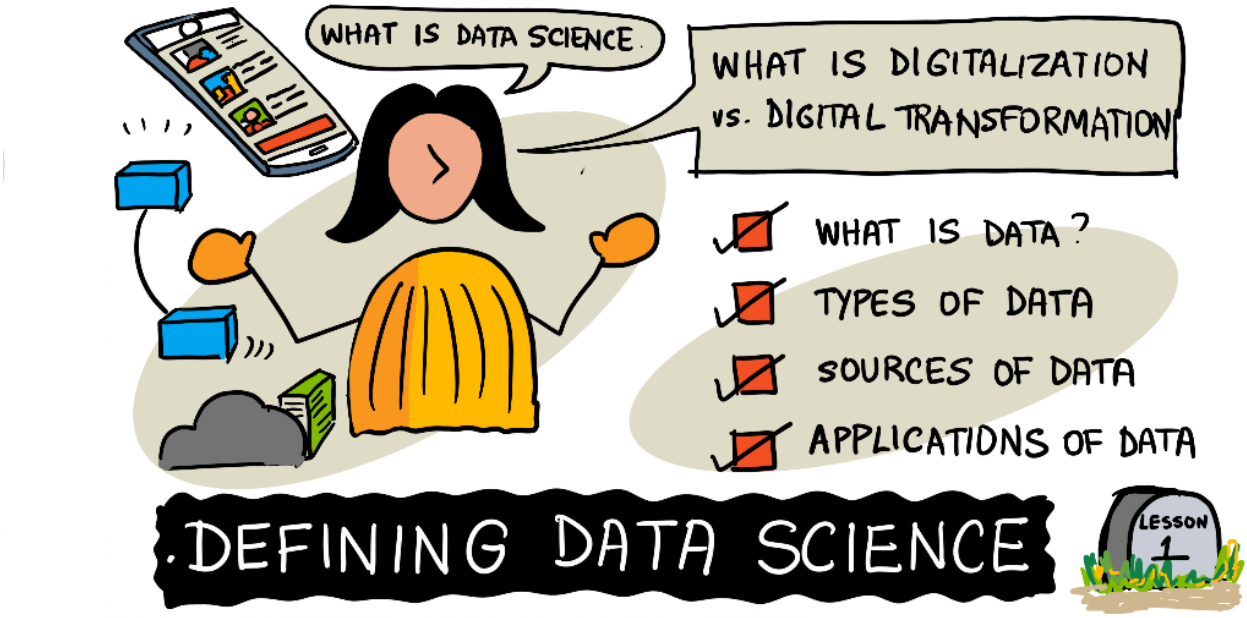
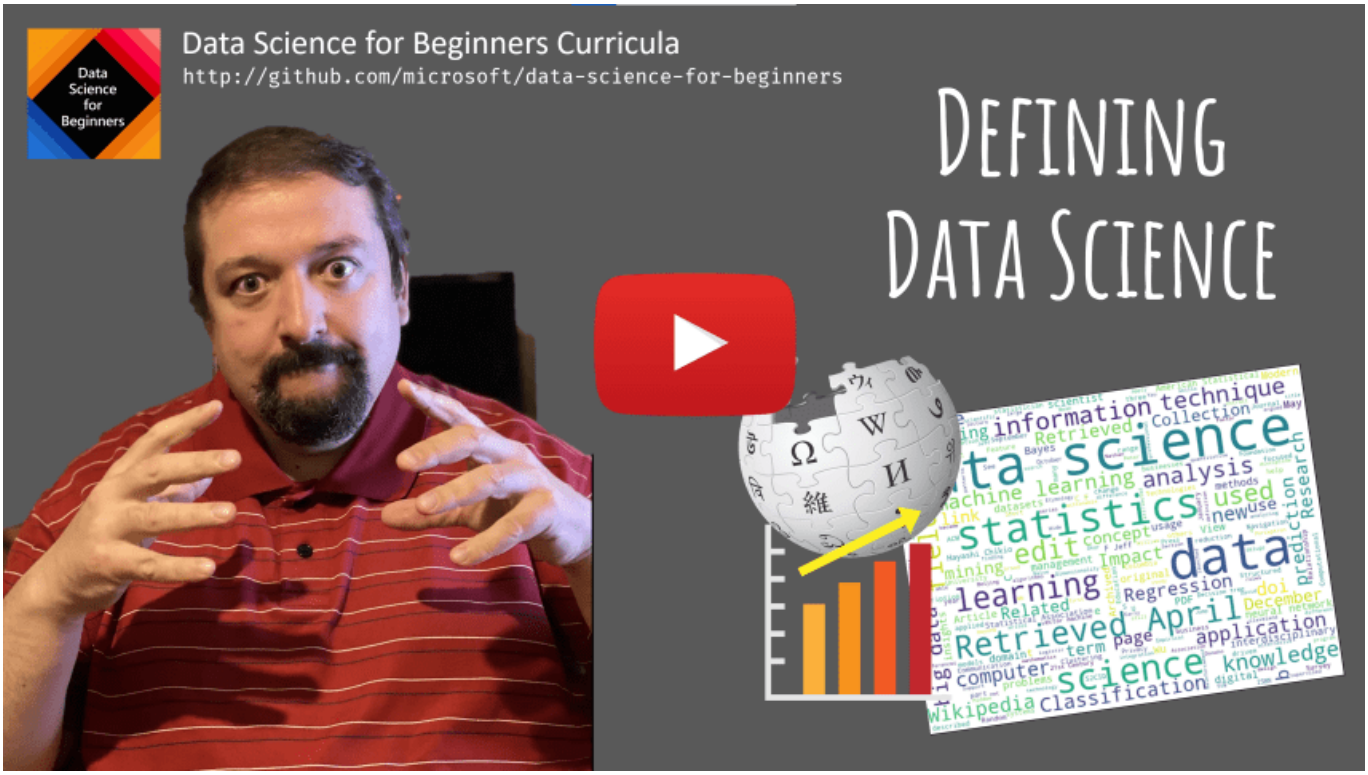
Defining Data Science



