Helping Users Automatically Find and Manage Sensitive, Expendable Files in Cloud Storage

Mohammad Taha Khan[†△], Christopher Tran[†], Shubham Singh[†], Dimitri Vasilkov[★],
Chris Kanich[†], Blase Ur[★], Elena Zheleva[†]
† *University of Illinois at Chicago,* △ *Washington & Lee University,* ★ *University of Chicago*

Abstract

With the ubiquity of data breaches, forgotten-about files stored in the cloud create latent privacy risks. We take a holistic approach to help users identify sensitive, unwanted files in cloud storage. We first conducted 17 qualitative interviews to characterize factors that make humans perceive a file as sensitive, useful, and worthy of either protection or deletion. Building on our findings, we conducted a primarily quantitative online study. We showed 108 long-term users of Google Drive or Dropbox a selection of files from their accounts. They labeled and explained these files' sensitivity, usefulness, and desired management (whether they wanted to keep, delete, or protect them). For each file, we collected many metadata and content features, building a training dataset of 3,525 labeled files. We then built Aletheia, which predicts a file's perceived sensitivity and usefulness, as well as its desired management. Aletheia improves over state-of-the-art baselines by 26% to 159%, predicting users' desired file-management decisions with 79% accuracy. Notably, predicting subjective perceptions of usefulness and sensitivity led to a 10% absolute accuracy improvement in predicting desired file-management decisions. Aletheia's performance validates a human-centric approach to feature selection when using inference techniques on subjective security-related tasks. It also improves upon the state of the art in minimizing the attack surface of cloud accounts.

1 Introduction

Since the introduction of Dropbox in 2007, cloud storage has become a convenient and affordable way to retain files over time and sync files across multiple devices with minimal user effort. However, with the passage of time, some files lose their relevance. Crucially, some files that are no longer useful may still contain sensitive information, creating risks due to data breaches, lost devices, and account takeovers [5, 53, 58].

The free versions of consumer cloud services provide gigabytes of storage, which is more than enough for thousands of documents and media files to pile up over the years. While making indefinite retention of files the default option has freed users from the risks of lost USB sticks and crashed hard drives, this policy also causes potentially sensitive information to accumulate in a single place. While this agglomeration of sensitive data is risky, manual management is far too time consuming and tedious to be practical. Thus, some form of automated assistance is quickly becoming necessary.

While researchers have characterized this need for retrospective management of consumer cloud archives [24, 49], they did not propose any concrete techniques for identifying which files users should revisit. Likewise, although Microsoft recently added a "Personal Vault" [41] to the OneDrive cloud platform that adds 2FA protection to a specific folder, deciding which files to put in such a folder is a manual process.

Because revisiting thousands of files that have accumulated over many years is time consuming, the foundation of any practical protection approach must be some form of automated inference. Even so, the subjective and human-centered nature of file management requires understanding what makes a file in the cloud sensitive, as well as what makes it expendable. Information rights management (*IRM*) [31] and data-loss prevention (*DLP*) [7, 18] have superficially similar goals of preventing the unwanted disclosure of information, though we hypothesized that identifying sensitive and useless files would differ between corporate and consumer domains. For instance, whereas industry focuses on identifiers (e.g., account numbers), consumers might also consider files sensitive if they cast them in a negative light or violate their self-presentation. Our results validated this hypothesis.

In this paper, we present a multi-part approach to developing an automated inference pipeline that predicts the perceived *sensitivity* and *usefulness* of files stored in the cloud. Due to the highly subjective nature of sensitivity and usefulness, as well as the incomplete understanding provided by prior work, we first explore users' mental models of these concepts qualitatively. With the goal of enumerating the many ways different people might consider a file sensitive or useful, we conducted 17 interviews. We found that participants considered files sensitive for objective reasons like the presence

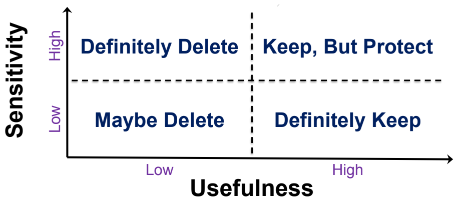


Figure 1: The file-management decisions we envision for files in the cloud based on their sensitivity and usefulness.

of financial data or personally identifiable information, as well as subjective reasons like the presence of content considered intimate in the participant's unique context. Participants considered files useful not only based on the recency of file access, but also based on sentimentality and relationships.

Subsequently, we acted on this holistic understanding by constructing and evaluating classifiers through primarily quantitative user studies. A key challenge is that sensitive files, our primary target, are very much a minority class within cloud archives. As a result, we conducted two rounds of online user studies. In each round, we showed participants dozens of files from their own Google Drive or Dropbox accounts, asking them to rate (and explain) the sensitivity and usefulness of each file. We collected numerous metadata and content features for each file, as well as for the cloud account overall.

In *Round 1* we showed 75 participants files selected from their account using heuristics inspired by our interviews. We trained a *preliminary classifier* using the data collected. To further mitigate the class imbalance for file sensitivity, in *Round 2* we showed 33 additional participants files selected based on our preliminary classifier. With the combined data, we trained and evaluated a final classifier, which we dub *Aletheia*.[1]

Aletheia performs three prediction tasks. It predicts whether a user will perceive a given file as (i) sensitive and (ii) no longer useful. Finally, it predicts (iii) a file-management decision specifying whether the file should be kept, deleted, or protected (e.g., requiring 2FA). Based on prior work [24], we hypothesized that users' file-management decisions would roughly correlate with file sensitivity and usefulness as shown in Figure 1. We compared Aletheia against typical baselines of assigning labels based on the majority class and randomly. In addition, for predicting file sensitivity, we also compared against a baseline that uses the output of Google's Data Loss Prevention API (*GDLP*, Section 2.2). For predicting file usefulness, we used the time since the file was last modified as an additional baseline. To the best of our knowledge, no prior work examines subjective classification of file usefulness and sensitivity, so we selected these heuristic baselines to reflect features that correlate most intuitively with those perceptions.

Aletheia is novel in its aim to identify files in consumer cloud accounts that users *perceive* as sensitive and no longer useful. Aletheia augments file access patterns and the objective identifiers (e.g., Social Security Numbers) captured by industry IRM [31] and DLP [7, 18] tools with numerous file metadata and content features that capture subjective characteristics our formative qualitative work found to be associated with human perceptions of a file as sensitive or useful.

We found that Aletheia substantially improves the state of the art for identifying files in the cloud that users are likely to perceive as sensitive and no longer useful. For predicting sensitivity, Aletheia showed an AUC improvement of 68% for documents and 153% for images over a classifier that used only GDLP features. Predicting files as no longer useful, Aletheia's AUC improvement was 26% for documents and 101% for images when compared to the *last modified* classifier. Predicting participants' desired file management, the accuracy improvement was 49% over the most sensible baseline, a majority label classifier.

In developing Aletheia, we made three key **contributions**:

- We performed 17 qualitative interviews to understand factors that lead a human to perceive a file in their cloud storage account as sensitive and no longer useful.
- We conducted online user studies of 108 users of Google Drive or Dropbox. Our quantitative and qualitative results characterize the relationship between desired file-management decisions and perceptions of file sensitivity and usefulness. While most participants nearly always elected to delete files they deemed not useful, some still preferred to retain files deemed not useful.
- We constructed classifiers that automatically identify files likely to be perceived as sensitive and not useful. These classifiers improve over state-of-the-art baselines by 26% to 159%. They can also identify users' desired file-management decisions with 79% accuracy. We further unpacked Aletheia's performance, analyzing mispredictions relative to participants' qualitative responses.

2 Related Work

Previous scholarship related to this project spans multiple subareas, including those that characterize the risks and harms of online data, user-focused understanding of data sensitivity, personal information management, and the design of automated tools to infer and manage online privacy preferences.

2.1 Risks of Online Data and Cloud Storage

The possibility of being harmed by data breaches is very real. Governments [11, 38] and academic researchers [14, 30, 34] have highlighted risks and damages potentially caused by malicious actors stealing users' private data. Researchers have evaluated how individuals perceive online privacy [21, 25], identifying associated risks [35] and the magnitude of those risks in various scenarios [10, 48, 57]. Entrusting data to a third party can increase the risk of data breaches, which have

[1] Aletheia is the Greek word for truth, which through the privative alpha literally means "un-forgetfulness" or " un-concealment."

become common [5, 53, 58, 58]. User-centered management of data retention has become a key part of online security.

Researchers have evaluated the risks of data breaches on multiple cloud storage systems [22]. Studies have focused on latent danger in the cloud in comparison to local storage [23], user perceptions of the inadvertent storage of sensitive data [9], and strategies for minimizing risk [32]. Complementary work has evaluated file-management practices [52, 56]. More recent user research has found that most users have forgotten-about data they wish to delete stored in the cloud [24]. Researchers have also studied users' mental models of cloud storage [4] and data retention [36, 45, 55]. We build upon these insights through user studies and classifier construction to automatically identify risky files in the cloud.

2.2 Data Sensitivity and Data-Leak Detection

Data sensitivity is subjective, and there is no universal definition of the term. There have been extensive efforts in the community to detect and quantify potentially sensitive data in various contexts. Peddinti et al. used anonymous Quora posts to understand the sensitivity of questions in a Q&A forum [42]. They found that, in addition to expected topics like religion, sex, and drugs, questions about emotions, relationships, and careers were also seen as sensitive. Researchers have also developed initial methods to detect potential nudity [47] and violence [37] in image and video files. The presence of nudity or violence commonly suggests that the data is sensitive. While we built on their understanding of sensitive questions, our investigation of files — including documents, images, and other media — required a far broader understanding of sensitivity.

