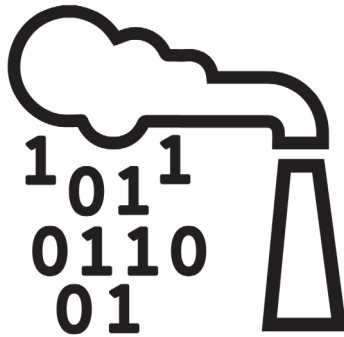


DATA POLLUTION



What's the Trend?

As we create more methods for quantifying the world around us, data gets manipulated, contaminated and sabotaged, making it harder to separate true insights from useless noise.

I don't consider myself an over-sharer on social media. At times weeks will go by before I remember to share a post on Facebook. On occasion, I will log in and get a desperate reminder from Facebook: "Did you know your followers haven't heard from you in 2 weeks? Post something today!"

Over the past decade, I have had a love-hate relationship with the platform. When I first joined, I loved the connections and essentially said yes to every friend request. Over the course of the next two years, that quickly led me to approach 5000 "friends" (the limit at that time) and my newsfeed was filled with stories from people I didn't care about or barely knew.

Frustrated with this lack of connection, I decided to declare "Facebook

bankruptcy” and shut my entire account down. I learned how to convert that account, to be a “Public Page” and I started over with a new private personal account which I started to use often.

In the process of writing this chapter, I started to wonder what Facebook had learned about me over those years of my constantly feeding it pictures, interactions, videos and moments of my life. So I decided to do something many people don’t realize they can do: I requested to download all of my data from Facebook.

The file I received back was 118 MB. It contained 1,732 items, including every photo, video and message I had shared during that time. While I saw this data and immediately looked at it in terms of messages, photos, videos and interactions, Facebook describes all this information as *data* – and they are not the only ones collecting a staggering amount of it.

In September of 2017, journalist Judith Duportail shared a story of requesting four years of her personal data from the relationship platform Tinder. What she got back was 800 pages of startlingly revealing personal information on everything from her racial preferences in men to transcripts of every conversation she had with a match—sometimes revealing intimate personal details.¹

The Hidden Side of Data Collection

This is chapter about data—but before I get too far, we need to have a shared understanding about what data really is. Often people think of data in terms of likes, views or conversion rates. The *numbers* are only one part of the data that is collected about each of us. The other side includes conversations, content, behaviors and intent. It includes all kinds of information that you might have assumed no one was tracking – until you learn they are. Here’s an example: on photo sharing site Flickr, most images are uploaded along with all sorts of meta data like camera exposure times and GPS locations for photos. Without realizing it, people are publicly sharing their exact whereabouts at a particular time.

An inevitable side effect of all this data collection is the fact that some of it will be questionably useful and hard to interpret. How can anyone separate the bad data – or misused data -- from the rest? The stakes are high, as IBM recently estimated that poor quality data costs businesses \$3.1 *trillion* annually in the US alone.²

The problem as it exists today, is two sided. On one side are unwilling consumers who reluctantly or naively shared their data without realizing the implications. On the other side are businesses struggling to leverage all this data in a way that doesn't annoy but rather delights. We call this problem *Data Pollution*, and as you'll learn in this chapter, it is a problem that affects us all.

To understand and perhaps solve this problem, my team and I took a deeper look at the many ways that data can be “polluted,” and searched for the root causes of this phenomenon so we could develop a model and insights for how to do better.

This research led us to define five elements of *Data Pollution*, but before I share them here – it is important to understand the sources for all of this data.

The 4 Sources For Data

Data is not measured by companies alone. To get a full picture of the data that is available for most organizations to analyze and use, we must also add data from three more sources. Here are all four, along with a short description of each:

1. **Open Data** – Typically collected by governments and NGOs on a wide variety of topics such as demographic shifts, labor and industry statistics, weather data, immigration data, etc.
2. **Scientific Data** – A broad category which describes data gathered by scientists and academic researchers as part of studies which are generally published in medical journals or online.
3. **Big Data** – Collected by an organization, typically for commercial purposes. It includes customer purchasing data, online behavioral data, customer loyalty program data and social media profile data.

4. **Small Data** – Enabled by the Internet of Things, this refers to data that individuals are collecting on their own through devices like wearables. Sometimes this type of data is “co-owned” with a service provider, but in most cases the consumer has the right to request it, use it and even ask for it to be deleted.

Now that we have looked at the sources of data, let’s consider the different ways that this data gets polluted – and what you can do about it.

