Machine Learning with Adversaries

Joseph E. Gonzalez

Co-director of the RISE Lab

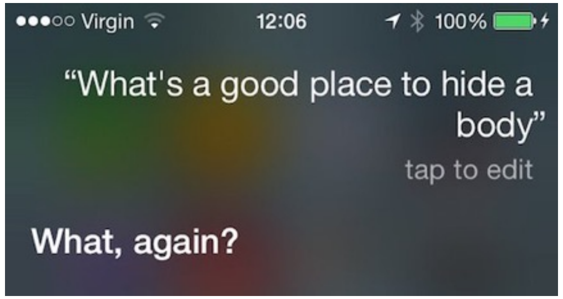
jegonzal@cs.berkeley.edu

Intelligence in **Sensitive Contexts**

AR/VR Systems

Home Monitoring

Voice Technologies

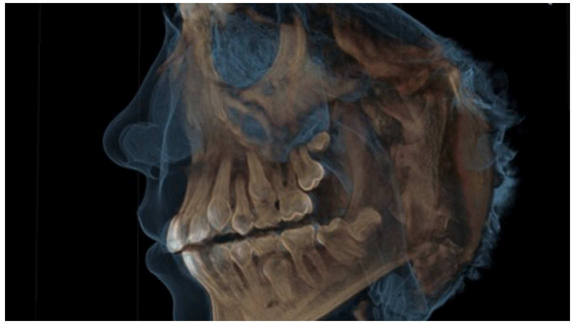


"What's a good place to hide a body"

tap to edit

What, again?

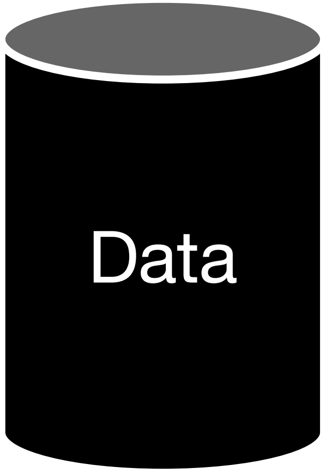
Medical Imaging



Protect the **data**, the **model**, and the **query**

Protect the **data**, the **model**, and the **query**

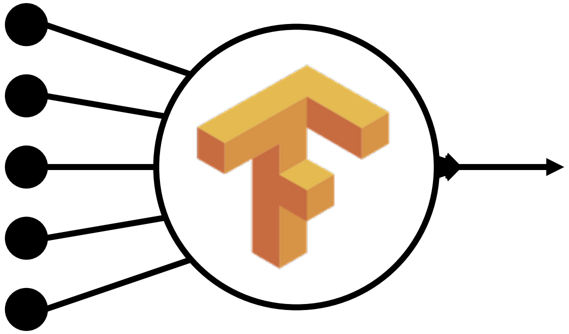
High-Value **Data** *is* ***Sensitive***



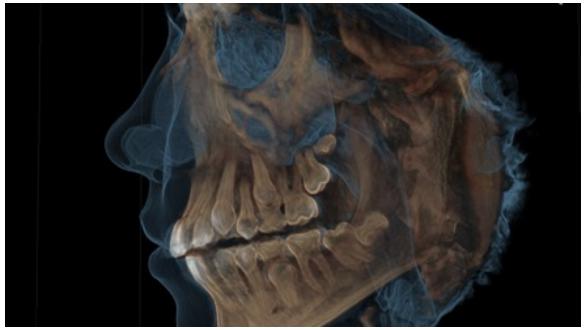
- Medical Info.
- Home video
- Finance

Models capture **value** in data

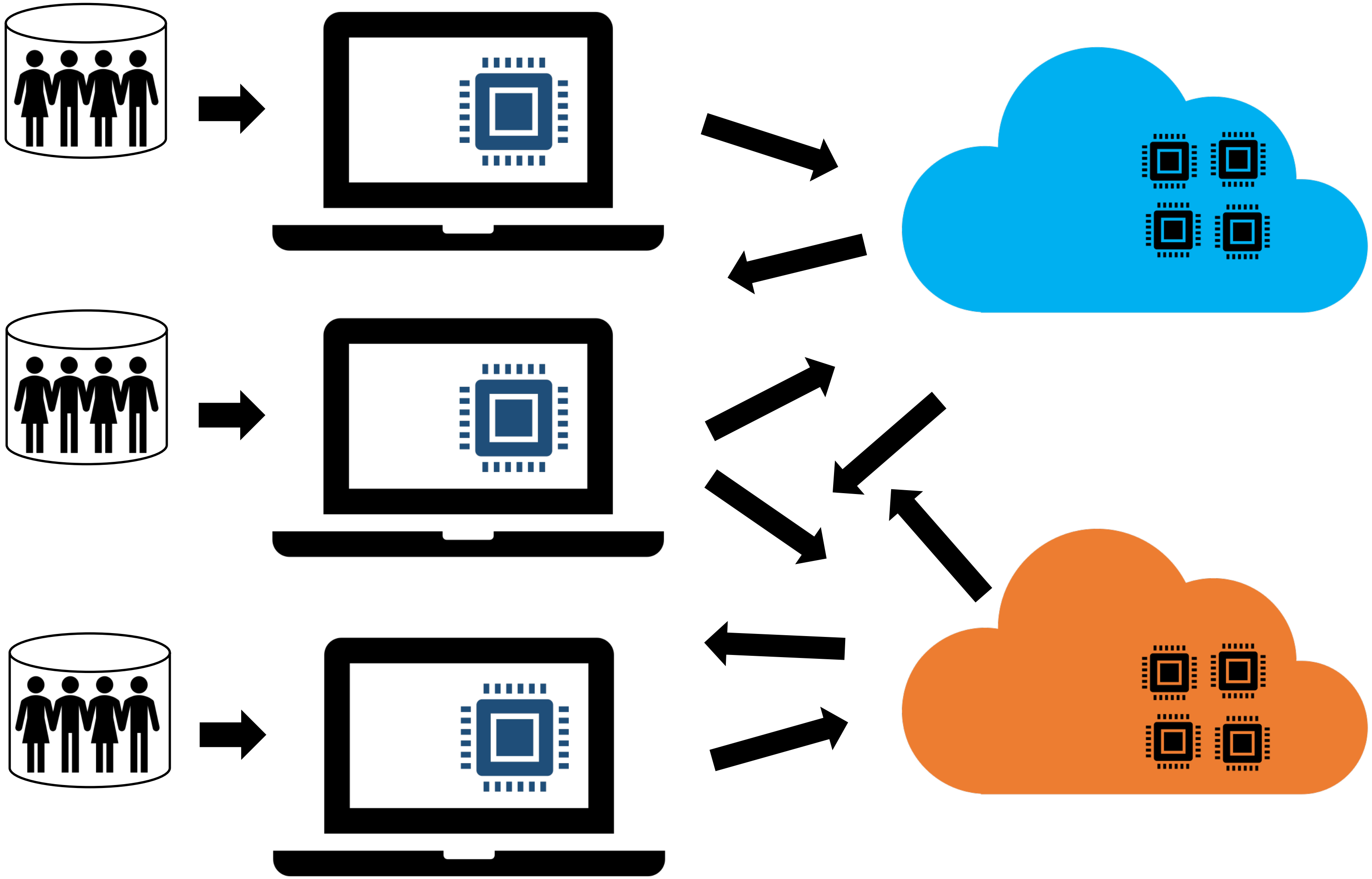
- Core Asset
- Sensitive



Queries can be as sensitive as the data



Who/What do you trust?



Attack Scenarios

Training: Data Poisoning Attack

- **Scenario:** *Spammers rating emails to affect classifier.*
  - *Real-world problem*
- *Kinds of Attacks*
  - Train model to misclassify their spam as ham.
  - Train model to misclassify everything (DoS attack). why?
  - Train model to classify competitor email as spam.
- *Responses*
  - Careful validation of training data
    - Reject on negative impact (RONI) defense
  - Personalization

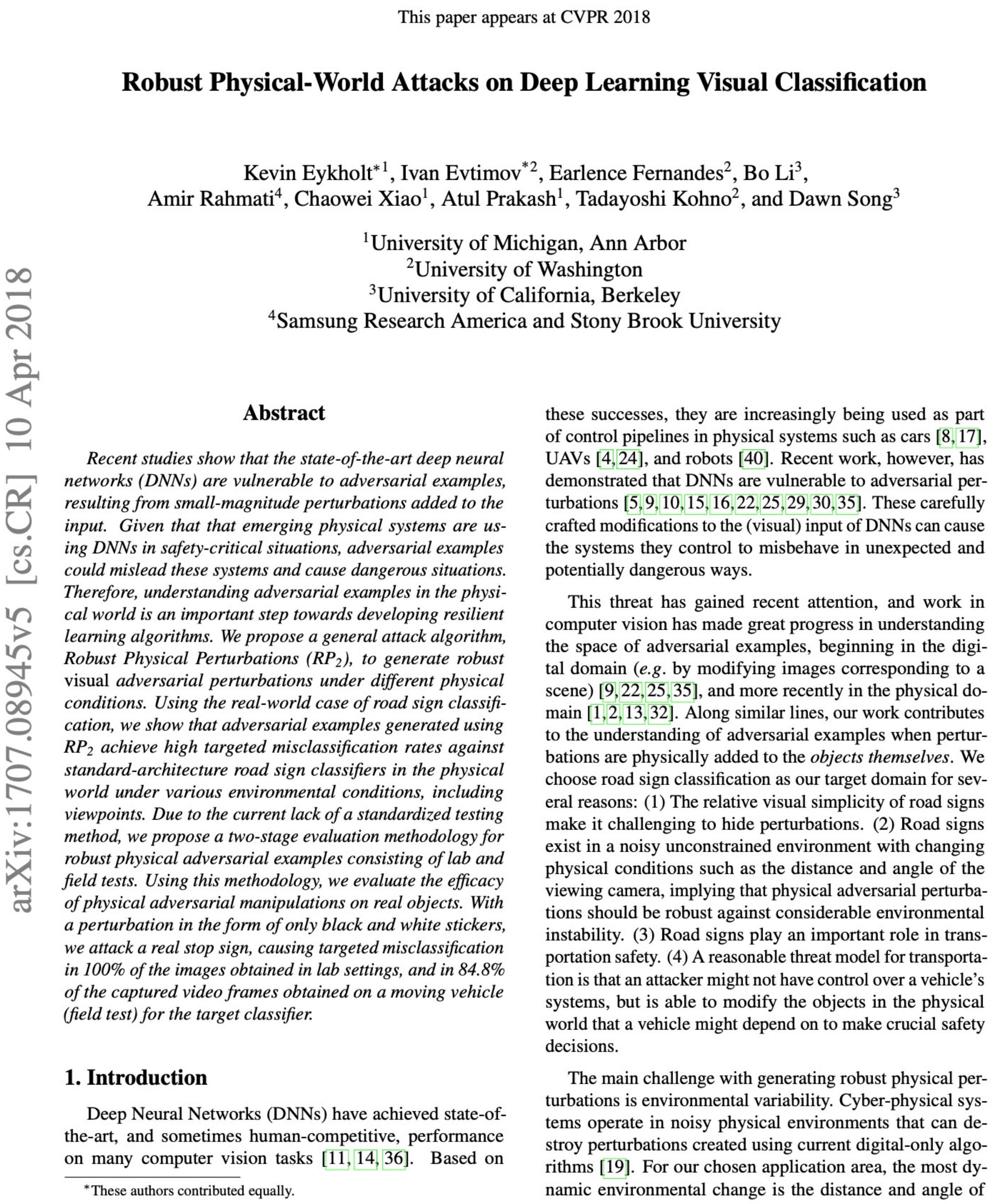
Training: Extracting Model, Alg., Data

- **Kinds of attacks:**
  - *Crafting labeled inputs to reveal important features*
  - *Modifying data/gradient to probe global loss*
  - *Tracking changes in model to learn about users*

