Pursuing Usable and Useful Data Downloads Under GDPR/CCPA Access Rights via Co-Design

Sophie Veys, Daniel Serrano, Madison Stamos, Margot Herman,
Nathan Reitinger[†], Michelle L. Mazurek[†], Blase Ur
University of Chicago, [†]University of Maryland

Abstract

Data privacy regulations like GDPR and CCPA define a *right of access* empowering consumers to view the data companies store about them. Companies satisfy these requirements in part via *data downloads*, or downloadable archives containing this information. Data downloads vary in format, organization, comprehensiveness, and content. It is unknown, however, whether current data downloads actually achieve the transparency goals embodied by the right of access. In this paper, we report on the first exploration of the design of data downloads. Through 12 focus groups involving 42 participants, we gathered reactions to six companies' data downloads. Using co-design techniques, we solicited ideas for future data download designs, formats, and tools. Most participants indicated that current offerings need improvement to be useful, emphasizing the need for better filtration, visualization, and summarization to help them hone in on key information.

1 Introduction

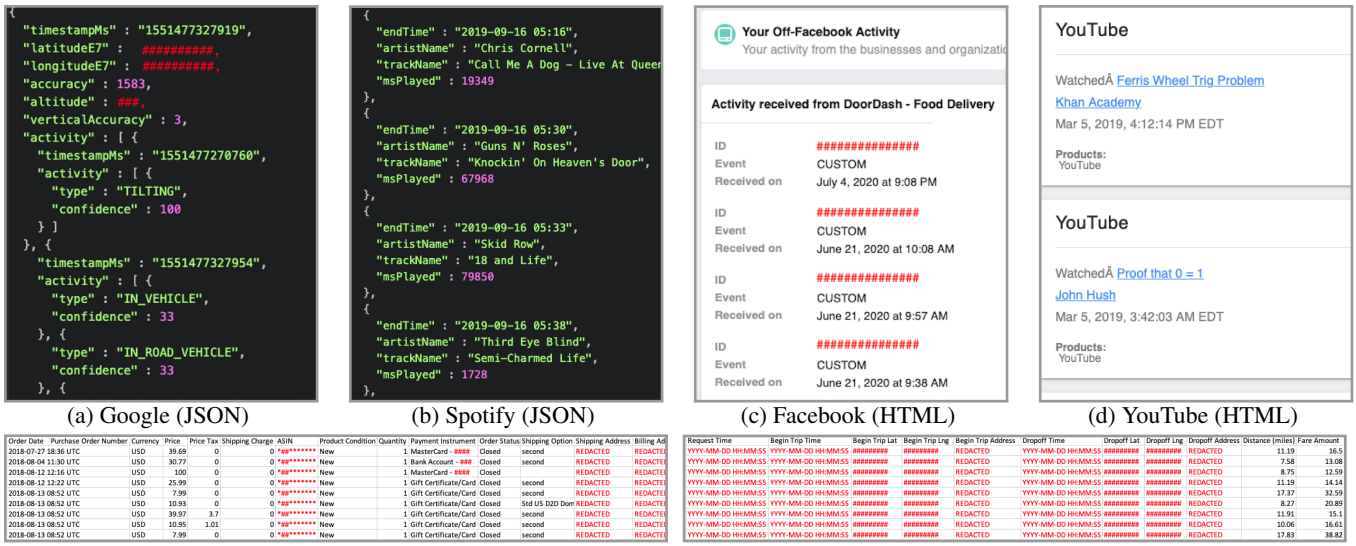
The principle of **data access** states that subjects should be able to obtain a copy of the data that has been collected about them. For decades, this principle has appeared in information privacy frameworks [24]. For example, access is one of the five core facets of the U.S. Federal Trade Commission's Fair Information Practice Principles (FIPPs) [24]. In past decades, while other FIPPs directly impacted consumers (e.g., the principle of notice underpins the ubiquity of privacy policies [66]), the principle of access was mostly ignored. In recent years, however, rights of access have been strengthened. In the European Union, Article 15 [79] of the General Data Protection Regulation (**GDPR**) enshrines a "right of access by the data subject." Similarly, under the California Consumer Privacy Act (**CCPA**), businesses must respond to consumer "requests to know" about data collected about them, enabling them " to access, view, and receive" a copy of that data [76].

Consumers might want access to their data for many reasons. First, data downloads can help users uncover distressing aspects of the online data ecosystem. Prior work has found that consumers can feel uneasy upon seeing evidence of online tracking and data collection [78, 81, 88]. Further, consumers often become upset when they feel that data has been misused or taken out of context [52], including for advertising [27] or politics [34]. In a widely discussed article, Hill used data downloads to expose "secret consumer scores" in which consumers' purchase histories and demographics impact their eligibility for refunds [31]. Access to data is a prerequisite for consumers to modify any incorrect information (the privacy principle of participation) [24]. Additionally, awareness of data collection might encourage users to exercise their right of erasure [9] or motivate other privacy-protective actions.

Privacy concerns aside, there are more practical reasons consumers might want access to their data. Many consumers have data spread across many platforms. For example, a consumer might have pictures published to Twitter, Instagram, and Tumblr. In the event they lose the device on which the original pictures are stored, they might try to reclaim as many photos as possible. Alternatively, a consumer might wish to move from one service (e.g., Spotify) to a competitor (e.g., Amazon Music), yet wish to seamlessly transfer their carefully curated playlists and other personal data. The pursuant right of **data portability**, which enables consumers to transfer personal data across services via interoperable formats, is also enshrined in both GDPR [79] and CCPA [76].

To comply with these legal rights of data access and portability, many companies have begun to offer what we term **data downloads**, which are either files or archives of files containing the identifiable data a business or other data processor has collected about a consumer. Figure 1 shows ex-

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.



(a) Google (JSON) (b) Spotify (JSON) (c) Facebook (HTML) (d) YouTube (HTML)

Order Date	Purchase Order Number	Currency	Price	Price Tax	Shipping Charge	ASIN	Product Condition	Quantity	Payment Instrument	Order Status	Shipping Option	Shipping Address	Billing Ad
2018-07-27 18:36 UTC		USD	39.69	0	0	*##*******	New	1	MasterCard - ####	Closed	second	REDACTED	REDACTE
2018-08-04 11:30 UTC		USD	30.77	0	0	*##*******	New	1	Bank Account - ###	Closed	second	REDACTED	REDACTE
2018-08-12 12:16 UTC		USD	100	0	0	*##*******	New	1	MasterCard - ####	Closed		REDACTED	REDACTE
2018-08-12 12:22 UTC		USD	25.99	0	0	*##*******	New	1	Gift Certificate/Card	Closed	second	REDACTED	REDACTE
2018-08-13 08:52 UTC		USD	7.99	0	0	*##*******	New	1	Gift Certificate/Card	Closed	second	REDACTED	REDACTE
2018-08-13 08:52 UTC		USD	10.93	0	0	*##*******	New	1	Gift Certificate/Card	Closed	Std US D2D Dom	REDACTED	REDACTE
2018-08-13 08:52 UTC		USD	39.97	3.7	0	*##*******	New	1	Gift Certificate/Card	Closed	second	REDACTED	REDACTE
2018-08-13 08:52 UTC		USD	10.95	1.01	0	*##*******	New	1	Gift Certificate/Card	Closed	second	REDACTED	REDACTE
2018-08-13 08:52 UTC		USD	7.99	0	0	*##*******	New	1	Gift Certificate/Card	Closed	second	REDACTED	REDACTE

(e) Amazon (CSV)

Request Time	Begin Trip Time	Begin Trip Lat	Begin Trip Lng	Begin Trip Address	Dropoff Time	Dropoff Lat	Dropoff Lng	Dropoff Address	Distance (miles)	Fare Amount
YYYY-MM-DD HH:MM:SS	YYYY-MM-DD HH:MM:SS	#########	#########	REDACTED	YYYY-MM-DD HH:MM:SS	#########	#########	REDACTED	11.19	16.5
YYYY-MM-DD HH:MM:SS	YYYY-MM-DD HH:MM:SS	#########	#########	REDACTED	YYYY-MM-DD HH:MM:SS	#########	#########	REDACTED	7.58	13.08
YYYY-MM-DD HH:MM:SS	YYYY-MM-DD HH:MM:SS	#########	#########	REDACTED	YYYY-MM-DD HH:MM:SS	#########	#########	REDACTED	8.75	12.59
YYYY-MM-DD HH:MM:SS	YYYY-MM-DD HH:MM:SS	#########	#########	REDACTED	YYYY-MM-DD HH:MM:SS	#########	#########	REDACTED	11.19	14.14
YYYY-MM-DD HH:MM:SS	YYYY-MM-DD HH:MM:SS	#########	#########	REDACTED	YYYY-MM-DD HH:MM:SS	#########	#########	REDACTED	17.37	32.59
YYYY-MM-DD HH:MM:SS	YYYY-MM-DD HH:MM:SS	#########	#########	REDACTED	YYYY-MM-DD HH:MM:SS	#########	#########	REDACTED	8.27	20.89
YYYY-MM-DD HH:MM:SS	YYYY-MM-DD HH:MM:SS	#########	#########	REDACTED	YYYY-MM-DD HH:MM:SS	#########	#########	REDACTED	11.91	15.1
YYYY-MM-DD HH:MM:SS	YYYY-MM-DD HH:MM:SS	#########	#########	REDACTED	YYYY-MM-DD HH:MM:SS	#########	#########	REDACTED	10.06	16.61
YYYY-MM-DD HH:MM:SS	YYYY-MM-DD HH:MM:SS	#########	#########	REDACTED	YYYY-MM-DD HH:MM:SS	#########	#########	REDACTED	17.83	38.82

(f) Uber (CSV)

Figure 1: Data download excerpts from the six services used in our study. Text in red or replaced with # indicates our redactions.

ample excerpts from data downloads. While data downloads provide unprecedented access to the data companies hold about consumers, their format, organization, and design is highly variable across companies. As we discuss further in Sections 2–3, data downloads can range from individual CSV files to sprawling archives containing hundreds of gigabytes of data. In many cases, consumers are left to decipher files intended to be processed by computers. Many files are in JSON or CSV formats. They frequently use UNIX timestamps (see Figure 1a), rather than human-readable dates and times. Some data downloads even come as files containing a single line millions of characters long. Even archives in more typically human-readable formats like HTML can be disorganized and riddled with both jargon and undefined terminology. These sorts of problems led one journalist to subtitle an article about GDPR data downloads as "138GB of data and no real answers" [58]. Most data downloads appear intended to address both access and portability rights, arguably coming up short at providing humans meaningful transparency about their data.

Motivated by the unique opportunities data downloads afford, but also their apparent usability shortcomings, we asked:

- **RQ 1**: How do users react to both the format and content of their own data downloads?
- **RQ 2**: What information is important for users to see in their data downloads? What practical uses are imagined for this information?
- **RQ 3**: How should data downloads be redesigned to improve transparency and best support users' goals?

To address these questions, we conducted 12 online focus groups with a total of 42 participants. Each focus group centered on one of six companies offering data downloads: Amazon, Facebook, Google, Spotify, Uber, or YouTube. Participants came to the session with their own data download, which they had requested in a previous step of our protocol. In our sessions, participants were given time to explore their files, during which we gathered their opinions about both the format and content of current data downloads. Using co-design techniques, we then led participants through a series of activities designed to elicit their ideas and preferences for making data downloads more intelligible for humans.

Most participants indicated that current offerings need improvement to be usable and useful. Participants were generally unsatisfied with either the format or content of their data downloads, if not both. Despite the usability barriers of current formats, most discovered information they found surprising in their data download, commenting on the unexpected nature of information retained or the lengthy retention period. Participants emphasized the need for better filtration, interactive visualization, more meaningful organization, and summarizations that help them hone in on key information. Participants were also interested in understanding the contents of their data downloads at a higher level, including seeing aggregate statistics and how data is synthesized into inferences. Based on these findings, we offer recommendations for improving the presentation and intelligibility of data downloads.

2 Background and Related Work

We begin by highlighting the legal basis for data access rights. We then discuss prior research on those laws' impacts before focusing on prior work studying data downloads. We also summarize transparency efforts and co-design techniques.

Legal Basis for Access and Portability: The right of access and the right to data portability are provided under both GDPR [79] and CCPA [76]. The right of access mandates

that companies give consumers a full view of the data they hold about them upon request [4]. Both GDPR and CCPA prescribe the content that should be released, but not the format in which to release it. While GDPR Article 12 [79] requires the use of plain and clear language, the focus is on communication regarding the request, rather than the response itself [65]. Whether the response itself should be comprehensible is not fully specified, though mandating intelligible responses would be consistent with data access as a foundational privacy right. The right to data portability encompasses the transferability of data and enables users to change platforms, helping to prevent vendor lock-in. To support data portability, both laws do specify that data downloads should be readable by computers via standardized and interoperable formats [76, 79].

The rights of access and data portability may seem similar at first glance; both stipulate that data be made available to consumers. In fact, in CCPA the right to data portability is included within the right of access. However, this conflation of access and portability impairs comprehensibility. Machine readability and human readability are different standards requiring distinct approaches and mechanisms. We develop recommendations for human-intelligible data downloads.

Studies of GDPR and CCPA's Impacts: While we focus on data downloads under rights of access, prior work has explored other requirements and implications of GDPR and CCPA. Degeling et al. and Utz et al. studied cookie notices, finding a lack of usability in the consent process [18] and discussing the impacts on consumer choice [84]. Politou et al. studied the right to be forgotten and the right to withdraw consent, showing that the need to keep data for legal investigations may conflict with these rights [57]. Bertram et al. reported longitudinal data on how Google complied with those rights [9]. Biega et al. investigated the feasibility of data minimization, which demands that companies collect only the data necessary to satisfy the purpose of collection [11].

Researchers have also explored the effectiveness of the laws [26, 30, 32, 41, 83]. Mahieu et al. argued that GDPR is weakly enforced and would be more effective were it executed on a collective, rather than individual, level [44]. De Hert et al. found a range of interpretations for GDPR's data portability requirements, hypothesizing that data controllers might use formatting loopholes to prevent the full exercise of consumers' rights [30]. Grundstrom et al. [26] and Labadie [41] highlighted some compliance challenges data controllers face.

