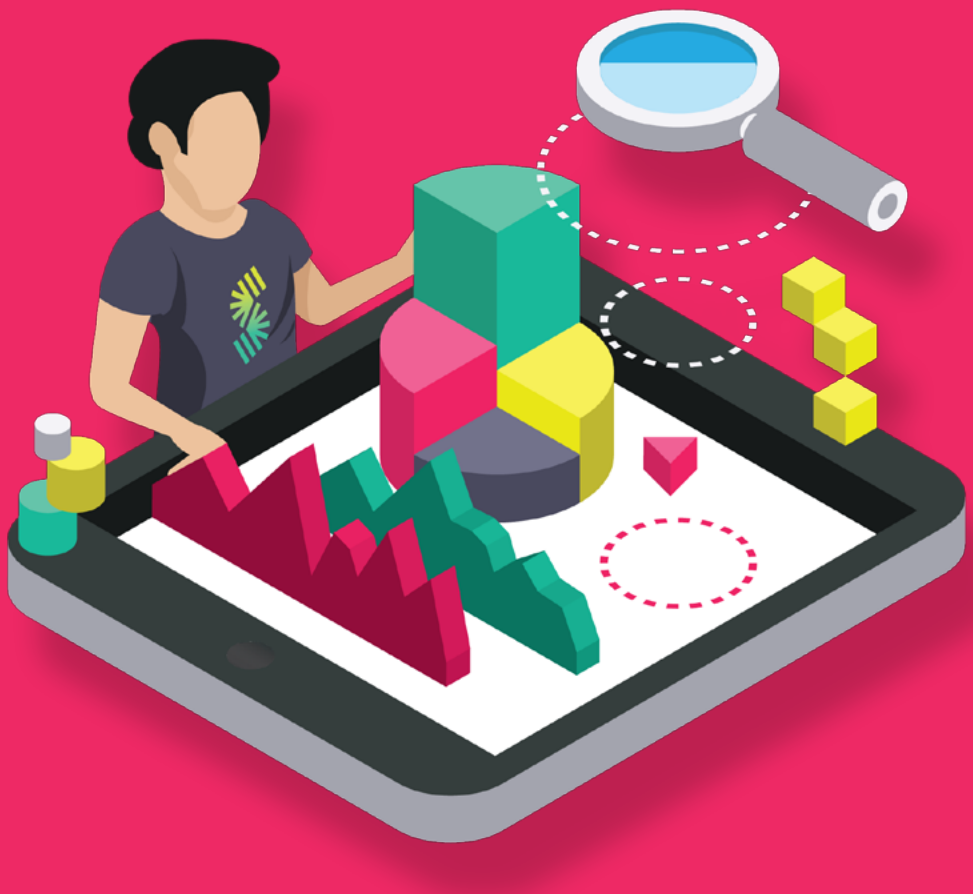
A Beginner's Guide to

Getting Your First Data Science Job



Springboard

Table of contents

Foreword

At the beginning of 2016, Glassdoor, one of the top careers websites in the world, released a report with the best jobs to pursue. Each job is ranked based on a composite score of median reported salary, job openings, and career opportunities. At the top of this list in 2016 was the relatively new profession called Data Scientist.

Such is the pace at which data is proliferating in the world, a phrase that barely existed a decade ago is now one of the most sought-after professions.

The new world economy needs a new approach to skills education. At Springboard, we're building an educational experience that empowers our students to thrive in this new world. Through our online workshops, we have prepared thousands of people for careers in Data Science, with 1-on-1 mentorship from industry experts.

As part of our mission to make high quality education accessible for all, we have created this guide to careers in data science. Through it, our goal is to bring you insight from our network of industry experts to demystify data science careers. Maybe we'll even inspire some of you to pursue a career in this fascinating field.

1 Introduction

1.1 An Unconventional Innovator

GiveDirectly is a not-for-profit organization that shouldn't work. The organization has built its success on giving unconditional cash transfers to the poorest people in the world. Charities aren't supposed to give their recipients unlimited leeway; they're supposed to provide certain goods for certain needs.

GiveDirectly is designed to break all the rules -- and it's working.

The organization's mandate is to transform international giving by attacking extreme poverty at its roots. People who are helped by GiveDirectly decide how to help themselves. This has led to one of the lowest percentages of money spent on administration, and recipients are close to doubling their assets and halving the rate of hunger.

It's hard to overstate how difficult GiveDirectly's mission is. The regions they work in are often neglected and forgotten. They not only have to provide for the poorest communities, they also have to find them.

Since census data is sparse or unreliable at a village level, GiveDirectly would often have to send somebody to manually scour each village for signs of obvious poverty. GiveDirectly representatives look for the presence of metal on home roofing, rather than the more plentiful thatch. People who can afford metal roofs typically buy them.

At a cost of $564 USD in a region where GDP per capita is around $1,700, a metal roof represents a significant capital investment, and the difference between extreme and relative poverty.

1.2 Data Science to the Rescue

Harvard Business Review referred to the role of data scientist as the sexiest career of the 21st century: one where you can earn a healthy salary, and maintain a great work-life balance. According to LinkedIn, Statistical Analysis & Data Mining were the hottest skills that got recruiters' attention in 2014, and Glassdoor ranked data scientist as the #1 job to pursue in 2016.

"The ability to take data—to be able to understand it, process it, to extract value from it, to visualize it, to communicate it —that's going to be a hugely important skill in the next decades."

- Hal Varian, Google's Chief Economist

GiveDirectly is a great example of how organizations win by using data to their advantage.

Sending people to each village would have taken several trips at a crushing expense, creating excessive overhead for an organization looking to operate leanly. To address this issue, a pair of industry experts from IBM and Enigma worked with GiveDirectly to see if data science could help.

Using satellite images provided by Google, they were able to use computers to classify which villages had metal roofs on their houses, and which ones had thatch. They were able to determine which villages needed the most help without sending a single person to the area.

These data scientists were required to collect the massive amounts of satellite photos of the area, something nearly impossible a decade ago. It required implementing machine learning algorithms, a cutting edge technology at the time, to train computers to recognize patterns within those photos.

Ultimately, they were able to pinpoint where GiveDirectly should operate through the type of roofing in the area, saving the organization hundreds of man-hours and allowing them to do what they do best: solve extreme poverty.

2 What is Data Science?

DJ Patil, the current Chief Data Scientist of the White House Office of Science and Technology, and the previous Head of Data Products at LinkedIn, first coined the term data science. However, a decade later, the definition remains contested.

"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."

- DJ Patil, Chief Data Scientist in the US

Despite the mass of data the world generates every day, most organizations are struggling to benefit from it. And according to McKinsey, the US alone faces a shortage of 150,000+ data analysts and an additional 1.5 million data-savvy managers.

