Olivia Dias

AI, Decision Making, and Society
6.3950/6.3952, Fall 2025
Pset 6 – Privacy

Due: November 12, 2025 (by 11:59 PM)

Algorithms are increasingly used to inform or make decisions about our lives in many settings, such as employment, lending, and healthcare. In this problem set, you will explore the privacy considerations in systems of algorithmic decision-making. This problem set has 3 parts:

- Problem 1: Notions of Data Privacy
- Problem 2: Differential Privacy & Unlearning with Census Data
- Problem 3: Differential Privacy and Governance

Please submit your assignment as a PDF compiled from this LaTeX template.

Problem 1: Notions of Data Privacy

Optional Background Reading:

- U.S. AI Bill of Rights: Data Privacy
- The Algorithmic Foundations of Differential Privacy (Dwork and Roth, 2014) (Chapters 1-3)
- Differential Privacy: A Primer for a Non-Technical Audience (Wood et al., 2018) (Pages 225-236, 250-259)

In the context of AI systems, privacy violations are often a repercussion of data-driven decision-making. This problem explores different notions of **data privacy**.

Privacy Fallacies

Consider the following fictional scenario: A project team in 6.395 plans to present a "data portrait" of TA Zekai during the final project presentation. They aggregate fragments from campus directories, court-accessible public records, course-evaluation snippets, news search, and social-media snapshots into a searchable prototype website. Before presenting, they ask you to serve as their "privacy consultant": Is this appropriate to present? What privacy and normative issues arise?

1. The students from the scenario above make the following claims. Briefly respond to each claim in 2–4 sentences. Answers should reference concepts in Lecture 13 - 15.

A. "Why should Zekai care if we make this information public? If he objects, he must have something to hide!"

B. "Many of these records are public anyway—we just consolidated what you could get from the courthouse or campus offices."

C. "Even if this makes Zekai feel humiliated, the law doesn't protect embarrassment as a harm."

A. Zekai still has the right to opt-out of data collection due to the fact that you aggregated his data and made it easily accessible without his consent. Additionally, the USA Freedom Act (2015) bans the bulk collection of metadata on US citizens. Therefore, Zekai is entitled to object to his information being made public, since the bare minimum data was not collected for this project. Finally, you need to make sure that if you make his information publicly available, that you get a release form where he acknowledges that he is okay with the use of his information in this manner.

B. Since the course-evaluation snippets are predominantly restricted to people within MIT's campus this information would be considered more private. Under the fourth amendment and the newly passed Massachusetts Data Privacy Act (MDPA), TA Zekai has the right to opt out of his information being collected from campus offices and courthouse. Again the final project must have some mechanism in place that demonstrates that they received Zekai's consent to collect his data and publicly share it online.

C. However, the law does protect Zekai from having his course evaluation snippets and campus directory information released in an easily accessible manner. This is because this data is stored within the confines of the institute. Based on the privacy torts derived from Warrens and Brandeis in 1890, it can be inferred that this is an invasion of Zekai's privacy

Online Data Privacy

There are many differences between notions of data privacy for offline and online contexts, but a central concern is that online systems collect orders of magnitude more data.

2. The following questions are about data privacy and online platforms (based on US laws). Briefly answer each question in a few sentences. Answers should refer to material from lecture.

A. What is the dead body problem?

B. Why is the accessibility of online information better understood as a *spectrum* rather than a binary public/private? Give two short examples that fall at different points on this range.

C. Roadside license-plate cameras can be networked and used to reconstruct driving histories. In 1–2 sentences, explain why aggregating many camera records over time differs from being observed once in public, and why the aggregated view may warrant stronger privacy safeguards (e.g., retention limits, higher access thresholds).

D. If you post content privately on social media, do you have a reasonable expectation of privacy?

A. The dead body problem is the problem where someone tries to prove a violation of privacy, however, there is no obvious evidence of privacy harms.

B. The accessibility of online information is better understood as a spectrum, because even though information is posted online does not mean it is accessible to everyone that is online. However, people with malicious intent can find out that information if it is posted online even if it is not publicly accessible to everyone. One example is a family groupchat on WhatsApp is online information, but it is encrypted so it would be difficult to access. So it is public to your family but private to the rest of the world. Another example of online information is a online news article written about you from your local newspaper. This online information is searchable by anyone.

C. Aggregating many camera records over time differs from being observed once in public, because aggregated data from the license-plate cameras is analogous to cellphone tracking which is against our fourth amendment rights. Whereas being observed once in public is random and can not be used to reconstruct a driving history.

The aggregated view many warrant stronger privacy safeguards because it would threaten our fourth amendment rights. It would require the removal of the license-plate camera data to be deleted after certain amount of time and require a warrant and a greater security clearance level for people to access.

D. If I post content privately on social media, I know most social media companies sell my online information so I would expect that my data would be used to train an AI model. So I would not have an expectation of privacy since my content would be public to my followers, and my followers might have their own distinct followers whom they can expose my content to.

Privacy in Datasets

Many datasets may infringe on the privacy of individuals to varying degrees. For example, datasets used to train AI models and datasets produced by AI systems might contain confidential or personally identifiable information. Even though we can restrict access to these datasets, we may still want to release the outputs of computations over these datasets (e.g. analytical queries or the outputs of AI models). Census data, for instance, has high analytical value, but the release of such aggregated information should protect the privacy of raw individual responses.

