J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

Data Security and Privacy: Key Concepts

Micah Altman
Director of Research
MIT Libraries

Goals for this Lecture

- Recognize the importance of information privacy and security in research
- Define key technical concepts in privacy and data security
- Design a an information privacy and security plan for your research project
- Select key protections across the research lifecycle

Lecture Overview

- Changing Landscape of Research Information
- Technical Concepts & Terminology
- Designing a Data Security and Privacy Plan
- Selecting and Applying Information Controls

Changing Landscape of Research Information

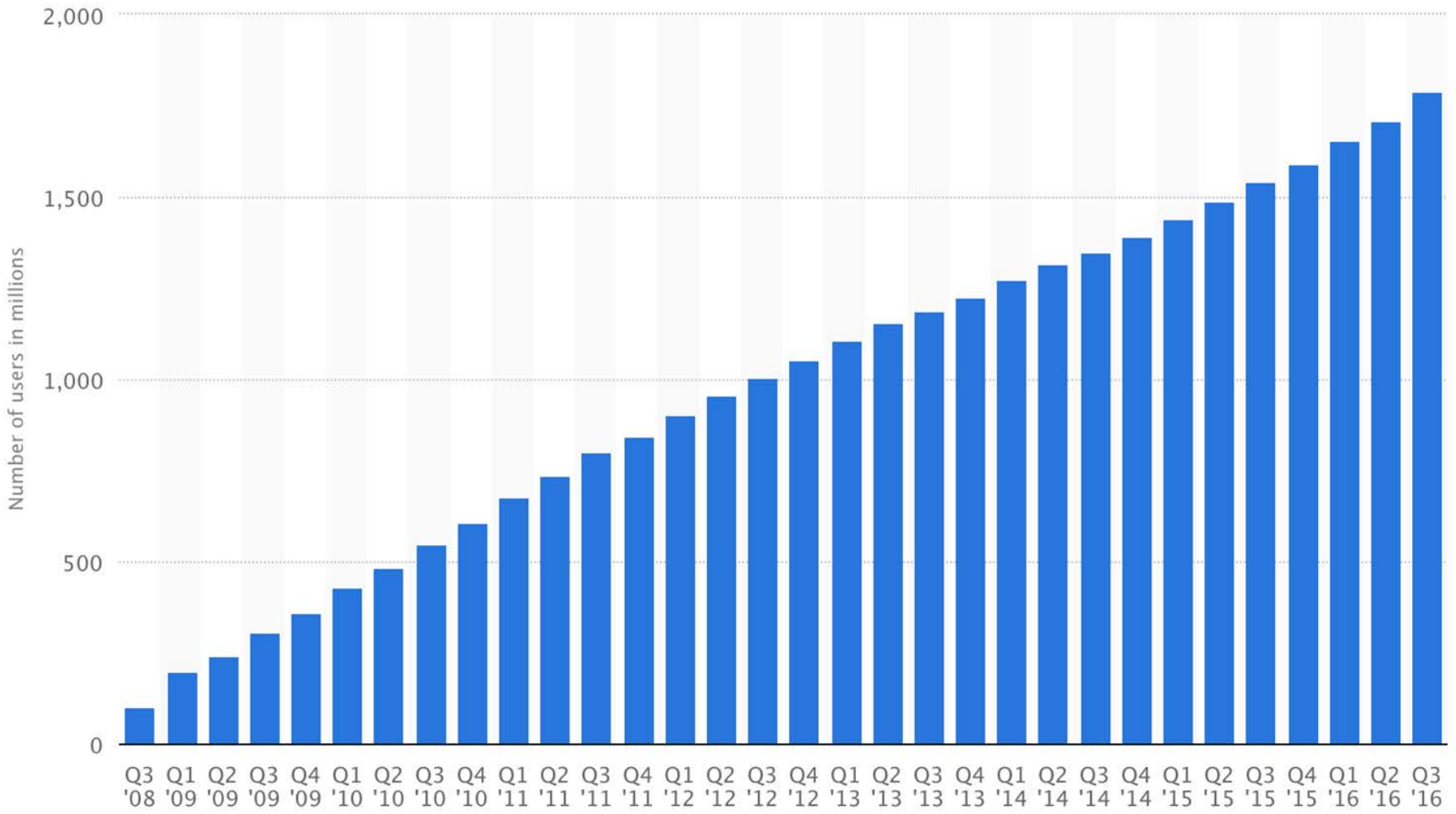
Information is Everywhere

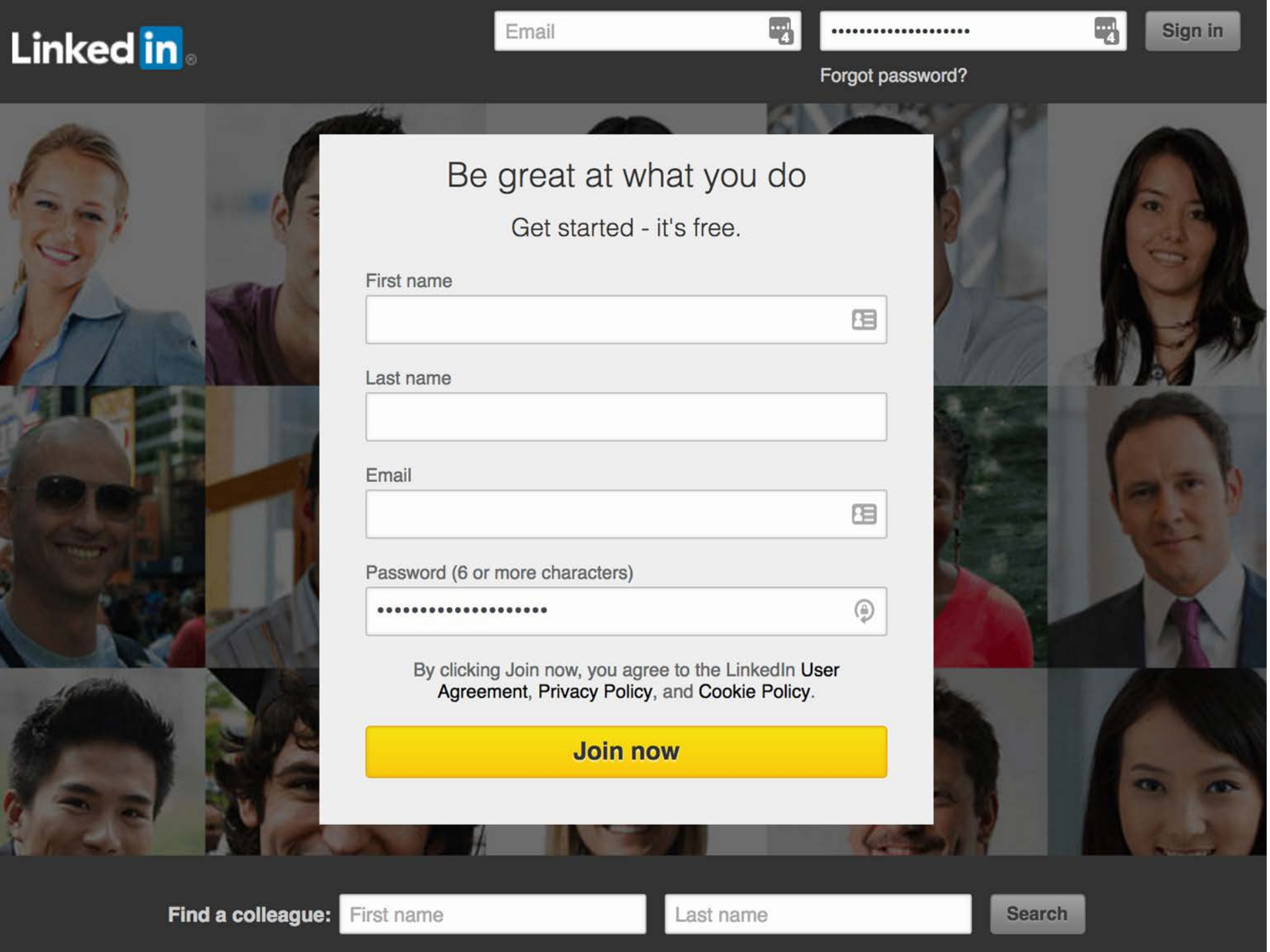
We all have information we care about

Information about:

- what we do
- what we say
- what we think
- where we go
- our history
- our health
- our property
- relationships with people
- relationships with organizations

Number of monthly active Facebook users worldwide as of 3rd quarter 2016 (in millions)

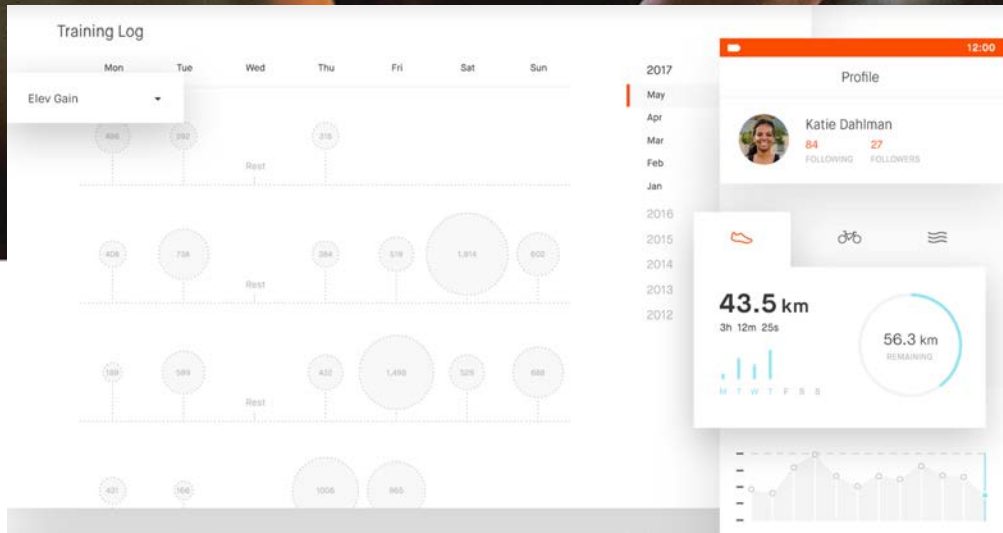



Linked in
Email
Sign in
Forgot password?
Be great at what you do
Get started - it's free.
First name
Last name
Email
Password (6 or more characters)
By clicking Join now, you agree to the LinkedIn User Agreement, Privacy Policy, and Cookie Policy.
Join now
Find a colleague:
First name
Last name
Search

STRAVA

Features Premium

Connecting the world's athletes.




3G
9:41 AM
Google
42 minutes to Work
Light traffic on US-101 S
San Francisco
Oakland
Hayward
Daly City
Livermore
Fremont
San Mateo
Palo Alto
Get directions
United Air Flight 217
Delayed (48 minutes)
SFO
PIT
Departed San Francisco, today
Scheduled 4:06 PM
Terminal
Gate
4:54 PM
3
85
Arrives Pittsburgh, tomorrow
Scheduled 11:58 PM
Terminal
Gate
12:46 AM
-
C59
Final, April 11
Fishery Field, Chicago
Clovers 7
Blaze 6
1 2 3 4 5 6 7 8 9 R H E
Clovers 0 0 0 4 3 0 0 0 0 7 14 4
Blaze 0 1 4 0 0 0 0 1 0 6 19 1
San Francisco
68°
CLEAR
13mph
0%
WED
THU
FRI
SAT
3G
9:41 AM
Google
Barbacco Restaurant
220 California St, San Francisco
Reservation in 1 hour
Journey time walking 30 minutes
Get directions
View email

FIND MUGSHOTS

INSTANT MUGSHOT SEARCH!

First Name | Last Name | Select State

ATTENTION: REPORTS MAY CONTAIN GRAPHIC IMAGES AND SHOCKING DETAILS

SEARCH

WHAT YOU'RE GETTING

FINDMUGSHOTS.COM MAY UNCOVER SHOCKING TRUTH ABOUT THE PERSON YOU ARE SEARCHING

MUGSHOTS

Mugshots of all times of the person you are searching for. New and Historical

POLICE REPORT

Police report related to the arrest report.

COURT RECORDS

Court Records related to the specific arrest.

FREE ACCESS

Free and unrestricted access to billions of mugshots and records.

Researchers are Responsible for Research Information

Informational Harms

Informational harms can occur when others use research results or data; learn about subjects as a result of their participation; and then violate the subjects rights or negatively affect the subjects interests

Information Privacy & Confidentiality

Information Privacy & Confidentiality denotes broadly the interests that individuals and groups have in controlling information about or from them

Privacy as a Human Right

- Many countries recognize privacy as a universal human right

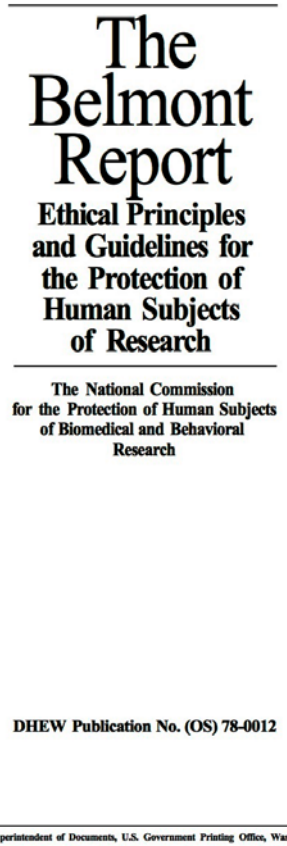
- European Convention on Human Rights, Article 12:

  "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation."

- EU Right to be forgotten

Researchers Have Ethical Responsibilities

- Information is a core concern of research
- Research information is supported by individuals and society as a public good
- Researchers have ethical responsibility for information they collect and share



Legal Risks and Requirements

- Researchers should treat law as a supplement to ethical responsibilities

- Researchers should incorporate data security & privacy planning during research design to minimize legal and ethical risks for you, your institution and subjects

Scope of Applicable Laws

- Many different laws protect information privacy and security in different contexts
- Laws place requirements on researchers, their institutions, and later use of research
- Laws may be triggered by a variety of characteristics

Trend: Privacy in the Law

- Increasing numbers of laws at state, national, and international level regulating private information
- Increasing recognition of need for data protection, redaction, ongoing controls
- Increasing recognition that de-identification is not a silver bullet solution
- Increasing recognition of more sophisticated scientific privacy measures

Trend: Privacy in the Spotlight

- Recognition that large amounts of information about people is publicly available
- Breaches and misuse of information is publicized widely – frequent media attention
- Individuals are increasing concerned about treatment of their information across sectors
- Vast majority of consumers agree consumers that they have lost control of how personal information is collected and used by companies.

