

Pinning Down “Privacy” in Statistical Databases

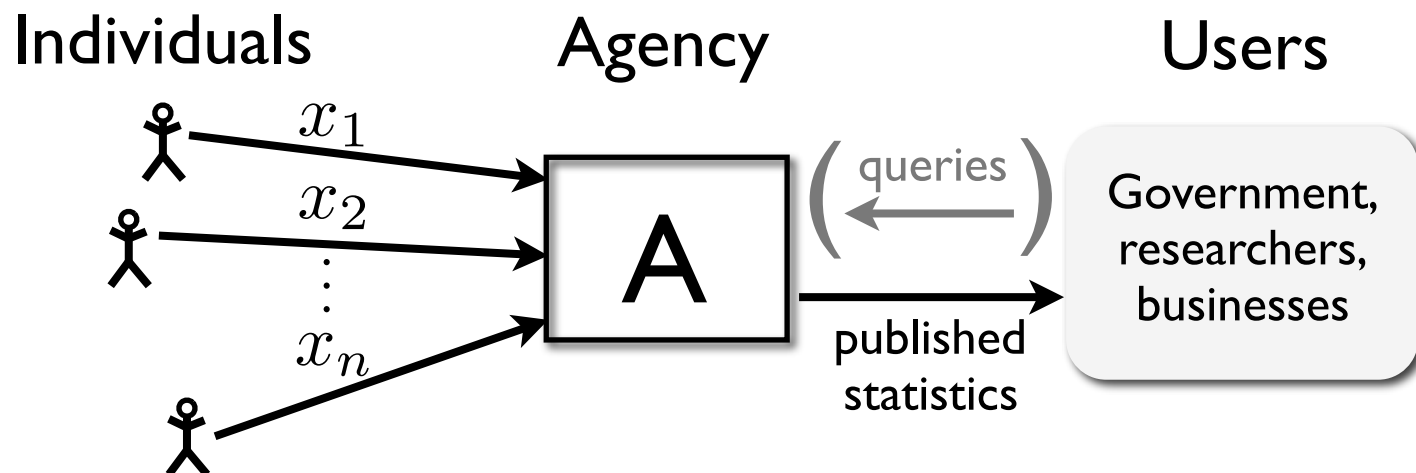
PENNSSTATE



Adam Smith

Computer Science & Engineering Department
Pennsylvania State University

Shared Statistical Data



Large collections of personal information

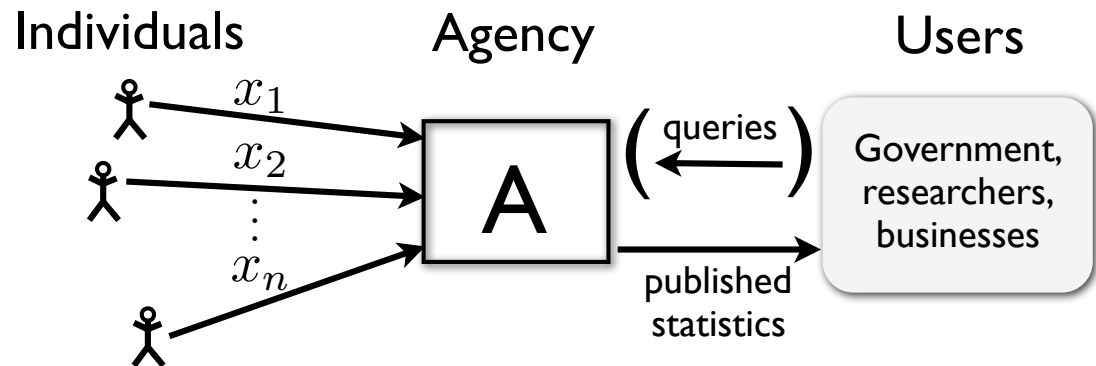
- census data
- medical/public health data
- social networks
- recommendation systems
- trace data: search records, etc
- intrusion-detection systems, botnet research, ...

Recently:

- larger data sets
- more types of data

Privacy in Statistical Databases

- **One** context



- Two conflicting goals
 - **Utility**: Users can extract “global” properties
 - **Confidentiality**: Individual information stays hidden
- (How) can we distinguish “global” from “individual” information?

Privacy in Statistical Databases

Privacy in Statistical Databases

- Variations on model studied in
 - **Statistics** (“statistical disclosure control”)
 - **Data mining / database** (“privacy-preserving data mining” *)

Privacy in Statistical Databases

- Variations on model studied in
 - **Statistics** (“statistical disclosure control”)
 - **Data mining / database** (“privacy-preserving data mining” *)
- No coherent theory and many failures

Privacy in Statistical Databases

- Variations on model studied in
 - **Statistics** (“statistical disclosure control”)
 - **Data mining / database** (“privacy-preserving data mining” *)
- No coherent theory and many failures
- **This project: rigorous foundations, coherent field**
 - new attacks (“cryptanalysis” for statistical data)
 - precise definitions
 - provably private algorithms/protocols

what is this **not**?

