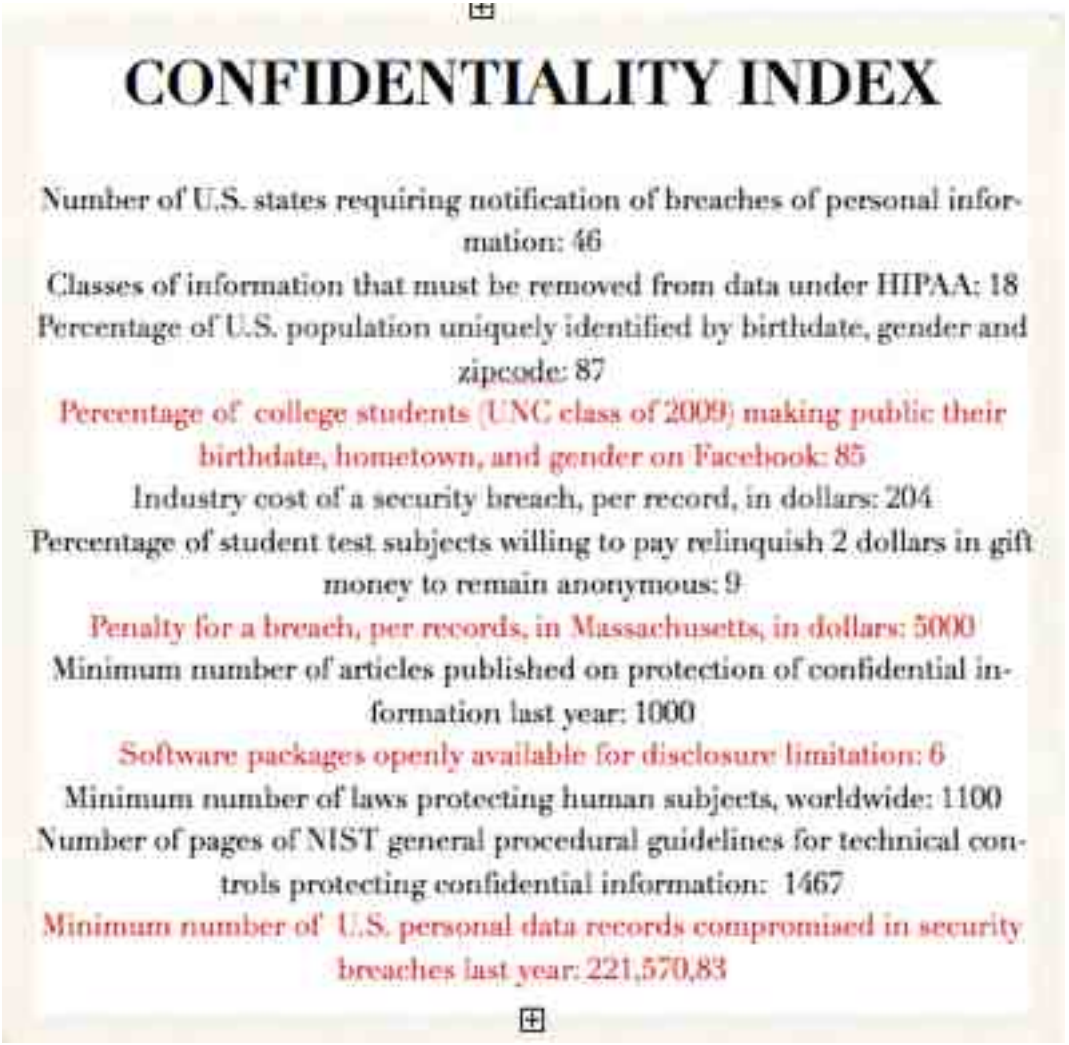
Managing Confidential Information in Research

Micah Altman
Senior Research Scientist
Institute for Quantitative Social Science
Harvard University

Confidential data matters in social science

- Social scientists often collect or use information from people – much of which is not clearly public
- Professional and ethical responsibility to manage subject privacy and confidentiality appropriately
- Possibility of huge administrative penalties



CONFIDENTIALITY INDEX

Number of U.S. states requiring notification of breaches of personal information: 46

Classes of information that must be removed from data under HIPAA: 18

Percentage of U.S. population uniquely identified by birthdate, gender and zipcode: 87

Percentage of college students (UNC class of 2009) making public their birthdate, hometown, and gender on Facebook: 85

Industry cost of a security breach, per record, in dollars: 204

Percentage of student test subjects willing to pay relinquish 2 dollars in gift money to remain anonymous: 9

Penalty for a breach, per records, in Massachusetts, in dollars: 5000

Minimum number of articles published on protection of confidential information last year: 1000

Software packages openly available for disclosure limitation: 6

Minimum number of laws protecting human subjects, worldwide: 1100

Number of pages of NIST general procedural guidelines for technical controls protecting confidential information: 1467

Minimum number of U.S. personal data records compromised in security breaches last year: 221,570,83

[Micah Altman, 3/10/2011]

Why protect confidential information?

- It's your money –
  - 8.3 Million identify theft victims/year
  - Lose a week of your life and $$
- It's your job –
  - Protection is **individual** responsibility
  - Violations can lead to sanctions
- It's your institution --
  - Institutional reputation
  - Loss of external funding
- It's your research --
  - Data management/sharing plans strengthen grant proposals
  - Data management/sharing plans can clarify research design
  - Data management/sharing plans can strengthen research program
  - Confidentiality can improve subject recruitment, responses
- It's the law --
  - Civil penalties
  - Criminal penalties
  - Administrative penalties
- It's the right thing to do --
  - Professional behavior
  - Obligations to research subjects

Researcher fights data breach pay cut

By Liam Tung on Oct 27, 2010 10:24 AM
Filed under Security

Feels "Scapegoated".

A US breast cancer researcher is battling a pay cut and demotion imposed on her by the University of North Carolina after a server she was responsible for was hacked, exposing some 180,000 patient details.

Epidemiologist Bonnie Yankaskas was demoted from full to associate professor and had her pay cut from US$178,000 to US$93,000 after the data breach was detected in 2009, some two years after the incident.

The university had reportedly attempted to fire Yankaskas before demoting her over the incident.

Yankaskas believed the IT department should be held responsible for the security of the server and has taken her case to the university's board.

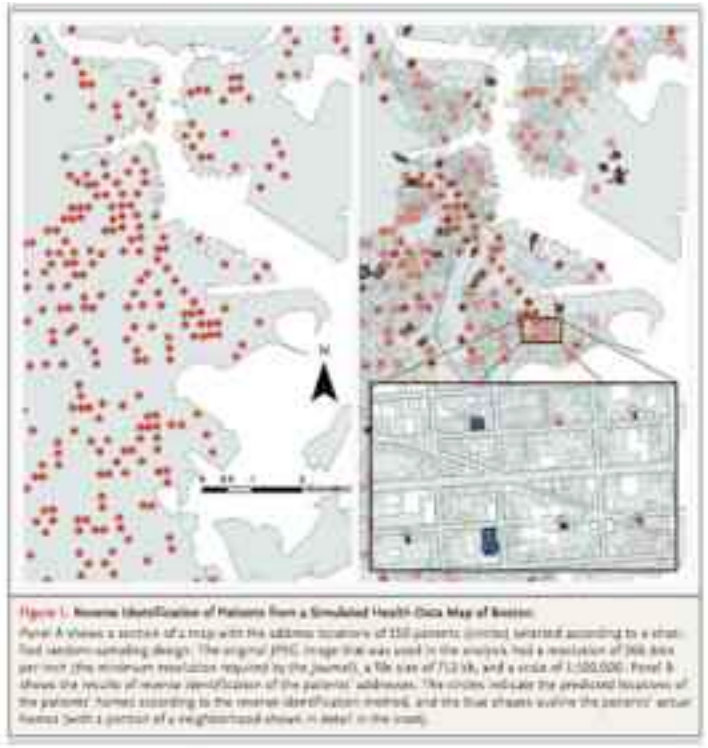
"I clearly have been scapegoated," she told North Carolina publication the News Observer.

"I bear the responsibility for my group doing what's right. But do I bear the responsibility for this machine not being secure? How do you lay that on me?"

[Micah Altman, 3/10/2011]

Personally identifiable private information is surprisingly common

- Includes information from a variety of sources, such as…
  - Research data, even if you aren't the original collector
  - Student "records" such as e-mail, grades
  - Logs from web-servers, other systems
- Lots of things are potentially identifying:
  - Under some federal laws: IP addresses, dates, zipcodes, …
  - Birth date + zipcode + gender uniquely identify ~87% of people in the U.S. [Sweeney 2002]
  - With date and place of birth, can guess first five digits of social security number (SSN) > 60% of the time. (Can guess the whole thing in under 10 tries, for a significant minority of people.) [Aquisti & Gross 2009]
  - Analysis of writing style or eclectic tastes has been used to identify individuals
- Tables, graphs and maps can also reveal identifiable information



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

Traditional approaches are failing

- Modal traditional approach:
  - removing subjects' names
  - storing descriptive information in a locked filing cabinet
  - publishing summary tables
  - (sometimes) release a public use version that suppressed and recoded descriptive information
- Problems
  - law is changing – requirements are becoming more complex
  - research computing is moving towards the cloud
  - social scientists are studying new forms of data that create new privacy issues
  - advances in the formal analysis of disclosure risk imply the impracticality of "de-identification" as required by law

And first, a word from our sponsor...

The Institute
for Quantitative Social Science
at Harvard University

IQSS (and affiliates) offer you support across all stages of your quantitative research:

- Research design, including:
  design of surveys, selection of statistical methods.
- Primary and secondary data collection, including:
  the collection of geospatial and survey data.
- Data management, including:
  storage, cataloging, permanent archiving, and distribution.
- Data analysis, including :
  statistical consulting, GIS consulting, high performance research computing.

http://iq.harvard.edu/

But wait … there's more!

The IQSS grants administration team helps with every aspect of the grant process. Contact us when you are planning your proposal.

- Assisting in identifying research funding opportunities
- Consulting on writing proposals
- Assisting IQSS affiliates with:
  - preparation, review and submission of all grant applications ("pre-award support")
  - management of their sponsored research portfolio ("post-award support")
  - Interpret sponsor policies
  - Coordinate with FAS Research Administration and the Central Office for Sponsored Programs

… And, of course, support seminars like this!

Goals for course

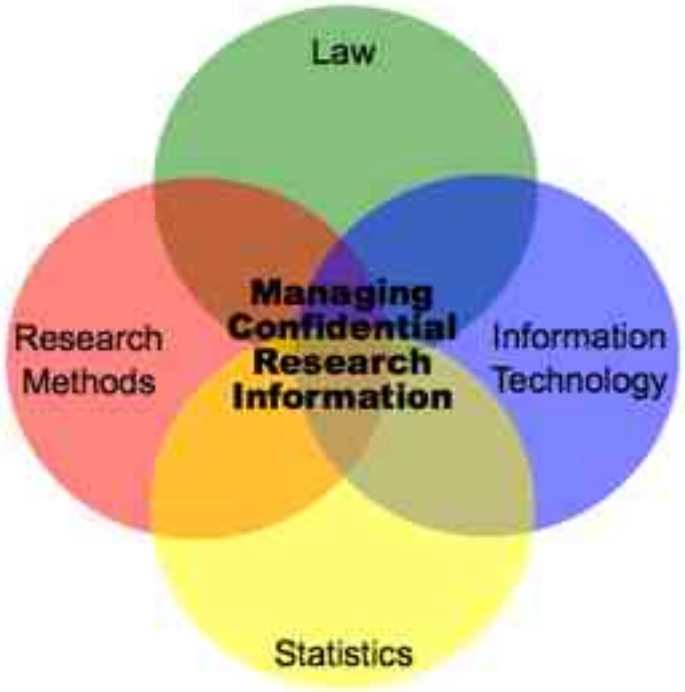
- Overview of key areas
- Identify key concepts & issues
- **Summarize Harvard policies, procedures, resources**
- Establish framework for action
- Provide connection to resources, literature

Outline

- [Preliminaries]

- Law, policy, ethics
- Research methods, design, management
- Information Security (Storage, Transmission, Use)
- Disclosure Limitation

- [Additional Resources & Summary of Recommendations]



Steps to Manage Confidential Research Data

- **Identify** potentially sensitive information in *planning*
  - Identify legal requirements, institutional requirements, data use agreements
  - Consider obtaining a certificate of confidentiality
  - Plan for IRB review
- **Reduce** sensitivity of collected data in *design*
- **Separate** sensitive information in *collection*
- **Encrypt** sensitive information in *transit*
- **Desensitize** information in *processing*
  - Removing names and other direct identifiers
  - Suppressing, aggregating, or perturbing indirect identifiers
- **Protect** sensitive information in *systems*
  - Use systems that are controlled, securely configured, and audited
  - Ensure people are authenticated, authorized, licensed
- **Review** sensitive information before *dissemination*
  - Review disclosure risk
  - Apply non-statistical disclosure limitation
  - Apply statistical disclosure limitation
  - Review past releases and publically available data
  - Check for changes in the law
  - Require a use agreement

[Micah Altman, 3/10/2011]

Handling Confidential Information @ Harvard

- 1. How to identify confidential information
  - Extremely Sensitive/ "Level 5"
    - Research data containing private extremely sensitive information about identifiable individuals
  - High Risk Confidential Information/HRCI/ "Level 4"
    - A person's *name* + *state*, *federal or financial identifiers*
    - *Or* research data containing private very sensitive information about identifiable individuals
  - Harvard Confidential Information/HCI/"Level 3"
    - Business information *specifically designated by the School as confidential*
    - *Or* identifiable *business information* that puts individuals at risk if disclosed
    - *Or* research data containing private sensitive information about identifiable individuals
    - *Or* student records (such as collections of grades, correspondence)
  - Benign Identified Information/ "Level 2"
- 2. **You are responsible** for confidential information you store, access, or share
  - Obtaining appropriate approval to access information
    - Approval from IRB to use or collect private identified information
    - Approval from PI for access to research data
    - Approval from OSP to sign any external data use agreements
  - Protect your system [Level 2+]
    firewall, virus scanner, limit accounts, good passwords, don't share accounts
  - Encrypt all laptops, portable storage, and network connections [Level 3+]
  - Keep hard-copy/media locked up when not in use [Level 3+]
  - Keep data *only* on specifically designated servers [Level 4+]
  - Keep isolated from any external network [Level 5]
- 3. Safely dispose of confidential information
  - When you change computers – contact IT for cleanup
  - Shred the rest
- 4. If unsure, seek help or approval
  - Consulting: http://www.iq.harvard.edu/data_collection_management_analysis
  - Research approvals: cuhs@fas.harvard.edu

[Micah Altman, 3/10/2011]

Law, Policy & Ethics

- Ethical Obligations
- Laws
- Fun and games ☺
- **Harvard Policies**
- [Summary]

Confidentiality & Research Ethics

Law, policy, ethics
Research design …
Information security
Disclosure limitation

- Belmont Principles
  - *Respect for Persons*
    - individuals should be treated as autonomous agents
    - persons with diminished autonomy are entitled to protection
    - implies "informed consent"
    - implies respect for confidentiality and privacy
  - *Beneficence*
    - research must have individual and/or societal benefit to justify risks
    - implies minimizing risk/benefit ratio

Scientific & Societal Benefits of Data Sharing

Law, policy, ethics
Research design …
Information security
Disclosure limitation

- Increases replicability of research
  - Journal publication policies may apply
- Increases scientific impact of research
  - Follow up studies
  - Extensions
  - Citations
- Public interest in data produced by public funder
  - Funder policies may apply
- Public interest in data that supports public policy
  - FOIA and state FOI laws may apply
- Open data facilitates…
  - Transparent government
  - Scientific collaboration
  - Scientific verification
  - New forms of science
  - Participation in science
  - Hands-on education
  - Continuity of research

Sources: Fienberg et. al 1985; ICSU 2004; Nature 2009

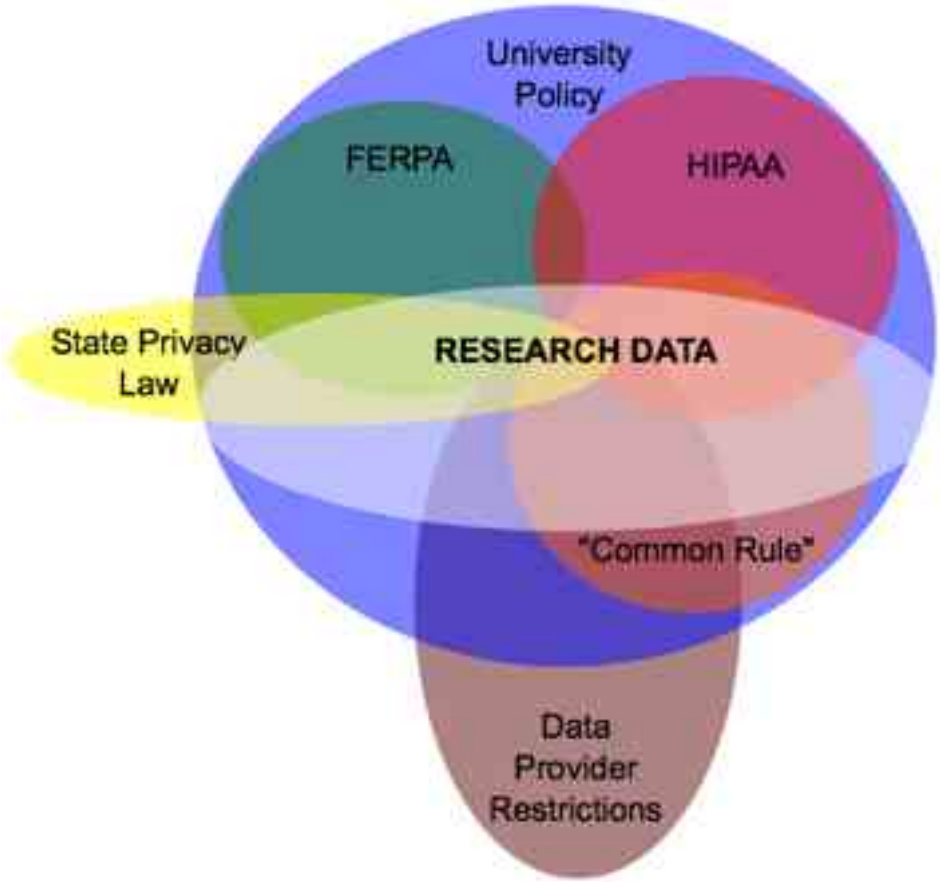
[Micah Altman, 3/10/2011]

Sources of Confidentiality Restrictions for University Research Data

Law, policy, ethics
Research design …
Information security
Disclosure limitation

- Overlapping laws
- Different laws apply to different cases
- All affiliates subject to university policy

(*Not included*: EU directive, foreign laws, classified data, …)



[Micah Altman, 3/10/2011]

45 CFR 46 [Overview]

Law, policy, ethics
Research design …
Information security
Disclosure limitation

"The Common Rule"

- Governs human subject research
  - With federal funds/ at federal institution
- Establishes rules for conduct of research
- Establishes confidentiality and consent requirement for for identified private data
- However, some information may be required to be disclosed under state and federal laws (e.g. in cases of child abuse)
- Delegates procedural decisions to Institutional Review Boards (IRB's)

HIPAA [Overview]

Law, policy, ethics
Research design …
Information security
Disclosure limitation

Health Insurance Portability and Accountability Act

- Protects personal health care information for 'covered entities'
- Detailed technical protection requirements
- Provides clearest legal standards for dissemination
- Provides a 'safe harbor'
- Has become an accepted practice for dissemination in other areas where laws are less clear
- *HITECH Act of 2009 extends HIPAA*
  - Extends coverage to associated entities of covered entities
  - Additional technical safeguards
  - Adds breach reporting requirement

HIPAA provides three dissemination options …

Dissemination under HIPAA [option 1]

Law, policy, ethics
Research design …
Information security
Disclosure limitation

- "safe harbor" -- remove 18 identifiers
  - [Personal identifiers]
    - Names
    - Social Security #'s; Personal Account #'s; Certificate/License #'s; full face photos (and comparable images); biometric id's; medical
    - Any other unique identifying number, characteristic, or code
  - [Asset identifiers]
    - fax #'s; phone #'s; vehicle #'s;
    - personal URL's; *IP addresses*; e-mail addresses
    - Device ID's and serial numbers
  - [Quasi identifiers]
    - dates smaller than a year (and ages > 89 collapsed into one category)
    - geographic subdivisions smaller than a state (except for 3 digits of zipcode, if unit > 20,000 people)

And

- Entity does not have *actual knowledge* [direct and clear awareness] that it would be possible to use the remaining information alone or in combination with other information to identify the subject

Dissemination under HIPAA [Option 2]

Law, policy, ethics
Research design …
Information security
Disclosure limitation

- "limited dataset" – leave some quasi-id's
  - Remove personal and asset identifiers
  - Permitted dates: dates of birth, death, service, years
  - Permitted geographic subdivisions: town, city, state, zip code

  And
  - Require access control and data use agreement.

