SAGE was founded in 1965 by Sara Miller McCune to support the dissemination of usable knowledge by publishing innovative and high-quality research and teaching content. Today, we publish more than 750 journals, including those of more than 300 learned societies, more than 800 new books per year, and a growing range of library products including archives, data, case studies, reports, conference highlights, and video. SAGE remains majority-owned by our founder, and after Sara's lifetime will become owned by a charitable trust that secures our continued independence.

Los Angeles | London | Washington DC | New Delhi | Singapore | Boston

DATA LITERACY

A User's Guide

David Herzog
Missouri School of Journalism

Los Angeles | London | New Delhi
Singapore | Washington DC | Boston

CHAPTER 1 DATA DEFINED

Before we start exploring our world of data, we need to have a solid grasp of exactly what data are and aren't. This might seem like a technicality that we could ignore, but it's important for us to develop an understanding, one that will prepare us as students or professionals to communicate effectively with others who maintain and share data. If we're able to express what we mean by data, we'll be more effective when we're trying to obtain them from the Internet or by using open-records laws.

The word **data**, of course, is the plural of the Greek word datum, and has been around for centuries. So it is nothing new. However, many people misunderstand and misuse the term data. When people think, talk or write about data, they often are referring to information generically. For instance, they might say, "Those are some interesting data points," or, "Do you have data to back up that claim?" In those cases they really mean, "That's some interesting information," or, "What evidence do you have to back up your claim?" In these cases, they might be referring to information that's stored as text, statistics, tables or even charts.

Dictionaries can help us get closer to the definition of data that we'll use throughout this book. It's true that some dictionaries define data more loosely. Merriam-Webster's (n.d.) primary definition of data is this: "factual information (as measurements or statistics) used as a basis for reasoning, discussion, or calculation." The *Oxford Advanced American Dictionary* (Oxford, n.d.) offers a similar primary definition: "facts or information, especially when examined and used to find out things or to make decisions."

If we read past those primary definitions, though, we get definitions that are more on point. Oxford's secondary definition of data is, "information that is stored by a computer" (Oxford, n.d.). This is correct, but not the whole story. After all, this definition would include any music files, Word documents or photos that you might have stored on your computer. Merriam-Webster gets a wee bit closer with its third definition: "information in numerical form that can be digitally transmitted or processed" (Merriam-Webster, n.d.). Data files can be digitally transmitted over computer networks and processed using programs, because data are composed of bits. **Bits** are the smallest units of computerized data.

None of these definitions nails it for our purposes or those of many professionals who work with data, however. So we will consider data to be any computerized file that uses columns and rows (a tabular structure) to organize data that are represented as text, numbers, dates and the like. In addition, these files can be manipulated using programs like

spreadsheets and database managers, and can be transmitted over computer networks. We'll learn more about these types of files soon.

	A	B	C	D	E
1	MFR_FIRM_FEI_NUM	LGL_NAME	LINE1_ADR5	LINE2_ADR5	CITY_NAME
2		3008347634 Agostino Recca Conserve Alimentari Srl	Contrada Santa Maria		Sciacca
3		3007364564 Banaful & Co.	397 Sk Mujib Road		Chittagong
4		3004398537 Prince Food Products	Commercial Plot No 2, Main Road 1,	Block B, Section 1, Mi	Dhaka
5		3007650858 Bengas ltd	Doulatdia,Chuadanga,bgangladesh.		Chuadanga
6		3008518445 Square Consumer Products	Meril Road	Pabna 6600	Salgaria
7		3004276258 Barnier	Zia Du Barier	34110 Frontignan	Montepellier
8		3004276258 Barnier	Zia Du Barier	34110 Frontignan	Montepellier
9		3004276258 Barnier	Zia Du Barier	34110 Frontignan	Montepellier
10		3004276258 Barnier	Zia Du Barier	34110 Frontignan	Montepellier
11		3004276258 Barnier	Zia Du Barier	34110 Frontignan	Montepellier
12		3004276258 Barnier	Zia Du Barier	34110 Frontignan	Montepellier
13		3004276258 Barnier	Zia Du Barier	34110 Frontignan	Montepellier
14		3004276258 Barnier	Zia Du Barier	34110 Frontignan	Montepellier
15		3004255266 Shanxi Changzhi Yunhai Foreign Trade Meat Co. Ltd	NO.41 Changan Road		Changzhi
16		3004255266 Shanxi Changzhi Yunhai Foreign Trade Meat Co. Ltd	NO.41 Changan Road		Changzhi
17		3008356772 Yiyuan Haida Foodstuffs Co Ltd	Lucun Town Yiyuan Countyzibo City		Shandong
18		3008356772 Yiyuan Haida Foodstuffs Co Ltd	Lucun Town Yiyuan Countyzibo City		Shandong
19		3004251160 Anhui Fuhuang Chaohu Sanzhen Co. Ltd	Huangjiu Town,		Chaohu
20		3004251160 Anhui Fuhuang Chaohu Sanzhen Co. Ltd	Huangjiu Town,		Chaohu
21		3009521500 THIEN MA SEAFOOD CO, LTD - FACTORY 3	2.11E STREET 9, TRA NOC II ZONE	O MON DISTRICT	Ho Chi Minh
22		3008518445 Square Consumer Products	Meril Road	Pabna 6600	Salgaria
23		3008518445 Square Consumer Products	Meril Road	Pabna 6600	Salgaria

