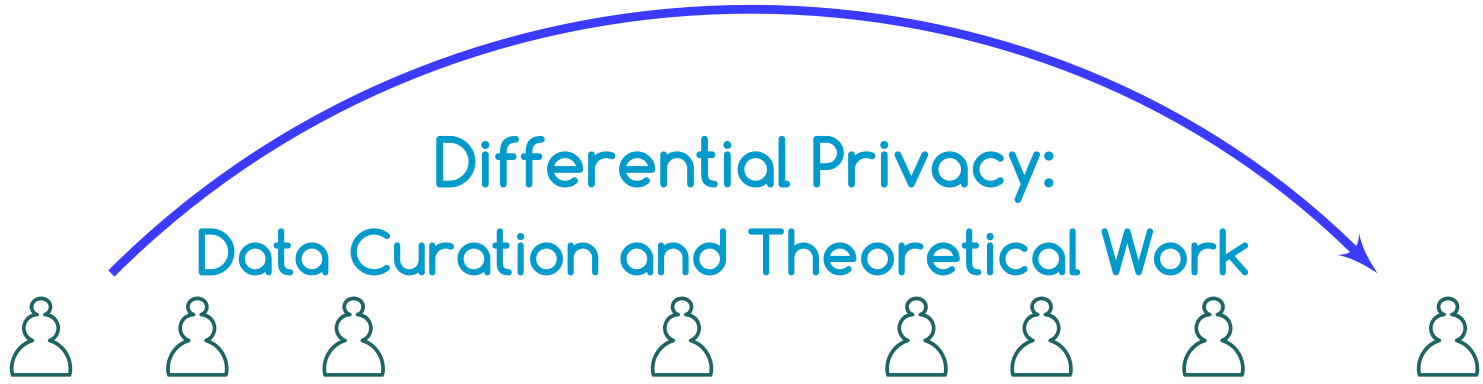


Differential Privacy:
Data Curation and Theoretical Work

James Honaker[1], Kobbi Nissim[2]

[1] Institute for Quantitative Social Science, Harvard University

[2] Center for Resrearch 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,

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

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

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:

- 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 as a proxy for deciding on applying for access.

Differential Privacy in Dataverse

- Our goal: Facilitate sharing of privacy sensitive data.
- How: Differential privacy as a proxy 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 as a proxy 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 as a proxy 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.