[Micah Altman, 3/10/2011]

Dissemination under HIPAA [Option 3]

Law, policy, ethics
Research design …
Information security
Disclosure limitation

- "qualified statistician" – statistical determination

  Have *qualified statistician* determine, using *generally accepted* statistical and scientific principles and methods, that the risk is *very small* that the information could be used, alone *or in combination with other reasonably available information*, by the anticipated recipient to identify the subject of the information.

- Important caveats
  - Methods and results of the analysis must be documented
  - No bright line for "qualified", text of rule is:
    "a person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable." [Section 164.514 (b)(1)]
  - No clear definitions for "generally accepted" or "very small" or "reasonably available information"…
    however, there are references in the federal register to statistical publications to be used as "starting points"

[Micah Altman, 3/10/2011]

FERPA

Law, policy, ethics
Research design …
Information security
Disclosure limitation

Family Educational Rights and Privacy Act

- Applies schools that receive federal (D.O.E.) funding
- Restricts use of student (not employee) information
- Establishes
  - Right to privacy of educational records
  - Right to inspect and correct records (with appeal to Federal government)
  - Definition of public "directory" information
  - Right to block access to public "directory" information, and to other records
- Educational records include:
  - Identified information about student
  - Maintained by institution
  - *Not …*
    - *Employee records*
    - *Some medical and law-enforcement records*
    - *Records solely in the possession and for use by the creator (e.g. unpublished instructor notes)*
- Personally identifiable information includes:
  - Direct identifiers
  - Indirect (quasi) identifiers
  - Indirectly linkable identifiers
  - "Information requested by a person who the educational agency or institution reasonably believes knows the identity of the student to whom the education record relates."

MA 201 CMR 17

Law, policy, ethics
Research design …
Information security
Disclosure limitation

Standards for the Protection of Personal Information

- Strongest U.S. general privacy protection law
- Has been delayed/modified repeatedly
- Requires reporting of breaches
  - If data is not encrypted
  - Or encryption key is released in conjunction with data
- Requires specific technical protections:
  - Firewalls
  - Encryption of data transmitted in public
  - Anti-virus software
  - Software updates

Inconsistencies in Requirements and Definitions

Law, policy, ethics
Research design …
Information security
Disclosure limitation

- Inconsistent definitions of "personally identifiable"
- Inconsistent definitions of sensitive information
- Requirements for to de-identify jibes with statistical realities

	FERPA	HIPAA	Common Rule	MA 201 CMR 17
Coverage	Students in Educational Institutions	Medical Information in "Covered Entities"	Living persons in research by funded institutions	Mass. Residents
Identification Criteria	-Direct -Indirect -Linked **-Bad intent (!)**	-Direct -Indirect -Linked	-Direct -Indirect -Linked	-Direct
Sensitivity Criteria	Any non-directory information	Any medical information	Private information – based on harm	Financial, State, Federal Identifiers
Management Requirements	- Directory opt-out - [Implied] good practice	- Consent - Specific technical safeguards - **Breach notification**	- Consent - [Implied] risk minimization	- Specific technical safeguards - **Breach notification**

Third Party Requirements

Law, policy, ethics
Research design …
Information security
Disclosure limitation

- Licensing requirements
- Intellectual property requirements
- Federal/state law and/or policy requirements
  - State protection of personal information laws
  - Freedom of information laws (FOIA & State FOI)
  - State mandatory abuse/neglect notification laws
- And … think ahead to publisher requirements
  - Replication requirements
  - IP requirements

- Examples
  - NSF requires data from funded research be shared
  - NIH requires a data sharing plan for large projects
  - Wellcome Trust requires a data sharing plan
  - Many leading journals require data sharing

[Micah Altman, 3/10/2011]

(Some) More Laws & Standards

Law, policy, ethics
Research design …
Information security
Disclosure limitation

- California Laws
  - Lots of rules
  - Applies any data about California *residents*
  - Privacy policy
  - Disclosure
  - Reporting policy
- EU Directive 95/46/EC
  - Data protection directive
  - Provides for notice, limits on purpose of use, consent, security, disclosure, access, accountability
  - Forbids transfer of data to entities in countries compliant with directive
  - U.S. is not compliant *but* …
    - Organizations can certify compliance with FTC
    - No auditing/enforcement !
    - Substantial criticism of this arrangement
- Payment Card Industry (PCI) Security Standards
  - Governs treatment of credit card numbers
  - Requires reports, audits, fines
  - Detailed technical measures
  - Not a law, but helps define good practice
  - Nevada law mandates PCI standards
- FISMA
  - Federal Information Security Management Act (FISMA), Public Law (P.L.) 107-347.
  - Is starting to be applied to NIH sponsored research
  - Detailed technical controls over inform
- Sarbanes-Oxley (aka, SOX, aka SARBOX)
  - Corporate and Auditing Accountability and Responsibility Act of 2002
  - Applies to U.S. **public company** boards, management and public accounting firms
  - Rarely applies to research in universities
  - Section 404 requires annual assessment of organizational internal controls – but does not specify details of controls
- Classified Data
  - Separate and complex rules and requirements
  - *The University does not accept classified data*
  - But, may have "Controlled But Unclassified"
    - Vaguely defined area
    - Mostly government produced area
    - Penalties unclear
  - And… export controlled information, under ITAR and EAR
    - Export control include technologies, software, documentation/design documents may be included
    - Large penalties
- **… and over 1100 International Human Subjects laws…**

[Micah Altman, 3/10/2011]

Predicted Legal Changes for 2011…

Law, policy, ethics
Research design …
Information security
Disclosure limitation

- Caselaw
  - "personal privacy" does *not* apply to information about corporations (a corporation is not a "person" for this purpose) *FCC vs. ATT 2011*
- Scheduled
  - EU "cookie privacy" directive 2009/136/EC goes into effect
  - Proposed updates to EU information privacy directives
- Very Likely
  - New information privacy laws in selected states in 2011
- Likely
  - Increased federal regulation of internet privacy

[Micah Altman, 3/10/2011]

What's wrong with this picture?

Law, policy, ethics
Research design …
Information security
Disclosure limitation

Name	SSN	Birthdate	Zipcode	Gender	Favorite Ice Cream	# of crimes committed
A. Jones	12341	01011961	02145	M	Raspberry	0
B. Jones	12342	02021961	02138	M	Pistachio	0
C. Jones	12343	11111972	94043	M	Chocolate	0
D. Jones	12344	12121972	94043	M	Hazelnut	0
E. Jones	12345	03251972	94041	F	Lemon	0
F. Jones	12346	03251972	02127	F	Lemon	1
G. Jones	12347	08081989	02138	F	Peach	1
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-Jones	12354	04041974	64800	F	Allergic	256

[Micah Altman, 3/10/2011]

What's wrong with this picture?

Law, policy, ethics
Research design …
Information security
Disclosure limitation

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	Californian
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	FERPA too?
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	Unexpected Response?

[Micah Altman, 3/10/2011]

Harvard: Categorizations

Law, policy, ethics
Research design …
Information security
Disclosure limitation

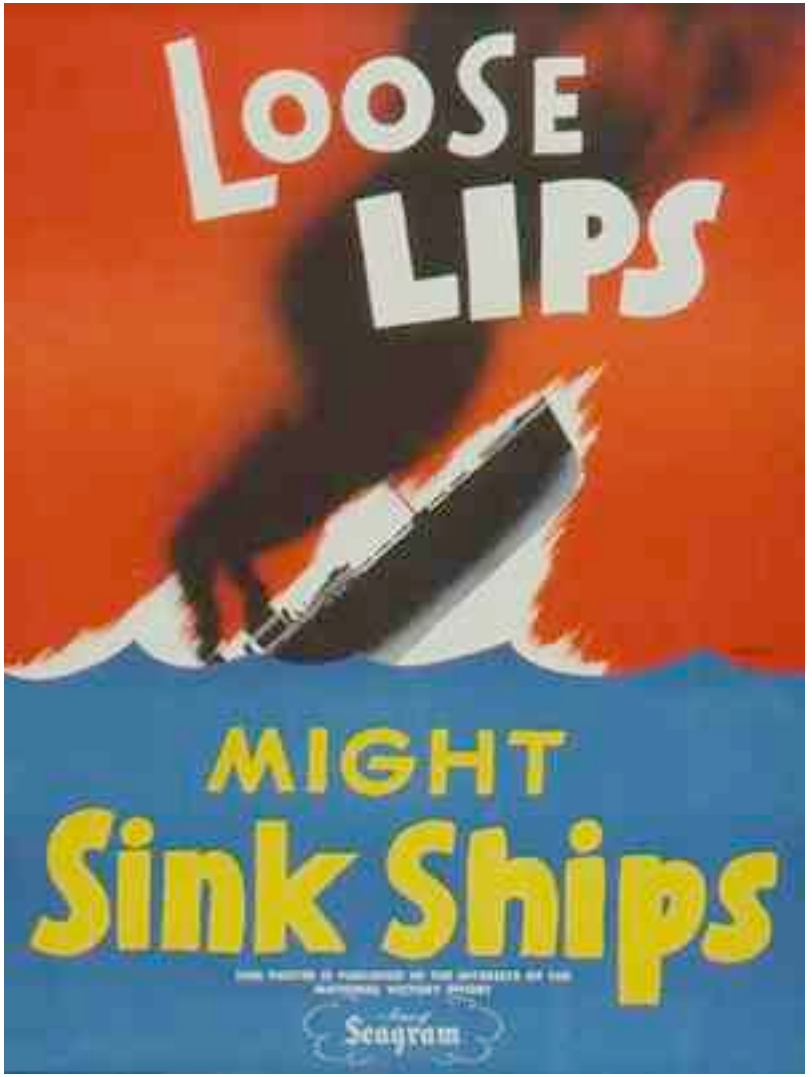
- High Risk, Extremely Sensitive (LEVEL 5)
  - Research data linked to individuals
  - Private information that creates a major risk of harm to subject if made public
  - **No access outside protected physical area allowed**
- High Risk Confidential Information (LEVEL 4 - HRCI)
  - A person's *name* + state, federal, or financial identifiers
    Or research data containing very sensitive private information about identifiable individuals
  - **Must be stored on designated professionally secured servers**
- Harvard Confidential Information (Level 3 - HCI)
  - Business information *specifically designated by the School as confidential*

    Or identifiable *business information* that puts individuals at risk if disclosed

    Or non-trivial student records (such as collections of grades, correspondence)

    Or research data containing sensitive private information about identifiable individuals

  - **Must be encrypted *or* stored in physically secure Harvard location**
- Benign Research Information about Identified Individuals (Level 2)
  - Research data, linked to individuals, but minimally harmful if disclosed
  - **Must use "good information hygiene" such as virus checkers, strong passwords, host firewalls**



[Micah Altman, 3/10/2011]

INFORMATION SECURITY & PRIVACY HARVARD

Security Mandates

Law, policy, ethics
Research design …
Information security
Disclosure limitation

1. Training
2. Comprehensive Communication
3. Laptop Encryption
4. Finding HRCI
5. Vulnerability Testing
6. Network Requirements
7. Remote Access
8. Standard File Transfer
9. Non-Administrative System Certification
10. Managing Security and Practices

On the Horizon

- FAS has drafted procedures for:
  - systems admin
  - remote access
  - scanning computers
  - data disposal
  - Vendors
  - Scanning
- More policies in development

[security.harvard.edu/university-security-mandates]

Harvard: Enterprise Security Policy (HEISP)

Law, policy, ethics
Research design …
Information security
Disclosure limitation

- Storing High Risk Confidential Information (HRCI)
  - Must not be stored on individual user computer or portable storage device
  - Must be stored on "target computers" or secure locked containers
- Human subject information
  - All research on human subjects must be approved by the IRB
  - All proposals must include a data management plan
- Personally identifiable medical information (PIMI)
  - "Covered entities" at Harvard are subject to HIPAA requirements
  - PIMI is to be treated as HRCI throughout the university
- Obtaining confidential information requires approval
- All confidential information must be encrypted when transported across any network
- Public directories must adhere to privacy preferences establishes by the individuals
- Identifying Users with Access to Confidential Information
  - System owners must be able to identify users that have access to confidential information
  - Strong passwords
  - No account/password sharing
- Inhibit password guessing with logging and lockouts
- Limit application availability time with timeouts
- Limit user access to confidential information based on business need

- Confidential information on Harvard computing devices
  - confidential information must be protected
  - confidential information on portable devices must be encrypted
  - laptops must have encryption (some schools require whole-disk encryption)
  - systems must be scanned annually
- Cannot save confidential information on computer directly accessible from the internet, open Harvard networks
- Employees who have access must annually agree to confidentiality agreements
- Access to lists and database of Harvard University ID numbers is restricted
- Each school must provide training
- Registrars have developed common definition of FERPA directory information
- Must adhere to student requests to block their directory information, per FERPA
- Accepting Payment Cards - Restricted to procedures outlined in HU Credit Card Merchant Handbook

[More on next page…]

HEISP – Part 2

Law, policy, ethics
Research design …
Information security
Disclosure limitation

- Physical Environment
  - All digital/non-digital media must be properly protected
- Computers must be physically secure
- Automatic logging must be consistent with written policies
- Vendor contracts
  - require approval by security officer
  - Include OGC contract rider
- Computer operator
  - computer must be regularly updated
  - operated securely
  - Only necessary application installed
  - annually certify compliance with university policies
- Computer setup - must filter malicious traffic
- "Target" systems and controllers
  - Private address space; locally firewalled
  - Annual vulnerability scanning
- Network take down
  - Network managers run vulnerability scans
  - May take computers off the network
- Service Resumption
  - Must have a service resumption plan if loss of confidential data is a substantial business risk
- Incident Response Policy
- Disposition and destruction of records
- Acquisition/use by unauthorized persons must be reported to OGC
- Interacting with legal authorities -- always refer to OGC unless imminent health/safety risk requires otherwise
- Web based surveys must have protections in place

[Micah Altman, 3/10/2011]

Law, policy, ethics
Research design …
Information security
Disclosure limitation

Harvard: Research Data Security Policy (HRDSP)

- Sensitivity of research data based on potential harm if disclosed:
  - Level 5 = "extremely sensitive"
  - Level 4 = "very sensitive" ~= HRCI
  - Level 3 = "sensitive" ~= HCI
  - Level 2 = "benign" ~= Good computer hygiene
  - Level 1 = anonymous and not business confidential
- Required protections based on sensitivity
  - Level 5: Entirely disconnected from network ("bubble security")
  - Level 4: Protections as per HRCI
  - Level 3: Protections as per HCI
  - Level 2: Good computer hygiene
- Designates procedures for treatment of external data use agreements [next section]
  - Legally binding
  - Can be both very detailed and not supported by Harvard security procedures
  - *Investigator should not sign these – forward to OSP*
- Designates responsibilities for IRB, Investigator, OSP, IT, Security Officers.

security.harvard.edu/research-data-security-policy

Harvard: Researcher Responsibilities

Law, policy, ethics
Research design …
Information security
Disclosure limitation

- … for knowing the rules
- … for identifying potentially confidential information *in all forms (digital/analogue; on-line/off-line)*
- … for notifying recipients of their responsibility to protect confidentiality
- … for obtaining IRB approval for any human subjects research
- … for following an IRB approved plan
- … for obtaining OSP approval of restricted data use agreements with providers, even if no money involved

… *and for proper*

- Storage
- Access
- Transmission
- Disposal

Confidentiality is not an "IT problem"

Harvard:
Staff – Personnel Manual

Law, policy, ethics
Research design …
Information security
Disclosure limitation

- Protect Harvard information and systems
- Keep your own information in Peoplesoft up to date
- Comply with copyrights and DMCA
- Comply with Harvard systems policies and procedures
- All information produced at work is Harvard property
- Attach only approved devices to the Harvard network

harvie.harvard.edu/docroot/standalone/Policies_Contracts/Staff_Personnel_Manual/Section2/Privacy.shtml

Key Concepts & Issues Review

Law, policy, ethics
Research design …
Information security
Disclosure limitation

- Privacy
  - Control over extent and circumstances of sharing
- Confidentiality
  - Treatment of private, sensitive information
- Sensitive information
  - Information that would cause harm if disclosed and linked to an individual
- Personally/individually identifiable information
  - Private information
  - Directly or indirectly linked to an identifiable individual
- Human subjects

  A living person …
  - who is interacted with to obtain research data
  - who's private identifiable information is included in research data
- Research
  - Systematic investigation
  - Designed to develop or contribute to generalizable knowledge
- "Common Rule"
  - Law governing funded human subjects research
- HIPAA
  - Law governing use of personal health information in covered and associated entities
- MA 201 CMR 17
  - Law governing use of certain personal identifiers for Massachusetts residents

