A Beginner's Guide to Getting Your First Data Science Job

2019 Edition

Springboard

Table of Contents

Introduction

Six years after the Harvard Business Review dubbed it "the sexiest job of the 21st century," data scientist remains one of the most sought-after roles in business.

Explosive growth in the volume of data has impacted every business sector, driving up the number of data science professionals required to help companies uncover the insights they need to stay competitive.

- Job postings for data scientists rose 75 percent on Indeed.com between January 2015 and January 2018.
- LinkedIn calculated that in August 2018 employers were seeking 151,717 more data scientists than existed in the U.S.
- By 2020, the number of jobs for all U.S. data professionals will increase by 364,000 openings to 2,720,000, according to research from IBM.

These often prove to be tricky job openings to fill. It takes an average of five days longer to close data scientist and data analyst roles than the market average, causing employers to pay premium salaries for qualified professionals.

A typical data scientist job pays about $119,000 at the midpoint of salaries and rises to $168,000 at the 95th percentile, according to staffing agency Robert Half Technology.

That's partly why Glassdoor named data scientist the best job in America for the third consecutive year in 2018 (with the number of job openings and overall job satisfaction ratings also weighed in).

Clearly, data science is still "sexy." But it's evolving, thanks in large part to the ongoing artificial intelligence (AI) revolution.

While it used to be the stuff of sci-fi, AI is now so widespread that most non-technical people are familiar with the term and making AI a part of their everyday lives—thanks to virtual assistants like Siri and Alexa, translation tools used while traveling, and other seamless product integrations.

Advances in AI and machine learning are expected to dramatically impact virtually every industry, creating as many as 2.3 million new jobs by 2020, according to Gartner.

AI is powering innovation, helping companies develop new business-critical solutions and services to optimize insights and improve the consumer experience. Businesses need more machine learning experts who can continually improve forecasting models to gain an edge over the competition.

To keep up with these changes in business and within data science, skills will need to shift.

At Springboard, we work to bridge the world's skills gap with flexible and affordable education. Through our self-paced, mentor-guided online workshops, we have helped thousands of people start and advance careers in data science, including machine learning roles.

As part of our mission to make high-quality education accessible for all and to help people advance their careers, we created this guide to careers in data science in 2016—and have comprehensively updated it for 2019. Our goal is to bring you insights from our network of industry experts that will demystify data science careers. Maybe we'll even inspire some of you to pursue a career in this fascinating field.

The Foundations of Data Science

DJ Patil, who built the first data science team at LinkedIn before becoming the first chief data scientist of the United States in 2015, coined the modern version of the term "data scientist" with Jeff Hammerbacher (Facebook's early data science lead) in 2008.

Patil has put it this way:

> "A data scientist is that unique blend of skills that can both unlock the insights of data and tell a fantastic story via the data."

A decade after it was first used, the term remains contested. There is some debate among practitioners and academics about what "data science" means and whether it's different from the data analytics and statistics that companies have long prioritized.

One of the most substantive differences, however, is the amount of data processed now as opposed to a decade ago. In 2020, the world will generate 50 times more data than we generated in 2011—on average, Google now processes more than 60,000 searches every second (5.5 billion per day).

With that in mind, data science can be considered an interdisciplinary solution to the explosion of data that takes old data analytics approaches and uses machines to augment and scale their effects on larger data sets.

So, what does a typical data scientist look like? Patil posits that "the dominant trait among data scientists is an intense curiosity—a desire to go beneath the surface of a problem, find the questions at its heart, and distill them into a very clear set of hypotheses that can be tested."

Springboard

Notice that there is no mention here of a strict definition of data science, nor of a profile that must fit it.

Think about this: Baseball players used to be judged by how good scouts thought they looked, not how many times they got on base—that was until 2002, when the Oakland A's won an all-time league record 20 games in a row with one of the lowest-paid rosters in the league. And elections used to swing from party to party with little semblance of predictive accuracy—that was until Nate Silver correctly predicted every electoral vote in the 2012 elections. (2016 was… a little more complicated.)

Data, and a systematic approach to uncover truths about the world around us, have changed the world.

"More than anything, what data scientists do is make discoveries while swimming in data. It's their preferred method of navigating the world around them," concludes Patil.

To do data science, you have to be able to find and process large data sets. You'll often need to understand and use programming, math, and technical communication skills.

Most importantly, you need to have a sense of intellectual curiosity to understand the world through data, and not be deterred easily by obstacles.

The Different Data Science Roles

Before we dive too deep into what skills you need to become a data scientist, you should be aware of the different roles within the expanding field of data science. While some small companies can call upon a jack-of-all-trades data scientist, more often a data science team will rely on different team members for accomplishing different tasks.

While there are some basics every data scientist should know (e.g., basic statistics), data science roles can vary significantly in their demands and expectations.

Let’s look at the some broad categories of roles that often get lumped under the umbrella term “data science.”

Data Scientists

One definition of a data scientist is someone who knows more programming than a statistician, and more statistics than a software engineer. Data scientists fine-tune the statistical and mathematical models that are applied onto that data. This could involve applying theoretical knowledge of statistics and algorithms to find the best way to solve a data problem. For instance, a data scientist might use historical data to build a model that predicts the number of credit card defaults in the following month.

A data scientist will be able to run data science projects from beginning to end. They can identify a business problem, store and clean large amounts of data, explore data sets to identify insights, build predictive models, and weave a story around the findings.

Springboard

Within the broad category of data scientists, you might encounter *statisticians* who emphasize statistical approaches to data and *data managers* who focus on running data science teams.

Data scientists are the bridge between programming and implementation of data science, the theory of data science, and the business implications of data.

Their average base pay is just below $140,000 USD, according to Glassdoor.

Data Engineers

Data engineers are software engineers who handle large amounts of data, and often lay the groundwork and plumbing for data scientists to do their jobs effectively. They are responsible for managing database systems, scaling the data architecture to multiple servers, and writing complex queries to sift through the data. They might also clean up data sets and implement complex requests that come from data scientists, e.g., they take the predictive model from the data scientist and implement it into production-ready code.

Data engineers, in addition to knowing a breadth of programming languages (e.g., Ruby or Python), will usually know some Hadoop-based technologies (e.g., MapReduce, Hive, and Pig) and database technologies like MySQL, Cassandra, and MongoDB.

