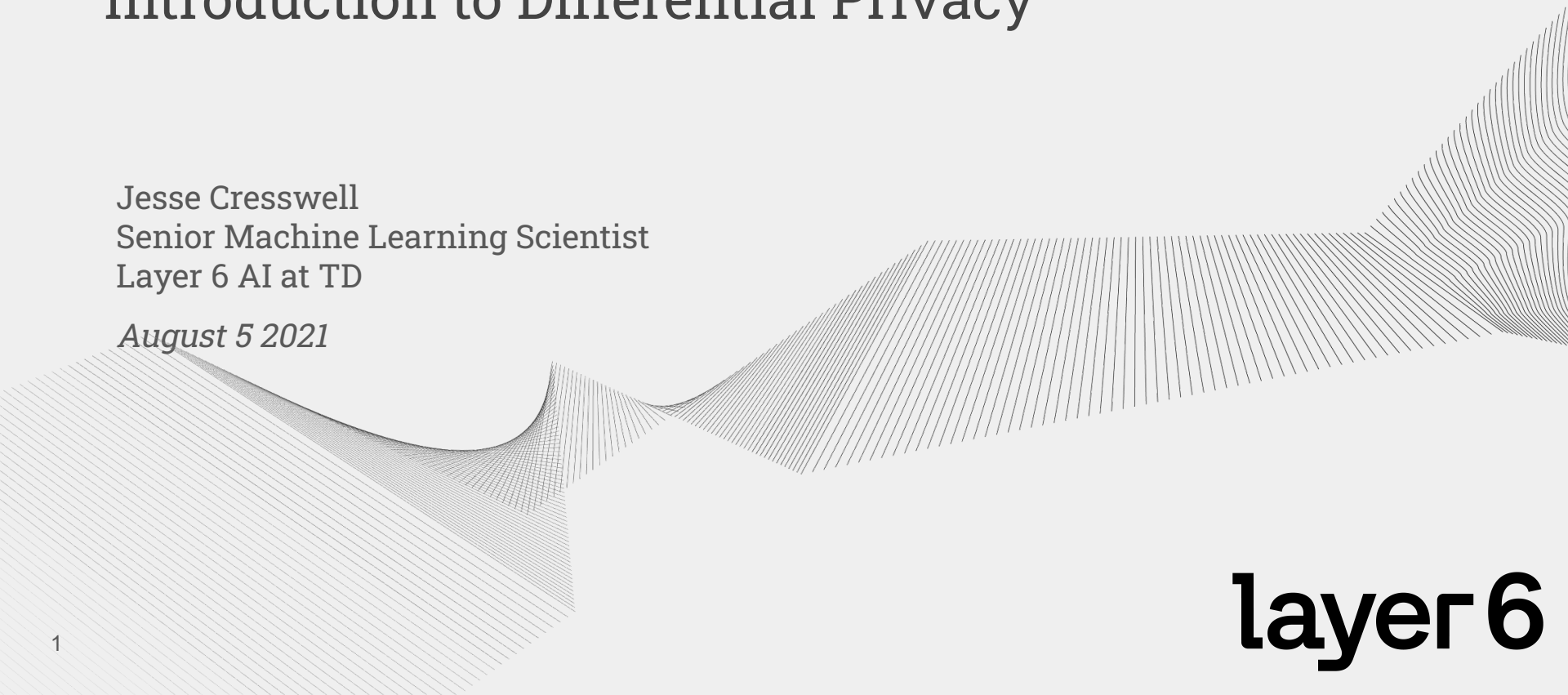
Protecting Privacy in ML: Introduction to Differential Privacy

Jesse Cresswell
Senior Machine Learning Scientist
Layer 6 AI at TD

August 5 2021



layer 6

Layer 6 - Who We Are

layer 6
AI at TD

Layer 6

ML in finance, healthcare

Active research groups:

CV, NLP, Deep Gen, RL, RecSys, Privacy

What do we mean by privacy?

layer 6

What do we mean by privacy?

Protecting people's privacy is a critical priority in healthcare and banking.

This term is often used, but rarely defined.

Canada has several laws about protecting "personal information" - data about an individual that can **identify the person**, on its own or combined with other data.

These laws cover collection, use, and disclosure of PI, but they may not capture the spirit of what privacy should mean in data analysis.

