Information Privacy and Data Governance when Working with Real Data

Hye-Chung Kum (kum@tamu.edu)
Associate Professor
Population Informatics Lab (https://pinformatics.org/)

PUBLIC HEALTH TEXAS A&M UNIVERSITY

POPULATION INFORMATICS

2

2

Agenda

- Concepts
- HIPAA
- Privacy: social & legal concepts
- Privacy: theoretical concepts
  - Differential privacy
- Approaches to Privacy

3

Do you want to drive on the information highway?

- Be prepared to get into accidents, especially as we all figure out how to do this more safely.
- Invest in personnel to do your due diligence to stay up to date with the fast paced security technology and privacy regulations
- Invest in technologies (i.e., seat belts & airbags) that can minimize the real harm when accidents do occur

4

The Digital World for you

- Live in the digital world
- Use privacy settings
- Know how to use privacy settings
- Felt privacy evaded

5

Vocab: Information Privacy

- What is *information* privacy?
- Privacy vs confidentiality
  - don't ask vs don't tell
- Privacy vs security
- Authorization (consent)
  - Opt in
  - Opt out
  - Blanket consent
- PHI: Protected Health Information
- PII: Personally Identifiable Information

6

Definitions

- Privacy (Don't ask)
  - An individual's right to be left alone and to limit access to his or her health care information
  - Who: person giving information
- Confidentiality (Don't tell)
  - Addresses the expectation that information shared with a health care provider during the course of treatment will be used only for its intended purpose and not disclosed otherwise
  - Who: person receiving information
- Security
  - The systems in place to protect health information and the systems within which it resides.
  - Technical tool to make privacy & confidentiality possible
- Authorization/Consent
  - Legal tools of use, disclosure, and sharing of data
  - Opt in, Opt out, Blanket consent, Broad Consent (Revised Common Rule, 2018). Wavier of Consent
- Accountability
  - So it is possible to determine whether a particular use is appropriate under a given set of rules
  - The system enables individuals and institutions to be held accountable for misuse
  - Transparency and accountability makes bad acts visible to all concerned
- PHI: Protected Health Information
- PII: Personally Identifiable Information
- Limited Data, DUA, Etc…

7

7

Vocab: Disclosure

- Identity disclosure
- Attribute disclosure
- Harm from disclosure
  - Identity theft: SSN, Name, DOB
  - HIV status
- Group disclosure
- Partial disclosure
- Incremental disclosure
- Minimum necessary standard
  - Cost of implementation?

8

Agenda

- Concepts
- HIPAA
- Privacy: social & legal concepts
- Privacy: theoretical concepts
  - Differential privacy
- Approaches to Privacy

9

HIPAA: Privacy Rule

- Health Insurance Portability and Accountability Act
- NOT HIPPA
- Privacy Rule & Security Rule
- http://www.hhs.gov/ocr/privacy/hipaa/understanding/
- Key elements of the Privacy Rule
  - who is covered
  - what information is protected
  - how protected health information can be used and disclosed

10

HIPAA: who is covered "Covered Entities"

- Health plans
- Health care providers
- Health care clearinghouse
- BA
  - business associate is a person or organization, other than a member of a covered entity's workforce, that performs certain functions or activities on behalf of, or provides certain services to, a covered entity that involve the use or disclosure of individually identifiable health information
- OLD: HSC as HIPAA entity.
  - NOW, HSC not HIPAA entity
  - TAMU is a hybrid entity
  - http://www.tamhsc.edu/institutional-compliance/hipaa-officials.html

11

HIPAA: what information is protected
PHI (Protected Health Information)

- "individually identifiable health information"
- held or transmitted by a covered entity or its business associate, in any form or media, whether electronic, paper, or oral
- The Privacy Rule calls this information "protected health information" (PHI)
- Other form of information
  - Limited Dataset
  - De-identified

12

PHI: List of 18 Identifiers

- 1. Names;
  2. All geographical subdivisions smaller than a State (except for the initial three digits of a zip code, if more than 20,000 people)
  3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89;
  4. Phone numbers;
  5. Fax numbers;
  6. Electronic mail addresses;
  7. Social Security numbers;
  8. Medical record numbers;
  9. Health plan beneficiary numbers;
  10. Account numbers;
  11. Certificate/license numbers;
  12. Vehicle identifiers and serial numbers, including license plate numbers;
  13. Device identifiers and serial numbers;
  14. Web Universal Resource Locators (URLs);
  15. Internet Protocol (IP) address numbers;
  16. Biometric identifiers, including finger and voice prints;
  17. Full face photographic images and any comparable images; and
  18. Any other unique identifying number, characteristic, or code (note this does not mean the unique code assigned by the investigator to code the data)

13

HIPAA: how PHI can be used and disclosed

- Basic Principle
  - A covered entity may not use or disclose protected health information, except either:
  - (1) as the Privacy Rule permits or requires;
  - or (2) as the individual who is the subject of the information (or the individual's personal representative) authorizes in writing.
- Exceptions
  - Public Interest and Benefit Activities

14

Important HIPAA Disclosure Exceptions

- Many HIPAA provisions allow the use and release of protected health information!!!
- No requirement for authorization or opportunity to object for:
  - Public health activities
    - PH authority authorized to collect information
      - ✓ i.e., state and local legal authorities
    - Other entities
  - Research
    - IRB or Privacy Board approval for Waiver
      - ✓ IRB must follow Common Rule
  - Others
    - E.g., uses and disclosures required by law, for law enforcement activities, health oversight activities

