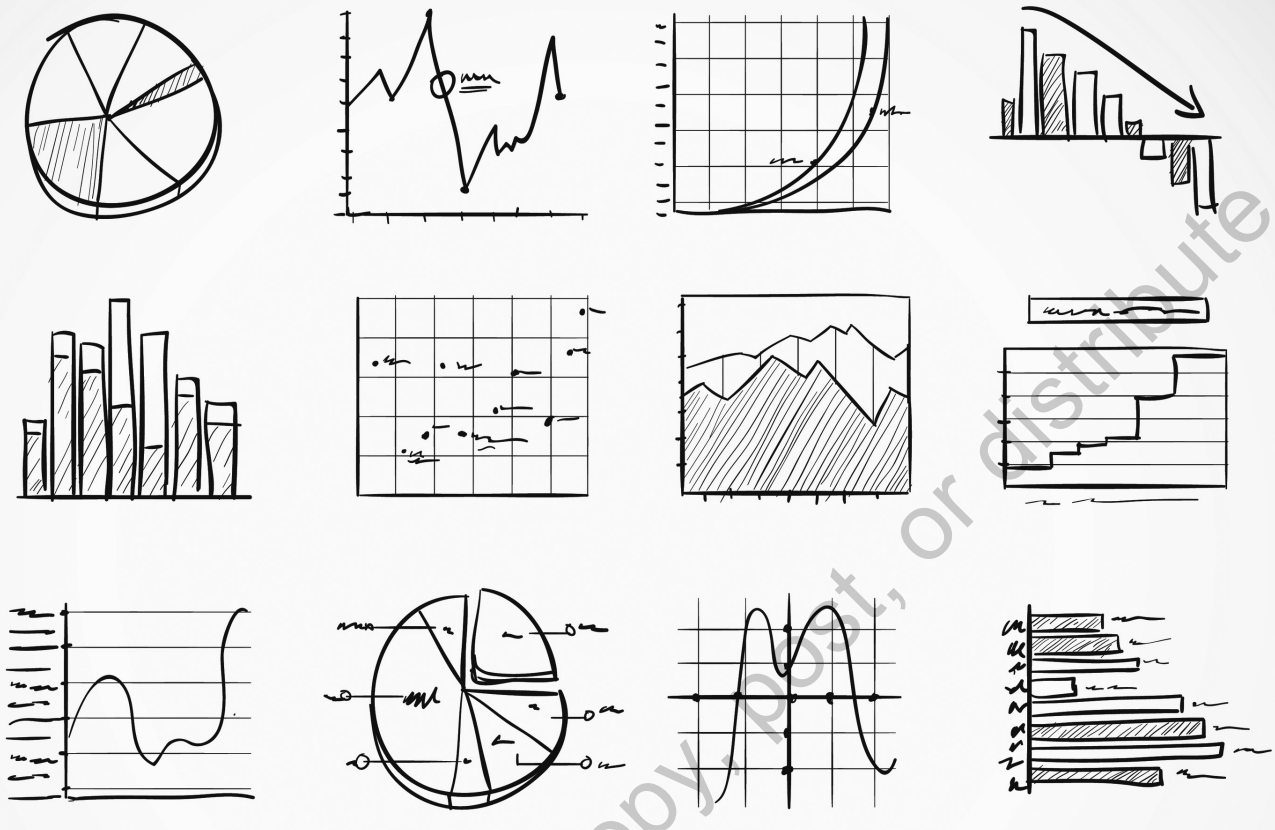


© iStockphoto.com/ SpiffyJ

INTRODUCTION: DATA SCIENCE, MANY SKILLS

LEARNING OBJECTIVES

- Articulate what data science is.
- Understand the steps, at a high level, of doing data science.
- Describe the roles and skills of a data scientist.

WHAT IS DATA SCIENCE?

For some, the term "data science" evokes images of statisticians in white lab coats staring fixedly at blinking computer screens filled with scrolling numbers. Nothing could be farther from the truth. First, statisticians do not wear lab coats: this fashion statement is reserved for biologists, physicians, and others who have to keep their clothes clean in environments filled with unusual fluids. Second, much of the data in the world is non-numeric and unstructured. In this context, unstructured means that the data are not arranged in neat rows and columns. Think of a web page full of photographs and short messages among friends: very few numbers to work with there. While it is certainly true that companies, schools, and governments use plenty of numeric information—sales of products, grade point averages, and tax assessments are a few examples—there is lots of other information in the world that mathematicians and statisticians look at and cringe. So, while it is always useful to have great math skills, there is much to be accomplished in the world of data science for those of us who are presently more comfortable working with words, lists, photographs, sounds, and other kinds of information.

