14.31x
Data Analysis for Social Scientists

Instructors:
Esther Duflo and Sara Ellison

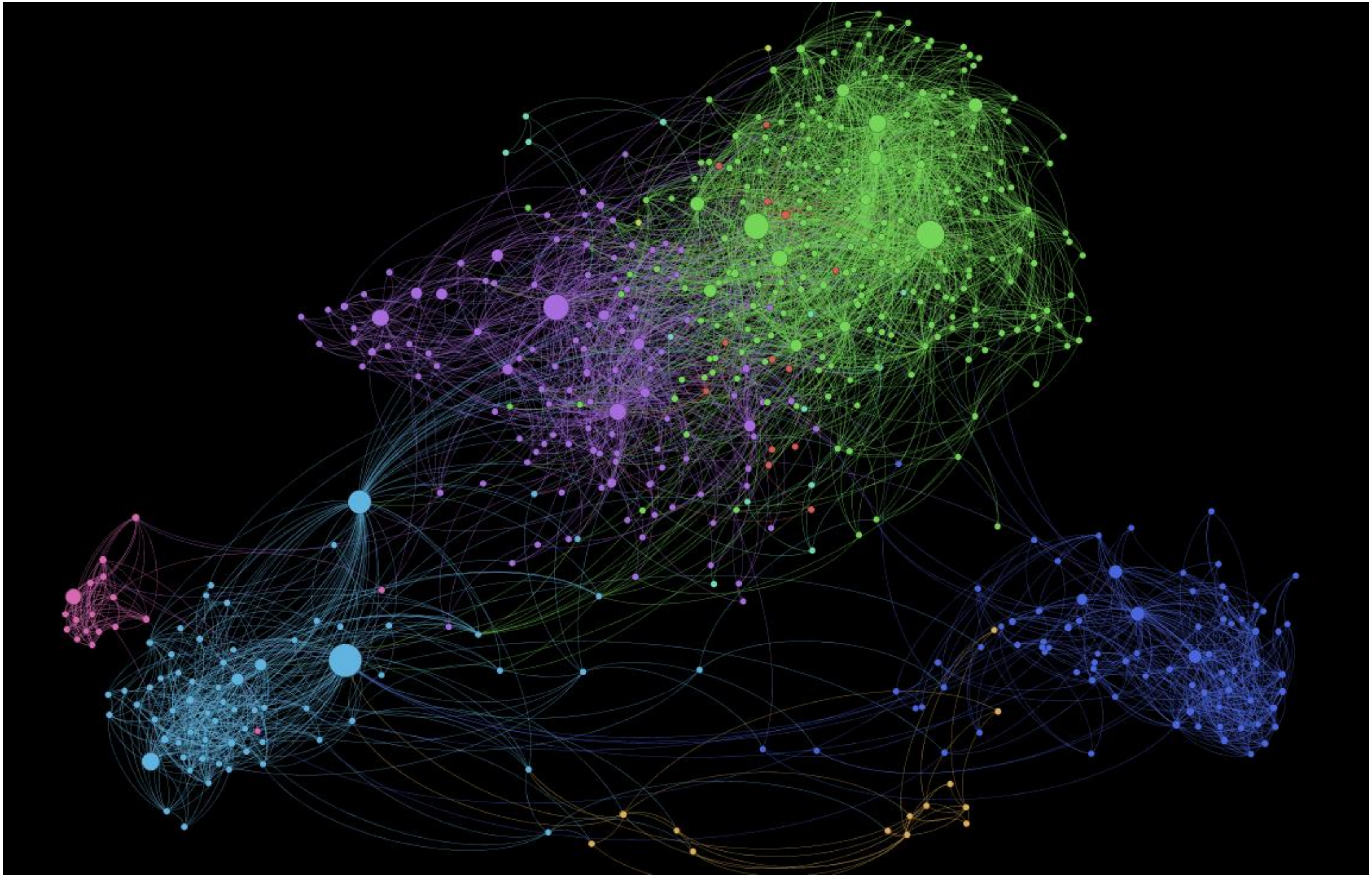
Massachusetts Institute of Technology

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

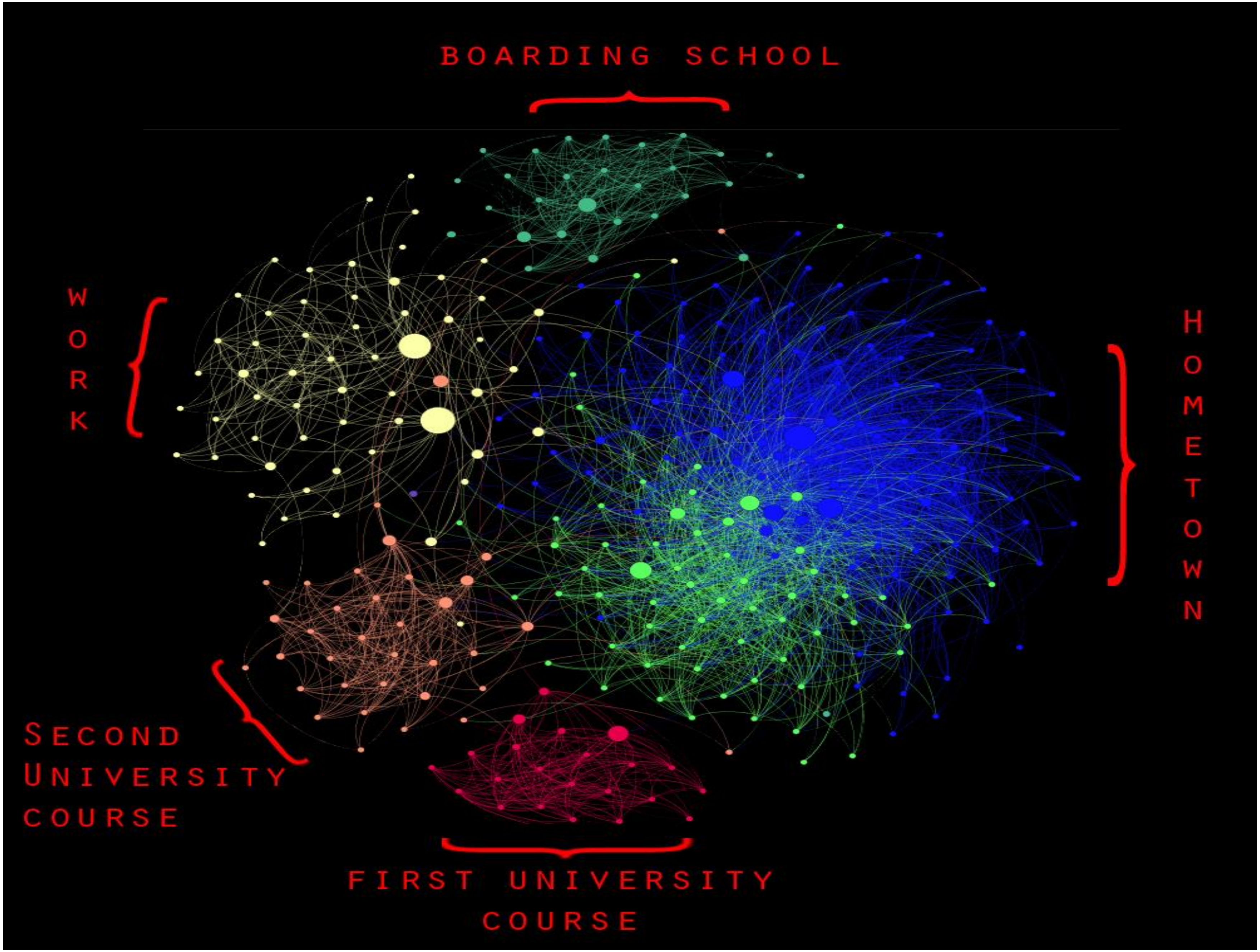
Data is Plentiful

Data is Beautiful

- Example: Mapping Facebook networks of individuals from Somalia living in Eastleigh



Courtesy of Kimo Quaintance. Used with permission.



