DSci529:
Security and Privacy
In Informatics

Expectations of Privacy
(student presentations)
Big Data
(student presentations and Lecture)
Measuring Privacy
Technology and Privacy

Prof. Clifford Neuman

Lecture 6
19 February 2021
Online

Course Outline

- Overview of Security and Privacy
- What data is out there and how is it used
- Technical means of protection
- Identification, Authentication, Audit
- Reasonable expectation of privacy
- **Big Data – Technology and Privacy**
- AI and Bias
- The Internet of Things and Security and Privacy
- Social Networks and the use of our Data
- Access to Data by Governments - Privacy in a Pandemic
- Privacy Regulation - GDPR, CCPA, CPRA
- Influence of Social Media – Free Speech – Disinformation
- CryptoCurrency - TOR - Privacy Preserving Technologies

Announcements

- Mid-term Friday February 26th - Noon to 1:40PM PST
  - Exam will be followed by Lecture from 2PM to 3:20PM.
  - Review for mid-term at end of today's lecture
  - Exam is online
  - Procedures discussed at end of today's lecture

Today's Outline

- 12:05-12:25 Student Presentations
  - Emily Christiansen – Right or Expectations of Privacy
  - Tian Yang – Expectations of Privacy of Health Data
- 12:30-1:20 Student Presentations – Big Data
  - Tingyi Guo – Big Data Security
  - Kung-Hsiang Huang – Secure Inference
  - Supreet Kaur Randhawa – Breaches of Big Data
  - Zheyu Ren –
  - Lingyu Ge – Big Data and Discrimination
- 1:35-2:10 Lecture Material on Big Data
- 2:10-2:40 Logistics and Review for Mid-term
- 2:40-3:20 Current Event Discussions

The Right or Expectation of Privacy in the United States

Emily Christiansen
Tian Yang

Introduction: Privacy in Large and Small Scale

Emily’s Topic: How can we discuss one of the most arcane documents in our country's history in conjunction with one of the fastest moving modern sciences?

Tian’s Topic: Reasonable expectation of privacy and health data. How is a pandemic changing the “reasonable expectation of privacy”? How digital contact tracing might affect us in the future?

Together these topics dissect the big and small, the opposing scales of cybersecurity and implications of privacy.

The Unenumerated Right to Privacy

- **Griswold vs Connecticut(1965):** The supreme court ruled that a law put in place in Connecticut restricting the purchase and use of contraceptives with married couples unconstitutional.
  - First time that the court recognized ***privacy as an unenumerated right*** and expressed that said right plays a very important role in restricting government intrusion

Griswold v Connecticut Criticism and Response

- Received criticism from the dissenting opinion because it appeared to have pulled a constitutional right out without text to support it
- “We have had many controversies over these penumbral rights of "privacy and repose." These cases bear witness that the right of privacy which presses for recognition here is a legitimate one.” - Justice Douglas

Lasting Effects of *Griswold v Connecticut*

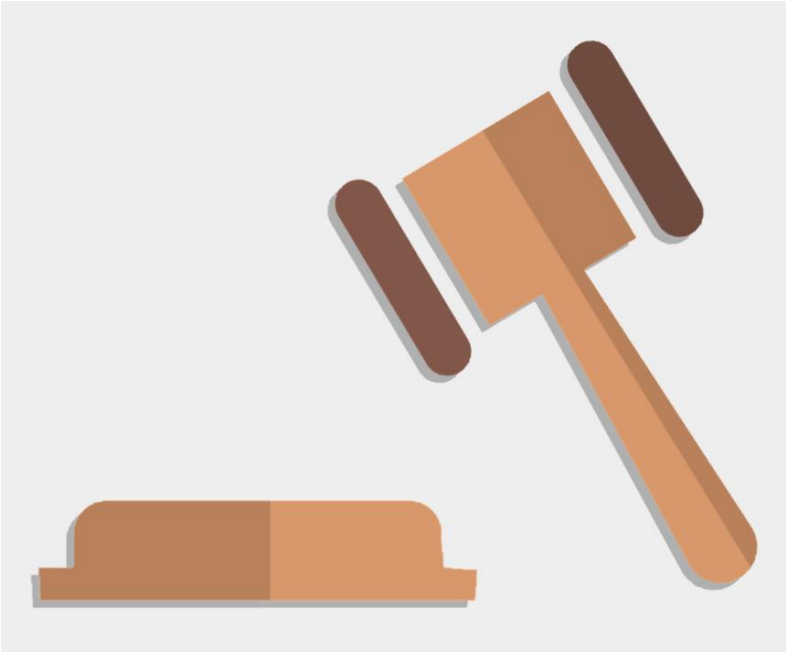
- An important definition to privacy in the Constitution is **penumbra**, which is described in the following way; “The **penumbra** includes a group of rights derived, by implication, from other rights explicitly protected in the Bill of Rights.”
- ***First Amendment Implications:*** The court explained in this case; “The right of freedom of speech and press includes not only the right to utter or to print, but the right to distribute, the right to receive, the right to read and freedom of inquiry, freedom of thought, and freedom to teach. (...) In other words, the First Amendment has a penumbra where privacy is protected from governmental intrusion.”

CISA vs The Due Process Clause of Fifth and Fourteenth Amendments

CISA (Cybersecurity Information Sharing Act)

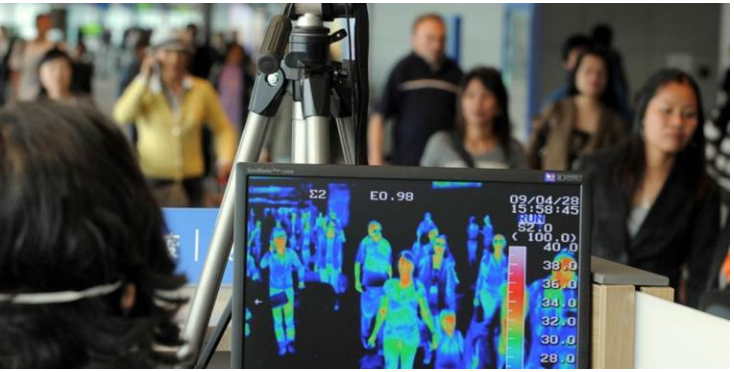
- “CISA encourages businesses to share information they collect on consumers with one another and the federal government without legal barriers (such as warrants), and without the risk of liability.”
- The information that is collected under CISA is not accessible via the Freedom of Information Act, claiming they do not want to reveal trade secrets of corporations

CISA vs The Due Process Clause of Fifth and Fourteenth Amendments



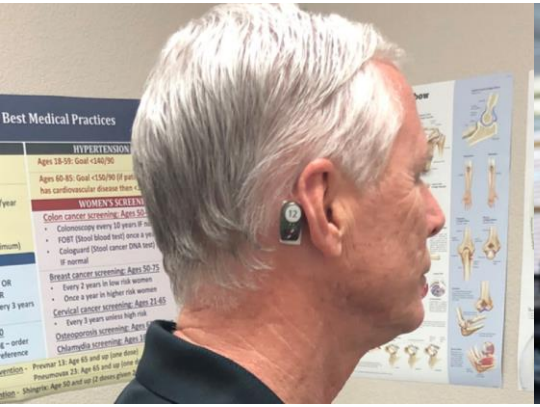
- 4th and 15th: “neither the federal government nor state governments shall deprive a person of “life, liberty, or property, without due process of law”

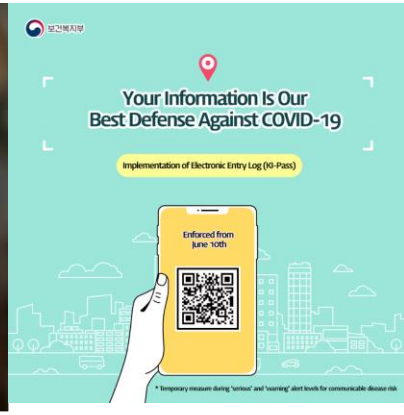
- “CISA allows private actors to identify and share information regarding cyber threats with the government without consideration of due process guarantees.”



