

Privacy Concerns -> Social Science Research

Frauke Kreuter

JPSM – Uni Mannheim – IAB

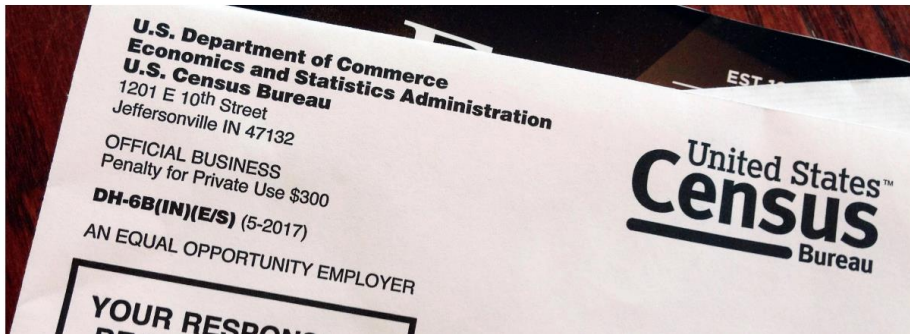
@fraukolos

Buzz

TheUpshot

To Reduce Privacy Risks, the Census Plans to Report Less Accurate Data

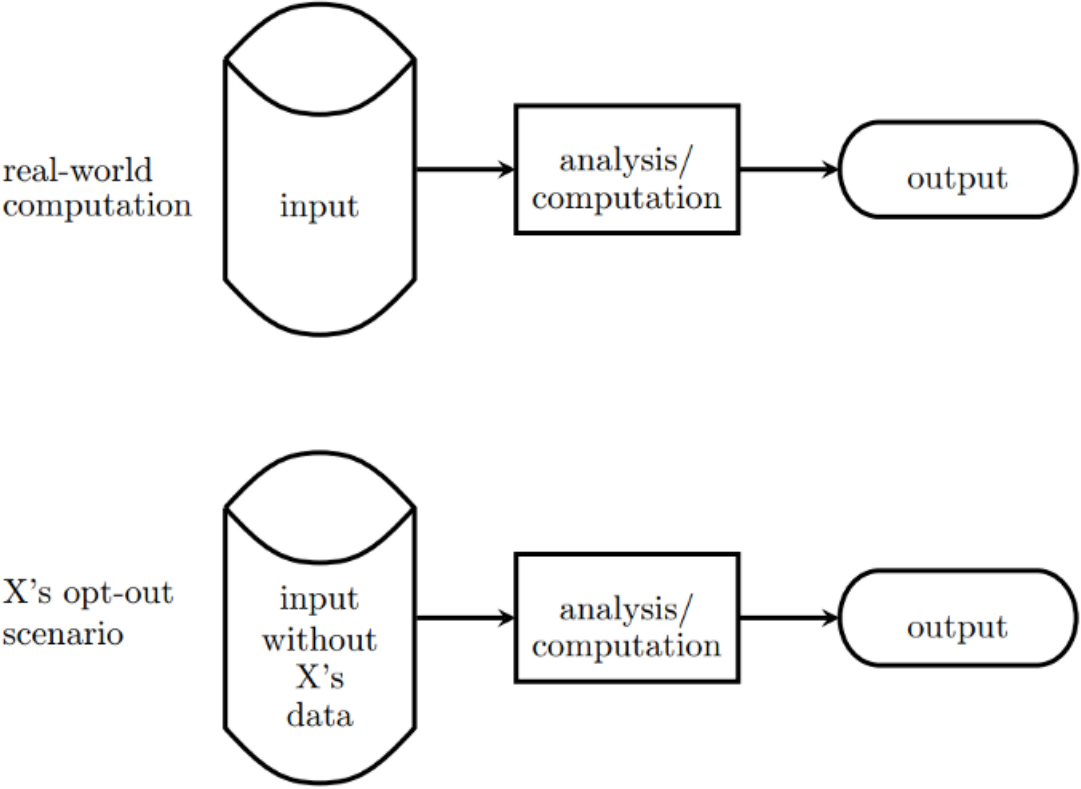
Guaranteeing people's confidentiality has become more of a challenge, but some scholars worry that the new system will impede research.



