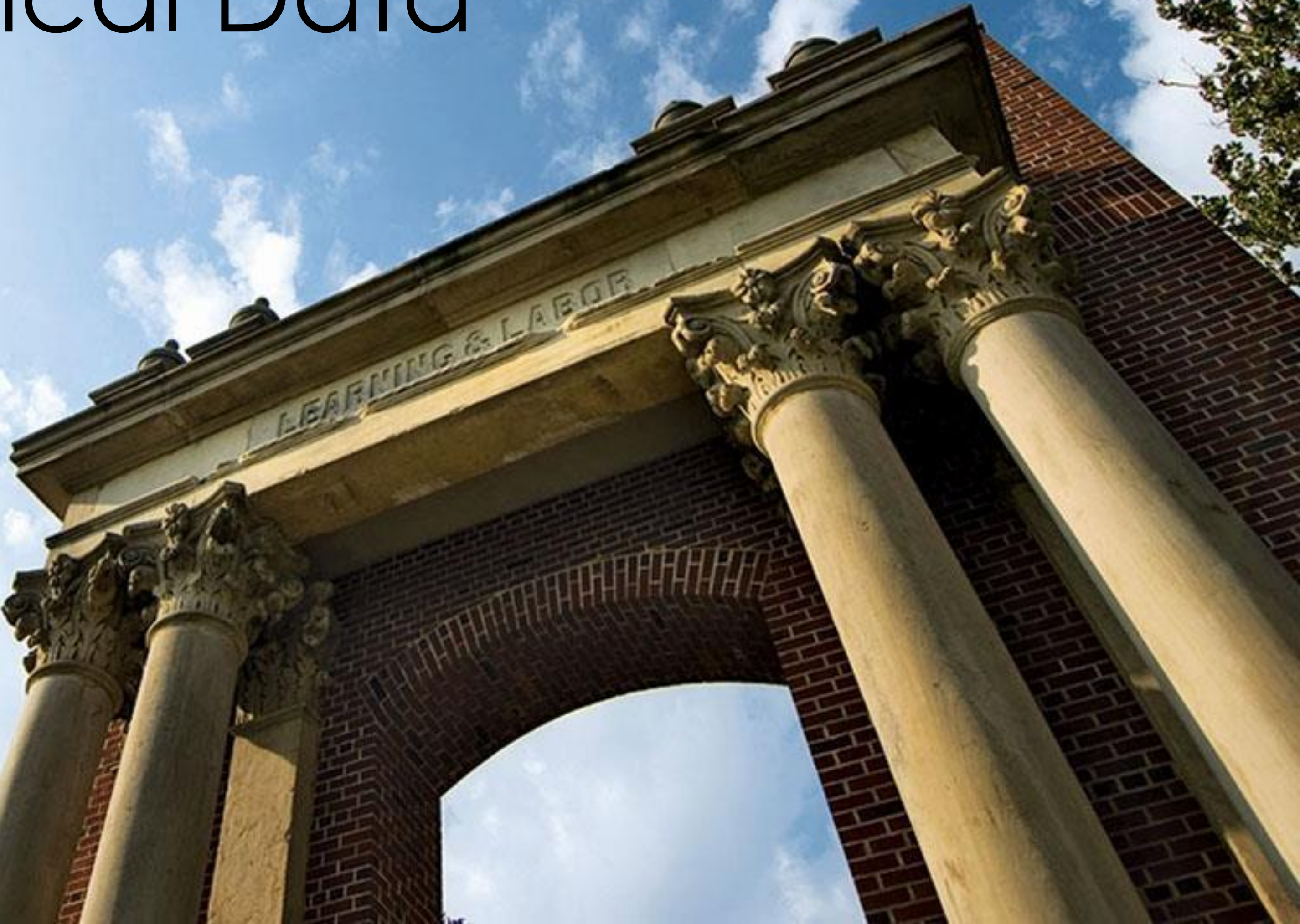
The Fundamental Limits of Statistical Data Privacy

Peter Kairouz
ECE Department
UIUC

ILLINOIS
illinois.edu



30 YEARS AGO

Pre-internet

Human to human

THEN CAME THE INTERNET

Pre-internet

Internet of content

Human to human

World Wide Web

Google

YAHOO!

+ smart networks, IT platforms and services

AND THEN THE INTERNET GOT BETTER

Pre-internet | **Internet of content** | **Smart internet**

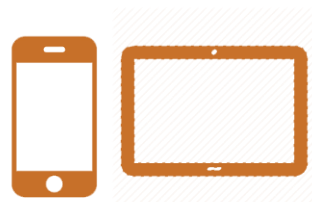
Human to human

World Wide Web

Google

YAHOO!

Smart phones and tablets



+ *smart* *networks, IT platforms and services*

+ *smart* *devices*

AND BETTER

Pre-internet	Internet of content	Smart internet	Internet of people
Human to human	***World Wide Web***	***Smart phones and tablets***	***Social Media***
SMS	Google YAHOO!		

+ *smart* networks, IT platforms and services

+ *smart* devices

+ *smart* phones and applications

UNPRECEDENTED LEVEL OF CONNECTIVITY

Pre-internet	Internet of content	Smart internet	Internet of people	Internet of things
Human to human	***World Wide Web***	***Smart phones and tablets***	***Social Media***	***Machine to machine***

+ *smart networks, IT platforms and services*

+ *smart devices*

+ *smart phones and applications*

+ *smart devices, objects and data*

WE'RE BEING WATCHED!

Pre-internet | Internet of content | Smart internet | Internet of people | Internet of things

SONY
hacked.again

TARGET

IT'S OKAY, OUR DATA IS ENCRYPTED

ebay

iCloud

CHASE

DON’T RELY EXCLUSIVELY ON ENCRYPTION

EVEN IF YOU'RE CAREFUL, THINGS CAN GO WRONG

de-anonymizing Netflix watch histories

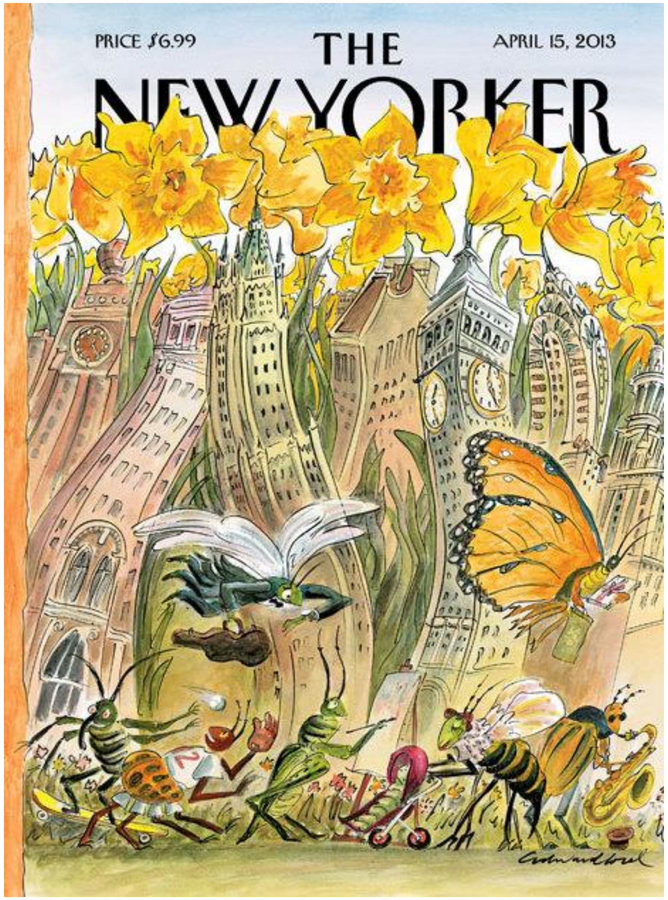
identifying surnames and ages from anonymized genomes

from anonymous faces to social security numbers

WE NEED CONTEXT FREE PRIVACY GUARANTEES

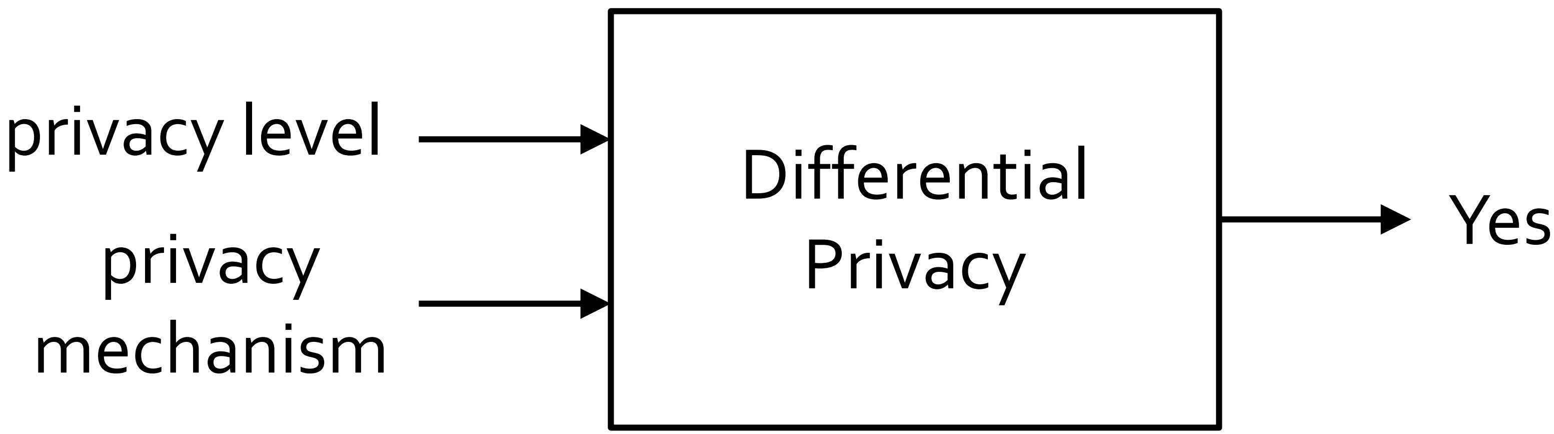
THE ULTIMATE PROTECTION

"the future of privacy is lying"



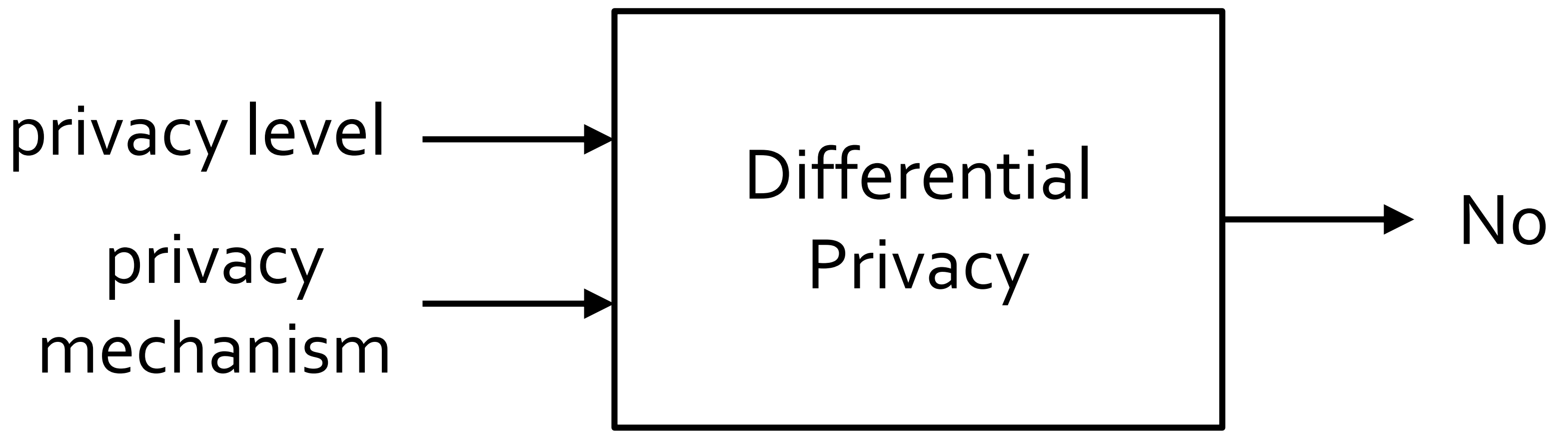
lying = adding noise to data

DIFFERENTIAL PRIVACY



[Dinur et al. 2003, *Dwork et al.* 2006]

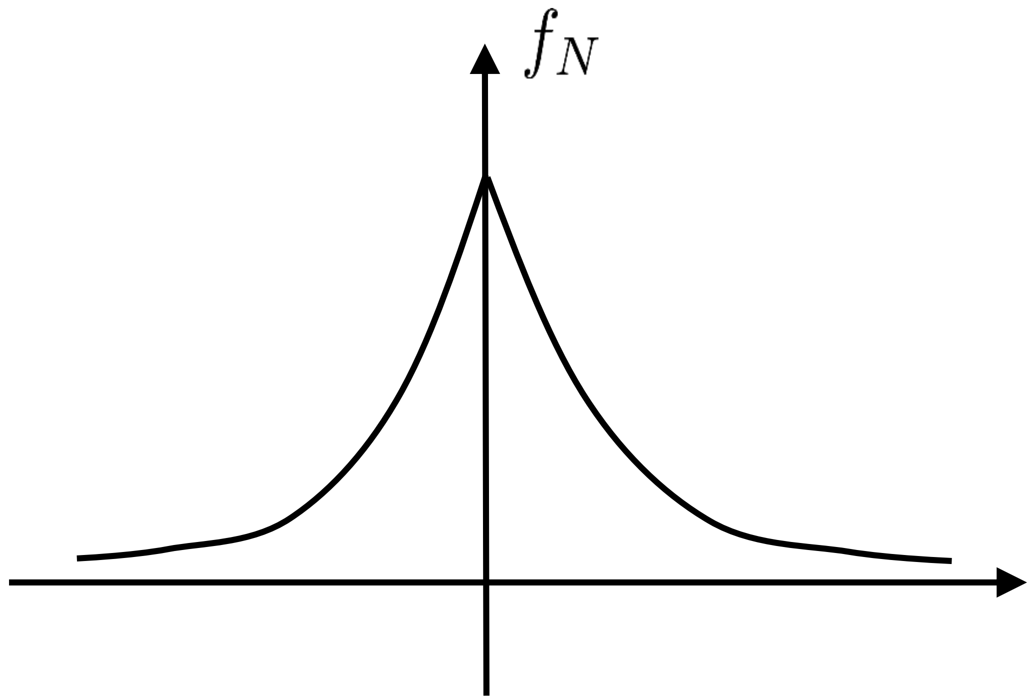
DIFFERENTIAL PRIVACY



[Dinur et al. 2003, *Dwork et al.* 2006]

