youth
delegate
programme
the idhouse

Who’s Afraid of the Big (Bad) Data?

An Introduction to Statistics

Jean-Christophe Spiliotis

August 13, 2018

University of Oxford

Table of contents

1. What is data?

2. A Data Hiker’s Guide

3. Letting the data speak

4. Making the data sing

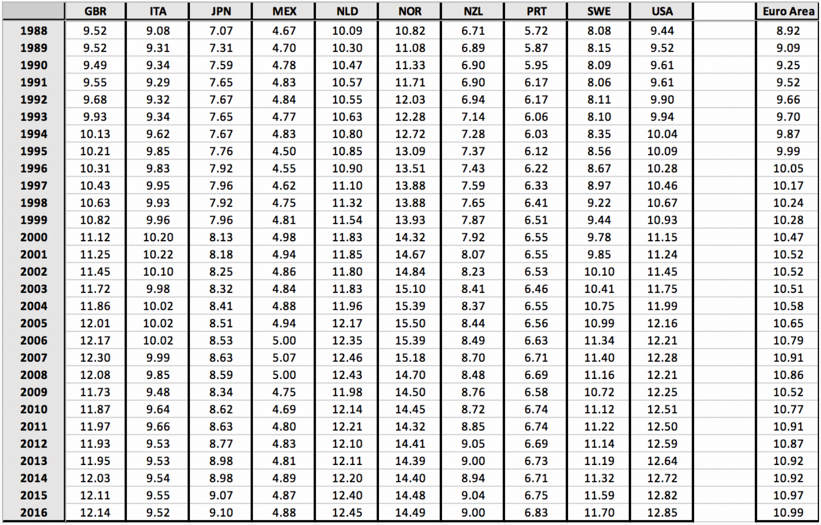
5. Don’t fall for the sirens’ song!

6. Where is data?

What is data?

What is data?

Spreadsheets, but not only...



