ALFRED P. SLOAN FOUNDATION

Reproducibility and Privacy:

Research Data Use, Reuse, and Abuse

Daniel L. Goroff

Opinions here are his own.

But based on the work of grantees.

Alfred P. Sloan

- Organized and ran GM
- Sloan School, Sloan Kettering, too
- Foundation (public goods business)
- Emphasized role of data in decision making

Privacy Protecting Research

How can work on private data be reproducible?

	Input	Computation	Output	Protocol Example
1				Open Data
2			X	Data Enclave
3	X		X	Nondisclosure Agreement
4		X		Anonymization
5	X			Randomized Response
6	X	X		Multiparty Computation
7	X	X	X	Fully Homomorphic Encryption
8		X	X	Differential Privacy

Protocols can impose obfuscation at 3 stages: input, computation, or output.

Data Enclave

- Work there
- Data stays
- Release write up if approved
- Irreproducible results!
- Same problem for NDA's.

Privacy Protecting Protocols

	Input	Computation	Output	Protocol Example
1				Open Data
2			X	Data Enclave
3	X		X	Nondisclosure Agreement
4		X		Anonymization
5	X			Randomized Response
6	X	X		Multiparty Computation
7	X	X	X	Fully Homomorphic Encryption
8		X	X	Differential Privacy

De-Identification?

- William Weld, while Governor of Massachusetts, approved the release of de-identified medical records of state employees.

- Latania Sweeney, then a graduate student at MIT, re-identified Weld's records and delivered a list of his diagnoses and prescriptions to his office.

- Try stripping out names, SSNs, addresses, etc.
- But 3 facts–gender, birthday, and zip code–are enough to uniquely identify over 85% of U.S. citizens.
- Including my assistant, a male from 60035 born on 9/30/1988 as above.
- See www.aboutmyinfo.org

Netflix Challenge

- Offered $1m prize for improving prediction algorithm.
- Release “anonymized” training set of >100m records.
- In 2007, researchers began identifying video renters by linking with public databas
- Suite settled in 2010.

NYC Taxi Records

- Last year, NYC's 187m cab trips were compiled and released, complete with GPS start and finish data, distance traveled, fares, tips, etc.
- But the dataset also included hashed but poorly anonymized driver info, including license and medallion numbers, making it possible to determine driver names, salaries, tips, and embarr

Privacy Protecting Protocols

	Input	Computation	Output	Protocol Example
1				Open Data
2			X	Data Enclave
3	X		X	Nondisclosure Agreement
4		X		Anonymization
5	X			Randomized Response
6	X	X		Multiparty Computation
7	X	X	X	Fully Homomorphic Encryption
8		X	X	Differential Privacy

Secure Multiparty Computing

- Am I well off? If make more salary than brother-in-law.
- Say I have two. Can we find our average salary without revealing our numbers?
- Call true salaries $S1, S2, S3$.

SMC Algorithm

- Each of us generates two random numbers and gives one to each of other two people.
- Person i reports Xi, which is Si plus random numbers received minus those given.
- I.e., if person i gives Rij to person j, we have

$$X1 = S1 + (R21 + R31) - (R12 + R13)$$

$$+ \quad X2 = S2 + (R12 + R32) - (R21 + R23)$$

$$+ \quad \underline{X3 = S3 + (R13 + R23) - (R31 + R32)}$$

$$= S1 + S2 + S3$$

SMC Features

Cassandra Hubbart

- Adding the X_i gives sum of the S_i without revealing them, assuming all follow the rules.
- But what if brothers-in-law conspire? They can compute my salary if they share theirs!
- Need a different algorithm for each operation.
- Being contemplated by financial regulators and by some repositories nevertheless.
- Hard to define what “privacy protecting research” should mean...

Privacy Protecting Protocols

	Input	Computation	Output	Protocol Example
1				Open Data
2			X	Data Enclave
3	X		X	Nondisclosure Agreement
4		X		Anonymization
5	X			Randomized Response
6	X	X		Multiparty Computation
7	X	X	X	Fully Homomorphic Encryption
8		X	X	Differential Privacy

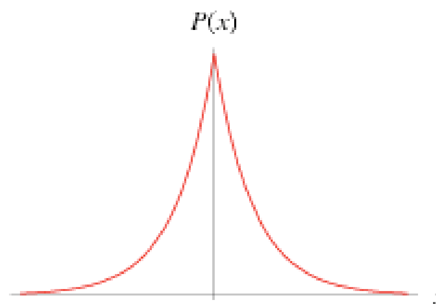
Differential Privacy

- A concept and procedures for allowing aggregate statistical queries while provable protecting individuals' privacy (see Dwork).

