ILLINOIS INSTITUTE OF TECHNOLOGY

CS 528 (Fall 2021)
Data Privacy & Security

Yuan Hong
Department of Computer Science
Illinois Institute of Technology

Syllabus and Introduction

OUTLINE

1. **Syllabus**

2. **Data Breach and Privacy**

3. **Security and Privacy**

COURSE OBJECTIVE

General Goals

- Learn what is meant by data privacy and security
- Why is it important (real world data breaches)
- How is it vulnerable
- How to develop specific secure and privacy preserving techniques
- How it will impact the future

Specific Goals

- Fundamental knowledge for data security and privacy, e.g., cryptography, and mathematical bound for protection
- Learn the attacks to data security/privacy and different adversarial models
- Design and implementation of cryptographic protocols
- Design and implementation of privacy preserving sanitization algorithms for large-scale datasets

TOPICS

- Attacks to Data Privacy and Security
  - De-identification, Semantic Attacks, Frequency Analysis, Temporal Attack, Background Knowledge Attacks, etc.
- Data Obfuscation for Data Protection
  - Anonymization Models, Uncertainty Models, Differential Privacy, Local Differential Privacy, etc.
- Cryptographic Techniques for Data Protection
  - Basic Cryptography, Authentication Protocols, Homomorphic Encryption, Fully Homomorphic Encryption, Secure Multiparty Computation, Garbled Circuit for Secure Computation, Private Information Retrieval, Searchable Encryption, etc.

ILLINOIS INSTITUTE OF TECHNOLOGY

COURSE CATEGORY

- A Core Security Course:
  - Theory 30%
  - Applications 70%

RELATIONSHIP TO OTHER COURSES

- Introduce the Attack and Defense at the **"Data and Applications"** Level

- Use basic knowledge (will be reviewed) from
  - CS458 Introduction to Information Security
  - CS422 Data Mining

- Very little overlap (Chap 5 Basic Cryptography for Data Security) with
  - CS549 Cryptography

- No overlap with other security courses
  - CS557 Cyber-Physical System Security and Design
  - CS558 Advanced Computer Security
  - CS595 Software Security

WEBPAGE AND FACULTY

Course Info

- **Blackboard**
  - Syllabus
  - Announcements
  - Slides and Handouts
  - Tools and Sample Source Codes
  - Homework & Course Project
  - Online Discussions
  - Other Resources

Faculty

- **Yuan Hong** http://cs.iit.edu/~yhong/
- **Email:** yuan.hong@iit.edu
- **Class Time**: MW, 2:00 pm-3:15pm
- **Office and Classroom: SB-216C or Zoom**
- **Office Hours**: Mondays, 3:30 pm-4:30pm (or by appointment)
- **TA and Office Hours**: TBD

ILLINOIS INSTITUTE OF TECHNOLOGY

ZOOM (FOR ONLINE)

Join Zoom Meeting

https://iit-edu.zoom.us/j/3123122405?pwd=NTZLRFR4alBQbjl5cU1WaklоNEZtZz09

Meeting ID: 312 312 2405

Passcode: CSIIT

Waiting room: disabled for live classes (easy join); enabled for office hours for individual meetings.

Videos will be recorded for each class

Videos will be available after classes!

WORKLOAD AND GRADING

ILLINOIS INSTITUTE OF TECHNOLOGY

Homework – 30%

- Four homework assignments
- Written assignments and/or small programming project

Course Project – 20%

- 2-3 students
- A list of topics will be announced
- A proposal will be due by 10/20/2021
- Design and implementation

Midterm and Final Exams – 25% + 25%

- A 80 or higher
- B 60 or higher
- C 50 or higher
- E less than 50

TENTATIVE SCHEDULE

Chapter	Dates	Topics	
1	8/23	Syllabus and Introduction	
2	8/25, 8/30	De-anonymization Attacks and Data Anonymization	
	9/1	Anonymizing Heterogeneous Data (**HW 1**), 9/6 Labor Day (No Class)	
3	9/8, 9/13	Differential Privacy (I)	Including data aggregation, data mining (machine learning) with DP and LDP (**HW 2**)
	9/15, 9/20	Differential Privacy (II)	
4	9/22, 9/27	Local Differential Privacy	
5	9/29, 10/4	Basic Cryptography for Data Security	
	10/6	Midterm Exam, 10/11 Fall Break Day (No Class)	
6	10/13, 10/18	Secure Multiparty Computation, Garbled Circuit (**HW 3**)	
7	10/20, 10/25	Homomorphic Encryption (HE) and Fully HE	
8	10/27, 11/1	Cryptographic Protocols for Data Mining (Machine Learning)	
9	11/3, 11/8	Zero-Knowledge Proof, Secret Sharing and Data Integrity (**HW 4**)	
	11/10, 11/15	Private Information Retrieval and Searching Encrypted Data	
	11/17	Blockchain and Zerocash	
	11/22, 11/29	11/24, Thanksgiving (No Class), Project Demos/Presentations (TBD)	
	12/1 or TBD	Final Exam	

READINGS

- No required textbook
- Readings will be posted on the blackboard

ILLINOIS INSTITUTE OF TECHNOLOGY

LATE SUBMISSION AND CHEATING

All work has to be original!