Salary trends have followed the impact of data science. With a national average salary of $118,000 (which increases to $126,000 in Silicon Valley), data science has become a lucrative career path where you can solve hard problems and drive social impact.

Since you're reading this guide, you're likely curious about a career in data science, and you've probably heard some of these facts and figures. You likely know that data science is a career where you can do good while doing well.

You're ready to dig beyond the surface, and see real-life examples of data science, and get real-life advice from practitioners in the field.

That's exactly why we wrote this guide--to bring data science careers to life for thousands of data-curious, savvy young professionals. We hope after reading this guide, you have a solid understanding of the data science industry, and know what it takes to get your first data science job. We also want to leave you with a checklist of actionable advice which will help throughout your data science career.

2.1 The Foundations of Data Science

A decade after the term data science was first used, there is continued debate among practitioners and academics about what data science means.

One of the most substantive differences is the amount of data you have to process now, as opposed to a decade ago. In 2020, the world will generate 50x more data than we generated in 2011. 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.

Baseball players used to be judged by how good scouts thought they looked, not how many times they got on base--that was until 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. 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.

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 datasets. 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.

You might not think you know anything about data science, but if you've ever looked for a Wikipedia table to settle a debate with one of your friends, you were doing a little bit of data science.

3 The Different Data Science Roles

Before we dive into what skills you need to become a data scientist, you should be aware that there are different roles in data science. Oftentimes, a data science team will rely on different team members for different skill sets, or the skill set needed may depend on the type of company and part of the organization you work in.

Let's look at some broad categories of roles that are lumped under the umbrella term "Data Science."

Data Scientists

One definition of a data scientist is someone who knows more about programming that a statistician, and more statistics than a software engineer. A data scientist will be able to run with data science projects from end-to-end: they will store and clean large amounts of data, explore data sets to identify potential insights, build predictive models, and weave a story around the findings.

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.

Data scientists are the bridge between programming and algorithmic thinking. A data scientist might use historical data to build a model that predicts the number of credit card defaults in the following month, and use their data engineering skills to implement a simulation of their model on some sample data.

Additionally, within the broad category of data scientists, you may encounter statisticians who focus on statistical approaches to data, and data managers who focus on running data science teams.

Data Analysts and Business Analysts

Data analysts sift through data and provide reports and visualizations to explain what the data can offer. 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 data science job.

Business analysts are adjacent to data analysts, but are more concerned with the business implications of the data. Should the company invest more in project X or project Y? Business analysts will leverage the work of data science teams to visualize and communicate what insight can be gained from the data to answer those questions.

Data Engineers

Data engineers are software engineers who handle large amounts of data, and often lay the groundwork 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 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 (e.g. 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.

Skills

You can roughly distinguish these different roles based on skill set: data scientists rely on their training in statistics and mathematical modeling, data engineers rely mostly on software engineering skills, and business analysts rely more heavily on their analytical skills and domain expertise. This blog post summarizes some of the differences between these roles, but you can be certain the people who occupy these roles will have a variety of skills outside their specialties.

It is important to keep these distinct roles in mind, however, when deciding on your own area of expertise. Data science encompasses many specialty roles, and each role comes with different needs and different salaries.

Salary Ranges

Data scientists need to have the broadest set of skills, covering the theory, implementation, and communication of data science. As such, they also tend to be the highest compensated group with an average salary above $115,000 USD.

Data engineers focus on setting up data systems and making sure code is clean, and technical systems are well-suited to the amount of data passing back and forth for analysis. They tend to be middle of the pack when it comes to compensation, with an average salary around $100,000 USD.

Data analysts often focus on querying information and communicating insight from the data that drives action within the organization. Their average salary is around $65,000 USD, partly because a lot of data analyst roles are filled by entry-level graduates with limited work experience.

To effectively solve a variety of data problems, you will need people within every one of these roles to form a complete data science team.

4 The Data Science Process

At Springboard, our data students often ask us questions like "What does a data scientist do?" or "What does a day in the data science life look like?"

These questions are tricky. The answer can vary by role and company.

So we asked Raj Bandyopadhyay, Springboard's Director of Data Science Education, if he had a better answer.

Turns out, Raj 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.

Raj calls it "the Data Science Process", which he outlines in detail in a short 5-day email course. Here's a summary of his insights.

Step 1: Frame the problem

Before you can solve a problem, you have to define the problem.

You'll often get ambiguous inputs from the people who have problems. You'll have to develop the intuition to translate scarce inputs into actionable outputs - and to ask the questions that nobody else is asking.

Say you're solving a problem for the VP of sales of your company. You should start by understanding their goals and the underlying 'why' behind their data questions. Before you can start thinking of solutions, you'll want to work with them to clearly define the problem.

To define the problem, you need to ask the right questions. For example,

1. Who are the customers?
2. Why are they buying our product?
3. How do we predict if a customer is going to buy our product?
4. What is different from segments who are performing well and those that are performing below expectations?
5. How much money will we lose if we don't actively sell the product to these groups?

You need as much context as possible for your numbers to become insights.

In response to your questions, the VP of Sales might reveal that they want to understand why certain segments of customers bought less than expected. Their end goal might be to determine whether to continue to invest in these segments, or deprioritize them. You'll want to tailor your analysis to that problem, and unearth insights that can support either conclusion.

It's important that at the end of this stage, you have all of the information and context you need to solve this problem.

Step 2: Collect the raw data needed for your problem

Once you've defined the problem, you'll need data to give you the insight needed to develop a solution. This part of the process involves thinking through what data you'll need, and finding ways to get that data, whether it's

querying internal databases or purchasing external datasets.

You might find out that your company stores all their sales data in a customer relationship management (CRM) software platform. You can export the CRM data in a CSV file for further analysis.

Step 3: Process the data for analysis

After you've collected all the raw data, you'll need to process it before you can do any analysis. Oftentimes, data can be messy, especially if it hasn't been well-maintained. You'll see errors that will corrupt your analysis: values set to null though they are actually zero, duplicate values, and missing values. It's up to you to go through and check your data and make sure you'll get accurate insights.

You'll want to check for the following common errors:

1. Missing values
2. Corrupted values
3. Timezone differences
4. Date range errors, such as data registered from before sales started

You'll need to look through aggregates of your file rows and columns, and sample some test values to see if your values make sense. If you detect something that doesn't make sense, you'll need to remove that data or replace it with a default value. You'll need to use your intuition here: if a customer doesn't have an initial contact date, does it make sense to say there was *no* initial contact date? Or do you have to hunt down the VP of Sales and ask if anybody has data on the customer's missing initial contact dates?

Once you're done working with those questions and cleaning your data, you'll be ready for exploratory data analysis (EDA).

Step 4: Explore the data

When your data is clean, you should start playing with it.

