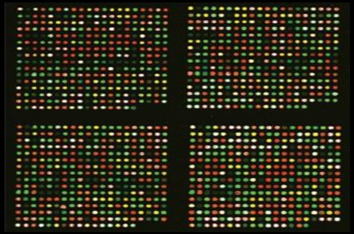
# Robust Traceability from Trace Amounts

Cynthia Dwork, Adam Smith, Thomas Steinke,
Jonathan Ullman, Salil Vadhan

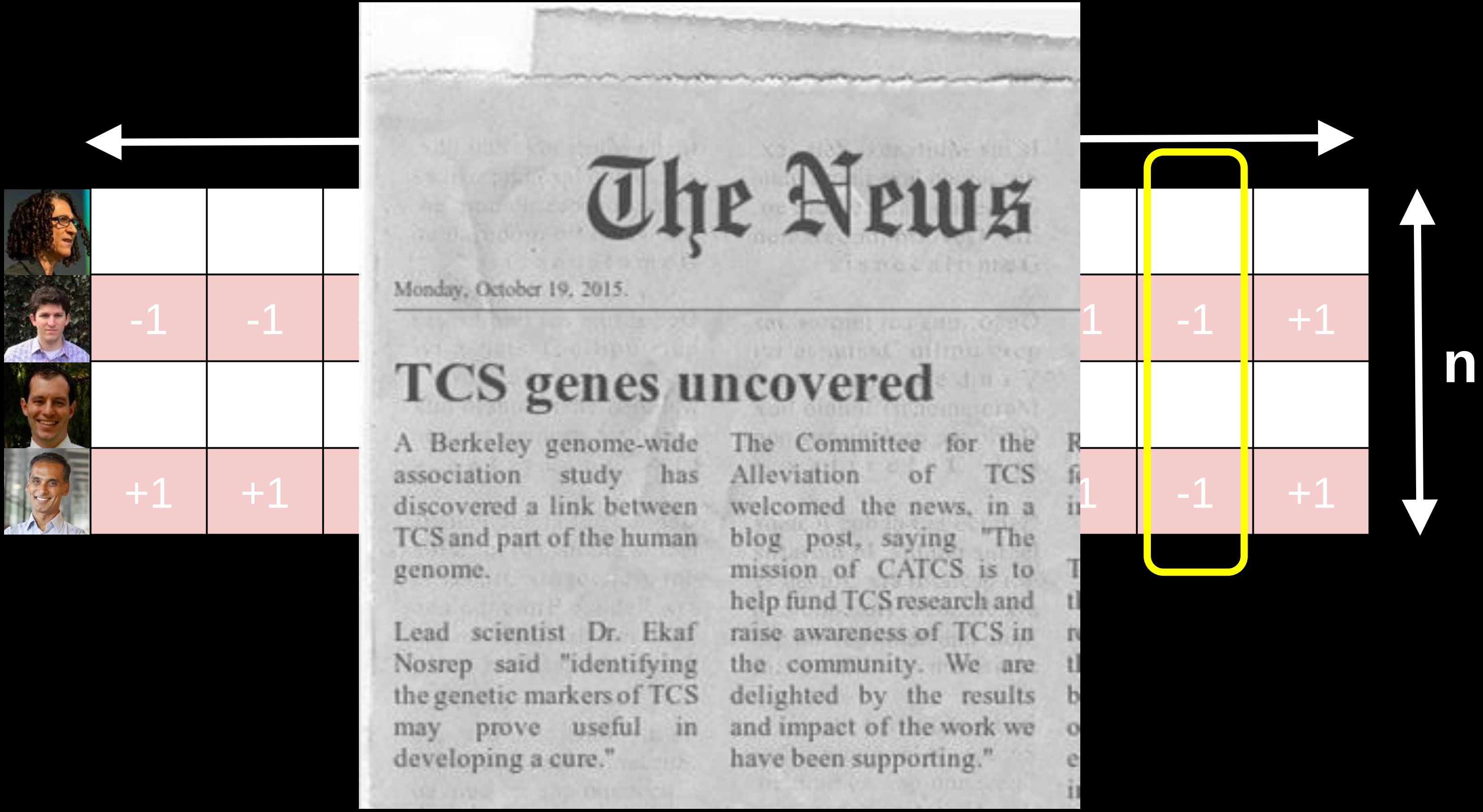
# A motivating story

Is TCS related to a genetic mutation?

**Case Group**

# What the data looks like (in theory)




The News

Monday, October 19, 2015.

## TCS genes uncovered

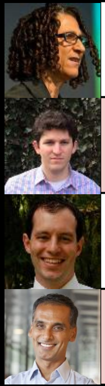
A Berkeley genome-wide association study has discovered a link between TCS and part of the human genome.

Lead scientist Dr. Ekaf Nosrep said "identifying the genetic markers of TCS may prove useful in developing a cure."

The Committee for the Alleviation of TCS welcomed the news, in a blog post, saying "The mission of CATCS is to help fund TCS research and raise awareness of TCS in the community. We are delighted by the results and impact of the work we have been supporting."