- Cheating = 0 points for assignment/exam
- Possibly E in course and further administrative sanctions
- Every dishonesty will be reported to office of academic honesty

Late policy:

- -20% per day

Course project (if in a group):

- Every student has to contribute in **every** phase of the project!

OUTLINE

1. **Syllabus**

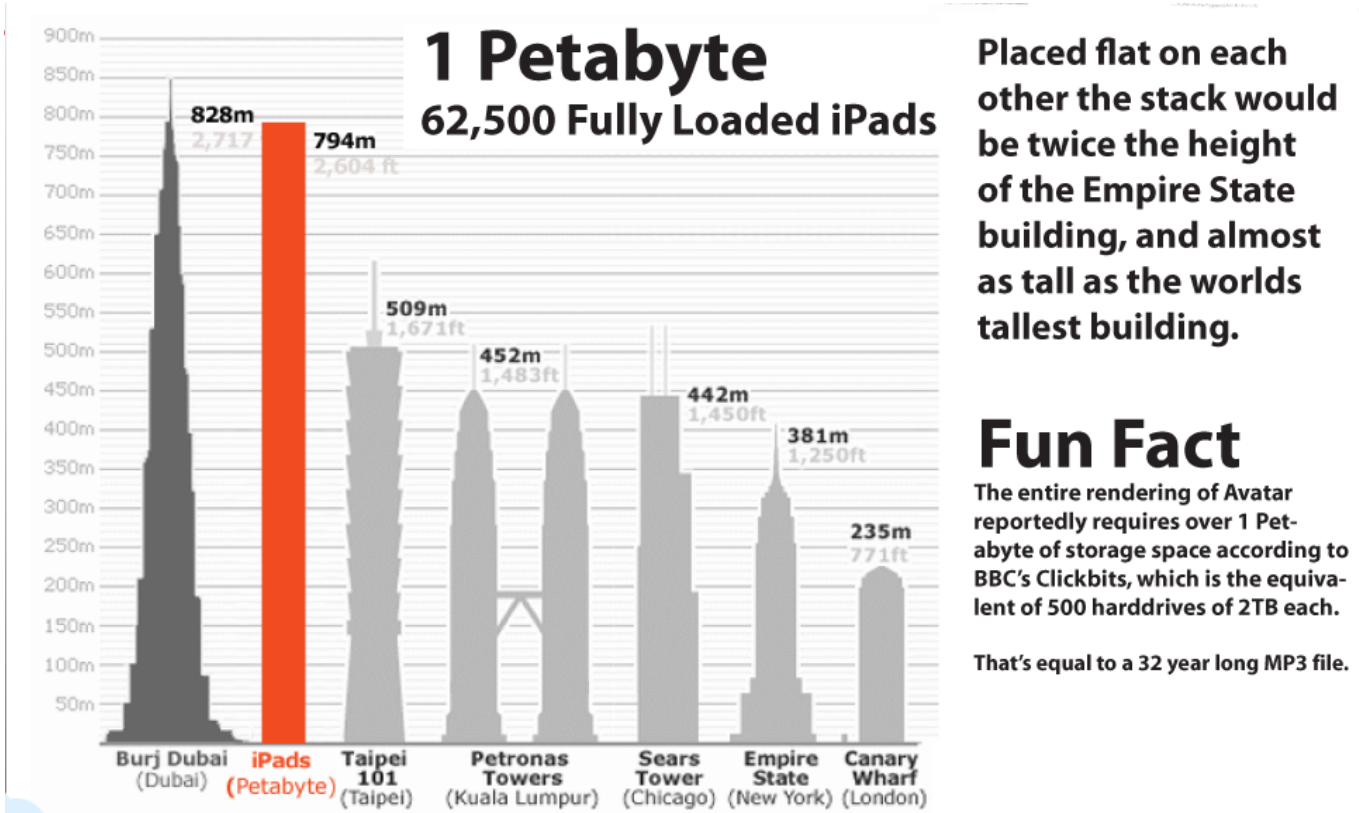
2. **Data Breach and Privacy**

3. **Security and Privacy**

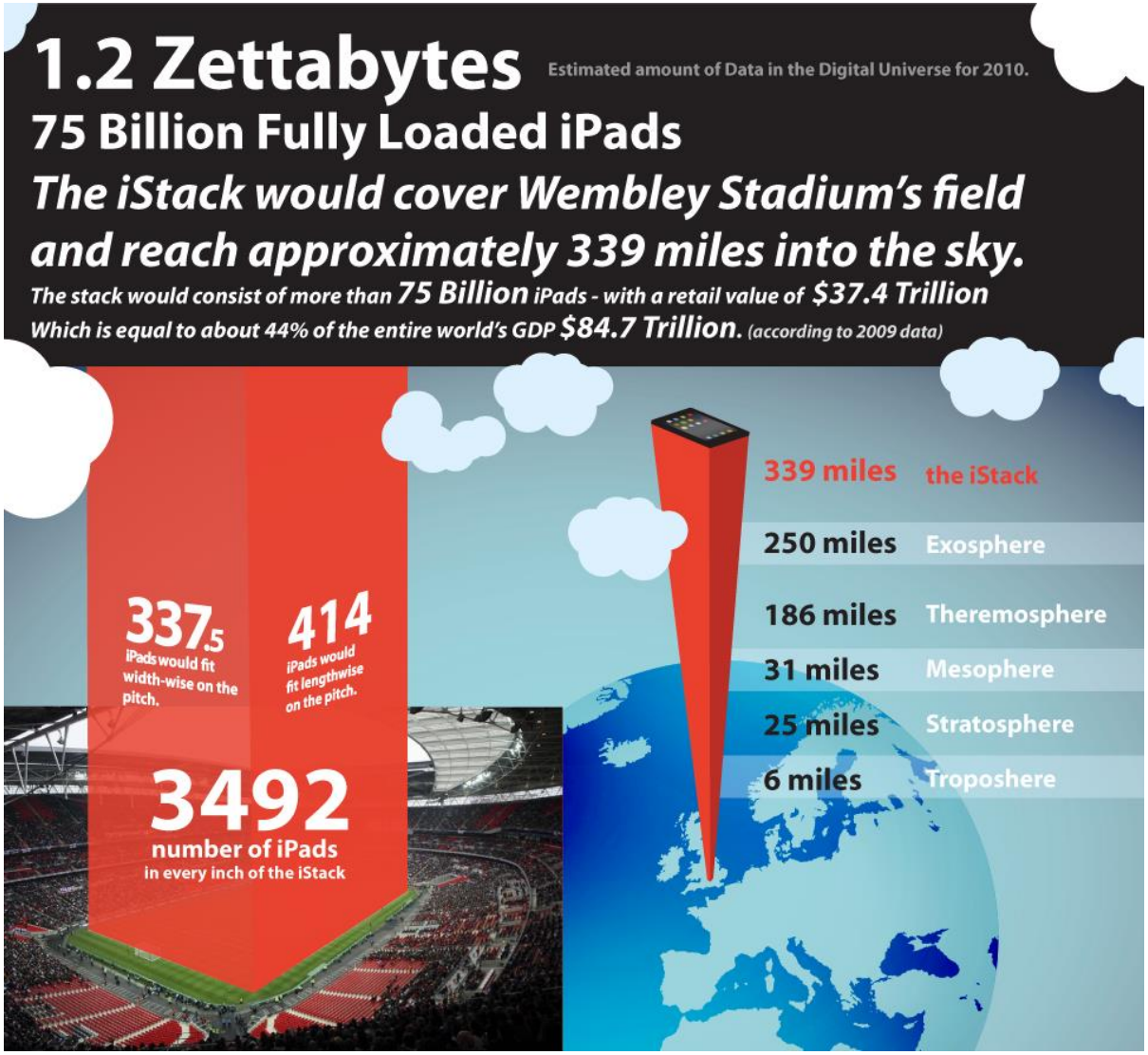
HOW MUCH DATA?

Estimated info added to digital universe each year approaches 40 ZB+ (zettabyte)

- 40x1000000000000000000000 (10^{21}) bytes
- From: https://blog.100tb.com/how-big-is-the-digital-universe, June 2019



THE GUARDIAN – INFOGRAPHIC



ILLINOIS INSTITUTE OF TECHNOLOGY

WHAT IS HAPPENING WITH THIS DATA?

Data will be ubiquitously collected, stored, and analyzed to benefit our society

