DATA SCIENTIST

The Definitive Guide to Becoming a Data Scientist

ZACHARIAS VOULGARIS, PHD

DATA SCIENTIST

The Definitive Guide to Becoming a Data Scientist

first edition

Zacharias Voulgaris, PhD

Published by:
Technics Publications, LLC
2 Lindsley Road
Basking Ridge, NJ 07920 USA

http://www.TechnicsPub.com

Cover design by Mark Brye
Cartoons by Sarah Silverberg

Edited by Carol Lehn

ISBN, print ed. 978-1-935504-69-6
ISBN, Kindle ed. 978-1-935504-74-0
ISBN, ePub ed. 978-1-935504-75-7

First Printing 2014

Library of Congress Control Number: 2014935091

Table of Contents

Introduction

A year and a half ago, I had no clear idea what a data scientist was and why it was an important role. Immersed in a dead-end job in an e-marketing company, I had started to forget all of the stuff I had learned through the many difficult years of my education. I am not sure what triggered my resolve to look into the matter more (at that time there were no decent books on the topic, and I had no-one to mentor me), but I do remember coming to the realization that this was my life's vocation. Naturally, there were problems with this new type of work – lots of things I hadn't learned and no idea of how to learn them, especially if you factor in my 50 hours per week schedule and the fact that there wasn't a decent data science course anywhere in the country in which I was living. But I did power through, my resolve fueled by the conviction that this was something worthwhile and enjoyable. And if I happened to fail in my pursuit, at least I would have picked up some useful skills in the process.

This book is for people who have the same desire to learn about this fascinating field. When I started my quest into the data science world, I had to learn the hard way, through trial and error, as well as through hard research via articles, videos and other sources on the Web. Fortunately, it will be much easier for you. That's why I wrote this book: so that you have a manual, of sorts, to provide you with guidelines for this challenging transition.

Data science is a very rewarding field that deals with a fascinating new entity in the data world: big data, something that constitutes a quite intriguing challenge since there is no straightforward way of dealing with it effectively. This leaves a lot of room for creativity

and a wider array of possibilities that you are called to explore as a data scientist. In addition, through this role you have the opportunity to develop aspects of yourself that no other role in the IT field provides: namely creativity, communication, direct links with the business world, etc. Through all this you have a chance of providing something useful to the organization you work for (which can be a company, government agency, or even a charity) through the intelligent use of the data that is available. Since this data is bound to be large, diverse, and quite messy, it is not something you would normally find in a tidy database. Hence the term big data and the role of the data scientist, the professional who deals with big data in a scientific, creative and understandable manner.

Over the past few years, there has been heightened awareness of big data and its implications in business, as well as its impact on the job market. But what is big data exactly? And how is it different from traditional data? The short definition of big data is "data that cannot be handled by a single computer." Although this is usually due to its very large size, there are a few other reasons. In general, it is defined by four main characteristics, usually referred to as the four Vs of big data:

- Volume. Contrary to "normal" data, big data is significantly larger; i.e., it ranges from a few Terabytes (TB) to a few Zettabytes (ZB). The latter is a billion TB, or a trillion Gigabytes (GB). That's a lot of data! In 2010, the data of the whole world was about 1 ZB – that's 125 million 8 GB media players! What's more, this number has been increasing rapidly over the past few years, and there is no sign of it stopping any time soon. This very high amount of data that characterizes big data, in combination with the fact that big data cannot be processed efficiently using a single machine (even a supercomputer), has brought about the use of parallel computing (a cluster of computers working together via a network connection), something

that is inherent in the vast majority of data science projects.

- **Variety**. Big data is also quite varied, coming from non-traditional as well as traditional sources. The data we are used to processing is *structured* data, the kind of data usually found in databases. We know what its data type and size are, and we generally know what's supposed to be in each field. Big data, however, includes *unstructured* and *semi-structured* data as well. Unstructured data lacks a pre-defined structure in its subcomponents (e.g., data found in Facebook posts, tweets, phone call transcripts, etc.), while semi-structured has some structure and is something in between structured and unstructured (e.g., data in machine logs and email address headers).

- **Velocity**. Another important characteristic of big data is velocity, or the rate at which it arrives at the enterprise and is processed. Traditional data is thought to be slower and fairly static in terms of how it is developed and transferred from the location it is generated to the location it is processed. Contrast this with big data, which is constantly moving, and moving fast (though there may be some exceptions to this rule). This means that it needs to be processed quickly, in real-time if possible, in order to harness its potential. For example, a financial services company may need to analyze over 5 million market messages every second, with a latency of about 30 microseconds.

- **Veracity.** This last one was added relatively recently, so there are still many references to the three Vs of big data in books and articles on the topic. Big data is also characterized by veracity, an attribute that relates to the quality (trustworthiness) of the data. As one would expect, there is a lot of noise in all of this data. Working with big

> data effectively means being able to discern the noise from the signals that may hide within. This is a challenging process that requires advanced analytical techniques. If one is not careful, it is easy to draw conclusions backed by statistical significance that don't have any real value, or that may lead to questionable decisions.

There are two more Vs that are sometimes included, *Variability* and *Visibility*, but there has not been consensus on these characteristics, yet.

It doesn't take much to realize that making effective use of big data is a challenge. Ignoring it is no longer an option in many industries as its information potential is becoming more and more evident and ways to make use of it constantly increase. Think of Amazon and Netflix, for example. Their clever use of big data has given them a competitive advantage and has opened new roads for their industries. If you were in the online shopping business, for example, and you had a large customer base that supplied you with large amounts of data, imagine what you could learn about buying patterns, the demographics of your customers, and the opportunities you could take advantage of by analyzing the data.

Building on this newly acquired knowledge, you could go one step further: namely, design a widget or an app that makes use of the insights you have derived and helps its user to gain similar insights into *their* experience with the environment of the data (in this case, the online shop). That's actually one of the reasons Amazon became so successful. It not only offered a large variety of products to its users, but made the whole experience of shopping easier and more enjoyable through the use of interesting features on its site, such as its recommender system. This and many other similar mini-programs that are based on intelligent analysis of big data are usually referred to as data products and constitute the goal of the majority of data scientists. There are data scientists, however, who are not directly involved in the creation of these products and

focus on engineering ways of facilitating other data scientists in their work. So the field is quite diverse in the particular tasks data scientists can undertake through the application of their specific skill-sets.

So the question is not whether or not to hop on the big data wagon, but *how*. This is where the data scientist comes in. The data scientist is a fairly new role in the industry, and since its introduction to the job market, it has grown in popularity. It involves all the different aspects of dealing with data, particularly big data, in an intelligent and very methodical manner, in order to create a useful product (the aforementioned data product). The product is usually a widget or an app that can provide meaningful information the users do not already know (the last part is something that is stressed by John Foreman, a very successful and experienced data scientist). Big data has brought about new paradigms in data processing and data visualization, equipping the data scientist with powerful tools that require a different mindset and a different skill-set to accompany it.

Many people confuse the data scientist with the data analyst. However, they are quite different roles, much like space flight is different from traditional flight. A data analyst uses techniques that may work with data that borders on being big data, but may be inefficient and lack the flexibility of the techniques employed by a data scientist. The former relies on a series of pre-made models to derive useful information from the data and creates reports for a businessperson to view. The latter often develops his own models or uses a completely data-driven approach in his analyses, often resulting in something that many other people can use, not just a businessperson in his company. The data analyst will create intuitive plots in his reports. The data scientist will create an interactive dashboard that will plot all the essential information in real-time.

