Enhancing Privacy through an Interactive On-demand Incremental Information Disclosure Interface: Applying Privacy-by-Design to Record Linkage

Hye-Chung Kum (kum@tamu.edu); Cason Schmit (schmit@sph.tamu.edu)
Gurudev Ilangovan; Mahin Ramezani; Qinbo Li
Population Informatics Lab (https://pinformatics.org/), Texas A&M University

Eric D. Ragan (ergan@ufl.edu)
INDIE Lab, University of Florida

Legitimate access to PII
Data Wrangling (cleaning & curation) is essential to data analytics

POPULATION INFORMATICS

INDIE

The New York Times | http://nyti.ms/1mZywng

TECHNOLOGY

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

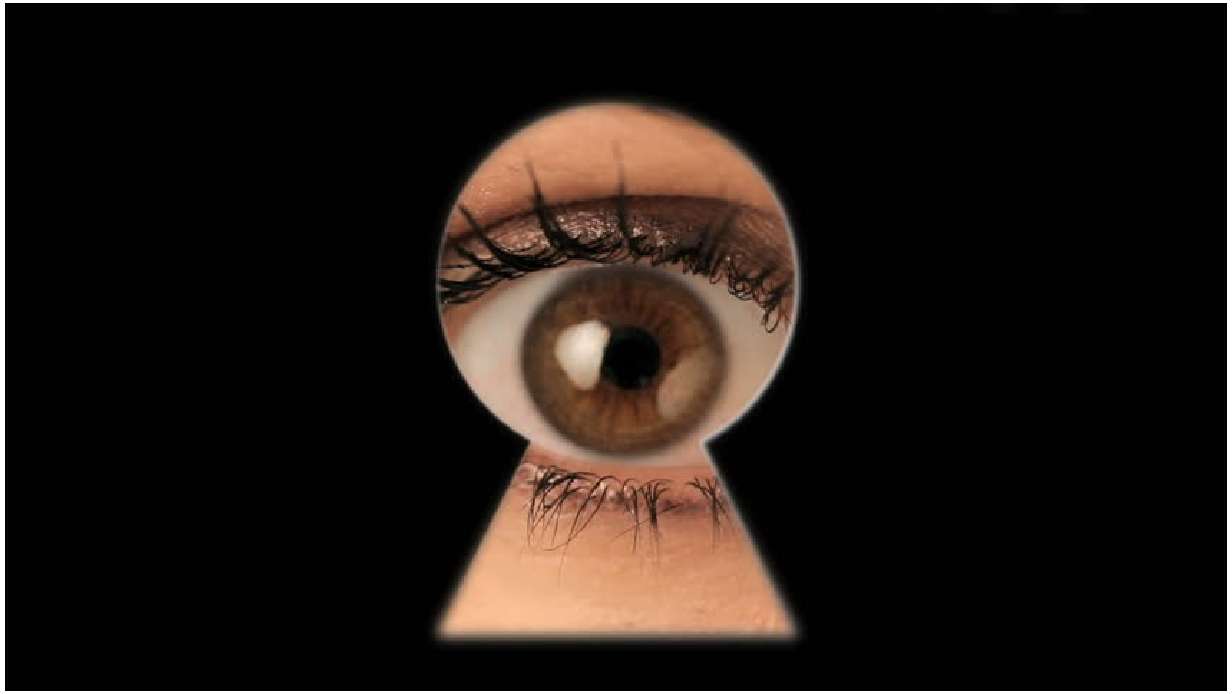
By STEVE LOHR AUG. 17, 2014

- Data Wrangling is a term that is applied to activities that make data more usable by changing their form but not their meaning
  - reformatting data: MDY vs YMD
  - mapping data from one data model to another: ICD9 vs CPT code
  - and/or converting data into more consumable forms: to graphs
- 30-80% of the work in using big data
- Once raw data is "wrangled" into the correct analytic data
  - Running statistics models are fairly simple and similar to what you do traditionally
  - There are new methods but, usually requires a LOT of data

Legitimate access to PII
Tuning parameters & building training data in ML

- Most all data analytics
  - Must tune parameters: Requires manual interaction with the data (even PII)
- Machine learning algorithms
  - Requires building training data

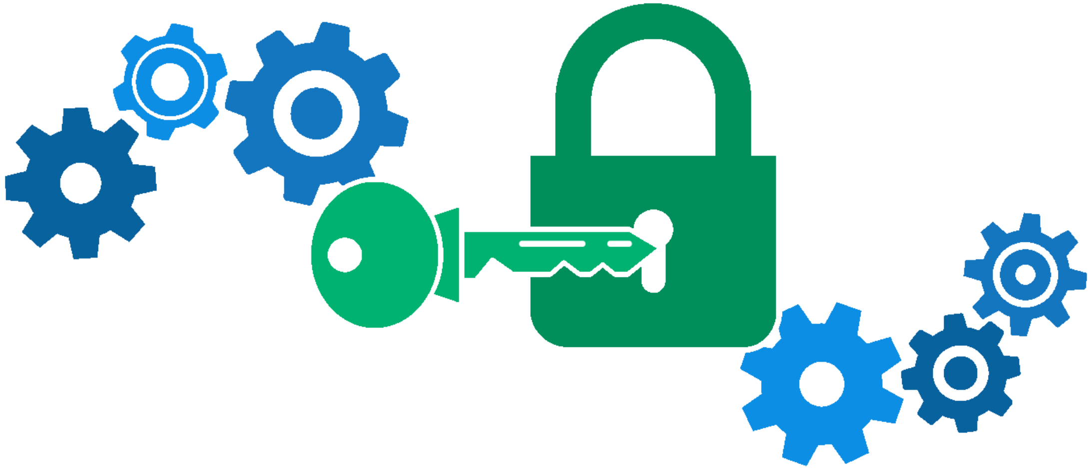
Personal Data and Privacy



- Not legitimate without explicit permission
  - Advertising tracking location
- Legitimate without informed consent
  - Track how many emergency department a patient visited
    - For better clinical care
    - To improve policies for reimbursement
  - Track use of opioid to assess relationship between addiction and treatment
  - Analyze relationship between cancer and HIV
  - Track outcomes to evaluate and improve public programs such as child welfare
    - Educational outcomes for children in foster care
    - Income outcomes
    - Incarceration outcomes

Partial Solutions
Sorry, data scientists can't do magic

- Restrict access
- Algorithms and automation
- Encryption
- Aggregation
- Synthetic data



Uncertainty + Human Judgement
Garbage in & Garbage out: Requires human in the loop

- Data cleaning, data wrangling
- Deduplication
- Record linkage
- Parameter tuning
- Building training datasets
- Anomaly investigation

NO FREE LUNCH!! Privacy vs. Utility

- Related background from literature on Differential Privacy
  - Research has demonstrated that information **privacy is a budget-constrained problem** that requires reasoning about the *tradeoff* between privacy and utility for a given context
  - Consequently, there is **no "one-size-fits-all" solution**, and there is **no way to benefit from using data without taking some privacy risks**.

We have Hope
Garbage in & Garbage out: Requires human in the loop

- Data cleaning, data wrangling
- Deduplication
- Record linkage
- Parameter tuning
- Building training datasets
- Anomaly investigation

- Key Insights
  1. **So EXACTLY how much data on average does a data scientist need dig into for high quality results?**
  2. **Who knows where to look?**
  3. **When do they know where to look?**
  4. **Can they tell you why they need to look where?**

Insight: How do you enhance privacy while maintaining effectiveness
What are key design elements for privacy enhanced systems?

