Analiza datelor. Corelație și cauzalitate

Proiectul învățământului superior din Moldova – DATASEA

Răzvan Rughiniș

razvan.rughinis@upb.ro

UNIVERSITATEA TEHNICĂ
A MOLDOVEI

Concepte principale

Co-incidență, asociere, corelație

Atribuiri robuste: criteriile lui Hill

Manipulări

Co-incidență, asociere, corelație

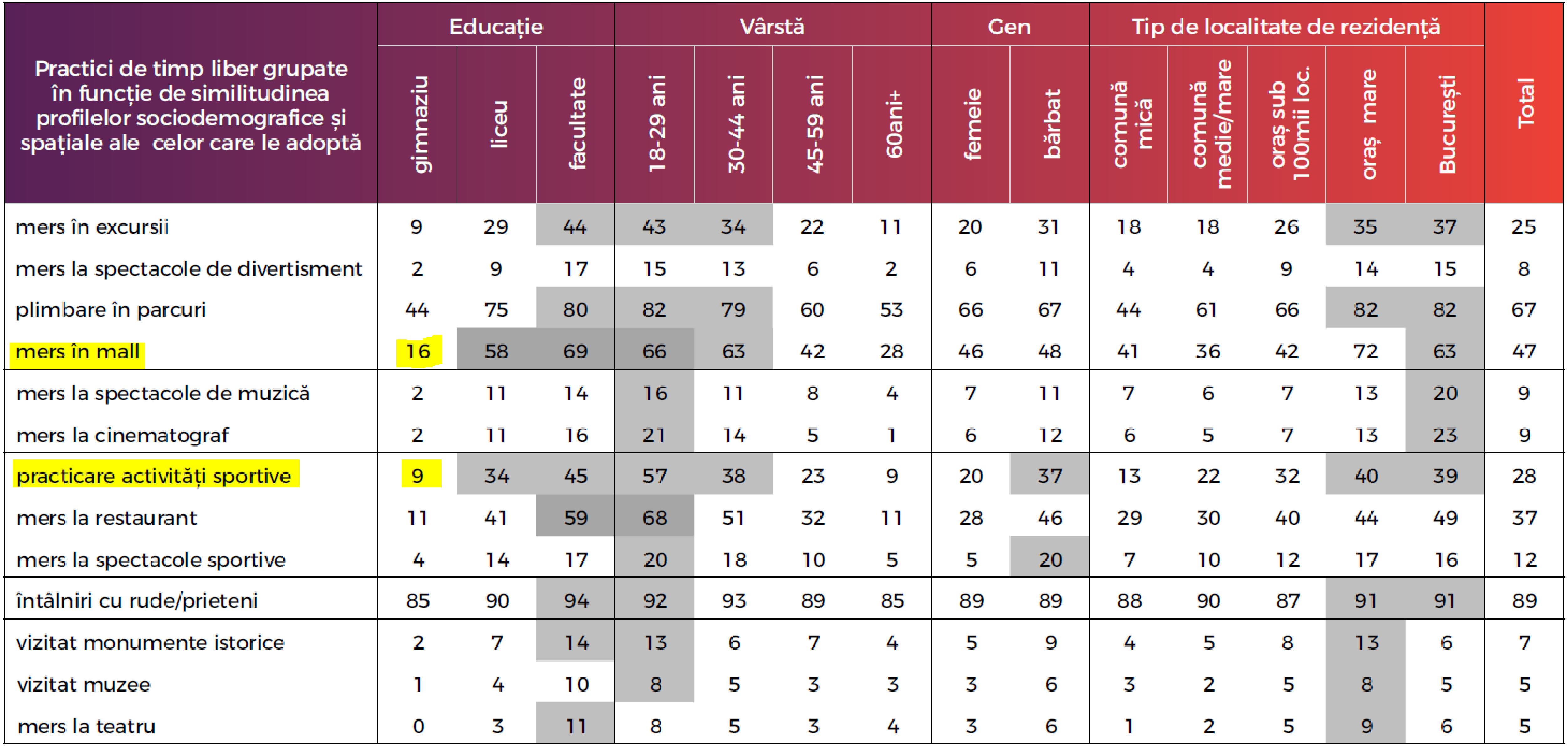
Atribuiri cauzale

Post hoc, ergo propter hoc

Cauzalitate vs. corelație

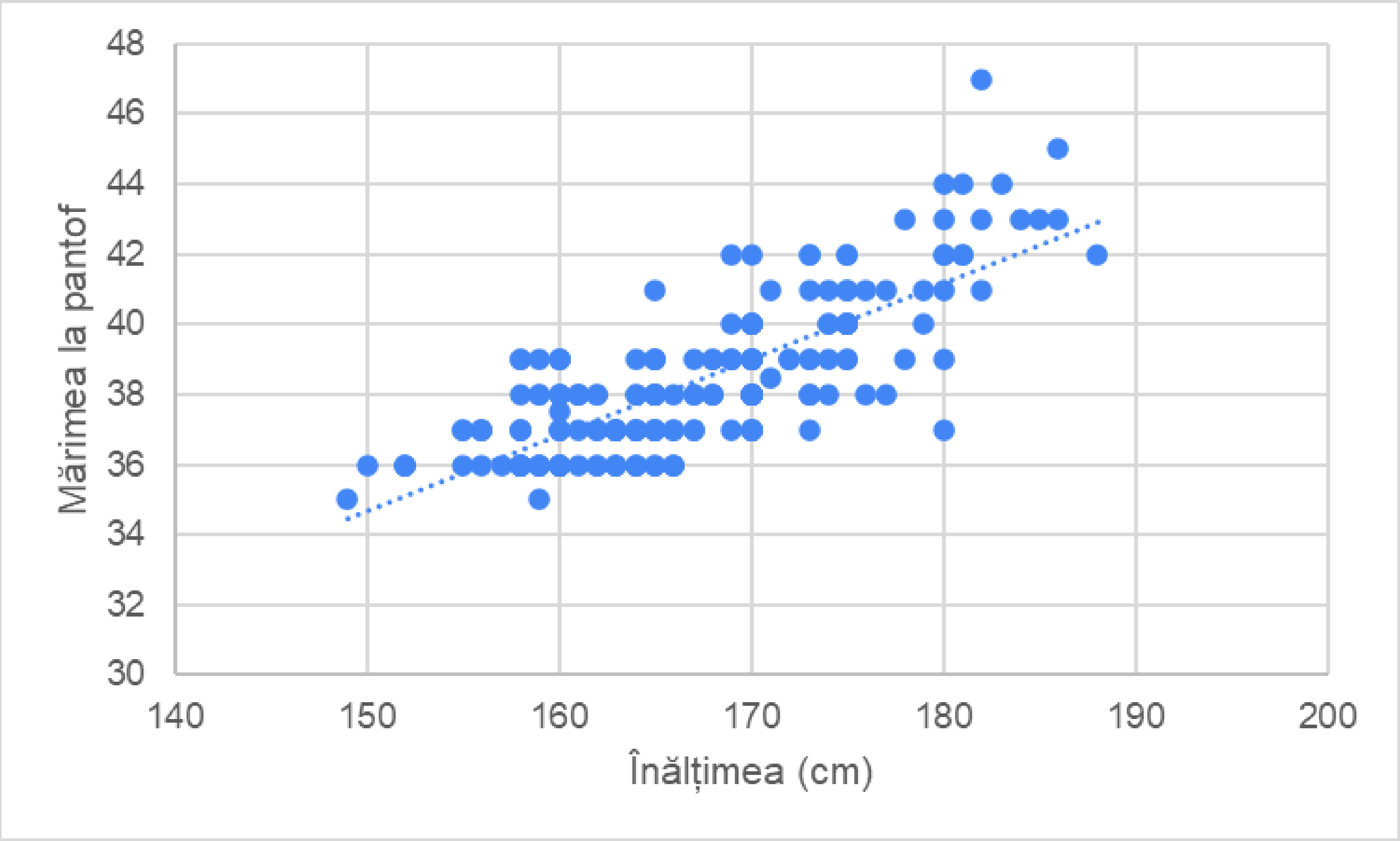
- Când există cauzalitate există și corelație, dar nu și invers!
- Eroarea „Post hoc, ergo propter hoc”
- Exemple:
  - Persoanele fără educație practică mult mai rar **sporturi**
  - Persoanele fără școală merg mult mai rar la **mall**
    - Este oare educația cauza, sau este o corelație care ascunde alte cauze?

Tabelul 2. Ponderea persoanelor care adoptă cel puțin lunar practici de timp liber pe categorii socio-demografice

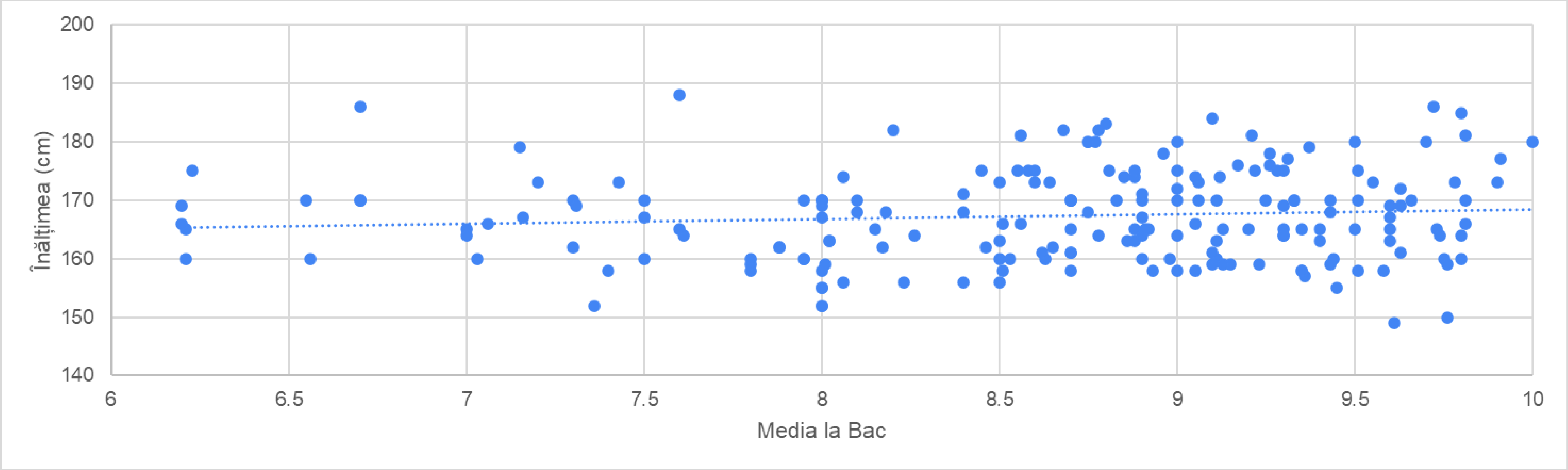


Practici de timp liber grupate în funcție de similitudinea profilelor sociodemografice și spațiale ale celor care le adoptă	Educație			Vârstă				Gen		Tip de localitate de rezidență					Total
	gimnaziu	liceu	facultate	18-29 ani	30-44 ani	45-59 ani	60ani+	femeie	bărbat	comună mică	comună medie/mare	oraș sub 100mii loc.	oraș mare	București	
mers în excursii	9	29	44	43	34	22	11	20	31	18	18	26	35	37	25
mers la spectacole de divertisment	2	9	17	15	13	6	2	6	11	4	4	9	14	15	8
plimbare în parcuri	44	75	80	82	79	60	53	66	67	44	61	66	82	82	67
mers în mall	16	58	69	66	63	42	28	46	48	41	36	42	72	63	47
mers la spectacole de muzică	2	11	14	16	11	8	4	7	11	7	6	7	13	20	9
mers la cinematograf	2	11	16	21	14	5	1	6	12	6	5	7	13	23	9
practicare activități sportive	9	34	45	57	38	23	9	20	37	13	22	32	40	39	28
mers la restaurant	11	41	59	68	51	32	11	28	46	29	30	40	44	49	37
mers la spectacole sportive	4	14	17	20	18	10	5	5	20	7	10	12	17	16	12
întâlniri cu rude/prieteni	85	90	94	92	93	89	85	89	89	88	90	87	91	91	89
vizitat monumente istorice	2	7	14	13	6	7	4	5	9	4	5	8	13	6	7
vizitat muzee	1	4	10	8	5	3	3	3	6	3	2	5	8	5	5
mers la teatru	0	3	11	8	5	3	4	3	6	1	2	5	9	6	5