3. This is about a video[1] that is part of Recitation 6 activity, which is about protecting privacy in Census data. Answer the following questions based on the video (answers can be short phrases).

A. The video describes an example reconstruction attack based on access to summary statistics of age, gender, and ice cream preferences. What does the output of this attack look like in terms of columns and rows?

B. What are the x and y axes of the "plausibilities plot"?

[1] `https://www.youtube.com/watch?v=pT19VwBAqKA`

C. What is one of the biggest benefits of differential privacy?

A. The attack is able to decompose the age, gender, whether participant loves ice cream, and whether the participant is married based on the median and mean age of the participants

The output from this attack corresponds to the ages of the survey (first column), the gender of the survey participants (2nd column), whether the participant likes ice cream (3rd column), and whether the participant is married (4th column). And the rows have the following values:

$$\begin{bmatrix} 8 & female & loves_{icecream} & no_{ring} \\ 18 & male & hates_{icecream} & no_{ring} \\ 24 & female & hates_{icecream} & no_{ring} \\ 30 & male & hates_{icecream} & married \\ 36 & female & loves_{icecream} & married \\ 66 & female & loves_{icecream} & married \\ 84 & male & loves_{icecream} & married \end{bmatrix}$$

B. The y-axis is accuracy and the x-axis is privacy for the plausibilities plot.

C. The biggest benefit of differential privacy is that it reliably compounds over multiple pieces of information, (based on Protecting Privacy with MATH (Collab with the Census) video)

4. Even if reconstruction of individual data is not feasible, sometimes even being a member of a dataset is sensitive information. Give two examples of datasets in which membership could be considered sensitive.

- Membership in a criminal dataset would be considered sensitive information.
- Membership in a LinkedIn dataset with salary, company, and job title information would be considered sensitive information.

Randomization for Privacy

Suppose you're working with a research group that says they're using a randomized response protocol to protect privacy of respondents for a yes/no survey question. Their protocol asks the respondent to do the following:

- With probability 2/5, the respondent should invert their true answer: if their true answer is "yes" they should say "no", and if their true answer is "no" they should say "yes".
- With the remaining 3/5 probability, the respondent should answer truthfully.

Let y denote what the researchers want to find: the unknown fraction of the population for whom the true answer is "yes." However, because of the randomized response protocol, they only have access to q: the fraction of answers received through the survey that are "yes."

5. Compute the following quantities (as mathematical expressions). Justify your answers (mathematically).

A. Solve for y in terms of q.

B. What is the probability that a respondent's true answer was "yes" given that they responded "yes" to the survey? Your answer should be in terms of y.

A. We are given in the problem statement that $P(said = yes|actual = no) = \frac{2}{5}$, $P(said = yes|actual = yes) = \frac{3}{5}$, y equals probability of actual yes, and q equals the probability of reporting yes. So by

the law of total probability, we have that

q = $\frac{2}{5}$ (1-y) + $\frac{3}{5}$y

Now we solve for y and we get

$q = \frac{2}{5} - \frac{2}{5}y + \frac{3}{5}y$

$q = \frac{2}{5} + \frac{1}{5}y$

$5q = 2 + y$

$y = 5q - 2$

B. And we want to find the probability that the respondent's true answer was "yes" given that they responded "yes" to the survey. So we will use Baye's theorem and we get

$P(actual = yes|said = yes) = \frac{P(said=yes|actual=yes)P(actual=yes)}{P(said=yes|actual=yes)P(actual=yes)+P(said=yes|actual=no)P(actual=no)}$

$= \frac{(\frac{3}{5})(y)}{(\frac{3}{5})(y)+(\frac{2}{5})(1-y)}$

$= \frac{\frac{3}{5}y}{\frac{3}{5}y+\frac{2}{5}-\frac{2}{5}y}$

$= \frac{\frac{3}{5}y}{\frac{1}{5}y+\frac{2}{5}}$

Therefore,

$P(actual = yes|said = yes) = \frac{\frac{3}{5}y}{\frac{1}{5}y+\frac{2}{5}}$

6. You ask the researchers about their use of the number 2/5. "Actually," they say, "this is the second thing we tried. We started with the same protocol[2], but using the number 1/2 instead of 2/5." They go on to say: "a funny thing happened when we tried this original protocol", but don't elaborate. What did the researchers mean by this?

The researcher meant that it would be impossible to learn the true percentage of the population for whom the true answer is "yes". This is because if we use the equation from 5.A for q we know that

$q = \frac{2}{5}(1 - y) + \frac{3}{5}y$

However, if you change the probabilities to $\frac{1}{2}$ you get

$q = \frac{1}{2}(1 - y) + \frac{1}{2}y$

$q = \frac{1}{2} - \frac{1}{2}y + \frac{1}{2}y$

$q = \frac{1}{2}$

Therefore, we are no longer able to solve for y if the probabilities become $\frac{1}{2}$.