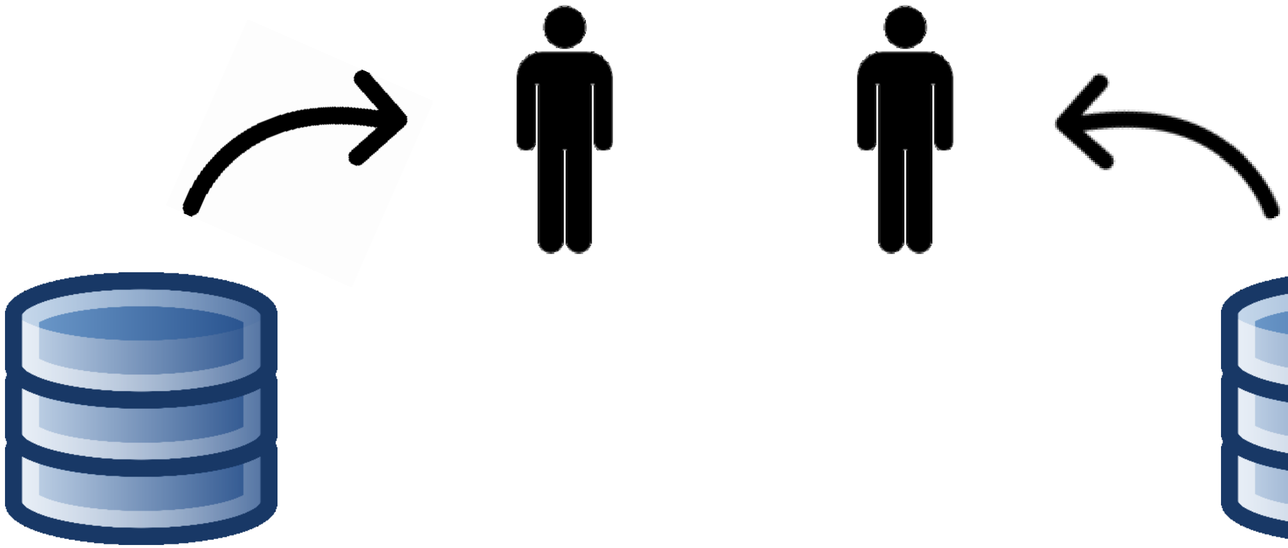
- Current approaches: All or Nothing
  - Either have approval to access EVERYTHING
  - OR access NOTHING
- **Need better ability to balance tradeoffs between privacy and utility**
  - **Partial Access**: only when needed, and only what is needed for good decisions (e.g., parameter tuning, data cleaning, validation etc)
    - Example: last four digits of SSN,
  - **Make just-in-time decision on what needs to be accessed**
  - **Monitoring on level of access**: (e.g. security cameras)
    - **Quantifying access level**: ability to compare, detect anomalies etc
  - **Be accountable** for what was accessed: audits (e.g., logs)

Problem Statement

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

Research Overview

- Goals:
  - Privacy goal: Limiting disclosure of personal information
  - Utility goal: But not reduce human effectiveness

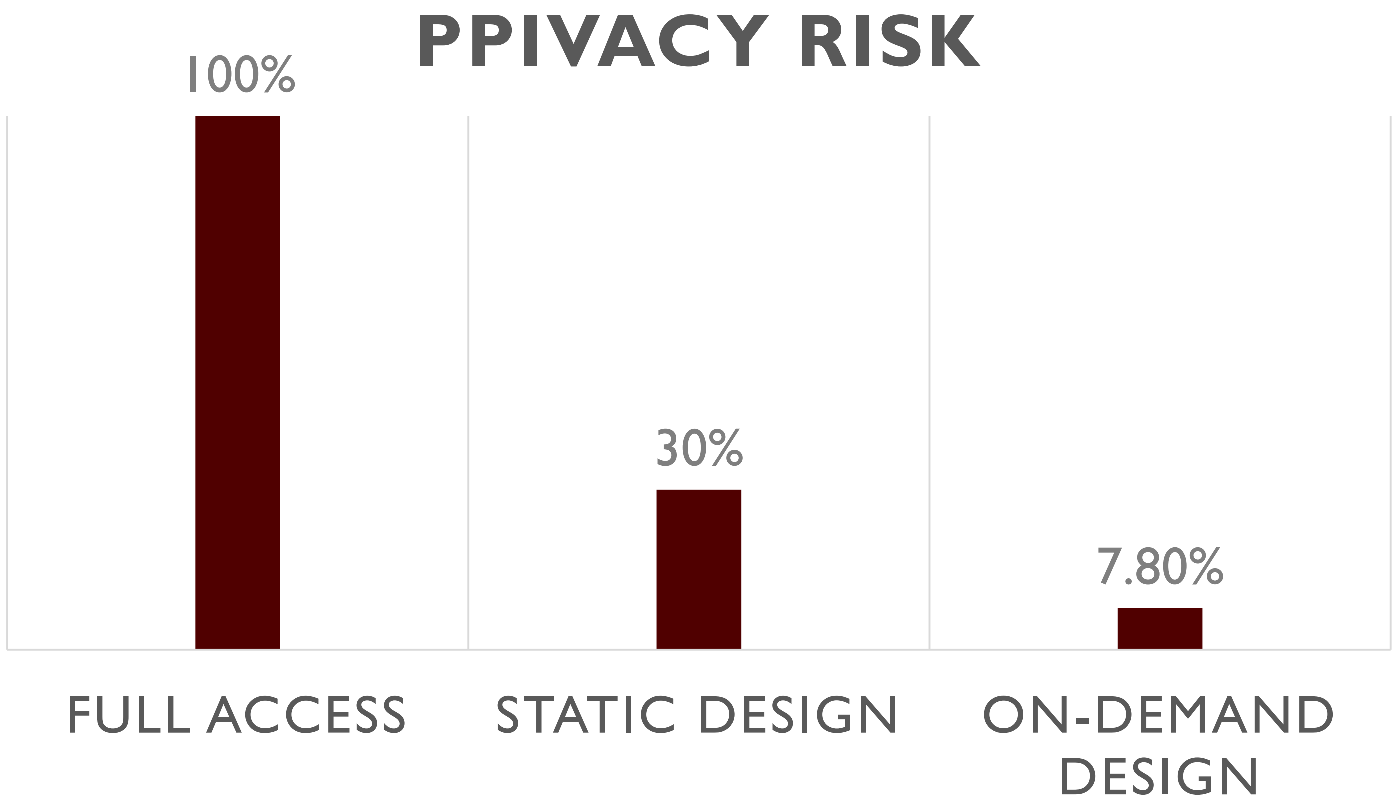
Real Question & Spoiler

- 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

YES!!

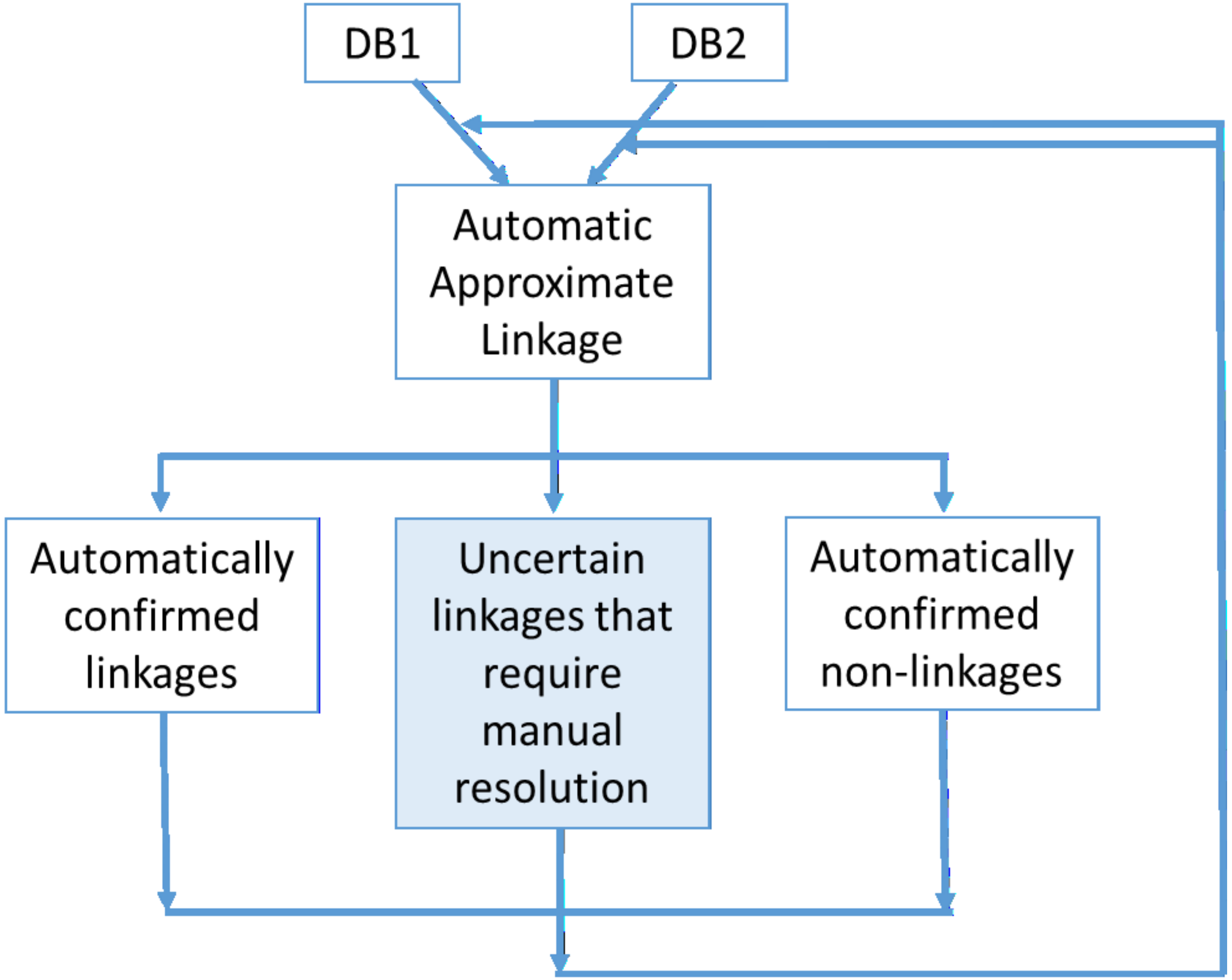
Privacy by Design Works

Significantly improved privacy for same quality of results



Background & Previous Work

Hybrid Human Computer Process



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

Application

- Uncertainty in data
  - Requires Human Judgement
  - Human Interaction With Data
    - ✓ Standardize Data
    - ✓ Clean Data
    - ✓ Build Training Data
    - ✓ Tune Model Parameters

- Common Issues
  - Typos
  - Nicknames
  - Switched characters
  - Name changes
  - Missing values
  - Family members

- Given multiple databases, determine if records refer to the same real world people or not
- Your job in this study is to:
  1) Look at pairs of rows of data about people
  2) Decide whether or not the pair refers to the same person.