Coronavirus

- Coronavirus disease (COVID-19)
- Sensors, QR health code, wearable sensor
- Two health surveillance technologies (contact tracing applications and health monitoring platforms)



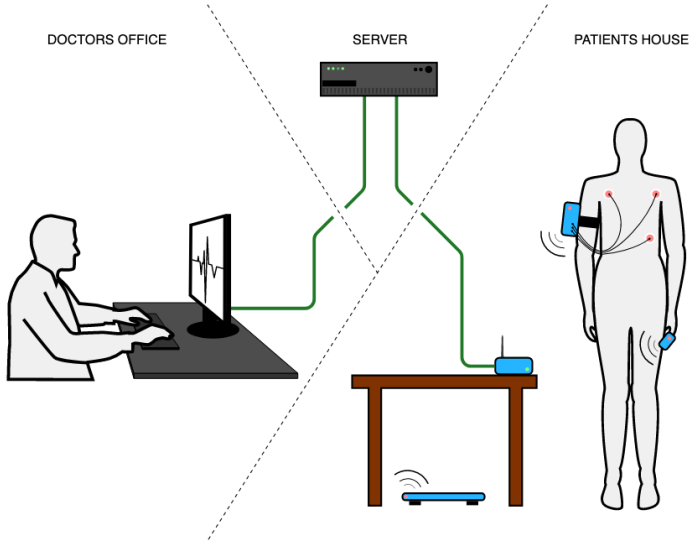


Contact Tracing Applications

- Manually interviews
- Smartphone contact tracing applications that utilize geolocation data, either by the phone's Bluetooth, WiFi, or GPS
- Apple and Google partnered to develop the "Exposure Notification System," ("ENS")
- In the United States, several contact tracing applications appeared on Google's Play and Apple's App Stores (some of which were not tied to a public health agency or authority)

Health Monitoring

- Health Monitoring provides software services that collect and analyze healthcare data
- Remote patient monitoring (RPM): a proven and effective modality that can be used to monitor patients avoiding health care facilities and to care for recovering patients at home given the shortages in acute care facility capacity.
- Some software platforms may not be designed for healthcare applications, yet they incorporate features and capabilities that are valuable to the pandemic response

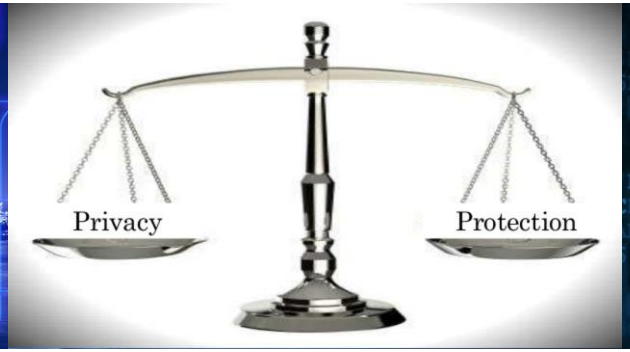


Privacy Risks

- The amount of data being collected, used and stored in response to COVID-19 brings serious risks
- From contact tracing applications and health monitoring, the data can be combined with data sourced from third party.
- The **willingness** to share data during the pandemic increases, but is the current regime provides **protections** for individuals,
- people have lowered their sensitivity to sharing their health data.

Finding the Right Balance

- The technologies and collection of personal data are necessary to combat the virus and slow the spread.
- The data privacy and security risks. What about after the pandemic? Could current legal standards and protections, the **reasonable expectation of privacy** standard, protect such personal data being collected during the pandemic?
- Enjoy the benefits and convenience while privacy is being protected



THANK YOU!

Big Data Security

Tingyi Guo

What is Big Data Security?

- Big data security: Tools and measures which are used to guard both data and analytics processes.
- Main purpose: Provide protection against the attacks, thefts, and other malicious activities that could harm the valuable data.

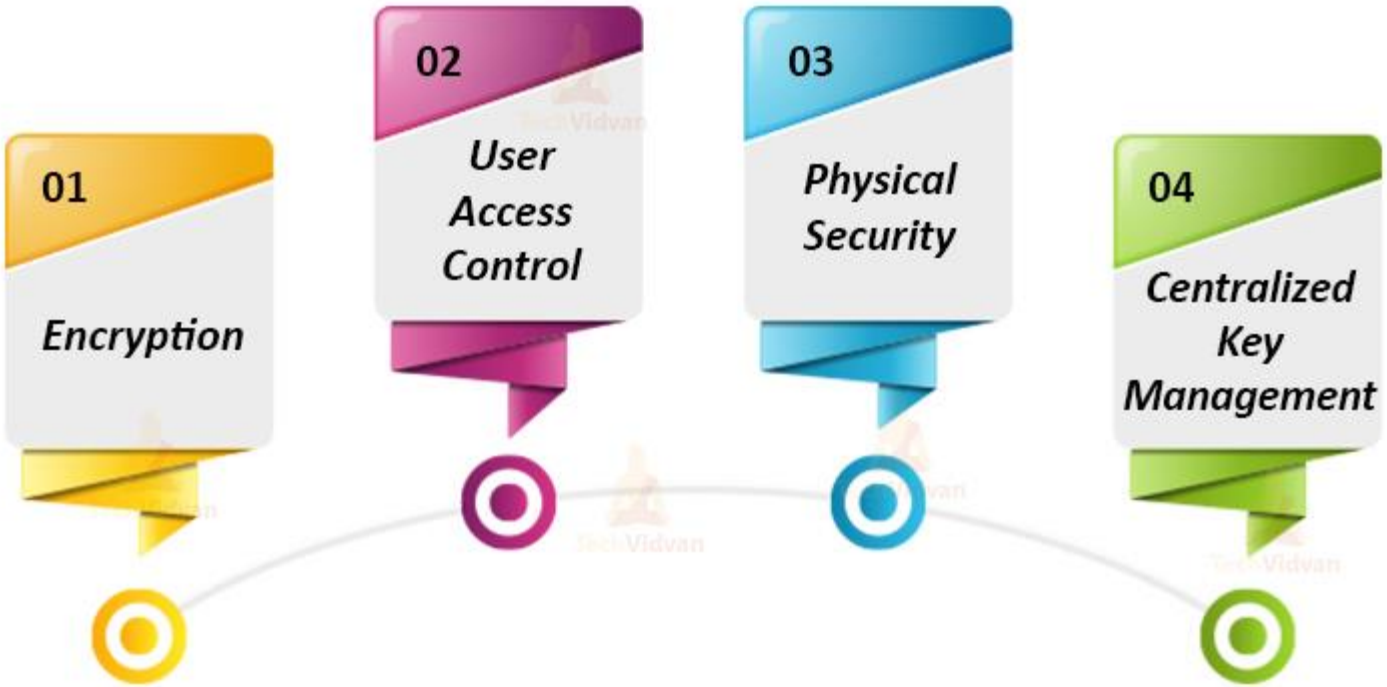
Big Data Security Challenge

- Distributed Data: Most big data frameworks distribute data processing tasks throughout many systems and ignore the data security problem. This undoubtedly increases the cost of data protection.
- Non-Relational Database: NoSQL databases favor performance and flexibility over security. If organizations want to use NoSQL, they still need additional security measures.
- Endpoint Vulnerabilities: Cybercriminals can manipulate data on the endpoint devices and transmit the false data to data lakes. Need to valid the authenticity of endpoint devices to protect data security.

