Privacy Challenges in Big Data and Industry 4.0

Jiannong Cao
Internet & Mobile Computing Lab
Department of Computing
Hong Kong Polytechnic University

Email: csjcao@comp.polyu.edu.hk
http://www.comp.polyu.edu.hk/~csjcao/

Outline

- What’s privacy about?
- Privacy issues
- Existing privacy measures
- New challenges by big data and industry 4.0
- Addressing new challenges
- Summary

What is Privacy?

- The desire of a person / entity to control the disclosure of personal / private information
  - Freedom from observation, intrusion, or attention of others

- Related by different from security
  - Protected from deliberate or accidental damage
  - Include protection of privacy

What is Privacy?

- Many privacy use cases studied
  - Retail and marketing
  - Healthcare
  - Medicine
  - Government
  - Industry
  - Education
  - Transportation
  - ……

Privacy Issues

- How is my data is collected?
- Where is my data?
- How is it used?
- Who sees it?
- Is anything private anymore?

Privacy Risks

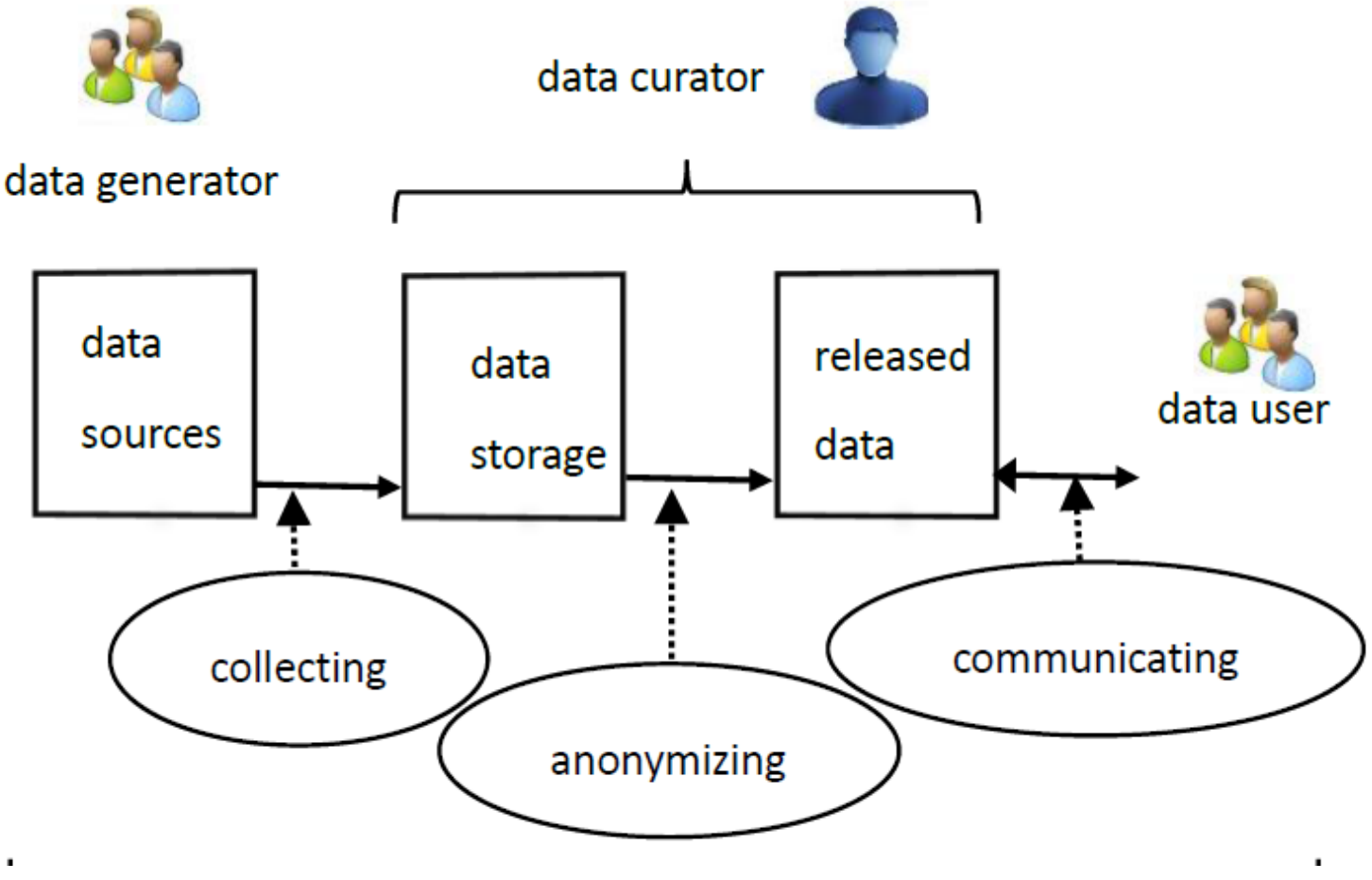
- Unauthorized use
  - Mismanagement of big data: big data constantly at employee fingertips
- Misuse
  - Sale of data for monetary reward
- Stealing
  - Business espionage
- Exploring
  - Searching and mining
- ……

