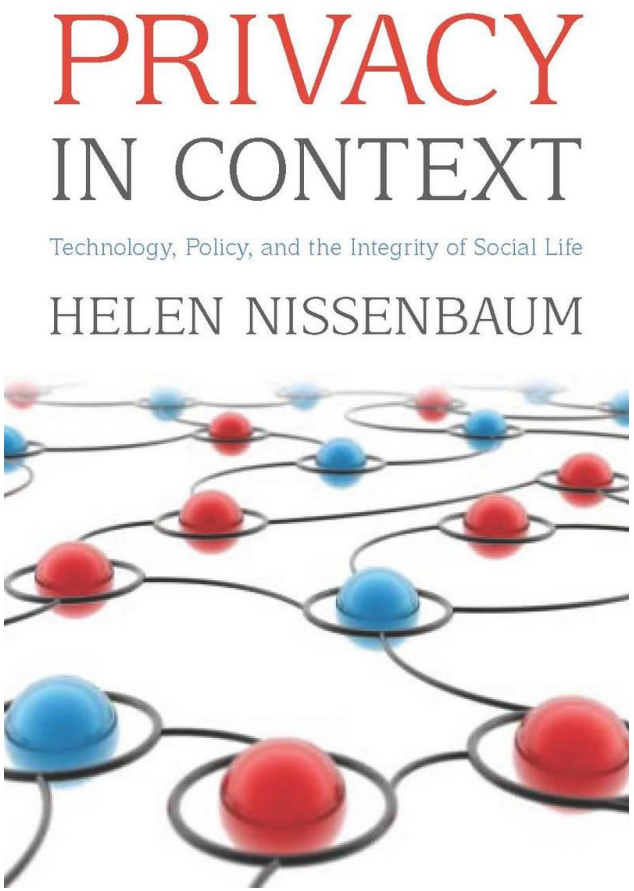
Data Privacy Introduction

Erman Ayday

WHAT IS PRIVACY?

Privacy: Definition

- **Privacy control** is the ability of individuals to determine when, how, and to what extent information about themselves is revealed to others
- **Tool:** Privacy-enhancing technologies
- **Goal**: Let personal data be used only in the context it has been released



Personal Data

- Any kind of information (a single piece of information or a set of information) that can personally identify an individual
  - Name, address, national identification number, date of birth, a photograph, hospital records, etc.
- Protection is crucial
- In an interconnected electronic world, individual pieces of data can no longer be regarded in isolation
- Can also be used to put people under complete surveillance, in breach of their fundamental rights

The Value of Privacy (1)

- Depends on the individual
- Common misconception:
  - If you aren't doing anything wrong, what do you have to hide?
- Privacy is not hiding the wrong!

If one would give me six lines written by the hand of the most honest man, I would find something in them to have him hanged.

- Cardinal Richelieu (1585-1642)

The Value of Privacy (2)

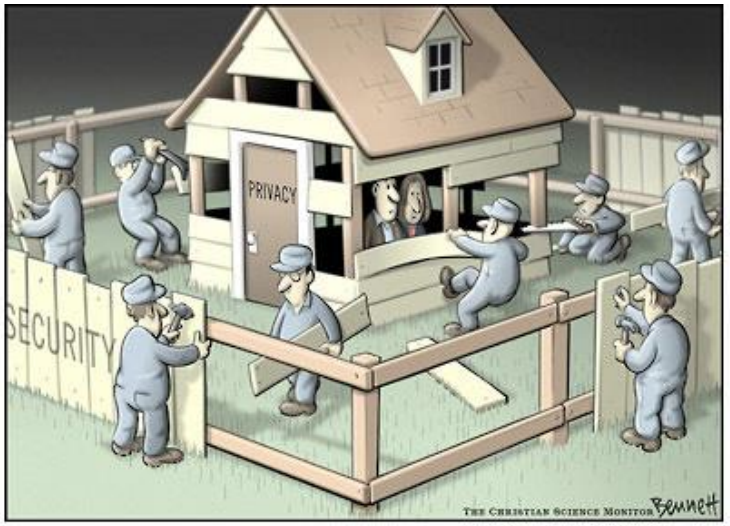
- Privacy is an inherent human right
  - A requirement for maintaining the human condition with dignity and respect
- Privacy is a basic human need
- People would lose their individuality knowing everything they do is observable and recordable
- Privacy accords us the ability to control who knows what about us and who has access to us
  - It allows us to vary our behavior with different people
  - So we can maintain and control our various social relationships, many of which will not be intimate
- Lack of privacy is equivalent to loss of freedom

Security vs. Privacy

- Which is more important?
- How much privacy are you willing to give up for security?
- How much liberty buys how much security?
- Can we even afford privacy in this age of insecurity?

https://www.schneier.com/blog/archives/2008/01/security_vs_pri.html

Security vs. Privacy – Common Beliefs

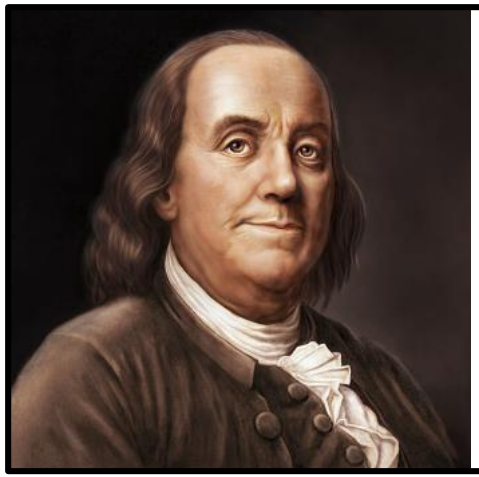


- Increased security can be bought with reduced privacy
- Privacy and security are a zero-sum game
  - If one gains, the other loses
- Security is vital to survival
  - To the defenders of the surveillance state, security means "saving American lives"
- Privacy is unique to humans, but it's a social need
- In 2008, 51% of Americans said security is more important than privacy [1]
  - Only 29% disagreed and said privacy is more important

[1] http://www.rasmussenreports.com/content/pdf/8522

Security vs. Privacy – Bottom line

- Security and privacy are not opposite ends of a seesaw
  - No need to accept less of one to get more of the other
- There is no security without privacy
- Liberty requires both security and privacy

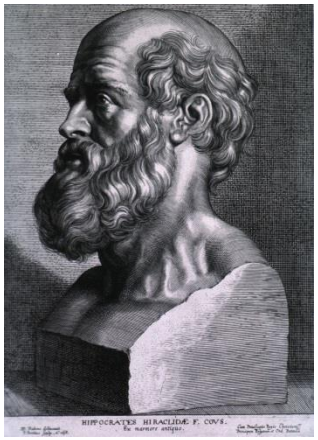


Those who would give up essential liberty to purchase a little temporary safety, deserve neither liberty nor safety

- Benjamin Franklin

History of Privacy

History of Privacy



Hippocrates
Ca. 460 to ca. 370 B.C.

Hippocratic oath

"…

All that may come to my knowledge in the exercise of my profession or in daily commerce with men, which ought not to be spread abroad, I will keep secret and will never reveal.

…"

Many centuries later…

“The Right to Privacy”

Warren and Brandeis

Harvard Law Review

Vol. IV Dec. 15, 1890 No. 5

First explicit declaration of a US right to privacy

Major concern: photography without consent

Privacy Concept

"The Right to Privacy" (1890)

- Laid the foundation for a concept of privacy known as control over information about oneself

William Prosser: different interests in privacy (1960):

- Intrusion upon a person's seclusion or solitude, or into his private affairs
- Public disclosure of embarrassing private facts about an individual
- Publicity placing one in a false light in the public eye
- Appropriation of one's likeness for the advantage of another

Views on the Meaning and Value of Privacy

Alan Westin described privacy as

- *the ability to determine for ourselves when, how, and to what extent information about us is communicated to others* (1967)

William Parent defined privacy as the condition of not having undocumented **personal information** known or possessed by others (1983)

- Personal information is documented, on Parent's view, only when it belongs to the public record, that is, in newspapers, court records, or other public documents

Today - Privacy Definitions

- There are many privacy definitions
  - k-anonymity, l-diversity, t-closeness, ε- Differential privacy (later in this course)
- Why can't we have a single definition for privacy?
- No Free Lunch Theorem [1]:
  - For every algorithm that outputs a D with even a sliver of utility, there is some adversary with a prior such that privacy is not guaranteed

[1] "No Free Lunch in Data Privacy", Kifer and Machanavajjhala, 2011

Why is Privacy Tricky to Understand?

- Abstract concept
- Diverging opinions
- Multi-disciplinary: computer science, law, economics, sociology, politics,…
- In computer science alone: applied crypto, applied statistics, HCI, inference techniques, machine learning, databases, signal processing, networking, operating systems,…
- There is no textbook on the technical aspects (and the field is evolving very fast, making the writing of such a book highly problematic)
  ➔ we'll rely on surveys/tutorials

Why is privacy difficult to promote?

Privacy Protection is at Odds with:

Security (e.g., homeland security)

Usability

Business (e.g.,
targeted advertisement)

System performance

Medical progress

Benefits of system usage are immediate, drawbacks
(in terms of privacy) usually are uncertain and come later

TOOLS WE USE THAT LEAK OUR DATA

Privacy Players

- Government
- Companies
  - Advertisers
- Organizations we work for
- Colleagues
- Family and friends
- Strangers

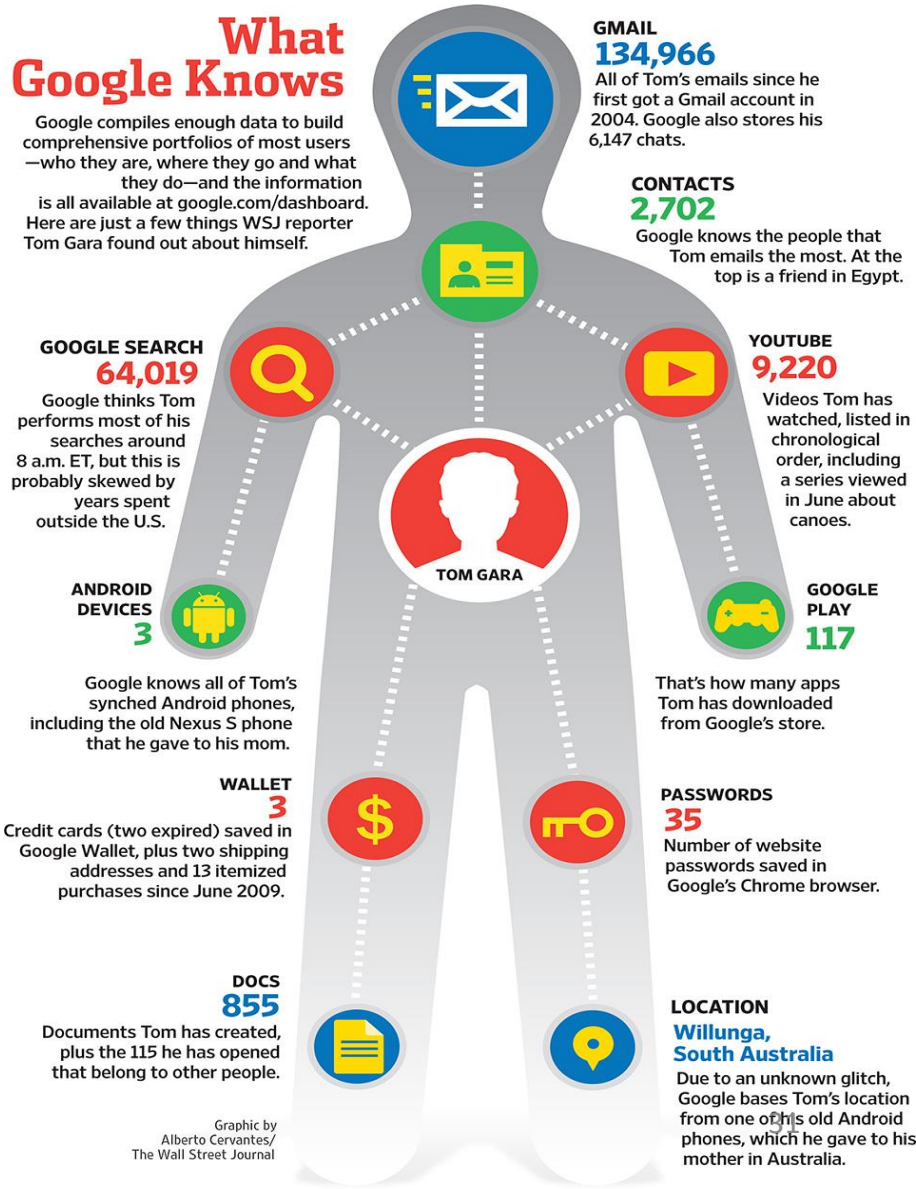
Why Do Some Companies Want Our Data?

- Information and communication technologies make life better
- For applications to improve their performance and offer us better services, they require information about us
- Social networking sites and most online services are usually free to join
  - Service providers collect our data and use it for "targeted advertising"

If you're not paying for it, you're part of the product

What Do Companies Know About Us?

- Companies have very detailed profiles of who we are
  - Can pose a problem if data is used to discriminate on the basis of assumed health status, age, gender, sexual orientation, etc.
  - Even more problematic if governments seek access to and use these data
- Most of us are happy to give out personal information in exchange for specific services
- What we object to is the improper collection of personal information, and the secondary use of information once it's collected
  - The buying and selling of our information behind our back



Ways Companies Collect Our Data

- Companies collect information about us:
  - Google collects personal information, data about the services you use and how you use them, server logs, location information, etc.
- We voluntarily share our information
  - Facebook's 1 billion active users share roughly 5 billion items a day, not counting the data Facebook collects about them

- *A child born today will grow up with no conception of privacy at all*

  Eric Snowden

- "Privacy died with the information age"

Cloud Computing

- Companies usually store/process data on the cloud
- Worries about control of the data and their geographical location
  - Who has access to it?
  - How can it be used?
  - How easy is it to move the data from one cloud service to another?
  - How secure are they?
  - Who is responsible if the data are lost or misused?

