Privacy for Quantitative Social Science

Gary King

Institute for Quantitative Social Science
Harvard University

October 19, 2015

Collaboration: {Computer, Social} Science

Collaboration: {Computer, Social} Science

"Accessible and reusable data are fundamental to science in order to continuously validate and build upon previous research. Progressive expansive scientific advance rests upon access to data accompanied with sufficient information for reproducible results, a scientific ethic to maximize the utility of data to the research community, and a foundational norm that scientific communication is built on attribution."

Crosas, King, Honaker, Sweeney (2015)

Collaboration: {Computer, Social} Science

"Accessible and reusable data are fundamental to science in order to continuously validate and build upon previous research. Progressive expansive scientific advance rests upon access to data accompanied with sufficient information for reproducible results, a scientific ethic to maximize the utility of data to the research community, and a foundational norm that scientific communication is built on attribution."

Crosas, King, Honaker, Sweeney (2015)

First paper crossing both fields presented at the Society for Political Methodology meetings

D'Orazio, Honaker, King (2015)

Attacks

Computer Science has destroyed the idea of "deidentification"

- anonymization techniques for data releases are generally open to reidentification attacks (Sweeney 1997, 2000, Narayanan & Shimatikov 2008);
- aggregated statistics can not have any privacy guarantee (Dinur and Nissim 2003) - fingerprinting (Bun, Ullman, Vadhan STOC 2014), (Dwork, Smith, Steinke, Ullman, Vadhan FOCS 2015);
- even statistical estimates can leak individual information (Ullman and Steinke 2013) - time variance.

What to do about it?

What to do about it?

- Option 1: use new differential privacy methods

What to do about it?

- Option 1: use new differential privacy methods
- Consequence: SE's and CI's are way too big to make any inferences;

What to do about it?

- Option 1: use new differential privacy methods
- Consequence: SE's and CI's are way too big to make any inferences; future work is done in air gap closets

What to do about it?

- Option 1: use new differential privacy methods
- Consequence: SE's and CI's are way too big to make any inferences; future work is done in air gap closets
- Option 2: recognize that Social Science has specialized goals and novel statistical methodologies

What to do about it?

- Option 1: use new differential privacy methods
- Consequence: SE's and CI's are way too big to make any inferences; future work is done in air gap closets
- Option 2: recognize that Social Science has specialized goals and novel statistical methodologies
  - Social Science is mostly about Causal Inference

What to do about it?

- Option 1: use new differential privacy methods
- Consequence: SE's and CI's are way too big to make any inferences; future work is done in air gap closets
- Option 2: recognize that Social Science has specialized goals and novel statistical methodologies
  - Social Science is mostly about Causal Inference
  - Most of the rest is description & prediction, using various forms of regression

What to do about it?

- Option 1: use new differential privacy methods
- Consequence: SE's and CI's are way too big to make any inferences; future work is done in air gap closets
- Option 2: recognize that Social Science has specialized goals and novel statistical methodologies
  - Social Science is mostly about Causal Inference
  - Most of the rest is description & prediction, using various forms of regression
- Consequences:

What to do about it?

- Option 1: use new differential privacy methods
- Consequence: SE's and CI's are way too big to make any inferences; future work is done in air gap closets
- Option 2: recognize that Social Science has specialized goals and novel statistical methodologies
  - Social Science is mostly about Causal Inference
  - Most of the rest is description & prediction, using various forms of regression
- Consequences:
  - SE's are reasonably sized, inference can proceed

What to do about it?

- Option 1: use new differential privacy methods
- Consequence: SE's and CI's are way too big to make any inferences; future work is done in air gap closets
- Option 2: recognize that Social Science has specialized goals and novel statistical methodologies
  - Social Science is mostly about Causal Inference
  - Most of the rest is description & prediction, using various forms of regression
- Consequences:
  - SE's are reasonably sized, inference can proceed
  - But where's the infrastructure to stand between data and users?

A repository for sharing, citing, analyzing, and preserving research data.

