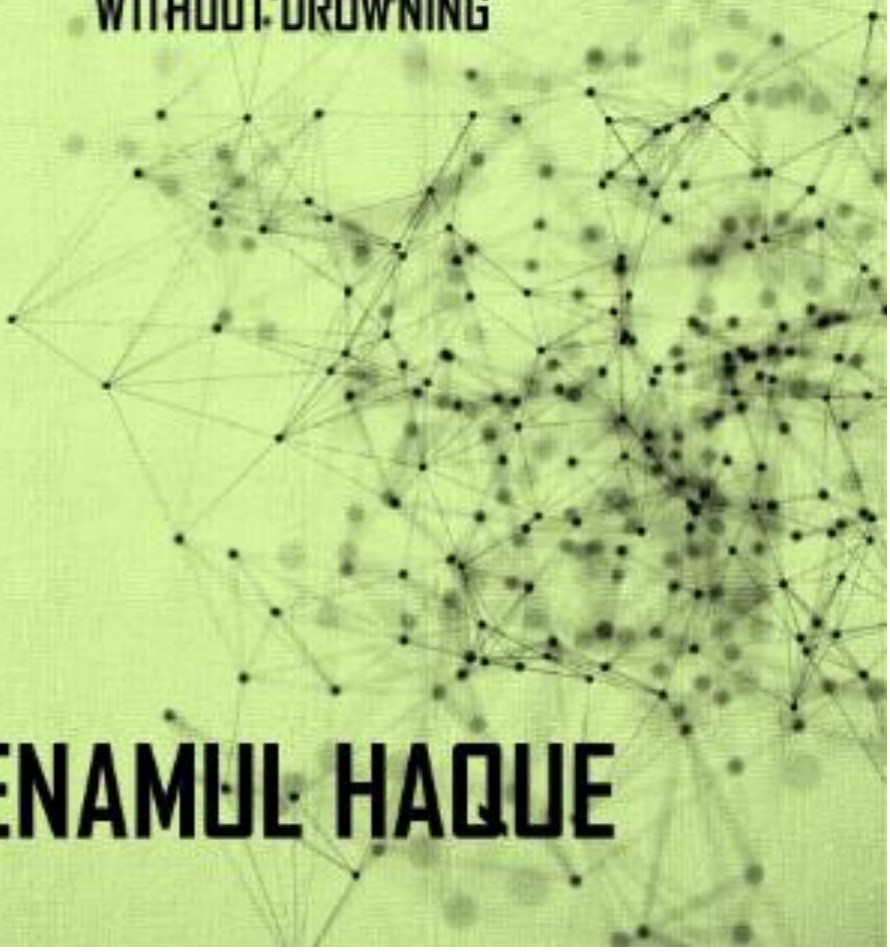
A BEGINNER'S GUIDE TO DATA SCIENCE

HOW TO DIVE INTO THE DATA OCEAN WITHOUT DROWNING



ENAMUL HAQUE

A Beginners Guide To DATA SCIENCE

How to dive into the data ocean without drowning

ENAMUL HAQUE

Enel Publications
London, UK
Amazon Kindle Direct Publishing

First Printing Edition, April 2021 (Revision 1)

ISBN 9798731261074

CHAPTER TWO:
THERE IS NO BUSINESS LIKE DATA BUSINESS

"For me, the thing about data science that makes it so exciting to the modern world is its unparalleled ubiquity — data science is everywhere. It is ultimately just a set of skills derived from computer science and mathematics, and this set of skills can be universally applied to learn from the past and improve future performances in any discipline you can think of. That's what makes data science so relevant: its enormous scope and potential to improve life across a wide variety of sectors. I'm excited to think about a future where data-driven decisions become more and more commonplace all around the world." - Vishnu Subramanian, Founder @ Jarvislabs.ai 1-click GPU cloud platform

Data Science Business Strategy

It's essential to precisely understand how to manage your data life cycle and maintain the corporate data model. You need to understand what data the company has to develop the right strategy for working with them. Then it will be valuable for business. In parallel with the first pilots, a roadmap for creating a platform for data science, including developing the storage platform and approaches to working with machine learning models, is being developed and refined.