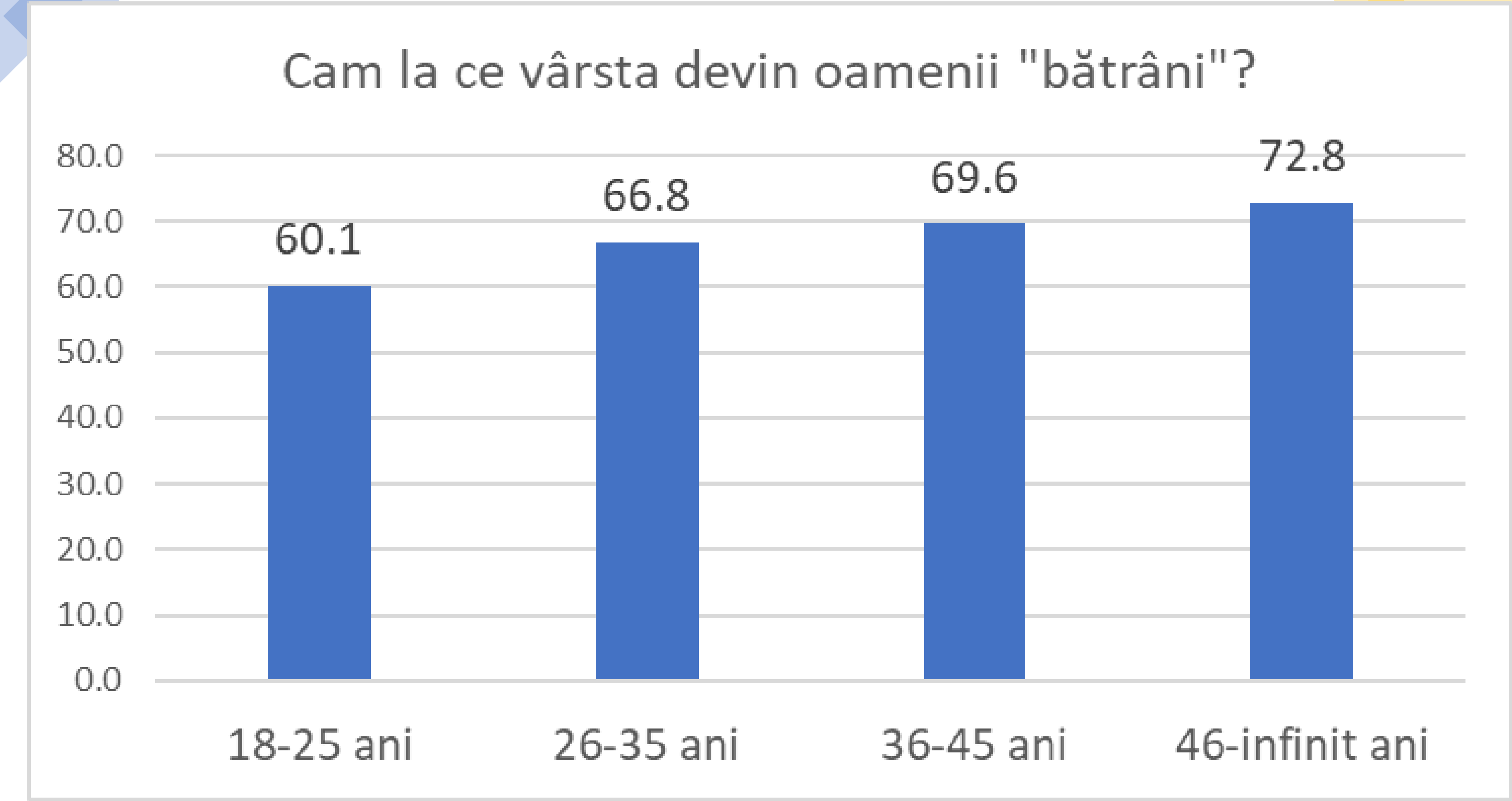
Sursă: Barometrul Consumului Cultural 2019



Coeficient Pearson de corelație: +0.78



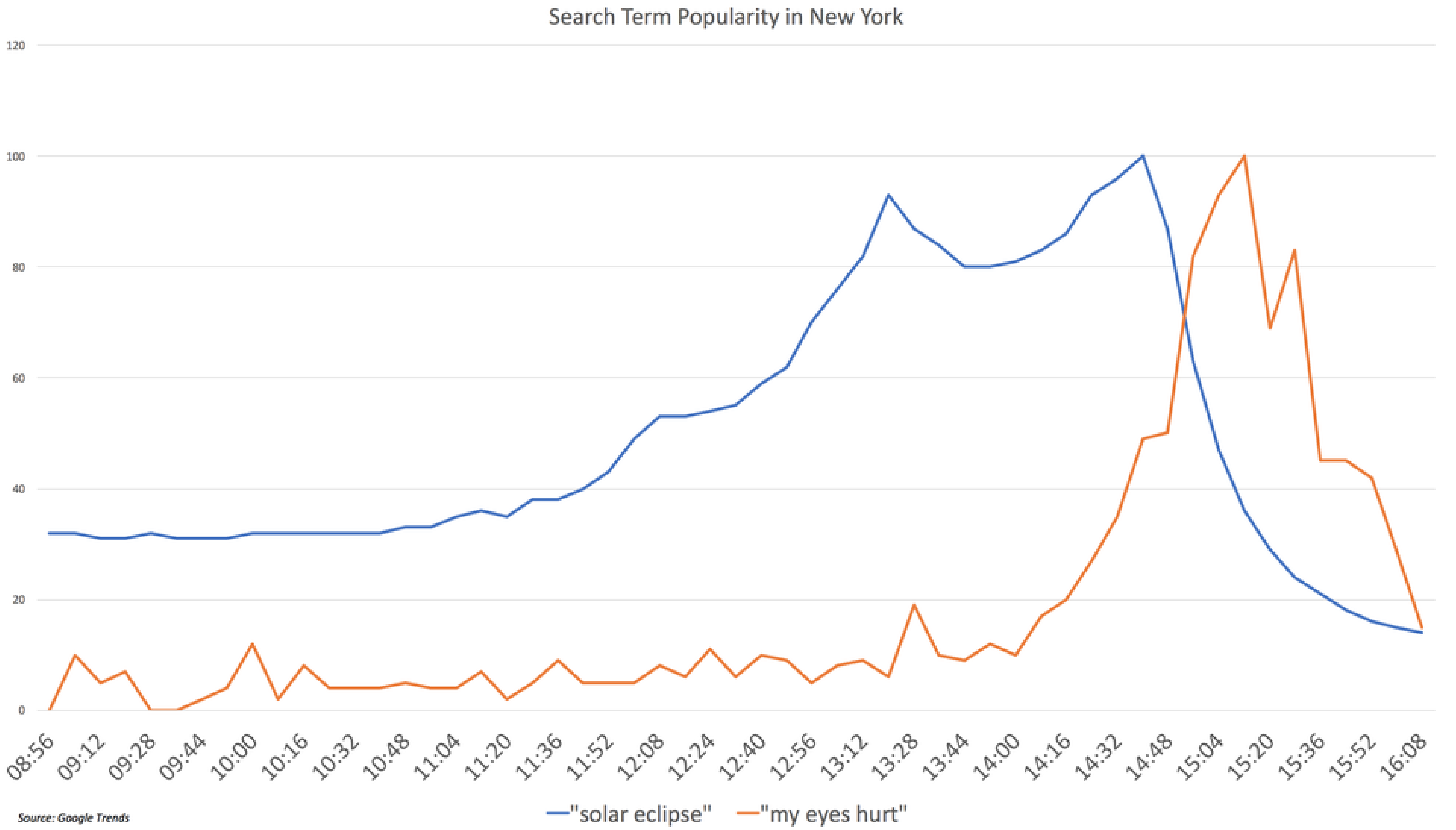
Coeficient Pearson de corelație: +0.09



Cam la ce vârsta devin oamenii "bătrâni"?

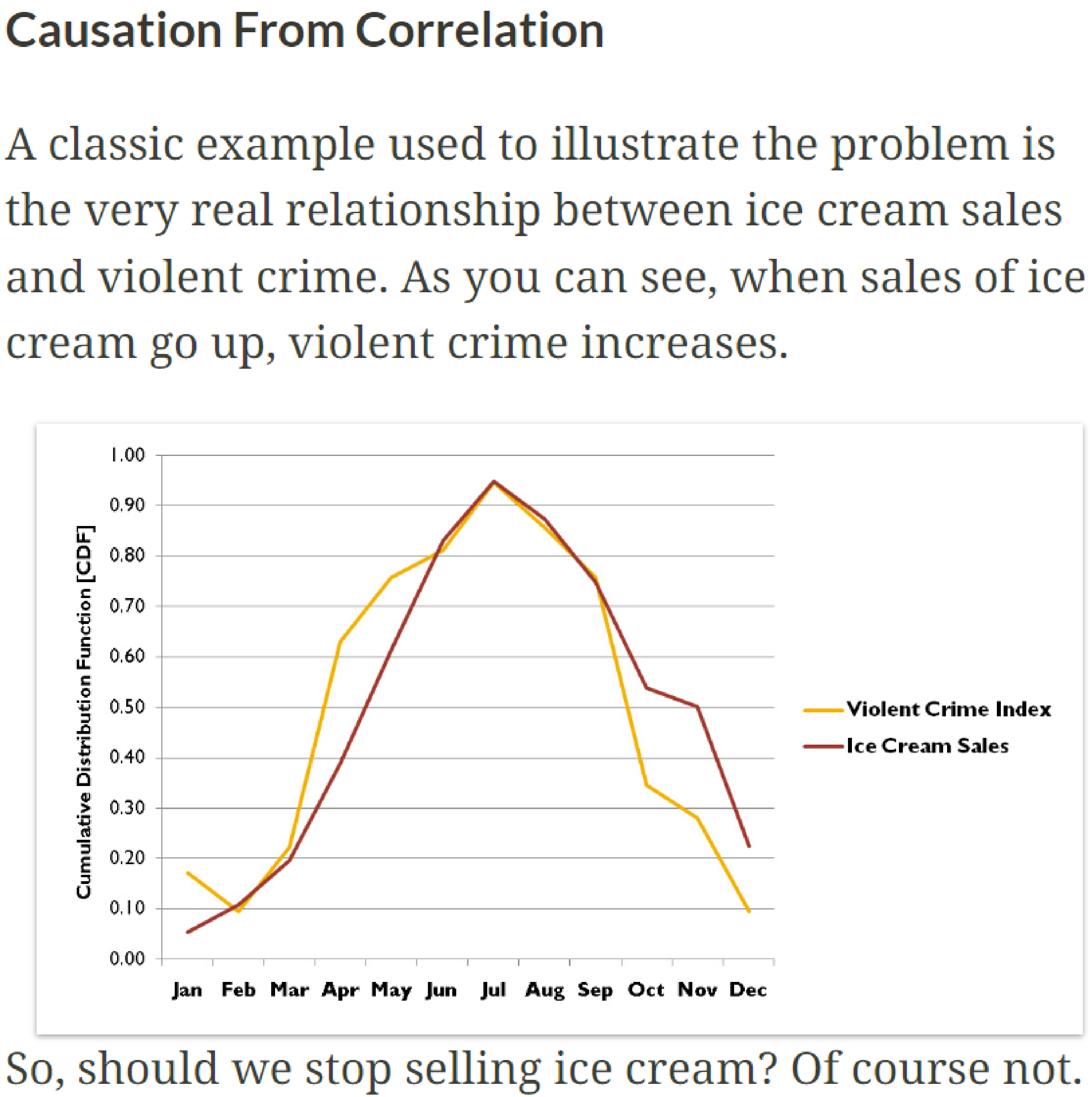
Co-incidență

2017



Sursa

Spurious correlation



Causation From Correlation

A classic example used to illustrate the problem is the very real relationship between ice cream sales and violent crime. As you can see, when sales of ice cream go up, violent crime increases.

