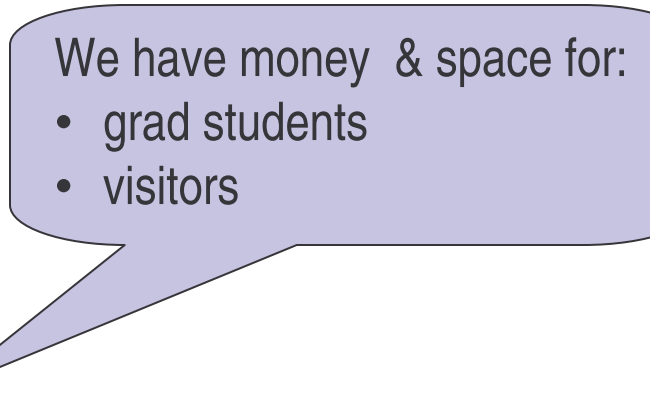
Interaction and Local Storage in Private Data Analysis



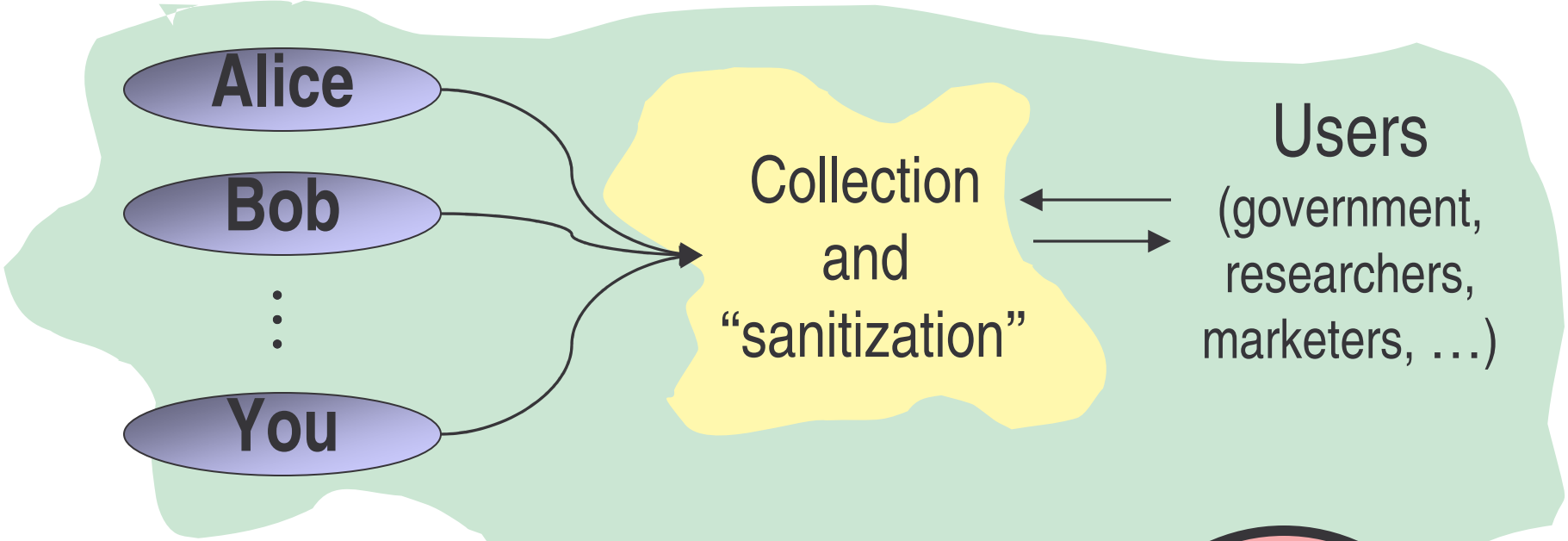
Adam Smith

Weizmann ⇝ **IPAM** ⇝ Penn State

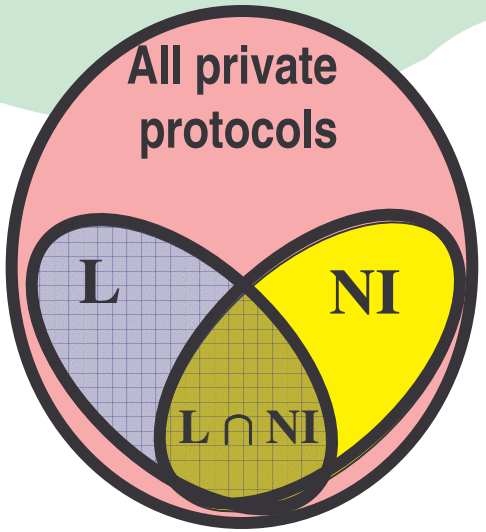
IPAM Workshop on Topics Vaguely Related to the Word “Privacy”

October 28, 2006

This talk



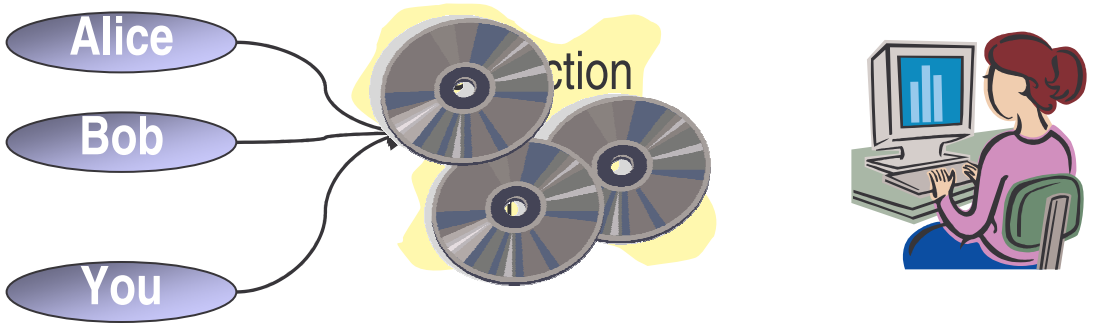
- Indistinguishability-based definition
- Limits on models of communication
  - Noninteractive protocols
  - Local protocols
- Where to from here?



