CONTENTS

Introduction

You have by now heard a lot about Big Data: the vast potential, the ominous consequences, the paradigm-destroying *new paradigm* it portends for mankind and his ever-loving websites. The mind reels, as if struck by a very dull object. So I don't come here with more hype or reportage on the data phenomenon. I come with the thing itself: the data, phenomenon stripped away. I come with a large store of the actual information that's being collected, which luck, work, wheedling, and more luck have put me in the unique position to possess and analyze.

I was one of the founders of OkCupid, a dating website that, over a very un-bubbly long haul of ten years, has become one of the largest in the world. I started it with three friends. We were all mathematically minded, and the site succeeded in large part because we applied that mind-set to dating; we brought some analysis and rigor to what had historically been the domain of love "experts" and grinning warlocks like Dr. Phil. How the site works isn't all that sophisticated—it turns out the only math you need to model the process of two people getting to know each other is some sober arithmetic—but for whatever reason, our approach resonated, and this year alone 10 million people will use the site to find someone.

As I know too well, websites (and founders of websites) love to throw out big numbers, and most thinking people have no doubt learned to ignore them; you hear millions of this and billions of that and know it's basically "Hooray for me," said with trailing zeros. Unlike Google, Facebook, Twitter, and the other sources whose data will figure prominently in this book, OkCupid is far from a household name—if you and your friends have all been happily married for years, you've probably never heard of us. So I've thought a lot about how to describe the reach of the site to someone who's never used it and who rightly doesn't care about the user-engagement metrics of some guy's startup. I'll put it in personal terms instead. Tonight, some thirty thousand couples will have their first date because

of OkCupid. Roughly three thousand of them will end up together long-term. Two hundred of those will get married, and many of them, of course, will have kids. There are children alive and pouting today, grouchy little humans refusing to put their shoes on *right now*, who would never have existed but for the whims of our HTML.

I have no smug idea that we've perfected anything, and it's worth saying here that while I'm proud of the site my friends and I started, I honestly don't care if you're a member or go create an account or what. I've never been on an online date in my life and neither have any of the other founders, and if it's not for you, believe me, I get that. Tech evangelism is one of my least favorite things, and I'm not here to trade my blinking digital beads for anyone's precious island. I still subscribe to magazines. I get the *Times* on the weekend. Tweeting embarrasses me. I can't convince you to use, respect, or "believe in" the Internet or social media any more than you already do—or don't. By all means, keep right on thinking what you've been thinking about the online universe. But if there's one thing I sincerely hope this book might get you to reconsider, it's what you think about yourself. Because that's what this book is really about. OkCupid is just how I arrived at the story.

I have led OkCupid's analytics team since 2009, and my job is to make sense of the data our users create. While my three founding partners have done almost all the hard work of actually building the site, I've spent years just playing with the numbers. Some of what I work on helps us run the business: for example, understanding how men and women view sex and beauty differently is essential for a dating site. But a lot of my results aren't directly useful—just interesting. There's not much you can do with the fact that, statistically, the least black band on Earth is Belle & Sebastian, or that the flash in a snapshot makes a person look seven years older, except to say *huh*, and maybe repeat it at a dinner party. That's basically all we did with this stuff for a while; the insights we gleaned went no further than an occasional lame press release. But eventually we were analyzing enough information that larger trends became apparent, big patterns in the small ones, and, even better, I realized I could use the data to examine taboos like race by direct inspection. That is, instead of asking people survey questions or contriving small-scale experiments, which was how social science was often done in the past, I could go and look at *what actually happens* when, say, 100,000 white men and 100,000 black women interact in private. The data was sitting right there on our servers. It was an irresistible sociological opportunity.

I dug in, and as discoveries built up, like anyone with more ideas than audience, I started a blog to share them with the world. That blog then became this book, after one important improvement. For *Dataclysm*, I've gone far beyond OkCupid. In fact, I've probably put together a data set of person-to-person interaction that's deeper and more varied than anything held by any other private individual—spanning most, if not all, of the significant online data sources of our time. In these pages I'll use my data to speak not just to the habits of one site's users but also to a set of universals.

The public discussion of data has focused primarily on two things: government spying and commercial opportunity. About the first, I doubt I know any more than you—only what I've read. To my knowledge, the national security apparatus has never approached any dating site for access, and unless they plan to criminalize the faceless display of utterly ripped abs or young women from Brooklyn going on and on about how much they like scotch, when, come on, you know they really don't, I can't imagine they'd find much of interest. About the second story, data-as-dollars, I know better. As I was beginning this book, the tech press was slick with drool over the Facebook IPO; they'd collected everyone's personal data and had been turning it into all this money, and now they were about to turn *that* money into even more money in the public markets. A *Times* headline from three days before the offering says it all: "Facebook Must Spin Data into Gold." You half expected Rumpelstiltskin to show up on the OpEd page and be like, "Yes, America, this is a solid buy."

As a founder of an ad-supported site, I can confirm that data *is* useful for selling. Each page of a website can absorb a user's entire experience—everything he clicks, whatever he types, even how long he lingers—and from this it's not hard to form a clear picture of his appetites and how to sate them. But awesome though the power may be, I'm not here to go over our nation's occult mission to sell body spray to people who update their friends about body spray. Given the same access to the data, I am going to put that user experience—the clicks, keystrokes, and milliseconds—to another end. If Big Data's two running stories have been surveillance and money, for the last three years I've been working on a third: the human story.

Facebook might know that you're one of M&M's many fans and send you offers accordingly. They also know when you break up with your boyfriend, move to Texas, begin appearing in lots of pictures with your ex, and start dating him again. Google knows when you're looking for a new car and can show the make and model preselected for just your psychographic. A thrill-seeking socially conscious Type B, M, 25–34? Here's your Subaru. At the same time, Google also knows if you're gay or angry or lonely or racist or worried that your mom has cancer. Twitter, Reddit, Tumblr, Instagram, all these companies are businesses first, but, as a close second, they're demographers of unprecedented reach, thoroughness, and importance. Practically as an accident, digital data can now show us how we fight, how we love, how we age, who we are, and how we're changing. All we have to do is look: from just a very slight remove, the data reveals how people behave when they think no one is watching. Here I will show you what I've seen. Also, fuck body spray.

∞

If you read a lot of popular nonfiction, there are a couple things in *Dataclysm* that you might find unusual. The first is the color red. The second is that the book deals in aggregates and big numbers, and that makes for a curious absence in a story supposedly about people: there are very few individuals here. Graphs and charts and tables appear in abundance, but there are almost no names. It's become a cliché of pop science to use something small and quirky as a lens for big events—to tell the history of the world via a turnip, to trace a war back to a fish, to shine a penlight through a prism *just so* and cast the whole pretty rainbow on your bedroom wall. I'm going in the opposite direction. I'm taking something big—an enormous set of what people are doing and thinking and saying, terabytes of data—and filtering from it many small things: what your network of friends says about the stability of your marriage, how Asians (and whites and blacks and Latinos) are least likely to describe themselves, where and why gay people stay in the closet, how writing has changed in the last ten years, and how anger hasn't. The idea is to move our understanding of ourselves away from narratives and toward numbers, or, rather, to think in such a way that numbers *are* the narrative.

This approach evolved from long toil in the statistical slag pits. *Dataclysm* is an extension of what my coworkers and I have been doing for years. A dating site brings people together, and to do that credibly it has to get at their desires, habits, and revulsions. So you collect a lot of detailed data and work very hard to translate it all into general theories of human behavior. What a person develops working amidst all this information, as opposed to, say, working for the wedding section of the Sunday paper, is a special kinship with the shambling whole of humanity rather than with any two individuals. You grow to understand people much as a chemist might understand, and through understanding come to love, the swirling molecules of his tincture.

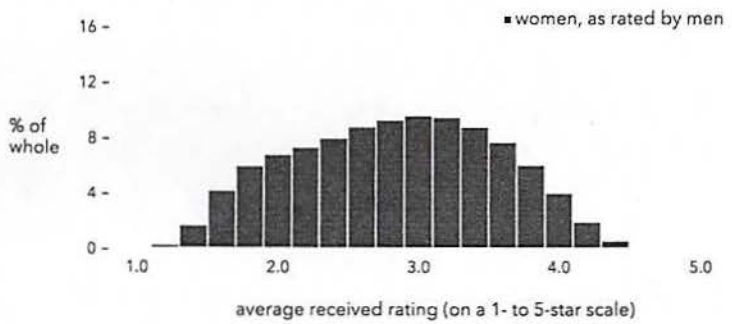
That said, all websites, and indeed all data scientists, objectify. Algorithms don't work well with things that aren't numbers, so when you want a computer to understand an idea, you have to convert as much of it as you can into digits. The challenge facing sites and apps is thus to chop and jam the continuum of human experience into little buckets 1, 2, 3, without anyone noticing: to divide some vast, ineffable process—for Facebook, friendship, for Reddit, community, for dating sites, love—into pieces a server can handle. At the same time you have to retain as much of the je ne sais quoi of the thing as you can, so the users believe what you're offering represents real life. It's a delicate illusion, the Internet; imagine a carrot sliced so cleanly that the pieces stay there in place on the cutting board, still in the shape of a carrot. And while this tension—between the continuity of the human condition and the fracture of the database—can make running a website complicated, it's also what makes my story go. The approximations technology has devised for things like lust and friendship offer a truly novel opportunity: to put hard numbers to some timeless mysteries; to take experiences that we've been content to put aside as "unquantifiable" and instead gain some understanding. As the approximations have gotten better and better, and as people have allowed them further into their lives, that understanding has improved with startling speed. I'm going to give you a quick example, but I first want to say that "Making the Ineffable Totally Effable" really should've been OkCupid's tagline. Alas.

Ratings are everywhere on the Internet. Whether it's Reddit's up/down votes, Amazon's customer reviews, or even Facebook's "like" button, websites ask you to vote because that vote turns something fluid and idiosyncratic—your opinion—into something they can understand and use. Dating sites ask people to rate one another because it lets them transform first impressions such as:

He's got beautiful eyes
Hmmm, he's cute, but I don't like redheads
Ugh, gross

...into simple numbers, say, 5, 3, 1 on a five-star scale. Sites have collected billions of these microjudgments, one person's snap opinion of someone else. Together, all those tiny thoughts form a source of vast insight into how people arrive at opinions of one another.

The most basic thing you can do with person-to-person ratings like this is count them up. Take a census of how many people averaged one star, two stars, and so on, and then compare the tallies. Below, I've done just that with the average votes given to straight women by straight men. This is the shape of the curve:

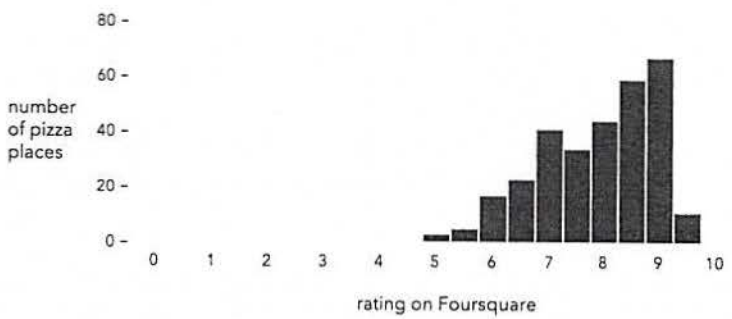


Fifty-one million preferences boil down to this simple stand of rectangles. It is, in essence, the collected male opinion of female beauty on OkCupid. It folds all the tiny stories (what a man thinks of a woman, millions of times over) and all the anecdotes (any one of which we could've expanded upon, were this a different kind of book) into an intelligible whole. Looking at people like this is like looking at Earth from space; you lose the detail, but you get to see something familiar in a totally new way.

So what is this curve telling us? It's easy to take this basic shape—a bell curve—for granted, because examples in textbooks have probably led you to expect it, but the scores could easily have gone hard to one side or the other.

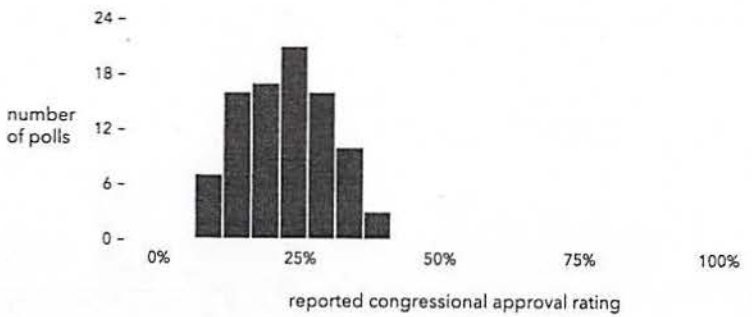
When personal preference is involved, they often do. Take ratings of pizza joints on Foursquare, which tend to be very positive:

user ratings of New York City pizza places on Foursquare's 0-10 scale



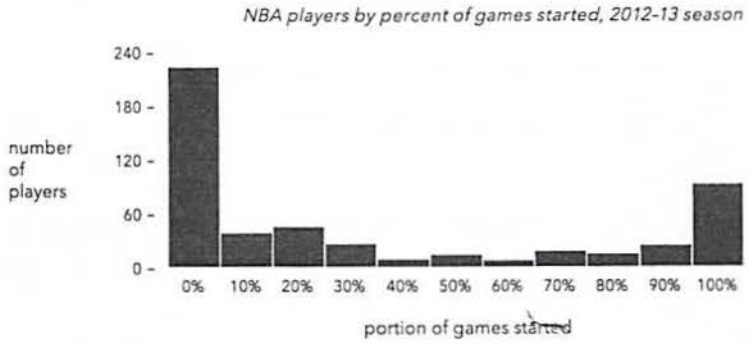
Or take the recent approval ratings for Congress, which, because politicians are the moral opposite of pizza, skew the other way:

congressional popularity in major media polls since November 2008

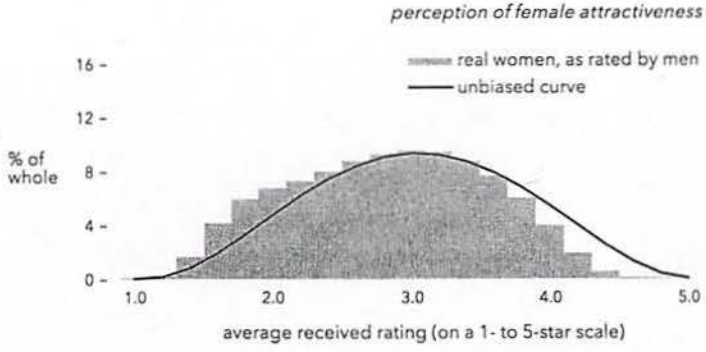


Also, our male-to-female ratings curve is *unimodal*, meaning that the women's scores tend to cluster around a single value. This again is easy to shrug at, but many situations have multiple modes, or "typical" values. If you plot NBA

