Challenges and Practices
of Local Differential Privacy in Real-World

by Pingchuan Ma & Zhiqiang Wang
@BESTI (KIS Lab)

About

- Pingchuan Ma (20162308@mail.besti.edu.cn)

He is a student from Beijing Electronic Science and Technology Institute and an intern of CNCERT.

- Zhiqiang Wang (wangzq@besti.edu.cn)

He is a lecturer from Beijing Electronic Science and Technology Institute and a post-doctoral of State Information Center. His research interests include vulnerability discovery and privacy preserving.

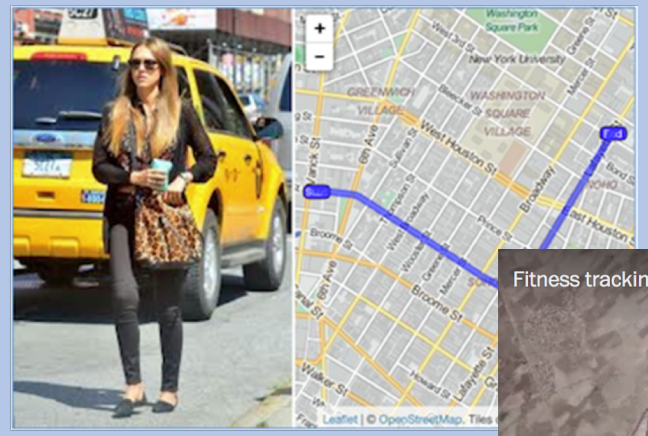
Overview

- Background & History
- Theoretic Foundation
- Applications
- Challenges & Future Directions

Background

Why we need LDP since we have anonymization?

Data Release & Privacy Leakage



Jessica Alba (actor)

Strava's fitness tracker heat map reveals the location of military bases

Geolocation isn't a new problem for the military

3:51pm EST

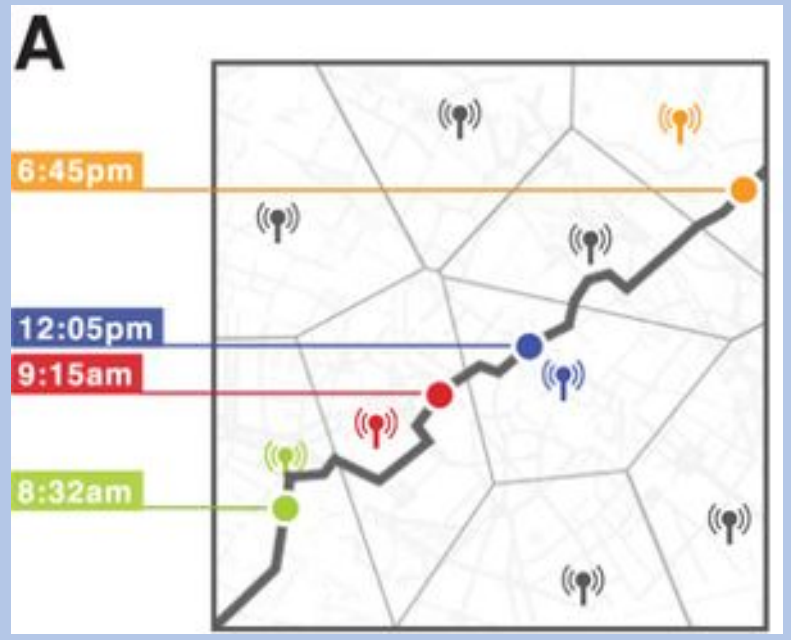
AOL

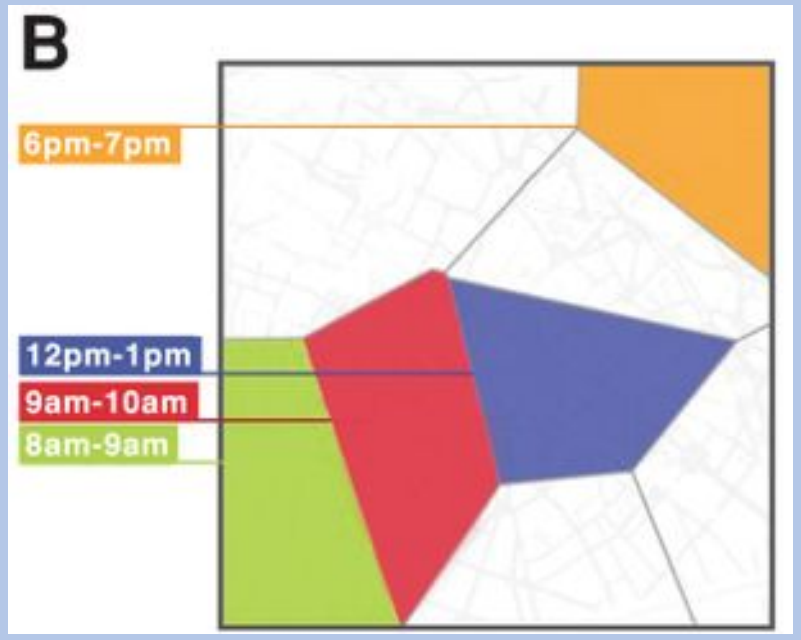
NETFLIX

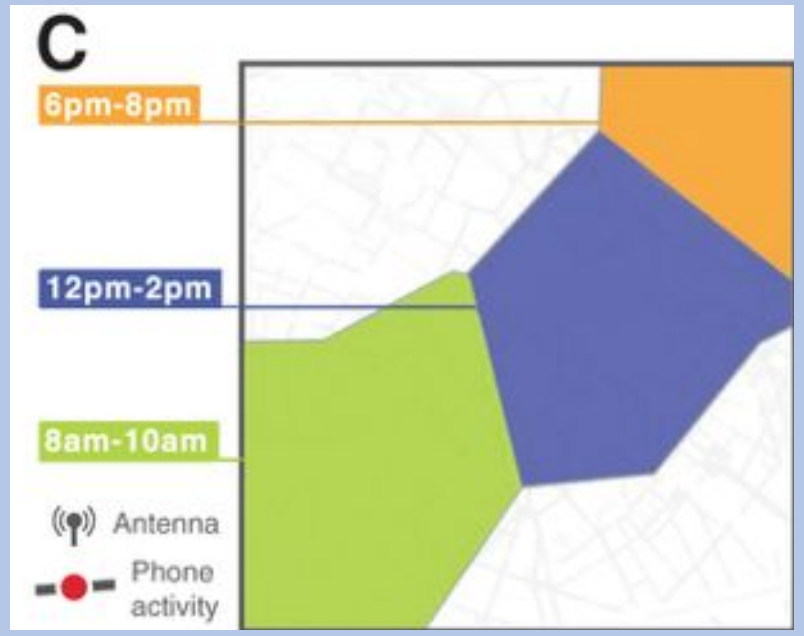
Commonwealth of Massachusetts
Group Insurance Commission
Your Benefits Connection

Slide from "Privacy at Scale: Local Differential Privacy in Practice"

De-anonymization & Linking Attack







Background

How to guarantee privacy?

