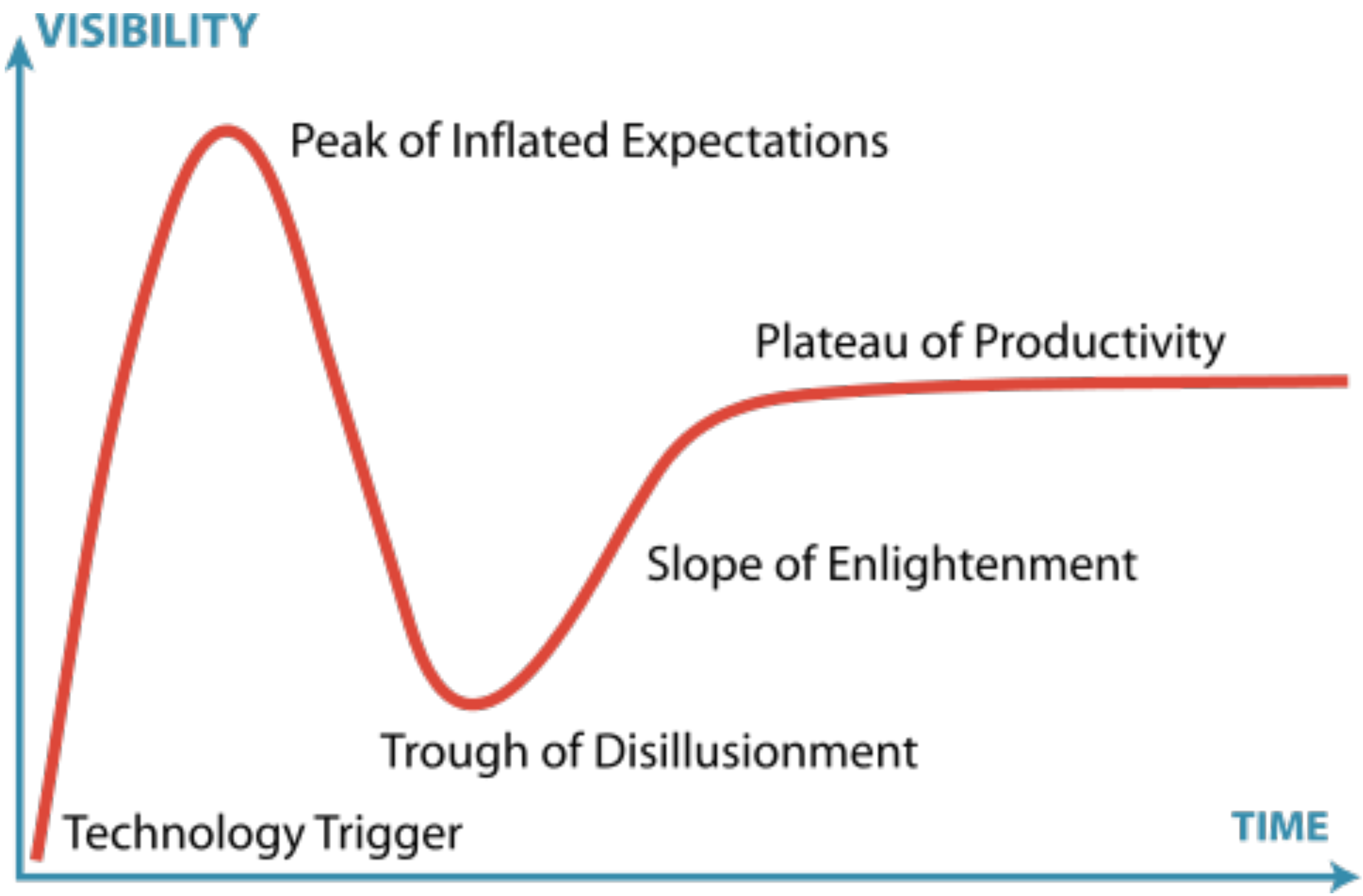
CS7280
TOPICS IN STATISTICS AND DATA ANALYSIS

GARTNER HYPE CYCLE



www.gartner.com www.wikipedia.org

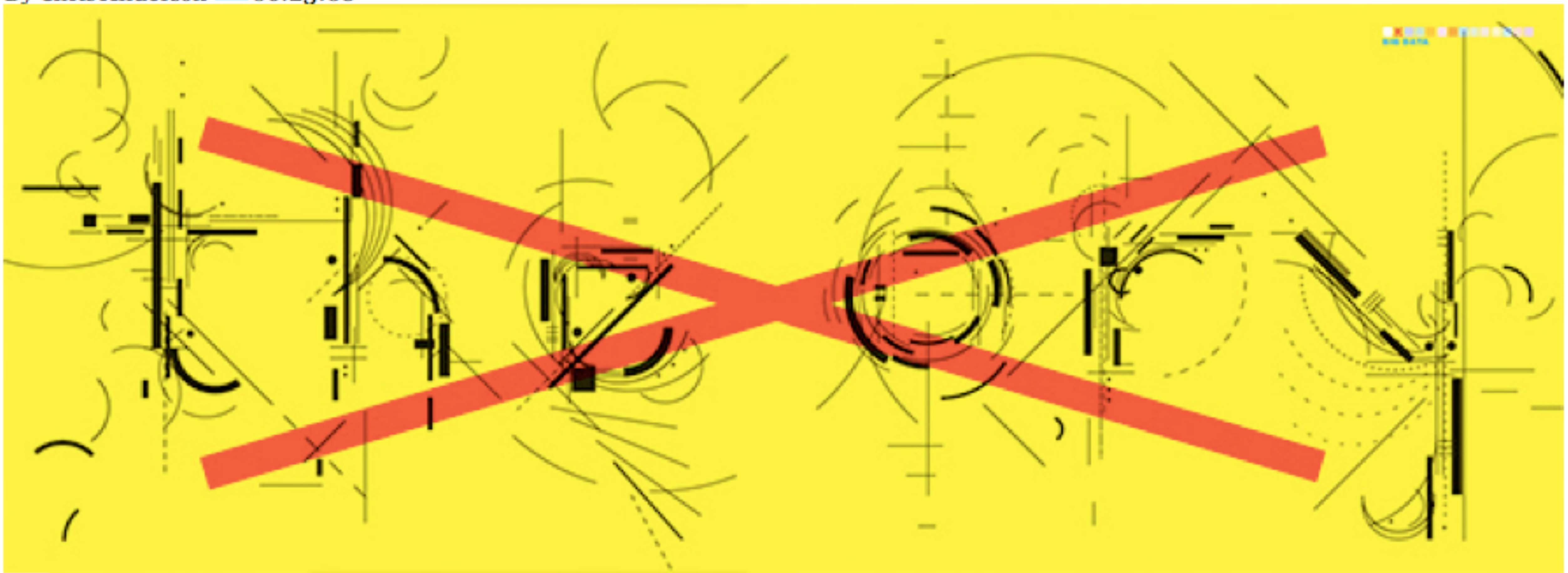
THE END OF THEORY?

WIRED MAGAZINE: 16.07

Science : Discoveries

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08



THE END OF THEORY?

- Old science: models
  - All models are wrong, but some are useful (George Box)
- New science: just data
  - Do not need to know culture and conventions
  - Do not need to know the underlying mechanisms
  - All models are wrong, and increasingly you can succeed without them
- What is the new scientific method?
  - Statistical tools will crunch the numbers and offer a new way of understanding the world
  - ‘There's no reason to cling to our old ways. It's time to ask: What can science learn from Google?’

The Parable of Google Flu: Traps in Big Data Analysis

Large er
avoidabl
of big da

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,3]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can

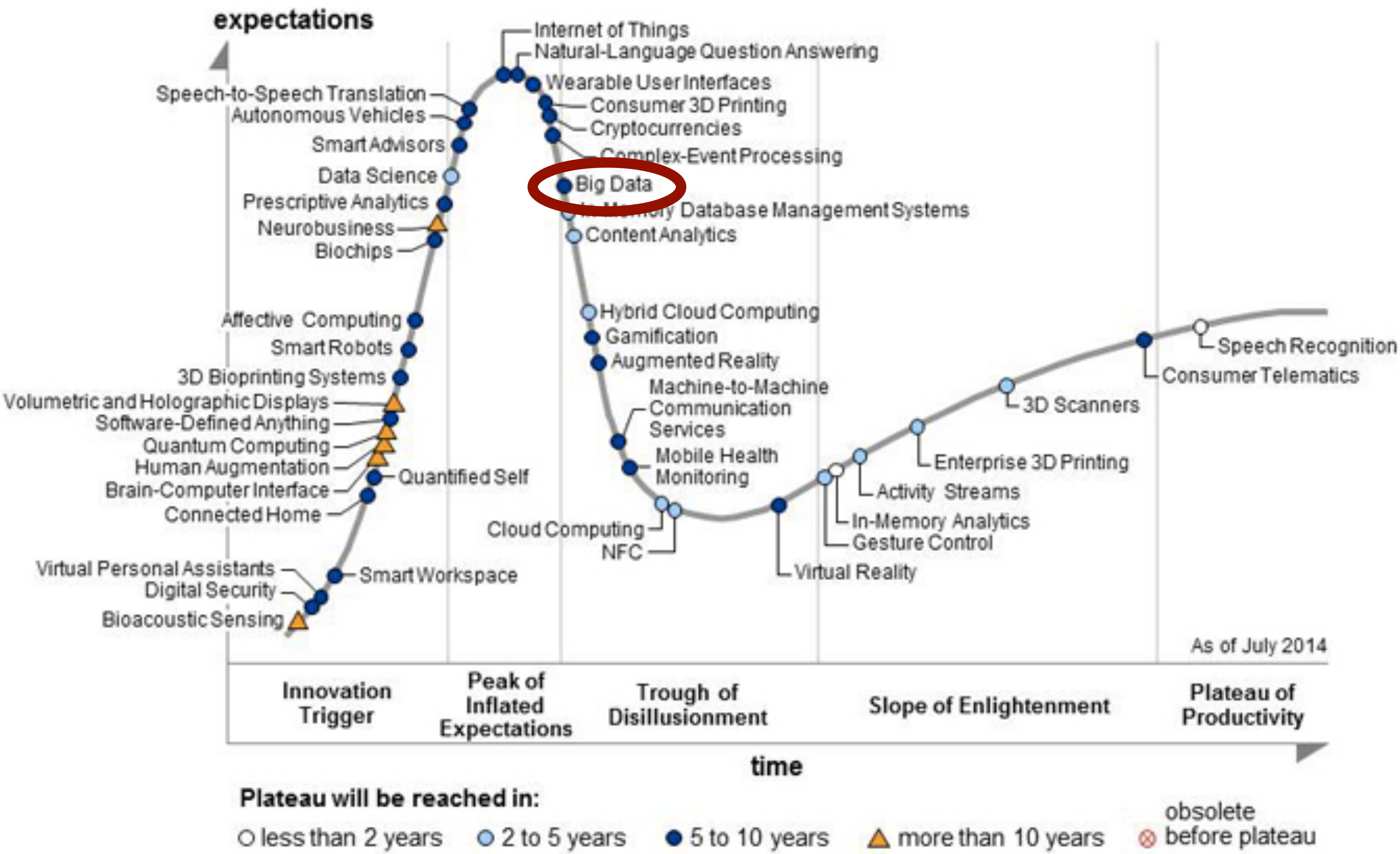
GOOGLE FLU TRENDS (GFT)

