541 Lab 2 - Privacy

Required readings

We covered some of these videos in lab, but including them here so that you have everything in one place. There might be questions on specifics from the required readings whereas the optional readings are more of a general help, optional questions, and if you are interested to explore further.

- **Lec 3** Madhumita Murgia "How data brokers sold my identity" (0:00 - 9:40)
- **Lec 3** Michal Kosinski "Part One: The End of Privacy, Data Scientists Know All Your Secrets" (13 min video)
- **Lec 3** Capture Behavioral Engagement "Marketing Automation for Higher Education" (2 min video)
- **Lec 4** Latanya Sweeney "Data Privacy in the Digital Age" (0:00 - 16:40)
- Sara Buhr "An Amazon Echo may be the key to solving a murder case" (3 min read)
- Emma Wollacott "70,000 OkCupid Profiles Leaked, Intimate Details And All" (4 min read)
- CPAC New clip "Clearview AI violated Canada's federal and provincial laws" (5 min video)

Optional readings

▶ Click to show

Submission instructions

rubric={mechanics:20}

You receive marks for submitting your lab correctly, please follow these instructions:

- Follow the general lab instructions.
- Click here to view a description of the rubrics used to grade the questions
- Push your `.ipynb` file to your GitHub repository for this lab (make at least three commits).
- Upload your `.ipynb` file to Gradescope.
- Include a clickable link to your GitHub repo for the lab just below this cell
  - It should look something like this https://github.ubc.ca/MDS-2022-23/DSCI_541_labX_yourcwl.

    - If you are working in a group, you can create you own (public) repo in the UBC-MDS organization and link that instead.
- All your written answers must be in your own words.
- You are not allowed to use generative AI tools to write your answers for you or simply paraphrase answer that you generate from these tools (that will lead to a failing grade), but you can use them to further understand the topics you are learning about.

https://github.com/UBC-MDS/541-lab-2

Overall writing quality

rubric={writing:20}

You will receive an overall writing grade for the entire lab instead of for each question. This is just a small part of your total grade, but please use the Jupyter Lab spell checker extension to catch typos and read through your text for grammatical errors before submitting (or paste it into Google Docs/MS Word/Grammarly. You don't need to type anything under this cell, it is just a placeholder to generate the grading rubric.

1. Short answer questions

Keep your replies brief, 1-3 sentences per question. Although these are short answer questions, don't copy answers from the readings, use your own words so that you practice learning these concepts. These will not be discussed during the lab.

Question 1.1

rubric={reasoning:60}

1. What is a data broker?
2. What are direct and indirect identifiers?
3. What is k-anonymity and l-diversity?
4. What are some weaknesses of k-anonymity (and l-diversity)? How could you re-identify individuals in these datasets?
5. If you have two differentially private datasets, one with and one without your data, what does differential privacy guarantee regarding your privacy?

6. How does differential privacy work on a conceptual level (watching the optional video from Minute Physics can help with this if you want more details and examples than what was given in the lecture)?
7. What are some additional approaches you could take to secure data stored in your organization?

1. A data broker is a business or person that gathers personal information about people from different places, like websites or public records, often without us knowing. They organize and analyze this data to build detailed profiles, which they then share or license to other companies for things like advertising, credit checks, or other decisions.

2. Direct identifiers are things like our name or email that can clearly point to who we are. Indirect identifiers, on the other hand, don't reveal our identity by themselves but could if combined with other info like our age, where we live, or what we do for work.

3. K-anonymity is a way to protect privacy by making sure each person's data looks the same as at least k-1 other people based on certain details. L-diversity takes it a step further by making sure there's variety in sensitive information within each group, so no one's private info can be guessed easily. Both methods work together to make it harder to trace data back to a specific person.

4. K-anonymity isn't perfect. If everyone in a group has the same sensitive detail, that info can still be exposed. L-diversity helps by adding variety, but it can still fall short if the values are too similar or easy to guess. It's also possible to re-identify people by combining anonymized data with other datasets that include matching details.

