Protecting Privacy in Mobile, Social Network, & Cloud Computing

Xiang-Yang Li

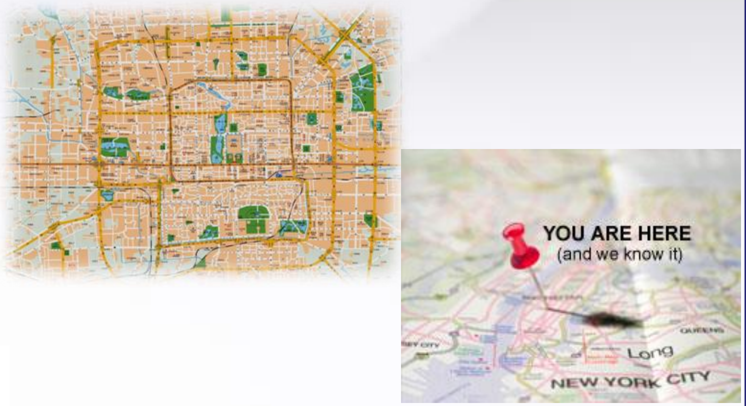
Illinois Institute of Technology,
Tsinghua University, EMC Visiting Chair Professor
www.cs.iit.edu/~xli
xli@cs.iit.edu

Results collaborated with: Taeho Jung, Lan Zhang, ShaoJie Tang, and many other students

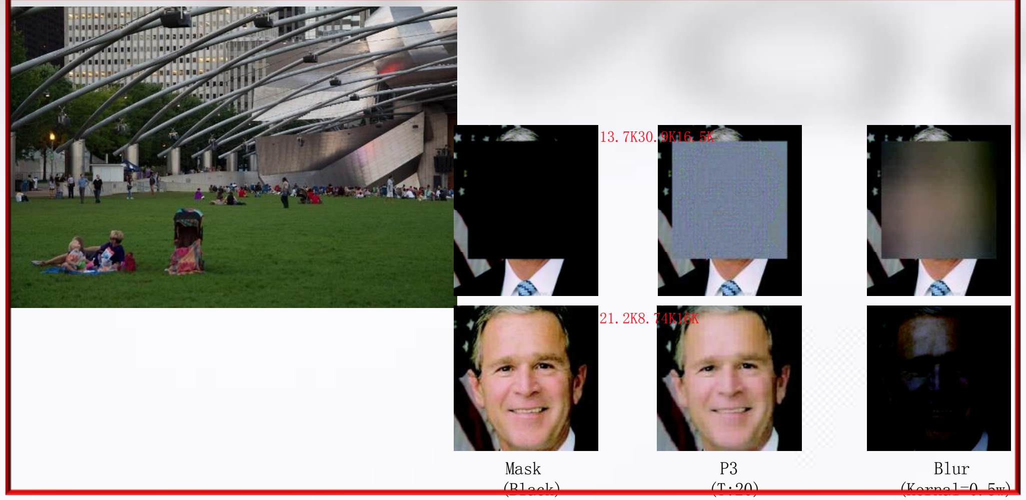
Acknowledgments

Students

Outline: Part I



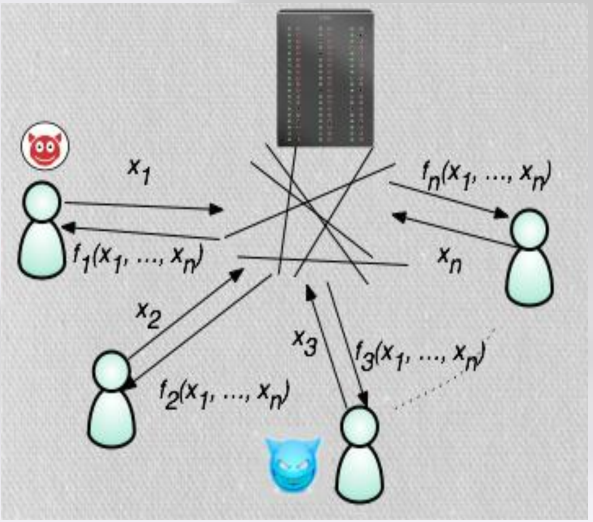
Trace and Location

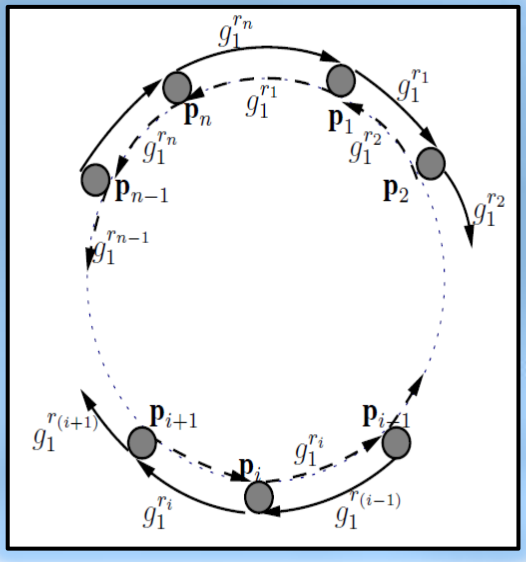


Picture and Image Search

Smart Devices

Outline: Part II

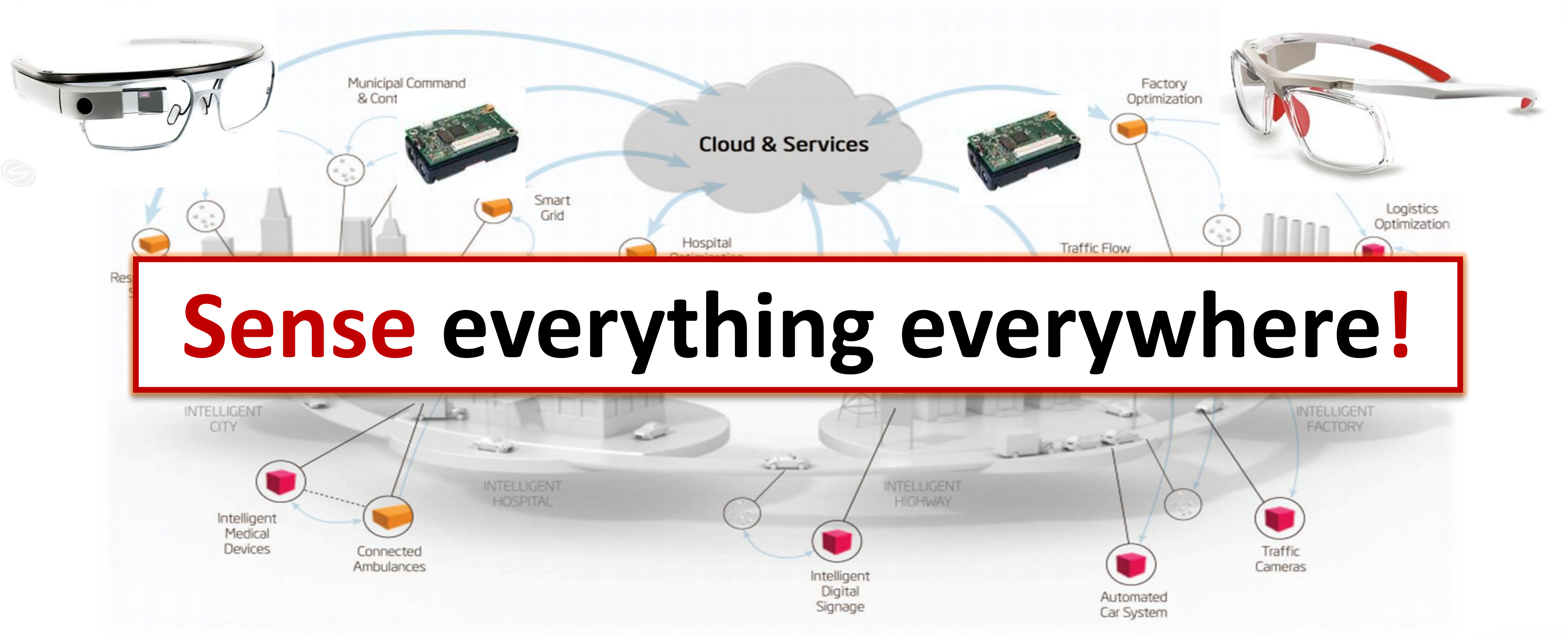




Theoretical Framework: Efficiency, Privacy, and Verifiable

Motivation:
Mobile, Social, Cloud and Privacy

Internet of Things



Sense everything everywhere!

Mobile Social Networks

Contacts

Schedule

Share everything everywhere!

Transportation

Activities

Online User Behavior

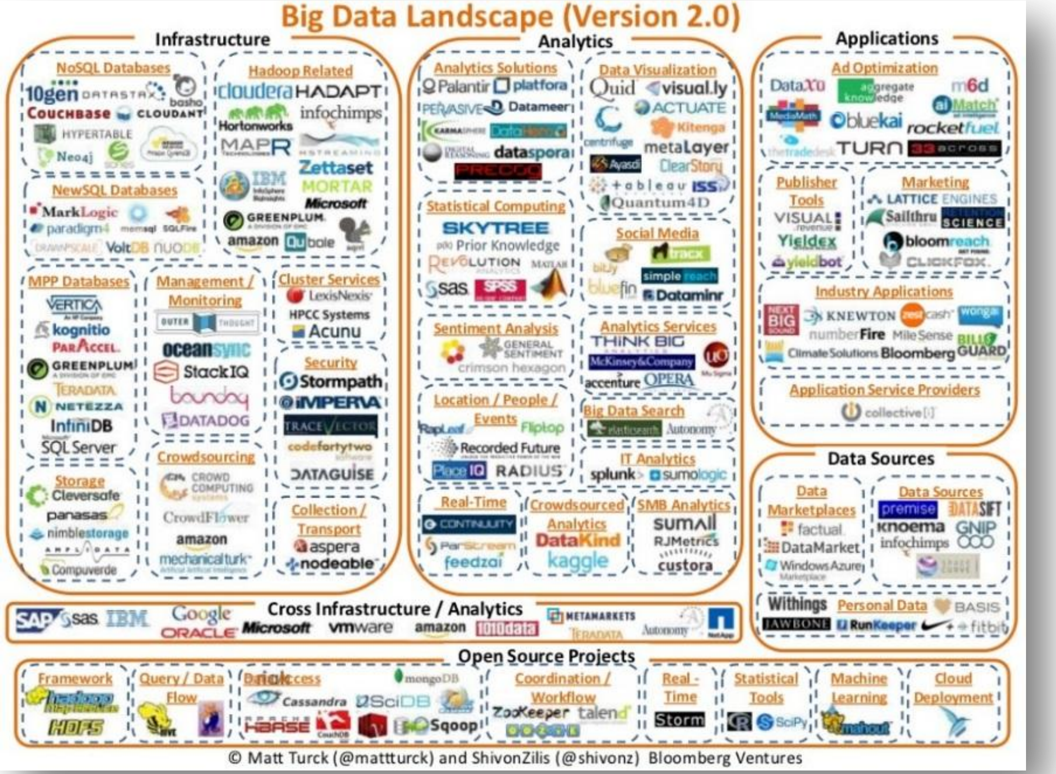
From
- Online payment
- Online browsing
- Electronic Medical record
- C
- S

Infer everything everywhere!

Infer
- Your age
- Your profession
- Your income
- Home address
- emotions
- ….

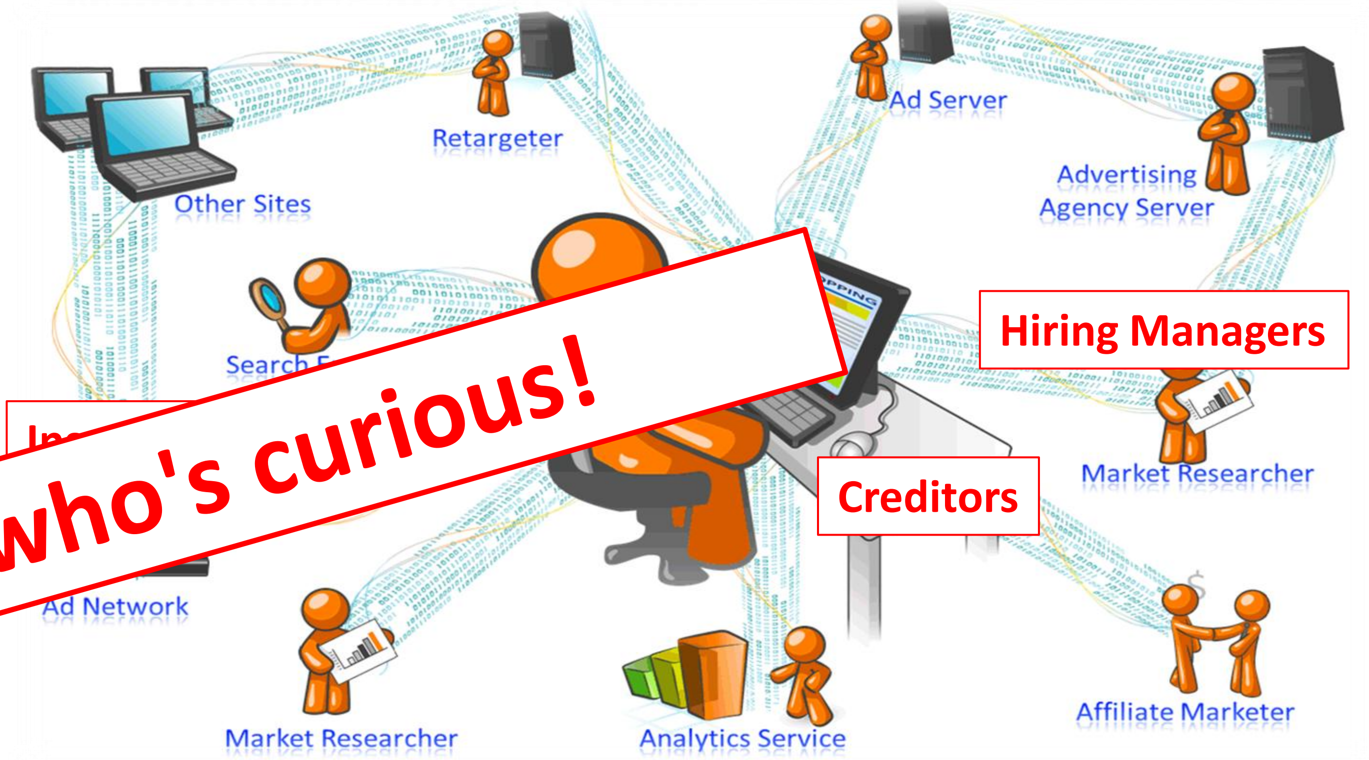
More risk in big data

- **12 TB** of Tweets
- **1G photos,10M videos** per week
- **5 million trade events**
- **3PB camera** data per day in Beijing
- **2.7ZB** data created in 2012
  =2,700,000,000TB



Big data may increase the power and prevalence of privacy leakages.

Who can observe those?

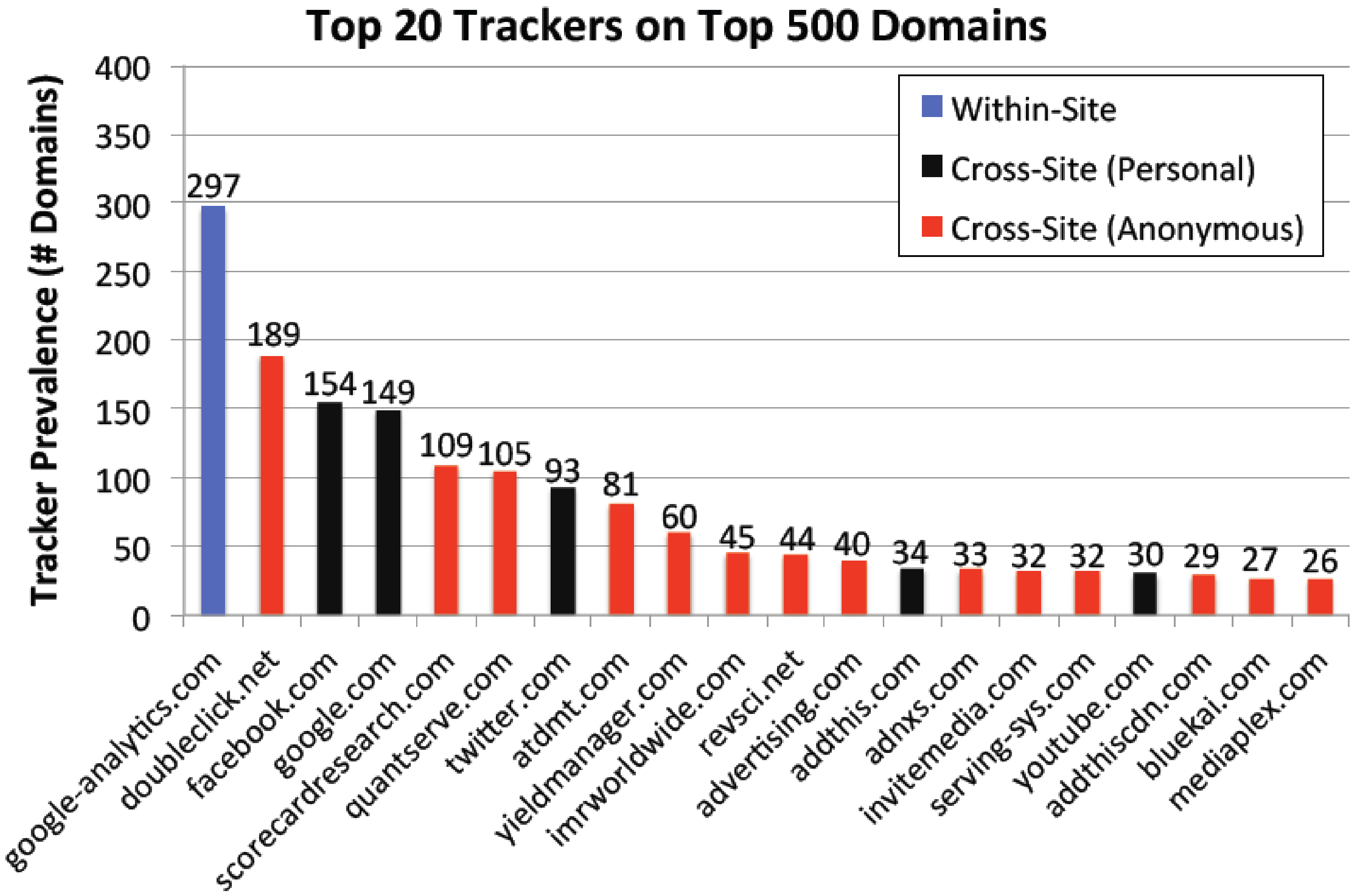


Hiring Managers

Creditors

Anyone who's curious!

Not Trustworthy Companies



Snowden Effect: We do care, a lot!

- > 70%: No one should ever be allowed to have access to my personal data or web behavior.
- < 25%: ok with trading some of personal information in exchange for more relevant advertising.

Scary?

Countermeasures??