K-anonymity & its variants

ID	Age	Zipcode	Diagnosis
1	28	13053	Heart Disease
2	29	13068	Heart Disease
3	21	13068	Viral Infection
4	23	13053	Viral Infection
5	50	14853	Cancer
6	55	14853	Heart Disease
7	47	14850	Viral Infection
8	49	14850	Viral Infection
9	31	13053	Cancer
10	37	13053	Cancer
11	36	13222	Cancer
12	35	13068	Cancer

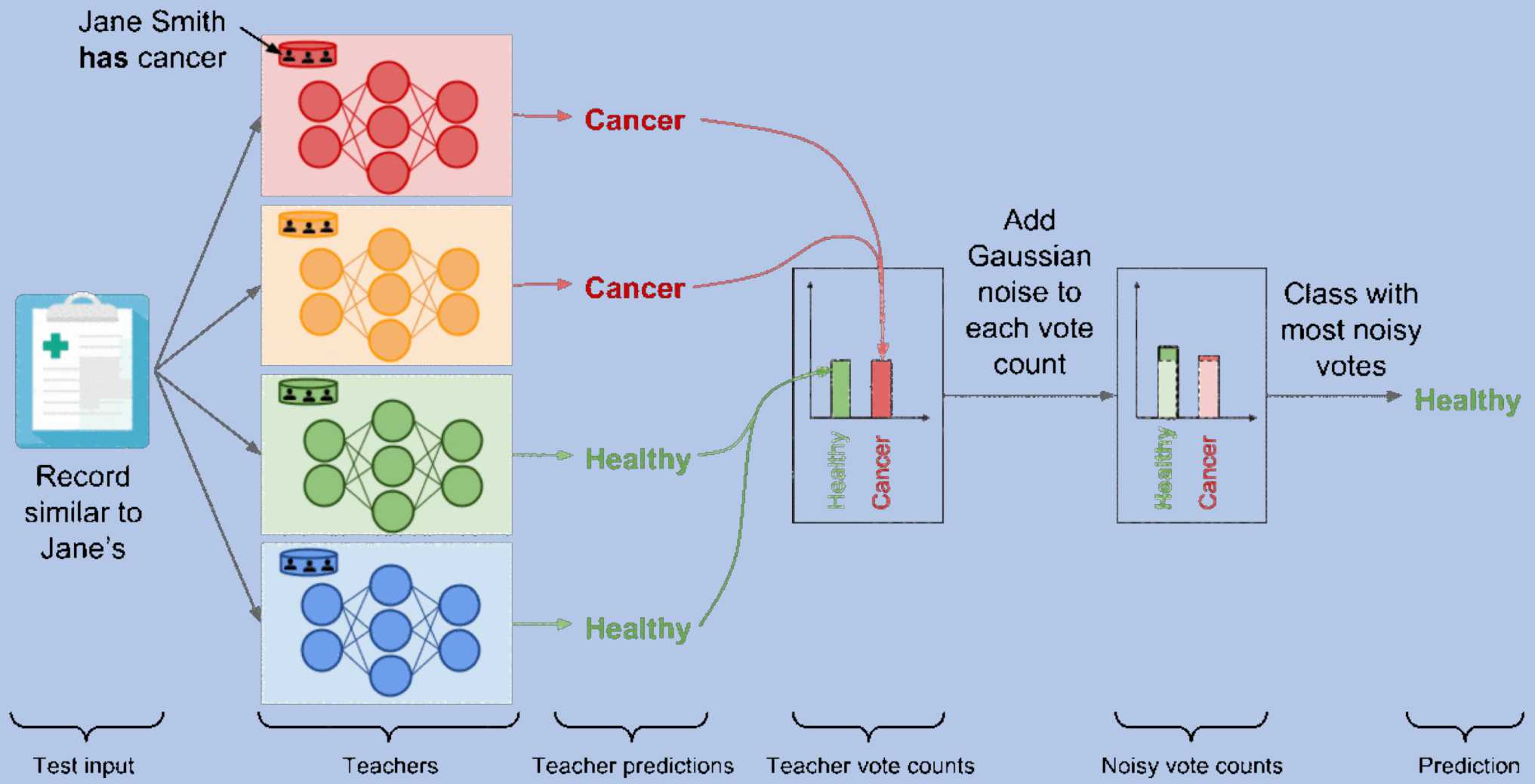
ID	Age	Zipcode	Diagnosis
1	[20-30]	130**	Heart Disease
2	[20-30]	130**	Heart Disease
3	[20-30]	130**	Viral Infection
4	[20-30]	130**	Viral Infection
5	[40-60]	148**	Cancer
6	[40-60]	148**	Heart Disease
7	[40-60]	148**	Viral Infection
8	[40-60]	148**	Viral Infection
9	[30-40]	13***	Cancer
10	[30-40]	13***	Cancer
11	[30-40]	13***	Cancer
12	[30-40]	13***	Cancer

Sweeney, Latanya. "k-anonymity: A model for protecting privacy." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10.05 (2002): 557-570.

Background

How to quantitatively guarantee privacy?

