Understanding and Improving Usability of Data Dashboards for Simplified Privacy Control of Voice Assistant Data

Vandit Sharma
Indian Institute of Technology Kharagpur
vanditsharma@iitkgp.ac.in

Mainack Mondal
Indian Institute of Technology Kharagpur
mainack@cse.iitkgp.ac.in

Abstract

Today, intelligent voice assistant (VA) software like Amazon's Alexa, Google's Voice Assistant (GVA) and Apple's Siri have millions of users. These VAs often collect and analyze huge user data for improving their functionality. However, this collected data may contain sensitive information (e.g., personal voice recordings) that users might not feel comfortable sharing with others and might cause significant privacy concerns. To counter such concerns, service providers like Google present their users with a personal data dashboard (called 'My Activity Dashboard'), allowing them to manage all voice assistant collected data. However, a real-world GVA-data driven understanding of user perceptions and preferences regarding this data (and data dashboards) remained relatively unexplored in prior research.

To that end, in this work we focused on Google Voice Assistant (GVA) users and investigated the perceptions and preferences of GVA users regarding data and dashboard while grounding them in real GVA-collected user data. Specifically, we conducted an 80-participant survey-based user study to collect both generic perceptions regarding GVA usage as well as desired privacy preferences for a stratified sample of their GVA data. We show that most participants had superficial knowledge about the type of data collected by GVA. Worryingly, we found that participants felt uncomfortable sharing a non-trivial 17.7% of GVA-collected data elements with Google. The current My Activity dashboard, although useful, did not help long-time GVA users effectively manage their data privacy. Our real-data-driven study found that showing users even one sensitive data element can significantly improve the usability of data dashboards. To that end, we built a classifier that can detect sensitive data for data dashboard recommendations with a 95% F1-score and shows 76% improvement over baseline models.

An extended version of this paper that includes the appendices is available for interested readers at https://osf.io/rgk95/.

1 Introduction

Voice assistants like Google's voice assistant (GVA), Amazon's Alexa, Microsoft's Cortana, or Apple's Siri are extremely popular today as they are well equipped to perform multiple tasks on users' voice requests (e.g., searching the internet, calling a friend, or playing music). However, these voice assistants also collect and analyze a lot of user data (e.g., timestamps, audio recordings, transcripts, etc.) to improve their infrastructure across multiple devices (e.g., in both smart speaker and smartphone). Unfortunately, this data can lead to a huge possible privacy nightmare since the voice assistant might be used in private situations. E.g., GVA collects three types of potentially sensitive data—audio clips of conversations, transcripts of conversations, and the ambient location of use. We refer to individual records of these three data types as data elements in this paper.

In this study, we take Google voice assistant (GVA) as our experimental testbed. Previous studies on understanding user perceptions and preferences for data collection by voice assistants (such as [29, 40]) have mainly focused their attention on smart speaker users. However, recent reports [23, 42] have highlighted the significantly greater popularity of smartphone-based voice assistants over their smart speaker counterparts. Intuitively, smartphones are easier to use in more contexts than smart speakers, multiplying potential privacy problems. To that end, the exact same GVA software runs in both Google smart speakers and Android smartphones, effectively aggregating data from both. So, we focus on GVA users and conduct a real user-data driven study to uncover user perceptions regarding GVA-collected data.

Specifically, to counter this problem of sensitive data collection, service providers like Google often provide a dashboard to the users showcasing their GVA collected data (Google's My Activity dashboard). We noted that the dashboard includes data from both smart speakers and smartphones without differentiating markers. However, the efficacy of these data dashboards for controlling privacy in the GVA context is not well-understood. To that end, we unpack user perceptions

and preferences regarding data collection by GVA as well as data dashboards through a two-part survey-based user study. Our overall goal is to assess the usefulness/efficacy of data dashboards. We specially focus on the context of data collected by voice assistants (in smart devices) and investigate the efficacy of these dashboards to enable the privacy goals of users in that context.

Our research questions (RQs), as stated below, are designed to unravel (i) whether data dashboards can indeed facilitate a better understanding of what (possibly sensitive) data VAs collect, and (ii) the particular helpful (or not so helpful) aspects of data dashboards from a user-centric view. Our RQs also investigate how to improve the usability of these data dashboards. In this study, we particularly contextualize our RQs with our focus on GVA. We selected GVA primarily because of the huge user base (boosted by the inclusion of GVA in all Android smartphones). Even though our choice of GVA poses some limitations, (e.g., GVA userbase and dashboard might not necessarily represent all VA users or dashboards), our approach is still useful—findings from our study answer broader questions about helpfulness of data dashboards in general and take a step forward towards improving their usability.

RQ1- How frequently do Android users leverage GVA? What is their GVA usage context?

Most (72.5%) of our participants had been using GVA for around two years or more. 73.75% of participants used GVA frequently, at least a couple of times a week in home, office and car. For the median participant, GVA collected 837.5 data elements. The context of using GVA ranged from getting information to entertainment.

RQ2- What are the user perceptions regarding the data collection and storage practices of GVA? What is their ideal access control preference for Google relative to their social relations for accessing GVA data?

Although the majority (78.75%) of participants were aware that Google collects and stores some form of data, around 40% of users were unclear about the type of data (e.g., audio clips) being collected, signifying superficial knowledge. Interestingly, statistical analysis shows that relative to various social relations (proxemic zones [21]), participants considered Google mainly as a public entity.

RQ3- Do users desire to restrict access of Google to GVA collected specific data elements? Do these access preferences correlate with the data element class or the medium of the data collection?

Participants wanted to restrict Google's access for 121 (18.08%) out of 669 audio clips and 61 (17.03%) out of 358 transcripts presented in our survey, a non-trivial fraction of all collected data. They had similar preferences for data collected by smartphones and smart speakers but felt significantly less comfortable viewing data elements where they did not know or could not recall the device through which it was collected. There were no significant differences in user privacy preferences for data elements from different control and possibly sensitive classes (prepared from previous work [7, 27, 29]), suggesting the inherent complexity of finding sensitive data.

RQ4- Do data dashboards help users to control the privacy of their GVA data? Can we further assist users by automated means to improve their privacy-preserving behavior by improving the data dashboard? How?

50% of our participants did not know about the Google-provided My Activity data dashboard. Most participants found the dashboard easy to use; however, more long-time GVA users expressed a need for assistance in using the dashboard, suggesting a lack of effectiveness for larger amounts of data. Showing users even one sensitive data element collected by GVA using our simple class-based sensitive content detection system made them highly (80%) likely to control their collected data. This suggests that assisting dashboard users through automated means might improve their privacy-preserving behavior. We took the first step in this direction by exploring an Machine learning (ML)-assisted human-in-the-loop (HITL) based design for data dashboards. We show that it is possible to create Machine learning-based systems to recommend sensitive content with more than 95% F1-score showing a concrete, feasible direction to improve data dashboards. We note that, although we used GVA as our experimental testbed, our findings regarding the efficacy of data dashboards as well on improving data dashboards could be extended to contexts concerning other VAs. For example, our results show that dashboards are indeed useful for tracking VA-collected data. However, there also is a need for automated assistance in using the dashboards, notably for long-term users to control the privacy of large amounts of accumulated data. These findings are potentially useful for designing improved data dashboards for any VAs.

The rest of the paper is organized as follows- The background and related work in Section 2. Our methodology is explained in Section 3. We describe the data analysis in Section 4. The survey results are presented in Section 5. In Section 6, we explore ML as a possible improvement to recommend sensitive data elements in data dashboards. Finally, we conclude the paper in Section 7.

2 Background and Related Work

GVA capabilities and usage: GVA is an intelligent voice-activated assistant software that Google introduced in May 2016 [28]. It allows users to perform a variety of actions such as getting local information, playing media, performing a Google search, managing tasks, and more [38] through simple voice-based commands. GVA supports cross-device functionality and is available on a wide range of devices such as smartphones, tablets, smartwatches, TVs, headphones, and more [38]. As of 2020, GVA is available on more than 1 billion devices, spread across 90 countries and supports over

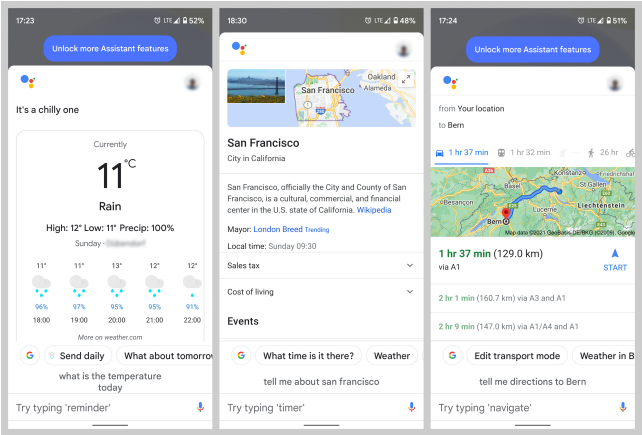

Figure 1: GVA interface on an Android smartphone.

30 languages. It has more than 500 million monthly active users [9], reflecting its immense popularity. Figure 1 shows a visual of the GVA interface and some functionalities.