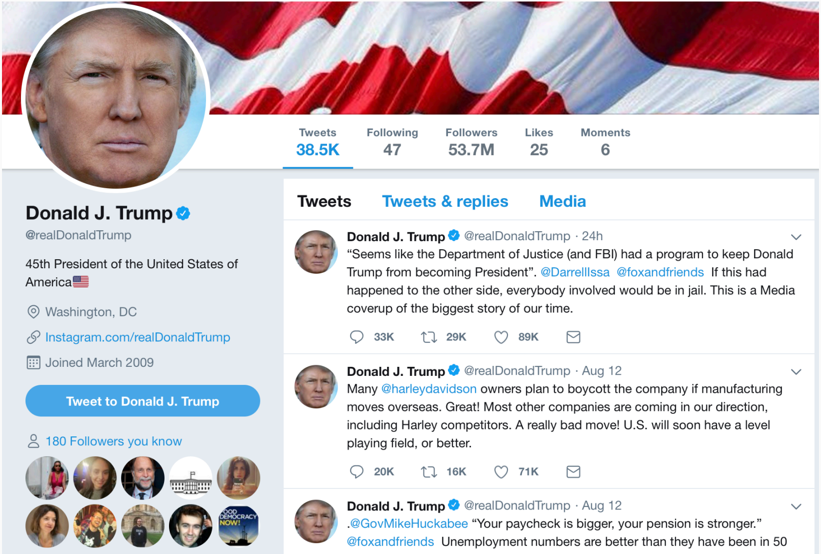
	GBR	ITA	JPN	MEX	NLD	NOR	NZL	PRT	SWE	USA		Euro Area
1988	9.52	9.08	7.07	4.67	10.09	10.82	6.71	5.72	8.08	9.44		8.92
1989	9.52	9.31	7.31	4.70	10.30	11.08	6.89	5.87	8.15	9.52		9.09
1990	9.49	9.34	7.59	4.78	10.47	11.33	6.90	5.95	8.09	9.61		9.25
1991	9.55	9.29	7.65	4.83	10.57	11.71	6.90	6.17	8.06	9.61		9.52
1992	9.68	9.32	7.67	4.84	10.55	12.03	6.94	6.17	8.11	9.90		9.66
1993	9.93	9.34	7.65	4.77	10.63	12.28	7.14	6.06	8.10	9.94		9.70
1994	10.13	9.62	7.67	4.83	10.80	12.72	7.28	6.03	8.35	10.04		9.87
1995	10.21	9.85	7.76	4.50	10.85	13.09	7.37	6.12	8.56	10.09		9.99
1996	10.31	9.83	7.92	4.55	10.90	13.51	7.43	6.22	8.67	10.28		10.05
1997	10.43	9.95	7.96	4.62	11.10	13.88	7.59	6.33	8.97	10.46		10.17
1998	10.63	9.93	7.92	4.75	11.32	13.88	7.65	6.41	9.22	10.67		10.24
1999	10.82	9.96	7.96	4.81	11.54	13.93	7.87	6.51	9.44	10.93		10.28
2000	11.12	10.20	8.13	4.98	11.83	14.32	7.92	6.55	9.78	11.15		10.47
2001	11.25	10.22	8.18	4.94	11.85	14.67	8.07	6.55	9.85	11.24		10.52
2002	11.45	10.10	8.25	4.86	11.80	14.84	8.23	6.53	10.10	11.45		10.52
2003	11.72	9.98	8.32	4.84	11.83	15.10	8.41	6.46	10.41	11.75		10.51
2004	11.86	10.02	8.41	4.88	11.96	15.39	8.37	6.55	10.75	11.99		10.58
2005	12.01	10.02	8.51	4.94	12.17	15.50	8.44	6.56	10.99	12.16		10.65
2006	12.17	10.02	8.53	5.00	12.35	15.39	8.49	6.63	11.34	12.21		10.79
2007	12.30	9.99	8.63	5.07	12.46	15.18	8.70	6.71	11.40	12.28		10.91
2008	12.08	9.85	8.59	5.00	12.43	14.70	8.48	6.69	11.16	12.21		10.86
2009	11.73	9.48	8.34	4.75	11.98	14.50	8.76	6.58	10.72	12.25		10.52
2010	11.87	9.64	8.62	4.69	12.14	14.45	8.72	6.74	11.12	12.51		10.77
2011	11.97	9.66	8.63	4.80	12.21	14.32	8.85	6.74	11.22	12.50		10.91
2012	11.93	9.53	8.77	4.83	12.10	14.41	9.05	6.69	11.14	12.59		10.87
2013	11.95	9.53	8.98	4.81	12.11	14.39	9.00	6.73	11.19	12.64		10.92
2014	12.03	9.54	8.98	4.89	12.20	14.40	8.94	6.71	11.32	12.72		10.92
2015	12.11	9.55	9.07	4.87	12.40	14.48	9.04	6.75	11.59	12.82		10.97
2016	12.14	9.52	9.10	4.88	12.45	14.49	9.00	6.83	11.70	12.85		10.99

What is data?

Text (nonfiction)

What is data?

Social media content



Tweets	Following	Followers	Likes	Moments
38.5K	47	53.7M	25	6

Donald J. Trump
@realDonaldTrump

45th President of the United States of America🇺🇸

Washington, DC
Instagram.com/realDonaldTrump
Joined March 2009

180 Followers you know

Donald J. Trump @realDonaldTrump · 24h
"Seems like the Department of Justice (and FBI) had a program to keep Donald Trump from becoming President". @DarrellIssa @foxandfriends If this had happened to the other side, everybody involved would be in jail. This is a Media coverup of the biggest story of our time.

