QUANTITATIVE ETHNOGRAPHY

DAVID WILLIAMSON SHAFFER

BOSWELL PRESS

AVAILABLE APRIL 2017

For my students

Who have taught me so much

We live, not by things, but by the meanings of things.

- Antoine de Saint-Exupéry,
Generation to Generation

Table of Contents

Foreword

Since the 1990s, the hard sciences have been undergoing a revolution. As scientists came more and more to study complex systems (such as the environment, stars and galaxies, atoms and electrons, cells and molecules, the human microbiome, the global economy, and the human brain), they realized that the problems they faced were far too hard and complex to be solved by any one discipline and by any single method. Scientists began to define what they did and who they were, not in terms of their discipline or sub-discipline per se, but in terms of a hard problem or big challenge they faced with others not in their own discipline. They began to work on complex systems with specialists outside their areas and to develop a new common language and shared methods. What resulted is not inter-disciplinarity (often just a form of academic parallel play), but a new cross-functional integrated approach to science.

A similar trend has emerged over the same span of time in high-tech workplaces, where work is now organized around deeply skilled teams that share a broad understanding of each other's specialties and can integrate their expertise to solve big and hard problems. These advances in science and at work have been facilitated, in part, by the growth of new digital technologies, Big Data, and new data-mining techniques.

Today, studying the brain takes chemists, physicists, biologists, psychologists, computer scientists, mathematicians, philosophers, and even graphic artists who together become "neural scientists" or, at least, contributors to "neural science." And, recently, these "brain people" have discovered we humans have a second brain in our gut and so they now have to talk to "gut people" as well—who would have thought! To make a video game today, it takes game designers, programmers, directors, engineers, visualization experts, content experts, motion-capture experts, artists, and writers, among others.

This revolution has not gone nearly as far in education, psychology, and much of the social sciences. The reason is that these areas are really not "hard sciences," but "hard hard sciences." They are areas that face hard problems where human beings are in the mix. Human beings, as individuals and as groups, are complex systems in their own right. Put them inside other complex systems and complexity grows exponentially.

The nasty problem humans (unlike atoms, cells, or stars) add to complexity is meaning-making. Human meaning-making is a computationally intractable problem (note how search engines clearly know nothing about meaning; they just know correlations among words). No machine yet can understand, let alone make, meaning in the way humans do (and the good bet here is no machine ever will). Scientists who study their fellow humans are interpreting beings that can interpret right back and can change their interpretive frameworks in the bat of an eye. It is as if the cells in a petri dish could plan with each other how to surprise the scientist studying them.

David Williamson Shaffer is, in this book, ringing the bell to announce the revolution is coming to education and the social sciences. A polymath if there ever was one, Shaffer has worked with some of the most diverse scientists and colleagues I know. This book is very much the result of that team work. While the signs on the doors in Shaffer's department say things like "psychology" and "learning science," he was trained at the famed Media Lab at the Massachusetts Institute of Technology and has the revolution in his blood. He doesn't just want to contribute to education, psychology, learning science, and understanding of humans as social actors, as he has done for years now; he wants to remake these disciplines, integrate them, discover new language and new methods, and bring the revolution home.

Shaffer integrates big data, data-mining, discourse analysis, social interactionism, cognition, learning science, statistics, and ethnography into a brand-new integrated human science. We see clearly how the teams of the future will need to be put together. We see, finally, a way to take on the hard hard humans-in-the-mix problems here. We see how to make science where we have too often been saddled by a stale trade-off between ungeneralizable "qualitative" anecdotes and vapid "quantitative" p-values with too little real power, save to furnish publication mills. Here we get numbers and meaning both, and they don't fight each other, rather they give birth to truly new ideas and innovative ways out of our old ruts. Those interested in teaching, learning, meaning-making, culture, social interaction, and human development will find here the first shot in a real revolution. It's a wild ride and a great read to boot.

— James Paul Gee
Mary Lou Fulton Presidential Professor of Literacy Studies
Arizona State University
March 2017

CHAPTER 1

Introduction: Captain's Log

Kirk's legacy

"Captain's log, Stardate 3614.9"

The plot device of my childhood was the captain's log. Like the national anthem at the beginning of a baseball game, each episode of my favorite television show, *Star Trek*, began with a log entry from James T. Kirk, the swashbuckling captain of the starship *Enterprise*. Kirk dictated a description of the ship's mission and stored it in the ship's computer, with the audience conveniently listening in to the captain's monologue as a voice-over.

In an episode in the second season, "Wolf in the Fold," a mysterious string of murders takes place on a planet the *Enterprise* is visiting, and one of the crew is a suspect. Kirk's log introduces the problem:

> Captain's log, Stardate 3614.9. Planet Argelius Two. While on therapeutic shore leave, Mister Scott has fallen under suspicion of having brutally murdered an Argelian woman. The chief city administrator, Mister Hengist, has taken charge of the investigation, but has learned little of value.

Even as a child I knew the log entries were just a way to give the viewer information about the story to come without using a narrator or having one character explain to another what was happening. But what I didn't realize at the time was that this simple plot device would be one of the most prescient of *Star Trek*'s predictions for the future.

Science fiction stories from the past have long been a source of inspiration for actual science in the present. Jules Verne and H.G. Wells both wrote about space travel a century before the first Apollo moon landings. In Verne's 1865 novel *De la terre à la lune* (*From the Earth to the Moon*), astronauts were shot in a hollow sphere from a cannon. The Martian invaders in Wells' 1897 serial novel *The War of the Worlds* used the same method to get from Mars to the Earth—although later in Wells' 1901 *The First Men in the Moon* an inventor and a businessman develop a more elegant method using "cavorite" to escape from Earth's gravity.[1]

Two years after *War of the Worlds* was published, a 17-year-old boy in Worcester, Massachusetts, wrote in his diary: "How wonderful it would be to make some device which had even the possibility of ascending to Mars." The boy was Robert Goddard, a pioneer in the science of rocketry who developed 214 patents for rocket engines, parts, and techniques. His first launch in 1926 has been

compared to the first flight of the Wright brothers at Kitty Hawk. Goddard developed the first liquid-fuel rocket, as well as many of the techniques still used today to stabilize and steer rockets in flight.