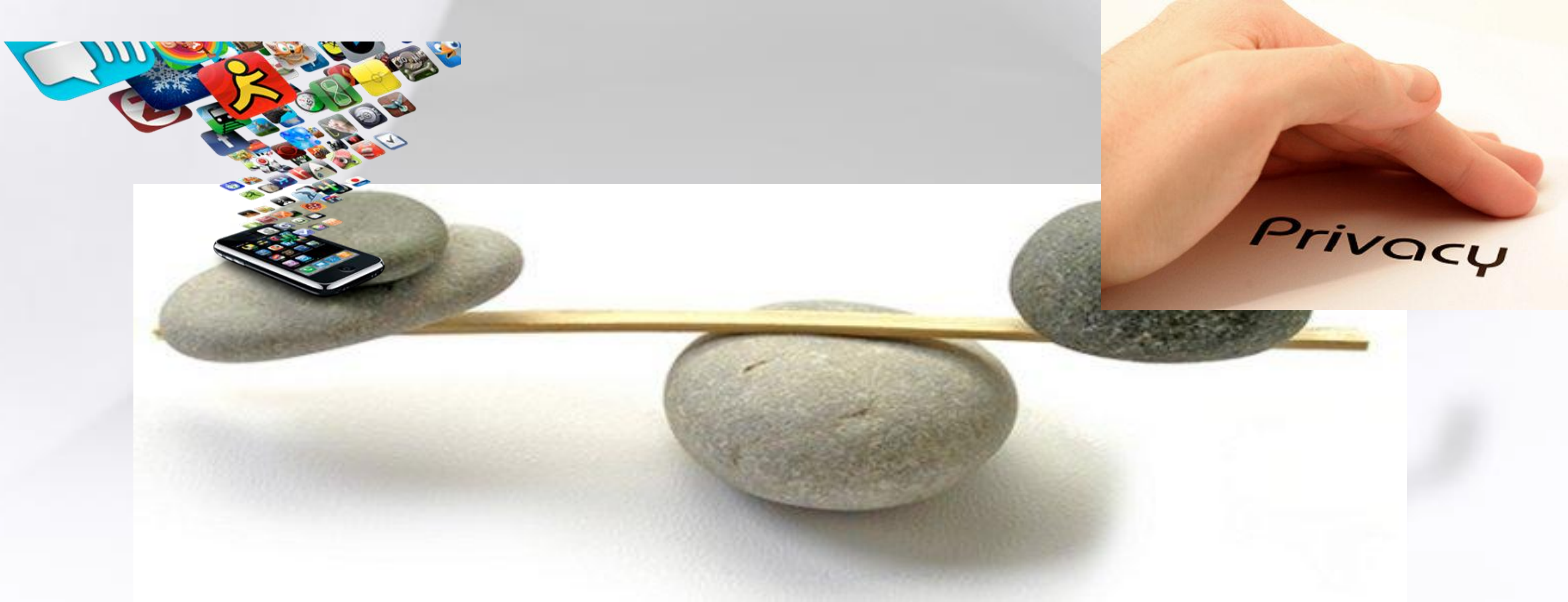
Lock Everything？

- **Example : computing and search functionalities disabled.**

Lock Everything？

- Problems : Data quality, utility, networking, etc.

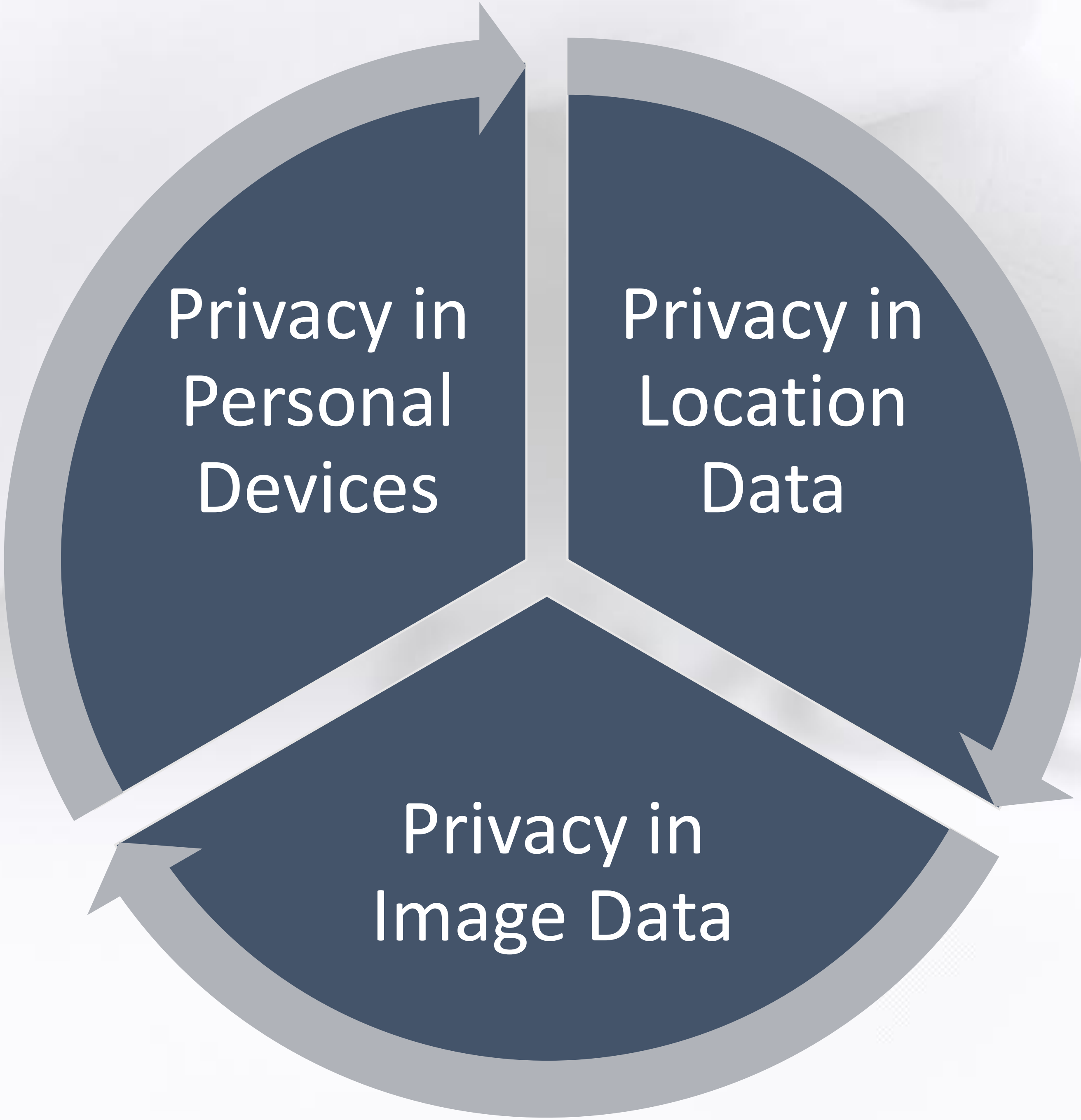
Our Goal

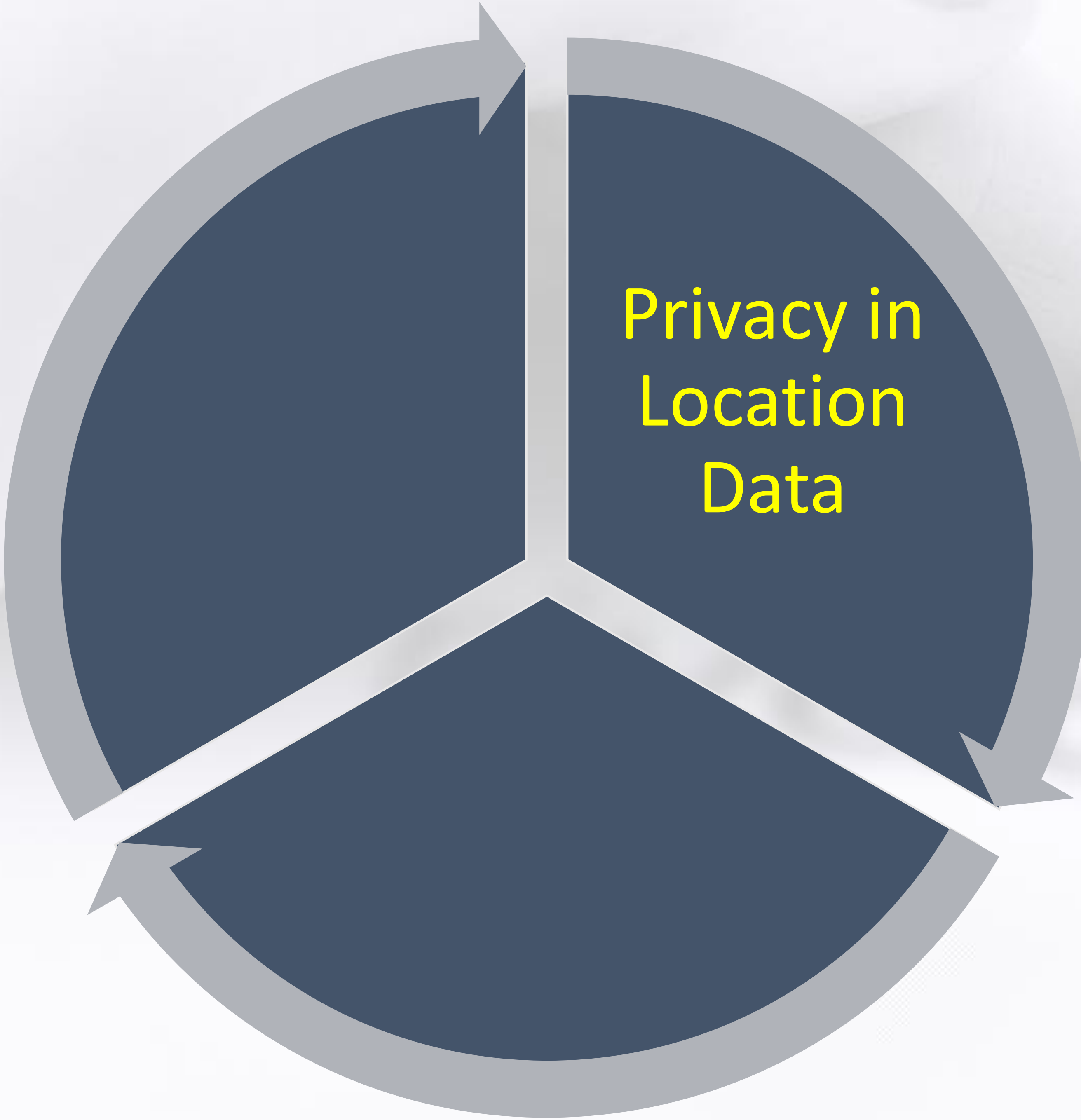


Privacy Issues in Real World

Protections on Data

in mobile social networks & mobile devices


Privacy in Personal Devices
Privacy in Location Data
Privacy in Image Data


Privacy in Location Data

Trace Leakage in Crowdsoured Map

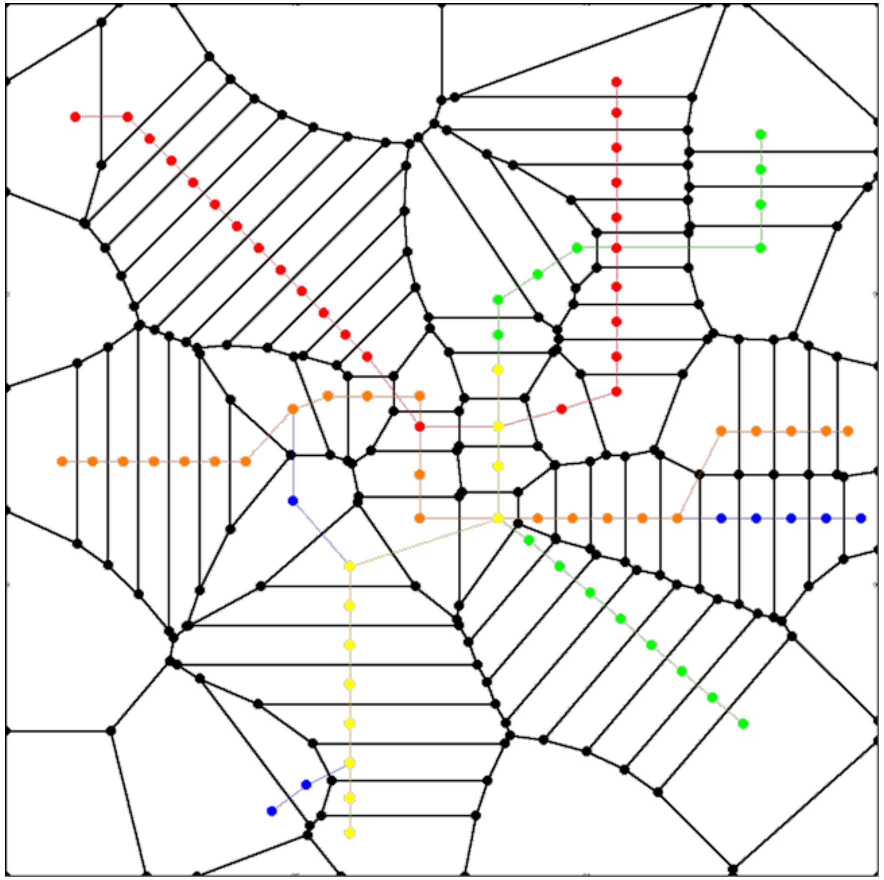
- Crowdsourcing using location data reported by users.

Privacy ⟷ **Trade-off** ⟷ **Accuracy**

Trace from location data

in Real World

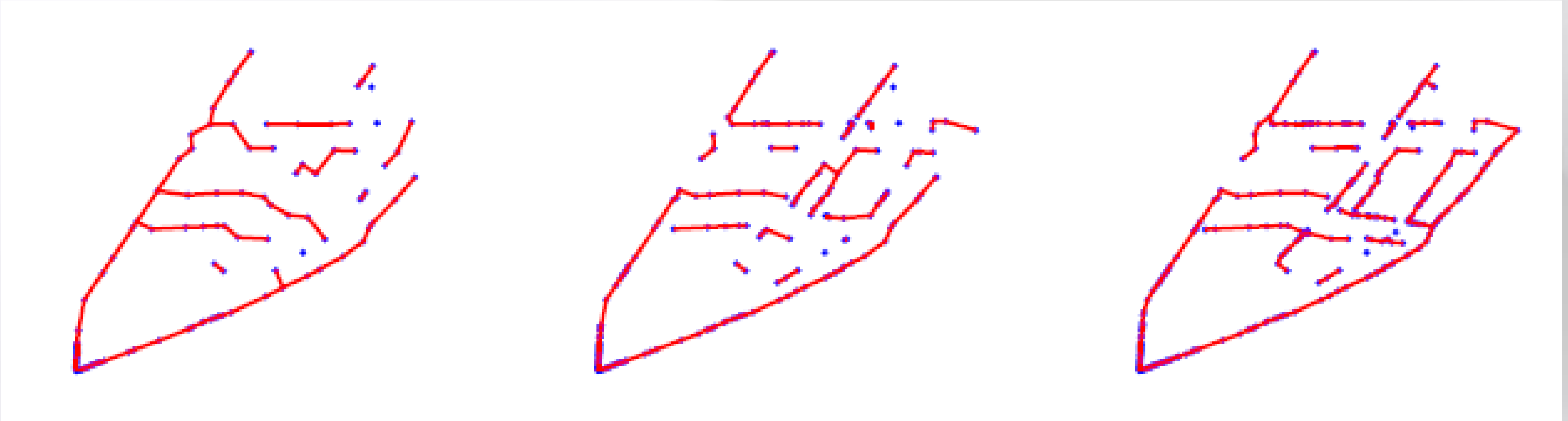
Make it impossible!

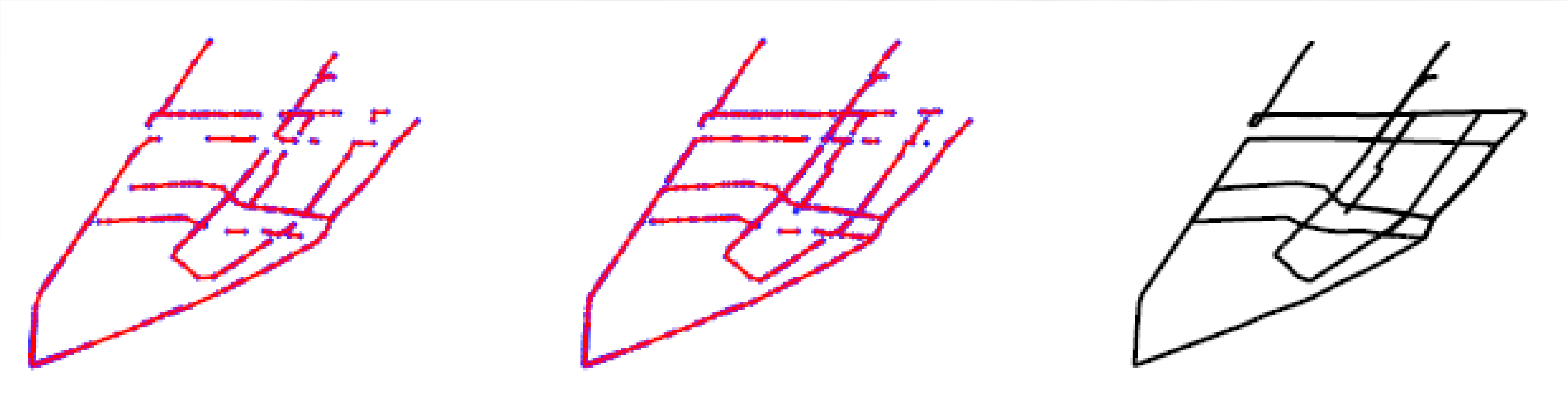


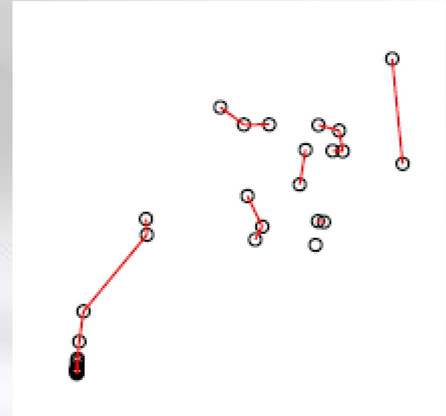
- Exploit the Voronoi diagram's properties and curve-reconstruction properties
- Manipulate the data publication
  - data density related to **curvature** of the route

Make Route reconstruction become an unsolvable problem!

Good accuracy, Good Privacy

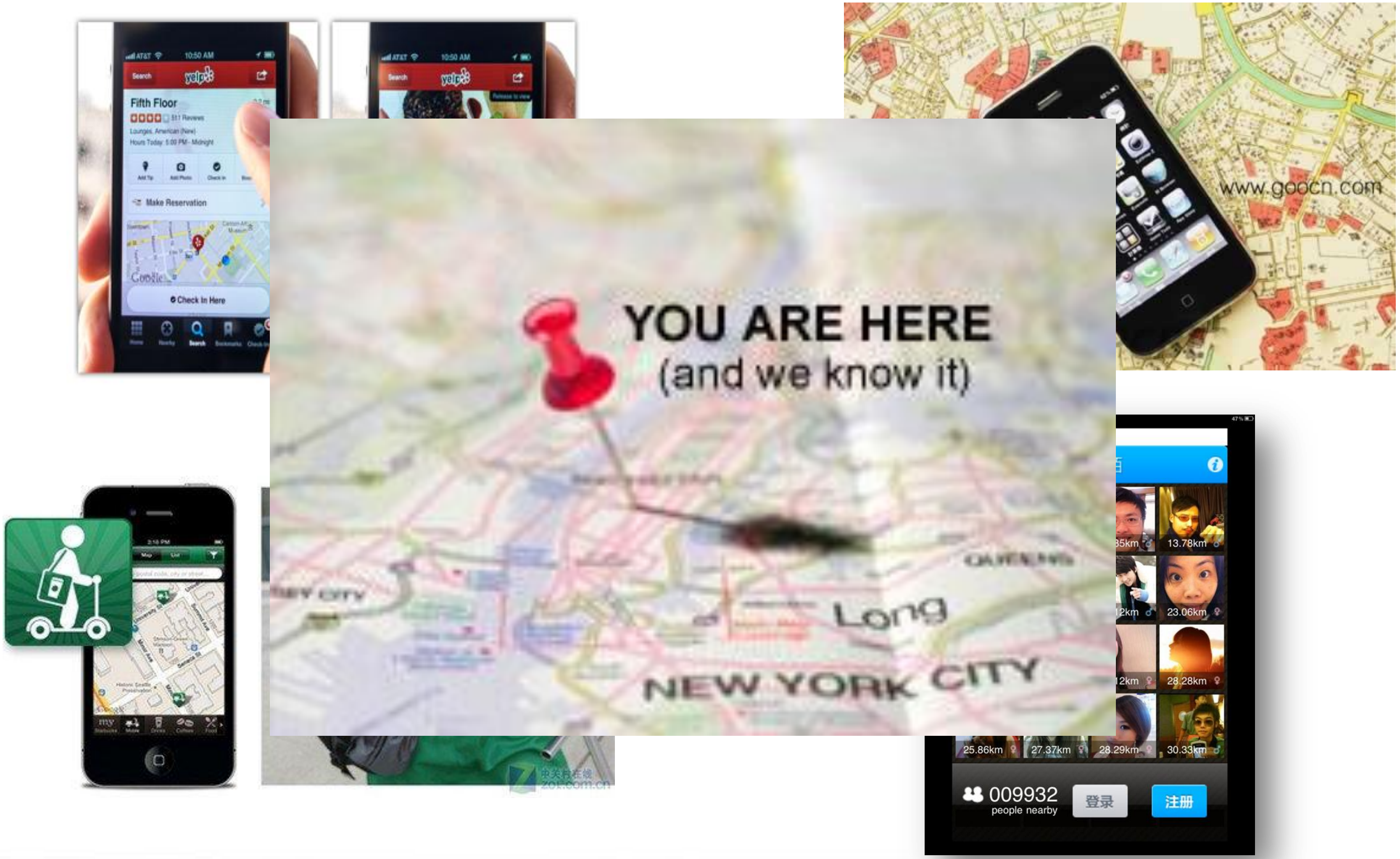






Location, Location, Location

Location leakage in LBS



Search Me If You Can: Privacy-Preserving Location Query Service（IEEE INFOCOM, 2013）

in Real World

Our Design Strategy

- You publish something,
  only **authenticated** people get
  only **tailored** information
  at your **specified context**.

