Understanding Data

When you ask people what data is, most reply with a vague description of something that resembles a spreadsheet or a bucket of numbers. The more technically savvy might mention databases or warehouses. However, this is just the format that the data comes in and how it is stored, and it doesn't say anything about what data is or what any particular dataset represents. It's an easy trap to fall in because when you ask for data, you usually get a computer file, and it's hard to think of computer output as anything but just that. Look beyond the file though, and you get something more meaningful.

WHAT DATA REPRESENTS

Data is more than numbers, and to visualize it, you must know what it represents. Data represents real life. It's a snapshot of the world in the same way that a photograph captures a small moment in time.

Look at Figure 1-1. If you were to come across this photo, isolated from everything else, and I told you nothing about it, you wouldn't get much out of it. It's just another wedding photo. For me though, it's a happy moment during one of the best days of my life. That's my wife on the left, all dolled up, and me on the right, wearing something other than jeans and a T-shirt for a change. The

FIGURE 1-1 *A single photo, a single data point*

pastor who is marrying us is my wife's uncle, who added a personal touch to the ceremony, and the guy in the back is a family friend who took it upon himself to record as much as possible, even though we hired a photographer. The flowers and archway came from a local florist about an hour away from the venue, and the wedding took place during early summer in Los Angeles, California.

That's a lot of information from just one picture, and it works the same with data. (For some, me included, pictures are data, too.) A single data point can have a who, what, when, where, and why attached to it, so it's easy for a digit to become more than a toss in a bucket. Extracting information from a data point isn't as easy as looking at a photo, though. You can guess what's going on in the photo, but when you make assumptions about data, such as how accurate it is or how it relates to its surroundings, you can end up with a skewed view of what your data actually represents. You need to look at everything around, find context, and see what your dataset looks like as a whole. When you see the full picture, it's much easier to make better judgments about individual points.

Imagine that I didn't tell you those things about my wedding photo. How could you find out more? What if you could see pictures that were taken before and after?

FIGURE 1-2 *Grid of photos*

Now you have more than just a moment in time. You have several moments, and together they represent the part of the wedding when my wife first walked out, the vows, and the tea drinking ceremony with the parents and my grandma, which is customary for Chinese weddings. Like the first photo, each of these has its own story, such as my father-in-law welling up as he gave away his daughter or how happy I felt when I walked down the aisle with my bride. Many of the photos captured moments that I didn't see from my point of view during the wedding, so I almost feel like an outsider looking in, which is probably how you feel. But the more I tell you about that day, the less obscure each point becomes.

Still though, these are snapshots, and you don't know what happened in between each photo. (Although you could guess.) For the complete story, you'd either need to be there or watch a video. Even with that, you'd still see only the ceremony from a certain number of angles because it's often not feasible to record every single thing. For example, there was about five minutes of confusion during the ceremony when we tried to light a candle but the wind kept blowing it out. We eventually ran out of matches, and the wedding planner went on a scramble to find something, but luckily one of our guests was a smoker, so he busted out his lighter. This set of photos doesn't capture that, though, because again, it's an abstraction of the real thing.

This is where sampling comes in. It's often not possible to count or record everything because of cost or lack of manpower (or both), so you take bits and pieces, and then you look for patterns and connections to make an educated guess about what your data represents. The data is a simplification—an abstraction—of the real world. So when you visualize data, you visualize an abstraction of the world, or at least some tiny facet of it. Visualization is an abstraction of data, so in the end, you end up with an abstraction of an abstraction, which creates an interesting challenge.

However, this is not to say that visualization obscures your view—far from it. Visualization can help detach your focus from the individual data points and explore them from a different angle—to see the forest for the trees, so to speak. To keep running with this wedding photo example, Figure 1-3 uses the full wedding dataset, of which Figure 1-1 and Figure 1-2 were subsets of. Each rectangle represents a photo from our wedding album, and they are colored by the most common shade in each photo and organized by time.

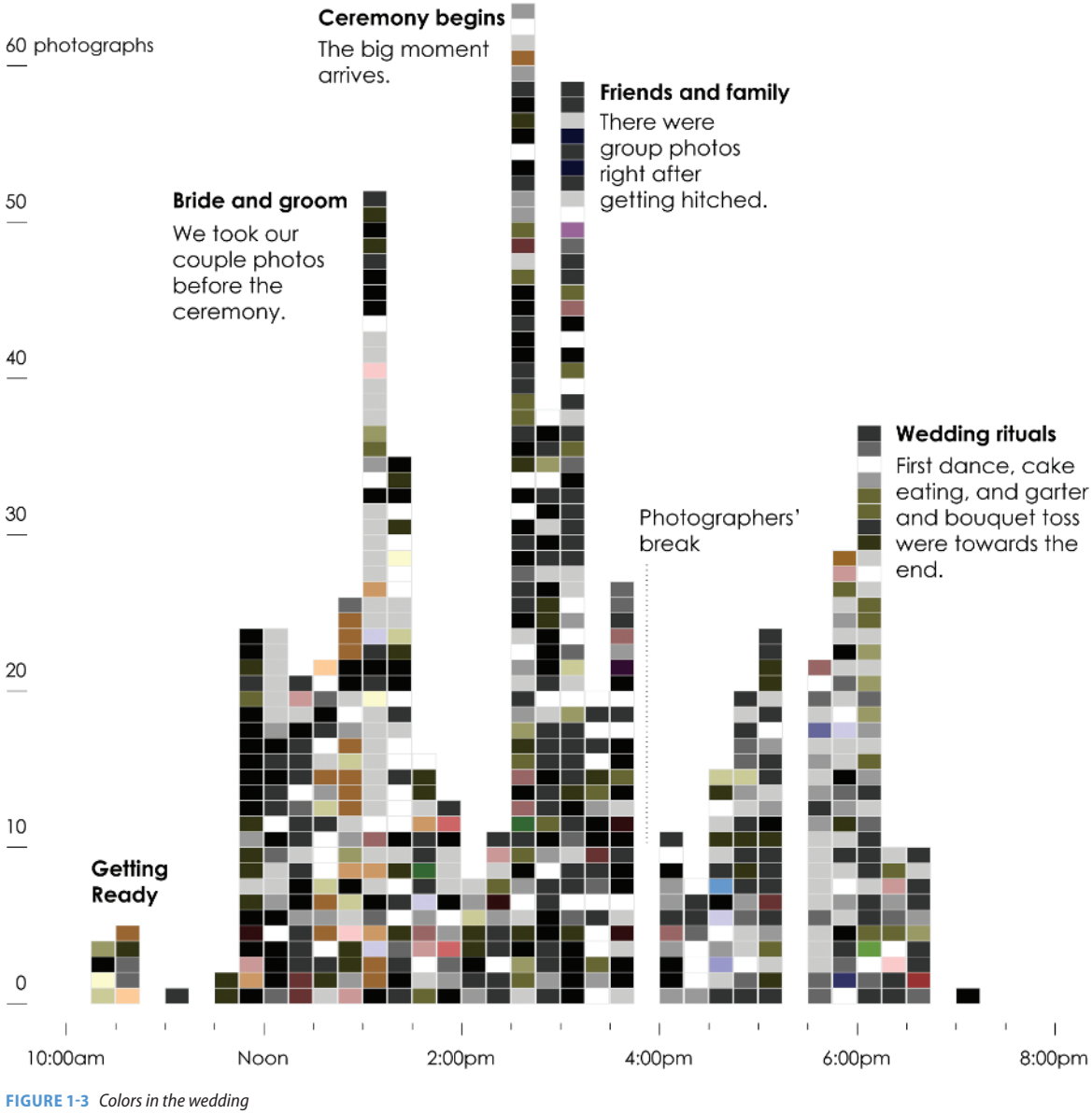


FIGURE 1-3 *Colors in the wedding*

With a time series layout, you can see the high points of the wedding, when our photographers snapped more shots, and the lulls, when only a few photos were taken. The peaks in the chart, of course, occur when there is something to take pictures of, such as when I first saw my wife in her dress or when the ceremony began. After the ceremony, we took the usual group photos with friends and family, so there was another spike at that point. Then there was food, and activity died down, especially when the photographers took a break a little before 4 o'clock. Things picked up again with typical wedding fanfare, and the day came to an end around 7 in the evening. My wife and I rode off into the sunset.

In the grid layout, you might not see this pattern because of the linear presentation. Everything seems to happen with equal spacing, when actually most pictures were taken during the exciting parts. You also get a sense of the colors in the wedding at a glance: black for the suits, white for the wedding dress, coral for the flowers and bridesmaids, and green for the trees surrounding the outdoor wedding and reception. Do you get the detail that you would from the actual photos? No. But sometimes that level isn't necessary at first. Sometimes you need to see the overall patterns before you zoom in on the details. Sometimes, you don't know that a single data point is worth a look until you see everything else and how it relates to the population.

You don't need to stop here, though. Zoom out another level to focus only on the picture-taking volumes, and disregard the colors and individual photos, as shown in Figure 1-4.

You've probably seen this layout before. It's a bar chart that shows the same highs and lows as in Figure 1-3, but it has a different feel and provides a different message. The simple bar chart emphasizes picture-taking volumes over time via 15-minute windows, whereas Figure 1-3 still carries some of the photo album's sentiment.

The main thing to note is that all four of these views show the same data, or rather, they all represent my wedding day. Each graphic just represents the day differently, focusing on various facets of the wedding. Interpretation of the data changes based on the visual form it takes on. With traditional data, you typically examine and explore from the bar chart side of the spectrum, but that doesn't mean you have to lose the sentiment of the individual data point—that single photo. Sometimes that means adding meaningful annotation that enables readers to interpret the data better, and other times the message in the numbers is clear, gleaned from the visualization itself.

Photographs over time

Our wedding photographers snapped more pictures during the significant events with a peak of 63 during a 15-minute span.

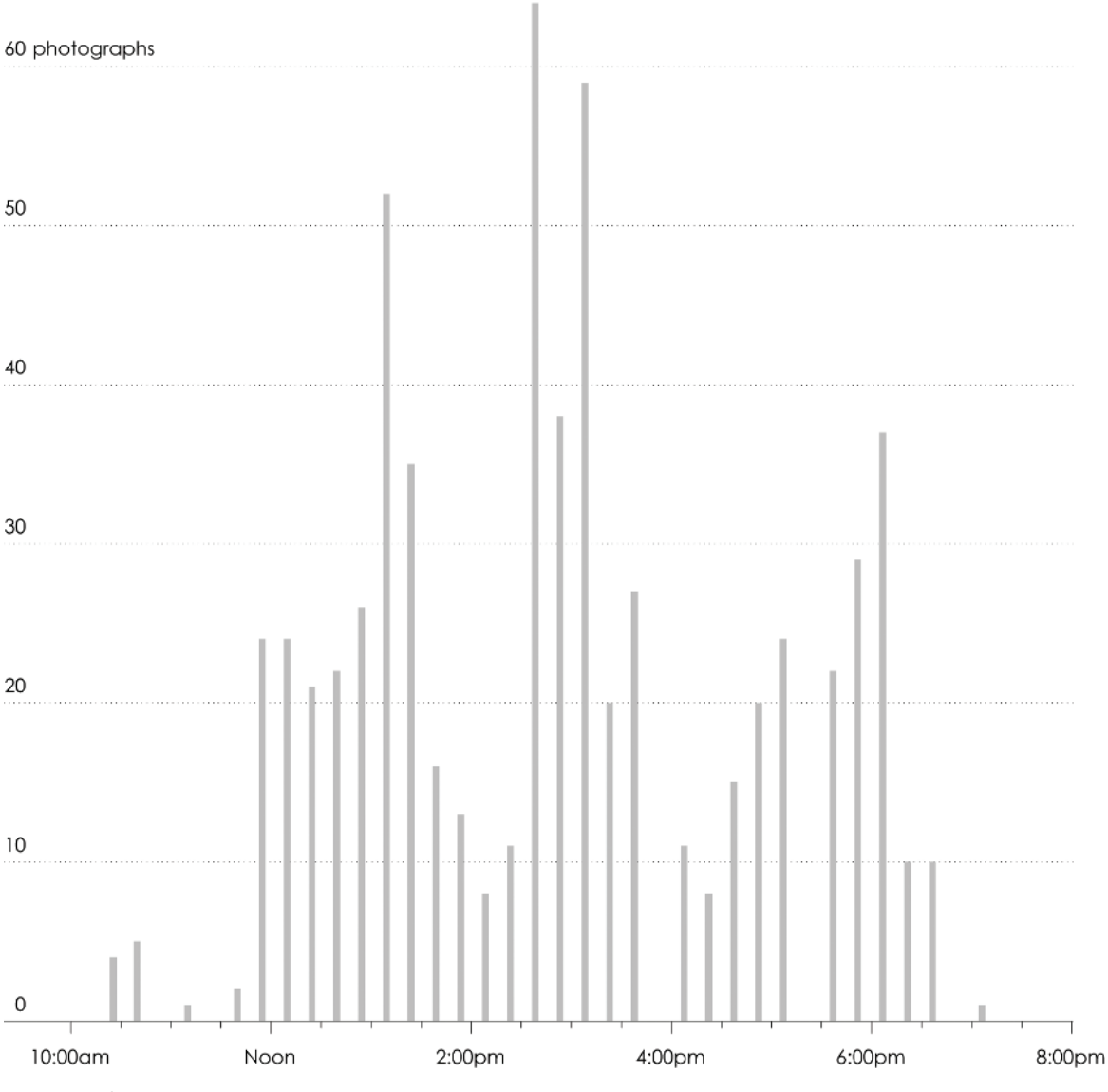


FIGURE 1-4 *Photos over time*

The connection between data and what it represents is key to visualization that means something. It is key to thoughtful data analysis. It is key to a deeper understanding of your data. Computers do a bulk of the work to turn numbers into shapes and colors, but you must make the connection between data and real life, so that you or the people you make graphics for extract something of value.