In addition, data science is much more than simply analyzing data. There are many people who enjoy analyzing data and who could happily spend all day looking at histograms and averages, but for those who prefer other activities, data science offers a range of roles and requires a range of skills. Let's consider this idea by thinking about some of the data involved in buying a box of cereal.

Whatever your cereal preferences—fruity, chocolaty, fibrous, or nutty—you prepare for the purchase by writing "cereal" on your grocery list. Already your planned purchase is a piece of data, also called a datum, albeit a pencil scribble on the back on an envelope that only you can read. When you get to the grocery store, you use your datum as a reminder to grab that jumbo box of FruityChocoBoms off the shelf and put it in your cart. At the checkout line, the cashier scans the barcode on your box and the cash register logs the price. Back in the warehouse, a computer tells the stock manager that it is time to request another order from the distributor, because your purchase was one of the last boxes in the store. You also have a coupon for your big box and the cashier scans that, giving you a predetermined discount. At the end of the week, a report of all the scanned manufacturer coupons gets uploaded to the cereal company so they can issue a reimbursement to the grocery store for all of the coupon discounts they have handed out to customers. Finally, at the end of the month a store manager looks at a colorful collection of pie charts showing all the different kinds of cereal that were sold, and, on the basis of strong sales of fruity cereals, decides to offer more varieties of these on the store's limited shelf space next month.

So the small piece of information that began as a scribble on your grocery list ended up in many different places, but most notably on the desk of a manager as an aid to decision making. On the trip from your pencil to the manager's desk, the datum went through many transformations. In addition to the computers where the datum might have stopped by or stayed on for the long term, lots of other pieces of hardware—such as the barcode scanner—were involved in collecting, manipulating, transmitting, and storing the datum. In addition, many different pieces of software were used to organize, aggregate, visualize, and present the datum. Finally, many different human systems were involved in working with the datum. People decided which systems to buy and install, who should get access to what kinds of data, and what would happen to the data after its immediate purpose was fulfilled. The personnel of the grocery chain and its partners made a thousand other detailed decisions and negotiations before the scenario described above could become reality.

THE STEPS IN DOING DATA SCIENCE

Obviously, data scientists are not involved in all of these steps. Data scientists don't design and build computers or barcode readers, for instance. So where would the data scientists play the most valuable role? Generally speaking, data scientists play the most active roles in the four A's of data: data architecture, data acquisition, data analysis, and data archiving. Using our cereal example, let's look at these roles one by one. First, with respect to architecture, it was important in the design of the point-of-sale system (what retailers call their cash registers and related gear) to think through in advance how different people would make use of the data coming through the system. The system architect, for example, had a keen appreciation that both the stock manager and the store manager would need to use the data scanned at the registers, albeit for somewhat different purposes. A data scientist would help the system architect by providing input on how the data would need to be routed and organized to support the analysis, visualization, and presentation of the data to the appropriate people.

Next, acquisition focuses on how the data are collected, and, importantly, how the data are represented prior to analysis and presentation. For example, each barcode represents a number that, by itself, is not very descriptive of the product it represents. At what point after the barcode scanner does its job should the number be associated with a text description of the product or its price or its net weight or its packaging type? Different barcodes are used for the same product (e.g., for different sized boxes of cereal). When should we make note that purchase X and purchase Y are the same product, just in different packages? Representing, transforming, grouping, and linking the data are all tasks that need to occur before the data can be profitably analyzed, and these are all tasks in which the data scientist is actively involved.

The analysis phase is where data scientists are most heavily involved. In this context, we are using analysis to include summarization of the data, using portions of data (samples) to make inferences about the larger context, and visualization of the data by presenting it in tables, graphs, and even animations. Although there are many technical, mathematical, and statistical aspects to these activities, keep in mind that the ultimate audience for data analysis is always a person or people. These people are the data users, and fulfilling their needs is the primary job of a data scientist. This point highlights the need for excellent communication skills in data science. The most sophisticated statistical analysis ever developed will be useless unless the results can be effectively communicated to the data user.

Finally, the data scientist must become involved in the archiving of the data. Preservation of collected data in a form that makes it highly reusable—what you might think of as data curation—is a difficult challenge because it is so hard to anticipate all of the future uses of the data. For example, when the developers of Twitter were working on how to store tweets, they probably never anticipated that tweets would be used to pinpoint earthquakes and tsunamis, but they had enough foresight to realize that geocodes—data that show the geographical location from which a tweet was sent—could be a useful element to store with the data.