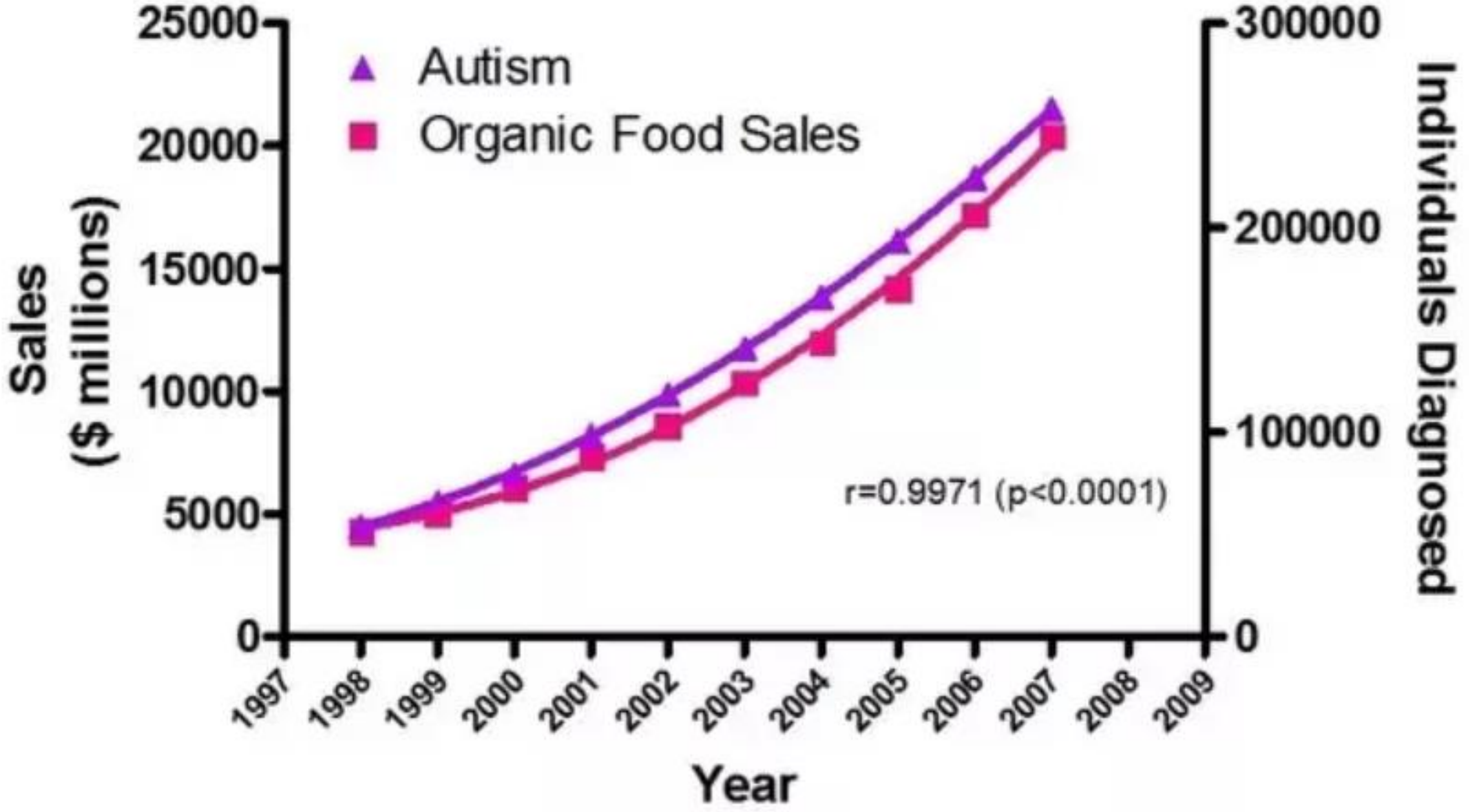
So, should we stop selling ice cream? Of course not.

Source

Spurious correlation

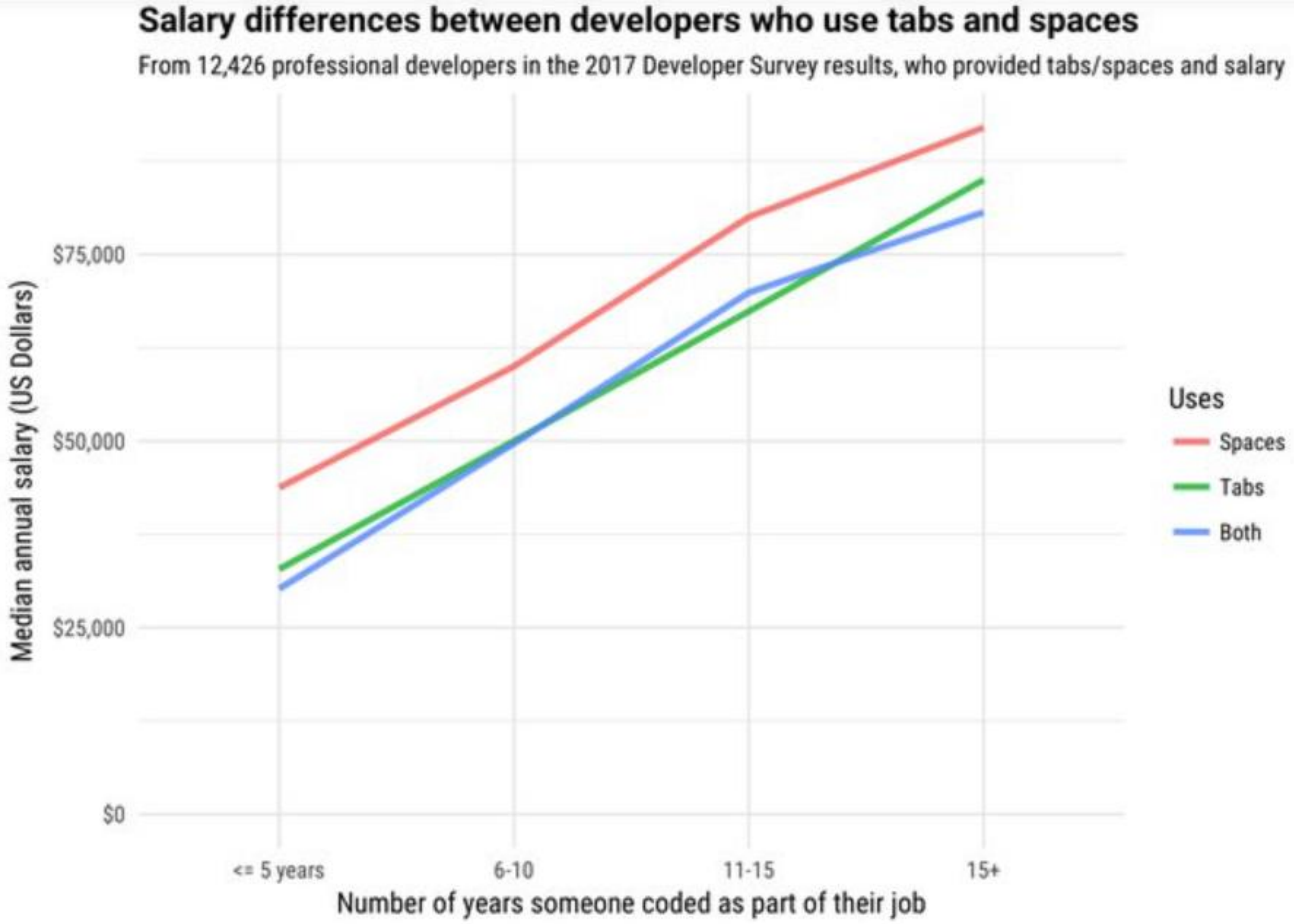
Rates of autism correlate very closely (r=.9971) with sales of organic produce.

The real cause of increasing autism prevalence?

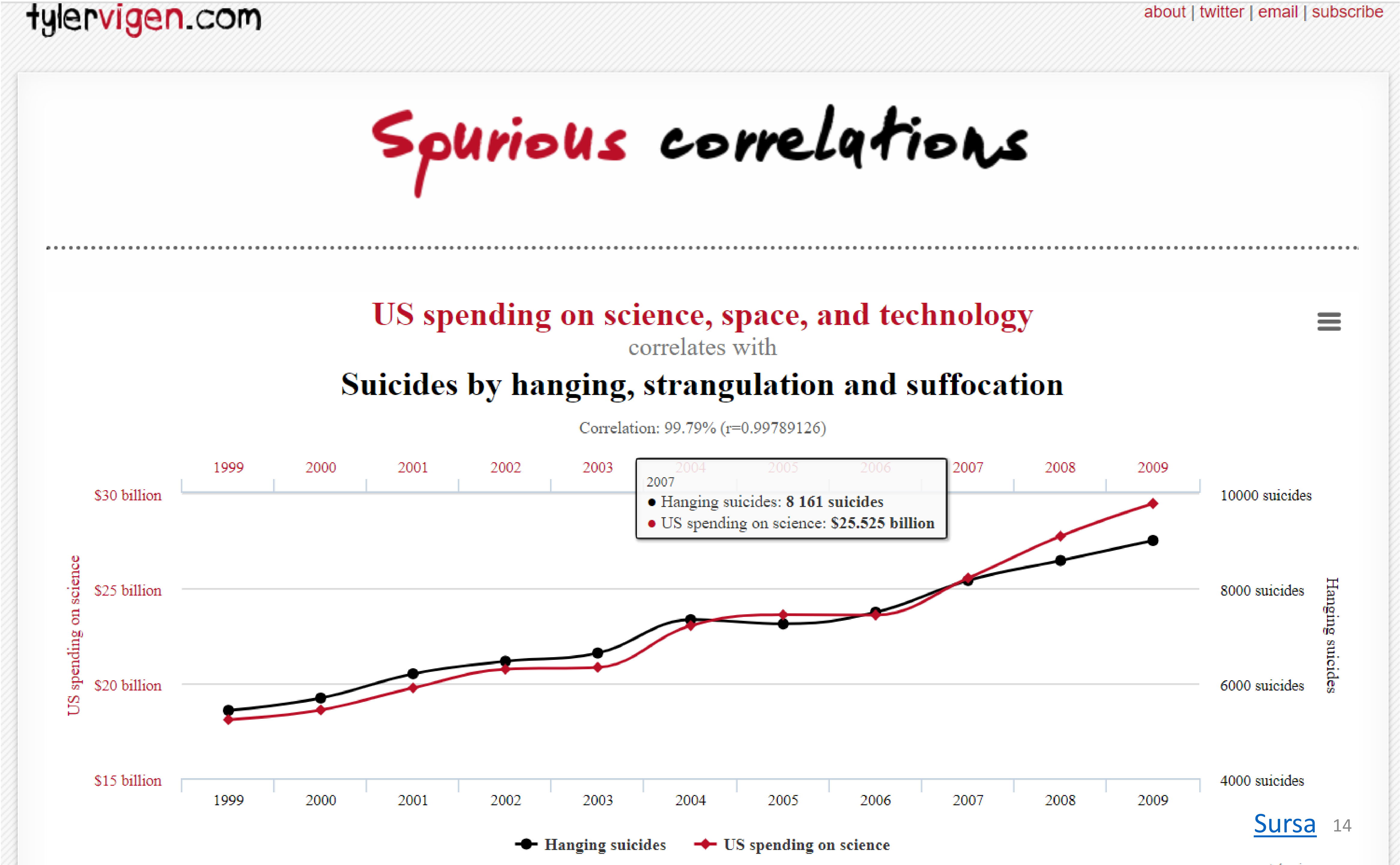


Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals
with Disabilities Education Act

