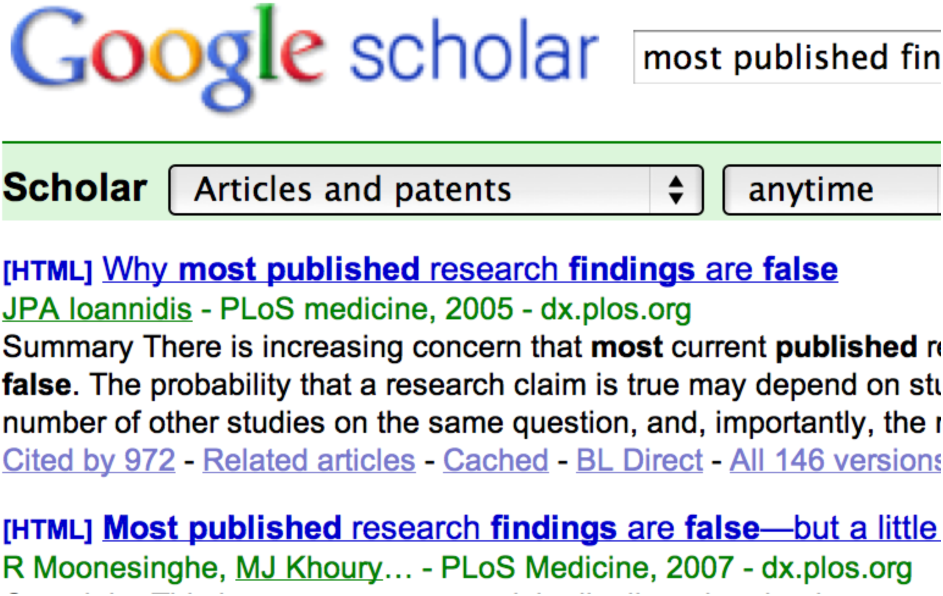
Little data: How traditional statistical ideas remain relevant in a big-data world

Andrew Gelman
Department of Statistics and Department of Political Science
Columbia University

10 Apr 2013



most published fin

[HTML] Why **most published** research **findings** are **false**

JPA Ioannidis - PLoS medicine, 2005 - dx.plos.org

Summary There is increasing concern that **most** current **published** r
false. The probability that a research claim is true may depend on stu
number of other studies on the same question, and, importantly, the r

Cited by 972 - Related articles - Cached - BL Direct - All 146 versions

[HTML] **Most published** research **findings** are **false**—but a little

R Moonesinghe, MJ Khoury… - PLoS Medicine, 2007 - dx.plos.org

Controversy in social psychology, medicine, . . .

- Fragile research findings
  - Joke research (Bem, Kanazawa, etc.)
  - Fraud, misconduct, and error (Hauser, Stapel, Anderson and Ones, etc.)
  - Systematic biases (selection, the statistical significance filter, etc.)
- Problems with the default model

Little data, big data

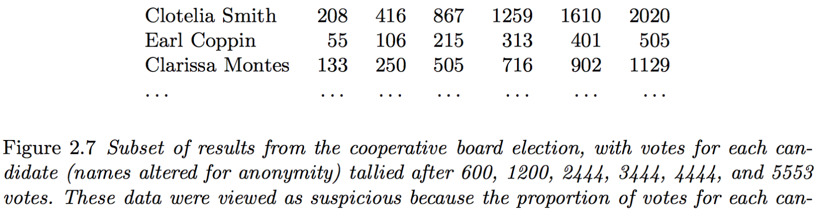
- Start with an example of traditional statistics with little data
- Then some big data

55,000 residents desperately need your help!

A fax arrives:

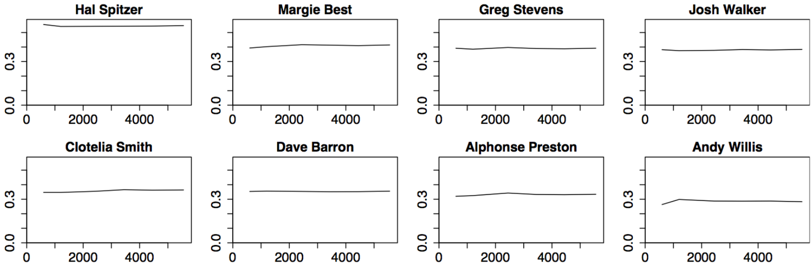
> *“Last week we had an election for the Board of Directors. Many residents believe, as I do, that the election was rigged . . . with fixed percentages being assigned to each and every candidate making it impossible to participate in an honest election. The unofficial election results I have faxed along with this letter represent the tallies. Tallies were given after 600 were counted. Then again at 1200, 2444, 3444, 4444, and final count at 5553. After close inspection we believe that there was nothing random about the count . . . Are we crazy? In a community this diverse and large, can candidates running on separate and opposite slates as well as independents receive similar vote percentage increases tally after tally . . . Does this appear random to you? What do you think? HELP!”*

