# Data Privacy in Data Mining
# CS 510 Final Project Report

Yiming Zhang
Computer Science
Portland State University
Portland, Oregon, USA
ymzhang@pdx.edu

## ABSTRACT

In recent years, the practice of data mining has become a popular computing technology. Data mining is a means to summarize useful information by analyzing big sets of data [1]. The growing popularity and development of data mining technologies bring serious threat to the security of individual's sensitive information. In this project, the details of data mining will be introduced, and how it influences people's daily life in both good ways and bad in terms of data privacy. In addition, we will look into how big companies like Google, Amazon and Facebook collect and use your data. Last but not least, a case study of COVID-19's impact on data privacy, protection and security will be examined.

**CCS CONCEPTS**

- Data mining
- Security and Privacy
- Sensitive Information
- Anonymization

**KEYWORDS**
Data mining, web application, data, mobile, privacy, security, COVID-19

**ACM Reference format:**
Yiming Zhang. 2020. Data Privacy. In *Proceedings of CS 410/510: Explorations of Data Science. Portland, OR, USA, 9 pages.*

## 1 Introduction

Data mining is a young, important, and increasingly popular field, with the first paper appearing only around 1992. Since then, database researchers have started working on huge amounts of data and scalable algorithm computations. Now data mining can be applied almost anywhere [2].

Amazon.com provides purchase recommendations for each user by using collaborative filtering algorithms, they say "People buying this book also buy other books" [3]. Some newborn tech companies, such as Ditto Labs Inc., use software to scan publicly posted photos [4]. For example, they might look for images of individuals holding a Dr. Pepper drink to determine what logos are in the picture, what facial expressions are present, like whether the person is smiling, and what the scene's context is [5].

In the structured data, people are usually looking for different patterns and interpret the hidden meaning from the numbers, whether to find clusters, regression or evolution [6]. The methods need raw data to support basic calculations, and personal data is sourced from both public databases and private sectors, this is where the ethics of data mining comes into the place.

Artificial intelligence, statistical computation and logistic regression are the basic algorithms for data mining [7], which make it possible to not only rely on numerical data sets, but also other types of datasets, such as text and photo. Data mining plays an important role in analyzing customer information and helps companies relate to their consumers better [8]. However, on the other hand, people concerned about their personal information may be invaded, analyzed and used unknowingly [9].

## 2 Process of Data Mining

The term data mining is often known as another term: knowledge discovery from data (KDD) which indicates the goal of the mining process [10]. To obtain useful knowledge from data, the following steps are performed in an iterative way. (See Figure 1)

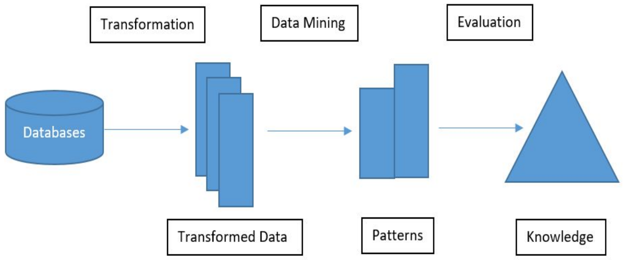


**Figure 1: Overview of Data Mining process**

Step 1: Data preprocessing. Basic operations include data selection (to retrieve data relevant to the KDD task from the database), data cleaning (to remove noise and inconsistent data, to handle the missing data elds, etc.) and data integration (to combine data from multiple sources).

Step 2: Data transformation. The goal is to transform data into forms appropriate for the mining task, that is, to find useful features to represent the data. Feature selection and feature transformation are basic operations.

Step 3: Data mining. This is an essential process where intelligent methods are employed to extract data patterns (e.g. association rules, clusters, classification rules, etc).

Step 4: Pattern evaluation and presentation. Basic operations include identifying the truly interesting patterns which represent knowledge, and presenting the mined knowledge in an easy-to-understand fashion [11].

## 3 Privacy Concerns and Ethics

We have to admit that data mining tools do make our lives easier, but sometimes the tools reveal too much about our personal information which make us feel unsafe, uncomfortable and even creepy sometimes. People have shown increasing concern about the privacy threats posed by data mining. For instance, you just searched some items on Amazon.com and you closed all the tabs. When you opened Amazon.com again, surprisingly, the items related to what you just searched appeared on recommendations. This is because the Amazon page reads your cookies and displays related advertisements (See Figure 2). This is also called Web Personalization or Web Mining. Web mining is a concept that gathers all techniques, methods and algorithms used to extract information and knowledge from data originating on the web (web data). A part of this technique aims to analyze the behavior of users in order to continuously improve both the structure and content of visited web sites. This technique may help the user feel comfortable when they visit a site through a personalization process. However, to some extent, the website may infringe the privacy of those who visit it [12-15].

