Dataclysm by Christian Rudder

Notes by Neekaan Oshidary

Introduction

Christian Rudder, the author, founded OkCupid along with his friends. They approached it with a strong mathematical foundation, as Rudder studied math at Harvard and led to Analytics at OkCupid. Big Data not only provides party conversation, like knowing that Belle and Sebastian is the least black band out there, but also enables seeing big trends and the universal. Today's big public discussion around big data focuses on (1) surveillance and spying and (2) commercial opportunity. Rudder wants to add (3) the focus on the human story it enables. It's trendy today to explore a topic through exemplary stories that illustrate a larger trend, but Dataclysm is a shift to numbers providing the narrative.

Online sites divide vast ineffable processes into quantifiable categories, much like a carrot retaining its shape after being cut on a cutting board. E.g.'s are Facebook likes, Reddit's up/down votes, Amazon's reviews, ratings on dating sites. Interesting insights can emerge from this, e.g. differences in how men rate women's attractiveness and vice versa (men over all rate women as far more attractive, falling on a usual bell curve, whereas women rate men more harshly, confirming that the two sexes have different sexual calculus (men regret sex they didn't have and women sex they had). Dataclysm will discuss data of (1) people connecting, (2) data of division, and (3) data of of people interacting and privacy implications.

The demographic base of Dataclysm's data is not perfect, but nonetheless wide and quite inclusive. A minority of the old and uneducated aren't online, but besides this its fairly representative. This is in contrast to behavioral data from academic institutions that suffers from representative bias of college students being WEIRD (white, educated, industrialized, rich, and democratic). Big data may not change the course of history like internal combustion, but it will change what history *is*, i.e. how history is done through data where everyone can more or less be counted and studied thoroughly through numbers. "Dataclysm" is a play on cataclysm from the Greek "Kataklysmos" from Old Testament for "flood." This deluge of data brings with it a possibility for greater understanding through numbers.

Part 1: What Brings Us Together

Chapter 1: Wooderson's Law

Longitudinal data can show changes across lifespans like evolution of your music preferences. When the Internet is old enough we'll have plenty of longitudinal data. But for now we can also look at people of various age groups to see age-based trends.

Women want a guy to be roughly as old as she is. Until 30, a woman prefers slightly older guys; afterward, she likes them slightly younger. At 40, she then anchors more on younger ages. Women's expectation of men mature as she grows older.

Men despite their age, constantly rate 20-year-olds as most attractive. Hence the title of chapter, "Wooderson's Law," after Matthew Mcconaughey's character in "Dazed and Confused". But despite being most attracted to younger women, older men an select age preference closer to their age. And men message women usually somewhat younger, but not too much, and at older ages there are anchors (late 30's → 30 year old women; Early 40s -44 → 35 year old women, 45+ → 40 year old. There's a balance between what men want, say, and do.

Men have more opportunities and options as they grow older through their 20s, while for women it's downhill as she gets older through her 20s.

Chapter 2: Death by a Thousand Mehs

Having haters makes people want you more. Women with greater variance in their ratings, even if they have lower overall ratings may get far more messages than someone in the 70s percentile. Theory is the guy thinks, "Some people don't like her (less competition), but I like her for her quirks and she'll appreciate that." It's better to be yourself and embrace your quicks than try to fit it in.

Chapter 3: Writing on the Wall

It's a writer's world now. While we no longer have the long, sentimental letters of say the iconic Civil War letters, people write more today than ever before. Despite claims that Twitter and tech have downgraded the knowledge of the English language, data shows for the most part words on Twitter are not too different and instead on average words are longer in length/characters per word. All metrics actually support that people don't "dumb down" their writing on Twitter, but just make it fit more tightly: we thought we'd see clear-cut trees in the forest, but rather we found a bonsai forest. Twitter doesn't so much change language, but rather the study of language. So does Google books, giving rise to the field of "culturometrics", studying language through data mining.

The size of messages people write each other on OkCupiid goes down with the rise of smartphones and apps. People use writing and functions like copy and paste to be efficient in date finding and this is an example of tech being used as a tool for "mass production" but now in the context of dating and human connection.

Chapter 4: You Gotta Be the Glue

Network analysis has been a field for 300 years going back to mathematics of Euler. 40 years ago Milgram had the theory of "six degrees of separation", which has been prove via Facebook. Some findings support the "strength of weak ties" postulated in the 70s and supported today: it says people you don't know very well in your life are the ones who help ideas spread (hence Pixar's architectural designs of the bathrooms located centrally to cause more "collisions" between employees).