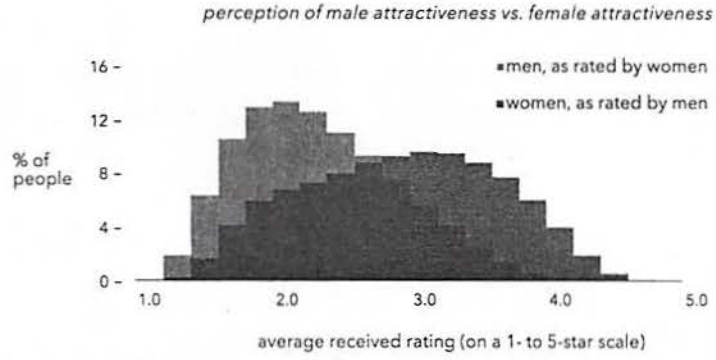
players by how often they were in the starting lineup in the 2012–13 season, you get a bunch of athletes clustered at either end, and almost no one in the middle:



That's the data telling us that coaches think a given player is either good enough to start, or he isn't, and the guy's in or out of the lineup accordingly. There's a clear binary system. Similarly, in our ratings data, men as a group might've seen women as "gorgeous" or "ugly" and left it at that; like top-line basketball talent, beauty could've been a you-have-it-or-you-don't kind of thing. But the curve we started with says something else. Looking for understanding in data is often a matter of considering your results against these kinds of counterfactuals. Sometimes, in the face of an infinity of alternatives, a straightforward result is all the more remarkable for being so. In fact, our graph is quite close to what's called a *symmetric beta distribution*—a curve often deployed to model basic unbiased decisions—which I'll overlay here:



Our real-world data diverges only slightly (6 percent) from this formulaic ideal, meaning this graph of male desire is more or less what we could've guessed in a vacuum: it is, in fact, one of those textbook examples I was making light of. So the curve is predictable, centered—maybe even boring. So what? Well, this is a rare context where boringness is something special: it implies that the individual men who did the scoring are likewise predictable, centered, and, above all, unbiased. And when you consider the supermodels, the porn, the cover girls, the Lara Croft–style fembots, the Bud Light ads, and, most devious of all, the Photoshop jobs that surely these men see every day, the fact that male opinion of female attractiveness is still where it's supposed to be is, by my lights, a small miracle. It's practically common sense that men should have unrealistic expectations of women's looks, and yet here we see it's just not true. In any event, they're far more generous than the women, whose votes go like this:



The red chart is centered barely a quarter of the way up the scale; only one guy in six is "above average" in an absolute sense. Sex appeal isn't something commonly quantified like this, so let me put it in a more familiar context: translate this plot to IQ, and you have a world where the women think 58 percent of men are brain damaged.

Now, the men on OkCupid aren't actually ugly—I tested that by experiment, pitting a random set of our users against a comparable random sample from a social network and got the same scores for both groups—and it turns out you get patterns like the above on every dating site I've seen: Tinder, Match.com,

DateHookup—sites that together cover about half the single people in the United States. It just turns out that men and women perform a different sexual calculus. As *Harper's* put it perfectly: "Women are inclined to regret the sex they had, and men the sex they didn't." You can see exactly how it works in the data. I will add: the men above must be absolutely full of regrets.

A beta curve plots what can be thought of as the outcome of a large number of coin flips—it traces the overlapping probabilities of many independent binary events. Here the male coin is fair, coming up heads (which I'll equate with positive) just about as often as it comes up tails. But in our data we see that the female one is weighted; it turns up heads only once every fourth flip. A large number of natural processes, including the weather, can be modeled with betas, and thanks to some weather bug's obsessive archiving, I was able to compare our person-to-person ratings to historical climate patterns. The male outlook here is very close to the function that predicts cloud cover in New York City. The female psyche, by the same metric, dwells in a place slightly darker than Seattle.

We'll follow this thread through the first of *Dataclysm*'s three broad subjects: the data of people connecting. Sex appeal—how it changes and what creates it—will be our point of departure. We'll see why, technically, a woman is over the hill at twenty-one and the importance of a prominent tattoo, but we'll soon move beyond connections of the flesh. We'll see what tweets can tell us about modern communication, and what friendships on Facebook can say about the stability of a marriage. Profile pictures are both a boon and a curse on the Internet: they turn almost every service (Facebook, job sites, and, of course, dating) into a beauty contest. We'll take a look at what happens when OkCupid removes them for a day and just hopes for the best. Love isn't blind, though we find evidence it should be.

Part 2 then looks at the data of division. We'll begin with a close look at that prime human divide, race—a topic we can now address at the person-to-person level for the first time. Our privileged data exposes attitudes that most people would never cop to in public, and we'll see that racial bias is not only strong but consistent—repeated almost verbatim (well, numeratim), from site to site. Racism can be an interior thing too—just one man, his prejudice, and a keyboard. We'll see what Google Search has to say about the country's most hated word—and what that word has to say about the country. We'll move on to explore the divisiveness of physical beauty with a data set thousands of times more powerful than anything previously available. Ugliness has startling social costs that we are finally able to quantify. From there, we'll see what Twitter reveals about our impulse to anger. The service allows people to stay connected up to the minute; it can drive them apart just as quickly. The collaborative rage that it enables brings a new violence to that most ancient of human gatherings: the mob. We'll see if it can provide a new understanding, as well.

By the book's third section, we will have seen the data of two people interacting, for better and for worse; here we will look at the individual alone. We'll explore how ethnic, sexual, and political identity is expressed, focusing on the words, images, and cultural markers people choose to represent themselves. Here are five of the phrases most typical of a white woman:

my blue eyes
red hair and
four wheeling
country girl
love to be outside

Haiku by Carrie Underwood, or data? You make the call! We'll explore people's public words. We'll also see how people speak and act in private, with an eye toward the places where labels and action diverge: bisexual men, for example, challenge our ideas of neat identity. Next, we'll draw on a wide range of sources—Twitter, Facebook, Reddit, even Craigslist—to see ourselves in our homes, both physically and otherwise. And we'll conclude with the natural question about a book like this: how does a person maintain his privacy in a world where these explorations are possible?

Throughout, we'll see that the Internet can be a vibrant, brutal, loving, forgiving, deceitful, sensual, angry place. And of course it is: it's made of human beings. However, bringing all this information together, I became acutely aware that not everyone's life is captured in the data. If you don't have a computer or a smartphone, then you aren't here. I can only acknowledge the problem, work around it, and wait for it to go away.

I will say in the meantime that the reach of sites like Twitter and Facebook, and even my dating data, is surprisingly thorough. If you don't use many of these services yourself, this is something you might not appreciate. Some 87 percent of the United States is online, and that number holds across virtually all demographic

boundaries. Urban to rural, rich to poor, black to Asian to white to Latino, all are connected. Internet adoption is lower (around 60 percent) among the very old and the undereducated, which is why I drew my "age line" well short of old age in these pages—at fifty—and why I don't address education at all. More than 1 out of every 3 Americans access Facebook *every day*. The site has 1.3 billion accounts worldwide. Given that roughly a quarter of the world is under age fourteen, that means that something like 25 percent of adults on Earth have a Facebook account. The dating sites in *Dataclysm* have registered some 55 million American members in the last three years—as I said above, that's one account for every two single people in the country. Twitter is an especially interesting demographic case. It's a glitzy tech success story, and the company is almost single-handedly gentrifying a large swath of San Francisco. But the service itself is fundamentally populist, both in the "openness" of its platform and in who chooses to use it. For example, there's no significant difference in use by gender. People with only a high school education level tweet as much as college graduates. Latinos use the service as much as whites, and blacks use it twice as much. And then, of course, there's Google. If 87 percent of Americans use the Internet, 87 percent of them have used Google.

These big numbers don't prove I have the complete picture of anything, but they at least suggest that such a picture is coming. And in any event the perfect should not be the enemy of the better-than-ever-before. The data set we'll work with encompasses thousands of times more people than a Gallup or Pew study; that goes without saying. What's less obvious is that it's actually much more inclusive than most academic behavioral research.

It's a known problem with existing behavioral science—though it's seldom discussed publicly—that almost all of its foundational ideas were established on small batches of college kids. When I was a student, I got paid like $25 to inhale a slightly radioactive marker gas for an hour at Mass General and then do some kind of mental task while they took pictures of my brain. It won't hurt you, they said. It's just like spending a year in an airplane, they said. No big deal, they said. What they didn't say—and what I didn't realize then—was that as I was lying there a little hungover in some kind of CAT-scanner thing, reading words and clicking buttons with my foot, I was standing in for the typical human male. My friend did the study, too. He was a white college kid just like me. I'm willing to bet most of the subjects were. That makes us far from typical.

I understand how it happens: in person, getting a real representative data set is often more difficult than the actual experiment you'd like to perform. You're a professor or postdoc who wants to push forward, so you take what's called a "convenience sample"—and that means the students at your university. But it's a big problem, especially when you're researching belief and behavior. It even has a name. It's called WEIRD research: white, educated, industrialized, rich, and democratic. And most published social research papers are WEIRD.*

Several of these problems plague my data, too. It will be a while still before digital data can scratch "industrialized" all the way off the list. But because tech is often seen as such an "elite field"—an image that many in the industry are all too willing to encourage—I feel compelled to distinguish between the entrepreneurs and venture capitalists you see on technology's public stages, making swiping gestures and spouting buzz talk into headset mikes, people who are usually very WEIRD indeed, from the users of the services themselves, who are very much normal. They can't help but be, because use of these services—Twitter, Facebook, Google, and the like—is the norm.

As for the data's authenticity, much of it is, in a sense, fact-checked because the Internet is now such a part of everyday life. Take the data from OkCupid. You give the site your city, your gender, your age, and who you're looking for, and it helps you find someone to meet for coffee or a beer. Your profile is supposed to be you, the true version. If you upload a better-looking person's picture as your own, or pretend to be much younger than you really are, you will probably get more dates. But imagine meeting those dates in person: they're expecting what they saw online. If the real you isn't close, the date is basically over the instant you show up. This is one example of the broad trend: as the online and offline worlds merge, a built-in social pressure keeps many of the Internet's worst fabulist impulses in check.

The people using these services, dating sites, social sites, and news aggregators alike, are all fumbling their way through life, as people always have. Only now they do it on phones and laptops. Almost inadvertently, they've created a unique

* An article in *Slate* noted: "WEIRD subjects, from countries that represent only about 12 percent of the world's population, differ from other populations in moral decision making, reasoning style, fairness, even things like visual perception. This is because a lot of these behaviors and perceptions are based on the environments and contexts in which we grew up."

archive: databases around the world now hold years of yearning, opinion, and chaos. And because it's stored with crystalline precision it can be analyzed not only in the fullness of time, but with a scope and flexibility unimaginable just a decade ago.

I have spent several years gathering and deciphering this data, not only from OkCupid, but from almost every other major site. And yet I've never quite been able to get over a nagging doubt, which, given my Luddite sympathies, pains me all the more: writing a book about the Internet feels a lot like making a very nice drawing about the movies. Why bother? That's the question of my dark hours.

∞

There's this great documentary about Bob Dylan called *Dont Look Back* that I watched a bunch back in college; my best friend, Justin, was studying film. Somewhere in the movie, at an after-party, Bob gets into an argument with a random guy about who did or who did not throw some glass thing in the street. They're both clearly drunk. The climax of the confrontation is this exchange, and it's stuck with me now for fifteen years:

DYLAN: I know a thousand cats who look just like you and talk just like you.
GUY AT PARTY: Oh, fuck off. You're a big noise. You know?
DYLAN: I know it, man. I know I'm a big noise.
GUY AT PARTY: I know you know.
DYLAN: I'm a bigger noise than you, man.
GUY AT PARTY: I'm a small noise.
DYLAN: Right.

And then someone breaks it up so they can all talk poetry. It's that kind of night. But here's the thing: rock star or no, big noises have been the sound of mankind so far. Conquerors, tycoons, martyrs, saviors, even scoundrels (especially scoundrels!)—their lives are how we've told our larger story, how we've marked our progression from the banks of a couple of silty rivers to wherever we are now. From Pharaoh Narmer in BCE 3100, the first living man whose name we still know, to Steve Jobs and Nelson Mandela—the heroic framework is how people order the world. Narmer was first on an ancient list of kings. The scribes have changed, but that list has continued on. I mean, the 1960s, power to the people and so on, is the perfect example: that's the era of Lennon and McCartney, Dylan, Hendrix, not "Guy at Party." Above all, Everyman's existence hasn't been worth recording, apart from where it intersects with a legend's.

But this asymmetry is ending; the small noise, the crackle and hiss of the rest of us, is finally making it to tape. As the Internet has democratized journalism, photography, pornography, charity, comedy, and so many other courses of personal endeavor, it will, I hope, eventually democratize our fundamental narrative. The sound is inchoate now, unrefined. But I'm writing this book to bring out what faint patterns I, and others, detect. This is the echo of the approaching train in ears pressed to the rail. Data science is far from perfect—there's selection bias and many other shortcomings to understand, acknowledge, and work around. But the distance between what could be and what is grows shorter every day, and that final convergence is the day I'm writing to.

