Which Type of Data Scientist are You?

Learn about the three most common types of data scientists and where you fit in.

Matthias Döring

2020-08-02

Contents

Chapter 1

Which Type of Data Scientist are You?

Why Data Science?

Data science is considered one of the hottest professions of the 21st century. There are many reasons why data science is currently so much in the limelight:

- **The data deluge:** According to Statista, merely 2 zettabytes ($2 \cdot 10^{21}$ bytes) of data were generated in 2010, while the estimated volume for 2020 is at a whopping 59 zettabytes. To make sense of these data, we need experts and tools that can transform these raw data into useful pieces of knowledge.
- **The AI revolution:** Recent advancements in artificial intelligence, particularly in deep learning, allow completely new applications. Examples include image classification (Russakovsky et al., 2015), face recognition (Taigman et al., 2014), and even gaming (Silver et al., 2016). For example, automative manufacturers worldwide are currently working hard on making autonomous cars a reality. That would have been unthinkable 20 years ago.
- **The internet of things (IoT)**: Nowadays most people are constantly connected to the internet, usually with multiple devices such as notebooks, mobile phones, or tablets. With the IoT, not only people are connected to the internet but there is also a network been individual devices, for example, in factories or in your own home. For instance, a smart home may use intelligent systems to control the temperature of your room.

What is Data Science About?

It is evident that data is becoming more and more important and that we need experts for analyzing these data. This is where the data science comes in. Data science is an interdisciplinary field that requires expertise in the following three areas:

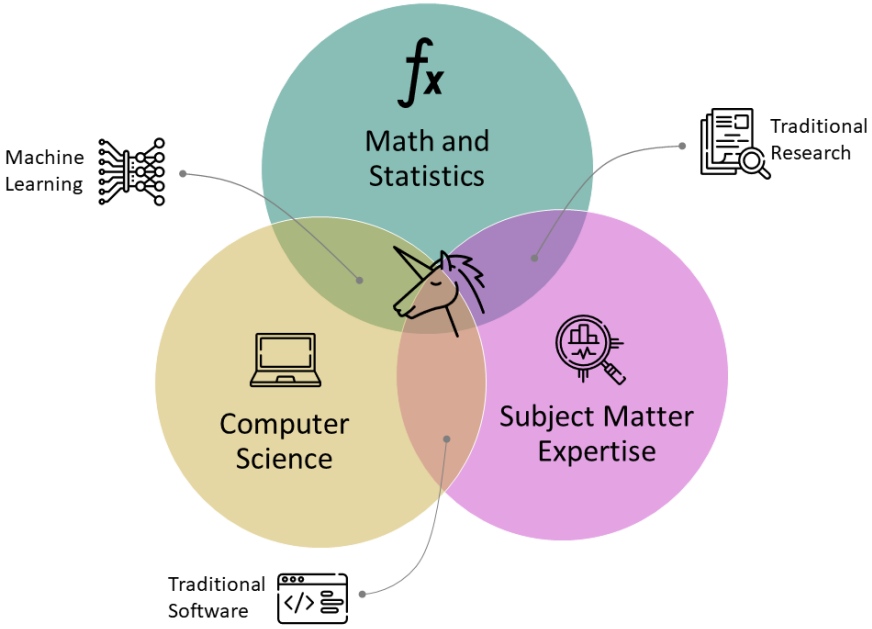

Figure 1.1: The Venn data science diagram. Adapted from the concept of Steven Geringer Raleigh, NC.

- **Math and Statistics:** Math forms the basis of all machine learning (ML) approaches. While linear algebra is particularly important for ML algorithms, statistics plays a more general role for two reasons. First, a good statistical understanding is required for reasoning about data. Second, many machine learning (i.e. statistical learning) methods are based on statistics.
- **Subject Matter:** Data science is relevant in all domains with challenging problems that can be solved with sufficient data such as pharmaceutical drug design (Méndez-Lucio et al., 2020) or recommendation systems for movies (Bennett et al., 2007). To find appropriate solutions, a data scientist needs to have a deep understanding of the subject matter that he is dealing with.
- **Computer Science:** Computer science skills are critical for data scientists because the complex calculations that are required for most data

science calculations cannot be performed by hand. Moreover, predictive models are usually integrated into automated workflows. Thus, models have to integrated into efficient software. This requires theoretical knowledge about algorithms and data structures as well as practical knowledge in software engineering.

Because data science requires knowledge in all of these three areas, there is the concept of the *unicorn data scientist* (see the center of the Venn diagram above). This person is an M-shaped professional who has expert knowledge in maths, computer science, and subject matter. However, we all know that unicorns do not exist, or, if they do, they are very rare.

