

# Balancing Privacy and Information Disclosure in Interactive Record Linkage with Visual Masking

Hye-Chung Kum<sup>1,3,4</sup>

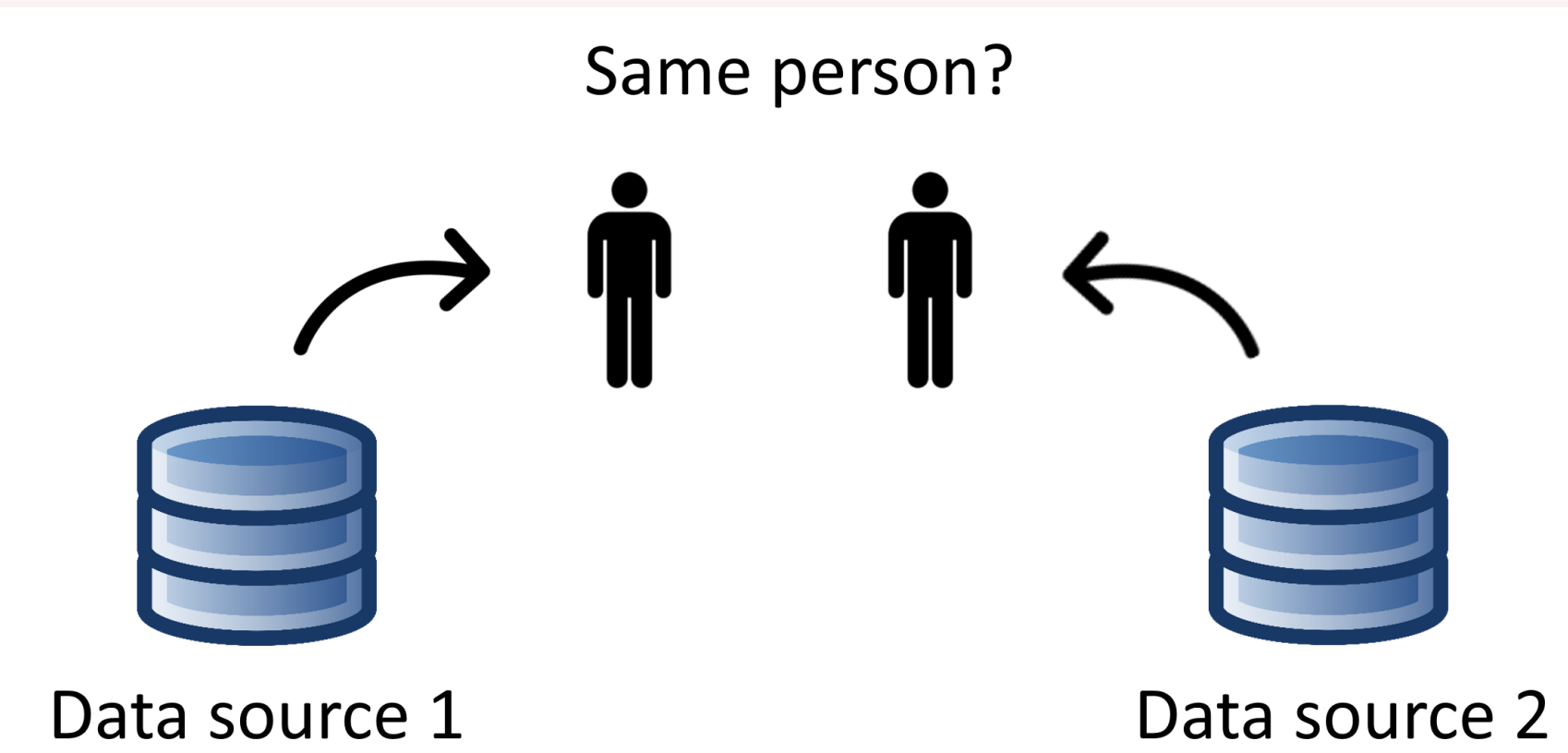
Eric D. Ragan<sup>2,3</sup>

Gurudev Ilangovan<sup>3,4</sup>

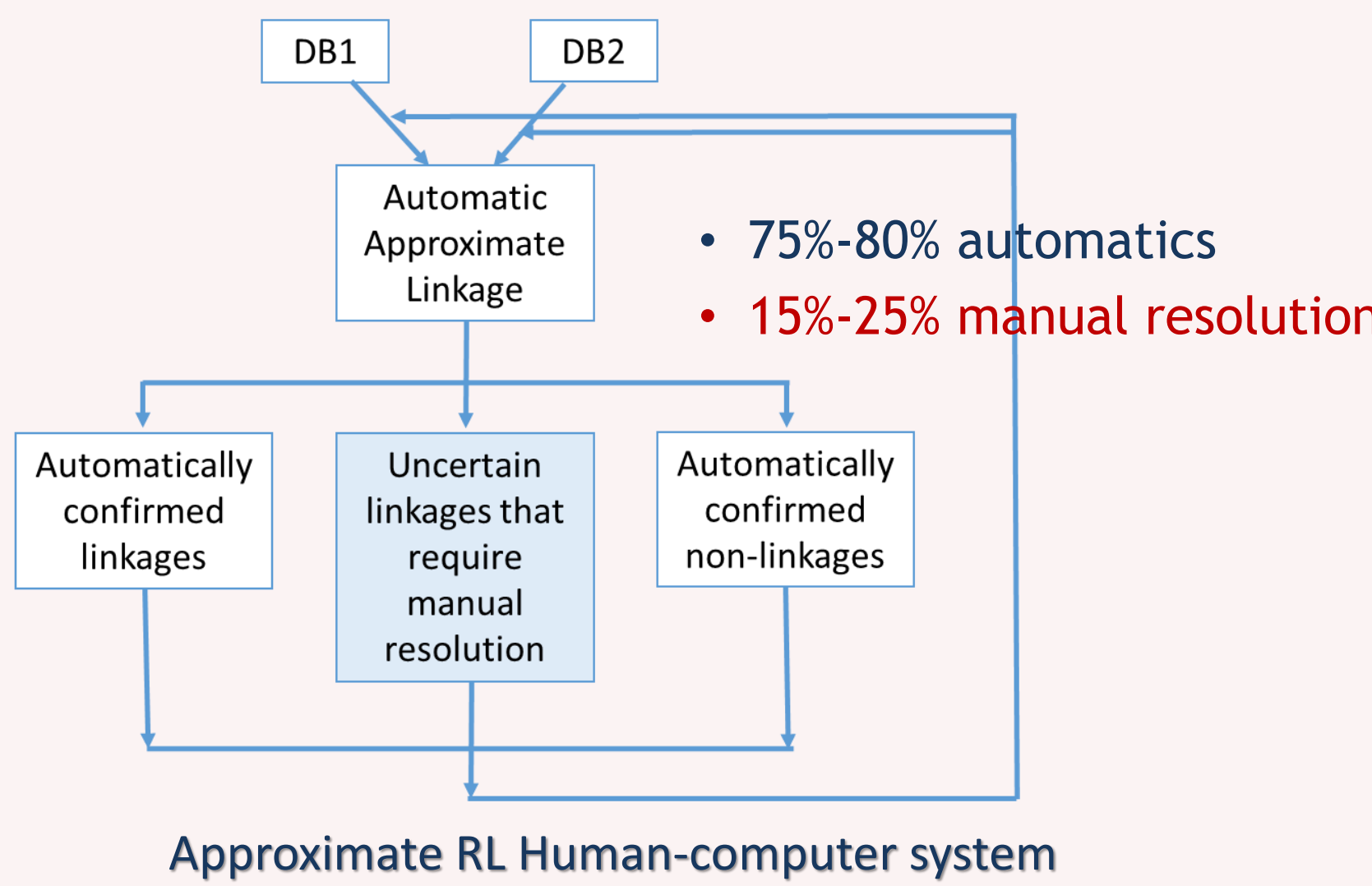
Han Wang<sup>3,4</sup>

<sup>1</sup> Department of Health Policy & Management, <sup>2</sup> Department of Visualization, <sup>3</sup> Department of Computer Science & Engineering, Texas A&M University  
<sup>4</sup> Population Informatics Lab (<https://pinformatics.org/>)