Researcher Responsibilities

- Researchers should understand individuals rights and interests and anticipate informational harms that may result from a subject's participation in research

- Researchers address potential harm through research design, data management planning, selection and use of protections

Protecting Information is Increasingly Challenging

It’s Challenging to Protect Information

- Cyberattacks are increasingly common and sophisticated

- Information shared carelessly can be rapidly and broadly distributed

- Popular platforms for information collection, storage, and analysis may expose research to additional threats

The EU Safe Harbor Agreement Is Dead

Facebook emotion study breached ethical guidelines, researchers say

Half of Americans expect to lose money to identity theft

France fines Google over 'right to be forgotten'

Mugshot websites: Free speech or extortion?

Mexico's Entire Voter Database Made Accessible on Internet

Challenges of Data Breaches

U.S. personnel management hack preventable, congressional probe finds

News › Business › Business News

Yahoo hack: World's biggest data breach could compromise Verizon deal and cost hundreds of millions of dollars

Challenges of Meaningful Consent

Boston College's Secret Tapes Could Bring IRA Exposure and Retribution

A court has ruled Boston College's taped interviews of dozens of ex-IRA members who were promised anonymity must be turned over to Northern Ireland police—exposing participants to retribution and perhaps even death.

THE CHRONICLE OF HIGHER EDUCATION

Harvard's Privacy Meltdown

Social-network project shows promise and peril of doing social science online

Reading the Privacy Policies You Encounter in a Year Would Take 76 Work Days

The Atlantic

Indian Tribe Wins Fight to Limit Research of Its DNA

Challenges of Learning

NETFLIX SPILLED YOUR *BROKEBACK MOUNTAIN* SECRET, LAWSUIT CLAIMS

Facebook Gaydar Emerges From Breakthrough MIT Project

Say 'Ahhh': A Simpler Way To Detect Parkinson's

Aggregate Information Can be Revealing

How unique am I?

- + Birth date: Aug. 31

  <2

- U.S. Zip Code: 02145

  < 25000

- + Gender: Male

  < 12000

It’s Not Just the Numbers...

Brownstein, et al., 2006 ,
***NEJM* 355(16),**

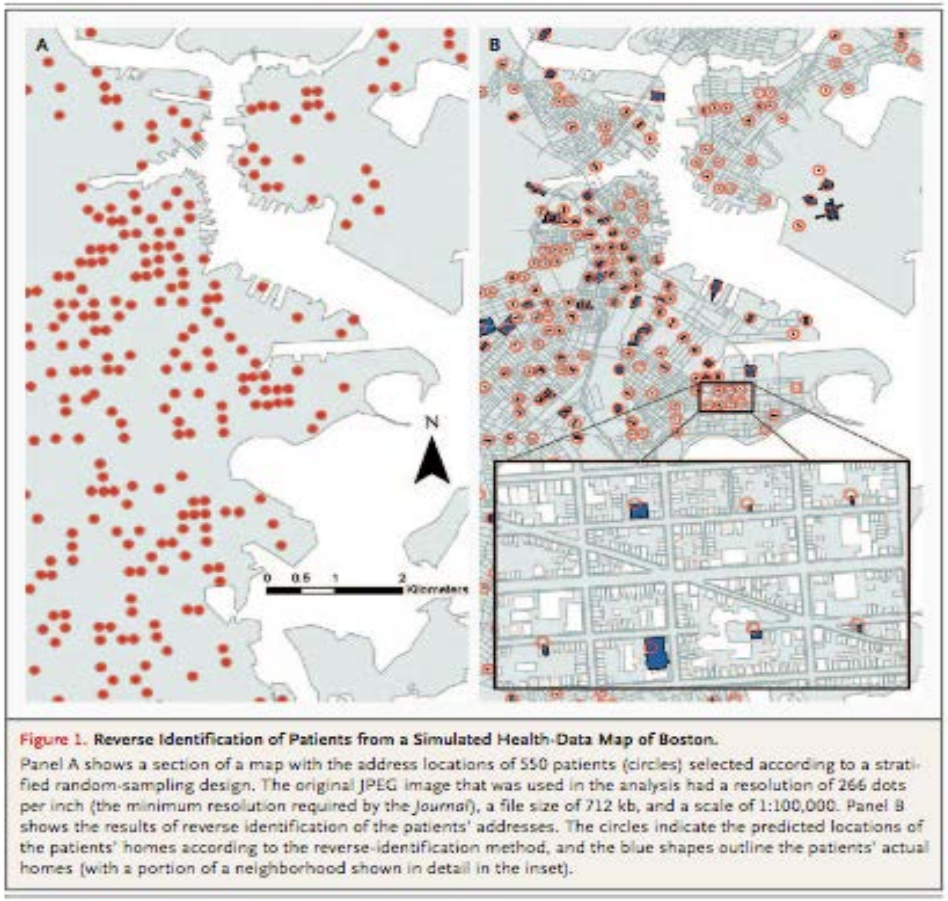


Figure 1. **Reverse Identification of Patients from a Simulated Health-Data Map of Boston.**
Panel A shows a section of a map with the address locations of 550 patients (circles) selected according to a stratified random-sampling design. The original JPEG image that was used in the analysis had a resolution of 266 dots per inch (the minimum resolution required by the *Journal*), a file size of 712 kb, and a scale of 1:100,000. Panel B shows the results of reverse identification of the patients' addresses. The circles indicate the predicted locations of the patients' homes according to the reverse-identification method, and the blue shapes outline the patients' actual homes (with a portion of a neighborhood shown in detail in the inset).

Technical Concepts & Terminology

Informational Harm

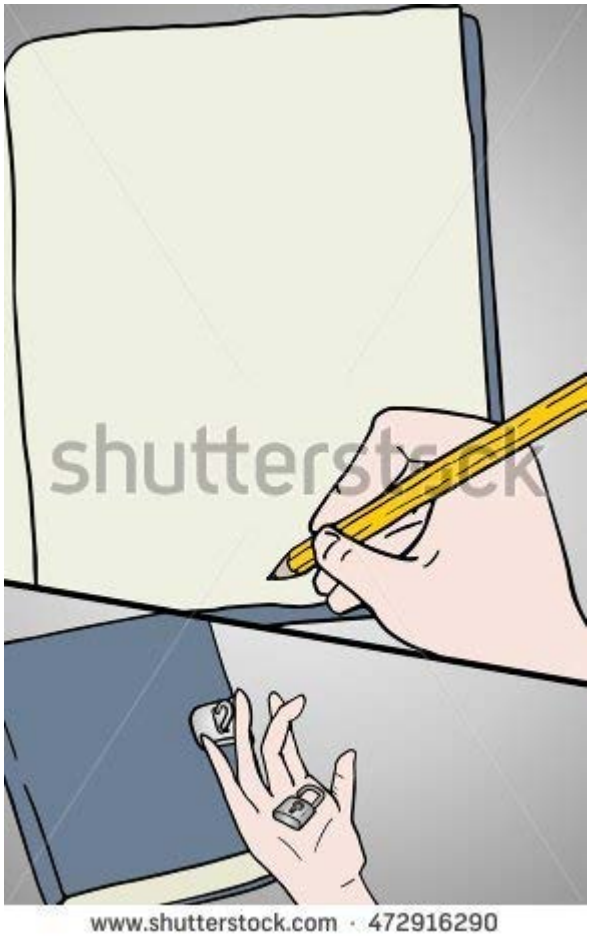
An *informational harm* occurs when others use research results or data; and then violate the rights of an individual or organization; or negatively impact their interests

Who might be harmed by information release?

Participant Harm

A informational harm *to research participants* occurs when others use research results or data; *learn about the individual as a result of their participation in the research*, and then violate their rights; or negatively impact their interests

Information Privacy and Security

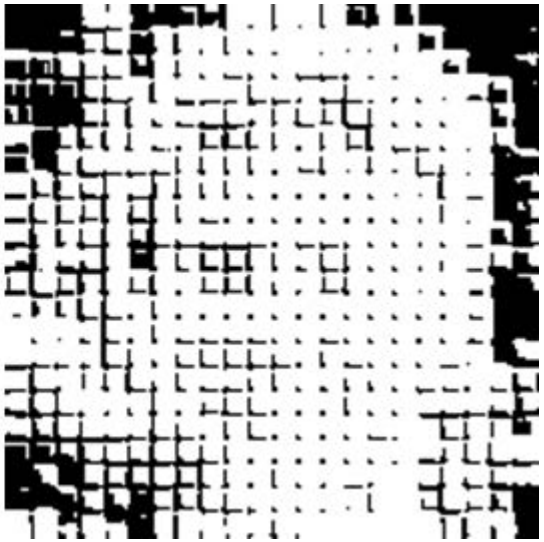


Information Security
Control and protection protection against unauthorized access, use, disclosure, disruption, modification, or destruction of information.

Information Privacy
Control and protection over the extent and circumstances of information collection, sharing, and use

https://image.shutterstock.com/display_pic_with_logo/419668/472916290/stock-photo-writing-in-diary-472916290.jpg

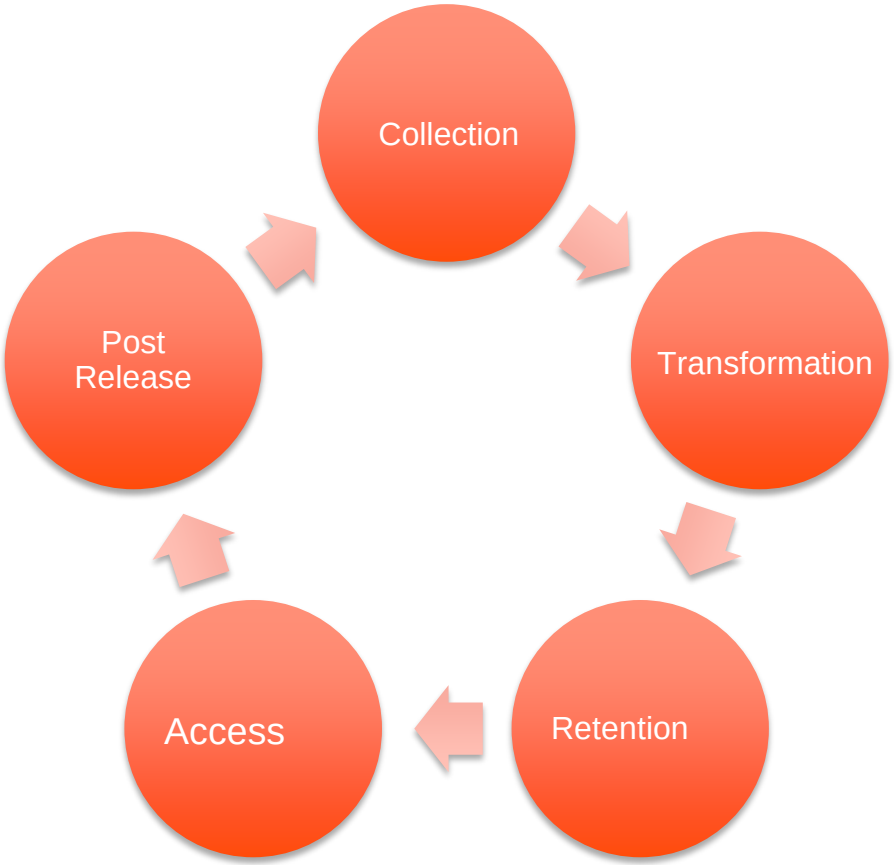
Information Utility



- Perfectly anonymous data is perfectly useless data
- There is no practical and universal measure of utility
- Privacy protections balance usefulness and privacy

Planning for Information Security and Privacy

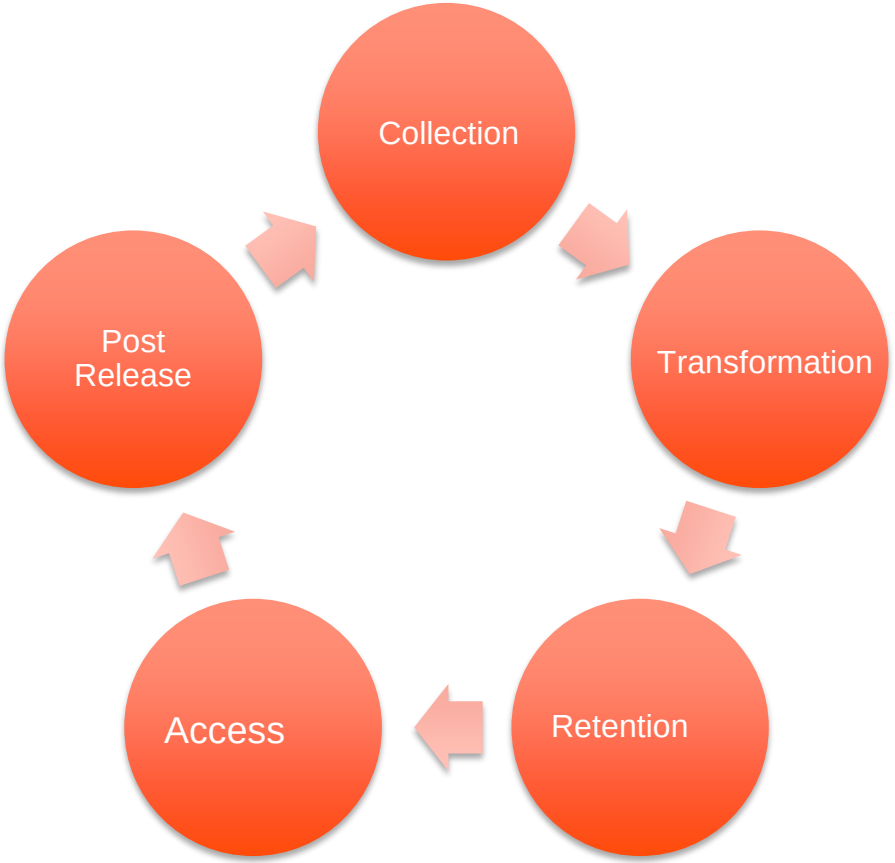
Information Lifecycle Planning



Aligning Research and Information Lifecycles

Research Phases

- Research Design
  - Evaluate privacy and security of measurement and data collection
  - Identify legal requirements for information management
  - Develop lifecycle data management plan
- Research Implementation
  - Data collection and transmission
  - Transformation
  - Retention
- Research Analysis
  - Internal data sharing and use
  - External data sharing
  - Disclosure limitation from research results
  - Data destruction



Measurement Choices & Information Risks

Identifying potential harms

What could occur to an individual subject if…

- Information collected during research was released, and
- that information was received by the "wrong hands", and
- that information was associated with an individual.