Big Data Challenge

- Data Mining Results: Data often contains personal and financial information. For this reason, companies and organizations need to add extra security layers to protect against external and internal threats
- Granular Access Controls: Companies sometimes need restrict access to sensitive data and provide grant granular access to users. However, in big data, such granular access is difficult to grant and control simply because big data technologies aren't initially designed to do so. Generally, as a way out, the parts of needed data sets, that users have the right to see, are copied to a separate big data warehouse and provided to particular user groups as a new 'whole'. Granular access issues can adversely affect the system's performance and maintenance because the number of data warehouses may grow exponentially.
- Reference: https://www.scnsoft.com/blog/big-data-security-challenges

Big Data Security Technologies



Physical Security:

- The most primitive method. It is generally built in when you deploy the Big data platform in your own center. They can deny data center access to strangers or suspicious visitors. Video surveillance and security logs are also used for the same purpose.
- Problem: No limitation to Internet, hackers cannot be prevented.

User Access Control:

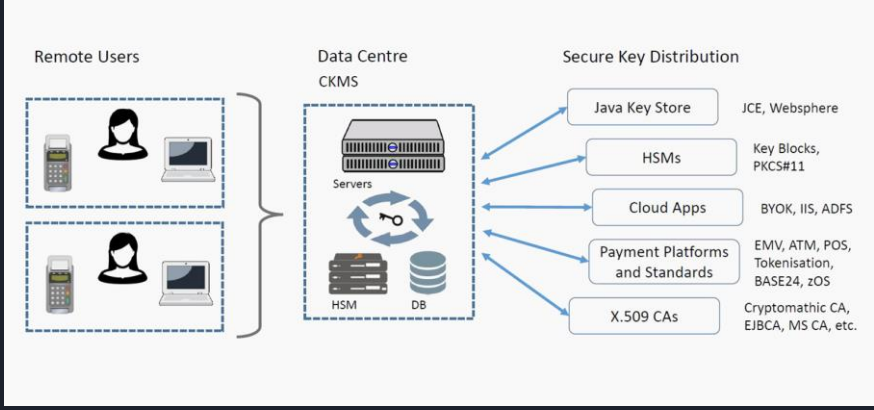
- Set access control to users. Different users have different access to data.
- It is the most basic network security tool.
- Problem:
  - it involves high management overhead, this can be dangerous at the network level and not good for the Big data platforms.
  - For institutions such as hospitals and banks, they hold a lot of sensitive data, and it is not easy to do user access control

Encryption:

- Encryption of data is generally done to secure a massive volume of data, different types of data. It can be user-generated or machine-generated code. Encryption tools along with different analytics toolsets format or code the data.
- Two types:
  - Symmetric Encryption: the same key is used to both encrypt and decrypt the data.
    - DES(Data Encryption Standard), 3DES(triple DES)
  - Asymmetric Encryption: A public key is used to encrypt the data, while a corresponding private key is required to decrypt the data.
    - RSA, SHA(Security Hash Algorithm)
- Problem : Encrypting data is the customer's responsibility. Increase the workload of companies and organization.
- Reference: https://www.mcafee.com/enterprise/en-us/security-awareness/data-protection/what-is-data-encryption.html

Centralized Key Management:

- Use a single point to secure key management, policies and access audit logs.
- It is one of the best security practices for many years. It is easily applied in Big data environments, especially on those having wide geographical distribution.
- Benefits:
  - Automatic key updates and distribution to any end-point
  - Reduces the risk of human errors
  - System-wide key control manages any key type and format
  - Offers High availability and scalability
  - Reduces costs
  - Simple backup and recovery
- Reference: https://www.cryptomathic.com/news-events/blog/the-benefits-of-an-automated-and-centralized-key-management-system



Benefits of Big Data Security:

- Boosts the security of non-relation data
- Help to implement endpoint security
- Enhances communication and availability of information
- Increases systems efficiency and robustness
- Avoid unauthorized access to protect the performance and security of the organization.
- Practice real-time security monitoring and compliance.

Cloud Security Monitoring: Big Data Security Use Case

- Concept: Cloud security monitoring supervises virtual and physical servers to continuously assess and measure data, application, or infrastructure behaviors for potential security threats. This assures that the cloud infrastructure and platform function optimally while minimizing the risk of costly data breaches.
- Why we need cloud security monitoring?
  - More and more sensitive data is stored on the cloud
  - Lack of standards about monitoring and reporting.
- Benefits: identify potential security vulnerabilities, prevent loss of business
- Reference: https://digitalguardian.com/blog/what-cloud-security-monitoring

Thank you!

Question?

Secure Inference

DSCI 529
Kung-Hsiang (Steeve) Huang

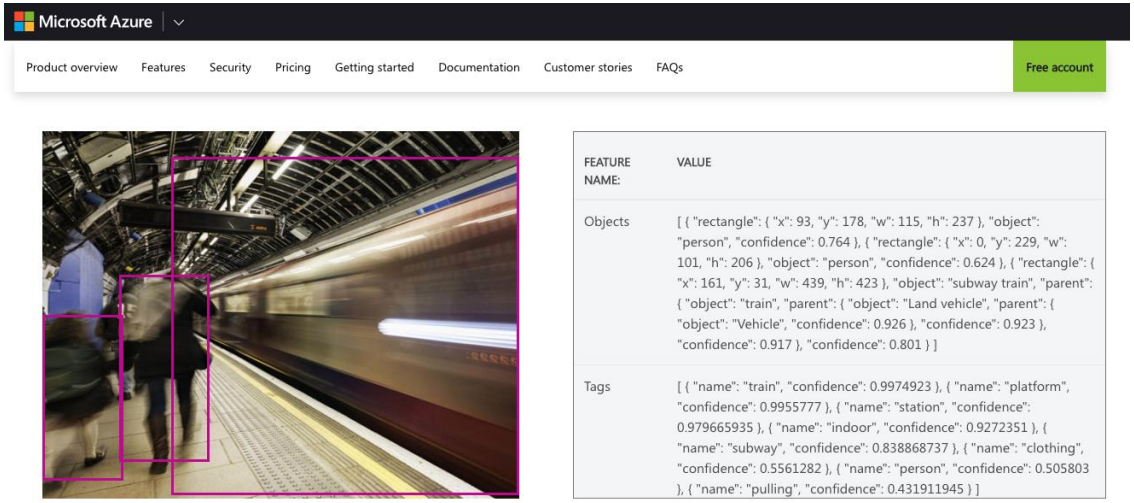
SAAS

SAAS

- Software as a service.