In other words, data analysis is a very useful tool, but if one is to make use of the data the world is immersed in today, one needs to not only be efficient with data analysis techniques, but also gain a working knowledge of other aspects of data science that will be described in this book. Being a data analyst is great, but it will limit you to a certain type of datasets that involve structured data only, and among these datasets you will only be able to deal with the relatively small ones. If you want to take a stab at the larger and more complicated ones, you'll need to learn the ways of the data scientist.

Being a data scientist is not only about know-how, though; to someone who's interested, it can also be a very enjoyable and intriguing occupation. The domain of the data scientist is constantly changing as new technologies are developed, making it a very dynamic field. He[1] is at the cutting edge of science and gets to communicate with interesting people, some of whom drive these changes. Data science is an inter-disciplinary field, so the data scientist expands his worldviews by learning to think in a more systemic way, integrating things from various fields. Most importantly, he often gets to be creative in the way he deals with the problems that arise and the ways data can be processed.

Being a data scientist is also a great profession. For example, given that it is a new role that can provide a strategic advantage to an organization (and there aren't many people trained to do the role properly), the data scientist can be very well paid, usually more than other IT professionals, according to Indeed.com, for the same years of experience. In addition, a data scientist has the opportunity to develop a wide variety of skills, making him a very versatile and adaptable professional who may have the opportunity to communicate with all kinds of people in the industry and the

[1] Although I use "he" to refer to a data scientist throughout the text, the role can be undertaken by both men and women.

scientific world and work in different industry sectors. This is particularly useful in times of financial turmoil, when job-hunting becomes challenging for specialized professionals.

This book is comprised of eighteen chapters, covering the basic aspects of the transition to the data science world. In the first few chapters you will learn more about what the field entails (what data science and big data are; why data science is very important, especially nowadays; and the different types of data scientists). Afterwards, you will have a chance to learn about what it takes to be a data scientist (the data scientist's mindset, his technical qualifications, the experience that is required for this role, and a few things about networking). Next, you will have an opportunity to learn about the everyday life of a data scientist (what software he uses, the importance of learning new things in this line of work, the kind of problems he encounters, and the main stages of the data science process). In the chapter that follows, you will be presented with the various migration paths from existing roles (what to do and learn if you are a programmer/software developer, if you are a statistician or machine learning practitioner, if you are a data-related professional, or if you are a student). Afterwards, you will be given some practical and down-to-earth advice on what you need to do to land your first data science job (where to look, how to present yourself as a would-be data scientist, and what you need to consider if you wish to follow the freelance track). Finally, you will have a chance to read about some real-world data scientists, their experiences and their views on the matter, as well as some real job posting examples for data scientist positions. At the end of the book, there is a glossary of the most important terms that have been introduced, as well as three appendices – a list of useful sites, some relevant articles on the Web, and a list of offline resources for further reading. There is also a comprehensive index at the end of this text.

Throughout the book, the Kea bird is used to represent the data scientist. The Kea is known for its intelligence, innovative attitude, curiousness, and is one of the rarest species of its category. These attributes are the discerning features of the Kea and are shared by the data science professional.

I sincerely hope that this book is useful and, perhaps, even enjoyable for you. The transition itself is quite demanding (especially if you are in the beginning of your professional life), but it is an intriguing and rewarding experience. And when you eventually become a data scientist, the field continues to be just as interesting. Not a role for the faint-hearted, being a data scientist is a wonderful experience on many levels and can be a fascinating journey. Are you ready to embark on it?

Dr. Zacharias Voulgaris

Chapter 1
Data Science and Big Data

Data science is a response to the difficulties of working with big data and other data analysis challenges we collectively face today. We examined this briefly in the introduction, but that was just scratching the surface. In fact, there is so much literature on big data that this whole chapter will still not be able to do it justice. It will, however, give you a good idea of its importance in today's world. Furthermore, it will help you understand what all the hype is about big data (a hype that has increased significantly over the past year), and why data science is so important.

Big data is a fundamental asset for today's businesses, and it is not a coincidence that the majority of businesses today are using, or are in the process of adopting, the corresponding technology. Despite all the hype about it in various media, this is not a fad. There are specific advantages to using this asset, and the fact that it is growing more abundant is an indication that it is imperative to do something about it, and do it fast! Perhaps it is not useful for certain industries right now as big data tends to be quite chaotic or even non-existent for them. Those who do have it and make intelligent use of it, though, reap its benefits and stand a good chance of being more successful in today's competitive economic ecosystems.

1.1 Digging into Big Data

Big data is abundant and contains information that is relevant to the business problems at hand. If you are a manager of an e-commerce company, for example, the data you collect on your servers regarding your customers and the visitors to your site are

rich with information that, when analyzed properly, can be used to increase your sales, enhance your site's design, and improve your customer service. It can also provide you with ideas on marketing strategies and ways to improve your company's overall strategy; all that from a bunch of ones and zeroes that dwell on your servers. You just need to extract the information from them, allocating a small part of your resources. Not a bad trade-off, for sure. We'll come back to this example later on.

Not every amalgamation of data qualifies for the term big data, although most Web-related data falls under this umbrella. This is because big data is characterized by the four Vs[2].

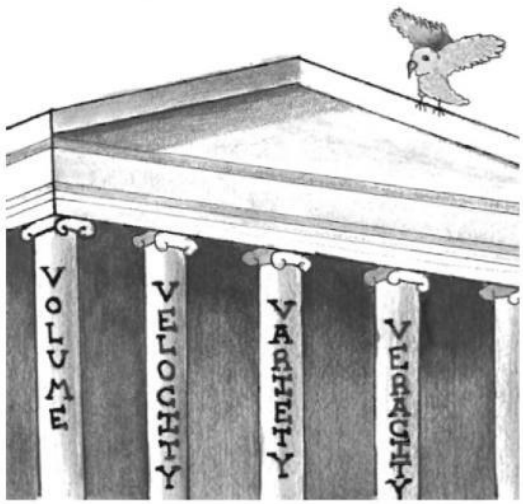

Fig. 1.1 The four Vs of big data.

As we have already seen, these are:

- **Volume** – Big data consists of large quantities of data. This translates into several TB up to a few ZB. This data may be distributed across various locations, often in several computer

[2] Actually, some people include an additional two Vs, variability and visibility, which refer to the fact that Big Data changes over time and is hardly visible to users.

networks connected through the Internet. Generally, any amount of data that is too large to be processed by a single computer satisfies the Volume criterion of big data. This alone is an issue that requires a different approach to data processing, something that gave rise to the parallel computing technology known as MapReduce.

- Velocity – Big data is also in motion, usually at high transfer speeds. It is often referred to as data streams, which are frequently too difficult to archive (the speed alone is a great issue, considering the limited amount of storage space a computer network has). That is why only certain parts of it are collected. Even if it were possible to collect all of it, it would not be cost effective to store big data for long, so the collected data is periodically jettisoned in order to save space, keeping only summaries of it (e.g., average values and variances). This problem is expected to become more serious in the near future as more and more data is being generated at higher and higher speeds.

- Variety – In the past, data used to be more or less homogeneous, which also made it more manageable. This is not the case with big data, which stems from a variety of sources and, therefore, varies in form. This translates into different structures among the various data sources and the existence of semi-structured and completely unstructured data. Structured data is what is found in traditional databases, where the structure of the content of the data is predefined in fields of specified sizes. Semi-structured data has some structure, but it is not consistent (see the contents of a .JSON file, for example), making it difficult to work with. Even more challenging is unstructured data (e.g., plain text) that has no structure whatsoever. In most cases big data is semi-structured, though rarely do its sources share the same form.

In the past few years, unstructured and semi-structured data have constituted the vast majority of all big data.