Data Breaches

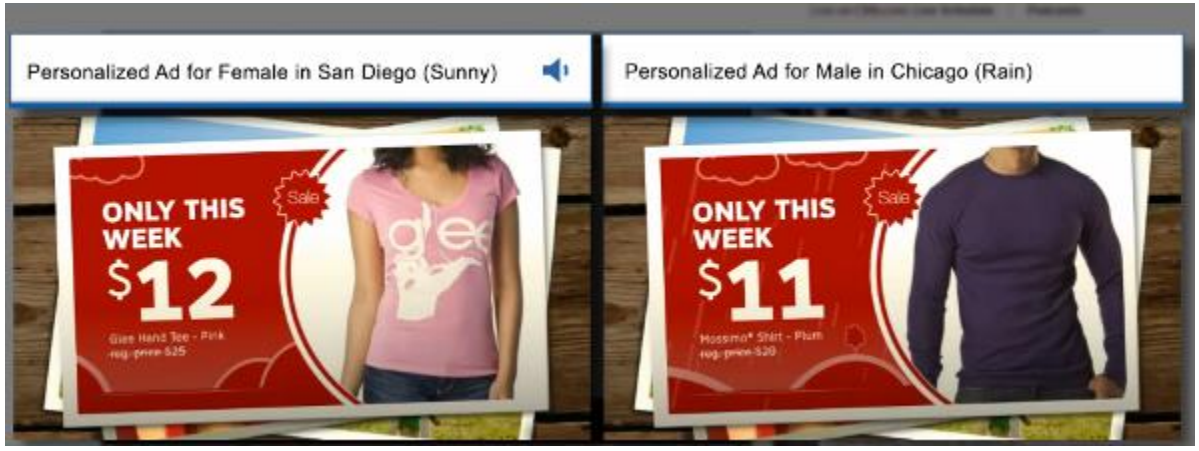
- Details of the average economically active adult in the developed world are located in around 700 major databases (as of 2002) [1]
  - Enough processed data to compile a formidable reference book for each person
- In 2013, in 2164 separate incidents, over 822 million records were exposed [2]

[1] http://www.guardian.co.uk/uk/2002/sep/07/privacy2

[2] https://www.riskbasedsecurity.com/2014/02/2013-data-breach-quickview/

How Companies Learn Your Secrets

- Target can buy data about:
  - Your ethnicity, job history, the magazines you read, if youve ever declared bankruptcy or got divorced, the year you bought your house, where you went to college, what kinds of topics you talk about online, your political leanings, reading habits, the number of cars you own, etc.
- All that information is meaningless without someone to analyze and make sense of it
- Targets Guest Marketing Analytics:
  - Find the customers who have children and send them catalogs that feature toys before Christmas
  - Look for shoppers who habitually purchase swimsuits in April and send them coupons for sunscreen in July and diet books in December



http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all&_r=0

Why Target Wants to Know You Are Pregnant?

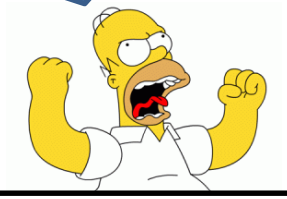
- Birth records are usually public
  - The moment a couple have a new baby, they are barraged with offers and advertisements from all sorts of companies
- The key is to reach them earlier, before any other retailers know a baby is on the way
  - When customers are going through a major life event, their shopping habits change in ways that are both predictable and potential gold mines for retailers
  - Marketers want to send specially designed ads to women in their second trimester of pregnancy
  - If companies can identify pregnant shoppers, they can earn millions

Target Knows You Are Pregnant

An angry man went into a Target outside of Minneapolis, demanding to talk to a manager

"My daughter got this in the mail!"

"She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?"



"We deeply apologize"

Manager called a few days later to apologize again

"We apologize again"

"I had a talk with my daughter"

"It turns out there's been some activities in my house I haven't been completely aware of. She's due in August. I owe you an apology"

Netflix Case

- Netflix movie recommendations

- Netflix provided a training data set of 100,480,507 ratings that 480,189 users gave to 17,770 movies

Researchers Reverse Netflix Anonymization [1]

- Researchers from the UT Austin identified two people out of all anonymized users whose movie ratings were released by online rental company Netflix
- The collection of movie ratings - combined with a public database of ratings - was enough to identify the people
  - Public reviews published by a few dozen people in Internet Movie Database (IMDb)

- Researchers found that one of the users had strong opinions about some liberal and gay-themed films

[1] Robust De-anonymization of Large Sparse Datasets - Narayanan and Shmatikov, 2008

AOL Search Logs

- 20 million Web search queries collected by AOL (anonymously) made public
- UserIDs were replaced by random numbers
- Examples (all provided by AOL)

(AOL stands for America Online, a major US ISP)

865712345	Uefa cup
865712345	Uefa champions league
865712345	Champions league final
865712345	Champions league final 2007
236712909	exchangeability
236712909	Proof of de Finetti's theorem
112765410	Zombie games
112765410	Warcraft
112765410	Beatles anthology
865712345	Grammy 2008 nominees

17556639 how to kill your wife
17556639 wife killer
17556639 how to kill a wife
17556639 poop
17556639 dead people
17556639 pictures of dead people
17556639 killed people
17556639 dead pictures
17556639 murder photo
17556639 steak and cheese
17556639 photo of death
17556639 death
17556639 dead people photos
17556639 photo of dead people
17556639 www.murderedpeople.com
17556639 decapitated photos
17556639 car crashes3
17556639 car crash photo

This user was not de-anonymized, fortunately for him…

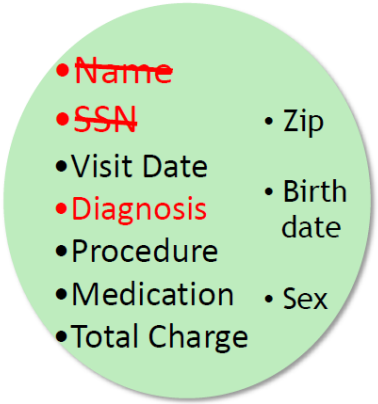
A Face Is Exposed for AOL Searcher No. 4417749 [1]

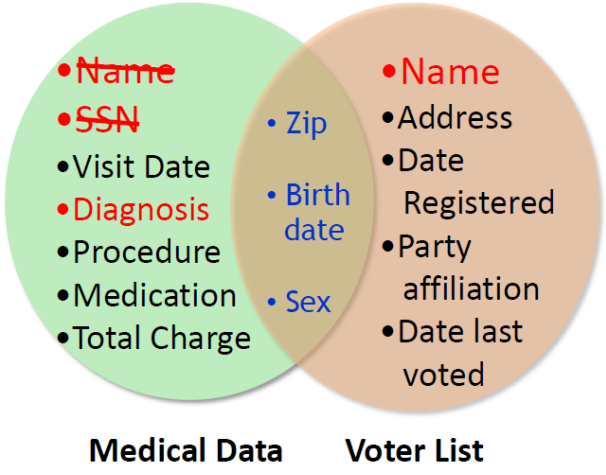
- Searches by No. 4417749:
  - landscapers in Lilburn, GA
  - people with the last name Arnold
  - dog that urinates on everything
- Data trail led to Thelma Arnold, a 62-year-old widow who lives in Lilburn, GA
- AOL removed the search data from its site over the weekend and apologized for its release

[1] The New York Times - August 9, 2006

The Massachusetts Governor Privacy Breach [1]

- Release of anonymized data is socially beneficial
  - E.g., the usage of hospital records greatly helps medical research
- Governor of MA uniquely identified using ZipCode, Birth Date, and Gender
  - 87% of US population can be uniquely identified using ZipCode, Birth Date, and Gender
- Name linked to Diagnosis



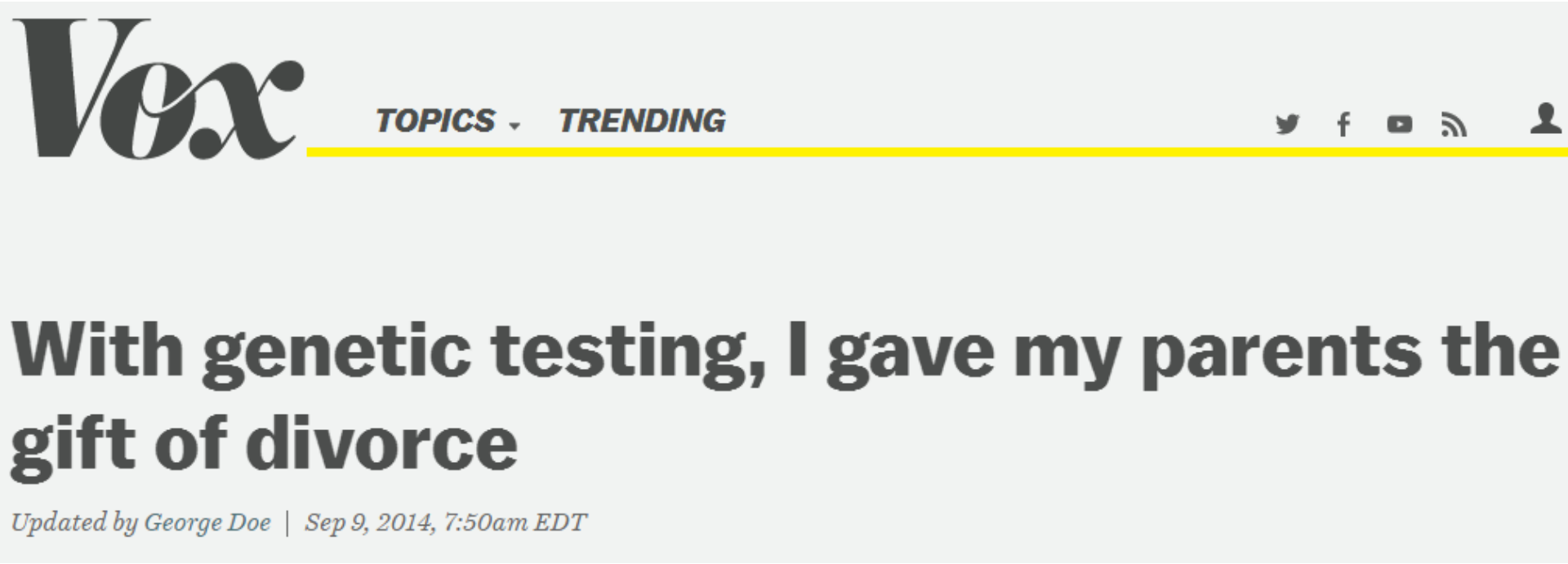


SSN: Social Security Number

[1] K-Anonymity: a model for protecting privacy - L. Sweeney, 2002

Figure: Ashwin Machanavajjhala

Innocent Leaks



Privacy vs. Law Enforcement

Public Datasets

- Still a lot of public datasets exists
  - Transactions
  - Driving habits
  - Expedia hotel searches
  - Yelp business ratings
  - Yandex searches
  - …

PRIVACY AND TECHNOLOGY

Privacy and Technology

Challenges for privacy with developing technology (before PRISM)

- Many people still view privacy as a valuable interest and realize it is now threatened more than ever by technological advances
- There are massive databases and Internet records of information about individual financial and credit history, medical records, purchases and telephone calls
- The ability for others to access and link the databases, with few controls on how they use, share, or exploit the information, makes ***individual control over information about oneself*** more difficult than ever before

Privacy and Technology

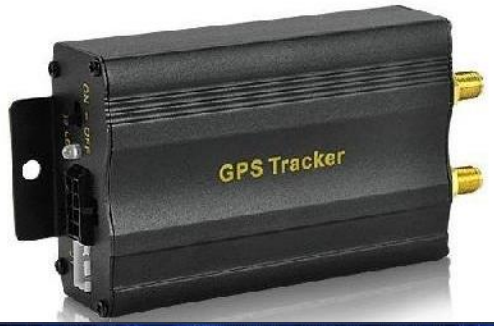
Challenges for privacy with developing technology (before PRISM)

- There is widespread mandatory and random drug testing of employees and others
  - Supreme Court has said policies requiring all middle and high school students to consent to drug testing in order to participate in extracurricular activities does not violate the Fourth Amendment
- Heat sensors aimed at and through walls to detect such things as growing marijuana could become acceptable
  - A court decided that thermal imaging devices that reveal information previously unknowable without a warrant does constitute a violation of privacy rights and the Fourth Amendment (*Kyllo v. U.S.* 2001)
- Surveillance photos are commonly taken of those using Fast Lane, resulting in tickets mailed to speeding offenders
  - Similar photos are also taken at red lights

Privacy and Technology

Challenges for privacy with developing technology (before PRISM)

- Face scanning at casinos and at large sporting events
  - Resulting in the capture of multiple offenders on the loose but also posing privacy issues for innocents photographed without their knowledge
- Rental car drivers are tracked by Global Positioning System (GPS) satellites
- There is a proliferation of biometric identification using faces, eyes, fingerprints, and other body parts for identifying specific individuals



PRISM

- Since 9/11, the US government has dramatically increased the ability of its intelligence agencies to collect and investigate information on both foreign subjects and US citizens

- Some of these surveillance programs capture the private data of citizens who are not suspected of any connection to terrorism or any wrongdoing

- In June 2013, a private contractor working for Booz Allen Hamilton, Edward Snowden, leaked classified presentation slides that detailed the existence of PRISM

PRISM – Who is Involved?