- Big Data (Volume, Velocity, Variety)
- Data are correlated

Privacy and Security are becoming increasingly important

Data (without meaningful Analysis) is useless

How can we resolve this conundrum?

- Do we utilize the data, or throw it away? If so, how?

DATA ANALYSIS

Data Analysis (a broad concept in this course)

- Scientific computation
- Statistical analysis
- Data mining
- Machine learning
- Artificial intelligence
- Computer vision
- Business analytics/intelligence
- Database queries
- Online analytical processing
- …

PRIVACY INCIDENTS

- Sony CD Spyware

- Samsung Smart TV Snooping

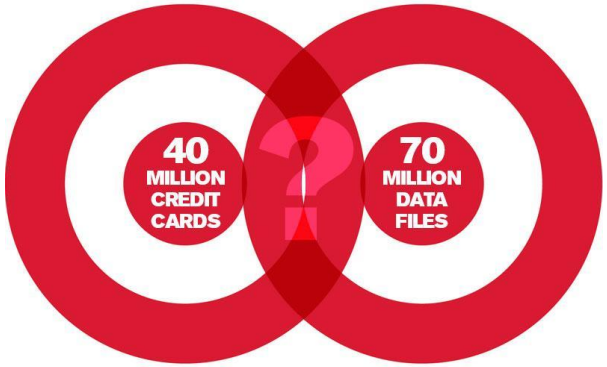
- Netflix Privacy Leakage

- AOL Privacy Leakage

RECENT DATA BREACHES

ILLINOIS INSTITUTE OF TECHNOLOGY

- Facebook 2018
- Equifax 2017
- Anthem 2015
- Target 2014
- Yahoo 2014
- Adobe 2013