Prototype Tool for Differentially Private Data Exploration

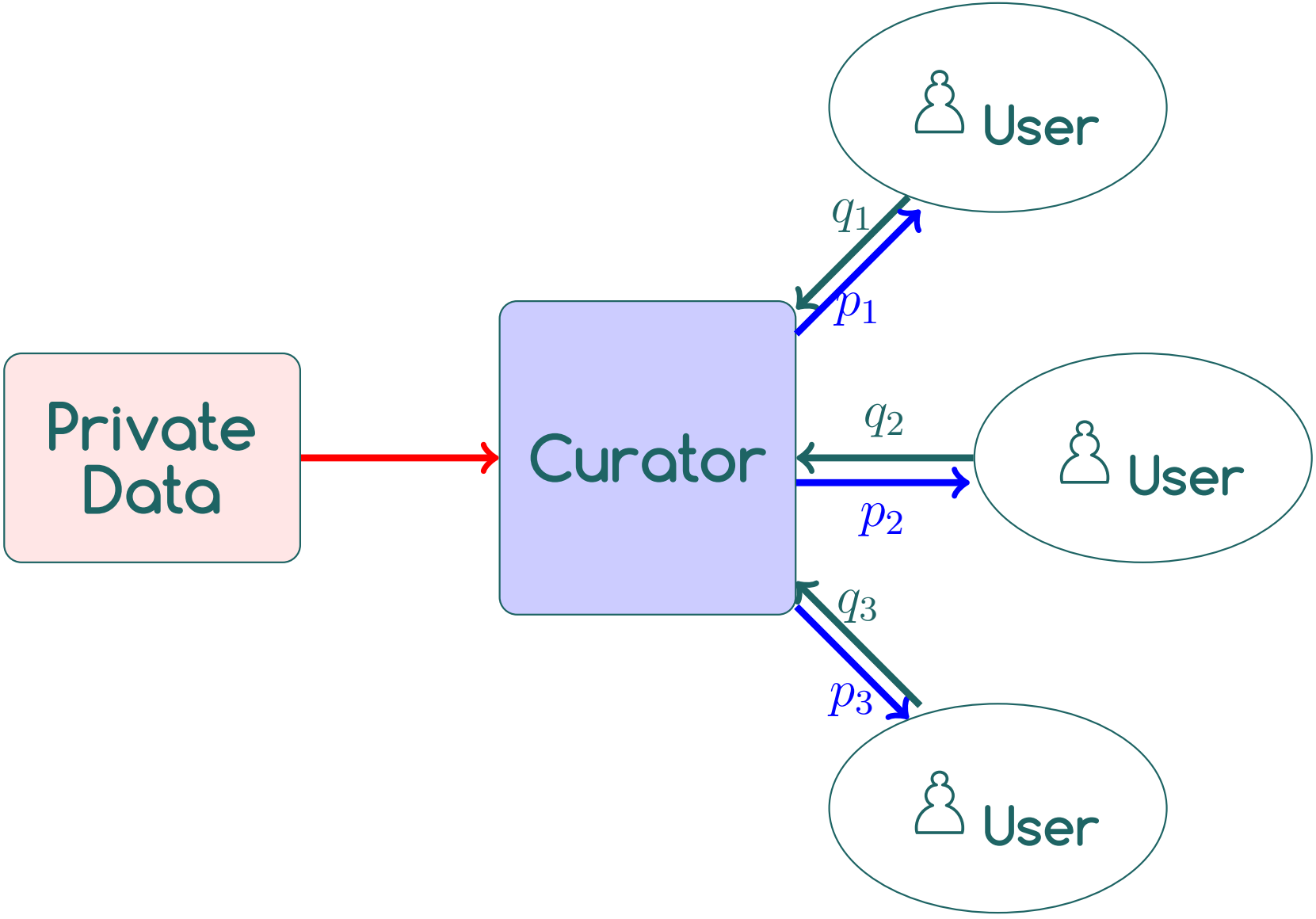


Figure: *The curator