This connection is sometimes hard to see when you look at data on a large scale for thousands of strangers, but it's more obvious when you look at data for an individual. You can almost relate to that person, even if you've never met him or her. For example, Portland-based developer Aaron Parecki used his phone to collect 2.5 million GPS points over 3½ years between 2008 and 2012, about one point every 2 to 6 seconds. Figure 1-5 is a map of these points, colored by year.

As you'd expect, the map shows a grid of roads and areas where Parecki frequented that are colored more brightly than others. His housing changed a few times, and you can see his travel patterns change over the years. Between 2008 and 2010, shown in blue, travel appears more dispersed, and by 2012, in yellow, Parecki seems to stay in a couple of tighter pockets. Without more context it is hard to say anything more because all you see is location, but to Parecki the data is more personal (like the single wedding photo is to me). It's the footprint of more than 3 years in a city, and because he has access to the raw logs, which have time attached to them, he could also make better decisions based on data, like when he should leave for work.

What if there were more information attached to personal time and location data, though? What if along with where you were, you also took notes during or after about what was going on at some given time? This is what artist Tim Clark did between 2010 and 2011 for his project *Atlas of the Habitual*. Like Parecki, Clark recorded his location for 200 days with a GPS-enabled device, which spanned approximately 2,000 miles in Bennington, Vermont. Clark then looked back on his location data and labeled specific trips, people he spent time with, and broke it down by time of year.

As shown in Figure 1-6, the atlas, with clickable categorizations and time frames, shows a 200-day footprint that reads like a personal journal. Select "Running errands" and the note reads, "Doing the everyday things from running to the grocery store all the way to driving 30 miles to the only bike shop in southern Vermont opened on Sundays." The traces stay around town, with the exception of two long ones that venture out.

There is one entry titled "Reliving the breakup," and Clark writes, "A long-term girlfriend and I broke up immediately before I moved. These are the times that I had a real difficult time coming to terms that I had to move on." Two small paths, one within the city limits and one outside, appear, and the data suddenly feels incredibly personal.

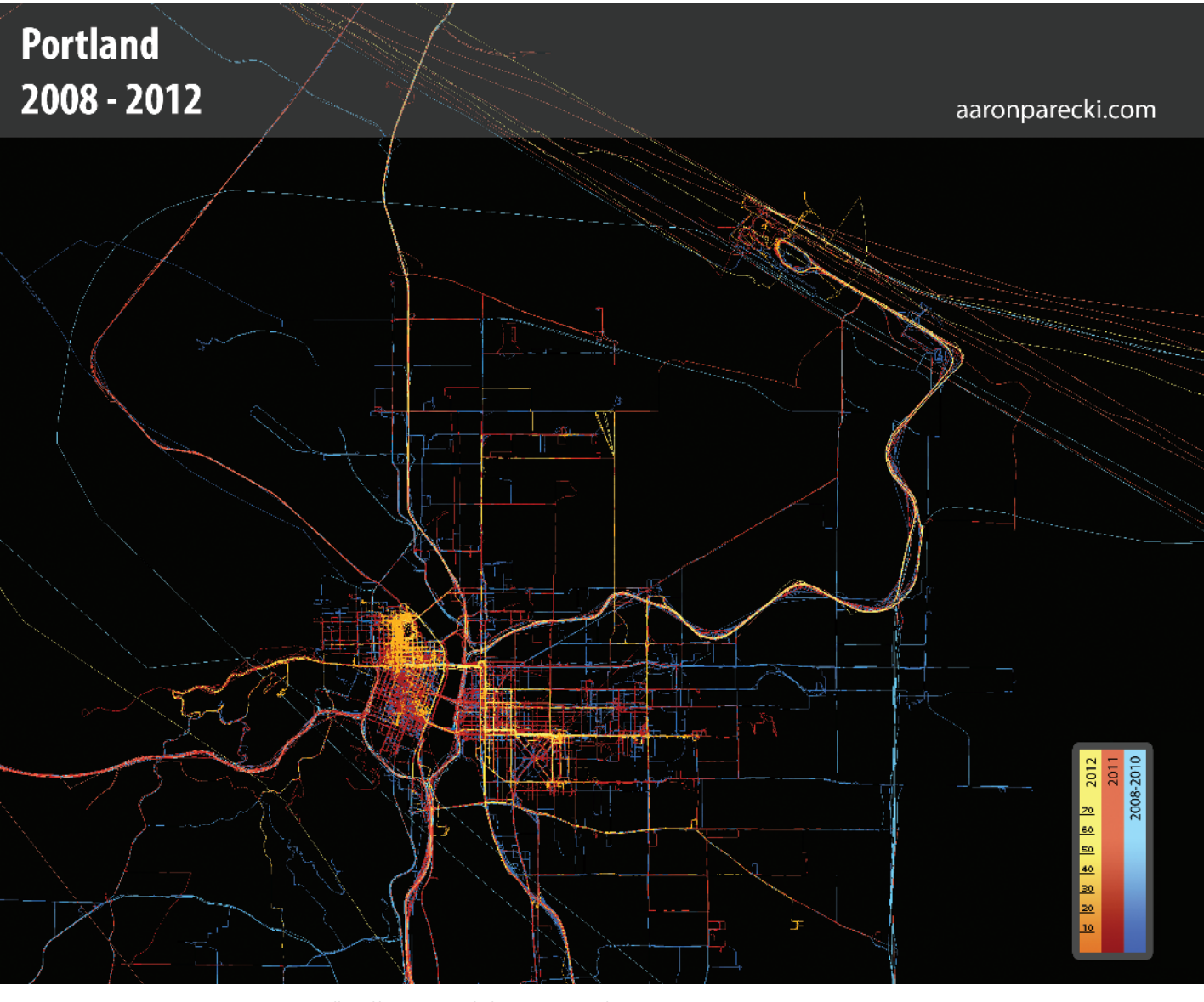


FIGURE 1-5 *GPS traces collected by Aaron Parecki, http://aaronparecki.com*

This is perhaps the appeal behind the Quantified Self movement, which aims to incorporate technology to collect data about one's own activity and habits. Some people track their weight, what they eat and drink, and when they go to bed; their goal is usually to live healthier and longer. Others track a wider variety of metrics purely as a way to look in on themselves beyond what they see in the mirror; personal data collection becomes something like a journal for self-reflection at the end of the day.

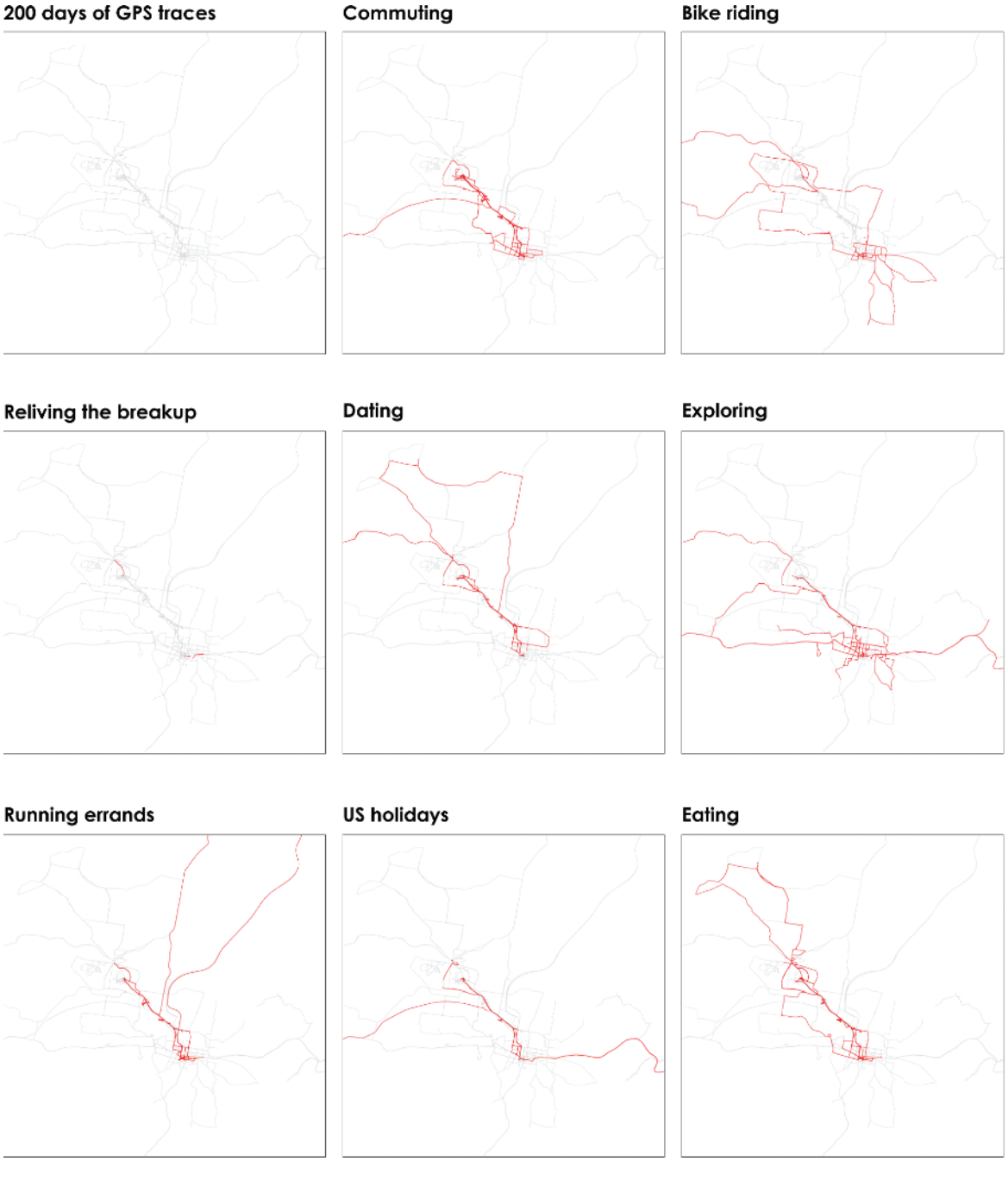


FIGURE 1-6 *Selected maps from* Atlas of the Habitual *by Tim Clark, http://www.tlclark.com/atlasofthehabitual/*

Nicholas Felton is one of the more well-known people in this area for his annual reports on himself, which highlight both his design skills and disciplined personal data collection. He keeps track of not just his location, but also who he spends time with, restaurants he eats at, movies he watches, books he reads, and an array of other things that he reveals each year. Figure 1-7 is a page out of Felton's 2010/2011 report.

Felton designed his first annual report in 2005 and has done one every year since. Individually, they are beautiful to look at and hold and satisfy an odd craving for looking in on a stranger's life. What I find most interesting, though, is the evolution of his reports into something personal and the expanding richness of data. Looking at his first report, as shown in Figure 1-8, you notice that it feels a lot like a design exercise in which there are touches of Felton's personality embedded, but it is for the most part strictly about the numbers. Each year though, the data feels less like a report and more like a diary.

This is most obvious in the *2010 Annual Report*. Felton's father passed away at the age of 81. Instead of summarizing his own year, Felton designed an annual report, as shown in Figure 1-9, that cataloged his father's life, based on calendars, slides, postcards, and other personal items. Again, although the person of focus might be a stranger, it's easy to find sentiment in the numbers.