- **Veracity** – This is one aspect of big data that is often neglected by the literature, partly because it is relatively new although equally important. It has to do with how reliable the data is, something that is taken into account in the data science process (which is different from the traditional data analysis process, as we will see in Chapter 11). Veracity involves the signal-to-noise ratio; i.e., figuring out what in the big data is valid for the business, which is an important concept in information theory. Big data tends to have varied veracity as not all of its sources are equally reliable. Increasing the veracity of the available data is a major big data challenge.

Note that a piece of data may have one or more of these characteristics and still not be classified as big data. Big data has all four of these. Big data is a serious issue as it is not easy, even for a supercomputer, to manage it effectively, let alone perform a useful analysis of it.

In the example we started with, a typical set of data that you would encounter would have the following qualities:

- The volume of data would be very large, with a tendency to become larger, especially if your site monitors several aspects of its visitors' behavior. This data may easily account for several TB a year.
- It would flow constantly as visitors come and go and new visitors pay a visit to your site. This translates to continuous network activity on your servers, which is basically a data stream from the Web flowing into your server logs.
- The data you would collect from your visitors would vary greatly, ranging from simple Web statistics (time spent on each page, time of the visit, number of pages visited, etc.) to text entered on the site (assuming you have some kind of review

system, like most serious e-commerce sites) and several other types of data (e.g., ratings from customers for various products, transaction data, etc.).

- Naturally, not everything you observe on your site's servers will be trustworthy. Some of your visitors may be bots sent by hackers or other users for shady purposes, while other visitors may be your competitors spying on you! Some visitors may have spelling errors in their reviews, or leave random or spam messages on the site for whatever reason. Even if you have some kind of filtering system, it is inevitable that your site will collect some useless data over time.

Based on all of the above observations, do you think that you are dealing with big data in this company or not? Why? If you have understood the above concepts, you should be confident in replying positively to this question. Each one of the bullet points describing the data situation in that company has to do with one of the Vs of big data.

1.2 Big Data Industries

Naturally, not all industries are equally affected by the big data movement. Depending on how much they rely on data and how profitable information is to them, they may be looking at a goldmine or one more asset that can wait. Based on recent statistics, the following industries appear to have benefited, or are inclined to benefit the most from big data:

- Retail (particularly in terms of productivity boost)
- Telecommunications (particularly in terms of revenue increase)
- Consulting
- Healthcare
- Air transportation
- Construction
- Food products

- Steel and manufacturing in general
- Industrial instruments
- Automobile industry
- Customer care
- Financial services
- Publishing
- Logistics

Note that the benefit is not always directly related to the bottom line, but it is definitely of significant business value. For example, by employing big data technologies in healthcare, physicians can use previous data to gain a better understanding of the patients' issues, yielding a better diagnosis and enabling them to take better care of their patients in general. This can eventually result in greater efficiencies in the medical system, translating into lower costs through the intelligent use of medical information derived from that data.

Another example comes from customer care, where big data can help leverage bad customer experiences. By effective use of big data technologies, companies can gain a better understanding of what their customers like and don't like in near real-time. This can help them amend their strategies in dealing with these customers and give them insight into how to improve their services in the future.

Note that there are many other industries that have the potential for gaining from big data, but based on their current status, it is not a worthwhile option for them. For example, the art industry is still not big on big data, since the data involved in this field is limited to descriptions of artwork and, in some cases, digitized forms of these works of art. However, it is possible that this may change in the future depending on how the artists act. For example, if a certain gallery makes use of sensors monitoring the number of people who view a certain painting, and in combination with other data (e.g., number of people who bought tickets to the various exhibitions

that hosted that painting), they could gradually build a large database that would contain data about the sensor readings, the ticket sales, and even the comments some people leave on the gallery's blog about the various paintings. All this can potentially yield useful information about which pieces of art are more popular (and by how much), as well as what the optimum ticket prices should be for the gallery's exhibitions throughout the year.

All this is great, but how is it of any real use to you? Well, higher profit margins and the potential to significantly boost productivity are not going to happen on their own. It is naïve to think that just installing a big data package and assigning it to an employee (even if they are a skilled employee) could result in measurable gains. In order to take advantage of big data, a company needs to hire qualified people who can undertake the task of turning this seemingly chaotic bundle of data into useful (actionable) information. This is the problem that all data scientists are asked to solve and one of the driving forces of all developments in the field that came to be known as data science.

1.3 Birth of Data Science

The field of data science resulted from the attempt to discover potential insights residing in big data and overcoming the challenges that were reflected in the four Vs described previously. This was possible through the combination of various technological advances of modern computing. Specifically, parallel computing, sophisticated data analysis processes (mainly through machine learning), and powerful computing at lower prices made this feasible. What's more, the continuously accelerating progress of the IT infrastructure and technology will enable us to generate, collect, and process significantly more data in the not-so-distant future. Through all this, data science addresses the issues of big data on a technical level through the application of the intelligence and creativity that is employed in the development and use of these technologies. That is, big data is somewhat manageable and

at least able to provide some useful information to make the whole process worthwhile.

It's important to note that data science is not a fad, but something that is here to stay and bound to evolve rapidly. If you were an IT professional when the World Wide Web came about, you might have seen it as a luxury or a fad that wouldn't catch on, but those who managed to see its real value and the potential it held made very lucrative careers out of it. Imagine being one of the first people to learn HTML, CSS and JavaScript, or one of the first to create digital graphics to be used for websites. It would be like holding a winning lottery ticket, especially if you were good at your job. This is the situation with data science today. It would probably not be so well-known if it weren't for so many people writing about its benefits. Still, most professionals and many students are not aware of what data science really means.

If you assimilate the aforementioned facts about big data, you will understand that data science is the solution to a real problem that is only going to become more pronounced in the years to come. This problem, as mentioned earlier, is reflected in the four Vs of big data, the characteristics that make it difficult to deal with using conventional technologies. As technology is on its side, data science is bound to become more robust and more diverse in the coming decade or so. There are already some post-graduate programs making an appearance in the academic world[3], and there are plenty of respectable researchers writing papers on data science topics. This is not a coincidence. It shows a trend for the development of an infrastructure of knowledge and know-how that will nourish this field.

---

[3] One of them, created by Berkley, costs around $60,000, which is significantly more than the high-priced MBAs you see elsewhere. This is a clear indication that people in the academic world as well as in the industry are taking data science quite seriously.

It is not very clear exactly when data science was born (there have been people working on this field as researchers for several decades), but the first conference where it received the spotlight was in 1996 ("Data Science, Classification, and Related Methods" by IFCS). It wasn't until September 2005, however, when the term "data scientist" first appeared in the literature. Specifically, in a report released that year[4], data scientists were defined as "the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection." In June, 2009, the importance of the role of the data scientist became more apparent, as Nathan Yau's article "Rise of the Data Scientist" in *FlowingData* was written[5]. Since then, references to and literature on data science have increased rapidly. Just take a look at how many conferences are being organized for it nowadays, appealing to both academics and people in the industry! What's more, as several large companies that are leaders in their sectors (e.g., Amazon) make use of data science in their everyday workflow, it is quite likely that this trend will continue. Also, as the role of the data scientist adapts to the ever-changing requirements of the data world, it has come to include several things such as the application of state-of-the-art data analysis techniques, not just the original responsibilities.

1.4 Key Points

- Big data is a recent phenomenon where there is a large quantity of data, in quick motion, varying from structured to unstructured (with everything else in between), and with

---

[4] Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century, available at http://www.nsf.gov/pubs/2005/nsb0540

[5] The article is still available online at the time of this writing. You can access it at http://flowingdata.com/2009/06/04/rise-of-the-data-scientist

different reliability levels. This is often referred as the four Vs of big data: Volume, Velocity, Variety, and Veracity.