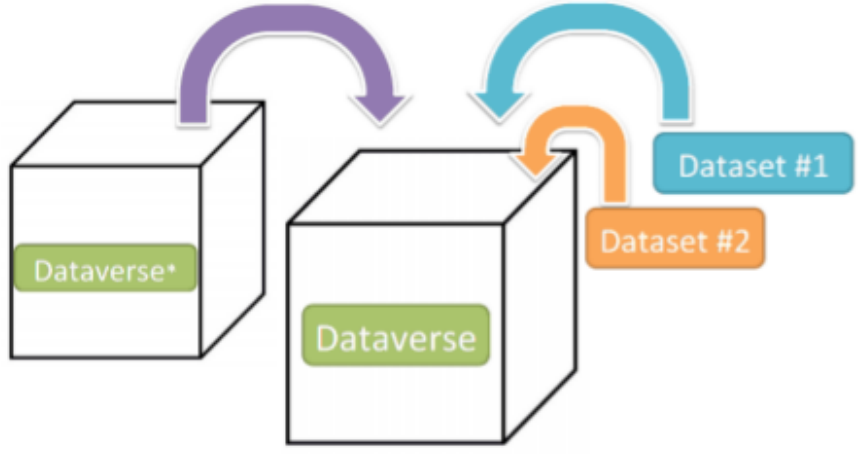
dataverse.org

A repository for sharing, citing, analyzing, and preserving research data.

dataverse.org

The largest collection of social science research data in the world

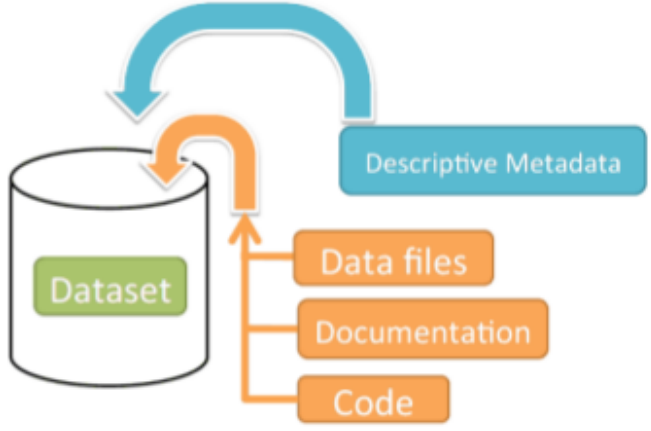
Schematic Diagram of a **Dataverse** in Dataverse 4.0



Container for your **Datasets** and/or **Dataverses***

* Dataverses can now contain other Dataverses (this replaces Collections & Subnetworks)

