DO YOU KNOW WHO YOU ARE?
Caitlyn Zawideh

| Home | Table of Contents | About |

About this project

"Do You Know Who You Are?" is a self-directed research project completed as part of the capstone course for the Sweetland Minor in Writing. I chose to pursue this project because I'm interested in the relationship between people and our technology. As data and algorithms increasingly shape the way we see the world and take in information, it's important that we take time to examine how our world is being shaped, by who, and why.

Thank you to T Hetzel, John Cheney-Lippold, and Simone Sessolo for your guidance, support, and encouragement over the course of the semester from the early planning stages to the final details.

About Me

Caitlyn Zawideh is graduating from the University of Michigan in December 2020 with BA in English Language and Literature and minors in Writing, Computer Science, and Digital Studies.

Contact: caitlynzawideh@gmail.com

Click me.

DO YOU KNOW WHO YOU ARE?
Caitlyn Zawideh

Home   Table of Contents   About

IN SEARCH
OF   DATA
IDENTITY

My sister is so convinced our phones are listening that she will censor what she says when she's near hers so as not to alter the content she sees on TikTok. I'll ask her, for example, if she's seen the election coverage from the Washington Post account on her feed, and she will pick up her phone and dictate her preferences into the microphone in case TikTok is listening: "Harry Styles, MCAT, Pre-med. No Washington Post. Harry Styles, MCAT, Pre-med." She is fully convinced this ritual has an effect on what type of content will appear on her algorithmically curated feed.

While there is no conclusive evidence that our technology is listening in on our conversations to help tech giants deliver personalized content, it is true that TikTok and other social media platforms collect massive amounts of information about their users in order to populate our feeds. Everything you do online, on your phone, in almost any digital capacity produces data. That data is then used to inform the algorithms that are responsible for targeted ads as well as placing the content you see on your newsfeeds, timelines, and search results.

The problems that have materialized from this practice have become a topic of increasing public concern in recent years. From the spread of misinformation and polarization over social media feeds, to security breaches, to the near banning of TikTok in the US, data is at the center of national discourse about privacy, truth, and power. Technology ethics advocate and founder of the Center for Humane Technology Tristan Harris likens the unchecked power and influence of the companies collecting our data to a threat on par with climate change [1].

I find it difficult to wrap my mind around the scale and severity of these problems while liking photos of my friend's new puppy on Instagram.

For all the media coverage and conversations about data and algorithmic recommendations and a need for transparency, most people, myself included, don't know what our data literally looks like or how we are understood by these platforms through our data. We only see the outcomes.

of transparency from any organization that collects data from people in the EU, has spurred platforms to provide information about how they're using our data [2]. Companies often make those features available to American users as well. For example, when I saw an ad on Facebook for sweatshirts adorned with the likeness of Supreme Court Justice Ruth Bader Ginsburg alongside the words "Notorious RBG," I had the option to click "Why am I seeing this ad?" The explanation provided was that the advertiser wanted to reach people that "Facebook thinks are interested in Shopping and fashion and Hoodie," as well as females ages 18 to 30 whose primary location is the United States. The explanation is also accompanied by a disclaimer that "there could be more factors not listed here."
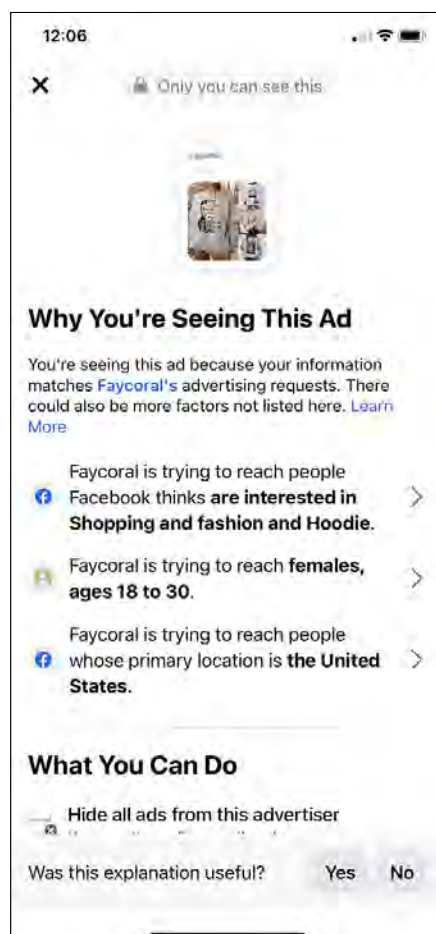
Facebook. Why you're seeing this ad.

Two of those factors, my age and location, are information I have provided to Facebook, but the third – the interest in shopping, fashion, and hoodies is something Facebook has inferred about me. With this explanation, I can understand why I was given an ad for a sweatshirt, but I have no information about why I got the ad for *this particular* sweatshirt. To be categorized as someone who likes Ruth Bader Ginsburg enough to buy items with her face on them and wear them feels like a very specific grouping. What has Facebook inferred about me to come to the conclusion that I am the right audience for it? Who does Facebook think I am?

data identity look like? I know there isn't a simple answer. The data-based approximations of who I am are constantly changing as I generate more information about myself through my digital habits. The answer will also be different depending on who is collecting my data. Facebook's definition of who is interested in "shopping and fashions and hoodie" is likely different than Google's.

Under GDPR, platforms like Facebook are required to allow users to download and view their data, a feature also available to American users, which means I can make an attempt to answer my questions. What does my data identity look like when I combine my information from many different platforms? How complete can an identity constructed through data really be? How much control do I have over my data as an individual user?

To find out, I've downloaded my personal data reports from the platforms I use the most (Facebook, Google, Snapchat, Twitter, Spotify, and TikTok) and consolidated my findings.

< Home

Facebook >

DO YOU KNOW WHO YOU ARE?
Caitlyn Zawideh

Home	Table of Contents	About

F A C E B O O K

Access Your Information

> "Last week showed how much more work we need to do to enforce our policies
> and help people understand how Facebook works and the choices they have
> over their data."
> - Erin Egan, VP and Chief Privacy Officer, Policy and Ashlie Beringer, VP and
>   Deputy General Counsel at Facebook. March 28, 2018 [3].

The "last week" casually mentioned in the opening paragraph of this March 2018 blog
post from Facebook is a reference to the more than 50 million profiles harvested by the
data analytics firm Cambridge Analytica, tied to Trump's 2016 presidential campaign,
one of the biggest known privacy breaches Facebook has had to date.

In an attempt to regain the trust of their users, Facebook announced the Access Your
Information tool, which is meant to simplify the process of downloading your Facebook
data. The download includes far from all of your information. Facebook provides limited
information about the advertisers who have shared user data with Facebook, only the
name of the company and no way to contact them. Additionally, Facebook stores data for
varying amounts of time, so older data is less likely to appear in the report.

As promised in their blog post, it's relatively easy to download my Facebook data. After
locating the Access Your Information tool in the website's privacy settings, I was asked to
verify my account password and thirty minutes later I was able to download a large file
with eleven years worth of information. It's divided into almost thirty folders, but the
content of all of them is displayed via a single, user-friendly html file.

For Facebook, winning back public trust means giving users the impression that they are
in control of their information, and that intention is present in the design of the interface
they created to show us our data beginning with the message stamped at the bottom of
every page of the report:

Literally speaking, this statement refers to the date and time I requested to download my information. However, the language brings to light a question of agency. The word "generating" seems to imply that the data was created by me at the time I requested it, but the data already existed, and it exists because Facebook has been collecting it since I opened my account in 2009. Even then, "collecting" suggests that data is out there in the world, naturally occuring, waiting to be harvested. In reality, data is the way we break down continuous occurrences into small, discrete parts. What gets broken down and how is up to the data collectors. As with every other platform I will be looking at, it's a choice that Facebook gets to make, and that is what I'm keeping in mind as I scroll through eleven years worth of information generated by those choices.

Note: Hovering over some highlighted sections will reveal data.
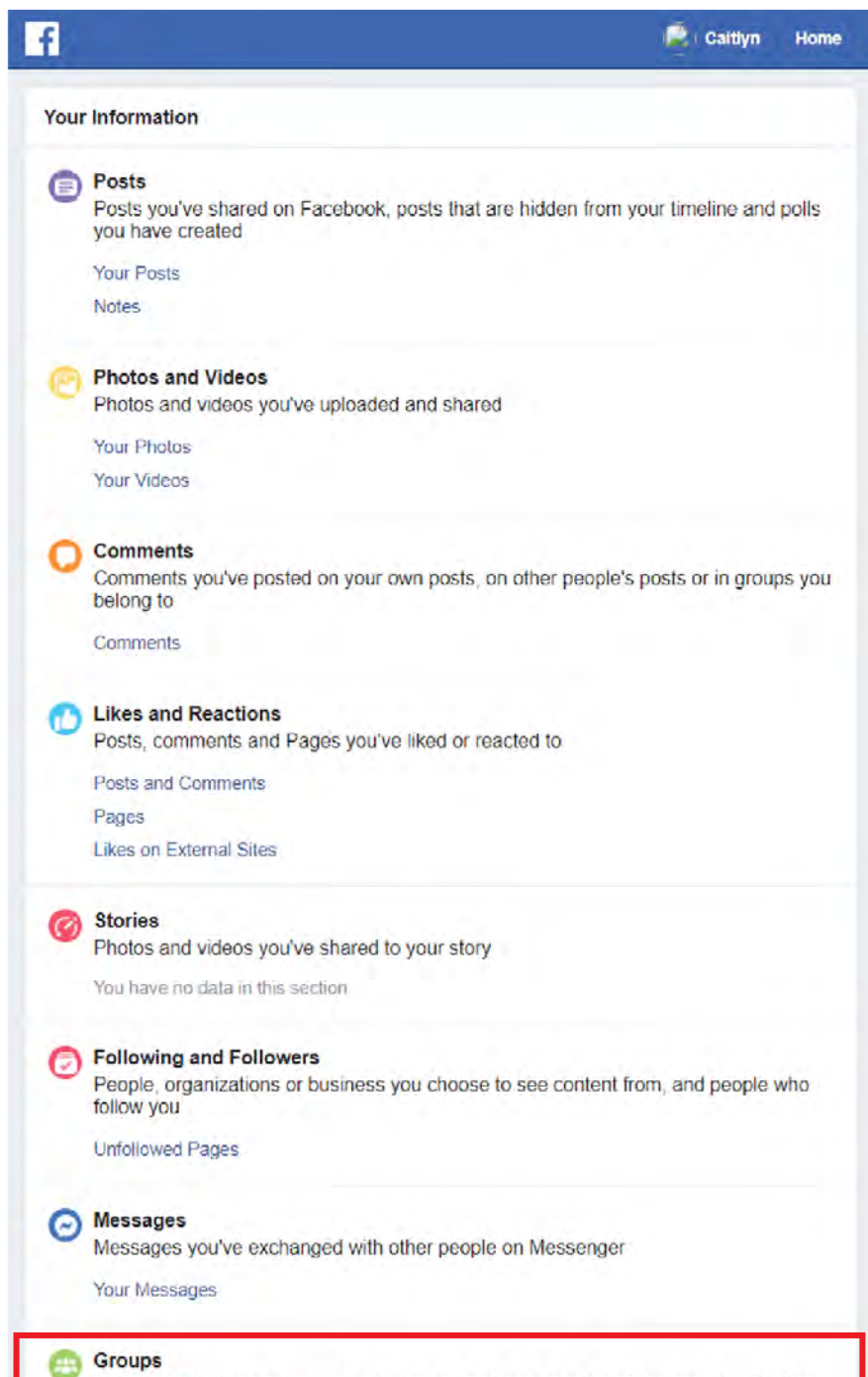
The Data