Differential Privacy

Differential privacy (DP)[3] is a formal technique for using randomization to protect the privacy of individuals in a dataset. The main idea behind DP is to ensure that adding or removing any single individual's data to or from the dataset does not significantly affect the outcome of any analysis or "function" computed on the

[2]Specifically, the protocol that they tried first was as follows: (1) with probability 1/2, the respondent should invert their true answer; (2) with the remaining 1/2 probability, the respondent should answer truthfully.

[3]DP was introduced in this seminal paper by Dwork et al. in 2006.

data. These computations could range from simple queries on the dataset (like averages or counts) to more complex tasks like training machine learning models. DP achieves privacy by adding carefully calibrated statistical noise to the outputs of these computations, making it difficult to identify any specific individual's data from the output.

The next few questions walk through the concept of differential privacy using the example dataset of student grades in Table 1.

Student Name	**GPA**
Alice	3.0
Bob	3.2
Charlie	3.5
David	3.8
Emma	4.0

Table 1: An example dataset of student GPAs (on a scale of 0.0 to 4.0).

Functions and Output Events

The formal definition of DP protects privacy for "functions" f computed over the data. For example, average GPA is a function that could be computed over student grades. The definition also relies on the concept of **output events** $T \in \text{Range}(f)$, where $\text{Range}(f)$ is the possible outputs of the function f. In other words, an output event could be a single possible value, or a range of values, that a function could take.

7. Answer the following questions about functions and output events based on Table 1.

A. Suppose f is average GPA. What is $f(\mathcal{D})$, where $\mathcal{D}$ is the dataset in Table 1?

B. Suppose no noise is added to output of f. What is the probability that $f(\mathcal{D}) \in T$, where $T = [3.0, 3.2]$?

C. Suppose noise drawn uniformly from the range $[-0.5, 0.5]$ is added to the output of f. What is the probability that $f(\mathcal{D}) \in T$, now that f includes the added noise?

A. $f(D) = 3.5$, the average GPA

B. If there is zero noise then the probability that $f(D) \in T$, where $T = [3.0, 3.2]$ is zero. Since the average GPA is 3.5 then there is a 0% chance that 3.5 is between [3.0, 3.2].

C. Now that f includes the added noise, the probability that $f(D) \in T$ is 20%. This is because there is noise drawn uniformly from $[-0.5, 0.5]$ and that means there is an equal chance of the output of f with noise being in the range $[3.0, 4.0]$.

Neighboring Datasets

The formal definition of DP further relies on the concept of neighboring datasets. Two datasets $\mathcal{D}$ and $\mathcal{D}'$ are neighbors if and only if they differ in the inclusion or exclusion of one data point (e.g. the data for one individual).

8. Answer the following questions about neighboring datasets based on Table 1.

A. Describe an example of a neighboring dataset in this scenario.

B. What is the difference in average GPA between the original table and the table without Bob?

C. What is the maximal difference in average GPA between the original table and any neighboring dataset?

A. An example of a neighboring dataset in this scenario is including an additional entry for student name, GPA with the value Frank, 3.6. Since this dataset D' includes one new data point then D'

is a neighboring dataset for D.

B. The average GPA without Bob is 3.575. So the difference in average GPA between the original table and he table without Bob is -0.075 (3.5 - 3.575 = -0.075).

C. The maximal difference in average GPA between the original table and any neighboring dataset is about 0.583. This is because if we choose the neighboring dataset that includes an additional entry (Steven, 0.0) then the average GPA becomes 2.9167. Thus, the difference in the original average GPA and the average GPA of the neighboring dataset is 0.583. This is the maximal difference between a GPA of 0.0 is the value, that is still within the GPA scale, with the greatest difference from the original average.

Noise Mechanism

The high-level idea behind DP is that the noise needed to protect privacy depends on the maximal difference between outputs of f across neighboring datasets (as you analyzed in 8(B)). This problem set omits the details, although graduate students will explore a specific noise mechanism in Question 11.

9. Suppose $f(\mathcal{D})$ is the average GPA of $\mathcal{D}$, where $\mathcal{D}$ is the dataset in Table 1, but now we are adding some noise z_i, $i = 1, 2$.

A. Suppose $\mathcal{D}'$ is $\mathcal{D}$ without Bob. What should $z_1 - z_2$ be so that $f(\mathcal{D}) + z_1 = f(\mathcal{D}') + z_2$?

B. Suppose $\mathcal{D}'$ is the neighboring dataset with the farthest average GPA from Table 1. What should $z_1 - z_2$ be so that $f(\mathcal{D}) + z_1 = f(\mathcal{D}') + z_2$?

A. From 8.B, we know that $f(D) = 3.5$ and $f(D') = 3.575$ so we solve for z_1 and z_2 such that $f(D) + z_1 = f(D') + z_2$. So if we plug in our values for $f(D)$ and $f(D')$ we have that

$3.5 + z_1 = 3.575 + z_2$

$z_1 = 0.075 + z_2$

So, we need to pick a value for z_1 and z_2 that satisfies this constraint. So we will choose $z_2 = 0.3$ and therefore we have that $z_1 = 0.375$.

Thus, we have that $z_1 - z_2 = 0.075$, $z_1 = 0.375$, and $z_2 = 0.3$.