33K 29K 89K

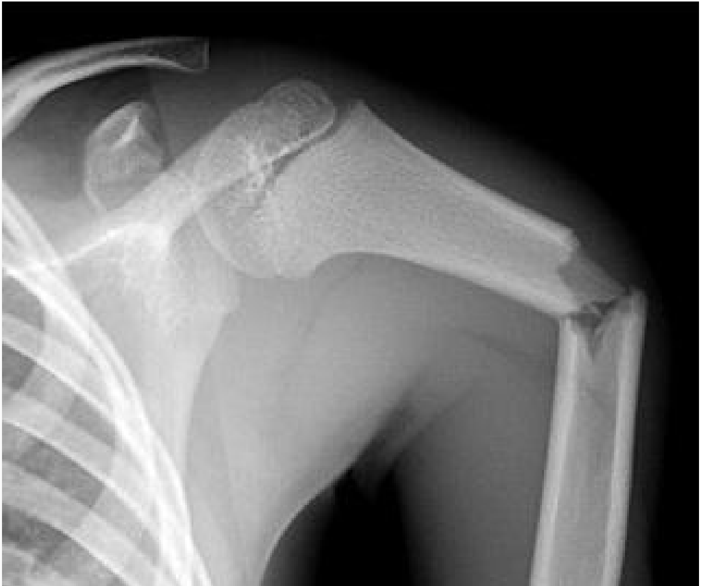
Donald J. Trump @realDonaldTrump · Aug 12
Many @harleydavidson owners plan to boycott the company if manufacturing moves overseas. Great! Most other companies are coming in our direction, including Harley competitors. A really bad move! U.S. will soon have a level playing field, or better.

20K 16K 71K

Donald J. Trump @realDonaldTrump · Aug 12
.@GovMikeHuckabee "Your paycheck is bigger, your pension is stronger." @foxandfriends Unemployment numbers are better than they have been in 50

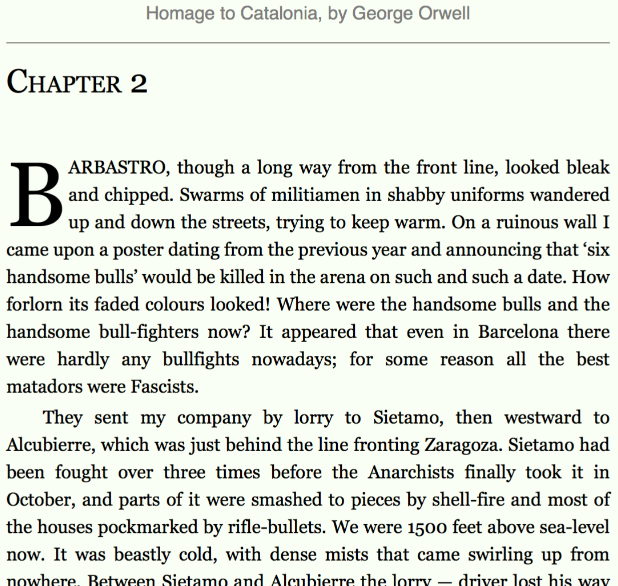
What is data?

Pictures



What is data?

Text (fiction)



