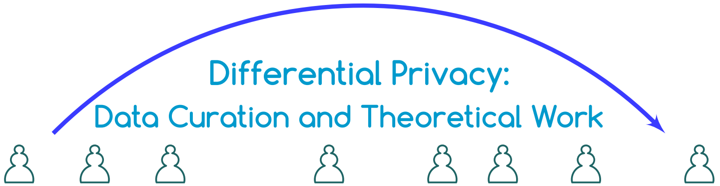
Differential Privacy:
Data Curation and Theoretical Work



James Honaker[1], Kobbi Nissim[2]

[1] Institute for Quantitative Social Science, Harvard University

[2] Center for Research on Computation and Society, Harvard University

October 19, 2015

Differential Privacy [DMNS 2006]

- A strong mathematical definition of individual privacy.

Differential Privacy [DMNS 2006]

- A strong mathematical definition of individual privacy.
- Controls the excess risk to an individual from participating in an analysis.

Differential Privacy [DMNS 2006]

- A strong mathematical definition of individual privacy.
- Controls the excess risk to an individual from participating in an analysis.
  - How: Hides the effect of every individual on the analysis outcome by injection of carefully designed random noise.

Differential Privacy [DMNS 2006]

- A strong mathematical definition of individual privacy.
- Controls the excess risk to an individual from participating in an analysis.
  - How: Hides the effect of every individual on the analysis outcome by injection of carefully designed random noise.
- Rich theoretical foundation; in prime time for testing and application.

Differential Privacy [DMNS 2006]

- A strong mathematical definition of individual privacy.
- Controls the excess risk to an individual from participating in an analysis.
  - How: Hides the effect of every individual on the analysis outcome by injection of carefully designed random noise.
- Rich theoretical foundation; in prime time for testing and application.
- Receives interest from many communities.

Differential Privacy [DMNS 2006]

Formally,

A randomized mechanism $M : X^n \to T$ is (pure) ϵ-differentially private if for all neighboring datasets $x, x' \in X^n$ and subset S of the outcome set T,

$$\Pr[M(X) \in S] \leq e^{\epsilon} \cdot \Pr[M(X') \in S].$$

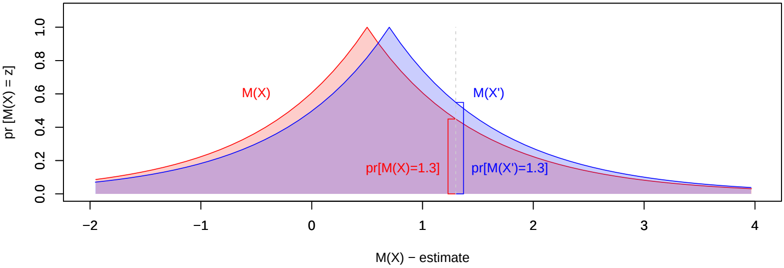
Differential Privacy [DMNS 2006]

Formally,

A randomized mechanism $M : X^n \to T$ is (pure) ϵ-differentially private if for all neighboring datasets $x, x' \in X^n$ and subset S of the outcome set T,

$$\Pr[M(X) \in S] \leq e^{\epsilon} \cdot \Pr[M(X') \in S].+\delta$$

Relaxation: *approximate* differential privacy also allows a (negligible) additive difference, δ.


M(X)
M(X')
pr[M(X)=1.3]
pr[M(X')=1.3]
pr [M(X) = z]
0.0
0.2
0.4
0.6
0.8
1.0
−2
−1
0
1
2
3
4
M(X) − estimate

What can be computed with DP?

A huge variety of computational tasks:

- Basic statistics.
  - Histograms, contingency tables, CDFs, ...
- Inferential statistics.
  - Regression, ...
- Machine learning.
  - Classification, clustering, SVD, convex optimization, ...
- Graph/social network analysis.
- Streaming algorithms.

What can be computed with DP?

A huge variety of computational tasks:

- Basic statistics.
  - Histograms, contingency tables, CDFs, ...
- Inferential statistics.
  - Regression, ...
- Machine learning.
  - Classification, clustering, SVD, convex optimization, ...
- Graph/social network analysis.
- Streaming algorithms.

Broader applications: Where privacy is not necessarily the goal

- Mechanism design, games.
- Preventing false detection.

Differential Privacy in Dataverse

- Our goal: Facilitate sharing of privacy sensitive data.

Differential Privacy in Dataverse

- Our goal: Facilitate sharing of privacy sensitive data.
- How: Differential privacy *at least* as a tool for deciding on applying for access.