The difficulty here isn't coming up with ideas to test, it's coming up with ideas that are likely to turn into useful insight. You'll have a fixed deadline for your data science project (your VP of Sales is probably waiting on your analysis), so you'll have to prioritize your questions.

You should look for interesting patterns that explain why sales are reduced for the segment of the populations you've identified as the problem. You may notice they're not very active on social media, with few having Twitter or

Facebook accounts. You may also notice that most people in this segment are older than your general audience. At this point, you can now begin to trace these patterns to analyze the data more deeply.

Step 5: Perform in-depth analysis

This step of the process is where you will need to apply your statistical, mathematical and technological knowledge, and leverage all the data science tools at your disposal to crunch the data and find every insight you can.

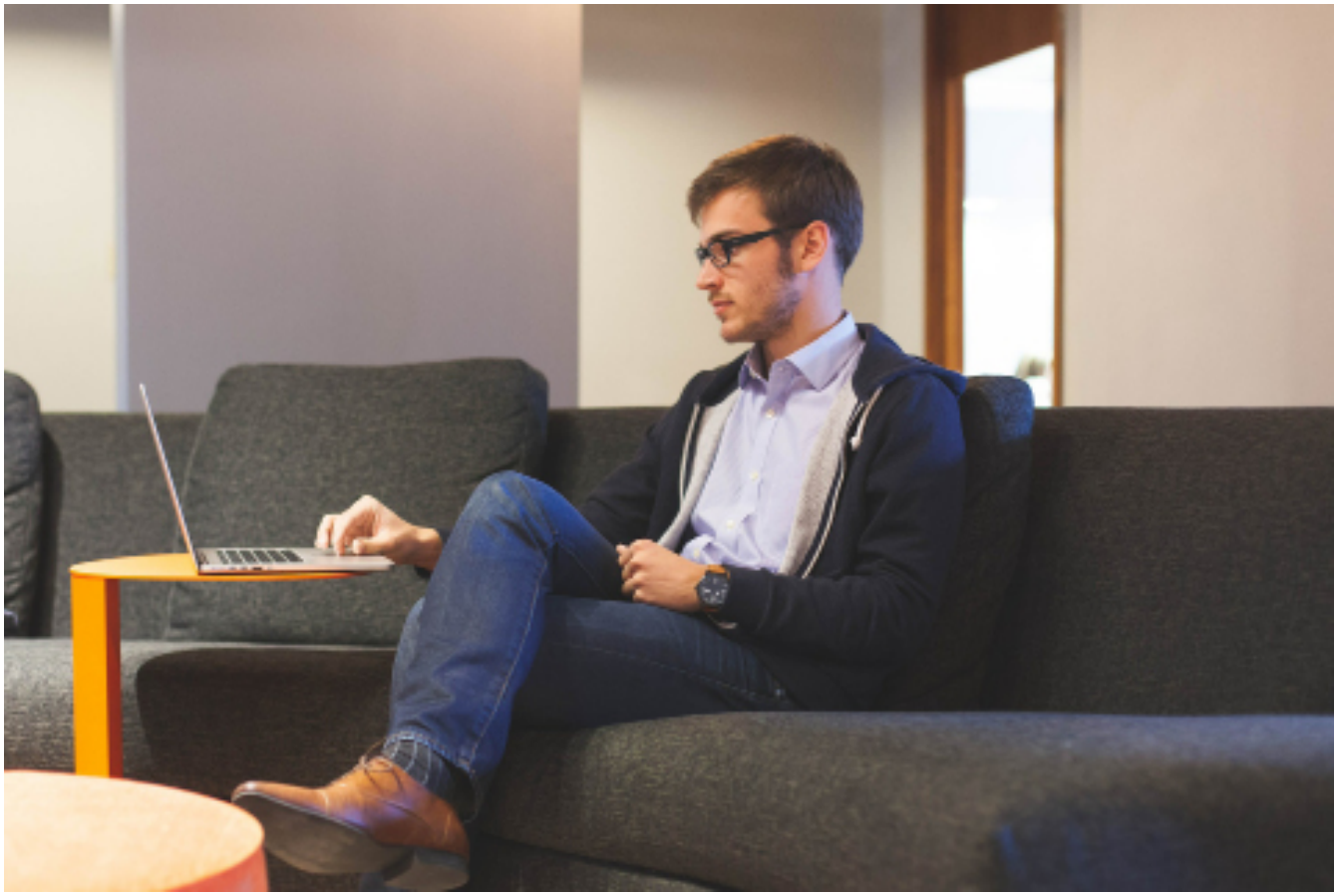
In this example, you may have to create a predictive model that compares your under-performing group with your average customer. You may find that age and social media activity are significant factors in predicting who will buy the product.

If you asked enough of the right questions while framing your problem, you might realize that the company has been concentrating heavily on social media marketing efforts, with messaging aimed at younger audiences. You would also know that certain demographics prefer being reached by telephone rather than by social media.

You will begin to see the way the product has been marketed is significantly affecting sales. However, maybe this under-performing group isn't a lost cause: a change in tactics from social media marketing to more in-person interactions 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 results of the analysis



It's important that the VP of Sales understands why the insights you've uncovered are important. Ultimately, you've been called upon to create a solution throughout the data science process. Your ability to properly communicate your results will define the difference between action and inaction on your proposals.

You need to craft a compelling story that ties in your data with their knowledge. You start by explaining the reasons behind the under-performance of the older demographic. And alongside answers your VP of Sales gave you and the insights you've uncovered from the data, you can move to concrete solutions that address the problem. One solution would be to shift some resources from social media to personal calls.

You tie it all together into a narrative that solves the problem for your VP of Sales: she now has clarity for how she can reclaim sales and hit her objectives.

She is ready to act on your proposals.

***

As a data scientist, you'll have to learn how to work through the entire data science process. Here's what that looks like from a day to day perspective.

5 Data Scientists in Action

We have a lot of mentors at Springboard who have shared their stories about the day-to-day happenings of data science. They're all practitioners in the field with real-life experience. Understanding what they do is the first step to fully understanding data science.

5.1 Day in the Life of a Data Scientist

This story is based on the day-to-day of an industry expert in the financial sector, who wishes to remain anonymous.

Data scientists in finance try to predict whether or not people will default on their credit due to certain predictive factors. They help classify which transactions seem fraudulent. All of this requires a look at millions of lines of data, and it involves extrapolation to the future, a skill set almost all human beings are notoriously bad at. However, the day-to-day isn't just spent looking through numbers.

9 am

There's a lot of legwork that goes into data science, like any other job. Nearly an hour is spent catching up on email and organizing for the day ahead.

10 am

A significant amount of time in data science is spent recruiting. Demand for data science skills is at an all-time high, so data science organizations are often evaluating potential recruits. Data scientists will often take time out of their day to do phone screens of potential new team members.

