



NYU

Center for
Data Science

12: Socio-cultural impact

DS-GA 1004: Big Data



This week

- Socio-cultural impact of recommender / information retrieval systems
- Privacy and de-anonymization
- Differential privacy

Socio-cultural impact of recommender and information retrieval systems

Claim: Recommender systems might have killed BuzzFeed

TECHNOLOGY

The Internet of the 2010s Ended Today

BuzzFeed News was more than a website: It defined an era.

By Charlie Warzel



The Associated Press @AP · Apr 20

Pulitzer Prize-winning digital media company **BuzzFeed** will shut down its news division as it cuts about 15% of its entire staff. The company will now have one remaining news brand, HuffPost.



APRIL 20, 2023

SHARE ▼ SAVED STORIES ↗ SAVE 📌

THE MEDIA

BuzzFeed News Is Dead

Who is really to blame?

BY HILLARY FREY APRIL 20, 2023 · 5:16 PM



Founder and CEO of BuzzFeed Inc. Jonah H. Peretti speaks during BuzzFeed Inc.'s listing day at Nasdaq on Dec. 6, 2021. Bennett Raglin/Getty Images

Echo chambers and filter bubbles

- Recommender systems rely on **similarity**
- Over time, this **reduces the diversity** items that a user sees
 - Users tend to herd together around similar items
- This is called a **filter bubble** [Pariser, 2011]
 - Best case: users get only what they like
 - Typical case: user gets **bored and leaves**
 - Worst case: user becomes **isolated and polarized (extreme echo chambers)**



Targeted advertising

- Not all personalization comes through explicit interactions
- Personalized advertising platforms rely on user features
- This can get you into trouble...
 - Age?
 - Gender?
 - Zip code?



[Washington Post, 2019-03-19](#)

Targeted advertising

Federal housing law [prohibits discrimination](#) based on race, color, religion, national origin, gender, disability or family status. Facebook said the new platform will also prevent advertisers from discriminating based on sexual orientation, age, ethnicity and other characteristics covered by state and local civil rights laws.

Facebook classified people according to their demographics, behaviors and interests using terms such as “English as a second language,” “disabled parking permit” or “Telemundo” — which critics argue are proxies for protected categories of people.

- Zip code?



[Washington Post, 2019-03-19](#)

The ethics can be complicated: Representation

- Recommender systems depend on past behavior of the *user population*
- Dealing directly with user data can be sensitive...
 - How is your *user population biased*?
 - How does your *model handle bias*?
 - How are *atypical users* treated?
- Always remember to ask:
 - *Who benefits* from a recommender system?
 - *Is personalization needed*?

Privacy and de-anonymization

Open data and anonymization

- Scientific progress is built on open data, but human data is often sensitive
- Datasets are often **anonymized** by obscuring identifiers, but this may not be enough to protect privacy!
- Even statistical summaries can violate privacy

Commonly used, but insufficient strategies for protecting users in open data

- Obfuscate identifiers:

Jean Valjean \Rightarrow 24601

10 March 2023

Some interesting discoveries in a shared dataset: Néma et al. (2022).

In this post I'm going to be discussing this article, but mostly its dataset:

Néma, J., Zdara, J., Lašák, P., Bavlovič, J., Bureš, M., Pejchal, J., & Schvach, H. (2023). Impact of cold exposure on life satisfaction and physical composition of soldiers. *BMJ Military Health*. Advance online publication. <https://doi.org/10.1136/military-2022-002237>

The first and most important problem in the file as it was uploaded, and was in place until a couple of days ago, is that a lot of PII was left in there. Specifically, the file contained *the first and last names and date of birth of every participant*. This study was carried out in the Czech

- Perturb observations: $X[100, 'height'] += N(0, 1)$
 - Can impact accuracy
 - Can sometimes be undone
- Limit per-user information
 - E.g., **k-anonymity** [Sweeney, 2002]: each attribute occurs in at least k rows, can't distinguish $k-1$ rows
- Statistical summaries instead of individual records
 - Can still leak individual information!

What is a de-anonymization attack?

- Say we release an “anonymized” dataset $R = (r_1, r_2, \dots, r_N)$
- Given some partial, potentially inaccurate observations for an individual:
 - Can we determine if the individual was included in R ?
 - Can we recover more information about the individual (that was not released)?
- Narayanan and Shmatikov (2008) demonstrated that this is possible in the Netflix prize dataset!



Internet Movie Database

Why is k-anonymity not enough?



- If each attribute occurs in at least k rows, then **no single attribute is informative**
- But **combinations of attributes are often informative!**
- In large, high-dimensional collections, **this is the rule, not the exception!**
- People are ultra-high-dimensional, and highly idiosyncratic, which is often reflected in data about people.

