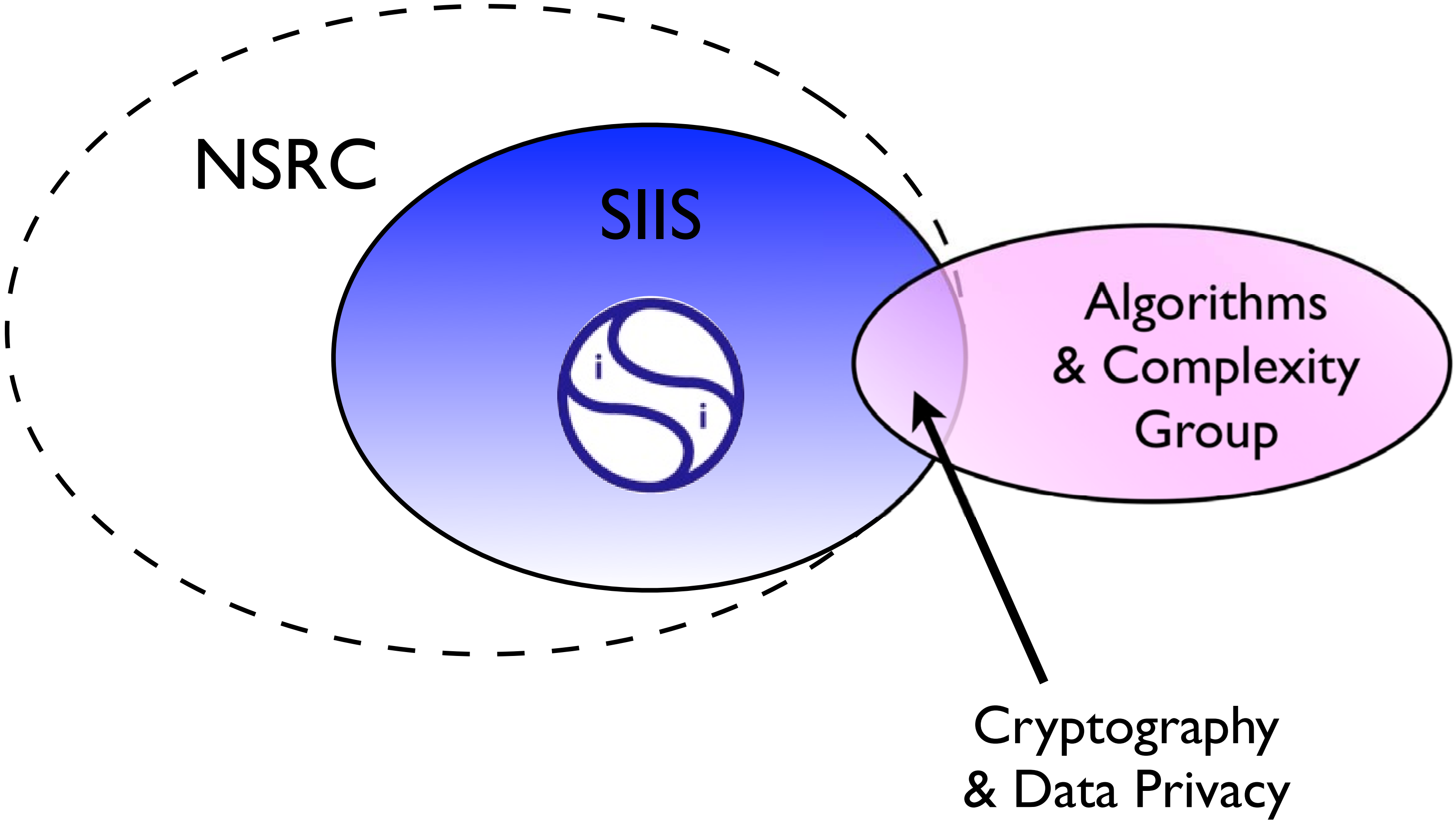
Cryptography & Data Privacy Research in the NSRC

Adam Smith

Assistant Professor
Computer Science and Engineering

Cryptography & Data Privacy @ CSE



Algorithms & Complexity

- Research on the theoretical foundations of computer science
  - Algorithm design
  - Complexity theory and lower bounds
  - Cryptography and information theory
  - Combinatorics and discrete mathematics