Models for Data Privacy
Interactive vs Noninteractive

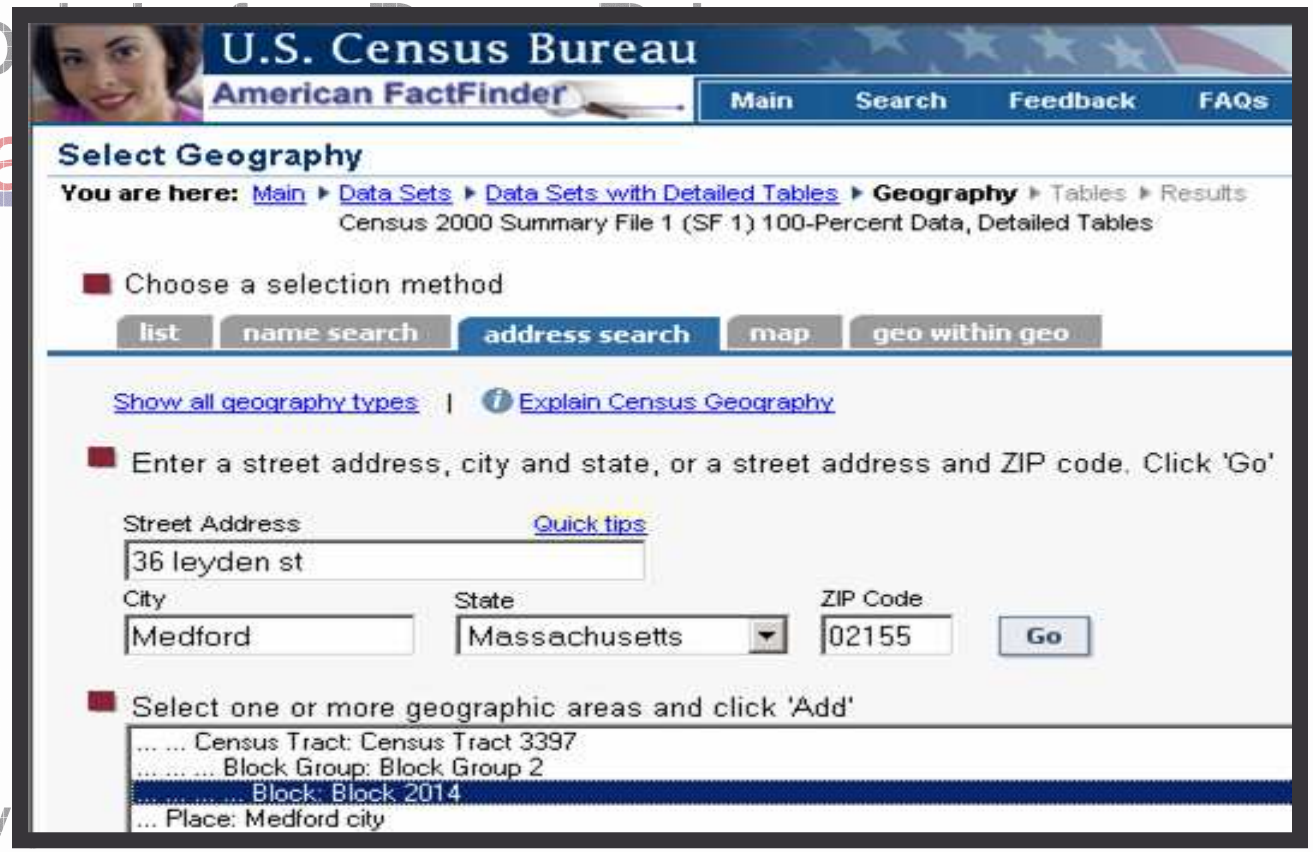
- Interactive:

- Noninteractive:

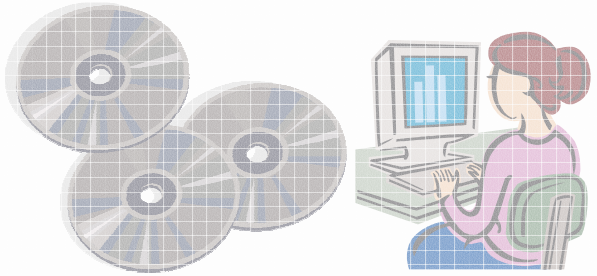


Mo
Intera

- Interactive:



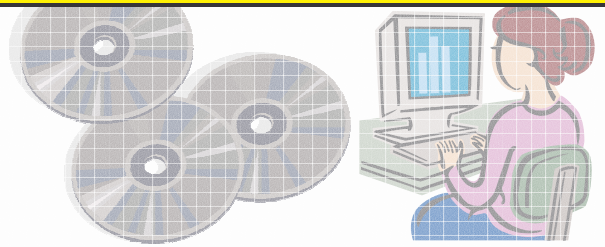
- Noninteractiv



Models for Data Privacy
Interactive vs Noninteractive

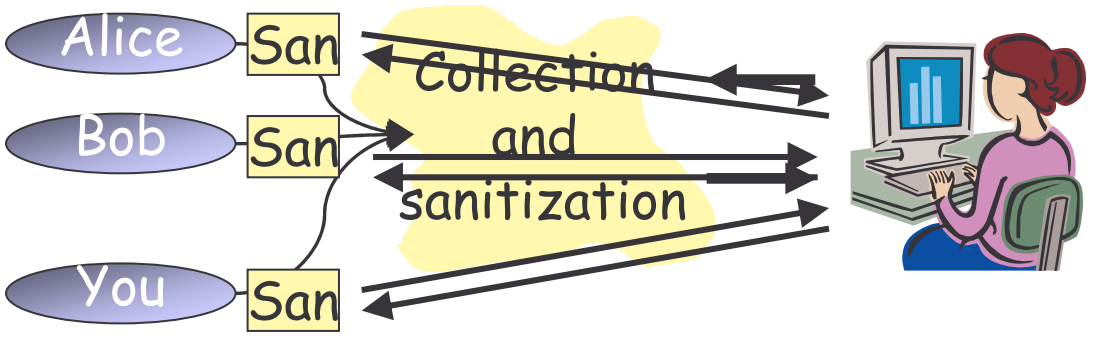
- Interactive
- Noninterac

- Noninteractive (vs. Interactive)
  - Easier distribution: web site, book, CD, …
  - More secure: can erase the data once it is processed
  - Almost all work in statistics, data mining is noninteractive!
  - Interactive solutions often illusory
  - Not (necessarily) true of recent crypto work