11 am

Data scientists spend a lot of time in meetings. Almost an hour is spent making sure every team is properly aligned with one another, and working on the right projects.

12 pm

Lunch offers the chance to relax a bit and catch up with colleagues, then it's back to the grind. One half of a typical day is spent coding an analysis or looking over someone else's code. This might involve building a graph to represent insights unearthed during a look through the data, or it might be about making sure your own code is clean so everybody on your team can read through it and understand what is going on.

4 pm

Data scientists will often discuss with groups of fellow data scientists ways in which they can collaborate and help one another. They'll often learn together and share the latest tool that can help improve productivity.

5.2 Infusing Data in Your Workplace: Chase Lehrman

Chase Lehrman works as a data analyst at a fast-growing education company called Higher Learning Technologies that helps dental and nursing students pass their board exams. He describes his day-to-day as being a data storyteller who looks to gain an understanding of how students are using the product Higher Learning Technologies sells. He also helps people across the organization get the data they need to make informed decisions: a recent example involved sizing a market.

Thanks to Chase, Higher Learning Technologies can change its static data into usable insights, something every data scientist should get their organization to embrace. Chase makes sure that data problems are framed the right way and that solutions are properly communicated and actionable.

Data scientists solve many different problems. A data scientist may hunt for raw data, create automated programs to process data quickly and efficiently, or communicate the impact of their results to the CEO of a company. You will have to learn a variety of tools and maintain a versatile skill set if you want to become an effective data scientist.

5.3 Understanding the Data: Sneha Runwal

Sneha Runwal works as a statistician at Apple, where she works in the AppleCare division. Her major work involves forecasting and time series analysis, in addition to anomaly detection.

Sneha feels that people are often too quick to delve into algorithms and computer code. You have to remember it's important to step back and understand your data before you get into implementation mode; even Sneha is trying to get more disciplined about this. Her advice? Understand as much of your data as possible, as early as you can.

6 What You Need to Learn to Become a Data Scientist

6.1 Introduction

This next section covers all of the data science skills you'll need to learn. You'll also learn about the tools you need to do your job.

Most data scientists use a combination of skills every day, some of which they have taught themselves on the job or otherwise. They also come from various backgrounds.

> There isn't any one specific academic credential that is required to be an effective data scientist.

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

6.2 Data Science Skills

6.2.1 An Analytical Mind

You'll need an analytical mindset to do well in data science.

Data science involves solving problems. You'll have to be adept at approaching those problems analytically, and methodically applying logic to solve them.

6.2.2 Mathematics

Make sure you know the basics of university math, from calculus to linear algebra. The more math you know, the better.

Mathematics is an important part of data science.

When data sets get large, it often gets unwieldy. You'll have to use mathematics to process and structure the data you're dealing with.

You won't be able to get away from knowing calculus and linear algebra. You'll need to know how to manipulate matrices of data and possess a general understanding of the math behind the algorithms.

6.2.3 Statistics

You must know statistics to infer insights from smaller data sets onto larger populations. This is a fundamental law of data science.

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. Understanding inferential statistics allows you to make general conclusions about an entire population from a smaller sample.

To understand data science, you must also know the basics of hypothesis-testing and experiment-design to comprehend the meaning and context of your data.

6.2.4 Algorithms

Algorithms are the ability 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.

In order for you to do any heavy lifting in data science, you'll have to understand the theory behind algorithm selection and optimization. You'll have to decide whether or not your problem demands a regression analysis, or an algorithm that helps classify different data points into defined categories.

You'll want to know many different algorithms, and you'll want to learn the fundamentals of machine learning. Machine learning is what allows for Amazon to recommend you products based on your purchase history without any direct human interventions. It is a set 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.

6.2.5 Data Visualization

Finishing your data analysis is only half the battle.

To drive impact, you will have to convince others to believe in, and adopt, your insights.

Human beings are visual creatures. According to 3M and Zabisco, almost 90% of the information transmitted to your brain is visual in nature, and visuals are processed 60,000 times faster than text.

Human beings have been wired to respond to visual cues. You'll need to find a way to convey your insights accordingly.

6.2.6 Business Knowledge

Data means little without its context. You have to understand the business you're analyzing.

Most companies depend on their data scientist not only to mine data sets, but also to communicate their results to various stakeholders and present recommendations that can be acted upon.

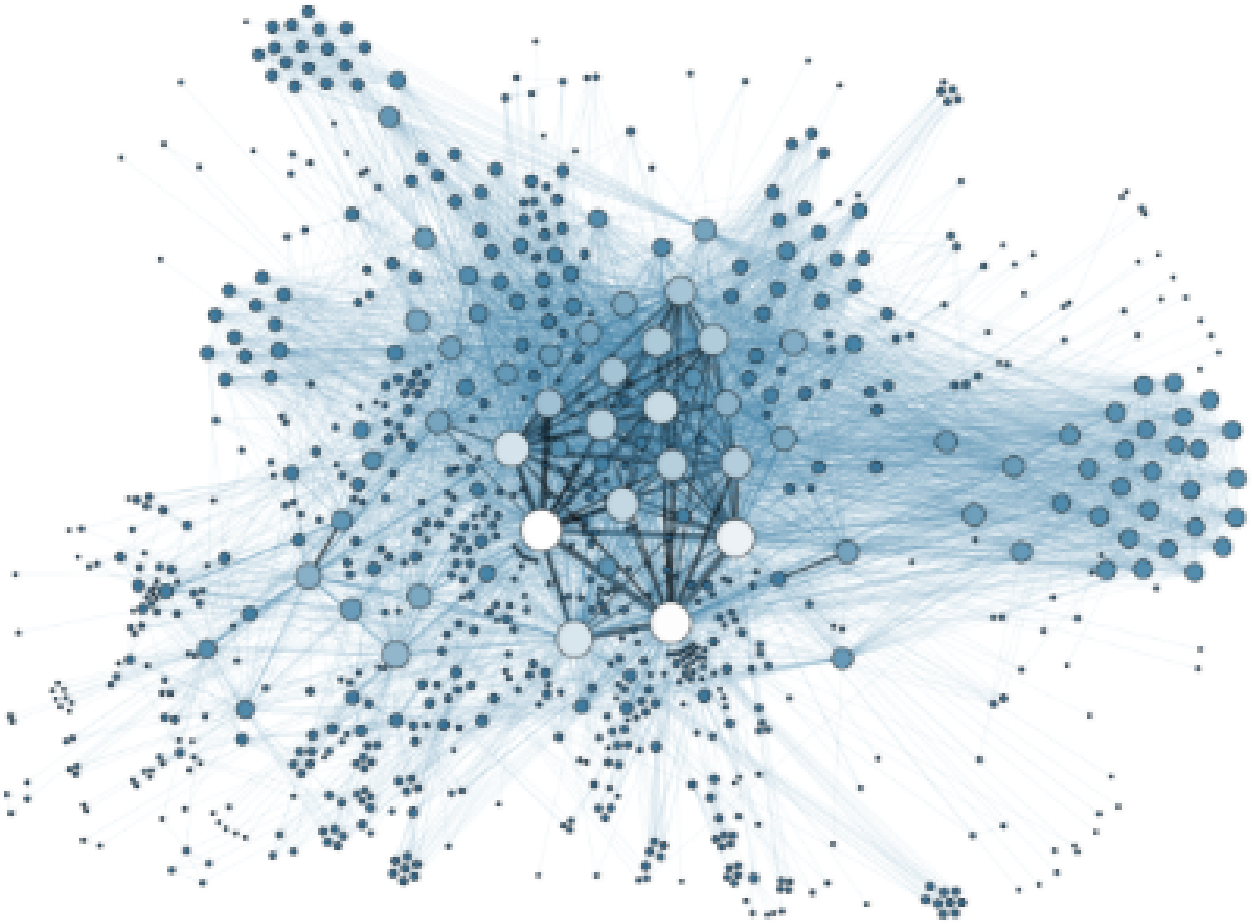
The best data scientists not only have the ability to work with large, complex data sets-- they also understand the intricacies of the business or organization they work for.