# Privacy concerns

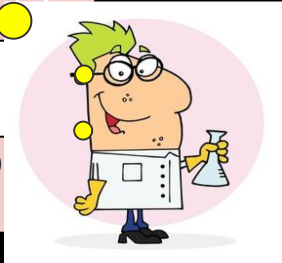
Can I share my data with other TCS researchers?



-1 -1 -1 +1

+1 +1 -1 -1

That’s not OK with us!



Perhaps I can just remove identifiers to protect subjects.

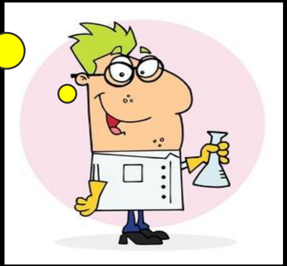
# Aggregate data

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
|  | | | | | | | | | | | |
|  | -1 | -1 | +1 | +1 | -1 | +1 | +1 | -1 | +1 | -1 | +1 |
|  | | | | | | | | | | | |
|  | +1 | +1 | -1 | +1 | -1 | -1 | +1 | -1 | -1 | -1 | +1 |
| | +.5 | 0 | -.5 | +.5 | 0 | 0 | +1 | -.5 | 0 | -1 | +.5 |

# Aggregate data

I can safely share the aggregates, right?

Wrong!
[HSR+08, SOJH09, DN03, BUV14, …, This Work]



| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| +.5 | 0 | -.5 | | | | | | 0 | -1 | +.5 |

Led to changes in how NIH deals with releasing genetic data.

# Fundamental law of information recovery

Releasing “overly accurate” estimates of “too many” aggregate statistics is not private.

[DN03,DMT07,HSR+08,DY08,SOJH09,MN12,BUV14,SU15,...]

# Genetics work [HSR+08, SOJH09, …]

Requires d=Θ(n) attributes.

Given the exact aggregate statistics for the case group

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| +.5 | 0 | -.5 | +.5 | 0 | | | | | | |

m=Θ(n) samples from the same population as the case group.

and the data of one individual

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | |

and a large reference sample,

“Tracing”

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| +.3 | -.1 | -.1 | +.9 | 0 | [illegible] | [illegible] | -.4 | -.1 | -.8 | +.2 |

I can determine whether that individual is in the case group.

# Differential privacy

By releasing <u>approximate</u> instead of <u>exact</u> aggregate statistics, we can prevent tracing (and other privacy attacks)

for up to $d=\tilde{\Theta}(n^2)$ attributes by using differential privacy [DMNS06, DKM+06,...].

# Limits of differential privacy [CFN94,BS95,Tar03,BUV14,SU15,...]

“Fingerprinting codes”

Given approximate aggregate statistics for the case group

Motivating Question: Is tracing possible when the database comes from a realistic distribution and the tracer has realistic side-information?

assuming $d \geq \tilde{O}(n^2)$

and an artificial population.

strong assumptions

# Our results

Given approximate aggregate statistics for the case group

and a single reference sample,

I can identify at least one person in the case group with high probability

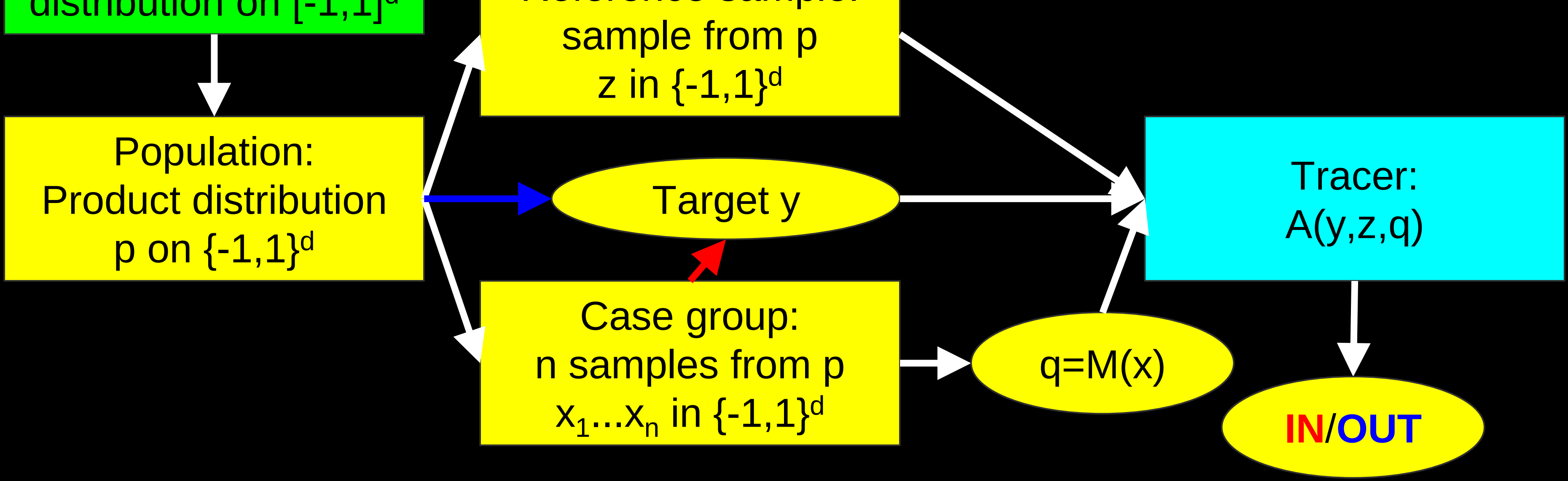
assuming $d \geq \tilde{O}(n^2)$

and a population drawn from a rich family of distributions.