- Dealing with big data is a challenging problem due to these four Vs. Data science is our response to the challenges that big data represents.

- Data scientists are the people that make sense of big data. Through the use of state-of-the-art technologies and know-how, they manage to derive actionable information from it, usually in the form of a data product.

- Big data occurs in a variety of industries; taking advantage of it can have a profound effect on them in terms of productivity boost and revenue increase.

- Data science has been around for over two decades but has only recently taken off as the corresponding technology was developed (parallel computing, intelligent data analysis methods, and powerful computing at a very low cost).

- The role of the data scientist first made an appearance in the literature in 2005, while it started becoming quite popular in 2009. In an article in *Harvard Business Review*, data science was called the "sexiest" profession of the 21st century.[6]

- Data science is expected to continue to grow in terms of business value, technology, available knowledge and know-how, and popularity in the years to come.

---

[6] Davenport, Thomas H., and D. J. Pattil. "Data Scientist: The Sexiest Job of the 21st Century." *Harvard Business Review*, October 2012.

Chapter 2
Importance of Data Science

In the previous chapter, we got a glimpse of how data science came about and how it is related to big data. We also looked into the major milestones of this field and why it has become popular in recent years. However, this was just scraping the surface, since data science has much to offer on many more levels. In order to get a better understanding, we will look into its history, the new paradigms it entails and the new mindset it brings about as well as the changes it brings.

2.1 History of the Data Science Field

The term "data science" was around before big data came into play (just like the term "data" preceded computers by four centuries or so). In 1962, when John W. Tukey[7] wrote his book *The Future of Data Analysis*[8], he foresaw the rise of new type of data analysis that was more of a science than a methodology. In 1974, Peter Naur published a book entitled *Concise Survey of Computer Methods*,[9] in both Sweden and the United States. Although this was merely an overview of the data processing methods of the time, this book contained the first definition of data science as "the science of

[7] Tukey was a remarkable statistician who invented the Tukey honest significance test, which is often used in combination with the well-known ANOVA method. You can find more on the Tukey method at the following site:
http://www.itl.nist.gov/div898/handbook/prc/section4/prc471.htm

[8] John W. Tukey: *The Future of Data Analysis,* Ann. Math. Statist. Volume 33, Number 1, 1962.

[9] Peter Naur: *Concise Survey of Computer Methods*, 397 p. Studentlitteratur, Lund, Sweden, ISBN 91-44-07881-1, 1974.

dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences." So back then, anyone proficient with computers who also understood the semantics of the data to some extent was a data scientist. No fancy tools, no novel paradigms, no new science behind it. It's no surprise that the term took a while to catch on.

As computer technology and statistics started to converge later that decade, Tukey's vision began to materialize, albeit quite subtly. It wasn't until the late 1980s, though, that it started to gain ground through one of data science's most well-known methods: data mining. As the years advanced, the scientific processing of data rose to new heights, and data science came into the spotlight of academic research through a conference in 1996 called "Data Science, Classification, and Related Methods." This conference, which was organized by the International Federation of Classification Societies (IFCS), took place in Kobe, Japan. It made data science more well-known to the circles of researchers and distinguished it from other data analysis terms, such as classification, which are not as broad as data science. This helped gradually make data science an independent field.

In the next year (1997), the *Data Mining and Knowledge Discovery* journal was launched, defining data mining as "extracting information from large databases." This was the first data science method to gain popularity and respect in the scientific community as well as in the industry. This method will be revisited in the data science process in Chapter 11.

The role of data science started to become more apparent at the end of the 1990s as databases grew larger. This was voiced very eloquently by Jacob Zahavi in December 1999 in his article "Mining

Data for Nuggets of Knowledge"[10]: "Conventional statistical methods work well with small data sets. Today's databases, however, can involve millions of rows and scores of columns of data… Scalability is a huge issue in data mining. Another technical challenge is developing models that can do a better job analyzing data, detecting non-linear relationships and interaction between elements… Special data mining tools may have to be developed to address web-site decisions." This depicted very clearly how the need for a new framework of data analysis was imperative, something that aided in the coming about of data science as a field to address that need.

In the 2000s, publications about data science started to appear at an increasing rate, though they were mainly academic. Journals and books on data science became more common and attracted interest among researchers. In September 2005, the term "data scientist" was first defined (albeit somewhat generically) in a government report, as we saw in the previous chapter. Later on, in 2007 the Research Center for Dataology and Data Science was established in Shanghai, China.

2009 was a great year for data science. Yangyong Zhu and Yun Xiong, two of the researchers in the aforementioned research center, declared in their publication "Introduction to Dataology and Data Science,"[11] that data science was a new science, distinctly different from natural science and social science. In addition, in January of that year, Hal Varian (Google's Chief Economist) stated for the press that the next sexy job in the coming decade would be statisticians[12] (a term sometimes used for data scientists when

---

[10] http://knowledge.wharton.upenn.edu/article/mining-data-for-nuggets-of-knowledge

[11] Yangyong Zhu, Yun Xiong. Introduction to Dataology and Data Science. 2009. This paper is available at the website: http://www.dataology.fudan.edu.cn/s/98/t/316/51/0d/info20749.htm

[12] http://www.mckinsey.com/insights/innovation/hal_varian_on_how_the_web_challenges_managers

addressing people who are not entirely familiar with the topic). Finally, in June of that year, Nathan Yau's article "Rise of the Data Scientist"[13] was published on *FlowingData*, making the role of the data scientist much more familiar to the non-academic world.

In the current decade (2010s), data science publications have become abundant, although there is still no decent source of information about how to effectively become a data scientist apart from this book you are reading. The term "data science" gained a more concrete definition, the essence of which was summarized in September 2010 by Drew Conway's Venn diagram (Fig. 2.1).

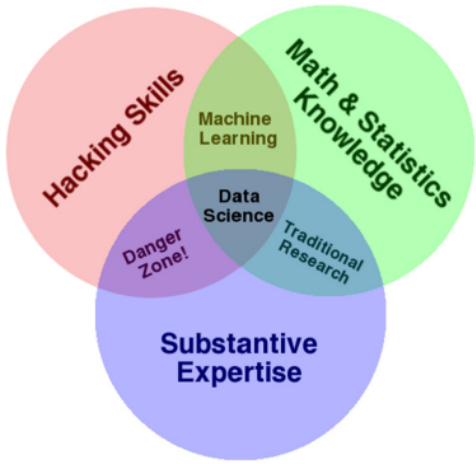

Fig. 2.1 Conway's Venn diagram about Data Science. This diagram illustrates the key components of data science as well as how it differs from the field of machine learning and traditional research. By "danger zone" he probably means the hackers/crackers that compromise the security of many computer systems today. Image source: Drew Conway.

His quote provides further understanding of the fundamentals for becoming a data scientist: "…one needs to learn a lot as they aspire

[13] http://flowingdata.com/2009/06/04/rise-of-the-data-scientist/#comment-30739

to become a fully competent data scientist. Unfortunately, simply enumerating texts and tutorials does not untangle the knots. Therefore, in an effort to simplify the discussion, and add my own thoughts to what is already a crowded market of ideas, I present the Data Science Venn Diagram… hacking skills, math and stats knowledge, and substantive expertise."[14]

Finally, in September of 2012, Hal Varian's quote about this decade's sexy job grew into a whole article in *Harvard Business Review* ("Data Scientist: The Sexiest Job of the 21st Century"[15]) making an even larger population aware of the importance of the role of the data scientist in the years to come.