Content Privacy

Interaction Privacy

Comm. Privacy

Privacy Risks



Privacy rights or obligations are related to the collection, use, disclosure, storage, and destruction of personal data

Existing Privacy Measures

- Encryption
  - Encryption of data and communication reduces risk of information theft
- Biometric authentication
- Software risk analysis and verification
- Privacy-preserving data publishing and processing
  - Maximize data utility while limiting disclosure risk to an acceptable level
  - Personal and population privacy
  - Data anonymization
  - Query and computing on encrypted data

Data clustering

- K-anonymity
  - If the information for each person contained in the release cannot be distinguished from at least k-1 other individuals in the release

Name	Job	Gender	Age	Disease	Other

Key Attributes

Job	Gender	Age	Disease	Other

Quasi-identifier

Job	Gender	Age	Disease	Other
Artist	F	[35-40)	Flu	NA
Artist	F	[35-40)	Cancer	NA
Professional	M	[30-35)	Flu	NA
Professional	M	[30-35)	Hepatitis	NA

K-anonymity table (k=2)

  - Works only on quasi-identifiers but not sensitive attributes
    - Subject to subtle but effective attacks, e.g., homogeneity attack due to lack of diversity in sensitive attributes, and background knowledge attack based an adversary's knowledge of the victims

Data clustering

- l-diversity
  - Extending k-anonymity by including the sensitive attributes: ‘‘guarantee there are at least l distinct values for the sensitive attributes in each *qid* group.”

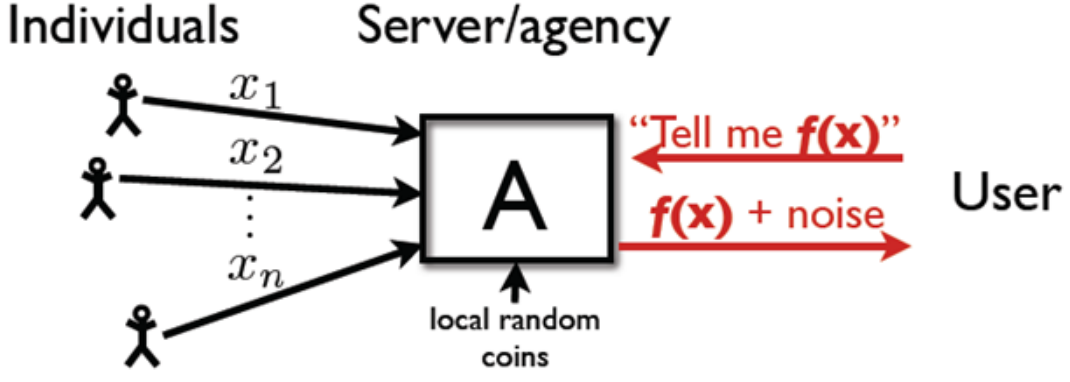
Job	Gender	Age	Disease	Other
Artist	F	[35-40)	Flu	NA
Artist	F	[35-40)	Flu	NA
Artist	F	[35-40)	Cancer	NA
Artist	F	[35-40)	Cancer	NA

K-anonymity table (k=2) obeying l-diversity (l=2)

- t-closeness
  - “guarantee distribution of data on a *qid* group is bounded by *t* against its corresponding distribution on the whole data set.”