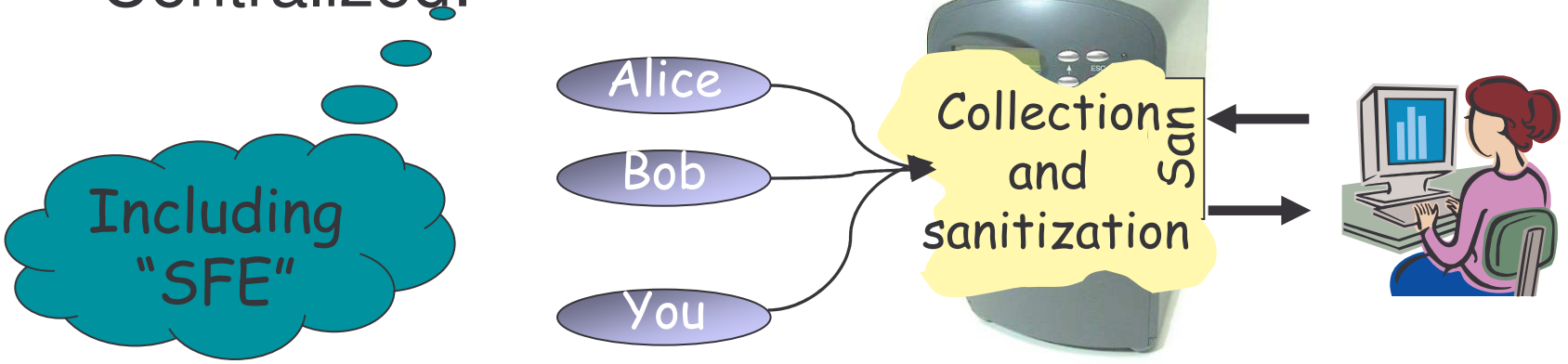


Models for Data Privacy - Local vs Centralized

- Local:

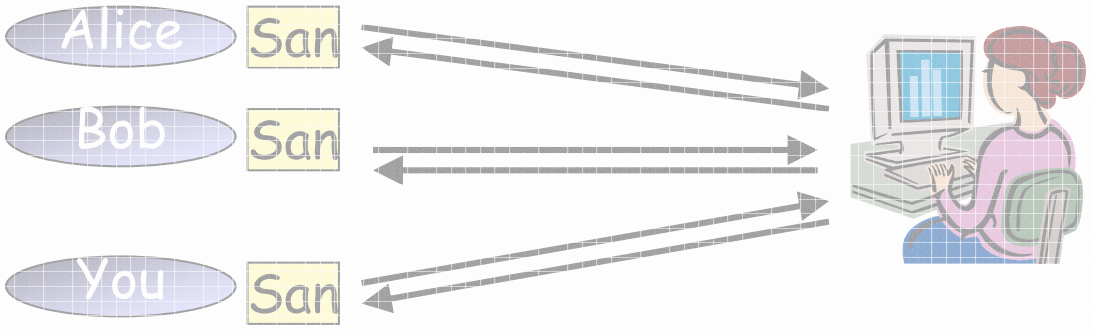


- Centralized:



Models for Data Privacy - Local vs Centralized

- Local:

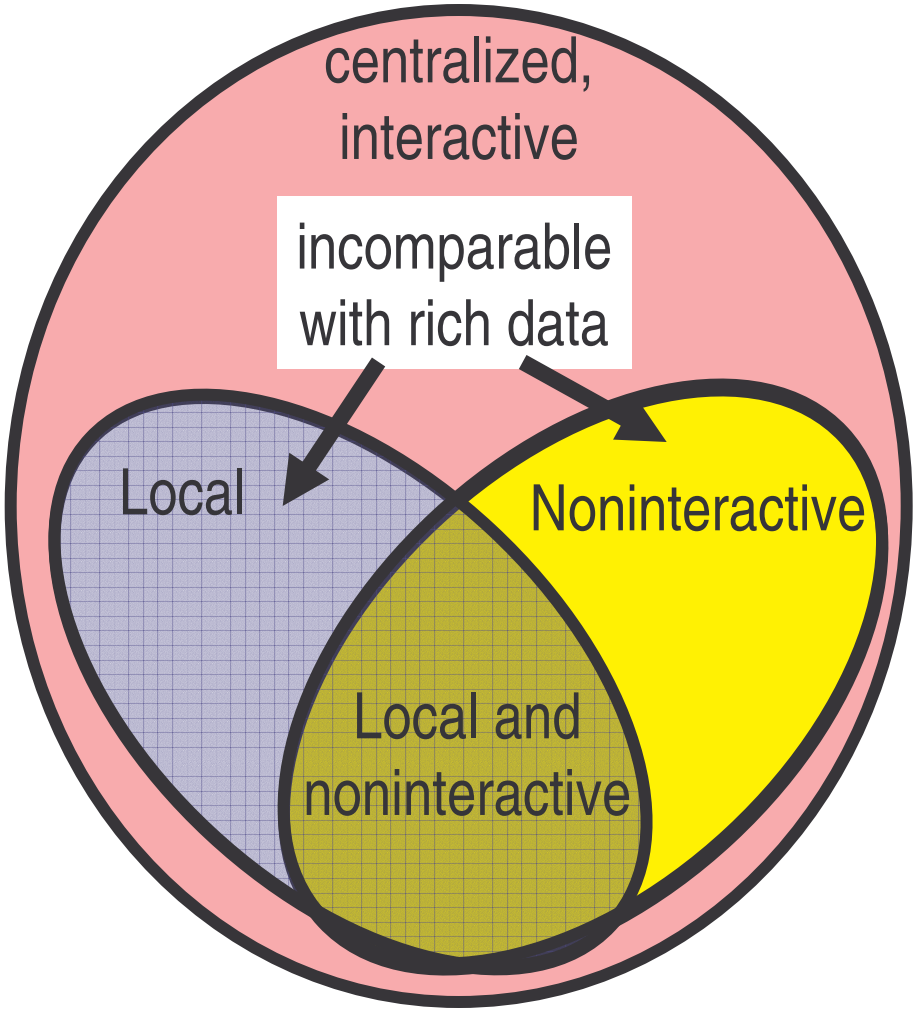


- Centralize

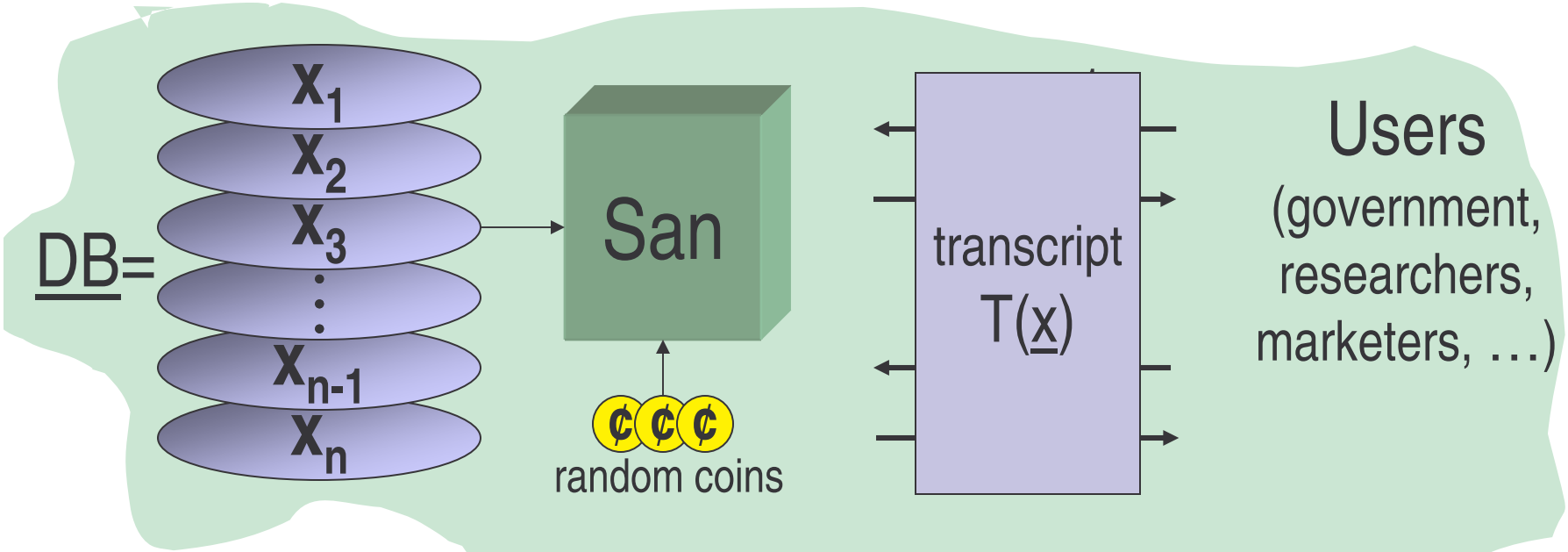
- Local (vs. centralized)
  - Non central trusted party
  - Individuals interact directly with (untrusted) user
  - Individuals control their own privacy
  - Most work in data mining is local
  - **Example**: randomized response

This talk

- Definitions
- Two types of Queries
  - Sum queries
  - Rank over $\mathbb{Z}_2$
- Protocols
- Impossibility results
- What it all means