Research shows more mutual friends two people share, the stronger their relationships. But a new and data-supported theory however says it's the people you don't know that are important for couples: when couples act as "glue" in between each other's networks and get to know each other's network via their partner, that is the best recipe for staying together. If everyone knew each other in the shared network to begin with, it can lead to competition for attention and "girls' nights out" or cliques. Assimilating each other to each other's network is best.

Chapter 5: There's No Success Like Failure

OkCupid tested disabling photos as part of a "Love is Blind" day promoting their new "Crazy Blind Date" app. The app was a failure since people insisted on knowing what they were getting into, but it led to very interesting findings. No matter which person was better looking or by how much, the percent of people giving the dates a positive rating was constant. People just didn't seem to care about appearance that much when they sat down and met someone. But in usual online dating this becomes a screening tool where people unfortunately turn down others even when in actuality looks wouldn't matter as much.

People tend to overemphasize big splashy things like politics, faith, or looks, but they don't matter as much as they seem. Often "simple" questions like "Do you like scary movies?" or "Have you ever traveled alone to another country?" can have very high predictive success.

Part 2: What Pulls Us Apart

Chapter 6: The Confounding Factor

The compatibility measure on OkCupid shows that race is no predictor of compatibility -- it is equivalent to the Zodiac sign which has no effect at all. But things change when someone's racial *opinions* come into play. Men and women tend to like others of their own race and on OkCupid, black women end up statistically being docked ¾ of a star (out of 5 stars). Asians also see somewhat of a discount, and when someone adds "white" as a component of their racial mix, their ratings on average go up.

If there's love at first sight, there's also dislike at first sight. The racism we see is not due to a few racists or a vocal minority, but rather a pervasive, cultural bias. Rudder says we can't necessarily fault individuals in the "flow" of rapid sifting through profiles, but rather it points to an overall unconscious bias, the "loaded dice" of our culture.

Chapter 7: The Beauty Myth in Apotheosis

Beauty operates on a Richter scale. It's exponential, and far more pronounced for women. More beautiful women get exponentially more messages, more Facebook friend requests, and more interview requests for jobs (even if it's women making the requests and doing the judging). Psychological research had a foundational paper, "What is Beautiful is Good", which established that good-looking people are seen as more intelligent, competent, and trustworthy. This bias can lead to hiring women based on looks, which just statistically guarantees poor performance.

Rudder shares an interesting personal anecdote likely shared by parents: parents want their daughters to be smart, successful, and beautiful -- but not *too* beautiful; just think of the difficult teenage years. But parents never have this worry for their boys.

When social media sites, including OkCupid, increased their profile picture size to give a more modern look, the rich got richer, i.e. the beautiful got more attention.

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

Polla data, surveys, and psychological research often run into the problem of "social desirability": people don't share their true beliefs but rather report what makes them look good. But data from Google searches by Americans provides a window into the darker side of people's biases and racism. Searches for the n-word correlated with Obama's 2008 presidential campaign. Search data shows there's a disconnect with what people publically say and privately believe or say on a topic like race. But we no longer have to wait for political slips to see this: we can see it in the data and Rudder says we're trending toward a better world, but we have a long way to go.

Obama's race cost him 3-5% points in 2008 to people who would have otherwise voted for a white Democrat.

Data ends anecdotes and shows large scale bias and attitudes that need facing. Historians like Howard Zinn who wrote "A People's History of the United States" could draw even more insight into their perceptive analysis if they had this data.

Chapter 9: Days of Rage

The collective discussion and rage that can emerge in response to tweets is the equivalent of a digital "stoning". People only share a target through their hate and the mob mentality diffuses the collective guilt. Rudder describes the response to two joking tweets, one by a 17 year old girl joking about the earth's 2014 year birthday on New Year's Eve and then a colleague of his who in more bad taste shared a joking tweet about Africa and being white, only to be met by a crazily disproportionate response by the angry masses on Twitter.

The internet provides social scientists the opportunity to see how negativity spreads. The spread of rumors and attention grabbing discourse is more about gaining an audience and social status rather than actually targeting the target of the gossip. But for the first time, the Internet gives us metrics for this social status via retweets, followers, likes, etc. Secondly everyone is now a public figure that can be built up and torn down. Importantly, while the negativity is not new, the Internet finally gives the opportunity to constructively respond and harness the chatter.

We can track the rise and decline of movements through types of words and messages being tweeted (e..g knowing Occupy was in decline when people began mentioning negative rather than hopeful words, or triflings like beer theft by fellow protesters). Additionally, we can track the susceptibility of a network or geography to the spread certain ideas. We can see in our written word that we're more politically divided than ever. Liberals, Rudder included, used to think of the disproportionate, sensational outrage of the right (e.g. "War on Christmas"), but incidents like these on Twitter show that the left can be equally self-righteously ignorant and uninformed. The data lets us study these contradictions, negativity, and hate.