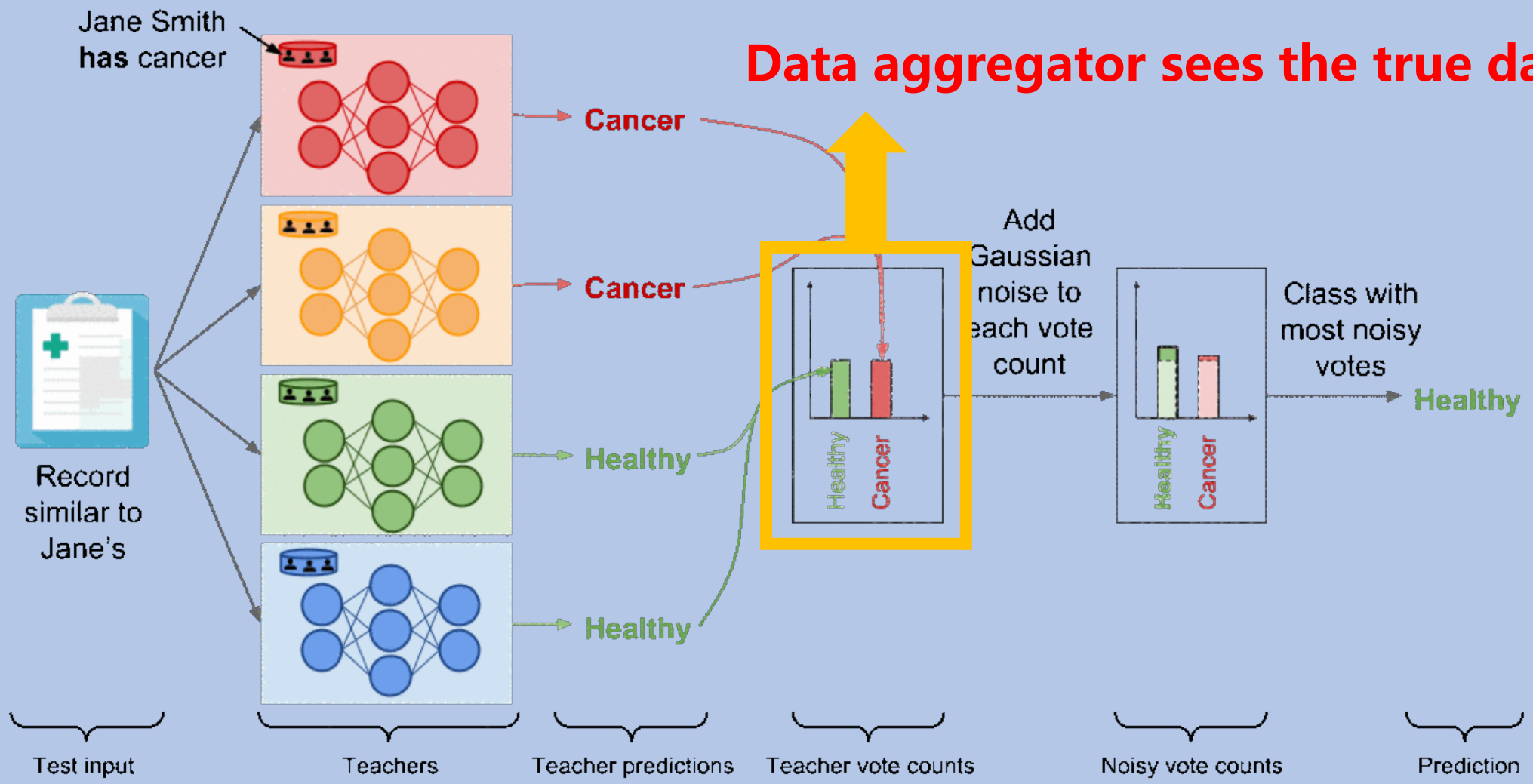
Differential Privacy



Background

How to preserve privacy and reduce trust?

Reduce Trust

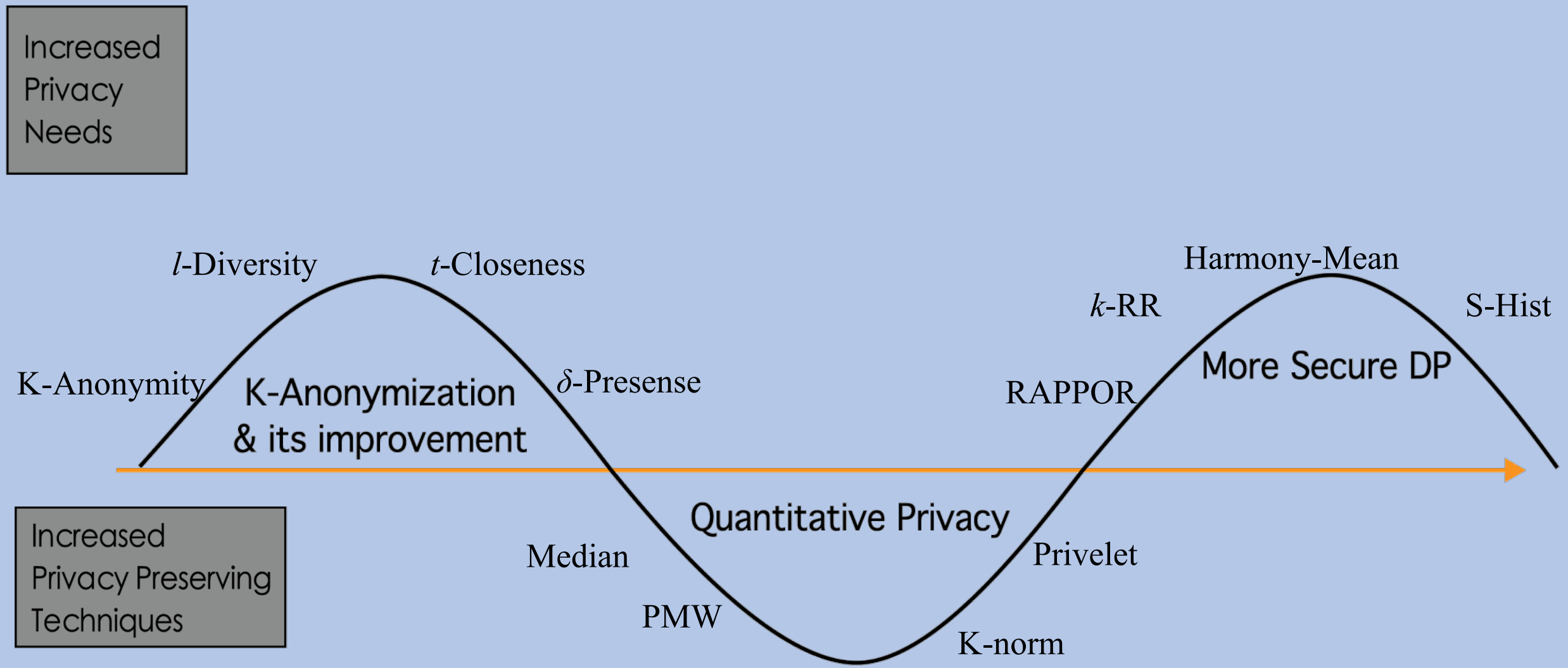


Reduce Trust

SMC? homomorphic encryption?

or run DP by each user?

History


Increased Privacy Needs
l-Diversity
t-Closeness
Harmony-Mean
k-RR
S-Hist
K-Anonymity
K-Anonymization & its improvement
δ-Presense
RAPPOR
More Secure DP
Quantitative Privacy
Median
Privelet
PMW
K-norm
Increased Privacy Preserving Techniques

Theoretical Foundations

A **randomized algorithm** K satisfies ϵ**-differential privacy** iff:
Given two data sets that differ by one individual called D and D', for any output S of K:

$$\frac{Pr[K(D) \in S]}{Pr[K(D') \in S]} \leq e^{\epsilon}$$

Epsilon is private budget.

Smaller epsilon comes with better privacy.