- *Possible Solutions (Federated setting)*
  - Secure multi-party computation
  - Trusted enclaves

Inference: Evasion Attacks & Adversarial Inputs

- **Scenarios**
  - *spammers modify content to bypass filters*
  - *adversary modifies signs to confuse autonomous vehicles*
- *Kinds of Attacks*
  - *Leveraging adversarial noise to minimally perturb inputs while substantially changing predictions.*
  - *Whitebox attacks seem to transfer to blackbox settings*
- *Possible Solutions*
  - *Robust model design*
  - *Noise elimination*
  - *Arms race …*



This paper appears at CVPR 2018

Robust Physical-World Attacks on Deep Learning Visual Classification

Kevin Eykholt*[1], Ivan Evtimov*[2], Earlence Fernandes[2], Bo Li[3],
Amir Rahmati[4], Chaowei Xiao[1], Atul Prakash[1], Tadayoshi Kohno[2], and Dawn Song[3]

[1]University of Michigan, Ann Arbor
[2]University of Washington
[3]University of California, Berkeley
[4]Samsung Research America and Stony Brook University

arXiv:1707.08945v5 [cs.CR] 10 Apr 2018

Abstract

Recent studies show that the state-of-the-art deep neural networks (DNNs) are vulnerable to adversarial examples, resulting from small-magnitude perturbations added to the input. Given that that emerging physical systems are using DNNs in safety-critical situations, adversarial examples could mislead these systems and cause dangerous situations. Therefore, understanding adversarial examples in the physical world is an important step towards developing resilient learning algorithms. We propose a general attack algorithm, Robust Physical Perturbations (RP$_2$), to generate robust visual *adversarial perturbations under different physical conditions. Using the real-world case of road sign classification, we show that adversarial examples generated using RP$_2$ achieve high targeted misclassification rates against standard-architecture road sign classifiers in the physical world under various environmental conditions, including viewpoints. Due to the current lack of a standardized testing method, we propose a two-stage evaluation methodology for robust physical adversarial examples consisting of lab and field tests. Using this methodology, we evaluate the efficacy of physical adversarial manipulations on real objects. With a perturbation in the form of only black and white stickers, we attack a real stop sign, causing targeted misclassification in 100% of the images obtained in lab settings, and in 84.8% of the captured video frames obtained on a moving vehicle (field test) for the target classifier.*

1. Introduction

Deep Neural Networks (DNNs) have achieved state-of-the-art, and sometimes human-competitive, performance on many computer vision tasks [11, 14, 36]. Based on these successes, they are increasingly being used as part of control pipelines in physical systems such as cars [8, 17], UAVs [4, 24], and robots [40]. Recent work, however, has demonstrated that DNNs are vulnerable to adversarial perturbations [5, 9, 10, 15, 16, 22, 25, 29, 30, 35]. These carefully crafted modifications to the (visual) input of DNNs can cause the systems they control to misbehave in unexpected and potentially dangerous ways.

This threat has gained recent attention, and work in computer vision has made great progress in understanding the space of adversarial examples, beginning in the digital domain (*e.g.* by modifying images corresponding to a scene) [9, 22, 25, 35], and more recently in the physical domain [1, 2, 13, 32]. Along similar lines, our work contributes to the understanding of adversarial examples when perturbations are physically added to the *objects themselves*. We choose road sign classification as our target domain for several reasons: (1) The relative visual simplicity of road signs make it challenging to hide perturbations. (2) Road signs exist in a noisy unconstrained environment with changing physical conditions such as the distance and angle of the viewing camera, implying that physical adversarial perturbations should be robust against considerable environmental instability. (3) Road signs play an important role in transportation safety. (4) A reasonable threat model for transportation is that an attacker might not have control over a vehicle's systems, but is able to modify the objects in the physical world that a vehicle might depend on to make crucial safety decisions.

The main challenge with generating robust physical perturbations is environmental variability. Cyber-physical systems operate in noisy physical environments that can destroy perturbations created using current digital-only algorithms [19]. For our chosen application area, the most dynamic environmental change is the distance and angle of

*These authors contributed equally.

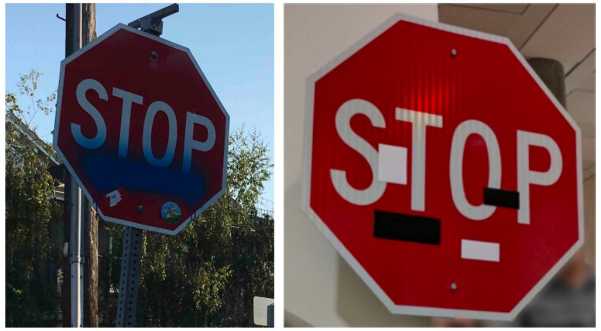


Figure 1: The left image shows real graffiti on a Stop sign, something that most humans would not think is suspicious. The right image shows our a physical perturbation applied to a Stop sign. We design our perturbations to mimic graffiti, and thus "hide in the human psyche."

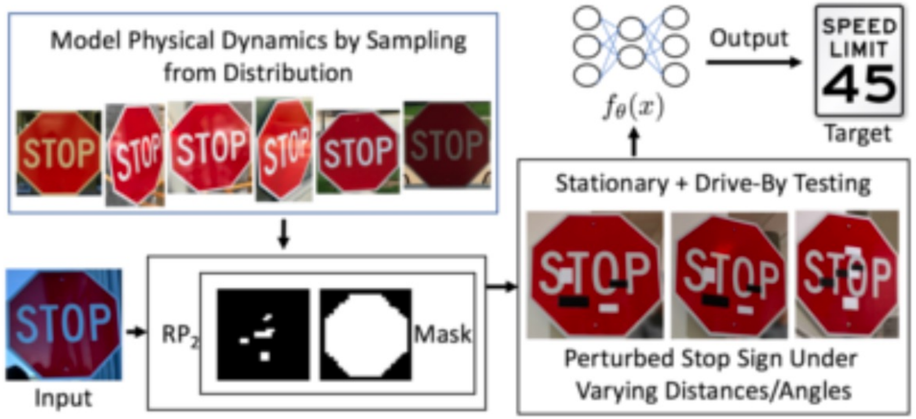

Model Physical Dynamics by Sampling from Distribution
$f_\theta(x)$
Output
SPEED LIMIT 45
Target
Stationary + Drive-By Testing
STOP
Input
RP$_2$
Mask
Perturbed Stop Sign Under Varying Distances/Angles

Figure 2: RP$_2$ pipeline overview. The input is the target Stop sign. RP$_2$ samples from a distribution that models physical dynamics (in this case, varying distances and angles), and uses a mask to project computed perturbations to a shape that resembles graffiti. The adversary prints out the resulting perturbations and sticks them to the target Stop sign.

Inference: Model + Data Inversion

- **Scenario:** *Query public prediction APIs to extract information from the training data or steal the model.*
- *Kinds of Attacks*
  - ***Active learning attack*** *to construct training data points that resolve decision boundary and allow model extraction.*
  - ***Membership inference*** *attacks determine if a piece of data was used in training*
  - ***Model inversion*** *attacks can construct training inputs given labels*
- *Possible Solutions*
  - Limit confidence information and query rates
  - Introduce noise in predictions
  - Model watermarks

Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures

Matt Fredrikson
Carnegie Mellon University

Somesh Jha
University of Wisconsin–Madison

Thomas Ristenpart
Cornell Tech

ABSTRACT

Machine-learning (ML) algorithms are increasingly utilized in privacy-sensitive applications such as predicting lifestyle choices, making medical diagnoses, and facial recognition. In a model inversion attack, recently introduced in a case study of linear classifiers in personalized medicine by Fredrikson et al. [13], adversarial access to an ML model is abused to learn sensitive genomic information about individuals. Whether model inversion attacks apply to settings outside theirs, however, is unknown.

We develop a new class of model inversion attack that exploits confidence values revealed along with predictions. Our new attacks are applicable in a variety of settings, and we explore two in depth: decision trees for lifestyle surveys as used on machine-learning-as-a-service systems and neural networks for facial recognition. In both cases confidence values are revealed to those with the ability to make prediction queries to models. We experimentally show attacks that are able to estimate whether a respondent in a lifestyle survey admitted to cheating on their significant other and, in the other context, show how to recover recognizable images of people's faces given only their name and access to the ML model. We also initiate experimental exploration of natural countermeasures, investigating a privacy-aware decision tree training algorithm that is a simple variant of CART learning, as well as revealing only rounded confidence values. The lesson that emerges is that one can avoid these kinds of MI attacks with negligible degradation to utility.

1. INTRODUCTION

Computing systems increasingly incorporate machine learning (ML) algorithms in order to provide predictions of lifestyle choices [6], medical diagnoses [20], facial recognition [1], over easy-to-use public HTTP interfaces. The features used by these models, and queried via APIs to make predictions, often represent sensitive information. In facial recognition, the features are the individual pixels of a picture of a person's face. In lifestyle surveys, features may contain sensitive information, such as the sexual habits of respondents.

In the context of these services, a clear threat is that providers might be poor stewards of sensitive data, allowing training data or query logs to fall prey to insider attacks or exposure via system compromises. A number of works have focused on attacks that result from access to (even anonymized) data [18,29,32,38]. A perhaps more subtle concern is that the ability to make prediction queries might enable adversarial *clients* to back out sensitive data. Recent work by Fredrikson et al. [13] in the context of genomic privacy shows a *model inversion attack* that is able to use black-box access to prediction models in order to estimate aspects of someone's genotype. Their attack works for any setting in which the sensitive feature being inferred is drawn from a small set. They only evaluated it in a single setting, and it remains unclear if inversion attacks pose a broader risk.

