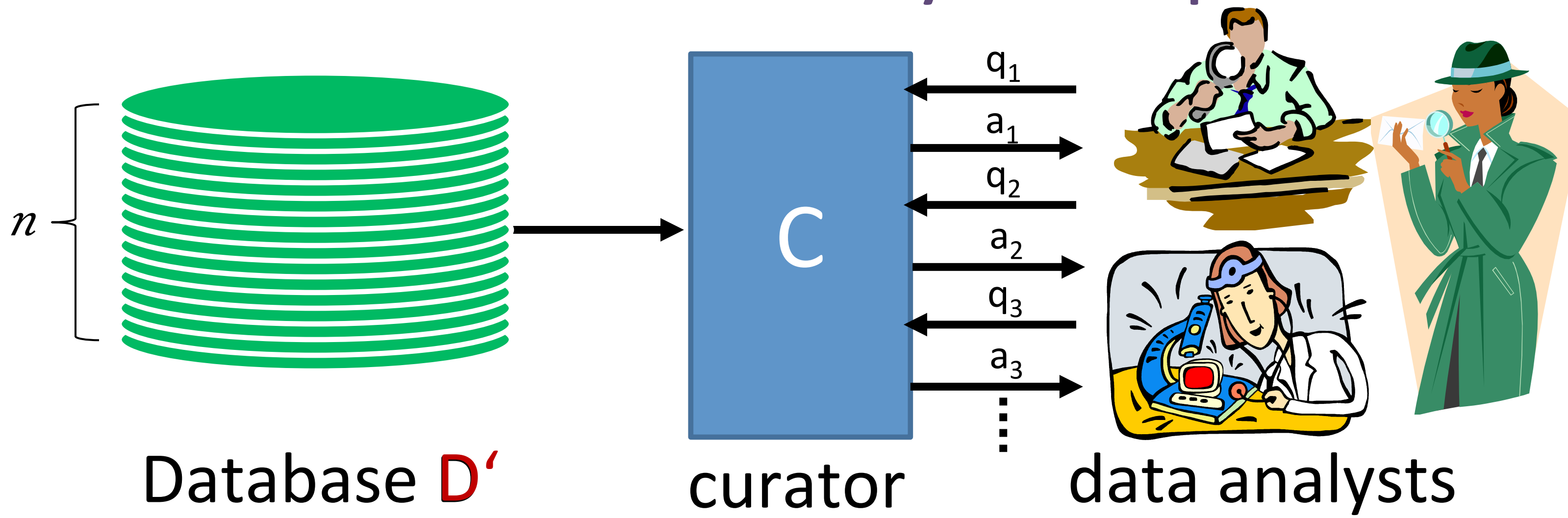
Current Developments in Differential Privacy

Salil Vadhan
Center for Research on Computation & Society
School of Engineering & Applied Sciences
Harvard University