Data Downloads: While (to our knowledge) we are the first to study data downloads from a design perspective, others have investigated data downloads in other contexts. Martino et al. demonstrated that the right of access can be abused by using forged or publicly available data to make illegitimate data subject access requests (**DSARs**) for other people's data [45]. Boniface et al. also identified vulnerabilities in the authentication process and presented guidelines for improvement [14]. Bufalieri et al. [15], Urban et al. [82], Kröger et al. [40], and Spiller [74] made data download requests to data controllers, quantifying the response time [15, 82], evaluating the completeness of the data [15, 40], and documenting shortcomings in the request and authentication processes [15, 74]. Wei et al. had participants request their Twitter data, which they used to characterize ad targeting on Twitter and personalize a related user study [87]. Alizadeh et al. asked participants to request data downloads from loyalty card providers, interviewing participants about the request process and contents of their files [3]. Our work focuses on the design of data downloads themselves, as opposed to the request process.

Transparency Tools and Data Visualization: Although many users are concerned about their online privacy [20, 35, 46, 70], most do not understand important elements of the data-aggregation process [8, 36, 59, 61, 85, 90]. Profit motives tend to disincentivize full transparency [1, 47]. Researchers have attempted to provide additional transparency without platform support via black-box tools [5, 6, 17, 42]. Even with good intentions, conveying complex technical information to users is challenging [19, 21, 23, 71]. Many researchers have created transparency- and privacy-enhancing tools (**TETs** and **PETs**), such as browser extensions and dashboards [7, 8, 12, 13, 37–39, 43, 49, 49, 51, 55, 56, 62, 64, 67–69, 80, 89, 91]. Some tools highlight the need for effective visualizations in improving user understanding [7, 88]. Researchers also emphasize the need to provide users with direct, fine-grained control [12, 16, 38, 49, 55, 56, 62, 89]. Others argue that focusing on control over personal information unduly burdens users [28, 53, 66].

We take the first step toward the creation of GDPR/CCPA data download TETs and PETs via co-design sessions. Most prior work pre-dates or is unrelated to GDPR/CCPA data downloads, instead focusing on visualizing the types of information readily available to consumers at the time those tools were created. Datta found that dashboards show only a portion of the existing data [17]. We do not pre-select the information we deem interesting, such as inferences [60] or advertising [87]. Instead, we ask participants to highlight their desired content. While we confirm some best practices of data visualization generally [29, 72, 73], our recommendations are specific to the data-collection ecosystem and emerge from co-design based on participants' actual data downloads.

Co-Design: The co-design research method (sometimes called participatory design) includes end users in the design process to leverage the knowledge and skills of end users in collaboration with the expertise of researchers and designers [75]. Co-design has been used in a few prior studies of security and privacy tools [25, 50, 63, 86]. Weber et al. emphasized the importance of establishing a "common language" between participants and researchers [86]. Our own sessions build on the lessons of these prior applications of co-design.

3 Selection and Overview of Data Downloads

To facilitate concrete discussions, we centered each focus group on participants' own data downloads from one of six companies. Here, we explain how we chose those six companies and briefly describe their data downloads for context.

Company Selection: To select popular companies, we examined the privacy policies of the Moz Top 500 Websites [48] to see which let users download their data. We excluded 105 websites that were not in English, illegal (e.g., ThePirateBay), potentially embarrassing, or were unable to be accessed by the researchers. We then filtered for companies that allowed non-California and non-EU residents to make data subject access requests, resulting in 109 websites. We categorized each site using the Alexa Top 500 categories [2], assigning categories based on the service the company provides and the type of data expected to be found in its data download.

We selected companies based on the following criteria:

- A simple request process via a clear, online portal (no emailing, mailing, or calling required)
- Relatively quick fulfillment (less than 10 days when members of the research team requested their own data)
- An easily recognizable and popular company, making it easier to find participants with an active account
- Belonging to a category of company participants would likely use (e.g., social media, entertainment)

Further, we selected companies meeting the above criteria such that the final slate would encompass both the breadth and depth of file information potentially available across all data downloads, ensuring a reasonably representative sample of the types of information available. We chose the following six companies: **Amazon**[1] (shopping); **Facebook** (social media); **Google**[2] (location and search); **Spotify** (entertainment); **Uber** (transportation); and **YouTube** (media).

Data Downloads' Characteristics: Members of the research team requested their own data downloads for these companies and many others, recording the data types and corresponding format of each type of information available. This achieved three goals: (i) it informed us about the types of data available in each download; (ii) it enabled us to identify variation in data formats (e.g., UNIX vs. UTC timestamps); and (iii) it provided us with an initial impression of how human-readable each download was. Although we were able to interpret most of the available data, there were a number of items we could not resolve. For example, we were unable to interpret Uber's "horizontal accuracy" column, which contained values like "30" and "10."

[1] While Amazon offers data downloads for all products, including Kindle and Audible, we omit all but order history to keep sessions focused.

[2] Similarly, Google downloads can be very large and variable, so we omit all Google products except location and search.

Table 1 summarizes key aspects of team members' data downloads from these six companies; results for others may vary. Note that user-uploaded files, such as Facebook photos, YouTube videos, and Google Drive files, retain their original file format in the data downloads and are excluded from Table 1. Informally analyzing our own data downloads, we identified eight classes of information. We found extensive variation in how different companies included and presented data within these classes. For example, five companies' data downloads (all but YouTube) contained some sort of location data. Spotify's location data included the user's address, payment country, payment card postal code, family plan address, and Car Thing accessory shipping address. Facebook, in contrast, included the user's primary location, the current city included in their profile, the IP addresses and locations from which they had ever logged in, and the places where the user had checked in. Google's location data included time-stamped locations, data presumably collected from a phone GPS (latitudes, longitudes, velocities, altitudes), and the type of activity performed at a location. Appendix C gives other examples. These examples are intended to provide context for participants' comments, rather than being exhaustive.

4 Co-Design Study: Method

We conducted a three-part study than ran from July 2020 to September 2020. Part 1 was a screening survey. Eligible participants were asked to download their data from one of the aforementioned six companies and were invited to Part 2. Part 2 determined eligibility for Part 3, a 75-minute co-design session hosted on Google Meet. We recruited participants on Prolific, a crowdsourcing platform that has many advantages [54] over Amazon Mechanical Turk. The appendix contains the text of all survey instruments and focus group guides.

4.1 Participant Selection

In Part 1, participants completed a demographic and screening survey in Qualtrics to provide information that would help us create the co-design sessions. Participants indicated their availability for a focus group and chose the companies on our list of six for which they had active accounts. We compensated $1 USD for this survey, which took on average 2.5 minutes.

Based on the Part 1 responses, we selected prospective participants. We assigned one of the six companies to each participant. In Part 2, we provided participants instructions to request a data download from their assigned company. For companies that provided both HTML and JSON options (see Table 1), we instructed participants to select HTML as it is more likely to be human-intelligible. Google offered location data in both JSON and KML, but we opted for JSON due to the complexity of opening a KML file. Participants completed a second survey to verify they had successfully requested their

Table 1: Key aspects of data downloads, as obtained by the research team. For Amazon and Google, we report on both the subset (*) of the download used in our study and the full (FULL) versions of these data downloads.

Company	File Formats	Includes "ReadMe"?	Time of Receipt	# Folders	# Files	Size
Amazon*	CSV	No	2-5 days	<10	<10	KBs
Amazon FULL	CSV	No	4 weeks	Tens	Tens	MBs
Facebook	JSON or HTML	Yes	Almost instantaneous	Hundreds	Thousands	GBs
Google*	JSON, HTML, KML	Yes	Almost instantaneous	<10	<10	KBs
Google FULL	JSON, HTML, KTML	Yes	Hours	Tens	Hundreds	GBs
Spotify	JSON	Yes	1-2 weeks	<10	<10	KBs
Uber	CSV	Yes	Hours	<10	<10	KBs
YouTube	JSON, HTML	Yes	Almost instantaneous	<10	10–20	KBs

data download by pasting in text (with no identifying information) from the notification email or data download page. The Part 2 survey also asked about their general sentiments toward data access and privacy. Participants were compensated $2 for completing Part 2, which took 11 minutes on average.

Participants who completed Part 2 were invited to Part 3, a 75-minute focus group and co-design session centering on the company for which they had downloaded their data. Group sizes ranged from 3–5 based on participant availability and turnout. We ensured there were no more than two participants per session who were students, and no more than one participant per session with CS or IT expertise. We held two focus groups for each of the six companies, resulting in 12 focus groups in total. Participants were compensated $25 for Part 3.

Due to COVID-19, we held all focus groups remotely as video calls, recording only the audio. We used Google Meet because it provides real-time captioning (transcription). A researcher listened to all audio recordings and corrected the transcripts. Meet requires participants to log in with a Google account, displaying the associated name on-screen. To protect participant privacy, we made five anonymous Google accounts for participants to use. We turned off those accounts' activity tracking and ad personalization. We logged participants out and changed the passwords between sessions.

4.2 Structure of Focus Group Sessions

Each 75-minute session included several activities designed to encourage discussion and inspire ideas about improving data downloads. A third survey was conducted concurrently with the session to facilitate giving participants instructions and collecting written responses. We iterated on the design of our focus group protocol through five pilot sessions with convenience samples. After each, we incorporated feedback from the previous pilot session to clarify the wording of questions and instructions, correct typos, and improve logistics. As suggested by a pilot tester, in our final protocol we screen-shared slides with bullet-point instructions. The survey and slides aimed to help participants stay on track even if they experienced connectivity issues or other interruptions.

Introductions and Guidelines: We began by directing participants to the third survey in order to consent to both participation and audio recording. To help participants get to know one another, we asked participants to introduce themselves with their first name (real or fake) and a fun fact about themselves. In the chat window, we mapped anonymous Google account names (e.g., Participant 1) to the first name provided by the participant, allowing participants to refer to each other by name during discussions. We then gave participants general instructions for the session: to turn cameras on (a requirement aimed to increase engagement), to mute when they were not speaking, and not to take screenshots or make recordings. We reminded participants they were not required to share specific information about themselves or their data.

GDPR/CCPA 101 and Free Exploration: The first activity, intended to provide context about data downloads, was a minute-long overview of GDPR and CCPA. We explained that data downloads are available in part as a right granted to residents of the EU and California. We answered (to the best of our knowledge) any questions participants posed about these laws. We then had participants freely explore their data download for five minutes. We asked participants to inspect the format, content, and organization of the files. For sites with Read Me or HTML overviews (all but Amazon), we gave participants 1–2 minutes to read them. We encouraged them to comment aloud about anything they found interesting.

Scavenger Hunt and Discussion: While free exploration avoids priming participants about what to look at, it can also lead to a lack of engagement. Thus, we next asked participants to complete a scavenger hunt with their data downloads. We provided a list of items to find, such as a deleted message or the timestamp of a purchase. We selected items such that:

- Collectively, the items spanned multiple folders
- Items were not too difficult to interpret
- Some items might interest the participant (e.g., "What 'life stage' does Facebook assign to your friends?")
- Items required scrolling (e.g., "Find an album... that starts with the same letter as your first name.")
- Some items required cross-referencing multiple files

All scavenger hunt items, 6–11 per company, met at least one of these criteria. While the scavenger hunt inevitably introduced some bias, we believe these items helped to expose participants to a broad range of their data, including information they may have overlooked during free exploration. The scavenger hunt lasted 5–7 minutes. For privacy, participants did not enter their answers in the survey, nor read them aloud.

For 10–15 minutes, participants then discussed their first impressions of data downloads. We debriefed the scavenger hunt, asking about experiences navigating the data download and looking for items. We asked about the content and format of these files, as well as data-collection practices in general.

Highlight Activity: In our Qualtrics survey, participants were then given a list of folder names associated with the relevant company's data download and were asked to highlight the categories they would be most interested in seeing. This was designed to identify content participants cared about. There was no limit on the number of items they highlighted.

Data Viz 101: To inform and inspire participants, we held a five-minute introduction to data visualization. We asked participants to browse Information Is Beautiful [33], which visualizes daily news. We chose this site because it offered many options, rather than endorsing one or two specific visualization approaches. We also wanted to avoid visualizations that would alienate participants. Instead, we wanted them to focus on data presentation, rather than content. We asked participants to share examples of visualizations they found particularly interesting or well-designed, as well as examples that synthesized multiple pieces of information. Participants pasted links to visualizations, briefly summarizing what they liked about each. We then used a basic example to show that even simple visualizations can be effective, showing a spreadsheet with two columns ("month" and "number of cats petted") and graphing the data as a line chart.

Sketch Activity: Finally, we asked participants to sketch, either on paper or digitally, their ideal version of a visualization tool for their data download. All prior activities were designed to build up to this activity, which directly supported our ultimate goal: to work with participants to reimagine data downloads. We provided guiding questions, referencing content, formatting, and menu options. We told participants they could use any approach they wanted, but mentioned two possible options: a high-level approach sketching the general layout of a tool and specifying its different options, and a low-level approach focused on representing a specific type of information (e.g., location data). After uploading their sketches to our server, participants were asked to explain them. With participants' permission, we screen-shared their sketches to the group; we have also made them available for download [77]. Participants were then redirected to Prolific for compensation.

4.3 Data Analysis

We analyzed the data from our co-design sessions using affinity diagramming, a method for consolidating qualitative data into emergent groups or themes [10]. Two researchers used Miro, an online whiteboard, to collaboratively affinity-diagram comments from all 12 sessions. We placed meaningful quotes from all the session activities on virtual Post-it Notes, then grouped them with other similar quotes. We determined meaningful quotes to be everything that was shared during a session with the exception of what researchers said and moments when participants required clarification or when they experienced technical difficulties. We framed our groupings around "what" (data content) and "how" (data format). We then isolated themes within those top-level groupings. As needed, we split quotes to ensure they did not contain more than one cohesive idea. If a quote fit into more than one grouping, we duplicated it as needed.

