Data Privacy

* Simplest privacy method: Anonymization.
  1. Remove "identifying" bits – eg, names, addresses, etc.
  2. Publish data.
* This has serious problems.

  eg. AOL 2006, Netflix 2008.
* Why? People's data tends to be very unique. For example:

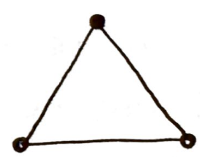
Gender	Position	Dept	Ethnicity
F	Faculty	CSE	S Asian

Only one person, Kamalika, fits description. "Linkage" information on CSE website.

* Statistics from data also problematic.

  eg, Wang et al study

  Histogram with outliers.



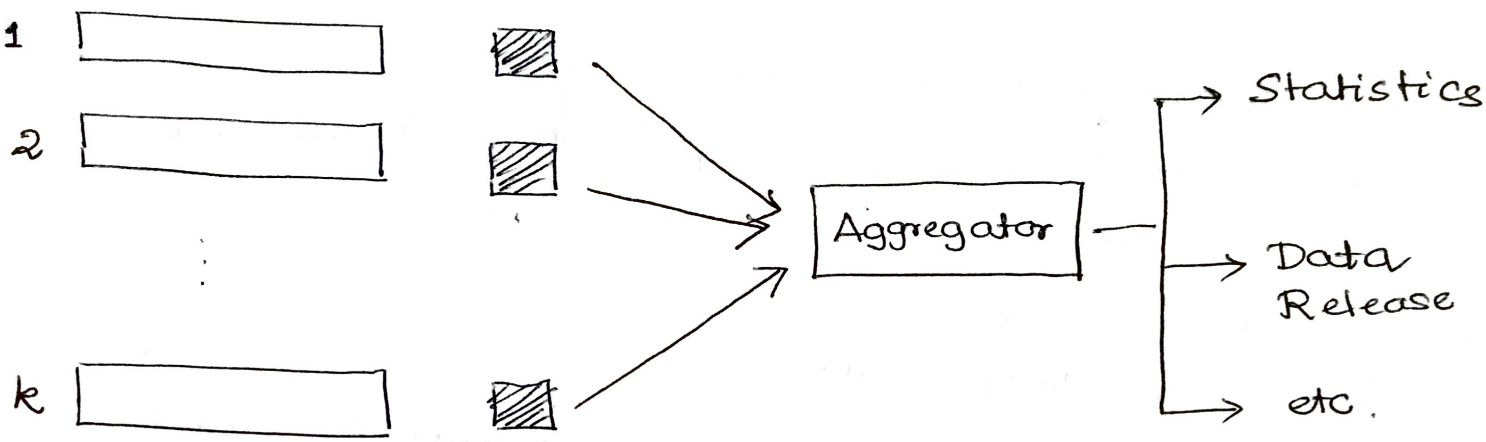
privacy – utility – sample size tradeoffs.

* Need robust, rigorous measures to preserve privacy.

Three Privacy Settings:

1. Local sanitizer
2. Centralized sanitizer
3. Consent management.

$\boxed{1}$ Local Sanitizer



1. Aggregator is untrusted.
2. Sanitization happens locally before passing on to Aggregator.

<u>Example:</u> Randomized Response. [Werner 1965]

Each person is asked if they use an illegal drug (Yes or No). Everyone takes their answer, flips w.p p and returns it.

- $\boxed{\text{Privacy offered : deniability.}}$

$$\Pr(\text{Output} = Y \mid \text{True value} = Y) = 1-p$$

$$\Pr(\text{Output} = N \mid \text{True value} = Y) = p$$

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

Using Bayes Rule,

$$\Pr(\text{True} = Y \mid \text{Output} = Y) = \frac{\Pr(\text{Output} = Y \mid \text{True} = Y)\Pr(\text{True} = Y)}{\Pr(\text{Output} = Y)}$$

$$\Pr(\text{Output} = Y \mid \text{True} = Y)\Pr(\text{True} = Y) + \Pr(\text{Output} = Y \mid \text{True} = N)\Pr(\text{True} = N)$$

$$= \frac{(1-p)\frac{1}{2}}{\frac{1}{2}\cdot(1-p) + \frac{1}{2}\cdot p} = 1-p.$$

What is the utility offered?