Facebook has kept a log of every post, picture, like, and comment, even the ones that I have since hidden from my timeline, but not deleted. It's all information I knowingly gave to the platform. Though at the time, I was thinking of it as content visible to my Facebook friends, rather than an entry into a personal database.

Every post is marked with the date, time, and page I posted it on. In recent years, my posts mostly consist of happy birthdays to my relatives and postings on housing pages in search of a summer subletter for my apartment.

Further back, around the time I would have been in middle school, are more frequent postings: inside jokes between old friends I no longer remember the punchline to, snow day predictions for winter 2011, chain messages threatening a haunting if I don't pass them on. I don't remember posting as often as the data shows me I have.

The record of my joining and leaving

Your Group Membership Activity

Your Posts and Comments in Groups

Events
Your responses to events and a list of the events you've created

Your Events

Your Event Responses

Event Invitations

Profile Information
Your contact information, information in your profile's About section, your life events, hobbies and music

Profile Information

Profile Update History

Pages
Pages you are the admin of, and pages you've recommended

Your Pages

Marketplace
Your activity on Marketplace

Items Sold

Payment History
A history of payments you've made through Facebook

Payment History

Saved Items and Collections
A list of the posts you've saved, and your activity within collections

You have no data in this section

Your Places
A list of places you've created

Places You've Created

Apps and Websites
Apps and websites you log into using Facebook and apps you admin

Apps and Websites

Posts From Apps and Websites

Other Activity
Activity associated with your account, such as Pokes given and received

Pokes

Polls You Voted On

Fundraisers Donated To

Support Correspondences

Facebook Gaming
Your profile for Facebook Gaming

Instant Games

Interactions
Actions you've taken on Facebook

Events

"You became of member of Class of 2016" marks my high school graduation.

"You became a member of Alice Zawideh Pictures." My dad created this group when his mom died. He wanted the extended family to be able consolidate all their photos of her in one place.

"You stopped being a member of MJR Waterford Coworkers" August 25, 2017. 9:52 PM. -- Twenty-two minutes after my final shift scooping popcorn at my hometown movie theater, my job the summer before I transferred schools.

Even the beginning of the pandemic and the move to online classes is noted by my joining a group called "Zoom Memes for Self Quarenteens" on March 16, 2020 at 11:54 PM EST.

Records of discontinued features like the poke live on in data.

Groups

Archive
Items you have moved to archive

You have no data in this section

Short Videos
Your activity with short videos on Facebook

You have no data in this section

Campus
Your campus activity and data

You have no data in this section

Information About You

Ads and Businesses
Ad topics that are relevant to you, advertisers who have collected information directly from you, information you've submitted to advertisers and your interactions with businesses and organizations you visit off of Facebook.

Ads Interests

Advertisers Who Uploaded a Contact List With Your Information

Your Off-Facebook Activity

Information You've Submitted to Advertisers

Advertisers You've Interacted With

Search History
A history of your searches on Facebook

Your Search History

Location
Information related to your location

Primary Location

Timezone

About You
Information associated with your Facebook account

Friend Peer Group

Face Recognition

Your Address Books

Notifications

Messenger

Visited

Viewed

Preferences

Security and Login Information
A history of your logins, logouts, periods of time that you've been active on Facebook and the devices you use to access Facebook.

Where You're Logged In

Account Activity

Logins and Logouts

Administrative Records

Used IP Addresses

Datr Cookie Info

Facebook's separation of my data into two groups, "Your Information" and "Information About You," is telling. There's a clear distinction between data I've provided directly, as in the previous section, and the inferences that have been made based on that data. One type is decidedly mine. I have the ability to add to or delete it at any time.

The other type was created by and for Facebook. It's the "about" data that tells me a bit more about who Facebook thinks I am.

Facebook provides short explanations of each of the subgroups under the About You section. For example, Friend Peer Group, which has been labeled University is a "description of the life stage of my friends."

One of the more puzzling inclusions is the Facial Recognition data, which I expected to be a file containing all the images on facebook containing my face. Instead, it's a totally indiscernible string of letters and numbers and no way to understand their meaning. The terms "raw data," "threshold," and "example count" don't do much in terms of explanation.
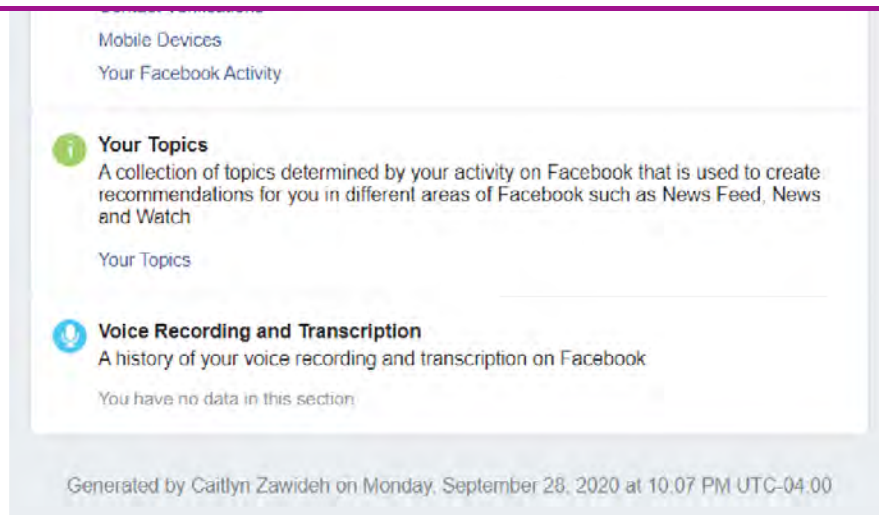
Outbreaks, and Writing, are not hard to understand. I listed Microsoft as my employer on my profile last year and even made a rare post about it - so it makes sense that it would take on disproportionate weight in my categorization. Additionally, most of my current Facebook activity is centered around running the page of a student literary journal. Since the name of the journal is Writer to Writer, it's not an insane logical leap for Facebook to conclude that I am in fact interested in Writing.

Disease Outbreaks is self-explanatory, given we're living in the midst of a worldwide pandemic. Of course my impulsive checks on news and numbers on coronavirus would translate into an interest in Disease Outbreaks. I think everyone has an interest in Disease Outbreaks these days.

As for the Musical Theater categorization, I had a short-lived enthusiasm for Hamilton in 2016 like the rest of the world, but that hardly qualifies the entire genre of Musical Theater one of my top interests, so I'm at a loss as to why that topic is there and featured so prominently.

While I'm not much of a baker or even a fan of desserts, I could attribute all the food, dessert, and baking-related categorization to Tasty cooking videos from Buzzfeed, which saw peak popularity a few years ago and were constantly on my feed during that time period, but for content I haven't thought about or interacted with in years to completely dominate Facebook's understanding of me in this context doesn't make a lot of sense.

Takeaways
Scrolling through old posts, comments, and even my group membership activity, I'm slightly tempted to veer into nostalgia and view my Facebook data download as a type of time capsule. However, findings like my facial recognition data and largely off-base interest topics remind me that this is not an exercise in reminiscing. Really, my group membership activity is just a list of dates, and the idea that they map onto certain significant life events is a meaning I have assigned to the data, knowing its context, and not a meaning the data creates on its own. In comparison, my facial recognition data is totally meaningless to me since I have no knowledge of what the list of letters and numbers represents.

Additionally, I reacted to my interest categories based on my own understanding of what they might mean, but I don't actually know how Facebook has defined "Musical Theatre" as a category, how I fit into it, or how being grouped into that category has affected my feed - which does not regularly feature musical-theatre-related content.

At first glance, the Access Your Information tool seems to accomplish Facebook's goal of

< Introduction

Google >

DO YOU KNOW WHO YOU ARE?
Caitlyn Zawideh

Home	Table of Contents	About

GOOGLE

caitlynz98

My longest standing gmail account is the one I made in 2010. I set it up on my parent's clunky desktop computer. The monitor was a giant cube, its depth unimaginable by today's standards, the screen a bulbous rounded surface. It wasn't my first email. There had been an AOL account before it and a different gmail account with a childhood nickname that at the mature age of eleven I felt I had outgrown, so I made my new grownup email address, caitlynz98. I was still at an age where the importance of internet anonymity was drilled into my brain right alongside not talking to strangers. Now, caitlynz98 is mainly my spam email. Having existed for so long, it's linked to Google's ever growing list of services that I have knowingly and unknowingly used over the years.