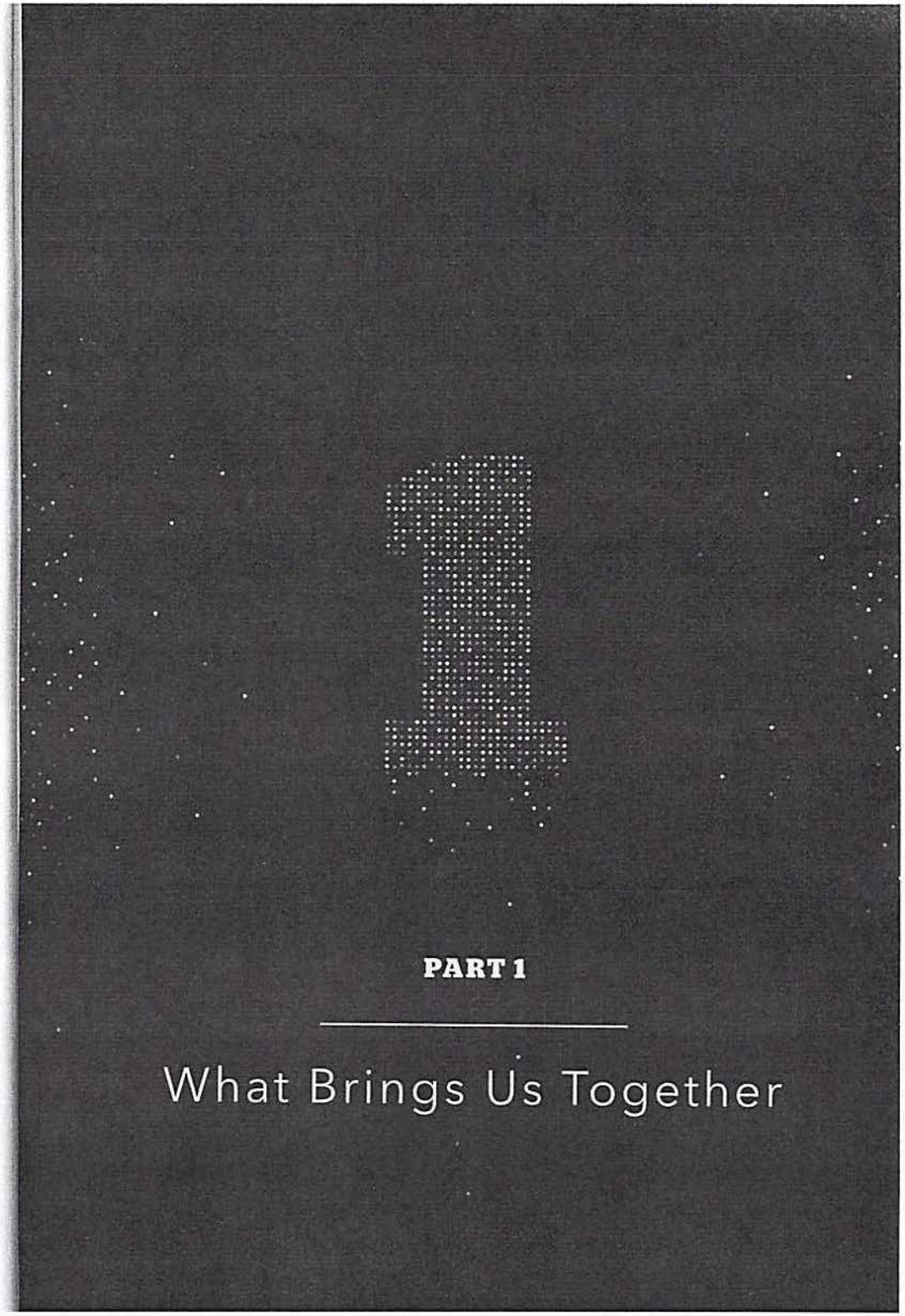
I know there are a lot of people making big claims about data, and I'm not here to say it will change the course of history—certainly not like internal combustion did, or steel—but it will, I believe, change what history is. With data, history can become deeper. It can become more. Unlike clay tablets, unlike papyrus, unlike paper, newsprint, celluloid, or photo stock, disk space is cheap and nearly inexhaustible. On a hard drive, there's room for more than just the heroes. Not being a hero myself, in fact, being someone who would most of all just like to spend time with his friends and family and live life in small ways, this means something to me.

Now, as much as I'd like me and you and WhoBeefed81 to be right there on the page with the president when future works treat this decade, I imagine everyday people will always be more or less nameless, as indeed they are even here. The best data can't change that. But we all will be counted. When in ten years, twenty, a hundred, someone takes the temperature of these times and wants to understand changes—wants to see how legalizing gay marriage both drove and reflected broader acceptance of homosexuality or how village society in Asia was uprooted, then created again, within its large urban centers—inside that story, even comprising its very bones, will be data from Facebook, Twitter, Reddit, and the like. And if not, our putative writer will have failed.

I've tried to capture all this with my mash-up title. *Kataklysmos* is Greek for the Old Testament Flood; that's how the word "cataclysm" came to English.

The allusion has dual resonance: there is, of course, the data as unprecedented deluge. What's being collected today is so deep it verges on bottomless; it's easily forty days and forty nights of downpour to that old handful of rain. But there's also the hope of a world transformed—of both yesterday's stunted understanding and today's limited vision gone with the flood.

This book is a series of vignettes, tiny windows looking in on our lives—what brings us together, what pulls us apart, what makes us who we are. As the data keeps coming, the windows will get bigger, but there's plenty to see right now, and the first glimpse is always the most thrilling. So to the sills, I'll boost you up.



PART 1

What Brings Us Together

2.

Death by a Thousand Mehs

In 2002, the Oscars hired the director Errol Morris to shoot a short film about why we love the movies. The Academy wanted to kick off the telecast with a rapid-fire montage of people, both celebrities and not, talking about their favorite films. My friend Justin was Morris's casting director, so he got me on the list. There was no guarantee that I'd end up in the final cut of the short, but I could do the interview on-camera and see how it went.

Having an in, I got scheduled the same day as the biggest names: Donald Trump, Walter Cronkite, Iggy Pop, Al Sharpton, Mikhail Gorbachev. Trump and Gorbachev were back to back, and somewhere out there there's a picture of the two of them, with me in the middle, photobombing before photobombing was a thing. I say "somewhere" because right after the flash, Trump snapped his fingers, and his bodyguard took Justin's camera. For his favorite movie, Trump picked *King Kong*, because he of course likes apes who try to "conquer New York." Gorbachev, through a translator whose mustache must've weighed ten pounds, chose *Gladiator*. At 2:01 in Morris's film, the wide eyes and the voice saying "*The Omen*" are mine.

Now, I like a good Antichrist movie more than most people, but I chose *The Omen* more or less at random. There are so many good movies, I'm actually not sure what my favorite one is. But I know my least favorite film with absolute certainty. *Pecker*, by John Waters. I walked out of it. Twice. I went once with some friends, couldn't deal with the mondo-trasho vibe, not to mention the exaggerated accents, and just had to leave. The next weekend, some *other* friends were going and since John Waters is a respected auteur, and hey I'm a cool guy who gets it, I figured there was at least some chance I was wrong the first time. Also I had nothing else to do. So I went again.

Such is the temporary madness of being twenty-two. I'm not saying John Waters makes objectively bad movies—they're just not for me. Or for a lot of people. And he embraces that fact, the rejection—it's practically his calling card as a director. Let me put it this way: nobody leaves *Pecker* thinking it was "meh";

either you loved it, or got the hell out after twenty minutes like I did, twice. That's by design.*

Waters's fans seem to love him all the more for being fewer in number. On OkCupid, a search through users' profile text returns more results for his name than George Lucas's and Steven Spielberg's combined. On Reddit, he has his own devoted page: */r/JohnWaters,*† and while it's not the most-trafficked URL ever, people actually put stuff there: news, old clips, questions about him, comments, and so on. There's a */r/GeorgeLucas*, too: it has one post, ever. If you enter */r/StevenSpielberg* into your address bar, you get "there doesn't seem to be anything here" from Reddit's server because, as good as his work is, no one's been enthusiastic enough to make a page. Even highly Internet-friendly directors like J. J. Abrams don't have their own page. It takes a certain special motivation to, say, make a fan site, and that motivation is often intensified by feeling like you're part of a special, embattled elect. Devotion is like vapor in a piston—pressure helps it catch.

Like many artists before and since, Waters understands exactly how it works: repelling some people draws others all the closer, and I bring him up not only because of my lifelong personal struggle with *Pecker*, but because Waters also gets the universality of the principle: it's not just true for art. He's got a lot of great quotes, but here's one that speaks right to me: "Beauty is looks you can never forget. A face should jolt, not soothe." He's completely correct, for as with music, as with movies, and as with a wide variety of human phenomena: a flaw is a powerful thing. Even at the person-to-person level, to be universally liked is to be relatively ignored. To be disliked by some is to be loved all the more by others. And, specifically, a woman's overall sex appeal is enhanced when some men find her ugly.

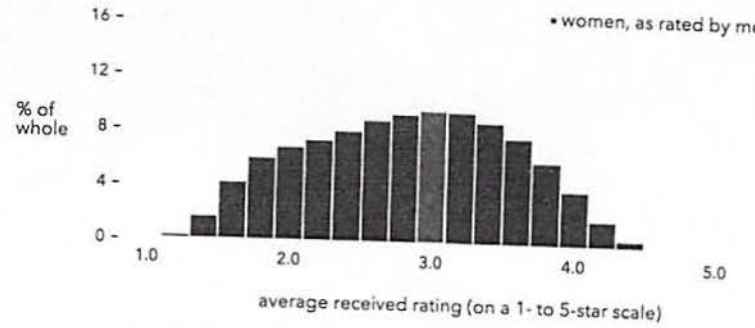
You can see this in the profile ratings on OkCupid. Because the site's rating system is 5 stars, the votes have more depth than just a *yes* or a *no*. People give degrees of opinion, and that gives us room to explore. To show this finding, we'll have to go on a short mathematical journey. These kinds of exercises are what make data science work. To put together puzzles, you have to lay out all the

* Waters on film: "To me, bad taste is what entertainment is all about. If someone vomits while watching one of my films, it's like getting a standing ovation."

† These pages on Reddit are called subreddits. I'll explain the site and its nuances in more detail later.

pieces and then just start trying things. In the absence of careful sifting, reduction, and parsimony, very little just "jumps out at you" from terabytes of raw data.

Consider a group of women with approximately the same attractiveness, let's just say the ones rated in the middle:



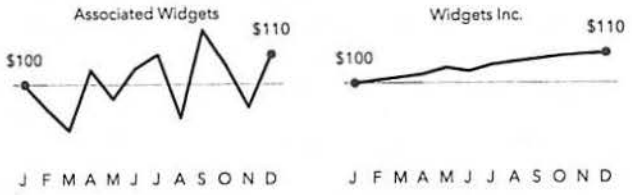
Now imagine a woman in that group and think of the many different votes men could've given her—basically think about how she ended up in the middle. There are thousands of possibilities; here are just a few I made up, combinations of 1s, 2s, 3s, 4s, and 5s, which all come to an average of 3:

	number of men who voted... "1"	"2"	"3"	"4"	"5"	pattern avg.
pattern A			100			3.0
pattern B		10	80	10		3.0
pattern C	10	20	40	20	10	3.0
pattern D	25	25		25	25	3.0
pattern E	50				50	3.0

As you might've noticed, the vote patterns I've chosen get more polarized as they go from Pattern A to Pattern E. Each row still averages out to that same central "3," but they express that average in different ways. Pattern A is the

embodiment of consensus. There, the men who cast the votes have spoken in perfect unison: *this woman is exactly in the middle*. But by the time we get to the bottom of the table, the overall average is still centered, yet no single individual actually holds that central opinion. Pattern E shows the most extreme possible path to a middling average: for every man awarding our theoretical woman a "1," someone else gives her a "5," and the total result comes out to a "3" almost in spite of itself. That's the John Waters way.

These patterns exemplify a mathematical concept called *variance*. It's a measure of how widely data is scattered around a central value. Variance goes up the further the data points fall from the average; in the table above, it is highest in Pattern E. One of the most common applications of variance is to weigh volatility (and therefore risk) in financial markets. Consider these two companies:



Both returned 10 percent for the year, but they are very different investments. Associated Widgets experienced large swings in value throughout the year, while Widgets Inc. grew little by little, showing consistent gains each month. Computing the variance allows analysts to capture this distinction in one simple number, and all other things being equal, investors much prefer the low score of that pattern on the right. Same return, fewer heart palpitations. Of course, when it comes to romance, heart palpitations *are* the return, and that gets to the crux of it. It turns out that variance has almost as much to do with the sexual attention a woman gets as her overall attractiveness.

In any group of women who are all equally good-looking, the number of messages they get is highly correlated to the variance: from the pageant queens to the most homely women to the people right in between, the individuals who get the most affection will be the polarizing ones. And the effect isn't small—being highly polarizing will in fact get you about 70 percent more messages. That means variance allows you to effectively jump several "leagues" up in the dating pecking order—for example, a very low-rated woman (20th percentile) with high variance in her votes gets hit on about as much as a typical woman in the 70th percentile.

Part of that is because variance means, by definition, that more people like you *a lot* (as well as dislike you a lot). And those enthusiastic guys—let's just call them the fanboys—are the ones who do most of the messaging. So by pushing people toward the high end (the 5s), you get more action.

But the negative votes themselves are part of the story, too. They drive some of the attention on their own. For example, the real patterns exemplified by C and D below get about 10 percent more messages than the ones shown in A and B, even though the top two women are rated far better overall:

	number of men who voted...					
	"1"	"2"	"3"	"4"	"5"	pattern avg.
woman A	2	22	27	29	20	3.4
woman B	10	13	31	28	18	3.3
woman C	32	22	12	16	18	2.7
woman D	47	13	6	19	15	2.4

I've been talking about messages as if they're an end unto themselves, but on a dating site, messages are the precursor to outcomes like in-depth conversations, the exchange of contact information, and eventually in-person meetings. People with higher variance get more of all these things, too. So, for example, woman D above would have about 10 percent more conversations, 10 percent more dates, and, likely, 10 percent more sex than woman A, even though in terms of her absolute rating she's much less attractive.

Moreover, the men giving out those 1s and 2s are not themselves hitting on the women—people practically never contact someone they've rated poorly.* It's that having haters somehow induces everyone else to want you more. People *not*

* Only 0.2 percent of the messages on the site are sent by users to a person to whom they awarded fewer than 3 stars.

liking you somehow brings you more attention entirely on its own. And, yes, in his underground castle, Karl Rove smiles knowingly, petting an enormous toad.

It only adds to the mystery of the phenomenon that OkCupid doesn't publish raw attractiveness scores (or a variance number, of course) for anyone on the site. Nobody is consciously making decisions based on this data. But people have a way of feeling the math behind things, whether they're aware of it or not, and here's what I think is going on. Suppose a guy is attracted to a woman he knows is unconventional-looking. Her very unconventionality implies that some other men are likely turned off; it means less competition. Having fewer rivals increases his chances of success. I can imagine our man browsing her profile, circling his cursor, thinking to himself: *I bet she doesn't meet many guys who think she's awesome. In fact, I'm actually into her for her quirks, not in spite of them. This is my diamond in the rough*, and so on. To some degree, her very unpopularity is what makes her attractive to him. And if our browsing guy was at all on the fence about whether to actually introduce himself, this might make the difference.