We analyzed all quotes using pseudonyms containing an abbreviation for the company under discussion, Session A or Session B for that company, and an assigned participant number (from 1 to 5) during the session. An example pseudonym is *G-A-1*: the first participant in Session A for Google.

4.4 Protection of Participants

Our protocol was reviewed by the University of Chicago IRB and determined to be exempt. We collected no personally identifiable information. Study-related communication was conducted via Prolific's internal messaging system, which uses pseudonyms to identify participants. As discussed above, we did not ask participants to share their data downloads with us, we created anonymous Google accounts to avoid participants exposing their personal information, and during the session participants identified themselves using only their first name or a pseudonym. We did not video record the session or take screenshots, and we instructed participants not to do so either. Participants consented to audio recording of sessions before completing Part 1 and again before completing Part 3. We reminded participants at the beginning of the session that they were under no obligation to share specific information about themselves or their data. We also told participants that if they said something they did not want on record, they could let us know afterwards and we would delete that portion.

4.5 Limitations

Due to the rich qualitative nature of our study, we had a relatively small sample size (42 participants). We recruited only participants located in the U.S. As is typical on Prolific, our participants skewed younger and more educated than the average population of the U.S. Additionally, our study made technical demands of participants. They needed to download their data (aided by our instructions), join a Google Meet call (requiring a webcam and microphone), and upload a photo of

their sketch. These requirements were listed in our recruitment ad's eligibility section. As a result, it is likely that our sample excluded people with limited technological experience. Finally, our sample likely excluded those with disabilities, particularly visual impairment. Future work should investigate the accessibility of data downloads to those with disabilities.

As with any qualitative study, a participant not making or responding to a statement does not mean they disagree with it. While, for context, we provide counts of participants who expressed specific sentiments, we do not intend them to indicate overall prevalence. Additionally, the scavenger hunt activity may have primed participants. Though we tried to offset this concern by starting with a free exploration, it is possible that some participants may have been led to believe some sections of their data were most important based on our scavenger hunt items. As a result, we make no claims about the generalizability of our study. Rather, we present initial findings and directions for the design of data downloads.

5 Results

We first summarize participant demographics. We then report key findings from our focus groups in four key areas: reactions to existing content; ideas for improving content; reactions to existing formats; and ideas for improving formats.

5.1 Participants

We recruited 272 participants for Part 1, 77 of whom completed Part 2 and 42 of whom completed Part 3. Among Part 1 participants who did not continue, 156 never responded that they were ready for Part 2, while 39 were deemed ineligible.

Among Part 3 participants, 25 identified as male, 16 as female, and one as non-binary. Our sample skewed young: eight were 18–24, 19 were 25–34, nine were 35–44, five were 45–54, and one was 55 or older. Participants reported their highest level of educational attainment: two completed high school, nine completed some college or an associate's degree, 18 had a bachelor's degree, and 13 had a graduate or professional degree. Twenty-eight participants self-reported as White, seven as Asian or Pacific Islander, six as Black or African American, and one marked "other." Five indicated they were of Hispanic or Latino origin. Six had an education or were employed in computer science or IT. Six were students.

5.2 Content of Existing Data Downloads

We first report on participants' reactions to the content of their data downloads. Participants were struck by the heavy amount of detail, sometimes reaching the level of creepy, and found the inclusion of certain content surprising. They also identified several practical uses for their data downloads.

Expectations: Before each free exploration, we asked if participants had looked at their downloads before the session; only six of the 42 had. We then asked participants about their expectations of the content of their data downloads. **Most commonly, participants expected data downloads to contain demographics and data generated via interaction with the site** (e.g., friends and messages for Facebook, playlists and watch history for YouTube). Seven participants (all in Amazon, Facebook, or YouTube focus groups) also expected to see inferences or data from third parties.

During the free exploration and scavenger hunt, we encouraged participants to comment aloud. Nine were surprised by the presence or absence of information. For instance, F-B-3 was surprised to see facial recognition data, and S-A-1 was surprised to see her full address associated with a music company. F-B-2 and S-B-1, in contrast, expected to see more ad interests and search queries, respectively. For 12 participants (all companies), at least some of their expectations of what would be contained in their data downloads matched reality.

Twenty participants (all companies but YouTube) commented on the **accuracy or inaccuracy of their data**. Eleven participants mentioned that at least one part of their data was accurate, and 13 participants mentioned that at least one part was inaccurate. Seventeen participants (all companies) **mentioned that there was information missing from their data downloads**, either time gaps or information omitted altogether. Y-B-1 said, *"I personally think there's information that they have that's not in these files. And it can be used depending on what they need."* F-B-1 was surprised to find data he was *"100% sure"* he had erased. F-B-2 also found data he believed he had deleted, which he attributed to a *"legacy issue."* In contrast, Y-A-4 recalled deleting specific searches. He was surprised to find they did not appear in his data download. Regardless of whether these participants are correct about their deleted data, these comments suggest a lack of clarity and perhaps distrust related to how data is stored and retained. Y-B-1 commented, *"I do think there is information that might not be in these files, but somehow when we signed up for these platforms, in the very small fine text, they're letting us know it's there and you may not know what the term is or exactly what it means . . . but I do think there's other information that they capture that we may not be aware of."*

Reactions to Content: Fourteen participants (all companies) either commented on how far their data went back in time or reacted to old content. F-A-2 said, *"It's kind of weird to like be pulled back into that space of when you set up the first ever Facebook account."* Eight participants (all companies but Facebook) noted the **level of detail** in the files. Five were surprised by how detailed the files were; one partially blamed difficulty navigating the data on the level of detail. Eight participants (Facebook, Google, YouTube) reported feeling **creeped out or scared** about the breadth, detail, and type of data being collected and stored about them. Feelings of

unease were not always related to a lack of awareness. G-A-4 noted, *"For me, nothing was surprising. Like, I knew Google is recording everything. It's just that seeing this in front of me and all the data that has been collected over all the years, it's like a rude realization that yeah, there is someone watching you all the time."* G-B-2 also expressed this sentiment. This quotation highlights the potential for data downloads to be used for promoting privacy-protective behaviors. While many users are aware of tracking and data collection in general, a data download situates these practices in a personal context. This personal context is perhaps more likely to inspire action than simply hearing about data collection in the abstract.

Five participants (Amazon, YouTube, Uber) commented on the large size of their data download. G-A-3 noted the presence of many product options on Google's data download page. G-B-3 made a similar point: *"There was also so many other things you could download, that also really scared me. I was like, this is only my location and search history. I can't imagine if everything else was included."* A-B-2 felt the lack of definitions for terms made navigation and comprehension harder. Asked about navigating data downloads, F-B-1 said, *"It was like reading a book about myself but not written by myself."* This quotation is perhaps emblematic of a lack of control by users over their data. We note that the right to be forgotten, the right to participation, and the right to rectification can help users reclaim control over their data.

Uses and Misuses: Twenty-nine participants spanning all companies discussed possible uses or misuses of data downloads. Eleven (all but Facebook) identified **practical reasons why they might want access to their own data downloads**, including accessing a lost record, budgeting, or finding and erasing problematic information. Eight participants (Amazon, Facebook, Spotify, Uber) imagined these files could be used for **privacy purposes**, namely keeping track of the collection of their personal information. U-A-1 observed, *"It's interesting to understand how exposed you are from a privacy perspective."* Four participants (Facebook and Spotify) went beyond awareness, suggesting privacy-protective actions they or others might take after viewing their data downloads. F-B-1 mentioned refraining from making sensitive searches on Facebook, and S-A-2 and S-A-3 considered using information from a data download to help them keep their accounts secure. S-B-5 said, *"I think if the company had a data breach, and I knew that I was in that data breach, being able to see what data was potentially accessed for myself is important . . . if it has some kind of impact on my credit or I need to freeze my credit."* These privacy and security concerns arose organically from looking at the data, without prompting from researchers.

Mentioned misuses included account compromise or inappropriate targeted ads (six participants). Four mentioned **data downloads being used by law enforcement, the government, or in court**; three others agreed with or commented on these statements. F-A-1 said, *"I'm curious to know if this information can be subpoenaed in a court, because there's a lot of information here. So I mean if there's any illegal activity going on you could definitely use this file to find out."* Concerns about misuse of downloads themselves are not entirely misplaced [15,45]. We note that while law enforcement could likely obtain information directly from companies regardless of the data download feature, the data download did raise awareness of how much information companies store.

5.3 Desired Content

To help focus the efforts of programmers and designers who might craft interfaces for data downloads, we examined the types of data most and least interesting to participants. Participants discussed demographics, data generated via interaction with the company, inferences, and aggregate information.

Demographics and Site Data: Fourteen participants (all companies but Spotify) were interested in data associated with their demographics and direct site interaction, as opposed to inferences. Participants mentioned search history (three, Amazon and Facebook), location data (two, Facebook and Google), and photos (two, Facebook). Y-B-3 wanted to see *"how much personal information they've collected."* Y-B-1 agreed. Six Spotify participants also cited personally identifiable information and payment details as among the most important things to see. Four participants (Amazon, Uber, Facebook) wanted to know how their information was being used and shared, and three participants (Google, Uber, YouTube) wanted to see everything the company had about them.

Inferences and Advertising: Ten participants (all but Uber) wanted to see inferences made about them or obtain insight into inferencing algorithms. F-A-2 said, *"What's interesting to me is how my online behavior is affecting how this company and all the affiliates see me. And in what category, say, they put me or don't put me. . . . That has a way broader implication than the actual things that I am looking for. . . . Who is programming these algorithms? . . . Do they represent a broader part of society or are they all from a very similar group, similar life experiences and backgrounds?"* Four participants (Amazon and YouTube) wanted insight into the company's recommendation algorithm and/or the data that powers it.

Seven participants (all but Spotify) wanted to see **advertising data**. Two (Facebook, YouTube) wanted data on advertising-related inferences. Furthermore, two (Google, Uber) mentioned things they said aloud being used for advertising, referencing a common folk belief that devices secretly listen to users [22]. U-B-2 said, *"There's nothing weirder than having talked to someone on the phone . . . and an ad pops up for something that you were talking to somebody on the phone about."*

Nine participants discussed things they did not want to see. Four (Amazon and YouTube) reported no interest in any of the information in their data downloads. Four others (Facebook, Spotify) named specific data types they found useless or irrelevant, including poke history, past aliases, or search queries. Three participants (Amazon, Facebook) wanted **less data retention, for privacy and security** purposes. For instance, F-B-3 did not want long-term location data saved because she feared that a malicious actor could use it to find her.

Aggregation and Synthesis: Participants wanted more than just raw data.[3] Nineteen participants (all but Facebook; seven unprompted) wanted to see aggregate data about their site usage or activity. Eight (Amazon, Uber, YouTube; two unprompted) mentioned wanting aggregate financial data for business, budgeting, or to examine spending habits. U-B-5, who drives for Uber, imagined using aggregate data to determine the most profitable times for her to drive. Six participants (Google, Spotify, YouTube) wanted a breakdown of how much time (relative or absolute) they spent listening to songs or artists, watching videos, or otherwise engaging.

During Data Visualization 101, we asked participants to look for examples synthesizing multiple data types. Participants quickly adopted this theme: 30 (all companies) used some kind of synthesis in their sketches. In addition, three participants (Amazon, Facebook, Spotify) brought up data synthesis, unprompted, earlier in the discussion. Most imagined using synthesis to learn about themselves and their use of the site. For example, S-B-5 proposed a graph of the correlation between music choices and the time of day and year.

5.4 Format of Existing Downloads

Participants identified benefits and drawbacks to the format and organization of the data downloads they examined.

Quantity of Data: Nine participants agreed that **accessing records was difficult due to the vastness of their data downloads**. Four (Facebook, Google, YouTube) described it as *"overwhelming,"* and five (Amazon, Facebook, Google, YouTube) described the challenge of moving through their data as *"tough," "tedious," "hard,"* or *"time-consuming."*

Navigation: Twenty-one participants (all but Amazon) felt that, **overall, their data download was easy to explore**. They attributed this to intuitive organization, nicely formatted files, and descriptive folder and file names. S-B-4 said, *"I think the names and the descriptions of the files was exactly what I expected them to be once you clicked on them."*

[3] We use "aggregate" for the collection of multiple instances of a single type of data (e.g., to summarize or identify trends). We use "synthesis" for combining multiple types of data to obtain insights unavailable in isolation.

Conversely, 13 participants (all companies) expressed **difficulty navigating through their data downloads**. Nine of these (all but Spotify) attributed their difficulties to a **lack of familiarity** with data downloads in general, and with file types such as JSON specifically. Three of these nine said that, despite initial difficulties, they expected they could learn to navigate the files over time. U-B-2 described *"a very small learning curve where you had to figure out how the information was set up. . . . Once you figure that out, it's pretty easy."* Six participants felt they needed more than a single read-through to understand their files. A-B-4 said, *"Everything was the same font and the same size, so there's nothing bolded that will jump out at you."* Y-A-5 wondered about deliberate obfuscation: *"Most of the interesting data is stored in these files, that as a non-specialist, I can't read. . . . We're effectively illiterate when it comes to reading this additional data that they've been collecting."*

Organization: Thirteen participants spanning all companies felt the **files were disorganized** or could be more usefully organized. Eight participants (Amazon, Facebook, Spotify, YouTube) attributed their difficulty finding information to data downloads' disorganization. Y-A-1 said, *"The top-level organization makes a lot of sense, but then when you try and go one layer deeper then it just turns into raw data."* Y-A-3 remarked, *"It's like they didn't even try [to organize the data]. They just kind of dumped it on you."* Five participants (Amazon, Spotify, Uber) felt that **related information was inconveniently spread across multiple files**. A-A-4 said, *"I'd prefer it if it was just a single file,"* and A-A-1 agreed.