Attribute Based Encryption

Functional Encryption

Big picture of the solution

1. Encrypted location
2. Encrypted query
3. Relayed encrypted query
4. Encrypted result

Homomorphic Operations

$$\boxed{\mathcal{E}(1), \mathcal{E}\left(\sum_{i=1}^{3} y_i^{\,2}\right), \mathcal{E}(-2y_i)_{\mathbf{i=1,2,3}}}$$

Encrypted query

$$\mathcal{E}(-2y_i)^{x_i} = \mathcal{E}(-2x_i y_i), \mathcal{E}\left(\sum_{i=1}^{3} y_i^{\,2}\right)^{\delta} = \mathcal{E}\left(\sum_{i=1}^{3} y_i^{\,2}\right)$$

$$\mathcal{E}(1)^{\sum_{i=1}^{3} x_i^{\,2}} = \mathcal{E}\left(\sum_{i=1}^{3} x_i^{\,2}\right)$$

$$\mathcal{E}\left(\sum_{i=1}^{3} x_i^{\,2}\right) = \mathcal{E}\left(\sum_{i=1}^{3} x_i^{\,2}\right)$$

$$\mathcal{E}\left(\sum_{i=1}^{3} x_i^{\,2}\right) \cdot \mathcal{E}\left(\sum_{i=1}^{3} y_i^{\,2}\right) \cdot \prod_{i=1}^{3} \mathcal{E}(-2x_i y_i) = \mathcal{E}(|\vec{x} - \vec{y}|^2)$$

