# PART I

# CONCEPTUAL INTRODUCTIONS

This part includes three chapters that serve as the foundations of data science. If you have never done anything with data science or statistics, I highly recommend going through this part before proceeding further. If, on the other hand, you have a good background in statistics and a basic knowledge of data storage, formats, and processing, you can easily skim through most of the material here.

Chapter 1 introduces the field of data science, along with various applications. It also points out important differences and similarities with related fields of computer science, statistics, and information science.

Chapter 2 describes the nature and structure of data as we encounter it today. It initiates the student about data formats, storage, and retrieval infrastructures.

Chapter 3 introduces several important techniques for data science. These techniques stem primarily from statistics and include correlation analysis, regression, and introduction to data analytics.

No matter where you come from, I would still recommend paying attention to some of the sections in Chapter 1 that introduce various basic concepts of data science and how they are related to other disciplines. In my experience, I have also found that various aspects of data pre-processing are often skipped in many data science curricula, but if you want to develop a more comprehensive understanding of data science, I suggest you go through Chapter 2 as well. Finally, even if you have a solid background in statistics, it would not hurt to at least skim through Chapter 3, as it introduces some of the statistical concepts that we will need many times in the rest of the book.

# 1 Introduction

*"It is a capital mistake to theorize before one has data.*
*Insensibly, one begins to twist the facts to suit theories, instead of theories to suit facts."*

*— Sherlock Holmes*

What do you need?
- A general understanding of computer and data systems.
- A basic understanding of how smartphones and other day-to-day life devices work.

What will you learn?
- Definitions and notions of data science.
- How data science is related to other disciplines.
- Computation thinking – a way to solve problems systematically.
- What skills data scientists need.

## 1.1 What Is Data Science?

Sherlock Holmes would have loved living in the twenty-first century. We are drenched in **data**, and so many of our problems (including a murder mystery) can be solved using large amounts of data existing at personal and societal levels.

These days it is fair to assume that most people are familiar with the term "data." We see it everywhere. And if you have a cellphone, then chances are this is something you have encountered frequently. Assuming you are a "connected" person who has a smartphone, you probably have a data plan from your phone service provider. The most common cellphone plans in the USA include unlimited talk and text, and a limited amount of data – 5 GB, 20 GB, etc. And if you have one of these plans, you know well that you are "using data" through your phone and you get charged per usage of that data. You understand that checking your email and posting a picture on a social media platform consumes data. And if you are a curious (or thrifty) sort, you calculate how much data you consume monthly and pick a plan that fits your needs.

You may also have come across terms like "data sharing," when picking a family plan for your phone(s). But there are other places where you may have encountered the notion of data sharing. For instance, if you have concerns about privacy, you may want to know if your cellphone company "shares" data about you with others (including the government).

And finally, you may have heard about "data warehouses," as if data is being kept in big boxes on tall shelves in middle-of-nowhere locations.

In the first case, the individual is consuming data by retrieving email messages and posting pictures. In the second scenario concerning data sharing, "data" refers to information *about* you. And third, data is used as though it represents a physical object that is being stored somewhere. The nature and the size of "data" in these scenarios vary enormously – from personal to institutional, and from a few kilobytes (kB) to several petabytes (PB).

In this book, we will consider these and more scenarios and learn about defining, storing, cleaning, retrieving, and analyzing data – all for the purpose of deriving meaningful insights toward making decisions and solving problems. And we will use systematic, verifiable, and repeatable processes; or in other words, we will apply scientific approaches and techniques. Finally, we will do almost all of these processes with a hands-on approach. That means we will look at data and situations that generate or use data, and we will manipulate data using tools and techniques. But before we begin, let us look at how others describe data science.

**FYI: Datum, Data, and Science**

Webster's dictionary (https://www.merriam-webster.com/dictionary/datum) defines *data* as a plural form of *datum* as "something given or admitted especially as a basis for reasoning or inference." For the purpose of this book, as is common these days, we will use *data* for both plural and singular forms. For example, imagine a table containing birthdays of everyone in your class or office. We can consider this whole table (a collection of birthdays) as data. Each birthday is a single point of data, which could be called *datum*, but we will call that *data* too.