Characterizing vulnerabilities

Of the potential harms identified…

- What is the likelihood that harm would occur **if** information was received by the "wrong hands"?
- Does the likelihood of harm in this case depend on some foreseeable characteristics of the subject?

How harmful is information, if identified?

- **… creates *minimal risk* of harm**
- … **creates a *non-minimal* risk of *minor* harm**
- **… creates *significant* risk of *moderate* harm**
- … **creates *substantial* risk of *serious* harm**
- **… creates *high* risk of *grave* harm**

Characterizing Vulnerable Populations

Does the likelihood of harm depend on some foreseeable characteristics of the subject?

- *Does the subject have reduced agency – reduced ability to make meaningful choices?*
- *Are some types of subjects predictably more vulnerable to specific threats, such as embarrassment, loss of employment?*
- *Are subjects member of vulnerable groups subject to stereotype, of threats to group dignity?*

Principles for Data Protection Planning

Fair Information Practice Principles:

- Notice/awareness
- Choice/consent
- Access/participation (verification, accuracy, correction)
- Integrity/security
- Enforcement/redress

"Five Safes"

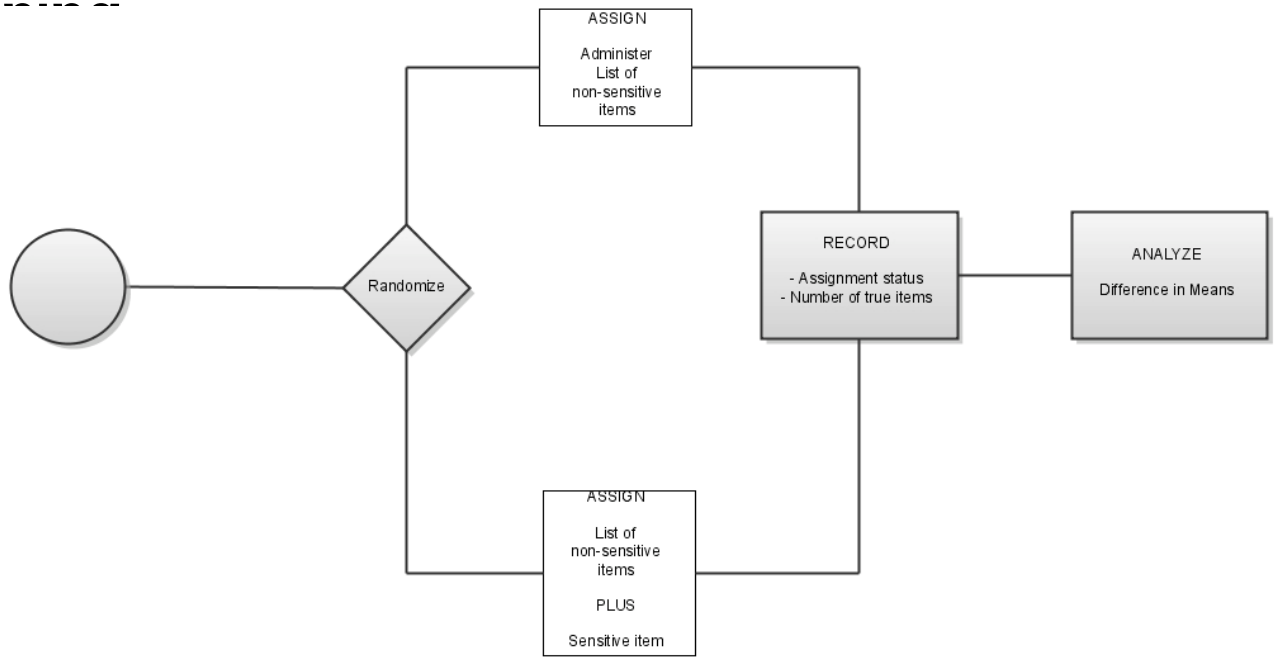
- Safe projects
- Safe people
- Safe settings
- Safe data
- Safe outputs

Modern privacy principles:

- Calibrate controls to use and risk
- Consider risks from inference
- Use a combination of controls
- Plan for tiered access
- Anticipate change

Measurement Choices: Mitigating Risks

- **Data minimization**
- **Alternative measurement design**
- **Lifecycle planning**



Recognizing Legal Requirements

There are many laws and regulations protecting information security and privacy

- Laws differ based on
  - General Scope of Law
  - Characterizing Protected Information
  - Protection requirements
- Most common laws affecting research
  - Human subjects regulations
  - Educational data restrictions
  - Health data protections
  - General data protection (non-US)

Types of Legal Requirements

- General strategies
  - Notice & consent
  - Limiting access
  - De-identification

- Typical mechanisms
  - Technical requirements
  - Process requirements
  - Civil and criminal liability
  - Certificates of confidentiality

General Triggers for Regulatory Concern

- Data collector / controller characteristics
- Data subject characteristics
- Data characteristics

The Role of IRBs

- IRBs review consent procedures and documentation

- IRBs may review data management plans
  - May require procedures to minimize risk of disclosure
  - May require procedures to minimize harm resulting from disclosure

- IRBs make determination of *sensitivity of information*
  - potential harm resulting from disclosure

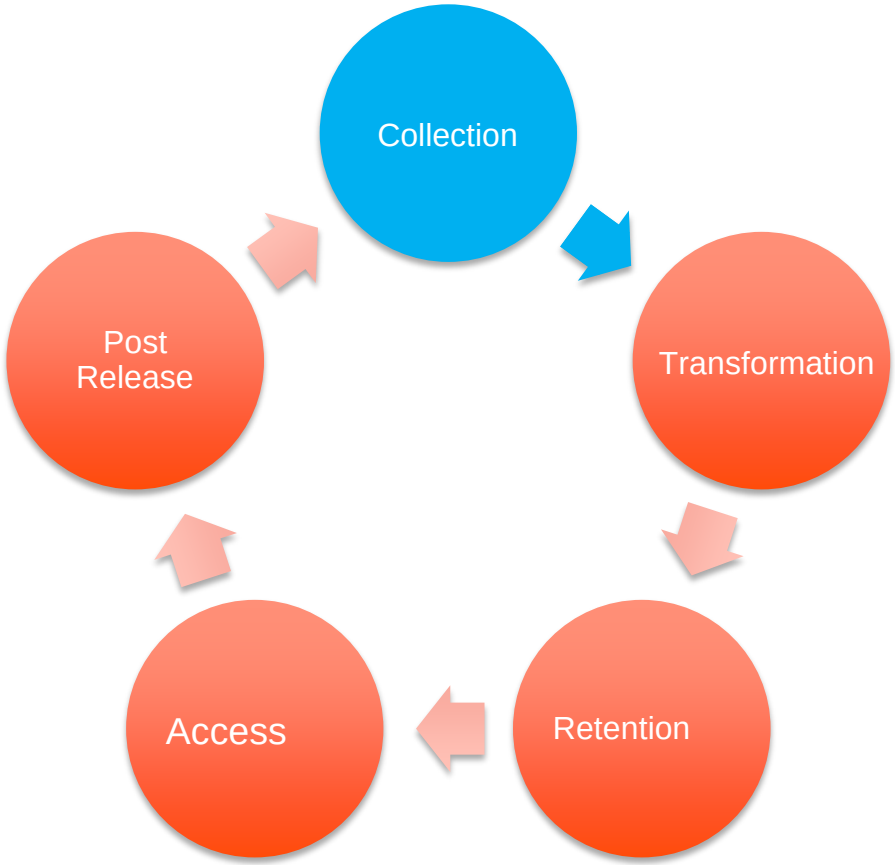
- IRBs make determination regarding whether data is de-identified for "public use"

Data Collection & Information Risks

Threats from direct observation

Threats from recording

Threats from transmission



Mitigating Data Collection Risks

- Consider separate channels for different measures
  - highly sensitive measures
  - Identifying information

- For sensitive data:
  - Collect on-line directly (with appropriate protections); *or*
  - Encrypt collection devices/media
  (laptops, usb keys, etc) or;
  - Consider trusted third-party for collection
  - Consider anonymous data collection

- For very/extremely sensitive data:
  - Collect with oversight directly; then
  - Store on encrypted device and;
  - Transfer to secure server as soon as feasible

Mitigating Data Transmission Risks

- Encrypt all data in motion
  - Strong keys
  - Client-side key
  - Verified algorithm and implementation
  - Client key never stored in cloud, never transmitted in clear

- Examples of encrypted protocols
  - SSL
  - SSH
  - VPN

Mitigating Online Data Collection Risks

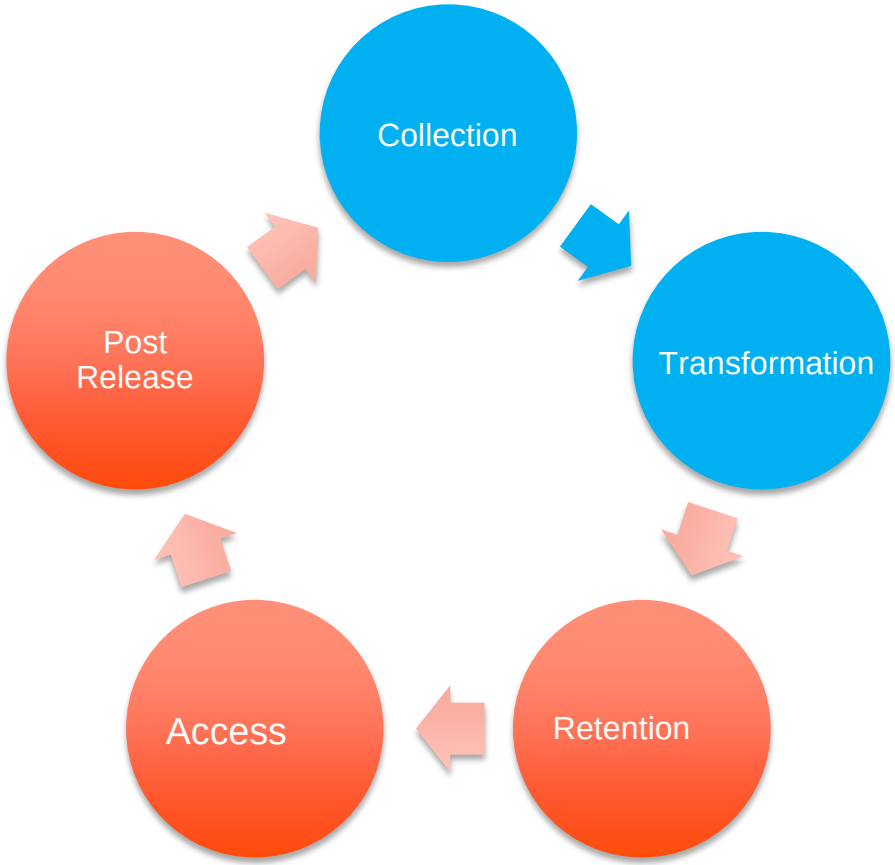
- Use only vendors that agree to comply with your policy
- Do not retain additional information beyond designed collections
- Encrypt data at rest on server
- Encrypt data in motion

Protective Transformations

Data partitioning

Redaction

Encryption



Partitioning Information for Protection

Name	SSN	Birthdate	Zipcode	LINK
A. Jones	12341	01011961	02145	1401
B. Jones	12342	02021961	02138	283
C. Jones	12343	11111972	94043	8979
D. Jones	12344	12121972	94043	7023

LINK	Favorite Ice Cream	Treat	# acts
1401	Raspberry	0	0
283	Pistachio	1	20
8979	Chocolate	0	0
7023	Hazelnut	1	12

- Reduces risk in information management
- Partition data information based on sensitivity
- Segregate for security
- Choose linking keys at random – or in a cryptographically secure way

De-Identification and Anonymization

De-identification is often accomplished by redaction or similar information removal.

This can satisfy legal requirements, and can reduce risk somewhat.

Anonymization and de-identification are legal concepts.

Example:
Different Definitions of Anonymized Data

	FERPA
Identification Criteria	- Direct - Indirect - Linked **- Bad intent**

Example:
Different Definitions of Anonymized Data

	FERPA	HIPAA
Identification Criteria	- Direct - Indirect - Linked **- Bad intent**	- direct/indirect: 18 identifier - OR statistician verifies minimal risk AND no actual knowledge of identified individual

Example:
Different Definitions of Anonymized Data

	FERPA	HIPAA	Common Rule
Identification Criteria	- Direct - Indirect - Linked **- Bad intent**	- direct/indirect: 18 identifier - OR statistician verifies minimal risk AND no actual knowledge of identified individual	- Direct - Indirect / Linked -- if "readily identifiable"

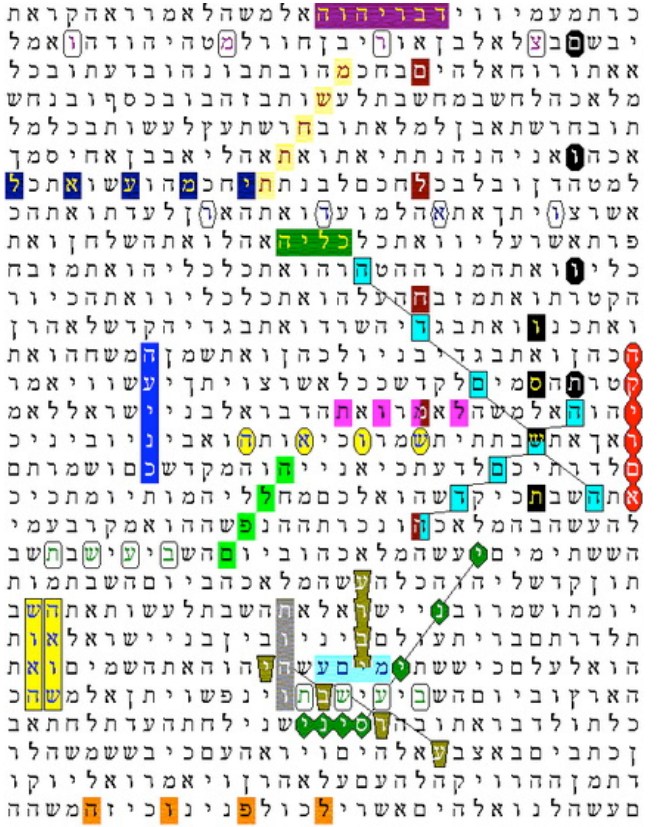
Example:
Different Definitions of Anonymized Data