In contrast, nine participants (all but Google) were satisfied with the organization of their files. U-B-4 said, *"It looked exactly the way I would organize it."*Two commented on the **usefulness of folder names and how files were ordered**. S-B-5 noted that *"the 'follow list' was alphabetized, and then you could kind of see other stuff was by most recent."*

File Formats: Twelve participants (Google, Spotify, Uber, YouTube) discussed **difficulties with JSON files** and how they might deter others. G-B-3 said, *"A JSON file to begin with is pretty inaccessible."* S-A-1 said, *"I was kind of surprised that it . . . comes down in a JSON file, which I think could feel really intimidating."* Five participants (Google, Spotify, Uber, YouTube) **weren't able to open the JSON files** on their computers. On the other hand, S-B-5 found the JSON files *"user-friendly to see, especially with the way they color coded it."* Note that the color-coding was a feature of the JSON viewer we provided. G-A-4, the tech expert of his group, pointed out that JSON would enable analysis scripts.

Three participants (Facebook and Google) felt **HTML files were usable and useful**, but not everyone agreed. Y-A-2 pointed out that an HTML file *"has a nice user interface and I can scroll through it all, but it's still not useful because*

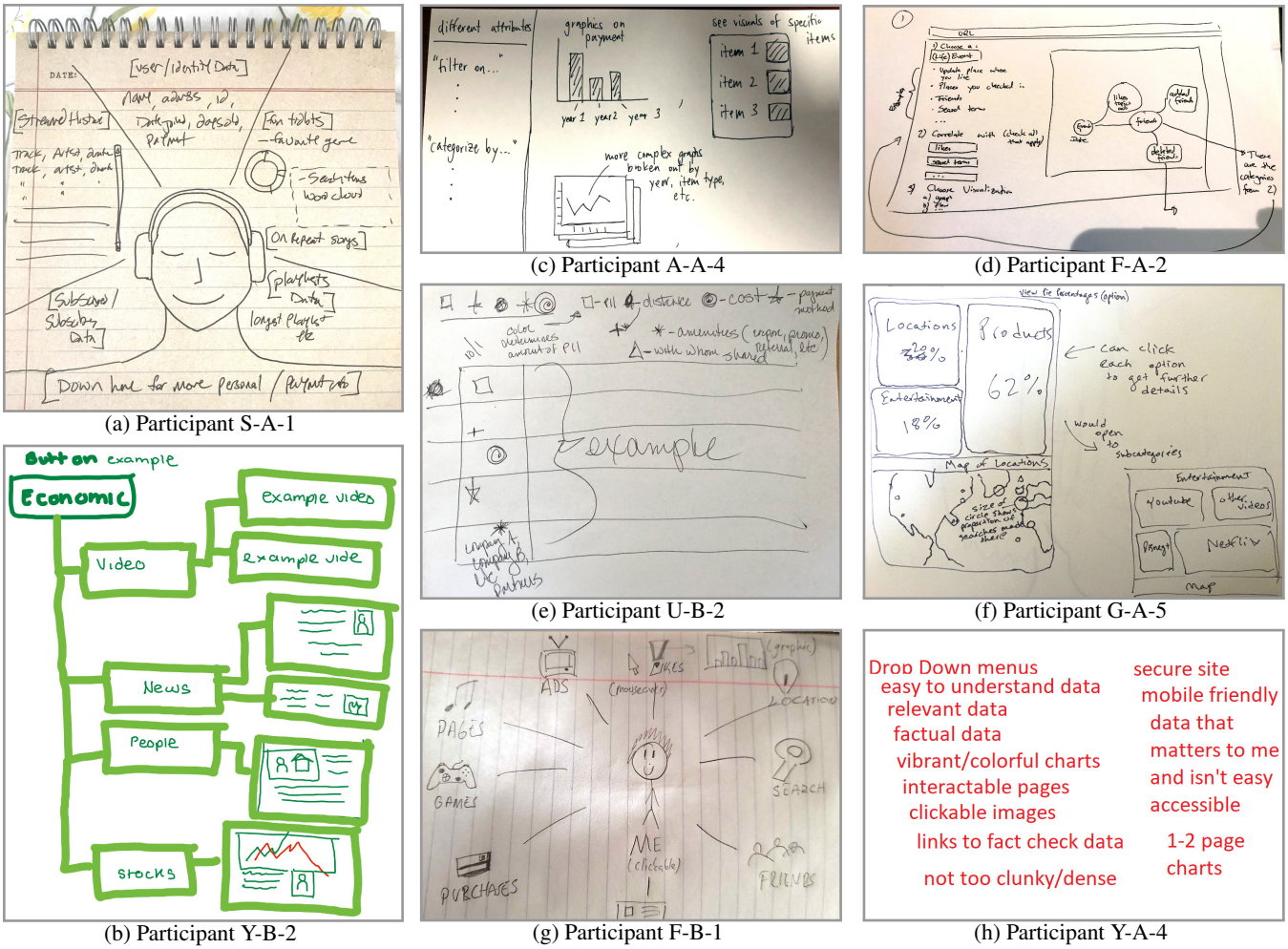


Figure 2: Excerpts from participants' sketches during the design activity.

it's a long list to scroll through. If you spent more than a couple days on YouTube you can incur a very long list, and I've actually had that file crash multiple times." A-B-3 and U-A-1 found the CSV format of their data downloads straightforward, though A-B-3 added that she worked with spreadsheets daily. In contrast, A-A-4 did not find CSV files convenient.

5.5 Desired Format

Finally, we report on participants' ideas for formatting, including meaningful organization, filtration, visual representation, and interactivity. As in any co-design exercise, participants' suggestions should be seen as inspirations for design professionals to build on, rather than direct specifications.

File Formats and Interfaces: Participants suggested various improvements to the file formats. Y-A-2 liked that HTML files could be opened easily in the browser, but also wished for a better user interface. Several participants expressed interest in **CSV or other spreadsheet-compatible formats**. G-B-3, who had prior experience downloading CSV Twitter data outside this study, said the Google data would have been easier to digest had it been formatted similarly. G-B-2, Y-A-5, and A-A-1 also discussed the merits of spreadsheet formats.

G-B-2 considered printing out his data download, citing research that people comprehend information better when it's printed on paper. G-B-1 agreed, adding that an older generation might feel more comfortable with paper.

Finding Important Information: Twenty-one participants (all groups) expressed a desire for a **high-level overview of their data, with the option of delving for more information.** Participants wanted to see either an overview of everything contained in the download or a summary of the most important information. S-A-1 and F-B-1 used this approach in their sketches (Figures 2a and 2g). Y-A-5 said, "*The idea is to give them as much information as possible as an option, but not to overwhelm them with this sort of first glance, first blush dashboard.*" This suggestion aligns with existing best practices in data visualization [29, 73].

Seventeen participants (all companies) emphasized **the importance of filtration**. F-A-1 commented, *"It would've been helpful to have filters so that you can organize the information by date or time, because if I'm looking for something specific, scrolling through that page of long history would be tedious."* F-A-2 agreed, and five other participants from Amazon and Google sessions made similar comments. Ten participants (all companies but Facebook) included filtration features in their sketches, including three who had already commented on filtration earlier in the session. Figure 2c shows A-A-4's sketch, which included a filtration feature.

Relatedly, 12 participants (all companies) included options to **sort by date** in their sketches. A-B-3 noted, *"I would only ever look at my stuff chronologically."* Most data downloads already organize relevant data by date. However, ten participants imagined extending this to filter by day, week, month, or year. F-A-2 imagined a different kind of sorting: an event-centered visualization in which the user selected a life event like *"[got] married, . . . moved to a new place, or got a new job."* Data would be displayed in relation to that event (Figure 2d). A-B-2 sketched another option: separating human-determined and computer-determined information.

In addition to sorting and filtering, several participants mentioned **prioritization**. Three Amazon participants mentioned reorganizing files so that meaningful information (e.g., item descriptions) appeared before less semantically useful data (e.g., order ID numbers). Three others (Facebook, YouTube) wanted to prioritize data types with the most entries.

Visualization: Nineteen participants (all companies) used **line graphs, bar graphs, pie charts, or tree graphs** in their sketches. Participants plotted information like payments vs. time, ads clicked on vs. ignored, and breakdowns of online activity by categories (e.g., entertainment, news, and people). Y-B-2's sketch (Figure 2b) illustrates the latter. This reflects participants' strong interest in synthesis, discussed above.

Nine participants' sketches (Amazon, Facebook, Google, Uber) used a **map to show location data**. Eight of these combined their map with other data, such as friends, search history, or frequency of the location. G-A-3 and S-A-3 used timelines in their sketches, consistent with the desire for chronological organization. Y-A-5 mentioned word clouds, and S-A-1 included a word cloud for search history in her sketch.

Fourteen participants (all companies) emphasized **interactivity** and/or included it when describing their sketches. Y-A-4, for example, wanted *"something that you'd be able to click on, whether you enlarge it or you control it with the mouse wheel . . . so you can see it a little more clearly or if you're looking for something specifically"* (Figure 2h). Others suggested buttons, menus, hovering, and clickable elements.

Several participants were also attracted to **simplicity** in visualization. Y-A-2 wanted to avoid *"very graphically interesting stuff that represents the data horribly. Because they don't do, like, bar graphs. They'll do, look at this zigzag graph, and you have no idea what that graph's trying to tell you or show you, because it doesn't actually tell you anything except for look pretty."* Five participants (Amazon and YouTube) emphasized simplicity in their sketch explanations.

Participants (19, all companies) also argued for using **color and element size** to distinguish data. Thirteen (all companies but Amazon) used color or size distinctions in their sketches. Six (Google, Spotify, Uber) used size and color to represent frequency, such as most-listened-to artists and most-visited locations. Figure 2f shows one example. These requests, which track good visualization practice, contrast heavily with the plain-text files participants viewed during the sessions.

Security, Privacy, and Accuracy: Four participants identified format-related security and privacy considerations. F-B-1 wished his data download had been **password-protected** so that someone with access to his computer could not read it, though presumably most information could also be accessed by visiting Facebook directly while logged in. Y-A-4 said he expected to access data via HTTPS and wished for some form of data encryption. The same participant also asked about fact-checking, or some other mechanism for verifying the provided data was accurate. G-B-2 and A-B-2 proposed in-band deletion of data directly after viewing it. This accords with the longstanding interaction principle of direct manipulation [72].

6 Discussion

We detail our design recommendations, followed by a brief discussion of the policy implications of our results.

6.1 Design Recommendations

Our co-design study provides insight into how to reimagine data downloads to be usable and useful. Our participants identified a variety of goals and use cases for data downloads. Some wanted easy access to artifacts, memories, or original content. Others wanted to know what personal information was collected and stored about them, or wanted insight into the inferences being made about them. A few were curious about aggregate statistics.

Participants also identified significant shortcomings of current data downloads that might hinder these goals: poor organization, sometimes-unfriendly file formats, and too many details with no way to filter. While these obstacles could perhaps be overcome with sufficient patience, they may deter users outside a research study. Furthermore, most data downloads do not include meaningful aggregation or synthesis, though it is highly likely such analysis is conducted internally to power recommendations and personalized advertising. The current state of data downloads thus prevents users from fully reaping the benefits that the right of access might provide.

It is therefore important to re-imagine data downloads for use by humans, separately from data designed for machine interpretation. Drawing from participants' responses, we make the following recommendations for data visualization tools:

- **Meaningful Organization:** Organize data chronologically, but with options for aggregating (e.g., by month). Group related data together. In line with visualization best practice, offer both a high-level overview and details on demand [73].
- **Filtration:** Enable filtering by type of data as well as other properties (e.g., payments over a certain amount).
- **Aggregation and Inferencing:** Provide insight into data aggregations and inferences made with the user's information, as well as the mechanisms behind them.
- **Interactivity and Exploration:** Enable rich interactions, such as selecting elements of high interest and zooming or hovering to reveal more information. Provide functions similar to simple spreadsheet or scripting tools to support synthesis. Current static formats (e.g., JSON, HTML) work with all common platforms, and interactive views should too. Web-browser-based interactivity may be appropriate.
- **Direct Manipulation:** Historically, rights of access have been associated with *participation*, the ability to contest or correct information [24]. Enact this principle by allowing correction via direct manipulation in the data download interface. Further, streamline the right to erasure (also defined in GDPR) via direct requests to delete specific information.

Efforts to improve data downloads could be made by the companies themselves, or by third parties. Companies know the most about their data, including its origin and schema, and are thus in the best position to provide explanations. Companies are also in the best position to enable direct manipulation for deleting or contesting data. However, companies may not have strong incentives to improve data downloads. From a legal compliance standpoint, data downloads are arguably sufficient currently. Additionally, some companies may wish to keep data downloads abstruse to hide unsavory data practices.

Thus, data downloads present an opportunity for third-party privacy and transparency advocates to design tools for user empowerment, continuing the mission of prior TETs and PETs. We found that viewing data downloads, even in their current not-very-usable state, raised organic privacy and security concerns. Third-party tool designers should consider how to make the content salient and digestible, while still leveraging the "creepy factor" [78, 88] to help users make better privacy choices. Additionally, third-party tools could offer cross-platform analysis and data-driven recommendations that promote privacy and support users in exercising control. Such tools could also serve as a GDPR/CCPA hub, allowing users to make data download and deletion requests with a click of a button. Further, third-party tools could (with proper consent and pseudonymization) aggregate data across users in order to characterize the data-collection ecosystem.

6.2 Policy Implications

The creation of data visualization tools should support, but not replace, legal intervention. While our sessions were designed to elicit ideas for data visualization tools, our findings also have implications for future iterations of data access laws:

- **Access vs. Portability:** We suggest that tensions between the right of access and right to data portability hinder the efficacy of the former. Future laws should better differentiate these requirements and include comprehensibility standards for data access rights. Specifically, data downloads should include a README-type file with an overview of the structure and content of the files, plus explanations for any technical or otherwise unintuitive fields.
- **Required Content:** We found that users were curious about their data, especially about inferences and aggregate data. It is at present ambiguous as to what data must be included in a data download. For instance, companies may argue that inference data embeds trade secrets and thus exclude inferences from data downloads. Notably, the Spotify and YouTube files did not include data about how recommended songs and videos were determined, which is related to the topic of algorithmic transparency. Policymakers should weigh companies' interests against users' right of access when deciding what data is within scope of a data subject access request. The law should clarify the required content.
- **Explanations for Missing Data:** Some participants felt data was missing from their files (e.g., gaps in time or the omission of certain categories). Data downloads should flag when and why data is missing.