Within the broad category of data engineers, you'll find *data architects* who focus on structuring the technology that manages data models and *database administrators* who focus on managing data storage solutions.

Springboard

Their average base pay is higher than data scientists', based on Glassdoor data: $151,000 USD.

Data / Business Analysts

Data analysts sift through data and provide reports and visualizations to explain what insights the data is hiding. When somebody helps people from across the company understand specific queries with charts, they are filling the data analyst (or business analyst) role. In some ways, you can think of them as junior data scientists, or the first step on the way to a traditional data science job.

Business analysts are a group that's adjacent to data analysts, and are more concerned with the business implications of the data and the actions that should result. Should the company invest more in project X or project Y? Business analysts will leverage the work of data science teams to communicate an answer.

Their average base pay is around $84,000 USD, according to Glassdoor, partly because many roles are filled by entry-level graduates with limited work experience.

Machine Learning Engineer

Machine learning engineers are highly sought after and command an annual median salary of $115,000, according to Glassdoor (note that this is a much narrower job function than the previous titles). They're mostly responsible for building, deploying, and managing machine learning projects.

Springboard

Most machine learning roles will require the use of Python or C/C++ (though Python is often preferred). Background in the theory behind machine learning algorithms and an understanding of how they can be efficiently implemented in terms of both space and time is critical.

The easiest path to a career as a machine learning engineer, though by no means the only one, is to start off with a software engineering background and then gain the statistics and machine learning knowledge needed to take on the role. Some also begin as academics more involved with machine learning theory who then develop their software engineering skills.

The Data Science Process

What exactly does a data scientist do? It's a big question with many possible answers. So we turned to Raj Bandyopadhyay, Springboard's data science expert-in-residence and former chief data scientist at Pindrop Security. Turns out, he employs an incredibly helpful framework that is both a way to understand what data scientists do and a cheat sheet to break down any data science problem.

He calls it "the data science process," which he outlines in an easy-to-follow, seven-day email course, updated in 2018. Here's a summary of his insights.

Step 1: Frame the problem

The first thing you have to do before you solve a problem is define exactly what it is. As a data scientist, you'll often get ambiguous inputs from the people coming to you for help, so you have to work with them and ask the right questions to get to the bottom of the issue.

Say you're solving a problem for your company's VP of sales. She wants you to help optimize the sales funnel. You'll want to start with the basics:

- Who are the customers and how do you identify them?
- What does the sales process look like right now?
- What kind of information does the company collect about potential customers?

Springboard

Once you have a reasonable grasp of the domain, ask more pointed questions. In this case, you ask, “What part of the funnel is not optimized right now?”

She tells you she wants her team to spend more time with the customers who are most likely to convert. She also wants to figure out why some customer segments are not converting well.

Bingo! You can now see the data science in the problem. And you can frame the VP’s request into data science questions:

- What are the important customer segments?
- How do conversion rates differ across these segments? Do some perform significantly better or worse than others?
- How can we predict if a prospective customer is going to buy the product?
- Can we identify customers who might be on the fence?

The next step for you is to figure out what data you have available to answer these questions.

Step 2: Collect the right data

In most data science projects, you will be using data that already exists and is being collected. In this case, you find out that most of the customer data generated by the sales department is stored in the company’s customer relationship management (CRM) software, and managed by the sales operations team. The backend for the CRM tool is a SQL database with several tables. However, the tool also provides a very convenient web-based API that returns data in the popular JSON format.

What data from the CRM database do you need? How should you extract it? What format should you store the data in to perform your analysis?

You decide to roll up your sleeves and dive into the SQL database. You find that the system stores detailed identity, contact, and demographic information about customers, in addition to details of the sales process for each of them. You decide that since the data set is not too large, you'll extract it to CSV files for further analysis.

Step 3: Process the data for analysis

Well, now you have your data. But before you draw any conclusions from it, you need to subject it to some data wrangling.

Data can be quite messy, especially if it hasn't been well-maintained. You'll see errors that will corrupt your analysis: values set to null though they really are zero, duplicate values, and missing values. It's up to you to go through and check your data to make sure you'll get accurate insights. You'll want to check for the following common errors:

1. Missing values

2. Corrupted values

3. Time zone differences

4. Date range errors, such as data registered from before sales started

Springboard

Finally, after a lot of wrangling, you're done cleaning your data set, and you're ready to start drawing some insights.

Step 4: Explore the data

Time to dive in! This step is called exploratory data analysis.

The difficulty here isn't coming up with ideas to test, it's coming up with ideas that are likely to turn into insights. You'll have a fixed deadline for your project, so the pressure is on.

You look at the original request: predict which future prospects are likely to convert. What if you split the data into two segments based on whether the customer converted or not and examine differences between the two groups?

Right away, you start noticing some interesting patterns. You notice that there is a large number of customers in their early 30s and far fewer in their 20s, even though the product was designed with the younger audience in mind. Furthermore, many of the customers who converted were targeted via email marketing campaigns as opposed to social media. It's also clear that customers in their 20s are being targeted mostly via social media.

Now you can begin to trace patterns you can analyze more deeply.

Step 5: Perform in-depth analysis

You have enough information to create a model to predict which customers will buy the product.

In order to create a predictive model, you must use techniques from machine learning. In this case, you have to create a predictive model that compares your underperforming group with your average customer. You might find out that the age and social media activity are significant factors in predicting who will buy the product.

You begin to see that the way the product has been has been marketed is significantly affecting sales: maybe this problem group isn't a lost cause! A change in tactics away from social media marketing could change everything for the better. This is something you'll have to flag to your VP of sales.

You can now combine all of those qualitative insights with data from your quantitative analysis to craft a story that moves people to action.

Step 6: Communicate your results

Communication is one of the most underrated skills a data scientist can have. While some of your colleagues can get away with being siloed in their technical bubbles, data scientists must be able to communicate with other teams and effectively translate their work for maximum impact. This set of skills is often called data storytelling.

So what kind of story can you tell based on the work you've done so far? First, you take the data on the current prospects that the sales team is pursuing, run it through your model, and rank them in a spreadsheet in order of most to least likely to convert. You provide the spreadsheet to your VP of sales.