Pair	ID	First name	Last name	DoB(M/D/Y)	Sex	Race
1	8000002767	JUDE	WILLIAM	09/09/1906	M	W
	8000003567	JUDE	WILLIAM JR	09/09/1960	M	B
2	0000006947	BRYANT	MADELINE	05/02/1962	F	W
	0000006947	MADELINE	BRYANT	05/02/1962	F	W
3	9000018540	SALLY	BYRD	07/04/1960	F	W
	6000008928	JOHN	BYRD	04/07/1960	M	

Maybe Father/Son

Probably data error

Maybe Twins

Status Quo: Access to ALL for approved personnel

Population Informatics

INDIE

Pair	ID	First name	Last name	DoB(M/D/Y)	Sex	Race
1	8000002767	JUDE	WILLIAM	09/09/1906	M	W
	8000003567	JUDE	WILLIAM JR	09/09/1960	M	B
2	0000006947	BRYANT	MADELINE	05/02/1962	F	W
	0000006947	MADELINE	BRYANT	05/02/1962	F	W
3	9000018540	SALLY	BYRD	07/04/1960	F	W
	6000008928	JOHN	BYRD	04/07/1960	M	

Maybe Father/Son

Probably data error

Maybe Twins

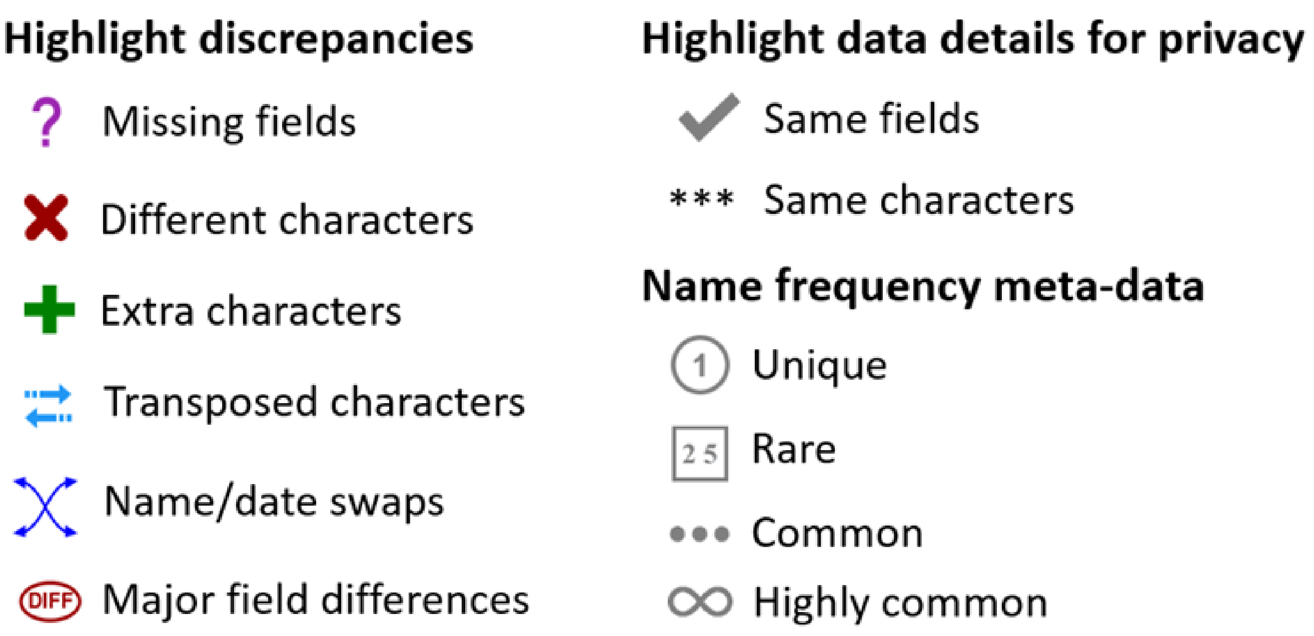
One Privacy Preserving Approach: Show NOTHING
Encrypted disclosure

Pair	ID	First name	Last name	DoB(M/D/Y)	Sex	Race
31	@@zz@@@@e@@@@@yc@tcflg==	JPHm/tFJf/Sa38z+PthPY				
	&&qw&p&&&&&&m&&&&v1m&&==	AgsX5d/vZ1tRukT6GTxCZ				
32	vx&+&&h&v&+&xnmyqmaa&&	AgsX5d/vZ1tRukT6GTxCZw				
	vx&+&&h&v&+&xnmyqmaa&&	@@@l@i@g@@@os@@bn@@@@g				
33	@@@l@i@g@@@os@@bn@@@@g==	HOMwdz8KpFKaTfPE+qr8Xw==	/KSKzJ2U5C/fpHmkMqZP	JPHm/tFJf/Sa38z+PthPYQ==	>	%
	vx&+&&h&v&+&xnmyqmaa&&==	o1fSci26GzxKx41n1lkRuQ==	bJupC1Skjj/bmw9DRq07	AgsX5d/vZ1tRukT6GTxCZw==	~	

- Are there ways to
  - Improve quality of linkage: Standardize Data, Tune Parameters, Build training data
  - Validate results
  - Monitor for drifts in linkage

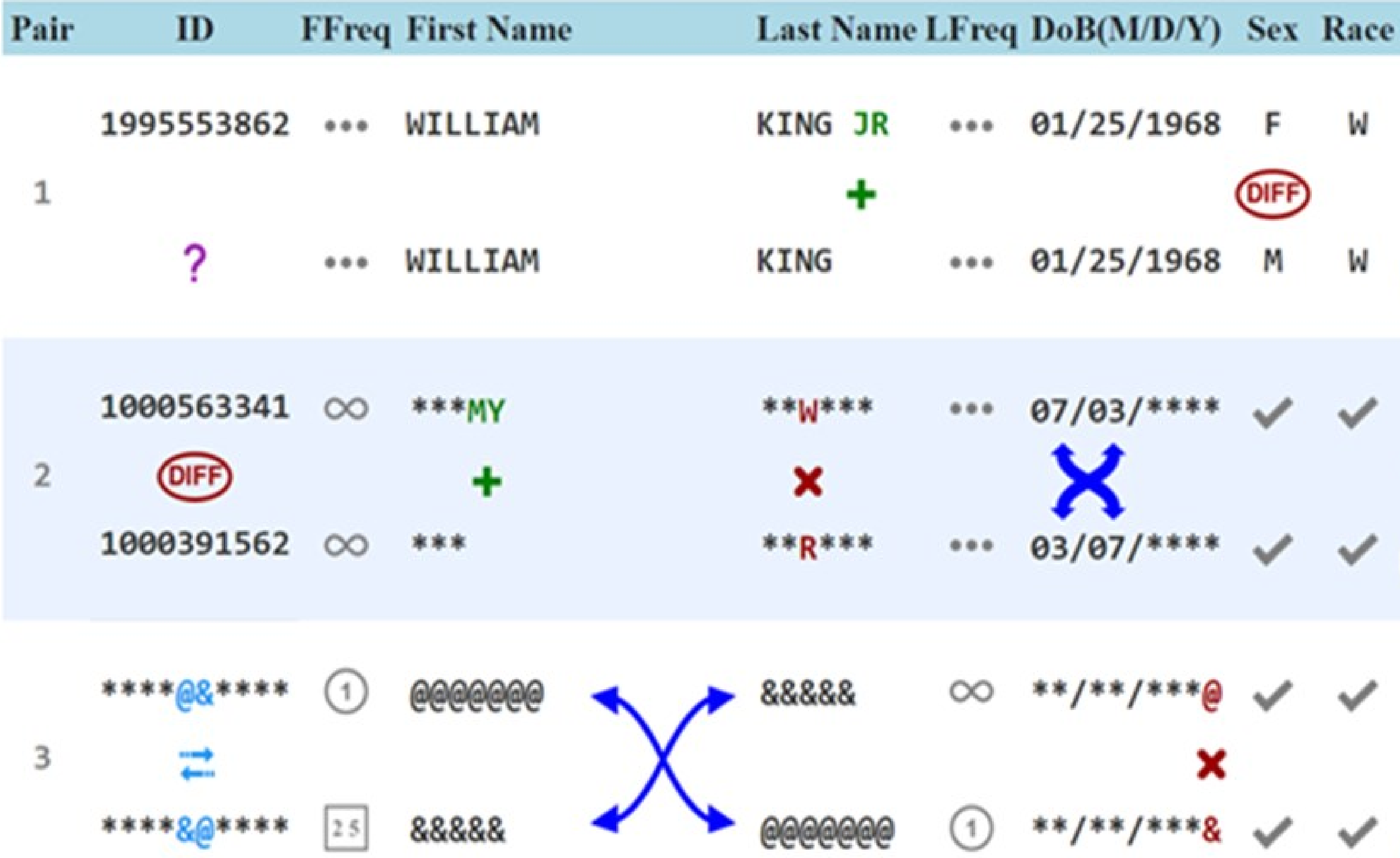
Previous work: What Works Best for Static Interface

