How to Lie with Big Data (and/or big computations)

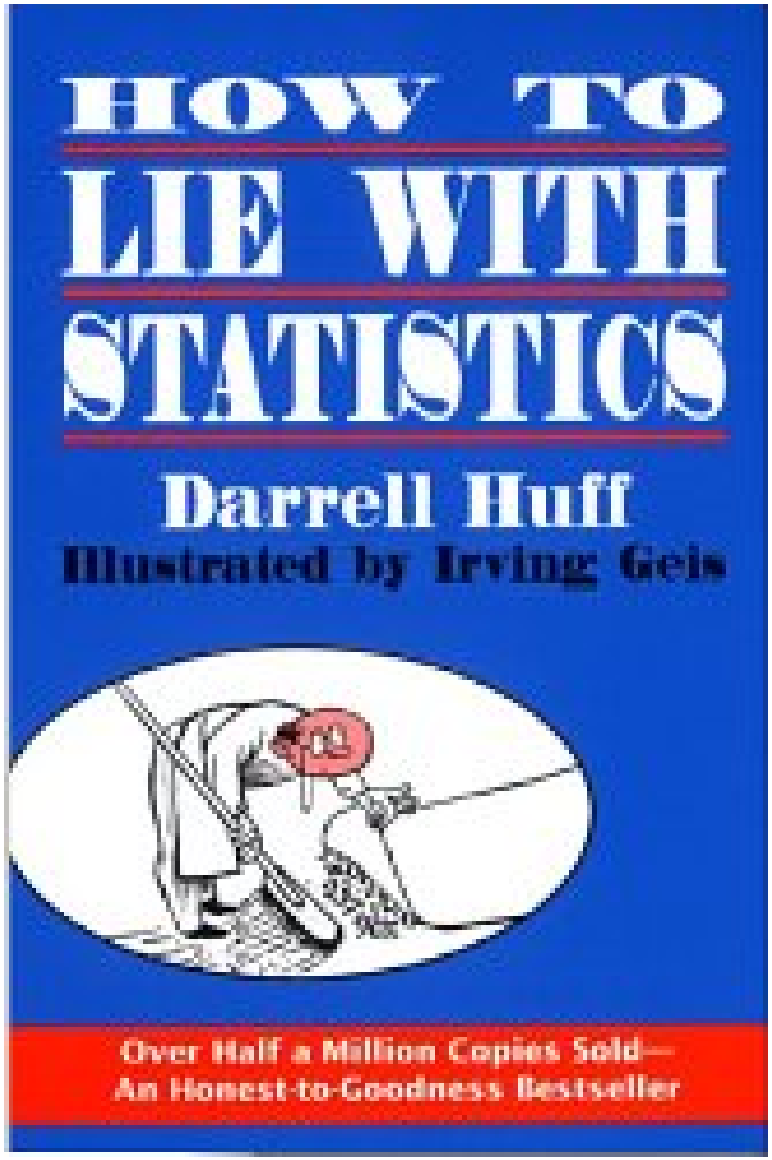
Philip B. Stark

Department of Statistics
University of California, Berkeley

MPE 2013+ Workshop on Global Change
Panel on Data Deluge or Drought (Quality and Quantity)
University of California
Berkeley, CA
19–21 May 2014

Huff, 1954. *How to Lie with Statistics*

1. The Sample with Built-in Bias
2. The Well Chosen Average
3. The Little Figures That Are Not There
4. Much Ado about Practically Nothing
5. The Gee-Whiz Graph
6. The One-Dimensional Picture
7. The Semi-Attached Figure
8. Post hoc rides again
9. How to Statisticulate



Data Deluge Delusions

- bigger is better
- if you have enough data, they speak for themselves
- if you have enough data, Statistics doesn't matter
- if you have enough data, everything has a Gaussian distribution
- "$n = \text{all}$" means experimental design doesn't matter

What's new in Big Data?