Schematic Diagram of a **Dataset** in Dataverse 4.0

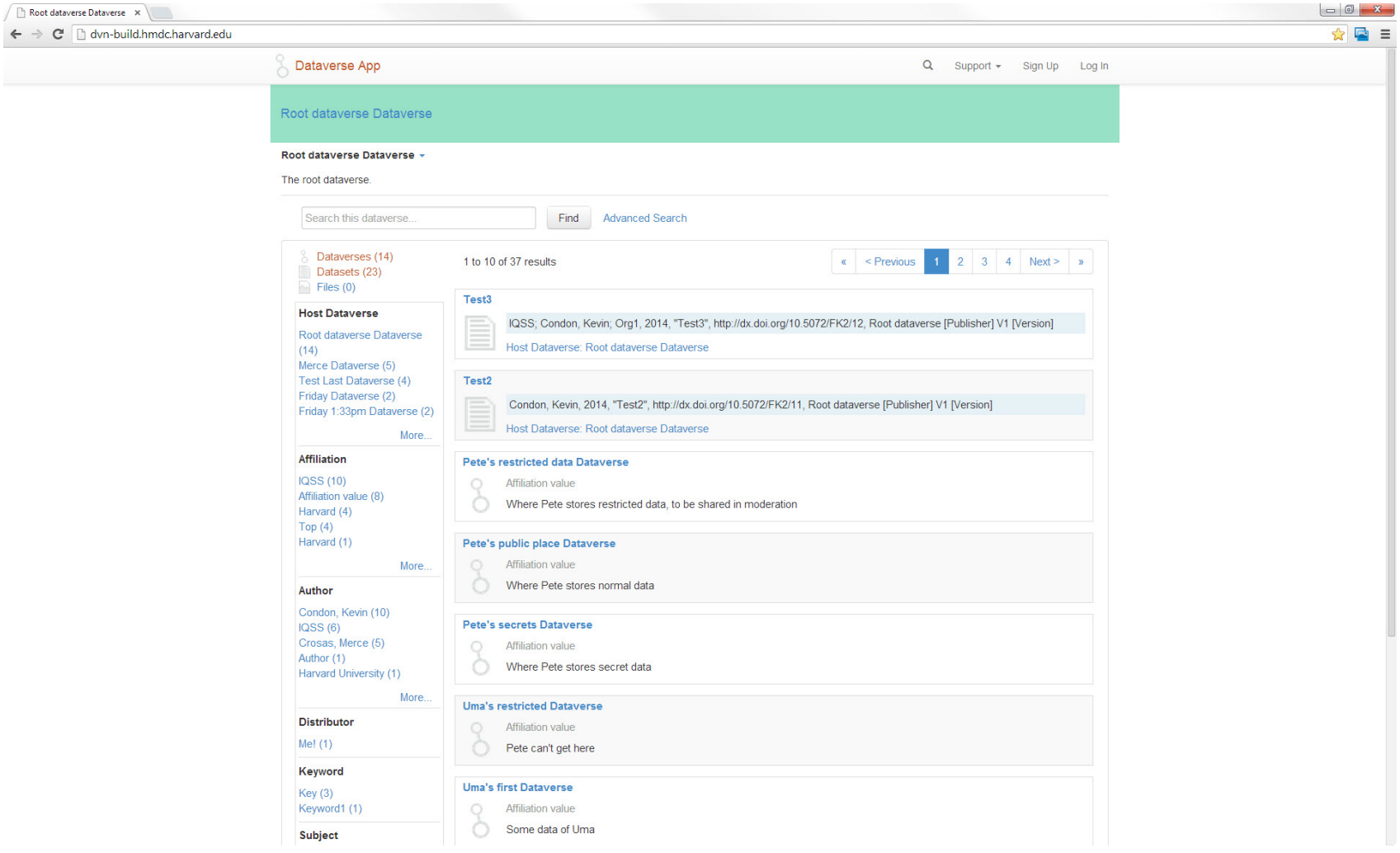


Container for your data, documentation, and code.

dataverse.org

Incentives Aligned for Scholars

Incentives Aligned for Scholars



Adjusting the Tools of Social Science

Adjusting the Tools of Social Science

- test experimental treatments (difference of means)

Adjusting the Tools of Social Science

- test experimental treatments (difference of means) — *privately*

Adjusting the Tools of Social Science

- test experimental treatments (difference of means) — *privately*
- match data to adjust for confounders

Adjusting the Tools of Social Science

- test experimental treatments (difference of means) — *privately*
- match data to adjust for confounders — *privately*

Adjusting the Tools of Social Science

- test experimental treatments (difference of means) — *privately*
- match data to adjust for confounders — *privately*
- compute summary statistics

