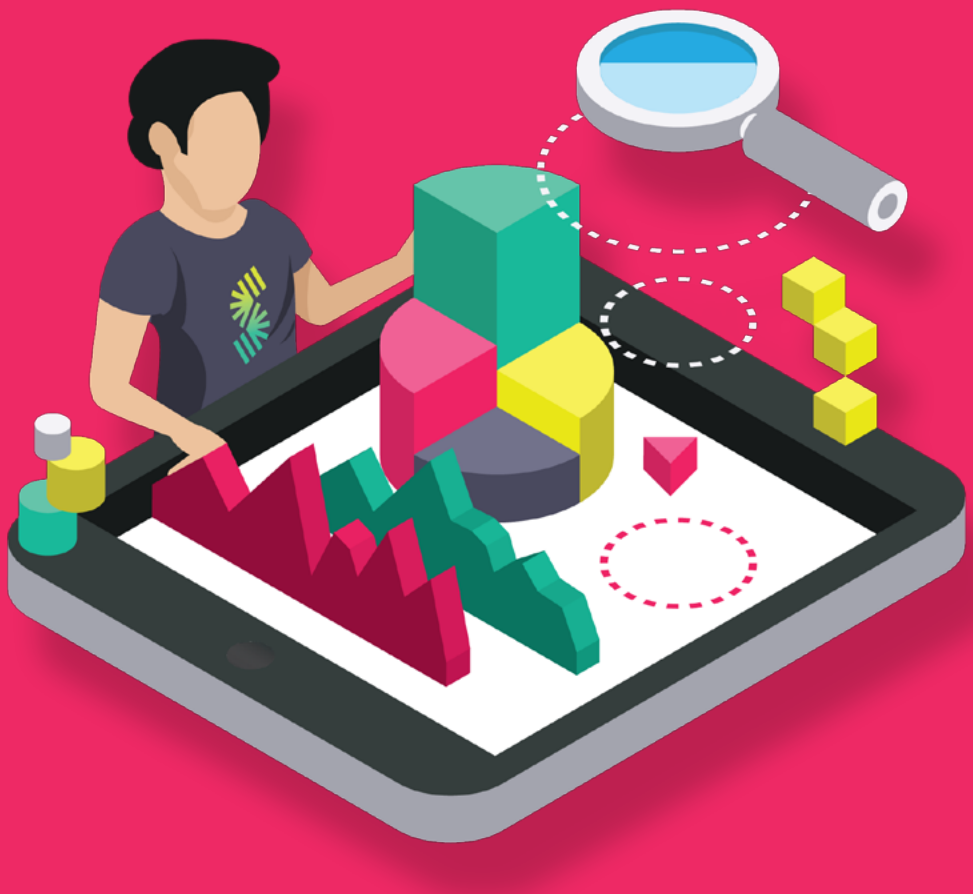
A Beginner's Guide to

Getting Your First Data Science Job



Springboard

Table of Contents

Introduction

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. And at the top of this list, the top job to pursue in 2016 was a relatively new profession called Data Scientist.

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

This 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 order. 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, and helping people advance their careers, 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 and demystify data science careers. Maybe we'll even inspire some of you to pursue a career in this fascinating field.

An Unconventional Innovator

GiveDirectly is a non-for-profit 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 only 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 stunning results. Recipients are well on their well to doubling their amount of assets. Their rate of hunger is almost halved. They earn 34% more.

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 very poorest, they 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.

One of the signs GiveDirectly representatives look for is 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 around $USD 564 in a region where GDP per capita is around $1,700, they represent a significant capital investment, and a good sign of the difference between extreme and relative poverty.

But sending people to each village could take several trips at a crushing expense, creating overheads for an organization looking to operate leanly.

Data Science to the Rescue

"The ability to take data—to be able to understand it, to 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

Liaising with GiveDirectly, a pair of industry experts from IBM and Enigma set out 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 top of 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.

This required mining satellite data and making sense of massive amounts of data, something that would have been impossible a decade ago. It required implementing machine learning algorithms, a cutting-edge technology at the time, to train computers to recognize patterns.

These data scientists were able to pinpoint where GiveDirectly should operate, saving the organization hundreds of man-hours and allowing it to do what it does best: solving extreme poverty.