It is noteworthy that parallel to these publications and conferences, there has been a lot of online social activity in terms of data science. The first official data science group was created on LinkedIn in June 2009 (known as Data Scientists group[16]), and currently also has an independent site (datascientists.net as well as datascientists.com, its original name). Other data science groups have been available online since 2008, although as of 2010, their number has risen at an increasing rate along with online postings for data scientist jobs. This will be covered in a bit more detail in Chapter 13. It should also be noted that over the past few years, there have been a lot of non-academic conferences on data science. These conferences are usually rich in workshops and are targeted at data professionals, project managers and executives.

2.2 The New Paradigms

Data science has brought about or popularized some new paradigms that constitute great tools for any data professional. The main ones are:

---

[14] http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

[15] http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

[16] http://www.linkedin.com/groups?home=&gid=2013423&trk=anet_ug_hm

- **MapReduce** – A parallel, distributed algorithm for splitting a complex task into a series of simpler tasks and solving them in a very efficient manner, thus increasing the speed of performing the complex task and lowering the cost of computing resources. Although this algorithm existed before, its wide use in data science has made it more well known.

- **Hadoop Distributed File System (HDFS)** – An open-source platform designed to make use of parallel computing technology, it basically makes dealing with big data manageable by breaking it into smaller chunks that are split over a network of computers.

- **Advanced Text Analytics** – Often referred to as Natural Language Processing (NLP), this is the field of data analysis that involves techniques for processing unstructured textual data to extract useful information and business intelligence from it. Before data science, this field didn't exist at all.

- **Large scale data programming languages (e.g., Pig, R, ECL, etc.)** – Programming languages that work with large datasets (especially big data) in an efficient manner. These were underdeveloped or completely absent before data science appeared.

- **Alternative database structures (e.g., HBase, Cassandra, MongoDB, etc.)** – Databases for archiving, querying and editing big data using parallel computing technologies.

You may be familiar with the New Technology File System (NTFS) employed by every modern Windows OS. This is a fairly satisfactory file system that works without too many problems for most PCs. It would be impossible to use in a network of connected computers, for handling large amounts of data, however; NTFS has a limit of 256 TB, which is insufficient for many big data applications. Unix-based file systems face similar restrictions, which is why when

Hadoop was developed, a new type of file system had to be created, one that was optimal for a computer cluster. HDFS allows the user to view all the files on the cluster and perform some basic operations on them as if they are on a single computer (even if most of these files are scattered across the entire network).

At the heart of Hadoop lies MapReduce, which is the paradigm that enables the network to crunch the data efficiently with limited risk of failure. All the data is replicated in case one of the computers of the cluster (usually called *nodes*) fails. There are a number of supervising nodes that are in charge of scheduling the tasks and managing the data flow. First, all of the data is mapped through a set of cluster nodes referred to as *mappers*. Once it is processed by the mappers, a set of nodes undertakes the task of reducing the resulting processed data into more useful outputs. This set of nodes, referred to as *reducers*, may include mappers that have finished their job as well. Everything is coordinated by the supervising node(s), ensuring that the outputs of every stage are stored securely (in multiple copies) across the cluster. Once the whole process terminates, the outputs are provided to the user. The MapReduce paradigm involves a lot of programming that can be quite tedious. Its big advantage is that it ensures the process finishes relatively quickly, making efficient use of all available resources, while at the same time minimizing the risk of data loss through hardware failure (something quite common for the largest clusters).

Text analytics have been around for a while, but data science introduced some advanced techniques that make the previous techniques seem almost primitive. Modern (advanced) text analytics allow the user to process large amounts of text data, pinpointing patterns in them very quickly while allowing for common problems such as misspelled words, multi-word terms split over a sentence, etc. Advanced text analytics may be able to pinpoint sentiment (!) in social media posts, recognizing if

someone's comments are literal or sarcastic, something that is extremely difficult for a machine to accomplish without the use of these advanced methods. This advancement was made possible via the application of artificial intelligence algorithms in a Hadoop environment.

Large scale data programming languages, such as Pig, R, and ECL, were developed to tackle big data and integrate well with the Hadoop environment (actually, Pig is part of the Hadoop ecosystem). R, which was developed before the advent of big data, underwent a major upgrade that allows it to connect with Hadoop and handle files in HDFS. As programming languages are not too difficult to develop nowadays, it is possible that at the time you are reading this book, other new languages in this category have been developed, so it is good to keep your eyes open. By the end of this decade, it is possible that the current languages will no longer be the first choice for a data scientist (although it is quite likely that R will be around for a while due to its immense user community).

New alternative database structures came about thanks to data science. These structures include Hash Table (e.g., JBoss data grid, Riak), B-Tree (e.g., MongoDB, CouchDB), and Log Structured Merge Tree (e.g., HBase, Cassandra). Unlike traditional databases, these types of schemas are designed for big data, so they are very flexible in how they read/write data records in a database. Each has its own advantages and disadvantages, but they are all better than traditional SQL databases, which fail when the number of records or the number of fields increases beyond a certain level. For example, if you have a very large database (big data warehouse) consisting of a million fields and a billion records, finding a simple maximum value of a given field using a traditional database will take longer than anyone is willing to wait. The same query in a columnar database (e.g., HBase) will take a fraction of a second.

All of these paradigms are based on the notion that a team of computers, in the form of a cluster, work significantly better than

any single (super)computer, given that there are enough members in that team. The innovation lies in the intelligent and customized approaches to planning the essential tasks so that they are efficiently handled by the computer cluster; in essence, optimizing the process of dealing with the problem at hand. It is no coincidence that these paradigms have exhibited increased popularity since their creation and that they continue to evolve rapidly. There is a lot of interest (and money) invested in these technologies; learning them now is bound to pay off in the near future.

2.3 The New Mindset and the Changes It Brings

By now, you've probably figured out that data science is not merely a set of clever tools, methodologies, and know-how. It is a whole new way of thinking about data altogether. Naturally, this paradigm shift brings about certain changes in the way people work on related projects, how they engage with the problems at hand and on how they develop themselves as professionals.

Data science requires us to think more systematically, combining an imaginative approach to problems with solid pragmatism. This translates into a way of thinking that resembles that of a good civil engineer, combining an artistic perspective (through design) with hard-core engineering and time management. Planning is a crucial aspect of working with big data as different ways of doing the same task may have vastly different demands on resources without any significant difference in the results.

The changes this new mindset brings are evident in the way a data scientist functions. The data scientist usually works as part of a varied team consisting of data modelers, businesspeople, and other professionals (depending on the industry). It is very rare to see a data scientist work on his own for long periods of time as a traditional waterfall model programmer would, for instance.

In addition, the data scientist handles problems by taking advantage of current literature, connecting with a variety of professionals who may be more knowledgeable on the problem he is facing, and breaking problems down into manageable sub-problems that he gradually solves.

The skills a data scientist needs to be successful are not uncommon individually. A data scientist should be able to learn new things easily. With the fast pace of development of big data technologies, a data scientist must have an agile mind that is quick to grasp new methods and familiarize itself with new tools.

A data scientist must also be proactive, anticipating things that will be needed in his work, problems that may arise, and anything else that will require his time. Existing methods may need to be fine-tuned or customized for the problem at hand, and changes in the method may be needed.

A data scientist needs to be flexible, adapting easily to a new business domain, new team members, and new tools (the software he uses when starting a job may be quite different from what he ends up using later in that job). He needs to be adept at networking and should understand the value of the skills he is missing so he takes steps to develop them. Overall, almost all of the skills that a data scientist has are highly transferable and applicable to a large variety of situations. As a result, he is a potent professional who can be an asset to any team, especially an IT one.