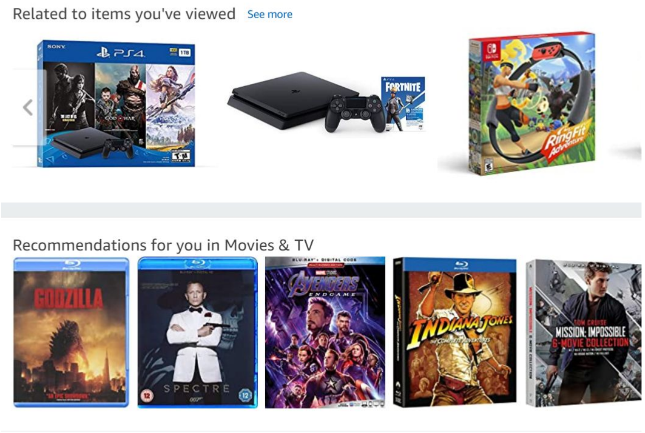

Related to items you've viewed
See more
Recommendations for you in Movies & TV

**Figure 2: Web Personalization on Amazon.com**

Individual's privacy may be violated due to the unauthorized access to personal data. the undesired discovery of one's embarrassing information, the use of personal data for purposes other than the one for which data has been collected [16]. For instance, the U.S. retailer Target once received complaints from a customer who was angry that Target sent coupons for baby clothes to his teenage daughter. However, it was true that the daughter was pregnant at that time, and Target correctly inferred the fact by mining its customer data [17]. From this story, we can see that the conflict between data mining and privacy security does exist.

All we have discussed above is the concerns from data providers. On one hand, the provider should be able to make his very private data which contains information that he does not want anyone else to know, inaccessible to the data collector. On the other hand, if the provider has to provide some data to the data collector, he wants to hide his sensitive information as much as possible and get enough compensation for the possible loss in privacy [18].

From data collectors' view, the data collected from data providers may contain individuals' sensitive information. Directly releasing the data to the data miner will violate data providers' privacy, hence data modification is required. On the other hand, the data should still be useful after modification, otherwise collecting the data will be meaningless. If the data gathered is not accurate enough, marketers are more likely to implement wrong business strategies and leads to business losses. As research shows, some participants in online data collection applications are distrustful and unreliable to the data collector. The reason why the respondents refuse to provide truthful data is because they are in fear of personal information leakage and collusion attacks. In order to get relatively accurate data, companies need to employ cryptographic and random shuffling techniques to preserve data accuracy. Therefore, the major concern of data collectors is to guarantee that the modified data contain no sensitive information but still preserve high utility [19-21].

## 4 Privacy Laws

The privacy law in Europe is rather strong in order to strengthen the rights of the consumers. However, the U.S. - E.U. Safe Harbor Principles, developed between 1998 and 2000, currently effectively expose European users to privacy exploitation by U.S. companies. As a consequence of Edward Snowden's global surveillance disclosure, there has been increased discussion to revoke this agreement, as in particular the data will be fully exposed to the National Security Agency, and attempts to reach an agreement with the United States have failed [22].

In the United States, privacy concerns have been addressed by the US Congress via the passage of regulatory controls such as the Health Insurance Portability and Accountability Act (HIPAA). The HIPAA requires individuals to give their "informed consent" regarding information they provide and its intended present and future uses. U.S. information privacy legislation such as HIPAA and the Family Educational Rights and Privacy Act (FERPA) applies only to the specific areas that each such law addresses. The use of data mining by the majority of businesses in the U.S. is not controlled by any legislation [23].

## 5 Google Data Collection

Google is not only the world's largest digital advertising company [24] but also provides No.1 web browser [25], the No.1 mobile platform, and the No. 1 search engine world wide [26]. Google has more than one billion monthly active users in its video platform, email service, and map application. Google collects a tremendous amount of data about people's online and real-world behaviors via its various products and then uses it to target people with paid advertising [27].

Google collects user data in two ways: active and passive. The obvious way is "active", in which users directly and consciously communicate information to Google. For example, people sign in to Google's widely used applications such as YouTube, Gmail, Search etc everyday. Less obvious ways for Google to collect data are "passive" which means that Google uses some applications to gather information from users without their notice. Google's passive data gathering methods arise from platforms (e.g. Android and Chrome), applications (e.g. Search, YouTube, Maps), publisher tools (e.g. Google Analytics, AdSense) and advertiser tools (e.g. AdMob, AdWords). We as users usually overlook the extent and magnitude of Google's passive data collection.