B. From 8.C, we know that $f(D) = 3.5$ and $f(D') = 2.9167$ (rounded). So we solve for z_1 and z_2 such that $f(D) + z_1 = f(D') + z_2$. So if we plug in our values for $f(D)$ and $f(D')$ we have that

$3.5 + z_1 = 2.9167 + z_2$

$z_1 = 0.583 + z_2$

So, we need to pick a value for z_1 and z_2 that satisfies this constraint. So we will choose $z_2 = 0.2$ and therefore we have that $z_1 = 0.783$.

Thus, we have that $z_1 - z_2 = 0.583$.

Formal Guarantee

The formal guarantee provided by an ϵ-differentially private noise mechanism is as follows. For a given $\epsilon > 0$, the function f satisfies differential privacy if, for all pairs of neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$, and all output events $T \in \text{Range}(f)$:

$$P(f(\mathcal{D}) \in T) \leq e^{\epsilon} \cdot P(f(\mathcal{D}') \in T)$$

Informally, this definition can be understood as "the results of the function f when applied to any neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$ are close to each other, in a way that an external observer cannot distinguish between

them." The parameter ϵ quantifies the privacy of the algorithm, as a bound on the maximum amount by which the inclusion or exclusion of a single data point can affect the results.

10. Answer the following questions about the formal guarantee provided by differential privacy, based on $\mathcal{D}$ being the dataset in Table 1.

A. Suppose once again that f is average GPA, and that no noise is added to f. If $T = 3.50$, does f satisfy differential privacy for any ϵ?

B. Suppose $\epsilon_1 > \epsilon_2$. Which ϵ provides more privacy? Which requires more noise to be added, on average?

A. No f does not satisfy differential privacy for any ϵ, because we are told that T = 3.50 and $f(D') \neq 3.5$. Then $P(f(D') \in T) = 0$, so the differential privacy inequality becomes

$P(f(D) \in T) \leq e^{\epsilon} * 0$

$P(f(D) \in T) \leq 0$

Since $P(f(D) \in T) = 1$, then the differential privacy inequality is no longer true. Thus f does not satisfy the differential privacy for any ϵ.

B. ϵ_2 provides more privacy since it tightens the upper bound, which means that D and D' are closer to each and an eternal observer would have a harder time distinguishing between the two datasets. This would also mean that more noise needs to be added for ϵ_2, on average, since this would increase the privacy and require ϵ_2 to be smaller.

Additional exercise for students in 6.3952

This question is only required for students enrolled in the graduate version of the class.

The Laplace mechanism is one way to add noise to satisfy differential privacy. Specifically, for any $f : \mathcal{D} \mapsto \mathcal{R}$, the Laplace mechanism outputs:

$$f(\mathcal{D}) + \text{Lap}(\Delta f/\epsilon)$$

where Δf is the maximum difference in f between $\mathcal{D}$ and any neighboring dataset:

$$\Delta f = \max_{\mathcal{D},\mathcal{D}'} |f(\mathcal{D}) - f(\mathcal{D}')|$$

and Lap(b) is a random draw from a Laplace random variable with mean 0 and scale parameter b. The probability density function of the Laplace distribution with mean 0 and scale parameter b is given by:

$$p(z \mid 0, b) = \frac{1}{2b} \cdot \exp\left(-\frac{|z|}{b}\right)$$

11. Prove that the Laplace mechanism satisfies ϵ-DP. Specifically, prove that the following expression is true for any output event $T \in$ Range(f):

$$\frac{P\left[f(\mathcal{D}) + \text{Lap}(\Delta f/\epsilon) = T\right]}{P\left[f(\mathcal{D}') + \text{Lap}(\Delta f/\epsilon) = T\right]} \leq e^{\epsilon}$$

Hints:

- Can you replace the ratio of probabilities with the ratio of PDFs?
- How do you simplify $\frac{e^a}{e^b}$?
- To simplify, you will need the reverse triangle inequality: $|a| - |b| \leq ||a| - |b|| \leq |a - b|$, where a is $|T - f(\mathcal{D}')|$ and b is $|T - f(\mathcal{D})|$.

We will define our scale parameter b to be $b = \Delta f/\epsilon$. Now, we will rewrite the probabilities as the following

$P[f(D) + Lap(b) = T] = P[Lap(b) = T - f(D)]$

$P[f(D') + Lap(b) = T] = P[Lap(b) = T - f(D')]$

So we can say that we are finding the probability that the Laplacian distribution equals $T - f(D)$ and $T - f(D')$. So using the definition of a Laplacian distribution, $p(z|0, b) = \frac{1}{2b}exp(-\frac{|z|}{b})$. We will define z as $z = T - f(D)$ and $z = T - f(D')$ for the different probabilities. Now we write

For $f(D)$
$Lap(\Delta f/\epsilon) = p(T - f(D)|0, \Delta f/\epsilon) = \frac{1}{2(\Delta f/\epsilon)}exp(-\frac{|T-f(D)|}{\Delta f/\epsilon})$

For $f(D')$
$Lap(\Delta f/\epsilon) = p(T - f(D')|0, \Delta f/\epsilon) = \frac{1}{2(\Delta f/\epsilon)}exp(-\frac{|T-f(D')|}{\Delta f/\epsilon})$