Recent efforts in industry have codified IRM and DLP methods for preventing data leaks [7, 18, 31]. While the goals of these efforts — preventing unwanted disclosures of information — are superficially similar to ours, the characteristics associated with file sensitivity and usefulness differ between corporate and consumer domains. Industry approaches focus on identifiers (e.g., bank account numbers). For instance, Google's Cloud Data Loss Prevention API [18] aims to classify and redact sensitive information from documents and is primarily marketed to organizations. The API categorizes numerous personal and financial identifiers as sensitive, partially through regex matching. Additional industry IRM solutions [5, 31] use both regex matching and access-management frameworks to tag sensitive data. However, such efforts are most applicable for information sharing within an organization. Critically, they do not incorporate subjective perceptions of file sensitivity or usefulness into the decision process. For instance, consumers might also consider a file sensitive if it violates their intended presentation of self, as might be the case for ill-advised poetry or embarrassing photographs. While we make use of Google's Cloud Data Loss Prevention API as a baseline for comparison and as one source of features for Aletheia, our approach takes a much broader view of sensitivity. We began with a qualitative study to characterize perceptions of data sensitivity, incorporating this understanding into subsequent parts of our project. Augmenting the Google Cloud Data Loss Prevention API features with others that capture subjective characteristics substantially improved classification accuracy (Section 7).

2.3 Automated Management of Privacy

The large quantity of forgotten-about data in the cloud requires semi-automated inference to help users revisit potentially sensitive files [24]. In the same spirit, researchers have proposed techniques for automated management of privacy settings. For social networks, Fang and LeFevre proposed a "privacy wizard" for automatic inference of privacy settings [12], while Ghazinour et al. used collaborative filtering to recommend privacy settings [16]. There have also been efforts to build classifiers around user-level privacy scores [29] and privacy risk [60]. Similar research focuses on inferring sensitive attributes and identity matching in online platforms [15, 17, 26, 28, 59]. Some researchers have also used classifiers to predict desired permissions for image files [51] and whether content should be private or public [13, 50]. To our knowledge, we are the first to develop a classifier for automated management of files in cloud storage, especially a classifier based on sensitivity ratings of private files collected from the owners of those files in user studies, as opposed to a third party rating publicly available files.

3 Approach

This section summarizes our approach (Figure 2) to automatically helping users find sensitive and unwanted files in the cloud. We elaborate on our process and high-level goals. We also explain how we dealt with the associated challenges.

1. Understanding Sensitivity and Usefulness: For files in the cloud, terms like sensitivity and usefulness can have subjective interpretations that vary across individuals. With the goal of enumerating the variety of these perceptions, we first conducted qualitative interviews. These interviews were conducted as open discussions to encourage individuals to highlight all possible file attributes associated with sensitivity and usefulness. Subsequently, we mapped these attributes to quantitative file features that can be collected programmatically. These interviews also influenced the design of our quantitative survey. Section 4 details these interviews.

2. Training Data Collection and Augmentation: A prerequisite for developing an automated classifier is collecting training features and labels. To this end, we performed a quantitative study of 108 long-term users of Google Drive and Dropbox. The study combined a user survey with automated collection of various features about participants' cloud accounts and files. These features included metadata provided

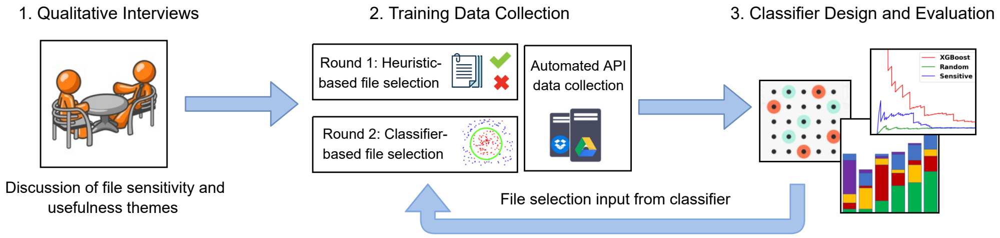

1. Qualitative Interviews
2. Training Data Collection
3. Classifier Design and Evaluation
Round 1: Heuristic-based file selection
Round 2: Classifier-based file selection
Automated API data collection
XGBoost
Random
Sensitive
Discussion of file sensitivity and usefulness themes
File selection input from classifier

Figure 2: Overview of our approach combining qualitative interviews and two rounds of quantitative data collection.

by cloud storage providers, as well as deeper content analysis using third-party services like Google Cloud Vision.

The survey centered on showing participants files from their Google Drive or Dropbox accounts and asking them to label and explain their sensitivity and usefulness. We also asked them to indicate a file-management decision: whether they would want to keep, delete, or protect each file. As it is not feasible to show a participant all files on their account, selecting the right subset of files to yield a well-suited distribution of training data was a challenge. To solve this, we conducted two rounds of data collection. In Round 1, we primarily used heuristic-based file selection leveraging insights from our interviews. Because only a handful of files on a typical account are sensitive, heuristic-based file selection yielded a small number of sensitive data points. Therefore, we trained a preliminary classifier, using its predictions for sensitive documents and images to select files in Round 2. Doing so let us oversample the minority class (sensitive files). Section 5 further details our method, while Section 6 summarizes the findings from both rounds of data collection.

3. Developing Aletheia, an Automated Classifier: Using the data collected from both rounds, we built classifiers to predict file (i) sensitivity, (ii) usefulness, and (iii) desired management. We formulated each prediction as a classification task. Note that file-management decisions are heavily influenced by file sensitivity and usefulness. As mentioned above, we used an initial version of the sensitivity classifier for Round 2 of data collection. Because decisions to delete data are highly subjective and consequential, we expect Aletheia to be used as part of a human-in-the-loop support system, rather than in a fully automated way. Therefore, we evaluated Aletheia with precision-recall analysis, which aligns with rankings of which files to present in a user interface or through recommendations. To quantify the accuracy of our models, we used the area under the precision-recall curve (AUC). Section 7 details Aletheia's experimental setup and performance results.

4 Qualitative Interviews

To gain an initial understanding of how people conceive of the sensitivity and usefulness of files in the cloud, we first conducted semi-structured interviews of cloud storage users. We aimed to build a formative understanding of factors that make someone perceive a file as sensitive or useful. This understanding underpins our online study, eventually enabling us to find files that may be sensitive, yet not useful, at scale.

Scenarios for Sensitivity
1. Files that would cause concern if they were hacked from the cloud
2. Cloud files that, if made public, would be embarrassing
3. Files that would cause worry if close family members viewed them
Scenarios for Usefulness
1. Files to be recovered if they were accidentally deleted from the cloud
2. Cloud files accessed and updated on a regular basis
3. Cloud files shared with friends and/or family

Table 1: Broad scenarios used as prompts in our interviews.

4.1 Methodology

Using Craigslist, we recruited participants who had a Google Drive or a Dropbox account over 3 months old and were willing to attend an in-person interview. We interviewed 17 participants from January through June 2019. Among participants, 10 identified as male and 7 as female. Their ages ranged from 20 to 45 years old. We prioritized participants without experience in an IT-related field. Six participants were full-time students, all from non-STEM majors. All other participants had completed a college education. The interview took approximately 30 minutes to complete, and compensation was a $20 Amazon gift card. This amount also accounted for the costs of participants commuting to the interview site.

Our protocol investigated participants' approaches to cloud storage both abstractly and concretely, where the latter was grounded in individual files in a participant's account. Appendix A in our online materials [1] contains our script.

The first half of the interview focused on general reasons for using cloud storage, followed by an open-ended discussion about broad classes and characteristics of sensitive and useful files stored in the cloud. To further spur participants' thinking, we also provided them the sensitivity and usefulness scenarios in Table 1. These specific scenarios were the research team's initial hypotheses about how sensitivity and usefulness manifest. Considering responses to both our broad questions and discussions following the scenario prompts, we began to conceptualize sensitive and useful files.

The second half of the interview investigated the same phenomena more concretely. Participants logged into a web app we built that used the Google Drive and Dropbox APIs to show ten files randomly selected from their account. For each file, the participant explained its sensitivity and usefulness, giving us concrete examples of files that were sensitive or useful, in addition to specific attributes that made them so. After the questions about specific files, participants were asked to provide overall feedback regarding draft questions from our quantitative survey (Section 5). These specifically focused on ways to elicit perceptions of file sensitivity and usefulness.

All interview responses were audio recorded with consent and then transcribed using the Google Speech to Text API [19]. One member of the research team open-coded these transcriptions to extract emergent themes. A second member of the team then independently coded the extracted quotes using that codebook. Cohen's κ, a measure of intercoder reliability, was 0.87. The two coders met and resolved conflicting codes. The final codebook, which is available in our online materials [2], contained thirty distinct codes across the sixteen prompts and questions.

We took care to ensure interviews were conducted ethically. We first obtained IRB approval for our protocol. Participants reviewed a consent form with opt-in permission for audio recording. Furthermore, to ensure participant privacy, we encouraged them to use their own personal device (computer or phone) to view the files selected for the study, though we also gave them the option of using a laptop we provided. During the part of the interview where they reviewed their own files, we instructed them to sit so that the contents of their screens were visible only to them.

Like all user studies, our protocol has limitations. One potential limitation was that we presented a fixed set of categories (Table 1) representing potential manifestations of file sensitivity and usefulness. While we intentionally provided these prompts only after a broad initial discussion of file sensitivity and usefulness, they may not have captured all possible conceptualizations of these ideas. Particularly, our prompts on sensitivity did not always align with the nuanced potential risks of a file being leaked. To minimize these biases, we provided these prompts only after participants gave us their initial open-ended thoughts about types and characteristics of sensitive and useful files. However, this approach may have discouraged participants from mentioning other categories of sensitivity and usefulness we did not anticipate. Additionally, as these were in-person interviews, we were also limited to individuals who were residents of a North American urban center, and participants represented a convenience sample of both students and members of the workforce. As a result, our formative understanding of file sensitivity and usefulness is likely to be situated in a particular culture and demographic.

4.2 Results

We now present interview participants' conceptions of the sensitivity and usefulness of files in the cloud.