Source: Food and Drug Administration.

Note: Data in a spreadsheet.

Data files may hold information that's been summarized or aggregated in some way. In this example, these data about bankruptcies filed in U.S. courts have already been summarized by federal court districts, circuits and type of bankruptcy (United States Courts, n.d.a). We can easily read the data in this table and see that, in the federal court district for Massachusetts, 3,207 people and businesses filed for bankruptcy during the first quarter of 2013. Of those, 33 filed for Chapter 11 bankruptcy, in which businesses are allowed to reorganize and continue operating (United States Courts, n.d.b).

The purpose of the Excel file that holds the bankruptcy summaries is to provide meaningful information to people who may not know how to manipulate raw data. Anyone who's able to download this Excel file of summaries can get an understanding of bankruptcy filing activity in this quarter. Think of it as a report, with the information already baked in.

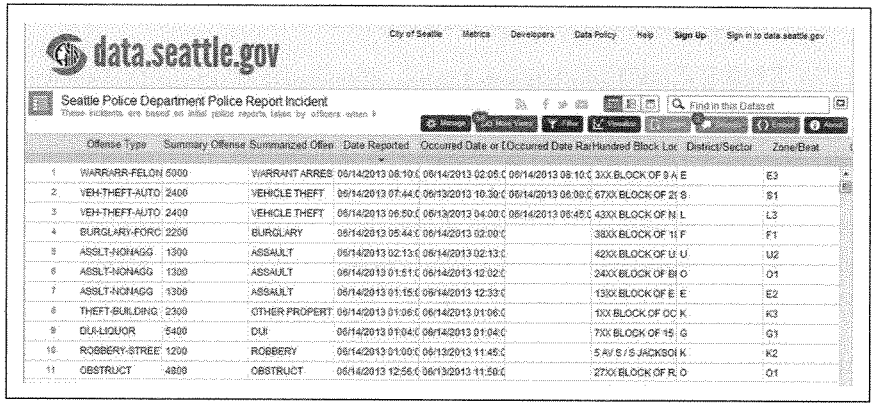
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Table F-2. U.S. Bankruptcy Courts—Business and Nonbusiness Cases Commenced, by Chapter of the Bankruptcy Code, During the Three-Month Period Ending March 31, 2013, Based on Data Current as of March 31, 2013														
2							Predominant Nature of Debt [1]								
3							Business Filings					Nonbusiness Filings			
4	Circuit and District	Total Filings	Total Chapter 7	Total Chapter 11	Total Chapter 12	Total Chapter 13	Total	Chapter 7	Chapter 11	Chapter 12	Chapter 13	Total	Chapter 7	Chapter 11	Chapter 13
5	TOTAL	272,296	189,083	2,345	103	80,737	8,512	5,763	1,996	103	689	263,784	183,380	355	80,048
6	DC	203	166	3	0	34	11	8	2	0	1	192	158	1	33
7	1ST	7,988	5,230	101	9	2,648	327	183	84	9	51	7,661	5,047	17	2,597
8	ME	514	409	5	0	100	35	24	5	0	6	479	385	0	94
9	MA	3,207	2,553	33	3	618	92	55	22	3	12	3,115	2,498	11	606
10	NH	869	631	10	0	228	65	36	7	0	22	804	595	3	206
11	RI	889	752	3	0	114	33	30	2	0	1	836	722	1	113
12	PR	2,529	885	50	6	1,588	102	38	48	6	10	2,427	847	2	1,578
13	2ND	11,190	9,131	243	2	1,803	564	283	237	2	31	10,626	8,848	6	1,772
14	CT	1,728	1,481	29	0	218	90	59	29	0	2	1,638	1,422	0	216
15	NY, N	2,016	1,576	13	2	425	49	26	13	2	8	1,967	1,550	0	417
16	NY, E	3,546	3,139	64	0	340	149	79	63	0	3	3,397	3,059	1	337
17	NY, S	2,246	1,758	125	0	356	219	82	120	0	10	2,027	1,676	5	348
18	NY, W	1,429	989	12	0	419	50	30	12	0	8	1,370	959	0	411
19	VT	234	189	6	0	45	7	7	0	0	0	227	182	0	45