Adjusting the Tools of Social Science

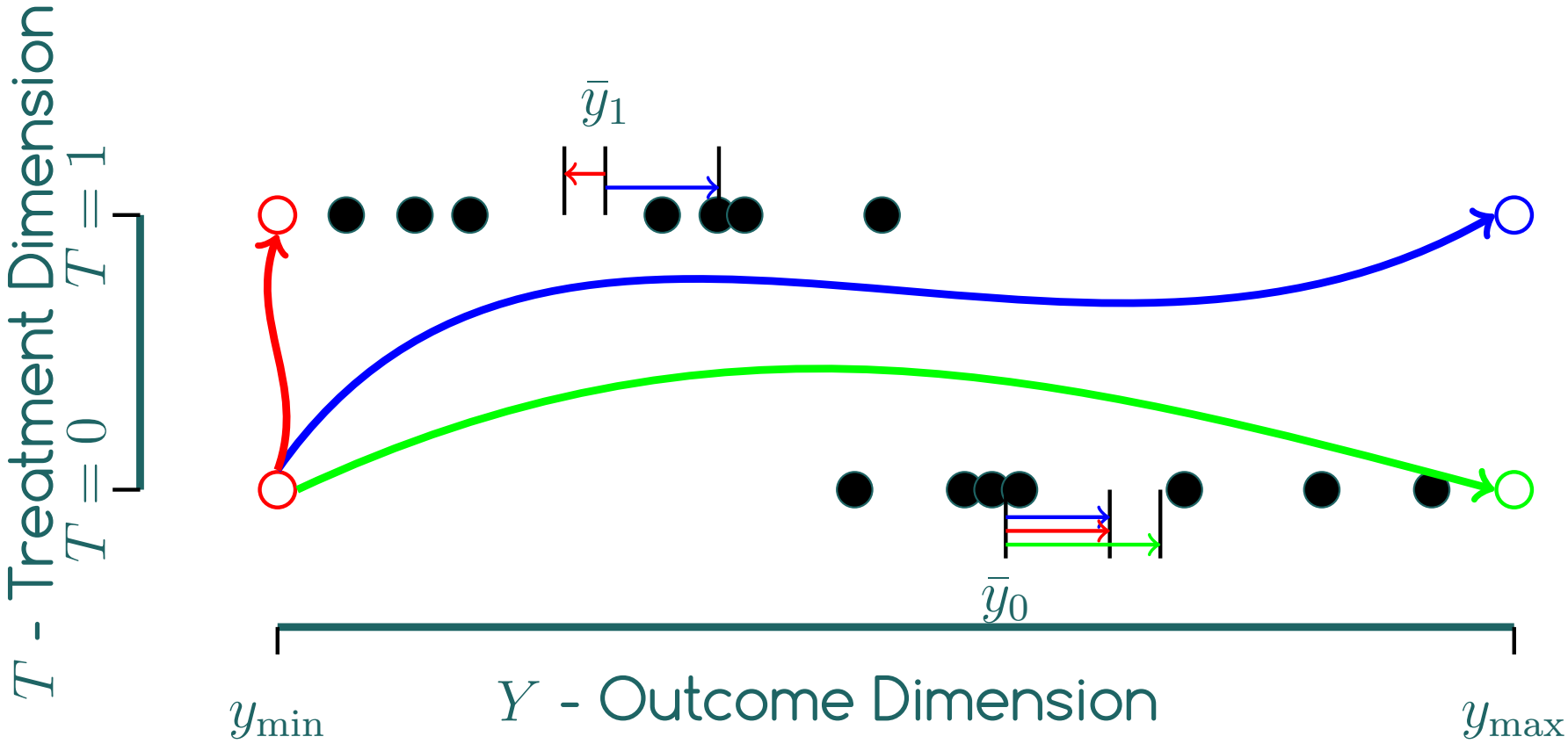
- test experimental treatments (difference of means) — *privately*
- match data to adjust for confounders — *privately*
- compute summary statistics — *privately*

Adjusting the Tools of Social Science

- test experimental treatments (difference of means) — *privately*
- match data to adjust for confounders — *privately*
- compute summary statistics — *privately*
- run regressions

Adjusting the Tools of Social Science

- test experimental treatments (difference of means) — *privately*
- match data to adjust for confounders — *privately*
- compute summary statistics — *privately*
- run regressions — *privately*


T - Treatment Dimension
$T = 1$
$T = 0$
$\bar{y}_1$
$\bar{y}_0$
$y_{\min}$
Y - Outcome Dimension
$y_{\max}$