Blogger	Calendar	Chrome	Cloud Print	Contacts
Drive	Fit	Google Account	Google My Business	Google Pay
Google Photos	Google Play Books	Google Play Movies _ TV	Google Play Music	Google Play Store
Google Shopping	Google Workspace Marketplace	Hangouts	Home App	Location History
Mail	Maps	My Activity	News	Profile
Tasks	YouTube and YouTube Music	archive_browser Chrome HTML Document 0.98 MB		

Google data. All folders.

Google divides up my data by service. Represented in the folder names are services Google has discontinued, like Hangout, and services I hadn't even realized were owned by Google, like Blogger. A good portion of the folders are empty or contain blank files. Eleven-year-old me didn't exactly have a disposable income, so folders like Shopping and Workspace Marketplace are void of data. I was using this account well before I even had a cellphone so my location data is sparse.

Search History

I know I shouldn't be surprised that Google has kept my search history. Under My

since making the account, Google has preserved every search I've conducted while logged in. In the three year period represented here, Google saved 1,471 searches.

Like with my Facebook data, I can fill in some of the context for my searches: October 2010, I was clearly on the American Revolution chapter of my seventh grade history book.

There are some searches that I imagine are from the span of time when everyone in my family was sharing one laptop. Whenever I'd forget to log out of my account, my other family members were browsing the internet as caitlynz98.

Google has no way to know that the search for the New York state bar directory, the search for realtor.com listings in Oakland county, Michigan, and the search an answer on the existence of pink dolphins were all conducted by separate human beings.

caitlynz98 is a singular identity that is interested in New York attorneys, Michigan real estate, and marine life all at once.
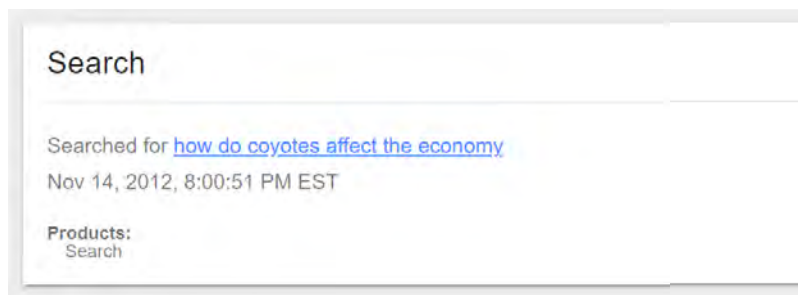
There's also plenty of search entries I can't begin to come up with an explanation for. Data doesn't preserve context.
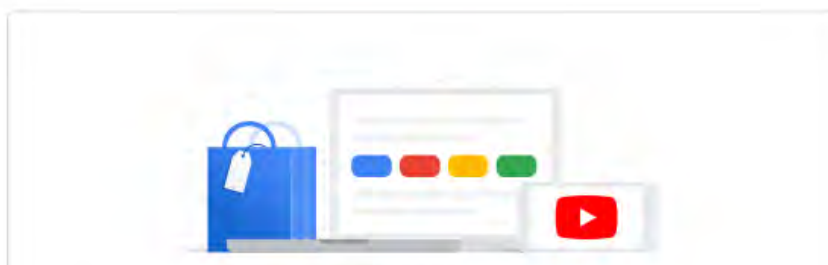
Ads

Google also provides a list of the ads I've seen and clicked on through Google searches, but the information about the ads I viewed is minimal.
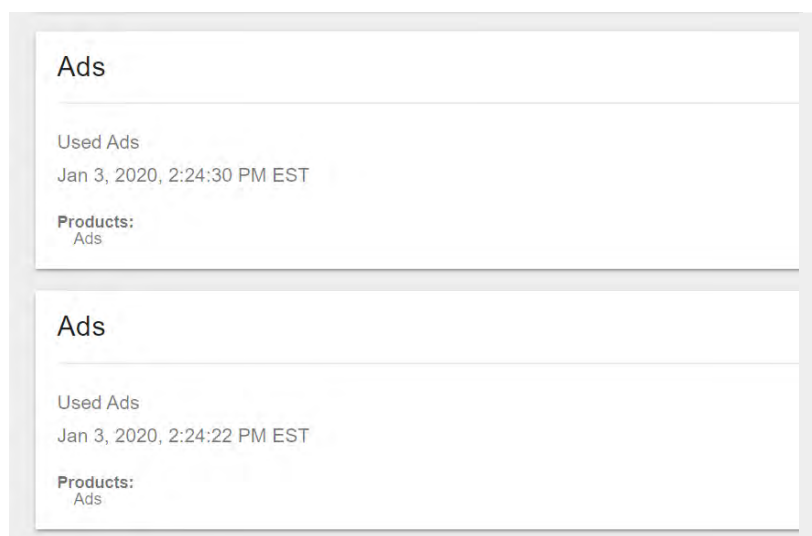
My Google data mostly consists of logs of dates and times events occured but very little context about the event itself. The most informative data in my Google download to me is my search history. I'm able to discern a lot about what I was looking for and even who conducted the search just by reading the query. In logs like ads I've viewed, there's no way to recover or even guess the context or content of what I viewed.

Ad Personalization

While there's no inferences in Google's data download, they're available for viewing online in Google's ad settings:

Google ad interests for caitlynz98 on November 21, 2020.

I learn more about the identity Google has constructed for me through their inferences than through the data itself. The same way I can look at my Google search log and remember the context for the searches, Google can apply their own context to my searches. This lists of my ad interests changes on a daily basis. For example, Auto & Vehicles normally wouldn't be on the list, but at the time of taking this screenshot, I had

approximate my gender, but on this day, my internet searches fit Google's definition of female.

The above list from November 21, 2020 on my caitlynz98 Google account had little in common with the lists of my ad interests less than a month later on December 14, 2020 from my school Google account:
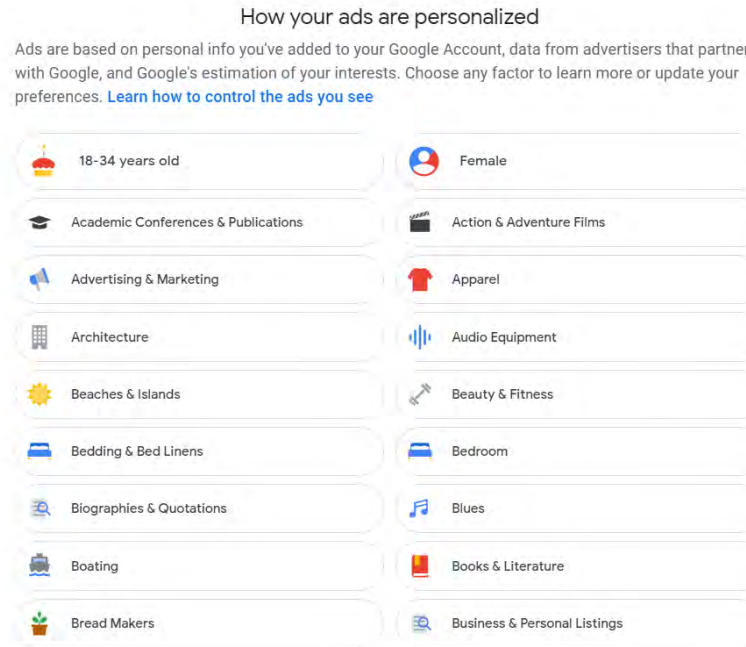
Google ad interests for my school account on December 14, 2020.

As my final semester of college is coming to an end, an interest in "Academic Conferences & Publications" has been added as I scramble to study for exams and write final papers. My age range has increased by ten years, from 18-24 to 18-34. While the end of the semester is stressful, I would say it hasn't aged me that much, but for Google's ad purposes, it has. Data identity is fluid and unstable. For Google, it's so they can place the right ad at the right time in order to optimize its effectiveness. For me, I primarily understand it as a categorization of the nature of my Google searches.

< Facebook

Twitter >

DO YOU KNOW WHO YOU ARE?
Caitlyn Zawideh

| Home | Table of Contents | About |

TWITTER

Recommended Tweets

I keep my Twitter account pretty small. It's set to private. I only follow 91 people, and my followers don't amount to a much higher number. Of those 91 accounts I follow, 17 of them, my data report informs me, I have muted, meaning that I still follow them, but their content doesn't show up on my feed. That means there's only 74 accounts whose posts are guaranteed to make it to my feed, and even then, many of those accounts haven't been active in years or only post rarely. In summary, I have less than 74 accounts whose content I regularly see, and yet I manage to spend more than an hour a day on Twitter, the majority of it spent on my feed and not on the explore tab. On their own, the people I follow are not tweeting often enough to fill the feed with over an hour's worth of content, so how does Twitter keep me scrolling? There is about as much, if not more, recommended content on my feed than there is content from the accounts I follow.

Of the first 8 tweets on my feed on the evening of November 19, 2020, only one is a tweet from an account I follow. One is a retweet from an account I follow. The rest are recommended content.
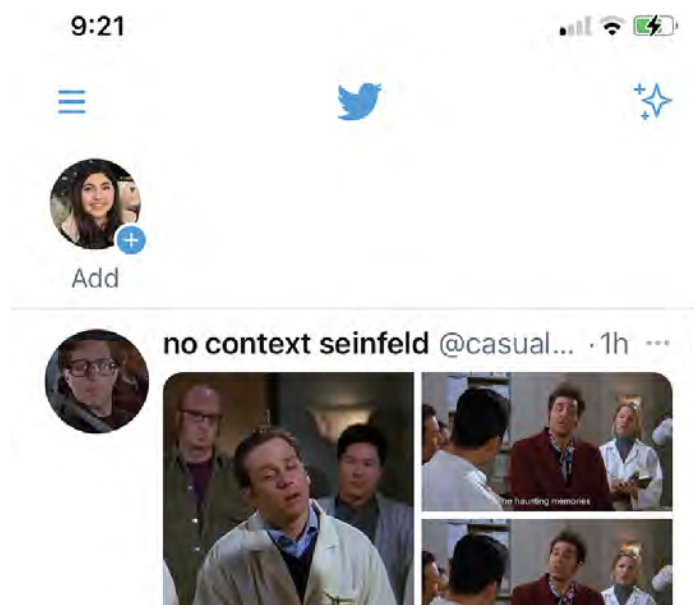
Tweet from an account I follow. Not a retweet or a quote tweet.

