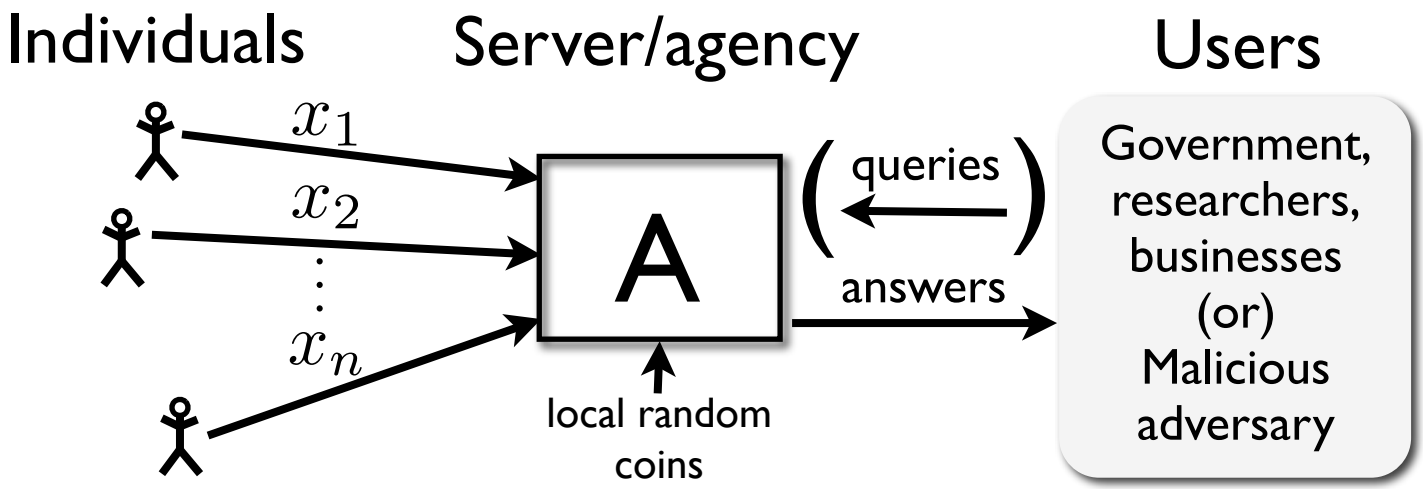
Integrating Differential Privacy with Statistical Theory

Adam Smith
Computer Science & Engineering Department
Penn State

http://www.cse.psu.edu/~asmith

Eagl II: October 4, 2009

Privacy in Statistical Databases



Large collections of personal information

- census data
- medical/public health data
- social networks
- recommendation systems
- trace data: search records, etc
- intrusion-detection systems

Recently:

- larger data sets
- more types of data

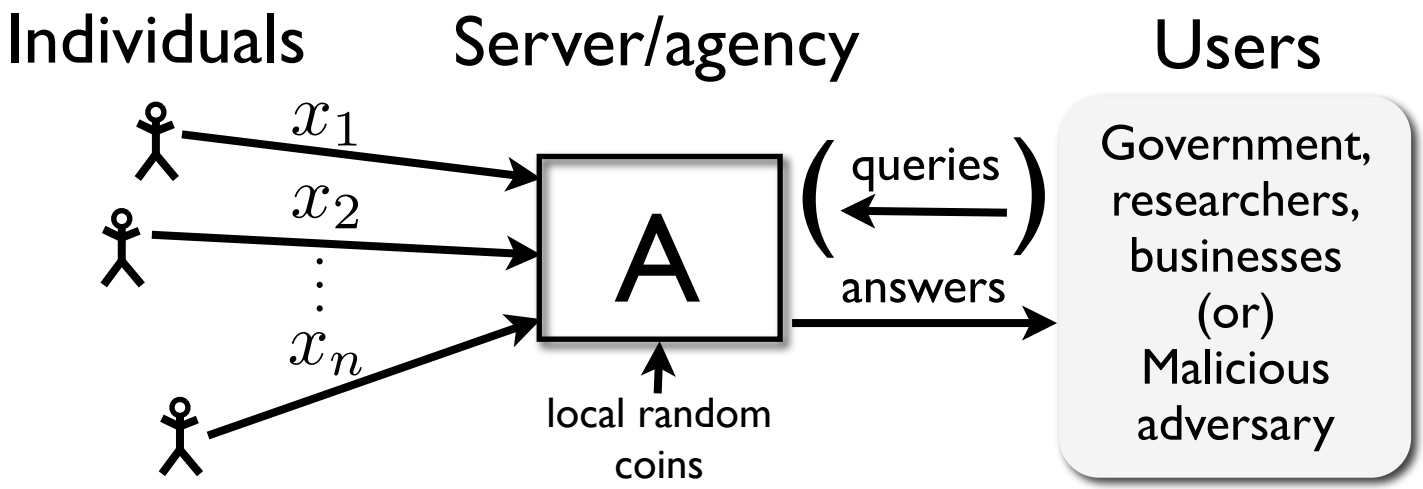
Privacy in Statistical Databases

Privacy in Statistical Databases

- Two conflicting goals
  - Utility: Users can extract "global" statistics
  - Confidentiality: Individual information stays hidden

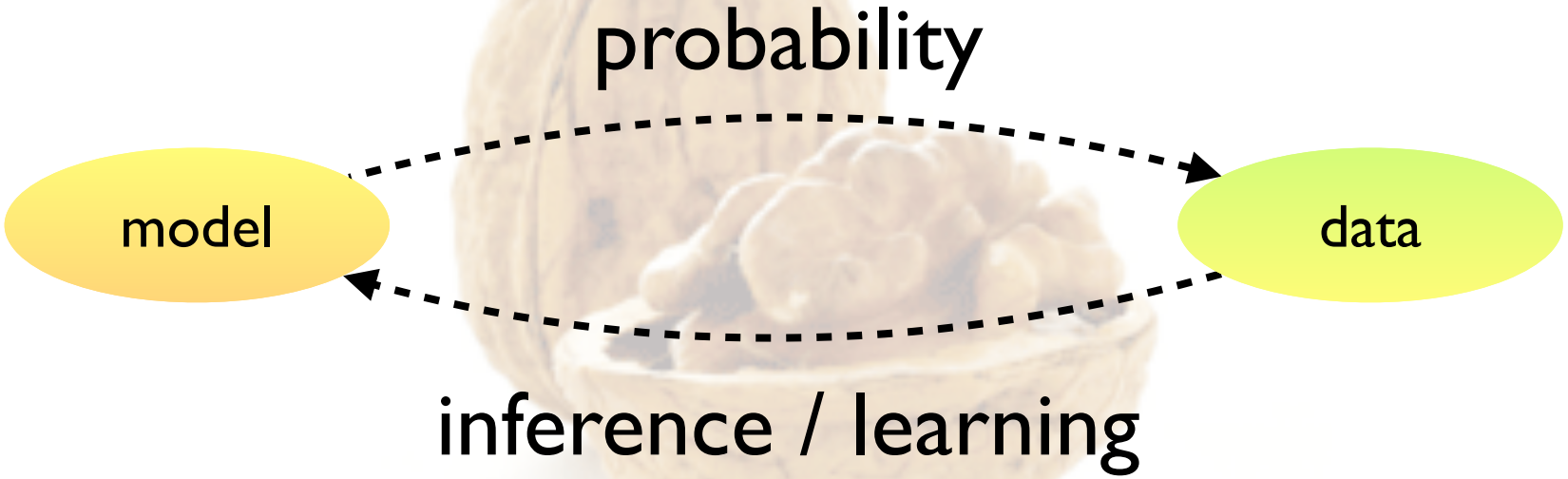
- How can we define these precisely?
  - Variations on model studied in
    - Statistics ("statistical disclosure control")
    - Data mining / database ("privacy-preserving data mining" *)
  - Recent: theoretical CS

Differential Privacy



- Definition of privacy in statistical databases
  - Imposes restrictions on algorithm A generating output
- If A satisfies restrictions, then output provides privacy no matter what user/intruder knows ahead of time
- **Question**: how **useful** are algorithms that satisfy differential privacy?

Statistics in a Nutshell



- This talk:
  - Differentially private alg's for recovering model
  - Performance = convergence to "true" model
- Goals:
  - interesting problems + bridge linguistic gap

This talk: Valid Statistical Inference

- Construct differentially private algorithms with same asymptotic error as best non-private algorithm
  - **Parametric**: for any* parametric model, there exists a private, efficient estimator (i.e. minimal variance)
  - **Nonparametric**: for any* distribution on [0,1], there is a private histogram estimator with same convergence rate as best (non-private) histogram estimator