Part 3: What Makes Us Who We Are

Chapter 10: Tall for an Asian

You can data mine to see what is most unique and most antithetical to a particular group, like white men. Trends emerge for most common words used only by a specific group and not others: white people differentiate themselves mostly by their eyes and hair (“my blue eyes”), Asians by their country of origin, Latinos by their music. This data isn’t available from say Google auto-complete but only emerges when you plot the data of different groups against each other. Antitheses (least used by group compared to other groups) are interesting and comical at times: Latino antithesis crops up stereotypical images of the white, “corn-fed” Midwesterner, where there’s often very little Latino presence. Asians show very little misspellings, or mentions of working class words or “underachievement” words like “single father” -- and then there’s “6’4”. Black women are most antithetical to “tanning” and in music “belle and sebastian”, “simon and garfunkel”, and “magnetic fields.” Such data can’t be discovered via Google Trends, but only by a specially designed algorithm.

When you compare differences between sexes on Twitter, you get expected differences: for men, there’s “bro” references, sports references (“fantasy football,” “iverson”), videogames (“ps4”), while women have “mani pedi”, “girls day”, “dress shopping”. But this is all the result of the algorithm, which is looking to find these differences. Consistently as a whole there are far more similarities than differences between the sexes. In the end, Rudder says, cultural differences, despite being comical sometimes, make the world a richer place.

Chapter 11: Ever Fallen in Love?

No one knows the exact % of gay people, but it’s probably around 5%. This is the % of people searching for gay porn on Google and it doesn’t matter on geography or how accepting a place is of homosexuality. It goes to show that when a state has a very low % of self reported gays (e.g. North Dakota), people are choosing to misreport. OkCupid data shows little difference overall in *how* gay vs. straight people love (there are same rates of drug use, racial prejudice, horniness); the difference is just in *who* they love.

When you look at most common words on OkCupid by sex and orientation, straight men and women singlemindedly focus on describing their (potential) partner and who they want. Lesbian women similarly describe the relation and person they want, while gay men reference pop culture more and less on the person and family.

Bisexuality is often seen as a “marginal” group within a minority: gays often see them as those who haven’t fully accepted their homosexuality. A 2005 psychological study stirred a controversy and a strong response when it claimed that bisexual men were actually almost always sexually stimulated by one sex, supporting the idea of true homosexuality or just curiosity. This is supported in OkCupid data where those identifying as bisexual often mostly message one sex. But Rudder says that people should be free to describe themselves as they choose, and this doesn't need to perfectly correspond to behavior, and this was the criticism of the psych study: sexuality and relationships don't fit neatly into a researcher's binary metrics of sexual stimulation.

Bisexuality for women is different because it is culturally mainstream and often sells.

Rudder says having more individuals coming out -- both celebrities and the nameless -- something happening since the late 19th Century -- is moving us toward a more open world where we'll no longer have to "guess" at the the numbers.

Chapter 12: Know Your Place

Big data and the Internet has created new geographies and interesting mappings. For example, you can map most common places people cite in Missed Connections on Craigslist -- it's Walmart for much of country. But Twitter for instance introduces the possibility of radically new mappings like knowing the real time "emotional epicenter" of the shock wave response of an earthquake; this could have vast implications like knowing where to offer aid after a disaster. This new data is in its infancy so we don't have longitudinal data, but Rudder describes how valuable a tool like this could have been to observe changes say during periods of geographical and cultural conquest.

Some data can be explained by other factors -- opinions on flag burning reflect urban vs. rural divides, same-sex searches on Google just reflect where people live most, how often you take a shower corresponds to how hot the weather is, etc.

Then there are new geographies of community with Reddit being the best example. Reddit and its subreddits provide a vast geography of interests and communities coming together online and even leading to support offline, e.g. inviting a lonely person who posted on Reddit over for Thanksgiving. While the previous examples of Twitter showcase the worst of the new mob-like darkness of the Internet, Reddit is brings out the Internet's better angels.

Lastly, throughout history groups of people have made large scale migrations in hopes of finding somewhere better and today the Internet and sites like Facebook can chart these movements of people, often moving from rural areas to cities in hope of financial opportunities.