# The model

Accuracy assumption:

$$\left\| q - \frac{1}{n}\sum_{i=1}^{n} x_i \right\| \leq \alpha$$




Our result: If y is **OUT**, tracer says **OUT** whp. Whp, for some of $y=x_i$, tracer says **IN**.

# Comparison

“Natural fingerprinting codes”

| Approach | “Statistics” [HSR+08,SOJH09] | “Fingerprinting” [CFN94,BS95,T03,BUV14] | This work |
|---|---|---|---|

# Comparison

“Natural fingerprinting codes”

| Approach | “Statistics” [HSR+08,SOJH09] | “Fingerprinting” [CFN94,BS95,T03,BUV14] | This work |
|---|---|---|---|
| Accuracy | Exact | Approximate | Approximate |

# Comparison

“Natural fingerprinting codes”

| Approach | “Statistics” [HSR+08,SOJH09] | “Fingerprinting” [CFN94,BS95,T03,BUV14] | This work |
|---|---|---|---|
| Accuracy | Exact | Approximate | Approximate |
| Reference | $m=\Theta(n)$ | $m\cong\infty$ | $m=1$ |

# Comparison

"Natural fingerprinting codes"

| Approach | "Statistics" [HSR+08,SOJH09] | "Fingerprinting" [CFN94,BS95,T03,BUV14] | This work |
|---|---|---|---|
| Accuracy | Exact | Approximate | Approximate |
| Reference | $m=\Theta(n)$ | $m\cong\infty$ | $m=1$ |
| Dimension | $d\geq O(n)$ | $d\geq\tilde{O}(n^2)$ | $d\geq\tilde{O}(n^2)$ |

# Comparison

“Natural fingerprinting codes”

| Approach | “Statistics” [HSR+08,SOJH09] | “Fingerprinting” [CFN94,BS95,T03,BUV14] | This work |
|---|---|---|---|
| Accuracy | Exact | Approximate | Approximate |
| Reference | $m=\Theta(n)$ | $m \cong \infty$ | $m=1$ |
| Dimension | $d \geq O(n)$ | $d \geq \tilde{O}(n^2)$ | $d \geq \tilde{O}(n^2)$ |
| Identifies | Everyone | 1 person | 1 person |

# Comparison

"Natural fingerprinting codes"

| Approach | "Statistics" [HSR+08,SOJH09] | "Fingerprinting" [CFN94,BS95,T03,BUV14] | This work |
|---|---|---|---|
| Accuracy | Exact | Approximate | Approximate |
| Reference | $m=\Theta(n)$ | $m\cong\infty$ | $m=1$ |
| Dimension | $d\geq O(n)$ | $d\geq\tilde{O}(n^2)$ | $d\geq\tilde{O}(n^2)$ |
| Identifies | Everyone | 1 person | 1 person |
| Population | No assumption | "Artificial" | "Rich Family" |

# Extension

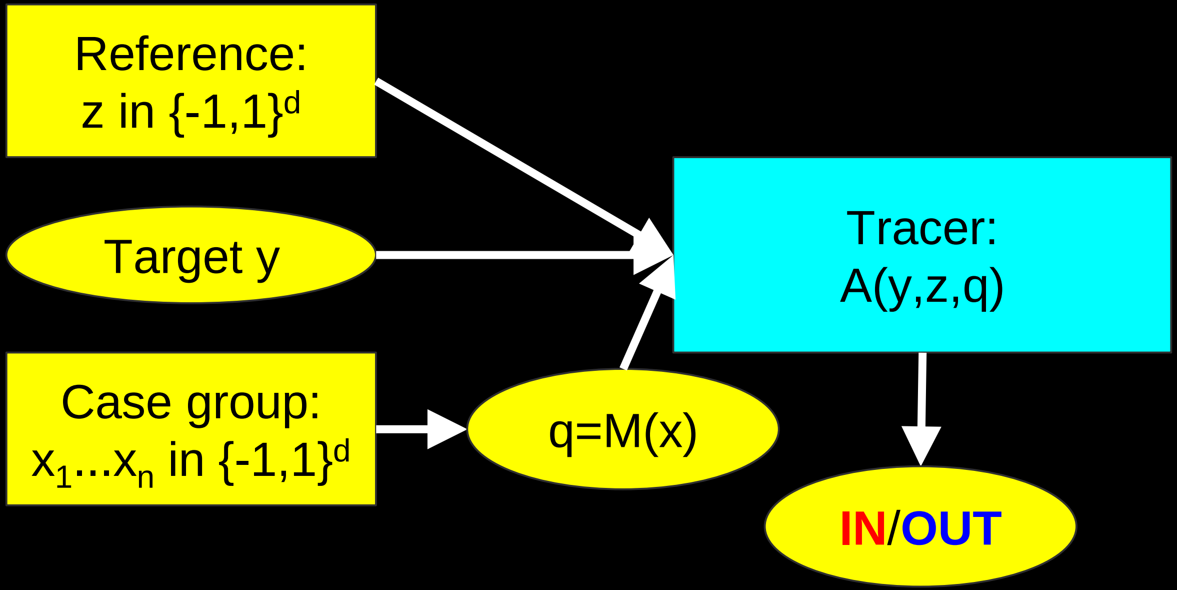
Accuracy assumption:

$$\left\| q - \frac{1}{n} \sum_{i=1}^{n} x_i \right\| \leq \alpha$$

| Approach | "Statistics" [HSR+08,SOJH09] | "Fingerprinting" [CFN94,BS95,T03, BUV14,SU15] | This work | |
|---|---|---|---|---|
| Accuracy | Exact | Approximate | Approximate | $\alpha \geq 1/\sqrt{n}$ |
| Reference | $m=\Theta(n)$ | $m \cong \infty$ | $m=1$ | $m=O(\log(n)/\alpha^2)$ |
| Dimension | $d \geq O(n)$ | $d \geq \tilde{O}(n^2)$ | $d \geq \tilde{O}(n^2)$ | $d \geq \tilde{O}(\alpha^2 n^2)$ |
| Identifies | Everyone | 1 person | 1 person | $\Omega(1/\alpha^2)$ people |
| Population | No assumption | "Artificial" | "Rich Family" | "Rich Family" |

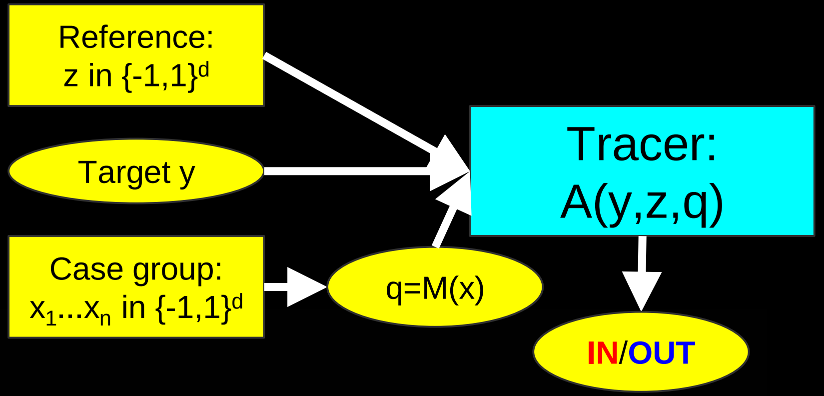
Smoothly interpolates between extremes

# Our tracer



Reference:
z in {-1,1}$^d$
Target y
Case group:
$x_1$...$x_n$ in {-1,1}$^d$
q=M(x)
Tracer:
A(y,z,q)
IN/OUT

# Very simple tracer




$A(y, z, q)$:

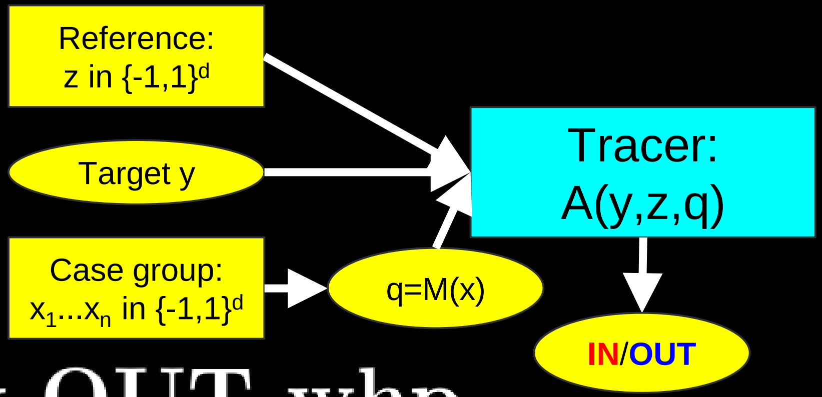
Input: $y, z \in \{-1, 1\}^d$, $q \in [-1, 1]^d$.

Compute $s = \langle y - z, q\rangle = \langle y, q\rangle - \langle z, q\rangle$.

If $s \geq \sqrt{8d \log(1/\delta)}$, output IN; else output OUT.

"Is y more correlated with q than z?" [HSR+08]

# Why does our tracer work?



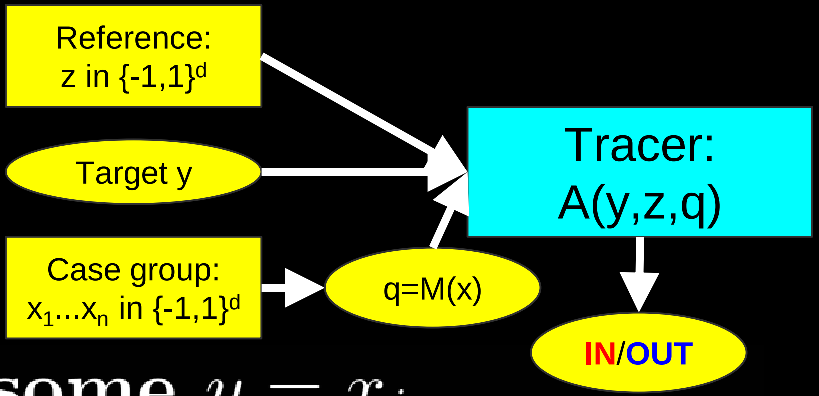

**Soundness: If OUT, say OUT whp.**

$\mathbb{E}[y] = \mathbb{E}[z]$ and $q$ is independent from $y$ and $z$.

