"If I press delete, it's gone" - User Understanding of Online Data Deletion and Expiration

Ambar Murillo, Andreas Kramm, Sebastian Schnorf, and Alexander De Luca, *Google*

https://www.usenix.org/conference/soups2018/presentation/murillo

This paper is included in the Proceedings of the Fourteenth Symposium on Usable Privacy and Security.

August 12–14, 2018 • Baltimore, MD, USA

ISBN 978-1-939133-10-6

Open access to the Proceedings of the Fourteenth Symposium on Usable Privacy and Security is sponsored by USENIX.

"If I press delete, it's gone" - User Understanding of Online Data Deletion and Expiration

Ambar Murillo, Andreas Kramm, Sebastian Schnorf, Alexander De Luca
Google
{ambarm, akramm, sebschnorf, adeluca}@google.com

ABSTRACT

In this paper, we present the results of an interview study with 22 participants and two focus groups with 7 data deletion experts. The studies explored understanding of online data deletion and retention, as well as expiration of user data. We used different scenarios to shed light on what parts of the deletion process users understand and what they struggle with. As one of our results, we identified two major views on how online data deletion works: UI-Based and Backend-Aware (further divided into levels of detail). Their main difference is on whether users think beyond the user interface or not. The results indicate that communicating deletion based on components such as servers or "the cloud" has potential. Furthermore, generic expiration periods do not seem to work while controllable expiration periods are preferred.

1. INTRODUCTION

With growing storage capabilities and the large amounts of data[1] that people store online, data deletion is a common practice for internet users these days [12]. Reasons for deletion are manifold and range from simple things such as cleaning up your account to more critical tasks like getting data out of the reach of others, i.e. privacy [12].

We know that incomplete understanding of online data deletion can cause problems such as mishandling personal data due to misinterpretation of the process [12]. Ultimately, this can lead to issues with maintaining user privacy. Despite this importance, understanding online data deletion practices from a user perspective is still an understudied topic that deserves more attention. It is important to study what users actually know, need, and want when it comes to online data deletion.

To fill this gap, we conducted a user study with 22 participants with varying demographic backgrounds. In addition, we ran two focus groups with 7 data deletion experts. The main focus of this workstream was on *deletion*, *retention*, and *expiration*. In this work, we define *deletion* as the process of a user-invoked event to remove user generated content from an account. *Retention* refers to how long it takes until data is removed from all entities after it has been deleted. Finally, with *expiration*, we explored if it makes sense to have certain data automatically disappear after a certain period of time (think for instance about Snapchat messages that disappear after a user-defined timeframe).

In this paper, we provide insights into users' understanding of online data deletion, retention, and expiration. Our results can help with designing and communicating deletion in a way that is graspable for users, and as such, help the community to create better user interfaces and user education for online data deletion. For example, we identified two major views on online deletion: one solely based on the user interface and the second about what is going on in the background (with different levels of detail). We also found that data expiration does not follow a chronological order but is rather context-dependent. This means that data that is considered worthless at a certain point in time can become useful again later due to certain events.

[1]Please note that in this work, we focused on user-generated content as opposed to automatically generated data such as different types of metadata (e.g., log data).

USENIX Symposium on Usable Privacy and Security (SOUPS) 2018.
August 12–14, 2018, Baltimore, MD, USA.

2. RELATED WORK

For a long time, humans' ability to remember relied on biological memory and media with limited storage and sharing capacities. Most things were forgotten, and only few were remembered [9]. Even most acts violating social norms were forgotten after some time [4]. However, modern technology, and especially the internet, provides us with new abilities to overcome forgetting. Data can easily be stored, distributed, searched and used. Despite its benefits, this presents new challenges, especially with respect to an individual's privacy. For example, in 2006, a student teacher posted a picture of herself in a pirate costume with the caption "Drunken Pirate" on MySpace. Based on this picture, she was later denied her teaching degree [13].

A lot of research work in the past years has focused on helping people to protect their privacy while still being able to live a digital life. Not surprisingly, much of this work is centered around the content of online social networks, and more precisely, deletion and permanence of this content. For example, Wang et al. [17] showed that regret is a major factor for deletion in Facebook. Similar results were found in research on regrets on Twitter [14]. Interestingly, despite regret, a large scale study on deletion on Twitter [1] found that the majority of deletion cases are rather for corrections/edits. They also showed that content on public social networks like Twitter might not really be gone after deletion due to replies, comments, and internet archives storing them. For example, the meaning of a deleted tweet can, in many cases, be recreated based on replies and mentions. To mitigate this issue, Wang et al. proposed a system to support social network users to post fewer regrettable posts by providing them hints on who will be able to see their posts [16].

In a field trial, this system indeed significantly reduced sharing of potentially regrettable content.

Another important dimension of online data privacy is permanence. Much online content is designed to remain available until it is actively removed by the user, raising questions of how data sharing preferences might change longitudinally. Ayalon and Toch [2] looked at sharing preferences of Facebook content over time, finding a meaningful decrease in willingness to share content as it ages. User behavior, however, did not directly align with these stated preferences, since users did not tend to delete old posts to the same degree that their sharing preferences would have implied. The authors suggest expiration controls as a method to manage longitudinal privacy, with users setting expiration dates for content as they post it. This assumes that people will be able to predict their sharing preferences for content with some degree of certainty. However, past research [3], has found that participants were not particularly good at predicting their privacy preferences over time, therefore, raising questions as to whether setting an expiration date for content as it is created would be in line with users' evolving privacy needs. Bauer et al. [3] also found that participants wanted constant access to posts over time, even if only for reminiscing purposes. Posts associated with changing privacy preferences seemed to be the exception.

Considering that users might not accurately predict their future privacy preferences for their data, Mondal et al. [10] suggested an alternative online data privacy preserving mechanism for older data that moves away from time-based deletion. The authors suggest that, after a given period of inactivity, the user could receive suggestions to remove online content (e.g., Twitter posts).

This past research suggests the use of deletion mechanisms, in the form of expiration, to help users manage their online data privacy. Although users might not be very accurate in their predictions for desired expiration dates for their data, they are quite familiar with the use of deletion as a privacy preserving mechanism. A recent study on deletion practices in cloud storage [12] showed that one of the main motivators for deleting data in the cloud is privacy. Moreover, the paper showed that many problems that came with deletion are grounded in incomplete "mental models". Other research has also established the connection between "mental models" and their impact on user behavior. Wash [18] has also researched user "folk models" about security, finding that users relied on their models to guide their choice of security software, what expert security advice to follow, and how to justify ignoring certain advice. Other "mental models of security and privacy" research has found that knowledge of "mental models" can also be used as a foundation to create better user communication [5].