	FERPA	HIPAA	Common Rule
Identification Criteria	- Direct - Indirect - Linked **- Bad intent**	- direct/indirect: 18 identifier - OR statistician verifies minimal risk AND no actual knowledge of identified individual	- Direct - Indirect / Linked -- if "readily identifiable"
Sensitivity Criteria	Any non-directory information	Any medical information	Private information, causing harm if known

Protection through Encryption

What does encryption do?
- transformed data so it is ‘as-if’ random
- can be understood by those with the ‘key’

What can be encrypted?
- files
- media
- file systems
- transmissions
- computations
- databases



Encryption considerations

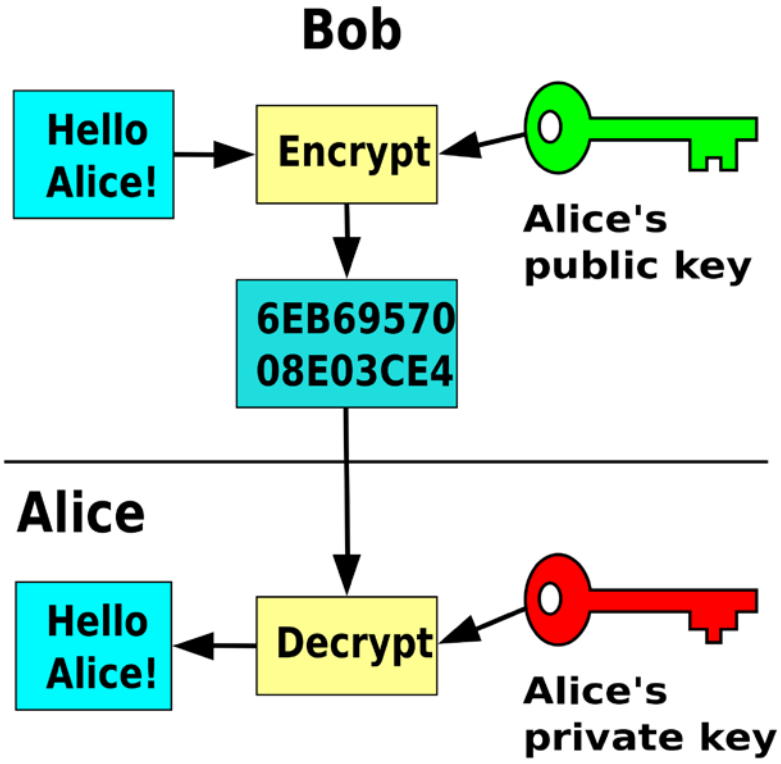
Algorithms

Implementation

Key structure

Key management

Implications for use, sharing, destruction, integrity

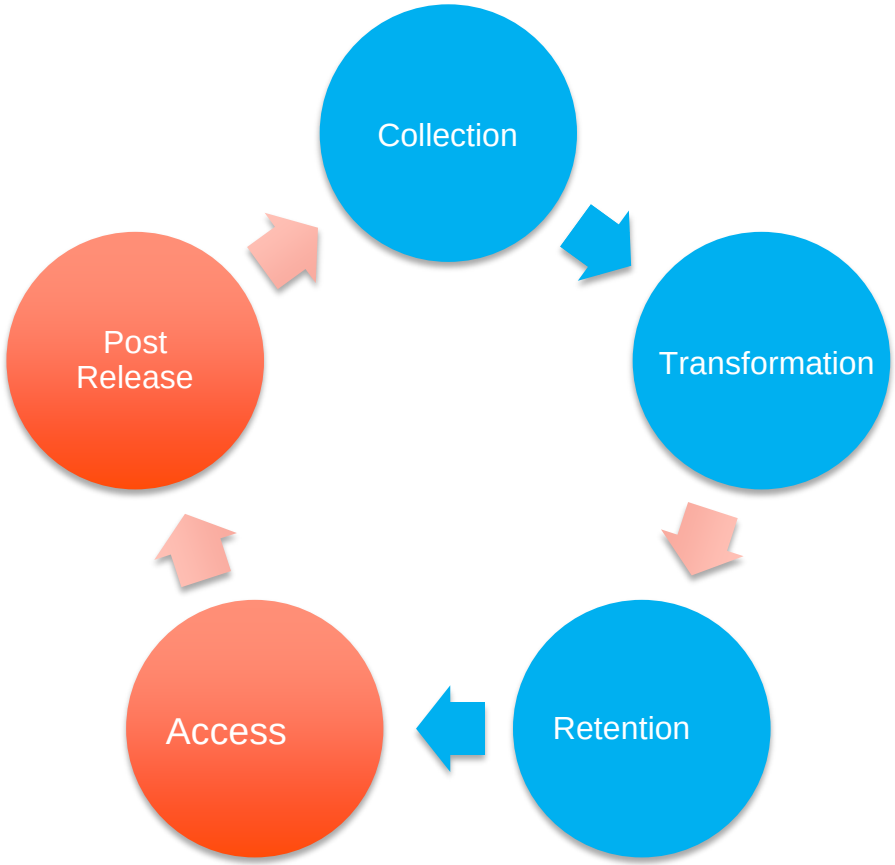


Data Retention & Information Risks

Violations of integrity

Violations of availability

Violations of confidentiality

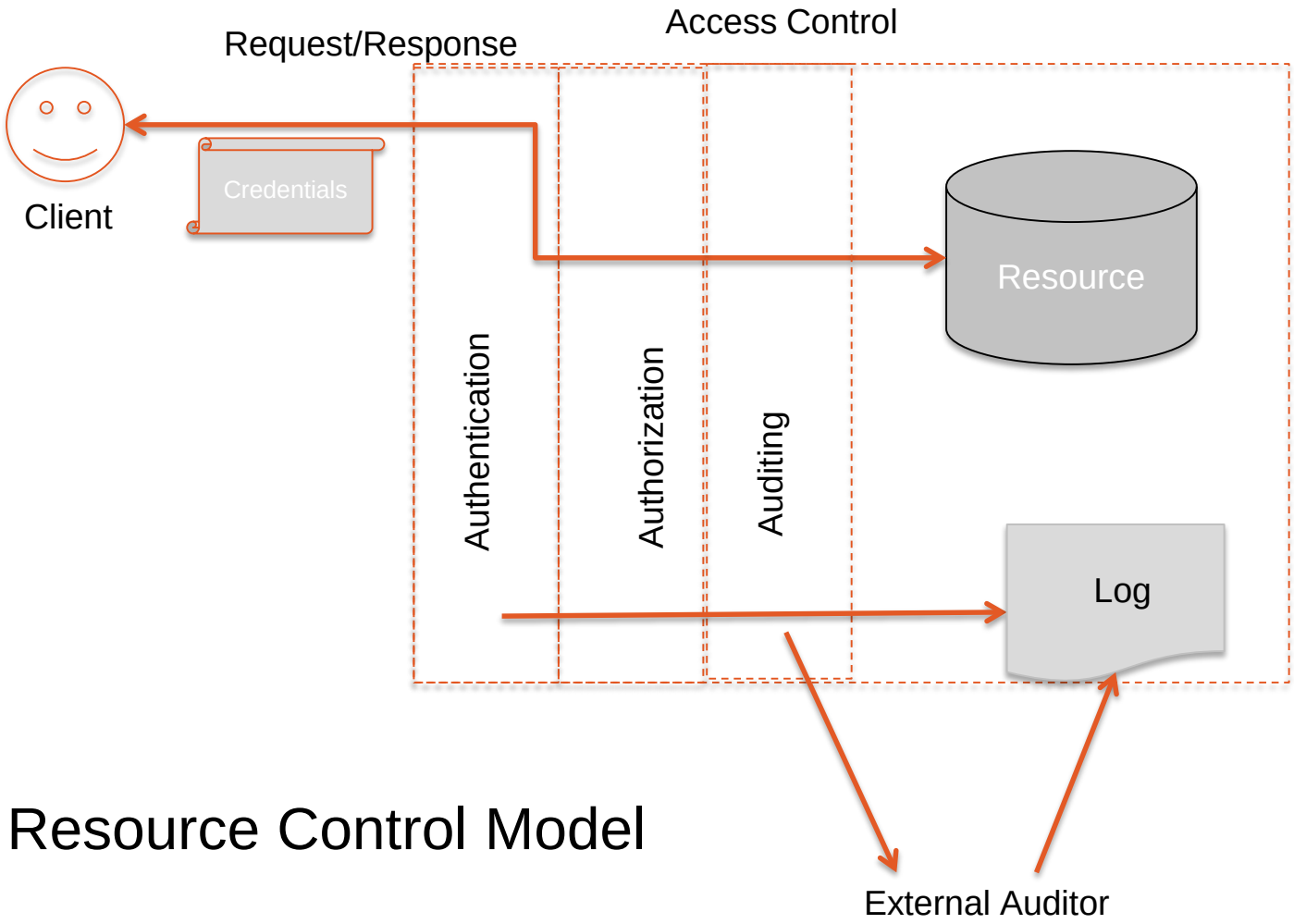


Information Security Threats

- Sources of threat
  - Natural
  - Unintentional Human
  - Intentional

- Areas of vulnerability
  - Logical
    - Data at rest in system
    - Data in motion across networks
    - Data being processed in applications
  - Physical
    - Computer systems
    - Network
    - Backups, disposal, media
  - Social
    - Social engineering
    - Mistakes
    - Insider threats

Simple Access Control Model



Resource Control Model

Internal Data Use and Data Retention

Operational

- Personnel security
- Physical and environmental protection
- Contingency planning
- Configuration management
- Maintenance
- System and information integrity
- Media protection
- Incident Response
- Awareness and training

Technical Controls

- Identification and authentication
- Access control
- Audit and accountability
- System and communication protection

High level security questions

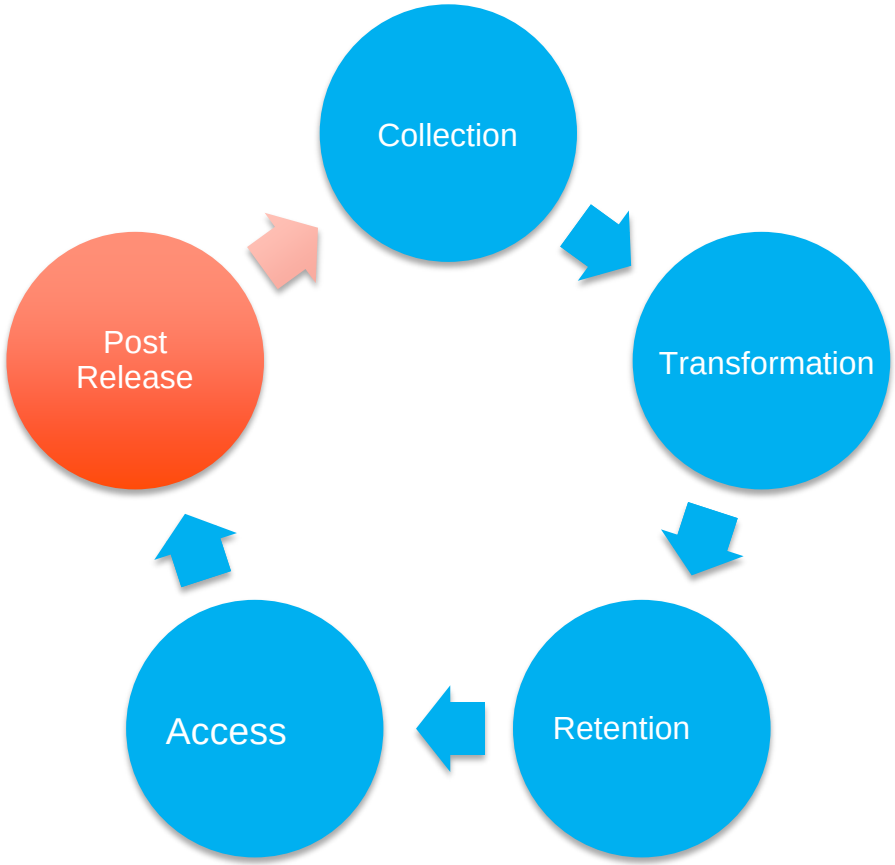
- What are goals for confidentiality, integrity, availability?
- What threats are envisioned?
- What controls are in place?
- What is your a checklist?
- Who is responsible for technical controls?
  - Do they have appropriate training, experience and/or certification?
- Who is responsible for procedural controls?
  - Have they received appropriate training?
- How is security monitored, audited, and tested?
  - E.g. SSAE 16/SAS Type -2 Audits; FISMA Compliance; ISO Certification
- What security standards are referenced?
  - E.g. FISMA, ISO, HEISP/HDRSP/PCI