Next we can plug in these values and get

$\frac{P[f(D)+Lap(\Delta f/\epsilon)=T]}{P[f(D')+Lap(\Delta f/\epsilon)=T]} = \frac{\frac{1}{2(\Delta f/\epsilon)}exp(-\frac{|T-f(D)|}{\Delta f/\epsilon})}{\frac{1}{2(\Delta f/\epsilon)}exp(-\frac{|T-f(D')|}{\Delta f/\epsilon})}$

$= \frac{exp(-\frac{|T-f(D)|}{\Delta f/\epsilon})}{exp(-\frac{|T-f(D')|}{\Delta f/\epsilon})}$

$= exp(-\frac{|T-f(D)|}{\Delta f/\epsilon} + \frac{|T-f(D')|}{\Delta f/\epsilon})$

$= exp(\frac{|T-f(D')|-|T-f(D)|}{\Delta f/\epsilon})|)$

Using the reverse triangle inequality we can say

$\leq exp(\frac{|T-f(D')-T+f(D)|}{\Delta f/\epsilon})|)$

$= exp(\frac{|f(D)-f(D')|}{\Delta f/\epsilon})|)$

$= exp(\frac{\epsilon|f(D)-f(D')|}{\Delta f})|)$

We know that Δf is defined as $\Delta f = max_{D,D'}|f(D) - f(D')|$, so we can say that

$\leq exp(\frac{\epsilon\Delta f}{\Delta f})) = e^{\epsilon}$

Thus, we have shown that

$\frac{P[f(D)+Lap(\Delta f/\epsilon)=T]}{P[f(D')+Lap(\Delta f/\epsilon)=T]} \leq e^{\epsilon}$

Problem 2: Differential Privacy & Unlearning with Census Data

Note: This problem is directly built on the activity in Recitation 6 on Oct 24.

Optional Background Reading:

- Why the Census Bureau Chose Differential Privacy (2020 Census Briefs)
- Differential Privacy for the 2020 Census (Gong et al. 2022)
- Membership Inference Attacks Against Machine Learning Models (Shokri et al. 2017)
- Algorithms that Approximate Data Removal (Suriyakumar & Wilson, 2022)

As you explored in Problem 1, differential privacy (DP) is a fundamental technique for protecting individual records in datasets. Recall that the formal guarantee provided by an ϵ-differentially private noise mechanism is as follows. For a given $\epsilon > 0$, the function f satisfies differential privacy if, for all pairs of neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$, and all output events $T \in \text{Range}(f)$:

$$P(f(\mathcal{D}) \in T) \leq e^{\epsilon} \cdot P(f(\mathcal{D}') \in T)$$

This problem builds upon the activity in Recitation 6, where we compared classifiers trained with and without DP on the Iris dataset and assessed their resilience to database reconstruction tasks. However, in this problem, we will apply differential privacy to the Census Adult dataset, and test resilience to a different type of attack: membership inference. We will also briefly explore the concept of unlearning, which is another way to ensure the privacy of individual records in datasets.

To complete this exercise, you will only need the Colab for certain parts below, and you will provide all your answers in this LaTeX document.

Exploratory Data Analysis

In this problem, we will use the Census Adult dataset, which contains demographic data about individuals across the US. One goal from using this dataset is to be able to predict whether or not the annual income of an individual exceeds \$50K/yr. Please upload the files "adult.data" and "adult.test" to your colab before starting the exercise.

1. In the notebook, run the cells in the "Exploratory Data Analysis" section (you do not need to change any code, just analyze the visualizations). Then, answer the following questions.

A. Based on the visualization of income and highest-level of education, which education levels have more individuals earning over \$50K compared to those earning \$50K or less?

B. Based on the visualization of income and gender, roughly what percentage of women earn over \$50K?

C. Based on the histogram of age, what is the shape of the distribution? Why is it shaped this way?

D. Based on the histogram for hours worked per week, what is the shape of the distribution? Why is it shaped this way?

A. Based on the visualization, individuals with a Masters, Doctorate, and Prof-school Degrees as the highest-level of education have more individuals earning over \$50K compared to those earning \$50K or less.

B. Based on the visualization of income and gender, roughly 10% of women earn over \$50K.

C. Based on the histogram of age, the distribution is skewed to left side of the ages with a peak at about 35 years old. It is shaped this way because young adults are (20-40) are the typical working age individuals and indivudal between the ages of 50 to 60 years old using start retiring. So there is less representation for older individuals since they typically are not working.

D. Based on the histogram for hours worked per week, the distribution has a peak at 40 hours with the majority of the individuals working 40 hours. Then there is a small spread of individuals that work less than 40 hours and individuals that work more than 40 hours.

Building Models with DP

Our goal is to build a differentially-private classifier of whether a person's annual income is greater or less than \$50K. For our implementation, we will use a **logistic regression** classifier. For logistic regression[4], differential privacy can be achieved by adding carefully calibrated noise to the learned vector of weights θ. The noise helps protect individual data points from being identified, while still keeping the model accuracy similar to the original model. Recall that logistic regression finds a vector θ such that, given a vector of features x_i, the model estimates the probability of $y_i = 1$ as follows:

$$\Pr(y_i = 1 \mid x_i) = \frac{1}{1 + \exp(-\theta^\top x_i)}. \tag{1}$$

In the notebook, run the code in the "Building Models with DP" section to answer the following questions.