DIFFERENTIAL PRIVACY

Laplace Mechanism



standard deviation proportional to privacy level

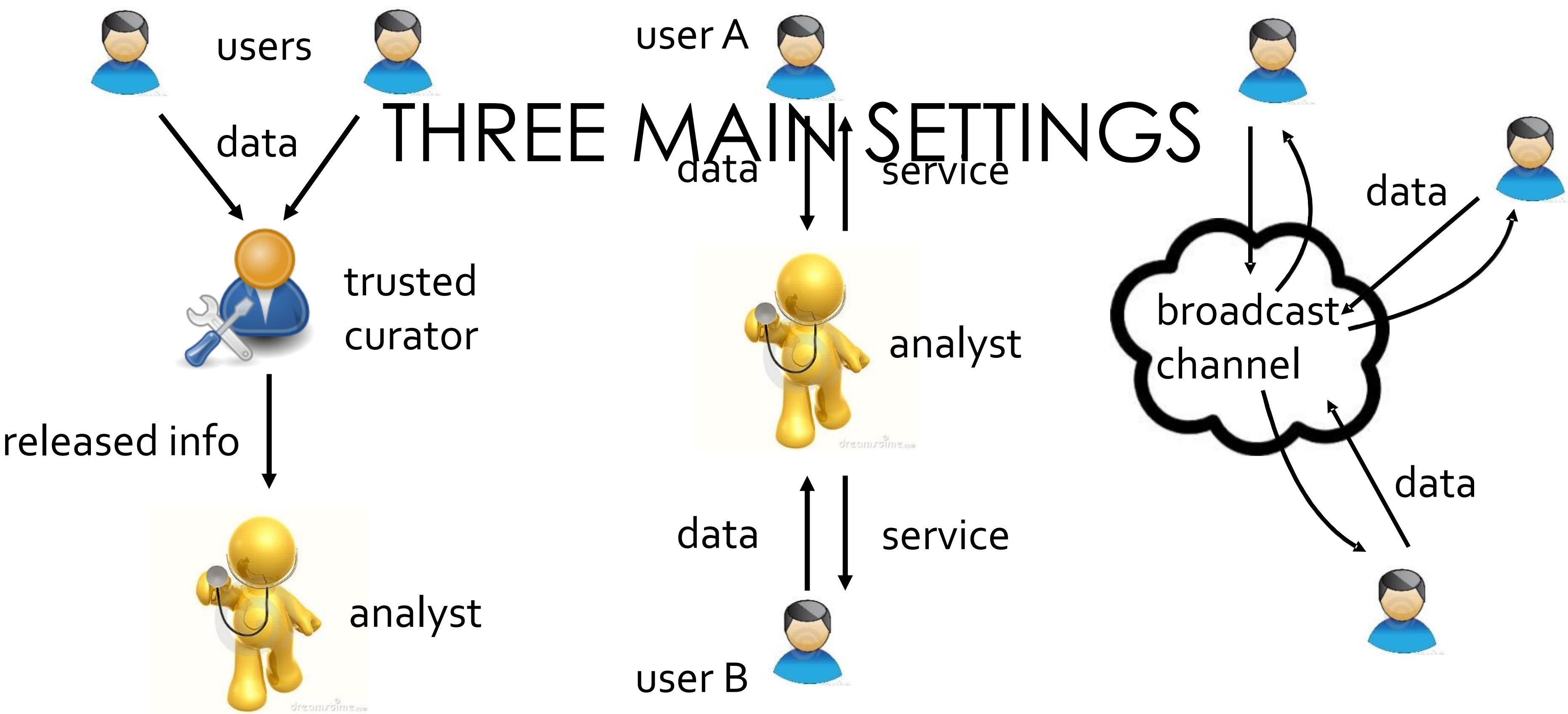
PRIVACY VS. UTILITY

GIVEN A PRIVACY LEVEL

FIND THE "BEST" PRIVACY MECHANISM UNDER DIFFERENTIAL PRIVACY

THREE MAIN SETTINGS

Global Privacy	Local Privacy	Multi-Party Privacy



OUR MAIN RESULT

Global Privacy	Local Privacy	Multi-Party Privacy

privacy mechanisms that achieve the best privacy-utility tradeoff

[NIPS 14, NIPS 15, ICML 15, TSTSP 15, CISS 16, JMLR 16, TIT 16]

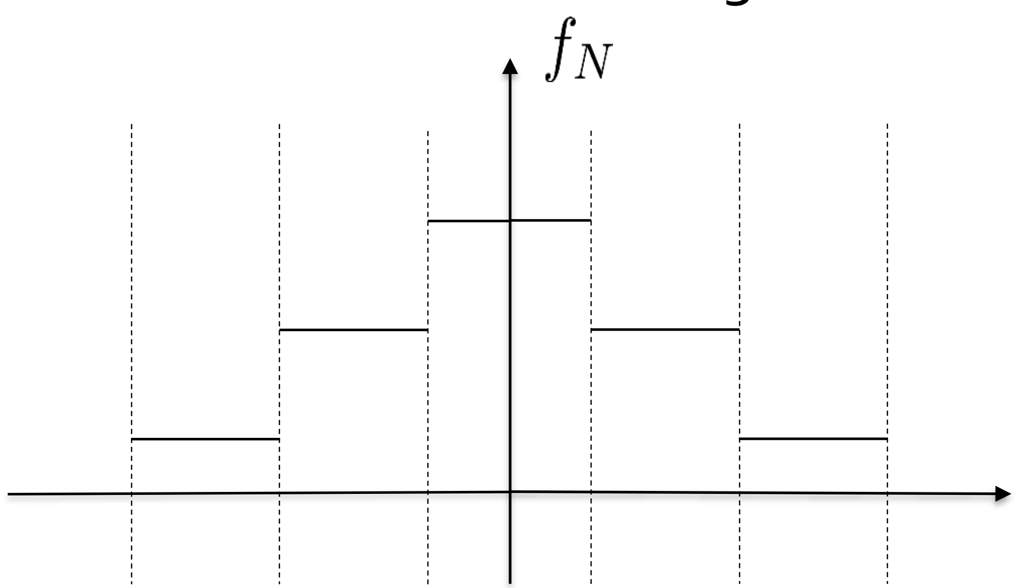
OUR MAIN RESULT

Global Privacy

Local Privacy

Multi-Party Privacy

the optimal mechanisms in all three settings have a **staircase shape**



[NIPS 14, NIPS 15, ICML 15, TSTSP 15, CISS 16, JMLR 16, TIT 16]

STAIRCASE MECHANISMS ARE OPTIMAL

Google

differential privacy

Scholar

About 2,560,000 results (**0.03** sec)

Articles

Case law

My library

Any time

Since 2016

Since 2015

Since 2012

Custom range...

Sort by relevance

Sort by date

include patents

include citations

Create alert

Differential privacy

C Dwork - Automata, languages and programming, 2006 - Springer

Abstract In 1977 Dalenius articulated a desideratum for statistical databases: nothing about an individual should be learnable from the database that cannot be learned without access to the database. We give a general impossibility result showing that a formalization of ...

Cited by 1744 Related articles All 22 versions Web of Science: 293 Cite Save

Differential privacy: A survey of results

C Dwork - Theory and applications of models of computation, 2008 - Springer

Abstract Over the past five years a new approach to **privacy**-preserving data analysis has born fruit [13, 18, 7, 19, 5, 37, 35, 8, 32]. This approach differs from much (but not all!) of the related literature in the statistics, databases, theory, and cryptography communities, in that **...**

Cited by 749 Related articles All 24 versions Cite Save

Mechanism design via **differential privacy**

F McSherry, K Talwar - … of Computer Science, 2007. FOCS'07. …, 2007 - ieeexplore.ieee.org

Abstract We study the role that **privacy**-preserving algorithms, which prevent the leakage of specific information about participants, can play in the design of mechanisms for strategic agents, which must encourage players to honestly report information. Specifically, we **...**

Cited by 573 Related articles All 24 versions Cite Save

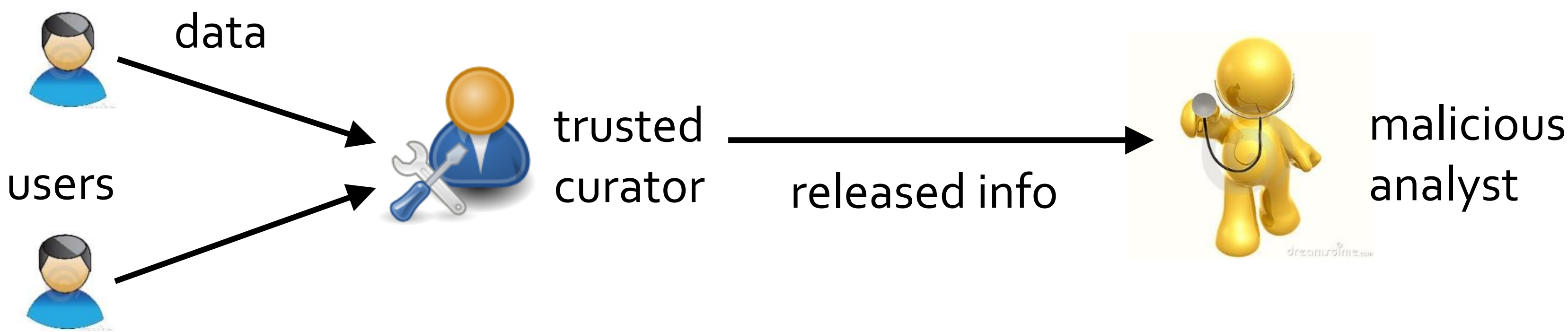
Differential privacy via wavelet transforms

X Xiao, G Wang, J Gehrke - Knowledge and Data Engineering, …, 2011 - ieeexplore.ieee.org

Abstract—**Privacy** preserving data publishing has attracted considerable research interest in recent years. Among the existing solutions, e-**differential privacy** provides the strongest **privacy** guarantee. Existing data publishing methods that achieve e-**differential privacy**, **...**

PART 1/3:
GLOBAL PRIVACY

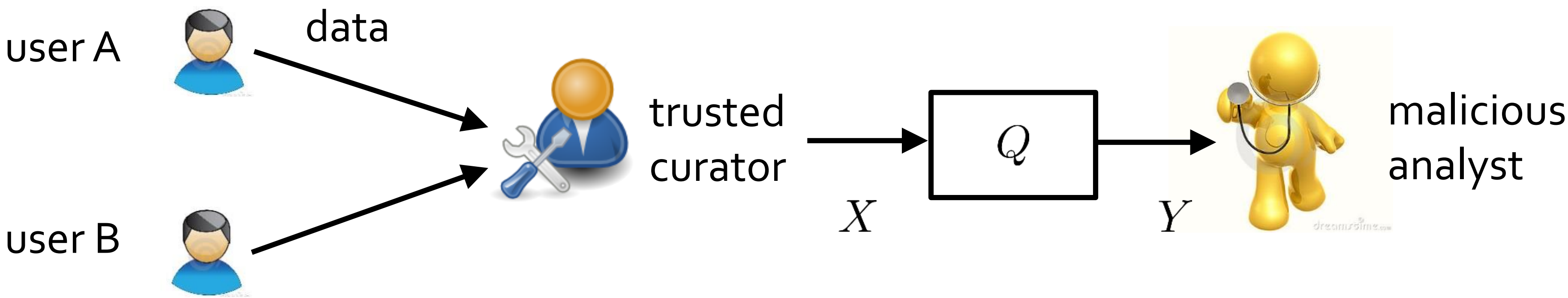
GLOBAL PRIVACY MODEL



NIH

National Institutes
of Health

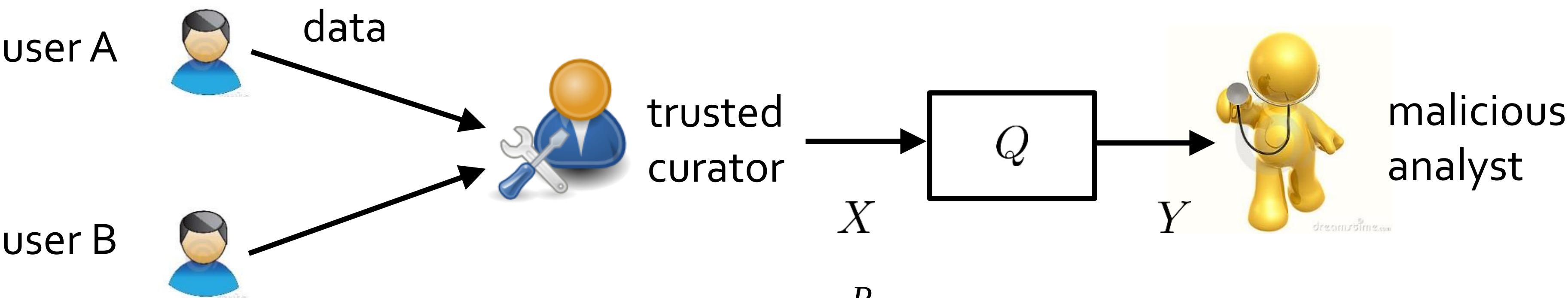
GLOBAL DIFFERENTIAL PRIVACY



$$e^{-\varepsilon} \leq \frac{\mathbb{P}\left(Y \middle| \text{user A present}\right)}{\mathbb{P}\left(Y \middle| \text{user A absent}\right)} \leq e^{+\varepsilon}$$