Most research has focused on the how and why of users' decision making process about deletion. That is, there is only little data on how users see deletion and whether this has consequences for their privacy. In addition, misunderstandings and unfounded expectations of deletion are grounded in incomplete understanding of the deletion process. With this work we provide first insights to fill this gap. This foundational research can then help us better design user education and implementation of data protection regulations.

3. STUDY

To uncover users' understanding of online data deletion and their expectations, we conducted semi-structured interviews in combination with a think-aloud drawing task. In addition, we conducted two expert focus groups to set a baseline to compare the interview study results against.

3.1 Interview Study

We conducted interviews in combinations with drawing tasks. Drawing tasks are a useful tool to uncover participants' understanding [8, 18]. They are particularly appropriate when researching underlying understandings which are hard to verbalize, as can be the case with abstract concepts where participants might lack technical vocabulary [11]. Additionally, drawing tasks are particularly well suited to generate reflective feedback as opposed to reactive feedback [15]. Drawing tasks are usually combined with the think-aloud protocol [7], meaning that participants verbalize what they are thinking as they are drawing, giving the observing researchers further insights into the meaning behind their drawings.

Each interview consisted of three main parts: *General Deletion*, *Deletion Scenarios*, and *Expiration*.

General Deletion - In this part, we explored the participants' online data use and understanding of deletion on a general level. For example, we asked them what online services they use that store data. At the end of this part, participants were asked to draw how they think online data deletion works in general (without a specific use case). As mentioned before, this included a think-aloud task.

Deletion Scenarios - This part explored two deletion scenarios: *Email* and *Social Media*[2]. We picked these two scenarios because they a) are very common, b) come with common deletion tasks, and c) are significantly distinct in how data is shown to users and how deletion works. This includes potential consequences such as the fact that social media data might still retain or be recoverable (literally or by meaning) after deletion due to shares, comments, archives etc. [1]. The two scenarios were counterbalanced, to mitigate learning effects.

Since there are plenty of different email and social media services, which could influence the results, we recruited for the following: We made sure that all participants used the online user interfaces of their respective email provider. All participants used either GMX, Web.de (the two most dominant email providers on the German market), or Gmail, or a combination of those. For social media, all participants were knowledgeable of Facebook (and referred to Facebook in their examples). Please be aware that this limits generalizability.

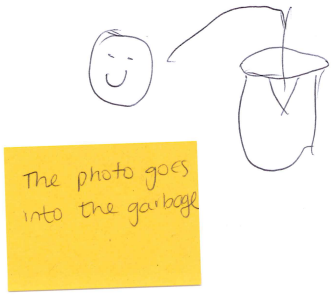
For each scenario, we asked the same questions, including why and when participants delete data on the respective platform. Similar to the general questions part, we also asked participants to create a drawing about how deletion works in each respective scenario, again, applying the think-aloud methodology. For details on the scenarios script, see Appendix A.

Three resulting drawings can be found in Figure 1.

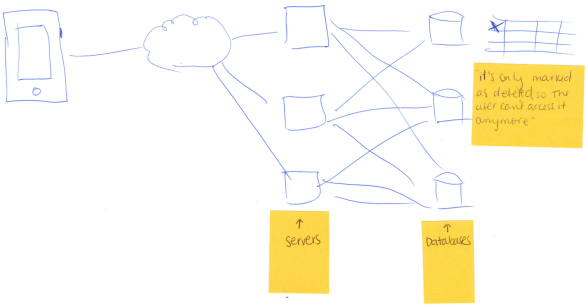
Expiration - The final part was about online data expiration. Here, we wanted to explore if and under what circumstances, participants thought specific data could or should be automatically deleted. We used four scenarios: Online shopping (data: address), email (data: email), social media (data: post/tweet), and search (data: search history). Online shopping and search were added in addition to the deletion scenarios to provide a wide spectrum of potential data. In addition, active deletion (as opposed to expiration) is rather rare in those two scenarios.

The main tool we used in this section was the graph shown in Figure 2. On the x-axis, participants were asked to add events which

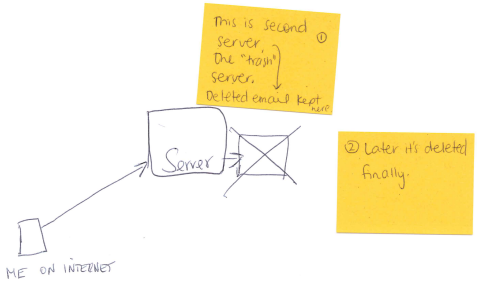
[2]All participants were recruited to be active email and social media users.



(a)



(b)



(c)

Figure 1: Participant diagrams explaining how deletion works: (a) in general, with no given scenario (Participant 6); (b) in an Email scenario (Participant 11); (c) in a Social Media scenario (Participant 8). Yellow notes were added by one of the researchers to clarify the diagrams.

would influence the usefulness of the data. On the y-axis, we asked them to add the respective usefulness rating (for them as users of the service). The question was: "*How useful it is for you that the service provider has this data?*". In the end, they were also asked if there would be an event, at which the data completely loses its usefulness (for details, see Appendix B).

3.1.1 Pilot Study

To verify and improve the study instrument, we ran an internal pilot study. To avoid or mitigate technical bias in the pilot study sessions, we recruited for co-workers from non-tech divisions of our company.

Figure 2: Graph used in the Expiration portion of the Interviews. Participants were given a blank graph (this one is from the Online Shopping scenario), and asked to add events that would influence the usefulness of the data on the x axis and the respective usefulness (for them) on the y axis.

In addition to simple wording improvements, the pilot study helped us to identify more significant changes: For instance, we used the results to identify appropriate scenarios for the expiration and deletion tasks, meaning the tasks covered a wide spectrum both in terms of how the service works and in how it is received by participants. The biggest change after the pilot was in the expiration graph which turned out to be much easier to understand with a time component involved in it as this seemed closer to how users perceive expiration.