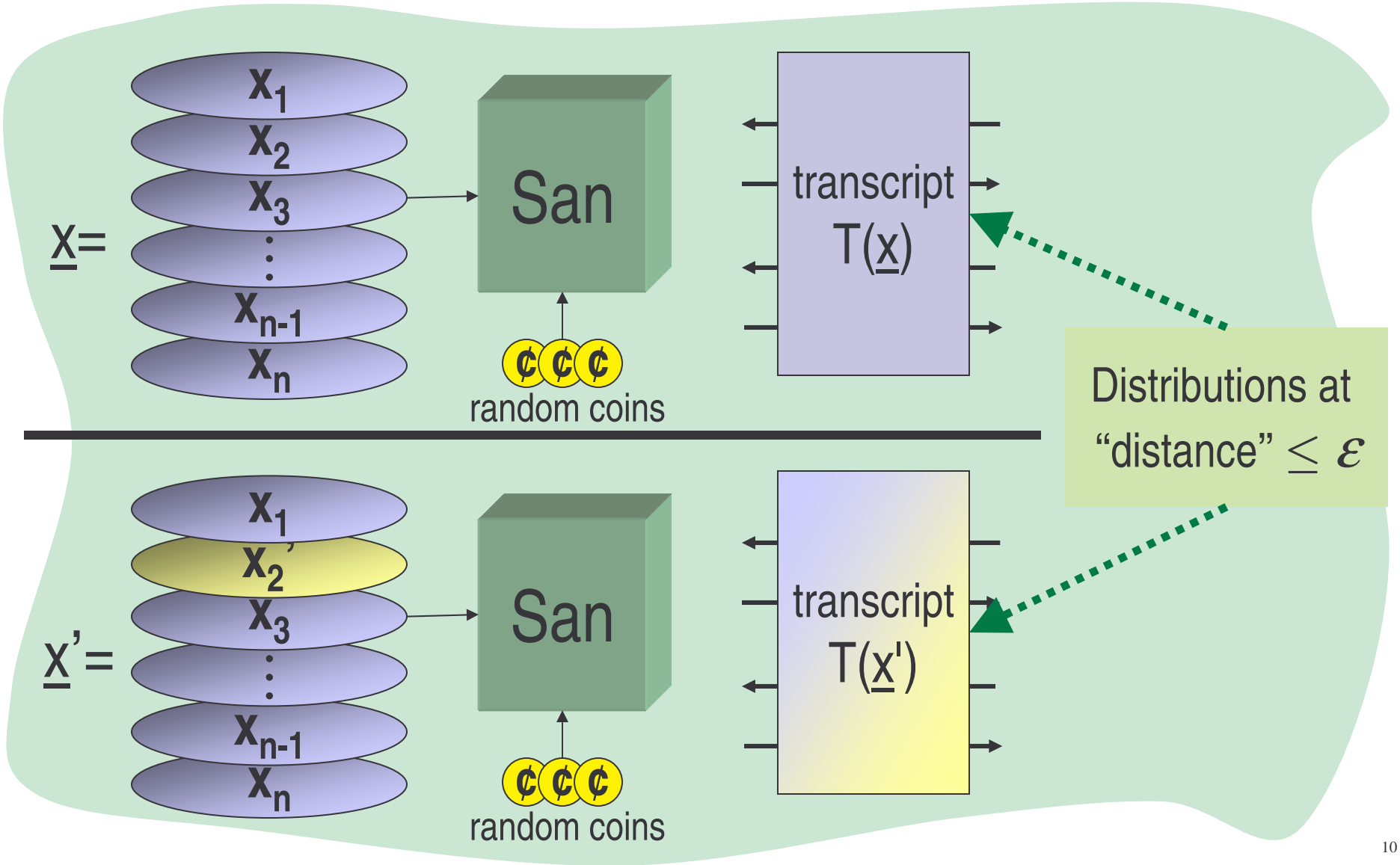
Basic Setting



- Database DB = table of n rows, each in domain D
  - D can be numbers, categories, tax forms, etc
  - This talk: D = $\{0,1\}^d$
  - E.g.: Married?, Employed?, Over 18?, …

Separations when
$n < \exp(d)$

Indistinguishability



Indistinguishability

Definition: *San* is ε-indistinguishable if

∀ A, ∀ $\underline{x}$, $\underline{x}$' which differ in 1 row, ∀ sets S of transcripts:

$$Pr(T(\underline{x}) \in S) \leq (1 + \varepsilon) Pr(T(\underline{x}') \in S) + \delta$$

non-negligible (small constant)

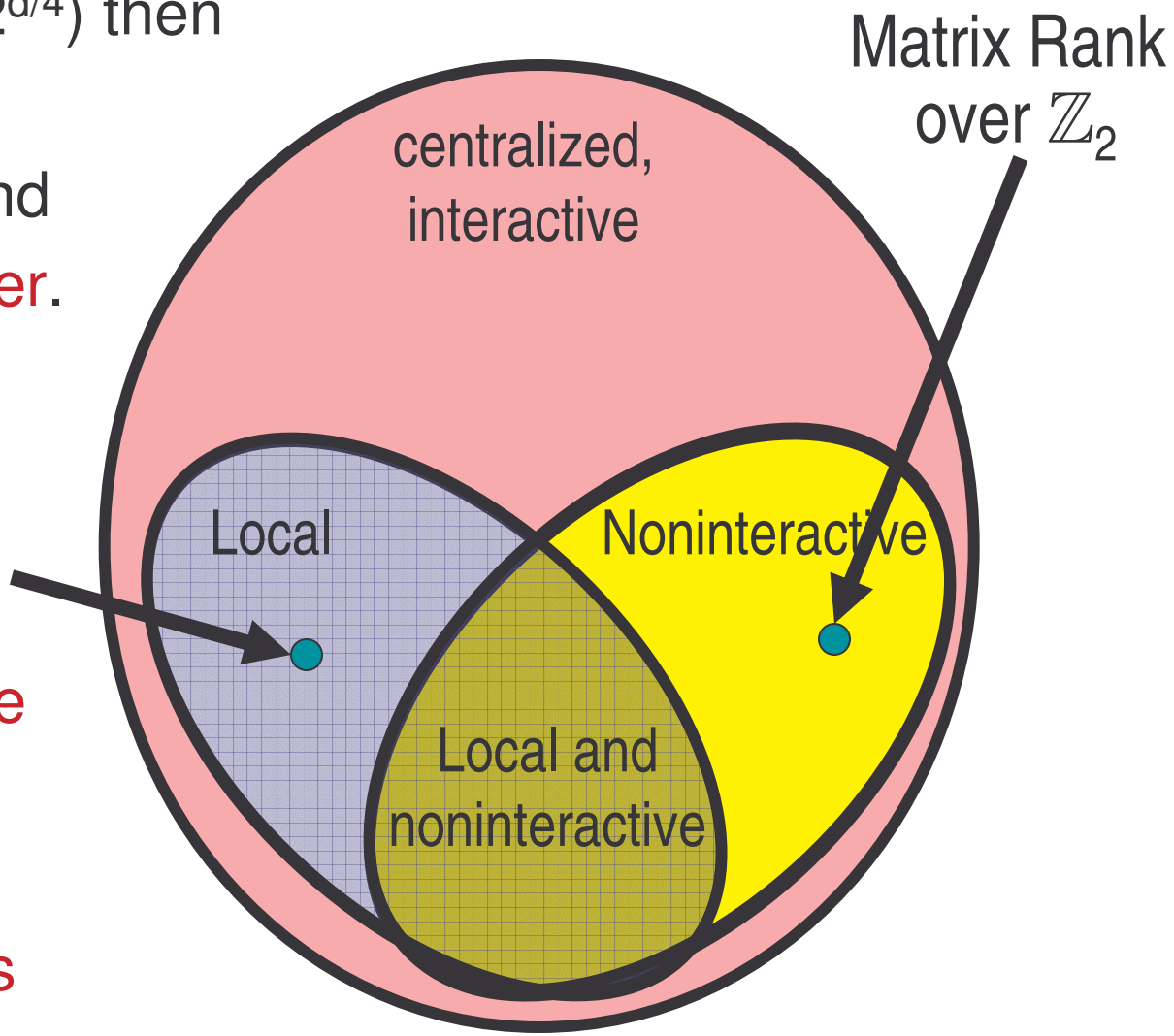
negligible

Models for Data Privacy

If data is "rich" ($n \leq 2^{d/4}$) then
$\exists$ data mining tasks
easy in one model and
impossible in the other.