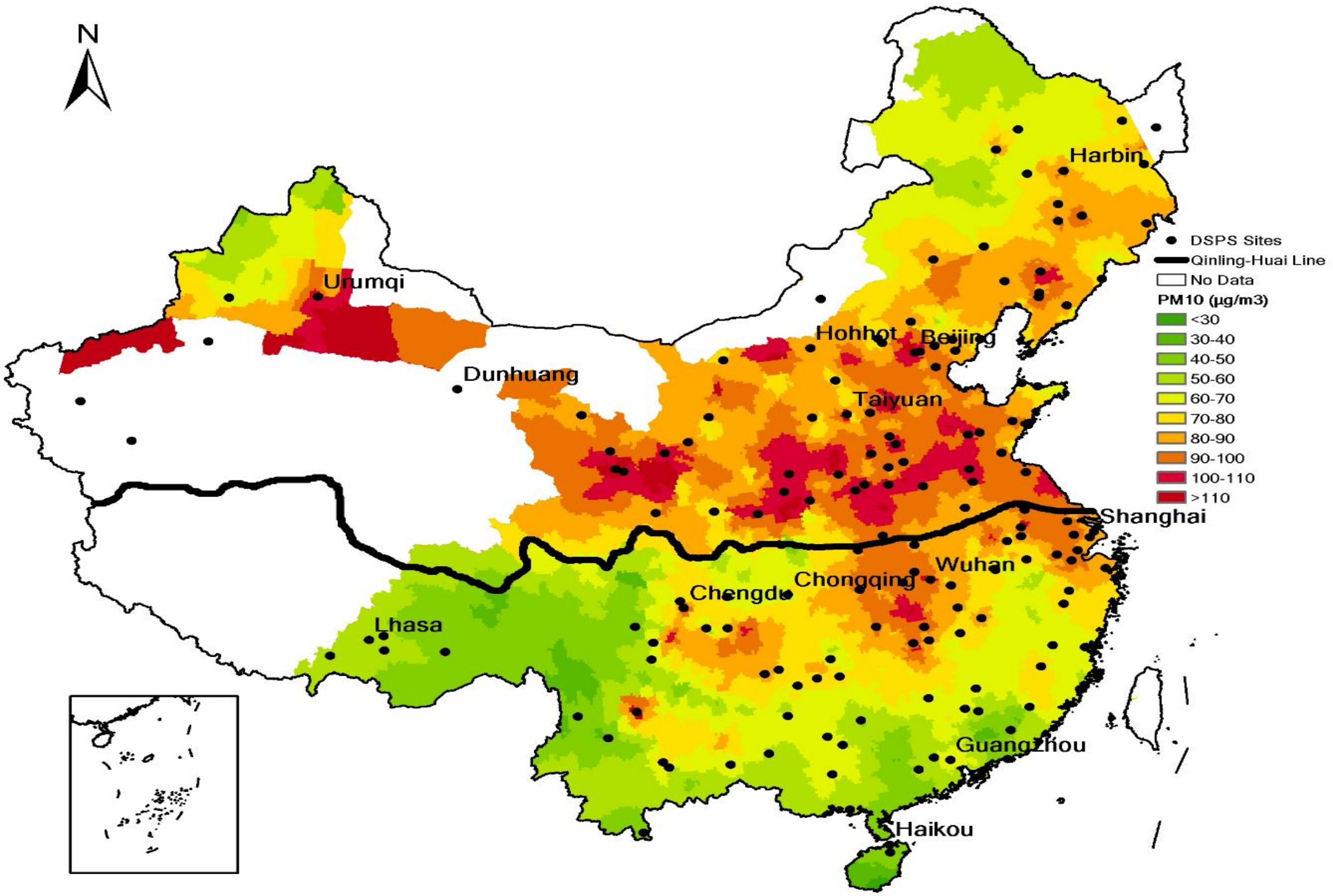
Courtesy of Kimo Quaintance. Used with permission.

Data is Insightful

- Example: Pollution in China

Figure 1

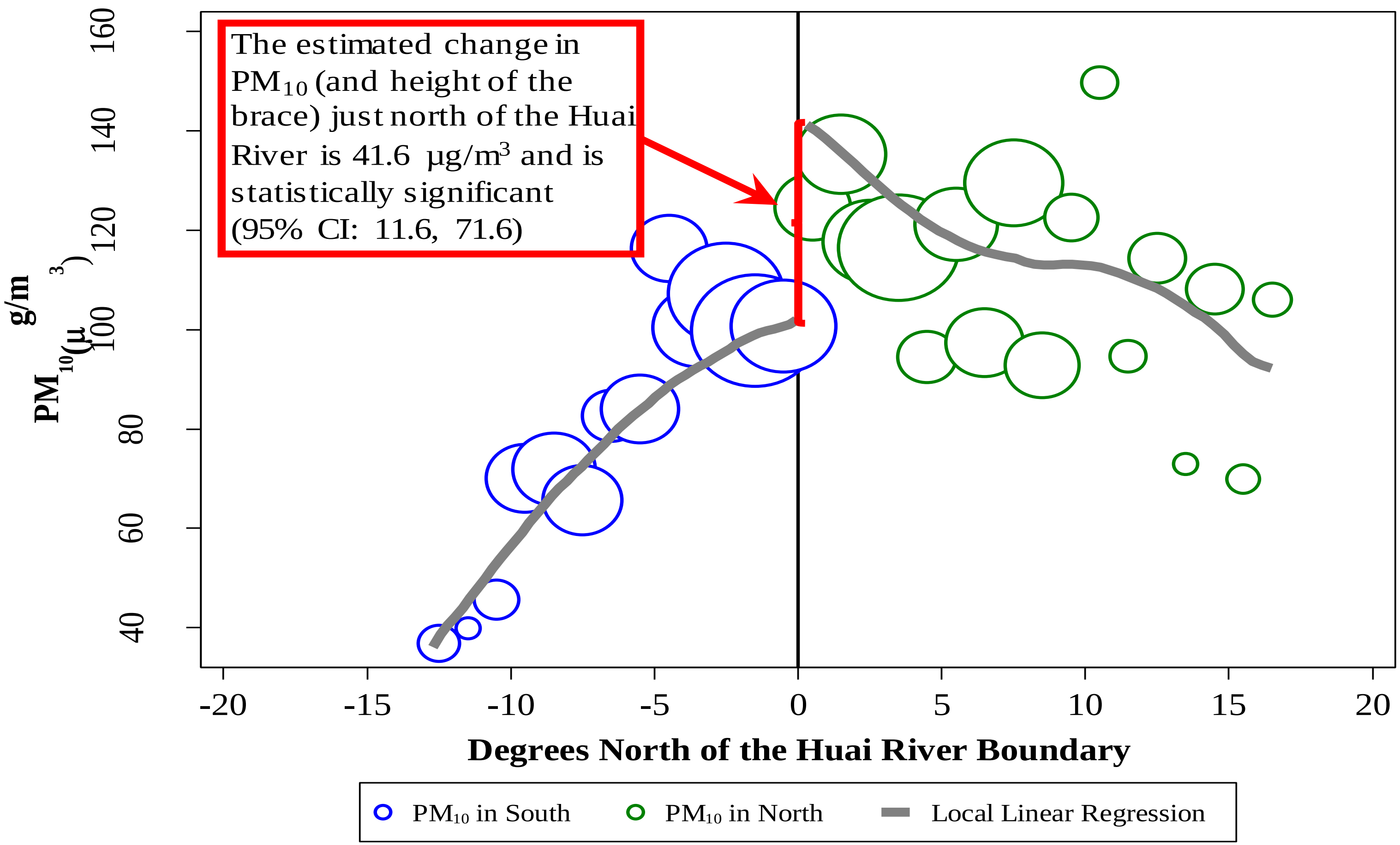
Pollution in China and the Huai River/Qinling Mountain Range



Notes : The cities shown are the locations of the Disease Surveillance Points. Cities north of the solid line were covered by the home heating policy. The figure coloring is generated by interpolating PM_{10} levels at the 12 nearest pollution monitoring stations to create a high resolution grid of pollution throughout China (.1 degree latitude cell width). Areas are left in white which are not within acceptable range of a station.