Is there a rigorous way to define privacy, relevant to how we use data in ML?

layer 6

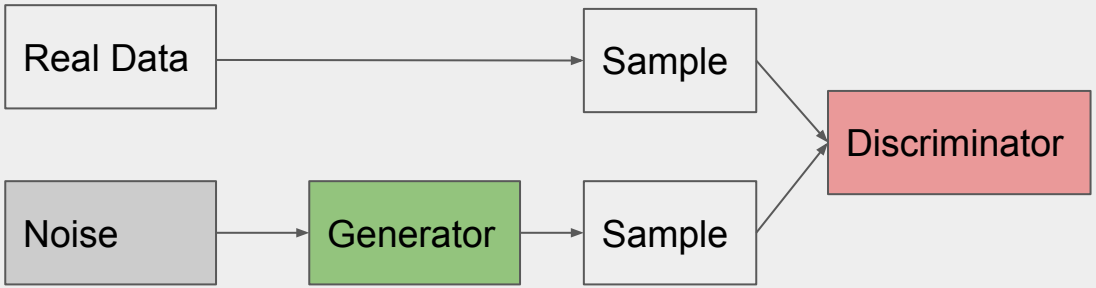
Is this private?

GitHub Copilot Autocomplete memorizes and repeats blocks of code, possibly from repos marked “private”. Is the code really private?

```c
// fast inverse square root

float Q_rsqrt(float number) {
    long i;
    float x2, y;
    const float threehalfs = 1.5F;
    x2 = number * 0.5F;
    y  = number;
```

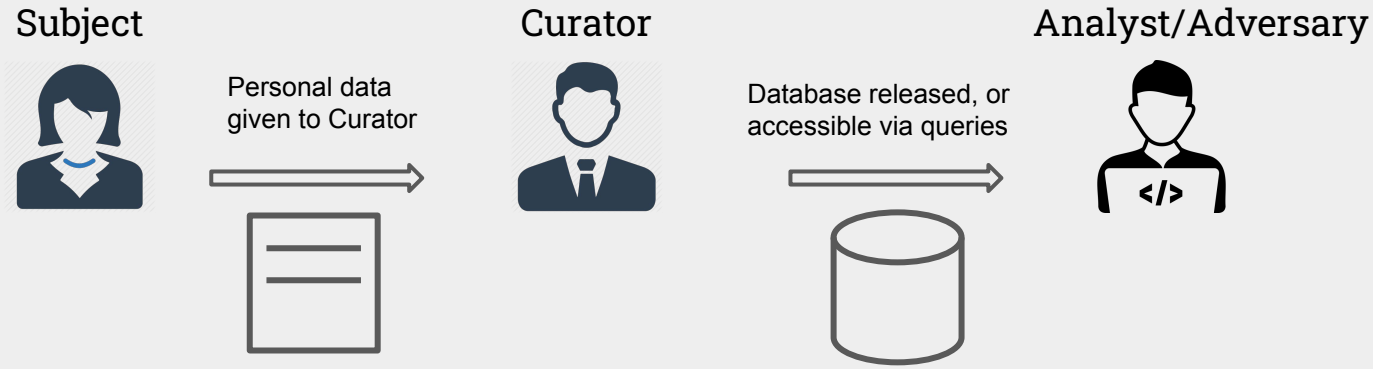
GANs learn to generate images that are similar to a training set. Has the privacy of training set samples been breached?



layer 6

Setting for defining privacy

Subjects - individuals who give, or refuse to give, their personal data
Curator - trusted centralized party that holds data and controls the database
Analysts/Adversaries - access the database for good or for evil



Privacy is a promise from the curator to the subjects that **no harm** will come to them by including their data in the database.

layer 6

The promise of privacy

Curator's promise - "*You will not be affected, adversely or otherwise, by allowing your data to be used in the database, no matter what other data sets are available*."

- How can we learn something about a population, while not learning anything about individuals?

Imagine an individual choosing to provide their data to a study on smoking.

Whether or not the individual gives their data, the study is likely to conclude that smoking causes health problems. Actions will be taken - insurers raise premiums on smokers, or our individual chooses to stop smoking.

However, it was **not the inclusion of the individual's data** that led to these outcomes, but the study as a whole. The promise was not broken.

layer 6

Attempts at privacy - Anonymization

Some approaches are commonly used, but don't preserve the promise.