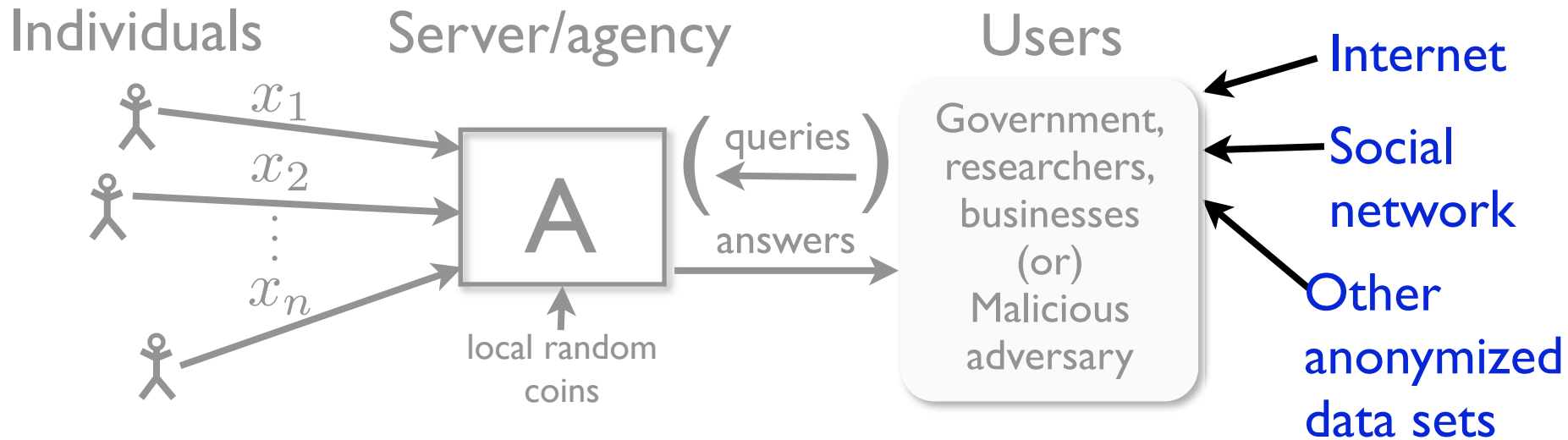
... **not** cryptography as usual

- Normally in crypto and access control:
clear boundary between **inside** and **outside**
 - Model: psychiatrist-patient records
 - Goal: technical enforcement of boundary
- Here: fluid boundary
 - Model: census publication
 - Goal: publish as much “global” information as possible

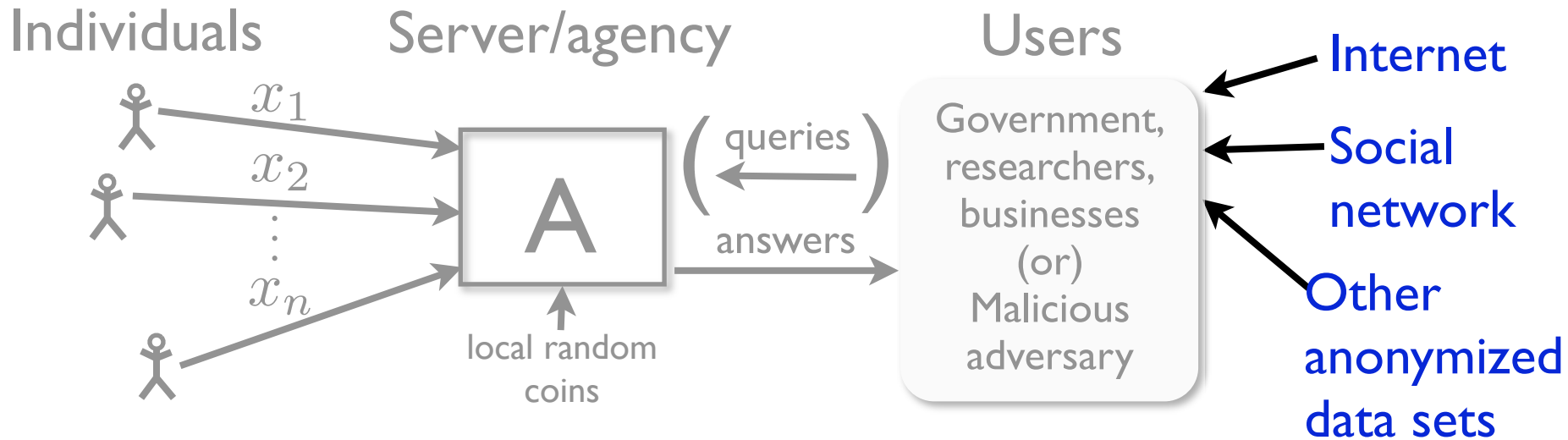
What makes this problem hard?

- External information
- Data are not atomic
- Related data appear in many places

Challenge: External Information

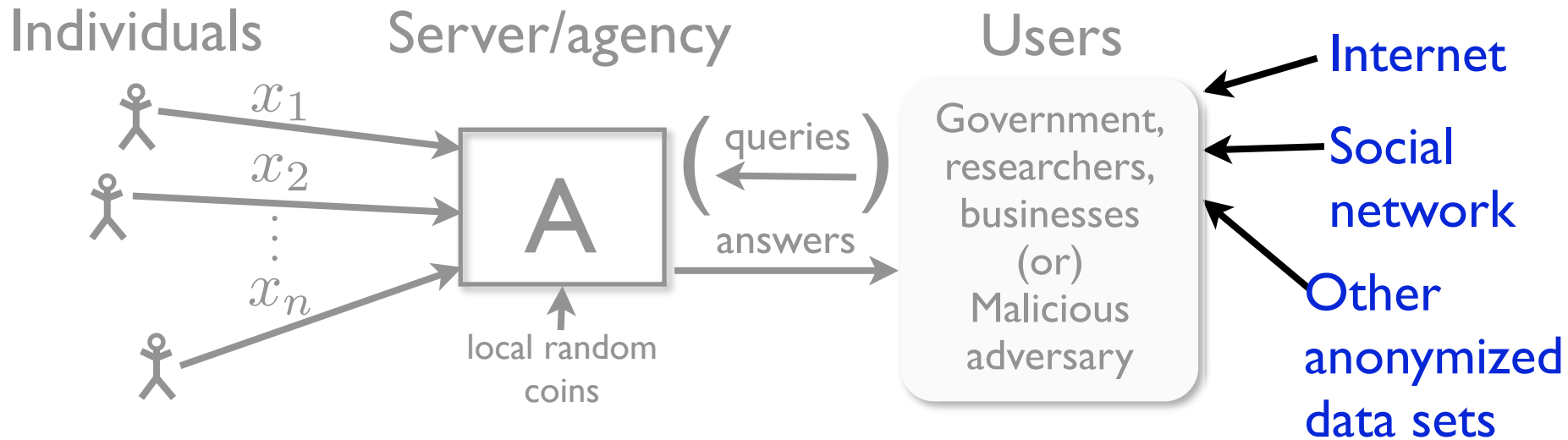


Challenge: External Information



- Users have external information sources
 - Can't assume we know the sources
















Challenge: External Information



- Users have external information sources
 - Can't assume we know the sources
- Anonymization schemes regularly broken, e.g.
 - Netflix via IMDB [Narayanan-Shmatikov]
 - “Composition attacks” [GKS]