The readability of data downloads is important not only for users, but also for technologists who will rely on README files with clear explanations to create data visualization tools. Thus, technology and law are both responsible for improving the transparency of the online data ecosystem.

7 Conclusion

We presented results from 12 focus groups with a total of 42 participants. We solicited participants' reactions to their own data downloads from one of six companies: Amazon, Facebook, Google, Spotify, Uber, or YouTube. Participants completed activities that familiarized them with their data downloads, elicited their opinions about content and format, and sparked inspiration for drawing their ideal data download visualization. Participants identified several key weaknesses in current data downloads, including that they were disorganized, unintuitive to navigate, and lacked usability features like filtration. These criticisms illuminate the need for companies themselves, or interested third parties, to reimagine data downloads to be usable and useful for humans, rather than simply machine-readable. This would better support the *right of data access*, as distinct from *data portability*. To this end, we presented associated design recommendations.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. CNS-2047827. We acknowledge funding from a UMIACS contract under the partnership between the University of Maryland and DoD. We thank Emma Veys, Joe Veys, Christopher John Boyle, Purrsephone, and Furdinand for their assistance. We also thank Lior Strahilevitz and the attendees of the 2021 Privacy Law Scholars Conference for their feedback and comments.

References

[1] Alessandro Acquisti, Curtis Taylor, and Liad Wagman. The Economics of Privacy. *Journal of Economic Literature*, 54(2):442–492, 2016.

[2] Alexa. The Top 500 Sites on the Web. https://www.alexa.com/topsites/category, 2020.

[3] Fatemeh Alizadeh, Timo Jakobi, Alexander Boden, Gunnar Stevens, and Jens Boldt. GDPR Reality Check – Claiming and Investigating Personally Identifiable Data from Companies. In *Proc. EuroUSEC*, 2020.

[4] David Alpert. Beyond Request-and-Response: Why Data Access will be Insufficient to Tame Big Tech. *Columbia Law Review*, 120:1215–1254, 2020.

[5] Athanasios Andreou, Márcio Silva, Fabrício Benevenuto, Oana Goga, Patrick Loiseau, and Alan Mislove. Measuring the Facebook Advertising Ecosystem. In *Proc. NDSS*, 2019.

[6] Athanasios Andreou, Giridhari Venkatadri, Oana Goga, Krishna Gummadi, Patrick Loiseau, and Alan Mislove. Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations. In *Proc. NDSS*, 2018.

[7] Julio Angulo, Simone Fischer-Hübner, Tobias Pulls, and Erik Wästlund. Usable Transparency with the Data Track: A Tool for Visualizing Data Disclosures. In *Proc. CHI*, 2015.

[8] Rebecca Balebako, Jaeyeon Jung, Wei Lu, Lorrie Faith Cranor, and Carolyn Nguyen. "Little Brothers Watching You:" Raising Awareness of Data Leaks on Smartphones. In *Proc. SOUPS*, 2013.

[9] Theo Bertram, Elie Bursztein, Stephanie Caro, Hubert Chao, Rutledge Chin Feman, Peter Fleischer, Albin Gustafsson, Jess Hemerly, Chris Hibbert, Luca Invernizzi, et al. Five Years of the Right to be Forgotten. In *Proc. CCS*, 2019.

[10] Hugh Beyer and Karen Holtzblatt. *Contextual Design: Defining Customer-Centered Systems*. Morgan Kaufmann, 1998.

[11] Asia J. Biega, Peter Potash, Hal Daumé III, Fernando Diaz, and Michèle Finck. Operationalizing the Legal Principle of Data Minimization for Personalization. In *Proc. SIGIR*, 2020.

[12] Christoph Bier, Kay Kühne, and Jürgen Beyerer. PrivacyInsight: The Next Generation Privacy Dashboard. In *Proc. APF*, 2016.

[13] Debmalya Biswas, Imad Aad, and Gian Paolo Perrucci. Privacy Panel: Usable and Quantifiable Mobile Privacy. In *Proc. ARES*, 2013.

[14] Coline Boniface, Imane Fouad, Nataliia Bielova, Cédric Lauradoux, and Cristiana Santos. Security Analysis of Subject Access Request Procedures How to Authenticate Data Subjects Safely When They Request for Their Data. In *Proc. APF*, 2019.

[15] Luca Bufalieri, Massimo La Morgia, Alessandro Mei, and Julinda Stefa. GDPR: When the Right to Access Personal Data Becomes a Threat. In *Proc. ICWS*, 2020.

[16] Johana Cabinakova, Christian Zimmermann, and Guenter Mueller. An Empirical Analysis of Privacy Dashboard Acceptance: The Google Case. In *Proc. ECIS*, 2016.

[17] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. In *Proc. PETS*, 2015.

[18] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. We Value Your Privacy ...Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *Proc. NDSS*, 2019.

[19] Claire Dolin, Ben Weinshel, Shawn Shan, Chang Min Hahn, Euirim Choi, Michelle L. Mazurek, and Blase Ur. Unpacking Perceptions of Data-driven Inferences Underlying Online Targeting and Personalization. In *Proc. CHI*, 2018.

[20] Serge Egelman, Adrienne Porter Felt, and David Wagner. Choice Architecture and Smartphone Privacy: There's A Price for That. In *Proc. WEIS*, 2012.

[21] Motahhare Eslami, Sneha R. Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. Communicating Algorithmic Process in Online Behavioral Advertising. In *Proc. CHI*, 2018.

[22] Bree Fowler. Is Your Smartphone Secretly Listening to You? Consumer Reports, July 10, 2019. https://www.consumerreports.org/smartphones/is-your-smartphone-secretly-listening-to-you/.

[23] Fatih Gedikil, Dietmar Jannach, and Mouzhi Ge. How Should I Explain? A Comparison of Different Explanation Types for Recommender Systems. *International Journal of Human-Computer Studies*, 72(4):367–382, 2014.

[24] Robert Gellman. Fair Information Practices: A Basic History. *SSRN 2415020*, 2019. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2415020.

[25] Peter Leo Gorski, Yasemin Acar, Luigi Lo Iacono, and Sascha Fahl. Listen to Developers! A Participatory Design Study on Security Warnings for Cryptographic APIs. In *Proc. CHI*, 2020.

[26] Casandra Grundstorm, Karin Väyrynen, Netta Iivari, and Minna Isomursu. Making Sense of the General Data Protection Regulation—Four Categories of Personal Data Access Challenges. In *Proc. HICSS*, 2019.

[27] Julia Hanson, Miranda Wei, Sophie Veys, Matthew Kugler, Lior Strahilevitz, and Blase Ur. Taking Data Out of Context to Hyper-Personalize Ads: Crowdworkers' Privacy Perceptions and Decisions to Disclose Private Information. In *Proc. CHI*, 2020.

[28] Woodrow Hartzog. The Case Against Idealising Control. *European Data Protection Law Review*, 4:423, 2018.

[29] Jeffrey Heer and Ben Shneiderman. Interactive Dynamics for Visual Analysis. *Communications of the ACM*, 55(4):45–54, April 2012.

[30] Paul De Hert, Vagelis Papakonstantinou, Gianclaudio Malgieri, Laurent Beslay, and Ignacio Sanchez. The Right to Data Portability in the GDPR: Towards User-Centric Interoperability of Digital Services. *Computer Law and Security Review*, 2018.

[31] Kashmir Hill. I Got Access to My Secret Consumer Score. Now You Can Get Yours, Too. *The New York Times*, 2019. https://www.nytimes.com/2019/11/04/business/secret-consumer-score-access.html.

[32] Nicholas F. Palmieri III. Who Should Regulate Data?: An Analysis of the California Consumer Privacy Act and Its Effects on Nationwide Data Protection Laws. *Hastings Science and Technology Law Journal*, 2020.

[33] Information Is Beautiful, 2021. https://informationisbeautiful.net.

[34] Jim Isaak and Mina J. Hanna. User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *IEEE Computer*, 51(8):56–59, 2018.

[35] Ruogu Kang, Stephanie Brown, Laura Dabbish, and Sara Kiesler. Privacy Attitudes of Mechanical Turk Workers and the U.S. Public. In *Proc. USENIX Security*, 2014.

[36] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. "My Data Just Goes Everywhere:" User Mental Models of the Internet and Implications for Privacy and Security. In *Proc. SOUPS*, 2015.

[37] Farzaneh Karegar, Tobias Pulls, and Simone Fischer-Hübner. Visualizing Exports of Personal Data by Exercising the Right of Data Portability in the Data Track - Are People Ready for This? *IFIP International Summer School on Privacy and Identity Management*, pages 164–181, 2016.

[38] Patrick Gage Kelley, Paul Hankes Drielsma, Norman M. Sadeh, and Lorrie Cranor. User-Controllable Learning of Security and Privacy Policies. In *Proc. AISec*, 2008.

[39] Jan Kolter, Michael Netter, and Günther Pernul. Visualizing Past Personal Data Disclosures. In *Proc. ARES*, 2010.

[40] Jacob Leon Kröger, Jens Lindermann, and Dominik Herrmann. How do App Vendors Respond to Subject Access Requests? A Longitudinal Privacy Study on iOS and Android Apps. In *Proc. ARES*, 2020.

[41] Clément Labadie and Christine Legner. Understanding Data Protection Regulations from a Data Management Perspective: A Capability-Based Approach to EU-GDPR. In *Proc. Wirtschaftsinformatik*, 2019.

[42] Mathias Lécuyer, Guillaume Ducoffe, Francis Lan, Andrei Papancea, Theofilos Petsios, Riley Spahn, Augustin Chaintreau, and Roxana Geambasu. Xray: Enhancing the Web's Transparency with Differential Correlation. In *Proc. USENIX Security*, 2014.

[43] Candice Louw. Modeling Personally Identifiable Information Leakage that Occurs through the Use of Online Social Networks. Master's thesis, University of Johannesburg, 2015.

[44] René L. P. Mahieu and Jef Ausloos. Harnessing the Collective Potential of GDPR Access Rights: Towards an Ecology of Transparency. *Internet Policy Review*, July 2020.

[45] Mariano Di Martino, Pieter Robyns, Winnie Weyts, Peter Quax, Wim Lamotte, and Ken Andries. Personal Information Leakage by Abusing the GDPR "Right of Access". In *Proc. SOUPS*, 2019.

[46] Aleecia M. McDonald and Lorrie Faith Cranor. Beliefs and Behaviors: Internet Users' Understanding of Behavioral Advertising. In *Proc. TPRC*, 2010.

[47] Jeremy B. Merrill and Ariana Tobin. Facebook Moves to Block Ad Transparency Tools — Including Ours. ProPublica, January 28, 2019. https://www.propublica.org/article/facebook-blocks-ad-transparency-tools.

[48] Moz. The Moz Top 500 Websites, 2021. https://moz.com/top500.

[49] Min Y. Mun, Donnie H. Kim, Katie Shilton, Deborah L. Estrin, Mark H. Hansen, and Ramesh Govindan. PDVLoc: A Personal Data Vault for Controlled Location Data Sharing. *ACM Transactions on Sensor Networks*, 10(4), 2014.

[50] Cosmin Munteanu, Calvin Tennakoon, Jillian Garner, Alex Goel, Mabel Ho, Clare Shen, and Richard Windeyer. Improving Older Adults' Online Security: An Exercise in Participatory Design. In *Proc. SOUPS*, 2015.

[51] Patrick Murmann and Simone Fischer-Hübner. Tools for Achieving Usable Ex Post Transparency: A Survey. *IEEE Access*, 5, 2017.

[52] Helen Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2009.

[53] Ehimare Okoyomon, Nikita Samarin, Primal Wijesekera, Amit Elazari Bar On, Narseo Vallina-Rodriguez, Irwin Reyes, Álvaro Feal, and Serge Egelman. On the Ridiculousness of Notice and Consent: Contradictions in App Privacy Policies. In *Proc. ConPro*, 2019.

[54] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research. *Journal of Experimental Social Psychology*, 2017.

[55] Marta Piekarska, Yun Zhou, Dominik Strohmeier, and Alexander Raake. Because We Care: Privacy Dashboard on FirefoxOS. ArXiv, 2015.

[56] Marco Pistoia, Omer Tripp, Paolina Centonze, and Joseph W. Ligman. Labyrinth: Visually Configurable Data-Leakage Detection in Mobile Applications. In *Proc. MDM*, 2015.

[57] Eugenia Politou, Efthimios Alepis, and Constantinos Patsakis. Forgetting Personal Data and Revoking Consent Under the GDPR: Challenges and Proposed Solutions. *Journal of Cybersecurity*, 2018.

[58] Jon Porter. GDPR Makes It Easier to Get Your Data, but That Doesn't Mean You'll Understand It. The Verge, January 27, 2019. https://www.theverge.com/2019/1/27/18195630/gdpr-right-of-access-data-download-facebook-google-amazon-apple.

[59] Emilee Rader. Awareness of Behavioral Tracking and Information Privacy Concern in Facebook and Google. In *Proc. SOUPS*, 2014.

[60] Emilee Rader, Samantha Hautea, and Anjali Munasinghe. "I Have a Narrow Thought Process": Constraints on Explanations Connecting Inferences and Self-Perceptions. In *Proc. SOUPS*, 2020.

[61] Emilee Rader and Janine Slaker. The Importance of Visibility for Folk Theories of Sensor Data. In *Proc. SOUPS*, 2017.

[62] Philip Raschke, Axel Kupper, Olha Drozd, and Sabrina Kirrane. Designing a GDPR-compliant and Usable Privacy Dashboard. *IFIP International Summer School on Privacy and Identity Management*, 2017.

[63] Elissa M. Redmiles, Everest Liu, and Michelle L. Mazurek. You Want Me To Do What? A Design Study of Two-Factor Authentication Messages. In *Proc. SOUPS*, 2017.

[64] Robert W. Reeder, Patrick Gage Kelley, Aleecia M. McDonald, and Lorrie Faith Cranor. A User Study of the Expandable Grid Applied to P3P Privacy Policy Visualization. In *Proc. WPES*, 2008.

[65] General Data Protection Regulation. Guidelines on Transparency under Regulation 2016/679. 2018. https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=622227.