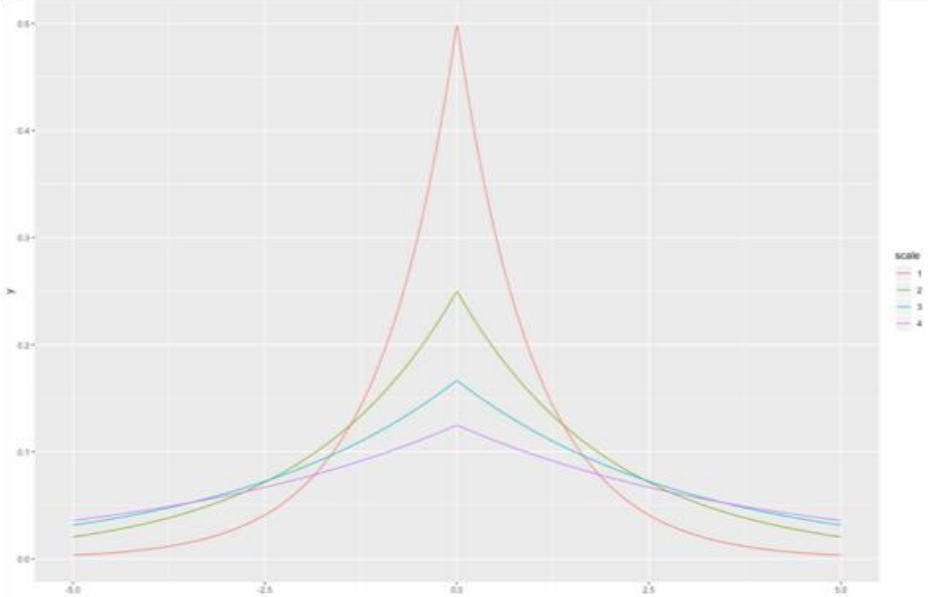
ANDY GREENBERG SECURITY 06.13.16 07:02 PM

APPLE'S 'DIFFERENTIAL PRIVACY' IS ABOUT COLLECTING YOUR DATA—BUT NOT YOUR DATA

Differential Privacy



"difference" at most ϵ



Kobbi Nissim, et al. [Differential Privacy: A Primer for a Non-technical Audience](#). February 14, 2018. <https://towardsdatascience.com/understanding-differential-privacy-85ce191e198a> An Nguyen

Randomized Response

Tail – Question A

Heads – Question B

A: Have you used marijuana in the last month?

B: Is your mother's birthday in June?

