Data Intimacy, Machine Learning, and Consumer Privacy

Prof. Michael Kearns[1]
University of Pennsylvania

Summary. *We discuss and analyze the data sources and practices at large consumer-facing technology companies such as Google and Facebook, and examine the central role of machine learning and artificial intelligence at such companies. We focus in particular on the notion of data intimacy --- the fact that machine learning enables companies to routinely draw accurate predictions and inferences about users that go far deeper than what is merely on the "surface" of the data collected. We discuss the consequences for consumer privacy, and briefly discuss broad implications for policy and regulation.*

[1] The author would like to thank AT&T for their support of this work. All analyses, exposition and opinions are exclusively those of the author.

1. Introduction

Over recent years there has been ongoing debate regarding the sensitivity and value of data collected by various types of companies on the Internet, including social networks, search engines, web browsers, advertising networks and Internet Service Providers (ISPs). A typical argument is that since ISPs carry all the network traffic generated by their customers, they must have a unique and unfettered ability to "see" and collect consumer data, and therefore should be subject to special privacy and other regulations and concerns.

However, an extensive 2016 study (Swire et al.[2]) details why ISP data is in fact more limited technically than may be believed, and is in many ways less comprehensive and valuable than the data collected by other consumer-facing technology companies. It was argued that applications such as search engines and social networking services collect data providing greater consumer insight than ISPs, and also excel at tracking their users across multiple devices and contexts. Furthermore, the types and diversity of data sources and platforms at such companies often insulates them from data limitations experienced by ISPs and other packet-level services, such as encryption.[3] By their very nature, most consumer-facing technology companies need to interact with their users in a direct, unencrypted fashion.

In this paper, we elaborate on and extend some of the points made in Swire et al., especially in light of the pervasive use of modern machine learning methodology and artificial intelligence (AI) in the leading consumer-facing technology companies. Using Google[4] and Facebook as case studies, we give particular attention not only to the sheer scale and volume of data collected by such companies, but to its diversity, and especially its unprecedented *intimacy*. We will further discuss the powerful efforts these companies have exerted in machine learning, artificial intelligence, and statistics to harness their data troves, resulting in the ability to learn, and deploy at scale, highly predictive models for both collective and individual consumer behavior, and to make subtle and accurate inferences that go far beyond the data actually collected. We shall argue that the combination of data volume, diversity, intimacy and modeling provides insights about consumers --- and privacy concerns --- that are historically unrivaled and are still rapidly expanding. The potential threats to privacy can be neither understood nor combated without accounting for the role played by machine learning and AI. We conclude with discussion of the implications for policy and regulation, including the need for a data- and

[2] "Online Privacy and ISPs: ISP Access to Consumer Data is Limited and Often Less Than Access by Others", by P. Swire, J. Hemmings, and A. Kirkland. Institute for Information Security and Privacy, Georgia Tech, May 2016. Available at http://peterswire.net/wp-content/uploads/Online-Privacy-and-ISPs-1.pdf.

[3] See e.g. Google Transparency Report. HTTPS Encryption on the Web, Google. Accessed: 2 May 2018 (as of April 28, 2018, 81% of web pages loaded on Chrome in the U.S. used HTTPS and 89% of web browsing time on Chrome in the U.S. used HTTPS).

[4] In 2015, Google re-organized its corporate structure and created Alphabet as the holding company for Google's core search business and all other affiliated businesses. For ease of reference, we will generally refer to Google throughout the paper.

technology-neutral privacy framework that anticipates the powerful uses of machine learning, and that can adapt to rapid changes in technology and markets.

While this paper shall focus on the history, data and practices of Google and Facebook, we note that many of the arguments presented also apply to other major consumer-facing technology companies such as Amazon, Apple, and Microsoft, all of whom also collect massive and ever-expanding data on their colossal user bases. Not coincidentally, Amazon and Apple have also rapidly expanded their machine learning and AI research, development and recruiting efforts in recent years.[5] It should be noted that in addition to having among the most advanced and mature machine learning and AI efforts in the technology industry, our choice of Google and Facebook as exemplars of the trend is in part due to their relative candor around their efforts, as their researchers often publish many of their findings in the open scientific literature. Privacy and consumer profiling concerns involving the companies, such as the ongoing investigations into how companies like Cambridge Analytica are using Facebook data for political advertising, have also been of broad public interest.[6]

2. Data Volume and Diversity

To provide the backdrop for our later remarks and arguments regarding the powerful intimacy of Google's and Facebook's data, and the central role that modern machine learning and related disciplines play in the exploitation of that data, we first detail the tremendous volume and diversity of consumer data sources the two companies have amassed. While there are distinct differences in the strategies the two firms have pursued to amass their data empires, and in the nature of their actual data sources, they have arrived at similar positions of data dominance. In Figure 1 below, we provide a partial taxonomy of the extensive data collection sources, platforms and networks of the companies that we shall discuss; further detail and documentation is provided in the Appendix.

We begin with a discussion of Google. Since its founding in 1998, Google has systematically created a data empire whose scale and scope are unprecedented. In its earliest years, Google was a highly focused and specialized company,[7] offering just their core search engine service. But as was quickly discovered at the dawn of the consumer Internet, search is the most data-intensive of businesses. Early efforts to "catalog" the content of the Web using hand-coded human expertise or knowledge (such as Yahoo!'s original hierarchical index) were quickly

[5] See Sawers, Paul. "Amazon to Open Machine Learning R&D Hub in Barcelona." *VentureBeat*, 7 September 2017; Richman, Dan. "Amazon Hires Carnegie Mellon Machine-Learning Expert as Google Expands Its Own AI Initiatives." *GeekWire*, 16 June 2016; and Mannes, John. "Apple Makes the Case that Even Its Most Banal Features Require a Proficiency in Machine Learning." *TechCrunch*, 19 October 2017 for just a small sampling of the frenzied activity in machine learning and AI hiring at Amazon and Apple recently.

[6] See e.g. Confessore, Nick. "Cambridge Analytica and Facebook: The Scandal and the Fallout So Far." *The New York Times*, 4 April 2018.

[7] "Google Acquires Usenet Discussion Service and Significant Assets from Deja.com." News from Google, 12 February 2001.

overwhelmed by the exponentially growing content. Since click-throughs provided an objective measure of search result quality, the race for data supremacy was on --- the more users a search engine had, the more feedback or "training data" it received to improve its search algorithms via machine learning, which would in turn bring more users and data.[8] This cycle was accelerated by the introduction of pay-per-click advertising against keyphrases as the dominant monetization model for search. It seems fair to say that Google effectively won this first data arms race sometime around 2003, and it has comfortably dominated the search market ever since. [9]

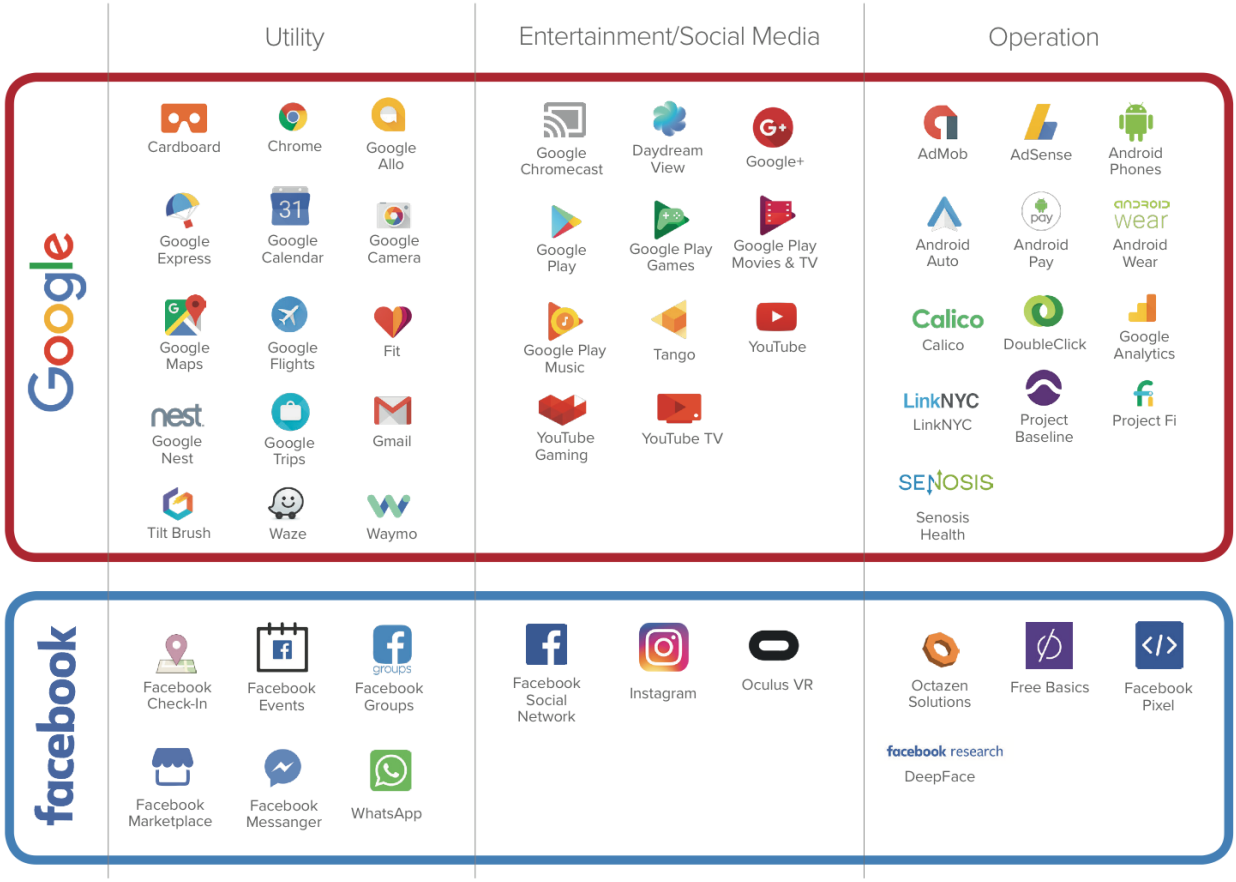


Figure 1. Partial taxonomy of the rich and varied consumer data sources collected by Google and Facebook.

Beginning in the early 2000s, Google gradually expanded the scope of its services and therefore the data it collects on consumers. The list of distinct Google products and services easily numbers in the hundreds, and was amassed through a combination of external corporate acquisitions and internal development efforts.[10] While many of these efforts, especially the earlier ones, were related to the core search business, most of them are not. Rather, these

[8] Elsewhere this author and others have detailed why this cycle essentially never terminates --- i.e. there is no such things as "too much" data --- due to the "heavy-tailed" statistical properties of language.

[9] Google's search engine market share is nearly 90% in the U.S. Stat Counter. StatCounter. "Search Engine Market Share United States of America April 2017 – April 2018." Stat Counter, 2018.

[10] "The Google Acquisition Tracker." CB Insights, 2018.

services and products span a great variety of human activity, and collectively provide Google with an expansive view of its users (and even non-users). It seems implausible that the scale and scope of this expansion[11] can be explained by occasional experimentation in new markets; rather, it appears to be a core part of corporate strategy.