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

6.2.7 Domain Expertise

As a data scientist, you should know the business you work for and the industry it lives in.

Beyond having deep knowledge of the company you work for, you'll also have to understand its field for your insights to make sense. Data from a biology study can have a drastically different context than data from a psychology study. You should know enough to cut through the industry jargon.



6.3 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 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 how to utilize them to their full potential.

6.3.1 File Formats

Data can be stored in many different file formats.

Here are some of the most common:

CSV

Comma separated values. You may have opened this sort of file with Excel. CSVs separate out data with a delimiter, a piece of punctuation that serves to separate out different data points.

SQL

Structured query language stores data in relational tables. If you go from the right of a column to the left, you'll get different data points on the same entity (e.g. a person will have a value in the AGE, GENDER, and HEIGHT columns associated with them).

JSON

Javascript Object Notation is a lightweight data exchange format that is both human and machine-readable. Data from a web server is often transmitted in this format.

6.3.2 Excel

Excel is often the gateway to data science, and every data scientist can benefit from learning it.

Introduction to Excel

Excel allows easily manipulation of data with what is essentially a 'What You See Is What You Get' editor that allows you to 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 in analystics 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 dataset on the statistics of NBA players and making a simple graph of the top scorers in the league.

6.3.3 SQL

SQL is the most popular programming language used to find data.

Introduction to SQL

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

Benefits of SQL

SQL is the most popular tool 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.

Who Uses This

Data analysts and some data engineers tend to use SQL.

Level of Difficulty

Beginner.

Sample Project

Using an SQL query to select the top ten most popular songs from an SQL database of the Billboard 100.

6.3.4 Python

Python is a powerful and versatile programming language for data science.

Introduction to Python

Once you download Anaconda, an environment manager for Python, and get set up on iPython Notebook, you'll quickly realize how intuitive Python is. A versatile programming language built 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.

Python is the most popular programming language taught in universities, and the community of Python programmers is only going to be larger in the years to come. 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: 40% of respondents to a data science survey conducted by O'Reilly used Python, which was more than the 36% who used Excel.

Who Uses This

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

Level of Difficulty

Intermediate.

Sample Project

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

6.3.5 R

R is a staple in the data science community because it is designed explicitly for data science needs.

Introduction to R

R is a programming environment designed for data analysis. R shines when it comes to building statistical models and displaying the results.

Benefits of R

R is slightly more popular than Python in data science, with 43% of data scientists using it in their tool stack, compared to the 40% who use Python.

It is an environment where a wide variety of statistical and graphing techniques can be applied.

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

Who Uses This

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

Level of Difficulty

Intermediate.

Sample Project

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

6.3.6 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, every two years. Imagine how much data Facebook has at any given time!

Any data set too large for conventional data tools, such as SQL and Excel, can be considered big data, according to McKinsey. Simply defined, big data is data which cannot fit onto one computer.

Here are tools to solve the big data problem:

6.3.7 Hadoop

By using Hadoop, you can store your data on multiple servers while controlling it from one.

Introduction to Hadoop

The solution to dealing with massive data sets is a technology called MapReduce. MapReduce is an elegant abstraction that treats a series of computers as if it were one central server.

Benefits of Hadoop

Hadoop is an open-source ecosystem of tools that allow you to MapReduce your data and store enormous datasets 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 datasets which update in real time, such as the number of likes Facebook users generate.

6.3.8 NoSQL

NoSQL allows you to manage data without unnecessary weight.

Introduction to NoSQL

Tables that bring all their data with them can be cumbersome. NoSQL includes a host of data storage solutions to separate out huge data sets into smaller, more manageable segments.

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 which can be manipulated like SQL tables, but store the data with less structure and density.

Who Uses This

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

Level of Difficulty

Advanced.

Sample Project

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

6.4 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.

	Excel	SQL	Python	R	Hadoop	NoSQL
Collect Data		X	X	X		
Process Data	X	X	X	X		X
Explore Data	X		X	X	X	X
Analyze Data	X		X	X	X	X
Communi-cate Data	X		X	X		

6.4.1 Collect Data

Sometimes doing the data analysis isn't the hard part; it's finding the data you need. Thankfully, there are many resources.

You can generate datasets by using data in an application programming interface (API), allowing you to take structured data from certain providers. You'll be able to query all kinds of data from Twitter, Facebook, and Instagram.

If you want to play around with public datasets, the United States government has made some free. And the most popular datasets are tracked on Reddit.

Dataset search engines, such as Quandl, allow you to search for the perfect dataset.

Springboard has compiled 19 of our favorite public datasets on our blog to help you out in case you ever need good data right away.

***

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

You can also create datasets from the web. The Python requests library scrapes data from different websites with a line of code. You'll be able to take data from Wikipedia tables, for example, clean the data with the beautifulsoup library, and then perform an in-depth analysis your new dataset.

R can take data from Excel, CSV, and from text files, and 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 beautifulsoup libraries in Python.

6.4.2 Process Data

Excel allows you to easily clean data with menu functions to clean 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.

Python uses the Pandas library for data analysis. It is much quicker to process larger data sets in Pandas than in Excel, and Pandas has more functionality. You can, for example, replace every error value in a dataset with a default value such as zero in one line of Pandas code.

R can help you add columns of information, reshape, and transform the data. Many of the newer R libraries, such as reshape2, allow you to play with different data frames and make them fit the criterion you've set.

NoSQL allows you the ability to subset large data sets and to change data according to your will, which you can use to clean your data.

6.4.3 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.

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 as SQL.

