Abstract
Online archives, including social media and cloud storage, store vast troves of personal data accumulated over many years. Recent work suggests that users feel the need to retrospectively manage security and privacy for this huge volume of content. However, few mechanisms and systems help these users complete this daunting task. To that end, we propose the creation of usable retrospective data-management mechanisms, outlining our vision for a possible architecture to address this challenge.

Author Keywords
retrospective data management; usable privacy and security; social media; cloud storage

ACM Classification Keywords
H.1.2 [User/Machine Systems]: Human factors

Motivation
Online services like social media (e.g., Facebook) and cloud storage (e.g., Dropbox) store content for billions of Internet users. These systems often accumulate substantial amounts of data over the years and effectively function as online archives. However, users of these systems often take a “set-it-and-forget-it” approach to managing access to this data. Even when access-control settings match the user’s intent when the content is initially uploaded, that setting
might no longer reflect the user’s intent years later. For example, intuitively, a Facebook post shared with “all friends” in 2008 might not be suitable for sharing with “all friends” in 2018. The author of the post, any subjects tagged in it, and even who is included in the “all friends” access-control setting have likely changed substantially in the intervening decade. The post and its settings, however, in most cases remain unchanged ten years later.

Recent work by Ayalon et al. found that users’ willingness to share past content on social media decreases with the age of the content [1]. This decrease is triggered by life changes, relationship changes, and decreases in the perceived relevance of the post. In a similar study, Bauer et al. found that, over time, users sometimes want to decrease the size of the audience who accesses their past content, yet sometimes want to increase the size of the audience to aid in reminiscence [2]. While those two studies focused on social media, Khan et al. examined analogous situations for cloud storage. They found that 48% of participants wanted to delete or encrypt at least half of the files from their Dropbox or Google Drive accounts presented in the study [4].

These studies together demonstrate a desire for retrospective management of online personal archives. However, in none of these formative studies did participants actually change their settings or the audience of their past content. Doing so in practice presents two major challenges. First, users accumulate a huge amount of content in their online archives. Khan et al. noted that some of their participants had tens of thousands of files in their Dropbox or Google Drive accounts [4]. Browsing through all of these files and manually changing access-control settings is infeasible.

Second, even when users aim to retrospectively change access to content retrospectively, they have a number of conceptual mechanisms for doing so. They can change the audience by modifying an access-control setting. They can often also delete the content or edit it. Some platforms (e.g., Tumblr) let users edit or delete content without a trace, while others (e.g., Facebook) make visible a full revision history. On some platforms, users can also encrypt content, archive it (keep a copy only for themselves), or anonymize a public post. Work by Mondal more fully mapped this design space of retrospective mechanisms [5].

With so many potential mechanisms for changing content’s visibility, or even the content itself, over time, there is a strong need to improve current retrospective data-management mechanisms. In this poster, we outline an infrastructure for potentially helping users retrospectively manage access to their old data.

Our Vision for Retrospective Management
We envision a retrospective data-management system that, given a user’s online archive, identifies potentially sensitive old content and recommends management decisions for that content. To do so, the system will leverage both automatically extracted features about users and individual content, as well as iterative user feedback. In short, our envisioned system will assist users in retrospective data management by reducing the cognitive burden of manually sifting through data.

We present a schematic of our system in Figure 1. At its core, our proposed system has three key components: (i) a feature-extraction engine, (ii) a machine-learning engine, and (iii) a visualization engine.

**Feature-extraction engine:** Users will enable the feature-extraction engine to access their online archive via APIs or similar platform-provided mechanisms. This engine will collect data about both users and their content. We imagine this data will include the features of the content itself,
content metadata (e.g., timestamp, existing access-control setting), user-uploaded profile information, and a temporal history of the user’s interactions with the online archive. This engine will then extract predefined features, passing this information to the machine-learning engine.

**Machine-learning engine:** Using the extracted features as input, the machine-learning engine aims to predict data-management and access-control decisions a user might want to make. Because such predictions are imperfect, the engine will also solicit the user’s feedback for a small, selected set of old content. This engine will iteratively query the user about desired access-control settings for this selected content, creating a model of the user’s desired privacy preferences. Finally, the machine-learning engine will recommend particular access-control settings.

**Visualization engine:** Because acting on so many files individually would be burdensome, the final component of our system is a visualization engine. The visualization engine presents the recommended access-control settings provided as output of the machine-learning engine. The user can then make the ultimate decision to adopt these recommendations or leave the content and its settings as-is.

Multiple interfaces might be suitable for visualizing these recommendations. Notably, there are several ways in which the set of old content can be ranked, including chronologically, clustered by time, or clustered by content similarity [3].

**Conclusion**
As online archives continue to grow over years, decades, and even generations, reevaluating access-control settings retrospectively will become increasingly important. We have presented our vision for a system to assist users in this retrospective management. In our ongoing research, we are working to implement and evaluate these tools.

**REFERENCES**