- Require that the addition or removal of a single individual from the dataset x should have a nearly zero effect on $M(x)$, the information released.

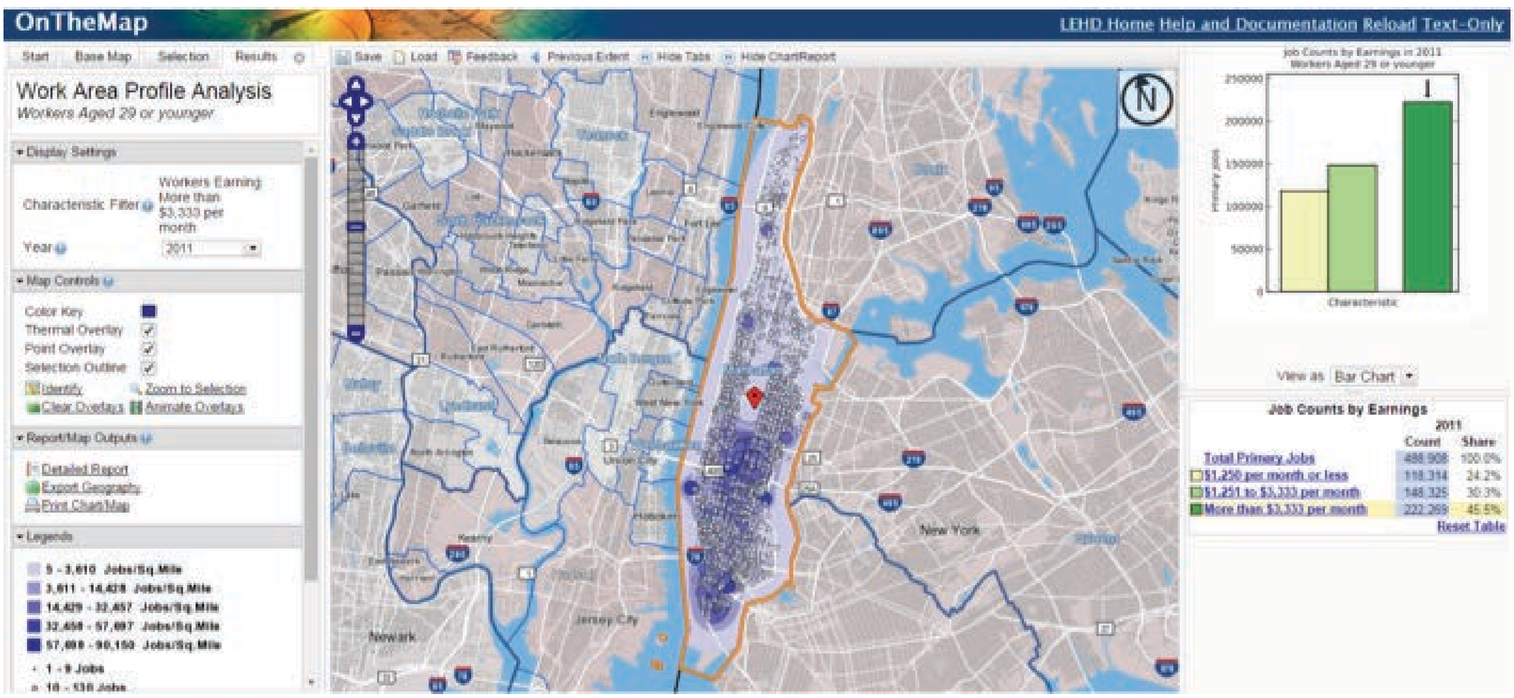
- I.e., you learn almost nothing new about individuals. So eliminates harm from participation (not findings).

DP Properties



- Such mechanisms exist, e.g., by adding Laplace noise so
- Protects against arbitrary risks, not just re-identification or linkage attacks.
- Quantifies privacy loss and tradeoffs.
- Closure under post processing.
- Behaves well under composition, e.g.

Census Data: On the Map



DP Characterization

Let x and y be two "neighboring" datasets, meaning they differ in at most one "row."

