

IN THIS CHAPTER

- » Deploying data science methods across various industries
- » Piecing together the core data science components
- » Identifying viable data science solutions to business challenges
- » Exploring data science career alternatives

Chapter 1

Wrapping Your Head Around Data Science

For over a decade now, *everyone* has been absolutely deluged by data. It's coming from every computer, every mobile device, every camera, and every imaginable sensor — and now it's even coming from watches and other wearable technologies. Data is generated in every social media interaction we humans make, every file we save, every picture we take, and every query we submit; data is even generated when we do something as simple as ask a favorite search engine for directions to the closest ice cream shop.

Although data immersion is nothing new, you may have noticed that the phenomenon is accelerating. Lakes, puddles, and rivers of data have turned to floods and veritable tsunamis of structured, semistructured, and unstructured data that's streaming from almost every activity that takes place in both the digital and physical worlds. It's just an unavoidable fact of life within the information age.

If you're anything like I was, you may have wondered, "What's the point of all this data? Why use valuable resources to generate and collect it?" Although even just two decades ago, no one was in a position to make much use of most of the data that's generated, the tides today have definitely turned. Specialists known as *data engineers* are constantly finding innovative and powerful new ways to

capture, collate, and condense unimaginably massive volumes of data, and other specialists, known as *data scientists*, are leading change by deriving valuable and actionable insights from that data.

In its truest form, data science represents the optimization of processes and resources. Data science produces *data insights* — actionable, data-informed conclusions or predictions that you can use to understand and improve your business, your investments, your health, and even your lifestyle and social life. Using data science insights is like being able to see in the dark. For any goal or pursuit you can imagine, you can find data science methods to help you predict the most direct route from where you are to where you want to be — and to anticipate every pothole in the road between both places.

Seeing Who Can Make Use of Data Science

The terms *data science* and *data engineering* are often misused and confused, so let me start off by clarifying that these two fields are, in fact, separate and distinct domains of expertise. *Data science* is the computational science of extracting meaningful insights from raw data and then effectively communicating those insights to generate value. *Data engineering*, on the other hand, is an engineering domain that's dedicated to building and maintaining systems that overcome data processing bottlenecks and data handling problems for applications that consume, process, and store large volumes, varieties, and velocities of data. In both data science and data engineering, you commonly work with these three data varieties:

- » **Structured:** Data that is stored, processed, and manipulated in a traditional *relational database management system (RDBMS)* — an example of this would be a MySQL database that uses a tabular schema of rows and columns, making it easier to identify specific values within data that's stored within the database.
- » **Unstructured:** Data that is commonly generated from human activities and doesn't fit into a structured database format. Examples of unstructured data is data that comprises email documents, Word documents or audio / video files.
- » **Semistructured:** Data that doesn't fit into a structured database system but is nonetheless organizable by tags that are useful for creating a form of order and hierarchy in the data. XML and JSON files are examples of data that comes in semi-structured form.

It used to be that only large tech companies with massive funding had the skills and computing resources required to implement data science methodologies to optimize and improve their business, but that's not been the case for quite a while now. The proliferation of data has created a demand for insights, and this demand is embedded in many aspects of modern culture — from the Uber passenger who expects the driver to show up exactly at the time and location predicted by the Uber application to the online shopper who expects the Amazon platform to recommend the best product alternatives for comparing similar goods before making a purchase. Data and the need for data-informed insights are ubiquitous. Because organizations of all sizes are beginning to recognize that they're immersed in a sink-or-swim, data-driven, competitive environment, data know-how has emerged as a core and requisite function in almost every line of business.

What does this mean for the average knowledge worker? First, it means that everyday employees are increasingly expected to support a progressively advancing set of technological and data requirements. Why? Well, that's because almost all industries are reliant on data technologies and the insights they spur. Consequently, many people are in continuous need of upgrading their data skills, or else they face the real possibility of being replaced by a more data-savvy employee.