In this paper we investigate commercial ML-as-a-service APIs. We start by showing that the Fredrikson et al. attack, even when it is computationally tractable to mount, is not particularly effective in our new settings. We therefore introduce new attacks that infer sensitive features used as inputs to decision tree models, as well as attacks that recover images from API access to facial recognition services. The key enabling insight across both situations is that we can build attack algorithms that exploit confidence values exposed by the APIs. One example from our facial recognition attacks is depicted in Figure 1: an attacker can produce a recognizable image of a person, given only API access to a

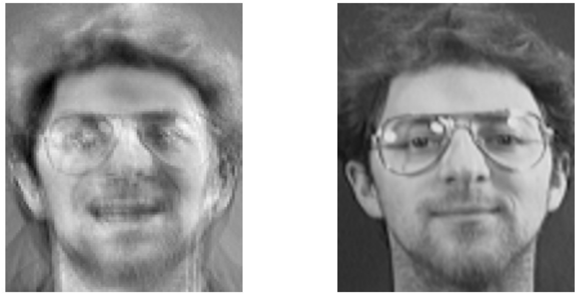


Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

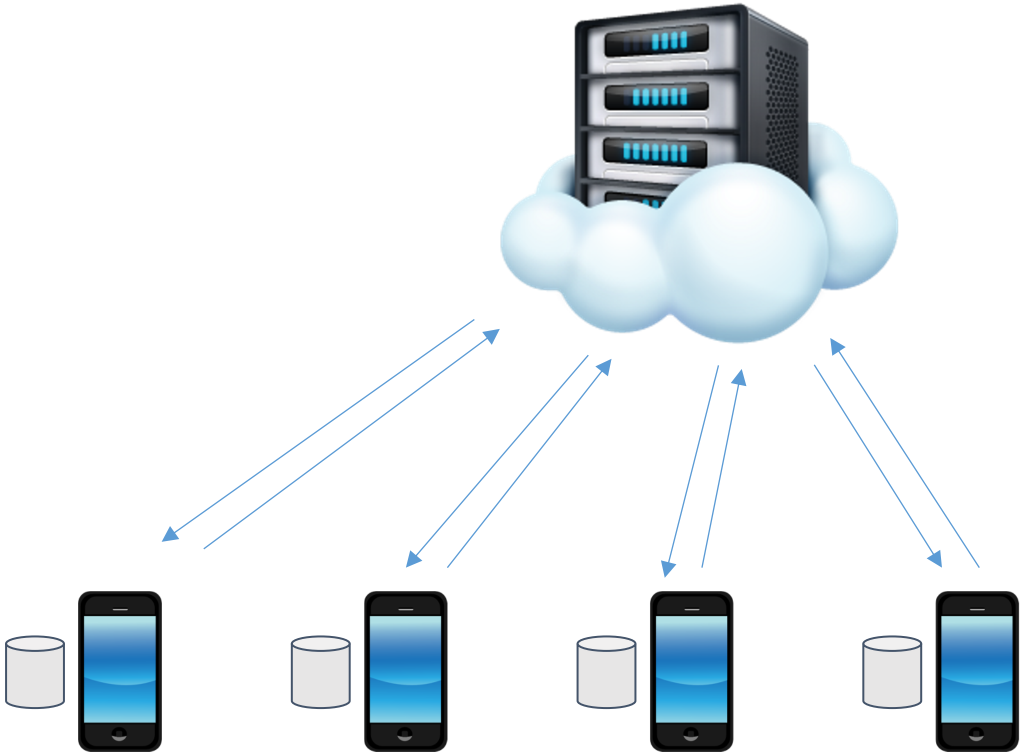
Big Concepts

Big Concepts in Secure ML

- **Adversarial Inputs:** *Inputs that are constructed to "trick" classifier into misclassifying (usually for evasion)*
- **Federated Learning:** A form of distributed learning in with limited comm. and non-iid data distribution
- **Differential Privacy:** A formal guarantee bounding the information leaked by a randomized algorithm
- **Security Technologies:**
    - **Secure Multi-party Computation:** Cryptographic protocols for shared computation without disclosing inputs.
    - **Trusted Execution Environments:** Hardware framework for secure computation.

Federated Learning

- Terminology was introduced in "Communication-Efficient Learning of Deep Networks from Decentralized Data" (AIStats'17)



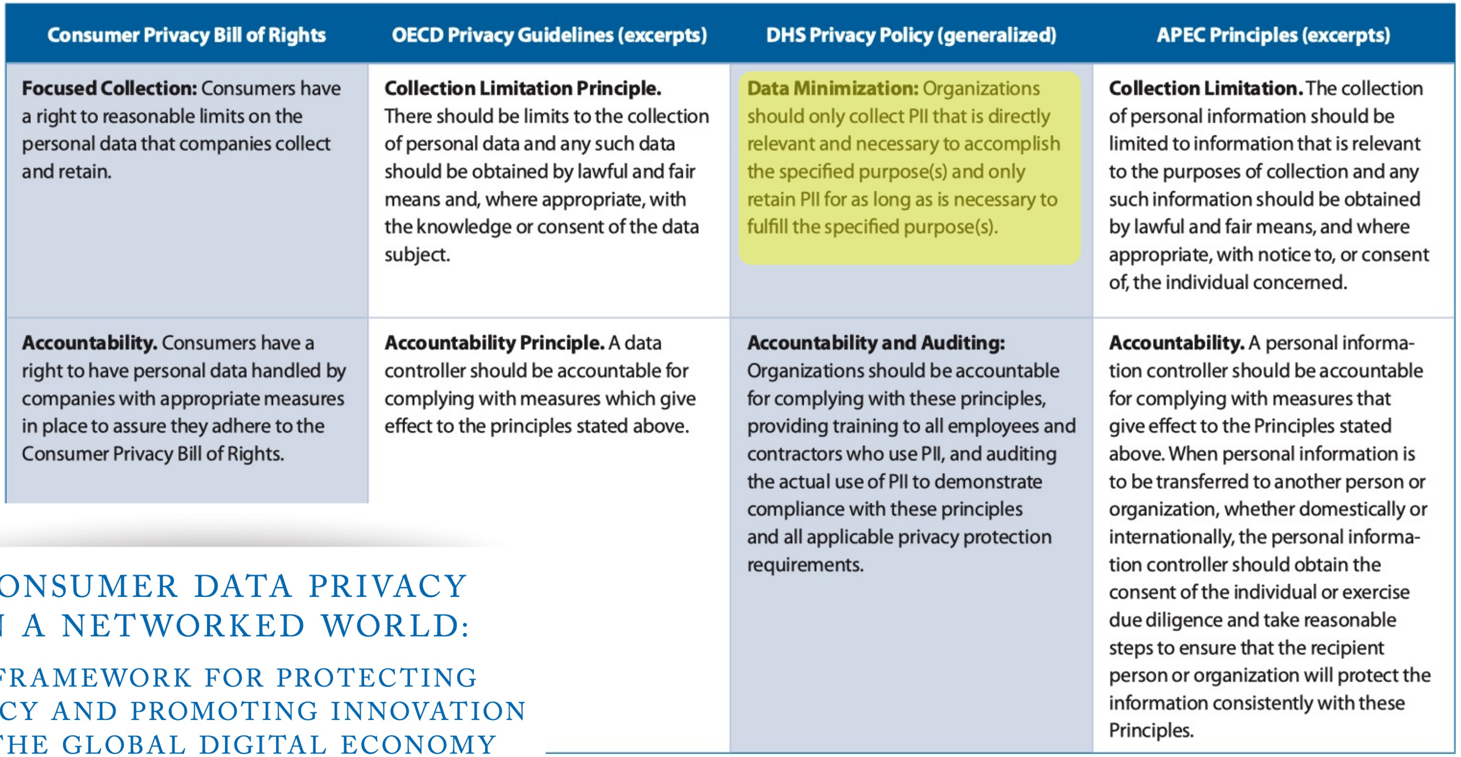
A form of distributed learning with:

- Original data does not leave device
- Limited communication (e.g., WAN)
- Data is not assumed to be iid
- More parties with far less data each
- Infrequent/random participation

Security?

Federated Learning and Security

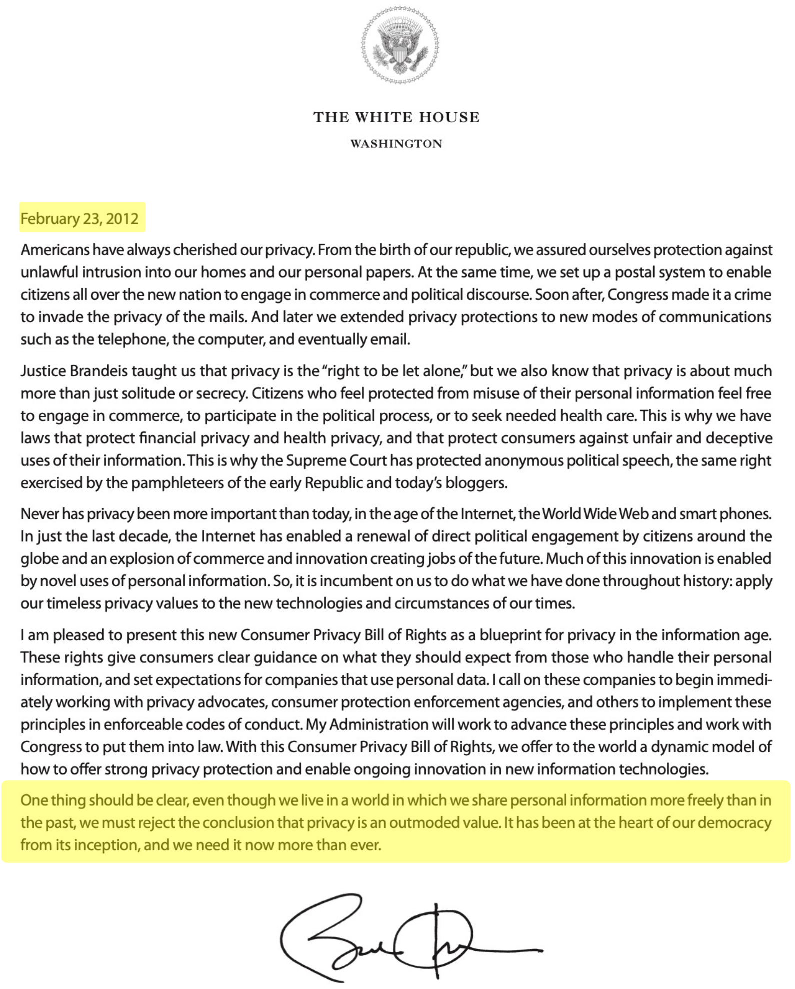
- ➢ Eliminates data collection in the cloud



Consumer Privacy Bill of Rights	OECD Privacy Guidelines (excerpts)	DHS Privacy Policy (generalized)	APEC Principles (excerpts)
Focused Collection: Consumers have a right to reasonable limits on the personal data that companies collect and retain.	**Collection Limitation Principle.** There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject.	**Data Minimization:** Organizations should only collect PII that is directly relevant and necessary to accomplish the specified purpose(s) and only retain PII for as long as is necessary to fulfill the specified purpose(s).	**Collection Limitation.** The collection of personal information should be limited to information that is relevant to the purposes of collection and any such information should be obtained by lawful and fair means, and where appropriate, with notice to, or consent of, the individual concerned.
Accountability. Consumers have a right to have personal data handled by companies with appropriate measures in place to assure they adhere to the Consumer Privacy Bill of Rights.	**Accountability Principle.** A data controller should be accountable for complying with measures which give effect to the principles stated above.	**Accountability and Auditing:** Organizations should be accountable for complying with these principles, providing training to all employees and contractors who use PII, and auditing the actual use of PII to demonstrate compliance with these principles and all applicable privacy protection requirements.	**Accountability.** A personal information controller should be accountable for complying with measures that give effect to the Principles stated above. When personal information is to be transferred to another person or organization, whether domestically or internationally, the personal information controller should obtain the consent of the individual or exercise due diligence and take reasonable steps to ensure that the recipient person or organization will protect the information consistently with these Principles.