Differential Privacy in Dataverse

- **Our goal:** Facilitate sharing of privacy sensitive data.
- **How:** Differential privacy *at least* as a tool for deciding on applying for access.
  - Depositors choose basic stats that best represent their data to be computed with DP.

Differential Privacy in Dataverse

- Our goal: Facilitate sharing of privacy sensitive data.
- How: Differential privacy *at least* as a tool for deciding on applying for access.
  - Depositors choose basic stats that best represent their data to be computed with DP.
  - DP stats integrate with TwoRavens, a data exploration GUI.

Differential Privacy in Dataverse

- Our goal: Facilitate sharing of privacy sensitive data.
- How: Differential privacy *at least* as a tool for deciding on applying for access.
  - Depositors choose basic stats that best represent their data to be computed with DP.
  - DP stats integrate with TwoRavens, a data exploration GUI.
  - Data users explore basic stats in TwoRavens and make further queries to determine interest in dataset.

Differential Privacy in Dataverse

- Our goal: Facilitate sharing of privacy sensitive data.
- How: Differential privacy *at least* as a tool for deciding on applying for access.
    - Depositors choose basic stats that best represent their data to be computed with DP.
    - DP stats integrate with TwoRavens, a data exploration GUI.
    - Data users explore basic stats in TwoRavens and make further queries to determine interest in dataset.
    - Try to support analyses most useful in social science (causal inference and regression)

Prototype Tool for Differentially Private Data Exploration

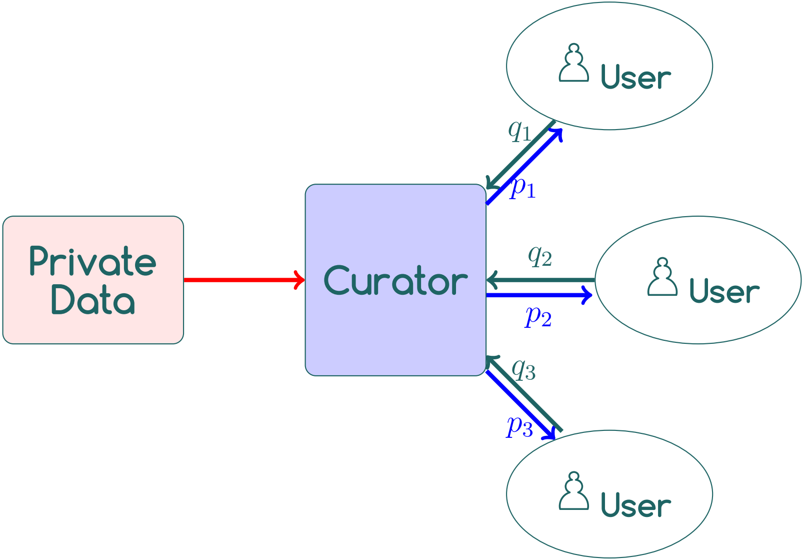


Figure: *The curator architecture for data privacy.*

workflow for private data

Upload Budget Release Explore Query

https://beta.dataverse.org/custom/DifferentialPrivacyPrototype/

Summer 2015 Efforts

	Social Science		Computer Science		Statistics
Problem	**Mentor**	**Intern**	**Intern**	**Mentor**	**Mentor**
Regression	Honaker	Antuca	Wang	Sheffet	Karwa
Utility	Honaker	Gooden		Sheffet	Karwa
Two-way Tables		Lim	Rogers	Gaboardi	Karwa
Visualiztn of Uncert.	Honaker		Bu		
Density and Trees		Muise		Bun & Nissim	
Security Architecture	(Durand)		Merrill	Chong	
Interactive Queries			Kaminsky	Vadhan & Murtagh	
Datalog Logic Engine			Bembenek	Chong	
Attacks on Agg. Data			Jiang	Steinke & Ullman	
Verification of DP			Farina	Gaboardi & Chong	