An only partial taxonomy of the various types of consumer and other data Google has acquired includes the following broad categories:

Search: In this oldest, original category, for its entire history Google has collected the queries entered by users of its search engine. Google can associate search histories with individuals either directly (if they log into a Google account), or indirectly (via information like IP address, device type and operating system, GPS coordinates, and techniques such as browser fingerprinting). Search is also special since the Web largely consists of free-form, natural language text, which must be crawled and indexed. As we shall discuss, the particular challenges of search data call for massive machine learning efforts, another area in which Google is dominant.

Browsing History: Similar comments hold for general web-surfing data (i.e. the sequence of URLs visited by a user, as opposed to the search terms they type into the Google search engine). Google collects browsing history for any user of its Chrome browser, which is by far the dominant browser globally,[12] and is the default installation on all Android devices. Chrome is also used to power third-party browsers such as Opera, further increasing Google's browsing market share and data collection.[13]

Media Preferences: Over time, Google has created a stable of services whose usage gives them a detailed profile of a user's tastes and preferences in media consumption of many varieties. These include online video consumption (via YouTube),[14] music (via YouTube and Google Play Music) and other services, including books (via Google Books), news (via Google News), photos (via Picasa and Google Image Search), and many other categories. Of course, information about media consumption from these services is amplified by user searches in these same categories of items, as well as shopping interests from services such as Google Shopping.

Location, Movement and Travel: From its inception Google has possessed information regarding the physical location of its users via IP addresses and their close association with

[11] In one two-year period, Google spent $17 billion on acquisitions, outspending Apple, Microsoft, Amazon, Facebook and Yahoo! combined. See D'Onfro, Jillian. "Google Has Spent More on Acquisitions than Its Top Five Rivals Combined." *Business Insider*, 15 January 2014.

[12] See StatCounter "Browser Market Share WorldWide April 2017 – April 2018." StatCounter, 2018. Chrome is the most popular browser and has been installed by more than 2 billion users worldwide.

[13] Shankland, Stephen. "A Nail in the Coffin for Firefox? Mozilla Struggles to Redefine Browser." *CNET*, 11 April 2016.

[14] YouTube currently has 1.5 billion users and an estimated 78% of U.S. video and multimedia site visits. The Statistics Portal. "Leading Multimedia Websites in the United States in November 2016, Based on Market Share of Visits." Statista, 2016.

geographic coordinates.[15] But such mappings can be imprecise compared to actual GPS data, and also are less applicable to mobile devices. So over time Google has developed and acquired a collection of services that provide much broader and more precise geolocation data on its users. Much of this data is collected by products that are infrastructure, rather than consumer-facing applications.

A notable effort in this regard is the Android Operating System (OS) that runs all Android devices, and which gathers the GPS coordinates of those devices on behalf of all location-based services. Android devices also collect location data based on nearby celltowers and Wi-Fi networks. Further, location data is gathered by the growing wireless network Google has been building out, including Google Fi,[16] Wi-Fi partnership with Starbucks, LinkNYC in New York City,[17] and many similar projects. In recent years, Google has made significant investment in services that not only determine where users *are*, but where they *will* be or *plan* to be in the future. The navigation services on Google Maps and the popular mobile app Waze (acquired for almost $1 billion in 2013) clearly invite users to share short-term driving or travel plans, and via GPS can track actual progress on trips. On a longer temporal and geographic scale, Google Flights (acquired as ITA Software for $676 million in 2010) provides views into users' commercial airline travel plans and purchases.

Social Activity: In addition to an explicit social networking application (Google+), there are myriad sources from which Google can extract information about the friends and other social, business and family relationships of users. These include Google Contacts, a cloud-based address and contact list management service; Google Hangouts, an Internet videoconferencing service; Google Voice, (call forwarding and voice messaging); and many other products with a social or sharing component, such as YouTube.

Communications and Documents: Some Google services permit a more unrestricted view into the lives of users --- effectively anything they "volunteer" to share with Google via the use of products like Gmail. At least until recently, Google algorithms inspected the content of Gmail messages for ad personalization purposes. Similarly, Google Docs provides another free-form data channel, and Google Calendar allows users to document their entire schedules on company servers.

Device Usage: A broad, rich and detailed source of behavioral data that Google enjoys derives from the dominance of the Android mobile OS, which provides data on virtually everything a user does on their device.[18] In addition to massive volumes of data in the categories above, this includes detailed data on app usage.

---

[15] As a demonstration, simply type "what is my IP address" into Google and visit a site like http://ip-api.com/ to see how much this address conveys.

[16] Project Fi. Project Fi, A Phone Plan from Google, Google. Accessed 2 May 2018.

[17] Google Fiber. Starbucks Wi-Fi from Google, Starbucks. Accessed 2 May 2018.

[18] Android currently has more than 2 billion monthly active users. Popper, Ben. "Google Announces Over 2 Billion Monthly Active Devices on Android." *The Verge*, 17 May 2017.

Residential Activity: One of Google's more recent and systematic forays seems to be products and services that provide the opportunity to collect data inside the home. Most notable in this regard was the acquisition of Nest Labs for $3.2 billion in 2014. Nest provides "smart" devices for home monitoring and control, such as thermostats, smoke alarms and cameras (via Dropcam, which was acquired shortly after Nest in 2014 for $555 million). There is also an extensive list of "Works with Nest" partners that provide a wide selection of home control, monitoring and surveillance products.[19] Google Fiber, Chromecast and Google Voice are other services that provide data on residential activity.

The Google data taxonomy we provide above mentions only a fraction of the hundreds of products Google offers and companies it has acquired. It also omits a number of notable past efforts that Google may significantly invest in again, such as health monitoring and data collection from its ambitious but discontinued Google Health platform. Related efforts include the active Google Fit Android exercise and health app platform, and ongoing efforts that open up entirely new categories of consumer data for Google, such as the Google Glass "ubiquitous computer" and Waymo self-driving cars.

Even though a typical user might use just a few Google products on a regular basis, it is worth contemplating the comprehensive view of an individual provided to Google by the data collected. Regular use of Google Search will already provide a tremendously detailed and private background profile of a user's personal, professional, social, medical and commercial interests. Use of Gmail will provide regular, free-form natural language data on the user's personal and professional activities. Navigating using Google Maps or Waze will track the user's movements. And use of any of the above on a mobile device, or the Android OS more generally, will tell Google accurate geographic coordinates for the user at frequent intervals. By integrating the data from just these few of the more common products, Google knows who you are, what you're doing, and where you are in real time throughout a typical day. As privacy expert Bruce Schneier puts it, "Google knows more about what I'm thinking than I do, because Google remembers all of it perfectly and forever."[20]

What about Facebook's data sources and acquisitions? Facebook is certainly more of a "walled garden" than Google, with a dominant primary service that has virtually no serious domestic rivals, and a host of supporting products, such as Instagram and Facebook Messenger. But despite its narrower offerings than Google and a less acquisition-oriented strategy, the very nature of Facebook's offering provides it with a similarly powerful trove of consumer data.

Once again, the sheer scale of this data is unprecedented, with 1.28 billion active daily users worldwide and more than 2 billion active monthly users as of September 2017.[21] More importantly, while Google obtains data diversity through multiple channels (e.g. shopping

[19] Works with Nest. Here's Everything that Works with Nest, Nest. Accessed: 2 May 2018.
[20] "Data and Goliath", Bruce Schneier (W.W. Norton and Company, 2014)
[21] Facebook Newsroom. Stats, Facebook. Accessed: 2 May 2018.

interests and habits through search, location data from Google Maps and Waze, mobile usage through Android, etc.), Facebook achieves the same via users simply willingly providing this information directly through the social network. Facebook users post their location and upload photos, indicate their social relationships by explicitly declaring and labeling them (e.g. spouses or partners), express their interests by liking content and ads or joining groups, and so on.

In addition to its core social networking site, Facebook has developed and acquired additional platform services with enormous user bases, including Facebook Messenger (1.3 billion active users), Instagram (800 million active users) and WhatsApp (1.5 billion active users). Facebook also uses its Facebook Connect log-in tool and Social Plugins (e.g. the Like, Send and Share buttons) to collect data about its users' activity on thousands (if not millions) of third-party websites and apps.

Like Google's users, Facebook users provide the company with a steady daily stream that provides data on the entire sweep and scope of an individual life: social interactions and friendships; physical location, movement and travel (from the pervasive use of Facebook's check-in feature, posting of location photographs, or simply user text posts describing where they are); private communications (via Messenger, WhatsApp, and Instagram); purchases and shopping interests (Marketplace); and calendar and schedule information (by posted and liked calendar events in a user's stream).

Of course, both Google and Facebook have created massive advertising platforms and use ad-tracking technology that gives them further insights into the online and offline behavior of their users. As is well known, together the two companies have over 50% of the US market for digital advertising, while the nearest competitor has less than 5%.[22] Furthermore, most of recent advertising revenue growth has been captured by the two companies,[23] so their lead is not only large, it is also growing rapidly. Both Google and Facebook perform tracking on third-party sites and services in ways that may not be obvious to consumers.[24] Thus, the more successful they are in monetizing users' data via advertising, the more reach they obtain to collect further data about users' activity on third-party sites and apps.

3. Data Intimacy

Despite the great volume and diversity of data collected by Google and Facebook summarized above, this is not alone what distinguishes them from so many other Internet services and applications. Rather, it is the *intimacy* of their data, and the powerful inferences that can be made from it when combined with large-scale AI and machine learning algorithms and models.

---

[22] McNair, Corey. "US Ad Spending: Google and Facebook to Capture over One-Quarter of the Market." *eMarketer Report*, 18 April 2018.
[23] See, e.g., O'Reilly, Lara. "The Race Is on to Challenge Google-Facebook 'Duoploy' in Digital Advertising." *The Wall Street Journal*, 19 June 2017 and Sullivan, Laurie, "Google, Facebook Duopoly Takes Between 60% and 70% of U.S. Ad Market Share." *MediaPost*, 2 November 2017.
[24] See Christl, Wolfie, "Corporate Surveillance in Everyday Life." A Report by Cracked Labs, June 2017.

Let us begin with a discussion of the highly personal --- and indeed, often private --- nature of the data that is routinely collected at the individual consumer level by just the core service of Google, which is of course web search. Far from mere data packets that may even be encrypted, Google directly sees the personal interests, plans, purchases, fears, hopes, fantasies and secrets of its users. Google users routinely conduct searches that reveal their fitness or medical condition (e.g. by searching for medications, specialists and terminology), financial health (e.g. by searching for wealth managers or bankruptcy advice), shopping and purchases, sexual interests, and so on.

There is overwhelming anecdotal and systematic evidence for the routine intimacy of search queries, that users search "as if unobserved". For example, the extensive empirical research of economist Seth Stephens-Davidowitz[25] has demonstrated quantitatively that large populations of users regularly conduct Google searches that reveal sexual orientation, underlying medical conditions such as depression, hidden social and cultural biases such as racism, and undesirable or even criminal behaviors such as child mistreatment. Further and more importantly, it is demonstrated that such conditions and behaviors are revealed on Google at a rate that is far higher than they are revealed in more public forums, such as surveys or via purchasing behavior related to those conditions or behaviors. In other words, Google's own data strongly demonstrates that not only do "private" conditions and behaviors exist (which we already knew), but that they are ritually shared by users with Google in a way they are not anywhere else.