A subset of the data

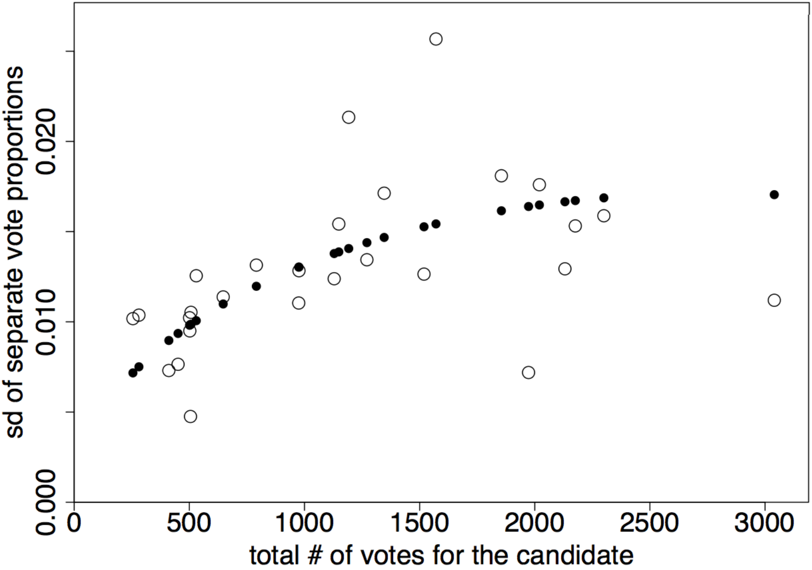


Clotelia Smith	208	416	867	1259	1610	2020
Earl Coppin	55	106	215	313	401	505
Clarissa Montes	133	250	505	716	902	1129
...	...	...	...	...	...	...

Figure 2.7 *Subset of results from the cooperative board election, with votes for each candidate (names altered for anonymity) tallied after 600, 1200, 2444, 3444, 4444, and 5553 votes. These data were viewed as suspicious because the proportion of votes for each can-*



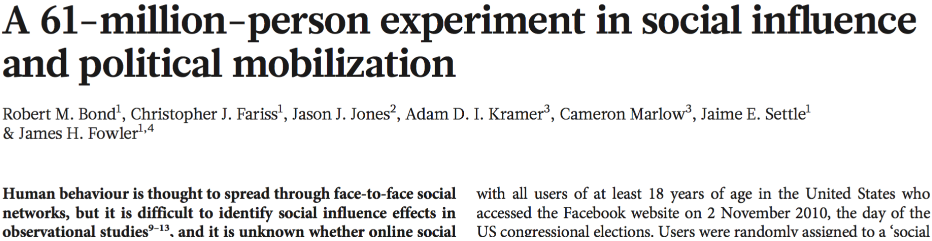
Comparing to random variation



Summary of election example

- The intermediate vote tallies are consistent with random voting
- Opinion polls of 1000 people are typically accurate to within 2%
- So, if voters really are arriving at random, it makes sense that batches of 1000 votes are highly stable

Big data!



A 61-million-person experiment in social influence and political mobilization

Robert M. Bond[1], Christopher J. Fariss[1], Jason J. Jones[2], Adam D. I. Kramer[3], Cameron Marlow[3], Jaime E. Settle[1] & James H. Fowler[1,4]

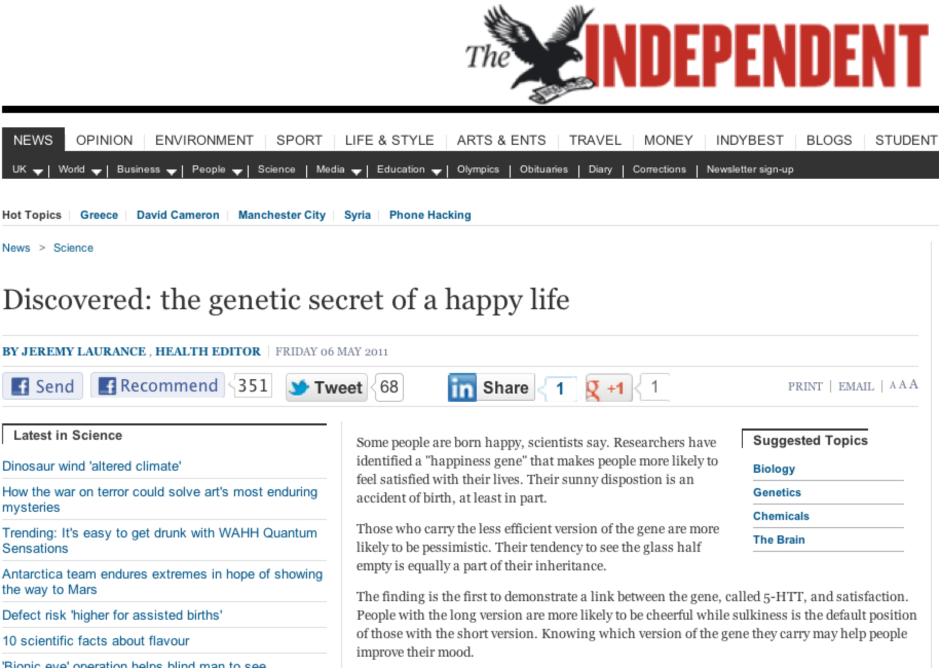
Human behaviour is thought to spread through face-to-face social networks, but it is difficult to identify social influence effects in observational studies[9–13], and it is unknown whether online social with all users of at least 18 years of age in the United States who accessed the Facebook website on 2 November 2010, the day of the US congressional elections. Users were randomly assigned to a 'social

- Facebook message directly increases voter turnout by 0.3%
  - Plausible small effect of innocuous advertisement
- Indirect (social) effect of 0.01%–0.1%
  - Lost in the noise—even if statistically significant

An example of research in behavior and genetics

“We expected reported anxiety to be significantly higher in the closeness condition compared to either of the other two treatments . . . There is no apparent main effect of the treatment . . .