6.4.4 Analyze Data

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

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

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

6.4.5 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 Matplot-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 porfolio.

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 the 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, such as jpg, or save them as separate PDFs. You can use ggplot2 for more advanced plots such as complex scatter plots with regression lines.

Python vs R

The data science community tends to use either Python or R. Here are some of the differences.

USAGE Python, as we noted above, is often used by computer programmers since it is the Swiss knife of programming languages, versatile enough so that you can build websites and do data analysis at the same time. R is primarily used by researchers.

SYNTAX Python has a nice clear "English-like" syntax that makes debugging and understanding code easier, while R has unconventional syntax that can be tricky to understand, especially if you have learned another programming language.

LEARNING CURVE R is slightly harder to pick up, especially since it doesn't follow the normal conventions other common programming languages have. Python is simple enough that it makes for a really good first programming language to learn.

POPULARITY Python has always been among the top 5 most popular programming languages on Github, a common repository of code that often tracks usage habits across all programmers quite accurately, while R typically hovers below the top 10.

FOCUS ON DATA SCIENCE Python is a general-purpose language, and there is less focus on data analysis packages then in R. Nevertheless, there are very cool options for Python such as Pandas, a data analysis library built just for it.

SALARY The average data scientist who uses R will receive a salary of $115k compared to the $95k average they would earn with Python.

Conclusion

Python is versatile, simple, and easy to learn. Python is powerful because of its usefulness in a variety of contexts, including those outside of data science. R is a specialized environment optimized for data analysis, but it is harder to learn than Python. However, you'll get paid more if you stick it out with R rather than strictly working with Python.

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 improve you as a data scientist. Approximately 23% of data scientists surveyed by DataCamp 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 R may be paid more than those who work in Python, data scientists who used 15 or more tools made $30,000 more than those who used 10 -14 tools.

As a data scientist, you'll often be called upon to do different tasks, and you'll need to know exactly which tool is best for each task.

7 Starting Your Job Search

7.1 How to Build a Data Science Portfolio and Resume

You need to make a great first impression to break into data science. This starts with your portfolio and your resume. Many data scientists have their own website which serves as both a repository of their work and a blog.

Your portfolio and resume will allow you to demonstrate your experience and your value in the data science community. In order for your portfolio to have the maximum 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 dilute projects.
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 everyone? Perhaps you have a tutorial you wrote? Showcase them here.
4. It should be easy to find your contact information.

This data portfolio of Trent Salazar exemplifies these four traits. Trent is a research assistant at Duke University who has had several analyst roles in investment banking. Impressively, when you google "Data Science Portfolio," his portfolio is one of the top results.

Here's how he ranks so high:

1. The website design is well-thought out: it doesn't look like a CV, it looks like a storybook. Solutions like Themeforest can help if you don't have the design skills.
2. Trent's portfolio tells a story of who he is, and places his work in its rightful context. You can see his interest in financial modeling and how his interest applies to his career.
3. He has a lot of his resources on his website. He's adding value and building a stronger personal brand both as a data scientist, and as a professional.
4. The website makes it easy to contact Trent, either by email, or through any one of his social channels.

Now take a look at our mentor Sundeep Pattem's personal portfolio for example projects. He's worked on complex data problems that resonate in the real world. He has five projects dealing with healthcare costs, labor markets, energy sustainability, online education, and world economies: fields where there are plenty of data problems to solve.

These projects are independent of any workplace. They show Sundeep innately enjoys creating solutions to complex problems with data science.

***

If you're short on project ideas, you can participate in data science competitions. Platforms like Datakind, Kaggle and Datadriven allow you to work with real corporate or social problems. By using your data science skills, you can use your ability to make a difference, and create the strongest portfolio asset of all: a demonstrated bias to action.

7.2 How to Network and Build a Personal Brand in Data Science

Once you have learned the 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. As we saw with Sundeep's example, solving challenging real-world problems will enable you to build a portfolio and a personal brand, ultimately helping you to land a job.

7.3 Finding a Mentor

One of the highest-value networking activities you can pursue is finding a mentor who can guide you along your 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!

You should remember that mentorship is a two-way street, and you can also generate 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 relationship where you can build value for yourself in a professional context. 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, which leads to significantly better outcomes through increased accountability and motivation.

7.4 Meetups and Conferences

In this section, we're listing some of the popular events and conferences. With a bit of searching, you can find great data science events in your area. These are great places to meet fellow aspiring data scientists, build connections with established scientists, and pick up the jargon.

At events and meetups, you'll network with fellow data scientists and unearth hidden job opportunities.

7.4.1 Conferences

Strata Conference

The Strata Conference is a big data science conference that takes place worldwide in different cities. 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.

Knowledge Discovery in Data Science (KDD)

Knowledge Discovery in Data Science is another large data science conference. This organization seeks to lead discussion and teaching of the science behind data science. Membership and attendance at these conferences offers an excellent way to contribute to growing trends in data science.

Neural Information Processing Systems (NIPS)

Neural Information Processing Systems 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 shake data science in the future.

7.4.2 Meetups.

We've listed the major conferences where the data science community assembles, but there are often smaller meetups connecting 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 America. 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 groups with influence 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: an invited speaker gives a talk, followed by a networking period where everyone is encouraged to mingle (usually over beers). The general data science meetups will often have an industry talk where someone will delve into a real-world data science problem and explain how it was solved. Specialized data science meetups, such as Python groups or R groups, will often focus on technical tutorials, teaching a specific tool or skill.

You should introduce yourself to the local data science community. Many of the best career opportunities are found by talking to people passionate about a certain field, many of whom will be with you at a data science meetup.

7.4.3 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'll often find data science podcasts to follow.

Podcasts such as the Talking Machines features interviews of prominent data scientists. Partially Derivative offers drunk data-driven conversations. 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, to 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 great programming resources and pieces on Hacker News.

7.5 Job Boards for Data Science

1. Kaggle offers a job board for data scientists.
2. You can find a list of open data scientist jobs at Indeed, the search engine for jobs.
3. Datajobs offers a listings site for data science.
4. Datasciencejobs scrapes data science jobs from around the web into one centralized location.

You can also find opportunities through networking and through your 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 exclusive to Y Combinator startups. Y Combinator is the most prestigious startup accelerator in the world. Angellist is a database for startups looking for funding, and it also has a jobs section.

7.6 Ace the Data Science Interview

An entire book can be written on the data science interview--in fact, it's likely we'll be releasing a book exclusively on the topic soon!

If you get an interview, what do you do next? First, acknowledge that a data science interview involves some degree of preparation. You should always anticipate a mixture of technical and non-technical questions in any data science interview. Prepare for questions about your background, coding, applied machine learning, and be ready to analyze data sets.