What is Data Science?

GiveDirectly is just one example of how organizations win by using data to their advantage.

Around the world, organizations are creating more data every day, yet most are struggling to benefit from it. According to McKinsey, the US alone will face a shortage of 150,000+ data analysts and an additional 1.5 million data-savvy managers.

According to LinkedIn, Statistical Analysis & Data Mining were the hottest skills that got recruiters' attention in 2014. Glassdoor ranked Data Scientist as the #1 job to pursue in 2016. Harvard Business Review even called it the sexiest career of the 21st century.

GiveDirectly was able to save thousands of dollars and put their money where their mission is thanks to a team of three data scientists. Within the mass of data the world generates every day, similar insights are hidden away. Each may have the potential to transform entire industries, or to improve millions of lives.

Salary trends have followed the impact data science drives. With a national average salary of $118k (which increases to $126k 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 that 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 you throughout your data science career.

The Foundations of Data Science

DJ Patil, the current Chief Data Scientist of the United States and previously the Head of Data Products at Linkedin, is the one who first coined the term data science.

A decade after it was first used, the term remains contested. There is a lot of debate among practitioners and academics about what data science means, and whether it's different at all from the data analytics that companies have always done.

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.

DJ 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." There is no mention here of a strict definition of data science, nor of a profile that must fit it.

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.

The Different Data Science Roles

Before we dive too deep 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. You don't have to become the world's best at everything.

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 all 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 with data science projects from end-to-end. They can store and clean large amounts of data, explore data sets to identify insights, build predictive models and weave a story around the findings.

Within the broad category of data scientists, you might encounter statisticians who focus on 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.

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 implements 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) 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.

Data Analysts and 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 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.

Skills

You can roughly say that data engineers rely most heavily on software engineering skills, data scientists rely on their training in statistics and mathematical modeling, and business analysts rely more heavily on their analytical skills and domain expertise. You can be sure that people who occupy these roles will have varying amounts of skills outside of their specialities.

It's important to keep this consideration in mind because data science can be a big tent, and you can pick and choose your spots, but each spot comes with different needs, and different salaries.

Salary Ranges

Data scientists need to have the broadest set of skills that covers 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 insights found to drive action within organizations. While their average salary is around $65,000 USD, this is partly because a lot of data analyst roles are filled by entry-level graduates with limited work experience.

Every one of these roles combines together into a whole data science team that can solve any data problem placed in front of them.

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

The first thing you have to do before you solve a problem is to define exactly what it is. You need to be able to translate data questions into something actionable.

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

Say you're solving a problem for the VP 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.

A great way to do this is to ask the right questions.

You should then figure out what the sales process looks like, and who the customers are. You need as much context as possible for your numbers to become insights.

You should ask questions like the following:

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?

In response to your questions, the VP Sales might reveal that they want to understand why certain segments of customers have bought less than expected. Their end goal might be to determine whether to continue to invest in these segments, or de-prioritize 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.

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

Step 2: Collect the raw data needed for your problem

Once you've defined the problem, you'll need data to give you the insights needed to turn the problem around with 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 of their sales data in a CRM or a customer relationship management software platform.You can export the CRM data in a CSV file for further analysis.

Step 3: Process the data for analysis

Now that you have all of the raw data, you'll need to process it before you can

do any analysis. Oftentimes, 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. 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 that there was NO initial contact date? Or do you have to hunt down the VP 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'll 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 insights. You'll have a fixed deadline for your data science project (your VP Sales is probably waiting on your analysis eagerly!), so

you'll have to prioritize your questions. '

You'll have to look at interesting patterns that explain why sales are reduced for this group. You might notice that they don't tend to be very active on social media, with few of them having Twitter or Facebook accounts. You might also notice that most of them are older than your general audience. From that you can begin to trace patterns you can analyze more deeply.