The effects in columns 2 and 4 (the models without an interaction term) suggest that genetic risk scores for negative affectivity decrease the probability of turnout, although these effects do not reach conventional levels of significance for Genetic Risk Index 1. This provides some qualified support for our first hypothesis . . . The interaction terms in Table 3 are both negative: the interaction term with Genetic Risk Score 1 is significant at the $p < .05$ *level and that with Genetic Risk Score 2 is significant at the* $p < .10$ *level . . . This confirms our proposition . . . ”*


The INDEPENDENT
NEWS | OPINION | ENVIRONMENT | SPORT | LIFE & STYLE | ARTS & ENTS | TRAVEL | MONEY | INDYBEST | BLOGS | STUDENT
UK | World | Business | People | Science | Media | Education | Olympics | Obituaries | Diary | Corrections | Newsletter sign-up
Hot Topics | Greece | David Cameron | Manchester City | Syria | Phone Hacking
News > Science
Send
Recommend 351
Tweet 68
in Share 1
+1 1
PRINT | EMAIL | A A A

Discovered: the genetic secret of a happy life

BY JEREMY LAURANCE , HEALTH EDITOR | FRIDAY 06 MAY 2011

Some people are born happy, scientists say. Researchers have identified a "happiness gene" that makes people more likely to feel satisfied with their lives. Their sunny dispostion is an accident of birth, at least in part.

Those who carry the less efficient version of the gene are more likely to be pessimistic. Their tendency to see the glass half empty is equally a part of their inheritance.

The finding is the first to demonstrate a link between the gene, called 5-HTT, and satisfaction. People with the long version are more likely to be cheerful while sulkiness is the default position of those with the short version. Knowing which version of the gene they carry may help people improve their mood.

Latest in Science

- Dinosaur wind 'altered climate'
- How the war on terror could solve art's most enduring mysteries
- Trending: It's easy to get drunk with WAHH Quantum Sensations
- Antarctica team endures extremes in hope of showing the way to Mars
- Defect risk 'higher for assisted births'
- 10 scientific facts about flavour
- 'Bionic eye' operation helps blind man to see

Suggested Topics

- Biology
- Genetics
- Chemicals
- The Brain

“Discovered: the genetic secret of a happy life”

From the news article:

> *“Researchers have identified a ‘happiness gene’ that makes people more likely to feel satisfied with their lives . . . The finding is the first to demonstrate a link between the gene, called 5-HTT, and satisfaction . . . Those with two long versions of the gene were 17 per cent more likely to say they were very satisfied. . . . ”*

From the research article by De Neve, Fowler, and Frey:

> *“Having one or two alleles . . . raises the average likelihood of being very satisfied with ones life by 8.5% and 17.3%, respectively.”*

Undiscovered . . .

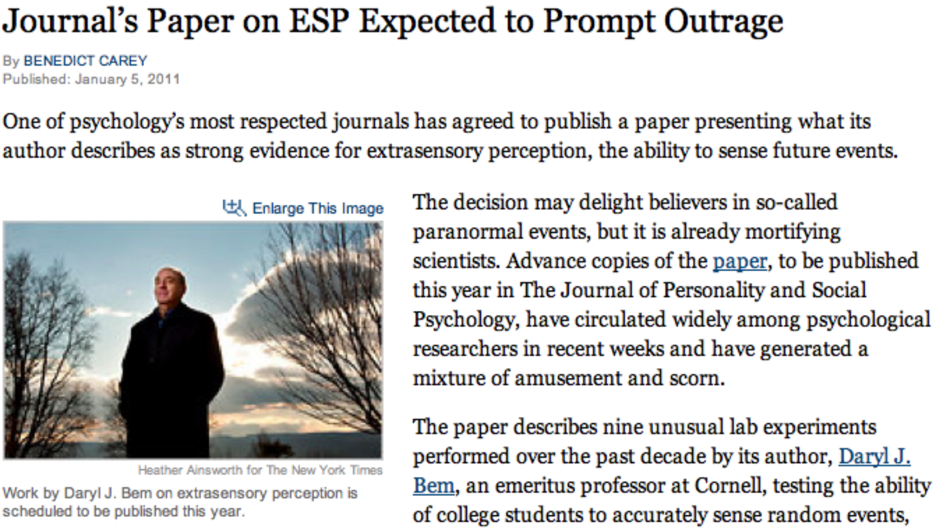
From the *text* of the research article:

> *"Having one or two alleles . . . raises the average likelihood of being very satisfied with ones life by 8.5% and 17.3%, respectively."*

From the *tables*:

- 46% of people who had two copies of the gene described themselves as satisfied and 41% described themselves as very satisfied. The corresponding percentages for those with no copies were 44% and 37%.
- Reported maximum difference is 4 percentage points (and not statistically significant), *not* 17%.