Source: United States Courts.

Note: Bankruptcy Court data. This spreadsheet summarizes bankruptcy information by type of filing and federal court district.

Other data files hold information that has not been summarized; these are considered **raw data**. We usually can tell if data are raw because one row contains data about one person, place or thing. With the bankruptcies, a raw data file would provide one row with detailed data about each bankruptcy filed by a business or a person. It would probably have, in the first row, headers to tell us what each column holds.

The snapshot of the city of Seattle Police Department's real-time incident reports provides a great example. Each row in the table represents a police report that an officer took when he or she responded to an incident. As you can see from the headers, the officers record data about the offense type, code, location, date, time and other details.



	Offense Type	Summary Offense	Summarized Offen	Date Reported	Occurred Date or I	Occurred Date Ra	Hundred Block Loc	District/Sector	Zone/Beat
1	WARRARR-FELON	5000	WARRANT ARRES	06/14/2013 08:10:0	06/14/2013 02:05:0	06/14/2013 08:10:0	3XX BLOCK OF 9 A	E	E3
2	VEH-THEFT-AUTO	2400	VEHICLE THEFT	06/14/2013 07:44:0	06/13/2013 10:30:0	06/14/2013 08:09:0	67XX BLOCK OF 2(	S	S1
3	VEH-THEFT-AUTO	2400	VEHICLE THEFT	06/14/2013 06:50:0	06/13/2013 04:00:0	06/14/2013 06:45:0	43XX BLOCK OF N	L	L3
4	BURGLARY-FORC	2200	BURGLARY	06/14/2013 05:44:0	06/14/2013 02:00:0		38XX BLOCK OF 11	F	F1
5	ASSLT-NONAGG	1300	ASSAULT	06/14/2013 02:13:0	06/14/2013 02:13:0		42XX BLOCK OF U	U	U2
6	ASSLT-NONAGG	1300	ASSAULT	06/14/2013 01:51:0	06/14/2013 12:02:0		24XX BLOCK OF BI	O	O1
7	ASSLT-NONAGG	1300	ASSAULT	06/14/2013 01:15:0	06/14/2013 12:33:0		13XX BLOCK OF E	E	E2
8	THEFT-BUILDING	2300	OTHER PROPERT	06/14/2013 01:06:0	06/14/2013 01:06:0		1XX BLOCK OF QC	K	K3
9	DUI-LIQUOR	5400	DUI	06/14/2013 01:04:0	06/14/2013 01:04:0		7XX BLOCK OF 15	G	G1
10	ROBBERY-STREE	1200	ROBBERY	06/14/2013 01:00:0	06/13/2013 11:45:0		5 AV S / S JACKSOI	K	K2
11	OBSTRUCT	4800	OBSTRUCT	06/14/2013 12:56:0	06/13/2013 11:50:0		27XX BLOCK OF R	O	O1

Source: https://data.seattle.gov/Public-Safety/Seattle-Police-Department-Police-Report-Incident/7ais-f98f.