E.g., say x contains my information but y doesn't.

A randomized statistical mechanism M satisfies ε– differential privacy if for all such neighbors

where $\Pr(A \mid B)$ is the conditional probability of A given B, and the exponential is

Using New Data

- Tom is either a Salesman or a Librarian.
- You find out he has a Quiet personality.
- Is S or L more likely? I.e., which conditional probability is bigger: Pr (S | Q) or Pr (L | Q) ?

(See Kahneman & Thaler)

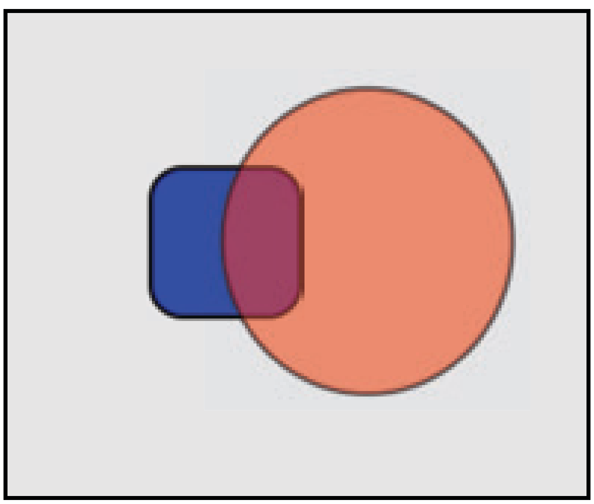
Salesman Problem

- Large "conditional probability" that Tom is Quiet given that he is a Librarian, of course.
- But say Fred is either a Salesman or a Librarian. You know nothing else.
- Now which is more likely for Fred, S or L?

Need one equation from the 1740's.

Conditional Probability

- Imagine two events: A and D
- Prob of A given D is:



= fraction of D that is A.

Pr(Red | Blue)= ?

Pr(Blue | Red)= ?

Bayesian Updating

- Prob of A given D is:
- Bayes Law:
- Odds = Bayes Factor x Base Rate

Base Rate Fallacy

- Fred is either a Salesman or a Librarian?
- There are ~100x as many S as L in the US.
- Maybe 1 in 10 salesmen are quiet.
- So which is more likely for Tom, S or L?
- Odds = Bayes Factor x Base Rate

Bayes and Differential Privacy

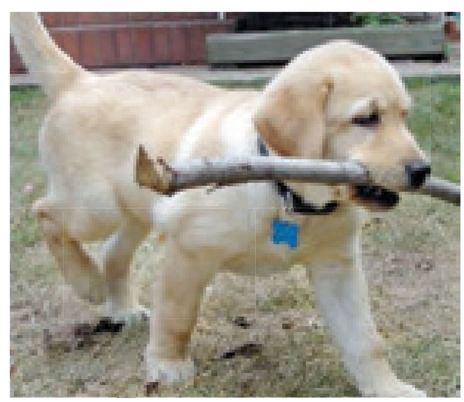
Can a DP study of dataset z tell if my info is there?

Say x has my info, y doesn't but otherwise same.

is the definition of a DP mechanism and implies:

Bayes Pays

- Management Focus Magazine Survey
- 85% of CEOs had a pet in high school!
- Pr(DOG | CEO) vs. Pr(CEO | DOG)
- Pet theories??



Hypothesis Testing

- Talk as if studying Pr(New Hypothesis | Data)
- Classically, study Pr(Data | Null Hypothesis) = p
- Reject Null if this p is small
- Call it “statistically significant”
- Publish if $p < .05$

Fishing, Dredging, P-Hacking

- May try many hypotheses H, H', H'', …
- Should not report p = $\Pr(D \mid \text{favorite } H)$
- Should report $\Pr(D \mid H \text{ or } H' \text{ or } H'' \text{ or } \ldots)$
- Much bigger!
- Hard to correct

for explorations.

- Preregistration?

Limiting False Discovery

- Publish if $p = \Pr(\text{Data} \mid \text{Null Hypothesis}) < .05$
- Often test multiple hypotheses but pretend not.
- Called p-hacking (see Simonsohn) or hypothesis fishing or data dredging. Test depends on data.
  C.f. Lasso Regression, Machine Learning, etc.
- Reuse of data is methodologically problematic.
- Overfitting on idiosyncratic observations.