Thus $\mathbb{E}[\langle y - z, q \rangle] = 0$.

Chernoff: $\mathbb{P}\left[\langle y - z, q \rangle < \sqrt{8d \log(1/\delta)}\right] \geq 1 - \delta$.

# Why does our tracer work?




**Completeness: Say IN for some $y = x_i$.**

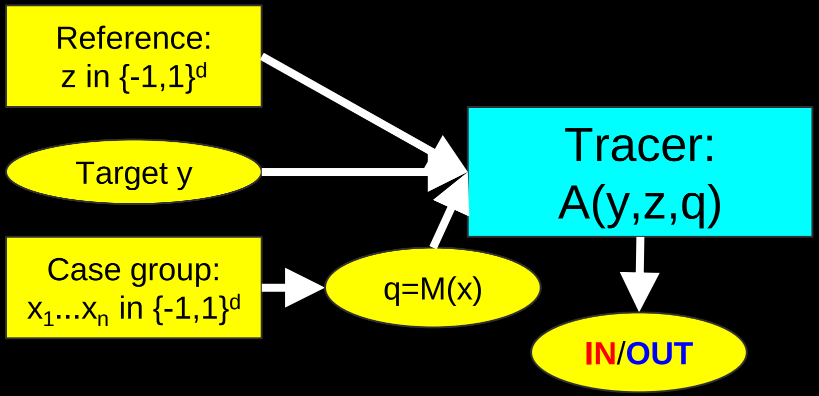
Lemma: $q$ accurate $\implies \sum_{i=1}^{n} \mathbb{E}\left[\langle x_i - z, q\rangle\right] \geq \Omega(d)$.

Azuma: $\mathbb{P}\left[\sum_{i=1}^{n} \langle x_i - z, q\rangle \geq \Omega(d)\right] \geq 1 - \delta$.

Thus $\exists i \; \langle x_i - z, q\rangle \geq \Omega(d/n) \geq \sqrt{8d\log(1/\delta)}$.

So, if $d = O(n^2 \log(1/\delta))$, say IN for some $y = x_i$ whp.

# Why does our tracer work?




Lemma: $q$ accurate $\implies \sum_{i=1}^{n} \mathbb{E}\left[\langle x_i - z, q\rangle\right] \geq \Omega(d)$.

Intuition ($d = 1$): $q = M(x)$. If $x_1 = \cdots = x_n = b \in \{\pm 1\}$,

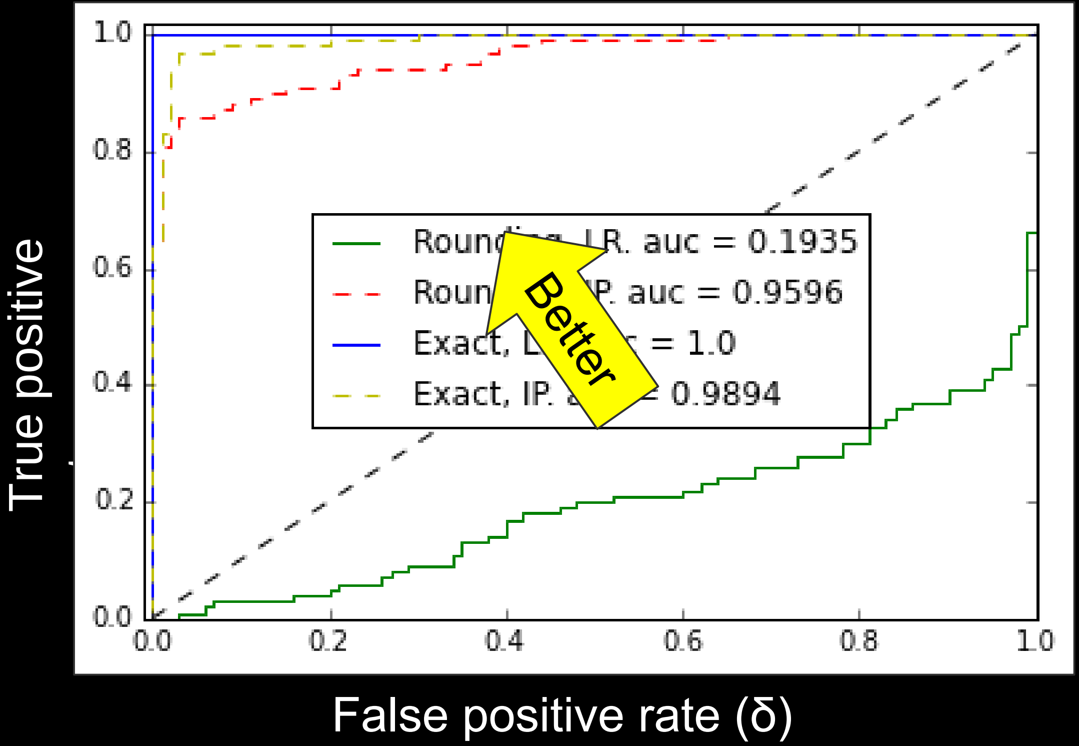
then $q \approx b$. So "on average" changing $x_i$ changes $q$ by $2/n$.