Differential Privacy: Recap

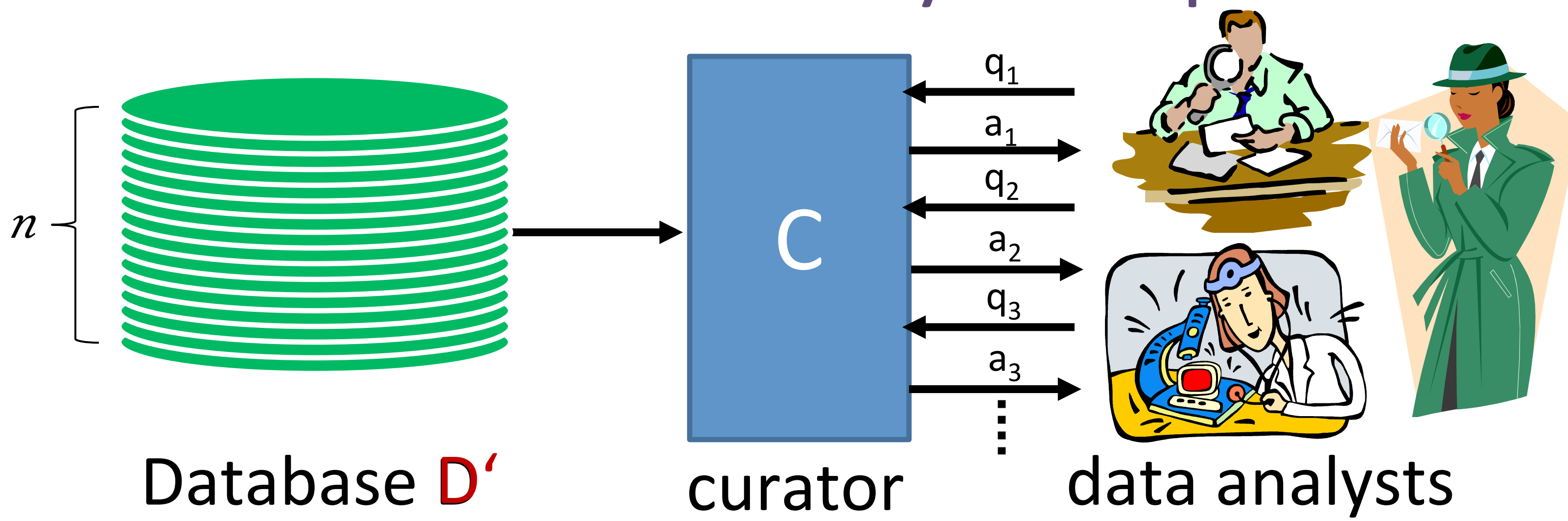


Def [DMNS06]: A randomized algorithm C is ε-differentially private iff for all databases D, D’ that differ on one row, and all query sequences $q_1,\ldots,q_t$

Distribution of $C(D,q_1,\ldots,q_t) \approx_{\varepsilon}$ Distribution of $C(D',q_1,\ldots,q_t)$

“My data has little influence on what the analysts see”

Differential Privacy: Recap

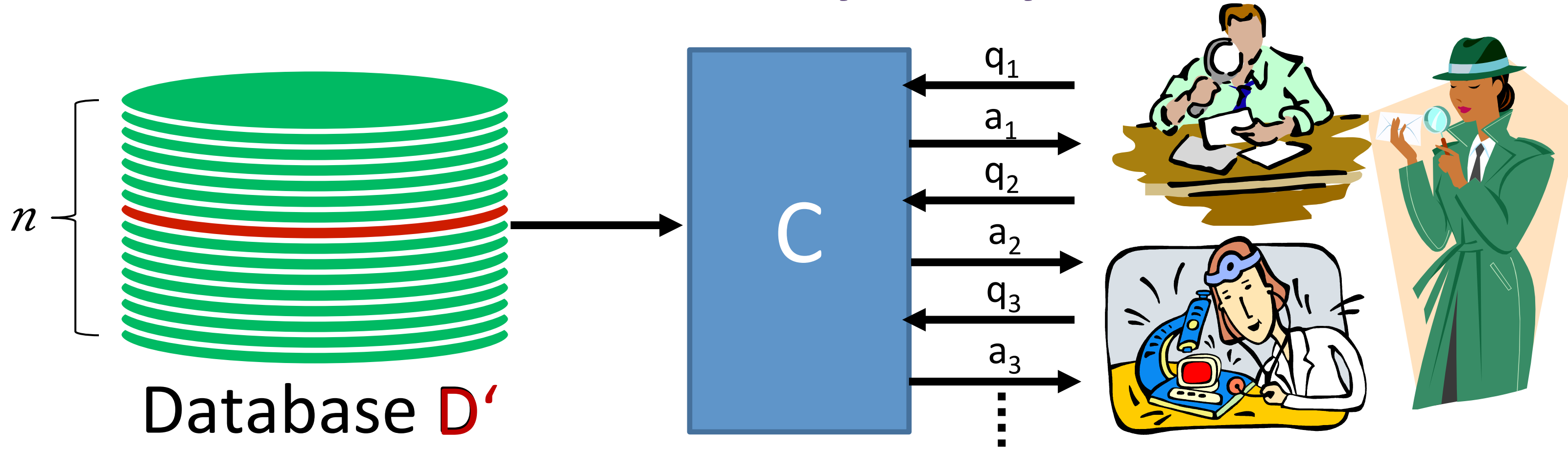


Def [DMNS06]: A randomized algorithm C is ε-differentially private iff
for all databases D, D' that differ on one row,
all query sequences $q_1,\ldots,q_t$, and
all sets $T \subseteq R^t$,

$$\Pr[C(D,q_1,\ldots,q_t)\in T] \lesssim (1+\varepsilon) \cdot \Pr[C(D',q_1,\ldots,q_t)\in T]$$