"Online"
sum queries

These models require
exponential amounts
of data to answer
all possible questions



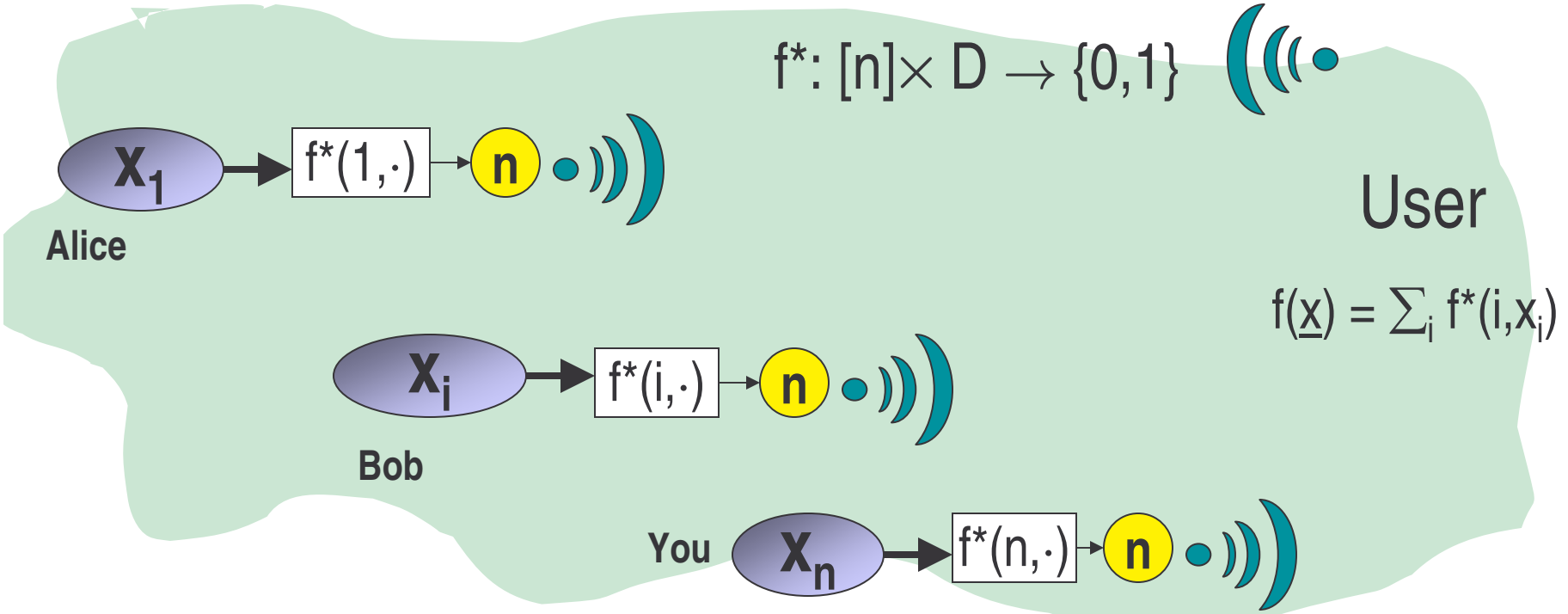
Strategy

- Two types of Queries
  - Sum queries
  - Rank over $\mathbb{Z}_2$
- Private protocols:
  - Interactive protocol for sum queries
  - Centralized protocol for rank
- Separations
  - "Online" sums **require** interaction
  - Rank **requires** centralized protocol

Queries

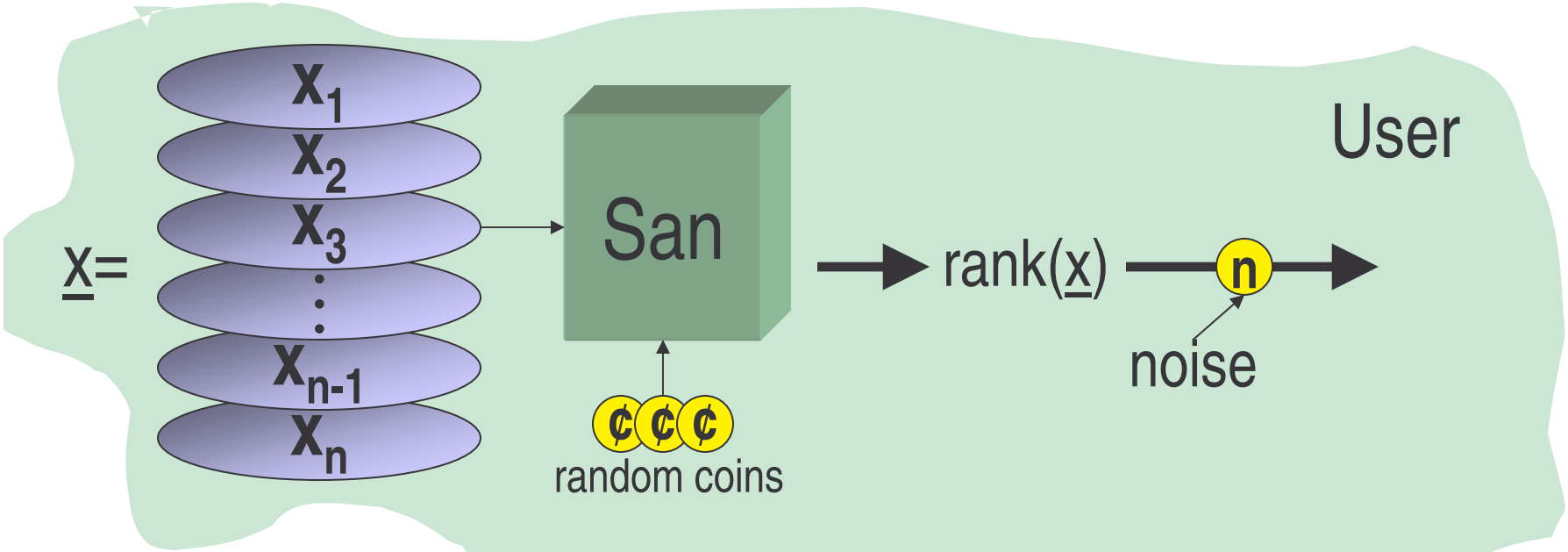
- Sum queries
  - Query: $f^*: [n] \times \{0,1\}^d \to \{0,1\}$
  - $f(\underline{x}) = \sum_i f^*(i, x_i)$
  - Captures natural counting tasks, linear algebra
    - Many other algorithms: clustering, perceptron, … [BDMN]
- Rank over $\mathbb{Z}_2$
  - View $\underline{x}$ as $n \times d$ matrix over $\{0,1\}$
  - $f(\underline{x}) = \text{rank}(\underline{x})$

Local & Interactive Sum Queries



- Each player computes $f(i,x_i)$ and adds noise $\approx T^{1/2}/\varepsilon$
  - Broadcast result
  - Total noise $\approx (Tn)^{1/2}/\varepsilon$
- Local
- Interactive: **users can choose queries online** ($\leq$ T of queries)

