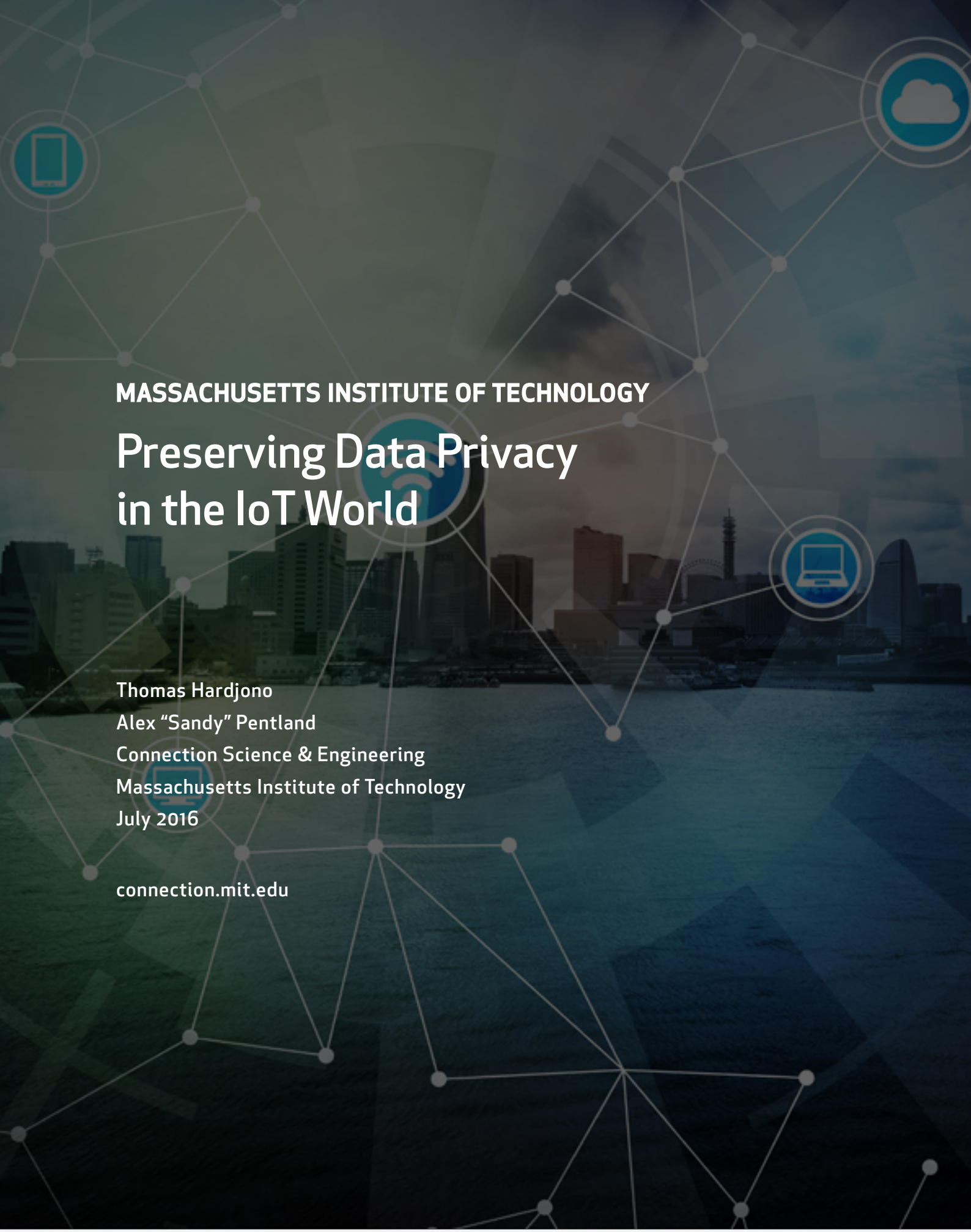

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Preserving Data Privacy in the IoT World

Thomas Hardjono
Alex "Sandy" Pentland
Connection Science & Engineering
Massachusetts Institute of Technology
July 2016

connection.mit.edu

Massachusetts Institute of Technology

MIT Connection Science
the technology of innovation

Disclaimer *The Massachusetts Institute of Technology may have financial or other relationships with one or more entities described in this document. No endorsement, implied or explicit, is intended by discussing any of the organizations or individuals mentioned herein, and is expressly disclaimed.*