Step 5: Perform in-depth analysis

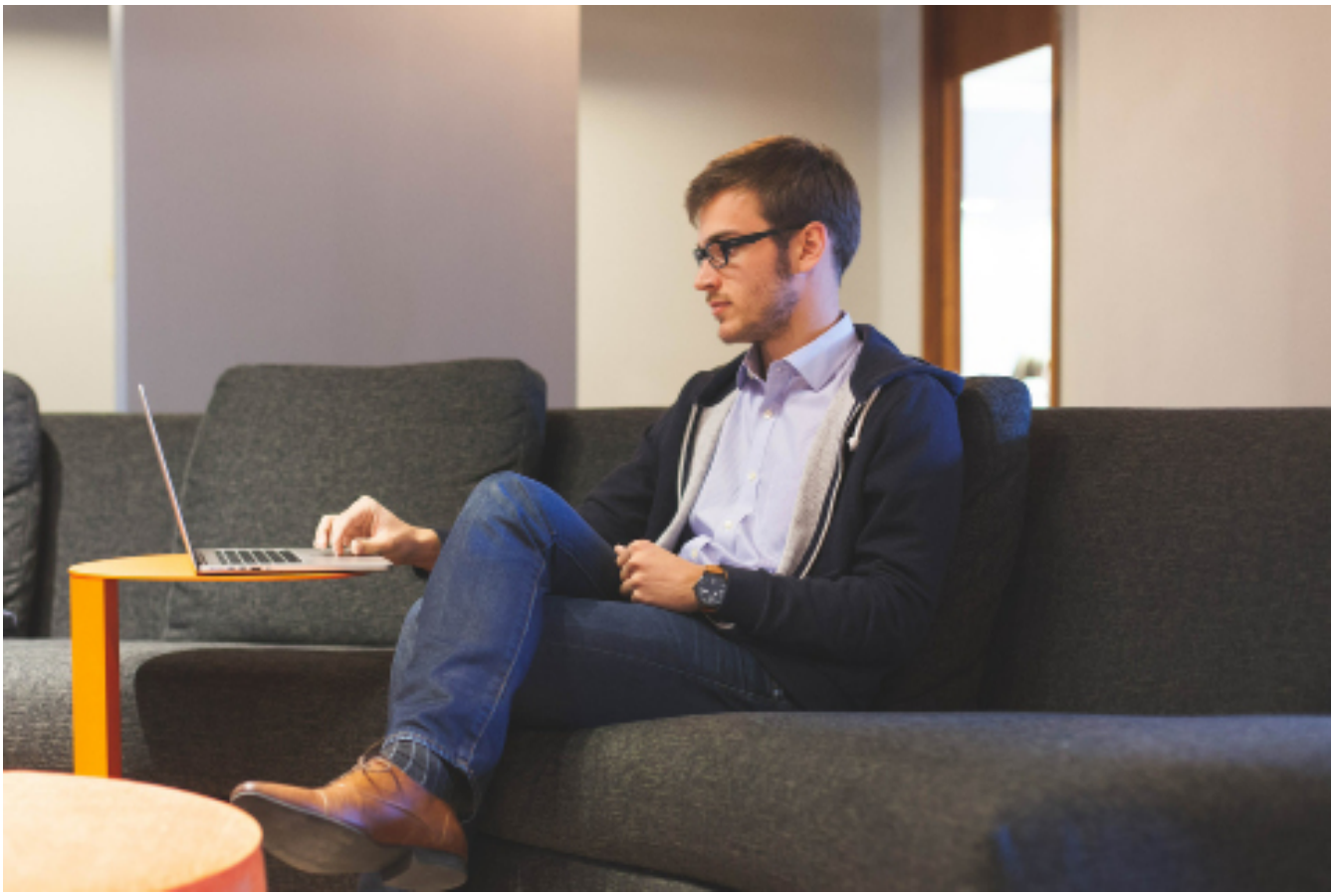
This step of the process is where you're going to have to apply your statistical, mathematical and technological knowledge and leverage all of the data science tools at your disposal to crunch the data and find every insight you can.

In this case, you might 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.

If you'd asked a lot 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 that is aimed at younger audiences.

You would know that certain demographics prefer being reached by telephone rather than by social media. You begin to see how 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 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 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 Sales understand why the insights you've uncovered are important. Ultimately, you've been called upon to create a solution throughout the data science process. Proper communication will mean the difference between action and inaction on your proposals.

Proper communication will mean the difference between action and inaction on your proposals.