The good news is that upgrading data skills doesn't usually require people to go back to college, or — God forbid — earn a university degree in statistics, computer science, or data science. The bad news is that, even with professional training or self-teaching, it always takes extra work to stay industry-relevant and tech-savvy. In this respect, the data revolution isn't so different from any other change that has hit industry in the past. The fact is, in order to stay relevant, you need to take the time and effort to acquire the skills that keep you current. When you're learning how to do data science, you can take some courses, educate yourself using online resources, read books like this one, and attend events where you can learn what you need to know to stay on top of the game.

Who can use data science? You can. Your organization can. Your employer can. Anyone who has a bit of understanding and training can begin using data insights to improve their lives, their careers, and the well-being of their businesses. Data science represents a change in the way you approach the world. When determining outcomes, people once used to make their best guess, act on that guess, and then hope for the desired result. With data insights, however, people now have access to the predictive vision that they need to truly drive change and achieve the results they want.

Here are some examples of ways you can use data insights to make the world, and your company, a better place:

» **Business systems:** Optimize returns on investment (those crucial ROIs) for any measurable activity.

- » **Marketing strategy development:** Use data insights and predictive analytics to identify marketing strategies that work, eliminate under-performing efforts, and test new marketing strategies.
- » **Keep communities safe:** Predictive policing applications help law enforcement personnel predict and prevent local criminal activities.
- » **Help make the world a better place for those less fortunate:** Data scientists in developing nations are using social data, mobile data, and data from websites to generate real-time analytics that improve the effectiveness of humanitarian responses to disaster, epidemics, food scarcity issues, and more.

Inspecting the Pieces of the Data Science Puzzle

To practice data science, in the true meaning of the term, you need the analytical know-how of math and statistics, the coding skills necessary to work with data, and an area of subject matter expertise. Without this expertise, you might as well call yourself a mathematician or a statistician. Similarly, a programmer without subject matter expertise and analytical know-how might better be considered a software engineer or developer, but not a data scientist.

The need for data-informed business and product strategy has been increasing exponentially for about a decade now, thus forcing all business sectors and industries to adopt a data science approach. As such, different flavors of data science have emerged. The following are just a few titles under which experts of every discipline are required to know and regularly do data science: director of data science-advertising technology, digital banking product owner, clinical biostatistician, geotechnical data scientist, data scientist-geospatial and agriculture analytics, data and tech policy analyst, global channel ops-data excellence lead, and data scientist-healthcare.

Nowadays, it's almost impossible to differentiate between a proper data scientist and a subject matter expert (SME) whose success depends heavily on their ability to use data science to generate insights. Looking at a person's job title may or may not be helpful, simply because many roles are titled data scientist when they may as well be labeled data strategist or product manager, based on the actual requirements. In addition, many knowledge workers are doing daily data science and not working under the title of data scientist. It's an overhyped, often misleading label that's not always helpful if you're trying to find out what a data scientist does by looking at online job boards. To shed some light, in the following sections I spell

out the key components that are part of any data science role, regardless of whether that role is assigned the data scientist label.

Collecting, querying, and consuming data

Data engineers have the job of capturing and collating large volumes of structured, unstructured, and semi structured *big data* — an outdated term that’s used to describe data that exceeds the processing capacity of conventional database systems because it’s too big, it moves too fast, or it lacks the structural requirements of traditional database architectures. Again, data engineering tasks are separate from the work that’s performed in data science, which focuses more on analysis, prediction, and visualization. Despite this distinction, whenever data scientists collect, query, and consume data during the analysis process, they perform work similar to that of the data engineer (the role I tell you about earlier in this chapter).

Although valuable insights can be generated from a single data source, often the combination of several relevant sources delivers the contextual information required to drive better data-informed decisions. A data scientist can work from several datasets that are stored in a single database, or even in several different data storage environments. At other times, source data is stored and processed on a cloud-based platform built by software and data engineers.

No matter how the data is combined or where it’s stored, if you’re a data scientist, you almost always have to *query* data — write commands to extract relevant datasets from data storage systems, in other words. Most of the time, you use Structured Query Language (SQL) to query data. (Chapter 7 is all about SQL, so if the acronym scares you, jump ahead to that chapter now.)