Checklist: *Identify Requirements*

Law, policy, ethics
Research design …
Information security
Disclosure limitation

Check if research includes …

- ✓ Interaction with humans → Common Rule & HEISP/HRDSP applies

Check if data used includes identified …

- ✓ Student records → FERPA & HEISP/HRDSP applies
- ✓ State, federal, financial id's → state law & HEISP/HRDSP applies
- ✓ Medical/health information → HIPAA (likely) & HEISP/HRDSP applies
- ✓ Human subjects & private info
  → Common Rule & HEISP/HRDSP applies

Check for other requirements/restriction on data dissemination:

- ✓ Data provider restrictions *and* University approvals thereof
- ✓ Open data requirements and norms
- ✓ University information policy

Resources

Law, policy, ethics
Research design …
Information security
Disclosure limitation

- **E.A. Bankert & R.J. Andur, 2006, *Institutional Review Board: Management and Function,* Jones and Bartlett Publishers**
- **P. Ohm, "Broken Promises of Privacy", SSRN Working Paper [ssrn.com/abstract=1450006]**
- D. J. Mazur, 2007. *Evaluating the Science and ethics of Research on Humans,* Johns Hopkins University Press
- *IRB: Ethics & Human Research* [Journal], Hastings Press
  www.thehastingscenter.org/Publications/IRB/
- *Journal of Empirical Research on Human Research Ethics*, University of California Press
  ucpressjournals.com/journal.asp?j=jer
- 201 CMR 17 text
  www.mass.gov/Eoca/docs/idtheft/201CMR17amended.pdf
- FERPA Website
  www.ed.gov/policy/gen/guid/fpco/ferpa/index.html
- HIPAA Website
  www.hhs.gov/ocr/privacy/
- Common Rule Website
  www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm
- State laws
  www.ncsl.org/Default.aspx?TabId=13489
- **Harvard Enterprise Information Security Policy/ Research Data Security Policy**
  www.security.harvard.edu
- **Harvard Institutional Review Board**
  www.fas.harvard.edu/~research/hum_sub/
- **Harvard FAS Policies and Procedures**
  www.fas-it.fas.harvard.edu/services/catalog/browse/39
- **IQSS Policies and Procedures**
  support.hmdc.harvard.edu/kb-930/hmdc_policies

Research design, methods, management

- Reducing risk
  - Sensitivity of information
  - Partitioning
- Decreasing identification
- Managing confidentiality and dissemination
- [Summary]

Trade-offs

- Anonymity vs. research utility
- Sensitivity vs. research utility
- (Anonymity * Sensitivity) vs. research costs/efforts

[Micah Altman, 3/10/2011]

Types of Sensitive Information

- Information is sensitive, if, once disclosed there is a "significant" likelihood of harm
- IRB literature suggests possible categories of harm:
  - loss of insurability
  - loss of employability
  - criminal liability
  - psychological harm
  - social harm to a vulnerable *group*
  - loss of reputational harm
  - emotional harm
  - dignitary harm
  - physical harm: risk of disease, injury, or death

Levels of sensitivity

- No widely accepted scale
- Publicly available data not sensitive under "common rule"
- Common rule anchors scale at "minimal risk":

  "if disclosed, the probability and magnitude of harm or discomfort anticipated are not greater in and of themselves than those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests"

- **Harvard Research Data Security Policy**
  - Level 5- Extremely sensitive information about individually identifiable people.

    Information that if exposed poses ***significant*** risk of **serious harm.**

    Includes information posing serious risk of criminal liability, serious psychological harm or other significant injury, loss of insurability or employability, or significant social harm to an individual or group.

  - Level 4 - Very sensitive information about individually identifiable people

    Information that if exposed poses a ***non-minimal*** risk of **moderate** harm.

    Includes civil liability, moderate psychological harm, or material social harm to individuals or groups, medical records not classified as Level 5, sensitive-but-unclassified national security information, and financial identifiers (as per HRCI standards).

  - Level 3- Sensitive information about individually identifiable people

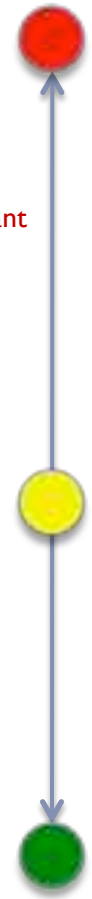
    Information that if disclosed poses a significant risk of minor harm.

    Includes information that would reasonably be expected to damage reputation or cause embarrassment; and FERPA records.

  - Level 2 – Benign information about individually identifiable people

    Information that would not be considered harmful, but that as to which a subject has been promised confidentiality.

  - Level 1 – De-identified information about people, and information not about people



[Micah Altman, 3/10/2011]

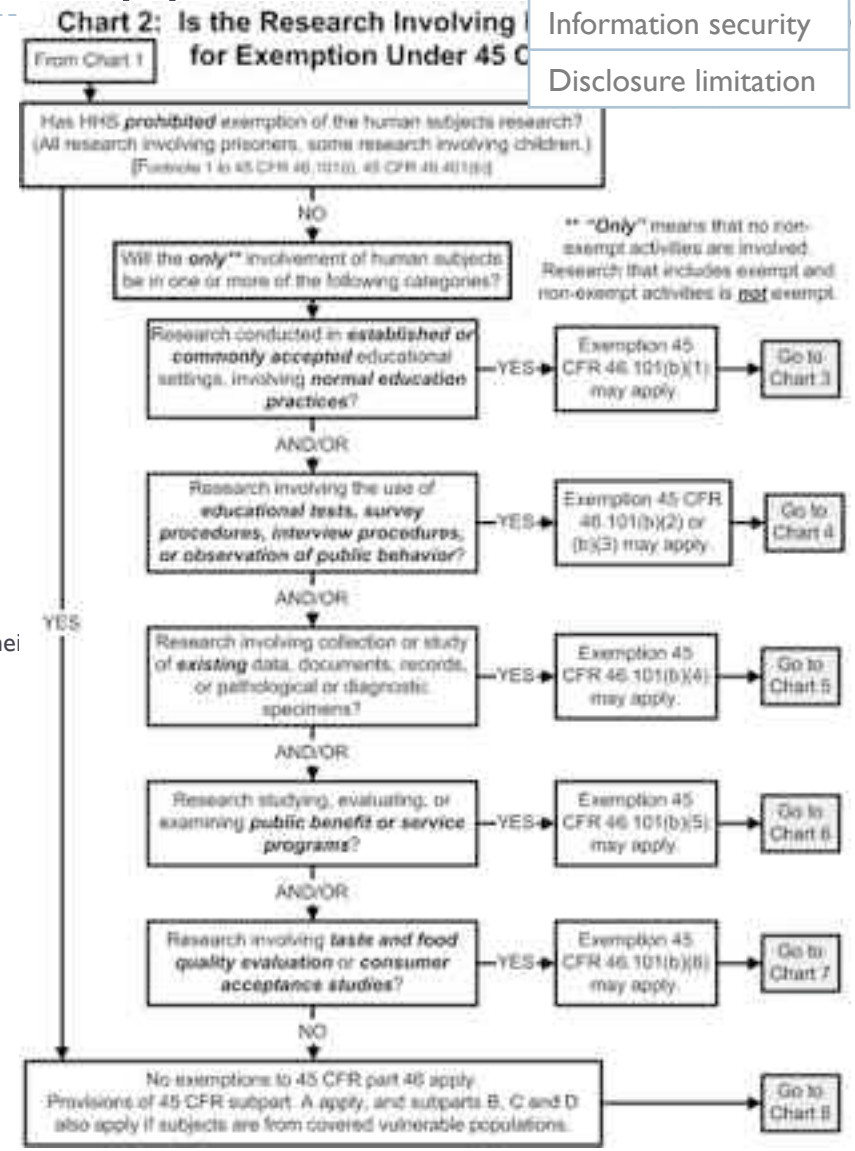
IRB Review Scope

- **IRB approval needed for all:**
  - **federally-funded research;**
  - **or any research at (almost all) institutions receiving federal funding that involve "human subjects".**
    (Any organization operating under a *general* "federal-wide assurance")
  - **All human subjects research at Harvard**
- **Human subject: individual about whom an investigator (whether professional or student) conducting research obtains**
  - **(1) Data through intervention or interaction with a living individual, or**
  - **(2) Identifiable private information about living individuals**
- **See**

www.hhs.gov/ohrp/

Research *not* requiring IRB approval

- Non-research:
  not generalizable knowledge & systematic inquiry
- Non-funded:
  institution receives no federal funds for research
- Not human subject:
  - *No living people described*
  - *Observation only **AND** no private identifiable information is obtained*
- Human Subjects, but "exempt" under 45 CFR 46
  - use of existing, publicly-available data
  - use of existing non-public data, if data is individuals cannot be *directly or indirectly* identified
  - research conducted in educational settings, involving normal educational practices
  - taste & food quality evaluation
  - federal program evaluation approved by agency head
  - observational, survey, test & interview of public officials and candidates (in thei formal capacity, or not identified)
- *Caution* not all "exempt" is exempt…
  - Some research on prisoners, children, not exemptable
  - Some universities require review of "exempt" research
- **Harvard requires review of all human subject research**
- See:
  www.hhs.gov/ohrp/humansubjects/guidance/decisioncharts.htm



[Micah Altman, 3/10/2011]

IRB's and Confidential Information

- IRB's review consent procedures and documentation
- IRB's may review data management plans
  - May require procedures to minimize risk of disclosure
  - May require procedures to minimize harm resulting from disclosure
- IRB's make determination of *sensitivity of information* -- potential harm resulting from disclosure
- IRB's make determination regarding whether data is de-identified for "public use"
  [see NHRPAC,
  "Recommendations on Public Use Data Files"]

Harvard IRB Approval

- The Harvard Institutional Review Board (IRB) must approve all *human subjects research* at Harvard prior to data collection or use
- Research involves human subjects if:
  - There is any interaction or intervention with living humans; *or*
  - If identifiable private data about living humans is used
- Some examples of human subject research in soc sci:
  - Surveys
  - Behavioral experiments
  - Educational tests and evaluations
  - Analysis of identified private data collected from people (your e-mail inbox, logs of web-browsing activity, facebook activity, ebay bids …)
- The IRB will:
  - Assess research protocol
  - Identify whether research is *exempt* from further review and management
  - Identify sensitivity level of data

[Micah Altman, 3/10/2011]

Harvard Responsibilities

- The Harvard Institutional Review Board (IRB) must approve all *human subjects research* at Harvard prior to data collection or use
- Research involves human subjects if:
  - There is any interaction or intervention with living humans; *or*
  - If identifiable private data about living humans is used
- Some examples of human subject research in soc sci:
  - Surveys
  - Behavioral experiments
  - Educational tests and evaluations
  - Analysis of identified private data collected from people (your e-mail inbox, logs of web-browsing activity, facebook activity, ebay bids …)
- The IRB will:
  - Assess research protocol
  - Identify whether research is *exempt* from further review and management
  - Identify sensitivity level of data

[Micah Altman, 3/10/2011]

HRDSP Responsibilities

- Responsibilities
  - Researchers are responsible for disclosing to IRB, and follow IRB approved plan
  - IRB is responsible for ensuring adequacy of Investigators plans; granting (lawful) variances from security requirements justified by research needs
  - IT is responsible for assisting with the identification of security level, and assisting in the implementation of security protections
  - Security Officer/CIO may review IT facilities and approve (give written designation) that they meet protections for a given level

Valuation of private information is uncertain

- Privacy valuations often inconsistent
  - Framing effects: ordering, endowment effect, possibly others
  - Non-normal/uniform distribution of valuations
  - One study: < 10% of subjects would give up $2 of a $12 gift card to buy anonymity of purchases

  [Aquesti and Lowenstein 2009]
- Cost benefit of information security may not be optimal for users [Herley 2009]
  - E.g. Loss from all phishing attacks is 100x less than time spent in avoiding them
  - Note, however weaknesses in this analysis:
    - Only loss of time modeled – no valuation of privacy made
    - Institutional costs not included – only personal costs
    - Very simplified model – not calibrated through surveys etc.
- Repeated surveys of students show they tend to disclose a lot, e.g.:
    - >80% of students sampled in several studies had public facebook pages with birthdays, home town and other private information
      - This information can easily be used to link to other databases!
      - Disclosure of extensive information on sexual orientation, private cell #'s, drinking habits, etc. etc. not uncommon

    [See Kolek & Saunders 2008]
- Emerging markets for privacy?
  - Micropayments for disclosures
  - http://www.personal.com/
  - http://www.i-allow.com/

[Micah Altman, 3/10/2011]

Reducing Risk in Data Collection

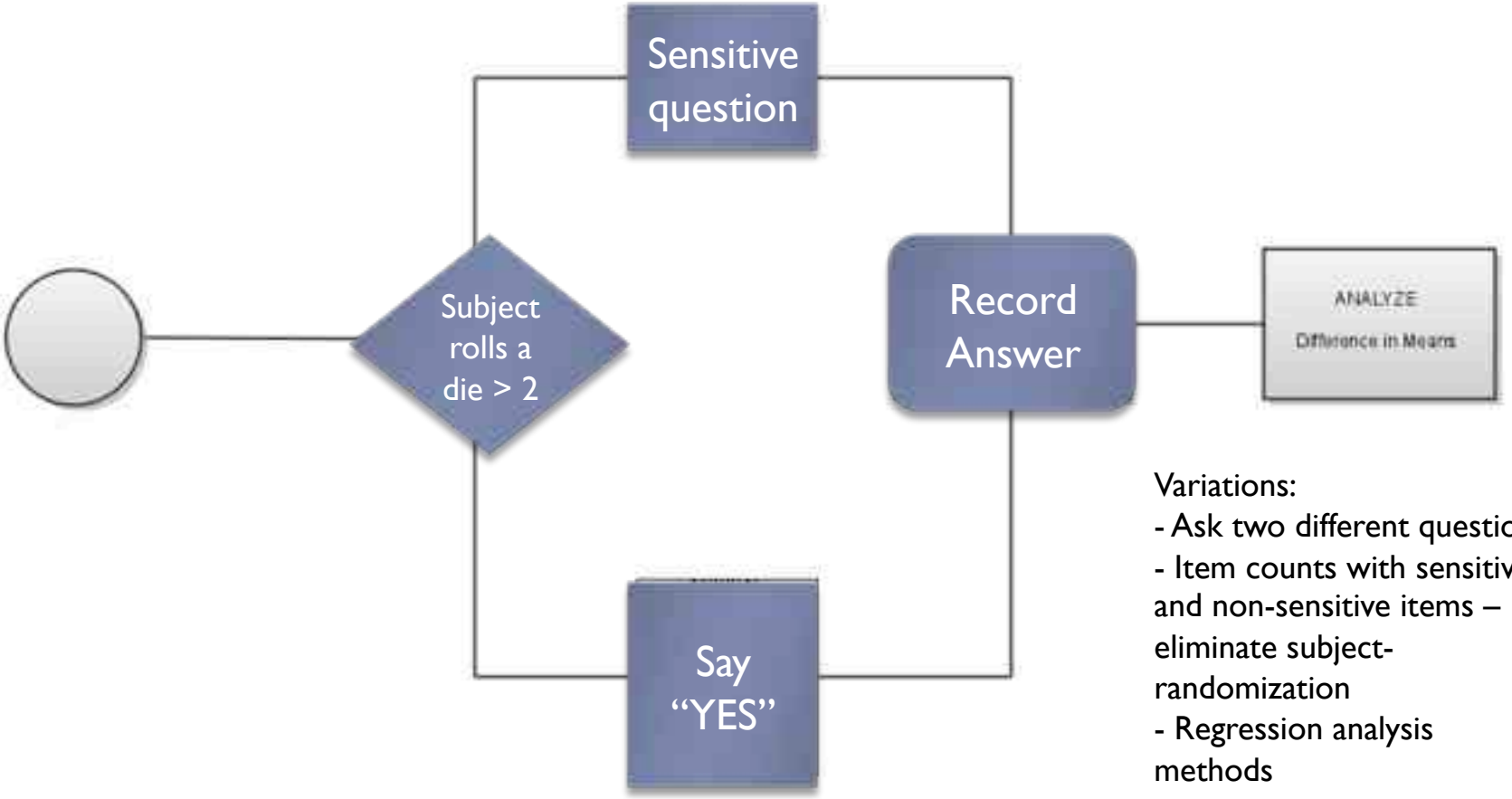
- Avoid collecting sensitive information, unless it is required by research design, method, or hypothesis
  - Unnecessary sensitive information → not minimal risk
  - Reducing sensitivity → higher participation, greater honesty
- Collect sensitive information in private settings
  - Reduces risk of disclosure
  - Increases participation
- Reduce sensitivity through indirect measures
  - Less sensitive proxies
    - E.g. Implicit association test [Greenwald, et al. 1998]
  - Unfolding brackets
  - Group response collection
  - Random response technique [Warner 1965]
  - Item count/unmatched count/list experiment technique

Managing Sensitive Data Collection

- Separate:
  sensitive measures, (quasi)-identifiers, other measures
- If possible avoid storing identifiers with measures:
  - Collect identifying information beforehand
  - Assign opaque subject identifiers
- For sensitive data:
  - Collect on-line directly (with appropriate protections); *or*
  - Encrypt collection devices/media (laptops, usb keys, etc)
- For very/extremely sensitive data:
  - Collect with oversight directly; then
  - Store on encrypted device and;
  - Transfer to secure server as soon as feasible

[Micah Altman, 3/10/2011]

Randomized Response Technique



Variations:
- Ask two different questions
- Item counts with sensitive and non-sensitive items – eliminate subject-randomization
- Regression analysis methods

Our Table – Less (?) Sensitive

						Less (?) Sensitive	
Name	**SSN**	**Birthdate**	**Zipcode**	**Gender**	**Favorite Ice Cream**	**Treat?**	**# acts***
A. Jones	12341	01011961	02145	M	Raspberry	0	0
B. Jones	12342	02021961	02138	M	Pistachio	1	20
C. Jones	12343	11111972	94043	M	Chocolate	0	0
D. Jones	12344	12121972	94043	M	Hazelnut	1	12
E. Jones	12345	03251972	94041	F	Lemon	0	0
F. Jones	12346	03251972	02127	F	Lemon	1	7
G. Jones	12347	08081989	02138	F	Peach	0	1
H. Smith	12348	01011973	63200	F	Lime	1	17
I. Smith	12349	02021973	63300	M	Mango	0	4
J. Smith	12350	02021973	63400	M	Coconut	1	18
K. Smith	12351	03031974	64500	M	Frog	0	32
L. Smith	12352	04041974	64600	M	Vanilla	1	65
M. Smith	12353	04041974	64700	F	Pumpkin	0	128
N. Smith	12354	04041974	64800	F	Allergic	1	256

* Acts = crimes if treatment = 0; crimes + acts of generosity if treatment =1

[Micah Altman, 3/10/2011]