Next, you decide to highlight a couple of your most relevant results:

Age: We're selling a lot more to prospects in their early 30s, rather than those in their mid-20s. This is unexpected since our product targets people in their mid-20s!

Marketing methods: We use social media marketing to target people in their 20s, but email campaigns to people in their 30s. This appears to be a significant factor behind the difference in conversion rates.

The following week, you meet with the VP of sales and walk her through your conclusions. She's thrilled with your work and ready to act on your proposals.

Boom! You've had a real impact on your company's revenue!

Data Scientists in Action

We have a large community of mentors and alumni at Springboard who have shared stories about what a day in the life of a data scientist is like. They're all practitioners in the field with real-life experience. Understanding what they do is a helpful step in fully understanding data science.

From Polyurethanes to Python

After graduating with a bachelor's degree in chemical engineering, **Melanie Hanna** worked in a variety of manufacturing settings, including a polyurethane plant. But lacing up steel-toed boots every morning was not the life for her.

When given the opportunity to work for a large pharmaceutical company in their pilot plant—an offer most chemical engineers would jump at—Melanie realized she had zero excitement to do the job and decided she needed to pivot.

Drawn toward analytical thinking, Melanie ultimately decided that she wanted to pursue data science. Studying chemical engineering in school gave her a base in data science concepts, but Melanie knew she needed further education, so she enrolled in a Springboard course.

In her new role, Melanie works on “productionalizing” models and turning model results into repeatable calculations that can be used throughout the company. She spends most of her day programming, meeting with business stakeholders, and working with data engineers and developers to ensure an error-free pipeline of data and transformations.

An Even Split

In his role as a data scientist at Mozilla, **Ryan Harter**’s typical day is split pretty evenly between analysis, meetings, and review.

He describes it this way:

Analysis is me actually doing computation, writing queries, trying to gauge something about the universe that isn't understood right now.

Meetings and scoping analyses—these are understanding the problem and communicating my results out. I spend a lot of time in meetings trying to draw value from the analyses that I've done.

And then separately I spend a lot of time reviewing and answering questions for other folks. This can be something as simple as writing a quick query for a question that doesn't need a lot of work or reviewing some more junior analyst's analysis or maybe even some PM's analysis.

This mix changes with seniority. More senior data scientists do a lot of reviews, mid-level professionals spend a significant amount of time meeting, communicating insights, and making sure the processes work. Junior members of the team tend to focus on analysis, Ryan said.

Springboard

The Sky's the Limit

From a young age, **Karen Masterson** loved learning about language. As she got older, she started making the connection between programming languages and ancient languages, and began to see parallels between programming languages and natural ones.

Karen was updating her Ph.D. dissertation on an information theoretic approach to languages when she realized her once-theoretical topic was now at the cutting edge of today's technology in natural language processing and artificial intelligence. That helped her make the decision to go back to work full-time in tech after raising her kids.

"I feel like the sky's the limit with data science," she said. "I especially like natural language processing and what can be done with it as far as language translation and artificial intelligence."

In her new role as a data analyst for Verizon Digital Media Services, Karen spends her days solving problems using a practical application of language processing, and tools like SQL and Hadoop.

What You Need to Learn to Become a Data Scientist

Data scientists are expected to know a lot and to possess a diverse skill set, from statistics to communication, coding to machine learning algorithms. They also come from various backgrounds. And there isn't any one specific academic credential that data scientists are required to have.

All the skills and tools we discuss are things you can teach yourself or learn with a Springboard mentor. We've also laid out some resources to get you started down that path.

Data Science Skills

Analytical Mindset

You'll need an overall analytical mindset to do well in data science. A lot of data science involves solving problems. You'll have to be adept at framing those problems and methodically applying logic to solve them.

Mathematics

When data gets large, it often gets unwieldy. You'll need to use mathematics to process and structure the data you're dealing with. Exactly how much and what kind depends on the specifics of your role. But it's safe to say the typical data scientist will have familiarity with statistics, linear algebra, and calculus.

Statistics

You need to know statistics to play with data. Statistics allows you to slice and dice through data, extracting the insights you need to make reasonable conclusions. You must know statistics to infer insights from smaller data sets onto larger populations. This is the fundamental law of data science.

Data Analysis

The process of turning numbers into insights is what it's all about. In the business world, a data analyst will focus on exploring large sets of data and connecting that data with actions that can drive business impact.

Data Visualization

Finishing your data analysis is only half the battle. To drive impact, you will have to convince others to believe and adopt your insights. Human beings are visual creatures. It's typically much easier for us to process information by examining a (thoughtfully created) chart or graph than by poring over a spreadsheet.

Algorithms

Put simply, an algorithm is a well-defined set of steps to solve a specific problem. Data scientists use algorithms to make computers follow a certain set of rules or patterns. Understanding how to use machines to do your work is essential to processing and analyzing data sets too large for the human mind to process.

Machine Learning

Machine learning is the set of algorithms used to make predictions based on a set of known information. Machine learning is what allows for Amazon to recommend you products based on your purchase history without any direct human intervention. It is a group of algorithms that will use machine power to unearth insights for you. In order to deal with massive data sets you'll need to use machines to extend your thinking.

Deep Learning

Deep learning typically refers to the set of machine learning algorithms that extends a basic neural network to much higher levels of complexity, making them capable of learning on much larger data sets and performing many more operations than standard models. The data usually gets this large in image processing and signal processing.

Natural Language Processing

Natural language processing (NLP) uses techniques from computer science, linguistics, and machine learning to process human language, typically in the form of unstructured text. Common applications of NLP include: text classification (e.g., is this news article fake or real?), sentiment analysis (e.g., how much do customers like my product?) and topic modeling (e.g., what are some common themes people are talking about?).

Business Acumen

Data means little without its context. Most companies depend on their data scientists not just to mine data sets, but also to communicate their results to various stakeholders and present recommendations that can be acted upon.

Communication is an underrated skill that can make or break a project.

The best data scientists not only have the ability to work with large, complex data sets, but also understand intricacies of the business or organization they work for.

Having general business knowledge allows them to ask the right questions, and come up with insightful solutions and recommendations that are actually feasible given any constraints that the business might impose.

Domain Expertise

As a data scientist, you should have deep knowledge of the company you work for and also understand the larger industry within which it operates for your insights to make sense. Data from a biology study can have a drastically different context than data gleaned from a well-designed psychology study. You should know enough to cut through industry jargon.