Homage to Catalonia, by George Orwell

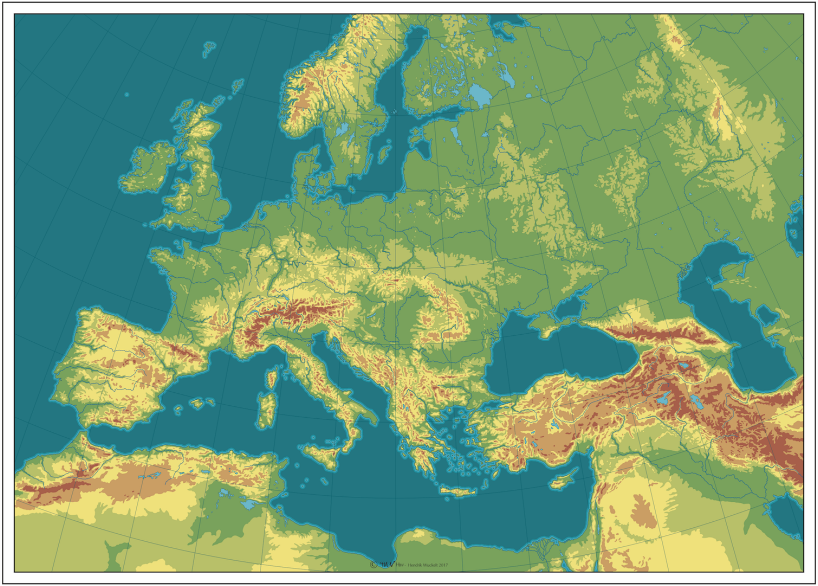
Chapter 2

BARBASTRO, though a long way from the front line, looked bleak and chipped. Swarms of militiamen in shabby uniforms wandered up and down the streets, trying to keep warm. On a ruinous wall I came upon a poster dating from the previous year and announcing that 'six handsome bulls' would be killed in the arena on such and such a date. How forlorn its faded colours looked! Where were the handsome bulls and the handsome bull-fighters now? It appeared that even in Barcelona there were hardly any bullfights nowadays; for some reason all the best matadors were Fascists.

They sent my company by lorry to Sietamo, then westward to Alcubierre, which was just behind the line fronting Zaragoza. Sietamo had been fought over three times before the Anarchists finally took it in October, and parts of it were smashed to pieces by shell-fire and most of the houses pockmarked by rifle-bullets. We were 1500 feet above sea-level now. It was beastly cold, with dense mists that came swirling up from nowhere. Between Sietamo and Alcubierre the lorry — driver lost his way

What is data?

Maps



© Hendrik Wuckelt 2017

Qualitative and quantitative data

Name	Eye color	Age	Gender
Anna	Brown	54	Female
Bob	Blue	17	Male
Charlie	Blue	16	Female
Charlie	Brown	25	Unknown

Code	Blue eyes	Brown eyes	Age	18 or above	Below 18	Gender
1	0	1	54	1	0	1
2	1	0	17	0	1	0
3	1	0	16	0	1	1
4	0	1	25	1	0	.

A Data Hiker’s Guide

The skill set

☐ A PhD
☑ A bit of logic

☐ Light-speed mental math skills
☑ Getting comfortable with numbers, but most importantly *concepts*

☐ A history of hacking into the Pentagon
☑ Willingness to spend hours staring at a computer screen, looking for a missing bracket

The tool set

1. A statistical software you’re comfortable with
    - User interface
    - Programming language
    - Price
    - Online resources and community
    - Performance
2. A textbook
3. Some contextual knowledge of the question at hand
4. A research buddy
    - To discuss procedures (data analysis is as much rhetoric as it is mathematics!)
    - To proofread outputs and make sure the message gets conveyed
5. A well formulated question…

Letting the data speak

Why visualize?

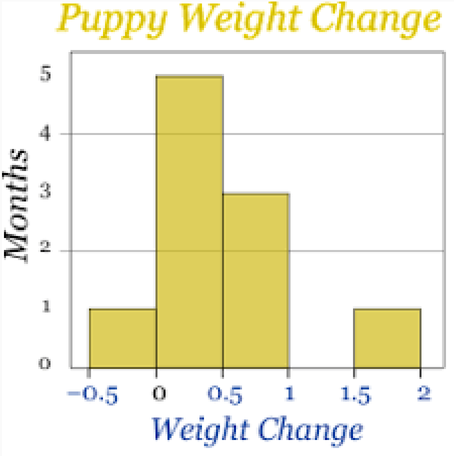
- Detecting patterns
- Finding anomalies:

Name	Eye color	Age	Gender
Anna	Brown	54	Female
Bob	Blue	170	Male
Charlie	Blue	16	Femmale
Charlie	Brown	25	Unknown

- Guiding further analysis
- Telling a story and communicating effectively to the general public