Chapter 13: Our Brand Could Be Your Life

In the late 90s two different marketing speakers introduced the idea of people being their own brand and soon thereafter "personal branding" took off. Personal branding often falls into the category of empowerment seminars ending with their aims of wealth and power, but today many smart, respectable people refer to their personal brand. With Twitter and other services you now have a potential audience and way to reach many with your brand. But its often hard to gain many followers and Twitter has a huge follower disparity, far higher than wealth disparity in the US: 1% of accounts have 72% percent of the followers, and 0.1% of accounts have just over half.

The want for followers is simple; the more popular you seem, the more popular you become -- which leads to many people (very often politicians) purchasing followers on Twitter. This leads to a chase of empty metrics for likes, favorites, followers, responses, interactions, and user engagement stats like counts, totals, and badges, which are all designed to get users to use these services more.

Klout reduces your social influence to a number and many were outraged when Salesforce required a Klout score of 35 on a job application. Reducing people to a number requires larger discussion. Algorithms are crude and this is how Big Data works, but Rudder stresses we shouldn't reduce an *individual* to a *number* -- a human can't be captured by a single number or metric. *But* when you take these little pieces of *many* people and aggregate it, you get a better picture of the *whole*, of all of *us*. The danger especially lies in people being dehumanized and becoming numbers in the means of building someone's brand.

Rudder concludes that while corporatism has invaded our online selves, some aspect of our humanity will be out of reach and ultimately numbers won't deny us our humanity, but the calculated decision to stop being human will.

Ch 14: Breadcrumbs

The predictive power of Big Data is only getting stronger. Just by your Facebook likes alone, software can relatively accurately predict sexual orientation, race (Caucasian or African American), gender, political affiliation (Republican or Democrat), drug use, and even if your parents divorced before you were 21. Imagine the darker side what this could predict if we had longitudinal data since childhood. You will only increasingly be seen with security cameras, satellites, drones, and whatever else the government uses. And then there's masterminds like Acxiom who have everything from bank and credit card records, retail history, and all kinds of online behavior on you, tracking you down with great accuracy.

One could also argue the surveillance state has led to counter-terrorism, both in preventing attacks (no major terrorist attacks in the US since 9/11) and in solving the mystery of the attacks (such as in the Boston Marathon bombing and London subway bombings). No doubt, those in the NSA doing the spying are some of the most brilliant math geniuses of our time, but we can only hope, Rudder says, like Einstein and Feynman, they temper the sheer inhumanly powerful scope of their work with a farsighted view of humanity. Government projects like PRISM work off metadata -- they only look at higher level data and patterns and only with a court order is content surfaced. But people leave trail of breadcrumbs that can easily be followed, including GPS coordinates in photos and our digital trails of *wheres* and *whens*, opening the doors to your demographic profile.

While some may be reticent to share their data, most are blasé when it come to privacy. Tech keeps pushing the boundaries on privacy and many just acclimate to the changes. People give away all kinds of personal data now: take menstruation apps that can easily determine when someone is pregnant, over exercising, getting older, or having unprotected sex. It's important when handling user data to always anonymize the data by viewing it in aggregate and to never reveal personally identifiable information (PPI).

Rudder says, while some claim such data should only be used for "demonstrated valid use", it's hard to justify this because we only learn what's possible and stumble upon the beneficial discoveries by exploring the data first. Google Flu has helped prevent and minimize disease and data scientists from tech and research universities were even able to determine unreported drug side effects before the FDA could, so these point to the very real benefits that come from giving up some privacy. Other projects like Google's Constitute help new countries write constitutions by looking at what has and has not worked with previous constitutions. And the possibilities only multiply with "social physics," where we get all kinds of detailed data on say an entire city, as is being done in Trento, Italy, where they've begun to track how much families spend, socialize, and even what preschools and doctors have the highest retention rates for families.

Rudder is doubtful regulating tech and privacy will work because the laws will already be outdated "as soon as the ink is dry", and most people won't use privacy controls you give them. Some proposals seem more tractionable like giving people the controls to delete your data from a service or move it elsewhere (copy and paste); this should be enabled, even if not everyone uses it. And other proposals consider compensation whenever you data is sold by services, but one could argue we're already being compensated by getting to use technologies like Google and Facebook for free. Unfortunately though, cultures and generations define privacy differently, and this whole debate may soon become an anachronism.

It's hard to fully say when "the waters are still churning," and we'll only know once the waters have settled. Our culture today often frames technology via the mythos of gods and Tech Titans, but in fact we are all flawed, mortal, and human, and we'll either be drowned or uplifted by the flood. Rudder concludes with a twist on Tennyson's Ulysses saying we must find *our* truth amidst this epic battle with technology, "to strive, to seek, to find, but then, always, to yield."