Thirty-three years after describing in his journal the dream of going to Mars, Goddard wrote about his passion for the field of rocketry to Wells, the author who had inspired his life's work. Later, Goddard became the director of the American Rocket Society—whose first president was the editor of the science fiction magazine *Science Wonder Stories*.

In other words, fictional stories of space travel starting in the 19th century inspired the science of rocketry and space exploration in the 20th. But the influence of fiction on science runs deep. James Bond first introduced the public to global positioning system (GPS) navigation in the 1964 movie version of *Goldfinger* in 1964 and underwater cameras in the 1965 film *Thunderball*. And in the early 1970s, when telephones were still connected to the network with cords, two telecommunications companies, Motorola and the giant AT&T, battled for market share, when a Motorola engineer named Martin Cooper stumbled on a *Star Trek* episode. "People are fundamentally, inherently mobile," he said, describing the incident years later. "They never, never would want to be leashed, tied to a desk or to their home or to their office if they have a choice. … And suddenly, there's Captain Kirk talking on his communicator." From this moment of inspiration, Cooper went on to develop the first cell phone, and it is no coincidence that until 2009 when smartphones came into wide use, the most popular phones—so-called "flip" phones like Motorola's famous RAZR—looked much like Kirk's communicator.[2]

Star Trek gave the world a preview of personal computers, tablets, portable computer memory, biometric scanning, and wireless headphones. So inspirational was the technology of *Star Trek* that the original model of the starship *Enterprise* hangs in the Boeing Milestones of Flight Hall at the Smithsonian Air and Space Museum. And long before William Gibson coined the term *cyberspace* in his short story "Burning Chrome," and Tim Berners Lee and his colleagues at CERN, the European Organization for Nuclear Research, invented the World Wide Web—even before the engineers at Bolt Beranek and Newman developed the first working network of computers—*Star Trek* provided a vision of what we now call *cloud computing*.[3]

The computers on the *Enterprise* stored a vast record of the knowledge of the species of the known galaxy that the crew could consult for the asking—it being more dramatically interesting to hear a character ask a question directly rather than watch someone type it into a computer. In "Wolf in the Fold," the mysterious murders are solved when the crew conducts a seance (yes, really) and the spirit medium uses the word "redjac." The ship's science officer, the brilliant Mr. Spock, then uses what we now think of as a simple Google search:

SPOCK: Computer, linguistic bank. Definition of following word: "redjac."

COMPUTER: Working. Negative finding.

SPOCK: There is no such word in the linguistics bank?

COMPUTER: Affirmative.

SPOCK: Scan all other banks.

COMPUTER: Working. Affirmative. A proper name.

SPOCK: Define.

COMPUTER: Redjac. Source Earth, nineteenth century. Language, English. Nickname for mass murderer of women. Other Earth synonym, Jack the Ripper.

One "Google" search later, the crew discovers the murderer is in fact (gasp) Mr. Hengist, the administrator who was in charge of the investigation in the first place!

What made *Star Trek*'s captain's log so prescient, though, was that it was part of a computerized record of *everything that happened on the ship*. In the episode "Court Marshal," Kirk is accused of causing the death of a member of the crew, and the evidence against him is a video recording of his actions that day, stored in the computer. Naturally, it turns out that Kirk was framed: The computer's files had been altered.

"Court Marshal" aired two years before the first computers were linked together in what would eventually become the Internet, when data was still stored on punch cards, which looked like stacks of index cards where each card had one

line of information on it. But on the *Enterprise*, everything that everyone said and did—all their actions, their decisions, and their explanations—everything was recorded and could be examined and re-examined, searched and sorted and analyzed.[4]

In 1967, *Star Trek* had already imagined the world of Big Data that we live in today.

Data, data, everywhere

It seems unimaginable in a world where almost everyone carries a digital camera as part of the phone in his or her pocket, but before 1900 almost no one took pictures at all. In that year, the Eastman Kodak Company made the first popular camera, the Brownie, and by 1905 (only five years later) 10 million people in the United States were taking "snapshots." Before the Brownie, though, most people in the US—indeed most people in the world—lived their entire lives without ever having their picture recorded. Goods and services were traded in barter or paid for with cash. Notes in a family bible might record births and deaths. But unless a person made a concerted effort to keep a diary or write and save letters, it was entirely likely that the only permanent trace of his or her life would be a few official records like marriage certificates, an entry in the census every 10 years if he or she lived in the United States, and a name on a tombstone.[5]

Now, though, we live among the towering mounds of data that accumulate as we move through life in the digital age. Even if we never post an entry to the modern day captain's logs of Facebook or Twitter or Snapchat or Instagram, every time we swipe our credit card, send an email, search for information, make a phone call, save a picture, or even walk down the security camera-filled streets, we create digital footprints that mark our path.

Eric Schmidt, chairman of Google's parent company, Alphabet, estimates the world records as much information *in two days* as was created from the beginning of written records 40,000 years ago through 2003. Nearly one third of the people on the planet has at least one social network account. Every minute they make 4 million posts, and upload 400 hours of video to YouTube. Every day, we create five exabytes of data, or enough to fill the hard drives of 5 million computers: a gigabyte of data for every man, woman, and child who is connected online. Two-thousand books worth of information, or 160 digital pictures, per person per day.[6]

In the digital age we are all Sherlock Holmes and cyberspace is our Dr. Watson, chronicling our every move; or if you prefer, we are Samuel Johnson to the Internet's Boswell.

Star Trek was prescient in predicting that our lives would be automatically and continuously recorded, although as is often the case with predictions, the consequences were not entirely clear. For one thing, the characters on *Star Trek* never seemed particularly worried about privacy. The world has quickly learned that Big Data is an invitation to Big Brother, whether in the form of the National Security Agency recording phone calls and screening email, Facebook choosing what it shows us to change our moods, or Google tracking our searches and the web pages we visit to decide which advertisements to show us.[7]

Nor, of course, was there much concern on the *Enterprise* about data security. Few people worried in the fictional future about problems that we encounter all the time in the age of Big Data: pictures being stolen, credit card numbers secretly recorded, websites hacked, and passwords compromised. In one notorious incident in 2014, a whole Twitter account was stolen from its user in an elaborate extortion scheme, and such incidents are sadly becoming almost commonplace today.[8]

Thefts certainly happened in the universe of *Star Trek*. Notable "hacks" include an alien carrying off Mr. Spock's brain to run the massive computer controlling the underground ecosystem of the far-off planet Sigma Draconis VI, and the Bynars gaining control of the *Enterprise* computer to hijack the ship. But the crew of the *Enterprise* overall seemed remarkably unconcerned about passwords, identity theft, online bullying, and other day-to-day worries of our own digital age.