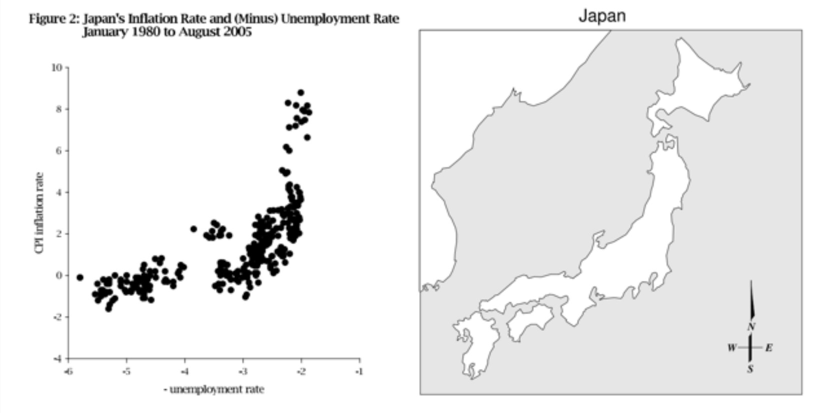
How to visualize?

Histograms



How to visualize?

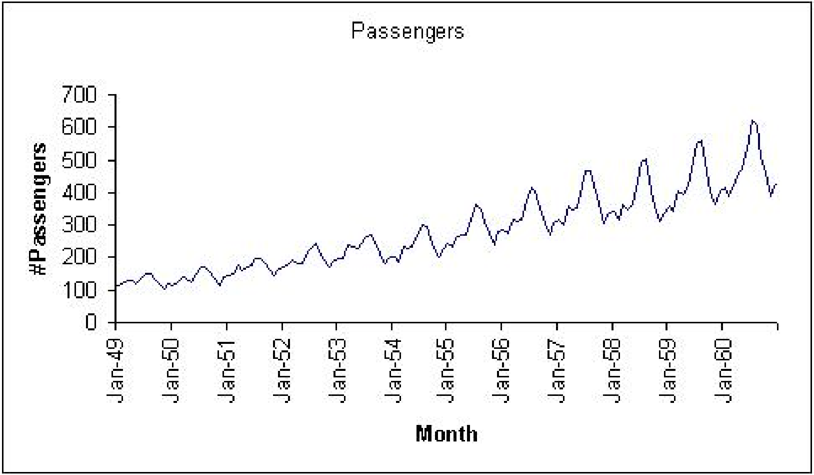
Plots



Smith, G. (2008), ”Japan’s Phillips Curve Looks Like Japan,”
Journal of Money, Credit and Banking.

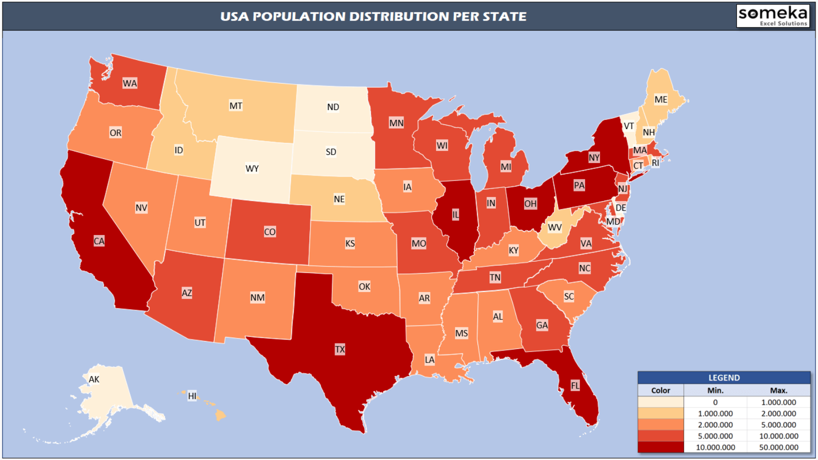
How to visualize?

Time plots



How to visualize?

