**Data Science Notes**

Data Science Jobs

Data science combines several disciplines, including statistics, data analysis, machine learning, and computer science. This can be daunting if you’re new to data science, but keep in mind that different roles and companies will emphasize some skills over others, so you don’t have to be an expert at everything. One important piece of advice for your job search is to read data science job descriptions carefully.  This will enable you to apply to jobs you’re already qualified for, or develop specific data skill sets to match the roles you want to pursue. “Data scientist” is often used as a blanket title to describe jobs that are drastically different.

**Data Analyst** visualize and communicates the data

**Data Engineer** sets up infrastructure to mine the data

**Machine Learning Engineer** produce data-driven products

**Data Scientist** uses a combination of all skills but mainly use tools to deal with raw data

<https://blog.udacity.com/2018/01/4-types-data-science-jobs.html>

The skills data scientists need are evolving (and experience with deep learning isn’t the most important one).In a conversation with Jonathan Nolis, a data science leader in the Seattle area who helps Fortune 500 companies, we posed the question, “Which skill is more important for a data scientist: the ability to use the most sophisticated deep learning models, or the ability to make good PowerPoint slides?” He made a case for the latter, since communicating results remains a critical part of data work.

One result of this rapid change is that the vast majority of my guests tell us that the key skills for data scientists are not the abilities to build and use deep-learning infrastructures. Instead they are the abilities to learn on the fly and to communicate well in order to answer business questions, explaining complex results to nontechnical stakeholders. Aspiring data scientists, then, should focus less on techniques than on questions. New techniques come and go, but critical thinking and quantitative, domain-specific skills will remain in demand.

Specialization is becoming more important.While there is no well-defined career path for data scientists, and little support for junior data scientists, we are starting to see some forms of specialization. Emily Robinson described the difference between Type A and Type B data scientists: “Type A is the analysis — sort of a traditional statistician — and Type B is building machine learning models.” Although many working data scientists are currently generalists and do all three, we are seeing distinct career paths emerging, as in the case of machine learning engineers.

Ethics is among the field’s biggest challenges.

Part of this movement involves a reemphasis on interpretability in models, as opposed to black-box models. That is, we need to build models that can explain why they make the predictions they make. Deep learning models are great at a lot of things, but they are infamously uninterpretable.

Data Science Skills

**Programming**

Statistical programming language, like R or Python, and a database querying language like SQL.

**Statistics** (*Most companies, but especially data-driven*)

Be familiar with statistical tests, distributions, maximum likelihood estimators, etc. This will also be the case for machine learning, but one of the more important aspects of your statistics knowledge will be understanding when different techniques are (or aren’t) a valid approach.

**Machine Learning** (*Companies with product that are data-driven*)

A lot of ML techniques can be implemented using R or Python libraries—because of this, it’s not necessary to become an expert on how the algorithms work. More important is to understand the broad strokes and really understand when it is appropriate to use different techniques.

**Multivariable Calculus and Linear Algebra** (*Companies defined by data*)

Understanding these concepts is most important at companies where the product is defined by the data, and small improvements in predictive performance or algorithm optimization can lead to huge wins for the company. In an interview for a data science role, you may be asked to derive some of the machine learning or statistics results you employ elsewhere. Or, your interviewer may ask you some basic multivariable calculus or linear algebra questions, since they form the basis of a lot of these techniques. You may wonder why a data scientist would need to understand this when there are so many out of the box implementations in Python or R. The answer is that at a certain point, it can become worth it for a data science team to build out their own implementations in house.

**Data Wrangling**

Often, the data you’re analyzing is going to be messy and difficult to work with. Because of this, it’s really important to know how to deal with imperfections in data. This will be most important at small companies where you’re an early data hire, or data-driven companies where the product is not data-related (particularly because the latter has often grown quickly with not much attention to data cleanliness), but this skill is important for everyone to have.

**Data Visualization & Communication**

Visualizing and communicating data is incredibly important, especially with young companies that are making data-driven decisions for the first time, or companies where data scientists are viewed as people who help others make data-driven decisions. When it comes to communicating, this means describing your findings, or the way techniques work to audiences, both technical and non-technical. Visualization-wise, it can be immensely helpful to be familiar with data visualization tools like matplotlib, ggplot, or d3.js. Tableau has become a popular data visualization and dashboarding tool as well. It is important to not just be familiar with the tools necessary to visualize data, but also the principles behind visually encoding data and communicating information.

**Software Engineering**

If you’re interviewing at a smaller company and are one of the first data science hires, it can be important to have a strong software engineering background. You’ll be responsible for handling a lot of data logging, and potentially the development of data-driven products.

**Data Intuition**

Companies want to see that you’re a data-driven problem-solver. At some point during the interview process, you’ll probably be asked about some high level problem—for example, about a test the company may want to run, or a data-driven product it may want to develop. It’s important to think about what things are important, and what things aren’t. How should you, as the data scientist, interact with the engineers and product managers? What methods should you use? When do approximations make sense?

<https://blog.udacity.com/2014/11/data-science-job-skills.html>

Data Science Workflow

* Identifying the data-analytics problems that offer the greatest opportunities to the organization
* Determining the correct data sets and variables
* Collecting large sets of structured and unstructured data from disparate sources
* Cleaning and validating the data to ensure accuracy, completeness, and uniformity
* Devising and applying models and algorithms to mine the stores of big data (Data Engineer)
* Analyzing the data to identify patterns and trends (Data Analyst)
* Interpreting the data to discover solutions and opportunities (Data Analyst)
* Communicating findings to stakeholders using visualization and other means (Data Analyst)

<https://datasciencedegree.wisconsin.edu/data-science/what-do-data-scientists-do/>

<https://blog.k2datascience.com/essential-checklist-for-any-data-analysis-or-science-project-7c4fa924e563>

<https://towardsdatascience.com/a-data-science-workflow-26c3f05a010e>

Data Science Process

* Frame the problem: Who is your client? What exactly is the client asking you to solve? How can you translate their ambiguous request into a concrete, well-defined problem?
* Collect the raw data needed to solve the problem: Is this data already available? If so, what parts of the data are useful? If not, what more data do you need? What kind of resources (time, money, and infrastructure) would it take to collect this data in a usable form?
* Process the data (data wrangling): Real, raw data is rarely usable out of the box. There are errors in data collection, corrupt records, missing values and many other challenges you will have to manage. You will first need to clean the data to convert it to a form that you can further analyze.
* Explore the data: Once you have cleaned the data, you have to understand the information contained within at a high level. What kinds of obvious trends or correlations do you see in the data? What are the high-level characteristics and are any of them more significant than others?
* Perform in-depth analysis (machine learning, statistical models, and algorithms): This step is usually the meat of your project, where you apply all the cutting-edge machinery of data analysis to unearth high-value insights and predictions.
* Communicate results of the analysis: All the analysis and technical results that you come up with are of little value unless you can explain to your stakeholders what they mean, in a way that’s comprehensible and compelling. Data storytelling is a critical and underrated skill that you will build and use here.

<https://medium.springboard.com/the-data-science-process-the-complete-laymans-guide-to-what-a-data-scientist-actually-does-ca3e166b7c67>

1) Ask Questions

2) Collect Data (NumPy, Pandas)

3) Clean Data (NumPy, Pandas)

4) Exploratory Data Analysis (Bokeh, Matplotlib)

5) Build a Model (Scikit-learn, Statsmodels)

6) Communicate/ Visualize (Jupyter, Matplotlib, Seaborn, Tableau)