List of personal attributes (there are others)

1. Height
2. Date of birth
3. Gender
4. Race
5. Sexual orientation
6. Eye color
7. Hair color
8. Chronic medical conditions
9. Education
10. Sibship status
11. Location of birth (zip code, if possible)
12. Location of current residence (zip code, if possible)
13. Marital status
14. Number of children
15. Job title
16. Approximate household income
17. Homeowner or renter
18. Primary language spoken at home
19. Religious affiliation, if any
20. Citizenship

Deanonymizing the Netflix prize: How the attack worked

[Narayanan & Shmatikov, 2008]

- Step 1: define similarity between rows as **average similarity** over their **support**

$$\text{Sim}(R_u, R_v) = [\sum_i \text{Sim}(R_{ui}, R_{vi})] / |R_u \cup R_v|$$

- Step 2: Given a partial observation q , compute similarity to each row.
- Step 3: Determine a threshold by comparing top score to second highest score
 - If sufficiently large difference, report a match, otherwise, report no match
- **Question: how much partial data do we need to succeed?**

The similarity measures

- Per-user similarity: $\text{Sim}(R_u, R_v) = [\sum_i \text{Sim}(R_{ui}, R_{vi})] / |R_u \cup R_v|$
- Boolean **observation similarity** if ratings are within a threshold:
 - $|R_{ui} - R_{vi}| = 0$ (identical ratings)
 - $|R_{ui} - R_{vi}| \leq 1$ (off by at most 1 star)
- For timestamps, thresholds of 3 or 14 days
- Matching threshold: $\text{Sim}(q, R_1) - \text{Sim}(q, R_2) > 1.5 \cdot \sigma_w(\text{Sim}(q, R_w))$

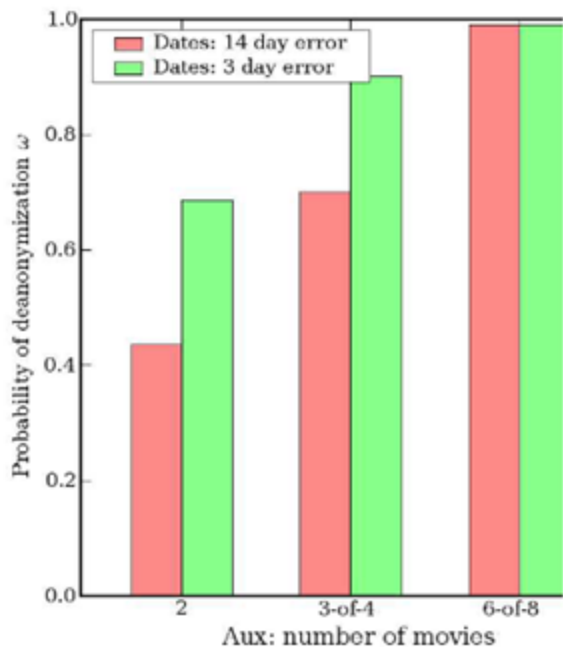


Figure 1: De-anonymization: adversary knows exact ratings and approximate dates.

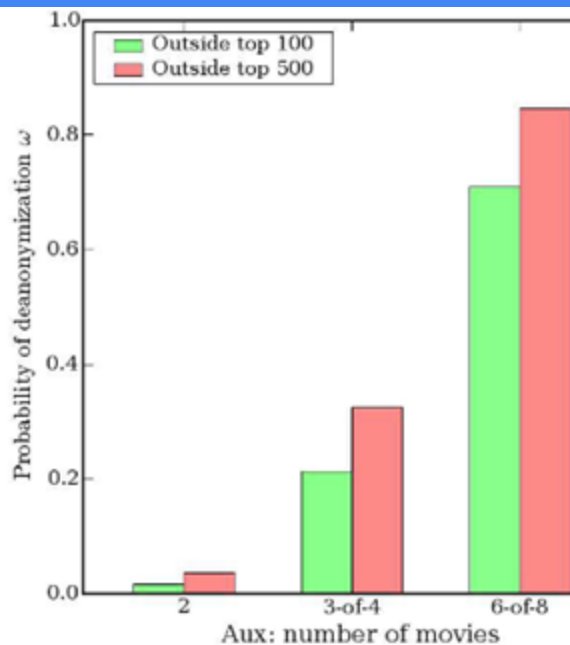


Figure 4: Adversary knows exact ratings but does not know dates at all.

With 8 ratings (2 may be wrong) and 14-day error on rating date, 99% of records can be uniquely identified!

Even without timestamps, unpopular items are highly informative!

How big of a deal is the fact that almost all records could be uniquely identified in this way?