## Interactive Record Linkage



- Common Issues
  - Typos
  - Nicknames
  - Switched characters
  - Name changes
  - Missing values
  - Family members
- Uncertainty in data
  - Requires Human Judgement
  - Human Interaction With Data
  - Standardize Data
  - Clean Data
  - Build Training Data
  - Tune Model Parameters



## Research Overview

- Goals:
  - Limit disclosure of personal information
  - Don't reduce human effectiveness
- Method:
  - Hide data values (when possible)
  - Add visual meta-data to help decision making

## Privacy vs. Utility



- How much can we hide data values without sacrificing decision quality?

## Visual Markup

- Highlight discrepancies
- Empty fields
  - Different characters
  - Extra characters
  - Transposed values
  - Name or Date Swaps
  - Major field differences
- Name frequency meta-data
- Unique
  - Rare
  - Common
  - Highly common

Pair	ID	FFreq	First name	Last name	LFreq	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	MADLINE	○	05/02/1962	F	W
	0000006947	○	BRYANT	MADLINE	○	05/02/1962	F	W
3	9000018540	○	SALLY	BIRD	○	07/04/1960	F	W
	6000008928	○	JENN	BIRD	○	04/07/1960	M	?

## Data Masking

Hide data details for privacy

- Same values
- && Different values

ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
1990443570	○	BOYLE	JASON	○	11/14/1980	M	W
1990443570	○	JASON	BOYLE	○	11/14/1980	M	W
1000027594	○	CHARLES	GREEN	○	07/10/1930	M	W
?	○	CHARLES	GREEN	○	07/10/1903	M	W

ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
✓	○	XXXX	XXXX	○	✓	M	✓
✓	○	XXXX	XXXX	○	✓	M	✓
*****	○	✓	✓	○	07/10/1930	M	✓
?	○	✓	✓	○	07/10/1903	M	✓

## Experiment

- 104 participants
- 90 minutes
  - Tutorial
  - Main trials (36 linkage pairs)
  - Additional practice and questionnaires

