Economics of Evidence:

Public Sector Problems and Solutions

Daniel L. Goroff

Vice President and Program Director, Alfred P. Sloan Foundation

Informed by grantees such as: Cynthia Dwork, Aaron Roth, Adam Smith, John Abowd, Jerry Reiter, Sahlil Vadhan, Micah Altman, Gary King, Alessandro Acquisti, and at ICPSR, IQSS, BITSS, COS, ADRN…

Opinions are not necessarily theirs or those of the Sloan Foundation.

What Problem are We Trying to Solve?

1. Public Goods Problem for Open Data

   Due to free riding and financial challenges.

2. Externality/Spillover Problem for Evidence

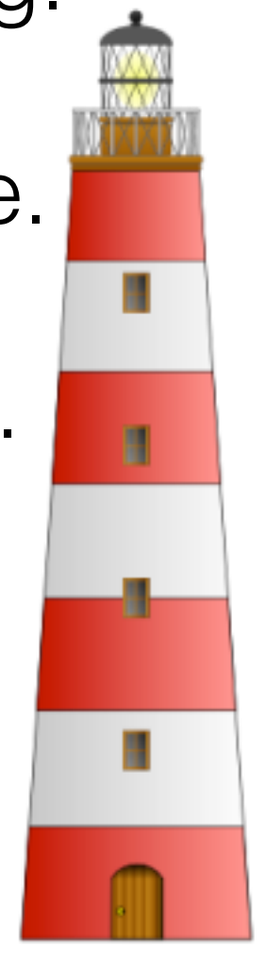
   Due to inevitable validity and privacy leakage.

3. Transaction Costs Problem for Researchers

   Due to lack of trusted intermediary platforms.

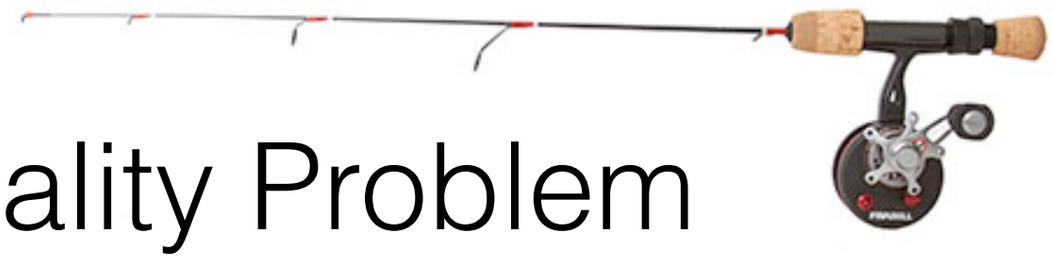
Data & the “Public Goods” Problem

- Open data is a “public good,” technically speaking.
- I.e., a commodity that’s non-rival & non-excludable.
- E.g., lighthouses, parks, discoveries, defense, etc.
- Problem is to finance and sustain public goods.
- Solutions to free riding are taxes or philanthropy.
- Works for look-up data: SDSS, Wikipedia, GPS, etc.



Evidence & the “Externality” Problem

- Data is not a “public good” (excludable).
- Evidence for policy isn’t either (actually rival).
  Need models, hypotheses, and causal inference.
- “Externality” or “spillover” is when you affect others without their choice, e.g., air or water pollution.
- Every query answered leaks *privacy* and *validity*!
- Solution: regulate bad behavior, facilitate good.



Accuracy & the Externality Problem

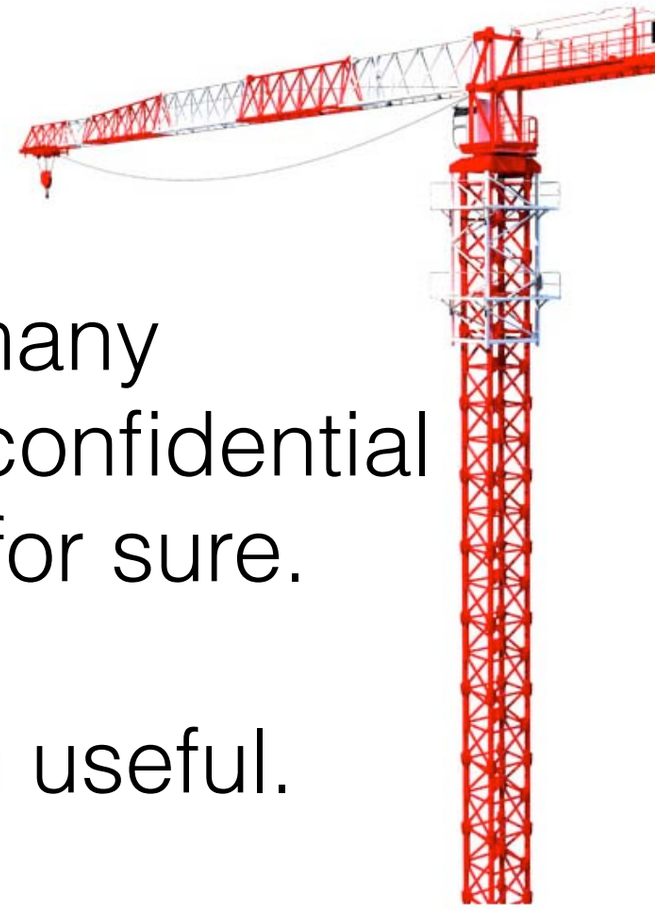
- Validity of testing a hypothesis against a null H?
  Reject null if p= prob of data D given H is < .05.