Thresholdout

- See Dwork et. al. in Science (2015)
- Use training set to explore hypotheses
- Use DP protocols to test on holdout set
- Can then reuse the holdout set many times (because DP estimates compose well)
- ML algorithms can overfit otherwise

Data Science Ideal

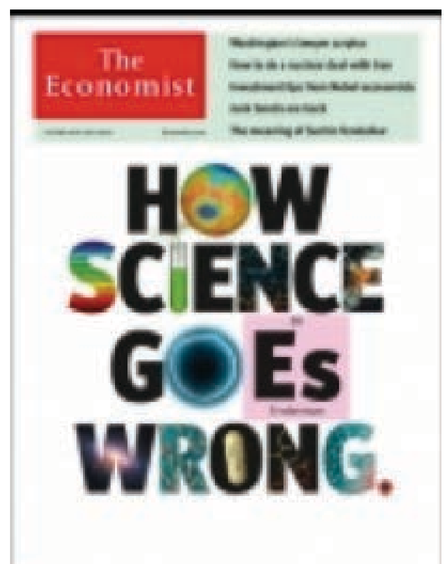
1. Start with hypothesis or model
2. Design a test and register plan
3. Collect unbiased data
4. Analyze data for significance as planned
5. Publish findings, code, and checkable data

Data Science Practice

1. Gain access to some administrativ
2. Try some hypotheses, models, methods, samples
3. Pick best (p-hacking), perhaps with machine help
4. Write it up as if you had followed ideal steps
5. Publish findings, maybe some data and

Trust Empirical Results?

- The Economist magazine says “no.” (2013)
- Says to expect >1/3 of data science is wrong!
- Even if no p-hacking.
- Even if no fraud.
- What is going on?
- What to do about it?



Many Findings Are Wrong?

Let A = finding is true, B = it is false.

Let D = data analysis says the finding is true.

The Economist takes the Base Rate as .

(Here D means where M is a 'mechanism' for analyzing a dataset x.)

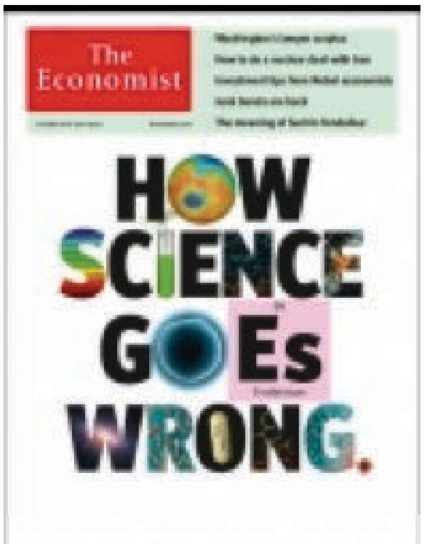
Scholarly Evidence Rules

Here Bayes Factor also called the likelihood ratio.

Want numerator (power) to be (ex ante)

Also want denominator to be (ex post)