	R to all			
Yes	20%			
No	80%			
	100%			

	R to all	Est. Response born in June		
Yes	20%	4%		
No	80%	46%		
	100%	50%		

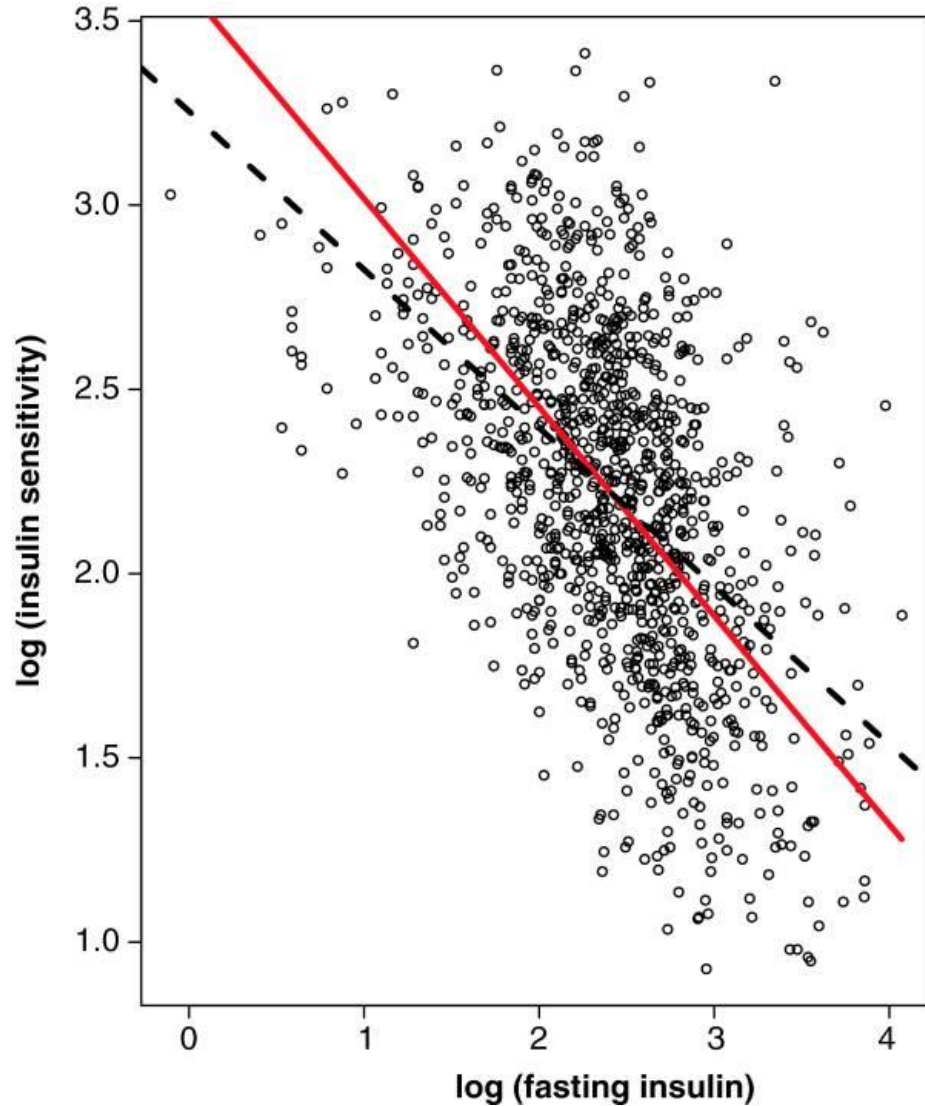
	R to all	Est. Response born in June	Inferred to target marijuana	Marijuana in pop. %
Yes	20%	4%	16%	32%
No	80%	46%	34%	68%
	100%	50%	50%	100%

In this talk

1. Four consequences for social scientists
2. Differential use of differential privacy – 3 data usage cases
3. Privacy as a social issue and the question if DP can solve that