The optimist in me likes to think that the reason data privacy and security are not big issues on *Star Trek* is that by the 23rd century these problems will be mostly resolved. Presumably biometric scanning and 200 years of progress on encryption algorithms will make data much harder to steal.

Of course privacy and protection of data are clearly important aspects of any ethical research—and really any ethical behavior, although sadly corporations and governments do not always see things that way. We'll touch on these issues in what follows, but the focus of this book is less on *whether it is OK* to use one or another kind of Big Data than it is on *what to do with the data* once you have it. Because although *Star Trek* provides a very clear image of the Big part of Big

Data—a world where everything everyone does is automatically recorded—the creators of *Star Trek* did not really foresee the limitations of Data itself.

Here is a simple example of what I mean.

Imagine I tell you that a certain person—let's just call her "Cassie"—used a credit card to buy $13.73 worth of unleaded gasoline on Saturday at 4:14pm in Gulfport, Mississippi. A piece of data about Cassie's life. OK, now, what sense would you make of it? Could you conclude that Cassie lives in Gulfport? Or in Mississippi? Or even in the United States? Probably not. She could be visiting Mississippi. Maybe on business. Maybe for vacation. Or maybe she isn't even there at all. Maybe someone stole her credit card. Even assuming that her card wasn't stolen, would you know she owns a car? Maybe she is renting a car. Or maybe she borrowed a car. Or maybe Cassie was taking a trip with a friend in the friend's car and paying for the gas.

So the first thing to notice is that it is almost impossible to conclude anything from this single piece of data. But let's imagine that we have some more data. Cassie's birth certificate is from 1996 in Austin, Iowa. She also purchased gas in Lake Charles, Louisiana, at 11:52am on Saturday, and in College Station, Texas, at 7:01am earlier that morning. Now, if you took the time to plot those points on a map, Cassie would certainly seem to be driving from Texas, through Louisiana, and into Mississippi. And if I told you that this Saturday was a day at the end of March during Texas A&M University's spring break … well, if you remembered that College Station is the home of Texas A&M, you might guess that Cassie is a college sophomore driving to Florida for vacation. You could not be sure, of course. There are an infinite number of other possible explanations. But if you accumulated more and more data consistent with your guess, you would be more and more confident about who Cassie was and what she was doing.

There are a number of technical ways of describing this kind of "guessing informed by data." Depending on how we were actually looking at the data we might describe the idea that Cassie is driving to Florida for spring break as a *hypothesis* or an *inference*. But whatever technical term we use, the point is the same: We are trying to make sense of the data, and the way to do that is to try to *understand what is going on*.

Until we do that—until we come up with an explanation for what we think is going on—the data itself is meaningless. Once we start to understand what the data *means* we actually transform it: the data becomes *information* that is part of a story about something that is happening in the world (or that has happened, or

that is likely to happen). All of which is just to say that actually the Internet is neither Dr. Watson nor Dr. Boswell at all. It is not telling the story of our lives. Rather, Big Data is like a medieval chronicle, recording disconnected details. We need another set of tools to make sense out of data. We need a method to create meaning from the exabyte mountains in which we travel.[9]

Information is always the combination of data and meaning, and at the most basic level this book is about how we can reliably and systematically convert Big Data into Big Information—and how we can use Big Information to get Big Understanding. It is about how we can use the incredible volume of data that computers let us collect without making superficial assumptions that lead to trivial or even misleading conclusions about what data tells us.

Pretty little data all in a row

The good news, of course, is that the same technological advances that create Big Data give us the tools to analyze it. Computers do not just store details about our lives; they can also examine those details and look for patterns. Computers are not overwhelmed by the mountains of data that we produce every day—although, to be fair the mountains are so vast that most computers can only handle one slope of a mountain at a time.

The term often used for this kind of digging for patterns in mountains of data is, appropriately enough, *data mining*. In the field of education, for example, an entire academic journal publishes articles about the use of computers to search through the piles of data collected in computer-based education: things like computer games, massively open online courses (better known as MOOCs), computer-based tests, and the like. It is called, not surprisingly, the *Journal of Educational Data Mining*. There is an educational data mining conference, run by the International Educational Data Mining Society. But there are also more general places where researchers write about how to find patterns in large collections of data: the *Journal of Data Mining*, for example, or the journal *Data Mining and Knowledge Discovery*.

The basic idea of data mining is actually pretty simple, even if some of the mathematical details are complex: You look for similarities in different sets of numbers. For example, here are two sets of numbers:

Height	Weight
72	161
70	152
67	150
66	149
73	162
70	156
67	145
69	154
74	167
78	170

One set of numbers is the height (in inches) of 10 people, and the other set is the weight (in pounds) of the same people. And the question is, what would it mean to say that these two different sets of numbers are in some way similar?

Mathematically, we could compare these two sets of numbers in many ways. Some methods are more sophisticated than others, but one easy thing to do is just graph the height and weight of each person we have measured. That would look something like this:

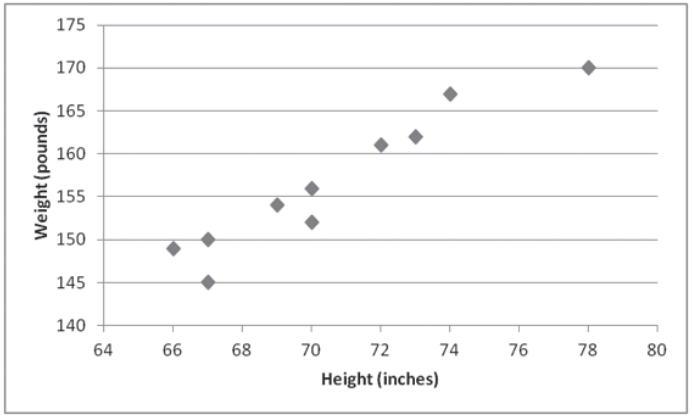

Figure 1.1 **The relationship of height and weight in 10 hypothetical men.**

On this graph, each point represents one of the people whose height and weight we measured, and it is pretty easy to see that the points make a kind of ragged line across the graph.