Recommended based on who I follow.

Retweet from an account I follow.

follow liked it.

An ad tweeted by Bloomberg promoted by American Express.

From this small sample of my feed, it is clear how much recommendations can affect a user's experience using Twitter. The algorithmic decision-making behind what content shows up on my feed is a more complex calculation than who I follow or the tweets people I follow like. That is evident in my data report.

Twitter has all the staples of personal data reports in a user-friendly html file similar to Facebook's.
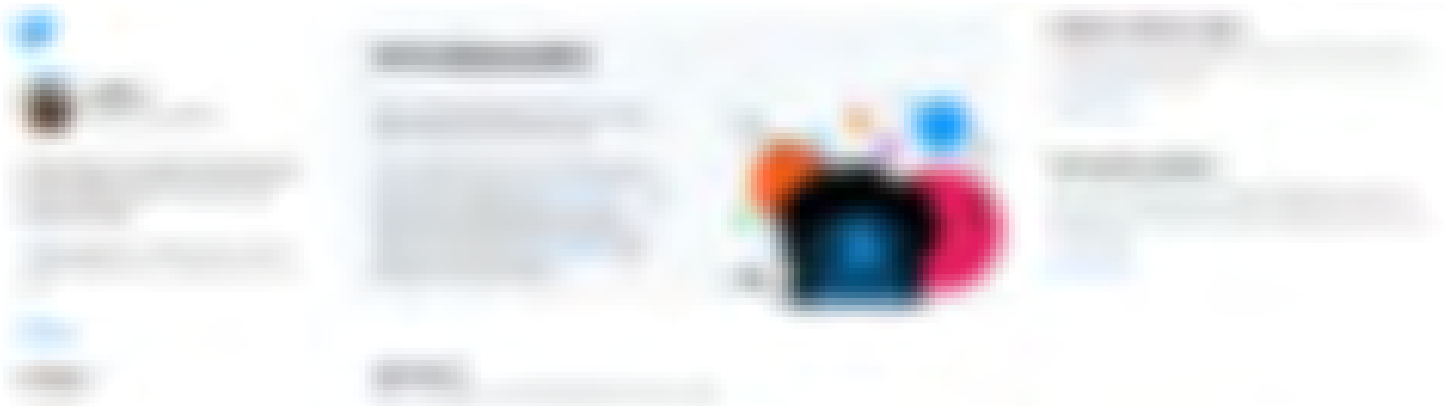
Targeted Ads and Data Dopplegangers

Besides a log of my tweets, likes, and direct messages. Twitter also includes a list of every ad that's been placed on my feed and how long I interacted with it. Twitter provides more detail than the other platforms about why they showed me certain ads.

The reasons behind an ad for an article about the 2020 presidential election from the Wall Street Journal are straightforward. The election was coming up at the time of the ad and my data says I'm an American of voting age.

Sometime in June 2020, I visited the Wall Street Journal's website, putting me on a list that WSJ then used to include me in the audience for this ad.

Microsoft Store @MicrosoftStore · September 26, 2020 · 1:14 PM
Post your top 5 emojis. We'll go first: 🙌💚🔥🎮💯
View ad on Twitter

Hide details
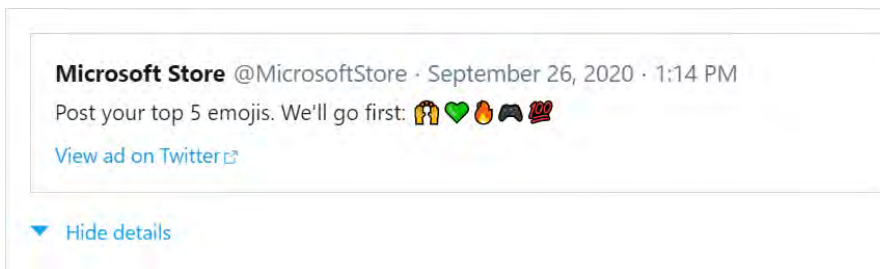
Despite a similar display, it's not nearly as clear how this ad from the

| Retargeting campaign engager | Retargeting campaign engager: 24294943 |
| Retargeting engagement type | Retargeting engagement type: 2 |

feed. The answer is coded in indecipherable terminology followed by an equally indecipherable string of numbers. Here, Twitter can only be as transparent as the ad owner allows them to be.

Kraft Macaroni & Che... @kraftmacnchees· · September 25, 2020 · 3:45 ...

Do you want to win a box of limited edition Pumpkin Spice Kraft Mac & Cheese? Of course you do! Tag us in a tweet with #PumpkinSpiceKMC +
#Sweepstakes for a chance to win. No purchase necessary. 18+. Ends: 9/29/20.
Rules:https://t.co/9jcGAJfec2 https://t.co/VggsDrsjAl

View ad on Twitter

▼ Hide details

Age	18 to 24
Conversation topics	Drake
	Spotify
	HBO
	Zendaya
	Harry Styles
	Netflix
Follower look-alikes	@chancetherapper
	@ShawnMendes
	@ArianaGrande
	@VictoriasSecret
	@FactsOfSchool
	@rihanna
	@Marvel
	@chrissyteigen
	@ESPNCFB
	@Zendaya
	@ladygaga
	@iamcardib
	@Harry_Styles
	@KylieJenner
	@kyliecosmetics
	@barstoolsports
	@justinbieber
	@Beyonce
	@sza
	@funnyordie
	@lizzo

Follower look-alikes are another way Twitter curates recommended content for their users. The advertiser identifies accounts whose audience they want to target, and then Twitter provides accounts that look similar the followers of the designated account.

Here, Kraft selected a long list of popular accounts to find look-alikes. Then Twitter determined that my activity and data are similar to the followers of these accounts, which made me part of the target audience for this ad.

While Kraft has chosen some of the most followed accounts on Twitter to identify follower look-alikes (i.e. Beyonce, Chrissy Teigen, Justin Bieber), Twitter actually advises against the practice of using extremely popular accounts to find an ad audience. Instead, Twitter Business suggests, it's more effective to "target the @handles that are most closely tied to your business. While very famous individuals or companies may have millions of followers, those millions are not necessarily aligned around a topic that's

@taylorswift13

@AOC

Locations	United States
Movies and TV shows	Euphoria
	College Football

Unlike Kraft, the pharmaceutical company Novartis seems to have followed Twitter's advice. For this ad about the "synergy between art and science," they only used one account to find look-alikes. I learn from his twitter bio that Atul Gawande is a surgeon, writer, and researcher. He is not followed by anyone I follow on Twitter, and I do not follow him myself. Yet, for reasons not apparent in my data download, I resemble his followers enough for the Norvartis ad to appear on my feed.

Nodes in a Web

My Twitter data, particularly the concept of follower look-alikes, is the best example so far of our how data identities do not exist independently. A chain of connections is what leads suggested content to my feed.

Tweets tagged with "[user] liked" or "[user] and 8 others follow" are simple instances of this. There is an established connection between my account and the accounts that I follow. When one of the accounts I follow likes a tweet, it establishes another connection between them and the tweet they liked. To determine what suggested tweets should show up on my feed, Twitter's recommendation algorithm looks at the established connections between users and content (i.e. a like) and then at the established connections those users have with other accounts, how often those accounts like similar content, and from that information and other factors like it, they identify the content I have the highest probability of engaging with and place it on my feed.

Identifying look-alikes is a similar process on a larger scale. The starting point is the account identified by the advertiser. The established connection is between that account and its followers. Look-alikes are found through the existing connections between the followers and the other content they like. By comparing those connections to the connections of other users who don't follow the starting point account, Twitter is able to find look-alikes. Our data identities are all nodes in an always-expanding interconnected web. The meanings and inferences made about our data is dependent on how each of our data identities relate to each other.

< Google

Snapchat >

DO YOU KNOW WHO YOU ARE?
Caitlyn Zawideh

Home   Table of Contents   About

SNAPCHAT

Feeling in Control

My Snapchat contains all the usuals that have been consistent across these platforms:
a list of my contacts, a log of everything posted, sent, received, and saved, every ad I've
ever viewed and when I viewed it, approximations of my location, my age, my gender, my
interests.

	User Profile	
	App Profile	
	Country:	US
	Creation Time:	2013-05-31 00:02:04 UTC
Login History and Account Information	Account Creation Country:	UNKNOWN
Snap History	Platform Version:	IOS
Chat History	In-app Language:	en_US
Our Story and Crowd-Sourced Content	**Demographics**	
Purchase History		
Shop History	Cohort Age:	AGE_21_TO_24
Snapchat Support History	Derived Ad Demographic:	FEMALE
User Profile		

Snapchat data download. User Profile.

The download includes an FAQ page, where Snapchat notes what information has been
left out this report, such as the log of my activity and engagements with ads. Where
Twitter provides what ads I saw, when, why I saw it, and how long I looked at it,
Snapchat only provides a list of advertisers I've interacted with.

Snapchat assures me that they typically retain the logs for 180 days, but "in some cases,
we may need to retain these logs for much longer. For example, we may need to store
some longs for the lifetime of the account."

Despite this, like Facebook, Snapchat presents the data in a way that's intended to give the user a sense that they are in control, although they do this far less effectively. The language is filled with a ambiguities and disclaimers. In one section of the FAQ aptly subheaded, "How can I control my information?" Snapchat writes that they want users to "feel in control when it comes to how [they] collect process your data."