AOL PRIVACY LEAK

- In August 2006, AOL released search history of 65k users over a 3-month period.
  - User IDs are replaced with random numbers (naïve anonymization)
  - 3 days later, data is available for public access

AOL searcher # 4417749
"landscapers in Lilburn, GA" *queries on last name "Arnold"* *"homes sold in shadow lake subdivision Gwinnett County, GA"* *"num fingers"* *"60 single men"* *"dog that urinates on everything"*

NYT

Thelman Arnold, a 62 year old widow who lives in Liburn GA, has three dogs, frequently searches her friends' medical ailments.

AOL INCIDENT IN 2006

user-ct-test-collection-06.txt

```
1998497 anthony burger  2006-03-05 13:01:36     2       http://www.anthonyburger.com
1998497 gaither 2006-03-05 13:02:22     4       http://www.bill.gaither.com-music.homepages.org
1998497 allegiant air   2006-03-05 15:27:59     1       http://www.allegiantair.com
1998497 gaithe  2006-03-05 17:07:32
1998497 gaither 2006-03-05 17:07:44     7       http://www.gaither.com
1998497 gaithe  2006-03-05 17:09:53
1998497 gaither 2006-03-05 17:10:03     7       http://www.gaither.com
1998497 allegiant air   2006-03-05 18:22:26     1       http://www.allegiantair.com
1998497 disney coronado springs resort orlando fl       2006-03-07 14:09:08     5       http://hotels.about.com
1998497 www.hli.com     2006-03-10 09:05:39
1998497 heritage lottery international  2006-03-10 09:06:56     1       http://blog.supersurge.com
1998497 googlemaps.com  2006-03-11 00:12:28     1       http://www.googlemaps.com
1998497 amy grant       2006-03-11 19:29:34     7       http://www.mindspring.com
1998497 amy grant       2006-03-11 19:29:34     2       http://www.amygrant.com
1998497 amy grant       2006-03-11 19:29:34     5       http://en.wikipedia.org
1998497 david phelps    2006-03-11 19:33:55     1       http://www.davidphelps.com
1998497 imercer.com socil security      2006-03-12 13:58:18
1998497 imercer.com social security     2006-03-12 13:58:30
1998497 www.uhc.com     2006-03-12 15:07:01     1       http://www.uhc.com
1998497 www.metlife.com 2006-03-12 15:31:06     2       http://www.metlife.com
1998497 www.vsp.com     2006-03-12 15:36:37     1       http://www.vsp.com
1998497 www.birdsandblooms.com   2006-03-15 20:06:15
1998497 www.birdsandblooms.com  2006-03-15 20:06:27     2       http://www.birdsandblooms.com
1998497 yahoo.com       2006-03-18 13:32:15     1       http://www.yahoo.com
1998497 google.com      2006-03-18 13:51:35     1       http://www.google.com
1998497 google.com      2006-03-18 14:13:57
1998497 google.com      2006-03-18 14:14:25
1998497 google.com      2006-03-18 14:14:52
1998497 google.com      2006-03-18 14:15:17
1998497 google.com      2006-03-18 14:15:54
1998497 google.com people       2006-03-18 14:16:17
1998497 www.bostonmarket.com    2006-03-20 19:48:30     1       http://www.bostonmarket.com
1998497 american heart association      2006-03-24 16:58:34     1       http://www.americanheart.org
1998497 american cancer society 2006-03-24 19:45:55     5       http://www.acs-tx.org
1998497 american cancer society 2006-03-24 19:49:13
1998497 american cancer society 2006-03-24 19:49:23
1998497 american cancer society 2006-03-24 19:50:08
1998497 american cancer society 2006-03-24 19:51:33
1998497 american cancer society 2006-03-24 19:51:54
1998497 american cancer society 2006-03-24 19:52:09
```

NETFLIX INCIDENT

Released 100 million supposedly anonymized movie ratings in "Rental Histories" for "Netflix Prize" (crowdsource the movie recommendation algorithm: $1M award)

Two university of Texas researchers identified many Netflix users from the data by matching their Netflix reviews with data from other sites like IMDb. They also found that if you knew a few movies a Netflix subscriber had rented in a given time period, you could reverse-engineer the data and find out the rest of their viewing history.

Knowing 6-8 approximate movie ratings and dates is able to uniquely identify a record with over 90% probability.

