

# Ensuring Privacy and Integrity against Untrusted Cloud for Internet of Things Applications

Haibo Hu 胡海波

Hong Kong Polytechnic University 香港理工大学

Oct 14 @ 隐私保护论坛 NDBC 2018

# Internet of Things Systems Are Vulnerable

- **Consumer IoT**

- BlackHat 2017: Remotely hack a Tesla Model S and gain full control of the vehicle.

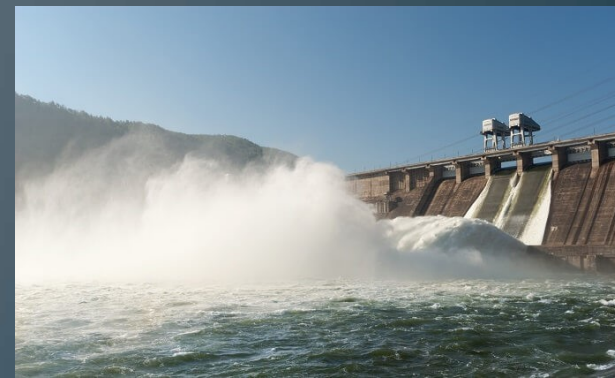


- present attack to turn Amazon Echo into a listening bug.



- **Industrial IoT**

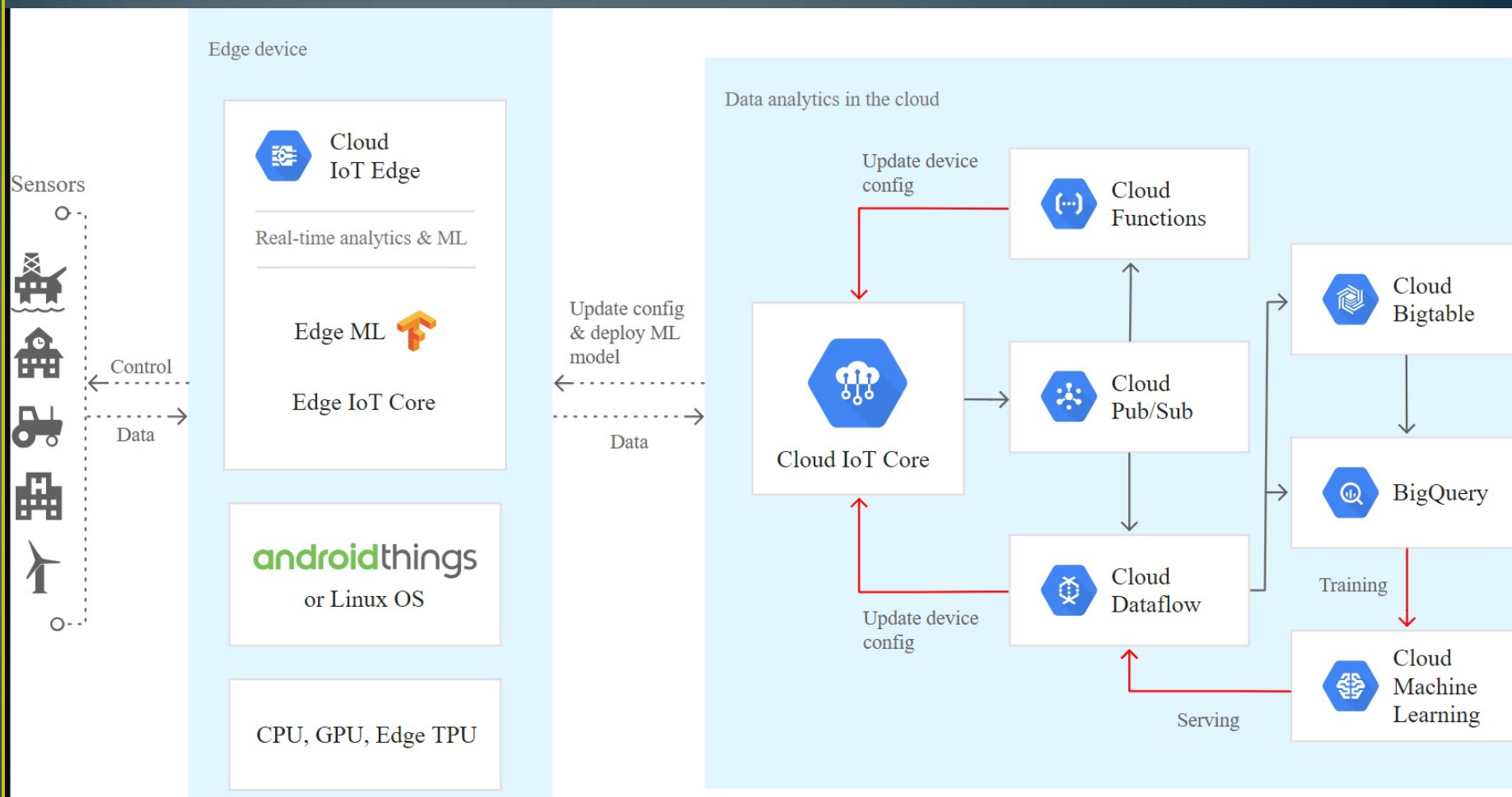
- BlackHat 2018: IBM X-Force Red researchers find vulnerabilities of various smart city systems from Libelium, Echelon and Battell.
- Exploit an IoT gateway connected to a dam, resulting in a flooded road.



# Cloud Databases Partially Solve This and More

- High Availability
    - Fault-tolerant
    - Instance failover
  - High Scalability
    - Elastic replica creation
  - High Performance
  - Inherited Security
  - Low Latency Network
    - Fiber network
  - Big Data and AI
- Main vendors
    - Microsoft Azure IoT Suite
    - Google Cloud IoT
    - GE Predix
    - AWS IOT
    - IBM Watson IoT
    - Salesforce IoT Cloud
    - More vendors are here:  
<https://www.postscapes.com/internet-of-things-platforms/>