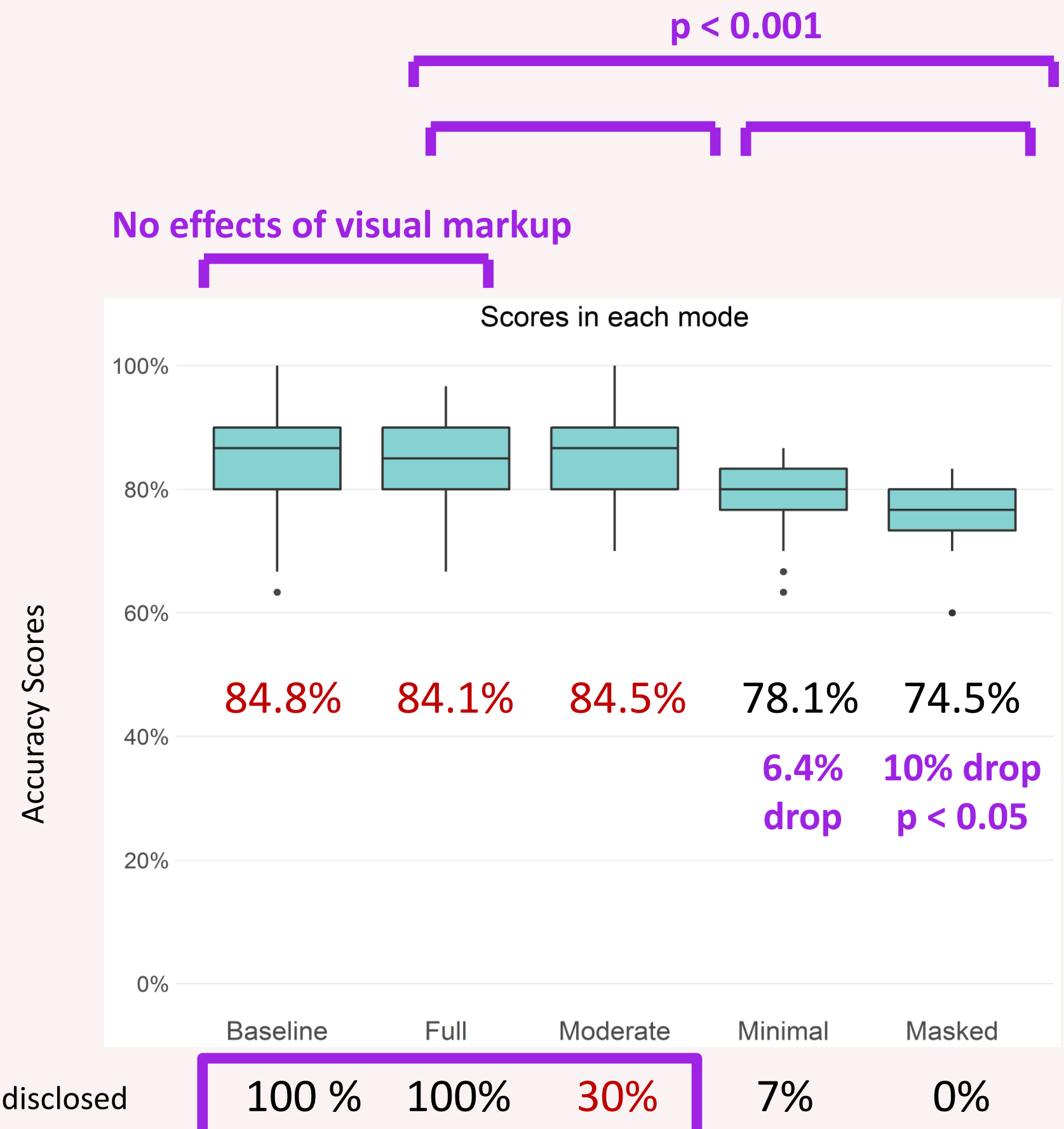
Pair	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race	Choice Panel
1	8000002767	○	JUDE	WILLIAM	○	09/09/1906	M	W	○
	8000003567	○	JUDE	WILLIAM JR	○	09/09/1960	M	B	○
2	0000006947	○	BRYANT	MADLINE	○	05/02/1962	F	W	○
	0000006947	○	BRYANT	MADLINE	○	05/02/1962	F	W	○
3	9000018540	○	SALLY	BIRD	○	07/04/1960	F	W	○
	6000008928	○	JENN	BIRD	○	04/07/1960	M	?	○