You need to craft a compelling story here that ties your data with their knowledge. You start by explaining the reasons behind the underperformance of the older demographic. You tie that in with the answers your VP Sales gave you and the insights you've uncovered from the data. Then you move to concrete solutions that address the problem: we could shift some resources from social media to personal calls. You tie it all together into a narrative that solves the pain of your VP Sales: she now has clarity on how she can reclaim sales and hit her objectives.

She is now 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 how that looks like from day to day.

Data Scientists in Action

We have a whole lot of mentors at Springboard who have shared their stories about the day-to-day 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.

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 or 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 skillset almost all human beings are notoriously bad at. All of this requires a closer look at the data. 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 just like any other job. Nearly an hour is spent just catching up on email and organizing for the day ahead.

10 am

A surprisingly high 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 days 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 just making sure that every team is properly aligned with one another, and working on the right things.

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 the typical day is spent coding an analysis or looking over somebody else's code. This might involve building a graph to represent insights unearthed during a look through the data, or it might just 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 that they can collaborate and help one another. They'll often learn together and share the latest tool that can help improve productivity.

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 users 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 useable 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.

> **Chase makes sure that data problems are framed the right way.**

Data scientists solve many different problems. A data scientist might hunt for raw data. They might be asked to create automated programs that can process

data quickly and efficiently. They might be asked to communicate their results and why they matter to the CEO of a company. You will have to learn a versatile skillset, and a variety of tools if you want to become one.

Understanding the Data: Sneha Runwal

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

She feels that people are often too quick to delve into algorithms and computer code, but it's important to step back and understand your data before you get into implementation mode. She says she is trying to get more disciplined about this herself. She's seen how if trying to implement ideas too quickly can lead you into haphazard and unproductive directions. Her advice? Take the time to understand as much of your data as possible, as early as you can.

What You Need to Learn to Become a Data Scientist

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 data scien-

tists are required to have.

There isn't any one specific academic credential that data scientists are required to have.

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

Data Science Skills

An Analytical Mind

> Takeaway
> You need to approach data science problems analytically to solve them. You'll need an analytical mindset to do well in data science.

You'll need an 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

Takeaway

Mathematics is an important part of data science. 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 gets 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 if you missed those topics in undergrad. You'll need to understand how to manipulate matrices of data and get a general idea behind the math of algorithms.

Statistics

Takeaway

You must know statistics to infer insights from smaller data sets onto

larger populations. This is the fundamental law of data science.

You must know statistics to infer insights from smaller data sets onto larger populations.

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 everybody in a population from a smaller sample.

To understand data science, you must know the basics of hypothesis testing, and experiment design in order to understand where the meaning and context of your data.

Algorithms

Takeaway

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.

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. You will 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 intervention. 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.

Data Visualization

> Takeaway
> Finishing your data analysis is only half the battle. To drive impact, you will have to convince others to believe and adopt your insights.

> **To drive impact, you will have to convince others to believe 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.

Business Knowledge

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

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.