The curator may remove PI, and publish the rest.

Banks and credit bureaus use anonymization to share customer information.
Hospitals release data on medical images and diagnoses without PI.

Linkage Attacks

It can be possible to de-anonymize records using outside information.

Netflix released data on sparse user movie ratings - de-anonymized by comparing to public ratings on IMDB. [Narayanan & Shmatikov 2008]

layer 6

Attempts at privacy - Restrict Queries

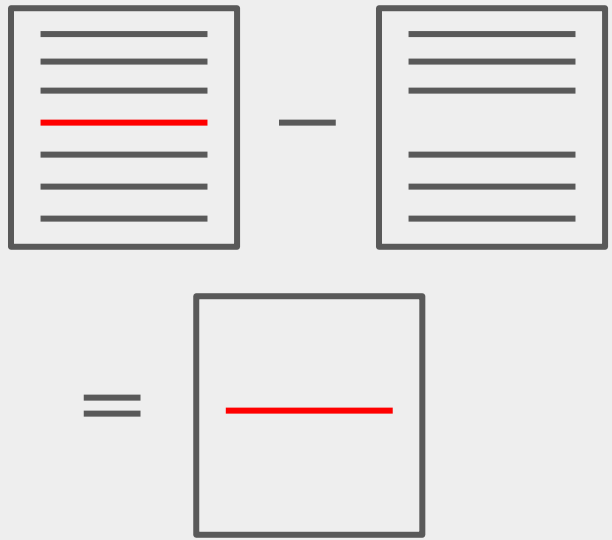
The curator does not release the database, but will answer certain queries.

To avoid harm to individuals, queries will only be answered if they return a large enough group.

Differencing Attacks

If an individual is known to be in the database then 2 queries over large sets will reveal info about the individual:

How many people in the database have disease Y?
How many people not named Person X have disease Y?



layer 6

Attempts at privacy - Federated Learning

Federated Learning is a distributed ML approach where data is not stored on a centralized server.

Instead, the model is trained on the device where data is collected (e.g. smart phone), and updates are aggregated.

Intuitively this seems “more private”.

Memorization

Except, there are many examples of NNs memorizing individual training examples.

XKCD 2169

layer 6

Fundamental Law of Information Recovery

Earliest definitions of privacy insisted that “nothing should be learned about an individual when data is released”.

This is almost comical, as if we cannot learn *anything* about *anyone*, then there is no information to be gained at all!

Fundamental Tradeoff: Learning anything from data analysis necessitates learning something about the underlying data.

Privacy may be thought of as a **resource**. The more information we gain from a data source, the more **privacy is used up**.

Differential privacy gives a formal definition of this resource.

layer 6

Differential Privacy

layer 6

Randomized Response

Survey on sensitive subject - Have you committed tax fraud?
- Nobody would truthfully respond to this survey and incriminate themselves.

Ask subjects to flip a coin so that they have plausible deniability

Answer
Truthfully

Flip again
If Heads: “Yes”
If Tails: “No”

Analyst gets an estimate of the true answer by understanding random process.

Randomness in **responses to queries** is the key to solving issues discussed above.

layer 6

Randomized Mechanisms

Adding randomness is a good way to achieve privacy.

Differential Privacy works with **randomized mechanisms** that query and return answers from a **database**.

Formally, a randomized mechanism $\mathcal{M} : A \to B$ will have some **probability of returning each output** for a given input.

On input $a \in A$ the mechanism returns $\mathcal{M}(a) = b$ with probability $(\mathcal{M}(a))_b$.

layer 6

Databases

We think of a database $\mathcal{D}$ as a collection of rows. Each row contains various points of information about one Subject.

Changing any bit of information in a row completely changes the identity of the row.

The **distance between databases** is essentially the number of rows that have been changed - think of Hamming distance.

We will write the distance using the L1 norm $\|x - y\|_1 = \sum_{i=1}^{|\mathcal{X}|} |x_i - y_i|$

Every possible row is an element of $\mathcal{X}$, so by comparing counts of all possible rows we see how much two databases differ.

layer 6

Differential Privacy [Dwork et al. 2006]

Consider 2 databases $\mathcal{D}$ and $\mathcal{D}'$ which differ by one record, $\|\mathcal{D} - \mathcal{D}'\|_1 = 1$,

and a randomized mechanism $\mathcal{M}[\mathcal{D}]$ which acts on databases to give a result.
The **mechanism is differentially private** if the results are almost indistinguishable for all such datasets.

