**3 Underappreciated Skills to Make You A Next-Level Data Scientist**

**Go beyond the standard curriculum to get the most out of your data.**

“Don’t get a PhD in History,” I was once told by a PhD Student in History. I wasn’t planning on it, but I must admit I found the tip to be a bit ironic and amusing, in particular because of where it was coming from.

I was given this advice by an old TA of mine, a graduate student instructor for an introductory history class I had to take during my first year of college. In his self-deprecating style, he was simply trying to convey that the skill set granted by an advanced history degree wasn’t what the job market needed these days.

That said, there was a silver lining, at least for undergraduates in history. That same TA also told me a separate, positive story about history. He detailed how he had a friend back in college who majored in history and as a result picked up stellar oral and written communication skills. Post-college, an old colleague working in the trenches of the tech industry ran into him, and upon seeing his skill set, offered him a position in a top-tier customer-facing role, which this history graduate excelled in (while making a boatload of money, I might add).

Both these stories illustrate an important point: in the job market, you expand your chances of getting hired if you have a skill an employer is in desperate need of. Oftentimes, the necessary skill might be a bit unorthodox —such as in the case of a history major breaking into the tech field above — but that’s exactly what makes it in demand. Standard members of the field haven’t mastered it — and thus it is coveted.

Data science is a vast and ever-growing field, drawing new proselytes by day and further hypnotizing existing members by night. However, despite its popularity, most people who come into data science roles tend to display the same standard set of skills: data management, proficiency in statistics and probability, user testing, basic data analysis, and foundations of machine learning, among others.

While these skills are undeniably important, you become increasingly attractive to an employer if you combine them with other, less common skills that are nevertheless beneficial to the data science workflow. In this article, I’ll discuss three underappreciated skills you can pick up to add a bit of glamor to your resume as it competes with all the others out there.

**Visualization**

One of the primary goals of data science in general is to glean meaningful insights from data and subsequently present those insights to some target audience. Although this is often done through the standard pipeline of data cleaning, analysis, and modeling, there is another related task which many lack tangible experience in: data visualization.

In the words of U.C. Berkeley professor Marti Hearst, a renowned information visualization researcher, “visual representations can communicate information more rapidly and effectively than text” [1]. Think about the average person sitting on their couch watching the news. Are they going to want to listen to a long lecture about the results of a model full of numbers, equations, and complexities?

No. They won’t. People like simple things and pretty things. Lucky for them, visualizations — if done well — have the potential to simplify data into pretty summaries. If you can master the skill of designing and implementing, you’ll attract attention in a data science setting, no doubt.

There are two parts to this: 1) **understanding what goes into good visualizations** and 2) **actually implementing these visualizations**.

Let’s start with the first part. For a detailed dive into these topics, you can check out my articles Distilling the Essential Principles of Data Visualization [Part 1](https://medium.com/digital-diplomacy/distilling-the-essential-principles-of-data-visualization-part-1-5df4302e401a) [2] and [Part 2](https://medium.com/mlearning-ai/distilling-the-essential-principles-of-data-visualization-part-2-644bd1b01a05) [3]. Here, I’ll just list a few high-level tips to get you started:

* **Don’t try to show everything**: There is no such thing as a perfect, all-encompassing visualization for a data set. Pick 1–2 aspects of the data to highlight, and design a visualization based on that.
* **Keep it simple**: Don’t overdo it — your job is to simplify the data, not make it even more complicated.
* **Choose wise representations**: For example, don’t use a discrete set of random colors to represent a continuous quantitative value like percentage score on an exam. Choose visual representations that are easy for the viewer to interpret.
* **Don’t lie about the data**: If I need to convince you on this one, it’s probably best that you stay away from data science in general.

Okay, but how do you actually make visualizations? Assuming you have absolutely no experience at all, here’s the general pipeline I recommend following, based on my own experience:

1. **Excel/Google Sheets**: These tools have a fixed set of charts you can automatically generate with various data that you pull in yourself. It’s a good way to explore visualizations and learn the basics in a way that doesn’t require too much effort.
2. [**Tableau**](https://www.tableau.com/) **[4]**: Tableau is an extremely useful tool which is used by practitioners everywhere to visualize their data. While it does have a learning curve, it doesn’t require any knowledge of programming, and allows you to explore a fairly wide range of visualizations in a convenient setting.
3. **Matplotlib/Seaborn:** If you’re already comfortable programming in Python, you might skip straight to this step. These are fairly easy-to-use Python libraries which allow you to program basic visualizations in code.
4. **Altair/Plotly/Vega-Lite [5, 6, 7]**: Here’s where things get interesting. If you truly want to be a visualization stud, you’ll need to come up with compelling graphics on your own which aren’t directly based on pre-generated ones. These declarative programming libraries (the first two in Python, the last one in JavaScript) provide a neat set of tools which are more difficult to manipulate than the simpler libraries above, but also provide a lot more freedom regarding what you can do.
5. [**D3**](https://d3js.org/) **[8]**: And finally, this brings us to D3: Data-Driven Documents — widely known in visualization circles as the gold standard for designing and implementing data visualizations. D3 is a JavaScript library which benefits from an extremely powerful characteristic: it can directly manipulate the DOM (Document-Object Model) of a webpage, allowing the programmer to make almost whatever they want and deploy it to the web with ease. It’s by far the hardest skill on this list to learn (and the stage I myself am currently working on developing). However, in return, you’ll be able to absolutely blow people away with the visualizations you design. For a primer, check out the [original research paper](https://www.researchgate.net/publication/51751977_D-3_Data-Driven_Documents#:~:text=Data%2DDriven%20Documents%20(D3),document%20object%20model%20(DOM).) [9] written by its inventor ten years ago. It recently won the Test of Time Award at VIS, the world’s premier academic conference on visualization.