So odds increase by a factor of at least

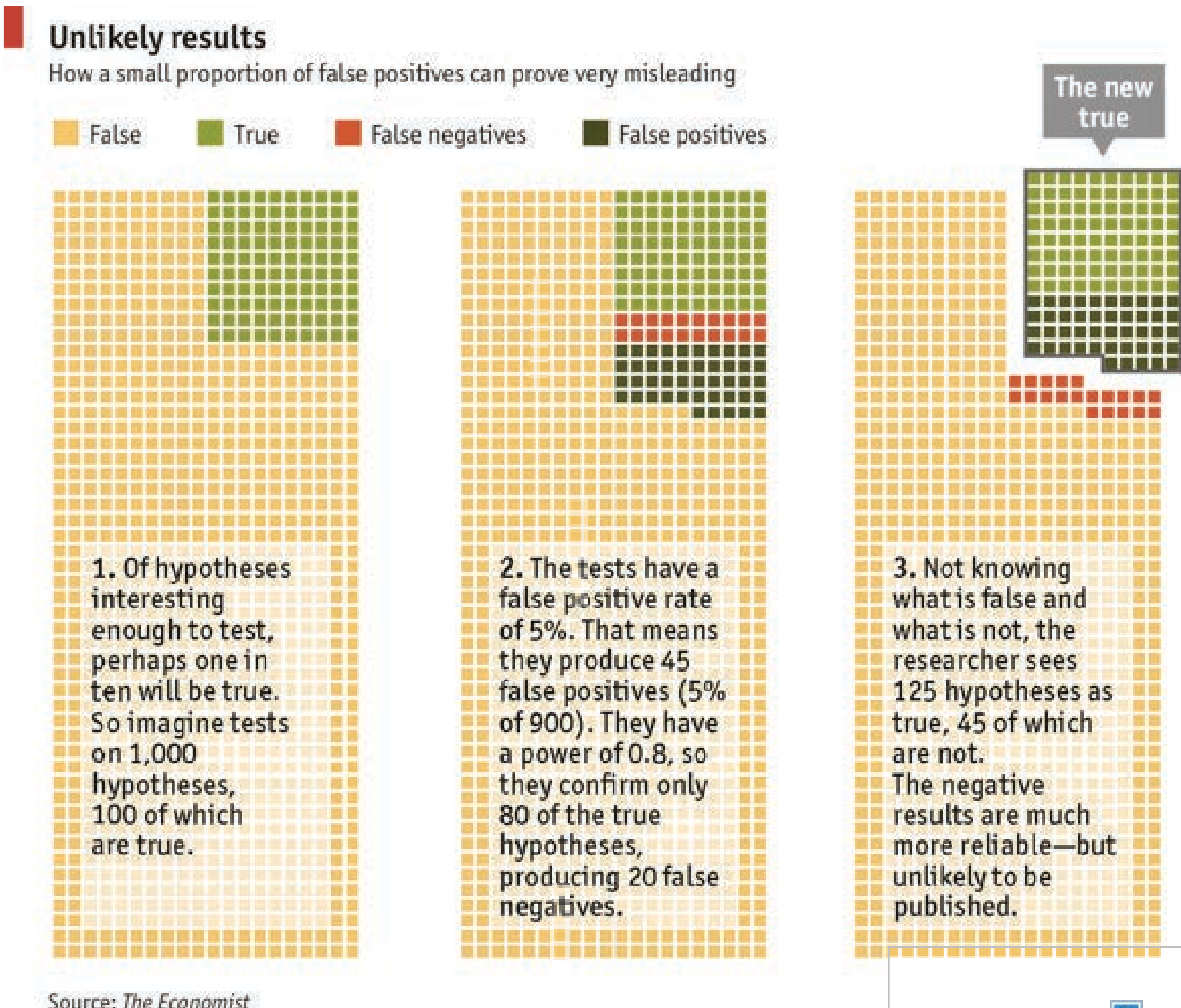


Example (October 19, 2013 issue):

1000	hypotheses to test empirically
100	of these are actually true
.80	acceptable "power" $< \Pr(D \mid T)$
.05	acceptable $p > \Pr(D \mid F)$

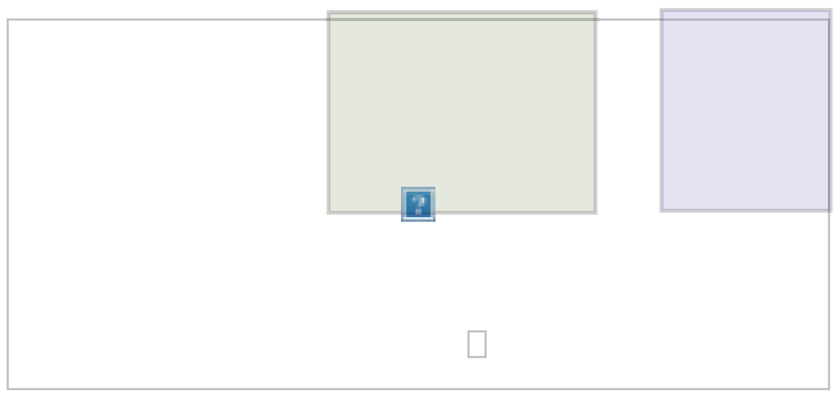
Expected Outcome:

80	confirmed true	=	80% of 100 that are true
+45	false positives	=	5% of 900 that are false
125	publishable	=	80 + 45
.64	fraction true	=	80/125 = 16/25 (16:9 odds)