ε small constant, e.g. $\varepsilon = .01$

Differential Privacy: Key Points

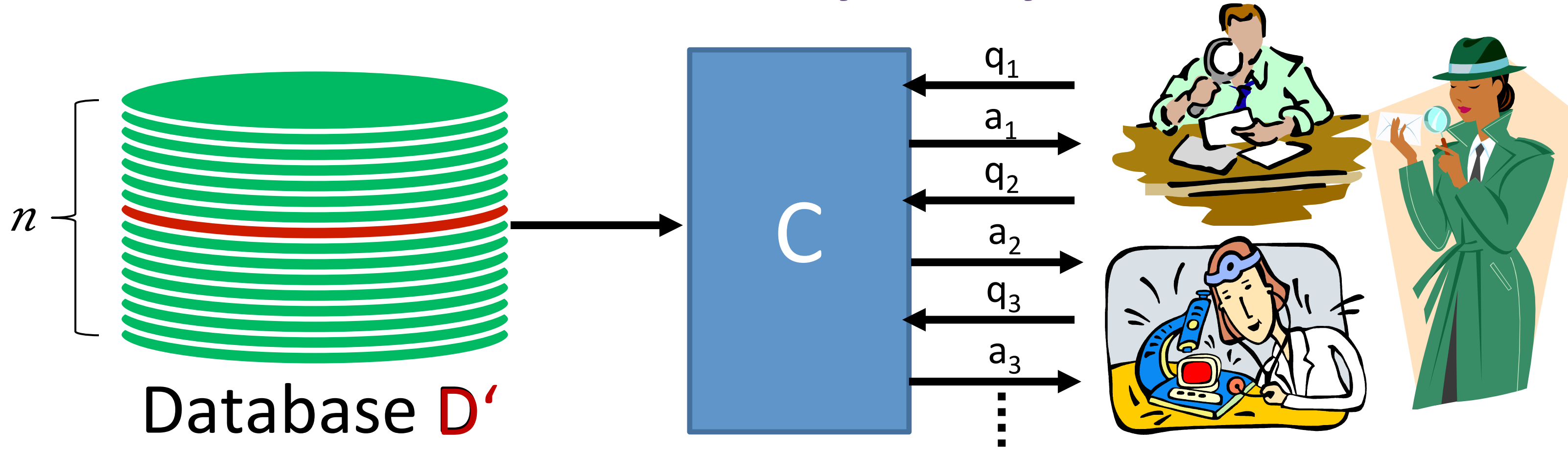


Distribution of $C(D,q_1,\ldots,q_t) \approx_{\varepsilon}$ Distribution of $C(D',q_1,\ldots,q_t)$

- Idea: inject random noise to obscure effect of each individual
  - Not necessarily by adding noise to answer!

- Good for Big Data: more utility and more privacy as $n \to \infty$.

Differential Privacy: Key Points



Distribution of $C(D,q_1,\ldots,q_t) \approx_{\varepsilon}$ Distribution of $C(D',q_1,\ldots,q_t)$

- Strong guarantees: for all databases, regardless of adversary's auxiliary knowledge

- Scalable: don't require privacy expert in the loop for each database or release

Some Differentially Private Algorithms

- histograms [DMNS06]
- contingency tables [BCDKMT07, GHRU11, TUV12, DNT14],
- machine learning [BDMN05,KLNRS08],
- regression & statistical estimation [CMS11,S11,KST11,ST12,JT13]
- clustering [BDMN05,NRS07]
- social network analysis [HLMJ09,GRU11,KRSY11,KNRS13,BBDS13]
- approximation algorithms [GLMRT10]
- singular value decomposition [HR12, HR13, KT13, DTTZ14]
- streaming algorithms [DNRY10,DNPR10,MMNW11]
- mechanism design [MT07,NST10,X11,NOS12,CCKMV12,HK12,KPRU12]
- …