2. The following questions are based on the cell that asks you to try different ϵ. Report accuracy to 5 decimal places.

A. What is the accuracy of the baseline model without DP?

B. What is the accuracy of the model with $\epsilon = 0.01$?

C. What is the accuracy of the model with $\epsilon = 0.1$?

D. What is the accuracy of the model with $\epsilon = 1.0$?

E. As ϵ increases, will the accuracy "increase", "decrease", or "stay the same" on average? Briefly justify your answer.

A. 0.82039 is the baseline model accuracy

B. The accuracy of a DP model with epsilon = 0.01 is 0.44523

C. The accuracy of a DP model with epsilon = 0.1 is 0.72056

D. The accuracy of a DP model with epsilon = 1.0 is 0.81597

E. As ϵ increases the accuracy will increase on average because this means less privacy which means less noise. With less noise and less privacy then this will result in the logistic regression model having better accuracy.

3. The following questions are based on the visualization of accuracy and different ϵ.

A. For what ϵ are the model predictions essentially a random guess?

B. For what ϵ does the accuracy match the baseline model without DP?

C. Your curve may have some fluctuations (e.g. it is not smooth). In general, why would we expect this to be the case?

D. How would you recommend that a real-world deployment choose ϵ? What are the trade-offs for different values of ϵ?

[4]For additional background, see this paper by Chaudhuri et al.

A. Values of $\epsilon \leq 10^{-9}$ have model predictions in which the predictions are essentially a random guess because the accuracy is about 50% or worse.

B. For an $\epsilon \geq 10^{0}$, the accuracy of the model with DP matches the accuracy of the baseline model without DP.

C. We expect this because we are using noise with the differential privacy logistic regression model, so the noise will potentially be different for each new epsilon value. And the noise is random so this causes the curve to have some fluctuations.

D. For a real-world deployment I would choose ϵ such that you optimize both privacy and accuracy. So in this case, based on the plot, I would choose an $\epsilon = 10^{-1}$. Since the accuracy is about 75% and this ϵ value provides some privacy compared to the baseline model without differential privacy.

Membership Inference Attacks

The goal of differential privacy is often to protect against adversarial attacks that may be able to glean some sensitive information about the dataset. One type of attack is **membership inference attacks** (MIA), which aim to determine whether a specific data point (e.g., an individual) is in a dataset.

For our setting, consider the salary prediction model that we have trained using the Census data. An adversary may want to know whether a particular individual was included in the training data used for this model. In this case, the salary prediction model that we have built is the "target model." The adversary's goal is to build an "attack model" which tries to infer whether a data point was included in the training set of the "target model". In order to do this, the adversary needs access to some information from the target model, such as its predictions. For our setting, we make two assumptions:

- Assumption 1: The adversary can make unlimited queries to the target model, and thus collect predictions for many data points.
- Assumption 2: The adversary has access to a subset[5] of the training data (i.e. it knows the membership status of some data points).

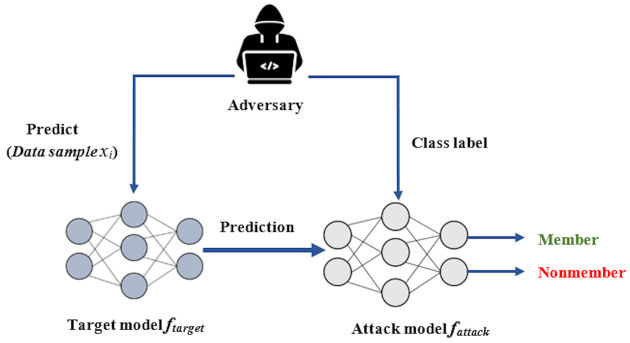

Figure 1: Membership inference attack: the adversary uses the target model's outputs to determine if a specific data point was a member of the training set.

4. In the notebook, use the code in the "Membership Inference Attack" section to answer the following questions. The code will report the accuracy of the attack model (i.e. how well it can determine if a point was in the training set). Report accuracy to 5 decimal places. Note that due to the noise in the process, you might see fluctuations in your results.

A. What is the attack model accuracy for a model trained with $\epsilon = 0.01$?

[5]In our implementation, we make the stronger assumption that the adversary knows a subset of data points in the training set and a subset of data points not in the training set, and that these subsets have data points that are sufficiently different.

B. What is the attack model accuracy for a model trained with $\epsilon = 0.1$?

C. What is the attack model accuracy for a model trained with $\epsilon = 1.0$?

D. As ϵ increases, will the attack model accuracy "increase", "decrease", or "stay the same" on average? Briefly justify your answer, and compare it to your answer in 2(E).

A. The attack model accuracy with epsilon = 0.01 is: 0.65690.

B. The attack model accuracy with epsilon = 0.1 is: 0.63520

C. The attack model accuracy with epsilon = 1.0 is: 0.77962

D. As ϵ increase the attack model accuracy will increase on average because their is less privacy as ϵ increases so it is easier for the attack model to accurately determine membership of a data point in the training set. This is similar to my answer from 2(E), because the logistic regression model accuracy increased as ϵ increased for similar reasons. As ϵ increases, privacy decreases, and therefore accuracy increases.