Data and Privacy in the Internet of Things: Motivation and Challenges

Recent developments in technology for the Internet of Things (IoT) promise a new world of highly interconnected devices, ranging from consumer electronic devices, to human-embedded medical devices to industrial sensors that control the operations of critical infrastructure. The lessons of big data from the past decade have taught us that *data increases in value when it is shared*[1]. There are numerous circumstances and use-cases today where data sharing would help communities and cities in finding solutions to societal challenges (e.g. spread of diseases, urban planning, climate change, etc.). There are a number of challenges, however, in data sharing – many of which become more acute and pressing in the face of the current pace of IoT technology developments:

- **Tension between data sharing and privacy:** There is an apparent tension or conflict between the needs of data sharing[1] and the need for preserving the privacy of the owner or source of the data[2].

- **Limited scalability of current distributed processing platforms:** The rise of big data analytics has ushered in new distributed storage and distributed processing platforms such as Hadoop[3] and Spark[4] that require all relevant raw data sets to be available. We believe this approach inherently does not scale for data sharing.

- **Limitations imposed by regulatory requirements:** Different legal jurisdictions have placed different regulations regarding cross-organization data sharing and cross-border data flows.

We believe a new set of design principles – architectural, governance and operational – are needed for the future IoT that addresses not only the infrastructure aspects of the IoT devices but also addresses the matter of data sharing and privacy.

Distributed Data Repositories

DESIGN PRINCIPLES:

- Move data as little as possible.
- Unalterably record patterns of communications between databases and human operators, make these patterns publically auditable.
- Unalterably record identity credentials and data operations, and make these credentials and operations auditable with one-time consensus permissions.
- All data is encrypted wherever possible.
- Use P2P architectures whenever possible.

A core design principle in achieving scalable sharing of data is to never release raw data from its repository. Placing control in the hands of the data/repository owner will create a sustainable data environment in which data owners will be incentivized to share data and consequently data uses will be more aligned with the public interest. This principle also combats the dangers of big brother surveillance, and provides the data owner with the ability to control the granularity of the query responses being released by the database.

In this distributed data repository architecture, different kinds of data should be stored separately. This reduces the risk associated with a data breach. The external or remote human querier (or query operator) needs to deploy tools that allow queries or sub-queries to be routed to the correct repository. Depending on the implementation, the query-response model for data processing can be performed in real time over these distributed repositories.

Distributing data across multiple repositories aids in enforcing individual privacy because it makes possible the tracking of the patterns of communications between each repository and the human operators/queriers. This capability arises from the observation that each category of data-analysis operation – whether it is searching for a particular item or computing some statistic – has its own characteristic pattern of communication. We refer to this pattern or signature as *metadata about metadata,* and it allows data owners to monitor the overall patterns of otherwise private communications.

Move the Algorithm to the Data

DESIGN PRINCIPLES:

- Keep data at its repository – never expose raw data.
- Use distributed query processing to send/route queries and sub-queries to correct repositories.
- Each data repository returns only anonymous aggregated results.
- Repository owner controls degree of privacy by controlling the granularity of answers.

Current data processing practice commonly results in privacy and security problems because different types of data are concentrated and then shared among many different data stores. At heart, current practice is due to the Von Neumann model of computing and the tendency of humans to view centralized processing as simpler. By reversing these biases and making distributed, privacy maximizing algorithms the default, we can maximize privacy and security for data processing.

As a corollary to the principle of never releasing raw data, the query algorithm or "logic" must therefore be moved to the data. That is, each data repository should locally provide query processing and data computation capabilities that enables remote queriers to send their query statements or algorithms to the repository.

This paradigm shift is in contrast to the current prevailing approach of the querier having to collect all raw data sets from various sources, loading them to a big data processing platform and performing analytics. We believe that as the need to share data increases together with increased concerns regarding individual privacy, these combined needs will outweigh the costs of processing power (meaning hardware and software) required to implement the model.

Data Always in Encrypted State: at Rest and in Computation

DESIGN PRINCIPLES:

- Raw data must remain encrypted during transit and storage.
- Computation performed on encrypted data.
- Provide controls to data-owner.

Today a major problem in many organizations and institutions is the growing liability (legal and financial) in holding large amounts of customer and user data. Recent hacking and data theft incidents (e.g. Sony and Anthem[5]) indicate that data theft by insiders is difficult to counter using prevailing techniques.

