Privacy for Quantitative Social Science

Gary King
Albert J. Weatherhead III University Professor at Harvard University

Director of the Institute for Quantitative Social Science

October 19, 2015

Openness and 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)

Attacks

Computer Science has destroyed the idea of "deidentification" but it is still the normal practice in social science data:

- 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.

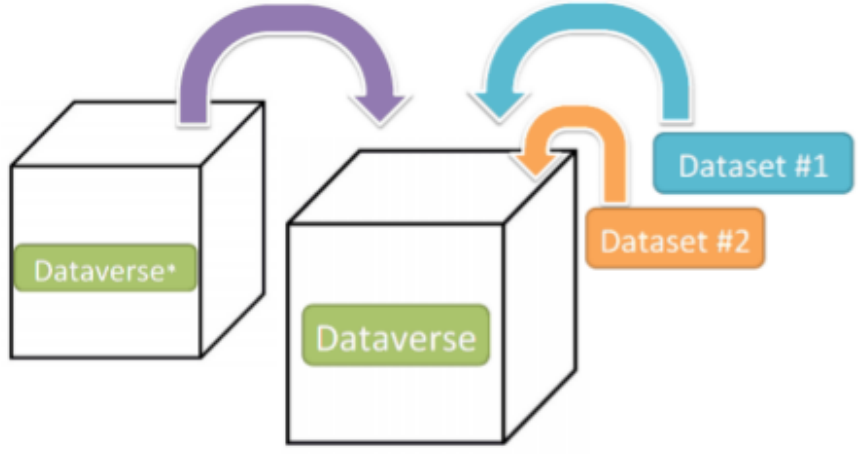
All of Social Science is Causal Inference.
King and Powell (2008); King Keohane and Verba (1994)

Regression is 90% of the rest.

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

http://dataverse.org

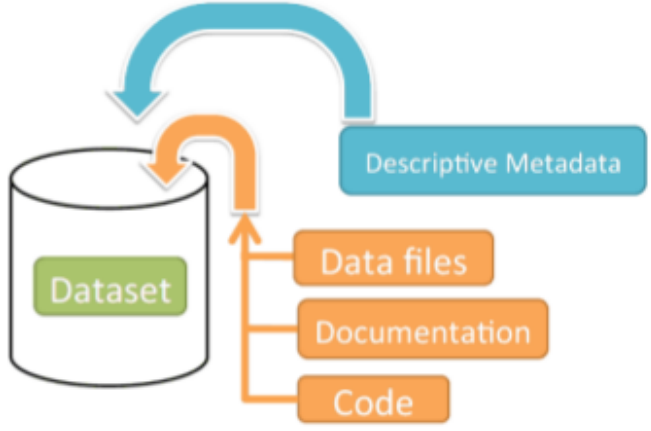
Schematic Diagram of a **Dataverse** in Dataverse 4.0