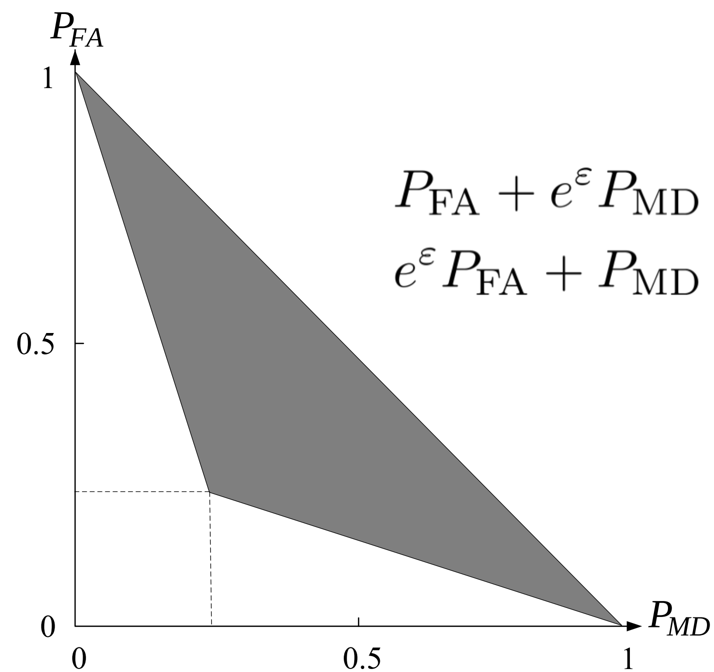
ε controls the level of privacy
large ε, low privacy
small ε, high privacy

OPERATIONAL INTERPRETATION



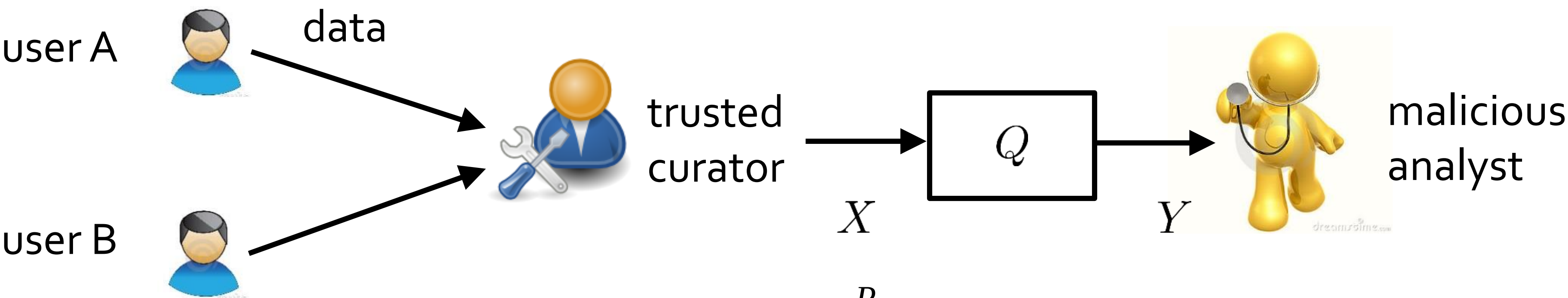
H0: user A is absent

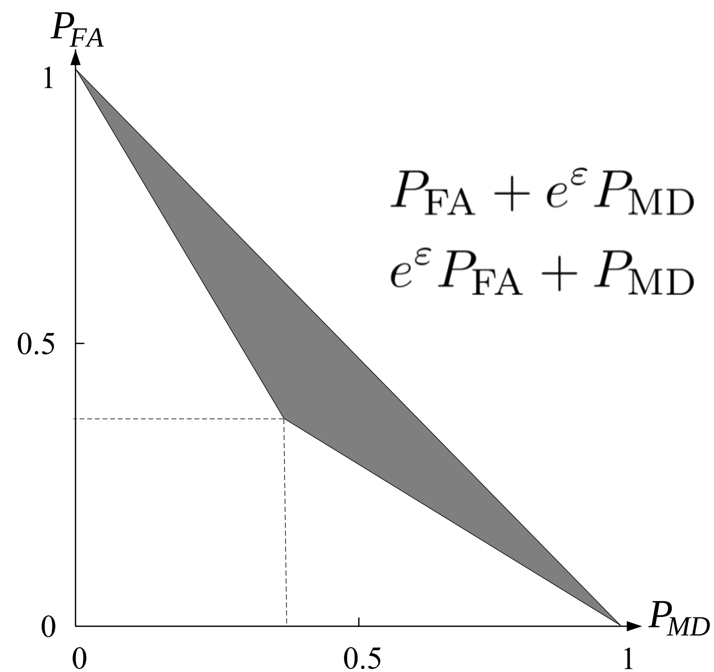
H1: user A is present



$$
\begin{aligned}
P_{\mathrm{FA}} + e^{\varepsilon} P_{\mathrm{MD}} &\geq 1 \\
e^{\varepsilon} P_{\mathrm{FA}} + P_{\mathrm{MD}} &\geq 1
\end{aligned}
$$

OPERATIONAL INTERPRETATION



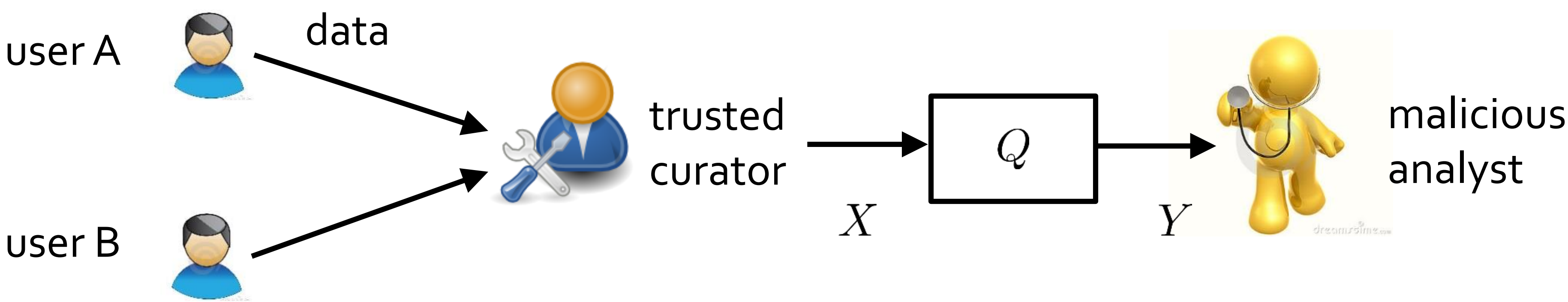


$$
\begin{aligned}
P_{\mathrm{FA}} + e^{\varepsilon} P_{\mathrm{MD}} &\geq 1 \\
e^{\varepsilon} P_{\mathrm{FA}} + P_{\mathrm{MD}} &\geq 1
\end{aligned}
$$