While most users of VAs use them on smartphones, tablets, and smart speakers [11], recent privacy studies have been paying increased attention to concerns surrounding the use of smart speakers [34], primarily because they are always listening devices. However, several works point out the popularity of VA residing in smartphones which can capture more diverse contexts and potentially private data. This work focuses on GVA, which runs on both smart speakers and smartphones.

Privacy concerns with VAs: Privacy concerns surrounding voice assistants have been studied extensively. Several researchers have proposed different approaches to launch privacy attacks against voice assistants [4, 10, 45, 46]. Schonherr et al. explored the accidental triggering of voice assistants [39], and Edu et al. conducted a detailed literature review of Smart Home Personal Assistants (SPA) from a security and privacy perspective [17]. They highlighted several key issues such as weak authentication, weak authorization, profiling, etc. Edu et al. also studied various attacks on SPAs, suggested countermeasures, and discussed open challenges in this area. Courtney further summarized various privacy concerns associated with voice assistants [15].

In recent years, there have been multiple instances of data leaks associated with voice assistants managed by prominent technology companies [31, 33]. Such data leaks can be a huge cause of privacy concern since VA collected data can include sensitive data such as audio recordings, location data, etc. Specifically, one interaction with GVA can lead to multiple data elements—audio clip, transcript, and location—depending on the controls set in Google-wide settings (e.g., Web & App Activity control, which is turned on by default and enables Google to store transcripts, location, and other metadata for all interactions). For instance, Kröger et al. discussed the threat of unexpected inferences (such as obtaining the speaker's identity, personality, age, emotions, etc.) from audio recordings stored by microphone-equipped devices through voice and speech analysis [24]. The two major classes of entities that can cause privacy violations with collected VA data are (i) the technology companies who own the voice assistants and store data on their servers (e.g., Google, Amazon, etc.), and (ii) external third parties with access to collected VA data, upon which technology companies might rely to review collected data. These two entities comprise our threat model.

Managing privacy of VA-collected data: A possible alternative to prevent privacy violations is limiting and controlling the data collected by voice assistants. Over the years, researchers have proposed several techniques for this purpose, involving both hardware and software. Champion et al. developed a device called the Smart^2 Speaker Blocker to address the privacy and security concerns associated with smart speaker use [12]. The device functioned by filtering and blocking sensitive information from reaching the smart speaker's microphone(s). However, such an intervention cannot be used in the case of smartphone-based voice assistants since smartphones are portable devices. Vaidya et al. proposed another technique to limit privacy concerns (such as voice spoofing) by removing certain features from a user's voice input locally on the device [41]. Qian et al. additionally presented VoiceMask, a robust voice sanitization and anonymization application that acts as an intermediate between users and the cloud [36]. Earlier work also developed a user-configurable, privacy-aware framework to defend against inference attacks with speech data.

While these techniques may be effective in checking privacy concerns, a significant downside is that they modify the collected data, rendering it unusable by developers. This defeats the purpose of collecting data in the first place since voice assistant developers need user data to train better ML models and improve their services. Tabassum et al. presented this as a privacy-utility trade-off, suggesting the development of privacy frameworks that allow users to control the amount of data collected by the voice assistant (in exchange for possibly limited services) [40]. The survey conducted by Malkin et al. on understanding the privacy attitudes of smart speaker users also highlighted a demand for effectively filtering accidental recordings and sensitive topics [29]. We take a first step in this direction by exploring a human-in-the-loop design to identify and recommend sensitive data to GVA users.

Privacy dashboards: Following up on the recommendations made by the Abramatic et al. for better user privacy control [2], Irion et al. advocated the use of privacy dashboards as a practical solution to enhance user control for data collected throughout the online and mobile ecosystem, including platforms such as GVA [22]. They also highlighted the potential of AI techniques and methods to users manage and enforce their privacy settings. In this area, Raschke et al. presented the design and implementation of a GDPR-compliant and

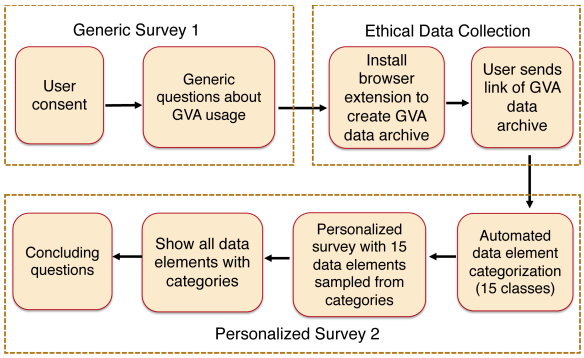

Figure 2: Three key sections of our study—generic Survey 1, ethical data collection and personalized Survey 2.

usable privacy dashboard [37]. In fact, to that end, Feth et al. proposed generic requirement and quality models to assist companies with developing privacy dashboards for different domains [20]. Our research is motivated by this prior work on the importance of privacy dashboards—we aim to uncover the efficacy and possible improvement of today's privacy dashboards by specifically focusing on a deployed system in our real-world data-driven study.

Google My Activity data dashboard: The Google My Activity dashboard is the primary data privacy control provided by Google for all its products and services. It is a hub where users can see and modify all of the key information that Google has been collecting about them over the years [35]. Figure 13 of Appendix C shows the user interface of the Google My Activity dashboard. Since GVA's launch in 2016, not much work has been done on studying user perceptions and the utility of such data dashboards to manage data privacy.

A recent and closely related study [19] investigated user perceptions and reactions to the Google My Activity dashboard. Through a survey, this study showed that viewing the My Activity dashboard significantly decreases concern about Google's data collection practices. However, the authors were unsure if the dashboard actually provided valuable assistance in reviewing the collected data and enforcing user privacy. The first part of our study partially revisits this work. However, we answer several additional questions about the perceptions and preferences of GVA users. We also answer some questions raised by this study, such as the effectiveness of the Google My Activity dashboard to enforce user privacy. We then present a possible solution to improve data dashboards—through recommending sensitive data elements to users. We also demonstrate a highly accurate proof-of-concept human-in-the-loop-based machine learning model for the same.

3 Methodology

We conducted a two-part survey-based user study to unpack perceptions and privacy preferences regarding GVA.

3.1 Recruitment and Inclusion Criteria

We deployed our study in the crowdsourcing platform Prolific Academic [1] during September 2020. We recruited 18+ years old US nationals with >95% approval rating on Prolific. Additionally, we required that our participants primarily used an Android smartphone, used GVA more than once per month in the past year, and were willing to install our browser extension to share their GVA-collected data. We took the help of a short screening survey (Appendix A.1) which took less than a minute with Prolific-suggested compensation of $0.11 to identify potential participants. Ultimately, we invited 249 participants (who satisfied our inclusion criteria) for participating in our actual two-part study. Survey 1 and Survey 2 of our study (seven days apart) took a total of 52 minutes and 42 minutes on average, respectively. We compensated participants who completed both the parts with $12 ($5 for part 1 and $7 for part 2). In total, 80 participants completed both surveys (out of 119 who responded to our initial invitation). This drop in the number of participants was potentially due to the task description and eligibility in the recruitment text. We consider the data from only these 80 participants in this paper as we wanted to combine data from both surveys (and, in effect, connect self-reported general perception from Survey 1 with the user feedback on real-world GVA-collected data in Survey 2).

3.2 Overview of Study Design

Figure 2 summarizes our institutional ethics committee-approved study procedure. The study consisted of three main sections- (i) generic Survey 1, (ii) Ethical GVA data collection and (iii) personalized Survey 2. First, participants were explained the study design as well as the exact data they needed to share. The participants who gave us informed consent first took the generic Survey 1. This survey contained generic questions (instrument in Appendix A.2) regarding user knowledge and usage of Android smartphones as well as GVA. After completing Survey 1, participants installed a browser extension developed by us for ethical GVA data collection. Our extension worked entirely on the client-side and helped users create an archive of GVA data and upload it to their own Google account. Then, the participants manually shared a link to the online archive with us. Next, we leveraged an end-to-end fully automated pipeline to fetch participants' shared GVA data and processed the data in a secure computer. No researcher ever manually saw or analyzed the raw data. This processing phase identified possibly sensitive data elements collected by GVA. Then, within seven days of completing Survey 1 and sharing data, we invited the participants to return for a personalized Survey 2 (instrument in Appendix A.3) In Survey 2, we elicited user perceptions of a stratified sample of these possibly sensitive data elements. Since Survey 2 was generated programmatically for each participant using

elements in their own GVA-collected data, we refer to it as a "personalized survey". We also showed each participant all of their possibly sensitive data elements identified by our data processing pipeline in a personalized Google Drive folder (with named files and subfolders for categories) created by us. Finally, we asked the participants about the utility of automatically detecting sensitive GVA-collected data elements. Next, we explain each of the three sections of our study.

