

Data Security & Privacy

CIS 545

Logistics and Intro

Birhanu Eshete
birhanu@umich.edu


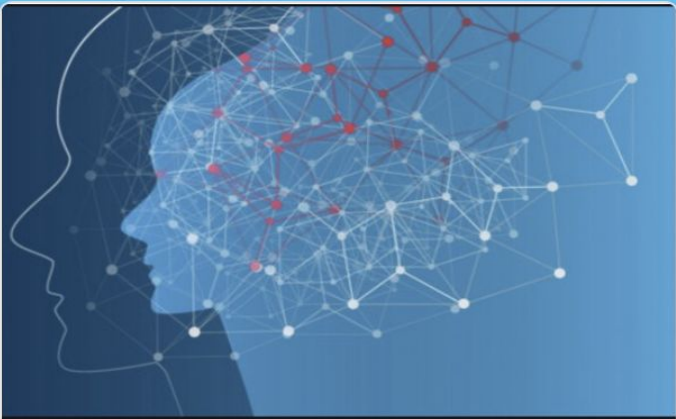



Lecture Goals

- ▶ Course objectives
- ▶ Topics overview
- ▶ Expected work and grading
- ▶ Course policies
- ▶ Introduction to data security & privacy

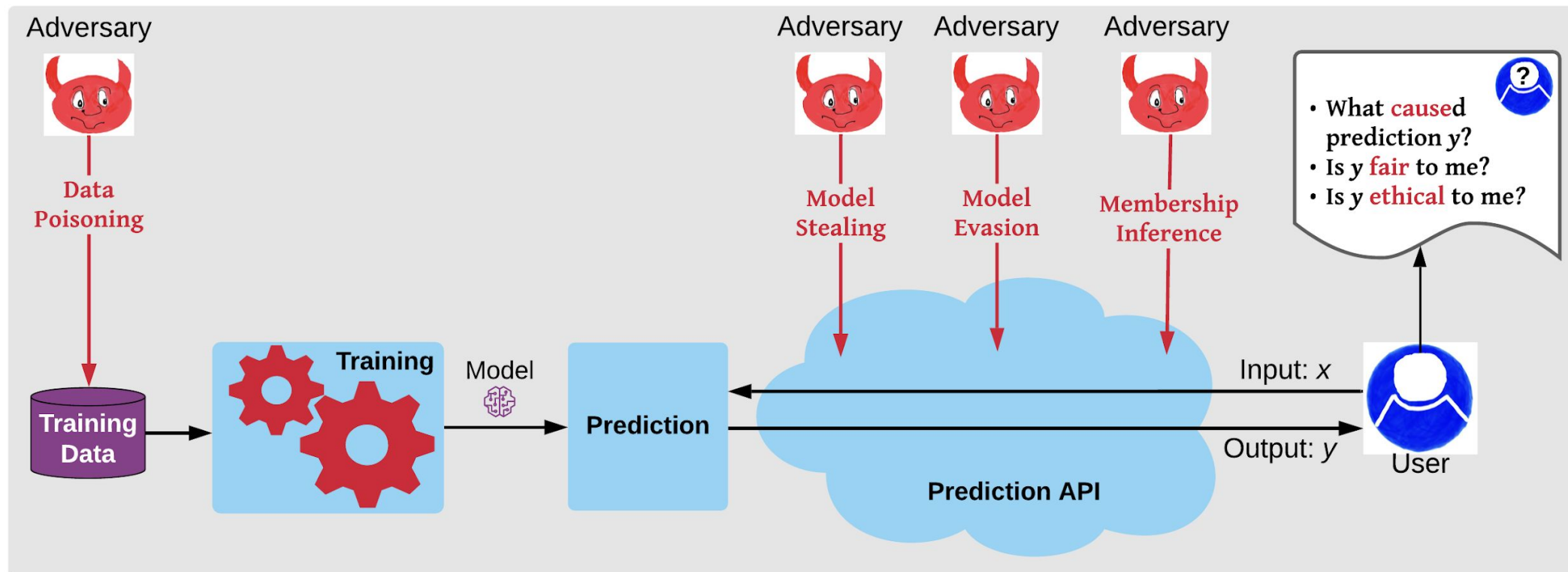
Research in my Lab

- Prof. Birhanu Eshete
- Email: birhanu@umich.edu
- Office: 229 CIS

		
CYBER THREAT INTELLIGENCE characterization, measurement, and forensics.	TRUSTWORTHY MACHINE LEARNING security, privacy, transparency, and fairness.	CYBERCRIME ANALYSIS reconstruction, measurement, and defense.

DSPLab: <http://www-personal.umd.umich.edu/~birhanu/dsplab>

Currently Active Research



- [Abderrehmen Amich, Birhanu Eshete, Vinod Yegneswaran, Nguyen Phong Hoang. DeResistor: Toward Detection-Resistant Probing for Evasion of Internet Censorship.](#) Proceedings of the 32nd USENIX Security Symposium (USENIX SEC'23).
- [Ismat Jarin, Birhanu Eshete. MIAShield: Defending Membership Inference Attacks via Preemptive Exclusion of Members.](#) Proceedings of the 23rd Privacy Enhancing Technologies Symposium (ACM PETS'23).
- [Abderrahmen Amich, Birhanu Eshete. EG-Booster: Explanation-Guided Booster of ML Evasion Attacks.](#) Proceedings of the 12th ACM Conference on Data and Application Security and Privacy (ACM CODASPY'22).
- [Abderrahmen Amich, Birhanu Eshete. Morphence: Moving Target Defense Against Adversarial Examples.](#) Proceedings of the 37th Annual Computer Security Applications Conference (ACM ACSAC'21).
- [Birhanu Eshete. Making Machine Learning Trustworthy.](#) Science, Vol. 373, Issue 6556, pp. 743-744, American Association for the Advancement of Science.
- [Ismat Jarin, Birhanu Eshete. PRICURE: Privacy-Preserving Collaborative Inference in a Multi-Party Setting.](#) Proceedings of the 7th International Workshop on Security and Privacy Analytics (IWSPA), co-located with ACM CODASPY'21.