A mechanism is (ϵ,δ)-differentially private if for all subsets of the output $\mathcal{S} \subseteq \text{Range}[\mathcal{M}]$ and all pairs of databases that differ by one row

$$\Pr(\mathcal{M}[\mathcal{D}] \in \mathcal{S}) \leq \exp(\epsilon)\Pr(\mathcal{M}[\mathcal{D}'] \in \mathcal{S}) + \delta$$

layer 6

Differential Privacy

A mechanism is (ϵ,δ)-differentially private if for all subsets of the output $\mathcal{S} \subseteq \text{Range}[\mathcal{M}]$ and all pairs of databases that differ by one row

$$\Pr(\mathcal{M}[\mathcal{D}] \in \mathcal{S}) \leq \exp(\epsilon)\Pr(\mathcal{M}[\mathcal{D}'] \in \mathcal{S}) + \delta$$

Counterexample:

Mechanism that returns the exact count of rows that satisfy property P.

On neighbouring databases

$$\Pr(\mathcal{M}[\mathcal{D}] = n) = 1 \quad \text{and} \quad \Pr(\mathcal{M}[\mathcal{D}'] = n) = 0$$

so this exact count mechanism cannot be DP unless ϵ=∞ or δ=1.

layer 6

Differential Privacy

A mechanism is (***ϵ***,***δ***)-differentially private if for all subsets of the output $\mathcal{S} \subseteq \text{Range}[\mathcal{M}]$ and all pairs of databases that differ by one row

$$\Pr(\mathcal{M}[\mathcal{D}] \in \mathcal{S}) \leq \exp(\epsilon)\Pr(\mathcal{M}[\mathcal{D}'] \in \mathcal{S}) + \delta$$

- Randomized - Any DP-mechanism must be randomized
- Quantitative - ***ϵ*** and ***δ*** are numerical, where lower means better privacy
- Worst Case - We assume nothing about what the input datasets are like

Differential Privacy

A mechanism is (***ϵ,δ***)-differentially private if for all subsets of the output $\mathcal{S} \subseteq \text{Range}[\mathcal{M}]$ and all pairs of databases that differ by one row

$$\Pr(\mathcal{M}[\mathcal{D}] \in \mathcal{S}) \leq \exp(\epsilon)\Pr(\mathcal{M}[\mathcal{D}'] \in \mathcal{S}) + \delta$$

Curator's promise - "*You will not be affected, adversely or otherwise, by allowing your data to be used in the database, no matter what other data sets are available.*"

By adding one Subject's data, differential privacy makes the promise that the likelihood of any result will change only by a small factor (expect with probability ***δ***).

layer 6

Guarantees

After accessing the database with an (ϵ,δ)-DP mechanism, can an attacker bring in outside info to weaken the privacy guarantee?

No. DP-mechanisms are **immune to post-processing**.

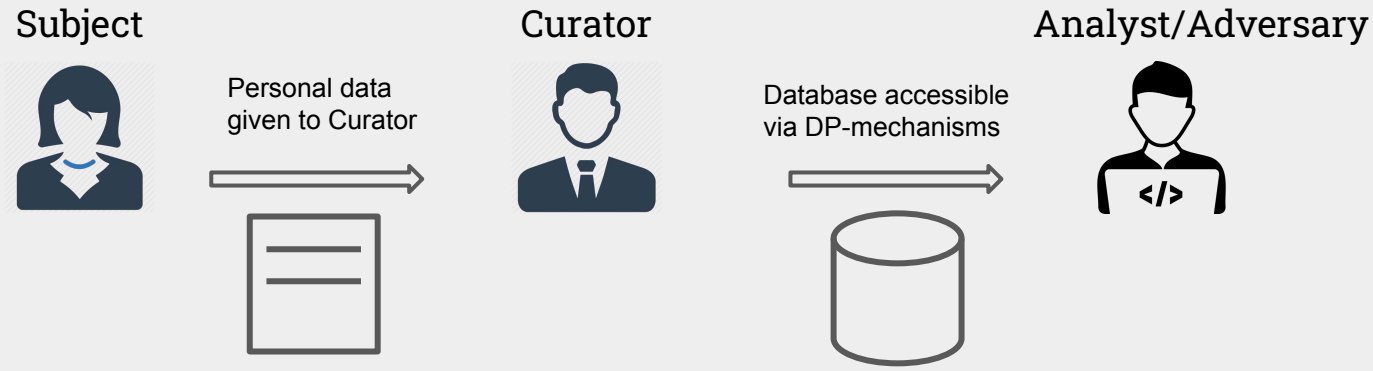
Formally, composing a (ϵ,δ)-DP mechanism with any other function (that does not involve the database) will give another (ϵ,δ)-DP mechanism.

