Introduction

CS 6220
'Data mining'

Professor Olga Vitek

September 10, 2015

Making sense of the terms

- Data mining
  - Analysis of (often large) observational datasets to *find unexpected relationships*
  - Often secondary, exploratory analysis of convenience (opportunity) datasets
- Machine learning
  - Specific tasks associated with class discovery (unsupervised learning), class prediction (supervised learning), and class comparison (testing)
- Statistics
  - Collection and analysis of data, to *make inference beyond the current dataset*.
  - Characterized by *measures of uncertainty* , and of decision making in presence of uncertainty
  - Often primary, confirmatory analysis of designed experiments or ad-hoc datasets
- Data science
  - Often used interchangeably with data mining
  - Often used in 'data-driven decision making'

Large observational datasets are increasingly common

- **Physics:** Large Hadron Collider
  - 150 million sensors, 600 million collisions/sec
- **Astronomy:** Sloan Digital Sky Survey
  - In 2000, collected more data in its first few weeks than all data in the history of astronomy
  - Now 200 GB per night, over 140 terabytes
  - In 2016, the Large Synoptic Survey Telescope will acquire that amount every five days
- **Genomics:** Sequencing human genome
  - First took 10 years, now in less than a day
- **Climate:** NASA Center
  - 32 petabytes of climate data & simulations
- **E-commerce:** Amazon
  - Millions of back-end operations / day
  - Queries from 1/2 million third-party sellers.
  - In 2005, databases of 7.8, 18.5, & 24.7 TB

en.wikipedia.org/wiki/Big_data

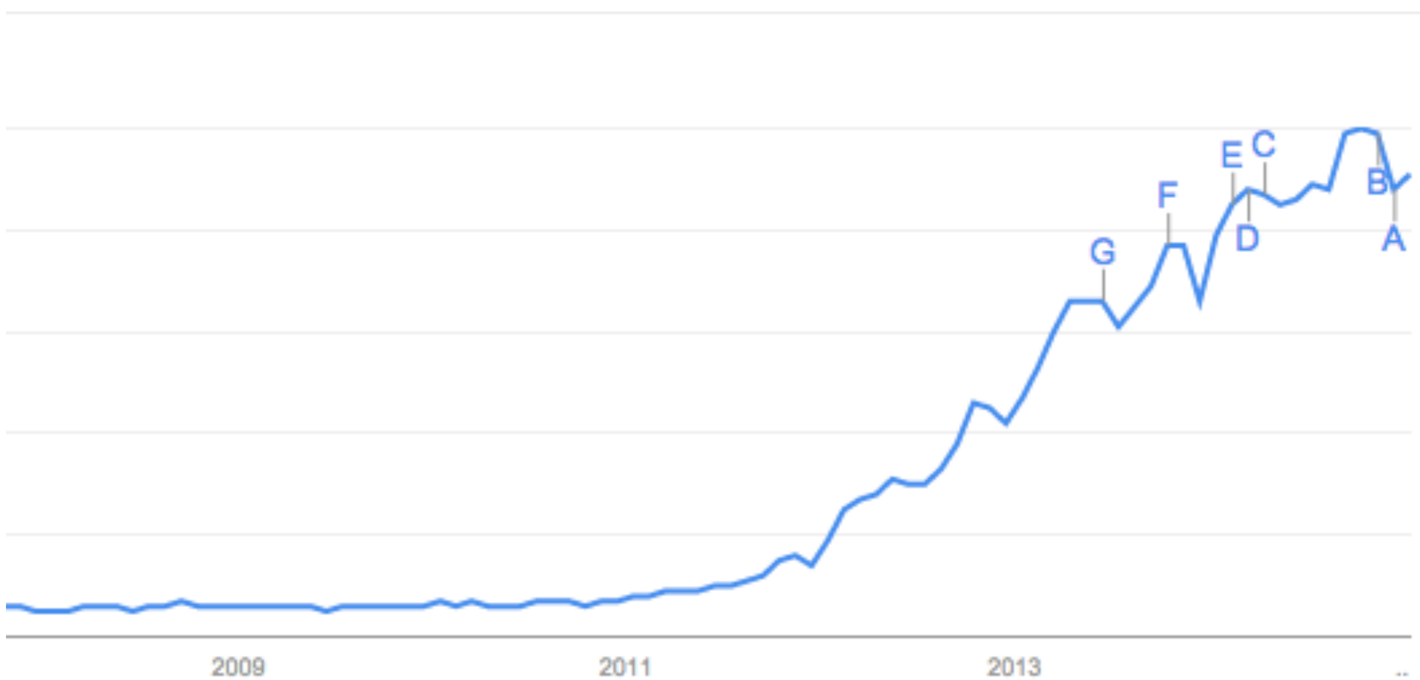
Mining big data: a great promise

Example: big data success stories (IBM marketing)

- Applies emerging technologies to deliver instantaneous people searches
- Analyzing huge volumes of customer comments in real time delivers competitive edge
- Analyzes real-time data streams to identify traffic patterns
- Putting real-time data to work and providing a platform for technology development
- Helping companies deliver the web experience their customers want.
- Streaming data technology supports covert intelligence and surveillance sensor systems
- Leveraging key data to provide proactive patient care
- Streaming real-time data supports large scale study of space weather
- Turning climate into capital with big data

public.dhe.ibm.com/software/data/sw-library/big-data/ibm-big-data-success.pdf

A great excitement



Google trends: 'Big Data' (01/11/2015)

Are big data 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

Chris Anderson. Wired Magazine: 16.07

archive.wired.com/science/discoveries/magazine/16-07/pb_theory

Are big data 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 underlying mechanisms
  - Do not need to settle for wrong models. We can succeed without them

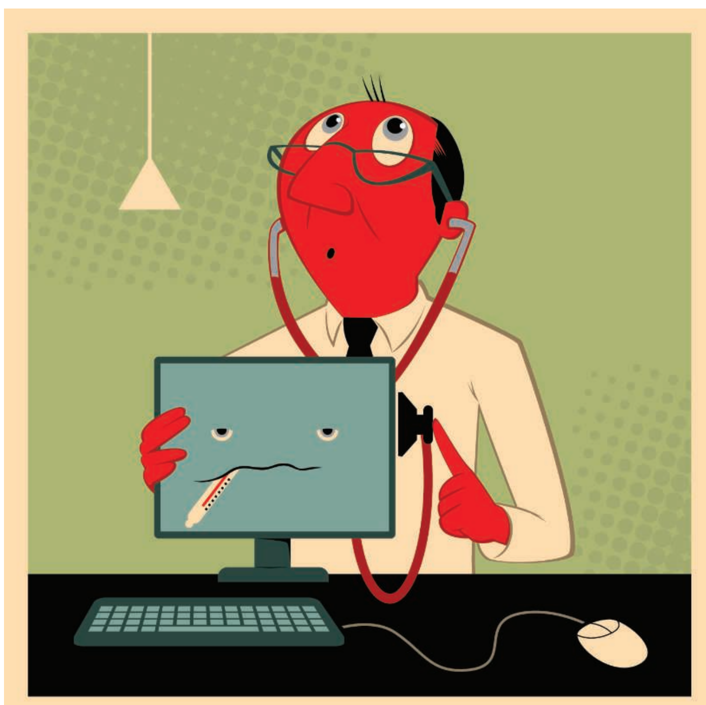
- What is the new scientific method?
  - The information is readable, reachable and queryable
  - Statistical tools will crunch the numbers and offer a new way of understanding the world
  - "There's no reason to cling to our old ways. It's time to ask: What can science learn from Google?"