### 5.1 A Day In the Life of A Google User

A study [28] has done an experiment to check how data is collected by Google on a normal day. The experiment was designed in a way that a researcher carried an factory reset Android mobile phone device and configured as a new device to avoid prior user information associated with the device. A new Google account was created and the researcher then went about a normal day using the mobile phone associated with the new Google account.

The study used two tools to check the data collection by Google: My activity [29] and Takeout [30]. My activity was used to show data collected by Google from any Search-related activities, use of Google applications (e.g. YouTube video plays, Maps search, Google Assistant), visits to 3rd-party web pages (while logged in to Chrome), and clicks on advertisements. The Google Takeout tool provides a more comprehensive information about all historical user data collected via Google's applications (e.g. it contains all past email messages on Gmail, search queries, location collection, and YouTube videos watched).

The key information collection events are shown in Figure 3.

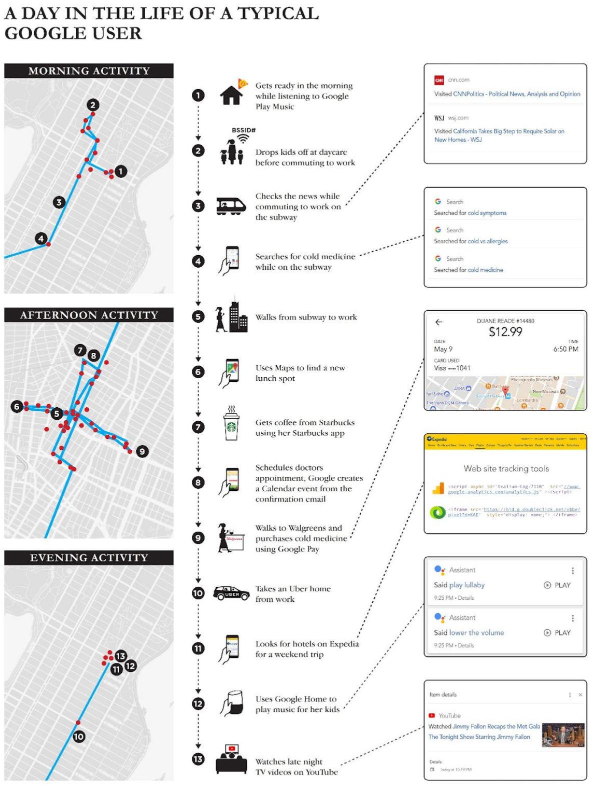
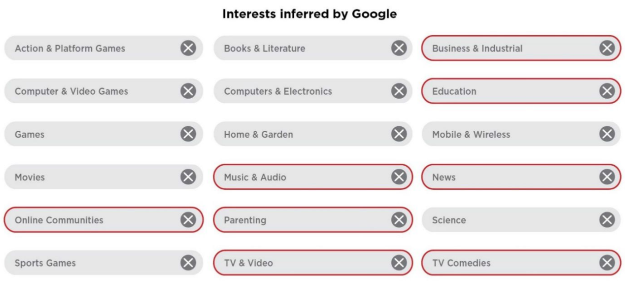

A DAY IN THE LIFE OF A TYPICAL GOOGLE USER
MORNING ACTIVITY
AFTERNOON ACTIVITY
EVENING ACTIVITY
1 Gets ready in the morning while listening to Google Play Music
2 Drops kids off at daycare before commuting to work
BSSID#
3 Checks the news while commuting to work on the subway
4 Searches for cold medicine while on the subway
5 Walks from subway to work
6 Uses Maps to find a new lunch spot
7 Gets coffee from Starbucks using her Starbucks app
8 Schedules doctors appointment, Google creates a Calendar event from the confirmation email
9 Walks to Walgreens and purchases cold medicine using Google Pay
10 Takes an Uber home from work
11 Looks for hotels on Expedia for a weekend trip
12 Uses Google Home to play music for her kids
13 Watches late night TV videos on YouTube
cnn.com
Visited CNNPolitics - Political News, Analysis and Opinion
WSJ wsj.com
Visited California Takes Big Step to Require Solar on New Homes - WSJ
Search
Searched for cold symptoms
Search
Searched for cold vs allergies
Search
Searched for cold medicine
DUANE READE #14480
$12.99
DATE
May 9
TIME
6:50 PM
CARD USED
Visa ····1041
Expedia
Web site tracking tools
Assistant
Said play lullaby
PLAY
9:25 PM • Details
Assistant
Said lower the volume
PLAY
9:25 PM • Details
Item details
YouTube
Watched Jimmy Fallon Recaps the Met Gala
The Tonight Show Starring Jimmy Fallon
Details
Today at 10:19 PM