3.1.2 Procedure

All interviews were conducted in-person at our premises. At the beginning of each session, participants were introduced to the study. First, they were asked to read and sign a consent form and NDA (was sent to all participants before the study so they had the chance to familiarize themselves with it). After this, the procedure was explained to them and the interviewer told them that they were free to stop the interview at any time or skip questions/parts they did not feel comfortable with (this option was not used by any participant). We also asked them for permission to make a video (and audio) recording of the session which was needed to analyze the data. To protect their privacy, the recordings were anonymized. For example, we only filmed participants' hands and drawings.

After the introduction, an anonymous ID was assigned to each participant, which was used during the analysis instead of their real data. This was followed by the interview. After the interview part was finished, the participants were debriefed and were given the chance to ask questions themselves. Each session lasted around 40 to 70 minutes. Since we always target to provide fair compensation for each respective country, participants received a compensation of around €60, which was based on their travel and time effort.

3.1.3 Participants

We recruited 22 interview participants from Germany. In order to recruit participants from the general population, we worked together with an external recruiting agency providing them with a detailed screener. The study was advertised as being about online data. The most important screening criteria were that they regularly engaged in online deletion activities and the categories of our scenarios: They had to use some sort of social network and own

ID	Age Range	M / F	Occupation
1	18 - 24	M	Student (Business Mgmt.)
2	45 - 54	F	Industrial Eng.
3	45 - 54	M	Self-employed (Tourism sector)
4	45 - 54	F	Freelance Manager in Public Health
5	45 - 54	M	Insurance Salesman
6	35 - 44	M	Hotel Clerk
7	35 - 44	M	Tour Guide
8	55 - 64	F	Clerk
9	25 - 34	M	Business Management
10	18 - 24	F	Student (Accounting)
11	25 - 34	M	Financial services
12	25 - 34	M	Electrician
13	45 - 54	M	Business Management
14	45 - 54	F	Office Manager
15	25 - 34	F	Automotive Engineer
16	45 - 54	M	Painter/Varnisher
17	35 - 44	F	Office Comm. Clerk
18	35 - 44	F	Real Estate Mgmt.
19	45 - 54	F	Florist
20	45 - 54	F	None
21	45 - 54	M	Insurance Salesman
22	35 - 44	F	Office Clerk

Table 1: Demographics of the interview study participants.

an email account and use it through its online interface. They were also familiar with online shopping and regularly performed online searches. With respect to diversity, we targeted for gender diversity, different professional backgrounds and education, as well as differing attitudes towards privacy.

Table 1 lists the demographics of all interview participants.

3.2 Expert Focus Groups

Instead of reproducing the interview study for the experts, we decided to run focus groups. This decision was made to enable discussion among the experts, which we identified as a vital step to come up with a solid baseline to compare the interview results against.

The focus groups were conducted in combination with a drawing task identical to the interview study. In contrast to the interview study, we asked focus group participants to do the 3 drawings (General (no scenario), Email scenario, Social Media scenario) as homework before the actual meeting. All necessary instructions were sent to them via email. They were also asked to bring these drawings with them to the focus group.

The actual focus group session consisted of 3 main parts (in this order): Data deletion in general, the Email scenario, and the Social Media scenario. For each part, each participant (counterbalanced per part) presented the respective drawing and discussed it with the rest of the group. Then, after everyone presented, the participants were asked to decide which parts of the presented drawings they thought were the most important ones that a lay person should know in order to have a good understanding about what is going on when deleting online data.

3.2.1 Procedure

Both focus groups were conducted at our premises in Switzerland. Two researchers conducted the focus groups together. One of them took notes and the other researcher was leading the focus group (including presentations and discussion).

Before attending the focus group, participants were introduced to the study via email. Moreover, they were asked to read and sign a consent form. At the beginning of the sessions, we ensured that all participants understood and signed the consent form, after which we explained the focus group procedure to them. The consent form mainly asked for permission to make a video (and audio) recording of the session which was needed to analyze the data. To protect participants' privacy, the recordings were anonymized like in the interview study. After the introduction, an anonymous ID was assigned to each participant, which was used during the analysis instead of their real data.

This was followed by the actual focus group. In the end, participants were debriefed and were given the chance to ask questions.

Both focus groups lasted around 60 minutes. Each expert received a compensation worth €30 with respect to the time they invested in being part of the study. Please note that they did not have to travel as we conducted the focus groups in their office spaces.

3.2.2 Participants

Overall, we recruited 7 participants from a major tech company, three for the first and four for the second focus group. Recruitment was done through the company's internal communication channels by specifically targeting pre-identified product areas that involve data deletion.

We targeted participants working in security and privacy and for whom online data deletion and retention are part of their daily job. Thus, we considered these participants experts in the technical parts of online data deletion. We aimed for a good mix of job level and nationalities. We also made sure they all worked on different types of products, and thus, types of online data deletion, to mitigate the influence of a certain type of application on the results. For example, occupations ranged from log specialists to data monitoring.

3.3 Data Analysis

Data analysis of the study results (both interviews and focus groups) took roughly two months from first to last session. Overall, three researchers were involved in the analysis process.

For both, the open-ended questions and the drawings, we used the same inductive coding approach: Two researchers independently coded the entire dataset and each separately came up with a codebook. Disagreements between both codebooks (<7%) were discussed by these researchers and resolved in two in-person sessions. The resulting codebook was then iterated on by both researchers by independently re-coding the dataset. Further disagreements were resolved in further in-person meetings. The final codebook was then used by one researcher to code the entire dataset.

Please note that for the drawings, the analysis did not only involve the actual drawings but also the transcripts of what participants said while drawing (think-aloud). Based on those two data sources (transcripts and drawings), we identified all elements that participants thought were part of the process as well as the elements' interdependencies (e.g. backup servers that are connected with each other). For the sketches, we did not differentiate between written elements (in words, e.g. "cloud") and drawn elements.

After the final codes were assigned, a third researcher joined the analysis process and took part in a two days analysis workshop and two additional refinement sessions. In those sessions, the data of the two studies (interviews and focus groups) was used by the three researchers to identify and discuss overarching themes. For instance, the final list of the expert focus group codes was used

Code	Email Scenario	Social Media Scenario	Total
Not needed anymore, old/outdated	10	6	16
Too much data, limited storage	10	–	10
Tidying inbox, avoiding cognitive overload	6	1	7
To remove Spam/Ads	6	–	6
To remove potentially embarrassing content	–	4	4
Don't delete data	3	7	10

Table 2: This table shows how many participants mentioned each of the following reasons for deletion in their responses for each scenario. Numbers do not add up to 22 because each participant can fall into several categories or none.

to iteratively go through the interview data (and codes) again to identify how they related to each other (i.e., how they were similar or different).

