**How I’d Learn Data Science If I Could Start Over (4 Years In)**

**A newer and more effective approach**

**A New Perspective**

Two years ago, I wrote a [similar article](https://towardsdatascience.com/how-id-learn-data-science-if-i-could-start-over-2-years-in-b821d8a4876c) explaining how I’d learn data science if I could start over. Now that I’m four years into my career, which is double the amount of time, I’ve realized that there is a much better approach to learning data science.

The problem with my previous guide is that it acts as a one-size-fits-all solution which simply isn’t the case. Because data science covers such a broad spectrum of skills and subjects, it’s only natural that particular skills matter a lot more for certain types of data scientists and a lot less for others.

And so, “How I’d Learn Data Science if I could start over” really starts with the question, “**what aspects of data science am Interested in?**” Is it statistical analysis? Is it deep learning? Is it building visualizations? Understanding this will help with prioritizing what skills to learn first. And if you’re unsure what aspects of data science you’re interested in, that’s completely okay because there are fundamental skills required by all types of data scientists that you can start with (as far I know).

Enjoying this article? [**Subscribe and become a member today**](https://terenceshin.medium.com/membership)to never miss another article on data science guides, tricks and tips, life lessons, and more!

**How I’d Re-Learn Data Science**

Below is a simplified and generalized flowchart that I’d use to guide my learning if I had to learn data science all over again. I want to re-emphasize the **simplicity** of this flowchart in exchange for 100% completeness to make it as comprehensive as possible.

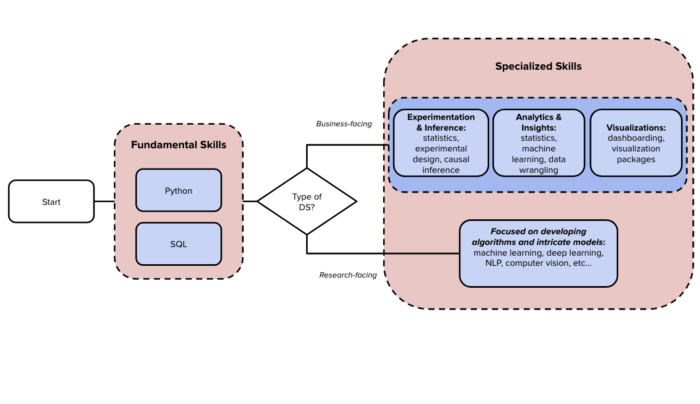


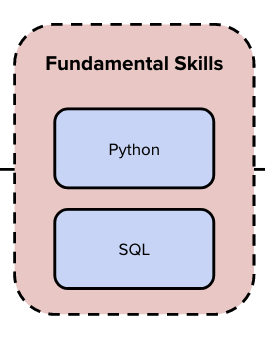
Image created by author

At a high level, the flowchart can be broken down into the following steps:

1. Start with fundamental skills, SQL and Python.
2. Decide whether your interest lies more in business-facing roles or research-facing roles.
3. Based on what you chose in Step 2, select a specialized subject that interests you and that you want to dive deeper into and repeat.

Let’s walk through each step in more detail…

**Fundamental Skills**



Regardless of what area of expertise you want to specialize in, it’s inevitable that you’ll have to know how to code in SQL and Python. And so, I recommend that you learn how to code as a starting point.

**SQL**

SQL is the universal language of data. Whether you’re a data scientist, a data analyst, a machine learning engineer, a data engineer, or a blend of any of these roles, you’re going to need to know SQL.

How I’d learn SQL is through a couple of resources in this order:

* [**Mode SQL Tutorial**](https://mode.com/sql-tutorial/): This is the best SQL course that I’ve ever come across. It’s free, it’s comprehensive, and it’s well written. Take the time to go through this and solidify your knowledge with the practice questions. You **don’t** need to memorize everything, but you should have a general idea of the tools at your disposal.
* [**DataLemur**](https://datalemur.com/): Once you have a fundamental understanding of SQL, DataLemur has a repository of Leetcode-like questions, but specifically for SQL! If you can complete the majority of these questions, you should feel confident in your ability to write relatively complex queries.

**Python/Pandas**

Python is important for data scientists, especially because there are so many packages and extensions of Python that are useful. R is equally as good of an alternative but doesn’t seem to be the main language that’s adopted in the data science world.  
Learning Python is a little less straightforward than SQL because I’ve found that Python is better learned by “doing”, as in trying to build projects. That being said, here are a few resources that I found helpful in my career:

* [**Codecademy**](https://www.codecademy.com/learn/learn-python-3): To learn the basics of Python, and programming in general, Codecademy is a friendly resource to learn the fundamentals.
* [**Pandas Practice Problems**](https://github.com/guipsamora/pandas_exercises): Pandas is a data manipulation language, like SQL. Similar to DataLemur, this repository has dozens of practice problems that you can dive into to learn how to use Pandas. My advice is that you learn Pandas by going through the questions and answers together.

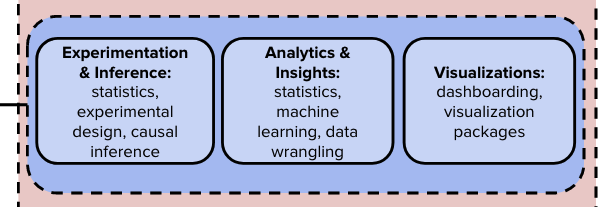
**Specialized Skills**

Once you learn the fundamentals, there are several subjects that you can specialize in. How I would determine what to focus on next first depends on whether I see myself as a Business-facing Data Scientist or a Research-facing Data Scientist.