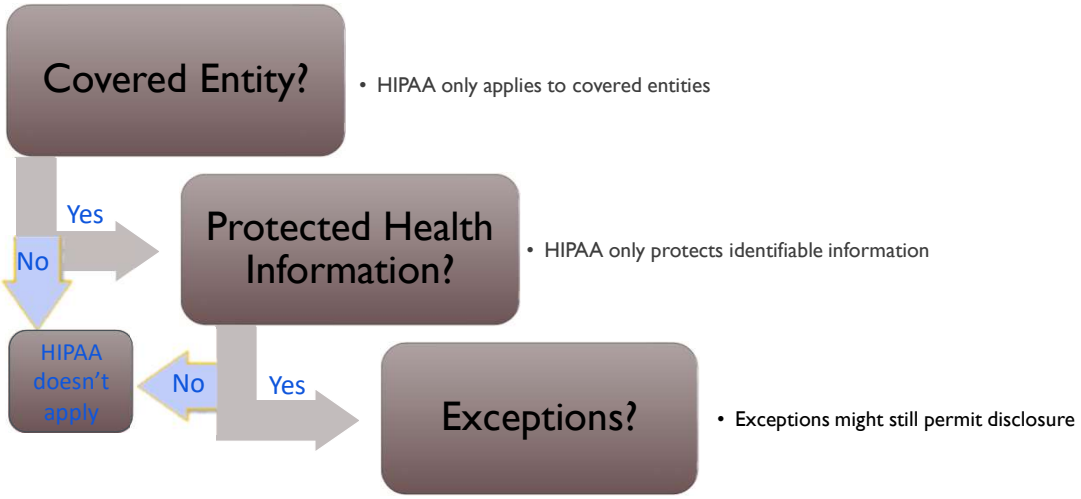
15

HIPAA: Penalties and enforcement

- HIPAA establishes both civil monetary penalties and federal criminal penalties for the impermissible use or disclosure of unsecured PHI in violation of HIPAA's Privacy and Security Rules.
- Civil penalties range from $100 per violation per incident, to $1,500,000 for all such violations of a single provision in a calendar year.
- Criminal penalties include fines up to $250,000 and up to ten (10) years imprisonment.
- HIPAA Hall of Shame: affecting 500+ people
  - https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf

16

Basic HIPAA Privacy Rule Flow Chart



17

Limited Dataset v. Minimum Necessary

- Limited Data set Disclosure
  - A specific type of disclosure
  - Specific data elements are excluded
  - Permitted uses:
    - Research
    - Public health
    - Health care operations
  - Requires a Data Use Agreement
- Minimum Necessary Standard
  - Standard for many permitted HIPAA disclosures
  - Get what you need
    - *Including data elements that would be excluded from a Limited Dataset disclosure*
  - Reasonable effort
    - Covered entity may rely (if reasonable) on PH official's representations

18

Agenda

- Concepts
- HIPAA
- Privacy: social & legal concepts
- Privacy: theoretical concepts
  - Differential privacy
- Approaches to Privacy

19

HEW Code of fair information practices (1973)

Openness	There must be no personal data record keeping systems whose very existence is secret.
Access	There must be a way for an individual to find out what information about him is in a record and how it is used.
Control	There must be a way for an individual to prevent information about him that was obtained for one purpose from being used or made available for other purposes without his consent.
Integrity	There must be a way for an individual to correct or amend a record of identifiable information about him.
Security	Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for their intended use and must take precautions to prevent misuse of the data.

U.S. Department of Health, Education and Welfare (HEW). 1973. In Report of the Secretary's Advisory Committee on Automated Personal Data Systems: Records, Computers and the Rights of Citizens.

20

Social Issues: Balance between

- Individual privacy
  - Secrecy does not work very well : accuracy of data
- Cost of integrity of data
  - incorrect analysis that lead to can lead to wrong decisions
- Organization transparency & accountability
- Freedom of speech
  - marketing is freedom to express why one should prescribe certain drugs
  - marketing is freedom to send junk mail & call
  - thus, getting more information to better target is acceptable and should be allowed

21

Privacy as contextual integrity (legal)

- Helen Nissenbaum (NYU Law School)
- *Washington Law Review, Vol. 79, No. 1, 2004*
- a conceptual framework for understanding privacy expectations and their implications developed in the literature on law, public policy, and political philosophy
- Privacy Protection / Violation
  - Social norms of expectation (on use, sharing etc)
  - Due diligence
  - Quantifying harm : loss of job

22

Sorrell v. IMS Health, Inc. : 6 to 3
Privacy of doctor vs right of speech (marketing)

- (2011), was a case in which the Supreme Court of the United States held that a Vermont statute that restricted the sale, disclosure, and use of records that revealed the prescribing practices of individual doctors violated the First Amendment.
- Prescription Confidentiality Law (2007)
  - Vermont, Maine, NH + 25 states
  - Vermont Medical Society resolution stating that using the prescribing history of doctors in marketing was an intrusion into the way doctors practice medicine
  - Vermont claimed that the law was necessary to protect medical privacy and achieve improved public healthcare
- Data mining companies and pharmaceutical manufactures
  - the law violated their First Amendment rights by restricting the speech of the companies without adequate justification

23

Voter Registry

- Why Voter Registration Data is Available
  - According to North Carolina law (General Statute 132), "The public records and public information compiled by the agencies of North Carolina Government or its subdivisions are the property of the people. Therefore, it is the policy of this State that the people may obtain copies of their public records and public information free or at minimal cost unless otherwise specifically provided by law." (Voter registration records are not exempt from this law.)