Randomized Response – Pros and Cons

- Pros
  - Can substantially reduce risks of disclosure
  - Can increase response rate
  - Can decrease mis-reporting
- Warning!
  - None of the randomized models uses a formal measure of disclosure limitation
  - Some would clearly violate measures (such as differential privacy) we'll see in section 4
  - *Do not use as a replacement for disclosure limitation*
- Other Issues
  - Loss of statistical efficiency (if compliance would otherwise be the same)
  - Complicates data analysis, especially model-based analysis
  - Leaving randomization up to subject can be unreliable
  - May provide less confidentiality protection if:
    - Randomization is incomplete
    - Records of randomization assignment are kept
    - Lists of responses overlap across questions
    - Sensitive question response is large enough to dominate overall response
    - Non-sensitive question responses are extremely predictable, or publicly observable

[Micah Altman, 3/10/2011]

Partitioning Information

- Reduces risk in information management
- Partition data information based on sensitivity
  - Identifying information
  - Descriptive information
  - Sensitive information
  - Other information
- Segregate
  - Storage of information
  - Access regimes
  - Data collections channels
  - Data transmission channels
- Plan to segregate as early as feasible in data collection and processing
- Link segregated information with artificial keys …

[Micah Altman, 3/10/2011]

Partitioned table

Name	SSN	Birthdate	Zipcode	Gender	LINK
A. Jones	12341	01011961	02145	M	1401
B. Jones	12342	02021961	02138	M	283
C. Jones	12343	11111972	94043	M	8979
D. Jones	12344	12121972	94043	M	7023
E. Jones	12345	03251972	94041	F	1498
F. Jones	12346	03251972	02127	F	1036
G. Jones	12347	08081989	02138	F	3864
H. Smith	12348	01011973	63200	F	2124
I. Smith	12349	02021973	63300	M	4339
J. Smith	12350	02021973	63400	M	6629
K. Smith	12351	03031974	64500	M	9091
L. Smith	12352	04041974	64600	M	9918
M. Smith	12353	04041974	64700	F	4749
N. Smith	12354	04041974	64800	F	8197

Not Identified

LINK	Favorite Ice Cream	Treat	# acts
1401	Raspberry	0	0
283	Pistachio	1	20
8979	Chocolate	0	0
7023	Hazelnut	1	12
1498	Lemon	0	0
1036	Lemon	1	7
3864	Peach	0	1
2124	Lime	1	17
4339	Mango	0	4
6629	Coconut	1	18
9091	Frog	0	32
9918	Vanilla	1	65
4749	Pumpkin	0	128
8197	Allergic	1	256

[Micah Altman, 3/10/2011]

Choosing Linking Keys

- Entirely randomized
  - Most resistant to relink
  - Mapping from original id to random keys is highly sensitive
  - Must keep and be able to access mapping to add new identified data
  - Most computer-generated random numbers are not sufficient by themselves
    - Most are PSEUDO random – predictable sequences
    - Use a *cryptographic secure PRNG: Blum Blum Shub, AES (or other block cypher) in counter mode* OR
    - Use real random numbers (e.g. from physical sources – see http://maltman.hmdc.harvard.edu/numal/) OR
    - Use a PRNG with a real random seed to random the order of the table; then another to generate the ID's for this randomly ordered table
- Encryption
  - More troublesome to compute
  - Same id's + same key + same "salt" produces same values → facilitates merging
  - ID's can be recovered if key is exposed, cracked, or algorithm weak
- Cryptographic Hash
  e.g. SHA-256
  - Security is well understood
  - Tools available to compute
  - Same id's produce same hashes → easier to merge new identified data
  - ID's cannot be recovered from hash because hash loses information
  - ID's can be *confirmed* if identifying information is known or guessable
- Cryptographic Hash + secret key
  - Security is well understood
  - Tools available to compute
  - Same id's produce same hashes → easier to merge new identified data
  - ID's cannot be recovered from hash because hash loses information
  - ID's cannot be confirmed unless key is also known
- *Do not choose arbitrary mathematical functions of other identifiers!*

[Micah Altman, 3/10/2011]

When is information anonymous?

- Remove all HIPAA ID's
  - Anonymous under HIPAA
  - Anonymous under Mass Law
  - Anonymous under HEISP
  - *Probably* anonymous under common rule, not guaranteed
  - *But* new/innovative forms of data collection may pose anonymity challenges
- **OR** anonymous observation only
  - No interaction with subject
  - No collection of identifying data
  - Not "human subjects" under common rule
- **OR** determined by a statistical expert to be anonymous

[Micah Altman, 3/10/2011]

Anonymous Data Collection: Pros & Cons

- Pros
  - Presumption that data is not identifiable
  - May increase participation
  - May increase honesty
- Cons
  - Barrier to follow-up, longitudinal studies
  - Can conflict with quality control, validation
  - Data still may be indirectly identifiable if respondent descriptive information is collected
  - Linking data to other sources of information may have large research benefits

[Micah Altman, 3/10/2011]

Anonymous Data Collection Methods

- Trusted third party intermediates
- Respondent initiates re-contacts
- No identifying information recorded
- Use id's randomized to subjects, destroy mapping

[Micah Altman, 3/10/2011]

Remote Data Collection Challenges

- Where network connection is readily available, easy to transfer as collected, or enter on remote system
  - Encrypted network file transfer (e.g. SFTP, part of ssh)
  - Encrypted/tunneled network file system (e.g. Expandrive)
- Where network connection is less reliable, high bandwidth
  - Whole-disk-encrypted laptop
  - Plus, Encrypted cloud backup solutions: CrashPlan, BackBlaze, SpiderOak
- Small data, short term
  - Encrypted USB keys (e.g. w/IronKey, TrueCrypt, PGP)
- Foreign Travel
  - Be aware of U.S. EAR export restrictions, use commercial or widely-available open encryption software only. Do not use bespoke software.
  - Be aware of country import restrictions (as of 2008): Burma, Belarus, China, Hungary, Iran, Israel, Morocco, Russia, Saudi Arabia, Tunisia, Ukraine
  - Encrypt data if possible, but don't break foreign laws. Check with department of state.

[Micah Altman, 3/10/2011]

Online/Electronic data collection challenges

- IP addresses are identifiers
  - IP addresses can be logged automatically by host, even if not intended by researcher
  - IP addresses can trivially be observed as data is collected
  - Partial ID numbers can be used for probabilistic geographical identification at sub-zipcode levels
- Cookies may be identifiers
  - Cookies provide a way to link data more easily
  - May or may not explicitly identify subject
- Jurisdiction
  - Data collected from subjects from other states / countries could subject you to laws in that jurisdiction
  - Jurisdiction may depend on residency of subject, availability of data collection instrument in jurisdiction, or explicit data collections efforts within jurisdiction
- Vendor
  - Vendor could retain IP addresses, identifying cookies, etc., even if not intended by researcher
- Recommendation
  - Use only vendors that certify compliance with your confidentiality policy
  - Do not retain IP numbers if data is being collected anonymously
  - Use SSL/TLS encryption unless data is non-sensitive and anonymous
- Some tools for anonymizing IP addresses and system/network logs

  www.caida.org/tools/taxonomy/anonymization.xml
- **Harvard policy**
  - **Recommendations as above**
  - **Plus: do not use or display Level 4+ for web surveys**

[Micah Altman, 3/10/2011]

Cloud computing risks

- Cloud computing decouples physical and computing infrastructure
- Increasingly used for core-IT, research computing, data collection, storage, and analysis
- Confidentiality issues
  - Auditing and compliance
  - Access and commingling of data
  - Location of data and services and legal jurisdiction

[Micah Altman, 3/10/2011]

Certificates of Confidentiality

- Issued by DHHS agencies such as NIH, CDC, FDA
- Protects against many types of forced legal disclosure of confidential information
- May not protect against all state disclosure law
- Does not protect against *voluntary* disclosures by researcher/research institutions

[Micah Altman, 3/10/2011]

Confidentiality & Consent

Best practice is to describe in consent form...

- Practices in place to protect confidentiality
- Plans for making the data available, to whom, and under what circumstances, rationale.
- Limitations on confidentiality (e.g. limits to a certificate of confidentiality under state law, planned voluntary disclosure)
- Consent form should be consistent with your:
  - Data management plan
  - Data sharing plans and requirements
- Not generally best practice to promise
  - Unlimited confidentiality
  - Destruction of all data
  - Restriction of all data to original researchers

[Micah Altman, 3/10/2011]

Data Management Plan

- **When is it required?**
  - *Any NIH request over $500K*
  - *All NSF proposals after 12/31/2010*
  - *NIJ*
  - *Wellcome Trust*
  - *Any proposal where collected data will be a resource beyond the project*
- **Safeguarding data during collection**
  - **Documentation**
  - **Backup and recovery**
  - **Review**
- **Treatment of confidential information**
  - **Overview: http://www.icpsr.org/DATAPASS/pdf/confidentiality.pdf**
  - **Separation of identifying and sensitive information**
  - **Obtain certificate of confidentiality, other legal safeguards**
  - **De-identification and public use files**
- **Dissemination**
  - **Archiving commitment (include letter of support)**
  - **Archiving timeline**
  - **Access procedures**
  - **Documentation**
  - **User vetting, tracking, and support**

One size does not fit all projects.

[Micah Altman, 3/10/2011]

Data Management Plan Outline

Law, policy, ethics
Research design …
Information security
Disclosure limitation

- Data description
  - nature of data {generated, observed, experimental information; amples; publications; physical collections; software; models}
  - scale of data
- Access and Sharing
  - Plans for depositing in an existing public database
  - Access procedures
  - Embargo periods
  - Access charges
  - Timeframe for access
  - Technical access methods
  - Restrictions on access
- Audience
  - Potential secondary users
  - Potential scope or scale of use
  - Reasons not to share or reuse
- Existing Data [If applicable]
  - description of existing data relevant to the project
  - plans for integration with data collection
  - added value of collection, need to collect/create new data
- Formats
  - Generation and dissemination formats and procedural justification
  - Storage format and archival justification
- Metadata and documentation
  - Metadata to be provided
  - Metadata standards used
  - Treatment of field notes, and collection records
  - Planned documentation and supporting materials
  - Quality assurance procedures for metadata and documentation
- Data Organization [if complex]
  - File organization
  - Naming conventions
- Quality Assurance [if not described in main proposal]
  - Procedures for ensuring data quality in collections, and expected measurement error
  - Cleaning and editing procedures
  - Validation methods
- Storage, backup, replication, and versioning
  - Facilities
  - Methods
  - Procedures
  - Frequency
  - Replication
  - Version management
  - Recovery guarantees
- Security
  - Procedural controls
  - Technical Controls
  - Confidentiality concerns
  - Access control rules
  - Restrictions on use
- Responsibility
  - Individual or project team role responsible for data management
- Budget
  - Cost of preparing data and documentation
  - Cost of permanent archiving
- Intellectual Property Rights
  - Entities who hold property rights
  - Types of IP rights in data
  - Protections provided
  - Dispute resolution process
- Legal Requirements
  - Provider requirements and plans to meet them
  - Institutional requirements and plans to meet them
- Archiving and Preservation
  - Requirements for data destruction, if applicable
  - Procedures for long term preservation
  - Institution responsible for long-term costs of data preservation
  - Succession plans for data should archiving entity go out of existence
- Ethics and privacy
  - Informed consent
  - Protection of privacy
  - Other ethical issues
- Adherence
  - When will adherence to data management plan be checked or demonstrated
  - Who is responsible for managing data in the project
  - Who is responsible for checking adherence to data management plan

[Micah Altman, 3/10/2011]

IQSS
Data Management Services

- **The Henry A. Murray Research Archive**
  - **Harvard's endowed permanent data archive**
  - **Assists in developing data management plans**
  - **Can provide cataloging assistance for public release of data**
  - **Dissemination of data through IQSS Dataverse Network**
- **The IQSS Dataverse Network**
  - **Standard data management plan for public, small data**
  - **Provides easy virtual archiving and dissemination**
  - **Data is catalogued and controlled by you**
  - **You theme and brand your virtual archive**
  - **Universally searchable, citable**
  - **Automatically provides data formatting and statistical analysis on-line**

http://dvn.iq.harvard.edu

Data Management Plans Examples (Summaries)

- **Example 1**
- The proposed research will involve a small sample (less than 20 subjects) recruited from clinical facilities in the New York City area with Williams syndrome. This rare craniofacial disorder is associated with distinguishing facial features, as well as mental retardation. Even with the removal of all identifiers, we believe that it would be difficult if not impossible to protect the identities of subjects given the physical characteristics of subjects, the type of clinical data (including imaging) that we will be collecting, and the relatively restricted area from which we are recruiting subjects. Therefore, we are not planning to share the data.
- **Example 2**
- The proposed research will include data from approximately 500 subjects being screened for three bacterial sexually transmitted diseases (STDs) at an inner city STD clinic. The final dataset will include self-reported demographic and behavioral data from interviews with the subjects and laboratory data from urine specimens provided. Because the STDs being studied are reportable diseases, we will be collecting identifying information. Even though the final dataset will be stripped of identifiers prior to release for sharing, we believe that there remains the possibility of deductive disclosure of subjects with unusual characteristics. Thus, we will make the data and associated documentation available to users only under a data-sharing agreement that provides for: (1) a commitment to using the data only for research purposes and not to identify any individual participant; (2) a commitment to securing the data using appropriate computer technology; and (3) a commitment to destroying or returning the data after analyses are completed.
- **Example 3**
- This application requests support to collect public-use data from a survey of more than 22,000 Americans over the age of 50 every 2 years. Data products from this study will be made available without cost to researchers and analysts. https://ssl.isr.umich.edu/hrs/
- User registration is required in order to access or download files. As part of the registration process, users must agree to the conditions of use governing access to the public release data, including restrictions against attempting to identify study participants, destruction of the data after analyses are completed, reporting responsibilities, restrictions on redistribution of the data to third parties, and proper acknowledgement of the data resource. Registered users will receive user support, as well as information related to errors in the data, future releases, workshops, and publication lists. The information provided to users will *not* be used for commercial purposes, and will *not* be redistributed to third parties.

- FROM NIH, [grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#ex]

External Data Usage Agreements

- Between provider and individual
  - Careful – you're liable
  - Harvard will not help if you sign
  - University review and OSP signature **strongly recommended**
- Between provider and University
  - University Liable
  - *Requires University Approved Signer : OSP*
- Avoid nonstandard protections whenever possible
  - DUA's can impose very specific and detailed requirements
  - Compatible in spirit does not apply compatibility in legal practice
  - *Use University policies/procedures as a template*

Controls on Confidential Information

Harvard University has developed extensive technical and administrative procedures to be used with all identified personal information and other confidential information. The University classifies this form of information internally as "Harvard Confidential Information" -- or HCI.

Any use of HCI at Harvard includes the following safeguards:

- Systems security. Any system used to store HCI is subject to a checklist of technical and procedural security measures including: operating system and applications must be patched to current security levels, a host based firewall is enabled, anti virus software is enabled and the definitions file is current
- Server security. Any server used to distribute HCI to other systems (e.g. through providing remote file system), or otherwise offering login access, must employ additional security measures including: connection through a private network only; limitation on length of idle sessions; limitations on incorrect password attempts; and additional logging and monitoring.

- Access restriction: an individual is allowed to access HCI only if there is a specific need for access. All access to HCI is over physically controlled and/or encrypted channels.
- Disposal processes: including secure file erasure and document destruction.
- Encryption: HCI must be strongly encrypted whenever it is transmitted across a public network, stored on a laptop, or stored on a portable device such as a flash drive or on portable media.

This is only a brief summary. The full University security policy can be found here:
http://security.harvard.edu/heisp

And a more detailed checklist used to verify systems compliance is found here:
http://security.harvard.edu/files/resources/forms/

These safeguards are applied consistently throughout the university, we believe that these requirements offer stringent protection for the requested data. And these requirements will be applied in addition to any others required by specific data use agreement.

[Micah Altman, 3/10/2011]

IQSS Data Management Services

- **The Henry A. Murray Research Archive**
  - Harvard's endowed permanent data archive
  - Assists in developing data management plans
  - Can provide cataloging assistance for public release of data
  - Dissemination of data through IQSS Dataverse Network
  - Provides letters of commitment to permanent archiving

  www.murray.harvard.edu

- **The IQSS Dataverse Network**
  - Provides easy virtual archiving and dissemination
  - Data is catalogued and controlled by you
  - You theme and brand your virtual archive
  - Universally searchable, citable
  - Automatically provides data formatting and statistical analysis on-line

  dvn.iq.harvard.edu

[Micah Altman, 3/10/2011]

Key Concepts & Issues Review

- Levels of sensitivity
- Anonymity criteria
- Sensitivity reduction
- Certificate of Confidentiality
- Data sharing plan
- Data management plan
- Information Partitioning
- Linking Keys

Checklist: Research Design …

- ✓ Does research involve human subjects?
- ✓ What are possible harms that could occur if identified information was disclosed?
- ✓ Is information collected benign, sensitive, very sensitive, or extremely sensitive? (IRB makes final determination)
- ✓ Can the sensitivity of the information be reduced?
- ✓ Can research be carried out with anonymity?
- ✓ Can research data be de-identified during collection?
- ✓ How can identifying information, descriptive information and sensitive information be segregated?
- ✓ Have you:
  - ✓ Completed NIH human subjects training
  - ✓ Harvard HETHR training
- ✓ Have you written the following to be consistent with final plans for analysis and dissemination:
  - ✓ data management plan?
  - ✓ consent documents?
  - ✓ application for certificate of confidentiality?