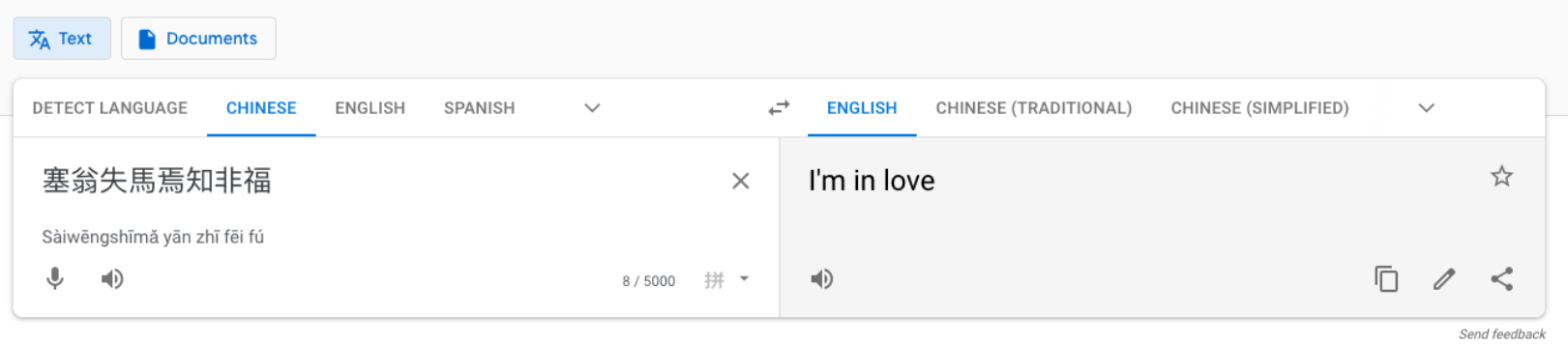
SAAS

- Software as a service.
- E.g. Azure Computer Vision, Google Translate, Rosetta.ai



SAAS

- Software as a service.
- E.g. Azure Computer Vision, Google Translate, Rosetta.ai



SAAS

- Software as a service.
- E.g. Azure Computer Vision, Google Translate, Rosetta.ai

Why do we care?

Sensitive data

- End users' private information.
- Service providers' machine learning model.

Why do we care?

Sensitive data

- End users' private information.
- Service providers' machine learning model.

Solution: secure inference!!

Secure inference

A mechanism that allows end users to query machine learning models (hosted on the cloud by service providers) without:

1. Service providers knowing the raw input data.
2. End user knowing the parameters of the model.

Machine Learning Backgrounds

Supervised learning:

- Train a model f_θ s.t. $y = f_\theta(x) \cong y^* \; \forall \; y \in Y$.
- Inference: $y = f_\theta(x)$

Problem Setup

A: End user who owns sensitive data x.

B: Service provider who owns the model f_θ.

Θ: the model parameters/ weights.

A wants to compute $f_\theta(x)$ without:

1. A knowing θ
2. B knowing x

Common secure inference approaches

1. 2-party Computation
2. (Fully) Homomorphic Encryption

2-party Computation

2 parties jointly compute a function h(m, n) over exclusively private values m and n.

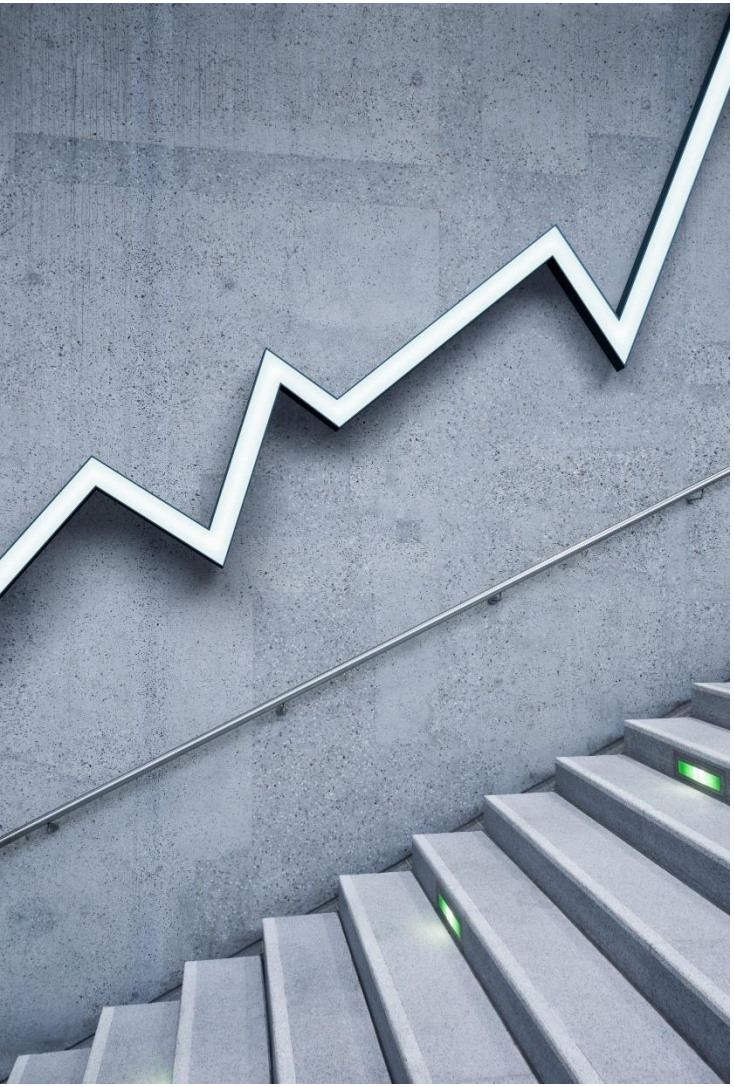
Garbled circuit:

- Both parties encrypt their shares using the same encryption method determined by A.
- B computes the encoded outputs based on the encoded x and encoded θ.
- A decodes the encoded outputs.

2-party Computation

1. f is represented as a set of boolean circuits (K).
2. A computes the encoded input enc(x) and a decoding table.
3. A computes garbled tables for each gate in K.
4. A sends enc(x) and the garbled tables to B.
5. B obtain enc(θ) using oblivious transfer from A.
6. B computes enc(f_θ(x)) using enc(x), enc(θ) and the garbled tables.
7. B sends enc(f_θ(x)) to A, and A obtains f_θ(x) using the decoding table.

Thank you!



DSCI 529: Risk of Eavesdropping and data breaches on Big Data

SUPREET KAUR RANDHAWA

USC ID: 2399553548

WHAT IS BIG DATA?

The term “big data” refers to data that is so large, fast or complex that it’s difficult or impossible to process using traditional methods.
Analyst Doug Laney articulated the definition of big data as the three V’s:

Volume: Organizations collect data from a variety of sources, including business transactions, smart (IoT) devices, industrial equipment, videos, social media and more.

Velocity: With the growth in the Internet of Things, data streams into businesses at an unprecedented speed and must be handled in a timely manner. RFID tags, sensors and smart meters are driving the need to deal with these torrents of data in near-real time.

Variety: Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, emails, videos, audios, stock ticker data and financial transactions to semi-structured that contain both the forms of data.

What is a Data Breach and Eavesdropping ?

The US Department of Health and Human Services defines a **data breach** as "a security incident in which sensitive, protected or confidential data is copied, transmitted, viewed, stolen or used by an individual unauthorized to do so."

The term '**eavesdropping**' is used to refer to the interception of communication between two parties by a malicious third party.

How can Big Data can be misused?

Large-scale cloud infrastructures, diversity of data sources and formats, the streaming nature of data acquisition and high-volume inter-cloud migration all create unique security vulnerabilities.

- The moral implications of improper profiling
- Discrimination
- Data, system and collection errors
- Data breaches or cyber attacks
- Political or social manipulation

How can Big Data be Breached?