Uses search keywords to predict reports by Center for Disease Control

- Methodology
  - First version: find best matches among 50 million searchers to explain 1152 flu cases
  - Later versions: improvements to eliminate other seasonal trend (e.g. basketball)
- Underwhelming results
  - 2009: missed nonseasonal 2009 H1N1 influenza
  - 2013: overestimated the proportion of doctor visits
  - Not better than simpler predictions
- Reasons for the challenges
  - Overfitting and confounding, lacks subject matter info
  - Algorithm dynamics
    - changes to both queries and algorithms
  - Cannot easily replicate the search results, poor documentation.

GARTNER HYPE CYCLE

Emerging technologies 2014



www.gartner.com

SPECIFIC ISSUES

- Large data hide true quantitative signal
- Large data generate spurious correlations
- Large data help mistake correlation for causation
- Large data amplify bias and confounding

'A big computer, a complex algorithm and a long time does not equal science'

— Robert Gentleman

SPECIFIC ISSUES

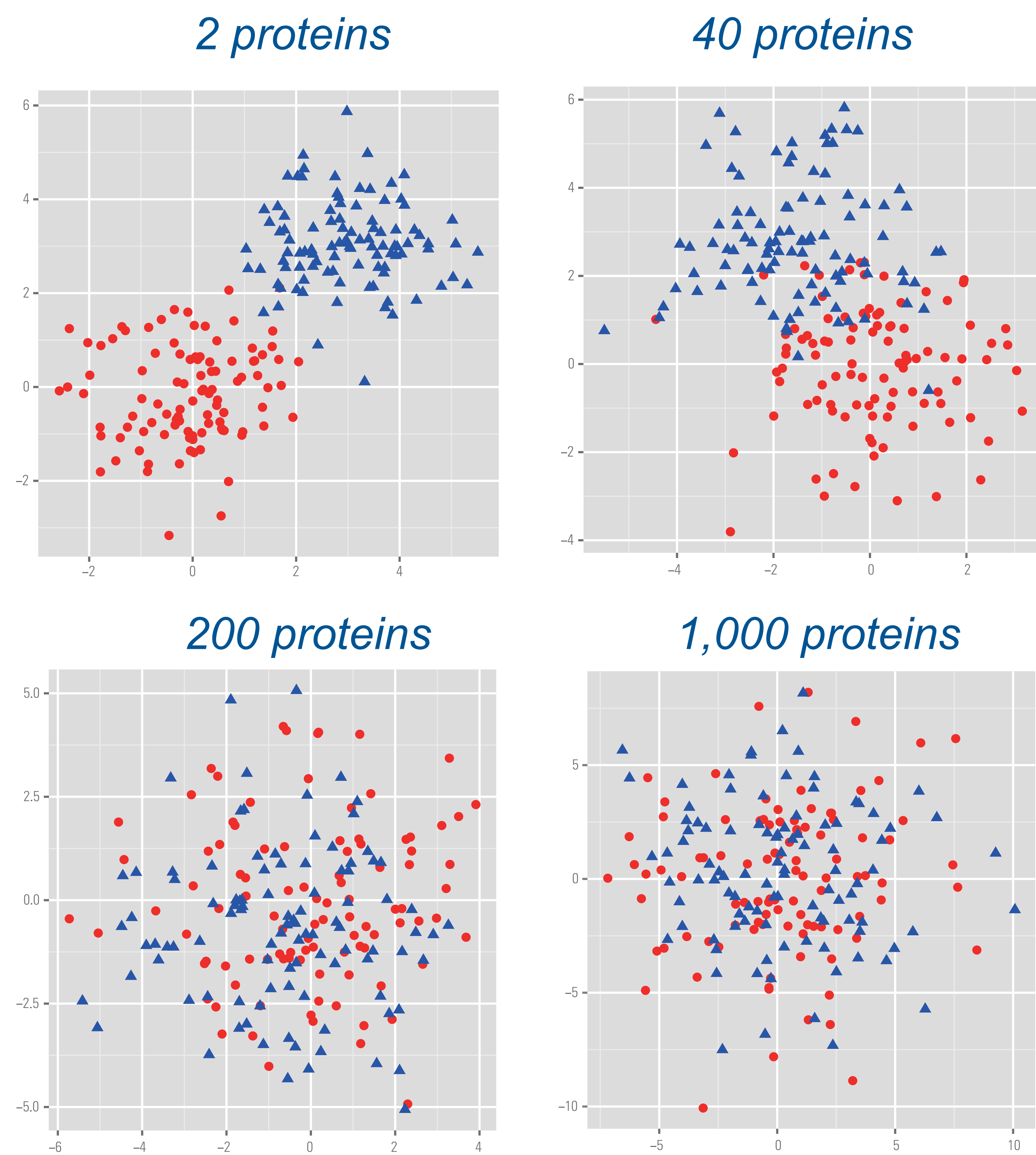
- $\boxed{\text{Large data hide true quantitative signal}}$
- Large data generate spurious correlations
- Large data help mistake correlation for causation
- Large data amplify bias and confounding

'A big computer, a complex algorithm and a long time does not equal science'

— Robert Gentleman

LARGE DATA HIDE SIGNAL

- A simulation study
  - *100 subjects*
  - *2 groups*
  - *10 differentially abundant proteins*
- Plot the first two principle components
  - *Expect good separation between the groups*



'We are drowning in information but starved for knowledge'
— John Naisbitt

Fan et al., National Science Review, 1:293, 2014

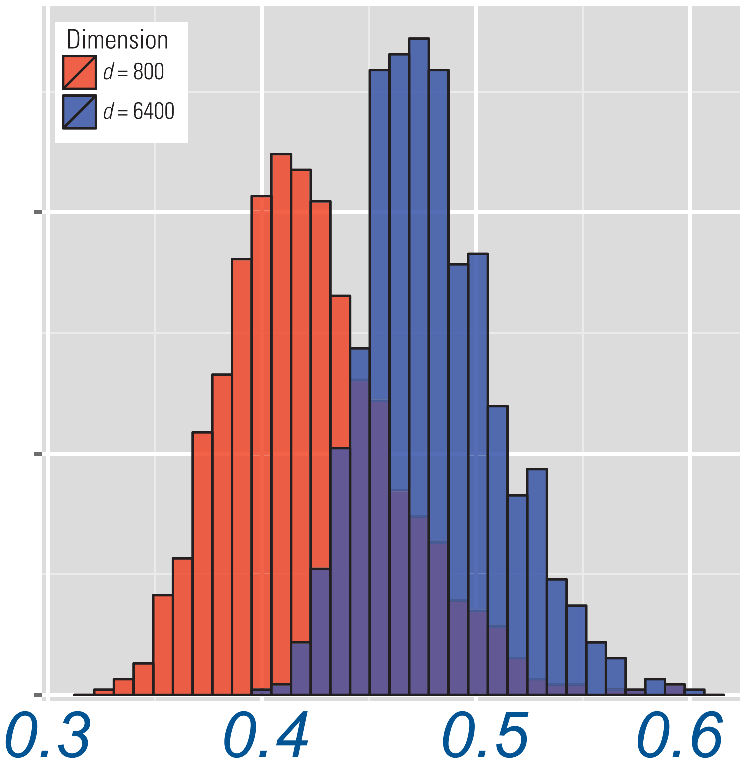
SPECIFIC ISSUES

- Large data hide true quantitative signal
- Large data generate spurious correlations
- Large data help mistake correlation for causation
- Large data amplify bias and confounding