When you see work like this, it's easy to understand the value of personal data to an individual, and maybe, just maybe, it's not so crazy to collect tidbits about yourself. The data might not be useful to you right away, but it could be a decade from now, in the same way it's useful to stumble upon an old diary from when you were just a young one. There's value in remembering. In many ways you log bits of your life already if you use social sites like Twitter, Facebook, and foursquare. A status update or a tweet is like a mini-snapshot of what you're doing at any given moment; a shared photo with a timestamp can mean a lot decades from now; and a check-in firmly places your digital bits in the physical world.

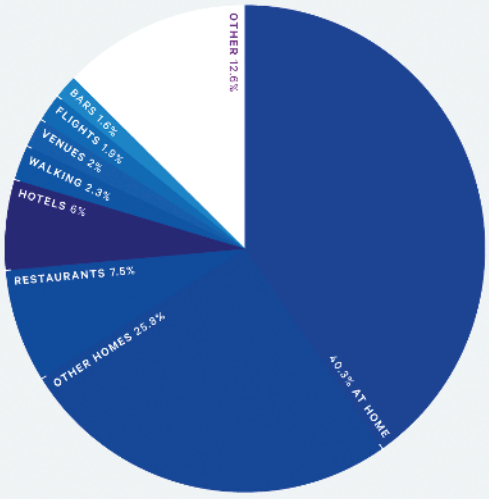
You've seen how that data can be valuable to an individual. What if you look at the data from many individuals in aggregate?

The United States Census Bureau collects the official counts of people living in the country every 10 years. The data is a valuable resource to help officials allocate funds, and from census to census, the fluctuations in population help you see how people move in the country, changing the neighborhood

FIGURE 1-7 *(following page) A page from* 2010/2011 Annual Report *by Nicholas Felton, http://feltron.com*

With Olga

EVERYWHERE



DAYS TOGETHER

191¼

315 different encounters

MOST TIME SPENT TOGETHER

- **MANHATTAN** — 83¾ DAYS
- **BROOKLYN** — 51½ DAYS
- **MILL VALLEY** — 9 DAYS
- **ANCHORAGE** — 7¾ DAYS
- **SYDNEY** — 4¼ DAYS

MOST VISITED PLACE TOGETHER

Old Apartment

194 visits

DIFFERENT CITIES VISITED TOGETHER

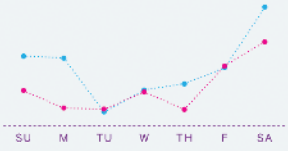
56

In 3 countries, 9 states and Washington D.C.

FAVORITE BEVERAGES WITH OLGA

- **FILTER COFFEE** — 111 SERVINGS
- **RED WINE** — 78 SERVINGS
- **DALE'S PALE ALE** — 35 SERVINGS
- **CHAMPAGNE** — 30 SERVINGS
- **LATTE** — 26 SERVINGS

TIME TOGETHER



BRIEFEST MONTH TOGETHER

June 2011

40¾ hours

MOST CONSECUTIVE HOURS TOGETHER

247

Australia trip — February 2010

TIME SPENT WITH OLGA AND...

- **SARAH** — 6¼ DAYS
- **MOM** — 6¼ DAYS
- **BRIAN** — 5¾ DAYS
- **OLGA'S MOM** — 5 DAYS
- **RYAN** — 4½ DAYS

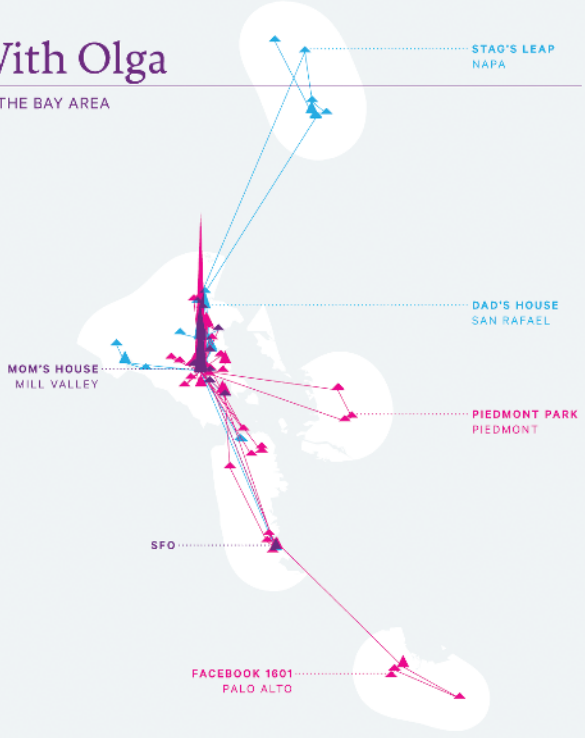
WEDDINGS ATTENDED TOGETHER

Seven

Aaron & Jessica, Charlie & Bret, Glenn & Mariana, Lewis & Ange, Randy & Allison, Rob & Elise and Toby & Harriet

With Olga

IN THE BAY AREA



DAYS TOGETHER IN THE BAY AREA

13½

Approximately 7% of total time together

BAY AREA PLACES VISITED TOGETHER

77

18 stores, 13 restaurants, 10 homes, 6 outdoor places, 3 coffee shops, 3 grocery stores, 2 airport terminals, 2 bars, 2 gas stations, 2 hospitals, 2 hotels, 2 liquor stores, 2 parking garages, 2 parking lots, a cinema, a deli, a drug store, a laundromat, a library, a museum, a park and work

FAVORITE BAY AREA BOTTLESHOP

Vintage Wine & Spirits

Visited twice

FAVORITE BAY AREA BEER WITH OLGA

Lagunitas IPA

5 servings

BAY AREA MUSEUMS VISITED TOGETHER

The Exploratorium

With Marina — July 9, 2011

MOST PLAYED ARTIST TOGETHER

The Beach Boys

25 songs listened to from *Christmas with the Beach Boys*

TIME TOGETHER IN THE BAY AREA

MOST FREQUENTED CITY TOGETHER

Mill Valley

68% of time in the Bay Area

MOST VISITED BAY AREA PLACES

- **MOM'S HOUSE** — 35 VISITS
- **MARIN GENERAL HOSPITAL** — 6 VISITS
- **CHEVRON MILL VALLEY** — 5 VISITS
- **SFO INTERNATIONAL TERMINAL** — 4 VISITS
- **DAD'S HOUSE** — 3 VISITS

CRISES INVOLVING A TICK

One

Spotted by Olga, removed by Mom

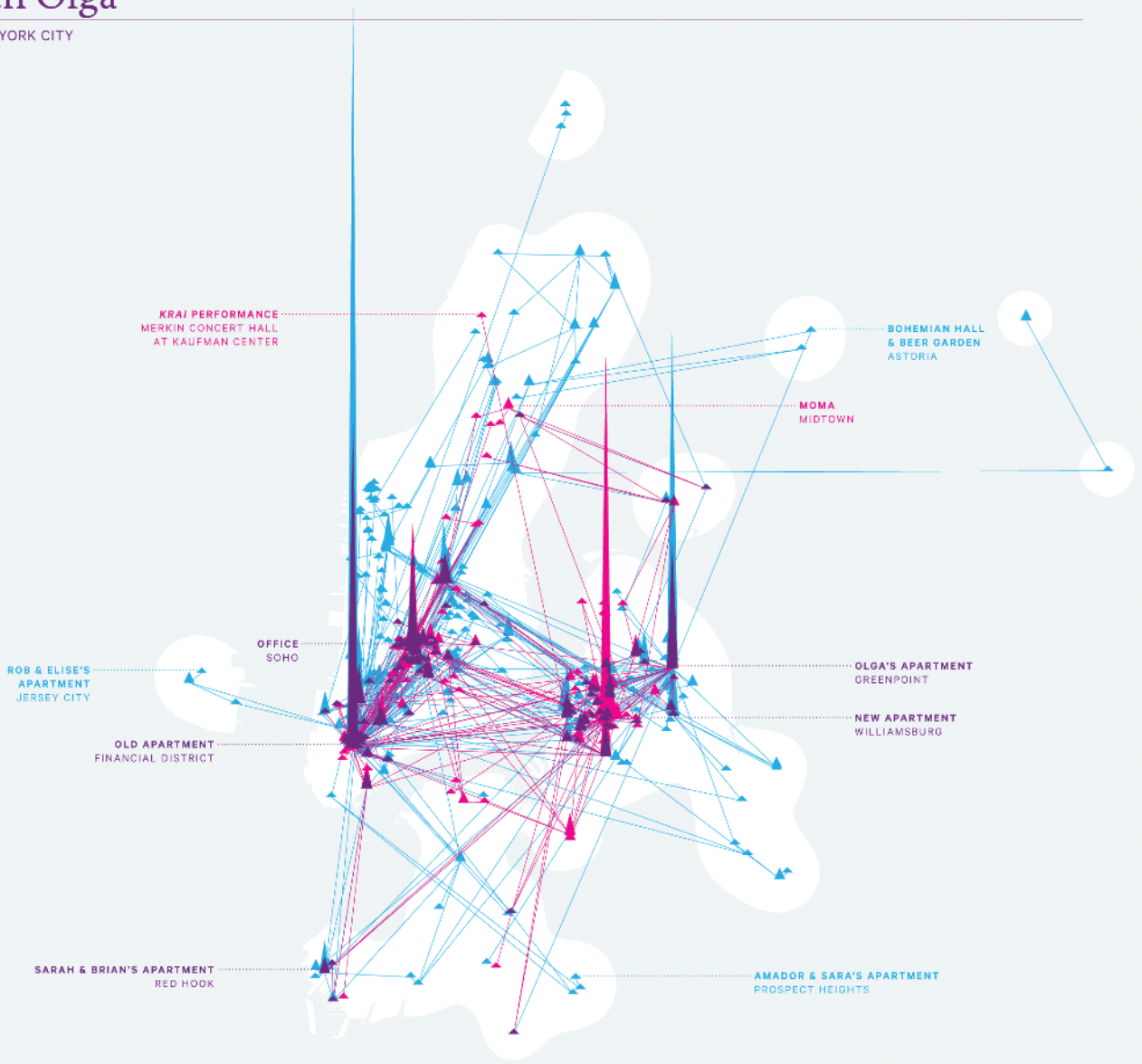
MOST VISITED RESTAURANTS TOGETHER

Le Garage, Picante and Sushi Ran

Each visited twice

With Olga

IN NEW YORK CITY



DAYS TOGETHER IN NEW YORK CITY

136¾

Approximately 72% of total time together

MOST VISITED NYC PLACES

OLD APARTMENT — 194 VISITS

OLGA'S APARTMENT — 84 VISITS

NEW APARTMENT — 67 VISITS

THE OFFICE — 35 VISITS

TAKAHACHI TRIBECA — 21 VISITS

TIME TOGETHER IN NEW YORK CITY

TIME IN NEW YORK SPENT WITH OLGA

31%

5% of time together spent in transit

MOST VISITED NYC RESTAURANTS

TAKAHACHI TRIBECA — 21 VISITS

LES HALLES ON JOHN STREET — 9 VISITS

DINER / ENID'S — 7 VISITS

MILLER'S TAVERN / FIVE LEAVES — 6 VISITS

RABBIT HOLE — 5 VISITS

FAVORITE NYC COCKTAIL WITH OLGA

Bloody Mary

22 servings

NYC PERFORMANCES WITH OLGA

Twenty-Eight

Bell (11), Bear in Heaven (3), Baths + How to Dress Well + Zola Jesus, Blonde Redhead + Pantha du Prince, Dexter Lake Club Band, Jason Nazary, Knights on Earth, Olga Bell *Krai*, Little Women, Nathan Fake + Four Tet, Now Ensemble + Matmos, Owen Pallett, Panda Bear, Pierre-Laurent Aimard, Sleigh Bells and *The Nose*

SIGNIFICANT NYC MISHAPS

Five

Abandoned keyboard stand, muddled dinner invitation date, missed ferry, shattered martini glass and smashed iPhone

composition, and how areas grow and shrink. In short, the data paints a picture of who lives in America. However, the data, collected and maintained by the government, can show only so much about the individuals, and it's hard to grasp who the people actually are.

What are their likes and dislikes? What kind of personality do they have? Are there major differences between neighboring cities and towns?

Media artist Roger Luke DuBois took a different kind of census, via 19 million online dating profiles in *A More Perfect Union*. When you join an online dating

FIGURE 1-8 *Selected pages from* 2005 Annual Report *by Nicholas Felton, http://feltron.com*

site, you first describe yourself: who you are, where you're from, and what you're interested in. After you uncomfortably fill out that information, and perhaps choose not to share a thing or two, you describe what your ideal mate is like. In the words of DuBois, in the latter, you tell the complete truth, and in the former, you lie. So when you aggregate people's online dating profiles, you get some combination of how people see themselves and how they want to be seen.