[66] Joel R Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T Graves, Fei Liu, Aleecia McDonald, Thomas B Norton, and Rohan Ramanath. Disagreeable Privacy policies: Mismatches between Meaning and Users' Understanding. *Berkeley Tech. LJ*, 30:39, 2015.

[67] Christopher J. Riederer, Daniel Echickson, Stephanie Huang, and Augustin Chaintreau. FindYou: A Personal Location Privacy Auditing Tool. In *Proc. WWW*, 2016.

[68] Florian Schaub, Aditya Marella, Pranshu Kalvani, Blase Ur, Chao Pan, Emily Forney, and Lorrie Faith Cranor. Watching Them Watching Me: Browser Extensions' Impact on User Privacy Awareness and Concern. In *Proc. USEC*, 2016.

[69] Roman Schlegel, Apu Kapadia, and Adam J. Lee. Eyeing Your Exposure: Quantifying and Controlling Information Sharing for Improved Privacy. In *Proc. SOUPS*, 2011.

[70] Carina Paine Schofield, Ulf-Dietrich Reips, Stefan Stieger, Adam N Joinson, and Tom Buchanan. Internet Users' Perceptions of 'Privacy Concerns' and 'Privacy Actions'. *International Journal of Human-Computer Studies*, June 2007.

[71] Nick Seaver. Knowing Algorithms. *Media in Transition*, 8, 2013.

[72] Ben Shneiderman. Direct Manipulation: A Step beyond Programming Languages. In *Proc. CHI*, 1981.

[73] Ben Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proc. VL/HCC*, 1996.

[74] Keith Spiller. Experiences of Accessing CCTV Data: The Urban Topologies of Subject Access Requests. *Urban Studies*, 53(13), 2016.

[75] Clay Spinuzzi. The Methodology of Participatory Design. *Technical Communication*, 2005.

[76] State of California. California Consumer Privacy Act. 2018. https://oag.ca.gov/privacy/ccpa.

[77] Study Participants. Sketches, 2021. https://www.blaseur.com/papers/soups21-sketches.zip.

[78] Omer Tene and Jules Polonetsky. A Theory of Creepy: Technology, Privacy and Shifting Social Norms. *Yale JL & Tech.*, 16:59, 2013.

[79] The European Parliament and the Council of the European Union. Regulation (EU) 2016/679 (General Data Protection Regulation), 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

[80] Slim Trabelsi and Jakub Sendor. Sticky Policies for Data Control in the Cloud. In *Proc. PST*, 2012.

[81] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising. In *Proc. SOUPS*, 2012.

[82] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. A Study on Subject Data Access in Online Advertising After the GDPR. In *Proc. DPM*, 2019.

[83] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. Measuring the Impact of the GDPR on Data Sharing in Ad Networks. In *Proc. ASIACCS*, 2020.

[84] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *Proc. CCS*, 2019.

[85] Jeffrey Warshaw, Nina Taft, and Allison Woodruff. Intuitions, Analytics, and Killing Ants: Inference Literacy of High School-educated Adults in the US. In *Proc. SOUPS*, 2016.

[86] Susanne Weber and Marian Harbach. Participatory Design for Security-Related User Interfaces. In *Proc. USEC*, 2015.

[87] Miranda Wei, Madison Stamos, Sophie Veys, Nathan Reitinger, Justin Goodman, Margot Herman, Dorota Filipczuk, Ben Weinshel, Michelle L. Mazurek, and Blase Ur. What Twitter Knows: Characterizing Ad Targeting Practices, User Perceptions, and Ad Explanations Through Users' Own Twitter Data. In *Proc. USENIX Security*, 2020.

[88] Ben Weinshel, Miranda Wei, Mainack Mondal, Euirim Choi, Shawn Shan, Claire Dolin, Michelle L. Mazurek, and Blase Ur. Oh, the Places You've Been! User Reactions to Longitudinal Transparency About Third-Party Web Tracking and Inferencing. In *Proc. CCS*, 2019.

[89] Zhi Xu and Sencun Zhu. SemaDroid: A Privacy-Aware Sensor Management Framework for Smartphones. In *Proc. CODASPY*, 2015.

[90] Yaxing Yao, Davide Lo Re, and Yang Wang. Folk Models of Online Behavioral Advertising. In *Proc. CSCW*, 2017.

[91] Angeliki Zavou, Vasilis Pappas, Vasileios P. Kemerlis, Michalis Polychronakis, Georgios Portokalidis, and Angelos D. Keromytis. Cloudopsy: An autopsy of data flows in the cloud. In *Proc. HCII*, 2013.

A Study Protocols

A.1 Consent Form (Shown Before Surveys in Parts 1 and 3)

Description: We are researchers at the University of Chicago doing a research study about data visualization. We hope to generate ideas for a data visualization tool based on your ideas and opinions. This is a three-part study.

- **Survey 1** – a short 5-minute screening survey.
- **Survey 2** – if selected, you will be asked to download your data from an online company. You will complete survey 2, a 10-minute survey in which you will verify that you have received your data and will choose a date and time for 75 minute online focus group with 2-4 other participants.
- **Survey 3 and Focus Group** – during this online session, we will lead you through a survey with a series of activities to inspire ideas for a tool that would visualize the data that you downloaded. These sessions will take place through Google Hangouts or a similar platform. The sessions will be audio-recorded.

Your participation is voluntary.

Incentives: You will receive $1 for completion of the first survey. You will receive $2 for completion of the second survey, which verifies that you have downloaded your data. You will receive $25 for completion of the third survey and participation in the video call.

Risks and Benefits: Your participation in this study does not involve any risks to you beyond those of everyday life. Taking part in this research study may not benefit you personally, but we may learn new things that could help others.

Confidentiality:

- No personally-identifiable information will be collected from you.
- If you decide to withdraw from this study, the researchers will ask you if the information already collected from you can be used.
- Any reports and presentations about the findings from this study will not include your name or any other information that could identify you. In some cases, you might provide personal stories or beliefs that we might quote or paraphrase as part of our research findings – any personally identifying information will be removed to protect your privacy.
- Identifiable data will never be shared outside the research team.
- De-identified information from this study may be used for future research studies or shared with other researchers for future research without your additional informed consent.

Contacts & Questions:

If you have questions or concerns about the study, you can contact Blase Ur, Assistant Professor, Department of Computer Science, University of Chicago. blase@uchicago.edu or (773)834-3034.

If you have any questions about your rights as a participant in this research, feel you have been harmed, or wish to discuss other study-related concerns with someone who is not part of the research team, you can contact the University of Chicago Social & Behavioral Sciences Institutional Review Board (IRB) Office by phone at (773) 702-2915, or by email at sbs-irb@uchicago.edu.

Consent:

Participation is voluntary. Refusal to participate or withdrawing from the research will involve no penalty or loss of benefits to which you might otherwise be entitled.

By clicking "Agree" below, you confirm that you have read the consent form, are at least 18 years old, and agree to participate in the research. Please print or save a copy of this page for your records.

◯ I agree to participate in the research
◯ I do NOT agree to participate in the research.

The transcriptions of the recordings taken as part of this research can be included in publications and presentations related to this research.

◯ Yes
◯ No.

A.2 Part 1 Survey (Demographics and Screening)

[Consent form]
Welcome to part 1 of the study. You will be asked a few demographic questions. If selected, you will be notified via Prolific with instructions for the next part.
What is your age? ◯ 18-24 ◯ 25-34 ◯ 35-44 ◯ 45-54 ◯ 55-64 ◯ 65 or older ◯ Prefer not to say
What is your gender? ◯ Male ◯ Female ◯ Non-binary ◯ Prefer to self-describe ◯ Prefer not to say
What is the highest degree or level of school you have completed? ◯ Some high school ◯ High school ◯ Some college ◯ Trade, technical, or vocational training ◯ Associate's degree ◯ Bachelor's degree ◯ Master's degree ◯ Professional degree or doctorate ◯ Prefer not to say
What is your race? Please select all that apply. ◯ White ◯ Black or African American ◯ American Indian or Alaska Native ◯ Asian or Pacific Islander ◯ Other (Please specify) ◯ Prefer not to say
Are you of Hispanic or Latino origin? ◯ Yes ◯ No
Which of the following best describes your educational background or job field? ◯ I have an education in, or work in, the field of computer science, computer engineering or IT. ◯ I do not have an education in, nor do I work in, the field of computer science, computer engineering or IT.
Are you currently a student? ◯ Yes ◯ No
As mentioned in the consent form and on Prolific, the final part of the study is a 75 minute online focus group for which you will be compensated $25. We will schedule this session based on your availability. Are you willing and able to participate in a Google Hangouts video call for this part of the study? ◯ Yes, I am willing to participate in a Google Hangouts call for the final part of the study. ◯ No, I am not interested in participating in the final part of the study.
Each of our focus groups will cover one of the sites below. We will use your answer to this question to place you into an appropriate group.

Please check all the sites for which the following is true:

1. You have an account.
2. You use the site frequently (at least once a month).
3. You have full ownership of the account. No one else has access.
4. You would be willing to download your data from this site. You will NOT be asked to send this data to us.

◯ Facebook ◯ YouTube ◯ Spotify ◯ Uber ◯ Amazon ◯ Google
When, in general, would you be available for a 75-minute video call? Please select all that apply. ◯ Monday morning ◯ Monday afternoon ◯ Monday evening

◯ Tuesday morning ◯ Tuesday afternoon ◯ Tuesday evening ◯ Wednesday morning ◯ Wednesday afternoon ◯ Wednesday evening ◯ Thursday morning ◯ Thursday afternoon ◯ Thursday evening ◯ Friday morning ◯ Friday afternoon ◯ Friday evening ◯ Saturday morning ◯ Saturday afternoon ◯ Saturday evening ◯ Sunday morning ◯ Sunday afternoon ◯ Sunday evening
Thank you for completing our screening survey. If selected, you will receive instructions for the next part on Prolific.

A.3 Part 2 Survey (Data Receipt and Knowledge)

Welcome to part 2 of the study. Today you will be asked to verify that you have downloaded and received your data. You will also be asked a few questions related to data, privacy, and the Internet. Finally, you will be asked to indicate your availability for a focus group session.
From which of the following companies did you request your data? This information can be found in your Prolific messages related to this study. ◯ Amazon ◯ Facebook ◯ Google ◯ Spotify ◯ Uber ◯ YouTube
[Participants were then asked to paste in text from an email related to their data download request or from the data download dashboard. We had company-specific screenshots and instructions to guide them. We did NOT ask them to provide any personally-identifiable information. We then asked these two questions:]
I have downloaded all the files that appear in the graphic above. ◯ Yes ◯ No
I know where these files are located on my computer. ◯ Yes ◯ No
Please write what you know about the General Data Protection Regulation (GDPR). Please do not look anything up. Your knowledge about this won't affect your eligibility or compensation in any way.
Please write what you know about the California Consumer Privacy Act (CCPA). Please do not look anything up. Your knowledge about this won't affect your eligibility or compensation in any way.
Which of the following terms have you heard of? Select all that apply. ◯ Data portability ◯ Right of access ◯ Right to be forgotten ◯ None of the above
Before this study, have you ever downloaded the data a company has collected about you? If so, which company?
Have you ever wanted to know what information a company has about you? ◯ Yes ◯ No ◯ Unsure
If companies gave you access to the information they had about you, what would you be most interested in seeing?
Have you ever been notified that data about you (passwords, emails, etc.) had been compromised? ◯ Yes ◯ No ◯ Unsure
What information, if any, do you think companies collect about you when you visit their sites?
Please follow this Doodle Poll link to schedule a time for part 3 of the study. The goal is to schedule the time that works for the most people.
Please observe the following guidelines:

- Enter your Prolific ID instead of your name
- Please choose ALL options that would work for you. This will increase your likelihood of being eligible to participate in part 3, a 75 minute focus group for which we offer compensation of $25.

Thank you for completing part 2 of the study. If you are eligible for part 3, based on your completion of this survey and your availability, we will message you on Prolific with more details.

A.4 Part 3 Survey (Focus Group)

Please do NOT start this survey until you have joined the Google Hangouts call. Check your Prolific inbox for information on how to join the call. Once you have joined, you may proceed to the next section.
[Consent form]
Please choose the company name designated on the PowerPoint. ◯ Amazon ◯ Facebook ◯ Google ◯ Spotify ◯ Uber ◯ YouTube
Please do not proceed to the next section until asked to do so by the session organizer.

Exploration of Files
Take 1-2 minutes to look at the index.html file. This is a visual overview of the folders and files contained in your data download. (Facebook sessions)
Take 1-2 minutes to look at the archive_browser.html file. This gives an overview of what is contained in the files, and also has links to settings related to your data. Make sure you look at all 3 tabs. (Google and YouTube sessions)
Take 1-2 minutes to look at the "Understanding My Data" link found in the Read Me First.pdf file. This gives an overview of what is contained in the files. (Spotify sessions)
Take 1-2 minutes to look at the readme.html file. This gives an overview of what is contained in the files, and also has links to settings related to your data. (Uber sessions)
Take 5 minutes to look through your data on your own. We encourage you to make comments aloud to us and to the other participants as you discover things that you find interesting.
While you look for these items, take time to familiarize yourself with your data, paying particular attention to the information that is included and the format and organization. Here are some things to thing about:

- What information is here?
- What information seems to be missing?
- How is this information presented?
- How is this information organized?
- How easy or difficult is it to find things that you are curious about?
- What, if anything, is confusing?

If you are unable to open your data file from your computer, use this online file viewer. Note: make sure you open this link in a **private browsing window.**
https://jsoneditoronline.org
To use this tool: [we included screenshots to supplement the written instructions]
Click the folder icon. Then click "Open from disk."
Find the data file you want to open and confirm.
You might want to open multiple tabs, one for each file in your data download. This way, you can refer back to a file without having to re-upload it.
Note: there is an option to save to cloud. To protect your privacy, do NOT use this feature.
Please do not proceed to the next section until asked to do so by the session organizer.