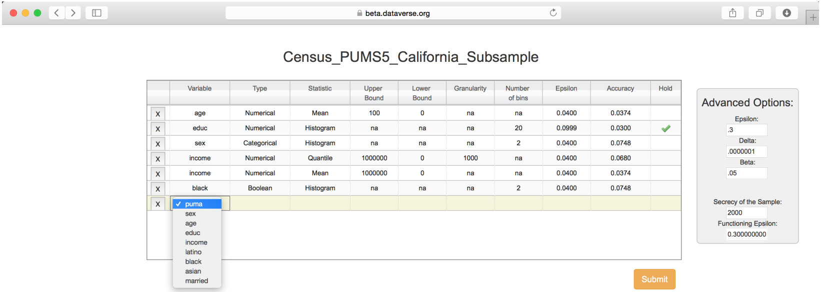
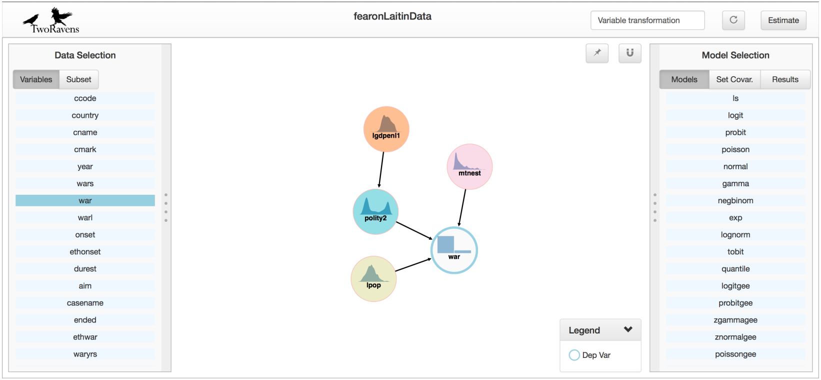


Figure: *Example screen from the interactive privacy budget allocation tool for data depositors.*

The TwoRavens Interface



TwoRavens

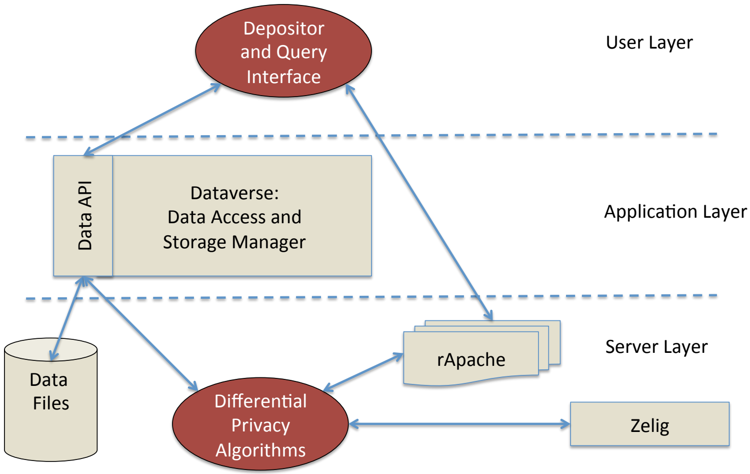
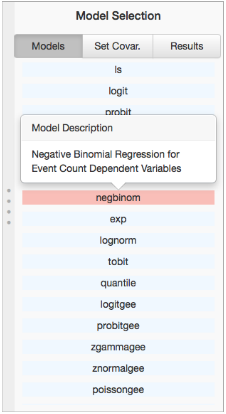
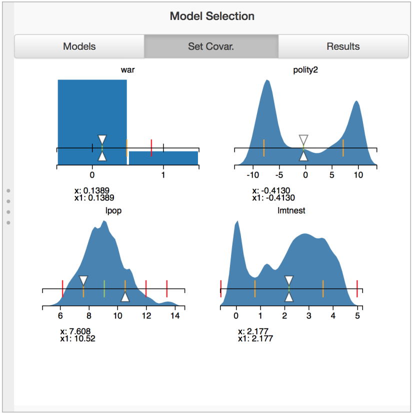


Figure: *Privacy architecture for secure curator interfaces.*

Integration with Zelig





Privacy Tools

Differentially Private Statistical Exploration

A National Science Foundation Secure and Trustworthy Cyberspace Project

This prototype system will allow researchers with sensitive datasets to make differentially private statistics about their data available through data repositories using the Dataverse platform.

Our prototype system will allow researchers to: [1] upload private data to a secured Dataverse archive, [2] decide what statistics they would like to release about that data, and [3] release privacy preserving versions of those statistics to the repository, [4] that can be explored through a curator interface without releasing the raw data, including [5] interactive queries.

This system was created by the Privacy Tools for Sharing Research Data project. Differential privacy is a mathematical framework for enabling statistical analysis of sensitive datasets while ensuring that individual-level information cannot be leaked. The project website contains resources for learning more about differential privacy.