Be honest

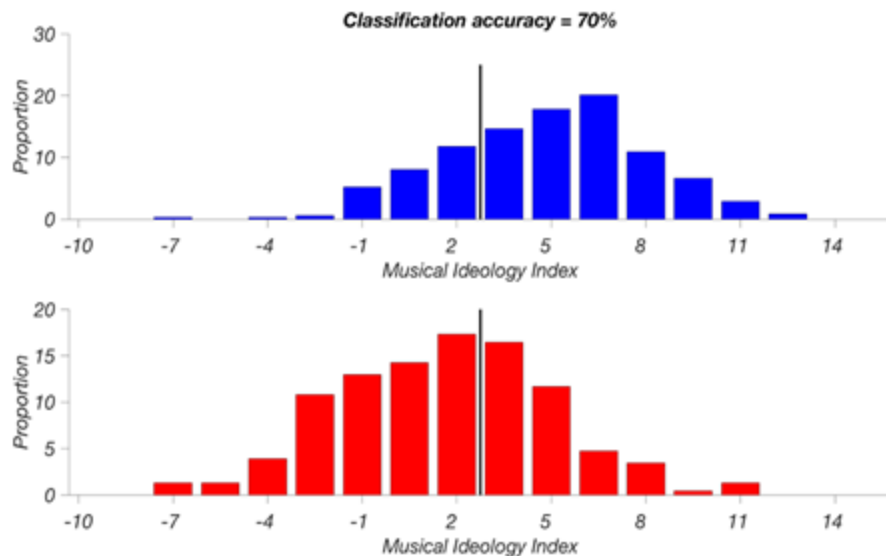
Why does this matter?

- Some comments from Narayanan and Shmatikov:
 - “The privacy question is *not* ‘Does the average Netflix subscriber care about the privacy of his movie viewing history?,’ but ‘Are there *any* Netflix subscribers whose privacy can be compromised by analyzing the Netflix Prize dataset?’”
 - Entertainment preferences can correlate with personal information (political leanings, sexual orientation, religious affiliation, etc...)
 - “The simple-minded division of information into **personal** and **non-personal** is a false dichotomy.”
- Breaches are **irrevocable** and may have implications for a **user’s future privacy**

It might be - there are all kinds of interesting correlations...

The Songs of My People: Appraisal Differences of Popular Music as a Function of Ideology

Myers, C.F., Spivack, S., Spilka, N.H., Philibotte, S.J., Passman, I.J., & Wallisch, P.



What we did

- Database reconstruction for all 308,745,538 people in 2010 Census
- Link reconstructed records to commercial databases: acquire PII
- Successful linkage to commercial data: putative re-identification
- Compare putative re-identifications to confidential data
- Successful linkage to confidential data: confirmed re-identification
- Harm: attacker can learn self-response race and ethnicity

What we found

- Census block and voting age (18+) correctly reconstructed in all 6,207,027 inhabited blocks
- Block, sex, age (in years), race (OMB 63 categories), ethnicity reconstructed
 - Exactly: 46% of population (142 million of 308,745,538)
 - Allowing age +/- one year: 71% of population (219 million of 308,745,538)
- Block, sex, age linked to commercial data to acquire PII
 - Putative re-identifications: 45% of population (138 million of 308,745,538)
- Name, block, sex, age, race, ethnicity compared to confidential data
 - Confirmed re-identifications: 38% of putative (52 million; 17% of population)
- For the confirmed re-identifications, race and ethnicity are learned correctly, although the attacker may still have uncertainty

Tracking user behavior / recommender systems as potential privacy violations

About a year after Pole created his pregnancy-prediction model, a man walked into a Target outside Minneapolis and demanded to see the manager. He was clutching coupons that had been sent to his daughter, and he was angry, according to an employee who participated in the conversation.

“My daughter got this in the mail!” he said. “She’s still in high school, and you’re sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?”

Tracking user behavior

“With the pregnancy products, though, we learned that some women react badly,” the executive said. “Then we started mixing in all these ads for things we knew pregnant women would never buy, so the baby ads looked random. We’d put an ad for a lawn mower next to diapers. We’d put a coupon for wineglasses next to infant clothes. That way, it looked like all the products were chosen by chance.

Tracking user behavior: Pulling back

- Recommenders rely on past interactions to predict future behavior
- Target's system *inferred and disclosed* pregnancy to a teenager's family
- **Later:** personalized ads were obfuscated and made to look *random*

Do you think Target's use of collaborative filtering was a violation of privacy?

Why?

Now



- Socio-cultural impact of recommender systems
- Privacy and de-anonymization
- **Differential privacy**

Cynthia Dwork

What are we exposing to the world?

- A whole dataset (raw data)?
- A “de-identified” dataset?
- A fixed set of statistics measured from a dataset?
- A statistical model derived from a dataset?
- An API to ask questions of a dataset?