Snapchat Has a Search History

Snapchat provides only a month's worth of searches from August 30, 2020 through September 26, 2020. These search terms are unique compared to other platforms because it's a log not of questions I searched the internet to answer or searches for people to add to my social network. The Snapchat queries are my searches for emoji's, stickers, and gifs to adorn snaps I sent to my friends or posted to my story.

search term ———
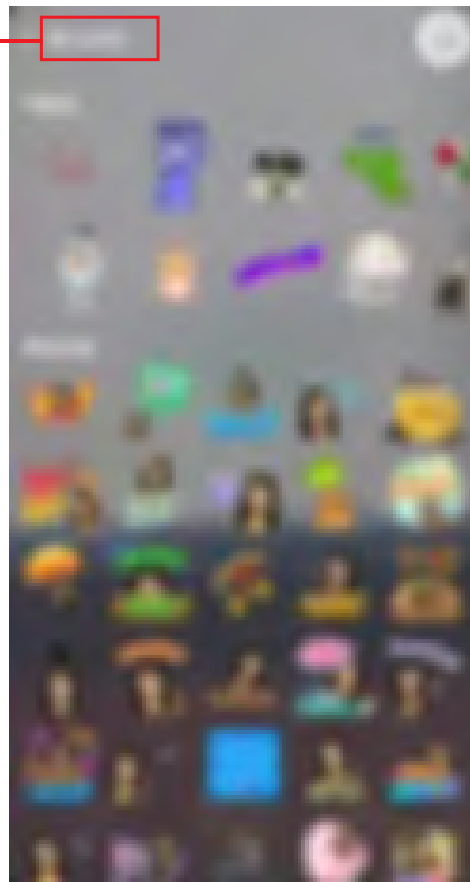
Snapchat sticker search results for the term "hello."

In the timeframe the search data is from, my snapchat memories (saved images from the same date over the years) were all from pre-pandemic days - friends, concerts, bars. Since the pandemic began, my friends and I have all been sending each other snap memories of happier times, and that nostalgia, longing to be go back or be somewhere else, and the anxieties of the present are all evident the some of the search entries.
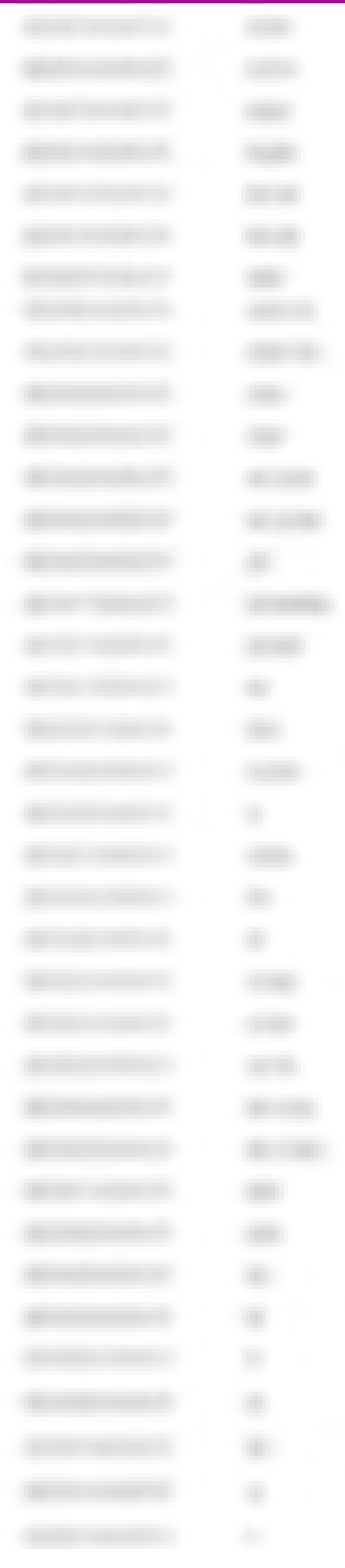
logged as individual terms - the result of the search loading before I had finished typing.

Many of these searches took place hours and days apart from each other. They are small pieces. Whatever context the individual searches existed in at the time they were logged is gone, but consolidated altogether as data points the fragments form a continuous narrative, a new context.

Ranking

In terms of information inferred about me, Snapchat has a list of topics similar to what Facebook provided called Ranking. However, the list only shares one commonality with Facebook, Cooking. Snapchat is another platform where recipe videos are widely watched via Snapchat's recommended stories feature. The other interest categories, however, range from oddly specific, like Naile_Art, to vague, as with Magazines.
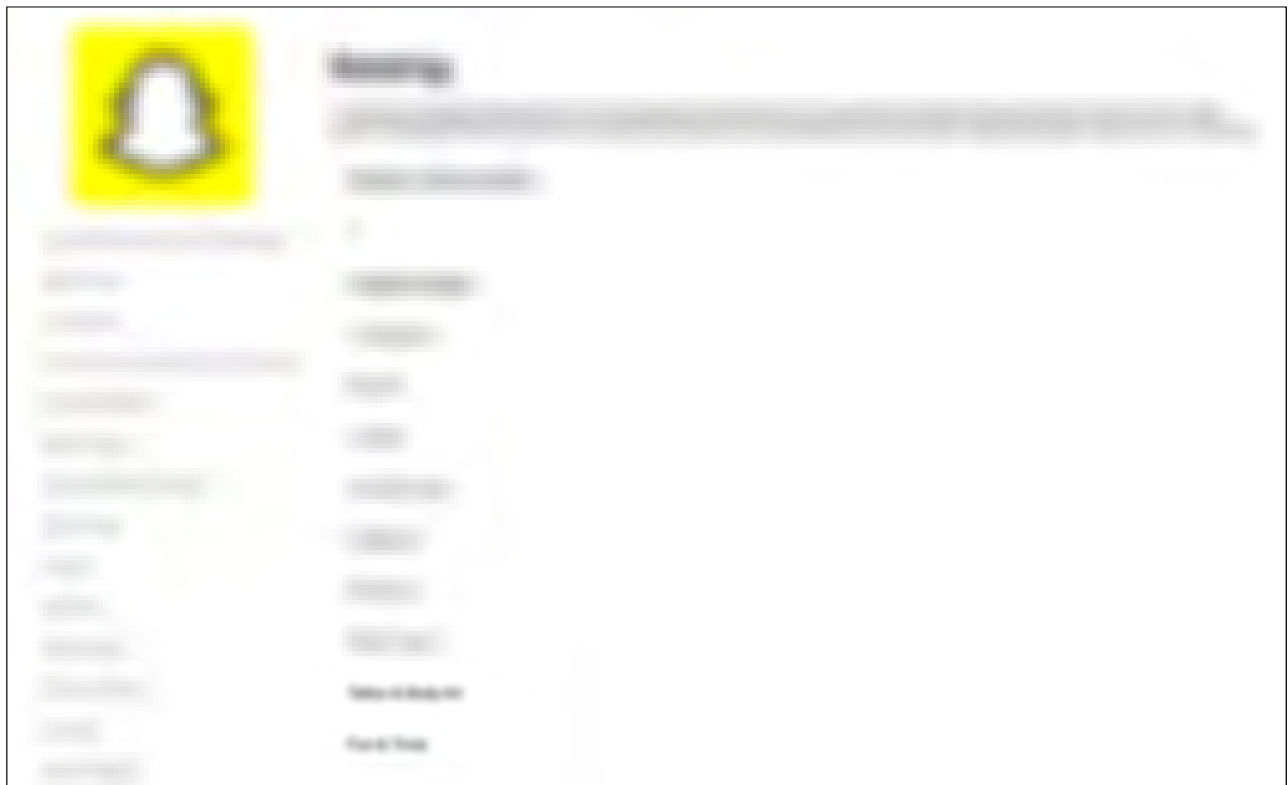
Snapchat data download. Interest rankings.

Work and Home

Snapchat is an app that always has my location. From that data, Snapchat has approximated the longitude and latitude of where I live and where I work within an error range of roughly 17 meters.
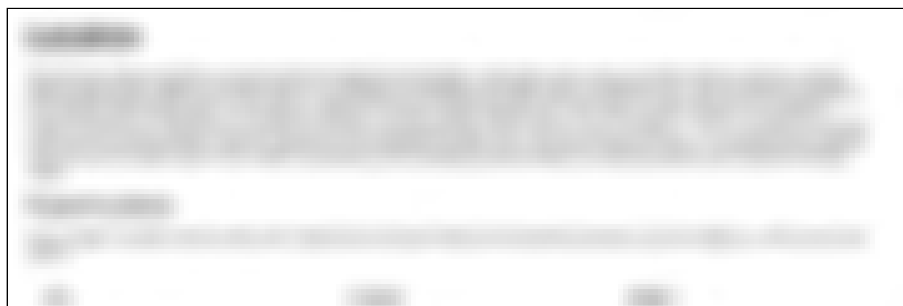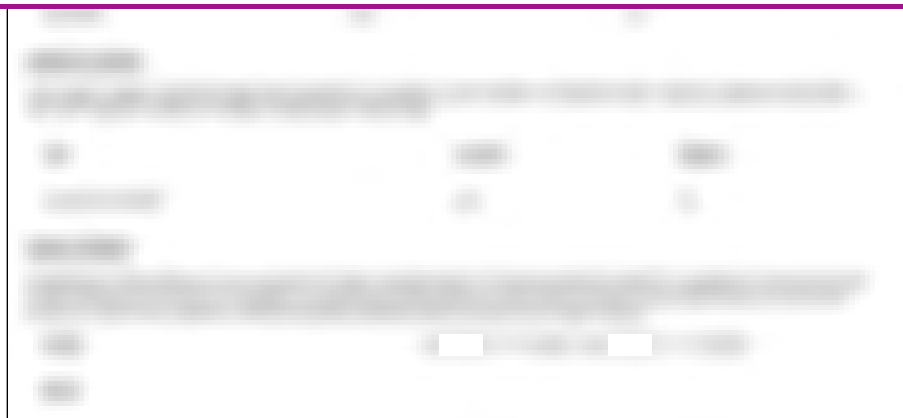
Snapchat location data. Work is unsurprisingly blank since I've been working from home since March. Snapchat hasn't updated it's collection and inference methods for the age of home offices.