- design still matters; experiments versus observational studies
- statistical reasoning still matters
- multiplicity (significance hunting) a bigger worry than ever: more opportunity to confuse correlation with causation
- confusing statistical and practical significance a bigger worry than ever
- reproducibility a bigger worry than ever
- numerical problems more difficult (or impossible) because of data volume, dimension, and velocity
- easy to confuse heroic computation with scientific truth
- algorithms that are polynomial or worse in the data become useless; need new algorithms (often based on sampling)

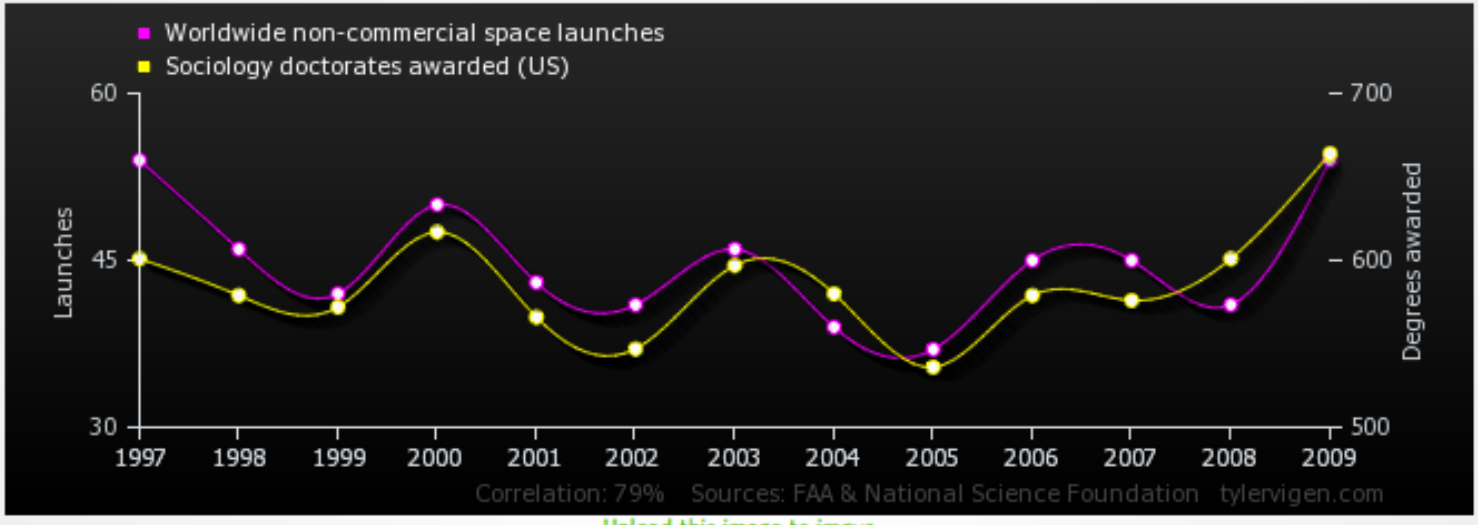
Sociology PhDs Produce Social Good

www.tylervigen.com/view_correlation.php?id=805

Worldwide non-commercial space launches

correlates with

Sociology doctorates awarded (US)



Upload this image to imgur

	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Worldwide non-commercial space launches Launches (FAA)	54	46	42	50	43	41	46	39	37	45	45	41	54
Sociology doctorates awarded (US) Degrees awarded (National Science Foundation)	601	579	572	617	566	547	597	580	536	579	576	601	664