[Micah Altman, 3/10/2011]

Resources

- E.A. Bankert & R.J. Andur, 2006, *Institutional Review Board: Management and Function,* Jones and Bartlett Publishers
- R. Groves, *et* al., 2004, *Survey Methodology,* John Wiley & sons.
- J.A. Fox, P.E. Tracy, 1986, *Randomized Response,* Sage Publications.
- R.M. Lee, 1993, *Doing Research on Sensitive Topics,* Sage Publications.
- D. Corstange, 2009, "Sensitive Questions, Truthful Answers? Modeling the List Experiment with LISTIT", *Political Analysis* 17:45–63
- ICPSR Data Enclave [www.icpsr.umich.edu/icpsrweb/ICPSR/access/restricted/enclave]
- Murray Research Archive [www.murray.harvard.edu]
- IQSS Dataverse Network [dvn.iq.harvard.edu/]
- NIH Certificate of Confidentiality Kiosk [grants.nih.gov/grants/policy/coc]

Information Security

- Security principles
- FISMA
- Categories of technical controls
- A simplified approach
- **Harvard Policies**
- [Summary]

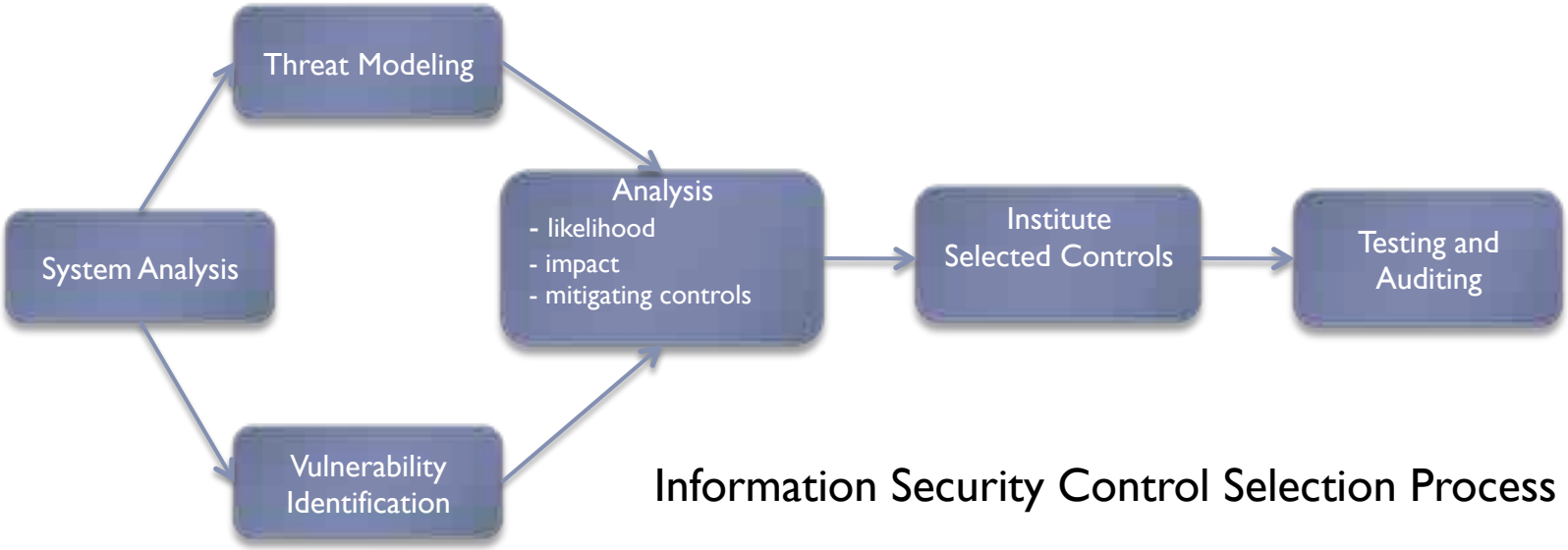
Core Information Security Concepts

- Security properties
  - Confidentiality
  - Integrity
  - Availability
  - [Authenticity]
  - [Nonrepudiation]
- Security practices
  - Defense in depth
  - Threat modeling
  - Risk assessment
  - Vulnerability assessment

[Micah Altman, 3/10/2011]

Risk Assessment

- [NIST 800-100, simplification of NIST 800-30]



Information Security Control Selection Process

[Micah Altman, 3/10/2011]

Risk Management Details

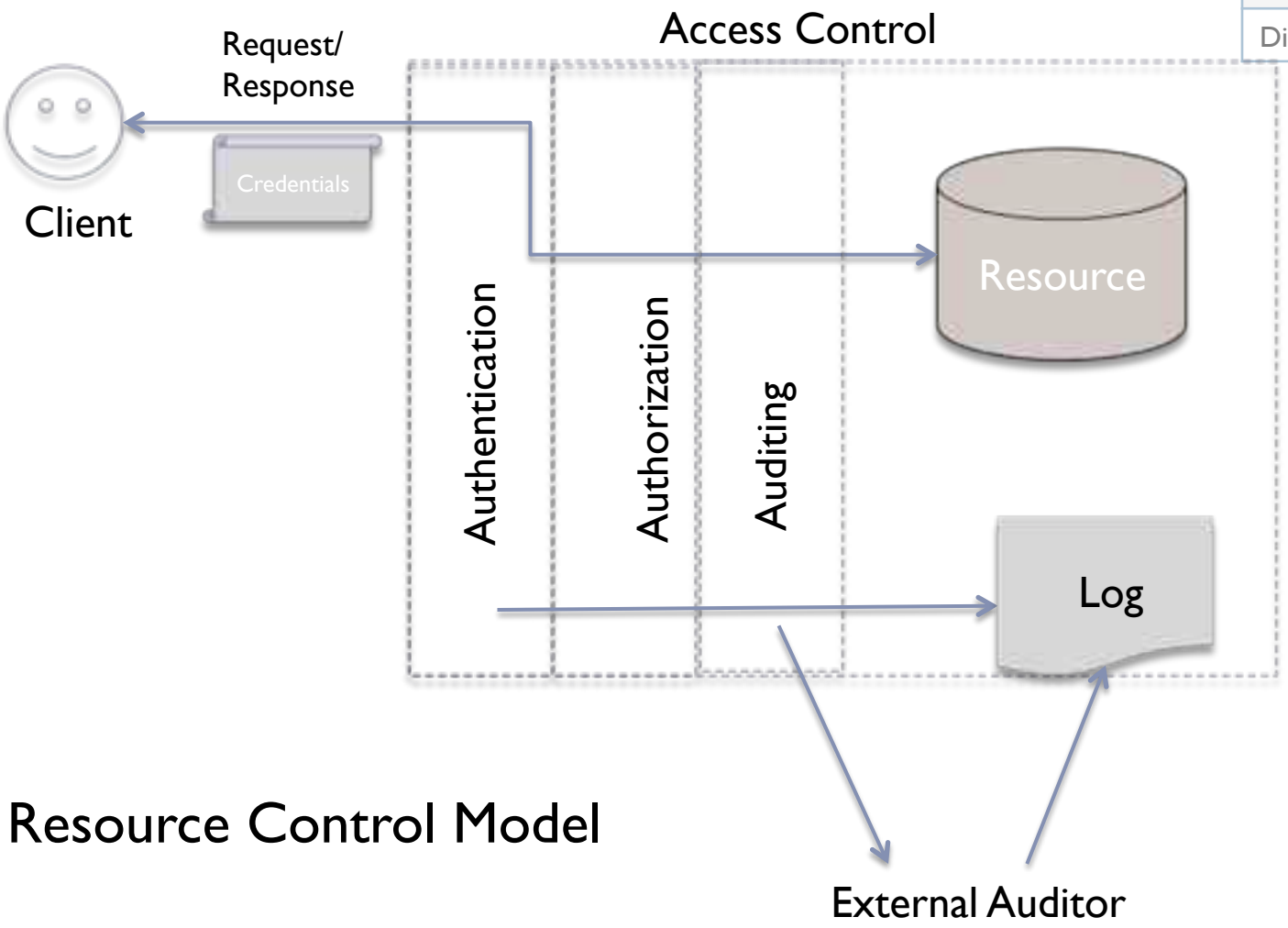
- System Characterization
- Threat Identification
- Control analysis
- Likelihood determination
- Impact Analysis
- Risk Determination
- Control recommendation
- Results documentation

[Micah Altman, 3/10/2011]

Classes of threats and vulnerabilities

- 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 Control Model



Resource Control Model

Operational and Technical Controls [NIST 800-53]

- 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

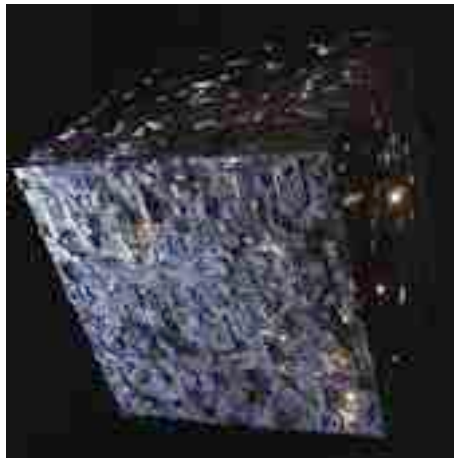
Key Information Security Standards

- **Comprehensive Information Security Standards**
  - **FISMA – framework for non-classified information security in federal government.**
  - ISO/IEC 27002 – framework of similar scope to FISMA, used internationally
  - PCI – Payment card industry security standards. Used by major payment card companies, processors, etc.
- Related Certifications
  - FIPS-compliance and certification
    - Establishes standards for cryptographic methods and modules
    - Be aware that FIPS-certification often limited to algorithm used, and *not* entire system
  - SAS 70 Audits – Type 2
    - Independent audit of controls and control objectives
    - *Does not establish sufficiency of control objectives*
  - CISSP -- Certified Information Systems Security Professional
    - Widely recognized certification for information security professionals

FISMA Overview

Federal Information Security Management Act of 2002

- All federal agencies required to develop agency-wide information security plan
- NIST published extensive list of recommendations
- Federal sponsors seem to be trending to FISMA as best practice for managing confidential data produced by award
- Identifies risk and impact level; monitoring; technical and procedural controls
- **Harvard HRCI controls: less than FISMA "low"**



[Micah Altman, 3/10/2011]

Security Control Baselines

Low (impact)	***Medium-High (impact), adds…***
Access Control	
Policies; Account management *; Access Enforcement; Unsuccessful Login Attempts; System Use Notification; Restrict Anonymous Access*; Restrict Remote Access*; Restrict Wireless Access*; Restrict Mobile Devices*; Restrict use of External Information Systems*; Restrict Publicly Accessible Content	Information flow enforcement; Separation of Duties; Least Privilege; Session Lock
Security Awareness and Training	
Policies; Awareness; Training; Training Records	
Audit and Accountability	
Policies; Auditable Events *; Content of Audit Records *; Storage Capacity; Audit Review, Analysis and Reporting *; Time Stamps *; Protection of Audit Information; Audit Record Retention; Audit Generation	Audit Reduction; Non-Repudiation
Security Assessment and Authorization	
Policies; Assessments* ; System Connections; Planning; Authorization; Continuous Monitoring	

[Micah Altman, 3/10/2011]

Security Control Baselines

Configuration Management	
Low (impact)	***Medium-High (impact), adds…***
Policies; Baseline*; Impact Analysis; Settings*; Least Functionality; Component Inventory*	Change Control; Access Restrictions for Change; Configuration Management Plan
Contingency Planning	
Policies; Plan * ; Training *; Plan Testing*; System backup*; Recovery & Reconstitution *	Alternate storage site; Alternate processing site; Telecomm
Identification and Authentication	
Policies; Organizational Users*; Identifier Management; Authenticator Management *; Authenticator Feedback; Cryptographic Module Authentication; Non-Organizational Users	Device identification and authentication
Incident Response	
Policies; Training; Handling *; Monitoring; Reporting*; Response Assistance; Response Plan	Testing
Maintenance	
Policies; Control*; Non-Local Maintenance Restrictions*; Personnel Restrictions*	Tools; Maintenance scheduling/timeliness

[Micah Altman, 3/10/2011]

Security Control Baselines

Low (impact)	***Medium-High (impact), adds…***
Media Protection	
Policies; Access restrictions*; Sanitization	Marking; Storage; Transport
Physical and Environmental Protection	
Policies; Access Authorizations; Access Control*; Monitoring*; Visitor Control *; Records*; Emergency Lighting; Fire protection*; Temperature, Humidity, water damage*; Delivery and removal	Network access control; Output device Access control; Power equipment access, shutoff, backup; Alternate work site; Location of information system components; information leakage
Planning	
Policies, Plan, Rules of Behavior; Privacy Impact Assessment	Activity planning
Personnel Security	
Policies; Position categorization; Screening; Termination; Transfer; Access Agreements; Third-Parties; Sanctions	
Risk Assessment	
Policies; Categorization Assessment; Vulnerability Scanning*	

[Micah Altman, 3/10/2011]

Security Control Baselines

System and Services Acquisition	
Low (impact)	***Medium-High (impact), adds…***
Policies; Resource Allocation; Life Cycle Support; Acquisition*; Documentation; Software usage restrictions; User installed software restrictions; External information System Services restrictions	Security Engineering; Developer configuration management; Developer security testing; supply chain protection; Trustworthiness
System and Communications Protection	
Policies; Denial of Service Protection; Boundary protection*; Cryptographic key Management; Encryption; Public Access Protection; Collaborative computing devices restriction; Secure Name resolution*	Application Partitioning; Restrictions on Shared Resources; Transmission integrity & confidentiality; Network Disconnection Procedure; Public Key Infrastructure Certificates; Mobile Code management; VOIP management; Session authenticity; Fail in known state; Protection of information at rest; Information system partitioning
System and Information Integrity	
Policies, Flaw remediation*; Malicious code protection*; Security Advisory monitoring*; Information output handling	Information system monitoring; Software and information integrity; Spam protection; Information input restrictions & validation; Error handling
Program Management	
Plan; Security Officer Role; Resources; Inventory; Performance Measures; Enterprise architecture; Risk management strategy; Authorization process; Mission definition	

HIPAA Requirements

- Administrative controls
  - Access authorization, establishment, modification, and termination.
  - Training program
  - Vendor compliance
  - Disaster recovery
  - Internal audits
  - Breach procedures
- Physical controls
  - Disposal
  - Access to equipment
  - Access to physical environment
  - Workstation environment
- Technical controls
  - Intrusion protection
  - Network encryption
  - Integrity checking
  - Authentication of communication
  - Configuration management
  - Risk analysis

[Micah Altman, 3/10/2011]

Delegating Systems Security

- What are goals for confidentiality, integrity, availability?
- What threats are envisioned?
- What controls are in place?
- Is there 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. SAS Type -2 Audits; FISMA Compliance; ISO Certification
- What security standards are referenced?
  - E.g. FISMA, ISO, HEISP/HDRSP/PCI

What most security plans **do not** do

- Protect against all insider threats
- Protect against all unintentional threats (human error, voluntary disclosure)
- Protect against the CIA, TEMPEST, evil maids, and other well-resourced, sophisticated adversaries
- Protect against prolonged physical threats to computer equipment, or data owner

[Micah Altman, 3/10/2011]

Information Security is Systemic

Not just control implementation but…

- Policy creation, maintenance, auditing
- Implementation review, auditing, logging, monitoring
- Regular vulnerability & threat assessment

[Micah Altman, 3/10/2011]

Simplified Approach for Sensitive Data

- Use whole-disk/media encryption to protect data at rest
- Use end-to-end encryption to protect data in motion
- Use core information hygiene to protect systems
- Scan for HRCI regularly
- Be thorough in disposal of information

Very sensitive/extremely sensitive data requires more protection.

[Micah Altman, 3/10/2011]

Plan Outline – Very Sensitive Data

- Protect very sensitive data on "target systems"
  - Extra physical, logical, administrative access control
    - Record keeping
    - Limitations
    - Lockouts
  - Extra monitoring, auditing
  - Extra procedural controls – specific, renewed approvals
  - Limits on network connectivity
    - Private network, not *directly connected to public network*
- Regular scans
  - Vulnerability scans
  - Scans for PII
- Extremely sensitive
  - Increased access control, procedural limitations
  - Not physically/logically connected (even via wireless) to public network, directly or indirectly

[Micah Altman, 3/10/2011]

Harvard: Identifying Confidential Information

- Has it been been designated by school CIO as confidential?
  - → Harvard Confidential Information
- Does it contain social security numbers, financial id's, credit card #'s, driver's license #'s, other federal id's?
  - → Treat as HRCI
- Is information completely anonymous?
  - → Not confidential
- Does it contain private information from research on people?
  - → Confidential
    IRB determines whether: Benign, HCI, HRCI, or **Extremely sensitive**.
- Does it contain private information about named/identified students that does not appear in the directory?
  - → Harvard Confidential Information

[but … if information is minimal risk & minimal size, custom & practice allows greater flexibility in transmission and storage]

- Would its release put individuals or organizations at significant risk of criminal or civil liability, or be damaging to financial standing, employability, or reputation?
  - → Treat as Harvard Confidential Information

[Micah Altman, 3/10/2011]

Harvard: Safe storage

- Extremely sensitive HRCI/Level 4 must be stored only in designated offline systems
- HRCI/Level 4 may be stored only on professionally managed, approved servers, on private network segments
  [no storage on user systems with limited exceptions for approved field data collection]
- *Cautions*
  - *Individual approval required*
  - *Additional rules for storing Credit Card numbers*
  - *Level 5 data and some level 4 systems may be required to be entirely disconnected from external networks*
- Confidential information may be stored on:
  - Managed File Services *OR*
  - Other approved server *OR*
  - On desktop/laptop/USB if…
    - Encrypted
    - Good information hygiene

Harvard: Safe use information

- Who can use Confidential Information?
  - Education records
    - Can be shared with other Harvard staff & faculty who have a specific need
  - Harvard business related CI & HRCI
    - Sharing must be approved by FAS Director of Security
  - Research CI & HRCI
    - Sharing must be approved by PI of research project.
    - PI must have data sharing policy approved by IRB.
  - **It is your responsibility to notify the recipient of their responsibility to protect confidentiality.**
- What to do?
  - Practice good information hygiene
  - Always Encrypt
  - Scan for HRCI
  - Use Strong Passwords

[Micah Altman, 3/10/2011]

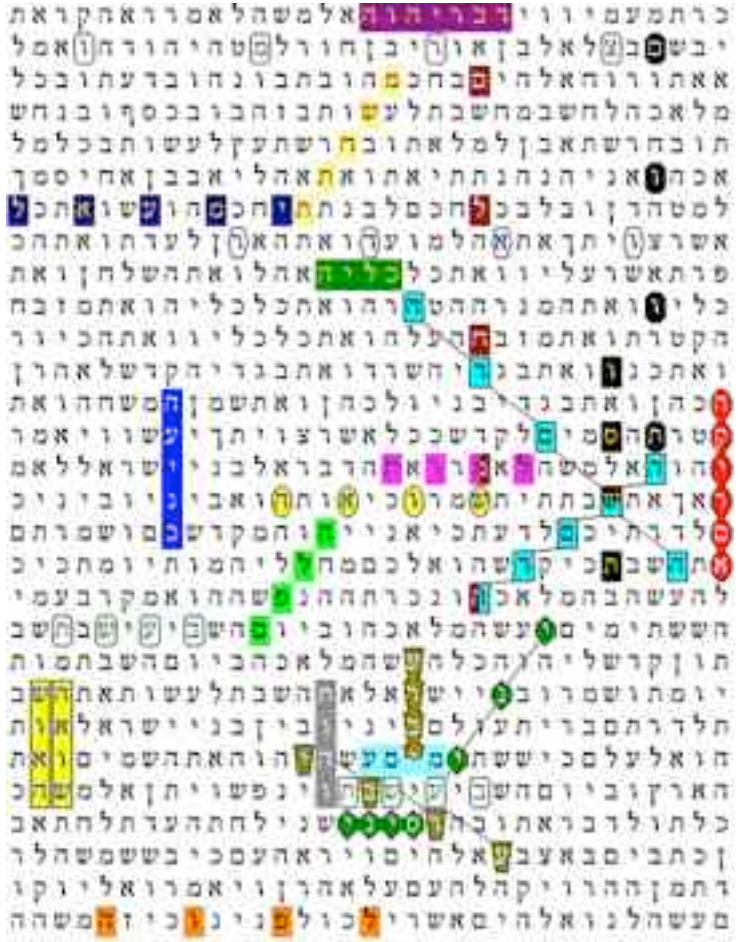
Harvard: good information hygiene (level 2)

- Computer setup
  - Use a virus checker
    - And keep it updated
  - Use a host-based firewall
  - Use a locking screen-saver
  - Lock default/open accounts
  - Regularly scan for sensitive information
  - Update your software regularly
    - Install all operating system and application security updates
- Server Setup
  - Password guessing restrictions
  - Idle session locking (or used on all client)
  - No password retrieval
  - Keep access logs
- Behavior
  - Don't share accounts or passwords
  - Don't use administrative accounts all the time
  - Don't run programs from untrusted sources
  - Don't give out your password to anyone
  - Have a process for revoking user access when no longer needed/ authorized (e.g. if user leaves university)
  - Documented breach reporting procedure
  - Users should have appropriate confidentiality training

[Micah Altman, 3/10/2011]

Harvard: use encryption

- Encrypt HCI *files* if they are not stored on an approved server
- Encrypt *whole disks* for all laptops and portable storage"
- Encrypt network transmissions:
  - Presenting your password
  - Transmitting HCI files
  - Remotely accessing confidential information
  - Do not send unencrypted files through mail – even if they are "passworded"
- Preferred tools
  - *Remote access generally:*
    - use the VPN
  - File transfer
    - Accellion: https://fta.fas.harvard.edu/
    - Winzip (must enable *encryption*)
    - Secure FTP (Secure FX, Expandrive, etc.)
    - PGP Encrypted mail



Harvard: Scan for HRCI

- University requires annual scanning
- FAS requires semi-annual scanning
- HMDC will scan information stored on servers
- FAS will provide scanning tools for laptops & desktops
  - http://www.fas-it.fas.harvard.edu/services/catalog/browse/39/761

[Micah Altman, 3/10/2011]

Harvard: Safe disposal

- Shred Paper, tapes, cd's:
  - Use a cross-cut shredder
  - Place in Harvard-designated, **locked**, shredder bin
- Erase confidential files when they are no longer needed
  - Information can hide in files, even if not visible
  - Files can remain on disk, even if not visible
  - Use an approved secure file eraser such as **PGP SHRED**
- Ask IT support when transferring, donating, or disposing of computers or portable storage
  - Information can hide in obscure places on an a computer.
    - Files on SSD may not be erased until later…
    - Information on standard disks may be preserved in "bad blocks", etc.
  - IT Support has special equipment to securely and complete erase disks.
  - Computers can then be safely reused.