4 Consequences of

Differential Privacy for Social Scientists

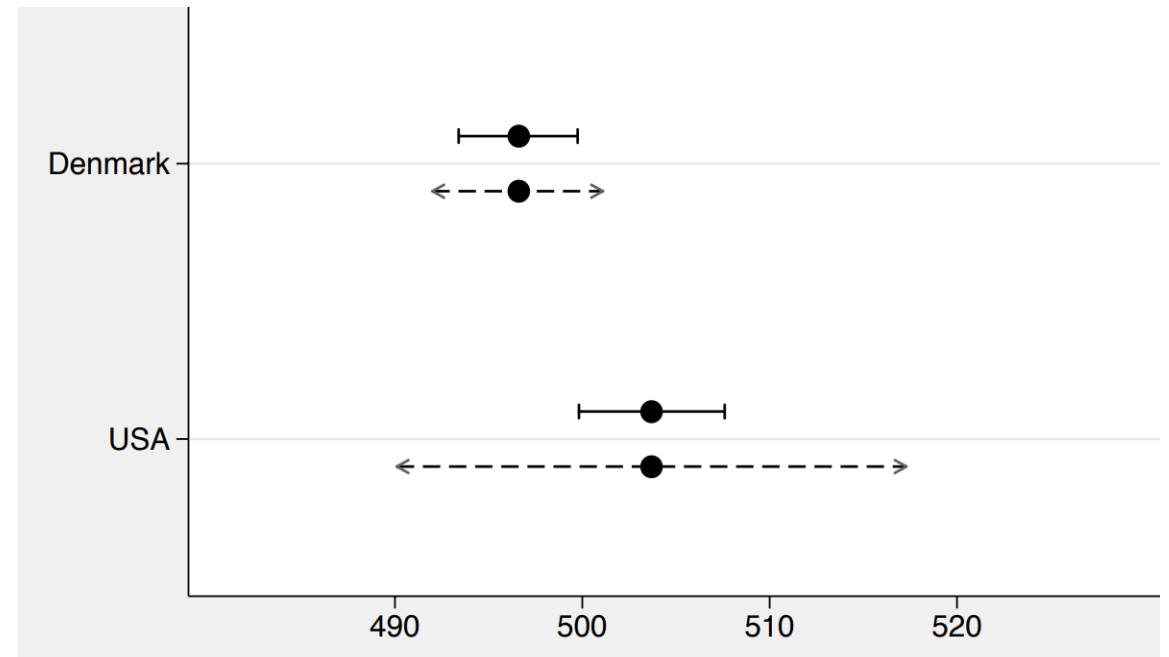


1. Differential privacy, through randomization of the data, sometimes creates bias in estimates of relationships

Social scientists will need to routinely employ measurement error corrections to obtain unbiased estimates.

2. Differential privacy adds layer of non-sampling error.

Social scientists will need to drastically increase the number of people in their samples to achieve both privacy and acceptably powerful tests of their theories.



Kreuter, Valliant (2007), PISA Test scores means and confidence intervals with and without complex sample design

3. Researchers that study small groups may find their current methods no longer sufficient.

Mail: \$50 per case

Phone: \$250 per case

Face-to-face: \$1,000 per case

Social scientists in these fields will likely require new research designs with increased costs.

4. to release data for general usage with differential privacy guarantees, the party that releases the data must weigh the privacy requirements against the foreseen usage of the data.

Social scientists will have to explicitly limit the type, scope, and/or number of questions they ask of any given dataset, ahead of time.



If implemented widely ...

- Differential Privacy will substantially transform social science.
- Researchers will need to use more complex statistical methods to account for nonsampling errors in their data;
- they will often need to drastically increase their sample sizes;
- they will have to explicitly limit the scope and complexity of their research questions to some extent; and
- they will sometimes need to target their data collection much more precisely to their questions.

Good or Bad?

Though privacy protections provide a benefit to the data subjects, they may be **detrimental to the researcher without additional funding**, since increasing sample sizes may be expensive.

They will inherently limit what can be learned from Census data, since there the sample size can't be increased.

And they might limit the types of research questions that can be answered.

However ...

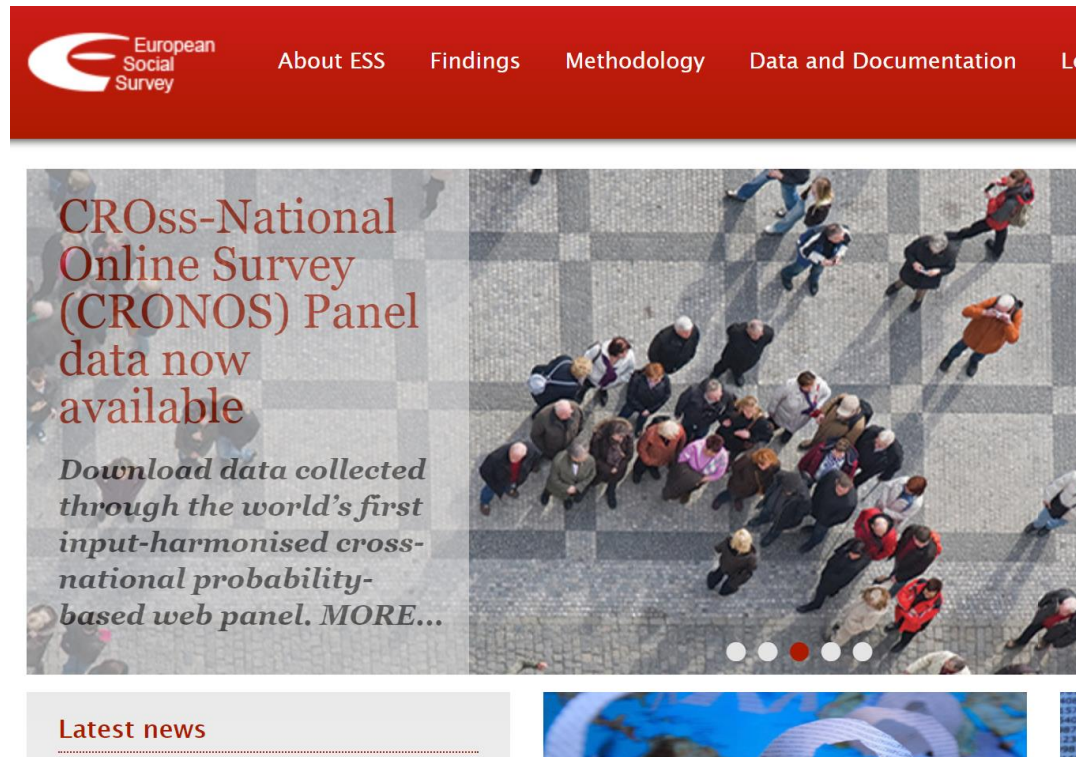
At its heart, differential privacy is about **limiting the sensitivity of one's conclusions** to the presence or absence of any one person in the analysis.