See Simons Institute Workshop on Big Data & Differential Privacy 12/13

Amazing Possibility I: Synthetic Data

n (rows), d (columns)

≥40	Smoke?	Cancer?
0	1	1
1	1	0
1	0	0
1	1	1
0	1	0
1	1	1

→ C →

≥40	Smoke?	Cancer?
1	0	1
1	1	1
0	1	0
0	1	1
1	1	0

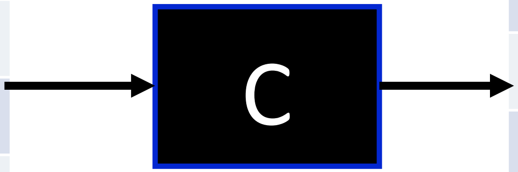
"fake" people

Theorem [BLR08,HR10]: If $n \gg d$, can generate diff. private synthetic data preserving exponentially many statistical properties of dataset (e.g. fraction of people w/each set of attributes).

- Computational complexity is a challenge [DNRRV09,UV11,U13]
- Practical implementations in [HLM12,GGHRW14]

Amazing Possibility I: Synthetic Data

≥40	Smoke?	Cancer?
0	1	1
1	1	0
1	0	0
1	1	1
0	1	0
1	1	1



≥40	Smoke?	Cancer?
1	0	1
1	1	1
0	1	0
0	1	1
1	1	0

“fake” people

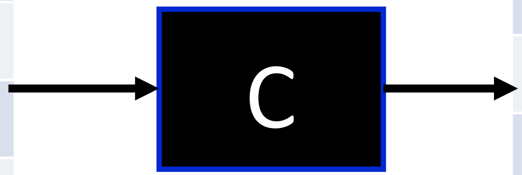
Q: How could this be possible?

Would be easy if we could compromise privacy of “just a few” people.

- A few random rows preserves many statistical properties.
- Differential privacy doesn’t allow this.

Amazing Possibility I: Synthetic Data

≥40	Smoke?	Cancer?
0	1	1
1	1	0
1	0	0
1	1	1
0	1	0
1	1	1



≥40	Smoke?	Cancer?
1	0	1
1	1	1
0	1	0
0	1	1
1	1	0

“fake” people

Q: How could this be possible?

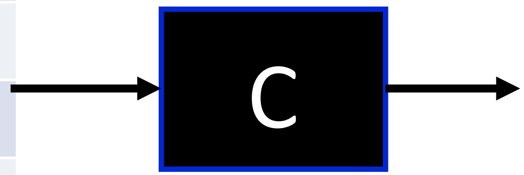
Construct a “smooth” distribution on synthetic datasets (via [MT07])

- Put higher probability on synthetic datasets that agree more with real dataset on statistics of interest.
- Ensure (Probability of each inaccurate synthetic dataset) $\times$(# of synthetic datasets) is very small.

Amazing Possibility II: Statistical Inference & Machine Learning

≥40	Smoke?	Cancer?
0	1	1
1	1	0
1	0	0
1	1	1
0	1	0
1	1	1

n (rows), d (columns)



Hypothesis h or model θ about world, e.g. rule for predicting disease from attributes

Theorem [KLNRS08,S11]: Differential privacy for vast array of machine learning and statistical estimation problems with little loss in convergence rate as $n \to \infty$.

- Optimizations & practical implementations for logistic regression, ERM, LASSO, SVMs in [RBHT09,CMS11,ST13,JT14].

Challenges for DP in Practice

- Accuracy for "small data" (moderate values of n)
- Modelling & managing privacy loss over time
  - Especially over many different analysts & datasets
- Analysts used to working with raw data
  - One approach: "Tiered access" – DP for wide access, raw data only by approval with strict terms of use (cf. Census PUMS vs. RDCs)
- Cases where privacy concerns are not "local" (e.g. privacy for large groups) or utility is not "global" (e.g. targeting)
- Matching guarantees with privacy law & regulation
- …