- PRISM: tool used by the NSA to collect private electronic data belonging to users of major internet services
  - Latest evolution of the US government's post-9/11 electronic surveillance efforts, which began under President Bush with the ***Patriot Act***, and expanded to include the ***Foreign Intelligence Surveillance Act (FISA)*** enacted in 2006 and 2007

PRISM - Scope

- NSA analysts are not allowed to specifically target someone "reasonably believed" to be a US person communicating on US soil
  - According to The Washington Post, an analyst must have at least "51 percent" certainty their target is foreign

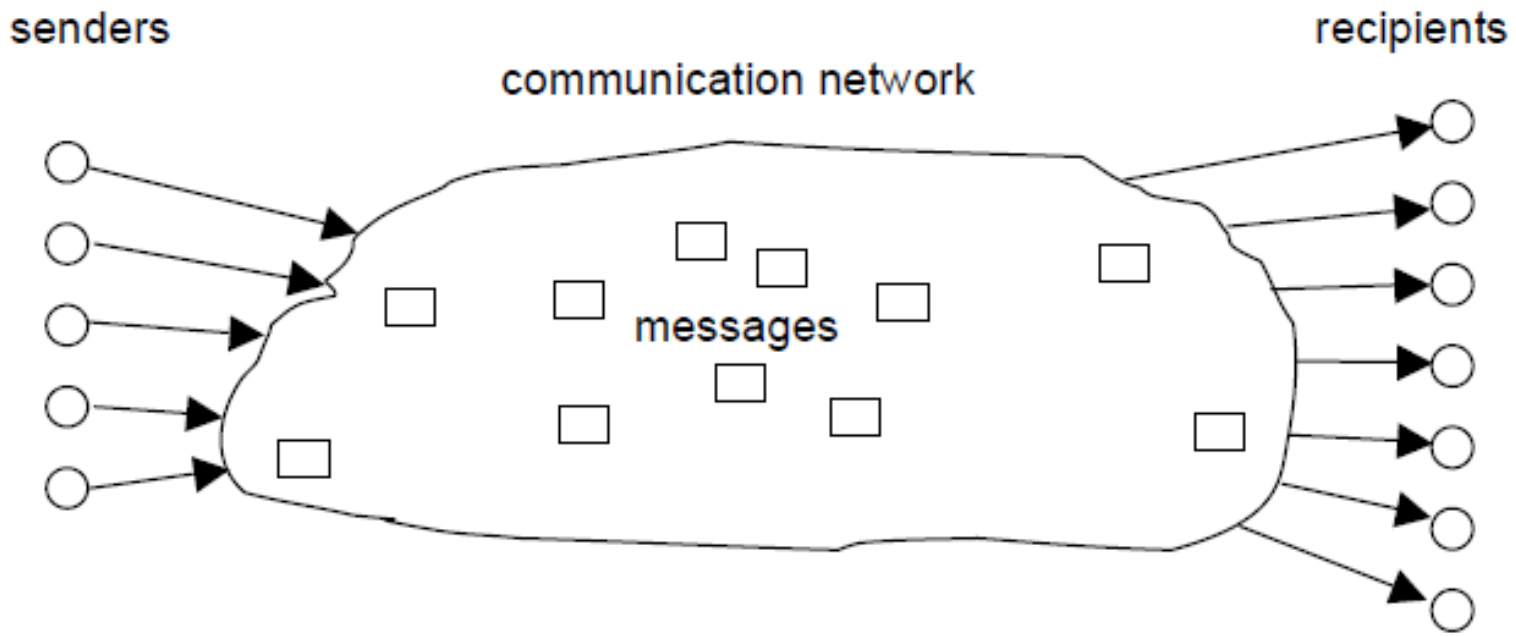
- But, NSA's "contact chaining" practices can easily cause innocent parties to be caught up in the process
  - An analyst collects records on a target's contacts, and their contacts' contacts


// A FILM BY ACADEMY AWARD® WINNER:
OLIVER STONE
THE ONLY SAFE PLACE IS ON THE RUN
SNOWDEN
JOSEPH GORDON-LEVITT // SHAILENE WOODLEY
// IN THEATERS SEPTEMBER 16
// DIRECTED BY:
OLIVER STONE
// SCREENPLAY BY:
KIERAN FITZGERALD & OLIVER STONE

SOME PRIVACY TERMINOLOGY

Some Privacy Terminology

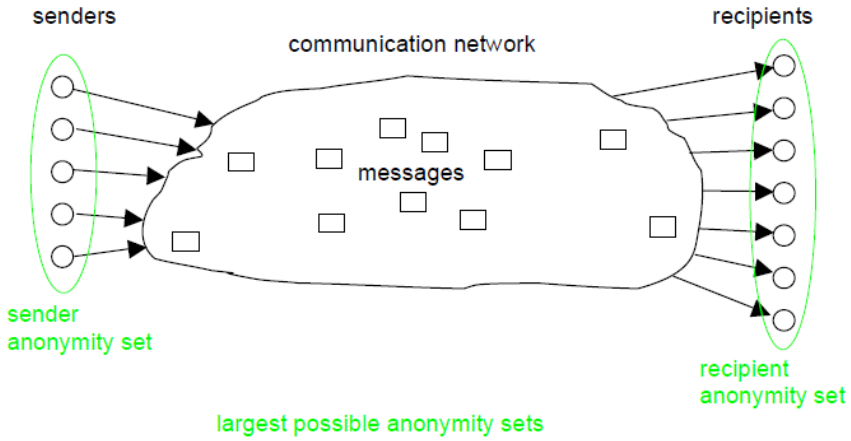
- Anonymity, unlinkability, unobservability, and related concepts
- Illustrative Example:

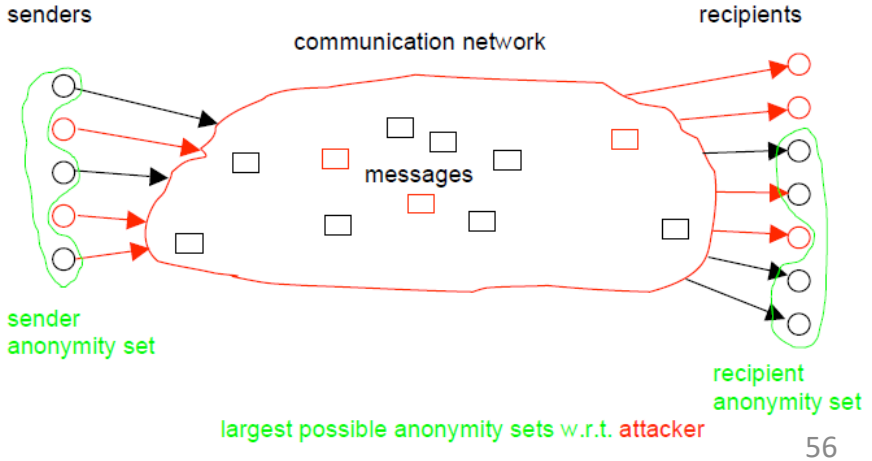


Anonymity, Unlinkability, Unobservability, Pseudonymity, and Identity Management – A Consolidated Proposal for Terminology – Pfitzmann and Hansen, 2005

Anonymity

- Anonymity: state of being not identifiable within a set of subjects (the anonymity set)
- All other things being equal, anonymity is the stronger if
  - the respective anonymity set is larger
  - the sending or receiving of the subjects within that set is more evenly distributed





Anonymity vs. Privacy

- Privacy is not anonymity

Bob	
Zipcode	**Age**
13039	33

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	< 30	*	AIDS
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

- Bob's record is indistinguishable from records of other cancer patients
  - Yet, we can infer Bob has Cancer

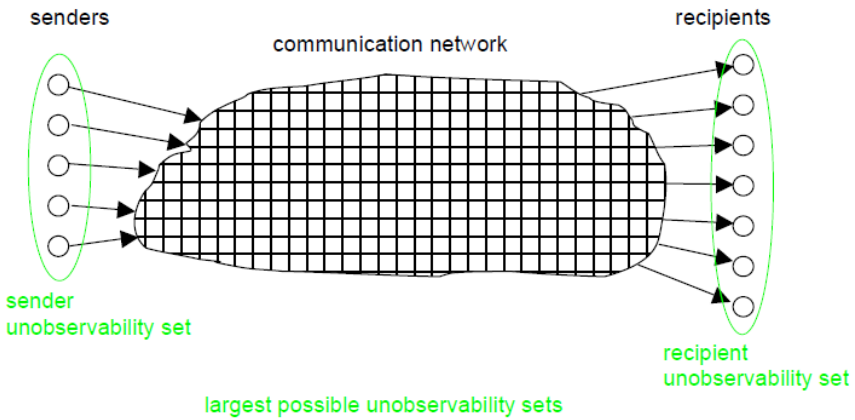
Unlinkability

- Unlinkability of two or more items of interest (IOIs, e.g., subjects, messages, events, actions, etc.) means that within the system, from the attacker's perspective, these IOIs are no more and no less related after his observation than they are related relying on his a-priori knowledge
- **Unlinkability ensures that a user may make multiple uses of resources or services without others being able to link these uses together**
- Unlinkability of items means that the ability of the attacker to relate these items does not increase by observing the system

- **Anonymity may be defined as unlinkability of an IOI and any identifier of a subject**

Unobservability

- Unobservability is the state of IOIs being indistinguishable from any IOI (of the same type) at all
- Sender unobservability means it is not noticeable whether any sender within the unobservability set sends
- Recipient unobservability means it is not noticeable whether any recipient within the unobservability set receives
- Desirable property of steganographic systems
- Also related to Oblivious RAM and Private Information Retrieval (PIR) (will be addressed later)

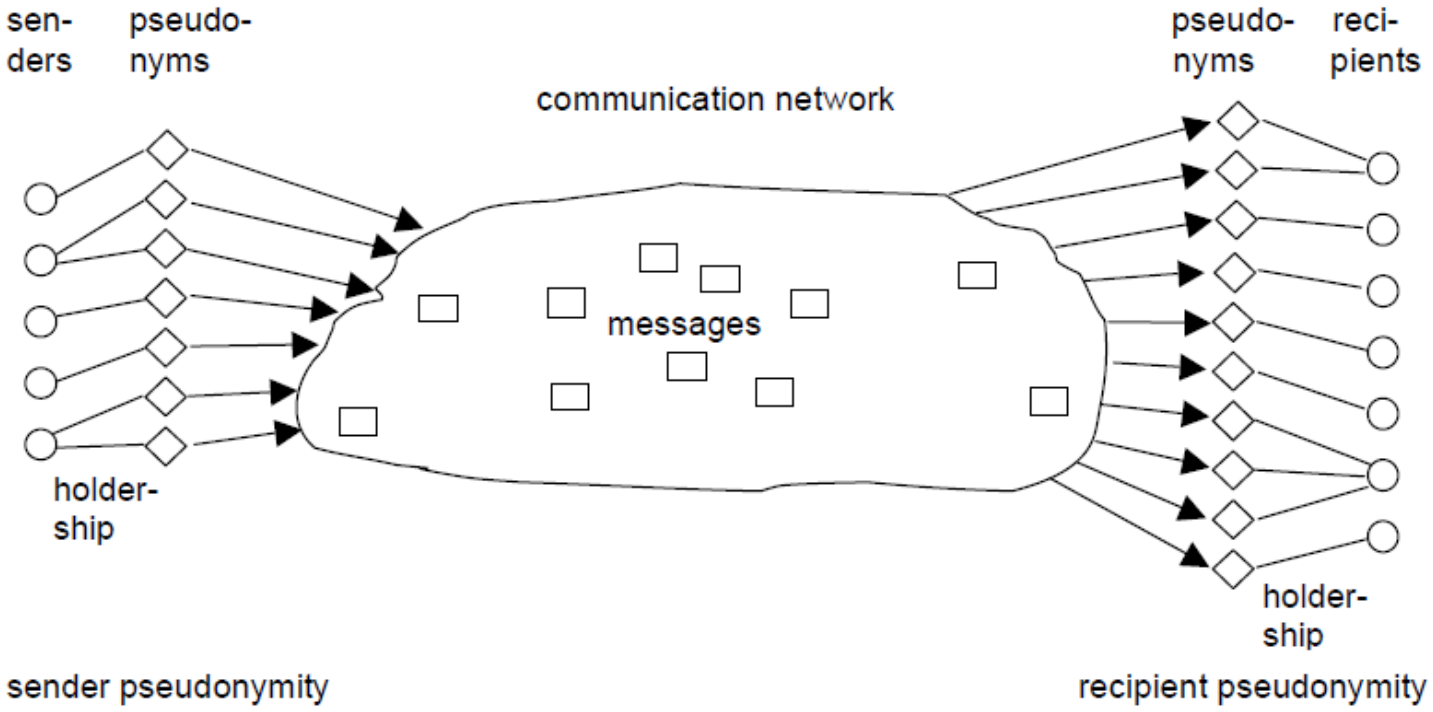


Unobservability vs. Anonymity

- Unobservability implies anonymity
- Anonymity does not imply unobservability
  - Anonymity only hides the identity of the sender/receiver, it does not guarantee unobservability

Pseudonymity

- Being pseudonymous is the state of using a pseudonym as ID



Pseudonymity with respect to Linkability

- For stronger anonymity:
  - Less personal data of the pseudonym holder should be linked to the pseudonym
  - Less often and the less context-spanning pseudonyms should be used
    - Therefore less data about the holder can be linked
  - Independently chosen, i.e., from an observer's perspective unlinkable
- Example: one's genome cannot be used as a pseudonym!

LEGAL FRAMEWORK

Policy and Consent

- Purpose limitation
  - The principle that a data controller can only collect and use personal data for a specific purpose
  - Without this, a data controller could collect personal data for a certain purpose and continue to use it any way it wishes
    - But, who reads “privacy statement” or “terms of use” when joining a service?
- Consent to data processing
  - Consent is one out of the legal grounds on which personal data can be processed

Information Privacy Laws

- Many countries have adopted comprehensive data protection laws
  - nearly every country in Europe and many in Latin America, Asia and Africa

- The US is notable for not having adopted a comprehensive information privacy law but rather having adopted limited sectorial laws in some areas: videotape rental records, healthcare information, etc.

Europe

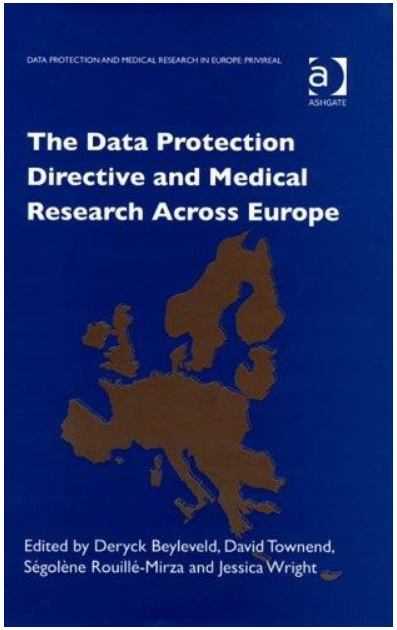
- The right to data privacy is heavily regulated and actively enforced in Europe
- Article 8 of the European Convention on Human Rights (ECHR) provides a right to respect for one's "*private and family life, his home and his correspondence*", subject to certain restrictions
  - Collection of information by officials of the state about an individual without his consent always falls within the scope of Article 8
- **Gathering information for the following has been judged to raise data privacy issues:**
  - Official census
  - Recording fingerprints and photographs in a police register
  - Collecting medical data or details of personal expenditures
  - Implementing a system of personal identification
- "The Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data" (1981)
  - Regulates automatic processing of personal data (by private companies)

Europe

Data Protection Directive

- Data protection law was harmonized within the EU via the *Data Protection Directive - DPD* (1995)
  - Anyone processing personal data must comply with the eight enforceable principles of good practice:

- Data must be:
  i. Fairly and lawfully processed
  ii. Processed for limited purposes
  iii. Adequate, relevant and not excessive
  iv. Accurate
  v. Kept no longer than necessary
  vi. Processed in accordance with the data subject's rights
  vii. Secure
  viii. Transferred only to countries with adequate protection



- All EU member states adopted legislation based on this directive
- Each country also has its own supervisory authority to monitor the level of protection

U.S.★EU SAFEHARBOR U.S. DEPARTMENT OF COMMERCE

Europe

Transfer of Personal Data

- DPD prohibits the transfer of personal data to non-EU countries that do not meet the EU "adequacy" standard for privacy protection
  - The transfer of personal information from the EU to the US is prohibited when equivalent privacy protection is not in place in the US
- **US companies that would work with EU data must comply with the framework principles**