One might argue that unless a Google user is conducting searches while logged into an actual Google account, and that their account is clearly linked to their real-world identity, these intimate queries nevertheless remain somehow "anonymous". Not surprisingly, this line of thinking has been debunked repeatedly due to the "fingerprinting" properties of a sequence of allegedly anonymous queries. As far back as 2006, it was demonstrated that the anonymous AOL queries over a few months of a specific but unknown individual were sufficient for a *New York Times* reporter to quickly identify, locate and contact that individual, who happened to be a 62-year old widow living in Lilbrun, Georgia.[26] Given that AOL's search service was dwarfed by Google's even at that time, it's implausible that the power of "anonymous" Google queries to identify specific real-world users isn't considerably greater. In the decade since this incident, there have been innumerable articles and studies documenting the powerful profiling and

[25] See e.g. "Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are", Harper Collins, 2017; "Essays Using Google Data", Seth Stevens-Davidowitz, Doctoral Thesis, Harvard University 2013 (available at https://dash.harvard.edu/bitstream/handle/1/10984881/StephensDavidowitz_gsas.harvard_0084L_11016.pdf?sequence=1).%25C2%25A0); and links to published research articles and *New York Times* pieces available at http://sethsd.com/.

[26] Barbaro, Michael and Zeller Jr., Tom. "A Face Is Exposed for AOL Searcher No. 4417749." *The New York Times*, 9 August 2006.

identifying properties of Google searches, as well as the potential exploitation and dangers from such intimate data.[27]

Indeed, the tangible monetary value to Google of even a single sufficiently intimate query is sometimes revealed through the prices in Google's dominant (and almost only) source of revenue, which is advertising. For example, for many years, one of the most expensive keywords for advertisers in Google's AdWords platform has been "mesothelioma" --- which is a form of cancer whose most prevalent cause is asbestos exposure --- with the cost of just a single click being many hundreds of dollars.[28] Why is this particular search term so much more valuable to advertisers than related ones such as "cancer" or "asbestos poisoning" or even "asbestos cancer"? The reason is that "mesothelioma" is a relatively technical and obscure medical term that one is most likely to hear from a doctor. So unlike related but less technical terms, this word is much more indicative of an actual diagnosis (as opposed to curiosity, research, etc.), and is thus much more valuable to the attorneys who bid on it, hoping to represent victims in suits brought against the employers who exposed them to asbestos. This is but one of the innumerable examples in which the choice of specific or intimate language in search terms acts as a "private signal" to Google and its advertisers about the very personal conditions of Google users. The intimacies revealed to Google are effectively limited only by the range of human behavior and interests, since so many users conduct searches as if they were entirely unobserved and use highly specific language.

What about the intimacy and value of the data collected by Facebook? Facebook users clearly and knowingly operate in a much more public forum than a search engine. While Google users may have varying degrees of realization of how much they are inadvertently "sharing" with Google itself, Facebook users are very deliberately broadcasting their posts, photos, locations and updates with at least their direct friends (which average in number in the hundreds, and frequently are in the thousands), and often beyond. Because of the very purposes of Google (information, research, shopping, etc.) and Facebook (socialization), and the private vs. public spheres, one might expect that the data collected by Facebook would be considerably less intimate than that collected by Google.

However, perhaps exactly because of the impression that they are sharing among "friends" or like-minded users, the intimacy and privacy of Facebook data is again striking, and not often replicated in other forums or sources. For example, it is widely known that users routinely express their emotional state in Facebook posts either through language or more basically through the use of emojis.[29] Emotion sharing on Facebook is so prevalent that there is a community of scientists who study the effects and user satisfaction derived from such

---

[27] Dewey, Caitlin. "You Are What You Google Search." *The Washington Post*, 16, December 2014.

[28] See e.g. Jones, Russ. "1,000,000 Top High Paying CPC, Adwords and Adsense Keywords for 2015." GrepWords, 3 January 2015.

[29] The rather lengthy list of emojis offered in Facebook's status update feature include ones labeled "excited", "blessed", "happy", "sad", "hopeful", "optimistic", "concerned", "nervous", "pissed off", "heartbroken", "stressed", and "overwhelmed".

interactions.[30] Indeed, a controversial scientific study conducted by Facebook researchers convincingly demonstrated that not only do users routinely reveal their emotional states, those emotional states are actually contagious within the Facebook network[31], so use of Facebook not only records but actually influences user emotions.

Like Google users, Facebook users also routinely reveal opinions, beliefs or affiliations that might carry social stigma, and that they would be more reluctant to reveal in everyday life. A recent *New York Times* survey of private groups on Facebook[32] discussed groups devoted to marital infidelity, marijuana growing, military veterans with behavioral health problems, believers in a flat earth, and myriad other self-interest groups. Reports of frequent racist posts, media and groups on Facebook are common, including amongst individuals who clearly would not "publicly" affiliate with such sentiments, such as a group of ten incoming Harvard freshmen whose acceptances to the university were recently revoked after it was revealed they had shared jokes and memes about race, sexual assault and child abuse on Facebook.[33] Allegedly "private" behaviors (including drunken or otherwise compromising photos) or attitudes on Facebook are sufficiently prevalent and informative that corporate human resource departments now systematically monitor the Facebook activity of both prospective and current employees.[34]

Perhaps even more striking than the direct revelations people make on Facebook about their moods, beliefs, attitudes and behaviors are the surprising *hidden* inferences that can be made from them. A 2013 scientific article[35] involving Facebook users and data begins:

> *"We show that easily accessible digital records of behavior, Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender."*

---

[30] E.g. for just one relatively recent example see "Social Sharing of Emotions on Facebook: Channel Differences, Satisfaction and Replies", N. Bazarova et al., *Computer Supported Cooperative Work,* 2015. Available at https://blogs.cornell.edu/socialmedialab/files/2013/12/Social-sharing-of-emotions-on-Facebook.pdf

[31] "Experimental evidence of massive-scale emotional contagion through social networks", A. Kramer, J. Guillory, J. Hancock. *Proceedings of the National Academy of Sciences,* 111(24), 2014. Available at http://www.pnas.org/content/111/24/8788.full. See also Meyer, Robinson. "Everything We Know about Facebook's Secret Mood Manipulation." *The Atlantic*, 8 September 2014.

[32] "Behind the Velvet Rope of Facebook's Private Groups", *The New York Times*, July 16, 2017. Available at https://www.nytimes.com/2017/07/16/business/behind-the-velvet-ropes-of-facebooks-private-groups.html?mcubz=0&_r=0. 201

[33] "Harvard Rescinds Admissions Offers Over Offensive Memes on Facebook", *The Guardian*, June 5, 2017. Available at https://www.theguardian.com/education/2017/jun/05/harvard-rescinds-admissions-offers-offensive-memes.

[34] See e.g. "Should Companies Monitor Their Employees' Social Media?" *The Wall Street Journal*, October 22, 2014. Available at https://www.wsj.com/articles/should-companies-monitor-their-employees-social-media-1399648685.

[35] "Private Traits and Attributes are Predictable from Digital Records of Human Behavior", M. Kosinski, D. Stillwell, T. Graepel. *Proceedings of the National Academy of Sciences*, 110(15), 2013. Available at http://www.pnas.org/content/110/15/5802.full.pdf.

In other words, using only a user's "likes", and not their explicit textual posts or other declarations, it is possible to accurately infer potentially highly personal and private attributes. The article continues:

> *"The model correctly discriminates between homosexual and heterosexual men in 88% of cases, African Americans and Caucasian Americans in 95% of cases, and between Democrat and Republican in 85% of cases. For the personality trait 'Openness,' prediction accuracy is close to the test–retest accuracy of a standard personality test."*

The methodology of this and similar studies (see Figure 2) is as revealing as the findings themselves. Starting with raw data given by a table or matrix of roughly 10 million entries indicating the "likes" of a population of 58,000 Facebook users, the researchers first used a sophisticated "dimensionality reduction" method known as singular value decomposition to automatically extract a much smaller set of informative "features" to represent each user. These feature representations were in turn given to a standard statistical algorithm to produce predictive models for each of the targeted categories (sexual orientation, race, political party, etc.). Thus, the raw data on the collective population is transformed into a higher-level model that permits accurate (and intrusive) inferences about specific individuals that were not present in their raw data at all. As we shall discuss further in the next section, the power of Google's and Facebook's data is realized largely via such machine learning methods.

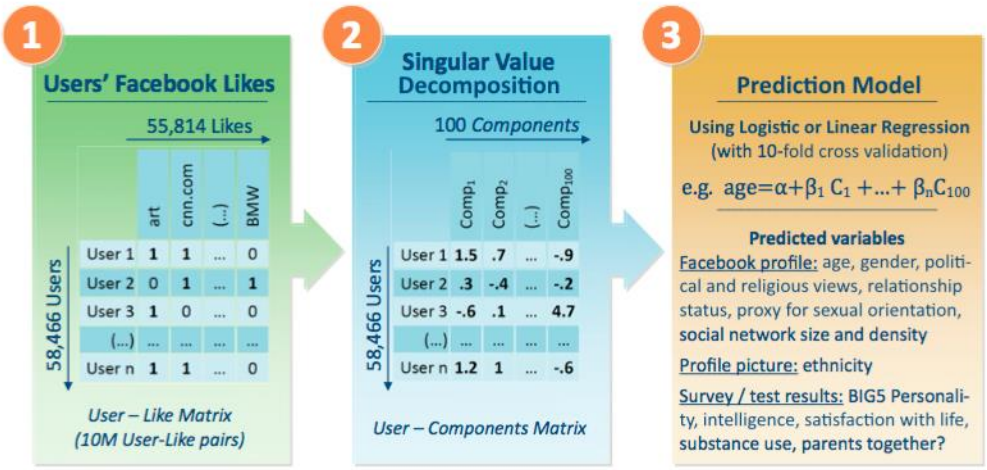
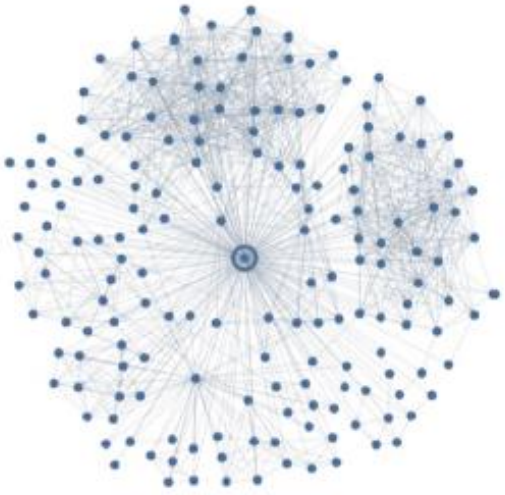


Figure 2. Reproduced from Kosinski et al.(http://www.pnas.org/content/110/15/5802.full.pdf), and describing machine learning methodology used to predict potentially sensitive attributes such as religious beliefs, sexual orientation, substance use and many others from Facebook users' "likes".