[Micah Altman, 3/10/2011]

Use Strong Credentials

- Strong Passwords, **at minimum**
  - At least 8 characters long for login/account passwords
  - **Recommended**: at least 12 characters for cryptographic keys
  - Combination of upper/lowercase, and at least one non-letter
  - Not be common words, names, or anything easily guessed
  - No dates, no long sequences (>4) of numbers
  - Not two words separated by another character
  - Not shared across accounts or people
  - **Never exposed without encryption**
- Good passwords are also
  - Cryptographically strong
  - Hard for a person to guess
  - Easy to remember
  - One method for choosing good passwords…
    - use phrases with punctuation
    - abbreviate
    - substitute characters
- Servers and Applications
  - Store passwords in a manner that can't be retrieved
    - Store hashes
    - Force password change of automatically generated passwords
  - Protect against password guessing
    - Brute-force guessing of a file of 8 character alphanumeric passwords takes less than $3 dollars of compute time!
      (12 character complex passwords > $1,000,000) [Campbell 2009]
    - Protect encrypted password files
    - Monitor failed login attempts

See security.harvard.edu/resources/best-practices/passwords

[Micah Altman, 3/10/2011]

Key Concepts Review

- Confidentiality
- Integrity
- Availability
- Threat modeling
- Vulnerability assessment
- Risk assessment
- Defense in depth
- Logical Controls
- Physical Controls
- Administrative Controls

Checklist: *Identify Requirements*

- ✓ Documented information security plan?
  - ✓ What are goals for confidentiality, integrity, availability?
  - ✓ What threats are envisioned?
  - ✓ What are the broad types of controls in place?
- ✓ Key protections
  - ✓ Use whole-disk/media encryption to protect data at rest
  - ✓ Use end-to-end encryption to protect data in motion
  - ✓ Use basic information hygiene to protect systems
  - ✓ Be thorough in disposal of information
- ✓ Additional protections for sensitive data
  - ✓ Extra logical, administrative, physical controls for very sensitive data?
  - ✓ Monitoring and vulnerability scanning for very sensitive data?
  - ✓ Check requirements for remote and foreign data collection
- ✓ Refer to security standards
  - ✓ FIPS encryption
  - ✓ FISMA / ISO practices
  - ✓ SAS-70 Auditing
  - ✓ CISSP certification of key staff
- ✓ Delegate implementation to information security professionals

[Micah Altman, 3/10/2011]

Resources

- S. Garfinkel, et *al.* 2003, *Practical Unix and Internet Security, 3rd ed.* , O'Reilly Media
- Shon Harris, 2001, *CISSP All-in-One Exam Guide*, Osborne
- NIST, 2009, DRAFT Guide to Protecting the Confidentiality of Personally Identifiable Information, Nist Publication 800-122.
- NIST, 2009, Recommended Security Controls for Federal Information Systems and Organizations v. 3, NIST 800-53. (Also see related NIST 800-53A, and other NIST Computer Security Division Special Publications) [csrc.nist.gov/publications/PubsSPs.html]
- NIST, 2006, Information Security Handbook: A Guide for Managers, NIST Publication 800-100.
- **Harvard Enterprise Security Checklists** [security.harvard.edu/resources/forms]

Recommended Software

- Whole Disk Encryption
  - Open Source: truecrypt.org
  - Commercial: pgp.com
- Scanning
  - Vulnerability scanner/assessment tool: www.nessus.org/nessus
  - Commercial version scans for (limited) PII: www.nessus.org/nessus
  - PII Scanning tool (open source), Cornell Spider: www2.cit.cornell.edu/security/tools
  - PII Scanning tool (commercial), Identity Finder: www.identityfinder.com
  - File integrity/intrusion detection engine, Samhain: la-samhna.de/samhain
  - Network intrusion detection, Snort: www.snort.org
- Encrypt transmission over network
  - Open SSL: http://openssl.org
  - Open SSH: http://openssh.org
  - VTUN: http://vtun.sourceforge.net
- Cloud backup services with encryption
  - Crashplan: http://crashplan.com
  - Spider oak: http://spideroak.com
  - Backblaze: http://backblaze.com

Disclosure Limitation

- Threat models
- Disclosure limitation methods
- Statistical disclosure limitation methods
- Types of disclosure
- Factors affecting disclosure protection
- SDL Caveats
- SDL Observations

Threat Models

- Nosy neighbor (nosy employer)
- Muck-raking Journalist (zero-tolerance)
- Business rival contributing to same survey
- Absent-minded professor
- …

[Micah Altman, 3/10/2011]

Non statistical Disclosure Limitation Methods

- Licensing
  - Used in conjunction with limited deidentification
  - Should prohibit reidentification & linking, dissemination to third parties; limit retention
  - Advantages: can decrease cost of processing, increase utility of research data
  - Disadvantages: licenses may be violated unintentionally or intentionally, difficult to enforce outside of limited domains (e.g. HIPAA)
- Automated de-identification
  - Primarily used for qualitative text medical records. Replaces identifiers with dummy strings.
  - Advantages: can decrease cost, increase accuracy of manual deidentification of qualitative information
  - Disadvantage: little available software, error rates still slightly higher than teams of trained human coders

[Micah Altman, 3/10/2011]

Automated De-identification

- Trained human sensitivity rates:
  - Single worker: [.63-.94] (.81)
  - Two-person team: [.89-.98] (.94)
  - Three-person team: [.98-.99] (.98)
    [Neamatullah 2008]
- State of the art algorithms approach recall of .95
  [Uzuner, et. al 2007]
  - Statistical learning of rule template features worked best
  - Simpler rules-based approach still did as well as median 2-person team
  - Rules for PII and local dictionary important

Text de-identification (HIPAA)

Law, policy, ethics
Research design …
Information security
Disclosure limitation

Cleaned

Name	SSN	Birthdate	Zipcode	Gender	Favorite Ice Cream	# of crimes committed
[Name 1]	*	*1961	021*	M	Raspberry	0
[Name 2]	*	*1961	021*	M	Pistachio	0
[Name 3]	*	*1972	940*	M	Chocolate	0
[Name 4]	*	*1972	940*	M	Hazelnut	0
[Name 5]	*	*1972	940*	F	Lemon	0
[Name 6]	*	*1972	021*	F	Lemon	1
[Name 7]	*	*1989	021*	F	Peach	1
[Name 8]	*	*1973	632*	F	Lime	2
[Name 9]	*	*1973	633*	M	Mango	4
[Name 10]	*	*1973	634*	M	Coconut	16
[Name 11]	*	*1974	645*	M	Frog	32
[Name 12]	*	*1974	646*	M	Vanilla	64
[Name 13]	*	*1974	647*	F	Pumpkin	128
[Name 14]	*	*1974	648*	F	[NONE]	256

Cleaned (by hand check)

[Micah Altman, 3/10/2011]

New Data – New Challenges

Law, policy, ethics
Research design …
Information security
Disclosure limitation

- How to deidentify without completely destroying the data?
  - The "Netflix Problem": large, sparse datasets that overlap can be probabilistically linked [Narayan and Shmatikov 2008]
  - The "GIS": fine geo-spatial-temporal data impossible mask, when correlated with external data [Zimmerman 2008]
  - The "Facebook Problem": Possible to identify masked network data, if only a few nodes controlled. [Backstrom, et. al 2007]
  - The "Blog problem" : Pseudononymous communication can be linked through textual analysis [Tomkins et. al 2004]

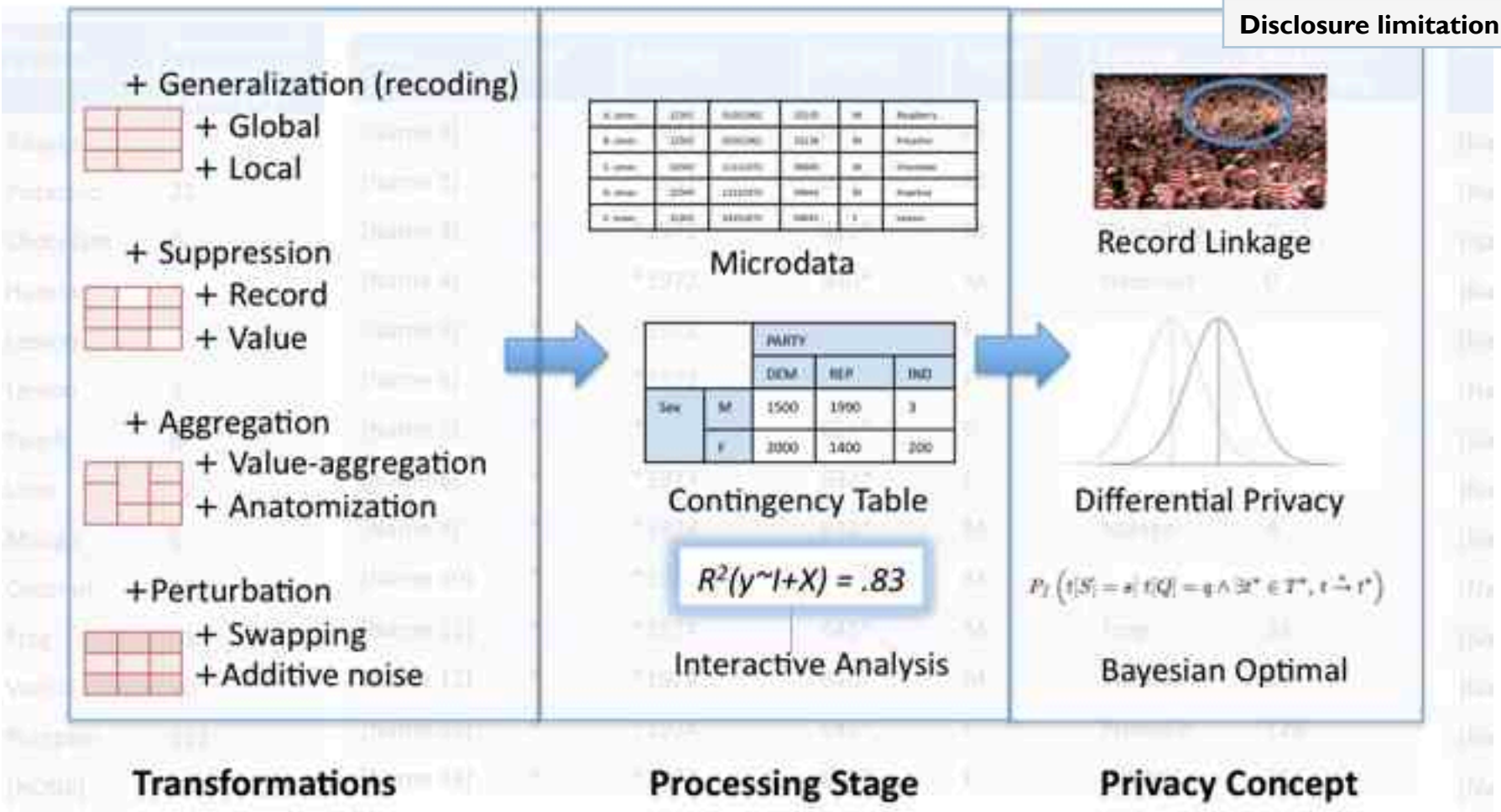
Source: [Calberese 2008; Real Time Rome Project 2007]

Hybrid Statistical/Non-statistical Limitation

- Data enclaves – physically restrict access to data
  - Examples: ICPSR, Census Research Data Center
  - May include availability of synthetic data as an aid to preparing model specifications
  - Advantages: extensive human auditing, vetting; information security threats much reduced
  - Disadvantages: expensive, slow, inconvenient to access
- Controlled remote access
  - Varies from remote access to all data and output to human vetting of output
  - Advantages: auditable, potential to impose human review, potential to limit analysis
  - Disadvantages: complex to implement, slow
- Model servers
  - Mediated remote access – analysis limited to designated models
  - Advantages: faster, no human in loop
  - Disadvantage: statistical methods for ensuring model safety are immature – residuals, categorical variables, dummy variables are all risky; very limited set of models currently supported; complex to implement
- Statistical Disclosure Limitation
  - Modifications to the data to decrease the probability of disclosure
  - Advantages/Disadvantages… to follow…

[Micah Altman, 3/10/2011]

Disclosure Limitation Approaches



[Micah Altman, 3/10/2011]

Pure Statistical Disclosure limitation techniques

- Data reduction
  - Removing variables (i.e. deidentifying)
  - Suppressing records
  - Sub-sampling
  - Global recoding (including top/bottom coding)
  - Local suppression
  - Global complete suppression ☺
- Perturbation
  - Microaggregation
    - Sorting based on similarity
    - Replace value of records in clusters with mean
  - Rule-based data swapping
  - Adding noise
  - Resampling
- Synthetic microdata
  - Bootstrap
  - Multiple imputation
  - Model based

[Micah Altman, 3/10/2011]

Help, help, I'm being suppressed…

- Law, policy, ethics
- Research design …
- Information security
- **Disclosure limitation**

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~~

Traditional Static Suppression

- **Data reduction**
  - Removing variables (i.e. deidentifying)
  - Suppressing records
  - Sub-sampling
  - Global recoding (including top/bottom coding)
  - Local suppression
  - Global complete suppression ☺
- **Perturbation**
  - Microaggregation
    - Sorting based on similarity
    - Replace value of records in clusters with mean
  - Rule-based data swapping
  - Adding noise
  - Resampling
- **Synthetic microdata**
  - Bootstrap
  - Multiple imputation
  - Model based

Row

[Micah Altman, 3/10/2011]

Suppression with R and sdcmicro


```
# setup
> library(sdcMicro)

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

[Micah Altman, 3/10/2011]

Suppression with R and sdcmicro


```
# 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
 -------------------------
```


[Micah Altman, 3/10/2011]

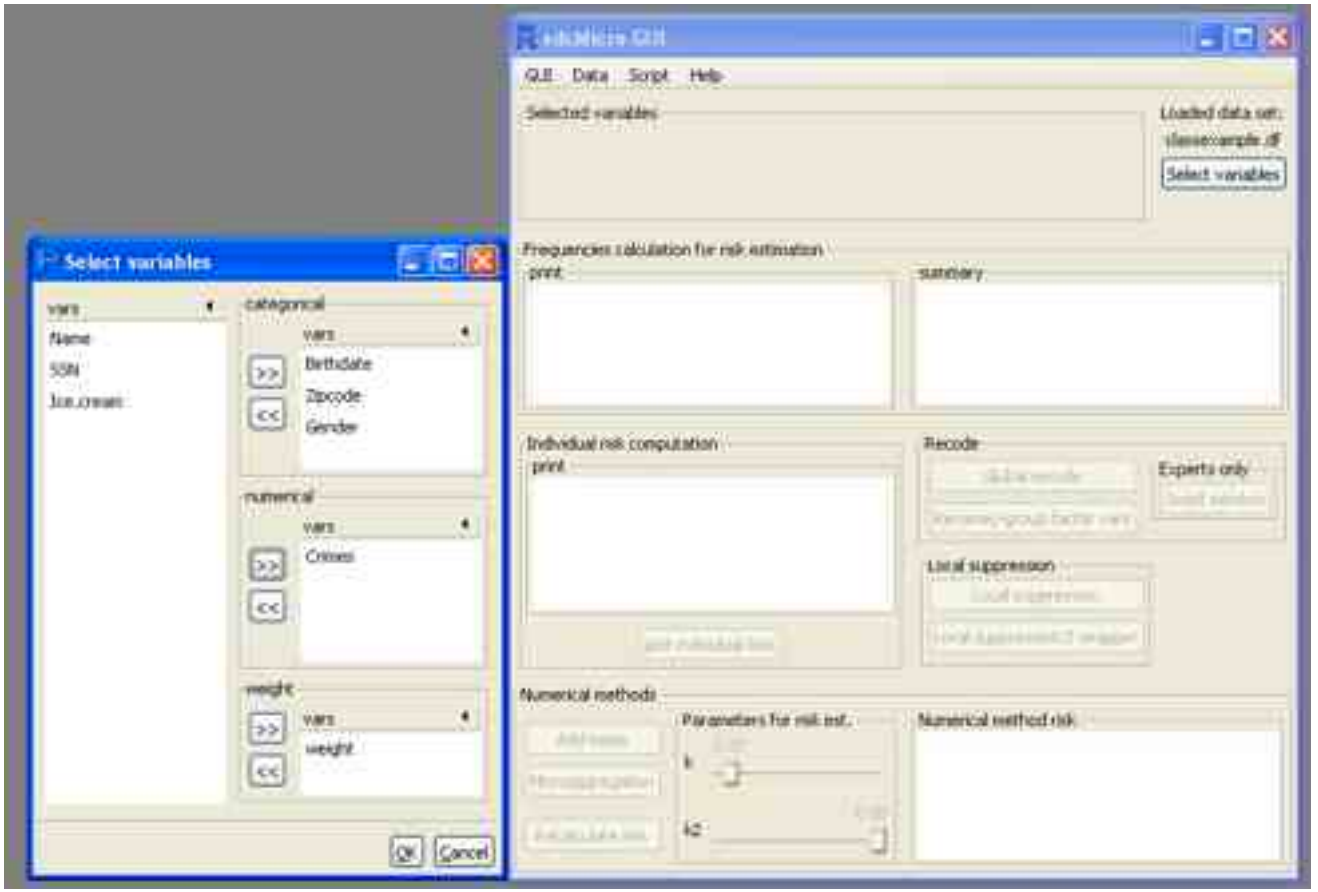
Suppression with R and sdcmicro