A **business-facing data scientist** is focused on initiatives that directly impact the business and tends to work with business stakeholders directly, almost like a consultant. Projects and required skills revolve more around solving business problems directly, the lifecycle of projects is relatively shorter and the impact of one’s work is consistently seen.

A **research-facing data scientist** acts more like a researcher or a Ph.D. student. He or she will work on longer-term projects, like building intricate models or conducting complex research questions. The lifecycle of projects is relatively much longer and the work may or may not be used by the business depending on the cost-benefit tradeoff.

If you choose to pursue a role that has more of a direct impact on the business, then there are three sub-categories that I would dive deeper into experimentation & inference, analytics & insights, and visualizations.



**Experimentation & Inference**

Experimentation and Inference refer to a set of techniques that are used to determine the cause-and-effect relationship between two variables. This is extremely important for a business to understand the drivers of success and ultimately what allows businesses to learn, iterate, and improve.

Initial resources to learn the fundamentals are provided below:

* [Statistics: Probability and Stats @ Georgia Tech](https://mediaspace.gatech.edu/playlist/dedicated/74258101/1_g5xwvbde/1_iw8fk73m)
* [Experimental Design: Design of Experiments @ PennState](https://online.stat.psu.edu/stat503/)
* [Causal Inference: Brady Neal](https://www.bradyneal.com/causal-inference-course)

**Analytics & Insights**

**Analytics** refers to organizing and examining data, while **insights** refer to discovering information, like patterns and anomalies, in data. Data Scientists focused on analytics and insights are required to answer vague and generally tough questions using a set of analytical and statistical tools.

Initial resources to learn the fundamentals are provided below:

* [Statistics: Probability and Stats @ Georgia Tech](https://mediaspace.gatech.edu/playlist/dedicated/74258101/1_g5xwvbde/1_iw8fk73m)
* [Machine Learning: Crash Course by Google](https://developers.google.com/machine-learning/crash-course)

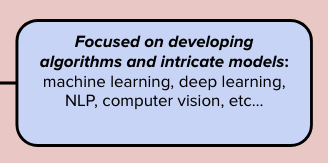
**Visualizations**

Data visualization is the graphical representation of information. Data scientists focused on visualizations are mainly focused on dashboarding, automated reporting, and developing visual insights.

Initial resources to learn the fundamentals are provided below:

* [Dashboarding: 4 Design Principles](https://www.sisense.com/blog/4-design-principles-creating-better-dashboards/)
* [Introduction to Matplotlib](https://www.datacamp.com/community/tutorials/matplotlib-tutorial-python)
* [Types of Visualizations in Plotly (beginner)](https://towardsdatascience.com/interactive-visualizations-with-plotly-ea3f8feb87d1)
* [Types of Visualizations in Plotly (beginner and advanced)](https://towardsdatascience.com/the-next-level-of-data-visualization-in-python-dd6e99039d5e)

**Algorithms**



On the other hand, if you’re more interested in diving into the intricacies of models, reading research papers to keep up with cutting-edge methods, and more interested in the deployment of models, then I recommend that you narrow in on a particular subject related to modeling. Some subjects include **machine learning**, **deep learning**, **NLP**, **computer vision**, **network science**, etc.

[**Saturn Cloud**](https://saturncloud.io/?utm_source=Medium+&utm_medium=TDS&utm_campaign=Terence+Shin&utm_term=how-id-learn-data-science-if-i-could-start-over-4-years)is a platform that allowed me to build computationally expensive models that I wouldn’t have been able to build locally. It’s a great solution if your specs are a bottleneck to your modeling.

**What’s next?**

Once you make it this far, it’s time to work on some data science projects and build your portfolio! Here’s a list of a couple of projects for inspiration if you don’t know where to start:

**[5 Practical Data Science Projects That Will Help You Solve Real Business Problems for 2022](https://towardsdatascience.com/5-practical-data-science-projects-that-will-help-you-solve-real-business-problems-for-2022-a5a3904ea39b" \t "_blank)**

**[A curated list of data science projects that mimic real-life problems](https://towardsdatascience.com/5-practical-data-science-projects-that-will-help-you-solve-real-business-problems-for-2022-a5a3904ea39b" \t "_blank)**

[towardsdatascience.com](https://towardsdatascience.com/5-practical-data-science-projects-that-will-help-you-solve-real-business-problems-for-2022-a5a3904ea39b" \t "_blank)

Some platforms that you can use to start building your own projects are below:

* [**Saturn Cloud**](https://saturncloud.io/?utm_source=Medium+&utm_medium=TDS&utm_campaign=Terence+Shin&utm_term=how-id-learn-data-science-if-i-could-start-over-4-years)is a platform that allowed me to build computationally expensive models that I wouldn’t have been able to build locally. It’s a great solution if your specs are a bottleneck to your modelling.
* [**Anaconda**](https://www.anaconda.com/products/distribution) is one of the most popular data science platforms where you can search and install thousands of Python/R packages.
* [**Kaggle**](https://www.kaggle.com/) offers a no-setup, customizable, Jupyter Notebooks environment. Access GPUs at no cost to you and a huge repository of community published data & code.

And with that, I wish you the best of luck in your endeavors!