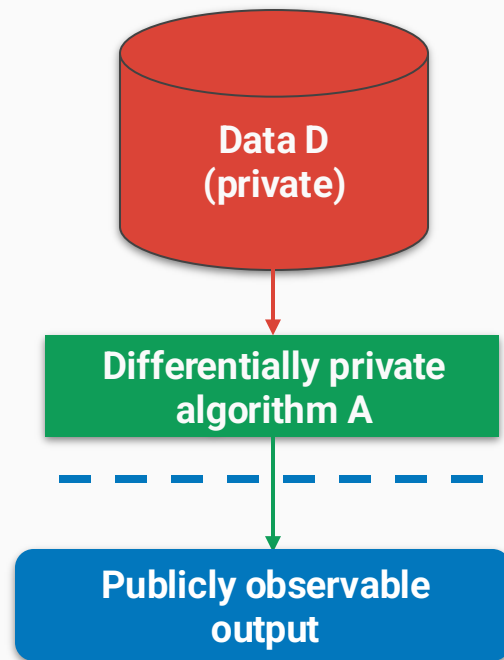
Differential privacy

[Dwork, McSherry, Nissim, Smith, 2006]

- High-level idea:

If an individual is **excluded from the data**,
the result of a computation **should not change**

- We'll achieve this by **randomizing** the computation
- DP is a property of an **algorithm**, not the data!



Differential privacy (definition)

[Dwork, McSherry, Nissim, Smith, 2006]

For any two datasets D and D' differing by one row (i.e., $D' = D + \{x\}$),
a **(stochastic/randomized) algorithm A** is **ϵ -differentially private** if

$$\Pr[A(D) \in S] \leq \exp(\epsilon) \cdot \Pr[A(D') \in S]$$

for any $S \subseteq \text{range}(A)$ (subset of possible outcomes of A)

Differential privacy (definition)

[Dwork, McSherry, Nissim, Smith, 2006]

For any two datasets D and D' differing by one row (i.e., $D' = D + \{x\}$),
a **(stochastic/randomized) algorithm A** is **ϵ -differentially private** if

$$\Pr[A(D) \in S] \leq \exp(\epsilon) \cdot \Pr[A(D') \in S]$$

for any $S \subseteq \text{range}(A)$ (subset of possible outcomes of A)

Definition uses subsets $\Pr[A(D) \in S]$ instead of specific values $\Pr[A(D) = v]$ so that we can support densities and continuous R.V.s.

If that's weird to you, just think of S as a small interval around a specific value v .

Differential privacy (definition)

[Dwork, McSherry, Nissim, Smith, 2006]

For any two datasets D and D' differing by one row (i.e., $D' = D + \{x\}$),
a **(stochastic/randomized) algorithm A** is **ϵ -differentially private** if

$$\Pr[A(D) \in S] \leq \exp(\epsilon) \cdot \Pr[A(D') \in S]$$

for any $S \subseteq \text{range}(A)$ (subset of possible outcomes of A)

Intuition:

If we observe an output value v , we can't reliably tell if it was computed by $v = A(D)$ or $v = A(D')$.

What does ϵ (epsilon) mean here?

- Differential privacy says:

$$\Pr[A(D) \in S] \leq \exp(\epsilon) \cdot \Pr[A(D') \in S]$$

- If ϵ is small (near 0), then $\exp(\epsilon)$ is near $\exp(0) = 1$, and we'll have

$$\Pr[A(D) \in S] \lesssim \Pr[A(D') \in S]$$

- DP is a *symmetric* definition, so we'd also have the reverse:

$$\Pr[A(D) \in S] \gtrsim \Pr[A(D') \in S]$$

- When ϵ is large, the bounds are looser

Tuning the noise:

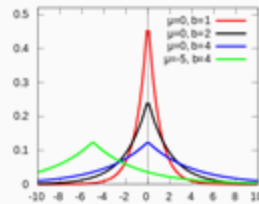
How much noise do we need to add to ensure privacy?

Sensitivity and the Laplace mechanism

- Say we have a vector-valued, private function $f: D \rightarrow \mathbb{R}^d$
 - Example: average across rows of some real-valued attributes
- How different are $f(D)$ and $f(D')$ if D and D' differ by a single row?
 - Define this as the **sensitivity** of f