Whether you’re using a third-party application or doing custom analyses by using a programming language such as R or Python, you can choose from a number of universally accepted file formats:

- » **Comma-separated values (CSV):** Almost every brand of desktop and web-based analysis application accepts this file type, as do commonly used scripting languages such as Python and R.
- » **Script:** Most data scientists know how to use either the Python or R programming language to analyze and visualize data. These script files end with the extension `.ply` or `.ipynb` (Python) or `.r` (R).
- » **Application:** Excel is useful for quick-and-easy, spot-check analyses on small- to medium-size datasets. These application files have the `.xls` or `.xlsx` extension.

» **Web programming:** If you're building custom, web-based data visualizations, you may be working in D3.js — or data-driven documents, a JavaScript library for data visualization. When you work in D3.js, you use data to manipulate web-based documents using `.html`, `.svg`, and `.css` files.

Applying mathematical modeling to data science tasks

Data science relies heavily on a practitioner's math skills (and statistics skills, as described in the following section) precisely because these are the skills needed to understand your data and its significance. These skills are also valuable in data science because you can use them to carry out predictive forecasting, decision modeling, and hypotheses testing.



REMEMBER

Mathematics uses deterministic methods to form a *quantitative* (or *numerical*) description of the world; *statistics* is a form of science that's derived from mathematics, but it focuses on using a *stochastic* (probabilities) approach and inferential methods to form a quantitative description of the world. I tell you more about both in Chapter 4. Data scientists use mathematical methods to build decision models, generate approximations, and make predictions about the future. Chapter 4 presents many mathematical approaches that are useful when working in data science.



REMEMBER

In this book, I assume that you have a fairly solid skill set in basic math — you will benefit if you've taken college-level calculus or even linear algebra. I try hard, however, to meet readers where they are. I realize that you may be working based on a limited mathematical knowledge (advanced algebra or maybe business calculus), so I convey advanced mathematical concepts using a plain-language approach that's easy for everyone to understand.

Deriving insights from statistical methods

In data science, statistical methods are useful for better understanding your data's significance, for validating hypotheses, for simulating scenarios, and for making predictive forecasts of future events. Advanced statistical skills are somewhat rare, even among quantitative analysts, engineers, and scientists. If you want to go places in data science, though, take some time to get up to speed in a few basic statistical methods, like linear and logistic regression, naïve Bayes classification, and time series analysis. These methods are covered in Chapter 4.

Coding, coding, coding — it's just part of the game

Coding is unavoidable when you're working in data science. You need to be able to write code so that you can instruct the computer in how to manipulate, analyze, and visualize your data. Programming languages such as Python and R are important for writing scripts for data manipulation, analysis, and visualization. SQL, on the other hand, is useful for data querying. Finally, the JavaScript library D3.js is often required for making cool, custom, and interactive web-based data visualizations.

Although coding is a requirement for data science, it doesn't have to be this big, scary *thing* that people make it out to be. Your coding can be as fancy and complex as you want it to be, but you can also take a rather simple approach. Although these skills are paramount to success, you can pretty easily learn enough coding to practice high-level data science. I've dedicated Chapters 6 and 7 to helping you get to know the basics of what's involved in getting started in Python and R, and querying in SQL (respectively).

Applying data science to a subject area

Statisticians once exhibited some measure of obstinacy in accepting the significance of data science. Many statisticians have cried out, "Data science is nothing new — it's just another name for what we've been doing all along!" Although I can sympathize with their perspective, I'm forced to stand with the camp of data scientists who markedly declare that data science is separate, and definitely distinct, from the statistical approaches that comprise it.

My position on the unique nature of data science is based to some extent on the fact that data scientists often use computer languages not used in traditional statistics and take approaches derived from the field of mathematics. But the main point of distinction between statistics and data science is the need for subject matter expertise.