If bias is well-spread, we get correlation on average.

# Comparing exact and approximate statistics

Simulation:
n=100
m=200
d=5000

Exact vs. Rounded




Compare to:
Likelihood ratio test
[SOJH09]

# Conclusion

We provide a <u>simple</u> and <u>robust</u> tracer that needs <u>less auxiliary information</u> than previous work.

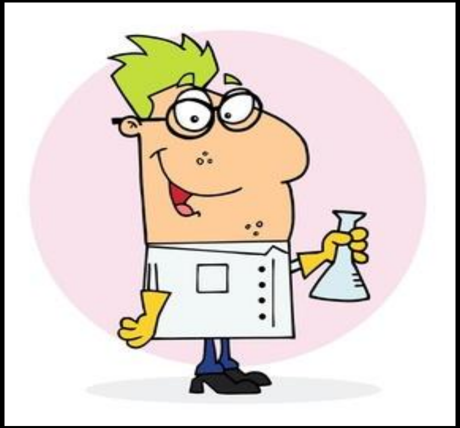
Build on work in genetics and cryptography. Simplified proofs.

Clearer picture of what can(not) be released privately.

Tells us differential privacy correctly quantifies privacy here.

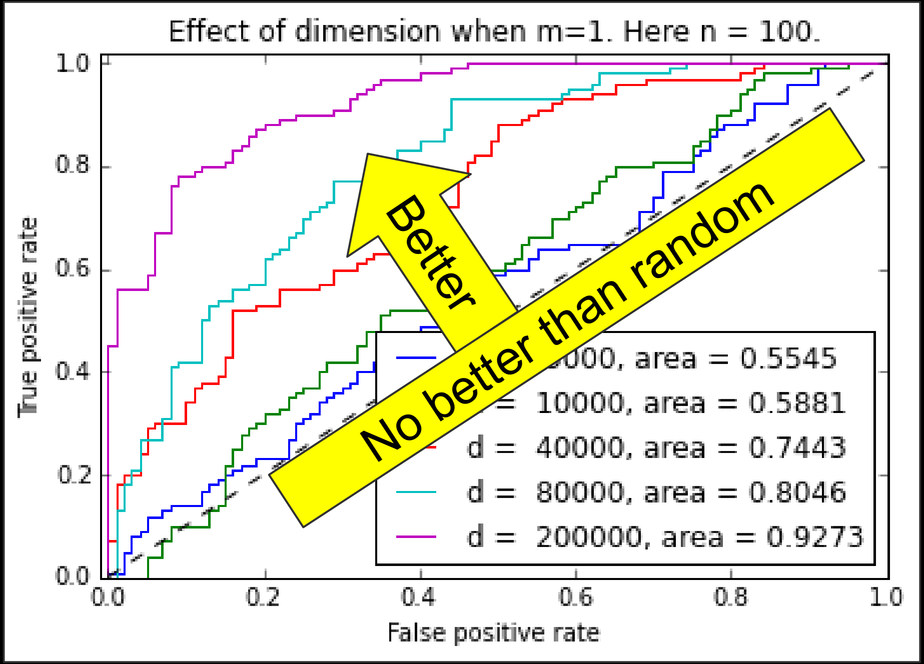
Releasing "overly accurate" estimates of "too many" aggregate statistics is not private.