External Dissemination & Information Risks

Identifiability

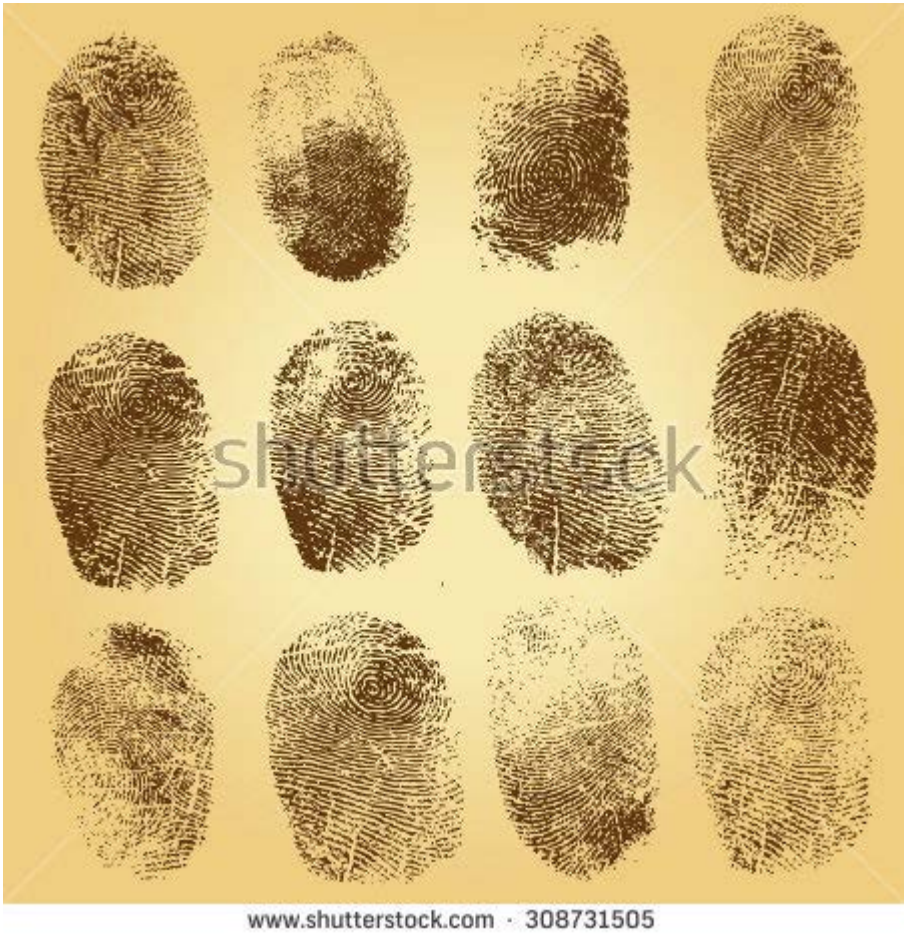
Disclosure limitation

Data sharing and access



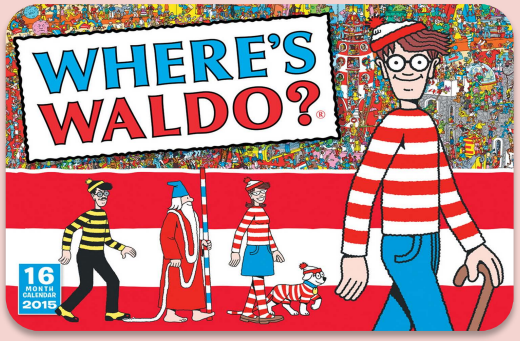
Privacy Core Concepts

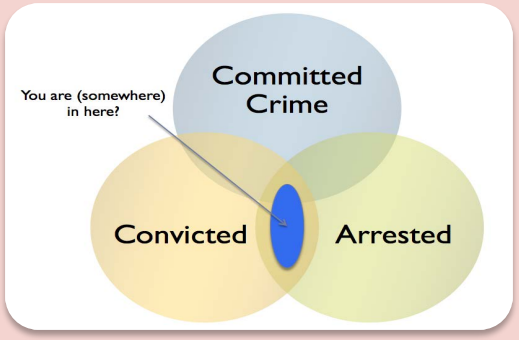
Identifiability
Potential for learning about individuals from computations based on data in which they are included

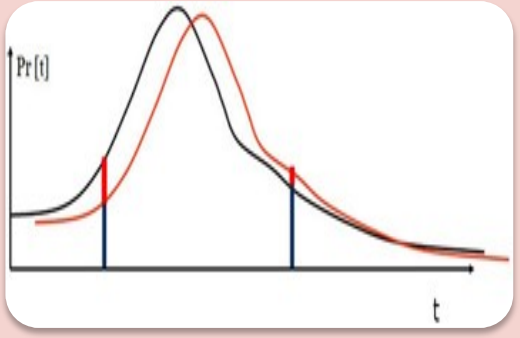


https://image.shutterstock.com/display_pic_with_logo/2664670/308731505/stock-vector-set-of-fingerprints-vector-illustration-isolated-on-vintage-background-308731505.jpg

Different measures of identifiability







Record-linkage

"where's waldo"

- Match a real person to precise record in a database
- *Examples*: direct identifiers.
- *Caveats*: Satisfies compliance for specific laws, but not generally; substantial potential for harm remains

Indistinguishability

"hiding in the crowd"

- Individuals can be linked only to a cluster of records (of known size)
- *Examples*: K-anonymity, attribute disclosure
- *Caveats*: Potential for substantial harms may remain, must specify what external information is observable, & need diversity for sensitive attributes

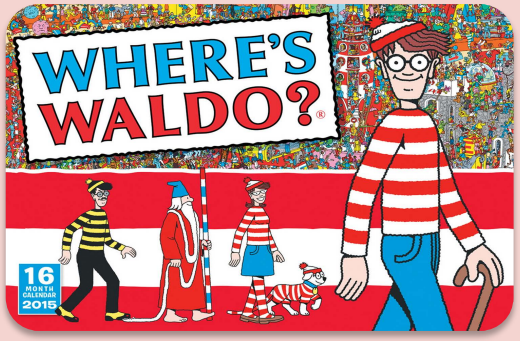
Limits on adversarial learning

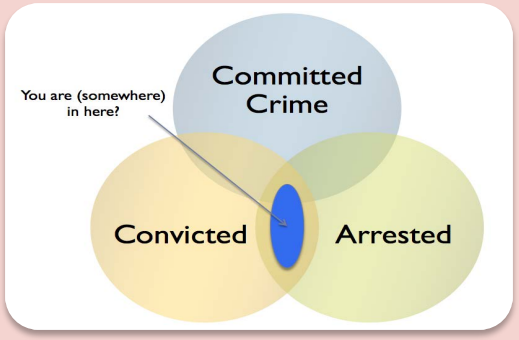
"confidentiality guaranteed"

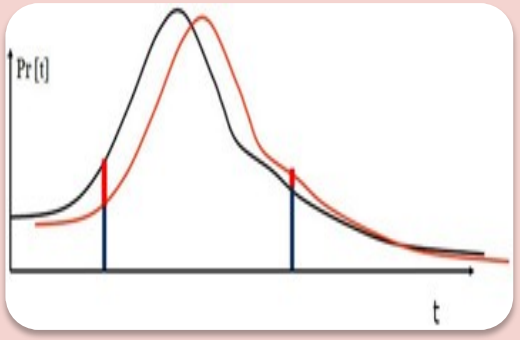
- Formally bounds the total learning about any individual that occurs from a data release
- *Examples*: differential privacy, zero-knowledge proofs
- *Caveats*: Challenging to implement, requires interactive system

Less Protection → More Protection

Different measures of identifiability







Record-linkage

"where's waldo"

- Match a real person to precise record in a database
- *Examples*: direct identifiers.
- *Caveats*: Satisfies compliance for specific laws, but not generally; substantial potential for harm remains

Indistinguishability

"hiding in the crowd"

- Individuals can be linked only to a cluster of records (of known size)
- *Examples*: K-anonymity, attribute disclosure
- *Caveats*: Potential for substantial harms may remain, must specify what external information is observable, & need diversity for sensitive attributes

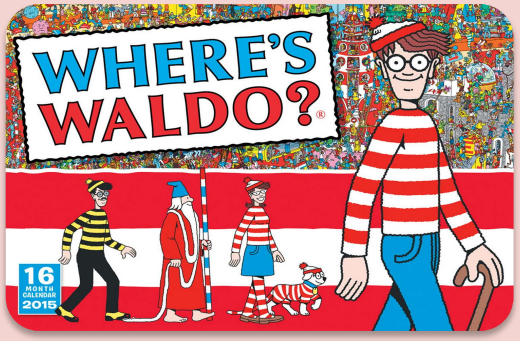
Limits on adversarial learning

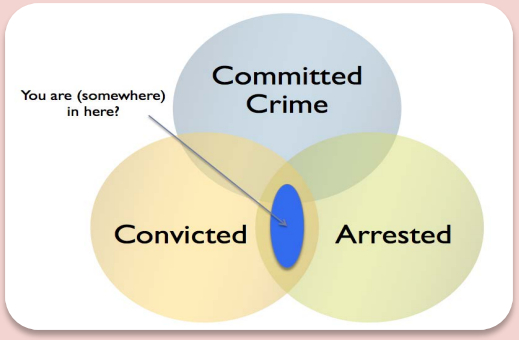
"confidentiality guaranteed"

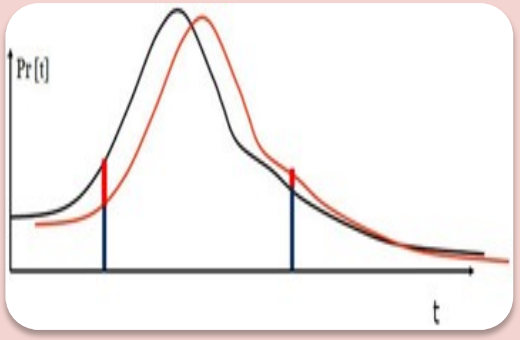
- Formally bounds the total learning about any individual that occurs from a data release
- *Examples*: differential privacy, zero-knowledge proofs
- *Caveats*: Challenging to implement, requires interactive system

Less Protection → More Protection

Different measures of identifiability







Record-linkage

"where's waldo"

- Match a real person to precise record in a database
- *Examples*: direct identifiers.
- *Caveats*: Satisfies compliance for specific laws, but not generally; substantial potential for harm remains

Indistinguishability

"hiding in the crowd"

- Individuals can be linked only to a cluster of records (of known size)
- *Examples*: K-anonymity, attribute disclosure
- *Caveats*: Potential for substantial harms may remain, must specify what external information is observable, & need diversity for sensitive attributes

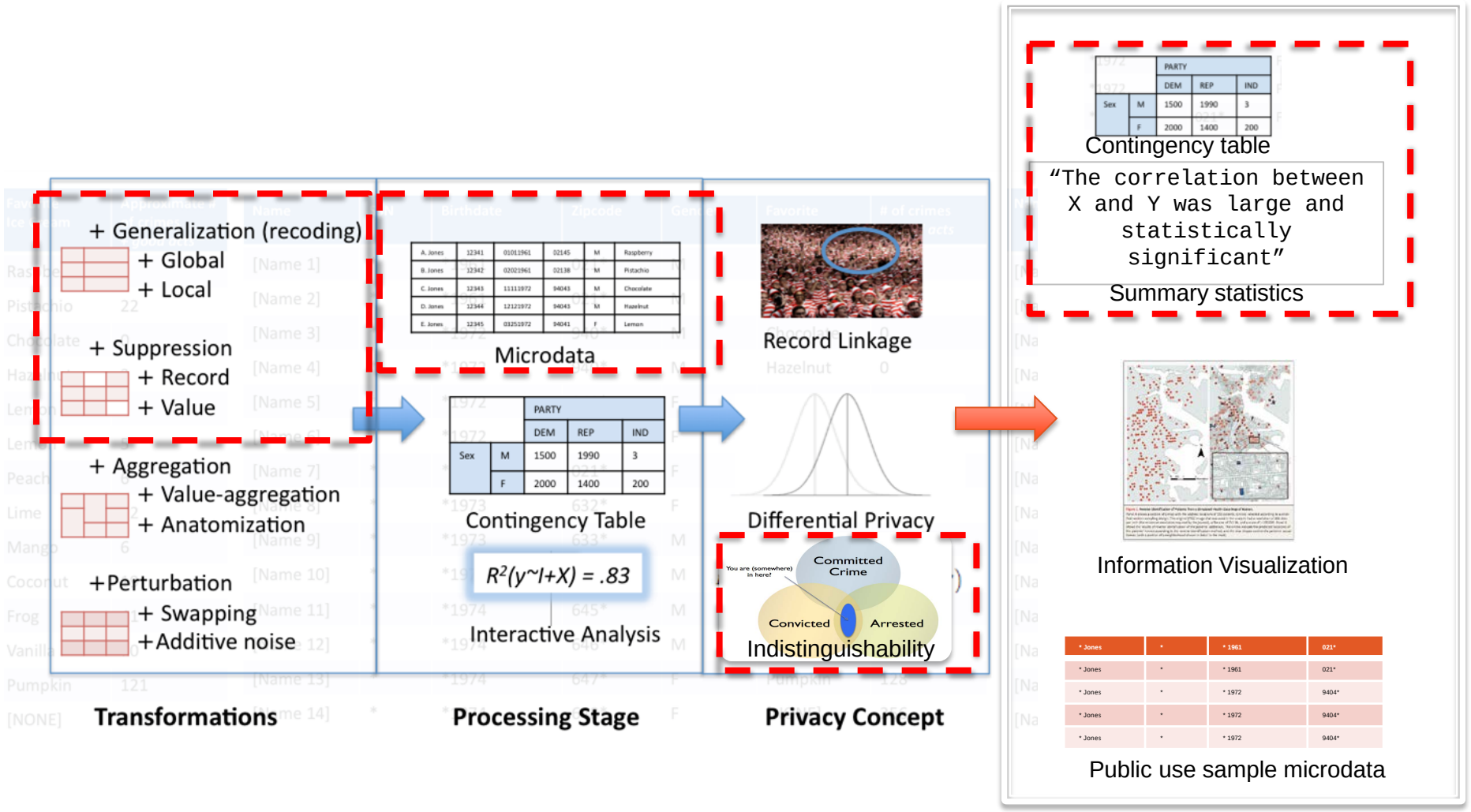
Limits on adversarial learning

"confidentiality guaranteed"

- Formally bounds the total learning about any individual that occurs from a data release
- *Examples*: differential privacy, zero-knowledge proofs
- *Caveats*: Challenging to implement, requires interactive system

Less Protection → More Protection

Disclosure Limitation for Data Privacy



Perfect Privacy*

Name	~~SSN~~	Birthdate	Zipcode	Gender	Favorite Ice Cream	# of crimes committed
*	*	*	*	*	*	*
*	*	*	*	*	*	*
*	*	*	*	*	*	*
*	*	*	*	*	*	*
*	*	*	*	*	*	*
*	*	*	*	*	*	*
*	*	*	*	*	*	*
*	*	*	*	*	*	*
*	*	*	*	*	*	*
*	*	*	*	*	*	*
*	*	*	*	*	*	*
*	*	*	*	*	*	*
*	*	*	*	*	*	*