5. Differential privacy means that whether our data is included or not, the results stay almost the same. This helps protect our privacy because no one can be sure if our data was part of the dataset in the first place.

6. Differential privacy works by adding some random noise to the results so that one person's data doesn't stand out. This way, the analysis reflects the group as a whole, not any one individual.

7. To keep data safe in our organization, we can use encryption both when it's stored and when it's being shared. It's also important to set clear access rules and do regular checks. On top of that, we can anonymize or remove personal details from the data to limit the damage if anything ever gets leaked.

2. Discussion questions

This section asks you to expand a bit on your reasoning, but still aim to write succinct replies around one paragraph per sub-question. The goal of lab discussions are not to provide you with the right answers, but to help your discussion along. Your TA will assist in this by bringing up topics that you might not have thought of, ask questions to break the silence or a dead end, and move the conversation along so that you have time to go through most questions. How useful the lab discussion is for your submission ultimately relies on that you actively contribute to the discussion and help your peers contribute and exchange ideas.

Some tips to make your discussions in lab more effective

It is easy to overlook the flaws of our own reasoning, so having a discussion with colleagues is an excellent opportunity to develop your thinking and receive feedback from someone who can provide an alternative perspective from your own. Nevertheless, many people don't know how to have an effective discussion, so I am sharing a few tips for you to be able to make the most out of this opportunity:

- Commit to learning, not "winning" debates.
- Comment in order to share information and develop arguments further, not to persuade.
- Listen respectfully, without interrupting, to try to understand each others' views.
  - Don't focus on what you are going to say next while someone else is talking.
- Challenge ideas, not individuals.
  - And be open to having your own ideas challenged.
- Think about as good arguments as possible against your position.
  - This is especially useful if many of your peers have the same opinion, help your group find angles that you might otherwise be missing.
- Allow everyone the chance to speak.
  - Politely ask members of your group about their opinion.
- Avoid assumptions about any member of the class or generalizations about social groups.
  - Be careful about asking individuals to speak on the behalf of their (perceived) social group.
- Be aware of logical fallacies, but avoid pointing them out in rude or disrespectful ways.

Question 2.1 -- Technology and privacy

rubric={reasoning:100}

1. **Acceptable exchange.** Do you consider it acceptable that companies collect information about your online behavior in exchange for using their services? Is the service they provide a reasonable compensation of your data or do you think they owe you monetary compensation if they profit from selling your data to a third party? Include an example of a service you think provides a reasonable or unreasonable trade off.
2. **Pay for privacy.** If you had the option would you pay to use email, social media, file storage, etc that met your preferred privacy standard? Are there specific online services which you would be more likely to pay for than others (and why do you think privacy is more important in these cases)? Is privacy a commodity that should only be available to those who can afford it, or is it a right that should be regulated and available to all for free?

1. It is acceptable for companies to collect users' data, however, they must clearly disclose what data they collect, how it will be used, and whether it will be sold to third parties in their Terms & Conditions. Users should have the option to opt out of the company's service anytime and at that time companies should stop collecting that user's data and delete previously collected data immediately.

   As for monetary compensation, no compensation owed if data is used solely to enhance the user's experience. For example, Google Maps collects location data to improve navigation, which benefits users directly. This is a reasonable trade-off because the utility is clear, and users can disable tracking. However, if data is sold to advertisers, or third parties, for example, Meta profits heavily from ad targeting but users receive no financial return, then this is an unreasonable trade-off as users, not the company, bears privacy risks without direct benefits. Thus, in such situations, profit -sharing should be debated, and users should be compensated.

   Therefore, user rights should be specifically specified in Terms of Conditions, opt-out must be effortless, with immediate data protection, and there is no penalty for refusing data collection. If user data is sold, companies must name the specific industries and allow users to exclude their data from certain uses.

2. Users should not pay to protect their privacy, it is companies' responsibility to ensure that their users' information is safely stored and used safely. Companies that profit from user data (e.g., Meta, Google) have an ethical obligation to safeguard privacy without charging extra and not offer it as a "premium" add-on.

