Dataclysm
Who We Are*
Christian Rudder

* When We Think No One’s Looking
Table of Contents

Cover
Title Page
Copyright
Introduction
Part 1.
What Brings Us Together
1. Wooderson’s Law
2. Death by a Thousand Mehs
3. Writing on the Wall
4. You Gotta Be the Glue
5. There’s No Success Like Failure
Part 2.
What Pulls Us Apart
6. The Confounding Factor
7. The Beauty Myth in Apotheosis
8. It’s What’s Inside That Counts
9. Days of Rage
Part 3.
What Makes Us Who We Are
10. Tall for an Asian
11. Ever Fallen in Love?
12. Know Your Place
13. Our Brand Could Be Your Life
14. Breadcrumbs
Coda
A Note on the Data
Notes
Acknowledgments
Index
CONTENTS

Cover
Title Page
Copyright

Introduction

Part 1.
What Brings Us Together

1. Wooderson’s Law

2. Death by a Thousand Mehs

3. Writing on the Wall

4. You Gotta Be the Glue

5. There’s No Success Like Failure

Part 2.
What Pulls Us Apart
6. The Confounding Factor

7. The Beauty Myth in Apotheosis

8. It’s What’s Inside That Counts

9. Days of Rage

**Part 3.**

What Makes Us Who We Are

10. Tall for an Asian

11. Ever Fallen in Love?

12. Know Your Place

13. Our Brand Could Be Your Life

14. Breadcrumbs

Coda

A Note on the Data
Notes
Acknowledgments
Index
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:
Or take the recent approval ratings for Congress, which, because politicians are the moral opposite of pizza, skew the other way:

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:

![Chart: Perception of Male Attractiveness vs. Female Attractiveness]

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.¹ 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 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.
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.”
PART 1

What Brings Us Together
1. Wooderson’s Law

2. Death by a Thousand Mehs

3. Writing on the Wall

4. You Gotta Be the Glue

5. There’s No Success Like Failure
Wooderson’s Law
Up where the world is steep, like in the Andes, people use funicular railroads to get where they need to go—a pair of cable cars connected by a pulley far up the hill. The weight of the one car going down pulls the other up; the two vessels travel in counterbalance. I’ve learned that that’s what being a parent is like. If the years bring me low, they raise my daughter, and, please, so be it. I surrender gladly to the passage, of course, especially as each new moment gone by is another I’ve lived with her, but that doesn’t mean I don’t miss the days when my hair was actually all brown and my skin free of weird spots. My girl is two and I can tell you that nothing makes the arc of time more clear than the creases in the back of your hand as it teaches plump little fingers to count: one, two, tee.

But some guy having a baby and getting wrinkles is not news. You can start with whatever the Oil of Olay marketing department is running up the pole this week—as I’m writing it’s the idea of “color correcting” your face with a creamy beige paste that is either mud from the foothills of Alsace or the very essence of bullshit—and work your way back to myths of Hera’s jealous rage. People have been obsessed with getting older, and with getting uglier because of it, for as long as there’ve been people and obsession and ugliness. “Death and taxes” are our two eternals, right? And depending on the next government shutdown, the latter is looking less and less reliable. So there you go.

When I was a teenager—and it shocks me to realize I was closer then to my daughter’s age than to my current thirty-eight—I was really into punk rock, especially pop-punk. The bands were basically snottier and less proficient versions of Green Day. When I go back and listen to them now, the whole phenomenon seems supernatural to me: grown men brought together in trios and quartets by some unseen force to whine
about girlfriends and what other people are eating. But at the time I thought these bands were the shit. And because they were too cool to have posters, I had to settle for arranging their album covers and flyers on my bedroom wall. My parents have long since moved—twice, in fact. I’m pretty sure my old bedroom is now someone else’s attic, and I have no idea where any of the paraphernalia I collected is. Or really what most of it even looked like. I can just remember it and smile, and wince. Today an eighteen-year-old tacks a picture on his wall, and that wall will never come down. Not only will his thirty-eight-year-old self be able to go back, pick through the detritus, and ask, “What was I thinking?,” so can the rest of us, and so can researchers. Moreover, they can do it for all people, not just one guy. And, more still, they can connect that eighteenth year to what came before and what’s still to come, because the wall, covered in totems, follows him from that bedroom in his parents’ house to his dorm room to his first apartment to his girlfriend’s place to his honeymoon, and, yes, to his daughter’s nursery. Where he will proceed to paper it over in a billion updates of her eating mush.