Differential Privacy

- Framework of differential privacy
  - Prevent attackers from obtaining private information by multiple queries on top of his knowledge of victims.
  - Strategy: for two data sets with a minimum difference, limit the difference between the queries on the two data sets: reducing information gain for attackers.
  - E.g. output perturbation: add noise to output.



Computing on Encrypted Data

- “I want to delegate processing of my data without giving away access to it”
  - Encrypt my data before uploading it to Cloud, allow Cloud to search / sort / edit the data on my behalf, keeping the data encrypted in Cloud, without needing to ship back and force to be decrypted.
  - Encrypt my queries to Cloud, while still allowing Cloud to process them, returning encrypted answers (I can decrypt)

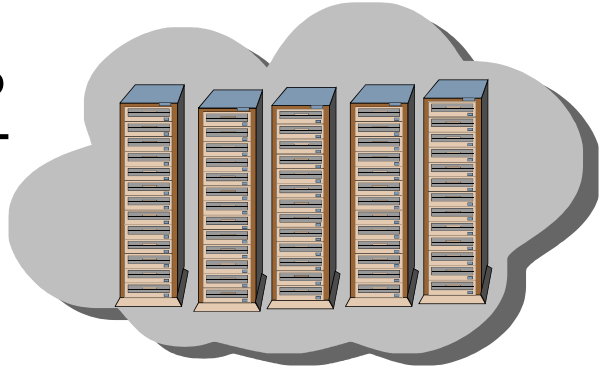
Computing on Encrypted Data

"I want to delegate the computation to the cloud, but the cloud shouldn't see my input"

Enc(x)

P

Enc[P(x)]

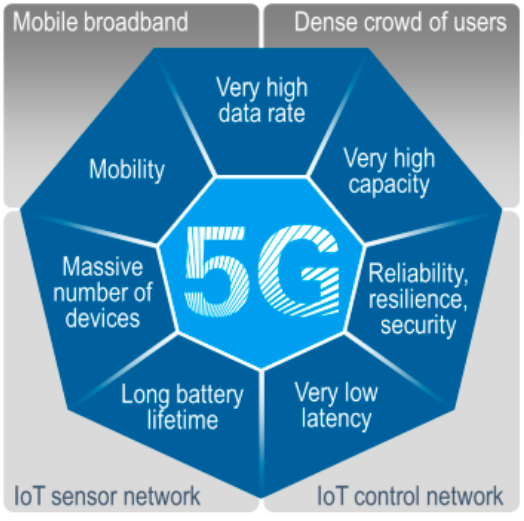


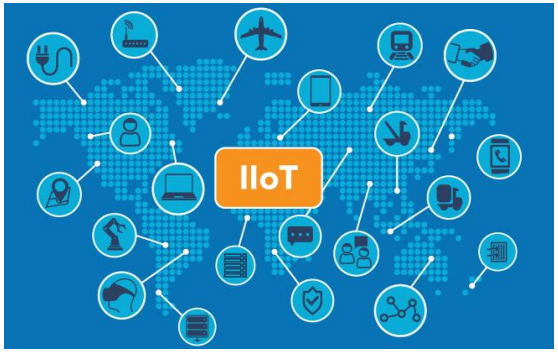
Client
(Input: x)

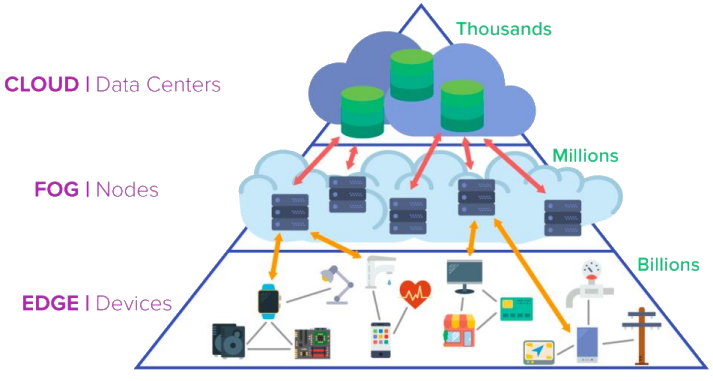
Server/Cloud
(Program: P)