Scavenger Hunt

To get you acquainted with your data, we have a short scavenger hunt for you to complete. If you can't find an item, skip it and move on. The goal of this activity is to get you acquainted with your data.While some items can be easily found by looking on the website or app, please only look for the answers in the files that you downloaded. You are welcome to use Windows Explorer, Finder, or any other search tool on your computer. You may also use the [index.html, Read Me First pdf file and the "Understanding My Data" link, archive_browser.html, or readme.html] file found in your data download.
Again, please comment aloud as you find scavenger hunt items or anything else you find interesting.

[Below we have included the scavenger hunt items for all of the companies. The answers (in purple) were not displayed.]

Amazon

1. Have you ever used a gift card to make a purchase? In Retail.OrderHistory csv, "Payment Instrument Type" (column L), search for "Gift Certificate"
2. How many refunds have you been issued? In Retail.OrdersReturned csv OR Retail.CustomerReturns csv, count the number of rows excluding the top row
3. What was the reason for your most recent return? In Retail.CustomerReturns csv, "ReturnReason" (column F). Go to the last row to find the most recent.
4. Around what fraction of your orders are taxed? In Retail.OrderHistory csv, "Price Tax" (column G). Count all orders that have a non-zero tax value, then divide it by the total number of orders, which is the number of rows minus 1.
5. Around what fraction of your orders do you pay shipping? In Retail.OrderHistory csv, "Shipping Charge" (column G). Count all orders that have a non-zero tax value, then divide it by the total number of orders, which is the number of rows minus 1.
6. Around what fraction of your orders are sold directly from Amazon? In Retail.OrderHistory csv, "Marketplace" (column A). Count all orders sold by Amazon.com, then divide it by the total number of orders, which is the number of rows minus 1.
7. What's the date of your most expensive order (excluding tax and shipping charges) this year? In Retail.OrderHistory csv, first scroll through "Order Date" (column C) until you find orders from 2020. Then, look at "Price" (column F) until you find the most expensive order.
8. Find the product name of your most recently returned item. [Hint: this might require looking in more than one file.] First, you need to get the order ID. There are two ways to do this. (i) In Retail.OrdersReturned csv, find the last order ID (which is the most recent) in the orderID column (column C). (ii) In Retail.CustomerReturns csv, find the last order ID (which is the most recent) in the orderID column (column A). Now, search for that order ID number in Retail.OrderHistory. Once you've located the appropriate row, scroll over to find the product name (column Q).

Facebook

1. Which file contains Facebook search history? In search_history folder, your_search_history.html file
2. Find a friend request you sent. [Hint: you might want to check the Friends folder!] In friends folder, sent_friend_requests.html file
3. Find a Facebook user whose friend request you rejected or who you removed as a friend. In friends folder, rejected_friend_requests.html OR removed_friends.html files
4. Find the first documented Facebook page you liked. [Hint: your likes are stored chronologically in an html document, can you find it?] In likes_and_reactions folder, pages.html, last entry
5. What are some of your ad interests? Does anything surprise you? In ads_and_business folder, ads_interests.html
6. Find an advertiser who uploaded information about you. Do you recall ever interacting with that advertiser?In ads_and_business folder, advertisers_who_uploaded_a_contact_list_with_your_information.html
7. In approximately how many cities have you logged into Facebook? [Hint: that seems like it might be related to security!] In security_and_login_information, where_you're_logged_in
8. When did you register for your Facebook account? [Hint: it's not in the about you folder!] In profile_information folder, profile_information.html, value of Registration Date
9. How many events have you responded to in the past 6 months? In events folder, your_event_responses.html
10. What 'life stage' does Facebook think your friends are at? In about_you folder, friend_peer_group.html
11. What was the last date you updated your profile picture? In profile_information folder, profile_update_history.html

Google

1. What is the date of the most recent search in your history? In search folder, MyActivity.html
2. Find a search for a restaurant or business. In search folder, MyActivity.html
3. Find a search where you asked a question. In search folder, MyActivity.html
4. Find a search for a product you wanted to buy. In search folder, MyActivity.html
5. Find a search you made late at night. In search folder, MyActivity.html
6. Find a trip for which the mode of transportation was most likely a vehicle. In location history folder, Location_history.json, find high confidence number for vehicle
7. Find the latitude and longitude of a location that you likely traveled to on foot. In location history folder, Location_history.json, find high confidence number for on foot, then find corresponding lat/long pair
8. Find the specific address of a place you visited. In semantic location history folder, value of "address"

Spotify

1. How many users are you following? In follow.json, value of followingUsersCount
2. What is your display name? In identity.json, value of displayName
3. Find an album name from your library that starts with the same letter as your first name. If you can't find one, choose another letter. In yourlibrary.json, value of album
4. Find a search you made for a song or artist. In SerachQueries.json, value of typedQuery OR selectedQuery
5. According to your data, was your Spotify account created from Facebook? In Userdata.json, value of createdFromFacebook
6. Find the name of one of your playlists in your data. In Playlist1.json, value of name
7. What is the first song in the playlist you found for #6? In Playlist1.json, value trackName of first item
8. For how many milliseconds did you listen to the song you found in #7? [Hint: You might want to look at Streaming History.] If you've listened to it multiple times, pick one instance. In SteamingHistory0.json, search for the trackName found in #7, then it's the value of msPlayed
9. Find a song in your library where the track name is the same as the album name. In yourlibrary.json, value of album and value of track

Uber

1. What is the user rating associated with your account? In profile_data.csv, "Rating" (column E)
2. Where did you take your last recorded Uber from? In trips_data.csv, top row (most recent), "Begin Trip Address" (column H)
3. Were you referred to Uber? In profile_data.csv, "Referred to Uber?" (column J)
4. How many payment methods are listed under your account? In payment_methods, it's the number of rows minus 1
5. What is the longest (in terms of distance) Uber ride you've taken? In profile_data.csv, "Distance (miles)" (column m), find the largest
6. If you have an Uber Eats account–what was your last order and where was it from? [Hint: you might have to piece together this question from the available information!] Last order: eats_order_details.csv, "Item Name" (column F). Where it was from: eats_restaurant_names.csv, "Restaurant Name" (column C)

YouTube

1. Find a song you listened to on YouTube. In history folder, watch-history.html, scroll until you find a song
2. What is the date and time of the most recent video you watched on YouTube that was NOT music? In history folder, watch-history.html, most recent is at the top
3. Have you ever commented on a video? If so, find your oldest comment. In my-comments folder, my-comments.html, find the oldest one
4. Find a video you have watched that starts with the same letter as your first name. If you can't find one, pick another letter. In history folder, watch-history.html, scroll until you find a song
5. Find a search you made during a summer month. If you can't find one, pick another season. In history folder, search-history.html, scroll until you find a search
6. Have you ever uploaded a video to YouTube? If so, how many views did it get? If you have uploaded multiple, pick one. In videos folder, videoName.json, value of viewCount
7. Do you have any videos in your watch later list? If so, find the description of one of the videos in that list. In playlists folder, watch-later.json, value of description
8. Do you subscribe to any channels? If so, find the description of one of the channels you subscribe to. In subscriptions folder, subscriptions.json, value of description

Please do not proceed to the next section until asked to do so by the session organizer.

Highlight Activity

Amazon: Alexa; Amazon Drive; Amazon Music; Amazon Lists Wishlist; Amazon Smile Customer Data; Appstore; Customer Communication Experience; DSAR Customer Retail Addresses; Devices Registration; Digital Action Benefit; Digital Content Ownership; Digital Customer Attributes; Digital Prime Video Customer Title Relevance Recommendations; Digital Prime Video Location Data; Digital Prime Video View Counts; Digital Prime Video Viewing History; Kindle Reading Insights; Outbound Notifications Amazon Application Update History; Outbound Notifications Email Delivery Status Feedback; Outbound Notifications Notification Engagement Events; Outbound Notifications Push Sent Data; Outbound Notifications Sent Notifications; Payment Options Amazon Pay Browser Behavior Data; Payment Options Payment Instruments; Physical Stores Whole Foods; Prime Acquisition; Retail Amazon Custom; Retail Cart Items; Retail Customer Attributes; Retail Customer Contacts; Retail Customer Profile; Retail Customer Returns; Retail Customer Service Chats; Retail Gift Certificates; Retail Order History; Retail Orders Returned Payments; Retail Orders Returned; Retail Promotions; Retail Region Authority; Retail Reorder; Retail Sports Fan Experience; Retail Website Authentication Tokens; Search Data; Subscription and Digital Order History

Facebook: About You; Ads; Apps and Websites; Comments; Events; Followers and Following; Friends; Groups; Likes and Reactions; Location; Marketplace; Messages; Other Activity; Pages; Payment History; Photos and Videos; Posts; Profile Information; Saved Items and Collections; Search History

Google: Android Device Configuration Services; Arts & Culture; Calendar; Chrome; Classroom; Contacts; Crisis User Reports; Data Shared for Research; Drive; Fit; Fusion Tables; G Suite Marketplace; Google Help Communities; Google Input Tools; Google My Business; Google Pay; Google Photos; Google Play Books; Google Play Games Services; Google Play Movies & TV; Google Play Music; Google Play Store; Google Shopping; Google Translator Toolkit; Groups; Handsfree; Hangouts on Air; Home App; Keep; Location History; Mail; Maps; Maps (your places); My Activity; My Maps; News; Posts on Google; Profile; Purchases & Reservations; Reminders; Saved; Search Contributions; Shopping Lists; Street View; Tasks; Textcube; Voice; YouTube and YouTube Music; YouTube Gaming

Spotify: Car Thing; Family Plan; Follow; Identity; Payments; Playlist; Search Queries; Streaming History; User Data; Your Library

Uber: Account and Profile; Driver; Eats; Jump; Regional Information; Rider

YouTube: All Playlists; Likes; My Comments; Search History; Subscriptions; Uploads; Videos; Watch History; Watch Later

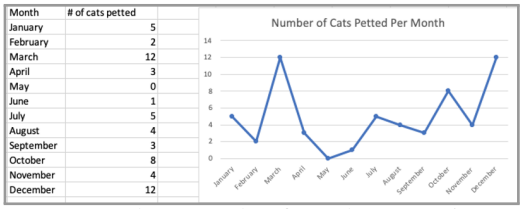
Please do not proceed to the next section until asked to do so by the session organizer.

Data Visualization 101

The ultimate goal of our project is to design tools that make it easier for you to understand your data downloads. One way to accomplish this is data visualization in which information is displayed using visuals like charts, graphs, and maps. Let's take a look at a couple of examples of data visualization.

When raw statistics or numbers are reported in the news, sometimes it can be hard to digest that information. Journalist David McCandless founded a site called informationisbeautiful.net, which offers visualizations of the daily news. Take a couple minutes to explore visualizations of the news you find most interesting. Open this link in a new tab: `https://informationisbeautiful.net`

Say I wanted to know how many cats I petted each month in 2019. One option would be to look at an excel spreadsheet with this information. However, if I wanted a visual representation of this data, I might take my spreadsheet and convert it to a line chart. I can easily look at this line chart and conclude that March was a great month for cat petting.



Month	# of cats petted
January	5
February	2
March	12
April	3
May	0
June	1
July	5
August	4
September	3
October	8
November	4
December	12

Figure 3: An example screenshot from the Data Viz 101 activity.

A.5 Focus Group Session Script

Hi. Thank you all for coming to this session. My name is ____ , and I will be co-leading this session with ____. Please begin the study on Prolific, which will take you to a Qualtrics survey. In part 1, you consented to participation in this study, which includes the audio recording of today's session. Please take a moment to confirm your consent to being recorded. Additionally, there will be a drawing activity. You will need a writing utensil and a piece of paper or the digital drawing tool of your choice. Please type "ready" in the chat once you have answered the consent questions and have a drawing tool on hand.

[Ensure that participants have completed consent form. Make sure everyone has pen and paper.]

Now, let's take a minute to introduce ourselves. Please say your first name, or the name by which you want to be referred during the session. Please also type your name in the chat when you are finished. For your protection, do not use your real last or middle name. Also, please share a non-sensitive fun fact about yourself. Finally, nominate someone to go next.

[One of the session leaders should go first to set the tone. "My name is ____ and I love cats. ____, go ahead!" Continue with introductions. We typed out a reference in the chat once everyone had introduced themselves: *Participant 1 - ____ Participant 2 - ____ . . . etc.*]

During today's session, you will be asked to look at your data and have the opportunity to answer questions aloud. Please remember that you are under no obligation to disclose specific information about you or your data during this session. We will also be recording the audio of this session. If you say something that you don't want on record, please let one of us know afterward, and we will delete that portion of the audio. We ask that everyone have their cameras turned on during the session. However, to protect your privacy, we will not record the video or take screenshots of the session. We ask that you do not do so either. Finally, to help make this session run smoothly, please mute your microphone when you are not speaking.

During this session, we will be generating ideas for a tool that will help people understand their data downloads. Here is an overview of today's activities.

- GDPR/CCPA Overview (2 minutes)
- Exploration of files, then Scavenger Hunt (12-15 minutes)
- Discussion (10-15 minutes)
- Highlight Activity (3-4 minutes)
- Data Visualization 101 (5-7 minutes)
- Sketch activity (10-15 minutes)

But first, let's talk about why you are able to download your data in the first place.

GDPR/CCPA Overview

In response to privacy concerns about online data, two major privacy laws were passed recently. The General Data Protection Regulation came into effect in the European Union in May 2018. GDPR grants users the right to access the data that an online company has about them–a right that you all have exercised as part of this study. Inspired by the GDPR, California produced a similar law, called the California Consumer Protection Act, that went into effect at the beginning of this year. These laws grant other rights, like the right to data portability, the right to erasure of your data, and the right to correct false information about yourself, but today we're going to focus on the right to access. Does anyone have any questions about GDPR or CCPA?

Exploration of Files

So let's talk about your data download. First, did anyone look at their data before this session?

What types of information were you/are you expecting to find in your data download?

What types of information do you want to see?

Please navigate to the next page of the Qualtrics survey.

First, we'll take 1-2 minutes to look at the [index.html, Read Me First pdf file and the "Understanding My Data" link, archive_browser.html, or readme.html] file. This is a visual overview of the folders and files contained in your data download. [This section was omitted for Amazon, which doesn't provide such a file in the order data download.]

