
INSAID

Table Of Contents

1. Introduction

1.1 An Unconventional Innovator

GiveDirectly

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 is achieving 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.

1.2 Data Science to the Rescue

You've probably heard that being a data scientist is the sexiest career of the 21st century, one where you can earn a healthy salary, and a great work-life balance.

Google's Chief Economist, Hal Varian, has said that,

"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".

GiveDirectly is just one example of how organizations win by using data to their advantage. 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. Some people have even called it the sexiest career of the 21st century.

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

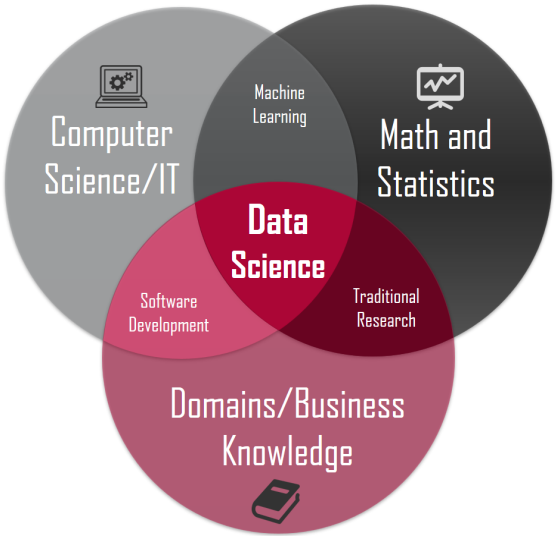
These data scientists were able to pinpoint where GiveDirectly should operate, saving the organization hundreds of man-hours and allowing it to do what it does best: solving extreme poverty.

2. 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.

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 to get 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 document. To bring data science careers to life, for thousands of data-curious, acute young professionals. We hope that after reading this document, you have a solid understanding of the data science industry, and know what it takes to get your first data science job.

2.1 Foundation 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.

In the upcoming years, the world will generate 50x more data than we generated in old ages. **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. 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, an important trait of sense of intellectual curiosity to understand the world through data, and not be deterred easily by obstacles is required.

3. 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 (For example, 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".

3.1 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.

3.2 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 and visualizations to explain what insights the data is hiding. When Chase from Higher Learning Technologies helps people from across the company understand specific queries with charts, he is filling the business analyst role.

3.3 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.

3.4 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.

3.5 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.

4. The Data Science Process

At INSAID various students often ask us questions like "What does a Data Scientist do?". So our Data Science Researchers made an simplified cheat sheet to understand the process in six simple steps.

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.

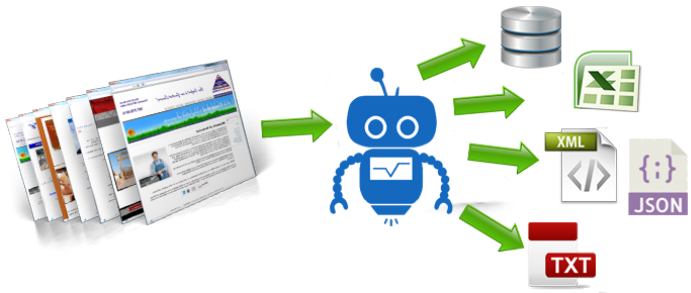
Say you're solving a problem for the Sales of your company. You should start by understanding their goals and the underlying why behind their data questions. Before you can start thinking of solutions, you'll want to work with them to clearly define the problem.

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 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: Data Acquisition



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 software or a customer relationship management software platform. You can export the CRM data in a CSV file for further analysis.

Step 3: Data Pre-processing 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. Time zone differences
4. Duplicate values
5. Outliers

You'll need to look through aggregates of your file rows and columns 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: Data Exploration

When your data is clean, you have to come up with ideas that are likely to turn into insights. You'll have a fixed deadline for your data science project , 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 analyse more deeply.

Step 5: 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 use 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! In-person interactions could change everything for the better. This is something you'll have to flag to your 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: Communication

It's important that the Sales Team understands 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.