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

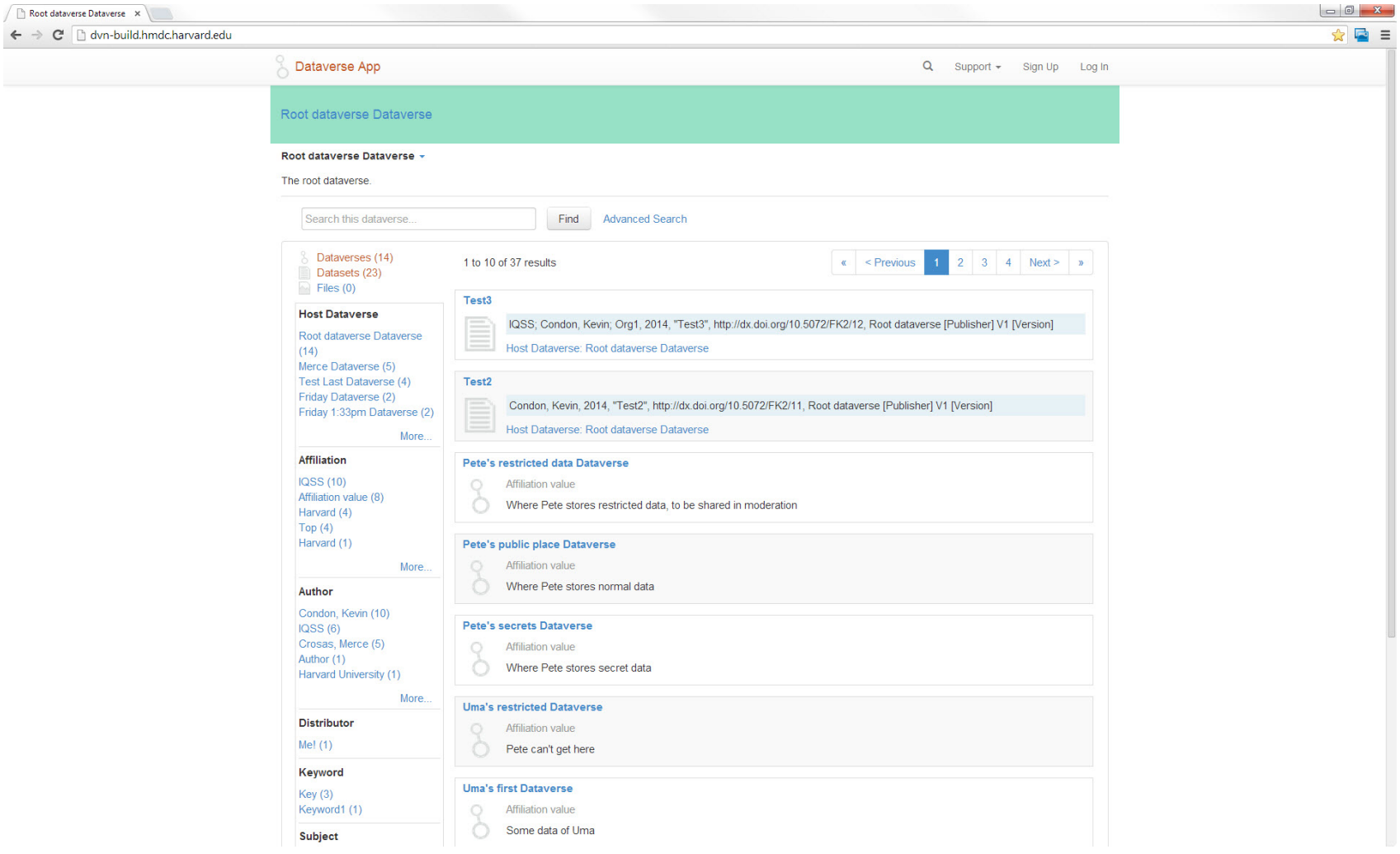
* Dataverses can now contain other Dataverses (this replaces Collections & Subnetworks)