Netflix canceled the second phase of the challenge.

http://www.cs.utexas.edu/~shmat/netflix-faq.html

PRIVACY

Privacy reflects the ability of a person, organization, government, or entity to control its own space, where the concept of space (or "privacy space") takes on different contexts

- Physical space, against invasion
- Bodily space, medical consent
- Computer space, spam
- Web browsing space, Internet privacy

PRIVACY IS NOT JUST FOR INDIVIDUALS

"CONFIDENTIALITY"

SOME U.S. PRIVACY LAWS.

Year	Title	Intent
1970	Fair Credit Reporting Act	Limits the distribution of credit reports to those who need to know.
1974	Privacy Act	Establishes the right to be informed about personal information on government databases.
1978	Right to Financial Privacy Act	Prohibits the federal government from examining personal financial accounts without due cause.
1986	Electronic Communications Privacy Act	Prohibits the federal government from monitoring personal e-mail without a subpoena.
1988	Video Privacy Protection Act	Prohibits disclosing video rental records without customer consent or a court order.
2001	Patriot Act	Streamlines federal surveillance guidelines to simplify tracking possible terrorists.

GDPR (2018)

- **<u>Protecting</u> individual data**
  - Personal data (*emails, physical address*, other identifiers such as *IP addresses*, …)
  - Sensitive personal data (*health, biometric* and *genetics*, etc.)
- **Guaranteeing <u>transparency</u> in data processing, fairness in the matchup between data processing and its description**
- **<u>Preventing</u> and <u>detecting</u> a data breach to evaluate on a periodic basis, the effectiveness of security practices**

CONVENIENCE VS PRIVACY

Legitimate Usage of Tracking technologies

- Safer streets
- Cheaper communications
- Better government services
- Easy and personalized shopping