A new parent is perhaps most sensitive to the milestones of getting older. It’s almost all you talk about with other people, and you get actual metrics at the doctor’s every few months. But the milestones keep coming long after babycenter.com and the pediatrician quit with the reminders. It’s just that we stop keeping track. Computers, however, have nothing better to do; keeping track is their only job. They don’t lose the scrapbook, or travel, or get drunk, or grow senile, or even blink. They just sit there and remember. The myriad phases of our lives, once gone but to memory and the occasional shoebox, are becoming permanent, and as daunting as that may be to everyone with a drunk selfie on Instagram, the opportunity for understanding, if handled carefully, is self-evident.

What I’ve just described, the wall and the long accumulation of a life, is
what sociologists call *longitudinal data*—data from following the same people, over time—and I was speculating about the research of the future. We don’t have these capabilities quite yet because the Internet, as a pervasive human record, is still too young. As hard as it is to believe, even Facebook, touchstone and warhorse that it is, has only been big for about six years. It’s not even in middle school! Information this deep is still something we’re building toward, literally, one day at a time. In ten or twenty years, we’ll be able to answer questions like … well, for one, how much does it mess up a person to have every moment of her life, since infancy, posted for everyone else to see? But we’ll also know so much more about how friends grow apart or how new ideas percolate through the mainstream. I can see the long-term potential in the rows and columns of my databases, and we can all see it in, for example, the promise of Facebook’s Timeline: for the passage of time, data creates a new kind of fullness, if not exactly a new science.

Even now, in certain situations, we can find an excellent proxy, a sort of flash-forward to the possibilities. We can take groups of people at different points in their lives, compare them, and get a rough draft of life’s arc. This approach won’t work with music tastes, for example, because music itself also evolves through time, so the analysis has no control. But there are fixed universals that can support it, and, in the data I have, the nexus of beauty, sex, and age is one of them. Here the possibility already exists to mark milestones, as well as lay bare vanities and vulnerabilities that were perhaps till now just shades of truth. So doing, we will approach a topic that has consumed authors, painters, philosophers, and poets since those vocations existed, perhaps with less art (though there is an art to it), but with a new and glinting precision. As usual, the good stuff lies in the distance between thought and action, and I’ll show you how we find it.

I’ll start with the opinions of women—all the trends below are true
across my sexual data sets, but for specificity’s sake, I’ll use numbers from OkCupid. This table lists, for a woman, the age of men she finds most attractive. If I’ve arranged it unusually, you’ll see in a second why.