architecture for data privacy.*

workflow for private data

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

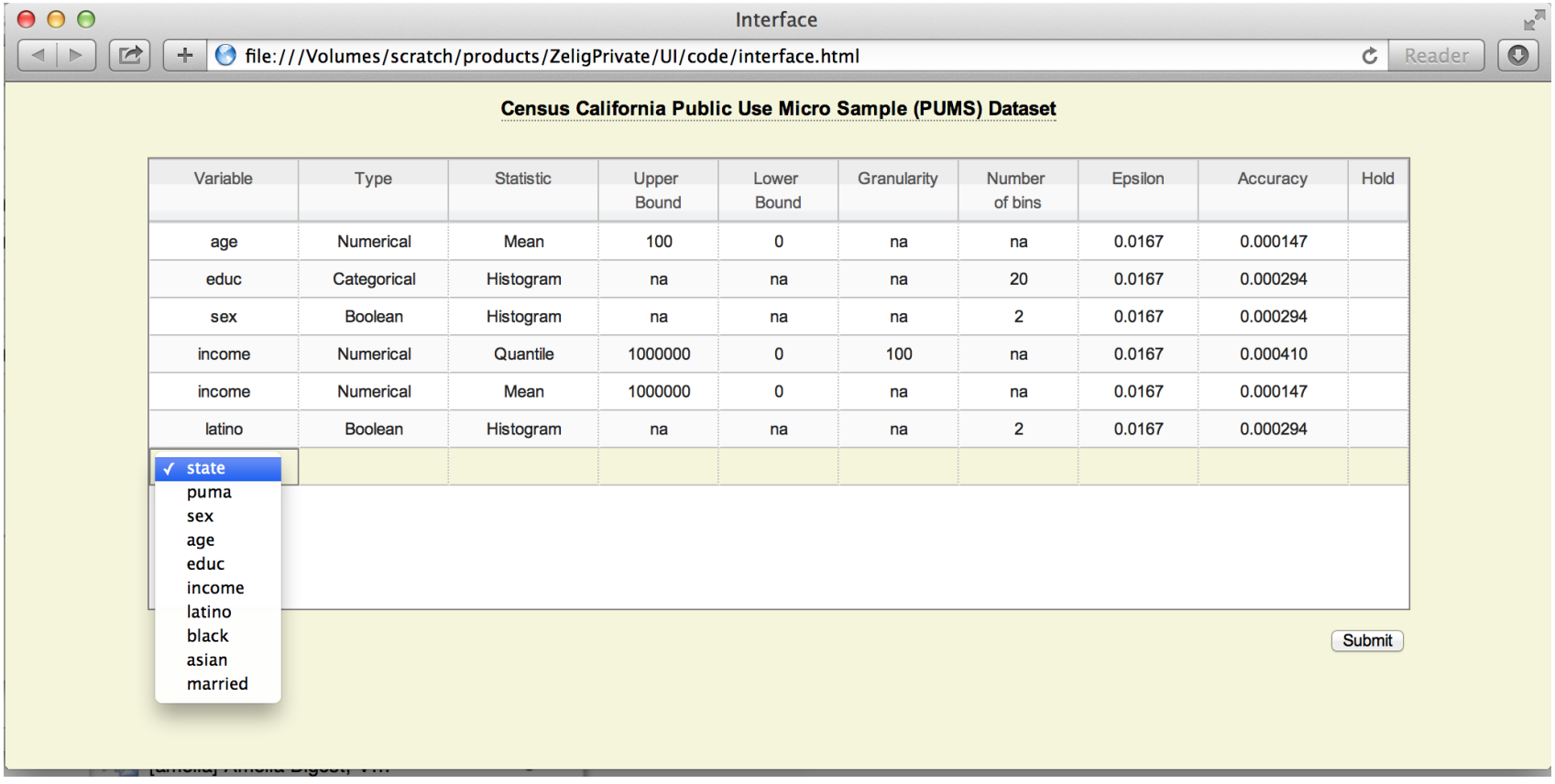
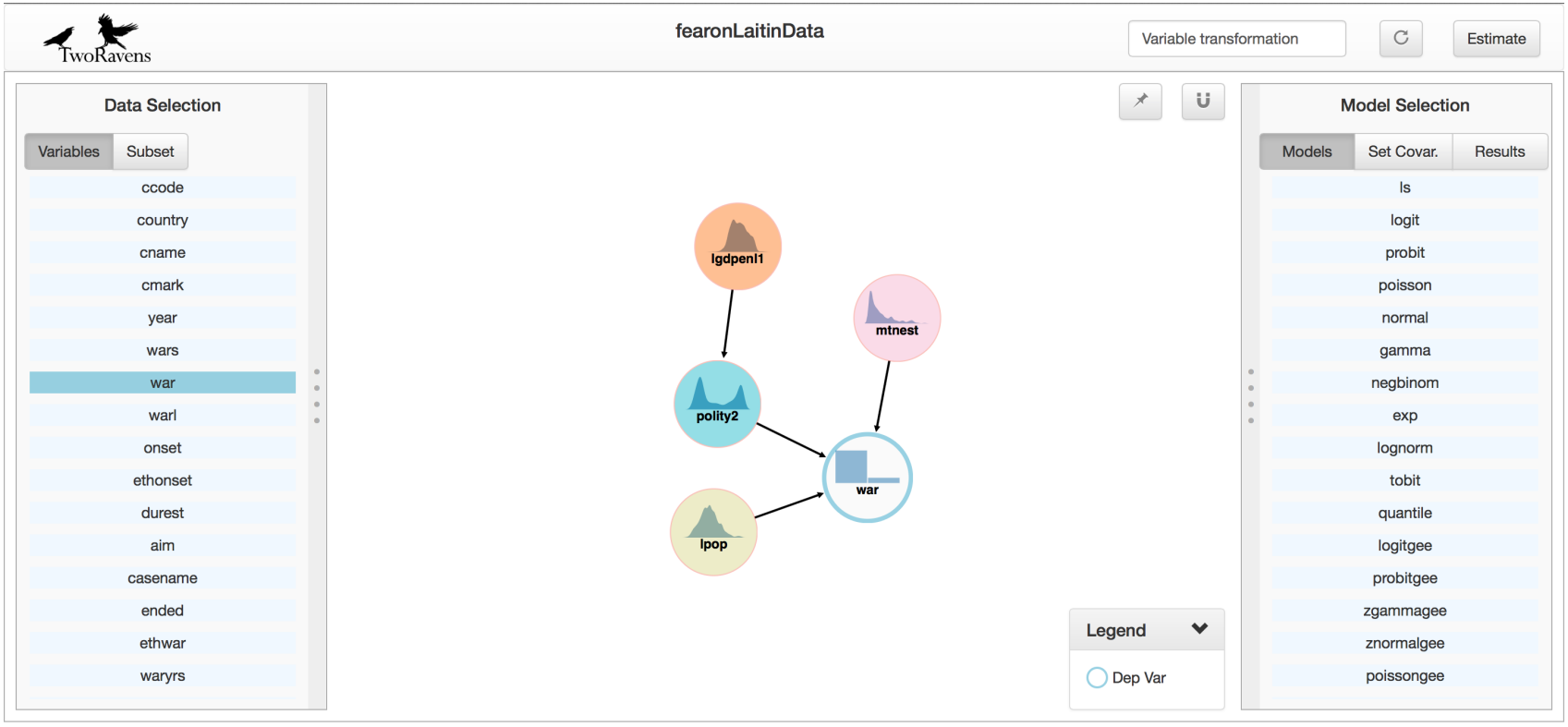
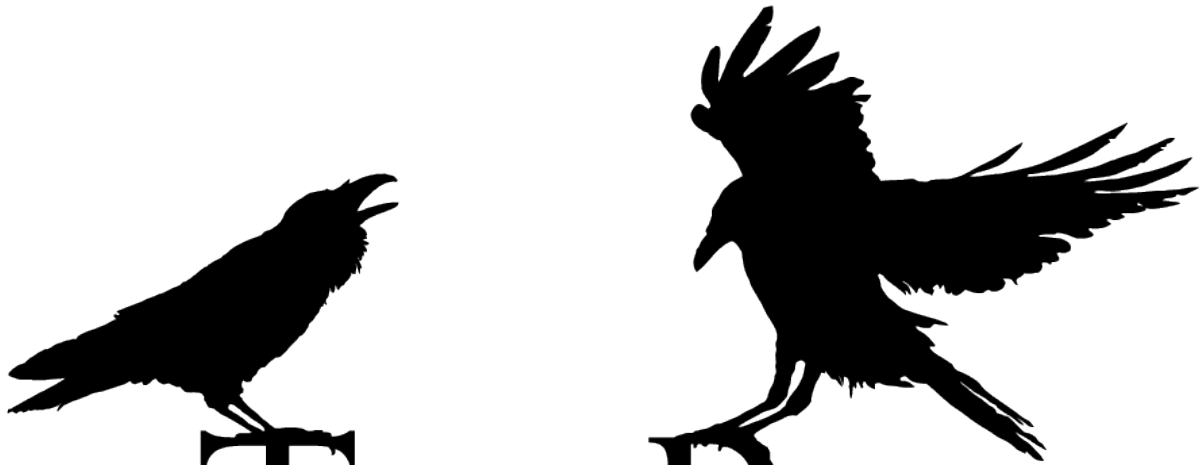