Why this Course?

- ▶ Data breaches & privacy violations on the rise
- ▶ No major sector we care about is left out
- ▶ Current trend: corporations and governments are increasingly harvesting **private** information
- ▶ Promise: "personalized services" (search, ads)
- ▶ [Current trend] + [poor default privacy settings] + [lack of awareness] => **security & privacy in jeopardy!**

Goals (Scope) of the Course

- ▶ Fundamentals of security & privacy
- ▶ Techniques for ensuring data security
- ▶ Techniques for preserving privacy
- ▶ Data protection regulations
- ▶ **Not about:** network security, systems security, software security

The Big Picture



- ▶ **Everybody wants to harvest data:** systems, platforms, corporations, governments
- ▶ **The problem:** data holds sensitive & private personal details
- ▶ **After completing this course:**
 - know-how & exposure to major data security & privacy issues
 - techniques to ensure data security and privacy (individual, enterprise)

Topics Covered

- ▶ Fundamentals
- ▶ Security Focus
- ▶ Privacy Focus
- ▶ Ethics and Regulation Focus

The Fundamentals

- ▶ **Data Security and Privacy Fundamentals:** data, breaches, data security, data privacy, security vs. privacy vs. anonymity
- ▶ **Security and Privacy Incidents:** cause, impact, high profile data breaches & privacy invasions, ...
- ▶ **Security and Privacy Implications:** social networks, online stores, sensitive domains (e.g., healthcare, finance), paradigms (e.g., cloud, IoT, big-data analytics)
- ▶ **Threats to Data Security and Privacy:** malware, spyware, ransomware, botnets

The Security Focus

- ▶ **Access Control Mechanisms:** access control lists, capability lists, discretionary, mandatory, role-based, attribute-based
- ▶ **Security Policies:** confidentiality, integrity, hybrid
- ▶ **Cryptography for data security:** symmetric key crypto, public key crypto, signatures, hashes, ...

The Privacy Focus

- ▶ **Privacy Enhancing Technologies:** anonymous communication, proxies, relays, servers, ...
- ▶ **Anonymization/De-identification Techniques:** k -anonymity, l -diversity, t -closeness
- ▶ **Privacy-preserving Data Analytics (mining, ML, ...):** differential privacy, secure multiparty computation, homomorphic encryption, ...

The Regulations Focus

- ▶ **General:**

- General Data Protection Regulation (GDPR)

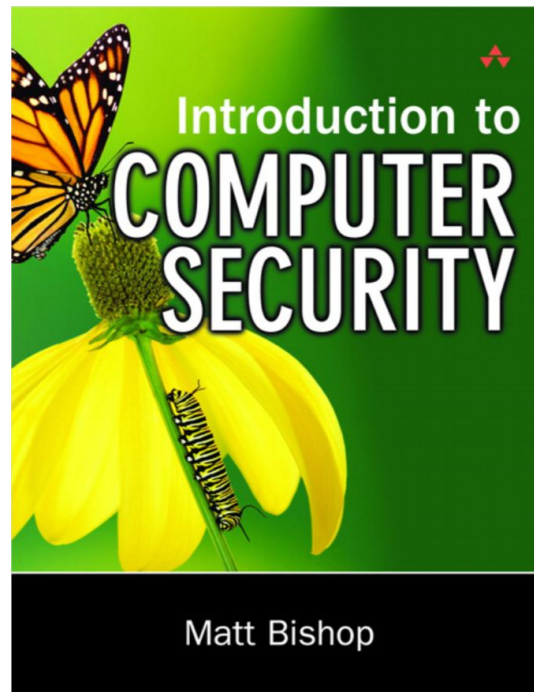
- ▶ **Healthcare:**

- Health Insurance Portability and Accountability Act (HIPAA)

- ▶ **Education:**

- Family Education Right and Privacy Act (FERPA)

Reference Materials



► Book(s):

- “Introduction to Computer Security” — Matt Bishop
- “Big Data Security and Privacy Handbook” — Cloud Security Alliance



► Research Papers:

- Will be made available as we go

Work Load & Grading

Homeworks (4 or 5)	50%
Course Project	40%
Final Exam	10%
Total	100%

Range	Letter Grade
[93,100]	A
[90,92]	A-
[87,89]	B+
[83,86]	B
[80,82]	B-
[77,79]	C+
[73,76]	C
[70,72]	C-
[67,69]	D+
[60,66]	D
[0, 59]	E

- ▶ Curving: if distribution is odd
- ▶ Fractions will be rounded

List of project topics will be provided, but your own (convincing) proposals are highly appreciated!

Course Policy

- ▶ List your **collaborators** for every homework and cite all references.
- ▶ Don't **copy** verbatim from others' work without citation.
- ▶ In **team projects**, your individual efforts will be evaluated.
- ▶ **Missed exams**: make-up exam for justified absence.
- ▶ **Late assignments**: 5% deduction for each day that passes.
- ▶ **Deadline extensions**: may be permitted for illness, family emergency, and travel.