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



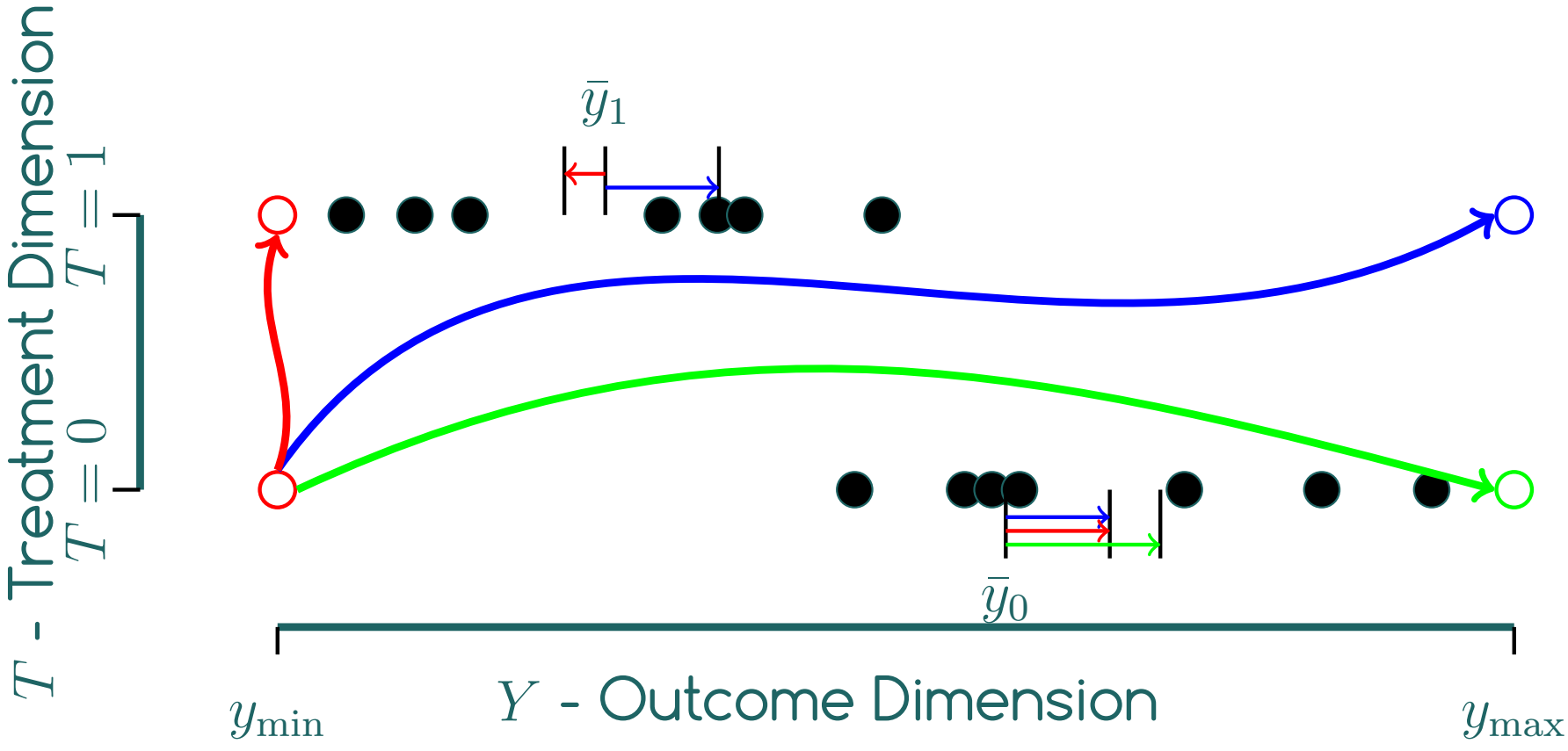
Container for your data, documentation, and code.

http://dataverse-demo.hmdc.harvard.edu


Root dataverse Dataverse
dvn-build.hmdc.harvard.edu
Dataverse App
Support
Sign Up
Log In
Root dataverse Dataverse
Root dataverse Dataverse
The root dataverse.
Search this dataverse...
Find
Advanced Search
Dataverses (14)
Datasets (23)
Files (0)
1 to 10 of 37 results
« < Previous 1 2 3 4 Next > »
Host Dataverse
Root dataverse Dataverse (14)
Merce Dataverse (5)
Test Last Dataverse (4)
Friday Dataverse (2)
Friday 1:33pm Dataverse (2)
More...
Affiliation
IQSS (10)
Affiliation value (8)
Harvard (4)
Top (4)
Harvard (1)
More...
Author
Condon, Kevin (10)
IQSS (6)
Crosas, Merce (5)
Author (1)
Harvard University (1)
More...
Distributor
Me! (1)
Keyword
Key (3)
Keyword1 (1)
Subject
Test3
IQSS; Condon, Kevin; Org1, 2014, "Test3", http://dx.doi.org/10.5072/FK2/12, Root dataverse [Publisher] V1 [Version]
Host Dataverse: Root dataverse Dataverse
Test2
Condon, Kevin, 2014, "Test2", http://dx.doi.org/10.5072/FK2/11, Root dataverse [Publisher] V1 [Version]
Host Dataverse: Root dataverse Dataverse
Pete's restricted data Dataverse
Affiliation value
Where Pete stores restricted data, to be shared in moderation
Pete's public place Dataverse
Affiliation value
Where Pete stores normal data
Pete's secrets Dataverse
Affiliation value
Where Pete stores secret data
Uma's restricted Dataverse
Affiliation value
Pete can't get here
Uma's first Dataverse
Affiliation value
Some data of Uma