**Figure 3: A typical day in the life of a google user [28].**

In the activity shown in Figure 3, the study [28] shows that the number of "passive" data collection events outnumbered the "active" events by approximately two to one. The study also observes that Google analyzes the collected data to assess user interests, which it then applies to target users with appropriate ads. For example, Google provides a list of interests that it has inferred from a user's activities, available via the "topics you like" section in Google's Ad Personalization [31] web page.



Interests inferred by Google

Action & Platform Games
Books & Literature
Business & Industrial
Computer & Video Games
Computers & Electronics
Education
Games
Home & Garden
Mobile & Wireless
Movies
Music & Audio
News
Online Communities
Parenting
Science
Sports Games
TV & Video
TV Comedies

**Figure 4: Google's assessment of user's interests at the end of the day [28].**

Figure 4 shows a list of interests Google associated with the user's account after a day's use of activity. In total, Google associated 18 interests to the user, eight of which (shown by colored borders) closely matched the user's usage and activities.

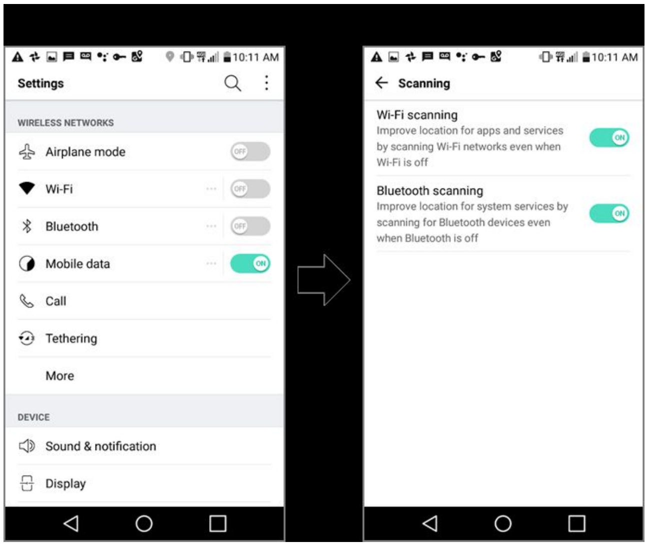
## 5.2 Data Collection Through Android And Chrome Platforms

Google's key platforms in terms of user data collection are Android and Chrome. Android captured 53% of the total US mobile OS market and there were more than 2 billion monthly active Android devices worldwide by January 2018. Google's Chrome browser held more than 60% share of all internet browser usage in the world with over 1 Billion monthly active users as reported in the 2017 Q4. Both platforms facilitate the use of Google and 3rd-party content which provide Google access to a wide range of personal, web activity, and location information [32 - 34].

Users must have a Google account in order to download and use apps from Google Play Store on an Android device. Therefore, Google accounts become a key gateway through which Google collects personal information, including user name, email, and phone number. In addition, if a user registers for services such as Google Pay, Google will also collect the user's credit card information, zip code, and birth date and all this information becomes part of a user's personal information associated with their Google Account [35].

Chrome on the other hand, collects a lot of personal information via its form "autofill" feature. The personal information typically includes user name, address, phone number, login name, and passwords. Chrome stores form fill information on a user's local drive, however, if the user logs in to Chrome using Google Account and enables its "Sync" feature, this information gets sent to and stored on Google servers [35]. Chrome and Android not only send Google information personal data but also send a user's web browsing and mobile app activities. If users are signed in to Chrome, Google will automatically track and collect their webpage visit. Chrome also collects information about a user's browsing history, passwords, website-specific permissions, cookies, download history, and add-on data .

Another important  personal data that Google collects is users' location data. Android and Chrome can collect an approximate location assessment using GPS coordinates on an Android phone or through a network's IP address on a desktop/laptop device. The user location accuracy can be improved through the use of nearby cell tower IDs or via scanning the device-specific BSSIDs or basic service set identifiers, assigned to the radio chipset used in nearby Wi-Fi access points [36]. Bluetooth beacons registered with Google's Proximity Beacon API can also be used to provide user's geolocation coordinates and even pinpoint exact floor levels in buildings. It is hard for an Android mobile user to disable location tracking. For example, on an Android device, even if a user turns off the Wi-Fi, the device's location is still tracked via its Wi-Fi signal. To prevent such tracking, Wi-Fi scanning must be explicitly disabled in a separate user action, as shown in Figure 5.