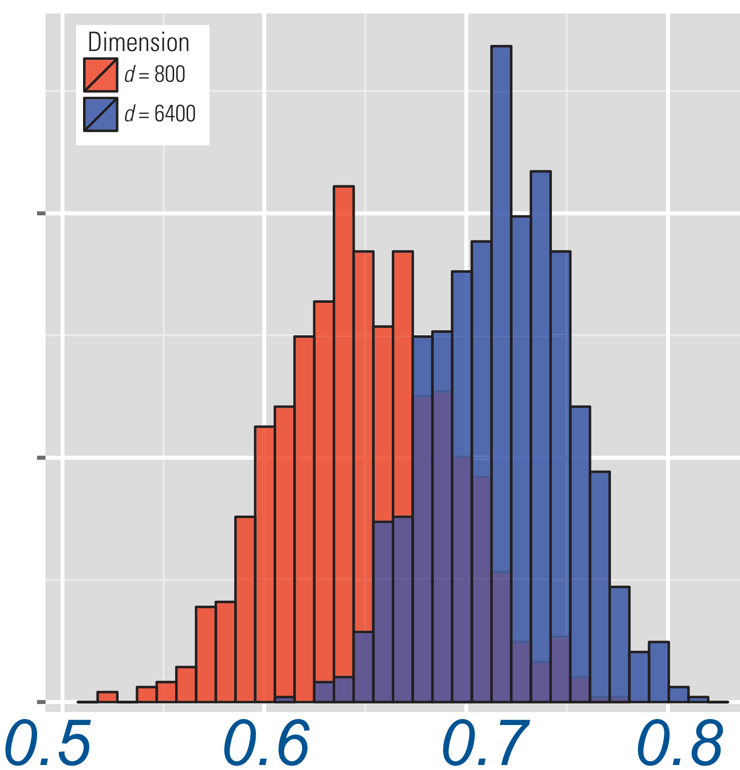
LARGE DATA HIDE SIGNAL

- A simulation study
  - *60 subjects with quantitative phenotype*
  - *red: 800 proteins unrelated to phenotype*
  - *blue: 6400 proteins unrelated to phenotype*
- *Repeat 1,000 times*

Max correlation between the phenotype and a protein



Max correlation between the phenotype and a linear combination of 4 proteins



'With four parameters I can fit an elephant, and with five I can make him wiggle his trunk'
— John von Neumann

Fan et al., National Science Review, 1:293, 2014

Drawing an elephant with four complex parameters

Jürgen Mayer
*Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstr. 108, (
Germany*

Khaled Khairy
European Molecular Biology Laboratory, Meyerhofstraße. 1, 69117 Heidelberg, Ge

Jonathon Howard
*Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstr. 108, (
Germany*

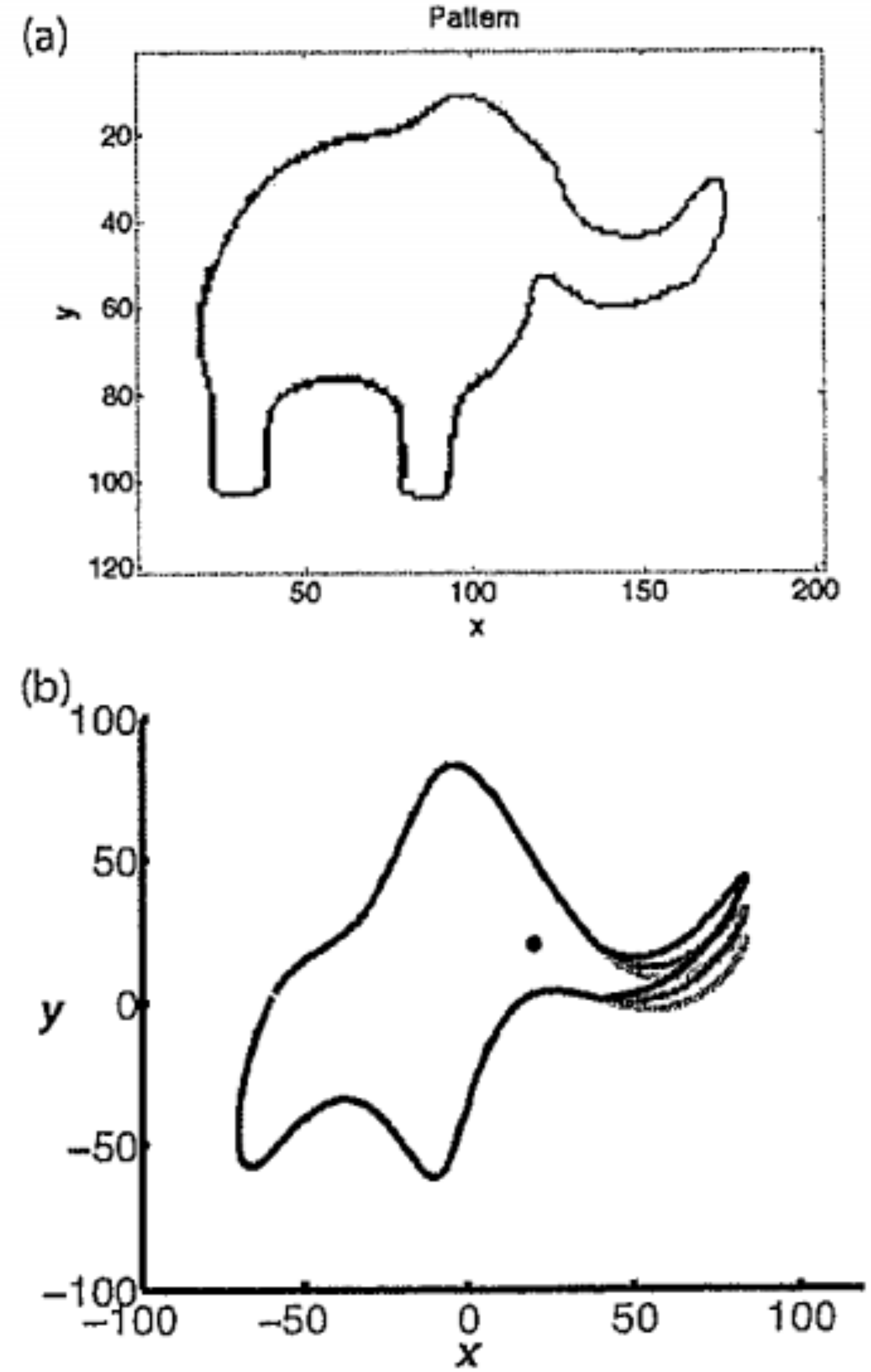
(Received 20 August 2008; accepted 5 October 2009)

$$x(t)=\sum_{k=0}^{\infty}(A_k^x\cos(kt)+B_k^x\sin(kt)),$$

$$y(t)=\sum_{k=0}^{\infty}(A_k^y\cos(kt)+B_k^y\sin(kt)),$$

Table I. The five complex parameters $p_1,\ldots,p_5$ that encode the elephant including its wiggling trunk.

Parameter	Real part	Imaginary part
$p_1=50-30i$	$B_1^x=50$	$B_1^y=-30$
$p_2=18+8i$	$B_2^x=18$	$B_2^y=8$
$p_3=12-10i$	$A_3^x=12$	$B_3^y=-10$
$p_4=-14-60i$	$A_5^x=-14$	$A_1^y=-60$
$p_5=40+20i$	Wiggle coeff.=40	$x_{eye}=y_{eye}=20$