Suppose we are aggregating N responses, and f is the fraction of drug uses.

$$\mathbb{E}[\#\text{Yes}] = N \Pr[\text{Output} = Y]$$

$$\Pr(\text{Output}=Y) = \Pr(\text{Output}=Y \mid \text{True}=Y)\ \Pr(\text{True}=Y) + \Pr(\text{Output}=Y \mid \text{True}=N)\ \Pr(\text{True}=N)$$

$$= (1-p)f + p(1-f) = p+f-2pf$$

$$\mathbb{E}[\#\text{Yes}] = N(p+f-2pf) = Np + N(1-2p)f$$

$$\text{Var}(\#\text{Yes}) \lesssim \sqrt{N}$$

Let $$T = \frac{\#\text{Yes} - Np}{N(1-2p)}$$

Then $\mathbb{E}[T] = f$

$\text{Var}(T) \lesssim \dfrac{1}{\sqrt{N}(1-2p)}$

Can estimate the fraction of drug users if p is not too close to $\frac{1}{2}$.

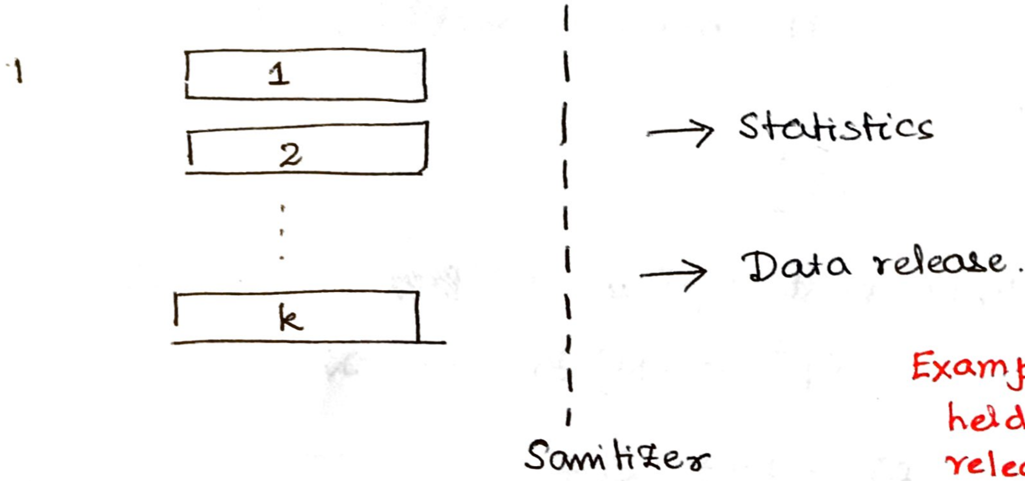
Note: Privacy – Utility – Datasize trade off

$p \approx \frac{1}{2}$: High privacy, but high variance of T, so low utility.

Trade off is better when N is $\boxed{\text{high}}$

__Applications:__ Data collection systems in companies, eg. Google, Apple, etc.

2 Centralized Sanitizer



Example: Medical data, held at a hospital, summaries released.

Example 2: Census (2020 Census)

1. Sanitizer is trusted.
2. Sanitizer sees / collects raw sensitive data and "privatizes" it and passes the privatized version to the public sphere.

What is a good notion of privacy for such a setting?
Differential privacy.

Main idea: Participation of a single person does not make a difference.

Alice + data → Randomized Algorithm → Output

Bob + data → Randomized Algorithm → Output

} similar.
[~~[illegible]~~ output distributions are similar]

Adversary: whatever she can learn about Alice from algorithm's output, she can learn even if Alice is not in data.