For all themes, saturation was reached after a maximum of 14 participants (excluding the experts), which indicates that we caught the main insights with the 22 participants that we recruited.

3.4 Results

In the following, we will outline the main themes that came out of the analysis. The results cover the interview study as well as the expert focus groups. For sake of ease, we will refer to the interview study participants as "participants" and to the focus group participants as "experts". For a discussion of the results, please refer to the discussion section.

3.4.1 Reasons for Deletion

We identified 5 main reasons for why participants delete data. Table 2 shows a frequency table of these themes split up by scenario, as well as the number of participants who stated they to not delete data in those two scenarios at all.

The data shows that deletion is much more frequent in the Email scenario with storage limitation being one of the major reasons. Participants also deleted emails because the data were no longer needed (10 participants). For 6 participants, deletion was carried out to keep a tidy inbox, and to avoid cognitive overload when checking emails.

In the case of Social Media, participants' main reason for deletion was to remove data which they considered outdated and no longer useful (mentioned by 6 participants), as well as potentially embarrassing content (mentioned by 4 participants and not mentioned at all in the Email scenario). For example, Participant 1 recalled deleting a few posts from a Social Media site because *"they were old, weird, embarrassing stuff I posted when I was 15"*. This is in line with reasons for deletion in social media as presented by Wang et al. [16] (we cover all of them under this category).

The number of participants who did not delete their online data differed in the two scenarios as well. While only 3 participants stated they did not delete emails, 7 participants stated that they do not delete social media data. For the Email scenario, essentially unlimited storage was one of the reasons mentioned why online data was not deleted, as stated by Participant 1: *"I just archive them [emails], in case I need them later on"*. For the Social Media scenario, data was not deleted because many participants declared to be passive users, therefore not having much of their own data added to the social media platforms they used, as exemplified by Participant 4: *"I mostly look at others' content, until now I have no need for that [referring to deletion], I don't have any information there that should be deleted"*.

Code	General Deletion	Email Scenario	Social Media Scenario	Total (Unique)
Components involved				
Servers	13	9	8	30 (15)
User Interface (e.g.,trash bin)	10	7	2	19 (11)
Databases (storage)	1	2	3	6 (3)
Internet	6	–	–	6 (6)
Cloud	3	–	1	4 (3)
User Account	–	1	2	3 (3)
Satellite	2	–	–	2 (2)
Finality				
Data is retained	10	8	4	22 (14)
Data is gone	6	3	9	18 (14)
Data remains in other places (e.g., recipient)	–	7	3	10 (10)
Only permanently deleted once deleted from Trash	–	7	1	8 (7)
Deletion is not entirely possible (permanent traces remain, data can be recovered)	4	2	–	6 (5)
Privacy Concerns Expressed				
Don't know if data is really gone from everywhere	–	3	10	13 (11)

Table 3: This table shows for each scenario, how many participants mentioned the following components, finality of deleted data, and whether they expressed privacy concerns regarding deletion. Numbers do not add up to 22 because each participant can fall into several categories or none.

3.4.2 Dimensions of Deletion

Participants mostly described deletion along two main dimensions: *components involved* (e.g., server, "the cloud"), and *finality of the deletion process* (e.g., the end state of the process). Table 3 shows the frequency with which participants mentioned each of the components involved, their understandings regarding the finality of the deletion, and if they expressed privacy concerns regarding the deletion process.

Ten participants in the general deletion scenario and 7 participants in the Social Media scenario associated the process of deletion with elements of the UI, using UI terminology (such as "trash can") to

Code	Email Scenario	Social Media Scenario	Total
Server/ Database/ Cloud/ Internet	13	18	31
With Recipient	11	3	14
Account	4	4	8
Device	4	3	7
No idea	2	1	3
Satellite	2	–	2

Table 4: This table shows how many participants thought data was stored at each of these locations. Numbers do not add up to 22 because each participant can fall into several categories or none.

explain deletion. The general understanding of deletion at the front end was that data is selected, a delete command is given (e.g., pushing a "delete" button), and then data is gone. We can see this deletion process explained by Participant 4: *"If I press delete, it's gone, not anymore inside, that's what I understand."* Four participants' view of online data deletion only included interactions which occurred at the UI front end.

The rest of the participants (18) described a second part of the deletion process which occurs in the back end. The major part of these participants were most familiar with servers as components, although they were not always clear on exactly what functions they served, often using the terms "server" and "cloud" interchangeably (please refer to Table 3 for the number of participants for each scenario). Participants who were aware of the backend also mentioned components such as databases (3 participants in the Social Media scenario), and the internet (6 participants, for the general scenario). In terms of the deletion process, these participants generally described a server or cloud as a place through where data transits, with data being stored on servers, the cloud or databases. Participant 6 describes this process: *"So my data is on the server, I am on the internet and I connect to the server and telling it to delete my data. Then the server isn't going to delete it completely. I think they have a second server, and they transfer the data there, the 'trash' server, and I don't know what will happen afterwards."*

Seven participants in the Email scenario and 3 in the Social Media scenario mentioned that data remains in other places, such as with the recipient, or on the provider's server. Ten participants mentioned that data is retained (general scenario), and four participants mentioned that deletion is not entirely possible (Email scenario), as explained by Participant 10: *"I think that no data is really deleted."* Most privacy concerns were mentioned for the Social Media scenario (10 participants). Also in the Social Media scenario, 9 participants mentioned that data is just gone after deletion.

3.4.3 Data Storage

Across the different interview parts, participants mentioned data storage before deletion as an essential part of the deletion process as its complexity influences whether or not data will be gone (immediately). The most common responses for both scenarios were server, database, cloud or "the internet." As we can see in Table 4, 13 (Email) and 18 participants (Social Media) thought that data was stored in these locations. Please note that participants often used the terms "server" and "cloud" interchangeably, so in their understanding they serve the same or similar purposes. Eleven participants also mentioned that data could be stored with the recipient in the Email scenario, but this was only mentioned by 3 participants in the Social Media scenario (please refer to Table 4). In both scenarios, participants referred to their accounts or their devices as places where data can be located as well.