Note: Seattle Police Department incident data.

Municipalities, counties and states provide data files like these as part of open-government initiatives that have gained momentum during the past several years. Seattle and other cities make data files like these available so citizens and other residents can better understand how government operates. In addition, the city makes these files available in formats that developers can use to build mobile applications and Web applications that display crime locations.

However, to someone who is untrained in the ways of data, this file would be nearly useless. All an untrained person could do is download the incidents file, open it in a spreadsheet program and then scroll through it, looking for information. No one has run any calculations yet, or generated a report.

Still, we usually prefer to get raw data. Raw data afford us more flexibility when it comes to analysis. We can make our own decisions about how to summarize. We could use the file from the Seattle Police Department to determine which block in the city has the greatest number of reported crimes, or the greatest number of reported burglaries. Additionally, raw data make it easier for us to examine our data and test the integrity—or quality—of our data, as we will do later in Chapter 5. For more-sophisticated data users, raw data are better.

In the U.S. Courts System summarized bankruptcy file seen earlier, we lack specific data about the individual cases, so we would be limited in what we could do in our

analyses. If we really wanted to perform more-sophisticated analyses, we would have to find raw data online. Failing that, we would need to request the data from the U.S. Courts System.

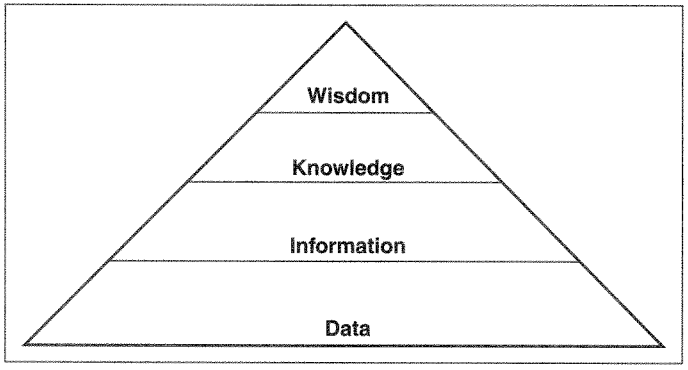
The data files that we will work with in this book all are considered **secondary data**, or data that are collected by individuals other than ourselves. Some academic disciplines and professions analyze data that they've collected themselves; these are considered **primary data**. Primary data are used heavily in the natural and social sciences, as well as in medicine. Such data are sometimes obtained by the process of sampling, or selecting a subset from the whole population of whatever it is you are studying. As you progress in your academic career, you might work more with primary data.

Here are some illustrations of primary data collection:

- A research psychologist interested in the effectiveness of televised antismoking advertisements enlists subjects to watch the ads. The psychologist collects data about the subjects' reactions by asking a series of questions.
- A medical researcher testing an experimental cancer drug enlists subjects, some of whom get the drug while others get a placebo. The researcher uses blood tests to record white cell counts as a way of assessing the effectiveness of the drug.
- A political polling consultant hired by a candidate running for office uses telephone interviews to ask questions of likely voters about what issues are important to them.

CLIMBING THE PYRAMID

Data are our starting point—the raw material that we use to understand our world better. As such, data sit at the bottom of the data-information-knowledge-wisdom (DIKW) hierarchy. Operations research expert R. L. Ackoff spoke about this hierarchy more than two decades ago (Ackoff, 1989).



Source: R. L. Ackoff; illustration by author.

Note: The DIKW pyramid.

"The implicit assumption is that data can be used to create information; information can be used to create knowledge, and knowledge can be used to create wisdom," wrote Jennifer Rowley, a professor of marketing and management at the University of Bangor in the United Kingdom (Rowley, 2007, 164.)

A lofty goal, but possible. Here's how the pieces fit together and how we can use them to ascend to the pyramid's peak.

Data, the bottom of the pyramid, are the unprocessed symbols that represent characteristics of objects or people. These data are unorganized and have no meaning or value in isolation (Rowley, 2007). Because data constitute the base of the pyramid and cover a greater area, we can conclude that we are awash in data relative to the other elements in the pyramid. Also, we can see that data serve as the foundation for all of the other levels.