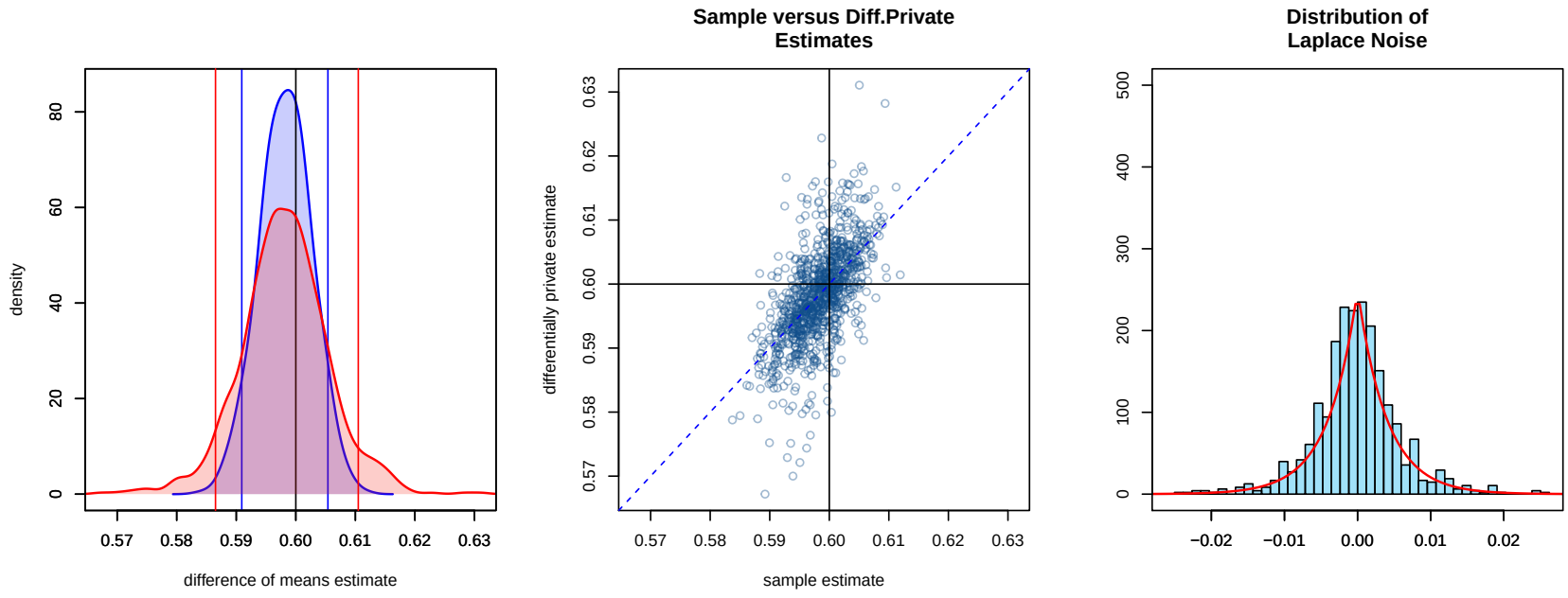


Figure: *Distributions of differentially private statistics of the difference of means estimate.*