- Markup Design



Ragan, E., **Kum, H.-C.,** Ilangovan, G.*, and Wang, H.* (2018). Balancing Privacy and Information Disclosure in Interactive Record Linkage with Visual Masking. Proceedings of the SIGCHI conference on Human factors in computing systems. ***ACM. CHI2018 Honourable Mention Best Paper Award (top 5% of all submissions).*** Also presented at the 14th Symposium on Usable Privacy and Security (SOUPS) Aug 2018 as invited poster.

Previous work: Our approach (static design)
Help people by highlighting differences: Add markup

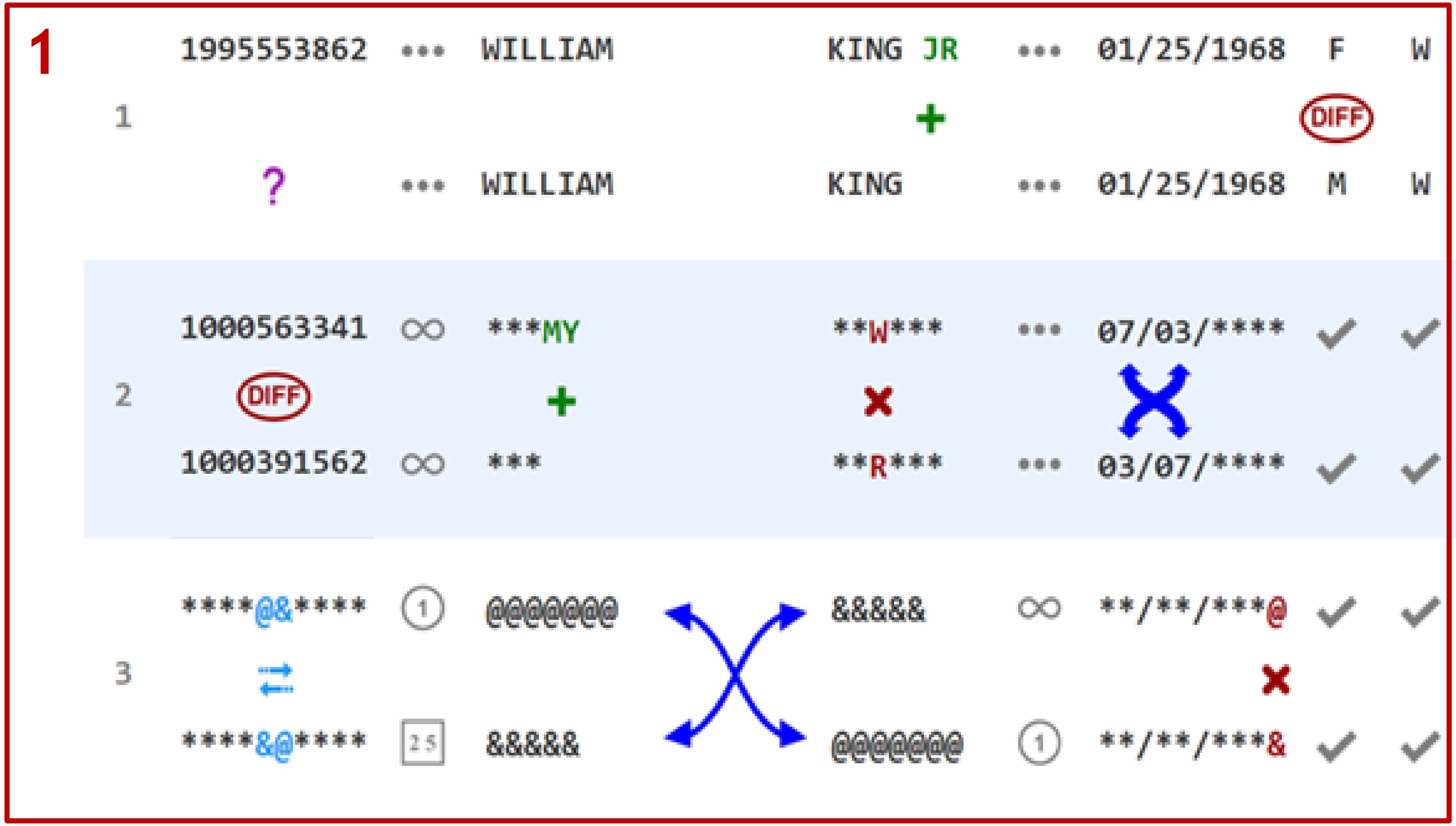


KEY FINDINGS

- High decision quality with only **30%** disclosure with appropriate masks
- Legally deidentified data?
  - Fully masked (0% disclosure) had 75% accuracy
- The quality of human decisions will suffer with low disclosure limits

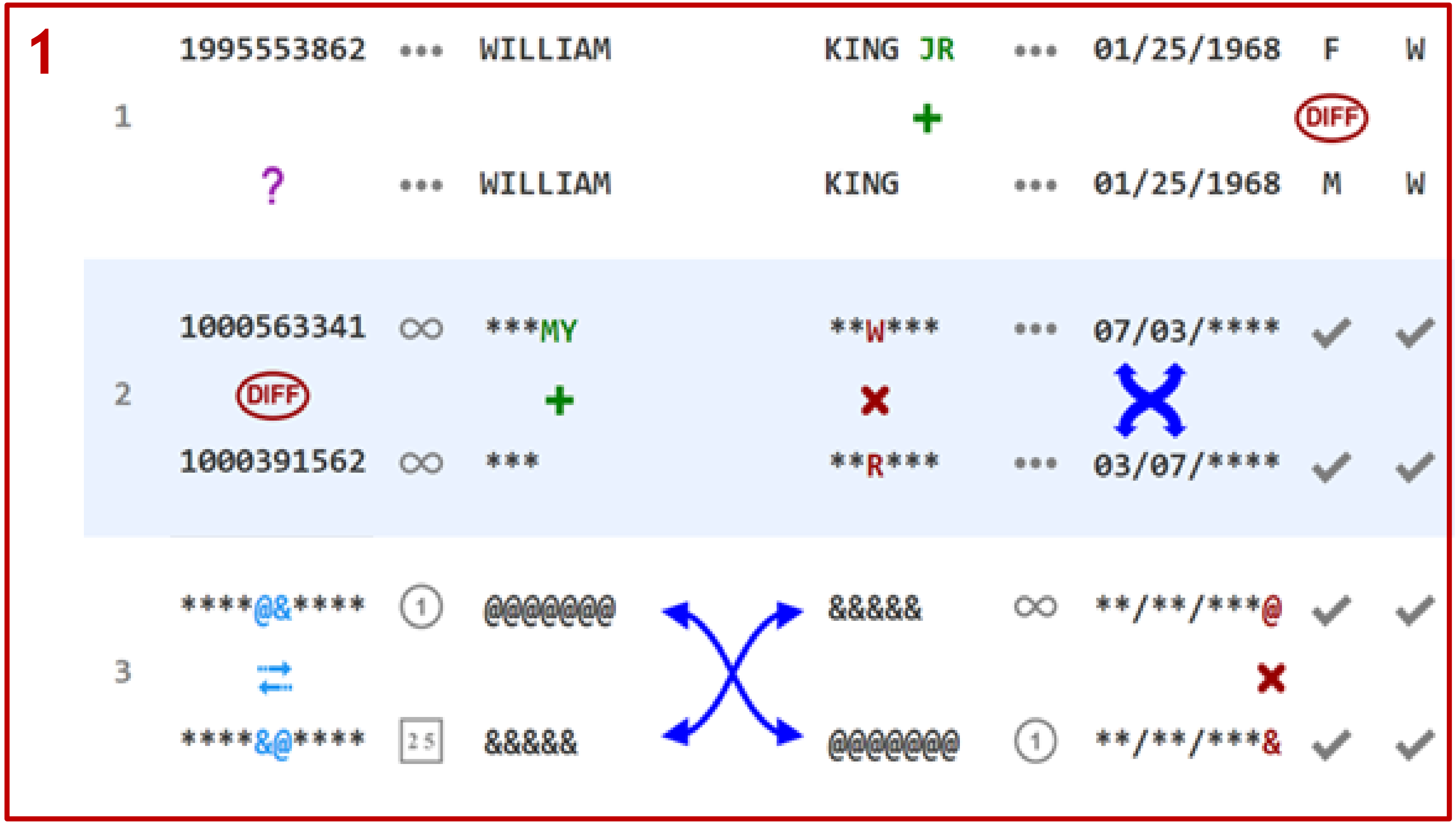
Ragan, E., **Kum, H.-C.**, Ilangovan, G.*, and Wang, H.* (2018). Balancing Privacy and Information Disclosure in Interactive Record Linkage with Visual Masking. Proceedings of the SIGCHI conference on Human factors in computing systems. ***ACM. CHI2018 Honourable Mention Best Paper Award (top 5% of all submissions)***. Also presented at the 14th Symposium on Usable Privacy and Security (SOUPS) Aug 2018 as invited poster.