CONSUMER DATA PRIVACY
IN A NETWORKED WORLD:

A FRAMEWORK FOR PROTECTING
PRIVACY AND PROMOTING INNOVATION
IN THE GLOBAL DIGITAL ECONOMY



THE WHITE HOUSE
WASHINGTON

February 23, 2012

Americans have always cherished our privacy. From the birth of our republic, we assured ourselves protection against unlawful intrusion into our homes and our personal papers. At the same time, we set up a postal system to enable citizens all over the new nation to engage in commerce and political discourse. Soon after, Congress made it a crime to invade the privacy of the mails. And later we extended privacy protections to new modes of communications such as the telephone, the computer, and eventually email.

Justice Brandeis taught us that privacy is the "right to be let alone," but we also know that privacy is about much more than just solitude or secrecy. Citizens who feel protected from misuse of their personal information feel free to engage in commerce, to participate in the political process, or to seek needed health care. This is why we have laws that protect financial privacy and health privacy, and that protect consumers against unfair and deceptive uses of their information. This is why the Supreme Court has protected anonymous political speech, the same right exercised by the pamphleteers of the early Republic and today's bloggers.

Never has privacy been more important than today, in the age of the Internet, the World Wide Web and smart phones. In just the last decade, the Internet has enabled a renewal of direct political engagement by citizens around the globe and an explosion of commerce and innovation creating jobs of the future. Much of this innovation is enabled by novel uses of personal information. So, it is incumbent on us to do what we have done throughout history: apply our timeless privacy values to the new technologies and circumstances of our times.

I am pleased to present this new Consumer Privacy Bill of Rights as a blueprint for privacy in the information age. These rights give consumers clear guidance on what they should expect from those who handle their personal information, and set expectations for companies that use personal data. I call on these companies to begin immediately working with privacy advocates, consumer protection enforcement agencies, and others to implement these principles in enforceable codes of conduct. My Administration will work to advance these principles and work with Congress to put them into law. With this Consumer Privacy Bill of Rights, we offer to the world a dynamic model of how to offer strong privacy protection and enable ongoing innovation in new information technologies.

One thing should be clear, even though we live in a world in which we share personal information more freely than in the past, we must reject the conclusion that privacy is an outmoded value. It has been at the heart of our democracy from its inception, and we need it now more than ever.

https://obamawhitehouse.archives.gov/sites/default/files/privacy-final.pdf

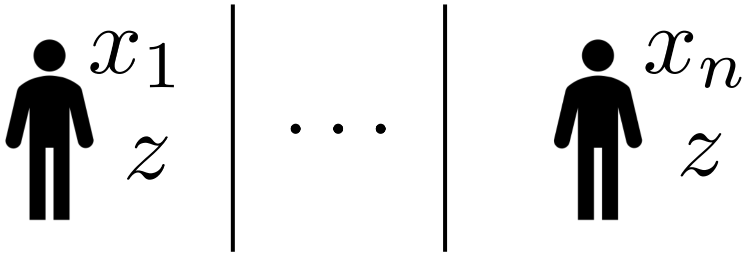
Federated Learning and Security

- Eliminates data collection in the cloud
  - Helps eliminate risk associated with centralized data
    - Warrants, rogue employee, curious executives …
    - “Data is a toxic asset” -- Bruce Schneier
- Privacy?
  - Does not really protect user privacy
  - Model may still reveal user information
    - Adversarial users could manipulate protocol to learn about other users
  - Learning procedure (e.g., FedAvg) leaks raw data
    - Adversarial cloud/coordinator can extract data
- Solutions?

Secure Multi-Party Computation

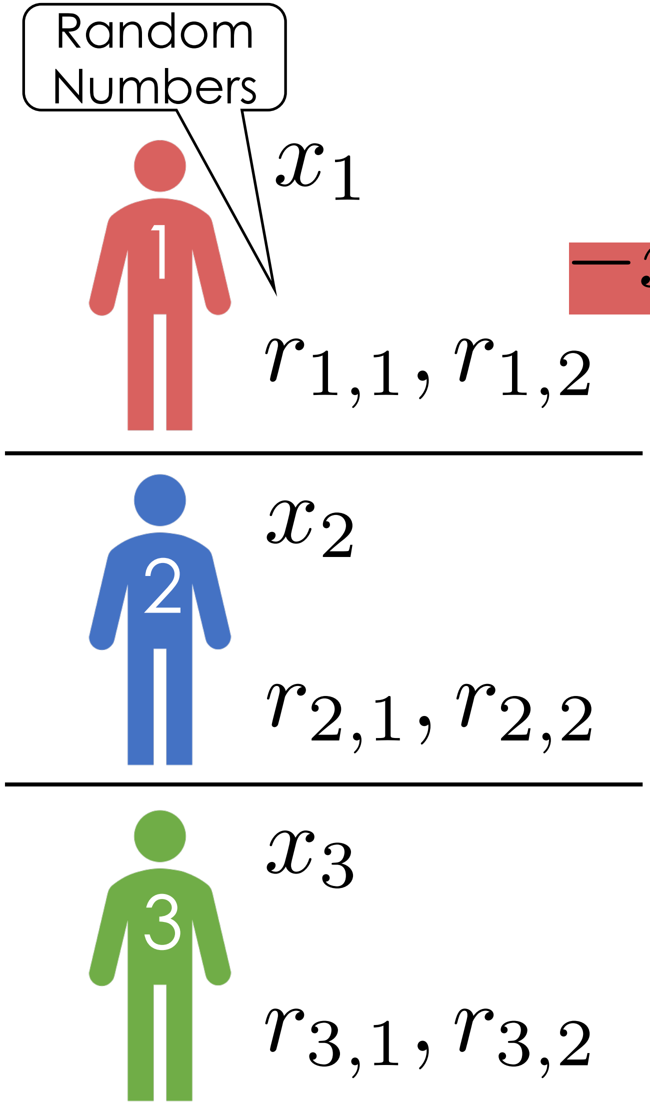
- Family of **cryptographic protocols** designed to allow **multiple parties** to compute a function over their inputs while keeping their inputs private

$$z = f\left(x_1, \ldots, x_n\right)$$



MPC Example: Additive Secret Sharing

$$f(x_1, x_2, x_3) = x_1 + x_2 + x_3$$



x_1

$r_{1,1}, r_{1,2}$

x_2

$r_{2,1}, r_{2,2}$

x_3

$r_{3,1}, r_{3,2}$

Secret Shares

$-x_1 - r_{1,1} - r_{1,2}$

$+x_1 + r_{1,1}$

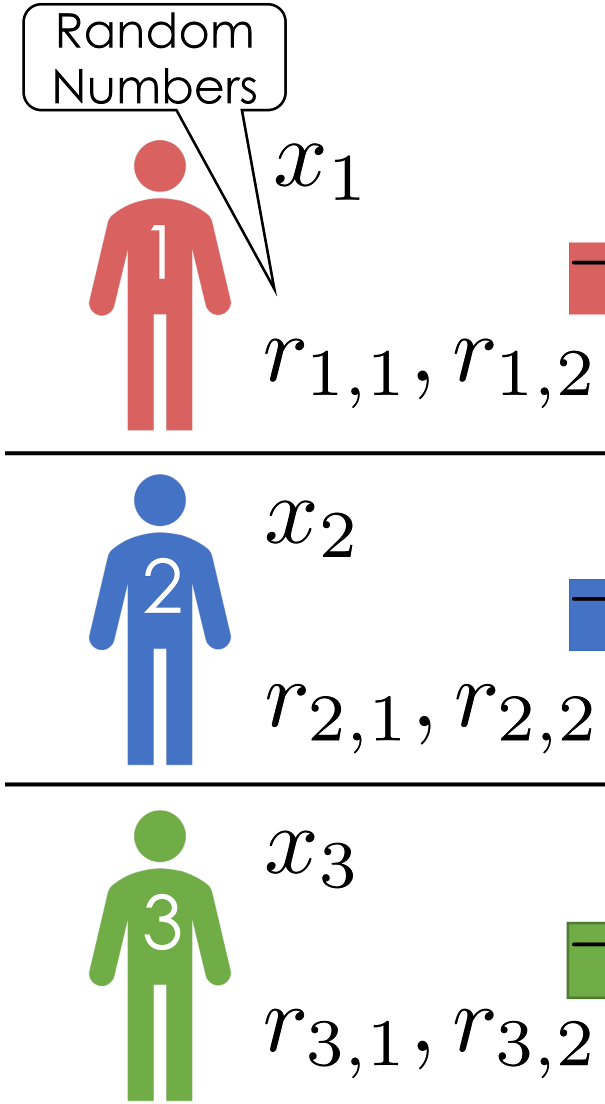
$+x_1 + r_{1,2}$

Note that:

- Neither person 2 or person 3 knows the value of x$_1$.
- If they collude, they cannot figure out the value of x$_1$.

MPC Example: Additive Secret Sharing

$$f\left(x_1, x_2, x_3\right) = x_1 + x_2 + x_3$$



$-x_1 - r_{1,1} - r_{1,2}$ $+x_2 + r_{2,1}$ $+x_3 + r_{3,1}$

$-x_2 - r_{2,1} - r_{2,2}$ $+x_1 + r_{1,1}$ $+x_3 + r_{3,2}$

$-x_3 - r_{3,1} - r_{3,2}$ $+x_1 + r_{1,2}$ $+x_2 + r_{2,2}$

MPC Example: Additive Secret Sharing

$$f\left(x_1, x_2, x_3\right) = x_1 + x_2 + x_3$$

Parties all share their partial sums

$$\begin{aligned}
&-x_1 - r_{1,1} - r_{1,2} \quad +x_2 + r_{2,1} \quad +x_3 + r_{3,1} \\
&+ \quad -x_2 - r_{2,1} - r_{2,2} \quad +x_1 + r_{1,1} \quad +x_3 + r_{3,2} \\
&+ \quad -x_3 - r_{3,1} - r_{3,2} \quad +x_1 + r_{1,2} \quad +x_2 + r_{2,2}
\end{aligned}$$

MPC Example: Additive Secret Sharing

$$f\left(x_1, x_2, x_3\right) = x_1 + x_2 + x_3$$

Rearranging and cancelling out terms:

$$-\cancel{x_1} - \cancel{r_{1,1}} - \cancel{r_{1,2}} \quad +\cancel{x_1} + \cancel{r_{1,1}} \quad +x_1 + \cancel{r_{1,2}}$$

$$+ \quad -\cancel{x_2} - \cancel{r_{2,1}} - \cancel{r_{2,2}} \quad +\cancel{x_2} + \cancel{r_{2,1}} \quad +x_2 + \cancel{r_{2,2}}$$