Spurious correlation



The data was amassed using the Stack Overflow 2017 Developer Survey ↗.



tylervigen.com

about | twitter | email | subscribe

Spurious correlations

US spending on science, space, and technology

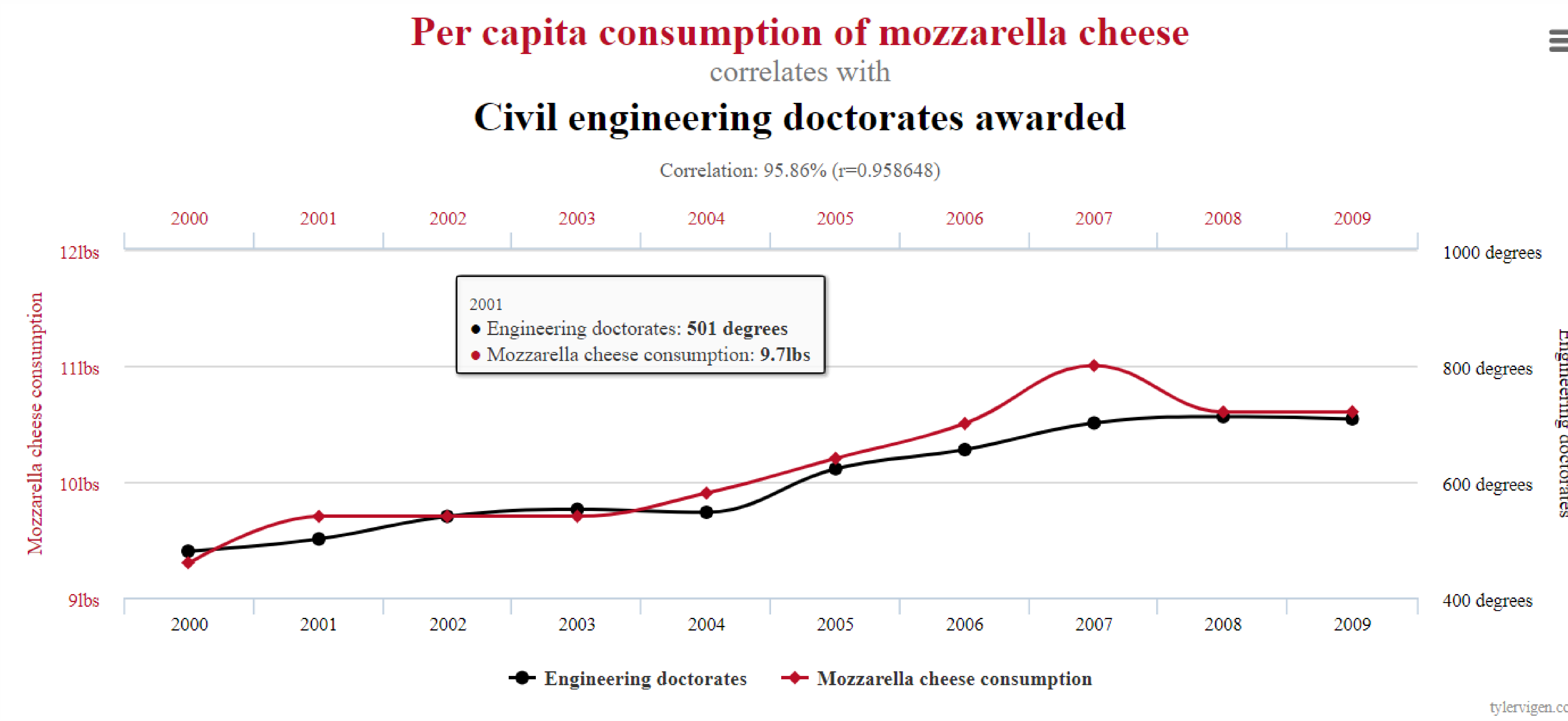
correlates with

Suicides by hanging, strangulation and suffocation

Correlation: 99.79% (r=0.99789126)

2007
● Hanging suicides: **8 161 suicides**
● US spending on science: **$25.525 billion**

Sursa



Per capita consumption of mozzarella cheese

correlates with

Civil engineering doctorates awarded

Correlation: 95.86% (r=0.958648)

Data sources: U.S. Department of Agriculture and National Science Foundation

Sursa

Asociere

2011

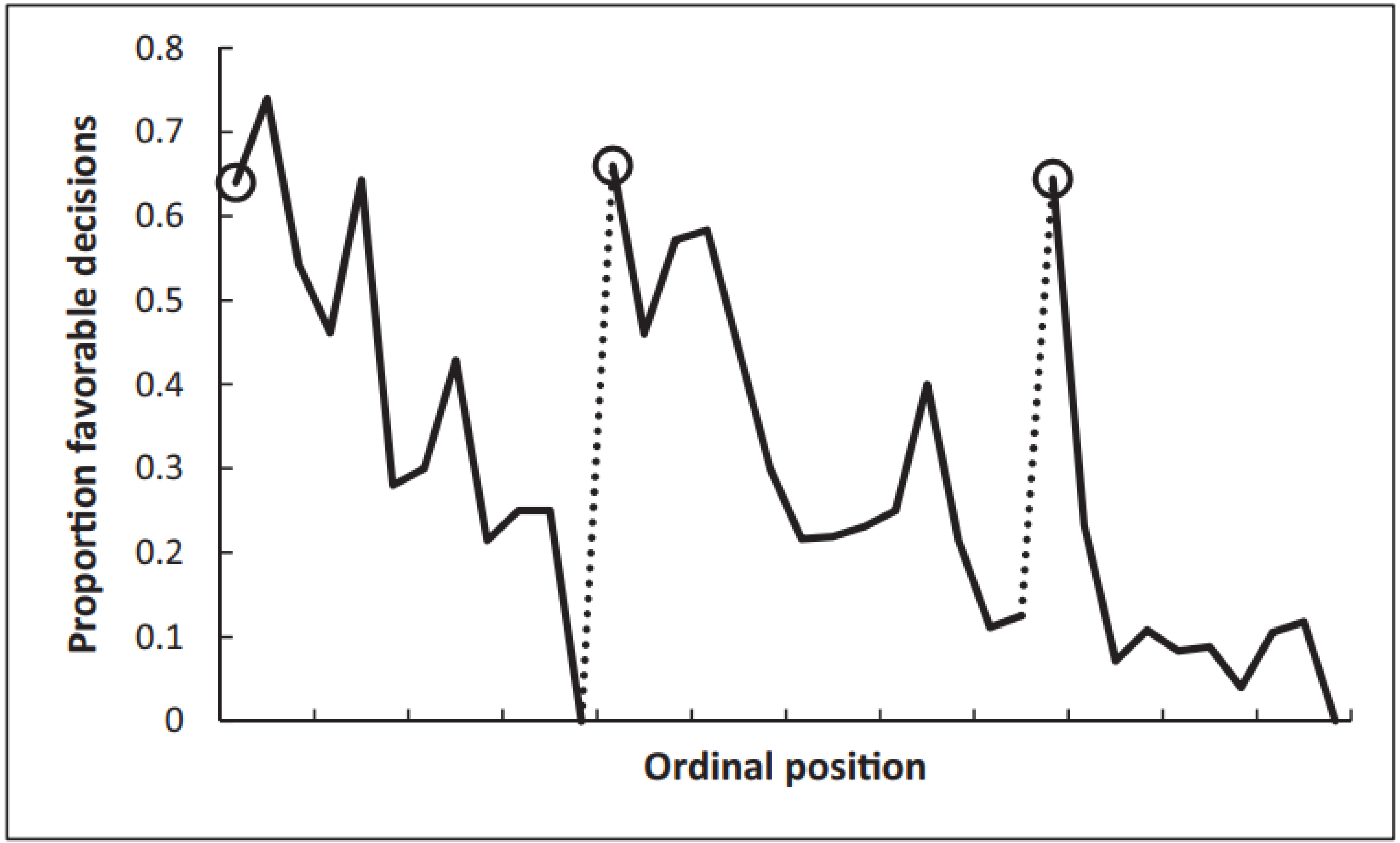
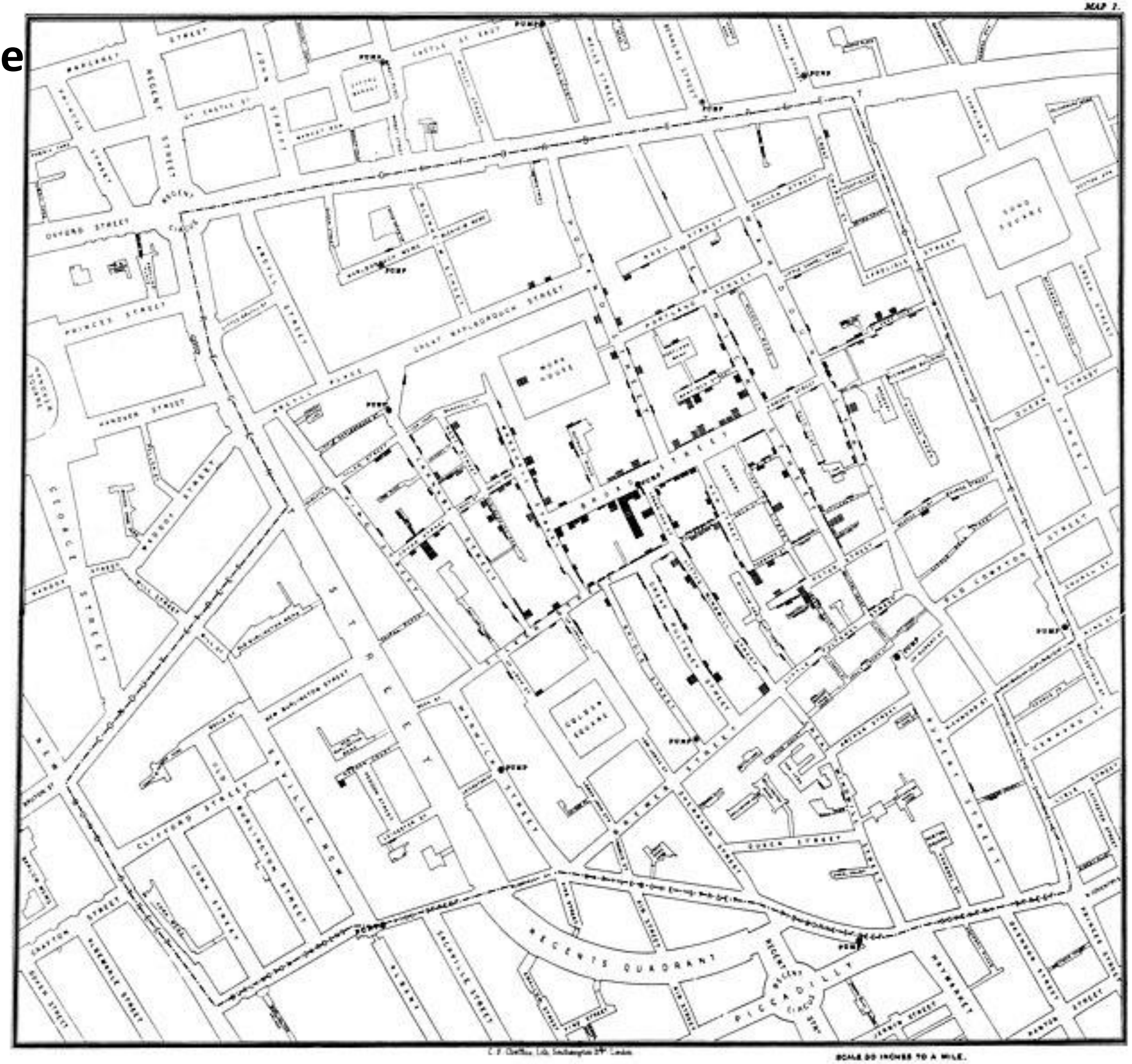