There is also often a debate about what is the difference between *data* and *information*. In fact, it is common to use one to define the other (e.g., "data is a piece of information"). We will revisit this later in this chapter when we compare data science and information science.

Since we are talking about sciences, it is also important to clarify here what exactly is *science*. According to the Oxford dictionary (https://en.oxforddictionaries.com/definition/science), science is "systematic study of the structure and behaviour of the physical and natural world through observation and experiment." When we talk about science, we are interested in using a systematic approach that can allow us to study a phenomenon, often giving us the ability to explain and derive meaningful insights.

Frank Lo, the Director of Data Science at Wayfair, says this on datajobs.com: "Data science is a multidisciplinary blend of data inference, algorithm development, and technology in order to solve analytically complex problems."[1] He goes on to elaborate that data science, at its core, involves uncovering insights from mining data. This happens through exploration of the data using various tools and techniques, testing hypotheses, and creating conclusions with data and analyses as evidence.

In one famous article, Davenport and Patil[2] called data science "the sexiest job of the twenty-first century." Listing data-driven companies such as (in alphabetical order) Amazon, eBay, Google, LinkedIn, Microsoft, Twitter, and Walmart, the authors see a data scientist as a hybrid of data hacker, analyst, communicator, and trusted adviser; a Sherlock Holmes for the

twenty-first century. As data scientists face technical limitations and make discoveries to address these problems, they communicate what they have learned and suggest implications for new business directions. They also need to be creative in visually displaying information, and clearly and compellingly showing the patterns they find. One of the data scientist's most important roles in the field is to advise executives and managers on the implications of the data for their products, services, processes, and decisions.

In this book, we will consider **data science** as a field of study and practice that involves the collection, storage, and processing of data in order to derive important insights into a problem or a phenomenon. Such data may be generated by humans (surveys, logs, etc.) or machines (weather data, road vision, etc.), and could be in different formats (text, audio, video, augmented or virtual reality, etc.). We will also treat data science as an independent field by itself rather than a subset of another domain, such as statistics or computer science. This will become clearer as we look at how data science relates to and differs from various fields and disciplines later in this chapter.

Why is data science so important now? Dr. Tara Sinclair, the chief economist at indeed.com since 2013, said, "the number of job postings for 'data scientist' grew 57%" year-over-year in the first quarter of 2015.[3] Why have both industry and academia recently increased their demand for data science and data scientists? What changed within the past several years? The answer is not surprising: we have a lot of data, we continue to generate a staggering amount of data at an unprecedented and ever-increasing speed, analyzing data wisely necessitates the involvement of competent and well-trained practitioners, and analyzing such data can provide actionable insights.

The "3V model" attempts to lay this out in a simple (and catchy) way. These are the three Vs:

1. Velocity: The speed at which data is accumulated.
2. Volume: The size and scope of the data.
3. Variety: The massive array of data and types (structured and unstructured).

Each of these three Vs regarding data has dramatically increased in recent years. Specifically, the increasing volume of heterogeneous and unstructured (text, images, and video) data, as well as the possibilities emerging from their analysis, renders data science ever more essential. Figure 1.1[4] shows the expected volumes of data to reach 40 zettabytes (ZB) by the end of 2020, which is a 50-fold increase in volume than what was available at the beginning of 2010. How much is that really? If your computer has 1 terabytes (TB) hard drive (roughly 1000 GB), 40 ZB is 40 billion times that. To provide a different perspective, the world population is projected to be close to 8 billion by the end of 2020, which means, if we think about data per person, each individual in the world (even the newborns) will have 5 TB of data.

## 1.2 Where Do We See Data Science?

The question should be: Where do we not see data science these days? The great thing about data science is that it is not limited to one facet of society, one domain, or one department of a university; it is virtually everywhere. Let us look at a few examples.