Statistical significance is not enough!



Journal's Paper on ESP Expected to Prompt Outrage

By BENEDICT CAREY
Published: January 5, 2011

One of psychology's most respected journals has agreed to publish a paper presenting what its author describes as strong evidence for extrasensory perception, the ability to sense future events.

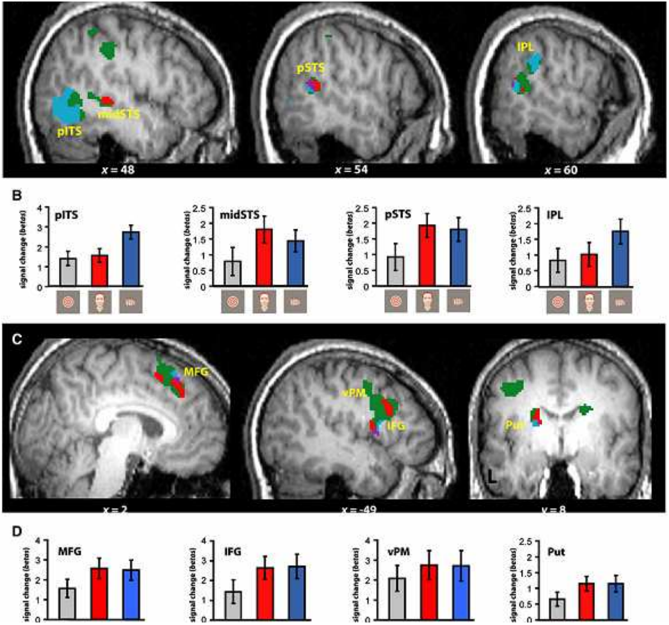
Heather Ainsworth for The New York Times

Work by Daryl J. Bem on extrasensory perception is scheduled to be published this year.

The decision may delight believers in so-called paranormal events, but it is already mortifying scientists. Advance copies of the paper, to be published this year in The Journal of Personality and Social Psychology, have circulated widely among psychological researchers in recent weeks and have generated a mixture of amusement and scorn.

The paper describes nine unusual lab experiments performed over the past decade by its author, Daryl J. Bem, an emeritus professor at Cornell, testing the ability of college students to accurately sense random events,

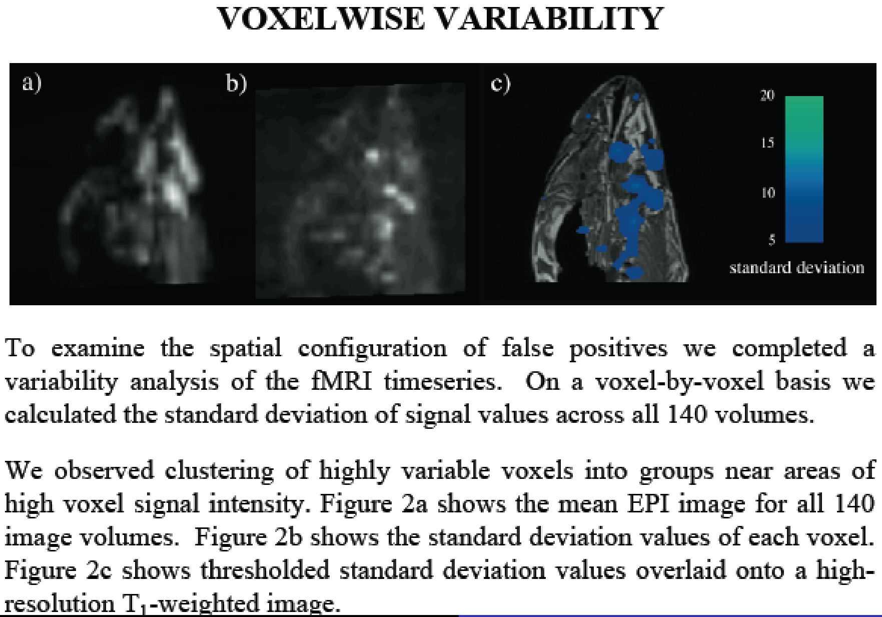
Finding statistical significance with big data



The “statistical significance filter”

- Vul, Harris, Winkelman, Pashler:
  - Correlations reported in medical imaging studies are commonly overstated because researchers select the highest values
  - These *statistical* problems are leading to *scientific* errors
- Statistical corrections for multiple comparisons do *not* solve the problem
- Discussion in *Perspectives in Psychological Science* (2009)