Case study: Google Flu

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

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

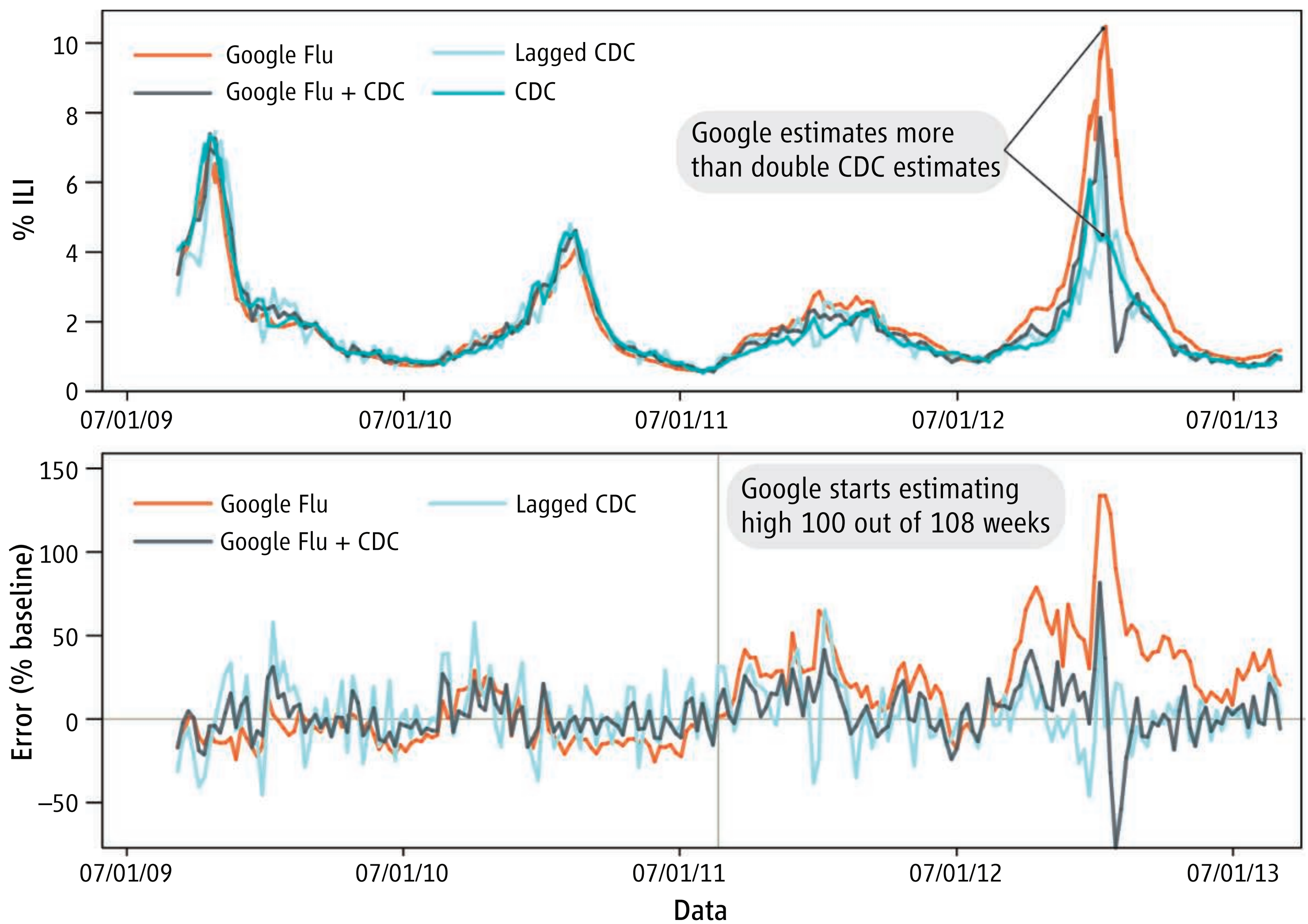


Science Vol 343, March 2014

Google Flu Trends

- Promising concept
  - Find best matches among 50 million searchers to explain 1152 flu cases
- Poor performance
  - 2009: missed nonseasonal 2009 H1N1 influenza
  - 2013: overestimated the % of doctor visits
- Later versions
  - 2009: One predictor is basketball season
  - Confounding between flu and winter
  - 2013: Eliminated basketball and other seasonal trends
  - Not better than simpler predictions

Google Flu Trends



Conclusion: we would have ran away with a wrong prediction

- overestimated the prevalence of flu in the 2012-2013 season
- overshot the actual level in 2011-2012 by > 50%

Sources of challenges

- Statistical challenges
  - Overfitting
  - Confounding
  - Lack subject matter info
- Algorithm dynamics
  - Changes to queries in real time
  - Changes to algorithms in real time
- Cannot easily replicate the results
  - Proprietary methods are poorly documented

Formally show the dangers

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

John P. A. Ioannidis.

PLoS Medicine, Volume 2, Issue 8, e124, 2005

Model: framework for false positive findings

Table 1. Research Findings and True Relationships

Research Finding	True Relationship		
	Yes	**No**	**Total**
Yes	$c(1-\beta)R/(R+1)$	$c\alpha/(R+1)$	$c(R+\alpha-\beta R)/(R+1)$
No	$c\beta R/(R+1)$	$c(1-\alpha)/(R+1)$	$c(1-\alpha+\beta R)/(R+1)$
Total	$cR/(R+1)$	$c/(R+1)$	c

c=number of relationships being probed

R= 'number of true relationship to no relationship'
α=Type 1 error
β=Type 2 error

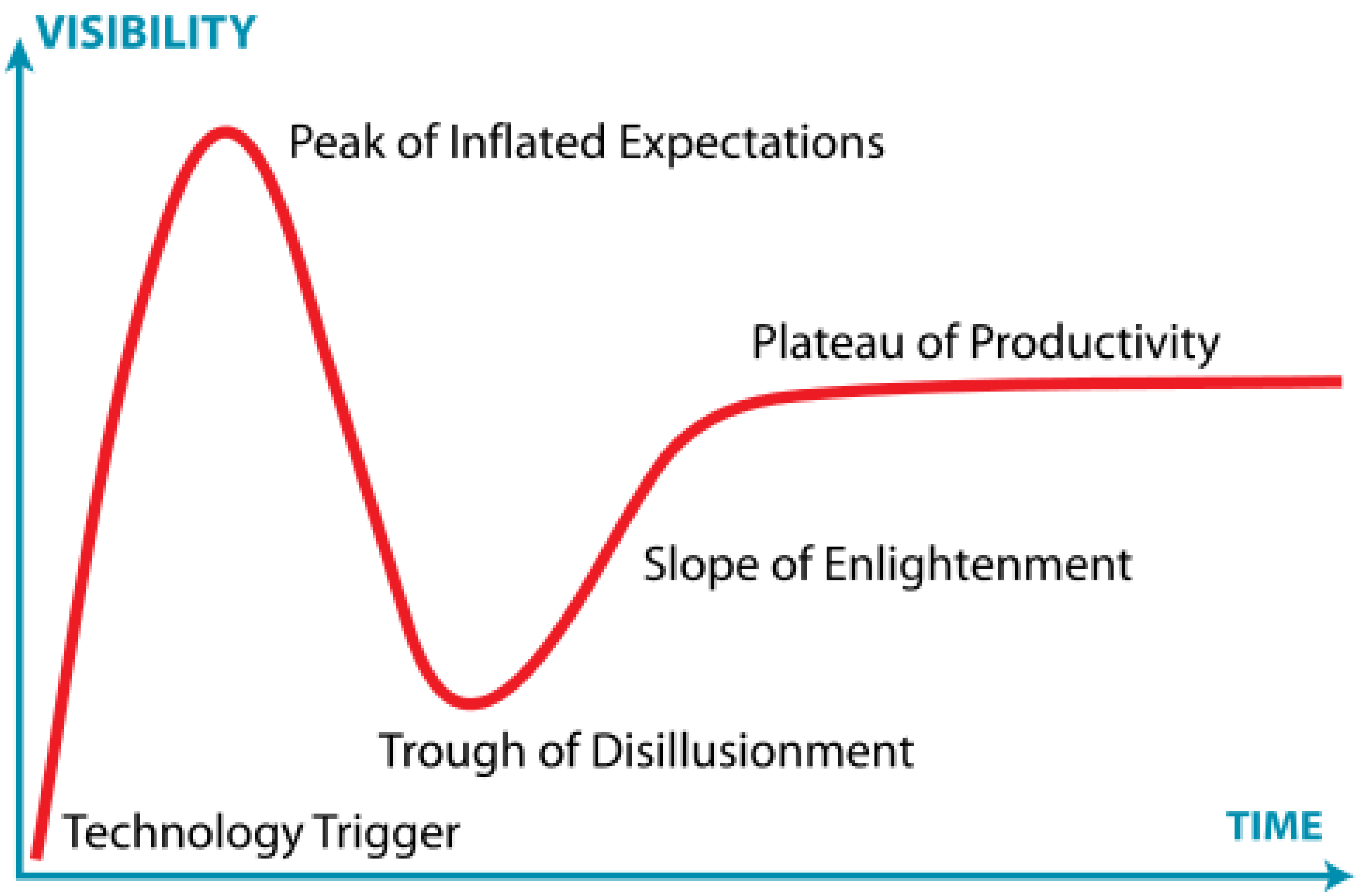
Repeat a similar analysis in presence of bias