Data breaches occur when information is accessed without authorization. In most cases, data breaches are the result of out-of-date software, weak passwords, and targeted malware attacks. Unfortunately, they can cost an organization a damaged reputation and a great deal of money. Few other reasons include:

- remote servers
- lax security

Three step approach to breach into big data:

1. Research
2. Attack(Network/Social attack)
3. Exfiltration

Eavesdropping affect privacy and security of big data

- What happened to the users of Amazon Alexa and Google home?
- However, if business fell victim to eavesdropping, it could experience the one or all following implications-

1. **Loss of privacy:** While eavesdropping, the attackers will absorb vital business information, ideas and conversations being exchanged within the organization, thereby affecting its privacy

2. **Identity theft:** The attacker who has been eavesdropping on a conversation between two employees has easy access to their credentials; and can steal all the important information.

3. **Financial loss:** Once the cyber attacker has vital business information, it can be used to full advantage by exposing the data or selling it to the competitors; the attackers will earn, and the organization will lose in millions.

Examples of data breaches on big data: Business

- Timehop (July 2018)
  Mobile App Vendor
  The data of the start-up's 21 million users was exposed for around 2 hours due to a network intrusion on 4 July.

- Reddit (June 2018)
  Content Aggregator
  Hackers gained access to an old database of users (the exact number of those affected has not been revealed) on 19 June.

- Dixons Carphone (June 2018)
  Retailer
  An estimated 10 million customers could be affected by the hacking attack on its network sometime last year. The compromised data may include personal information like names, addresses, and email addresses. Some 5.9 million payment card records (nearly all of which are protected by the chip-and-PIN system though) may have been accessed as well.

- Ashley Madison (July 2015)
  Social Media Website
  Hacktivists stole and dumped 10GB worth of data on the Deep Web. This included the account details and personally identifiable information (PII) of some 32 million users, as well as credit card transactions.

- Target (January 2014)
  Retailer
  Hackers penetrated the vendor's network and infected all of its point-of-sale (PoS) machines. They were able to expose nearly 40 million debit and credit cards to fraud. The information stolen included PINs, names, and banking information.

- Equifax (July 2017)
  Information Solutions Provider
  The major cybersecurity incident affected 143 million consumers in the U.S. Initially discovered on 29 July, the breach revealed the names, Social Security numbers, birth dates, and addresses of almost half of the total U.S. population. With investments in 23 other countries worldwide, around 400,000 U.K. customers were also reportedly affected. Final findings revealed a total of 145.5 million exposed records.

Examples: Medical/Healthcare

- SingHealth (July 2018)
  Medical/Healthcare Service Provider
  The nonmedical personal data of 1.5 million patients was reportedly accessed and copied, including their national identification number, address, and date of birth as part of the attack. The stolen data also included the outpatient medical data of 160,000 patients.

- Hong Kong Department of Health (July 2018)
  Federal Agency
  The government agency was hit by a ransomware attack that rendered its systems inaccessible for two weeks starting 15 July.

- Anthem (May 2015)
  Medical/Healthcare Service Provider
  An attack that started in April 2014 resulted in the theft of more than 80 million records of current and former customers. The data stolen included names, birthdays, social IDs, email addresses, and employment information

Examples: Government/Military

- U.K. military contractor (May 2017)

  Military Contractor

  Sensitive data from a military contractor was extracted by a targeted attack group from the military contractor's network using a backdoor identified as RoyalDNS.

- U.S. OPM (April 2015)

  Federal Agency

  Hackers gained access to more than 18 million federal employee records, including Social Security numbers, job assignments, and training details.

Examples: Banking/Credit/Financial

- Deloitte (October/November 2016)

  Accountancy Firm

  The firm was targeted by a sophisticated hack that compromised the confidential emails and plans of some of its blue-chip clients. The attack was discovered in March 2017 though findings revealed though the hack may have been launched as early as October or November 2016.

- JP Morgan Chase & Co. (October 2014)

  Credit Service Provider

  The data of an estimated 76 million households and 7 million small businesses was compromised. The information included names, addresses, phone numbers, email addresses, and others.

Examples: Educational

- University of Maryland (March 2014)

  Educational Institution

  More than 300,000 student, faculty, and staff records going as far back as 1998 were compromised though no financial, medical, or academic information was included. The stolen data included names, birth dates, university ID numbers, and Social Security numbers.

- University of Greenwich (2004)

  Educational Institution

  The university was fined £120,000 for exposing the personal data of students, including names, addresses, dates of birth, signatures, and in some cases even medical information, on a microsite that was left unsecured since 2004.

Why??

- To make money by duplicating credit cards
- Using personal information for fraud, identity theft, and even blackmail.
- Bragging Rights
- Revenge / to inflict damage
- Terrorism and Extortion
- Financial / Criminal enterprises
- Nation State motivations

Defense against breaches and eavesdropping on big data

- **Military-grade encryption:** By using a 256-bit, also known as military-grade encryption, the attacker may gather the data via eavesdropping, but the data will still be safe as it will take him around 500 billion years to decode it.
- **Spread awareness:** training and informing the employees of the organization about cybersecurity is of utmost importance. The employee should have complete knowledge about eavesdropping attacks before he/she downloads an application, software or connects over a weak network.
- **Network segmentation:** Network division or segmenting helps in decongesting the network traffic, improves security and prevents unwanted connectivity.

- **Cryptography** : create building blocks that provide privacy, authentication, anonymity; "*secure multiparty computation*"
- **Wireless device detection system** : identify access points on network which shouldn't be there
- **Degaussing:** Degaussers demagnetize the hard disk drive, making it completely inoperable.
- **Bring in a Cybersecurity specialist**
- **Refer to the FINRA checklist**
- **Destroy Before Disposal**

Conclusion:

Big data is a double-edged sword for customers. It provides many benefits. But it can be a serious problem if your data is compromised.

Big data undoubtedly has its plus points for any business needing to process large volumes of data on a regular basis. But, in spite of this, the increasingly sophisticated techniques utilized by cybercriminals are becoming ever-more difficult to combat. Taking this into account, it's safe to say that optimal cybersecurity practices need to be put in place by businesses. Otherwise, instances of large data breaches will only continue to take their toll.

BIG DATA AND DISCRIMINATION

LINGYU GE

DATA-DRIVEN DISCRIMINATION

- Big data analytics may lead to discrimination of certain groups of people.

CHALLENGES

- Unintended data bias
- Intended data bias
- Digital Divide

UNINTENDED DATA BIAS

- Biased data.
- Sample size.
- Bad data quality.
- High dimensionality.
- Weak correlation of parameters.
- Algorithms.

INTENDED DATA BIAS

- Personal data of individuals or groups can be aggregated to detailed profiles.
- Personalized services or advertisements are exclusive to certain groups.
- Wrong statistical analysis.

DIGITAL DIVIDE

- Different levels of access and skill in technology.
- Second-level Digital Divide
  - Different degrees of skill, time, knowledge, and usage possibilities.
  - Social status.

OPPORTUNITIES

- Big data helps to decrease social inequity.
- Personalized or individualized services.
- MobiDig - services in rural areas in order to increase social inclusion.

THANK YOU

Upcoming Presentations: Internet of Things – March 5th

- Pratheek Athreya
- Arzu Karaer
- Bolong Pan
- Danielle Sim
- Haipeng Yu
- Xihao Zhou
- Jinyu Zhao
- Pu Zhao
- Junbo Sheng
  - This group will have one hour 30 minutes to present.