This talk: Valid Statistical Inference

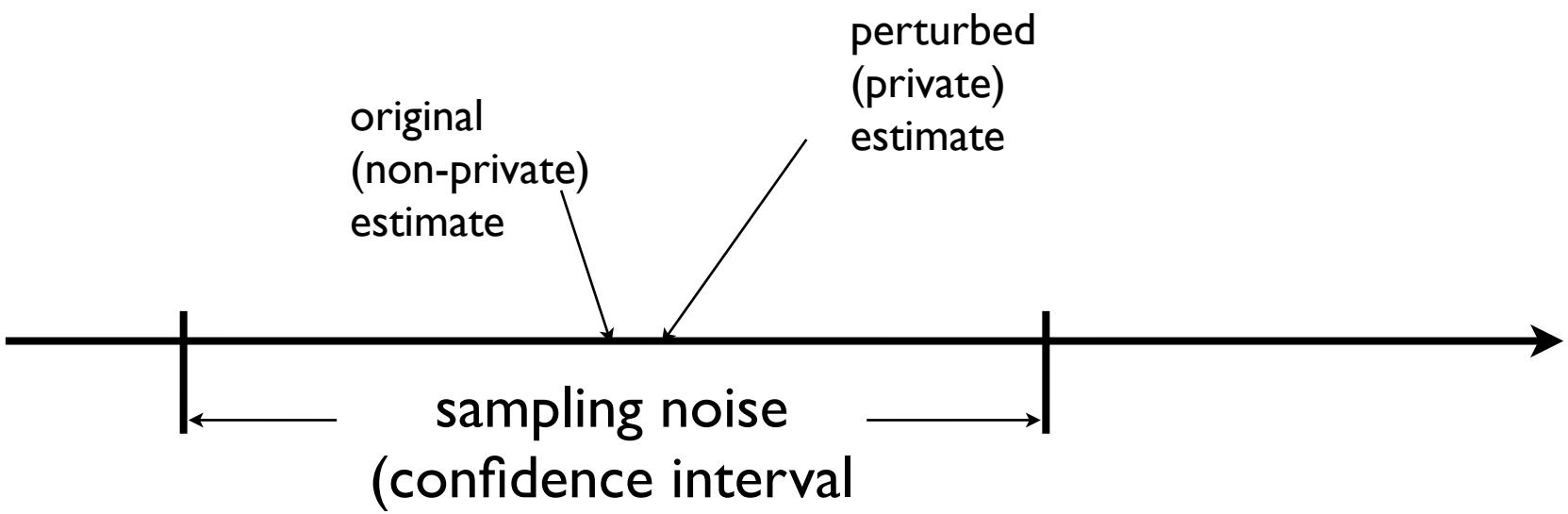
- Construct differentially private algorithms with same asymptotic error as best non-private algorithm
  - **Parametric**: for any* parametric model, there exists a private, efficient estimator (i.e. minimal variance)
  - **Nonparametric**: for any* distribution on [0,1], there is a private histogram estimator with same convergence rate as best (non-private) histogram estimator

Other recent work along these lines:
- Dwork, Lei 2009
- Wasserman, Zhou 2008
- Chaudhuri, Monteleoni 2008
- McSherry, Williams 2009
- ...

Main Idea for both cases

- Add noise to carefully modified estimator
  - Several ways to design differentially private algorithms
  - Adding noise is the simplest

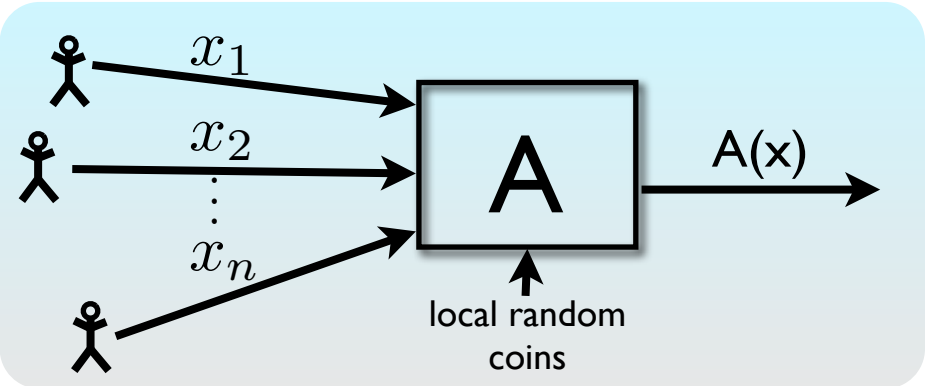
- Prove that required noise is less than inherent variability due to sampling



Reminder: differential privacy

- Intuition:
  - Changes to my data not noticeable by users
  - Output is "independent" of my data

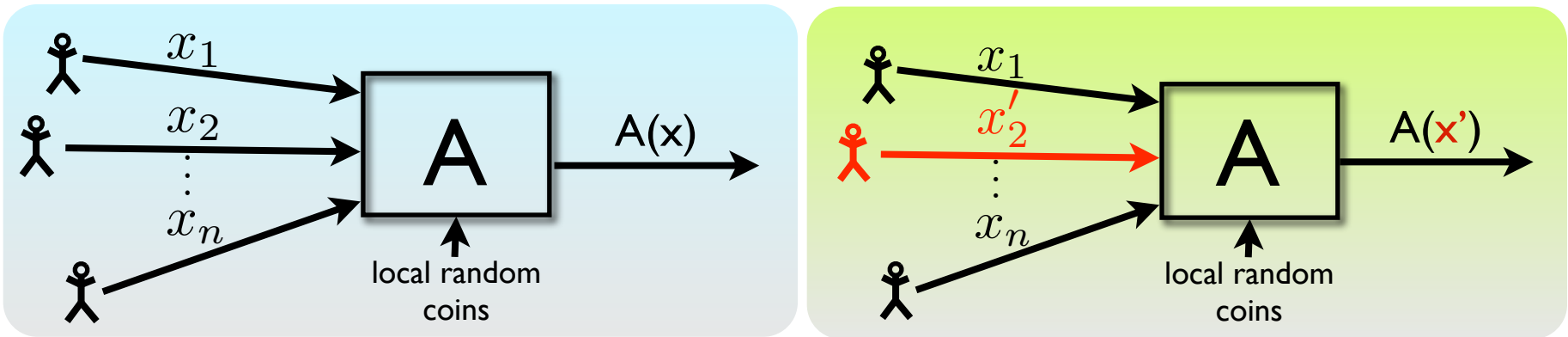
Defining Privacy [DiNi,DwNi,BDMN,DMNS]



- Data set $\mathbf{x} = (x_1, \ldots, x_n) \in D^n$
  - Domain D can be numbers, categories, tax forms
  - Think of x as **fixed** (not random)
- A = **randomized** procedure run by the agency
  - A(x) is a random variable distributed over possible outputs

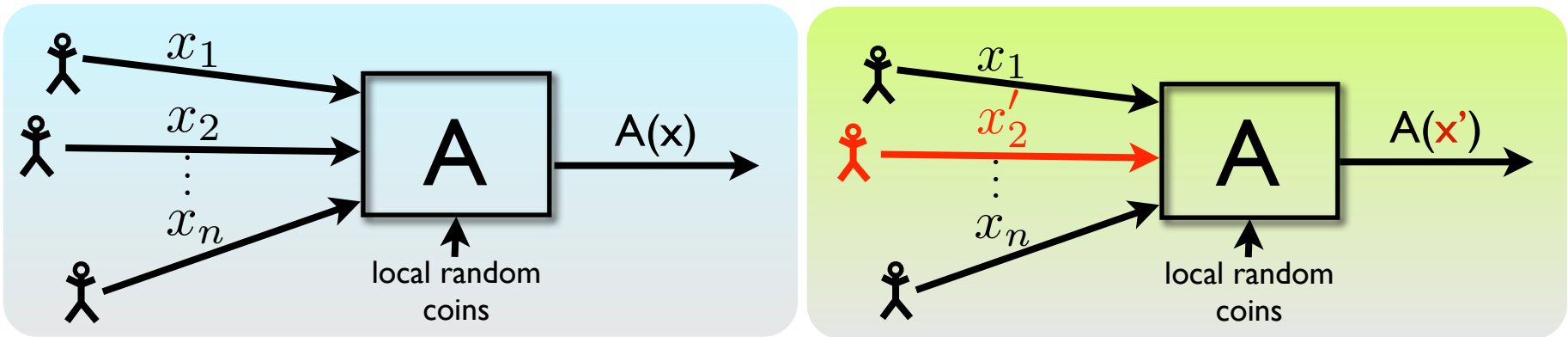
    Randomness might come from adding noise, resampling, etc.

Defining Privacy [DiNi,DwNi,BDMN,DMNS]



x' is a neighbor of x
if they differ in one data point

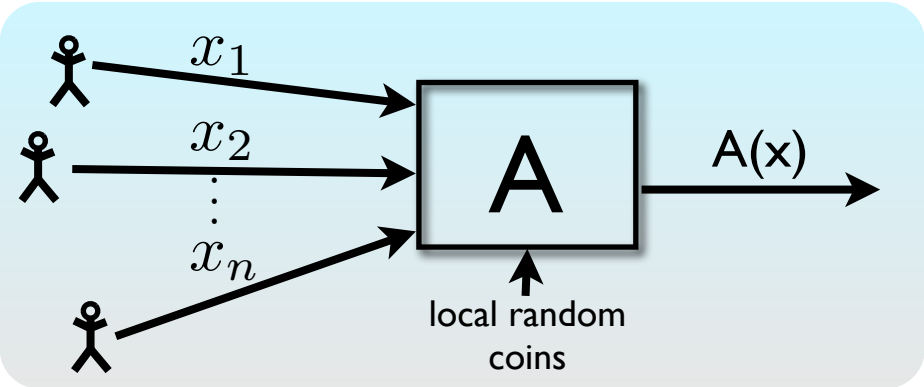
Defining Privacy [DiNi,DwNi,BDMN,DMNS]