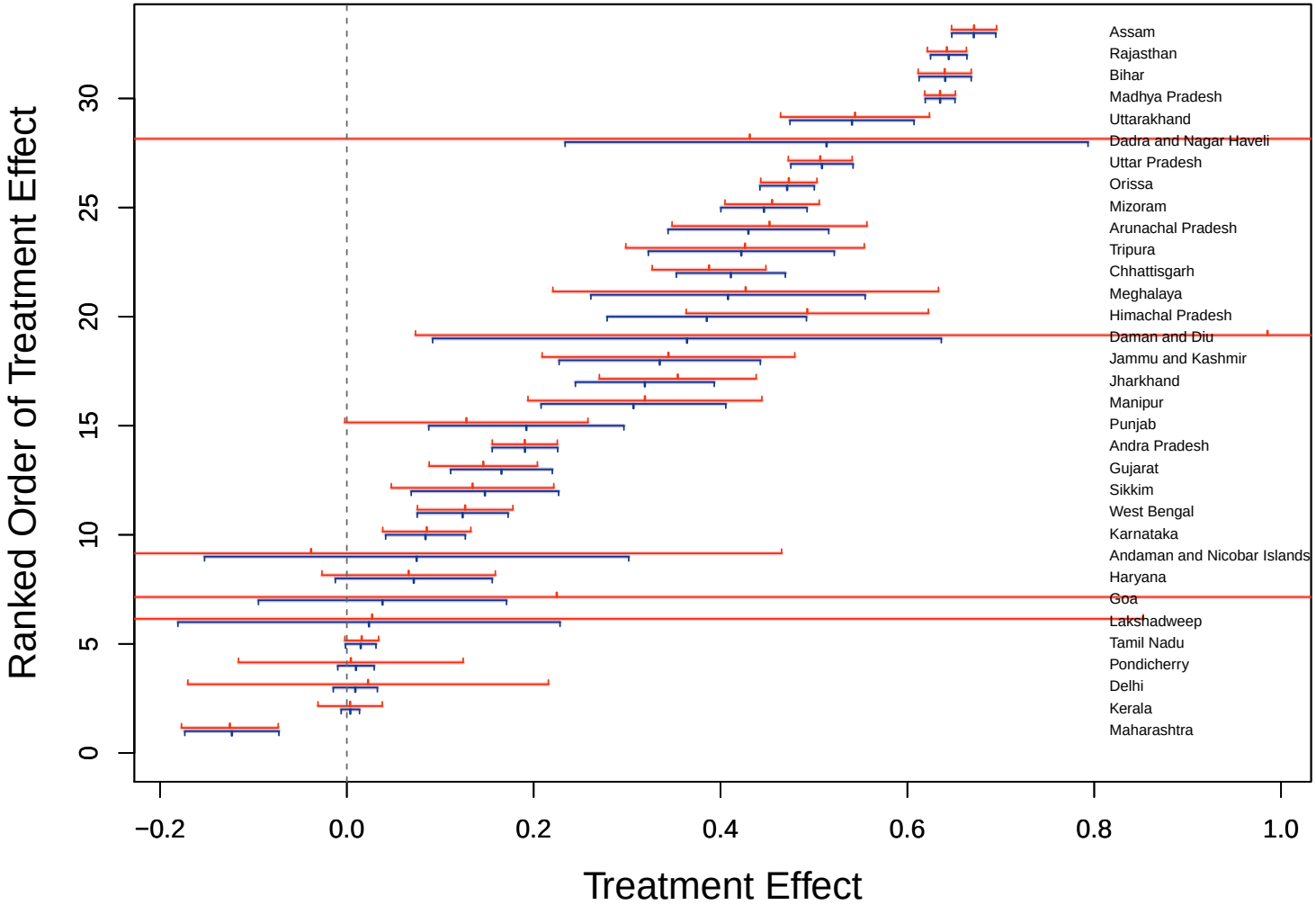


Figure: *Diff. of means estimates across 33 Indian states for the treatment effect of JSY cash transfers to women on the probability of delivering at a birthing center.*