10:11 AM

Settings

WIRELESS NETWORKS

Airplane mode OFF

Wi-Fi OFF

Bluetooth OFF

Mobile data ON

Call

Tethering

More

DEVICE

Sound & notification

Display

10:11 AM

Scanning

Wi-Fi scanning
Improve location for apps and services by scanning Wi-Fi networks even when Wi-Fi is off ON

Bluetooth scanning
Improve location for system services by scanning for Bluetooth devices even when Bluetooth is off ON

**Figure 5: Android collects data even if Wi-Fi is turned off by the user [28].**

Google can track if a user is still, walking, running, bicycling, or riding on a train or a car by tracking an Android mobile user's location coordinates at frequent time intervals in combination with the data from accelerometer sensors on mobile phones.

To assess frequency of occurrence of active and passive data collection by Android or Chrome platforms, an experiment has been conducted by a study and for comparison's sake, the analysis of data sent to Apple via an iPhone device is also included. For simplicity, the phones were kept stationary, with no user interaction. In the experiment, on the Android phone a single Chrome browser session remained active in the background, whereas on the iPhone the Safari browser was used. Figure 6 shows a summary of the results obtained from this experiment. The x-axis indicates the number of times the phones communicated with Google (or Apple) servers, whereas the y-axis indicates the phone type (Android or iPhone) and server domain type (Google or Apple) with which data packets were exchanged by the phones. The colored legend describes the type of data requests identified by the domain address of the server.

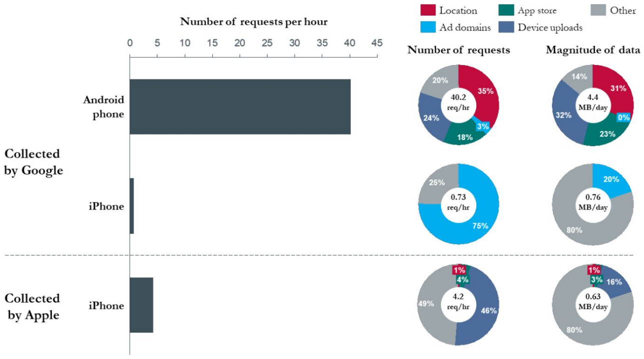


**Figure 6: Traffic data sent from idle Android and iPhone [28].**

Figure 6 indicates that the iPhone device communicated with Google server at a much lower frequency than the Android device. And Google did not collect any user location during the experiment time frame via iPhone which indicates that the Android and Chrome platforms play an important role in Google's data collection.

According to the results of the experiment. The Android device communicated approximately 900 data samples to a variety of Google server endpoints during a 24-hour time period. About 35% of the data communication was location related. And about 62% of communications are with the Google server domains requests to Google's Play App store, Android's uploads of device related data. Finally, Google ad domains received only about 3% of the traffic which is due to the fact that the browser was not actively used during the experiment.

## 6 COVID-19's impact on data privacy

The current coronavirus pandemic may be the first global pandemic for many of us to face. Technology is being used in many fields to fight for the pandemic: It is being used to measure the progress of the pandemic; Deliver possible digital health solutions; Help us keep social distancing; It has also enabled a fair portion of the world to stay up and running, albeit working from home and via online conference calls. However, numerous privacy, data protection, security questions related to technologies are raised during the global spread of COVID-19. For example, people are curious about how the transfer, remote access to, and analysis of their data (medical and otherwise). And a series of recent cyber attacks raise some data privacy concerns: who has access to which data and for how long? We are now facing the dilemma that we need privacy and data protection on one hand and the protection of public health on the other. There are a number of reasons why we should be wary of the new information-sharing reality that the global pandemic has created, but we are lucky to have versatile technology and cybersecurity on our side to help us assess risks and make the right decisions to stay safe [37].

Around the world, governments have been applying technology to conduct mobile location tracking of infected, or potentially infected, citizens and to enforce quarantines. At the same time, experts are cautious about this new level of visibility into our lives. Since everything we do is exposed and documented, it increases the risk of attackers hacking or gaining access to email and web accounts, thereby gaining the ability to completely obstruct someone's personal or professional life. This is something that could potentially happen with increased data collection by governments, or even hostile entities that could access our accounts due to the new state of data exchange during the pandemic [37].