Wait! That's a pattern!

The points make a line because there is a relationship between a person's height and his or her weight. In general, people who are taller weigh more than people who are shorter. Of course, the "in general" part of that last sentence is critical. There are plenty of exceptions: people who are very tall and skinny, or people who are short but weigh a lot for their height. That is why the line looks ragged or imprecise. And of course the real "line" that relates height and weight is actually quite a bit more ragged.[10]

The mathematical relationship between height and weight was first quantified in the mid 19th century by Belgian statistician Adolphe Quetelet. Quetelet was one of the founders of the social sciences. His most famous work, Sur l'homme et le développement de ses facultés (On Man and the Development of his Faculties) published in 1835 was about what he called l'homme moyen, or the average man. Quetelet imagined the average man as a fictitious person whose characteristics were the mathematical average of everyone in society—something like the Half Boy whom Milo meets in the Phantom Tollbooth. The Half Boy is the .58 of a child in the average family that has 2.58 children, and therefore is the only one who can drive the three tenths of the average family's 1.3 automobiles.11

One of the chapters of Quetelet's Treatise of Man examined changes in height and weight as people get older, as well as the relatively stable relationship between height and weight once people reach adulthood. Based on this work, Quetelet developed a measure of the amount that a person is above or below the "normal" weight for his or her height, called the Quetelet index for many years. Today we use a similar formula, body mass index, or BMI, to determine whether a person is too heavy for his or her height (that is, of he or she is overweight) or if someone is too light for his or her weight (which might suggest malnutrition, problems with digestion, or an eating disorder).

The statistical term for this kind of relationship between two things—in this case how much people weigh and how tall they are—is a *correlation*. That is just a mathematical way of saying that *when height goes up, all other things being equal, weight goes up as well*. Interestingly, and somewhat distressingly, ideas about correlation (and many other statistical ideas) actually arose from measures like Quetelet's Index and more generally in the late 19th century from the field of

anthropometry, the statistical analysis of the human mind and body. Anthropometry, also called *biometry*, was inspired by Charles Darwin's theory of evolution by natural selection. Led by Darwin's cousin Francis Galton, and later by the British statistician Karl Pearson, biometricians tried to use statistics to determine who was "fit" and who was not. The idea of statistical correlation was born and bred, in other words, in the service of eugenics: to understand the differences between the "weak and feeble" and the "better stocks" as a way to get an advantage in the struggle for survival between races and nations.[12]

Along the way, Pearson and others created some of the key concepts in modern statistics, including *Pearson's r*, which measures the strength of a correlation. So in our example above, Pearson's r would tell us how strongly height and weight are related to one another in the group of people we measured. If Pearson's r (usually just referred to as r because it is so commonly used) is 1, the relationship between the two sets of numbers is perfect: In this case, when height goes up so does weight. If r = -1 the relationship is perfectly backward: When height goes up, weight goes down. Values for r range from 1 to -1—from perfectly correlated to inversely correlated and everything in between. When r = 0 there is no relationship at all.[13]

In our little collection of height and weight data, the correlation has r = .95, which is very high, and explains why the points look so much like a line. In data drawn from one actual group of more than 500 people real people who were in good shape, the correlation between height and weight was closer to r = .70. That is still a relatively high correlation, but there is more variation, so the line would look more jagged.[14]

At this point you might be wondering why Pearson's r is so important. After all, the heights and weights line up quite nicely in our little example. We can *see* the pattern without calculating any fancy statistics. But not every pattern is quite so easy to see.

Let me give you another two sets of numbers:

Set 1	Set 2
7	8
2	4
9	6
4	7
4	3
5	2
8	4
6	4
1	1
8	9

Once again we could look for a pattern by creating a graph, and once again the points for Set 1 and Set 2 form a kind of line. The line is a little more imprecise this time, so I added an actual line to the graph that approximates the underlying relationship between the two sets of numbers.

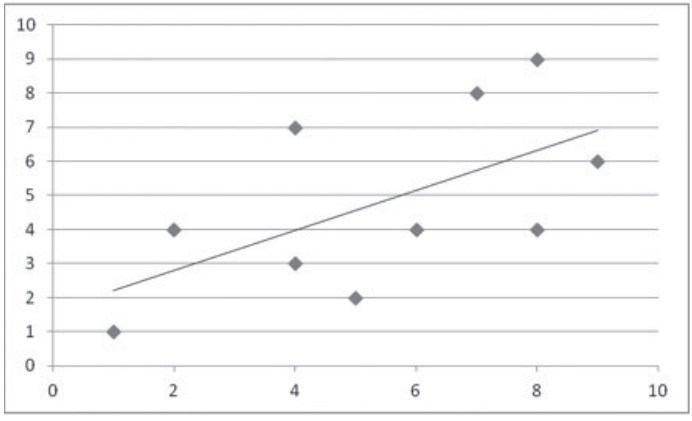

Figure 1.2 **The relationship of two arbitrary sets of numbers where Pearson's r = .60.**

The points in this second graph look more ragged than in the first graph because for this correlation r = .60, which is lower than r = .95 in the first graph. Actually, to be precise, r is lower in the second graph because the points are not as well aligned: The r value describes the strength of the correlation. Pearson's r is not the *reason* the graph looks ragged, it is a *measure of how strong the pattern is*. In fact, even a relatively strong mathematical relationship might be hard to see by eye alone. The graph below has a correlation with r = .50. That may not seem like a very strong relationship, but to put that in perspective, some studies suggest income and happiness are correlated with r = .50. People with more money tend to be happier, although the relationship is not perfect because a lot of things can make you unhappy if you are rich or happy if you are poor. Money alone can't buy happiness, but there clearly is some connection between the two.[15]