We will go into how the shift of mindset and the skills required manifest themselves in practice in much more detail in Chapter 4.

2.4 Key Points

- Data science is older than most people think, but it only started gaining ground in the past decade (2000s).

- Drew Conway's well-known Venn diagram, created in September 2010, effectively summarizes the essence of data science.

- Data science has brought about some new paradigms that change the way we deal with data, the main ones being:
    - MapReduce
    - Hadoop Distributed File System (HDFS)
    - Advanced Text Analytics
    - Large scale data programming languages (e.g., Pig, R, ECL, etc.)
    - Alternative database structures (e.g., HBase, Cassandra, MongoDB, etc.)

- Data science's paradigm shift in the way we deal with data caused certain important changes in our lives as data professionals as it brought about a whole new mindset that is essential for dealing with big data.

- The new mindset that data science promotes brings about several changes in the data scientist's professional life and in the way he interacts with others.

Chapter 3
Types of Data Scientists

Just as there are no two snowflakes that are exactly the same, there are also no two data scientists who have identical skill-sets or identical roles. The big data world has a wide variety of problems, causing some natural differentiation in the specific roles that a data scientist may undertake. In addition, the profession has not been properly defined yet, so depending on various aspects of one's background, such as education, the data scientist role can be further differentiated. Based on some research that was done on the topic by a group of scientists (Harlan Harris, Sean Murphy, and Marck Vaisman, who recently published the book *Analyzing the Analyzers*[17]), there are four types of data scientists: data developers, data researchers, data creatives, and data businesspeople. Often encountered among the most experienced professionals of the field is a fifth type, a mixed/generic combination of these. While there is a certain overlap among all of these categories (e.g., they are all familiar with data analysis methodologies, big data technology, and the data science process), they are generally quite different from one another in several ways. Let's examine each one of them in more detail.

3.1 Data Developers

Data developers usually focus on the more technical issues of data management and data analysis. In other words, their day-to-day work involves getting the data from various sources and organizing

[17] Harlan Harris, Sean Murphy, Marck Vaisman: *Analyzing the Analyzers*, O'Reilly June 2013. http://www.oreilly.com/data/free/analyzing-the-analyzers.csp

it in large databases, querying those databases for meaningful results, and analyzing the results to derive useful information from them. Data developers have a tendency to be programmers with strong coding and machine learning skills, since these are the skills that are most essential for this particular specialty. Their business or statistics skills may be relatively immature, depending on their education and work experience. Data developers are ideal for certain parts of the data science work, the bigger picture of which we will examine later on once the specific parts become clear (see Chapter 11). Data developers may not produce the most robust analyses, which is why they usually team up with other data professionals and designers. Still, they provide value for the companies for which they work, and they can always develop the skills they lack through courses, workshops, etc.

Data developers can be found in a variety of industries and are often employed in smaller companies or as part of a data science team in larger companies. People coming from an IT background may tend to become this type of data scientist since it comes naturally to them. They can enhance their skills by taking courses in business and statistics, parallel to acquiring experience in the industry. A data developer is usually found in entry-level (junior) data scientist roles, although he may take up more managerial roles as he develops his skill-set.

3.2 Data Researchers

Data researchers usually come from the academic world, demonstrating a strong background in statistics or any of the sciences that employ statistics (e.g., social sciences). They also tend to have PhDs in a significantly higher proportion than any other types of data scientists. Business skills are usually not their strong suit, but they are excellent analysts. This particular attribute of theirs is great in cases where a lot of groundbreaking work needs to take place (e.g., in the case of an organization that has never

done data science before and/or has no clear idea of what to do with the data it has).

Data researchers are often a very good asset for larger organizations as part of a data science team along with other professionals who complement this type of data scientist by contributing programming and business skills, things that are essential for the creation of useful data products. As data researchers are adept at learning new things, they can quickly pick up additional skills, expanding their skill-set and becoming more flexible professionals if there is a need for it.

3.3 Data Creatives

Data creatives usually have considerable academic experience and are exceptionally good at big data technologies (i.e., software designed for big data governance and analysis), machine learning, and programming. They tend to be devoted users of open source software and boast a broad-based skill-set. This enables data creatives to move with little effort from one role to another, acting like the Swiss Army knives of the data science field. Not the most business savvy of professionals, they are good at doing the day-to-day work of a data scientist but may require help in making others see its value.

Data creatives are a great asset for smaller companies, where flexibility is of fundamental importance in an employee. Yet they can easily work in a larger company, particularly if they team up with more business-oriented professionals. Missing skills can usually be acquired through work experience.

3.4 Data Businesspeople

Data businesspeople are usually the senior data scientists who lead data science teams (which they sometimes build from scratch). They are adept in business skills and are great project managers.

Their focus is mainly on increasing the revenue of a company, and they are concerned with the bigger picture. Nevertheless, they can also be down-to-earth since they have substantial technical expertise.

Data businesspeople tend to be found in larger organizations or their own start-ups. They are great at dealing with other professionals, particularly businesspeople, and often have extensive experience in every aspect of the data science process. This kind of data scientist usually has other data scientists and data professionals working for him and has a project management role in the data science projects in which he is involved.

3.5 Mixed/Generic Type

Mixed/generic data scientists are like data businesspeople but without the broad experience or the intense business focus. They are more balanced than the other types of data scientists and are more likely to grow into the higher echelons of the field faster than the first three types. Their skill-set includes programming, statistics, and business skills, and they are very flexible, much like the data creatives but with better understanding of the business world. Most new data scientists who study data science at a younger age tend to be of this type since they develop their skills in a more holistic manner (something that is reflected in the syllabus of the data science courses).

Mixed/generic data scientists are good for any kind of company, can work very well independently as well as part of a team, and are quite enthusiastic about the field (which is why they have acquired this wide variety of skills). Based on the growing supply of data science courses and the maturing of the field, it is expected that many data scientists in the future will be of this type, even though they may have other types of differentiations, just like programmers today are more balanced and versatile than programmers in the early days of computing.

3.6 Key Points

- There are five different types of data scientists:
    - Data developers
    - Data researchers
    - Data creatives
    - Data businesspeople
    - Mixed/generic

- The data developers are experts in programming, but may lack other parts of the data scientist skill-set. They usually come from the IT industry.

- The data researchers are experts in data analysis techniques and possess state-of-the-art knowledge in machine learning and other fields. They usually have a PhD and have been or are involved in academic research.

- The data creatives are more holistically developed as data science professionals than the other two types, have a bias towards using open-source software, and are very versatile. They come from all kinds of industries, though usually they are computer scientists already.

- The data businesspeople (aka senior data scientists) are the highest level of data scientist and usually have managerial roles, closer to the business world than to data science per se. They usually come from a mixed background that includes a degree in management.

- The mixed/generic type of data scientists are the most balanced, having developed all of the aspects of data science more or less equally. They have less breadth of experience than data businesspeople, are very versatile, and come from all types of backgrounds. Usually, the mixed/generic data scientist evolves into the data businesspeople type.

Chapter 4
The Data Scientist's Mindset

People tend to have a very superficial view of what a data scientist is (if they can even distinguish the term from the data analyst or from the traditional scientist). This is clearly reflected in the books and articles that are available today on this role[18]. Rarely will you find a text that attempts to go deeper into what a data scientist really is.

A data scientist is a person characterized by a particular set of traits, qualities, a way of thinking, and ambition, just like every profession, not just by a set of skills. Let us look at each one of these key aspects of this mindset one by one in order to obtain a better understanding of it and create a framework about what being a data scientist really is.