Unlearning

So far, this problem set has focused on differential privacy as a way to protect against membership inference attacks. But if we know certain individuals might be at risk, another approach might be to have the model "unlearn" these training examples.

Unlearning involves taking a trained model and removing the influence of certain training examples (dubbed the "forget set"). Due to computational limitations, unlearning is often more practical than re-training a model from scratch without the "forget set". However, the goal is to make the "unlearned" model as close as possible to the model that could be re-trained from scratch. This process is illustrated in Figure 2.

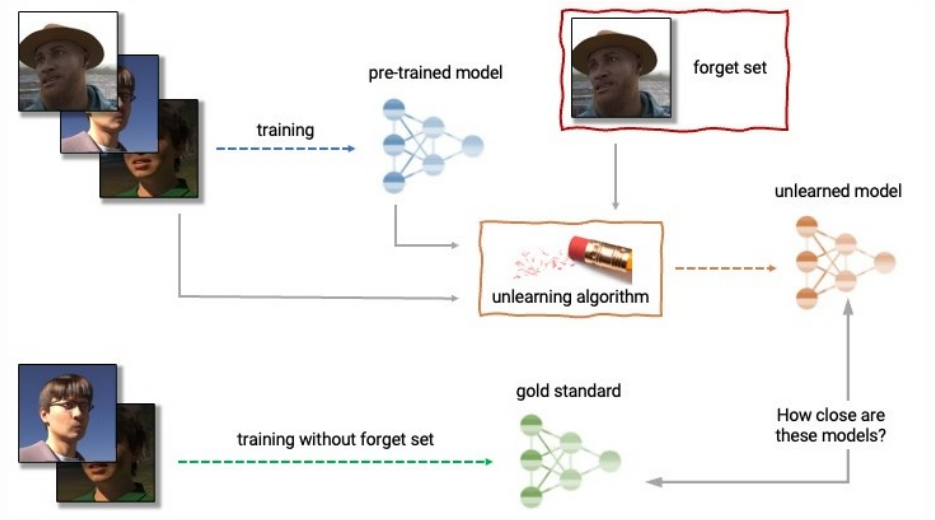

Figure 2: An overview of how models "unlearn" training samples.

In this exercise, you will test a simple unlearning algorithm that continues training the model on a subset of the training set that excludes the "forget set" (also known as the "retain set"). You will then test the performance of the membership inference attack to check if unlearned samples are *indistinguishable* from samples that were not in the training set. Specifically, the notebook will report the accuracy of the membership inference attack on the forget set and samples not from the training set.

5. In the notebook, run the code in the "Unlearning" section to answer the following questions. You do not need to change any of the code in this section, apart from choosing the number of samples to unlearn.

A. After unlearning 500 training examples, what is the accuracy of the membership inference attack?

B. After unlearning 5000 training examples, what is the accuracy of the membership inference attack?

C. As the number of samples to unlearn increases, will the accuracy of the membership inference attack "increase", "decrease", or "stay the same"? Briefly justify your answer.

A. After unlearning 500 training samples:

The model accuracy after unlearning is: 0.8212621435905537

The accuracy of the membership inference attack is: 0.74.

B. After unlearning 5000 training samples:

The model accuracy after unlearning is: 0.8237106073769844

The accuracy of the membership inference attack is: 0.79

C. As the number of samples to unlearn increases, the accuracy of the membership inference attack will increase. This is because there are less samples for the attacker to have to conduct membership inference on. Since the total number of training samples decreases (since the model is unlearning training samples) then the Additionally, I tested the membership inference attack accuracy after unlearning 10,000 training samples and got the following results:

The model accuracy after unlearning is: 0.8256851749466867

The accuracy of the membership inference attack is: 0.83

This showcases the fact that as the number of samples to unlearn increases, the accuracy of the membership inference attack increases.

Problem 3: Differential Privacy and Governance

Differential privacy is often highlighted in governance discussions around AI and privacy, as you will explore in the following reflection questions. **Your responses can be brief (a few sentences for each part).**

1. The following questions are based on this article: Differential Privacy: Issues for Policy-makers . Read the article, then answer the following questions.

A. Why does the article claim the trade-off between privacy and accuracy is not straightforward?

B. What are "linking attacks"? Why do they pose a challenge for legal requirements about privacy?

C. The article claims that "any deployment of differential privacy requires a value judgment." What does this mean? Provide an example based on the Census.

A. The article claims that the trade-off between privacy and accuracy is not straightforward, because smaller ϵ means more noise, which leads to worse accuracy. If the accuracy is worse then this means the statistics released or model trained from a noisy dataset will be less accurate and might invalidate some results. That is why the tradeoff between having more or less privacy vs accuracy is complex and depends on the context in which the model is being used.