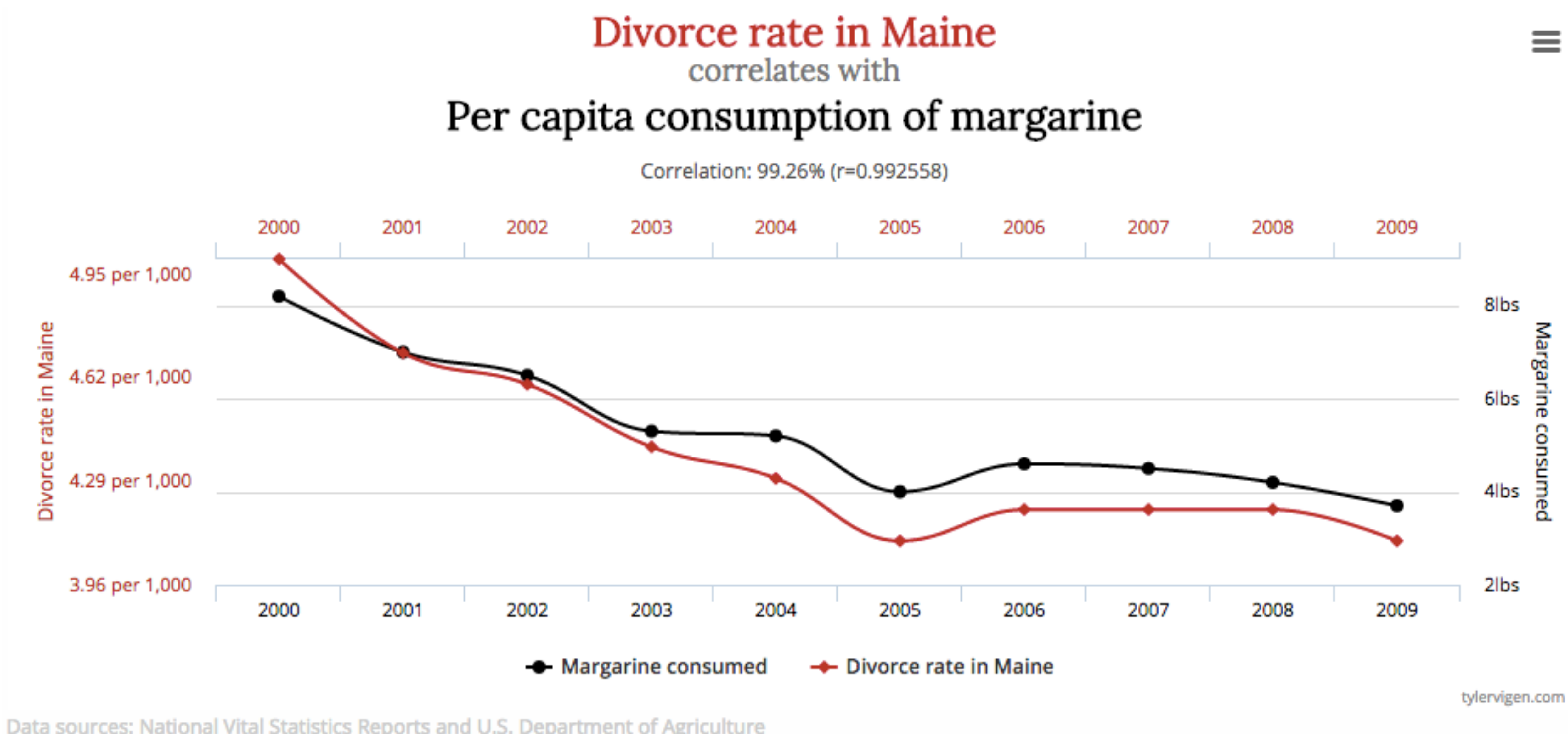


Fourier coordinate expansion with complex numbers as parameters

SPECIFIC ISSUES

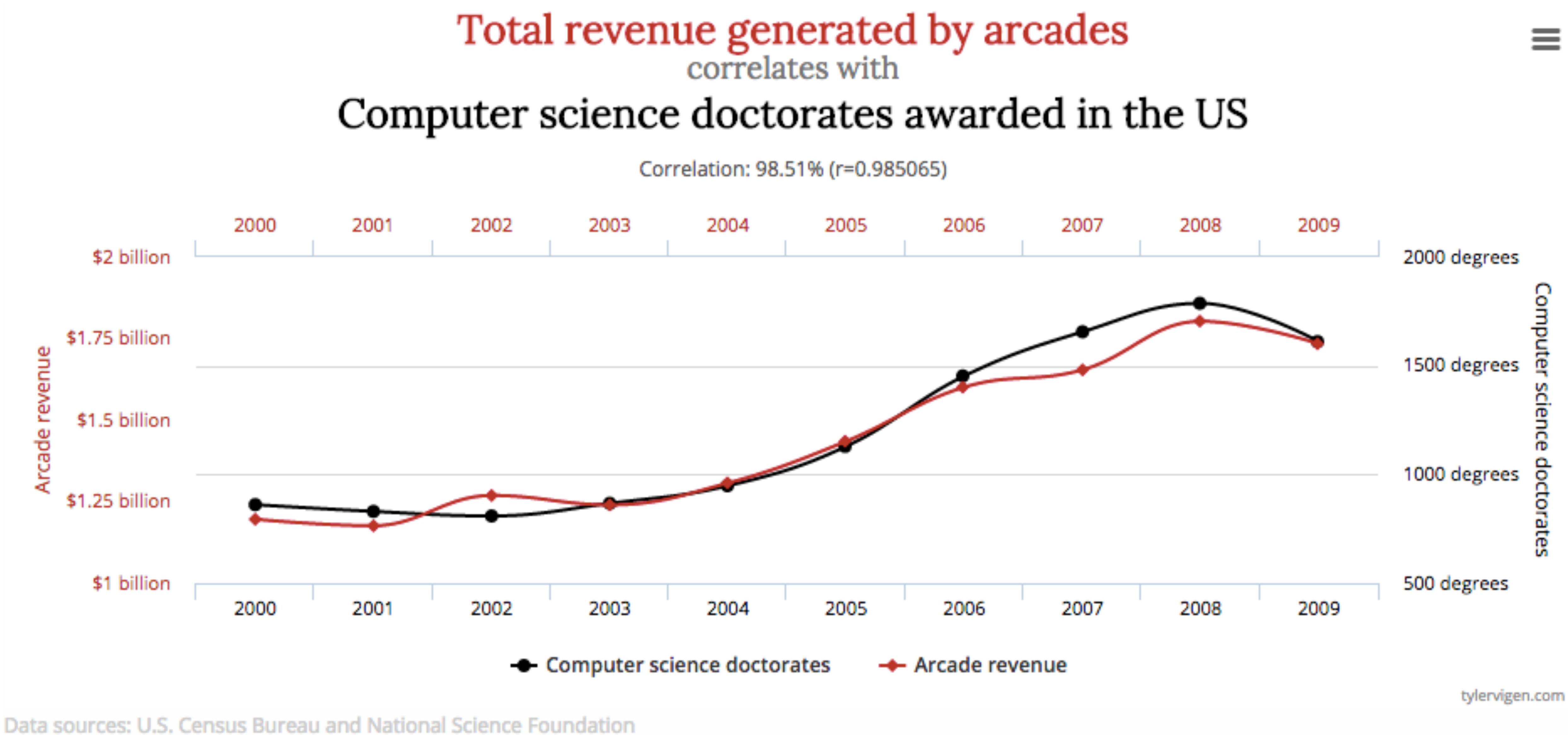
- Large data hide true quantitative signal
- Large data generate spurious correlations
- Large data help mistake correlation for causation
- Large data amplify bias and confounding

SPURIOUS CORRELATIONS ABOUND



tylervigen.com/spurious-correlations

SPURIOUS CORRELATIONS ABOUND



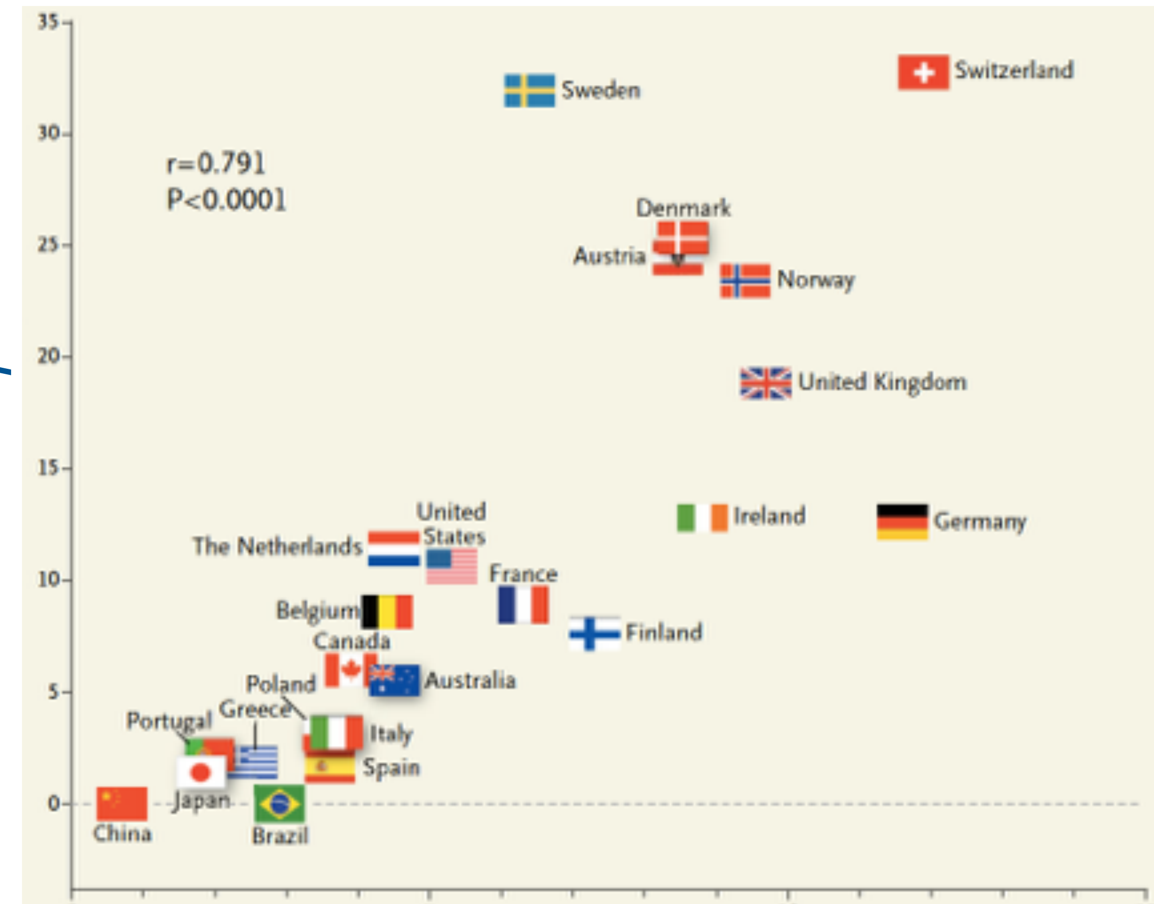
tylervigen.com/spurious-correlations

SPURIOUS CORRELATIONS ABOUND

Easy to dismiss when we understand the context

- Premier medical journal
  - *Nobel prize is related to cognitive ability*
  - *flavanols (organic molecules present in chocolate) are linked to cognitive ability*
- Technical flows
  - *Nobel prize winners between 1900-2011*
  - *Chocolate consumption after 2002*
  - *Countries with many Nobel prizes have a high Human Development Index and high per capita income*

Nobel laureates per 10 mio



Chocolate consumption (kg/yr/capita)

New England Journal of Medicine, 367:1562 (2012)
A. Jogalekar, Scientific American, 2012