Some Efforts to Bring DP to Practice

- CMU-Cornell-PennState "Integrating Statistical and Computational Approaches to Privacy"
  - See http://onthemap.ces.census.gov/
- UCSD "Integrating Data for Analysis, Anonymization, and Sharing" (iDash)
- UT Austin "Airavat: Security & Privacy for MapReduce"
- UPenn "Putting Differential Privacy to Work"
- Stanford-Berkeley-Microsoft "Towards Practicing Privacy"
- Duke-NISSS "Triangle Census Research Network"
- Harvard "Privacy Tools for Sharing Research Data"
- …

A project at Harvard: "Privacy Tools for Sharing Research Data"

http://privacytools.seas.harvard.edu/

- Computer Science, Social Science, Law, Statistics

- Goal: to develop technological, legal, and policy tools for sharing of personal data for research in social science and other fields.

- Supported by an NSF Secure & Trustworthy Cyberspace "Frontier" grant and seed funding from Google.

CRCS Center for Research on Computation and Society

IQSS
The Institute for Quantitative Social Science
HARVARD UNIVERSITY

VE RI TAS
Berkman
The Berkman Center for Internet & Society at Harvard University

DATA PRIVACY LAB

IQSS Dataverse Network >

POWERED BY THE Dataverse Network™ PROJECT v. 3.2

Murray Research Archive Original Collection Dataverse

Create Account Log In

INTERGENERATIONAL STUDIES, 1932-1982

hdl:1902.1/00627UNF:3:jYQzhUZ5MxpaKGMvlojITA==

Version: 5– Released: Tue Jun 19 13:50:23 EDT 2012

Cataloging Information | **DATA & ANALYSIS** | Comments (0) | Versions

Use the check boxes next to the file name to download multiple files. Data files will be downloaded in their default format. You can also download all the files in a category by checking the box next to the category name. You will be prompted to save a single archive file. Study files that have restricted access will not be downloaded.

Access to some files is restricted, and those files are not available for downloading. Check the Terms of Use for more information.

Select all files | Download All Selected Files

1. Documentation

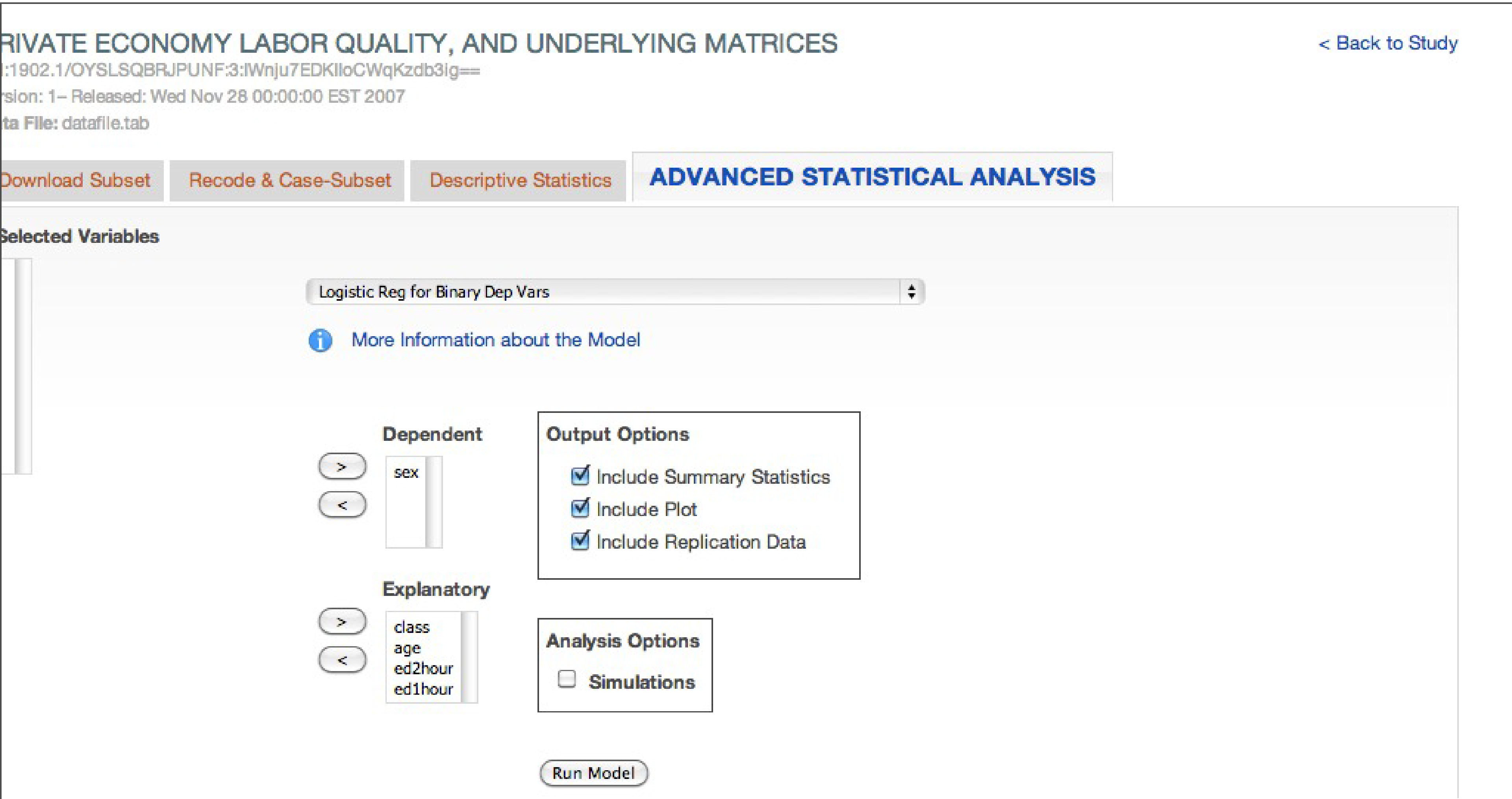
File	Access	Description
00627IHD-InterGenerational-C... Adobe PDF - 41 MB - ...	Download	Description of coded data variables
00627IHD-InterG... Adobe PDF - 7 M...	Download	Blank measures for study
00627IHD-InterG... Adobe PDF - 173 K...	Download	Overview: abstract, research methodology, publications, and other info.