H0: user A is absent

H1: user A is present

PRIVACY-UTILITY TRADEOFF

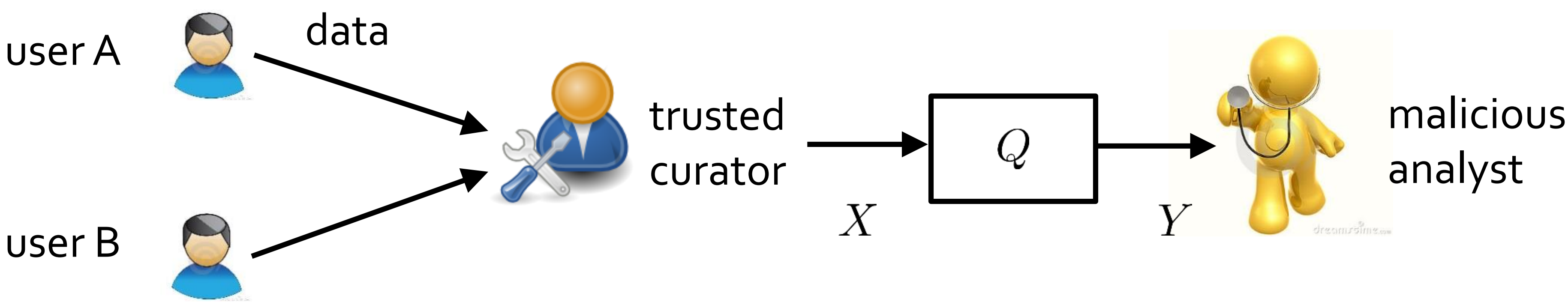


loss = $|X - Y|$

average loss = $\mathbb{E}|X - Y|$

worst case average loss

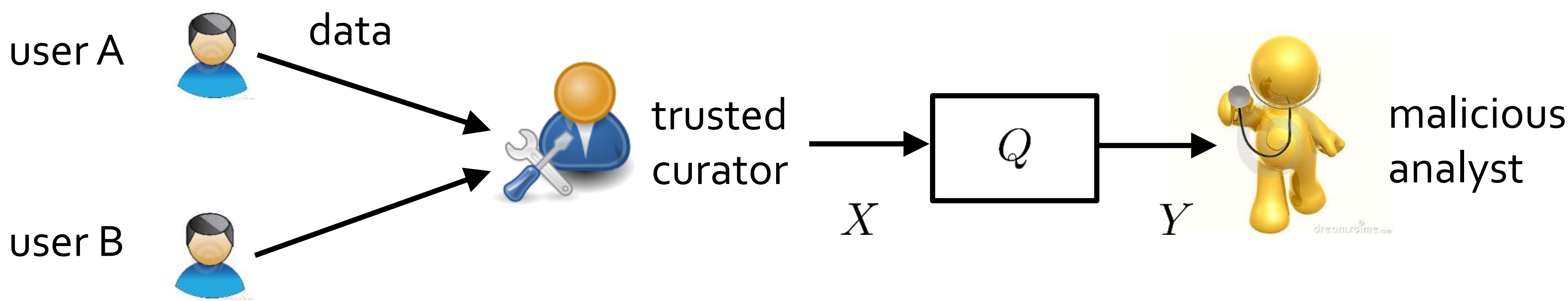
PRIVACY-UTILITY TRADEOFF

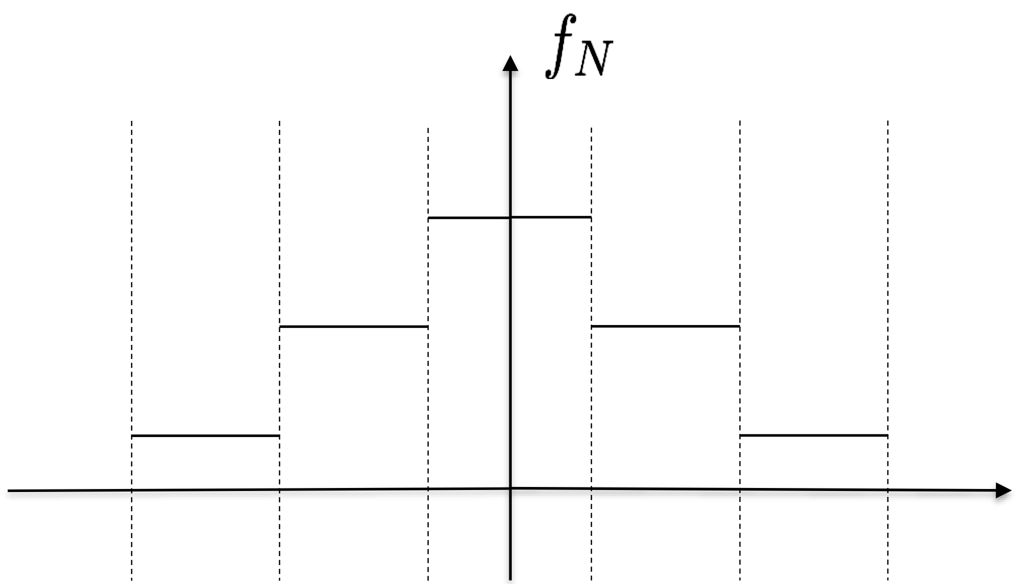


minimize the worst case average loss

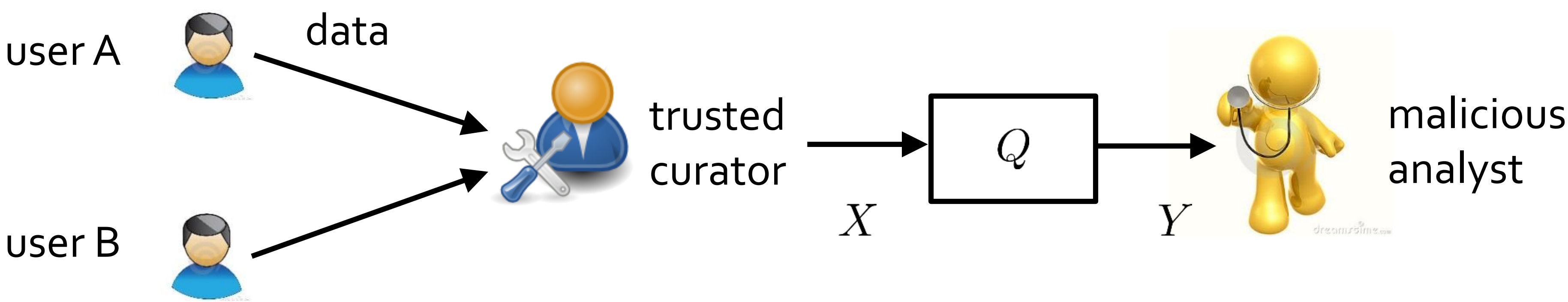
subject to differential privacy

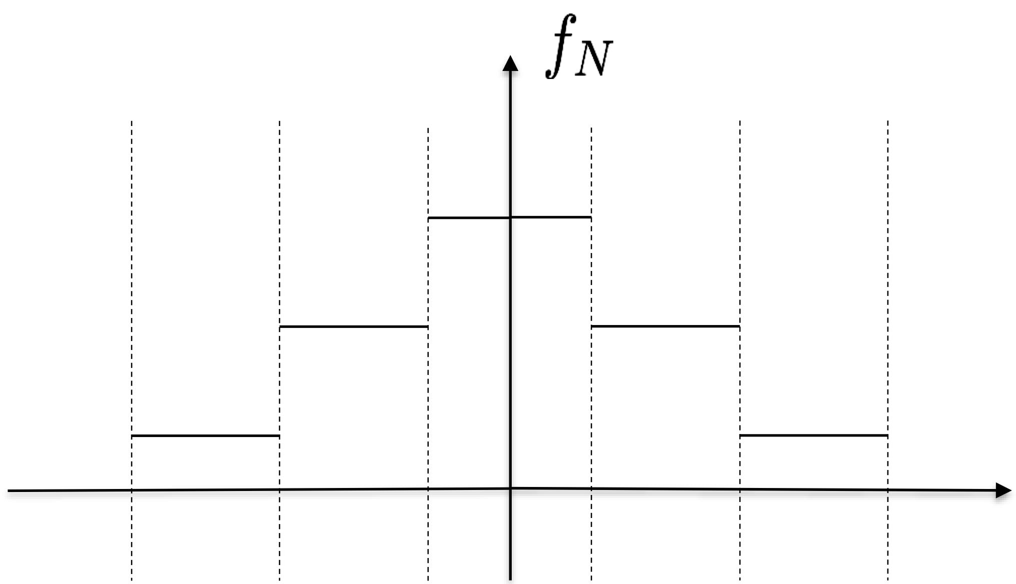
OPTIMALITY OF STAIRCASE MECHANISM



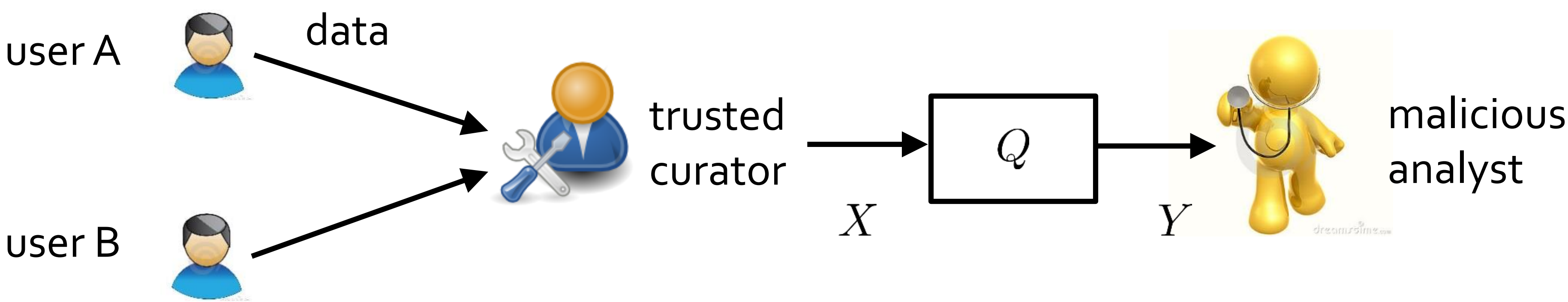


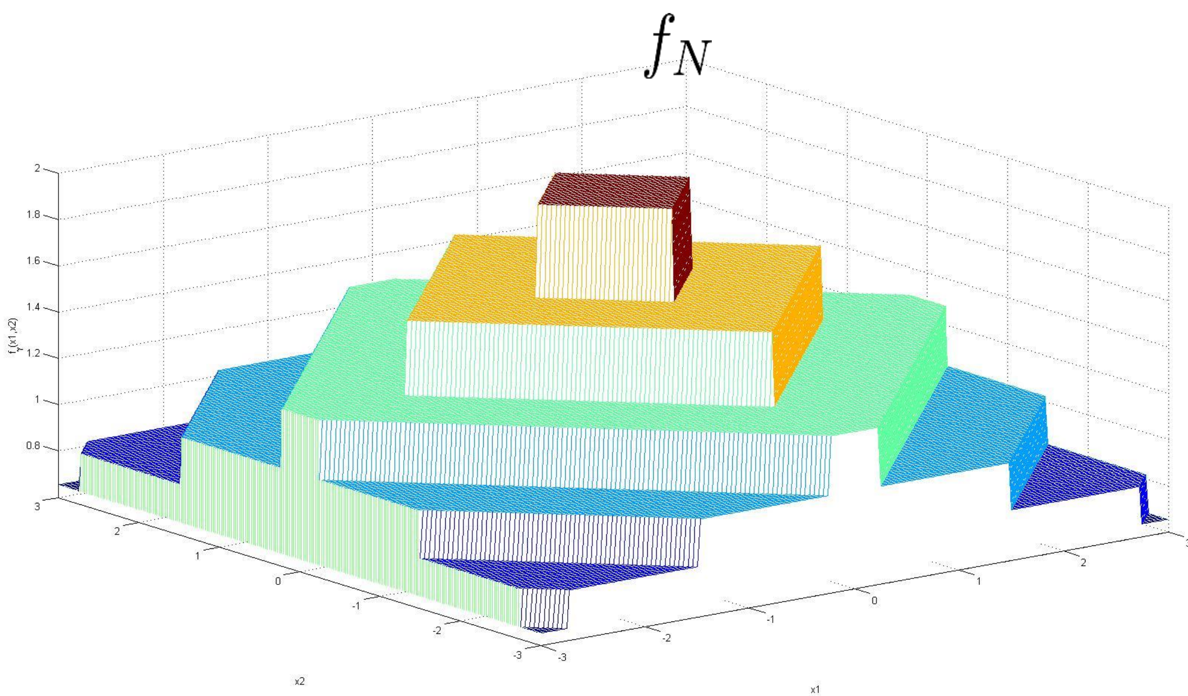
WHAT ABOUT OTHER LOSSES





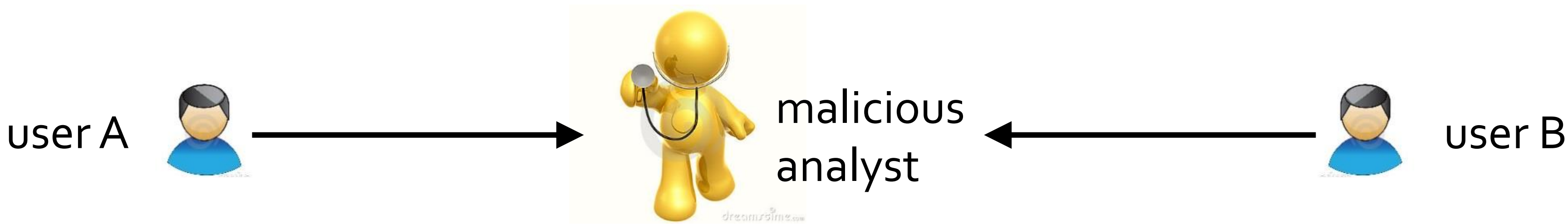
WHAT ABOUT 2 DIMENSIONAL DATA





PART 2/3:
LOCAL PRIVACY

LOCAL PRIVACY MODEL



LOCAL PRIVACY MODEL

have you ever used illegal drugs?

answer truthfully

answer wrongly

[*Warner* 1965]

LOCAL DIFFERENTIAL PRIVACY

$$e^{-\varepsilon} \leq \frac{\mathbb{P}\left(Y|X\right)}{\mathbb{P}\left(Y|X'\right)} \leq e^{+\varepsilon}$$

ε controls the level of privacy
large ε, low privacy
small ε, high privacy

[*Duchi et al.* 2012]

PRIVACY-UTILITY TRADEOFF

maximize utility

subject to differential privacy

BINARY DATA

answer truthfully

$$\frac{e^{\varepsilon}}{e^{\varepsilon}+1}$$

answer wrongly

$$\frac{1}{e^{\varepsilon}+1}$$

WARNER'S RESPONSE IS OPTIMAL

answer truthfully

answer wrongly

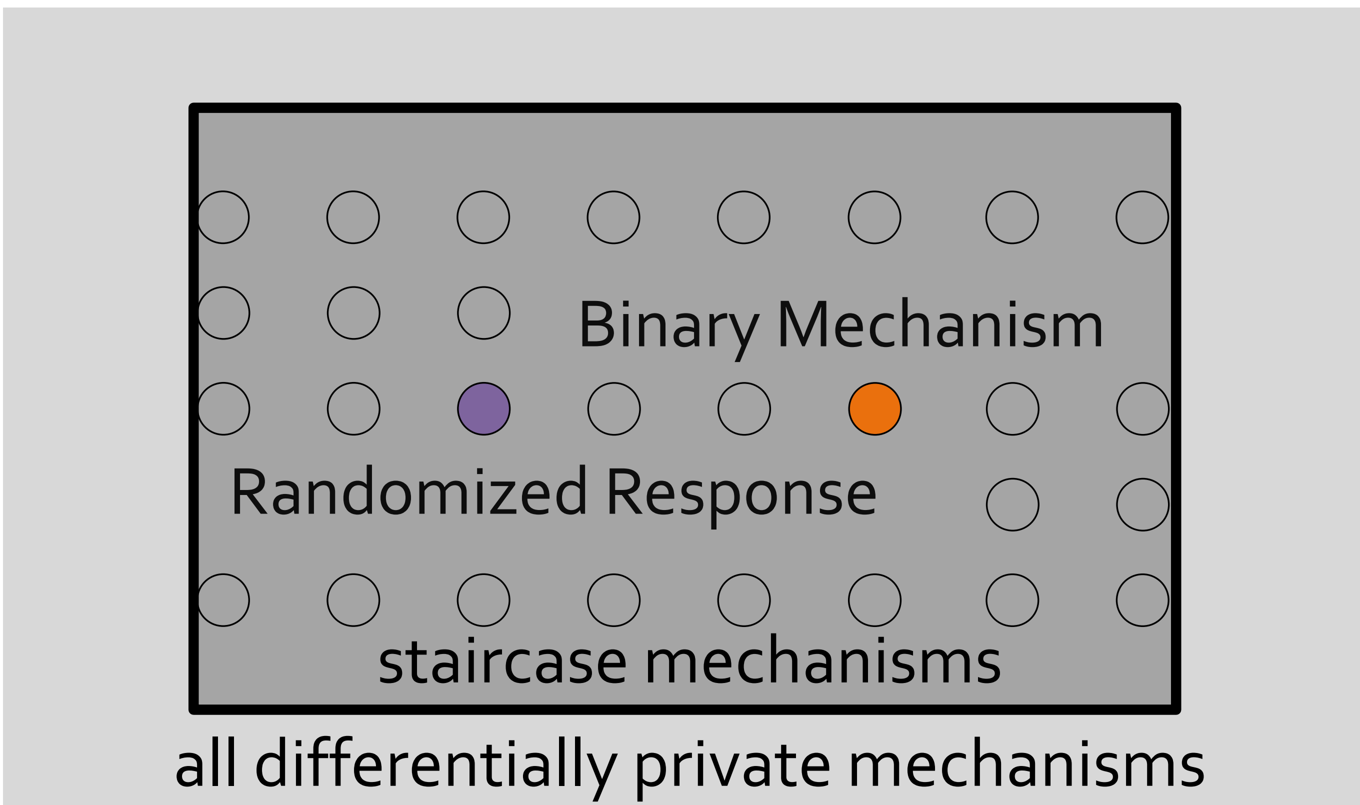
optimal for all privacy levels & all well behaved utilities