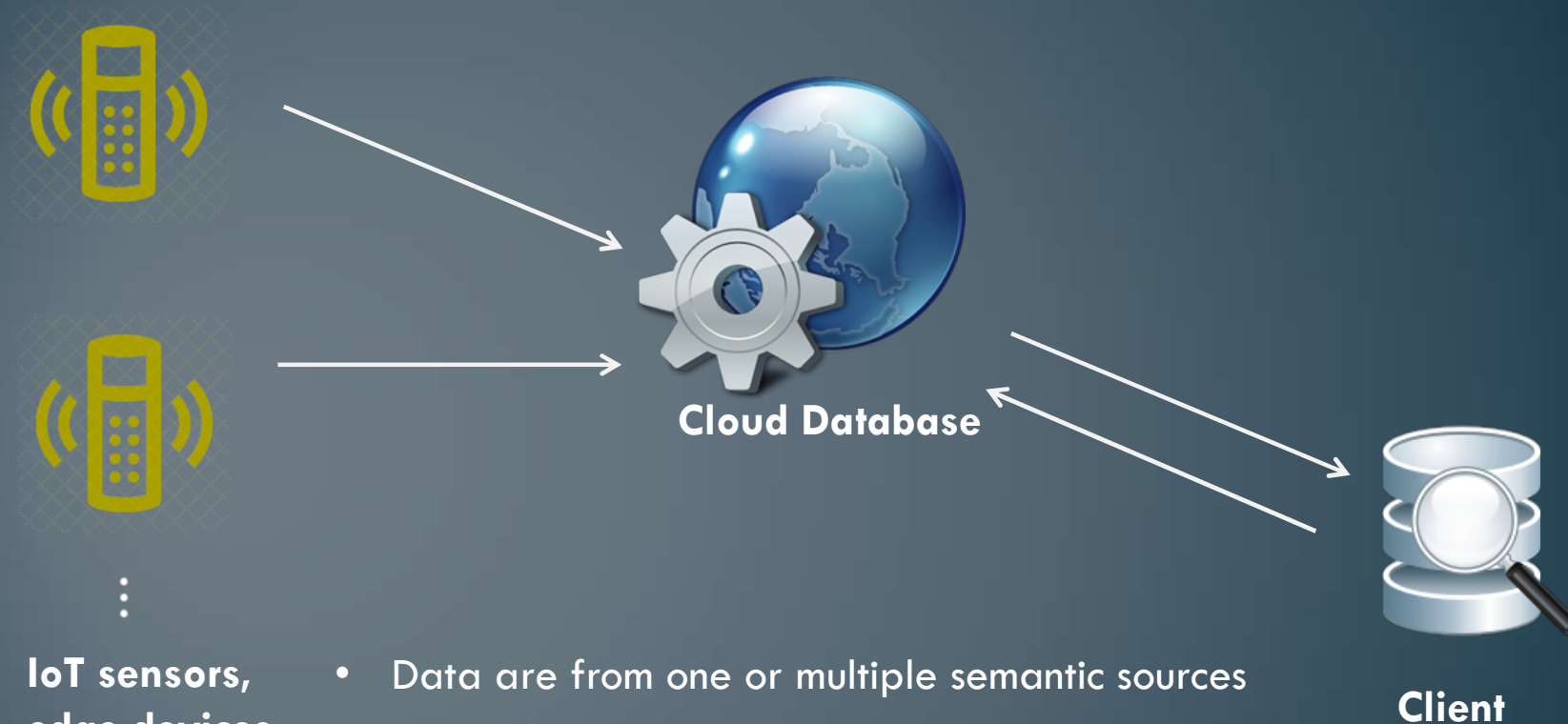
# Reference Architecture in Google Cloud IoT



# Roadmap

- ✓ IoT and Outsourced Database
- System, Security and Privacy Model
- Case Studies
  - 1: Single-Source Privacy-Preserving Integrity Assurance
  - 2: Integrity Assurance with Access Control and Zero-knowledge Proof
  - 3: Local Differential Privacy for Key-Value Pairs
- Conclusion

# System, Security and Privacy Model



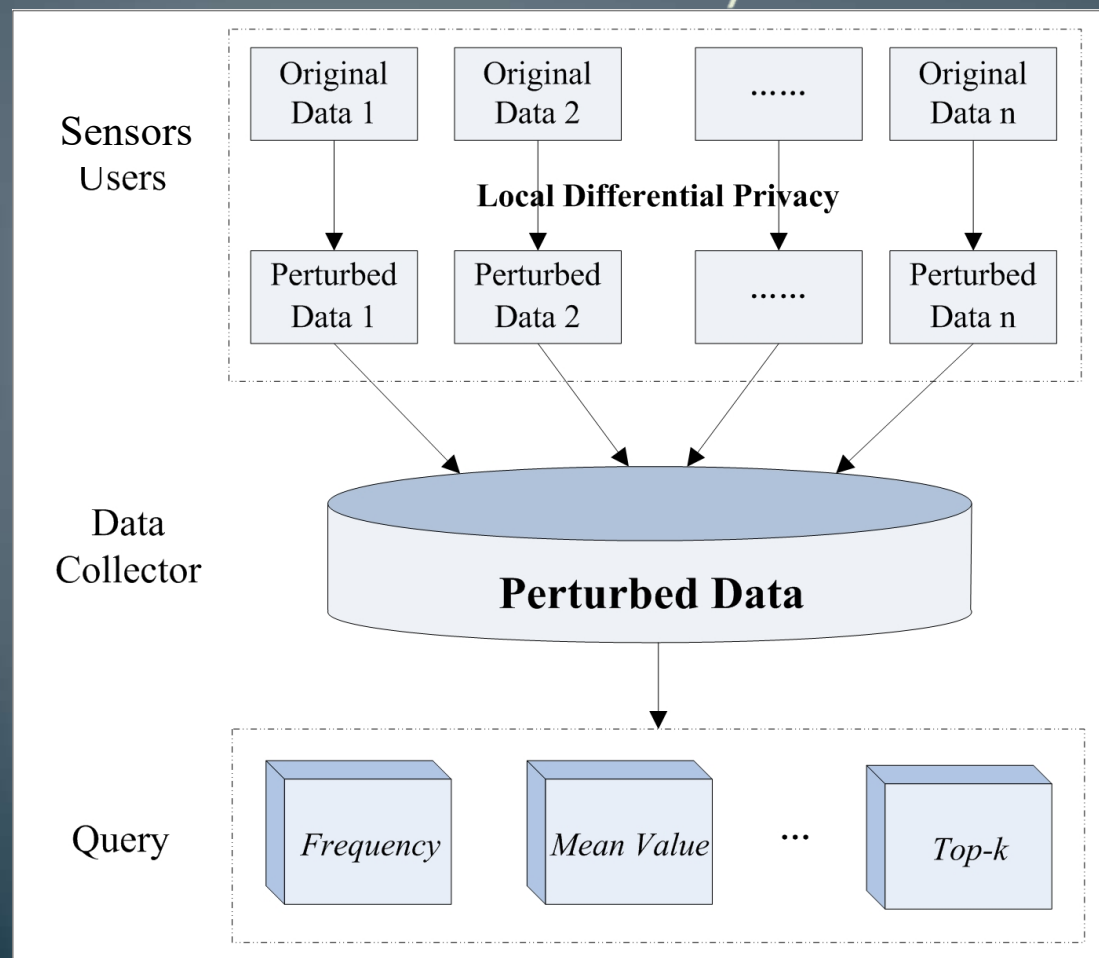
- Data are from one or multiple semantic sources
- Cloud database provides a unified query interface
- Integrity: the query results can go wrong
- Privacy: sensor values can be private

# Security Model

- Why the cloud database cannot be trusted?
  - Compromised by hackers
  - Incomplete database search
  - Program/human error
  - Business dishonesty (e.g., in favor of sponsors)
- How shall we ensure the integrity of query results?
  - Soundness: no false positive
  - Completeness: no false negative
  - Freshness: no obsolete result (the most challenging!)