The Fundamental Tools of Social Science

- How to test experimental treatments (difference of means) privately King and Powell (2008)
- How to match data privately
  Iacus, King, Porro (2011); Ho, Imai, King, Stuart (2007)
- How to compute summary statistics privately
  Imai King Lau (2008)
- How to run regressions privately


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

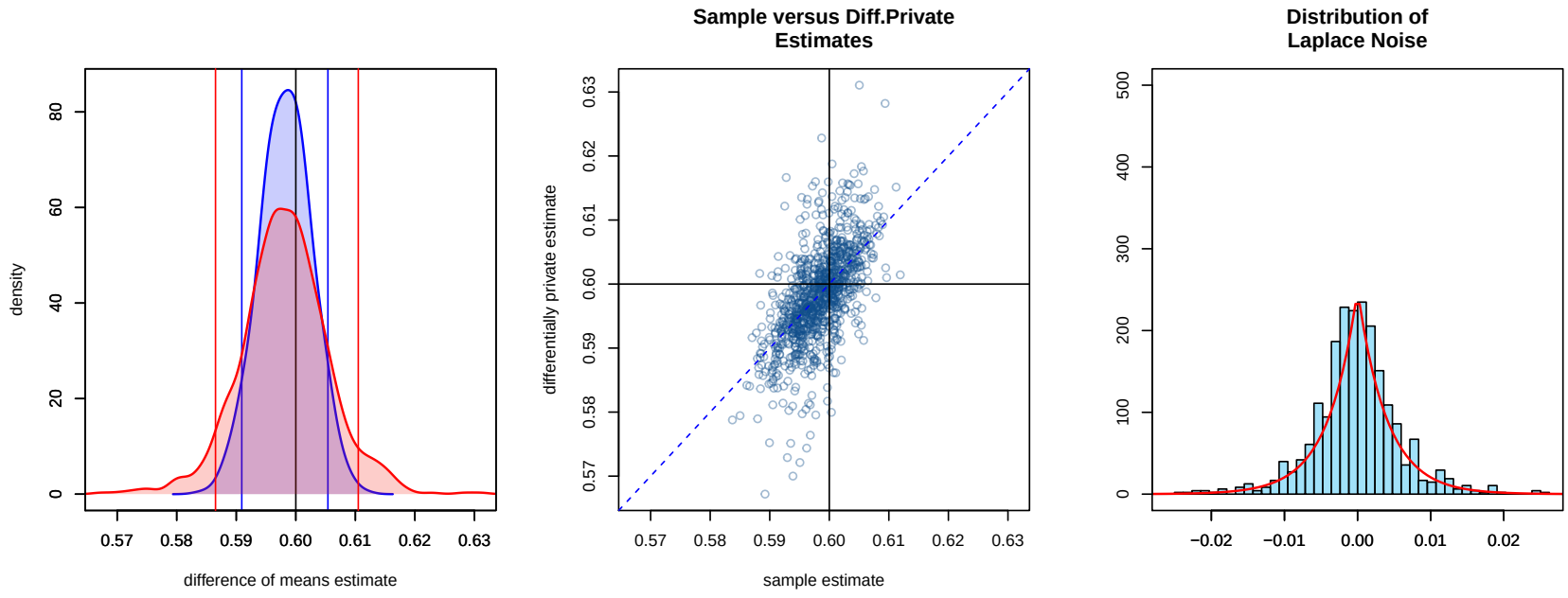


Figure : *Distributions of differentially private statistics of the difference of means estimate.*