From DP to Local DP

A **randomized algorithm** K satisfies ϵ**-differential privacy** iff:
Given two data sets that differ by one individual called D and D', for any output S of K:

Run on the server

$$\frac{Pr[K(D) \in S]}{Pr[K(D') \in S]} \leq e^{\epsilon}$$

A **randomized algorithm** K satisfies ϵ**-local differential privacy** iff:
Given any two inputs x and x' and for any output y of K,

Run on the client

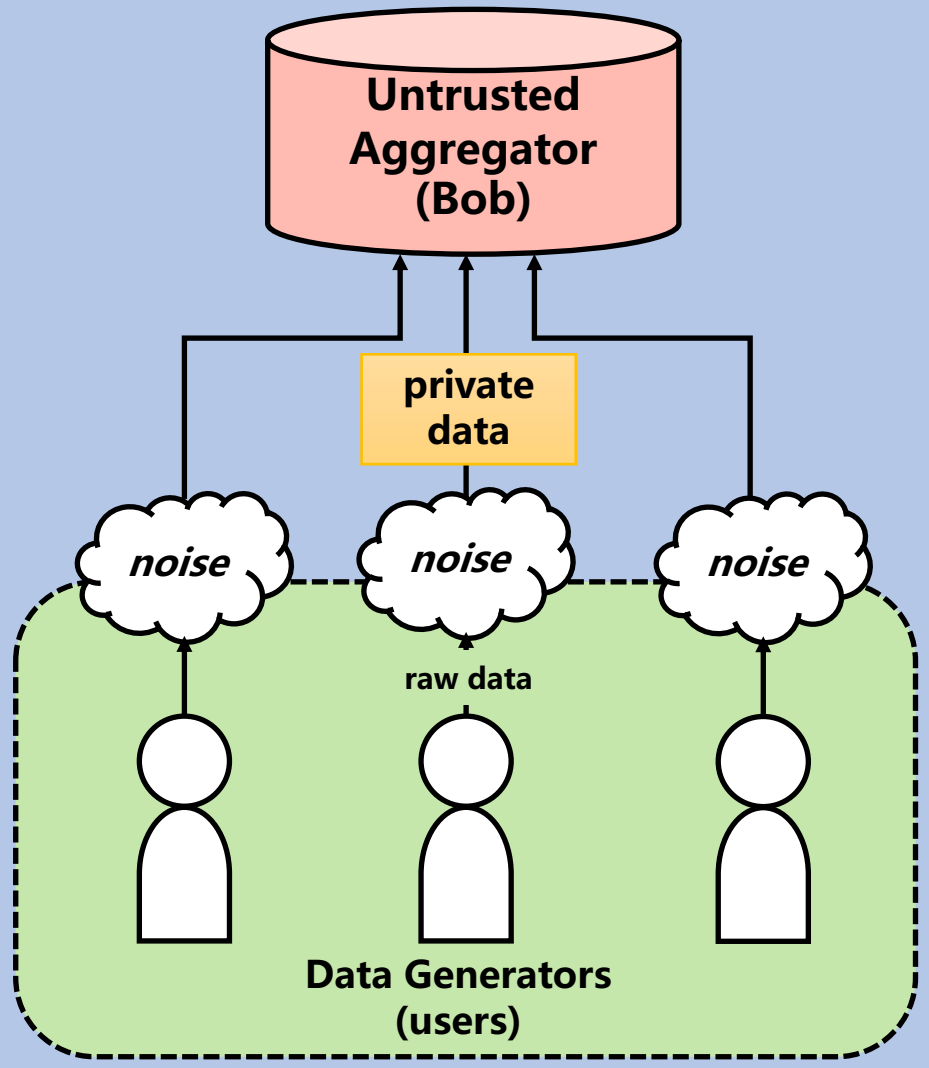
$$\frac{Pr[K(x) = y]}{Pr[K(x') = y]} \leq e^{\epsilon}$$

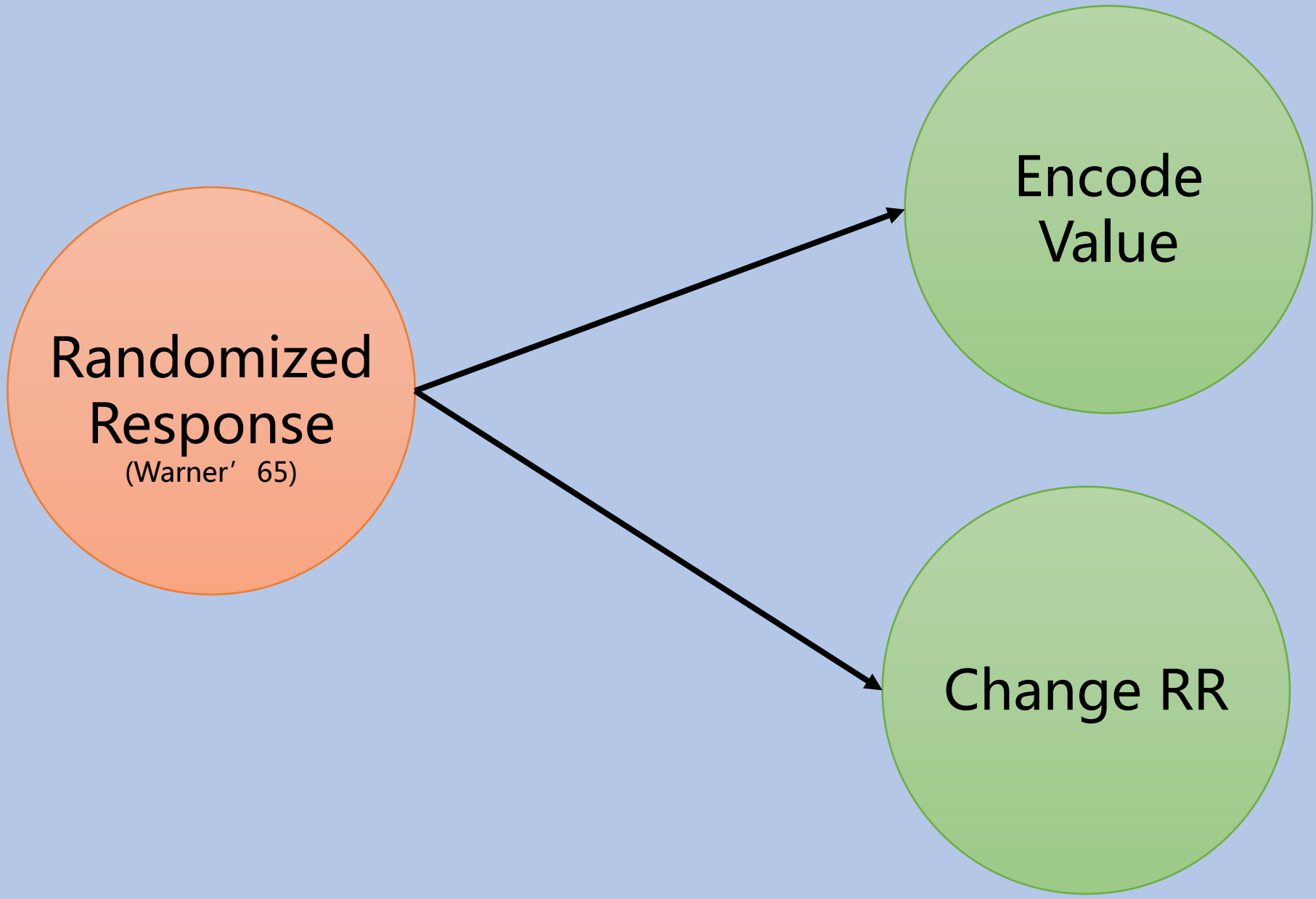
Randomized Response

- **Step1** Ask user a question whose answer can be "yes" or "no".
- **Step3** User answer the true result if head and answer randomly if tail.