4.2.1 Why a File Might Be Perceived as Sensitive

In our general discussions of what makes a file sensitive, participants invoked the following seven classes of sensitivity:

Personally Identifiable Information (PII): Files that contained names, contact details, dates of birth, passports, or driver's licenses were considered sensitive. Many participants cited their resume as an example. P01 explained, *"Anything that can easily identify you, like your name, your birthday, your phone number, your address. It's all on my resume."*

Confidential Information: Distinct from PII, participants mentioned that some data should never be released publicly because of its proprietary or confidential nature. Students mentioned original work that could be plagiarized. P05 said, *"If it's like an essay or something that I'm turning in, I don't think I necessarily want a bunch of people to read it."* Three participants also mentioned files containing passwords.

Financial Information: Participants mentioned tax documents, pay stubs, and files with Social Security Numbers (SSN) as very sensitive. They also worried about statements for bank accounts and credit/debit cards, as well as other documents containing those numbers. Nine participants explicitly mentioned their SSN as particularly sensitive, yet also found on their cloud accounts due to backups of files like tax returns.

Intimate Content: Participants described broad conceptions of content that could be considered intimate or personal, and thus sensitive. Photos, videos, and similar media files were most commonly mentioned, particularly individuals' own photos (both in adult situations and in general), as well as adult content they had downloaded. P16 included among their embarrassing files *"porn, anything that's not for the public's eyes. Pictures of myself or significant other."*

Personal Views: Files that contained personal views or opinions were also identified as sensitive. P09 explained, *"I'm a religious person and so there are times when I would make audio recordings or save videos that are of a religious nature. People may not particularly subscribe to it, or some people may deem it offensive."* Participants also mentioned files that contain political opinions and anti-government views.

Self-Presentation: Participants found files related to their self-presentation as sensitive. For example, P11 talked about *"unflattering photos and videos."* Other participants said files that revealed activities they hoped to hide from specific people were sensitive. For example, P14 said, *"If there was a photo of me smoking weed, my parents would freak out."*

Content That May Be Misinterpreted: Participants also said files that could be misconstrued by others were sensitive. A participant who was in the military discussed a specific picture they saw during the study by explaining, *"This is a picture of some of my soldiers at a cemetery. Even though it's innocent, I don't want people to associate this with, like, death."* In contrast to data like financial documents, this type of sensitivity is particularly contextual and subjective.

4.2.2 Why a File Might Be Perceived as Useful

Participants most commonly considered files in the cloud useful if they might need to access them in the future. The specific reasons for this future access spanned five categories:

Reminiscence: Participants frequently invoked photos' sentimental nature and value for reminiscence as a key reason they are useful. P09 explained, *"Pictures are useful because they capture memories. You want to have some memory of good times, good events, or different things."* P16 explained why a specific picture of her kids was useful by saying, *"I would show my children what they looked like when they were younger."* Expanding this definition, P09 explained, *"I share photos and videos of deceased family members that we like to reminisce about."* Broadly, participants explained that files with sentimental value will likely remain useful forever.

Active Projects: Participants explained that files related to projects at work or school were useful, but many would not remain useful indefinitely. When asked to think about files they would prioritize recovering if accidentally deleted from the cloud, 13 participants mentioned work- or school-related files. For example, P12 said, *"I would try to recover my resume and any school work that needed to be turned in."* Similarly, P09 said, *"Documents are useful because… you always have to deal with documents online in school, at work."*

Recent Files For Reference: Some documents remained useful for reasons other than their initial purpose. For example, P04 described a recent cover letter being useful for future job applications to additional employers by saying, *"This version is very current. I just recently updated it, so it will be very useful for me."* In general, participants said files that had been recently accessed or modified were more likely to be useful, yet some older files might also be needed for reference.

Files Frequently Updated Over Time: Participants said cloud files that are frequently modified are useful. While some work- or school-related files fell into this category, journals and other evolving documents were key examples.

Sharing: Five participants mentioned that shared files were useful. For example, P03 (a student) explained: *"Midterm or final papers I usually store in the cloud if I need to share them with somebody else or have someone else look at it."*

We used this qualitative understanding both to develop closed-form survey questions (Section 5.1) and to identify metadata and content features to collect about the files in participants' cloud accounts to train our classifiers. Section 5.2 lists these features and their relationship to these findings.

5 Quantitative Online User Study: Method

Building on the insights from our qualitative interviews, we conducted an online user study combining a survey and automated data collection from participants' cloud accounts. Our core goal was to collect rich data about participants' perceptions alongside quantitative features of files in the cloud to train an automated tool for aiding cloud file management. Appendix B, online [1], contains the survey instrument.

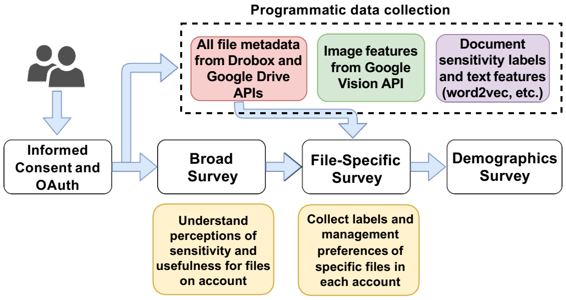

Programmatic data collection
All file metadata from Drobox and Google Drive APIs
Image features from Google Vision API
Document sensitivity labels and text features (word2vec, etc.)
Informed Consent and OAuth
Broad Survey
File-Specific Survey
Demographics Survey
Understand perceptions of sensitivity and usefulness for files on account
Collect labels and management preferences of specific files in each account

Figure 3: Overview of the survey and data-collection process.

We first built a tool that allows us to survey participants about specific files in their cloud storage account while simultaneously collecting metadata and content-based features about those files. We collected data across two rounds. For each round, we recruited a separate set of participants to complete both the generic and file-specific surveys described below. In Round 1, we used a heuristic-based approach to select files. From the results of Round 1, we trained a preliminary classifier, which we used to select files in Round 2. We used data from both rounds to build and evaluate Aletheia.

5.1 Study Overview and Survey Structure

We recruited participants on Amazon's Mechanical Turk (Mturk) and Prolific Academic.[2] We recruited American participants age 18+ with a platform approval rating of 95%+. Participants were also required to have a Google Drive or Dropbox account that was 3+ months old with 100+ files.

We first presented participants with a consent form and a visualization of the data we would collect from their cloud account. Afterwards, we asked participants to authorize our tool to programmatically scan their account. Figure 3 summarizes the overall study flow and back-end data collection. The survey contained three sections: (i) broad questions about their use of cloud storage; (ii) file-specific questions about the sensitivity and usefulness of particular files on their account; and (iii) questions about their demographics and the protection mechanisms used to secure their accounts.

File-Specific Survey: The focus of our survey was its second part, in which we queried participants about particular files stored in their accounts. Participants' responses, paired with the file features we collected, formed the training data for Aletheia. As shown in Table 2, our file-selection strategies differed across two rounds of data collection.

Round 1: We first selected files with heuristics defining different categories of files. For category #1, we looked for the presence of sensitive keywords in the filename. We chose

[2] While we initially used Mturk for data collection, we found Prolific more successful for recruitment as it is designed for academic user studies.

Category	# of Files	File Description
Round 1 (File selection based on heuristics)		
1	5	Files containing a sensitive keyword in file name
2	8	Document files (.txt, .docx, .pdf, .xlsx, .ppt, etc.)
3	8	Media files (.jpg, .png, .mp4, .mpeg, etc.)
4	4	Files other than documents or media
Round 2 (File selection based on preliminary classifier)		
5	25	Top sensitive documents
6	25	Top sensitive images

Table 2: File-selection categories for the quantitative survey.

keywords (e.g., "resume," "passport," "tax") based on our interviews. The other three categories were documents (#2), media files (#3), and additional files (#4). We chose this diversified approach to file selection to capture a variety of file types, particularly those that our qualitative interviews suggested were potentially sensitive. We showed participants these files in randomized order. In comparison to a purely random selection, this approach provided a broader perspective, especially for accounts with a skewed file distribution (e.g., one with 10 documents and 500 images).

Round 2: We used the data from Round 1 to train a preliminary classifier for identifying sensitive documents and images. Because sensitive files are a clear minority class (most files are not sensitive), in Round 2 we used this classifier to select only potentially sensitive documents and images. We also doubled the number of selected files to 50. In particular, we ranked documents (#5) and all images (#6) based on their predicted sensitivity score. We selected the top 25 images and 25 documents, showing them in randomized order.

For each file shown in either round, participants rated their agreement (on a five-point Likert scale) that *"I consider this file worth keeping,"* which was the proxy we developed for usefulness based through our qualitative interviews. Similarly, agreement that *"it would be risky, harmful, or otherwise dangerous if this file were accessed without my consent"* was our proxy for sensitivity. Because our eventual goal was to train binary classifiers for finding files that are not useful, yet sensitive, we aggregated "strongly disagree" and "disagree" responses to the former statement as *not useful* and "strongly agree" and "agree" responses to the latter as *sensitive*.

We also asked participants to choose how they desired to manage the file from among the following three options:

- **Keep as-is:** The file will remain in your cloud storage account in its current state.
- **Delete:** The file will be removed from your cloud storage account.
- **Protect:** The file will remain in your cloud storage account. However, you will need to take extra security steps to access the contents of the file.

Aletheia (Section 7) aims to predict the answers to the three dimensions above. To better diagnose incorrect predictions, we also asked participants to justify each answer in free text.

5.2 File Feature Collection

Table 3 lists the features we collected. We chose these features primarily based on insights from the qualitative interviews. Because many interview participants mentioned personal and financial identifiers as sensitive, we used the Google Cloud Data Loss Prevention (*GDLP*) API [18] to find such identifiers in files. Likewise, because interview participants mentioned concerns about specific types of images, we used the Google Vision API to collect image object labels and binary labels corresponding to the presence or absence of adult, racy, medical, and spoofed content within images. For documents, we performed local text processing to extract features including TF-IDF vectors, topic models, word2vec vectors, and bags of words. Finally, we collected metadata about file activity and sharing. Section 7 details how Aletheia uses these features.