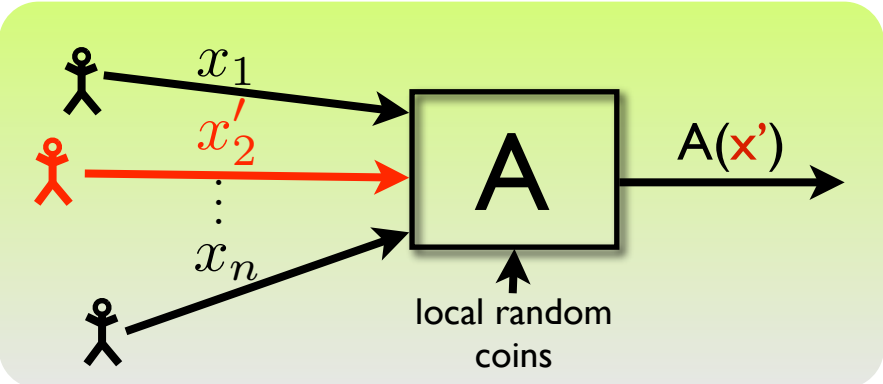


x' is a neighbor of x
if they differ in one data point

Neighboring databases induce **close** distributions on outputs

Defining Privacy [DiNi,DwNi,BDMN,DMNS]





x' is a neighbor of x
if they differ in one data point

Neighboring databases induce **close** distributions on outputs

Definition: A is ε-differentially private if,
for all neighbors x, x',
for all subsets S of outputs

$$\Pr(A(x) \in S) \leq e^{\epsilon} \cdot \Pr(A(x') \in S)$$

Defining Privacy [DiNi,DwNi,BDMN,DMNS]

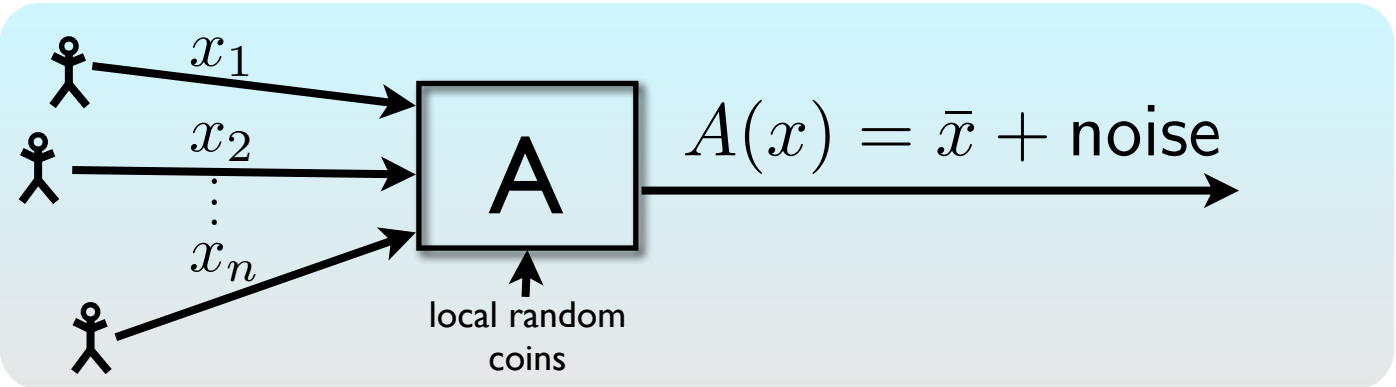
- This is a condition on the **algorithm** (process) A
  - Saying "this output is safe" doesn't take into account how it was computed
- Semantics:
  - no matter what user knows ahead of time, learn the same things about me whether or not my data is present

Neighboring databases induce **close** distributions on outputs

Definition: A is ε-differentially private if,
for all neighbors x, x',
for all subsets S of outputs

$$\Pr(A(x) \in S) \leq e^{\epsilon} \cdot \Pr(A(x') \in S)$$

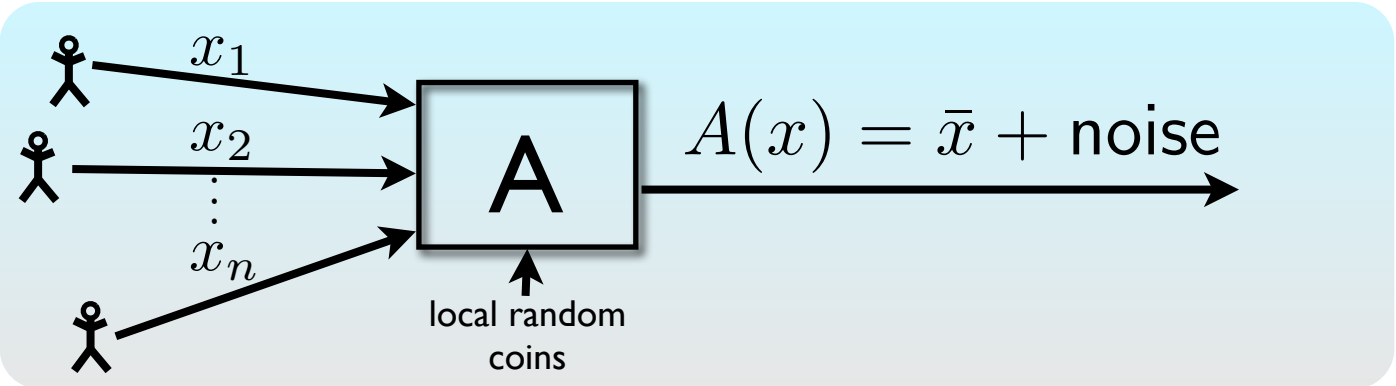
Example: Perturbing the Average



$$x_i \in \{0, 1\}$$

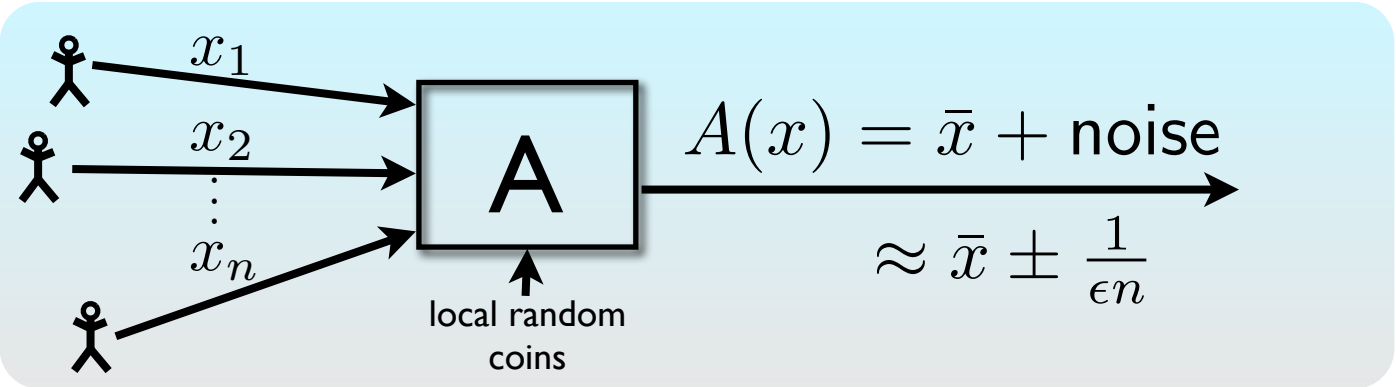
$$\bar{x} = \frac{1}{n} \sum_i x_i$$

Example: Perturbing the Average



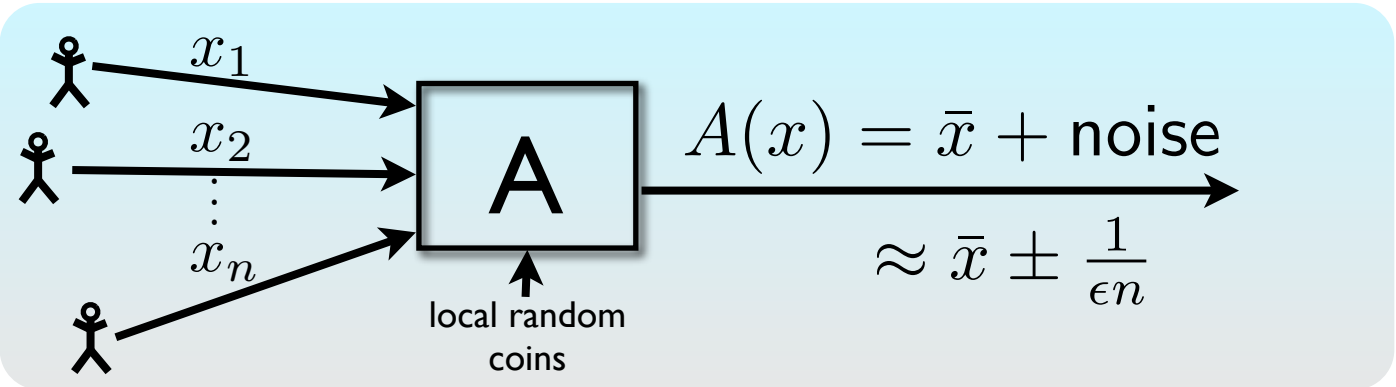
- Data points are binary responses $x_i \in \{0, 1\}$
- Server wants to release average $\bar{x} = \frac{1}{n} \sum_i x_i$