3.3 Survey 1

Our participants provided their online informed consent before starting Survey 1. In the consent form, we highlighted the purpose of our study, the specific data we would ask to share, and our privacy-preserving data collection and processing approach. Then, in Survey 1, we first asked participants some general questions to uncover their usage of Android smartphones and GVA. Next, drawing from earlier studies on privacy concerns surrounding voice assistants, we designed a set of GVA usage scenarios to ground the user and uncover experiences with sensitive and even privacy-violating data collected by GVA [7, 27, 29]. These scenarios ranged from "Using inappropriate language" and "Using GVA in places with audible background sounds" to "Accidental activation of GVA" (complete list is in Appendix D). Then, our participants self-reported whether they recalled using GVA in these scenarios and their comfort in such contextual GVA usage. After this, participants responded to questions about their perceptions regarding GVA data collection (in general and under different transmission principles [5]), storage, and access, including a few questions specifically about Google My Activity dashboard. Finally, we concluded Survey 1 by asking questions related to general privacy attitudes.

3.4 Ethical Collection of GVA Data

Given the sensitive nature of the GVA-collected data elements, we wanted to collect it in the most ethical manner possible, as we will describe next. Our data collection protocol and analysis plan were thoroughly evaluated and approved by our Institutional Ethics Committee (equivalent to an IRB). Participants were briefed about the data collection process through the consent form at the beginning of the study.

Deciding on an ethical data collection protocol: We explored several options to collect GVA data ethically from users along with their downsides—a client-side data-analysis approach was infeasible due to the scale of data and computation, a Google password sharing approach encouraged oversharing private data, and approaching Google to analyze user data and performing our study could potentially be perceived as diminishing user agency. We finally asked our participants to use Google Takeout [1], create an archive of *only* GVA data by selecting specific options in the Google Takeout interface, and share the archive file with us after reviewing the data. We created a Firefox browser extension to facilitate data collection—(i) The extension automatically selected the right options in the Google takeout interface (in client browser) to create an archive with *only* GVA data by choosing the right options in the Google Takeout interface). This approach diminished the chances of oversharing (e.g., chances of accidentally adding all their emails). (ii) The extension automatically selected the option provided in Google Takeout to create an archive in a user's own Google cloud storage [2] (associated with Google account). A participant shared their unique link with us to allow processing of their archive file.

Ensuring privacy of our collected data: In our protocol, participant GVA data could only be accessed using unique individual links Moreover, we informed the participants that they could revoke access anytime, All GVA-collected data was anonymous since it did not include any email or names of users. On receiving a link, an automated pipeline checked the validity of the data (using data type and folder structure of the shared data) and invited only participants with valid data for Survey 2. All data processing was automated (no manual exploration of raw GVA-data) and was done in password-protected computers accessible only to the researchers.

Class	Description	med. #
audio-noise	Audio with high background noise	3
audio-non-bkgd	Audio with low background noise	94
audio-multi-spkr	Audio with multiple speakers	36
audio-non-gend	Audio with non-dominant gender speaker	57
audio-grammar	Audio with grammatical error	9
audio-non-eng	Audio with non-standard English word	51
audio-regret	Audio with regret word	25
audio-neg-sent	Audio with negative sentiment	32
audio-rand	Audio not in the above categories	-
transcript-typo	Transcript with grammatical error	5
transcript-non-eng	Transcript with non-standard English word	28
transcript-regret	Transcript with regret word	20
transcript-neg-sent	Transcript with negative sentiment	19
transcript-rand	Transcript not in the above categories	-
location	Location data (e.g., latitude-longitude)	10

Table 1: Description of fifteen classes from Survey 2 for classifying GVA collected audio clips, transcript and location data. The last column signifies the median number of data elements per user for our participants.

3.5 Survey 2

Our personalized Survey 2 primarily involved eliciting user reactions regarding specific data elements collected and stored

[1] A Google service that enables users to export part or all of data elements stored in Google servers in an archive file [43].

[2] We took due consent to store and share this data in participants' personal cloud storage. The consent form is in Appendix A.2.

by GVA. We start with our data processing pipeline to select data elements for Survey 2.

Creating a classifier to categorize data elements: We identified (and leveraged in Survey 1) a set of privacy-violating scenarios where according to earlier work, potentially sensitive data might be collected by GVA [7, 27, 29]. We analyzed these scenarios to create twelve classes that encompass potentially sensitive GVA-collected data elements. These data elements were broadly of two types—GVA-collected audio clips and transcripts of the commands given to GVA. Aside from these twelve classes, we considered three additional classes—a separate class "location" for location data, and two separate classes ("audio-rand" and "transcript-rand") which identify audio clips and transcripts not belonging to any of the twelve classes and act as a baseline for data elements. These total fifteen classes are presented in Table 1.

We then created automated classifiers to categorize data elements in each of these classes for each user. These classifiers primarily relied on off-the-shelf signal processing (e.g., measuring Signal to Noise Ratio or detecting the number and gender of speakers) and NLP techniques (finding a grammatical error, non-English word or negative sentiment). We created one binary classifier for each of the above-mentioned twelve classes in Table 1 (aside from "location", "audio-rand" and "transcript-rand" classes). These classifiers categorized GVA-collected audio clips and transcripts into one or more of these classes. The motivation and detailed description of each classifier is in Appendix D.

Selecting individual data elements for Survey 2: Once we classified each data element into one or more categories (with the help of classifiers) from Table 1, we used a stratified sampling approach. In short, we randomly selected one data element from each category (without replacement) and used them to create the Survey 2 questionnaire. We also created a personalized Google Drive folder for each participant to review in Survey 2. The folder contained all possibly sensitive data elements found in their GVA data, arranged in thirteen respectively named files and subfolders (excluding "audio-rand" and "transcript-rand" categories). Note that our pipeline handled all of the above tasks automatically. Once personalized Survey 2 was generated, one researcher manually invited the corresponding participant (within seven days of data upload) to participate in Survey 2 via messaging on Prolific.

Overview of Survey 2: We created a personalized Survey 2 (instrument in Appendix A.3) for each participant using at most fifteen selected data elements, depending on the presence/absence of a particular class. During the survey, we first showed these data elements randomly to the participants and correspondingly asked some related questions, e.g., what are the contents of the data element and how comfortable is the participant in sharing it with people in different proxemic zones [21] as well as Google. Note that participants were not provided with any clue about the possibly sensitive nature of these data elements at this stage of Survey 2. Next, we gave participants a brief explanation of the respective classes from which the data elements for their personalized survey were selected. The participants also rated the accuracy of those explanations. Then, to demonstrate the possible output of automated techniques to uncover sensitive data elements, we asked participants to review a personalized Google Drive folder with named files and subfolders containing categorized possibly sensitive GVA data. Then we asked questions to measure user awareness about GVA after seeing this folder. We concluded by asking about the utility of an automated system for detecting sensitive GVA-collected data. In the end, we gave instructions to uninstall the browser extension.

3.6 Limitations

First, our study is limited in recruitment since we recruited only US Prolific users who primarily use Android smartphones and are familiar with GVA. In other words, we might have chosen primarily English speaking users who are also more frequent Android and GVA users than average. However, US-based users are still an important portion of the GVA user base and any privacy issue uncovered by exploring experienced GVA users possibly also affects lesser experienced users. Second, we focused on GVA users who also used an Android device as their primary smartphone. Since GVA is also available in iOS and third-party IoT devices, we might have missed those users. However, this is expected since, in this study, we aimed to investigate the most prominent users of GVA—Android users (GVA is installed by default in Android, unlike iOS). Consequently, some of our survey participants' perceptions about data collection might not be representative of data collected by other voice assistants, which might be used in a different context (e.g., a voice assistant integrated into a children's toy Third, a few of our participants might consider some of our questions as probing based on both language of the question and their prior experience—introducing bias in some of our self-reported data-based results. Lastly, our results might have underestimated privacy needs as very privacy-sensitive individuals would be unlikely to participate in a study that aimed to investigate their GVA data. Not covering such privacy-sensitive individuals is a common concern with user studies related to privacy [30]. However, we strongly feel that this work is still valuable since we unpack common privacy perceptions of GVA users regarding their data and identify possible avenues to improve data dashboards and simplify privacy controls for this data.

4 Data Analysis

We performed both quantitative and qualitative analyses of participants' survey responses. In this section, we briefly elaborate on our data analysis process.

Qualitative open coding: We performed qualitative open coding to analyze free-text responses [25]. First, an author analyzed the responses to each question and created a codebook. Next, two researchers independently coded the responses using this shared codebook. Across all questions, Cohen's kappa (inter-rater agreement) ranged from 0.769 to 1.0 signifying near-perfect agreement. At last, the coders met to resolve disagreements and finalized a code for each response.

Quantitative statistical analysis: To gain more insight into the collected quantitative data, we performed several statistical tests [16, 18]. When the independent variable was categorical, and the dependent variable was numerical, we found all distributions were non-normal (using the Shapiro Wilkes test) and nearly all independent variables with more than two levels. Therefore, we decided to opt for the Kruskal Wallis test for comparing distributions in such cases. When both independent and dependent variables were categorical, we used either the χ^2 test or Fisher's exact test (when individual cell values in the contingency table were < 5) to find significant correlations. We also used difference in proportions as a measure of effect size in our analysis. Apart from statistical tests, we used standard evaluation metrics such as accuracy, precision, and recall to test our prediction model [3, 8].