Figure 2

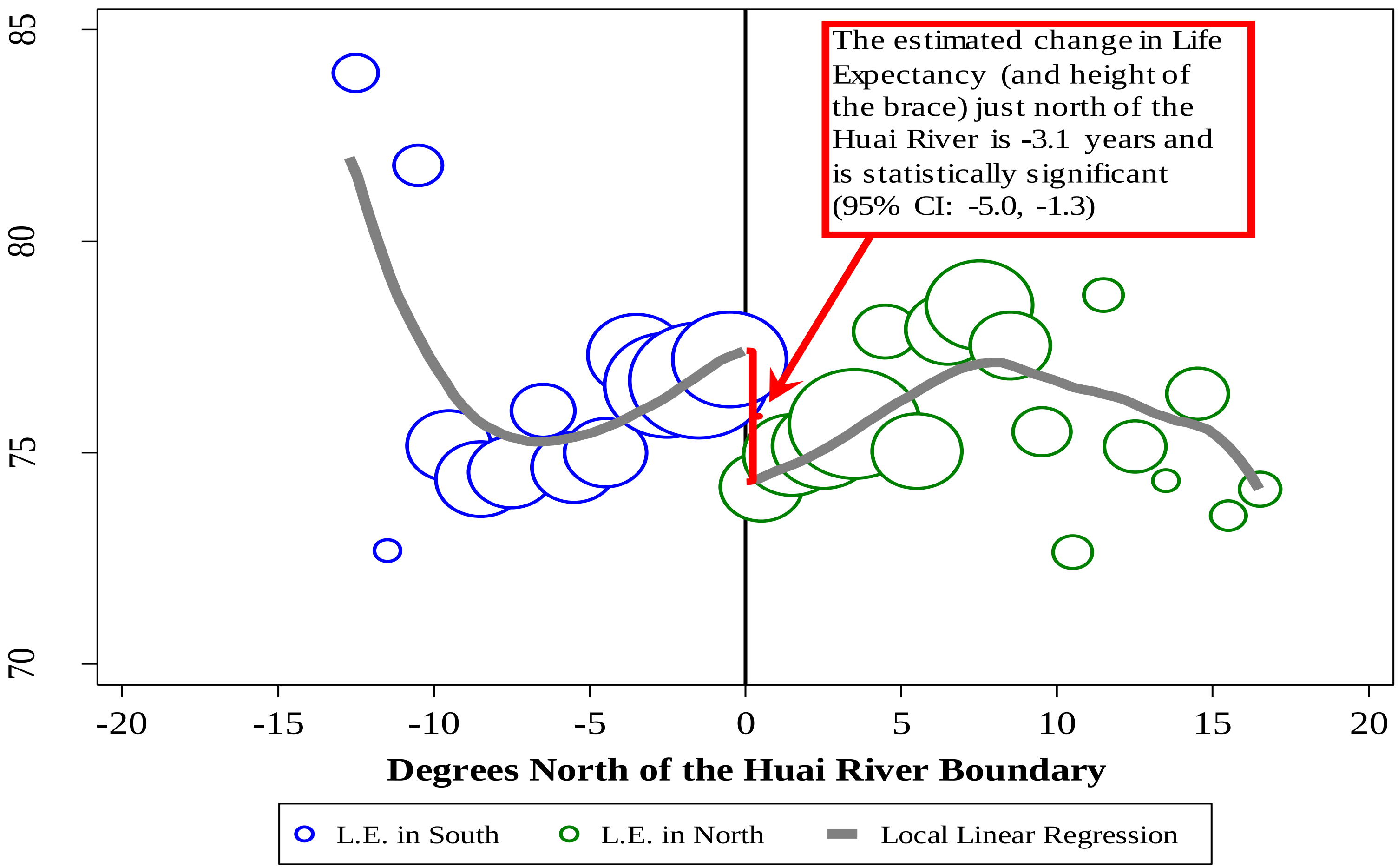
Particulate Matter Levels (PM_{10}) South and North of the Huai River Boundary



Surveillance Point locations within a 1 degree latitude range, weighted by the population at each location. The size of the circle is in proportion to the total population at DSP locations within the 1 degree latitude range. The plotted line reports a local linear regression plot estimated separately on on each side of the Huai River.

Figure 3

Life Expectancy South and North of the Huai River Boundary



Notes : Each observation (circle) is generated by averaging life expectancy across the Disease Surveillance Point locations within a 1 degree latitude range, weighted by the population at each location. The size of the circle is in proportion to the total population at DSP locations within the 1 degree latitude range. The plotted line reports a local linear regression plot estimated separately on on each side of the Huai River.

Data is Powerful

- Example: Changing regulation in India

Figure 2: Audit and Backcheck Readings for Suspended Particulate Matter (SPM, mg/Nm^3), Midline

A. Control Plants

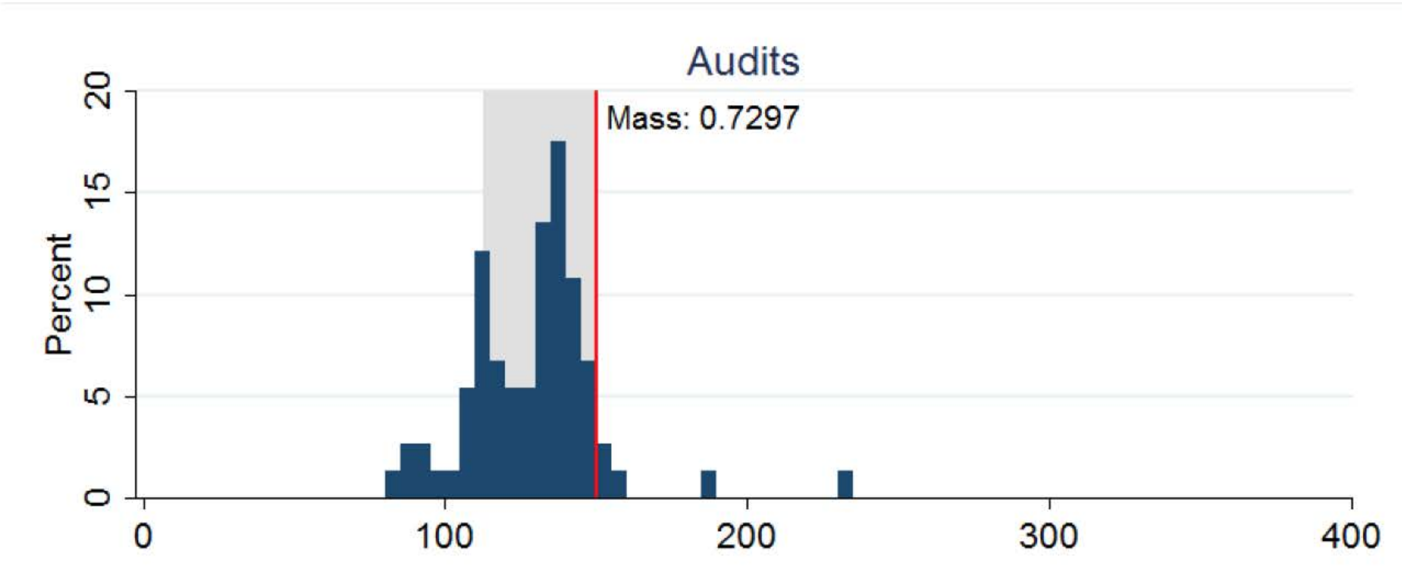


Figure 2: Audit and Backcheck Readings for Suspended Particulate Matter (SPM, mg/Nm^3), Midline