Perfectly Anonymous Data

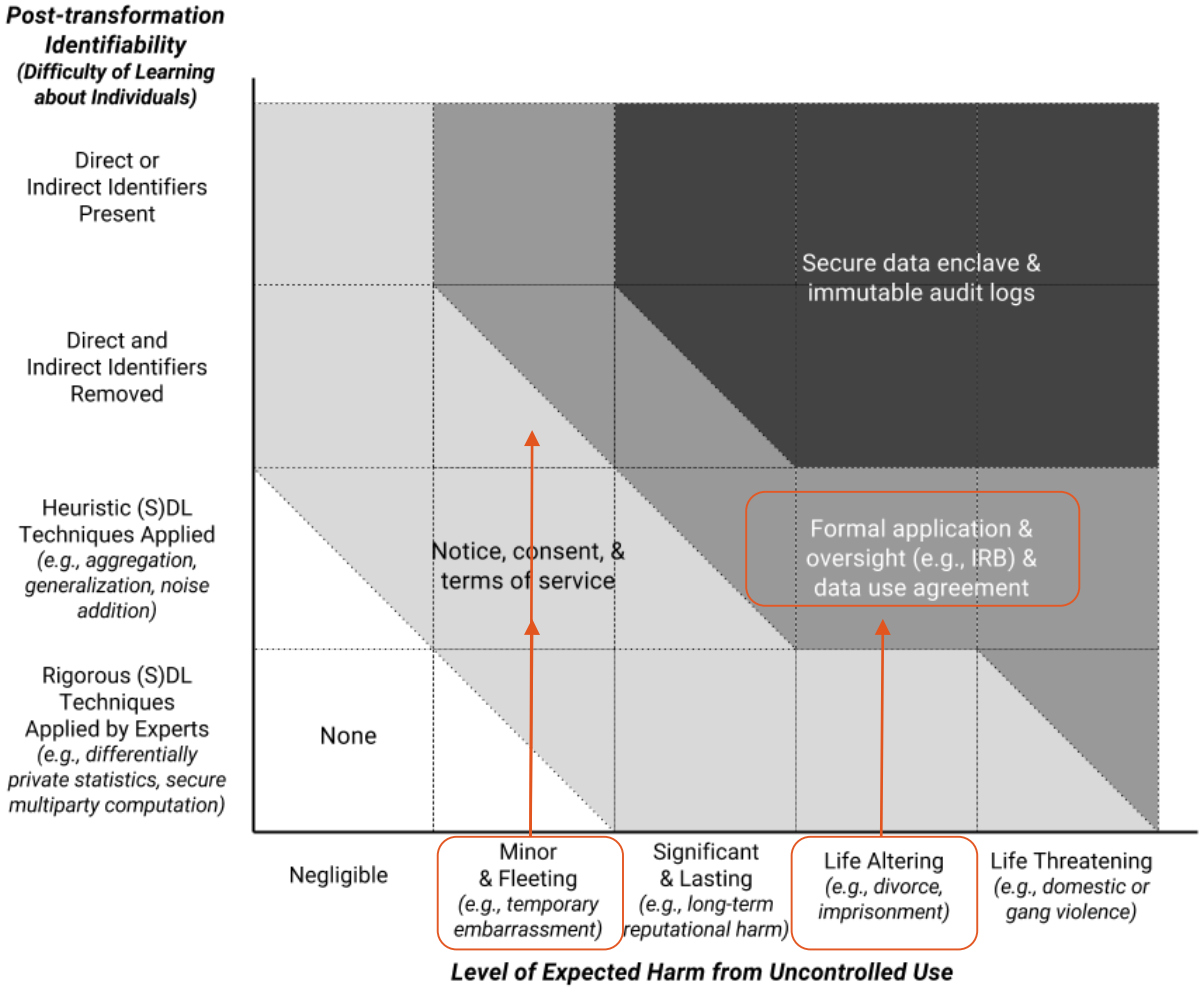
**** Global Tabular Suppression***

No-Free-Lunch for Privacy

Any data analysis that is useful leaks *some measurable private information*

Calibrating Controls

Exemplifying how to choose combination of controls and disclosure limitation based on information sensitivity



Supporting Data Sharing and Access

- Public Data Archives
  - Broad access
  - Examples: Dataverse, Data.Gov
- Data enclaves
  - physically restrictions on access
  - Example: ICPSR, national statistical agencies
- Controlled remote access
  - Virtual control
  - Example: US Census RDC, NORC
- Model servers
  - Mediated remote access – analysis limited to designated models

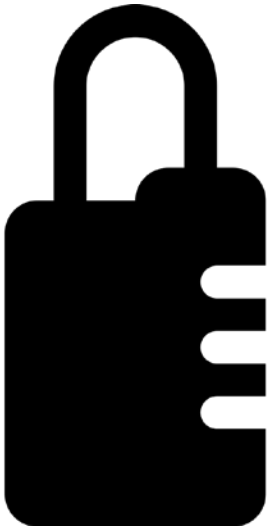
Auditing and Monitoring

- Provide subjects with rights to action over their data
- Look for changes in auxiliary data affect identifiability and sensitivity
- Plan to discover and notify re-identified subjects
- Plan to detect and and respond to improper use

Selecting and Applying Key Information Controls

What are Information Controls?

- Information controls are *safeguards or countermeasures*
- Their aim is to *avoid, detect, counteract, or reduce* information risks
- There are different types of controls



Select Controls from Information Security & Privacy Frameworks

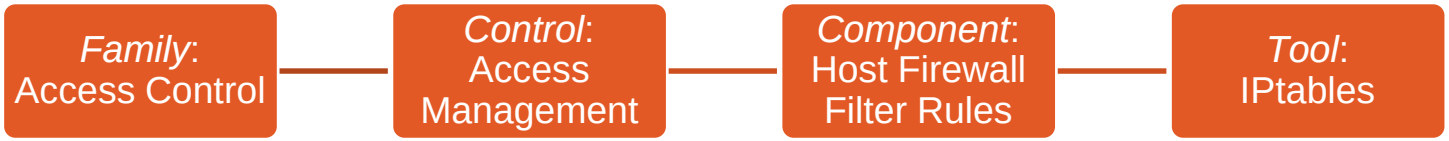
- Frameworks support compliance
  - Many laws incorporate or recognize standards
- FISMA
  - Federal Information Security Management Act of 2002
- ISO/IEC 27000 Series
  - Broad international privacy and Security Standards
- FIPPs
  - Federal Trade Commission's Fair Information Practice Principles

Use Frameworks for Design & Interoperability

- Frameworks support design
  - guidance in matching controls to vulnerabilities
  - guidance in calibrating controls to risk

- Frameworks support interoperability
  - interoperability across organizations
  - interoperability across tools

Example control hierarchy

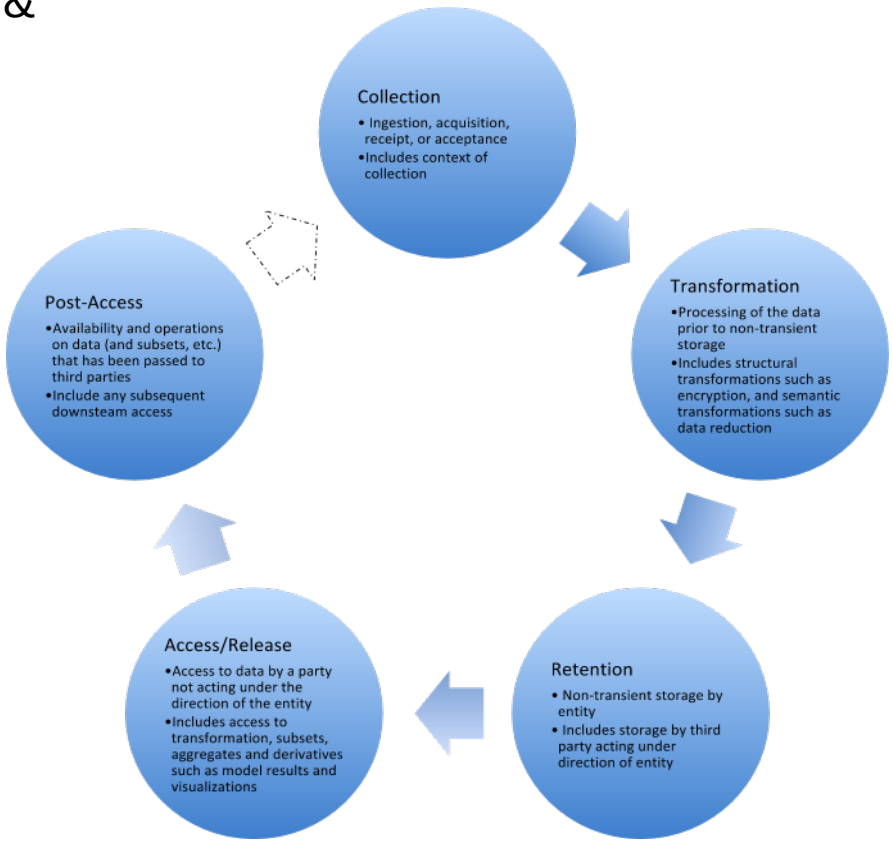


Embed Controls in the Information Lifecycle

Research Design Phase
- Identify major information threats & sensitive information
- Identify privacy and security framework
- Identify key control families
- Create data management plan

Collection, Storage, Retention, Access
- Select individual controls
- Select tools
- Implementation
- Monitoring and auditing

Post-Dissemination
- Monitoring & auditing



Align Controls & Documentation

- Documentation as Guidance
  - risks, vulnerabilities
  - requirements
  - framework and design choices
- Documentation as Communication
  - for institutions, researchers, research participants

Key Documentation

Information Privacy & Security Plan Outline

- Informational Risks
- Legal and Policy Requirements
- Standards and Frameworks
- Data collection
- Data retention
- Access/sharing
- Adherence

Functions of Consent Document

- Ensure informed consent to participate in research
- Communicate foreseeable risks
- Communicate potential societal benefits
- Identify mechanisms for questions and for withdrawal
- Communicate procedures for protecting the confidentiality of personal information about the participants
- Communicate limits to confidentiality
- Communicates processes and benefits of sharing data

Consent Document – Practices to avoid

- X Assurances of complete anonymity / privacy
- X No plans to share data
- X No plans to disseminate/archive data
- X Data will be shared only in anonymized form
- X All data destroyed at end of research

Functions of Service Level Agreement

- Service Level Agreement (SLA) is an official commitment between a service provider and a customer
- SLA's generally specify type, cost, quality, and availability of service
- Services for research data should also specify information security and privacy controls

Key Elements to include in an SLA

- Privacy and security standards compliance
- Additional information controls
- Access control policies
- Backup and retention policies
- Usage logging and sharing policies
- Breach notification policies
- *For Highly Sensitive Data*
  - Duty of Care
  - Information residency
  - Incident response policies
  - External auditing
  - Policies on responding to legal records request

Data Use Agreement Functions

- Communicate requirements to third parties
- Reduce risks to institution
- Support transparency and reuse
- Legal compliance
- Grant subjects individual right of action

Data Use Agreement – Key Elements

- Key elements
  - Who may receive and use information
  - Permitted/prohibited uses
  - Intellectual property restrictions
  - Restrictions on further sharing
  - Safeguards for use and disclosure:
    Storage and transmission controls; Access controls
  - Data destruction requirements
  - Auditing/review requirements
  - Reporting/notification requirements
  - Duration and revocation

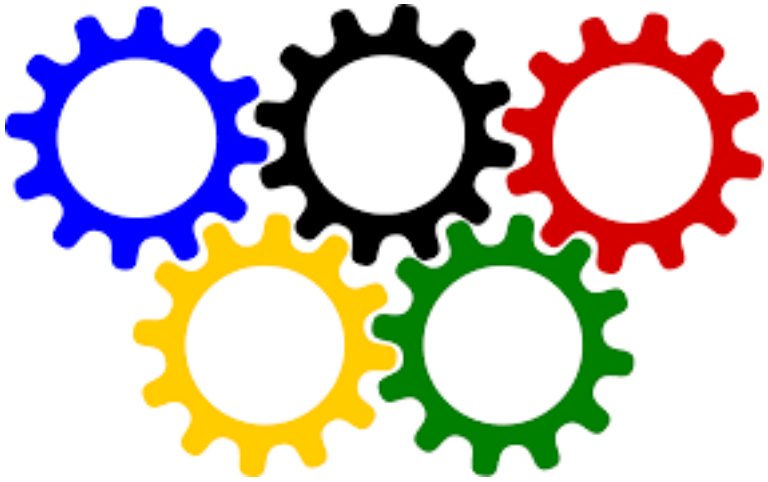
- Consider
  - Indemnification
  - Rights of actions by individual subjects
  - Transparency – citation and replication requirements
  - Openness – least restrictive conditions – equivalent to that required for original research

Functions of Implementation Checklists

- Ensures consistent processes
- Provides guidance on specific controls to be applied
- Assign clear responsibility for review
- Establish records that controls were applied

Documentation Good Practices

- Prepare key documents prior to implementation
  - *information privacy and security plan should be prepared during research design*
- Calibrate documentation to key information risks, requirements, and frameworks
- Use a well-specified vocabulary
- Align policies and controls across documents



Systems & File Security Tools -- Controls on Data Retention

What is a hardened system?