Defining Data Science - *Sketchnote by @nitya*



Pre-lecture quiz

What is Data?

In our everyday life, we are constantly surrounded by data. The text you are reading now is data. The list of phone numbers of your friends in your smartphone is data, as well as the current time displayed on your

watch. As human beings, we naturally operate with data by counting the money we have or by writing letters to our friends.

However, data became much more critical with the creation of computers. The primary role of computers is to perform computations, but they need data to operate on. Thus, we need to understand how computers store and process data.

With the emergence of the Internet, the role of computers as data handling devices increased. If you think about it, we now use computers more and more for data processing and communication, rather than actual computations. When we write an e-mail to a friend or search for some information on the Internet - we are essentially creating, storing, transmitting, and manipulating data.

> Can you remember the last time you have used computers to actually compute something?

What is Data Science?

In Wikipedia, **Data Science** is defined as *a scientific field that uses scientific methods to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains*.

This definition highlights the following important aspects of data science:

- The main goal of data science is to **extract knowledge** from data, in other words - to **understand** data, find some hidden relationships and build a **model**.
- Data science uses **scientific methods**, such as probability and statistics. In fact, when the term *data science* was first introduced, some people argued that data science was just a new fancy name for statistics. Nowadays it has become evident that the field is much broader.
- Obtained knowledge should be applied to produce some **actionable insights**, i.e. practical insights that you can apply to real business situations.
- We should be able to operate on both **structured** and **unstructured** data. We will come back to discuss different types of data later in the course.
- **Application domain** is an important concept, and data scientists often need at least some degree of expertise in the problem domain, for example: finance, medicine, marketing, etc.

> Another important aspect of Data Science is that it studies how data can be gathered, stored and operated upon using computers. While statistics gives us mathematical foundations, data science applies mathematical concepts to actually draw insights from data.