Conclusions

- True for small data:
  - The smaller the study, the less likely the research findings are to be true
  - The smaller the effect size, the less likely the research findings are to be true
- True for big data:
  - The greater the number and the lesser the selection of tested relationships, the less likely the research findings are to be true
  - The greater the flexibility in designs, definitions, outcomes, and analytical modes, the less likely the research findings are to be true
  - The greater the financial and other interests and prejudices, the less likely the research findings are to be true
  - The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true

The Gartner Hype Cycle

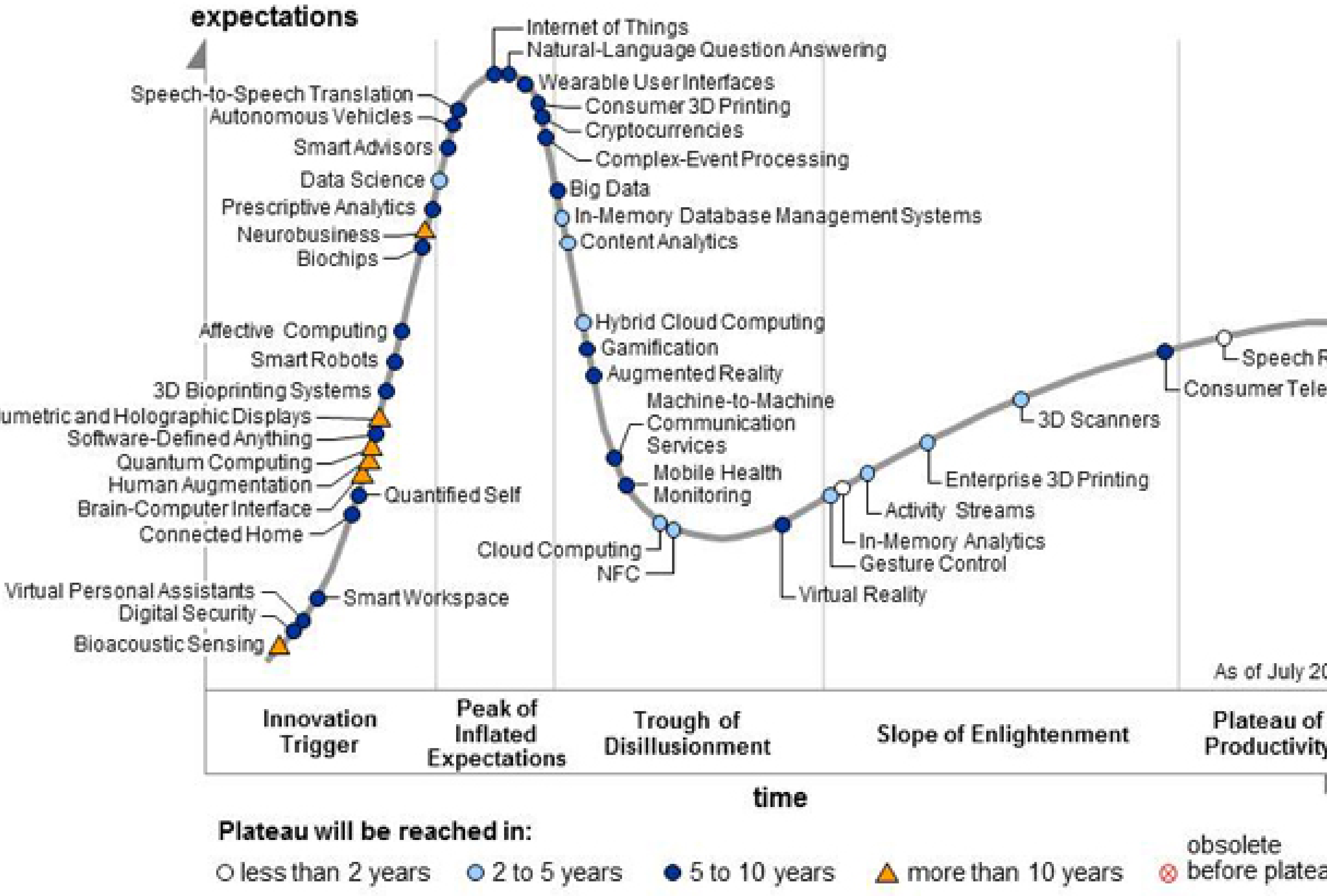
Big data passed its peak of inflated expectations



www.wikipedia.org

The Gartner Hype Cycle

Big data passed its peak of
inflated expectations



www.gartner.com

Task at hand:

Understand the strengths and the limitations of the methods to move to the productivity stage

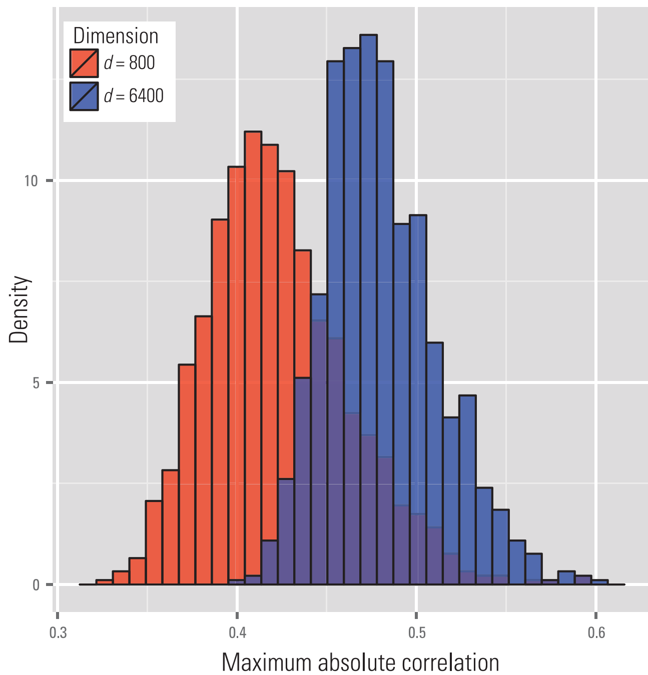
Challenges of modern data

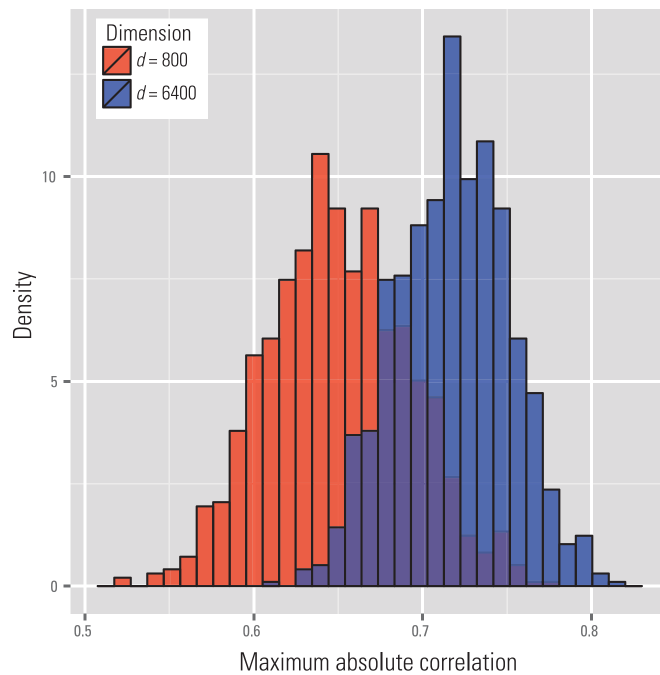
- Many datasets form an array with n observations, and p variables
  - **Large and complex in p**
    - Large p
    - Complex dependencies between predictors
  - **Large and complex in n**
    - Large n
    - Heterogeneity of observations
  - Hard to compute, visualize, summarize
- Many datasets are not arrays
  - Networks, sequences, time series
  - Even more complex dependencies
- Often the mechanism underlying the associations is unknown