WHAT ABOUT NON-BINARY DATA

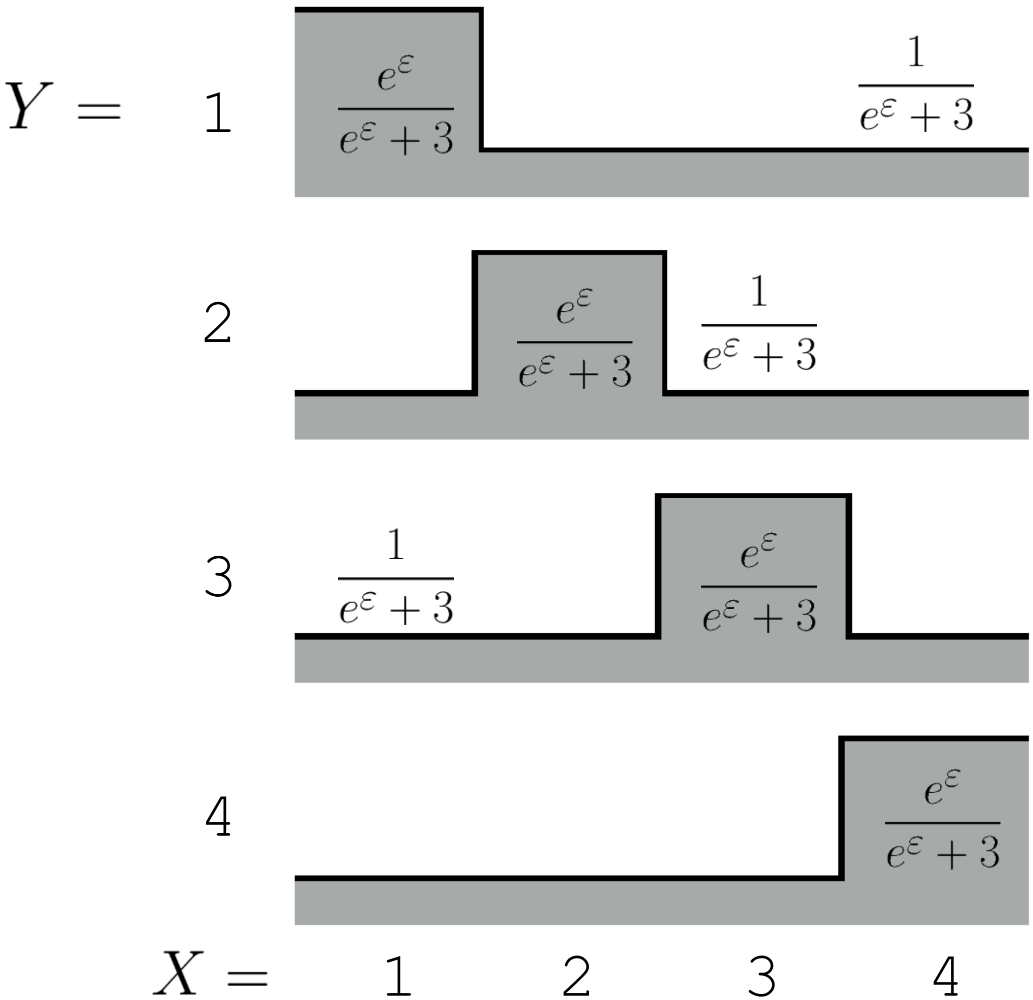
maximize utility

subject to differential privacy

MAIN RESULTS

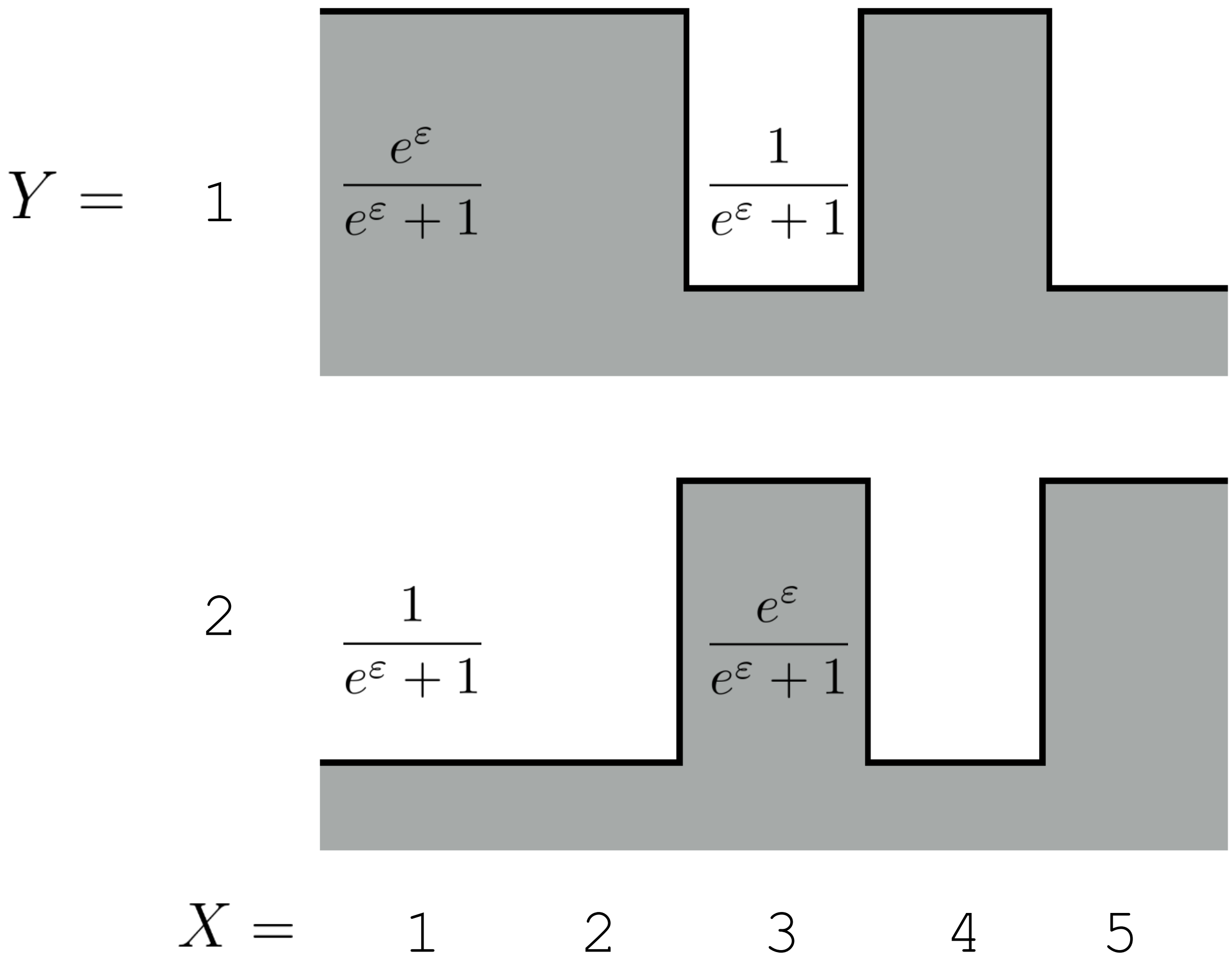


RANDOMIZED RESPONSE



optimal in the low privacy regime

BINARY MECHANISM

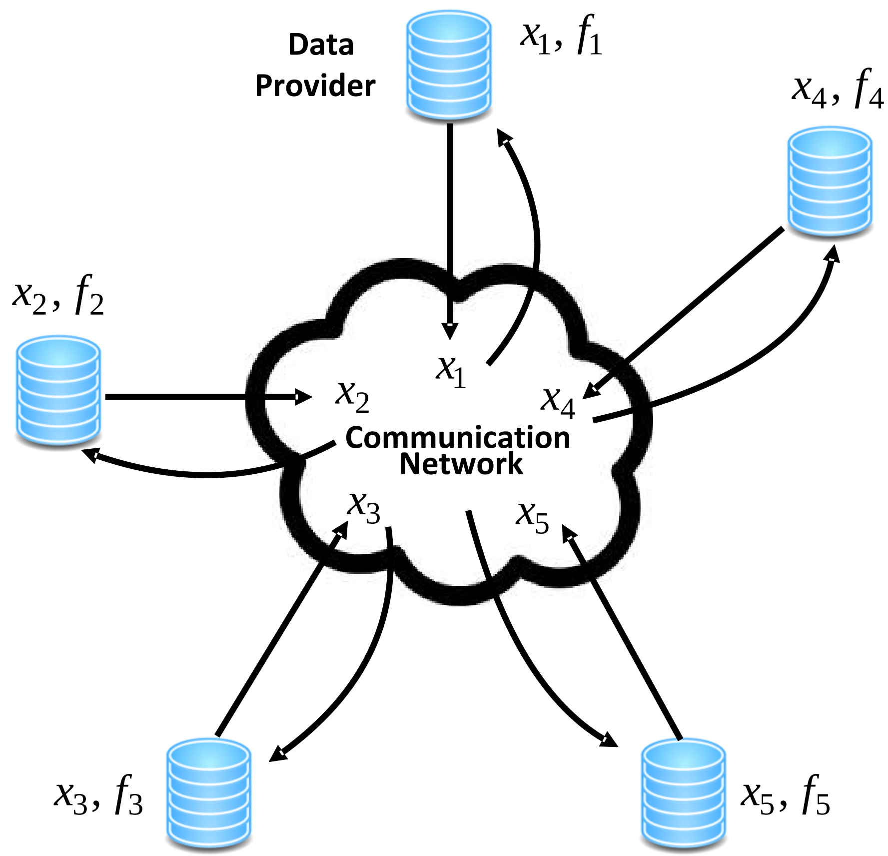


optimal in the high privacy regime

@Google

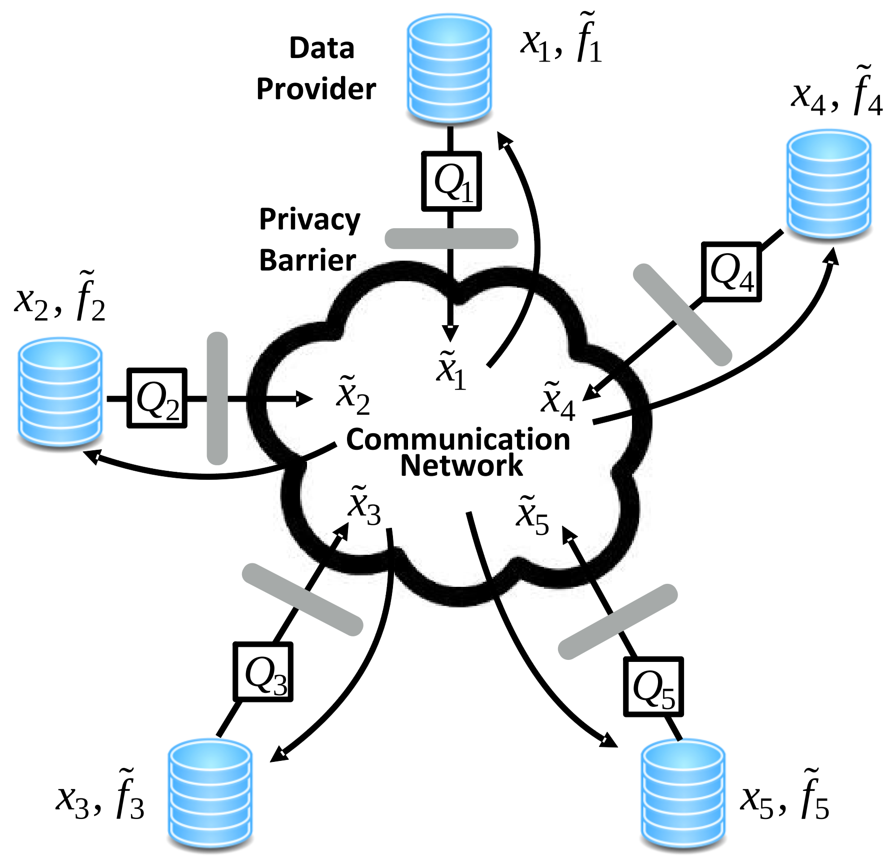
PART 3/3:
MULTI-PARTY PRIVACY

MULTI-PARTY COMPUTATION



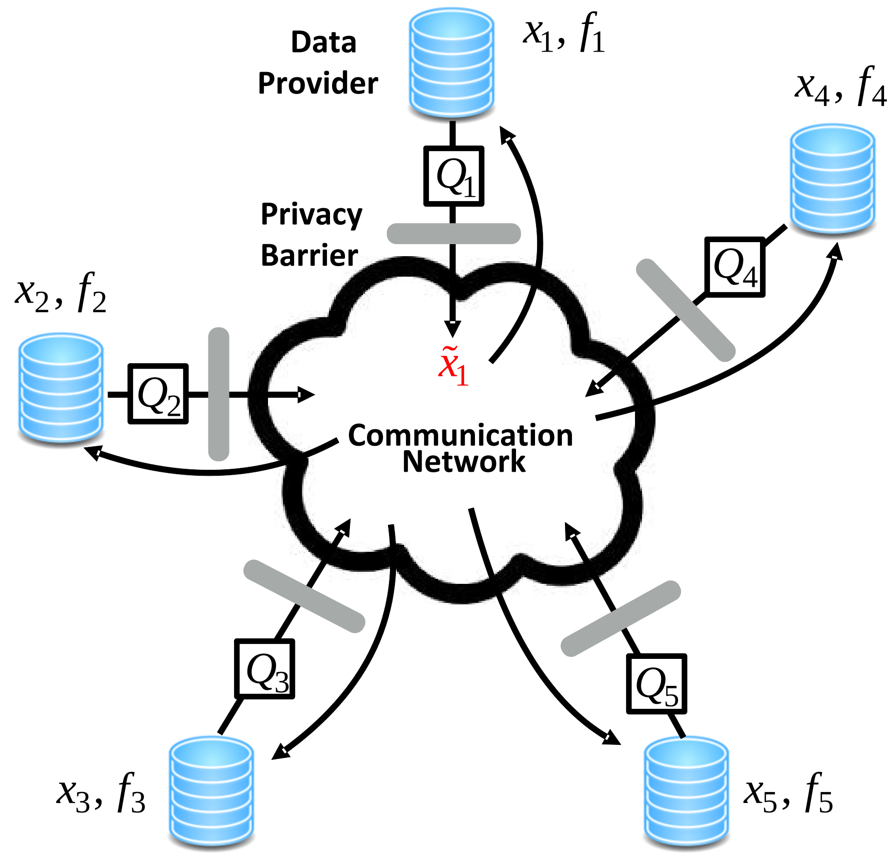
an important setting in distributed systems

PRIVATE MULTI-PARTY COMPUTATION

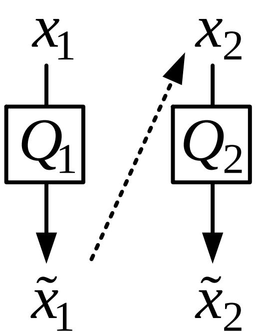


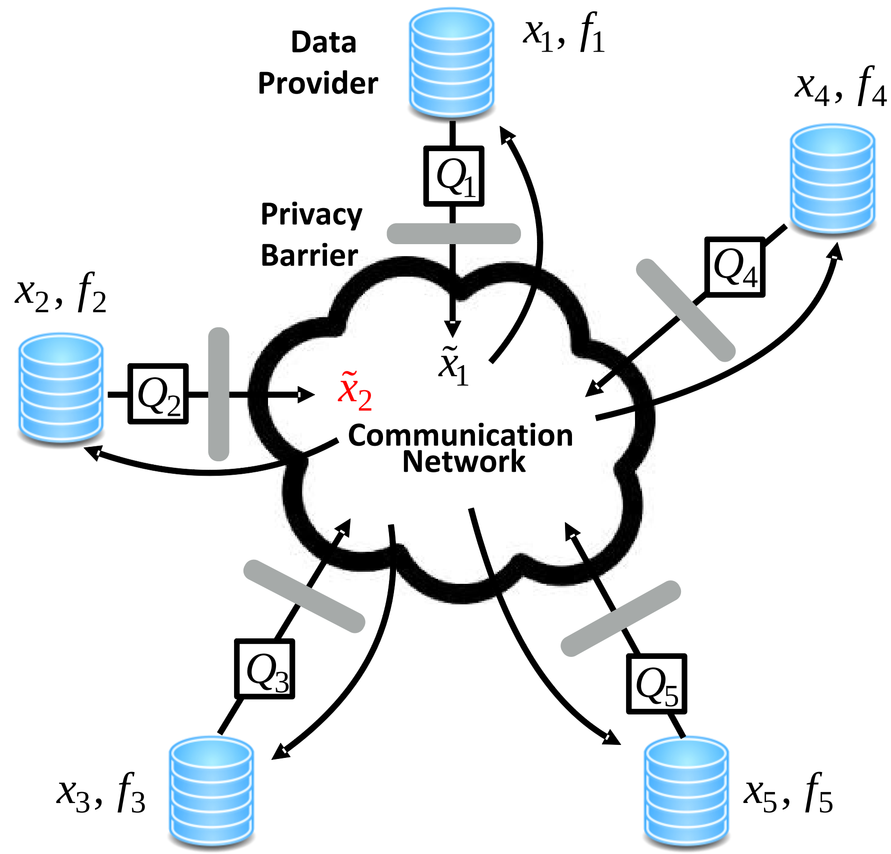
each party shares a noisy version of its data

