20 Privacy

As companies, and governments, and researchers are getting better at collecting data, what if this data is about you?! Google, Facebook, Yahoo! and most other web companies make money by understanding their customer base, and applying targeted advertising. The government would like to identify terrorists before they strike. Researchers would like to study properties of large data sets. But if *you* are the subject of the data, are you ok with your data being available to be mined?

How can we preserve the "privacy" of individual data points while also being able to mine the data?

This goal has a long history. Since privacy is hard to define, there have been many attempts which have come up short in achieving their intended goal. As researchers devise new standards by which data should be private, others find weaknesses that allow more information than one might have thought. Yet, these attempts are useful for demonstrating the problems and issues.

Ethics and Empathy. Before we get into technical definitions and examples, its worth noting that what follows will not necessary provide solutions to ethical dilemmas – just tools and examples to think about them. As data scientists, **you will at some point face an ethical dilemma!** The only mechanism to prevent bad ethical behavior is societal pressure (shame people for doing bad things, or at least let them know it is bad), and laws. The other way of course, is to hold yourselves to high ethical standards. But one may argue these standards may be evolving, and there is no way to know what the right choice is. I argue that there is a simple rule that you can follow, "if you were the data point, would you want to be treated that way." Basically don't hide behind technology to justify your ethical decisions. The best way I've seen this expressed is in the following video:
`https://www.youtube.com/watch?v=5HbYScltf1c`

20.1 Attempts to Define Privacy

What does it mean to have privacy?

Here is the prototypical problem. A company C wants to release a dataset D about its costumers habits, first to its own data scientists, but then they may want to share with researchers in universities. The goal is to develop new ways to extract information to improve business (e.g. advertising, product pricing).

So the data scientists want to compute some statistics / mine some structure from D, but not to be able to identify individual's information.

Example: health records *Consider an "anonymized" survey of hospital patients. Only provide zip code to identify location. Some have cancer. What if a CEO went for routine treatment on a treatable cancer, but was the only person in that zip code (e.g. the one with big houses ... she is a CEO after all). If someone knows the CEO was part of the study, they may know they have cancer, and it can cause the stock to plummet!*

STORY TIME:

- In 2000, Massachusettes released all stated employee's medical records in an effort for researchers to be able to study them.
- They wiped all ids, but kept zip codes, birthday, gender. Was declared anonymized by the government.
- In Massachusettes, it was possible to buy voter data for $20. It has names, birthday, zip codes, and birthdays of all voters.

- A grad student, Latanya Sweeney combined the two to identify the governor of Massachusettes. Story is, she mailed him his own health records!
- Dr. Sweeney now teaches at Harvard.

This demonstrates the danger of other data sets D' which are released or available independently of the one D you release. How can we circumvent this? There was a series of several attempts:

k-anonymity: With datasets D, one can only identify someone up to k other people, for some parameter k. That means, there must be at least k people with the same public set of tests in the same zip code / birthday / gender if D is released. Otherwise, some data is withheld.

Teacher evaluations work like this, I only see them if there are at least $k = 4$ people who respond.

ℓ-diversity: In this setting, for each of the anonymization groups (with at least k people), these exists someone from each one of ℓ-well separated classes. If you cannot identify a single person (the governor), but know they are one of k people, and *all* of those people have cancer, then you still know the governor has cancer. Instead, ℓ-diversity ensures that some person in each such set of k is in one of ℓ distinct clusters, and thus has the hidden attributes different.

t-closeness: This starts with ℓ-diversity, but also ensures that within each anonymization group, the set needs to look like the full data set D among the hidden parameters (less than distance t in Earth Movers Distance).

But none of these can completely prevent corruption from outside data. You never know what other information is available.

What if height was an important quantity? (say for movie star, Silvester Stallone?)

- Information: *Sly Stallone is same height as average New Jersey man*
- Independent survey: *Average New Jersey man is 5' 8"*

This gives away Stallone's height? (Or did he have surgery to become taller ?)

STORY TIME: Netflix Prize

- In 2006, Netflix released awesome data sets $D_1 = \{\langle$user-id, movie, date of grade, grade$\rangle\}$. And another set $D_2 = \{\langle$user-id, movie, date of grade$\rangle\}$. Wants researchers to develop algorithm to predict grade on D_2. (Had another similar private data D_3 to evaluate grades – cross validation.)
- If certain improvement over Netflix's algorithm, get \$1 million!
- Led to lots of cool research!

- Raters of movies also rate on IMDB (with user id, time stamp)
- Researchers showed that by linking who rated similar sets of movies, with similar scores and times, they could identify many people.
- (maybe watched embarrassing films on Netflix, not listed on IMDB)
- Class action lawsuit filed (lated dropped) against Netflix.
- Netflix Prize had proposed sequel, dropped in 2010 for more privacy concerns.