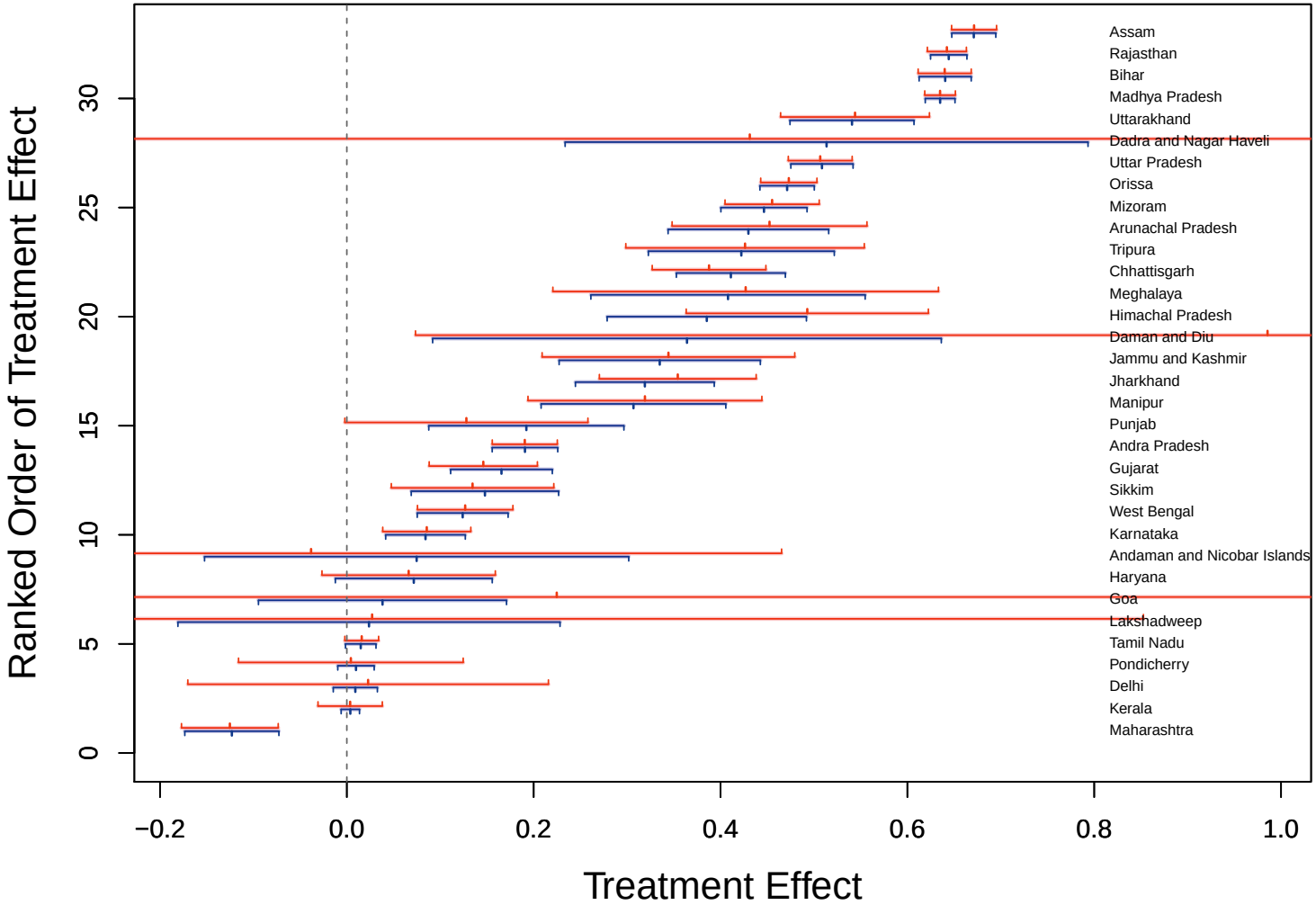


Figure : *Diff. of means estimates across 33 Indian states for the treatment effect of JSY cash transfers to women on the probability of delivering at a birthing center.*