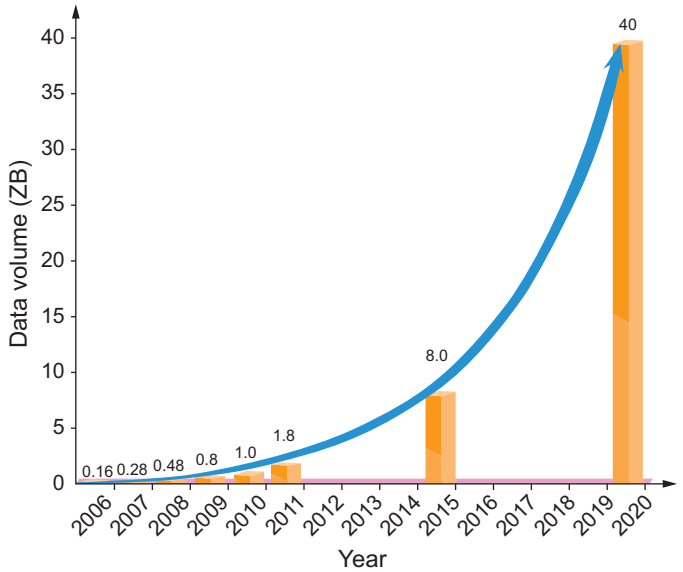


**Figure 1.1** Increase of data volume in last 15 years. (Source: IDC's Digital Universe Study, December 2012.[5])

### 1.2.1 Finance

There has been an explosion in the velocity, variety, and volume (that is, the 3Vs) of financial data, just as there has been an exponential growth of data in almost most fields, as we saw in the previous section. Social media activity, mobile interactions, server logs, real-time market feeds, customer service records, transaction details, and information from existing databases combine to create a rich and complex conglomeration of information that experts (*cough, cough*, data scientists!) must tackle.

What do financial data scientists do? Through capturing and analyzing new sources of data, building predictive models and running real-time simulations of market events, they help the finance industry obtain the information necessary to make accurate predictions.

Data scientists in the financial sector may also partake in fraud detection and risk reduction. Essentially, banks and other loan sanctioning institutions collect a lot of data about the borrower in the initial "paperwork" process. Data science practices can minimize the chance of loan defaults via information such as customer profiling, past expenditures, and other essential variables that can be used to analyze the probabilities of risk and default. Data science initiatives even help bankers analyze a customer's purchasing power to more effectively try to sell additional banking products.[6] Still not convinced about the importance of data science in finance? Look no further than your credit history, one of the most popular types of risk management services used by banks and other financial institutions to identify the creditworthiness of potential customers. Companies use machine learning algorithms in analyzing past spending behavior and

patterns to decide the creditworthiness of customers. The credit score, along with other factors, including length of credit history and customer's age, are in turn used to predict the approximate lending amount that can be safely forwarded to the customer when applying for a new credit card or bank loan.

Let us look at a more definitive example. Lending Club is one of the world's largest online marketplaces that connects borrowers with investors. An inevitable outcome of lending that every lender would like to avoid is default by borrowers. A potential solution to this problem is to build a predictive model from the previous loan dataset that can be used to identify the applicants who are relatively risky for a loan. Lending Club hosts its loan dataset in its data repository (https://www.lendingclub.com/info/download-data.action) and can be obtained from other popular third-party data repositories[7] as well. There are various algorithms and approaches that can be applied to create such predictive models. A simple approach of creating such a predictive model from Lending Club loan dataset is demonstrated at KDnuggets[8] if you are interested in learning more.

### 1.2.2 Public Policy

Simply put, public policy is the application of policies, regulations, and laws to the problems of society through the actions of government and agencies for the good of a citizenry. Many branches of social sciences (economics, political science, sociology, etc.) are foundational to the creation of public policy.

Data science helps governments and agencies gain insights into citizen behaviors that affect the quality of public life, including traffic, public transportation, social welfare, community wellbeing, etc. This information, or data, can be used to develop plans that address the betterment of these areas.