In *A More Perfect Union*, DuBois categorized online dating profiles, digital encapsulations of hopes and dreams, by postal code, and then looked for the word that was most unique to each area. Using a tracing of a Rand McNally map, DuBois replaced each city name with the city's unique word and painted a different picture of the United States: a more recognizable and personal one.

In Figure 1-10, around southern California, where they make the talkies, words such as *acting*, *writer,* and *entertainment* appear; on the other hand, in Washington, DC, shown in Figure 1-11, words like *bureaucrat*, *partisan*, and *democratic* appear. These mostly pertain to professions, but in some areas the words describe personal attributes, favorite things, and major events.

In Louisiana, shown in Figure 1-12, *Cajun* and *curvy* pop out at you, as does *crawfish*, *bourbon*, and *gumbo*, but in New Orleans, the most unique word is *flood*, a reflection of the effects of Hurricane Katrina in 2005.

People are defined by common demographic data such as race, age, and gender, but they also identify themselves with what they like to do in their spare time, what has happened to them, and who they hang around with. The great thing about *A More Perfect Union* is that you can see that in the data on a countrywide scale.

The same sentiment—where data points are recollections and reports are portraits and diaries—is seen in Felton's reports, Clark's atlas, and Parecki's GPS traces. Statisticians and developers call this analysis. Artists and designers call this storytelling. For extracting information from data, though—to understand what's in the numbers—analysis and storytelling are one and the same.

Just like what it represents, data can be complex with variability and uncertainty, but consider it all in the right context, and it starts to make sense.

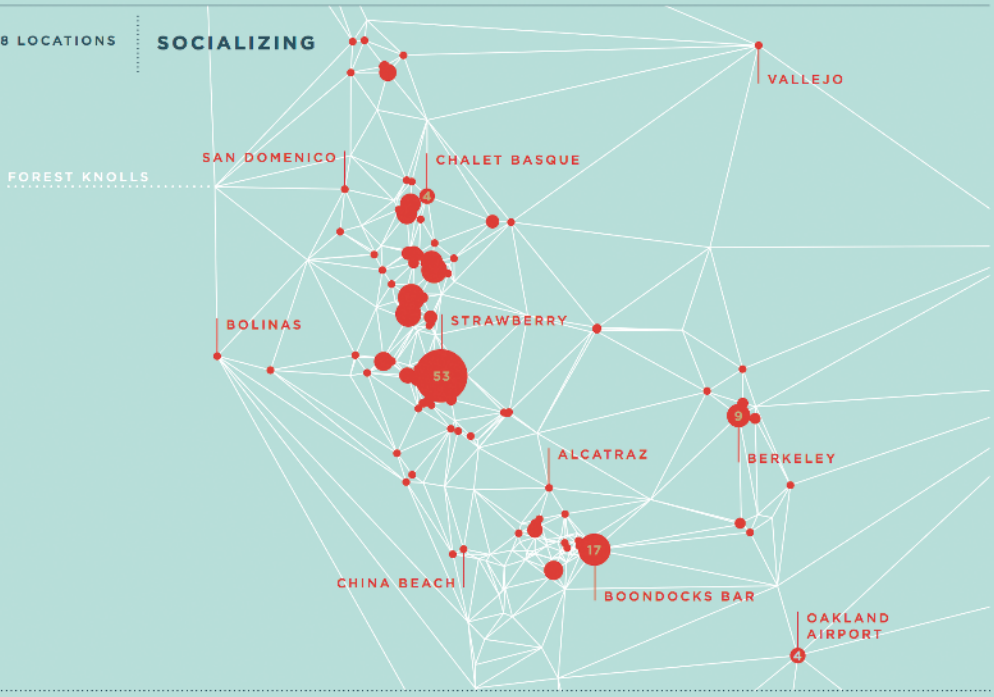
FIGURE 1-9 *(following page) Selected pages from* 2010 Annual Report *by Nicholas Felton, http://feltron.com*

2000-2010
SAN RAFAEL

THE 21ST CENTURY

138 LOCATIONS

SOCIALIZING



JAN 5, 2001
71 YEARS,
6 MONTHS
AND 1 DAY

PERSON SEEN THE MOST

MARINA
117 TIMES

BLACK PANTHERS MET

ONE
BOBBY SEALE

WALKS RECORDED

THIRTY-FIVE
AND 1 HIKE

2009-2010
GOLDEN GATE
TOLL BOOTH
PREFERENCE

LANE **6**

14 VISITS

ENTERTAINMENT

- 123 MOVIES
- 45 MUSIC
- 38 LECTURES
- 29 DANCES
- 29 POKER
- 22 PLAYS
- 19 TELEVISION
- 8 SLIDE SHOWS
- 5 ACROBATS
- 3 MIMES

MOST WATCHED TV SHOW

THE OSCARS
8 TIMES

LAST DAY

SEP 12, 2010
81 YEARS, 2 MONTHS
AND 8 DAYS OLD

WEATHER
SEP 12, 2010
3:20 PM

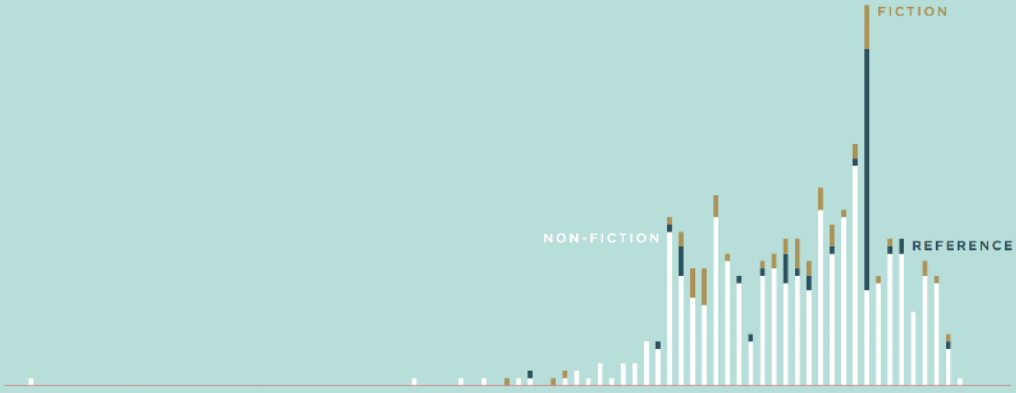
49.8° F AND OVERCAST
LARKSPUR, CALIFORNIA

1848–2009

BOOKS

536 BOOKS

DATE PUBLISHED



BOOKS

561

SPANNING 161 YEARS

MOST REPRESENTED AUTHOR

MARTIN GILBERT

5 BOOKS

TYPES OF BOOKS

- 88 TRAVEL
- 77 HISTORY
- 42 MACHINES
- 37 GEOGRAPHY
- 32 ENCYCLOPEDIA
- 26 RELIGION
- 23 HEALTH
- 23 NOVEL
- 22 SCIENCE
- 17 HOW-TO

BIOGRAPHIES

8

FROM ARMSTRONG TO STALIN

MEDIAN PUBLISHING DATE

1983

11 BOOKS

REGION WITH MOST TRAVEL BOOKS

RUSSIA

6 BOOKS

TRAVEL BOOKS FOR UNVISITED PLACES

SIX

AUSTRALIA, ICELAND, GREENLAND, IRAN, PAKISTAN AND VENEZUELA

COOKBOOKS

FIVE

WAR-RELATED BOOKS

51

35 BOOKS ABOUT WORLD WAR 2

ELEVATOR BOOKS

TWELVE

1941–1991

HOW-TO TOPICS

FOURTEEN

BICYCLES, CLEANING, CROSS COUNTRY SKIING, DOING EVERYTHING RIGHT, HANDICRAFT, HOME REPAIR, PEST CONTROL, PHOTOGRAPHY, PREVENTING AND SURVIVING FIRES, SAILING, SURVIVAL AND TAI CHI

FIGURE 1-10 *California map from* A More Perfect Union *(2011) by R. Luke DuBois, courtesy of the artist and bitforms gallery, New York City, http://perfect.lukedubois.com*