# Privacy Model

- Indistinguishability (semantic security)
- $\epsilon$ -Local Differential Privacy



$$\frac{\Pr[M(\mathbf{t}) = \mathbf{t}^*]}{\Pr[M(\mathbf{t}') = \mathbf{t}^*]} \leq \exp(\epsilon)$$

for any two **tuples**:  $\mathbf{t}$ ,  $\mathbf{t}'$

At home?		At home?
1	→	1
0	→	1
1	→	0
0	→	0
1	→	1

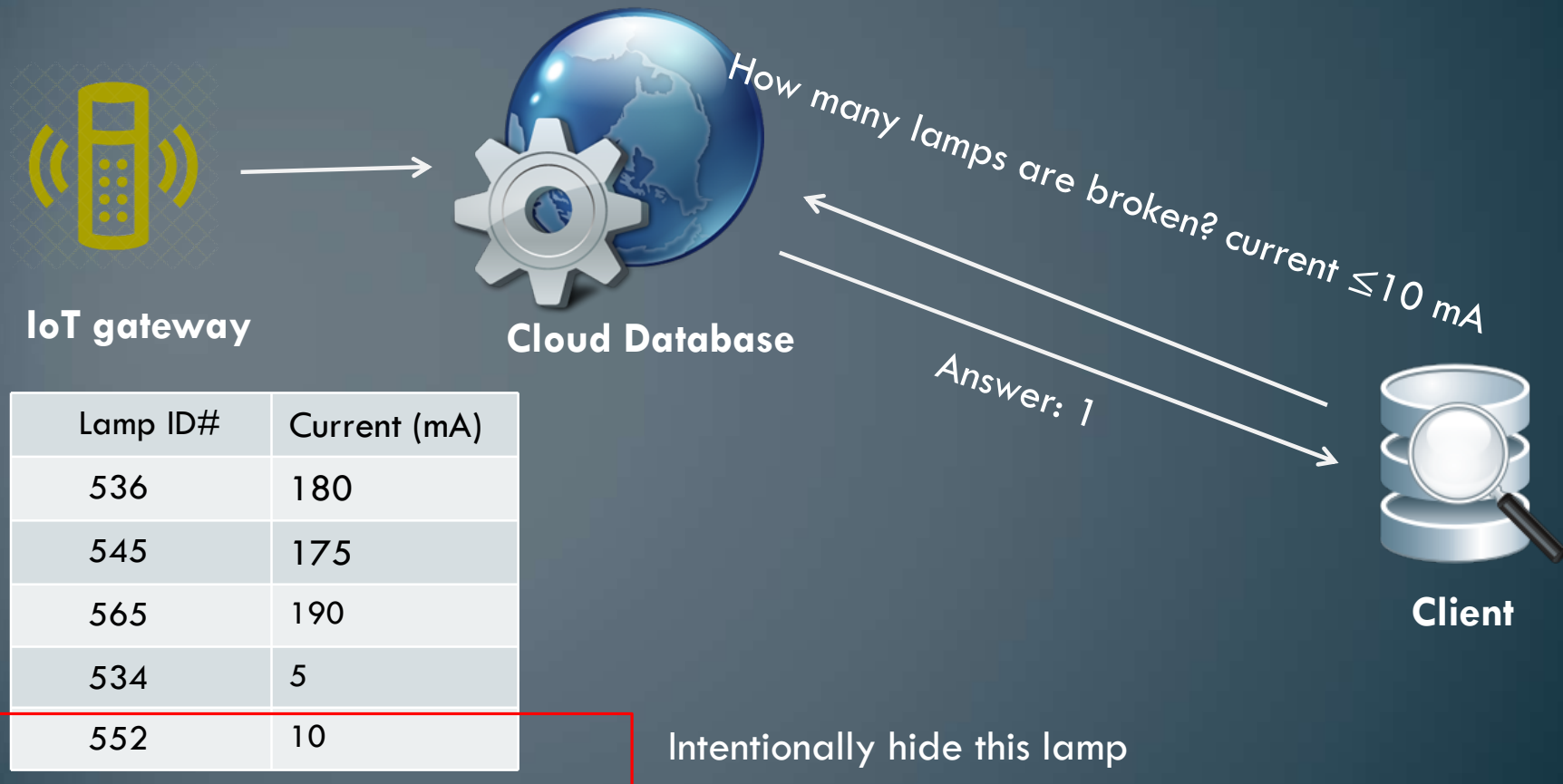


# Roadmap

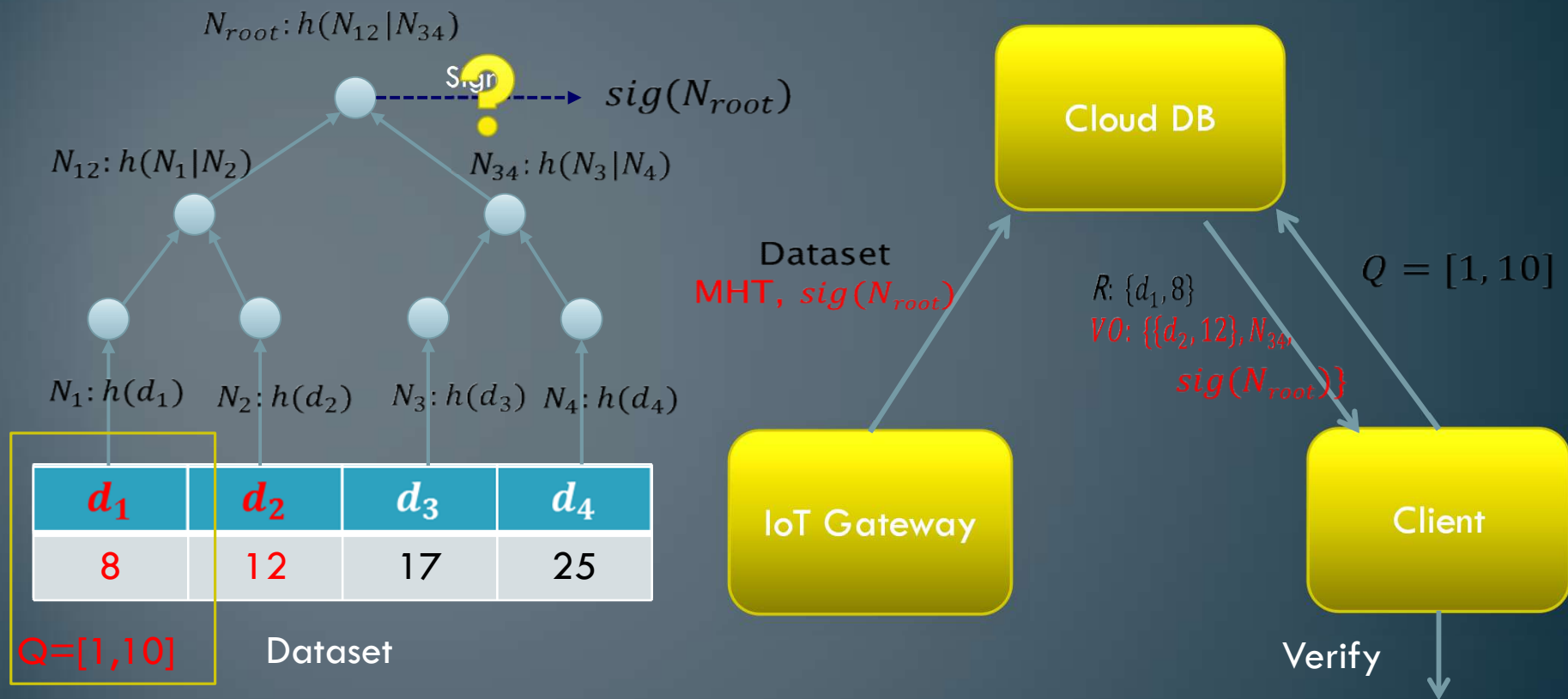
- ✓ IoT and Outsourced Database
- ✓ System, Security and Privacy Model
- Case Studies
  - 1: Single-Source Privacy-Preserving Integrity Assurance
    - 2: Integrity Assurance with Access Control and Zero-knowledge Proof
    - 3: Local Differential Privacy for Key-Value Pairs
- Conclusion