Ethics and the Law

- ▶ We cover both offensive and defensive techniques
- ▶ **Goal:** understand capabilities and motivations of attackers to build better defense
- ▶ Do not, under any circumstance, use the offensive insights and skills you gain in this class for attacking individuals, organizations, or nations!

Introduction to Data Security & Privacy

Terminology

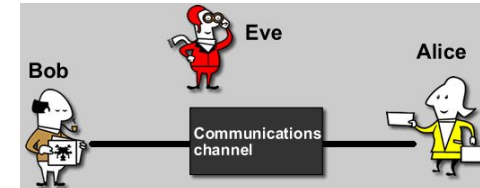
- ▶ Confidentiality
- ▶ Integrity
- ▶ Availability
- ▶ Authentication
- ▶ Authorization
- ▶ Auditing/Accountability
- ▶ Non-repudiation
- ▶ Security
- ▶ Privacy
- ▶ Anonymity
- ▶ Cyber attack
- ▶ Data Breach
- ▶ Personally Identifiable Information

Classic Security Characters

- ▶ Alice & Bob – “good guys” trying to get some useful work done
- ▶ Eve – a “passive” eavesdropper (intercept)
- ▶ Mallory – an “active” eavesdropper (intercept and modify)
- ▶ Trent – trusted by Alice & Bob (“Trusted Third Party”)



Ideal Security for Alice & Bob



- ▶ don't want their communication to be known by Eve
 - ▶ don't want Mallory to temper with the messages they exchange or data they store
 - ▶ need to use the communication channel whenever they need to
 - ▶ need to verify the identity of one another before exchanging something critical/sensitive
 - ▶ have specific permission (e.g. writing) on certain resources (e.g. files)
 - ▶ Trent keeps track of all the transactions between Alice and Bob and they both trust him
 - ▶ don't want one party deny the fact that it communicated with the other
- ▶ Integrity
 - ▶ Authorization
 - ▶ Non-Repudiation
 - ▶ Availability
 - ▶ Authentication
 - ▶ Confidentiality
 - ▶ Accountability/Auditing

Personally Identifiable Information

- ▶ **PII**: information that can be used on its own or with other information to identify, contact, or locate a single person, or to identify an individual in context
- ▶ **Strictly PII (NIST)**: home address, national IDs (e.g., SSN), passport number, vehicle registration plate number, driver's license number, fingerprints, handwriting, credit card numbers, digital identity, date of birth, birthplace, genetic information, ...
- ▶ **Potentially PII** (may be combined with other personal information to identify an individual): e.g., first or last name (if common), country, state, post-code or city of residence, age gender or race, ...

Data Security

- ▶ **Operational definition:** protecting data **at rest** or **in transit** from **destructive forces** or from **unwanted actions** of unauthorized users
- ▶ **Alternative definition:** guarantee **confidentiality**, **availability**, and **integrity** of data (ensure data isn't being used or accessed by unauthorized individuals or parties)

Let's Unpack the Definition

- ▶ **Unwanted actions** (e.g., cyberattacks, data breaches)
- ▶ **Cyber attacks:** offensive maneuvers that target cyber-infrastructures and services
 - could be performed by individuals, groups, nation-states
 - Examples: ransomware, exploit kits, botnets, APTs (more on this)
- ▶ **Data breaches:** *intentional/unintentional* release of secure or private/confidential information to an untrusted entity/environment
 - organized cybercrime, political activism, national governments
 - breach targets: CCNs, PII, trade/military secrets, IPs ...etc
- ▶ **Protection Mechanisms:** encryption, data masking, access control, security policies (*a big chunk of this course!*)

Consequential Data Breaches



2016: 19,252 emails & 8,034 attachments



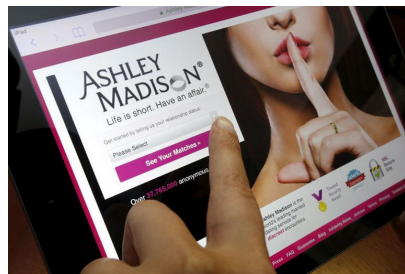
2016: >360M accounts



2013: 72.2M customer records



2014: 145M customer records



2015: 37M records



2013: ~50M accounts



2014: 76M customer records



2008: 134M customer records



2016: 412M customer records



2011: 40M customer records



TARGET

2013: 70M customer records



2014: 500M email accounts



2017: 143M customer records



2012: 22M customer records



2013: 38M customer records



2011: 77M gamers' profiles



2015: 56M customer records

What Caused the Breaches?

- ▶ **Target:** stolen credentials from a third-party HVAC vendor
- ▶ **Home Depot:** stolen credentials from a third-party vendor
- ▶ **DNC Hacks:** spear phishing followed by credential stealing exploitation
- ▶ **Anthem:** lack of data protection (missing encryption)
- ▶ **RSA Security:** targeted phishing used a Flash object embedded in an Excel file

Privacy

- ▶ RFC2828: “The right of an entity (normally a person), acting in its **own behalf**, to determine the **degree** to which it will interact with its environment, including the degree to which the entity is willing to **share information about itself with others.**”



The actor's identity might be public, but their actions remain unknown (restricted).