Upcoming Presentations
Social Media – March 19th

- Addison Allred
- Yixiang Cao
- Lei Gao
- Brianna Hefferin
- Mingliao Xu
- Shengwang Zhang
- Zixin Zheng
- Hehan Xie
- Chengyuan Zhou

- This group will have 90 minutes to present.

Upcomming Presentations
Pandemic/Govt Data Use – March 26th

Pandemic (40 minutes)

- Yuemeng Gao
- Tanmay Ghai – Privacy Preserving Contact Tracing
- Yi Lin – Big Data in China related to the COVID Pandemic
- Gan Xin – Health QR Code in China

Other government use of data (50 min)

- Yi Jin – How US and China collect and use personal data
- Congrui Li
- Michelle Muldoon – Law Enforcement and Privacy w.r.t. Data Brokers
- Griffin Weinhold – Decentralized Search and Search Histories in Court
- Xihao Zhou – Use of Data by Governments
- Jinglun Chen – Use of location data

Upcoming Presentations
Privacy&Security Regulation – April 2nd

- Jia Yu Lee
- Yansong Wang
- Kaifan Lu – Assessing China's Cybersecurity Law

- 30 minutes for this group to present

Upcoming Presentations – April 9th Free Expression - Disinformation

- Adriana Nana – Deep Fakes and Privacy
- Resherle Verna – Should Social Media company's have right of censorship

  – This group will have 20 minutes to present.

Upcoming Presentations Healthcare – Date T.B.D.

- Vartan Batmazyan
- Tingyi Guo
- Phuong Ngo
- Sharad Sharma (DNA Databases)
- Ye Zheng - Fitness apps

- This group will have 50 minutes to present.

Upcoming Presentations Privacy and Finance – April 16th

- Jonathan De Leon – Privacy in Finance
- Sidong Wang – History and Technologies for Cryptocurrencies
- Saurabh Jain – Privacy of Credit Card/Payment card information
- Yifeng Shi -Financial value of data gathered through free services

- 40 minutes

Secure Communication – Privacy Preserving Technologies – April 16th

- Zihuan Ran – Privacy Preserving Database Technologies
- Aziza Saulebay – 5G and Data Privacy
- Carol Varkey – Messaging Application Privacy
- Francisco Ventura – Encryption Technologies and implications
- Zixin Zheng – Privacy Preserving Technologies

- 50 minutes

Upcoming Presentations
Other Security Topics – Date TBD

- Yo-Shuan Liu – User experience and Multi-Factor Authentication
- Philana Williams – Security for Web App Development
- Haonan Xu – Privacy issues in Cloud Computing
- Pratishtha Singh – Card privacy Concerns in India

Current Event Discussion

- http://csclass.info/USC/INF529/s21-lec6-ce.html

Privacy and Big Data

Required reading:

Big Data and the Future of Privacy
Epic.org

Will Democracy Survive Big Data and Artificial Intelligence?
Scientific American – 25 February 2017

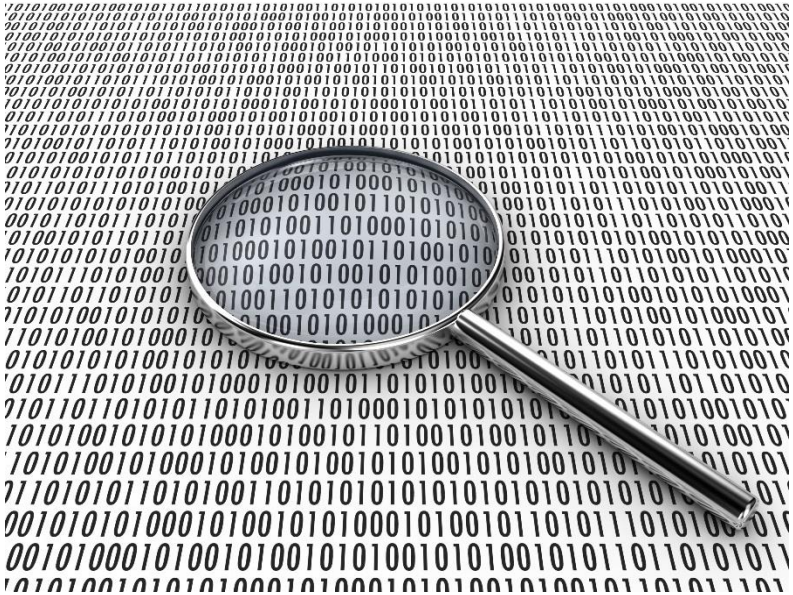
"Muslim registries", Big Data and Human Rights
Amnesty International – 27 February 2017.

What is Big Data

Processing of large and complex data sets.

- **Often with multiple structures.**
- **Data is mined to find trends, relationships, and correlations.**

- **Danger**
  - **By combining information from multiple sources more can be inferred than specifically disclosed.**



Inferences are imprecise

- **The algorithms learn discrimination**

What is Big Data

- Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.
- Use of Big Data promotes
  - cost reductions
  - time reductions
  - new product development and optimized offerings
  - smart decision making

What Data Mining Can Tell Us

Quite a lot, and acting on that information can cause problems.

FEB 16, 2012 @ 11:02 AM 3,122,087 VIEWS

Forbes

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill, FORBES STAFF

Welcome to The Not-So Private Parts where technology & privacy collide FULL BIO

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target TGT +0.21%, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Charles Duhigg outlines in the New York Times how Target tries to hook parents-to-be at that crucial moment before they turn into rampant -- and loyal -- buyers of all things pastel, plastic, and miniature. He talked to Target

Who uses it

- Banking
  - finding new and innovative ways to manage big data
  - understand customers and boost their satisfaction
  - minimize risk and fraud
- Education
  - identify at-risk students
  - make sure students are making adequate progress
  - implement a better system for evaluation and support of teachers and principals
- Government
  - managing utilities
  - running agencies
  - dealing with traffic congestion
  - preventing crime
- Health Care
  - patient records
  - treatment plans
  - Prescription information
- Manufacturing
  - solve problems faster
  - make more agile business decisions
- Retail
  - the best way to market to customers
  - the most effective way to handle transactions
  - the most strategic way to bring back lapsed business