# Case Study 1: Privacy Preserving Single Source Integrity Assurance

- Motivation: IoT data analytics

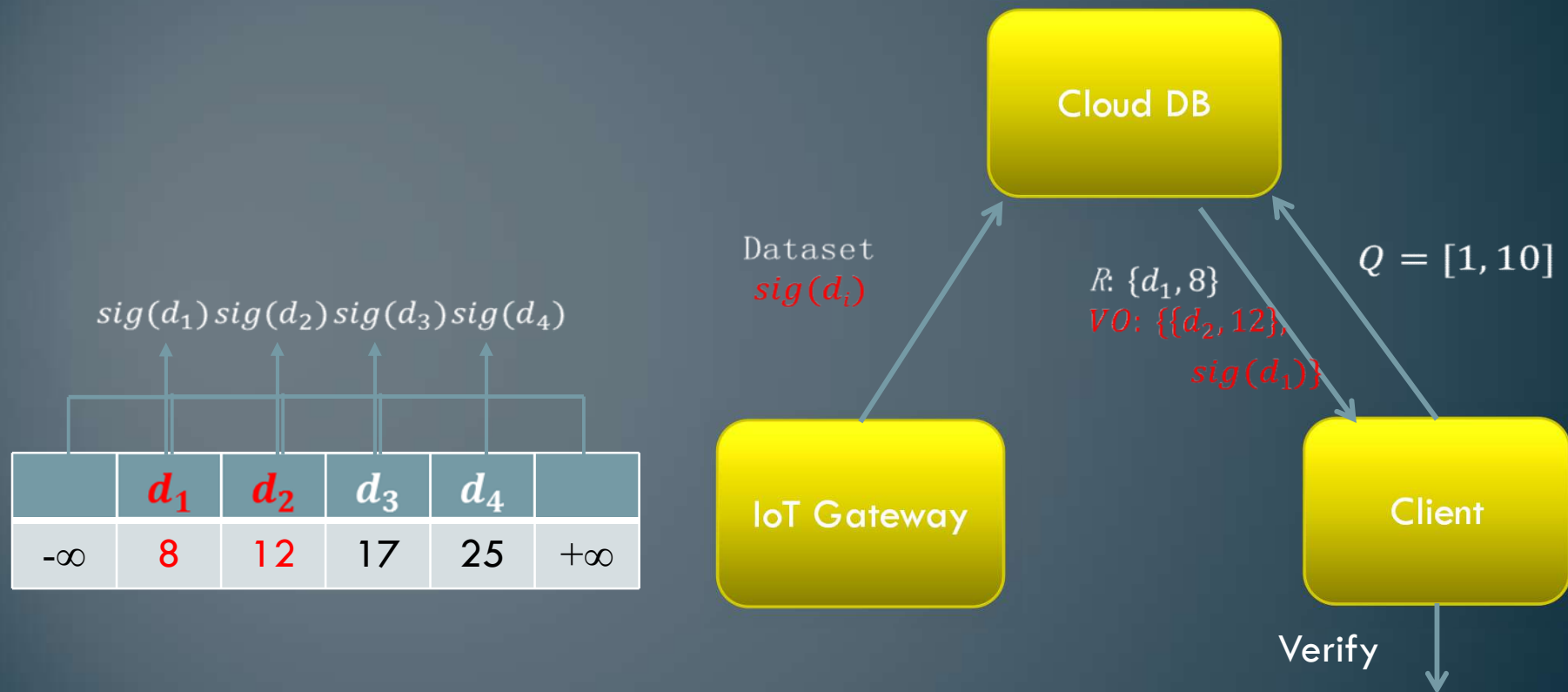


# Integrity Assurance Tool: Merkle Hash Tree



- Soundness:  $8 \in [1, 10]$ ; root sig
- Completeness:  $12 \notin [1, 10]$

# Integrity Assurance Tool: Signature Chaining



- Soundness:  $8 \in [1, 10]; sig(d_1)$
- Completeness:  $12 \notin [1, 10]$

# Comparison

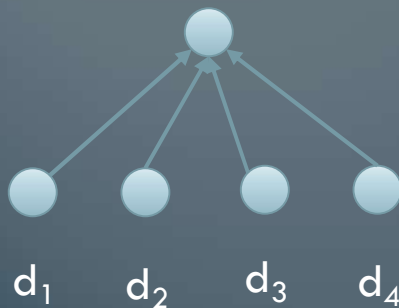
- Pros of MHT

1. Scalable to large ranges
2. Data do not need to be sorted