## Experimental Design

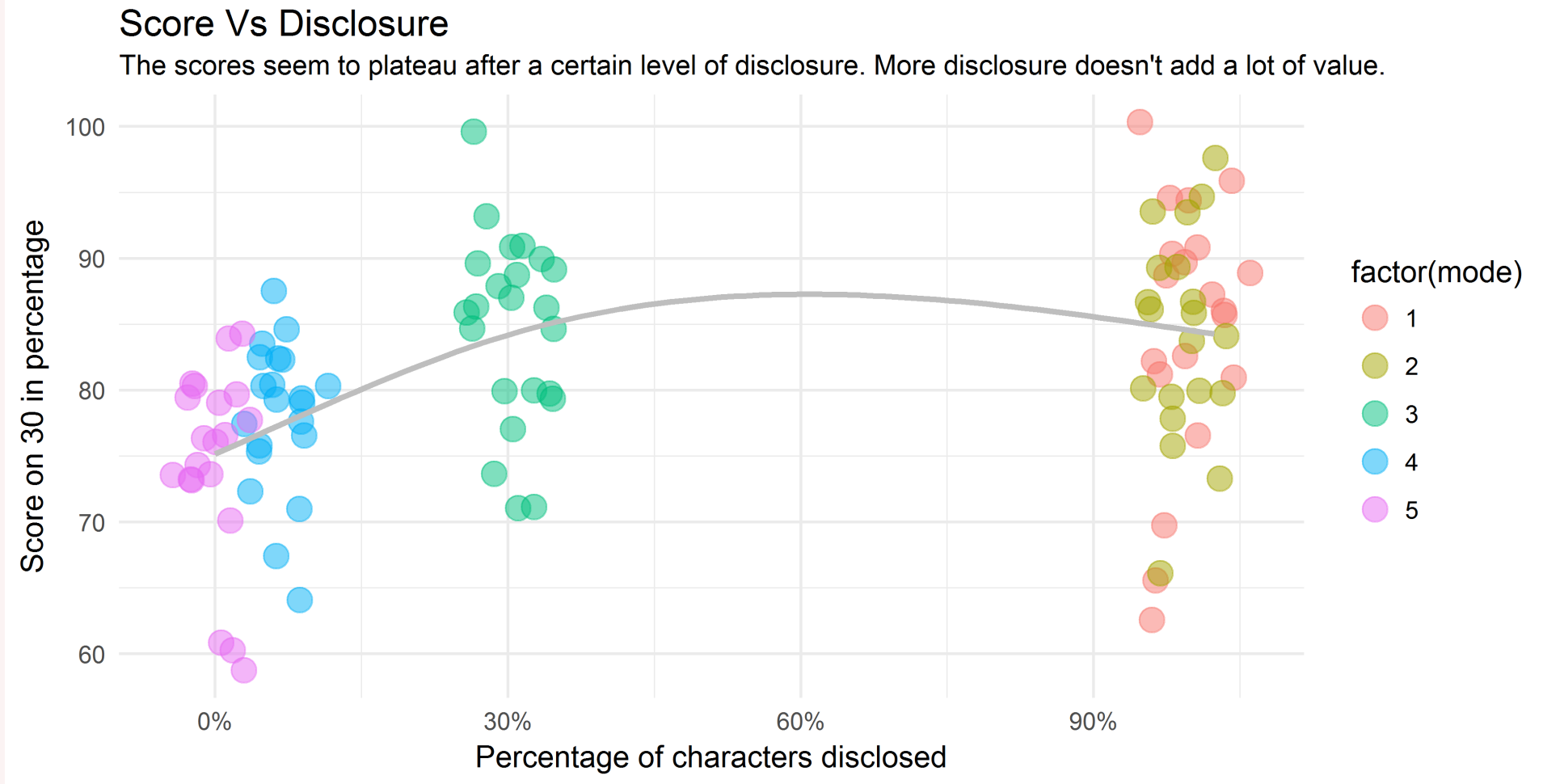
- Data: Perturbed from real voter registration data with known ground truth
- Between-subjects design (x5 conditions)
- Lab study with group sessions

Baseline	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
	8000002767	○	JUDE	WILLIAM	○	09/09/1906	M	W
	8000003567	○	JUDE	WILLIAM JR	○	09/09/1960	M	B
Full	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
	8000002767	○	JUDE	WILLIAM	○	09/09/1906	M	W
	8000003567	○	JUDE	WILLIAM JR	○	09/09/1960	M	B
Moderate	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
	*****2767	○	✓	WILLIAM	○	09/09/1906	M	W
	*****3567	○	✓	WILLIAM JR	○	09/09/1960	M	B
Low	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
	*****2767	○	✓	*****	○	09/09/1906	M	W
	*****3567	○	✓	*****	○	09/09/1960	M	B
Masked	ID	FFreq	First name	Last name	LFreq	DoB(M/D/Y)	Sex	Race
	*****2767	○	✓	*****	○	09/09/1906	M	W
	*****3567	○	✓	*****	○	09/09/1960	M	B