THE SKILLS NEEDED TO DO DATA SCIENCE

All in all, our cereal box and grocery store example helps to highlight where data scientists get involved and the skills they need. Here are some of the skills that the example suggested:

Learning the application domain: The data scientist must quickly learn how the data will be used in a particular context.

Communicating with data users: A data scientist must possess strong skills for learning the needs and preferences of users. The ability to translate back and forth between the technical terms of computing and statistics and the vocabulary of the application domain is a critical skill.

Seeing the big picture of a complex system: After developing an understanding of the application domain, the data scientist must imagine how data will move around among all of the relevant systems and people.

Knowing how data can be represented: Data scientists must have a clear understanding about how data can be stored and linked, as well as about metadata (data that describe how other data are arranged).

Data transformation and analysis: When data become available for the use of decision makers, data scientists must know how to transform, summarize, and make inferences from the data. As noted above, being able to communicate the results of analyses to users is also a critical skill here.

Visualization and presentation: Although numbers often have the edge in precision and detail, a good data display (e.g., a bar chart) can often be a more effective means of communicating results to data users.

Attention to quality: No matter how good a set of data might be, there is no such thing as perfect data. Data scientists must know the limitations of the data they work with, know how to quantify its accuracy, and be able to make suggestions for improving the quality of the data in the future.

Ethical reasoning: If data are important enough to collect, they are often important enough to affect people's lives. Data scientists must understand important ethical issues such as privacy, and must be able to communicate the limitations of data to try to prevent misuse of data or analytical results.

The skills and capabilities noted above are just the tip of the iceberg, of course, but notice what a wide range is represented here. While a keen understanding of numbers and mathematics is important, particularly for data analysis, the data scientist also needs to have excellent communication skills, be a great systems thinker, have a good eye for visual displays, and be highly capable of thinking critically about how data will be used to make decisions and affect people's lives. Of course, there are very few people who are good at all of these things, so some of the people interested in data will specialize in one area, while others will become experts in another area. This highlights the importance of teamwork, as well.

In this *Introduction to Data Science* book, a series of data problems of increasing complexity is used to illustrate the skills and capabilities needed by data scientists. The open source data analysis program known as R and its graphical user interface companion RStudio are used to work with real data examples to illustrate both the challenges of data science and some of the techniques used to address those challenges. To the greatest extent possible, real data sets reflecting important contemporary issues are used as the basis of the discussions.

Note that the field of big data is a very closely related area of focus. In short, big data is data science that is focused on very large data sets. Of course, no one actually defines a "very large data set," but for our purposes we define big data as trying to analyze data sets that are so large that one cannot use RStudio. As an example of a big data problem to be solved, Macy's (an online and bricks and mortar retailer) adjusts its pricing in near real

time for 73 million items, based on demand and inventory (http://searchcio.techtarget.com/opinion/Ten-big-data-case-studies-in-a-nutshell). As one might guess, the amount of data and calculations required for this type of analysis is too large for one computer running RStudio. However, the techniques covered in this book are conceptually similar to how one would approach the Macy's challenge and the final chapter in the book provides an overview of some big data concepts.

Of course, no one book can cover the wide range of activities and capabilities involved in a field as diverse and broad as data science. Throughout the book references to other guides and resources provide the interested reader with access to additional information. In the open source spirit of R and RStudio these are, wherever possible, web-based and free. In fact, one of guides that appears most frequently in these pages is Wikipedia, the free, online, user-sourced encyclopedia. Although some teachers and librarians have legitimate complaints and concerns about Wikipedia, and it is admittedly not perfect, it is a very useful learning resource. Because it is free, because it covers about 50 times more topics than a printed encyclopedia, and because it keeps up with fast-moving topics (such as data science) better than printed sources, Wikipedia is very useful for getting a quick introduction to a topic. You can't become an expert on a topic by consulting only Wikipedia, but you can certainly become smarter by starting there.

Another very useful resource is Khan Academy. Most people think of Khan Academy as a set of videos that explain math concepts to middle and high school students, but thousands of adults around the world use Khan Academy as a refresher course for a range of topics or as a quick introduction to a topic that they never studied before. All of the lessons at Khan Academy are free, and if you log in with a Google or Facebook account you can do exercises and keep track of your progress.

While Wikipedia and Khan Academy are great resources, there are many other resources available to help one learn data science. So, at the end of each chapter of this book is a list of sources. These sources provide a great place to start if you want to learn more about any of the topics the chapter does not explain in detail.

It is valuable to have access to the Internet while you are reading, so that you can follow some of the many links this book provides. Also, as you move into the sections in the book where open source software such as the R data analysis system is used, you will sometimes need to have access to a desktop or laptop computer where you can run these programs.