Neural activity in a dead fish

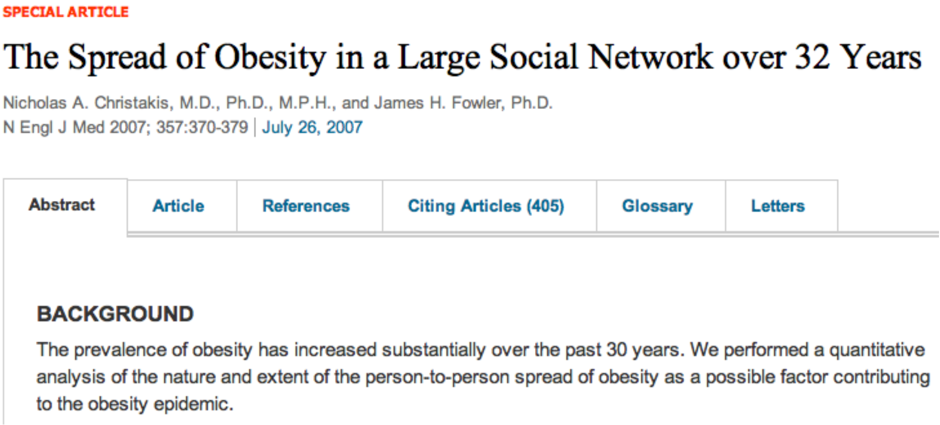


VOXELWISE VARIABILITY

To examine the spatial configuration of false positives we completed a variability analysis of the fMRI timeseries. On a voxel-by-voxel basis we calculated the standard deviation of signal values across all 140 volumes.

We observed clustering of highly variable voxels into groups near areas of high voxel signal intensity. Figure 2a shows the mean EPI image for all 140 image volumes. Figure 2b shows the standard deviation values of each voxel. Figure 2c shows thresholded standard deviation values overlaid onto a high-resolution T_1-weighted image.

Sampling is important, even with big data



SPECIAL ARTICLE

The Spread of Obesity in a Large Social Network over 32 Years

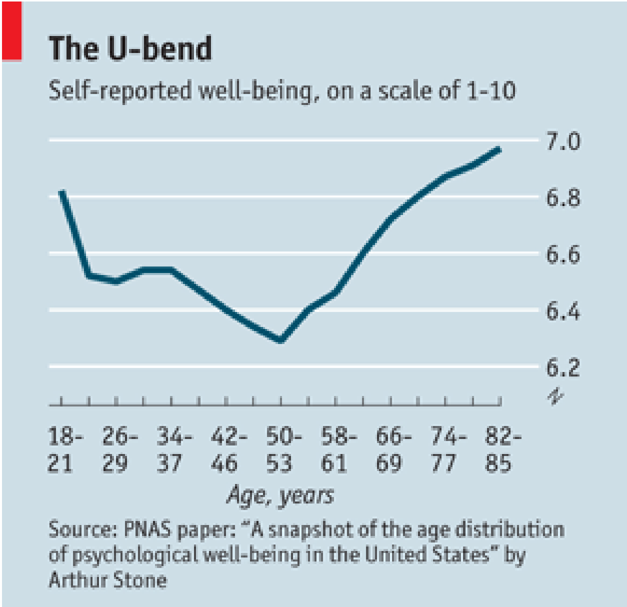
Nicholas A. Christakis, M.D., Ph.D., M.P.H., and James H. Fowler, Ph.D.

N Engl J Med 2007; 357:370-379 | July 26, 2007

BACKGROUND

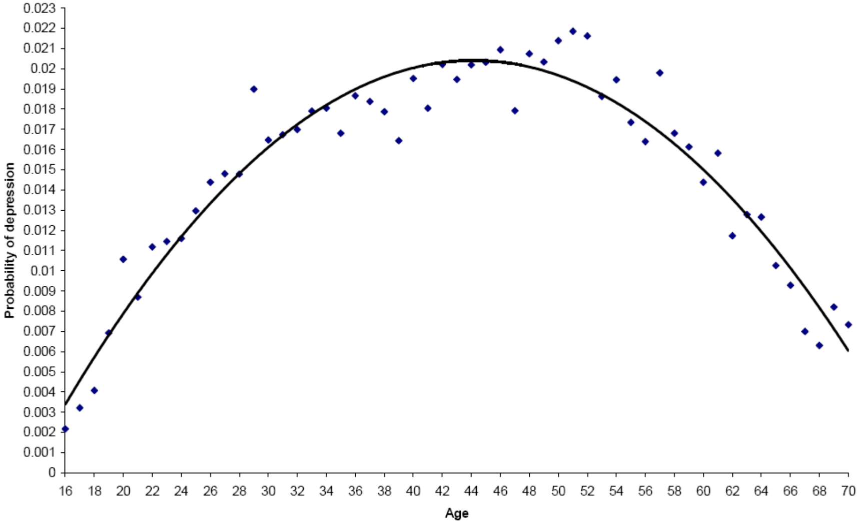
The prevalence of obesity has increased substantially over the past 30 years. We performed a quantitative analysis of the nature and extent of the person-to-person spread of obesity as a possible factor contributing to the obesity epidemic.