Proposed Design Elements



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

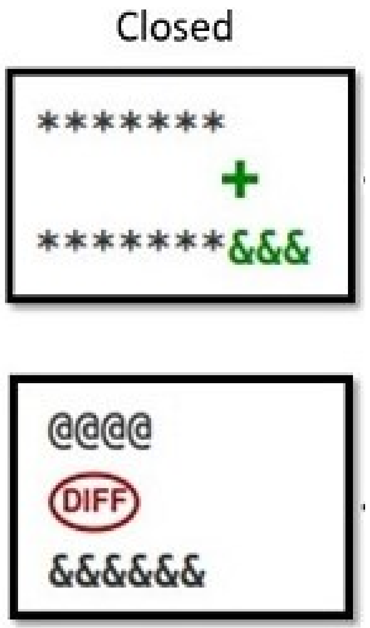


Our Proposed Key Design Elements

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

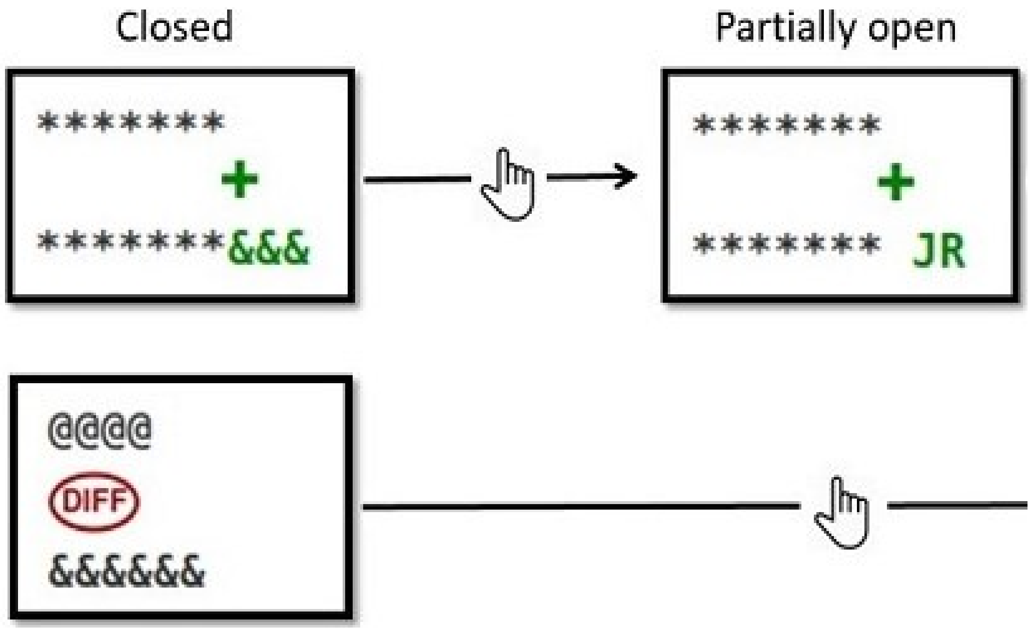
Our proposed approach 1: Interactive Interfaces
Dynamic On-demand Incremental Disclosure

- 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



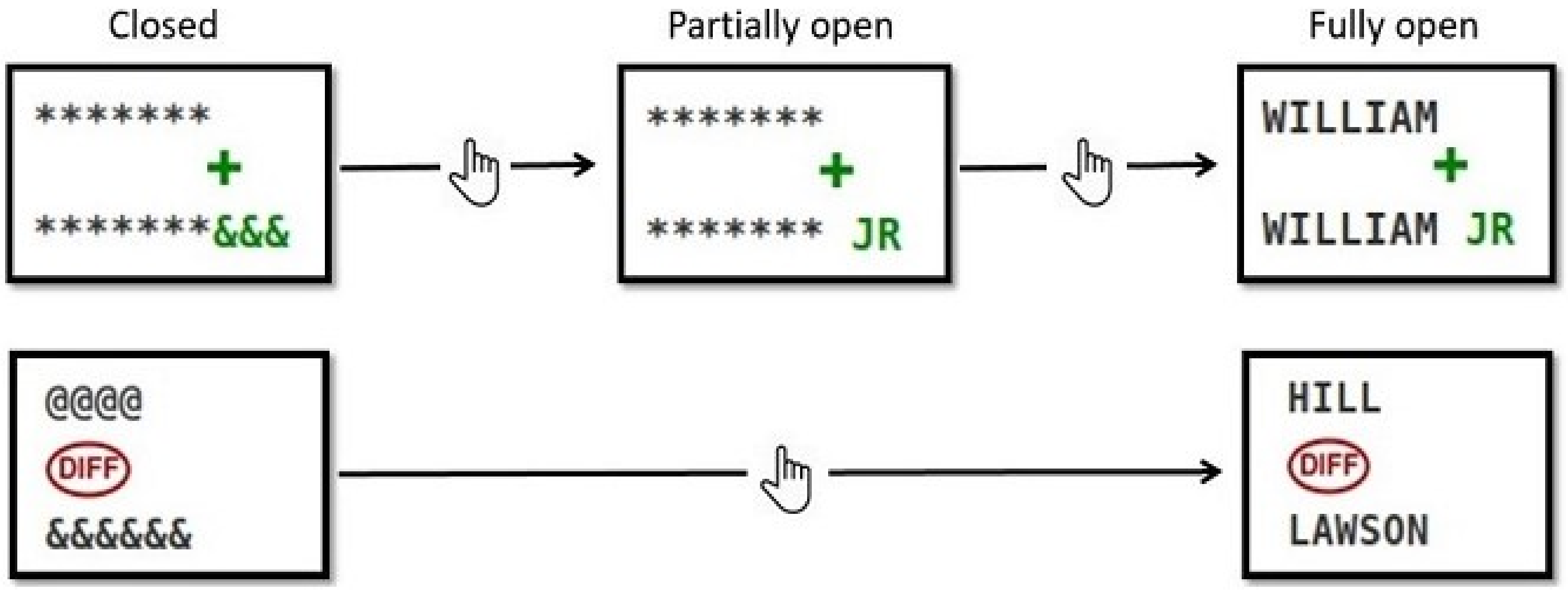
Our proposed approach 1: Interactive Interfaces Dynamic On-demand Incremental Disclosure

- 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



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

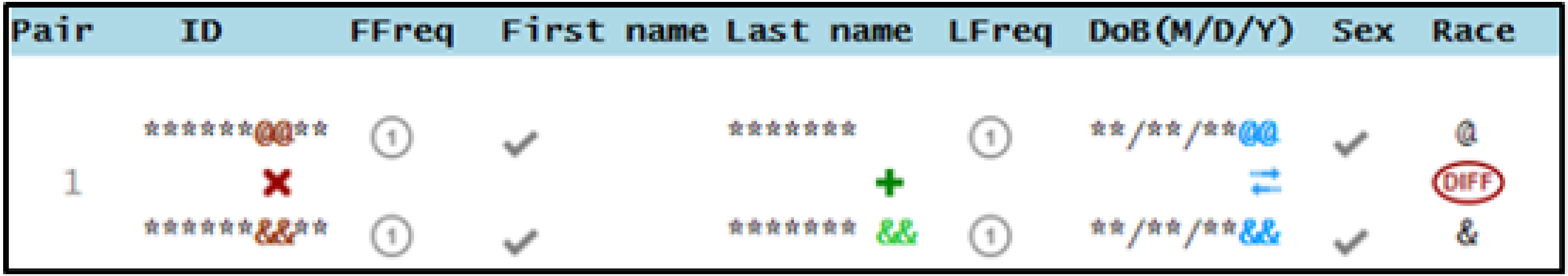


Status quo: Access to ALL

Pair	ID	First name	Last name	DoB(M/D/Y)	Sex	Race
1	8000002767	JUDE	WILLIAM	09/09/1906	M	W
	8000003567	JUDE	WILLIAM JR	09/09/1960	M	B