- Hardened systems are systems intentionally managed and modified to be more trustworthy
- Hardening generally involves identifying and reducing security vulnerabilities
- Reducing vulnerabilities often involves
  - disabling unnecessary software and services
  - restricting access channels and user access
  - managing configuration and updates
  - managing authorized users
  - instrumenting logging and auditing processes

Key Hardened Systems Tools & Practices

Low-Moderate Risk

- Configuration checklist
- Host-based firewall
- Virus scanner
- Idle lockout mechanism
- Password lockout mechanism
- Whole-disk / filesystem encryption
- Open account removal
- Default password change
- Automated update mechanism

Moderate High-Risk adds…

- Physical/protected access
- Continuous professional management
- Intrusion detection systems
- Vulnerability scans
- Multi-factor authorization
- Password complexity enforcement
- Passwords stored only in encrypted form
- Account management and revocation
- Protected network connections
- Minimizing services and software
- Audit/access/security logs are maintained on separate system

Selecting Encryption tools for Filesystems, Media, and Files

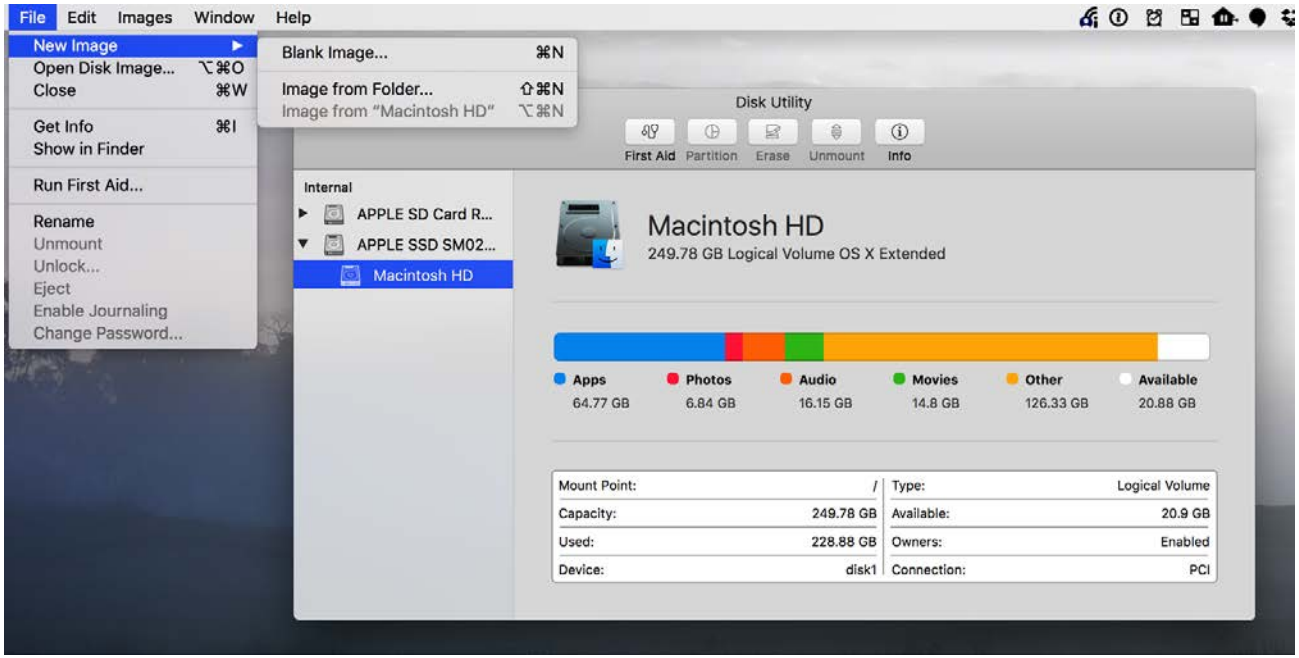
- What is encryption?
- What can it do for you and what can't it do for you?
- What are different encryption types? Symmetric and Asymmetric encryption.
- What are private keys? What are public keys?
- Criteria Selecting and configuring encryption tools
  - Public, well-known algorithm
  - Vetted software
  - Key size (256 bit symmetric, 2048 public key)
  - Password complexity and entropy
  - Implications for backup and recovery

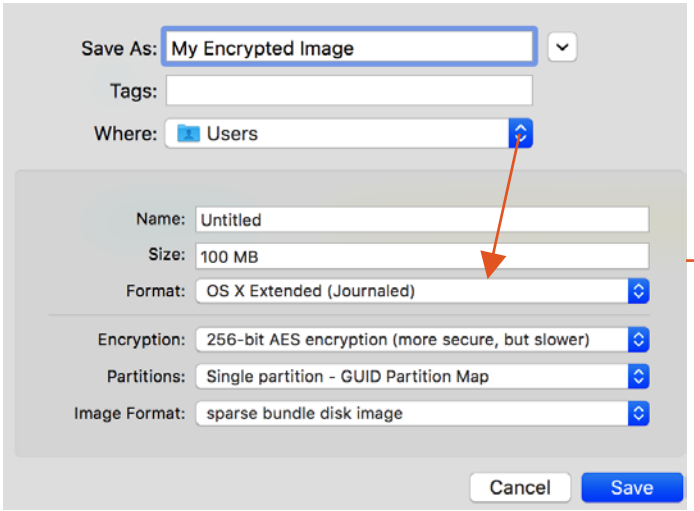
Major File/Filesystem Encryption Tools

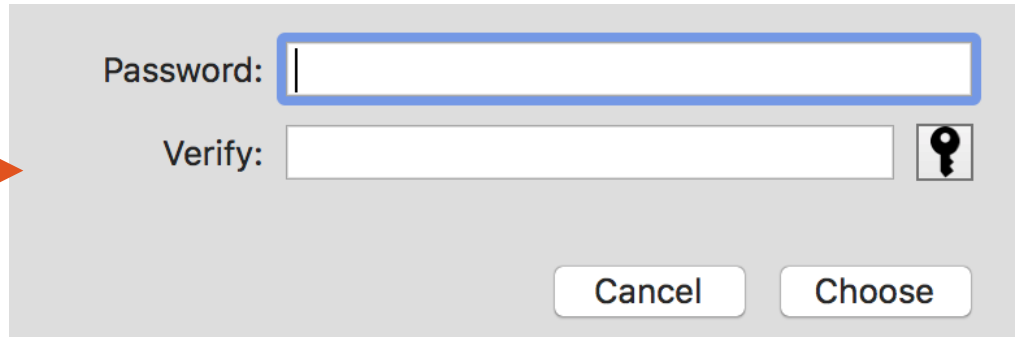
	Operating System	Encryption Level	Caveats
VeraCrypt	Windows, OS X, Linux	Filesystem, Virtual Filesystem	Not for multiuser networked filesystems
Bitlocker/FIleVault	Windows/OS X	FileSystem, Virtual Filesystem	Not for multiuser networked filesystems Not open source
7Zip	Windows/OS X/Linux	File, Bundle files	Separate secure mechanism needed for password transmission
GNU Privacy Guard	Windows/OS X/Linux	File, Bundle files, Filesystems	Steep learning curve
Crashplan	Windows/OS X/Linux	Encrypted Backup	Backups only Configure to use client side key for highly sensitive data

Encryption Example

Disk Utility

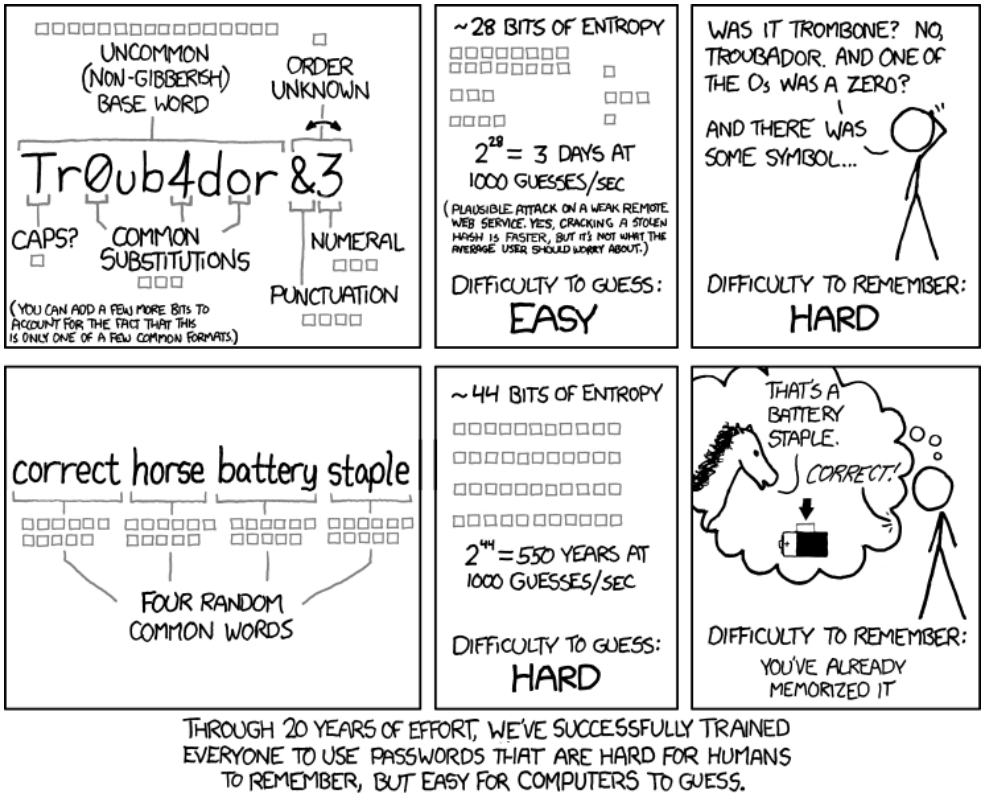






Password Protection and Tools

- Good passwords are easy to remember…
  - Personally crafted phrases
  - Don't change frequently
- Good passwords are hard to guess
  - High entropy - resist brute force attacks
  - Not common names
  - Not shared
  - Resist brute force attacks
- Passwords work best with complementary protections
  - Stored and transmitted encrypted
  - Multi-factor authentication
    - Token management, recovery, and channels



Password Management Tools

	Operating System	Function	Caveats
OnePassword	Windows, OS X, Linux, Android, IOS, Browser	Password vault, Breach alerting service	Not open source, No multifactor authentication
LastPass	Windows, OS X, Linux, Android, IOS	Password vault	
KeePass	Windows, OS X, Linux, Android, IOS	Password vault	
Google Authenticator	Android	Token Generator	Single device only
Authy	Android/IOS/Browsers	Token Generator, Multi-device synchronization	

Selecting tools for Information Hosting

- Publish & read vs. multiple writers
- Cloud service vs. locally hosted service
- Vetted software
- Data residency
- SLA
- Key and password management
- Access control
- Revocation

Major Data Hosting Tools

	Hosting Model	Features	Caveats
TahoeLAFS	Self-Hosted	Fully distributed multiuser encrypted filesystem	Reduced support, deployment complexity
EncFS	Self-Hosted	Encrypted file system	Reduced support, deployment complexity
Git-Annex + GPG	Self/Cloud Hosted	Git-based revision control model	High learning curve for use
AWS FISMA-compliant	Infrastructure as Service	File storage and computing infrastructure in a secure environment	Deployment complexity, cost
Dropbox, Google Drive, …	Cloud	Easy, encrypted file sharing	Encryption does not use client-supplied key; can be decrypted by host, closed source
SpiderOak, Tresor	Cloud	Zero knowledge encrypted sharing Encryption uses client-side key – host does not have access	Limited formal legal certifications, small compoanies, closed source

Approaches at the Leading Edge of Access Control

- Secure multi-party computing
- Homomorphic encryption
- Functional encryption
- Personal data stores
- Virtual enclaves
- Blockchain

Disclosure Limitation Tools

Sensitive & identified data

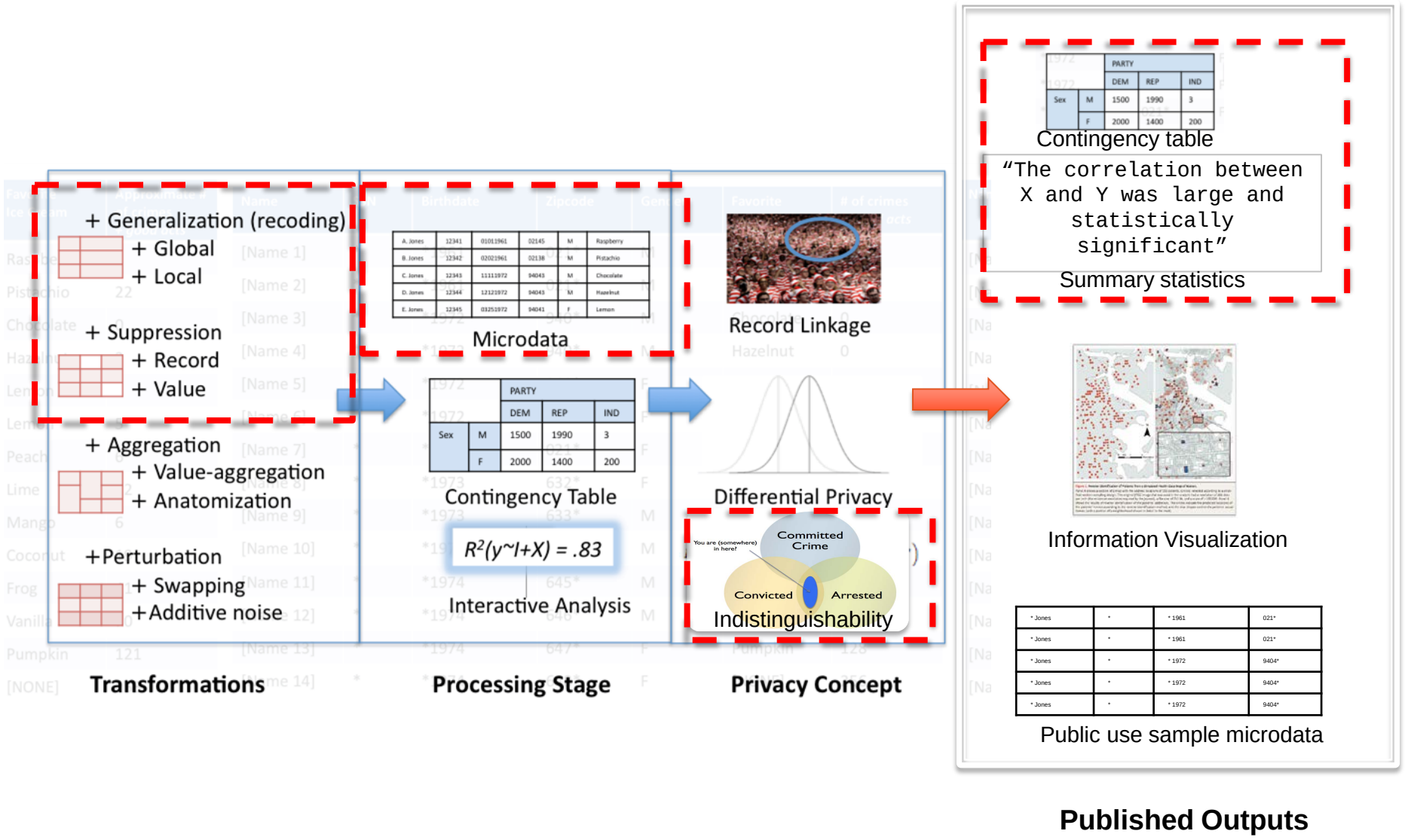
Identifier	*Sensitive* Private Identifier	Private Identifier	Identifier			Sensitive	
Name	SSN	Birthdate	Zipcode	Gender	Favorite Ice Cream	# of crimes committed	
A. Jones	12341	01011961	02145	M	Raspberry	0	Mass resident
B. Jones	12342	02021961	02138	M	Pistachio	0	
C. Jones	12343	11111972	94043	M	Chocolate	0	California resident?
D.Jones	12344	12121972	94043	M	Hazelnut	0	
E. Jones	12345	03251972	94041	F	Lemon	0	Twins, separated at birth?
F. Jones	12346	03251972	02127	F	Lemon	1	
G.Jones	12347	08081989	02138	F	Peach	1	Student record?
H. Smith	12348	01011973	63200	F	Lime	2	
I. Smith	12349	02021973	63300	M	Mango	4	
J. Smith	12350	02021973	63400	M	Coconut	16	
K. Smith	12351	03031974	64500	M	Frog	32	
L. Smith	12352	04041974	64600	M	Vanilla	64	
M.Smith	12353	04041974	64700	F	Pumpkin	128	
N. Smith	12354	04041974	64800	F	Allergic	256	Health information?