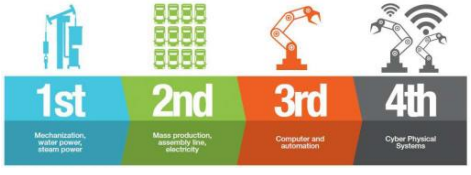
$$f(m) = D_k(f'(E_k(m)))$$

Big Data and Industry 4.0

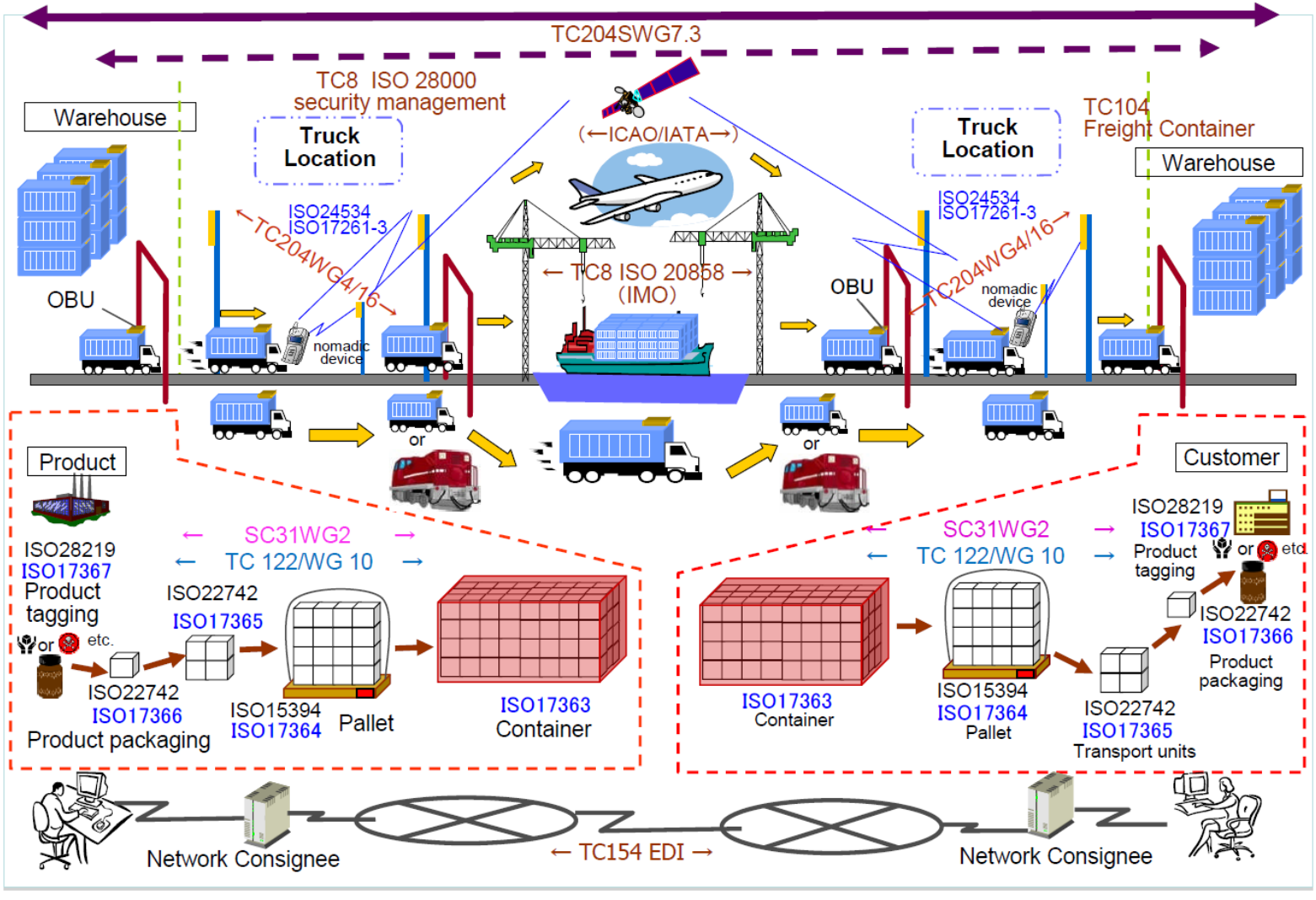






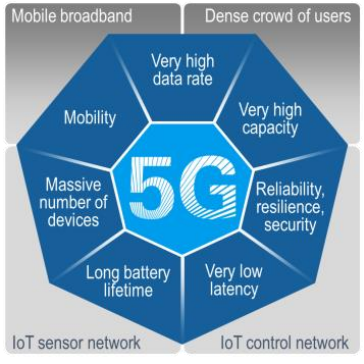


Big Data and Industry 4.0



New Privacy Risks

- Big data & industry 4.0 create intelligent environments and a massive concentration of resources
- Gather significant amounts of data about inhabitants
  - Behavior and usage patterns
  - Personal preferences
  - Inhabitant data
- A massive concentration of risk



  - expected loss from a single breach can be significantly large
- Increased connectivity and interactions of devices
- Complex systems and inextricable problems involving both cyber and physical worlds.

New Privacy Risks

- Information collection and discovery
  - Often voluntary, but also …
  - Legal, involuntary sources
  - Surveillance
  - Monitoring
  - Aggregation
  - Finding missing / new pieces of information
    - Big Data can violate your privacy without any one piece of data violating your privacy.
    - Threats of massive data mining!

New Privacy Risks

- Massive number of devices and higher mobility
  - Devices (vehicles, smart homes, watches) with Cloud / mobile interfaces are vulnerable to attack
  - Tiny objects with simple artifact and scarce resource
    - e.g., potentially Trillions of Tag Reads per Day with basically zero security on the tags.
  - Smart watch communications trivially intercepted / transmitted without encryption.

New Privacy Risks

- Remote access facilities
  - Intelligent environments can frequently be accessed remotely over the network
- Computer-enabled access to the site
  - Intruders can falsify access authentications
- Large, multi-sourced databases
  - Multiple, large amounts of private information represent a target for intruders
- Wireless communications
  - Wireless communications are easy to intercept and hard to regulate due to diversity