There are several kinds of questions that are always asked:

1. Questions centered around your programming knowledge.
2. Questions that test what you know about data science algorithms, and make you share your real-life experience.
3. Questions focused on your prior work with data science.
4. Questions testing your knowledge of the interviewer's business.

Make sure you brush up on your programming and data science--and try to interweave it with your own personal story. To prepare for the coding questions, you'll have to treat interviews on data science partly as a software engineering exercise. You should brush up on all coding interview resources, many of which are found online.

Here is a list of specific data science questions you might encounter, including the following:

- Python vs R: which language do you prefer for [x] situation?
- What is K? Describe when you would use it.
- Tell me a bit about the last data science project you worked on.
- What do you know about the key growth drivers for our business?

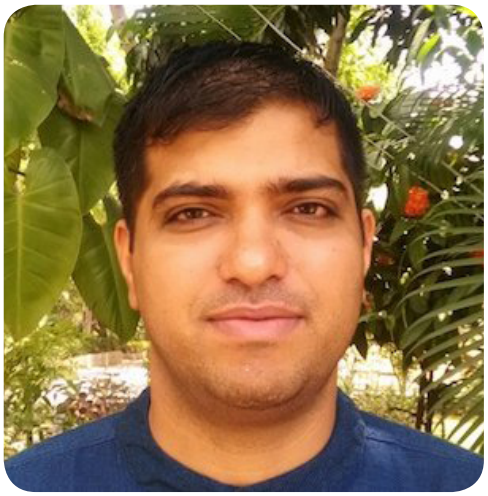
If you can demonstrate how your data science work can help move the needle for your potential employers, you'll impress them. They'll know you are someone who cares enough to learn about what they're doing, and who already knows a lot about the industry, even prior to them having to provide a detailed education.

8 Paths into Data Science

You might wonder where you'll find people who work in data science. What background do they come from? What skills do they have to pick up? Who are these people and how can I become one of them them?

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

8.1 Engineering & Business = Data Stories - Amit Kapoor



You can become a great data storyteller without a computer science degree.

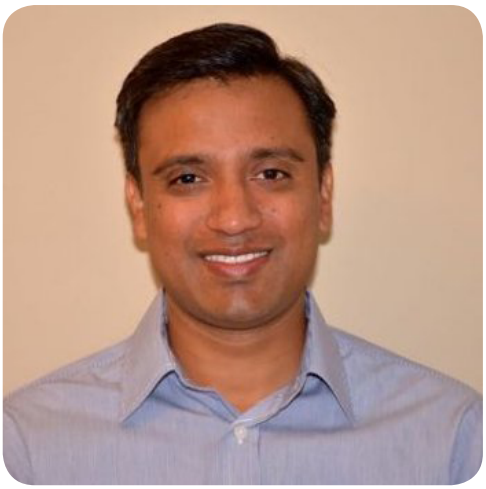
Amit Kapoor is the founding partner of NarrativeViz Consulting, an agency that helps clients with data visualization problems. He describes himself as a visual storyteller, somebody who can convey information using cold hard numbers turned into appealing visual stories.

Amit first got a degree in mechanical engineering, and went on to get an MBA. He then worked as a consultant for A.T Kearney, and then Booz & Company. It was during his time as a consultant that he realized the importance of being able to communicate data properly. He leveraged his background in both engineering and business to transition into managing his own data visualization consultancy.

Amit's background proves you don't have to learn computer science to do wonderful things with data.

8.2 Drive Real-World Impact to Get Into Data Science - Sundeep Pattem

Solve hard real-world problems to break into a data science career.



Sundeep Pattem is a data innovation leader at the California Department of Justice. He's mentored for several data science courses, and as a data scientist he works on creating end-to-end solutions to extract value from data. He has his own personal websites with different data science projects.

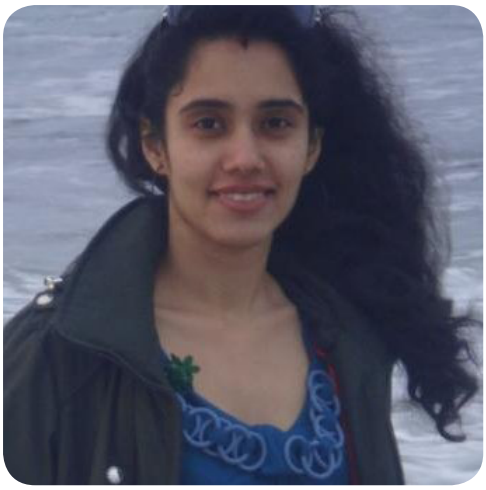
He comes from a traditional data science background with a PhD in electrical engineering and a job at Cisco as a software engineer--mostly because he didn't feel like there were many opportunities in electrical engineering.

It was the Machine Learning course at Coursera, however, that truly inspired him to do what he loves now: data science. After getting an initial spark of inspiration from online learning. Sundeep started attending data science meetups and interviewing for different positions without much success.

His breakthrough came when he found an unsolved problem in energy sustainability, and worked to solve it. He was soon a published author at a prestigious academic conference, and shortly thereafter, he was hired to become a practicing data scientist.

Sundeep's path shows that if you work on hard, real-world problems outside of work, you'll drive social impact and find data science jobs waiting for you.

8.3 Gaining Data Science Experience - Sneha Runwal



You can come from a computer science background and gain valuable data science experience by finding the right jobs.

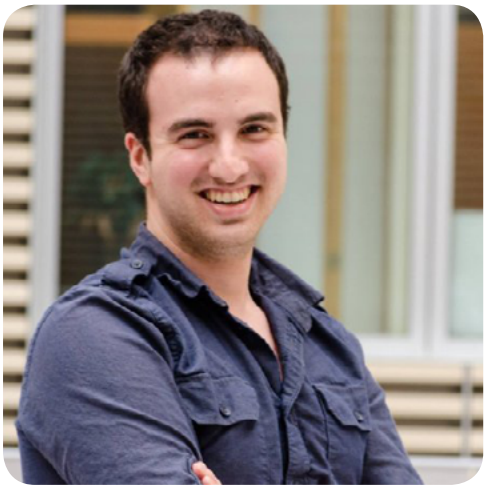
Sneha graduated with a bachelor's degree in computer science, and she did a short stint at Cisco as a software engineer. After pursuing her MBA with a focus in analytics and strategy, she took up several data science internships.

After working extensively with Infosys on their HR analytics data, she realized this was something she wanted to pursue.

Sneha moved to the Bay Area and took a data science internship with a startup in the area. After working on their psychometrics data to determine what candiddates would be a good fit for certain jobs, she transitioned to her current role as a Statistician at Apple.

Sneha's path shows that you can enter data science with a computer science degree, and a few great data science experiences.