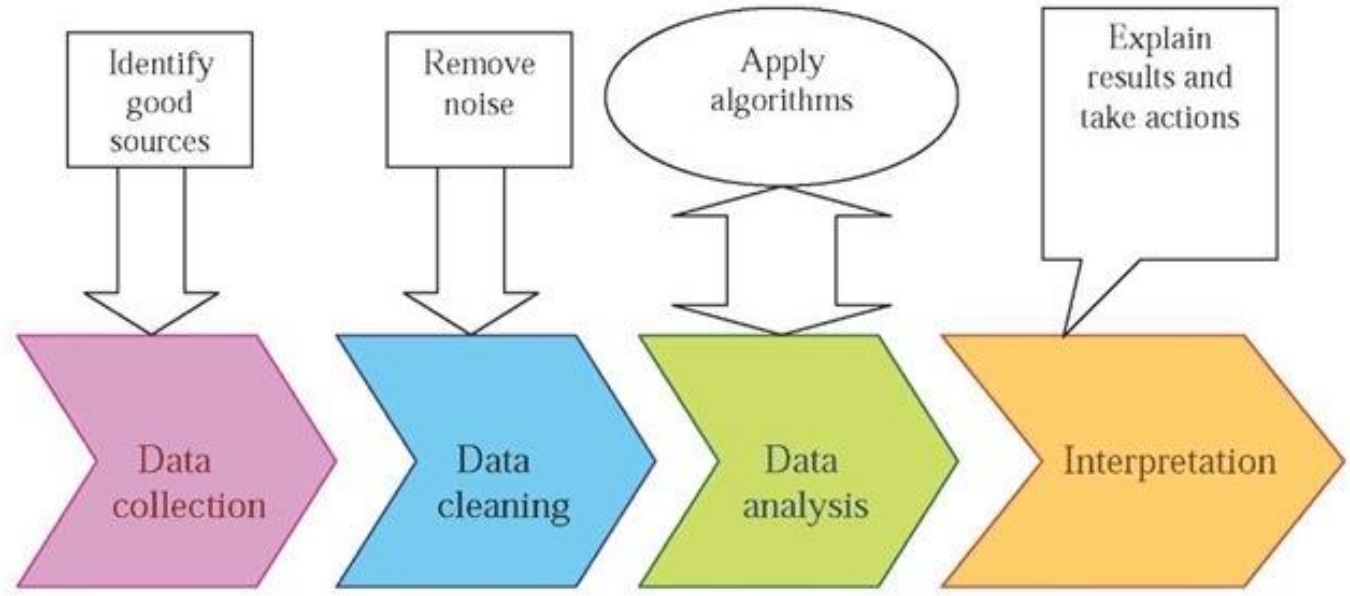
Who uses it

- Case study:

One classic example of the success of big data is the success of House of Cards. Netflix, the distributor of this TV show, collects data from its users and analyze those data. For example, they analyze what kind of show or movie did the users watch, share, and subscribe, therefore make inference about which type of show, which director and actors will be preferred by the users. That's how the director and actors of house of cards are decided. Then, they use algorithm to rank and recommend shows to the users, and most of the time, users will like it.

Steps of Data Mining

- The process of analyzing data from different perspectives and summarizing it into useful information.



Privacy Consideration

- Are users concerned?
  - According to a survey in 2017, about 49% of consumers are less willing to share their personal information. Many consumers are now aware of the dangers of sharing their personal information and the security issues involved by consenting to the sharing of their personal information online.

Privacy Consideration

- (Big data) breaches are big (data breaches).
- The more information used in big data, the more likely it includes personal or sensitive information.
- Sources of information vary greatly, allowing multiple opportunities for exfiltration.
- The distributed processing of big data (e.g. cloud services) increases the attack surface for this data.

Some Area for Risk

- **Personal data protection**
  - Existing methods for protecting identity might be thwarted by Big Data Analysis
- **Financial and legal liabilities**
  - The data you hold may now be more sensitive because of what can be derived through big data analysis.
  - Disovery requests
- **Ethical dilemmas**
  - New ethical dilemmas are being created by the analysis of Big Data

Anonymization in big data

- Modifying personal data so individuals cannot be reidentified and no information about them can be learned

- Techniques to prevent re-identification :
  - Controlled linkability
  - Composability
  - Anonymization of dynamic/streaming data
  - Computability for large data volumes
  - Decentralized anonymization

  - (rewrite)

Privacy preservation methods

- **K anonymity**
  - **P**rocess of modifying data before it is given for data analytics so that de identification is not possible and will lead to K indistinguishable records if an attempt is made to de identify by mapping the anonymized data with external data sources.
  - K anonymity can optimized if the minimal generalization can be done without huge data loss.
    - Homogeneity attack.
    - Back ground knowledge attack.
- **L diversity**
  - L well represented values for the sensitive attribute (disease) in each equivalence class.
  - Prone to skewness attack
- **T closeness**
  - **A**n equivalence class is considered to have ‘T closeness’ if the distance between the distributions of sensitive attribute in the class is no more than a threshold and all equivalence classes have T closeness.
  - T closeness may ensure attribute disclosure but implementing T closeness may not give proper distribution of data every time.
- **Differential Privacy**
- **(rewrite)**

Privacy preservation methods

- **Randomization technique**
  - **P**rocess of adding noise to the data which is generally done by probability distribution
  - randomization on large datasets is not possible because of time complexity and data utility
- **Data distribution technique**
  - **Horizontal -** data is distributed across many sites with same attributes
  - **Vertical -** specific information is distributed across different sites under custodian of different organizations
  - Distribution of data will not ensure privacy preservation but it closely overlaps with cryptographic techniques.
- **Cryptographic techniques**
  - encrypting large scale data using conventional encryption techniques is highly difficult and must be applied only during data collection time.
  - Data utility will be less if encryption is applied during data analytics
- **Multidimensional Sensitivity Based Anonymization (MDSBA)**
  - data is split into different bags based on the probability distribution of the quasi identifiers by making use of filters in Apache Pig scripting language. Appropriate for large scale data but only when the data is at rest
  - makes it difficult to map the data with external sources to disclose any person specific information.
  - (rewrite)

Mid-Term Exam is Friday February 26th

Exam will be 100 minutes, from Noon to 1:40PM PST.

- For students in distant time zones, an alternative time will be 6PM-7:40PM PST. You MUST contact me in advance to arrange for this alternate time.
- A lecture will follow from 3PM to 4:20PM

- Format of the exam
  - The exam is open book, open note and online
  - Previous exams are posted on the class website http://ccss.usc.edu/529
    - The 2020 exams will be posted soon
  - Material to be covered will be the start of the semester through Today's lecture.
  - You are responsible for material discussed in lecture, and also the several assigned readings (including those listed in today's slides).

Mid-Term Exam Logistics

Full Instructions will be sent by Monday through email.

By 10 minutes before the time of the exam, three versions of exam (PDF,TXT,and Word) will be sent to students.

Students will complete exam by editing the exam files (the word file is preferred; the other formats are provided in case students do not have ability to use the word version).

At conclusion, exam will be uploaded through the D2L dropbox.

Students will self cerify that you completed the exam in the allotted time and neither received or provided assistance.

Mid-Term Outline of Material

- Overview of security and privacy
  - What are they, why we have neither
  - Relationship between the two
- Understanding our data in the cloud
  - What data exists and who can access it
    - Both officially and unofficially
  - What is the data used for
  - What can it be potentially used for

Mid-Term Outline of Material

- Overview of Technical Security
  - Confidentiality, Integrity, Availability
  - The role of Policy
  - Risk Management from multiple perspectives
- Mechanisms
  - Encryption/Key Management, Firewalls, Authentication, Digital Signatures, Authorization, Detection, Trusted hardware
- Attacks
  - Malicious Code
  - Social Engineering
  - Attack Life Cycle

Mid-Term Outline of Material

Identity Management and Privacy

Expectations of Privacy

Big-Data and Security and Privacy

I will ask opinions on the predominant current events with respect to how they relate to the topics above.

You will be asked to argue BOTH sides of at least one Privacy issue

Mid-term Format

One questions focused on a Sample service sector
- Description of the service

Questions for you
- Analyze the information requirements
  - And the policies to apply to preserve privacy.
  - Discuss ethical issues around that policy.
  - What are the expectations of users.
- Discuss the vulnerabilities that likely exist and how attacks might be facilitated
- Discuss technical and design measures one might use to preserve security and privacy in the system.