Looking at the phenomenon from the opposite angle—the low-variance side—a relatively attractive woman with consistent scores is someone any guy would consider conventionally pretty. And she therefore might seem to be more popular than she really is. Broad appeal gives the impression that other guys are after her, too, and that makes her incrementally less appealing. Our interested but on-the-fence guy moves on.

This is my theory at least. But the idea that variance is a positive thing is fairly well established in other arenas. Social psychologists call it the "pratfall effect"—as long as you're generally competent, making a small, occasional mistake makes people think you're *more* competent. Flaws call out the good stuff all the more. This need for imperfection might just be how our brains are put together. Our sense of smell, which is the most connected to the brain's emotional center, prefers discord to unison. Scientists have shown this in labs, by mixing foul odors with pleasant ones, but nature, in the wisdom of evolutionary time, realized it long before. The pleasant scent given off by many flowers, like orange blossoms and jasmine, contains a significant fraction (about 3 percent) of a protein called indole. It's common in the large intestine, and on its own, it smells accordingly. But the flowers don't smell as good without it. A little bit of shit brings the bees. Indole is also an ingredient in synthetic human perfumes.

You can see a public implementation, as it were, of the OkCupid data in the rarefied world of modeling. The women are all professionally gorgeous—5 stars out of 5, of course. But even at that high level it's still about distinguishing yourself through imperfection. Cindy Crawford's career took off after she stopped covering her mole. Linda Evangelista had the severe hair—you can't say it made her *prettier*, but it did make her far more interesting. Kate Upton, at least according to the industry standard, has a few extra pounds. Pulling a few examples from the data set, perhaps ones that are more relatable than swimsuit models, will help you see how it works for a normal person. Here are six women, all with middle-of-the-road overall scores, but who tend to get extreme reactions either way: lots of Yes, lots of No, but very little Meh:

Thanks to each of them for having the confidence to agree to be displayed and discussed here. What you see in the array is what you get throughout the corpus. These are people who've purposefully abandoned the middle road: with body art, a snarky expression, or by eating a grilled cheese like a badass. And you find many relatively normal women with an unusual trait: like the center woman in the bottom row, whose blue hair you can't see in black and white. And you

especially see women who've chosen to play up their particular asset/liability. If you can pull off, say, a 3.3 rating despite the extra pounds or the people who hate tattoos or whatever, then, literally, more power to you.

So at the end of it, given that everyone on Earth has some kind of flaw, the real moral here is: be yourself and be brave about it. Certainly trying to fit in, just for its own sake, is counterproductive. I know this is dangerously close to the kind of thing that gets put on a quilt, and quilts, being the PowerPoint presentations of an earlier time, are the opposite of science. It also sounds a lot like the advice a mother gives, along with a pat on the head, to her big-nosed and brace-faced son when he's fourteen and can't figure out why he isn't more popular. But either way, there it is, in the numbers. Like I said, people can feel the math behind things, especially, thankfully, moms. I just wish she'd told me that by ninth grade bears aren't cool.

5.

There's No Success Like Failure

There's a great Tumblr called "Clients from Hell," where anyone can submit their service-industry horror stories. There are all kinds of cluelessness and oblivion on display, and new posts go up every few hours. Here's a typical submission, from someone doing a photo spread:

CLIENT: Can we have a heading on the photo as well?

DESIGNER: Well, it already has a caption.

CLIENT: If the reader misses the caption, then they will still see the heading.

DESIGNER: It would be quite unusual to have both a heading and a caption on a photo.

CLIENT: That makes sense. Just put a heading next to the caption, then.

My favorite client quote on the site right now is: "I don't like the dinosaur in this graphic. It looks too fake. Use a real photo of a dinosaur instead." The blog mostly gets submissions from graphic designers, but Clients from Hell's popularity speaks to a universal truth. People hate their customers.

I don't mean hate on an individual level but, en masse, customers, like any rabble, are to be feared. Anyone who tells you otherwise, from the cupcake-shop owner down the street to the CEO in the boardroom, is lying. Part of it is the "...is always right" thing—nobody likes a person with that much power. But by far the biggest cause of frustration is that people don't understand and can't articulate what they actually need. As Steve Jobs said, "People don't know what they want until you show it to them." What he didn't say is that showing them, especially in tech, means playing a game of Pin the Tail on the Donkey with several million people shouting advice.

If you are, say, a car company and people don't like some part of your product, they mostly tell you indirectly, by not buying it. There's historically been no open channel between Ford and the folks who want the cup holders to be green or who think it would be better if the steering wheel were a square, because, you know, most turns are 90 degrees. That's why traditional companies spend so much on market research—they have to stay way ahead of these kinds of things, because by the time a company like Ford would naturally hear about a problem, via Accounts Receivable, it's way too late.

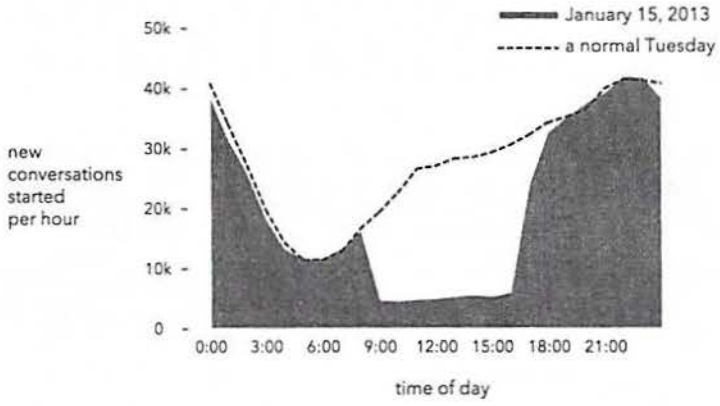
A website is different: if people have a cockamamie idea, someone at the com-

pany is just an e-mail away. And if people don't use something, the site notices immediately. Measurements are tracked in real time, down to the finest grain, everywhere. Whenever you see something new on your favorite site—Google, Facebook, LinkedIn, YouTube, or anywhere—and you click it, know that someone, probably wearing headphones and eating Doritos, just saw a little counter go up by 1. That's when the richness of data can drive a person crazy: one of Google's best designers, the person who in fact built their visual design team, Douglas Bowman, eventually quit because the process had become too microscopic. For one button, the company couldn't decide between two shades of blue, so they launched all forty-one shades in between to see which performed better. *Know thyself:* It was etched into a footstone of the Temple of Apollo at Delphi. But like the rest of the best wisdom that time has to offer, it goes right out the window as soon as anyone turns on a computer.

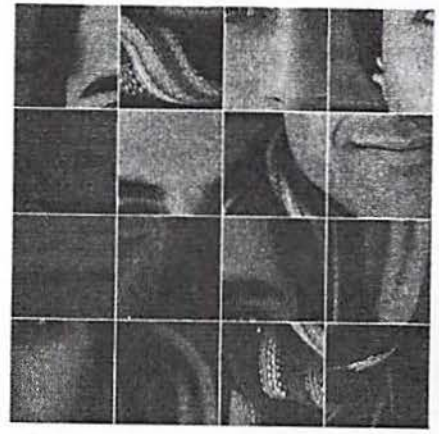
Not knowing what customers need from a car, or even from a particular website interface—those are matters for a business school or a design workshop. It's when people don't understand their own hearts that I get interested. People saying one thing and doing another is pretty much par for the course in social science, but I had a rare opportunity to see people *acting* in two contradictory ways. And it all happened because *I* didn't know what they wanted either.

∞

On January 15, 2013, OkCupid declared "Love Is Blind Day" and removed everyone's profile photos from the site for a few hours. The idea was to do something different and get a little attention for a new service we were launching at the same time. The programmers "flipped the switch" at nine a.m.:



It was a bona fide pit of despair—rare in the wild! The new service OkCupid was trying to promote was a mobile app called Crazy Blind Date. With a couple taps on the screen, it would pair you with a person and select a place nearby and a time in the near future for the two of you to meet. The app provided an interface to let both parties confirm, but there was no way for anyone to directly communicate before the date. The only information it gave you about the other person was a first name and a scrambled thumbnail, like the one below. You were just supposed to show up and hope for the best.



a CBD-style scramble of a stock photo

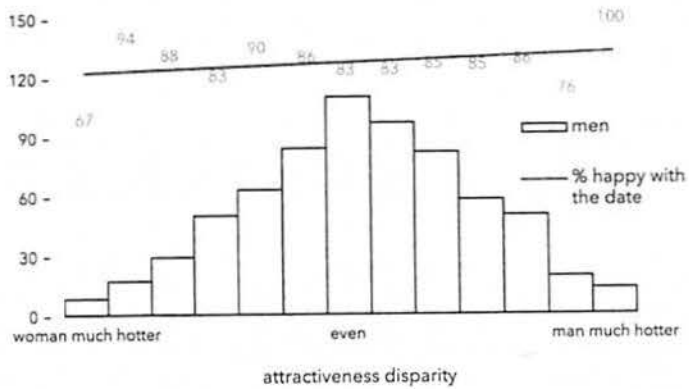
You've probably already noticed that I'm speaking of Crazy Blind Date in the past tense. Even after a quarter million downloads, it failed, because in the end people insist on seeing what they're getting into. The app was one of those ideas that looks great on a whiteboard and miserable in the full color of creation—it was like one long "Love Is Blind Day," and with no way to flip the switch back to normal. A few months after launch, we shut the service down, but before Crazy Blind Date went off to the great app store in the sky (little-known fact: there are no bugs in heaven, just sweet features), about 10,000 people used it to share a beer or a cup of coffee with someone they'd never seen or spoken to before.

From these intrepid few, the app bequeathed the world a rare data set. Crazy Blind Date recorded not only the fact that dater A and dater B met in person but also their opinions of each other. After each completed date, like a nosy roommate, the app asked how it went. Because most of the users also had OkCupid accounts, we were able to cross-reference this data with all kinds of demographic details. We suddenly had in-person records to combine with our massive collection of digital interactions. When you merge the two sources you find something remarkable: the two people's looks had almost

no effect on whether they had a good time. No matter which person was better-looking or by how much—even in cases where one blind-dater was a knockout and the other rather homely—the percent of people giving the dates a positive rating was constant. Attractiveness didn't matter. This data, from real dates, turned everything I'd seen in ten years of running a dating site on its head.

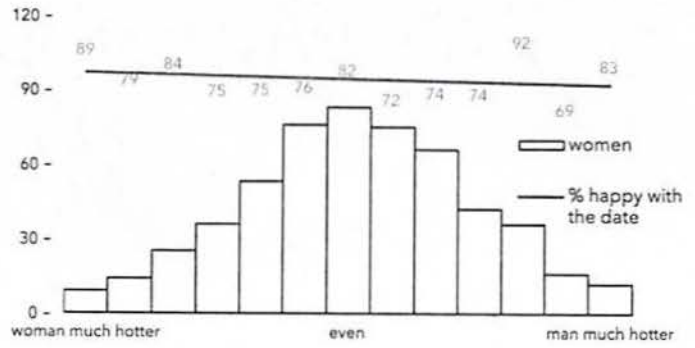
Here are the numbers for men. I've expressed attractiveness below as the *relative difference* in a couple's individual ratings, rather than as absolutes. I did this to capture the fact that a person's happiness at finding himself across the table from, say, a "6" is highly dependent on his own looks. If he's a "1," he might be thrilled with that arrangement—it means he's dating up. A "10" would feel differently. I've included the counts of dates as the bars to show that the balance in attractiveness between the men and women going on the dates was about what you'd expect if they were randomly paired. There was no evidence of people gaming the system by, say, somehow unscrambling the pictures beforehand or showing up to the date venue and then leaving on the sly when their blind date arrived and didn't pass muster. The satisfaction numbers (for males) are the percentages in red:

how attractiveness affects male date satisfaction



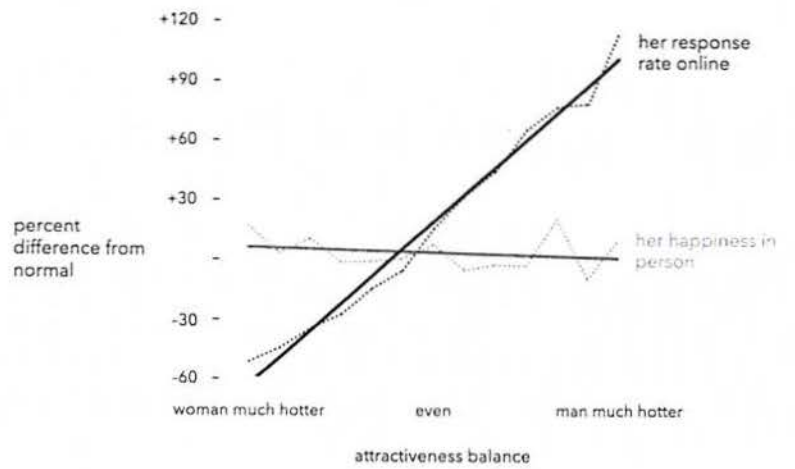
And following is the same data for women:

how attractiveness affects female date satisfaction



Through both Crazy Blind Date data sets, people just didn't seem to care that much about the other person's physical appearance. Women had a good time 75 percent of the time, men 85 percent. The rest of the variation is basically noise. That indifference to looks is just about the opposite of what you see in the OkCupid data. For example, I've plotted the in-person satisfaction data above (the numbers in red) alongside those same women's *reply rates* to messages online. To make it easier to compare them, the lines show change against the average of their respective quantities:

female response to male attractiveness



The male comparison chart is very similar to this one, and, to be clear, the data underpinning the two lines above is from the *same set of people*. The black line is their OkCupid experience, the red from Crazy Blind Date. In short, people appear to be heavily preselecting online for something that, once they sit down in person, doesn't seem important to them.

That kind of superficial preselection is everywhere. In fact, there's a lot of money to be made off it. You know what the difference between Tylenol and Kroger's store-brand acetaminophen is? The box. Unless you take medicine like a king snake and plan to just swallow the package whole, there's really no reason to pay twice as much for the "name" molecules, whose properties are determined by immutable chemical law. And yet, I have a big red Tylenol bottle on my dresser.