Anonymity

- ▶ **Anonymity:** when you opt to have your **actions seen**, but keep your **identity hidden** (e.g., voting in free elections)
- ▶ In short, “**they**” can see what you do, but not who you are
- ▶ **Privacy:** “**they**” can see who you are, but not what you do



Anonymous Communication

- ▶ **Sender anonymity:** e.g., whistleblowers
 - the identity of the party who sent a message is hidden, while its receiver (and the message itself) might not be
- ▶ **Receiver anonymity:** e.g., BCC emails
 - the identity of the receiver is hidden
- ▶ **Unlinkability of sender and receiver:** e.g., Doctors-Patients Chat
 - although the sender and receiver can each be identified as participating in some communication, they cannot be identified as communicating with each other

More on specific technologies later in the course

Contexts of Privacy: Informational

► **Generally:**

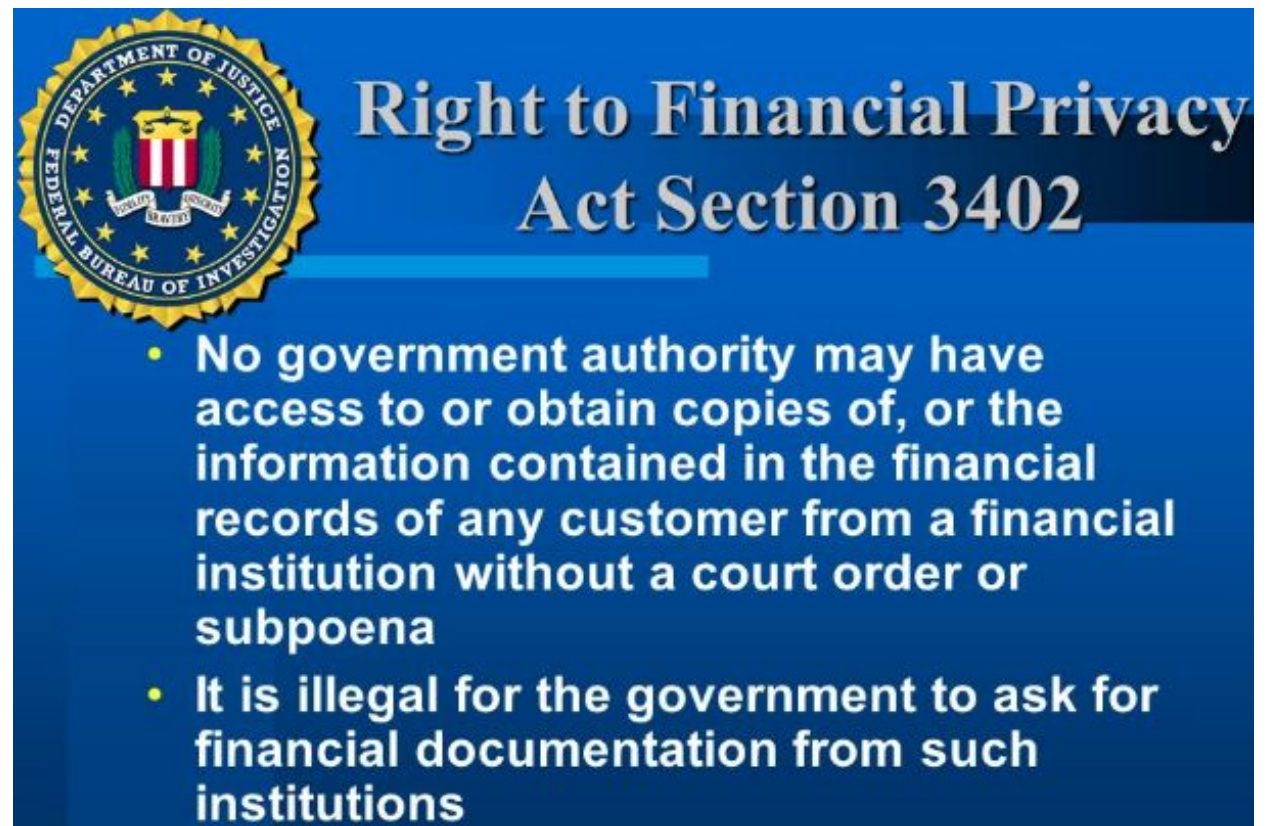
- wherever uniquely identifiable data relating to a person or persons are collected, stored, and analyzed in digital form or otherwise

► **Specifically:**

- how data are collected, stored, and analyzed
- who is given access to information
- whether an individual has any ownership rights to data about them, and/or the right to view, verify, and challenge that information

Contexts of Privacy: Financial

- guarding against fraud including identity theft
- purchases could imply preferences, places visited, contacts, products (such as medications), activities & habits
- privacy over the bank account opened by individuals



Contexts of Privacy: Internet

- ▶ The ability to determine what information one reveals or withholds about oneself over the Internet
- ▶ Who has access to such information?
- ▶ For what purposes one's information may or may not be used?
- ▶ **Examples:**
 - web sites collect, store, and possibly share PII
 - email users generally consider their emails to be private
- ▶ **Tools used to protect privacy on the Internet:** encryption tools (TSL, SSL) and anonymizing services like I2P and ToR

Contexts of Privacy: Medical

- ▶ HIPAA (Health Insurance Portability and Accountability Act of 1996):
 - allows a person to **withhold their medical records** and related information from others
- ▶ Why?
 - affects **insurance coverage** or **employment**
 - avoid implications of **revealing medical conditions** or treatments
 - could also reveal **other aspects** of one's personal life (e.g., sexual preferences)



Contexts of Privacy: Political

- ▶ **Voting systems:** secret ballot helps to ensure that voters **cannot be coerced** into voting in certain ways, since they can allocate their vote as they wish in the privacy and security of the voting booth while maintaining the anonymity of the vote