Correlation: 0.78915

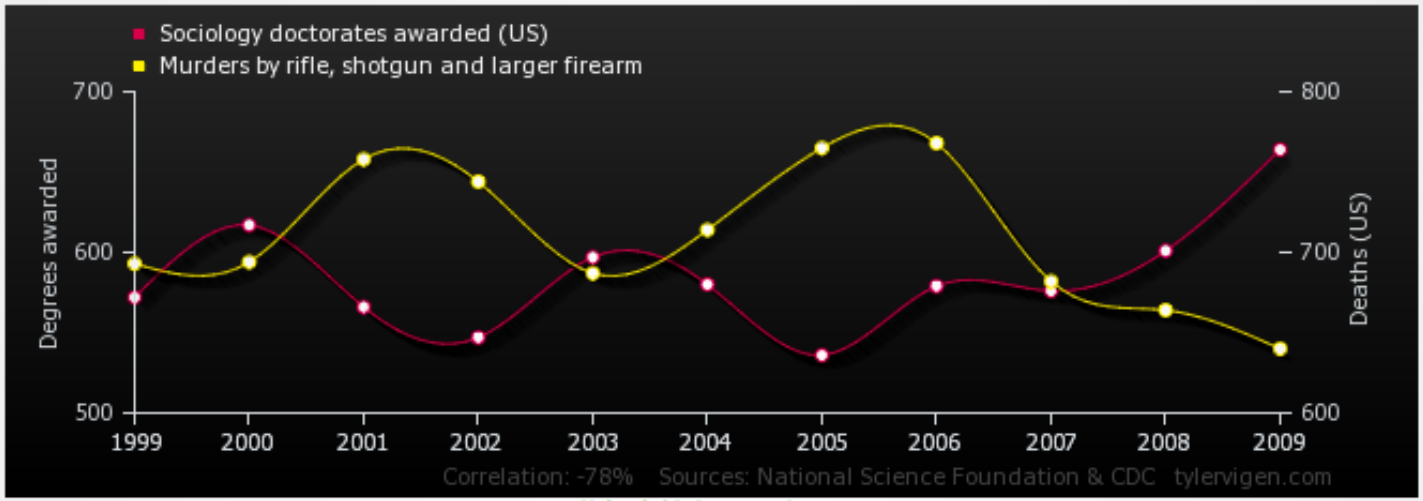
Sociology PhDs Prevent Social Harm

www.tylervigen.com/view_correlation.php?id=1960

Sociology doctorates awarded (US)

inversely correlates with

Murders by rifle, shotgun and larger firearm



Upload this image to imgur

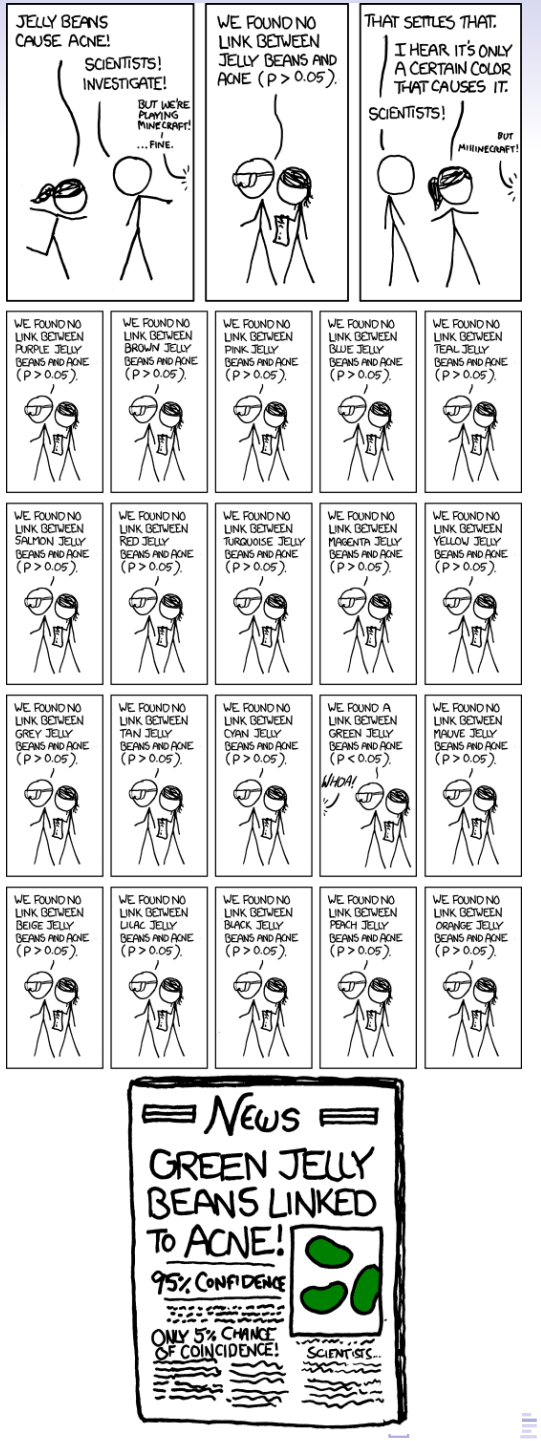
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Sociology doctorates awarded (US) *Degrees awarded (National Science Foundation)*	572	617	566	547	597	580	536	579	576	601	664
Murders by rifle, shotgun and larger firearm *Deaths (US) (CDC)*	693	694	758	744	687	714	765	768	682	664	640