It has become easier than ever to obtain useful data about policies and regulations to analyze and create insights. The following open data repositories are examples:

(1) US government (https://www.data.gov/)
(2) City of Chicago (https://data.cityofchicago.org/)
(3) New York City (https://nycopendata.socrata.com/)

As of this writing, the data.gov site had more than 200,000 data repositories on diverse topics that anyone can browse, from agriculture to local government, to science and research. The City of Chicago portal offers a data catalog with equally diverse topics, organized in 16 categories, including administration and finance, historic preservation, and sanitation. NYC OpenData encompasses datasets organized into 10 categories. Clicking on the category City Government, for instance, brings up 495 individual results. NYC OpenData also organizes its data by city agency, of which 94 are listed, from the Administration for Children's Services to the Teachers Retirement System. The data is available to all interested parties.

A good example of using data to analyze and improve public policy decisions is the Data Science for Social Good project, where various institutions including Nova SBE, Municipality of Cascais, and the University of Chicago will participate in the program for three months, and which will bring together 25 data analytics experts from several

countries who will be working on using the open public policy dataset to find clues to solve relevant problems with impact on society, such as: how does an NGO use data to estimate the size of a temporary refugee camp in war zones to organize the provision of help, how to successfully develop and maintain systems that use data to produce social good and inform public policy, etc. The project usually organizes new events in June of every year.[9]

### 1.2.3 Politics

Politics is a broad term for the process of electing officials who exercise the policies that govern a state. It includes the process of getting policies enacted and the action of the officials wielding the power to do so. Much of the financial support of government is derived from taxes.

Recently, the real-time application of data science to politics has skyrocketed. For instance, data scientists analyzed former US President Obama's 2008 presidential campaign success with Internet-based campaign efforts.[10] In this *New York Times* article, the writer quotes Ariana Huffington, editor of *The Huffington Post*, as saying that, without the Internet, Obama would not have been president.

Data scientists have been quite successful in constructing the most accurate voter targeting models and increasing voter participation.[11] In 2016, the campaign to elect Donald Trump was a brilliant example of the use of data science in social media to tailor individual messages to individual people. As Twitter has emerged as a major digital PR tool for politics over the last decade, studies[12] analyzing the content of tweets from both candidates' (Trump and Hillary Clinton) Twitter handles as well as the content of their websites found significant difference in the emphasis on traits and issues, main content of tweet, main source of retweet, multimedia use, and the level of civility. While Clinton emphasized her masculine traits and feminine issues in her election campaign more than her feminine traits and masculine issues, Trump focused more to masculine issues, paying no particular attention to his traits. Additionally, Trump used user-generated content as sources of his tweets significantly more often than Clinton. Three-quarters of Clinton's tweets were original content, in comparison to half of Trump's tweets, which were retweets of and replies to citizens. Extracting such characteristics from data and connecting them to various outcomes (e.g., public engagement) falls squarely under data science. In fact, later in this book we will have hands-on exercises for collecting and analyzing data from Twitter, including extracting sentiments expressed in those tweets.

Of course, we have also seen the dark side of this with the infamous Cambridge Analytica data scandal that surfaced in March 2018.[13] This data analytics firm obtained data on approximately 87 million Facebook users from an academic researcher in order to target political ads during the 2016 US presidential campaign. While this case brought to public attention the issue of privacy in data, it was hardly the first one. Over the years, we have witnessed many incidents of advertisers, spammers, and cybercriminals using data, obtained legally or illegally, for pushing an agenda or a rhetoric. We will have more discussion about this later when we talk about ethics, bias, and privacy issues.

### 1.2.4 Healthcare

Healthcare is another area in which data scientists keep changing their research approach and practices.[14] Though the medical industry has always stored data (e.g., clinical studies, insurance information, hospital records), the healthcare industry is now awash in an unprecedented amount of information. This includes biological data such as gene expression, next-generation DNA sequence data, proteomics (study of proteins), and metabolomics (chemical "fingerprints" of cellular processes).

While diagnostics and disease prevention studies may seem limited, we may see data from or about a much larger population with respect to clinical data and health outcomes data contained in ever more prevalent electronic health records (EHRs), as well as in longitudinal drug and medical claims. With the tools and techniques available today, data scientists can work on massive datasets effectively, combining data from clinical trials with direct observations by practicing physicians. The combination of raw data with necessary resources opens the door for healthcare professionals to better focus on important, patient-centered medical quandaries, such as what treatments work and for whom.