## 6.1 Zoom and Zoombombing

With the social distancing applied in many countries, lots of businesses and schools deserted the offices and campuses and began working remotely from home. For large organizations that were equipped for telecommuting prior to the crisis, only minor adjustments had to be made. But for the majority of companies, sufficient frameworks, policies and solutions to protect organizational data were not in place, potentially exposing employees' personal and professional data to hackers eager to manipulate the current situation.

Employees now use remote VPN connections on household Wi-Fi networks with weak security controls instead of working under the security umbrella of their employers' protected networks. This exposes new threats to many organizations.

Zoom, which is the most popular online conference platform during the pandemic, is involved in the recent security breaches. Underground hacker forums circulated and sold thousands of users' credentials, and the unfortunate practice of "Zoombombing," or the hijacking of meetings, to display profane or malicious content.

An Israeli cybersecurity company called BitDam, has developed a solution that protects Zoom calls. Its advanced threat protection solution blocks content-borne attacks across all enterprise communication channels, including email, cloud storage, and other platforms.

## 6.2 Digital QR Health Code [38]

During the pandemic, the Chinese government has used a color-based "health code" system to control people's movements and control the spread of the coronavirus. Relying on mobile technology and big data, the automatically generated quick response codes, commonly abbreviated to QR codes, are assigned to citizens as an indicator of their health status [38].

### 6.2.1 How does QR Health Code work

In many cities of China, citizens without the app wouldn't be able to leave their residential compounds or enter most public places. Imagine your daily routine being entirely dependent on a smartphone app. Leaving your home, taking the subway, going to work, entering cafes, restaurants and shopping malls. Each move, dictated by the color shown on your screen. Green: you're free to proceed. Amber or Red: you're barred from entry. QR codes are shown in Figure 7.

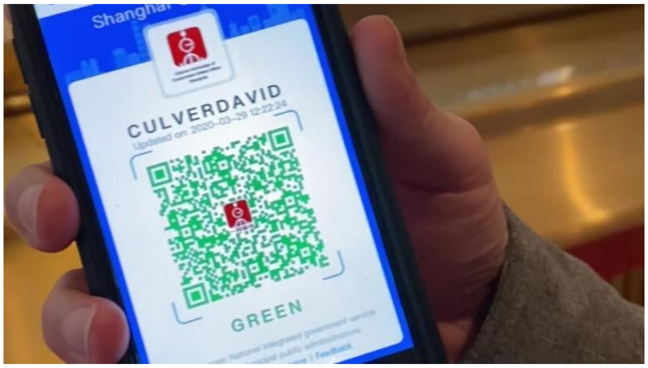


**Figure 7: CNN International Correspondent David Culver shows his health QR code in Shanghai. A green code means he's healthy and safe to travel [38].**

The two tech giants Alibaba and Tencent help Chinese government to host the health code systems on their popular smartphone apps. Alibaba's mobile payment app Alipay and Tencent's messaging app Wechat are both ubiquitous in China, each used by hundreds of millions of people. Placing the health codes on these platforms means easy access for many.

With the virus largely contained and lockdown measures gradually lifted across most of China in February, the small barcodes have remained in place and are still ruling people's lives in China. Hangzhou was among the first cities to use the health codes to decide which citizens should go into quarantine. The system was launched on February 11 by Alipay. To obtain a health code, citizens have to fill in their personal information including their name, national identity number or passport number, and phone number on a sign-up page. They're then asked to report their travel history and whether they have come into contact with any confirmed or suspected Covid-19 patients in the past 14 days. They also need to check the boxes for any symptoms they might have: fever, fatigue, dry cough, stuffy nose, running nose, throat ache or diarrhea. After the information is verified by authorities, each user will be assigned a QR code in red, amber or green.

Users with a red code have to go into government quarantine or self-quarantine for 14-days, users with an amber code will be quarantined for seven days, while users with a green code can move around the city freely. The health codes can also serve as a tracker for people's moves in public areas, as residents have their QR codes scanned as they enter public places. Once a confirmed case is diagnosed, authorities are able to quickly backtrack where the patient has been and identify people who have been in contact with that individual.