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

- Incremental disclosure: No Access
  - Start with nothing opened, click to see more



Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
	******@@**	①	✔	*******	①	**/**/**@@	✔	@
1	✘			+		⇄		DIFF
	******&&**	①	✔	******* &&	①	**/**/**&&	✔	&

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

- Incremental disclosure: Partial Information
  - Start with nothing opened, click to see more

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
	******27**	①	✔	*******	①	**/**/**06	M	@
1	✖			+		⇄		DIFF
	******35**	①	✔	******* JR	①	**/**/**60	M	&

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

- Incremental disclosure: Full Access
  - Start with nothing opened, click to see more

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
	8000002767	1	JUDE	WILLIAM	1	09/09/1906	M	W
1	✗			+		⇄		DIFF
	8000003567	1	JUDE	WILLIAM JR	1	09/09/1960	M	B

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

Our approach 2: Accountability & Transparency
Quantify the Risk: Add privacy risk meter

Privacy risk: 6.9% + 3%

- Behavior Triggers, Nudges
- Proactive

Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	******27**	①	✓	WILLIAM	①	09/09/1906	M	W
	✗			+		⇄		DIFF
	******35**	①	✓	WILLIAM JR	①	09/09/1960	M	B
2	✓	①	&&&&&&	@@@@@@@@	①	✓	F	✓
	✓	2-5	@@@@@@@@	&&&&&&	•••	✓	F	✓
3	@@@@@@@@@@@	•••	SALLY	✓	•••	07/04/1960	F	*
	DIFF		DIFF			✗	DIFF	
	&&&&&&&&&&&	∞	JOHN	✓	•••	04/07/1960	M	?

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.

KAPR (k-anonymity privacy risk) score properties
Work in progress

Population Informatics

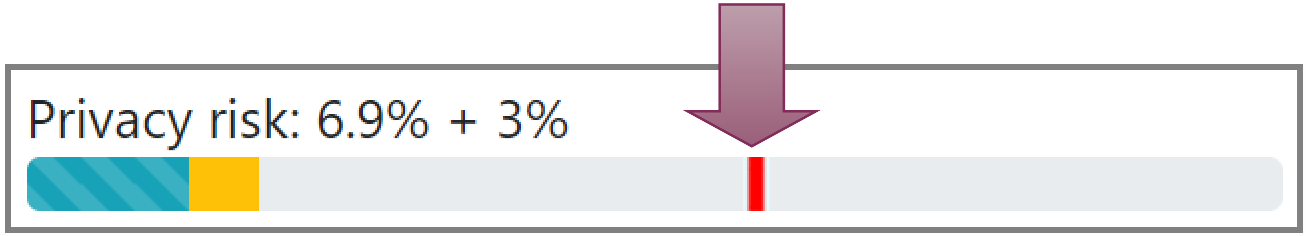
INDIE

- Risk of identity disclosure
- 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.
- Qinbo Li, Adam D'Souza, Cason Schmit, and Hye-Chung Kum. Increasing Transparent and Accountable Use of Data by Quantifying the Actual Privacy Risk in Interactive Record Linkage. Poster presentation at *Proceedings of the AMIA Symposium 2019*, Full technical report available on [arXiv:1906.03345 cs.DB] http://arxiv.org/abs/1906.03345

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

Our approach 3: Accountability & Transparency
Limiting Privacy Risk via Budget: Add limit on meter



- Forced

Pair	ID	FFreq	First name		Last name	LFreq	DoB(M/D/Y)	Sex	Race
	******27**	①	✓		WILLIAM	①	09/09/1906	M	W
1	✗				+		⇄		DIFF
	******35**	①	✓		WILLIAM JR	①	09/09/1960	M	B
	✓	①	&&&&&&	⤨	@@@@@@@@	①	✓	F	✓
2									
	✓	2-5	@@@@@@@@		&&&&&&	•••	✓	F	✓
	@@@@@@@@@@	•••	SALLY		✓	•••	07/04/1960	F	*
3	DIFF		DIFF				✗	DIFF	
	&&&&&&&&&&	∞	JOHN		✓	•••	04/07/1960	M	?

Evaluation: Hypothesis & Experimental Design

Controlled Experiment

- Basics
  - Record linkage task
  - Data: Perturbed from real voter registration data with known ground truth
  - Between-subjects design (5 conditions)
  - Lab study with group sessions
- **122 participants**
- 90 minutes
  - Tutorial
  - Practice trial (36 linkage pairs)
  - Main trials (36 linkage pairs)
  - Additional practice and questionnaires
- **Bonferonni-adjusted $\alpha = 0.0125$**
  - **4 hypothesis tests**

Experimental Design: Five Conditions

Condition	Default Masking	On-demand Interface	Meter & Limit
Fully open	⊘	⊘	⊘
No meter	✓	☝	⊘
Unlimited meter	✓	☝	 ∞
High Limit	✓	☝	
Low limit	✓	☝	

H1: Effects of On-demand Interface

Condition	Default Masking	On-demand Interface	Meter & Limit
Fully open	⃠	⃠	⃠
No meter	✓	☝	⃠

⬆

H1: We hypothesize that an appropriate on-demand and incremental disclosure interface can significantly reduce disclosure without compromising decision quality

H2: Effects of Privacy Risk Meter

Condition	Default Masking	On-demand Interface	Meter & Limit
No meter	✔	👆	⃠
Unlimited meter	✔	👆	∞

H2: The second hypothesis is that the addition of the feedback mechanism, which quantifies and provides a real-time display of consequences of the click, can better inform the decision to access information, and hence encourage only the most needed disclosure

H3: Effects of Pre-specified Budget

Condition	Default Masking	On-demand Interface	Meter & Limit

When providing feedback on disclosure, enforcing a limit on privacy disclosure through a pre-specified budget
will change disclosing behavior to tend toward the given limit ↓

Condition	Default Masking	On-demand Interface	Meter & Limit
Unlimited meter	✓	☝	∞
High Limit	✓	☝	
Low limit	✓	☝	

H3: Effects of Pre-specified Budget
H3.1: Effects of High Pre-specified Budget

Condition	Default Masking	On-demand Interface	Meter & Limit

If the limit is set high, then **higher levels of disclosure will occur**

Unlimited meter	✓		∞
High Limit	✓		

H3: Effects of Pre-specified Budget
H3.2: Effects of Low Pre-specified Budget

On the other hand, if the limit is set too low, disclosure levels will be forced to be lower, but decision quality will be negatively affected

Condition	Default Masking	On-demand Interface	Meter & Limit
High Limit	✓		
Low limit	✓		