Centralized & Noninteractive Rank



- View $\underline{x}$ as $n \times d$ matrix over $\mathbb{Z}_2$
  - $f(\underline{x}) = \text{rank}(\underline{x})$
- Sensitivity 1
  - Release with noise $\approx 1/\varepsilon$
- Noninteractive but centralized

Strategy

- Two types of Queries
  - Sum queries
  - Rank over $\mathbb{Z}_2$
- Private protocols:
  - Interactive protocol for sum queries
  - Centralized protocol for rank
- Separations
  - "Online" sums **require** interaction
  - Rank **requires** centralized protocol

Main Results

If $n < 2^{d/4}$ (and $n < 1 / \delta^{1/4}$)

Theorem 1 [DMNS]: ∃ distrib over sum queries f*

∀ noninteractive San(·)

W.h.p. over sum queries f*(·,·)

San($\underline{x}$ s.t. $f(\underline{x})=0$ **)** ≈ **San(** $\underline{x}$ s.t. $f(\underline{x})=n$ **)**

No entry satisfies predicate

All entries satisfy predicate

- Mechanism must be tailored to specific functions (without lots of data)
- Quantifiers matter

Main Results

If $n < 2^{d/4}$ (and $n < 1 / \delta^{1/4}$)

Theorem 1 [DMNS]: $\exists$ distrib over sum queries f^*

$\forall$ noninteractive San(·)

W.h.p. over sum queries $f^*(\cdot,\cdot)$

San($\underline{x}$ s.t. $f(\underline{x})=0$ **)** $\approx$ **San(** $\underline{x}$ s.t. $f(\underline{x})=n$ **)**

Theorem 2 [NSW]:

$\forall$ local San(·) with $2^{o(d)}$ rounds

San($\underline{x}$ s.t. $\text{rank}(\underline{x})=d/2$ **)** $\approx$ **San(** $\underline{x}$ s.t. $\text{rank}(\underline{x})=d$ **)**

- Locality is very restrictive (without crypto [DKMMN])

Proof Idea

- For each index i:

$$Z(\mathbf{y}) = \mathbf{San}(x_1,\ldots,x_{i-1}, \mathbf{y}, x_{i+1},\ldots,x_n)$$

- $\forall$ y, y': **Pr(** Z(y)=z **)** $\leq$ 2 **Pr(** Z(y')=z **)**

- **Main Lemma:**
  Let V $\subseteq$ D=$\mathbb{Z}_2^d$ be random subspace of dimension k
  With prob. 1-exp(-k) : Z(**V**) $\approx_{\mathbf{exp(-k)}}$ Z($\{0,1\}^d$)

- Chebyshev bound for **distributions**
  - Think of V as pairwise-independent sample
  - Bound does not depend on size of Z's output!

Sums Require Interaction

- Say **San**(·) is noninteractive
- Pick $r_1,\ldots,r_n \in \{0,1\}^n$
- $f^*(i,x) = r_i \odot x$
- $V_i = \{ x : x \odot r_i = 0\}$
- $Z_i(\mathbf{y}) = \text{San}(V_1,\ldots,V_{i-1}, \mathbf{y}, D)$
- **Main lemma + hybrid**: With prob 1-exp(-d)

  $\text{San}(V_1,\ldots,V_n) \approx \text{San}(D^n)$
- Symmetric arg.: $\text{San}(D\setminus V_1,\ldots,D\setminus V_n) \approx \text{San}(D^n)$

  $\Rightarrow \text{San}(V_1,\ldots,V_n) \approx \text{San}(D\setminus V_1,\ldots,D\setminus V_n)$

Crucial that V_i is different for each player

Q.E.D.

Rank Requires Centralized Protocol

- **Distribution** 1:
  - $\underline{x} \leftarrow D^n$
- **Distribution** 2:
  - Choose $V \subseteq D$ of dimension n/2
  - $\underline{x} \leftarrow V^n$

Assume n > 2d…

- Rank d w.h.p.
- Rank d/2 w.h.p.

- Adversary's job:
  - Distinguish Distrib. 1 from Distrib. 2 via **local San**(\cdot)
  - But does not know V !
- Break **San** protocol into rounds $\mathbf{San}_1$, $\mathbf{San}_2$,…
- **Main lemma + hybrid**: Each round reveals no info about V
  - Adversary can't use adaptivity to tailor questions to V
  - Tricky: need to build aversary's code into Z

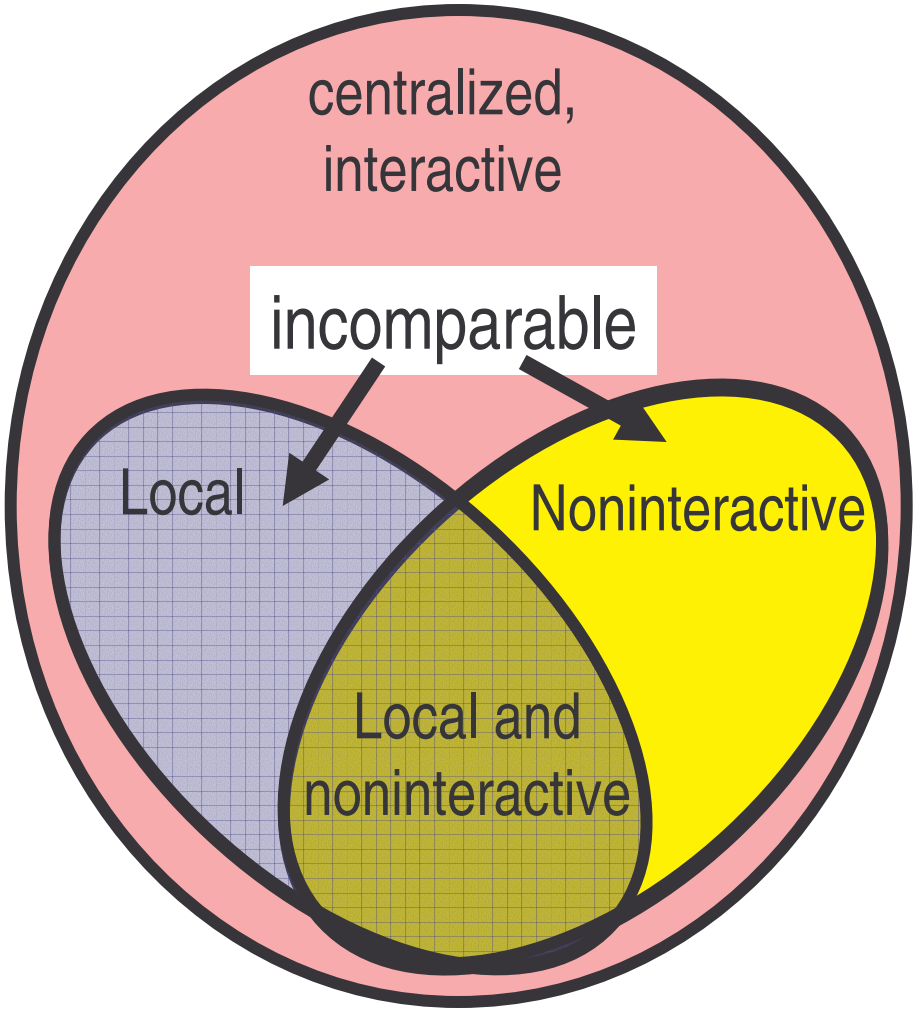
Q.E.D.