Because statisticians usually have only a limited amount of expertise in fields outside of statistics, they're almost always forced to consult with a SME to verify exactly what their findings mean and to determine the best direction in which to proceed. Data scientists, on the other hand, should have a strong subject matter expertise in the area in which they're working. Data scientists generate deep insights and then use their domain-specific expertise to understand exactly what those insights mean with respect to the area in which they're working.

The following list describes a few ways in which today's knowledge workers are coupling data science skills with their respective areas of expertise in order to amplify the results they generate.

- » **Clinical informatics scientists** combine their healthcare expertise with data science skills to produce personalized healthcare treatment plans. They use healthcare informatics to predict and preempt future health problems in at-risk patients.
- » **Marketing data scientists** combine data science with marketing expertise to predict and preempt customer *churn* (the loss of customers from a product or service to that of a competitor's, in other words). They also optimize marketing strategies, build recommendation engines, and fine-tune marketing mix models. I tell you more about using data science to increase marketing ROI in Chapter 11.
- » **Data journalists** *scrape* websites (extract data in bulk directly from the pages on a website, in other words) for fresh data in order to discover and report the latest breaking-news stories. (I talk more about data storytelling in Chapter 8.)
- » **Directors of data science** bolster their technical project management capabilities with an added expertise in data science. Their work includes leading data projects and working to protect the profitability of the data projects for which they're responsible. They also act to ensure transparent communication between C-suite executives, business managers, and the data personnel on their team who actually do the implementation work. (I share more details in Part 4 about leading successful data projects; check out Chapter 18 for details about data science leaders.)
- » **Data product managers** supercharge their product management capabilities with the power of data science. They use data science to generate predictive insights that better inform decision-making around product design, development, launch, and strategy. This is a classic type of data leadership role, the likes of which are covered in Chapter 18. For more on developing effective data strategy, take a gander at Chapters 15 through 17.
- » **Machine learning engineers** combine software engineering superpowers with data science skills to build predictive applications. This is a classic data implementation role, more of which is discussed in Chapter 2.

Communicating data insights

As a data scientist, you must have sharp verbal communication skills. If a data scientist can't communicate, all the knowledge and insight in the world does *nothing* for the organization. Data scientists need to be able to explain data insights in a way that staff members can understand. Not only that, data scientists need to

be able to produce clear and meaningful data visualizations and written narratives. Most of the time, people need to see a concept for themselves in order to truly understand it. Data scientists must be creative and pragmatic in their means and methods of communication. (I cover the topics of data visualization and data-driven storytelling in much greater detail in Chapter 8.)

Exploring Career Alternatives That Involve Data Science

Not to cause alarm, but it's fully possible for you to develop deep and sophisticated data science skills and then come away with a gut feeling that you know you're meant to do something more.

Earlier in my data career, I was no stranger to this feeling. I'd just gone and pumped up my data science skills. It was the “sexiest” career path — according to *Harvard Business Review* in 2012 — and offered so many opportunities. The money was good and the demand was there. What's not to love about opportunities with big tech giants, start-ups, and multiple six-figure salaries, right?

But very quickly, I realized that, although I had the data skills and education I needed to land some sweet opportunities (including interview offers from Facebook!), I soon realized that coding away and working only on data implementation simply weren't what I was meant to do for the rest of my life.

Something about getting lost in the details felt disempowering to me. My personality craved more energy, more creativity — plus, I needed to see the big-picture impact that my data work was making.

In short, I hadn't yet discovered my inner data superhero. I coined this term to describe that juicy combination of a person's data skills, coupled with their personality, passions, goals, and priorities. When all these aspects are in sync, you'll find that you're absolutely on fire in your data career. These days, I'm a data entrepreneur. I get to spend my days doing work that I absolutely adore and that's truly aligned with my mission and vision for my data career and life-at-large. I want the same thing for you, dear reader.



TIP

Over on the companion site to this book (<https://businessgrowth.ai/>), you can find free access to a fun, 45-second quiz about data career paths. It helps you uncover your own inner data superhero type. Take the quiz to receive personalized data career recommendations that directly align with your unique combination of data skills, personality, and passions.

