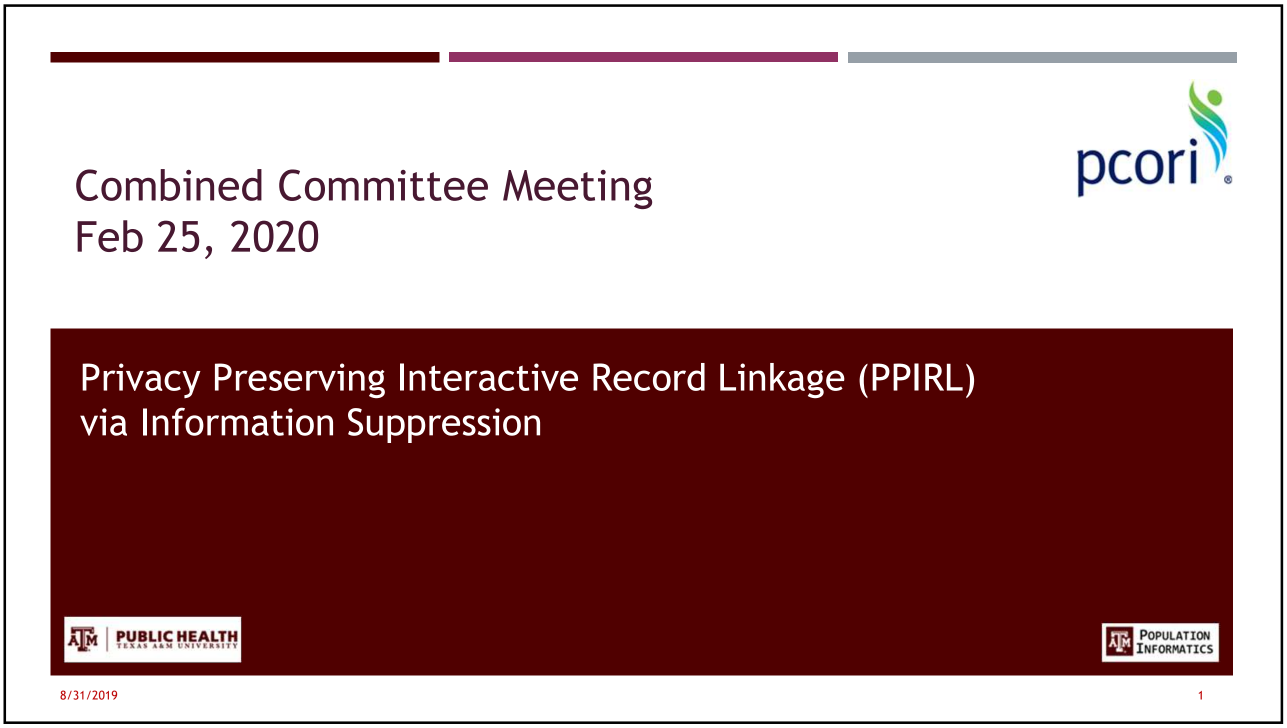
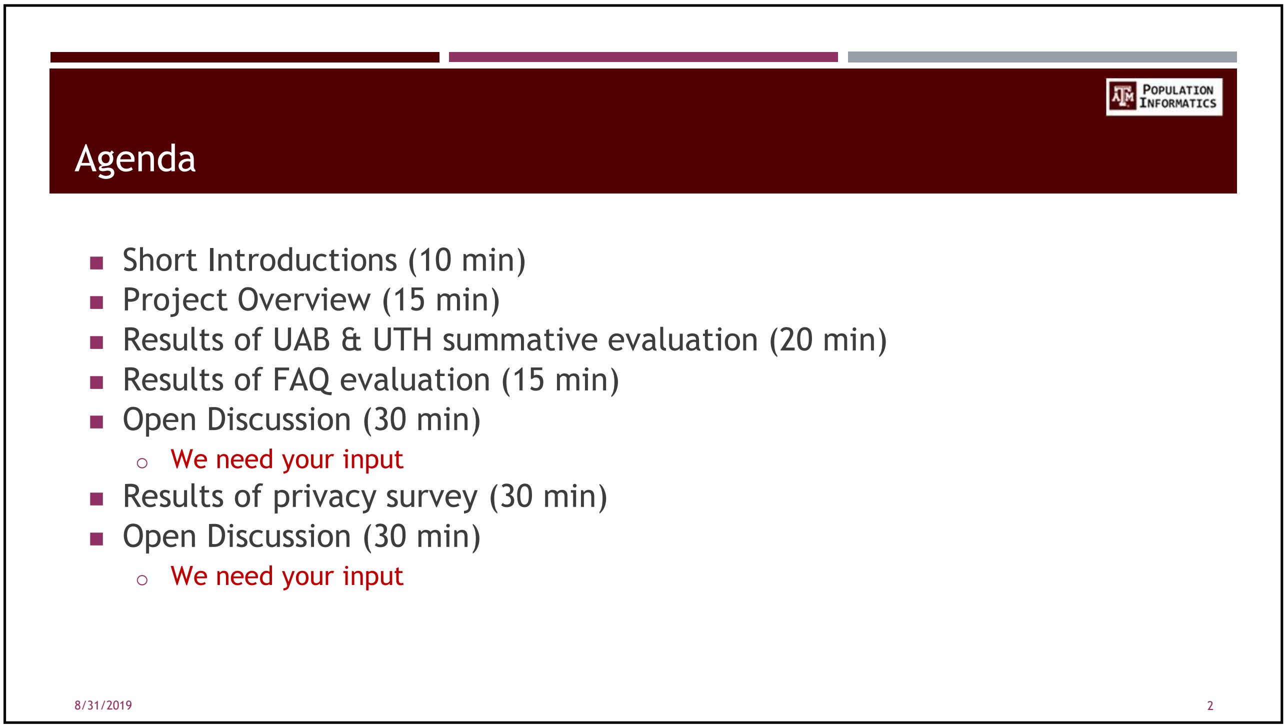


Combined Committee Meeting
Feb 25, 2020

Privacy Preserving Interactive Record Linkage (PPIRL) via Information Suppression

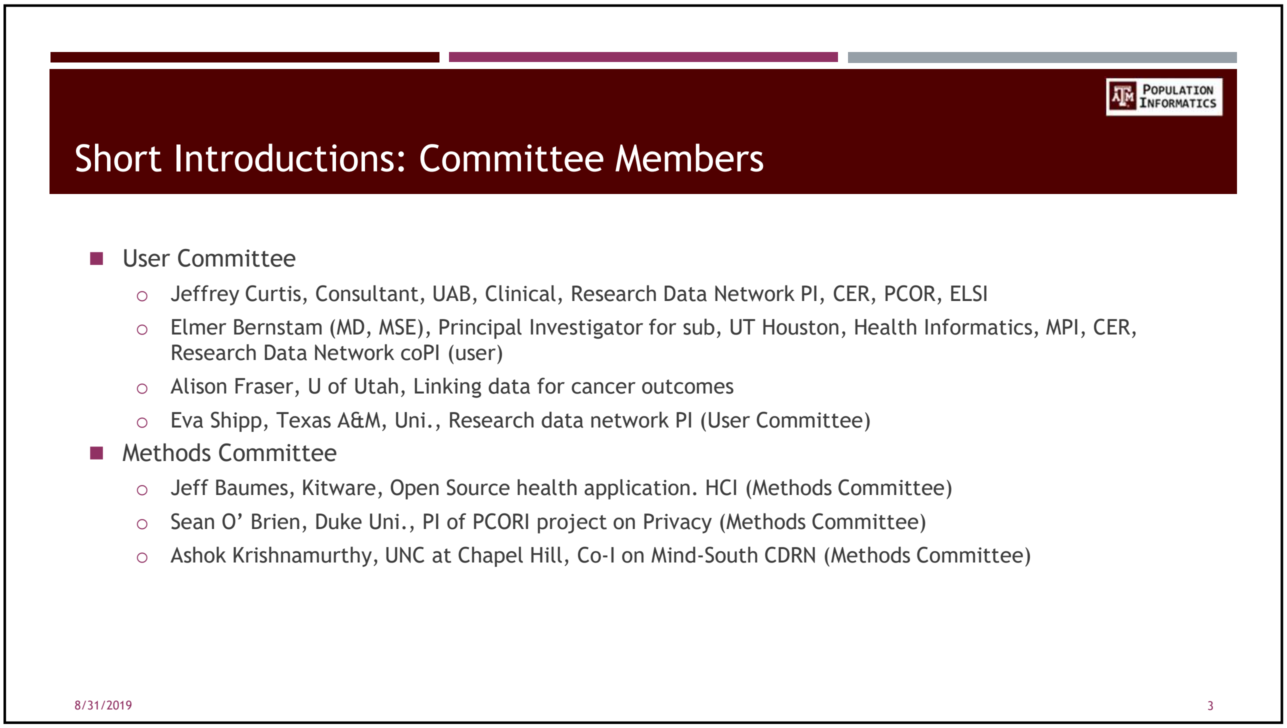
1



Agenda

- Short Introductions (10 min)
- Project Overview (15 min)
- Results of UAB & UTH summative evaluation (20 min)
- Results of FAQ evaluation (15 min)
- Open Discussion (30 min)
  - We need your input
- Results of privacy survey (30 min)
- Open Discussion (30 min)
  - We need your input

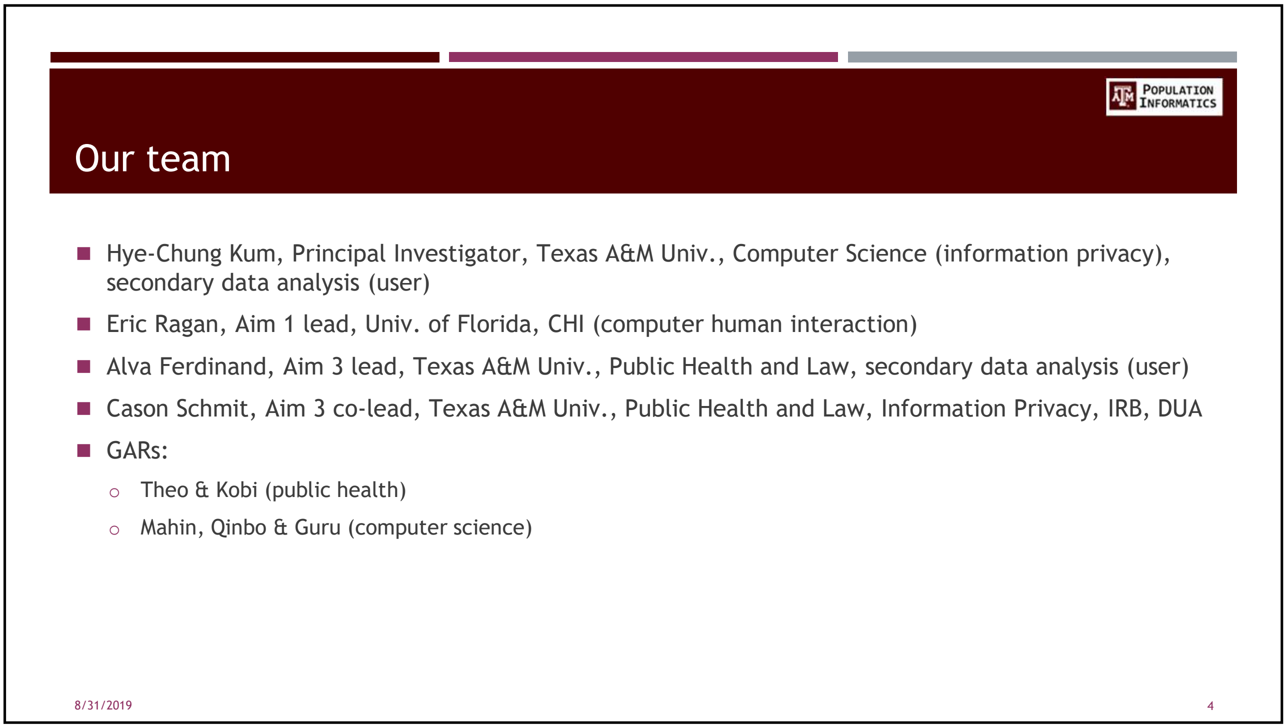
2



Short Introductions: Committee Members

- User Committee
  - Jeffrey Curtis, Consultant, UAB, Clinical, Research Data Network PI, CER, PCOR, ELSI
  - Elmer Bernstam (MD, MSE), Principal Investigator for sub, UT Houston, Health Informatics, MPI, CER, Research Data Network coPI (user)
  - Alison Fraser, U of Utah, Linking data for cancer outcomes
  - Eva Shipp, Texas A&M, Uni., Research data network PI (User Committee)
- Methods Committee
  - Jeff Baumes, Kitware, Open Source health application. HCI (Methods Committee)
  - Sean O' Brien, Duke Uni., PI of PCORI project on Privacy (Methods Committee)
  - Ashok Krishnamurthy, UNC at Chapel Hill, Co-I on Mind-South CDRN (Methods Committee)

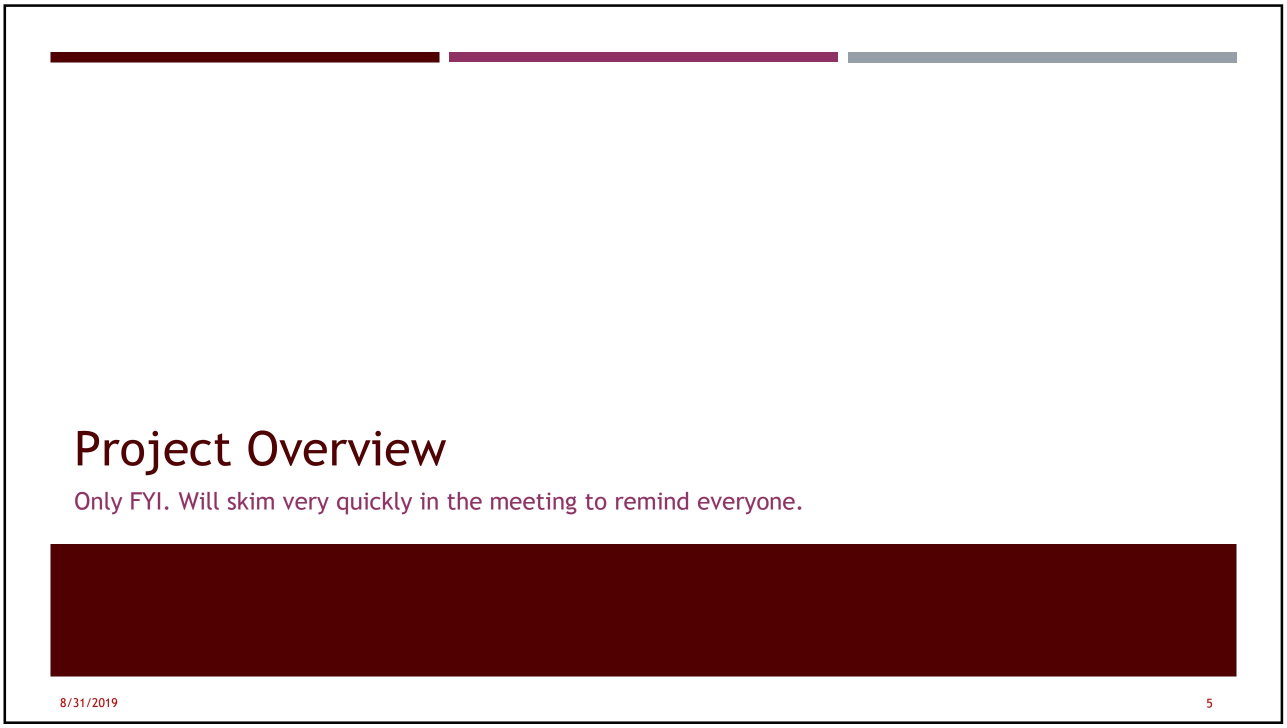
3



Our team

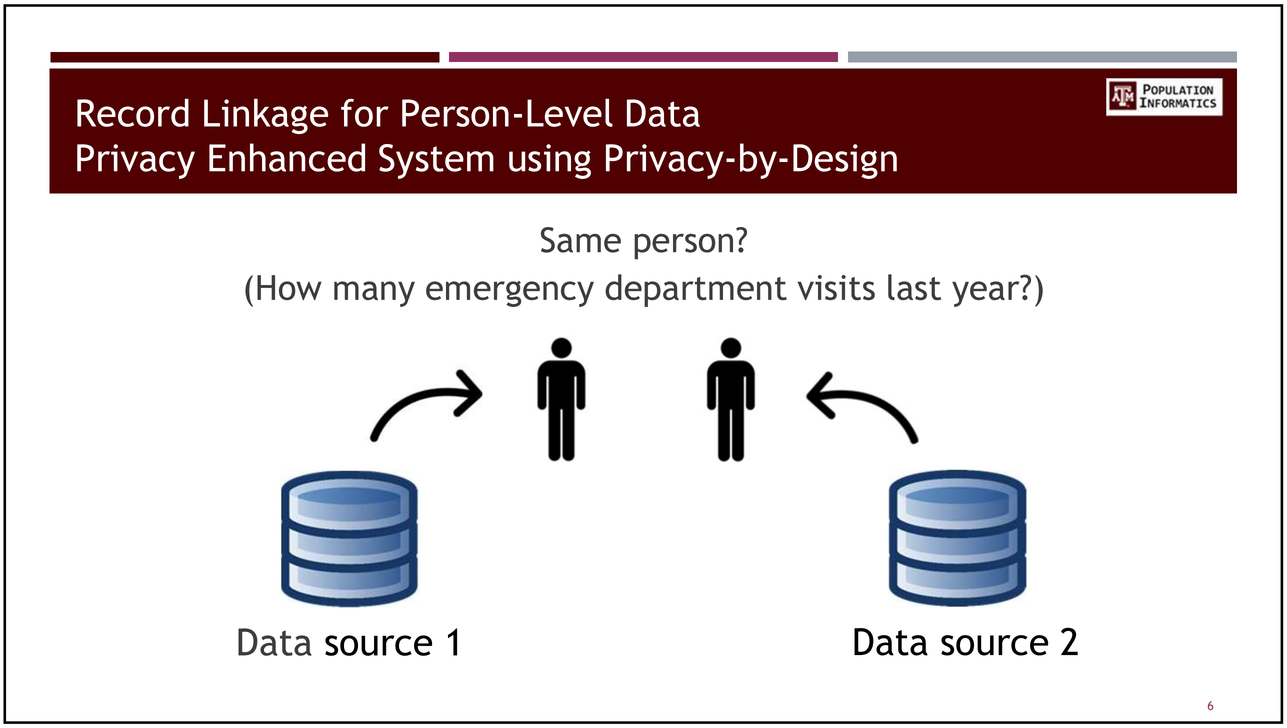
- Hye-Chung Kum, Principal Investigator, Texas A&M Univ., Computer Science (information privacy), secondary data analysis (user)
- Eric Ragan, Aim 1 lead, Univ. of Florida, CHI (computer human interaction)
- Alva Ferdinand, Aim 3 lead, Texas A&M Univ., Public Health and Law, secondary data analysis (user)
- Cason Schmit, Aim 3 co-lead, Texas A&M Univ., Public Health and Law, Information Privacy, IRB, DUA
- GARs:
  - Theo & Kobi (public health)
  - Mahin, Qinbo & Guru (computer science)

4



Project Overview

Only FYI. Will skim very quickly in the meeting to remind everyone.



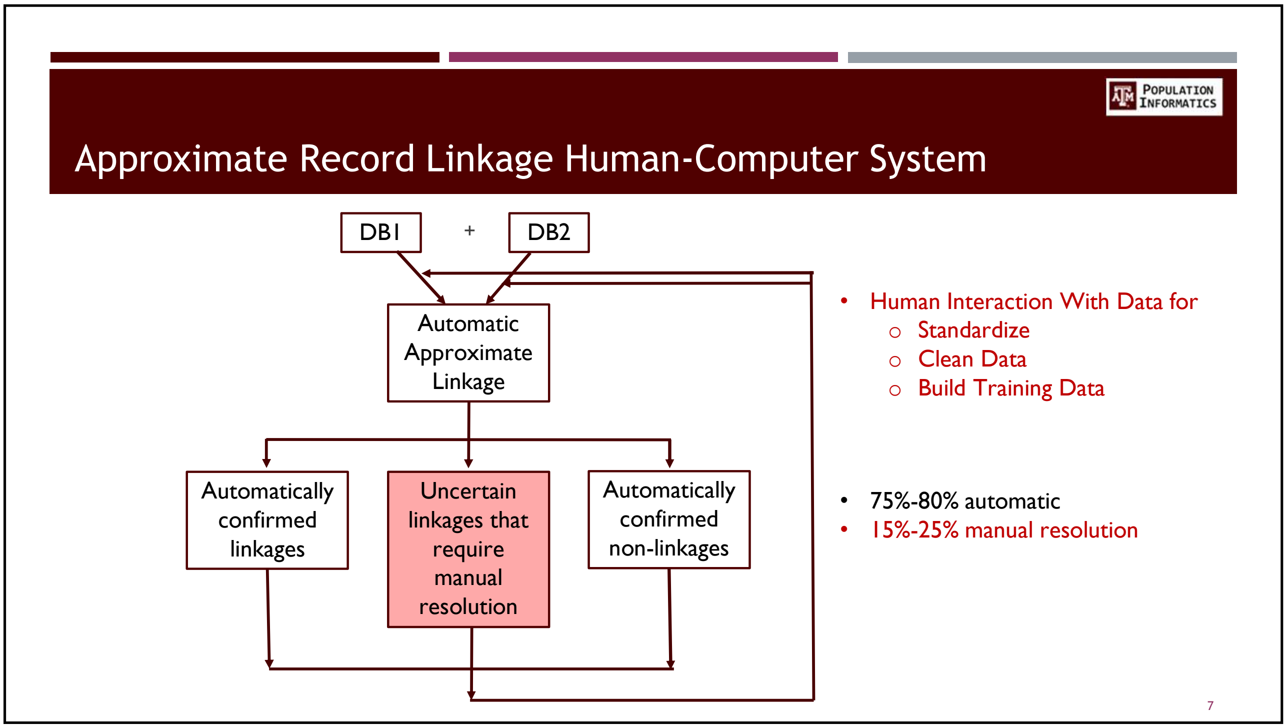
Record Linkage for Person-Level Data
Privacy Enhanced System using Privacy-by-Design

Same person?

(How many emergency department visits last year?)

Data source 1

Data source 2



Approximate Record Linkage Human-Computer System

DB1 + DB2

Automatic Approximate Linkage

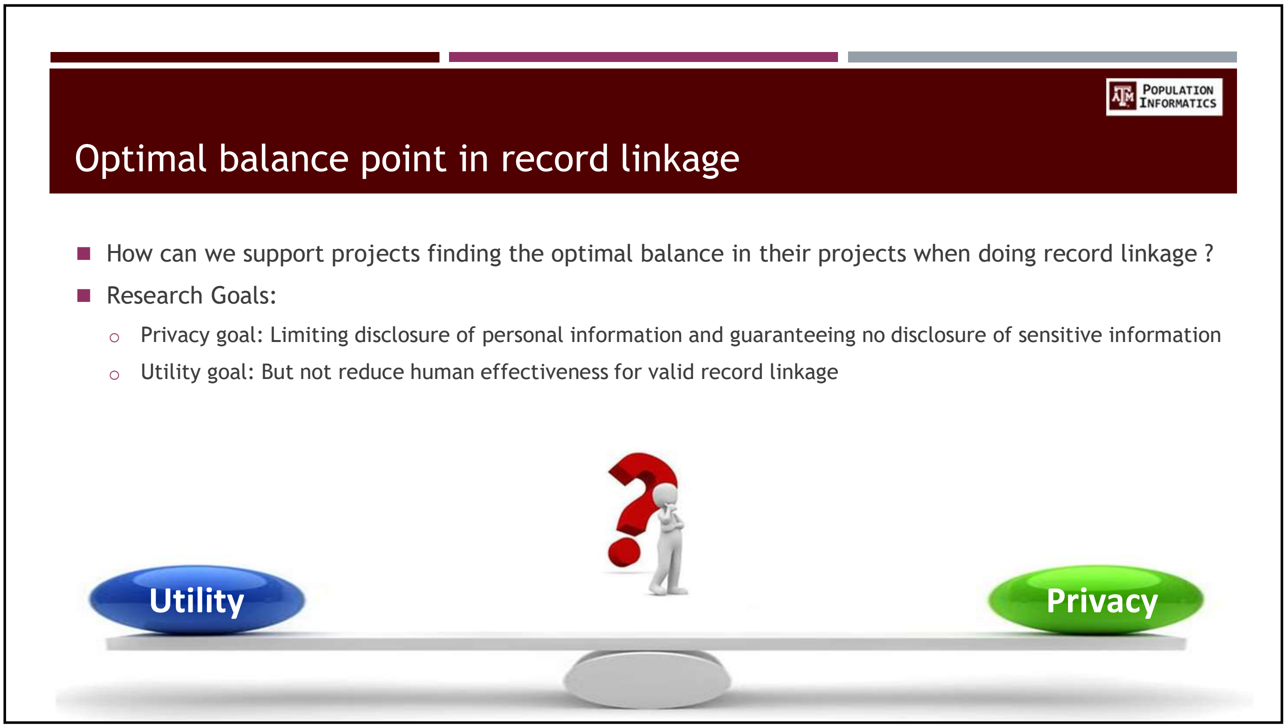
Automatically confirmed linkages

Uncertain linkages that require manual resolution

Automatically confirmed non-linkages

- Human Interaction With Data for
  - Standardize
  - Clean Data
  - Build Training Data

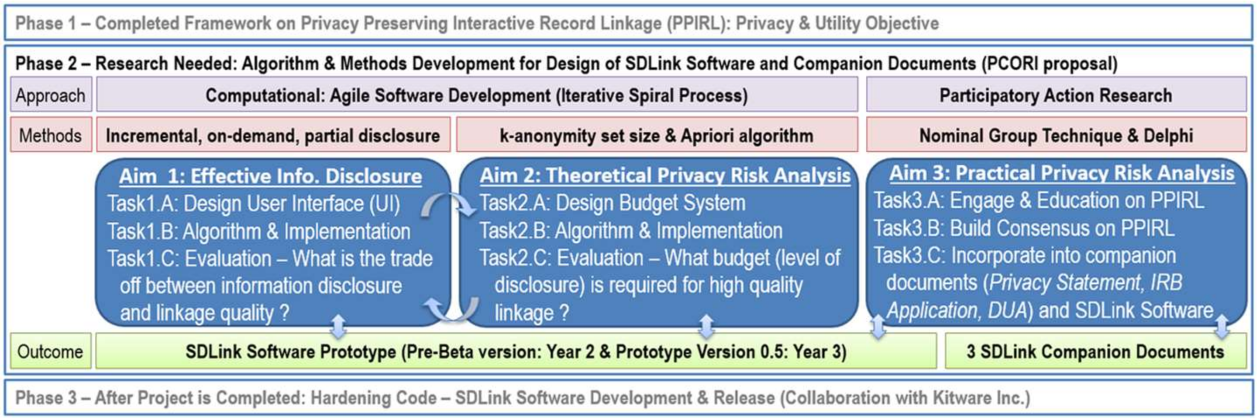
- 75%-80% automatic
- 15%-25% manual resolution



Optimal balance point in record linkage

- How can we support projects finding the optimal balance in their projects when doing record linkage ?
- Research Goals:
  - Privacy goal: Limiting disclosure of personal information and guaranteeing no disclosure of sensitive information
  - Utility goal: But not reduce human effectiveness for valid record linkage

Aims & Outcomes
Prototype software & companion documents



Phase 1 – Completed Framework on Privacy Preserving Interactive Record Linkage (PPIRL): Privacy & Utility Objective

Phase 2 – Research Needed: Algorithm & Methods Development for Design of SDLink Software and Companion Documents (PCORI proposal)

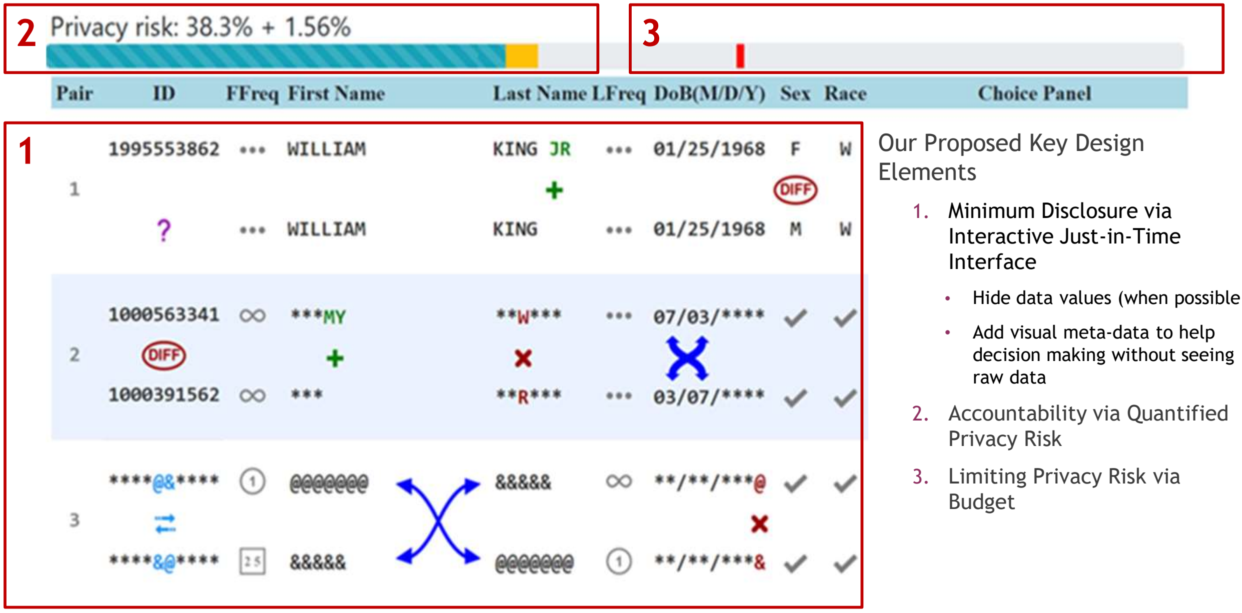
Approach	Computational: Agile Software Development (Iterative Spiral Process)		Participatory Action Research
Methods	Incremental, on-demand, partial disclosure	k-anonymity set size & Apriori algorithm	Nominal Group Technique & Delphi
	Aim 1: Effective Info. Disclosure Task1.A: Design User Interface (UI) Task1.B: Algorithm & Implementation Task1.C: Evaluation – What is the trade off between information disclosure and linkage quality ?	**Aim 2: Theoretical Privacy Risk Analysis** Task2.A: Design Budget System Task2.B: Algorithm & Implementation Task2.C: Evaluation – What budget (level of disclosure) is required for high quality linkage ?	**Aim 3: Practical Privacy Risk Analysis** Task3.A: Engage & Education on PPIRL Task3.B: Build Consensus on PPIRL Task3.C: Incorporate into companion documents (*Privacy Statement, IRB Application, DUA*) and SDLink Software
Outcome	SDLink Software Prototype (Pre-Beta version: Year 2 & Prototype Version 0.5: Year 3)		3 SDLink Companion Documents

Phase 3 – After Project is Completed: Hardening Code – SDLink Software Development & Release (Collaboration with Kitware Inc.)

9

9

Three Design Elements for Implementing the Minimum Necessary Standard



Privacy risk: 38.3% + 1.56%

Pair	ID	FFreq	First Name	Last Name	LFreq	DoB(M/D/Y)	Sex	Race	Choice Panel
1	1995553862	•••	WILLIAM	KING JR	•••	01/25/1968	F	W	
	?	•••	WILLIAM	KING	•••	01/25/1968	M	W	
2	1000563341	∞	***MY	**W***	•••	07/03/****	✓	✓	
	1000391562	∞	***	**R***	•••	03/07/****	✓	✓	
3	****@&****	1	@@@@@@@	&&&&&	∞	**/**/***@	✓	✓	
	****&@****	25	&&&&&	@@@@@@@	1	**/**/***&	✓	✓	

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

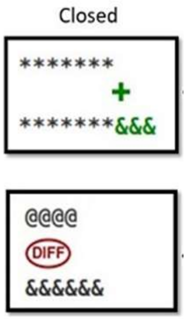
10

10

Our proposed approach 1: Interactive Interfaces Dynamic On-demand Incremental Disclosure

POPULATION INFORMATICS

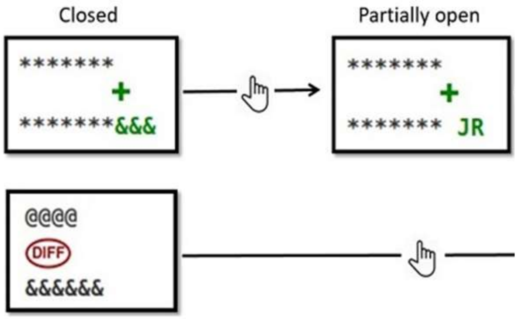
- Dynamic: Click to see more
- On-demand: When needed
  - Just-in-time decision
- Incremental: As needed
  - Not all at once
- Allow for easy accountability in information Use



Our proposed approach 1: Interactive Interfaces Dynamic On-demand Incremental Disclosure

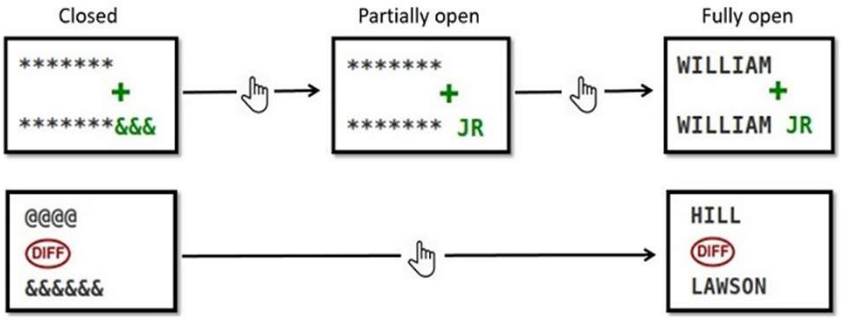
POPULATION INFORMATICS

- Dynamic: Click to see more
- On-demand: When needed
  - Just-in-time decision
- Incremental: As needed
  - Not all at once
- Allow for easy accountability in information Use



Our proposed approach 1: Interactive Interfaces Dynamic On-demand Incremental Disclosure

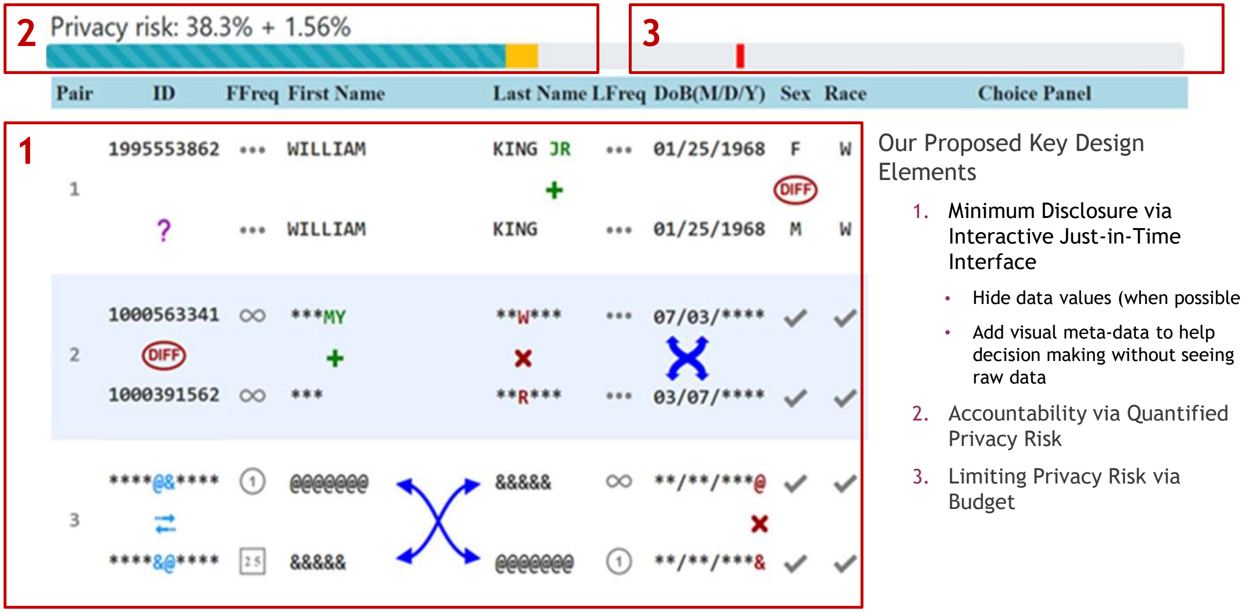
- Dynamic: Click to see more
- On-demand: When needed
  - Just-in-time decision
- Incremental: As needed
  - Not all at once
- Allow for easy accountability in information Use



13

13

Three Design Elements for Implementing the Minimum Necessary Standard



Privacy risk: 38.3% + 1.56%

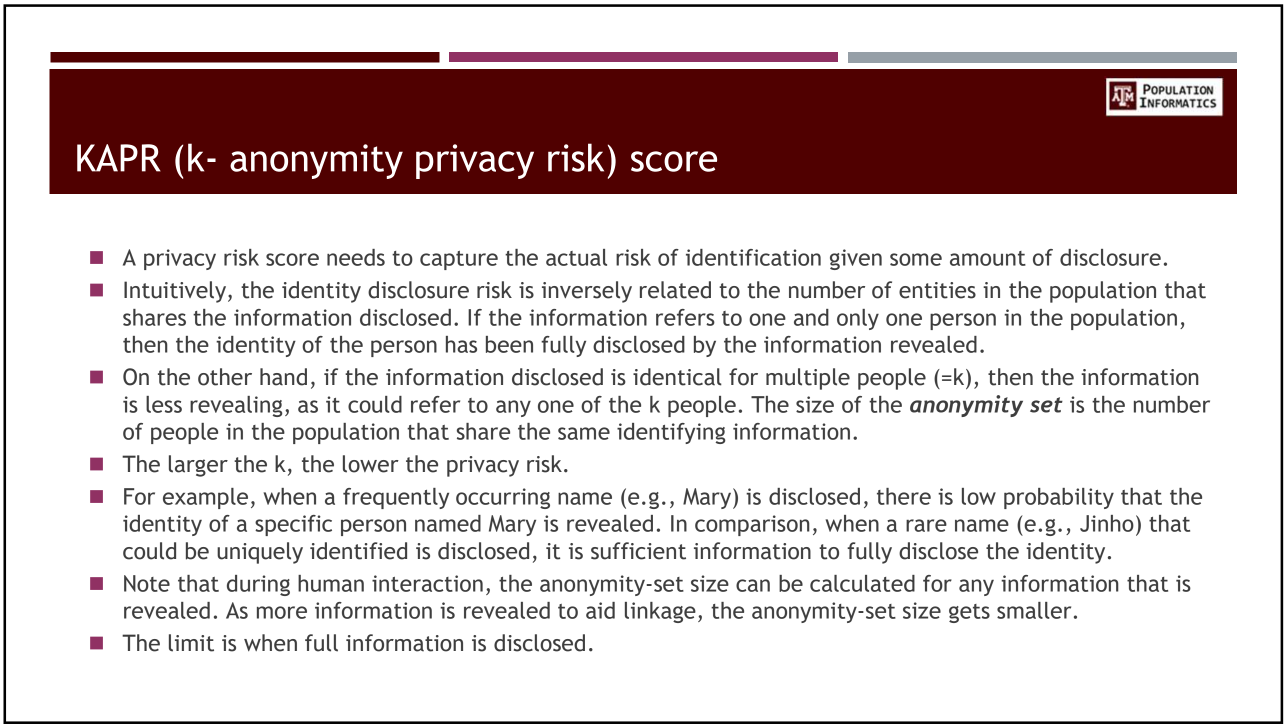
Pair	ID	FFreq	First Name	Last Name	LFreq	DoB(M/D/Y)	Sex	Race	Choice Panel
	1995553862	•••	WILLIAM	KING JR	•••	01/25/1968	F	W	
1				+			DIFF		
	?	•••	WILLIAM	KING	•••	01/25/1968	M	W	
	1000563341	∞	***MY	**W***	•••	07/03/****	✓	✓	
2	DIFF		+	✗		✗			
	1000391562	∞	***	**R***	•••	03/07/****	✓	✓	
	****@&****	①	@@@@@@@	&&&&&	∞	**/**/***@	✓	✓	
3	⇄					✗			
	****&@****	25	&&&&&	@@@@@@@	①	**/**/***&	✓	✓	

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

14

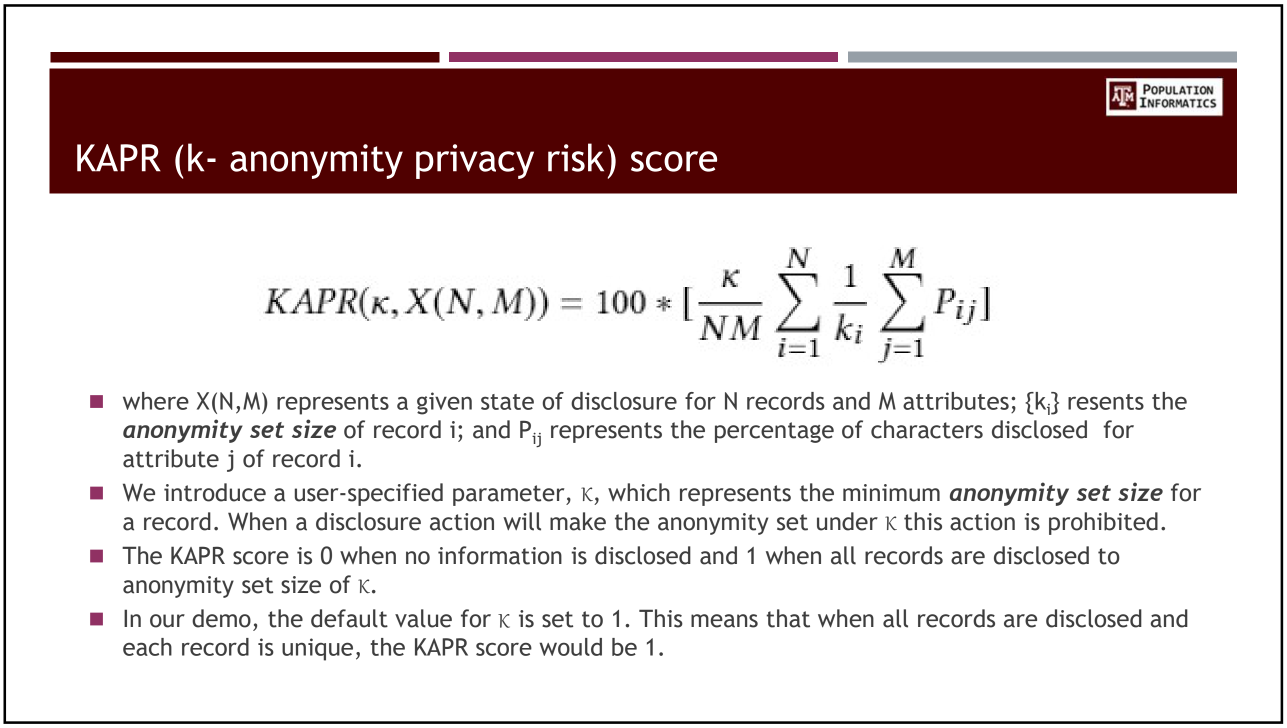
14



KAPR (k- anonymity privacy risk) score

- A privacy risk score needs to capture the actual risk of identification given some amount of disclosure.
- Intuitively, the identity disclosure risk is inversely related to the number of entities in the population that shares the information disclosed. If the information refers to one and only one person in the population, then the identity of the person has been fully disclosed by the information revealed.
- On the other hand, if the information disclosed is identical for multiple people (=k), then the information is less revealing, as it could refer to any one of the k people. The size of the ***anonymity set*** is the number of people in the population that share the same identifying information.
- The larger the k, the lower the privacy risk.
- For example, when a frequently occurring name (e.g., Mary) is disclosed, there is low probability that the identity of a specific person named Mary is revealed. In comparison, when a rare name (e.g., Jinho) that could be uniquely identified is disclosed, it is sufficient information to fully disclose the identity.
- Note that during human interaction, the anonymity-set size can be calculated for any information that is revealed. As more information is revealed to aid linkage, the anonymity-set size gets smaller.
- The limit is when full information is disclosed.

15

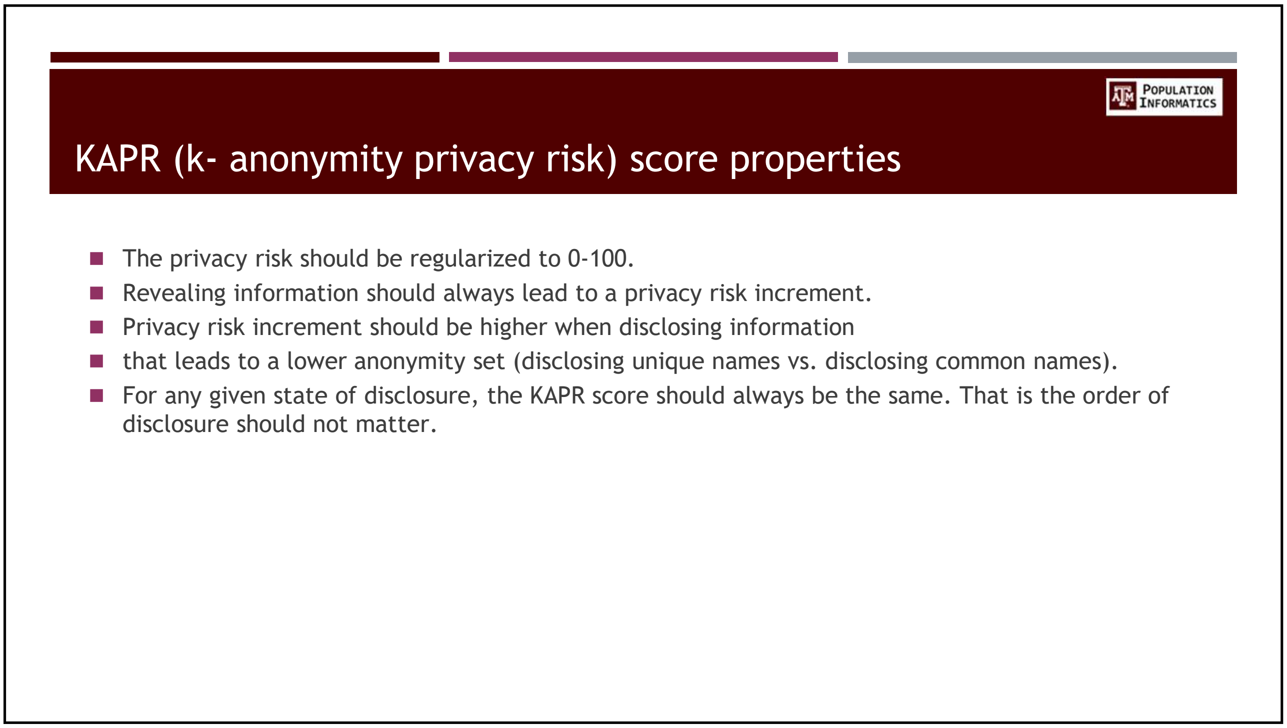


KAPR (k- anonymity privacy risk) score

$$KAPR(\kappa, X(N,M)) = 100 * [\frac{\kappa}{NM}\sum_{i=1}^{N}\frac{1}{k_i}\sum_{j=1}^{M}P_{ij}]$$

- where X(N,M) represents a given state of disclosure for N records and M attributes; $\{k_i\}$ resents the ***anonymity set size*** of record i; and P_{ij} represents the percentage of characters disclosed for attribute j of record i.
- We introduce a user-specified parameter, κ, which represents the minimum ***anonymity set size*** for a record. When a disclosure action will make the anonymity set under κ this action is prohibited.
- The KAPR score is 0 when no information is disclosed and 1 when all records are disclosed to anonymity set size of κ.
- In our demo, the default value for κ is set to 1. This means that when all records are disclosed and each record is unique, the KAPR score would be 1.

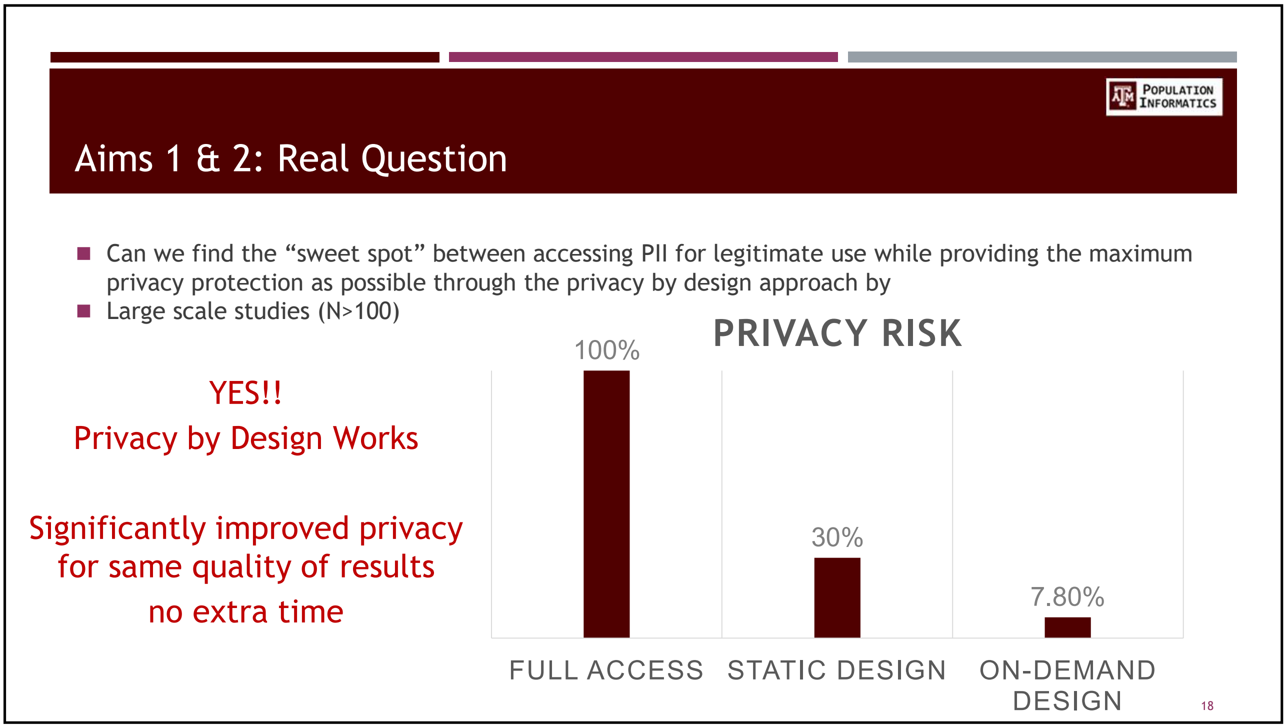
16



KAPR (k- anonymity privacy risk) score properties

- The privacy risk should be regularized to 0-100.
- Revealing information should always lead to a privacy risk increment.
- Privacy risk increment should be higher when disclosing information
- that leads to a lower anonymity set (disclosing unique names vs. disclosing common names).
- For any given state of disclosure, the KAPR score should always be the same. That is the order of disclosure should not matter.

17



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

18

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.

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

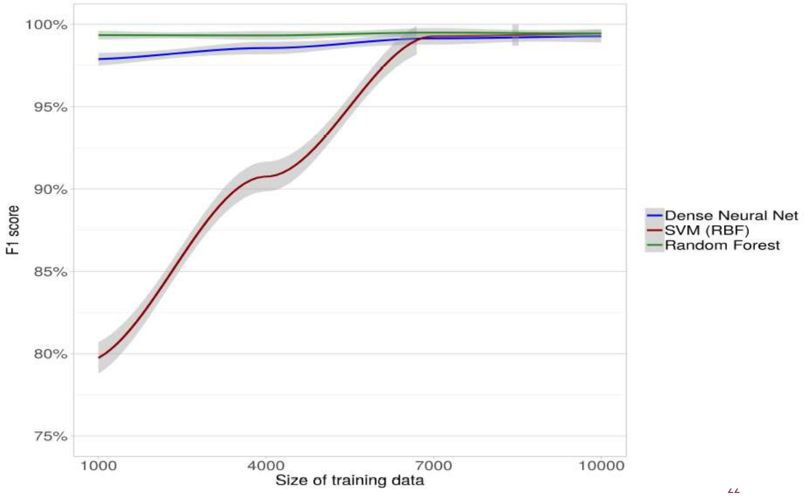
UTH & UAB Summative Study Results

21

Automatic Record Linkage

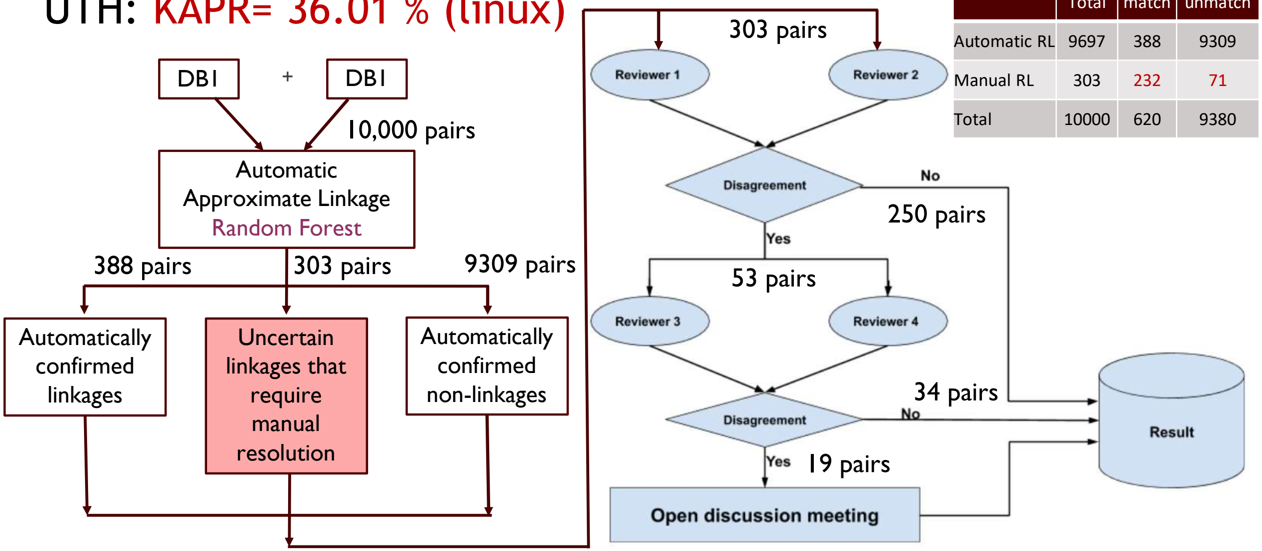
POPULATION INFORMATICS

- Random Forest
- Joffe E, Byrne MJ, Reeder P, Herskovic JR, Johnson CW, McCoy AB, Sittig DF, Bernstam EV. A benchmark comparison of deterministic and probabilistic methods for defining manual review datasets in duplicate records reconciliation. Journal of the American Medical Informatics Association. 2014 Jan 1;21(1):97-104.
  - 10,000 Training data
  - 10,000 Test data
- Record linkage benchmarking system



22

UTH: KAPR= 36.01 % (linux)



DB1 + DB1

10,000 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

250 pairs

53 pairs

34 pairs

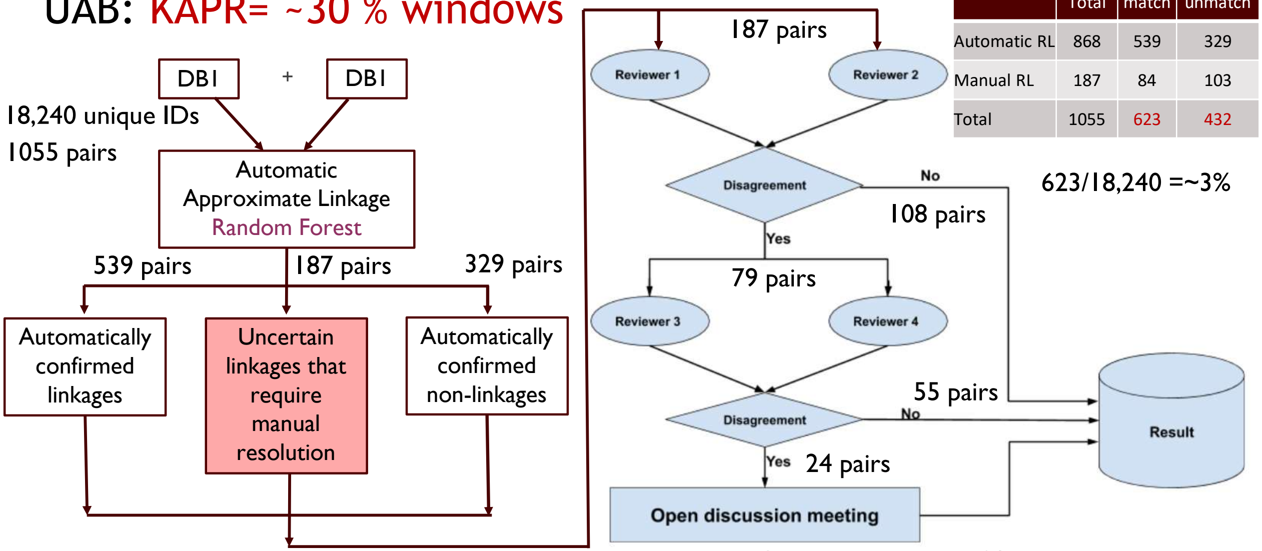
19 pairs

	Total	match	unmatch
Automatic RL	9697	388	9309
Manual RL	303	232	71
Total	10000	620	9380

Consensus reached: 19 Pairs
Consensus not reached: 0 Pairs

EHR

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

108 pairs

79 pairs

55 pairs

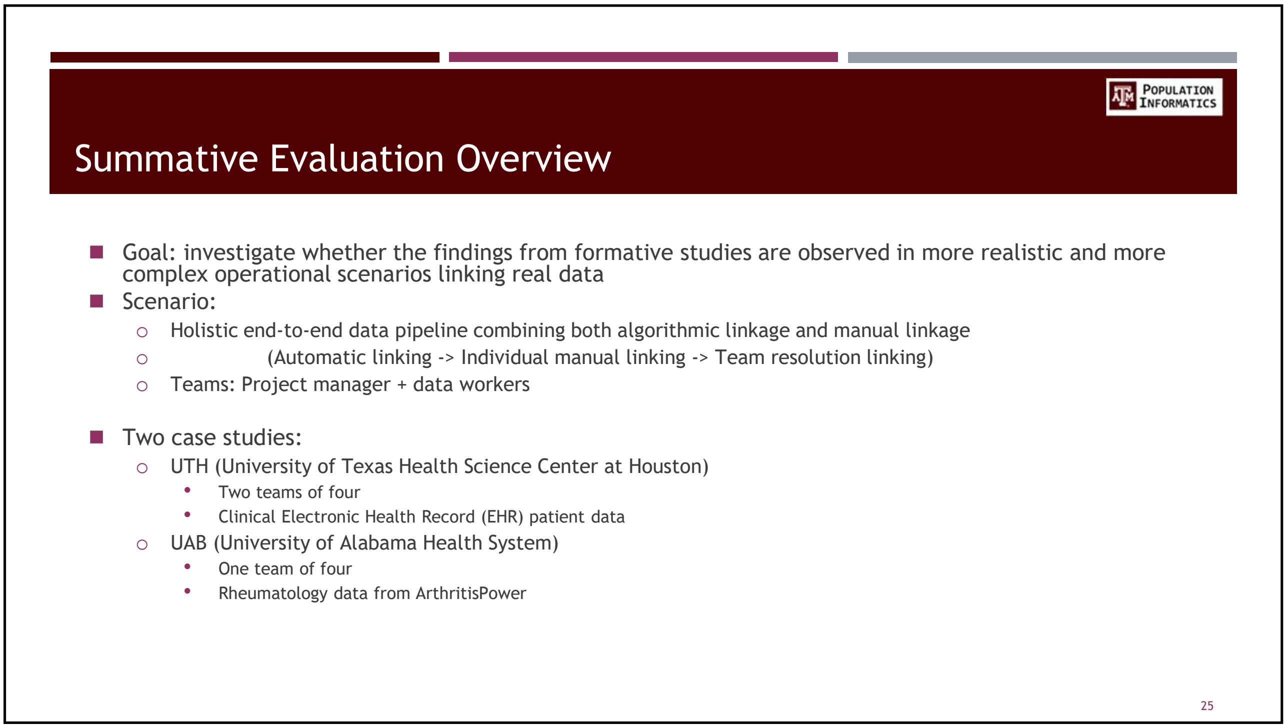
24 pairs

	Total	match	unmatch
Automatic RL	868	539	329
Manual RL	187	84	103
Total	1055	623	432

623/18,240 =~3%

Consensus reached: 23 Pairs
Consensus not reached: 1Pairs

Patient Generated Data



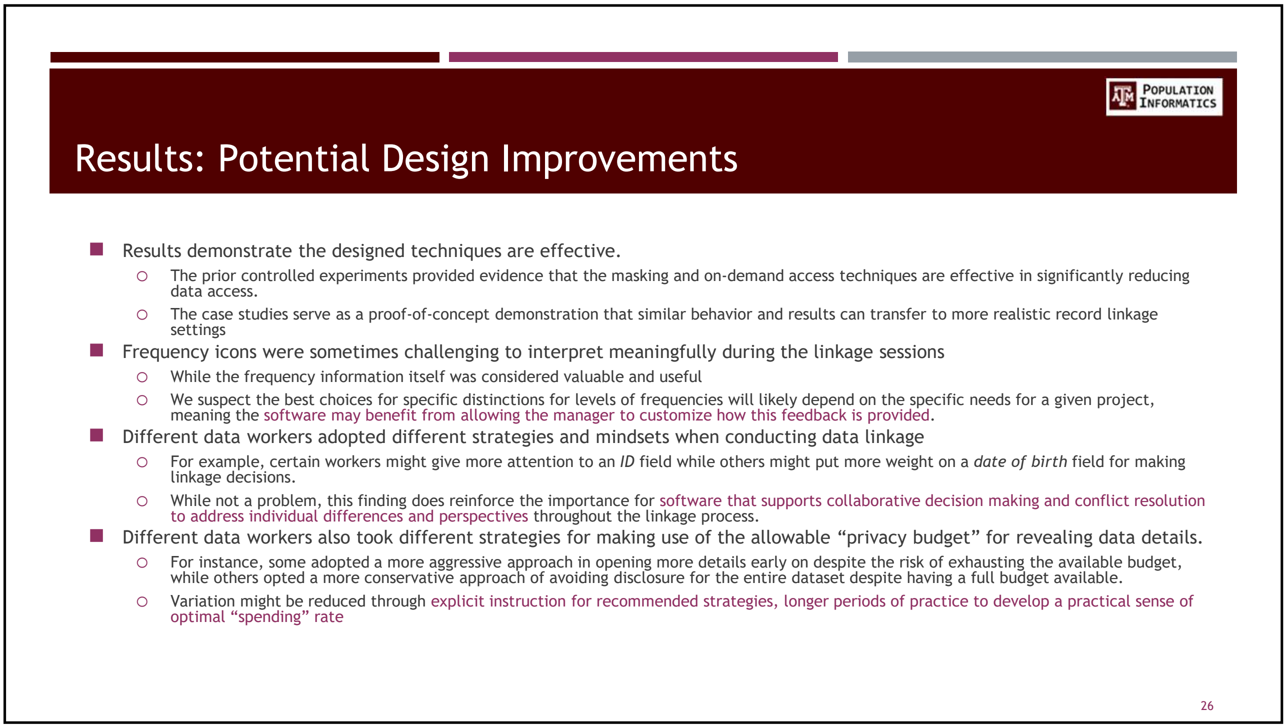
Summative Evaluation Overview

- Goal: investigate whether the findings from formative studies are observed in more realistic and more complex operational scenarios linking real data
- Scenario:
  - Holistic end-to-end data pipeline combining both algorithmic linkage and manual linkage
  - (Automatic linking -> Individual manual linking -> Team resolution linking)
  - Teams: Project manager + data workers

- Two case studies:
  - UTH (University of Texas Health Science Center at Houston)
    - Two teams of four
    - Clinical Electronic Health Record (EHR) patient data
  - UAB (University of Alabama Health System)
    - One team of four
    - Rheumatology data from ArthritisPower

25

25



Results: Potential Design Improvements

- Results demonstrate the designed techniques are effective.
  - The prior controlled experiments provided evidence that the masking and on-demand access techniques are effective in significantly reducing data access.
  - The case studies serve as a proof-of-concept demonstration that similar behavior and results can transfer to more realistic record linkage settings
- Frequency icons were sometimes challenging to interpret meaningfully during the linkage sessions
  - While the frequency information itself was considered valuable and useful
  - We suspect the best choices for specific distinctions for levels of frequencies will likely depend on the specific needs for a given project, meaning the software may benefit from allowing the manager to customize how this feedback is provided.
- Different data workers adopted different strategies and mindsets when conducting data linkage
  - For example, certain workers might give more attention to an *ID* field while others might put more weight on a *date of birth* field for making linkage decisions.
  - While not a problem, this finding does reinforce the importance for software that supports collaborative decision making and conflict resolution to address individual differences and perspectives throughout the linkage process.
- Different data workers also took different strategies for making use of the allowable “privacy budget” for revealing data details.
  - For instance, some adopted a more aggressive approach in opening more details early on despite the risk of exhausting the available budget, while others opted a more conservative approach of avoiding disclosure for the entire dataset despite having a full budget available.
  - Variation might be reduced through explicit instruction for recommended strategies, longer periods of practice to develop a practical sense of optimal “spending” rate

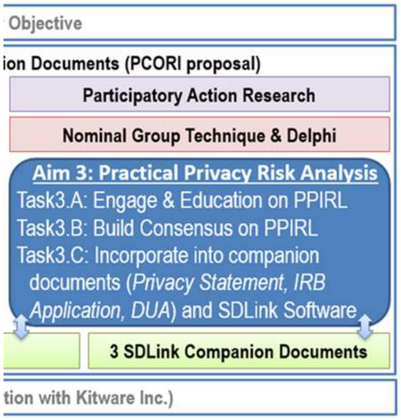
26

26

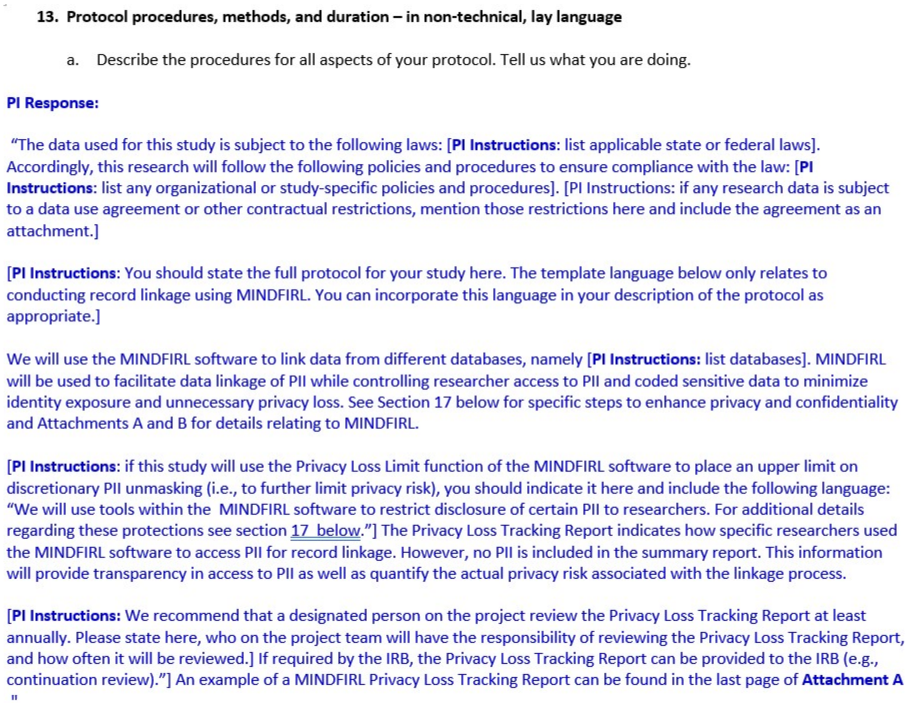
Aims & Outcomes
Prototype software & companion documents

POPULATION INFORMATICS

- Three companion documents
  - IRB template application:
    - NGT + Delphi with ELSI experts
  - Privacy statement (FAQ):
    - NGT + Delphi with patients
  - DUA:
    - adapted HHS DUA for data covered by the Privacy Act of 1974
    - Cason drafted with input from Hye-Chung
    - Under review by three other lawyers
      - ✓ UTH
      - ✓ UAB
      - ✓ Alva Ferdinand



Objective
ion Documents (PCORI proposal)
Participatory Action Research
Nominal Group Technique & Delphi
Aim 3: Practical Privacy Risk Analysis
Task3.A: Engage & Education on PPIRL
Task3.B: Build Consensus on PPIRL
Task3.C: Incorporate into companion documents (Privacy Statement, IRB Application, DUA) and SDLink Software
3 SDLink Companion Documents
tion with Kitware Inc.)



13. Protocol procedures, methods, and duration – in non-technical, lay language

a. Describe the procedures for all aspects of your protocol. Tell us what you are doing.

PI Response:

"The data used for this study is subject to the following laws: [**PI Instructions**: list applicable state or federal laws]. Accordingly, this research will follow the following policies and procedures to ensure compliance with the law: [**PI Instructions**: list any organizational or study-specific policies and procedures]. [PI Instructions: if any research data is subject to a data use agreement or other contractual restrictions, mention those restrictions here and include the agreement as an attachment.]

[**PI Instructions**: You should state the full protocol for your study here. The template language below only relates to conducting record linkage using MINDFIRL. You can incorporate this language in your description of the protocol as appropriate.]

We will use the MINDFIRL software to link data from different databases, namely [**PI Instructions:** list databases]. MINDFIRL will be used to facilitate data linkage of PII while controlling researcher access to PII and coded sensitive data to minimize identity exposure and unnecessary privacy loss. See Section 17 below for specific steps to enhance privacy and confidentiality and Attachments A and B for details relating to MINDFIRL.

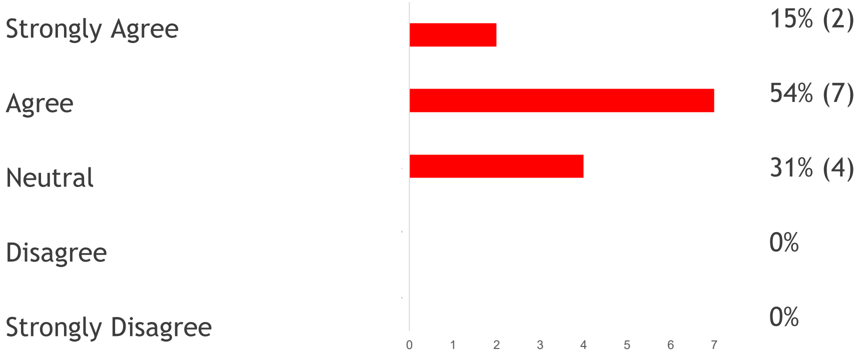
[**PI Instructions**: if this study will use the Privacy Loss Limit function of the MINDFIRL software to place an upper limit on discretionary PII unmasking (i.e., to further limit privacy risk), you should indicate it here and include the following language: "We will use tools within the MINDFIRL software to restrict disclosure of certain PII to researchers. For additional details regarding these protections see section 17 below."] The Privacy Loss Tracking Report indicates how specific researchers used the MINDFIRL software to access PII for record linkage. However, no PII is included in the summary report. This information will provide transparency in access to PII as well as quantify the actual privacy risk associated with the linkage process.

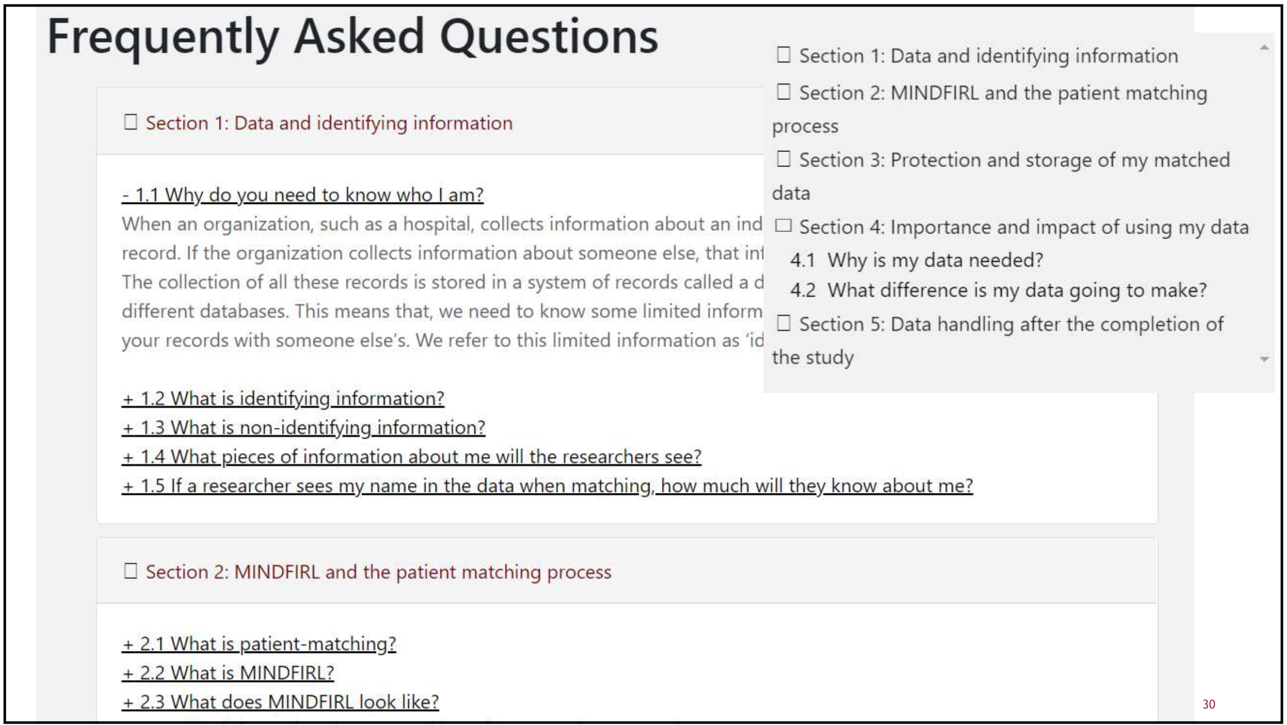
[**PI Instructions:** We recommend that a designated person on the project review the Privacy Loss Tracking Report at least annually. Please state here, who on the project team will have the responsibility of reviewing the Privacy Loss Tracking Report, and how often it will be reviewed.] If required by the IRB, the Privacy Loss Tracking Report can be provided to the IRB (e.g., continuation review)."] An example of a MINDFIRL Privacy Loss Tracking Report can be found in the last page of **Attachment A**
"

- Full IRB application
- 10 pgs

Aim 3 IRB template Highlights (N=13)

- We asked ELSI experts about their opinion on risk reduction to minimum when using MINDFIRL
- “The use of the MINDFIRL software will further reduce risk to the minimum necessary to conduct reliable record linkage.”





Frequently Asked Questions

☐ Section 1: Data and identifying information

- 1.1 Why do you need to know who I am?

When an organization, such as a hospital, collects information about an ind record. If the organization collects information about someone else, that int The collection of all these records is stored in a system of records called a d different databases. This means that, we need to know some limited inform your records with someone else’s. We refer to this limited information as ‘id

+ 1.2 What is identifying information?
+ 1.3 What is non-identifying information?
+ 1.4 What pieces of information about me will the researchers see?
+ 1.5 If a researcher sees my name in the data when matching, how much will they know about me?

☐ Section 2: MINDFIRL and the patient matching process

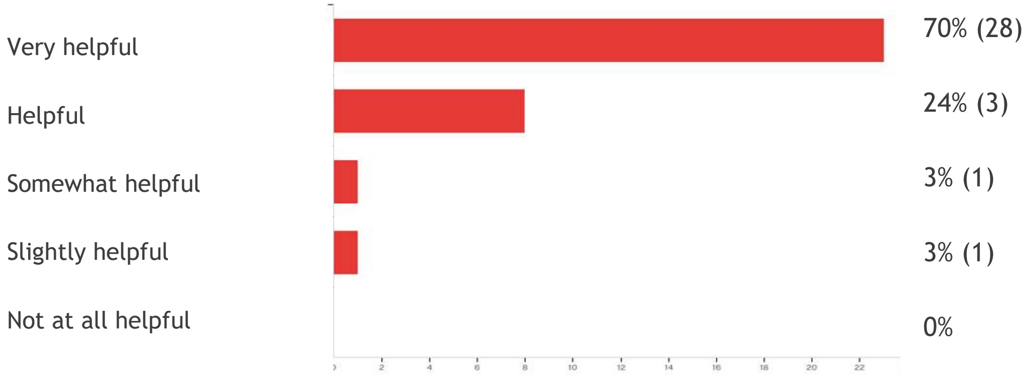
+ 2.1 What is patient-matching?
+ 2.2 What is MINDFIRL?
+ 2.3 What does MINDFIRL look like?

☐ Section 1: Data and identifying information
☐ Section 2: MINDFIRL and the patient matching process
☐ Section 3: Protection and storage of my matched data
☐ Section 4: Importance and impact of using my data
4.1 Why is my data needed?
4.2 What difference is my data going to make?
☐ Section 5: Data handling after the completion of the study

30

Aim 3 FAQ: Highlights I (N=33)

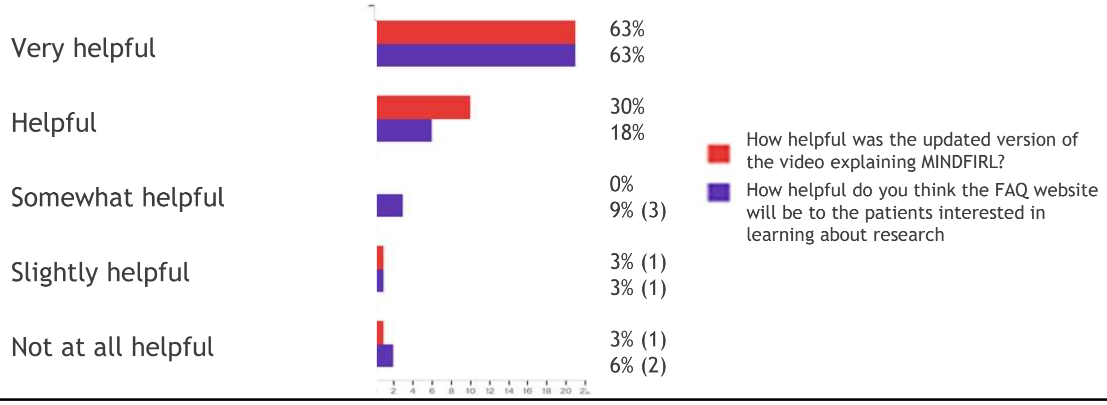
- With respect to the new website format of the FAQ, most participants found the updated format to be very helpful as shown below:



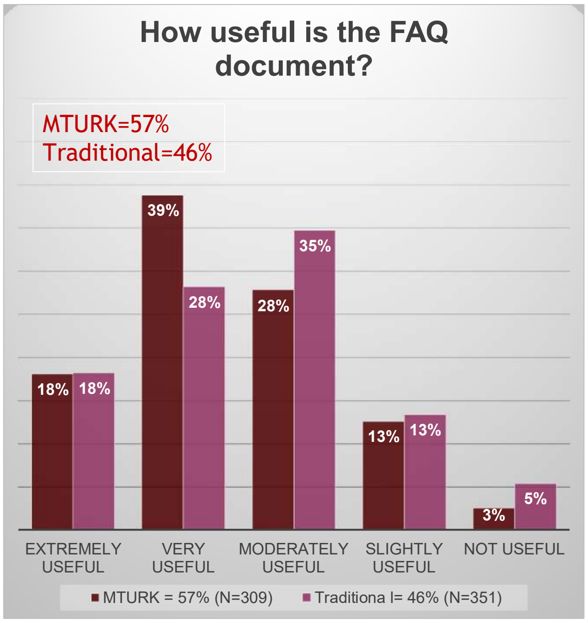
31

Aim 3 FAQ: Highlights II

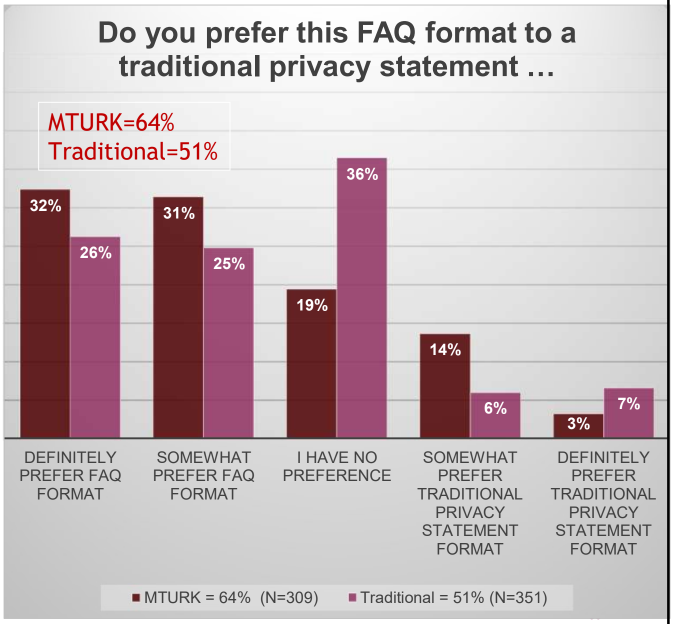
- We developed a dynamic website based FAQ (http://mindfirl-uth.herokuapp.com/faq)
- We also created and shared a video demonstrating MINDFIRL



32



How useful is the FAQ document?



Do you prefer this FAQ format to a traditional privacy statement …

Open ended feedback: Positive

- Easy to navigate

> **"I really like the FAQ layout because it's not as cumbersome to read as a traditional privacy policy. It's easier to open up each section as I like."**

- Patient centered voice

> **"Definitely like the sections being broken apart into questions I might have, I think it reframes the document into a user-centered POV and I think that shows consideration."**

- Liked the comprehensive detailed explanations
  - Tension between those who want more/less detail

> **"I like how thorough this FAQ is and the in-depth responses. I also like being able to choose the topics that most interest me, or that I have less an understanding of."**

POPULATION INFORMATICS

Open ended feedback: Negative

- Too much information
  - Tension between those who want more/less detail

"I think they are just too long and no one will actually read them."

- Still have concerns on privacy risk

"While I like the format, it doesn't change the problem of a system getting compromised. So it's helpful in providing answers, but would not eliminate my concerns. Best intentions don't always lead to good results."

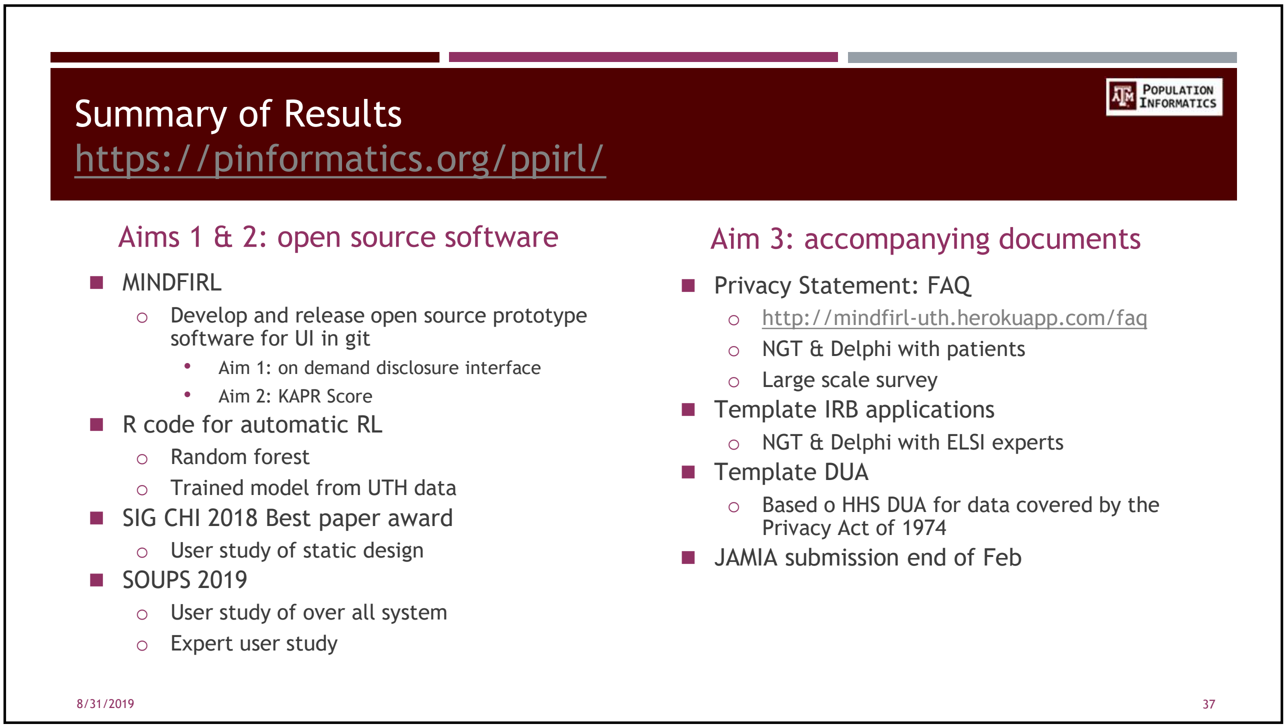
- Missing information: Details on what happens after data breach

"I would like to see what would happen if there would be a data breach. How would a company be accountable."

POPULATION INFORMATICS

Suggested Improvements

- Add a search box
  - "I like the way it is setup, it is easy to follow and navigate. It might be nice if there was a search box since there is so much information, it could take awhile to find the exact answer you are looking for. "
- Multiple languages
  - "Make sure it is available in multiple languages. Otherwise it looks fine to me."



Summary of Results
https://pinformatics.org/ppirl/

Aims 1 & 2: open source software

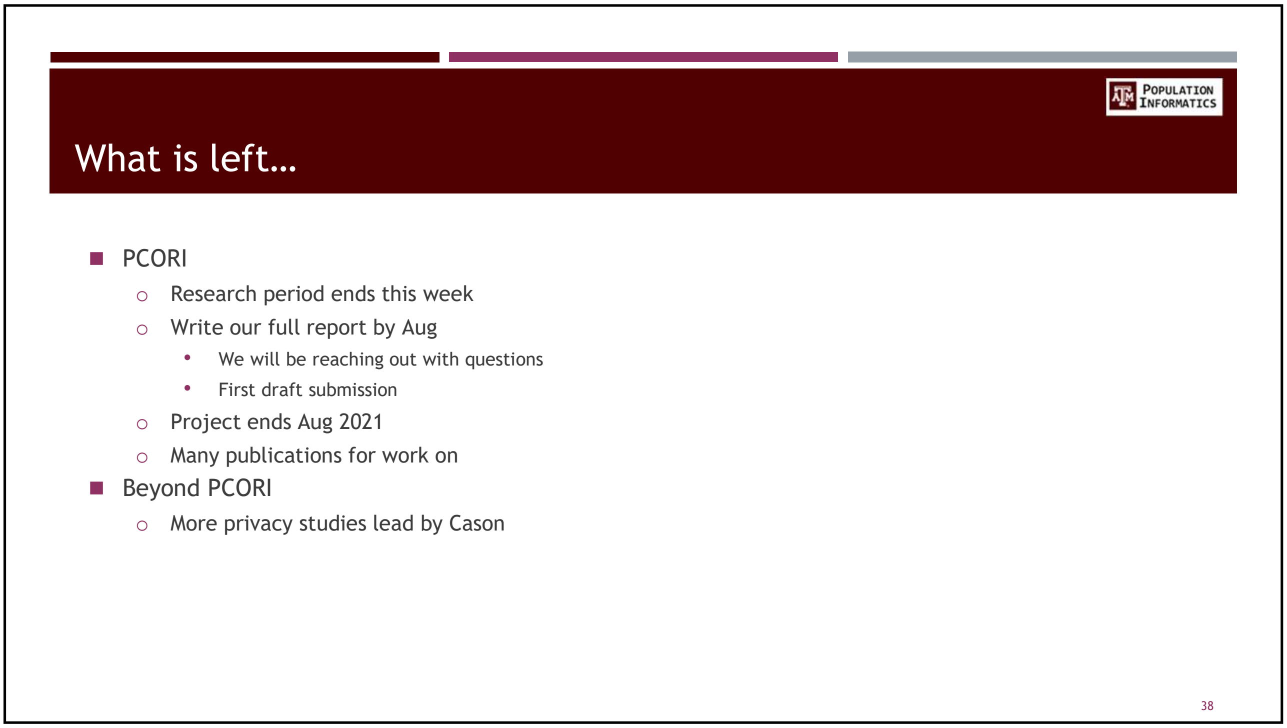
- MINDFIRL
  - Develop and release open source prototype software for UI in git
    - Aim 1: on demand disclosure interface
    - Aim 2: KAPR Score
- R code for automatic RL
  - Random forest
  - Trained model from UTH data
- SIG CHI 2018 Best paper award
  - User study of static design
- SOUPS 2019
  - User study of over all system
  - Expert user study

Aim 3: accompanying documents

- Privacy Statement: FAQ
  - http://mindfirl-uth.herokuapp.com/faq
  - NGT & Delphi with patients
  - Large scale survey
- Template IRB applications
  - NGT & Delphi with ELSI experts
- Template DUA
  - Based o HHS DUA for data covered by the Privacy Act of 1974
- JAMIA submission end of Feb

8/31/2019 37

37



What is left…

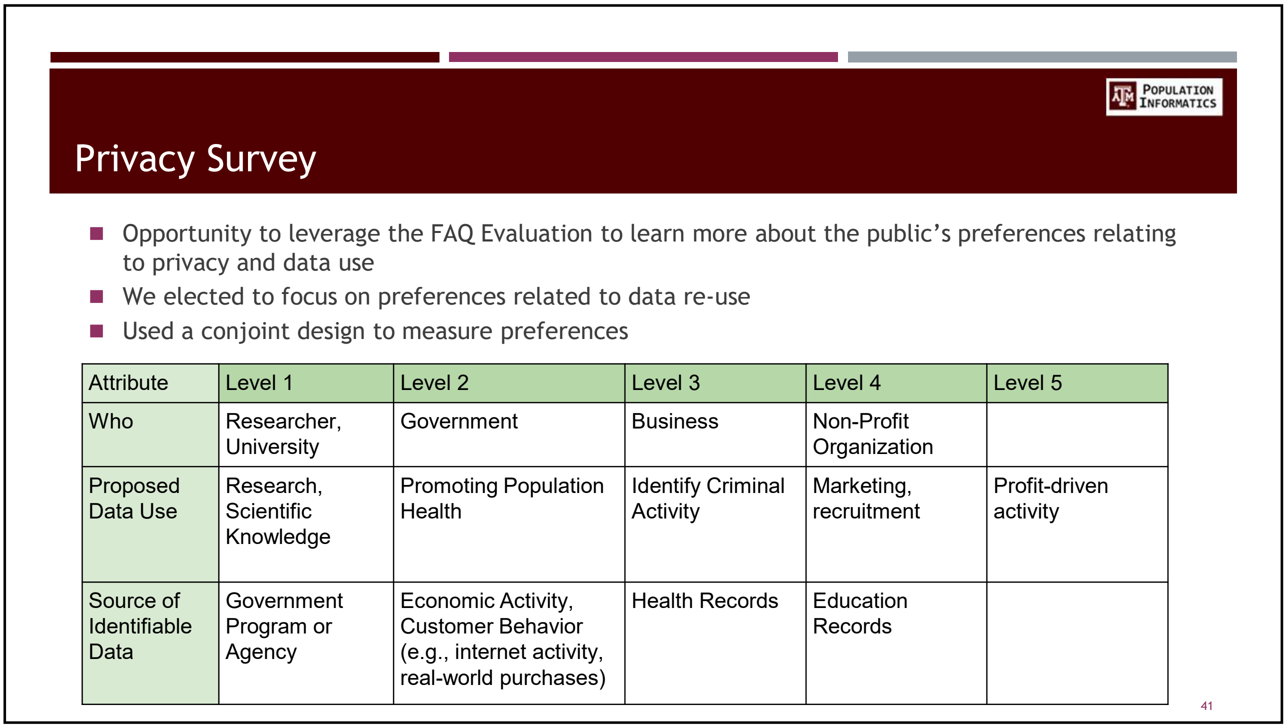
- PCORI
  - Research period ends this week
  - Write our full report by Aug
    - We will be reaching out with questions
    - First draft submission
  - Project ends Aug 2021
  - Many publications for work on
- Beyond PCORI
  - More privacy studies lead by Cason

38

38

Privacy Survey Results

Cason Schmit

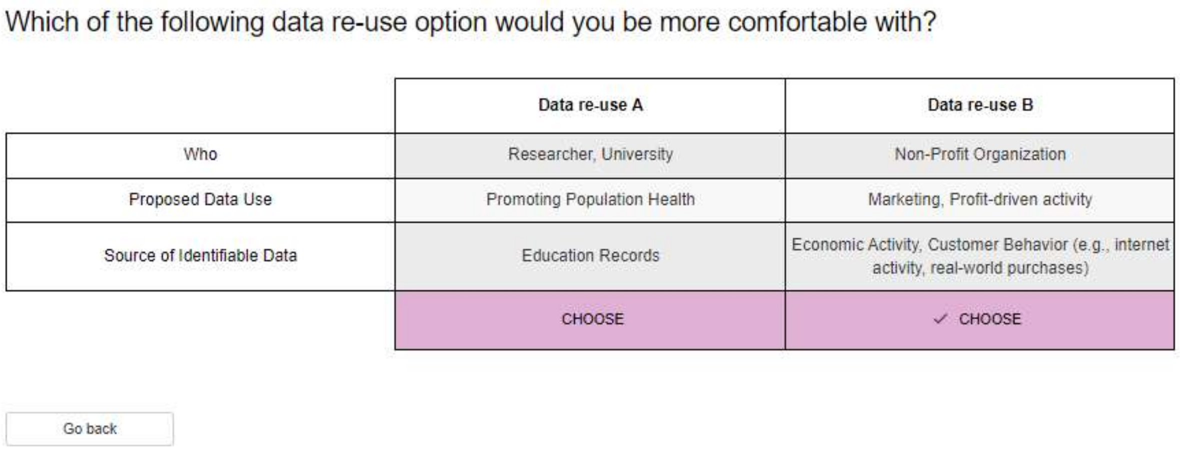


Privacy Survey

- Opportunity to leverage the FAQ Evaluation to learn more about the public's preferences relating to privacy and data use
- We elected to focus on preferences related to data re-use
- Used a conjoint design to measure preferences

Attribute	Level 1	Level 2	Level 3	Level 4	Level 5
Who	Researcher, University	Government	Business	Non-Profit Organization	
Proposed Data Use	Research, Scientific Knowledge	Promoting Population Health	Identify Criminal Activity	Marketing, recruitment	Profit-driven activity
Source of Identifiable Data	Government Program or Agency	Economic Activity, Customer Behavior (e.g., internet activity, real-world purchases)	Health Records	Education Records	

Sample Conjoint Question



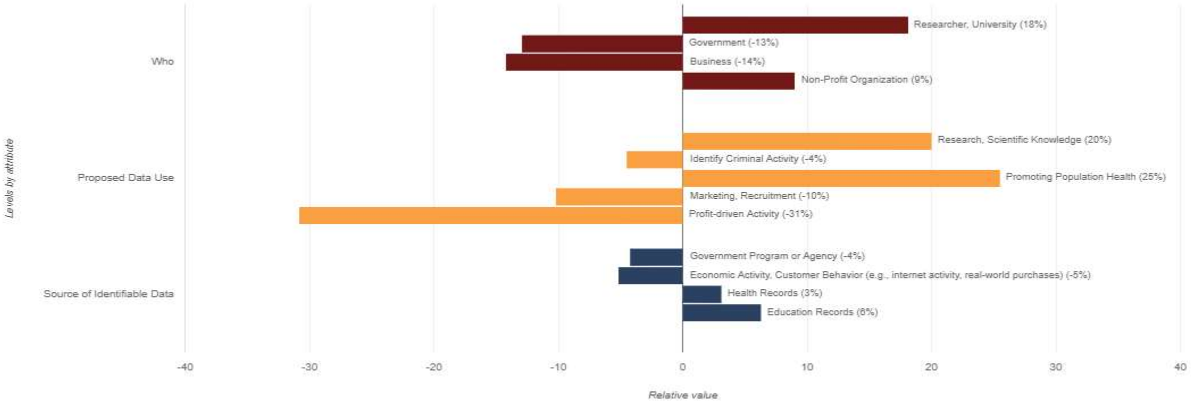
Which of the following data re-use option would you be more comfortable with?

	Data re-use A	Data re-use B
Who	Researcher, University	Non-Profit Organization
Proposed Data Use	Promoting Population Health	Marketing, Profit-driven activity
Source of Identifiable Data	Education Records	Economic Activity, Customer Behavior (e.g., internet activity, real-world purchases)
	CHOOSE	✓ CHOOSE

42

42

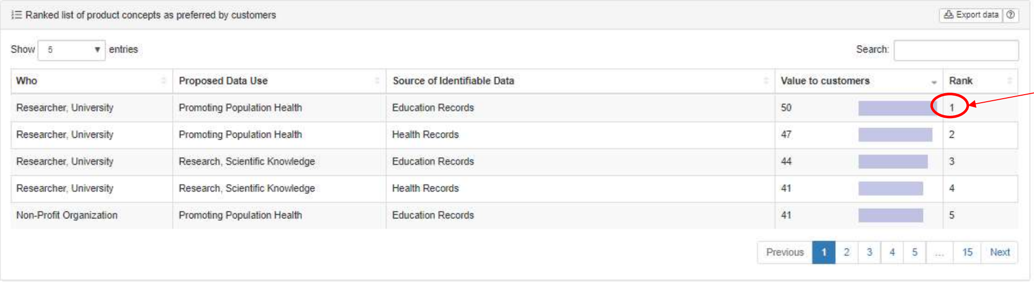
Preliminary Results from Privacy Conjoint



43

43

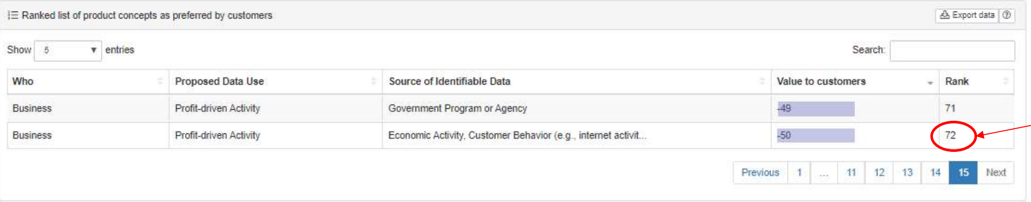
Preliminary Results (Cont.)



Ranked list of product concepts as preferred by customers

Who	Proposed Data Use	Source of Identifiable Data	Value to customers	Rank
Researcher, University	Promoting Population Health	Education Records	50	1
Researcher, University	Promoting Population Health	Health Records	47	2
Researcher, University	Research, Scientific Knowledge	Education Records	44	3
Researcher, University	Research, Scientific Knowledge	Health Records	41	4
Non-Profit Organization	Promoting Population Health	Education Records	41	5

Illegal



Ranked list of product concepts as preferred by customers

Who	Proposed Data Use	Source of Identifiable Data	Value to customers	Rank
Business	Profit-driven Activity	Government Program or Agency	-49	71
Business	Profit-driven Activity	Economic Activity, Customer Behavior (e.g., internet activit...	-50	72

Legal

Input on Additional Surveys: Research Values

	Attribute	Level 1	Level 2	Level 3
Given	Speed	Fast	Typical	Slow
	Cost of Research	Expensive	Typical	Cheap
Variable	Precision/ Quality	High	Medium	Low
	Data Protections (Privacy/ Security)	High	Medium	Low
	Probability of Success (Probability of Waste)	High (Low)	Medium (medium)	Low (High)
	Benefit (Utility)	High	Medium	Low

Input on Additional Surveys: Research Priorities

- You have a $1000 to spend on data project. How should you spend the money?
  - Research
  - Population health
  - Evaluate government program
  - Audit program
  - Identify/investigate criminal activity
  - Others?

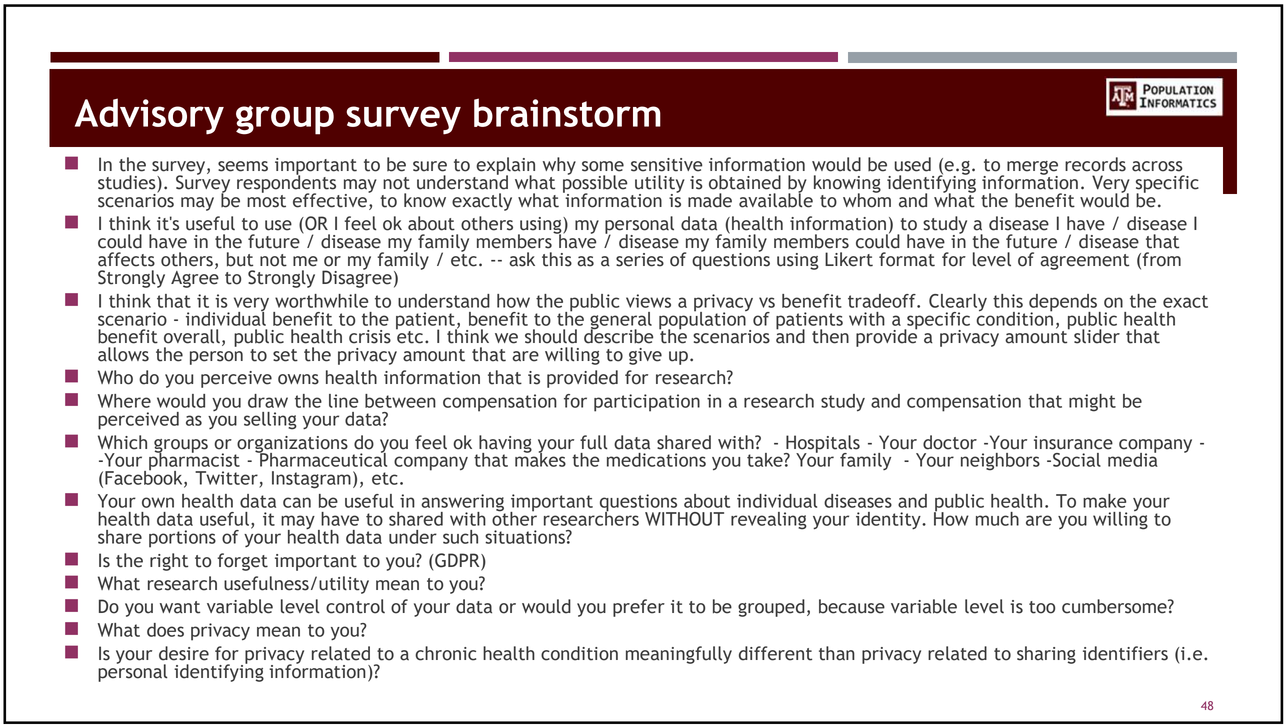
46

46

Input on Additional Surveys: Research (Big Data) Ethics

Size of activity	Small (data from 500 people)	Medium (data from 20,000 people)	Large (data from 1,000,000 people)	
Respect for persons	The project lead met with members of the public and relevant community groups to understand their perspectives. The project was designed with these perspectives in mind.	The project is not risky, and it will be very difficult to get informed consent from so the project lead is asking for your permission to skip the informed consent process		
Harms	#1a Equity+ (activity might reduce the burdens and risks that threaten health or opportunity of a group)	#1b Justice+ (activity is fair to potential participants relating to anticipated risks and benefits)	#2aEquity - (There is some concern that the activity might increase the burdens and risks that threaten health or opportunity of a group)	#2bJustice - (There is some concern that the activity might expose participants to risks, and the participants (and others like them) are unlikely to benefit from the activity)
Good governance	Takes steps for transparency, accountability, and data protection	Takes steps to protect data as required by law or the organization's policy .		
Common Good	The activity promotes population health or other common good	The activity mostly benefits the user or organization, but might have some anticipated societal benefits	The activity mostly benefits the user or organization	
Beneficence	Some risk of harm to participants	Minimal risk of harm to participants		

47

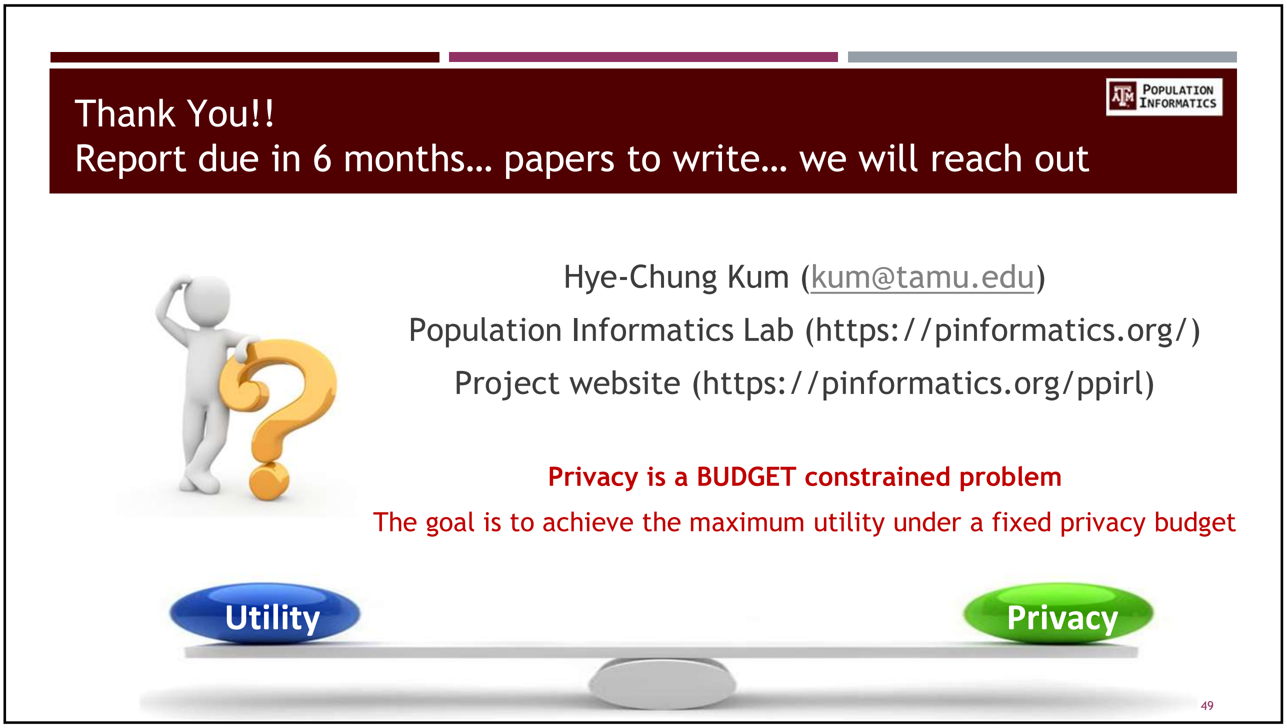


Advisory group survey brainstorm

- In the survey, seems important to be sure to explain why some sensitive information would be used (e.g. to merge records across studies). Survey respondents may not understand what possible utility is obtained by knowing identifying information. Very specific scenarios may be most effective, to know exactly what information is made available to whom and what the benefit would be.
- I think it's useful to use (OR I feel ok about others using) my personal data (health information) to study a disease I have / disease I could have in the future / disease my family members have / disease my family members could have in the future / disease that affects others, but not me or my family / etc. -- ask this as a series of questions using Likert format for level of agreement (from Strongly Agree to Strongly Disagree)
- I think that it is very worthwhile to understand how the public views a privacy vs benefit tradeoff. Clearly this depends on the exact scenario - individual benefit to the patient, benefit to the general population of patients with a specific condition, public health benefit overall, public health crisis etc. I think we should describe the scenarios and then provide a privacy amount slider that allows the person to set the privacy amount that are willing to give up.
- Who do you perceive owns health information that is provided for research?
- Where would you draw the line between compensation for participation in a research study and compensation that might be perceived as you selling your data?
- Which groups or organizations do you feel ok having your full data shared with? - Hospitals - Your doctor -Your insurance company - -Your pharmacist - Pharmaceutical company that makes the medications you take? Your family - Your neighbors -Social media (Facebook, Twitter, Instagram), etc.
- Your own health data can be useful in answering important questions about individual diseases and public health. To make your health data useful, it may have to shared with other researchers WITHOUT revealing your identity. How much are you willing to share portions of your health data under such situations?
- Is the right to forget important to you? (GDPR)
- What research usefulness/utility mean to you?
- Do you want variable level control of your data or would you prefer it to be grouped, because variable level is too cumbersome?
- What does privacy mean to you?
- Is your desire for privacy related to a chronic health condition meaningfully different than privacy related to sharing identifiers (i.e. personal identifying information)?

48

48



Thank You!!
Report due in 6 months… papers to write… we will reach out

Hye-Chung Kum (kum@tamu.edu)

Population Informatics Lab (https://pinformatics.org/)

Project website (https://pinformatics.org/ppirl)

Privacy is a BUDGET constrained problem

The goal is to achieve the maximum utility under a fixed privacy budget

49

49