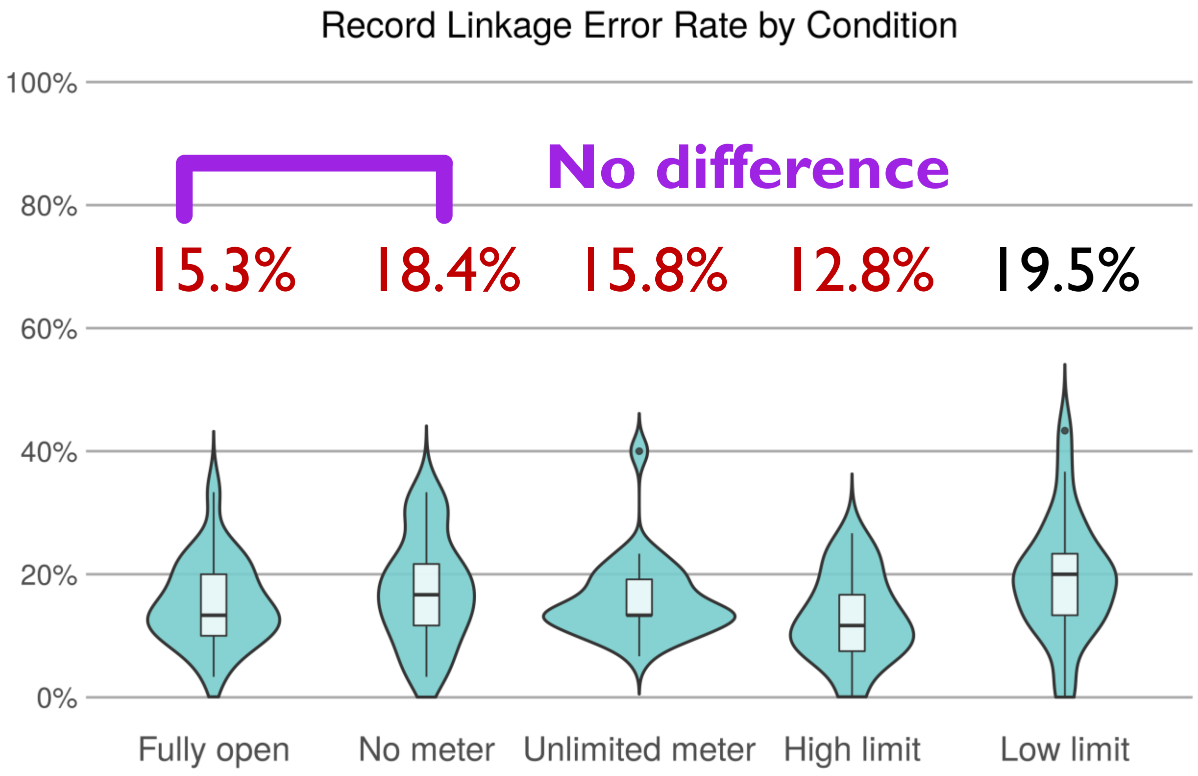
Expert Study

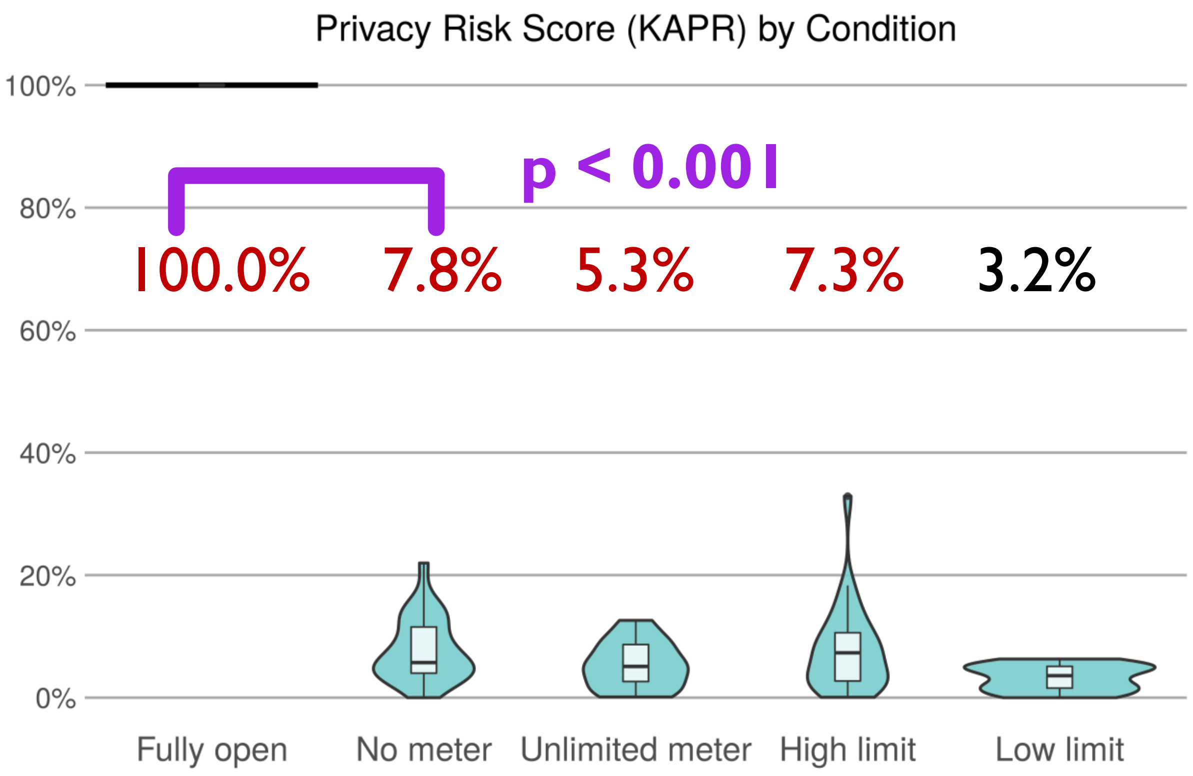
Condition	Default Masking	On-demand Interface	Meter & Limit
High Limit	✔		

- Six experts who regularly conduct record linkage and work with PII (5-10 years of experience)
- All experts completed an abbreviated version of the *high limit* condition
- The experts then answered questions about the potential utility and limitations of the approach and system