Unlikely results
How a small proportion of false positives can prove very misleading
False
True
False negatives
False positives
The new true
1. Of hypotheses interesting enough to test, perhaps one in ten will be true. So imagine tests on 1,000 hypotheses, 100 of which are true.
2. The tests have a false positive rate of 5%. That means they produce 45 false positives (5% of 900). They have a power of 0.8, so they confirm only 80 of the true hypotheses, producing 20 false negatives.
3. Not knowing what is false and what is not, the researcher sees 125 hypotheses as true, 45 of which are not. The negative results are much more reliable—but unlikely to be published.
Source: *The Economist*

Improving the Odds

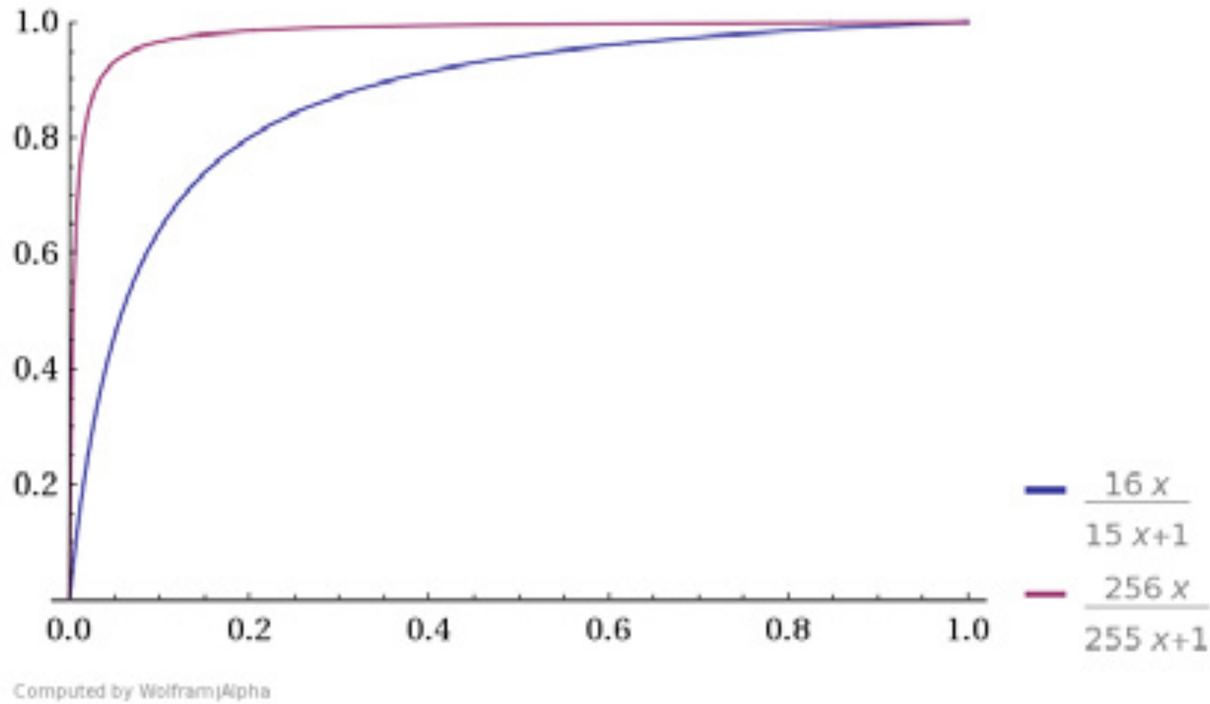


- Prior odds get multiplied by Bayes Factor >16
- What to do if this is not good enough?
- If take p = .01 and alpha = .90, can get BF >90