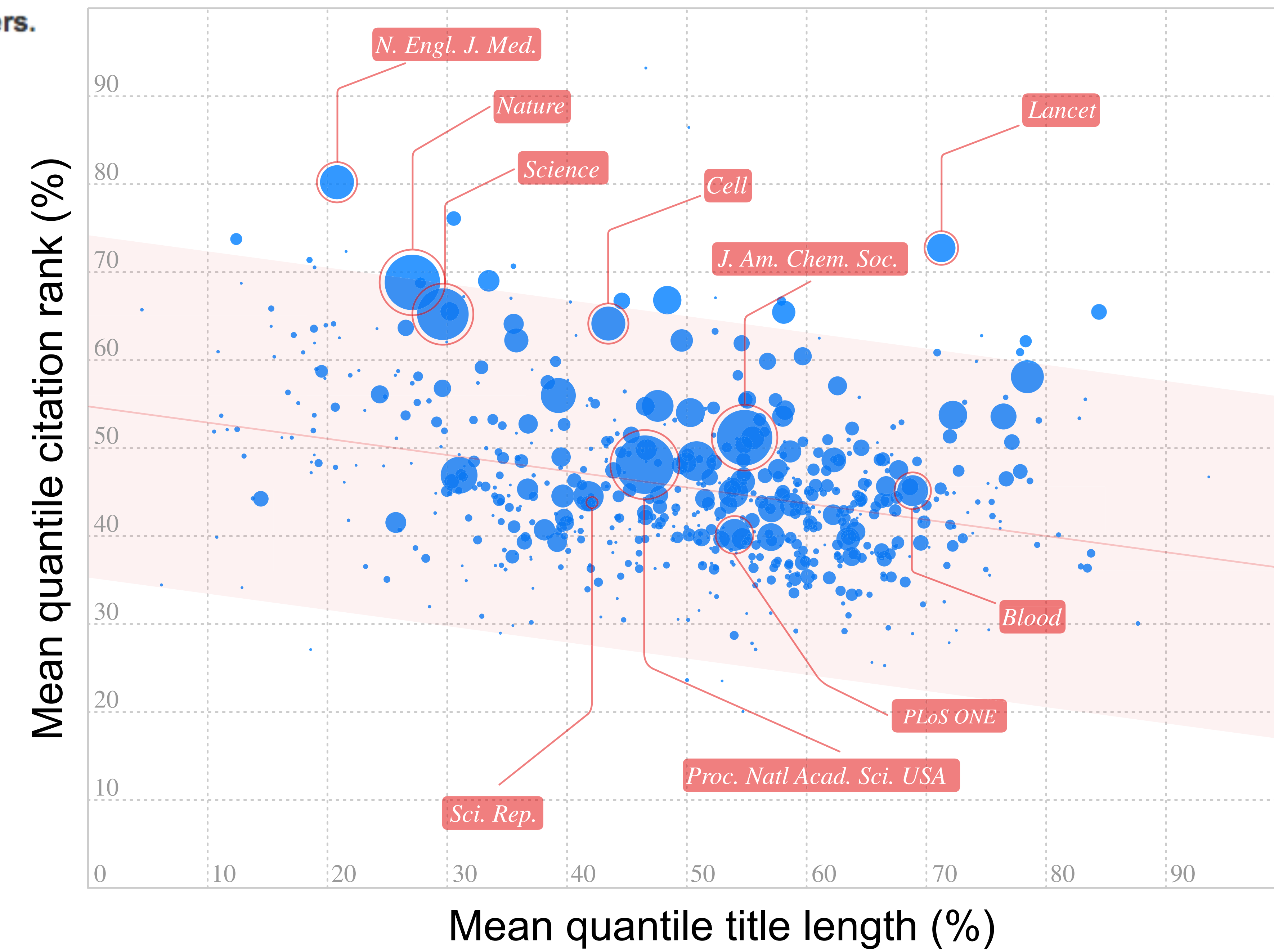
nature International weekly journal of science

NATURE | NEWS

Papers with shorter titles get more citations

Intriguing correlation mined from 140,000 papers.

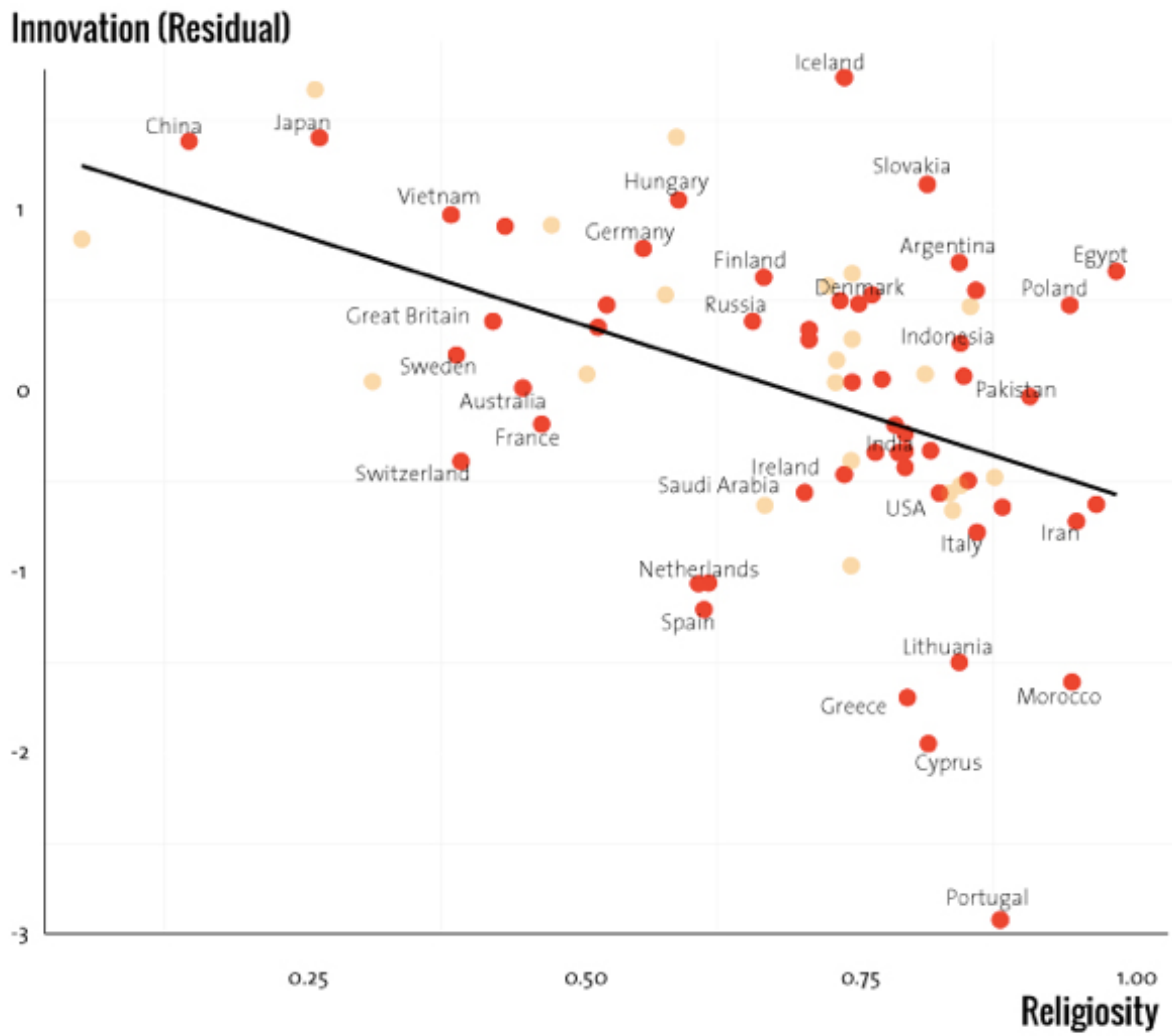
Boer Deng



A. Letchford et al., Royal Society Publishing, 2015

SPURIOUS CORRELATIONS ABOUND

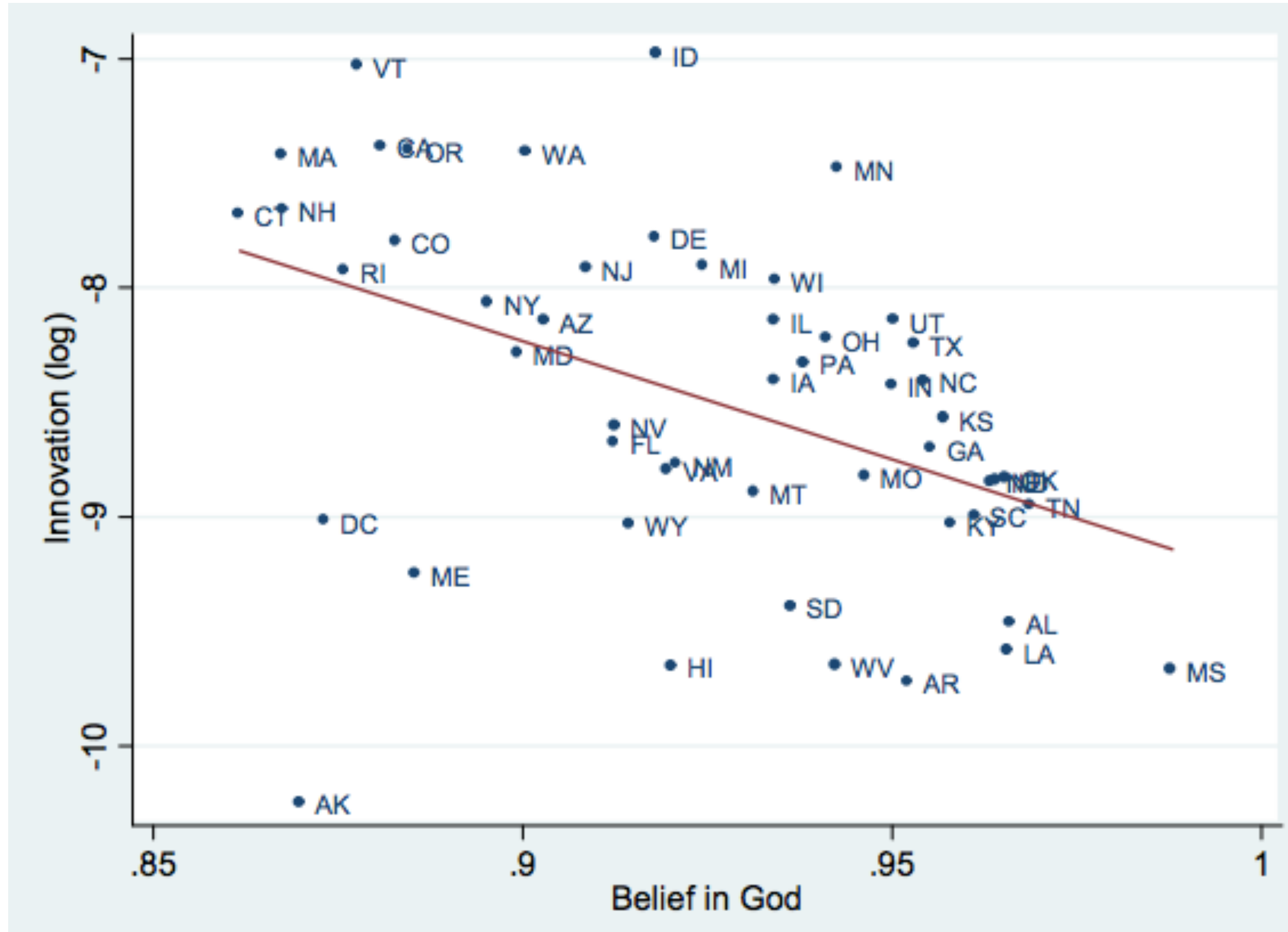
Not easy to dismiss when the context is unknown