Code	Email Scenario	Social Media Scenario	Total
Backups for provider, because they can store everything	3	11	14
Law enforcement	8	6	14
To learn about/profile users for marketing purposes	2	5	7
Data not stored indefinitely, provider keeps data for retention period	6	1	7
No idea/ no reason given	2	2	4
Data sold to 3rd parties	–	3	3
Backups to help user recover data	2	1	3
Deletion in the world wide web isn't possible	–	2	2

Table 5: This table shows how many participants thought that data was stored for these reasons, in each of the scenarios. Numbers do not add up to 22 because each participant can fall into several categories or none.

Participants also discussed reasons for data being stored at these locations. As shown in Table 5, for the Email scenario, 6 participants noted that data is stored for a given retention period (the exact duration of which could not be specified), but not indefinitely. In the case of Social Media, this was not the case, with only 1 participant mentioning that data was not stored indefinitely. In terms of reasons why Email data was stored, law enforcement (e.g., as evidence in a criminal case) was the most often mentioned reason (8 participants), such as stated by Participant 17: *"[data is stored] under certain circumstances like legal enforcement."* Backups were another prominent reason why data was kept by the provider (3 participants). Only two participants mentioned that Email data was retained to profile users, possibly for marketing purposes.

In the case of Social Media, as shown in Table 5, backups by the service provider were the most commonly cited reason (given by 11 participants) explaining why providers keep data. The next most commonly given reasons were law enforcement, mentioned by 6 participants, and 5 participants mentioned profiling users for marketing purposes, as explained by participant 8: *"I think they are collecting all data. I don't know where they store it, but they keep it for sending commercials or something like that to your profile, to see your habits and what you like."* In this scenario, two participants thought a consequence of this was that deletion in the world wide web is not possible, and three participants thought that their data was sold to third parties.

Although several participants in both scenarios mentioned data being kept by providers in the form of backups, only 2 participants in the Email scenario and 1 participant in the Social Media scenario thought that these backups were kept to help the user recover data which was accidentally deleted. The other participants saw backups as a part of business processes, and these backups were not necessarily accessible by users.

3.4.4 Automatic Deletion

As mentioned before, one part of the interview study was dedicated to data expiration, i.e. automatic deletion of data. While expiration was mentioned as a theme across the study, the results in this section are mainly based on the expiration exercise.

We explored expiration by having participants consider how useful it is to them that a specific service provider has their data, and how this value evolves over time. It turned out that all participants had major issues thinking about changes to this value over time, for all 4 different scenarios. The overwhelming majority of participants thought that changes to this value were related to specific events which were not time bound, for example canceling their account with the service provider.

In some instances, participants could think about specific situations in which it was no longer useful for them that the service provider held their data. However, this did not necessarily mean the overall end of its usefulness. Certain events were able to "revive" data and increase its value again. For example, several participants thought that it was always useful for websites from which they shop online to have their address for delivery. In the short term, after a particular delivery is received, it was no longer as immediately useful that the service provider has their address data. However, as participants put up another order with those shops, it was once again useful that the provider has their data. Therefore, as opposed to our assumption, the value to users that service providers have their data does not change in a linear fashion but comes in waves or short bursts of usefulness because it is highly context-dependent.

3.4.5 Supportive Deletion Knowledge

Based on their detailed knowledge of online data deletion, experts agreed on six major topics they thought would be beneficial for users to know. "Beneficial" refers to the fact that experts thought that knowing these things will help users to make appropriate decisions that help them better maintain their data privacy. They did not expect users to have such detailed knowledge. However, they assumed that users would benefit from this knowledge. How users could acquire this knowledge was not part of the discussion.

The six topics are:

Backend - This refers to knowing that something is happening beyond the interface. Data will be sent to different servers and will be stored. There will also be copies of the data. This was considered a crucial aspect for the understanding of online data deletion.

Time - Data is not immediately deleted after pressing the delete button. Data may still rest somewhere, even though the users might not be able to see it on their screen.

Backup - Identical data may exist in different places for data storage and data security reasons. In addition, the same information may be stored in different services except the service where it was deleted. For example, travel information might be deleted from an email account but could still be available in a calendar service.

Derived Information - If data is deleted, its essence might still exist. For instance, a user might have deleted a song from a playlist, but the musical interest profile still has this information. Unlearning of derived information like this takes time and thus, deleting data might not immediately change the corresponding profile.

Anonymization - In many cases, a first step of deletion is removing the connection between the data item and the user. After this point, the data might still exist for a while but cannot be related to the user anymore.

What experts consider helpful for users to know	Participants' mentions across all scenarios (N=22)
Backend	18
Time	16
Backup	7
Derived information	1
Anonymization	1
Shared Copy	7

Table 6: Concepts experts think users should know in order to better understand deletion, and the number of participants who are at least slightly aware of these.

Shared Copies - Experts added that users should know about shared copies. Other users might have a copy of the data, e.g., a deleted email. As one expert put it: "Better think before posting and regretting it later."

3.4.6 Expert and Participant Knowledge

In the last rounds of data analysis (e.g., during the workshop days), this list was used to analyze overlap with what participants mentioned throughout the study. We counted an overlap if the participant had mentioned the item at least once during the whole interview (including the drawing tasks). Please note that degrees of knowledge between participants varied significantly. Some participants briefly or inaccurately touched one of the topics and did not further elaborate - even when prompted to do so. However, we wanted to learn, if participants are generally aware of the topics experts mentioned. Thus, we did not differentiate whether participants thoroughly discussed these topics or only briefly mentioned them.

In this analysis, we observed a huge discrepancy between the different topics experts consider helpful for users to know about deletion. Most interview participants expressed awareness for two of the topics: 18 out of 22 participants were aware that something is happening in the Backend and 16 participants acknowledged that it will take some time until data is finally deleted (see Table 6). However, only few participants brought up the topics of Backup (7/ 22), Derived Information (1/ 22) and Anonymization (1/ 22). In the Email and Social Media scenarios, only 7 participants mentioned that other users might still hold a copy of the data they deleted.

3.4.7 Views and Understanding of Deletion

Overall, we found that participants differed in their view of online data deletion across five parameters. The first two were *components involved* in deletion, and the *terminology* used to refer to them. The third was how these *components interact*, and the fourth was whether a *backend* (anything beyond the UI) was identified. Finally, their understanding of online data deletion was also different in the *duration* of the deletion process.