- Say another project tests D against another null H'.
  But should publish only if prob of D given H *or* H'<.05.

- Or try 100 tests. Noise should make 5 look significant.
  If put other 95 away, literature will differ from evidence.
  Called *p-hacking*, *hypothesis fishing*, or *data mining*.

- Solutions: Limit access. Or pre-register hypotheses.
  Or use some data to explore, set-asides for testing.
  Or control validity-leakage rate using DP methods…

Facts about Privacy [from DR14]



- Database Reconstruction Theorem: Too many statistics answered too accurately from a confidential database will expose the entire database for sure.

- Data cannot be fully anonymized & remain useful.

- Re-identifying anonymized data is not the only risk.

- Queries over large sets are not protective.

- Query auditing is problematic & provably impractical.

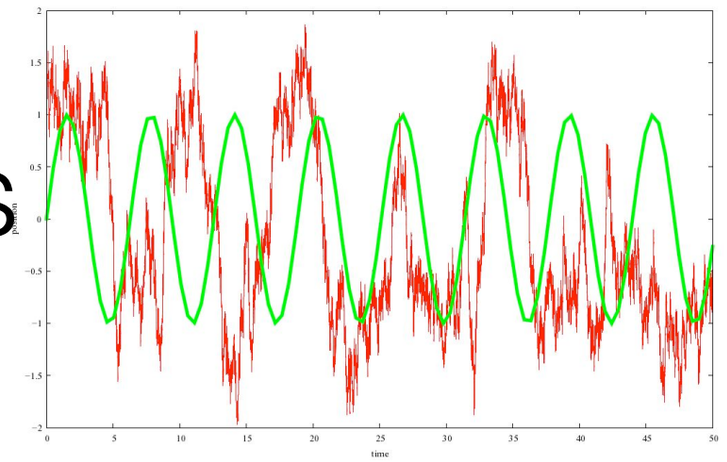
- Neither summary statistics nor ordinary facts are safe.

Privacy Solutions [DMNS 06]

- Idea: allow researchers to ask certain questions about a dataset D to a *mechanism* M that adds noise to the true answer, then gives an approximate answer M(D).

- Definitions: Let $\varepsilon>0$ and let U be a database I cannot see. It has a row for each individual's information. Call a pair of datasets D and D' *neighbors* if they differ in at most one row. Before learning M(U), I have prior beliefs about the odds that U=D vs. U=D'. We say M satisfies ε-*differential privacy* if learning M(U) cannot change those odds by more than a factor of *exp*(ε).

Differential Privacy Properties



- Note: Because *exp*(ε)~1+ε for small ε, this means M(U) tells you almost nothing new about U=D vs D'.

- DP Theorem: There exist useful M that satisfy ε-DP.
  E.g., given a standard statistical question about U, compute the answer then add noise of "size" ε.

- Participation: Anything learned from M(U) or after is essentially the same whether or not your info is in U.

- Composition: Doing M1 then M2 is (ε1+ε2)-DP.

Privacy & the Externality Problem

- Only shows how to regulate the leakage of privacy. Still can't answer too many questions, or researchers could average out the noise. Need a *privacy budget*.

- Small ε means more privacy. But requires more noise. So can ask more questions, but get less accuracy.

- Synthetic Dataset Theorem: Given D, you can run an M that approximately answers certain statistical questions in such a way that researchers can hardly ever tell M(D) from M(D'), even after many queries.

Produce Evidence but Limit Externalities?

- Let data scientists explore away at synthetic data.
- Given a hypothesis so generated, access data to test it using DP to control privacy *and* validity leaks.
- Yes, Differentially Private methods also control overfitting and false positive rates by ignoring D vs D'.
- Thus distinguish between *exploratory* work on data vs. *confirmatory* research that can produce evidence.
- Who will help facilitate all this for researchers?

High Transaction Costs for Researchers

- Gov't can try to reduce such costs: currency, FOIA.
- Administrative data use is now *ad hoc*: Hard to obtain, prepare, protect, supply, sustain, study, link.
- Need trusted intermediaries with sector expertise. Call these *Administrative Data Research Facilities*.
- For gov't *or* proprietary data, e.g., IRIS, Kilts, AISP, CDRC. ADRF's may also help with federal statistics.
- Make a network, call it the ADRN, to share standards and best practices for producing reliable evidence.

Basic References

- Dinur and Nissim (2003) [link]
- Dwork, McSherry, Nissim, and Smith (2006) [link]
- Machanavajjhala, Kifer, Abowd, Gehrke, and Vilhuber (2008) [link]
- Dwork and Roth (2014) [link]
- Dwork, Feldman, Hardt, Petassi, Reingold, and Roth (2015) [link]
- Goroff (2015) [link]