What is Data Science?

“Data Science” is a new term for an old activity, analyzing data. Scientists have analyzed data for centuries. From time to time, they have developed new toolkits for examining data, such as the scientific method (1600’s), descriptive statistics (1800’s), and statistical inference (1900’s). In recent years, however, scientists have been spurred on by the personal computer, a device that allows researchers to accumulate, store, and process very large data sets. As a result, data analysis has become more complicated in every way; the data that we analyze has become more complex, the theory that guides us has become more sophisticated, and the methods that we can use have become more numerous and technical.

These developments have transformed data analysis from a common task done by every scientist, to a specialized field with its own rules, theories, and experts. Data analysis is now something that you can — and perhaps must — specialize in if you want to do it correctly. Modern analysts call this new field “Data Science,” and it focuses on the unique problems posed by data.

Three problems posed by data

Every scientist faces the same three problems when they wish to learn from data. You can think of these problems, and their solutions, as the logistical, tactical, and strategic components of data science.

A logistical problem - How can you store and manipulate data without making errors?

To understand these problems, put yourself in a scientist's shoes. The first problem that you will face is that data is difficult to work with. Modern data sets are too large to memorize and impossible to manipulate in your head. You could try to manipulate a data set with paper and a pencil, but this would take a long time and invite errors.

Moreover, many of the things that you'll want to do to data are repetitive and complicated, which also invites errors. A thorough simulation may require you to repeat a task thousands or even millions of times. How long could you do this without becoming bored or careless?

Computers provide an efficient way to handle the logistics of data science. You can store an entire data set in a commputer's memory and retrieve parts of the data at will. You can also use computers to manipulate the data and run simulations. However, to use a computer, you need to know how to program.

A cognitive problem - How can you discern the information contained in data?

The next problem will appear when you try to understand what the data says. You will face a struggle because the human mind does not easily consume tabular data. If you don't believe me, try the experiment in the side bar.

This is a cognitive problem, as in cognitive science. The human mind turns input into insights by identifying relationships and context. When it comes to data, this means that the meaning of each value in a data set will depend on the rest of the values in the set. For example, you may notice that Berkshire-Hathaway stock costs $169,000 while looking at a data set of stock prices. However, you would not understand that this price is an extreme outlier until you notice that *every other* stock price is much lower. Relationships between variables also only become apparent when you consider many values at the same time. You could not guess whether the temperature in Dallas has much to do with the price of tea in China, until you look at many measurements of both taken over a period of time.

Unfortunately, it is difficult for the human mind to attend to multiple values at once. Our working memory creates a bottleneck that only lets us consider 4-7 new values at the same time. If a relationship in a data set depends on more than 4-7 values, it is unlikely that you will spot it. There are cognitive tricks that you can use to get around this bottleneck — that's why you do not notice it on a daily basis — but tabular data does not take advantage of any of them.

You can make data easier to comprehend by changing the format of your data sets. For example, you can reduce your data to a few summary statistics, or you can transform it into a plot, ???, which will make the data almost immediately transparent. Each of these transformations pre-processes the data in a way that makes it graspable by the human cognitive system. This deceptively simple activity is the basis of many academic fields including Descriptive Statistics, Exploratory Data Analysis, infoVis, sciVis, and Visual Analytics. I hesitate to add the general field of Data Analysis here because Data analysis has come to mean different things in different places. However, in its literal sense, "Data Analysis" is a cognitive task. Analysis means breaking a thing into smaller parts, which are then easier to understand.

The cognitive problem imposed by data applies to more than just your mind; computers can struggle to comprehend data as well. Most computer programs cannot correctly identify relationships between values in a data set unless the data is formatted in a structure that the software expects or understands. Usually this format will parallel the real relationships contained in the data, relationships created by variables and observations.

Computers also have a cognitive bottleneck of their own; a computer cannot process a data set that exceeds its memory. This has led to the concept of Big Data. Big Data is a set of information that exceeds a computer's ability to process it, just as Data is a set of information that exceeds the human mind's ability to process it. In each case, you can use various strategies to manage cognitive limitations while working with overwhelming amounts of data.

What does it say?

It is hard to comprehend the information in tabular data. For example, the 120 (x, y) points in the table below reveal a strong relationship, but can you discern what it is?

(0.31, 2.03)	(0.39, 2.12)	(0.03, 1.22)	(1, 3.19)	(0.84, 0.38)	(0.95, 2.73)
(0.86, 3.47)	(0, 0.24)	(0.76, 3.69)	(0.03, 0.35)	(0, 0.98)	(0.29, 1.89)
(-0.03, 1.3)	(0.86, 2.44)	(0.61, 1.18)	(0.77, 0.71)	(0.09, 2.03)	(0.02, 2.45)
(0.77, 3.55)	(0.3, 2.01)	(0.97, 2.79)	(0.17, 2.05)	(0.57, 1.3)	(0.79, 2.37)
(1, 0)	(0.99, 2.95)	(0.37, 1.74)	(0.32, 3.95)	(0.58, 2.23)	(0.54, 2.17)
(0.92, 0.1)	(0.81, 3.53)	(0.02, 1.97)	(0.77, 0.55)	(0.91, 0.18)	(0.45, 1.7)
(0.95, 2.54)	(0.03, 3.46)	(0.95, 2.92)	(0.48, 1.42)	(0.02, 2.26)	(1.03, 3.04)
(0, 1.6)	(0, 0.87)	(0.96, 2.84)	(0.93, 2.57)	(-0.01, 2.56)	(0.76, 3.65)
(-0.04, 2.9)	(0.76, 3.65)	(0.13, 3.97)	(0.69, 2.27)	(0.42, 1.57)	(0.95, 3.12)
(0.82, 0.49)	(0.11, 1.97)	(0.69, 2.22)	(0.74, 2.29)	(0.01, 0.12)	(0.79, 2.42)
(0.6, 1.07)	(-0.04, 0.56)	(0.52, 1.51)	(0.57, 1.35)	(0.48, 2.15)	(0.92, 0.28)
(0.05, 3.88)	(-0.04, 4.01)	(0.87, 0.27)	(0.65, 0.88)	(0.03, 2.81)	(0.03, 4.02)
(0.67, 0.86)	(0.38, 1.82)	(-0.01, 3.81)	(-0.01, 1.42)	(-0.04, 1.8)	(0, 3.22)
(-0.01, 4.03)	(0.02, 2.13)	(0, 2.67)	(-0.05, 3.07)	(0.89, 2.5)	(0.87, 3.42)