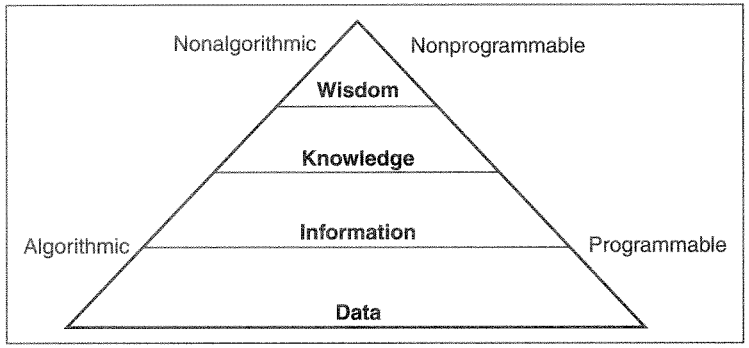
Data can be used to step up to the next level of the pyramid: information. We create information by processing our data using computer programs. For example, we might use a spreadsheet like Microsoft Excel to create totals or averages—descriptive statistics that are more meaningful to us. Or, as we will do later in this book, we might calculate rates or ratios using our data. We could also use database managers like Microsoft Access to perform more-sophisticated analyses, something that is outside the scope of this book. Information is descriptive: it answers questions like Who? What? Where? When? (Rowley, 2007).

After we've generated information, we aim to generate knowledge, which requires us to reflect on the information. Knowledge is created when we take information and turn it into something that can be acted upon (Rowley, 2007). Knowledge helps us and all kinds of professionals—such as teachers, political researchers or business analysts—make better decisions.

As Nate Silver, the founder and editor in chief of ESPN's FiveThirtyEight, put it, "The world has come a long way since the day of the printing press. Information no longer is a scarce commodity; we have more of it than we know what to do with. But relatively little of it is useful. We perceive it selectively, subjectively, and without much self-regard for the distortions that this causes. We think we want information when we really want knowledge" (Silver, 2012, 17).

Helping you learn how to get from data to knowledge is one of the central goals of this book. After you have experience getting from data to knowledge, you may even be able to climb to the peak and attain wisdom. But that takes experience, along with the exercise of values and judgment (Rowley, 2007).

One important point to keep in mind is that working effectively with data requires more than computing. It requires thought and reflection on our own parts. As we climb the pyramid, we rely less on computers and more on our own experiences and thoughts. E. M. Awad and H. M. Ghaziri accounted for this dynamic when they modified Ackoff's



Source: Awad and Ghaziri; illustration by author.

Note: Modified DIKW pyramid.

DIKW pyramid, shown above, for their business knowledge management textbook (Awad and Ghaziri, 2004).

As you can see, this modified pyramid shows that the data at the base can be manipulated by computer programs or **algorithms**—a set of steps for solving a mathematical problem. The steps up to information and knowledge are less reliant on programs and algorithms, and wisdom is not reliant on them at all.

A BRIEF HISTORY OF THE DATA WORLD

With all the recent buzz over open-data portals, big data and data science, it's easy to overlook the roots of data and computing, which stretch back to 20,000 B.C. (The following discussion is based on Wolfram Alpha n.d.) That's when our predecessors invented arithmetic as a tool for calculating numbers of objects. Then, in the years from 2150 to 1700 B.C., they created standards for measurement and multiplication. Later, in 500 B.C., Greek scholar and mystic Pythagoras promoted the idea that the world could be understood by numbers.

Much later, during the Renaissance, Nicole Oresme developed the idea of representing numbers by using graphs, which continue to be important tools for communicating data. Nearly three centuries later, Wilhelm Schickard invented a wooden machine that could add up to six digits.

The Industrial Revolution brought even more innovations that helped create our world of data, including Joseph Marie Jacquard's use of punch cards to control the looms at his weaving mill in France, Charles Babbage's early mechanical computers and Herman Hollerith's use of punch cards to automatically tabulate results from the decennial U.S. census.