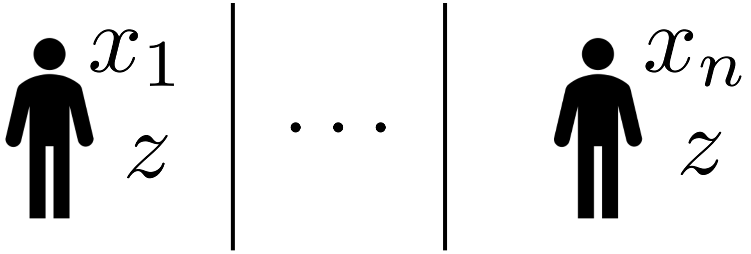
$$+ \quad -\cancel{x_3} - \cancel{r_{3,1}} - \cancel{r_{3,2}} \quad +\cancel{x_3} + \cancel{r_{3,1}} \quad +x_3 + \cancel{r_{3,2}}$$

$$= x_1 + x_2 + x_3$$

Secure Multi-Party Computation

- Family of **cryptographic protocols** designed to allow **multiple parties** to compute a function over their inputs while keeping the inputs private

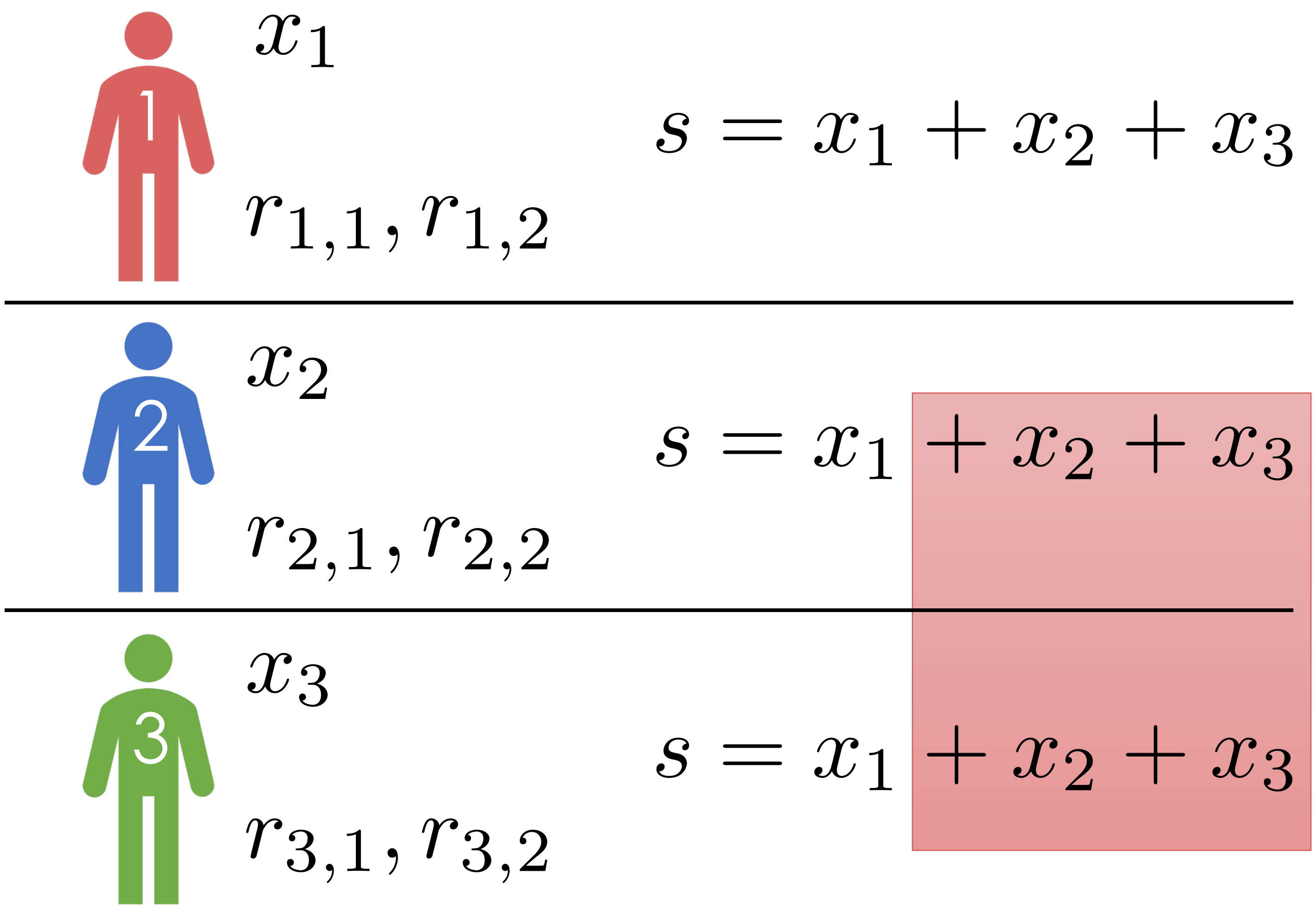
$$z = f\left(x_1, \ldots, x_n\right)$$



- More sophisticated protocols
  - **Shamir secret sharing**: uses evaluation of random polynomials
    - Supports **addition** and **multiplication**
    - User specified fault tolerance (not all parties must participate)
- Privacy?

MPC Example: Additive Secret Sharing

$$f\left(x_1, x_2, x_3\right) = x_1 + x_2 + x_3$$



x_1

$s = x_1 + x_2 + x_3$

$r_{1,1}, r_{1,2}$

x_2

$s = x_1 + x_2 + x_3$

$r_{2,1}, r_{2,2}$

x_3

$s = x_1 + x_2 + x_3$

$r_{3,1}, r_{3,2}$

Note that:

- Neither person 2 or person 3 knows the value of x_1.
- If they collude, they cannot figure out the value of x_1.

The computation still reveals information.

In secure MPC we assume that each party is willing to share outcome of computation.

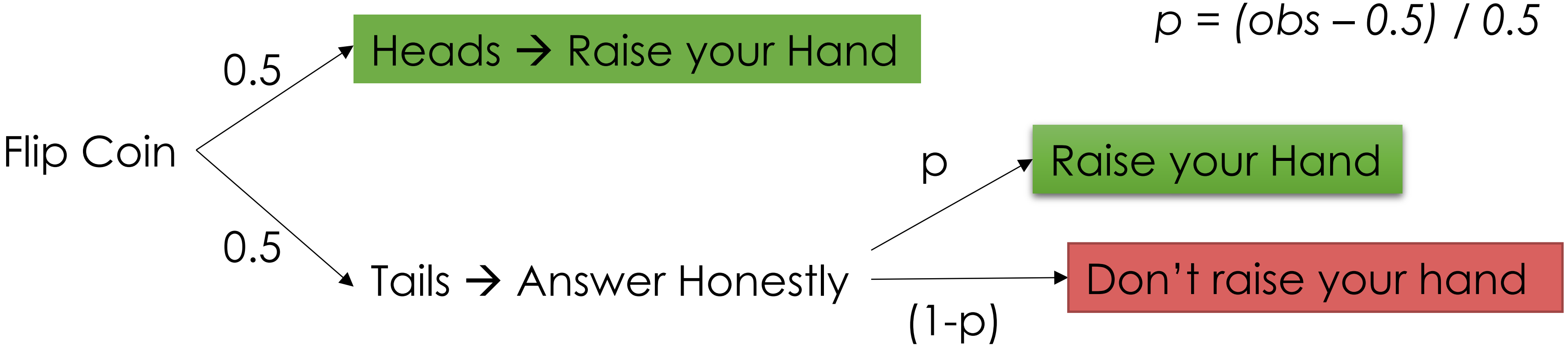
Differential Privacy

- Guarantees that only a **limited** amount of **information** is leaked about data involved in a computation.
  - Formalized in 2006 by Dwork, McSherry, Nissim, and Smith
  - Builds on randomized response (1965)

Have you cheated on an exam?

$\text{obs} = 0.5 + 0.5\,p$

$p = (\text{obs} - 0.5)\ /\ 0.5$



Flip Coin

0.5

Heads → Raise your Hand

0.5

Tails → Answer Honestly

p

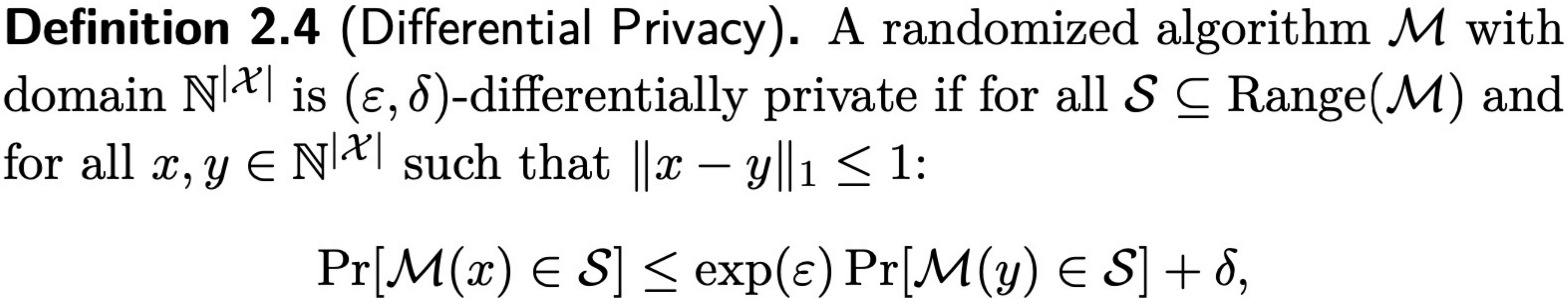
Raise your Hand

(1-p)

Don't raise your hand

Differential Privacy

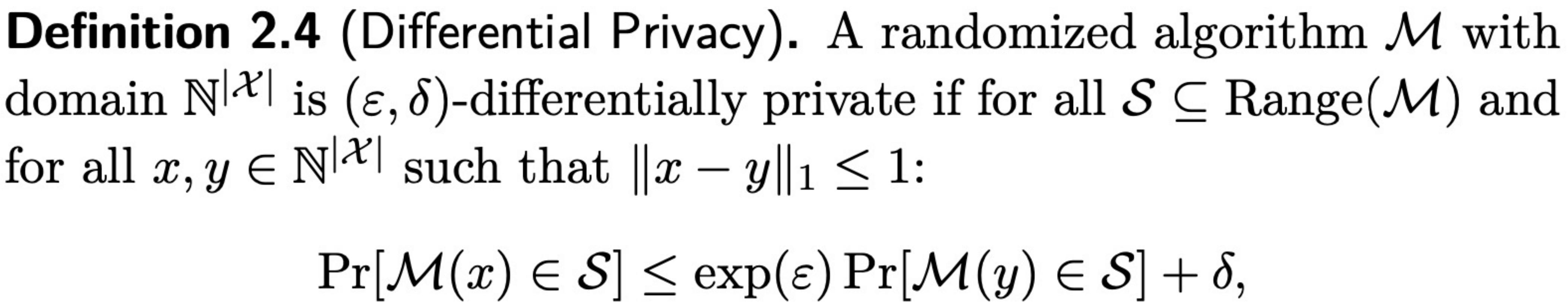
- Guarantees that only a **limited** amount of **information** is leaked about data involved in a computation.
  - Formalized in 2006 by Dwork, McSherry, Nissim, and Smith
  - Builds on randomized response (1965)