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 Sales Team 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.

5. Requirement to Become a Data Scientist

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 scientists are required to have.

5.1 Skills Required

5.1.1 An Analytical Mind:

You need to approach data science problems analytically to solve them. 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.

5.1.2 Mathematics:

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.

5.1.3 Statistics:

You must know statistics to infer insights from smaller data sets onto larger populations. This is the fundamental law of data science. Understanding inferential statistics allows you to make general conclusions about everybody in a population from a smaller sample.

5.1.4 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 analysing data sets too large for the human mind to process.

5.1.5 Data Visualization:

Finishing your data analysis is only half the battle. To drive impact, you will have to convince others to believe and adopt your insights. Human beings are visual creatures. 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.

5.1.6 Business Knowledge:

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

5.1.7 Domain Expertise:

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

5.2 Data Science Tools

With your skill set developed, you'll now need to learn how to use modern data science tools. 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.

5.2.1 File Formats

Data can be stored in different file formats. Some common file formats are Comma Separated Values (CSV), Structured Query Language (SQL), Java Script Object Notation (JSON), Excel (XLSX).

5.2.2 Microsoft 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.

5.2.3 SQL Database

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

5.2.4 Python Language

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.

5.2.5 R Language

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.

5.2.6 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. 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.

The solutions are given below

5.2.7 Hadoop

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

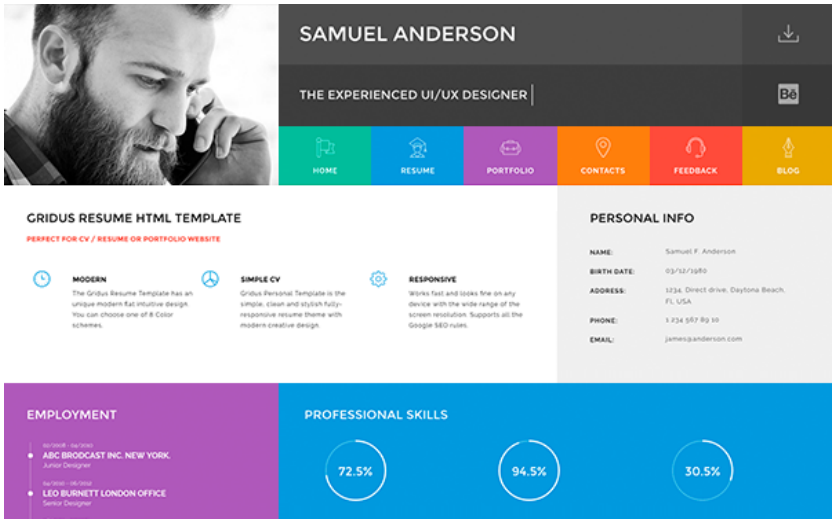
5.2.8 NoSQL

NoSQL allows you to manage data without unneeded weight. Tables that bring all their data with them can become cumbersome. NoSQL includes a host of data storage solutions that separate out huge data sets into manageable chunks.