A. Control Plants

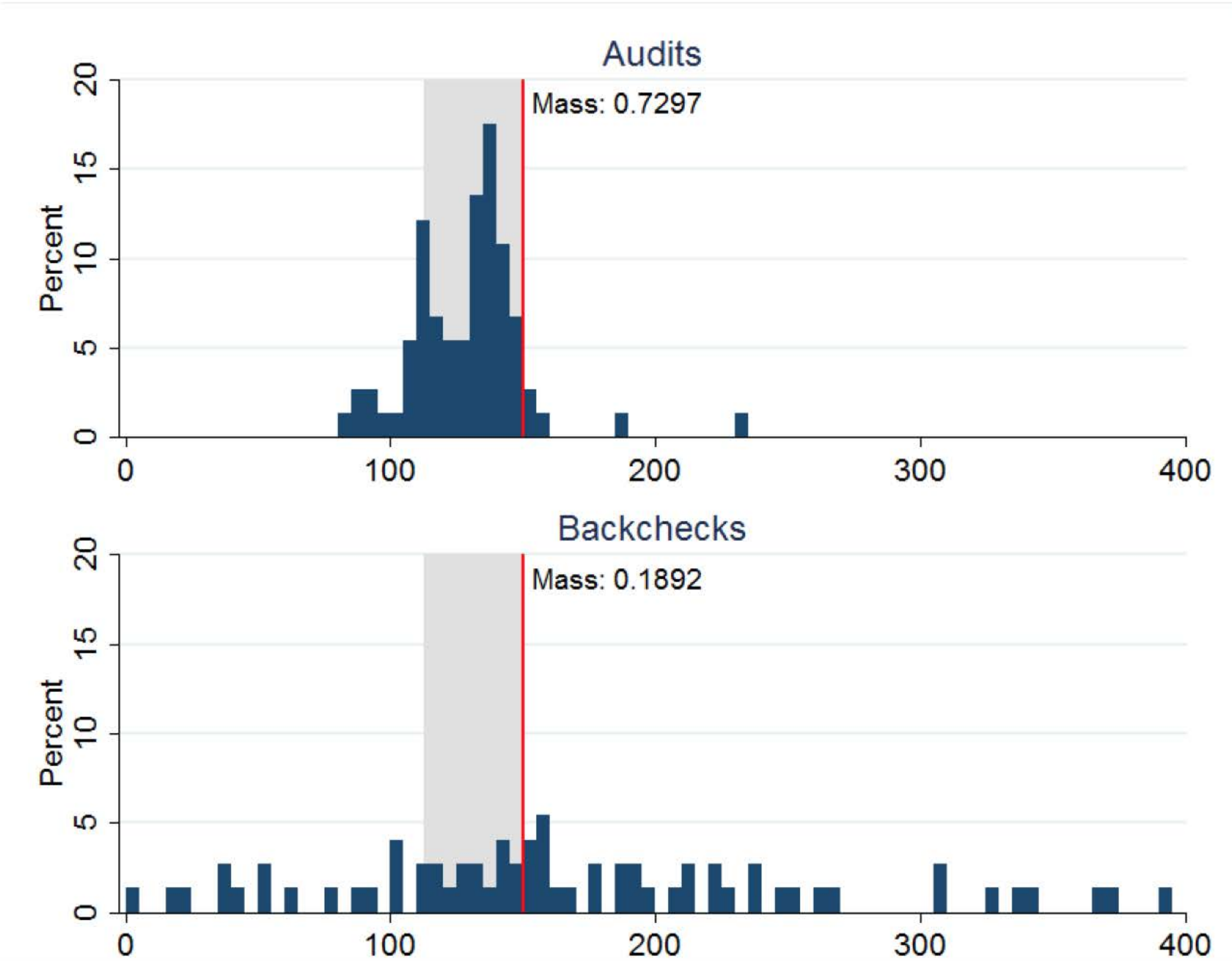
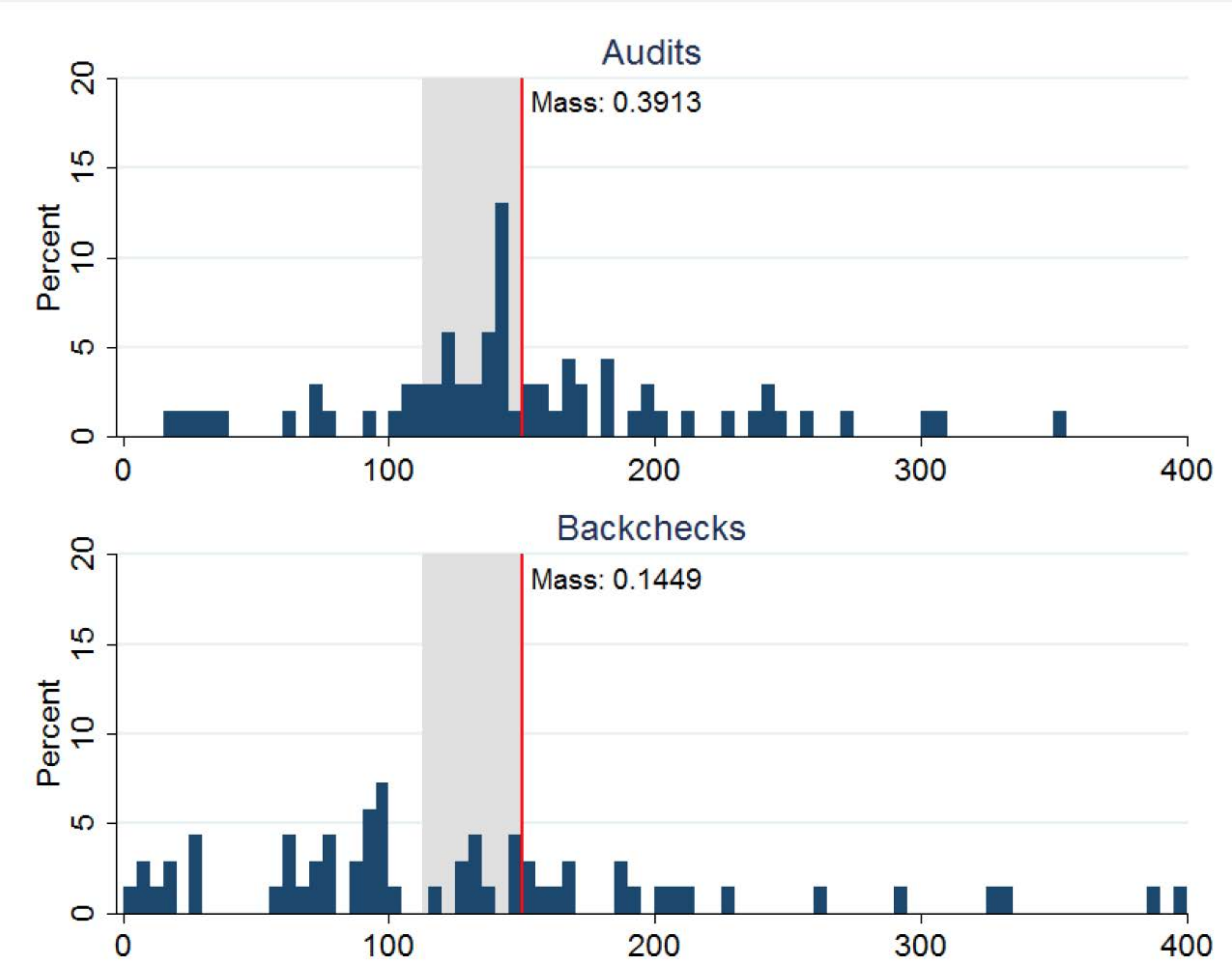


Figure 2: Audit and Backcheck Readings for Suspended Particulate Matter (SPM, mg/Nm^3), Midline