The example above shows the surprising power and intimacy of even the most rudimentary behavioral data in an environment as rich as Facebook --- after all, "likes" are mere binary approvals, presumably impoverished compared to the language in posts and updates or the

information in shared photographs. Yet when situated in the context of the data provided by all the other users, they can be extraordinarily predictive of even the most private details. As another example, consider the apparently impossible problem of accurately predicting which of a Facebook user's friends is their romantic partner, using *only* the pattern of connectivity amongst their friends --- thus ignoring all user profile or identify information, their posts or "likes", etc. In Figure 3 below, the data for each user X consists only of dots representing X's friends and links representing pairs of X's friends that are friends themselves, the so-called "ego network" of user X.

Another remarkable study,[36] again employing advanced machine learning methods, showed that such anonymous social relationship data permits accurate identification of romantic partners for over 55% of users --- orders of magnitude higher than random guessing, since the typical Facebook user has hundreds of friends. (See Figure 3.). Thus, for myriads of users who are Facebook friends with their romantic partner, but have chosen not to identify them as such, they are easily inferred anyway. Even further, Facebook's data, algorithms and models are capable of identifying social relationships that its users *are themselves unaware of*, as in a recent case in which the "People You May Know" feature suggested an unknown distant relative to a user.[37]



type	embed	rec.disp.	photo	prof.view.
all	0.247	0.506	0.415	0.301
married	0.321	0.607	0.449	0.210
married (fem)	0.296	0.551	0.391	0.202
married (male)	0.347	0.667	0.511	0.220
engaged	0.179	0.446	0.442	0.391
engaged (fem)	0.171	0.399	0.386	0.401
engaged (male)	0.185	0.490	0.495	0.381
relationship	0.132	0.344	0.347	0.441
relationship (fem)	0.139	0.316	0.290	0.467
relationship (male)	0.125	0.369	0.399	0.418

Figure 4. The performance of different measures for identifying spouses and romantic partners: the numbers in the table give the *precision at the first position* — the fraction of instances in which the user ranked first by the measure is in fact the true partner. Averaged over all instances, recursive dispersion performs approximately twice as well as the standard notion of embeddedness, and also better overall than measures based on profile viewing and presence in the same photo.

Figure 3. Reproduced from Backstrom and Kleinberg (https://arxiv.org/pdf/1310.6753v1.pdf), and illustrating how the local friendship networks of Facebook users (such as that on the left) can be used to accurately identify a user's romantic partners via a machine learning approach.

[36] "Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook", L. Backstrom and J. Kleinberg. Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing, 2014. Available at https://arxiv.org/pdf/1310.6753v1.pdf. See also Stinson, Elizabeth. "Facebook Inches Closer to Figuring Out the Formula for Love." *Wired*, 12 November 2013.

[37] Hill, Kashmir. "Facebook Figured Out My Family Secrets, and It Won't Tell Me How." *Gizmodo*, 25 August 2017.

4. The Centrality and Ubiquity of Machine Learning

Above we have already given some concrete examples of the power of machine learning approaches in making accurate predictions and intimate inferences about Google and Facebook users. We now elaborate on these methods, discuss their critical importance to the two companies, and discuss their longstanding and aggressive recruiting, hiring and acquisition practices in machine learning. As we shall detail, despite the scale and automation involved, the science and engineering of machine learning is a very human-centric process, and there are strong rich-get-richer effects not only from the data, but from the recruiting and hiring of the highly talented and specialized personnel involved. There is thus a very broad competitive cycle at work that strongly favors these companies. As Google and Facebook have increased their market shares and growth capture, they have doubled down on machine learning hiring, which in turn permits them to extend their advantages in digital advertising and market share in their core services and elsewhere. This in turn brings them ever-more data to complete the cycle.

Machine learning is the modern science underlying the construction of large-scale predictive models from massive data sets. It is a mixture of topics from areas as diverse as statistics, probability theory, pattern recognition, algorithms, artificial intelligence and most recently, distributed systems. While its origins lie in the 1980s, in recent years the data explosion enabled by the Internet has made machine learning one of the most important scientific fields, and one that has even entered the popular consciousness. A detailed examination of the manifold ways in which machine learning is central to practically everything that Google and Facebook do is beyond our scope, but we will try to provide some brief context.[38]

We begin by briefly sketching the engineering pipeline underlying the construction of large-scale predictive models from raw consumer data. In its broad outline this pipeline is not particular to Google and Facebook, but they are certainly among the companies that have perfected its deployment at massive scale.

[38] Levy, Steven. "How Google is Remaking Itself as a "Machine Learning First" Company," *Wired*, 22 June 2016 for a recent and extensive article about the long technical and hiring history of machine learning at Google. The practices at Facebook are more recent but similar in scale and centrality.

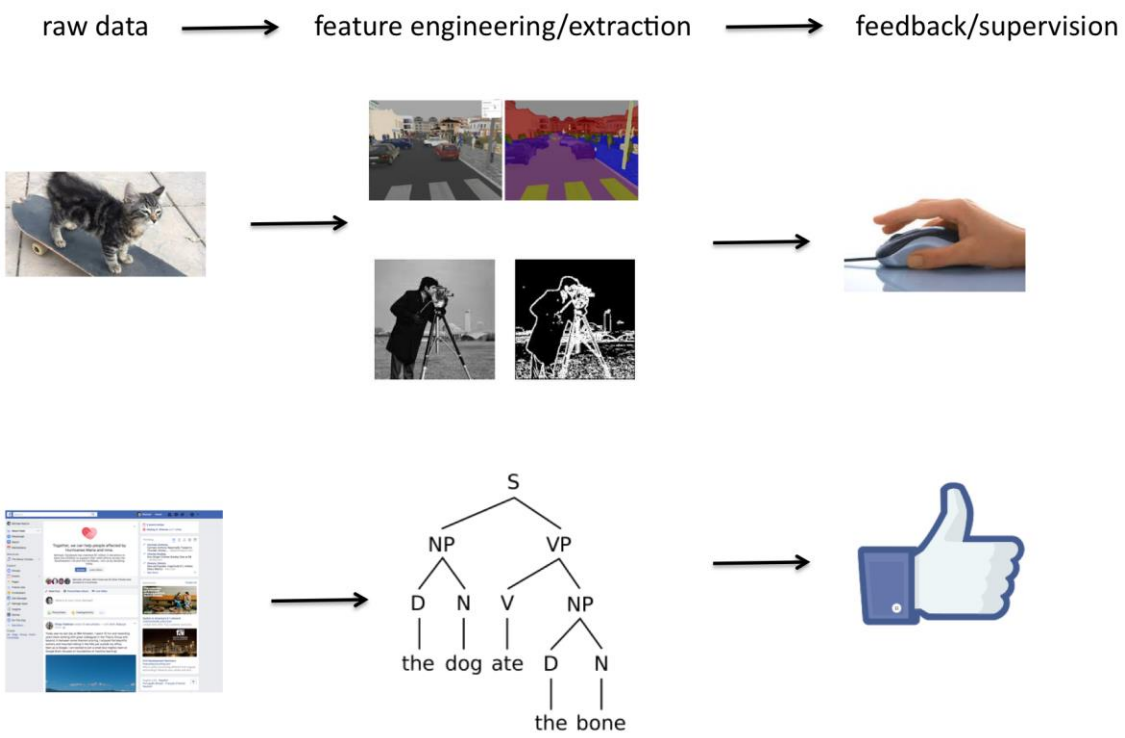


Figure 4. The machine learning pipeline. Raw data (such as digital images or Facebook posts, left) are pre-processed by often sophisticated algorithms to extract higher-level properties or features (such as objects or edges in images, or grammatical parsing of sentences, middle), which is then annotated with various forms of explicit or implicit user feedback (such as clicks on ads or Facebook likes, right).

The pipeline (see Figure 4) begins with the raw consumer data collected by the companies --- e.g. the streams of user activity on Facebook (status updates, photo and content sharing, likes, etc.) and the search queries (and myriad other sources) from Google users. In its raw forms, much of this data is inconveniently represented as words and potentially complex natural language phrases and sentences, or as pixel color values in images. In the parlance of machine learning, the raw data is "unstructured" and difficult to describe or extract meaning from directly.

The first step is thus known as "feature extraction" or "feature engineering", which are the terms used to describe processes that re-represent the raw data streams into higher-level abstractions that have more structure and encode more directly the underlying meaning and intent in the data. Examples would be identifying objects and edges in images, or parsing an English sentence in a Facebook post. The development of algorithms for such feature extractions is actually extremely challenging, and itself the source of many decades of intense research.[39] Many of these subtasks have entire sciences of their own, and both Google and Facebook are flush with influential researchers whose careers have been devoted to such subtasks.

[39] It is worth noting that the striking advances in so-called "deep learning" sometimes permit the automation of feature engineering in certain problems (notably image classification and speech recognition) by building sufficiently complex, layered models that learn internal representations of the most relevant features.

Feature engineering turns the raw, unstructured data streams into structured objects with more meaningful and informative representations that are also much more amenable to machine understanding. For many machine learning tasks, the next step is to annotate such data with user feedback, which in the field's terminology is sometimes referred to as "labels" or "supervision". The basic idea is that if individual data items or events (such as sentences, photos, documents, web pages, etc.) can be identified as relevant or irrelevant, good or bad, etc. then one can use sample data to train a predictive statistical model. Both companies use myriad forms of explicit and implicit user feedback, but common examples are Google users clicking on ads or organic search results, and user "likes" on Facebook.

The combination of feature extraction with user feedback or supervision sets up a classical statistical modeling problem: the raw user streams have now been transformed into <x,y> pairs, where x is some structured representation of complex data items like documents, sentences or images, and y is a signal indicating whether x is "good", "bad" or in between. The challenge is then to take a (very) large sample of such data pairs, and build a predictive model --- i.e. a model that given a *new*, previously unseen x, can accurately predict the associated feedback y. This challenge is precisely the domain of modern machine learning. The science behind principled, accurate and computationally efficient algorithms for this problem is beyond our scope, but has been the focus of the field for 30 years, and of statistics for decades before that.

The pipeline does not end here, because deploying such models and algorithms at the scale of Google and Facebook also requires herculean systems engineering. At these companies one cannot store "the data" in a single computer or even a single geographic location --- it is simply too massive, and arriving too fast and in too many places, due to the colossal worldwide user bases. So the algorithms and models of machine learning have to be implemented in a distributed fashion and coordinated in a computationally efficient yet consistent fashion. This presents all manner of challenges in network communication, data consistency and integrity, synchronization, etc. Not surprisingly, these companies also have armies of stellar and experienced network and systems engineers who work closely with machine learning scientists.

The sketch above describes the powerful machine learning pipelines developed by Google and Facebook, but does not explicitly address why they are actually necessary and in fact the core of the companies' services and businesses. After all, we have already documented the massive volumes and variety of raw data they hold; why isn't this data alone sufficient? Why can't Google and Facebook simply "look up" anything they need to know about users in their raw data? Why is it important to use machine learning to make "predictions" or "inferences" beyond what is in a user's history?