4.1 Traits

A data scientist has a variety of professional characteristics and traits that usually reflect the kind of work he specializes in, so this list is not set in stone and is more of a guideline to understand this role better. First and foremost, a data scientist has a healthy *curiosity* about the things he observes, such as potential patterns or relationships between two attributes or features, unusual

[18] Recently, the author came across a post on Quora (a forum for geeks) where the poster mentioned a series of 10 steps that you need to do in order to become a data scientist. Most of them were focused on specific skills, the majority of which are of questionable quality. This clearly illustrates the limited understanding many people have about what being a data scientist involves and how this misinformation propagates.

distributions, etc. If you want to be a data scientist worth the money you earn, you need to have an inquiring mind.

This does not mean that you need to be curious about everything and get lost in perpetual random quests for answers. Curiosity has to be accompanied by the discipline to focus on down-to-earth, long-term interests that are more grounded than a fleeting curiosity, which can be impulsive and superficial. A data scientist is interested in the phenomena he observes in the data he deals with, wanting to get to the bottom of them. A statistical analysis of what's there may be a good first step for him, but he is not satisfied until he has a good answer for the reason of these phenomena, the root cause behind the statistical metrics he calculates. This allows him to explain the root cause to other people in the company in the form of a story.

Fig. 4.1 Curiosity is a very useful trait to have as a data scientist.

This leads to another trait that is somewhat akin to curiosity: an interest in *experimentation*. Namely, the data scientist has the courage and the imagination to try out new things, develop new ideas and put them into practice, design experiments and validate new notions that he develops. He is not afraid to build a model that no one else has built before, always being fully aware of the risks in terms of resource usage, etc. All this is a disciplined and practical

form of experimentation where the ideas stem from the data available. Otherwise, there is the risk of misusing the data to project notions that are not there, a common mistake among data analysts lacking scientific discipline in their work. Experimentation is crucial, though, because it allows the data scientist to find new ways of interpreting the data and helping it transmute into information that can be useful to other people. This is an important point. The output of the experiments needs to be understandable to the non-technical members of his team; otherwise, it is probably immature. So experimentation is applied on many levels. Representation of results is one of them, and although not the most intellectually challenging to the data scientist, it is definitely no less important than the other tasks he undertakes.

Other traits that the data scientist has are *creativity* and *systematic work*. These are mentioned together because they are often applied together in data science and are equally important. The data scientist is an artist of sorts, in the sense that he is involved in design and other creative endeavors in his line of work. He values out-of-the-box thinking and regularly applies it to the problems he tackles. Although knowledgeable in various data analysis methods, he is not restricted by this palette of methodologies. Instead, he may use a combination of them, or even something completely new, tailored for the particular problem he faces. This is an important aspect of the data scientist that distinguishes him from a traditional data analyst and statistician. Creativity goes hand in hand with experimentation, making it an organic growth approach to tackling problems. Without creativity, experimentation may not quickly lead to results (think of a scientist researching a treatment for a disease; without creativity he may have to spend a large amount of time and other resources trying out potential solutions, many of which he could avoid testing altogether by applying a more creative and efficient approach). Creativity is, therefore, invaluable to the data scientist and a fundamental aspect of his thinking.

The data scientist is not, however, an artist per se. That's why every creative thought he has is accompanied by several not-so-creative actions. This is where *systematic work* comes in. Think of any inventor (e.g., Thomas Edison) and how many hours of often tedious work were spent on honing and applying their creative insight. In a sense, having a particularly creative idea is not all that difficult. Finding one that is applicable to a given problem is very creative but still not too challenging, either. However, putting this idea into practice, working out all the engineering details of it, and getting useful results in a manageable timeframe: that is a real accomplishment. This is feasible through systematic work, which is not just hard work, but work done in a methodical and efficient way, something typical of any type of scientific endeavor. The trait of working systematically expresses itself as the discipline, organization and rhythm through which the data scientist manages to ground the creative ideas he comes up with.

Last, but certainly not least, of the essential skills of the data scientist is that of *communication*. Data science is not an *ad hoc* field. It is an interdisciplinary one, and, as such, it is closely connected to other fields. In the data scientist role, this translates into a series of connections or collaborations with other professionals in the organization. These professionals are usually in a variety of specialties and may have a different understanding of the various levels of the information the data scientist deals with. For the data scientist to be good at his role, he needs to be able to explain not only his methodology and results to his colleagues and his managers, but also the value of the whole process. It is this connectivity to other people that gives value to the data scientist's role. You don't do data science on your own (unless you are just practicing). Besides, requirements and problem parameters are not always clear cut, needing to be defined through a process of interviews with other professionals in the organization as well as communiqués with middle/upper management. The data scientist needs to be able to not only communicate his results, but also

understand clearly what is expected of him and engage in a constructive conversation to determine the best possible parameters of the projects he undertakes. He needs to be able to manage the ensuing expectations and make sure that others, especially those in managerial roles, see practical value in what he can provide (without expecting miracles).

Although there are other traits that a data scientist may have, the traits described above are the most essential ones for a good data scientist. When applied with discernment and intelligence, they can help his role develop organically and effectively.

4.2 Qualities and Abilities

Hand in hand with traits are the qualities and abilities of a data scientist, which often depend on his particular specialty. However, there are certain ones that are found in every type of data scientist. The most important of these are the following:

Model Building. This is a fundamental ability of a data scientist, involving the design and implementation of mathematical models that can be used to solve the data-related problems he is asked to tackle. Stemming from the need to scientifically explain and predict certain phenomena that are reflected in the available data, model building is a key skill for every data scientist. This is also one of those things that differentiate him from most statisticians and the majority of data professionals. It involves understanding and creativity as well as a great deal of imagination. The models built by a data scientist are implemented in an interactive environment, so it goes without saying that a certain amount of programming also takes place and that the models created take into account the available resources, using them in a very effective and efficient manner.

Building a model, though, is not that easy. The model has to be as simple as possible without being too simple. For example, a simple

model may be able to predict how many people will attend a football match based on how large the fan clubs of the participating teams are, the expected weather on that day, and the time of the year, while an overly simple model may try to predict the same thing using only one of these features. The model has to be able to generalize so that it can predict a lot of different cases that may not be entirely akin to the ones that were used to create it. It has to be easy to change and understood by everyone who uses it, especially those who may need to fine-tune it. Model building can be based on mathematics, a computational algorithm, or, more often, a combination of both. The key thing is efficiency, so this is something that the data scientist needs to factor into the whole process. What good would a perfect model be if it took weeks to provide any results, or if it required a huge number of computers to run it? Also, it goes without saying that a data scientist needs to be able to evolve and fine-tune his models, customizing them to different circumstances and adapting them to the data when it changes.

Planning. This is an obvious quality for anyone in the data-related professions, but it is especially useful for a data scientist as it is very easy to get carried away with analyzing the available data, experimenting with various models, and not dedicating sufficient time for other tasks such as documenting the process and the results or creating the corresponding visuals, comprehensive presentations and reports. In addition, a data scientist needs to be able to factor in potential delays, technical issues, communication lags, etc. in order to make sure that he can meet all the deadlines of the projects he undertakes. He needs to be able to think like a project manager and have a practical approach to assessing time durations of different tasks and plotting a realistic and efficient plan of action for all the projects he undertakes.

Problem Solving. This is a key quality for any scientist, particularly a data scientist; it involves being able to focus on solutions rather

than on the restrictions that a problem presents. Often, the data scientist has not encountered these solutions before, so it requires a certain amount of imagination and creativity. It means being able to look at the problem at hand from different angles, with different eyes and an open mind.