Example: Perturbing the Average



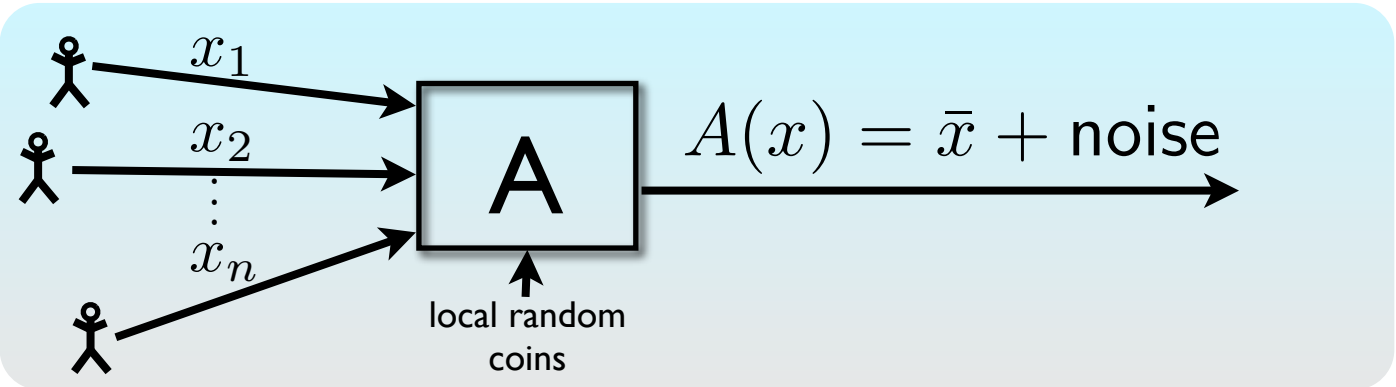
- Data points are binary responses $x_i \in \{0, 1\}$
- Server wants to release average $\bar{x} = \frac{1}{n} \sum_i x_i$
- **Claim**: Can obtain ε-differential privacy with noise $\approx \frac{1}{\epsilon n}$

Example: Perturbing the Average



- Data points are binary responses $x_i \in \{0, 1\}$
- Server wants to release average $\bar{x} = \frac{1}{n} \sum_i x_i$
- **Claim**: Can obtain ε-differential privacy with noise $\approx \frac{1}{\epsilon n}$
- Is this **a lot**?
  - If x is a random sample from a large underlying population, then **sampling noise** $\approx \frac{1}{\sqrt{n}}$
  - Statistical error swamps noise required for privacy

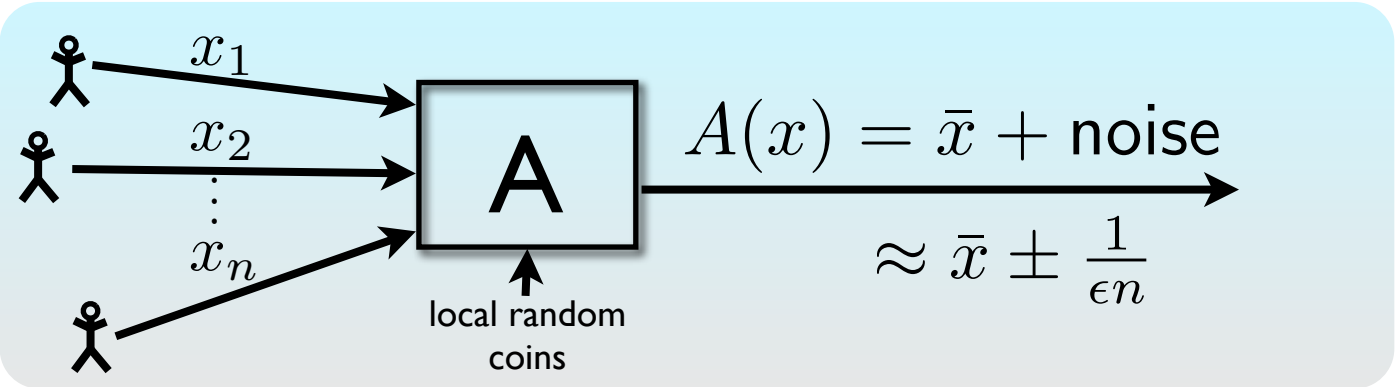
Example: Perturbing the Average



$$x_i \in \{0, 1\}$$

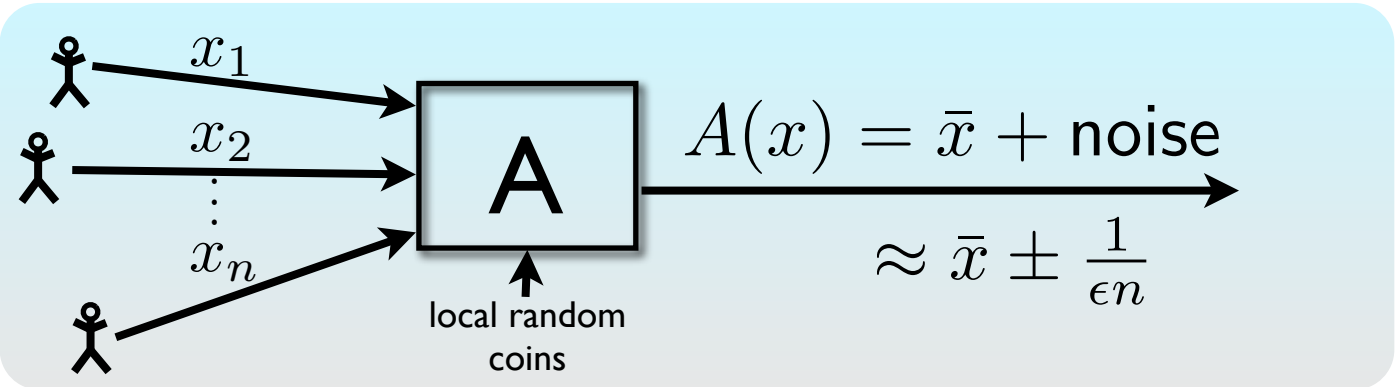
$$\bar{x} = \frac{1}{n} \sum_i x_i$$

Example: Perturbing the Average

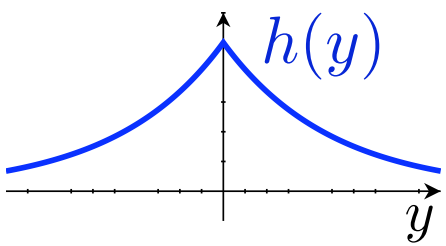


- Data points are binary responses $x_i \in \{0, 1\}$
- Server wants to release average $\bar{x} = \frac{1}{n} \sum_i x_i$
- **Claim**: Can obtain ε-differential privacy with noise $\approx \frac{1}{\epsilon n}$

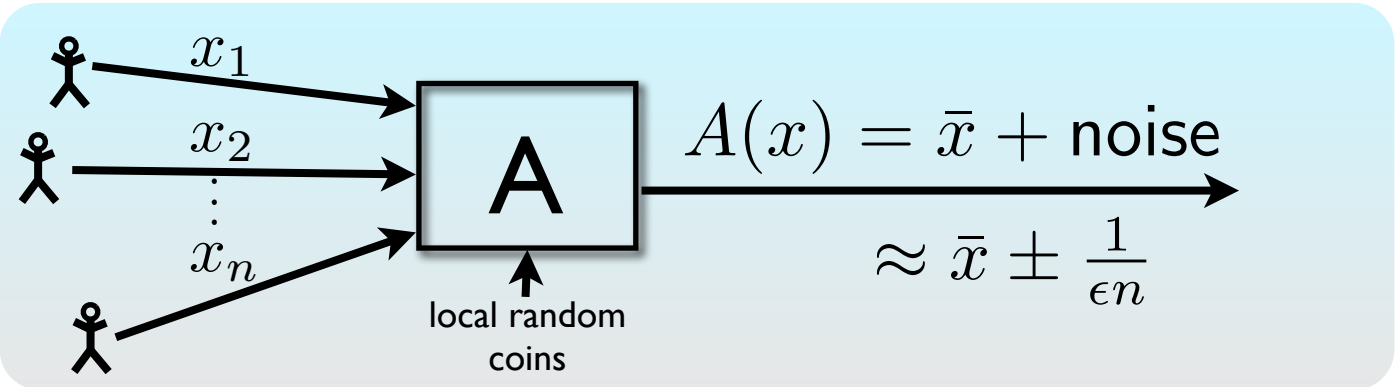
Example: Perturbing the Average



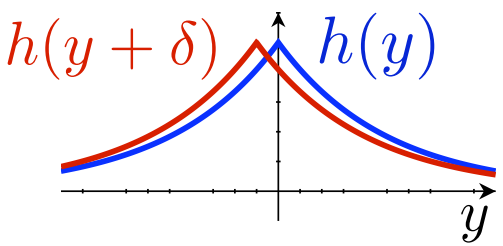
- Data points are binary responses $x_i \in \{0, 1\}$
- Server wants to release average $\bar{x} = \frac{1}{n} \sum_i x_i$
- **Claim**: Can obtain ε-differential privacy with noise $\approx \frac{1}{\epsilon n}$
  - Laplace distribution $\text{Lap}(\lambda)$ has density $h(y) \propto e^{-|y|/\lambda}$
  - Sliding property: $\frac{h(y)}{h(y+\delta)} \leq e^{\delta/\lambda}$