RR only support binary attribute.

In that case, the data aggregator cannot infer exact answer of a certain user.




Randomized Response
(Warner’ 65)
Encode Value
Change RR

More...

PCE

Harmony-Mean

LoPub

S-Hist

RAPPOR

O-RR

O-RAPPOR

k-Subset

K-RR: From binary to N

The generalized randomized response mechanism is that for any input x and its output y:

$$Pr(y|x) = \begin{cases} \frac{e^{\epsilon}}{|\mathcal{X}|-1+e^{\epsilon}} & if\ y = x \\ \frac{1}{|\mathcal{X}|-1+e^{\epsilon}} & if\ y \neq x \end{cases}$$

where $\mathcal{X}$ is the true data set, $x \in \mathcal{X}$.

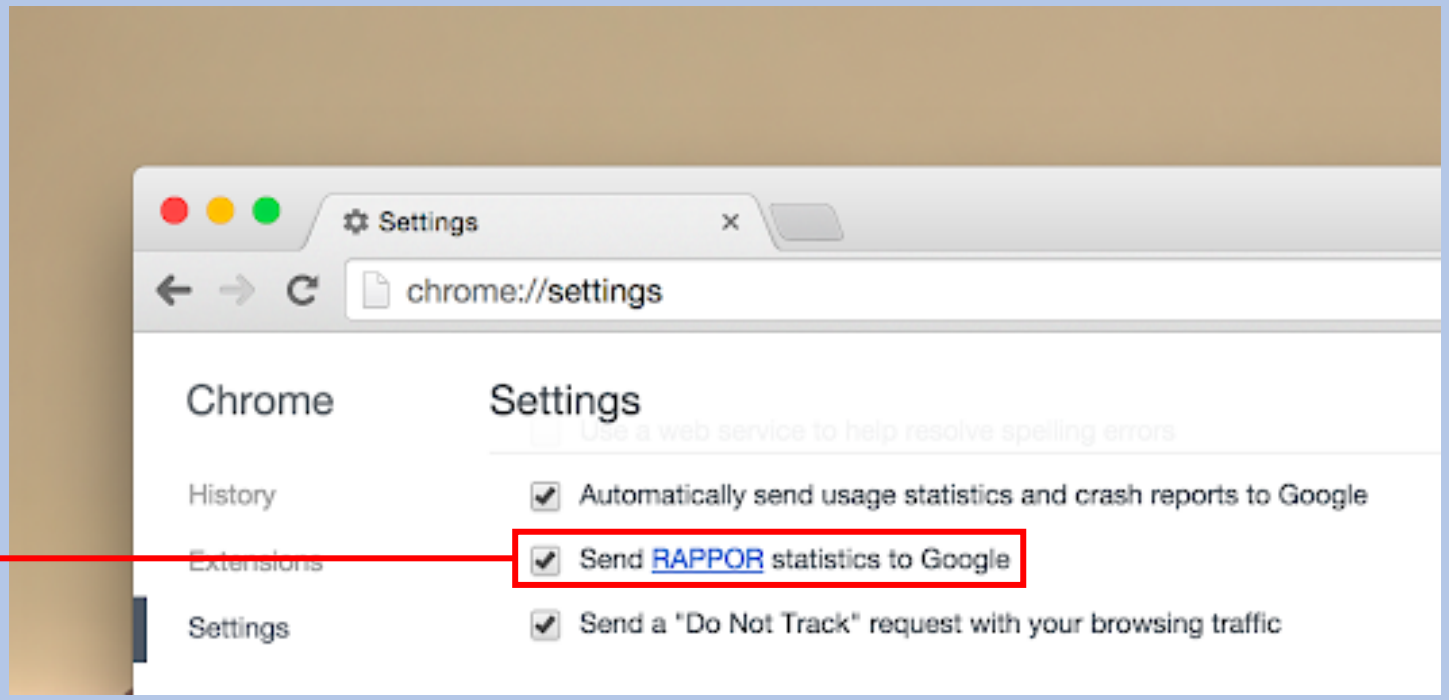
Randomized Response is included in a special case when $|\mathcal{X}| = 2$

RAPPOR: LDP In Google

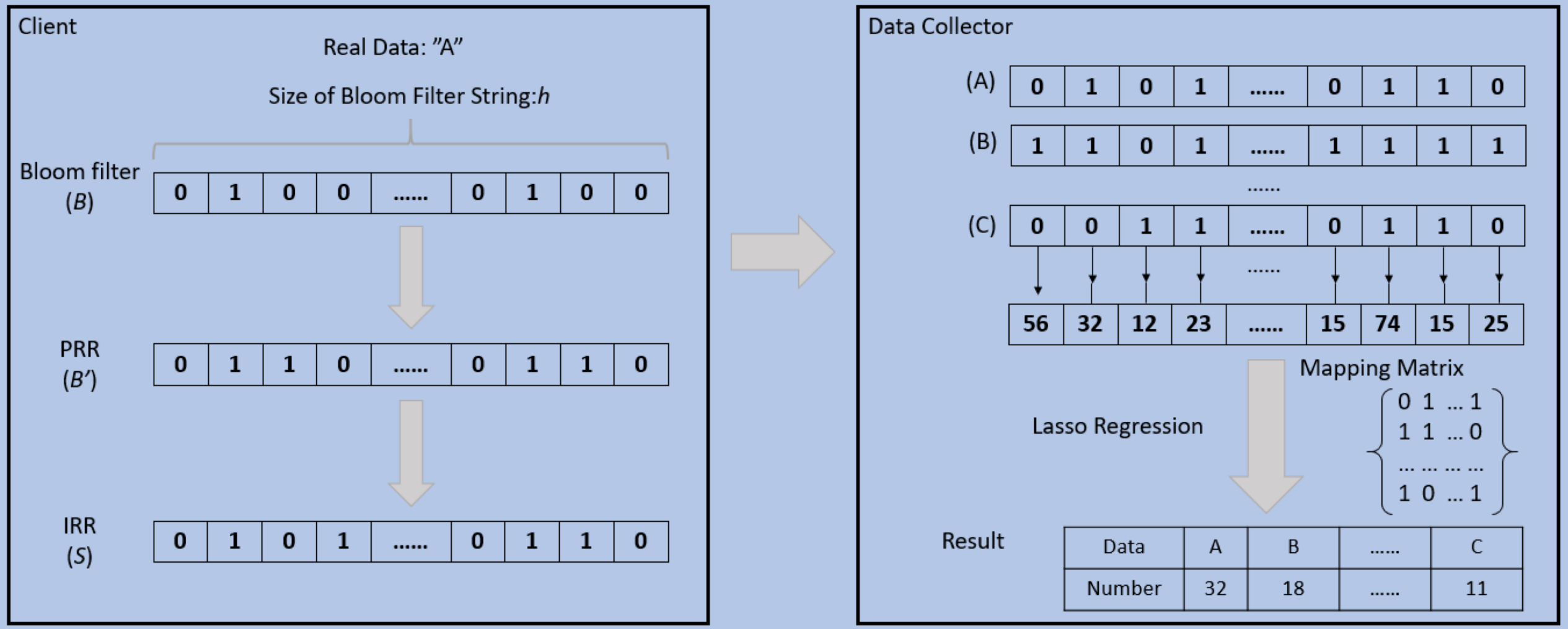
Tracks inputs in the Chrome browser **(URLs).**