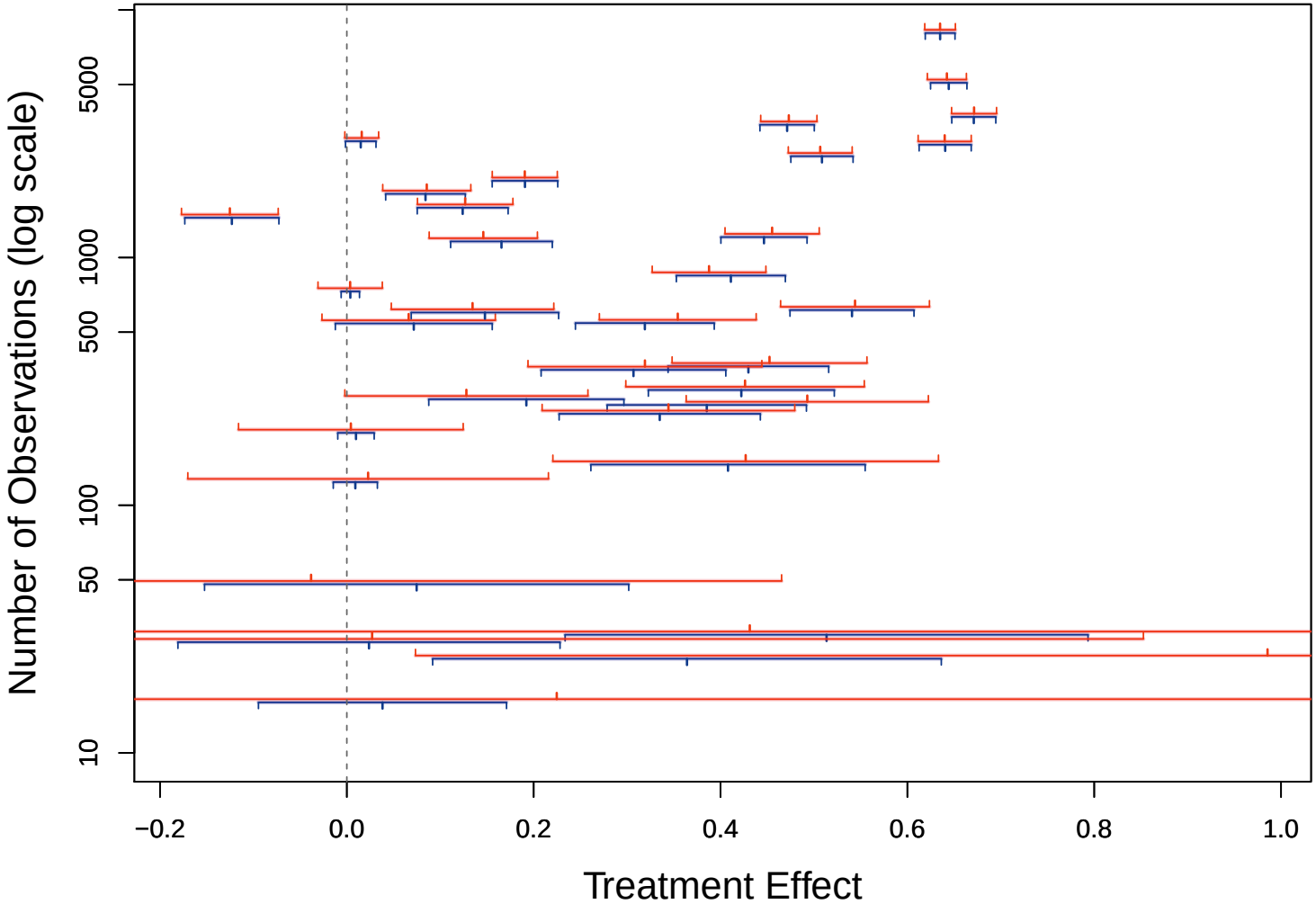


Figure: *Diff. of means estimates across 33 Indian states for the treatment effect of JSY cash transfers to women on the probability of delivering at a birthing center.*

Conclusions

- The threat of reidentification is endemic in social science research

Conclusions

- The threat of reidentification is endemic in social science research
- Access to data is central to open science and progressive reuse

Conclusions

- The threat of reidentification is endemic in social science research
- Access to data is central to open science and progressive reuse
- Social science exploration revolves around causal inference, summary statistics, and regression – all of which we've made strides in

Conclusions

- The threat of reidentification is endemic in social science research
- Access to data is central to open science and progressive reuse
- Social science exploration revolves around causal inference, summary statistics, and regression – all of which we've made strides in
- Going forward:

Conclusions

- The threat of reidentification is endemic in social science research
- Access to data is central to open science and progressive reuse
- Social science exploration revolves around causal inference, summary statistics, and regression – all of which we've made strides in
- Going forward:
  - sensitive data can be shared through the same repositories now used, such as Dataverse

Conclusions

- The threat of reidentification is endemic in social science research
- Access to data is central to open science and progressive reuse
- Social science exploration revolves around causal inference, summary statistics, and regression – all of which we've made strides in
- Going forward:
  - sensitive data can be shared through the same repositories now used, such as Dataverse
  - There's an urgent need to expand the realm of social science-specific statistical methods that be made (a) differentially private but (b) with reasonably sized uncertainty estimates