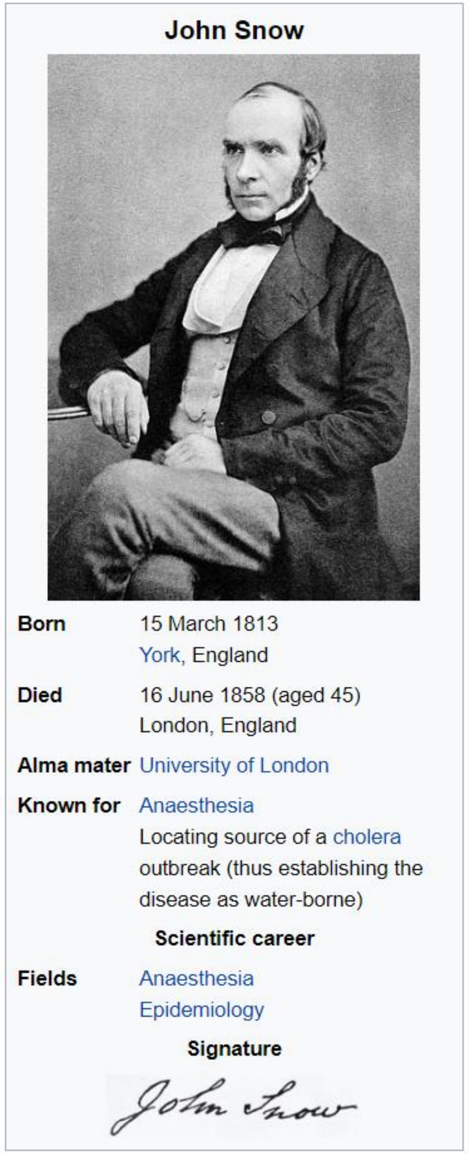
Fig. 1. Proportion of rulings in favor of the prisoners by ordinal position. Circled points indicate the first decision in each of the three decision sessions; tick marks on *x* axis denote every third case; dotted line denotes food break. Because unequal session lengths resulted in a low number of cases for some of the later ordinal positions, the graph is based on the first 95% of the data from each session.

Sursa

Asociere

1854





John Snow

Born	15 March 1813 York, England
Died	16 June 1858 (aged 45) London, England
Alma mater	University of London
Known for	Anaesthesia Locating source of a cholera outbreak (thus establishing the disease as water-borne)

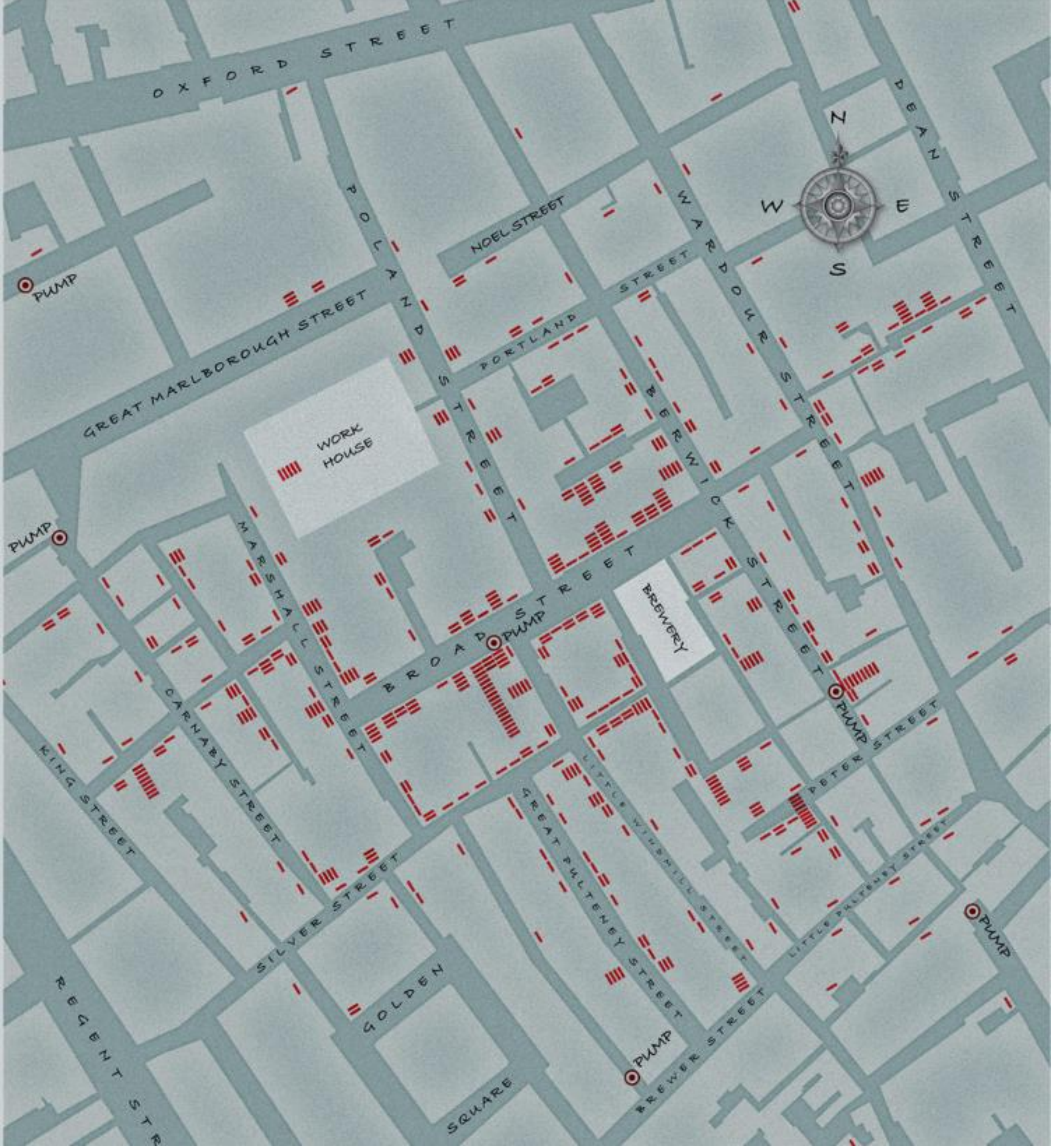
Scientific career

Fields	Anaesthesia Epidemiology

Signature

John Snow

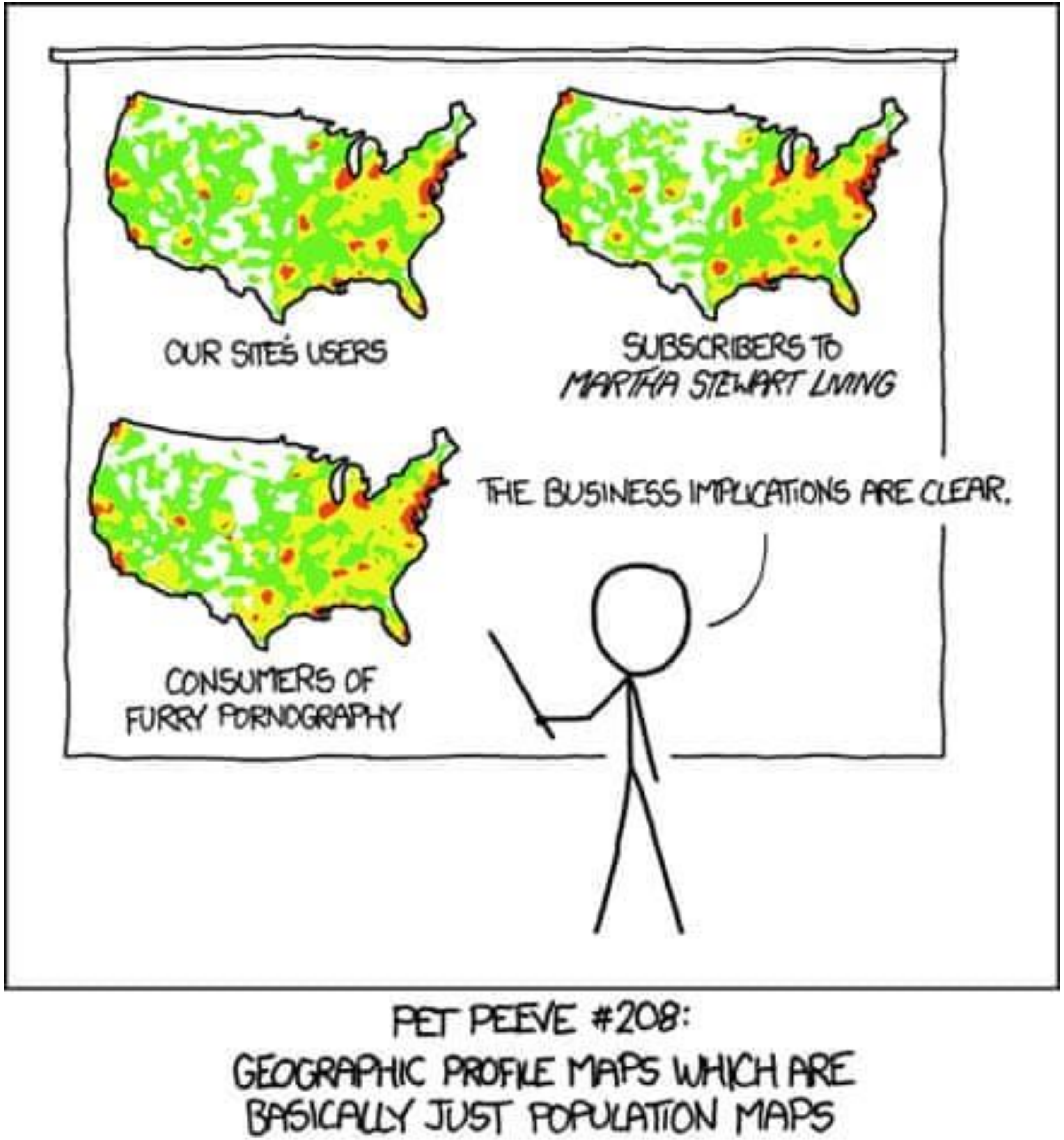
Sursa



„At a local brewery, the workers were allowed all the beer they could drink - it was believed they didn't drink water at all. But it had its own water supply too and there were consequently fewer cases."