Data Science Tools

With your skill set developed, you'll now need to learn how to use modern data science tools. Each tool has its strengths and weaknesses, and each plays a different role in the data science process. You can use just one of them, or you can use all of them. What follows is a broad overview of the most popular tools in data science as well as the resources you'll need to learn them properly if you want to dive deeper.

File Formats

Data can be stored in different file formats. Here are some of the most common:

CSV

Comma separated values. You've likely opened this sort of file with Excel before. CSVs use a delimiter (a piece of punctuation) to separate out different data points.

SQL

Structured Query Language is a programming language that works well for editing and querying information stored in a relational database. SQL can also be used for advanced analytical operations and for transforming the queried database's structure. You can add or delete tables of data, for example. There are open-source framework implementations of SQL, including the most popular one: MySQL.

JSON

JavaScript Object Notation is a lightweight data exchange format that can be read by both humans and machines.

Excel

Introduction to Excel

Excel is often the gateway to data science, and something that every data scientist can benefit from learning. It allows you to easily manipulate data with what is essentially a What You See Is What You Get editor that lets you perform equations on data without working in code at all. It is a handy tool for data analysts who want to get results without programming.

Benefits of Excel

Excel is easy to get started with, and it’s a program that anybody who is in analytics will intuitively grasp. It can be very useful to communicate data to people who may not have any programming skills: they should still be able to play with the data.

Who Uses This

Data analysts tend to use Excel.

Level of Difficulty

Beginner.

Sample Project

Importing a small data set on the statistics of professional basketball players and making a simple graph of the top scorers in the league.

SQL

Introduction to SQL

Data science needs data. SQL is a programming language specially designed to extract data from databases.

Benefits of SQL

SQL remains one of the most popular tools used by data scientists. Most data in the world is stored in tables that will require SQL to access. You'll be able to filter and sort through the data with it.

Springboard

Who Uses This

Data analysts and some data engineers tend to use SQL.

Level of Difficulty

Beginner.

Sample Project

Using a SQL query to select the top 10 most popular songs from a SQL database of the Billboard 100 music chart.

Python

Introduction to Python

Once you download Anaconda, an environment manager for Python, and get set up on Jupyter Notebook, you'll quickly realize how intuitive it is. A versatile programming language designed for everything from building websites to gathering data from across the web, Python has many code libraries dedicated to making data science work easier.

Benefits of Python

Python is a versatile programming language with a simple syntax that is easy to learn.

Springboard

The average salary for jobs with Python in their description is around $102,000. Python is one of the fastest-growing programming languages in the world. The Python community is passionate about teaching Python and building useful tools that will save you time and allow you to do more with your data.

Many data scientists use Python to solve their problems: 63 percent of respondents to the annual data science survey conducted by O'Reilly used Python, an increase from 58 percent the previous year.

Who Uses This

Data engineers and data scientists use Python for medium-sized data sets.

Level of Difficulty

Intermediate.

Sample Project

Using Python to source tweets from celebrities, then doing an analysis of the most frequently used words by applying programming rules.

R

Introduction to R

R is a staple in the data science community because it is designed explicitly for data analysis. R shines when it comes to building statistical models and displaying the results.

Benefits of R

R has fallen slightly behind Python in popularity among data scientists in recent years, but it remains an important tool. It is an environment where a wide variety of statistical and graphing techniques can be applied.

The community contributes packages that, similar to Python, can extend the core functions of the R codebase so that it can be applied to very specific problems, such as measuring financial metrics or analyzing climate data.

Who Uses This

Data engineers and data scientists will use R for medium-sized data sets.

Level of Difficulty

Intermediate.

Sample Project

Using R to graph stock market movements over the last five years.

Big Data Tools

Big data comes from Moore's Law, a theory that computing power doubles every two years. This has led to the rise of massive data sets generated by millions of computers.

Put simply, big data is a collective term that describes data that is too large to fit on a single computer. Conventional tools like SQL and Excel are typically unable to handle big data, so new ones have been developed to take their place.

Here are tools that help solve that problem:

Hadoop

Introduction to Hadoop

Hadoop MapReduce is a software framework that treats multiple computers as if they were one, which enables a single set of commands to utilize the storage and processing power of as many computers as you want (or can afford to pay for).

Benefits of Hadoop

Hadoop is an open-source ecosystem of tools that allow you to MapReduce your data and store enormous data sets on different servers. It allows you to manage much more data than you can on a single computer.

Who Uses This

Data engineers and data scientists will use Hadoop to handle big data sets.

Level of Difficulty

Advanced.

Sample Project

Using Hadoop to store massive data sets that update in real time, such as the number of likes Facebook users generate.

NoSQL

Introduction to NoSQL

Tables that bring all their data with them can become cumbersome. NoSQL includes a host of data storage solutions that separate out huge data sets into manageable chunks.

Benefits of NoSQL

NoSQL was a trend pioneered by Google to deal with the impossibly large amounts of data they were storing. Often structured in the JSON format popular with web developers, solutions like MongoDB have created databases that can be manipulated like SQL tables, but which can store the data with less structure and density.

Who Uses This

Data engineers and data scientists will use NoSQL for big data sets, often website databases for millions of users.

Level of Difficulty

Advanced.

Sample Project

Storing data on the users of a social media application that is deployed on the web.

Bringing Tools Into the Data Science Process

Each one of the tools we've described is complementary. They each have their strengths and weaknesses, and each one can be applied to different stages in the data science process.

Collect Data

Sometimes it isn't doing the data analysis that is hard, but rather finding the data you need. Thankfully, there are many resources.

You can create data sets by extracting data from an API (application programming interface). These allow developers to programmatically interact with data. Twitter, Facebook, and Yelp are among the large tech companies with the most popular APIs.

Springboard has compiled 19 of our favorite public data sets on our blog.

Python supports most data formats. You can play with CSVs or with JSON sourced from the web. You can import SQL tables directly into your code.

You can also create data sets from the web. Python's Requests library scrapes data from different websites with a line of code. You'll be able to take data from Wikipedia tables, and once you've cleaned the data with the Beautiful Soup library, you'll be able to analyze them in-depth.