It should be noted that these parameters are reflecting the complexity of the participants' views of deletion, and not their technical accuracy. Thus, the following should not be interpreted as a quality rating of the responses.

By analyzing participants' responses across these parameters, we identified two general distinct categories of understanding of deletion[3] as shown in Figure 3. The first category reflects a UI-centric understanding of deletion. Therefore, we refer to it as the *UI-Based*

[3]During the iterative analysis process, we at first identified 4 categories that we then narrowed down to the 2 presented here.

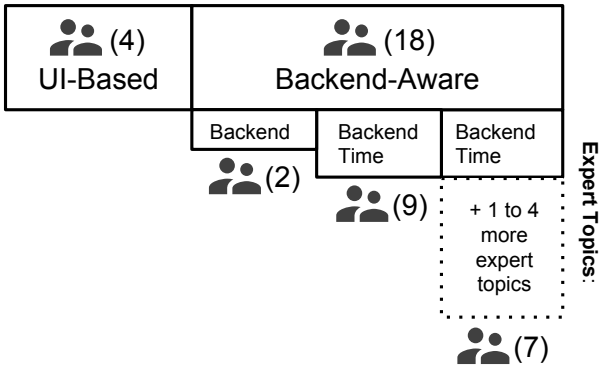


Figure 3: Two categories of user understanding of online data deletion: UI-Based and Backend-Aware. The second category can be subdivided using the topic list of the expert focus group: Backend, Time, Backup, Derived Information, Anonymization, Shared Copy.

Figure 4: A UI-Based view of online data deletion, explaining how email deletion works (Participant 4).

category. The 4 participants that fell into this category displayed an understanding of deletion and a terminology completely based on the UI components they were most familiar with, such as checkboxes to select emails, and then pressing the delete button, so that data ends up in the trash bin. Backend knowledge was not part of this view. Consequently, participants in the UI-Based category described the deletion process as being completed within seconds of clicking the delete button. Figure 4 shows a sample diagram of the UI-Based category.

The second category we identified was more complex. The 18 participants that fell in this category identified more components involved in the deletion process, particularly backend components such as servers or the cloud. Therefore, we refer to this as the *Backend-Aware* category. The components mentioned by the participants also interact with each other, such as sending a delete command to a server from a device where the user is accessing their email account. However, the terminology used to describe these components was often inconsistent, with terms such as "cloud" and "server" being used interchangeably. Since a backend to the deletion process was identified, participants tended to understand the deletion process as taking longer than a few seconds, even if the exact duration of the process could not be identified. Please refer to Figure 5 for two sample drawings of the Backend-Aware category.

Participants that fell into this category distinguished between deletion at the UI level as opposed to data being purged from backend. No participant mentioned the risk of data being stolen nor the advantages of retention after deletion for recovery. However, this advantage was mentioned for data in the trash.

Unsurprisingly, all experts fell into the Backend-Aware category expressing varying degrees of knowledge about what exactly goes on in the background. In general, experts' knowledge around deletion surpassed all interview participants' knowledge by far.

While the understandings and drawings categorized as the UI-Based category were rather homogeneous, this was not the case for the Backend-Aware category. Since all topics mentioned by the experts fall into this more complex view, we used those six topics to further divide this category into three sub-categories, based on how many of these dimensions were included.

Two participants had a view which only included "backend" (see Figure 3). Nine participants fell into the next sub-category, which includes both concepts of a backend and time. Seven participants fell into the more complex sub-category, which includes both a backend, time, and at least one of the other dimensions. Of all the participants, only one mentioned all the dimensions, and was the only one to include the more complex concepts of derived information and anonymization.

4. DISCUSSION

Our study results revealed a plethora of reasons, views and understanding, and needs when it comes to online data deletion. In this section, we will provide some lessons learned and implications based on these results. Please note that while the results are based on two specific use cases (plus two for expiration), our recommendations go beyond these two instances.

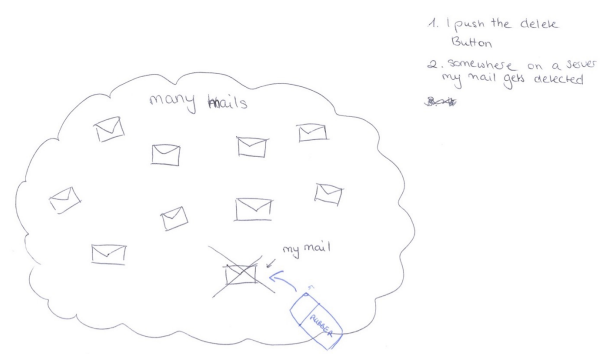
4.1 No One-Size-Fits-All Solution

As mentioned before, we used email and social media as scenarios because we hypothesized that they represent different ends of the deletion spectrum (i.e., different types of data generated in different ways). Our results show that this assumption held true. Understanding as well as views and needs for the two scenarios differed to a great extent. For instance, reasons for deletion had little to no overlap and were highly service-dependent.

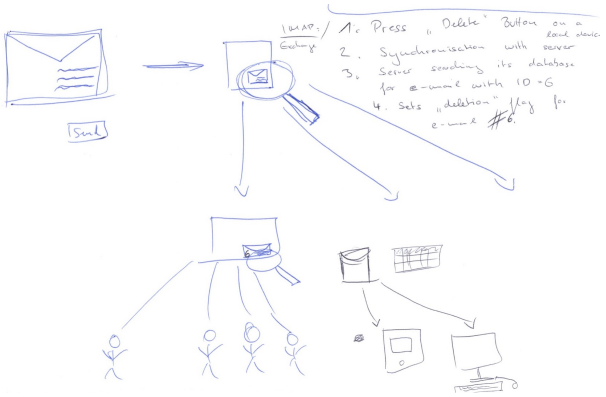
This shows that there is likely no one-size-fits-all solution when it comes to deletion strategies (from both a UI and technological point of view) which means that these individual differences need to be taken into account when designing deletion for a specific online service. For instance, understanding of (what happens during and after) deletion depends to a great extent on how a service handles its data and deletion should be reflective of this.

4.2 No Generalization of Data Deletion Needs

Related to this, we observed a great number of reasons to delete data, including privacy issues. The most prominent one (and also the only one consistent across the two scenarios) was getting rid of old or outdated data that is not needed anymore. Another interesting reason, which is related to the value of data, is deletion to tidy (or clean) an account. Participants mentioned that certain data would pollute their accounts and they wanted to get rid of this data.