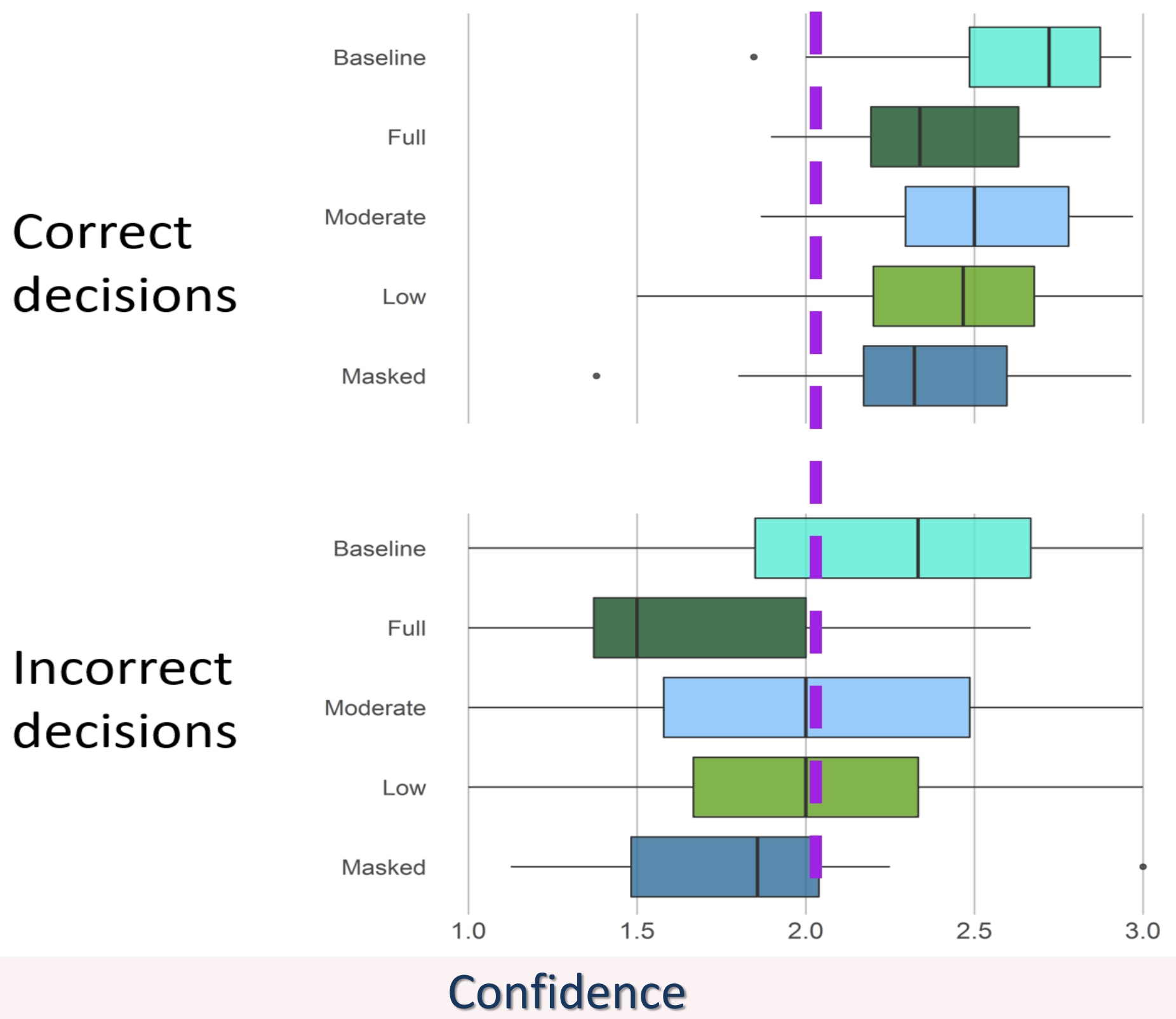
## Results



- 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



## Accuracy Score by Disclosure Mode



## Conclusion

- For legitimate data work such as data integration and verification using PII data, different people need to have access to personal information, which sacrifices the personal privacy of those whose data is stored.
- Often, the primary methods for handling privacy concerns are either to restrict data access at the expense of data utility, or to open the data to more people to improve throughput and utility at the expense of reduced privacy.
- Our study results demonstrate that it is possible to significantly reduce PII disclosure without noticeably affecting decision accuracy with appropriate meta-data.
- Moreover, when legal requirements only allow for deidentified data access, use of well-designed interface can significantly improve data utility.

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