Address New Challenges

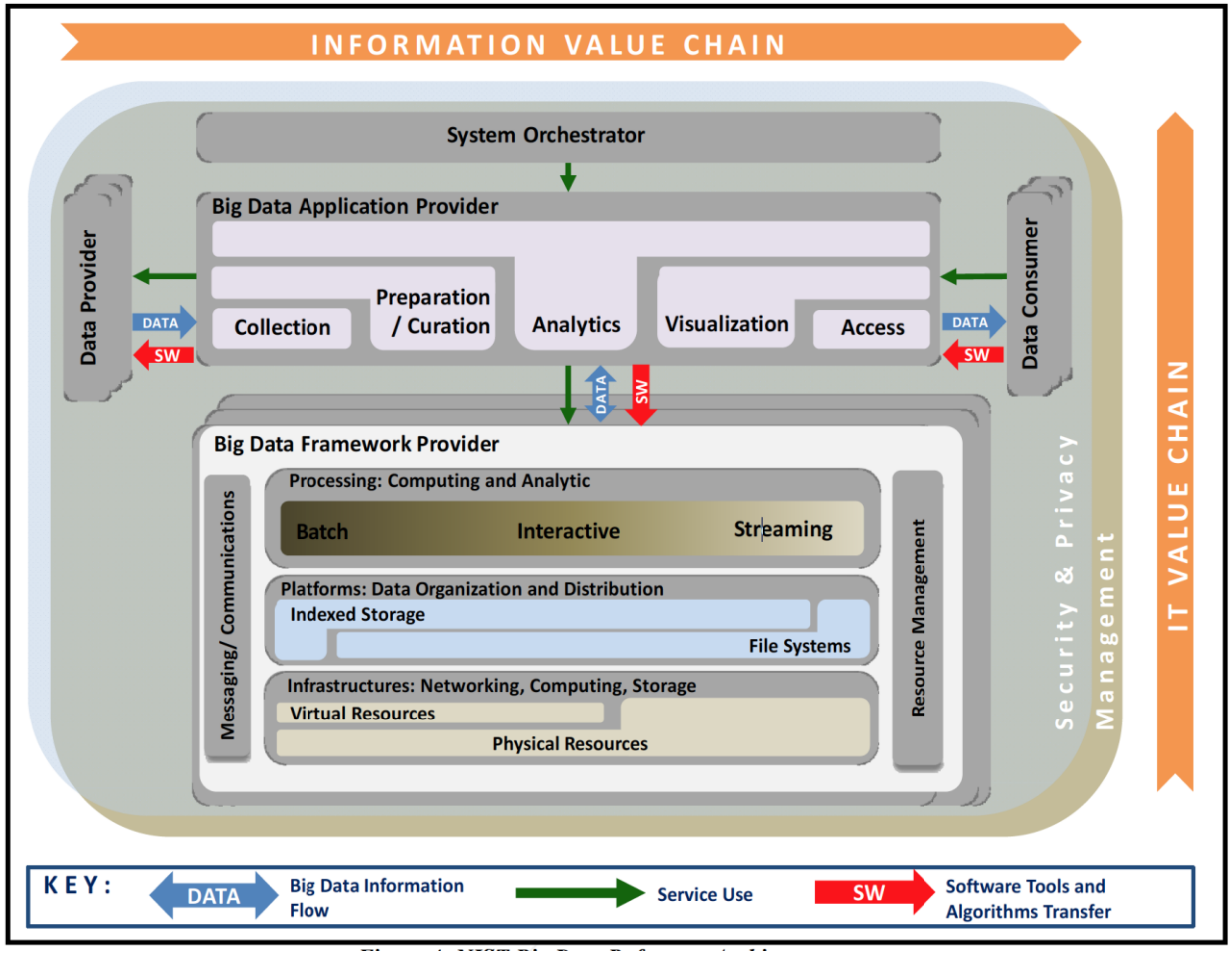
- Seeing research on security for IoT and Cloud but not much on big data and large-scale IoT like Industry 4.0 scenarios.
- Improve existing privacy protection methods to meet new requirements of big data & Industry 4.0.
- Applying big data management and analytics for cyber security and privacy.
- Integrating cross-disciplinary research with social scientists in sociology, psychology, and public policy studies.
  - Having a better understanding of the privacy problem from different perspectives

Address New Challenges

- Need new *privacy framework* including privacy policies, standards, procedures, practices, mechanisms and technology can be implemented, supported and audited
  - Legislation and regulation
  - Data ownership
  - Data governance
  - Technologies guided by laws and policies to address security and privacy issues throughout the lifecycle of the data

Address New Challenges

- NIST Big Data Working Group



Remaining Open Challenges

- The concept of privacy varies widely among (and sometimes within) countries, cultures, and jurisdictions.
  - shaped by public expectations and legal interpretations;
  - a concise definition is elusive if not impossible.
- Measurement of privacy
  - Not only technical, also social and psychological
- New theoretical privacy framework for big data
- Scalability, lightweight and efficiency of privacy algorithms
- Heterogeneity of data sources

Remaining Open Challenges

- Big data and IoT are designed to share information but security and privacy were an afterthought (and it continues to be today)
  - Privacy by design: It is much easier and far more cost-effective to build in privacy, up-front, rather than after-the-fact
  - View privacy as a business issue, not a compliance issue

Summary

- Intelligent environments powered by big data and IoT pose many security and privacy issues
  - Inhabitant privacy has to be protected
  - Access has to be restricted to authorized individuals
  - Communication links have to be secured
  - Software has to be reliable
  - Need scalable and practical solutions to security and privacy
- Very limited privacy research in context of big data and Industry 4.0

Summary

- Practical recommendations (e.g., OWASP):
  - Only collect data the device needs to function
  - Try not to collect sensitive data
  - De-identify or anonymize
  - Ensure the Thing and its components protect personal information
  - Only give access to authorized individuals
  - “Notice and Choice” for end-users if more data is collected than would be expected
- Need greater effort to improve existing work, especially the theoretical attempt and put them into practical use.

Open Web Appllication Security Project (slightly edited)
https://www.owasp.org/index.php/OWASP_Internet_of_Things_Top_Ten_Project

Big data, Industry 4.0 and Privacy – We can have them all!


thank
you!