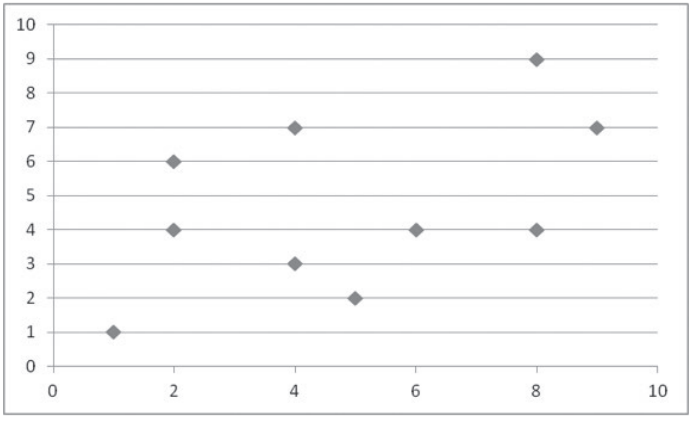

Figure 1.3 **The relationship of two arbitrary sets of numbers where Pearson's r = .5.**

Without a line to illustrate the relationship it is pretty hard to see. Which is part of the reason that Pearson's r is useful: it helps find patterns that are hard to see. There are many other ways to measure the strength of a pattern, of course. And like Pearson's r, many of the methods for quantifying the strength of a pattern are useful because computers are much better at calculating and comparing numbers than they are at deciding whether something does or does not look like a line.

Being able to measure the strength of a pattern makes it possible to use computers to analyze the Big Data that they collect. We can feed a computer data and let it search for patterns by using correlations and other statistical tests to measure how strongly one kind of data is related to another. The science of data mining is the use of statistics to find relationships hidden in massive collections of data—the modern descendant in method, though hopefully not in spirit, of the eugenics movement a century before the birth of the Internet.

Mining shrimp

There is a problem with this kind of data analysis, though, which always reminds me of a riddle my father told me when I was a child:

How do you tell the live shrimp from the dead shrimp in a bucket of shrimp?

The answer, of course, is:

Throw them against the wall and the ones that say "Ouch!" are alive.

To be fair to my father, the original joke was funnier when he told it, I suspect because it was more scatological. But more to the point here: Why is *looking for relationships by using statistical tests to measure the strength of patterns* like throwing shrimp against a wall?

To see why, consider for a moment the last set of numbers that I gave you. It was relatively easy to find a correlation in the data, to measure the correlation, and to conclude that there was a relationship between the two sets of numbers. But here is the thing: I did not say anything about what those numbers mean! So yes, we could use statistics to find a pattern, but without knowing what the numbers represent or where they came from, the pattern is meaningless.

OK. Allow me to remedy that problem:

The first string of numbers (7, 2, 9, 4, 4, 5, 8, 6, 1, 8) represents the number of cherry tomatoes I collected on day in August last year from each of 10 tomato plants in our garden. I collected 7 from the first plant, 2 from the second plant, and so on.

> The second string of numbers (8, 4, 6, 7, 3, 2, 4, 4, 1, 9) represents the height in Lego blocks of the beds my daughter built for her Polly Pocket dolls in a "setup" she made in the basement on the same day. The first doll's bed was 8 Lego bricks tall; the second doll's bed was 4 bricks tall; and so on.

On the same day, in the same house, two seemingly unconnected things were related to one another. More than that, the two things—the number of tomatoes on my plants and the number of Legos in my daughter's doll beds—were not just related to one another, but related to one another quite strongly ($r = .60$). They were related more strongly than income influences happiness ($r = .50$), although less strongly than height is related to weight among physically fit people ($r = .70$).

But now let us make two observations about what just happened. First, this example illustrates the power of data mining to find a relationship where we might never have expected to find one—and not only to find a relationship, but to show us the strength of the pattern in the data. And even more than that, to compare the strength of this unexpected relationship to other patterns and connections in the world. This is just a small, small example. But it shows very clearly what statistical methods can do if they are applied to data. It is easy to imagine what patterns we might find if and when we mine the mountains of data generated by computers in the age of the Internet.

So that is the first point: Data mining has the power to find new and unexpected relationships in Big Data.

But there is a second point that is actually more important to be made about this example: This statistical analysis found a relationship that is completely meaningless. It would be extremely strange indeed if somehow there was a connection between the number of cherry tomatoes in my garden and the number of Lego bricks my daughter used in her doll beds on any particular day. In fact, it would even be a little strange to discover that my daughter had somehow (consciously or unconsciously) made the same number of doll beds as tomato plants in our yard—or equally strange, that I deliberately chose to plant one tomato vine for each of my daughter's Polly Pocket doll. One of these things might have happened purely by chance, but it is hard to imagine this pattern has any more meaning than that.

In fact, the sheer absurdity of the idea that this pattern is meaningful at all is easy to see if we think about how the data was organized in the first place.

There is no obvious order for counting tomato plants: Even if the plants are in a row, you could start counting at either end. Similarly, once a "setup" is made, it can be pretty hard to decide which doll bed came first and which one last: Like everything else in the setup, beds are moved around, changed, taken apart, and rebuilt as play goes on and the story develops and changes.

To see why this matters, we can rearrange the order of the doll beds. Instead of the heights being 8, 4, 6, 7, 3, 2, 4, 4, 1, and 9, we could order them 4, 8, 4, 3, 6, 4, 1, 7, 2, and 9. Same numbers (one each of the numbers 1, 2, 3, 6, 7, 8, and 9, and three 4s) but in a different order. Now the data looks like this:

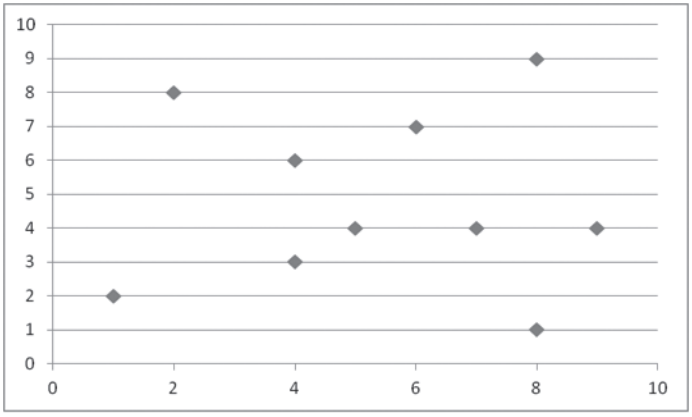

Figure 1.4 **The relationship of the two sets numbers used in Figure 2, but with the numbers reordered.**

With this new ordering, there is no easily discernible pattern. Indeed, the tomato plants and doll beds now have almost no correlation at all. Pearson's r = .01, and as you may recall Pearson's r is always a value between -1 and 1, with a value of 0 indicating no relationship at all. On that scale, r = .01 is about as close to no pattern as you are likely to get.