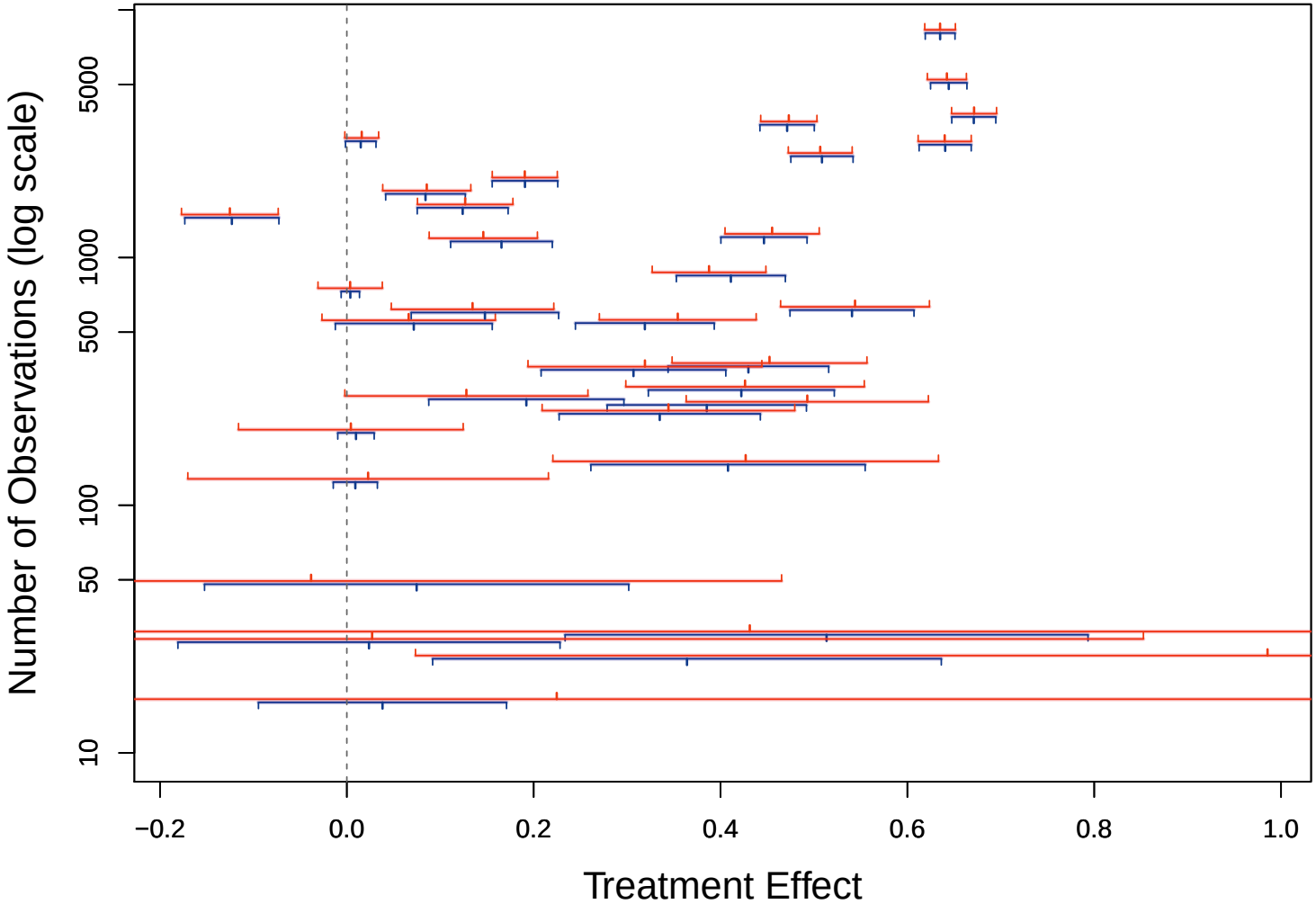


Figure : *Diff. of means estimates across 33 Indian states for the treatment effect of JSY cash transfers to women on the probability of delivering at a birthing center.*

Conclusion

- 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 repositories such as Dataverse with these tools