Safe Harbor principles must provide:

- **Notice** - Individuals must be informed that their data is being collected and about how it will be used
- **Choice** - Individuals must have the option to opt out of the collection and forward transfer of the data to third parties
- **Onward Transfer** - Transfers of data to third parties may only occur to other organizations that follow adequate data protection principles
- **Security** - Reasonable efforts must be made to prevent loss of collected information
- **Data Integrity** - Data must be relevant and reliable for the purpose it was collected for
- **Access** - Individuals must be able to access information held about them, and correct or delete it if it is inaccurate
- **Enforcement** - There must be effective means of enforcing these rules

Europe vs US

Main Differences

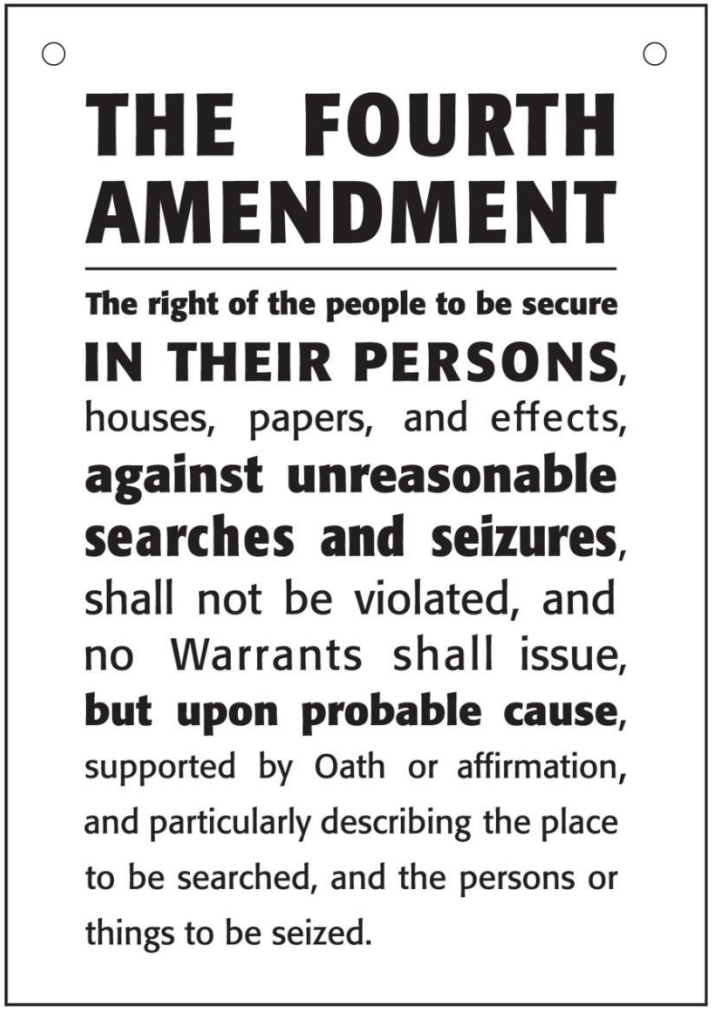
- US has developed a limited system of privacy protection that focuses on self-regulation within industry and government
  - Hence, personal information is often readily available
- EU has adopted an alternative vision highlighting consumer protection and individual privacy against the economic interests of firms and public officials
- US has generally stood behind efficiency arguments that business and government need unlimited access to personal data to guarantee economic growth and national security
- EU has sent a coherent signal that privacy has critical value in a robust information society
  - Because citizens will only participate in an online environment if they feel their privacy is guaranteed against ubiquitous business and government surveillance

United States

The Fourth Amendment

- The Fourth Amendment **prohibits unreasonable searches and seizures and requires any warrant to be judicially sanctioned and supported by probable cause**

- Fourth Amendment protection against search and seizure was extended later in the twentieth century to cover telephone wiretaps and electronic surveillance



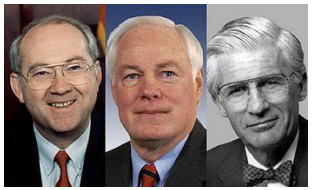
THE FOURTH AMENDMENT

The right of the people to be secure IN THEIR PERSONS, houses, papers, and effects, **against unreasonable searches and seizures**, shall not be violated, and no Warrants shall issue, **but upon probable cause**, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.

United States

Other Laws on Privacy

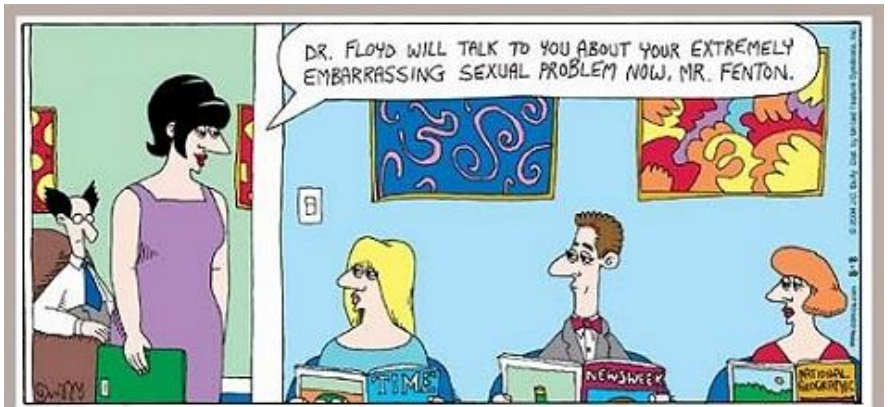
- There is no all-encompassing law regulating the acquisition, storage, or use of personal data in the US
- Private data contained in third-party credit reports may be sought when seeking employment or medical care, or making automobile, housing purchases
- There are laws protecting specific pieces of personal data but nothing like the broad privacy protection laws in Europe
  - HIPAA (Health Insurance Portability and Accountability Act, 1996)
    - Protecting personal health information
  - GLBA (Gramm-Leach-Bliley-Act, 1999)
    - Protecting personal information held by financial service institutions
  - COPPA (Children ‘s Online Privacy Protection Act, 1998)
    - Protecting information posted online by children under 13



HIPAA (1996)

HIPAA
Health Insurance Portability and Accountability Act

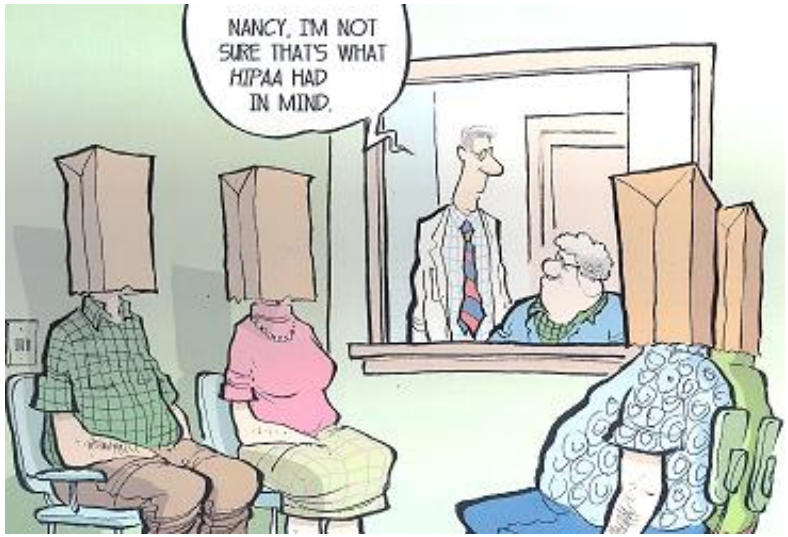
- Sets rules and limits on who can look at and receive your health information
- You have the right to:
  - Receive a notice that tells you how your health information may be used and shared
  - Decide if you want to give your permission before your health information can be used or shared for certain purposes
  - Get a report on when and why your health information was shared for certain purposes
- What Information is Protected?
  - Information in medical records
  - Conversations your doctor has had about your health
  - Information about you in your health insurer's system
  - Billing information from your clinic
- Your information can be used and shared
  - For your treatment and care coordination
  - To pay doctors and hospitals for your healthcare
  - With your family, relatives, friends or others you identify who are involved with your healthcare or your healthcare bills, unless you object
  - To protect the public's health, such as reporting when the flu is in your area
  - To make required reports to the police, such as reporting gunshot wounds



HIPAA

What is Protected?

- A major purpose of the rule is to define and limit the circumstances in which an individual's protected heath information may be used or disclosed by covered entities
- **Protected Health Information (PHI):** HIPAA protects all *individually identifiable health information*
- **De-Identified Health Information:** There are no restrictions on the use or disclosure of de-identified health information
  - De-identified health information neither identifies nor provides a reasonable basis to identify an individual
  - To de-identify information either:
    - A formal determination by a qualified statistician is required, or
    - The removal of specified identifiers of the individual and of the individual's relatives, household members, and employers is required



HIPAA Identifiers

- Some level of anonymization can be reached by removing all 18 HIPAA identifiers:
  - Names
  - Geographic subdivisions smaller than a state (except the first three digits of a zip code if the geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people and the initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000)
  - All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, and date of death and all ages over 89 and all elements of dates (including year) indicative of such age (except that such ages and elements may be aggregated into a single category of age 90 or older)
  - Telephone numbers
  - Fax numbers
  - Electronic mail addresses
  - Social security numbers
  - Medical record numbers
  - Health plan beneficiary numbers
  - Account numbers
  - Certificate/license numbers
  - Vehicle identifiers and serial numbers, including license plate numbers
  - Device identifiers and serial numbers
  - Web Universal Resource Locators (URLs)
  - Internet Protocol (IP) address numbers
  - Biometric identifiers, including finger and voice prints
  - Full face photographic images and any comparable images
  - Any other unique identifying number, characteristic, or code (excluding a random identifier code for the subject that is not related to or derived from any existing identifier)

Law is not Enough

- The law alone cannot protect privacy
- It is not self executing
- Worldwide, there are 5 major legal systems each with different definitions for privacy

Figure: http://www.droitcivil.uottawa.ca/world-legal-systems/eng-monde-large.html

TECHNICAL COUNTERMEASURES: PRIVACY-ENHANCING TECHNOLOGIES (PETS)

Technical Countermeasures

- Anonymization
  - Information theoretical privacy
- Cryptographic techniques
  - Computational privacy, based on cryptographic assumptions