First pilot projects

At this stage, pilot models of machine learning are being built, including recommendation and evaluation. In the pilot's implementation, recommendations are made on whether to use machine learning models for this task and possible ways to improve the proposed models' quality. They are assessed in terms of the potential economic impact and complexity of implementation. It also determines how tasks are prioritised. In the development of pilots, the requirements for the future ecosystem of data and models are clarified. Data and IT specialists are immersed in the specifics of production, and the production itself is introduced to new tools.

Introducing data lake as one of the first steps

Cheaper and more functional than traditional lake storages, the lakes allow for rapid data processing through analytics and machine learning. Their concept will enable you to start accumulating data before specific tasks are defined. This, in turn, allows historical data to be used for machine learning models. Deploying a storage layer and downloading an accessible history leads to faster testing and the introduction of new models.

Accumulation of data

It goes either way, even if the data is not used in existing models. It is important to organise storage space and engage in minimal structuring; otherwise, the "lake" will turn into a useless "swamp". In addition, data lake needs to be linked to the company's analytical ecosystem and information security: it should not leak or cause problems with regulators.

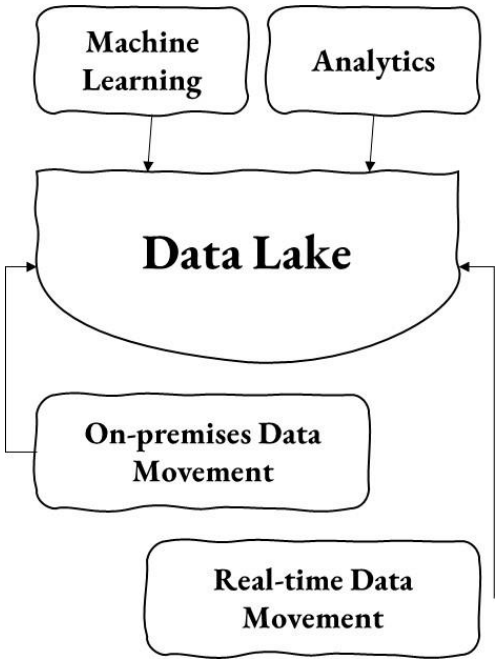


Figure 1 - A data lake is a centralised repository that allows you to store all your structured and unstructured data at any scale.

Create a layer of models and put them into production

You don't have to wait for a long accumulation and time-consuming structuring of data to use the lake. Using data lake and modern virtualisation technologies, we can quickly deploy the layer for models and create them in the target architecture. Over time, technology and data composition change, the quality of the model can fall because it needs to be modified or created a new one. Over time, there may be several models that may be more or less effective in different situations. Therefore, simultaneously with the development of service models, tools are created to manage their life cycle.

Figure 2 - Data modelling creating opportunities

Comprehensive data work, machine learning initiatives, and process digitisation enable any manufacturing company to become more efficient. This provides an opportunity to maximise profits by reducing the cost of production, facilitating and accelerating specialists' work, improving production safety, improving the situation with overspending of raw materials, the percentage of optimisation, and equipment maintenance. And in the long term - to ensure the transition to fully autonomous production.

Approaches and Methods of Social Media Data Analysis

Social media is a good source of data, and it is vital to be able to work effectively with that data. Let's take a look at a few features and approaches to how social media data works. It is worth noting that there is a separate direction - Social Mining. This is applying data mining methods and algorithms to find and detect dependencies and knowledge on social networks (or areas of knowledge where data can be presented as networks/graphs). The applications are pretty comprehensive.

In general, almost all the practical tasks of analysing social media data are reduced to the following basic:

- Analysis of social network information flows, structure and metrics
- Analysis of the tone of messages (emotional colouring)
- Analysis and extraction of topics (as written in social networks)
- Image analysis

There are also combinations of these tasks.