Sursă

Discuție in
The Guardian


OUR SITE'S USERS
SUBSCRIBERS TO MARTHA STEWART LIVING
CONSUMERS OF FURRY PORNOGRAPHY
THE BUSINESS IMPLICATIONS ARE CLEAR.
PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

Sursă

Asociere

1858

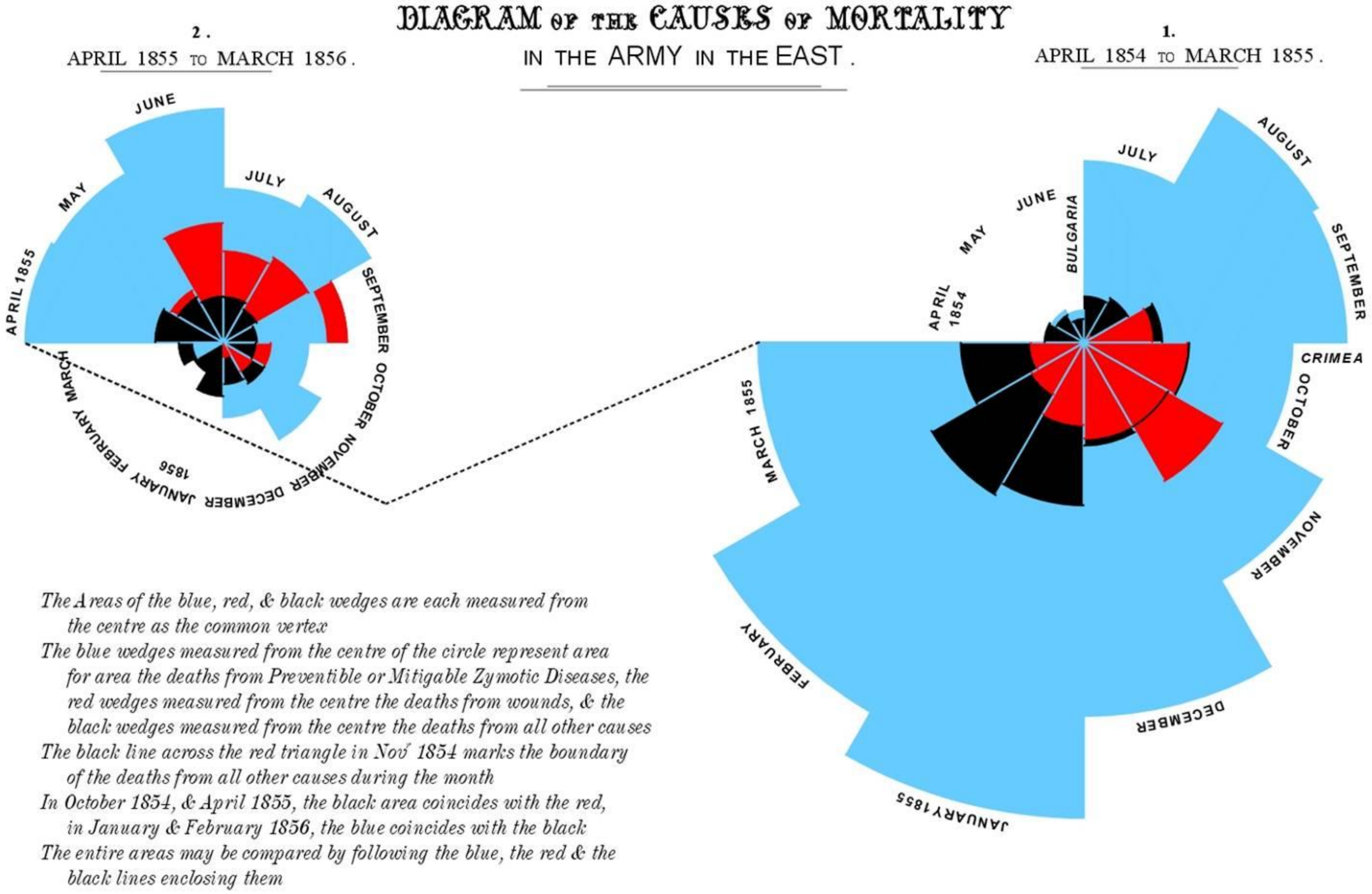


DIAGRAM OF THE CAUSES OF MORTALITY IN THE ARMY IN THE EAST.

2. APRIL 1855 TO MARCH 1856.

1. APRIL 1854 TO MARCH 1855.

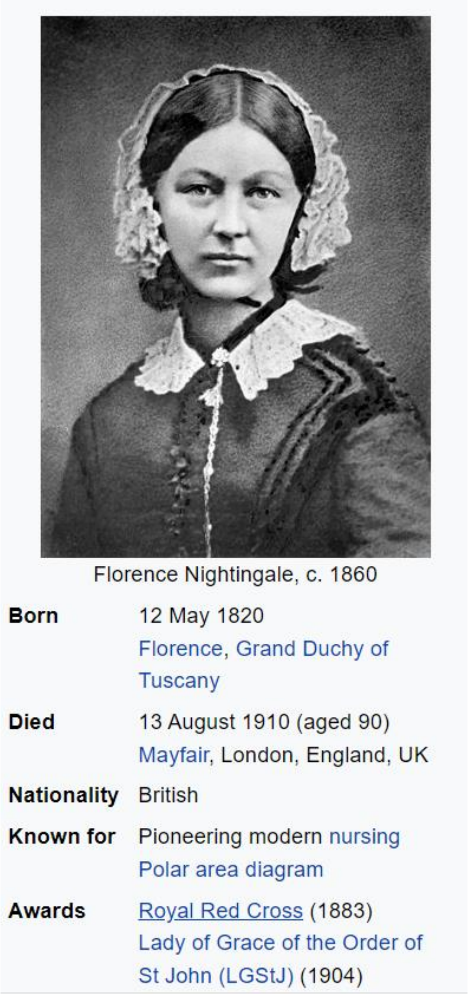
The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex

The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic Diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes

The black line across the red triangle in Novr 1854 marks the boundary of the deaths from all other causes during the month

In October 1854, & April 1855, the black area coincides with the red, in January & February 1856, the blue coincides with the black

The entire areas may be compared by following the blue, the red & the black lines enclosing them

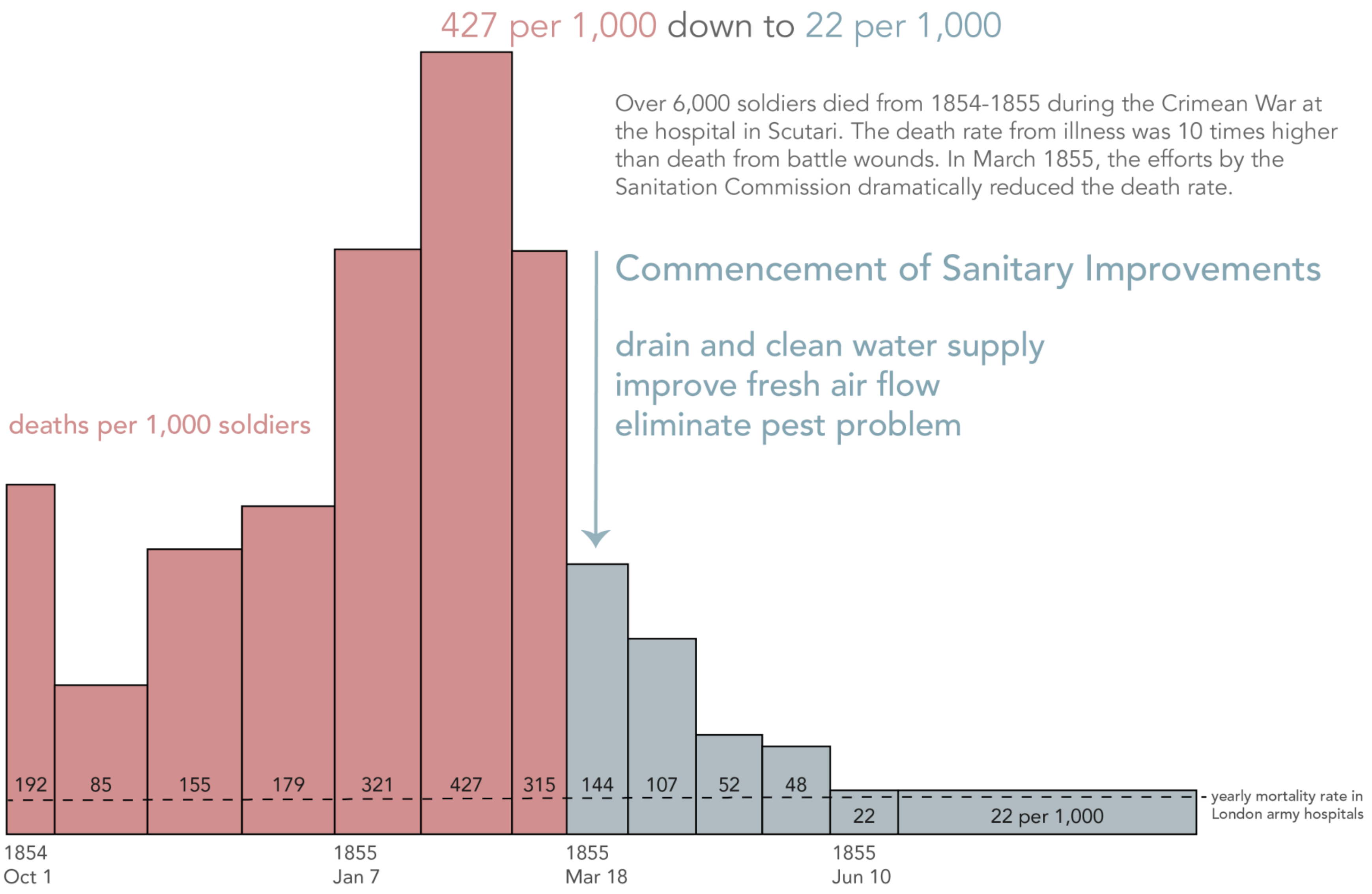


Florence Nightingale, c. 1860

Born	12 May 1820 Florence, Grand Duchy of Tuscany
Died	13 August 1910 (aged 90) Mayfair, London, England, UK
Nationality	British
Known for	Pioneering modern nursing Polar area diagram
Awards	Royal Red Cross (1883) Lady of Grace of the Order of St John (LGStJ) (1904)

The Scutari Death Camp

simple sanitation improvements can save lives



Over 6,000 soldiers died from 1854-1855 during the Crimean War at the hospital in Scutari. The death rate from illness was 10 times higher than death from battle wounds. In March 1855, the efforts by the Sanitation Commission dramatically reduced the death rate.