5 Results

In this section, we present results from our study on understanding user perceptions and privacy preferences regarding GVA data. Unless otherwise specified, results in this section will correspond to self-reported data and not actual usage data. In specific analyses (e.g., sharing comfortability) involving audio clips and transcripts from Survey 2, we sometimes discounted very few elements due to lack of user feedback.

5.1 Participants

A total of 80 participants completed both Survey 1 and Survey 2. We start by checking the basic demographics of those participants in this section.

Basic demographics: Our participant pool had a slight gender bias—68.8% self-identified as male, 30% as female, and 1.2% as non-binary. In terms of age—30% were 18-24 years old, 31.3% were 25-34 years old, and 26.3% of the participants were between 35 and 44 years. Our participants self-identified themselves with several ethnicities—66.3% reported themselves as White, 13.8% as Asian or Pacific Islander, 8.8% as Black or African American and 6.3% as Hispanic or Latino. The rest had mixed ethnicity. The majority of our participants were employed—47.5% employed full-time and only 20% identified as students. In our sample, 53.75% of participants had a bachelor's degree or higher, and only 30% were associated with computer science or a related field. Overall, our participants came from a wide demographic spread.

Usage of Android smartphones: Even though we did not specifically attempt to recruit long-time Android users, 91.3% of our participants reported using an Android smartphone for three years or more. Furthermore, 90% of participants also mentioned using their current Google Account on Android smartphones for three years or more. We had an active Android-user sample— 61.3% of participants used their smartphones daily for 2 to 6 hours and 26.3% for 6 to 10 hours and 6.3% daily for more than 10 hours. The participants used different smartphone applications—54.5% participants had more than 50 apps on their phones at the time of the study. The majority of our participants were familiar with advanced Android features such as rooting, developer options, and launchers (over 70% participants for each). In our sample, 87.5% of participants owned devices running recent Android versions (9 or 10) manufactured by nine different manufacturers. Overall, our participants were long term Android users, well aware of advanced features, and had moderate to high daily usage.

5.2 Characterizing GVA usage (RQ1)

In this subsection, we present results on general usage patterns of GVA as well as the context for such usage.

General usage: 72.5% of our participants were long-time GVA users, with 43.8% participants using GVA for three years or more, and 28.8% using it for two years and 17.5% for a year. In terms of usage frequency, 43.8% of participants used GVA at least once a day, 30% used it a couple of times per week, and the remaining 26.2% of participants used it once a week to a couple of times per month. Participants used different methods to activate GVA (with some using multiple methods)—76.3% of participants used a hotword (e.g., "OK Google"), and 56.3% activated GVA by touching, pressing, or holding buttons on their device. Interestingly, 97.5% of participants used GVA in three broad zones: home, office, and car encompassing both professional and personal lives. Additionally, 38.8% of participants also reported using Google smart speakers. Using actual usage data collected in part 1 of the study (as described in Section 3.4), we found that interactions with GVA resulted in 138,874 data elements stored in Google's servers. The median participant had 837.5 data elements, signifying non-negligible usage of GVA. An overview of participants' GVA data is in Table 2 and year-wise statistics are in Table 10 of Appendix C.

Understanding context of GVA usage: To understand the context for using GVA, we analyzed participant responses to the question-*For what purposes do you use Google Assistant on your Android smartphone?* from Survey 1. The common reasons for using GVA were getting local information (50), communicating with others (29), resolving a query (28), playing audio and video files (27), navigating through devices (25), controlling other devices (24), entertainment such as

Data element	Total	Min.	Median	Max.
# Audio w/ transcript	83,635	12	354	19,451
# Only transcripts	55,243	0	273.5	4671
# Ambient location	84,309	1	407.5	12,504
Total # data elements	138,878	16	837.5	22,073
Age of data (yrs.)	NA	1	3	8

Table 2: Overview of participants' GVA data.

games, jokes, etc. (16), and planning their day (14). Thus, participants used GVA for a wide number of purposes.

Usage of GVA in smartphones: For each of the 1,027 data elements (audio and transcript) presented in Survey 2, we asked participants to choose the device that, according to them, collected each data element (GVA can run in multiple devices). Participants reported that 494 (73.9%) out of 668 audio clips were collected by GVA installed on smartphones, whereas smart speakers collected only 92 clips, indicating a bias towards smartphones for GVA usage. For a non-trivial 81 clips, participants either did not recall or even did not know. The results are similar for transcripts where 229 (63.78%) out of 359 transcripts were collected by GVA on smartphones, and smart speakers collected 57 transcripts; the participants could not recall or didn't know the source for the rest. Thus, most data elements presented in Survey 2 were collected by GVA on smartphones. We note that this bias towards GVA use on Android smartphones could be because of our inclusion criteria since we recruited users of Android, which has GVA pre-built into it. Still, our finding hints at an important domain of heavy data collection by GVA on smartphones in a wide variety of contexts.

Summary: 73.8% of our participants used GVA frequently (at least a couple of times per week or more). The majority of GVA usage happened in smartphones in multiple contexts, and our median participant contributed a total of 837.5 data elements. The median age of GVA data was 3 years.

5.3 User Perceptions of GVA Data (RQ2)

Next, we check whether participants understood how Google handled any GVA-related data. Specifically, we investigate user perceptions regarding GVA data collection and storage using data from Survey 1.

Perceptions of overall data collection: First, we identified (using the Google account of the authors) that there are seven different types of data collected (at most) by GVA. We verified these types in our automated data collected phase too. Table 3 shows the seven different types of data collected (at most) by GVA. The top three are the most obviously sensitive data types (audio, transcript, location), whereas the rest can be considered metadata. Recall that we referred to these three obviously sensitive data types as data elements in this work. To understand the awareness about GVA data collection, we

Data Type	Description
audio	Audio clip of conversation
transcript	Transcript of conversation
location	Ambient location at the time of use
date	Date of conversation
epoch	Time of conversation
noti	Notifications sent by GVA
trig	Activation method (w/ or w/o hotword)

Table 3: Types of data collected by GVA

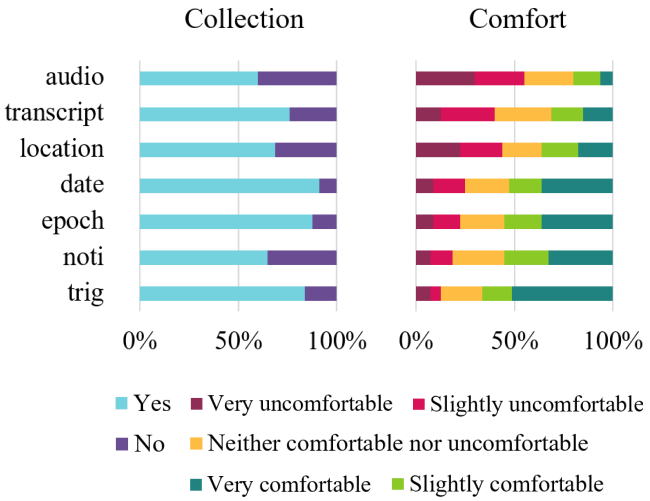

Figure 3: Participant awareness of collected data types and comfort in sharing them with Google. Data types perceived to be collected/not-collected correlated with participant comfort in sharing with Google (Fisher's exact, $p < 0.0001$).

asked participants-*Do you think that Google Assistant on your Android smartphone collects any kind of data while you are using it?* in Survey 1. 78.8% of our participants responded affirmatively with "Yes", and an additional 20% responded "Maybe", signifying the participants are well-aware of possible data collection by GVA.

Perceptions of specific data collection: However, then we dug deeper and asked -*What pieces of data do you think are collected when you use Google Assistant on your Android smartphone?*, showing participants the list given in Table 3. Most participants expressed that GVA collected data such as the date (89.1%) and time (85.7%) of conversations as well as the activation method (81.5%). Interestingly, comparatively fewer participants were aware that GVA collected sensitive data types such as the transcripts of conversations (73.9%) and the ambient location (70.58%). Just 61.3% of participants believed that GVA collected audio clips of conversations, implying that a non-trivial 38.7% of users were unaware about the collection of audio clips by GVA.

Correlation between awareness of data collection and comfort in sharing data with Google: Next, we asked partici-

pants to indicate how comfortable they would feel if GVA collected data from each of the seven data types. The top three data types where most participants felt most uncomfortable to share with Google were audio clips of conversations (58.8%), transcripts of conversations (45.4%), and ambient location (45.4%). The top three data types where most participants felt comfortable with data collection were activation method (66.4%), notifications (56.3%), and time of conversations (48.7%). Next, to check if the data types that most participants felt uncomfortable with being collected were also the ones participants were least aware of, we performed a correlation analysis. Figure 3 presents the analysis result—participant awareness about the collection of each data type and their comfort level in sharing the data type positively correlated. These results signify that the data types for which people were less aware of collection (e.g., audio clips), people were also less comfortable with them being collected. This result indicates a superficial understanding of GVA data collection. We surmise that this shortcoming might cause a decreasing interest in GVA users to delete the GVA collected data via existing privacy controls—e.g., deleting or even browsing their stored data through the data dashboard.