Problem solving often involves finding ways to hack existing technologies to work around a problem. Data science is rarely clearly defined (similar to most academic endeavors), and every problem it deals with is unique. That's why a data scientist is often more akin to the hacker than the scientist as he may have to tackle problems through a lateral thinking approach (see next subchapter for details) and walk outside the beaten path. Also, he may need to develop new tools for tackling the problems he faces (i.e., making sense of chaotic big data), building code from scratch or doing major modifications to the existing code.

Learning Fast. Being able to learn new things and learn them fast is a priceless quality for any profession. However, in a field with constant and rapid changes such as data science, it is particularly useful. It also attests to mental agility and promotes creativity, both invaluable aspects of the mindset suitable for someone who wants to tackle big data problems. Learning fast means being very methodical, selective, and able to assess different sources of knowledge. It requires great discipline and mental plasticity. Almost anyone can learn like that at a relatively early age, but being able to maintain this openness throughout adulthood is a challenge to most people. A data scientist accepts this challenge and does not let age dictate what he can or cannot learn, nor how fast he can do so. His disciplined and nimble mind makes sure of that.

Key elements for learning fast are motivation and being able to perceive the applicability of new material. If you keep this in mind, it will be easier to develop this ability and use it effectively in your journey as a data scientist.

been encountered before. It hones a can-do attitude that allows him to deal with novel situations effectively, efficiently and creatively. This simple quality is the glue that ties all the other qualities and abilities together, enabling the data scientist to organically evolve his techniques and even his thinking, thus making him an invaluable asset to his organization.

Research. This has nothing to do with academic research, although it is scientific in essence. The data scientist is able to understand and evaluate the current state of the art in his field and find all the knowledge resources that are required for his tasks. This entails more than looking things up on a search engine or a knowledge base, though. Finding quality sources is crucial for tackling the challenging problems of big data, and it requires a trained eye to see which methods are applicable and efficient when applied to a specific problem. It also entails putting together documents describing new methods he develops in a concise, scientifically robust and replicable way. Whether or not these documents are publishable is another matter and not related to how useful the described techniques are.

This ability ties very well with learning fast as it enables the data scientist to be self-sufficient when it comes to learning. In addition, it makes it possible for him to train others as well as have something to share at data science conferences and other relevant events if he so chooses. Needless to say, it is particularly vital in the initial stages of his career, especially if he has a disposition towards innovation.

Attention to Detail. A data scientist needs to be attentive to details since that is usually where useful information lurks. Also, a small detail may cause syntactical or, even worse, logical errors in his programs, slowing him down and compromising his deadlines. Apart from the efficiency boost, this ability is very useful in other ways as well. For example, certain details in the available data may hint at using one or another data analysis approaches or towards a

Fig. 4.3 Thinking is an important aspect of the data scientist's mindset.

4.4 Ambitions

It seems a bit unconventional for a book like this to talk about a professional's ambitions as this is something that is very personal and somewhat relative. However, there are certain aspirations that are more or less common to data scientists; understanding them may provide useful insight into his mindset.

A data scientist aspires to master big data in its many forms. Being able to deal with a particular data set in this domain is great, but often not enough. Someone who cares for data science finds ways, often through interaction with other professionals in this field, to be on top of the data that is out there, meaning that he comprehends fully what each data type can offer to an organization, what useful information he can potentially derive from it and what costs acquiring each data type entails. This stems from the dream of continuous improvement, which is quite feasible in fields like this where more and more tools become available as new data analysis methods are developed all the time.

Data scientists also constantly want to learn new things. This wish ties quite well with the previous ambition of mastering big data since learning, especially when related to diverse things that include the realm of big data, has been proven to aid in the development of creativity and mental agility. These are essential aspects of the role of the data scientist, and cultivating them makes perfect sense. A data scientist's interests are not limited to the data science techniques that he may use in his everyday work. He is also interested in new developments in artificial intelligence, distributed computing, information security, new programming languages and machine learning, among other fields.

Fig. 4.4 A data scientist is not without ambitions.

Finally, a data scientist aspires to familiarize himself with the open problems and challenges that exist in the big data world as well as the opportunities that are available through the intelligent processing of company data. He may want to research new ways of tackling problems through the use of new technologies, development of new methods, etc., or he may look into how specific business requirements can be fulfilled through the use of certain kinds of data that are available or can be acquired in a cost-effective manner.

Chapter 5
Technical Qualifications

Similar to many other jobs nowadays, a robust set of technical qualifications is essential before you can opt for a data science job. The mindset of the data scientist, which was described in the previous chapter, is like an operating system you need to have installed in your mind, but it needs to be augmented with particular software (i.e., your technical skills) to enable you to get the job done. These skills fall into three broad categories: general programming, scientific background and specialized know-how (software and techniques). Naturally, all these qualifications will vary greatly from one company to another, but having a core set of skills across all of these categories may help you qualify for most data science jobs.

In this chapter, we will look into the most commonly expected qualifications for a data scientist position today. We'll look into the general programming skills required, the scientific background you will be expected to have and the specialized know-how you need to possess related to data analysis and data engineering.

5.1 General Programming

Unlike other branches of science, programming is a must have for any data scientist. Professionals in academia may be able to get by without knowing any coding, but in data science you need to know languages that are:

- Robust
- Popular in the industry
- Scalable, especially when it comes to large data sets


image
not
available

use of the data science process (see Chapter 11) and have a thorough understanding of the results. Moreover, you will be able to fine-tune your methods, know where something has gone wrong and come up with alternative approaches to a problem. It is very hard to overestimate the importance of having a scientific background.

5.3 Specialized Know-How

Being a data scientist requires some specialized know-how that distinguishes him from other professionals. It is important that you have mastery of at least one of these statistics tools:

- R (the most advanced statistical analysis platform; open-source)
- SPSS (another great statistical tool; proprietary)
- SAS (a very popular statistical tool in the industry; proprietary)
- Stata (another good statistical tool; proprietary)

Some employers might also include Matlab in the list, since Matlab enables you to do any data analysis conceivable with minimal code and comes with its own advanced integrated development environment (IDE) that makes debugging and development a walk in the park. The big drawback of Matlab is that its license is quite expensive, especially for commercial applications.

If you are not sure on which tool to focus, it is recommended that you go with R. Over the past few years, R has become more popular for several good reasons: R is open-source (and therefore completely free), it has a very large user-community, it is easy to install and customize, it is fairly easy to learn, there is ample documentation for it as well as several books for a variety of levels, and it comes with a wide variety of libraries (known as packages) that enable you to do many complex tasks easily without having to do much coding. Note that although R has all the characteristics of an OO language (and all of the data structures in its workspace are

the marketing endeavors that employ this information. It is particularly useful if you are working for a company in the retail industry.

- Big data integrated processing system (e.g., IBM's BigInsights, Knime, Alpine, and Pivotal, just to name a few) – although it is not likely that this will be a requirement for a data scientist job posting, being familiar with a system like this provides you with a better understanding of the bigger picture of big data processing and allows you to focus on the most creative aspects of your job since it does all the low-level work for you and helps you deal with the problem using a high-level approach.

As the data science field matures, it is likely that additional specialized know-how will be required in order to be a data scientist. However, the qualifications identified in this chapter are bound to remain essential, particularly the data analysis tools. It is recommended that you keep up to date with the newest developments in the field so that you know how to adjust your training strategy and avoid wasting your resources on things that you may not need.

5.4 Key Points

- As a data scientist, you need a specific set of technical skills that are the tools you will use in your everyday job.
- You need to be familiar with one or more object-oriented programming languages such as Java or Perl. Having mastery of at least one of them is imperative.
- You need to have a solid scientific background (even if your education is non-technical), making you adept in the following:
    - The scientific process
    - The theory behind various data analysis techniques


image
not
available


image
not
available