One last thing: The book presents topics in an order that should work well for people with little or no experience in computer science or statistics. If you already have knowledge, training, or experience in one or both of these areas, you should feel free to skip over some of the introductory material and move right into the topics and chapters that interest you most.

Sources

http://en.wikipedia.org/wiki/E-Science

http://en.wikipedia.org/wiki/E-Science_librarianship

http://en.wikipedia.org/wiki/Wikipedia:Size_comparisons

http://en.wikipedia.org/wiki/Statistician

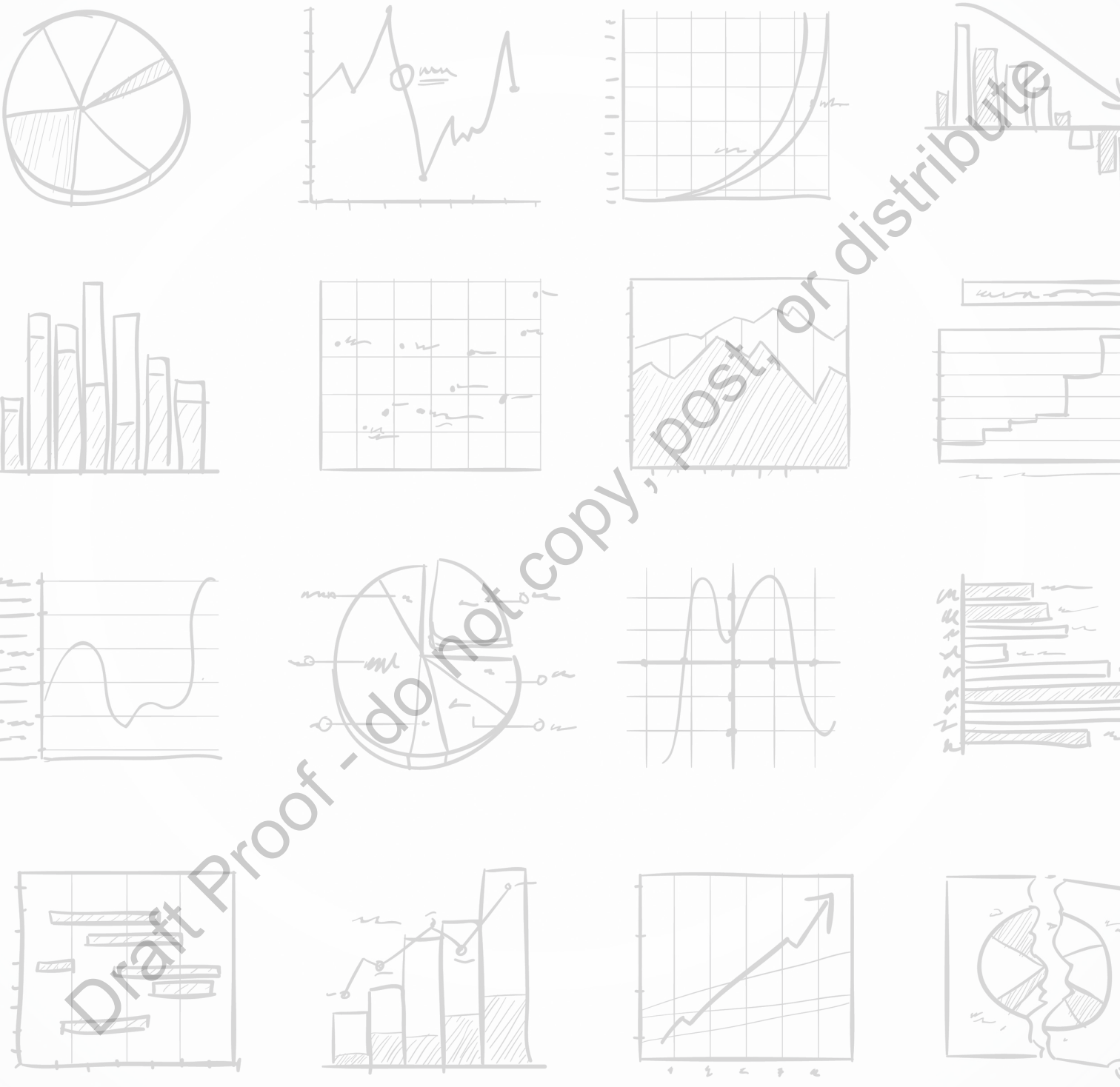
http://en.wikipedia.org/wiki/Visualization_(computer_graphics)

http://www.khanacademy.org/

http://www.r-project.org/

http://readwrite.com/2011/09/07/unlocking-big-data-with-r/

http://rstudio.org/



© iStockphoto.com/ SpiffyJ