Perceptions of data storage: We asked our participants *Where do you think the data, if collected by Google Assistant on your Android smartphone (and voice-enabled Google smart speakers) is stored?* 86.3% of our participants correctly responded that the data is stored on Google data storage facilities (servers). However, 10% of participants responded that the data is stored only on the respective device, whereas 3.7% of participants responded that the data is stored completely or partially in both places. So, the majority of participants had a clear idea of about data storage practices of GVA.

Summary: Most (78.8%) participants thought that Google collected some data using GVA, and the majority were aware of where this data is stored. However, their awareness about the type of data stored was lacking—a non-trivial fraction was unaware of the collection of sensitive data types. In fact, the participants were uncomfortable sharing the data elements they were not aware GVA was collecting (e.g., audio clips).

5.4 Unpacking Preferred Access Control for Google to Collect GVA Data (RQ2)

Most participants were aware that Google collects and stores some data using GVA in their servers. Thus, we investigated the desired access control rules for Google in the context of specific classes of GVA data elements. We analyzed participant responses to this question in Survey 2 for specific data elements—*After going through the audio clip/Google Assistant command, how comfortable would you feel if someone in your intimate/private/social/public relations/Google heard it/came to know about it?*. This question checked the sharing (i.e., access control) preferences for GVA data with people in four proxemic zones–intimate, private, social public [21] as well as Google. Then, we used statistical analysis to check the proxemic zone closest to Google in terms of these sharing preferences. A Fisher's exact test found that there was a statistically significant correlation between desired access rules for Google and all proxemic zones (Fischer's exact $p < 0.05$) across all classes of data elements from Table 1. Then we used *difference in proportions* as a measure of effect size on 2×2 contingency tables containing comfort data elements between a proxemic zone and google (one table for each class of data element) [18]. For each class, the zone(s) with the smallest effect size had the closest sharing preference with Google. The average effect size for each proxemic zone across all classes revealed that participants associated Google most closely with the public zone (average effect size 0.81) and farthest from the private zone (average effect size 0.92). Table 9 (Appendix C) contains all the effect sizes.

Summary: Across all specific data elements, our participants expressed that the access control rules for Google should be similar to a public entity. To understand how this observation translates to actual user behaviour, we now analyze user privacy preferences for sharing specific files with Google.

5.5 Desired Privacy Preferences of GVA Data (RQ3)

For specific data elements across different classes presented in Survey 2, we asked participants if they are comfortable sharing specific data elements with Google today.

Users want to restrict access of Google for specific GVA data: Our participants were uncomfortable sharing 121 (18.1%) out of 669 audio clips and 61 (17%) out of 358 transcripts (presented to them in Survey 2) with Google. These numbers indicate that participants felt uncomfortable sharing a non-trivial fraction of their GVA-collected data. Next, we will check the correlation between this preference with the medium of data collection and class of data.

Correlation with medium of data collection: We checked the correlation between the device of data collection with user comfort to share data. Figure 4 presents this result. Our statistical test did not reveal any significant difference in comfort across data collected via GVA on phones or smart speakers. However, we did find a significant correlation ($p = 0.00$) between user knowledge of the medium of data collection and user comfort in sharing the data element (both audio and transcript) with Google. Participants felt significantly more comfortable sharing data elements where they knew about or could recall the origin of the collected data element. Interestingly, today, Google My Activity Dashboard only shows whether a data element was collected by a Google smart speaker, completely ignoring smartphone devices. Future dashboard designs could add these missing details to make users more comfortable while viewing their data.

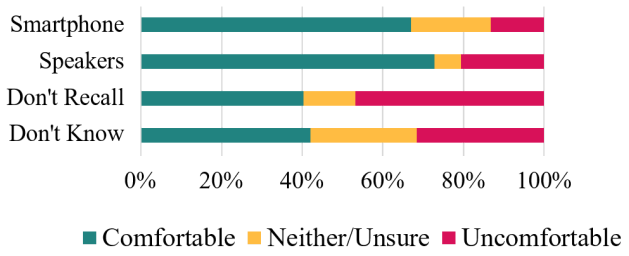

Figure 4: User preferences for sharing audio clips collected by different devices with Google. Preferences for transcripts followed a similar trend.

Correlation with auto-detected classes of data element: Participants felt most uncomfortable sharing audio clips containing regret words (24.4% of all elements from that class), followed by audio clips having multiple speakers (21.1%) and transcripts containing regret words (20.8%). However, a Kruskal Wallis test revealed no significant differences between privacy preferences for data elements belonging to different classes. This result implies that perhaps these simple classes (often based on word matching) were insufficient to accurately identify the GVA-collected data elements where participants wanted to restrict access.

Summary: Participants want to restrict access for a non-trivial fraction of their GVA-collected data, which underlines a need to simplify data dashboards for identifying such data elements. Interestingly, showing users the origin of their collected data made them more comfortable sharing their data. Our simple NLP and signal processing based classes were unable to capture sensitive data elements. Thus, we need more complex tools to identify such sensitive data elements that users are uncomfortable sharing.

5.6 Understanding Utility of Data Dashboard for privacy control (RQ4)

As highlighted in the previous subsection, GVA users wanted to restrict access to 17.7% of GVA-collected data. Today the *My Activity Dashboard* is the only Google-provided way (aside from legal interventions) for the users to delete this data (albeit post-facto) and control privacy. Therefore, we sought to understand user perceptions regarding the utility of this dashboard.

Perceptions of data accessibility: First, we checked whether participants were aware that they could access the data collected by GVA. 61.3% participants believed that the data could be accessed, while 32.5% responded that it might be accessed, indicating the majority are at least aware of the possibility of a tool like the data dashboard.

Popularity and usage of data dashboard: Now, we check if participants knew about the My Activity dashboard. 40 (50%) participants responded that they had heard of it, 10% were not sure, whereas a surprisingly high 40% of participants had never heard of the data dashboard. Among the 40 participants who had heard of the My Activity dashboard, 4 participants had never visited it, and 33 of the remaining 36 participants visited their dashboards less than once per month. In effect, only 3 (3.8%) out of 80 participants visited their dashboards more than once a month. According to the 36 participants who had visited their respective dashboards before our study, the top reasons for visiting it were–(i) To simply check it out (30), (ii) To view collected data (18), (iii) To change activity settings (11), and (iv) To delete some data (9). We asked 50 participants who were unsure/unaware to visit the data dashboard to check their GVA-collected data before continuing with the study.

Unpacking perceptions of data dashboard: Since all participants had explored their dashboards at least once by this point in the study, we asked them how comfortable they felt while viewing the data on their dashboard and why. On a scale of 1 to 5 (1 being very uncomfortable), the average comfortability rating was 3.26 (σ = 1.06, median = 3), indicating that most participants felt neither comfortable nor uncomfortable viewing the data. In fact, our qualitative analysis revealed that participants had mixed reactions to GVA data presented in the dashboard. 37.5% of our participants were either bothered or surprised by the information collected. For instance, P31 said, "*I know that google is collecting information, but I am not 100% comfortable to see the amount they collect. There is really no privacy.*" 15% of participants were glad that the data was available, and 6.25% of participants were unsure of their choice. The remaining 41.25% of participants were neither bothered nor surprised. For example, P12 told, "*I already know Google was collecting all of the information I saw.*" Using a Kruskal Wallis test ($p = 0.029$), we found that participants who had heard of the Google My Activity dashboard before the study were more comfortable (N = 40, μ = 3.525) in viewing collected data, as compared to participants who had never heard of the dashboard (N = 32, μ = 2.875). Therefore, participants grew more comfortable with the dashboard as they became more familiar with it. Our results strongly support the need for data dashboards since participants feel more in control (and thus comfortable) when they can see and manage their data.

Understanding usability of data dashboard: To check the usability of the data dashboard, we asked our participants the question-*How easy was it to reach and find what you were looking for?*, using a 5-point scale (1 being very difficult and 5 being very easy). The average rating was 4.025 (σ = 0.899, median = 4). Thus, most participants found My Activity Dashboard easy to reach. To get a better idea of any difficulties faced by participants during navigation, we then asked them the question-*Did you face any difficulties or problems in navigating through your Google My Activity Dashboard?* Qualita-

tive analysis showed that 8 (10%) out of 80 participants found it hard to navigate through the dashboard to find their data. For example, P62 said, "*There is so much data it is a little overwhelming.*" 2 participants did not check their dashboards, and 1 participant faced some problems with navigation but did not elaborate on it. The remaining 69 participants did not report any difficulties with navigation. 16.25% of participants said they would like some assistance in using the dashboard, and another 10% told that they might want some assistance. The remaining 73.75% of participants indicated that they would not like any assistance. We found a positive correlation (Fisher's exact test, $p = 0.048$) between the duration of using GVA (less or more than around 2 years) with the need for assistance in using the dashboard, highlighting a possible cognitive overload for long-time GVA users.

Summary: We found an interesting knowledge gap within our participants—93.8% of participants thought their GVA-collected data can or may be accessed. However, only 50% of the participants were aware of the data dashboard, showing a lack of actionable knowledge. Even the people who knew about data dashboards, just 3.75% visited it regularly. In fact, more than one-third participants (37.5%) felt bothered or surprised while viewing the collected data. Quite assuringly, most participants found the dashboard easy to use; however, 10% of participants also found it difficult to access their data. We observed that more long-time GVA users expressed a need for assistance in using the dashboard, suggesting the more they become familiar with the dashboard, the more overwhelmed they might get with the huge data collected by GVA over time.