$$\Delta f = \max_{D, D'} \sum_i |f(D)[i] - f(D')[i]|$$

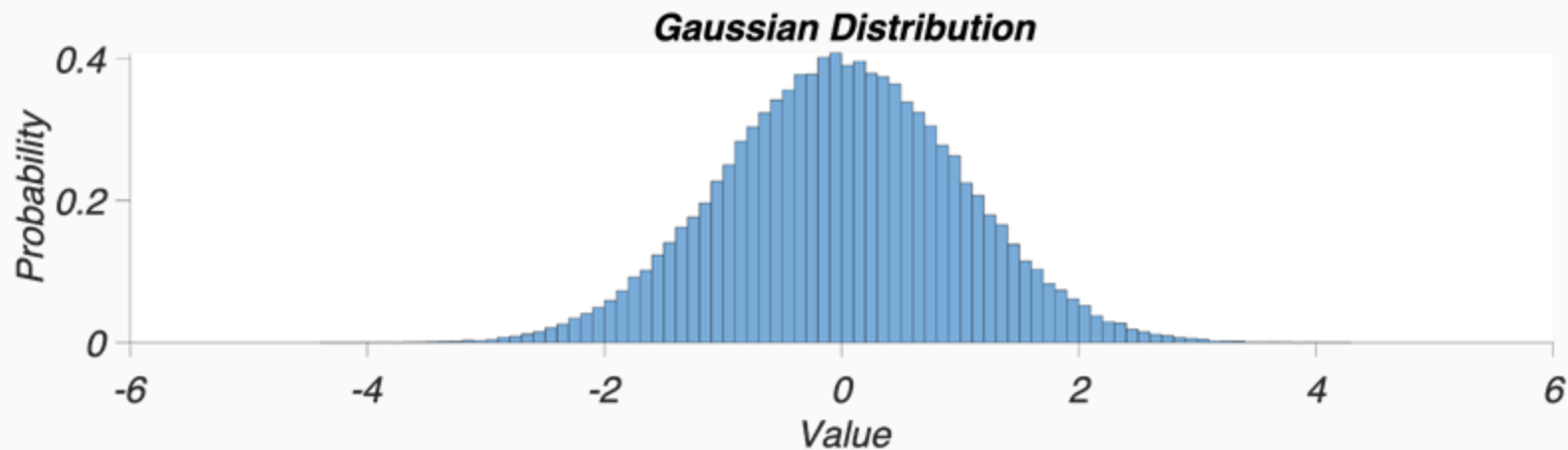
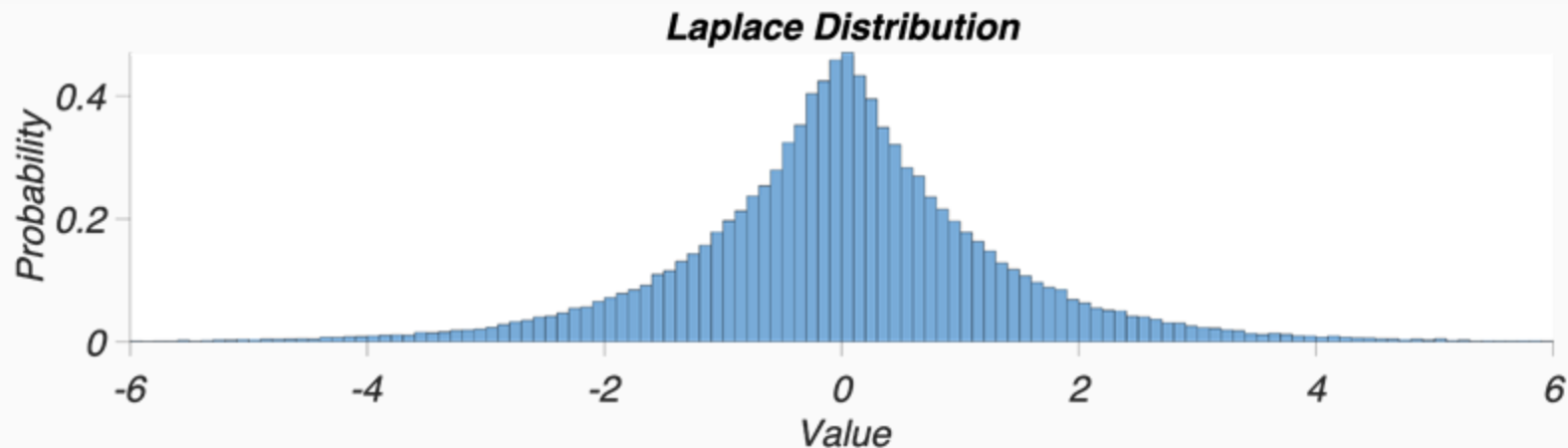
- Let $A(D) = f(D) + z$ where $z[i] \sim \text{Laplace}(0, \Delta f / \epsilon)$.
Then A is ϵ -differentially private!



$z \sim \text{Laplace}(\mu, b)$:

$$p(z) = \exp(-|z - \mu|/b) / 2b$$

Brief aside: What's the right kind of noise to add?



Aside: Be careful with “AI” (a paradigmatic case of hallucinations I saw in the wild)

What can you tell me about the laplace distribution, relative to the gaussian distribution?

wait, i thought the gaussian is falling off exponentially. is this not true?

The Laplace distribution and the Gaussian (or normal) distribution are both continuous probability distributions that describe the behavior of random variables. The key differences.

Laplace distribution:

The Laplace distribution, also known as the double exponential distribution, is characterized by its peakedness at the mean and faster decay in the tails compared to the Gaussian distribution.

$$f(x | \mu, b) = (1 / 2b) * \exp(-|x - \mu| / b)$$

Here, μ is the location parameter (mean), and b is the scale parameter. The distribution is symmetric about the mean, and its tails decay exponentially.