We have already suggested the first answer to such questions, which is that there is obvious value (both monetary and in terms of improving functionality) in being able to make inferences that lie well outside the immediate confines of the services provided. While your relative appetite for kitten photos might be directly revealed by your Facebook "likes", your political

affiliation and sexual orientation may not be --- but they can be *inferred* (as the studies we have cited have demonstrated), and knowing such information is much more valuable to advertisers (and therefore to Facebook). Similarly, being able to infer or predict that the word "bike" is a synonym for "bicycle" when modified by "mountain", but not when modified by "Harley Davidson" is of obvious functional and monetary value to Google, since it benefits both users and advertisers who are either searching for or selling bicycles or motorcycles.

So one reason that machine learning is important is that it enables valuable predictions that go beyond the raw data generated by users. A related but more technical reason is that many consumer data sets --- such as Google search queries, or Facebook "likes" or repostings --- have what are called *heavy-tailed* distributions. Informally, this means there are innumerable events in the data that are individually rare (and thus hard to estimate or model) but collectively frequent (and thus cannot be ignored). Examples include highly specific search queries such as "dentist open Saturdays Fairmount neighborhood Philadelphia" or very precise user demographic targeting on Facebook, such as female users living in the Midwest ages 18-25 who have "liked" the *Wall Street Journal* and the online game Overwatch. While each of these rather specific events might be relatively infrequent, in the aggregate events "similar" to them are extremely common and combinatorially explosive in number, and cannot be modeled or even identified "by hand". Instead Google and Facebook use advanced machine learning techniques to build extensive, large-scale language, social and other models that automatically discover and build predictive models, often by sharing or generalizing data across similar events.

Virtually all of the myriad inferences Google and Facebook make regarding their users (some of which we have mentioned) are made possible by machine learning. While we might be able to articulate and measure particular relationships that we are curious about --- such as whether there is a correlation between one's political views and one's choice of a Mac or PC --- these might not be the ones that turn out to be meaningful, predictive or profitable, and there are far too many to exhaustively enumerate and test them. Far better to have highly scalable, data-driven machine learning algorithms that can directly find the most important relationships without much or any human intervention.

The algorithms of machine learning and the models they produce are largely automated once in operation. But the development of these algorithms, their improvement and evolution, their implementation in a distributed, cloud-based computing environment, and their specialization to the idiosyncrasies of new and ever-changing data sets remains a highly technical, research-intensive, and human-centric activity. Not surprisingly, this is again an area where Google and Facebook have maintained competitive dominance for many years through aggressive and systematic recruiting practices. Google is without question one of the earliest and largest employers (probably *the* largest, though of course any such information is proprietary) of machine learning PhDs, including not only fresh graduates, but many senior, tenured academics

they have lured from top universities.[40] Facebook has rapidly followed suit in recent years.[41] Both companies are well-known in the academic machine learning research community; not only do their own scientists often publish in and attend the top journals and conferences, they are well-represented on the editorial committees of such venues, as well as reliable financial sponsors at the most generous levels.[42] Google's acquisitions in recent years have had a strong focus on machine-learning and AI-based companies such as DeepMind (purchased for $625 million, and whose recent success in the development of a learning-based, world-class Go-playing program made headlines when it defeated a champion player from Korea[43]).

There is a cyclical, rich-get-richer dynamic with respect to the machine learning recruiting efforts of Google and Facebook. Competition for PhD-level scientists and engineers in areas like machine learning and computer science is more intense than ever, in the technology industry and beyond. In addition to established competitors like Amazon and Microsoft, virtually every consumer-facing Silicon Valley startup is desperately seeking expertise in these areas, as well as more traditional industries such as finance. But especially the smaller companies find luring top candidates away from Google and Facebook exceedingly difficult, because research and engineering remain social activities, and thus the best people want to be where the other best people already are, where the data is the most plentiful and most diverse, where the computing infrastructure has the largest scale, and so on. Thus, Google's and Facebook's data empires reinforce their dominant recruiting positions, and vice-versa.

5. Policy Implications and Recommendations

We conclude by providing some perspectives on privacy policy and regulation raised by the arguments and observations made above.

From a privacy perspective, perhaps the most important overarching conclusion is that *the "intimacy" of consumer data cannot be measured by the number of bits crossing a pipe*, or similarly crude metrics that fail to account for the nature, diversity and content of the data, and its potential uses for modeling and inference. It is both possible and common that the highest-volume data sources (such as the fragmented and possibly encrypted packets passing through a core router in the Internet) can reveal relatively little about the end-users who generate that traffic, while much lower-volume and more specialized data sources (such as Google search queries or Facebook likes) can both directly and indirectly reveal the most private and personal

[40] See https://research.google.com/pubs/MachineIntelligence.html for just a sampling of the hundreds of machine learning PhD researchers and engineers at Google, and their staggering output of publications, software and systems.

[41] Facebook's research site at https://research.fb.com/research-areas/ lists hundreds PhD researchers in the closely related areas of applied machine learning, data science, and artificial intelligence.

[42] Google had top billing among dozens of corporate sponsors at NIPS, the premier machine learning conference, in 2015 (https://nips.cc/Conferences/2015/Sponsors), as well as at many other prominent machine learning conferences and workshops. The same is true of Facebook.

[43] See Sang-Hun, Choe and Markoff, John. "Master of Go Board Game Is Walloped by Google Computer Program." *The New York Times*, 9, March 2016 for one example of the extensive coverage.

details about end-users. The widespread application of machine learning to specialized consumer data sources is deliberately designed to extract personal and actionable insights about both individual users and collective behaviors.

The facts we have described here render it nonsensical to have a fragmented privacy framework in which different policies and regulations govern different types of online consumer data. While there are specialized privacy regulations for domains such as health and financial data, it is important to have consistency in the privacy treatment of online consumer data --- especially in an era in which much of the actionable (and potentially sensitive) insight in the data is not explicit, but is inferred or discovered by sophisticated algorithms and statistical models, as we have explained. In particular, it is both naïve and misleading to formulate privacy policy or metrics based only on the amount or apparent "source" of data --- one must also anticipate how private or intimate the *inferences* that could be made from the data might be. And such anticipation is infeasible for a variety of reasons --- not least that it would itself already require machine learning expertise comparable to that present in companies such as Google and Facebook. The fact that the machine learning methods employed in such companies are amongst their most tightly guarded intellectual property is yet a further barrier.

This argues for a regulatory privacy framework that comprehensively covers the diverse range of data being used commercially and applies consistent privacy requirements. Policymakers should also take a forward-looking approach to privacy, and not overly focus on specific data types or practices (which are likely to become obsolete quickly due to the rapidly changing nature of technology). A technology-neutral approach that can adapt quickly to new technical and market developments is called for.

The proprietary nature of the features, algorithms, and models used by consumer-facing Internet companies is sometimes presented as an argument against privacy and other regulations, the claim being that the revelation of these methods would compromise valuable trade secrets. However, we would note that many other data- and algorithm-intensive industries are subjected to notable regulatory restrictions and audits that do not seem to significantly compromise intellectual property. Examples include laws preventing the use of "protected attributes" (such as race or gender) in lending decisions[44] which are often made partially or entirely algorithmically; regulations governing the behavior and robustness of trading algorithms in financial markets (such as the MiFID regulations[45] regarding algorithmic and high-frequency trading in the European Union); and the explicit development of predictive models for criminal justice that avoid discrimination while seeking to protect individual privacy.[46]

In these cases, laws and regulations seek a balance between limiting and auditing algorithmic behavior, and still providing protection and incentives for intellectual property. There is no

[44] See e.g. https://www.fdic.gov/regulations/compliance/manual/4/iv-1.1.pdf
[45] See e.g. https://www.fca.org.uk/mifid-ii/8-algorithmic-and-high-frequency-trading-hft-requirements
[46] See e.g. http://www.nycja.org/resources/details.php?id=1388

reason to believe that similar regulations and auditing mechanisms could not be developed for consumer privacy protections in Internet services. We encourage policymakers, scientists and engineers to explore and develop technologies for the automated monitoring of consumer-facing services on matters of privacy, data and modeling practices, in a way that preserves the intellectual property of proprietary algorithms.

Appendix

Large online platforms collect extensive data about consumers from a variety of sources and build and deploy large-scale predictive models for virtually every aspect of consumer behavior. The privacy debate tends to focus on individual services (e.g., search or social network service) and not the extensive advertising networks and platform capabilities that collect data from third-party apps and websites. Online companies have very personal data from a variety of sources that is more valuable to marketers than general web browsing data. They also have data that is considerably more goal- and intent-oriented than raw Internet traffic data.

While the paper focuses on the impacts of data collection and analyzation by integrated online platforms, this Appendix focuses specifically on the kinds of information that these platforms are monitoring, collecting and selling regarding consumers' behavior. Companies like Google and Facebook have invested billions of dollars acquiring and building online services that will attract consumers and generate valuable data. They also have implemented policies that are designed to enable the creation of comprehensive profiles, while limiting the user's ability to turn off data collection.

Platform Data Collection

As discussed further below, consumers cannot avoid data collection from the integrated online platforms as a practical matter. Google's Android Operating System (OS) has more than 2 billion monthly active users[47] and Google operates seven other service platforms with at least 1 billion users – Google Maps, YouTube, Chrome, Gmail, Search and Google Play.[48] Facebook's social network service has more than 2.1 billion users and Facebook operates three other giant service platforms – WhatsApp (1.5 billion users), Messenger (1.3 billion users) and Instagram (800 million users).[49] Indeed, eight of the ten most popular mobile apps in the U.S. are owned by Google and Facebook.[50] Given the massive scale of these platforms, it is very difficult in practice for consumers to change their social network service or mobile Operating System (OS), or to avoid using any of Google's and Facebook's affiliated services.

Google and Facebook also collect data from millions of third-party websites and have the most extensive reach of any ad networks.[51] According to a 2016 Princeton University study, Google trackers were found on 75% of the top million websites on the Internet and Facebook trackers were found on 25% of these websites.[52] As discussed further below, Google operates the industry's largest online ad network, covering over two million websites and 650,000 apps that reach over 90% of Internet users.[53] Facebook's Audience Network extends the reach of its advertising platform to thousands of third-party

---

[47] *Id*.
[48] Popper, Ben. "Google Announces Over 2 Billion Monthly Active Devices on Android." *The Verge*, 17, May 2017.
[49] Molla, Rani. "WhatsApp is now Facebook's Second Biggest Property, Followed by Messenger and Instagram." *Recode*, 1 February 2018.
[50] Frommer, Dan. "These are the 10 Most Popular Mobile Apps in America." *Recode*, 24, Aug. 2017.
[51] Englehardt, Steven and Arvind Narayanan. "Online Tracking: A 1-Million Measurement and Analysis." Princeton University, 2016.
[52] *Id.*
[53] Google AdWords. *Reach Customers on the Web and in Apps – Across Devices* Google. Accessed: 2 May 2018.

apps and websites.[54] Facebook reports that over 1 billion people see an ad through its Audience Network every month.[55]

By virtue of the massive audiences they have consolidated, their extensive data collection activities and sophisticated machine learning analysis, Google and Facebook are leading the market for online advertising. Google and Facebook are expected to capture a combined 56.8% of digital ad spending and 27.6% of total media expenditures in the U.S. this year.[56] They have an even larger combined share (84%) of the global digital advertising market, excluding China.[57] And by 2019, Google and Facebook are expected to have combined advertising revenues greater than that of all TV ad spending.[58] These massive advertising networks, in turn, facilitate data collection across millions of third-party websites and apps, as evidenced in the examples to follow. From a data analytics perspective, the comprehensive and highly personalized consumer profiles compiled by large online platforms are more valuable than the sum of the parts.