Definition 2.4 (Differential Privacy). A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is (ε, δ)-differentially private if for all $\mathcal{S} \subseteq \mathrm{Range}(\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $\|x - y\|_1 \leq 1$:

$$\Pr[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(y) \in \mathcal{S}] + \delta,$$

https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf



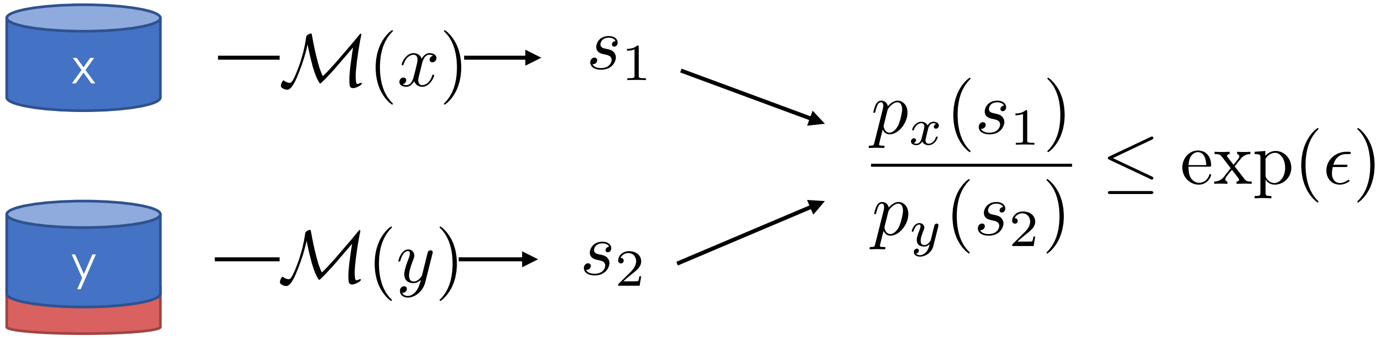
Definition 2.4 (Differential Privacy). A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is (ε, δ)-differentially private if for all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $\|x - y\|_1 \leq 1$:

$$\Pr[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(y) \in \mathcal{S}] + \delta,$$

$x, y \in \mathbb{N}^{|\mathcal{X}|}$: Represent databases x and y as histograms.

$||x - y||_1 \leq 1$: Databases differ by only one record.

$\delta = 0$: Special case, (ϵ)-differential private.



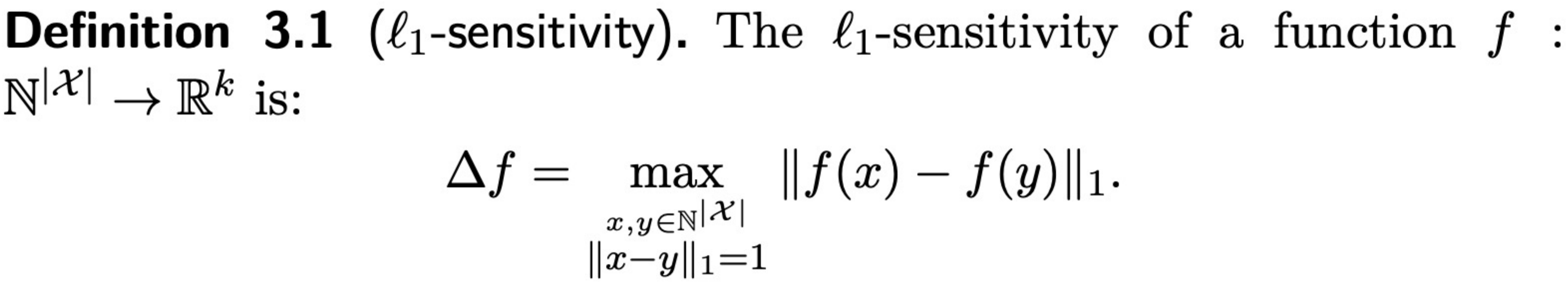
$$\frac{p_x(s_1)}{p_y(s_2)} \leq \exp(\epsilon)$$

https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

Achieving Differential Privacy

- Several mechanisms built around the addition of noise
- Laplace mechanism is most common/simple

Laplace Mechanism

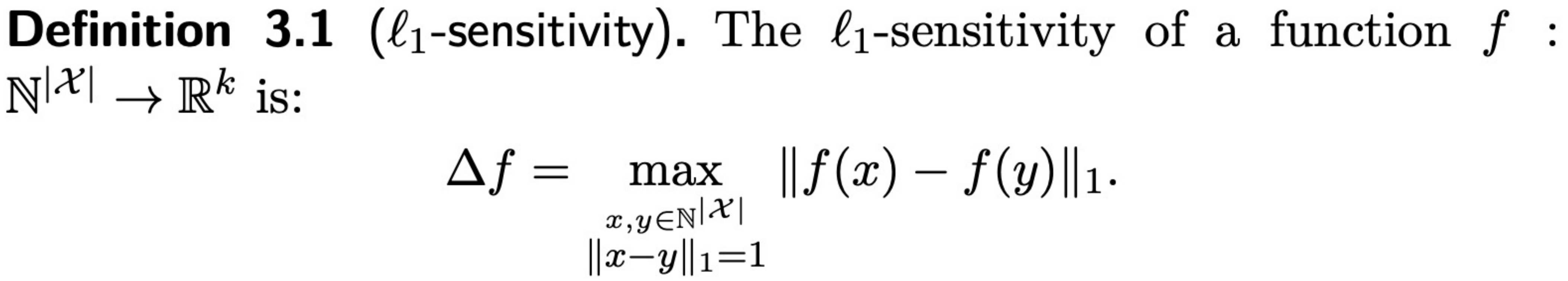


Definition 3.1 (ℓ_1**-sensitivity).** The ℓ_1-sensitivity of a function $f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^k$ is:

$$\Delta f = \max_{\substack{x,y \in \mathbb{N}^{|\mathcal{X}|} \\ \|x-y\|_1 = 1}} \|f(x) - f(y)\|_1.$$

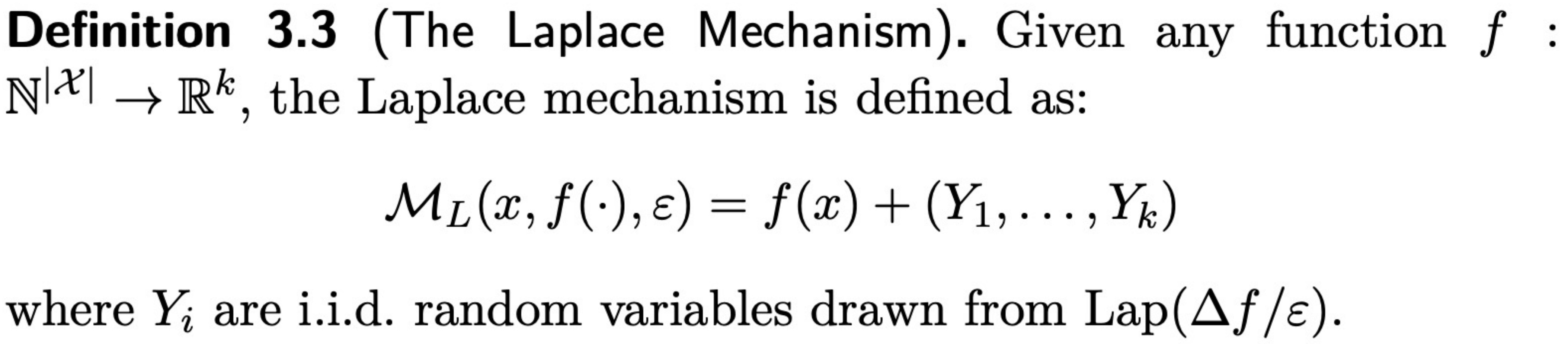
https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

Laplace Mechanism



Definition 3.1 (ℓ_1-sensitivity)**.** The ℓ_1-sensitivity of a function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$ is:

$$\Delta f = \max_{\substack{x,y \in \mathbb{N}^{|\mathcal{X}|} \\ \|x-y\|_1 = 1}} \|f(x) - f(y)\|_1.$$



Definition 3.3 (The Laplace Mechanism)**.** Given any function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$, the Laplace mechanism is defined as:

$$\mathcal{M}_L(x, f(\cdot), \varepsilon) = f(x) + (Y_1, \ldots, Y_k)$$

where Y_i are i.i.d. random variables drawn from $\mathrm{Lap}(\Delta f/\varepsilon)$.

https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

Definition 3.3 (The Laplace Mechanism). Given any function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$, the Laplace mechanism is defined as:

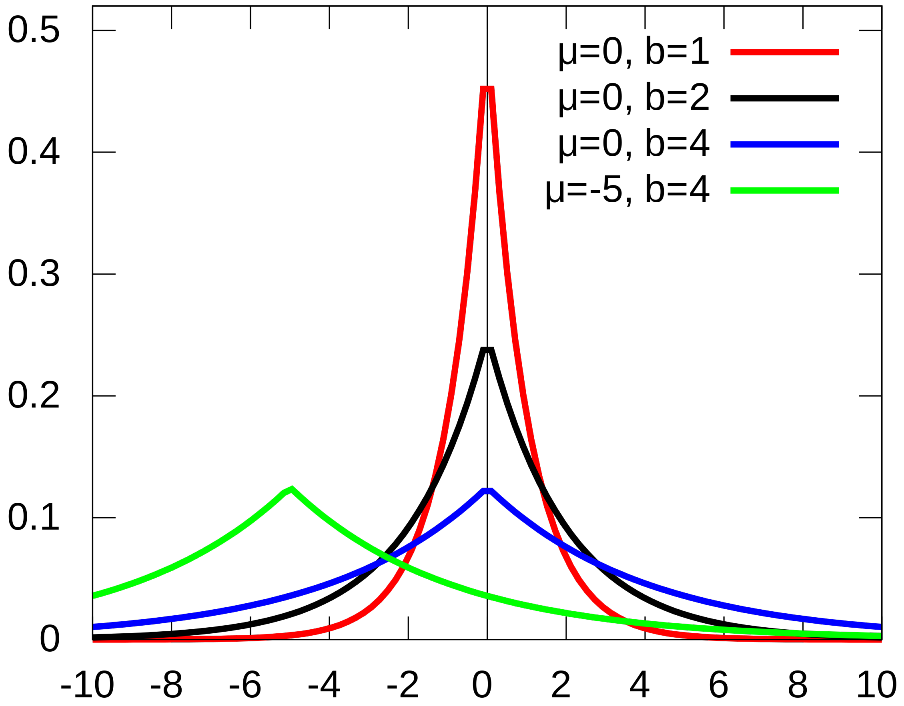
$$\mathcal{M}_L(x, f(\cdot), \varepsilon) = f(x) + (Y_1, \ldots, Y_k)$$

Zero Mean

$\text{var.} = 2\left(\Delta f/\epsilon\right)^2$

where Y_i are i.i.d. random variables drawn from $\text{Lap}(\Delta f/\varepsilon)$.