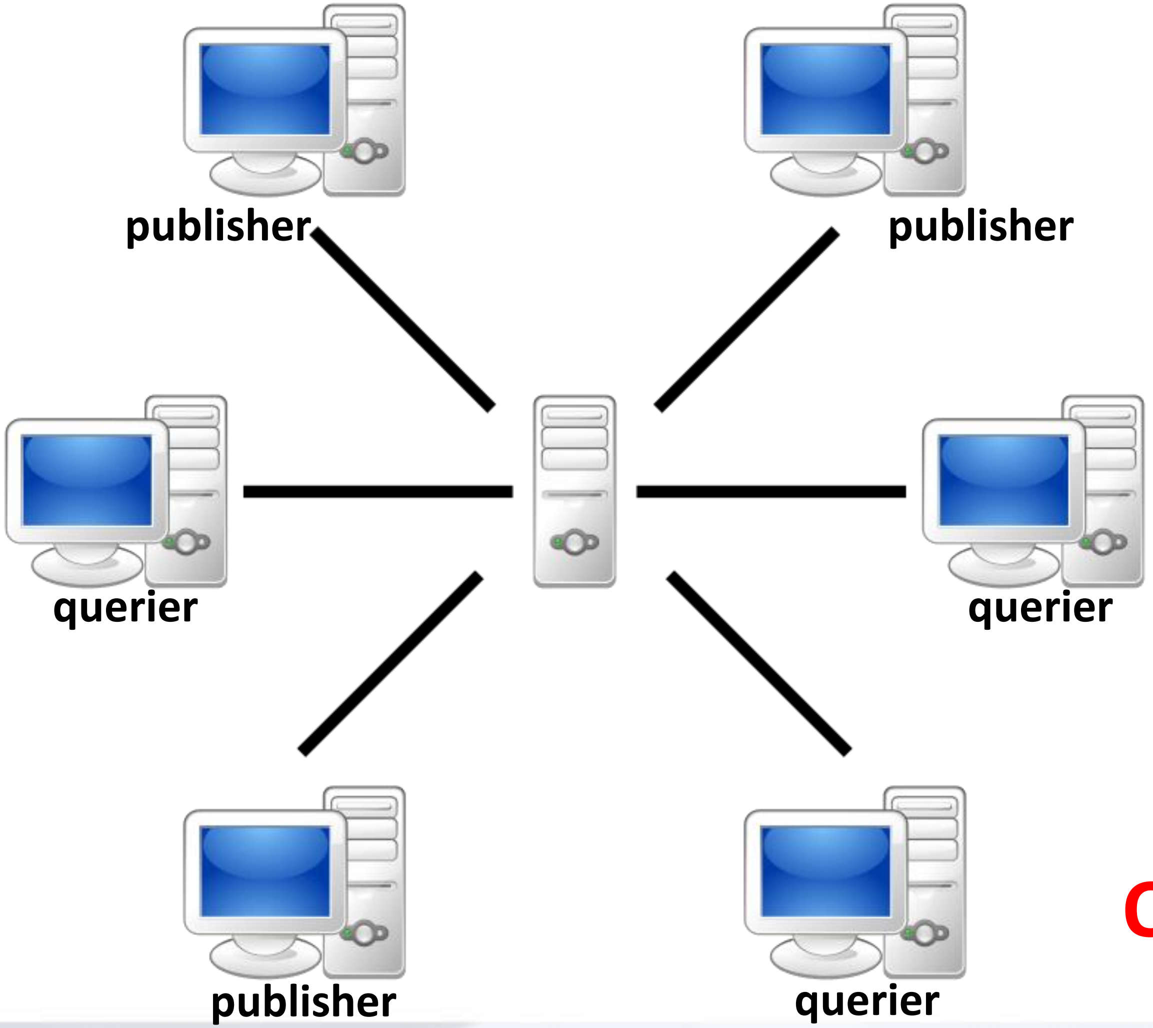
Publisher

Querier

$$\boxed{= \mathcal{E}(|\vec{\boldsymbol{x}} - \vec{\boldsymbol{y}}|^2)}$$

Encrypted result

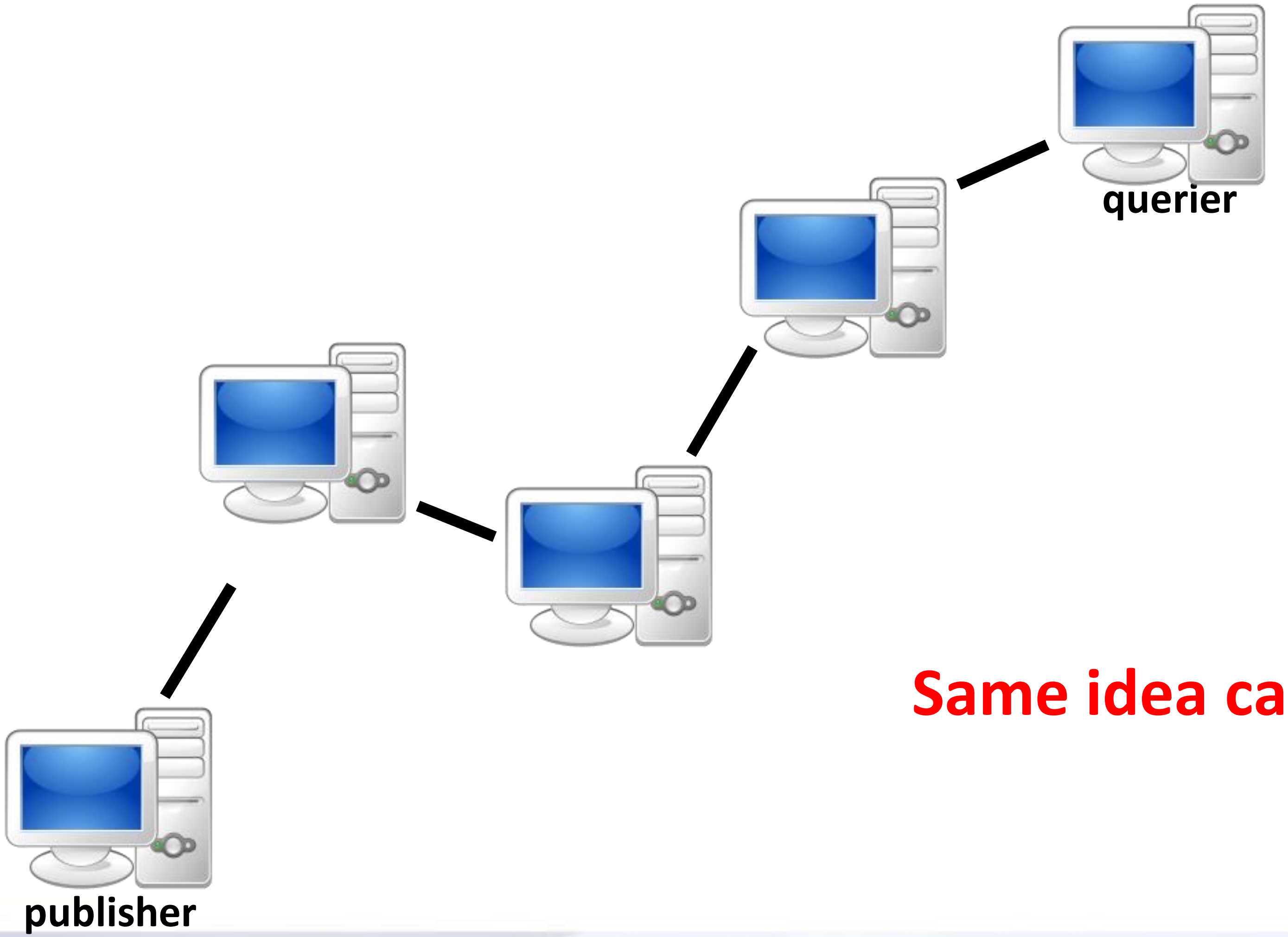
System Model



C/S model

Our solution relies on C/S model

Different System Model

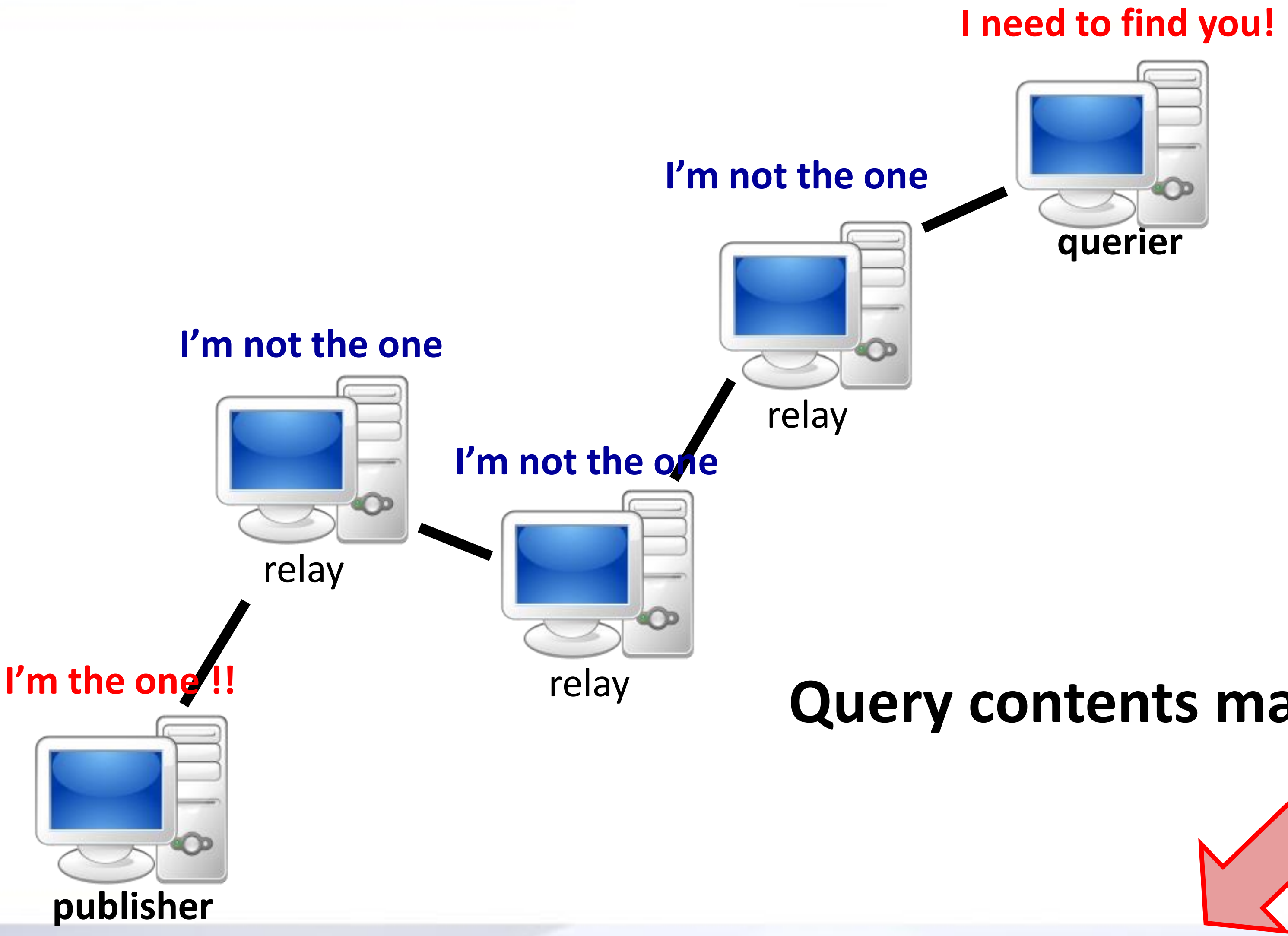


C/S model

P2P model

Same idea can be applied in a P2P network too

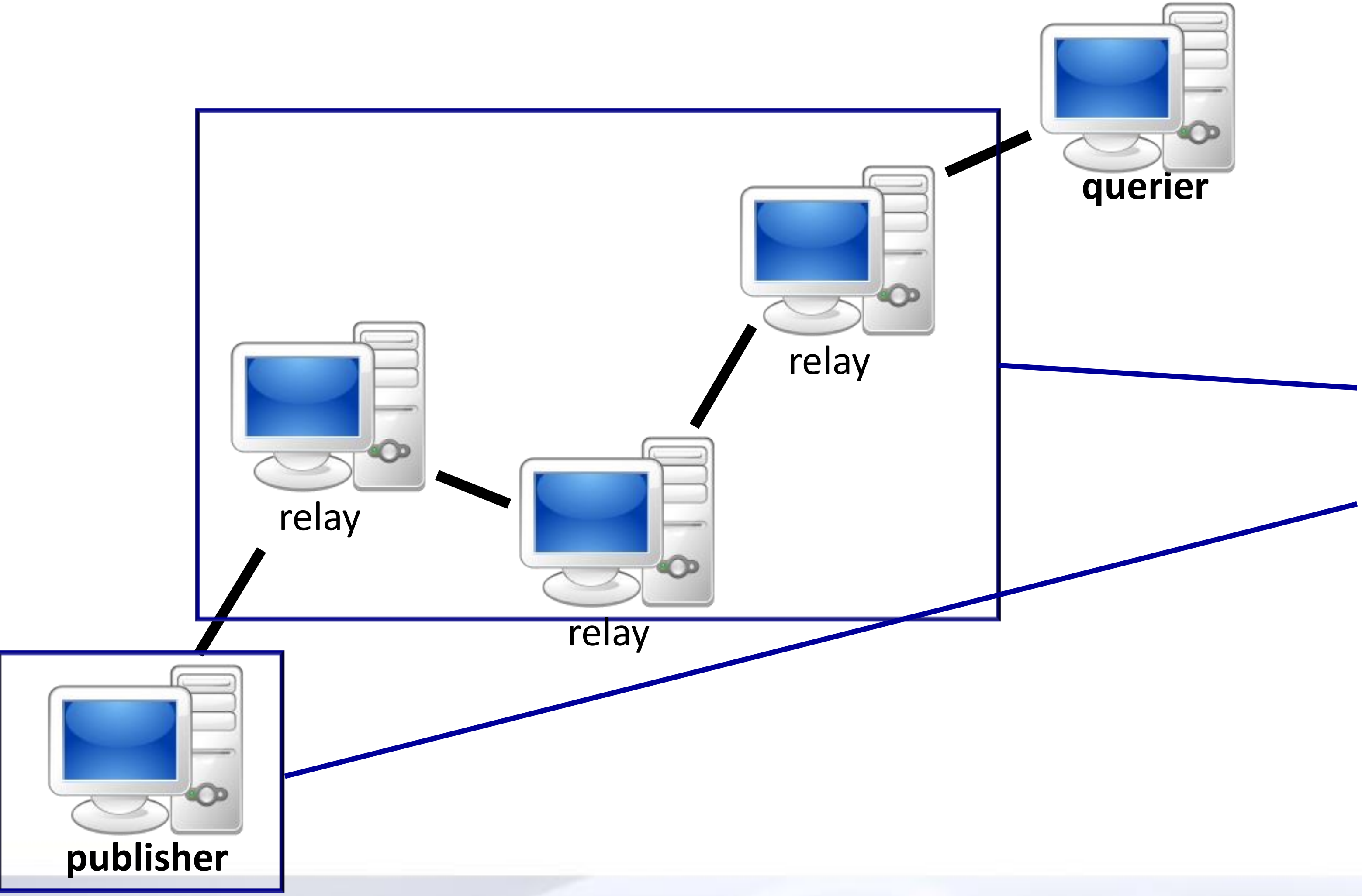
Sketch of Our Solution in P2P



P2P model

Query contents may be general profile information

Sketch of Our Solution in P2P

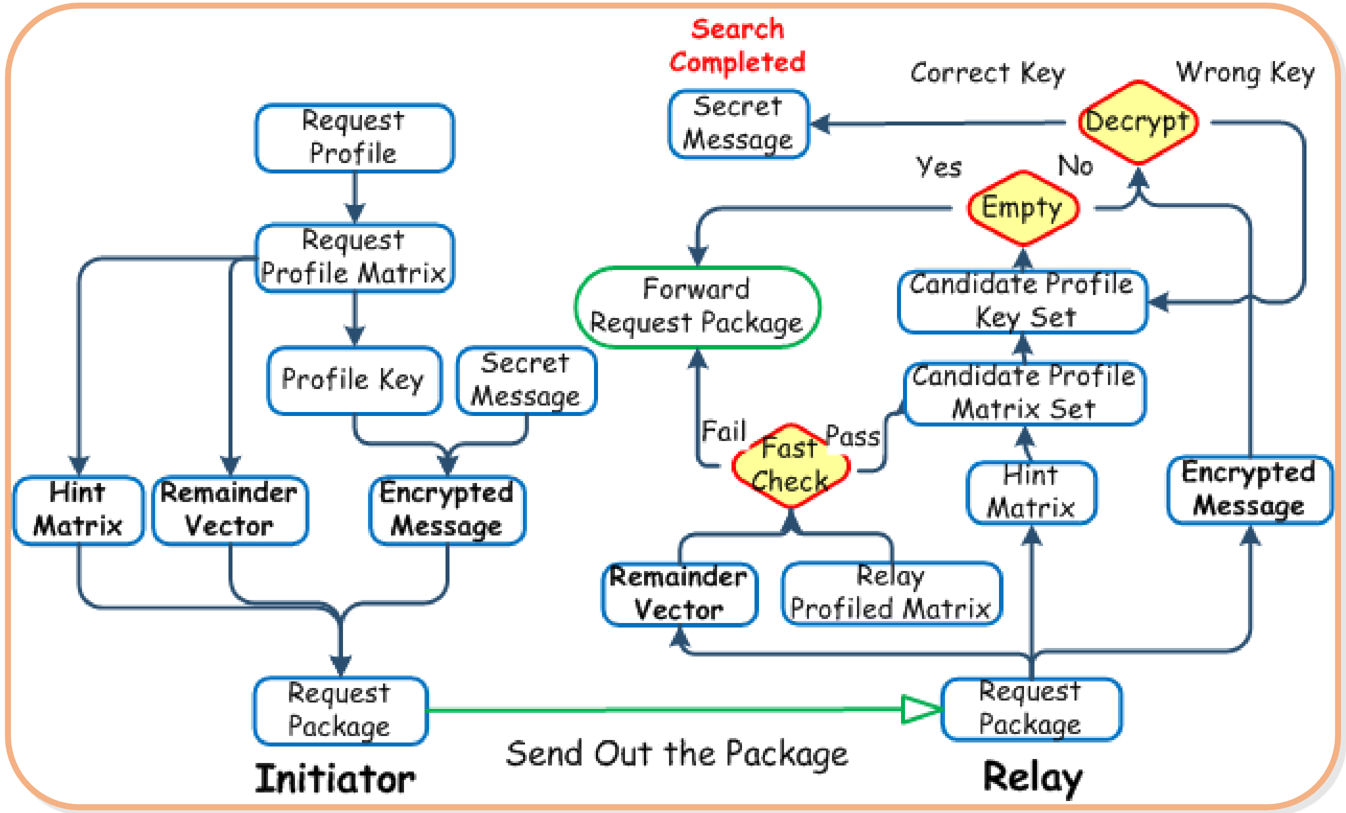


Two Critical Parts

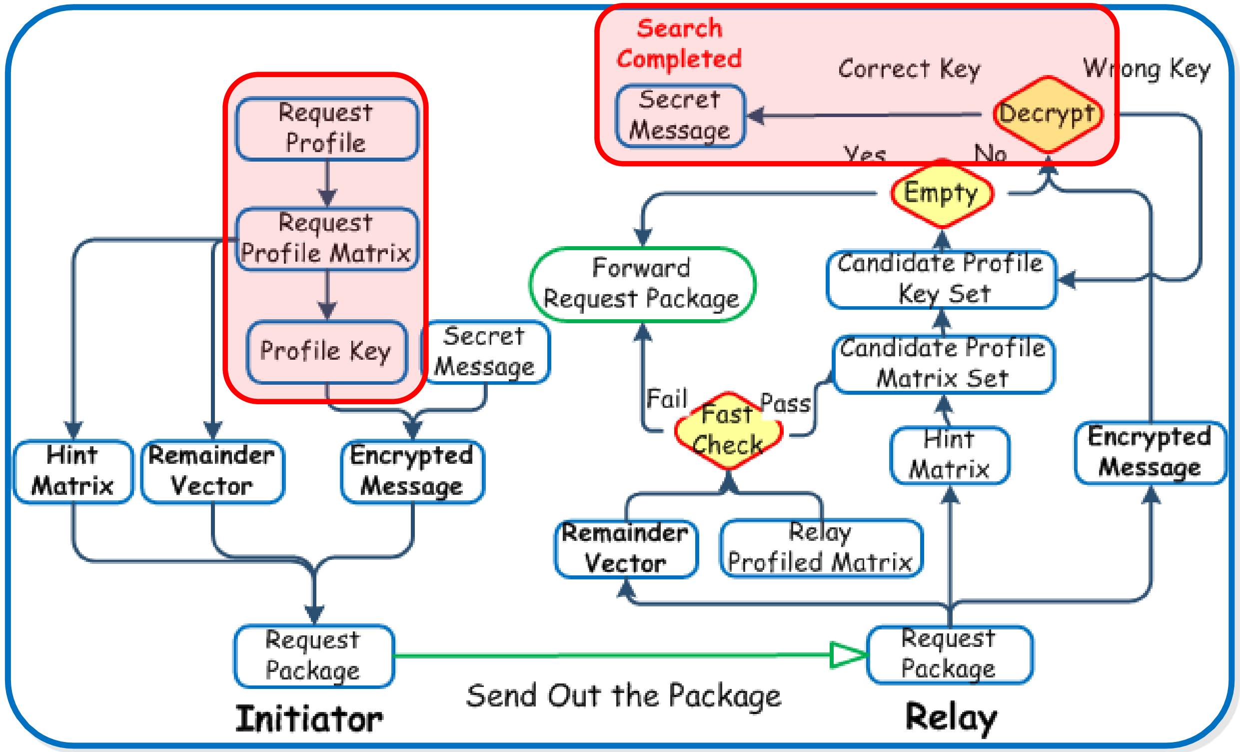
1. ID examination
2. Query response

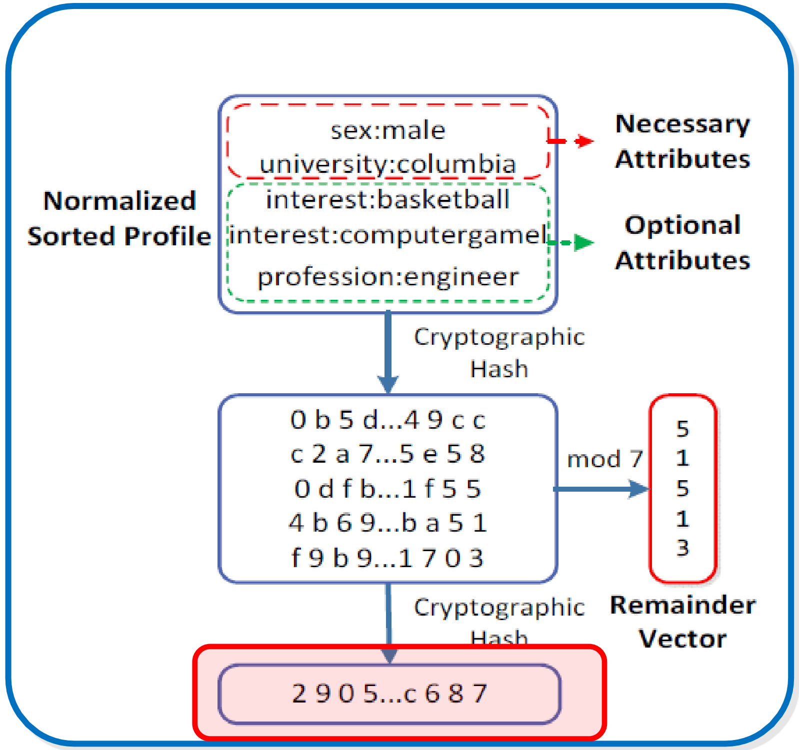
Basic Mechanism

- Use **common attributes between matching users** to encrypt a message with a secret channel key
- Only a matching user can decrypt the message **efficiently**.
- In one round simultaneously
  - privacy-preserving matching
  - secure channel construction



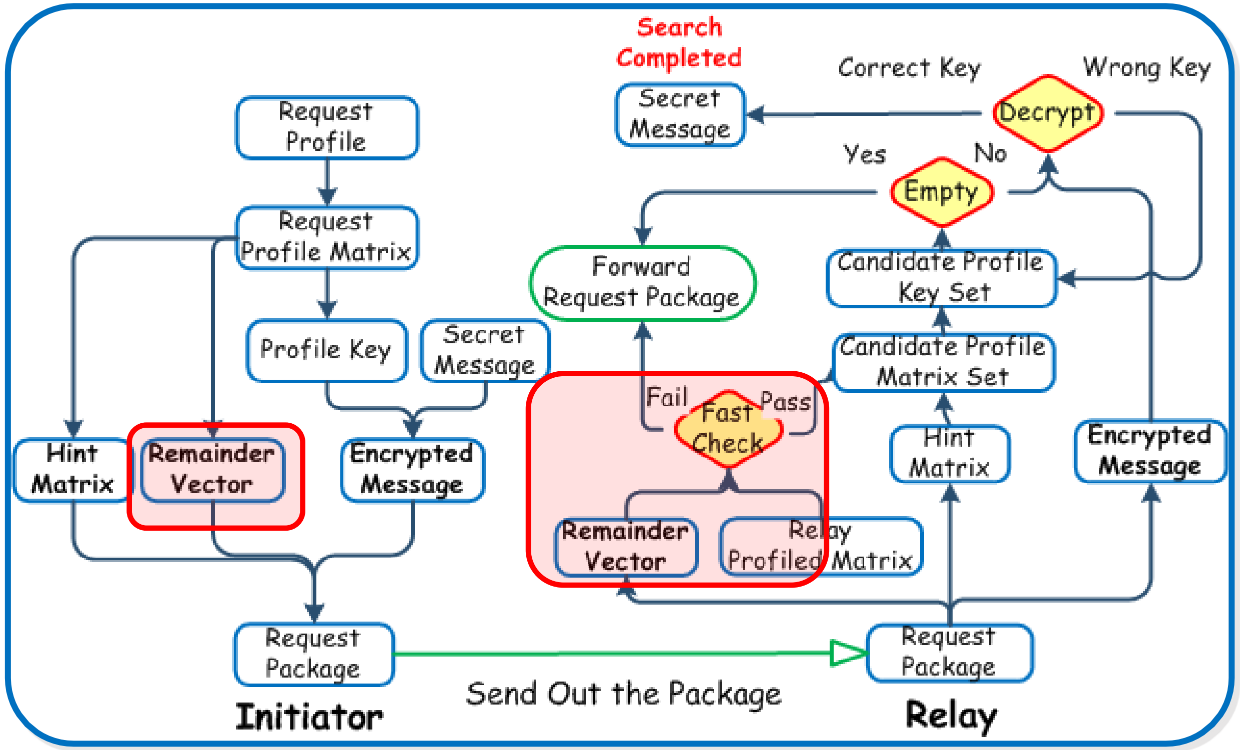
1. Profile Key Generation

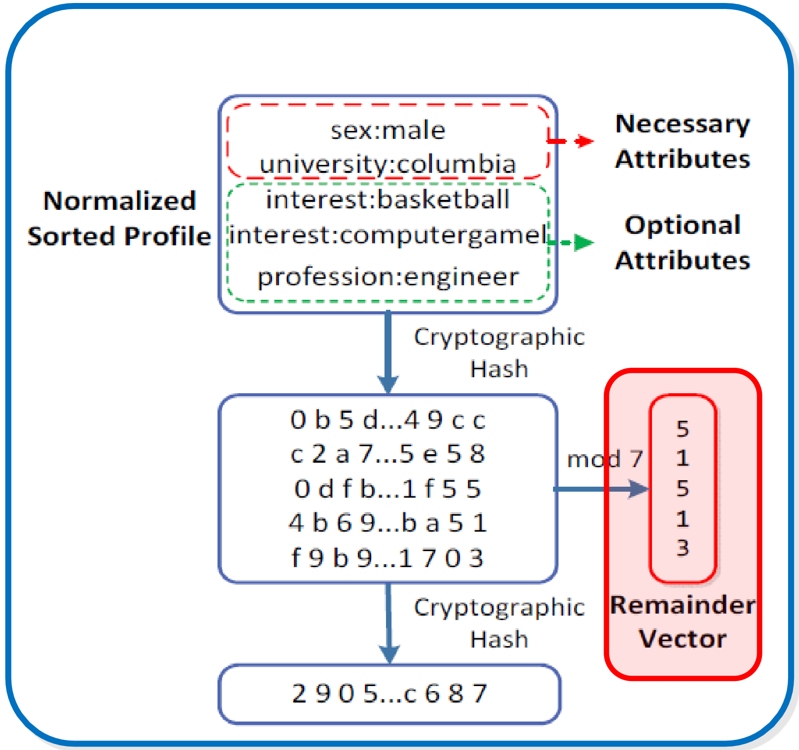




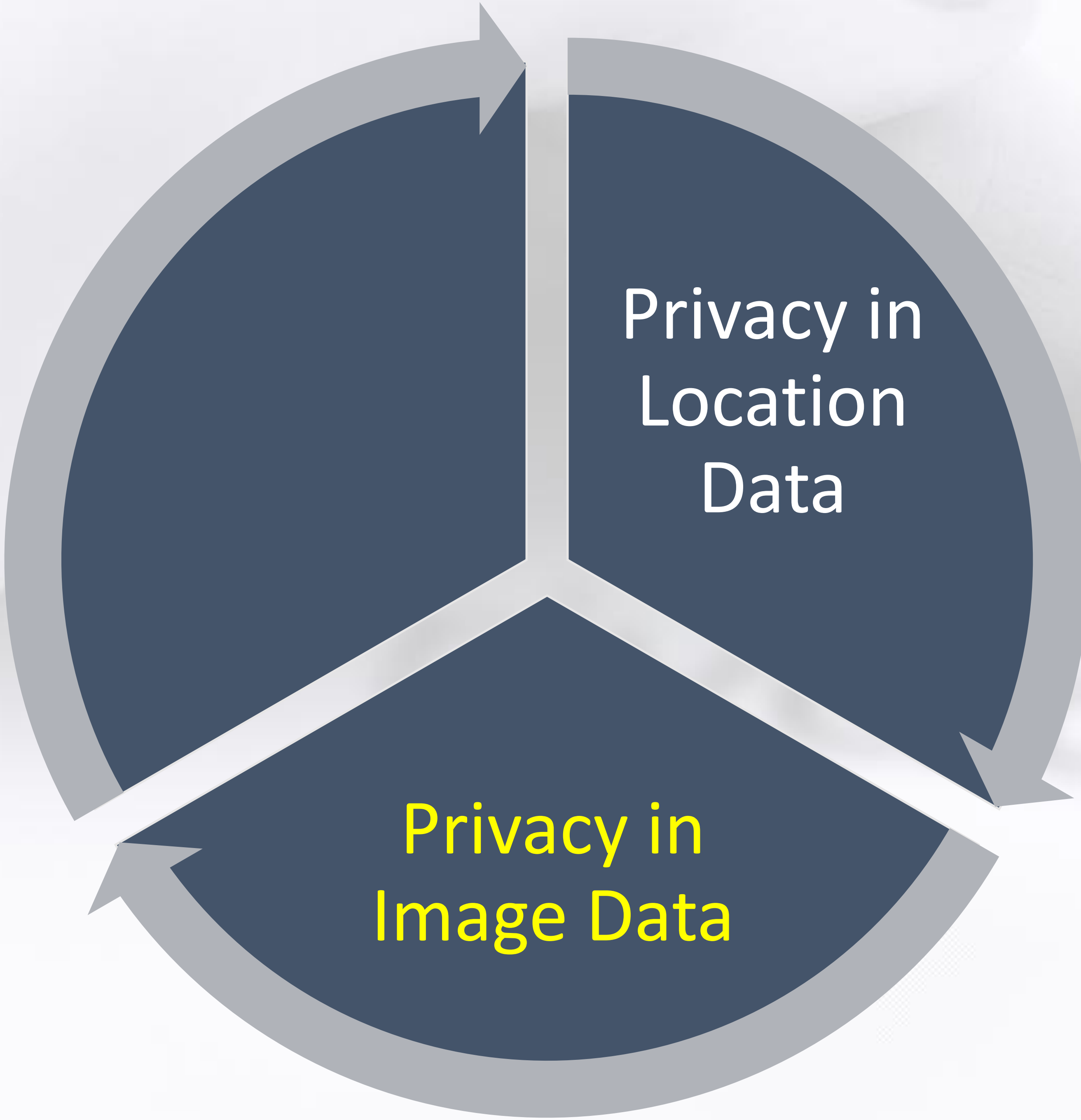
- **Profile key** is generated from request profile.
- Used to **encrypt communication key**.

2. Remainder Vector





- **Remainder vector** of the profile vector is yield for **fast exclusion by a large portion of unmatched persons**.


Privacy in Location Data
Privacy in Image Data

Privacy in images

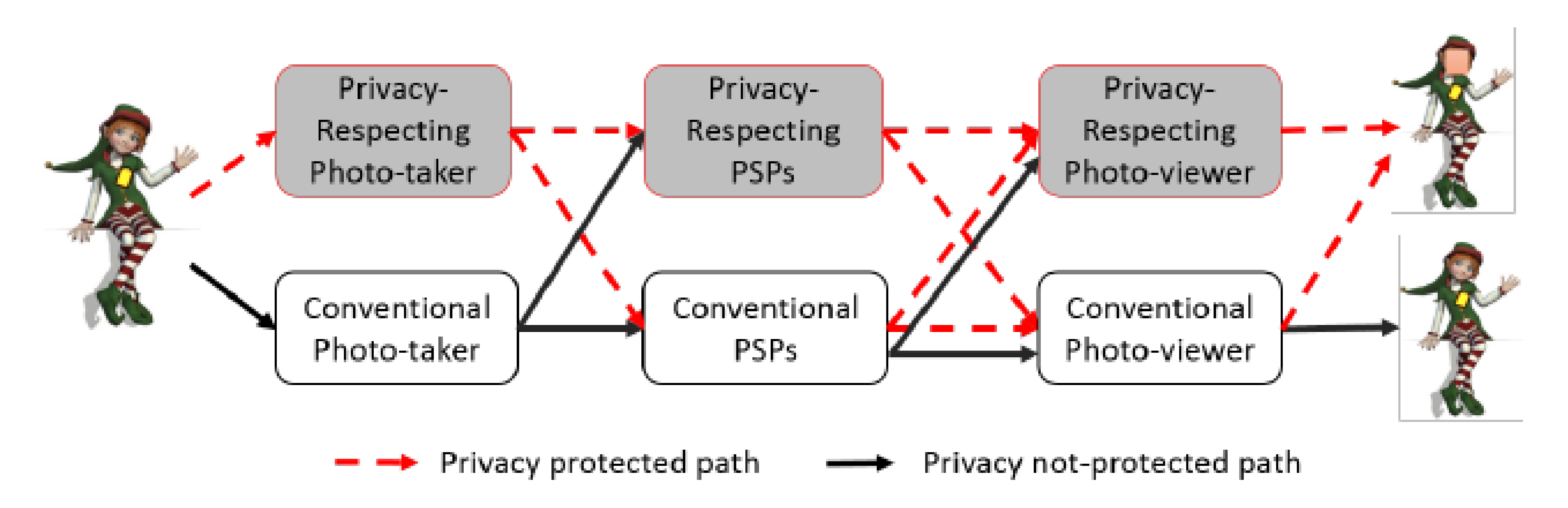
Captured

Strangers may be in my photo ↔ I may be in stranger's photo as well!