Google

Advertising: Google offers a wide variety of digital advertising services. AdSense allows website publishers to place targeted display ads on their site, while AdMob enables mobile app developers to place similar display ads. The AdWords program for advertisers includes Display Ads embedded in websites, mobile apps, online videos and Gmail messages, and Search Ads appearing next to Google search results. The DoubleClick Digital Advertising Solutions provides an integrated ad-technology platform for both publishers and advertisers. Google Analytics is a freemium web analytics service that tracks and reports website traffic for advertising and other purposes. Google acquired DoubleClick in 2007 and AdMob in 2010 to help build its industry-leading digital advertising business.

Google's advertising services enable it to collect data about web browsing on millions of third-party websites and apps. In particular, Google uses cookies and similar technology to customize ads on Google properties (e.g., Google Search, YouTube) and to track users on other websites and apps for customized advertising.[59] The Google Display Network alone reaches 2 million websites and 90% of Internet users.[60] More than 12 million websites utilize AdSense[61] and more than 1 million apps use AdMob.[62] There are estimates that as many as 30-50 million websites use Google Analytics.[63] Google's Customer Match

[54] Facebook Business. *Create an Ad on Facebook*, Facebook. Accessed 2 May 2018.
[55] Facebook Business. *Businesses Can Now Connect Over 1 Billion People Through Audience Network*, Facebook, 12 January 2018.
[56] McNair, Corey. "US Ad Spending: Google and Facebook to Capture over One-Quarter of the Market." *eMarketer Report*, 18 April 2018.
[57] Sullivan, Laurie. "Google, Facebook Account for 84% of Digital Investments, Excluding China." *MediaPost*, 5 December 2017.
[58] *Id*.
[59] Google Privacy & Terms. *Types of Cookies Used by Google*, Google. Accessed: 2 May 2018.
[60] Google AdWords. *Choose How You Want to Reach Your Customers*, Google. Accessed: 2 May 2018.
[61] BuiltWith. *Websites Using Google Adsense*, BuiltWith. Accessed 2 May 2018.
[62] AdMob by Google. *Why Choose AdMob?* Google. Accessed 2 May 2018.
[63] McGee, Matt. "As Google Analytics Turns 10, We Ask: How Many Websites Use It?," *Marketing Land*, 12 November, 2015.

program allows advertisers to use their online and offline data to target customers across Google services.[64]

In 2016, Google announced it would combine third-party web browsing data with other data from a user's Google account, which reversed a commitment Google made to the FTC when it acquired the DoubleClick advertising network.[65] Privacy groups filed an FTC complaint challenging the change as deceptive and unfair.[66]

Android Operating System (OS): Google's Android OS has more than 2 billion monthly active users.[67] Android OS collects device and usage data, including device identifiers, web browsing, voice and text messaging logs and cell tower and Wi-Fi location. Google requests access to all this data as part of the Android set-up process.[68] Google also offers Android Auto for connected cars and Android TV for smart TVs.[69]

Android Pay: Android Pay tracks consumers' spending both online and offline. Android Pay is supported on many third-party apps and mobile sites, so vendors can accept payment directly from a user's Google-linked cards.[70]

Chrome Web Browser: The Chrome web browser is the most popular web browser in the U.S.[71] It has been installed by more than 2 billion users worldwide.[72] The Chrome browser collects web browsing history and bookmarks across devices by default if a user logs in with their Google account. Google Chrome can also remotely install code that activates the microphone without obtaining permission first.[73] If a user activates Data Saver mode in Chrome, Google can collect data about a user's visits to all HTTP websites by routing traffic through a Virtual Private Network (VPN).

Chromecast: Google is branching out into new platforms, including Google Chromecast, which is technology that is built into TVs and mobile devices to stream entertainment.[74]

---

[64] AdWords Help. About Customer Match, Google. Accessed: 2 May 2018.
[65] Drozdiak, Natalia and Jack Nicas. "Google Privacy-Policy Change Faces New Scrutiny in EU." *The Wall Street Journal*, 24 January 2017.
[66] Complaint, Request for Investigation, Injunction, and Other Relief Submitted by Consumer Watchdog and Privacy Rights Clearinghouse filed 16 December 2016.
[67] Popper, Ben. "Google Announces Over 2 Billion Monthly Active Devices on Android." *The Verge*, 17 May 2017.
[68] Swire, Peter, et al. *Online Privacy and ISPS*. Georgia Tech Institute for Information Security & Protection Working Paper, 2016.
[69] See e.g. Vaughn, Scott. "Android Auto, CarPlay, and Data Tracking." *Berla*, 23 November 2016.
[70] Profis, Sharon. "The Android Pay Details Google Didn't Tell You." *C-Net*, 3 June 2015.
[71] StatCounter. *Browser Market Share United States of America April 2017 – April 2018,* StatCounter, 2018.
[72] Smith, Craig. "40 Google Chrome Statistics and Google App Statistics (August 2017)." *DMR*, 12 November 2017.
[73] Gibbs, Samuel. "Google Eavesdropping Tool Installed on Computers Without Permission." *The Guardian*, 23 June 2015.
[74] Chromecast Built-In. *TVs With Chromecast Built-In*, Google. Accessed 2 May 2018.

Education Services: Google offers K-12 education services as free services. Google has been criticized for using these services to attract new users from a young age and track students when they use their credentials to log in to non-educational Google services.[75]

Fitness Tracking: Google has entered the fitness tracking space with the introduction of their health-tracking platform Google Fit[76] and the Android Wear smartwatch.[77] These products allow Google to track the user's specific location and the time and length of visits to specific locations.

Gmail: Google's Gmail service has 1.2 billion users.[78] Google announced in June of 2017 that it would stop scanning Gmail user's emails to target ads to them, but it still collects information about who is sending and receiving Gmail messages and it could choose to begin using email content for advertising again in the future.[79]

Google+ Social Network: Google launched its own social network, Google+, in 2011. Google+ offers many of the same features as other popular social networks. The service has 111 million users and gives Google access to a user's activity on the site, including their connections and posts.[80]

Google Assistant and Google Home Smart Speaker: Google Assistant[81] provides voice-enabled search capabilities across multiple devices. Google Assistant is now available on more than 400 million devices, including smartphone, TVs and the Google Home smart speaker.[82]

Google Drive: Google Drive is a cloud storage and file back-up service. It now has more than 800 million users.[83] It includes the Google Docs collaboration tool for writing, editing and sharing documents. Concerns have been raised that Google is automatically scanning documents in Google Drive, which Google claims is to prevent abuse and protect users.[84]

Google Maps: Google Maps has over 1 billion users.[85] Google's Travel and Maps services provide detailed information on the physical movements of consumers over time and space, including future travel plans.[86] Google is now using a visual positioning system (VPS) to increase location accuracy for

---

[75] Herold, Benjamin. "Google Acknowledges Data Mining Student Users Outside Apps for Education." *Digital Education*, 17 February 2017.
[76] Google Fit. *Step Up Your Fitness* Google. Accessed 2 May 2018.
[77] Graziano, Dan. "Google is Finally Taking Fitness Seriously With Android Wear 2.0." *C-Net*, 19 February, 2017.
[78] Smith, Craig. "18 Amazing Gmail Facts and Statistics (August 2017) by the Numbers." *DMR*, 22 November 2017.
[79] Bergen, Mark. "Google Will Stop Reading Your Emails for Gmail Ads." *Bloomberg*, 23 June 2017.
[80] Denning, Steve. "Has Google+ Really Died?" *Forbes*, 23 April 2015.
[81] Google Assistant. *Meet Your Google Assistant*, Google. Accessed: 2 May 2018.
[82] Google the Keyword. "How Google Home and the Google Assistant Helped You Get More Done in 2017," Google, 5 January 2018.
[83] Popper, Ben. "Google Announces Over 2 Billion Monthly Active Devices on Android." *The Verge*, 17 May 2017.
[84] Salam, Maya. "Google Docs Glitch That Locked Out Users Underscores Privacy Concerns." *The New York Times*, 31 October 2017.
[85] Popper, Ben. "Google Announces Over 2 Billion Monthly Active Devices on Android." *The Verge*, 17 May 2017.
[86] Sullivan, Laurie. "When Google VPS Becomes Next Search Targeting Option." *Media Post*, 22 May 2017.

marketers. Google is tracking parking spots on Google Maps and can utilize the information to determine how long a user was at a location.[87]

Google Play: The Google Play app store, which also includes Google Play Games, Google Play Movies & TV, Google Books and Google Music, allows Google to collect data about all app usage on Android phones.[88] More than 3.3 million apps are available in the Google Play store.[89]

Health Site Tracking: Google struck a deal with Ancestry.com for access to DNA data in 2013.[90] Google bought a health monitoring start-up called Senosis Health, which turns smartphones into medical devices and collects various health statistics.[91] Alphabet's Project Baseline collected health data from 10,000 people in a study, including but not limited to, heart rate data, sleep data and blood test data.[92] Google founded Calico, which is a new start-up in the health research and development space that uses technologies to understand the process of aging and diseases.[93]

LinkNYC: Through its Sidewalk Labs affiliate, Google is deploying 7,500 LinkNYC kiosks throughout New York City's five boroughs that provide free Wi-Fi and other services supported by advertising revenues.[94] Privacy concerns have been raised about the how data from the kiosks will be collected and used.[95]

Nest: Google's Nest devices (e.g., connected thermostats, cameras, doorbells), and the myriad of devices that network with the Nest, provide detailed data on consumers' physical activities within their own homes. Like the Amazon Echo, the Google Home device enables in-home surveillance in the form of a voice-enabled service. Google continues to aggressively try to move further into the home.[96] Google Home can now manage multiple accounts[97] and uses voice recognition to distinguish[98] users.

---

[87] Murphy, Margi. "Google Maps' Parking Feature Tells You Where You Left Your Car – and How Long You Have Left on Your Meter." *The Sun*, 21 March 2017.
[88] Google Play. *New Movie Releases*, Google. Accessed 2 May 2018.
[89] The Statistics Portal. "Number of Available Applications in the Google Play Store from December 2009 to September 2017." Statista, 2017.
[90] Bergen, Mark. "The Long Game: Google-Backed Calico Partners With Ancestry to Beat the Specter of Aging." *Recode*, 26 July 2017.
[91] The Business Standard. "Google Just Bought an Indian-Origin Professor's Health Monitoring Start-Up." *The Business-Standard*, 16 August 2017.
[92] Rogers, Adam. "That Google Spinoff's Scary, Important, Invasive, Deep New Health Study." *Wired*, 20 March 2017.
[93] Calico. *We're Tackling Aging*, Calico. Accessed 2 May 2018.
[94] LinkNYC. *Frequently Asked Questions*, LinkNYC. Accessed 2 May 2018.
[95] Buttar, Shahid and Amul Kalia. "LinkNYC Improves Privacy Policy, Yet Problems Remain." Electronic Frontier Foundation, 4 October 2017.
[96] The Washington Post, "Why It Matters That Google Home Can Now Identify You by Voice." *The Washington Post*, 4 March 2017.
[97] Ruddock, David. "Google Will Introduce Google Wi-Fi, a $129 Home Wi-Fi Router, on October 4th." *Android Police*, 23 September 2016.
[98] Shields, Nicholas. "Google Boosts Home's Monetization Potential." *The Business Insider*, 17 August 2017.