5.3 Ethics

We obtained IRB approval prior to data collection. We took additional steps to protect participant privacy and ensure informed, affirmative consent. Our consent page provided textual and visual examples (shown in online Appendix B [1]) of the type of data we collected about participants' files. In addition, to further address privacy-related concerns, we provided participants with a link to our privacy policy, which comprehensively detailed our data-collection process and how data was stored and used during the research process. Participants were also provided with the contact information for the IRB office and the researchers themselves. Our web apps were reviewed and verified by Google Drive and Dropbox, and our OAuth scopes were set precisely to those required for the survey. We did not retain any personally identifiable information, and we only stored high-level labels, counts, features, and similarity-based hashes. We also guided participants on revoking access to our tool following completion of the study.

6 Quantitative Online User Study: Results

We had a total of 108 participants, 75 for Round 1 and 33 for Round 2. We collected free-text justifications alongside participants' Likert-scale perceptions of a file's sensitivity, its usefulness, and how the participant wished to manage the file. Thus, our dataset is rich with insights that we leveraged in designing Aletheia (Section 7). Except as noted, we aggregate results across both rounds of data collection because the distributions of responses were similar in most cases.

6.1 Demographics and Security Hygiene

Table 4 summarizes participant demographics. 78% of participants primarily used Google Drive, and 22% Dropbox. Participants were diverse in age and profession, which included engineers, freelancers, office assistants, salespeople,

Category	Collection Method	List of Features
Metadata	Google Drive/Dropbox API	account size, used space, file size, file type (img, doc, etc.), extension (jpg, txt, etc.), last modified date, last modifying user, access type (owner, editor, etc.), sensitive filename, sharing status
Documents	Local text processing	bag of words for top 100 content keywords, LDA topic models, TF-IDF vectors, word2vec representations, table schemas for spreadsheets
Images	Google Vision API [20]	image object labels, adult, racy, medical, violent, logos, dominant RGB values, average RGB value
Sensitive Identifiers	Google DLP API [18]	***counts*** of the following identifiers in a file: name, gender, ethnic group, address, email, date of birth, drivers license #, passport #, credit card, SSN, bank account #, VIN

Table 3: A list of the features we automatically collected for each file using multiple APIs and custom code.

Gender		Age		Technical Background	
Male	63	18–34	75	Yes	25
Female	44	35–50	29	No	82
Non-binary	1	51+	4	Not answered	1

Table 4: Participant demographics (combined across rounds).

Categories Implying Sensitivity	% of Participants
Files containing the participant's PII	62%
Files containing PII of other than the participant	31%
Files with intimate or embarrassing content	30%
Files with original or creative content	84%
Files with proprietary information	23%
Categories Implying Usefulness	**% of Participants**
Files stored for future referencing	96%
Files with content of sentimental value	87%
Files which serve as backup	91%

Table 5: The percentage of participants who reported having files in categories implying they might be sensitive or useful.

and retailers. Participants were also well-established cloud storage users; 81% had used their account for 3 years or more. We observed both free and paid cloud accounts. Some participants used paid accounts provided by their work/school. All participants reported using their account for personal purposes, and 82% also used it for work/school. Participants were also reasonably frequent users of cloud storage; 22% of them used their account weekly, and 33% used it monthly.

Most participants were privacy-aware. Over 50% of them reported that they would be moderately or extremely concerned if their cloud files were stolen in a data breach. While 43% had enabled 2FA, nearly one-fourth of participants reported taking additional steps to protect their accounts. These included using strong passwords, backing up information, and monitoring for malicious activity.

6.2 Categories of Sensitive and Useful Files

In the first section of the survey, we asked participants to provide specific examples of files in various categories of potentially sensitive or useful files. Table 5 summarizes these categories and the fraction of participants who reported that they had files belonging to that category in their account.

Files Considered Sensitive: More than half of participants stated that their account had files containing PII. Files in this category were related to bank accounts (20%), taxes (19%), their resume (11%), and IDs (11%). Discussing financial documents, one participant wrote, *"When I was buying a house I might have uploaded some of the documents I needed for the mortgage onto the drive."* While the presence of others' PII was not very common (only 31%), such PII was typically that of school/work collaborators or family members. For example, P30 described *"tax returns that would have my family's Social Security Numbers and addresses."*

For intimate and embarrassing content, all participants who had such files mentioned it being an image or video file. 76% of participants specifically referenced nudity or porn. In this regard, one participant explained, *"I have nude photos of my wife on there, and I might have some of myself."*

Creative content was the most common category deemed sensitive. When asked about the specific type of creative work, participants mentioned school-related work (43%), art work (23%), and original writing (15%). Only 23% of participants expected that they had proprietary information in their account. Of those who did, 86% specifically identified it as work-related. For example, one participant wrote, *"There might be an NDA there but it is old and hopefully not any more of use."* This sentiment and the mortgage document quote above exemplify the interplay between long-term archives and file sensitivity. While enabling long term storage is helpful, it can also accumulate sensitive files that are no longer useful.

Files Considered Useful: Files that were in the cloud for future reference were the most common category of useful files, with 96% of participants mentioning such files. Common examples in this category were personal photos (21%), followed by documents for school (14%) and work (11%).

Among participants, 87% reported retaining files because of their sentimental value. For example, one participant wrote, *"I have a lot of my son's first milestones, Christmas photos. I have photos of my wife and me before kids. It helps me to remember how fast time flies."* Another common category was videos and personal writings that belonged to the participant.

Files retained as backups were most likely to consist of many different file types. Common examples included images (21%), work (16%) and school documents (8%), and miscellaneous backup items (14%). Participants also mentioned files

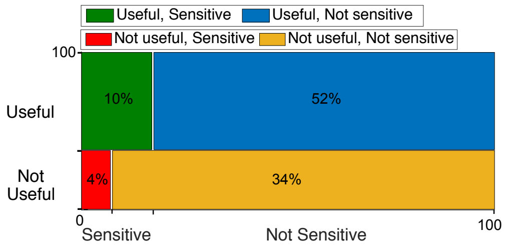


Figure 4: The distribution of sensitivity and usefulness labels. The percentages in each box represent the proportion of files belonging to each {sensitivity, usefulness} tuple.

Description (Selection Category #)	% Sensitive	% Useful
Sensitive keyword in file name (#1)	25%	65%
Document files (#2)	14%	61%
Media files (#3)	7%	67%
Other files (#4)	8%	51%
Top sensitive documents from classifier (#5)	15%	56%
Top sensitive images from classifier (#6)	15%	66%

Table 6: The percentage of files participants labeled as sensitive and useful, divided by the reason they were selected.

related to personal hobbies, such as music and games. For instance, one participant wrote, *"I am a hobbyist musician, so I like to keep previous versions of songs I make on Drive. There have been occasions where I make a mistake later on and it's nice to have a previous version I can go back to."*

Overall, 82% of files identified as sensitive or useful were images or documents. Note that our categorization of the latter included not just text-focused files, but also presentations and spreadsheets. Other file types considered sensitive included audio and video files (5%), as well as our miscellaneous category (13%) that encompassed saved web pages, computer code, database files, executables, and OS config files. This fact, combined with the additional filetype-specific features available for these files, led us to focus our prediction task (Section 7) specifically on images and documents.

6.3 Distribution of Sensitive and Useful Files

After we asked about useful and sensitive files in general, we showed each participant dozens of files from their own account, asking them to label and explain the usefulness, sensitivity, and the desired management decision for those files. This provided us with labels for a total of 3,525 files across rounds. Among the files we selected (biased towards those that are sensitive), 62% were deemed useful and 14% were deemed sensitive. Although the overall number of files perceived to be sensitive was low, 78% of our participants identified at least one file as sensitive. This observation aligns with previous studies that found a non-trivial fraction of the files stored in the cloud are potentially sensitive [9, 24].

Table 6 summarizes perceived usefulness and sensitivity across the file-selection categories. In Round 1, files with sensitive keywords in their file names and documents were more likely to be labeled as sensitive compared to other selection categories. Meanwhile, the distribution of file usefulness was fairly consistent across all categories. Figure 4, an area plot, summarizes the distribution of file usefulness and sensitivity.

6.4 Management of Sensitive and Useful Files

Figure 5 shows participants' desired file-management decisions broken down by whether they perceived the file as sensitive and/or useful. For files deemed useful and not sensitive, participants wanted to keep 93% of such files as-is. For files that were not useful, in the vast majority of cases the participant wanted to delete them, regardless of their sensitivity. This result is somewhat at odds with informal wisdom regarding digital packrats wanting to keep all data by default, but is consistent with the proposed management decision. In 94% of cases, participants wanted to delete files they considered not sensitive and not useful, while in 90% of cases they wanted to delete files they considered not useful, yet sensitive. These quantitative results directly align with the hypothesized management model we presented in Figure 1. When asked why they wanted to delete these files, the most common response was that they no longer needed them or that the files had served their purpose. The high likelihood of removing files shows both a willingness to reduce digital risk/clutter, as well as a lack of previous management that could have already deleted useless files from the account.