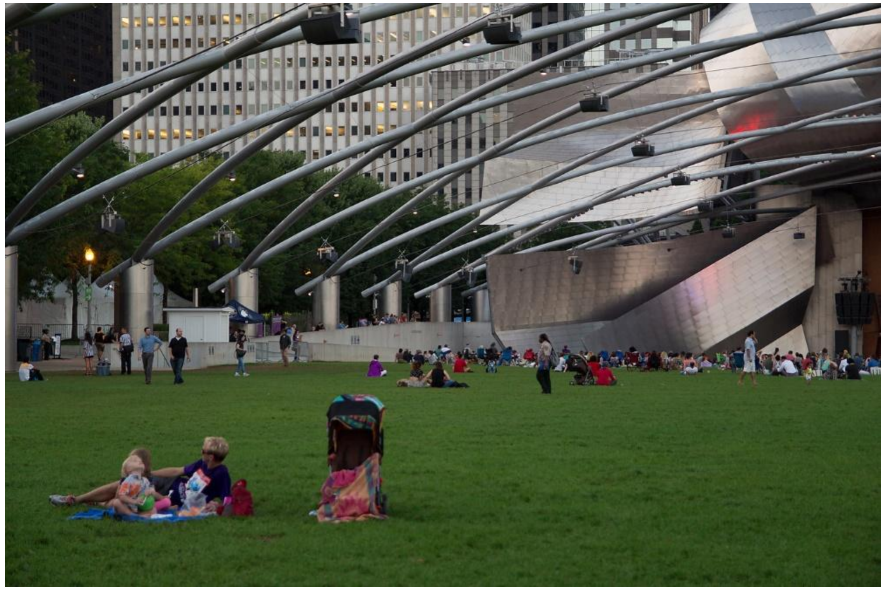
Too many cameras these days…

Current protection against cameras

Privacy Concern Expressed & Respected



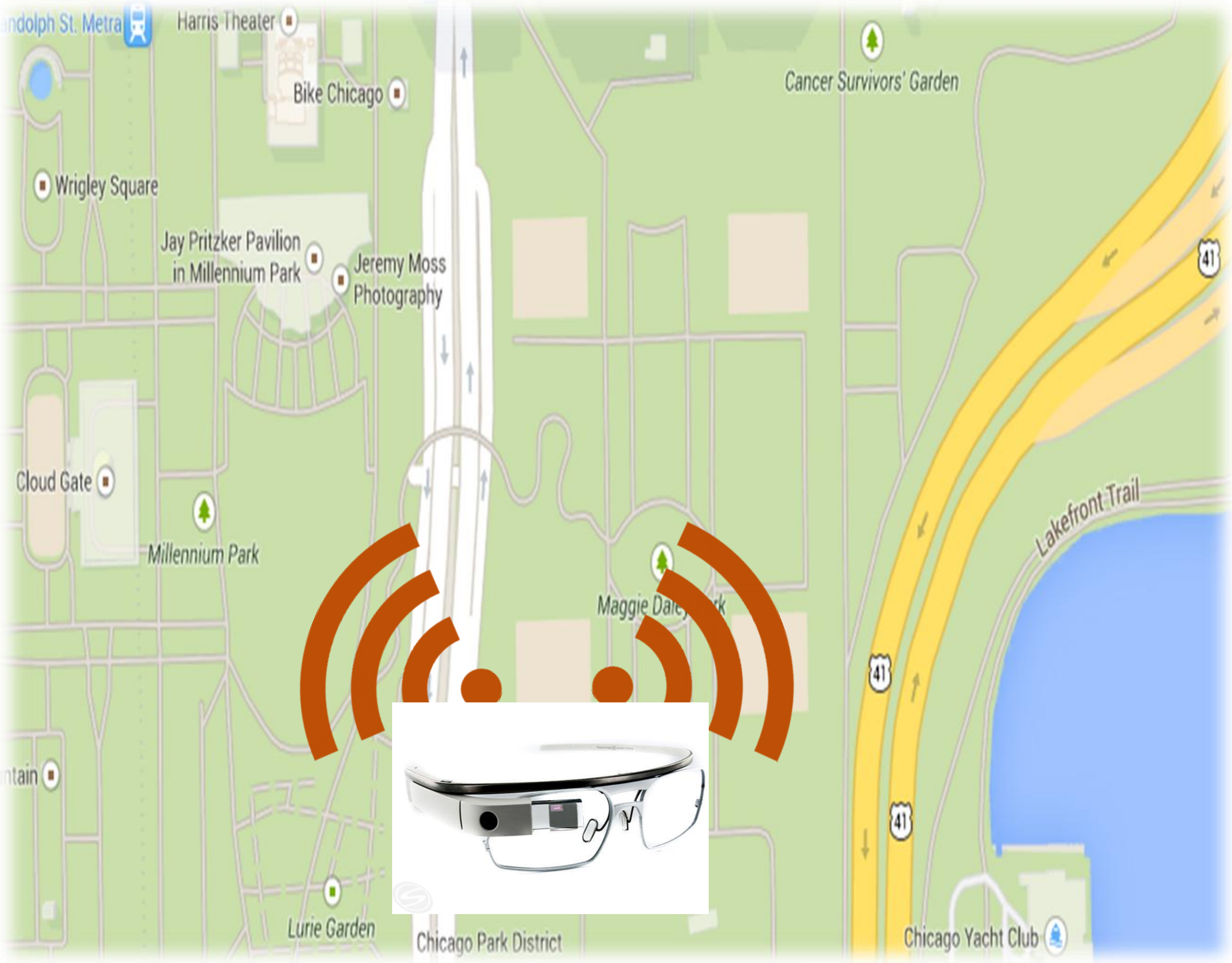
Our Interactive solution



1. Photo taken

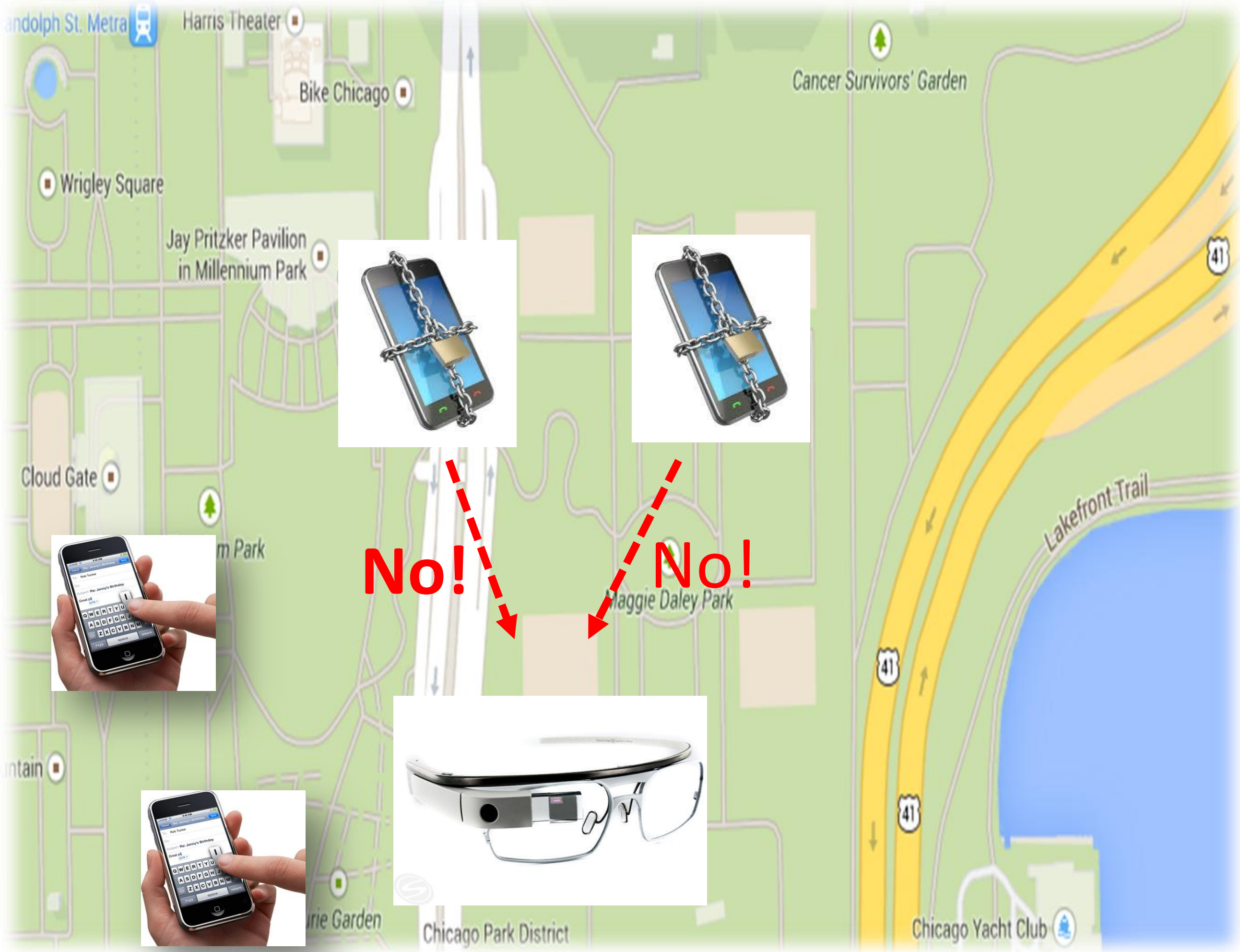
in Real World

Interactive solution



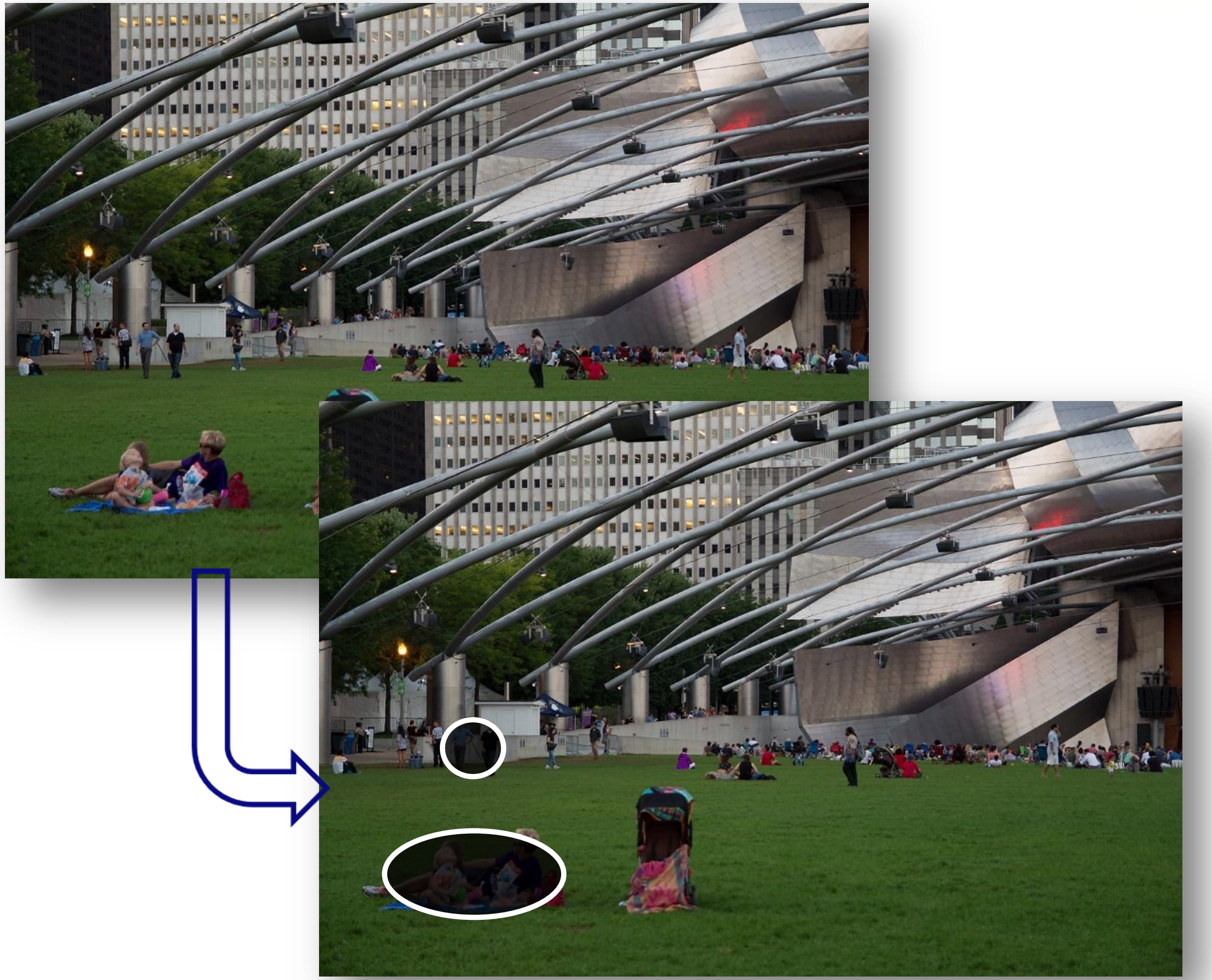
1. Photo taken
2. Broadcast

Interactive solution



1. Photo taken
2. Broadcast
3. Privacy Request
   - Sending his photo using face features

Interactive solution



Various sanitization in reality (eg: blur)

1. Photo taken
2. Broadcast
3. Privacy Request
4. Sanitize Image

Evaluation setting

- Networking workshop with ≥ 50 people in $200m^2$
- 10 volunteers, 4 female 6 male, acted as invisible users and photographer
  - Took photos freely in 1 day
  - 208 photos are taken
  - 1326 pedestrian detected (belong to 42 people)
  - 412 faces are detected

Diversity and consistency

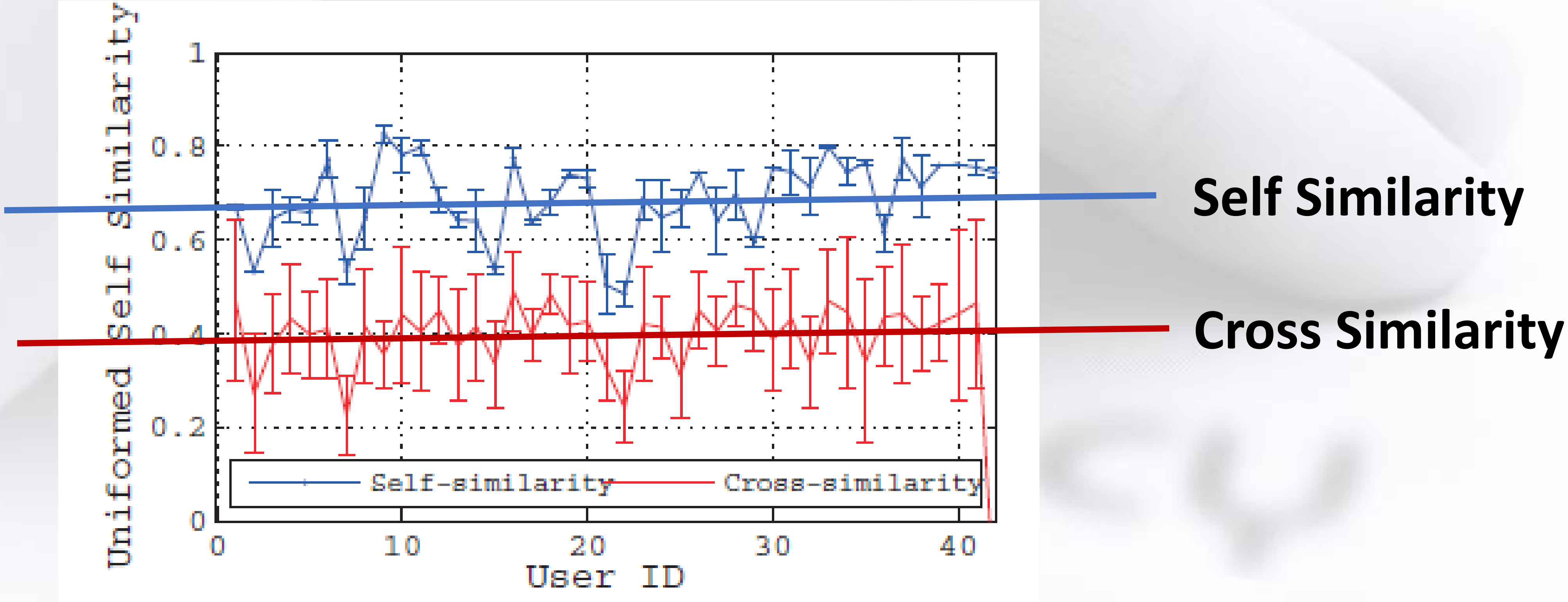


Figure 7: Portrait similarity variances.

Performance

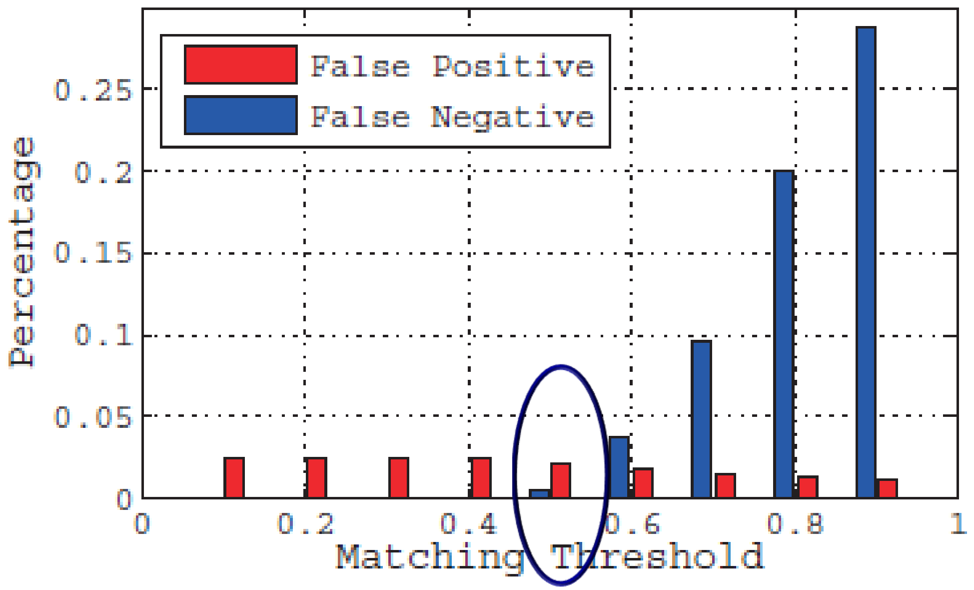


Figure 8: FP and FN in basic scheme

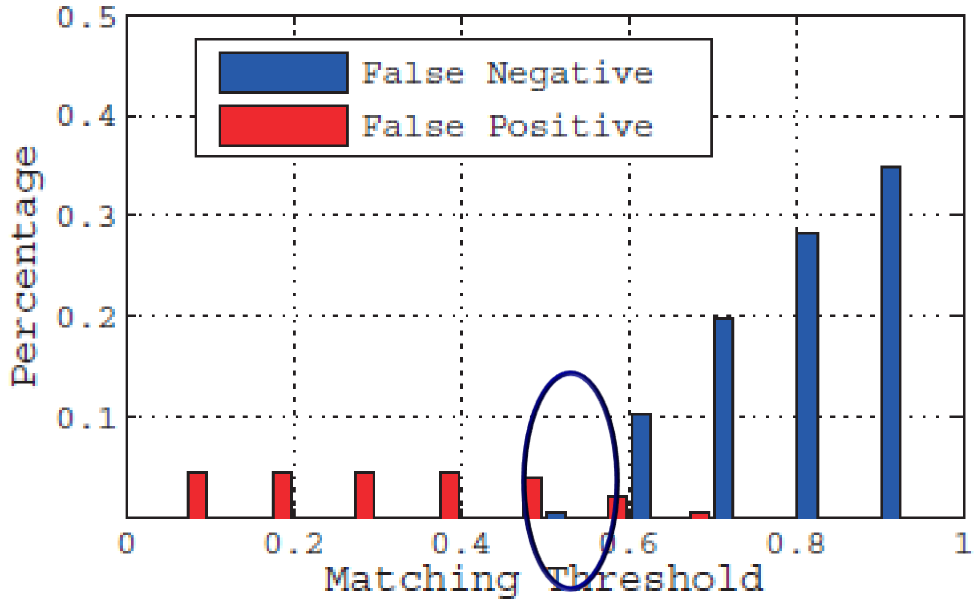


Figure 9: FP and FN in advanced scheme

Communication overhead is

less than 1KB for each neighbor

Less than 10KB for the photographer

Non-interactive solution

1. Photo taken

Non-interactive solution

1. Photo taken

2. Privacy Seeker?

Non-interactive solution

1. Photo taken
2. Privacy Seeker?
3. Enforce privacy

Conceal image: blur

Search on concealed images?