Heat maps



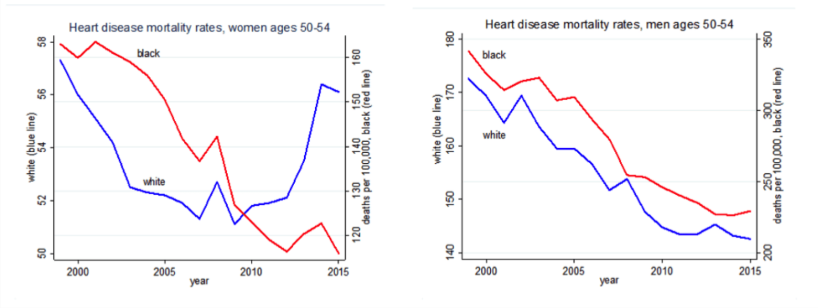
USA POPULATION DISTRIBUTION PER STATE

someka Excel Solutions

LEGEND

Color	Min.	Max.
⬜	0	1.000.000
🟨	1.000.000	2.000.000
🟧	2.000.000	5.000.000
🟥	5.000.000	10.000.000
🟫	10.000.000	50.000.000

Some visualization caveats



Case, A. and A. Deaton (2017), ”Mortality and Morbidity in the 21st Century,” *Brookings Papers on Economic Activity*.

Making the data sing

Beyond description: prediction and causality

Does A cause B?

Beyond description: prediction and causality

If A causes B, then *under the same conditions*:

- when A happens, then B happens
- when A doesn’t happen, then B doesn’t happen

The trouble with counterfactual reasoning

- Under the *same* conditions, really?
- What happens on ”The Road Not Taken”?

> James Salter, *Light Years*
>
> *Acts demolish their alternatives, that is the paradox.*

Beyond description: the role of theory

The power of theory

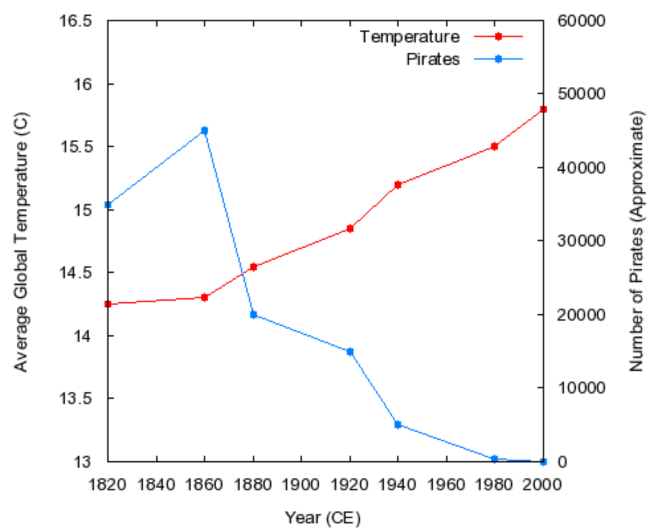
1. Simplifying
2. Hypothesizing
3. Falsifying

$\Rightarrow$ Theory is to data as orchestra is to an opera singer:

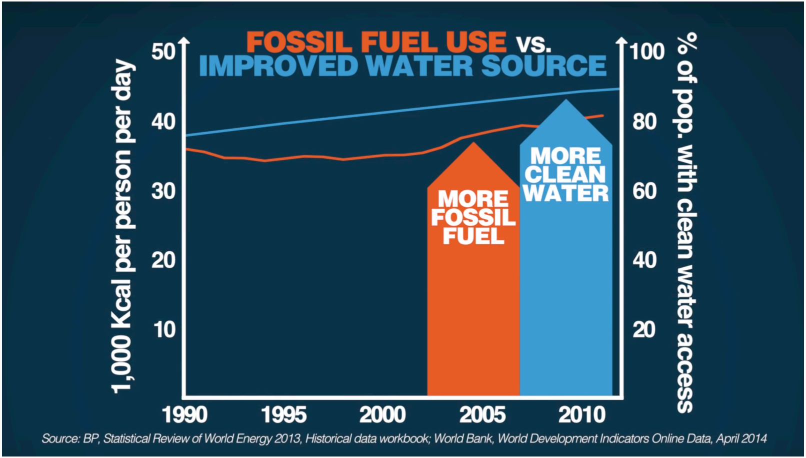
- It provides structure and guidance;
- It helps appreciate what the data is singing…
- … but shouldn’t cover its voice!

Don’t fall for the sirens’ song!