THE 5 TYPES OF DATA POLLUTION

1. **Data Overflow** – When too much data is captured leaving organizations confused about what to focus on or prioritize.
2. **Data Manipulation** – When the results and insights of data are twisted to selectively support biased arguments.
3. **Data Sabotage** – When people intentionally share incorrect or incomplete information with the intent to cause harm.
4. **Data Contamination** – When data is collected from multiple sources, mixed together, and reclaimed or deleted by creators.
5. **Data Expiration** – When data is not updated as frequently as it should be and therefore loses its value because it is not current.

Why Data Overflows

Our 2016 Non-Obvious Trend Report, featured a trend we called *Data Overflow* to describe the fact that many organizations and institutions were becoming buried in the sheer volume of all the analytics they had access to. It is a sad irony that among business leaders, the idea of data-driven decision making is viewed as a magic tactic capable of addressing every problem. It rarely achieves this ideal because so few of those leaders know how to truly use data correctly.

Part of the reason connecting the dots is so difficult is because

non-scientists often lack the willingness to learn how to process the data they are given. For example, many website marketing teams focus blindly on trying to get people to spend more time on their site and reducing the “bounce rate” (how many of your visitors leave your site immediately after arriving).

Imagine you owned a restaurant website and a potential customer wanted to find your address to come in and dine. If they visit your site and find your address quickly, they will “bounce” off to Google to get directions and hopefully become a paying diner (the highest conversion you can get). Now imagine if that potential customer looked around for thirty seconds (an eternity online) to try and find your address, failed and left your site, frustrated.

That person most likely did not become a customer, yet if you only measure “time on site,” that second experience looks like a success while the first one looks like a failure. This is the problem with using data incorrectly, and it happens with far more frequency because of *Data Pollution*.

The Overquantified Self

Another issue of *Data Overflow* is how it affects the industries where we desperately wish data was more useful but doesn’t yet seem to be. Healthcare is a perfect example of this. Four years ago when I published a book about the future of healthcare called *ePatient 2015* with my co-author Fard Johnmar, we first introduced the trend *Overquantified Self*. It described the gap between all the excitement in healthcare for tools that could track our health and the reality that most of these tools were simply generating disconnected data that wasn’t connected in a meaningful way to actual healthcare. Weill Cornell Medicine Internist Dr. Fred N. Pelzman describes this issue as “drowning in data” and writes about how he struggles to make sense of it:

“Every morning when I log onto our electronic health record, one group of messages are from patients with their self-recorded health information

*data that they've sent to me for review. Here's my home blood pressure over the past 24 hours. Here are my fingerstick glucose readings over the past 24 hours. Here's my weight over the past 24 hours. Here is my depression score over the past 24 hours. A lot of data -- oftentimes a lot of data that I don't quite know how to react to, what actions I'm supposed to take."*³

Data Overflow is also prevalent in the Open Data that is publicly dumped by governments into the marketplace in huge quantities. One report from the index that tracks open data reported that of the 1 million datasets which have been made available by governments worldwide, less than 7% are in a machine-readable format – rendering them almost useless.

The bottom line is, we collect too much data and lack the tools or expertise to analyze it properly. This is a bad combination. Unfortunately, this leads directly to another problem: in the absence of any real insights the conclusions from data are manipulated in order to promote an agenda.

A Universe of Manipulated Data

When the results of a *scientific* study are under suspicion, the globally recognized standard test for validity is to repeat the experiment. If the results hold and the study is replicable, it passes the test. Unfortunately, less than half of all research papers would pass this standard of proof.⁴

Outside the scientific world this figure is likely even lower. The problem is, there are so many easy ways that data can be manipulated at the time of collection – for example by asking leading questions or isolating participants.

The report referenced above also found that “the greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true.” For a perfect example of this, consider the biased and now widely debunked studies into the nutritional effects of sugar published in the 1960s and referenced in the opening story of Chapter 6.

Research has frequently been published with this sort of ulterior motive and it can take decades to effectively refute. As a result, this form of *Data Pollution* is widespread, long lasting and very difficult to remove.

Why Everyone Has a Fake Email Address

Whoever owns the email address bob@bob.com probably deserves an apology from me.

For the last twenty years, poor Bob has provided me with my go-to email address to sign up for free wi-fi, appease particularly insistent retail employees at checkouts and fill out long waivers at kids' birthday parties.

Usually I do this to save time, but sometimes I do it to exact my revenge on aggressive and annoying marketers by providing them with intentionally bad data. As any online marketer will already tell you, I am hardly the only one to take this low road.

The fact is, people take sadistic amounts of delight in providing intentionally bad data to organizations who dare to ask them for it without earning it or providing a good reason for why they need it.

This type of intentional *Data Sabotage* can feel like a justified form of payback, but not all Data Sabotage comes from revenge-seeking consumers. Sometimes there are profiteers at work.