8.4 Competing to Get Into Data Science - Sinan Ozdemir



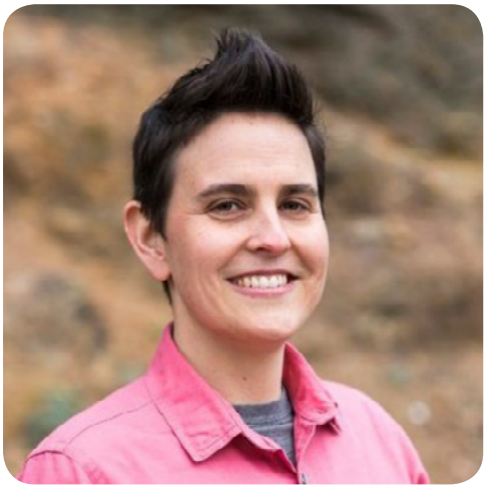
You can participate in online competitions to practice and demonstrate your data science skills.

Sinan Ozdemir followed a slightly unconventional path to a data science career. He got a Master's in Theoretical Mathematics, and then became a lecturer at Johns Hopkins University in Business Intelligence. It was there, he became fascinated by data science competitions.

Becoming a regular at Kaggle, an online platform for data science competitions, he soon demonstrated an extraordinary capacity at creating accurate predictive models. Soon, he was working for data science startups and teaching data science to others.

Sinan's path shows that you can come from an academic background and compete your way to a job.

8.5 A Psychology PhD on the Path to Data Science – Erin Baker



You can unearth insights from many different fields in data science.

Erin is a social psychologist by training who was focused on her PhD on how to get people to be more pro-social. She wanted to get more people engaged in volunteering with their community.

It was during an internship at Hewlett Packard where she learned about data science analytics, which propelled her into a hybrid role of using quantitative methods in data to examine the why behind human behavior.

Her ultimate motivation behind moving into data science after grad school came from the idea that a lot of academia was focused on understanding the theory of human behavior. Erin wanted to take that theory and apply it to real-world situations.

9 Final Advice

On the work of data science: "As a data scientist [...], it's important to recognize that the solution may not be something that you already know or something that just fits nicely with the problem." - Raj B., Director of Data Education at Springboard

On the data science process: "Acquiring and cleaning data takes about 75% of the time in a project." - Amit K., NarrativeViz Consulting

On what interviewers are looking for when they hire: "When I'm looking for a candidate, the first thing that I want to understand is, what is their thought process?" - Sneha R., Statistician at Apple

10 Checklist

- [ ] 1. Assemble a learning plan. Springboard has one for you
- [ ] 2. Brush up on linear algebra and calculus
- [ ] 3. Learn the theory of statistics and machine learning
- [ ] 4. Install Python, play around with it
- [ ] 5. Install R, play around with it
- [ ] 6. Do a few SQL queries
- [ ] 7. Assemble your own data set from a website
- [ ] 8. Register for a data science community online such as KDNuggets
- [ ] 9. Attend a data science meetup
- [ ] 10. Network with at least five data scientists
- [ ] 11. Look for a data science mentor
- [ ] 12. Build a portfolio filled with your data science projects!

11 Conclusion

Getting your first data science job might be challenging, but it's possible to achieve this goal with a diligent approach to picking up skills, working on projects, building a portfolio, and getting 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.

At Springboard, we hope this is the beginning of your adventure, and we hope it ends with the data science job you desire.

If you thought this guide was valuable, share a free copy with your friends and co-workers on Facebook and LinkedIn.

12 Resources

12.1 What is Data Science?

This article on KDNuggets visually shows the difference between data science roles.

12.2 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 for data science 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 basics. This tutorial goes into the exact steps you'll need to perform to get clean data in R.

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

12.3 Getting Data

This Quora thread goes over many of your options for getting public data sets.

12.4 Algorithms

The top ten data algorithms you'll have to use can be quite complex (and there are many more algorithms, but this blog post will explain most of them in plain English. You'll be able to understand all of the options you have around you when you're confronted with a data problem.

12.5 Machine Learning

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

12.6 Data Visualization

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

12.7 Data Science Interview

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

12.8 Building a Data Science Portfolio

This piece gets into how you should build great data products that resonate with users.

13 Stuff Data Scientists Say

Now that you've been flooded with a wealth of resources to look over and the tools you need to understand data science, it's important to take a step back. This is a brief glossary of terms we covered before, and some we haven't, to ensure you never get lost in any data science discussion.

Algorithm

A set of instructions a computer must follow, typically implemented in computer code such as Python or SQL.

API

Application Programming Interface. A set of standards for collecting data from different web sources.

Bit

The basic unit of computer data, which can either be a 0 or 1 value. It's a shortened version of a binary unit.

Byte

A byte is composed of 8 bits and is the second smallest unit of computer data. Historically, it is the amount of data required to encode a character in the computer.

Kilobyte

A kilobyte is 1024 bytes.

Megabyte

A megabyte is 1024 kilobytes.

Gigabyte

A gigabyte is 1024 megabytes.

Terabyte

A terabyte is 1024 gigabytes.

Petabyte

A petabyte is 1024 terabytes.

Data wrangling

Cleaning up your raw data so you can perform analysis on it by, for example, adding new columns of data, or transforming certain columns of data. An example of this would be replacing all error values (such as NaN) with 0.

Big data

Defined in many ways. The simplest explanation is an amount of data that conventional computing methods, such as SQL or Excel, cannot process.

Feature engineering

When you define the independent variables you want to dive deeper into for your data analysis.

Hadoop

An open source framework that allows for distributed data analysis across multiple hardware components; commonly associated with big

data and for the analysis of large data sets supported by the Apache Foundation.

Hive

An SQL-like query language that allows you to access big data records.

Latency

The amount of time it takes to deliver data from one point to another.

Python

Versatile programming language often used by data scientists.

Pandas

A library of code you can access in Python to simplify data analysis.

R

Programming language designed for statistical analysis.

Scalability

The ability of a system to maintain performance as its workload (e.g. the volume of data) increases.

Schema

A set of rules to define how data is organized in a database.

SQL

Structured Query Language. Programming language specifically designed to get data out of relational databases, which have tables of data with columns related to one another.

NoSQL

A set of data storage languages that are not SQL. Created by companies such as Google to deal with scaling issues when it comes to tables of relational data. Typically deals with JSON, a data format popular with web developers.

Machine Learning

Using data-driven algorithms to direct machines to identify certain features in data, thus allowing the machine to learn and detect patterns.

Overfitting

The cardinal sin of machine learning and statistical analysis. This is when you take random variations in the data and overstate their importance in your predictive model, which can generate wildly inaccurate results.

Supervised Learning

Using human-labeled inputs to get machine outputs. An example of this is a program that classifies faces based on a dataset of faces already labeled by humans.

Unsupervised Learning

This is letting the machine classify features without any human inputs. An example of this is a program that can classify faces from pictures of food without any human labeling of the data.