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

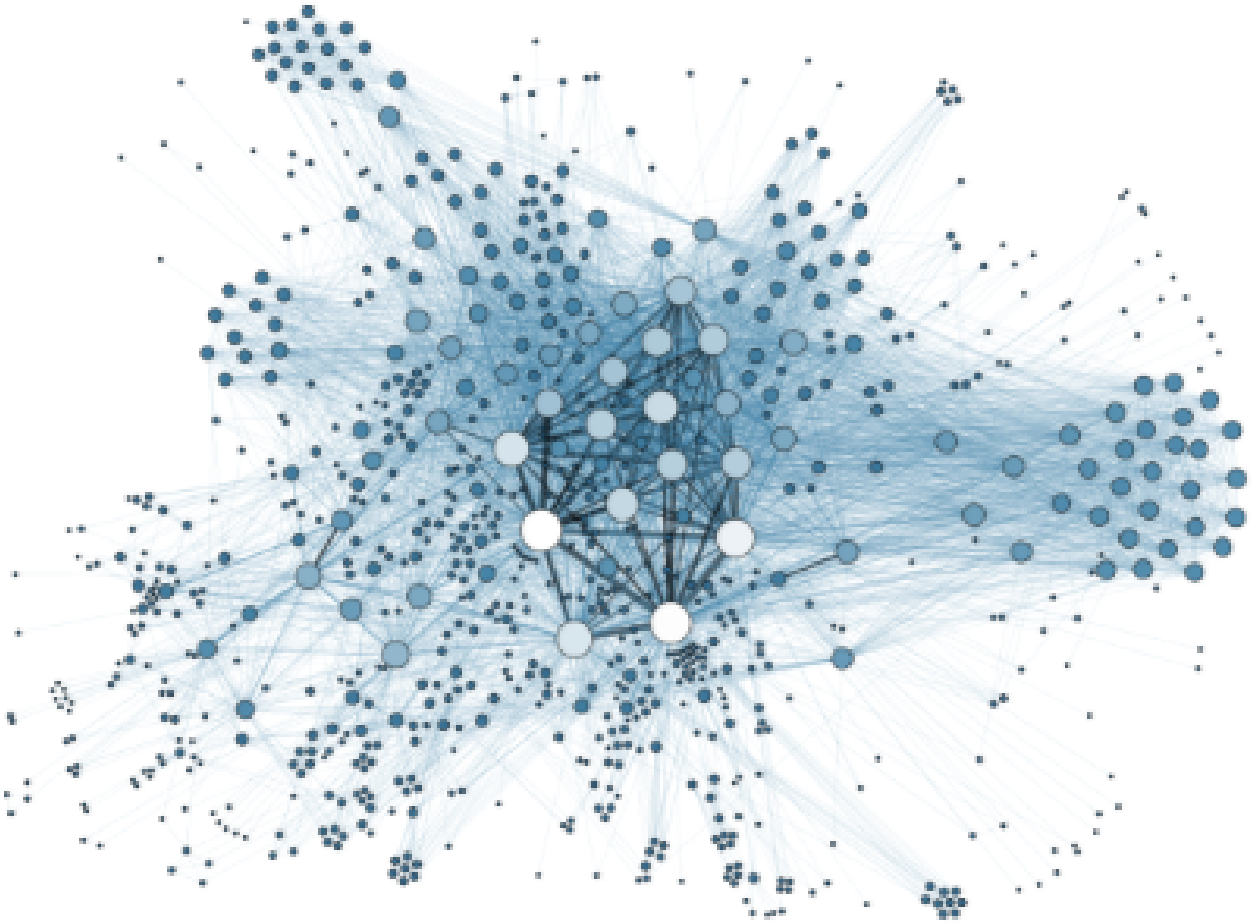
> Takeaway
> As a data scientist, you should know the business you work for and the industry it lives in.

Beyond having deep knowledge of the company you work for, you'll also have to understand its field 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 their 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 may have opened this sort of file with Excel before. CSVs separate out data with a delimiter, a piece of punctuation that serves to separate out different data points.

SQL

SQL, or structured query language, stores data in relational tables. If you go from the right from a column to the left, you'll get different data points on the same entity (for example, a person will have a value in the AGE, GENDER, and HEIGHT categories).

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.

Excel

Takeaway

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

Excel is often the gateway to data science.

Introduction to Excel

Excel allows you to easily manipulate 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 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 dataset on the statistics of NBA players and making a simple graph of the top scorers in the league

Introduction to SQL

Takeaway

SQL is the most popular programming language to find data.

SQL is the most popular programming language 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 with it.

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 ten most popular songs from a SQL database of the Billboard 100.

Python

Takeaway

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

Python is a powerful, 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.

The average salary range for jobs with Python in their description is around

$102,000. Python is the most popular programming language taught in universities: 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.

R

Takeaway

R is a staple in the data science community because it is designed explicitly for data science needs. It is the most popular programming environment in data science with 43% of data professionals using it.

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 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-size 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. Imagine how much data Facebook has at any give time!

Any data set that is too large for conventional data tools such as SQL and Excel can be considered big data, according to McKinsey. The simplest definition is that big data is data that can't fit onto your computer.

Here are tools to solve that problem:

Hadoop

> ### Takeaway
> By using Hadoop, you can store your data in multiple servers while controlling it from one.
>
> **By using Hadoop, you can store your data in multiple servers while controlling it from one.**

Introduction to Hadoop

The solution is a technology called MapReduce. MapReduce is an elegant abstraction that treats a series of computers as it were one central server. This allows you to store data on multiple computers, but process it through one.

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 that update in real time, such as the number of likes Facebook users generate.