```
# 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

Suppression with R and sdcmicro


```
# launch gui if you like
sdcGui

# and play around some more
```



[Micah Altman, 3/10/2011]

How SDL Methods Reduce Utility

	Issues
Removing variables	Model misspecification
Suppressing records	Induced non-response bias
Sub-sampling	Weak protection
Global recoding (generalization)	Censoring
Local suppression	Non-ignorable missing value bias
Rules-based swapping	Biased, must keep rules for secret
Random swapping	Weakens bivariate, multivariate relationships
Adding noise	Weak protection
Resampling	Weak protection
Synthetic microdata	Destroys unmodeled relationships, not currently widely accepted

[Micah Altman, 3/10/2011]

Types of Disclosure

- Identity disclosure (re-identification disclosure) – associate an individual with a record and set of sensitive variables
- Attribute disclosure (prediction disclosure) – improve prediction of value of sensitive variable for an individual
- Group disclosure -- predict the value of a sensitive variable for a known group of people

Factors affecting disclosure protection

- Properties of the sample
  - Measured variables
  - Realizations of measurements
    - Outliers
    - Content of qualitative responses
- Distribution of population
- Adversarial knowledge
  - Variables
  - Completeness
  - Errors
  - Priors

- Individual reidentification occurs when:
  - Respondent is unique on values of the key
  - Attacker has access to measurements of key
  - Respondent is in attackers set of measurements
  - Attacker comes across disclosed data
  - Attacker recognizes respondent

[Willenborg & DeWaal 1996]

Disclosure protection: k-anonymity [Sweeney 2002]

- Operates on *micro-data*
- Designate subset of variables as *key's* – variables that the attacker could use to identify individual
- For each combination of key variables in the *sample* – there must be *k* rows taking on that combination
- *k* is typically desired to be in 3-5

Our table made 2-anonymous (one way)

Cleaned Quasi-keys

Name	SSN	Birthdate	Zipcode	Gender	Favorite Ice Cream	# of crimes committed
* Jones	*	* 1961	021*	M	Raspberry	0
* Jones	*	* 1961	021*	M	Pistachio	0
* Jones	*	* 1972	9404*	*	Chocolate	0
* Jones	*	* 1972	9404*	*	Hazelnut	0
* Jones	*	* 1972	9404*	*	Lemon	0
* Jones	*	*	021*	F	Lemon	1
* Jones	*	*	021*	F	Peach	1
* Smith	*	* 1973	63*	*	Lime	2
* Smith	*	* 1973	63*	*	Mango	4
* Smith	*	* 1973	63*	*	Coconut	16
* Smith	*	* 1974	64*	M	Frog	32
* Smith	*	* 1974	64*	M	Vanilla	64
* Smith	*	04041974	64*	F	Pumpkin	128
* Smith	*	04041974	64*	F	Allergic	256

Both more and less than HIPAA default

[Micah Altman, 3/10/2011]

k-anonymous – but not protected

	Sort Order/ Structure				Additional background	
Name	**SSN**	**Birthdate**	**Zipcode**	**Gender**	**Favorite Ice Cream**	**# of crimes committed**
* Jones	*	* 1961	021*	M	Raspberry	0
* Jones	*	* 1961	021*	M	Pistachio	0
* Jones	*	* 1972	9404*	*	Chocolate	0
* Jones	*	* 1972	9404*	*	Hazelnut	0
* Jones	*	* 1972	9404*	*	Lemon	0
* Jones	*	*	021*	F	Lemon	1
* Jones	*	*	021*	F	Peach	1
* Smith	*	* 1973	63*	*	Lime	2
* Smith	*	* 1973	63*	*	Mango	4
* Smith	*	* 1973	63*	*	Coconut	16
* Smith	*	* 1974	64*	M	Frog	32
* Smith	*	* 1974	64*	M	Vanilla	64
* Smith	*	04041974	64*	F	Pumpkin	128
* Smith	*	04041974	64*	F	Allergic	256

Homogeneity

[Micah Altman, 3/10/2011]

More than one way to de-identify (but don't release both…)

- Law, policy, ethics
- Research design …
- Information security
- **Disclosure limitation**

Name	SSN	Birthdate	Zipcode	Gender
* Jones	*	*1961	021*	*
* Jones	*	*1961	021*	*
* Jones	*	*1972	94043	*
* Jones	*	*1972	94043	*
* Jones	*	03251972	*	*
* Jones	*	03251972	*	*
*	*	*	*	*
*	*	*	*	*
* Smith	*	02021973	6*	*
* Smith	*	02021973	6*	*
* Smith	*	03031974	6*	*
* Smith	*	04041974	6*	*
* Smith	*	04041974	6*	*
* Smith	*	04041974	6*	*

Name	SSN	Birthdate	Zipcode	Gender
* Jones	*	* 1961	021*	M
* Jones	*	* 1961	021*	M
* Jones	*	* 1972	9404*	*
* Jones	*	* 1972	9404*	*
* Jones	*	* 1972	9404*	*
* Jones	*	*	021*	F
* Jones	*	*	021*	F
* Smith	*	* 1973	63*	*
* Smith	*	* 1973	63*	*
* Smith	*	* 1973	63*	*
* Smith	*	* 1974	64*	M
* Smith	*	* 1974	64*	M
* Smith	*	04041974	64*	F
* Smith	*	04041974	64*	F

[Micah Altman, 3/10/2011]

Vulnerabilities of k-anonymity

- Sort order [Sweeney 2002]
  - Information in structure of data, not content!
- Contemporaneous release [Sweeney 2002]
  - overlap of information under different anonymization schemes→ disclosure
- Information in suppression mechanism, may allow recovery → – e.g. rules based swapping
- Temporal changes
  - "barn door" -- deletion of tuples can subvert k-anonymity → can't "unrelease" records
  - Additions of tuples, information can yield disclosures if you re-do anonymization → must anonymize these based on the past data release [Sweeney 2002]
- Variable Background Knowledge [Machanavajjhala 2007]
  - Incorrect assumption about what variables are in quasi-key
  - This may change over time
- Homogeneity [Truta 2006]
  - Sensitive values may be homogenous, even if not literally individually identified

Strengthening k-anonymity vs. homogeneity

- Ensure each k-anonymous set also satisfies some measure of attribute diversity
  - P-sensitive k-anonymity [Truta 2006]
  - Fixed l-diversity, Entropy l-diversity, Recursive (c,l) diversity [Machanavajjhala 2007]
  - T-closeness [Li 2007]
- Diversity measures may be too strong or too weak
- And sometimes attribute disclosure is not justifiable
  - It does not literally (legally?) identify an individual
  - Research may be explicitly designed to make attribute more predictable
  - In some cases, study would probabilistically identify an attribute, even if participant weren't in it!

Sometimes k-anonymity is too strong

- Embodies several worst case assumptions -- safer, but more information loss:
  - Sample unique → population unique
  - Attacker discovers your data with certainty
  - Attacker has complete database of non-sensitive variables and their links to identifiers
  - Attacker database and sample are error-free

[Micah Altman, 3/10/2011]

Research Areas

- Standard SDL approaches are designed to apply to dense s… tables of quantitative data… use caution & seek consultation with the following

- Dynamic data
  - Adding new attributes
  - Incremental updates
  - Multiple views
- Relational data
  - Multiple relations that are not easily normalized
- Non tabular data
  - Sparse matrices
  - Transactional data
  - Trajectory data
  - Rich text
  - Social networks

[Micah Altman, 3/10/2011]

Problem 2: Information loss

- No free lunch: anonymization → information loss
- Various approaches none satisfactory or commonly used
  - Count number of suppressed values
  - Compare data matrix before & after anonymization
    - Entropy, MSE, MAE, mean variation
  - Compare statistics on data matrix before & after
    - Variance, Bias, MSE
  - Weight by (ad-hoc) importance of variable
- Optimal (information loss) k-anonymity is NP-hard [Meyerson & Williams 2004]
- **Utility degrade very fast in increased privacy**
  - **See [Brickell and Shmatikov 2008; Ohm 2009,; Dinur & Nissim 2004; Dwork et al 2006, 2007]**

Alternative risk limitation–non-microdata approaches

- Models and tables can be safely generated from anonymized microdata, however information loss may be less when anonymization is applied at the model/table level directly
- Model servers
  - Compute models on full microdata
  - Limit models being run on data
  - Limit specifications of models
  - Synthesize residuals; perturb results
- Table-based de-identification
  - Compute tables on full micro-data
  - Perturb (noise, rounding), suppress cells (and complementary cells, if marginals computed), restructure tables (generalization, variable suppression), synthesize value
  - Disclosure rule: number of contributors to a cell (similar to k-anonymity); proportion of largest group of contributors to a cell total; percentage decrease in upper/lower bounds on contributor values
- Limitations
  - Feasible (privacy protecting) multi-dimensional table/multiple table protection is NP-hard
  - Model/table disclosure requires evaluating entire history of previous disclosures
  - Dynamic table servers, model servers should be considered open research topics, not mature.

[Micah Altman, 3/10/2011]

Alternate solution concept – probabilistic record linkage

- Apply disclosure rule to population based on threshold probability, and estimated population distribution
- E.g. for **3**-anonymity – *probability* < ***.02*** that there exists a tuple of quasi-identifier values that occurs < 3 time *in the population*
- Advantages
  - When sample is small, population risk model will result in far less modification & information loss
- Disadvantages
  - Harder to explain.
  - Does not literally prevent individual reidentification.
  - Need to justify reidentification risk threshold
  - Need to justify population distribution model
  - *Assumes that background knowledge of attacker does not include whether each identified individual is the sample*

[Micah Altman, 3/10/2011]

Alternate Solution Concept – Bayesian Optimal Privacy

- Possibly…
  - Minimal distance between posterior and prior distribution for some all priors…

$$\Delta\left(\pi(.),\pi(.\mid B)\right) \leq \varepsilon$$

- Limitations… [See A. Mchanavajjala, et. al 2007]
  - Insufficient knowledge about distributions of attributes
  - Insufficient knowledge about distributions of priors
  - Instance-level knowledge not modeled well
  - Multiple adversaries not modeled
- Possible limitations
  - Complexity of computation not known
  - Implementation mechanisms not well-known
  - Utility reduction not well-known

Alternate Solution Concept– Differential Privacy

$$\pi\left(K_i(D), \forall i \in C\right) \leq e^{\varepsilon} \pi\left(K_i\left(D^*\right), \forall i \in C\right),$$

$$\Delta\left(D, D^*\right) = 1 \text{ record}$$

- Based on cryptography theory (traitor tracing schemes) & provides formal bounds on disclosure risk across all inferences -- handles attribute disclosure well [Dwork 2006]
- Roughly, differential privacy guarantees that all inferences made from the data with a subject included will differ only by epsilon if subject is removed.
- Analysis is accompanied by formal analysis of estimator efficiency – differential privacy can be achieved in many cases with (asymptotic) efficiency
- DP is essentially Frequentist … possible Bayesian interpretation
  - Prior: n-1 complete records, and distribution over nth record
  - DP criterion implies Hellinger distance [Fienberg 2009]

Implementing Differential Privacy

- Currently, almost all realizations of differential privacy rely on noise applied to q numeric tabular databases – unknown how to apply it to new forms of data such as networks. [Dwork 2008]
- Static sanitization is possible … BUT limited
  - If possible number of queries in analysis family is superpolynomial in size of data no efficient anonymization exists [Dwork et al 2009]
- Differential privacy methods need to be develop for the type of analysis being performed.
  - Currently differentially private versions of datamining queries exist, but
  - … development of differentially private versions of common statistical methods is just beginning. [Dwork & Smith 2009]
- Differential privacy may be *too* strong, in some cases..
  - identity disclosure may be the appropriate measure
  - disclosing attributes that are the explicit topic of research may be appropriate
  - allowing for greater than epsilon gains in information may be appropriate
- There is only one publicly available software tool that supports these methods (PLINQ)
  - Test use only
  - Restricted domain of queries
- Researchers may need access to data not just coefficients – e.g."show me the residuals"!

DP is probably not the final solution concept..

- DP is only practical and effective when *applied interactively during analysis* of the data for many types of analysis.
  - Implies that to implement needs to be part of security infrastructure with…
  - Testing
  - Multiple implementations
  - Auditing
  - Monitoring
  - Vulnerability and threat assessment in whole system context
- It does not provide strong protection against disclosure of information about groups.
  - No clean model for "sensitive groups"
- Addresses incentives to disclose rather than guarantee privacy of release
  - Does not solve problem of individual privacy protection, just individual incentive to disclose
  - Nash-equilibrium of individual disclosures not necessarily societally optimal
  - Need a game theoretic/social choice theoretic view of privacy!
- Other solution concepts may provide stronger guarantees:
  - Bayesian concepts
  - Learning theory– distributional privacy [Blum, et. al 2008]
- Protection of Distributed Databases – w/untrusted parties?
  - See Karr 2009

MIND THE GAPS – Future Research

Law, policy, ethics
Research design …
Information security
Disclosure limitation

- Reconcile Bayesian an Frequentist notions of privacy
- Model privacy from game theoretic/social choice & policy analysis point of view
- Reconcile "random response"/sensitive survey methods and statistical disclosure concepts
- Disclosure limitation methods needed for new forms of data
- Differential Privacy methods needed for many more statistical models
- Bridge gap between regulatory and statistical views
  - Update regulations/law based on statistical concepts
  - Educate IRB's on statistical disclosure control
  - Integrate permission for data sharing and some disclosure in consent & design of experiments
- Bridge gap between mathematics and implementation
  - Very few software packages available for disclosure limitation and analysis
  - Interactive disclosure limitations require not just software, but validated, audited software infrastrucure
- Data sharing infrastructure needed for managing confidentiality effectively:
  - Applying interactive privacy automatically
  - Implementing limited data use agreements
  - Managing access & logging – virtual enclave
  - Providing chokepoint for human auditing of results
  - Providing systems auditing, vulnerability & threat assessment
  - Ideally:
    - Research design information automatically fed into disclosure control parameterization
    - Consent documentation automatically integrated with disclosure policies, enforced by system

[Micah Altman, 3/10/2011]

What to do – for now…

- (1) Use only information that has already been made public, is entirely innocuous, or has been declared legally deidentified; *or*
- (2) Obtain informed consent from research subjects, at the time of data collection, that includes acceptance of the potential risks of disclosure of personally identifiable information; *or*
- (3) Pay close attention to the technical requirements imposed by law:
  - Remove all 18 HIPAA factors; *or*
  - Use suppression and recoding to achieve k-anonymity with l-diversity on data before releasing it or generating detailed figures, maps, or summary tables.
  - Supplement data sharing with data-use agreements.
  - Apply extra caution & use consultation with "non-traditional" data – networks, text corpuses, etc.

Preliminary Recommendations

- Avoid complexities of table and model SDL
  - Apply SDL to microdata
  - Tables and models based on deidentified microdata are de-identified
- Use substantive knowledge to guide disclosure limitation
  - Globally recode using natural categories
  - Use local suppression – check suppressed observations
  - Estimate substantively interesting statistics from original and modified data as a check

[Micah Altman, 3/10/2011]

Key Concepts Review

- Text de-identification
- License and access control restrictions
- K-anonymity
- Suppression
- Attribute homogeneity
- Risk/utility tradeoff

Checklist

- ✓ Will license be used to limit disclosure?
- ✓ Will enclave or remote access limit disclosure?
- ✓ Are there natural categories for global recoding?
- ✓ Is there a natural measure of information loss, or natural weighting for importance of variables?
- ✓ What level of reidentification risk is acceptable?
- ✓ What is expected background knowledge of attacker?

Available Software

- Deidentification of text
  - Regular expression, lookup tables, template matching
    [www.physionet.org/physiotools/deid]
- Deidentification of IP addresses and system/network logs
  www.caida.org/tools/taxonomy/anonymization.xml
- Interactive Privacy
  - PINQ – Experimental interactive differential privacy engine
    [research.microsoft.com/en-us/projects/PINQ/]
- Tabular Data – Tau Argus
  - Cell suppression, controlled rounding
    [neon.vb.cbs.nl/casc]
- Microdata
  - Mu-Argus
    - Microaggregation, local suppression, global recoding, PRAM
      [neon.vb.cbs.nl/casc]
  - SDCmicro
    - Microaggregation, local suppression, global recoding, PRAM, rank swapping
    - Heuristic k-anonymity (using local suppression)
    - R module
      [cran.r-project.org/web/packages/sdcMicro]
  - NISS Data Swapping Toolkit (DSTK)
    - Data swapping in risk/utility framework
    - Implemented in Java

    [nisla05.niss.org/software/dstk.html]

Resources

- **FCSM, 2005. "Report on Statistical Disclosure Limitation Methodology", FCSM Statistical Working Paper Series**
  [www.fcsm.gov/working-papers/spwp22.html]
- **L. Willenborg, T. de Waal, 2001. *Elements of Statistical Disclosure Control,* Springer.**
- **ICPSR Human Subjects Protection Project Citation Database**
  [www.icpsr.umich.edu/HSP/citations]
- **A. Hundepool, et *al.* 2009, *Handbook of Statistical Disclosure Control,* ESSNET**
  [neon.vb.cbs.nl/casc/..%5Ccasc%5Chandbook.htm]
- **Privacy in Statistical Database Conference Series**
  [unescoprivacychair.urv.cat/psd2010/]
  (See *Springer's Lecture Notes in Computer Science series for previous proceedings volumes*)
- ASA Committee on Privacy and Confidentiality Website
  [www.amstat.org/committees/pc]
- National Academies Press, Information Security Book Series
  [www.nap.edu/topics.php?topic=320]
- National Institute of Statistical Sciences, Technical Reports
  [www.niss.org/publications/technical-reports]
- *Transactions on Data Privacy*, IIIA-CSIC [Journal]
  [www.tdp.cat]
- *Journal of Official Statistics*, Statistics Sweden:
  [www.jos.nu]
- *Journal of Privacy and Confidentiality*, Carnegie-Mellon
  [jpc.cylab.cmu.edu]
- *IEEE Security and Privacy*
  [*www.computer.org/**security***]
- Census Statistical Disclosure Control checklist
  [www.census.gov/srd/sdc]
- B. C.M. Fung, K. Wang, R. Chen, P.S. Yu, 2010, Privacy Preserving Data Publishing: A Survey of Recent Developments, ACM CSUR 42(4)

Additional Resources

- Final review
- Additional training resources
- **Harvard Consulting**
- **Handout for Harvard staff**
- **Harvard IQSS Research Support**
- Additional references

Final Review: 7 Steps

- **Identify** potentially sensitive information in *planning*
  - Identify legal requirements, institutional requirements, data use agreements
  - Consider obtaining a certificate of confidentiality
  - Plan for IRB review
- **Reduce** sensitivity of collected data in *design*
- **Separate** sensitive information in *collection*
- **Encrypt** sensitive information in *transit*
- **Desensitize** information in *processing*
  - Removing names and other direct identifiers
  - Suppressing, aggregating, or perturbing indirect identifiers
- **Protect** sensitive information in *systems*
  - Use systems that are controlled, securely configured, and audited
  - Ensure people are authenticated, authorized, licensed
- **Review** sensitive information before *dissemination*
  - Review disclosure risk
  - Apply non-statistical disclosure limitation
  - Apply statistical disclosure limitation
  - Review past releases and publically available data
  - Check for changes in the law
  - Require a use agreement

Preliminary Recommendation: Choose the Lesser of Three Evils

- (1) Use only information that has already been made public, is entirely innocuous, or has been declared legally deidentified; *or*
- (2) Obtain informed consent from research subjects, at the time of data collection, that includes acceptance of the potential risks of disclosure of personally identifiable information; *or*
- (3) Pay close attention to the technical requirements imposed by law:
  - Use suppression and recoding to achieve k-anonymity with l-diversity on data before releasing it or generating detailed figures, maps, or summary tables.
  - Supplement data sharing with data-use agreements.