INTERACTIVE MECHANISMS

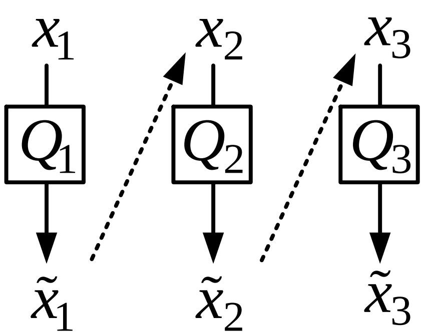


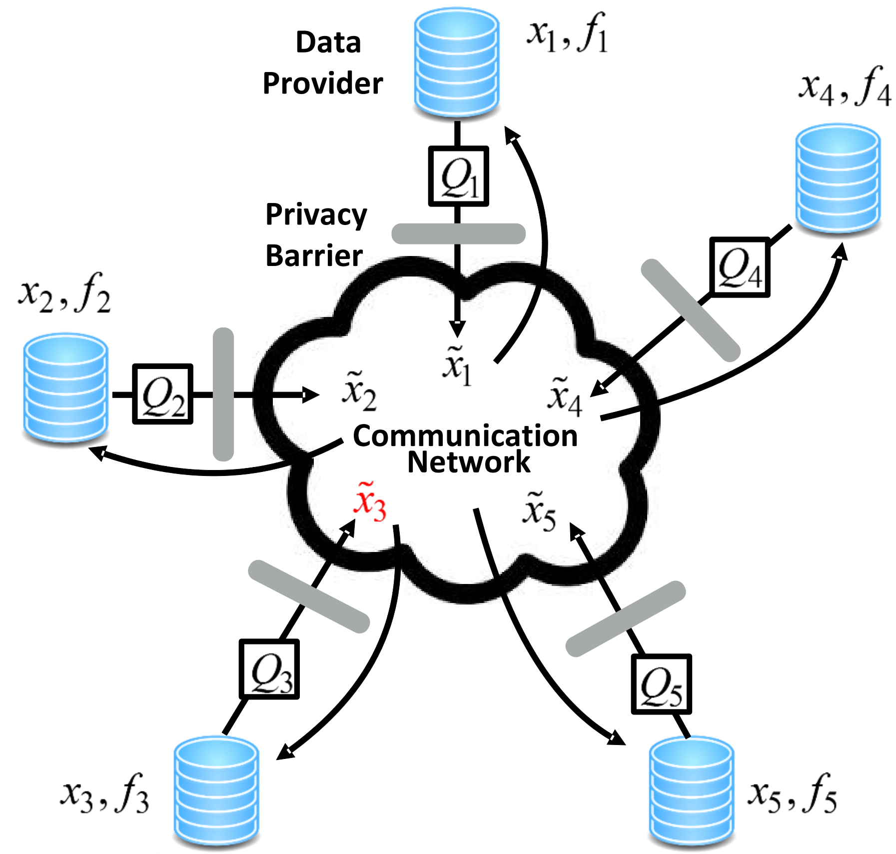
INTERACTIVE MECHANISMS



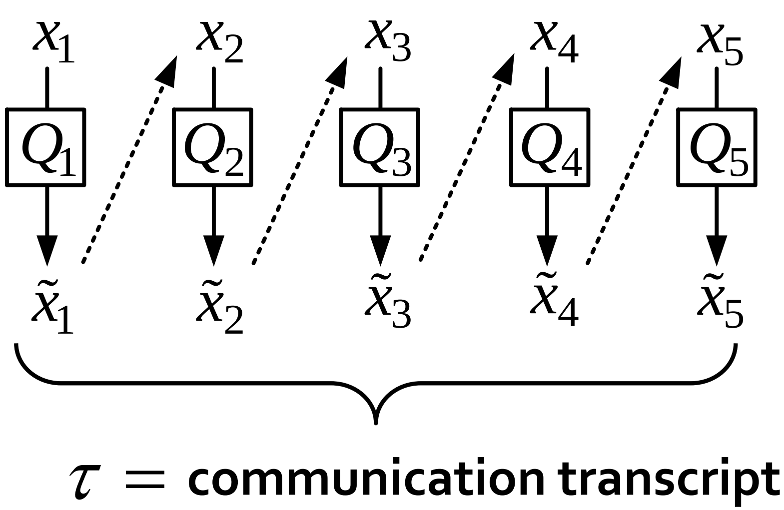


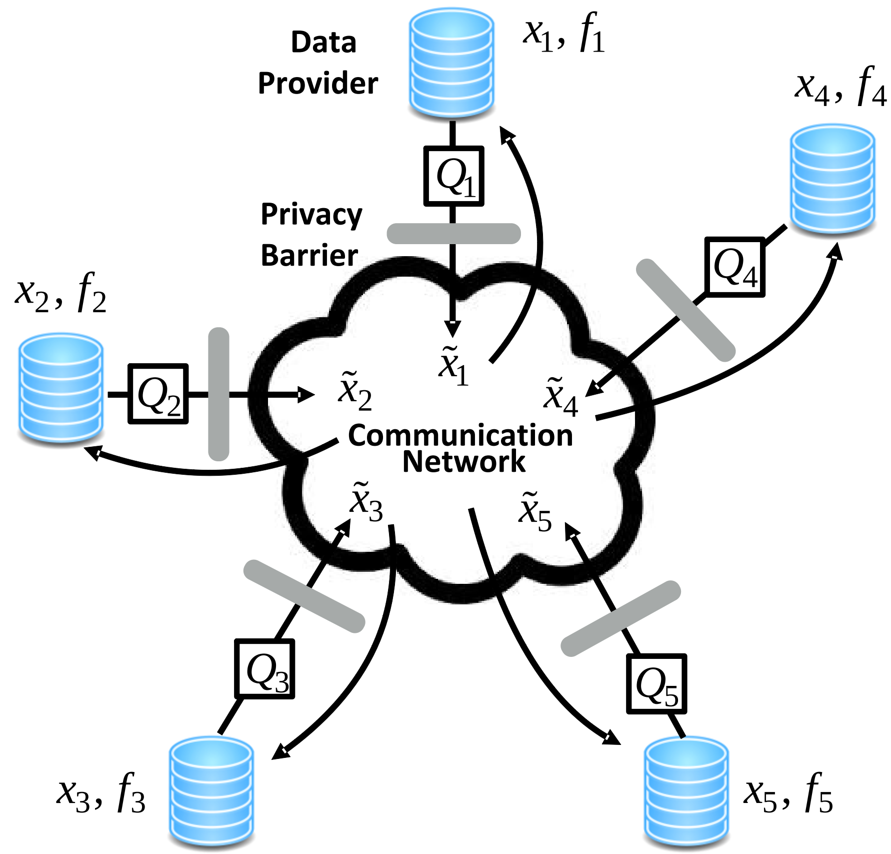
INTERACTIVE MECHANISMS



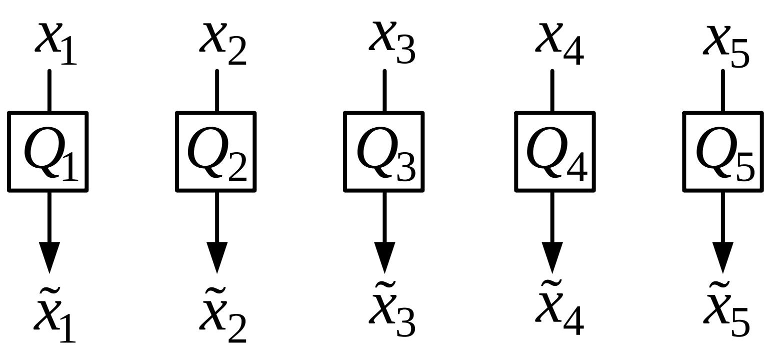


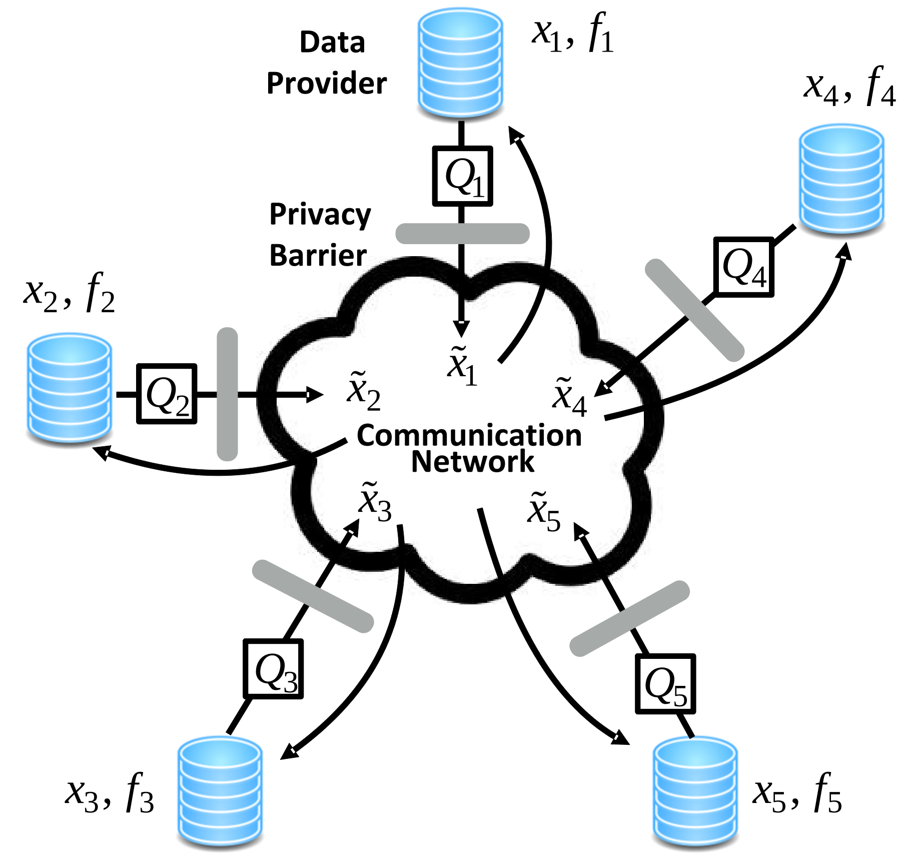
INTERACTIVE MECHANISMS



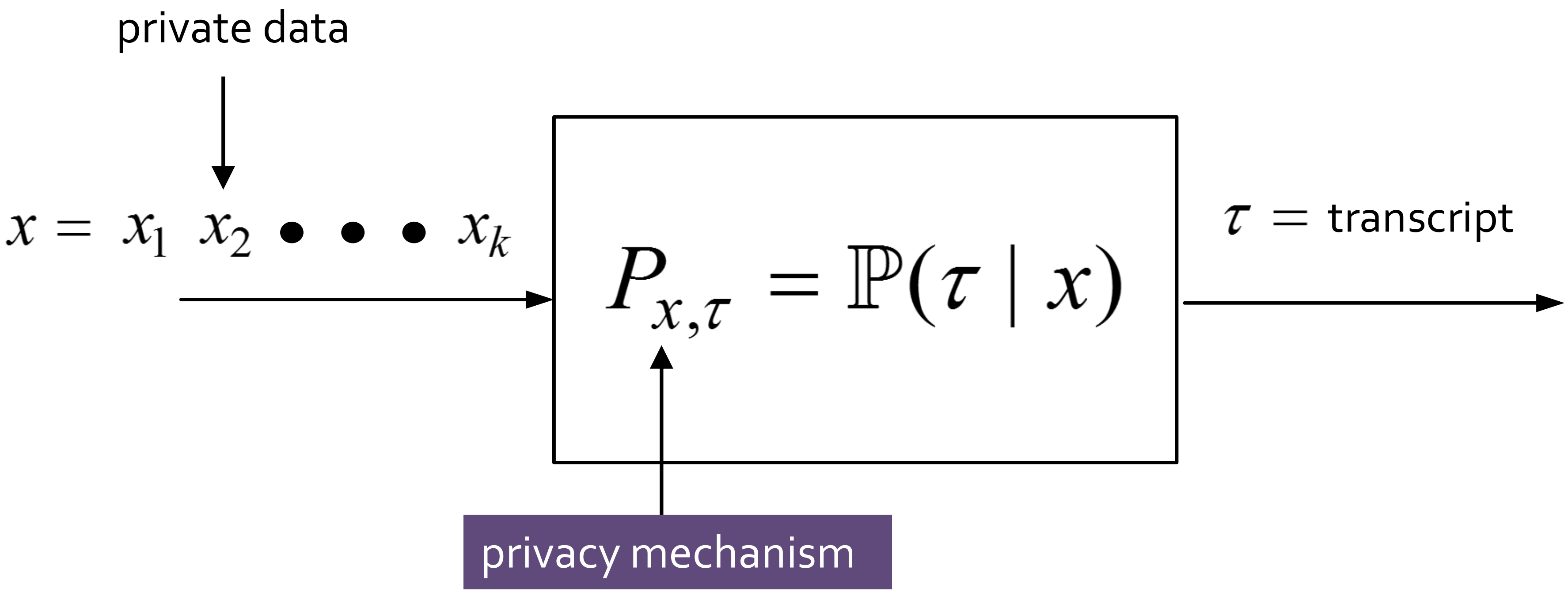


NON-INTERACTIVE MECHANISMS





GENERAL REPRESENTATION



MULTI-PARTY DIFFERENTIAL PRIVACY

$x = x_1 \; x_2 \bullet \bullet \bullet \; x_k$

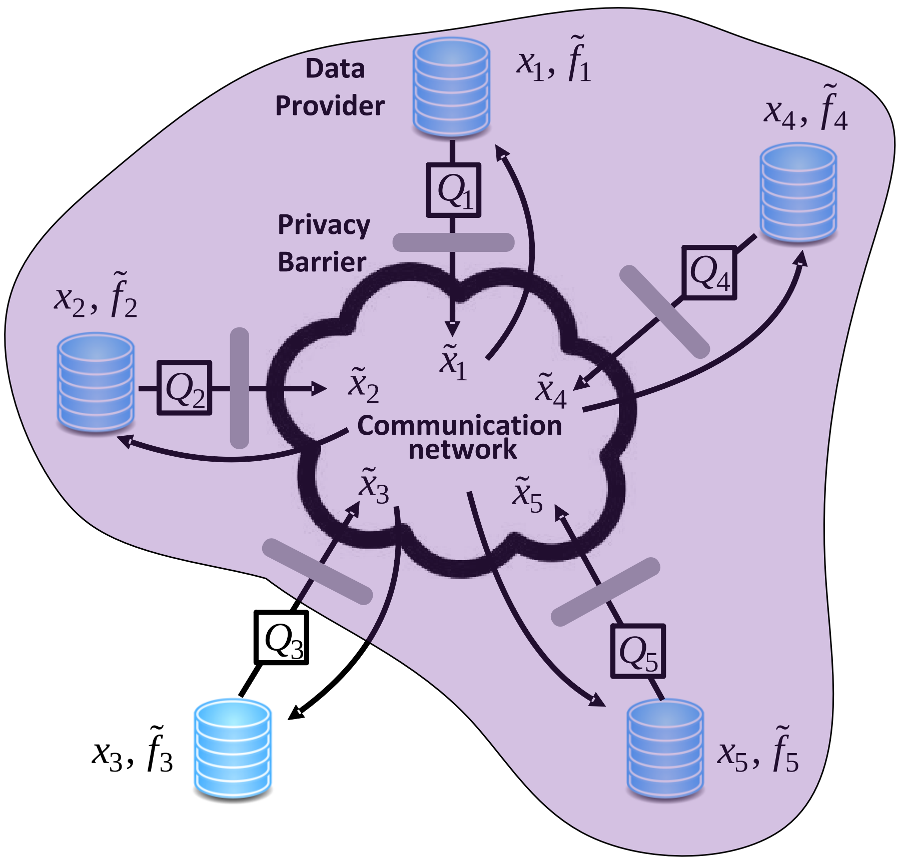
$$P_{x,\tau} = \mathbb{P}(\tau \mid x)$$

$\tau =$ **transcript**

$$e^{-\varepsilon_i} \le \frac{\mathbb{P}(\tau \mid x_i = 0, x_{-i})}{\mathbb{P}(\tau \mid x_i = 1, x_{-i})} \le e^{\varepsilon_i}$$

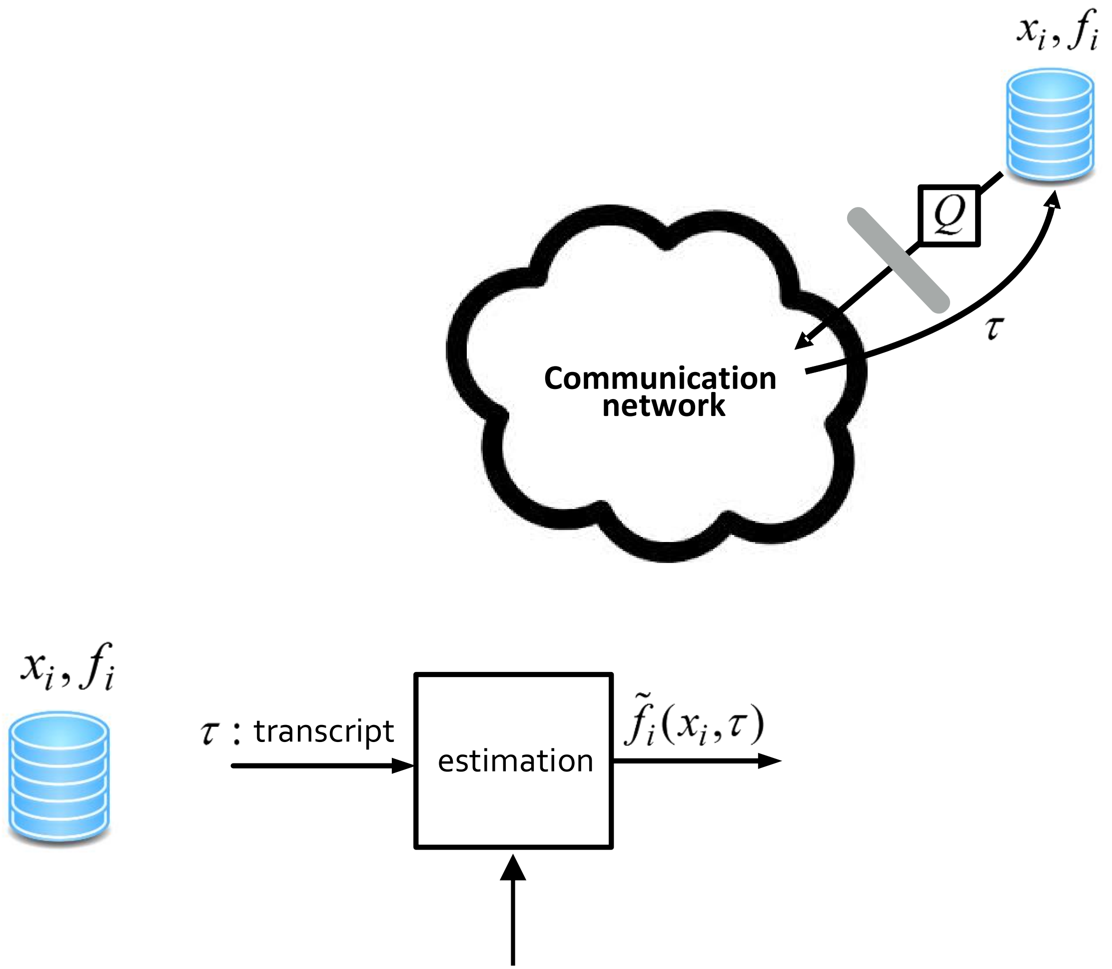
$x_{-i} = (x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_k)$

CAN'T SAY MUCH EVEN IF…



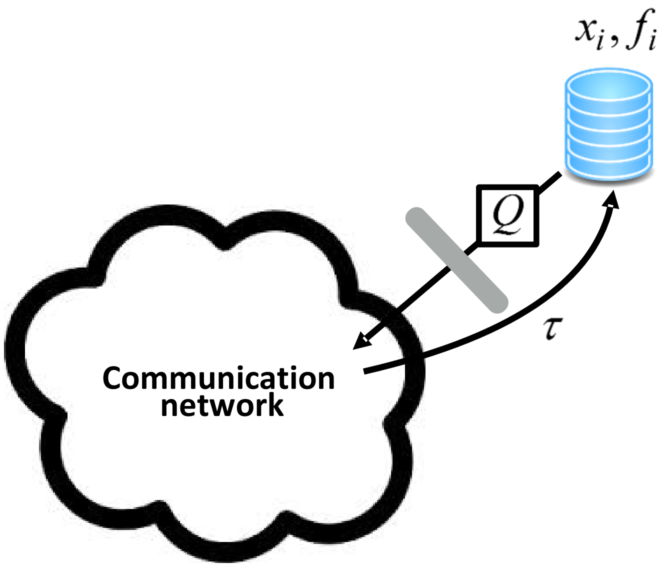
all parties but one collude to figure out a party's bit