Netflix Data Release [Narayanan-Shmatikov 2008]

- Ratings for subset of movies and users
- Usernames replaced with random IDs
- Some additional perturbation

	Item 1	Item 2			Item M	
User 1						
User 2						
						
						
						
User N						

Use Public Reviews from IMDB.com

👍		👎	👍		
	👍				
👍		👎		👍	👍
👍			👎		
	👍		👎	👎	
		👎	👍		

+

👍			👍		
	👍				
👍					👍
👍			👎		
				👎	
		👎			

Alice
Bob
Charlie
Danielle
Erica
Frank

Anonymized
Netflix data

Public, incomplete
IMDB data

=

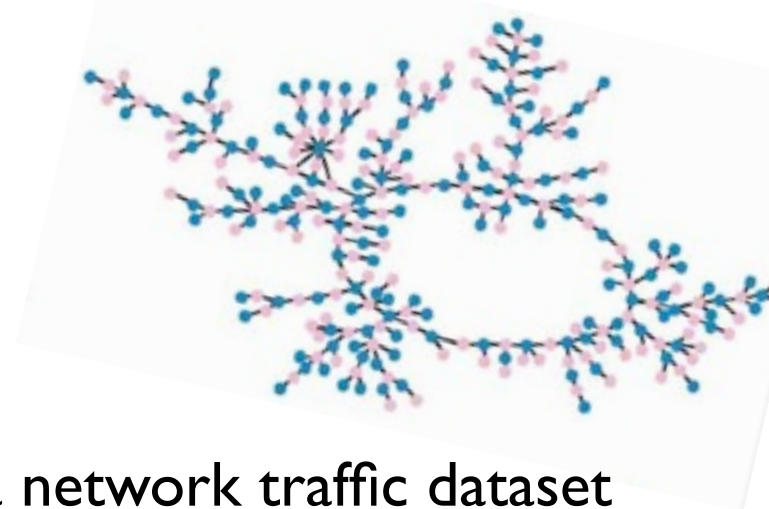
👍		👎	👍		
	👍				
👍		👎		👍	👍
👍			👎		
	👍		👎	👎	
		👎	👍		

~~Alice~~
~~Bob~~
~~Charlie~~
~~Danielle~~
~~Erica~~
~~Frank~~

Identified Netflix Data

Some other attacks...

- Web search:
 - AOL release [New York Times, 2006]
 - De-tokenizing search [Kumar et al, 2007]
- Social networks [Backstrom-Dwork-Kleinberg'07, NarayananS'09]
 - Reidentifying individuals in a social network with node labels removed
- Computer networks
[Coull et al., NDSS '07, Ribeiro et al, NDSS '08]
 - Relabeling computers and users in a network traffic dataset
- Genetic data [Homer et al. PLoS Genetics '08]
 - Identifying individuals in aggregate genomic data



What makes this problem hard?

- External information
- Data are not atomic
- Related data appear in many places

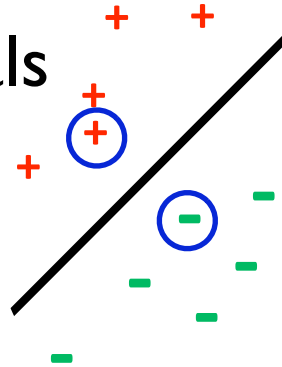
Data are not atomic

Data are not atomic

- Myth: PII, sensitive, nonsensitive information
 - geographic location + d.o.b. predict SSN

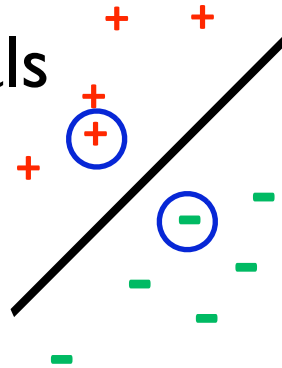
Data are not atomic

- Myth: PII, sensitive, nonsensitive information
 - geographic location + d.o.b. predict SSN
- Some aggregate statistics leak **lots** about individuals
 - **Support vector machine** described by subset of data

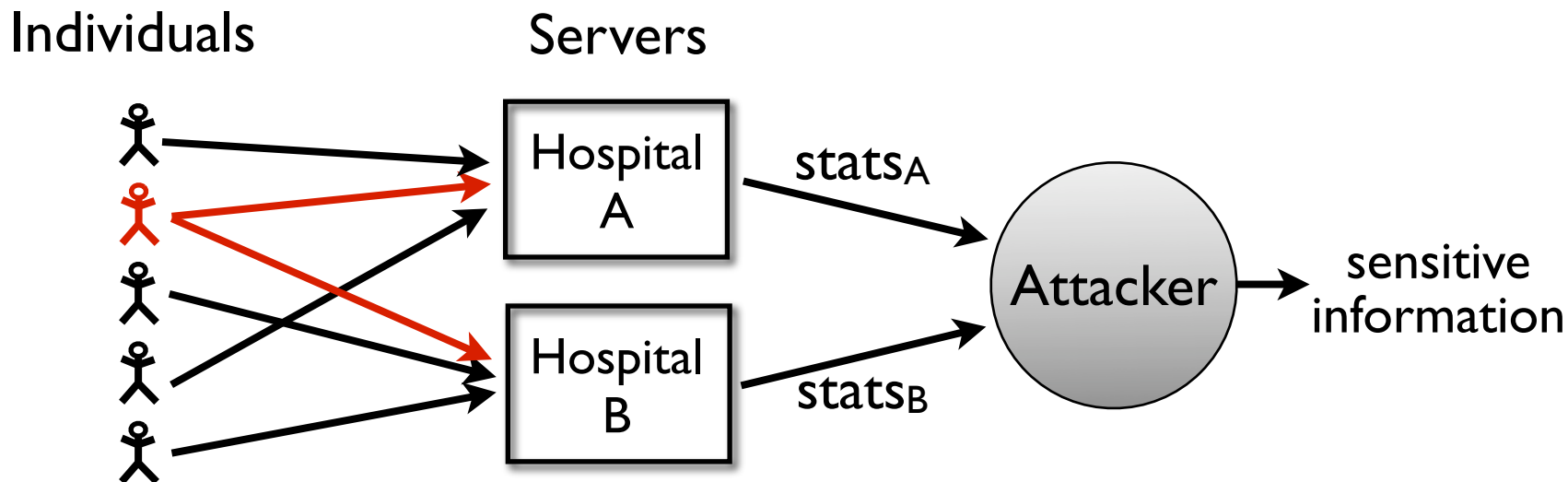


Data are not atomic

- Myth: PII, sensitive, nonsensitive information
 - geographic location + d.o.b. predict SSN
- Some aggregate statistics leak **lots** about individuals
 - **Support vector machine** described by subset of data
- Every aggregate statistic leaks **something**
 - “Innocuous” pieces of information can be combined
 - How many diabetics in Manhattan?
 - How many diabetics in Manhattan not named “Adam Davison Smith”?
 - Even unrelated questions cause problems [Dinur-Nissim, *PODS* '03]
 - Answering **too many** questions **too precisely** reveals almost the whole data set