B. Treatment Plants

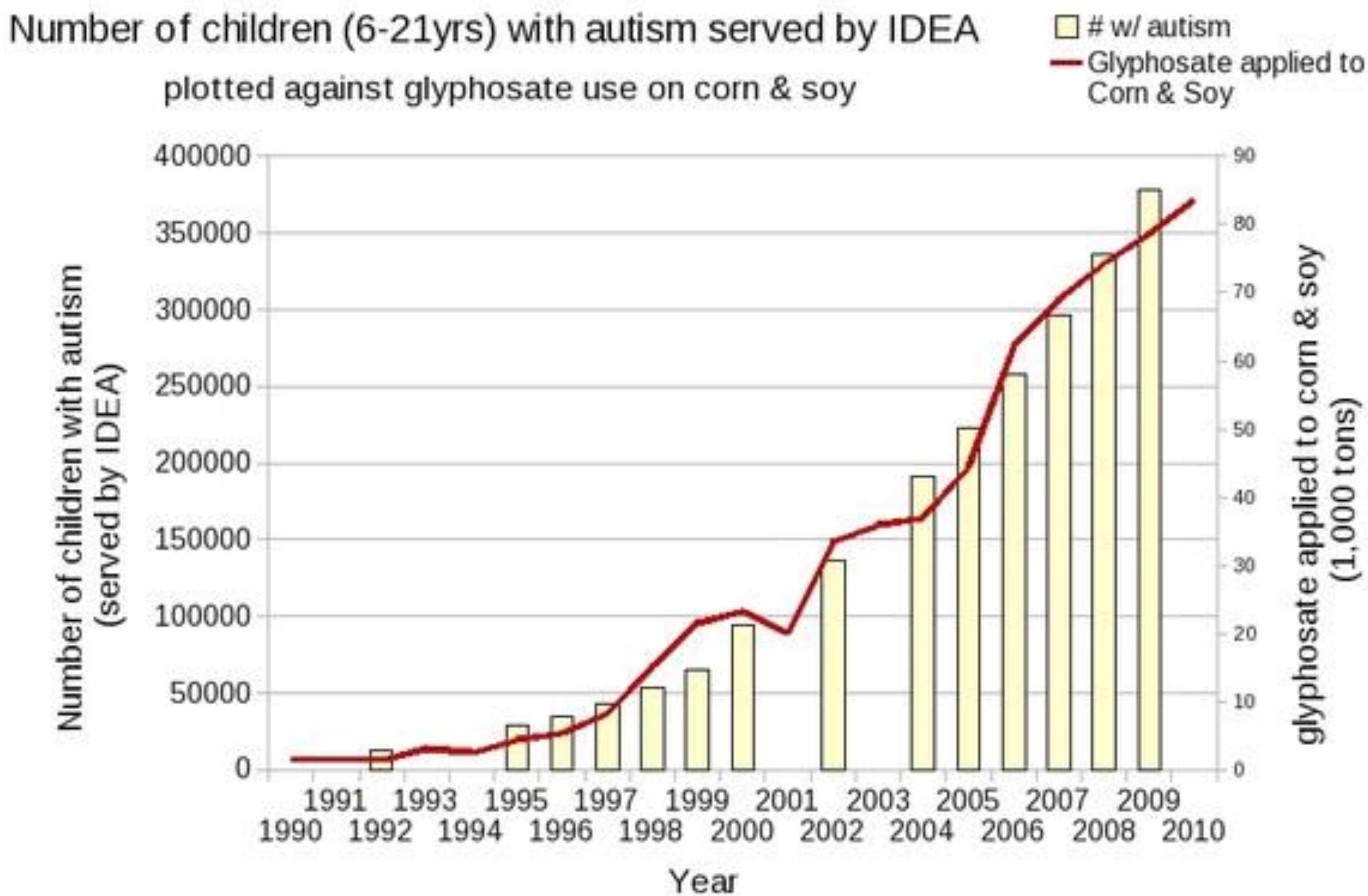


Lessons

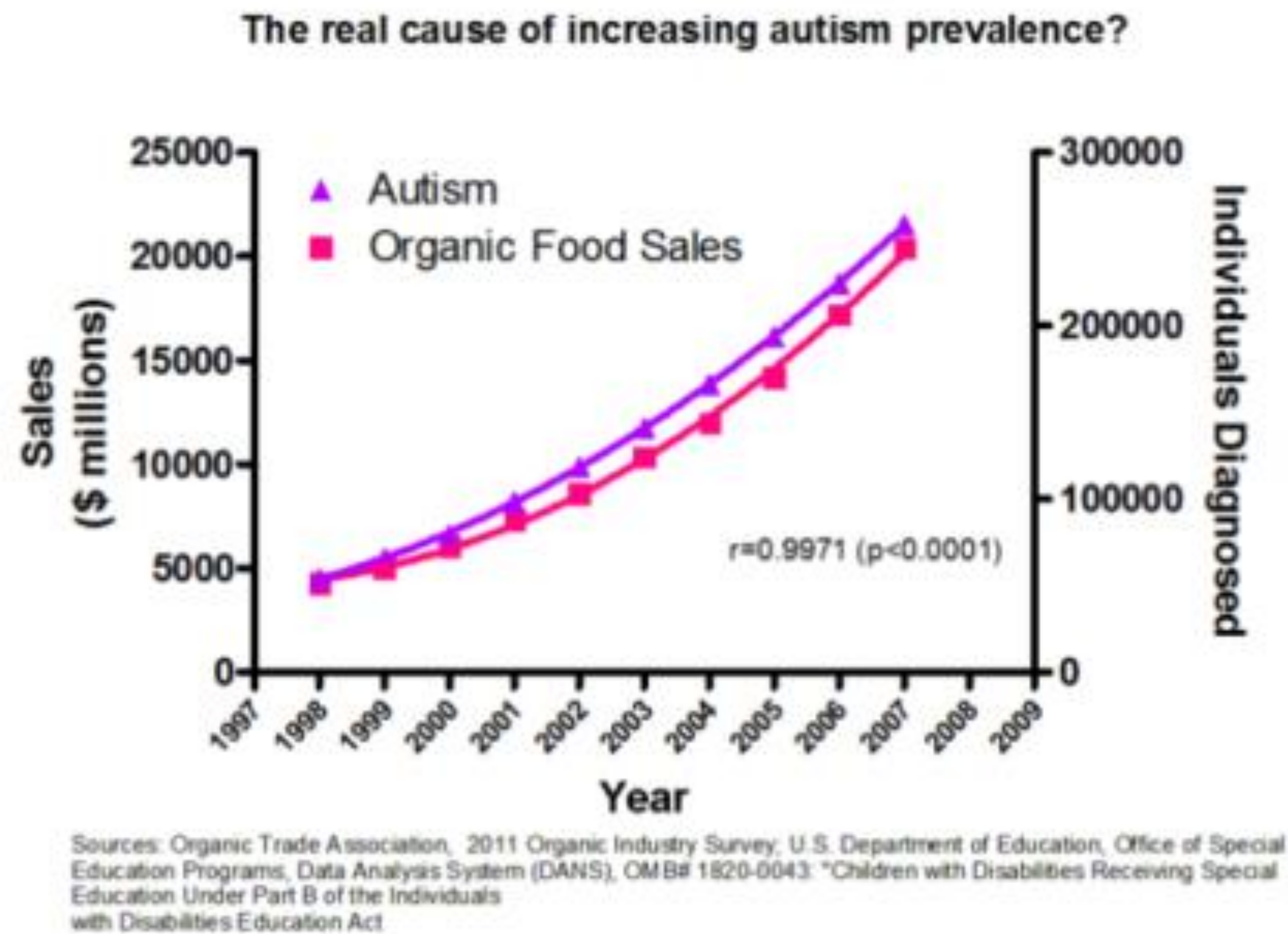
- Conflict of interest leads auditor to cheat on the data they report to the government
- An experiment that changes the reporting structure to eliminate the conflict of interest largely solves the problem.
- This demonstration leads the government of Gujarat to change their policy!
- To date 207 million people have been touched by programs that J-PAL has shown to be effective based on RCT