The role of data science in healthcare does not stop with big health service providers; it has also revolutionized personal health management in the last decade. Personal wearable health trackers, such as Fitbit, are prime examples of the application of data science in the personal health space. Due to advances in miniaturizing technology, we can now collect most of the data generated by a human body through such trackers, including information about heart rate, blood glucose, sleep patterns, stress levels and even brain activity. Equipped with a wealth of health data, doctors and scientists are pushing the boundaries in health monitoring.

Since the rise of personal wearable devices, there has been an incredible amount of research that leverages such devices to study personal health management space. Health trackers and other wearable devices provide the opportunity for investigators to track adherence to physical activity goals with reasonable accuracy across weeks or even months, which was almost impossible when relying on a handful of self-reports or a small number of accelerometry wear periods. A good example of such study is the use of wearable sensors to measure adherence to a physical activity intervention among overweight or obese, postmenopausal women,[15] which was conducted over a period of 16 weeks. The study found that using activity-measuring trackers, such as those by Fitbit, high levels of self-monitoring were sustained over a long period. Often, even being aware of one's level of physical activities could be instrumental in supporting or sustaining good behaviors.

Apple has partnered with Stanford Medicine[16] to collect and analyze data from Apple Watch to identify irregular heart rhythms, including those from potentially serious heart conditions such as atrial fibrillation, which is a leading cause of stroke. Many insurance companies have started providing free or discounted Apple Watch devices to their clients, or have reward programs for those who use such devices in their daily life.[17] The data collected through such devices are helping clients, patients, and healthcare providers to better monitor, diagnose, and treat health conditions not possible before.

### 1.2.5 Urban Planning

Many scientists and engineers have come to believe that the field of urban planning is ripe for a significant – and possibly disruptive – change in approach as a result of the new methods of data science. This belief is based on the number of new initiatives in "informatics" – the acquisition, integration, and analysis of data to understand and improve urban systems and quality of life.

The Urban Center for Computation and Data (UrbanCCD), at the University of Chicago, traffics in such initiatives. The research center is using advanced computational methods to understand the rapid growth of cities. The center brings together scholars and scientists from the University of Chicago and Argonne National Laboratory[18] with architects, city planners, and many others.

The UrbanCCD's director, Charlie Catlett, stresses that global cities are growing quickly enough to outpace traditional tools and methods of urban design and operation. "The consequences," he writes on the center's website,[19] "are seen in inefficient transportation networks belching greenhouse gases and unplanned city-scale slums with crippling poverty and health challenges. There is an urgent need to apply advanced computational methods and resources to both explore and anticipate the impact of urban expansion and find effective policies and interventions."

On a smaller scale, chicagoshovels.org provides a "Plow Tracker" so residents can track the city's 300 snow plows in real time. The site uses online tools to help organize a "Snow Corps" – essentially neighbors helping neighbors, like seniors or the disabled – to shovel sidewalks and walkways. The platform's app lets travelers know when the next bus is arriving. Considering Chicago's frigid winters, this can be an important service. Similarly, Boston's Office of New Urban Mechanics created a SnowCOP app to help city managers respond to requests for help during snowstorms. The Office has more than 20 apps designed to improve public services, such as apps that mine data from residents' mobile phones to address infrastructure projects. But it is not just large cities. Jackson, Michigan, with a population of about 32,000, tracks water usage to identify potentially abandoned homes. The list of uses and potential uses is extensive.

### 1.2.6 Education

According to Joel Klein, former Chancellor of New York Public Schools, "when it comes to the intersection of education and technology, simply putting a computer in front of a student, or a child, doesn't make their lives any easier, or education any better."[20] Technology will definitely have a large part to play in the future of education, but how exactly that happens is still an open question. There is a growing realization among educators and technology evangelists that we are heading toward more data-driven and personalized use of technology in education. And some of that is already happening.

The Brookings Institution's Darrell M. West opened his 2012 report on big data and education by comparing present and future "learning environments." According to West, today's students improve their reading skills by reading short stories, taking a test every other