In this sense, it serves simply as a reminder of the importance of robustness (Dwork et al., 2015).

When using robust methods to analyze data, we already limit the effect an individual observation can have on the result.

3 Types of Data Use

Sample Surveys – publicly funded



The screenshot shows the top navigation bar of the European Social Survey website with the logo and menu items: About ESS, Findings, Methodology, Data and Documentation, and Le. Below the navigation bar is a large banner image of a crowd of people from an aerial perspective. On the left side of the banner, there is text: 'CROss-National Online Survey (CRONOS) Panel data now available' and 'Download data collected through the world's first input-harmonised cross-national probability-based web panel. MORE...'. Below the banner is a 'Latest news' section with a small image of a globe and a list of news items.

Usually available publicly at no cost, for example, with well-known publicly funded studies such as the European Social Survey or World Values Survey.

For this type of data, differential privacy can only hurt, since there is no evidence of privacy risks and therefore no known benefit from implementing privacy guarantees.

Confidential Microdata via Enclaves



The screenshot shows the top navigation bar of the United States Census Bureau website. It includes the logo, a search bar, and four main menu items: BROWSE BY TOPIC, EXPLORE DATA, LIBRARY, and SURVEYS/ PROGRAMS. Below the navigation bar is a breadcrumb trail: // Census.gov > About the Bureau > Federal Statistical Research Data Centers. The main heading is "Federal Statistical Research Data Centers". On the left side, there is a list of links: "About this Section", "Available Data", "Federal Partners", and "Research Data Centers". The main content area contains a paragraph describing the centers as partnerships between federal agencies and research institutions, providing secure access to restricted-use microdata. A "Read More" link is located at the bottom of the paragraph.

United States
Census
Bureau

Search

BROWSE BY TOPIC EXPLORE DATA LIBRARY SURVEYS/ PROGRAMS

// Census.gov > About the Bureau > Federal Statistical Research Data Centers

Federal Statistical Research Data Centers

About this Section Available Data Federal Partners Research Data Centers

Federal Statistical Research Data Centers are partnerships between federal statistical agencies and leading research institutions. These secure facilities providing authorized access to restricted-use microdata for statistical purposes only.

[Read More](#)

Example: Administrative Records

Trusted researchers can access the data in a secure computing environment after passing an often impressive number of hurdles, including binding agreements to honor the participants' confidentiality. Such agreements can include disclosure protections such as differential privacy guarantees.

Differential privacy can therefore be beneficial here, but only if it happens while also preserving the integrity of the data enclave approach. This approach requires additional funding.

Interactions with platforms and devices

SOCIAL SCIENCE ONE
Building Industry-Academic Partnerships

CONTACT US

HOME About Us ▾ Our Facebook Partnership ▾ People Blog FAQ ▾

Our Facebook Partnership

- Grant Process
- Research Ethics
- Data Security & Privacy
- Funders

Data Security & Privacy

The security and privacy of sensitive data is extremely important to everyone involved in Social Science One. Social scientists cannot accomplish their goals of learning about and addressing societal challenges unless they can be trusted with important information about human characteristics, behaviors, and opinions. The opportunities afforded by our partnership with companies, and the valuable information our approach makes possible, make these stakes even more substantial.