Data can be Deceitful

- Example: Correlations with autism

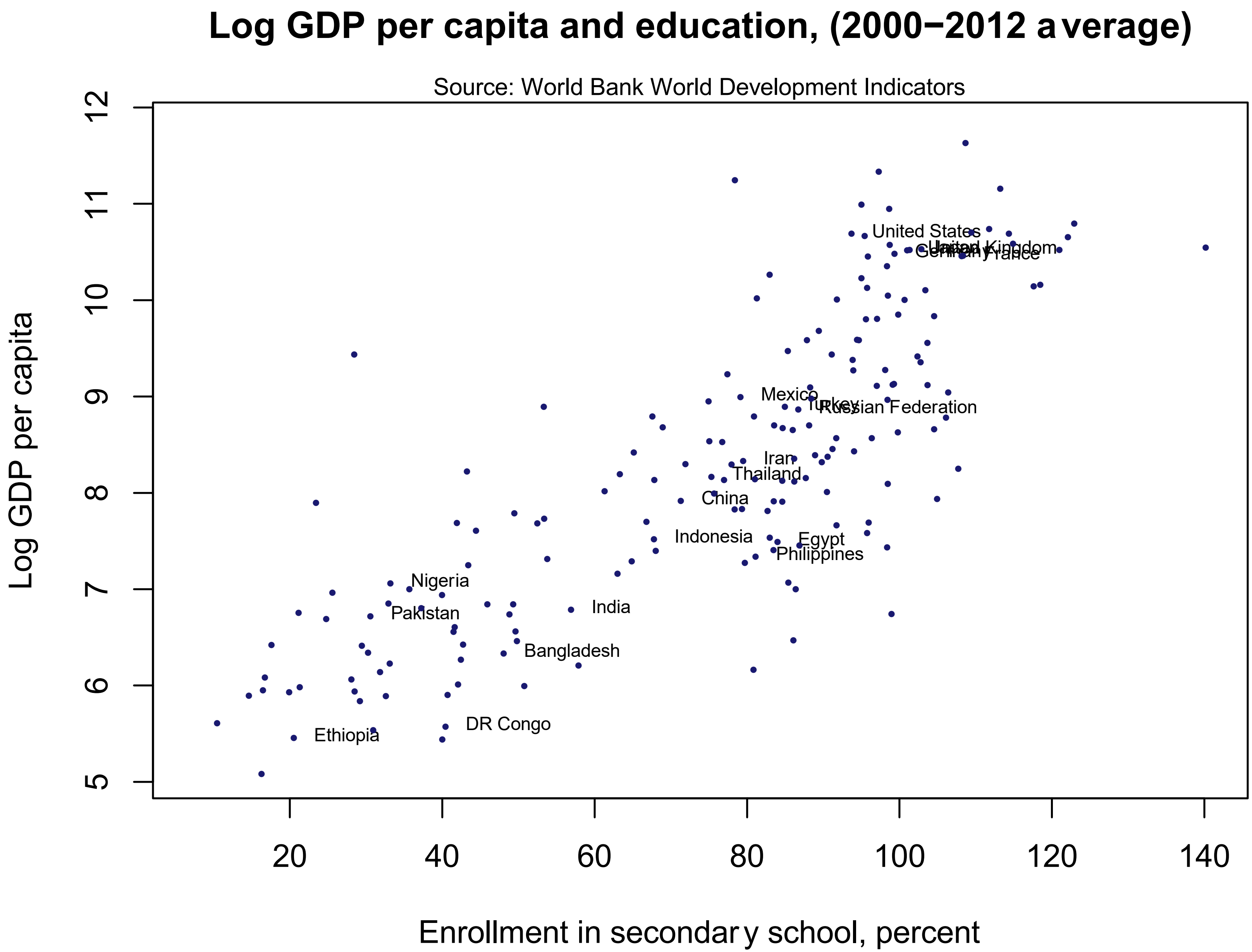


Number of children (6-21yrs) with autism served by IDEA
plotted against glyphosate use on corn & soy
w/ autism
Glyphosate applied to Corn & Soy
Number of children with autism (served by IDEA)
glyphosate applied to corn & soy (1,000 tons)
0
50000
100000
150000
200000
250000
300000
350000
400000
0
10
20
30
40
50
60
70
80
90
1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010
Year


The real cause of increasing autism prevalence?
Autism
Organic Food Sales
Sales ($ millions)
Individuals Diagnosed
25000
20000
15000
10000
5000
0
300000
200000
100000
0
r=0.9971 (p<0.0001)
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Year
Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act

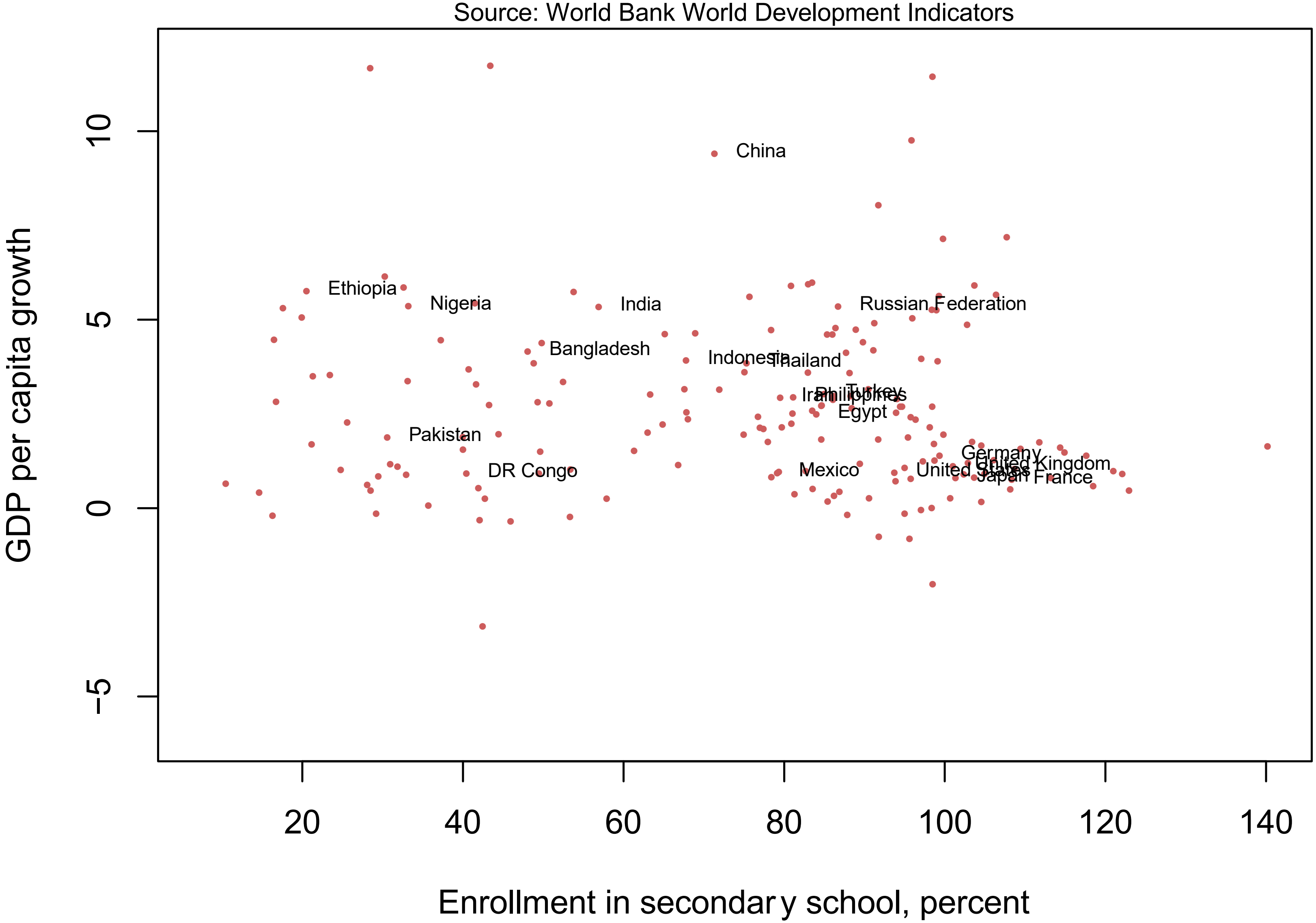
Data can be Deceitful

- That one is trivial but… how about some less obvious ones?