Opensource implement @https://github.com/google/rappor

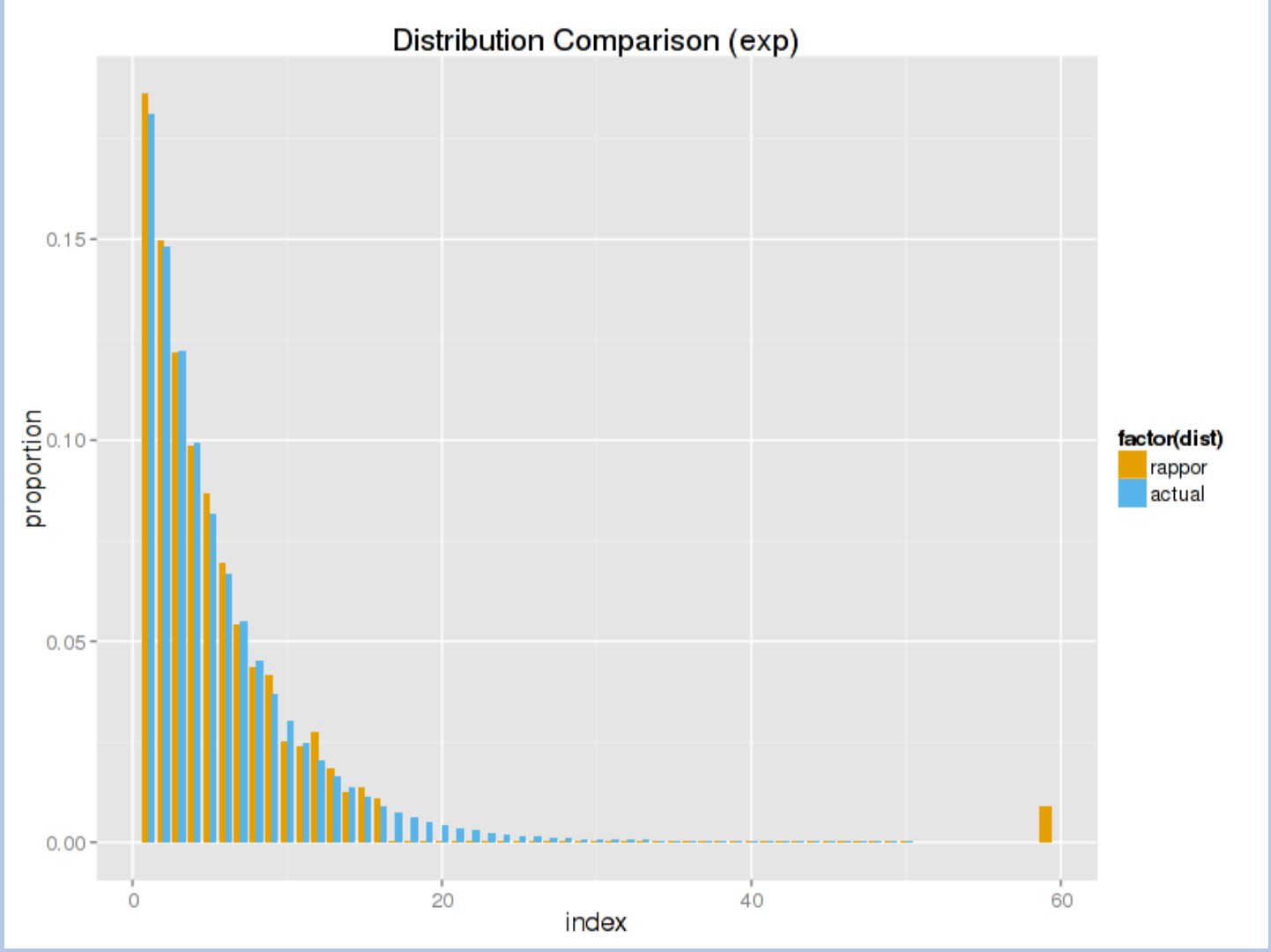
RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response