Transformed data

Synthetic	Var	Global Recode		Local Suppression		Aggregation + Perturbation
Name	SSN	Birthdate	Zipcode	Gender	Favorite Ice Cream	# of crimes committed
Name 1	~~12341~~	*1961	021*	M	Raspberry	.1
Name 2	~~12342~~	*1961	021*	M	Pistachio	-.1
Name 3	~~12343~~	*1972	940*	M	Chocolate	0
Name 4	~~12344~~	*1972	940*	M	Hazelnut	0
Name 5	~~12345~~	*1972	940*	F	Lemon	.6
Name 6	~~12346~~	*1972	021*	F	Lemon	.6
Name 7	~~12347~~	*1989	021*	*	Peach	64.6
Name 8	~~12348~~	*1973	632*	F	Lime	3
Name 9	~~12349~~	*1973	633*	M	Mango	3
Name 10	~~12350~~	*1973	634*	M	Coconut	37.2
Name 11	~~12351~~	*1974	645*	M	*	37.2
Name 12	~~12352~~	*1974	646*	M	Vanilla	37.2
Name 13	~~12353~~	*1974	647*	F	*	64.4
Name 14	~~12354~~	*1974	648*	~~F~~	~~Allergic~~	~~256~~

Redaction

Disclosure Limitation for Data Privacy

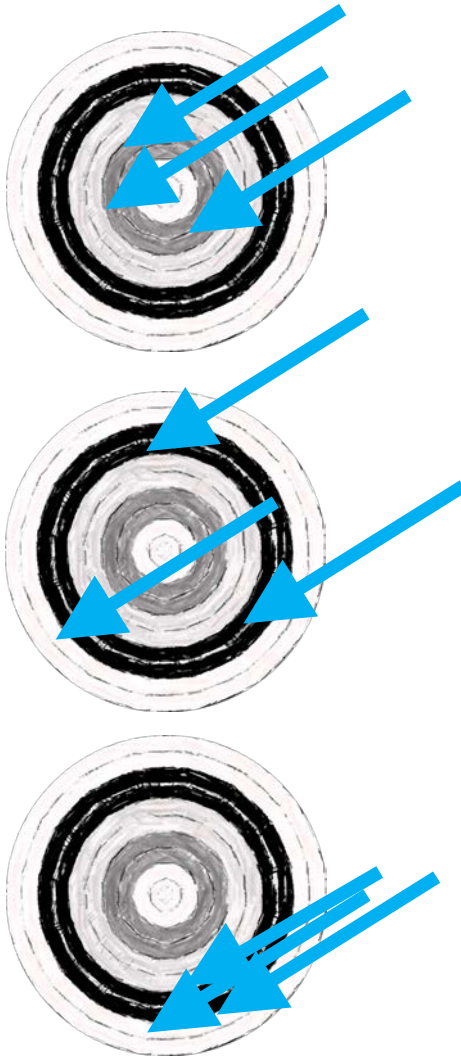


Transformation reduces utility

- Common approach of anonymizing/suppressing/redacting data **reduces usefulness**

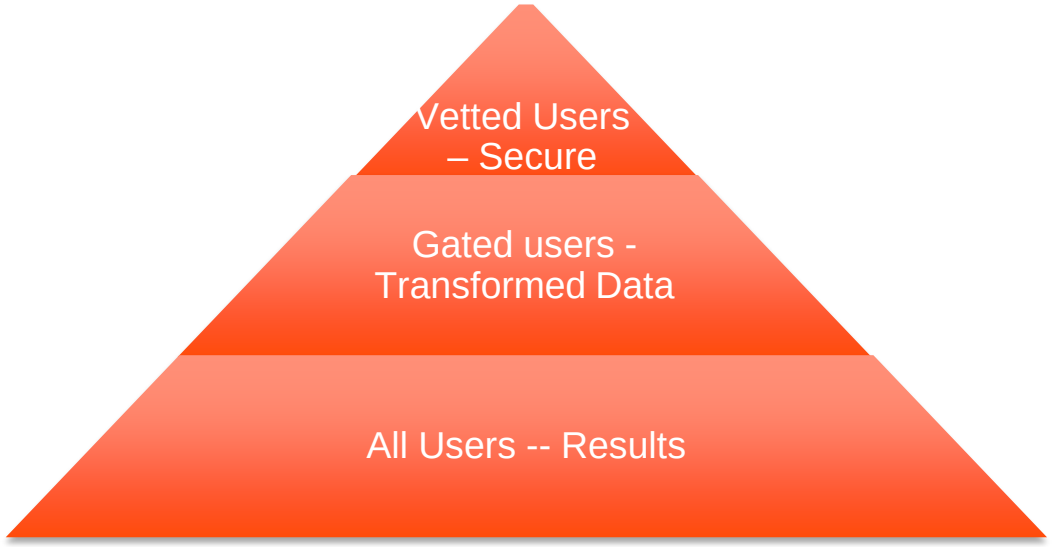
- Minimizing disclosure in the presence of large external data sources **reduces usefulness a lot**

- Anonymized data is not simply less informative -- it typically yields **biased analyses**



Plan for Tiered Access

- Modern approaches to privacy requires planning for **tiered modes of access**



Major Disclosure Limitation Tools
(for tabular Microdata)

	Operating System	Functions	Caveats
sdcMicro	R (Windows/OS-X/Linux)	- Suppression - Perturbation - Micro-aggregation - Synthetic data - Record-linkage risk - Utility analysis - K-anonymity	- No general indistinguishability measures - No differential privacy
ARX	Windows/OS-X/Linux	- Suppression - Perturbation - Micro-aggregation - Synthetic data - Record-linkage risk - Utility analysis - K-anonymity and Indistinguishibility	- Standalone tool - No synthetic data - Limited differential privacy options
µ-ARGUS	Windows/Linux	- Suppression - Perturbation - Micro-aggregation - Synthetic data - Record-linkage risk	- Standalone tool - Steeper learning curve - No utility measures - No k-anonymity or indistinguishability - No differential privacy

Example Suppression with R and sdcmicro

sdcMicro GUI | About/Help | Microdata | Anonymize | Risk/Utility | Export Data | Reproducibility | Undo

Reset the inputdata

Reset inputdata

What do you want to do?

Display Microdata

Explore variables

Reset variables

Use subset of microdata

Convert numeric to factor

Convert variables to numeric

Modify factor variable

Create a stratification variable

Set specific values to NA

Hierarchical data

Loaded microdata

In this tab you can manipulate the data to prepare for setting up an object of class `sdcMicroObj` in the Anonymize tab. The loaded dataset is `testdata` and consists of `4580` observations and `15` variables.

Show 20 entries Search:

urbrur	roof	walls	water	electcon	relat	sex	age	hhcivil	expend	income	savings	ori_hid	sampling_weight	household_weights
2	4	3	3	1	1	1	46	2	90929693	57800000	116258.5	1	100	25
2	4	3	3	1	2	2	41	2	27338058	25300000	279345	1	100	25
2	4	3	3	1	3	1	9	1	26524717	69200000	5495381	1	100	25
2	4	3	3	1	3	1	6	1	18073948	79600000	8695862	1	100	25
2	4	2	3	1	1	1	52	2	6713247	90300000	203620.2	2	100	16.6666666666667
2	4	2	3	1	2	2	47	2	49057636	32900000	1021268	2	100	16.6666666666667
2	4	2	3	1	3	2	13	1	63386309	22700000	8119166	2	100	16.6666666666667
2	4	2	3	1	3	2	19	1	1106874	89100000	9881406	2	100	16.6666666666667
2	4	2	3	1	3	1	9	1	32659507	2087324	7043642	2	100	16.6666666666667
2	4	2	3	1	3	2	16	1	34347609	44100000	4783134	2	100	16.6666666666667
2	4	3	3	1	1	1	65	2	71883547	55500000	7942221	3	100	33.3333333333333
2	4	3	3	1	2	2	60	2	55174345	41200000	4318171	3	100	33.3333333333333
2	4	3	3	1	5	2	6	1	46002021	99600000	2680967	3	100	33.3333333333333
2	4	3	3	1	1	1	34	2	33042094	98400000	3662611	4	100	33.3333333333333
2	4	3	3	1	2	2	31	2	22328588	68900000	6668614	4	100	33.3333333333333
2	4	3	3	1	3	1	3	1	49958473	45600000	8158939	4	100	33.3333333333333
2	4	2	3	1	1	1	40	2	57681859	42800000	7296617	5	100	25
2	4	2	3	1	2	2	40	2	67311078	93200000	3944082	5	100	25
2	4	2	3	1	3	2	9	1	55539882	27900000	8443666	5	100	25
2	4	2	3	1	3	1	16	1	25734672	28500000	9693944	5	100	25
2	4	3	3	1	1	1	44	2	36358184	7954067	7859047	6	100	25
2	4	3	3	1	2	2	37	2	83713597	39900000	2553958	6	100	25
2	4	3	3	1	3	2	7	1	61897713	73000000	4009506	6	100	25
2	4	3	3	1	3	1	12	1	34961268	27000000	4759842	6	100	25
2	4	3	3	1	1	1	48	2	9420392	40700000	4475361	7	100	20

Example: Using SDCMicro from R

```
# setup
> install.packages("sdcMicro",dependencies=TRUE)
> library(sdcMicro)

# load data
> classexample.df<-read.csv("examplesdc.csv", as.is=T,
stringsAsFactors=F,colClasses=c("character","character","character","character","fa
ctor","factor","numeric")

# create a weight variable if needed
> classexample.df$weight<-1


# simple frequency table shows that data is uniquely identified
> ftable(Birthdate~Zipcode,data=classexample.df)
```

Birthdate	01/01/1973	02/02/1973	03/25/1972	04/04/1974	08/08/1989	10/01/1961	11/11/1972	12/12/1972	20/02/1961	30/03/1974
Zipcode										
02127	0	0	1	0	0	0	0	0	0	0
02138	0	0	0	0	1	0	0	0	1	0
02145	0	0	0	0	0	1	0	0	0	0
63200	1	0	0	0	0	0	0	0	0	0
63300	0	1	0	0	0	0	0	0	0	0
63400	0	1	0	0	0	0	0	0	0	0
64500	0	0	0	0	0	0	0	0	0	1
64600	0	0	0	1	0	0	0	0	0	0
64700	0	0	0	1	0	0	0	0	0	0
64800	0	0	0	1	0	0	0	0	0	0
94041	0	0	1	0	0	0	0	0	0	0
94043	0	0	0	0	0	0	1	1	0	0

Using SDCmicro to check identifiability

```
# global recoding

> recoded.df<-classexample.df
> recoded.df$Birthdate<-substring(classexample.df$Birthdate,7)
> recoded.df$Zipcode<-substring(classexample.df$Zipcode,1,3)


# Check if anonymous?
# NOTE makes sure to use column numbers and w=NULL

> print(freqCalc(recoded.df,keyVars=3:5,w=NULL))

 ----------------------------
10 observation with fk=1
4 observation with fk=2
 ----------------------------
```

Using sdcmicro to anonymize data

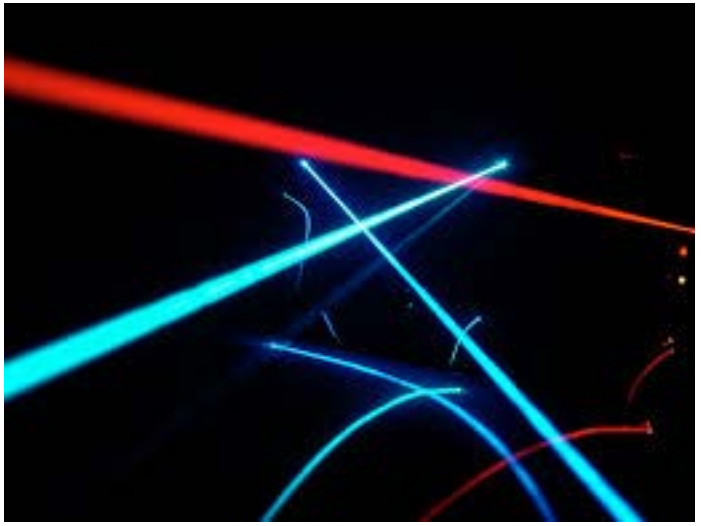
```
# try local suppression with preference for suppressing Gender
> anonymous.out <-
localSupp2Wrapper(recoded.df,3:5,w=NULL,kAnon=2,importance=c(1,1,100))
...
[1] "2-anonymity after 2 iterations."

# look at the data
> as.data.frame(anonymous.out$xAnon)
```

	Name	SSN	Birthdate	Zipcode	Gender	Ice.cream	Crimes	weight
1	A. Jones	12341	1961	021	<NA>	Raspberry	0	1
2	B. Jones	12342	1961	021	<NA>	Pistachio	0	1
3	C. Jones	12343	1972	940	M	Chocolate	0	1
4	D. Jones	12344	1972	940	M	Hazelnut	0	1
5	E. Jones	12345	1972	940	<NA>	Lemon	0	1
6	F. Jones	12346	<NA>	021	<NA>	Lemon	1	1
7	G. Jones	12347	<NA>	021	<NA>	Peach	1	1
8	H. Smith	12348	1973	<NA>	<NA>	Lime	2	1
9	I. Smith	12349	<NA>	633	<NA>	Mango	4	1
10	J. Smith	12350	<NA>	634	<NA>	Coconut	16	1
11	K. Smith	12351	1974	<NA>	<NA>	Frog	32	1
12	L. Smith	12352	<NA>	646	<NA>	Vanilla	64	1
13	M. Smith	12353	<NA>	647	<NA>	Pumpkin	128	1
14	N. Smith	12354	<NA>	648	<NA>	Allergic	256	1

Approaches at the Leading Edge of Disclosure Limitation

- Interactive differentially private analysis systems
  - PSI Tool at Harvard
  - pinq

- Secure virtual data enclaves
  - ICPSR
  - NORC
  - Census



END