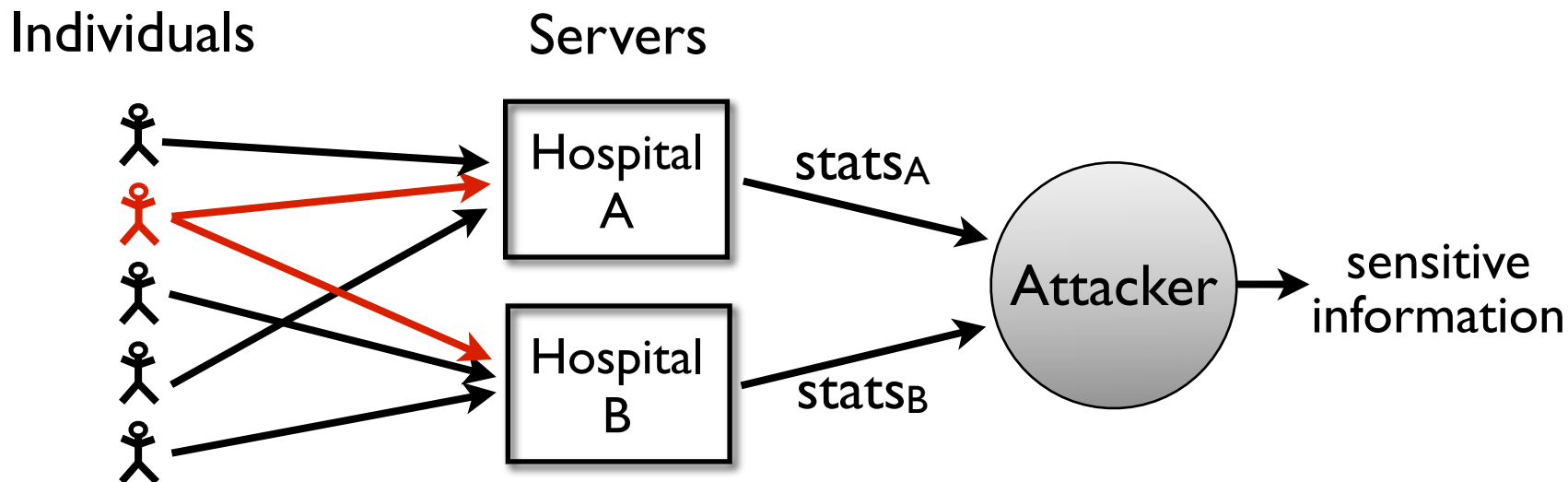


“Composition” Attacks [Ganta-Kasiviswanathan-Smith, KDD 2008]



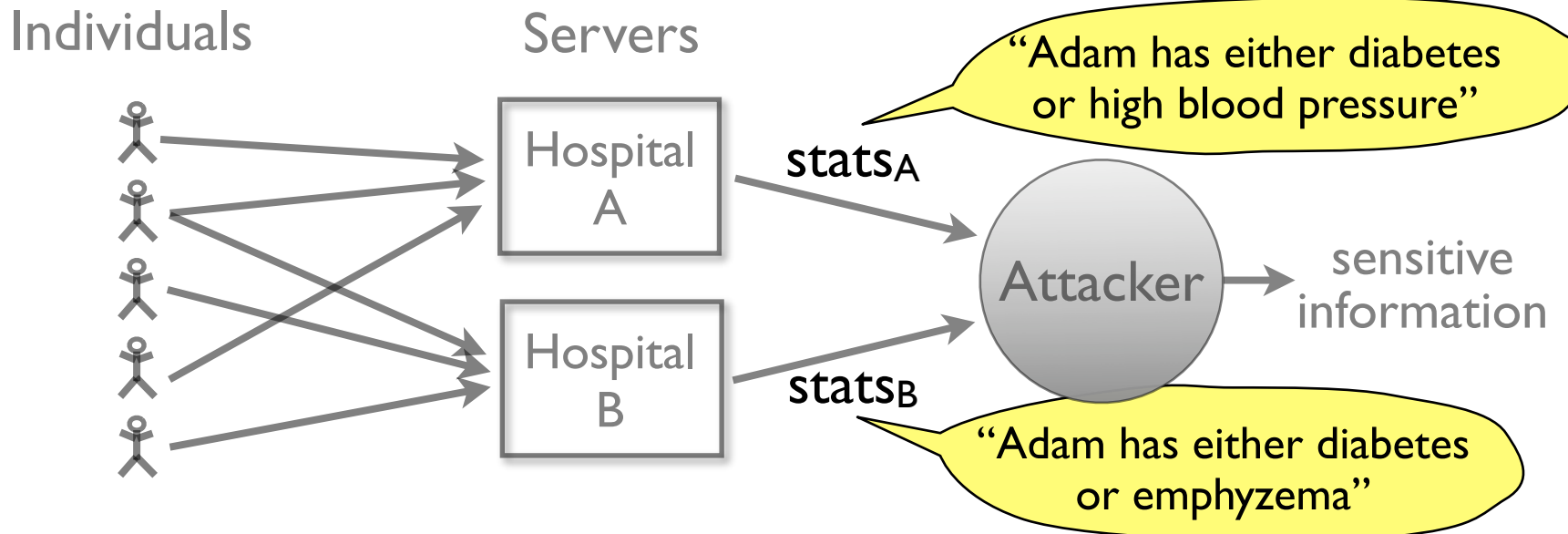
- **Example:** two hospitals serve overlapping populations
 - What if they **independently** release “anonymized” statistics?
- **Composition attack:** Combine independent releases
 - popular anonymization schemes leak lots of information
 - Litmus test for a proposed scheme’s reasonability?