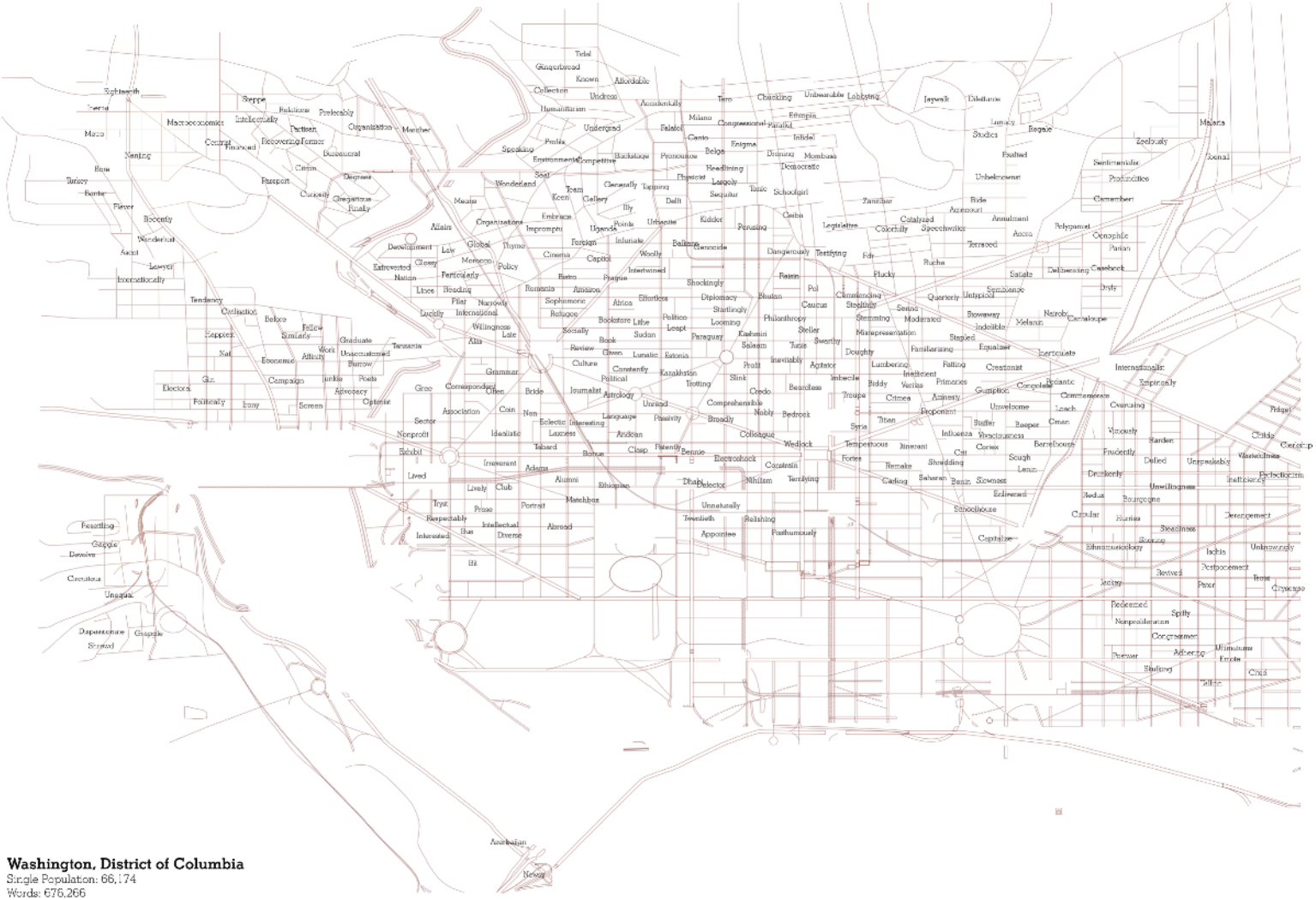


FIGURE 1-11 *Washington, DC map from* A More Perfect Union *(2011)*

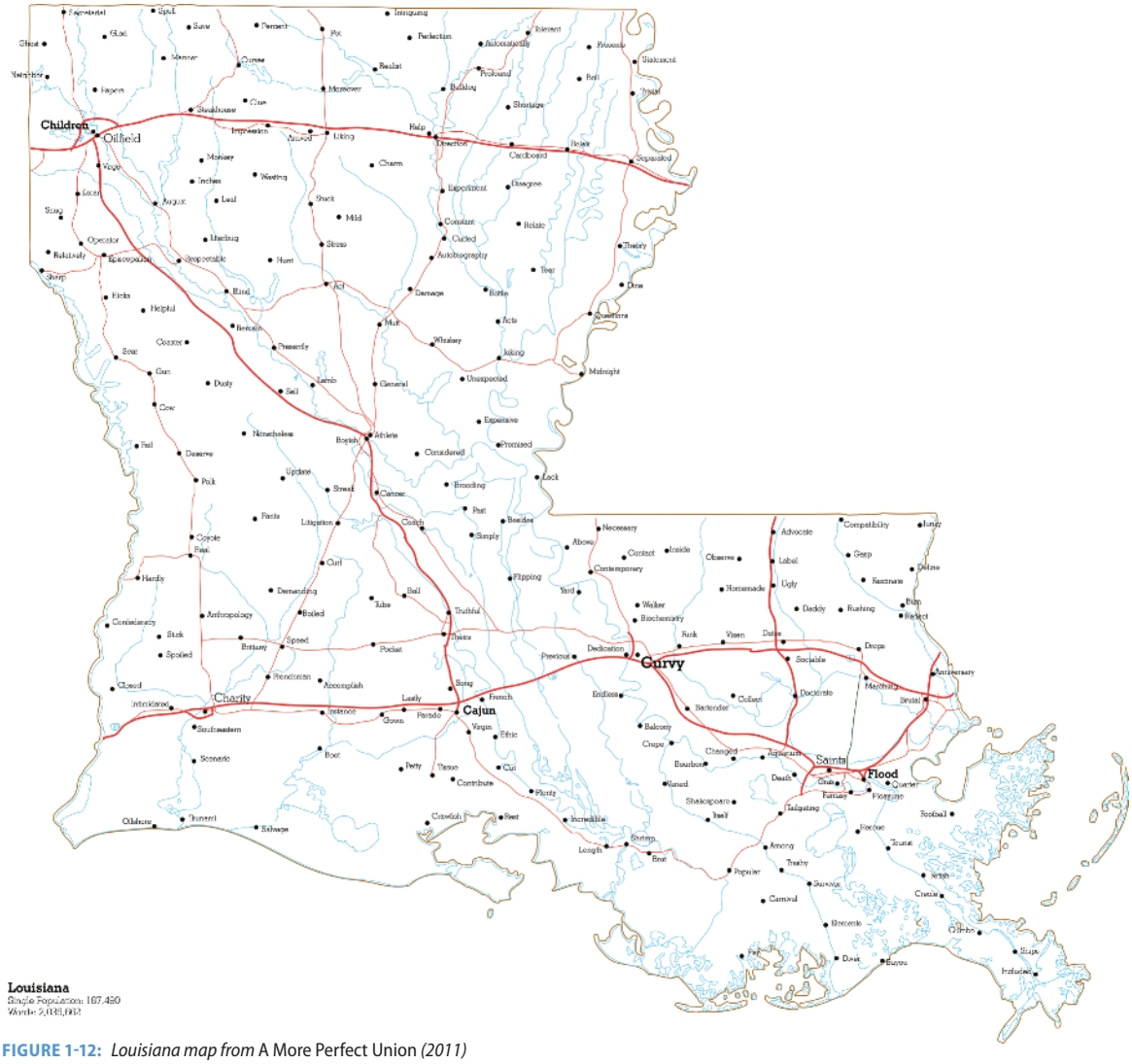


FIGURE 1-12: *Louisiana map from* A More Perfect Union *(2011)*

VARIABILITY

In a small town in Germany, amateur photographer and full-time physicist Kristian Cvecek heads out into the forest at night with his camera. Using long-exposure photography, Cvecek captures the movements of fireflies as they prance between the trees. The insect, as shown in Figure 1-13, is tiny and barely noticeable during the day, but in the dark, it's hard to look elsewhere.

FIGURE 1-13 *A firefly in the night by Kristian Cvecek, http://quit007.deviantart.com/*

Although each moment in flight seems like a random point in space to an observer, a pattern emerges in Cvecek's photos, as shown in Figure 1-14. It's as if the fireflies move along the walking path and circle around the trees with a predetermined destination.

There is randomness, though. You can guess where a firefly goes next based on its flight path, but how sure are you? A firefly can bolt left, right, up, and down at any moment, and that variability, which makes each flight unique, is what makes fireflies so fun to watch and the picture so beautiful. The path is what you care about. The end point, start point, and average position don't mean nearly as much.

With data, you can find patterns, trends, and cycles, but it's not always (rarely, actually) a smooth path from point A to point B. Total counts, means, and other aggregate measurements can be interesting, but they're only part of the story, whereas the fluctuations in the data might be the most interesting and important part.

Between 2001 and 2010, according to the National Highway Traffic Safety Administration, there were 363,839 fatal automobile crashes in the United States. No doubt this total count, over one-third of a million, carries weight because it represents the lost lives of even more than that. Place all focus on the one number, as in Figure 1-15, and it makes you think or maybe even reflect on your own life.

However, is there anything you can learn from the data, other than that you should drive safely? The NHTSA provides data down to individual accidents, which includes when and where each occurred, so you can look closer.

In Figure 1-16, every fatal crash in the contiguous United States between 2001 and 2010 is mapped. Each dot represents a crash. As you might expect, there is a higher concentration of accidents in large cities and major highways; there are fewer accidents where there are fewer people and roads.

FIGURE 1-14 Path of a firefly *by Kristian Cvecek, http://quit007.deviantart.com/*

FIGURE 1-15 *(facing page) One aggregate*

Fatal Crashes, 2001–2010

363,839

Source: National Highway Traffic Safety Administration

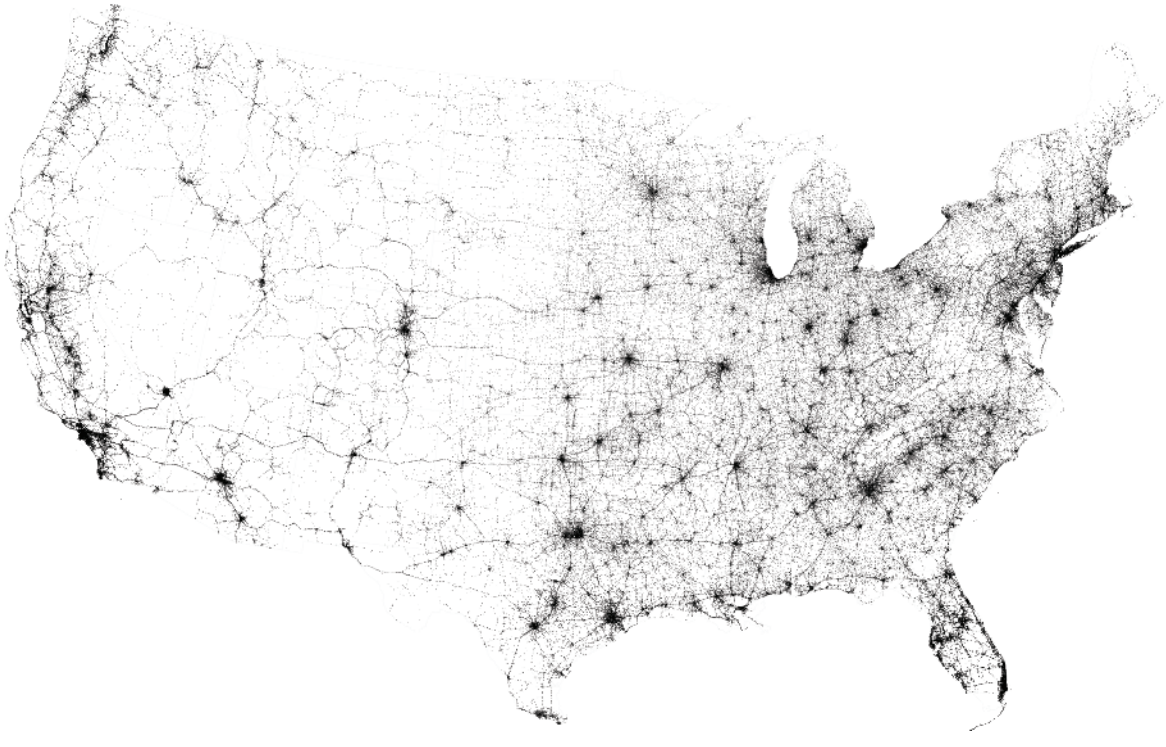


FIGURE 1-16 *Everything mapped at once*

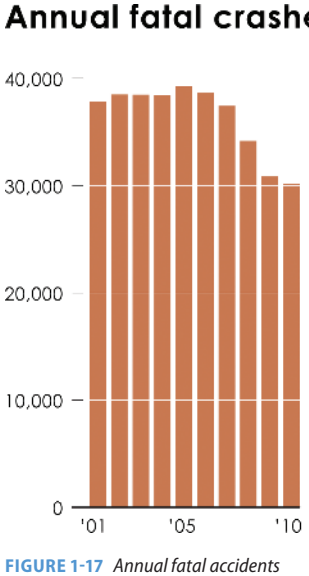


FIGURE 1-17 *Annual fatal accidents*

Again, although not to be taken lightly, the map tells you more about the country's road network than it does the accidents.

A look at crashes over time shifts focus to the events themselves. For example, Figure 1-17 shows the number of accidents per year, which tells a different story than the total in Figure 1-15. Accidents still occurred in the tens of thousands annually, but there was a significant decline from 2006 through 2010, and fatalities per 100 million vehicle miles traveled (not shown) also decreased.

Seasonal cycles become obvious at month-by-month granularity, as shown in Figure 1-18. Incidents peak during the summer months when people go on vacation and spend more time outside, whereas during the winter, fewer people drive, so there are fewer crashes. This happens every year. At the same time, you can still see the annual decline overall between 2006 and 2010.

However, there's variability when you compare specific months over the years. For example, in 2001, the most crashes occurred in August, and there was a small, relative drop the following month. The same thing happened in 2002 through 2004. However, in 2005 through 2007, July had the most accidents. Then it was back to August in 2008 through 2010.

On the other hand, February, the month with the fewest days had the least accidents every year, with the exception of 2008. So there are seasonal variations and variation within the seasons.

Go down another level to daily crashes, as shown in Figure 1-19, and you see even higher variability, but it's not all noise. There still appears to be a pattern of peaks and valleys. Although it's harder to make out the seasonal patterns, you can see a weekly cycle with more accidents during the weekends than during the middle of the week. The peak day each week fluctuates between Friday, Saturday, and Sunday.