Benabou et al., Princeton University

SPURIOUS CORRELATIONS ABOUND

Not easy to dismiss when the context is unknown



‘Correlation doesn’t imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing ‘look over there’

Benabou et al., Princeton University

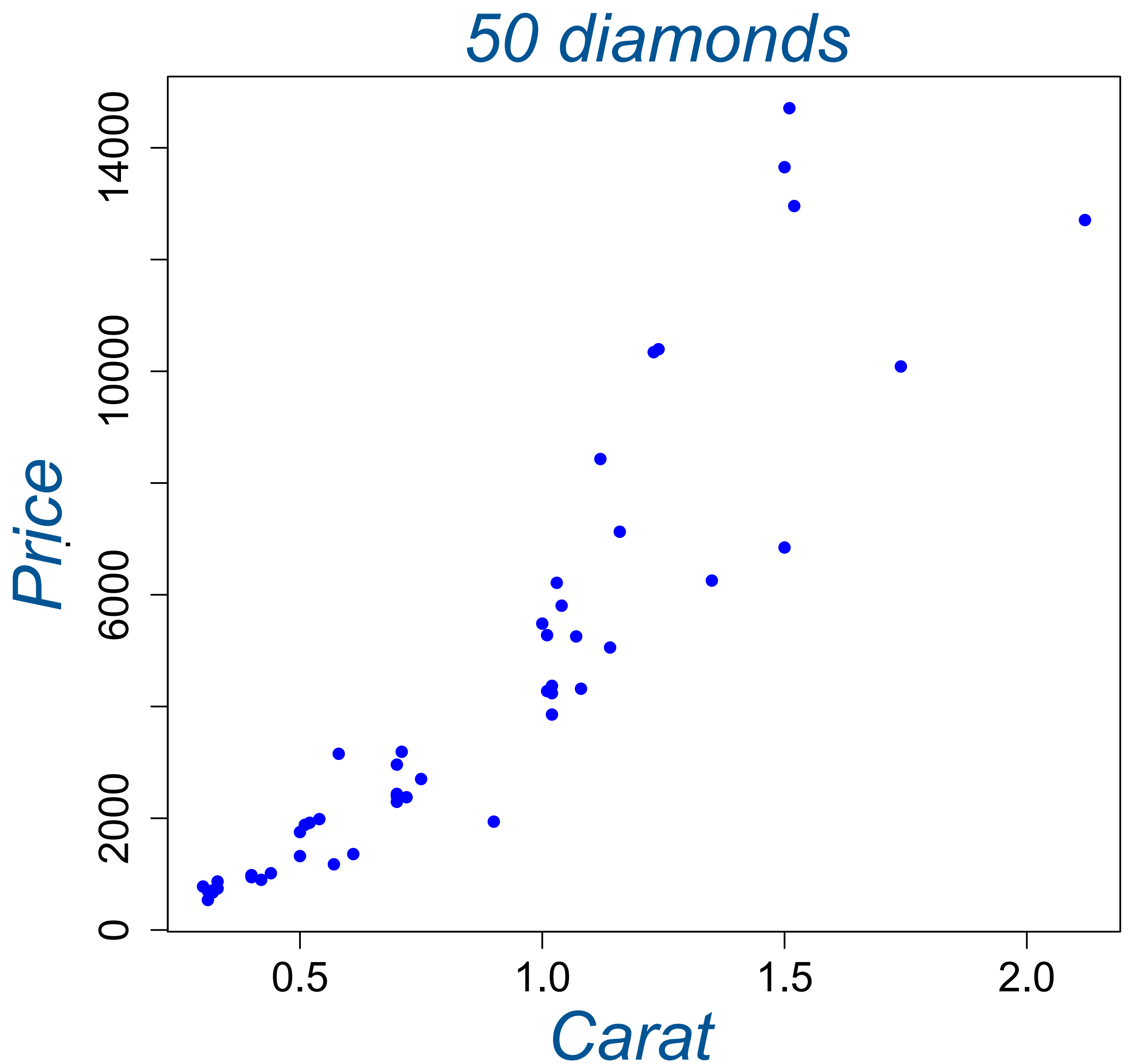
SPECIFIC ISSUES

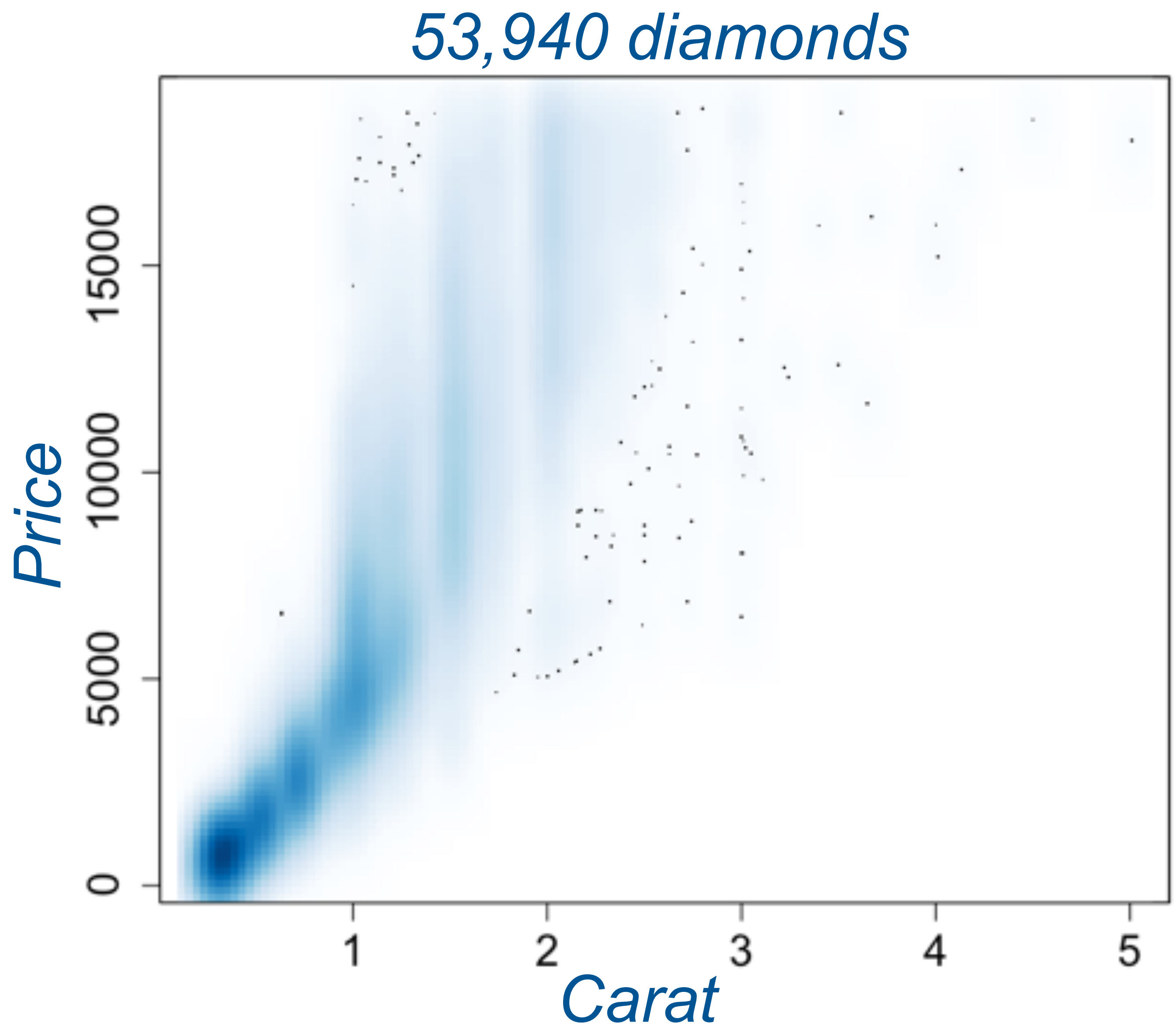
- Large data hide true quantitative signal
- Large data generate spurious correlations
- Large data help mistake correlation for causation
- Large data amplify bias and confounding

EXAMPLE

53,940 diamonds

carat	color	price
0.23	E	326
0.21	E	326
0.23	E	327
0.29	I	334
0.31	J	335
....		