Hard to measure the value of privacy

NEW THREATS TO PRIVACY

DNA Databases for medical research

Cheap tiny microphones

Small video cameras

Smart phones & Apps

Biometrics

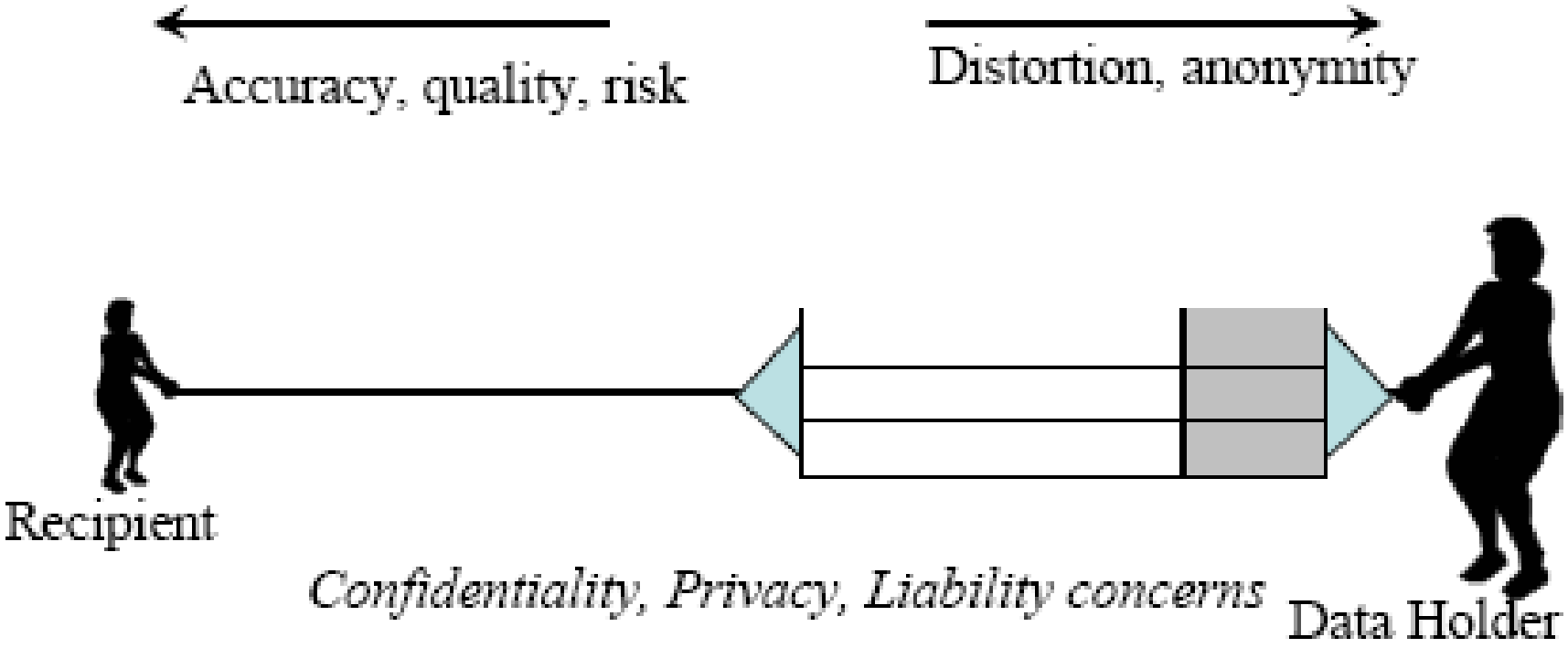
RFID chips

IoT devices

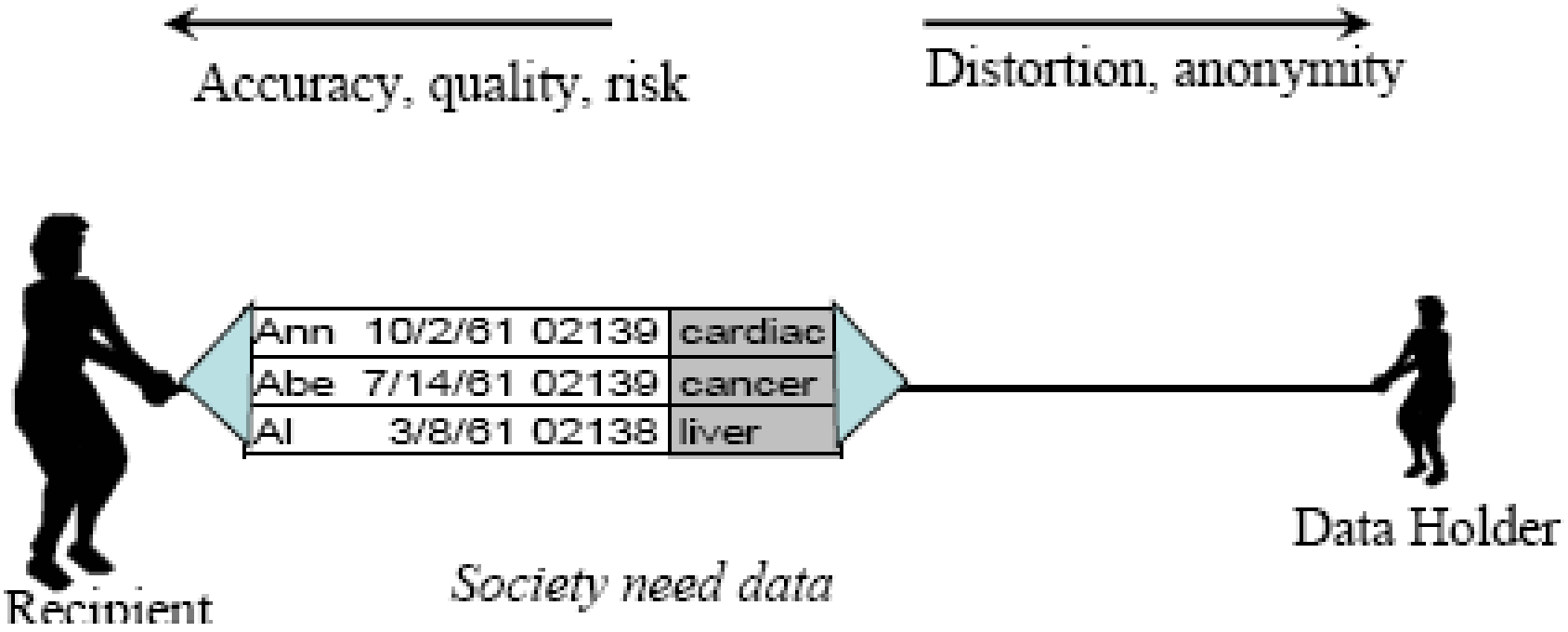
Smart grid & smart homes

…..