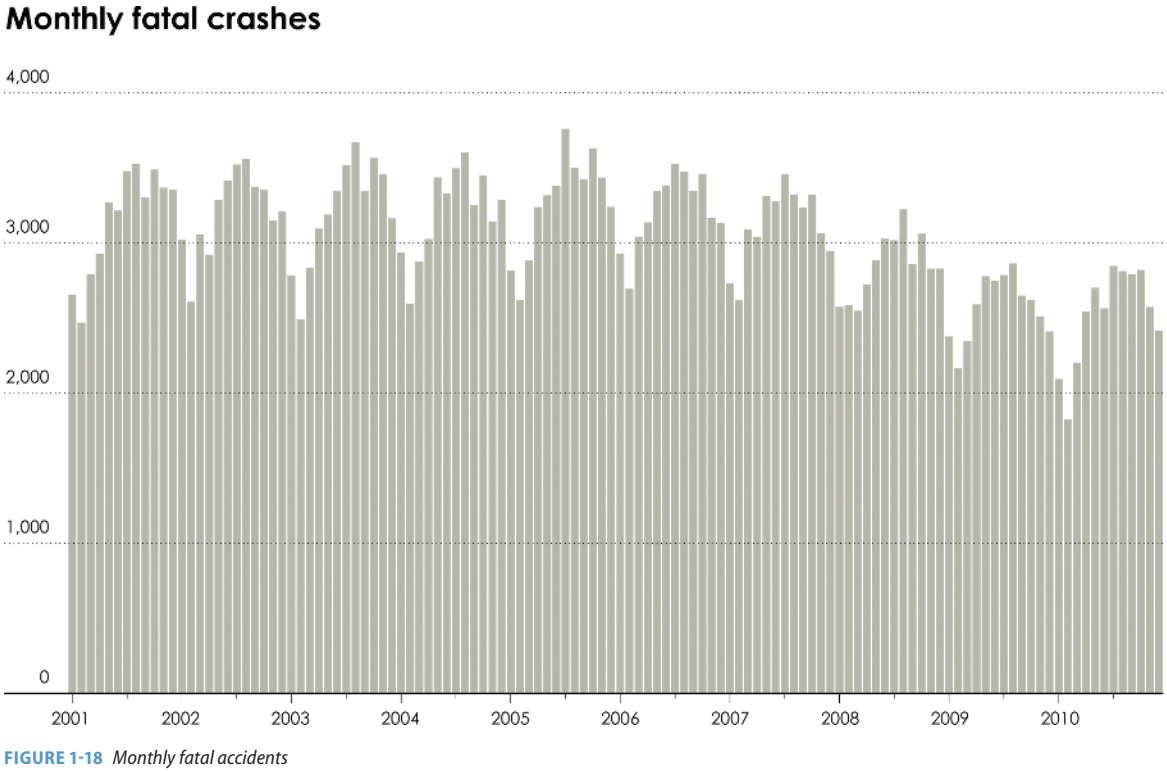


FIGURE 1-18 *Monthly fatal accidents*

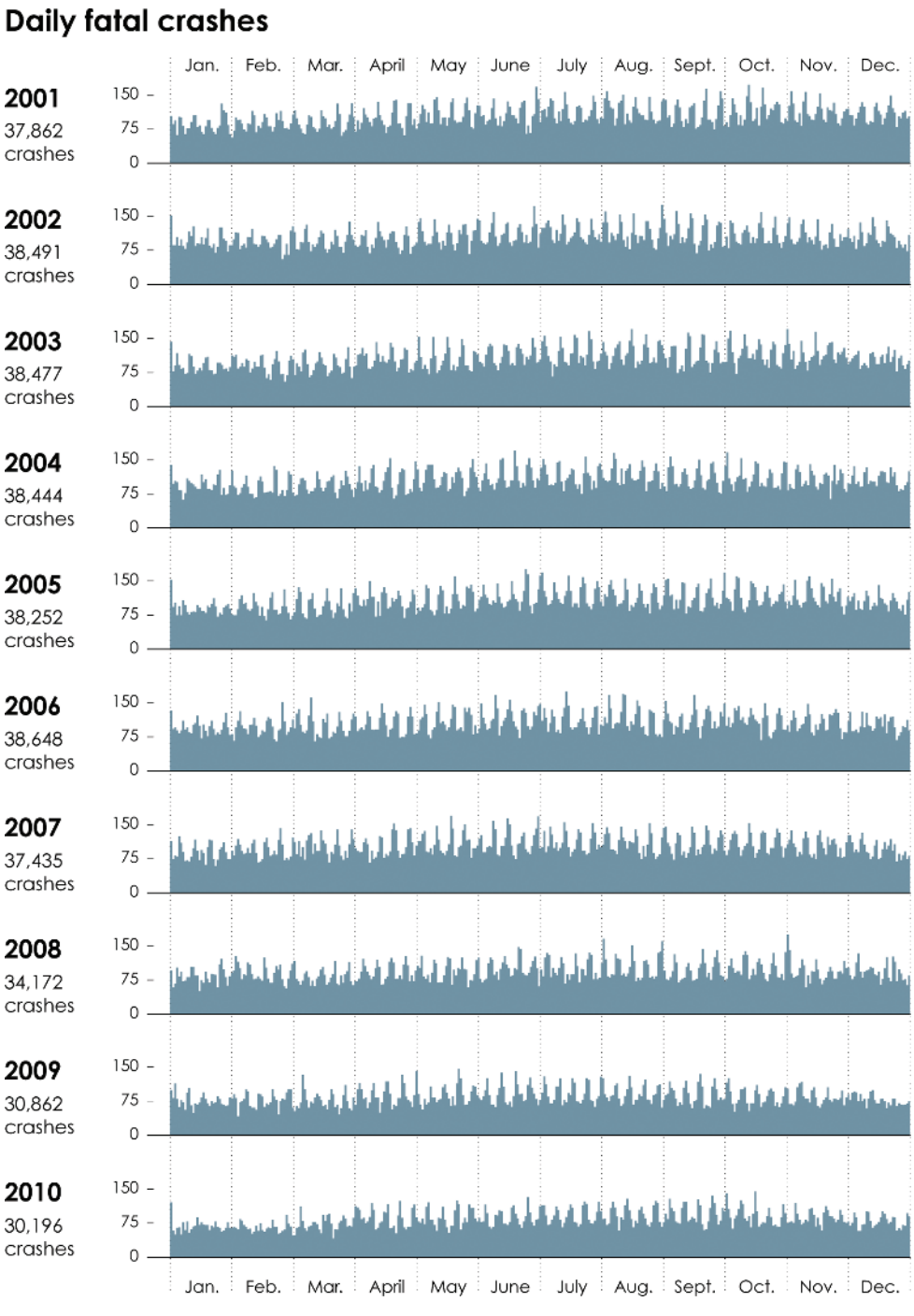


FIGURE 1-19 *Daily fatal accidents*

But guess what: You can increase granularity to crashes by the hour. Figure 1-20 breaks it down. Each row represents a year, so each cell in the grid shows an hourly time series for the corresponding month.

With the exception of a new year's spike during the midnight hour, it's hard to make out patterns at this level because of the variability. Actually, the monthly chart is hard to interpret, too, if you don't know what you're looking for. There are clear patterns, though, if you aggregate, as shown in Figure 1-21. Instead of showing values at every hour, day, or month, you can aggregate on specific time segments to explore the distributions.

What was hard to discern, or looked like noise before, is easy to see here. There's a small bump in the morning when people commute to work, but most fatal crashes occur in the evening after work. As you saw in Figure 1-19, there are more crashes during the weekend, but summed up, it's more obvious. Finally, you can see the seasonal patterns, but more clearly, with a greater number of accidents during the summer than in the winter.

The main point is that there's value in looking at the data beyond the mean, median, or total because those measurements tell you only a small part of the story. A lot of the time, aggregates or values that just tell you where the middle of a distribution is hide the interesting details that you should actually focus on, for both decision making and storytelling.

An outlier that stands out from the crowd could be something that you need to fix or pay special attention to. Maybe the changes over time are a signal that something good (or bad) is happening in your system. Cycles or regular occurrences could help you prepare for the future. However, sometimes it isn't helpful to see so much variability; in which case you can dial back the granularity for generalizations and distributions.

You lose this information—the juicy bits—when you step too far away from the data.

Think of it this way: When you look back on your life, would you rather just remember what your days were like on average, or is it the highs and the lows that are most important? I bet it's some combination of the two.

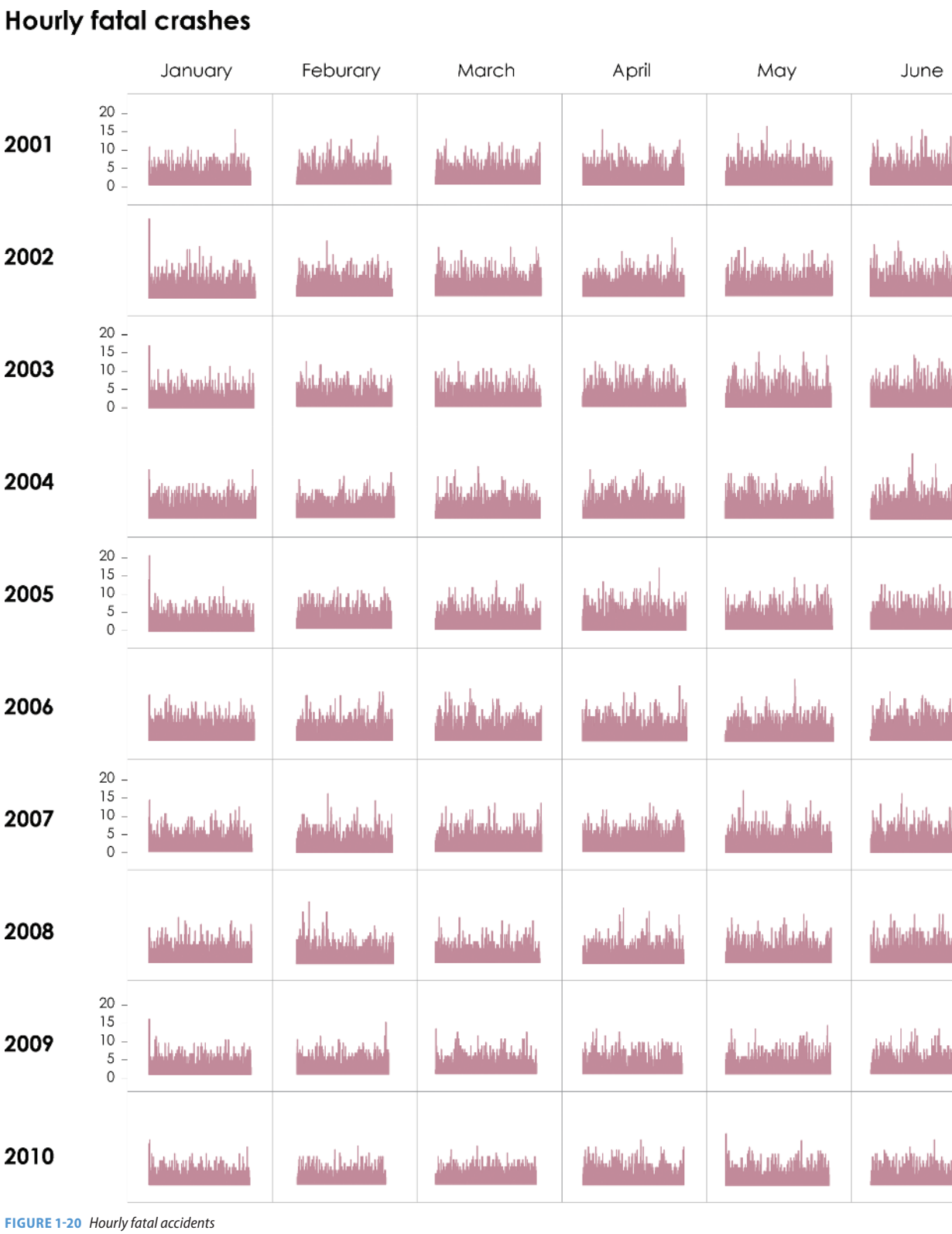


FIGURE 1-20 *Hourly fatal accidents*


July
August
September
October
November
December
20
15
10
5
0

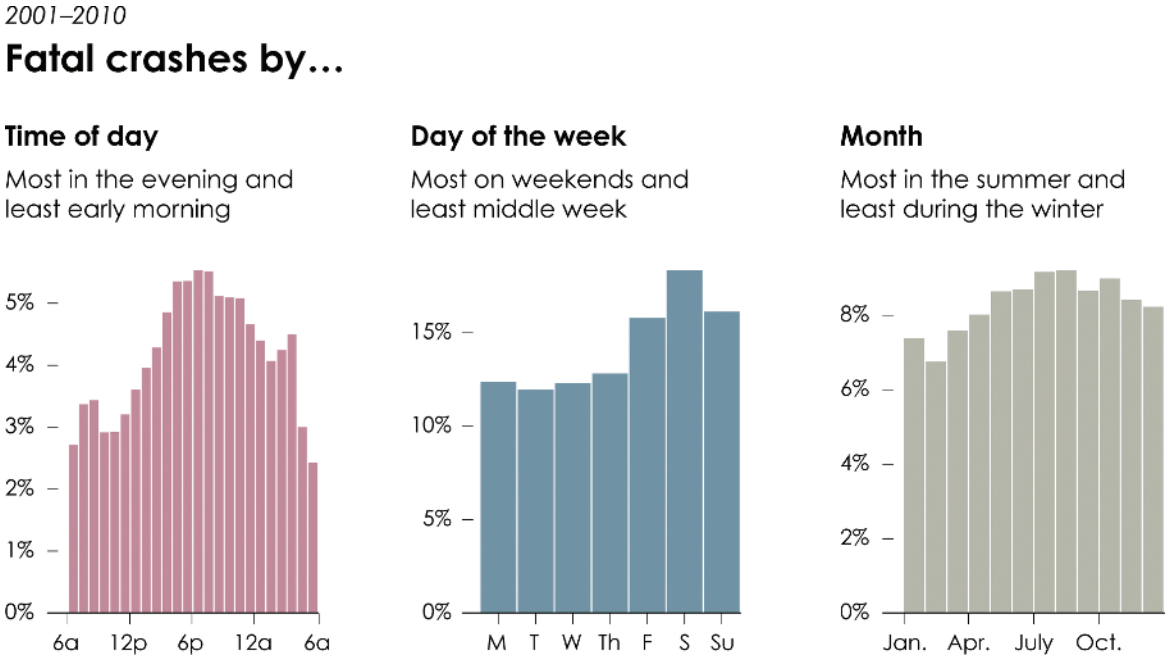


FIGURE 1-21 *Accident distributions over time*

UNCERTAINTY

A lot of data is estimates rather than absolute counts. An analyst considers the evidence (such as a sample), and makes an education guess about a full population. That educated guess has uncertainty attached to it. You do this all the time in your day-to-day. You make a guess based on what you know, read, or what someone told you, and you can say with some (possibly rough) certainty that you're right. Are you absolutely positive or are you basically clueless? It works the same with data.