“Composition” Attacks [Ganta-Kasiviswanathan-Smith, KDD 2008]



- **Example:** two hospitals serve overlapping populations
 - What if they **independently** release “anonymized” statistics?
- **Composition attack:** Combine independent releases
 - popular anonymization schemes leak lots of information
 - Litmus test for a proposed scheme’s reasonability?

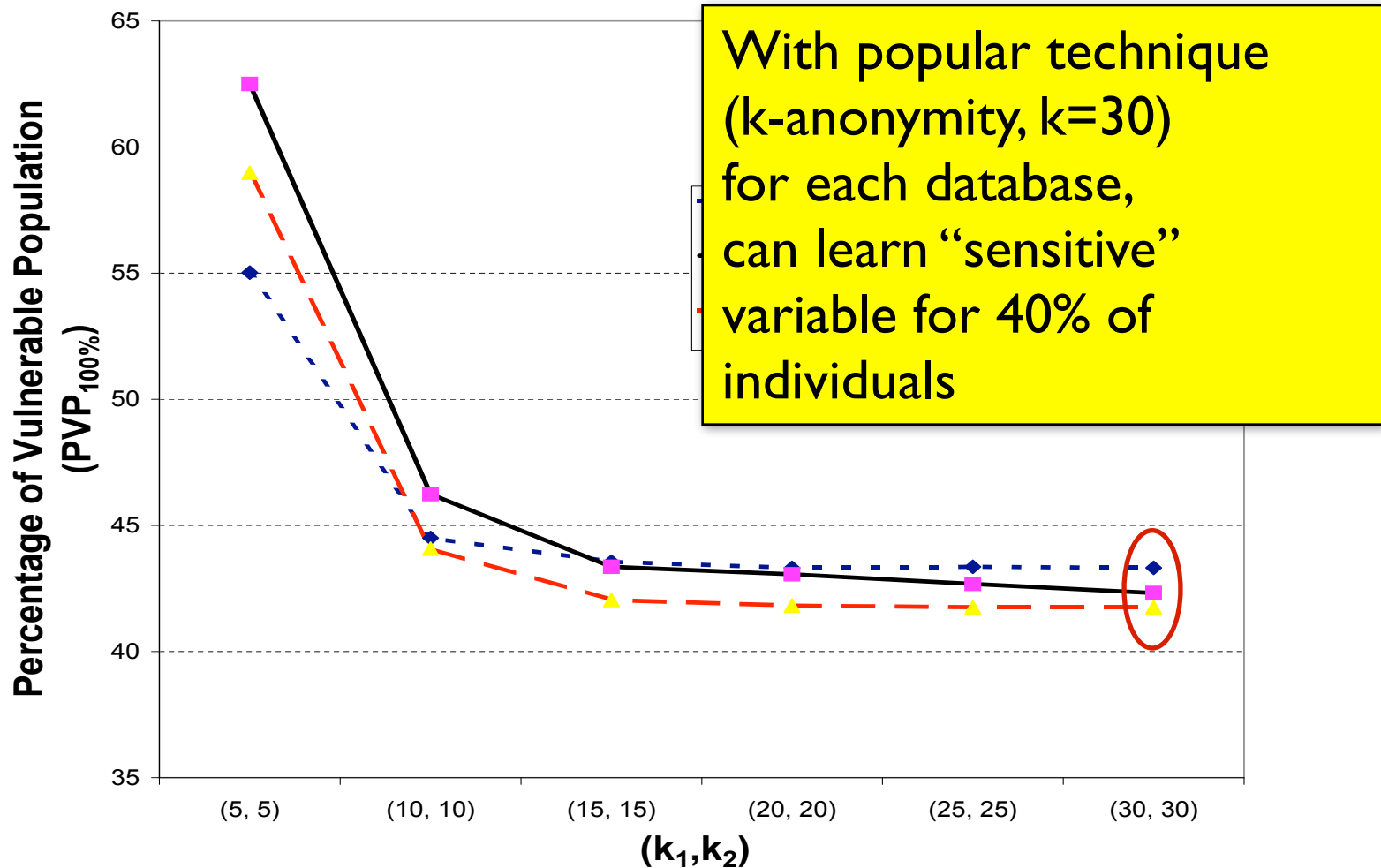
“Composition” Attacks [Ganta-Kasiviswanathan-Smith, KDD 2008]



- **Example:** two hospitals serve overlapping populations
 - What if they **independently** release “anonymized” statistics?
- **Composition attack:** Combine independent releases
 - popular anonymization schemes leak lots of information
 - Litmus test for a proposed scheme’s reasonability?

“Composition” Attacks [Ganta-Kasiviswanathan-Smith, KDD 2008]

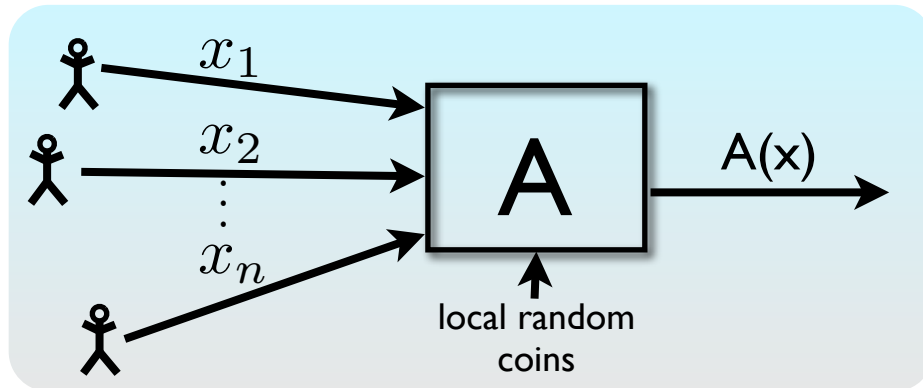
- “IPUMS” census data set. 70,000 people, randomly split into 2 pieces with overlap 5,000.