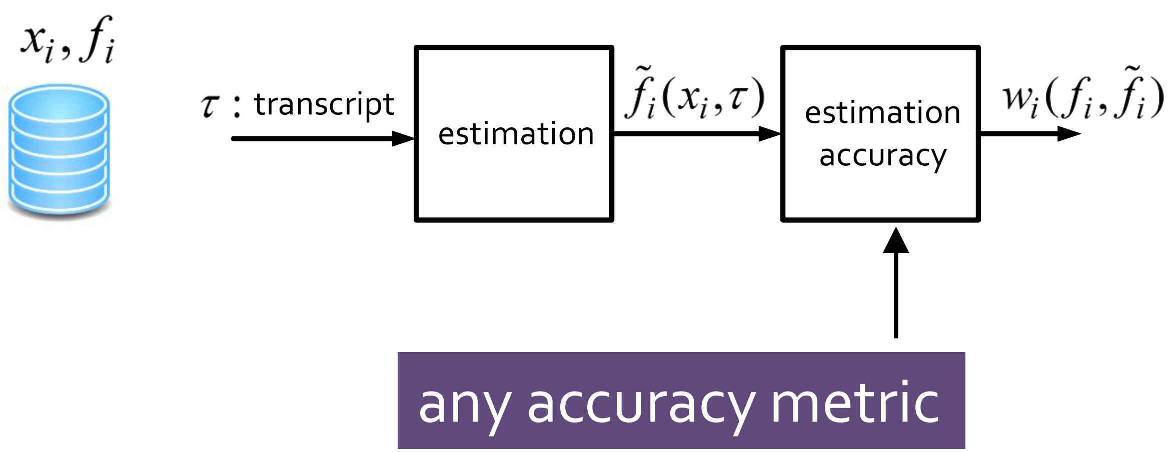
FUNCTION ESTIMATION



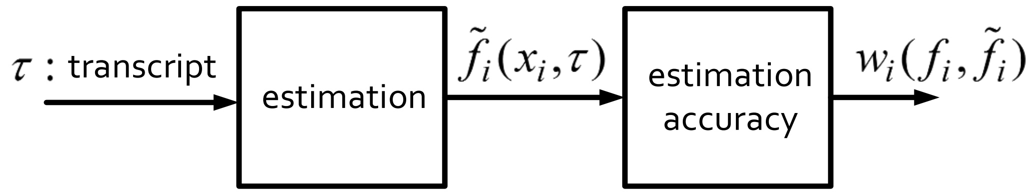
any estimation rule (potentially randomized)

FUNCTION ESTIMATION





ACCURACY-PRIVACY TRADEOFF



$$\mathrm{ACC}_{\mathrm{ave}} \equiv \underbrace{\frac{1}{2^k} \sum_{x \in \{0,1\}^k}}_{\text{average over all possible inputs}} \mathbb{E}_{\hat{f}_i, P_{x,\tau}} [w_i(f_i(x), \tilde{f}_i(\tau, x_i))]$$

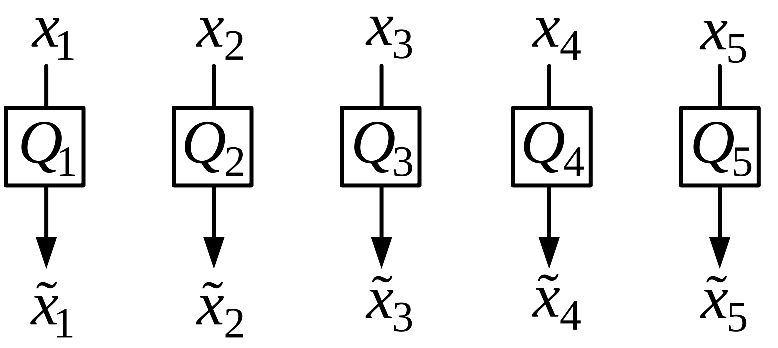
ACCURACY-PRIVACY TRADEOFF

$$
\begin{aligned}
\underset{P,\tilde{f}_i}{\text{maximize}} \quad & \mathrm{ACC}_{\mathrm{ave}}(P, w_i, f_i, \tilde{f}_i), \\
\text{subject to} \quad & P \text{ and } \tilde{f}_i \text{ are row-stochastic matrices} \\
& P \text{ satisfies the differential privacy constraints} \\
& \text{for all parties}
\end{aligned}
$$

- heterogeneous privacy levels across users
- each party possesses a single bit
- the functions can vary from one party to the other
- the accuracy metrics can vary from one party to the other
- interactive & non-interactive mechanisms