24

Build a social consensus around privacy expectations for doing research

- Easier than business, surveillance
- IRB approvals based on (Belmont Report)
  - Benefit to society
  - Risk of harm
    - **To individuals: privacy violation & inaccurate results**
    - To society : inaccurate analysis resulting in wrong decisions
- Important to build trust in certain type of research
  - The public is willing to let doctors give them drugs blindly trusting that potential harms have been properly assessed to be safe
    - Public needs to be able to trust analytics will not harm
    - Need a better framework for assessing potential harm in IRB
  - Benefit is clear : find cure for cancer!
    - Benefit needs to be better emphasized

25

IRB: Risk of privacy violation

- Risk of attribute disclosure
  - Group disclosure
  - Linkage attack using auxiliary information
- Risk of identity disclosure
- Given?
  - Kinds of data elements used in the study
    - Name/dob/cancer status/ etc… (are there $$)
  - What system the data resides in : HW/SW
    - Risk of outsiders intruding / insider attack / negligence
  - What can users do with the data on the system
    - Take data off / look at everything / only do limited queries

26

Protection beyond disclosure … Actual Real Harm

- Identity theft
- Typically requires decisions & actions
  - To deny insurance, not hire someone
- There are costs to making wrong decisions
  - Lost business for no reason
- Harm requires fairly high level of confidence in correctness of information
  - Introducing enough uncertainties will prevent from actual harm due to making the cost of decision/action high

27

Agenda

- Concepts
- HIPAA
- Privacy: social & legal concepts
- Privacy: theoretical concepts
  - Differential privacy
- Approaches to Privacy

28

Myths and Fallacies of PII
Narayanan and Shmatikov (2010)

- "There is no silver bullet to privacy preserving computation"
- Developing an effective model for privacy-preserving linkage on sensitive data for research requires
  - a well orchestrated system
  - with strong fine grained access control,
  - regular privacy audits,
  - and good IRB approval guidelines

29

Data Privacy: Theoretical perspective

- Although several privacy and security challenges arise from unauthorized access or malicious dissemination of data
- The results of valid data analyses can also lead to the disclosure of sensitive information about individuals, and thus a confidentiality breach.
- There is a fine line between an adversary's ability to infer sensitive attributes of an individual and a researcher's ability to learn trends in the population.
- Hence, mathematically formulating what it means for some data analysis to not breach the privacy of individuals is a challenging task. Understanding these risks well is especially important for data released as open access or as monitored access in the four-level model discussed later.
- Disclosure limitation methods
  - Use statistics to limit disclosed information

30

Disclosure through multiple queries Background Knowledge Attack

- Another challenge in private data analyses is that even if one result does not disclose sensitive information about any individual, a collection of these tasks could potentially lead to a breach.
- For instance, consider two queries: number of unemployed males in Durham, and the number of males in Durham other than Bob who are unemployed.
- While Bob's employment status is not disclosed by either query in isolation, it can be inferred by combining the answers to both queries.
- Recent work has shown that many supposedly safe methods of releasing data can lead to disclosure of individual information by combining multiple invocations of these algorithms

31

Example

- Suppose a malicious user (often termed an adversary) wants to find whether Chandler has diabetes or not.
- As a side information he knows in which row of the database Chandler resides.
- Now suppose the adversary is only allowed to use a particular form of query, Q(i) which returns the partial sum of first i rows of column X in the database.
- In order to find Chandler's diabetes status the adversary simply executes Q(3)-Q(2).
- One striking feature this example highlights is: individual information can be compromised even without explicitly querying for the specific individual information.

Name	Diabetes
Lisa	1 (Y)
Sam	1 (Y)
Chandler	0 (N)

32

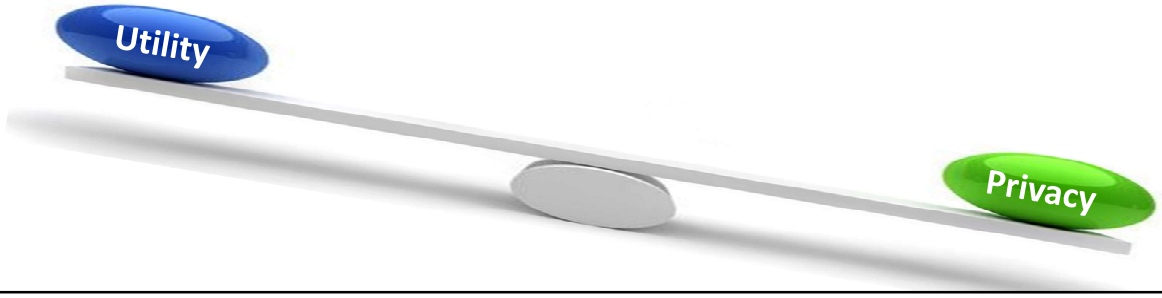
Information Privacy 101: Privacy is a BUDGET constrained problem

- Differential privacy literature proves each query leads to some privacy loss while providing some utility in terms of data analysis
- Current protection mechanism in database research is not effective
  - de-identified data cannot be linked
  - Not sharing enough details: leads to bias, and invalid results
- The goal is to achieve the maximum utility under a fixed privacy budget

33

Too Focused on Privacy

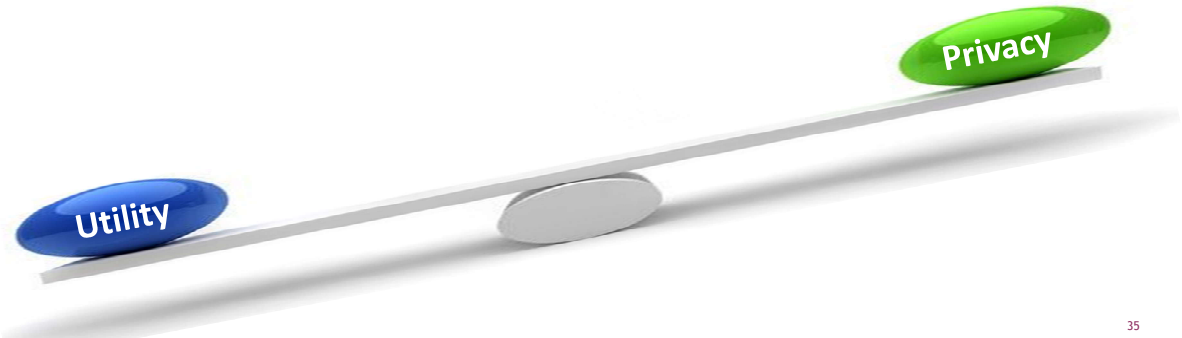
- Not enough information to make good linkage decisions
  - Consequences 1: incorrectly link different people
  - Consequences 2: missing linking same people
- Ultimately: research results are not correct