Typical image search

Face search

Image feature search

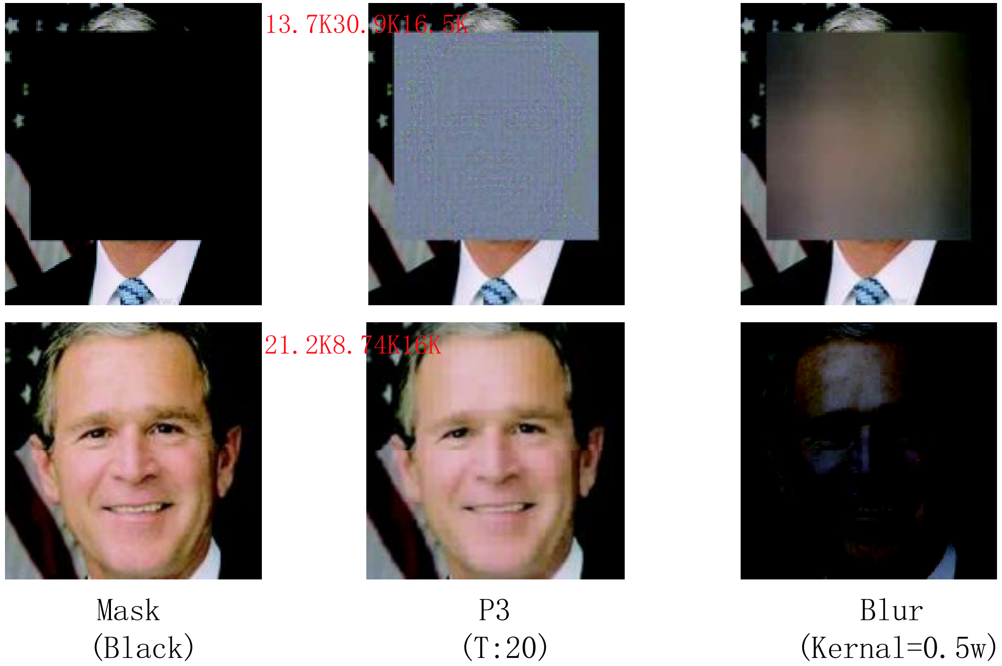
Metadata search

Concealing with image search enabled

1. Image separation

2. Search key encryption & access control

3. PP Vector search using HE



Publish public part, control secret part

Concealing with image search enabled

1. Image separation

2. Search key encryption & access control

3. PP Vector search using HE

ABE

Key is required to conduct ‘search’

Concealing with image search enabled

1. Image separation

2. Search key encryption & access control

3. PP Vector search using HE

$\widehat{x}_1$

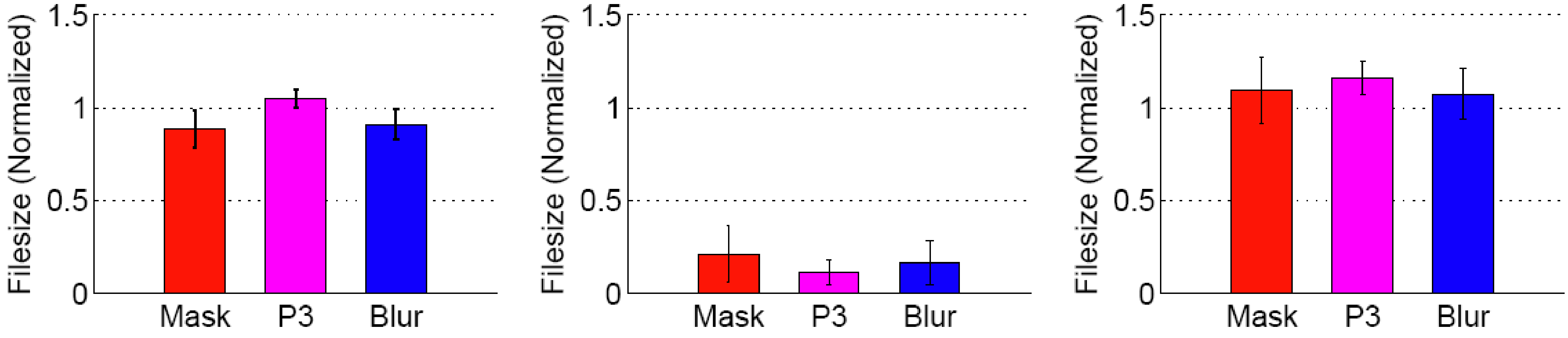
Searched

$(a_1, a_2, a_3, \cdots, a_n)$
$(b_1, b_2, b_3, \cdots, b_n)$
$(c_1, c_2, c_3, \cdots, c_n)$

$$dist(\overline{x}, \overline{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

Performance: Image File Size

- 3000 real-life photos.



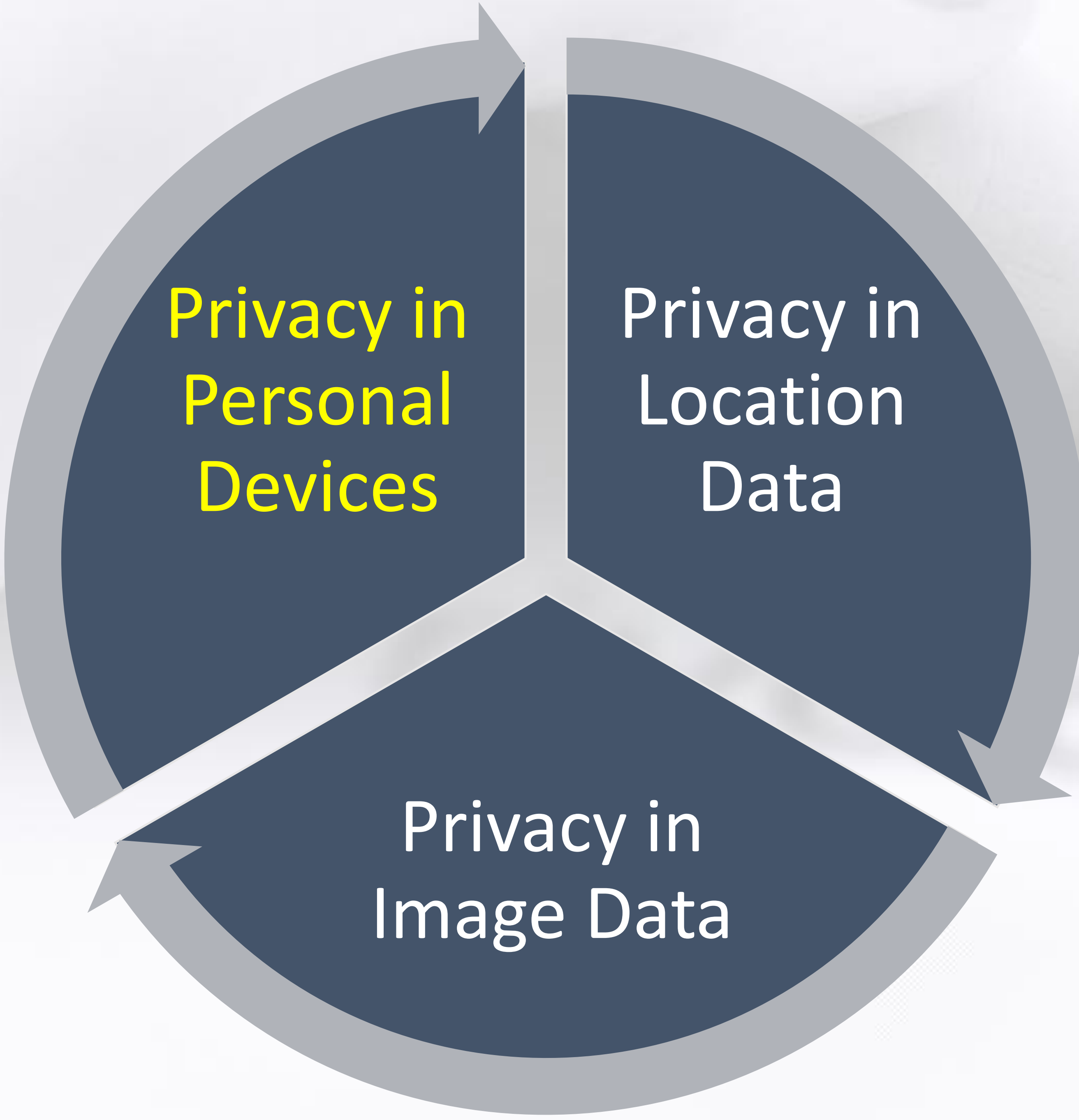
(a) Public (b) Secret (c) Public+Secret

Performance: Processing Time

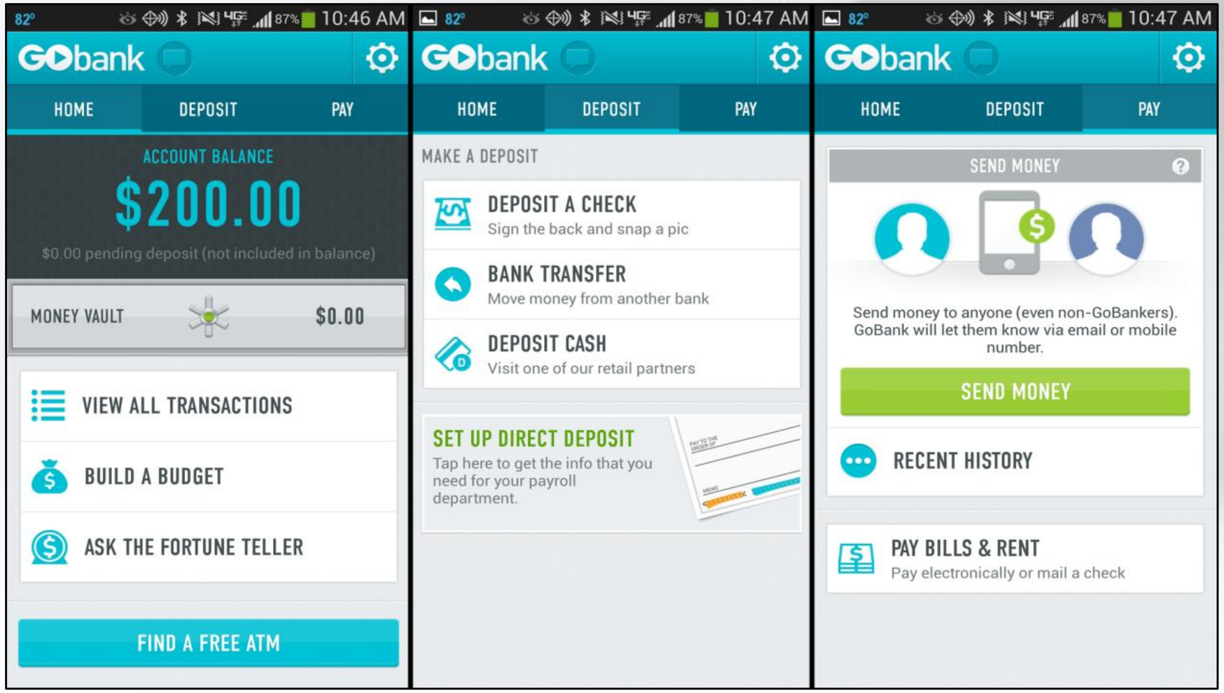
(b) Image search (average run time)

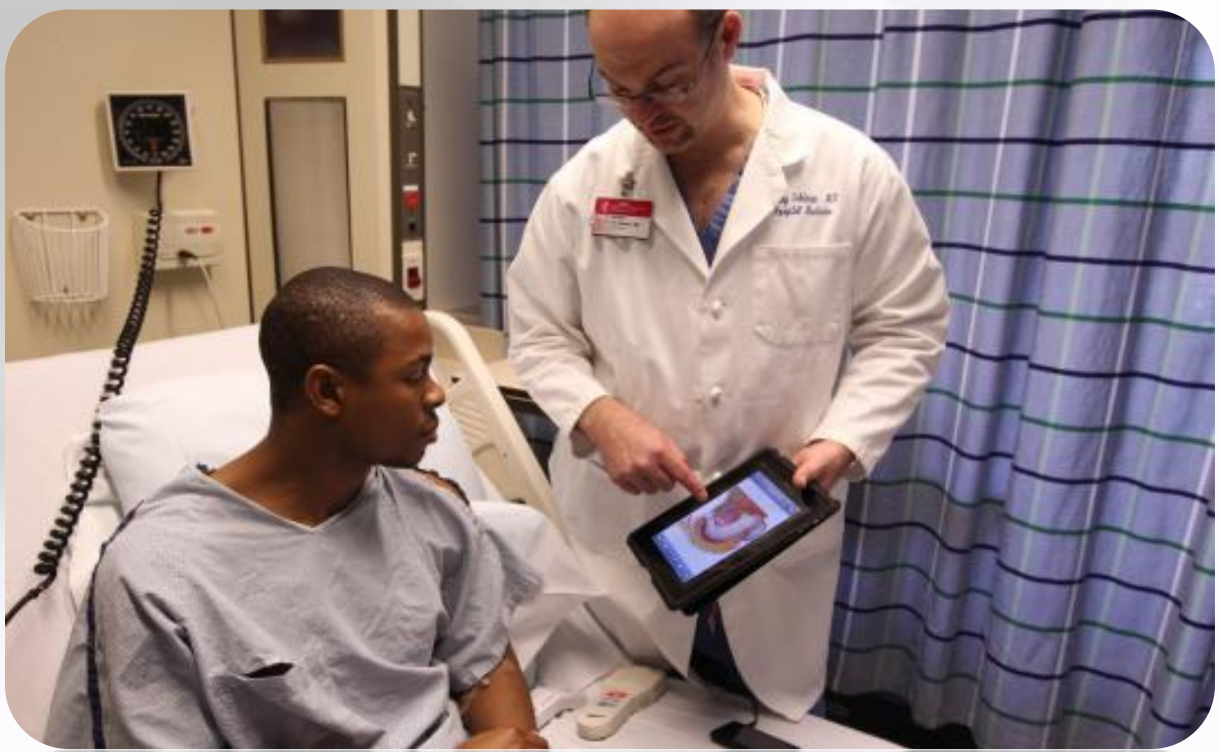
Laptop (sec)		
SouTu	64 dimension	128 dimension
Encrypt Vector (owner)	1.02	2.01
Encode Vector (querier)	0.55	1.12
Decrypt Distance	0.016	0.016
SouTu$_{bin}$	64 dimension	128 dimension
Encrypt Vector (owner)	0.51	1.03
Encode Vector (querier)	< 0.001	< 0.001
Decrypt Distance	0.016	0.016
Smartphone (sec)		
SouTu	64 dimension	128 dimension
Encrypt Vector (owner)	1.85	3.91
Encode Vector (querier)	0.64	1.37
Decrypt Distance	0.024	0.024
SouTu$_{bin}$	64 dimension	128 dimension
Encrypt Vector (owner)	0.56	1.33
Encode Vector (querier)	< 0.001	< 0.001
Decrypt Distance	0.024	0.024

- Client side
  - About 0.5s per image using labtop
- Cloud Side
  - About 0.2s per image using labtop


Privacy in Personal Devices
Privacy in Location Data
Privacy in Image Data

Continuous and Oblivious Authentication



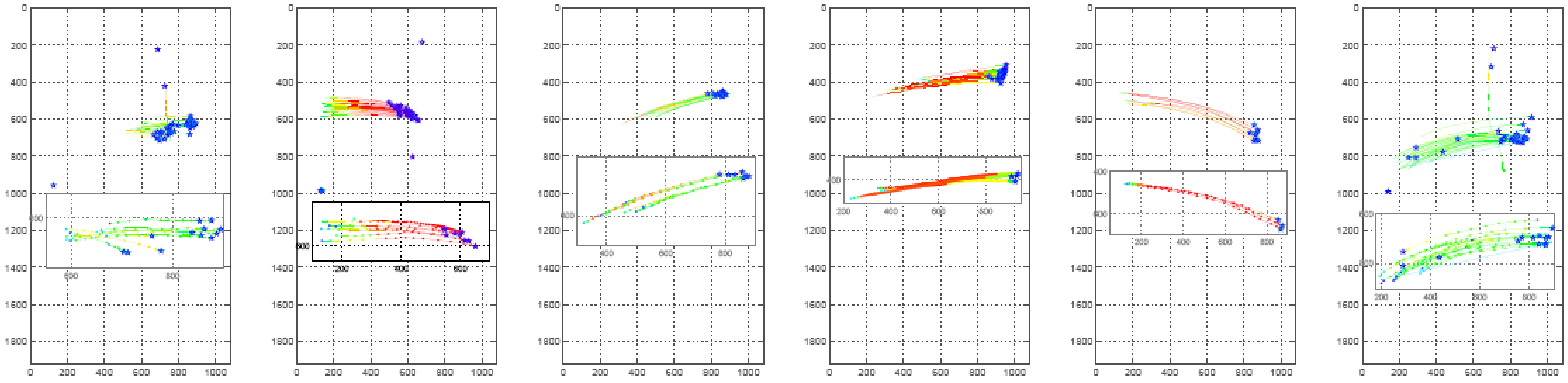


in Real World

Biometric feature as evidence

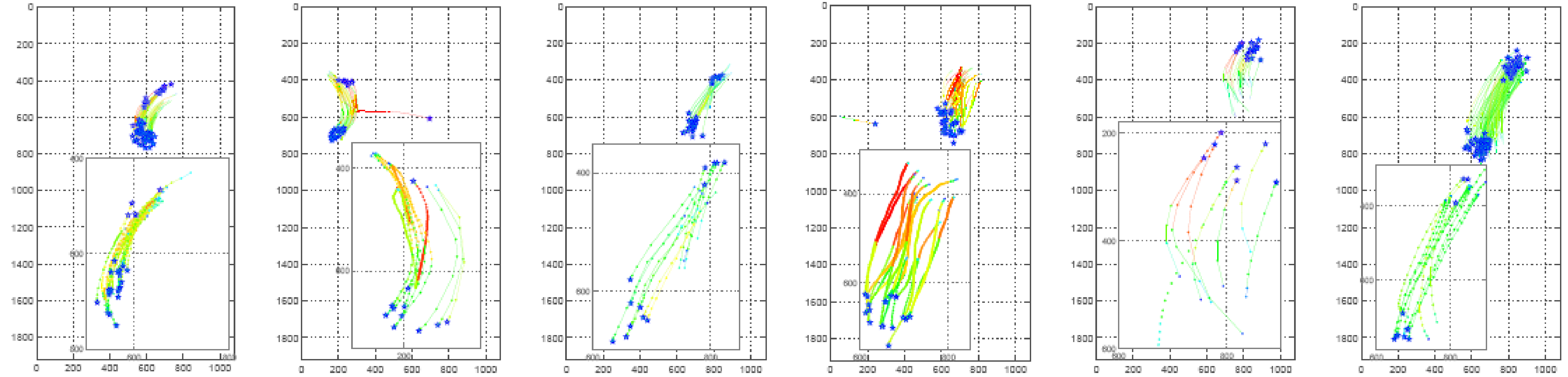
in Real World

Micro-behavior difference



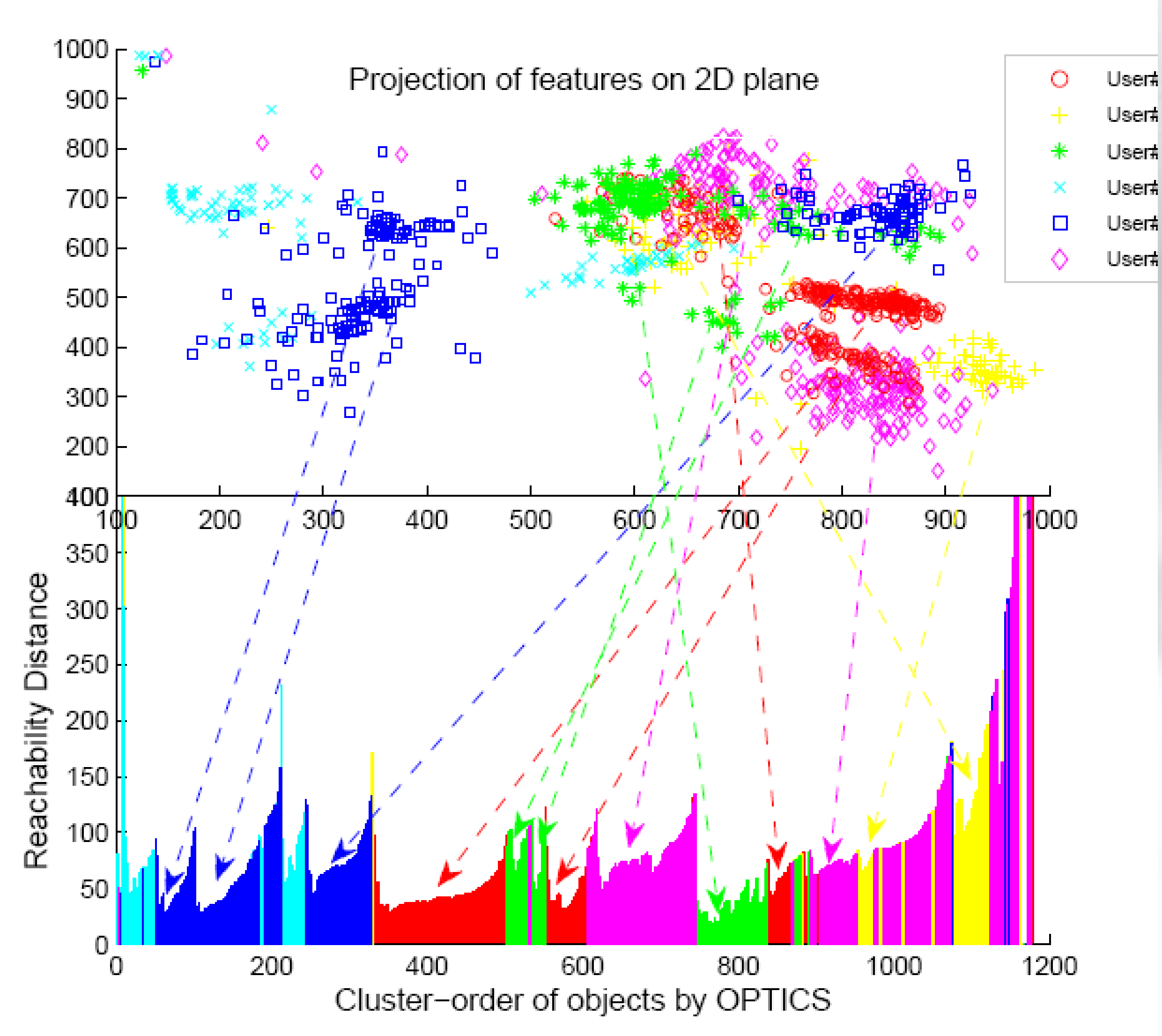
(a) BrowsePhotos (User#1) (b) BrowsePhotos (User#2) (c) BrowsePhotos (User#3) (d) BrowsePhotos (User#4) (e) BrowsePhotos (User#5) (f) BrowsePhotos (User#6)