Gaussian distribution:

The Gaussian distribution, also known as the normal distribution, is characterized by its bell-shaped curve, with the mean, median, and mode all coinciding at the center.

$$f(x | \mu, \sigma) = (1 / (\sigma * \sqrt{2 * \pi})) * \exp(- (x - \mu)^2 / (2 * \sigma^2))$$

Here, μ is the mean and $\sigma > 0$ is the standard deviation. The Gaussian distribution is also symmetric about the mean, but its tails decay more slowly compared to the Laplace distribution.

Relative comparison:

1. Tails: The Laplace distribution has heavier tails, which means it assigns more probability to extreme values compared to the Gaussian distribution. The tails of the Laplace distribution decay exponentially, whereas the tails of the Gaussian distribution decay quadratically.

The laplace distribution has heavier tails than the gaussian distribution, however

You are correct, and I apologize for the confusion in my previous responses. The Laplace distribution does indeed have heavier tails compared to the Gaussian distribution.

To reiterate:

- The Laplace distribution's tails decay at a rate proportional to $\exp(-|x - \mu| / b)$. This is a simple exponential decay. Despite its faster decay rate compared to the Gaussian distribution, it is more prone to extreme values, leading to heavier tails.
- The Gaussian distribution's tails decay at a rate proportional to $\exp(- (x - \mu)^2 / (2 * \sigma^2))$. This is a squared exponential decay, which leads to a slower decay rate and thinner tails compared to the Laplace distribution.

In summary, the Laplace distribution has heavier tails compared to the Gaussian distribution, which makes it more suitable for modeling data with more extreme values or noise. I apologize for any confusion caused by my earlier responses.

Apologies for the confusion. I understand why my previous response might have been misleading. The Gaussian distribution does indeed have tails that fall off exponentially, but more specifically, they fall off as a squared exponential. To clarify, let's revisit the probability density function (PDF) of the Gaussian distribution:

$$\frac{1}{\sqrt{2 * \pi}} * \exp(- (x - \mu)^2 / (2 * \sigma^2))$$

In the formula, the exponent in the Gaussian distribution is $- (x - \mu)^2 / (2 * \sigma^2)$, which is a squared exponential decay. This means that the tails of the Gaussian distribution fall off as a squared exponential, which is slower than the simple exponential decay of the Laplace distribution's tails.

So, while the Gaussian distribution's tails do fall off exponentially, they do so as a squared exponential, which is slower than the exponential decay of the Laplace distribution's tails.

What do you mean slower? It's much faster. The laplace distribution has heavier tails

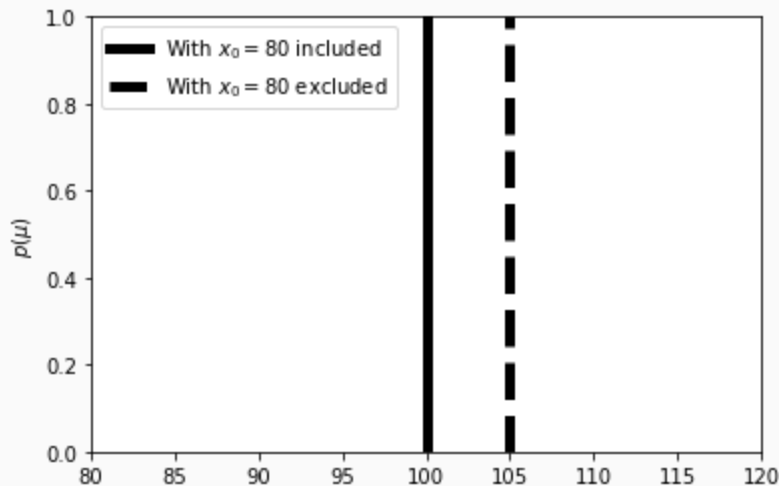
I understand your confusion, and I apologize for the confusion in my previous response. The tails of the Laplace distribution indeed decay faster than those of the Gaussian distribution. To be precise:

- The Laplace distribution's tails decay at a rate proportional to $\exp(-|x - \mu| / b)$. This is a simple exponential decay, which makes the tails of the Laplace distribution lighter and decay faster than the Gaussian distribution's tails.
- The Gaussian distribution's tails decay at a rate proportional to $\exp(- (x - \mu)^2 / (2 * \sigma^2))$. This is a squared exponential decay, which results in slower decay and heavier tails compared to the Laplace distribution.

Thank you for pointing out the inconsistency, and I apologize for any confusion that may have arisen. To summarize, the Laplace distribution has faster decaying, lighter tails than the Gaussian distribution.