Next, take 5 minutes to look through your data on your own. We encourage you to make comments aloud to us and to the other participants as you discover things that you find interesting.

While you look for these items, take time to familiarize yourself with your data, paying particular attention to the information that is included and the format and organization. There are a few guiding questions on Qualtrics. [See Survey 3]

What are your initial reactions?

What surprised you?

What was it like navigating this file?

Did your expectations match the reality of what was contained in the file?

Is there anything you wanted to see but didn't?

Scavenger Hunt

Now we're going to do a short scavenger hunt to help get you acquainted with your data. Please proceed to the next section. You will see a list of items to search for in your data. If you can't find an item, skip it and move on. It's possible that an item may not be in your data at all. While some items can be easily found by looking on the website or app, please only look for the answers in the files that you downloaded. However, you are welcome to use Windows Explorer, Finder, or any other search tool on your computer. You may also use the [index.html, Read Me First pdf file and the "Understanding My Data" link, archive_browser.html, or readme.html] file found in your data download. The goal of this activity is to get you acquainted with your data download. You don't need to write anything down, but you can if you'd like. We'll spend around 5 minutes on this activity.

Again, please comment aloud as you find scavenger hunt items or anything else you find interesting.

Discussion Questions

Next, we have some discussion questions.

Scavenger Hunt

1. How many scavenger hunt items did you find?
2. Did the [index.html, Read Me First pdf file and the "Understanding My Data" link, archive_browser.html, or readme.html] help you with the scavenger hunt?
3. What was it like navigating this file?
4. Was there any information collected about you that you were surprised by? Why?
5. Is there any data you think the company has about you that is missing from these files?

General

1. What are some reasons, if any, you might want to have access to your data?
2. From which websites or apps (social media, online shopping, ride share, etc.) would you be most interested in downloading your data?
3. What pieces or types of data are most important for you to see in a data download?
4. What pieces or types of data are not important for you to see in a data download?
5. How was the process of requesting your data?
6. How did you navigate to the page that gives you access to your data?
7. How long did it take for you to be able to access your data?
8. Were you previously aware of how to navigate through a csv/json/txt file?
9. What records were you looking to find from your data download? Were you able to access them?

Privacy

1. Was there any information collected from you that made you uncomfortable? Why? Do you think this information is useful or important for the company to have?
2. If after seeing this data download, you wanted to share less info with the website, what steps would you take?
3. Do you feel the data about you is accurate?

Design

1. What elements of the data download layout are most intuitive to you, and which were the most difficult to navigate?
2. How was your data separated into folders? Does this organization make sense to you? Can you think of other ways to organize?
3. Was any terminology used in the data download unclear? If so, which terms?
4. Are the file names descriptive?
5. When you think about your interaction with this platform, is it easy to trace your online activity through this data file?

Other

1. How would you feel about adding aggregate statistics to your data–for instance, your average ride cost (Uber), number of 'liked' pages per month (Facebook)
2. How would you feel about a setting that lets you choose different levels of specificity for your report?
3. How would you feel about a tool that helped you make privacy-protective choices based on your data?
4. How would you feel about reminders to do things like delete your data or modify your settings?

Highlight Activity

Please advance to the next page of the survey. For this next activity, we will give you a list of the categories or folder names in your data. Please highlight the ones that would be most important for you to see and understand in your data download. To highlight an item, double click in, then click the word important. You may highlight as many as you'd like. Do not think too hard about your answers. Some categories have confusing or unclear names. Go with your gut.

Data Visualization 101

Please navigate to the next page of the survey. The ultimate goal of our project is to design tools that make it easier for you to understand your data downloads. One way to accomplish this is data visualization in which information is displayed using visuals like charts, graphs, and maps. Let's take a look at a couple of examples of data visualization.

Beautiful News Daily

When raw statistics or numbers are reported in the news, sometimes it can be hard to digest that information. Journalist David McCandless founded a site called information is beautiful dot net, which offers visualizations of the daily news. Take a couple minutes to explore visualizations of the news you find most interesting. As you explore, pay close to attention to examples that synthesize multiple pieces of information to give a more complete or interesting account.

https://informationisbeautiful.net/beautifulnews/

Does anyone want to share a visualization they found particularly interesting or well designed? [If so, ask them to drop the link in the chat.]

Does anyone have an example where multiple types of information were synthesized?

Excel Line Graph

Here is a more basic example of data visualization. Say I wanted to know how many cats I petted each month in 2019. One option would be to look at an excel spreadsheet with this information. However, if I wanted a visual representation of this data, I might take my spreadsheet and convert it to a line chart. I can easily look at this line chart and conclude that March was a great month for cat petting. Data visualization doesn't have to be super complex–it could be a simple graph!

Sketch Activity

Please advance to the next page of the survey. For this last activity, we would like you to imagine that someone designed a tool that generated a visualization of your data. Please sketch your ideal version of this visualization on a piece of paper or using your favorite drawing tool. Do not feel limited to what has been discussed in this session. Don't worry about the quality of your sketch. The goal is to get your ideas across. For example, if you can't draw a unicorn, simply write "picture of unicorn."

You can take several approaches. You could sketch the overall layout of the tool, like the website layout and the different options that the tool provides. Or you could focus on representing a specific type of data, for example, location data. You might also consider how to synthesize multiple pieces of information like we saw in the daily news data example. You are also welcome to take more than one approach. There are a few guiding questions on Qualtrics. [See Survey 3]

Once you're done, please scan in or take a photo of your drawing and upload it to the survey. If you're doing it from your computer, you can click the link. If you're doing it from a phone or other device, you can type in the URL or open your camera to scan the QR code. Type "ready" in the chat when you're done.

Now we're going to share our ideas. Please explain your sketch. If you'd like, we can share your drawing with the group, but you can opt for a verbal explanation only. Who would like to go first?

Closing Remarks

Thank you for participating in our study. You all have been fabulous! Please advance to the final page of the survey to get the completion code, which you will use on Prolific to receive compensation for your participation. If you have any questions about the study, please ask now, or refer to the consent form for contact details. We will stick around for a few minutes.

B Instructions for Downloading Data

Amazon

Part 1: Request Your Data

1. Go to the following URL: https://www.amazon.com/gp/help/customer/display.html?nodeId=G5NBVNN2RHXD5BUW
2. Click the "Request My Data" link.
3. Log in with your Amazon username and password. Then select "Your Orders" from the drop-down menu. Then click "Submit Request."
4. You should see this message: [screenshot of message]
5. Log in to the email associated with your Amazon account. Find the email with the subject line "Your Data Request Confirmation." Click the "Confirm Data Request" button.
6. You should see this message: [screenshot of message]
7. It may take anywhere from a couple hours to a couple days for your data to be ready. Amazon will notify you by email when your data is ready.

Part 2: Download Your Data

8. Login to the email associated with your Amazon account. Find the email from Amazon with the subject line “Your Data Request.” Click the yellow “Download Data” button in the body of the email.
9. You will be redirected to a new page. You may be asked to login to your Amazon account. Click the “Download” button next to all of the files.
10. Make sure you remember where you saved these files. You will need them for part 3 of the study.

Facebook

Part 1: Request Your Data

1. Go to facebook.com.
2. Login with your username and password.
3. Click the blue triangle in the upper right corner.
4. Click "Settings" from the drop down menu.
5. Click "Your Facebook Information" on the left column
6. Click "View" under "Download Your Information."
7. Ensure that "All of my data," "HTML," and "High" are selected. Then click "Create File."
8. You should see this message: [screenshot of message]
9. It may take anywhere from a couple hours to a couple days for your data to be ready.

Part 2: Download Your Data

10. Facebook will notify you when your data is ready either by email or via a Facebook notification.
    - **Option 1**: Login to the email associated with your Facebook account. Find the email with the subject line “Your Facebook information file is ready.” Click the “Download Your Information” link found in the body of the email.
    - **Option 2:** Click on the Facebook notification that looks like this: [screenshot of notification]
11. You will be redirected to a new page. You may be asked to login to your Facebook account. Click “Download” on your most recent file.
12. Make sure you remember where you saved this folder. You will need it for part 3 of the study.

Google

Part 1: Request Your Data

1. Go to https://myaccount.google.com/?utm_source=sign_in_no_continue
2. Log in with your username and password. Please use your primary Google account. Note: if you are already signed in, you can skip this step.
3. Click “Data & Personalization” from the left column
4. Scroll down until you see “Download, delete, or make a plan for your data.” Click “Download your data.”
5. Click “Deselect all.”
6. Scroll down until you see “Location History.” Check the box.
7. Scroll down until you see “My Activity.” Check the box.
8. Click the “All activity data included” button.
9. Click the “Deselect All” button.
10. Check the “Search” box.
11. Press the “OK” button.
12. Click “Next Step” at the bottom right corner.
13. Leave all the presets alone. The page should look like this: [screenshot of page]
14. Click “Create export.”
15. It may take anywhere from a couple hours to a couple days for your data to be ready. Google will notify you by email when your data is ready.
16. Note: Some people have reported that they didn’t receive an email. If you haven’t received an email after a couple days, go to https://takeout.google.com to see if your data is ready.

Part 2: Download Your Data

17. You will receive a link to the email address associated with your account. Follow this link and press “Download.”
18. Make sure you remember where you saved this folder. You will need it for part 3 of the study.

Spotify

Part 1: Request Your Data

1. Go to spotify.com.
2. Log in with your username and password.
3. Click "Account" under the "Profile" menu in the top right corner.
4. Click "Privacy settings" from the left column.
5. Scroll down to the "Download your data" section. Click the "Request" button.
6. Log in to the email associated with your Spotify account. Find the email with the subject line "Confirm your Spotify data request." Click the "confirm" button.

7. You should see this message: [screenshot of message]
8. It may take several days for your data to be ready. Spotify will notify you by email when your data is ready.

Part 2: Download Your Data

9. Login to the email associated with your Spotify account. Find the email with the subject line " Your Spotify personal data is ready to download." Click the green "Download " button in the body of the email.
10. Type in the password to your Spotify account and click “verify.” The download will start automatically.
11. Make sure you remember where you saved this file. You will need it for part 3 of the study.

Uber

Part 1: Request Your Data

1. Go to the following URL: https://auth.uber.com/login/?breeze_local_zone=dca1&next_url=https%3A%2F%2Fmyprivacy.uber.com%2Fprivacy%2Fexploreyourdata%2Fdownload%3F_ga%3D2.160201528.441384756.1587066962-1367774538.1587066962&state=K5fXVafN4vyOBujPSOoPLCsftsZFkaPRRmI81J_NvwY%3D
2. Enter your email address.
3. Enter your password.
4. Enter your phone number, and then the 4-digit code.
5. Click "Request Your Data."
6. You should see this message: [screenshot of message]
7. It may take several days for your data to be ready. Uber will notify you by email when your data is ready.

Part 2: Download Your Data

8. Login to the email associated with your Uber account. Find the email with the subject line “Your Uber data is ready for download.” Click the green “Go to Download Page” button in the body of the email.
9. You will be redirected to a new page. You may be asked to login to your Uber account. Click the blue “Download” button.
10. Make sure you remember where you saved this folder. You will need it for part 3 of the study.

YouTube

Part 1: Request Your Data

1. Go to https://myaccount.google.com/?utm_source=sign_in_no_continue
2. Sign in with your username and password. Please use the primary Google account you use to access YouTube.
3. Click "Data & Personalization" from the left column.
4. Scroll down until you see "Download, delete, or make a plan for your data." Click "Download your data."
5. Click "Deselect all."
6. Scroll down until you see "YouTube and YouTube Music." Check the box.
7. Click "Next Step" at the bottom right corner.
8. Leave all the presets alone. The page should look like this: [screenshot of page]
9. Click "Create export"
10. It may take anywhere from a couple hours to a couple days for your data to be ready. Google will notify you by email when your data is ready.
11. Note: Some people have reported that they didn’t receive an email. If you haven’t received an email after a couple days, go to https://takeout.google.com to see if your data is ready.

Part 2: Download Your Data

12. You will receive a link to the email address associated with your account. Follow this link and press “Download.”
13. Make sure you remember where you saved this folder. You will need it for part 3 of the study.

C Contents of Data Downloads

Category	Amazon	Facebook	Google	Spotify	Uber	YouTube
Communications	Gift Messages	Comments, Sent friend requests, Posts, Like and reactions, Messages, Pokes, Stories	–	–	Info on support conversations with Uber	Comments
Inferences	–	Off-Facebook activity, Ad interests, Advertisers interacted with, Information submitted to advertisers, Advertisers with a contact list of your info	–	List of market segments user is associated with	–	–
Locations	Billing address, Shipping address	Primary location, Profile current city, IP where you've logged in, IP addresses of user device for login, your places	Location, Latitude, Longitude, timestamp, velocity, altitude, activity at location, type of activity	User address, Payment country and card postal code, Family plan address, Car thing shipping address	Locations and times at which a trip (either using Uber Rider or Uber Jump) was started and ended	–
Payment Data	Payment instrument type for orders and subscriptions	Facebook Pay payment history and payment methods	–	Details of payment data	Payment method info	–
Primary Usage Data	Orders and Subscriptions info	Events, Posts, Stories, Following, Friends, Groups,Like and reactions, Marketplace activity, Pokes, polls voted on, support correspondence, Photos and videos uploads, search history, account activity	Searches	Following/Followers data, search queries, streaming history, playlists	Uber rider trips history, Uber jump bike rides history, Uber eats order history	Likes, playlists, video uploads, subscriptions
Search History	–	Time-stamped searches	Time-stamped searches	List of searches with date and time, type of device/ platform used to make search	–	Timestamped searches
User Profile	–	Name, Previous names, Emails, Birthday, Gender, Current City, Hometown, Education, Work experiences, Phone numbers, Bio, Registration timestamp, Profile update history	–	Username, email address, address, mobile number, mobile operator, mobile brand, gender, birthday, registration date, Facebook user ID	Name, email address, mobile number, ratings, and registration date	–
Voice Data	–	Voice recording and transcript	–	List of voice input commands	–	–

Table 2: Our informal categorization of data contained in Amazon, Facebook, Google, Spotify, Uber, and YouTube data downloads.