For files deemed sensitive and useful, participants wanted to protect 58% of them. In our model, we posited that users would be likely to protect all sensitive and useful files. Nonetheless, participants wanted to keep 39% of them as-is despite their sensitivity. A potential reason behind this decision is the subjective relationship between sensitivity and how risky the file is. Our assumption in Figure 1 was that all sensitive files are risky in some way and hence required management. However, our results revealed greater nuance. We asked participants why they wanted to protect these files. Popular reasons included that the file contained PII or financial information, the file had sentimental value, and that the file contained intellectual property. Most of the reported reasons were consistent with the understanding of sensitivity we developed during the qualitative interviews. We observed a strong correlation between sentimental value and sensitivity. For instance, one participant wrote, *"This is a photo of a loved one I would like to keep private."* Prominent reasons for the participant wanting to keep sensitive files as-is were that they were satisfied with the overall level of protection of their cloud account or that they did not consider the content to be sensitive enough to warrant additional protection. Representing the latter category, one participant described a file by writing, *"While it does contain proprietary information, it is not sensitive enough to prompt additional security."*

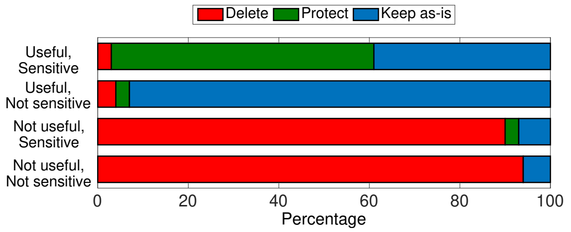


Figure 5: Desired file management by sensitivity/usefulness.

Overall, these file-management preferences and accompanying reasoning shed light on how participants conceptualize and operationalize file management in the cloud based on files' perceived sensitivity and usefulness. In Section 7, we leverage both our collected training data and these qualitative observations to build Aletheia, an automated inference approach to predict a file's usefulness, sensitivity, and management decision. Aletheia's ultimate goal is to assist users in protecting (or deleting) the files most likely to be in need of reconsideration.

6.5 Consistency of Decisions Over Time

A potential concern for self-report surveys like ours is that participants' answers might not be consistent over time and thus not represent a meaningful preference. To evaluate the stability of responses over time, we conducted a follow-up survey approximately eight months after the initial study. This follow-up survey was specifically designed for the 33 participants who had participated in Round 2 of our online study and thus had answered questions about a larger number of files than those who had participated in Round 1. Because we wanted to ask about a non-trivial number of decisions to either delete or protect files, we invited the 23 participants from that round who had desired to either delete or protect at least 10 of the 50 files shown to them. Of these 23 qualified participants, 16 participated in the follow-up study.

Similar to the initial study, participants were informed of our privacy and data-collection policies as part of the consent process. In the survey, we asked each participant to revisit a random selection of 10 files that they had previously wanted to delete or protect. We presented them with their previous file-management decision and asked them to select an updated decision and explain why they chose either the same decision or a different decision. Our 16 participants saw a total of 160 files, among which they initially wanted to delete 136 files and protect 24 files. The survey took approximately 15 minutes to complete, and the additional compensation was $7. Before we conducted this follow-up study, our IRB approved our request for a protocol modification.

After 8 months, participants reported the same management decision for 81% of the files. For files that participants initially wanted to delete, participants wanted to continue to delete 86% of such files. In explaining why their decisions remained the same, participants mentioned that the files were either embarrassing or no longer needed. For example, one participant wrote, *"Same as before: total junk."* While participants had been made aware of the presence of these files a few months ago in the initial study, only one participant had actually manually deleted the files in the interim. Doing so required them to log into their account through their normal interface outside of our study system. To facilitate file management, automated tools that are part of users' workflow can potentially increase the feasibility of the management process and reduce manual overhead. For a smaller portion of these files, participants wanted to revert their initial decision from delete to keep as-is (13%) or protect (1%). Participants stated two prominent reasons for changing their decision. For 37% of files for which the decision differed, participants mentioned that the file had sentimental value. For instance, one participant said, *"Upon seeing the photo again, it brings back good memories. It's been some time since I've seen these photos."* For the remaining 63% of files, participants reported realizing the file was potentially useful. One wrote, *"I thought that I wouldn't need this anymore, but now I think that I may."*

Among the files that participants had initially desired to protect, they wanted to continue to do so for 48% of them. Participants' free-text justifications mentioned that the files continued to be useful, yet contained sensitive information. However, participants now wanted to delete 42% of these files they initially wanted to protect. In a matter of months, these files had lost their utility in participants' eyes. These changed decisions are consistent with longitudinal file management; a sensitive file that is initially protected can easily be deleted once it is no longer deemed useful, whereas the opposite is impossible. The overall stability of participants' preferences over time provides further motivation for the development of advanced mechanisms that can keep track of dynamic file attributes longitudinally, which we elaborate on in Section 8.

7 Predicting File-Management Decisions

Because users can have hundreds or thousands of files in their cloud storage accounts, a core goal of ours was to alleviate the burden of manual file management with automated tools. In this section, we formulate the task of predicting file-management decisions based on features automatically collected from individual files and user accounts as a whole. To inform the classifier for file-management decisions, we also predict user perceptions of file usefulness and sensitivity.

7.1 Prediction Tasks and Baselines

Aletheia has three prediction tasks: predicting whether a user will perceive a file as sensitive (Task 1); predicting whether a user will perceive a file as no longer useful (Task 2); and predicting what management decision a user will choose among keeping, deleting, and protecting a file (Task 3). To perform classification for each task, we compared several established supervised learning algorithms: Decision Trees (*DT*), Logistic Regression (*LR*), Random Forests (*RF*), Deep Neural Networks (*DNN*) with the Adam optimizer using scikit-learn [43], and XGBoost (*XGB*) [8]. All model parameters were optimized using grid search on the training set in each fold in cross validation, and tested on the testing set. We use the best performing classifier, which turned out to be XGBoost for both the preliminary classifiers trained on Round 1 data and the final classifiers trained on Round 2 data. We report results only on the final classifiers, which we refer to as *Aletheia w/ all features*, or *Aletheia* for short.

We compared Aletheia to multiple baselines. The first was a random classifier (*Random*), which randomly assigned a management decision for each file. The second was a majority classifier (*Majority*), which always predicted the most frequent class. For the task of predicting whether a file would be perceived as sensitive, we employed a more meaningful third baseline, *GDLP feature count*, leveraging Google's Cloud Data Loss Prevention API [18] (see Table 3). This baseline ranked documents based on the number of sensitive GDLP features identified in each document. We also tested a variant of our model that used only the GDLP output as features for predicting sensitivity: *Aletheia w/ only GDLP features*.

For predicting whether a file would be perceived as useful, we again used the Random and Majority baselines. We also tested two additional baselines centered on how recently the file was last modified and how useful files of its type were considered overall. We ordered all files by last modification date, from oldest to newest, and assigned them a "staleness" score between 1 (oldest) and 0 (newest) by normalizing the last modification date. The *Last Modified* baseline predicted the most stale files (those not modified recently) as *not useful*. The *Last Modified, File Type* baseline augmented the staleness score with overall statistics about the perceived usefulness of other files with the same file extension. For every file type, a "not-useful-type" score between 1 and 0 was calculated by considering all files of that file type (e.g., PDFs) in the training data and calculating what percent of them were marked as *not useful*. The *Last Modified, File Type* baseline ranked files based on the product of their staleness and not-useful-type scores. It allowed for files whose type is generally perceived as less useful to be ranked higher than files whose type is generally considered more useful. To the best of our knowledge, no prior work has attempted to predict perceptions of sensitivity and usefulness or file-management decisions for files in the cloud. We thus chose these baseline to represent common machine learning baselines and additional baselines capturing the most intuitive features for sensitivity and usefulness.

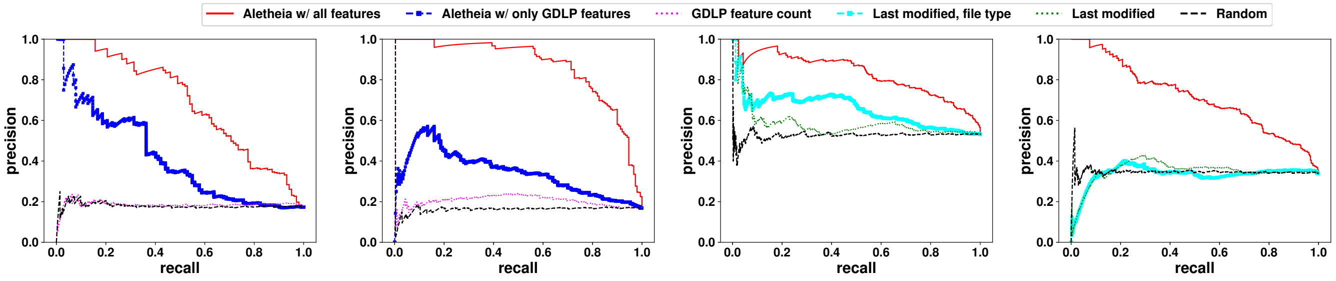
7.2 Dataset Description

We used the final dataset collected in Round 2. Our dataset consisted of tuples $(\mathbf{X}, Y)$, where X_i was the feature vector and Y_i was our target for prediction. The feature vector X_i, included metadata and information on files and user accounts. For accounts, we had the total amount of storage and the amount used. For files, we had the size of the file, whether or not the file was shared, the link access (view or edit), whether or not the file was last modified by the user, and the access type (owner, editor, viewer). For documents and images that contained text, we extracted counts of sensitive information discovered using the GDLP API [18]. In addition, we collected a bag of words on a heuristic set of keywords. For documents, we collected an average word2vec embedding of each document using Google News word2vec embeddings [33]. Doing so enabled us to approximate text context without breaching the privacy of participants by having actual interpretable text from their files. For images, we used the Google Vision API [20] to obtain multiple image features. We additionally converted the labels from the API, including the "best guess label," to one-hot encoding representations, as well a word2vec embedding representation, which were added to the feature vector. These features are listed in Table 3. In addition, we also computed the following user-level statistics as features: (1) the percentage of files in a participant's account with each sensitive feature (e.g., the fraction of files tagged as adult); and (2) the percentage of files labeled as sensitive in the training data that contained each sensitive feature. Compared to *Aletheia w/ only GDLP features*, we considered a broader set of file-based and user-based features.