- Pros of Signature Chaining

1. Efficient for small ranges on large datasets
2. Verification object does not need to include any intermediate nodes

$h(\text{mbr} \mid h(d_1) \mid h(d_2) \mid h(d_3) \mid h(d_4))$



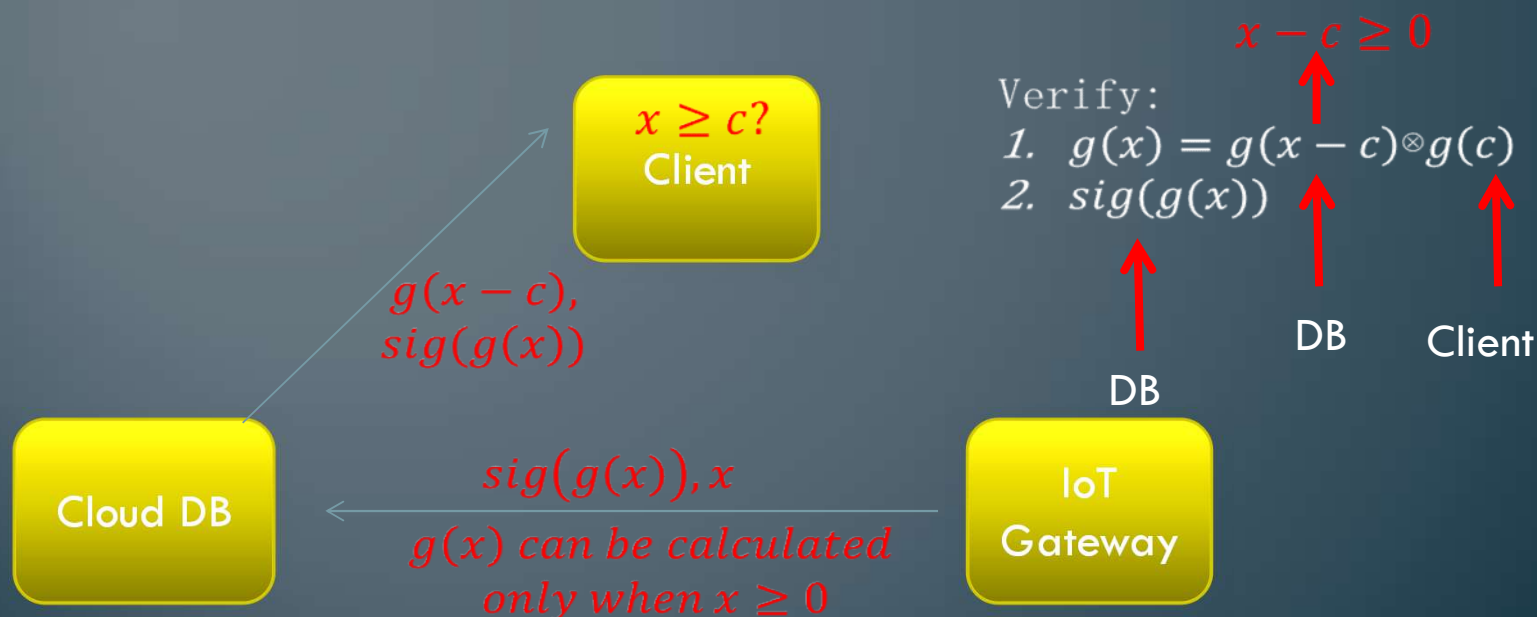
# How Privacy Comes into Play?

- Goal: hide all non-result data for verification

$d_1$	$d_2$	$d_3$	$d_4$
8		17	25

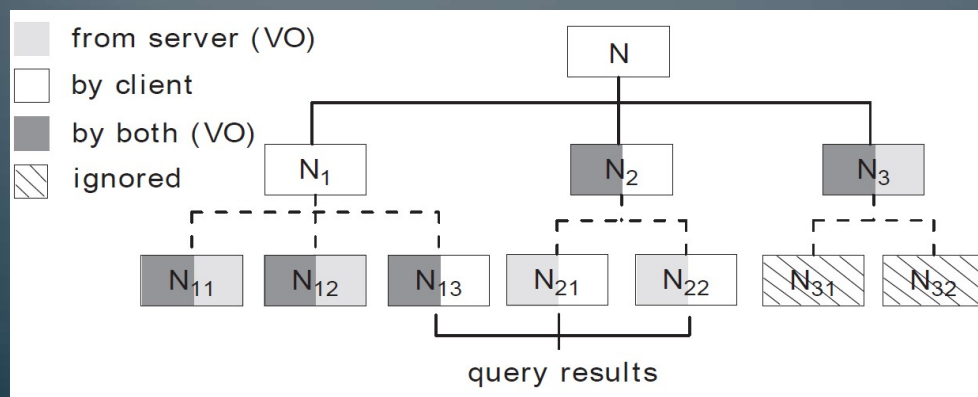
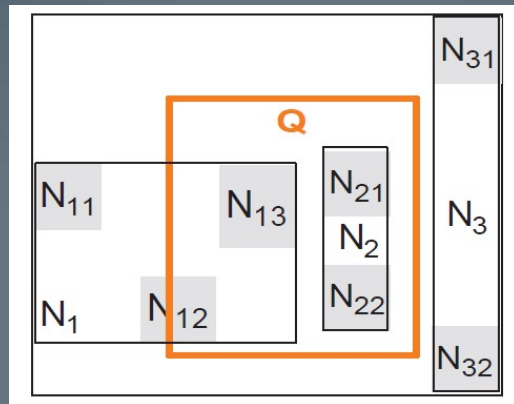
$Q=[1,10]$

- How to verify  $x \geq c$  without knowing  $x$  (as above  $x = d_2, c = 10$ )