As Snapchat explains on the Locations page of my data download, they make these guesses about my location, so my Bitmoji, an avatar you can create for yourself in the app, can reflect my activities. For example, when I'm at the library, my Bitmoji will appear on the Snapchat map sitting at a desk with a stack of books beside her.

What's interesting about how Snapchat has presented why and how they use location data is that they make the trade-off very clear. Denying Snapchat access to my location data means giving up some of the app's services. In the digital, data-driven space, we are always choosing between convenience and privacy. On one hand, I like that if I want to post a snapchat story, I need only swipe through a few options to find a nicely designed sticker announcing my location. On the other hand, the idea that Snapchat is always monitoring my location and making approximations of where I live and work is unsettling.

Even more unsettling to me is their disclaimer about the location services: "Even without location permissions, Snapchat can sometimes guess your approximate location based on your IP address." Really, we don't get a choice as users of Snapchat. Whether we actively participate or not, Snapchat tracks location data either way. We just get to decide if we want to see it ourselves in the form of cute graphics while they meet their "important business needs." We always stand to lose more in the tradeoff than Snapchat does. They are always in a position to gain.

< Twitter

Spotify >

DO YOU KNOW WHO YOU ARE?
Caitlyn Zawideh

| Home | Table of Contents | About |

SPOTIFY

Spotify is distinct from the other platforms included because it's not necessarily a social networking platform, and it doesn't employ the use of any kind of newsfeed or postings. If you're a Spotify Premium subscriber, your streaming experience is never interrupted by ads. Spotify recommends music in the app and site interface for all users and provides algorithmically generated playlists for its users based on their data. My data download revealed inferences categorizing me for targeted ads far beyond the scope of music.

Read_Me_First   Follow   Inferences   Payments   Playlist1   SearchQueries   StreamingHistory0   Userdata   YourLibrary

The content of every file with the exception of Inferences is familiar. It's a lists of the artists I follow, every song I've listened to, my search queries and saved songs. All of these are visible in Spotify's user interface, although they appear differently in the data report. The in-app view is very polished and designed for ease of use and understanding. The data view is a json, a file type used in web and app development that provides information to be rendered on the webpage or in the app. In other words, it's made for ease of use and understanding of computers, not people.

In app playlist view
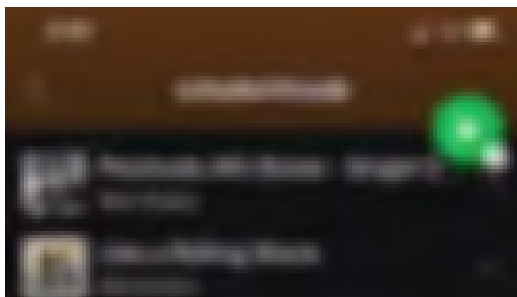
Playlist as data
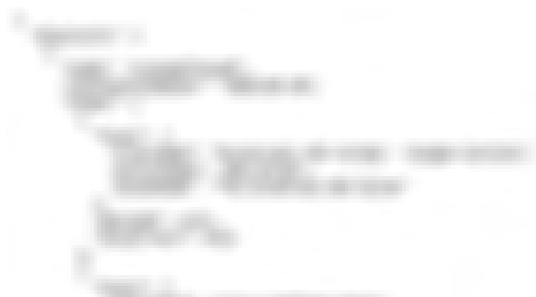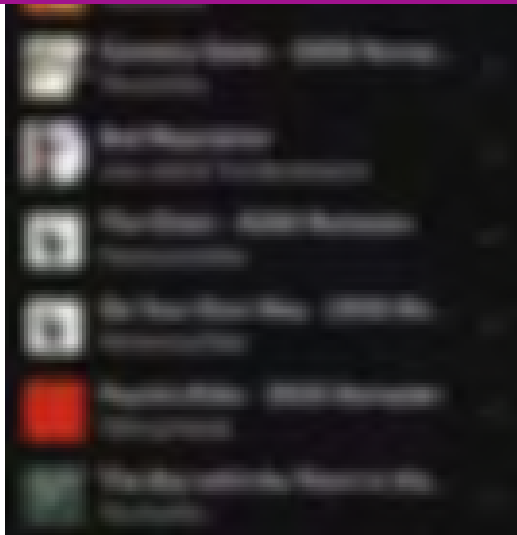
Inferences

The only new-to-me information in my Spotify data download exists in the Inferences.json folder.

The list is extensive. It's filled with hundreds and hundreds of consumer categories that I've been placed in. Below is just a small segment of them.
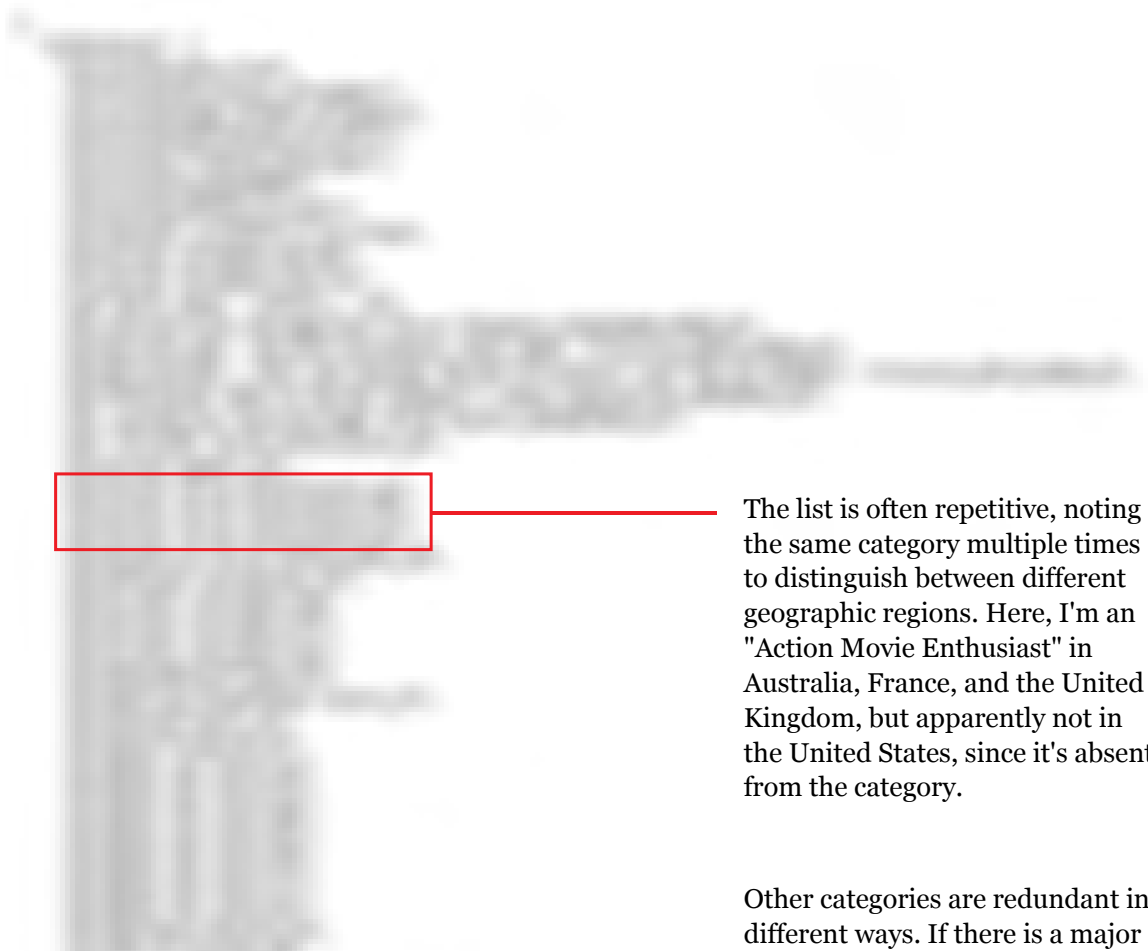
The list is often repetitive, noting the same category multiple times to distinguish between different geographic regions. Here, I'm an "Action Movie Enthusiast" in Australia, France, and the United Kingdom, but apparently not in the United States, since it's absent from the category.

Other categories are redundant in different ways. If there is a major distinction between "Art &
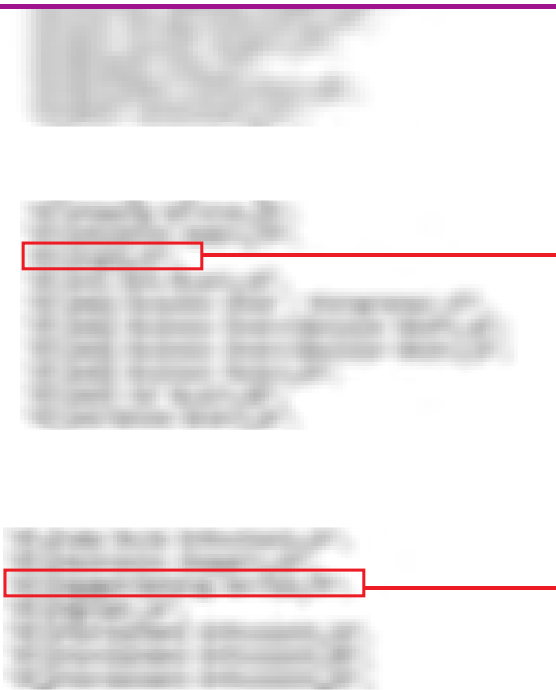
Enthusiasts" it's not clear what that difference is here.

There are also categories that completely contradict each other. A person cannot be both engaged and single at once in the real world. In the world of market categorization, nothing has to be mutually exclusive. In data form I am labeled "Single" and "Engaged" and as far as the data is concerned, both of those things are true because my data has allowed Spotify to infer that it is.