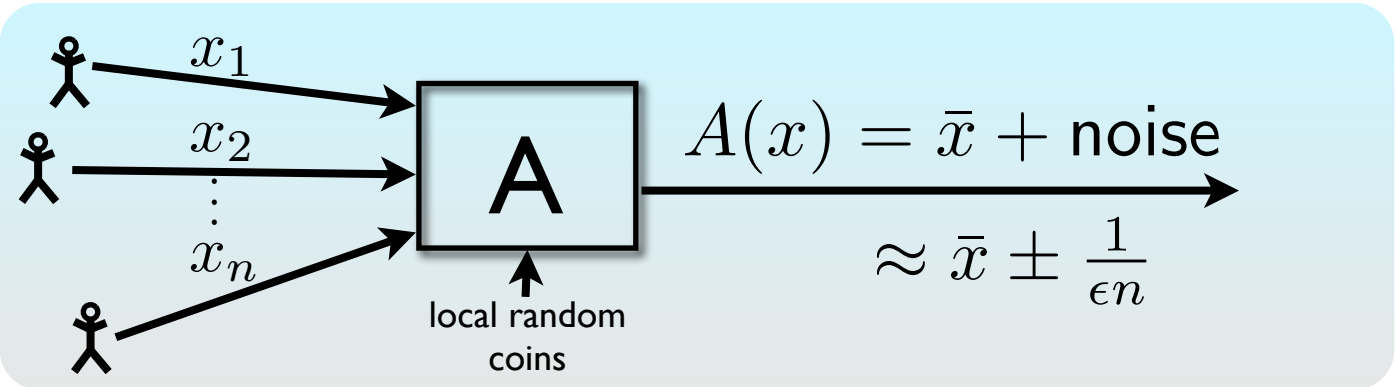
Example: Perturbing the Average



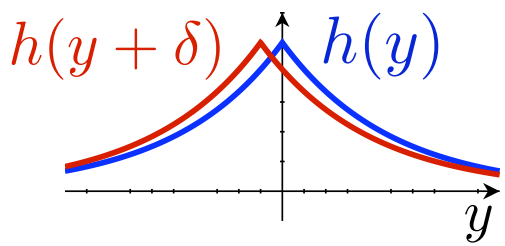
- Data points are binary responses $x_i \in \{0, 1\}$
- Server wants to release average $\bar{x} = \frac{1}{n} \sum_i x_i$
- **Claim**: Can obtain ε-differential privacy with noise $\approx \frac{1}{\epsilon n}$
  - Laplace distribution $\text{Lap}(\lambda)$ has density $h(y) \propto e^{-|y|/\lambda}$
  - Sliding property: $\frac{h(y)}{h(y+\delta)} \leq e^{\delta/\lambda}$



Example: Perturbing the Average

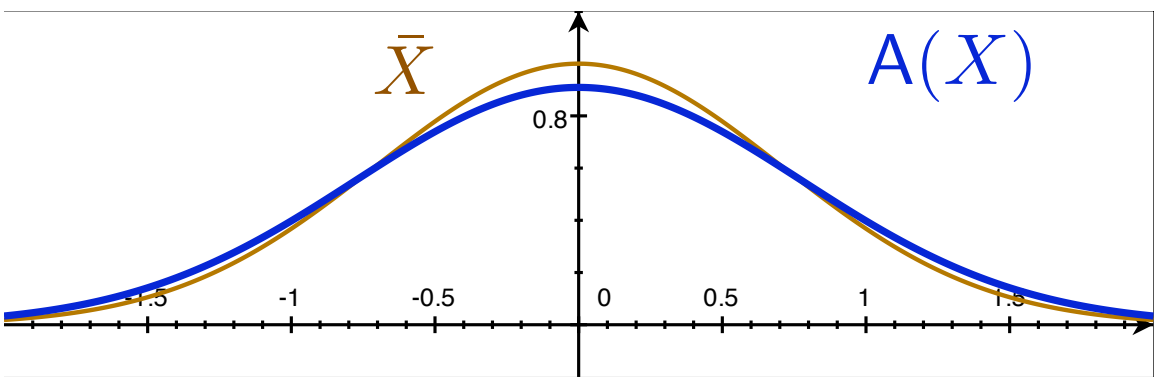


- Data points are binary responses $x_i \in \{0, 1\}$
- Server wants to release average $\bar{x} = \frac{1}{n}\sum_i x_i$
- **Claim**: Can obtain ε-differential privacy with noise $\approx \frac{1}{\epsilon n}$
  - Laplace distribution $\text{Lap}(\lambda)$ has density $h(y) \propto e^{-|y|/\lambda}$
  - Sliding property: $\frac{h(y)}{h(y+\delta)} \leq e^{\delta/\lambda}$
  - A(x) = blue curve, A(x') = red curve
  - $\delta = |\bar{x} - \bar{x}'| \leq \frac{1}{n} \implies \frac{\text{blue curve}}{\text{red curve}} \leq e^{\epsilon}$



When Does Noise **Not** Matter?

- Average: $\mathrm{A}(x) = \bar{x} + \mathrm{Lap}(\frac{1}{\epsilon n})$
  - Suppose $X_1, X_2, X_3, \ldots, X_n$ are i.i.d. random variables
  - $\bar{X}$ is a random variable, and $\sqrt{n} \cdot (\bar{X} - \mu) \xrightarrow{\mathcal{D}} \mathrm{Normal}$
  - $\dfrac{\mathrm{A}(X) - \bar{X}}{\mathrm{StdDev}(\bar{X})} \xrightarrow{P} 0 \quad \text{if } \epsilon \gg \dfrac{1}{\sqrt{n}}$
  - No "cost" to privacy:
    - A(X) is "as good as" $\bar{X}$ for statistical inference*



When Does Noise **Not** Matter?

When Does Noise **Not** Matter?

- Mean example generalizes to other statistics
- **Theorem:** For any* exponential family, can release "approximately sufficient" statistics
  - Suff. stats T(X) are sums, add noise $\frac{d}{\epsilon n}$ for dimension d
  - $\frac{\mathrm{A}(X) - T(X)}{\mathrm{StdDev}(T(X))} \xrightarrow{P} 0$

When Does Noise **Not** Matter?

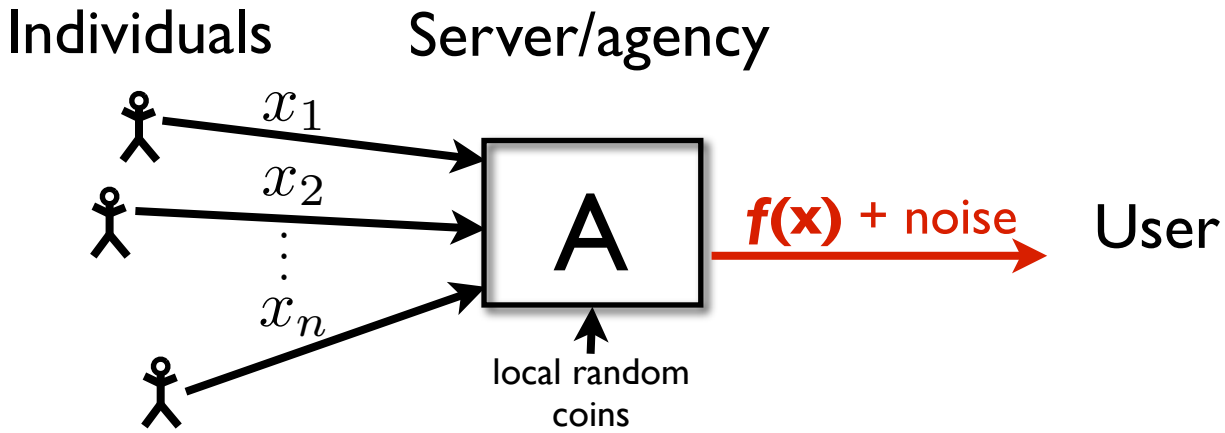
- Mean example generalizes to other statistics
- **Theorem:** For any* exponential family, can release "approximately sufficient" statistics
  - Suff. stats T(X) are sums, add noise $\frac{d}{\epsilon n}$ for dimension d
  - $\frac{\mathrm{A}(X) - T(X)}{\mathrm{StdDev}(T(X))} \xrightarrow{P} 0$
- Asymptotic result: Indicates that useful analysis possible
  - Requires more sophisticated processing for small n

When Does Noise **Not** Matter?

- Mean example generalizes to other statistics
- **Theorem:** For any* exponential family, can release "approximately sufficient" statistics
  - Suff. stats T(X) are sums, add noise $\frac{d}{\epsilon n}$ for dimension d
  - $\frac{\mathrm{A}(X) - T(X)}{\mathrm{StdDev}(T(X))} \xrightarrow{P} 0$
- Asymptotic result: Indicates that useful analysis possible
  - Requires more sophisticated processing for small n
- Noise degrades with dimension
  - More information $\Rightarrow$ less privacy
  - Research question: Is this necessary?