For file-management decisions (Task 3), we used all files for which we collected survey data. For Task 1 and Task 2, we separated the evaluation by image files and document files since they had different features. The labels Y for each task were obtained from participants' answers to questions S-1, U-1 and M-1 (as labeled in the survey instrument in online Appendix B [1]). Questions S-1 and U-1 asked participants to rate a file's sensitivity and usefulness, respectively, on a Likert scale. Question M-1 inquired how participants wanted to manage the file by either deleting it, protecting it, or keeping it as-is. Based on the answers to S-1, we created binary labels for Task 1: *sensitive* ("strongly agree," "agree") and *not sensitive* ("neutral," "disagree," "strongly disagree"). A total of 15% of files were sensitive. Based on the answers to U-1, we created binary labels for Task 2: *not useful* ("strongly disagree," "disagree") and *useful* ("neutral," "agree," "strongly agree"). A total of 38% of the files were not useful. Note that for both S-1 and U-1, "neutral" responses were assigned to the categories that we were *not* interested in finding (*not sensitive* and *useful*). From the answers to M-1, we had three labels for Task 3: *delete* (40%), *protect* (8%), and *keep* (52%).

7.3 Experimental setup

Tasks 1 and 2 for predicting sensitivity and usefulness had the same setup, while predicting file-management decisions in Task 3 used a different setup.



(a) Sensitivity of documents

(b) Sensitivity of images

(c) Usefulness of documents

(d) Usefulness of images

Figure 6: Precision vs. recall for predicting sensitivity and usefulness. We compared two versions of *Aletheia* (red and blue) against a random baseline (black) and a baseline using Google's Data Loss Prevention (GDLP) tool (magenta) for the sensitivity dataset. For the usefulness dataset, we compared against two heuristic baselines using the last modification date (cyan and green).

7.3.1 Task 1 and 2: Sensitivity and usefulness

We performed a *nested cross-validation* [40, 46, 54] and report averaged results across five test folds. To this end, we first created five training and test folds. Within each training fold, we further performed a five-fold cross-validation to tune and select hyperparameters. Finally, each tuned model was evaluated on the respective test folds and performance was averaged across all five test folds. This allowed us to see how the model may perform in the general setting, and it also reduced bias from selecting a single random test set. Note that we *did not* tune any hyperparameters on the separated test fold; it was used exclusively for evaluation.

Since we focused on finding files that participants wished to delete or protect, we ordered examples in the test data by the probability of being $Y_i = 1$ (*sensitive* for Task 1, *not useful* for Task 2), and assessed the precision and recall. This is a common setup for evaluating binary classification where one label (e.g., *sensitive*) is more important than the other (e.g., *not sensitive*). Since there were significantly fewer *sensitive* and *not useful* labels, we had a "needle in the haystack" problem. We aimed for both high precision and high recall, but there is typically a trade-off between them. Precision was computed as $TP/(TP+FP)$, where TP was the number of true positive examples (actual label positive, predicted label positive), and FP was the number of false positive examples (actual label negative, predicted label positive). In other words, precision was the proportion of examples predicted as bearing the label of interest that were correctly predicted. It is also known as the positive predictive value in information retrieval [6, 39] or Bayesian detection rate in intrusion detection [3]. Recall was computed as $TP/(TP+FN)$, where FN was the number of false negative examples (actual label positive, predicted label negative). A precision-recall curve (PRC) allows us to see the trade-off between precision and recall when different possible cutoffs for positive classifications are used. For example, if we predicted the top 20% most likely files as positive, then a point on the PRC is created which shows the exact trade-off between including false positives in the files predicted as positive (top 20%) and not including the false negatives, files that should be predicted as positive but fall in the lower 80%.

7.3.2 Task 3: File-Management decision

In this task, we had three classes: *delete* ($Y_i = 1$), *protect* ($Y_i = 2$), and *keep* ($Y_i = 3$). Since perceptions of sensitivity and usefulness correlated highly with file-management decisions, we wanted to leverage them in our classification. However, one typically does not have these labels for all files in a user's account. Thus, we predicted these labels using the classifiers for Task 1 and 2, adding the predicted labels as two additional features for predicting the file-management decision. We compared the performance of adding these two features against a classifier that does not use them, a majority classifier, and an *oracle* with the actual perceived sensitivity and usefulness of a file as reported by the participant.

7.4 Results

Here, we present the precision-recall curves for sensitivity and usefulness, separated into image and document files. We also analyze the top features for predicting sensitivity and usefulness. For both Task 1 and Task 2, a majority classifier performed the same as a random classifier for precision and recall, so we do not report results for random classifiers.

7.4.1 Task 1: Sensitivity

We first tried to predict if a user would perceive a document or image as *sensitive*. Figures 6a–6b show precision vs. recall curves (PRC) for the **sensitivity dataset**. Figure 6a shows the PRC for predicting the sensitivity of documents. *Aletheia* performed the best overall, while *Aletheia w/ only GDLP features* performed worse. The *GDLP feature count* classifier did not perform better than the *Random* baseline in this setting. *Aletheia* had an AUC 0.68, while *Aletheia w/ only GDLP features* had an AUC of 0.40, an improvement of 68%.

Figure 6b shows the PRC for predicting the sensitivity of images. *Aletheia* had an AUC of 0.86, compared to *Aletheia*

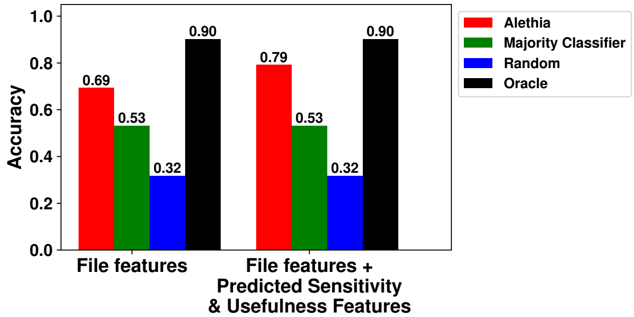


Figure 7: A comparison between directly predicting the file-management decision versus first predicting sensitivity and usefulness for all files. *Aletheia* is compared to a majority classifier and an oracle that knows participants' responses about a file's sensitivity and usefulness.

w/ only GDLP features with an AUC of 0.34, an improvement of 153%. The *GDLP feature count* classifier had an AUC of 0.20. Compared to prediction for documents, we observed much better performance in terms of the PRC for *Aletheia*, but not for *Aletheia w/ only GDLP features*.

From the sensitivity results, we found that a broader set of features besides counts of sensitive information provided more accurate results. We also found that *Aletheia* performed better at predicting the sensitivity of images than documents. This makes sense because we had additional image features capturing adult, racy, spoofed, medical, and violent content, which may be indicative of sensitivity for images.

7.4.2 Task 2: Usefulness

We next tried to predict if a user would perceive a document or image as being *not useful*. Figures 6c–6d show precision vs. recall curves (PRC) for predicting that specific documents and images in the **usefulness dataset** were *not useful*. *Aletheia* performed the best in both tasks, significantly outperforming the baseline classifiers.

Figure 6c shows the PRC for predicting *not useful* on documents. The best baseline classifier was the *Last Modified, File Type* heuristic, which performed reasonably well for predicting that a document was *not useful*. *Aletheia* achieved an AUC value of 0.82, compared to the best baseline AUC value of 0.65, an improvement of 26%. For images, all baseline classifiers performed similarly, with the *Last Modified* heuristic performing the best among them with an AUC of 0.35. *Aletheia* was more accurate in predicting *not useful* than the baselines, with an AUC of 0.71, an improvement of 101%.

We found that predicting images as *sensitive* was easier than predicting documents as *sensitive*, while predicting documents as *not useful* was easier than predicting images as *not useful*. This may be due to varying perceptions of usefulness, which are less likely to be captured from image features.

Decision	Correct Predictions	Incorrect Predictions
Keep	91%	9% (delete)
Delete	75%	25% (keep)
Protect	37%	56% (keep), 7% (delete)

Table 7: Accuracy per file-management decision. For incorrect predictions, we show our classifier's prediction in parentheses. For example, we predicted 91% of *keep* labels correctly, while incorrectly predicting 9% of *keep* labels as delete.

7.4.3 Task 3: File-Management Decision

Finally, we tried to predict file-management decisions, and Figure 7 shows our overall accuracy in doing so. We show results for both predicting the file-management decision directly from file features, as well as the aforementioned two-step process in which we first predicted the file's sensitivity and usefulness, subsequently using these predictions as additional features in predicting the file-management decision. The oracle shows that if the classifier knew participants' actual responses for perceived usefulness and sensitivity, we could achieve 90% accuracy for predicting file-management decisions. In the more realistic scenario of using only file features and (possibly incorrect) predictions of sensitivity and usefulness in our two-step process, we saw a roughly 10% increase in accuracy compared to the single-step process. This result shows that even without an oracle of participants' actual responses for a file's sensitivity and usefulness, leveraging our predictions of those perceptions boosted the accuracy of predicting the file-management decision.

Table 7 compares accuracy across file-management decisions. We were most accurate on *keep* decisions, the majority class. For cases in which we mispredicted *keep* decisions, *Aletheia* always instead predicted them as *delete*. On the other hand, for *delete* decisions, *Aletheia* had 75% accuracy, mispredicting *delete* decisions as *keep*. With only 37% accuracy, *Aletheia* did not perform very well on *protect* labels. Interestingly, the majority of mispredictions for *protect* were mislabeled as *keep*. This result shows that *Aletheia* considered *protect* decisions as closer to *keep* decisions than *delete*. The number of *protect* labels was significantly smaller than the other two labels, making it harder to predict.

7.5 Understanding Prediction Results

We also examined which features were important for each prediction task. Table 8 shows the top features identified by each classifier for the sensitivity and usefulness tasks, in order of importance. Generally, word2vec had high feature importance in the classifiers. However, since word2vec features are not easy to interpret [27], we do not show them on the list. For documents in the sensitivity task, we noticed that user-level statistics like the fraction of files in the account containing

Task		Features
Sensitivity	Documents	gender; fraction of ethnic/VIN/location files; credit card; date of birth; email
	Images	fraction of gender/SSN/ethnic/location files; adult; credit card; racy; passport
Usefulness	Documents	access type; last modifying user; finance keywords; report & journal keywords
	Images	file size; finance keywords; access type; last modifying user; medical keywords
File Management	All Files	usefulness; sensitivity; spoof; account size; used space; finance keywords; medical keywords

Table 8: Top features for prediction tasks. Italicized *keywords* were top terms identified via the bag of words collections.