Commencement of Sanitary Improvements

drain and clean water supply
improve fresh air flow
eliminate pest problem

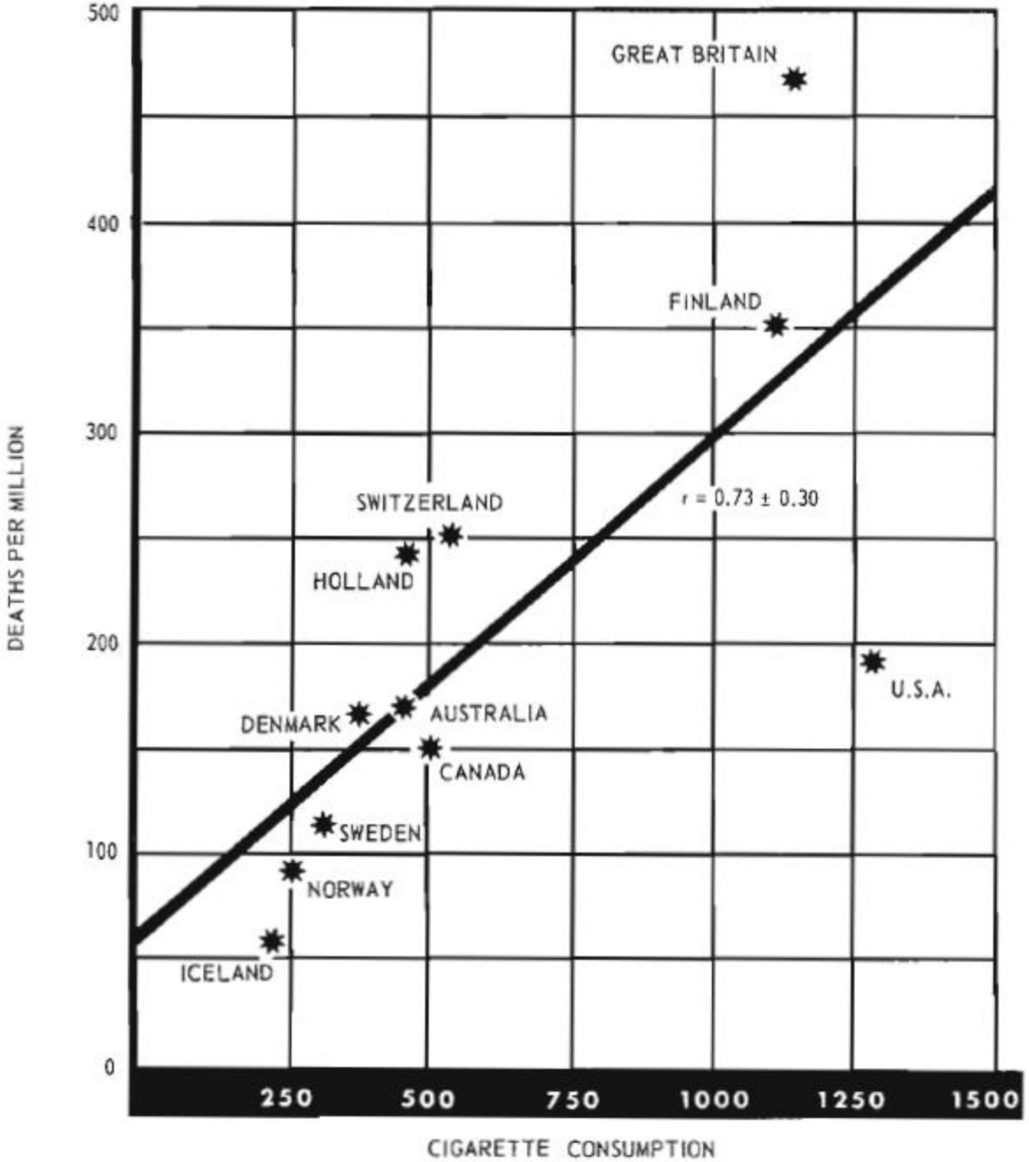
A redesign of Florence Nightingale's Rose Chart created by Jeffrey A. Shaffer | DataPlusScience.com

Sursa

Corelație

1964

CRUDE MALE DEATH RATE FOR LUNG CANCER IN 1950 AND PER CAPITA CONSUMPTION OF CIGARETTES IN 1930 IN VARIOUS COUNTRIES.

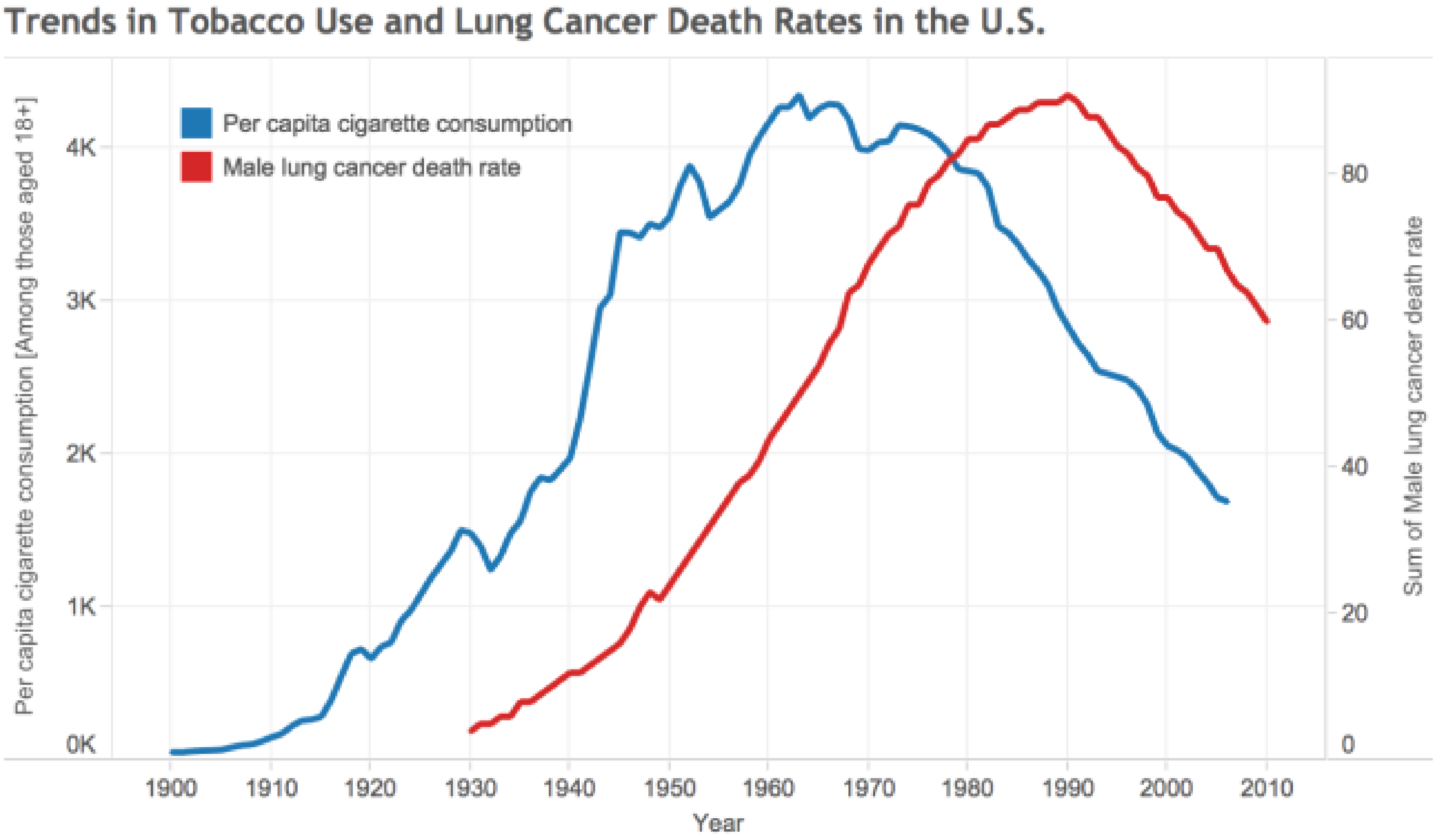


Report of the Advisory Committee to the Surgeon General, *Smoking and Health* (Washington, D.C., 1964), p. 176; based on R. Doll, "Etiology of Lung Cancer," *Advances in Cancer Research*, 3 (1955), 1–50.

Sursă:
Tufte, 2001

Corelație

2010

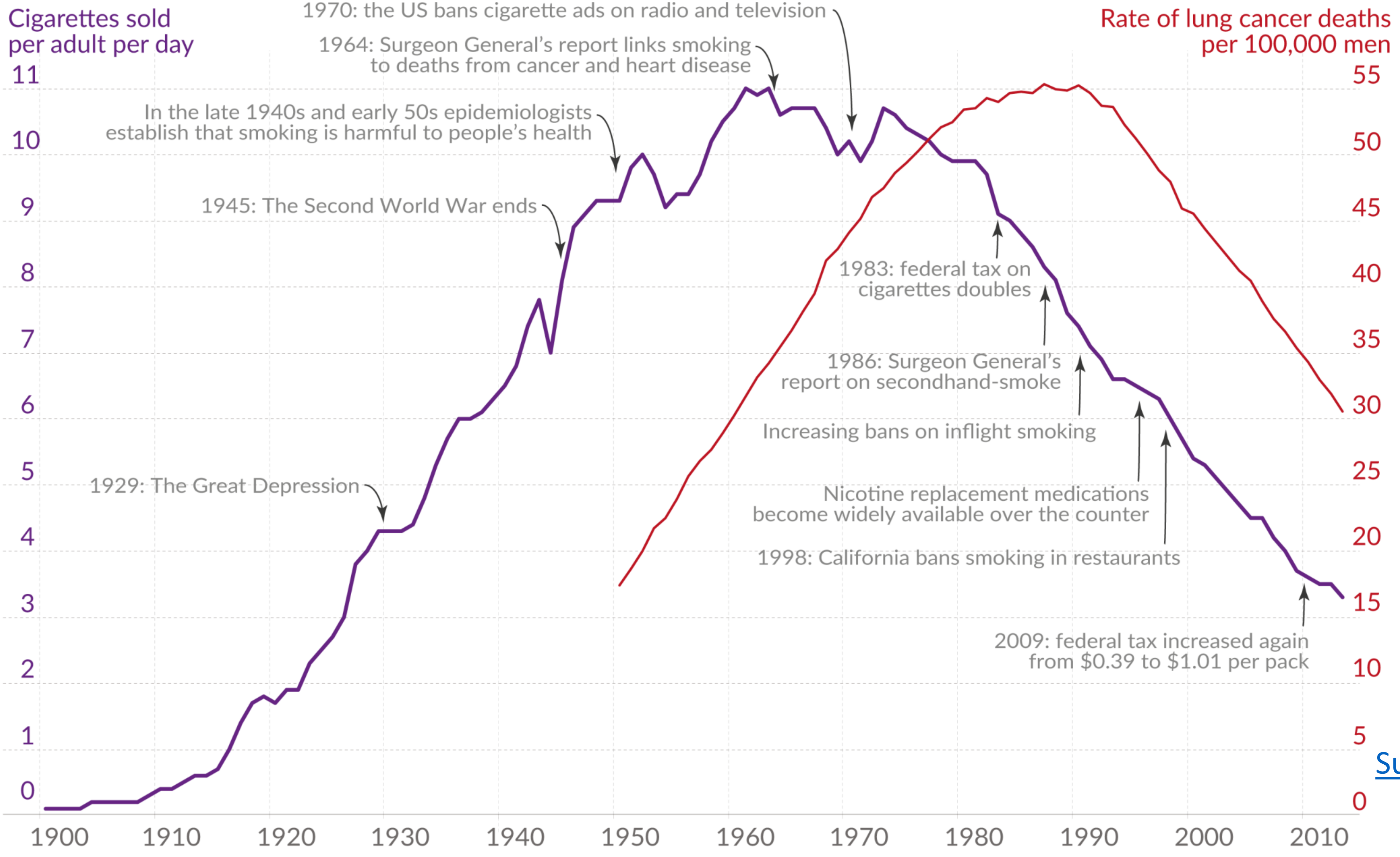


Death rates source: US Mortality Data, 1960-2010, US Mortality Volumes, 1930-1959, National Center for Health Statistics, Centers for Disease Control and Prevention.
Cigarette consumption source: US Department of Agriculture, 1900-2007.

Sursa

Cigarette sales and lung cancer mortality in the US

Our World in Data



1970: the US bans cigarette ads on radio and television

1964: Surgeon General's report links smoking to deaths from cancer and heart disease

In the late 1940s and early 50s epidemiologists establish that smoking is harmful to people's health

1945: The Second World War ends

1929: The Great Depression

1983: federal tax on cigarettes doubles

1986: Surgeon General's report on secondhand-smoke

Increasing bans on inflight smoking

Nicotine replacement medications become widely available over the counter

1998: California bans smoking in restaurants

2009: federal tax increased again from $0.39 to $1.01 per pack

Sursa

Data sources: International Smoking Statistics (2017); WHO Cancer Mortality Database (IARC). The death rate from lung-cancer is age-standardized.