EXAMPLE

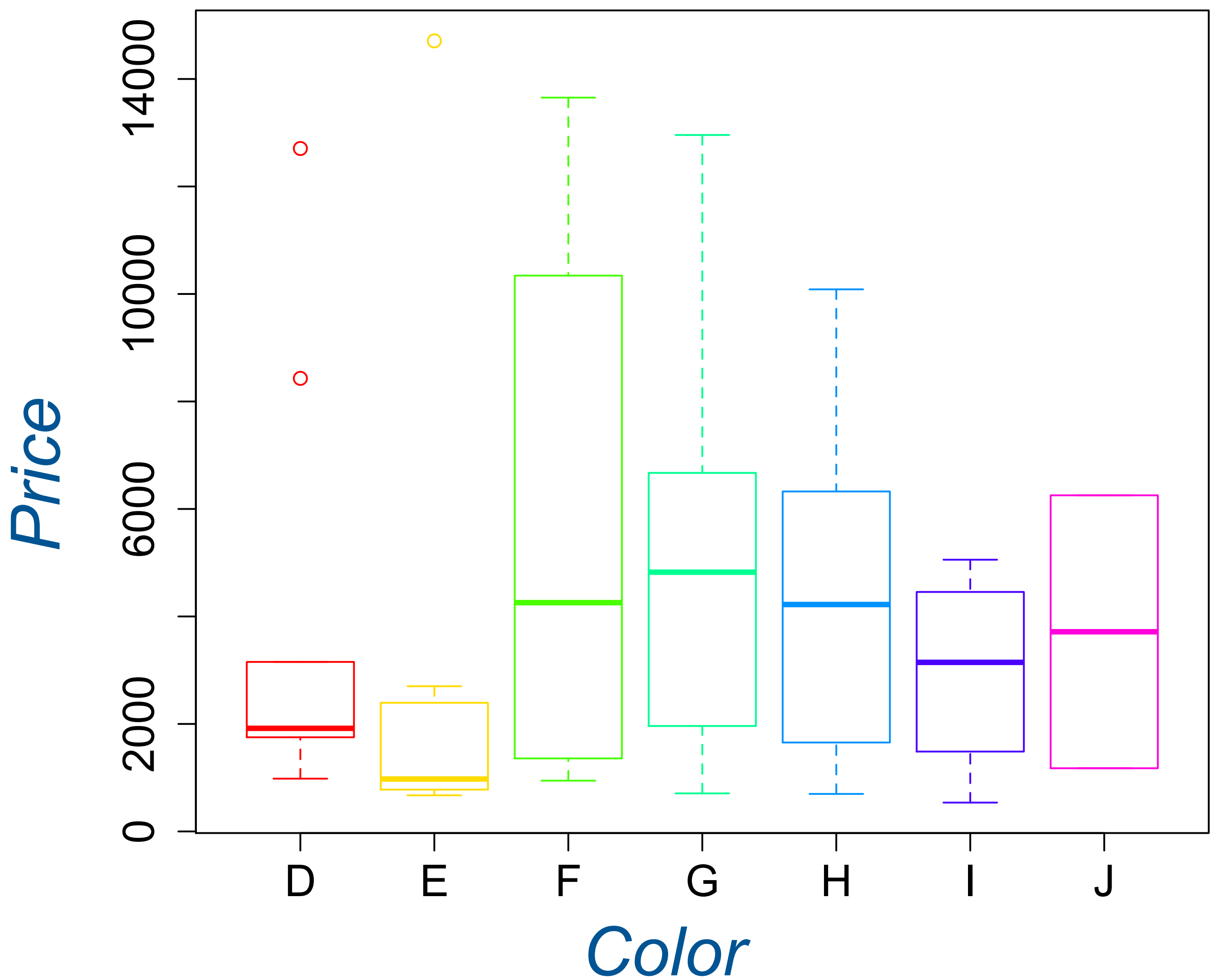
53,940 diamonds

- New discovery!
  - later colors cost more!

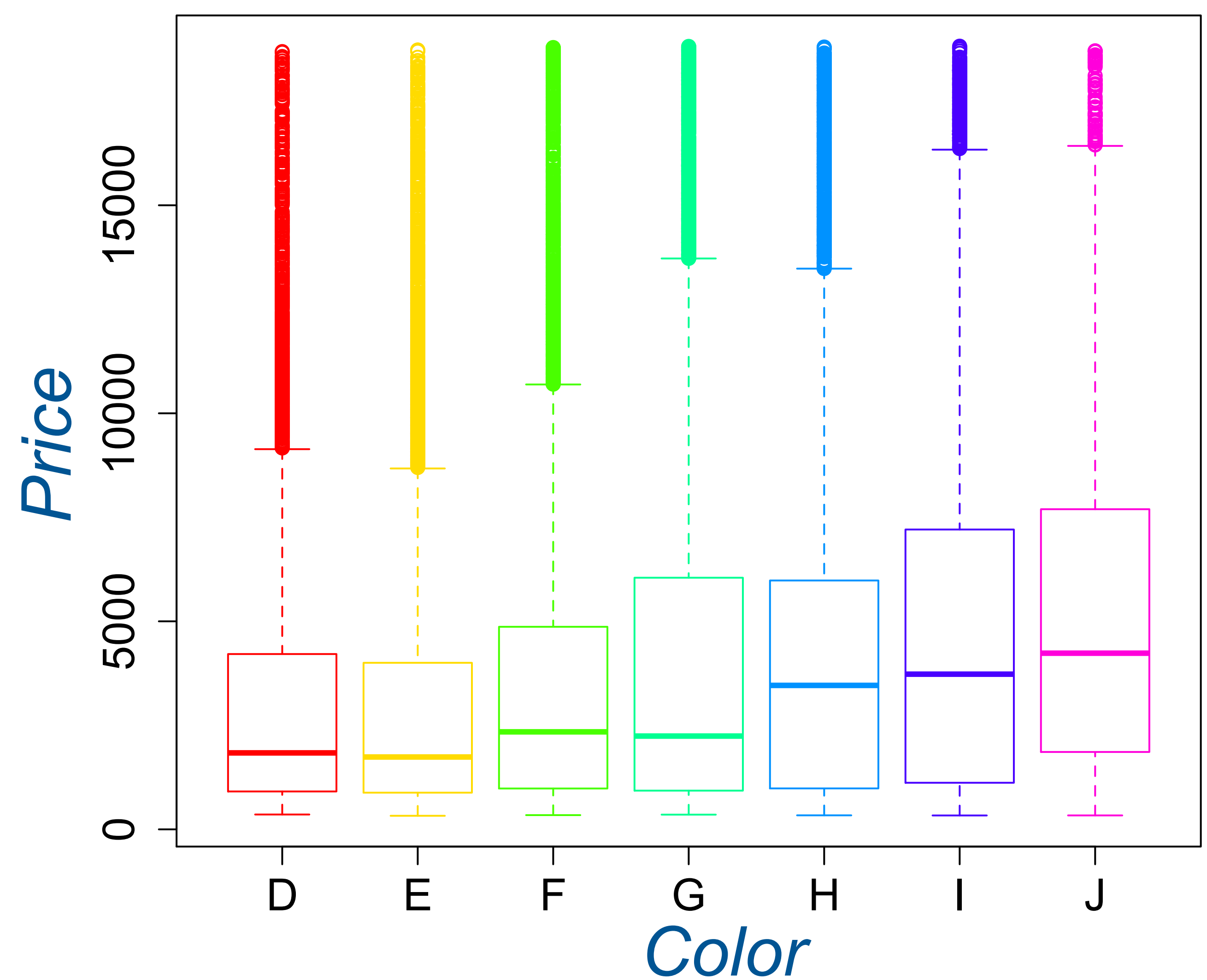
carat	color	price
0.23	E	326
0.21	E	326
0.23	E	327
0.29	I	334
0.31	J	335

....