Fake Profiles And Data Hacking Bots

At the 2017 Strata Data Conference in New York, Data & Society Founder and President danah boyd⁶ shared a dire warning about the fact that data is being actively sabotaged by groups of “adversarial actors” who are focused on exposing the flaws of our overreliance on it. Some of them are motivated by the recognition that comes from being able to hack the data or sabotage existing systems.

The implications, according to boyd, are that we need to reconsider how to secure all of this data:

“If we believe that data can and should be used to inform people and fuel technology, we need to start building the infrastructure necessary to limit the corruption and abuse of that data—and grapple with how biased and problematic data might work its way into technology and, through that,

into the foundations of our society.”⁷

A good place to start may be the problematic data which is being continually created on the world’s largest social media platforms through profiles that are not attached to real people.

In April 2017, when Facebook conducted a purge of millions of fake accounts, large brands saw significant drops in their follower counts, proving that much of their “audiences” consisted of fake accounts. USA Today was the brand that felt the largest impact, losing over six million followers (more than a third of their overall audience) overnight.⁸

It is hard to tell exactly how many social media accounts might be fake when many platforms are reluctant to share those numbers, but some independent researchers estimate that at least 15% of *all* social media profiles may be autogenerated without a real person behind them.⁹ If you asked *Daily Beast* reporter Joseph Cox, however, he would probably tell you that number seems low.

Cox had his account suspended on Twitter because a small army of (most likely) bots had started to follow him at the same time. Though it has yet to be proven, he suspected that the bots automatically followed him due to algorithms that identified him as the author of an article covering the rise of bots originating in Russia. Because of this, the fake accounts mounted a retaliatory attack and caused his account to be suspended.

These fake profiles created by scammers, bots, or unscrupulous humans are a symptom of a larger problem that is fueled by (but not limited to) the Internet. Data has become contaminated.

Data Contamination

I follow my 13-year-old son on Instagram and last month I noticed something strange on his account. After spending weeks traveling on the West Coast over the summer and sharing lots of images, his account was down to just four. Worried that he had lost some of his content, I asked him about it.

“I deleted them,” he told me.

“Why?” I wondered.

“I didn’t need them.”

Most of his posts get about ten times the number of likes that mine do, but even then, at a certain point those older images just didn’t matter to him anymore. It turns out, he’s not alone. A story in the *Washington Post* reported how teenagers were “curating” their accounts on Instagram and deleting old photos that either did not get that many likes or that they no longer wanted to have on their profiles anymore.

I use Instagram and other social media platforms as a gallery of my life to be saved, shared and occasionally to reminisce through. Teenagers today are using it as an article of clothing: something to be worn today and changed tomorrow. In fact, this deleting was so prevalent that Instagram introduced a feature to allow users to archive and save those old images instead of deleting them.¹⁰

When users delete old posts, it can reduce page views and engagement time, as well as skew the data which most sites are trying to monetize.

It’s no wonder the team at Instagram are desperate to change this behavior. The less accurate their data, the less opportunities they have to sell advertising.

This is *Data Contamination* and it’s something you can expect to hear plenty more about in the coming year.

Why It Matters

When researching this trend and going through the Haystack Method for naming it, the brainstorm was particularly short because *Data Pollution* immediately felt like the perfect name. Like clean air in the atmosphere, data is fundamental to how businesses are run—and yet it is under attack from negative forces like manipulation, sabotage, and contamination. We intentionally selected inflammatory words to describe the nature of this attack, because it relays the gravity of the situation.

In a world where we can no longer trust the data that we are collecting

or use it to make informed decisions, we must operate on blind instinct and hope even more blindly for luck. Understanding the severity of the problem and how it is happening is the first step toward a solution. Building and implementing better tools for analytics will help as well. The most helpful thing, however, will be the one thing that many of us are reluctant to do: spend more time learning how to become data literate ourselves. The only way to fight pollution (in any form) is to stop making the same mistakes and to get better educated about what we need to do differently.

How to Use This Trend

- ✓ **Find your hero metrics.** Rather than trying to make sense of large batches of irrelevant data, one trick smart executives use all the time is finding a small group of “hero metrics” that help them measure what is working and what isn’t. Revenue per store, emails forwarded, number of social media shares within the first hour of posting ... these are all examples of metrics that could be used to get a more nuanced view of what is happening in your business, and what to do about it.
- ✓ **Audit your data for more than numbers.** The best way to realize value from your data is to better understand what you have and what is missing. For that purpose, conducting a detailed audit is a valuable place to start, but this has to look at more than just the numbers. Instead the best audits will review what is being collected in the first place, what sources are being used and conduct a gap analysis of what is missing in order to use data more effectively.