   We would consider paying only for services handling highly sensitive data (e.g., financial records, health info). For example, the service that encrypts files so even

the company itself can’t access them, or the service that rejects ads and data harvesting, which relies on donations and subscriptions.

Privacy is not a commodity, it is a Human Right. Like the Fifth Amendment of US law, a person shouldn’t need to be wealthy to exercise his or her rights. Regulatory frameworks such as below can be legally enforced.

- Bans on "dark patterns" (tricking users into consent)
- Mandatory data minimization (collect only what’s necessary)

Question 2.2 -- Technology and privacy

rubric={reasoning:100}

1. **Terms of service.** When we sign up for online services, we often have to agree to terms of conditions to use the services. Sometimes these do indeed mention that your data is being collected by the company, but studies have shown that the language used in terms and conditions often requires college level reading apprehension, while many users of such platforms are children in middle and high school. Could you think of some ways to improve the presentation of terms of services to convey the information more effectively?
2. **Alternative options.** A common argument is that we can just opt out of services we don't like and use others instead (i.e. it is the indvidual's responsibility), but is this really true? Or are we getting to the stage where we "need" certain online services for us to integrate in society and opting out is not really an option? For example, is there a widely used replacement alternative for services such as Facebook, Instagram, YouTube or TikTok that does not center their business model on mininig user information?
3. **Beyond online tracking.** What about extending the data collection outside the online environment? Which of the following do you find reasonable and unreasonable privacy-wise? Are these examples different from tracking online (why/why not)?
    - Is it ok if an Amazon Echo/Google Home device records your voice at home?
    - What if you have friends over, do you need to inform them about your echo device and should their voices be recorded too?
    - What about a pair of glasses that automatically record video and sound as you walk around in public places?
    - Or using WiFi signals to detect people moving in their homes?

1. Terms of service usually use very specfic and formal language, which makes them long and difficult to understand, as their main purpose is to protect the service provider from legal risks. Still, there are some improvements that could be made to optimize the user experience:

    - Alongside the original full version of the TOS, provide a short summary with links to the details.
    - Provide a teen/senior version with simplified wording and shorter content.
    - Highlight important sentences for users.
2. Sometimes we do not have alternative options. For example, in China, Alipay is closely connected with public services. If users don't agree with the TOS, they are expected not to use the app or not to have a cellphone, which can make it nearly impossible to get things done or too troublesome.

    If we're just talking about social media, big platforms have more content creators and consumers, providing richer entertainment to explore. The good news is, there are some widely used non-profitable alternatives, as long as we bring our old friends along or make new ones.

    Still, from the perspective of social integration, someone might be seen as a weirdo if they don't use Twitter or Instagram. And it can be harder to find a job if they don't use LinkedIn.

3. Extending the data collection outside the online environment is unacceptable. None of them are reasonable privacy-wise. These examples are different from tracking online, because we may only use online services 5 hours one day and nobody is supposed to and has no rights to track our rest 19 hours. We may just give out our interests of purchase or clicking or watching online, but we can lose much more offline.

    - Is it ok if an Amazon Echo/Google Home device records your voice at home?
        - Not Ok. Firstly, we do not give permission to online companies to track our offline information (which is not even information that we voluntarily expose). Secondly, at home we may be talking about more private and sensitive information that could have more serious consequences if this information is compromised.
    - What if you have friends over, do you need to inform them about your echo device and should their voices be recorded too?
        - We will inform our friends about the Echo device. Since their words and information are involved, their preferences about being recorded should be taken into account. They also have the right to ask not to be recorded.
    - What about a pair of glasses that automatically record video and sound as you walk around in public places?

    - Though it is common, it's not good behavior and sometimes even illegal. People who are recorded have the right to prevent their publicity or even call the police. Informed consent needs to be sought from everyone.
- Or using WiFi signals to detect people moving in their homes?
    - It is not common and sounds like a military context. It could lead to serious consequences if any group or individual were able to do it.