Budget Tool

The first part of this system is a tool that helps both data depositors and data analysts distribute a global privacy budget across many statistics. Users select which statistics they would like to calculate and are given estimates of how accurately each statistic can be computed. They can also redistribute their privacy budget according to which statistics they think are most valuable in their dataset. This work has motivated new theoretical results from our group that maximize the utility achievable when using differential privacy to share many statistics about a research dataset.

Explore Budget Tool Demo

Curator Interface

When the data depositor has distributed their privacy budget, the second portion of our tool system draws differentially private versions of those statistical summaries selected by the data depositor from a library of differentially private routines (which we created in the R statistical language, and also make available for use by the R community) and stores them in metadata associated with that file on Dataverse. Future researchers who wish to explore restricted social science data can then access these privacy-preserving summary statistics either from the metadata, or through the TwoRavens graphical data exploration tool built for Dataverse, which we have adapted for differentially private statistics.

Explore TwoRavens Interface Demo

Interactive Queries

Our system will allow some of the privacy budget to be reserved for future data analysts to choose their own differentially private statistics to calculate (selected from the library of differentially private algorithms provided by the system). Differential privacy will ensure that even if these queries are chosen adversarially, individual-level information will not be leaked. This currently works through a command-line interactive system, and we are developing a future user interface.

Watch Video Tour of Interactive Queries Demo

This system was created by the Privacy Tools for Sharing Research Data project. Here are ways you can follow or contribute to this project.

https://beta.dataverse.org/custom/DifferentialPrivacyPrototype/

Differential Privacy Theoretical Research

Main Focus Areas

- DP and statistics [D'Orazio, Gaboardi, Honaker, Karwa, King, Lim, Rogers, Sheffet, Vadhan, Zheng].
- Private machine learning [Bun, Nissim].
- Bounds on DP [Bun, Nissim, Vadhan].
- DP and false discovery [Nissim, Smith, Steinke, Ullman].
- Programming languages techniques for DP [Gaboardi].
- Composition of DP mechanisms [Murtagh, Vadhan].
- A new real-life application [Kantarcioglu, Sweeney].
- Estimating privacy risk [Dwork, Jiang, Smith, Steinke, Ullman, Vadhan].
- DP as an equilibrium of economic games [Chen, Nissim, Sheffet, Vadhan].

DP and Statistics

DP and Statistics

- Linear Regression & Casual Inference:
  - [Sheffet]: New DP algorithms for 2^{nd} moment matrix of a dataset and least-squares regression for statistical inference.
  - [D'Orazio, Honaker, King] and [Karwa, Vadhan]: Work in progress.

DP and Statistics

- Linear Regression & Casual Inference:
  - [Sheffet]: New DP algorithms for 2^{nd} moment matrix of a dataset and least-squares regression for statistical inference.
  - [D'Orazio, Honaker, King] and [Karwa, Vadhan]: Work in progress.
- [Gaboardi, Lim, Rogers]: New results on goodness-of-fit testing and independence testing with DP.
  - In particular, how to calculate "significance level" of χ^2 test, taking into account noise added for DP.

DP and Statistics

- Linear Regression & Casual Inference:
  - [Sheffet]: New DP algorithms for 2^{nd} moment matrix of a dataset and least-squares regression for statistical inference.
  - [D'Orazio, Honaker, King] and [Karwa, Vadhan]: Work in progress.
- [Gaboardi, Lim, Rogers]: New results on goodness-of-fit testing and independence testing with DP.
  - In particular, how to calculate "significance level" of χ^2 test, taking into account noise added for DP.
- [Vadhan, Zheng]: Traditional synthetic data generation methods achieve differential privacy in many cases.
  - Zheng's thesis won the Hoopes Prize for outstanding undergraduate work.

* Karwa is a post-doc researcher with Airoldi & Vadhan. Airoldi & Nissim were readers on Zheng's Thesis.

New Real-Life Applications of DP

- In first site visit Sweeney introduced a re-identification of bicycle routs in Hubway contest.

New Real-Life Applications of DP

- In first site visit Sweeney introduced a re-identification of bicycle routs in Hubway contest.
- [Kantarcioglu, Sweeney]: Compared DP techniques to create DP synthetic data set.
  - showed that DP would have sufficed for most entries in Hubway data contest with $\epsilon = 0.9$.

New Real-Life Applications of DP

- In first site visit Sweeney introduced a re-identification of bicycle routs in Hubway contest.
- [Kantarcioglu, Sweeney]: Compared DP techniques to create DP synthetic data set.
  - showed that DP would have sufficed for most entries in Hubway data contest with $\epsilon = 0.9$.
- Over the next year, they will test the actual utility of these methods actual against contest entries.
  - Goal: provide software to contest organizers worldwide to use.