Challenge:

Large data generate spurious associations

Spurious correlations





A simulation study

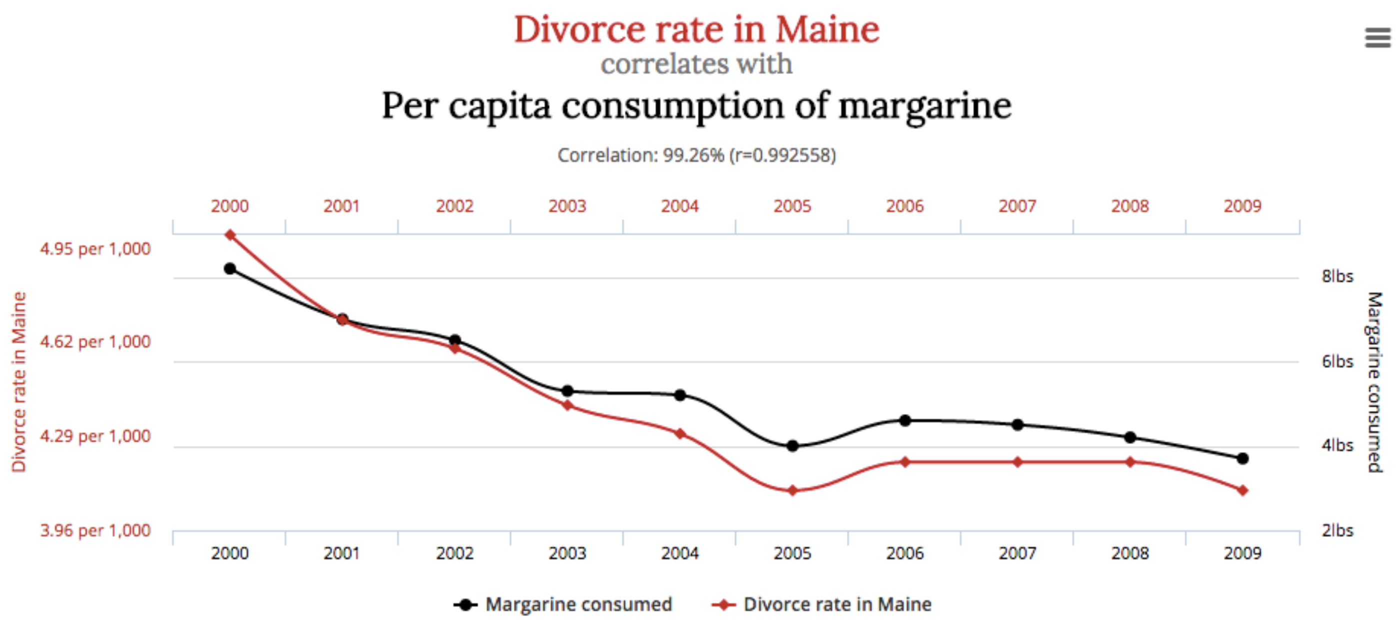
- Simulate $n = 60$ independent observations
- Each observation is in $d = 800, 6400$ dimensions
- Left: max absolute correlation between the first dimension and any other dimensions
- Right: max absolute correlation between the first dimension and a linear combination of any 4 other

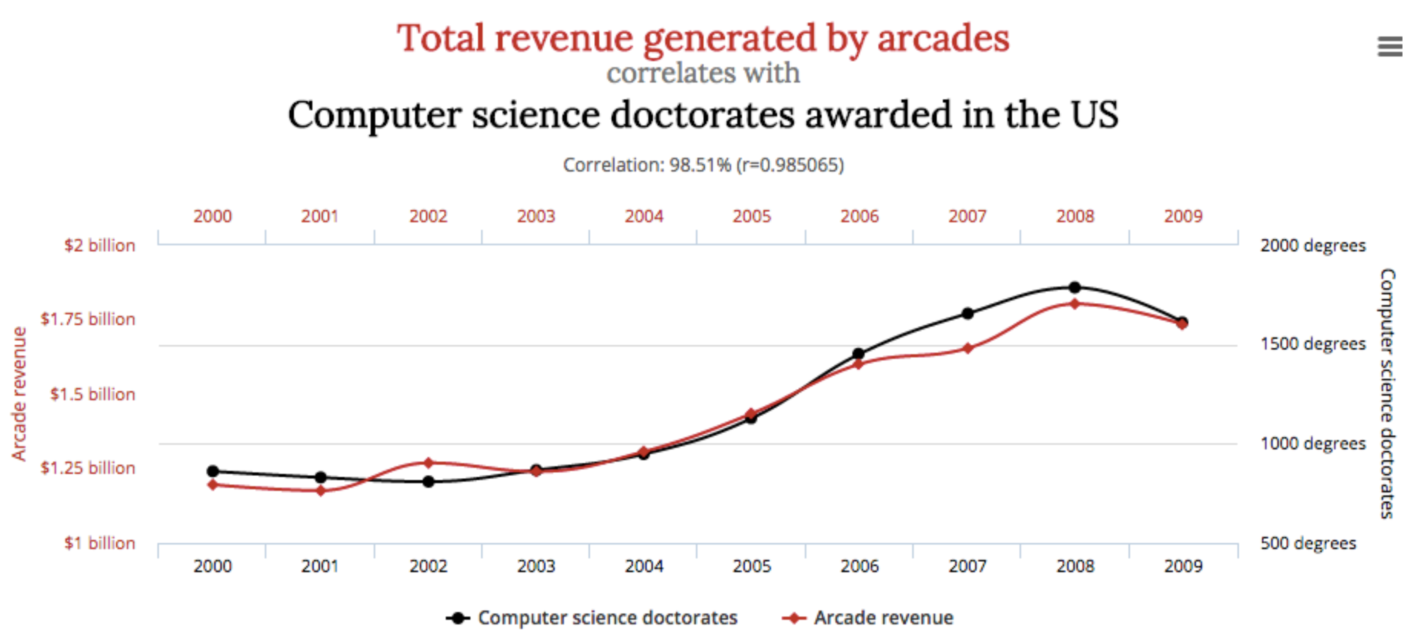
Conclusion: If we look hard enough,
we end up finding associations

Challenge:

Not to mistake a newly discovered association for causality

Example 1: Random





tylervigen.com/spurious-correlations

Example 2: Medicine (?)