RAPPOR: LDP In Google



RAPPOR: LDP In Google



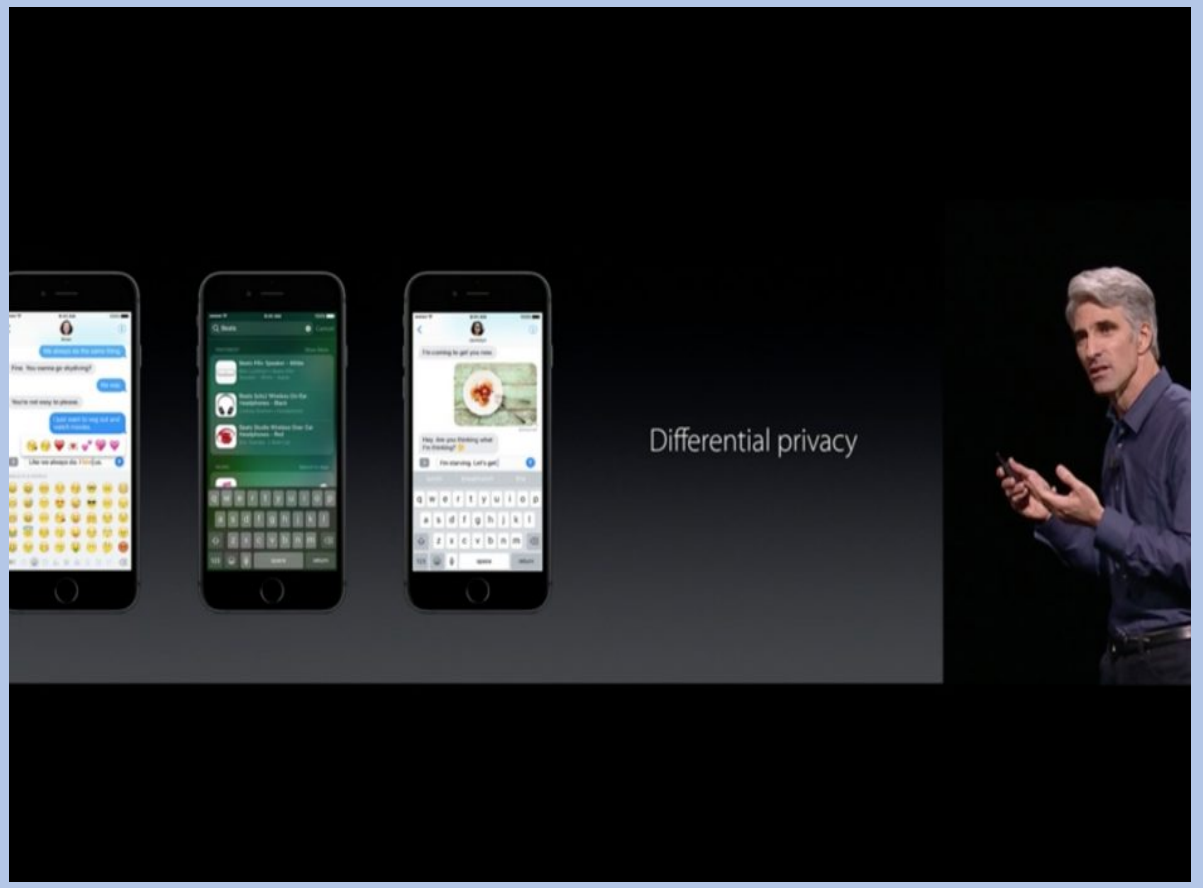
LDP In

Objective:
count frequencies of many items

1:57

DifferentialPrivacy_2018-10-20-1...


```
{
"version": 21,
"segments": [
{
"algorithm": "MultiBitHistogram",
"key": "com.apple.health.datatypes.usage.monthly",
"parameters": {"epsilon":2,"p":104},
"records": [
"160C147051F441031127341891"
]
},
{
"algorithm": "CountMedianSketch",
```

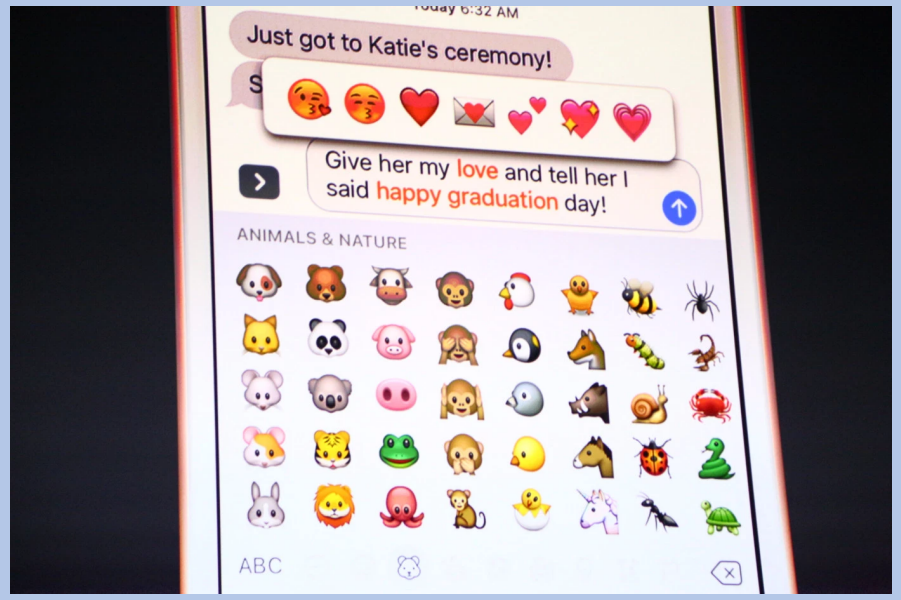



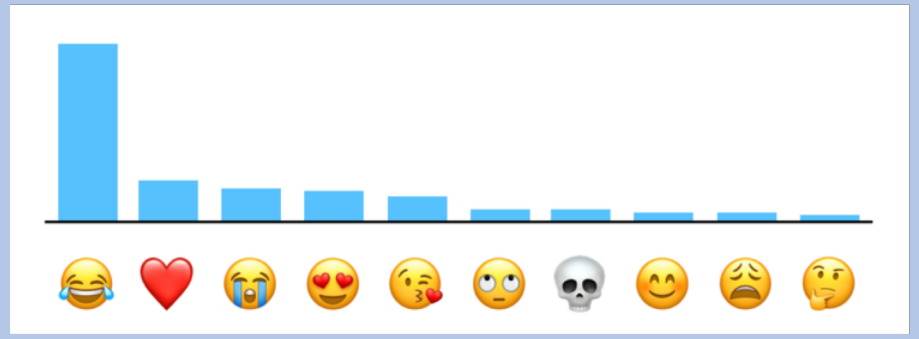
WWDC 2016

LDP In

Apple uses their system to collect data from iOS and MacOS users.

- **Popular emojis**: (heart) (laugh) (smile) (crying) (sadface)
- **New words**: bruh, hun, bae, tryna, despacito, mayweather
- **Which websites to mute, which to autoplay audio on!**