2019 Mid-Term

How did they get my data? (30 points)

Privacy breaches involve inappropriate access to or use of personally identifiable information. Such inappropriate access typically takes one of two forms. Either data held legitimately is disclosed through the actions of criminals that breach the security of a system, or alternatively, the holder of the information gives the data to someone that should not have access or uses the data in ways that are not authorized or collects data they shouldn't be collecting to begin with.

a) List the three primary ways that adversaries can get hold of your personally identifiable data in the systems that you use. (10 points)
[Hint, of the three different ways, two of them are probably your own fault]

b) Explain the role that malware, malicious apps, or apps that exceed their legitimate authority play in mis-use or release of our PII. (10 points) [Hint, it can play a role in any of the three ways covered in 1a, but you must explain how it does so]

c) If you were designing a system that used PII, what are some of the steps you would take to minimize the risk of inappropriate disclosure of PII to others. (10 points)

2019 Mid-Term

2) Much of the information collected about us has been collected and stored for many years. The first photograph was taken around 1827, the first video (moves) were recorded in 1888. The earliest transaction receipts (records of goods traded) go back at least as far as the Mesopotamian civilizations. Given that such data has been recorded for years, what has changed about our technology that makes things different in terms of its impact on our privacy? (10 points)

2019 Mid-Term

- The primary focus in class for our discussion on expectations of privacy was on access to our private data by our government (e.g. search and seizure, wiretaps, our encrypted data, messages, email, as well as transaction records, information from security cameras, and even D.N.A.). The discussion was very much focused on the expectations of individuals within the United States. There are equally legitimate arguments on both sides of the issue regarding what kind of access is to be permitted and what should be the conditions under which the data may be used. These arguments attempt to balance potential rights of privacy with the need of government to stop crime and protect its citizens. These arguments have been made for and against proposals in the United States, and they have been made in other countries as well, sometimes resulting in different outcomes in terms of the laws that apply.
- In this question you are to make arguments in favor of the rights of individual privacy over the need for governments to have access to significant private information for the purpose of public safety. You are ALSO to make arguments in favor of the need for government to have access to private information, even at the cost of diminishing individual privacy. I want you to make equally compelling arguments on each side of this issue, and you should provide example scenarios or real world examples that support each of the opposing arguments.
- You do not need to tell me where your personal beliefs fall in terms of these arguments. I am not grading your viewpoint. If, however, you object to arguing the opposing side as your own, then you may cast those arguments as "Others ague that", or with similar wording. (30 points)

2019 Mid-Term

You have been hired by a joint commission comprising the FTC in the United States, and investigators from the E.U. to analyze the security and privacy practices of Facebook, Google, Apple, and similar data brokers In particular, you are asked to check whether the practices of these organizations are consistent with the terms and conditions / privacy policies of the organizations, and with applicable law in Europe and the U.S.

- List some of the actions (and inaction) by these three organizations that have come under fire by regulators for demonstrating a lack of concern for the privacy of individuals. [hint, most of these items were the subject of multiple current event discussions] (10 points)

- Discuss policy, technical, and procedural recommendations that you have for these three organizations, and other organizations that process this kind of PII, that will help them to address these concerns. (10 points)

- Discuss your recommendations for elements that you feel should be part of a comprehensive U.S. privacy law in order to address the potential misuse of our PII by these and other private organizations. (10 points)

Current Event Discussion

- http://csclass.info/USC/INF529/s21-lec6-ce.html

Bias in Big Data

- Confirmation bias
  - Relying on data to confirm a certain hypothesis
- Availability heuristic/availability bias
  - Relying on only data that is readily available or recent.
- Selection bias
  - Sample not representative of the general population
- Confounding variables
  - Relationship between variables is only true when combined with a third (overlooked) variable.

Examples of big data bias

"Predictive policing" in Chicago

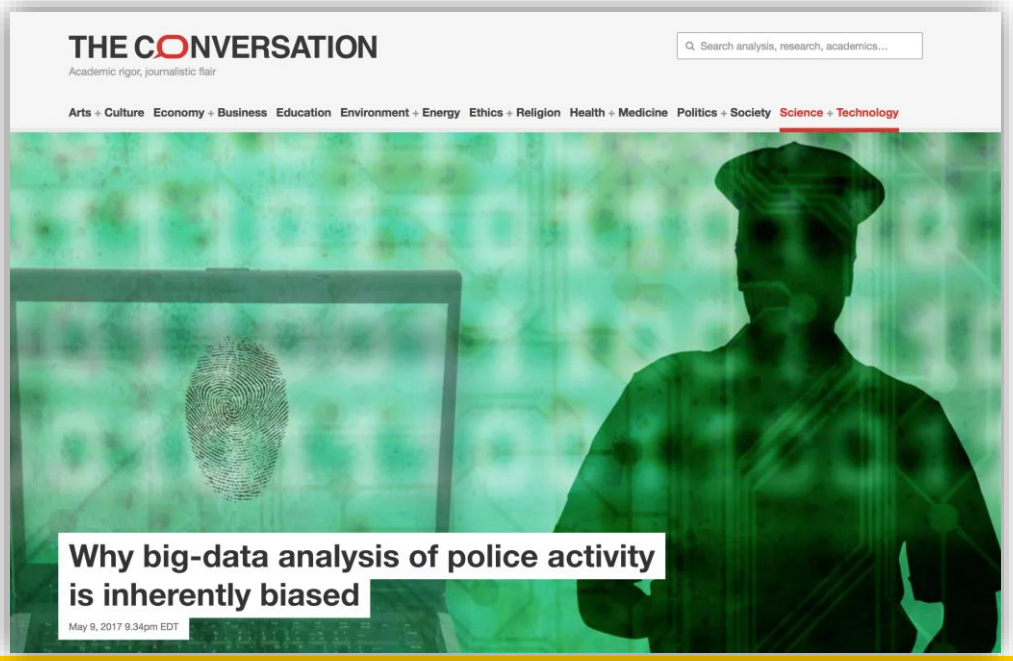
"The Chicago police will use data and computer analysis to identify neighborhoods that are more likely to experience violent crime, assigning additional police patrols in those areas. In addition, **the software will identify individual people who are expected to become – but have yet to be – victims or perpetrators of violent crimes**. Officers may even be assigned to visit those people to warn them against committing a violent crime."

Why big-data analysis of police activity is inherently biased, *The Conversation,* May 9, 2017

Examples of big data bias

Why big-data analysis of police activity is inherently biased, *The Conversation,* May 9, 2017

"Neighborhoods with lots of police calls aren't necessarily the same places the most crime is happening. They are, rather, where the most police attention is – though where that attention focuses can often be biased by gender and racial factors."



Can algorithms illegally discriminate

CNBC – and Whitehouse report

But when it comes to systems that help make such decisions, the methods applied may not always seem fair and just to some, according to a panel of social researchers who study the impact of big data on public and society.

The panel that included a mix of policy researchers, technologists, and journalists, discussed ways in which big data—while enhancing our ability to make evidence-based decisions—does so by inadvertently setting rules and processes that may be inherently biased and discriminatory.

The rules, in this case, are algorithms, a set of mathematical procedures coded to achieve a particular goal. Critics argue these algorithms may perpetuate biases and reinforce built-in assumptions.

Also

http://www.nextgov.com/big-data/2017/02/cfpb-wants-know-how-alternative-data-changes-credit-scores/135695/

Critics allege big data can be discriminatory, but is it really bias?

Pradip Sigdyal | @PSigdyal
Sunday, 8 May 2016 | 4:00 PM ET

CNBC

Getty | 187131740

Big data is increasingly viewed as a strategic asset that can transform organizations through its use of powerful predictive technologies.

But when it comes to systems that help make such decisions, the methods applied may not always seem fair and just to some, according to a panel of social researchers who study the impact of big data on public and society.