DP and False Discovery

- A new surprising line of research shows that DP can be used as a tool for preventing false discovery with adaptively used data.

DP and False Discovery

- A new surprising line of research shows that DP can be used as a tool for preventing false discovery with adaptively used data.
- [Steinke, Ullman COLT15]: extend lowerbound techniques for DP to give tight bounds on computational hardness of preventing false discovery.

DP and False Discovery

- A new surprising line of research shows that DP can be used as a tool for preventing false discovery with adaptively used data.
- [Steinke, Ullman COLT15]: extend lowerbound techniques for DP to give tight bounds on computational hardness of preventing false discovery.
- [Nissim, Smith, Steinke, Ullman+]: Give improved upperbounds for false discovery and a tight characterization of the generalization of DP.

Bounds on DP

- Part of our long-term research on understanding what can be computed with DP, and with what costs.
- [Bun, Nissim, Vadhan+ FOCS15]: Estimating basic statistics such as quantiles, learning distributions wrt Komogorov distance on domain D.
  - Requires between $\log^* |D|$ and $2^{\log^* |D|}$ samples.
  - Impossible when information is taken from a continuous domain.

Private Machine Learning

- Part of our long-term research of possibility and limitations of DP machine learning.
- [Nissim+ SODA15]: Semi-supervised learning (where some examples are unlabeled) for mitigating the higher sample complexity of DP learning.
  - Number of labeled examples matches non-private learning.
- [Bun, Nissim+]: Upper- and lower-bounds on the cost of simultaneously learning k concepts.

PL techniques for DP

- Part of our long-term research into PL tools for ensuring differential privacy.
- [Gaboardi+ POPL15]: Semi-automated techniques for verifying a program is DP.
  - based on a type system able to express properties of two runs of a program
- [Gaboardi+ SNAPL15]: Formal program logic techniques for reasoning about randomized algs.
  - Useful, in particular, for expressing and verifying accuracy properties of DP mechanisms.

Composition of DP mechanisms

- Part of our long-term goal of understanding and using composition of DP mechanisms.
  - Composition is one of the properties making DP programmable.
- [Murtagh, Vadhan] Optimal composition theorems for DP.
  - Hardness of exactly computing the optimal composition.
  - Poly-time approximation of optimal composition.

Estimating Privacy Risk

- [Dwork, Smith, Steinke, Ullman, Vadhan FOCS15]: New attacks on releases of aggregate stats.
  - Require less auxiliary information than previous similar attacks.
  - Use very simple stats (column sums).
  - Robust to choice of perturbation technique.
- [Jiang, Steinke] perform experimental evaluation of the attacks.

DP as an equilibrium of economic games

- Does DP appear naturally in games?
- [Chen, Sheffet, Vadhan WINE14]: Analyzed a simple game-theoretic model where an agent balances benefits and risks of revealing sensitive information.
- Research in this vein continues (+Nissim).

DP - New papers (since Jan 2015)

- Mark Bun, Jonathan Ullman, Salil Vadhan. *Fingerprinting Codes and the Price of Approximate Differential Privacy.* SICOMP.
- Amos Beimel, Kobbi Nissim, Uri Stemmer. *Learning Privately with Labeled and Unlabeled Examples.* SODA.
- Mark Bun, Kobbi Nissim, Uri Stemmer, Salil Vadhan. *Differentially Private Release and Learning of Threshold Functions.* FOCS.
- Or Sheffet. *Private Approximations of the 2nd-Moment Matrix Using Existing Techniques in Linear Regression.*
- Or Sheffet. *Differentially Private Least Squares: Estimation, Confidence and Rejecting the Null Hypothesis.*
- Jack Murtagh, Salil Vadhan. *The Complexity of Computing the Optimal Composition of Differential Privacy.*
- Mark Bun, Mark Zhandry. *Order-Revealing Encryption and the Hardness of Private Learning.*
- Thomas Steinke, Jonathan Ullman. *Interactive Fingerprinting Codes and the Hardness of Preventing False Discovery.* COLT.
- Thomas Steinke, Jonathan Ullman. *Between Pure and Approximate Differential Privacy.*
- Raef Bassily, Adam Smith, Thomas Steinke, Jonathan Ullman. *More General Queries and Less Generalization Error in Adaptive Data Analysis.*
- Kobbi Nissim, Uri Stemmer. *On the Generalization Properties of Differential Privacy.*
- Kobbi Nissim, Uri Stemmer, Salil Vadhan. *Locating a Small Cluster Privately.*
- Mark Bun, Kobbi Nissim, Uri Stemmer. *Simultaneous Private Learning of Multiple Concepts.*
- Kobbi Nissim, David Xiao. *Mechanism Design and Differential Privacy.* Encyclopedia of Algorithms.
- Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, Salil Vadhan. *Robust Traceability from Trace Amounts.* FOCS.
- Xianrui Meng, Seny Kamara, Kobbi Nissim, George Kollios. *GRECS: Graph Encryption for Approximate Shortest Distance Queries.* ACM CCS.