34

Too Focused on Utility

- Unnecessarily exposure, risk



35

35

Differential Privacy

- Differential privacy is a methodology that lets us concretely reason about privacy-budgeted data analysis
- Differential Privacy is a condition on the release mechanism (algorithm) and not on the dataset
- This means that for any two datasets which are close to one another (that is, which differ on a single element) a given differentially private algorithm will behave approximately the same on both data sets.
- The definition gives a strong guarantee that presence or absence of an individual will not affect the final output of the query significantly.
  - Roughly same answer regardless of 1 record

36

Differential Privacy

- An algorithm, A, for releasing data satisfies differential privacy if,
  - for any two datasets D1 and D2 that differ in one row,
  - the ratio of the likelihood of the algorithm resulting in the same output starting from D1 and D2 is bounded by at most e^{ε}.
  - Thus, if each row in a database corresponds to an individual, then using a differentially private algorithm provably ensures that the output is not sensitive to an arbitrary change in any one individual's input.
  - Differential privacy is powerful because it can be composed—two algorithms that satisfy differential privacy with parameters $\varepsilon 1$ and $\varepsilon 2$ results in $(\varepsilon 1 + \varepsilon 2)$ differential privacy, thus allowing us to apportion a total privacy budget of ε across many subtasks.
  - Differential privacy can allow accurate analyses in certain cases.
  - For instance, one of the LEHD data products boasts of provable differentially private protection in the released data. (http://onthemap.ces.census.gov/).
  - How much noise and what kind of noise?

37

Agenda

- Concepts
- HIPAA
- Privacy: social & legal concepts
- Privacy: theoretical concepts
  - Differential privacy
- Approaches to Privacy

38

Information Privacy 101: Information Accountability (Transparency) Works

- Secrecy : Hiding information does not support legitimate use
  - In reality, has limited power to protect privacy
  - Severe Consequences related to
    - Accuracy of data and decisions, use of data for
    - legitimate reasons, transparency & democracy
- Information Accountability support effective use (Credit Report)
  - Very clear transparency in the use of the data
  - Disclosure : Declared in writing, so when something goes wrong the right people are held accountable (data use agreements)
  - IT WORKS! Primary method used to protect financial data
  - Internet : crowdsourced auditing (public access IRB)
  - Logs & audits : what to log, how to keep tamperproof log

- D.J. Weitzner et al., Information Accountability, Comm. ACM, vol. 51, no. 6, [illegible], pp. 82–87.

39

39

Privacy by design
A trusted third party computer system

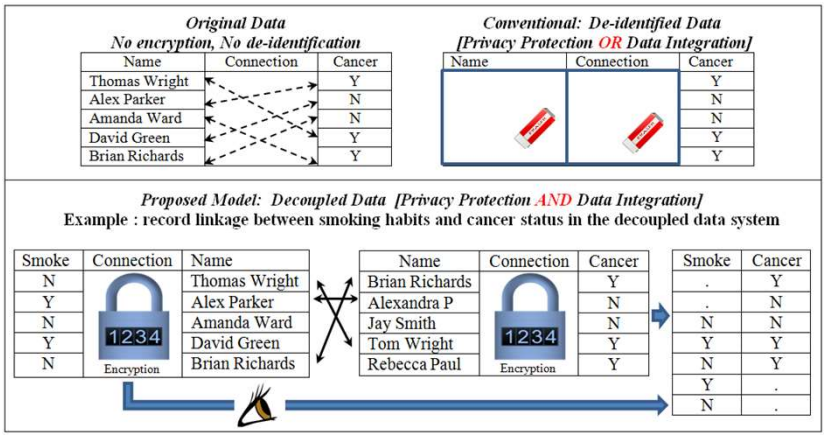
POPULATION INFORMATICS

- "At least three tools would enhance the infrastructure. First, an automated honest broker approach would allow researchers to request access to data and to perform integrated analyses on multiple data sets housed by different providers."
  - Standards and Infrastructure for Innovation Data Exchange Laurel L. Haak, David Baker, Donna K. Ginther, Gregg J. Gordon, Matthew A. Probus, Nirmala Kannankutty, and Bruce A. Weinberg Science 12 October 2012: 338 (6104), 196-197.

40

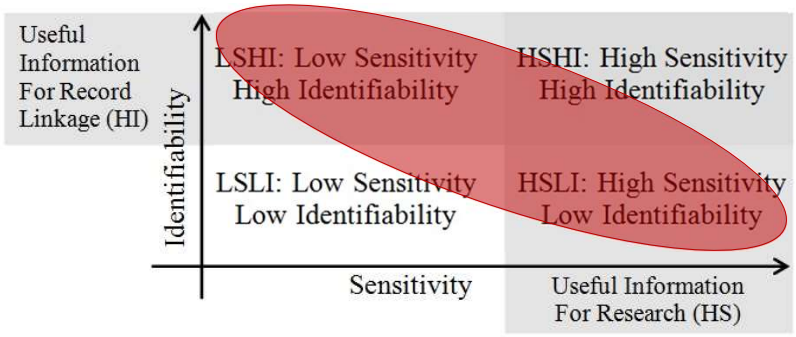
Decoupled Data :
Identity disclosure w/o attribute disclosure