Reading through the list of inferences, it's not clear how they were derived from my listening habits alone. How does a Bob Dylan playlist inform Spotify that I fall into the category "3p_Engineer_UK?" Spotify gives a brief explanation of how they draw inferences on the Understanding my Data page of their support site:

> "We draw certain inferences about your interests and preferences based on your usage of the Spotify service and using data obtained from our advertisers and other advertising partners. This includes a list of market segments with which you are currently associated. Depending on your settings, this data may be used to serve interest-based advertising to you within the Spotify service."[5]

While it's not unlikely that Google, for example, has information similar to this with my inferences grouped by advertiser rather than interest, they don't present it that way. Notably absent from Spotify's inferences are the one thing I had expected to find there: music genres - something I could easily map back to my activity on the app. With my Google inferences, I could understand how those inferences came to be based on the information I knew I had been giving to Google. With Spotify, those types of connections are not evident.

(< Snapchat)   (TikTok >)

DO YOU KNOW WHO YOU ARE?
Caitlyn Zawideh

Home   Table of Contents   About

TIKTOK

Getting Acquainted

My sister convinced me to download TikTok some time in December 2019. Its main feature is the "for-you" page – an endless queue of algorithmically-curated videos, each less than a minute long. The for-you page is entirely made up of suggested content based on your user data. When I first made an account, the only information I provided was my location and my phone number. My user data didn't exist yet. I didn't fill out a questionnaire profiling my interests. TikTok didn't have much to go off of in terms of approximating what content I would want to see, so when I first encountered my own personal for-you page, I was confronted with cat video after cat video. On instinct, every time one came up on my feed, I made sure to flag with the "not interested" button and kept scrolling. After just a few days on the app, tapping the heart icon for videos I liked, quickly scrolling past the ones I didn't care for, and marking the content I really didn't want to see with "not interested," the content of my for-you page slowly started to accurately reflect my interests.

Tiktok's accuracy has crossed the border into creepy territory plenty of times. One night, over dinner, my mom and I had a brief conversation about Marriage Story, specifically the scene where Laura Dern goes on a long monologue railing against the unfair expectations placed on mothers and its basis in what she calls "our Judeo-Christian whatever." The whole exchange couldn't have been longer than a minute or two. The next morning, I opened Tiktok. Just a few videos into my for-you page, there was a clip of the exact scene from Marriage Story we'd talked about the night before, down to the one line I'd said out loud in the exchange with my mom, the Judeo-Christian whatever.

I can try to rationalize this eerie encounter with some more context about my Tiktok and online activity: for one, the movie I referenced isn't exactly niche. I follow a couple film accounts, in the past I've liked videos featuring actors that are in Marriage Story, I've liked TikToks about Little Women - which Laura Dern is also in, and the film was directed by Greta Gerwig who is married to the director of Marriage Story. If I don't want to jump to the conclusion that TikTok is listening in on my private conversations, there's plenty of dots for them to connect that would eventually lead to that one specific scene in Marriage Story just by following the chain of connections in my data.

When I downloaded my TikTok data, I expected to find at least some evidence of the inferences they have made to populate my for-you page, but it is the most opaque and least forthcoming with explanations of all the platforms I've downloaded data from.

The Download

To download my TikTok data, I had to go through the app's Privacy settings, where I found a Personalization and Data section. From there, I found the option to request a data file. TikTok warns that it can take up to thirty days to process data requests, but my file was ready less than forty-eight hours later:

.
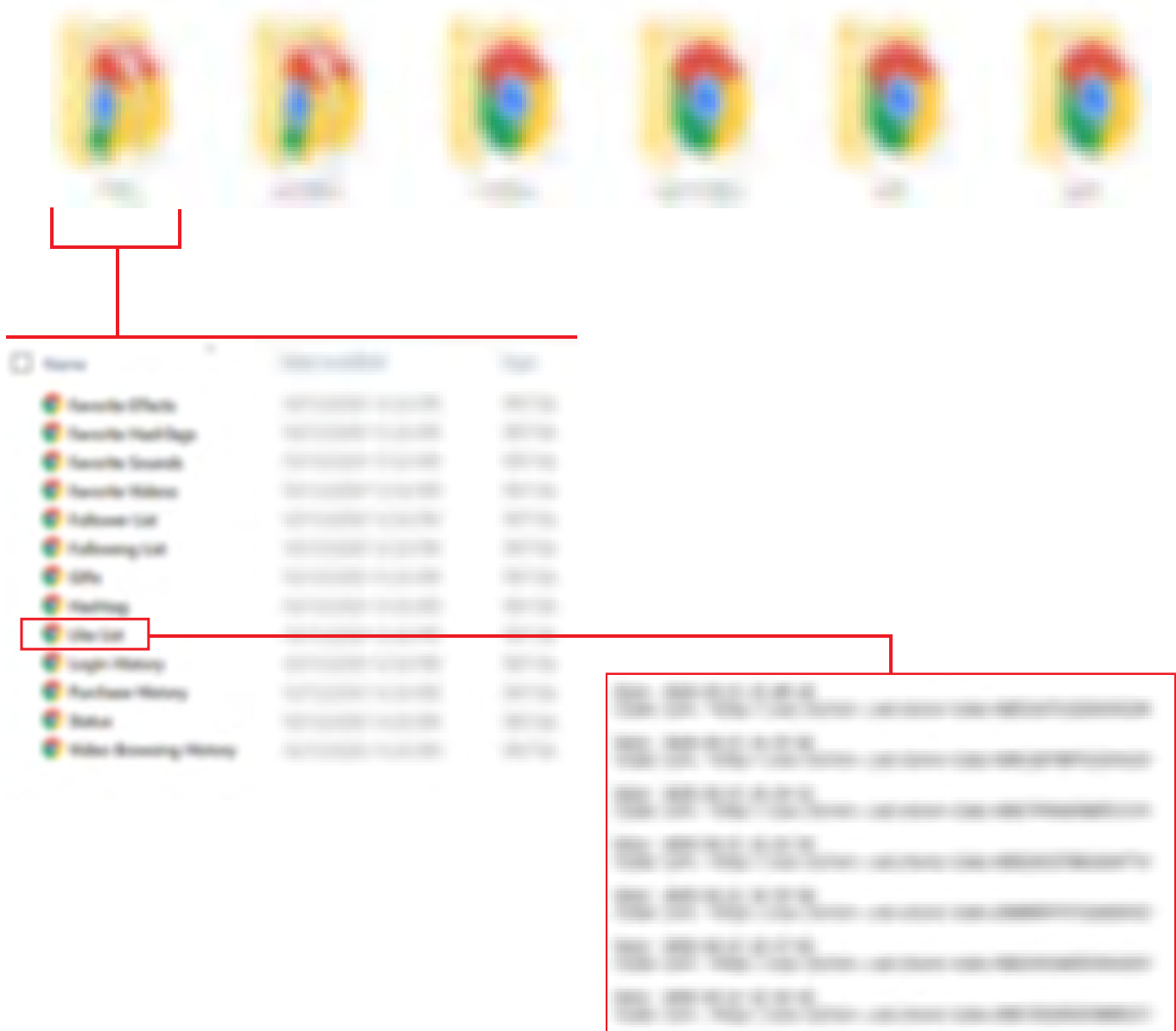
A log every TikTok that I've marked as "liked."

The information that is there is not exciting or different from what the other platforms have provided. There's a log of every time I've logged into the app. A record of my privacy settings. Every comment I've ever left on a video. "Like list" is a log of every video I've ever liked. Even the Hashtag document under Activity, which I thought might contain a

hashtags I've used in my own posts. However, unlike the other platforms, TikTok doesn't provide any inferences it's made about me in order to generate content my for-you page.

In a recent blog post, TikTok gave a rundown of how their for-you 'recommendation system' works:

> Recommendations are based on a number of factors, including things like:
> - User interactions such as the videos you like or share, accounts you follow, comments you post, and content you create.
> - Video information, which might include details like captions, sounds, and hashtags.
> - Device and account settings like your language preference, country setting, and device type. These factors are included to make sure the system is optimized for performance, but they receive lower weight in the recommendation system relative to other data points we measure since users don't actively express these as preferences.
>
> All these factors are processed by our recommendation system and weighted based on their value to a user. A strong indicator of interest, such as whether a user finishes watching a longer video from beginning to end, would receive greater weight than a weak indicator, such as whether the video's viewer and creator are both in the same country. Videos are then ranked to determine the likelihood of a user's interest in a piece of content, and delivered to each unique For You feed. [6]

So it's from all this simple, raw data that they have been able to approximate my interests. Since none of those approximations appear in my data file, I can only guess what inferences they've made based on the ever evolving content of my own For You page.

For Me

After almost a year of regular visits to the app, the TikTok algorithm has gotten eerily specific in its understanding of what content I'm most likely to watch. My for you page has become increasingly personalized and niche. There is no scenario outside the context of TikTok where I would watch twenty-second videos matching Le Creuset cookware to different Hozier songs and accompanying kitchen remodels. It's not something I would ever think to seek out myself, but one day, one of those videos showed up on my for-you page. I was mildly entertained, so I hit the like button, and when more videos of the same nature subsequently showed up, I liked those too. The algorithm has gotten to know me and my digital habits so well that it knows what videos I'm going to like before I do.

From my for-you page. TikTok user @ave.abe matches Le Creuset's meringue set with Spanish hacienda kitchens and sets the mood with a song by Irish folk artist Hozier. Why is this so calming?

While TikTok doesn't provide interest categories themselves, their users do. To be posed the question, "What side of TikTok are you on?" is to ask what kind of person TikTok has approximated you to be.

Typically, the users asking this question aren't thinking about data at all. They're asking what subcommunity of the app you've been grouped into based on the type of content that's been algorithmically placed on your for-you page. The side of TikTok that you're on — and you can be on more than one "side" at once — says something about you, your social standing, your taste, and how interesting you are as a person. In this realm, there's no greater insult than accusing someone of being on "Straight TikTok," which means your for-you page is full of conventionally attractive influencers, dance challenges, and lifestyle vlogs, occasionally sponsored by Dunkin Donuts. In other words, it is content that is dull, unoriginal, and uninteresting, which in turn means that you are also dull, unoriginal, and uninteresting, at according to the people who are not on Straight Tiktok.