Spurious relationships



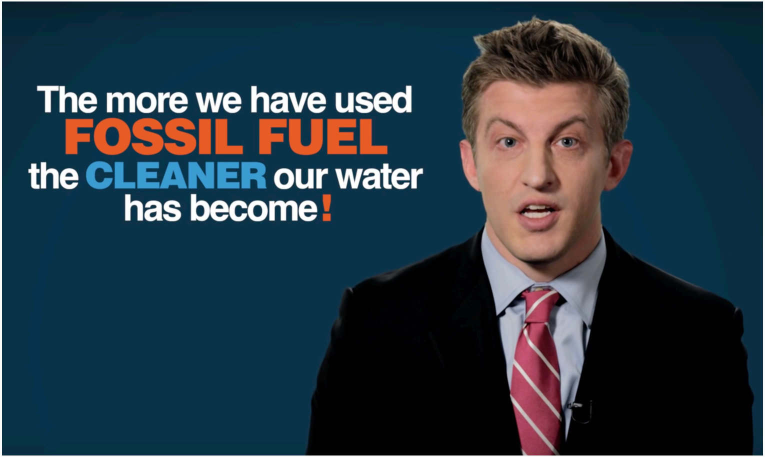
Spurious relationships



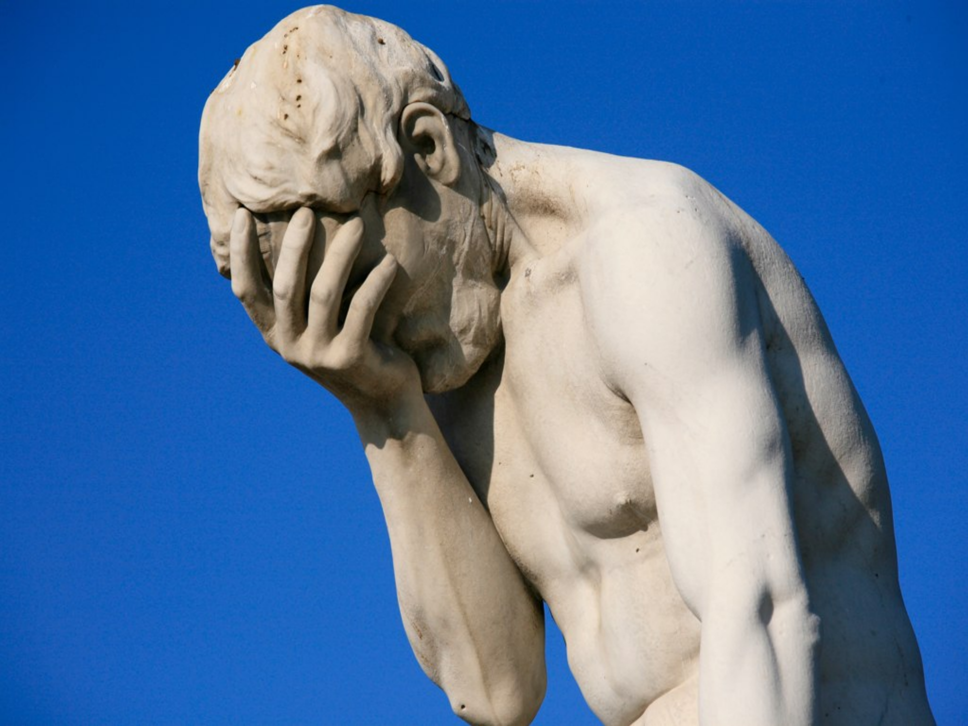
Source: BP, Statistical Review of World Energy 2013, Historical data workbook; World Bank, World Development Indicators Online Data, April 2014

Surely no one would interpret this as a causal relationship... *Right?*

Spurious relationships



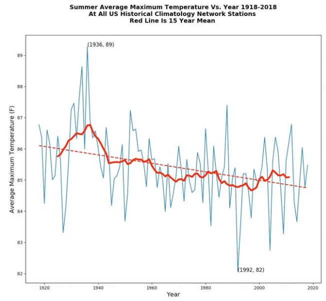
https://www.youtube.com/watch?v=BJWq1FeGpCw

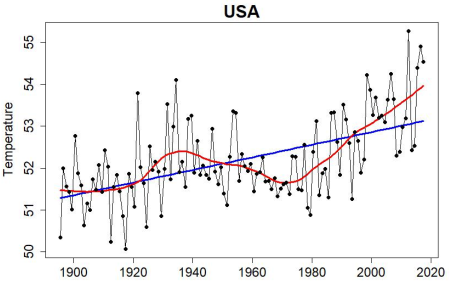


Researcher bias

Ronald H. Coase

If you torture the data long enough, it will confess.