(0.05, 3.33)	(0.91, 3.25)	(0.95, 3.38)	(0.68, 3.73)	(0.48, 3.86)	(0.99, 2.68)
(0.04, 3.62)	(-0.02, 1.07)	(0.02, 1.57)	(-0.05, 0.67)	(0.28, 3.96)	(0.32, 2.1)
(0.19, 2.01)	(0.96, 3.23)	(0.72, 0.73)	(1.01, 3.08)	(0, 1.9)	(0.02, 0.44)
(0.52, 2.2)	(-0.04, 2.37)	(0.62, 3.84)	(0.64, 0.97)	(0.57, 3.82)	(0.29, 4.01)
(0.04, 2)	(0.26, 2.06)	(0.68, 2.26)	(0.52, 3.88)	(-0.03, -0.01)	(0.64, 3.82)
(0.15, 3.96)	(0.91, 3.41)	(0.46, 3.94)	(0.24, 3.99)	(1.01, 2.75)	(-0.04, 3.53)

You can make the same information clear by displaying it in a human friendly way. It is no secret that data visualizations do just this. See Figure P-4 at the end of the preface for a visualization of the points above. Compare how easy it is to spot the relationship in the visualization vs. in the table.

A reasoning problem - What does the data imply?

Once you've understood all of the information to be found within your data, you face a final problem: what does the data imply? There is a large gap between data, which describes specific cases, and the type of general rules that describe reality and comprise scientific hypotheses. Data can sometimes show that a hypothesis is not true for some cases, and therefore cannot be true for *every* case. However, data cannot show that a hypothesis *is true* for every case. Nor can data differentiate with certainty between two competing, but plausible, hypotheses. Yet these are the things that you will want to do as a scientist.

Scientists proceed by pairing data with logical reasoning. They use the data as observations to draw inferences from. Many of these inferences are not valid in the logical sense, which requires that an inference *must* be true. So scientists compensate by measuring and then comparing the probabilities that inferences are true.

These inferences form the final product of Data Science: an argument for or against a conclusion, and the entire Data Science process is built around creating strong inferences. Data scientists employ the scientific method, which helps ensure that data is free of error and bias, two things that undermine inferences. Data scientists visualize and explore their data to find information that can suggest, corroborate, or refute an inference. And, data scientists rely heavily on technical methods, such as hypothesis testing, modeling, and machine learning, which are all advanced ways of identifying logical inferences that have a high probability of being true. Many of these advanced inference techniques are so complicated that they pose a logistical problem of their own. They can only be performed with a computer, which brings the problems of data science full circle.

How to excel at Data Science

This book will teach you a pragmatic version of data science that is organized around three skill sets: computer programming, data comprehension, and scientific reasoning, Figure P-1. These are the skill sets that solve the problems posed by data.

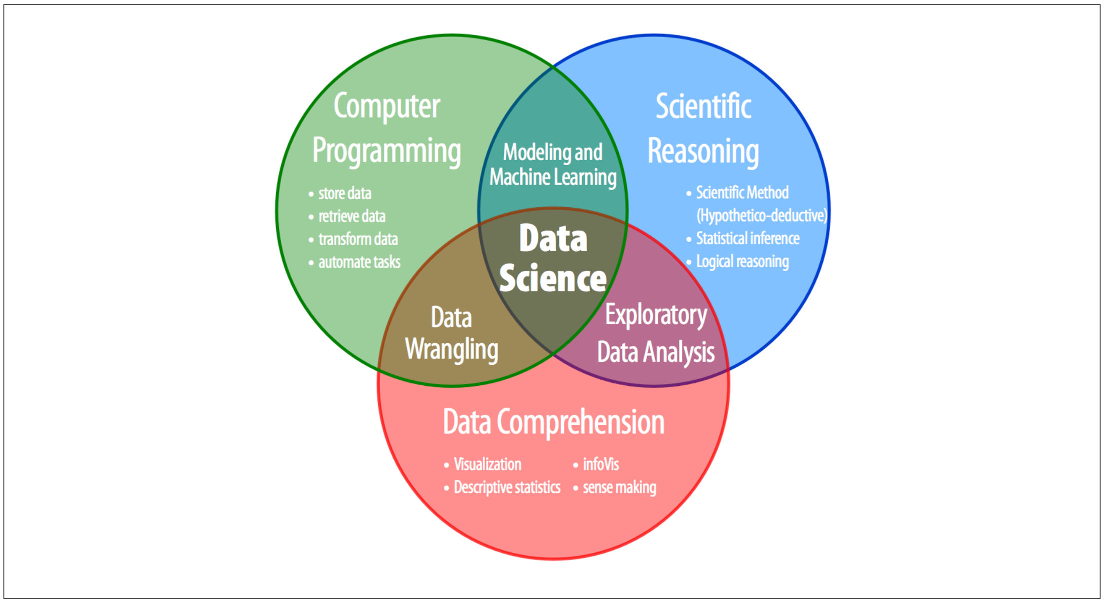

Figure P-1. The three core skill sets of data science