Info stream analysis

This class of methods allows identifying opinion leaders in social networks, managing the media campaign, and evaluating users' attitude to this or that information. The challenges to this are:

- Search for most communication objects.
- Search for objects with the most connections.
- Search for the most "authoritative" objects.
- Search for objects that serve as a bridge between communities.

The most commonly used tool for analysis and visualisation in a given area is a graph where nodes (actors) are people or groups. The edges demonstrate relationships (connections) or information flows between nodes.

One of the most essential tasks in the analysis of social networks is to find "important" (from different points of view) participants of the social graph. To do this, the researchers calculate different types of metrics: Degree centrality (by the number of related nodes; necessary to those with many friends; useful for highlighting opinion leaders), Closeness centrality (by proximity; how close a participant is to everyone else on the network; most often used in the task of finding influence groups and "grey cardinals", betweenness centrality (in between; the number of shortcuts passing through the participant; how often the information passes through that person).

Tonality analysis

This class of methods allows you to evaluate users' attitude to certain information (object, person, event, etc.). The tasks to be solved here: assessing the emotional colouring of messages, highlighting named entities and assessing their emotional colouring.

Theme analysis

This class of methods allows you to identify the most popular topics in the community and most often discussed in it (at a particular time). Solved tasks: highlighting topics (topic modelling), assessing emotional colouring by themes, highlighting entities related to the topic.

Image analysis

It allows you to identify what types of photo-content place different segments of users. Solved tasks: the kind of object in the photo, the type of location in the image, people's emotions, verification and identification (to compare the person found in a physical location with his profile on the social network).

If the task is aimed at the level of analysis of a particular person, that is, such directions:

- Personalisation of proposals
- Analysis of the structure of the social network
- Analysis of human content on the social network

Offering personification allows you to provide the user with the content that is most relevant to them. Tasks: collecting and enriching user information; Clustering and segmenting users User classification based on the built model personalised provision of information.

What does Google know about you? You can find information about yourself here: google.com/settings/ads

Current and promising research in the field of social media analysis

- Semi-supervised learning on social media
- Social media sustainability and design
- Predicting the spread of information on social networks
- A synergy of spatial data and social media data

Data Science Project Management

As the volume of data increases day by day in all areas and industries, it is essential for any company, industry, or domain to know about it and use it appropriately to grow enormously. No business wants to slow down growth, and then they do not know what the root of the problem is and how to solve it and develop it. Often when we talk about data science projects, it seems that no one can provide a clear explanation of how the whole process is going. From data collection to analysis and presentation of results. In the previous section, we saw the data science lifecycle, and now we will apply them in the data science project

Problem statement

There are two ways in the problem statement-based data science approach: dive into the problem and solve. First, you need to know if your goal in this data is a numerical or categorical decision. For example, your problem statement is whether a drug has shown the desired results or not, whether customers are satisfied with a new product released, or whether sales will rise or fall in the future. This is a definite answer, i.e., simply yes or no, possibly or not. If your job is to predict the future sales price or home prices, or what dosage is required. They all give numerical values based on the data provided. So, first, you need to identify the problem and find the best solution for it.

Understanding data or business

The problem arises in different areas or areas, and understanding the terminology and having experience in the area of understanding helps us find a better solution. In this way, we learn many other suggestions based only on business understanding or business knowledge in this region.