NoSQL

Takeaway

NoSQL allows you to manage data without unneeded weight.

NoSQL allows you to manage data without unneeded weight.

Introduction to NoSQL

Tables that bring all their data with them can become cumbersome. NoSQL includes a host of data storage solutions that optimize for speed for certain queries. When Google is looking through millions of websites for certain words, they

often use NoSQL to store and access the results.

Benefits of NoSQL

NoSQL was a data trend that pioneered by technology companies to speed up the millions of queries done on their data. 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 store the data with less structure, and less reliability. The trade-off is a great amount of speed querying data.

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 users of a social media application that is deployed on the web, and which needs to be accessed rapidly by other users of the application.

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		

Collect Data

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

You can create datasets by taking data from what is called an API or an application programming interface that allows you to take structured data from certain providers. You'll be able to query all kinds of data from among others, Twitter, Facebook, and Instagram.

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

set search engines such as Quandl that allow you 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 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, and once you've cleaned the data with the beautifulsoup library, you'll be able to analyze them in-depth.

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

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.

Process Data

Excel allows you to easily clean data with menu functions that can 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 from them.

Python uses the Pandas library for data analysis. It is much quicker to process larger data sets than Excel, and 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 dataset 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 such as dplyr and tidyr 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 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.

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 HTML documents. This can help you embed snippets of code into interactive websites or your online portfolio. Many people have used this function to create

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 such as jpg., 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

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, 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. You'll get paid more if you stick it out with R rather than 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. 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 that work in Python, those that used 15 or more tools made 30k more than those that just used 10 to 14.
>
> ### Takeaway
>
> 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 what tool is best.

Starting Your Job Search

How to Build a Data Science Portfolio and Resume

How to Build a Data Science Portfolio and Resume

You need to make a great first impression to break into data science. That 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.

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 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? Showcase them here.
4. It should be easy to find your contact information.

We found this portfolio of Trent Salazar to show you what a data science portfolio looks like with 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 is one of the top results to come up!

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 you easily look just as good 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 the interest he has in financial modelling and how that has applied to his career.
3. You'll notice he has a lot of his resources on his website. He's adding value to his visitors 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 that 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 show your ability to make a difference, and create the strongest portfolio asset of all: a demonstrated bias to action.

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, and land a job based on that.

Finding 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!

One of the highest-value networking activities you can pursue is finding a mentor who can guide you along your data science career.

What most people don't get 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, which leads to significantly better outcomes through increased accountability and motivation.

Meetups and Conferences

In this section, we're listing some of the popular events and conferences we know of. 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 and pick up the jargon. At some of these events, you will get to hear from and build connections with established data scientists, and even unearth hidden job opportunities.

> **At events and meetups, you'll network with fellow data scientists, and interview for hidden job opportunities.**

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 net-

working events.

KDD (Knowledge Discovery in Data Science)

KDD or Knowledge Discovery in Data Science is another large data science conference. 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.

NIPS (Neural Information Processing Systems)

NIPS, or Neural Information Processing Systems, is a largely academic data science conference, which is 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.

Meetups

We've listed the major conferences where the data science community assembles, but there are often 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 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 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 networks with each other (usually over beers). 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 groups for data science 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! 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.

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 interview 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 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 discussion, and you'll always find a lot of great programming resources and pieces on Hacker News!

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 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 the most prestigious startup accelerator in the world. Angellist is a database for startups looking to get funding and it also has a jobs section.

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? The first thing you have to consider is that a data science interview involves some degree of preparation. There are several kinds of questions that are always asked: your background, coding questions, and applied machine learning ones. You should always anticipate a

mixture of technical and non-technical questions in any data science interview. Make sure you brush up on your programming and data science--and try to interweave it with your own personal story!

You'll also often be asked to analyze datasets. You'll likely be asked culture fit and stats questions.

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, a lot of which are around online.

Here is a list of data science questions you might encounter.

Among some of these questions, you'll see common ones like:

1-Python vs R: which language do you prefer for [x] situation?

2- What is K-means (a specific type of data science algorithm)? Describe when you would use it.

3- Tell me a bit about the last data science project you worked on.

4- What do you know about the key growth drivers for our business?

***

1-The first type of question tests your programming knowledge.