Releasing Private Data

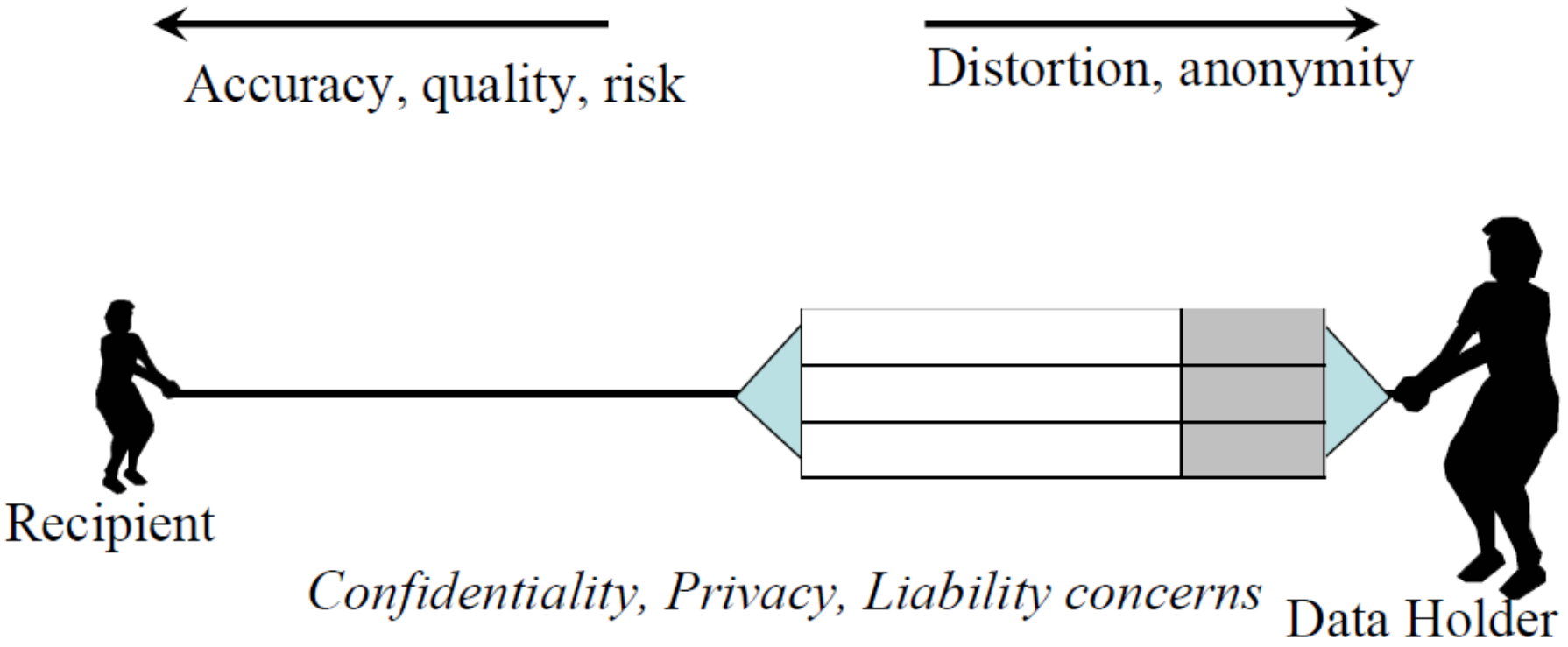
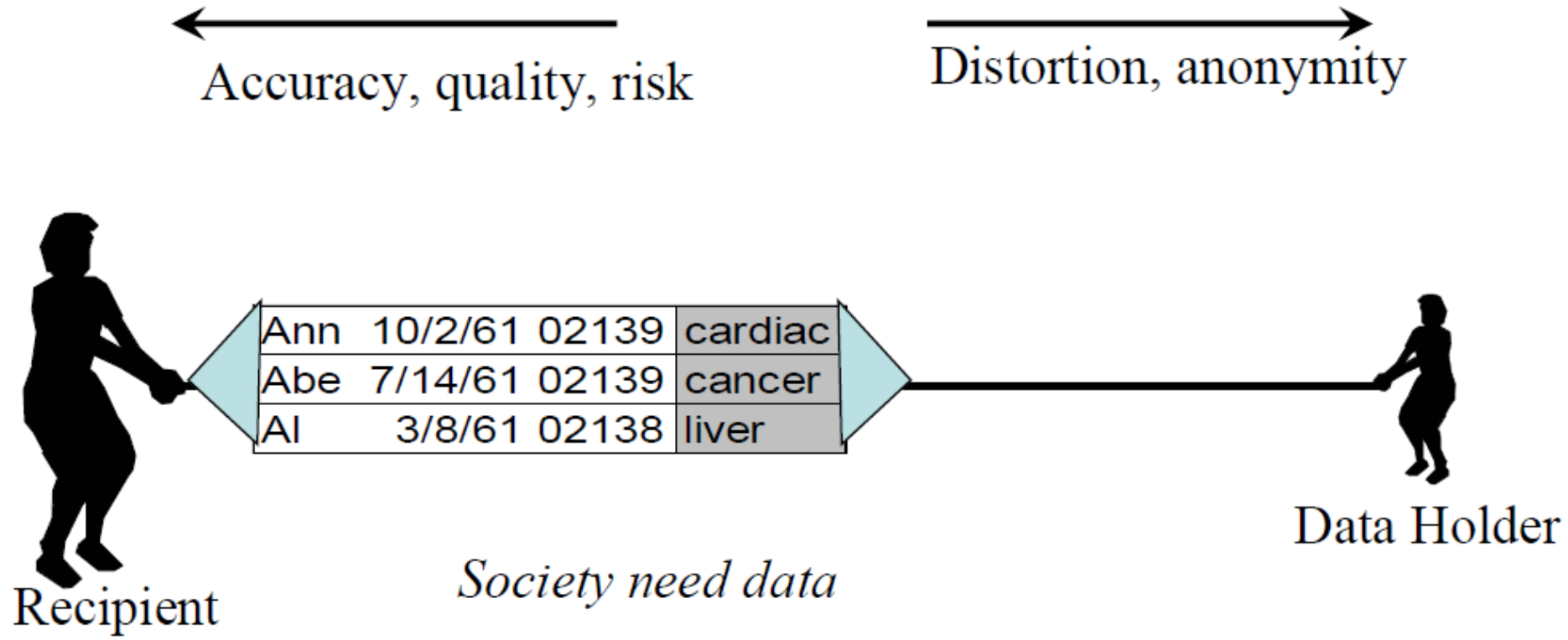
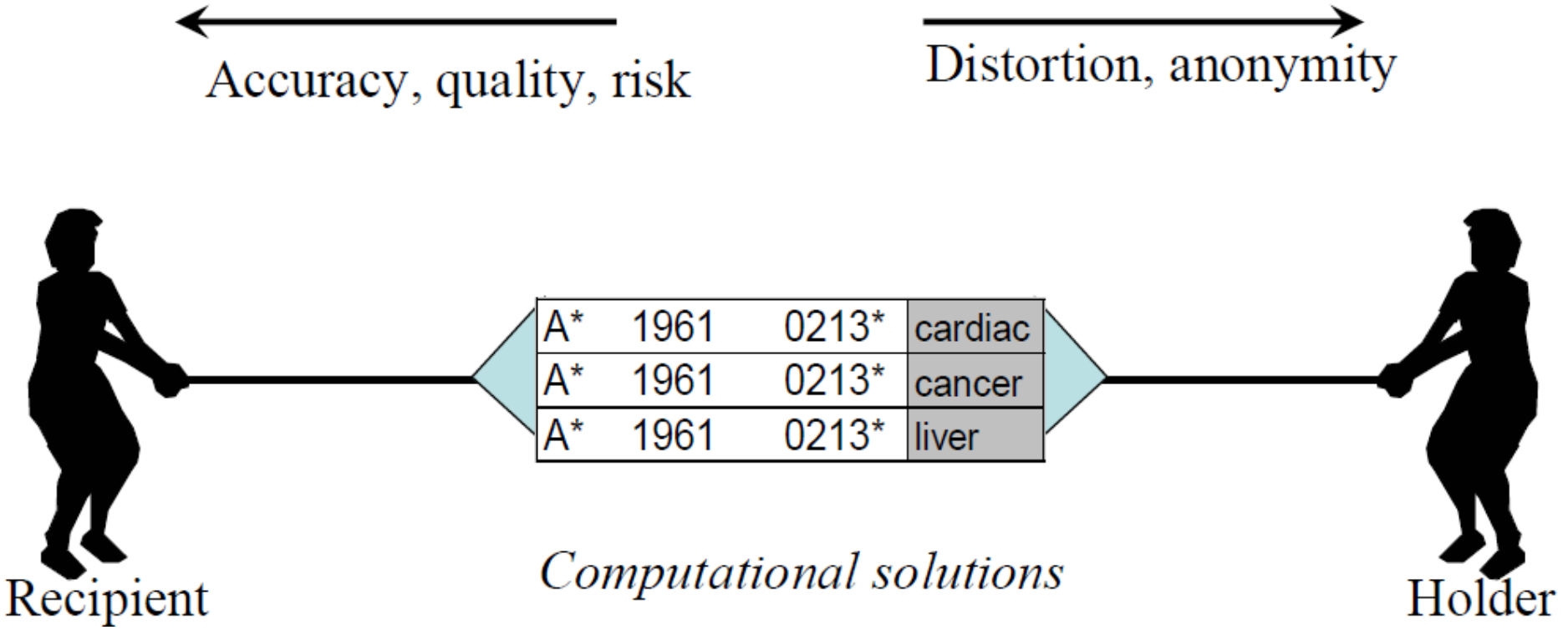


Figure: Latanya Sweeney

Privacy is Dead



Anonymization



Anonymization

- Removing or obscuring information from (electronic) traces that would allow direct or indirect identification of a person

Age	Gender	Zipcode
34	male	81667
45	female	81675
66	male	81925
70	female	81931

Age	Gender	Zipcode
<50	$\star$	816$\star\star$
<50	$\star$	816$\star\star$
≥ 50	$\star$	819$\star\star$
≥ 50	$\star$	819$\star\star$

- Advantage:
  - Allows research that would otherwise not be possible due to privacy concerns
- Major misconception:
  - Governments, industry and researchers tend to claim that effective anonymisation of personal data is possible and can help society to ensure the availability of rich data resources whilst protecting individuals' privacy
- Unfortunately, this is simply not the case
  - Netflix, MA Governor medical records, DNA
- Pseudonymisation (replacing the name and other direct identifiers with a new identifier) does not solve this problem either

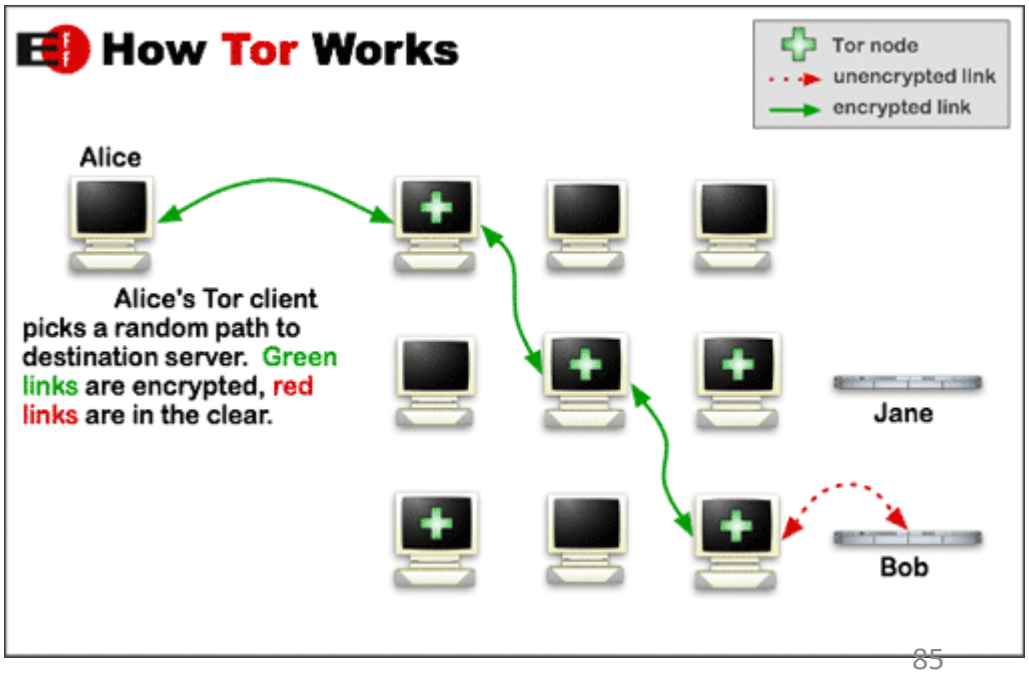
Cryptographic Mechanisms for Privacy Protection

- Anonymous communication
  - Tor
- Anonymous credentials
- Blind signatures
- Secure multiparty computation
  - Garbled circuits
- Searchable encryption
- Deterministic encryption
  - Order-preserving encryption
- Computing on encrypted data
  - Homomorphic encryption
  - Functional encryption
- Oblivious RAM
- Private information retrieval
- Zero-knowledge proofs
- Etc.

Anonymous Communication

- Anonymity of participants is usually achieved by special routing overlay networks that hide the physical location of each node from other participants
- More on this later

Enables online anonymity and censorship resistance



https://www.torproject.org/about/overview.html.en

Anonymous Credentials (1)

- Allow users to authenticate themselves in a privacy-preserving manner
- Traditional credentials:
  - Alice obtains credentials from an organization
  - At some later point, she proves to the organization (or any other party) that she has been given appropriate credentials
- Anonymous credentials:
  - Alice can do the same without revealing anything else about her identity
  - If she uses her credentials a second time, no one will be able to tell that the two interactions involved the same user

Anonymous Credentials (2)

Organization

Alice

Service

"I have a cred from Org saying
WA resident
Age >18"

Cred from Org
Name Alice
Address
Birthdate
Birthplace
Citizenship
…

- Cannot
  - Identify Alice
  - Learn anything beyond the info she gives
  - Distinguish two users with the same attributes
  - Link multiple uses of the same credentials

Figure: Melissa Chase – Microsoft Research

Blind Signatures

- Content of a message is blinded before it is signed
- Resulting blind signature can be publicly verified against the original (unblinded) message
- Cryptographic voting systems
  - Authority checks the credentials of the voter to ensure that he is allowed to vote, and that he is not submitting more than one vote
  - Authority does not learn the voter's selections

Secure Multiparty Computation

- Alice and Bob compute a function of their private data, without exposing anything about their data besides the result

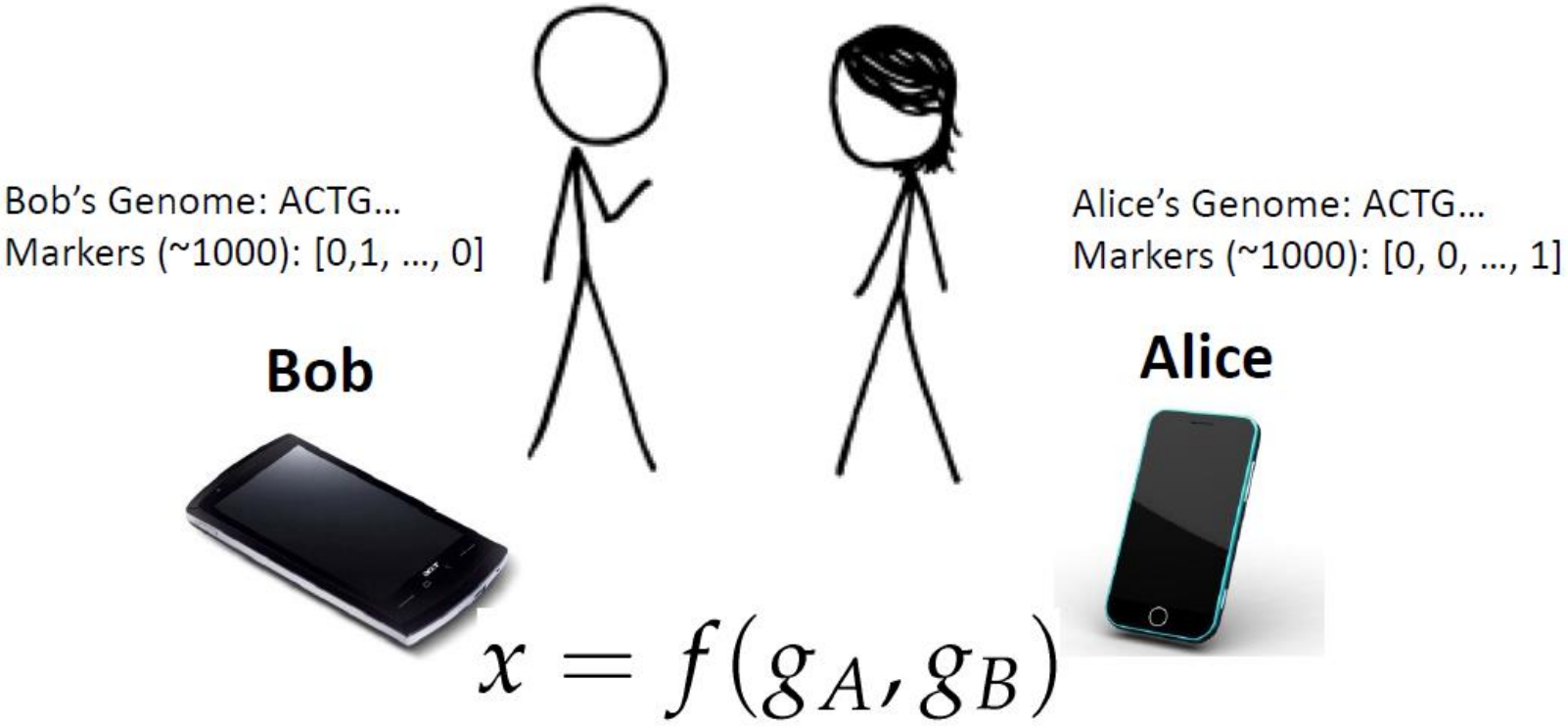
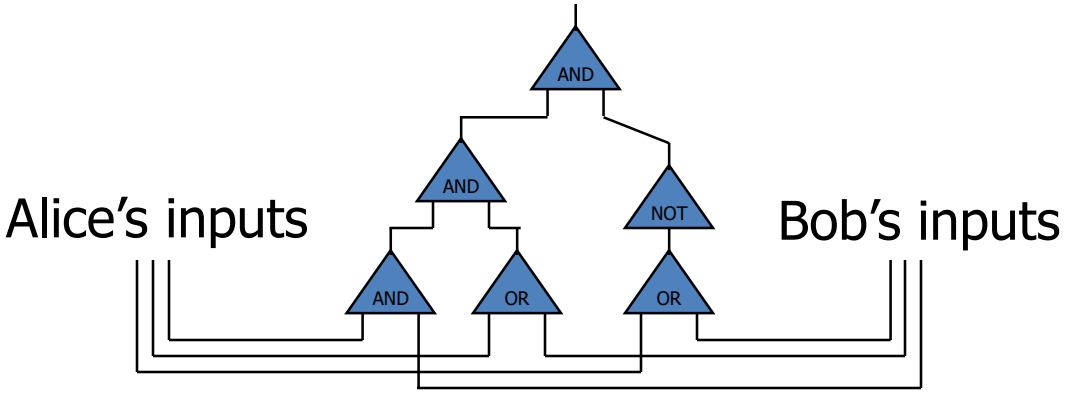


Figure: David Evans et al.