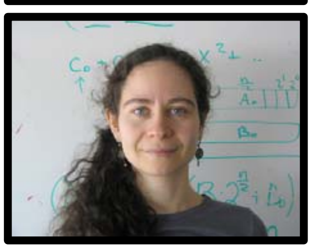
- Collaboration with other research groups
  - **Give**: abstractions, modeling, applications of algorithmic techniques
  - **Get**: new theoretical, mathematical challenges

Algorithms & Complexity: Faculty

- Piotr Berman
  - ➢ Combinatorial optimization, computational biology

- Martin Fürer
  - ➢ Complexity theory, combinatorics

- Sean Hallgren (Sep. '07)
  - ➢ Quantum computing, computational complexity



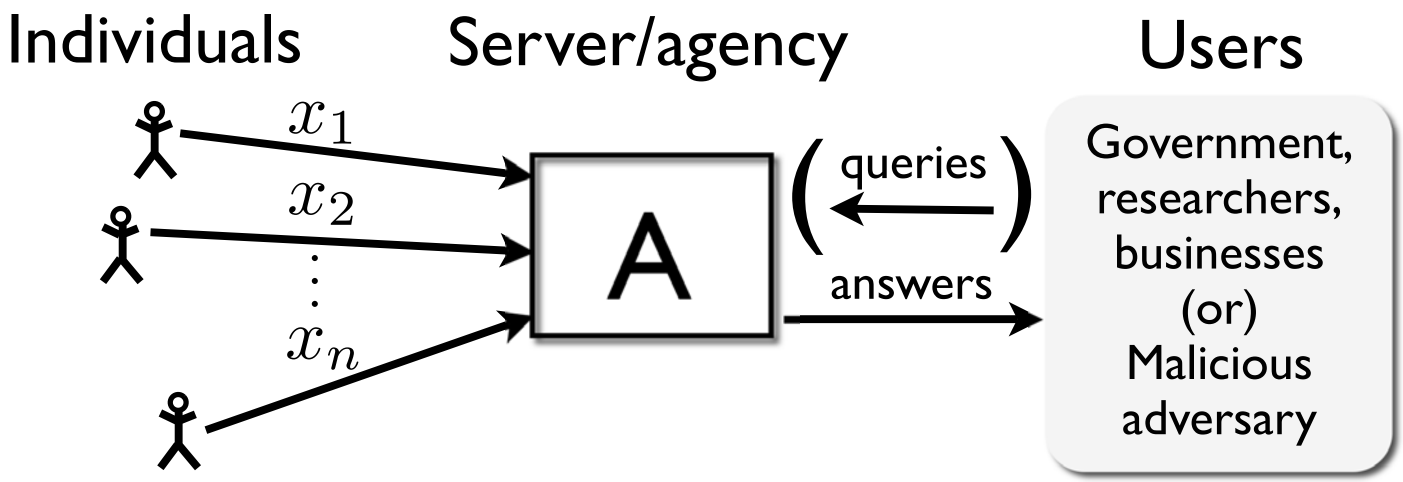
- Sofya Raskhodnikova (Jan. '07)
  - ➢ Sublinear algorithms, complexity, data privacy

- Adam Smith (Jan. '07)
  - ➢ Cryptography, data privacy, quantum information

Cryptography & Data Privacy

- Foundations of cryptographic protocols
  - Efficient Protocols for Multi-party Computations [*ICALP* 2004, J.-K.-M. *ESORICS* 2005,*Eurocrypt* 2008]
  - (Im)possibility of Deniable Authentication [Dodis, Katz, S., Walfish, Y. Youn, in progress]
- Key Extraction from Noisy Secrets
  - biometrics, voiceprint [*Eurocrypt* 04/05, *STOC* 05, *Crypto* 2006, SICOMP2008]
- Quantum cryptography
  - Understanding how recent technology impacts security and deniability [*STOC* 02, *FOCS* 02, *Eurocrypt* 05, *STOC* 06]
- Privacy in Statistical Databases