This admittedly cartoonish example of data mining illustrates the critically important principle of GIGO or *garbage in, garbage out*. The term GIGO goes back to the early days of computing. It first appeared in a newspaper article in 1963 about the Internal Revenue Service of all things. GIGO describes what happens when someone enters bad data into a computer program: If you type in something that does not make sense, the answer you get back from the computer will not make sense either.[16]

Now obviously real data mining techniques are quite a bit more sophisticated than this example. And of course real data mining uses collections of data with many more than 10 points of data in each of two sets. But unfortunately the tomato/doll bed problem is not as far removed from real data mining as you might think. Don't believe me? Here's what Chris Anderson, the editor-in-chief of Wired magazine wrote in a famous 2008 column:

> Petabytes [of data] allow us to say: "Correlation is enough"…. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.[17]

In other words: Throw the doll beds and tomatoes and cost of gas and people's ages and income and weight and height all into one big hopper with anything else you have and look at the patterns.

The ones that say "Ouch!" are alive.

This example, in other words, is very much at the heart of the use—and abuse—of statistics for data mining: If you put nonsense in, you get nonsense out. And this is even before we deal with the problem of where the data comes from in the first place, and how it gets collected. Why was I counting doll beds and tomatoes to begin with? Whose height and weight are being measured, by whom, when, and for what purpose?

If these seem like silly objections, consider that Lawrence Kohlberg, who more or less founded the field of moral psychology with his work on stages of moral development, based his framework on studies with male participants. It was not until 20 years after Kohlberg's original work that psychologist Carol Gilligan pointed out the problem: Kohlberg's tools for measuring moral development, created from studies of boys, led to the conclusion that girls, on average, reach a lower level of moral development. This research design is a little like giving students who speak French a test written in Japanese and concluding that French speakers are, on average, less articulate than Japanese speakers.[18]

And if all that is not enough, Anderson is actually quite wrong about what happens when we have more and more data, and more and more powerful computers. The GIGO problem actually gets *much worse* when we use statistics on Big Data. For reasons we'll discuss later, Little Data is not the same as Big Data from a statistical point of view. Analyzing 100 or 1,000 pieces of data presents

different issues and problems than analyzing 100,000 or 100 million pieces of data—and not just because it takes longer. Sadly, not everyone who uses data mining understands the pitfalls that Big Data presents.

There is an even more important lesson in the tomato/doll bed data, though. In a very real sense, these correlations are not any more or less simplistic than the data we saw earlier about Cassie's road trip from Texas to Florida. Both tiny, invented examples illustrate how we analyze much more complex sets of data. Both were facts that could have been plucked from the stream of data that are recorded about us in our *Star Trek*-like world of ever-present computers.

But they seem different.

The difference is not about whether one example involved numbers and statistics and the other did not. Our analysis of Cassie's data was about time and spatial location—not to mention the cost of gas—which are just as much about numbers as counting Lego bricks in a doll's bed. We did not use statistics in Cassie's simple example, but we could have. The difference is not that one analysis involved computers and the other did not. I used a computer in different ways, but I used a computer in both cases.

No, the difference between Cassie and the Doll Beds (which would make a great name for a punk band, by the way) is that one of these analyses made sense and the other did not. One turned data into information; the other turned data into confusion. That difference—between Cassie's road trip and the meaningless correlation of tomato plants and doll beds, between finding a story and finding a pattern—is what this book is about. That is also, by the way, what data mining *should* be about.

The goal of this book is to lay out the conceptual and practical tools that distinguish Cassie from the Doll Beds: to create a set of techniques and concepts for sensibly analyzing Big Data. My approach to this challenge is to bring together the methods of ethnography and the tools of statistics to form a science of Quantitative Ethnography, and show how to use it to find meaningful insights from the hills and valleys of data through which we travel every day.

Understanding people

In some ways it seems quite natural that we use quantitative tools—that is, statistics—to analyze Big Data. Computers collect Big Data, and computers are digital tools, which means *by definition* anything that can be stored in a computer

can be represented by numbers. That basic fact seems obvious if we look at information like credit card numbers, or Social Security Numbers, or the $13.73 Cassie spent on unleaded gasoline. It makes sense that the date Cassie bought the gas is stored as a number. Even Cassie's name is actually recorded as a number: Each letter in the alphabet is stored as a number in the computer's memory.

But if Cassie takes a picture on her iPhone and sends it to a friend on Snapchat, the picture is stored as a string of numbers as well, at least until the picture is automatically deleted. Each pixel—each point in the picture—has a color value associated with it, and the computer represents each color value with a number from 0 to 16,777,216. So a 4x6 inch, high-resolution picture (say, 600 pixels per inch) is really just 8.6 million numbers. Actually that is not quite true. Many of the pixels in a picture will have the same color value, so to save space a computer will often record the color value once and then keep track of how many pixels in a row have the same value. But the basic point remains the same: names, addresses, prices, locations, images, music, videos, and even captain's log entries: Everything in the world of Big Data is recorded as numbers, and thus can be analyzed mathematically.[19]

So the "Quantitative" part of "Quantitative Ethnography" makes sense. But why "Ethnography"? A science that originated in studies of the traditional culture of African tribes and Pacific Islands may seem like an odd choice for making sense of our hyper-modern digital society. But in fact it is precisely because ethnography is the study of culture that it is a critical part of a sensible analysis of Big Data.