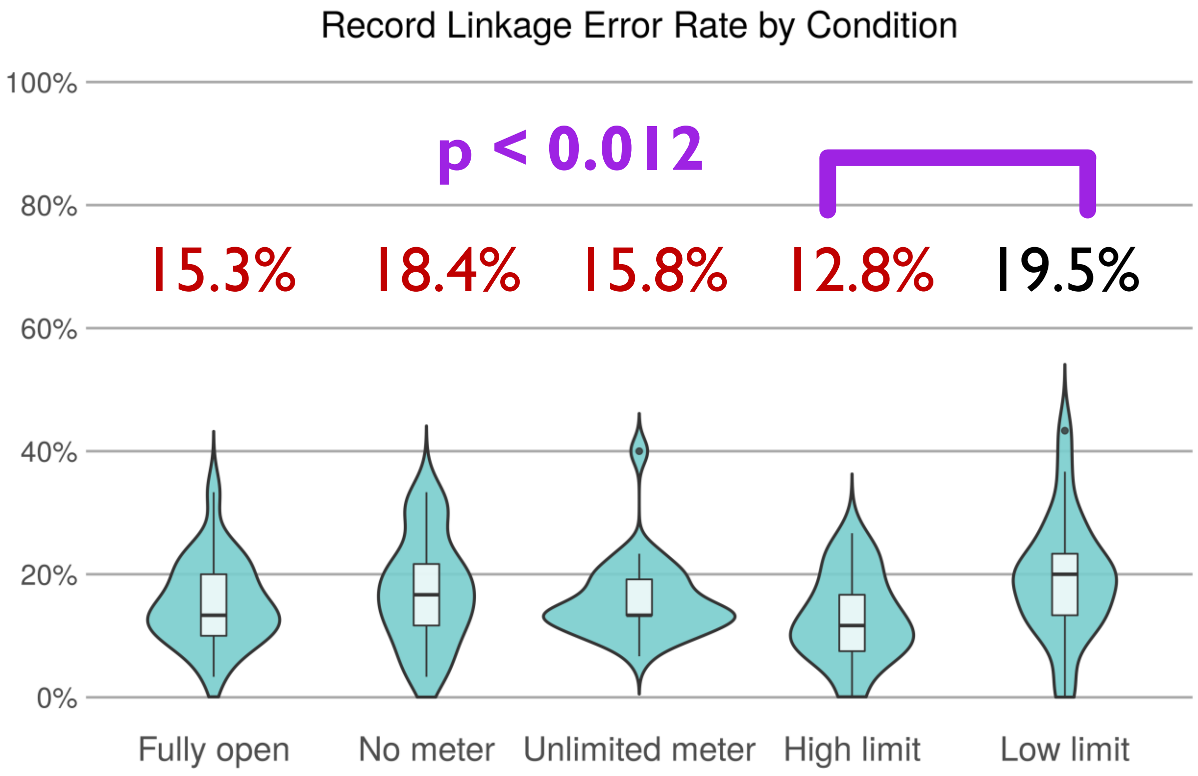
Results

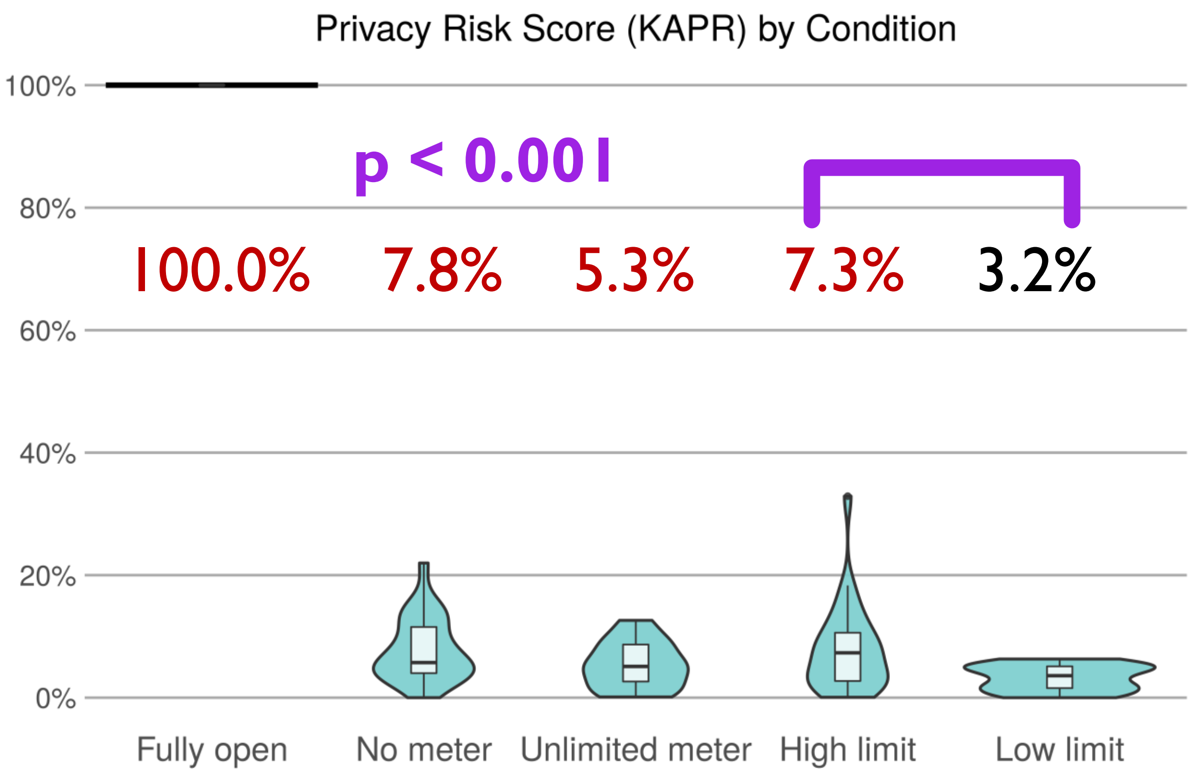
H1: Effects of On-demand Interface





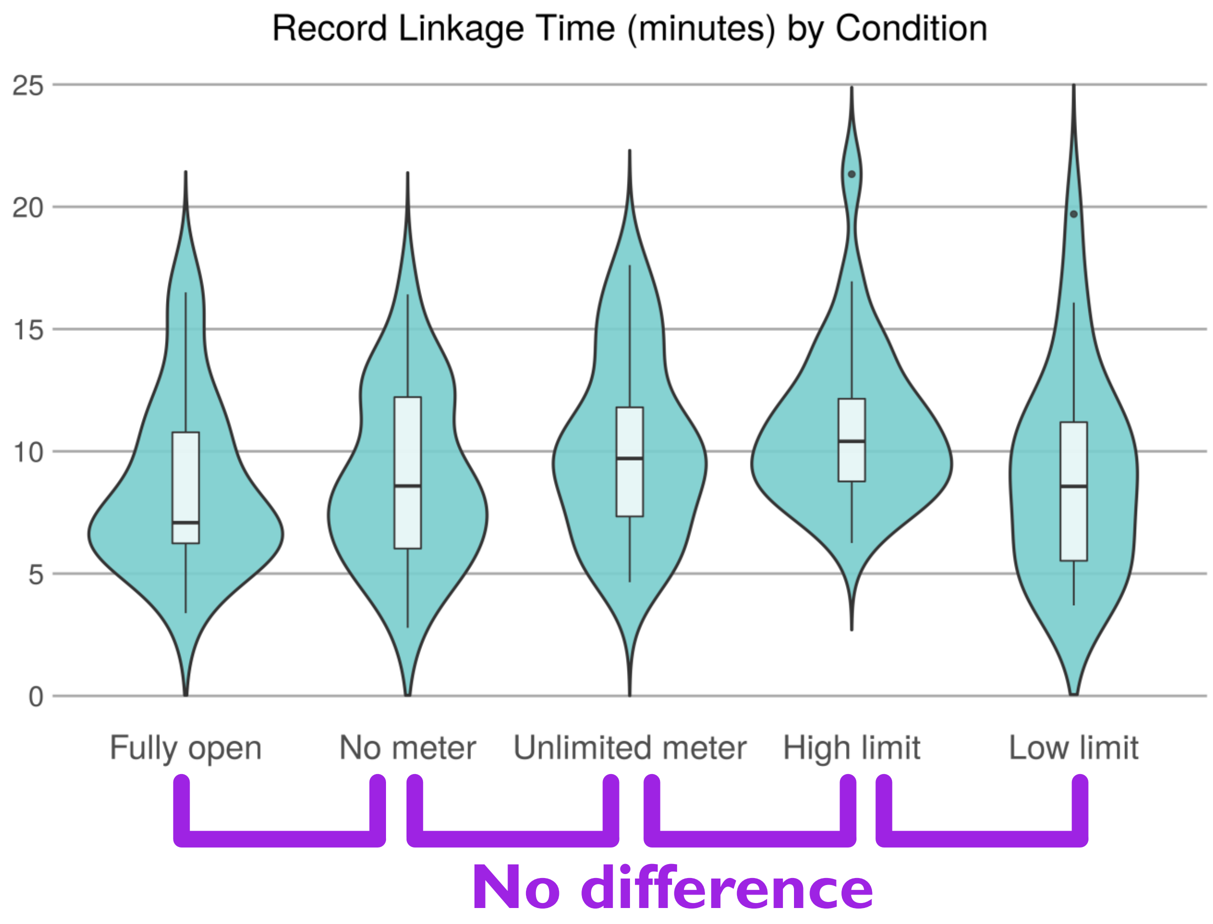
H3: Effects of Privacy Risk Pre-specified Budget
H3.2: Effects of Low Pre-specified Budget





Time to complete the task: 36 pairs

- No significant difference
- Needs further work



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"

Expert Study Results
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.

Highlights on On-Demand & Just-in-Time Interface Model

POPULATION INFORMATICS

INDIE

- 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

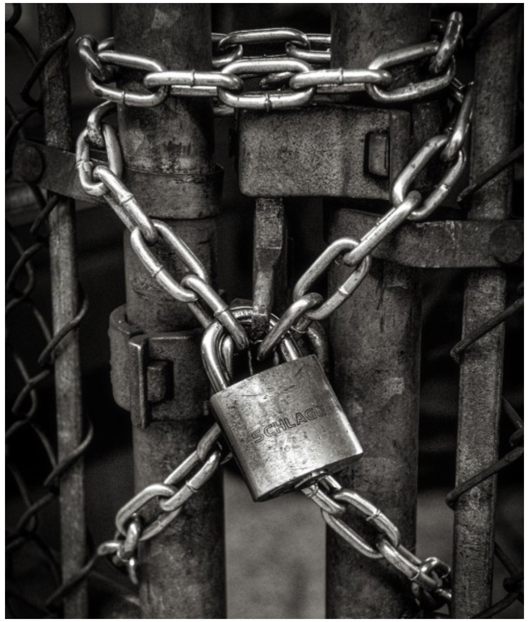
Closing Thoughts

Closing thoughts and discussion on Information Privacy Threat model: Insider threat

- Insider Threat
  - system goals are to minimize any incidental knowledge from legitimate access
  - discourage against access for unauthorized purposes by authorized users
- Incidental
  - MUCH less information disclosed to significantly reduce incidental inferences (e.g. co-workers)
- Negligent (curious but honest)
  - What is the effect of a surveillance camera in discouraging bad behavior?
  - KEY: people must know that their behavior is being recorded AND audited
- Malicious
  - Limitation: Not full guarantees like encryption
  - Some guarantees on total level of disclosure

Key Technology Used in Physical Security Systems

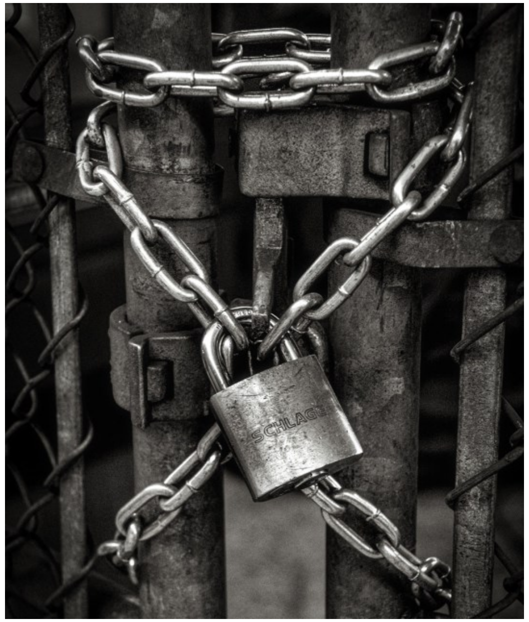
- Locks: Control Access



- Surveillance Camera: monitoring & information accountability/transparency

Tools for information privacy: virtual secure systems
More research needed in CHI

- Locks: Control Access (=Encryption)



- Surveillance Camera (monitoring) (=CHI)
  - **LOGS: How ???**
    - Interactive Interface: Just-in-time incremental access

Team



Research supported by the Patient Centered Outcomes Research Institute ME-1602-34486.

- Hye-Chung Kum: *Population Informatics Lab, Texas A&M University*
- Eric D. Ragan: *INDIE Lab, University of Florida*
- Cason Schmit, JD: *Population Informatics Lab, Texas A&M University*
- Students: *Population Informatics Lab, Texas A&M University*
  - Gurudev Ilangovan
  - Mahin Ramezani
  - Qinbo Li

Thank You!!

Hye-Chung Kum (kum@tamu.edu)

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

Privacy is a BUDGET constrained problem

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