(a)



(b)

Figure 5: The differing complexity of the Backend-Aware categories can be seen by comparing diagrams (a) and (b), drawn by Participants 10 and 11, respectively.

Whether data is still needed is decided on a highly individual level depending on different factors such as context, service, and usefulness, and it cannot be generalized. Even within a participant, no consistent reasons for deleting data across services (or cases) could be identified as they decide these issues on a case-to-case base.

Similar to the previous insight (no one-size-fits-all solution), a major consequence of this is that we cannot generalize user needs of online data deletion across services. For instance, while providing unlimited storage space can make many cases of deletion unnecessary in the email space, this does not translate to social networks in which publicness and embarrassment are much bigger factors.

4.3 Communicating Deletion

The study results showed that certain concepts related to online data deletion were highly present in the participants' view of deletion even though they were not necessarily correctly used from a technical point of view.

Terms and functionality related to different components in the backend (or the backend in general) were mentioned by the majority of participants. In many cases, they referred to them as the reason for increased data retention periods, i.e., the fact that data is not deleted immediately. The participants that connected these technical constraints with data retention were also more likely to find it acceptable or understandable as opposed to the participants who thought that data was solely retained for business purposes.

Other concepts were harder to understand and thus seldom part of their mindsets. One example is anonymization, which was only mentioned once.

A consequence of this is that communicating (and explaining) online data deletion using these more common concepts has the potential to positively affect users' attitude towards technological constraints of deletion. As an example, based on this, a promising direction for explaining retention periods might be to built it around technical complexity of removing it from servers, backup servers, and the like.

4.4 Deletion in the UI

Related to the previous section on communicating deletion, we think that our results can have direct influence on how deletion user interfaces are designed.

For instance, a common practice for services with a trash folder is to highlight that fact in the deletion dialogue (e.g., along the lines of "This file has been moved to the trash."). Similar to this, one could imagine that when deleting a file for good by removing it from the trash, the following procedure could be teased, again, based on parts of the process that users understand (e.g., "This file will now be deleted from our servers" to indicate technical complexity).

That said, we do not have data to judge how this should look like exactly and thus argue that this would have to be evaluated in further studies, especially with a focus on how upfront such messages would have to be to provide the best effect.

4.5 Control of Expiration

Expiration is a special use case of deletion: automatic deletion after a certain time. We worked based on the assumption that expiration for data could be represented on a timeline together with certain events that mark the end of its usefulness to a user.

However, the study results showed that this did not hold true for any of the scenarios. While there are single instances (or single participants) that could identify such an event, it was highly context-dependent. In addition, for each data item (and scenario), participants could identify events or situations which would give new value to information that was previously marked as useless.

This indicates that enforcing specific expiration periods on undeleted user data is likely to create situations in which useful (or wanted) data is not available anymore. A potential consequence of this would be a reduction in service quality from a user's point of view.

Participants indicated that control, especially self-selected expiration conditions, would be a better way to approach this issue. One participant proposed the following approach for email deletion: *"You could have a folder which allows you to set an expiration date for items in this folder. Like when I move an email in there, it could be automatically deleted after 30 days or whatever amount I decide."*

This type of control mechanism highlights another result of this research: users can relate even abstract concepts very well to the UI. Therefore, we can leverage this to communicate with users through well-known concepts and metaphors, such as the "trash can".

Summed up, this means that, instead of automatic (default) data expiration, allowing control over how data expiration is handled on an individual level is a more promising direction. This would also give users more control over their data (preferences).

4.6 Shared/Implicit Copies Not Well Understood

Participants understood the concept of shared copies for emails rather well. It is easy to comprehend that when you send out an email, a copy of it will exist with the recipients. As a consequence, they pay more attention to what they write due to the fact that control about the data will be lost [6].

As opposed to this, for social networks, this problem was not well understood, despite it being similarly likely as shown by related work on Twitter and Facebook [1]. For instance, only 3 out of 22 participants mentioned that data might be stored with a recipient (or the like). Thus, the idea of (not necessarily verbatim) copies based on retweets and other ways of interacting with a post seems to be less graspable. This is even worse as implicit copies can be a challenge to the user privacy as the user loses control over the content but might not even be aware of the existence of the copies.

While our data provides further insights into this being an issue that potentially affects user privacy, we do not have data to make recommendations on how to mitigate this risk. However, we argue that this is an important topic for the research community to study and want to highlight its necessity.

5. LIMITATIONS

The main limitation of this work is the limited sample size of the interview study. While we made sure to recruit participants from a wide spectrum of society, the data should not be interpreted as generalizable to the whole (internet) population but rather as trends. That said, we are confident that we cover the most relevant themes, which is supported by the fact that we reached saturation of themes after (max) 14 participants.

Furthermore, despite carefully selecting the scenarios to cover a wide range, the results are limited to the two tested contexts (plus two for expiration). As mentioned in the discussion, results might have been different had we tested other services (e.g., cloud storage), and thus, recommendations in this paper should be handled with care in these contexts. Since the selected scenarios cover different ends of the deletion spectrum, we argue that the major insights of this work are still (partially) applicable to online deletion overall.

6. CONCLUSION

In the present work, we explored users' understanding of online data deletion which is essential to maintaining user privacy and protecting their data. We identified two main views on how deletion works: UI-Based and Backend-Aware. We found that a large majority of participants were aware of a backend to the deletion process. Although participants' understanding of the backend processes of deletion varied in their complexity, explanations of online data deletion can build off of this understanding to explain the technical constraints of deletion in conceptual terms. Our results indicate that doing so could also have the potential to positively affect users' attitudes toward these constraints and be more accepting of certain retention periods.

Our results also provide insights into expiration preferences. We found that participants considered the usefulness of their online data to be very context-dependent, as opposed to time bound. Consequently, participants did not envision their online data having an expiration date that could be set on a chronological scale. Participants therefore favored control over the expiration of their data, such as moving data to a specific folder where they can manually set expiration dates.

A challenge raised by this work relates to user understanding of shared copies of online data for services where it is not well understood and can be problematic in terms of privacy. While the concept of a shared copy is clear for email (i.e., the recipient has a copy), it is not so clear in the social media contexts, where different ways of interacting with the data could lead to different copies (e.g., re-posts) or traces of it (e.g., comments referencing a post). Future work should explore these understandings, and how to best communicate to users this concept of shared copies in complex settings.