Micro-behavior difference

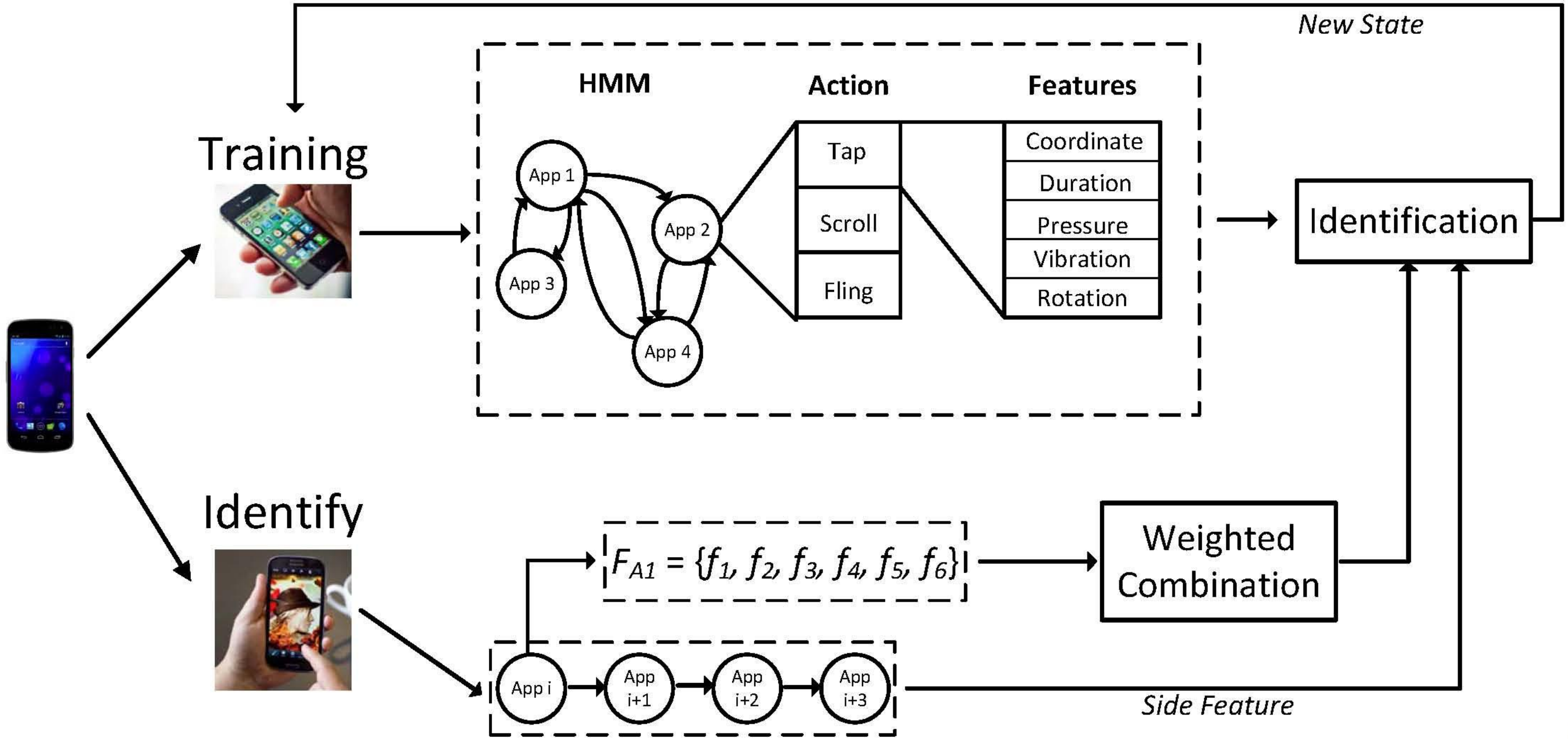


(g) BrowseTweets (User#1) (h) BrowseTweets (User#2) (i) BrowseTweets (User#3) (j) BrowseTweets (User#4) (k) BrowseTweets (User#5) (l) BrowseTweets (User#6)

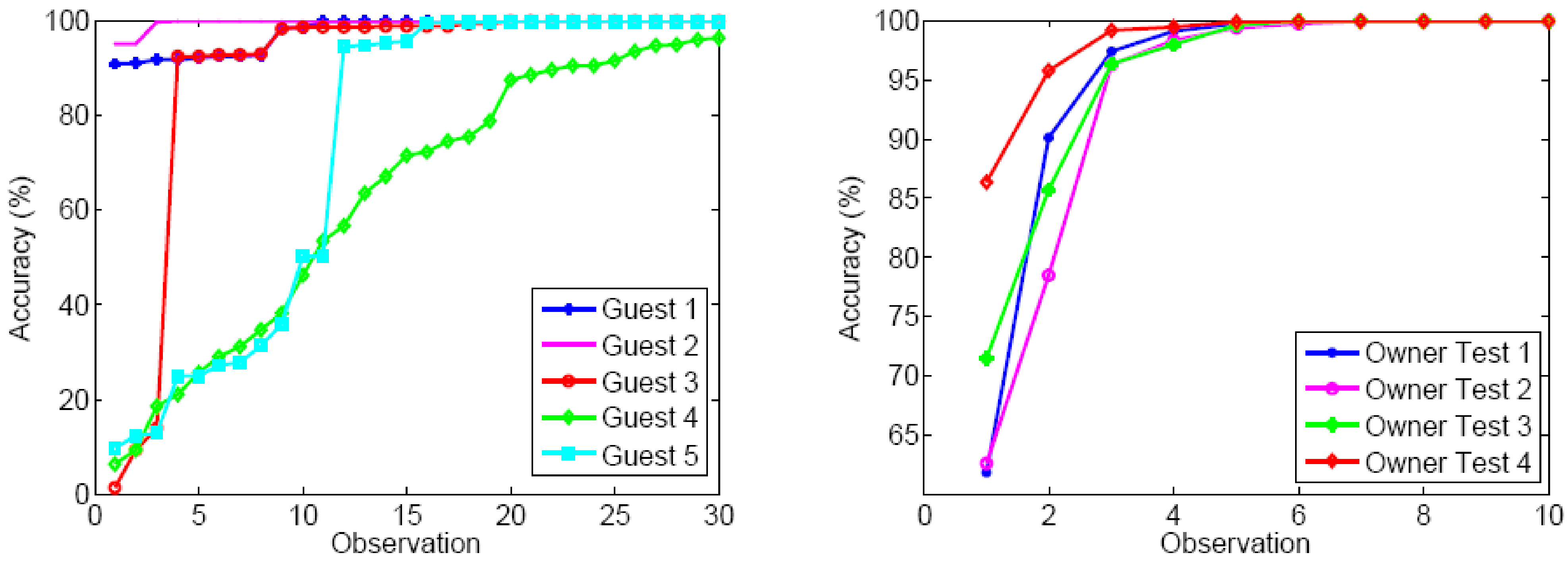
Diversity and consistency



Flow of our predictive model

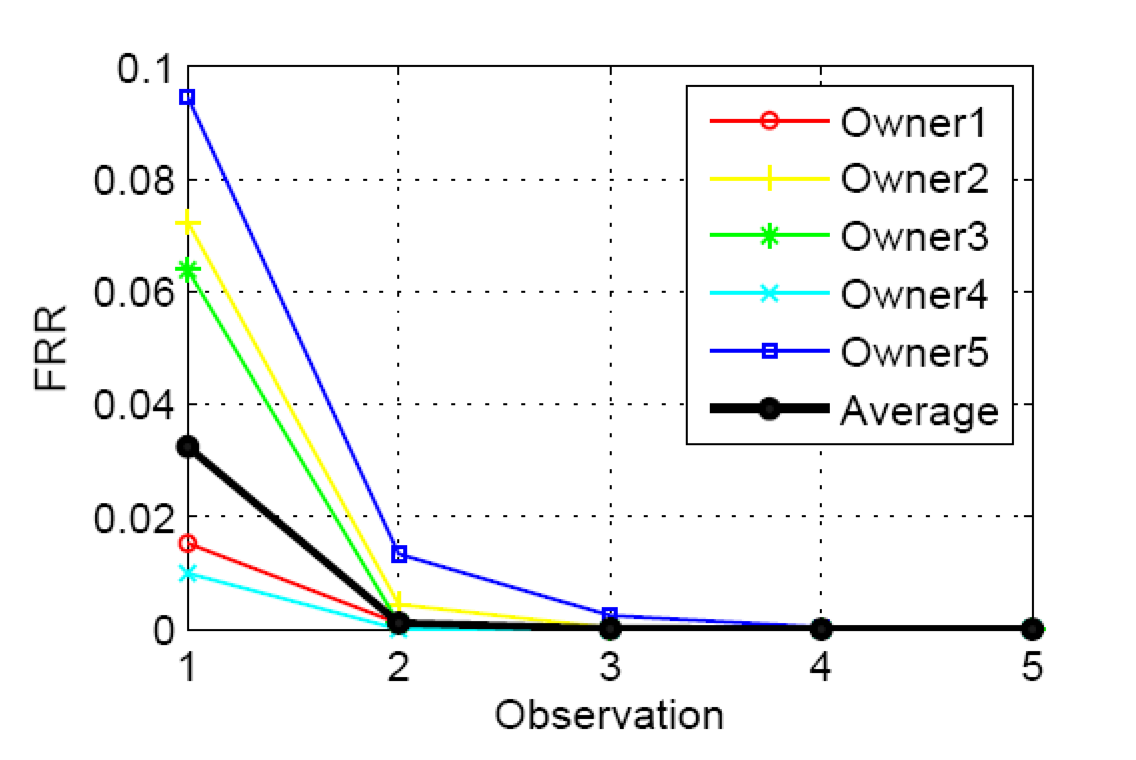


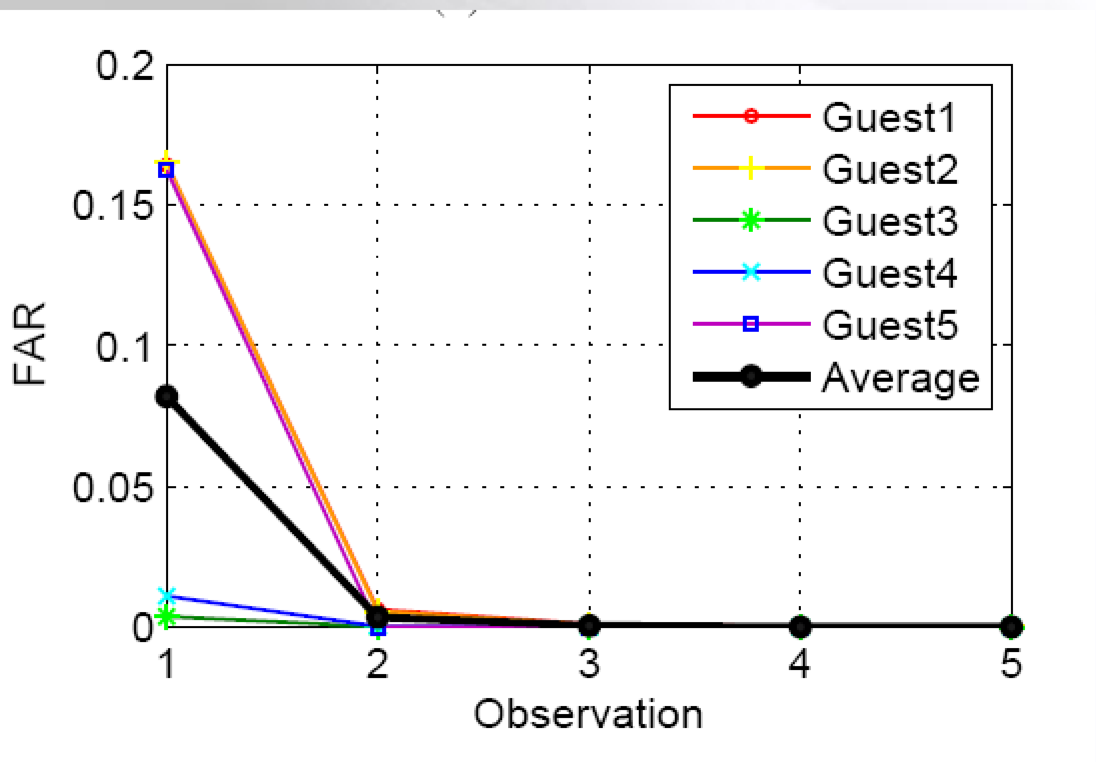
Accuracy performance



(a) Guest Accuracy (b) Owner Accuracy

Rejection and acceptance accuracy





Theoretic Frameworks for Data Sharing

Data mining everywhere

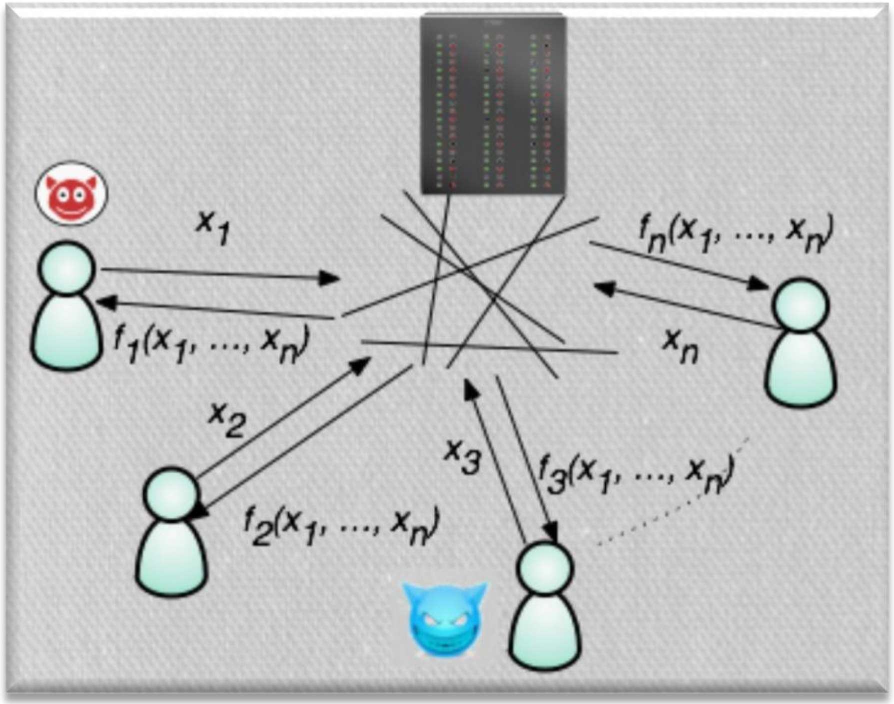
- Calculating average salary of a company?
- Finding the most frequent events, places?
- Analyze statistics on sensitive individual data?

Privacy-Preserving Data Aggregation without Secure Channel: Multivariate Polynomial Evaluation (IEEE INFOCOM, 2013)
Collusion-Tolerable Privacy-Preserving Sum and Product Calculation without Secure Channel (IEEE TDSC, 2014)

Modeling Privacy-Preserving Data Mining

Evaluate $f(\{x_1, \cdots, x_n\}) = \boxed{\sum_{k=1}^{m} c_k \left(\prod_{i=1}^{n} x_i^{d_{ik}} \right)}$ without disclosing x_i to each other

General polynomial



Privacy-Preserving Data Aggregation without Secure Channel: Multivariate Polynomial Evaluation (IEEE INFOCOM, 2013)
Collusion-Tolerable Privacy-Preserving Sum and Product Calculation without Secure Channel (IEEE TDSC, 2014)

Adversaries

Malicious semi-active adversary

- Deviate from the protocol specification
  - Without affecting final result.
  - Eg: passive rushing attacker

Existing approaches (practice & academic)

- Cryptographic approaches
  - SMC, secure secret sharing

slow

- Change the data precision & accuracy
  - Perturbation
  - Value distortion (e.g. differential privacy in database)
  - Add dummy data, dummy users

approximated

- Change the data owners
- Anonymization

data is open &
de-anonymization works

Privacy-Preserving Data Aggregation without Secure Channel: Multivariate Polynomial Evaluation (IEEE INFOCOM, 2013)
Collusion-Tolerable Privacy-Preserving Sum and Product Calculation without Secure Channel (IEEE TDSC, 2014)

Our contributions

✓Unsecured channel: Our communication channels are open to anyone, and we can still achieve privacy and security.

✓Theoretically provable privacy

✓Low computation overhead: Running time (computation only) is 10-1000 times less than SMC.

Privacy-Preserving Data Aggregation without Secure Channel: Multivariate Polynomial Evaluation (IEEE INFOCOM, 2013)
Collusion-Tolerable Privacy-Preserving Sum and Product Calculation without Secure Channel (IEEE TDSC, 2014)

Simple observation

Inspired by the observation :

- Polynomial = **Multiplications (*) & Additions (+)**

Design two novel protocols

- Multi-party Product & Sum calculation protocols

Product Protocol

Integers, modulo P

$$\prod x_i \Rightarrow \prod (x_i R_i) = \prod x_i \cdot \prod R_i$$

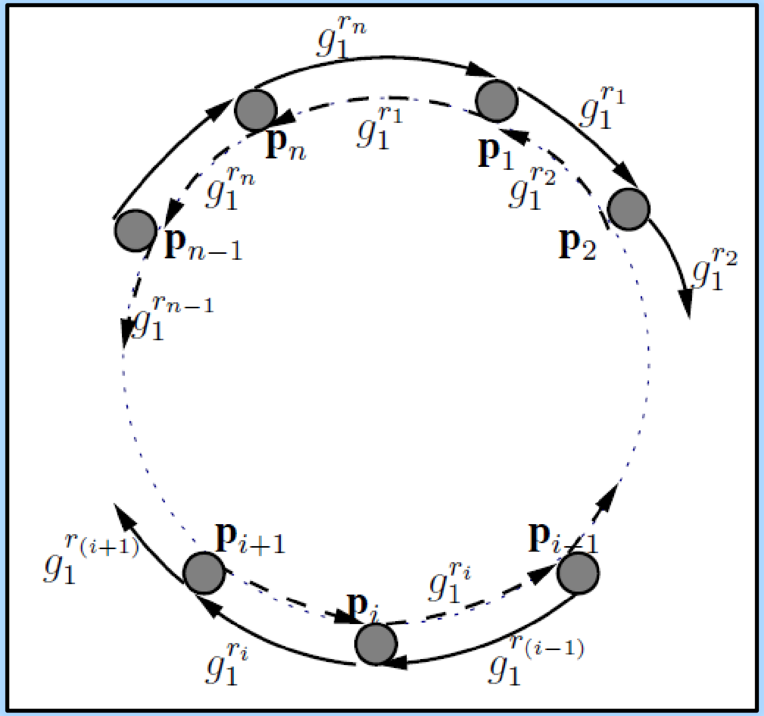
Random mask

- Every participant i computes

$$R_i = (g^{r_{i+1}} / g^{r_{i-1}})^{r_i}$$
$$= (g^{r_i r_{i+1}} / g^{r_{i-1} r_i})$$

Random, selected by i

- $\prod x_i R_i = \prod x_i$



Sum Protocol

$$(1+p)^x = \sum_{i=0}^{x} \binom{x}{i} p^i = 1 + xp \pmod{p^2}$$

$$\Rightarrow \frac{\prod (1+p)^{x_i} - 1}{p} \pmod{p^2}$$

$$= \frac{(1+(\sum x_i)p) - 1}{p} \pmod{p^2}$$

$$= \sum x_i \pmod{p}$$

Use product protocol

Put All Together

- Combine product and sum protocls to achieve general multivariant polynomial operation:

$$f(\{x_1, \cdots, x_n\}) = \sum_{k=1}^{m} c_k \left(\prod_{i=1}^{n} x_i^{d_{ik}}\right)$$

- Provable privacy preservation
  - Entropy, hardness

Privacy-Preserving Data Aggregation without Secure Channel: Multivariate Polynomial Evaluation (IEEE INFOCOM, 2013)
Collusion-Tolerable Privacy-Preserving Sum and Product Calculation without Secure Channel (IEEE TDSC, 2014)

Run time comparison

FairplayMP by Ben et al. (SMC implementation)

Gates	64	128	256	512	1024
Run time (ms)	130	234	440	770	1394

26 additions in our schemes are equivalent to a 1066-gate circuit.