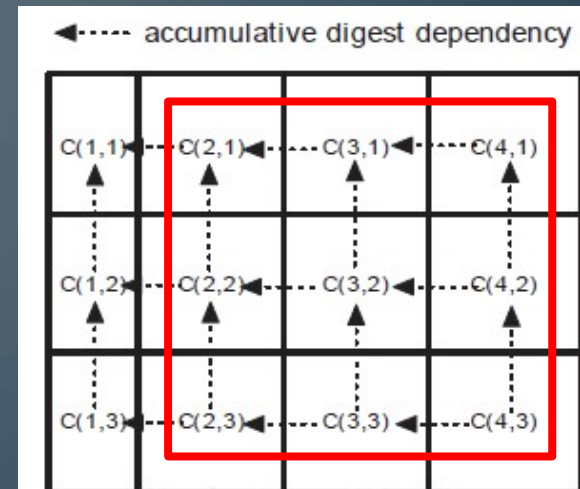
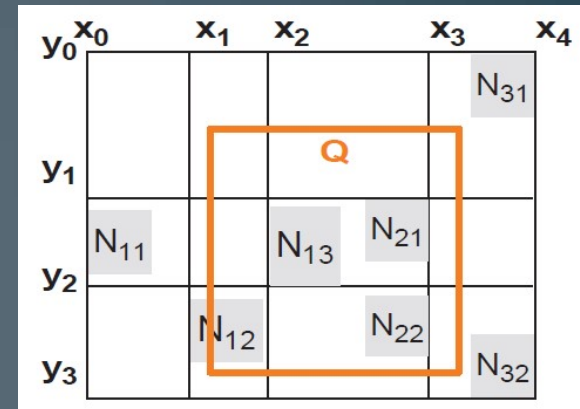


# Two-Dimensional Case (ACM SIGMOD '12)

- MHT-based



- Signature Chaining



# Top-k (PVLDB '14) and Skyline Query(IEEE TKDE' 14)

- Dataset with multi-attributes
  - For example, (1) current, (2) temperature, (3) luminance
- A **top-k query** returns the best k results according to a linear ranking function
  - $\text{Rank}(p) = \alpha \cdot \text{current}(p) + \beta \cdot \text{temperature}(p) + \gamma \cdot \text{luminance}(p)$
- A **skyline query** returns all results that are not worse than other results in at least one attribute.

Lamp ID#	Current	Temperature	Luminance
536	180	67	1500
545	175	72	1400
565	190	70	1600

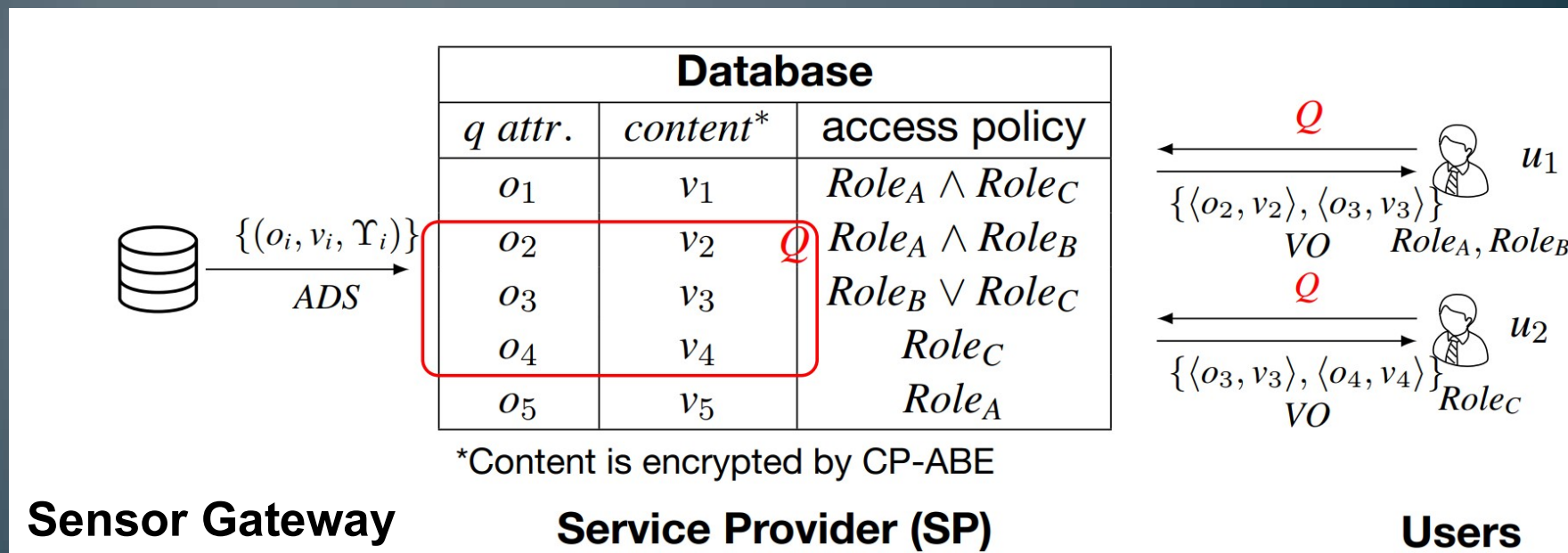
Top-2 most abnormal results: #536 and #545

Skyline results: all three



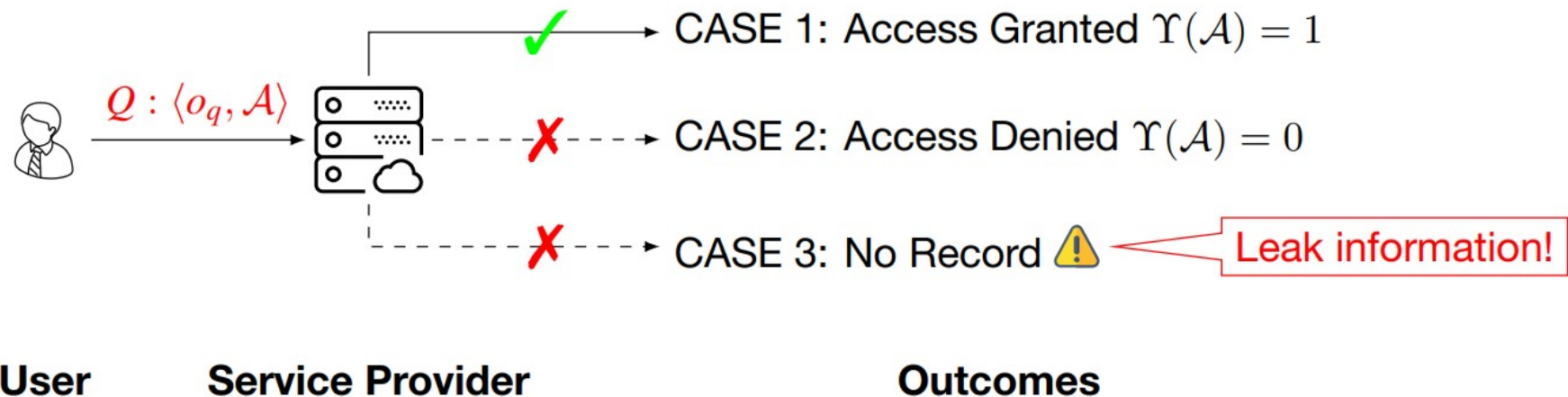
## Case Study 2: Access Control and Zero-knowledge Proof (ACM SIGMOD '18)