R can take in data from Excel, CSVs, and text files. Files built in Minitab or in SPSS format can be turned into R data frames.

The Rvest package will allow you to perform basic web scraping, while magrittr will clean and parse the information for you. These packages are similar to the Requests and Beautiful Soup libraries in Python.

Process Data

Excel allows you to easily clean data with menu functions that can remove duplicate values, filter and sort columns, and delete rows or columns of data.

SQL has basic filtering and sorting functions so you can source exactly the data you need. You can also update SQL tables and clean certain values from them.

Python uses the pandas library for data analysis. It is much quicker to process larger data sets than Excel, and it has more functionality.

You can clean data by applying programmatic methods to the data with pandas. You can, for example, replace every error value in the data set with a default value, such as zero, in one line of code.

R can help you add columns of information, reshape, and transform the data itself. Many of the newer R libraries allow you to play with different data frames and make them fit the criteria you've set.

NoSQL allows you to subset large data sets and to change data at will, which you can use to clean through your data.

Explore Data

Excel can add columns together, get the averages, and do basic statistical and numerical analysis with pre-built functions.

Python and pandas can take complex rules and apply them to data so you can easily spot high-level trends.

You'll be able to do deep time series analysis in pandas. You could track variations in stock prices to their finest detail, for example.

R was built to do statistical and numerical analysis of large data sets. You'll be able to build probability distributions, apply a variety of statistical tests to your data, and use standard machine learning and data mining techniques.

NoSQL and Hadoop both allow you to explore data on a similar level to SQL.

Analyze Data

Excel can analyze data at an advanced level. Use pivot tables that display your data dynamically, advanced formulas, or macro scripts that allow you to programmatically go through your data.

Python has a numeric analysis library: NumPy. You can do scientific computing and calculation with SciPy. You can access a lot of pre-built machine learning algorithms with the scikit-learn code library.

R has plenty of packages out there for specific analyses, such as the Poisson distribution and mixtures of probability laws.

Communicate Data

Excel has basic chart and plotting functionality. You can easily build dashboards and dynamic charts that will update as soon as somebody changes the underlying data.

Python has a lot of powerful options to visualize data. You can use the Matplotlib library to generate basic graphs and charts from the data embedded in your Python. If you want something that's a bit more advanced, you could try Plot.ly and its Python API.

You can also use the nbconvert function to turn your Python notebooks into websites or your online portfolio. Many people have used this function to create online tutorials on how to learn Python.

R was built to do statistical analysis and demonstrate the results. It's a powerful environment suited to scientific visualization with many packages that specialize in graphical display of results. The base graphics module allows you to make all of the basic charts and plots you'd like from data matrices. You can then save these files into image formats or you can save them as separate PDFs. You can use ggplot2 for more advanced plots such as complex scatter plots with regression lines.

Python vs. R

These languages are always in competition to be the language of choice among data scientists.

Python is versatile, simple, easier to learn, and powerful because of its usefulness in a variety of contexts, some of which have nothing to do with data science. R is a specialized environment that looks to optimize for data analysis, but which is harder to learn.

While the Python vs. R debate is often framed as a zero-sum game, in reality it's not. Learning both tools and using them for their respective strengths can only help you improve as a data scientist. Twelve percent of data scientists polled by KDnuggets in 2017 used both R and Python.

O'Reilly found in their survey of data scientists that using many programming tools is correlated with increased salary. While those who work in Python may be paid more than those who work in R ($100,000 to $90,000, according to the most recent data), those who used 15 or more tools made $30,000 more than those who used 10 to 14.

The Python vs. R debate really doesn't need to happen unless you really want to confine yourself to one programming language. The reality is that as a data scientist, you'll often be called upon to do different tasks, and you'll only be able to do them better if you know exactly which tool is best.

We recently asked two working data scientists for their thoughts on this debate.

Springboard mentor and Mozilla data scientist Ryan Harter:

"Python, I think, is more general purpose. I characterize this by saying that Python is maintained by computer scientists and R is maintained by statisticians. Python can be more general purpose, better if you need to do some scraping or something general like that. But R is great for running statistical analyses. There are a lot of things built in that haven't caught up in the Python world. If you're trying to build things like visualizations or you're trying to make very quick progress as a data scientist, R can be a very useful tool."

Mansha Mahtani, a data scientist at Instagram:

"R and Python are very different languages, but I wouldn't say that there's one that's better than the other. It really depends on preference and what you're trying to do. So, if you have had more object-oriented programming experience, you probably are going to hate R. But if you are someone who likes doing ad hoc analysis, if you want to just keep checking your code in the middle or coming up with quick visualizations, R might be the right place for you. I wouldn't get too married to choosing a specific language. I would say just go with what you're comfortable with. What's great is both languages are trying to compete with each other to be the preference for data scientists, so they're kind of evolving to do the same things."

Springboard

Starting Your Job Search

How to Build a Data Science Portfolio and Resume

No matter the field, it's important to make a great first impression. But this is particularly important when you're transitioning into a new career. It all starts with your portfolio and your resume.

Many data scientists have their own websites, which serve as both a repository of their work and a blog. This allows them to demonstrate their experience and the value they create in the data science community.

In order for your portfolio to have the same effect, it must share the following traits:

1. Your portfolio should highlight your best projects. Focusing on a few memorable projects is generally better than showing a large number of them.
2. It must be well-designed and tell a captivating story of who you are beyond your work.
3. You should build value for your visitors by highlighting any impact you've had through your work. Maybe you built a tool that's useful for a general audience? Perhaps you have a tutorial? Showcase them here.
4. It should be easy to find your contact information.

Data scientist, instructor, and Springboard mentor David Yakobovitch recently shared his thoughts on what makes a data science portfolio stand out:

What should be included in a data science portfolio?

If a recruiter or someone who wants to hire you is going to look at your link, it's important that what you share be able to stand alone without you.

In a competitive job market, it's important that if you send the link, someone could look at it and understand the business case, they can read the whole problem. So I think the one area that often people can add more to is that business case, writing up those paragraphs.

In data science, you can write that in Markdown. After every single line of code, write a thoughtful interpretation: what's occurring here, what's the meaning of the analysis, how does it relate to your client or customer?

How many projects should be in a data science portfolio?