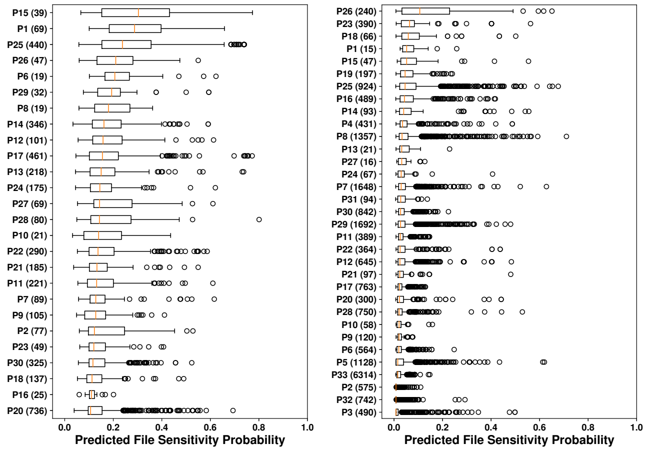
ethnic terminology, VIN numbers, and locations played a role in prediction, as did specific features like credit card numbers, dates of birth, and email addresses. For images, we saw some of the same important features, but also sensitive image features, such as whether content was potentially adult or racy.

For the usefulness dataset, we saw some similar top features as in the sensitivity dataset, including access type and financial keywords. For documents, report and journal keywords were important in predicting usefulness. For images, medical content was also predictive of usefulness.

Besides the top features in Table 8, word2vec embeddings were also identified as important features. This means that text content is central to these prediction tasks. For documents, one word2vec embedding represented the entire document. However, for images, we considered additional one-hot encoding and word2vec features based on the automatically identified image labels. Of those, only word2vec features were identified as important, probably because one-hot encodings of "best guess labels" were too sparse.

Table 8 also shows the most important features for predicting file-management decisions. The top two features were the predicted labels for file usefulness and sensitivity. Using XGBoost, the feature importances of usefulness and sensitivity predictions were 0.40 and 0.11, respectively, confirming our earlier observation about the two-step prediction process being superior to the one-step process. Sensitive information in the file, such as medical content, was an important feature.

To better understand the distribution of sensitive files in a single account, Figure 8 shows box plots of the predicted probability a file was sensitive for all documents and images in each Round 2 participant's accounts. We omitted participants with fewer than 10 documents or images in their accounts. These predictions came from our preliminary classifier, which was trained on Round 1 data. On average, the preliminary classifier predicted the majority of files as having a low probability of being sensitive. Only a small subset of files with high probability of being sensitive were selected for each participant. For many participants, we were selecting only a small number of files that the preliminary classifier deemed sensitive with high probability. This resulted in a high percentage of potentially sensitive files in our Round 2 dataset.



(a) Documents

(b) Images

Figure 8: Predicted sensitivity probability for each document and image for every participant. On average, our classifier predicts low sensitivity for a majority of files, and high sensitivity on a small number of files.

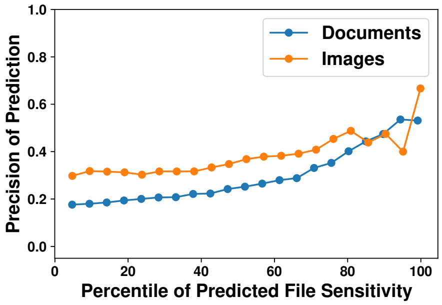


Figure 9: Preliminary classifier prediction precision as a function of predicted file sensitivity. The increasing precision at higher predicted sensitivity scores indicates that predictions of high sensitivity are more accurate.

To explore our preliminary classifier's accuracy in picking sensitive files, we looked at the relationship between high probability predictions a file was sensitive and the precision of prediction. We ranked the selected files in order of predicted probability of being sensitive, classifying files based on a sliding threshold for which everything above the threshold was classified as sensitive. Finally, we computed the precision for the files above the threshold on our ground truth file sensitivity, reporting the results for both documents and images in Figure 9. A higher percentile means a higher threshold for the predicted probability of sensitivity. When the threshold for predicted probability was low, we had lower precision (around 30%). With higher predicted probability, our preliminary classifier had better precision. This shows that the preliminary classifier produced meaningful sensitivity predictions.

To better understand mispredictions, we also performed a qualitative evaluation of files that were false positives in our final classifier, alongside the reasons why participants did not consider them sensitive. Specifically, we looked at false positives in the top five documents and images by rank.

Within documents, 6% of such files contained PII that was obsolete. One participant wrote, *"It's just a cover letter I had written several years ago and doesn't contain any good info because the address and phone aren't good anymore.*" As highlighted in Table 8, phone numbers and addresses are both important in predicting sensitivity. However, accurately classifying files requires more temporal information and context. Similarly, 3% of the files contained sensitive information belonging to someone other than the participant, so they did not consider the file sensitive. Regarding an ex-partner's resume, a participant wrote, *"It might be slightly sensitive to my ex, but not really."* This particular finding supports prior work [44] suggesting that life experiences impact data-privacy valuations. For a majority of the other documents (70%), participants' responses did not indicate a strong element of sensitivity. They mentioned that the files contained information they did not feel could compromise them in any manner, or details that were already publicly available.

Most images that were misclassified as sensitive were pictures with faces, memes, or some form of artwork or original content. However, participants did not perceive them as sensitive. For a family photo, a participant wrote, *"This does not reveal any personal information about me, or the person in the photo."* In another example of original artwork, a participant mentioned, *"There is nothing sensitive in the file, but I would not want someone stealing the image to use as their own."* Pictures containing adult content that did not directly affect the participants were also not considered sensitive by some participants. Regarding a nude photo, a participant mentioned it was not compromising as they were not in the photo.

This investigation revealed that files shown to participants conformed with our broader definition of sensitivity listed in Table 1. However, different participants had varying sensitivity thresholds, which eventually weighed more into the decision-making process of how they wanted to manage the files. Better understanding this phenomenon requires both additional data collection and the development of personalized classifiers that account for such personalization. We note this as a limitation of our current study, discussing possible future work in this direction in the next section.

8 Discussion and Future Work

Decisions about file management are predicated on several factors, some internal to the user and some based on the contents of the file. The design of Aletheia focuses not on directly predicting that decision, but rather on predicting perceptions regarding these files that can be inferred using passively collected file metadata, which can then in turn be useful in predicting the ultimate file-management decision. To this end, we applied the usefulness/sensitivity model from Figure 1.

Our findings in Section 7.4 were particularly encouraging for the usefulness part of this model, as using automated inference techniques to first build an understanding of participants' conceptualization of usefulness significantly improved our ability to predict their file-management decision; the predicted usefulness was the single most predictive feature for the file-management-decision classification. This holistic, human-centered approach to automated inference highlights the importance of deep qualitative engagement with users during the design of such classifiers.

Not only does this human-centered understanding improve the performance of automated inference, but this approach can also develop a deeper understanding of perceived usefulness and sensitivity for files. Perceptions of usefulness are strongly correlated with future access, while perceptions of sensitivity correlate with the existence of PII, financial information, intimate content, and sentimental value.

While Figure 5 shows a very strong correlation between usefulness and desire to delete a given file, as well as keeping non-sensitive useful files as-is, two more subtle points arise. First, participants' preferences for how to manage useful, sensitive files did not map onto our hypothesized model; decisions to protect useful files were nearly evenly split between sensitive and not-sensitive files. Second, while not-useful files were nearly always deleted, participants still wanted to retain a nontrivial minority of files deemed not useful. This phenomenon suggests that using the concept of usefulness is very helpful for determining whether to retain a given file. Nonetheless, automated systems should not use such a prediction to make file-retention decisions automatically on behalf of the user, but rather should seek confirmation from the user.

While predicting file usefulness was incredibly helpful for subsequently predicting file-management decisions, predicting file sensitivity was both less successful and less helpful for predicting file-management decisions. Beyond being harder to accurately predict because the base rate of sensitive files is low (13%), these phenomena suggest the relationship between sensitivity and file management is more complex than our hypothesized model. Future work could explore whether classifiers tuned to individual users' preferences would be able to improve performance on using sensitivity predictions to underpin file-management-decision predictions.

Within the sensitivity prediction task, our classifier performed better for images than for documents. While this can be an artifact of the underlying data, we hypothesize that some of the significant features for images, such as the "adult," "racy," and "violent" features, are evidently easier to automatically detect among images of different users. For documents, we observed that while there were standardized classes of significant and clearly identifiable features (e.g., PII and financial information) that are straightforward to detect, qualitative responses from participants suggest the presence of a strong temporal relevancy of these features. Our classifier does not account for contexts, temporal or otherwise. Similarly for some images, participants described sensitive pictures as having sentimental value (e.g., pictures of children, loved ones). Directly predicting sensitivity of this kind from our features is not feasible. This task certainly merits deeper investigation.

Future work to further improve our understanding of file sensitivity and file management should focus on longitudinal studies. This will enable us to passively observe participants' actions over time, rather than actively asking the participant to make management decisions. Longitudinal data will enable building a *sensitivity persona* that can account for the variation in sensitivity perceptions among individuals. The success of a classifier depends to a large extent on the training and testing data coming from the same distribution. If the covariate distribution changes over time, a problem known as concept drift, then the classifiers would need to be updated and account for this concept drift in order to perform well. While in this work we lacked longitudinal data and were thus unable to check for concept drift, a quantitative evaluation of the drift effect on classifier performance for retrospective file management would be a fruitful direction for future work.