Our run time : 72.2 microseconds.

Privacy-Preserving Data Aggregation without Secure Channel: Multivariate Polynomial Evaluation (IEEE INFOCOM, 2013)
Collusion-Tolerable Privacy-Preserving Sum and Product Calculation without Secure Channel (IEEE TDSC, 2014)

In arbitrary user groups

- In previous approaches, we are given a fixed user group.
  - What happens if user group changes?
  - Shall we distribute keys for EVERY different group?
    - NO, too much (2^N groups for N users).

- A protocol that only needs $O(N)$ key space for each user
  - That can be used to evaluate any polynomial among any subgroup of N users.

Inspired by secret sharing!

In Shamir's secret sharing for polynomial $y = f(x)$ of degree $k - 1$,

- k data points y_i are needed to re-construct it

$$f(x) = \sum_i y_i \boxed{l_i(x)}$$

Lagrange coefficients

$$l_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}$$

- Sharing arbitrary k values $\Rightarrow$ fixed value $f(x_0)$ can be derived

Goal is : $\prod_i R_i = C$

- Core idea
  - We want R_i's such that $\prod R_i x_i = C \prod x_i$ for some constant C
  - We can distribute $y_i, l_i(x_i)$
    $\Rightarrow$ Any k set of $< y_i, l_i(x_i) >$'s will lead to the same $f(x_0)$
  - Proper initialization among n people
    $\Rightarrow$ Any subgroup of k people can privately share their data.
  - Security parameters must be carefully chosen to guarantee semantic security.

Detailed Protocol Description

- Key distribution
  - Let user i possess $EK_i = \boxed{\left(q^{(2)}(i), q^{(3)}(i), \cdots, q^{(n-1)}(i)\right)}$ having $n-2$ parameters.

Secret parameters

- Data publication
  - When a polynomial evaluation is needed among m users
    - User i publishes $\boxed{C(x_i) = x_i H(t)^{q^{(m-1)}(i) l_i(0)}}$

x_i masked by secret parameter

- Data aggregation

$$\prod_i C(x_i) = \left(\prod_i x_i\right) \cdot H(t)^{\sum_i q^{(m-1)}(i) l_i(0)}$$

$$= \left(\prod_i x_i\right) \cdot H(t)^{\boxed{q^{(m-1)}(0)}} \qquad \text{(Polynomial interpolation)}$$

  - If $\boxed{q^{(m-1)}(0)}$ is set as 0, the entire product is equal to $\prod x_i$

Illegally altered inputs?

Will **data** from a user be trustable?

Will **results** from other users be trustable?

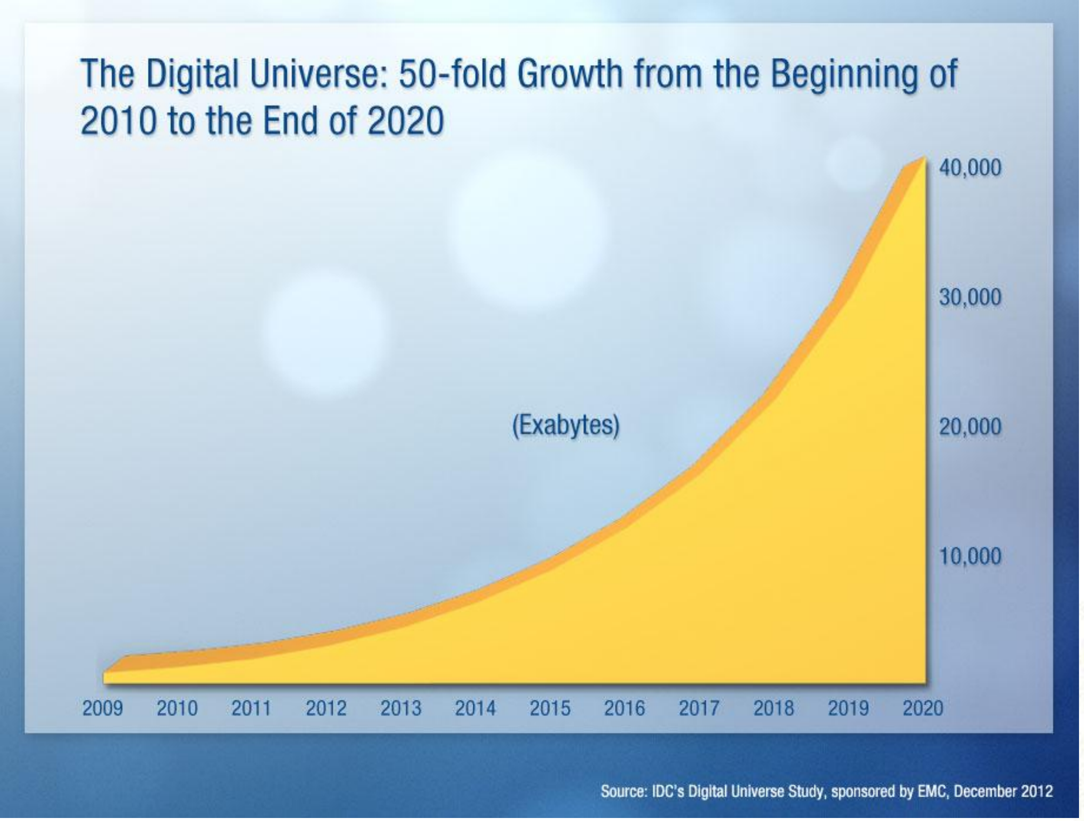
Data Providers

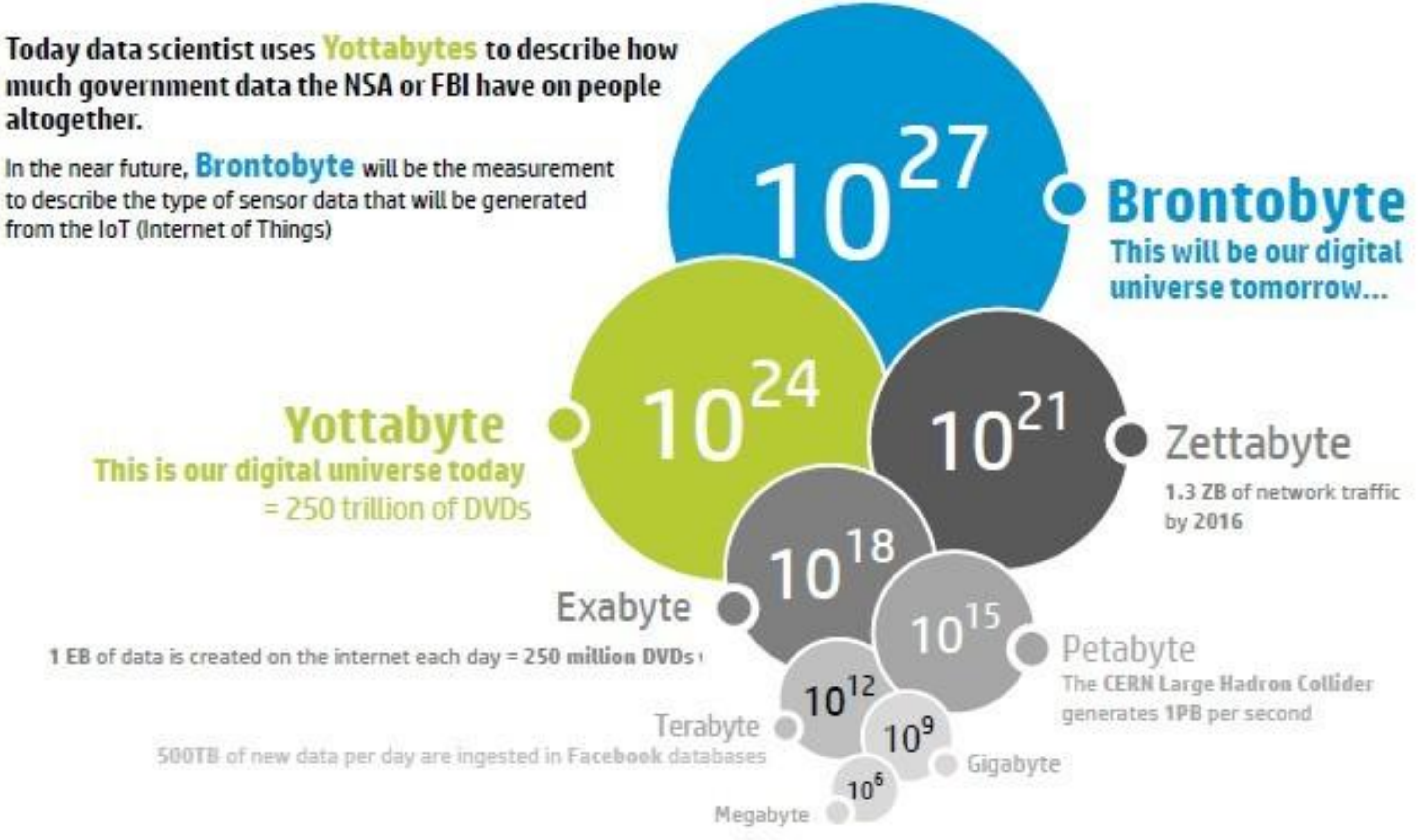
End- Users

Will **computations** from “cloud” be trustable?

Cloud: Storage/Computing

Challenges:
Secure yet Efficient Computation for Big Data Era


The Digital Universe: 50-fold Growth from the Beginning of 2010 to the End of 2020
40,000
30,000
20,000
10,000
(Exabytes)
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Source: IDC's Digital Universe Study, sponsored by EMC, December 2012


Today data scientist uses Yottabytes to describe how much government data the NSA or FBI have on people altogether.
In the near future, Brontobyte will be the measurement to describe the type of sensor data that will be generated from the IoT (Internet of Things)
10^27
Brontobyte
This will be our digital universe tomorrow...
10^24
Yottabyte
This is our digital universe today
= 250 trillion of DVDs
10^21
Zettabyte
1.3 ZB of network traffic by 2016
10^18
Exabyte
1 EB of data is created on the internet each day = 250 million DVDs
10^15
Petabyte
The CERN Large Hadron Collider generates 1PB per second
10^12
Terabyte
500TB of new data per day are ingested in Facebook databases
10^9
Gigabyte
10^6
Megabyte

Efficiency, Efficiency

- Computing is not powerful enough
- Effi
  as s
  - Security often requires efficiency drop

Efficiency is the key!

Even our own methods

- Even our super-efficient microsecond-level operation may be unacceptable.
- Needless to talk about other millisecond-level or even second-level operations.

New Security/Privacy Metrics?

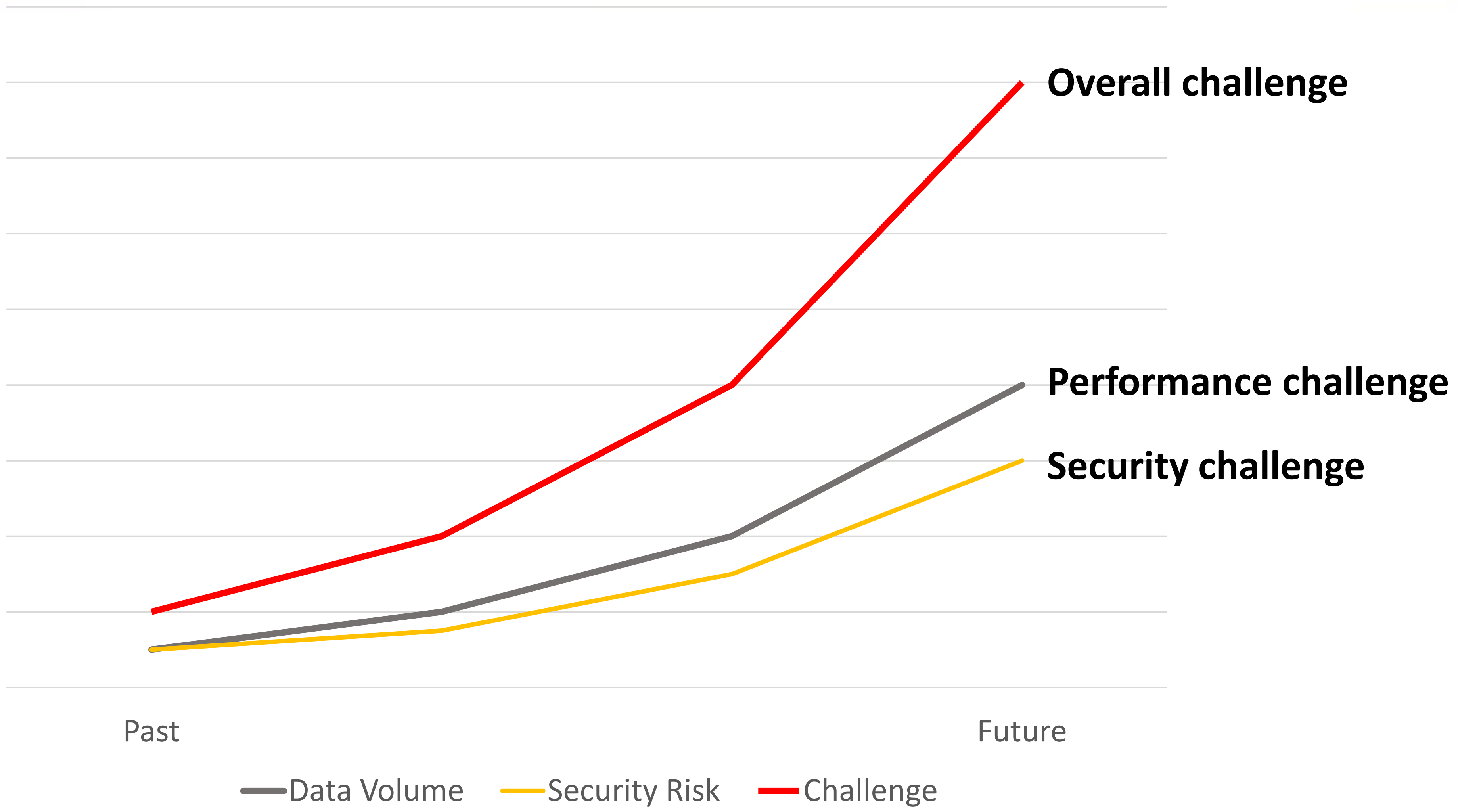
Current security guarantees

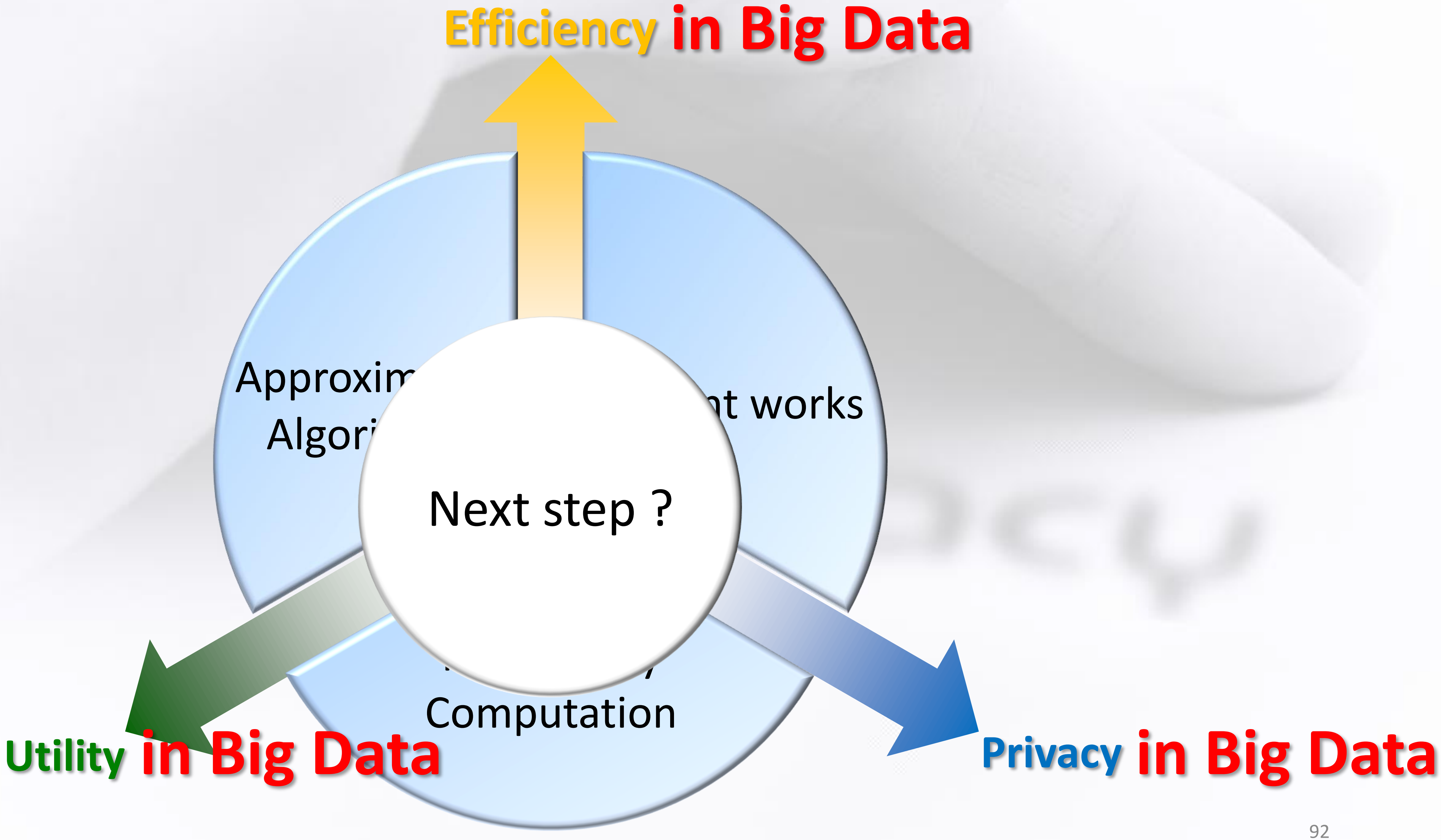
- Randomness or indistinguishability may become obsolete

Current security guarantees are not enough!

  - ➢ Variety & volume of big data are attacker-friendly

Still a very long way to go…




Efficiency in Big Data
Approxim
Algori
nt works
Next step ?
Computation
Utility in Big Data
Privacy in Big Data

Thanks for you attention!

Thank you !

Xiang-Yang Li (李向阳)
Professor, IIT, USA
www.cs.iit.edu/~xli
www.cs.iit.edu/~winet
xli@cs.iit.edu