**6.2.2 How widely is QR Health Code used**

Within a week of its launch, the Alipay health codes were rolled out in more than 100 cities across the country. By late February, more than 200 cities had adopted these QR codes, according to Alipay. And Tencent's health code system had also expanded to more than 300 cities as of last month, according to the state-run Science and Technology Daily. On March 1st, Beijing launched its version of the three-colored QR code, accessible via both Alipay and Wechat. In addition to providing their name and ID number, users also need to register with facial recognition to obtain their colored code. The health codes have also played a central role in the gradual lifting of travel restrictions in Hubei province, where around 60 million people had their movement restricted following lockdowns in late January. On March 10, the province issued its health codes for residents who want to travel within the province. The colors are assigned according to the provincial epidemic control database: people who have been diagnosed as confirmed, suspected or asymptomatic cases, or people with a fever will receive the red color code; their close contacts will receive the yellow code; and people without any record in the database will get the green code which means they're healthy and safe to travel. The colors of the QR codes decide people's freedom of movement: green code holders are allowed to travel within the province, amber code holders are not allowed to travel, and red code holders will be treated and quarantined. All residents and visitors leaving Hubei and Wuhan need to have a green QR code on their phones.

**Figure 8: Passengers check their Wuhan Health code in front of the Hankou Train Station in Wuhan on April 8 [38].**

#### 6.2.3 Problems with Health Code

As with all products of technology, the health app is not perfect. It can make mistakes and assign users the wrong color code, and force the wrong people into quarantine. In Hangzhou, the city where the Alipay health codes were first introduced, some residents have complained on social media that they were given the red code for the wrong reason such as checking "stuffy nose" or "fatigue" on the sign-up page, despite they are also symptoms for the common cold and flu. A few days after its launch, the Hangzhou authorities said in a statement that the mayor's hotline had received too many calls from people who have questions over their codes, and had thus set up an online application for people demanding a review of their assigned codes.

As Chinese people resume traveling under the lifting of lockdown measures, another problem has arisen: not all cities and provinces recognize each other's health codes. Although the QR codes all come in the same three colors and are developed by the same companies, they are based on different Covid-19 databases set up by local authorities. Because the databases are not shared among local governments, and because different governments might have different standards for assigning the colors, some have been reluctant to recognize health codes from other places. To address the issue, the central government has launched a national "epidemic prevention code." It also uploaded a nationwide database of confirmed and suspected Covid-19 cases and their close contacts on a centralized platform, hoping that local governments can recognize each other's health codes through data sharing.

#### 6.2.4 Privacy Concerns with Health Code

There are also concerns about privacy. The health codes rely on troves of data the authorities have collected from individuals including their personal information, location, travel history, recent contacts and health status. Chinese citizens are concerned about whether their personal information will be leaked, and whether our information security can be ensured.

Jason Lau, a privacy expert and professor at Hong Kong Baptist University, said Chinese authorities need to make sure the health codes meet the typical data privacy principles. For instance, the data collected should be "proportionate with the purpose to be achieved." He also raised the question of whether the codes and all the personal information collected will be here to stay even after the pandemic has passed.

## 7 Conclusion

Data mining is an indispensable analytical tool for answering both big and small business questions, and raw data is a necessary foundation for data mining implementation. On the contrary, this technology arouses privacy concerns as social networks and big data are booming, especially in this global pandemic. People don’t want to share their personal information without knowing it, they don’t want their personal information to get misused or revealed. But actually every company is keeping records of their customers. Also, information leaks are a harmful threat to people’s normal lives. In recent years, there have been many personal information leak cases. Gradually, people are unwilling to provide personal information, which is a huge drawback to commerce, since companies don’t have sufficient data to analyze and target their consumers. In addition, there are possible ways to reduce risks of privacy invasion from both the user side and company side, but extra methods and efforts are required. As a result, the utilization of data mining still stays controversial.

## ACKNOWLEDGMENTS

We would like to thank Professor Kristin Tufte, the instructor for this course, for providing advice, ideas, and resources for this project.

# REFERENCES

[1] Christiansen, Linda. "Personal Privacy and Internet Marketing: An Impossible Conflict or a Marriage Made in Heaven?" Business Horizons 54.6 (2011): 509-14. Web.

[2] Winslett, Marianne, and Braganholo, Vanessa. "Jiawei Han Speaks out On Data Mining, Privacy Issues and Managing Students." ACM SIGMOD Record SIGMOD Rec. 40.4 (2012): 28. Web.

[3] Dwoskin, Elizabeth, and MacMillan, Douglas. "Smile! Marketing Firms Are Mining Your Selfies." WSJ. Web. 19 May 2014.

[4] Gorry, G. Anthony, and Robert A. Westbrook. "Can You Hear Me Now? Learning from Customer Stories." Business Horizons 54.6 (2011): 575-84. Web.

[5] Park, Yong Jin, Scott W. Campbell, and Nojin Kwak. "Affect, Cognition and Reward: Predictors of Privacy Protection Online." Computers in Human Behavior 28.3 (2012): 1019-027. Web.