Privacy is a resource - it gets used up as the database is accessed more.

(ϵ,δ)-DP mechanisms quantify how much privacy is consumed via composition.

Applying the same mechanism k times can be considered as a single mechanism which is (kϵ, kδ)-DP

Differential Privacy in Practice



The private database is accessible only through DP-mechanisms. Each access consumes (ϵ,δ) worth of the privacy budget.

We need to design *useful* mechanisms that are differentially private.

Deterministic queries $f[\mathcal{D}]$ on a database like counts, sums, and averages will not maintain privacy. We can modify any query by **adding random noise to the result**.

layer 6

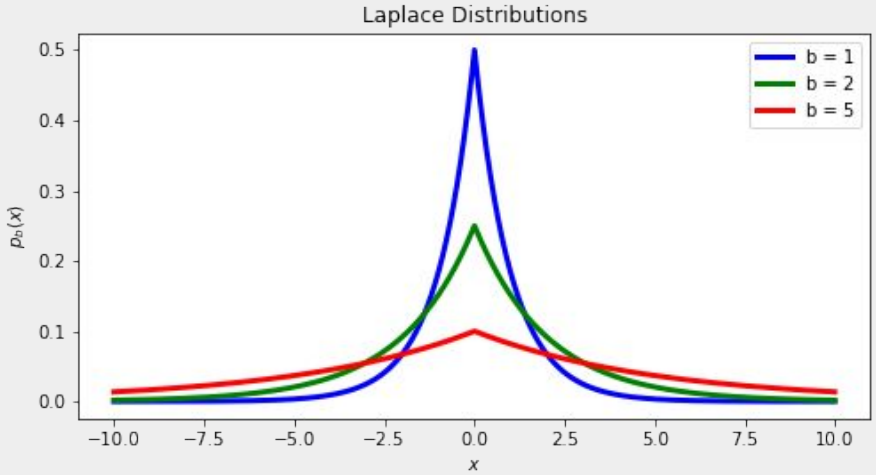
Our First DP-Mechanism

The **Laplace mechanism** adds noise from a Laplace distr. to a deterministic query

$$\mathcal{M}[\mathcal{D}] = f[\mathcal{D}] + \xi$$

Laplacian noise has “width” b.

This mechanism is (***ϵ,δ***)-DP with $\epsilon = \Delta f / b$ and ***δ***=0



Δf is the sensitivity of a function, the maximum difference between its result on two neighbouring databases

$$\Delta f = \max_{\mathcal{D},\mathcal{D}'} \|f[\mathcal{D}] - f[\mathcal{D}']\|_1$$