Garbled Circuits

- Bob creates a garbled circuit and sends the circuit to Alice
- Alice evaluates the circuit with her inputs and returns the result to Bob
- The result of the circuit evaluation with Alice's inputs is the output of the function Alice and Bob wish to compute
- Bob does not send his inputs to Alice, instead his inputs are encoded into the garbled circuit such that Alice cannot determine what they are



Searchable Encryption

- Allows a party to outsource the storage of his data to another party in a private manner, while maintaining the ability to selectively search over it

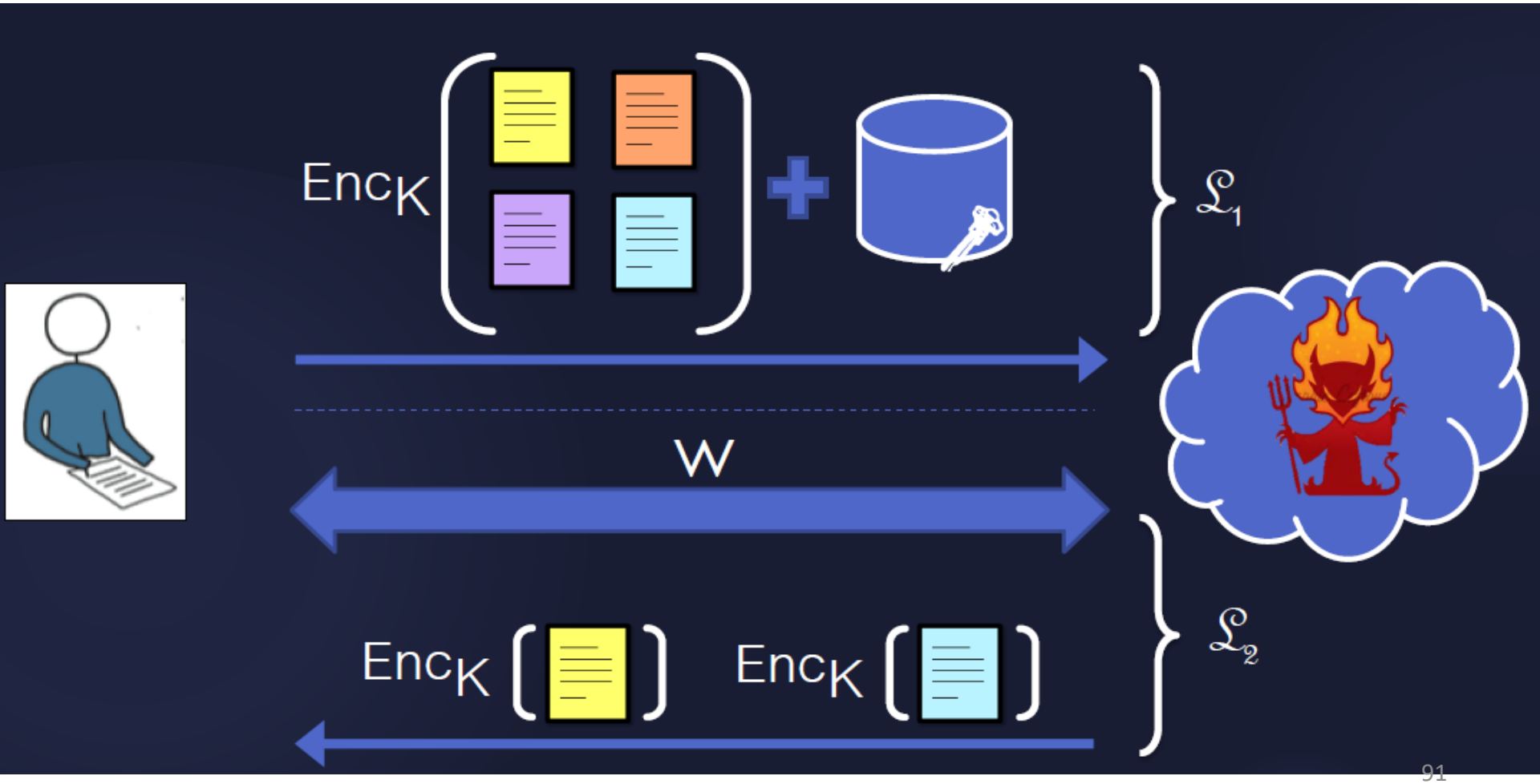
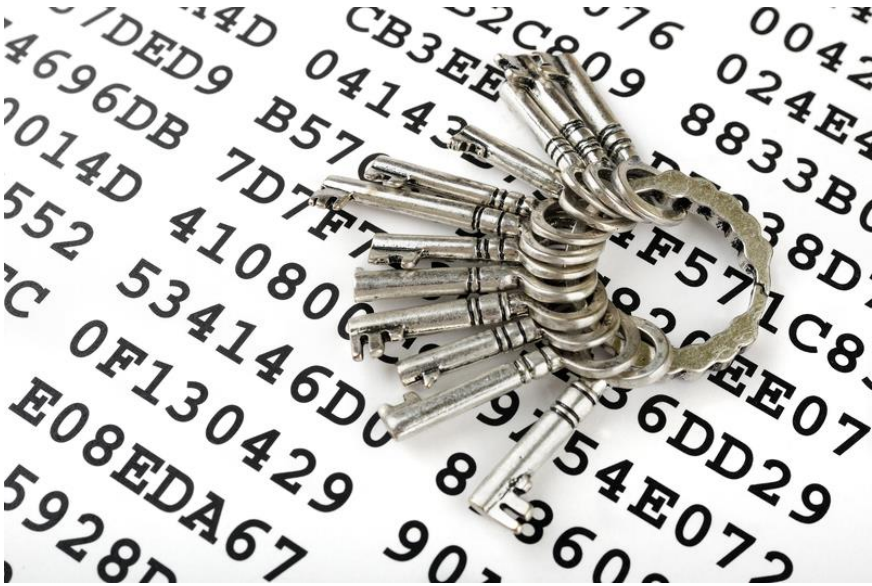


Figure: Seny Kamara

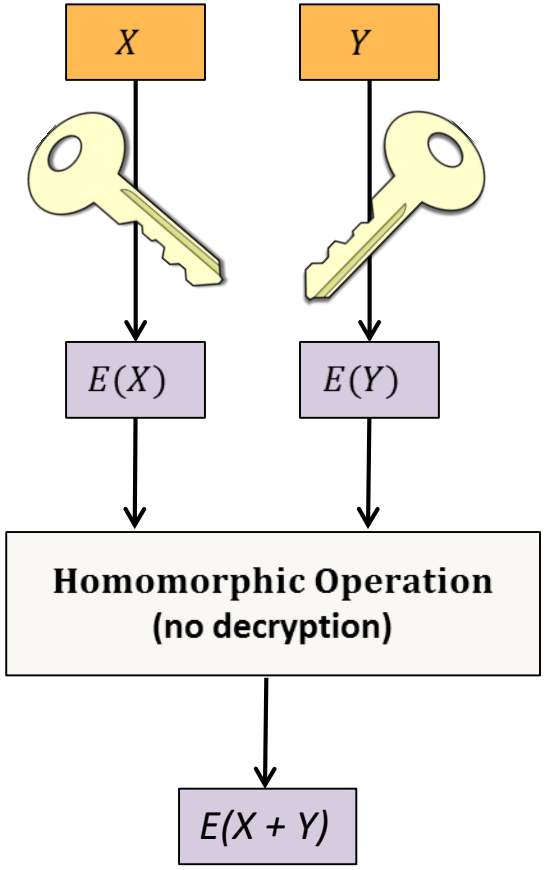
Deterministic Encryption

- Always produces the same ciphertext for a given plaintext and key
  - It is efficient in searching of encrypted data
- Order-preserving encryption
  - M>N → E(M)>E(N)
  - Can leak information to an eavesdropper, who may recognize known ciphertexts
  - Does not guarantee what is known as “semantic security”

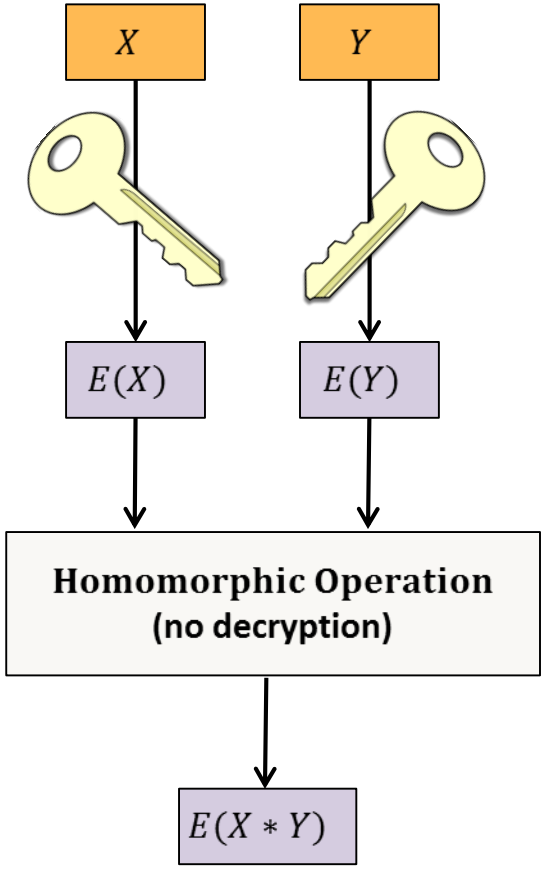


Homomorphic Encryption

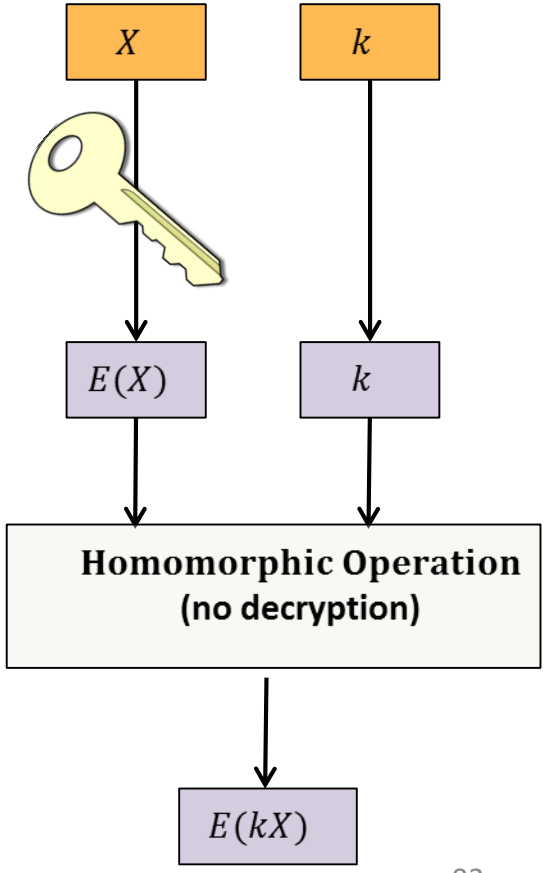
- Allows specific types of computations to be carried out on ciphertext



Benaloh, Paillier



Unpadded RSA, ElGamal



Functional Encryption

- Possessing a secret key allows one to learn a function of the ciphertext. Examples:
  - Attribute-based encryption
  - Identity-based encryption

The email recipient, who has a master secret key sk, gives a spam-filtering service a key sk[*f*] for the functionality *f*; this *f* satisfies $f(x) = 1$ whenever message *x* is marked as spam by a specific spam predicate, otherwise $f(x) = 0$. A sender encrypts an email message *x* to the recipient, but the spam filter blocks the message if it is spam. The spam filter learns nothing else about the contents of the message.

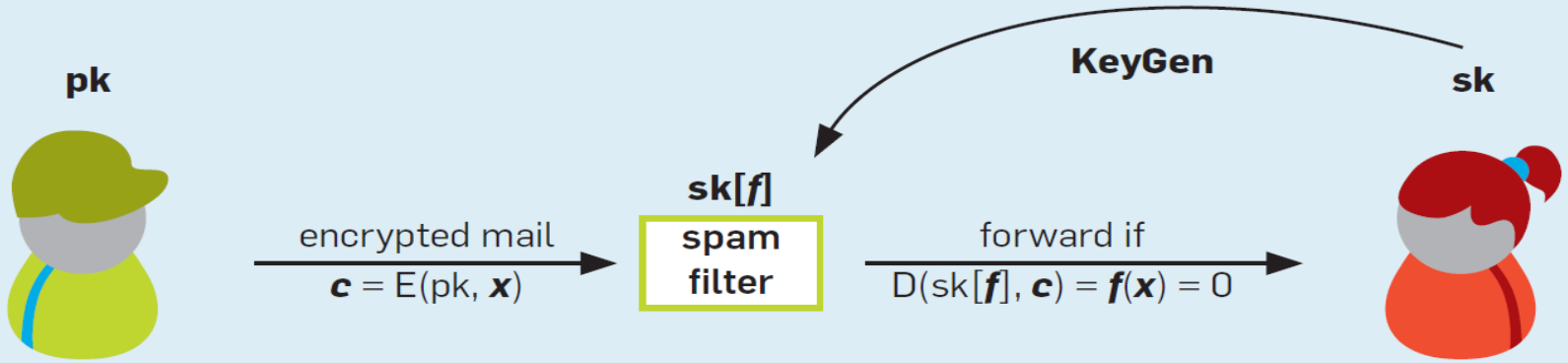


Figure: Dan Boneh et al.