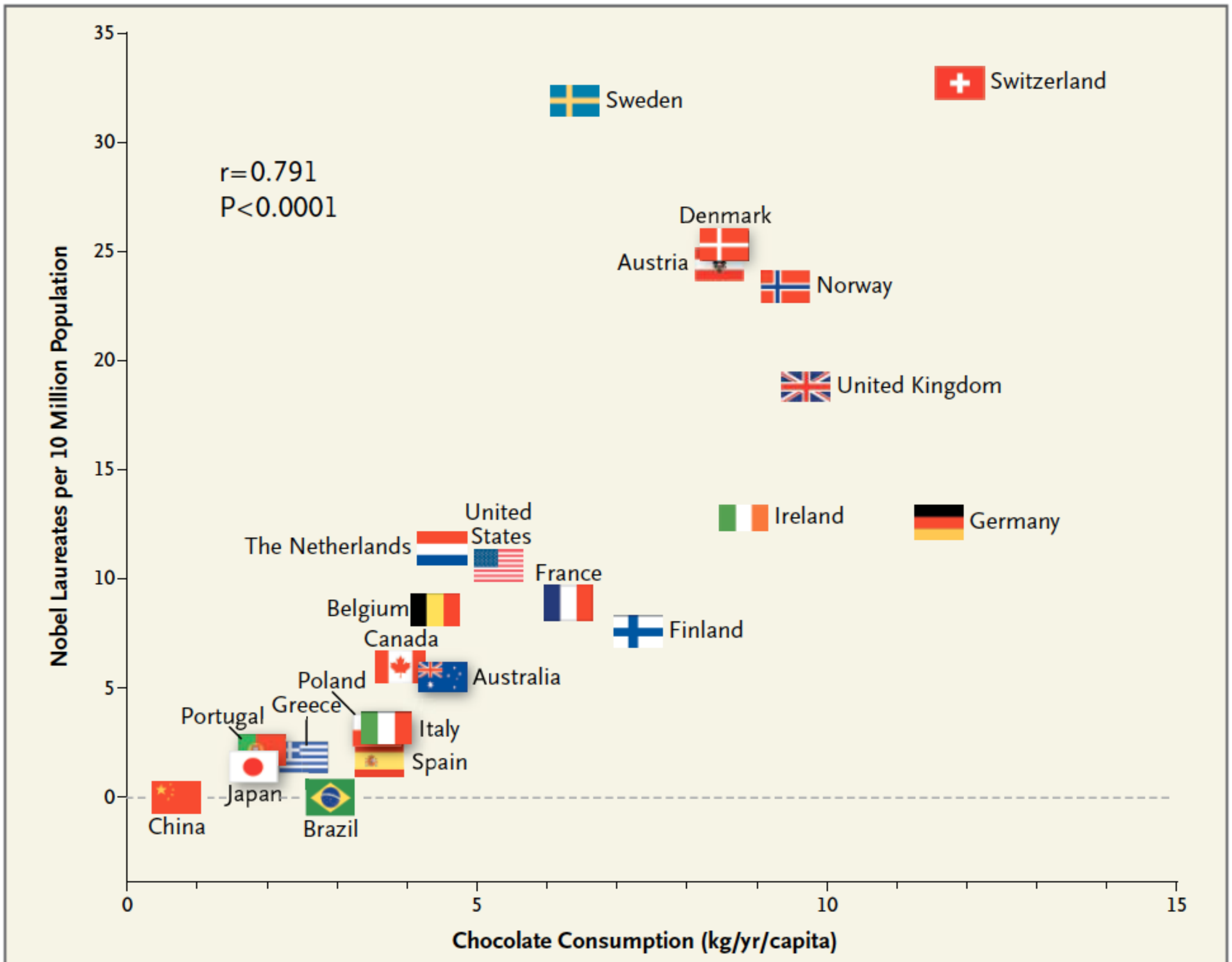


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

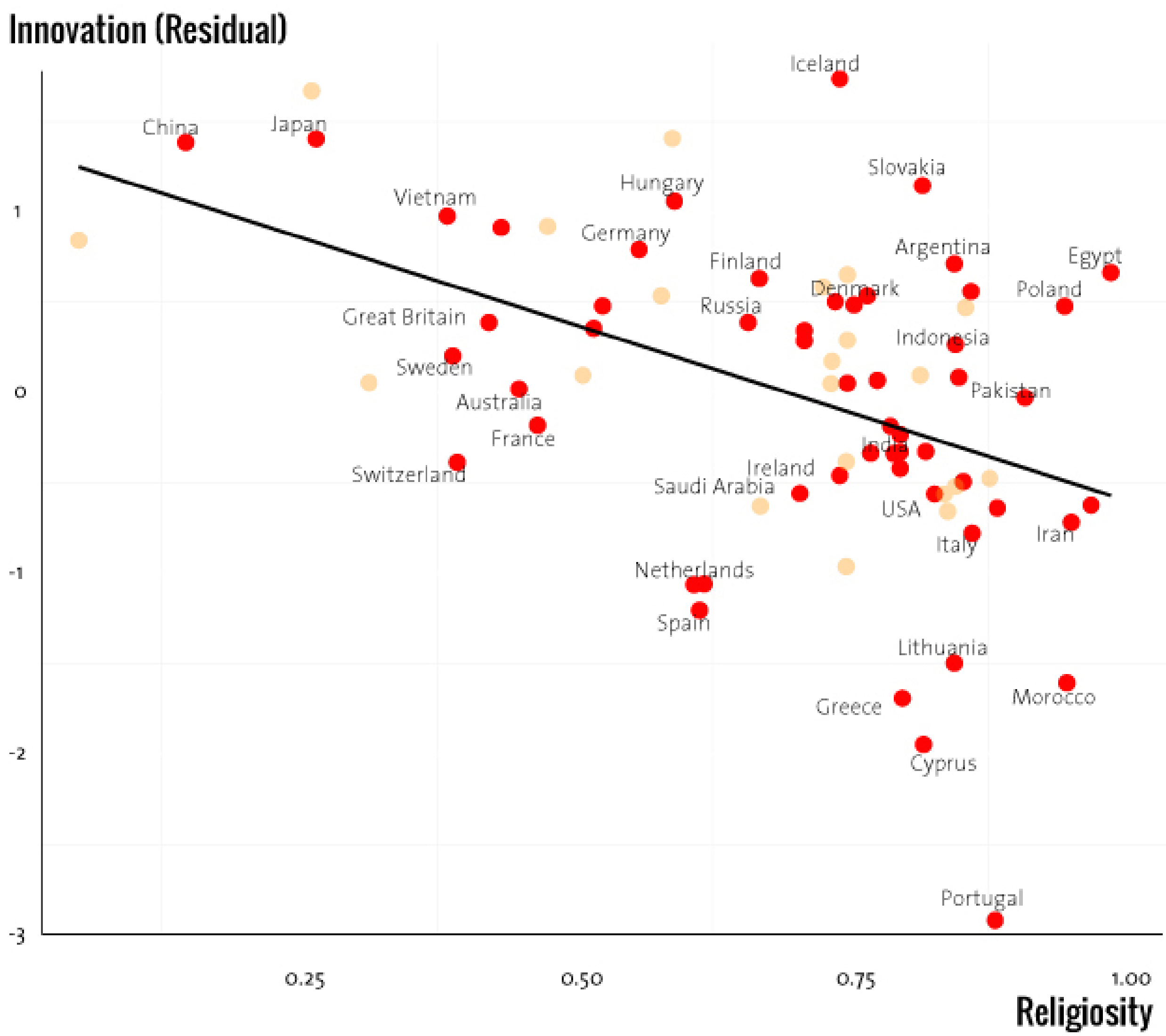
Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population

New England Journal of Medicine, 367:1562 (2012)

Example 2: Medicine (?)

- Premier journal of medical research
- Explains the association
  - Nobel prize is related to cognitive ability
  - Flavanols (organic molecules present in chocolate) are related to cognitive ability
- Technical flaws:
  - Nobel prize winners between 1900-2011
  - Chocolate consumption after 2002
  - Countries with many Nobel prizes have high Human Development Index (HDI) and high per capita income.
- **Conclusion:** The study is easy to dismiss, because we understand the context