5.7 Moving towards Improving Utility of Data Dashboards (RQ4)

Currently, Google's My Activity Dashboard provides two ways to delete collected data- (i) users can either inspect and delete each data element individually, or (ii) delete all data elements stored within a specified date-time range. The former is particularly not useful from a privacy perspective because inspecting a large number of collected data elements (most of which are non-sensitive) is quite time-consuming and laborious to be practically feasible, as seen in the previous subsection. On the other hand, the latter can help enforce privacy but is not a good option for users who might want to retain some/all of their collected data for future reference, assisting product development, etc. To assist users with finding their possibly sensitive data collected by GVA, we did a simple proof-of-concept test—around the end of Survey 2, we showed them a personalized Google drive with data elements divided into subfolders according to their auto-detected classes from Table 1 (with classes as subfolder names) and asked if a system which can show such classifications will be useful. Recall that these classes were constructed with privacy-violating scenarios in mind, and hence some of the data elements were expected to be sensitive.

Recommending elements in data dashboards: After participating in the study, 56.3% of participants reported that they were very likely to delete some of their data collected by Google. 65% of participants said that our classifier provided valuable assistance in finding sensitive data on Google servers, and 72.5% told that they would recommend others to try it out if made publicly available. This percentage was 50% higher than the 27.5% of participants who expressed a need for assistance in finding sensitive data on Google servers in Survey 1, indicating a strong demand for such a recommendation system in data dashboards.

The efficacy and challenge in providing recommendations: The primary challenge that we faced while developing our sensitive content detection system was related to the accuracy of the system in assigning classes to the data elements. On a 5-point Likert scale (1 is very inaccurate and 5 is very accurate), the average rating provided by participants to our classifier was just 2.67 (σ = 0.96, median = 3), suggesting that most participants did not find it highly accurate. Additionally, 20 participants provided qualitative feedback regarding the study. 9 out of these 20 participants pointed out that the system accuracy could be improved. For example, P60 stated: "*Overall I found the sensitive content system not to be very accurate. It some cases it was correct, but in more cases it was rather incorrect.*" Despite the low perceived accuracy of the classification, we found it surprising that 65% of participants found it useful to find sensitive content. P71 further explained the connection between classification accuracy and helpfulness of our recommendation system: "*Perhaps try improving the accuracy. I noticed that while it did get some things right, it'd periodically get things wrong. I'm not expecting the system to be perfect though but if you can improve the accuracy at all that'd be great.*"

To better understand this result, we looked at participant accuracy scores and sharing preferences for individual data elements presented in Survey 2. Out of 52 (65%) participants who believed that our system provided valuable assistance in finding sensitive data (i.e. who liked our *classification presentation* potentially irrespective of accuracy), we focused on 36 participants who rated at least one encountered data element $>= 3$ for accuracy and also felt neutral or uncomfortable sharing it with Google. We observed that 29 (80.6%) of these 36 participants found our system to be helpful (i.e. they found our classification accurate). So, our results hinted that even when our classifier was able to detect at least one possibly sensitive file, most participants found it useful, signifying a need for such recommendations.

Summary: Accurate recommendations of possibly sensitive data elements help users restrict access and protect the privacy of their VA-collected data. Encouragingly, recommending users to revisit even one accurately sensitive data element

collected by GVA made them highly (80.6%) likely to control their collected data, highlighting the demand for accurate, usable dashboards. This finding strongly underlines the efficacy of a highly accurate sensitive-element recommendation system to improve the utility of data dashboards. In the next section, we present the feasibility of building such an automated, highly accurate, sensitive content detection system.

6 Feasibility of Accurately Recommending Sensitive Data in Data Dashboards

Earlier, we saw that voice assistants like GVA collect and store large amounts of user data. While deleting all collected data in bulk can help avoid privacy violations, in Survey 2, participants mentioned not deleting the shown data elements for 64.8% (674 out of 1040 data elements shown) of cases. Our open coding of their explanations revealed interesting themes—the prominent reasons for not immediately deleting these 674 data elements was that these data elements were non-sensitive (24.0% of data elements), improving Google Assistant or Google services in general (8.2%). For eg., P48 said, "*I don't mind Google having access to clips like this to improve their services.*" Other reasons for not deleting collected data included having the choice to view or delete previously collected data at will (1.4%) and personalized recommendations from Google (0.4%). In the case of better personalized recommendations, P32 explained, "*I don't mind if Google knows hat music I listen to, especially if it improves it's music suggestion service.*" For 17.7% of data elements, participants did not mention any specific reason, but they wanted Google to carry out their processing and delete this data within a time frame (e.g., after 3 months).

Even though companies provide data dashboards for users to access this data (e.g., My Activity Dashboard by Google), current dashboard designs do not offer mechanisms for users to efficiently sift through and restrict access to specific data elements. To that end, our results (section 5.7) hint at a need for an improved human-in-the-loop (HITL) GVA data dashboard design—we envision an interface that can prioritize potentially sensitive content in the dashboard interface and assist users in controlling the privacy of their GVA-collected data. However, auto-detecting and restricting access to sensitive data elements to help users is also challenging as sensitivity can depend on external factors (e.g., user's age, frequency of use, other personal reasons, etc.) aside from the content of data elements. To that end, we explored the feasibility of *recommending sensitive data elements* in data dashboards in a HITL scenario where the recommendation is only to help users find such data elements and not to take away their control. Companies could leverage such recommendations to improve their data dashboards by presenting possibly sensitive data to their users for review. We envision that such data elements can be presented either by creating a separate review section in a dashboard or changing the default ranking of shown content.

6.1 Modelling Sensitive Content Detection as a Supervised Prediction Task

Our prediction task involved predicting whether a user will perceive a particular data element collected by GVA as sensitive. For classification, our training dataset consisted of tuples (X_i, Y_i), where X_i represents the feature vector, and Y_i represents the sensitivity label of a data element i. The prediction task involved binary classification, where $Y_i = 1$ corresponded to the 'Yes' label (sensitive class) and $Y_i = 0$ corresponded to the 'No' label (not sensitive class). The feature vector X_i included audio-based features, text-based features, and user-based features, all of which were captured through user survey responses and shared GVA data. The audio-based features that we used were Mel-Frequency Cepstral Coefficients (MFCC) [44], spectral contrast, tempo, and SoundNet-based features [6]. The text-based features included LIWC-based features, sentence embedding, presence of swear words, presence of regret words, sentiment-based features, emotion-based features, and presence of top 100 unigrams and bigrams. The user-based features consisted of age range and gender of users, age of Google Account, frequency and span of GVA usage, and association with computer science or a related field. The survey responses were included either as one-hot encoding or binary indicators for multiple-choice answers. A detailed description of all features is in Appendix E.

To perform this classification, we explored several established supervised ML algorithms such as Support Vector Machines (SVMs), Logistic Regression (LR), Random Forest (RF), Multi-layer Perceptron (MLP), each from the scikit-learn library [32], along with XGBoost (XGB) [14]. We compared the performance of these classifiers with two baselines. The first one was a random classifier that randomly assigned a label to each data element, where prediction probabilities for labels were chosen based on their prevalence in our dataset. For our second baseline, we used the preliminary categorization (Table 1) of each data element as the input feature to train another XGBoost model called XGB-Class. All model hyperparameters were optimized using grid search with 10-fold cross-validation. We found the XGB model to perform the best and thus use it to report our final performance metrics.

6.2 Our Dataset

Our dataset consisted of 542 audio clips and 412 transcripts. Each data element was associated with one of three sensitivity labels by the users during Survey 2- 'Yes' (sensitive class), 'No' (not sensitive class), and 'I am not sure' (ambiguous). Since data sensitivity is subjective, we considered these user-assigned labels as accurate ground truth for our predictions. The distribution of labels in the dataset was as follows-

18.34% 'Yes,' 63.94% 'No,' and 17.72% 'I am not sure.' We were specifically interested in sensitive data elements (labeled 'Yes') for our prediction task. Since the neutral label, 'I am not sure,' constituted a non-trivial fraction of our dataset, we performed four different experiments, treating it as a separate label each time. In each experiment, we trained our best performing XGB classifier using the 'I am not sure' label as a proxy for one of these four labels- 'Yes' (sensitive class), 'No' (not sensitive class), 'Removed' (do not consider in training process), and 'I am not sure' (treat as a separate class). The first three experiments were binary classification problems, whereas the last one was a three-class classification problem. The best results were obtained associating the 'I am not sure' label with the not sensitive class ('No' label), which semantically also implies a conservative prediction—recommending only data element where the classifier is certain that its sensitive. Hence, we report the results of only this experiment in the paper. The rest of the results are in Appendix B.