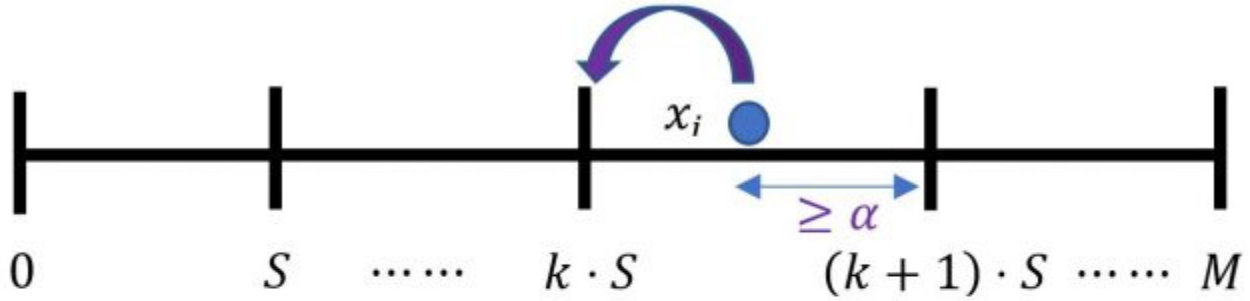


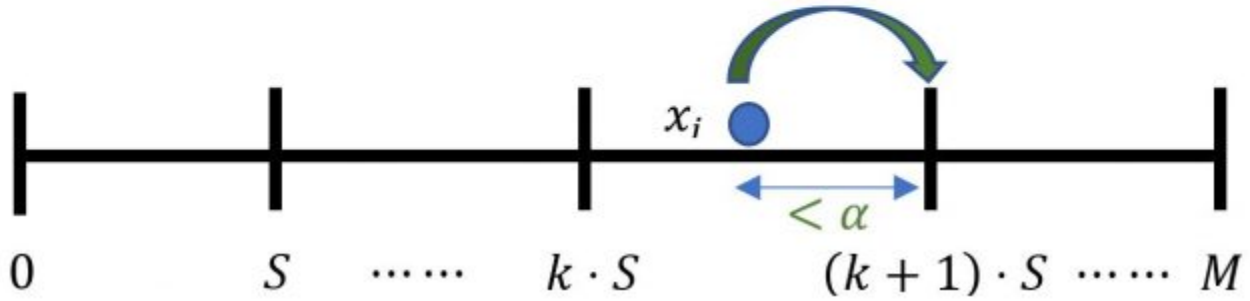
Slide from "Privacy at Scale: Local Differential Privacy in Practice"

Telemetry Collection: LDP In Microsoft

Objective:
Collects number of seconds users spend in different apps

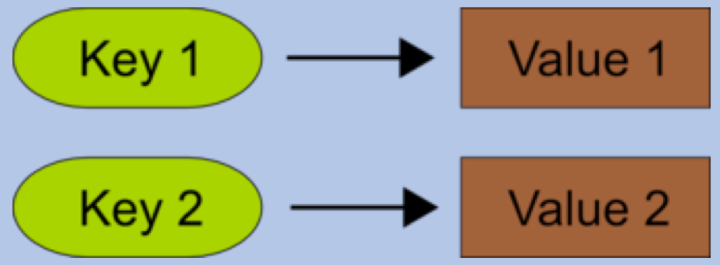
Deployed in Windows 10 Fall Creators Update 2017 Oct.



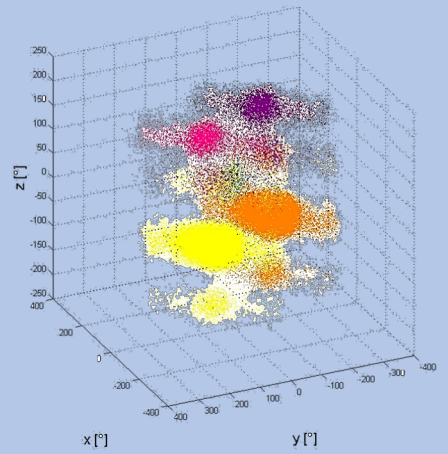


0	$A(0)$
S	$A(S)$
$2S$	$A(2S)$
$\cdots$	$\cdots$
$k \cdot S$	$A(k \cdot S)$
$(k+1) \cdot S$	$A((k+1) \cdot S)$
$\cdots$	$\cdots$
M	$A(M)$

Challenges: Complicated Data Structure



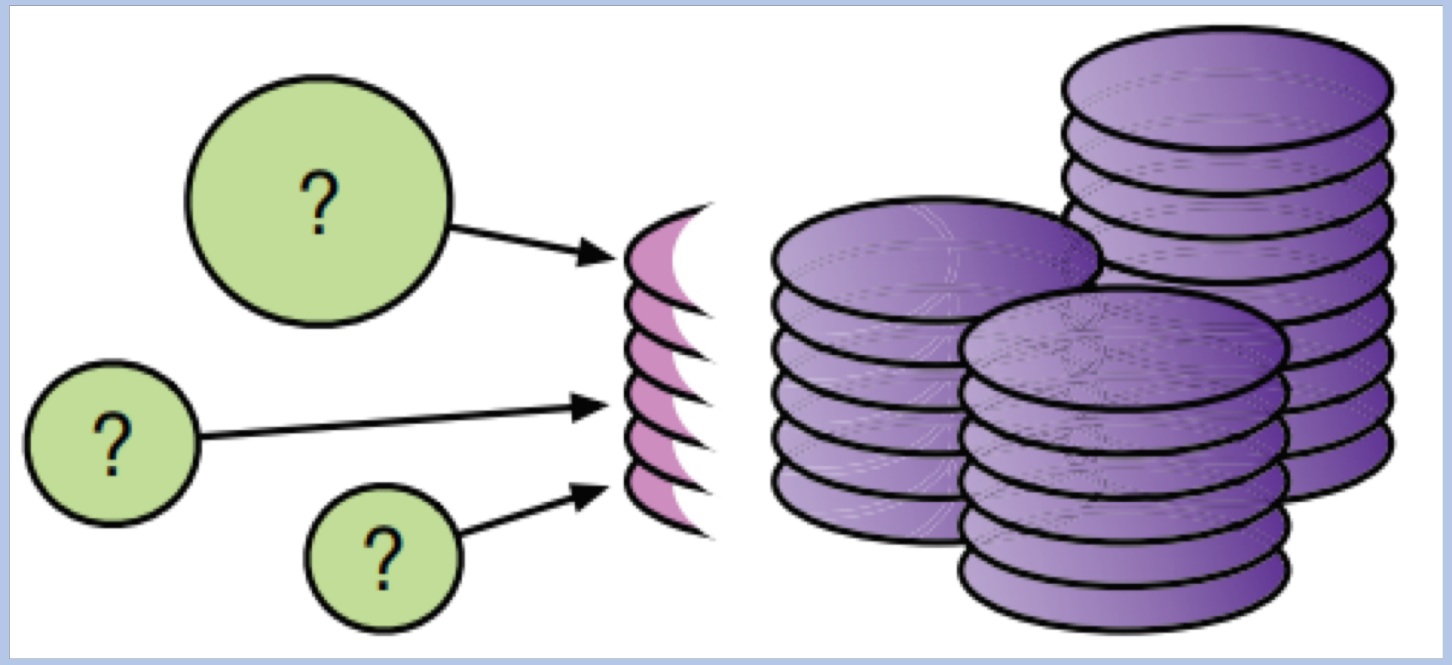
Key-Value Data



High-Dimensional Data

Graph Data

Challenges: Complicated Query Task



The End