B. Linking attacks are situations in which an adversary aggregates multiple data releases. So even though a single dataset might be anonymized, a linking data can aggregate multiple anonymized datasets together to reconstruct personal information. This poses a challenge for legal requirements about privacy because there is an assumption that anonymizing a dataset is enough to prevent a person's personal information being released. However, a linking attack breaks that assumption. It makes it difficult to protect individuals when anonymizing a single dataset would ultimately not prevent their information from be identifiable.

C. This means that deployment of DP requires people to incorporate societal norms and policies into the DP process to know whether to value privacy or accuracy more. An example of this based on the Census is that the Census might prioritize acquiring more accurate demographic information in order to inform policy. This would result in having less noise and less privacy for individuals.

2. The following questions are based on the National Institute of Standards and Technology (NIST) SP 800-226 (March 2025) guidelines for differential privacy. Read Section 3.3 "Bias" (PDF pp. 35–39), then answer the following.

A. How can differential privacy magnify disparate impacts on small groups? Refer to **Figure 14** (group counts with DP confidence intervals) and **Figure 15** (classifier accuracy vs. ϵ).

B. In the context of differential privacy, how does the document distinguish among *systemic bias*, *human bias*, and *statistical bias*?

C. Why can post-processing of DP outputs (e.g., clamping negative counts to zero) introduce *statistical bias*? How does this bias change with ϵ? Refer to **Figure 16**.

A. Differential privacy can magnify disparate impacts on small groups by trading off accuracy and usability of the model for small groups. This is becaused from figure 14 we see that for small group, like American Indian people, the 95% confidence interval is very big meaning the accuracy of the model will be terrible for them. Whereas for a bigger group, like white people, the 95% confidence interval is very small so the model will be fairly accurate for this group. Similarly for figure 15, the minority race confidence interval for accuracy is large and thus the model would have worse accuracy. Therefore, differential privacy magnifies disparate impacts on small groups causing systematic and human bias.

B. In the context of differential privacy, the document says systematic bias as "results from rules, processes, or norms that advantage certain social groups and disadvantage others". It describes

human bias as bias that "results from the heuristics that humans use to make decisions based on data". Finally, it describes statistical bias as a mechanism refers to a difference between the true query result $f(x)$ and the expected value (i.e., the average over many samples) of the mechanism's output". The distinguishing factor is that human bias is derived from data backed decisions based on societal norms, whereas systematic bias are laws implemented to help some social groups and disadvantage other social groups. Finally, statistical bias occurs when the true data value is different from the average value.

C. Post-processing of DP outputs can introduce statistical bias because changing the negative counts to zero would affect the average or the expected value of the population. By clamping the negative counts then the new average is different then the actual count so this introduces statistical bias. Based on Figure 16 the bias decreases as ϵ increases. So in order to reduce the statistical bias a bigger ϵ value is better.

3. Read the news then answer the following:

A. **Wayback Machine and the privacy–accountability tradeoff**
Pre-read: Reddit will block the Internet Archive's Wayback Machine from indexing most content (Aug. 2025).
In **4-5 sentences**, explain how web archiving changes the privacy calculus for users who later delete posts (i.e., why accessibility is not a simple public/private binary). Do you think platforms should be able to block archiving to reduce privacy harms and AI scraping, or does that undermine transparency and research? State one concrete safeguard (technical or policy) you would require of an archiving service.

B. **From car sensors to insurers: safeguarding driver privacy**
Pre-read: FTC takes action against GM for sharing precise location and driving data (Jan. 2025).
In **4-5 sentences**, summarize the FTC's allegations and the key requirements in the proposed order. Then answer the **Question:** In California, should sharing precise location and driving behavior with insurers be *opt-in* by default? What is the *minimum* set of items the sign-up screen must show (plain-language notice; a separate checkbox to allow insurer sharing; a link/button to opt out or limit later; a short "no penalty for saying no" note)?

A. Web archiving changes the privacy calculus for users who later delete posts because a user's deleted post may still exist and be accessible indefinitely. Even though a user might have deleted a post on their end, someone else could have archived a copy of the post. This means that the deleted copy could still be somewhere else on the platform. That is why the content may persist in an archived state and accessibility is not a simple public/private binary state.

I think platforms should not block archiving. I think this would undermine transparency of individuals and potential research opportunities.

One safeguard that I would require of an archiving service is to have an opt-in and opt-out feature, which allows users to opt-in or out of having their posts, deleted or not, used for training AI models.

B. The FTC's allegations are about General Motors(GM) and its affiliate OnStar collecting data about a person's precise driving habits and location. The FTC alleges that GM and OnStar collected this data in a frequent manner and sold it to consumer reporting agencies and insurers. They sold this data without sufficient consent from the driver. The key requirements in the proposed order are obtain affirmative express consent before data collection and sharing, provide consumers access to data, allow consumers to be able to opt-out of data collection/sharing, and they are banned from sharing anymore data to consumer agencies for five years.

In California, sharing precise location and driving behavior should NOT be opt-in by default, because a consumer may not be aware of the fact that their data is being used without their

consent. The minimum set of items the sign-up screen must show is text explaining that there is no penalty for saying no to data sharing/collection, text saying what will happen with their data and who will be able to view their data, a toggle switch that indicates whether the consumer wants to opt-out or opt-in to data collection/sharing, and their data must be deleted every 30 days.