2. Berkeley Data

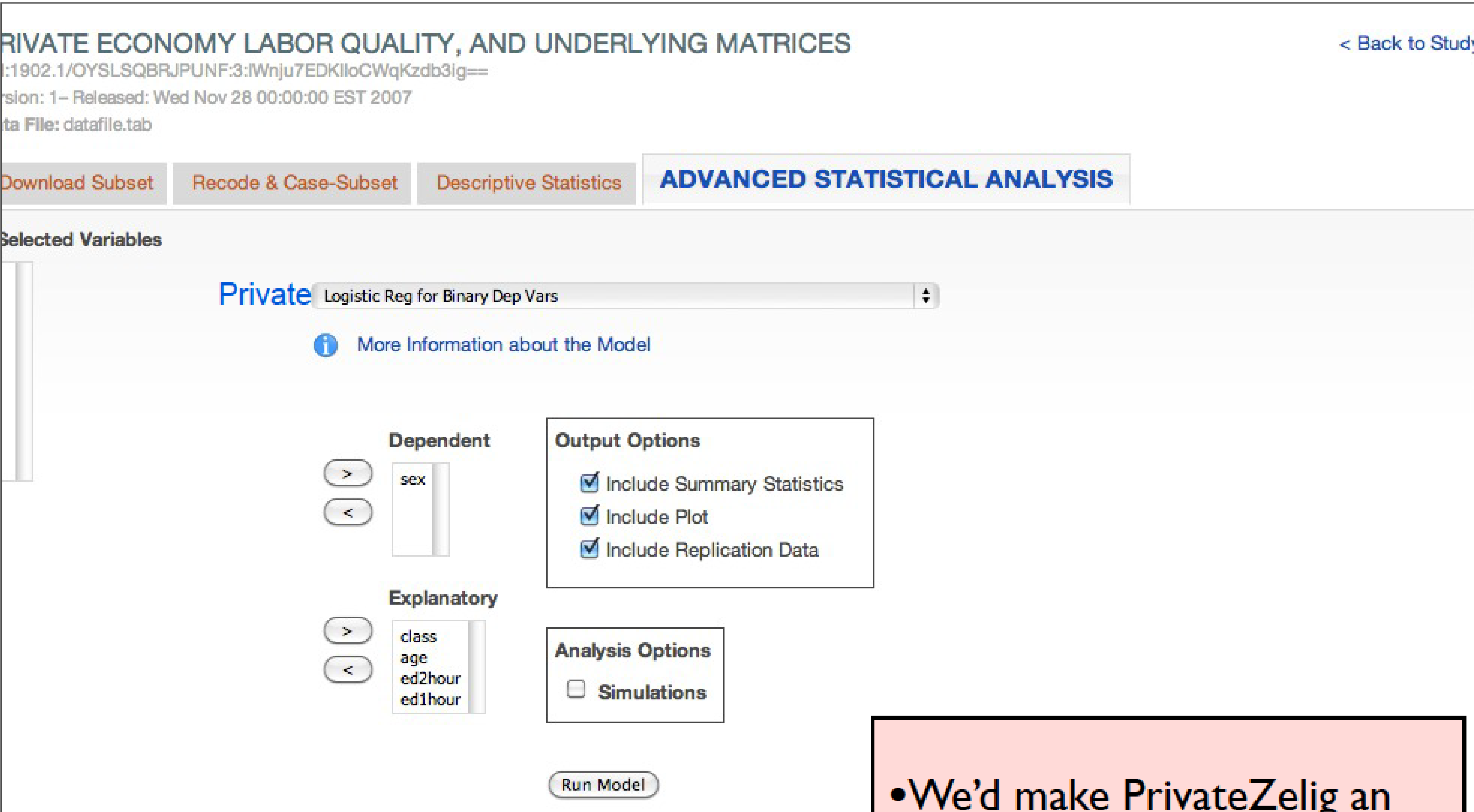
File	Access	Description
00627IHD-InterGenerational-BerkSpou-Data.por SPSS Portable - 29 KB - 0 downloads	Restricted	Data on Spouses in Berkeley Sample in SPSS Portable Format
00627IHD-InterGenerational-BerkSpou-Data.tab Tab Delimited - 22 KB - ... TABULAR DATA 47 ...	Restricted	Data on Spouses of Berkeley Sample in Tab ...
00627IHD-InterGenerational-BerkSubj-Data.por SPSS Portable - 217 KB - 0 downloads	Restricted	Data on Subjects in Berkeley Sample in SPSS Portable Format

Many datasets are restricted due to privacy concerns

Goal: use differential privacy to widen access



For non-restricted datasets, can run many statistical analyses ("Zelig methods") through the Dataverse interface, without downloading data.



- We'd make PrivateZelig an option, the interface would stay roughly the same
- For sensitive datasets PrivateZelig might be the only option

Dataverse Analysis

You could get information about what alg we ran, the privacy param, etc.

The following are the results of your requested analysis.

Summary Results

privatezelig(formula=..., model="logit", DPalg="smith", eps=0.1)