Log GDP per capita and education, (2000−2012 average)
Source: World Bank World Development Indicators
Log GDP per capita
Enrollment in secondary school, percent
5
6
7
8
9
10
11
12
20
40
60
80
100
120
140
United States
United Kingdom
Germany
France
Mexico
Turkey
Russian Federation
Iran
Thailand
China
Indonesia
Egypt
Philippines
Nigeria
Pakistan
India
Bangladesh
Ethiopia
DR Congo

GDP per capita growth and education, (2000−2012 average)



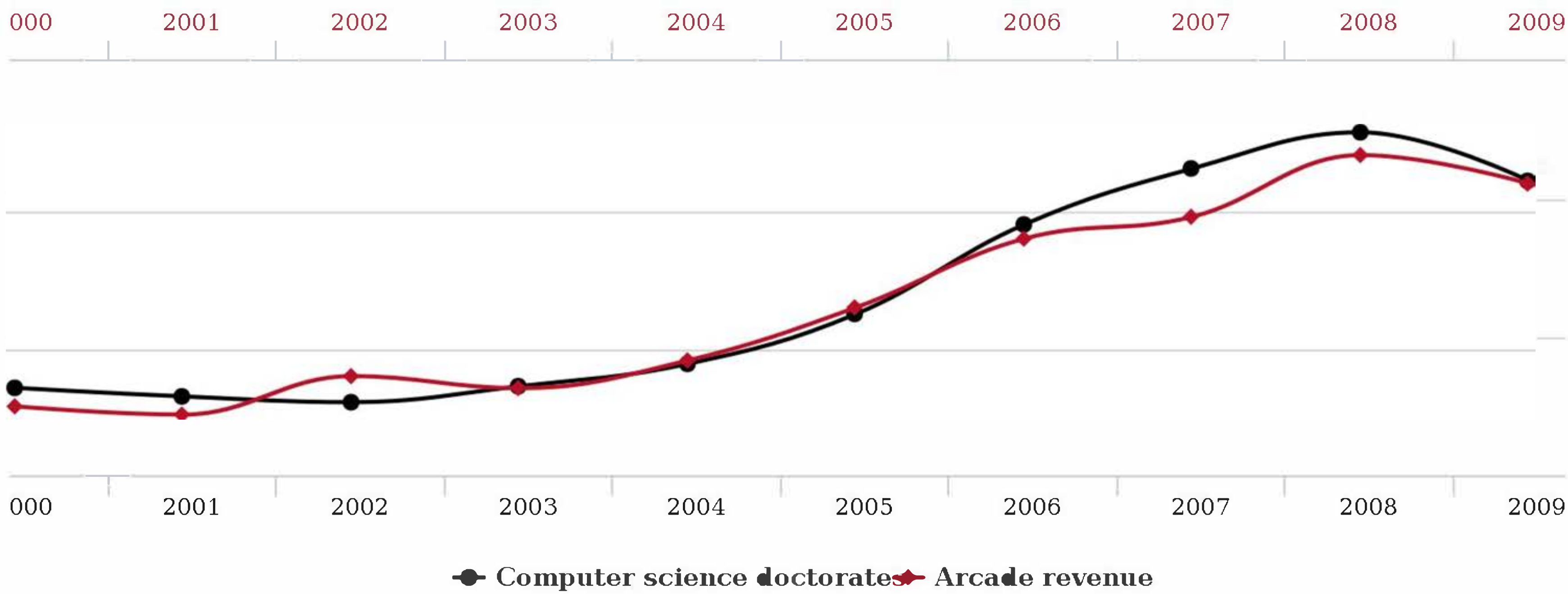
Causation versus Correlation

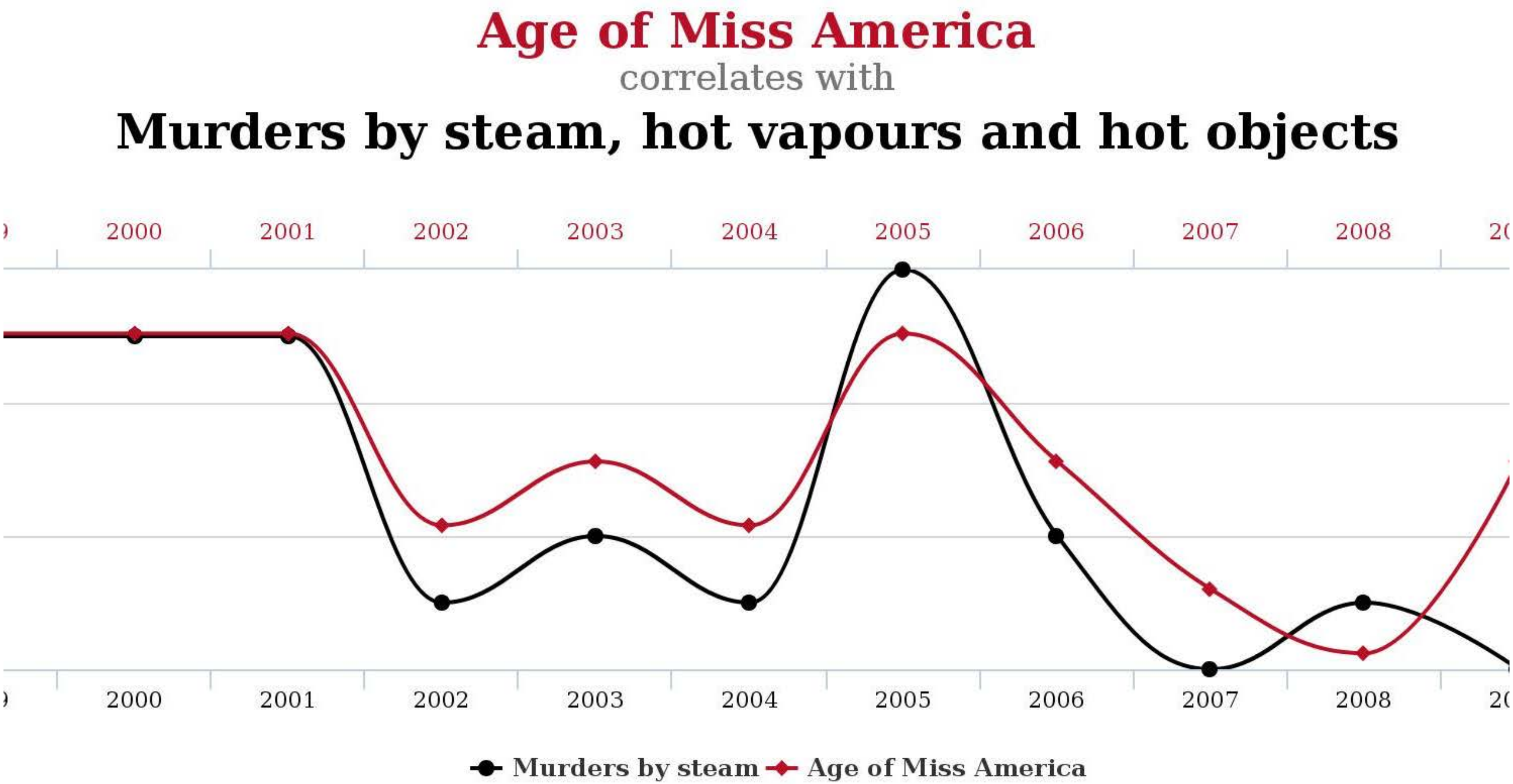
- Correlation is not causality
- A causal *story* is not causality either…
- Even more sophisticated data use may still not be causality.