# Experimental results



# Experimental results

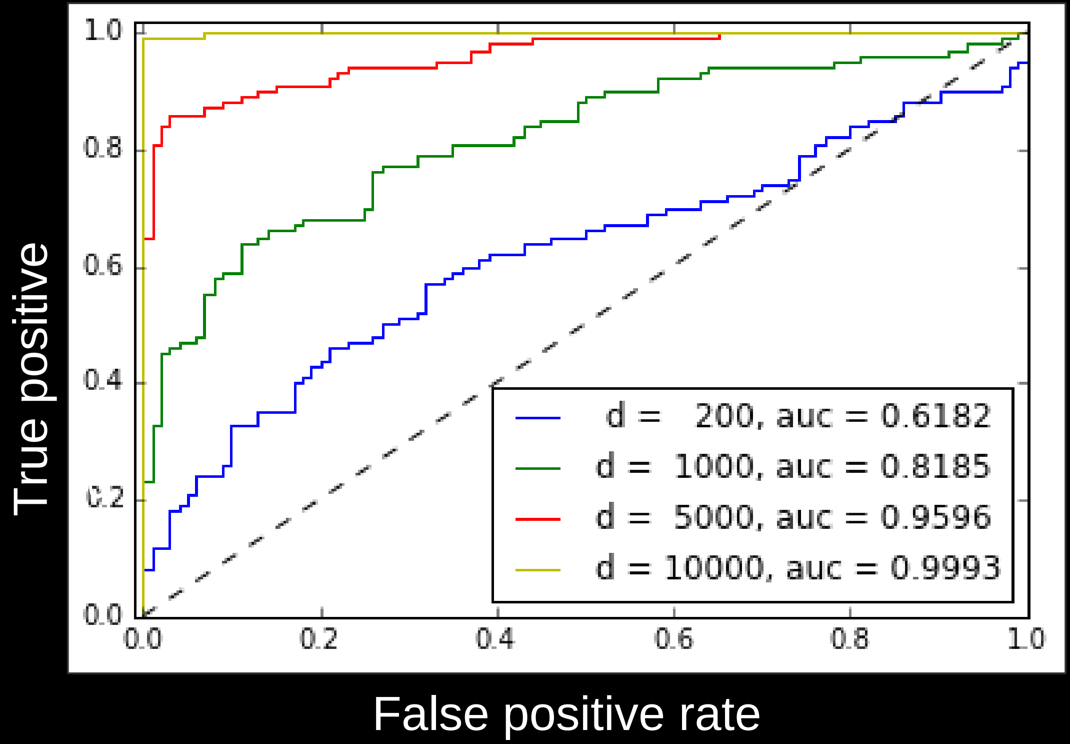
Simulation:
n=100
m=1
Rounded
to 0.1



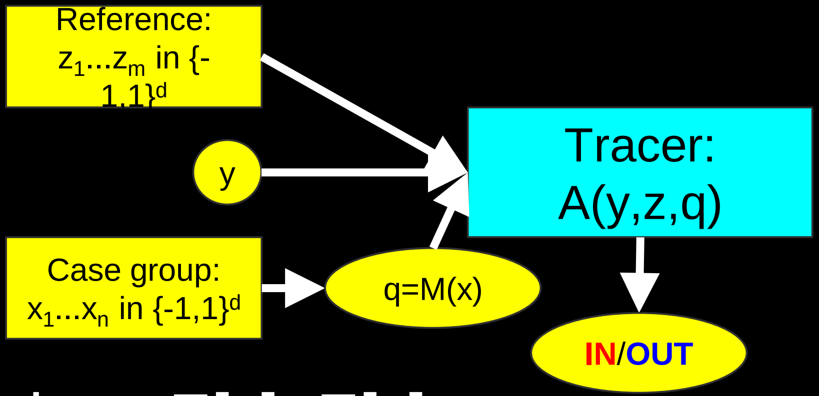

(Here we are varying the IN/OUT threshold.)

# More experimental results

Simulation:
n=100
m=200
Rounded
to 0.1



# Why does our tracer work?



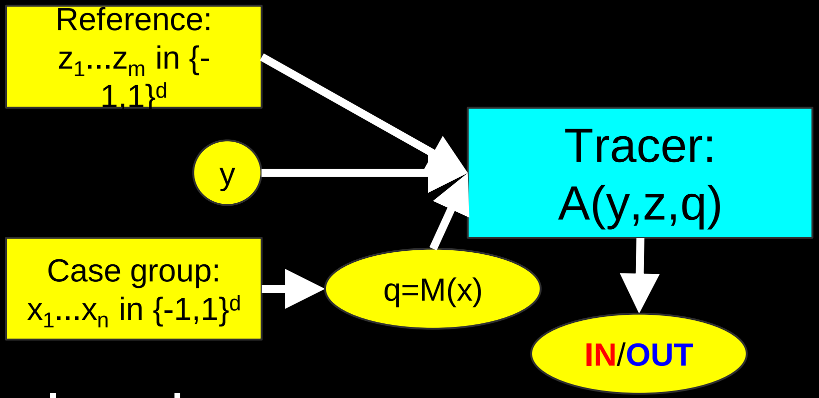

Soundness:

In **OUT** case, y-z and q are independent. E[y]=E[z].

Thus s=<y-z,q>≅0 whp by Chernoff-Hoeffding bound.

P[A(y,z,q) says **OUT**] ≥ 1-$\delta$.

# Why does our tracer work?




## <u>Completeness:</u>

In **IN** case, $y=x_i$ and $q=M(x)$ are correlated.

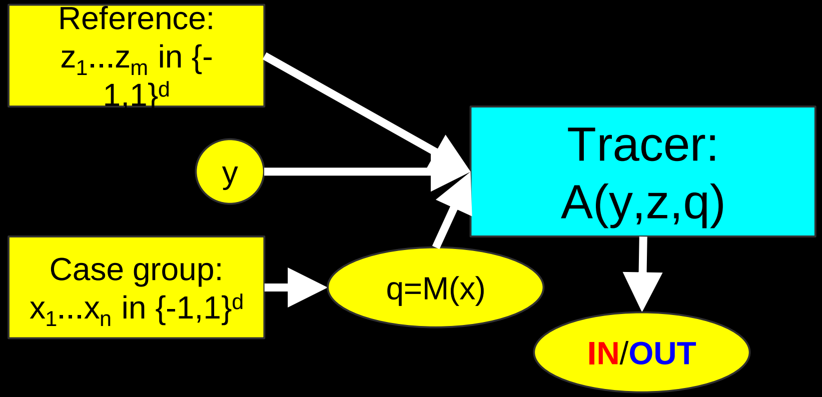
Lemma: $\Sigma_i \, E[\langle x_i - z, q\rangle] \geq \Omega(d)$.