Happiness and life satisfaction

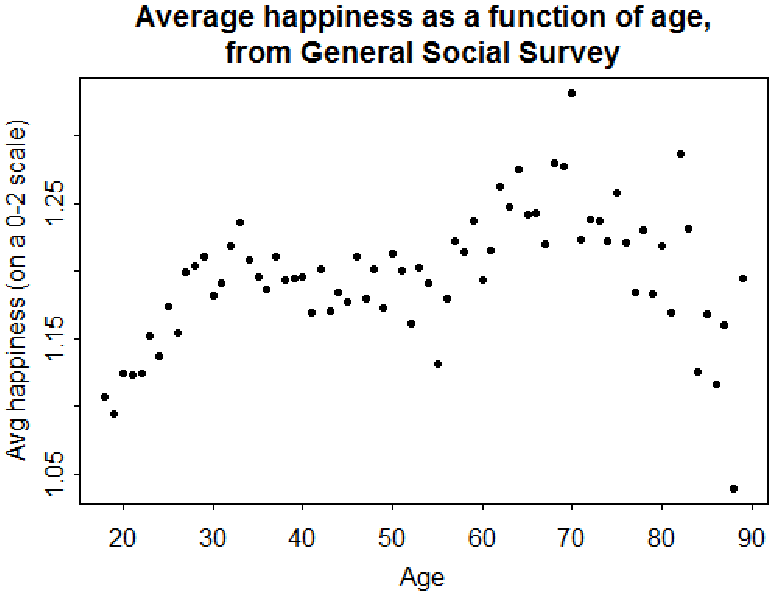


Source: PNAS paper: "A snapshot of the age distribution of psychological well-being in the United States" by Arthur Stone

Data!


Probability of depression
0.023
0.022
0.021
0.02
0.019
0.018
0.017
0.016
0.015
0.014
0.013
0.012
0.011
0.01
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
50
52
54
56
58
60
62
64
66
68
70
Age

More data


Average happiness as a function of age,
from General Social Survey
Avg happiness (on a 0-2 scale)
1.05
1.15
1.25
20
30
40
50
60
70
80
90
Age

The Perils of Pooling

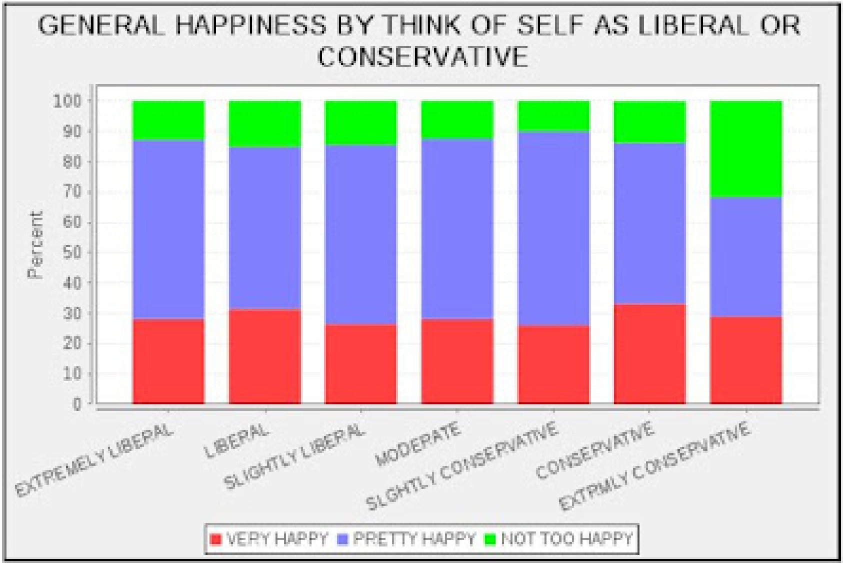
Arthur "not David" Brooks in the *New York Times*:

> *"People at the extremes are happier than political moderates. . . . none, it seems, are happier than the Tea Partiers . . . "*

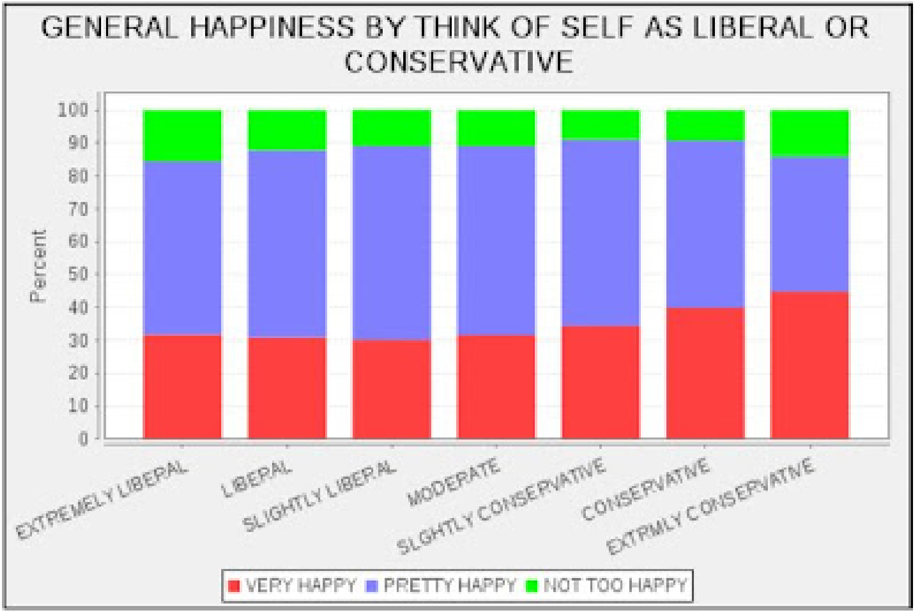
Jay Livingston (sociology, Montclair State University) looks up the data in the General Social Survey . . .