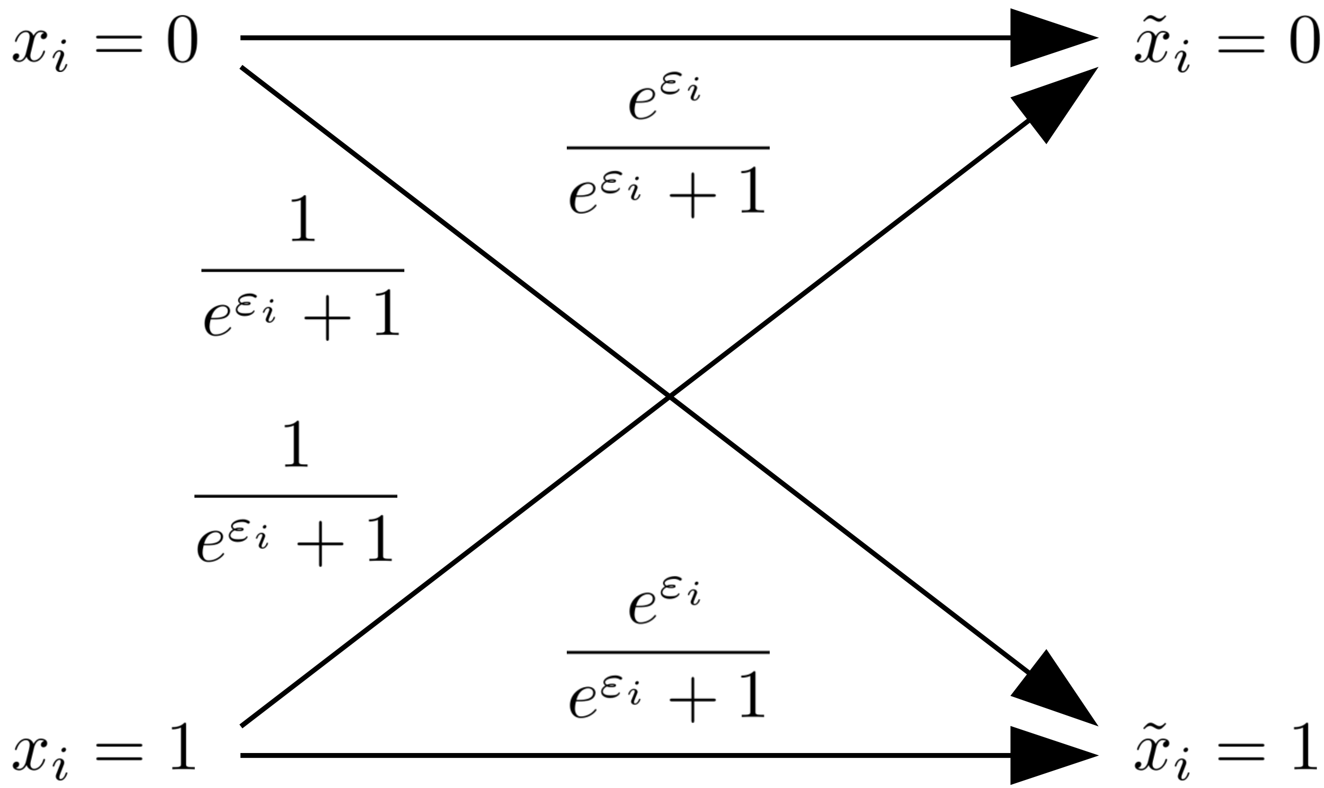
OUR RESULT

non-interactive mechanisms are optimal

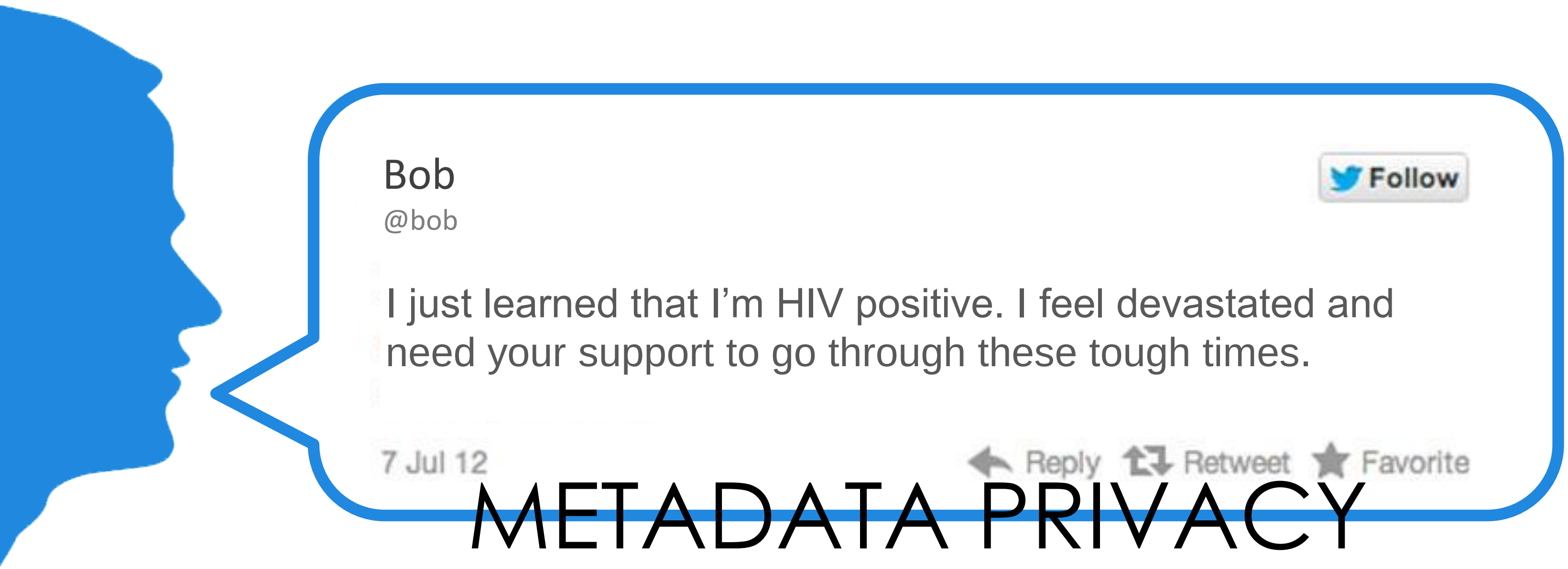


OUR RESULT

Warner's randomized response is optimal

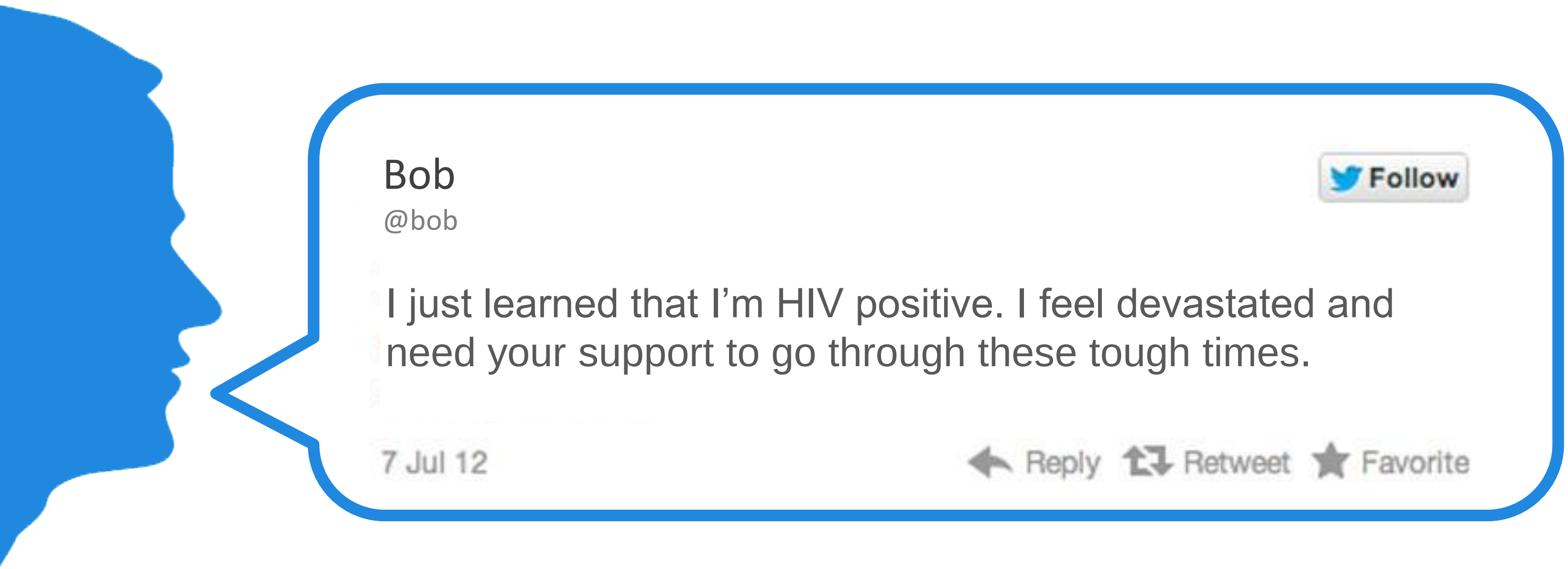


NON-BINARY DATA



METADATA PRIVACY

METADATA PRIVACY

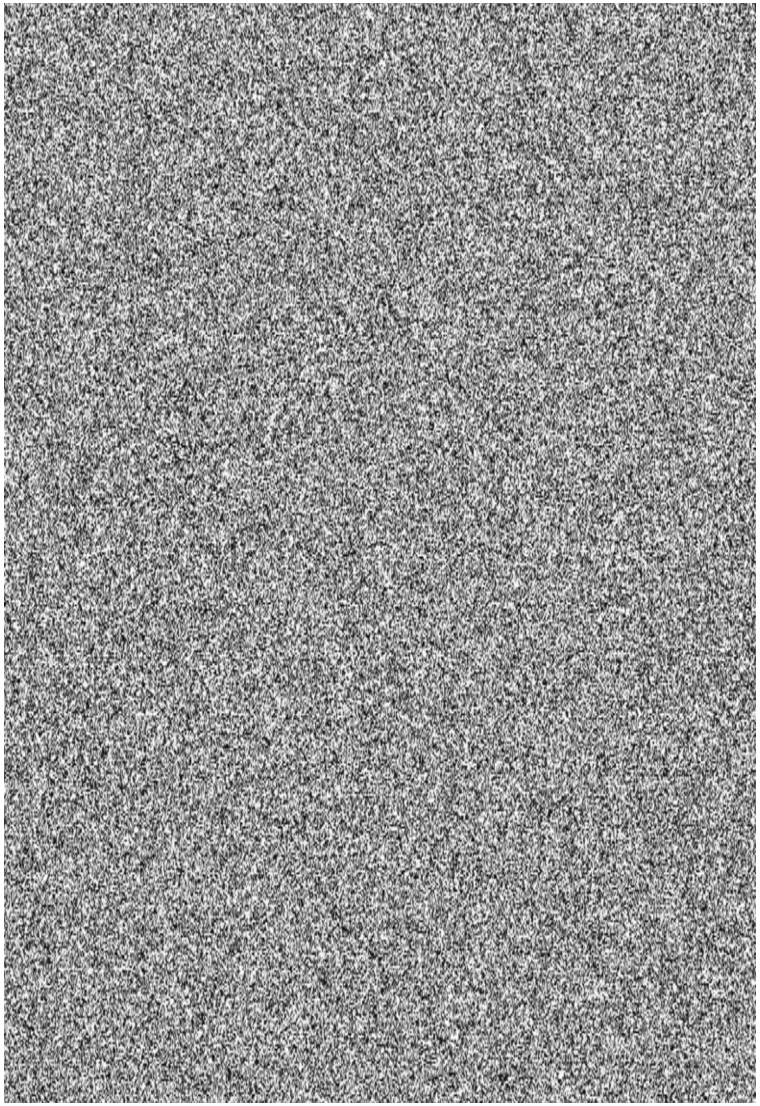


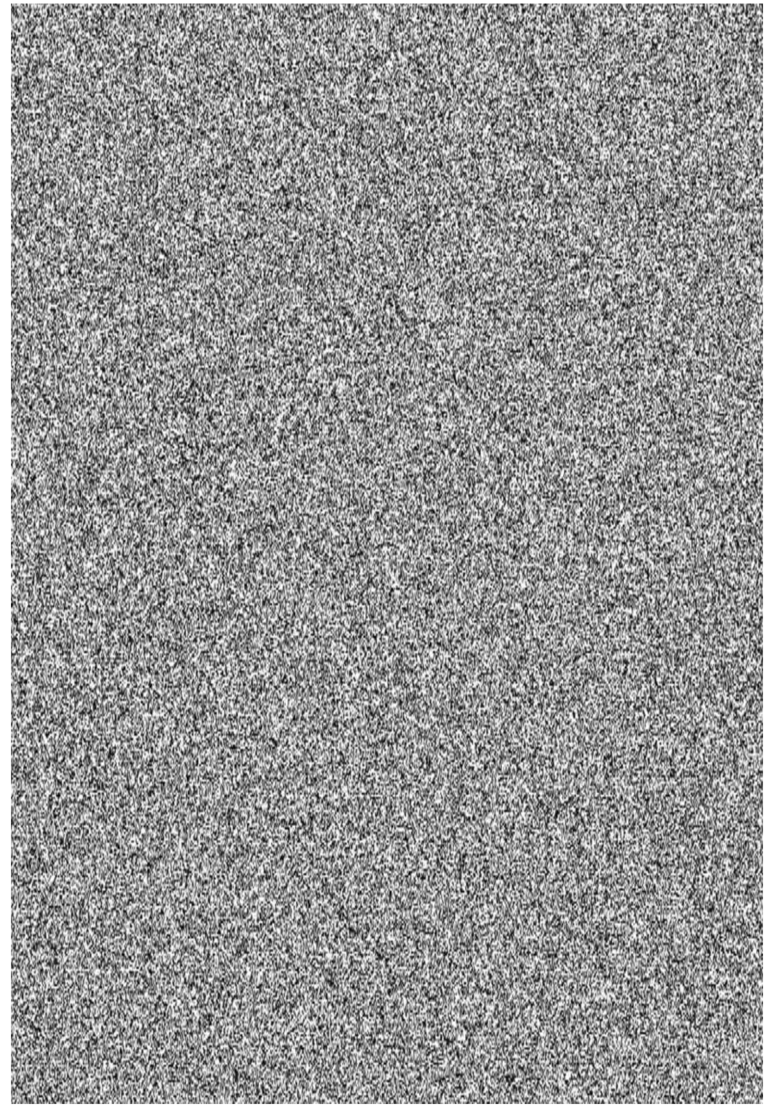
[Best Paper Award at SIGMETRICS 15, SIGMETRICS 16]

first fully distributed, truly **anonymous social network**

THANK YOU!

A VERY BIG THANK YOU!





A VERY BIG THANK YOU!

Sewoong Oh

Pramod Viswanath

A VERY SPECIAL THANK YOU!

A VERY SPECIAL THANK YOU!

SELFIE EVERYONE?