DP - Selected Presentations

- Privacy Tools participates in TPDP Workshop as program chair, committee, and invited speaker.
- Kobbi Nissim: *Privacy: How theory can influence reality*. UCSD Distinguished Lecturer Series.
- Salil Vadhan: *Keynote talk*. Big Data/Social Informatics 2015.
- Vito D'Orazio, James Honaker, Garry King: *Differentially private methods*. Annual Meetings of the American Political Science Association *and* Meetings of the Midwest Political Science Association.
- Thomas Steinke: *Differential Privacy*. China Theory Week.
- Kobbi Nissim: *The Theory of Bringing Privacy into Practice*. Caltech.
- Kobbi Nissim: *private learning*. Charles River Crypto Day
- Mark Bun: *Differentially Private Release and Learning of Threshold Functions*. FOCS.
- Thomas Steinke: *Robust Traceability from Trace Amounts*. FOCS.
- Vito D'Orazio, James Honaker, Gary King and co-PI King presented on differentially private methods at the Meetings of the Midwest Political Science Association.

Some future research directions

Some future research directions

- Linear regression and causal inference.

Some future research directions

- Linear regression and causal inference.
- Improvements on S&A – a basic DP construction technique.

Some future research directions

- Linear regression and causal inference.
- Improvements on S&A – a basic DP construction technique.
- Open problems w.r.t. private learning in approx. DP: Characterization of sample complexity, improper learning, …

Some future research directions

- Linear regression and causal inference.
- Improvements on S&A – a basic DP construction technique.
- Open problems w.r.t. private learning in approx. DP: Characterization of sample complexity, improper learning, ...
- Adaptive choice of parameters.

Some future research directions

- Linear regression and causal inference.
- Improvements on S&A – a basic DP construction technique.
- Open problems w.r.t. private learning in approx. DP: Characterization of sample complexity, improper learning, ...
- Adaptive choice of parameters.
- Data-based choice of DP mechanism.

Some future research directions

- Linear regression and causal inference.
- Improvements on S&A – a basic DP construction technique.
- Open problems w.r.t. private learning in approx. DP: Characterization of sample complexity, improper learning, ...
- Adaptive choice of parameters.
- Data-based choice of DP mechanism.
- Controlling false discovery in adaptive data analysis.

Some future research directions

- Linear regression and causal inference.
- Improvements on S&A – a basic DP construction technique.
- Open problems w.r.t. private learning in approx. DP: Characterization of sample complexity, improper learning, . . .
- Adaptive choice of parameters.
- Data-based choice of DP mechanism.
- Controlling false discovery in adaptive data analysis.
- DP where privacy is not the goal.

Conclusion

- We advocate differential privacy as part of our approach to privacy.

Conclusion

- We advocate differential privacy as part of our approach to privacy.
- We are building an open library of differentially private tools in R, the most commonly used language in applied quantitative analysis.

Conclusion

- We advocate differential privacy as part of our approach to privacy.
- We are building an open library of differentially private tools in R, the most commonly used language in applied quantitative analysis.
- We have designed and implemented an architecture for social science researchers to use our tools.

Conclusion

- We advocate differential privacy as part of our approach to privacy.
- We are building an open library of differentially private tools in R, the most commonly used language in applied quantitative analysis.
- We have designed and implemented an architecture for social science researchers to use our tools.
- The use of differentially private tools requires new ways of thinking about our statistical estimators.

Conclusion

- We advocate differential privacy as part of our approach to privacy.
- We are building an open library of differentially private tools in R, the most commonly used language in applied quantitative analysis.
- We have designed and implemented an architecture for social science researchers to use our tools.
- The use of differentially private tools requires new ways of thinking about our statistical estimators.
- Our theoretical work has helped establish and advance the rich theory that makes differential privacy a strong privacy concept.