20.2 Differential Privacy

These scenarios has led to a new approach to release data, called *differential privacy*. The goal is that given two very similar data sets D_1 and D_2,

(a) both D_1 and D_2 have similar global statistical properties, so one can run global analysis on them and get *roughly* the same result.

(b) given a particular data point in a data set D_1, one cannot know its true value.

The main way this is achieved is by carefully adding random (usually Laplace noise) to a data set D before it is released; so it could be D_1 or D_2.

More formally lets consider a concrete example. Let the two similar databases D_1 and D_2 each consist of n bits in $\{0, 1\}$. We enforce that D_1 and D_2 only differ in one bit.

Now consider a query q which we can apply to D_1 or D_2.

If $\mathsf{Ham}(D_1, D_2) = 1$ (by above assumption) then

$$\frac{\mathbf{Pr}[q(D_1) \in R]}{\mathbf{Pr}[q(D_2) \in R]} \leq \exp(\varepsilon) \approx 1 + \varepsilon.$$

and we say the query q is ε-differentially private. Here R is some event we want to condition on (say the number of 1 bits in an interval is above 10).

Interactive Approach: The first main option of how to apply differential privacy is an interactive option. Do not publish a database D_1 directly for all to use. Rather, one keeps it behind a firewall, and allow a public interface that only allows fixed sorts of queries, known to be safe (e.g. ε-differential private). Then in each returned answer, the database can figure out the true result, and add a small amount of noise.

For many types of queries, if one is able to ask enough queries, then they can eventually reconstruct the entire database. So the database needs to keep track of which queries have been asked, and prevent associated informative queries. Or, just limit the number of queries. Ideally, one would need an exponential number of queries to accurately reconstruct the database.

Non-Interactive Approach: In this option, we actually publish a data set D_1, but not the original dataset D. Typically, $D_1 = D + \mathsf{Lap}(\varepsilon)$ where $\mathsf{Lap}(\varepsilon) = c \exp(-\varepsilon)$ is some form of Laplacian noise. The details are subject to the particular type of data in D. The goal is so that any set query $q \in R$, including ones that would identify a single point, may give different answers on another database D_2, but that as a whole D_1 and D_2 are similar.

Here is a concrete but very simple example. Let D be a single number (e.g. Sylvester Stallone's height, perhaps 68 inches). We consider two cases, $D_1 = 68$ and $D_2 = 67$. We release one of these two with Laplacian noise (and announce this to the world), to create either D_1' or D_2'.

Now we want that

$$\mathbf{Pr}[D_1' \geq 70] \approx \exp(-2\varepsilon)$$

and

$$\mathbf{Pr}[D_2' \geq 70] \approx \exp(-3\varepsilon)$$

where the range R is ≥ 70 (e.g. 5 foot 10 inches). *So the probability Stalone is above 5'10".*

Now we see that

$$\frac{\mathbf{Pr}[D_1' \geq 70]}{\mathbf{Pr}[D_2' \geq 70]} = \frac{\exp(-2\varepsilon)}{\exp(-3\varepsilon)} = \exp(\varepsilon) \approx 1 + \varepsilon,$$

And thus D_1 is ε-differentially private.

20.2.1 Adding Your Data

Another view of this, is not the value of your data, but is your data in the database. Can an adversary not tell if you data is in the data base or not. You want that any allowable query will change by at most $(1+\varepsilon)$.

An example is *subset sum*. Let $D = [01010101000111001]$ be an n-bit vector (saying whose data is inside with 1). Let a query be a bit mask $q = [1011001101011010 1]$. Then $|D \text{ XOR } q| = 0+0+0+1+0+0+0+1+0+0+0+1+1+0+0+0+1 = 5$.

If we enforce that $\|D - D'\| \leq 1$ (that is they differ by at most 1 bit, then it follows that $|q(D) - q(D')| \leq 1$ also.

We can make the answer to a query differentially private by adding Laplacian noise. $A(D,q) = q(D) + \mathsf{Lap}(1/\alpha)$. Now

$$\frac{\mathbf{Pr}[A(D,q)=x]}{\mathbf{Pr}[A(D',q)=x]} = \frac{\mathsf{Lap}(|x-q(D)|)}{\mathsf{Lap}(|x-q(D')|)} = \exp(\alpha(|x-q(D)|-|x-q(D')|)) \leq \exp(\alpha|q(D)-q(D')|) \leq \exp(\alpha) \approx 1+\alpha.$$

Big Picture. The final picture should be clear.

More noise implies that data is more private, but less informative.

The underlying hypothesis, is that if you are trying to learn a model, you should want your model to be robust. If it is robust, then adding removing a small amount of low-level noise should hopefully not affect the result. Then you should be able to apply differential privacy to hide individual elements without spoiling the global analysis.

It is currently the subject of active research to understand the effective of this trade-off, and when and how this can be made practical and useful.