Question 2.3 -- Scraping data and facial recognition software

rubric={reasoning:100}

1. **Facilitating queries on scraped data.** Is it fair game to scrape public information from the internet, and then facilitate queries on that information or are there any additional concerns when the data becomes easier to access? What about scraping data that is only public for anyone logged in to that platform but otherwise not accessible (e.g. Facebook or OKCupid) and then making it available outside the platform?
2. **Privacy and control of scraped data.** Clearview AI is a company that provides facial recognition software, which is used by private companies, law enforcement agencies, universities and individuals. The database consists of more than three billion images scraped from the internet, including from social media applications. Clearview claims 99% accuracy for most photos and one of their goals is to find criminals more accurately than current approaches used by law enforcement. One of its biggest technical advantages over previous law enforcement tools is that they have a huge set of images including the general public, not just photos of previous convicts. A recent data breach revealed that Clearview AI is employed among American and Canadian law enforcement, including by the Vancouver Police Department.
    - How do you think such powerful facial recognition technology should be regulated? For example, is it fair use to scrape data that was public at the time of scraping and store it permanently in a database or do we have a a right to be forgotten even in downloaded data? Should we take into consideration that they are using this data for a good cause, such as catching criminals (e.g. Clearview is claimed to have been used to identify participants in the Capitol hill riot and to identify dead soldiers in the Russia/Ukraine war)? Also think about how the scale of implementation matters, even for algorithms that are 99% accurate.

1. Scraping public data is often in a legal gray area, but it's generally allowed if done ethically by respecting terms of service, robots.txt, and user privacy. Scraping data behind a login wall is more risky and could break laws like the CFAA or GDPR, as well as platform rules. Making scraped data searchable can also raise serious privacy concerns, especially if it includes personal information. We believe it's important to always consider the legal, ethical, and potential harm aspects before going ahead with any scraping.

2. Clearview AI's technology might be helpful for law enforcement, but it brings up major privacy concerns because it scrapes public photos without consent to create a permanent facial recognition database. Even with 99 percent accuracy, mistakes at scale could affect millions of people. We believe there should be strict regulations that ensure transparency, set clear usage limits, and give people the right to opt out. Good intentions alone are not enough to justify this level of surveillance.

Question 2.4 (Challenging)

rubric={reasoning:20}

Note: Since this exercise is a bit longer than the regular challenging questions, it will count ***both for the ~5% challenging points for this lab, and the 2% extra credit assignment for your overall grade in 541.***

In previous years, students have asked for an exercise on how to implement differential privacy in a data analysis. It's difficult to fit this into the curriculum because many of the differential privacy libraries are not trivial to use and would take a fair amount of time to learn properly.