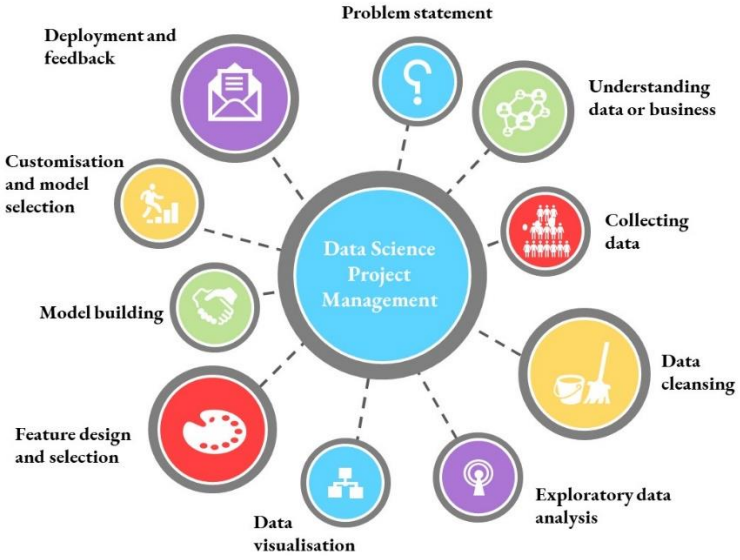


Figure 3 - Data science project deliverables

Collecting data

Now the data processing starts, and the data is collected from various sources and placed in a specific location (database). All data required to solve this problem is collected.

Data cleansing

The collected data is correctly installed and checked for any missing data, anomalies and data distribution. The data is cleared and processed with all the payload.

Exploratory data analysis

As all data is cleared and the necessary part is removed, leaving unnecessary things. The data is now analysed and studied along with all statistics.

Data visualisation

Since most of the collected data is now cleaned up, explored and well understood and presented visually with some graphs, graphs using the Scikit-learn library in Python or visualisation can be created in Tableau and in some visualisation software or in something else. In this way, ideas are well extracted with perfect images that anyone can see, which can be well explained.

Feature design and selection

Implement some statistical or dimensionality reduction techniques or some other techniques, as appropriate, to add some useful columns from existing or new columns and provide only the data you need here and not any others. Otherwise, there is a possibility of misinterpretation.

Model building

During the model building phase, the data is split into two parts, one of which is used for training and the other for validation, because if you use the same data, there is a chance that the machine will be overfitted (instead of just study the data, ideally studying the subject or data theory). Machine learning comes in different types and is used differently, depending on the data and requirements. Types: supervised, unsupervised, and reinforcement learning. So, the required models are implemented, and the best model is selected.

Customisation and model selection

We do not know which model is suitable and the right one to choose from. So, after building the model, they are evaluated and additionally adjusted with some other parameters, and then a model is selected that has proven itself well.

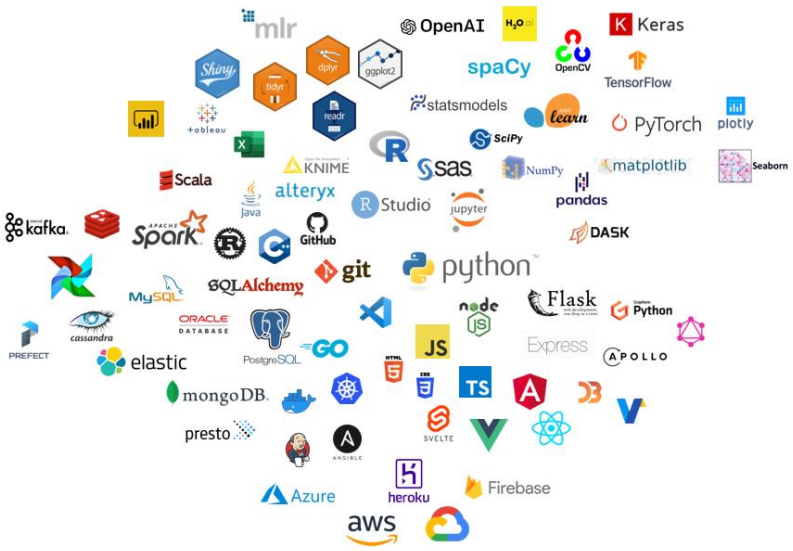


Figure 4 - Data science technology stack

Deployment and feedback

The required machine learning algorithm has been selected and is now deployed. This can be done in many different ways, and there are many tools for that, like Flask, AWS, Google Cloud, Django etc. Once deployed, it is used by a company or customers and feedback is collected; if it works well, the problem is resolved; otherwise, it will be retrieved by the data analysis team again for further improvements, so this will be done by rechecking.