6.3 Experimental Setup

We performed 10-fold cross-validation and reported macro-averaged precision, recall, and F1 scores for each classifier. Here, precision is defined as the ratio, TP/(TP+FP), where TP refers to the number of true positive predictions, and FP refers to the number of false positive predictions. Similarly, recall is defined as TP/(TP+FN), where FN refers to the number of false negative predictions. Since our dataset was highly skewed away from the class of our interest, we used the Synthetic Minority Oversampling TEchnique (SMOTE) [13] to balance our dataset before training the models. SMOTE aims to balance imbalanced datasets by oversampling or randomly replicating samples from the minority class. We used the implementation of SMOTE provided by the imbalanced-learn [26] Python library.

Next, we plotted precision-recall (PR) curves (averaged over 10-folds) for each classifier to analyze the trade-off between showing a larger number of sensitive data elements to the users and the accuracy in finding such elements, reflected by recall and precision values, respectively. Maximizing both precision and recall is often not mathematically possible. In practice, a classifier with high precision and low recall returns fewer but relevant results, whereas a classifier with high recall and low precision returns many but irrelevant results. Therefore, achieving a balance between precision and recall is crucial for such classifiers to recommend as many sensitive data elements as possible to the user. A valuable heuristic to capture this trade-off between precision and recall is precision-recall area under curve (PR-AUC). A higher value of PR-AUC for a classifier shows its ability to achieve both good precision and recall. We calculated the PR-AUC values for different classifiers from their PR curves and used it as a metric to quantify their overall performance.

Finally, we used precision@k to assess different classifiers from a recommendation system's perspective. In a practical scenario, it is unlikely that a user will go over all suggestions of sensitive data elements presented on their data dashboard. In such cases, a classifier must sort their recommendations and minimize the number of false positives within top k recommendations. We report this value using precision@k. Precision@k is simply the proportion of correct classifications within top k recommendations. A higher value of precision@k signifies good quality of recommendations.

Model	Precision	Recall	F1 score
Proposed Feature-based Models			
SVM	0.90	0.89	0.89
LR	0.92	0.91	0.91
RF	0.83	0.83	0.83
MLP	0.91	0.91	0.91
XGB	**0.96**	**0.95**	**0.95**
Baseline Models			
Random	0.51	0.52	0.44
XGB-Class	0.54	0.54	0.54

Table 4: Macro-averaged Precision, Recall, F1-score for all models. The highest values in each column are boldfaced.

6.4 Results

Our ML models tried to predict whether a particular data element will be perceived by the user as sensitive or not. Table 4 shows the macro-averaged precision, recall, and F1 scores for all models. Across all models, XGB offered the best performance with an F1 score of 0.95, followed by LR and MLP, each of which achieved an F1 score of 0.91. The best performing baseline model was XGB-Class which achieved an F1 score of 0.54. Our XGB model outperformed the best-performing baseline model by approximately 76%.

Figure 5 shows the PR curves generated for all models. Once again, the XGB model performed the best, achieving a near-perfect PR-AUC value of 0.9894, followed by LR that achieved a PR-AUC value of 0.9283. The XGB model showed a significant improvement over the XGB-Class baseline model (PR-AUC = 0.5452), outperforming it by approximately 81%.

Figure 6 shows the precision@k curves generated for all models. Looking at the top 30 predictions, the XGB and RF models performed the best, achieving a perfect precision@30 value of 1. They were followed by MLP, which achieved a precision@30 value of 0.9. Other models such as SVM and LR had relatively poor precision@30 values comparable to the precision@30 value of 0.57 for the XGB-Class baseline model. To distinguish between the performance of XGB and RF models, we looked at their precision@k values for large values of k. We observed a slight drop in performance while varying k from 1 to 500 for the RF model (precision@500 = 0.954), whereas the XGB model retained its performance (precision@500 = 1) even for larger values of

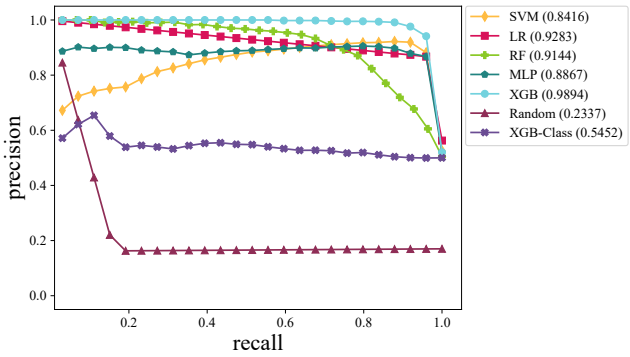

Figure 5: PR curves while classifying data elements (SVM, LR, RF, MLP, and XGB are evaluated ML models, whereas Random and XGB-Class are baseline models)

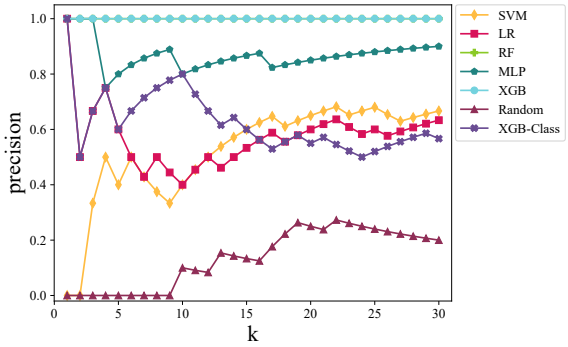

Figure 6: Precision@k curves while classifying data elements (SVM, LR, RF, MLP, and XGB are evaluated ML models, whereas Random and XGB-Class are baseline models)

k, highlighting the stability of XGB model (Figure 15 of Appendix C). Despite achieving a lower F1 score and PR-AUC value than models such as LR, the RF model offered better performance in the scenario where only a few data elements should be presented to users. Although our best-performing XGB model achieved a high F1-score of 95%, resourceful organisations like Google might be able to further improve accuracy in real-world deployments with additional labeled data.

6.5 Understanding Prediction Accuracy

Finally we analyzed the features that played the most important role in our prediction task.

Many of the important features are user-based: Table 5 shows the top ten features identified by our best performing XGB classifier, in decreasing order of importance. Three out of the top five features were user-based, which highlights that user details are crucial in predicting the perceived sensitivity of data elements. Five out of the top ten features were text-based, implying that the text content of data elements is also central to the prediction task. This is in contrast with the result in Section 5.5, where we did not find significant differences in user privacy preferences across simple lexicon-based classes. We believe this contrast is because of using more involved textual features (e.g., LIWC, sentence embedding, sentiment).

Rank	Feature Type	Name
1	User-based	Age of Google Account
2	User-based	Frequency of GVA usage
3	Text-based	Sentiment-based
4	User-based	Age range of user
5	Audio-based	SoundNet-based
6	Text-based	LIWC-based
7	Text-based	Presence of top 100 unigrams
8	Text-based	Sentence Embedding
9	User-based	Association with CS or a related field
10	Text-based	Presence of regret words

Table 5: Top 10 features as decided by the XGB model in decreasing order of importance

7 Concluding Discussion

In this work, we present the first study on understanding users' privacy attitudes and preferences regarding data collection by GVA. Specifically, using a real-world data-driven approach, we unpacked users' knowledge of the data collection practices of GVA. Previous work [19] has looked into general user perceptions and reactions towards the Google My Activity data dashboard. However, we, in contrast, focused on using real-world GVA-collected data elements to elicit specific user responses. We seek to understand whether such data dashboards actually provide utility in controlling data privacy through an 80-participant user study grounded into actual GVA-collected data. Recent studies have paid increased attention to voice assistants on smart home speaker devices. Given the pervasiveness of smartphones, our results show that smartphone voice assistants can collect data in a variety of scenarios different from stationary smart speaker devices. Thus, our work sheds light on the data-centric ecosystem of voice assistants, with GVA as our test case. Furthermore, in spite of using GVA data dashboards as our test case, many of our findings on assessing the efficacy of data dashboards and improving their usability are generalizable to dashboards of other voice assistants.

A new direction towards usable data dashboards: Our results establish a definite need for better data dashboards while acknowledging the utility of the current one. As a first step, our results hint at the fact that users have just superficial knowledge about the data collection and storage practices of GVA. Although data dashboards help to raise awareness about the total collected data by GVA, the huge amount of data does not help. Long-term users would need automated

assistance to review more sensitive data elements. Thus, our results underline a need to make these data dashboards more usable by helping users uncover sensitive data elements. Our user feedback and accurate classification results identify that machine-learning based human-in-the-loop systems might significantly help the cause. To that end, we identified the top ten most important features for this prediction task. In addition to text-based and audio-based features already available to the VA platform, our results highlight that user-based features can also play an important role in identifying sensitive content. We believe that a handful of these user-based features unavailable to the VA platform could be collected by directly asking the users as part of an initial setup process (while mentioning this will assist the users in finding their sensitive data elements). In fact, our second survey, which aimed to raise awareness about different types of data collected by GVA and stored by Google increased user awareness about GVA collected data.

However, there is much left to explore in this direction, including the presentation of these recommendations to the users and checking the efficacy of interface nudges using these recommendations. For instance, our HITL design to improve usability focused on assisting users in uncovering potentially sensitive elements. A potential future work is creating and evaluating a query system in parallel to this recommender system. Such a system could assist users in sifting through the GVA-collected data efficiently and further improve the usability of data dashboards. Thus, we strongly feel our work paves the way to build more usable data dashboards for better assisting users and takes a step forward to bringing transparency to the data ecosystem of voice assistants.