Oblivious RAM

- Client outsources the storage of his data to a cloud
- Client stores only a small amount of data locally
- Client accesses (read/write) his data while hiding the identities of the items being accessed
- More about this later

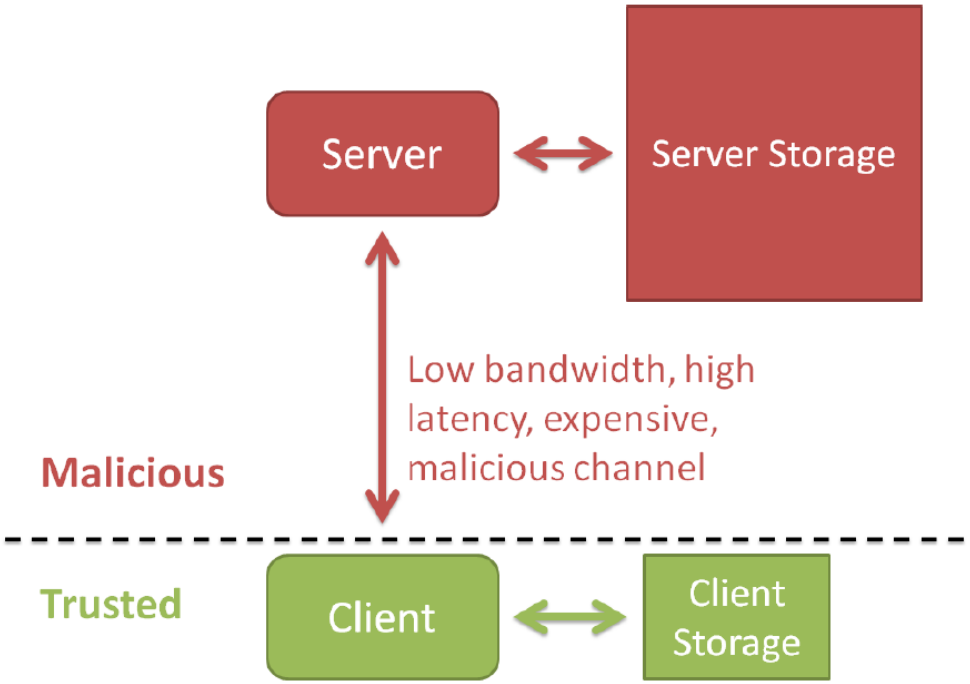
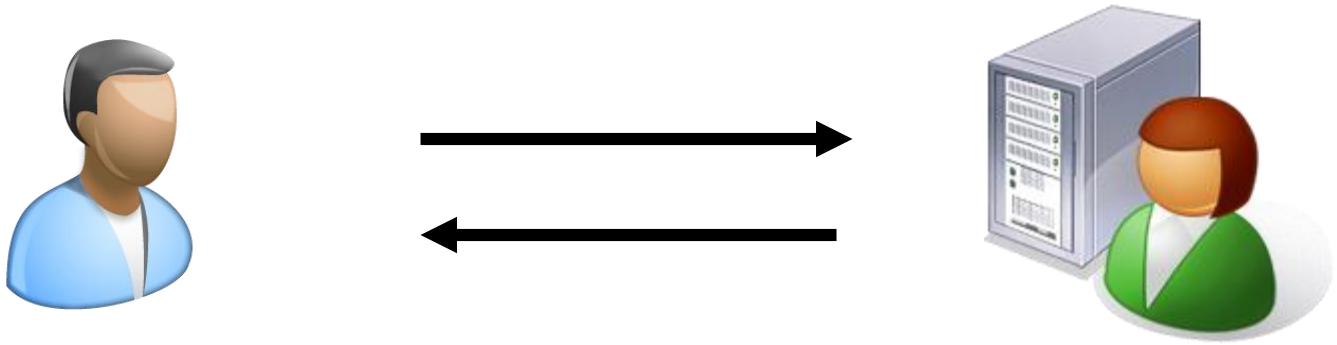


Figure: Emil Stefanov et al.

Private Information Retrieval

- Allows a user to retrieve an item from a server in possession of a database without revealing which item is retrieved
- More about this later



User U

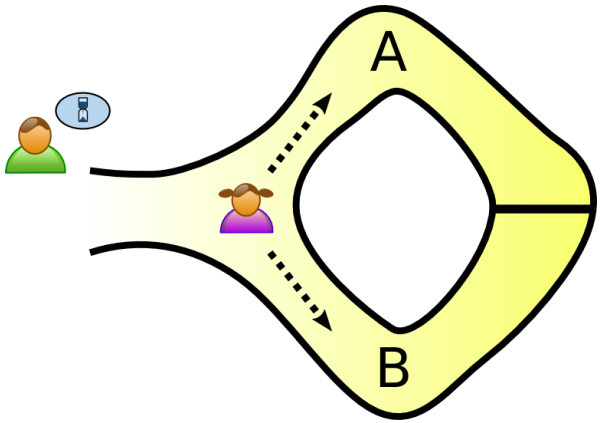
wants to retrieve some data from D

Database D

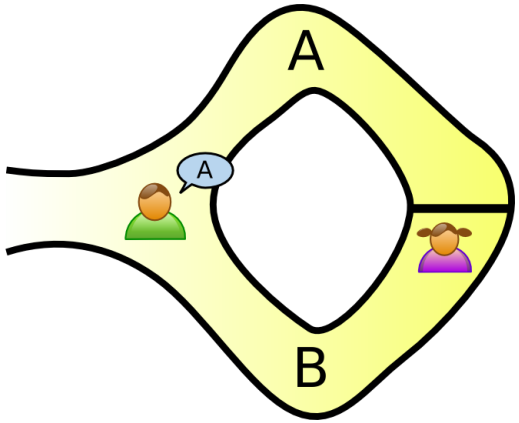
shouldn't learn what U retrieved

Zero-knowledge proofs

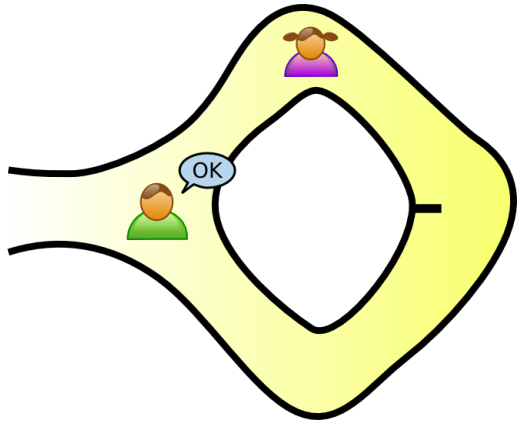
- Peggy has uncovered the secret word used to open a magic door in a cave
- Victor wants to know whether Peggy knows the secret word; but Peggy, being a very private person, does not want to reveal the fact of her knowledge to the world in general



Peggy randomly takes either path A or B, while Victor waits outside



Victor chooses an exit path



Peggy reliably appears at the exit Victor names

Figure: Wikipedia

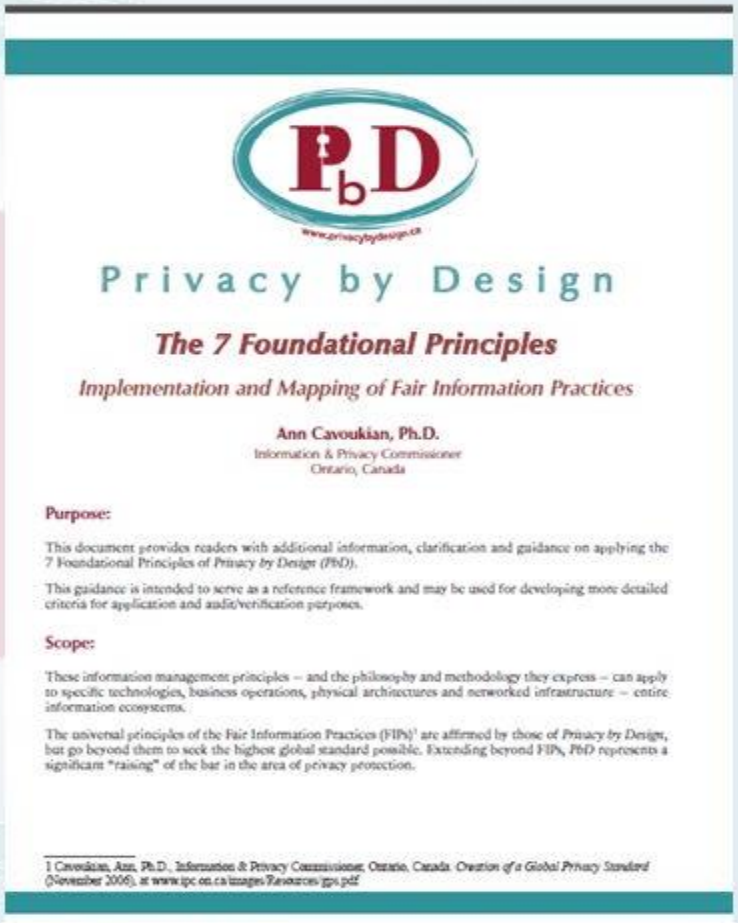
PRIVACY BY DESIGN

Privacy by Design (PbD) - Motivation

- To facilitate a flourishing Internet economy, consumers need to be able to trust the services they use online
- They should not need to worry about their giving companies more data than necessary for the service being used
- It is crucial to ensure that privacy protections are built into the design and implementation of the products and services

7 Foundational Principles of PbD

1. ***Proactive*** not Reactive: Preventative, not Remedial
2. Privacy as the ***Default***
3. Privacy ***Embedded*** into Design
4. *Full* Functionality: ***Positive-Sum***, not Zero-Sum
5. End-to-End ***Security***: Full Lifecycle Protection
6. Visibility and Transparency: ***Keep it Open***
7. Respect for User Privacy: Keep it ***User-Centric***



PbD

www.privacybydesign.ca

Privacy by Design

The 7 Foundational Principles

Implementation and Mapping of Fair Information Practices

Ann Cavoukian, Ph.D.
Information & Privacy Commissioner
Ontario, Canada

Purpose:

This document provides readers with additional information, clarification and guidance on applying the 7 Foundational Principles of *Privacy by Design (PbD)*.

This guidance is intended to serve as a reference framework and may be used for developing more detailed criteria for application and audit/verification purposes.

Scope:

These information management principles – and the philosophy and methodology they express – can apply to specific technologies, business operations, physical architectures and networked infrastructure – entire information ecosystems.

The universal principles of the Fair Information Practices (FIPs)[1] are affirmed by those of *Privacy by Design*, but go beyond them to seek the highest global standard possible. Extending beyond FIPs, *PbD* represents a significant "raising" of the bar in the area of privacy protection.

1 Cavoukian, Ann, Ph.D., Information & Privacy Commissioner, Ontario, Canada. *Creation of a Global Privacy Standard* (November 2006), at www.ipc.on.ca/images/Resources/gps.pdf

7 Foundational Principles of PbD

- **Proactive** not Reactive; **Preventative** not Remedial
  - PbD anticipates and prevents privacy invasive events before they happen
- Privacy as the **Default Setting**
  - Privacy is built into the system, by default
- Privacy **Embedded** into Design
  - Privacy is integral to the system, without diminishing functionality
- Full Functionality — **Positive-Sum**, not Zero-Sum
  - PbD seeks to accommodate all legitimate interests and objectives in a positive-sum "win-win" manner
  - Avoids misconceptions such as privacy vs. security
- End-to-End Security — **Full Lifecycle Protection**
  - All data are securely retained, and then securely destroyed at the end of the process, in a timely fashion
- **Visibility** and **Transparency** — Keep it **Open**
  - Its component parts and operations remain visible and transparent, to users and providers
- **Respect** for User Privacy — Keep it **User-Centric**

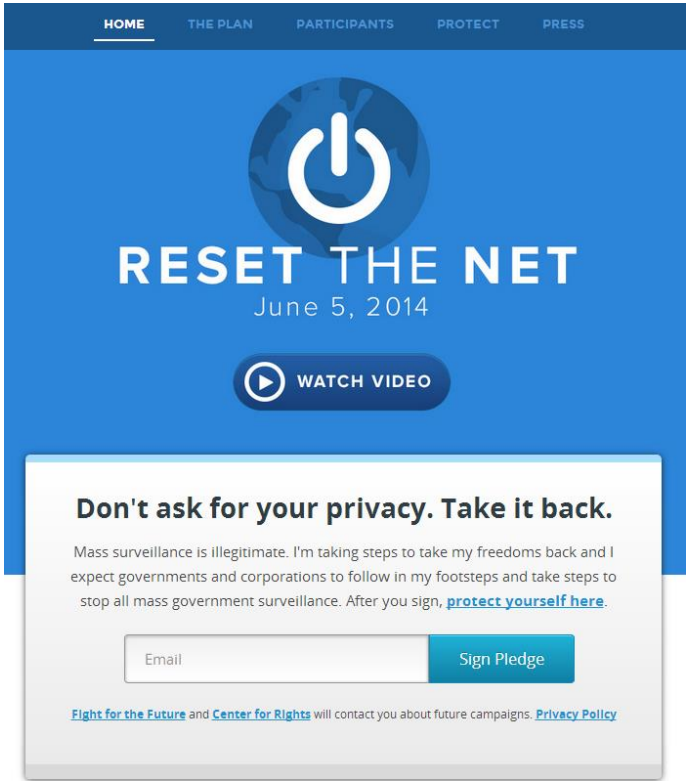
Privacy by Design and Default

- PbD means controllers of data take a positive approach to protecting privacy, by embedding it into both technology and into their organizational policies
- When such protections are built-in from the beginning, they can help to prevent invasions of privacy rights
- *Privacy by default*: when a user receives a product or service, privacy settings should be as strict as possible, without the user having to change them
- Sharing does not inherently mean an end to privacy
  - With effective privacy by design and by default, you can have both

PbD - Examples

Founded in summer 2013 at CERN by scientists who were drawn together by a shared vision of a more secure and private Internet

Ixquick does *not* collect or share *any* personal information



Goal: provide guidance for your online protection; supported by the Electronic Frontier Foundation

Google announced its support for the campaign and published the source code for an "end-to-end" encryption service for its Chrome browser

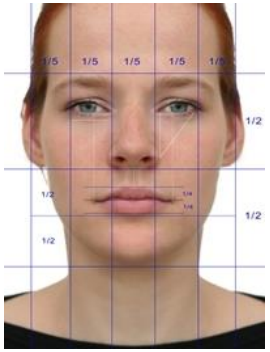
THE FUTURE OF PRIVACY

Some Modern Observation Tools

Cellular phones

Web browsing

Online Social Networks



Face recognition

Genomic data

Challenges for Privacy

- Big databases
  - Data is never deleted, it is sometimes sold, and may be used in different contexts
  - Re-identification is becoming more and more feasible as the data is becoming detailed
- Data aggregation
  - Data is increasingly aggregated, collected and matched from multiple sources
  - Cross layer attacks
- Social networks

Privacy is rapidly becoming a collective value in that technology and market forces are making it hard for any one person to have privacy without all persons having a similar minimum level of privacy