“None, it seems, are happier than the Tea Partiers . . .” ??

Chart for YEAR = 4(2009-2010)

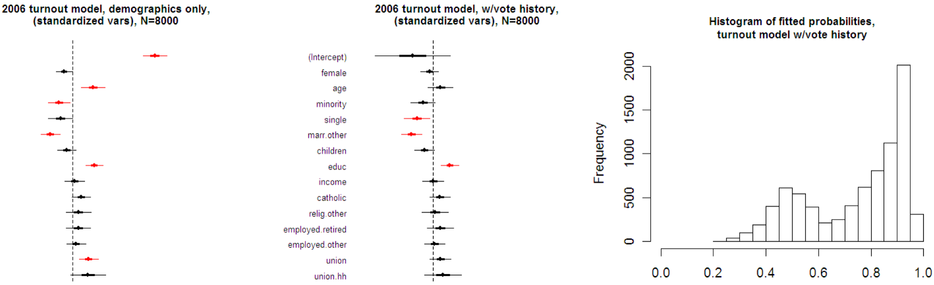


Pooling, 1972–2010



When to aggregate or break up the data?

- Always always always a concern
- A “big data” example:



2006 turnout model, demographics only, (standardized vars), N=8000

2006 turnout model, w/vote history, (standardized vars), N=8000

Histogram of fitted probabilities, turnout model w/vote history

More happiness

From *USA Today*:

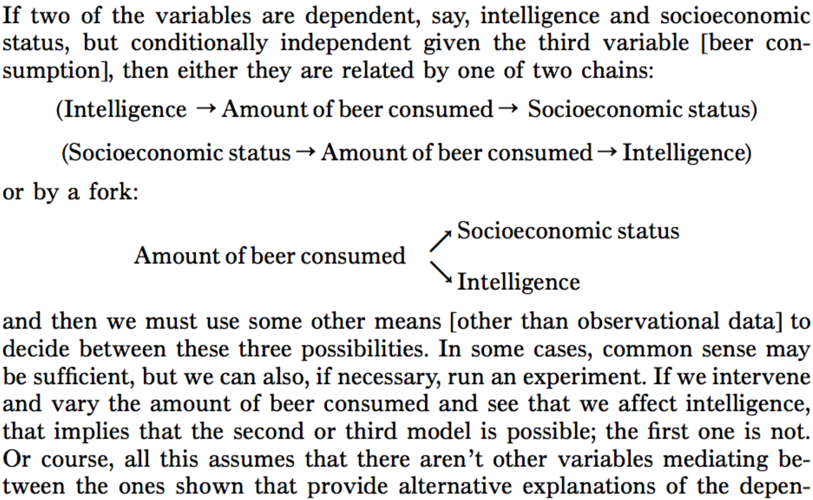
> *"Conventional wisdom . . . has said parents are less happy, more depressed and have less-satisfying marriages than their childless counterparts. But . . . newer analyses . . . based on data from almost 130,000 adults around the globe . . . say that parents today may indeed be happier than non-parents . . . The other study, of some 120,000 adults . . . finds that parents were indeed less happy than non-parents in the decade 1985-95, but parents from 1995 to 2008 were happier . . ."*

From an author of the cited research papers:

> *"We find that globally, happiness decreases with the number of children . . . the association between happiness and fertility evolves from negative to neutral to positive above age 40 . . . The first child increases happiness quite a lot. The second child a little. The third not at all."*

Challenges of causal reasoning are not going away

From a recent book by a cognitive scientist:



If two of the variables are dependent, say, intelligence and socioeconomic status, but conditionally independent given the third variable [beer consumption], then either they are related by one of two chains:

(Intelligence → Amount of beer consumed → Socioeconomic status)

(Socioeconomic status → Amount of beer consumed → Intelligence)

or by a fork:

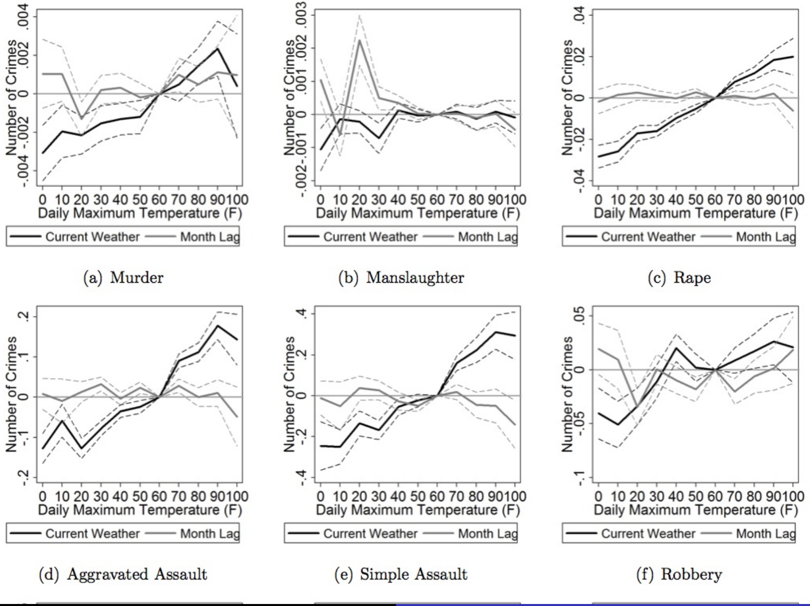
Amount of beer consumed ↗ Socioeconomic status
↘ Intelligence