We of course pay the most attention to labels when they're attached to people. In terms of superficial compatibility, self-described Democrats and Republicans get along the least of all major groups on OkCupid—worse even than Protestants and Atheists. I know this through the many match questions the site asks: they cover pretty much everything, and the average user answers about three hundred of them. The site lets you decide the importance of each question you answer, and you can pinpoint the answers that you would (and would not) accept from a potential match. Despite all this control, in the political case, the system breaks down. When you look beyond the labels, at who actually messages whom, and who replies (and therefore who ends up going on actual dates), it's *caring* about politics, one way or the other, that is actually more important to mutual compatibility than the details of any particular belief. We confirmed this in a summer-long experiment in 2011.

People tend to run wild with those match questions, marking all kinds of stuff as "mandatory," in essence putting a checklist to the world: I'm looking for a dog-loving, agnostic, nonsmoking liberal who's never had kids—and who's good in bed, of course. But very humble questions like *Do you like scary movies?* and *Have you ever traveled alone to another country?* have amazing predictive power. If you're ever stumped on what to ask someone on a first date, try those. In about three-quarters of the long-term couples OkCupid has ever brought together, both people have answered them the same way, either both "yes" or both "no." People tend to overemphasize the big, splashy things: faith, politics, and certainly

looks, but they don't matter nearly as much as everyone thinks. Sometimes they don't matter at all.

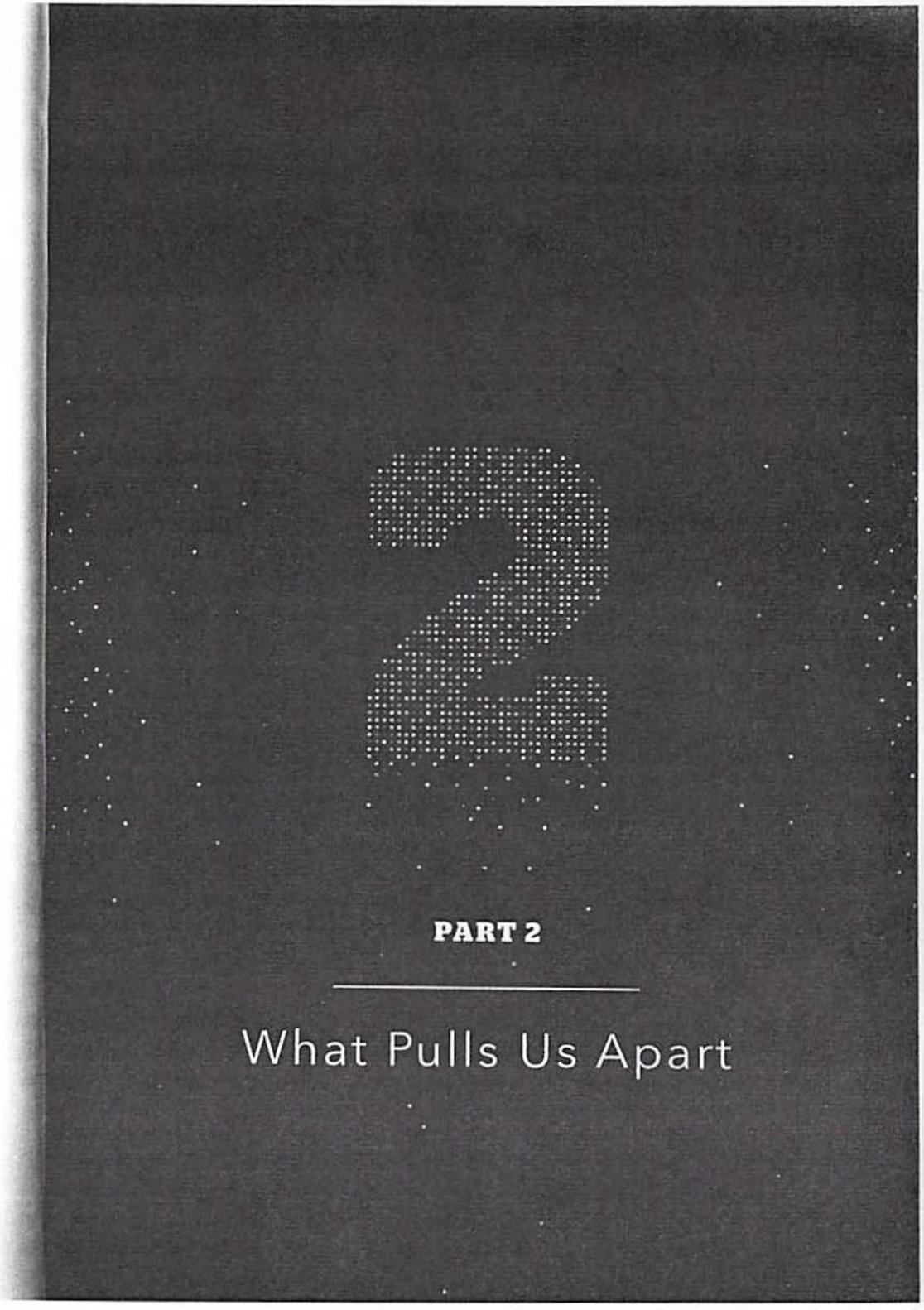
Fiasco though it was, Love Is Blind Day gave us a visceral example of what people do in the absence of information. In hiding pictures but changing nothing else, we created a real-time experiment to set against the site's usual activity. For seven hours our users acted without the very thing our previous data had indicated was the single most important piece of knowledge OkCupid could offer: what everyone else looked like.

Some of the upshot was predictable. People sent messages without the typical biases, or racial and attractiveness skews. What a user couldn't see, he couldn't judge. But of the 30,333 messages sent blindly, eventually 8,912 got replies, a rate about 40 percent higher than usual. And in the dark, for those who were there, something astounding happened. Twenty-four percent of the pairs of people talking when the photos were hidden had exchanged contact info before pictures were turned back on. That was in only the seven-hour window of Love Is Blind Day. The expected number in that amount of time is barely half that. So not only were people writing messages that were far more likely to get replies, they were giving out phone numbers and e-mail addresses at a higher rate—to people they'd never even seen.

For the couples who began talking and were still getting to know each other when we restored photos at four p.m., however, the day had a reverse effect. The two people had been in the dark, then suddenly the lights came on, and, in the data, you can actually see them spook. Threads straddling the moment we flipped the switch lasted an average of 4.4 more messages. When you compare them against a control data set, they should've lasted 5.6. Eventual contact-info exchanges in those "lights on" threads were down by a similar amount.

Dating sites are designed to give people the tools and the information to get whatever they want out of being single—casual sex, a few fun dates, a partner, a marriage . . . anything. Stuff like height, political views, photos, essays, all of it is right there, easily sortable, easily searchable. It's there to help people make judgments and fulfill their desires, and as fascinating as those judgments and desires may be to pick apart, there's a side of it that I think does love a disservice. People make choices from the information we provide because they *can*, not because they necessarily should.

I can't help think of the many people getting turned down because of some perceived "deal-breaker" that actually no one cares about and wonder if the Internet has changed romance in the way it's changed so much else—and for the same reason. If I may channel my inner anti-Jagger: Online, you *can* always get what you want. But what you need, that's a much harder thing to find.



PART 2

What Pulls Us Apart

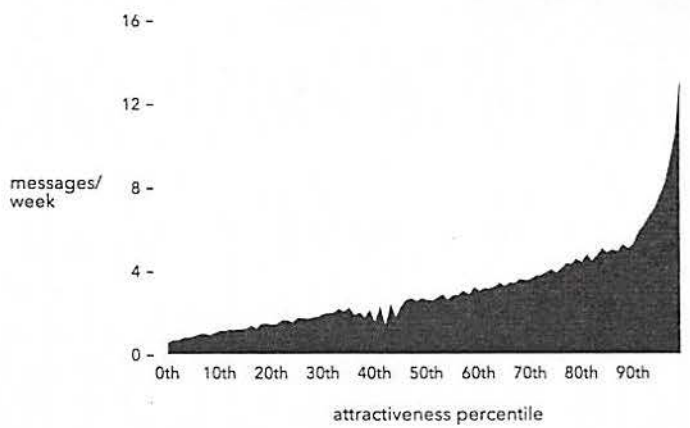
7.

The Beauty Myth in Apotheosis

I work in a universe where people identify themselves along almost every conceivable axis—as smokers and non-; as Christians and atheists; as nerds or geeks, or maybe dorks; to say nothing of black or white or Asian or gay or straight, or neither, or both. Mankind is tribes within tribes. Or, putting it more beautifully, like the Korean proverb: "Over the mountains, mountains." That's the ruggedness of their peninsula and the endless difficulty of our fractured human terrain.

Running a dating site you become aware of a subdivision that on the one hand seems frivolous but on the other is as inborn as a person's race or sexuality, and like those latter traits it's often resistant to direct analysis. On OkCupid—as on Match, as on Tinder—a prime divide, perhaps the deepest, is between the beautiful and the rest. These are our haves and have-nots, our rich, our poor, and when it comes to sexual attention, the haves reap the benefit of their inheritance just as surely as any heir, while the have-nots largely go without. Not unlike race, beauty is a card you're dealt, and it has huge repercussions.

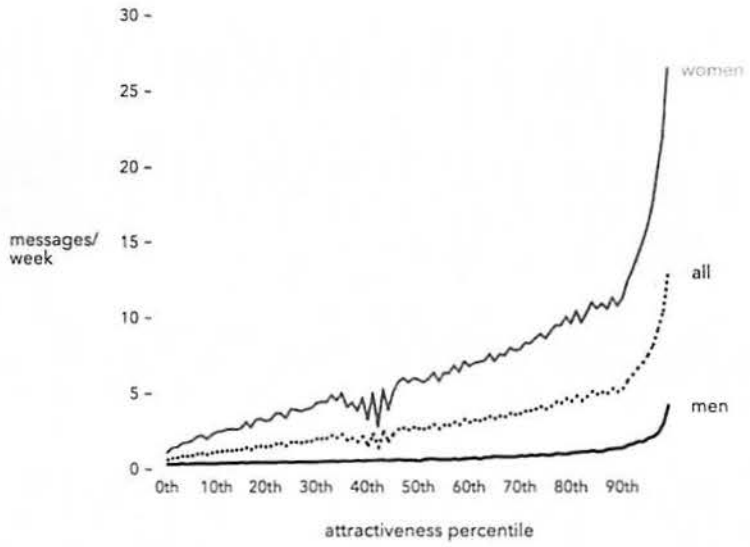
Below I've plotted new messages received per week, by the recipient's physical attractiveness:



The sharp rise out at the right smashes down the rest of the curve, so its true nature is a bit obscured, but from the lowest percentile up, this is roughly an

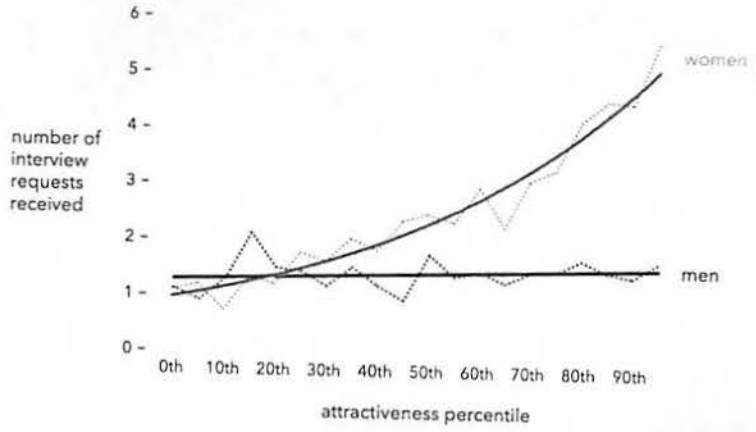
exponential function. That is, it obeys the same math seismologists use to measure the energy released by earthquakes: beauty operates on a Richter scale. In terms of its effect, there is little noticeable difference between, say, a 1.0 and 2.0—these cause tremors that vary only in degree of imperceptibility. But at the high end, a small difference has cataclysmic impact. A 9.0 is intense, but a 10.0 can rupture the world. Or launch a thousand ships.

What you definitely can't see in the chart above, because I aggregated the data to obscure it, is that men and women experience beauty unequally. Here is that OkCupid message density, split out by gender, with the aggregates as the dotted line in the middle.

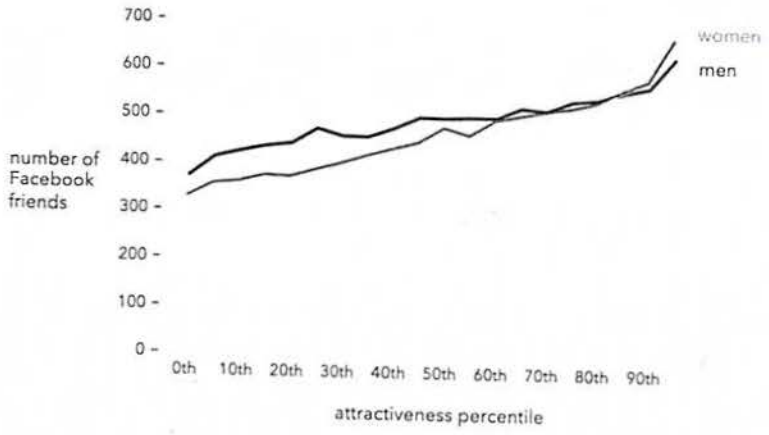


It's hard for me to convey how much attention the upper-right corner of this curve entails, short of tracking you down and screaming in your face about my hobbies. Especially in larger cities, where the message flow is 50 percent higher than even what you see above, a woman at the top of the scale has something like a term paper's worth of hey-what's-up-do-you-like-motorcycles-because-I-like-motorcycles waiting for her every time she comes to the site. A dudeclysm, if you will. However, neither beauty's effects, nor the male/female split, are confined to the sexual realm.

Here is data for interview requests on Shiftgig, a job-search site for hourly and service workers:*



And for friend counts on Facebook:



* I foreground trend lines here because the data is slightly sparser and therefore more noisy than usual. This sample is ≈5,000 people.