[6] Sainani, Kristin L. "Logistic Regression." Pm&r 6.12 (2014): 1157-162. Web.

[7] Wu, Kuang-Wen, Shaio Yan Huang, David C. Yen, and Irina Popova. "The Effect of Online Privacy Policy on Consumer Privacy Concern and Trust." Computers in Human Behavior 28.3 (2012): 889-97. Web

[8] Ashrafi, Mafruz Zaman, and See Kiong Ng. Collusion-resistant Anonymous Data Collection Method. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '09 (2009): n. pag. Web.

[9] Bal, Gökhan, Kai Rannenberg, and Jason I. Hong. Styx: Privacy Risk Communication for the Android Smartphone Platform Based on Apps' Data-access Behavior Patterns. Computers & Security 53 (2015): 187-202. Web.

[10] Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2014-01-27.

[11] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases"

[12] Seltzer, William (2005). "The Promise and Pitfalls of Data Mining: Ethical Issues"

[13] Pitts, Chip (15 March 2007). "The End of Illegal Domestic Spying? Don't Count on It". Washington Spectator. Archived from the original on 2007-11-28

[14] Taipale, Kim A. (15 December 2003). "Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data". Columbia Science and Technology Law Review. 5(2). OCLC 45263753. SSRN 546782

[15] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques.San Mateo, CA, USA: Morgan Kaufmann, 2006

[16] L. Brankovic and V. Estivill-Castro, ``Privacy issues in knowledge discovery and data mining," in Proc. Austral. Inst. Comput. Ethics Conf., 1999, pp. 8999.

[17] R. Agrawal and R. Srikant, ``Privacy-preserving data mining," ACM SIGMOD Rec., vol. 29, no. 2, pp. 439450, 2000.

[18] Y. Lindell and B. Pinkas, ``Privacy preserving data mining," in Advances in Cryptology. Berlin, Germany: Springer-Verlag, 2000, pp. 3654.

[19] C. C. Aggarwal and S. Y. Philip, A General Survey of Privacy Preserving Data Mining Models and Algorithms. New York, NY, USA: Springer-Verlag, 2008.

[20] M. B. Malik, M. A. Ghazi, and R. Ali, ``Privacy preserving data mining techniques: Current scenario and future prospects," in Proc. 3rd Int. Conf. Comput. Commun. Technol. (ICCCT), Nov. 2012, pp. 2632.

[21] S. Matwin, ``Privacy-preserving data mining techniques: Survey and challenges," in Discrimination and Privacy in the Information Society. Berlin, Germany: Springer-Verlag, 2013, pp. 209221.

[22] UK Researchers Given Data Mining Right Under New UK Copyright Laws. Archived June 9, 2014, at the Wayback Machine Out-Law.com. Retrieved 14 November 2014.

[23] Judge grants summary judgment in favor of Google Books – a fair use victory". Lexology.com. Antonelli Law Ltd. Retrieved 14 November 2014.

[24] “Google and Facebook tighten grip on US digital ad market,” eMarketer, Sept. 21, 2017.

[25] “Market share or leading internet browsers in the United States and worldwide as of February 2018,” Statista, February 2018

[26] “Global OS market share in sales to end users from 1st quarter 2009 to 2nd quarter 2017,” Statista, August 2017.

[27] “Worldwide desktop market share of leading search engines from January 2010 to October 2017,” Statista, Feb. 2018.

[28] Douglas C. Schmidt, “Google Data Collection”, Digital Content Next, 2018.

[29] “My Activity,” Google, available at https://myactivity.google.com/myactivity

[30] “Download your data,” Google, available at https://takeout.google.com/settings/takeout?pli=1

[31] “Ads personalization,” Google, last accessed 2018, available at https://adssettings.google.com/authenticated

[32] “Subscriber share held by smartphone operating systems in the United States from 2012 to 2018,” Statista, May 2018

[33] Dave Burke, “Android: celebrating a big milestone together with you,” Google, May 17, 2017.

[34] Google 10K filings with the SEC.

[35] “Google Chrome privacy whitepaper,” Google, March 6, 2018.

[36] “Google beacon platform, proximity beacon API,” Google, last accessed on August 15 2018.

[37] https://www.forbes.com/sites/startupnationcentral/2020 /05/07/data-privacy-cybersecurity-public-health-covid19-coronavirus /#534dbcd18f29

[38] https://www.cnn.com/2020/04/15/asia/china-coronavirus -qr-code-intl-hnk/index.html