Charli D'Amelio, currently the most followed account on TikTok, promotes her new Dunkin Donuts drink, 'the Charli.'

There was about a three week stretch in September where I found myself grouped into a more niche TikTok community, WitchTok, which is exactly what it sounds like: my for-you page was overwhelmed with videos of people in black pointy hats, people in flowy white dresses explaining how to cleanse bad energy from your home by burning sage while Stevie Nicks played in the background, and one guy who'd made more than a dozen videos dressed as a 17th century Puritan woman reinventing scenes from *the Crucible*. My only explanation for why this might have happened is that sometime at the beginning of the semester, I was looking for an elective on my college's course guide and came across a class on the History of Witchcraft. Suddenly, TikTok was crying witch. I might as well have been Goody Proctor dancing with the devil. While TikTok has denied having access to users' Google data, it's the only online activity I can think of that would link me to an interest in WitchTok.

WitchTok. TikTok user @bequietjoe puts a modern spin
on an old classic.

Of course, the repercussions of miscategorization in this instance were minimal. It's not
like I was burned at the stake because of my status as a data witch. I was just stuck in a
particular corner of the app for a few weeks. In other cases of mistaken data identity, the
consequences cross over from the digital to the real world in damaging and alarming
ways.

In 2019, the NYU Law Review published a study by race, technology, and law scholar
Rashida Richardson that called for skepticism toward data-driven predictive policing
practices. Predictive policing is a system that uses past data in order to predict how likely
it is that a person will be a perpetrator or victim of a crime, or how likely it is that a crime
will occur in a specific location. The study found that the use of data by law enforcement
to practice predictive policing in jurisdictions with a history of flawed, racially biased,
and unlawful practices and policies has resulted in inaccurate predictions and reified
existing biases in the prediction model 7.

In practice, that's not the case. To datafy something, whether it's the probability that an individual will like a video or the likelihood that they will commit a crime, will always leave gaps in understanding and lead inaccuracies. While data can be informative, it is never whole objective truth, even when it's properly and accurately collected. Bad data, when fed into prediction models, often leads to the magnification of it's flaws and biases.

< Spotify

Conclusion >

DO YOU KNOW WHO YOU ARE?
Caitlyn Zawideh

| Home | Table of Contents | About |

ACCEPTING
PARTIALITY

"If you were to look at a complete model of your digital self, it would be a complex relational web. At the most granular level of that web are nodes, each representing actions (a text, a selfie, a purchase...). The connections between those nodes are formulas that infer relationships, record patterns, and predict behavior. If you zoom out, you get the sub-web of a given service. [...] These sub-webs then join together to form the larger web that is your digital identity." – Nik Milanovic, "The Next Revolution Will Be Reclaiming Your Digital Identity" [8]

Opinion writer for Tech Crunch Nik Milanovic imagines a future in which individuals will have total control of their data identities. Total control, as he sees it, would mean that everyone would own their personal data, have it all consolidated in one place instead of fragmented across platforms, and have the ability to grant or deny access to third-party services. In terms of how such a model of digital identity could exist, Milanovic turns to blockchain technology, which is a system of information that is highly secure, decentralized, and distributed in structure. [8]

Considering the way personal data is used by companies now, for advertising and personalization, it's hard to imagine what the function of a singular data identity would be in this context, besides the ability to have control over who accesses it. While Milanovic's vision for the future of data identity sounds idyllic, the idea that an identity, even a datafied one can be complete and cohesive across the many contexts our data is generated in, is unattainable.

Despite the large amount of personal data that I've been able to download and manually consolidate, it is far from a complete model of my identity. There were plenty of gaps, missing information, and incorrect approximations. Many of the data points from different platforms and even within the same platform at times directly contradicted each other. To have a singular complete model of my digital self would be to imply that identity is singular and completely understood and can be accurately and objectively

The idea that identity is singular and completely understood is false both on and offline. Offline, I am understood as a different person in a classroom than I am in the TSA line at the airport or with friends. Even then, I could be understood as a certain type of friend in one social circle, and a different type in another. Identity is as much about the way we are understood by others as it is about the way we understand ourselves. It's highly dependent on context and relationships. The same way an algorithm takes in data about my actions and uses its pre-existing framework to interpret what those actions mean, people are always trying to understand each other using prior knowledge, and there are always gaps in understanding.

For example, if I came home from work with red puffy eyes, my roommate might ask me if I was upset about something, but the reality could be that I have bad seasonal allergies and the pollen count was high that day. We misinterpret each other all the time. Being a social human being means constantly trying to bridge those gaps in understanding.

Gaps in understanding just take a more tangible form in data. To Google, around 3:30 in the afternoon on October 25, 2020, I was understood as a Single 18- to 24-year-old Female interested in, among about 140 other things, Apparel, Breakfast Foods, Clipart, Home Automation, and Parenting. At the same time, to Twitter, I was an English-speaking 13- to 54-year-old Female interested in, among hundreds of other things, Abba, California, Comedians, Jordan (the country), and Writing. Some of the inferences are accurate, some are too broad to be completely false, and all of them are an attempt at understanding who I am based on Google's and Twitter's own definitions of what an interest in Abba or Jordan or Clipart actually means.

The mistake in accepting Milanovic's model of an ideal complete digital identity is to accept data as somehow more objective and therefore more true than human understanding. On the idea of objective knowledge, feminist theorist Donna Haraway writes, "the codes of the world are not still, waiting to be read. The world is not raw material for humanization." It is more valuable to understand knowledge as something rooted in your perspective, through the lens of your senses. [9] The multitude of different ways my data has been interpreted and the ways I've interpreted it myself have made it abundantly clear that data is not objective knowledge. In coming to terms with the question of whether I can construct a complete identity for myself out of my data, it's best to accept Haraway's concept of objectivity through partiality. You cannot have a complete, objective data identity any more than you can have a complete objective embodied, non-data identity.

The problem that distinguishes these platform's datafied interpretations from our regular person-to-person ones is their impact on how we see the world. In 2012, Facebook released the results of a study that found that users that received informational messages about voting on their feeds had "a direct impact on political self-expression, information seeking, and real-world voting behavior." The study illustrates how the content we see on our algorithmically curated news feeds, whether that's Facebook or

The risk is that the lack of awareness puts users in a passive role. We accept the interpretations of the world that appear in our feeds as in line with our own because, based on our data, the algorithms decided that they they are. This is where the large-scale problems of misinformation and polarization come into play.

Two studies from the Conference on Human Factors in Computing Systems found that when users were made aware of how algorithms populate their feeds or form their own theories about how it works, they alter their interactions with the platform. When there is an increase in awareness, users can exercise some control over the content they see. [10] [11] Knowing what you look like as data, at least in part, and trying to recognize how that data translates to what you see when you browse recommendation-algorithm driven platforms, the experience of scrolling through your feed doesn't have to be completely passive. You can be intentional about the data you create through your engagement with your feed.

Having downloaded my data, I start to understand the content I see on my feeds a little better. When I see an ad on Twitter for Novartis pharmaceuticals, I now know it's because Novartis wanted to reach an audience that looks like followers of @Atul_Gawande, which for data-related reasons that weren't provided to me, meant that I was the audience they were trying to reach. Now when I like a tweet or scroll past a certain ad, I can guess the effect it will have on my feed and make decisions about my digital habits accordingly.

It's not possible to escape the reach of data collection or the influence of algorithmic recommendations unless you want to delete all your social media, throw out your cell phone, and move to a remote cabin deep in the woods to live off the land. Solutions in the form of legislation or technology like blockchain are on the horizon, but not necessarily promising. It's hard for policy to keep up with quickly evolving technology, and I'm skeptical of a technology-reliant solution to a problem created by our reliance on technology. In the meantime, this is what I can do: take the time to be reflective and critical of the content appears on my feed and why it might be there, be intentional about the data create with my digital habits, and in that small way, I get to regain some control over my data identity.

< TikTok   Table of Contents   Notes   About

Want to download your data? Learn how: Facebook, Google, Twitter, Snapchat, Spotify, TikTok

DO YOU KNOW WHO YOU ARE?
Caitlyn Zawideh

| Home | Table of Contents | About |

Notes

1. *The Social Dilemma*. Directed by Jeff Orlowski,Exposure Labs, 2020. Netflix.
2. "Art. 20 GDPR – Right to Data Portability." *General Data Protection Regulation (GDPR)*, 28 Mar. 2018, gdpr-info.eu/art-20-gdpr/.
3. Egan, Erin, and Ashlie Beringer. "It's Time to Make Our Privacy Tools Easier to Find." *About Facebook*, 28 Mar. 2018, about.fb.com/news/2018/03/privacy-shortcuts/.
4. "Interest and Follower Look-Alikes Targeting." *Twitter*, Twitter, business.twitter.com/en/help/campaign-setup/campaign-targeting/interest-and-follower-targeting.html.
5. "Understanding My Data." Spotify, 26 Nov. 2020, support.spotify.com/us/article/understanding-my-data/.
6. TikTok. "How TikTok Recommends Videos #ForYou." *Newsroom*, TikTok, 16 Aug. 2019, newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you.
7. Richardson, Rashida and Schultz, Jason and Crawford, Kate, Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice (February 13, 2019). 94 N.Y.U. L. REV. ONLINE 192 (2019), Available at SSRN: https://ssrn.com/abstract=3333423
8. Malanovic, Nik. "The Next Revolution Will Be Reclaiming You Digital Identity." Tech Crunch, 17 October 2017, https://techcrunch.com/2017/10/17/the-next-revolution-will-be-reclaiming-your-digital-identity/
9. Haraway, Donna. "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective." Feminist Studies, vol. 14, no. 3, 1988, p. 593., doi:10.2307/3178066.
10. Eslami, Motahhare, et al. "First I 'like' It, Then I Hide It." Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016, doi:10.1145/2858036.2858494.
11. Eslami, Motahhare, et al. "'I Always Assumed That I Wasn't Really That Close to [Her]"." Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15, Apr. 2015, doi:10.1145/2702123.2702556.