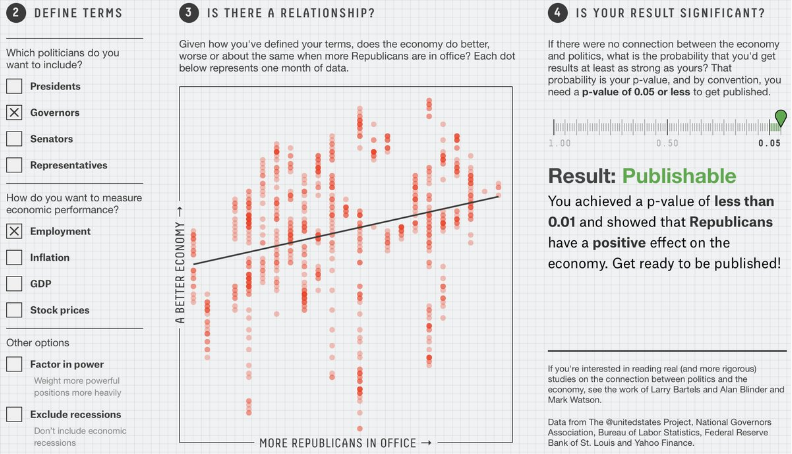




Researcher bias

Try it yourself!

https://projects.fivethirtyeight.com/p-hacking/



2 DEFINE TERMS

Which politicians do you want to include?

□ Presidents

☒ Governors

□ Senators

□ Representatives

How do you want to measure economic performance?

☒ Employment

□ Inflation

□ GDP

□ Stock prices

Other options

□ Factor in power

Weight more powerful positions more heavily

□ Exclude recessions

Don't include economic recessions

3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Republicans are in office? Each dot below represents one month of data.

4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.

Result: Publishable

You achieved a p-value of **less than 0.01** and showed that **Republicans** have a **positive** effect on the economy. Get ready to be published!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

Where is data?

Internal data

Bertold Brecht, *Reading Book for City Dwellers: Poems*

Whatever you say, don't say it twice
If you find your ideas in anyone else, disown them.
The man who hasn't signed anything, who has left no picture
Who was not there, who said nothing:
How can they catch him?
Cover your tracks.

⇒ Internal data (i.e. about yourself or your organization) is in every word spoken or written, every decision and every person around; you just have to *keep track*!

Big Data is just a reflection of how digitization allows to keep track of virtually everything.

Internal data

The N.E.R.D. framework

4 pillars of effective organizational data collection:

1. **Needs**: adapt the scope and frequency of data collection to its purpose;
2. **Encryption**: protect from external threats (hacking) but also internal threats (carelessness);
3. **Routines**: automate or simplify the process as much as possible;
4. **Deletion**: know what to keep and what to get rid of.

Finding external data

Secondary data

- Collected by someone else;
- Most often *observational*, i.e. the source of variation is not under the researcher's control.

- Identify your needs;
- Verify your sources (and cite them!);
- Find the right format for the methods you will use.

Finding external data

A few sources

- https://ourworldindata.org
- https://fred.stlouisfed.org
- https://data.worldbank.org
- https://dataverse.harvard.edu
- Companion data sets for (recent) academic publications

Finding external data

Primary data collection

Primary data

- Collected by YOU!
- Leaves more room for *experimental* procedures: designing and implementing an intervention.

- Surveys
- Field experiments
- Lab experiments
- Randomized-control trials

GAME

OVER