One of the ways (attributed to Jim Gray) to look at the data science is to consider it to be a separate paradigm of science:

- **Empirical**, in which we rely mostly on observations and results of experiments
- **Theoretical**, where new concepts emerge from existing scientific knowledge
- **Computational**, where we discover new principles based on some computational experiments
- **Data-Driven**, based on discovering relationships and patterns in the data

Other Related Fields

Since data is pervasive, data science itself is also a broad field, touching many other disciplines.

Databases
: A critical consideration is **how to store** the data, i.e. how to structure it in a way that allows faster processing. There are different types of databases that store structured and unstructured data, which we will consider in our course.

Big Data
: Often we need to store and process very large quantities of data with a relatively simple structure. There are special approaches and tools to store that data in a distributed manner on a computer cluster, and process it efficiently.

Machine Learning
: One way to understand data is to **build a model** that will be able to predict a desired outcome. Developing models from data is called **machine learning**. You may want to have a look at our Machine Learning for Beginners Curriculum to learn more about it.

Artificial Intelligence
: An area of machine learning known as artificial intelligence (AI) also relies on data, and it involves building high complexity models that mimic human thought processes. AI methods often allow us to turn unstructured data (e.g. natural language) into structured insights.

Visualization
: Vast amounts of data are incomprehensible for a human being, but once we create useful visualizations using that data, we can make more sense of the data, and draw some conclusions. Thus, it is important to know many ways to visualize information - something that we will cover in Section 3 of our course. Related fields also include **Infographics**, and **Human-Computer Interaction** in general.

Types of Data

As we have already mentioned, data is everywhere. We just need to capture it in the right way! It is useful to distinguish between **structured** and **unstructured** data. The former is typically represented in some well-structured form, often as a table or number of tables, while the latter is just a collection of files. Sometimes we can also talk about **semi-structured** data, that have some sort of a structure that may vary greatly.

Structured	Semi-structured	Unstructured
List of people with their phone numbers	Wikipedia pages with links	Text of Encyclopedia Britannica
Temperature in all rooms of a building at every minute for the last 20 years	Collection of scientific papers in JSON format with authors, data of publication, and abstract	File share with corporate documents
Data for age and gender of all people entering the building	Internet pages	Raw video feed from surveillance camera

Where to get Data

There are many possible sources of data, and it will be impossible to list all of them! However, let's mention some of the typical places where you can get data:

- **Structured**

    - **Internet of Things** (IoT), including data from different sensors, such as temperature or pressure sensors, provides a lot of useful data. For example, if an office building is equipped with IoT sensors, we can automatically control heating and lighting in order to minimize costs.
    - **Surveys** that we ask users to complete after a purchase, or after visiting a web site.
    - **Analysis of behavior** can, for example, help us understand how deeply a user goes into a site, and what is the typical reason for leaving the site.
- **Unstructured**
    - **Texts** can be a rich source of insights, such as an overall **sentiment score**, or extracting keywords and semantic meaning.
    - **Images** or **Video**. A video from a surveillance camera can be used to estimate traffic on the road, and inform people about potential traffic jams.
    - Web server **Logs** can be used to understand which pages of our site are most often visited, and for how long.
- Semi-structured
    - **Social Network** graphs can be great sources of data about user personalities and potential effectiveness in spreading information around.
    - When we have a bunch of photographs from a party, we can try to extract **Group Dynamics** data by building a graph of people taking pictures with each other.

By knowing different possible sources of data, you can try to think about different scenarios where data science techniques can be applied to know the situation better, and to improve business processes.

What you can do with Data

In Data Science, we focus on the following steps of data journey:

1) Data Acquisition

    The first step is to collect the data. While in many cases it can be a straightforward process, like data coming to a database from a web application, sometimes we need to use special techniques. For example, data from IoT sensors can be overwhelming, and it is a good practice to use buffering endpoints such as IoT Hub to collect all the data before further processing.