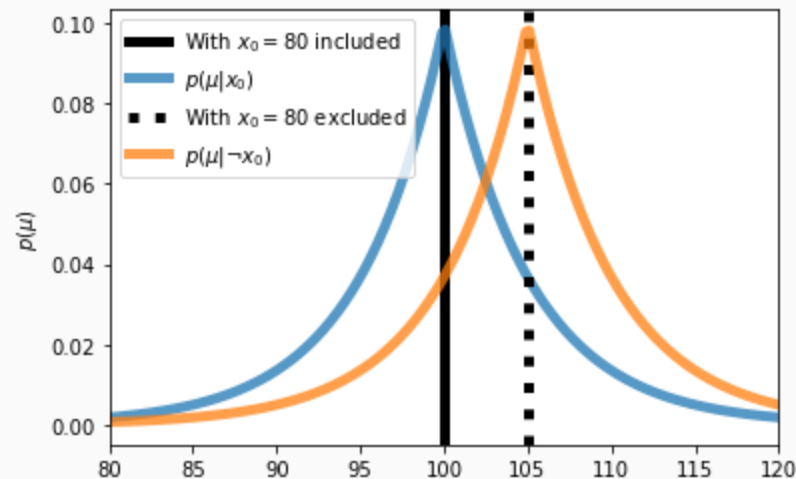
Why is adding Laplacian noise reasonable?

- Say we have data $X = [80, 90, 100, 110, 120]$
- We want to compute the mean
- What happens if $x_0=80$ is included or not?



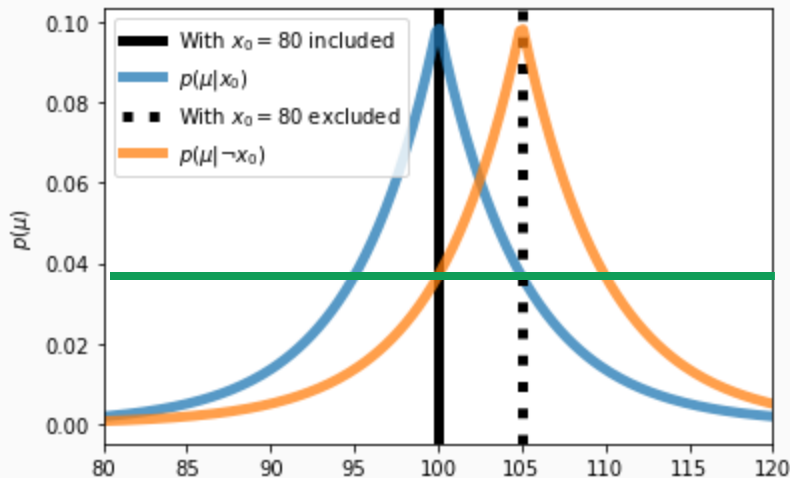
Why is adding Laplacian noise reasonable?

- Say we have data $X = [80, 90, 100, 110, 120]$
- We want to compute the mean
- What happens if $x_0=80$ is included or not?



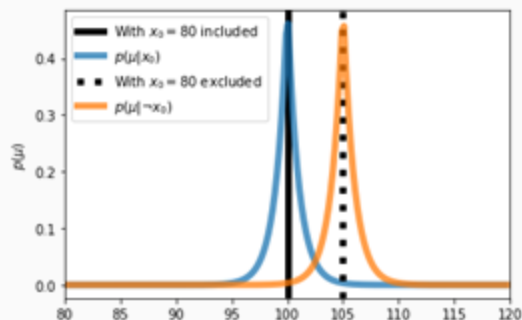
Why is adding Laplacian noise reasonable?

- Say we have data $X = [80, 90, 100, 110, 120]$
- We want to compute the mean
- What happens if $x_0=80$ is included or not?
- If we observe $A(D) = 105$:
 - Not much evidence that x_0 was (not) included!



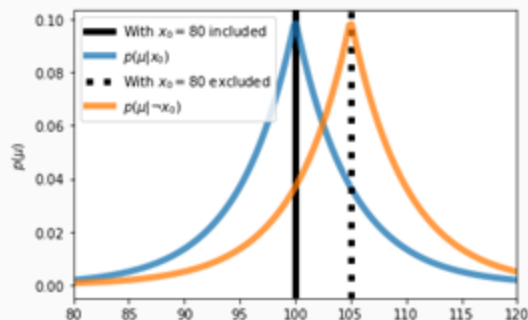
High noise \Rightarrow high privacy, but there are tradeoffs, as it lowers accuracy

Low noise ($b=1$)

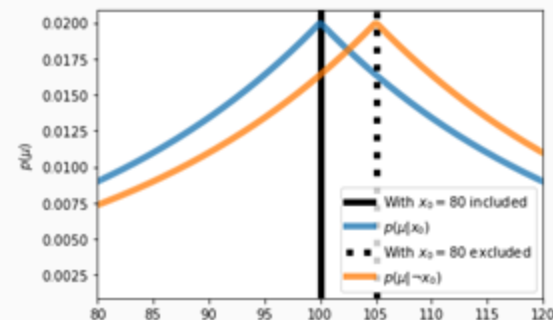


Large $\epsilon \Rightarrow$ High accuracy

Medium noise ($b=5$)