Example 1: Study shows smoking causes cancer, Adversary knows Alice smokes → infers Alice may have cancer.

$\boxed{\text{NOT a privacy violation.}}$

xample 2: Study in Wong et al with Alice in data. Adversary knows about Alice's genome → solves equations and finds Alice in Cancer group

Privacy Violation

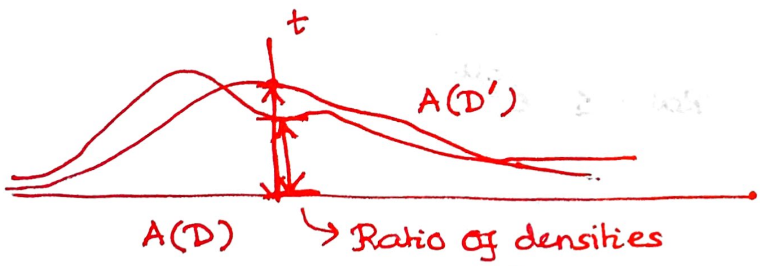
Formally, Differential privacy definition:

A mechanism A is ε-DP if for all datasets D and D' that differ in the private value of a single person and for all t,

$$Pr(A(D) = t) \leq e^{\varepsilon} Pr(A(D') = t).$$

ε = privacy budget

Pr = over randomization of the algorithm A.

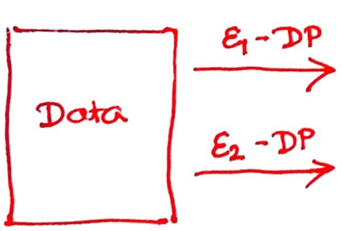


→ Ratio of densities Bounded.

Properties: $\boxed{1}$ Post processing Invariance.

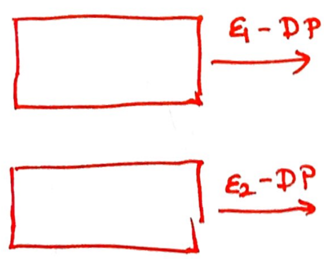
Data ⇏ ε-DP answer ⇏ ε-DP (for any transformation)

$\boxed{2}$ (Graceful) Composition



$\varepsilon_1 + \varepsilon_2$-DP

Sequential



$\max(\varepsilon_1, \varepsilon_2)$-DP

Parallel

How to get DP?

- The Global Sensitivity Mechanism.
- Global sensitivity of a function f:

$$GS(f) = \max_{\substack{D, D' \\ |D \setminus D'| = 1}} |f(D) - f(D')|$$

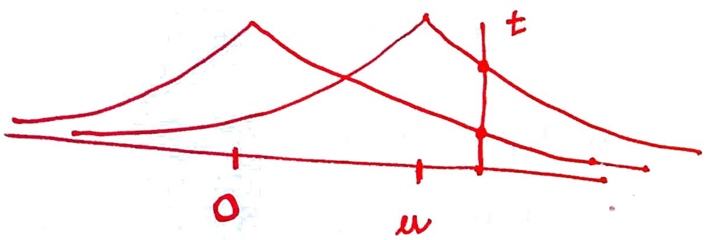
Mechanism:

Output $f(D) + \frac{GS(f)}{\varepsilon} Z$ where $Z \sim Lap(\frac{1}{\varepsilon})$.

Laplace distribution:

$$f(z) = \frac{1}{2} e^{-|z|/b}$$

[mean 0, stdev: $\sqrt{2}b$]



Ratio $\leq e^{\varepsilon u}$

Use this mechanism + composition to get more complex ones.

Utility: Back to the drug user problem, if we use DP, then

GS of mean of n bits $= \frac{1}{n}$.

So stdev of noise added $= \frac{\sqrt{2}}{n\varepsilon}$ (as opposed to $\frac{1}{\sqrt{n}(1-2p)}$ for RR)

So better privacy - utility - data size tradeoff

4)