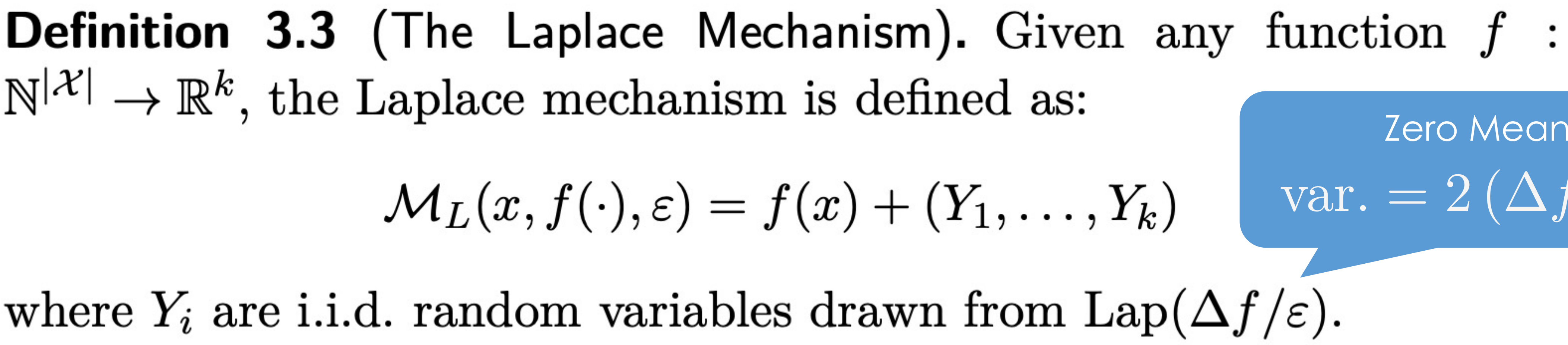
Samples from Laplace distribution.



Proof

$$\frac{\Pr\left(f(x) + \text{Lap}\left(\frac{\Delta f}{\epsilon}\right) = z\right)}{\Pr\left(f(y) + \text{Lap}\left(\frac{\Delta f}{\epsilon}\right) = z\right)}$$

https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf



Definition 3.3 (The Laplace Mechanism). Given any function $f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^k$, the Laplace mechanism is defined as:

$$\mathcal{M}_L(x, f(\cdot), \varepsilon) = f(x) + (Y_1, \ldots, Y_k)$$

where Y_i are i.i.d. random variables drawn from $\mathrm{Lap}(\Delta f/\varepsilon)$.

Zero Mean

$\text{var.} = 2\left(\Delta f/\epsilon\right)^2$

Proof

$$\frac{\Pr\left(f(x) + \mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right) = z\right)}{\Pr\left(f(y) + \mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right) = z\right)} = \frac{\Pr\left(\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right) = z - f(x)\right)}{\Pr\left(\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right) = z - f(y)\right)}$$

https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

Proof

$$\frac{\Pr\left(f(x)+\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)=z\right)}{\Pr\left(f(y)+\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)=z\right)}=\frac{\Pr\left(\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)=z-f(x)\right)}{\Pr\left(\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)=z-f(y)\right)}$$

Product of K iid draws from
A Laplace Distribution

$$=\prod_{i=1}^{k}\left(\frac{\exp(-\frac{\varepsilon|f(x)_i-z_i|}{\Delta f})}{\exp(-\frac{\varepsilon|f(y)_i-z_i|}{\Delta f})}\right)$$

$$=\prod_{i=1}^{k}\exp\left(\frac{\varepsilon(|f(y)_i-z_i|-|f(x)_i-z_i|)}{\Delta f}\right)$$

$$\leq\prod_{i=1}^{k}\exp\left(\frac{\varepsilon|f(x)_i-f(y)_i|}{\Delta f}\right)$$

$$=\exp\left(\frac{\varepsilon\cdot\|f(x)-f(y)\|_1}{\Delta f}\right)$$

$$\leq\exp(\varepsilon),$$

https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

Proof

$$\frac{\Pr\left(f(x)+\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)=z\right)}{\Pr\left(f(y)+\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)=z\right)}=\frac{\Pr\left(\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)=z-f(x)\right)}{\Pr\left(\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)=z-f(y)\right)}$$

Product of K iid draws from A Laplace Distribution

$$=\prod_{i=1}^{k}\left(\frac{\exp(-\frac{\varepsilon|f(x)_i-z_i|}{\Delta f})}{\exp(-\frac{\varepsilon|f(y)_i-z_i|}{\Delta f})}\right)$$

Algebra

$$=\prod_{i=1}^{k}\exp\left(\frac{\varepsilon(|f(y)_i-z_i|-|f(x)_i-z_i|)}{\Delta f}\right)$$

$$\leq\prod_{i=1}^{k}\exp\left(\frac{\varepsilon|f(x)_i-f(y)_i|}{\Delta f}\right)$$

$$=\exp\left(\frac{\varepsilon\cdot\|f(x)-f(y)\|_1}{\Delta f}\right)$$

$$\leq\exp(\varepsilon),$$

https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

Proof

$$\frac{\Pr\left(f(x) + \mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right) = z\right)}{\Pr\left(f(y) + \mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right) = z\right)} = \frac{\Pr\left(\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right) = z - f(x)\right)}{\Pr\left(\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right) = z - f(y)\right)}$$

Product of K iid draws from A Laplace Distribution

$$= \prod_{i=1}^{k} \left(\frac{\exp(-\frac{\varepsilon |f(x)_i - z_i|}{\Delta f})}{\exp(-\frac{\varepsilon |f(y)_i - z_i|}{\Delta f})} \right)$$

Algebra

$$= \prod_{i=1}^{k} \exp\left(\frac{\varepsilon(|f(y)_i - z_i| - |f(x)_i - z_i|)}{\Delta f} \right)$$

Triangle Inequality

$$\leq \prod_{i=1}^{k} \exp\left(\frac{\varepsilon |f(x)_i - f(y)_i|}{\Delta f} \right)$$

$$= \exp\left(\frac{\varepsilon \cdot \|f(x) - f(y)\|_1}{\Delta f} \right)$$

$$\leq \exp(\varepsilon),$$

https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

Proof

$$\frac{\Pr\left(f(x)+\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)=z\right)}{\Pr\left(f(y)+\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)=z\right)}=\frac{\Pr\left(\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)=z-f(x)\right)}{\Pr\left(\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)=z-f(y)\right)}$$

Product of K iid draws from A Laplace Distribution

$$=\prod_{i=1}^{k}\left(\frac{\exp(-\frac{\varepsilon|f(x)_i-z_i|}{\Delta f})}{\exp(-\frac{\varepsilon|f(y)_i-z_i|}{\Delta f})}\right)$$

Algebra

$$=\prod_{i=1}^{k}\exp\left(\frac{\varepsilon(|f(y)_i-z_i|-|f(x)_i-z_i|)}{\Delta f}\right)$$

Triangle Inequality

$$\leq\prod_{i=1}^{k}\exp\left(\frac{\varepsilon|f(x)_i-f(y)_i|}{\Delta f}\right)$$

Algebra

$$=\exp\left(\frac{\varepsilon\cdot\|f(x)-f(y)\|_1}{\Delta f}\right)$$

$$\leq\exp(\varepsilon),$$

https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

Proof

$$\frac{\Pr\left(f(x) + \mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right) = z\right)}{\Pr\left(f(y) + \mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right) = z\right)} = \frac{\Pr\left(\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right) = z - f(x)\right)}{\Pr\left(\mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right) = z - f(y)\right)}$$

Product of K iid draws from A Laplace Distribution

$$= \prod_{i=1}^{k} \left(\frac{\exp(-\frac{\varepsilon |f(x)_i - z_i|}{\Delta f})}{\exp(-\frac{\varepsilon |f(y)_i - z_i|}{\Delta f})} \right)$$

Algebra

$$= \prod_{i=1}^{k} \exp\left(\frac{\varepsilon(|f(y)_i - z_i| - |f(x)_i - z_i|)}{\Delta f} \right)$$

Triangle Inequality

$$\leq \prod_{i=1}^{k} \exp\left(\frac{\varepsilon |f(x)_i - f(y)_i|}{\Delta f} \right)$$

$$\Delta f = \max_{\substack{x,y \in \mathbb{N}^{|\mathcal{X}|} \\ \|x-y\|_1 = 1}} \|f(x) - f(y)\|_1$$

Algebra

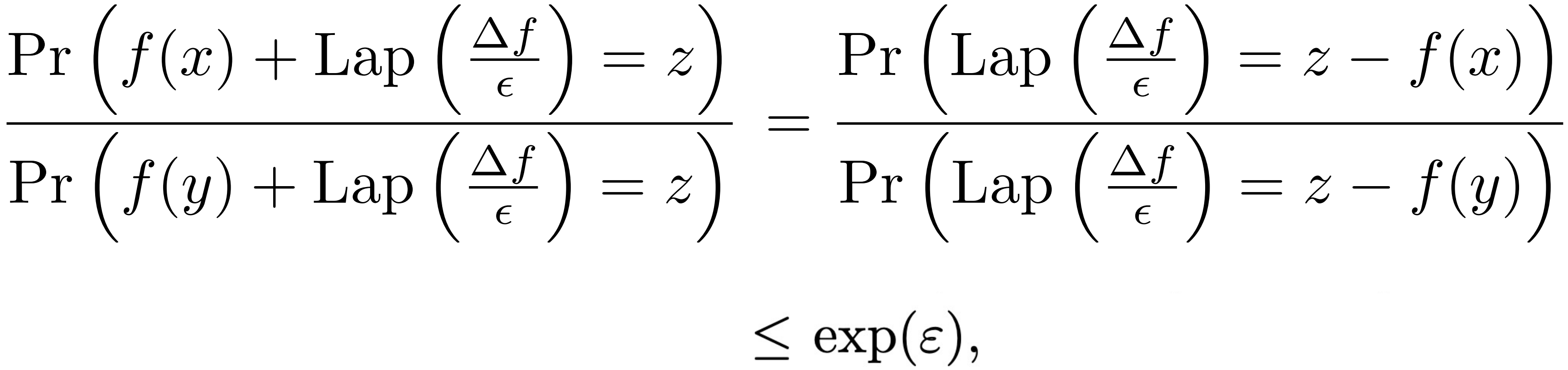
$$= \exp\left(\frac{\varepsilon \cdot \|f(x) - f(y)\|_1}{\Delta f} \right)$$

Definition of Sensitivity

$$\leq \exp(\varepsilon),$$

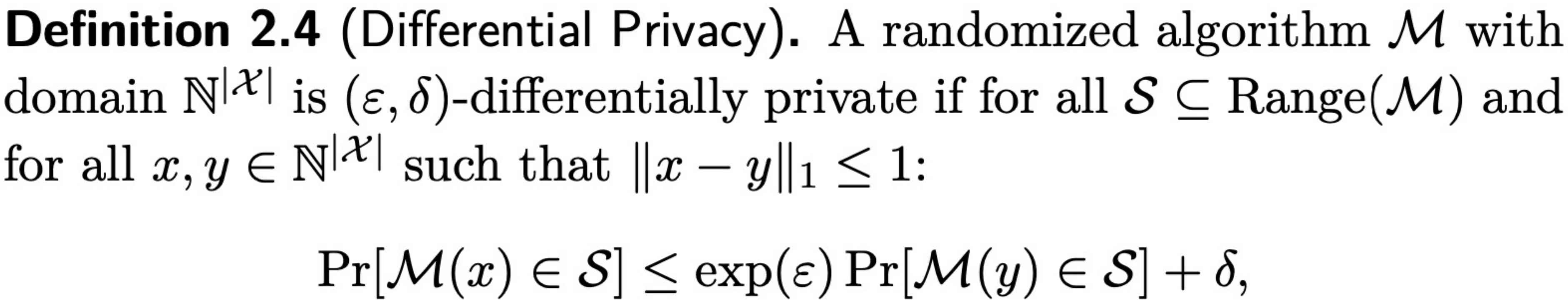
https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

Proof



$$\frac{\Pr\left(f(x) + \operatorname{Lap}\left(\frac{\Delta f}{\epsilon}\right) = z\right)}{\Pr\left(f(y) + \operatorname{Lap}\left(\frac{\Delta f}{\epsilon}\right) = z\right)} = \frac{\Pr\left(\operatorname{Lap}\left(\frac{\Delta f}{\epsilon}\right) = z - f(x)\right)}{\Pr\left(\operatorname{Lap}\left(\frac{\Delta f}{\epsilon}\right) = z - f(y)\right)}$$

$$\leq \exp(\varepsilon),$$



Definition 2.4 (Differential Privacy). A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is (ε, δ)-differentially private if for all $\mathcal{S} \subseteq \mathrm{Range}(\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $\|x - y\|_1 \leq 1$:

$$\Pr[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(y) \in \mathcal{S}] + \delta,$$

https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

Differential Privacy: Issues

- **Privacy-Utility Tradeoff:** What are the right values of (ε, δ) to balance privacy and accuracy?