The term culture has many meanings, of course. Some definitions, like *a collection of bacteria grown in a Petri dish*, are obviously not relevant here. But neither is a definition that limits "culture" to a description of arts and literature: the songs, stories, and images—highbrow and low—that a group of people tell about themselves. Those things are *part* of a culture because they are part of how people understand the meaning of things that happen in their lives. But culture is about much more than just the arts.[20]

Culture matters because while computers can mine in a mountain of data, human beings swim in a sea of significance. We traffic in symbols: in action, in talk, in writing, and in making things that *mean something* to ourselves and to others. The things people say and make and do are interpreted by others who share their culture, and ethnography is the science of understanding those inter-

pretations. Culture is how people understand the meaning of things—and not just the meaning of things themselves, but the web of meanings that connect things to each other, and things to people, and thus ultimately connect us to one another.

Culture is what makes data into information by adding meaning, and thus the "Ethnography" part of Quantitative Ethnography is as important as the "Quantitative" part in moving from Big Data to Big Information. We need a method for analyzing meaning to make sensible analyses of Big Data if we want to shed light on what people do and why. Put another way, the mountains of data that we see around us are more like a chain of islands in the middle of the cultural ocean of meaning. Ethnography is a way to chart that ocean, and Quantitative Ethnography is a way to use statistical tools to make better charts by finding landmarks amidst the mountains of data.

To do anything less—to pretend that the mountains of data do not exist in sea of cultural significance—may be mathematically rigorous, but in the end is conceptually empty. Or as Clifford Geertz, one of the best-known ethnographers of the last century, said succinctly: "Nor … have I been impressed with claims that structural linguistics, computer engineering, or some other advanced form of thought is going to enable us to understand men without knowing them."[21]

This is, of course, just a more rigorous way of saying that throwing shrimp at a wall is not a very good way to understand shrimp (or anything else, for that matter) no matter how big the wall is and how many shrimp you are able to throw. Understanding something—particularly something as complex and interconnected as human beings and the cultures in which they live—requires more sophisticated analysis than throwing things at a wall, or throwing data blindly into a statistical model. The data has to have meaning, which can only come, ultimately, from knowing something about the people and situations being analyzed.

Now obviously ethnography is not the only way to understand people by knowing them. But ethnography is a tool particularly well-honed for this challenge since its principal focus is the interpretation of cultural material. Ethnography is the science of understanding how a system of symbols works—and more particularly how to find the meaning that people attach to things they say and do and make. But there are other methods for making sense of human endeavor: humanities like history, and literary analysis; other social sciences like sociology and psychology; and many other forms of meaning-based (the technical term is *qualitative*) data analysis. As a result, many of the principles and ideas that

are the foundation of Quantitative Ethnography will apply equally well to other approaches for making sensible interpretations of Big Data.

One reason for starting with ethnography here is that when we combine ethnographic and statistical tools to analyze Big Data, we also get a larger set of tools for making sense of smaller data: the kind of data ethnographers, and historians, and journalists, and a host of other scholars use to study all manner of art, literature, and social interaction. Making meaning of Big Data gives us insight into how to use statistics to understand cultural material of all kinds, and the techniques of Quantitative Ethnography work quite well for Quantitative History, or Quantitative Journalism, or Quantitative Literary Analysis.

But one of the most basic tenets of good ethnography is that it works best to move from the specific to the general rather than the other way around. For me—and hence for us here—those specifics are in the practices of ethnography, which is the form of cultural interpretation that I know best.

In the end, though, this book is not about ethnography, any more than it is about statistics. We will look at some important statistical issues and discuss some of the important principles of ethnography—and of interpretive or qualitative research in more generally. But fundamentally this book is about how to use ethnographic techniques to guide statistical analyses of Big Data. At the same time, it explores how to use statistical techniques to increase the scope and power of ethnographic and other qualitative methods of research.

This is a book about understanding why, in the digital age, the old distinctions between qualitative and quantitative research methods, between the sciences and humanities, and between numbers and understanding, limit the kinds of questions we can ask, in some cases, and lead us accept superficial answers in others. Quantitative Ethnography is a research method that goes beyond those distinctions to help us understand how to make sense of our increasingly data-rich world.

Going forward

The remainder of the book, then, fleshes out the key concepts, tools, and methods of Quantitative Ethnography in more or less in three parts.

Before we can talk about integrating qualitative and quantitative approaches to research, we first have to look at the fundamental logic by which each method

operates. We cannot build a common language without understanding something of the individual ways of talking that it will bridge.

The first part of what follows, then—chapters 2, 3, and 4—looks at the foundations of qualitative and quantitative methodologies. Chapter 2 lays out the basic concerns of ethnography: the issues that ethnographers think about, the kinds of data they use, the concerns they raise. Chapter 2 considers problems of bias and subjectivity, and how ethnographers frame those challenges in their work. Chapter 3 looks at the mechanics of ethnography: what ethnographers do, and the reasoning behind those practices. Chapter 3 covers concepts of thick description and coding, and how ethnographers develop, structure, and defend arguments about what people do and why. Finally, Chapter 4 takes on the same two tasks for quantitative methods. Chapter 4 looks at the concept of generalization—how researchers use statistics to make claims about similarities and differences between groups—and focuses on the logic of sampling and statistical significance.

I have made this point already, and will make it many times in the pages to follow, but the goal of these chapters is, of course, not to pretend that one part of one book could cover all of the key ideas in two large, diverse, and complex fields. Rather, the goal is to lay out the basic frameworks of these two different research methods so we can begin the task of connecting them in Quantitative Ethnography.

The second part of the book—chapters 5, 6, and 7—describes the key theories and practices that link quantitative and qualitative methods together. Chapter 5 looks at how to organize qualitative data so that it can be analyzed using statistical tools. This question is pragmatic (How should the data be arranged in a file?) but more important, it is conceptual: What are the underlying structures of human interaction that we can use to organize the data? Chapters 6 and 7 take the basic quantitative process of constructing quantitative models and show how it can be applied to data that is organized using the approach described in chapter 5. Again, this question is both practical (What are the components of a model?) and philosophical. The focus of chapter 6 is on the logic of modeling from a quantitative perspective. Chapter 7 looks at using that logic in an ethnographic context.