2- The second type of question tests what you know about data science algorithms, and makes you share your real-life experience with them.

3- The third question is a deep dive into the work you've done with data science before.

4- Finally, the fourth type of question will test how much you know about the business you're interviewing with.

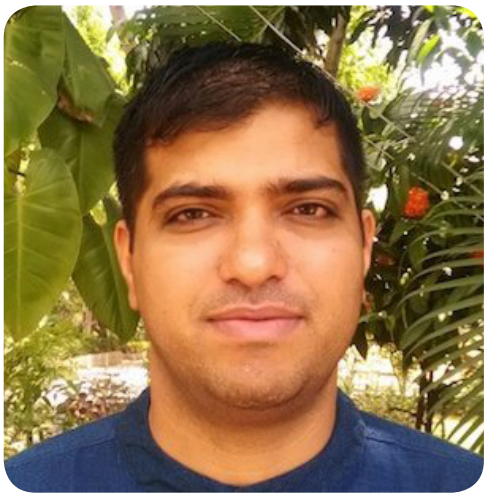
If you can demonstrate how your data science work can help move the needle for your potential employers, you'll impress them. They'll know they have somebody who cares enough to look into what they're doing, and who knows enough about the industry that they don't have to teach you everything.

Paths into Data Science

You might wonder where you'll find people who can do data science. What background do they come from? What skills did 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.

Engineering+Business= Data Storytelling - Amit Kapoor



Takeaway

Amit's path through business and engineering shows that you can become a great data storyteller without a computer science degree.

Amit's path through business and engineering shows that 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.

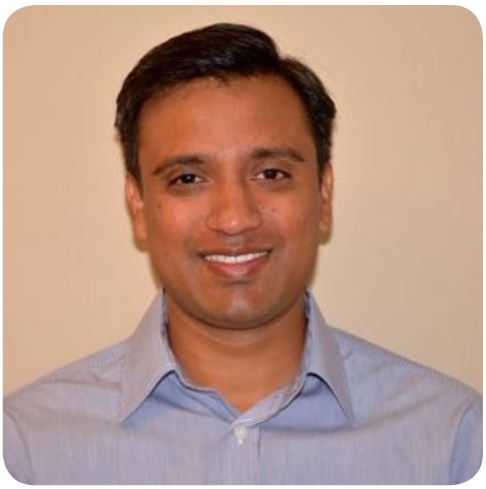
Amit first got a degree in mechanical engineering, then he went on to get a

MBA. He then worked as a consultant for both 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, and so he leveraged his background in both engineering and business to transition into managing his own data visualization consultancy.

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

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

Takeaway



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

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 also mentored for several data science courses and as a data scientist, he works on creating end-to-end solutions that extract value from data. He has his own personal websites with different data science projects.

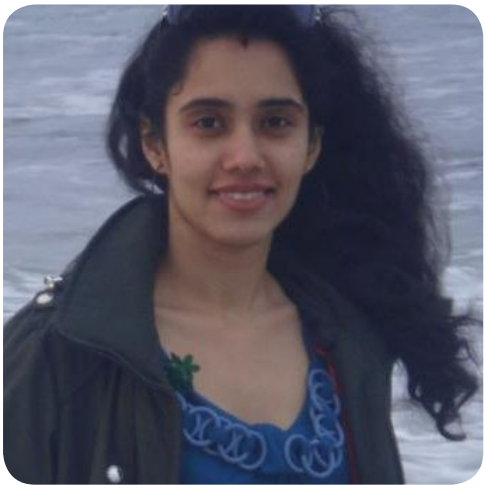
He comes from a traditional data science background with a PhD. in electrical engineering and a job at Cisco as a software engineer--mostly because he didn't feel like there were many opportunities in electrical engineering.

It was really the Machine Learning course at Coursera that helped inspire 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.