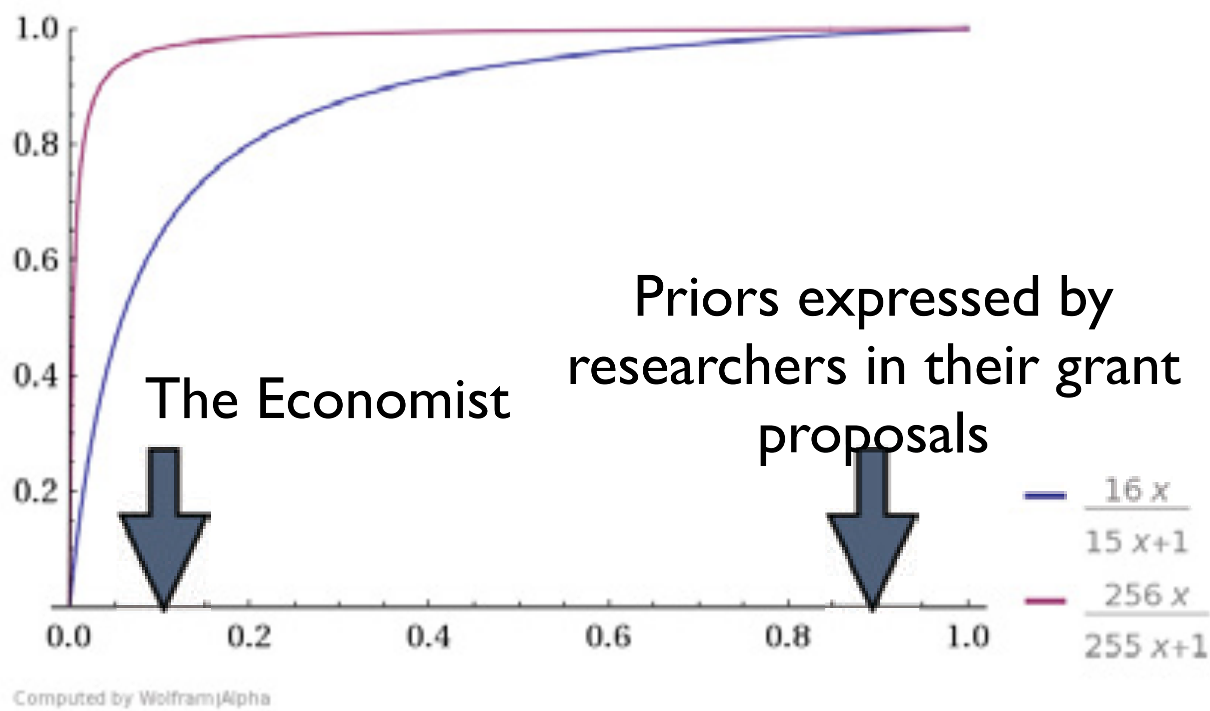
Reproduction Helps

- Prior odds get multiplied by >16 first time.
- Original odds multiplied by > 256 next time.
- In terms of x = fraction true findings/look true, $f(x) = 16x / (1+15x)$ = fraction after one test.
- And $f(f(x)) = 256x / (1+255x)$ after two tests.
- Second tests of 125 that initially look true yield

Fraction of apparently true findings that really are after one test or two



Fraction of apparently true findings that really are after one test or two



In Sum

- Many data science “discoveries” are false.
- Reproduction solves many such problems.
- New tools make reproducibility easier, even with private data.
- Bonus: same privacy tools can also limit false discovery, even with public data.

And Sloan grants are helping to make research results more robust...

Data Science Ideal

1. Start with hypothesis or model: guidelines, exploration
2. Design a test and register plan: RCT's, methodologies
3. Collect unbiased data: administrative, privacy
4. Analyze data for significance as planned: transparency
5. Publish findings, code, and checkable data: repositories

Sloan Reproducibility Projects

- plication of economics lab experiments
- Center for Open Science
- Berkeley Initiative for Transparent Social Science
- Institute for Quantitative Social Science
- DataCite, DataVerse, ICPSR, CNRI
- ode, Research

Sloan Administrative Data Projects

- Council of Professional Associations on Federal Statistics
- LinkedIn, EBay, Mint, etc.
- Software Carpentry, iPython/Jupyter Notebooks
- Open Corporates, Legal Entity Identifiers

Sloan Methodological Projects

- Stan: Open Source Bayesian Software
- Moore/Sloan Data Initiative
- AEA Registry: Study Design & Analysis Plans
- Peer Review of Registered Reports
- Fully Homomorphic Encryption Research

ALFRED P. SLOAN FOUNDATION

- Basic Research
  - Deep Carbon Observatory
  - Microbiology of the Built Environment
- Economic Performance and the Quality of Life
  - Economic Institutions, Behavior, and Performance
  - Working Longer
- STEM Higher Education
  - The Science of Learning
  - Advancement for Underrepresented Groups
- Public Understanding of Science, Technology, & Economics
  - Radio, Film, Television, Books, Theater, New Media
- Digital Information Technology:
  - Data and Computational Research
  - Scholarly Communication
  - Universal Access to Knowledge
- Sloan Research Fellowships
- Civic Initiatives