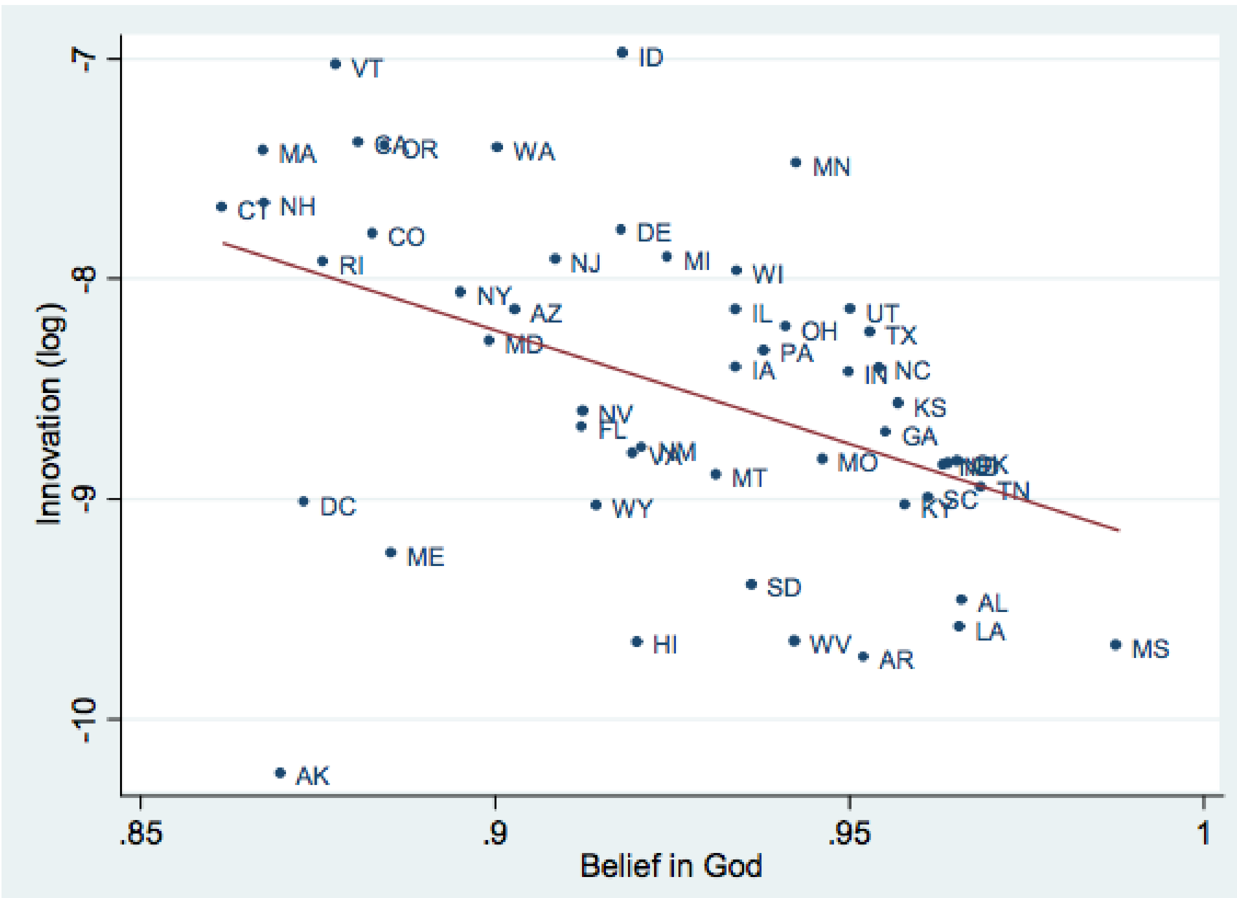
Example 3: Sociology?



After controlling for education and income levels, correlations between religion in the region, and number of patents per capita

Bénabou *et al.*, Princeton Univ., 2013

Example 3: Sociology?



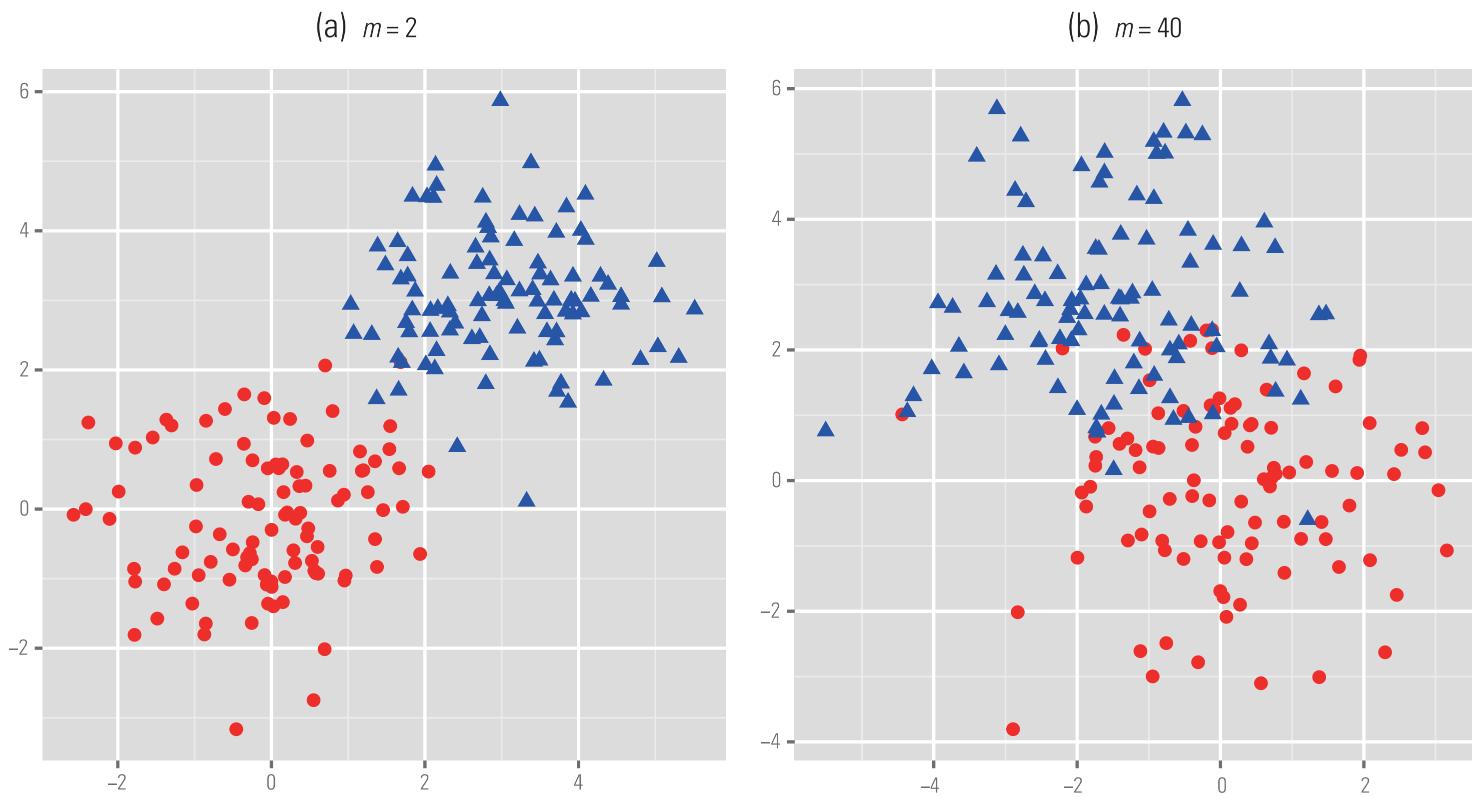
As before, for US states

'Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing ?look over there' - xkcd

Conclusion: the more complex the phenomenon, the more likely we are to mistake correlations for causality