POPULATION INFORMATICS



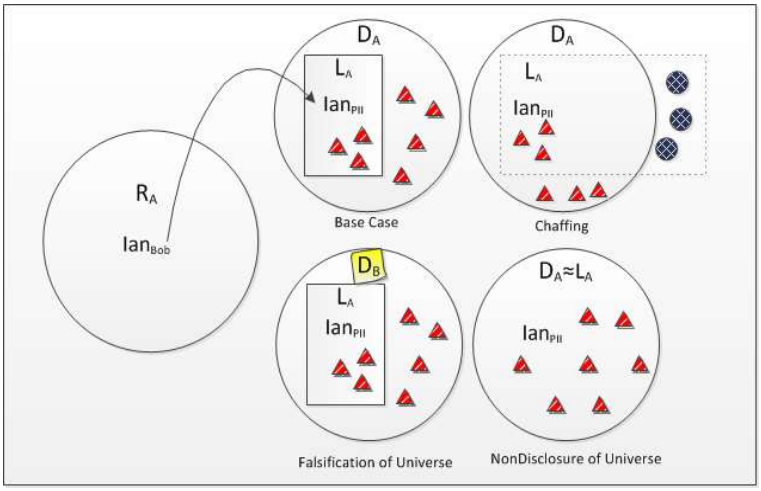
41

Identifiability and Sensitivity of Information



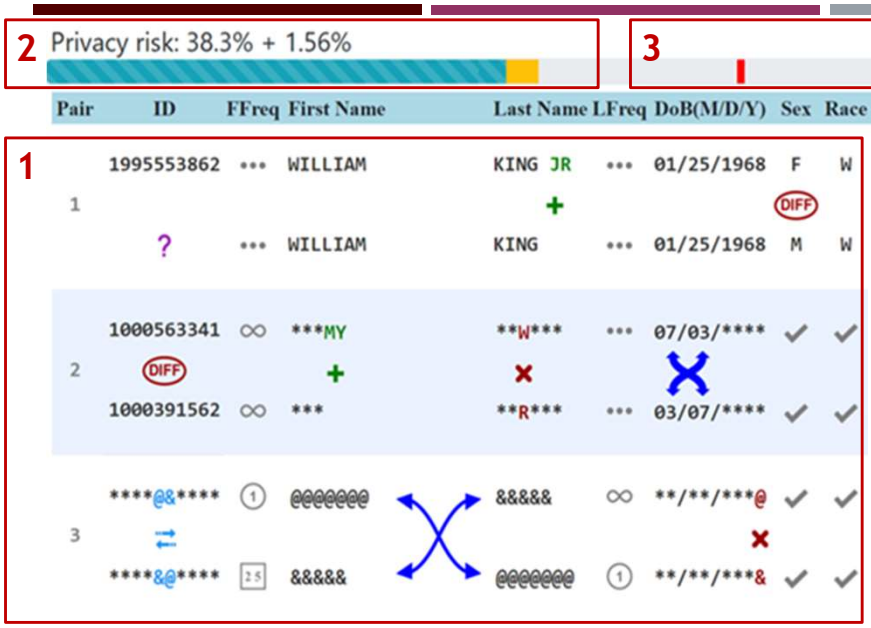
42

Chaffing & Universe Manipulation
Protection against group disclosure



43

Three Design Elements for Implementing the Minimum Necessary Standard



Our Proposed Key Design Elements

1. Minimum Disclosure via Interactive Just-in-Time Interface
   - Hide data values (when possible)
   - Add visual meta-data to help decision making without seeing raw data
2. Accountability via Quantified Privacy Risk
3. Limiting Privacy Risk via Budget

For more details check out

https://pinformatics.org/ppirl/faq/faq.htm

44

44

Aims 1 & 2: Real Question

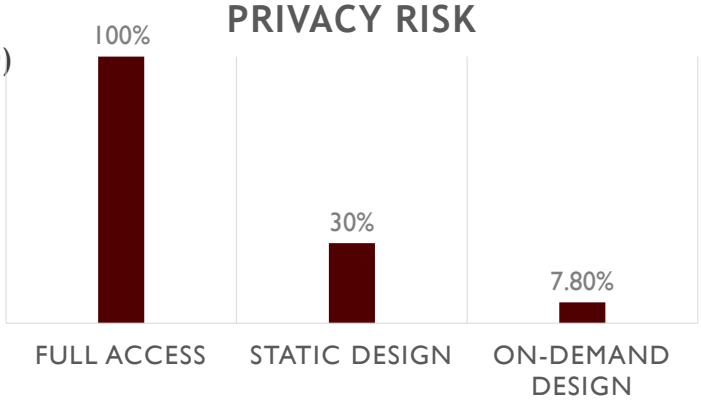
- Can we find the "sweet spot" between accessing PII for legitimate use while providing the maximum privacy protection as possible through the privacy by design approach by
- Large scale studies (N>100)

YES!!

Privacy by Design Works

Significantly improved privacy for same quality of results

no extra time



45

45

Aims 1 & 2: Expert Study Results

Compared to Full access to PII

- Five of the experts normally conducted record linkage with full access to PII
- They perceived that this system
  - offered more privacy protection
  - with little to no impact on accuracy in the linkage
  - but may take more time
- Evidence for improving linkage (i.e., more consistent linkage decisions) by providing better processed information for decision making in place of raw data

"**Once I got used to the coding, allowing partial disclosure helped in decision making**"

Compared to Encryption Based No Access to PII

- One expert had prior experience using encryption-based methods of data hiding for private record linkage with no access to PII.
- Compared to the encryption-based method, this participant perceived our system
  - to have less protection
  - and require more time
  - but to also allow for much better accuracy

"**I never know how well the hashing worked, or how accurate it is. It would be helpful to use this method to spot check a random sample (e.g., 5%)**"

- This seems to agree with our goal of providing a level of access between the all or nothing that provides better accuracy than no access, but more protection than full access.