Search Engine: Google's search engine market share is nearly 90% in the U.S.,[99] and Google earned almost 80% of U.S. search advertising revenues last year.[100] Google pays Apple 34% of advertising revenues to be the default search engine on Safari, which amounted to $1 billion in 2014. It's estimated that Google paid Apple $3 billion last year.[101] Google uses its search engine to direct users to its own products.[102] More than 85% of Google's advertising revenue now comes from its own sites.[103]

Shopping: Google tracks online shopping and what consumers buy while in stores to sell more digital advertising. Since 2014, Google has measured more than 5 billion store visits.[104] In 2017, Google began using data from roughly 70% of credit-card and debit-card transactions in the U.S. to link online and offline data.[105] Even YouTube has been added to this tracking,[106] which has sparked an FTC complaint.[107] Google even launched its own online marketplace called Google Express as a home delivery service for groceries and other items.[108]

Starbucks Wi-Fi: Google provides free Wi-Fi services at thousands of Starbucks locations across the country.[109]

Virtual Reality/Artificial Reality: Google has entered the virtual reality space with several products and platforms, including the platform Daydream,[110] a painting product called Tilt Brush,[111] Google Earth VR,[112] the education tool Google Expeditions,[113] and its own device called the Cardboard.[114] These services will allow Google to have access to customer habits and behavior while utilizing the platforms and devices.

Waymo: Alphabet launched a self-driving tech company that is working towards making self-driving cars available to the public. Waymo is testing the software and will collect data about how people use the cars and how users will want their future transportation to exist.[115]

---

[99] StatCounter. Search Engine Market Share United States of America. StatCounter, 2018.
[100] Marvin, Ginny. "Report: Google Earns 78% of $36.7B US Search Ad Revenues, Soon to be 80%." *Search Engine Land*, 14 March 2017.
[101] Heisler, Yoni. "Google Pays Apple $3 Billion to be the Default Search Engine on the iPhone." *BGR*, 14 August 2017.
[102] Nicas, Jack. "Google Uses Its Search Engine to Hawk Its Products." *The Wall Street Journal*, 19 January 2017.
[103] Dawson, Jan. "Nearly 85% of Google's Ad Revenues Now From Its Own Sites." Twitter, 27 April 2017.
[104] Bergen, Mark. "Google Adds YouTube to Suite of Ad Tools Tracking Retail Suites." *Bloomberg*, 23 May 2017.
[105] Associated Press. "Google Starts Tracking Offline Shopping – What You Buy at Stores in Person." *LA Times*, 23 May 2017.
[106] Bergen, Mark. "Google Adds YouTube to Suite of Ad Tools Tracking Retail Suites." *Bloomberg*, 23 May 2017.
[107] The Washington Post. "Google's New Program to Track Shoppers Sparks a Federal Privacy Complaint." *The Washington Post*, 30 June 2017.
[108] Google Express. *Google Express*, Google. Accessed 2 May 2018.
[109] Google Fiber. *Starbucks Wi-Fi from Google*, Starbucks. Accessed 2 May 2018.
[110] Daydream. *Dream with Your Eyes Open*, Google. Accessed 2 May 2018.
[111] Tilt Brush by Google. *Painting From a New Perspective*, Google. Accessed 2 May 2018.
[112] Google Earth VSR. *Your World Awaits*, Google. Accessed 2 May 2018.
[113] Google Expeditions. *Bringing the World Into the Classroom*, Google. Accessed 2 May 2018.
[114] Google Cardboard. *Experience Virtual Reality in a Simple, Fun, and Affordable Way*, Google. Accessed 2 May 2018.
[115] Laris, Michael and Steven Overly. "Waymo is Giving Away Hundreds of People Access to Their Own Self-Driving Cars." *The Washington Post*, 25 April 2017.

Waze: Google purchased[116] the mapping service Waze to enhance their search capabilities and gain access to more data within the application and Google as a whole.[117] Waze has shared aggregate user data, including driving history and habits, with governments in exchange for real-time information on highways, construction data and city events.[118]

YouTube: YouTube currently has 1.5 billion users.[119] In November 2016, YouTube had an estimated 78.8% share of U.S. video and multimedia site visits.[120] YouTube has also launched YouTube Gaming, YouTube TV and YouTube TV, which all capture ad revenue from AdSense.

Facebook

Advertising: Facebook gives advertisers the ability to run ads across Facebook, Instagram, Messenger and on third-party websites and apps that are part of the Facebook Audience Network.[121] Facebook Audience Network extends the reach of its advertising platform to thousands of third-party apps and websites.[122] Facebook reports that over 1 billion people see an ad through the Audience Network every month.[123]

Facebook Pixel is Facebook's own analytics tool that allows an advertiser on Facebook to measure the effectiveness of their advertising by understanding the actions that people take on their website. [124] The tool allows advertisers to show ads to people who have recently viewed pages or specific products on their website.

Apps on Facebook: Third-party apps on Facebook integrate with a user's Facebook profile to pull various personal data, from work history to timeline posts to birthdates. [125] This occurs when a user logs into a product using the Facebook Login. Until recently, even if a user does not allow permission, the app could utilize information from a "Facebook Friend" about another person. This can also be collected from Facebook Groups within the social network.

---

[116] Lunden, Ingrid. "Google Bought Waze for $1.1B, Giving a Social Data Boost to its Mapping Business." *Tech Crunch*, 11 June 2013.
[117] Nahar, Anish. "Google Maps – The Most Expansive Data Machine." Digital Innovation and Transformation, 5 April 2017.
[118] Olson, Parmy. "Why Google's Waze is Trading User Data with Local Governments." *Forbes*, 7 July 2014.
[119] McNamee, Roger. "I Invested Early in Google and Facebook. Now They Terrify Me." *USA Today*, 8 August 2017.
[120] The Statistics Portal. "Leading Multimedia Websites in the United States in November 2016, Based on Market Share of Visits." Statista, 2016.
[121] Facebook Business. *Introducing Facebook's Audience Network*. Facebook, 30 April 2014.
[122] Facebook Business. *Create an Ad on Facebook*, Facebook. Accessed 2 May 2018.
[123] Facebook Business. *Businesses Can Now Connect Over 1 Billion People Through Audience Network.* Facebook, 12 January 2017.
[124] Facebook Business. The Facebook Pixel, Facebook. Accessed: 2 May 2018.
[125] Komando, Kim. "Facebook is Watching and Tracking You More Than You Probably Realize." *USA Today*, 18 March 2016.

DeepFace: Facebook's facial recognition software DeepFace collects and stores the user's biometric information.[126] DeepFace holds "the largest facial dataset to date." [127] The service has an accuracy rate of 97.35% on the "Labeled Faces in the Wild" dataset. This could allow Facebook and its third-party partners to tailor ads to specific customers based on their mood, age, eye gaze, emotion or other personal features.[128]

Facebook Social Network: Facebook has more than 2 billion active users for its social networking service. [129] Facebook has penetrated 79% of U.S. Internet users, while its Instagram service is second with 32%.[130] It generates almost 80% of mobile social traffic.[131]

Instagram: In 2012, Facebook purchased Instagram, a photo- and video-sharing application.[132] Instagram currently has 700 million active users.[133] Facebook and Instagram share data to better target advertising to consumers, including location data, interests and past searches.

Location Tracking: Unless a user disables the location feature on their mobile device, Facebook can track a user's location even while the app is not actively being used. On many of their platforms and services, like Messenger, a user's location is tracked by default.[134]

Marketplace: Facebook Marketplace allows users to buy and sell within their local community. The platform allows more space for businesses to advertise their products and services to consumers.[135]

Messenger: Facebook Messenger has 1.2 billion active users.[136] Per Facebook's privacy policy, the Messenger service can read contact data, including who is called or messaged, and how often the user communicates with them. [137] Facebook can turn on the device microphone to collect information and connect a user's location data to a message if their location settings are enabled.

Octazen Solutions: Facebook purchased Octazen Solutions, a company focused on contact importer software. Octazen was combined with Google's own contact software to collect and store user credentials that are utilized to sign into Facebook and third-party websites.

---

[126] Glaser, April. "Facebook is Using an 'NRA Approach' to Defend its Creepy Recognition Programs." *Slate*, 4 August 2017.
[127] Taigman, Yaniv, et al. "DeepFace: Closing the Gap to Human-Level Performance in Face Verification." Facebook Research, 24 June 2014.
[128] Glaser, April. "Facebook is Using an 'NRA Approach' to Defend its Creepy Recognition Programs." *Slate*, 4 August 2017.
[129] Constine, Josh. "Facebook Now has 2 Billion Monthly Users … And Responsibility." *Tech Crunch*, 27 June 2017.
[130] Chaffey, Dave. "Global Social Media Research Summary 2017." *Smart Insights*, 27 April 2017.
[131] Kolbert, Elizabeth. "Who Owns the Internet?" *The New Yorker*, 28 August 2017.
[132] Rusli, Evelyn. "Facebook Buys Instagram for 1 Billion." *The New York Times*, 9 March 2012.
[133] Constine, Josh. "Facebook Now has 2 Billion Monthly Users … And Responsibility." *Tech Crunch*, 27 June 2017.
[134] King, Hope. "Facebook Messenger Tracks Your Location by Default." *CNN Money*, 28 May 2015.
[135] Advertisemint. "Facebook Advertising Opportunities Grow as Marketplace Expands to Europe." *Advertisemint*, 1 September 2017.
[136] Constine, Josh. "Facebook Now has 2 Billion Monthly Users … And Responsibility." *Tech Crunch*, 27 June 2017.
[137] Facebook. *Data Policy*, Facebook. Accessed 2 May 2018.

Onavo Virtual Private Network (VPN): In 2013, Facebook purchased the Onavo VPN service, which has been downloaded by millions of Android and iOS users. The app allows Facebook to collect data about the user's web browsing and app usage history.[138]

Shopping: Facebook has started tracking the brick-and-mortar stores that consumers visit and from where they purchase items. Facebook will use phones' location services[139] to track whether people actually[140] walk into the stores after seeing an ad that has been targeted to them.