Success and beauty are correlated for both sexes, but you can see that the slope of the red line is always steeper. On Facebook, every percentile of attractiveness gives a man two new friends. It gives a woman three. On Shiftgig, the curves aren't even comparable in this way. The female curve is exponential and the male is linear. Moreover, they hold whether the *hiring manager*, the person doing the interviewing, is a man or a woman. In either case, the male candidates' curves are a flat line—a man's looks have no effect on his prospects—and the female graphs are exponential. So these women are treated as if they're on OkCupid, even though they're applying for a job. Male HR reps weigh the female applicants' beauty as they would in a romantic setting—which is either depressing or very, very exciting, depending on whether you're a lawyer with a litigation practice. And female employers view it through the same (seemingly sexualized) lens, despite there (typically) being no romantic intent.

It is hardly fresh intellectual ground that beauty matters, and that it matters more for women. For example, a foundational paper of social psychology is called "What Is Beautiful Is Good." It was the first in a now long line of research to establish that good-looking people are seen as more intelligent, more competent, and more trustworthy than the rest of us. More attractive people get better jobs. They are also acquitted more often in court, and, failing that, they get lighter sentences. As Robert Sapolsky notes in the *Wall Street Journal*, two Duke neuropsychologists are working on why: "The medial orbitofrontal cortex of the brain is involved in rating both the beauty of a face and the goodness of a behavior, and the level of activity in that region during one of those tasks predicts the level during the other. In other words, the brain . . . assumes that cheekbones tell you something about minds and hearts." On a neurological level, the brain registers that ping of sexual attraction—*Ooh, she's hot*—and everything else seems to be splash damage.

To my second point, that beauty affects women in particular, Naomi Wolf's bestseller *The Beauty Myth* showed that better than I ever could. In short, my raw findings here are not new. What is new is our ability to test ideas, established ones, famous ones even, against the atomized actions of millions. That granularity gives strength and nuance to previous work, and even suggests ways to build on it.

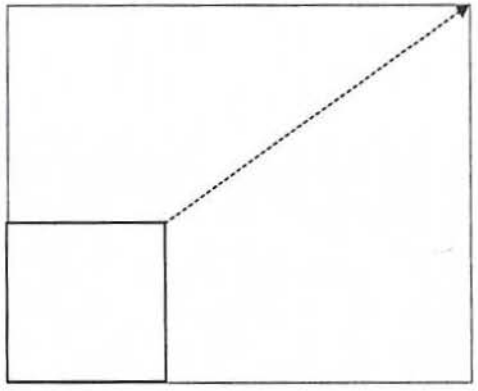
The paper "What Is Beautiful" was based on a research sample of only 60 subjects—barely adequate to prove the effect, let alone its many facets.* But now we can go from "What Is Beautiful Is Good" to asking "How Good?" and in what contexts. In sex, beauty is very good. In friendship, it's only somewhat good, and when you're looking for a job, the effect really depends on your gender. As for Wolf's seminal work, we can confirm the truth behind her broad observation that "today's woman has become her 'beauty'"—three robust research sets agree that the correlation is strong. And, better, we can extend some of her most cogent arguments about beauty being a means of social control. Think about how the Shiftgig data changes our understanding of women's perceived workplace performance. They are evidently being sought out (and exponentially so) for a trait that has nothing to do with their ability to do a job well. Meanwhile, men have no such selection imposed. It is therefore simple probability that women's failure rate, as a whole, will be higher. And, crucially, the criteria are to blame, not the people. Imagine if men, no matter the job, were hired for their physical strength. You would, *by design*, end up with strong men facing challenges that strength has nothing to do with. In the same way, to hire women based on their looks is to (statistically) guarantee poor performance. It's either that or you limit their opportunities. Thus Ms. Wolf: "The beauty myth is always actually prescribing behavior and not appearance." She was speaking primarily in a sexual context, but here, we see how it plays out, with mathematical equivalence, in the workplace.

As I've mentioned before, I have a young daughter, and in our rare downtime, Reshma and I will speculate about her and her life and where it might lead. All parents do this—give them a quiet moment and it's inevitable, just like two drunks in a bar will always argue. Every family must have their own particular flights of fancy, but ours go more or less like most, I imagine. My wife or I will start, it doesn't really matter who: Our little girl's going to be so smart. Oh yes, we'll teach her everything we can. She'll be so gentle, so good-hearted. These

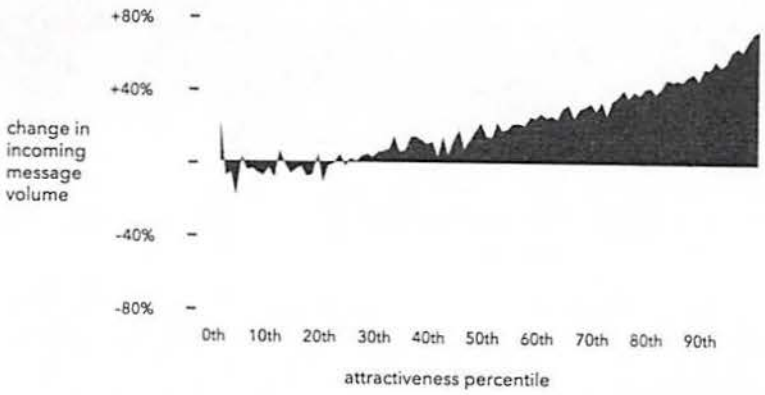
* The study of beauty by traditional methods is especially susceptible to the problem of insufficiency. If your research topic is, say, wealth, you can very easily get a measure of someone's net worth or income and then move on to the dependent trait you want to look at. But to study beauty, first you have to determine how good-looking your subjects are, which is a resource-intensive process. Beauty being so wildly subjective (as opposed to, say, hair color, where if you crowdsourced it, you might get slight variations—*brown, brunette, chestnut*—that are essentially synonymous), you get wide swings in opinion that can only be absorbed by sampling a large, diverse research set. As we've seen with WEIRDness earlier, that has not been a strength of past academic research.

things are very important to a good life, we agree. And of course, look at that skin, like chai, those eyes, she'll be so pretty. I mean, wow. Yeah, we'll have to put locks on the doors when she's a teenager. And there the conversation takes a little turn. But not too pretty, right? Yeah, we wouldn't want that. We both sit back, and the conversation moves on to something else. This is what it comes down to: I can't imagine anyone wishing limits on a son.

Unfortunately, it's a problem the Internet is surely making worse: for *The Beauty Myth*, social media signals Judgment Day. Your picture is attached to practically everything, certainly every résumé, every application, every byline. If people care about what you are doing, they will find out what you look like. Not because they should, but because they can—Facebook and LinkedIn have essentially extended OkCupid's Love Is Blind problem to everything. Even just ten years ago, it was almost impossible to tie the average person's name to her photograph; now you just Google the words—everyone does—and up pops a thumbnail from a social network. We've all had to pick through snapshots for that "best" one. Choose wisely, friends, because it defines you in a way it never has before. There's a momentum to the trend that might not be obvious to people who work outside the industry. The new design standard of the last two or three years, more open and more photocentric—what I think of as "Pinteresty"—is making not just pictures, but *beauty specifically* more important. OkCupid recently made a change for some photo displays, going from the size of the black box to that of the red, below:



The designers just wanted the page to look more modern. What they didn't anticipate (and later had to mitigate) was the following: all those extra pixels allowed the pretty faces to outshine the others all the more. The rich got richer. It was the web-design equivalent of American domestic policy.



Given this pressure it's no wonder that body-image blogs are so prevalent. And that posts tagged like #thinspiration #thinspo #loseweight #keeplosing #proana #thighgap became so common that both Tumblr and Pinterest (independent of each other) had to alter their Terms of Service to ban this kind of content. If you're wondering what the last two hashtags are, #proana is short for "pro anorexia"—people *in favor of* starvation as a weight-loss technique. Meanwhile, #thighgap refers to having thighs so thin that they do not touch when you stand with your feet and knees together. It's a trait fetishized by teenage girls. Quite apart from the questionable desirability, it's biologically impossible for most of them. The full depravity of the phenomenon can't hit you until you search for these tags yourself and are confronted with an unending page of broken bodies tilting at the camera—not only are the "inspiring" women deathly thin, they are also frequently in lingerie, bikinis, underwear. The blogs, created by women, are truly the epitome of the male gaze—and I say this as a person reflexively skeptical of the language of the academic left.

Tumblr and Pinterest banning the content didn't solve anything, of course, least of all their users' body-image issues, so the sites are now taking another approach.

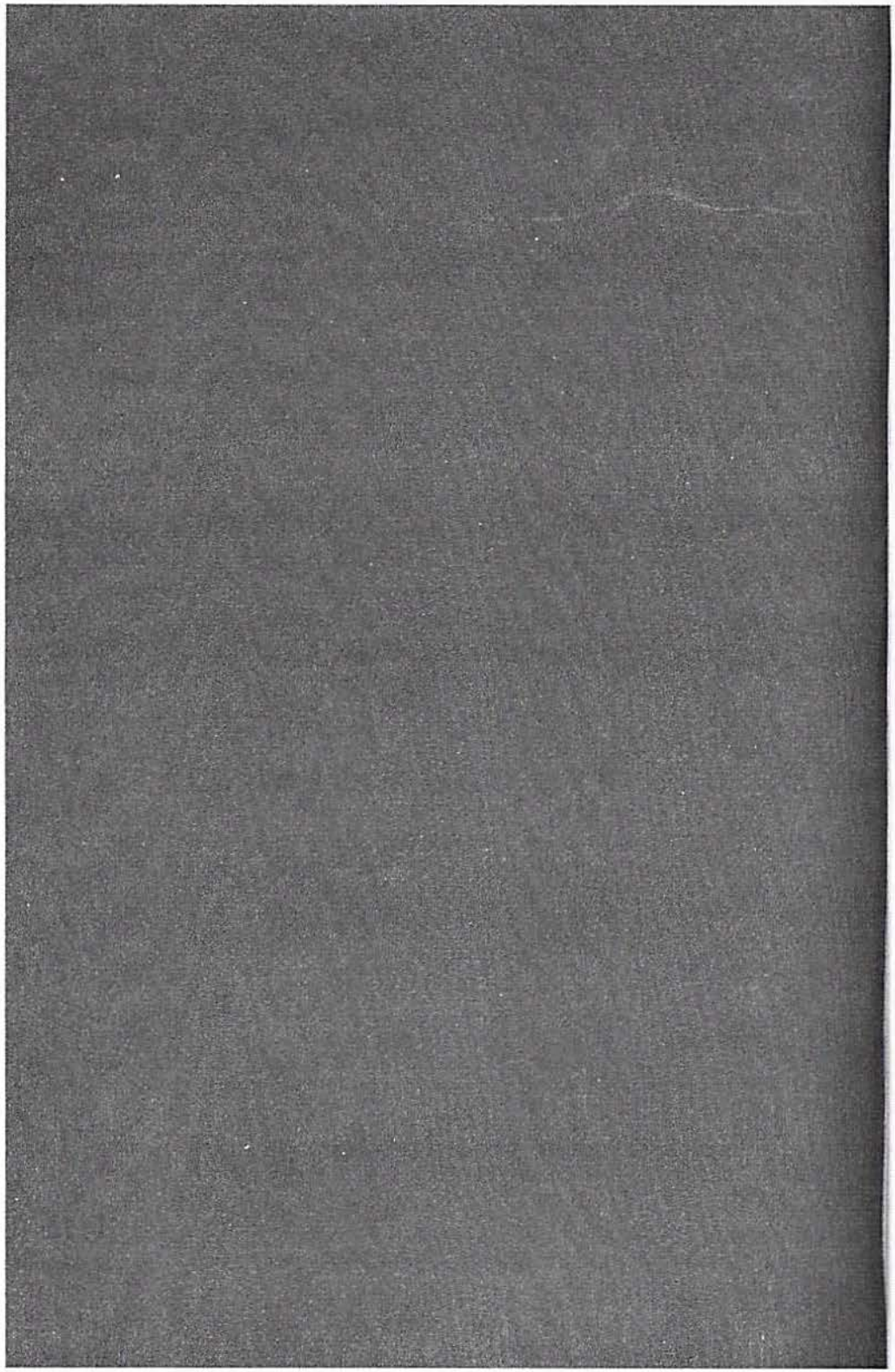
Because these blogs are tagged, they are able to intervene algorithmically—search for thighgap on Tumblr and the screen goes blank, an overlay appearing:

"if you or someone you know is dealing with an eating disorder . . ."

A link to help and resources follows. It is a small measure, but before the behavior was digitized, there was practically no way to get directly at this problem, at least not until visible damage had already occurred. There was only rumor—an ear at the bathroom door, perhaps a parent's sad suspicion. Data is about how we're really feeling—feeling about one another, yes, but also about ourselves. If it finds divides in our culture, our politics, our habits, our tribes, it finds divides within us, too. And that's a hopeful thought, because for anything to be made whole, the first step is to know what's missing.

8.

It's What's Inside That Counts



A Note on the Data

Numbers are tricky. Even without context, they give the appearance of fact, and their specificity forbids argument: *20,679 Physicians say "LUCKIES are less irritating."* What else is there to know about smoking, right? The illusion is even stronger when the numbers are dressed up as statistics. I won't rehash the old wisdom there. But behind every number there's a person making decisions: what to analyze, what to exclude, what frame to set around whatever pictures the numbers paint. To make a statement, even to just make a simple graph, is to make choices, and in those choices human imperfection inevitably comes through. As far as I know, I've made no motivated decision that has bent the outcome of my work—the data of people acting out their lives is interesting enough without me needing to lead it one way or another. But I have made choices, and those choices have affected the book. I'd like to walk you through a few of them.

My first choice was probably my most difficult: the decision to focus on male-female relationships when I talk about attraction and sex. Space, of course, was a factor—to include same-sex relationships would've meant repeating each graph or table in triplicate. But more than that was the discovery that same-sex relationships aren't exceptional—they follow all the same trends. Gay men, for example, prefer younger partners just like straight men do. For issues that have to do with sex only indirectly, such as ratings from one race to another, gays and straights also show similar patterns. Male-female relationships allowed for the least repetition and widest resonance per unit of space, so I made the choice to focus on them.

My second decision, to leave out statistical esoterica, was made with much less regret. I don't mention confidence intervals, sample sizes, p values, and similar devices in *Dataclysm* because the book is above all a popularization of data and data science. Mathematical wonkiness wasn't what I wanted to get across. But like the spars and crossbeams of a house, the rigor is no less present for

being unseen. Many of the findings in the book are drawn from academic, peer-reviewed sources. I applied the same standards to the research I did myself, including a version of peer-review: much of the OkCupid analysis was performed first by me and then verified independently by an employee of the company. Also, I separated the analysis from the selection and organization of the data to make sure the former didn't motivate the latter. One person would extract the information, another would try to figure out what it meant.