- A commercial db usually has complex role sets and access policies (e.g., an IoT database of health care and biomedical sensors)



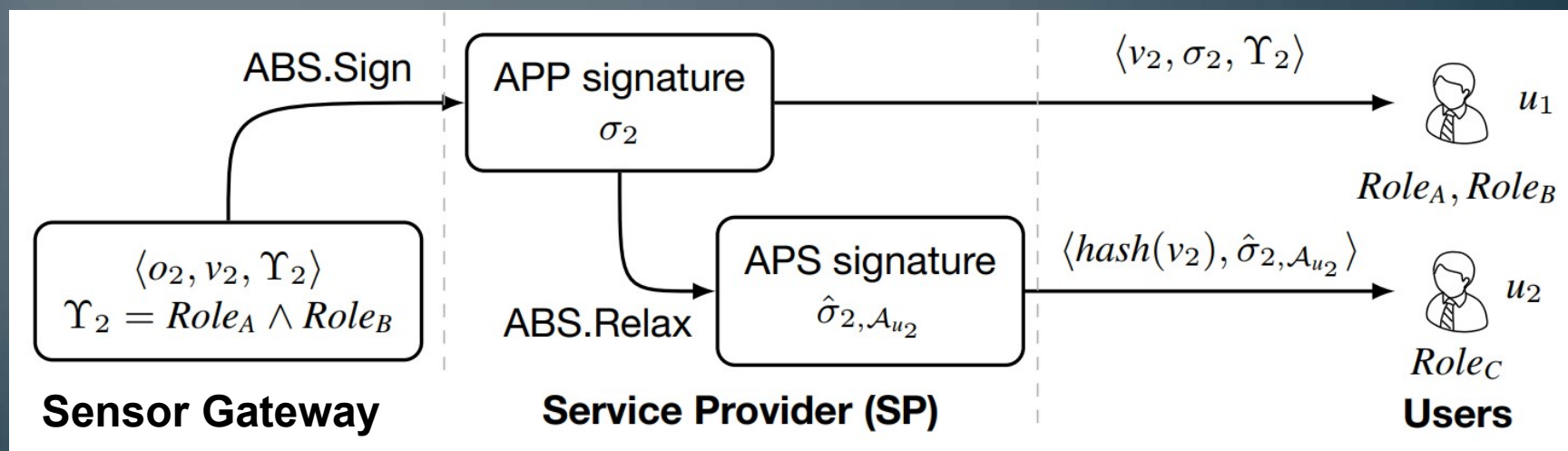
# Why Zero-Knowledge?

- We don't want the user to tell the difference between case 2 and 3.



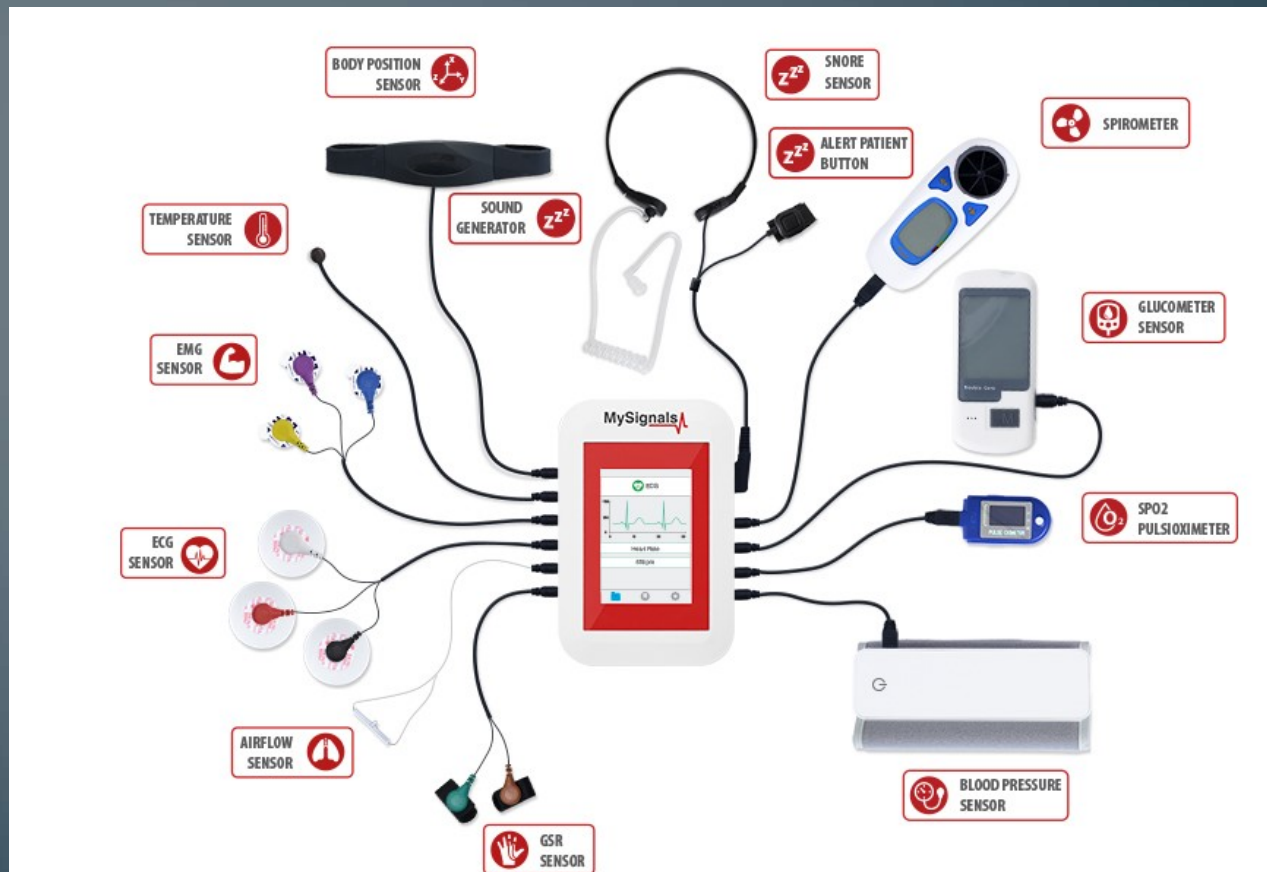
# Solution by Attribute Based Signature

- If the user has all roles, the ABS signature signed by the data owner can be returned directly as VO.
- Otherwise, the cloud db (SP) can perform a relaxation of the signature, which is sufficient to tell the user the access is denied either because the record does not exist, or the roles are insufficient.