Note: It's tempting to look at data as absolute truth, because we associate numbers with fact, but more often than not, data is an educated guess. Your goal is to use data that doesn't have large levels of uncertainty attached.

When I was a young lad, a recent engineering graduate with a statistics minor, I had a 9-month gap in between college and graduate school. I took a few temporary jobs that paid a little more than minimum wage, and they were mind-numbingly boring, so naturally my mind wandered to more engaging things.

One day I thought to myself, "Hey, I have some statistics and probability know-how and a deck of cards. I'm going to become an expert blackjack player like those kids from MIT. Forget this stupid job. I'm gonna be rich!" And my 1-month obsession with blackjack began. (To save you the suspense, I didn't get rich, and it's not nearly as exciting as they make it look in the movies.)

In case you're unfamiliar with the game, here's a quick rundown. There's a dealer and a player. The dealer deals two cards to each (one of his is face down), and the goal is to get a card total as close to 21 as possible, without going over. You can choose to take additional cards (called a *hit*) or not (*stay*). In some cases, you can also *split* your hand of two cards, as if you're playing two separate hands; you can also *double down*, which means to double your bet. The more you bet, the more you can win. If you go over 21, to *bust,* you automatically lose, and if not, the dealer hits or stays, and whoever is closer to 21 wins.

By design, the dealer has the advantage, but if you hit and stay when you're supposed to, you can decrease that advantage. These rules are based on averages, but as anyone who has played blackjack can tell you, there is uncertainty in each hand of cards. You can still lose even when you make the right move. For example, imagine you are dealt a 5 and a 6 for a total of 11, and the dealer shows a 6. The right move is to double your bet because it's impossible for you to bust with an additional card, and there's a decent chance of getting 21. There's also a good chance the dealer will bust with a 6 showing.

So you double down, and you get a 3, for a total of 14. Ouch. That's not good. Your only hope is for a dealer bust. So he flips his hidden card, and it's a 10 for a total of 16. By rule, he has to hit, and it's a 5. Dealer total: 21. You lose.

Had you not double downed, you would have lost only half the money that you did playing the right way. But if it were that easy to win, the casino wouldn't bother putting the game on the floor.

There's uncertainty in each hand because you are playing against distributions, or rather, you know only the approximate probabilities of drawing cards. You might have an idea of what cards are in the deck, but you can make only an educated guess about what card comes next.

Of course, uncertainty applies to things outside of cards, and it comes in a variety of forms. Take the weather for example. How many times have you looked up the forecast for the

Note: If you count cards, or keep track of what's left in the deck, the probabilities change as you modify your bet based on your advantage, but uncertainty remains.

next day or for the next week as you pack for a trip, only to find, when the time comes, that the weather isn't how you expected it to be?

What about the meter in cars that tells you how much farther you can drive with the current amount of fuel in your tank? I was running errands with my wife, and the meter said I could drive an estimated 16 more miles, but home was about 18 miles away. Dilemma. Instead of stopping at the nearest gas station, I drove toward the one nearest home, and the meter said I had zero miles left for about 2 miles, but we made it. (Good thing because someone kept insisting that I would be the one to push the car.)

Weigh yourself more than once, and you might get different readings; typically though, breathing for a few seconds does not lead to weight loss or gain. The estimated battery life on your laptop can jump around by hour increments when only minutes have passed. The subway announcement says a train will arrive in 10 minutes, but it comes in 11, or a delivery is estimated to arrive on Monday, but it comes on Wednesday instead.

When you have data that is a series of means and medians or a collection of estimates based on a sample population, you should always wonder about the uncertainty.

Note: Numbers seem concrete and absolute, but estimates carry uncertainity with them. Data is an abstraction of what it represents, and the level of exactness varies.

This is especially important when people base major decisions, which affect millions, on estimates, such as with national and global demographics. Program creation and funding is often based on these numbers, so even a small margin of error can make a big difference.

The United States Census Bureau releases data about the country on topics such as migration, poverty, and housing, which are estimates based on samples from the population. (This is different from the decennial census, which aims to count every person in the United States.) A margin of error is provided with each estimate, which means that the actual count or percentage is likely within a given range. For example, Figure 1-22 shows estimates about housing. The margin of error for total households is almost one-quarter of a million.

To put it differently, imagine you have a jar of gumballs that you can't see into, and you want to guess how many of each color there are. (Why do you care about gumball distribution? I don't know. Use your imagination. You're a gumball connoisseur who works for a gumball factory, and you bet your snotty statistician friend that every jar on your watch is uniformly distributed, so it's a matter of pride and cash.)

If you were to pour all the gumballs on to the table and count every one, you wouldn't have to guess because you would get the full tally.

But say you can grab only a handful, and you have to guess the contents of the entire jar, based on what you have in your hand. A larger handful would make it easier to guess because it's more likely a better representation of the entire jar. On the other side of the spectrum, you could take just one gumball out, and it'd be much harder to guess what else is in the jar.

With one gumball, your margin of error would be high; with a large handful of gumballs, your margin of error would be lower; and if you counted all the gumballs, you would have zero margin of error.

Apply that to millions of gumballs in thousands of differently sized jars, with different distributions and big and small handfuls, and estimation

	Estimate	Margin of error
Total households	114,235,996	± 248,114
Total families	76,254,318	± 230,785
Average family size	3.17	± 0.01
Married-couple family households	56,655,412	± 293,638
Married, 15 and over	50.2%	± 0.2
Divorced, 15 and over	10.5%	± 0.1

Source: 2010 American Community Survey

FIGURE 1-22 *Household estimates in 2010*

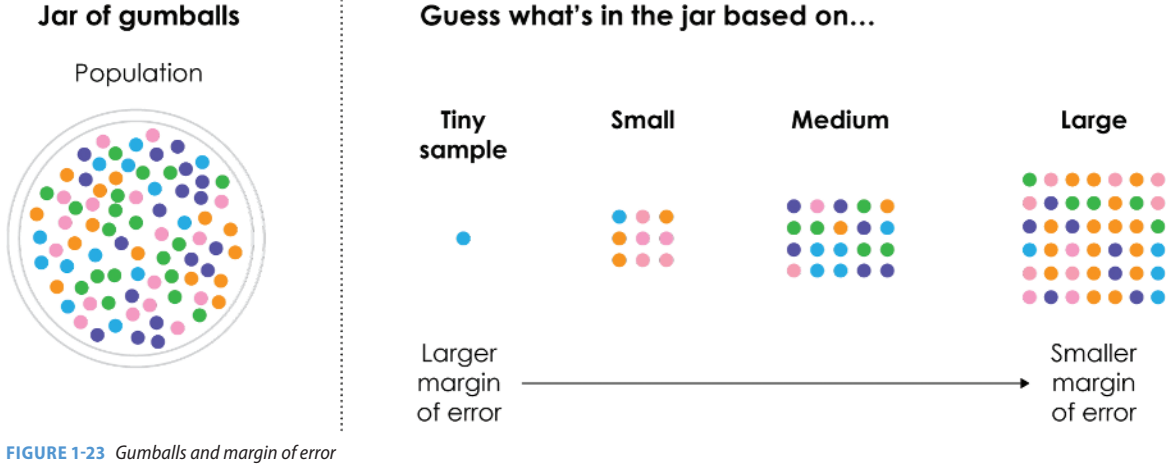


FIGURE 1-23 *Gumballs and margin of error*

grows more complex. Then substitute the gumballs for people, the jars for towns, cities, and counties, and the handfuls for randomly distributed surveys, and a mean with a margin of error carries more weight.

According to Gallup, 48 percent of Americans disapproved of the job Barack Obama was doing from June 11 through 13 in 2012. However, there was a 3 percent margin of error, which means the difference between more than half and less than half of the country disapproving. Similarly, during election season, polls estimate which candidates lead, and if the margin of error is wide, the results can put more than one person in front, which kind of defeats the purpose of the poll.

Estimates get tricky when you rank people, places, and things, especially when you combine measurements (and create statistical models with multiple variables).

Take education evaluation, for example, which is under constant scrutiny. Cities, schools, and teachers are often compared against one another, but what defines a good education or makes an entire city smart? Is it the percentage of high school students who graduate? The percentage of students who go to college? Is it the number of universities, libraries, and museums per capita? If it's all of this, is one count more important than the other, or do you give all of them equal weight? Answers change depending on who you ask, as do ratings.

Note: My hometown was ranked the "dumbest" city in America by a publication that shall go unnamed. The rankings were estimates, which were based on estimates with questionable uncertainty.

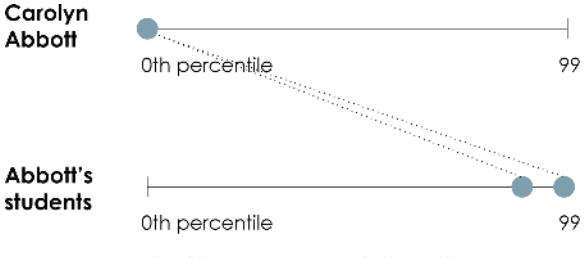


FIGURE 1-24 *Carolyn Abbott's rating compared to her students'*

In 2011, the New York City Department of Education released Teacher Data Reports that tried to measure teaching quality. The reports were originally given only to schools and teachers but were later made publicly available in early 2012. The estimates took several factors into account, but one of the main ones was the change in test percentiles from the seventh to eighth grade.

This is how seventh- and eighth-grade math teacher Carolyn Abbott became known as the worst math teacher in the city, placed in the 0^{th} percentile. However, her seventh-grade students scored in the 98^{th} percentile. What?

Those students were predicted to score in the 97^{th} percentile in the eighth grade, but they

instead scored in the 89th percentile, which according to the statistical model, was not progress. Most would agree that students wouldn't earn the scores they did with a poor teacher. The challenge is that there's uncertainty and variability within teacher ratings. A rating represents a distribution of teachers, who are ranked based on estimates with uncertainty attached, but the ratings are treated as absolute. A general audience won't understand that concept, so it's your responsibility to and communicate it clearly.

When you don't consider what your data truly represents, it's easy to accidently misinterpret. Always take uncertainty and variability into account. This is also when context comes into play.

CONTEXT

Look up at the night sky, and the stars look like dots on a flat surface. The lack of visual depth makes the translation from sky to paper fairly straightforward, which makes it easier to imagine constellations. Just connect the dots. However, although you perceive stars to be the same distance away from you, they are actually varying light years away.

If you could fly out beyond the stars, what would the constellations look like? This is what Santiago Ortiz wondered as he visualized stars from a different perspective, as shown in Figure 1-25.

The initial view places the stars in a global layout, the way you see them. You look at Earth beyond the stars, but as if they were an equal distance away from the planet.

Zoom in, and you can see constellations how you would from the ground, bundled in a sleeping bag in the mountains, staring up at a clear sky.

The perceived view is fun to see, but flip the switch to show actual distance, and it gets interesting. Stars transition, and the easy-to-distinguish constellations are practically unrecognizable. The data looks different from this new angle.

This is what context can do. It can completely change your perspective on a dataset, and it can help you decide what the numbers represent and how to interpret them. After you do know what the data is about, your understanding helps you find the fascinating bits, which leads to worthwhile visualization.