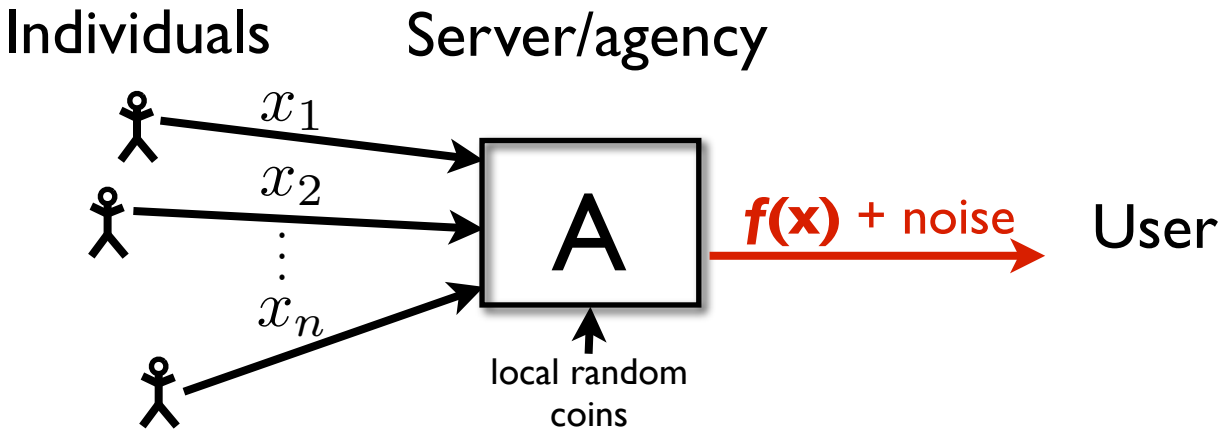
- Histogram Density Estimation
  - Calibrating noise to sensitivity
- Maximum Likelihood Estimator
  - Sub-sample and aggregate

Global Sensitivity [DMNS06]



- Intuition: ***f*(x)** can be released accurately when ***f*** is insensitive to individual entries $x_1, x_2, \ldots, x_n$
- Global Sensitivity: $$GS_f = \max_{\text{neighbors } x,x'} \|f(x) - f(x')\|_1$$
- Example: $GS_{\text{average}} = \frac{1}{n}$

Global Sensitivity [DMNS06]



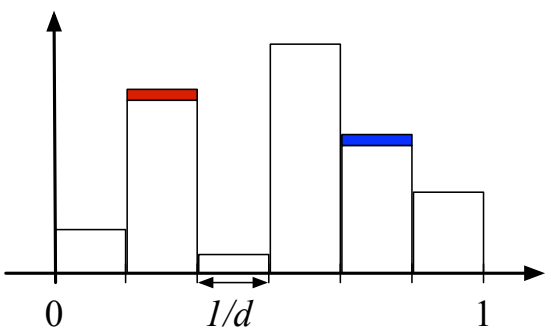
- Intuition: ***f*(x)** can be released accurately when ***f*** is insensitive to individual entries $x_1, x_2, \ldots, x_n$
- Global Sensitivity: $\mathrm{GS}_f = \max_{\text{neighbors } x,x'} \|f(x) - f(x')\|_1$
- Example: $\mathrm{GS}_{\text{average}} = \frac{1}{n}$

Theorem: If $\mathrm{A}(x) = f(x) + \mathrm{Lap}\left(\frac{\mathrm{GS}_f}{\epsilon}\right)$, then A is ϵ-differentially private.

Example: Histograms

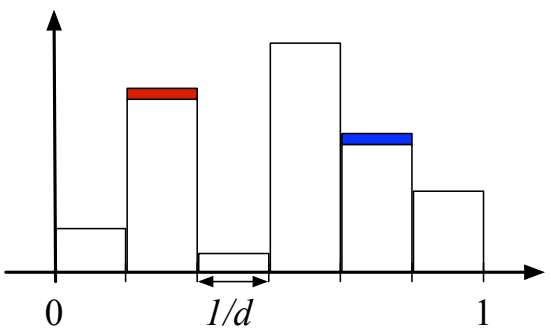
$f(x) = (n_1, n_2, ..., , n_d)$ where $n_j = \#\{i : x_i$ in j-th interval$\}$

$\mathsf{Lap}(1/\epsilon)$



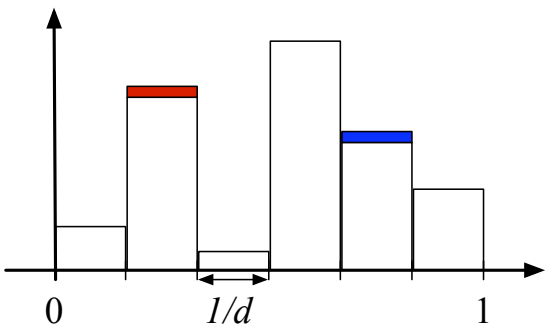
Example: Histograms

- Say $x_1, x_2, \ldots, x_n$ in [0,1]
  - Partition [0,1] into d intervals of equal size
  - $f(x) = (n_1, n_2, \ldots, , n_d)$ where $n_j = \#\{i : x_i \text{ in } j\text{-th interval}\}$
  - $GS_f = 2$
  - Sufficient to add noise $\text{Lap}(1/\epsilon)$ to each count
    - Independent of the dimension



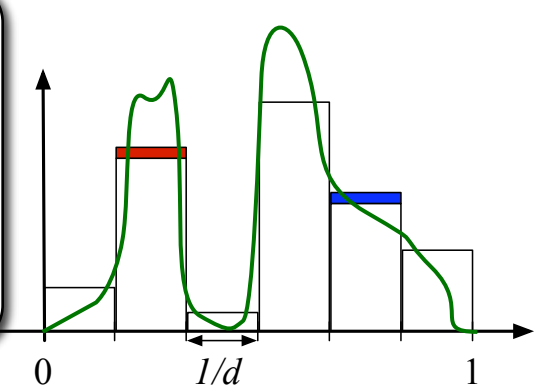
Example: Histograms

- Say $x_1, x_2, \ldots, x_n$ in ~~[0,1]~~ arbitrary domain D
  - Partition ~~[0,1] into d intervals of equal size~~ into d disjoint "bins"
  - $f(x) = (n_1, n_2, \ldots, , n_d)$ where $n_j = \#\{i : x_i$ in j-th ~~interval~~ bin
  - $GS_f = 2$
  - Sufficient to add noise $\mathrm{Lap}(1/\epsilon)$ to each count
    - Independent of the dimension



Example: Histograms

- Say $x_1, x_2, \ldots, x_n$ in [0,1]
  - Partition [0,1] into d intervals of equal size
  - $f(x) = (n_1, n_2, \ldots, , n_d)$ where $n_j = \#\{i : x_i \text{ in } j\text{-th interval}\}$
  - $GS_f = 2$
  - Sufficient to add noise $\mathrm{Lap}(1/\epsilon)$ to each count
    - Independent of the dimension
- For any smooth density h, if X_i i.i.d. ~ h, noisy histogram converges to h
  - Expected L_2 error $O(\frac{1}{\sqrt[3]{n}})$ if $n \gg \frac{1}{\epsilon^3}$
  - Same as "best" non-private estimator



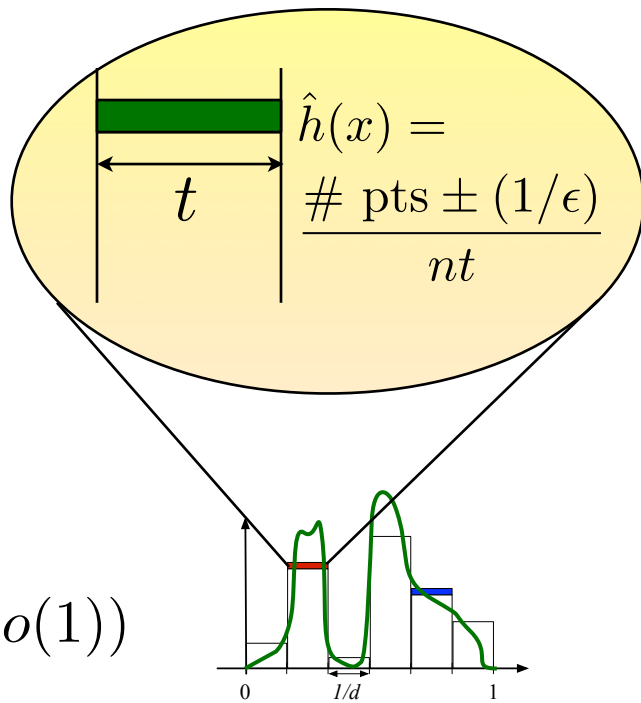
Proof

- L_2 error: $IMSE = \mathbb{E}_{\hat{h}} \left\{ \|h - \hat{h}\|_2^2 \right\} = \int_0^1 (h - \hat{h})^2 dx$
- If bins have fixed width t
  - Theorem [Scott]: $IMSE = \frac{1}{nt} + t^2 R(h') + \frac{1}{n}$
  - Minimized for $t = \sqrt[3]{n}$
- Additional square error due to noise
  - Additional error per bin $t \times \frac{1}{t^2 n^2 \epsilon^2}$
  - Total additional error:

$$\underbrace{1/t}_{\#\text{ bins}} \times \underbrace{1/(tn^2\epsilon^2)}_{\text{error per bin}} = \frac{1}{t^2 n^2 \epsilon^2}$$

  - Total error (by additivity of variance):

$$\frac{1}{nt} + t^2 + \frac{1}{t^2 n^2 \epsilon^2} = (\text{original error}) \times (1 + o(1))$$

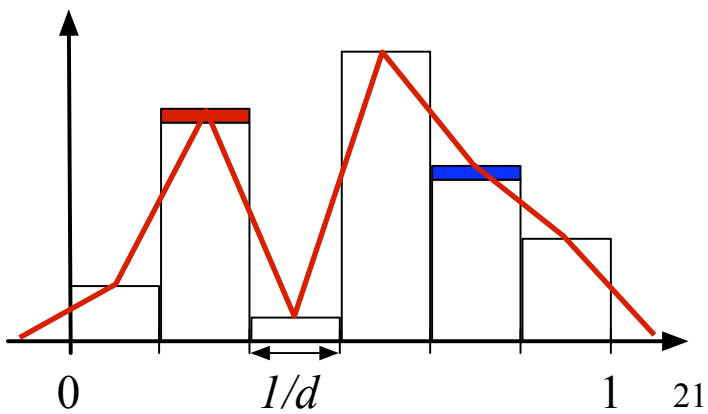


Frequency Polygon

- Frequency polygon
  - linear interpolation through midpoints of histogram

- For any smooth density h, if X_i i.i.d. ~ h, frequency polygon of noisy histogram converges to h
  - Expected L_2 error $O(n^{-4/5})$ if $n \gg \frac{1}{\epsilon^{5/2}}$
  - Same as non-private estimator