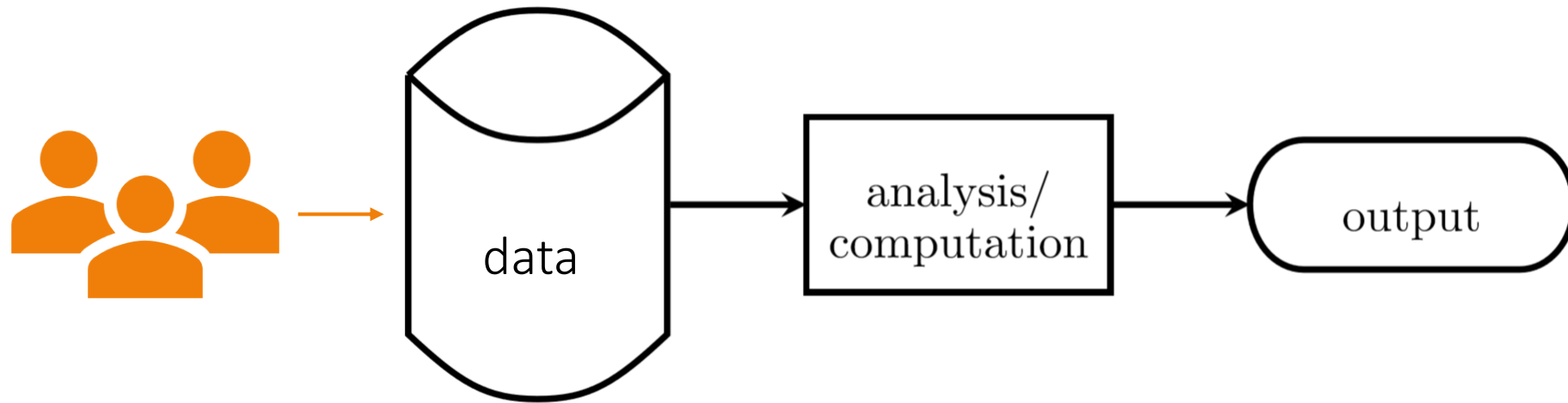


Content: geolocations, purchases, ad viewing, clicking, etc.

Raw data - out of researchers' reach. By applying the principles of differential privacy, much of the useful social-scientific information in them can be rescued.

Social Science One is exploring applying the principles of differential privacy to allow researchers access to Facebook data.

Privacy a Social Issue



Coutts & Jann [2011, SMR]