The final part of the book looks at what it means to use statistical models to understand ethnographic data—and therefore to model how people make sense of the world. Again the approach is to consider both the theoretical implications of using statistics this way and also the mechanical details of how to actually do it. Chapter 8 focuses on how researchers identify what people mean as they talk and act, including questions of automated coding and reliability. Chapter 9 looks at the structure of meaning-making, and how we can model the way people express their understanding about the world. The 10th and final chapter ties these strands together, and places Quantitative Ethnography in the context of other approaches to the analysis of Big Data.

Or at least that is what these chapters try to do. This book is very much written as an introduction to thinking about research in an age where the sheer volume of data strains the capabilities of methods developed when life was less thoroughly recorded.

It is intended to be the first words for a reader interested in Quantitative Ethnography, not the last word on the field.

In fact, the book you have in your hands—or, more likely, on the screen—comes from two main sources. The first is work done in my own research lab and in other labs by a number of very capable students and scholars over many years. I have mentioned some of these people specifically in the pages that follow, but there are many others who have contributed to my own understanding of Quantitative Ethnography in ways big and small. The particular words here—and all of the mistakes they contain—are my own. But I would be remiss if I did not point out that the ideas were, and continue to be, very much a collaborative effort.

The other source for the material in this book is a course I have taught for several years. My goal in writing this book has, in part, been to set out the fundamental concepts and practices of Quantitative Ethnography for those approaching the subject for the first time.

To help readers who are near the beginning of their research training—and perhaps others who may be helping train new researchers—I have included at the end of each chapter some suggestions for further reading and activities that may be helpful in seeing how to put concepts from the page into practice. These are, more or less, the same readings and activities that I use when teaching Quantitative Ethnography, but obviously these short sections can be skipped without losing the key points of each chapter.

Finally, more technical papers about the ideas here are available online. In these pages, I have tried to write in a way that will be enjoyable and easy to read, while still being authoritative and accurate. My hope is that the result is an overview of Quantitative Ethnography that is accessible—and ideally even inspiring—to readers from many backgrounds.

Chapter 1. Introduction: Captain's Log

1. *Cavorite*, named for the fictional Dr. Cavor, is a metal that acted as a gravity shield, such that objects encased within it were freed from the influence of gravity.
2. For more on the influence of *Star Trek* on the mobile phone industry, see "Brain scan" (2009), "How Star Trek inspired an innovation" (2012), Handel Productions (2009), and Choney (2009). For information on James Bond's technological contributions, see Dyce (2012).
3. For a list of *Star Trek's* contributions, see Farrington (2009) and Handel Productions (2009). The place of the *Enterprise* in aviation history is discussed in Catlin (2015). The term "cyberspace" appeared first in Gibson (1982).
4. Linguistic purists will note that this is the first use I have made of the word data, which technically is plural: Data is a collection of pieces of data, each of which is technically called a datum. Thus it would be grammatically correct to write "data were" rather than "data was." However, common usage refers to data as a collective noun rather than a plural—the team likes to speak in American English rather than the team like to speak in British English. So here and throughout I use singular verbs with data, not because I am ignorant of the difference, but because it sounds less affected to my ear as a writer. If you prefer "data were," please feel free to set your universal translator accordingly.
5. For more on the history of the Brownie camera, see Olivier (2007).
6. Schmidt's estimate is described in Siegler (2010). The prevalence of cell phones is described in "Mobile phone access reaches three quarters of planet's population" (2012). For benchmarks on information storage, see "How much is 1 byte, kilobyte, megabyte, gigabyte, etc.?" (n.d.). More on social media statistics can be found in Newcomb (2016).
7. For more on Facebook manipulation, see Popkin (2014).
8. The Twitter ransom is described in Gayomali (2014).
9. The idea that information is the combination of data and meaning comes from (Devlin, 1995).
10. For more on the relationship of height and weight, see Eknoyan (2008).
11. In making the comparison to Juster's Half Boy, I do not mean to trivialize Quetelet's contribution to science, which is in some ways far greater—and far more problematic—than most people realize. Based on the work of Quetelet and his contemporaries, medicine transformed over the course of the 19th century from a practice based on what is *natural* (patients are evaluated against their prior selves) to what is *normal* (patients are evaluated against the distribution of patients). For more

on Quetelet's work, see Quetelet (1835), Hacking (1990), Hankins (1908), Tanner (1981). The Half Boy is from Juster (1961).

12. For more on statistics and eugenics, see Hacking (1990), Norton (1978), Tanner (1981).
13. Technically a correlation of r = 1 means that every time the height goes up or down by some amount—say, 1 inch—then weight goes up (or down) by some fixed amount—say, 1 pound. If height goes up 2 inches, weight goes up 2 pounds. If height goes down 5 inches, weight goes down 5 pounds. This means when the two values are graphed the relationship is a straight line.
14. In the sample referenced is from Heinz, Peterson, Johnson, and Kerk (2003). In their sample, correlation between height and weight is r = 0.72, but the sample consisted of "physically active individuals." The overall correlation is probably closer to r = 0.50, and for Olympic athletes it may be as high as r = 0.77. See, for example, "SOCR Data Dinov 020108 HeightsWeights" (1993) and "Your Olympic athlete body match" (2012). Regardless, height alone is not actually a good predictor of weight. We can, however, get correlations very close to r = 1 if we include other body measurements.
15. On the relationship between money and happiness, see Matthews (2013).
16. The origins of GIGO are described in "World Wide Words: Garbage in, garbage out" (n.d.).
17. The quotation is from Anderson (2008).
18. Gilligan's landmark work can be found in Gilligan (1982).
19. The form of image compression described here, sometimes known as *duplicate string elimination*, is only one of many techniques computer scientists have developed for reducing the storage requirements for data files. The subject is quite beyond the scope of this book, other than to point out that the amount of information a person can get from a data file is not always directly proportional to the size of the file, for the obvious reason that humans and computers process information differently. For those interested in the mathematics, a 4x6 inch image at 600dpi has 600 x 600 = 360,000 pixels per inch. So the 24 square inches of the photo have 8,640,000 pixels, each of which is represented by a number.
20. This point is made in Geertz (1973c) p. 30.
21. The quotation is from Geertz (1973c) p. 30.