Differential privacy

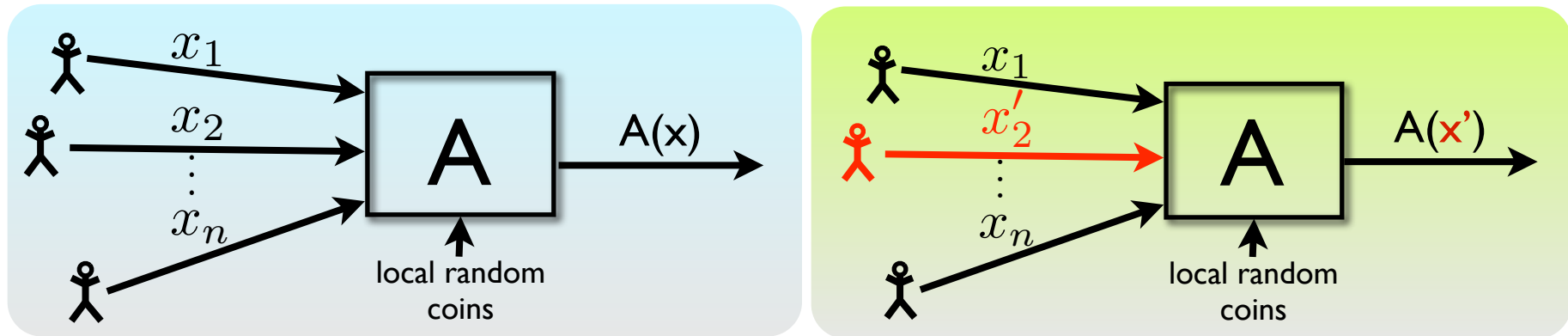
(How) can we distinguish **global**
from **individual** information?

Defining Privacy [Dwork-McSherry-Nissim-Smith 2006]



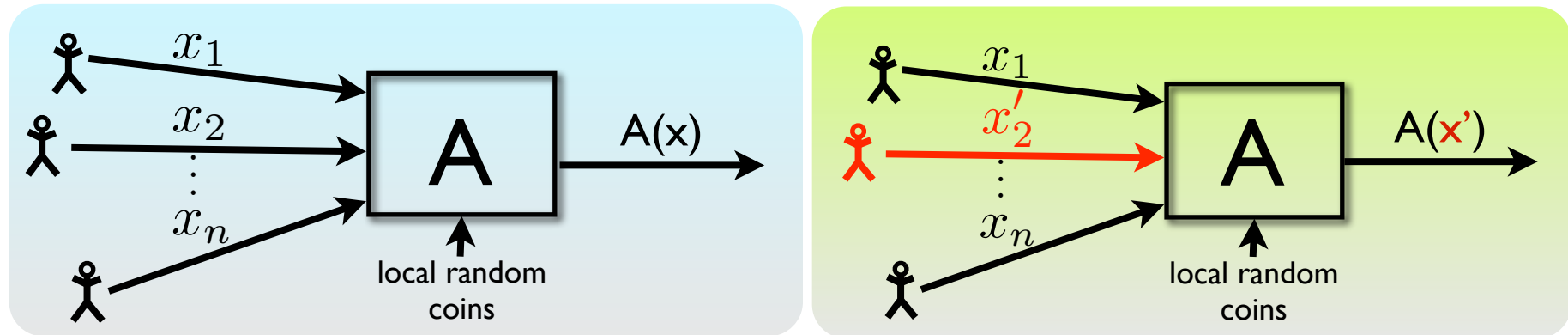
- Intuition:
 - Changes to one person's data not noticeable by users
- Data set $\mathbf{x} = (x_1, \dots, x_n) \in D^n$
 - Domain D can be numbers, categories, tax forms
 - Think of \mathbf{x} as **fixed** (not random)
- $A =$ **randomized** procedure run by the agency
 - $A(\mathbf{x})$ is a random variable distributed over possible outputs

Defining Privacy [Dwork-McSherry-Nissim-Smith 2006]



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

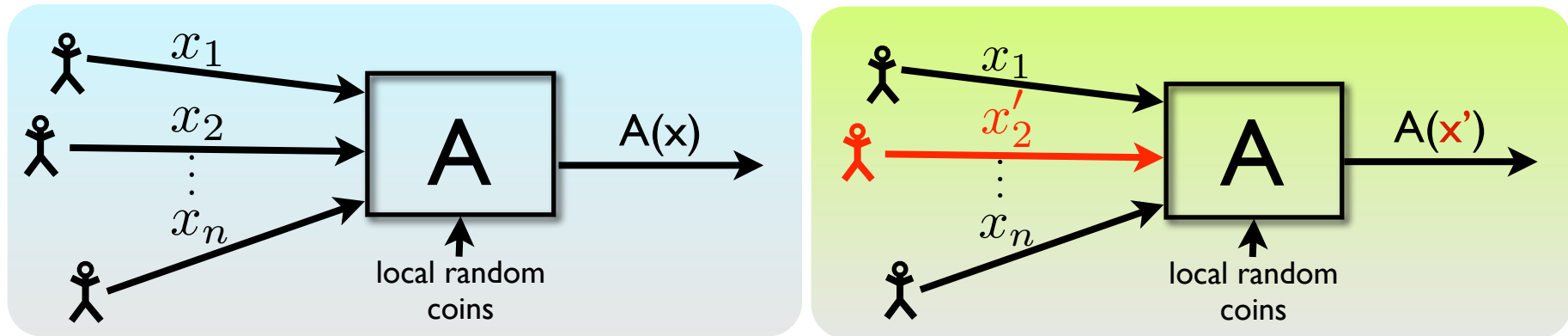
Defining Privacy [Dwork-McSherry-Nissim-Smith 2006]



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

Neighboring databases
induce **close** distributions
on outputs

Defining Privacy [Dwork-McSherry-Nissim-Smith 2006]



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

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

Neighboring databases
induce **close** distributions
on outputs

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

Why is this a good approach?