layer 6

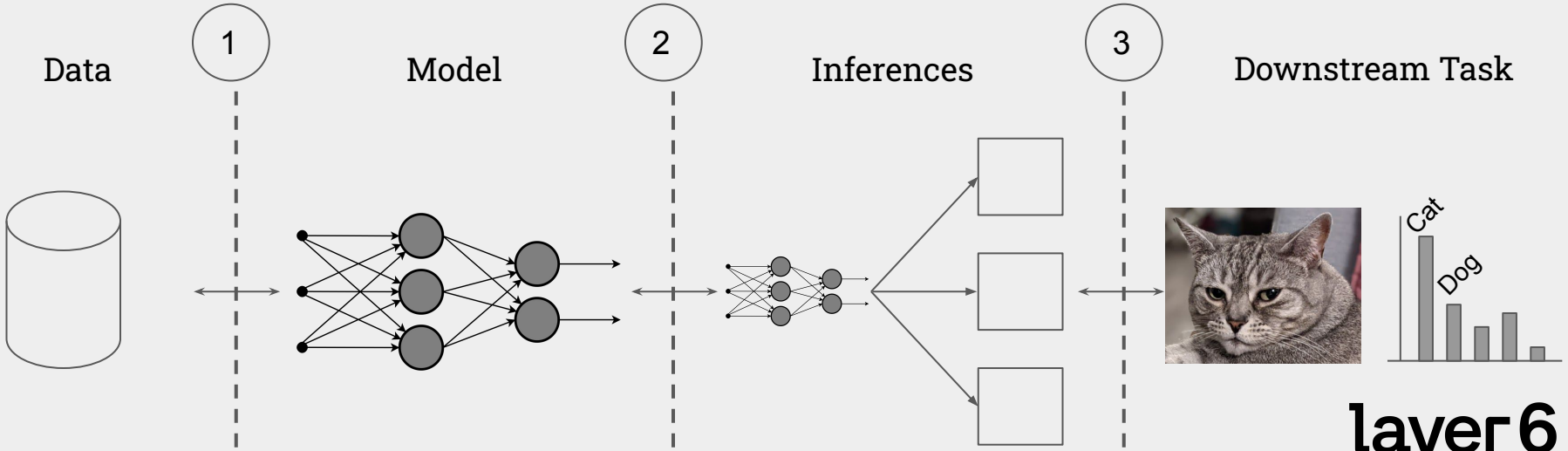
Differential Privacy in ML

layer 6

How to train your model

Machine learning poses several challenges from a differential privacy perspective.

- Memorization of individual examples
- Many accesses to the database are needed (training steps)
- Inference needs to be run many times

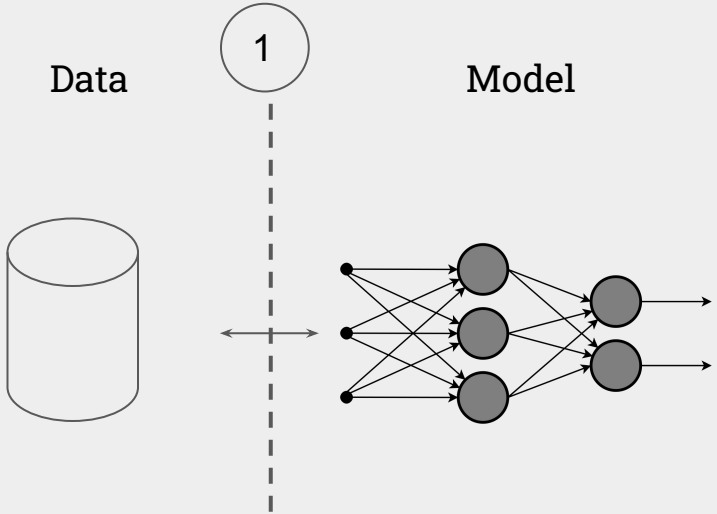


layer 6

How to train your model

DP-SGD [Abadi et al. 2016] - Control privacy via gradient updates during training

- Gradients have unbounded sensitivity - **clip** to provide hard bound
- **Add noise** as in the Laplace Mechanism
- Aggregate gradients and take an SGD step



Gradient clipping and noising are standard regularization technique.

Once model is trained, unlimited inference can be done due to the post-processing guarantee.

layer 6

Advanced Topics

Strong Composition Theorems

- Theoretical analysis can provide sublinear composition guarantees

---

Privacy Accountants

- Further strengthens composition for large scale ML training

---

Relaxations of Differential Privacy

- Weaken definition of DP while maintaining many desirable properties

---

Mechanism Design

- Application specific analysis of privacy guarantees

Conclusions

layer 6

Conclusions

Privacy cannot be perfect for every individual, while providing useful information.

Randomized results are essential for protecting privacy.

Privacy is a resource that is used up with every query.

Differential privacy aims to quantify how much privacy a query uses.

Our job is to develop mechanisms that minimize privacy use, while maximizing the utility/accuracy of results.

layer 6

Resources

If you like learning from a textbook:

Dwork & Roth - Algorithmic foundations of Differential Privacy
The Algorithmic Foundations of Differential Privacy

If you want a quick blog post overview (see also part II)

Borealis AI - Tutorial #12: Differential Privacy I: Introduction

If you want a university course with video lectures

Waterloo CS860 - Prof. Gautam Kamath - A Course In Differential Privacy

Much of these slides is built from these sources

layer 6

We are open to collaboration!

jesse@layer6.ai