Randomized Response Techniques are problematic because of

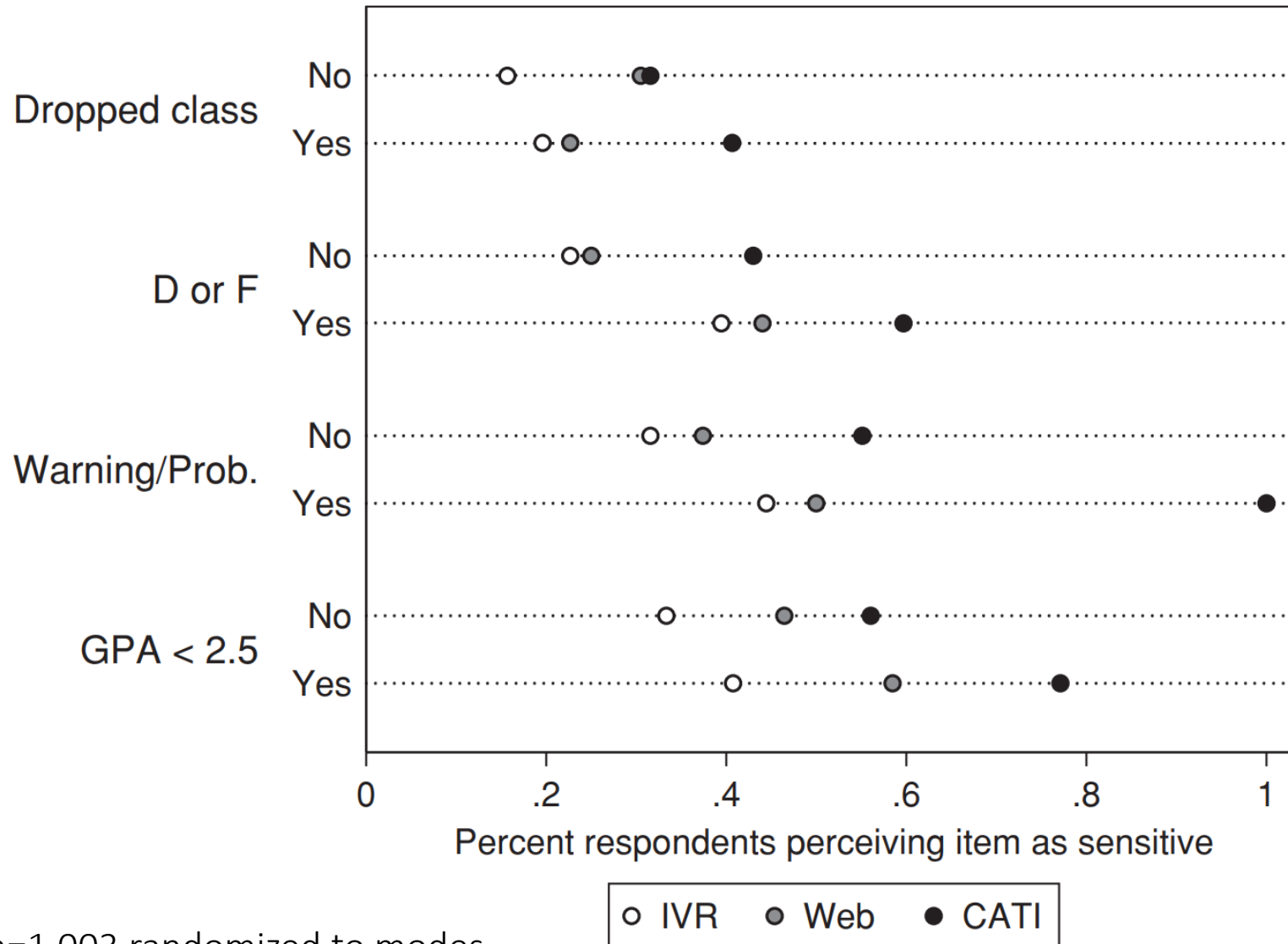
- limited trust
- high variance due to false negative tendency (especially for more sensitive questions)

Kirchner [2015]

No improvement of reporting accuracy with RRT compared to direct questioning (using administrative data for benchmark validation)