If you do choose to learn this skill, there is a long path ahead. But in return, the rewards you will reap in the pursuit of data science excellence will also be immense.

**Not being afraid to go back to the data**

This past year, I spent some time working on a machine learning research project spearheaded by a colleague of mine. As with many such projects, we needed the model to reach a certain performance threshold on the training data.

However, try as we might, no amount of parameter-tuning or SK-Learn model-searching would give us the numbers we needed. We were forced to deal with a frustrating reality: our ground-truth data simply wasn’t good enough, and we would need to go back to square one.

In other words, our research team had to comb through many, many rows of the training data by hand and re-evaluate the initial labels we’d assigned (in technical terms, this is known as auditing the data). This was a real pain, as it added nearly six months of additional work to the project. However, we simply didn’t have a choice, because the model wasn’t going to improve in any other way.

And though I’ve chosen the specific sub-field of machine learning as a personal anecdote, this is an issue which extends throughout the whole of data science. No matter what you’re doing — building a model, designing a visualization, or setting up a database — *the quality of your data matters*.

This seems like a simple enough statement, especially when you consider the name “data science.” But it’s a surprisingly easy fact to overlook when you’re in the thralls of an involved, important project which has been in the works for many months (or even years). It’s hard to acknowledge that you have to start over. Simply doing so is a skill in itself — an underrated and valuable one.

C’est la vie. The solution is not always in “improving the model,” whatever that means. Sometimes, no amount of fancy Python modules or incoherent statistical maneuvering will get you the result you need, and you will need to accept that perhaps your data is insufficient or inaccurate.

Yeah, it sucks. But at least it’s right.

**Unorthodox Forms of Data**

One of the most coveted skills you can pick up as a data scientist is familiarity and comfort working with unorthodox forms of data. The average data science student learns to manipulate and work with what is arguably the most common form of data: tables of numbers. And this generally works out fine, since myriad jobs require this very skill of their workers.

However, there are plenty of data formats which far fewer people know how to work with — if you make one of them your specialty, you could become a hot commodity on the job market. Here is a (very incomplete) list of examples:

* **Text Data**: A good portion of interesting human data is in the form of words. By “human data” I mean any data which humans produce — think Twitter, Facebook, text messages, etc. As data science continues to make headway as the primary outlet for approaching humanity’s modern problems, understanding how humans think, feel, and communicate (from a technical lens) will be of the utmost importance. Accordingly, learning to work with text data will be an invaluable skill.
* **Image Data**: The majority of people — data scientists included — probably don’t know much about how images are encoded by a computer below the abstraction barrier. Something about pixels and RGB values, right? This is a somewhat harder form of data to work with, and thus it will be all the more beneficial to you should you choose to learn about it.
* **Geo-spatial data**: This is a fun one. One of the most common methods of data communication with the general public is through maps (if you don’t believe me, consider the most recent election season). And yet, few people know how to perform transformations which will take numerical data and turn it into a map. You could be one of them — to get started, take a look at the GeoPandas module in Python, a wonderful tool that integrates well with traditional Pandas.

Most everything in the world can be viewed as data, and thus embodies a potential avenue for the field of data science to explore. It’s therefore unfortunate that we often neglect potentially exciting areas of study in favor of the comfort provided by rows and columns of numbers.

It’s a simple case of supply and demand. Not enough people know how to deal with the immense supply of unorthodox data — become one of them, and you’ll be in high demand.

**Recap and Final Thoughts**

As data science continues to rise in popularity, it will become increasingly more important for individual data scientists to learn currently underappreciated skills. By mastering them, you’ll get the most out of your data (not to mention your resume).

Here’s a cheat sheet for future reference:

1. People like pretty things that make sense. Get good at **visualization**.
2. No one likes hacky solutions. **Go back and fix the data when necessary.**
3. Numbers aren’t everything. **Learn to deal with other data formats.**

I wish you the best of luck in your data science endeavors.

*My name is Murtaza Ali, and I am a PhD Student at the University of Washington studying human-computer interaction. I enjoy writing about education, programming, life, and the occasional random musing.*

*Please consider using my referral link below to sign up as a full-fledged Medium member. You will be able to read unlimited stories a month, and your membership fee directly supports me and other writers.*