# Case Study 3: Local Differential Privacy for Key-Value Pair (IEEE S&P 2019)

- Motivation 1: Wearable devices and biomedical sensor data
- Keys: blood pressure, glucose, heart rate, ECG, EMG, body position, body temperature, etc.



## Motivation 2: Mobile Usage Data Collection



Active in dating activities



Active in online shopping

# Canonical Problem Formulation

- Given

Users	Items
Alice	<id1, 7> <id2, 3> <id4, 1> <id6, 2>
Bob	<id3, 8> <id4, 3> <id5, 3> <id7, 1> <id8, 9> <id9, 8>
Chris	<id2, 5> <id3, 2> <id5, 7> <id8, 9>
Denise	<id2, 7> <id7, 2>

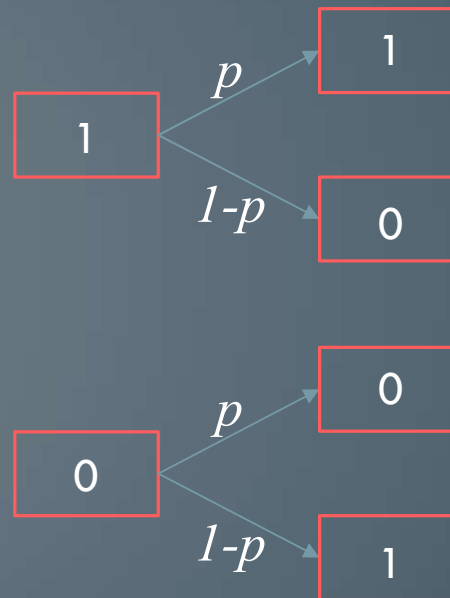


Users	Items
Alice	<1, 7> <1, 3> <0, 0> <1, 1> <0, 0> <1, 2> <0, 0> <0, 0> <0, 0>
Bob	<0, 0> <0, 0> <1, 8> <1, 3> <1, 3> <0, 0> <1, 1> <1, 9> <1, 8>
Chris	<0, 0> <1, 5> <1, 2> <0, 0> <1, 7> <0, 0> <0, 0> <1, 9> <0, 0>
Denise	<0, 0> <1, 7> <0, 0> <0, 0> <0, 0> <0, 0> <1, 2> <0, 0> <0, 0>

- Tasks:**
1. frequency estimation on keys
  2. mean estimation on values for each key

# Straw-man Randomization

Users	Item
Alice	$\langle 1, 0.6 \rangle$
Bob	$\langle 0, 0 \rangle$
Chris	$\langle 0, 0 \rangle$
Denise	$\langle 1, 0.8 \rangle$
Frank	$\langle 1, 0.4 \rangle$
Sally	$\langle 1, 0.3 \rangle$
Tom	$\langle 0, 0 \rangle$
Mike	$\langle 1, 0.3 \rangle$



Original values would reveal the trace of perturbation.

$\langle 1, 0.6 \rangle \longrightarrow \langle 1, 0.6 \rangle$

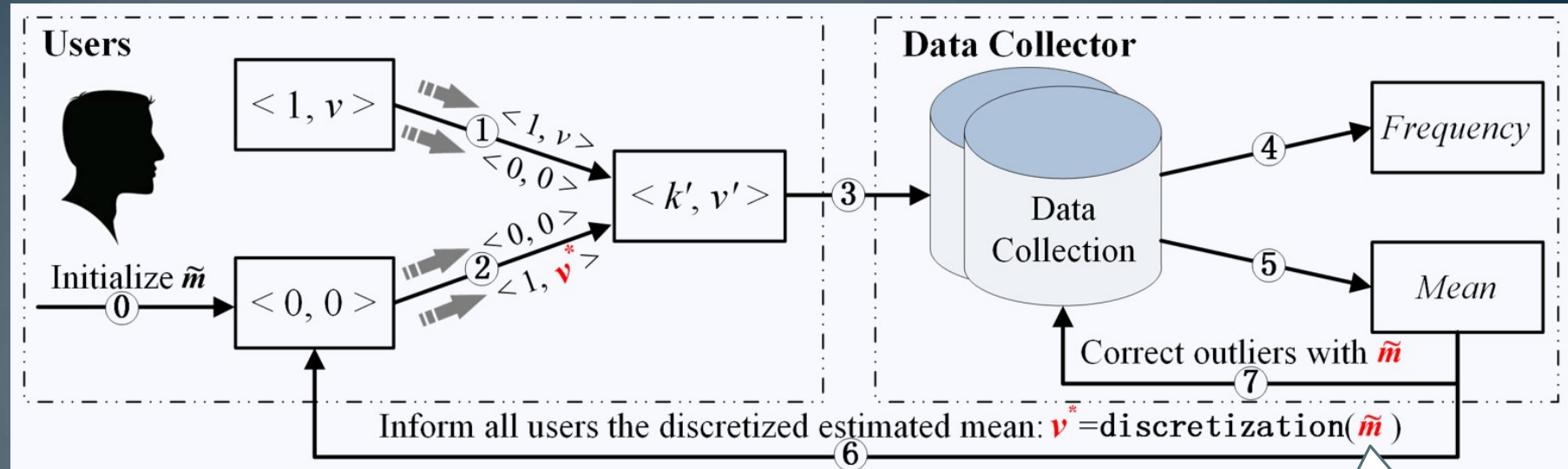
$\langle 1, 0.6 \rangle \longrightarrow \langle 0, 0 \rangle$

$\langle 0, 0 \rangle \longrightarrow \langle 0, 0 \rangle$

$\langle 0, 0 \rangle \longrightarrow \langle 1, ? \rangle$

We need mean value here!

# PrivKVM: An Iterative Execution Model



## Discretization:

Intermediate mean won't be disclosed to users



# Concluding Remarks

- Privacy and integrity are two key security problems in IoT
- There are no on-the-shelf solutions due to the decentralized nature and low capacity of IoT devices.
- Cryptographic-only approach cannot scale well.
- Integration with data engineering techniques are rewarding.



