What is Data Science

Definitions

From Wikipedia[1]:

> **Data Science** is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured.

Problem. The use of terms **data science** and **data scientist** over a short period of time reached the levels where these terms are perceived as no more than just **buzzwords**. This makes explaining the precise nature of these terms difficult.

Popular theories. A lot of people postulate that

Data Science = `<Insert your term here>`

or

Data Scientist = `<Insert your job title here>`

The terms used to equate to **data science** often are:

- Statistics
- Machine learning
- Business analytics
- Knowledge discovery in data

[1] https://en.wikipedia.org/wiki/Data_science

Similarly, the following professions are mentioned as being equivalent to that of a **data scientist**:

- Statistician
- Data analyst
- Business analyst
- Database analyst

This leads to snide remarks and jokes that a data scientist is a statistician in San Francisco, or data science is statistics on a Mac.

A somewhat more insighful notion of data science is that it lies in the intersection of three fields:

- Statistics
- Computer Science
- Domain expertise

From Statistics, data science borrows hypothesis testing and data analytical techniques. From computer science data science borrows approaches to data manipulation and machine learning algorithms, while domain knowledge directs the combined machinery of statistics and computer science to proper use.

- This understanding of **data science** is colored by the fact that first data scientists were neither computer scientists, nor statisticians.
- Rather, they were pure scientists who had to learn statistics and computer science in order to be able to deal with data analytical tasks they have encountered in their studies.
- This set of skills wound up being extremely useful outside of scientific domain.
- What these people were doing for the companies that hired them was different enough from either pure software development or pure data analysis, that terms "software engineer", "data analyst", "statistician" were not applicable to their job description.
- Out of necessity, the term **data scientist** was re-purposed.

Data Science as a set of skills. Another way to define data science is to outline a set of skills that a pratitioner in the field (i.e., a data scientist) is expected to possess. This approach has its advantages and drawbacks.

Among the **advantages** are:

- **Clear picture.** Skills are easy to evaluate - one either has them or one does not.
- **Reduction from job descriptions.** One can study job descriptions for **data scientist** positions and reverse-engineer the necessary skills[2].

The **disadvantages** of this approach, however, are significant:

- **Such definitions are not stable.** Skills, especially when they are ground in specific technologies, are very present-day. Tomorrow's set of skills necessary for a data scientist may be very different than today's. This means that an effective definition of data science will have to shift with the times. *This is not very convenient.*
- **Such definitions ignore the "science" part of data science.** We do not define what carpentry or medicine is by the sets of tools carpenters or medics have to work with. We define different professions and disciplines by the nature of the work and/or the nature of the product produced. In case of data science a skills-based definition lacks the ability to explain *what these skills are used for.*

As a result:

- Skills-based definitions of data science are very useful in determining the specific portions of the curriculum (what skills to teach, what technologies to present in class).
- But such definitions cannot be all-encompassing.

Data Science as a process. Our final approach to describing what **data science** is, is to observe that productive work with data is a process.

When comparing **what** data scientists do on a day-to-day basis, with what statisticians or computer scientists are taught to do, we can notice significant differences. This leads to an understanding that

> **Data Science** can be thought of as the discipline that studies and the full cycle of work with data and incorporates all stages this work from data acquistion, through data analysis and all the way to presentation of the obtain insight.

Data Science Process

In this class, we will use the term data science process a lot, and we will spend considerable time talking about the specific steps of this process. In a nutshell, the data science process consists of the following steps:

[2] In fact, this is, in large part, how we have built the curriculum for the Data Science Minor at Cal Poly.

1. Formulation of questions.
2. Data acquistion.
3. Data cleaning/pre-processing.
4. Data modelling.
5. Data analysis.
6. Visualisation of results.
7. Presentation of insight/results.

Not every step is present in every specific instance of a data science process, but, each of these steps is present (and is crucial) to some instances.

Formulation of questions. On this step, data scientists determine what needs to be studied. Often this step is split into two stages: one that precedes the data acquisition step, and presents the challenges/questions in a general form, and one that succeeds the data acquistion and data cleaning stages, and makes the questions precise.

Data acquistion. On this step, we go from not having data to having data. Data acquistion is the process of collecting the data that is needed to answer the questions posed.

Data cleaning. Not all collected data is needed to answer the questions. Not all collected data contains information that is useable. Collected data may contain duplicate information. The data cleaning step of the process discovers imperfections in the data, and when possible eliminates/fixes them.

Data modelling. The shape in which the data is acquired may not necessarily be the shape in which the data is easiest to analyze. On the data modelling stage, the data is transformed into the "shape" that matches both its underlying nature (semantics) and the needs of the subsequent data analysis.

Data analysis. The heart of the data science process. On this step the actual insight is generated. This is aslo the step that is often thought of as being *the only one that matters.*

Visualization of results. Once the data analysis produces results, the results need to be somehow prepared for presentation. Often, this involves some form of visualization.

Presentation of results. The final step is the delivery of the insight. This step involves explanation of what was observed.