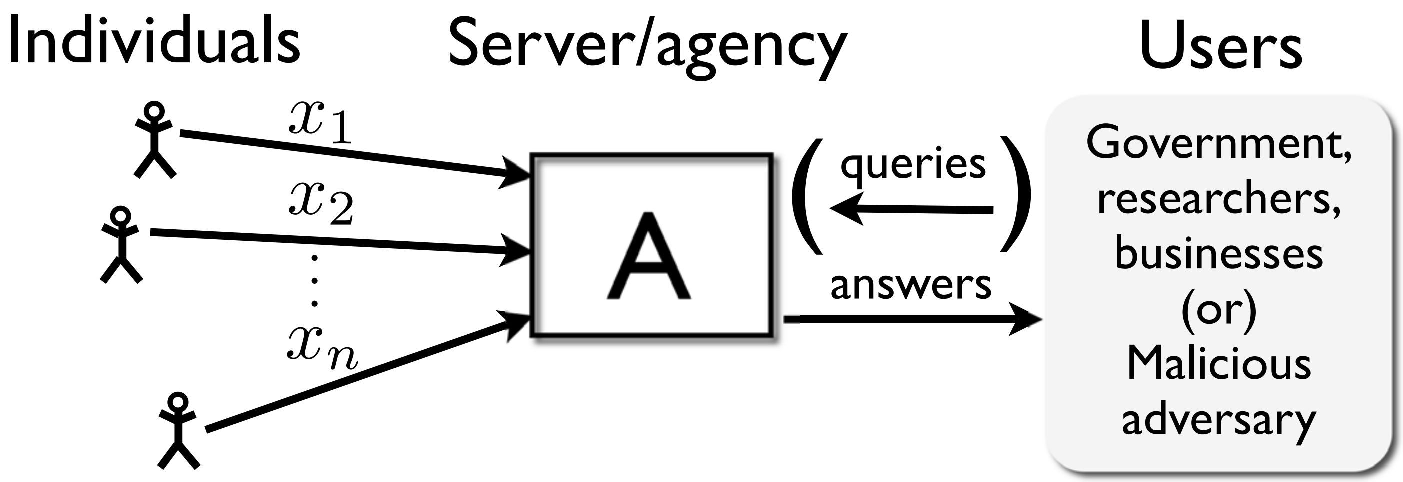
Privacy in Statistical Databases



Large collections of personal information

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

Privacy in Statistical Databases



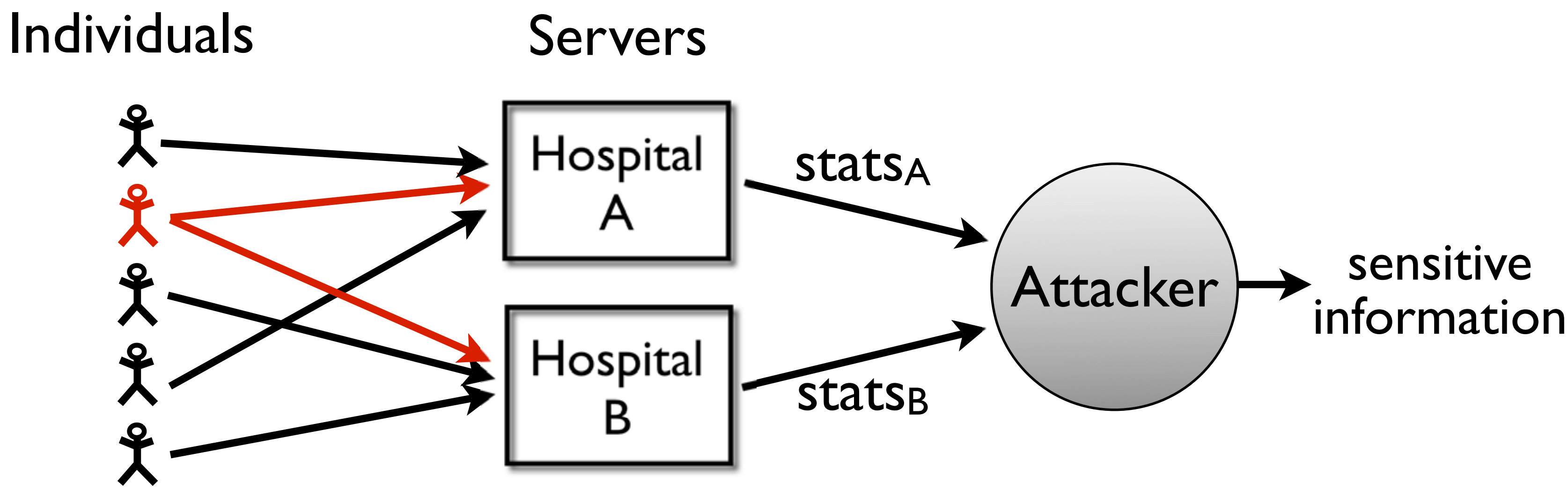
- Two conflicting goals
  - Utility: Users can extract "global" statistics
  - Privacy: Individual information stays hidden

Our Work

- Unify approaches from disparate fields
  - statistics, data mining, database theory, cryptography,...
- Rigorous formulations of "privacy"
  - Want **provable guarantees** that sensitive info. is not leaked
  - Should be secure against **arbitrary side information**
- New protocols / techniques [TCC'06,STOC'07,FOCS'08,...]
- New attacks [Ganta-Kasivswanathan-Smith, KDD 2008]

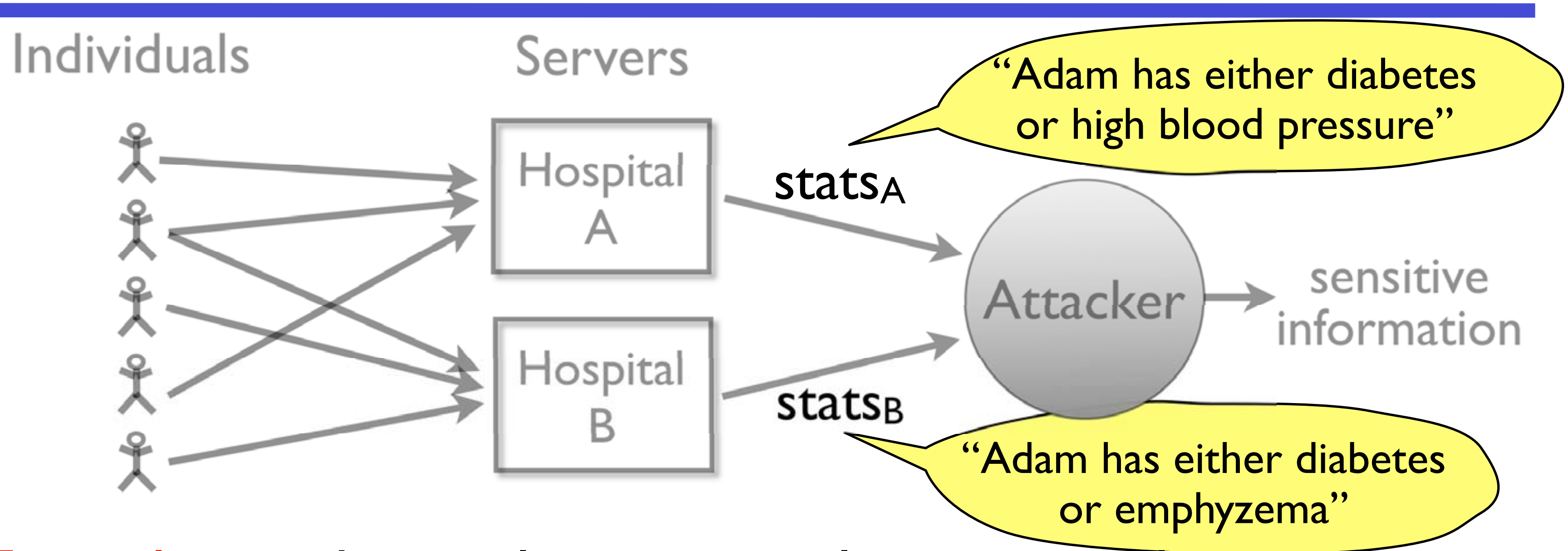
NSF

"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 schemes leak lots of information
  - Litmus test for a proposed scheme's reasonability?

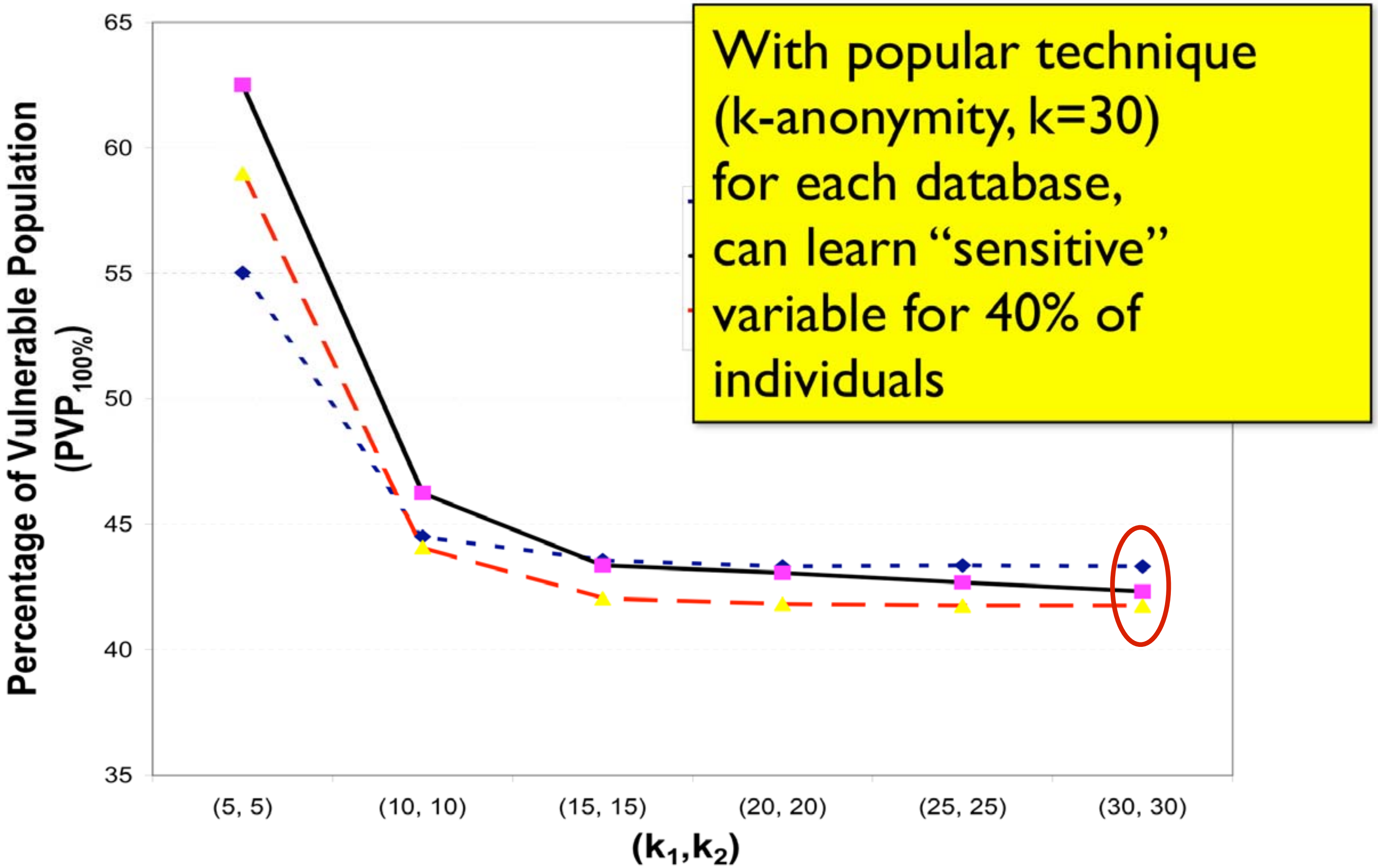
"Composition" Attacks [Ganta-Kasiviswanathan-Smith, KDD 2008]



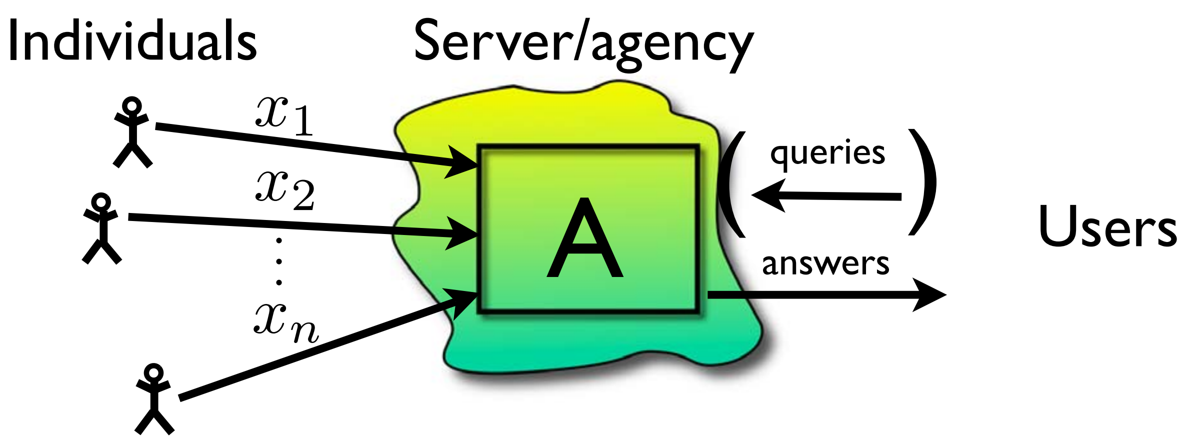
- Example: two hospitals serve overlapping populations
  - ➢ What if they independently release "anonymized" statistics?
- Composition attack: Combine independent releases
  - ➢ popular schemes leak lots of information
  - ➢ Litmus test for a proposed scheme's reasonability?

Does it work for real?

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



New Protocols *[TCC '06, STOC '07,FOCS'08]*

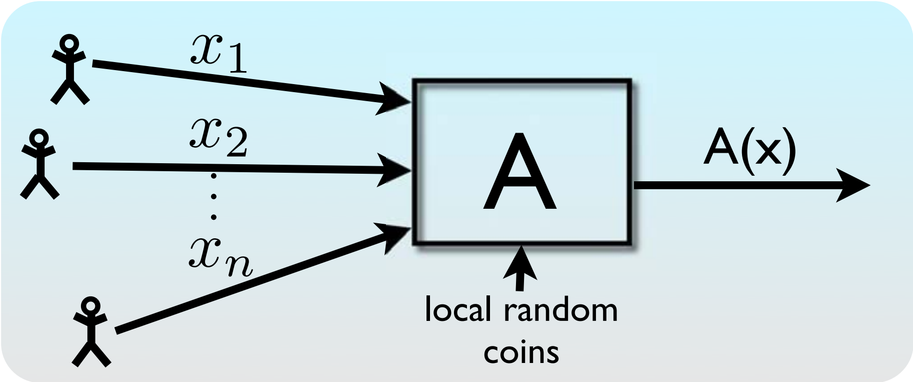


- **New notion:** Differential privacy
  - Roughly: any single individuals' data does not affect the release significantly
- Robust against very strong attacks
  - Correlation with arbitrary outside data collections
  - Composition attacks
- Practical...?
  - Common data mining algorithms can be modified to be D.P.
  - Apply current statistical methodology almost "as is"

Defining Privacy [D-M-N-S '06]

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

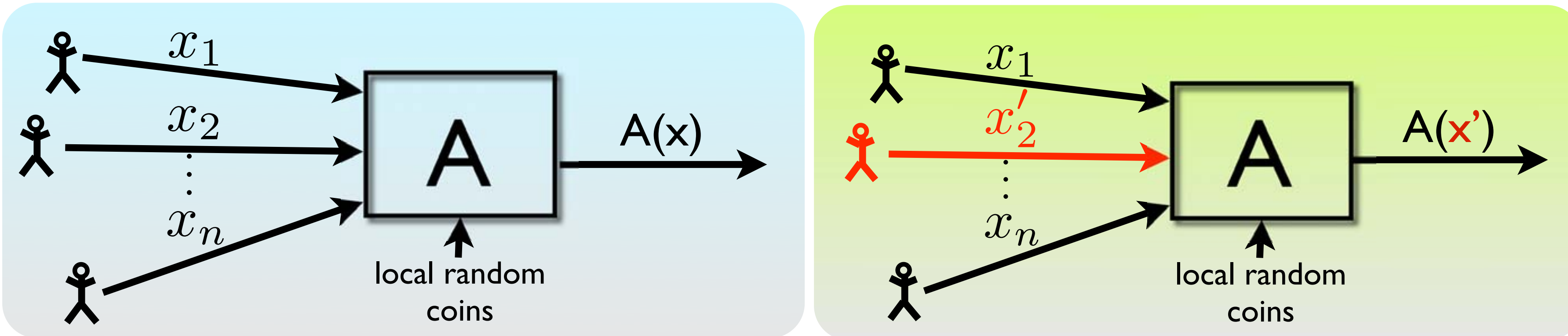
Defining Privacy [DiNi,DwNi,BDMN,DMNS]



- Data set $\mathbf{x} = (x_1, ..., x_n) \in D^n$
  - ➢ 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

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

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

Why is this a good definition?

- [DM] Differential privacy implies:
  No matter what you know ahead of time,

  You learn the same things about me
  whether or not I am in the database

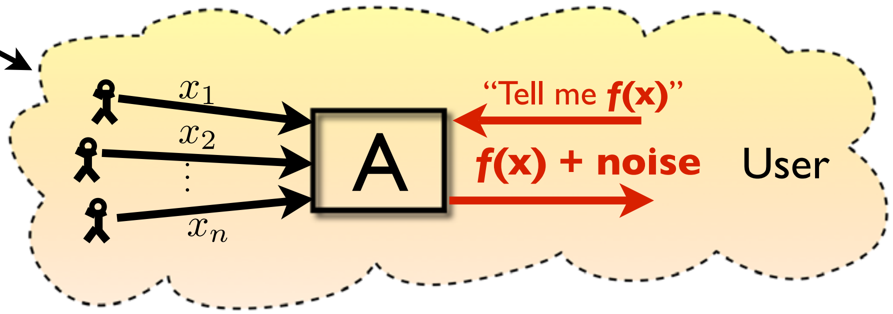
- Suppose you know I have a history of diabetes
  - You could learn that I have a high probability of early heart attack
    It doesn't matter whether or not my data is part of it.
  - Has the DB compromise my privacy?
    - **No**: it didn't have my data.
  - **Theorem** (Dwork-Naor): Learning things about individuals is unavoidable in the presence of external information

What can we compute privately?

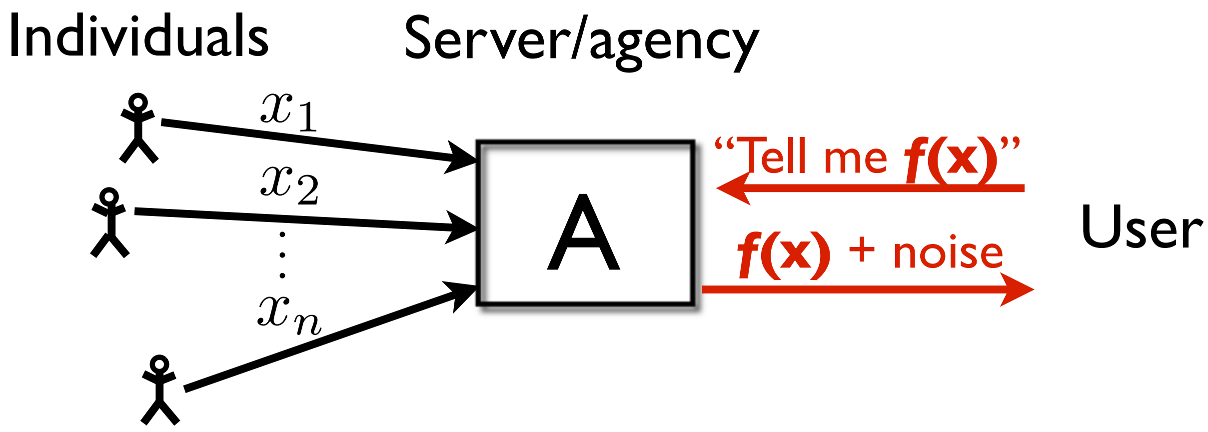
- "Privacy" = change in one input leads to small change in output distribution

  What computational tasks can we achieve privately?

- Research so far
  - ➢ Function approximation [DN, DN,BDMN,DMNS,NRS,BCDKMT,BLR]
  - ➢ Mechanism Design [MT]
  - ➢ Learning [BDMN,KLNRS]
  - ➢ Statistical estimation [S]
  - ➢ Synthetic Data [MKAGV]
  - ➢ Distributed protocols [DKMMN,BNO]
  - ➢ Impossibility results / lower bounds [DiNi,DMNS,DMT]



Our work [*TCC'06, STOC'07, FOCS'08*]



- Rigorous proofs of security
- Robust against very strong attacks
  - Correlation with arbitrary outside data collections
  - Composition attacks
  - Multiple releases of the same data statistics
- Suitable for applications where individual/organizational privacy is paramount

Summary

- **Foundations of Cryptography**
  - Efficient protocols
  - Basic (im)possibility questions
- **Privacy in Statistical Databases**
  - New protocols, new attacks
- In progress (we're still fairly new!):
  - Integration, collaboration with other CSE projects

Thank you

SIIS Lab:

`http://siis.cse.psu.edu`

A & C group:

`http://www.cse.psu.edu/theory`

Me:

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