Main Lemma

- $\forall$ y, y': **Pr(** Z(y)=z **)** $\leq$ 2 **Pr(** Z(y')=z **)**
- **Main Lemma:**
  Let V $\subseteq$ D=$\mathbb{Z}_2^d$ be random subspace of dimension k
  With prob. 1-exp(-k) : Z(**V**) $\approx_{exp(-k)}$ Z(**{0,1}^{d}**)
- For each z:
  - Pr(Z(V)=z) = $\sum_{y \in V}$ Pr(Z(y)=z)
  - Pairwise-independent estimator for Pr(Z(D)=z)
  - Variance of estimator
    - proportional to Pr(Z(D)=z) $\leftarrow$ use bounded ratio
    - Scales with $|V|^{1/2}$ = exp(k) $\leftarrow$ Chebyshev
  - Markov: for most z get estimate $\in$ p(z) (1 $\pm$ exp(-k))
  - **Sum of errors** $\approx \sum_z$ p(z) (1 $\pm$ exp(-k)) $\leq$ 1 + exp(-k)

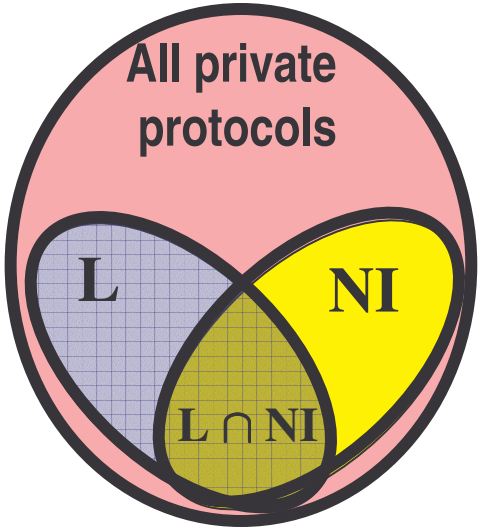
This talk

- ✓ Definitions
- ✓ Two types of Queries
  - ➢ Sum queries
  - ➢ Rank over $\mathbb{Z}_2$
- ✓ Protocols
- ✓ Impossibility results
- • What it all means



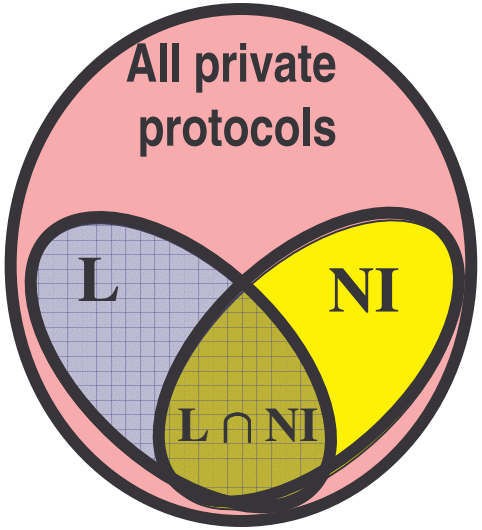
Conclusions

- Limits of differential-style definitions
  - Indistinguishability is “necessary” for handling arbitrary auxiliary information [DN]
- Separations based on communication model
- Perspective: **random sampling**
  - Random sample of size k
    - Answers sum queries
    - Compute rank of matrices $\ll$ k
    - Noninteractive & local (not private)
  - Restricted private protocols **cannot** simulate functionality of random sample
  - [NRS] Centralized, global protocols **can***



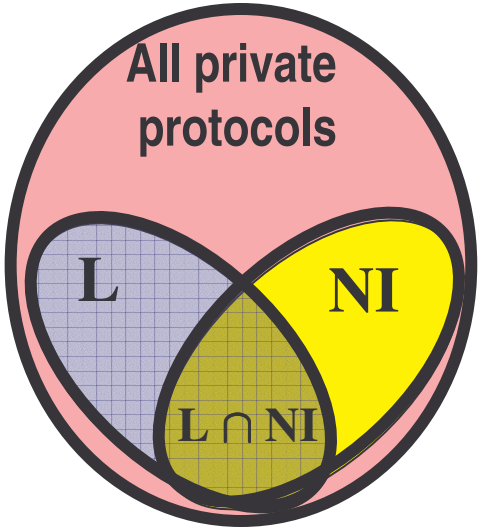
Where to go from here?

- Lower bounds for more natural queries?
  - Characterization of local / NI protocols?
- Computational security:
  - SFE circumvents locality [DKMMN]
  - More efficient solutions?
  - What about interactivity? (do as well as a sample?)
- Entropy assumptions?
- Design secure interactive implementation
- All the questions Kobbi asked…



Where to go from here?

- Big picture emerging
  - Definitions, feasibility, separations
  - Many technical questions still to be answered
- Crypto with non-negligible error?
  - Anonymity?
- Big(ger) picture
  - Economic / game-theoretic perspective
    - When will self-interested players insist on privacy?
    - Who decides on type/level of privacy?
  - Can this help other fields?
    - Getting good data is hard: IM data, cell phone records, health, etc
  - Settings where this notion of privacy is inappropriate?



Thank you

Ack's:
Cynthia Dwork
Frank McSherry
Moni Naor
Kobbi Nissim
Sofya Raskhodnikova
Gil Segev
Enav Weinreb