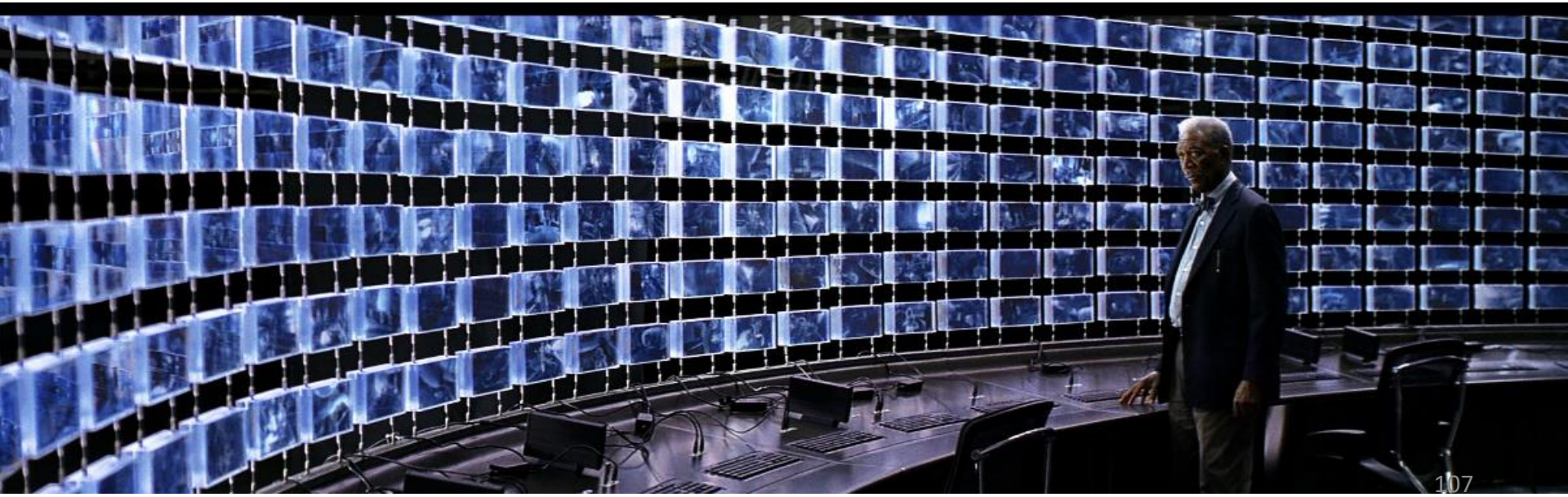
- Priscilla Regan, 1995

Analogy: road safety

- Location-based services

Surveillance

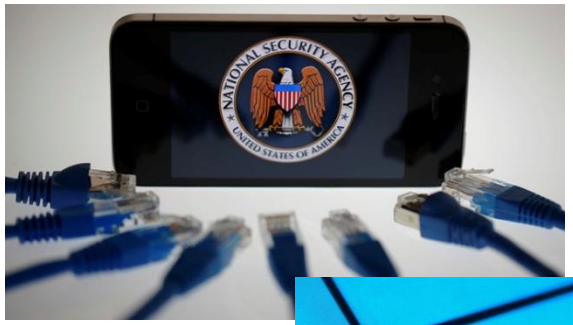
- The pervasiveness of computers has resulted in the almost constant surveillance of everyone
- Corporations and the police are both using this new trove of surveillance data
- Today's surveillance: *wholesale surveillance*
  - Previously: “follow that car”
  - Now: “follow every car”
- Tomorrow’s surveillance: even more extensive



https://www.schneier.com/blog/archives/2006/03/the_future_of_p.html

Wholesale Surveillance

- NSA can eavesdrop on most phone calls and read most e-mails
- Most of our browsing, search, and purchases are recorded
- EZ Pass can regularly record the location of our car
- Video cameras capture our images several times a day



Surveillance and Technology

- As computer memory becomes cheaper, more and more of these electronic footprints are being saved
  - 100 megabytes: to record everything the fastest typist input to his computer in a year
  - 4 to 8 gigabytes: to record everything the average user does on the Internet in a year
  - 5 gigabytes: to save the yearly phone calls of a typical person who uses 500 cell phone minutes a month
  - 200 gigabytes: to constantly record all audio of an individual per year
  - 700 gigabytes: to constantly record all video of an individual per year (life recorder)
- As processing becomes cheaper, more and more of it is being cross-indexed and correlated
- There are companies in the business of buying and reselling customer databases

Drones are Coming

- The Federal Aviation Administration (US) is relaxing its restrictions around the domestic use of “unmanned aerial systems,” leading to greater use of drones by public agencies and the private sector

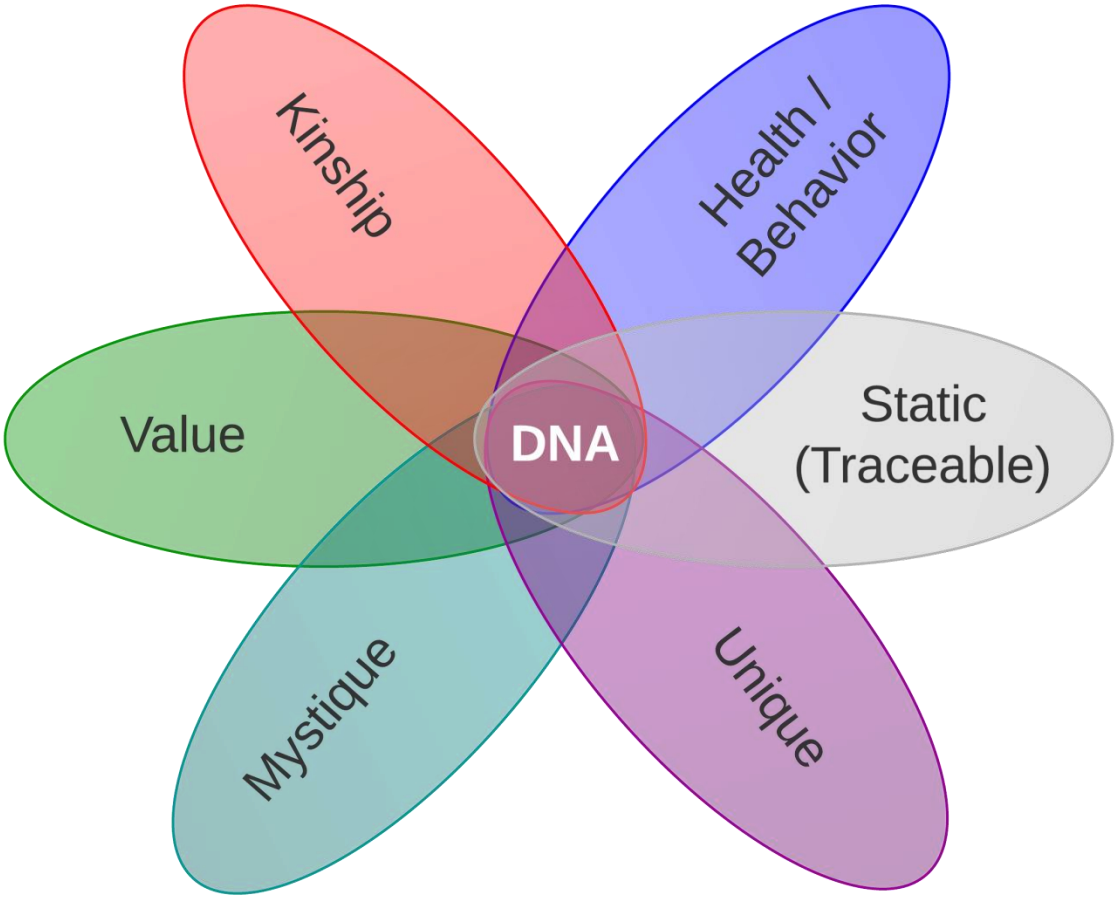
Drones are Coming

- Congress authorized the FAA to open the nation's airspace to widespread drone flights by 2015
- The FAA estimates that more than 7,000 civilian drones could be surfing the sky by 2020

New Era of Surveillance

- (Online mass surveillance: PRISM)
  - Trivial solution: do not go online?
- Supreme Court rejects GPS tracking
  - Police need a warrant before affixing a GPS device to a car and following a suspect for a prolonged period (United States v. Jones)
  - The FBI reportedly turned off thousands of GPS devices in response to the ruling
- But,
  - Drones can follow a car without the need to attach anything
  - Law enforcement can equip a drone with thermal or chemical sensors and let it loose to roam a neighborhood in search of invisible infractions such as indoor marijuana cultivation

Another New Era: Genomic Privacy

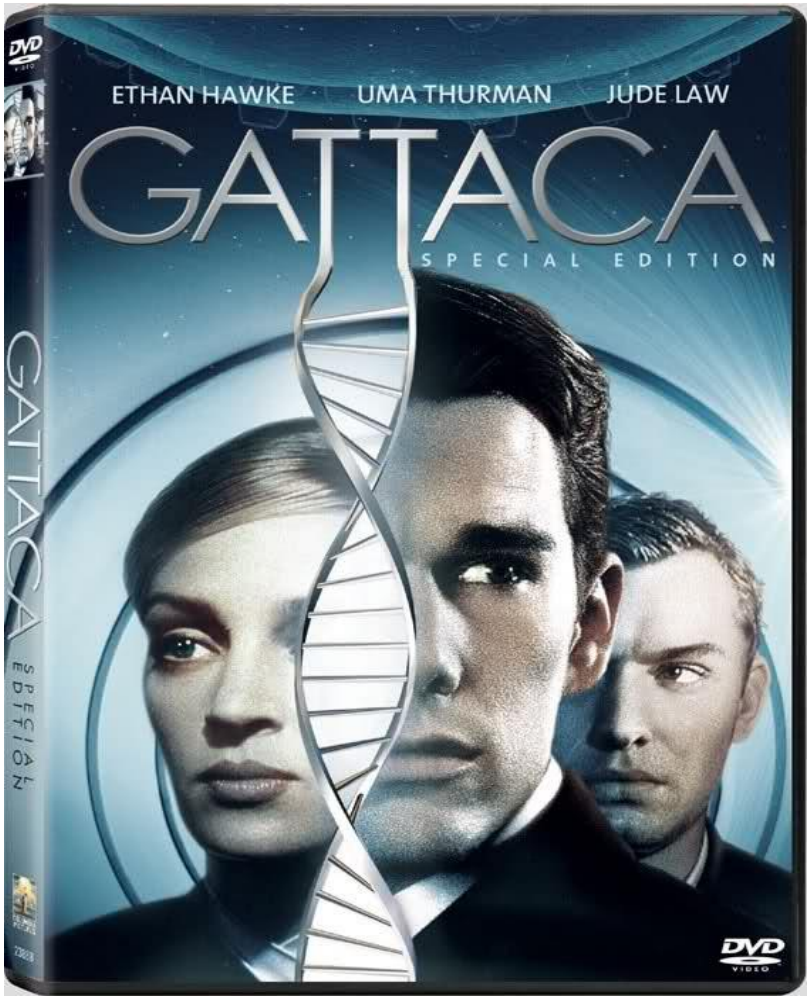


Why Protect Genomic Data?

- Genome carries information about a person's genetic condition and predispositions to specific diseases
  - Leakage of such information could cause *genetic discrimination*
  - Denial of access to health insurance, mortgage, education, and employment

Why Protect Genomic Data?

- Genome carries information about a person's genetic condition and predispositions to specific diseases
  - Leakage of such information could cause *genetic discrimination*
  - Denial of access to health insurance, mortgage, education, and employment



Why Protect Genomic Data?

- Genome carries information about a person's genetic condition and predispositions to specific diseases
  - Leakage of such information could cause *genetic discrimination*
  - Denial of access to health insurance, mortgage, education, and employment
- Anonymisation is ineffective

Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

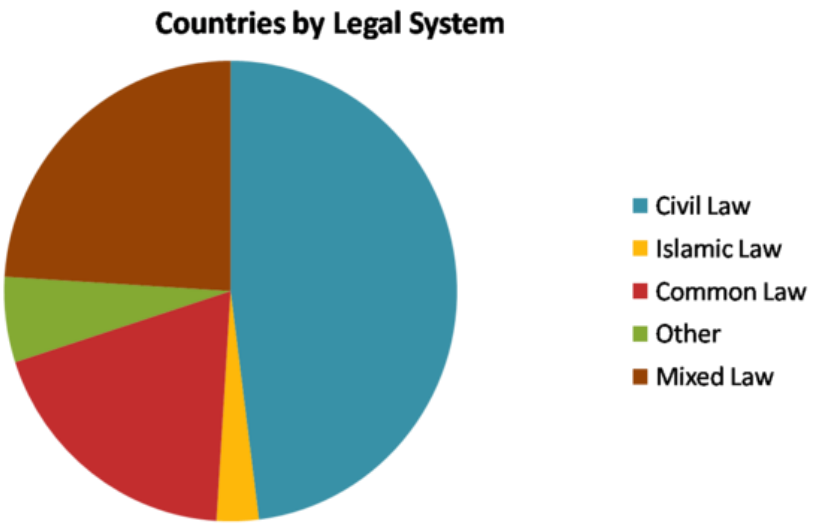
M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich. *Identifying personal genomes by surname inference*. Science: 339 (6117), Jan. 2013.

Why Protect Genomic Data?

- Genome carries information about a person's genetic condition and predispositions to specific diseases
  - Leakage of such information could cause *genetic discrimination*
  - Denial of access to health insurance, mortgage, education, and employment
- Anonymisation is ineffective
- Genome carries information about family members
  - Cross-layer attacks
    - Using privacy-sensitive information belonging to a victim retrieved from different sources

Why Protect Genomic Data?

- Genome carries information about a person's genetic condition and predispositions to specific diseases
  - Leakage of such information could cause *genetic discrimination*
  - Denial of access to health insurance, mortgage, education, and employment
- Anonymisation is ineffective
- Genome carries information about family members
  - Cross-layer attacks
    - Using privacy-sensitive information belonging to a victim retrieved from different sources
- Genomic data is non-revocable
- Law is not universal; it is hard to enforce



Other Concerns

- A recent poll, conducted by market-research firm Toluna, found 72% of Americans cited privacy concerns as the biggest reason for not wanting to wear the Glass
- Microsoft had to fully detail what the Xbox One Kinect sensor sees and sends
  - A new privacy statement page gets into what Kinect data is collected and how it will be used

http://mashable.com/2014/04/07/google-glass-privacy/

Computer Scientists Must Save the World [1]

- **Policy** is limited by that which can be expressed in words
- **IT** can provide glue technology, but heavily relies on existing technology
- Laws can change and **lawyers** often lack understanding of ways technology will continue to change
- **Computer scientists** construct tomorrow's machines, and can do so with privacy as part of their problem definition, so that new technology can be deployed and easily adopted

[1] Latanya Sweeney