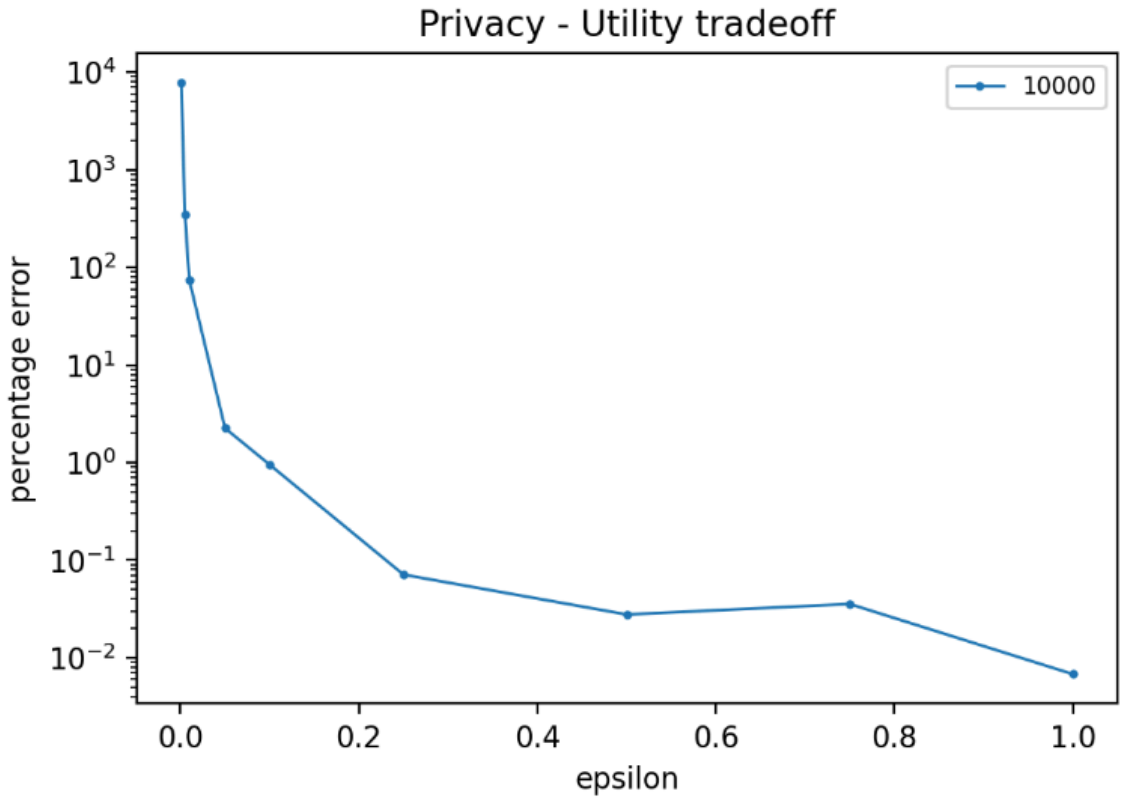
This challenging exercise is an experiment where I collaborated with the team behind Antigranular, a community-led open-source platform that combines confidential computing with differential privacy. They have put together a notebook with an introduction to a couple of differential privacy libraries that are easier to get started with. Antigranular is aiming to develop a general curriculum for privacy concerned data analysis, so as part of this exercise you will give feedback on how easy it was to follow along in the tutorial notebook and what you think could be improved.

You can access the notebook in the student repo. You can either study the output of the notebook without running it, or re-run the cells and change things around if you want to do more experimentation. You task is to read through the notebook and answer the following questions:

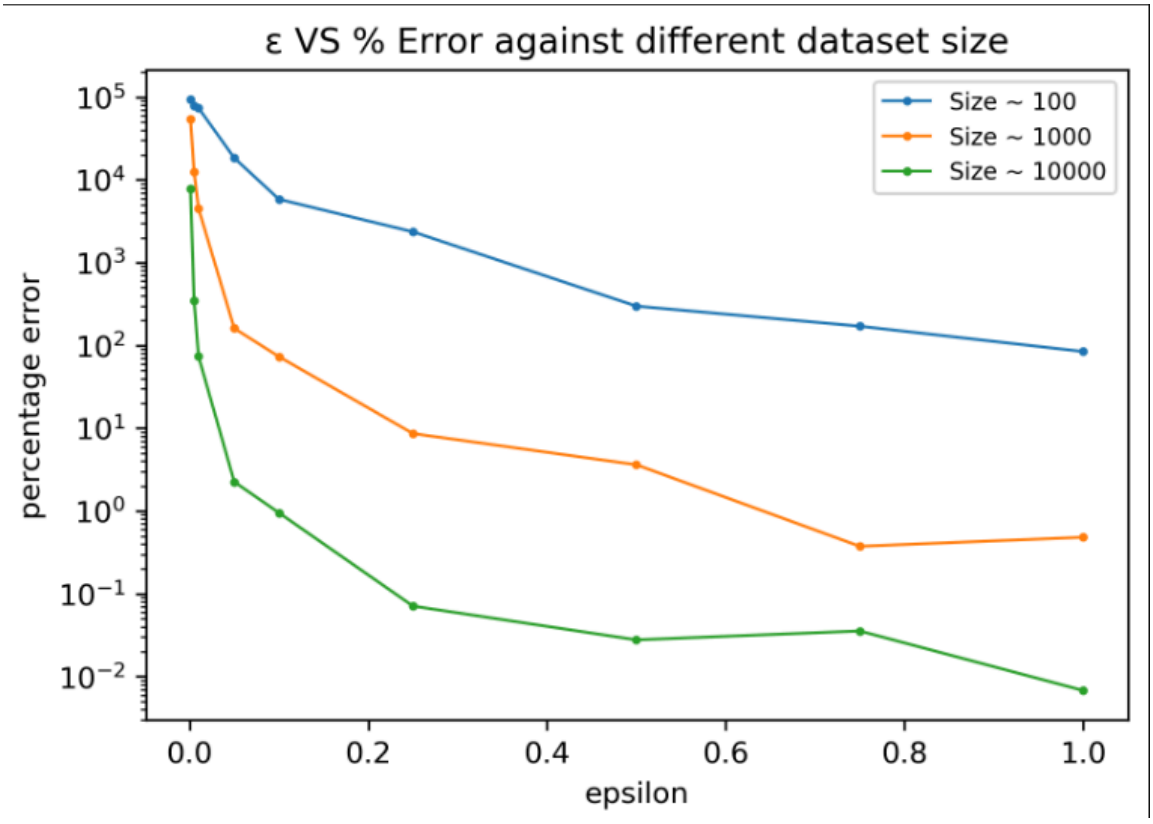
1. Why is the salary that is printed for the new employee so different when using differential privacy versus when we are not using it? Would you say this is a feature of differential privacy or a drawback?
2. What does it mean that there is a tradeoff between privacy and utility? Describe this in the context of differential privacy specifically referring to the figure with the same name in the notebook and comment on why the size of the data matters.
3. Provide feedback on your general experience working through the notebook. What worked well and what do you think could be improved? Was it suitable for someone with your background level of experience with differential privacy (a brief high level introduction to the main concepts during lecture)? Did you find that seeing examples in actual code helpful for your understanding of what differential privacy is and how it can be employed in a data analysis?

1. The salary calculated for the new employee is very different when using differential privacy because noise is added to the average salary both before and after the new hire. When we use this noisy average in the salary formula, it results in a large error in the estimated salary. This is a fundamental aspect of differential privacy, meaning it is designed to prevent adversaries from being able to accurately infer individual-level data by adding random noise. While it reduces precision, this is an intentional trade-off to protect privacy.

2. The tradeoff between privacy and utility refers to the fact that stronger privacy (i.e., lower epsilon) introduces more noise, which reduces the accuracy (utility) of the output. In the "**Privacy - Utility tradeoff**" graph, we see that as epsilon increases, the percentage error in the sum decreases, meaning we get more accurate results but less privacy.



The "**ε vs % Error against different dataset size**" graph shows that larger datasets have lower percentage errors for the same epsilon. This happens because the impact of noise on aggregate values becomes smaller when there are more records, improving utility without compromising individual privacy.



3. Exploring the notebook gave us a valuable learning experience. The organization of the notebook, combining theoretical background, code examples, and easy-to-understand visualizations, made complex concepts like epsilon-DP and the Laplace mechanism much clearer. We found setting up using Antigranular smooth, and the `op_pandas` library made it easy to run differentially private queries without needing to understand low-level implementation details. For students like us with just a lecture-level introduction to differential privacy, the notebook was well-paced and approachable. One suggestion for improvement would be to explain the significance of the outputs right after each plot or calculation, so it's easier to interpret the results step by step. Overall, seeing real code and data helped us better understand both the power and the limitations of differential privacy in practical analysis. So, we would like to say thank you.

Help us improve the labs

The MDS program is continually looking to improve our courses, including lab questions and content. The following optional questions will not affect your grade in any way nor will they be used for anything other than program improvement:

1. Approximately how many hours did you spend working or thinking about this assignment (including lab time)?

Ans: 6 hrs

2. Were there any questions that you particularly liked or disliked?

Ans: Question 2.1

Ans: [Questions you disliked]

In [ ]: