pcornet

ASIS CRG

Autoimmune and Systemic Inflammatory Syndromes
Collaborative Research Group (ASIS CRG)

Privacy Preserving Interactive Record Linkage (PPIRL) via Information Suppression

Hye-Chung Kum (PI)

Alva Ferdinand Eric Ragan Cason Schmit

Population Informatics Lab

https://pinformatics.org/ppirl/index.php

ArthritisPower™

GHLF creakyjoints®

Privacy Preserving Interactive Record Linkage (PPIRL) via Information Suppression

Hye-Chung Kum (PI)

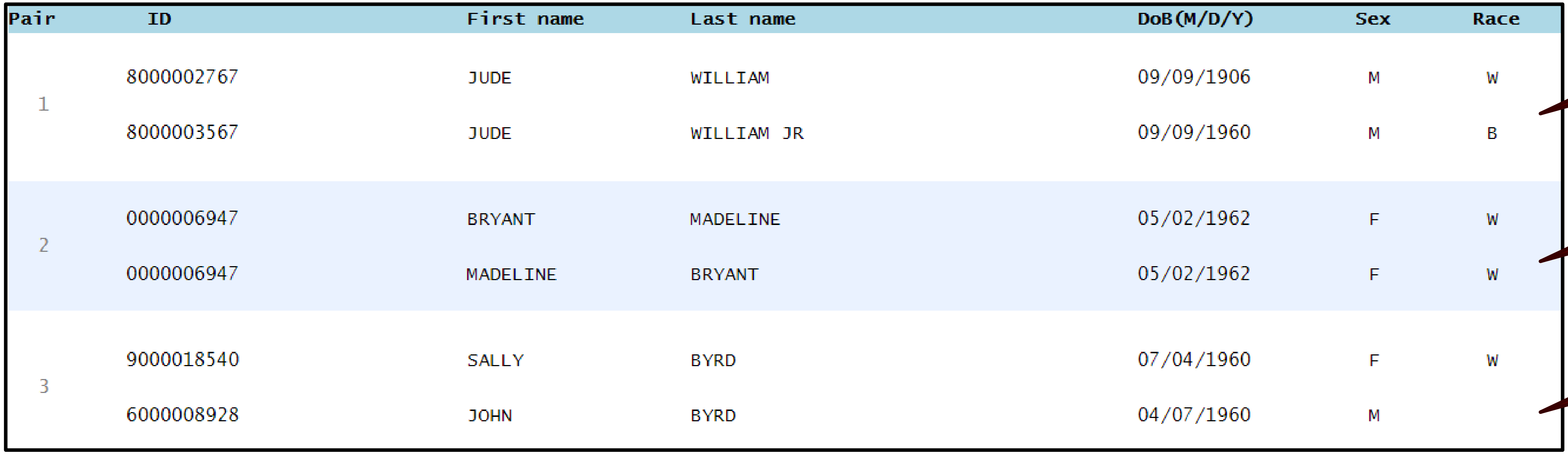
Alva Ferdinand Eric Ragan Cason Schmit

Population Informatics Lab

https://pinformatics.org/ppirl/index.php

Record Linkage: Same or Different People?

- 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

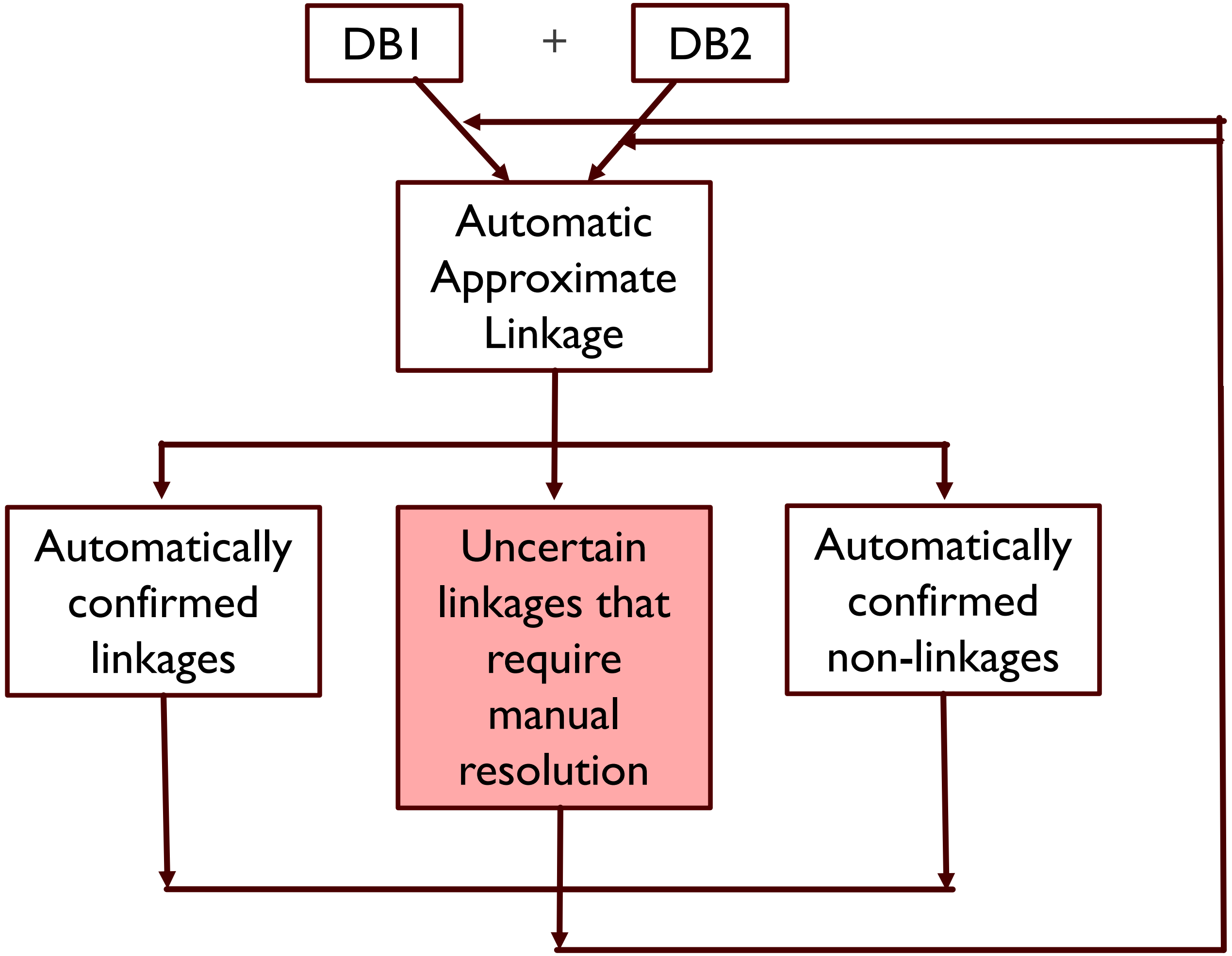
Maybe Twins

Common Issues with Data about People Make Record Linkage Difficult to do Fully Automatically

POPULATION INFORMATICS

- **Data are expressed differently**
  - Nick Names (Elizabeth & Beth)
- **Data change over time**
  - Women get married and change their last name
- **Data are not unique attributes**
  - John Smith (there are different people that have the same name)
  - Twins & Family members have similar identifying information such as DOB & last name
  - Same names in Families with different suffix (Jr and Sr)
- **Data are sometimes missing**
  - SSN are often missing
- **Data have errors**
  - Inserting/deleting extra characters
  - Typing in the wrong character
  - Transposing two characters
  - First name and last name are mixed up
  - Day and month is mixed up

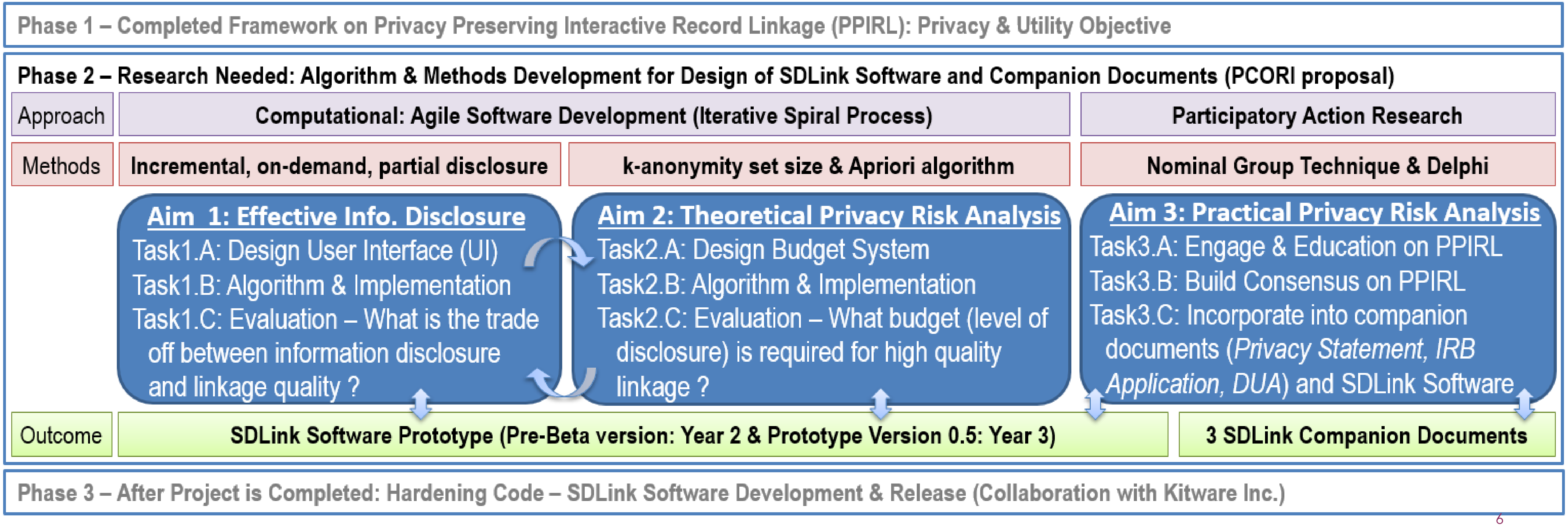
Approximate Record Linkage Human-Computer System


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% automatics
- 15%-25% manual resolution

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 ?

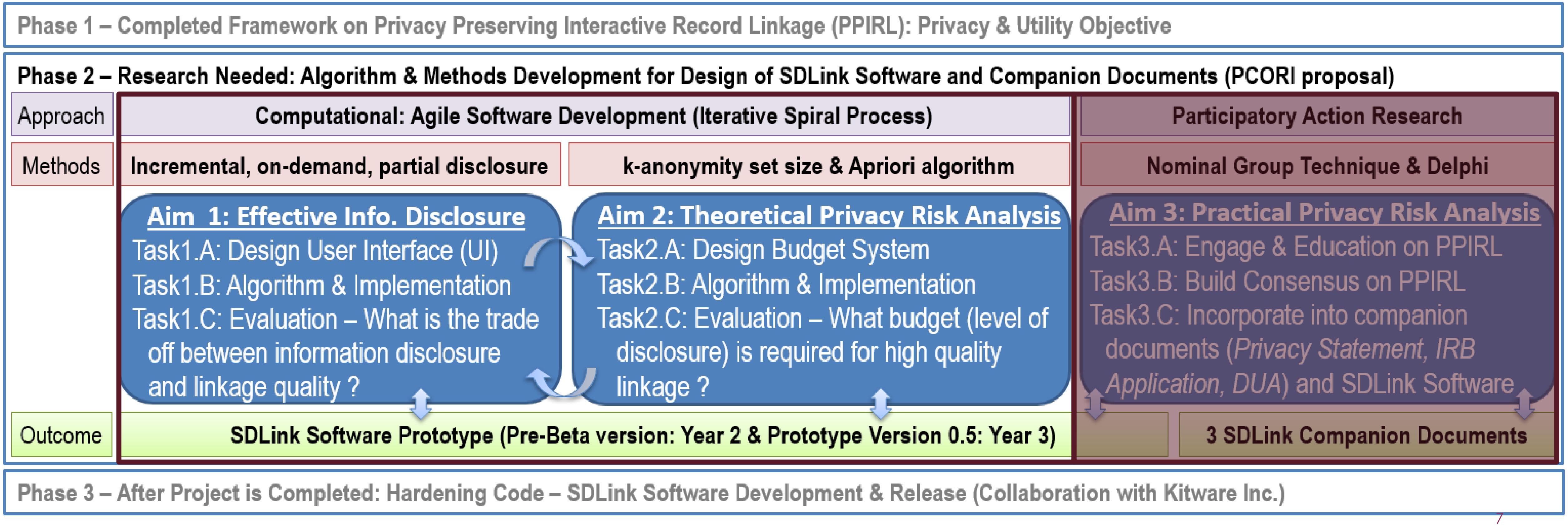
Aim 3: Practical Privacy Risk Analysis
Task3.A: Engage & Education on PPIRL
Task3.B: Build Consensus on PPIRL
Task3.C: Incorporate into companion documents (*Privacy Statement, IRB Application, DUA*) and SDLink Software

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.)

Aims 1&2: Outcomes – Prototype Software Privacy Preserving Interactive Record Linkage (PPIRL)

POPULATION INFORMATICS



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

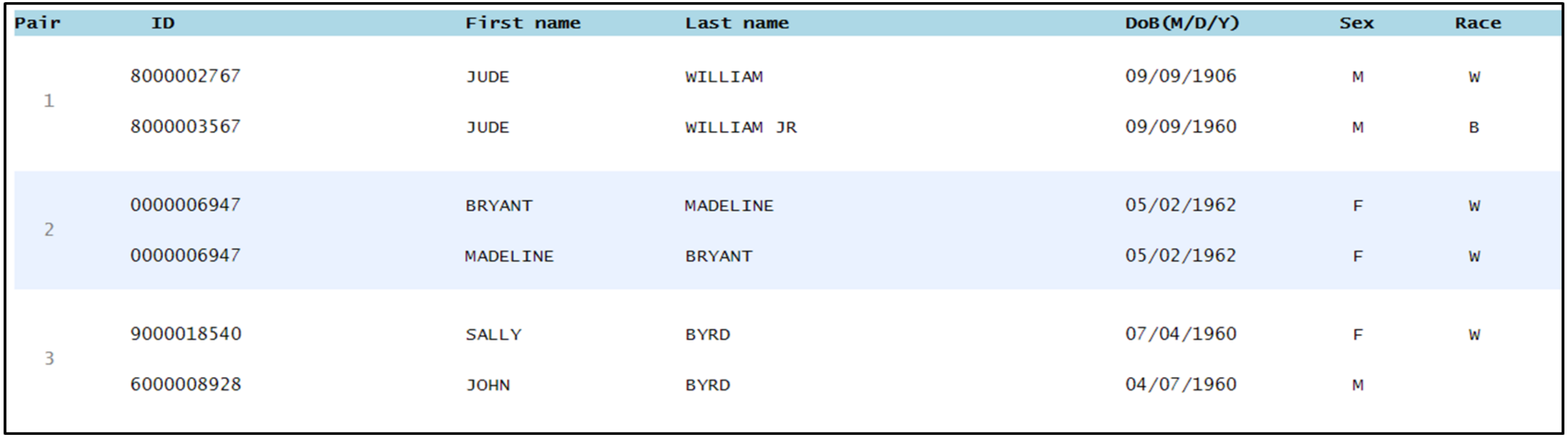
Task3.B: Build Consensus on PPIRL

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.)

Status Quo: Show everything



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	

- Are there ways to enhance privacy during record linkage ?

Information Privacy 101: Point One
Privacy is a BUDGET constrained problem

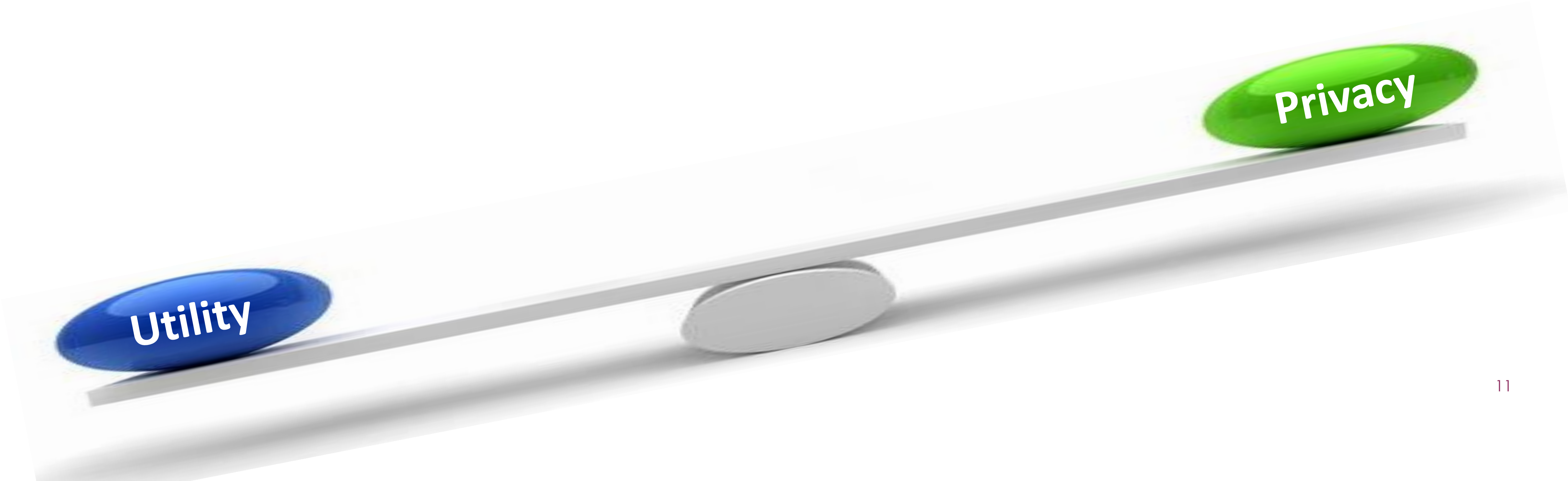
- Differential Privacy proves each query leads to some privacy loss while providing some utility in terms of data analysis
- Current protection mechanism in database research is not effective
  - de-identified data cannot be linked
  - Not sharing enough details: leads to bias, and invalid results
-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

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

Too Focused on Utility

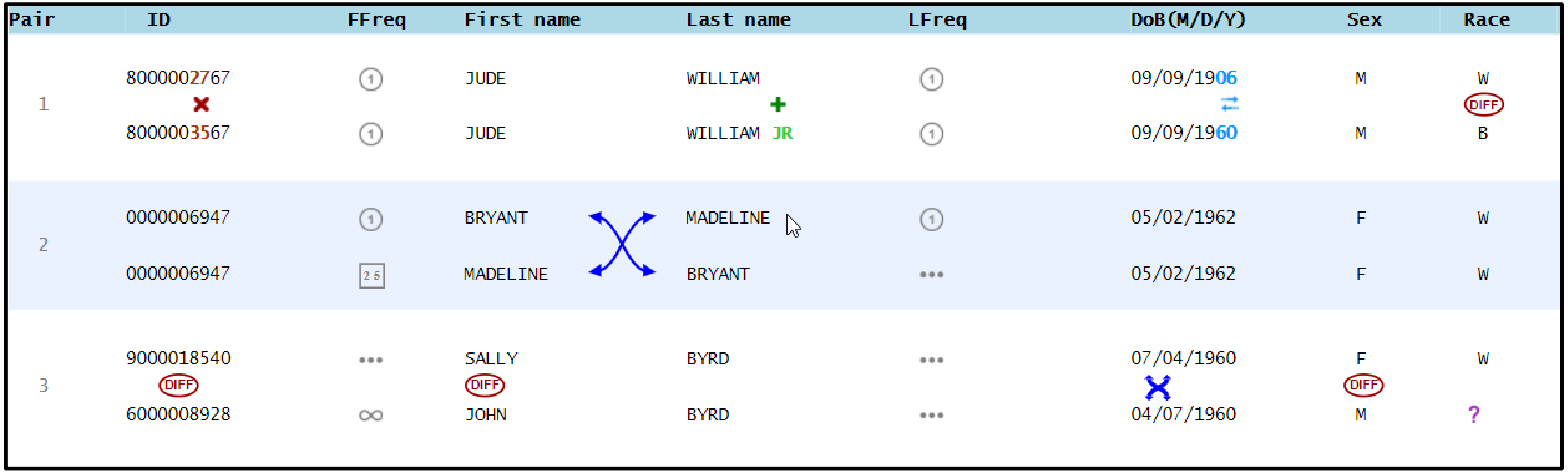
- Unnecessarily exposure, risk



Optimal balance point in record linkage

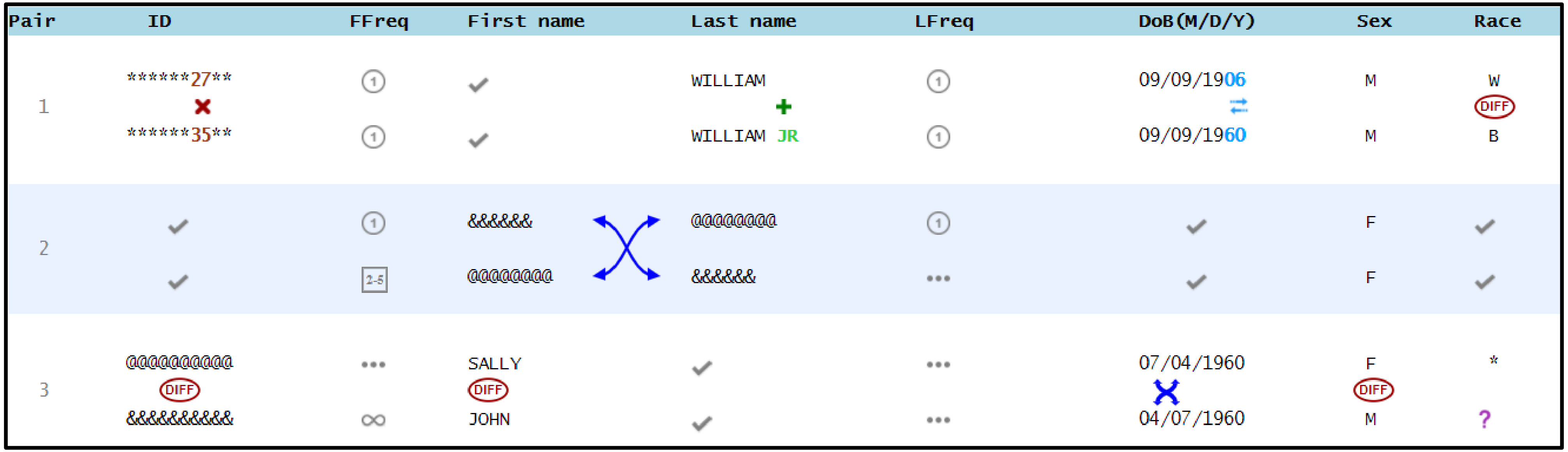
- How can we support projects finding the optimal balance in their projects?

Our approach 1
Help people by highlighting differences: Add markup



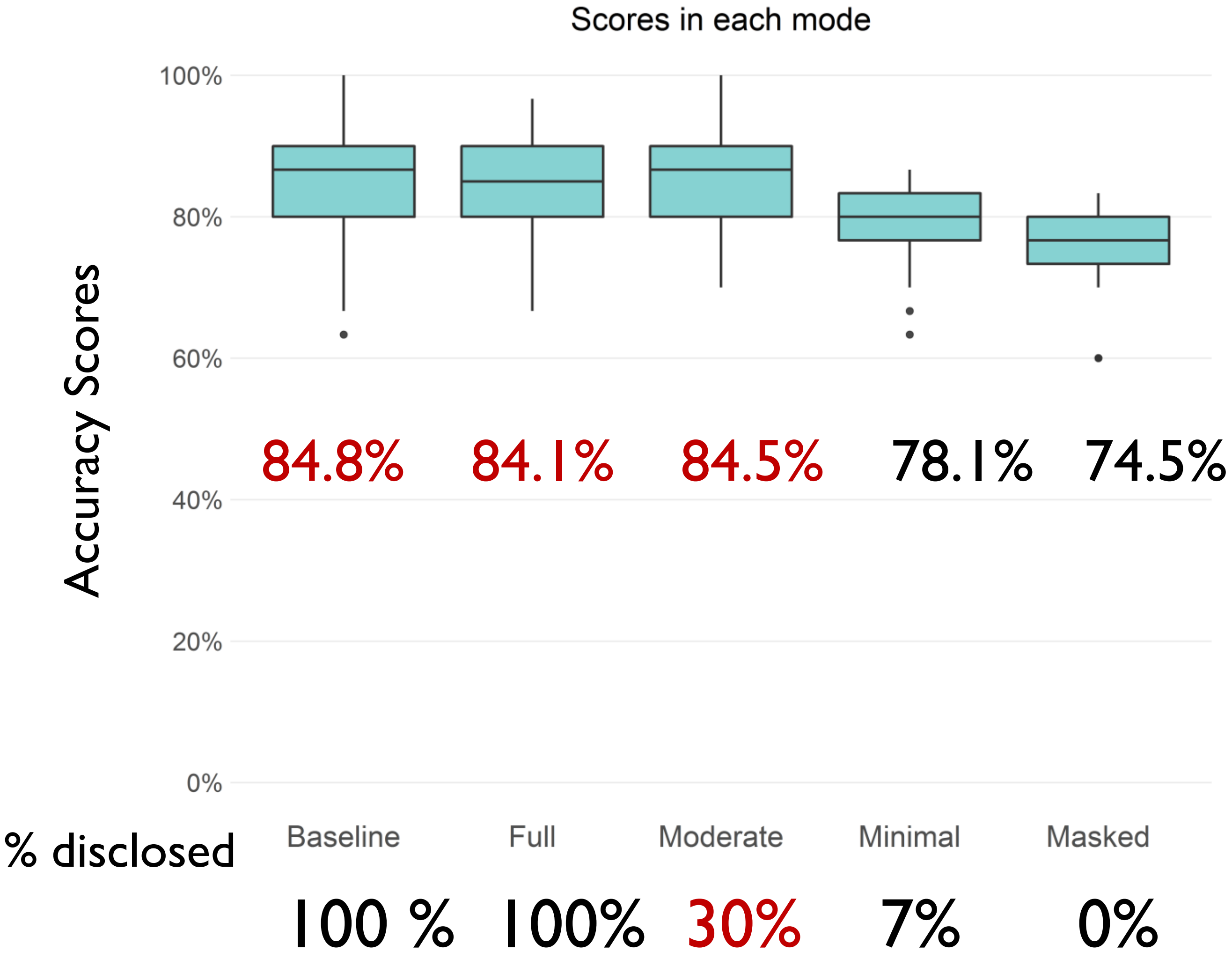
Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	8000002767 ✖ 8000003567	① ①	JUDE JUDE	WILLIAM ✚ WILLIAM JR	① ①	09/09/1906 ⇄ 09/09/1960	M M	W DIFF B
2	0000006947 0000006947	① 25	BRYANT ⤨ MADELINE	MADELINE BRYANT	① •••	05/02/1962 05/02/1962	F F	W W
3	9000018540 DIFF 6000008928	••• ∞	SALLY DIFF JOHN	BYRD BYRD	••• •••	07/04/1960 ✖ 04/07/1960	F DIFF M	W ?

Our approach 2
Minimum Necessary Disclosure



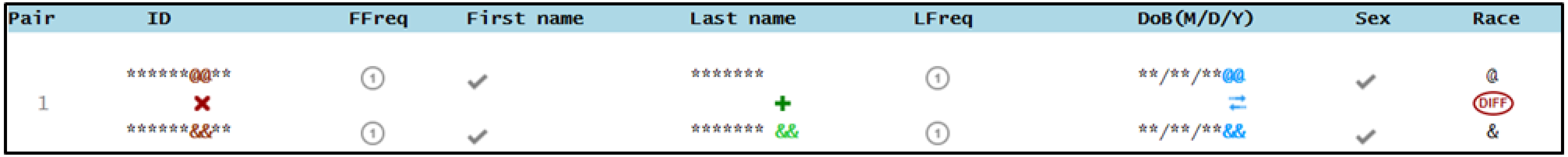
Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
	******27**	①	✓	WILLIAM	①	09/09/1906	M	W
1	✖			✚		⇄		DIFF
	******35**	①	✓	WILLIAM JR	①	09/09/1960	M	B
	✓	①	&&&&&&	@@@@@@@@	①	✓	F	✓
2			⤨					
	✓	2-5	@@@@@@@@	&&&&&&	•••	✓	F	✓
	@@@@@@@@@@	•••	SALLY	✓	•••	07/04/1960	F	*
3	DIFF		DIFF			✖	DIFF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

Accuracy Score by Disclosure Mode



- We can get comparable results to full mode with only 30% disclosure with appropriate masks (moderate mode)
- As we mask more values for privacy, quality of results start to suffer (p<0.001)
- However, even legally de-identified data with proper masks can be linked properly for most situations
  - 0% disclosure still had 75% accuracy
- **Incremental disclosure can significantly improve privacy protection with negligible impact on quality of linkage**

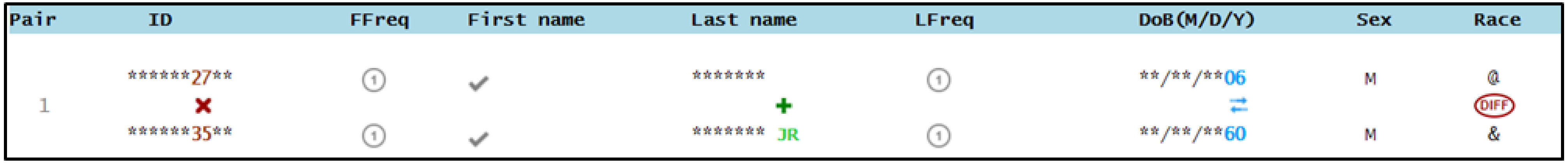
Our approach 3 – Open on Demand
Click to Open: Only Open When Needed for Good Decision



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

Nothing Opened

click

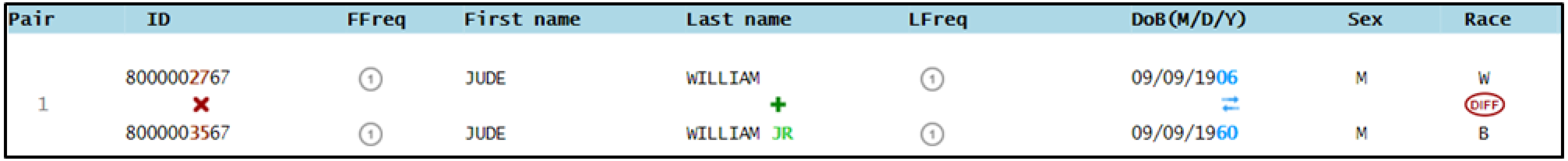


Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1	******27** ✗ ******35**	① ①	✓ ✓	******* + ******* JR	① ①	**/**/**06 ⇄ **/**/**60	M M	@ DIFF &

Partially Opened

That is open only different characters if not too different

click



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

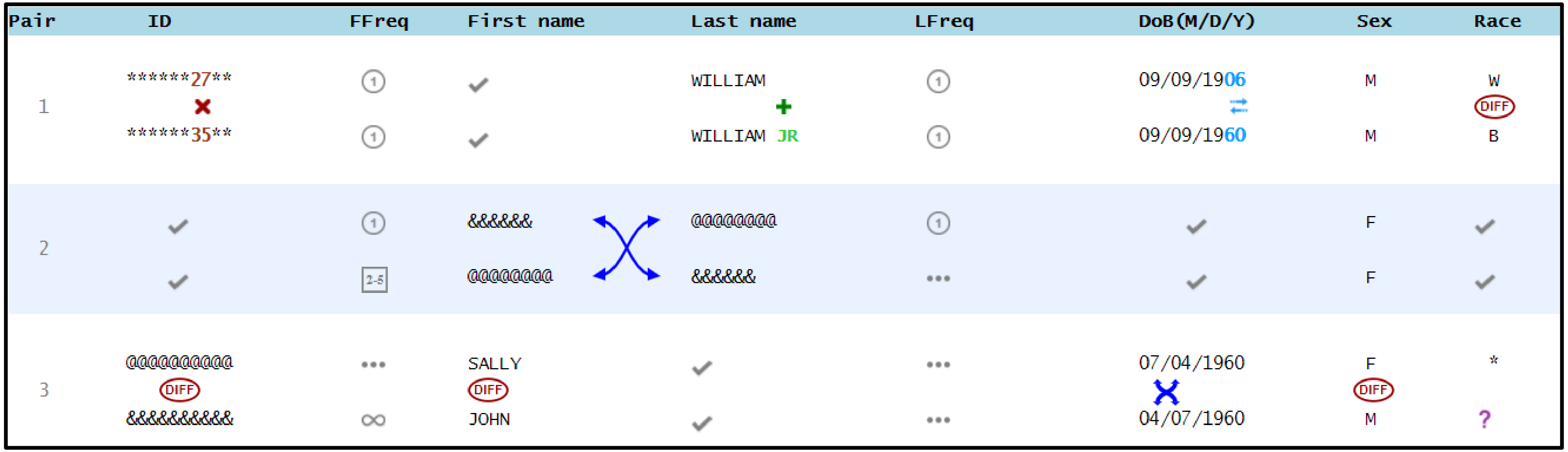
Fully Opened

Information Privacy 101: Point two
Information Accountability (Transparency) Works

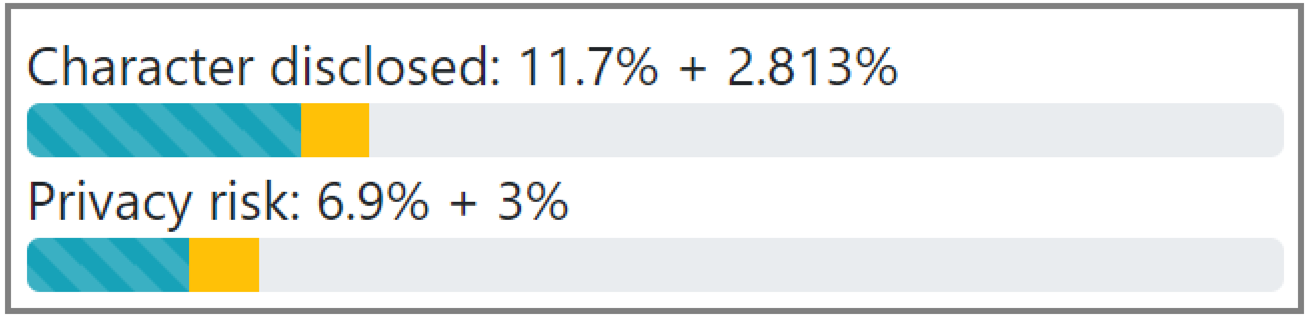
- Secrecy : Hiding information does not support legitimate use
  - In reality, has limited power to protect privacy
  - Severe Consequences related to
    - Accuracy of data and decisions, use of data for
    - legitimate reasons, transparency & democracy
- Information Accountability support effective 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, 2008, pp. 82–87.

Our approach 4
Quantify the Risk: Add privacy risk meter



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	?



Character disclosed: 11.7% + 2.813%

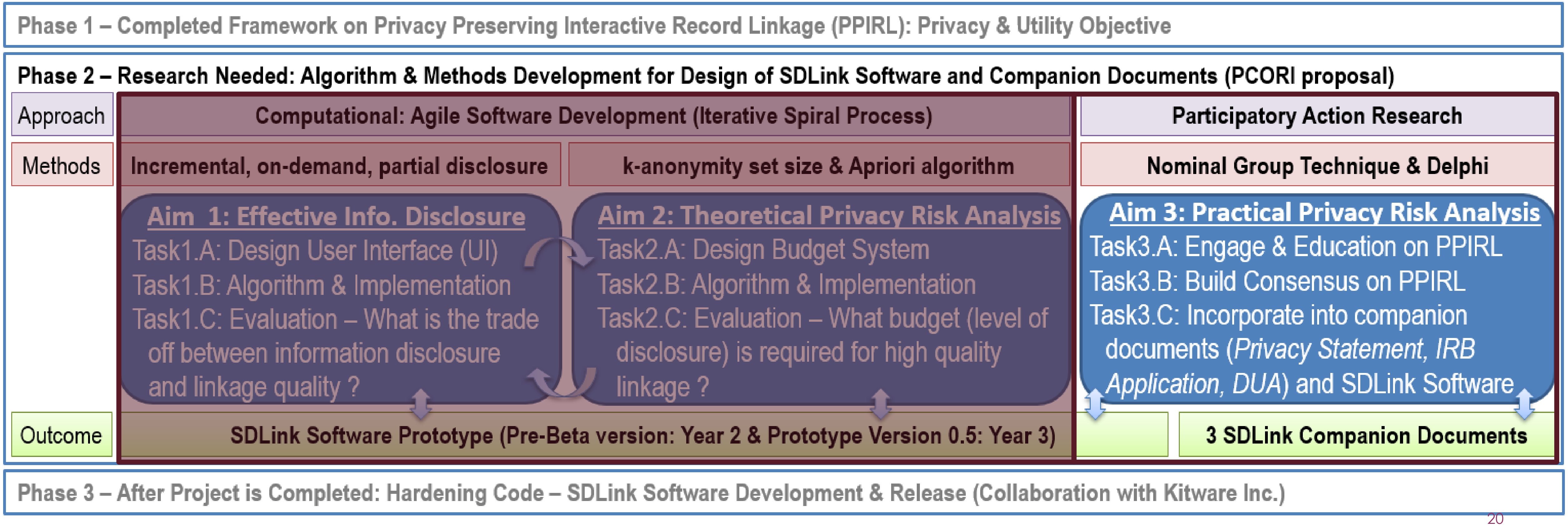
Privacy risk: 6.9% + 3%

- Protection through transparency
  - Measure how much was disclosed
  - And the actual risk of identification that results from the disclosure

Try it!

- http://ppirl-dev.herokuapp.com/
- http://ppirl-tutorial-g.herokuapp.com/

Aim 3 Outcomes: we need your help! 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.)

Aim 3: Companion Documents for the Software Working with patients and stakeholders

- Privacy Statement
  - In lieu of informed consent: Posted on project websites that use the software
  - Simple language to describe how protection is provided when using the software
- Template IRB applications
  - Good IRB language to describe the risk and benefits when using the software
- Template DUA
  - Good legal language to describe the protection provided by the software

ArthritisPower/CreakyJoints and other PPRNs: Privacy Statement

- Help us convey in plain language to patients
  - How use of PPIRL can enhance privacy
  - What potential risk might still remain when using PPIRL
    - Maybe fundamental risk of doing studies that require record linkage
    - How to interrupt the Privacy Risk Score for a project
  - What patients should know about record linkage projects using PPIRL
    - What might you want to see in an informed consent form (if we could have one)?

Acknowledgements

- Ben Nowell, Global Healthy Living Foundation
- Jeffery Curtis, UAB

Thank you

- Participate in our study:
  - 4/27 (Friday): 6-8 pm ET
  - https://ppirl-tutorial.herokuapp.com/
- Stay Informed
  - https://pinformatics.org/ppirl/index.php
- Questions?
  - Hye-chung Kum, kum@tamu.edu