Correlation: -0.784176

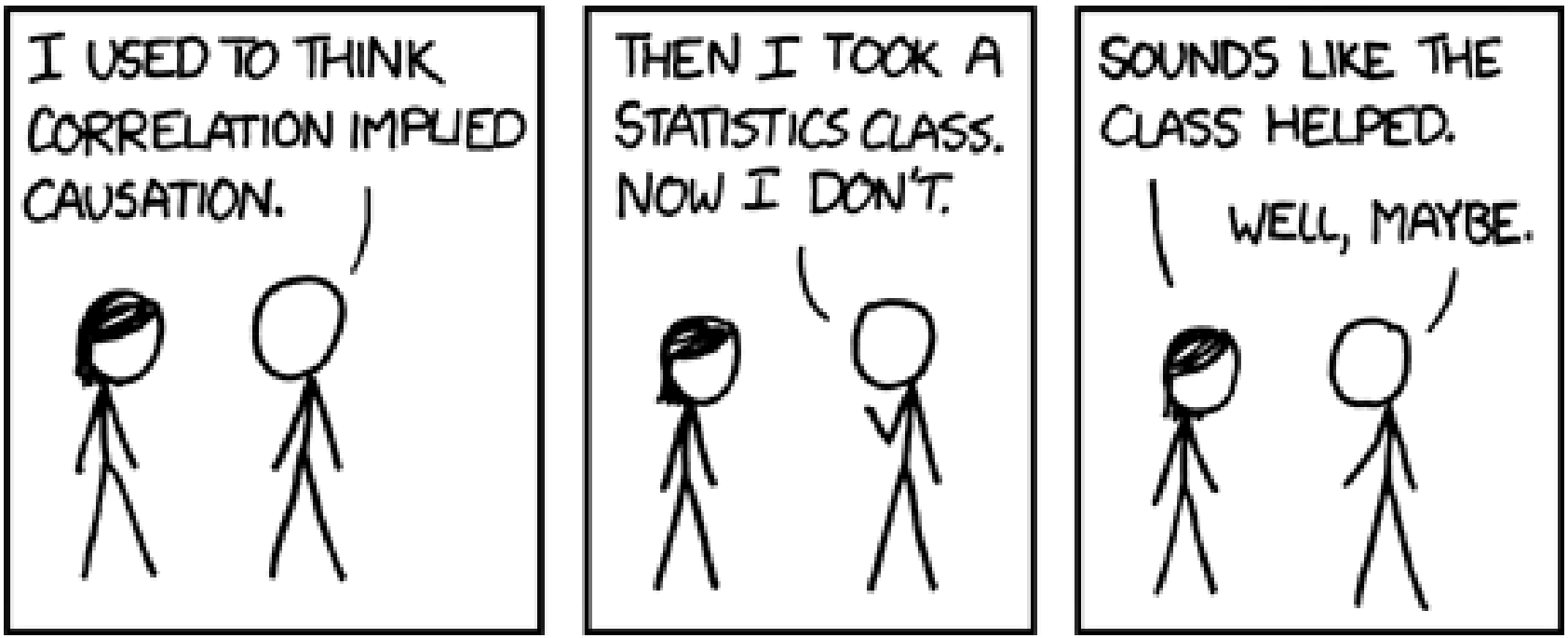
Google flu trends

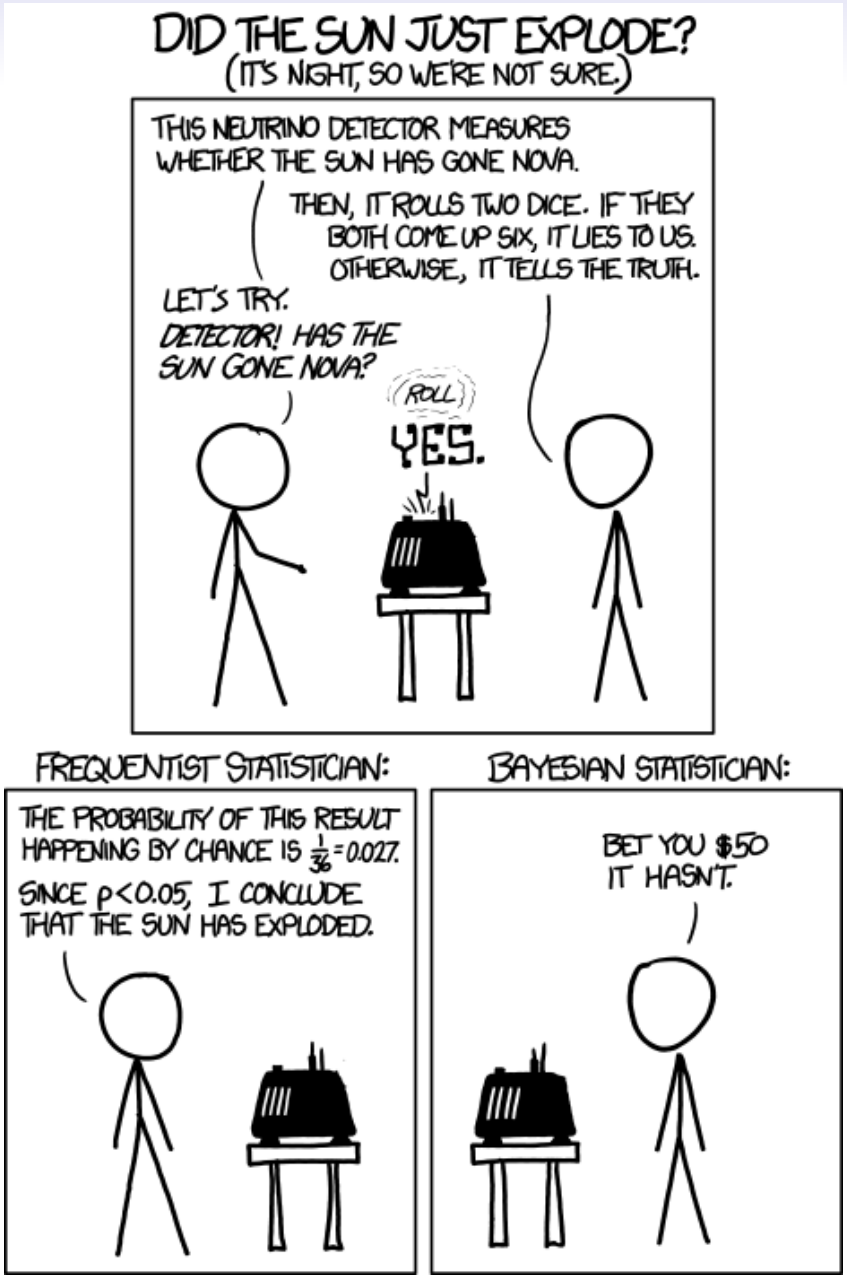
http://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/

Significance hunting:
`http://xkcd.com/882/`



Better: http://xkcd.com/552/





http://xkcd.com/1132/

Directions for research:
Post-selection inference,
conditional confidence intervals

Uncertainty Quantification Strategic Initiative–LLNL

- Uncertainty Quantification Strategic Initiative at LLNL: 1154 climate simulations using the Community Atmosphere Model (CAM).
- $p = 21$ parameters scaled so that $[0, 1]$ has all plausible values.
- f is global average upwelling longwave flux (FLUT) approximately 50 years in the future.
- Each run took several days on a supercomputer.
- Several approaches to choose $X \subset [0, 1]^p$: Latin hypercube, one-at-a-time, and random-walk multiple-one-at-a-time.
- 1154 simulations total.

CAM calculations

Empirical lower bound on Lipschitz constant: $\hat{K} = 14.20$.

$$M > \epsilon^{-21} \times 10^{26}$$

If ϵ is 1% of $\hat{K}$, then $M \geq 10^{43}$.

Even if ϵ is 50% of $\hat{K}$, $M > 10^{8}$.

Parker's Rule of Epistemology

The more you assume, the less you know.

Philip's Rule of Uncertainty Quantification

To quantify the uncertainty of your fancy model takes at least three orders of magnitude more computational power than you have.