<table>
<thead>
<tr>
<th>a woman’s age vs. the age of the men who look best to her</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
</tr>
<tr>
<td>21</td>
</tr>
<tr>
<td>22</td>
</tr>
<tr>
<td>23</td>
</tr>
<tr>
<td>24</td>
</tr>
<tr>
<td>25</td>
</tr>
<tr>
<td>26</td>
</tr>
<tr>
<td>27</td>
</tr>
<tr>
<td>28</td>
</tr>
<tr>
<td>29</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>31</td>
</tr>
<tr>
<td>32</td>
</tr>
<tr>
<td>33</td>
</tr>
<tr>
<td>34</td>
</tr>
<tr>
<td>35</td>
</tr>
<tr>
<td>36</td>
</tr>
<tr>
<td>37</td>
</tr>
<tr>
<td>38</td>
</tr>
<tr>
<td>39</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>41</td>
</tr>
<tr>
<td>42</td>
</tr>
<tr>
<td>43</td>
</tr>
<tr>
<td>44</td>
</tr>
<tr>
<td>45</td>
</tr>
<tr>
<td>46</td>
</tr>
<tr>
<td>47</td>
</tr>
<tr>
<td>48</td>
</tr>
<tr>
<td>49</td>
</tr>
<tr>
<td>50</td>
</tr>
</tbody>
</table>
Reading from the top, we see that twenty- and twenty-one-year-old women prefer twenty-three-year-old guys; twenty-two-year-old women like men who are twenty-four, and so on down through the years to women at fifty, who we see rate forty-six-year-olds the highest. This isn’t survey data, this is data built from tens of millions of preferences expressed in the act of finding a date, and even from just following along the first few entries, the gist of the table is clear: a woman wants a guy to be roughly as old as she is. Pick an age in black under forty, and the number in red is always very close. The broad trend comes through better when I let lateral space reflect the progression of the values in red:
That dotted diagonal is the “age parity” line, where the male and female years would be equal. It’s not a canonical math thing, just something I overlaid as a guide for your eye. Often there is an intrinsic geometry to a situation—it was the first science for a reason—and we’ll take advantage wherever possible. This particular line brings out two
transitions, which coincide with big birthdays. The first pivot point is at thirty, where the trend of the red numbers—the ages of the men—crosses below the line, never to cross back. That’s the data’s way of saying that until thirty, a woman prefers slightly older guys; afterward, she likes them slightly younger. Then at forty, the progression breaks free of the diagonal, going practically straight down for nine years. That is to say, a woman’s tastes appear to hit a wall. Or a man’s looks fall off a cliff, however you want to think about it. If we want to pick the point where a man’s sexual appeal has reached its limit, it’s there: forty. The two perspectives (of the woman doing the rating and of the man being rated) are two halves of a whole. As a woman gets older, her standards evolve, and from the man’s side, the rough 1:1 movement of the red numbers versus the black implies that as he matures, the expectations of his female peers mature as well—practically year-for-year. He gets older, and their viewpoint accommodates him. The wrinkles, the nose hair, the renewed commitment to cargo shorts—these are all somehow satisfactory, or at least offset by other virtues. Compare this to the free fall of scores going the other way, from men to women.
This graph—and it’s practically not even a graph, just a table with a couple columns—makes a statement as stark as its own negative space. A woman’s at her best when she’s in her very early twenties. Period. And really my plot doesn’t show that strongly enough. The four highest-rated
female ages are twenty, twenty-one, twenty-two, and twenty-three for every group of guys but one. You can see the general pattern below, where I’ve overlaid shading for the top two quartiles (that is, top half) of ratings. I’ve also added some female ages as numbers in black on the bottom horizontal to help you navigate:
Again, the geometry speaks: the male pattern runs much deeper than just a preference for twenty-year-olds. And after he hits thirty, the latter half of our age range (that is, women over thirty-five) might as well not exist. Younger is better, and youngest is best of all, and if “over the hill”
means the beginning of a person’s decline, a straight woman is over the hill as soon as she’s old enough to drink.
Of course, another way to put this focus on youth is that males’ expectations never grow up. A fifty-year-old man’s idea of what’s hot is roughly the same as a college kid’s, at least with age as the variable under consideration—if anything, men in their twenties are more willing to date older women. That pocket of middling ratings in the upper right of the plot, that’s your “cougar” bait, basically. Hikers just out enjoying a nice day, then bam.
In a mathematical sense, a man’s age and his sexual aims are independent variables: the former changes while the latter never does. I call this Wooderson’s law, in honor of its most famous proponent, Matthew McConaughey’s character from Dazed and Confused.

Unlike Wooderson himself, what men claim they want is quite different from the private voting data we’ve just seen. The ratings above were submitted without any specific prompt beyond “Judge this person.” But
when you ask men outright to select the ages of women they’re looking for, you get much different results. The gray space below is what men tell us they want when asked:

**a man’s age vs. the age of the women who look best to him**

---

**age range entered as search preference**
Since I don’t think that anyone is intentionally misleading us when they give OkCupid their preferences—there’s little incentive to do that, since all you get then is a site that gives you what you know you don’t want—I see this as a statement of what men imagine they’re supposed to desire, versus what they actually do. The gap between the two ideas just grows over the years, although the tension seems to resolve in a kind of pathetic compromise when it’s time to stop voting and act, as you’ll see. The next plot (the final one of this type we’ll look at) identifies the age with the greatest density of contact attempts. These most-messaged ages are described by the darkest gray squares drifting along the left-hand edge of the larger swath. Those three dark verticals in the graph’s lower half show the jumps in a man’s self-concept as he approaches middle age. You can almost see the gears turning. At forty-four, he’s comfortable approaching a woman as young as thirty-five. Then, one year later … he thinks better of it. While a nine-year age difference is fine, ten years is apparently too much.
It’s this kind of calculated no-man’s-land—the balance between what you want, what you say, and what you do—that real romance has to occupy: no matter how people might vote in private or what they prefer.
in the abstract, there aren’t many fifty-year-old men successfully pursuing twenty-year-old women. For one thing, social conventions work against it. For another, dating requires reciprocity. What one person wants is only half of the equation. When it comes to women seizing the initiative and reaching out to men, because of the female-to-male attraction ratio we saw at the beginning of the chapter (1 year:1 year), plus the nonphysical motivations that push women toward older men—economics, for example—women send more, rather than fewer, messages to a man as he gets older, up until the early thirties. From there, the amount of contact declines, but no faster than the general number of available females itself is shrinking. Think about it like this: imagine you could take a typical twenty-year-old guy, who’s just starting to date as an adult (definition: no solo cups present during at least one of courtship/consummation/breakup), and you could somehow note all the women who would be interested in him. If you could then track the whole lot over time, the main way he’ll lose options from that set is when some of them just stop being single because they’ve paired off with someone else. In fact, his total “interested” pool would actually gain women, because as he gets older, and presumably richer and more successful, those qualities draw younger women in. In any event, his age, of itself, doesn’t hurt him. Over the first two decades of his dating life, as he and the women in his pool mature, the ones who are still available will find him as desirable an option as they did when they were all twenty. If you could do the same thing for a typical woman at twenty, you’d get a different story. Over the years, she, too, would lose men from her pool to things like marriage, but she would also lose options to time itself—as the years passed, fewer and fewer of the remaining single men would find her attractive. Her dating pool is like a can with two holes—it drains on the double.
The number of single men shrinks rapidly by age: per the US Census there are 10 million single men ages twenty to twenty-four, but only 5 million at thirty to thirty-four, and just 3.5 million at forty to forty-four. When you overlay the preference patterns we see above to those shrinking demographics, you can get a sense of how a woman’s real options change over time. For a woman at twenty, this is the actual shape of the dating pool:

![Graph showing the number of men interested by their age, with a peak in their early twenties and a rapid decline with age.](chart.png)

Her peers (guys in their early twenties) form the biggest component, and the numbers slope off rapidly—thirty-year-old men, for example, make up only a small part. They are less likely to actually contact someone so young, despite their privately expressed interest, and in addition many men have already partnered off by that age. By the time the woman is fifty, this is who’s left (and still interested), presented on the same scale. It’s Bridget Jones in charts.
Comparing the areas, for every 100 men interested in that twenty-year-old, there are only 9 looking for someone thirty years older. Here’s the full progression of charts like the two above, rendered from a woman’s perspective for each of the ages twenty to fifty:
So often in my line of work, I’ll see two individuals, both alone but for whatever reason not connecting. In this case, for this facet of the experience, it’s two whole groups of people searching for each other at cross-purposes. Women want men to age with them. And men always head toward youth. A thirty-two-year-old woman will sign up, set her age-preference filters at 28–35, and begin to browse. That thirty-five-year-old man will come along, set his filters to 24–40, and yet rarely contact anyone over twenty-nine. Neither finds what they are looking for. You could say they’re like two ships passing in the night, but that’s not quite right. The men do seem at sea, pulled to some receding horizon. But in my mind I see the women still on solid ground, ashore, just watching them disappear.

This, in my opinion, is what distinguishes a true data visualization from, say, a plain graph or an impressionistic work of art that happens to include numbers. In a visualization, the physical space itself communicates relationships.