50 diamonds



53,940 diamonds



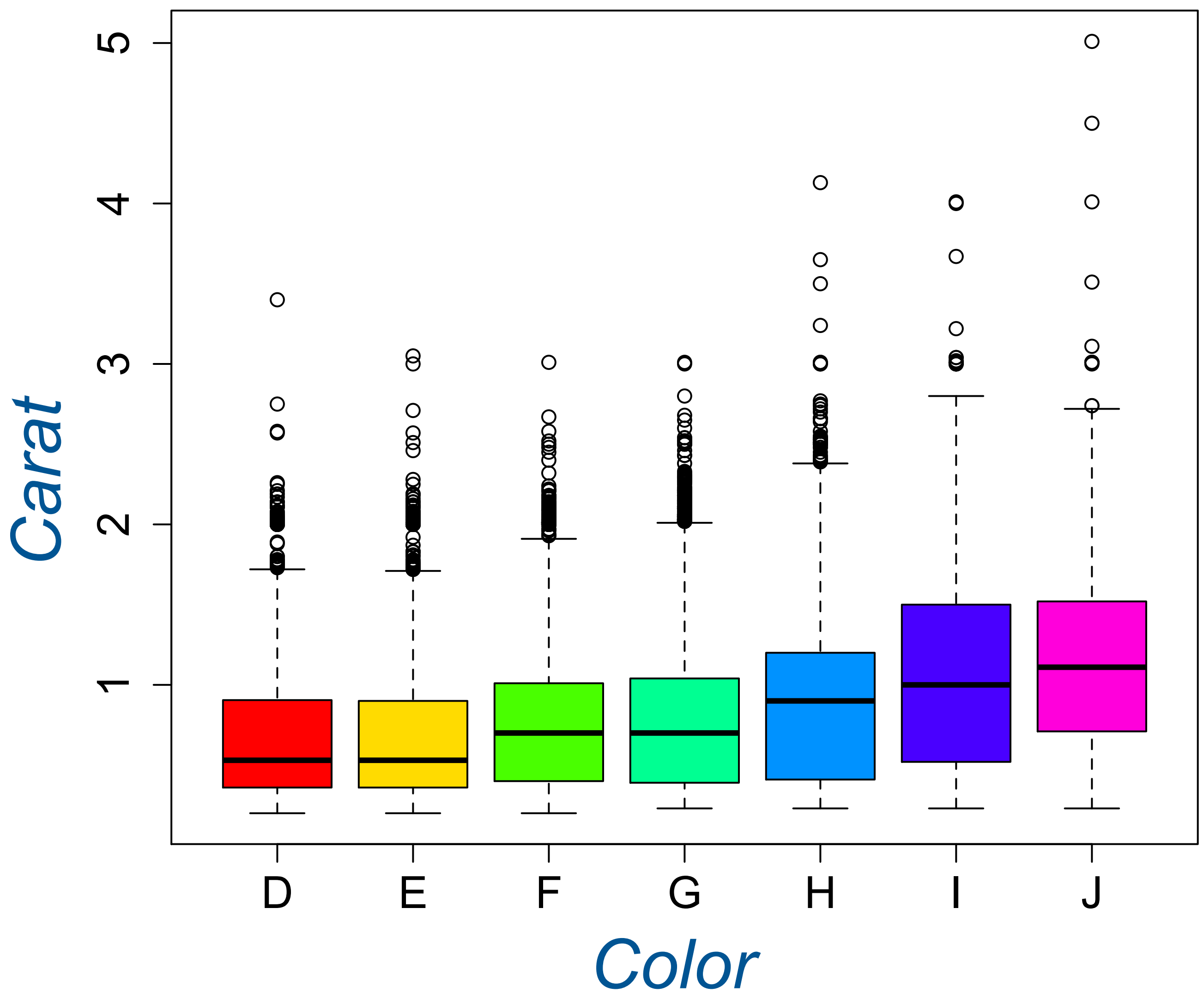
EXAMPLE

- Subject matter knowledge
  - later colors are cheaper
  - they also weigh more
  - Both color and weight affect price

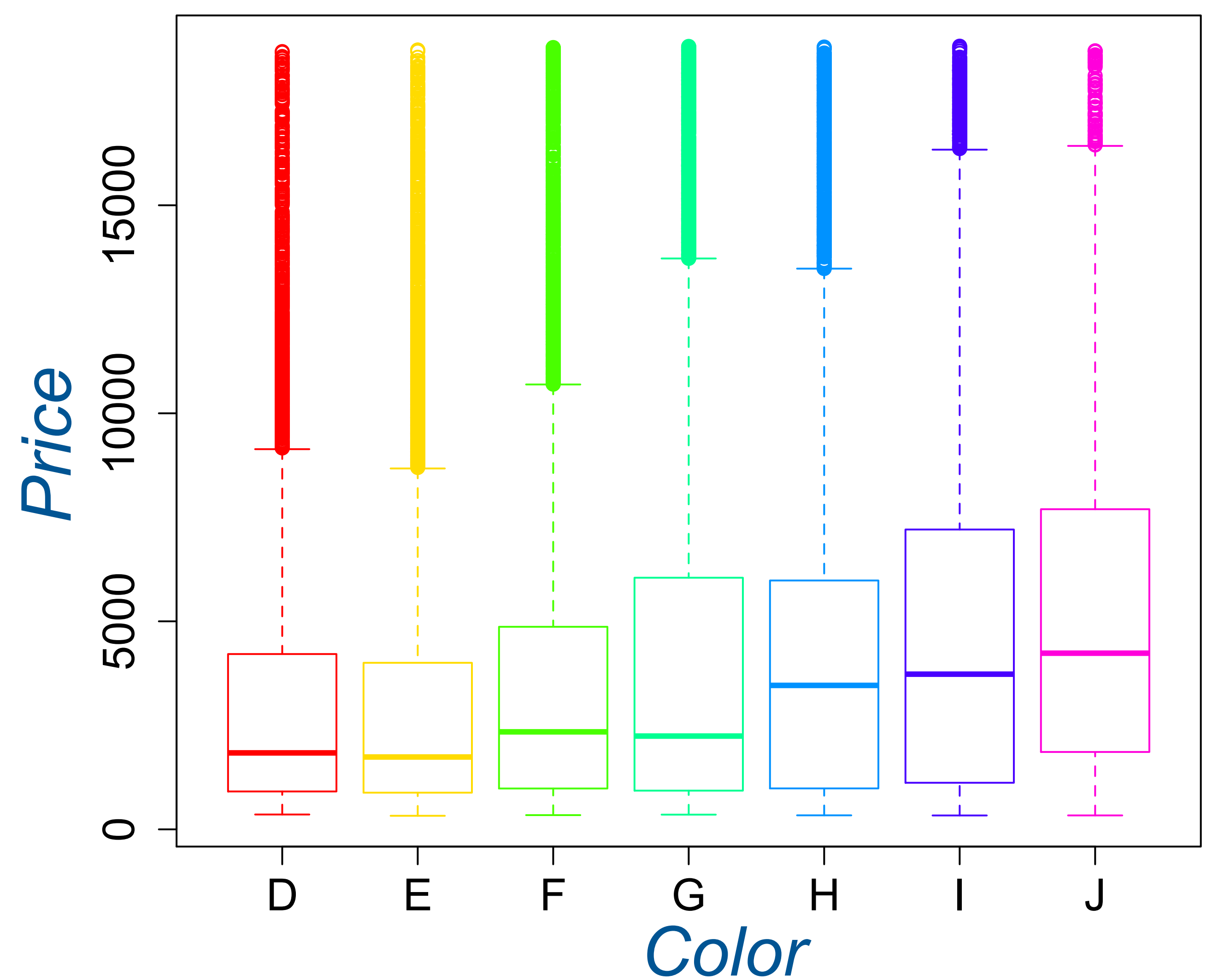
carat	color	price
0.23	E	326
0.21	E	326
0.23	E	327
0.29	I	334
0.31	J	335

....

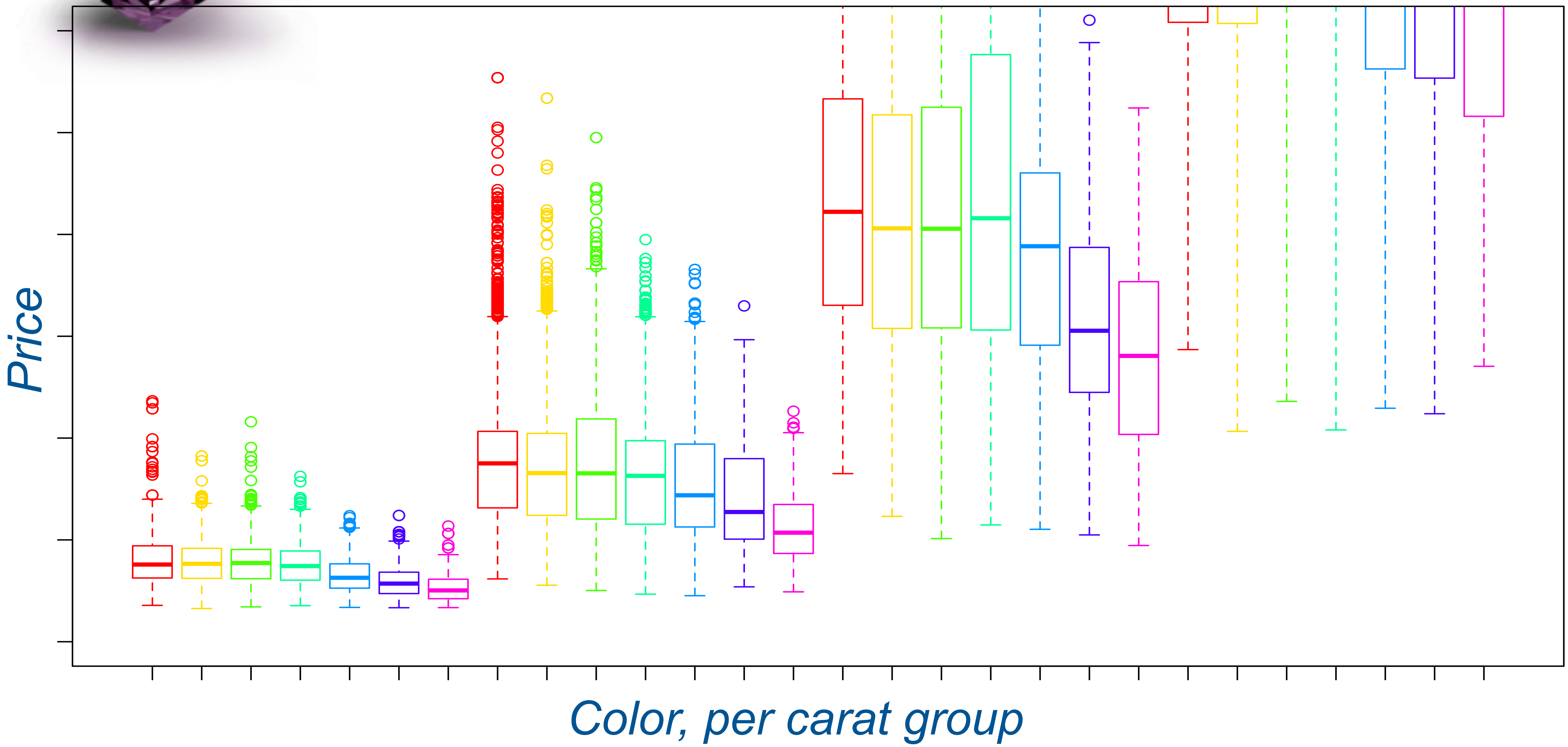
53,940 diamonds



53,940 diamonds



EXAMPLE



'To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of'
— Ronald Fisher

SUMMARY

- More data ≠ more information
- We should:
  - state clearly the scientific question
  - follow the fundamental principles of experimental design
  - select methods that are appropriate for the question
    - more complexity does not mean more insight!
  - use problem-specific information
- Data and algorithms do not substitute thinking through the problem

'There are no routine statistical questions, only questionable statistical routines'

— D. R. Cox