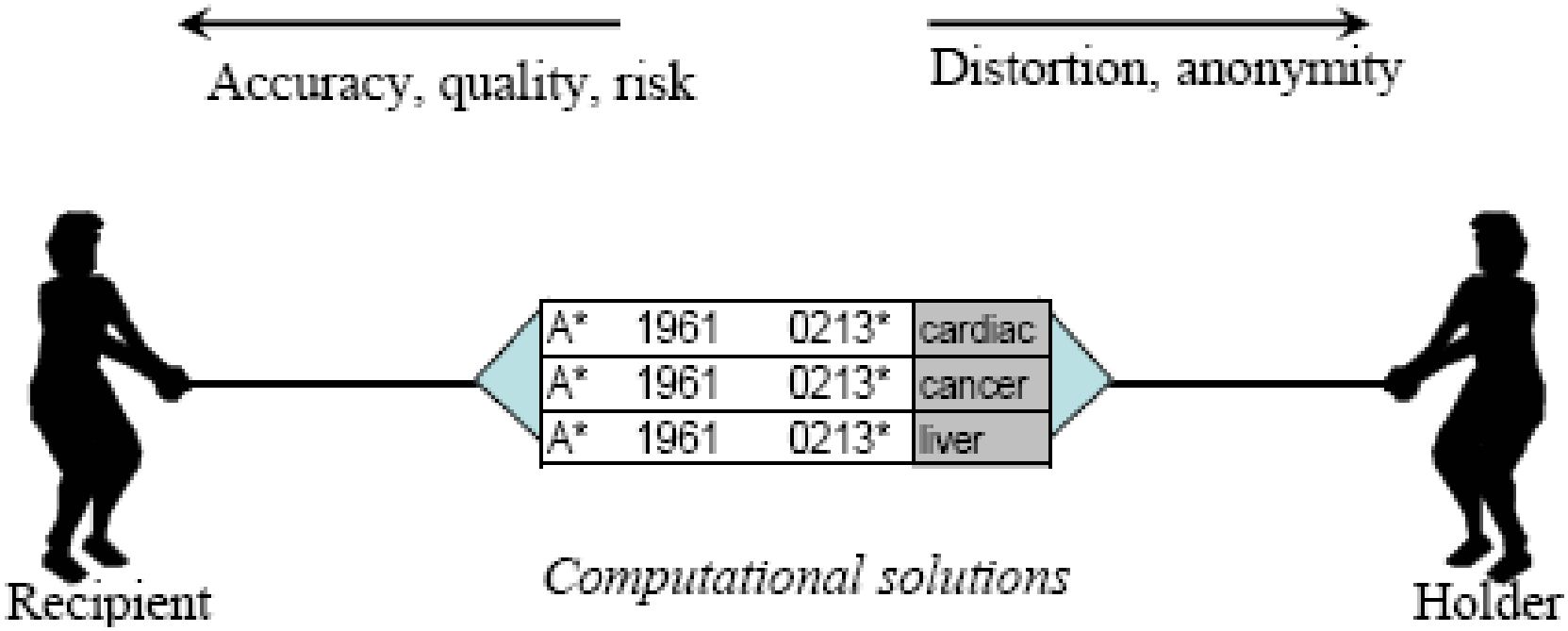
"CANNOT RELEASE DATA"



"PRIVACY IS DEAD, GET OVER IT"



"SHARING DATA WHILE PROVIDING GUARANTEES OF PRIVACY"



SOLVING THE PRIVACY PROBLEM

ILLINOIS INSTITUTE OF TECHNOLOGY

The emergence of many new technologies becomes increasingly hampered by privacy concerns because these technologies leave society vulnerable to privacy abuses.

Current situation

- Let society choose between benefiting from the technology and maintaining privacy protections

Information technology solution → privacy technology.

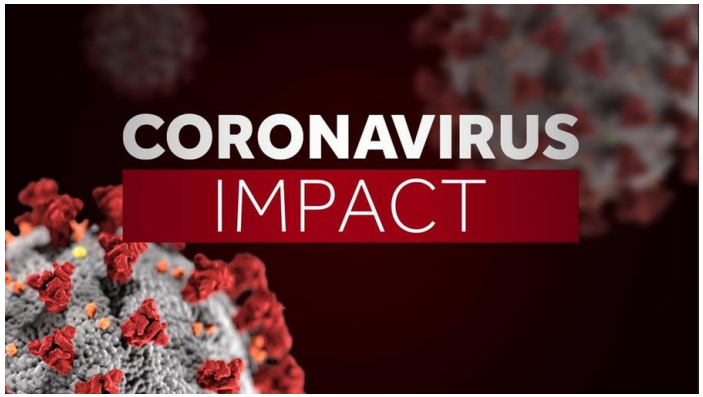
- Construct privacy technology with provable guarantees of privacy protection while allowing society to collect and share person-specific information for many worthy purposes

PRIVACY AND COVID-19

ILLINOIS INSTITUTE OF TECHNOLOGY

Amid the coronavirus pandemic, many tech companies are also stepping up to do their part.

One example is Google's COVID-19 Community Mobility Reports, which are taking aggregate data from those who have turned on *Location History* and using Google Maps to determine how busy certain places are.



PRIVACY AND COVID-19

ILLINOIS INSTITUTE OF TECHNOLOGY

People who have Location History on are already having their location tracked. The only difference now is that this information will be part of the aggregate that is published in the reports.

Instead, they use differential privacy to collect data that grants useful insights into the group, without compromising the privacy of individuals.

PRIVACY AND COVID-19

ILLINOIS INSTITUTE OF TECHNOLOGY

MIT develops privacy-preserving COVID-19 contact tracing inspired by Apple's "Find My" feature.

Automated contact tracing that taps into the *Bluetooth signals* sent out by everyone's mobile devices, tying contacts to random numbers that aren't linked to an individual's identity in any way.

Automate check-ins against the positive chirp database and provide alerts to individuals who should get tested or self-isolate.

PRIVACY AND COVID-19

ILLINOIS INSTITUTE OF TECHNOLOGY

If tests positive, upload a full list of the chirps that their phone has broadcasted over the past 14 days.

Those go into a database of chirps associated with confirmed positive cases, which others can scan against to see if their phone has received one of those chirps during that time.