- “**Composition**”: **If** algorithms A_1 and A_2 are ϵ -differentially private **then** the outputting results of both algorithms $A_1(x), A_2(x)$ is 2ϵ -differentially private
- Meaningful in the presence of **arbitrary external information**

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

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

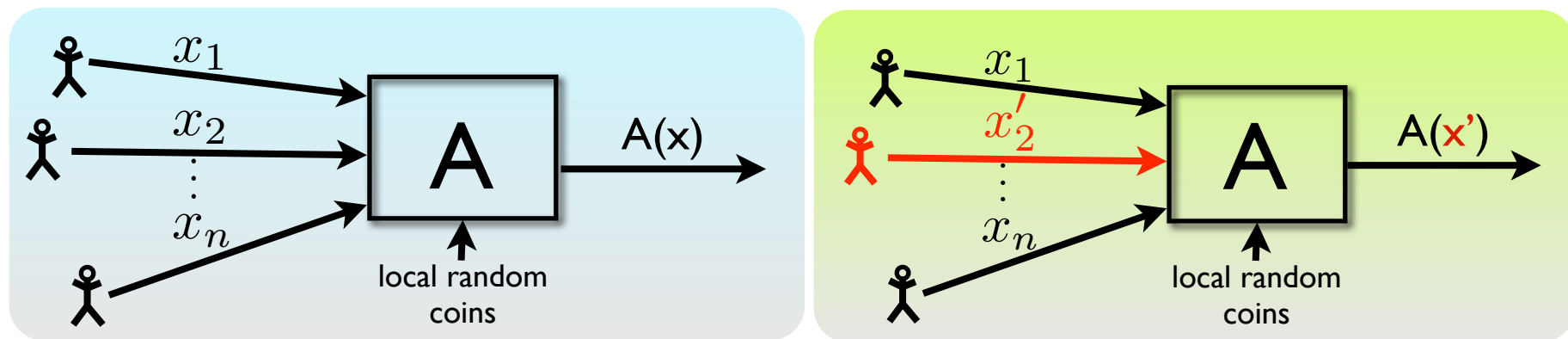
Neighboring databases induce **close** distributions on outputs

A differential guarantee

- Suppose you know that I used to smoke
 - A public health study could teach you that I am at risk for cancer
 - But it didn't matter whether or not my data was part of it.
 - Has my “privacy” been compromised?

- Differential privacy implies:
 - Whatever you know ahead of time, you learn (almost) the same things about me
whether or not my data is used

What can we **compute** privately?



- “Privacy” = change in one input leads to small change in output distribution

What computational tasks can we achieve privately?

- Lots of recent work, open questions

A sampling of results (STOC,FOCS,PODS,VLDB,KDD,CRYPTO,...)

- **Basic Techniques**, e.g.
 - noise addition [Dwork-McSherry-Nissim-S.'06, NRS'07, ...]
 - Optimizations [BCDKMT'07, HRMS'10, LiHRMM'10, ...]
 - Programming tool “PInQ” [McSherry'09]
 - “exponential sampling” [McSherry-Talwar'07]
- **Broad feasibility statements**, e.g.
 - PAC Learning / classification [KasiviswanathanLNRS '08]
 - Statistical estimation [Wasserman-Zhou '09, S. '09, Dwork-Lei '09]
 - Synthetic data [Blum-Ligett-Roth '08, DNRRV '09, DRV '10]
- **Distributed models** [DKMMN'06, BNO'08, DMPR'10]
- **Impossibility results** [Dinur-Nissim'03, DMT'07, DY'08, KRSU'10, HT'10]

Example: Statistical Estimation [S'08]

- Given a parametric model $\{f_\theta : \theta \in \Theta\}$
- **Goal:** given samples x_1, x_2, \dots, x_n , estimate θ
- Maximum likelihood estimator (MLE) = $\operatorname{argmax}_\theta(f_\theta(x))$
 - MLE has **optimal error (asymptotically)**
for well-behaved families of distributions
- **Theorem:** For any well-behaved parametric family, one can construct a diffe. private **efficient** estimator A
 - $A(X)$ converges to MLE
- Question:
 - How well does error scale with **dimension** (# of parameters)?

Technique: “sample-and-aggregate” [NRS]

- Break data into k blocks of n/k points

- For block i , ($i=1, 2, \dots, k$)
compute MLE $\hat{\theta}_i$

- Correct for bias:

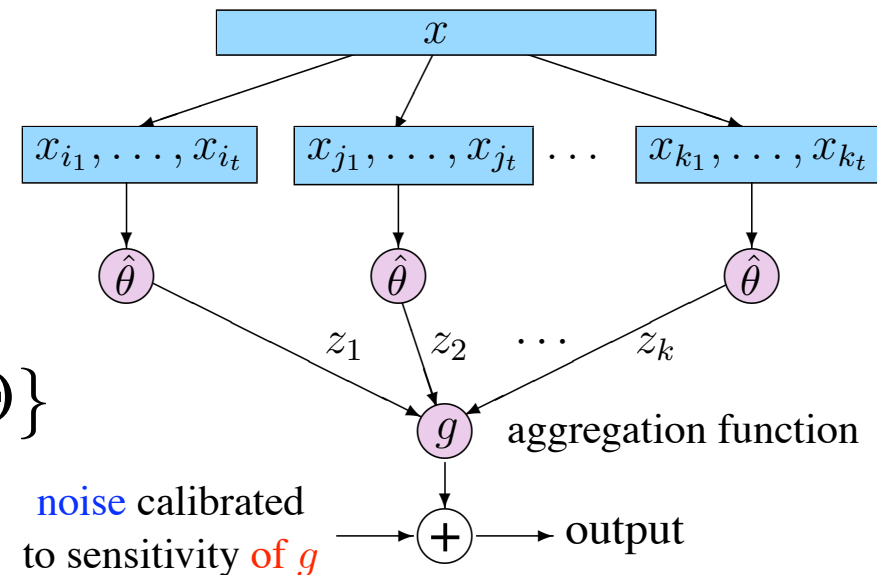
$$\tilde{\theta}_i \leftarrow \hat{\theta}_i - \text{bias}(\hat{\theta}_i)$$