Figure: *Example screen from the interactive privacy budget allocation tool for data depositors.*

The TwoRavens Interface





TwoRavens

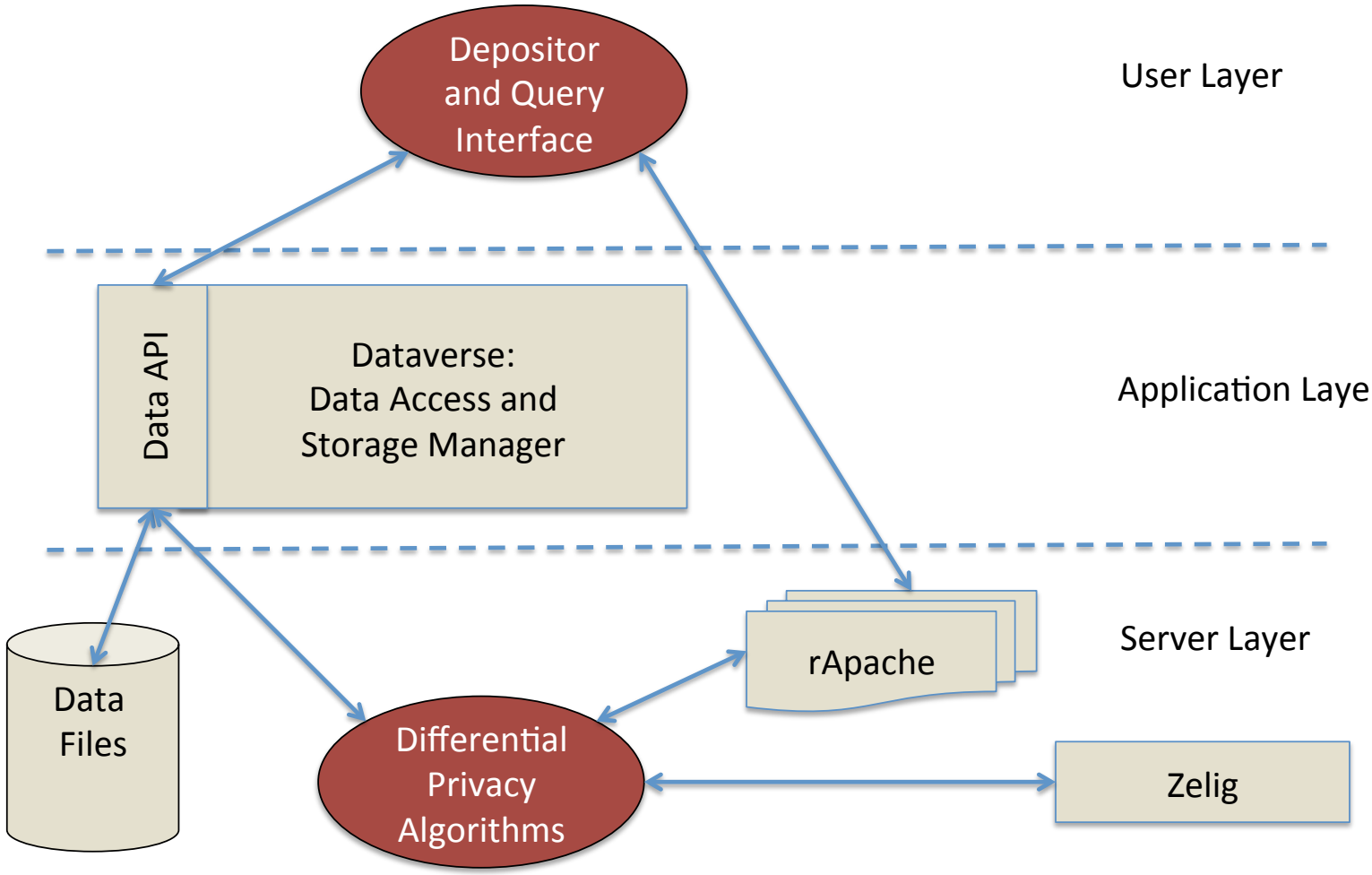
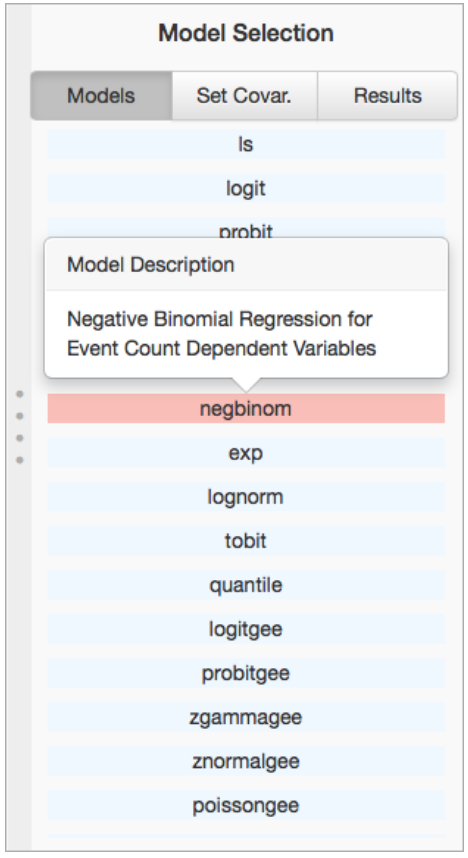
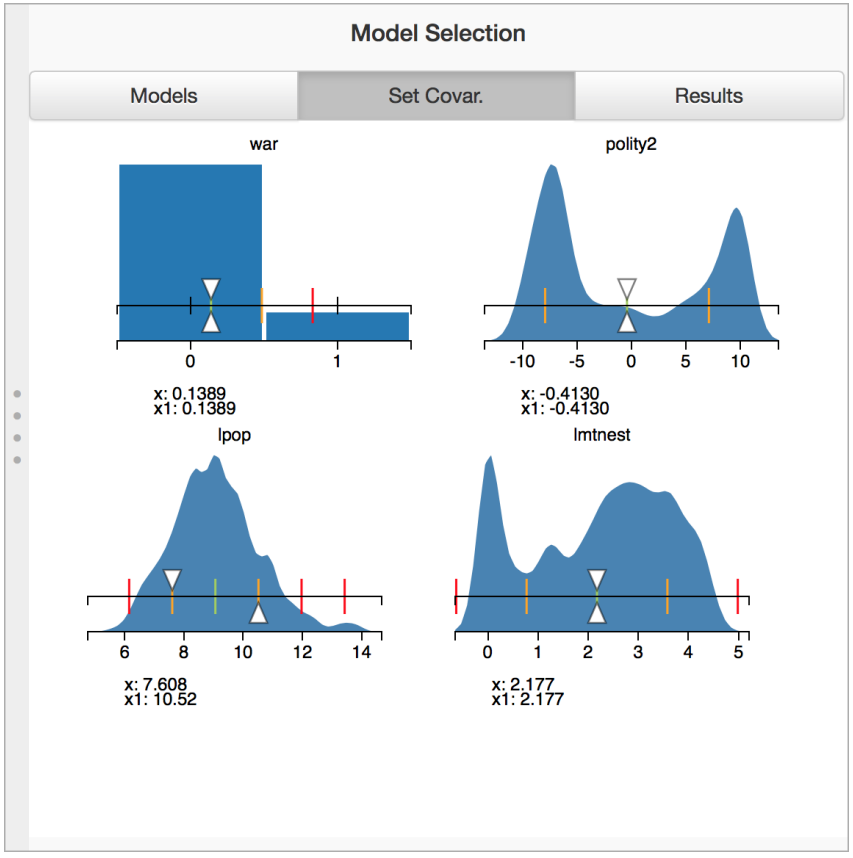


Figure: *Privacy architecture for secure curator interfaces.*

Integration with Zelig





conclusion

- Privacy is fundamental to maintaining the trust of subjects in a big data world

conclusion

- Privacy is fundamental to maintaining the trust of subjects in a big data world
- Differential Privacy is one important tool that offers the strongest privacy guarantees
- It requires new ways of thinking about our statistical estimators
- It requires a new architecture for interacting with data