Causation versus Correlation

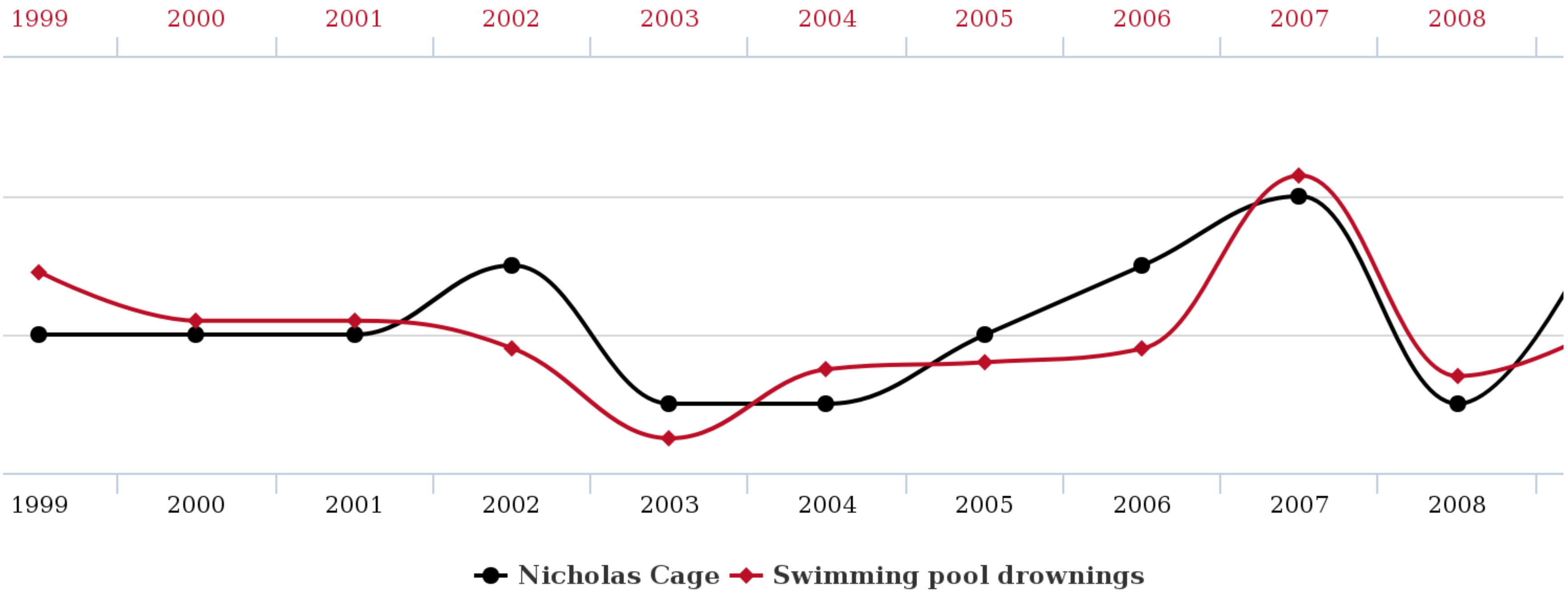
- Data by the chokefull
  - There is so much data available that it is possible to infer from the data very powerful predictive patterns:
    - What do people who live in Boston, search for capoeira classes video and websites for children before going on the spurious statistics web site to download a couple of graphs, and buy PlanToys doll house may want to buy next?
    - Are people with a specific gene more likely to be patient?
  - But you want to be careful of patterns you observe in the data… they are not always meaningful.

Total revenue generated by arcades
correlates with
Computer science doctorates awarded in the US




Age of Miss America
correlates with
Murders by steam, hot vapours and hot objects
2000
2001
2002
2003
2004
2005
2006
2007
2008
Murders by steam
Age of Miss America

Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in



What we need to learn

- How do we model the processes that might have generated our data?
  - Probability
- How do we summarize and describe data, and try to uncover what process may have generated it?
  - Statistics
- How do we uncover pattern between variables?
  - Exploratory data analysis
  - Econometrics
  - Machine Learning

What we need to learn

- How do we think of causality?
  - A causal framework
  - RCTs, AB/testing, etc.
  - Regressions
- How do we do all this in practice?
  - R
  - Experiment design
  - Where to get data?
- How do we present our results in a compelling (and truthful!) way?
  - Beautiful graphs: GIS, networks, etc.
  - Insightful tables
  - Enlightening text!

5. Course Outline: The number of lectures devoted to each topic is approximate and subject to change.

Introduction and Motivation	1 lecture
Probability	8 lectures
Definitions	
Random variables	
Distributions of RVs	
Functions of RVs	
Expectation, variance	
Basic estimation and inference	3 lectures
Definitions	
Estimators	
CLT	
Confidence intervals	
Hypothesis testing	
Randomized controlled trials	2 lectures
Nonparametric estimation	1 lecture
Causality	1 lecture
Regression analysis	4 lectures
Design of experiment	2 lectures
Machine learning	2 lectures
Assorted topics, such as visual display	1 lecture

5. Course Outline: The number of lectures devoted to each topic is approximate and subject to change.

Introduction and Motivation	1 lecture
Probability	8 lectures
Definitions	
Random variables	
Distributions of RVs	
Functions of RVs	
Expectation, variance	
Basic estimation and inference	3 lectures
Definitions	
Estimators	
CLT	
Confidence intervals	
Hypothesis testing	
Randomized controlled trials	2 lectures
Nonparametric estimation	1 lecture
Causality	1 lecture
Regression analysis	4 lectures
Design of experiment	2 lectures
Machine learning	2 lectures
Assorted topics, such as visual display	1 lecture

Spend a chunk of time on probability---this provides necessary foundation for all of the data analysis we will do later on

5. Course Outline: The number of lectures devoted to each topic is approximate and subject to change.

Introduction and Motivation	1 lecture
Probability	8 lectures
Definitions	
Random variables	
Distributions of RVs	
Functions of RVs	
Expectation, variance	
Basic estimation and inference	3 lectures
Definitions	
Estimators	
CLT	
Confidence intervals	
Hypothesis testing	
Randomized controlled trials	2 lectures
Nonparametric estimation	1 lecture
Causality	1 lecture
Regression analysis	4 lectures
Design of experiment	2 lectures
Machine learning	2 lectures
Assorted topics, such as visual display	1 lecture

To give you some idea of topics---will not stick to this order or allocation

5. Course Outline: The number of lectures devoted to each topic is approximate and subject to change.

Introduction and Motivation	1 lecture
Probability	8 lectures
Definitions	
Random variables	
Distributions of RVs	
Functions of RVs	
Expectation, variance	
Basic estimation and inference	3 lectures
Definitions	
Estimators	
CLT	
Confidence intervals	
Hypothesis testing	
Randomized controlled trials	2 lectures
Nonparametric estimation	1 lecture
Causality	1 lecture
Regression analysis	4 lectures
Design of experiment	2 lectures
Machine learning	2 lectures
Assorted topics, such as visual display	1 lecture

Throughout semester, we will be mixing in instruction on R, information about data sources, empirical techniques, such as web-scraping, online surveys, etc.

Sources

- Chen, Yuyu, Avraham Ebenstein, Michael Greenstone, and Hongbin Li. "Evidence on the Impact of Sustained Exposure to Air Pollution on Life Expectancy from China's Huai River Policy. MIT Working Paper No 13-15. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2291154&download=yes.
- Duflo, Esther, Michael Greenstone, Rohini Pande, Nicholas Ryan (2013). "Truth-telling by third-party auditors and the response of polluting firms: Experimental evidence from India. NBER working paper 19259. http://economics.mit.edu/files/10713.
- Gorski, David (aka ORAC). "Oh, no! GMOs are going to make everyone autistic!" Dec 31, 2014. http://scienceblogs.com/insolence/2014/12/31/oh-no-gmos-are-going-to-make-everyone-autistic/.
- Quantaince, Kimo. "What can we learn about Somalis from their social networks?" http://kimoquaintance.com/2011/08/22/what-can-we-learn-about-somalis-from-their-facebook-networks/
- Seneff, Stephanie. Most Popular Herbicide Glyphosate Causes Autism. April 28, 2014. https://people.csail.mit.edu/seneff/California_glyphosate.pdf.
- Vigen, Tyler. "Spurious Correlations" http://www.tylervigen.com/spurious-correlations. https://creativecommons.org/licenses/by/4.0/legalcode.
- World Bank World Development Indicators. http://data.worldbank.org/data-catalog/world-development-indicators.