8 Acknowledgments

We thank the anonymous reviewers and our shepherd Camille Cobb for their valuable feedback. We also thank Shalmoli Ghosh for her help with an earlier iteration of this work and Niloy Ganguly for the discussion early in the project. The experiments in this work were funded by Huawei Technologies India Private Limited via the ADUL project.

References

[1] Prolific. https://www.prolific.co/, 2021.

[2] J-F Abramatic, B Bellamy, ME Callahan, F Cate, P van Eecke, N van Eijk, E Guild, P de Hert, P Hustinx, C Kuner, et al. Privacy bridges: Eu and us privacy experts in search of transatlantic privacy solutions. 2015.

[3] Charu C. Aggarwal. *Recommender Systems: The Textbook*. Springer-Verlag, 2016.

[4] E. Alepis and C. Patsakis. Monkey says, monkey does: Security and privacy on voice assistants. *IEEE Access*, 5:17841–17851, 2017.

[5] Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster. Discovering smart home internet of things privacy norms using contextual integrity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2), July 2018.

[6] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, pages 892–900, 2016.

[7] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. Understanding the long-term use of smart speaker assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(3), September 2018.

[8] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

[9] Manuel Bronstein. A more helpful google assistant for your every day. https://blog.google/products/assistant/ces-2020-google-assistant/, 2020. (Last accessed on October 2021).

[10] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden voice commands. In *Proceedings of the 25th USENIX Security Symposium (USENIX)*, Sec'16, pages 513–530, August 2016.

[11] Pew Research Center. Nearly half of americans use digital voice assistants, mostly on their smartphones. https://www.pewresearch.org/fact-tank/2017/12/12/nearly-half-of-americans-use-digital-voice-assistants-mostly-on-their-smartphones/, 2017. (Last accessed on October 2021).

[12] Christopher Champion, Ilesanmi Olade, Konstantinos Papangelis, Haining Liang, and Charles Fleming. The smart^2 speaker blocker: An open-source privacy filter for connected home speakers. *CoRR*, abs/1901.04879.

[13] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[14] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016.

[15] M. Courtney. Careless talk costs privacy [censorship digital assistants]. *Engineering Technology*, 12(10):50–53, 2017.

[16] W.J. Dixon and Jr. Massey, FJ. *Introduction to statistical analysis*. McGraw-Hill., 1951.

[17] Jide S. Edu, Jose M. Such, and Guillermo Suarez-Tangil. Smart home personal assistants: A security and privacy review. *ACM Computing Surveys*, 53(6), 2020.

[18] Paul D. Ellis. *The Essential Guide to Effect Sizes*. Cambridge University Press, 2010.

[19] Florian Farke, David G. Balash, Maximilian Golla, Markus Dürmuth, and Adam J. Aviv. Are privacy dashboards good for end users? evaluating user perceptions and reactions to google's my activity. In *Proceedings of the 30th USENIX Security Symposium*, Sec'21, August 2021.

[20] Denis Feth and Hartmut Schmitt. Requirement and quality models for privacy dashboards. In *2020 IEEE 7th International Workshop on Evolving Security Privacy Requirements Engineering (ESPRE)*, 2020.

[21] Edward Twitchell Hall. *The hidden dimension*, volume 609. Garden City, NY: Doubleday, 1966.

[22] Kristina Irion, Svetlana Yakovleva, Joris van Hoboken, and Marcelo Thomson. A roadmap to enhancing user control via privacy dashboards, 2017. https://hdl.handle.net/11245.1/aec2f33f-5c89-492e-b45c-8b953351754e.

[23] Bret Kinsella. Voice assistant use on smartphones rise, siri maintains top spot for total users in the u.s. https://voicebot.ai/2020/11/05/voice-assistant-use-on-smartphones-rise-siri-maintains-top-spot-for-total-users-in-the-u-s/, 2020. (Last accessed on October 2021).

[24] Jacob Leon Kröger, Otto Hans-Martin Lutz, and Philip Raschke. Privacy implications of voice and speech analysis – information disclosure by inference. In Michael Friedewald, Melek Önen, Eva Lievens, Stephan Krenn, and Samuel Fricker, editors, *Privacy and Identity Management*, pages 242–258. Springer, 2020.

[25] Jonathan Lazar, Jinjuan Feng, and Harry Hochheiser. *Research Methods in Human-Computer Interaction*. Morgan Kaufmann, 2017.

[26] Guillaume Lemaître, Fernando Nogueira, and Christos K Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, 18(1):559–563, 2017.

[27] Yuting Liao, Jessica Vitak, Priya Kumar, Michael Zimmer, and Katherine Kritikos. Understanding the role of privacy and trust in intelligent personal assistant adoption. In *Proceedings of Information in Contemporary Society*, iConference 2019, pages 102–113. Springer, 2019.

[28] Matthew Lynley. Google unveils google assistant, a virtual assistant that's a big upgrade to google now. https://techcrunch.com/2016/05/18/google-unveils-google-assistant-a-big-upgrade-to-google-now/, 2016. (Last accessed on October 2021).

[29] Nathan Malkin, Joe Deatrick, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. Privacy attitudes of smart speaker users. *Proceedings on Privacy Enhancing Technologies*, 2019(4):250 – 271, 01 Oct. 2019.

[30] Mainack Mondal, Günce Su Yilmaz, Noah Hirsch, Mohammad Taha Khan, Michael Tang, Christopher Tran, Chris Kanich, Blase Ur, and Elena Zheleva. Moving Beyond Set-It-And-Forget-It Privacy Settings on Social Media. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, New York, NY, USA, November 2019.

[31] Lily Hay Newman. An alexa bug could have exposed your voice history to hackers. https://www.wired.com/story/amazon-alexa-bug-exposed-voice-history-hackers/, 2020. (Last accessed on October 2021).

[32] 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. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[33] Sarah Perez. Google is investigating the source of voice data leak, plans to update its privacy policies. https://techcrunch.com/2019/07/11/google-is-investigating-the-source-of-voice-data-leak-plans-to-update-its-privacy-policies/, 2019. (Last accessed on October 2021).

[34] PRWeb. Over Half of U.S. Adults Are Now Using Voice Assistants Like Siri and Alexa on Smartphones - New Market Data from Voicebot Research. https://www.prweb.com/releases/over_half_of_u_s_adults_are_now_using_voice_assistants_like_siri_and_alexa_on_smartphones_new_market_data_from_voicebot_research/prweb17536545.htm, 2020. (Last accessed on October 2021).

[35] Sarah Jacobsson Purewal. Everything you need to know about google's my activity page. https://www.cnet.com/how-to/everything-you-need-to-know-about-googles-my-activity-page/, 2016. (Last accessed on October 2021).

[36] Jianwei Qian, Haohua Du, Jiahui Hou, Linlin Chen, Taeho Jung, Xiang-Yang Li, Yu Wang, and Yanbo Deng. Voicemask: Anonymize and sanitize voice input on mobile devices, 2017. https://arxiv.org/abs/1711.11460.

[37] Philip Raschke, Axel Küpper, Olha Drozd, and Sabrina Kirrane. Designing a gdpr-compliant and usable privacy dashboard. pages 221–236, 2018.

[38] Mitja Rutnik. Google Assistant guide: Make the most of your virtual assistant. https://www.androidauthority.com/google-assistant-838138/, 2021. (Last accessed on October 2021).

[39] Lea Schönherr, Maximilian Golla, Thorsten Eisenhofer, Jan Wiele, Dorothea Kolossa, and Thorsten Holz. Unacceptable, where is my privacy? exploring accidental triggers of smart speakers, 2020. https://arxiv.org/abs/2008.00508.

[40] Madiha Tabassum, Tomasz Kosiński, Alisa Frik, Nathan Malkin, Primal Wijesekera, Serge Egelman, and Heather Richter Lipford. Investigating users' preferences and expectations for always-listening voice assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(4), December 2019.

[41] T. Vaidya and M. Sherr. You Talk Too Much: Limiting Privacy Exposure Via Voice Input. In *Proceedings of 2019 IEEE Security and Privacy Workshops (SPW)*, pages 84–91, 2019.

[42] Voicebot. Voice assistant consumer adoption report. https://voicebot.ai/wp-content/uploads/2018/11/voice-assistant-consumer-adoption-report-2018-voicebot.pdf, 2018. (Last accessed on October 2022).

[43] Web Webster. Google takeout: Why you need it and how to use it. https://www.lifewire.com/what-is-google-takeout-4173795, 2020. (Last accessed on October 2021).

[44] Min Xu, Ling-Yu Duan, Jianfei Cai, Liang-Tien Chia, Changsheng Xu, and Qi Tian. HMM-Based Audio Keyword Generation. In *Advances in Multimedia Information Processing*, pages 566–574, 2005.

[45] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. DolphinAttack: Inaudible Voice Commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, page 103–117, 2017.

[46] Rongjunchen Zhang, Xiao Chen, Sheng Wen, and James Zheng. Who Activated My Voice Assistant? A Stealthy Attack on Android Phones Without Users' Awareness. In *Machine Learning for Cyber Security*, pages 378–396. Springer, 2019.