Additionally, future work should focus on broadening the participant pool to minimize sources of potential bias and better account for cultural diversity, as well as understanding the trade-off between file management and the associated risk of sensitive files. This can be achieved by developing and widely deploying an effective user management interface with additional surveys, which can surface these ideas efficiently. Overall, these efforts would minimize our current limitations and operationalize the results of our work to improve Aletheia's performance.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants CNS-1801644 and CNS-1801663. We thank Will Brackenbury for his assistance with our data-collection infrastructure, as well as Noah Hirsch and Michael Tang for their assistance with our interviews.

References

[1] Interview scripts and survey instruments, 2021. `https://bit.ly/usenix21appendix`.

[2] Qualitiative inteview questions codebook, 2021. `https://bit.ly/usenix21codebook`.

[3] Stefan Axelsson and David Sands. *Understanding Intrusion Detection Through Visualization*. Springer Science & Business Media, 2006.

[4] Benett Axtell and Cosmin Munteanu. Back to real pictures: A cross-generational understanding of users' mental models of photo cloud storage. *PACM IMWUT*, 3(3):74, 2019.

[5] Russell Brandom. The Capital One breach is more complicated than it looks. The Verge, Jul 2019. `https://www.theverge.com/2019/7/31/20748886/capital-one-breach-hack-thompson-security-data`.

[6] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. The binormal assumption on precision-recall curves. In *Proc. ICPR*, 2010.

[7] Anthony M. Butler. Data leak prevention enforcement based on learned document classification. International Business Machines Corporation, US Patent 9,626,528, April 18, 2017.

[8] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proc. KDD*, 2016.

[9] Jason W. Clark, Peter Snyder, Damon McCoy, and Chris Kanich. I saw images I didn't even know I had: Understanding user perceptions of cloud storage privacy. In *Proc. CHI*, 2015.

[10] David M. Douglas. Doxing: A conceptual analysis. *Ethics and Information Technology*, 18(3):199–210, 2016.

[11] European Parliament and Council of the European Union. Regulation (EU) 2016/679. *Official Journal of the European Union (OJ)*, 59(1-88), 2016.

[12] Lujun Fang and Kristen LeFevre. Privacy wizards for social networking sites. In *Proc. WWW*, 2010.

[13] Casey Fiesler, Michaelanne Dye, Jessica L. Feuston, Chaya Hiruncharoenvate, Clayton J. Hutto, Shannon Morrison, Parisa Khanipour Roshan, Umashanthi Pavalanathan, Amy S. Bruckman, Munmun De Choudhury, and Eric Gilbert. What (or who) is public?: Privacy settings and social media content sharing. In *Proc. CSCW*, 2017.

[14] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. "A stalker's paradise": How intimate partner abusers exploit technology. In *Proc. CHI*, 2018.

[15] David Garcia. Leaking privacy and shadow profiles in online social networks. *Science Advances*, 3(8), 2017.

[16] Kambiz Ghazinour, Stan Matwin, and Marina Sokolova. Monitoring and recommending privacy settings in social networks. In *Proc. EDBT Workshops*, 2013.

[17] Neil Zhenqiang Gong and Bin Liu. You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors. In *Proc. USENIX Security*, 2016.

[18] Google. Cloud data loss prevention, 2021. https://cloud.google.com/dlp/.

[19] Google. Cloud speech to text, 2021. https://cloud.google.com/speech-to-text/.

[20] Google. Cloud vision, 2021. https://cloud.google.com/vision/.

[21] 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.

[22] Wenjin Hu, Tao Yang, and Jeanna N. Matthews. The good, the bad and the ugly of consumer cloud storage. *ACM SIGOPS Operating Systems Review*, 44(3):110–115, 2010.

[23] I. Ion, N. Sachdeva, P. Kumaraguru, and S. Čapkun. Home is safer than the cloud!: Privacy concerns for consumer cloud storage. In *Proc. SOUPS*, 2011.

[24] Mohammad Taha Khan, Maria Hyun, Chris Kanich, and Blase Ur. Forgotten but not gone: Identifying the need for longitudinal data management in cloud storage. In *Proc. CHI*, 2018.

[25] Spyros Kokolakis. Privacy attitudes and privacy behaviour. *Comput. Secur.*, 64(C):122–134, January 2017.

[26] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *PNAS*, 110(15):5802–5805, 2013.

[27] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proc. ACL*, 2014.

[28] Jack Lindamood, Raymond Heatherly, Murat Kantarcioglu, and Bhavani Thuraisingham. Inferring private information using social network data. In *Proc. WWW*, 2009.

[29] Kun Liu and Evimaria Terzi. A framework for computing the privacy scores of users in online social networks. *TKDD*, 5(1):6, 2010.

[30] Sonia Livingstone, Leslie Haddon, Anke Görzig, and Kjartan Ólafsson. Risks and safety on the internet: The perspective of European children. LSE, London: EU Kids Online, 2011. https://resourcecentre.savethechildren.net/library/risks-and-safety-internet-perspective-european-children-full-findings-and-policy.

[31] Microsoft. Information rights management, June 30, 2020. https://docs.microsoft.com/en-us/exchange/information-rights-management-exchange-2013-help.

[32] Adriana Mijuskovic and Mexhid Ferati. User awareness of existing privacy and security risks when storing data in the cloud. In *Proc. e-Learning*, 2015.

[33] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proc. NIPS*, 2013.

[34] Kimberly J. Mitchell, Lisa M. Jones, David Finkelhor, and Janis Wolak. Trends in unwanted online experiences and sexting. Crimes Against Children Research Center, 2014.

[35] Anthony D. Miyazaki and Ana Fernandez. Consumer perceptions of privacy and security risks for online shopping. *Journal of Consumer Affairs*, 35(1):27–44, 2001.

[36] Ambar Murillo, Andreas Kramm, Sebastian Schnorf, and Alexander De Luca. "If I press delete, it's gone": User understanding of online data deletion and expiration. In *Proc. SOUPS*, 2018.

[37] Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar. Violence detection in video using computer vision techniques. In *Proc. CAIP*, 2011.

[38] Maureen K. Ohlhausen. Painting the privacy landscape: Informational injury in FTC privacy and data security cases. FTC Public Statement, 2017. https://www.ftc.gov/system/files/documents/public_statements/1255113/privacy_speech_mkohlhausen.pdf.

[39] Brice Ozenne, Fabien Subtil, and Delphine Maucort-Boulch. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of Clinical Epidemiology*, 68(8):855–859, 2015.

[40] Saeid Parvandeh and Brett A. McKinney. EpistasisRank and EpistasisKatz: Interaction network centrality methods that integrate prior knowledge networks. *Bioinformatics*, 35(13):2329–2331, 2019.

[41] Seth Patton. OneDrive Personal Vault brings added security to your most important files and OneDrive gets additional storage options, 2019. https://www.microsoft.com/en-us/microsoft-365/blog/2019/06/25/onedrive-personal-vault-added-security-onedrive-additional-storage/.

[42] Sai Teja Peddinti, Aleksandra Korolova, Elie Bursztein, and Geetanjali Sampemane. Cloak and swagger: Understanding data sensitivity through the lens of user anonymity. In *Proc. IEEE S&P*, 2014.

[43] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[44] Yu Pu and Jens Grossklags. Valuating friends' privacy: Does anonymity of sharing personal data matter? In *Proc. SOUPS*, 2017.

[45] Kopo M. Ramokapane, Awais Rashid, and Jose M. Such. "I feel stupid I can't delete...": A study of users' cloud deletion practices and coping strategies. In *Proc. SOUPS*, 2017.

[46] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

[47] Clayton Santos, Eulanda M. dos Santos, and Eduardo Souto. Nudity detection based on image zoning. In *Proc. ISSPA*, 2012.

[48] Peter Snyder, Periwinkle Doerfler, Chris Kanich, and Damon McCoy. Fifteen minutes of unwanted fame: Detecting and characterizing doxing. In *Proc. IMC*, 2017.

[49] Peter Snyder and Chris Kanich. Cloudsweeper: Enabling data-centric document management for secure cloud archives. In *Proc. CCSW*, 2013.

[50] Fred Stutzman, Ralph Gross, and Alessandro Acquisti. Silent listeners: The evolution of privacy and disclosure on Facebook. *Journal of Privacy and Confidentiality*, 4(2):7–41, 2013.

[51] Ashwini Tonge and Cornelia Caragea. Dynamic deep multi-modal fusion for image privacy prediction. In *Proc. WWW*, 2019.

[52] Lars Arne Turczyk, Oliver Heckmann, and Ralf Steinmetz. File valuation in information lifecycle management. In *Proc. AMCIS*, 2007.

[53] Karen Turner. Hacked Dropbox login data of 68 million users is now for sale on the dark web. The Washington Post, Sep 2016. https://www.washingtonpost.com/news/the-switch/wp/2016/09/07/hacked-dropbox-data-of-68-million-users-is-now-or-sale-on-the-dark-web/.

[54] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1):91, 2006.

[55] Francesco Vitale, William Odom, and Joanna McGrenere. Keeping and discarding personal data: Exploring a design space. In *Proc. DIS*, 2019.

[56] Fons Wijnhoven, Chintan Amrit, and Pim Dietz. Value-based file retention: File attributes as file value and information waste indicators. *Journal of Data and Information Quality*, 4(4), 2014.

[57] Pamela Wisniewski, Heng Xu, Mary Beth Rosson, Daniel F. Perkins, and John M. Carroll. Dear diary: Teens reflect on their weekly online risk experiences. In *Proc. CHI*, 2016.

[58] Kim Zetter. Hackers finally post stolen Ashley Madison data. Wired, Jun 2017. https://www.wired.com/2015/08/happened-hackers-posted-stolen-ashley-madison-data/.

[59] Elena Zheleva and Lise Getoor. To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles. In *Proc. WWW*, 2009.

[60] Elena Zheleva, Evimaria Terzi, and Lise Getoor. Privacy in Social Networks. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 3(1):1–85, 2012.