and then we must use some other means [other than observational data] to decide between these three possibilities. In some cases, common sense may be sufficient, but we can also, if necessary, run an experiment. If we intervene and vary the amount of beer consumed and see that we affect intelligence, that implies that the second or third model is possible; the first one is not. Or course, all this assumes that there aren't other variables mediating between the ones shown that provide alternative explanations of the depen-

The problem understanding the world using “stylized facts”

- Problems with is-it-there-or-is-it-not models of correlations and effects
- Problems with the concept of “false positives”
- Accepting variation (as distinct from measurement error)
- Don’t fool yourself!

Another example: Hot weather and crime



(a) Murder

(b) Manslaughter

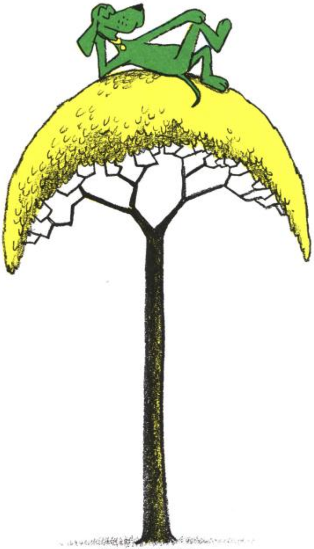
(c) Rape

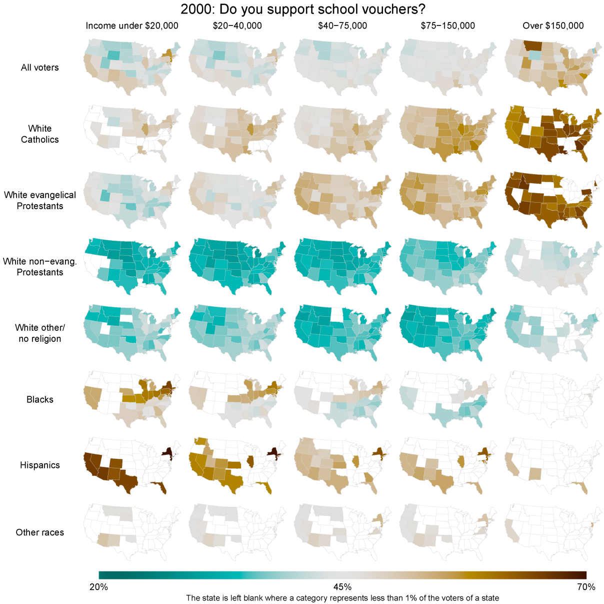
(d) Aggravated Assault

(e) Simple Assault

(f) Robbery

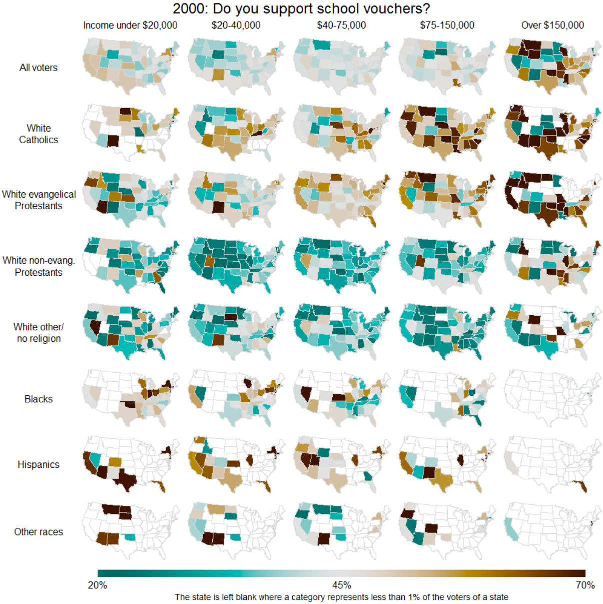
Big Data and Big Model





The state is left blank where a category represents less than 1% of the voters of a state

Data don't always "speak for themselves"


2000: Do you support school vouchers?
Income under $20,000
$20-40,000
$40-75,000
$75-150,000
Over $150,000
All voters
White Catholics
White evangelical Protestants
White non-evang. Protestants
White other/ no religion
Blacks
Hispanics
Other races
20%
45%
70%
The state is left blank where a category represents less than 1% of the voters of a state

Conclusions

- My goal is not to “debunk”
- The central question of traditional, “little-data” statistics:
    - Here’s a pattern in the data
    - Is it “real”? That is . . .
    - Is something similar going on in the general population?
- Ideas of statistical sampling are central to science