Challenge:
Large data hide true quantitative signal

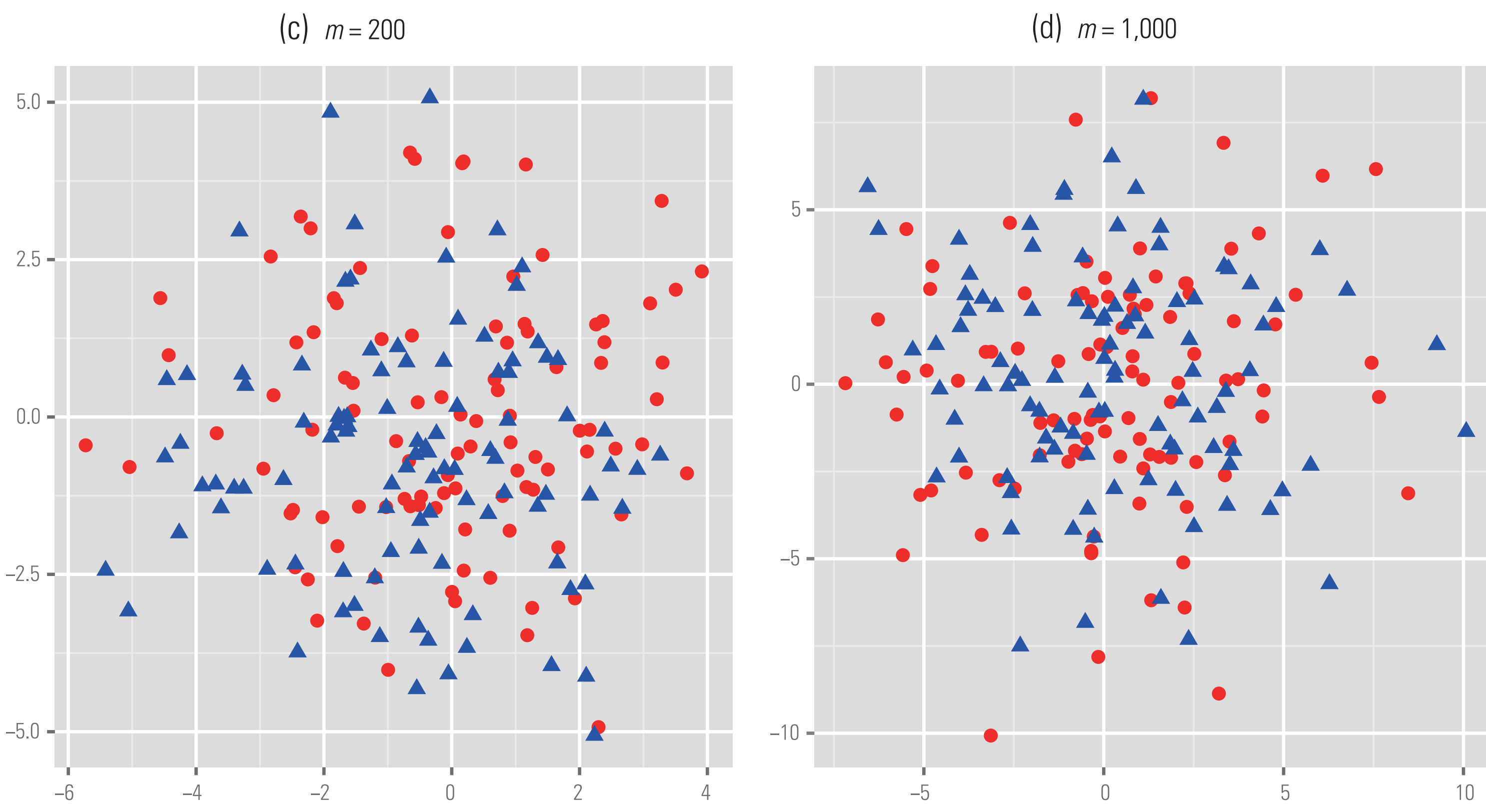
Noise accumulation



A simulation study

- Simulate $n = 100$ observations from 2 classes
- Each observation is a point in $m = 2, 40, 200, 1000$ dimensions
- Only first 10 dimensions are informative
- Plot first 2 principle components (i.e., eigenvectors)
- Informative data should show a good separation between the two classes

Noise accumulation

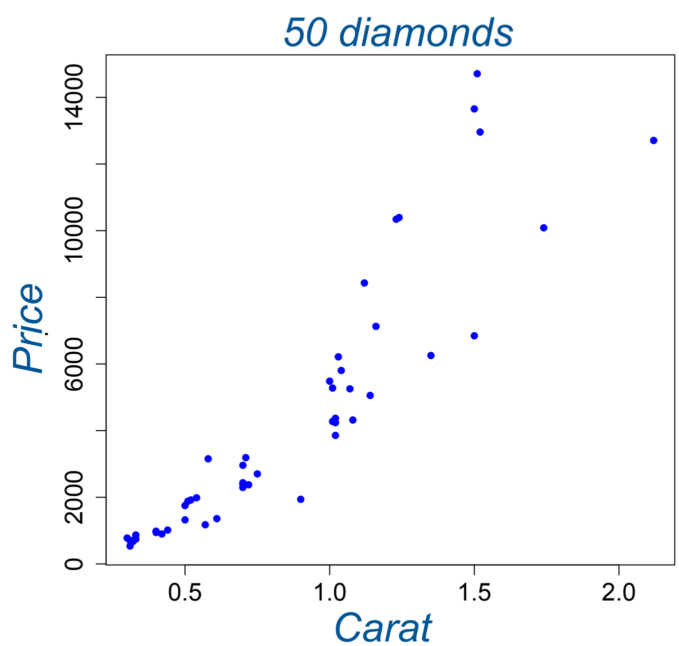


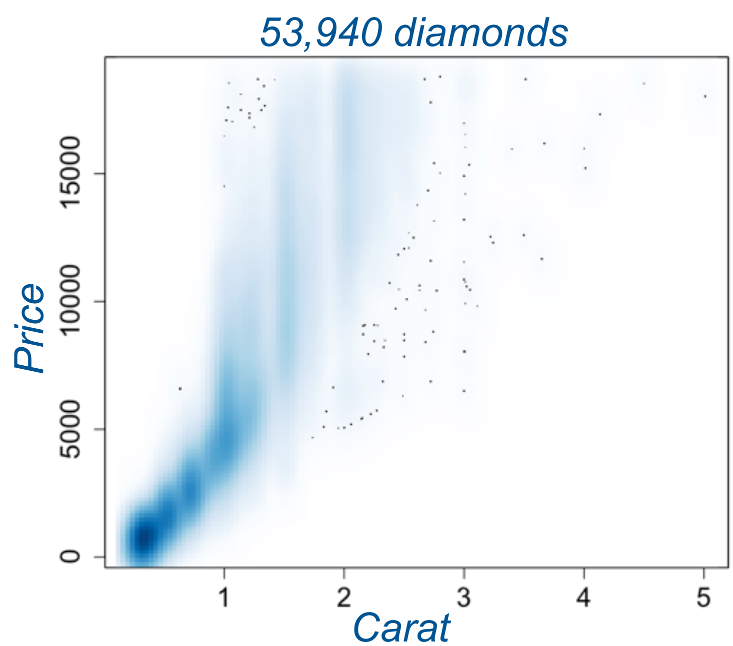
Conclusion: As we add new unrelated variables, we lose information

Challenge:
Large datasets amplify bias and confounding

Case study: Diamonds

```
> library(ggplot2); data(diamonds); head(diamonds)
carat color price
0.23  E     326
0.21  E     326
0.23  E     327
0.29  I     334
     ...............
```

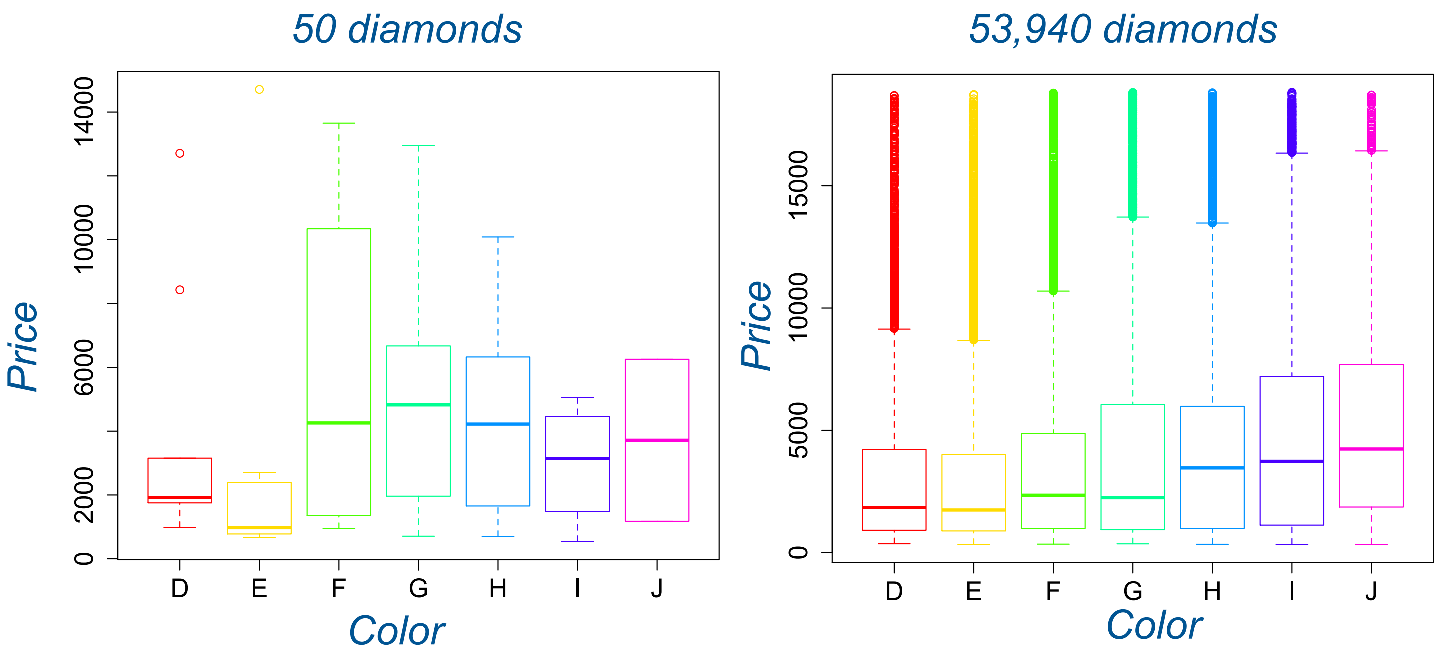




Large data amplifies true signal

- Heavier and pricier diamonds exist
- Large and expensive diamonds are rare
- Price increases exponentially with carat
- Not just a curve: also increase in variance