- Similar, more complicated proof



More detail

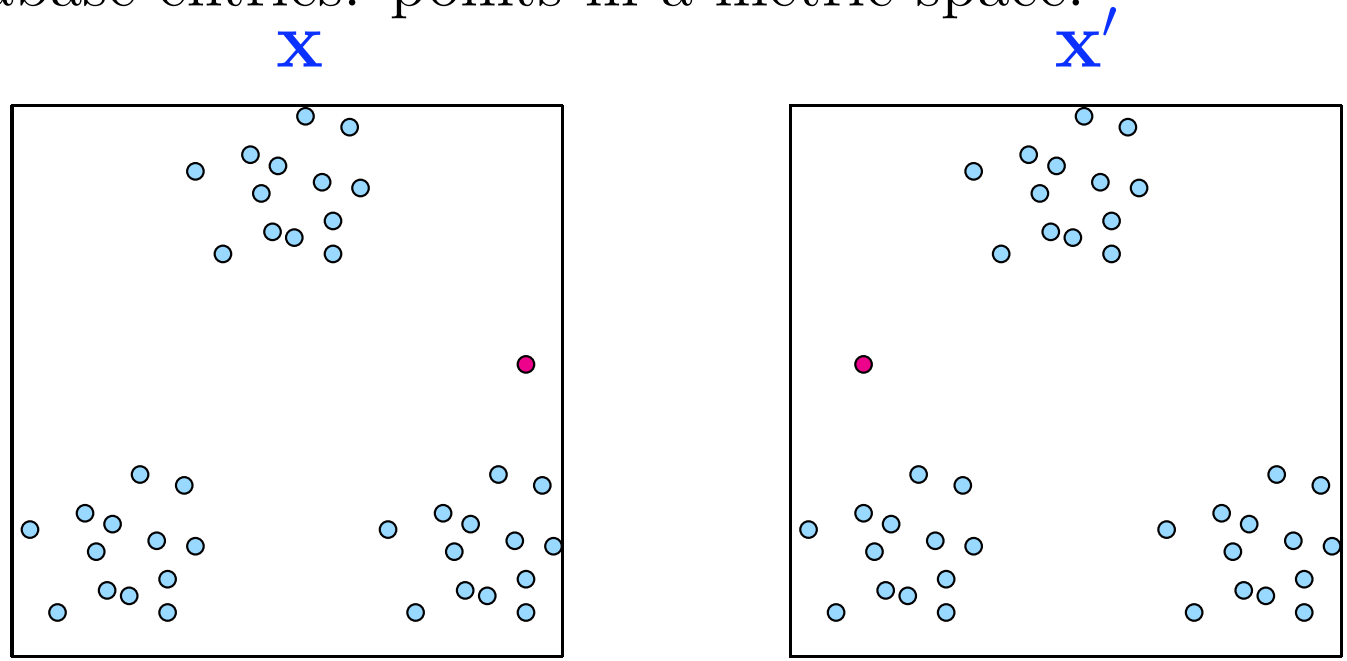
- This actually shows that for any given bin width, can find noisy estimator that is close to non-noisy estimator

- Does not address how to choose exact bin width
  - Subject to extensive research
  - Common "bandwidth selection" criteria can be approximated privately:
    - compute many different candidates
    - use exponential mechanism to select best candidate(s)

- Histogram Density Estimation
  - Calibrating noise to sensitivity
- Maximum Likelihood Estimator
  - Sub-sample and aggregate

High Global Sensitivity: Learning Mixtures

Database entries: points in a metric space.

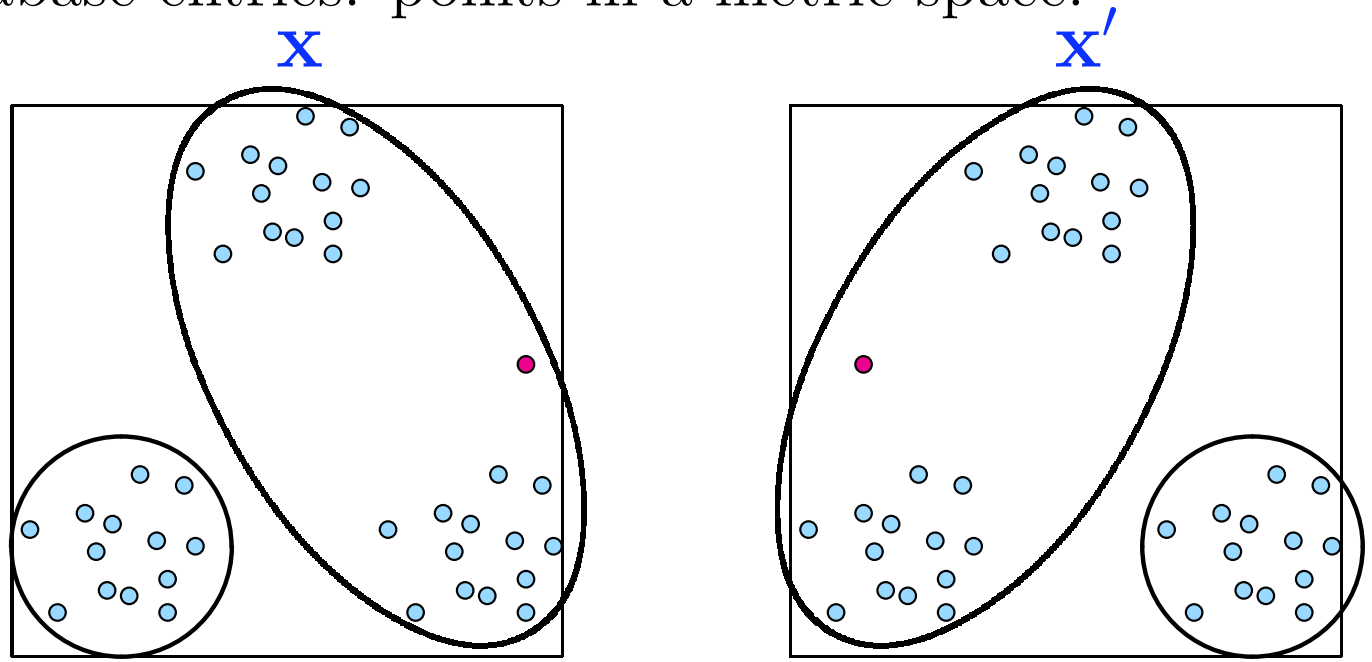


Global sensitivity of cluster centers is roughly the diameter of the space.

- But intuitively, if clustering is "good", cluster centers should be insensitive.

High Global Sensitivity: Learning Mixtures

Database entries: points in a metric space.

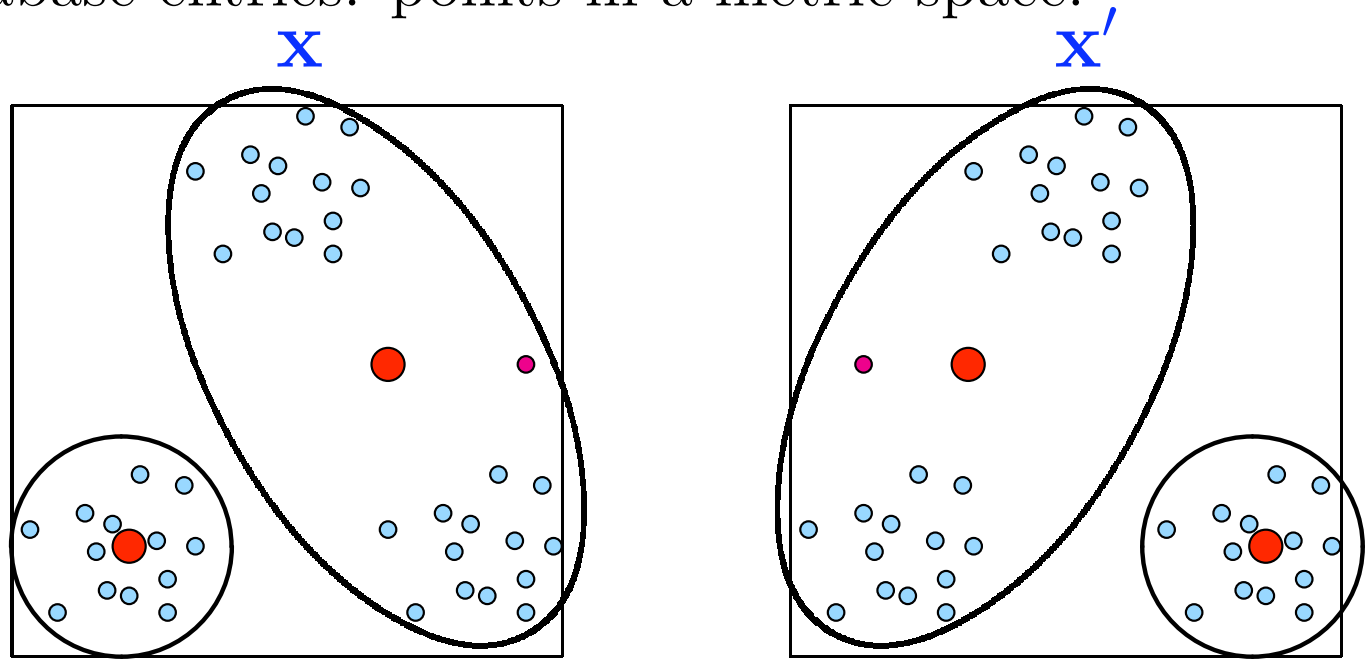


Global sensitivity of cluster centers is roughly the diameter of the space.