2) Data Storage

    Storing data can be challenging, especially if we are talking about big data. When deciding how to store data, it makes sense to anticipate the way you would like to query the data in the future. There are several ways data can be stored:

    - A relational database stores a collection of tables, and uses a special language called SQL to query them. Typically, tables are organized into different groups called schemas. In many cases we need to convert the data from original form to fit the schema.
    - A NoSQL database, such as CosmosDB, does not enforce schemas on data, and allows storing more complex data, for example, hierarchical JSON documents or graphs. However, NoSQL databases do not have the rich querying capabilities of SQL, and cannot enforce referential integrity, i.e. rules on how the data is structured in tables and governing the relationships between tables.
    - Data Lake storage is used for large collections of data in raw, unstructured form. Data lakes are often used with big data, where all data cannot fit on one machine, and has to be stored and processed by a cluster of servers. Parquet is the data format that is often used in conjunction with big data.

3) Data Processing

    This is the most exciting part of the data journey, which involves converting the data from its original form into a form that can be used for visualization/model training. When dealing with unstructured data such as text or images, we may need to use some AI techniques to extract **features** from the data, thus converting it to structured form.

4) Visualization / Human Insights

    Oftentimes, in order to understand the data, we need to visualize it. Having many different visualization techniques in our toolbox, we can find the right view to make an insight. Often, a data scientist needs to "play with data", visualizing it many times and looking for some relationships. Also, we may use statistical techniques to test a hypotheses or prove a correlation between different pieces of data.

5) Training a predictive model

    Because the ultimate goal of data science is to be able to make decisions based on data, we may want to use the techniques of Machine Learning to build a predictive model. We can then use this to make predictions using new data sets with similar structures.

Of course, depending on the actual data, some steps might be missing (e.g., when we already have the data in the database, or when we do not need model training), or some steps might be repeated several times (such as data processing).

Digitalization and Digital Transformation

In the last decade, many businesses started to understand the importance of data when making business decisions. To apply data science principles to running a business, one first needs to collect some data, i.e. translate business processes into digital form. This is known as **digitalization**. Applying data science techniques to this data to guide decisions can lead to significant increases in productivity (or even business pivot), called **digital transformation**.

Let's consider an example. Suppose we have a data science course (like this one) which we deliver online to students, and we want to use data science to improve it. How can we do it?

We can start by asking "What can be digitized?" The simplest way would be to measure the time it takes each student to complete each module, and to measure the obtained knowledge by giving a multiple-choice test at the end of each module. By averaging time-to-complete across all students, we can find out which modules cause the most difficulties for students, and work on simplifying them.

> You may argue that this approach is not ideal, because modules can be of different lengths. It is probably more fair to divide the time by the length of the module (in number of characters), and compare those values instead.

When we start analyzing results of multiple-choice tests, we can try to determine which concepts that students have difficulty understanding, and use that information to improve the content. To do that, we need to design tests in such a way that each question maps to a certain concept or chunk of knowledge.

If we want to get even more complicated, we can plot the time taken for each module against the age category of students. We might find out that for some age categories it takes an inappropriately long time to complete the module, or that students drop out before completing it. This can help us provide age recommendations for the module, and minimize people's dissatisfaction from wrong expectations.

🚀 Challenge

In this challenge, we will try to find concepts relevant to the field of Data Science by looking at texts. We will take a Wikipedia article on Data Science, download and process the text, and then build a word cloud like this one:

Visit `notebook.ipynb` to read through the code. You can also run the code, and see how it performs all data transformations in real time.

> If you do not know how to run code in a Jupyter Notebook, have a look at this article.

Post-lecture quiz

Assignments

- **Task 1**: Modify the code above to find out related concepts for the fields of **Big Data** and **Machine Learning**
- **Task 2**: Think About Data Science Scenarios

Credits

This lesson has been authored with ♥ by Dmitry Soshnikov