- Call: zelig(formula = sex ~ class + age + ed1hour + ed2hour, model = "logit", data = data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-8.4904	0.0000	0.0000	0.0001	8.4904

Coefficients:

Analysis would come back in the same format

	Estimate	Std. Error	z value	Pr(>|z|
(Intercept)	2.0761e+13	2.5442e+13	0.8160	0.4145
class	5.9152e-03	3.9310e-01	0.0150	0.9880
age	-2.0761e+13	2.5442e+13	-0.8160	0.4145
ed1hour10012835	4.1522e+13	5.0883e+13	0.8160	0.4145
ed1hour100285552	8.3044e+13	1.0177e+14	0.8160	0.4145
ed1hour1004600704	6.2283e+13	7.6325e+13	0.8160	0.4145
ed1hour100926200	6.2283e+13	7.6325e+13	0.8160	0.4145
ed1hour1011177792	1.0381e+14	1.2721e+14	0.8160	0.4145
ed1hour1011535104	1.0381e+14	1.2721e+14	0.8160	0.4145

Conclusions

Differential Privacy offers

- Strong, scalable privacy guarantees
- Compatibility with many types of “big data” analyses
- Amazing possibilities for what can be achieved in principle

There are some challenges, but reasons for optimism

- Intensive research effort from many communities
- Some successful uses in practice already
- Differential privacy easier as $n \to \infty$