Gaining Data Science Experience - Sneha Runwal



Takeaway

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

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 in computer science, and after that, she did

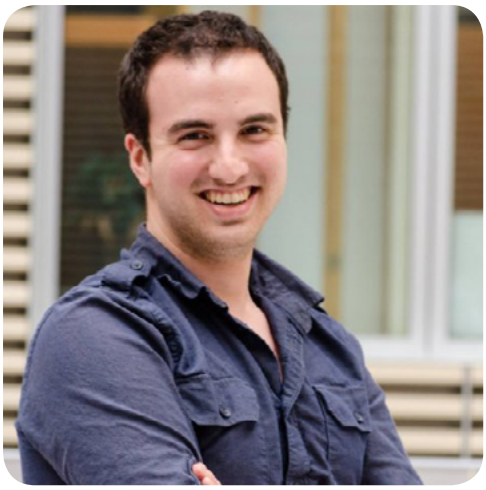
a short stint at Cisco as a software engineer. After pursuing her MBA with a focus in analytics and strategy, she took up several data science internships. After working extensively with Infosys on their HR analytics data, she realized this was something she wanted to pursue.

She ended up moving to the Bay Area after graduating from her MBA, and taking a data science internship with a startup in the area. After working on their psychometrics data to determine what candidates 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.

Competing to Get Into Data Science - Sinan Ozdemir

Takeaway



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

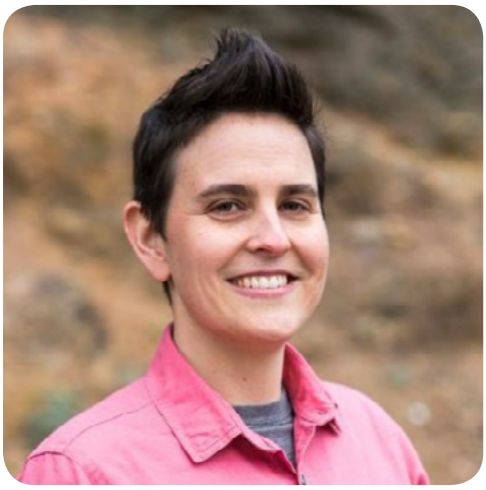
You can participate in online competitions to practice and demonstrate your data science skills.

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. That was when 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.

A Psychology PH.D. on the Path to Data Science - Erin Baker



Takeaway

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

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

Erin is a social psychologist by training who was focused in her PhD. on how to get people to be more pro-social. She wanted to get more people engaged in volunteerism, and in engagement 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 from 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.

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 Science 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

Conclusion

Getting your first data science job might be challenging, but it's possible to achieve with a diligent approach to picking up skills, and 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. We have provided a checklist that might be useful as you continue your journey towards that goal.

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

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

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!

Resources

What is Data Science?

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

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.

Getting Data

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

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.

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.

Data Visualization

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

Data Science Interview

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

Building a Data Science Portfolio

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

Stuff Data Scientists Say

Now that you've been flooded with all of the resources you need to look over and the ideas you need to understand, it's a good time to take a brief break, and step back a little bit. Here's a brief glossary of data science terms we covered before and some we haven't to ensure you never get lost in any data science discussion--or interview.

Algorithm

A set of instructions a computer must follow, typically implemented in computer code such as Python or SQL (not an invention of Al Gore)

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 that 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 in a column) 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

A framework that allows for distributed data analysis across multiple hardware components, commonly associated with big data. It is an open

source framework for analysis of large data sets supported by the Apache Foundation. (not the sound made in a football stadium.)

Hive

A SQL-like query language that allows you to access big data records. (not what bees make when they group together.)

Latency

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

Machine Learning

The use of algorithms to detect patterns in data.

Python

Versatile programming language often used by data scientists (not the giant snake.)

Pandas

A library of code you can access in Python that helps make data analysis easy (not the furry cute animal.)

R

Programming language designed for statistical analysis (not the versatile letter.)

Scalability

The ability of a system to maintain performance as its workload (in this case the volume of data) increases.

Schema

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

SQL

Programming language specifically designed to get data out of relational databases, which have tables of data with columns that relate to one another (not a weird electro group.)

NoSQL

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

Machine Learning

Using data-driven algorithms that direct machines to identify certain features in data, thus allowing the machine to learn and detect patterns (this is what is pretty much sounds like.)

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-labelled inputs to get machine outputs. An example of this is a program that can classify faces based on a dataset of faces already labelled by humans (this is pretty much what it sounds like.)

Unsupervised Learning

This is letting the machine classify features without any human inputs.

An example of this is a program that can classify faces vs. pictures of food from a random set of images without any human labelling of the data (this is pretty much what it sounds like.)