Sometimes, I present a trend and attribute a cause to it. Often that cause is my best guess, given my understanding of all the forces in play. To interpret results—a necessity in any book that isn't just reams of numbers—I had to choose one explanation from a variety of possibilities. Is there some force besides age behind what I call Wooderson's law (the fact that straight men of all ages are most interested in twenty-year-old women)? Perhaps. But I think it is very unlikely. "Correlation does not imply causation" is a good thing for everyone to keep in mind—and an excellent check on narrative overreach. But a snappy phrase doesn't mean that the question of causation isn't itself interesting, and I've tried to attribute causes only where they are most justified.

For almost all the parts of *Dataclysm* that overlap with posts on OkCupid's blog, I chose to redo the work from scratch, on the most recent data, rather than quote my own previous findings. I did so because, frankly, I wanted to double-check what I'd done. The research published there from 2009 through 2011 was put together piecemeal. Many different people—I can count at least five—had pulled male-female message-reply rates for me over those three years, just to name one frequently used data point, and going back through my records of this data, there was no way to be sure what data set had generated the results. Doing it again myself, I could be sure. I could also enforce a uniform standard across all my research (for example, restricting analysis to only people ages twenty to fifty—a choice I made because those are the ages where I knew I had representative data).

Because the research is new, the numbers printed in *Dataclysm* are different from the numbers on the blog. Curves bend in slightly new ways. Graphs are a bit thicker or perhaps a bit thinner in places. The findings in the book and on the blog are nonetheless consistent. Ironically, with research like this, precision is often less appropriate than a generalization. That's why I often round findings to the nearest 5 or 10 and the words "roughly" and "approximately" and "about" appear frequently in these pages. When you see in some article that "89.6 percent" of people do x, the real finding is that "many" or "nearly all" or "roughly 90 percent" of them do it, it's just that the writer probably thought the decimals sounded cooler and more authoritative. The next time a scientist runs the numbers, perhaps the outcome will be 85.2 percent. The next time, maybe it's 93.4. Look out at the churning ocean and ask yourself exactly which whitecap is "sea level." It's a pointless exercise at best. At worst, it's a misleading one.

If you trace the findings in *Dataclysm* back to the original sources, the OkCupid data isn't the only place you'll see discrepancies. This data of our lives, being itself practically a living thing, is always changing. For example, my Klout score, which is holding steady at 34 as I write these words, will have no doubt gone up by the time you read them, since part of my obligation to Crown will be to tweet about this book. User engagement, ho!

Sometimes the numbers shift for no obvious reason. My copy editor and I had a mess of a time pinning down the Google autocompletes for prompts like "Why do women..." Google had given each of us slightly different results ("...wear thongs?" was my third result to the above, presumably because that's a typically male question [?]. Hers was "...wear bras?"). Then when I checked a few weeks later, I myself saw something different: "...wear high heels?" Since it was the most recent result, that's what ended up in the book.

As interesting a tool as it is, the black box of Google's autocomplete (and Google Trends, for that matter) is an example of one of the worst things about today's data science—its opaqueness. Corroboration, so important to the scientific method, is difficult, because so much information is proprietary (and here OkCupid is as guilty as anyone). Even as most social media companies trumpet the hugeness and potential of their data, the bulk of it has stayed off-limits to the larger world. Data sets currently move through the research community like yeti—*I have a bunch of interesting stuff but I can't say from where; I heard someone at Temple has tons of Amazon reviews; I think L has a scrape of Facebook.* That last is something I was told by three unrelated academics; they referred to another scientist by name, which I've here obscured. L does in fact have that rogue Facebook scrape—I met him and confirmed—but he can't show it to anyone. He's really not supposed to have it at all. Data is money, which means companies treat it as such—and though some digital data sits out in the open, it's secured

behind legal walls as thick as any vault's. If you look at your friend Lisa's Facebook page, observe that her name is Lisa, and publish that fact (anywhere!)—you have technically stolen Facebook's data. If you've ever signed up for a website and given a fake zip code or a fake birthday, you have violated the Computer Fraud and Abuse Act. Any child under thirteen who visits newyorktimes.com violates their Terms of Service and is a criminal—not just in theory, but according to the working doctrine of the Department of Justice.* The examples I've laid out are extreme, sure, but the laws involved are so broadly written as to ensure that, essentially, every Internet-using American is a tort-feasing felon on a lifelong spree of depraved web browsing. Whether anyone penalizes you for your "crime" is another matter, but, legally, you are prostrate, a boot on your neck. A company's general counsel, or a district attorney looking to please an important corporate donor, can destroy your life simply by deciding to press. When it suits, they do. So social scientists are very cagey with data sets; actually, more than yeti, they treat them like big bags of weed—possessive, slightly paranoid, always curious who else is holding and how dank that shit is.

Increasingly the preferred practice is to bring researchers in-house rather than release information outside.† And that approach has yielded, among many fruits, the novel research by Facebook's data team and Seth Stephens-Davidowitz's fine work at Google, both of which I've drawn on here. I hope more companies follow this model, and that eventually we, the owners of the sites, will find a way to release our data for the public good without jeopardizing our users' privacy in the act.

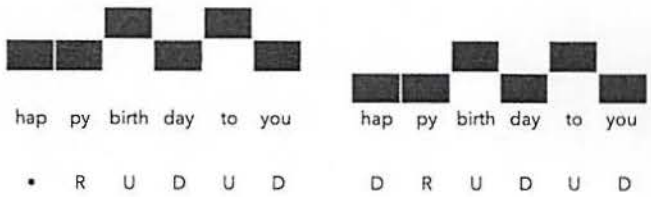
∞

It's old hat now, but the app Shazam was, to me, one of the first great wonders of the iPhone. It's a little program for identifying music—if some song is playing, and you want to know what it is, you just turn on the app and hold up your phone. Shazam listens through the microphone, and, like, two seconds later, it tells you what you're listening to. The first time someone did it in front of me, I was just

* For more on the Kafkaesque implications of the CFAA, please see "Until Today, If You Were 17, It Could Have Been Illegal to Read Seventeeen.com Under the CFAA" and "Are You a Teenager Who Reads News Online? According to the Justice Department, You May Be a Criminal," both published by the Electronic Frontier Foundation.

† I wish this were called hotboxing, but sadly, no.

blown away, not only at how little the software needed to get the song right (it can often work through walls or above the din of a bar), but at how fast it worked. It was the closest thing I'd seen to magic, at least until I came to know a certain able necromancer who, at a whim, could summon fees and add them to my goddamn kitchen renovation. But anyway, as I later found out, Shazam relies on an incredible principle: that almost any piece of music can be identified by the up/down pattern in the melody—you can ignore everything else: key, rhythm, lyrics, arrangement . . . To know the song, you just need a map of the notes' rise and fall. This melodic contour is called the song's Parsons code, named after the musicologist who developed it in the 1970s. The code for the first two lines of "Happy Birthday" is •RUDUDDRUDUD, with U meaning "melody up," D meaning "melody down," and R for "repeated note." The dot • just marks the beginning of the tune, which of course isn't up or down from anything. Hum it to yourself to check:



As crazy as it seems, the code for "Happy Birthday" is practically unique across the entire catalog of recorded music, as is the code for almost all songs. And it's because these few letters are such a concise description that Shazam is so fast: instead of a guitar, Paul McCartney, and just the right amount of reverb, "Yesterday" starts with •DRUUUUUUDDR. That's a lot easier to understand.

Like an app straining for a song, data science is about finding patterns. Time after time, I—and the many other people doing work like me—have had to devise methods, structures, even shortcuts to find the signal amidst the noise. We're all looking for our own Parsons code. Something so simple and yet so powerful is a once-in-a-lifetime discovery, but luckily there are a lot of lifetimes out there. And for any problem that data science might face, this book has been my way to say: I like our odds.

Notes

We no longer live in a world where a reader depends on endnotes for "more information" or to seek proof of facts or claims. For example, I imagine any reader interested in Sullivan Ballou will have Googled him long before she consults these notes and transcribes into her browser the links I've provided. So I have used this section to focus on the many sources that have contributed not only facts but ideas to this book. I've also used it to substantiate or explain claims about my own proprietary data.

Since the subject of *Dataclysm* is changing almost daily, I've decided to enhance this section online at dataclysm.org/endnotes, where you will find additional source material and findings from emerging research.

Introduction

9 ***10 million people will use the site*** For this number, I counted every person who logged into OkCupid in the twelve months trailing April 2014: 10,922,722.

9 ***Tonight, some thirty thousand couples*** It's the great unknowable of running an online dating site: How many of the users actually meet in person? And what happens next? This passage represents my best guesses at some basic in-person metrics. I used two separate methods:

1. I assumed someone who's actively using OkCupid goes on one date every other month. I think this is conservative. At roughly 4,000,000 active users each month, that means roughly 65,000 people go on dates each day, meaning roughly 30,000 couples.

2. Every day 300 couples wind their way through our "account disable" interface to let us know that they no longer need OkCupid specifically because they have found a steady relationship on OkCupid. These are couples who (a) are dating seriously enough to shut down their OkCupid accounts,

and who (b) are willing to go through the trouble of filling out a bunch of forms to let us know their new relationship status. I estimate that Group B represents only 1 in 10 of the long-term couples actually created by the site. And I estimate that Group A represents the outcome of only 1 in 10 first dates. Therefore, there must be 3,000 long-term couples, from 30,000 first dates each day. Of every 3,000 long-term couples, I believe something less than 1 in 10 go on to get married. One way to look at this: How many serious relationships did you have before you found the person you settled down with? I imagine the average number is roughly 10.

These appraisals together are mutually supporting, at least of the "first dates" number, and even if it's approximate, I think the deeper metrics follow plausibly.

15 *ratings of pizza joints on Foursquare* Ratings from a random sample of 305 New York City pizza places accessed through Foursquare's public API.

15 *the recent approval ratings for Congress* These were collected from the 529 polls measuring "congressional job approvals" listed on the site real clearpolitics.com from January 26, 2009, through September 14, 2013. See realclearpolitics.com/epolls/other/congressional_job_approval-903.html#polls.

15 *NBA players by how often* The chart shows percent of games started for each of the players listed on a team roster for the 2012–2013 season on espn.com. Yes, I'm counting the 76ers as an NBA team.

17 *6 percent* This number comes from taking the geometric mean of the distances between each of the 21 discrete data points along the curves. So, for curves *a* and *b*, I calculated:

$$\sqrt{\sum_{k=1}^{21}(a_k - b_k)^2}$$

Which equals 0.056.

17 *58 percent of men* The male attractiveness curve is centered more than a whole standard deviation below the female. Translating the same disparity to IQ means that the median male IQ would be slightly lower than 85, which is the threshold for "borderline intellectual functioning." For example, the US Army doesn't accept applicants with IQs below 85. I say "brain damaged" as a bit of hyperbole meant to capture this shift. Strictly speaking, I mean that 58 percent of men would have IQs lower than 85.

18 *half the single people in the United States* Specifying the reach of the dating data I have was a challenge. I've strived to do so in broad, easy-to-grasp terms because, unlike Facebook or Twitter, I know much of my reading audience has never used a dating site. If you've been married or in a relationship since the late '90s or before, you have never needed online dating. According to the 2011 Census numbers, there are 103 million single people ages fifteen to sixty-four in the United States—that counts *everyone* who isn't legally married, including many people who are actually in long-term relationships and nearly every gay person. Together, Tinder, OkCupid, DateHookup, and Match.com registered 57 million US accounts from 2011 to 2013, and 23 million in the last of those three years alone. "Half" is my approximation of 57/103, minus the 10 to 15 percent wastage in overlap and duplicate accounts.

18 *"Women are inclined to regret"* This quote is from the "Findings" section of the February 2014 issue of *Harper's* by Rafil Kroll-Zaidi.

18 *A beta curve plots* My data researcher, Tom Quisel, helped me put the binomial nature of beta curves into simple terms. He also pointed out that they're used to model weather, and ran the comparisons to the by-city patterns on weatherbug.com.

19 *Some 87 percent of the United States is online* See Susannah Fox and Lee Rainie, "Summary of Findings," Pew Research Internet Project, Pew Research Center, February 27, 2014, pewinternet.org/2014/02/27/summary-of-findings-3/.

19 *that number holds* ... For example, Internet use among white, African American, and Hispanic Americans is 85, 81, and 83 percent, respectively. One can only assume adoption among Asian Americans is similar. Adoption is above 80 percent for all age groups, save people sixty-five and older. Susannah Fox and Lee Rainie, "Internet Users in 2014," Pew Research Internet Project, Pew Research Center, February 27, 2014, pewinternet.org/files/2014/02/12-internet-users-in-2014.jpg.

20 ***More than 1 out of every 3 Americans access Facebook*** Facebook reported 128 million US users in August 2013. Facebook had at least 1.26 billion users worldwide in September 2013. World and US population statistics are from Wikipedia. See expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats/.

20 ***fundamentally populist*** This is something like common knowledge among people who study social media adoption beyond the Google Glasshole/Technocrat use case. See Pew Research Center's "Demographics of Key Social Networking Platforms" (2013). The report shows no statistically significant difference in rates of Twitter use between the "high school grad or less" and "College +" educational cohorts (coming in at 17 percent and 18 percent, respectively). Pew surveys a random cross-section of Americans eighteen years old or older, so very few of the "high school grad or less" cohort are that way simply because they're still *in* high school. By ethnicity, Pew reports adoption rates of 29 percent among blacks and 16 percent among both whites and Hispanics. The full report, by Maeve Duggan and Aaron Smith, is here: pewinternet.org/2013/12/30/demographics-of-key-social-networking-platforms/.

21 ***It's called WEIRD research*** This fact and my general take on the phenomenon are adapted from "Psychology Is WEIRD," by Bethany Brookshire, in *Slate*. See also "The Roar of the Crowd," *The Economist*, May 24, 2012, economist.com/node/21555876.

22 ***Pharaoh Narmer*** As you can imagine, this is up for debate, though Narmer, also known as Serket, is a defensible choice. In earlier drafts I had Gilgamesh, the Akkadian hero, in this place because J. M. Roberts, in his *History of the World* (New York: Oxford University Press, 1993), chooses Gilgamesh. I eventually went with Narmer because his life is dated several centuries earlier, and he seemed to me as likely to have actually lived. Yahoo! Answers also mentions Elvis Presley.