46

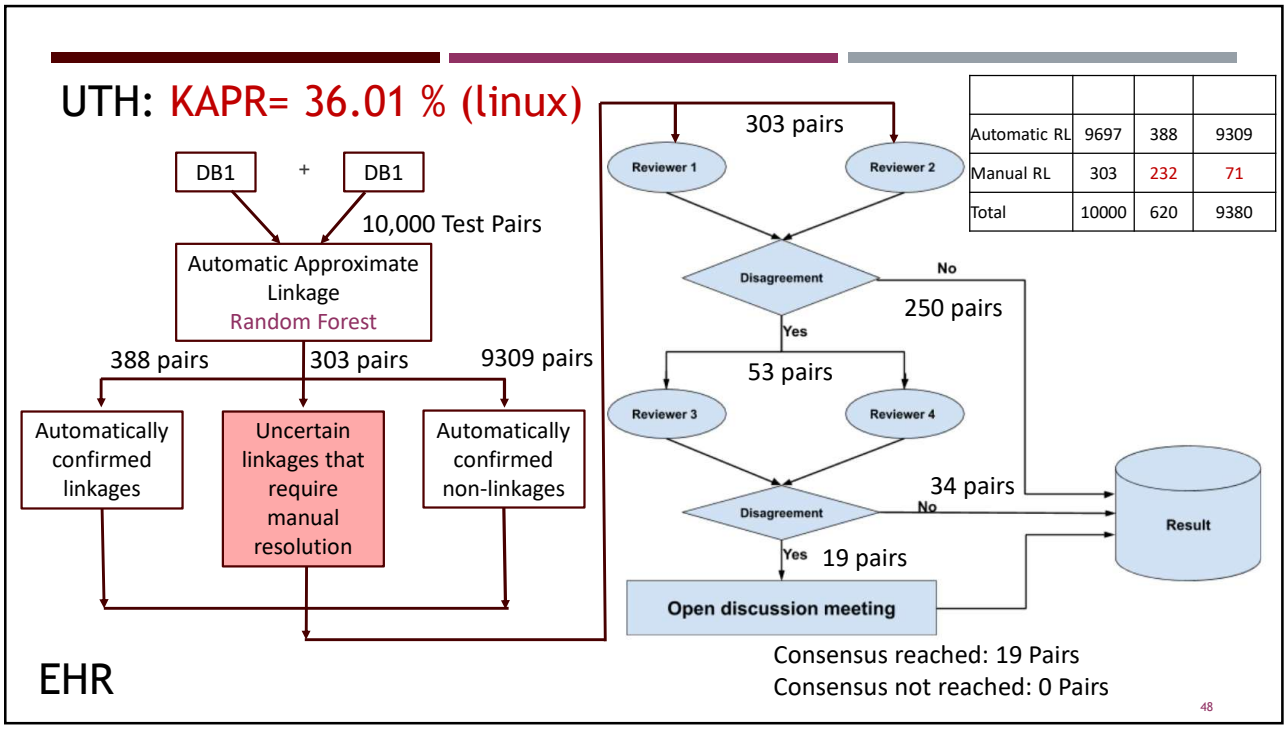
46

Aims 1 & 2: Highlights
On-Demand & Just-in-Time Interface Model

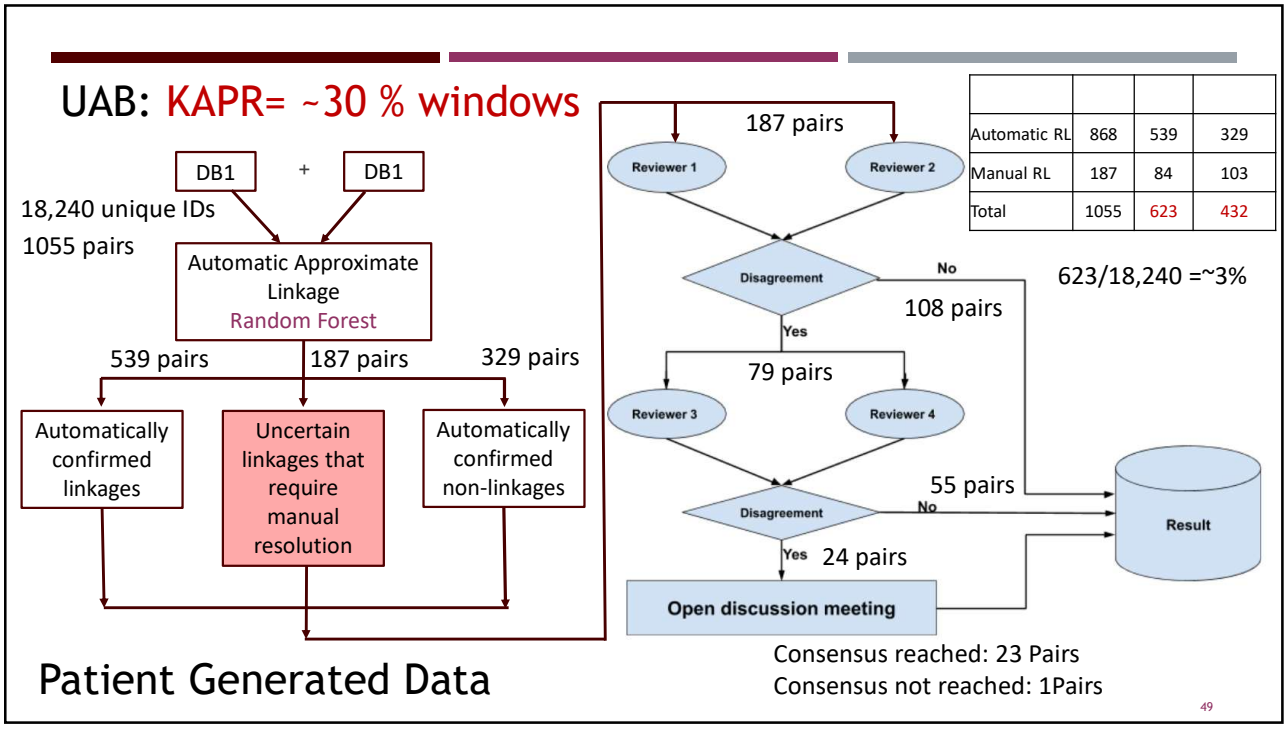
- User Study
  - On-demand model to **satisfy minimum-necessary legal requirement** (e.g., GDPR, HIPAA)
  - On-demand interface **reduced privacy risk to 7.85% compared to 100% when all data is disclosed with little impact on decision quality or completion time**
  - **To have high quality results, you must have sufficient budget**: The error results indicate that the quality of human decisions will suffer if low disclosure limits are enforced
- Expert Study: Positive reactions from experts in intended user population
  - **Evidence for improving linkage** (i.e., more consistent linkage decisions) by providing better processed information for decision making in place of raw data
  - **Potential to validate results when used in conjunction with encryption based no access methods**
- Future Works
  - Need to refine privacy risk score
  - Need to refine design considerations for possible time costs

47

47


UTH: KAPR= 36.01 % (linux)
DB1
+
DB1
10,000 Test Pairs
Automatic Approximate Linkage
Random Forest
388 pairs
303 pairs
9309 pairs
Automatically confirmed linkages
Uncertain linkages that require manual resolution
Automatically confirmed non-linkages
303 pairs
Reviewer 1
Reviewer 2
Disagreement
No
250 pairs
Yes
53 pairs
Reviewer 3
Reviewer 4
Disagreement
34 pairs
No
Yes
19 pairs
Open discussion meeting
Result
Automatic RL 9697 388 9309
Manual RL 303 232 71
Total 10000 620 9380
EHR
Consensus reached: 19 Pairs
Consensus not reached: 0 Pairs
48

48


UAB: KAPR= ~30 % windows
DB1
+
DB1
18,240 unique IDs
1055 pairs
Automatic Approximate Linkage
Random Forest
539 pairs
187 pairs
329 pairs
Automatically confirmed linkages
Uncertain linkages that require manual resolution
Automatically confirmed non-linkages
187 pairs
Reviewer 1
Reviewer 2
Disagreement
No
108 pairs
Yes
79 pairs
Reviewer 3
Reviewer 4
Disagreement
55 pairs
No
Yes
24 pairs
Open discussion meeting
Result
Automatic RL 868 539 329
Manual RL 187 84 103
Total 1055 623 432
623/18,240 =~3%
Patient Generated Data
Consensus reached: 23 Pairs
Consensus not reached: 1Pairs
49

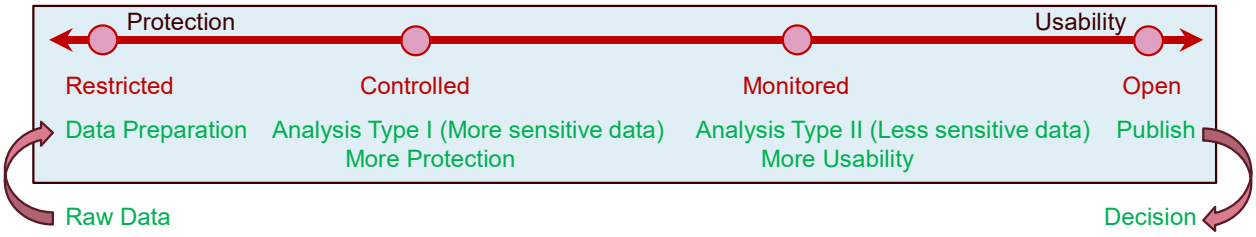
49

Secure Workflow

Safe Platform for Data to Decision

50

Use Published Data for Good Decision Making



Deployed together the four data access models can provide a comprehensive system for privacy protection, balancing the risk and usability of secondary data in population informatics research

51

The start …

- Write up a research plan on
  - What data you need
  - What you want to do with them
  - Determine access levels for each data
- Submit to IRB process

52

IRB: Risk of privacy violation vs. Benefit to Society

- Risk of attribute disclosure
  - Group disclosure
  - Linkage attack using auxiliary information
- Risk of identity disclosure
- Given?
  - Kinds of data elements used in the study
    - Name/dob/cancer status/ etc… (are there)
  - What system the data resides in : HW/SW
    - Risk of outsiders intruding / insider attack / negligence
  - What can users do with the data on the system
    - Take data off / look at everything / only do limited queries

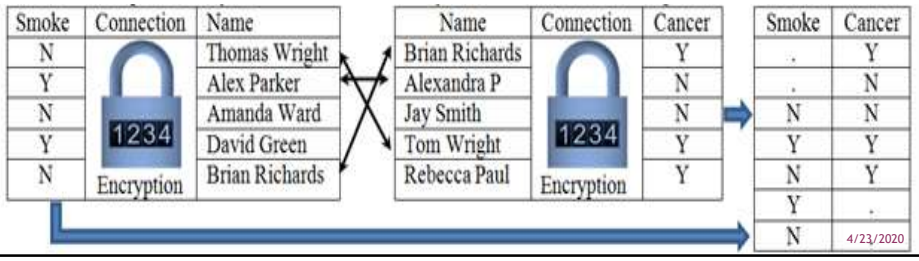
53

Restricted Access : Prepare the customized data

- Decoupled Data (Kum 2012)
  - Automated Honest Broker SW
- Sample selection
- Attribute selection
- Data integration (access to PII)
- Some data cleaning
- Full IRB
- Example: RDC (TX census RDC)

Privacy Preserving Interactive Record Linkage

- Decouple data via encryption
- Automated honest broker approach via computerized third party model
- Chaffe to prevent group disclosure
- Kum, H.C., Krishnamurthy A., Machanavajjhala A., Reiter M., and Ahalt S. Privacy Preserving Interactive Record Linkage (PPIRL). J Am Med Inform. Assoc. 2014;21:212-220. doi:10.1136/amiajno-2013-002165



Controlled Access : Model using given tools

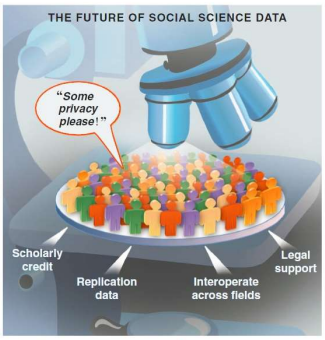


Fig. 1. New types of research data about human behavior and society pose many opportunities if crucial infrastructural challenges are tackled.

Gary King. Ensuring the Data-Rich Future of the Social Sciences, Science, vol 331, 2011, pp 719-721.

- With approved deidentified data
- Locked down VM: customized appliances
- only approved software
- Remote access via VPN
- Very effective for threats from HBC
- Full IRB
- U Chicago-NORC , UNC-Tracs (CTSA), UCSD-iDASH, SAIL

Monitored Access : Freely Repurpose

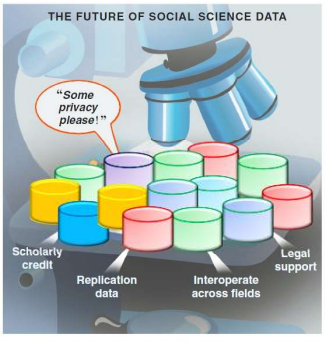
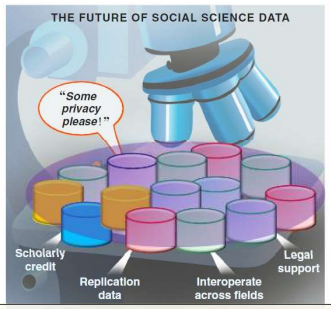


Fig. 1. New types of research data about human behavior and society pose many opportunities if crucial infrastructural challenges are tackled.

Gary King. Ensuring the Data-Rich Future of the Social Sciences, Science, vol 331, 2011, pp 719-721.

- Information Accountability model
- Exempt IRB: Explicit data use agreement (5 big Q)
  - Public online (crowdsource)
- Any software & auxiliary data
- Remote Access via VPN
- Less sensitive data(e.g. Aggregate data)
- SHRINE, Secure Unix servers

Open Access : No restriction on use



Package with filter (disclosure limitation methods) & take out of lab

Gary King. Ensuring the Data-Rich Future of the Social Sciences, Science, vol 331, 2011, pp 719-721.

- Anyone : Publish information for others
- No IRB
- No monitoring use
- Publish data use terms
- Disclosure Limitation Methods (filter)
  - Be careful of incorrect use
- Sanitized data
- Public websites, publications

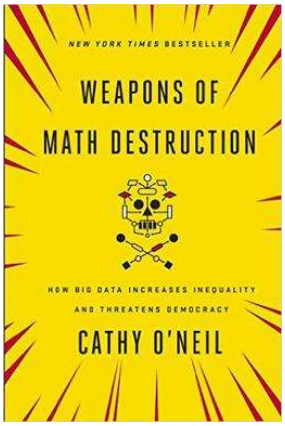
MAJOR CHALLENGE
INFORMATION PRIVACY & DATA GOVERNANCE

FAIRNESS IN ALGORITHMS

Fairness in blackbox algorithms
Criteria for evaluation?

POPULATION INFORMATICS

- Some algorithms reinforce discrimination that exist in our real world
- She posits that these problematic mathematical tools
  - Are opaque
  - Unregulated and difficult to contest
  - And scalable
- Amplify any inherent biases to affect increasingly larger populations



Fairness in blackbox algorithms

POPULATION INFORMATICS

- Examples
  - Google Photos mistakes in labeling
  - Facebook requires extra work for some native Americans to get an account
- Take away:
  - Human critical thinking and judgment is very important to using algorithms appropriately
  - There must be humans who will take on the responsibility for the decision

62

Information Privacy 101: Takeaway 1
Do you want to drive on the information highway?

POPULATION INFORMATICS

- Be prepared to get into accidents, especially as we all figure out how to do this more safely.
- Invest in personnel to do your due diligence to stay up to date with the fast paced security technology and privacy regulations
- Invest in technologies (i.e., seat belts & airbags) that can minimize the real harm when accidents do occur

4/23/2020 63

63

Information Privacy 101: Takeaway 2
Privacy is a BUDGET constrained problem

POPULATION INFORMATICS

- Differential privacy literature proves each query leads to some privacy loss while providing some utility in terms of data analysis
- Current protection mechanism in database research is not effective
  - de-identified data cannot be linked
  - Not sharing enough details: leads to bias, and invalid results
- The goal is to achieve the maximum utility under a fixed privacy budget

64

64

Information Privacy 101: Takeaway 3
Information Accountability (Transparency) Works

POPULATION INFORMATICS

- Secrecy : Hiding information does not support legitimate use
  - In reality, has limited power to protect privacy
  - Severe Consequences related to
    - Accuracy of data and decisions, use of data for
    - legitimate reasons, transparency & democracy
- Information Accountability support effective use (Credit Report)
  - Very clear transparency in the use of the data
  - Disclosure : Declared in writing, so when something goes wrong the right people are held accountable (data use agreements)
  - IT WORKS! Primary method used to protect financial data
  - Internet : crowdsourced auditing (public access IRB)
  - Logs & audits : what to log, how to keep tamperproof log

- D.J. Weitzner et al., Information Accountability, Comm. ACM, vol. 51, no. 6, [illegible], pp. 82–87.

65

65