where the bias function

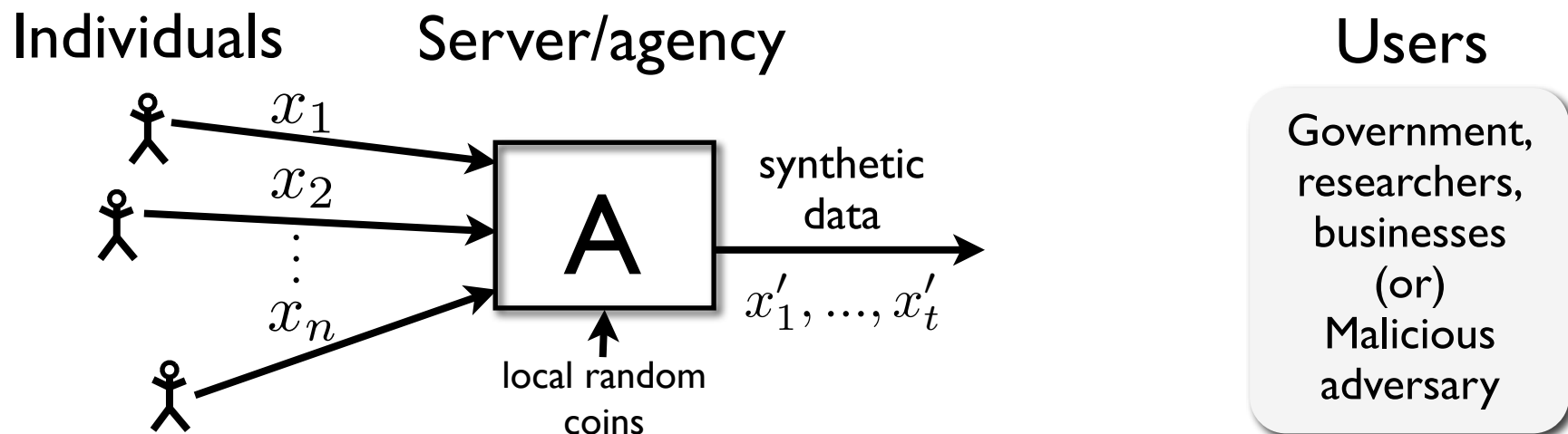
depends on family $\{f_\theta : \theta \in \Theta\}$

- Output aggregate:

$$A(x) = \frac{1}{k} \left(\sum_i \tilde{\theta}_i \right) + \text{noise}$$



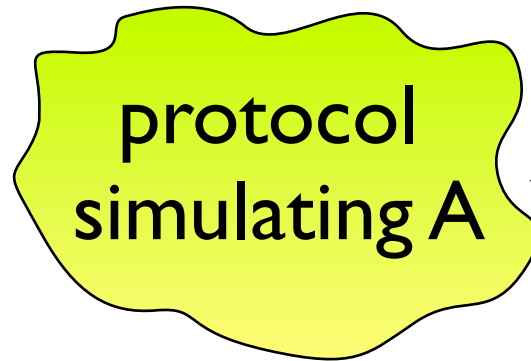
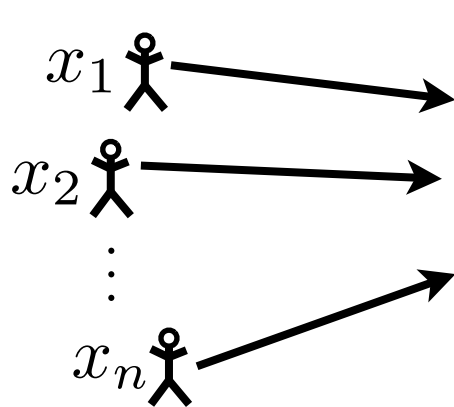
Synthetic Data



- **Goal:** new data set with “similar” statistical properties
 - Specify precisely the set of preserved properties
 - [Blum, Ligett, Roth 2008, Dwork-Naor-Reingold-Rothblum-Vadhan 2009]
broad theoretical possibility results
 - [Machanavajjhala, Kifer, Abowd, Gehrke, Vilhuber 2008, McSherry-Talwar 2008]
Differentially private geographic data, in use at the US Census

Distributed Private Data Mining

Individuals



Users

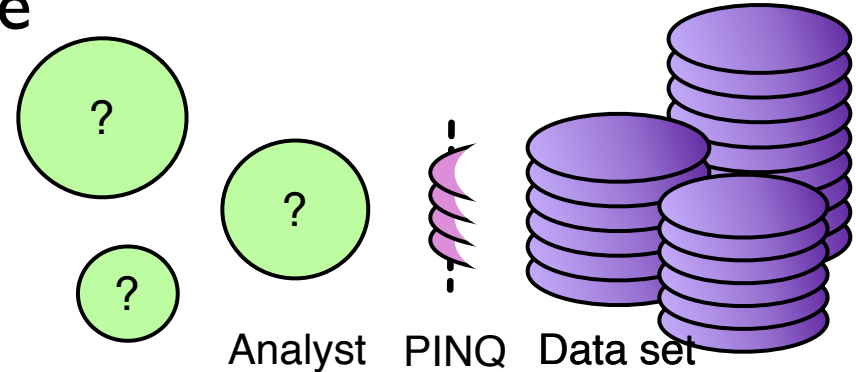
Government,
researchers,
businesses
(or)
Malicious
adversary

- Eliminate the trusted server/agency
- Use cryptographic protocols to jointly mine shared data
 - Individuals retain data
 - Mining algorithm still needs to respect (differential) privacy
- “Our ~~Bodies~~, Ourselves” (Dwork)
Data

A programming tool [McSherry '09]

- PInQ: database query language based on LInQ (~ SQL)

➤ Ever valid query is differentially private



- Available online

<http://research.microsoft.com/en-us/projects/PINQ/>

- Demo by Common Data Project

<http://demos.commondataprotect.org/PINQDemo.html>



Perspective

- **Goal: rigorous foundations for privacy** of statistical data
 - Differential privacy is one approach
- **What other notions provide similar guarantees?**
 - Can we exploit uncertainty about data, computational limits?
- **When are aggregate statistics a problem?**
 - E.g. gov't stats may leak classified info
 - How can we distinguish **good** information from **bad**?
- **How to reason about privacy & anonymity more generally?**
 - Where else can a similar perspective be applied?