- **Privacy Budget:** each differentially private operation on the data leaks information.
  - At some point you must stop querying the data

Trusted Execution Environments

- **Hardware primitives** that **isolate data** and **computation** from other **processes**, the **OS**, and even **hardware attacks** while **attesting** to the correct execution.
  - **isolation:** typically encrypt data in memory and prevent other processes including the OS from observing computation
  - **attestation:** verify the integrity of the computation as well as all inputs.

- Examples:
  - Intel SGX (Software Guard Extensions)
  - AMD Memory Encryption
  - ARM TrustZone

Trusted Execution Environments

- **Advantages**
  - Minimal impact on compute performance.
  - Support for fully generic computation
  - Leverage industry standard encryption rather than PHE/FHE
- **Disadvantages**
  - Need to trust hardware manufacturer
    - Sign certificates and verify code
    - Bugs: exploits like Spectre and Meltdown can compromise SGX
  - Susceptible to a wide range of side channel attacks
  - Current implementations have some performance limitations
    - SGX Enclave Page Cache (EPC) is limited to ~100MB → very slow to move data in and out of EPC
    - Single socket desktop processor

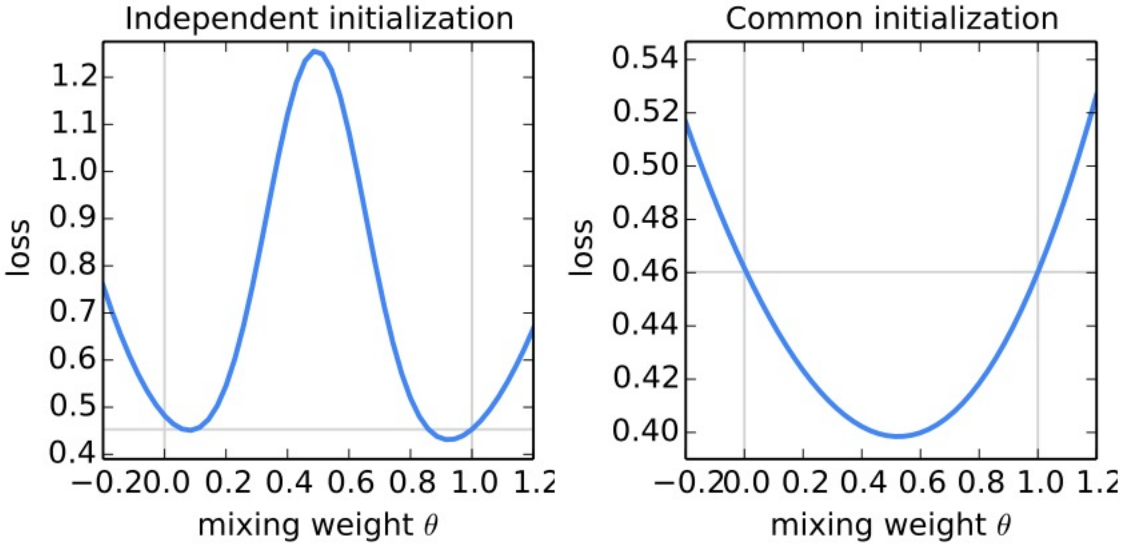
Reading This Week

Reading for the Week

- Communication-Efficient Learning of Deep Networks from Decentralized Data (AIStat'17)
  - Introduced Federated Learning setting and federated averaging
- Privacy Accounting and Quality Control in the Sage Differentially Private ML Platform (SOSP'19)
  - Differential privacy ML platform which doesn't expire
- ~~Slalom: Fast, Verifiable and Private Execution of Neural Networks in Trusted Hardware(ICLR'19)~~
  - ~~Combine trusted execution environment (SGX) with untrusted hardware (GPU) for inference~~
- Robust Physical-World Attacks on Deep Learning Models [CVPR' 18]
  - Highly influential paper on on adversarial attacks on computer vision models

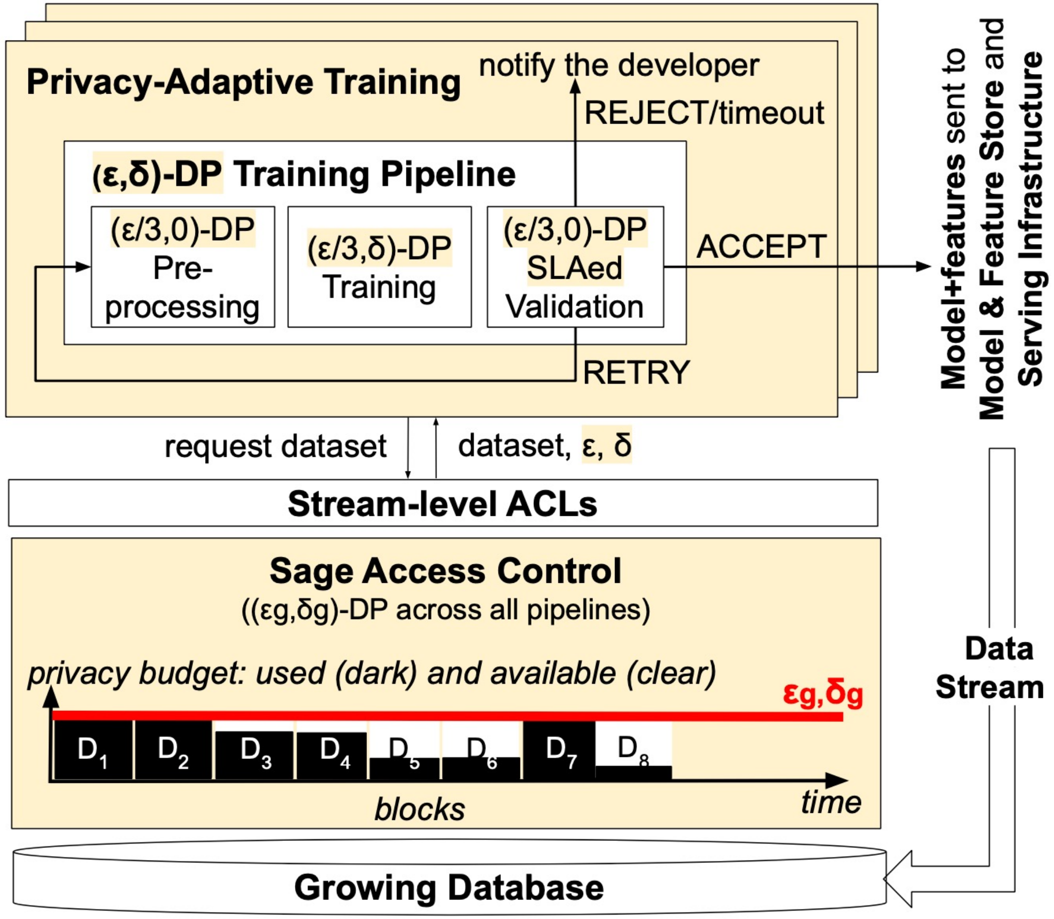
Communication-Efficient Learning of Deep Networks from Decentralized Data (AIStats'17)

- Early work framing the federated setting
  - Identified several key challenges: comm., non-iid, participation
- Advocating for distributed model averaging in the federated setting
  - Counter intuitive (at the time)
    - Non-iid assumption
    - Non-convexity of deep learning



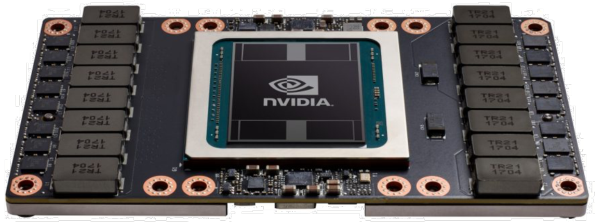
Privacy Accounting and Quality Control in the Sage Differentially Private ML Platform (SOSP'19)

- Platform for differential privacy accounting that interposes in existing ML frameworks.
- Addresses the problems:
  - **Managing** a privacy budget and **extending** the privacy budget
  - Balancing **privacy** and **accuracy** objectives with retraining
- **Big Idea:** Break data into blocks with independent privacy budgets


Privacy-Adaptive Training
notify the developer
REJECT/timeout
(ε,δ)-DP Training Pipeline
(ε/3,0)-DP Pre-processing
(ε/3,δ)-DP Training
(ε/3,0)-DP SLAed Validation
ACCEPT
RETRY
Model+features sent to Model & Feature Store and Serving Infrastructure
request dataset
dataset, ε, δ
Stream-level ACLs
Sage Access Control
((εg,δg)-DP across all pipelines)
privacy budget: used (dark) and available (clear)
εg,δg
D1 D2 D3 D4 D5 D6 D7 D8
blocks
time
Data Stream
Growing Database

Slalom: Fast, Verifiable and Private Execution of Neural Networks in Trusted Hardware (ICLR’19)

- Use cryptographic techniques to combine TEE with untrusted accelerator during inference.
- **Big Idea:** Leverage linearity operators to compute over pre-computed noise



Slalom with integrity & privacy

```
TEE(F, x_1)                                                        S(F)
----------------------------------------------------------------------
Preproc: for $i \in [1, n]$ do $r_i \xleftarrow{R} \mathbb{F}^{m_i}$, $u_i = r_i W_i$
----------------------------------------------------------------------
for $i \in [1, n]$ do
    $\tilde{x}_i = x_i + r_i$                  $\xrightarrow{\tilde{x}_i}$
                                               $\xleftarrow{\tilde{y}_i}$        $\tilde{y}_i = \tilde{x}_i W_i$
    $y_i = \tilde{y}_i - u_i$
    assert Freivalds($y_i, x_i, W_i$)
    $x_{i+1} = \sigma(y_i)$
return $y_n$
```

Done!