Contexts of Privacy: Legal

- ▶ Government agencies, corporations or organizations: **keep activities or secrets from** other organizations or individuals
- ▶ Organizations may seek **legal protection** for their secrets
- ▶ **Examples:**
 - governments: **invoke or declare certain information to be classified**
 - corporations: protect valuable proprietary info. as **trade secrets**

Executive Privilege

- Claim by a president that he has the right to decide that the national interest will be better served if certain information is withheld from the public, including the Courts and Congress
- *United States v. Nixon* (1973) – presidents do NOT have unqualified executive privilege (Nixon Watergate tapes)

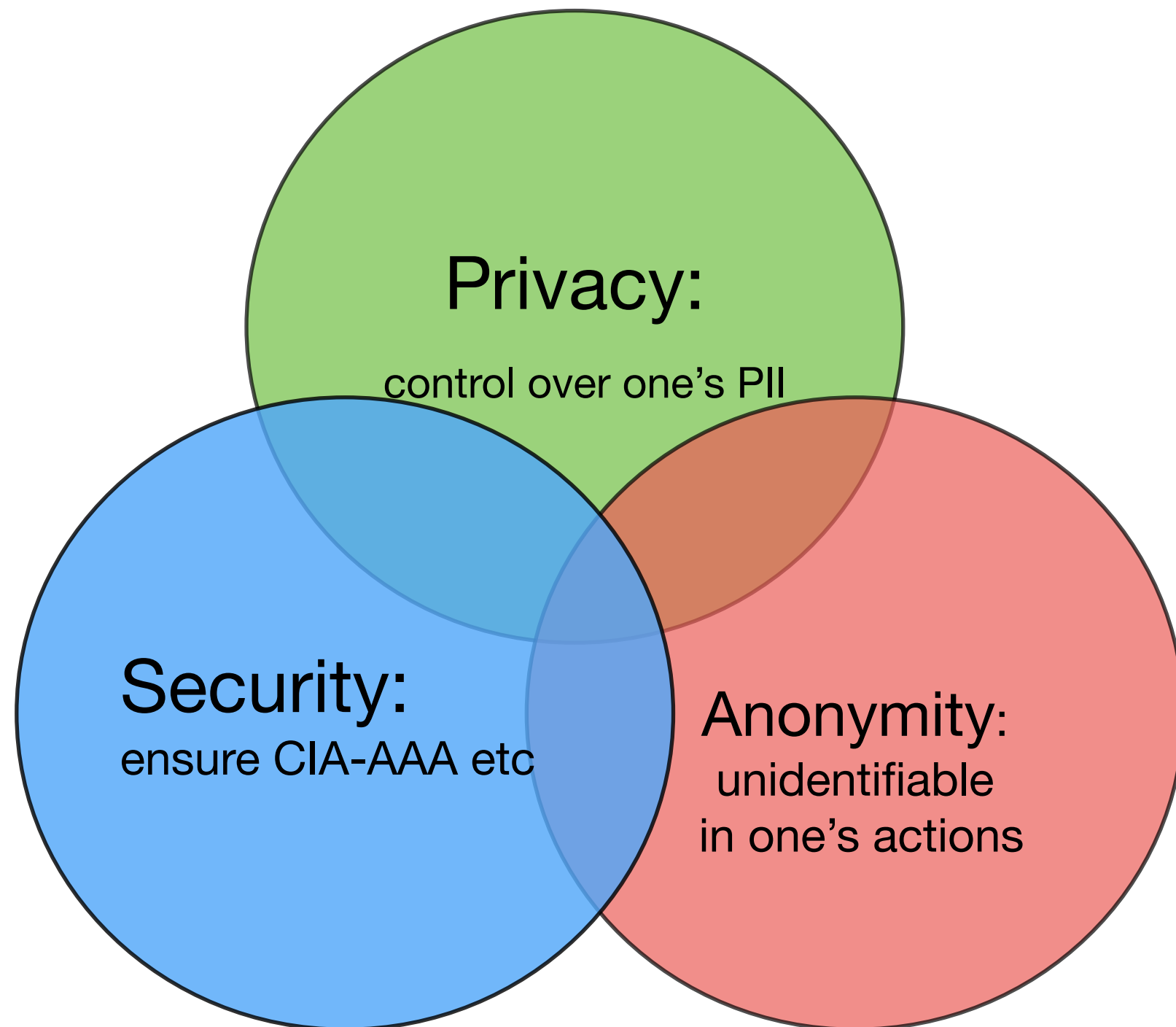


Attorney-Client Privilege

- Elements of the Privilege
 - Attorney-Client Relationship
 - Communication between lawyer and client
 - Made in confidence
 - For the purpose of facilitating legal advice
 - Kept in confidence (e.g., not waived through disclosure)
- An “absolute” privilege – no exceptions if it is applicable
- But privilege may be destroyed by conduct such as waiver or fail to apply. Most common example of this is the “crime-fraud” exception – communications are not privileged if lawyer’s services used to effect a crime or fraud



Security vs. Privacy vs. Anonymity



The Era of Big Data

► Think about your daily data footprint:

- websites you visit
- queries you issue to search engines
- the smartphone you carry
- credit card transactions you make
- toll booths you drive through
- GPS signals that direct you
- pictures, blog entries and status updates you post

► Think about the vast amount of information about you:

- interests, hobbies, and routines captured on a daily basis



► Big Data Analytics:

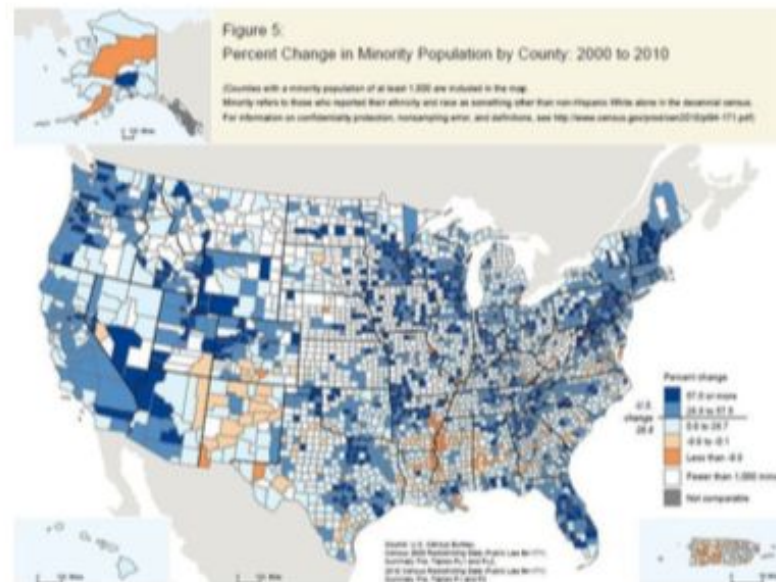
- “the ability to **capture**, **store**, **manage** and **analyze** information that would otherwise be considered **too big**, **too fast** or **too unstructured** for traditional database and analytical software”

Aggregated Personal Data

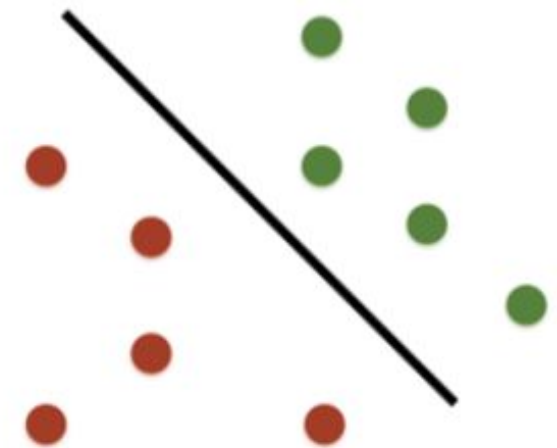
De-identified records
(e.g., medical)



Statistics
(e.g., demographic)

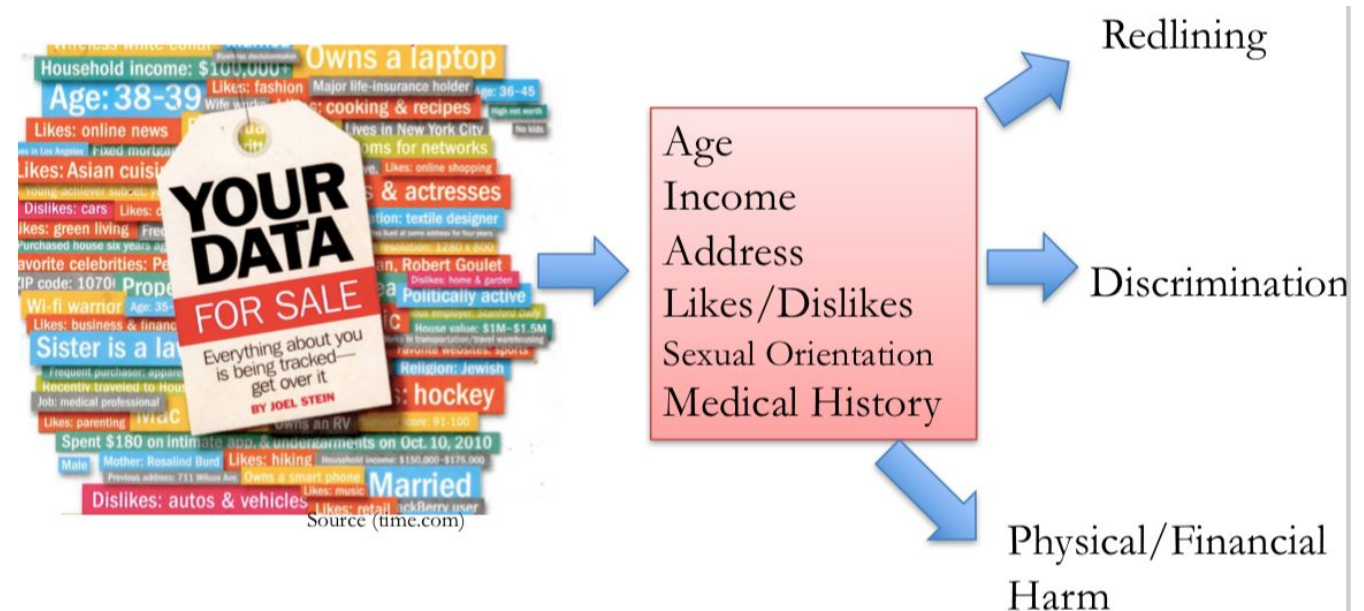


Predictive models
(e.g., advertising)



Collect -> Analyze->Target

- ▶ web data, e-commerce
- ▶ purchase at department/grocery stores
- ▶ bank/credit card transactions
- ▶ social networks
- ▶ surveillance devices and systems
- ▶ embedded systems and IoT
- ▶ drones



Data Utility vs. Privacy

- ▶ **Data analytics utility:** analytics and prediction to derive business decisions
- ▶ **Security & privacy:** ensure security & preserve privacy in the analytics pipeline
- ▶ **Reasonable goal:** maximize data utility without compromising privacy
- ▶ **Trade-off:** between utility and privacy

More on specific techniques later in the course

Who Cares, I'm Anonymous!