Whp $\Sigma_i \langle x_i - z, q\rangle \geq \Omega(d)$.

Whp $\langle x_i - z, q\rangle \geq \Omega(d/n)$ for at least one $i \in \{1,2,\ldots,n\}$.

$P[\exists i \; A(x_i, z, q) = \textbf{IN}] \geq 1 - \delta$.

# Very simple tracer (m=1 case)




A(y,z,q):

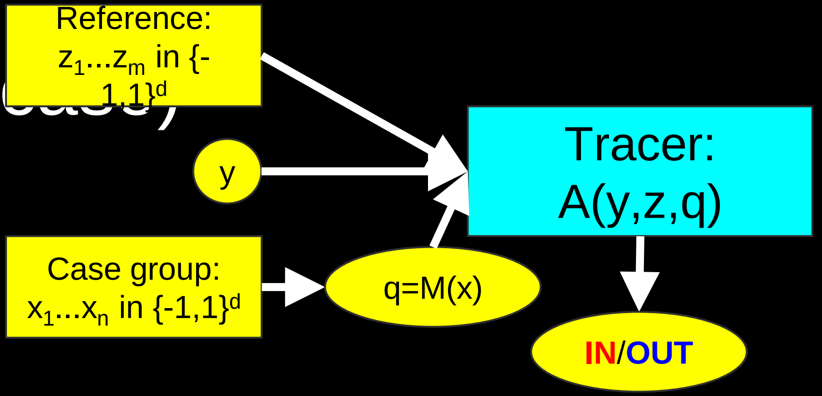
Input: $y,\ z \in \{-1,1\}^d$, $q \in [-1,1]^d$

Compute $s=\langle y-z,q\rangle=\langle y,q\rangle-\langle z,q\rangle$

If $s \geq \sqrt{8\ d\ \log(1/\delta)}$, output **IN**; otherwise output **OUT**.

"Is y more correlated with q than z?"

# Simple tracer ($m=O(\log(n)/\alpha^2)$ case)




A(y,z,q):

Input: $y, z_0, z_1, \ldots, z_m \in \{-1,1\}^d$, $q \in [-1,1]^d$

Compute $s=\langle y-z_0,[q-\hat{z}]\rangle$

If $s \geq 4\alpha\sqrt{d \log(1/\delta)}$, output **IN**; otherwise output **OUT**.

# More precisely...

“Tracing”

Given the aggregate statistics for the case group

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| +2 | 0 | -2 | +2 | 0 | 0 | [illegible] | -2 | 0 | -4 | +2 |

and the data of one individual,

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | |

I can determine whether that individual is in the case group.*

Requires auxiliary information and accuracy assumption.

# Tightness

Differential Privacy [DMNS06,DKM+06,...]

There exists a method to release α-approximate aggregate statistics

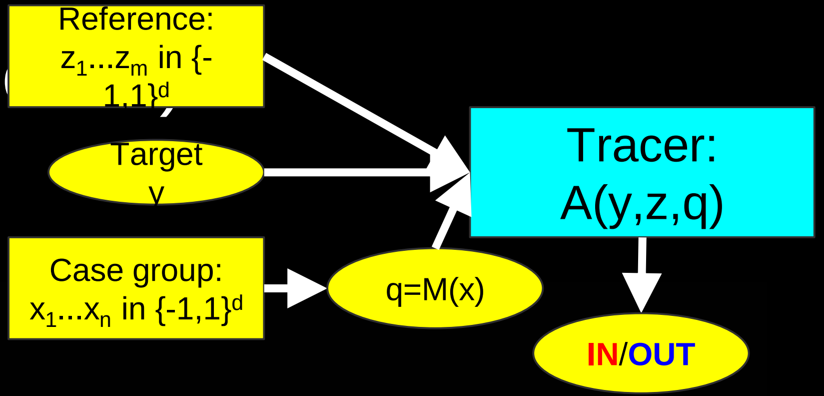
even when $d=\tilde{O}(\alpha^2 n^2)$

that prevents tracing and other attacks.

i.e. Differential privacy is tight.

Corollary: We cannot do better than differential privacy in this setting.

# Simple tracer (m=O(log(n)/$\alpha^2$)




$A_\alpha(y, z, q)$:

Input: $y, z_0, z_1, \cdots, z_m \in \{-1,1\}^d$, $q \in [-1,1]^d$.

Let $\overline{z} = \frac{1}{m}\sum_{i=1}^{m} z_i$.

Compute $s = \langle y - z_0, \lceil q - \overline{z} \rfloor \rangle$.

If $s \geq 4\alpha\sqrt{d\log(1/\delta)}$, output IN; else output OUT.