- But intuitively, if clustering is "good", cluster centers should be insensitive.

High Global Sensitivity: Learning Mixtures

Database entries: points in a metric space.



Global sensitivity of cluster centers is roughly the diameter of the space.

- But intuitively, if clustering is "good", cluster centers should be insensitive.

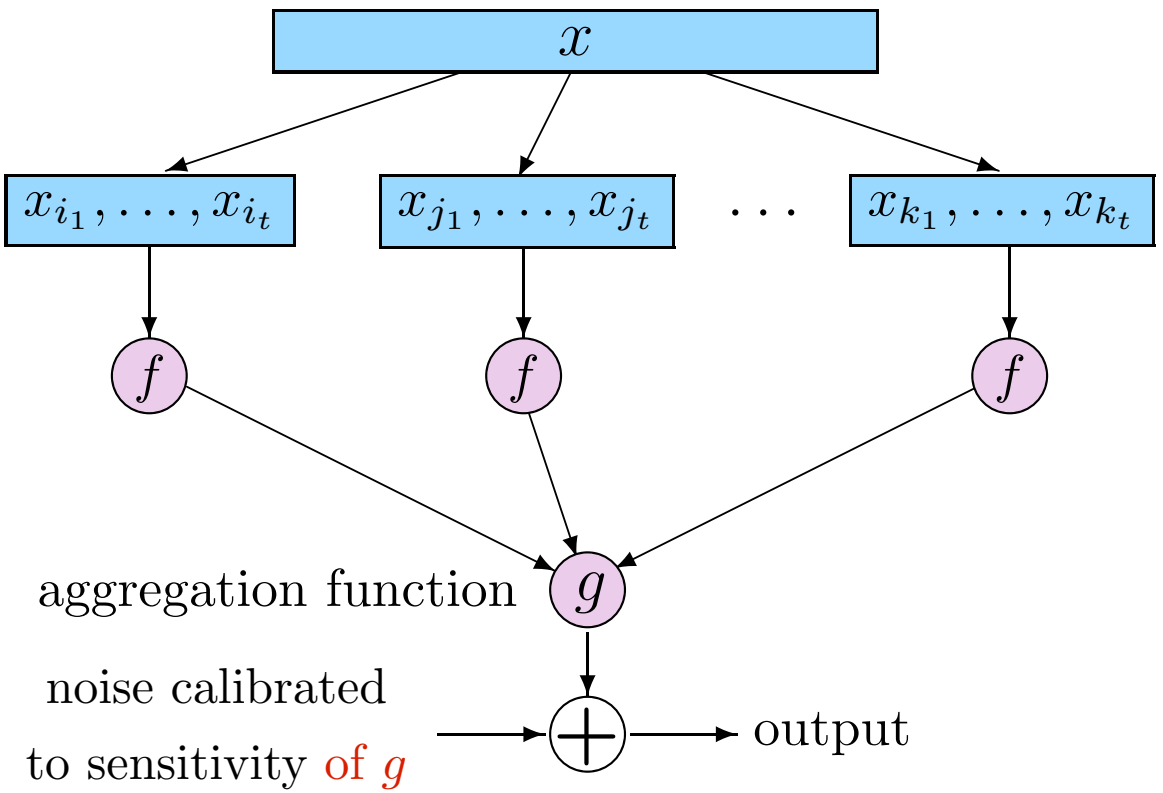
What about maximum likelihood estimates?

- Sometimes MLE is well-behaved,
  - e.g. observed proportion for binomial
- Sometimes we have no idea
  - e.g. no closed form expression for most loglinear models
    - (loglinear = popular class of statistical models for categorical data)
  - Can have arbitrarily bad sensitivity

Sample-and-Aggregate Methodology [NRS]

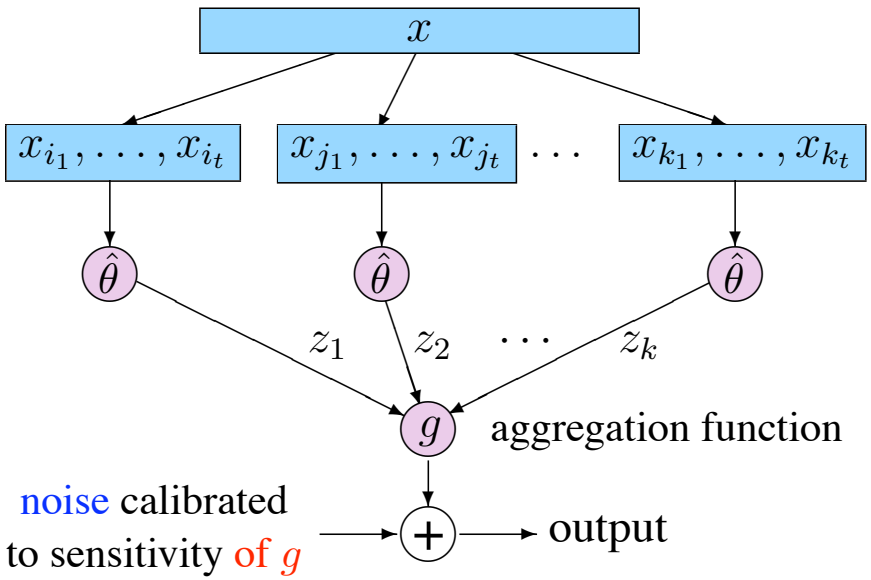
Intuition: Replace f with a less sensitive function $\tilde{f}$.

$$\tilde{f}(x) = g(f(sample_1), f(sample_2), \ldots, f(sample_k))$$



Example: Efficient Point Estimates

- Given a parametric model $\{f_\theta : \theta \in \Theta\}$
- MLE $= \text{argmax}_\theta(f_\theta(x))$
- Converges to Normal
  - Var(MLE) = $(I_F(\theta)n)^{-1}$
  - Can be corrected so that bias($\hat{\theta}$) = $O(n^{-3/2})$

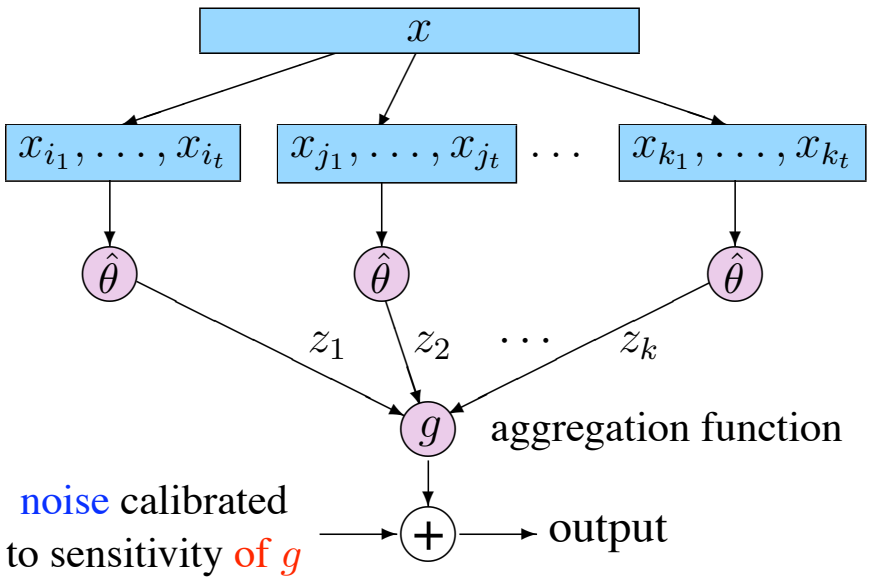


- **Theorem**: If model is well-behaved, then sample-aggregate using $\hat{\theta}$ gives efficient estimator if $n = \omega(1/\epsilon^6)$

- Basic version: Average MLE from different samples
  - If parameter space is bounded, sensitivity of average is O(1/k)

Proof idea

- Each MLE ≈ N((k/n)$^{3/2}$, k/I(n))
- Output ≈ Gaussian since each MLE ≈ Gaussian
- Variance of average = (k/n) / k = 1/n
- Bias of average = bias of each term = (k/n)$^{3/2}$
- Total squared error = variance + bias^2 + E(noise^2)



$$= \frac{1}{n} + (\frac{k}{n})^3 + (\frac{1}{k\epsilon})^2 \quad = \quad \frac{1+o(1)}{n} \qquad \text{when } n = \omega(1/\epsilon^6) \text{ (set } k = o(n^{2/3}))$$

Higher Dimension

- Theorem holds in any fixed dimension
  - actually up to poly(n)
- Higher dimension requires significant extra work
  - need to find an ellipse-shaped "envelope" for data set that captures 1-o(1) fraction of points in set...
  - ... differentially privately
  - Use Dunagan-Vempala '01
- Open: understanding exact dependence on dimension

Conclusions

- Define privacy in terms of my effect on output
  - Meaningful despite arbitrary external information
  - I should participate if I get benefit
- What can we compute with rigorous guarantees?
  - This talk: statistical estimators that are "as good" as optimal non-private estimators
    - First step: basic asymptotic theory
    - New aspect to "curse" of dimensionality
  - Future / ongoing work:
    - More sophisticated analyses: linear/logistic regression, kernel density estimates, etc
    - Multiple analyses (see Blum-Ligett-Roth)