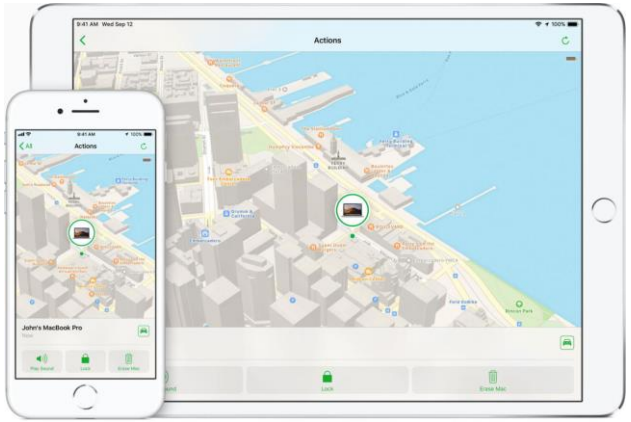
A positive match with one of those indicates that an individual could be at risk, since they were at least within 40 feet or so of a person who has the virus.

A good indicator that they should seek for a test if available, or at least self-quarantine for the recommended two-week period.

PRIVACY AND COVID-19

The system would work through an app they install on their phone, and its design was inspired by Apple's "Find My" system for locating lost Mac and iOS hardware, as well as keeping track of the location of devices owned by loved ones.

"Find My" also uses chirps to broadcast locations to passing Apple hardware.



OUTLINE

1. **Syllabus**

2. **Data Breach and Privacy**

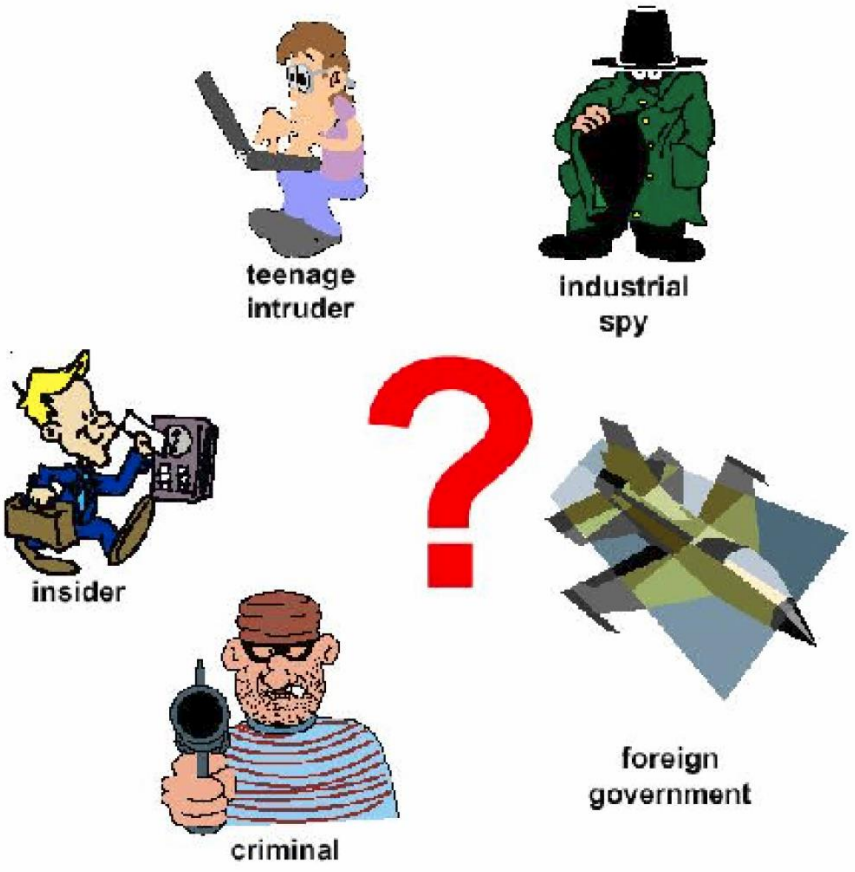
3. **Security and Privacy**

SECURITY

Main Security Properties

- **Confidentiality**
- **Integrity**
- **Availability**
- **Anonymity**
- **Atomicity**
- **Authentication**
- **Fairness**
- **Non-repudiation**
- **…**

Attackers



SECURITY (CONT'D)

Confidentiality

- **Secrecy**
- **Related to privacy**

Integrity

- **Not be accidentally or maliciously altered or destroyed**
- **Related to privacy**

Availability

- **Availability of the resources, e.g., data, device, system.**

Anonymity

- **Unable to identify**
- **Related to privacy**

SECURITY (CONT'D)

Atomicity

- **E.g., transferring money completes entirely or not at all**

Authentication

- **Message authentication**
- **Principal authentication**

Fairness

- **Avoid one of participants being able to gain some advantage over another**

Non-Repudiation

- **Cannot deny**

TWO ASPECTS OF SECURITY AND PRIVACY

Protection: Ensuring Privacy = Improving Security

- Security and Privacy share some similarities, e.g., confidentiality
- Considering privacy protection (e.g., anonymity) as a part of security
- Complement each other

Can security and privacy contradict each other?

- Tradeoff between security and privacy? (any example)