High noise ($b=25$)



Small $\epsilon \Rightarrow$ Low accuracy

$z \sim \text{Laplace}(\mu, b)$:

$$p(z) = \exp(-|z - \mu|/b) / 2b$$

Example of differential privacy in action: counts

- Say our data are binary vectors in \mathbf{R}^5 , e.g. $x_i = (0, 1, 1, 0, 0)$
- Let $f(D)$ be the sum over rows in D . What's the sensitivity?
 - Let $D' = D + \text{one row } x_N$
 - For each attribute i , we have
$$f(D')[i] = f(D)[i] + x_N[i] \quad \Rightarrow \quad |f(D')[i] - f(D)[i]| = |x_N[i]| \leq 1$$
 - So for 5 attributes, we have a **sensitivity of 5**
- To get ϵ -privacy, we need Laplace noise of scale $5/\epsilon$

Say D has 50 records, each x in D is a single number between -10 and 10, and $f(D) = \max(\text{abs}(x))$
Sensitivity: $f(D) = \max(|x|)$

- If $x \sim [-10, 10]$, then $|x| \in [0, 10]$
- What's the biggest gap that $f(D)$ and $f(D')$ could have?
 - Worst case: $D = [0, 0, 0, 0, \dots]$, and $D' = D + \{10\}$
 - $f(D) = 0, f(D') = 10$
- Sensitivity $\Delta f = |10 - 0| = 10$
- Take-away, extremal aggregators (min, max) are very sensitive!
 - Safer alternatives exist, eg 1%- and 99%-percentile

In contrast: Same setting as before, but now the aggregator is the mean

Sensitivity: $f(D) = \text{mean}(x)$, $N = 50$

- If $x \sim [-10, 10]$, then $\text{mean}(x) \in [-10, 10]$
- $f(D) = \sum x / (N-1)$
 $f(D') = [(N-1) \cdot f(D) + x_N] / N$
- $|f(D') - f(D)| = |[(N-1) \cdot f(D) + x_N] / N - f(D)|$
 $= |[(N-1) \cdot f(D) + x_N] - N \cdot f(D)| / N$
 $= |x_N - f(D)| / N$

Sensitivity, sample size, scale and privacy

- If $x \sim [-10, 10]$, then $\text{mean}(x) \in [-10, 10]$

- $f(D) = \sum x / (N-1)$

$$f(D') = [(N-1) \cdot f(D) + x_N] / N$$

- $|f(D') - f(D)| = |[(N-1) \cdot f(D) + x_N] / N - f(D)|$
 $= |[(N-1) \cdot f(D) + x_N] - N \cdot f(D)| / N$
 $= |x_N - f(D)| / N$
 $\leq |10 - -10| / N$
 $= 20 / N$

Sensitivity decreases as N increases!

Privacy is easier to achieve in larger datasets.



In small samples,
Waldo is easy to find

**WHERE'S
WALDO?**



What about multiple queries?

- Each time you ask for $A(\cdot)$, you'll get a different answer (it's randomized!)
- Differential privacy **composition** theorem:

If you make a sequence of queries A_i , each being ϵ_i -DP, then the result is only guaranteed to be $(\epsilon = \sum_i \epsilon_i)$ -differentially private

- The good news: any **deterministic** post-processing preserves privacy!

Extra resources

-  [Cynthia Dwork & Aaron Roth \(2014\)](#)
[monograph on Differential Privacy](#)
-  [Kamalika Chaudhuri & Anand Sarwate \(2017\)](#)
[NeurIPS tutorial on Differentially Private Machine Learning](#)
-  [John Abowd \(2019\)](#)
[ICML Keynote on the 2020 US Census](#)
-  [danah boyd & Jayshree Sarathy \(2022\)](#)
[Summary / updates to ongoing DP controversy w/ 2020 Census](#)

Differential privacy: Summary

- Being able to release data is critical for **reproducibility** but needs to be balanced with concerns about **privacy**.
- Removing **identifiers** and providing **k-anonymity** is not sufficient to guarantee privacy, particularly if the data is high-dimensional.
- In that case, **de-anonymization attacks** are highly accurate, even if based only on crude similarity search and relatively little data (e.g. the Netflix prize attack)
- **Differential privacy** keeps the raw data private, but lets the user interact with it via an **API**.
- Idea: Add carefully calibrated **Laplacian** noise to the data returned to give plausible deniability (as to whether the results came from a dataset containing a given row)
- **Sensitivity**: Maximal difference in output given a single row difference.
Extremal aggregators are highly sensitive, but higher N reduces sensitivity.
- **Multiple queries** reduce differential privacy - privacy loss aggregates.