Source (<http://xkcd.org/834/>)

Anonymity is not Enough

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

 SIGN IN TO E-
THIS



Why 'Anonymous' Data Sometimes Isn't

By Bruce Schneier  12.13.07

Last year, Netflix published 10 million movie rankings by 500,000 customers, as part of a challenge for people to come up with better recommendation systems than the one the company was using.

The Scientist » The Nutshell

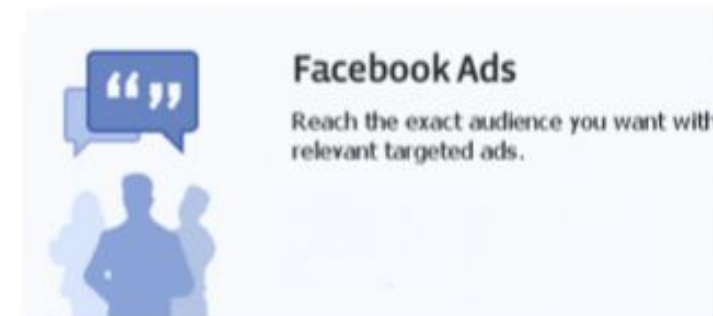
“Anonymous” Genomes Identified

The names and addresses of people participating in the Personal Genome Project can be easily tracked down despite such data being left off their online profiles.

By Dan Cossins | May 3, 2013



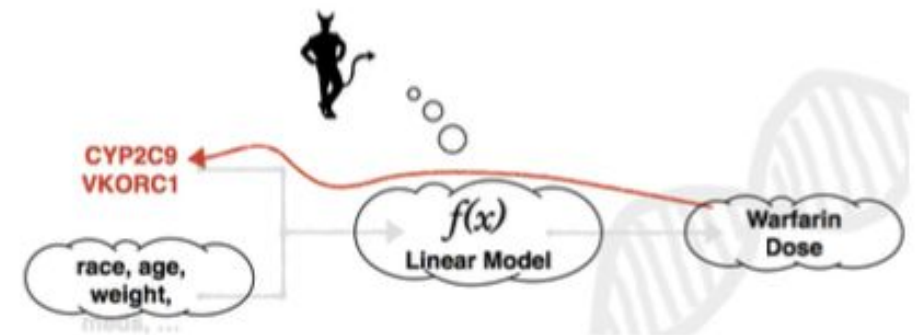
Analytics can Breach Privacy



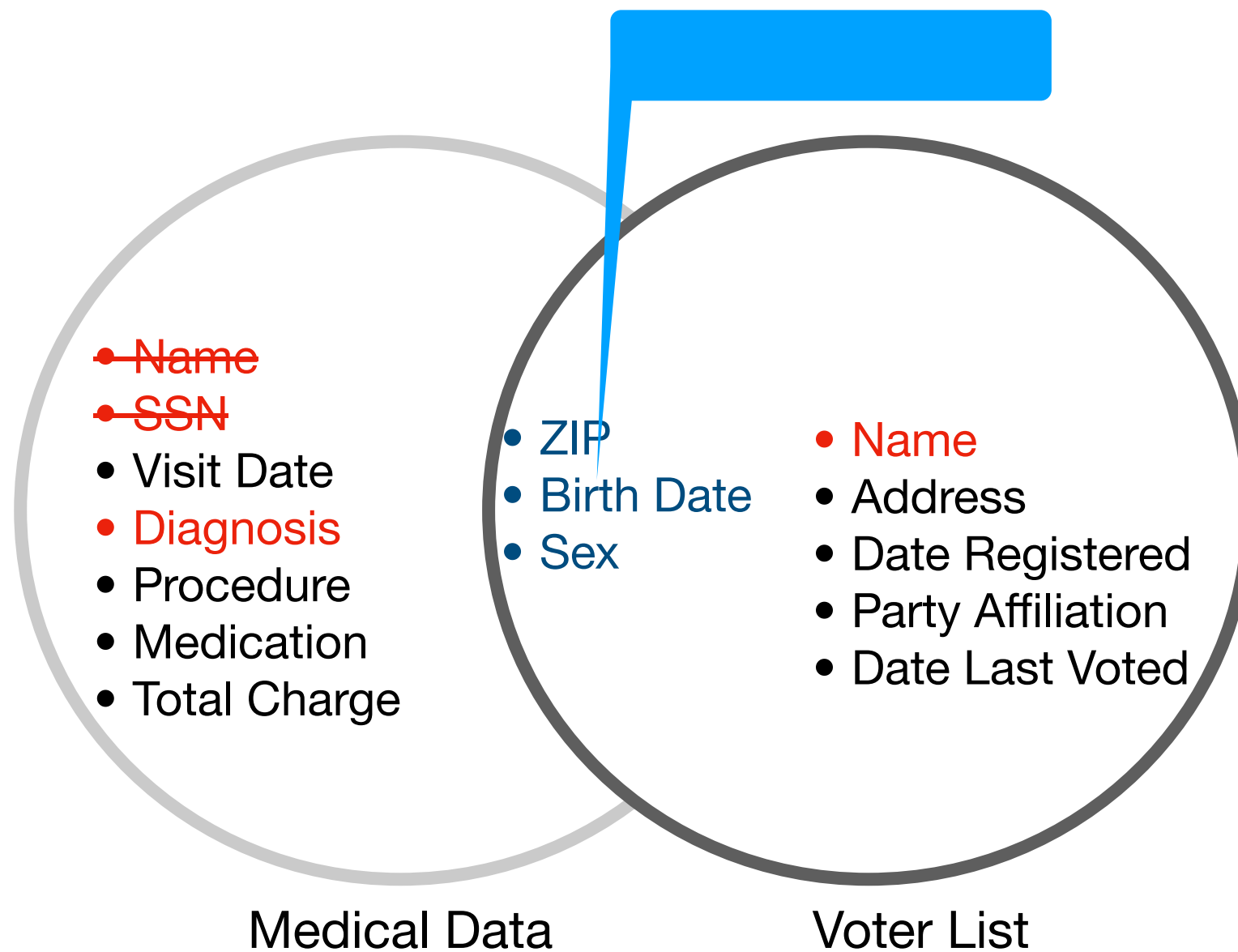
TECH | 2/16/2012 @ 11:02AM | 837,678 views