Social Plugins: Facebook uses its Social Plugins (e.g., the Like, Send and Share buttons) and the Facebook Connect log-in tool, to collect data about users and non-users across the Internet. Nearly half of the top 100,000 most visited websites include one or more Facebook technologies within the site.[141] Facebook Social Plugins alone are used by an over 17 million websites.[142] And, on average, the Like and Share buttons are viewed across 10 million websites daily.[143]

Facebook collects website and app data from people, even if they are logged out of Facebook or do not have a Facebook account.[144] Under Facebook's Data Policy, Facebook may use this information to improve or target the content and ads shown on Facebook.[145]

Virtual Reality: Facebook acquired Oculus VR in 2014 to enter into the virtual reality technology space.[146] The privacy policy of the service states that when a user agrees to the terms and conditions of the policy, there is "information automatically collected about you when you use our services," including when, where and how the user interacts with content on the platform.[147]

WhatsApp: Facebook purchased WhatsApp in 2014 and it was announced in 2016 that WhatsApp would, for the first time, allow the sharing of its user data with parent company Facebook to allow better ad targeting to users.[148] European regulators have voiced concern with this efforts and Facebook suspended the sharing of European user data.[149] WhatsApp currently has 1.2 billion active users.[150]

---

[138] Seetharaman, Deepa and Betsy Morris. "Facebook's Onavo Gives Social-Media Firm Inside Peek at Rivals' Users." *The Wall Street Journal*, 13 August 2017.
[139] Swant, Marty. "Facebook Will Track Whether Ads Lead to Store Visits and Offline Purchases." *Ad Week*, 14 June 2016.
[140] Facebook Business. "In-Store, Meet Mobile: New Ways to Increase and Measure Store Visits and Sales." Facebook, 14 June 2016.
[141] Similar Tech. "Facebook Reach Outside Facebook." *Similar Tech Blog*, 27 October 2016.
[142] *Id.*
[143] Zephoria. "*The Top 20 Valuable Facebook Statistics – Updated April 2018*." Zephoria. Accessed 2 May 2018.
[144] Facebook, Hard Questions: What Data Does Facebook Collect When I'm Not Using Facebook and Why?. Facebook, 16 April 2018.
[145] *Id.*
[146] Zuckerberg, Mark. Facebook post, 25 March 2014.
[147] Oculus. *Oculus Privacy Policy*, Oculus. Accessed 2 May 2018.
[148] Lomas, Natasha. "WhatsApp's Privacy U-Turn on Sharing Data with Facebook Draws More Heat in Europe." *Tech Crunch*, 30 September 2016.
[149] The Independent. "WhatsApp Temporarily Suspends Sharing European User Data with Parent Company Facebook." *The Independent*, 16 November 2016.
[150] Constine, Josh. "Facebook Now has 2 Billion Monthly Users … And Responsibility." *Tech Crunch*, 27 June 2017.

Amazon

Advertising: Amazon Marketing Services allows businesses to advertise products on Amazon with pay-per-click ads.[151] Amazon's digital advertising business is growing rapidly and its revenues from display and search ads on Amazon are expected to rise 63.5% in the U.S. this year.[152]

Alexa Personal Assistant: Utilized in Amazon's Echo and Echo Look devices, Amazon is considering giving transcripts of Alexa's audio recordings to third-party app developers.[153] Alexa is an open source platform that can be utilized throughout the IoT and will now be coupled with Cortana, Microsoft's AI.

Amazon Shopping: As of 2016, Amazon accounted for nearly $1 of every $2 Americans spend shopping online. Research has found that using data on what consumers browse, Amazon selectively raises prices and frequently steers customers towards its own products.[154] Amazon gathers data on everything sold on its platform and matches the data to individual consumers, giving the company insight into what consumers search for and purchase, as well as browse but do not buy.[155]

Amazon Web Services: Amazon enables companies to create scalable big data applications and secure them without using hardware or maintaining infrastructure through its Web Services remote computing platform.[156] Big data applications such as clickstream analytics, data warehousing, recommendation engines, fraud detection, event-driven ETL and Internet-of-Things (IoT) processing are done through cloud-based computing at AWS.

Anticipatory Shipping Model: Amazon's anticipatory shipping model uses big data for predicting what products the user is likely to purchase and when users may buy products.[157] The items are sent to a local distribution center or warehouse, so they will be ready for shipping once the order is placed.

Comprehensive Collaborative Filtering Engine: Amazon is a leader in utilizing a Comprehensive Collaborative Filtering Engine to make product recommendations for buyers.[158] The retail giant's recommendation system is based on a number of elements: what a user has bought in the past, which items they have in their virtual shopping cart, items they've rated and liked and what other customers have viewed and purchased. Amazon calls this "item-to-item collaborative filtering," and has used this algorithm to customize the browsing experience for returning customers.

---

[151] Amazon Marketing Services. *Advertise on Amazon*, Amazon. Accessed: 2 May 2018.
[152] McNair, Corey. "US Ad Spending: Google and Facebook to Capture over One-Quarter of the Market," *eMarketer Report*, 18 April 2018.
[153] CBS News. "If Amazon Starts Sharing Alexa Recordings, Should We Be Concerned?" *CBS News*, 14 July 2017.
[154] LaVecchia, Olivia and Mitchell, Stacy. "Amazon's Stranglehold: How the Company's Tightening Grip Is Stifling Competition, Eroding Jobs, and Threatening Communities." Institute for Local Self-Reliance, November 2016.
[155] *Id.*
[156] Miller, Ron. "Yawn: Amazon Cloud Business Just Keeps Rolling Along." *Tech Crunch*, 27 April 2018.
[157] Bensigner, Greg. "Amazon Wants to Ship Your Package Before You Buy It." *The Wall Street Journal Blog*, 17 January 2014.
[158] Fatourechi, Mehrdad. "The Evolving Landscape of Recommendation Systems." *Tech Crunch*, 28 September 2015.

Echo Look: Amazon's new Echo Look allows owners to place orders through Alexa while the camera sits in their residence.[159] Video and voice data is stored in the Amazon cloud until deleted by the user. According to the company, "designated Amazon personnel may view photos and video to provide and improve our services, for example to provide feedback through Style Check."

Kindle: Amazon acquired Goodreads, a social networking tool which allows users to highlight and share portions of the books they read and share with others.[160] Now integrated with Kindle, Amazon regularly reviews words highlighted in the Kindle to determine a user's interest. The company can then send additional e-book recommendations.

Microsoft

Bing: Bing search engine captures 33% market share in the United States with over five billion monthly search requests.[161] The search engine customizes ads based on consumer behavior on other web sites.[162]

Cortana: Microsoft's Cortana AI assistant, a voice-activated component of Windows OS in phones and computers, is nearly omnipresent throughout Windows 10.[163]
Edge: Edge web browser accounts for nearly 8% of market share for internet browsers in the United States.[164] By default, "Do Not Track" mode is turned off in Microsoft Edge.[165]

Explorer: Internet Explorer web browser accounts for nearly 13% of market share for Internet browsers in the United States.[166] By default, "Do Not Track" mode is turned off in Internet Explorer 11.[167]

LinkedIn: With the purchase of LinkedIn in 2016, Microsoft gained access to millions of LinkedIn users and their profiles. LinkedIn's privacy policy specifically notes that they target ads to people "on and off of our Services through a variety of ad networks and exchanges."[168] It continues by noting that providing

---

[159] Heater, Brian. "Amazon's Camera-Equipped Echo Look Raises New Questions About Smart Home Privacy." *Tech Crunch,* 26 April 2017.
[160] Carmody, Tim. "Amazon to Acquire Goodreads, a Social Network for Book Recommendations." *The Verge*, 28 March 2013.
[161] Smit, Dreyer. "Bing Grabs 33% Market Share in the U.S According to Data {resented by Microsoft." Neowin, 19 August 2017.
[162] Melendez, Steven. "Apple Sends the Ad Industry Scrambling to Preserve Web Tracking ." *Fast Company*, 4 October 2017.
[163] Freedman, Andrew E. "How to Restrict Cortana's Ever-Present Listening in Windows 10." *Laptop Magazine*, 29 July 2016.
[164] Statista, "Market share held by leading desktop internet browsers in the United States from January 2015 to November 2017." Statista, 2017.
[165] Kishhore, Aseem. "Enable Do Not Track and Tracking Protection in IE 11 and Edge." *Online Tech Tips*, 13 March 2018.
[166] Statista, "Market share held by leading desktop internet browsers in the United States from January 2015 to November 2017." Statista, 2017.
[167] Kishhore, Aseem. "Enable Do Not Track and Tracking Protection in IE 11 and Edge." *Online Tech Tips*, 13 March 2018.
[168] LinkedIn. *Privacy Policy*, LinkedIn. Accessed 2 May 2018.

information to LinkedIn "enables you to derive more benefit from our Services" and "it also enables us to serve you ads and other relevant content on and off of our Service."

Skype: Microsoft's Skype services allows users to send and receive voice, video and instant message communications. Microsoft collects usage data about communications, such as time and date, and the numbers are usernames involved in the communications.[169]

Windows 10: Windows 10 has a feature called "Getting to Know You," which can collect "typing history."[170] Simply turning off this feature does not remove the data from the cloud, which must be done separately.

Apple

Apple iOS: Apple iOS captures 40% market share in the United States.[171] The company has reached 1.3 billion devices in use worldwide, including iPhone, iPod touch, iPad, Mac, Apple TV, and Apple Watch models.[172] Apple collects location and other data from users.[173]

Apple Pay: To use Apple Pay at all, a user must unlock the system via a biometric scan with the Touch ID fingerprint system. Once a user's identity is verified, Apple Pay releases payment information via data token, an encrypted bit of data that stands in for a card number in a transaction.

Find My Friends: Previously a downloadable app, the Find My Friends feature is now standard on all iPhones and allows iPhone users to track their contacts if the features are enabled.[174]

Location Services: Apple's location services track the user's location and patterns of movement to make recommendations on commutes, routes, parking, destinations, time spent at destination and more.[175] This feature is connected to Bluetooth, which can recognize when the device is connected to a car.

Safari: Safari accounts for nearly 14% market share of browsers worldwide.[176] Apple's Safari captures 58.8% market share for tablet only browsing.[177]

Siri Personal Assistant: Siri will reveal personal details, including name, telephone number and recent calls, even while the phone is in lock screen.[178]

---

[169] Microsoft. "Microsoft Privacy Statement," Microsoft. Accessed: 2 May 2018.
[170] Gordon, Whitson. "What Windows 10's 'Privacy Nightmare' Settings Actually Do." *Life Hacker*, 5 August 2015.
[171] Hollander, Rayna. "Apple Has Lost iOS Market Share in the US, Europe, and Japan." *Business Insider*, 16 January 2018.
[172] Clover, Juli. "Apple Now Has 1.3 Billion Active Devices Worldwide." *Mac Rumors*, 1 February 2018.
[173] Pangburn, DJ. "How—And Why—Apple, Google, And Facebook Follow You Around In Real Life." *Fast Company*, 22 December 2017.
[174] Arguilar, Nelson. "How to Secretly Track Someone's Location Using Your iPhone." *Gadget Hacks*, 4 October 2015.
[175] Waterson, Jim. "Your iPhone Knows Exactly Where You've Been and This Is How to See it." *Buzz Feed*, 29 April 2014.
[176] StatCounter. "Browser Market Share Worldwide April 2017 – April 2018." StatCounter, 2018.
[177] StatCounter. "Tablet Browser Market Share Worldwide July 2017." StatCounter, 2017.
[178] Brynes, Jeff. "Privacy Concerns with Siri Are Bad News, But She's So Convenient." *App Advice*, 27 February 2017.