6. Initiation Of Your Job Hunt

6.1 Building of 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. Take a following example.

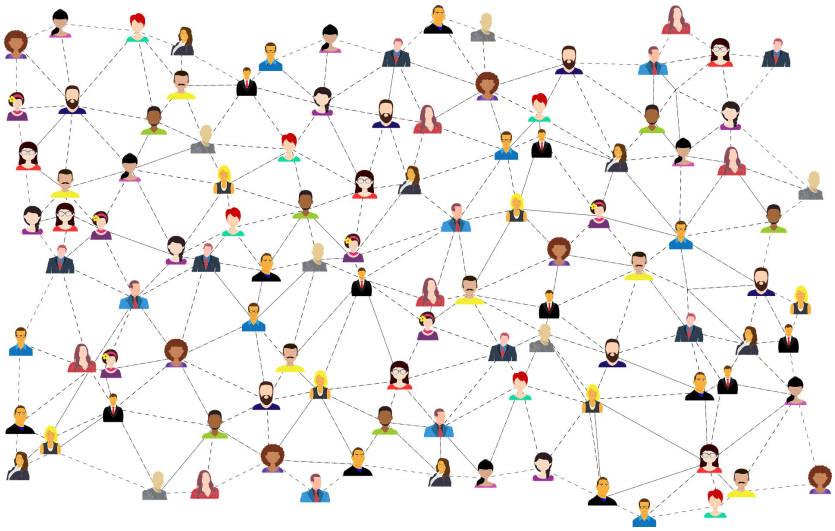


This allows them to demonstrate their experience and the value they create in the data science community. In order for your portfolio to have the same effect, it must share the following traits:

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. 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.

6.2 Building Data Science Network



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.

6.3 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. 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.

6.4 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. Your network will grow and you get some insights what usually is done by data scientists.

6.4.1 Conferences

6.4.1.1 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.

6.4.1.2 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.

6.4.1.3 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.

6.4.2 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

6.4.2 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!

6.5 Job Boards for Data Science

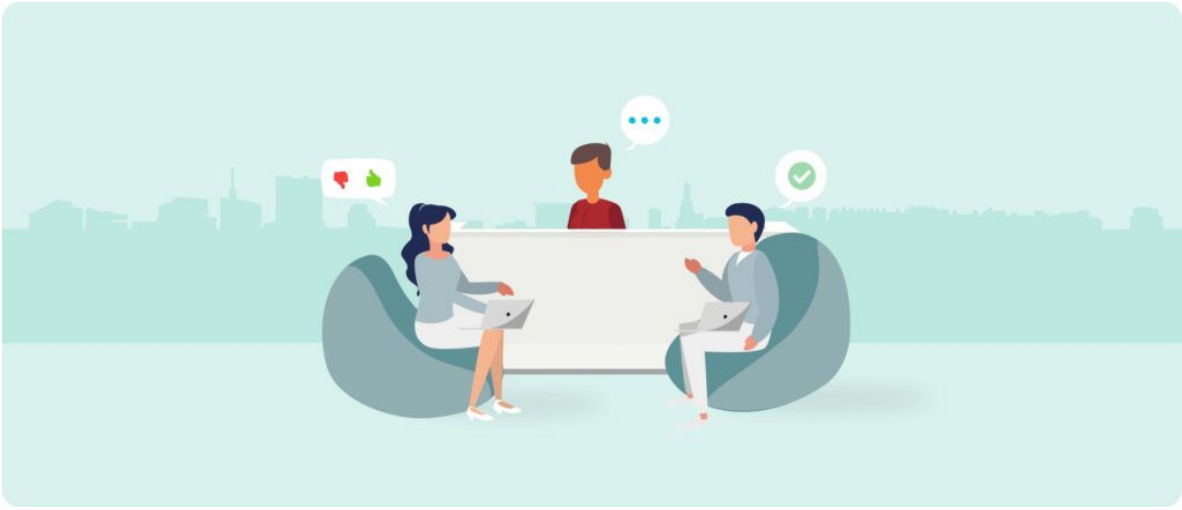
1. Kaggle offers a job board for data scientists.
2. You can find a list of open data scientist jobs at Indeed, the search engine for jobs.
3. Datajobs offers a listings site for data science.
4. Datasciencejobs scrapes data science jobs from around the web into one centralized location.

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

6.6 Data Science Interview

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 are some frequently asked questions.

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.

7. 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.

On the data science process: Acquiring and cleaning data takes about 75% of the time in a project.

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 his thought process?

8. Checklist

- Assemble a learning plan.
- Brush up on linear algebra and calculus.
- Learn the theory of statistics and machine learning.
- Install Python, play around with it.
- Install R, play around with it.
- Do a few SQL queries.
- Assemble your own data set from a website.

- Register for a data science community online such as Kaggle.
- Attend a data science meetup.
- Network with at least five data scientists.
- Look for a data science mentor.
- Build a portfolio filled with your data science projects!

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