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

Privacy in Pharmacogenetics:
An End-to-End Case Study of
Personalized Warfarin Dosing



MA Governor Privacy Breach



William Weld, Governor of MA uniquely identified using ZIP Code, Birth Date, and Sex

How?

-he lived in Cambridge MA

-only 6 people had his birth date

-only 3 of them were men

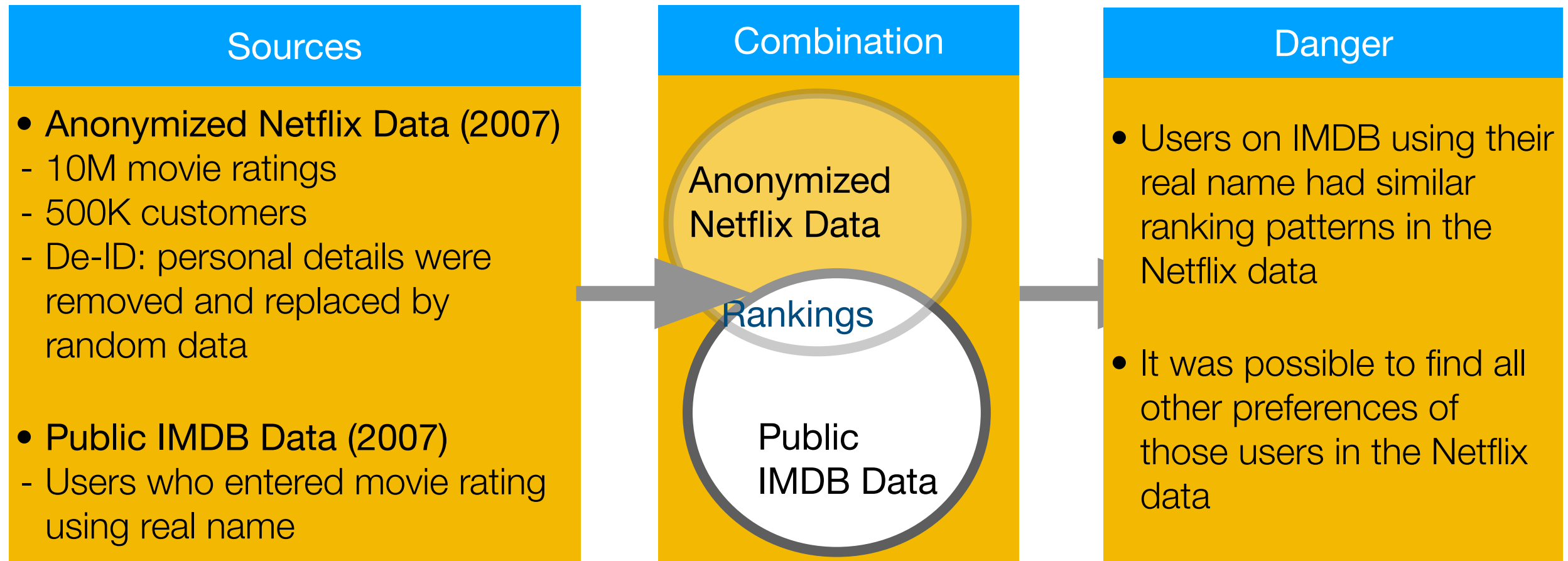
-he was the only one in his 5-digit ZIP code

87% of US population uniquely identified using ZIP Code, Birth Date, and Sex

Public by Group Insurance Commission of MA
Voter registration list of MA purchased for \$20

Sweeney, 2002: "*k-Anonymity: A Model for Protecting Privacy*"
Paper: https://epic.org/privacy/reidentification/Sweeney_Article.pdf

Netflix + IMDb De-anonymization



► **Lesson Learned:** you can never fully estimate the anonymity of your data using standard de-identification approaches

Narayanan & Shmatikov, 2008: "*Robust De-anonymization of Large Sparse Datasets*"

Paper: https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

Lecture Summary

- ▶ **Fundamentals:** confidentiality, integrity, availability, authentication, authorization, auditing/accountability, non-repudiation, security, privacy, anonymity, cyber-attack, data breach, PII, ...
- ▶ **Contexts of privacy:** informational, financial, online, medical, political, legal
- ▶ **Security, privacy, and anonymity:** have different meanings
- ▶ **The Big Data Phenomenon:** utility-privacy trade-off
- ▶ **Inadequacy of De-identification:** MA governor's re-identification, Netflix+IMDb de-anonymization