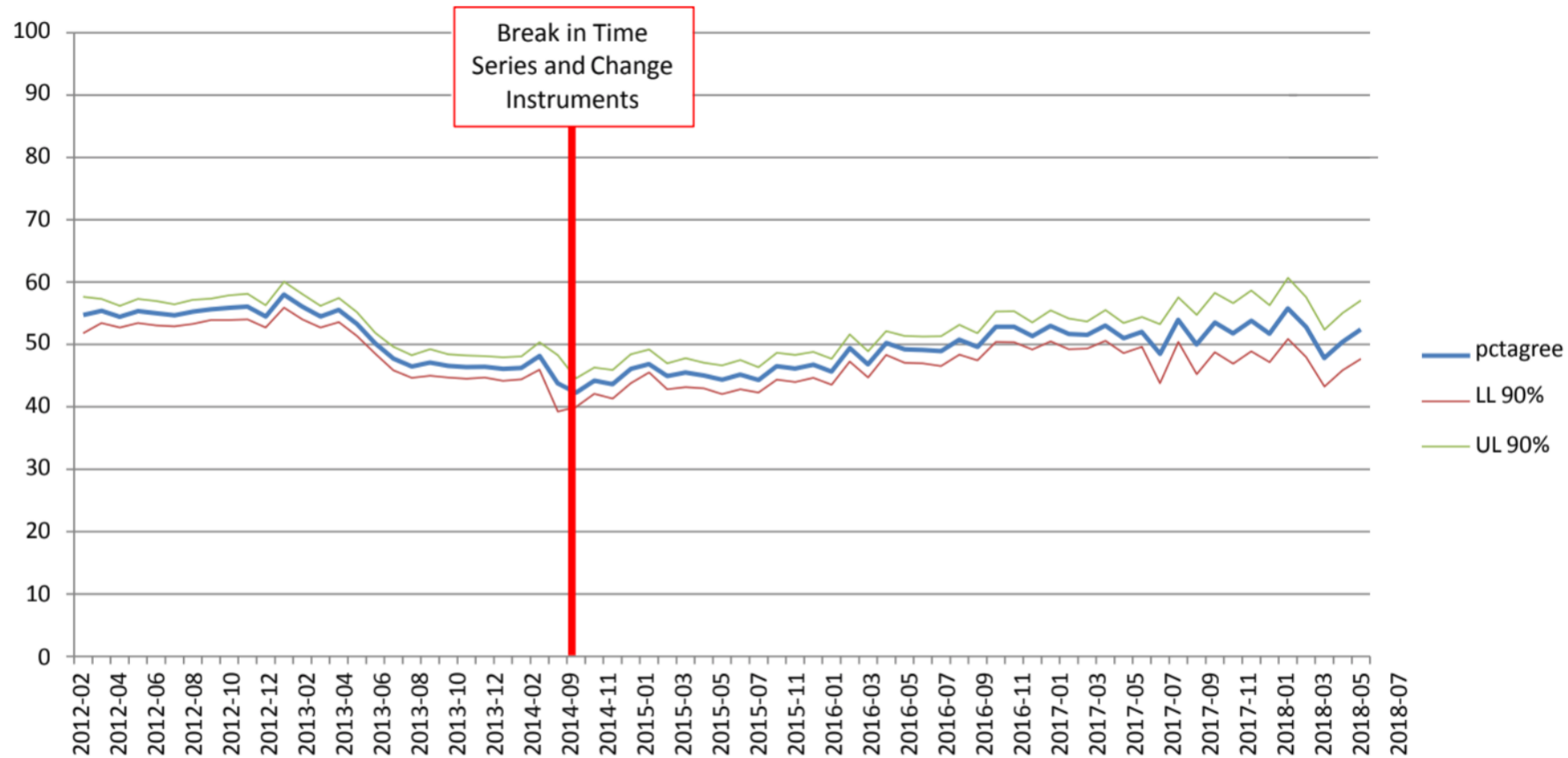
Caution shared by others [e.g. Holbrook and Krosnik 2010; Coutts et al. 2011; Wolter and Preisendoerfer 2013; Hoeglinger, Jann, and Diekmann 2014] though not all [e.g. Blair, Imai, Zhou 2015]

Reported Sensitivity of Survey Questions by True State and Mode of Data Collection [Kreuter, Presser, Tourangeau 2008, POQ]



Survey of UMD alumni n=1.003 randomized to modes

Reported Believe – Data are Kept Confidential in the Federal Statistical System [Childs, Eggeleston, Fobia 2018, BigSurv]



* Change in instruments coincided with a 4.8% decrease in reported belief.

Your participation is vital to our effort. Domestic terrorism preparedness transcends any single level of government, including the Federal government. It is a national issue that can only be effectively addressed through close cooperation at all levels—Federal, state, and local. The work of this Panel concerns nothing less than the security of our nation, the protection of our citizens' civil liberties, and the ideals of our democratic society.

Your organization has been randomly selected to represent «ORG_TYPE_TEXT» throughout the United States. The survey is being

“The estimates will be the same with or without you in the data”.

Summary

- DP implemented widely will change social science research
- DP implementation can be useful for specific data types / access
- DP implementation will like not solve the perception problem
- Many open questions
 - What if **data linkage is desired**?
 - What if **missing or misreports were intended answers**?
 - How can we get **designed unbiased estimates** for complex samples?
 - How do we deal with hierarchies (**household estimates/longitudinal data**)?