The 1940s ushered in the era of digital computers, those that used vacuum tubes to perform calculations on data stored as a series of numbers. In 1963, the American

Standards Association developed the data-encoding system that still dominates computing in the United States. This system is called **ASCII**—short for the American Standard Code for Information Interchange—and allows us to easily share data. ASCII text files are the most portable data files available. The 1970s brought other advances, such as relational **database manager** programs, interactive computing and the first personal computers.

DATA FILE FORMATS

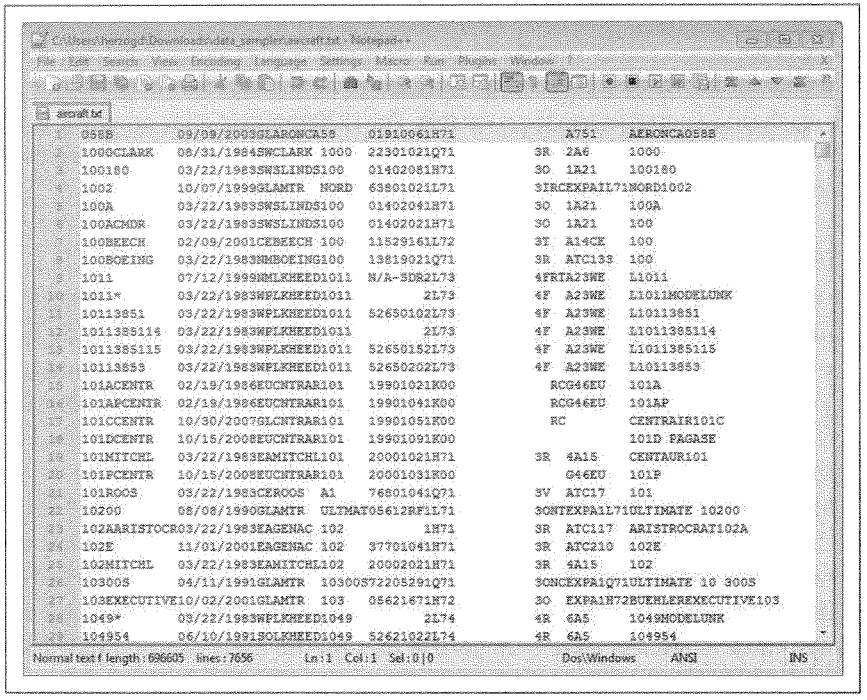
The data files that you store on your computer, whether it's a Mac or PC, can come in hundreds of formats. Audio, video and graphics files all are stored differently and have their own **file extensions** (those characters that come after the file name and period) (FileInfo.com, n.d.). For this book, we'll focus on using just a few different types of data files that we can analyze using **spreadsheet programs** such as Excel or OpenOffice Calc.

Microsoft has two file formats for Excel workbooks, which can hold multiple worksheets. Excel 1997–2003 format workbooks have an **.xls** file extension. A more recent file format, introduced with Office 2007 for Windows, has an **.xlsx** extension and is based on **XML**, or Extensible Markup Language. Microsoft says the XML-formatted files are smaller, more robust and more interoperable than the .xls files (Microsoft Developer Network, n.d.).

Aside from using Excel's native files, we can use files in other formats. Excel and Calc can open OpenDocument spreadsheet format files. The files have an .ods extension and were developed as an XML-based open-source alternative to Excel's proprietary formats. Also, both spreadsheet programs can work with the proprietary **dBASE** database file format. dBASE files have a .dbf extension