Instead, most data scientists are T-shaped individuals who have a strong general foundation but excel in a specific aspect of data science. The specialization towards two of the three data science areas are indicated by the three pairwise circle intersections indicated in the Venn diagram: two circles in the Venn diagram:

- **Computer science + subject matter expertise → traditional software**: The less you are concerned with maths and statistics, the more you are venturing into traditional software development. The data science roles that are most associated with software engineering are the *machine learning engineer* and the *data engineer*. While ML engineers, integrate ML models into production systems, data engineers are mainly concerned with big data engineering (e.g. databases).
- **Maths and statistics + computer science → machine learning**: The less you are concerned with subject matter, the more you are moving away from applications to theory. For example, ML researchers are usually not very interested in the applications of the methods they develop and they do not require domain knowledge because they develop approaches that are generally applicable and backed by theory. The data science role that is most associated with this direction is the *machine learning expert*. This is a person with a depth of knowledge about ML and can develop ML models that are particularly useful for a given application scenario.
- **Subject matter expertise + maths and statistics → traditional research**: If you forego computer science skills then you move towards a traditional research role. This is because most research positions require profound subject matter expertise (e.g. in biology) but also quantitative skills for evaluating the results of experiments. The role that fits this direction best is probably that of the *data analyst* or *statistician*. The difference between the two roles is that data analsyis is data-driven, while the work of statisticians is more general and often more theory-driven.

You may be wondering if all of the three areas are equally important. The answer is: it depends on the subject domain and the specific problem at hand. Let us consider the impact of the role of the domain first. Different domains inherently have different levels of complexity. Let us revisit the example of movie recommendations and molecular drug design from earlier. While the

subject matter for movie recommendations is easily comprehensible (e.g. movies, genres, customers), the concepts of drug design are not. For example, you probably know the difference between a thriller and a comedy but do you know the difference between the Van-der-Waals force and ionic forces? If you are not working in the field, the answer is probably *no*.

The second aspect are the specifics of the data science project. If the problem is very hard, you probably cannot simply apply standard tools, so you need to have expert machine learning skills. On the other hand, once a well-performing model has been developed, engineering aspects become more and more important.

Which Types of Data Scientist are Out There?

Based on my experience, these are the three most common personas in data science:

Persona	**Likes**	**Dislikes**	**Go-To Tools**
The *ML* Data Scientist	Formulating, tweaking, and evaluating ML models	Data wrangling	ML libraries (e.g. Tensorflow)
The *Analyst* Data Scientist	Statistical analysis/modeling, data visualization	Coding standards	Jupyter Notebooks
The *Tech* Data Scientist	Model deployment and software design	Math formalism	IDE, CI/CD, APIs

What I would like you to take away from this section is that data science is a diverse field and that there are data scientists in all shapes and colors. If you are at the beginning of your journey, I would recommend you to build a solid foundation and then to dive deeper into those areas that bring you the most joy. For example, if you're interested in developing ML models, become an ML expert, or, if you're more interested in the software engineering aspects, become an ML engineer. And, who knows? Maybe you will become a unicorn at one point in the future?

Which Type of Data Scientist are You?

To find out which specialization suits you the most, note down your answers to the following self-asessment questions.

(1) After evaluating several models, you find that none of the models achieve the desired performance. What do you do?

   (A) Analyze the models in greater detail and develop an adjusted prediction algorithm.

(B) Take a deeper look at the data. Maybe there is something you have missed?
(C) Improve the runtime of your evaluation pipeline in order to test additional models.

(2) The tests of your application are failing. What do you do?

(A) Try to fix the tests.
(B) Report the problem.
(C) Fix the tests.

(3) If someone asked you which model you would select for a specific prediction task, would you give a well-founded answer?

(A) Yes
(B) No
(C) No

(4) Your manager asks you to integrate a machine learning model into an existing object-oriented codebase. How do you feel about the task?

(A) I'd rather be doing something else. Moreover, what is object orientation once again?
(B) I'd rather be doing something else. Moreover, what is object orientation once again?
(C) Great, finally I can show my coding skills.

(5) A new deep learning paper has just been published. The application is in drug discovery. You discuss the paper with your team. What is your contribution to the discussion?

(A) You are very excited about the newly developed method and you already have some ideas how the approach could be adopted.
(B) You are amazed about the high performance of the model and wonder how it could impact drug discovery.
(C) The paper sounds interesting but you are wondering whether the model would scale well in the cloud.

(6) A set of new data have just come in. Your manager asks you to form a hypothesis about the data generation process. What do you do?

(A) You have recently read about a Bayesian approach that could work well on the data, so you start doing some research in that area.
(B) You visualize the data and use your subject knowledge to reason about the data generation process.
(C) You ask the manager to give the task to someone else.

Evaluation

You can determine your inclination for the individual aspects of data science by summing up the amount of time you selected each letter:

(A) Machine Learning: you love tinkering with machine learning methods
(B) Data Analysis: you are excited about diving deep into data sets and producing insights
(C) Software Development: your are passionate of getting machine learning models into production

If there is one letter that you selected considerably more than the others, you should think about specializing in that area. If you selected each letter approximately equally, you may be a jack of all trades.

Copyright

Bibliography

Bennett, J., Lanning, S., et al. (2007). The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. New York.

Méndez-Lucio, O., Baillif, B., Clevert, D.-A., Rouquié, D., and Wichard, J. (2020). De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nature communications*, 11(1):1–10.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–503.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708.