For the [Data Science Career Track](), we have two capstones that students work on, so I like to say a minimum of two projects in your portfolio. Often when I work with students and they've finished the capstones and they're starting the job search, I say, "Why not start a third project?" That could be using data sets on popular sites such as [Kaggle]() or using a passion project you're interested in or partnering with a non-profit.

When you're doing these interviews, you want to have multiple projects you can talk about. If you're just talking about one project for a 30-to-60-minute interview, it doesn't give you enough material. So that's why it's great to have two or three, because you could talk about the whole workflow—and ideally, these projects work on different components of data science.

What's the best order?

The order really depends on what message you'd like to convey and the strength of what you've accomplished. I believe that you should show your projects in order of complexity. So if you have three capstones, the one that has the more advanced machine learning or the bigger scale you could show first, and then the others could follow.

If you're going to show your projects on LinkedIn or on your resume, you don't need to show many paragraphs. You could put one to two sentences and the link: "This is the supervised machine learning model on credit default risk and here's the GitHub link." It could be that simple.

What's the optimal format?

I think GitHub is the best resource to showcase this. You can set up your own repositories, where it has full support from Markdown, which lets you format your reports. And you can also have all your code—your Jupyter notebooks, your R files—they could all be shared there.

In addition, as part of the portfolio, it's not only important to show the code and to show the Markdown files, but if you can make a business case for it presentable to an audience, that's really helpful.

Two things I encourage all students to do: the first is to create a PowerPoint. You have 10 to 20 slides that summarize it, so anyone can look at that PowerPoint or PDF and see what you've accomplished and the business case surrounding that project.

And secondly, going the extra mile is creating a YouTube video of you doing the presentation. You can keep it as an unlisted video and if a recruiter ever says, “Hey, I want to see an example of your portfolio or a project you’ve done,” not only can you point them to GitHub, but you can also show them your personality and your presentation skills by sharing that YouTube link.

How to Network and Build a Personal Brand in Data Science

Once you have learned the required skills and developed a strong portfolio, the next step is to connect with people who can help you leverage those strengths into a data science job.

Building your network among data scientists will substantially increase your odds of breaking into the field. Many of the best opportunities aren’t posted on job boards.

Find a Mentor

One of the highest-value networking activities you can pursue is finding a mentor who can guide you as you seek and pursue a data science career. Somebody who has been in a hiring position can tell you exactly what companies are looking for and how to prepare for interviews. She can also introduce you to other people in the data science community, or in the best of cases, even end up hiring you!

What some people don’t understand is that mentorship is a two-way street, and you can always create value for your mentor in different ways, whether it’s sharing your story, or giving them some perspective on problems they see. Mentorship is a special category of a relationship where you can build value for yourself in a professional context—but never forget the golden rule of relationships: you get what you give.

We've seen the benefits of mentorship first-hand at Springboard. In all of our courses, students are paired with a mentor from the industry they hope to break into, which leads to significantly better outcomes through increased accountability and motivation.

Go to Conferences

At some of these events, you will get to hear from and build connections with established data scientists, and even unearth hidden job opportunities. With a bit of searching, you can find great data science events in your area.

Here are a few to consider:

> The Strata Data Conference, created in 2012, is the largest data conference series in the world. Speakers come from academia and private industry. The themes tend to be oriented around cutting-edge data science trends in action. Practical workshops are provided if you want to learn the technology behind data science, and there are plenty of networking events.
>
> KDD (Knowledge Discovery in Data Mining) is a large interdisciplinary conference bringing together researchers and practitioners from data science, data mining, knowledge discovery, large-scale data analytics, and big data. It's also an organization that seeks to lead discussion and teaching of the science behind data science. Membership and

attendance at these conferences offers an awesome way to contribute to growing trends in data science.

NeurIPS, or Neural Information Processing Systems (previously known as NIPS), is a largely academic data science conference focused on evaluating cutting-edge science papers in the field. Attending will give you a sneak preview of what will shape data science in the future.

The International Conference on Machine Learning (ICML) is supported by the International Machine Learning Society and brings together some of the best minds in machine learning to present research and discuss new ideas. It was first held in 1980.

Attend Meetups

We've listed the major conferences where the data science community assembles, but there are many smaller meetups that serve to connect the local data science community.

The San Francisco Bay Area tends to have the most data meetups, though there is usually one in every major city in the U.S. You can look up data science meetups near you with Meetup.com. Some of the largest data science meetups, with more than 4,000 members, are SF Data Mining, Data Science DC, Data Science London, and the Bay Area R User Group.

Most data science meetups are organized by influencers in the local data science community. If you really want to make a splash, you should consider volunteering at a data science event.

Most events follow the same format, with an invited speaker who gives a talk and then a networking period where everybody connects with each other. The general data science meetups will often have an industry talk where somebody will delve into a real-world data science problem and how it was solved. Specialized data science meetups, such as Python or R groups, will often focus on technical tutorials that teach a specific tool or skill.

You should introduce yourself to the local data science community! You may just walk away with a career opportunity.

Other Ways to Network

We live in a digital world, so you shouldn't feel confined to offline networking! Some of the best data scientists are on Twitter, and you can discover influencers worth following (and potentially connecting with) through podcasts and other outlets.

Talking Machines includes interviews with prominent data scientists. Partially Derivative has been described as "'Car Talk' for the data community." The O'Reilly Data Show is the equivalent of a graduate seminar delivered in podcast form.

You'll also find online blogs, newsletters, and communities such as O'Reilly and KDnuggets that will help you connect with data scientists online.

Make sure to check out Reddit and Quora, where you can engage in trending data science discussions, and you'll always find a lot of great programming resources and related content on Hacker News.

Job Boards

1. Kaggle offers a job board for data scientists.

2. You can find a list of open data scientist jobs at Indeed and Glassdoor.

3. Datajobs is a listings site for data science.

4. Data Science Central has a frequently updated jobs section.

As previously mentioned, you can also find job opportunities through networking and through finding a mentor. We continue to emphasize that the best job positions are often found by talking to people within the data science community.

You'll also be able to find opportunities for employment in startup forums. Hacker News has a job board that is exclusive to Y Combinator startups. Y Combinator is arguably the most prestigious startup accelerator in the world. AngelList is a database for startups looking to get funding and it also has a robust jobs section.