Case study: Diamonds



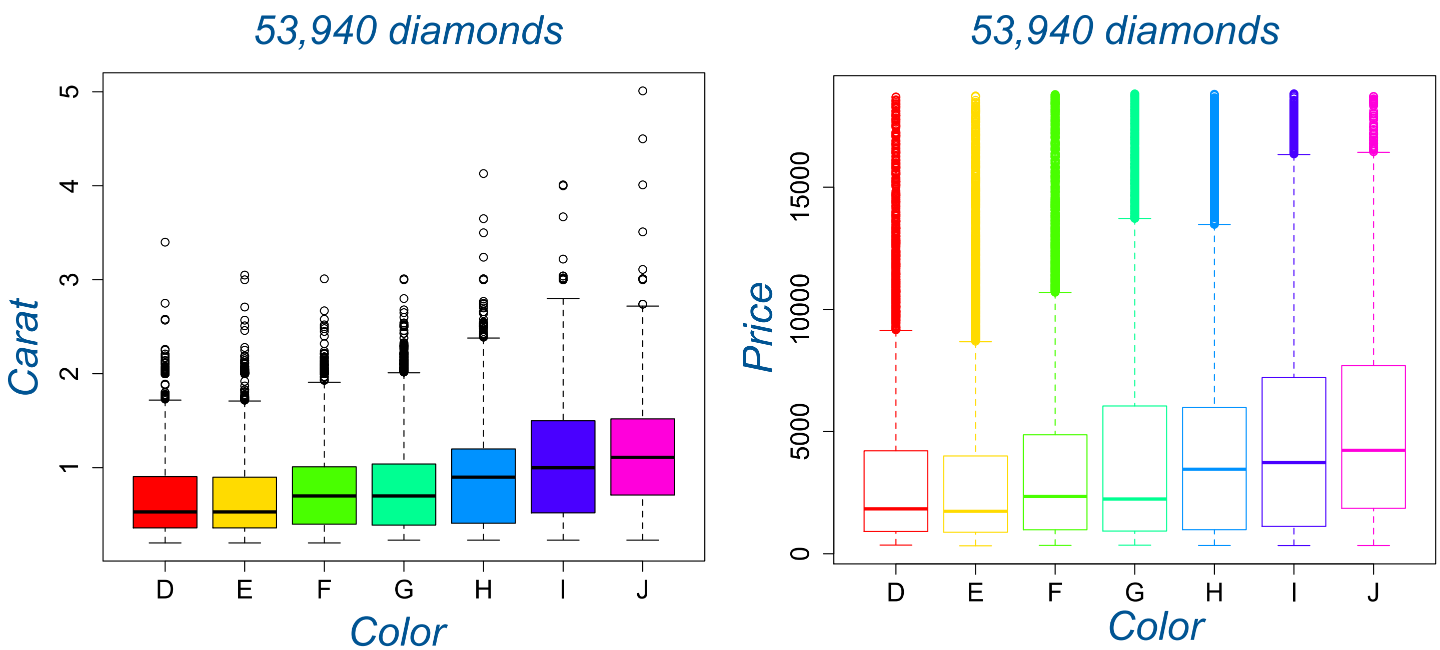
Large data amplifies wrong signal too

- 50 diamonds: no apparent trend in color
- The differences in price are consistent with variation
- All diamonds: discovered a new trend!
- Later colors are more expensive!

Contradiction

- Diamonds expert: later colors are cheaper

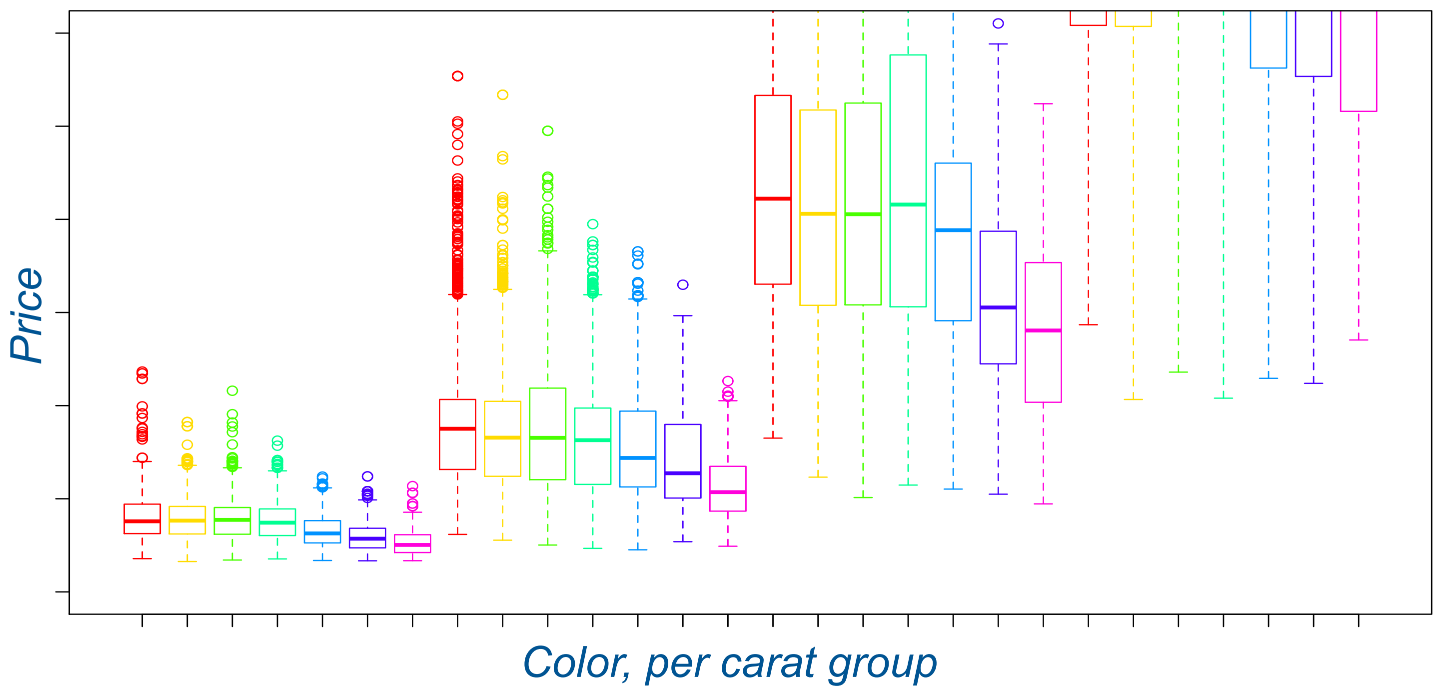
Case study: Diamonds



A closer look:

- Confounding between color and carat
- Later colors are heavier
- Carat is likely contributing to the price more than the color
- The right way to look at the problem is to stratify by range of carat

Case study: Diamonds



Conclusion: We would have ran away with a wrong discovery

- if we did not have the domain knowledge
- if we did not measure the right variables

Context and human insight is key

Summary of the challenges

- Due to observational nature
  (not designed / controlled experiments)
  - Confounding
    Diamonds dataset: color & carat confound price
  - Latent variables
    Google flu: seasons affect both flu & basketball
  - Heterogeneity
    Aggregating data from distinct subpopulations
    Google flu: changing searches and algos
- Due to high dimensionality
  - Noise accumulation
  - Spurious associations
- Poorly defined, unstructured problems

Statistical methods aim at distinguishing the artifacts from the systematic signals

J. Fan, F. Han, H. Liu, 'Challenges in big data analysis', National Science Review, 1:293, 2014