Chapter 5 What is Data Science?

5.1 What is a Data Scientist?

Data science is an exciting field with a lot of expected growth and interesting opportunities. But what exactly is a data scientist? What do they do?

One writer gave some insight into the day-to-day life of a data scientist in **this article**.

Some questions you may want to ask yourself as you read this article that might help you see how a data scientist working in industry thinks and works, and how you would fit in to this kind of a job:

- What questions is the writer asking himself?
- What questions is he asking others? (colleagues, employers, etc)
    - Think of this perhaps as what types of questions is he asking, rather than the literal questions. What information is he trying to gain through the questions? Why those questions? What's good about them?
- What surprised you about the day-to-day life of a data scientist?
- What interests or excites you?
- What are you not looking forward to, or what would you not enjoy if this was your job?
- What aspects of the job that he describes do you feel like you need more information about?
- What are the skills (both hard and soft) that the data scientist possesses?

To be an effective data scientist, you must become a problem solver. Everything you will do in your career will be about solving some kind of problem using data. This usually requires learning to think about problems in an organized way. This article discusses the practice of **structured thinking**:

The art of structured thinking and analyzing

Some things you may want to ponder:
- How would you define structured thinking?
- How could past projects or work that you've done have been helped through more structured

thinking?
- How can you demonstrate to potential employers that you've developed a structured thinking approach to your work?

Another article to help show how data scientists work and think:
What I do when I get a new data set as told through tweets

5.2 Languages of Data Science

Being a good data scientist requires a lot more than just being able to write code well, but not being able to code well is a sign of a poor data scientist. Currently, three programming languages drive the data science community. If you want to argue that you are a data scientist, you need to be proficient in at least one and able to use all three.[1]

- **R:** - A successor to the S language with it's first beta release in 2000. Heavily used by trained statisticians and researchers. Thanks to RStudio (established in 2010), data scientists also use this software for their work. (ref)
- **Python:** - Version 2.0 was released in 2000, with version 3.0 arriving in 2008. Pandas is the foundation for data science in Python, and it started development in 2008. Python is heavily used by software engineers as well. (ref1 and ref2)
- **SQL:** - Has been around much longer. In the early 1970s, IBM implemented the language. Oracle created the first commercially available implementation. It is built to handle relational databases but has also been leveraged for other big data database constructs. IT departments heavily use SQL. (ref)

5.2.1 R for data science

- Why you should learn R first for data science
- Why Learn R? 10 Handy Reasons to Learn R programming Language
- R for Data Science Introduction

BYUI students can take MATH 325 to be introduced to R for statistics and MATH 335 to learn R for data wrangling and visualization.

5.2.2 Python for data science

- Advantages of Learning Python for Data Science
- WHY SHOULD YOU LEARN PYTHON FOR DATA SCIENCE?
- A Beginner's Guide to Python for Data Science

BYU-I students can take CSE 110 to be introduced to Python and CSE 250 to be introduced to Python for data science.

5.2.3 SQL for data science

- Is SQL needed to be a data scientist?
- Why do you need to learn SQL?

BYU-I students can take CIT 111 or CIT 225 to be introduced to SQL.

5.3 Requesting and Communicating Data

It's important to remember that most of the people that you'll interact with in your career won't be data scientists, and may not have any experience working with data in the way that a data scientist does. You may be the only "data person" on a team, and will need to communicate with your teammates about your work and present it in a way that they will understand. Read **this article again** for an example of this.

Similarly, you will need to get access to the data that you will be working with. Usually that will come from people who either aren't familiar with your project, aren't data people, or both. Without good work doesn't happen without good data. There's an art to requesting data from and communicating with people unfamiliar with what you are doing. Part of getting good at doing so can only come through time and practice, but there are things you can do from the get-go. Listed below are links to some articles that offer some good advice:

- **How to get the right data? Trying asking for it.**
- **How to ask for datasets**

Here is a real example of some of the pains of requesting data that you should be prepared to handle in your career. Note the actions that the data scientist took to ensure that he had the data that he needed to solve the client's problem. - What questions did he have to ask? - What data was important to him? What didn't matter? Why? - What principles can you learn from this about requesting and communicating data?

Fundamental Statistical Concepts in Presenting Data

Apart from being a good demonstration of what it's like to acquire data in real-world data science work, this is also an excellent example of great data science work in solving a client's problem.

5.4 Marketing Yourself as a Data Scientist

While data science is a rapidly growing field with a lot of opportunities for employment, it's also very competetive. Simply having a degree isn't enough to land the best jobs, you will need to be able to show employers that you're capable of meeting their data needs. Three of the best tools for accomplishing this are your resume, personal Github repository, and LinkedIn profile.

5.4.1 Resumes

In many cases, your resume will be the first thing that a potential employer will see about you. It should showcase your skills, contributions to past projects, and value as a data scientist. Here are some resources to help guide you through gearing your resume geared towards data science jobs, as well as some tools for helping you make a resume in general.

Guides to data science resumes

- How to Write a Great Data Science Resume
- How to Build an Effective Data Science Resume?
- How to Write the Perfect Data Scientist Resume

Resume builder websites

- cvmaker.com
- resume.com

- kickresume.com
- BYU-I Resume support
- Zety Resume Templates

5.4.2 Github

Github has become a staple in the software world for collaborative work, and data scientists also make great use of it. It's a great way to show potential employers the work that you've done in the past so that they can get see a concrete example of some of your technical abilities. Use it as a place to store your work for projects, both professional (where appropriate - don't share anything sensitive in a public place) and personal. Do work that isn't required by work or school to show that data science work is something that you enjoy and post it to your personal Github repository.

If you are unfamiliar with Github, here is a good place to start:
What Is GitHub, and What Is It Used For?

This article explains the kind of material that you should share on your Github repo, and how you can use Github as a tool to present your work and talent:
What do job-seeking developers need in their GitHub?

5.4.3 Linkedin

You've probably heard of LinkedIn before, it's a popular social media platform designed to help people connect with potential employers and other people in a professional setting. Many recruiters will use it as a tool for finding potential employees to fill openings in companies, so it's worth your time to build up a good LinkedIn profile. It's a place where you can post your resume, experience, skills, a link to your Github repostiory, and establish a professional presence as a data scientist. You should also use it as a networking tool to connect with potential employers and interact with professionals already in the industry who can offer advice and further connections and opportunities.

Here is a guide to developing your LinkedIn footprint, as well as how to use it as a tool to further your professional pursuits:
The Complete Data Science LinkedIn Profile Guide

5.4.4 Resources for Finding and Landing a Job

Data science job postings

- **Indeed**
  - Indeed: Data Science
  - Indeed: Analyst
  - Indeed: Statistician
  - Indeed: Data Analysis
  - Indeed: Data Visualization

- **Glassdoor**
  - glassdoor: Data Scientist Intern
  - glassdoor: Data Visualization

- **Chegg Internships**
  - Data Science Internships

Interview tips

109 Data Science Interview Questions and Answers

1. Knowing a language doesn't make you a data scientist, just like knowing English doesn't make you a poet. You will also need to have analytics and visualization capabilities.↩