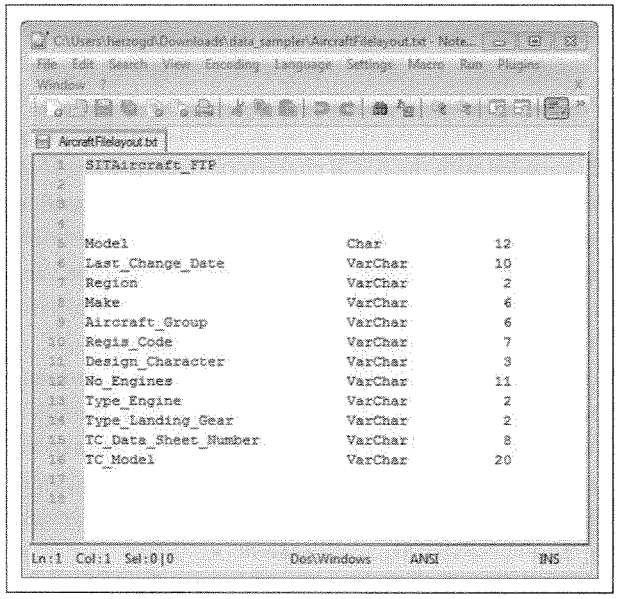
Text files are perhaps the most useful format of all because they can be read, processed and exported by all computers and data analysis programs. In the United States, these files are usually encoded with the ASCII characters, making them readable by mainframe computers, servers, Macs and PCs alike. These text files come in two different flavors: **fixed-width** and **delimited**. You can examine both types using a **text editor**, such as **Notepad++**, which is a free and **open-source software** Windows program that's able to handle large files. **TextWrangler** is a good choice for Macs that is also free.

The data inside fixed-width text files already are nicely arranged into columns and rows. The table looks just how we'd expect a data table to look. The following example is a table of aircraft types compiled by the Federal Aviation Administration. Each of the 7,655 lines contains data about one type of aircraft. Eyeballing the file, we might make some solid, educated guesses about where column breaks should go, but we really can't be sure.



Source: Federal Aviation Administration.

Note: Federal Aviation Administration aircraft data. The data are arranged in columns, which tells us this is a fixed-width text file.

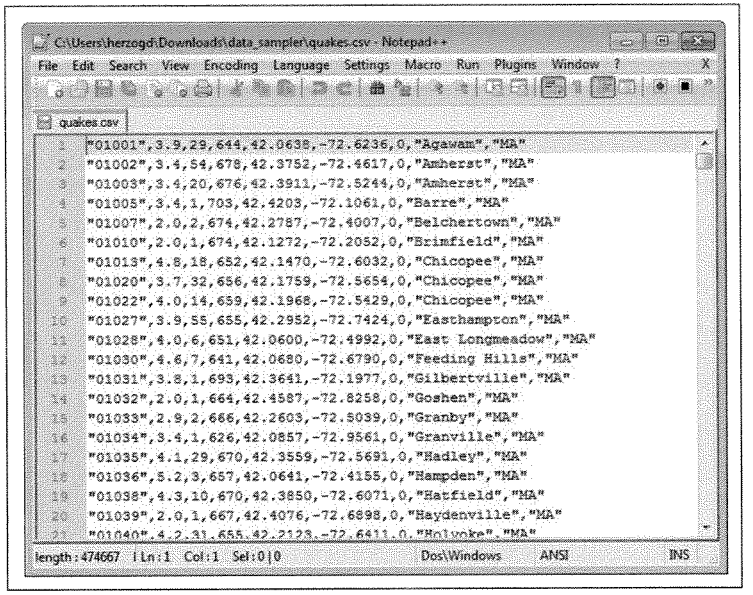


Source: Federal Aviation Administration.

Note: Documentation for FAA aircraft file.

Fortunately, the FAA provides documentation that tells us what the names for the columns should be, and where the breaks go. In this case, we see Model is the name of the first column, and that it is a character column 12 characters wide.

In contrast, delimited text looks like a mess, as you can see from the following earthquake report data released by the U.S. Geological Survey. The data are all smashed together—it looks like a train wreck. However, if you look closely, you will see that commas separate different pieces of data in each row. We call these comma-separated or comma-delimited text files. Data can be delimited by other ASCII characters such as the tab, pipe (|), tilde (~), exclamation point (!) or carat (^). Even though these data look gnarly, they are easier to import than data in a fixed-width file, because computer programs can read the **delimiters** and automatically determine where the column breaks go.



Source: Geological Survey.

Note: A comma-delimited text file. This delimited file uses commas as column separators and double quotation marks to denote text.

There is no convention for specifying extensions with text files, but you will often see .txt, .csv (comma-separated values), .tsv (tab-separated values), .tab, .prn or .dat. If you ever get a text file with an extension that Excel is unable to recognize, simply rename it with one of the extensions it does recognize.

Now that you know what data are, we're going to learn about clues that will help you discover the data you need.