[Micah Altman, 3/10/2011]

Preliminary Recommendations Planning and methods

- Review research design for sensitive identified information
  - Information which would cause harm if disclosed
  - HIPAA identifiers
  - Other indirectly identifying characteristics
- Design research methods to reduce sensitivity
  - Eliminate sensitive/identifying information not needed for research questions
  - Consider randomized response, list experiment design
- Design human subjects plan with information management in mind
  - Recognize benefits of data sharing
  - Ask for consent to share data appropriately
  - Apply for a *certificate of confidentiality* where data is very sensitive
- Separate sensitive information
  - Separate sensitive/identifying information at collection, if feasible
  - Link separate files using cryptographic hash of identifiers *plus* secret key; or *cryptographic-strength random number*
- Incorporate extra protections for on-line data collection
  - Use vendor agreements that specify anonymity and confidentiality protections
  - Do not collect IP addresses if possible, regularly anonymize and purge otherwise
  - Restrict display of very sensitive information in user interfaces
  - Limit on-line collection of very sensitive information
  - **Harvard prohibits display/collection of HRCI online**

[Micah Altman, 3/10/2011]

Preliminary Recommendations Information Security

- Use FISMA as a reference for baseline controls
- Document:
  - Protection goals
  - Threat models
  - Types of controls
- Delegate implementation to IT professionals
- Refer to standards
  - Gold standards: FISMA / ISO practices, SAS-70 Auditing, CISSP certification of key staff
- Strongly recommended controls
  - Use whole-disk/media encryption to protect data at rest
  - Use end-to-end encryption to protect data in motion
  - Use core information hygiene to protect systems
    - Use a virus checker, and keep it updated
    - Use a host-based firewall
    - Update your software regularly
    - Install all operating system and application security updates
    - Don't share accounts or passwords
    - Don't use administrative accounts all the time
    - Don't run programs from untrusted sources
    - Don't give out your password to anyone
  - Scan for HRCI regularly
  - Be thorough in disposal of information
    - Use secure file erase tools when disposing of files
    - Use secure disk erase tools when disposing/repurposing disks

[Micah Altman, 3/10/2011]

Preliminary Recommendations

Very Sensitive/Extremely Sensitive Information security

- Protect very sensitive data on "target systems"
  - Extra physical, logical, administrative access control
    - Record keeping
    - Limitations
    - Lockouts
  - Extra monitoring, auditing
  - Extra procedural controls – specific, renewed approvals
  - Limits on network connectivity
    - Private network, not *directly connected to public network*
- Regular scans
  - Vulnerability scans
  - Scans for PII
- Extremely sensitive
  - Increased access control, procedural limitations
  - Not physically/logically connected (even via wireless) to public network, directly or indirectly

Preliminary Recommendations non tabular data disclosure

- Use licensing agreements – even if they are "clickthroughs"
  Reason: They provide additional protection without limiting legitimate research.
- For qualitative text information
  - Use software for the first pass
  - Supplement with localized dictionary of place names, common last names, etc
  - Have a human review results

  Reason: Software more effective than single human coder. However error rate high enough that human still necessary.
- For emerging forms of data (networks, etc.)
  - Use remote access, and user authentication, if feasible
    Reason: Greater auditability to compensate for less well understood statistical de-identification.
  - Pay careful attention to structure of data.
    Reason: Identifying information may be present in structure of information (word ordering, prose style, network topology, sparse matrix missingness) rather than in the primary attribute information

[Micah Altman, 3/10/2011]

Preliminary Recommendations Tabular data disclosure

- Use licensing agreements – even if they are "clickthroughs"
  Reason: They provide additional protection without limiting legitimate research.
- Use HIPAA default variable suppression and recoding *if according to the PI's best judgment this does not seriously degrade the research value of the data.*
  Reason: Clearest legal standard
- For quantitative tabular data
  - Use generalization, local suppression, variable suppression.
    Reason: These are effective, commonly used in HIPAA and in statistical disclosure control
  - Use k-anonymity
    Reason: k-anonymity appears to be current good practice; provably eliminates literal individual re-identification; works if attacker has knowledge of sample participation
  - Choose k in [3-5]
    Reason: Best practice at federal agencies for *table* suppression requires table cells to have 3-5 contributors. Tables derived from k-anonymous microdata will also fulfill this.
  - Choose quasi-identifiers based on plausible threat models
    Reason: Too broad a definition of quasi-identifiers renders de-identification impossible. Background knowledge is pivotal, and threat model is the only source for this.
  - Use micro-data anonymization, rather than tabular/model anonymization
    Reason: (1) Table/model methods become computational intractable. (2) Analysis of model-anonymization is immature. (3) Anonymizing microdata implies derived tables and models are also anonymized. (4) Administratively harder to track and evaluate entire history of previous models/tables than history of previously released versions of single micro-data set.
  - Use domain knowledge in choosing recodings and testing the resulting anonymization for information loss.
    Reason: MSE , etc. probably not a good proxy for research value of data. Use standard measures, but also consider planned uses and simulate possible analyses.
  - Inspect data for attribute diversity, use PI's judgement regarding suppression
    Reason: (1) Some attribute disclosures are not avoidable if research is to be conducted at all, some would occur even if subject had not participated. (2) Disclosures that would not have resulted if subject had opted out, and are not substantially based on representative causal/predictive relationships revealed by the research, should be eliminated. (3) All current diversity measures are likely to severely reduce the utility of the anonymized data if applied routinely.

On-line training

- NIH Protecting Human Subject Research Participants
  - Provides minimal testing and certification
  - Required for human subjects research at NIH

    [phrp.nihtraining.com]
- NIH Security and Privacy Awareness
  - Includes basics of information security, review of privacy laws

    [irtsectraining.nih.gov/]

- **Harvard Staff Training**
  - Provides compact training for staff members in handling of confidential information

    [http://www.security.harvard.edu/resources/training]

- **Collaborative Institute Training Initiative**
  - Provides testing, certification, continuing education credits
  - Required for human subjects research at Harvard
  - Includes basic training on confidentiality, and informed consent

    [https://www.citiprogram.org/]

Handling Confidential Information @ Harvard

- 1. How to identify confidential information
  - Extremely Sensitive/ "Level 5"
    - Research data containing private extremely sensitive information about identifiable individuals
  - High Risk Confidential Information/HRCI/ "Level 4"
    - A person's *name* + *state*, *federal or financial identifiers*
    - *Or* research data containing private very sensitive information about identifiable individuals
  - Harvard Confidential Information/HCI/"Level 3"
    - Business information *specifically designated by the School as confidential*
    - *Or* identifiable *business information* that puts individuals at risk if disclosed
    - *Or* research data containing private sensitive information about identifiable individuals
    - *Or* student records (such as collections of grades, correspondence)
  - Benign Identified Information/ "Level 2"
- 2. **You are responsible** for confidential information you store, access, or share
  - Obtaining appropriate approval to access information
    - Approval from IRB to use or collect private identified information
    - Approval from PI for access to research data
    - Approval from OSP to sign any external data use agreements
  - Protect your system [Level 2+]
    firewall, virus scanner, limit accounts, good passwords, don't share accounts
  - Encrypt all laptops, portable storage, and network connections [Level 3+]
  - Keep hard-copy/media locked up when not in use [Level 3+]
  - Keep data *only* on specifically designated servers [Level 4+]
  - Keep isolated from any external network [Level 5]
- 3. Safely dispose of confidential information
  - When you change computers – contact IT for cleanup
  - Shred the rest
- 4. If unsure, seek help or approval
  - Consulting: http://www.iq.harvard.edu/data_collection_management_analysis
  - Research approvals: cuhs@fas.harvard.edu

[Micah Altman, 3/10/2011]

Harvard: Consultation and Approvals

[Consultation]

University security policy: http://security.harvard.edu/

Consultation on data management, dissemination, plans:
http://www.iq.harvard.edu/data_collection_management_analysis

[FAS Approval of IT, vendors, use of business confidential info]

Jay Carter,
Director of Information Security, FAS

jcarter@fas.harvard.edu

[Approval for obtaining confidential information in research]

Harvard Institutional Review Board

www.fas.harvard.edu/~research/hum_sub/

The PI of the research project is responsible for further approvals for use of that information, consistent with the research plan approved by the IRB.

[Approval to obtain external confidential research information]

Office of Sponsored Programs

http://security.harvard.edu/files/resources/forms/

Tel: 617.495.5501

Harvard IQSS Research Support

- IQSS supports your research design:
  - Research design, including:
    design of surveys, selection of statistical methods.
- IQSS supports your research process:
  - Primary and secondary data collection, including:
    the collection of geospatial and survey data.
  - Data management, including:
    storage, cataloging, permanent archiving, and distribution.
  - Data analysis, including :
    statistical consulting, GIS consulting, high performance research computing
- IQSS supports your projects
  - Dissemination: web site hosting, scholars website
  - Research computing infrastructure and hosting
  - Conference/seminar/event planning and facilities

Strengthen your proposal through:

- Consultation on research design, statistical issues, GIS, research computing
- Including relevant resources in "facilities" etc.
- Obtaining IQSS letters of support

[Micah Altman, 3/10/2011]

Additional References

- *A Aquesti, L John, G Lowestein, 2009, "What is Privacy Worth", 21rst Rowkshop in Information Systems and Economics.*
- *A. Blum, K. Ligett, A Roth, 2008. "A Learning Theory Approach to Non-Interactive Database Privacy", STOC'08*
- L. Backstrom, C. Dwork, J. Kleinberg, 2007, Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography. Proc. 16th Intl. World Wide Web Conference., KDD 008
- J. Brickell, and V. Shmatikov, 2008. The Cost of Privacy: Destruction of Data-Mining Utility in Annoymized Data Publishing
- P. Buneman, A. Chapman an.d J. Cheney, 2006, 'Provenance Management in Curated Databases', in Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, (Chicago, IL: 2006), 539-550. http://portal.acm.org/citation.cfm?doid=1142473.1142534;
- Calabrese F., Colonna M., Lovisolo P., Parata D., Ratti C., 2007, "Real-Time Urban Monitoring Using Cellular Phones: a Case-Study in Rome", Working paper # 1, *SENSEable City Laboratory*, MIT, Boston http://senseable.mit.edu/papers/, [also see the Real Time Rome Project [http://senseable.mit.edu/realtimerome/]
- Campbell,. D. 2009, reported in D. Goodin 2009, Amazon's EC2 brings new might to password cracking, *The Register,* Nov 2, 2009, http://www.theregister.co.uk/2009/11/02/amazon_cloud_password_cracking/
- Dinur and K. Nissim. Revealing information while preserving privacy. Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pages 202–210, 2003.
- C. Dwork, M Naor, O Reingold, G Rothblum, S Vadhan, 2009. *When and How Can Data be Efficiently Released with Privacy*, STOC 2009.
- C Dwork, A. Smith, 2009. Differential Privacy for Statistics: What we know and what we want to learn, Journal of Privacy and Confdentiality 1(2)135-54
- C Dwork 2008, Differential Privacy, A Survey of Results. TAMC 2008, LCNS 4978, Springer Verlag. 1-19
- C. Dwork. Differential privacy. Proc. ICALP, 2006.
- C. Dwork, F. McSherry, and K. Talwar. The price of privacy and the limits of LP decoding. Proceedings of the thirty-ninth annual ACM Symposium on Theory of Computing, pages 85–94, 2007.
- C. Dwork, F. McSherry, K. Nissim, and A. Smith, Calibrating Noise to Sensitivity in Private Data Analysis, Proceedings of the 3rd IACR Theory of Cryptography Conference, 2006
- A. Desrosieres. 1998. *The Politics of Large Numbers*, Harvard U. Press.
- S.E. Fienberg, M.E. Martin, and M.L. Straf (eds.), 1985. Sharing Research Data, Washington, D.C.: National Academies Press.
- S. Fienberg, 2010. Towards a Bayesian Characterization of Privacy Protection & the Risk-Utility Tradeoff, IPAM--Data 2010
- B. C.M. Fung, K. Wang, R. Chen, P.S. Yu, 2010, Privacy Preserving Data Publishing: A Survey of Recent Developments, ACM CSUR 42(4)
- Greenwald, A. G. McGhee, D. E. Schwartz, J. L. K., 1998, "Measuring Individual Differences In Implicit Cognition: The Implicit Association Test", Journal of Personality and Social Psychology 74(6):1464-1480
- C. Herley, 2009, So Long and No Thanks for the Externalities: The Rational Rejection of Security Advice by Users; NSPW 09
- A. F. Karr, 2009 Statistical Analysis of Distributed Databases, journal of Privacy and Confidentiality (1)2:

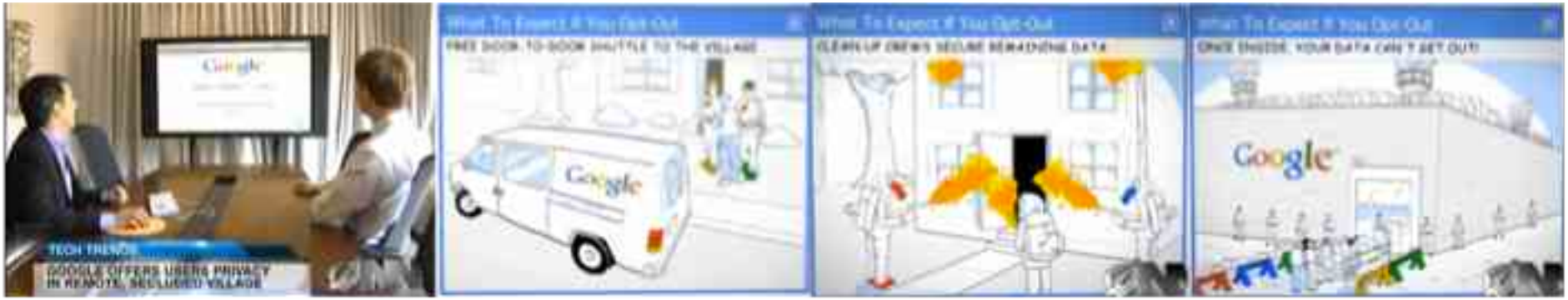
Additional References

- International Council For Science (ICSU) 2004. ICSU Report of the CSPR Assessment Panel on Scientific Data and Information. Report.
- J. Klump, et. al, 2006. "Data publication in the open access initiative", Data Science Journal Vol. 5 pp. 79-83.
- E.A. Kolek, D. Saunders, 2008. Online Disclosure: An Empirical Examination of Undergraduate Facebook Profiles, NASPA Journal 45 (1): 1-25
- N. Li, T. Li, and S. Venkatasubramanian. T-closeness: privacy beyond k-anonymity and l-diversity. In Pro- ceedings of the IEEE ICDE 2007, 2007.
- A. MachanavaJJhala, D Kifer, J Gehrke, M. Venkitasubramaniam, 2007,"l-Diversity: Privacy Beyond k-Anonymity" *ACM Transactions on Knowledge Discovery from Data*, 1(1): 1-52
- A. Meyerson, R. Williams, 2004. "On the complexity of Optimal K-Anonymity", ACM Symposium on the Principles of Database Systems
- Nature 461, 145 (10 September 2009) | doi:10.1038/461145a
- A. Narayanan and V. Shmatikov, 2008, "Robust De-anonymization of Large Sparse Datasets" , *Proc. of 29th IEEE Symposium on Security and Privacy* (Forthcoming)
- I Neamatullah, et. al, 2008, Automated de-identification of free-text medical records, *BMC Medical Informatics and Decision Making* 8:32
- J. Novak, P. Raghavan, A. Tomkins, 2004. Anti-aliasing on the Web, Proceedings of the 13th international conference on World Wide Web
- National Science Board (NSB), 2005, Long-Lived Digital Data Collections: Enabling Research and Education in the 21rst Century, NSF. (NSB-05-40).
- A Qcquisti, R. Gross 2009, "Predicting Social Security Numbers from Public Data", *PNAS 27(106): 10975–10980*
- Sweeney, L., (2002) k-Anonymity: A Model for Protecting Privacy, International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems, Vol. 10, No. 5, pp. 557 – 570.
- Truta T.M., Vinay B. (2006), Privacy Protection: p-Sensitive k-Anonymity Property, International Workshop of Privacy Data Management (PDM2006), In Conjunction with 22th International Conference of Data Engineering (ICDE), Atlanta, Georgia.
- O. Uzuner, et al, 2007, "Evaluating the State-of-the-Art in Automatic De-identification", *Journal of the American Medical Informatics Association* 14(5):550
- W. Wagner & R. Steinzor, 2006. *Rescuing Science from Politics,* Cambridge U. Press.
- Warner, S. 1965. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association 60(309):63–9.
- D.L. Zimmerman, C. Pavlik , 2008. "Quantifying the Effects of Mask Metadata, Disclosure and Multiple Releases on the Confidentiality of Geographically Masked Health Data", *Geographical Analysis* 40: 52-76

Questions?

Web:

maltman.hmdc.harvard.edu



[Micah Altman, 3/10/2011]

Creative Commons License