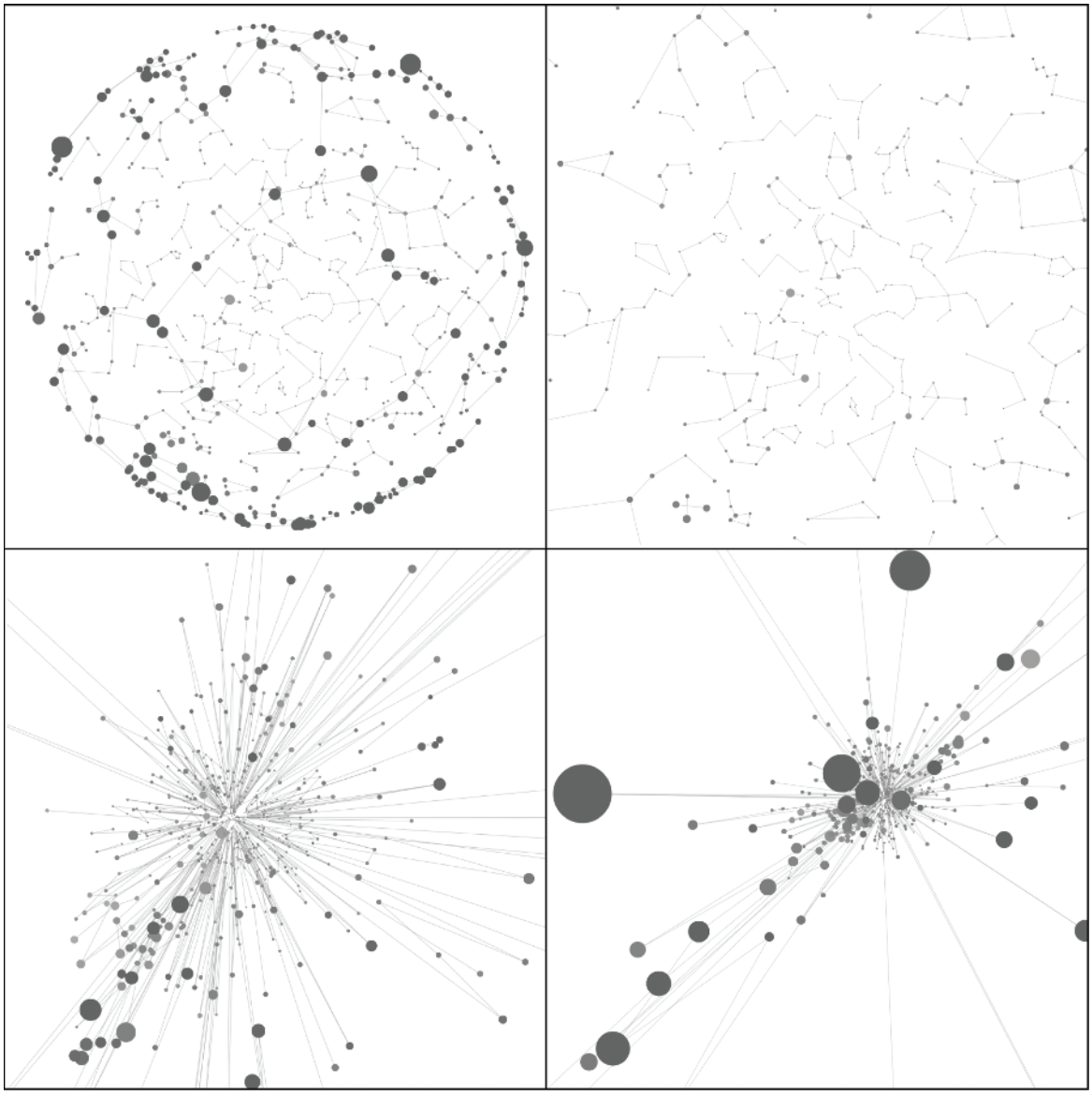


FIGURE 1-25 View of the Sky *by Santiago Ortiz, http://moebio.com/exomap/viewsofthesky/2/*

Without context, data is useless, and any visualization you create with it will also be useless. Using data without knowing anything about it, other than the values themselves, is like hearing an abridged quote secondhand and then citing it as a main discussion point in an essay. It might be okay, but you risk finding out later that the speaker meant the opposite of what you thought.

You have to know the who, what, when, where, why, and how—the metadata, or the data about the data—before you can know what the numbers are actually about.

Who: A quote in a major newspaper carries more weight than one from a celebrity gossip site that has a reputation for stretching the truth. Similarly, data from a reputable source typically implies better accuracy than a random online poll.

For example, Gallup, which has measured public opinion since the 1930s, is more reliable than say, someone (for example, me) experimenting with a small, one-off Twitter sample late at night during a short period of time. Whereas the former works to create samples representative of a region, there are unknowns with the latter.

Speaking of which, in addition to who collected the data, who the data is about is also important. Going back to the gumballs, it's often not financially feasible to collect data about everyone or everything in a population. Most people don't have time to count and categorize a thousand gumballs, much less a million, so they sample. The key is to sample evenly across the population so that it is representative of the whole. Did the data collectors do that?

How: People often skip methodology because it tends to be complex and for a technical audience, but it's worth getting to know the gist of how the data of interest was collected.

If you're the one who collected the data, then you're good to go, but when you grab a dataset online, provided by someone you've never met, how will you know if it's any good? Do you trust it right away, or do you investigate? You don't have to know the exact statistical model behind every dataset, but look out for small samples, high margins of error, and unfit assumptions about the subjects, such as indices or rankings that incorporate spotty or unrelated information.

Sometimes people generate indices to measure the quality of life in countries, and a metric like literacy is used as a factor. However, a country might not have up-to-date information on literacy, so the data gatherer simply uses an estimate from a decade earlier. That's going to cause problems because then the index works only under the assumption that the literacy rate one decade earlier is comparable to the present, which might not be (and probably isn't) the case.

What: Ultimately, you want to know what your data is about, but before you can do that, you should know what surrounds the numbers. Talk to subject experts, read papers, and study accompanying documentation.

In introduction statistics courses, you typically learn about analysis methods, such as hypothesis testing, regression, and modeling, in a vacuum, because the goal is to learn the math and concepts. But when you get to real-world data, the goal shifts to information gathering. You shift from, "What is in the numbers?" to "What does the data represent in the world; does it make sense; and how does this relate to other data?"

A major mistake is to treat every dataset the same and use the same canned methods and tools. Don't do that.

When: Most data is linked to time in some way in that it might be a time series, or it's a snapshot from a specific period. In both cases, you have to know when the data was collected. An estimate made decades ago does not equate to one in the present. This seems obvious, but it's a common mistake to take old data and pass it off as new because it's what's available. Things change, people change, and places change, and so naturally, data changes.

Where: Things can change across cities, states, and countries just as they do over time. For example, it's best to avoid global generalizations when the data comes from only a few countries. The same logic applies to digital locations. Data from websites, such as Twitter or Facebook, encapsulates the behavior of its users and doesn't necessarily translate to the physical world.

Although the gap between digital and physical continues to shrink, the space between is still evident. For example, an animated map that represented the "history of the world" based on geotagged Wikipedia, showed popping dots for each entry, in a geographic space. The end of the video is shown in Figure 1-26.

The result is impressive, and there is a correlation to the real-life timeline for sure, but it's clear that because Wikipedia content is more prominent in English-speaking countries the map shows more in those areas than anywhere else.

Why: Finally, you must know the reason data was collected, mostly as a sanity check for bias. Sometimes data is collected, or even fabricated, to serve an agenda, and you should be wary of these cases. Government and elections might be the first thing that come to mind, but so-called information graphics around the web, filled with keywords and published by sites trying to grab Google juice, have also grown up to be a common culprit. (I fell for these a couple of times in my early days of blogging for FlowingData, but I learned my lesson.)

Learn all you can about your data before anything else, and your analysis and visualization will be better for it. You can then pass what you know on to readers.

FIGURE 1-26 A History of the World in 100 Seconds *by Gareth Lloyd, http://datafl.ws/24a*

However, just because you have data doesn't mean you should make a graphic and share it with the world. Context can help you add a dimension—a layer of information—to your data graphics, but sometimes it means it's better to hold back because it's the right thing to do.

In 2010, Gawker Media, which runs large blogs like Lifehacker and Gizmodo, was hacked, and 1.3 million usernames and passwords were leaked. They were downloadable via BitTorrent. The passwords were encrypted, but the hackers cracked about 188,000 of them, which exposed more than 91,000 unique passwords. What would you do with that kind of data?

The mean thing to do would be to highlight usernames with common (read that poor) passwords, or you could go so far as to create an application that guessed passwords, given a username.

A different route might be to highlight just the common passwords, as shown in Figure 1-27. This offers some insight into the data without making it too easy to log in with someone else's account. It might also serve as a warning to others to change their passwords to something less obvious. You know, something with at least two symbols, a digit, and a mix of lowercase and uppercase letters. Password rules are ridiculous these days. But I digress.

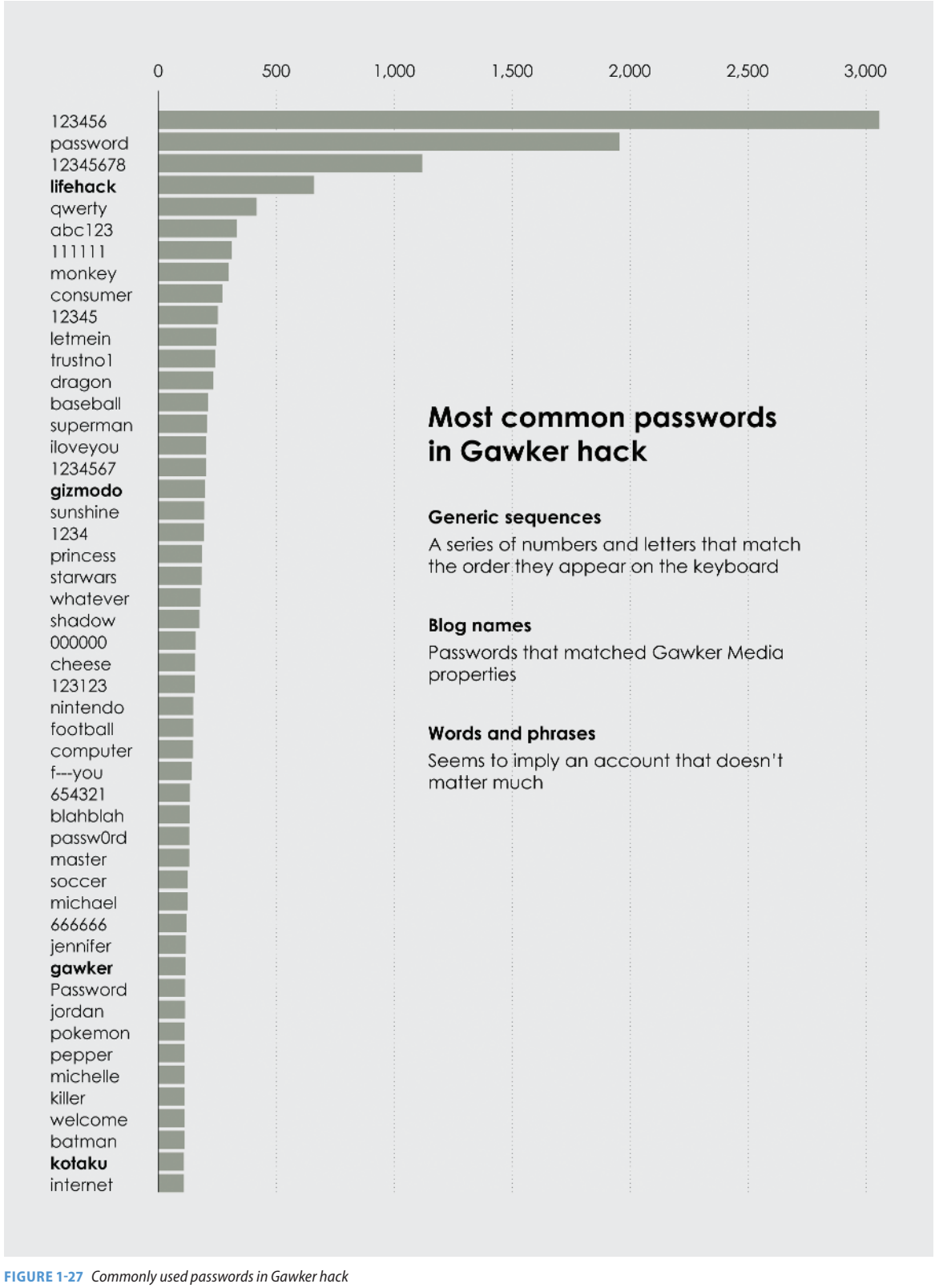


FIGURE 1-27 *Commonly used passwords in Gawker hack*

With data like the Gawker set, a deep analysis might be interesting, but it could also do more harm than good. In this case, data privacy is more important, so it's better to limit what you show and look at.

Whether you should use data is not always clear-cut though. Sometimes, the split between what's right and wrong can be gray, so it's up to you to make the call. For example, on October 22, 2010, Wikileaks, an online organization that releases private documents and media from anonymous sources, released 391,832 United States Army field reports, now known as the Iraq War Logs. The reports recorded 66,081 civilian deaths out of 109,000 recorded deaths, between 2004 and 2009.

The leak exposed incidents of abuse and erroneous reporting, such as civilian deaths classified as "enemy killed in action." On the other hand, it can seem unjustified to publish findings about classified data obtained through less than savory means.

Maybe there should be a golden rule for data: Treat others' data the way you would want your data treated.

In the end, it comes back to what data represents. Data is an abstraction of real life, and real life can be complicated, but if you gather enough context, you can at least put forth a solid effort to make sense of it.

WRAPPING UP

Visualization is often thought of as an exercise in graphic design or a brute-force computer science problem, but the best work is always rooted in data. To visualize data, you must understand what it is, what it represents in the real world, and in what context you should interpret it in.

Data comes in different shapes and sizes, at various granularities, and with uncertainty attached, which means totals, averages, and medians are only a small part of what a data point is about. It twists. It turns. It fluctuates. It can be personal, and even poetic. As a result, you can find visualization in many forms.