We believe that the best approach to counter internal data theft and external hacking is to maintain data (e.g. from an IoT device) always in an encrypted state, both when data is in storage (e.g. in the file system) and during computations. New cryptographic algorithms and approaches, such as Enigma[6] and homomorphic encryption, allow operations to be carried out on encrypted data without needing to decrypt it first.

As such, we see this as another core principle to achieving data privacy while allowing data to be shared at a global level. The combined use of these new cryptographic approaches with distributed repositories, where data is physically distributed across a large peer-to-peer (P2P) network offers a solution to the issue of resiliency against attacks while at the same time addressing the growing data privacy concerns.

Encode Data Usage Agreements in Legal Trust Networks

DESIGN PRINCIPLES:

- Develop and deploy operational trust networks as the legal foundation for data access and data sharing.
- Maintain strong audit of all data access and usage modes, with a tamper-proof history of provenance and permissions, to ensure that data usage agreements are being honored.
- Ensure a high degree of interoperability with existing trust frameworks that address relevant aspects of data sharing.

Large scale data sharing in an ecosystem – ranging from many small data repositories to a few large repositories – requires all parties to agree to a common operation framework. Such a framework – also referred to as a *trust network* – defines not only a common technical and legal terminology, but also defines the obligations and liabilities of each entity in the data sharing ecosystem. Trust networks combine a computer network that keeps track of user permissions for each piece of data within a legal framework that specifies what can and cannot be done with the data — and what happens if there is a violation of the permissions.

By maintaining a tamper-proof history of provenance and permissions, trust networks can be automatically audited to ensure that data usage agreements are being honored. The emergence of new kinds of P2P networks that affect a distributed ledger system (i.e. blockchain technologies) allows tamper-proof history to be recorded, where the ledger acts as a "public notary" of events that have occurred. Such a solution not only offers a degree of nonrepudiation on the part of the actors, but also achieves scalability and reliability of infrastructure.

Examples of trust networks can be found today in the legal trust framework underlying identity-federation schemes (e.g. FICAM[7], SAFEBioPharma[8], OIX[9]), where often competing entities have to obtain a shared degree of assurance regarding the authenticity of digital identities which members of the organization wield.

DISCOVER A NEW WAY TO THINK ABOUT BIG DATA ANALYTICS

If you want to unlock the potential of big data in your organization, consider strengthening your theoretical knowledge and technical understanding in this **8-week certificate course from MIT.**

EXperimental Learning

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

BIG DATA AND SOCIAL ANALYTICS CERTIFICATE COURSE

Presented in collaboration with leaders in online education, GetSmarter.

FIND OUT MORE

View all offerings from MIT and GetSmarter at mit.getsmarter.edu

We are getting beyond complexity, data science and web science, because we are including people as a key part of these systems. That's the promise of Big Data, to really understand the systems that make our technological society...then you can build systems that are better."

ALEX "SANDY" PENTLAND: FOUNDING FACULTY DIRECTOR OF MIT CONNECTION SCIENCE; COURSE INSTRUCTOR

REFERENCES

[1] Alex Pentland, Todd G Reid, and Tracy Heibeck, Revolutionizing medicine and Public Health – Report of the Big Data and Health Working Group 2013, available at: http://wish-qatar.org/big-data/big-data.

[2] A. Pentland, "Personal Data: The Emergence of a New Asset Class," Report from World Economic Forum, 2011, available on http://www.weforum.org/reports/personal-data-emergence-new-asset-class.

[3] Hadoop, Apache Foundation. http://hadoop.apache.org

[4] Spark, Apache Foundation. http://spark.apache.org

[5] Wall Street Journal, Anthem: Hacked Database Included 78.8 Million People, 24 February 2015, http://www.wsj.com/articles/anthem-hacked-database-included-78-8-million-people-1424807364

[6] MIT Enigma, Enigma: Decentralized Computation Platform with Guaranteed Privacy. Available at: http://enigma.media.mit.edu/enigma_full.pdf

[7] FICAM, U.S. Federal Identity, Credential and Access Management (FICAM) Program, http://info.idmanagement.gov

[8] SAFE-BioPharma Association, Trust Framework Provider Services, http://www.safe-biopharma.org/SAFE_Trust_Framework.htm

[9] OIX, OpenID Exchange, http://openidentityexchange.org