Ace the Data Science Interview

An entire book could be written on the data science interview—in fact, we have one of those! But here's a condensed guide to help prepare you to nail the various parts of the interview process.

The Phone Screen

Your point of entry typically will be the human resources department. Sometimes there will be basic technical questions to screen out unqualified candidates, but most of the time, this screen involves establishing the beginnings of a culture fit and making sure that you have the communication skills to come off well in a subsequent interview with the hiring manager.

Many candidates underestimate this process because it's not technical and it's with a recruiter. But if you don't do well here, you won't go to the next stage. So come prepared with thoughtful, practiced responses to some of the typical questions:

1. *Why do you want to be a data scientist/analyst?*
    a. The best kind of answer is one that shows you're connecting dots from your past such that data science is the obvious next step. Maybe you were a political science major and economics minor who had an internship in government and found you loved working with large data sets. It's clear, then, that data science is the immediate next step.
2. *Why do you want to work at this company?*
    a. Mission alignment is important. You might be looking at jobs across many industries, but you'll come out on top here if you can tie your decision-making to the company's current problems.
    b. Show that you've done your research; point to the team size and stage, recent press you've read, positive impressions from Glassdoor or Muse profiles, etc.
    c. Try to show genuine motivation—not just "I need a job," but "I need this job."

3. *Talk about a time you had to handle a conflict with your colleagues or manager.*
    a. They're looking for smarts, good communication skills, problem-solving ability, an absence of ego, and above all else, self-awareness and introspection.
    b. Read between the lines: tell them not just the situation and what you did to resolve it, but also what you learned from it for the future.
    c. Making yourself look good at others' expense actually doesn't reflect well on you! Speak positively about past coworkers and companies.

Remember not to treat the person talking to you on the phone as "just a recruiter." Maintain full professional courtesy with everyone in the interview process, no matter what their role in the organization.

Finally, during this call you'll want to get a sense of what problems the data team is facing and the organizational structure of the team you'd be joining. This will help you later in the process.

The Take-Home Assignment

After the phone screen, companies often send a prepared assignment to candidates, with some time pressure applied to screen out people who may be technically weak or who may not be committed to the recruitment process.

Common example assignments include:

- A deep analysis on a specific data set provided for you. Here, you'd be expected to storytell around insights found in the data.

- Cleaning a data set with significant errors.
- Working with a specific problem relevant to the business (e.g., building a job recommendation system for applicants based on data from job descriptions).

Take the time needed to do the assignment well, and try to see how it relates to problems the company is solving. Don't just answer the technical question. Focus on the story and providing concrete recommendations based on your answers. It's better to provide a simple analysis with practical business recommendations than a complex technical analysis with no real business implications.

The Call With the Hiring Manager

This will likely be the final evaluation before a company invites you to an on-site interview. The call typically is split into three components:

Mathematics/statistics - You'll be evaluated on core concepts here, but the specifics will depend on the role and company. Web companies may focus on p-values and statistical significance. Energy companies might test you on regression and linear algebra. No matter what type of interviewer you're talking with, sketch out your thought process. Treat the question like a mathematical proof and a test of your ability to reason, but also try to tell a coherent story about why this matters to the company.

Springboard

Coding - You'll be evaluated on your ability to solve coding challenges by presenting pseudo code or compile-ready code. If you're applying for a data analyst position, this will swing more toward how you'd think about querying data with SQL. Otherwise, you'll be asked questions in the programming and scripting languages you've claimed experience in, from Java to Python. You will likely be asked about data structures more than anything else. Know hashmaps, trees, stacks, and queues very well.

> *Note:* Your interviewer may also use tools like HackerRank or Collabedit to evaluate you live. In this case, your hiring manager will watch you as you type out your solution. Train with those tools if you can!

Communication/culture - The hiring manager will try to get a feel for your character, your motivation, and your fit with their team. Most hiring managers have a mental model for who they are looking for. The closer you fit to it, the more likely you'll pass to on-site interviews. This is where your work with the recruiter will shine. The more you know about the problems the hiring manager is facing and the kind of person they're looking for, the better you'll be able to present yourself as a fit.

> *Tip:* Networking in advance with people who work or have worked for the company will give you an idea what the culture is like.

On-Site Interview - With the Hiring Manager

Finally meeting face to face, the hiring manager will be evaluating you from both a technical and nontechnical perspective. They're looking to ascertain if you're a culture fit, and they may test you on your technical chops by having you whiteboard different scenarios.

In general, hiring managers appreciate when you demonstrate:

1. Passion for the company and data science in general

2. An ability to get along well with everybody, which may even help you balance weaknesses in your technical ability

3. Strong willingness to learn and demonstrated ability to do so rapidly

4. A strong record of previous projects and the ability to relate them with impact driven

5. Strong analytical ability

On-Site Interview - Technical Challenge

Prepare to be challenged on your technical skills in one form or another, especially for roles that lean more toward data engineering. You'll often find that this is similar to a software engineering interview, where you will be asked to whiteboard and write down how you'd implement certain algorithms or solve problems.

This is where strong knowledge of software engineering concepts, such as time complexity or Big O notation, and a strong grasp of the mathematics and statistics behind data algorithms can truly shine.

On-Site Interview - With an Executive

If you pass the bar for your hiring manager, you'll likely have a final interview with a senior executive. In a startup, this will often be the founder. Normally, only candidates who have passed the technical assessment will get here, so now you need to emphasize how you can drive impact with your knowledge of the business itself and the problems it faces.

Tips:

1. Dress accordingly. A great rule is to dress one notch higher than the employees normally do.

2. Research your interviewer and the company. Come up with relevant questions to ask. The importance of this cannot be overstated. Here are some examples:

- What does success look like for this position?
- Which projects do you have in mind for this role?
- What is the last project you shipped? How long did it take? What were the stumbling blocks?
- What will my first 90 days in this role look like?
- Why did the last person who quit this team or company leave?
- What is the current state of the data infrastructure?
- What's the riskiest/most audacious project the team has taken on in the past year, and what was the result?

- If you had unlimited headcount, how many people would you hire, and what roles would you hire them for?
- How are tasks prioritized? When was the last time you had to circumvent that process?