7. ACKNOWLEDGMENTS

First of all, we would like to thank our interview and focus group study participants for their valuable time. Furthermore, we are very grateful for the input provided by both the internal Google reviewers as well as the external reviewers whose input significantly helped to improve this work and this paper.

8. REFERENCES

[1] H. Almuhimedi, S. Wilson, B. Liu, N. Sadeh, and A. Acquisti. Tweets are forever: A large-scale quantitative analysis of deleted tweets. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 897–908, New York, NY, USA, 2013. ACM.

[2] O. Ayalon and E. Toch. Retrospective privacy: Managing longitudinal privacy in online social networks. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, pages 4:1–4:13, New York, NY, USA, 2013. ACM.

[3] L. Bauer, L. F. Cranor, S. Komanduri, M. L. Mazurek, M. K. Reiter, M. Sleeper, and B. Ur. The post anachronism: The temporal dimension of facebook privacy. In *Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society*, WPES '13, pages 1–12, New York, NY, USA, 2013. ACM.

[4] M. Bishop, E. R. Butler, K. Butler, C. Gates, and S. Greenspan. Forgive and forget: Return to obscurity. In *Proceedings of the 2013 New Security Paradigms Workshop*, NSPW '13, pages 1–10, New York, NY, USA, 2013. ACM.

[5] L. J. Camp. Mental models of privacy and security. *IEEE Technology and Society Magazine*, 28(3):37–46, Fall 2009.

[6] A. De Luca, S. Das, M. Ortlieb, I. Ion, and B. Laurie. Expert and non-expert attitudes towards (secure) instant messaging. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 147–157, Denver, CO, 2016. USENIX Association.

[7] K. A. Ericsson and H. A. Simon. Verbal reports as data. *Psychological review*, 87(3):215, 1980.

[8] R. Kang, L. Dabbish, N. Fruchter, and S. Kiesler. "my data just goes everywhere:" user mental models of the internet and implications for privacy and security. In *Symposium on Usable Privacy and Security (SOUPS)*, pages 39–52. USENIX Association Berkeley, CA, 2015.

[9] V. Mayer-Schönberger. *Delete: The virtue of forgetting in the digital age*. Princeton University Press, 2011.

[10] M. Mondal, J. Messias, S. Ghosh, K. P. Gummadi, and A. Kate. Forgetting in social media: Understanding and controlling longitudinal exposure of socially shared data. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 287–299, Denver, CO, 2016. USENIX Association.

[11] E. S. Poole, M. Chetty, R. E. Grinter, and W. K. Edwards.

More than meets the eye: Transforming the user experience of home network management. In *Proceedings of the 7th ACM Conference on Designing Interactive Systems*, DIS '08, pages 455–464, New York, NY, USA, 2008. ACM.

[12] K. M. Ramokapane, A. Rashid, and J. M. Such. "i feel stupid i can't delete...": A study of users' cloud deletion practices and coping strategies. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 241–256, Santa Clara, CA, 2017. USENIX Association.

[13] J. Rosen. The web means the end of forgetting, 2010.

[14] M. Sleeper, J. Cranshaw, P. G. Kelley, B. Ur, A. Acquisti, L. F. Cranor, and N. Sadeh. "i read my twitter the next morning and was astonished": A conversational perspective on twitter regrets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 3277–3286, New York, NY, USA, 2013. ACM.

[15] M. Tohidi, W. Buxton, R. Baecker, and A. Sellen. User sketches: A quick, inexpensive, and effective way to elicit more reflective user feedback. In *Proceedings of the 4th Nordic Conference on Human-computer Interaction: Changing Roles*, NordiCHI '06, pages 105–114, New York, NY, USA, 2006. ACM.

[16] Y. Wang, P. G. Leon, A. Acquisti, L. F. Cranor, A. Forget, and N. Sadeh. A field trial of privacy nudges for facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 2367–2376, New York, NY, USA, 2014. ACM.

[17] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, and L. F. Cranor. "i regretted the minute i pressed share": A qualitative study of regrets on facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, SOUPS '11, pages 10:1–10:16, New York, NY, USA, 2011. ACM.

[18] R. Wash. Folk models of home computer security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, pages 11:1–11:16, New York, NY, USA, 2010. ACM.

APPENDIX

The following two sections list the study instruments used for the deletion scenarios and the expiration exercise. Please note that: a) They have been slightly adapted (e.g., the scenarios script actually consists of 3 parts that we merged for the appendix); b) They are listed out of context.

A. DELETION SCENARIOS SCRIPT

1. Do you sometimes delete [emails/tweets/posts]?
   (a) If yes: Why?
   (b) If no: Why not?
2. Now let's imagine you just deleted an [email/tweet/post]. Please draw what happens after you press the "Delete" button. (instruction: hand participant pen and paper)
3. You just named a few things that occur when you delete an [email/tweet/post]. Please write them down as a list in the order in which they occur. Use "Press the delete button" as the first item on your list.
4. Imagine that you pressed the delete button now. When would the last item on your list take place?
5. After this last point on the list, is it possible for you to recover the deleted [email/tweet/post]?
   (a) If yes: Why?
   (b) If no: Why not?
6. Is it possible for the [service provider] to recover the deleted [email/tweet/post]?
   (a) If yes: Why? For what purpose is the data stored?
   (b) If no: Why not?

B. EXPIRATION GRAPH SCRIPT

1. Here is a card with an online context, and a type of personal data associated with that context written on it. (instruction: hand participant one of the cards in counterbalanced order)
2. Here is a screenshot of what this online context would look like. (instruction: hand participant screenshot, read description)
3. On a scale from 1-5, with 1 being the least sensitive and 5 being the most sensitive, how sensitive is this type of data to you? Please write your rating on the card.
4. Now we will be referring to this graph. (instruction: hand participant the expiration graph)
5. On the horizontal axis, please add different events which can occur in this scenario that could have an influence on the usefulness of this data. Usefulness refers to how useful it is for you that the service provider has this data.
6. The usefulness might change over time. Let me give you an example: It is most useful for your dentist to know the time of your appointment before it happens, and still quite useful on the day of the appointment. After the day of the appointment, it is perhaps less useful that your dentist has this information.
7. After the point when this data is no longer useful to you, what should happen to it, if anything? (instruction: if participant added an event with a usefulness rating of 0/1, refer to that point)