Chapter 1: Wooderson's Law

34 ***This isn't survey data*** This is a good place to point out that for anyone's attractiveness to have been considered in my analysis in this book, that person needed to have received votes from at least twenty-five other people. For something as idiosyncratic as attraction, I felt an average score comprising fewer than twenty-five votes wasn't reliable.

39 ***per the US Census*** These numbers are from the US Census Bureau's "Marital Status of People 15 Years and Over, by Age, Sex, Personal Earnings, Race, and Hispanic Origin, 2011."

Chapter 2: Death by a Thousand Mehs

46 ***"Beauty is looks you can never forget"*** John Waters, *Shock Value: A Tasteful Book About Bad Taste* (Philadelphia: Running Press, 2005), p. 128.

48 ***concept called*** variance I used standard deviation to measure variance throughout this chapter.

50 ***the "pratfall effect"*** A Google search for "pratfall effect" will yield many examples. I particularly relied on the précis "The Positive Effect of Negative Information" by Bill Snyder and the original paper he summarizes, "When Blemishing Leads to Blossoming: The Positive Effect of Negative Information," by Danit Ein-Gar, Zakary Tormala, and Shiv Tormala, *Journal of Consumer Research* 38, no. 5 (2012): 846–59.

50 ***Our sense of smell*** For this passage, I relied on Fabian Grabenhorst et al., "How Pleasant and Unpleasant Stimuli Combine in Different Brain Regions: Odor Mixtures," *Journal of Neuroscience* 27, no. 49 (2007): 13532–40, doi: 10.1523/JNEUROSCI.3337–07.2007. Wikipedia's "Indole" entry describes its "intense fecal smell." For more on indole's role in perfumes and in naturally occurring flower scents, see, as I did, perfumeshrine.blogspot.com/201%5/jasmine-indolic-vs-non-indolic.html.

51 ***Here are six women*** We received these permissions using a double-blind system, to protect user privacy. I submitted criteria (women, high variance scores, midrange overall attractiveness) to OkCupid's data team. The data team generated a list of possible names, which they passed on to our admin. She then had a list of names, with no other information attached, and was told to contact them for blanket photo authorization. (We commonly receive press requests for user photos, so this type of outreach isn't unusual.) A photo and its unique attributes were only connected once permission was granted.

77 ***Another long-held idea in network theory*** Though embeddedness was first proposed by Granovetter in 1985, my remaining discussion of embeddedness and of interpersonal network theory is drawn from the primary source behind this chapter, Backstrom and Kleinberg's "Romantic Partnerships." I apply their heuristic to my own networks and somewhat simplify their original work for a nonacademic audience.

79 ***an astounding 75 percent of the time*** Backstrom and Kleinberg define many subtly different mathematical kinds of dispersion. My number here refers to the accuracy they reported with the method they call "recursive dispersion."

79 ***50 percent more likely*** This is drawn from the following passage in Backstrom and Kleinberg's paper: "We find that relationships on which recursive dispersion fails to correctly identify the partner are significantly more likely to transition to 'single' status [that is, break up] over a 60-day period. This effect holds across all relationship ages and is particularly pronounced for relationships up to 12 months in age; here the transition probability is roughly 50% greater when recursive dispersion fails to recognize the partner."

80 ***Have a meeting with Microsoft people*** This might not be broadly true of all Microsoft employees; however, the teams responsible for Microsoft's mobile and tablet products are, in my experience, dogfooders of the first order. Windows mobile is so rare as to be especially noteworthy, so you remember it when you see it. This is a good place to point out that I am a lifelong user of Microsoft Office, and all the charts and much of the analysis in this book were done in Excel.

Chapter 5: There's No Success Like Failure

86 ***one of Google's best designers*** Douglas Bowman leaving Google is a famous event in tech circles. See his own post "Goodbye, Google" at stopdesign.com/archive/2009/03/20/goodbye-google.html.

88 ***no evidence of people gaming the system*** It was fairly simple to unscramble a Crazy Blind Date photo; we knew this would be the case. Sure enough, about a week after launch a few hackers had built apps to de-anonymize the photos. However, these apps never caught on, mostly because they were difficult to use and even then only worked part of the time. These unscramblers were not a factor in Crazy Blind Date's product trajectory or the data it generated. The scrambled example photo printed in the book is a stock photo, licensed from Getty Images.

Chapter 6: The Confounding Factor

99 ***of a certain type*** See, for example, "Blacks Still Dying More from Cancer Than Whites," by Jordan Lite, *Scientific American*, February 2009. Also see the Sentencing Project's "Criminal Justice Primer for the 111th Congress," which details many depressing disparities in the sentences handed down to whites, compared to minority defendants: sentencingproject.org/doc/publications/cjprimer2009.pdf.

100 ***conclusions like this*** The headline cited is from ThinkProgress.org. "Study: Black Defendants Are at Least 30% More Likely to Be Imprisoned Than White Defendants for the Same Crime," by Inimai Chettiar, August 30, 2012, thinkprogress.org/justice/2012/08/30/770501/study-black-defendants-are-at-least-30-more-likely-to-be-imprisoned-than-white-defendants-for-the-same-crime.

100 ***in the 97,000 results*** It's a bit of a hack to get Google to give you a number here. My exact query was for " 'black quarterback' –adsffsdada." Using the minus sign with the nonsense word keeps the page from automatically returning images instead of the "about 97,000 results" text. I'm sure without the browser in front of you, this all sounds mystifying. Try it yourself if you care, and you'll see immediately what I mean. Also, this is another example of a raw number that has changed during the course of writing this book. I've also gotten "89,800 results" returned to me.

100 ***I found only one article*** See Jason Lisk, "Quarterbacks and Whether Race Matters," *The Big Lead*, December 2, 2010, thebiglead.com/2010/12/02/quarterbacks-and-whether-race-matters/. Of course, the fact that I found only one writer who calculates quarterback rating by race is hardly proof that no other writer has made the calculation. However, I spent several hours combing results and found only Lisk.

101 ***the four largest racial groups*** 15 percent of OkCupid users who select an ethnicity select more than one race; 3 percent select a race other than the

four largest. These people are excluded from the analysis, as are people who neglected to choose a race at all.

102 *"normalize" each row* I normalized against the simple average in each row, rather than the weighted average. Because of the preponderance of white people, the latter technique would've skewed the matrix, functionally using what everyone thinks of white people as the "norm." A simple average captures the following: "When a person of race A meets an arbitrary person of race B, how does A appraise B, relative to A's appraisals of other races?" That's the interesting question, and what we want to investigate.

103 *There is no cadre of racists* An analysis of individual bias applied by non-black men to black female profiles shows a median deduction of 0.6 stars, with most of the sample applying a deduction from 0.2 to 1.0 stars. 82 percent of the sample shows at least some consistent anti-black bias.

103 *Here are our numbers* Though the numbers I list for OkCupid here were generated from internal data, you can see those numbers corroborated and compared to Quantcast's national averages by visiting https://www.quantcast.com/okcupid.com?country=US. Select "Ethnicity" from the Demographics menu and expand the "US average" feature.

109 *OkCupid users putting it in their own words* These excerpts are from user-submitted "Success Stories" published on the site. Bella and Patrick's is here: https://www.okcupid.com/success/story?id=2855. Dan and Jenn's is here: https://www.okcupid.com/success/story?id=2587.

110 *"There are very few"* Barack Obama's quote is excerpted from his comments on the George Zimmerman verdict: whitehouse.gov/the-press-office/2013/07/19/remarks-president-trayvon-martin.

110 *One paper asked* See "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," by Marianne Bertrand and Sendhil Mullainathan, *American Economic Review* 94, no. 4 (2004): 991–1013, doi: 10.1257/0002828042002561.

111 *Osagie K. Obasogie* My discussion of Obasogie's work relies on Francie Latour's *Boston Globe* article "How Blind People See Race," January 19, 2014. Latour provides a précis of Obasogie's book *Blinded by Sight: Seeing Race Through the Eyes of the Blind* (Redwood City, CA: Stanford University Press, 2014), and interviews him.

113 *Baywatch* I was in Japan in 1992. *Baywatch* was popular worldwide by then, but didn't arrive in the Japanese mainstream until a year later. Nonetheless, surf culture, California, and sun-kissed blondness were already everywhere. When you walked into a "cool" clothing store, they'd be playing the Beach Boys. In 1992. Stuff like "Surfin' Safari," not "Kokomo."

Chapter 7: The Beauty Myth in Apotheosis

117 *Korean proverb* I got this from William Manchester's biography of Douglas MacArthur, *American Caesar* (New York: Little, Brown, 1978), which, in the death throes of this book, I was reading to get my mind off data.

118 *beauty operates on a Richter scale* I was already familiar with the logarithmic nature of the Richter scale, but relied on the Wikipedia entry for "Richter magnitude scale" to understand the implications of the benchmark magnitudes. In comparing beauty to the scale, I am, of course, employing a bit of poetic license; the functions are not exactly the same.

119 *Here is data for interview requests* The Shiftgig data was provided by their data team and with the gracious cooperation of founder Eddie Lou.

119 *And for friend counts* These are the aggregated and anonymized friend counts for OkCupid users who've elected to connect their OkCupid accounts to their Facebook accounts.

120 *a foundational paper of social psychology* See "What Is Beautiful Is Good," by Karen Dion, Ellen Berscheid, and Elaine Walster in *Journal of Personality and Social Psychology* 24 (1972): 285–90.

120 *It was the first in a now long line . . .* This passage adapts conclusions from and directly quotes "Pretty Smart? Why We Equate Beauty with Truth," by Robert M. Sapolsky, in the *Wall Street Journal*, January 17, 2014. The Duke neuropsychologists alluded to are Takashi Tsukiura and Roberto Cabeza. See also "Jurors Biased in Sentencing Decisions by the Attractiveness of the Defendant" at *Psychology and Crime News* for an overview of the effects of physical attractiveness in the criminal justice process: crimepsychblog.com/?p=1437, posted by user EmmaB, April 3, 2007.

123 *both Tumblr and Pinterest* See "A New Policy Against Self-Harm Blogs," Tumblr's staff blog, March 1, 2012, staff.tumblr.com/post/18132624829/self-harm-blogs.

See also "Pinterest 'Thinspiration' Content Banned According to New Acceptable Use Policy," by Ellie Krupnick, *Huffington Post*, March 26, 2012, huffingtonpost.com/2012/03/26/pinterest-thinspiration-content-banned_n_1380484.html.

The *Huffington Post* has actively covered the "thinspiration" phenomenon. See "The Hunger Blogs: A Secret World of Teenage 'Thinspiration,'" by Carolyn Gregoire, February 8, 2012, huffingtonpost.com/2012/02/08/thinspiration-blogs_n_1264459.html.

For more on "thighgap" (and for evidence that altering the Terms of Service did not solve the problem), see "The Sexualization of the Thigh Gap," by Allie Jones, on *The Wire*, November 22, 2013, thewire.com/culture/2013/11/sexualization-thigh-gap/355434/.

Chapter 8: It's What's Inside That Counts

127 ***That's been the popular standard since*** These basic facts on the origins of Gallup were found on the "Gallup (company)" Wikipedia entry.

127 ***surveys have historically*** As I mention in the text and in the footnotes to this chapter, the idea of using Google Trends to look at taboos is the brainchild of Seth Stephens-Davidowitz. His June 9, 2012, article in the *New York Times*, "How Racist Are We? Ask Google," and his 2013 Harvard PhD dissertation, "Essays Using Google Data," http://nrs.harvard.edu/urn-3:HUL.InstRepos:10984881, were the inspiration for this chapter. For the question of exactly how much Obama's race cost him in the 2008 election, picked up later in the chapter, I rely directly on Stephens-Davidowitz's work. For the over-time use of the word "nigger" and in the other direct citations of Google Trends findings in the chapter, the work is my own, though I am adapting a method he first suggested.

Though Stephens-Davidowitz now works at Google, I emphasize that his search research is always based on public and anonymous sources, not on privileged access to anyone's personal search history. My own search research is similarly based on a public, anonymous source, namely Google Trends: google.com/trends.

127 ***This tendency is called*** I used Wikipedia's "Social desirability bias" entry as my source for basic details here.

127 ***The most famous case*** The Bradley effect first came to my attention during the 2008 campaign, as pundits wondered how it would affect Obama's polling on Election Day. Here, I relied on the Wikipedia entry "Bradley effect" for basic facts surrounding Tom Bradley's defeat.

128 ***Since the service launched*** See Nick Bilton, "Google Search Terms Can Predict Stock Market, Study Finds," *New York Times* Bits blog, April 26, 2013. See also Casey Johnston, "Google Trends Reveals Clues About the Mentality of Richer Nations," *Arstechnica*, April 5, 2012, arstechnica.com/gadgets/2012/04/google-trends-reveals-clues-about-the-mentality-of-richer-nations/; and Tobias Preis et al., "Quantifying the Advantage of Looking Forward," *Scientific Reports* 2, no. 350 (2012), doi: 10.1038/srep00350.

128 ***track epidemics of flu*** Google Flu was first developed in the paper "Detecting Influenza Epidemics Using Search Engine Query Data," by Jeremy Ginsberg et al. in *Nature* 457 (2009): 1012–14, doi:10.1038/nature07634. Recently, Flu's efficacy has been found wanting: see Kaiser Fung, "Google Flu Trends' Failure Shows Good Data > Big Data," *Harvard Business Review* Blog Network, March 25, 2014.

128 ***included in 7 million searches a year*** Stephens-Davidowitz, "How Racist Are We?"

129 ***more American than "apple pie"*** Google Trends index for US searches, January 2004–September 2013, for "apple pie": 25. For "nigger": 32.

129 ***And, tellingly*** The ratio of "nigga":"nigger" is thirty times higher in tweets sent from my Twitter corpus than reflected in Google Trends. That is, on Twitter "nigger" appears thirty times less frequently.

130 ***roughly 1 in 100 searches for "Obama"*** Stephens-Davidowitz shared this fact with me over e-mail.

130 ***25 percent below the pre-Obama status quo*** Stephens-Davidowitz, "How Racist Are We?" This is also confirmable firsthand through Google Trends.

131 ***Other awful terms*** These racial epithets are far less common on Twitter, in private messages to OkCupid, and in Google search, as confirmed by Stephens-Davidowitz via e-mail.

131 ***If you're not familiar with autocomplete*** The algorithm that supplies Google