3. Keep answers detailed but concise. The interviewer can lose interest if your answers are long-winded and unfocused.

4. Smile and be confident. Don't come in stressed. Meditate, stretch, or read—do whatever it takes to get you to your peak.

We've put together an exhaustive list of 109 data science interview questions that will help prepare you for these on-site meetings.

Paths Into Data Science

For people interested in becoming data scientists, few things are as interesting or inspiring as the stories of other people who made the transition.

What background do they come from? What skills did they have to pick up? Who are these folks and how can you become one of them them?

Fortunately, at Springboard we have a long list of mentors and alumni who can tell you about their journey into data science and help you with yours. Here are some of their stories.

Turning Frustration Into Success

As an analyst for the Federal Energy Regulatory Commission, **Aaron Pujanandez** focused on market power studies of the U.S. electric industry. He worked with data, but on the level of making recommendations on future events based on assumptions that they would match historical trends. For Aaron, that wasn't satisfying.

Aaron's interest in more advanced data science topics grew from his frustration with the limitations of the data tools that were available to him.

He decided to explore new techniques so he could develop better answers to the questions he was struggling with.

Armed with the skills he learned through Springboard's course, Aaron created a brand new position for himself at the Federal Energy Regulatory Commission and then moved on to Accenture Federal Services, where he's an analytics and modeling manager.

"In my new position, I'm able to use my data science skills to help government agencies derive insights from their data to drive business decisions," he said. "As a manager, I'm able to work more as a mentor and advisor for junior data scientists just starting their career."

Aaron's background shows that if you find yourself limited, you can create a path toward professional satisfaction.

Escaping Academia for Startup Life

Meghan Thomason has a Ph.D. in ecology from the University of California, Davis. During her academic career, she focused on invaded plant communities, trying to understand why non-native plants were successful, often from a perspective of studying their competitive interactions with resident species. She studied grasslands, rangelands, chaparral, and riverine wetlands and brackish marshes all over Northern California. Over the course of her studies, she learned increasingly complex statistical techniques, leading her to consider data science as a career.

"Although I enjoyed research, it became clear the job market was severely oversaturated, and I began researching other opportunities for someone with quantitative and programming skills," said Meghan, who began working for a digital publishing startup shortly after completing Springboard's Data Science Career Track.

Meghan's background proves that you can unearth insights from many different fields in data science.

Hooked by a Book

Srdjan Santic had a solid foundation from which to begin building a full career in data science. He holds bachelor's and master's degrees in economics and quickly found work as a statistician, spending six years at a large marketing research consultancy. There, he cut his teeth on analytics, statistical inference, and predictive modeling.

He connects his full transition to a book he often used at work that relied on the commercial software SPSS for its hands-on examples. One day, he noticed that the same authors (Andy Field and Jeremy Miles) had released a new edition using the statistical programming language R for examples.

"It intrigued me, I ordered it, and after the first few chapters, and only a few exercises of writing code to accomplish a task, I was instantly hooked," Srdjan said. "I continued to work through the book and learn R, also did some online courses on R and Python, and decided that this was the toolset I want to keep using in my career."

He remembers the exact day in the summer of 2013 when he made the decision: “I suddenly thought to myself, ‘I think I’m more than ready to get a full-time job using just these tools.’”

A week later, he landed a part-time gig; six months later, he started his first full-time data science job.

Srdjan’s story shows that listening to your gut can pay off.

The Intersection of Art and Science

James Flint describes himself as “a kid from Stratford-upon-Avon, England, who grew up wanting to be a writer but kept getting distracted by computers.” After his science- and tech-inflected studies, he began writing for the U.K.’s first newspaper technology supplement at The Independent, helped launch the country’s first dedicated art and tech magazine, Mute, and worked for Wired magazine before becoming head of digital development at The Telegraph.

Firmly planted in the intersection of art and science, James took a leap into the startup world, eventually co-founding an online community platform designed for health teams and patients.

“What I have come to realize over the years, and what is so wonderfully refreshing (and reassuringly human) about the current AI inflection in the evolution of computation," he says, "is that data science and machine learning require statistics and storytelling, deduction and induction, coding and writing."

James' experience proves that there's an art to data science.

Conclusion

Getting your first data science job might be challenging, but it's an achievable goal. It takes a diligent approach to picking up skills, working on projects, building a portfolio, and getting your work (and yourself) in front of the right people.

We hope that going through this guide has brought you a little bit closer to your goal of breaking into a career in data science.

For additional support, check out our Data Science Career Track, a mentor-guided, career-focused bootcamp designed to get students a job as a data scientist within six months of completing the course—or their tuition is refunded.

If you thought this was valuable, share a free copy of this guide to your friends or coworkers on Facebook, Twitter, or Linkedin.

Additional Resources

Data Science Roles

This article from Hacker Noon, written by Google's chief decision intelligence engineer, runs through the different data science roles from a unique perspective.

This Springboard blog post also provides a deep dive into the different career paths.

Skills and Tools

This Quora post is a broad overview of many of the essential skills you need to become a data scientist, and resources to go about learning them.

The following introduction to Python will get you set up on the basics.

This blog will help you with all of the latest news in Excel data visualization.

This interactive tutorial to R will help you grasp the fundamentals.

W3Schools has an excellent interactive tutorial on SQL that will get you started on how to select parts of a database for further analysis.

Getting Data

This Springboard blog post lists many of your options for finding high-quality public data sets.

Springboard

Algorithms

The 10 most influential data mining algorithms can be quite complex (and there are many others), but this blog post explains them in plain English.

Machine Learning

This repository on machine learning offers a solid definition and working examples you can get started on right away. If you're more of a visual learner, this introduction to ML concepts will fill the gap for you.

Data Visualization

Flowing Data is a blog that focuses on data communication and the design of appealing data visualizations.

Data Science Interviews

Here is a list of data science interview questions and how to prepare for them.

Building a Data Science Portfolio

Check out this webinar for data science portfolio guidance from a Springboard mentor.

Springboard

Learn more about data science!

Introduction to Data Science

Build a foundation in R programming and statistics as you prepare for data analytics and junior data scientist roles.

Enroll now

Intermediate Data Science: Python

Work on real-world projects to practice Python for data science; delve into machine learning and inferential statistics.

Enroll now

Data Science Career Track

Get a complete education in data science along with personalized career coaching—plus a job guarantee!

Apply now