For now, let's take a look at the three main data superhero archetypes that I've seen evolving and developing over the past decade.

The data implementer

Some data science professionals were simply *born* to be implementers. If that's you, then your secret superpower is building data and artificial intelligence (AI) solutions. You have a meticulous attention to detail that naturally helps you in coding up innovative solutions that deliver reliable and accurate results — almost every time. When you're facing a technical challenge, you can be more than a little stubborn. You're able to accomplish the task, no matter how complex.

Without implementers, none of today's groundbreaking technologies would even exist. Their unparalleled discipline and inquisitiveness keep them in the problem-solving game all the way until project completion. They usually start off a project with a simple request and some messy data, but through sheer perseverance and brainpower, they're able to turn them into clear and accurate predictive data insights — or a data system, if they prefer to implement data engineering rather than data science tasks. If you're a data implementer, math and coding are your bread-and-butter, so to speak.

Part 2 of this book are dedicated to showing you the basics of data science and the skills you need to take on to get started in a career in data science implementation. You may also be interested in how your work in this area is applied to improve a business's profitability. You can read all about this topic in Part 3.

The data leader

Other data science professionals naturally gravitate more toward business, strategy, and product. They take their data science expertise and apply it to lead profit-forming data science projects and products. If you're a natural data leader, then you're gifted at leading teams and project stakeholders through the process of building successful data solutions. You're a meticulous planner and organizer, which empowers you to show up at the right place and the right time, and hopefully keep your team members moving forward without delay.

Data leaders love data science just as much as data implementers and data entrepreneurs — you can read about them in the later section “The data entrepreneur.” The difference between most data implementers and data leaders is that leaders generally love data science for the incredible outcomes that it makes possible. They have a deep passion for using their data science expertise and leadership skills to create tangible results. Data leaders love to collaborate with smart people across the company to get the job done right. With teamwork, and some

input from the data implementation team, they form brilliant plans for accomplishing any task, no matter how complex. They harness manpower, data science savvy, and serious business acumen to produce some of the most innovative technologies on the planet.

Chapters 7 through 9 and Chapters 15 through 17 in this book are dedicated to showing you the basics of the data science leadership-and-strategy skills you need in order to nail down a job as a data science leader.

That said, to lead data science projects, you should know what's involved in implementing them — you'll lead a team of data implementers, after all. See Part 2 — it covers all the basics on data science implementation. You also need to know prominent data science use cases, which you can explore over in Part 3.

The data entrepreneur

The third data superhero archetype that has evolved over the past decade is the data entrepreneur. If you're a data entrepreneur, your secret superpower is building up businesses by delivering exceptional data science services and products.

You have the same type of focus and drive as the data implementer, but you apply it toward bringing your business vision to reality. But, like the data leader, your love for data science is inspired mostly by the incredible outcomes that it makes possible. A data entrepreneur has many overlapping traits and a greater affinity for either the data implementer or the data leader, but with one important difference:

Data entrepreneurs crave the creative freedom that comes with being a founder.

Data entrepreneurs are more risk-tolerant than their data implementer or data leader counterparts. This risk tolerance and desire for freedom allows them to do what they do — which is to create a vision for a business and then use their data science expertise to guide the business to turn that vision into reality.

For more information on how to transform data science expertise into a profitable product or business, jump over to Part 3.

Using my own data science career to illustrate what this framework looks like in action, (as mentioned earlier in this chapter) I started off as a data science implementer, and quickly turned into a data entrepreneur. Within my data business, however, my focus has been on data science training services, data strategy services, and mentoring data entrepreneurs to build world-class businesses. I've helped educate more than a million data professionals on data science and helped

grow existing data science communities to more than 650,000 data professionals — and counting. Stepping back, you could say that although I call myself a data entrepreneur, the work I do has a higher degree of affinity to data leadership than data implementation.



TIP

I encourage you to go to the companion site to this book at <https://businessgrowth.ai/> and take that career path quiz I mention earlier in this section. The quiz can give you a head-start in determining where you best fit within the spectrum of data science superhero archetypes.