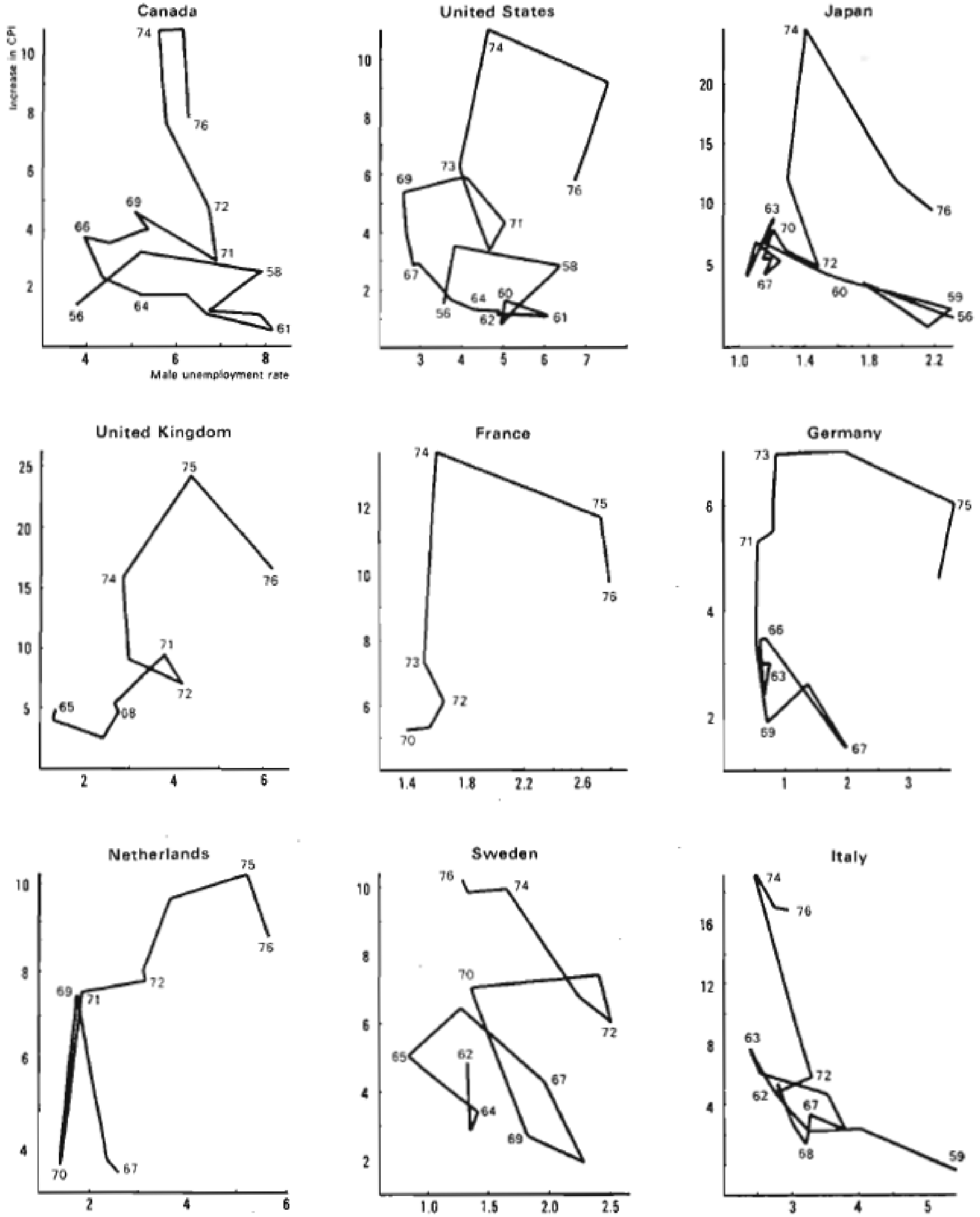
OurWorldinData.org – Research and data to make progress against the world's largest problems.

Lipsa corelației

1977

Inflation and Unemployment Rates

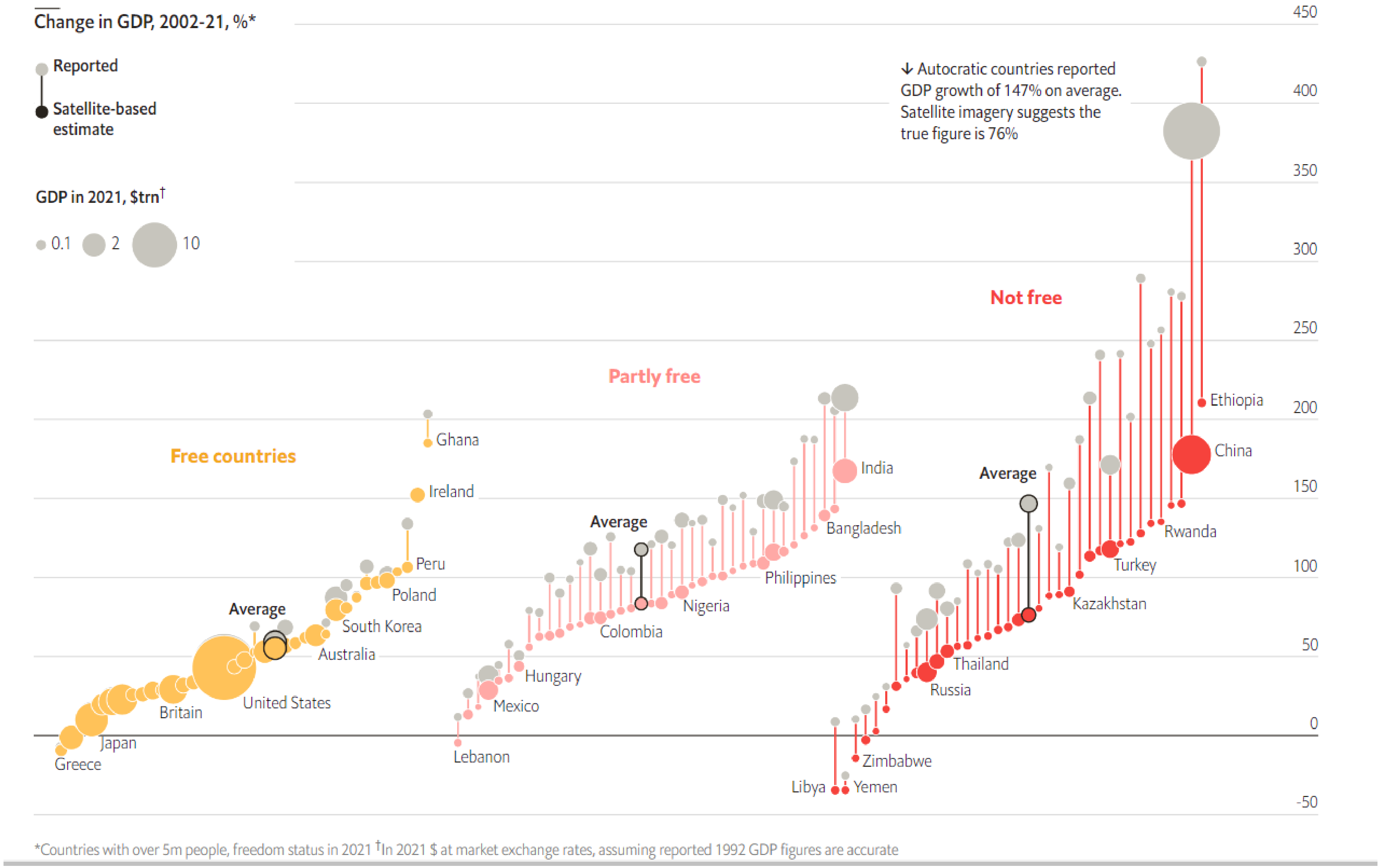
Per cent



Paul McCracken, et al., *Towards Full Employment and Price Stability* (Paris, 1977), p. 106.

Sursă: Tufte, 2001

2022



*Countries with over 5m people, freedom status in 2021 †In 2021 $ at market exchange rates, assuming reported 1992 GDP figures are accurate

Sursa: The Economist

Atribuiri cauzale **robuste**

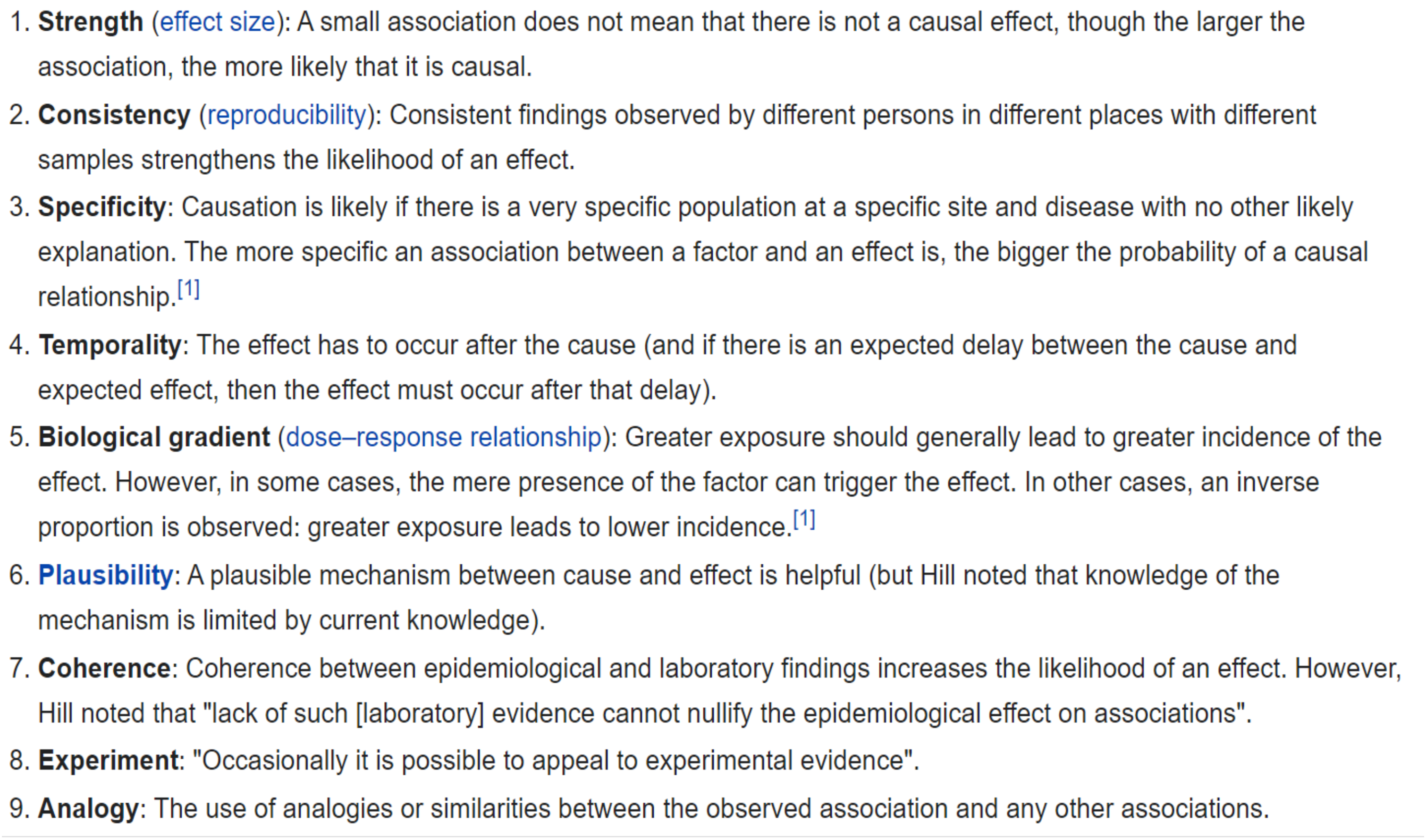
Criteriile lui Hill

Criteriile cauzalității @ Austin Bradford Hill (1965)

Strength	Intensitatea asocierii
Consistency	Consistență între surse diferite
Specificity	Specificitate
Temporality	Temporalitate (înainte / după)
Biological gradient	Doză / răspuns
Plausibility	Plauzibilitate
Coherence	Coerență experiment - epidemiologie
Experiment	Dovezi experimentale
Analogy	Dovezi prin analogie

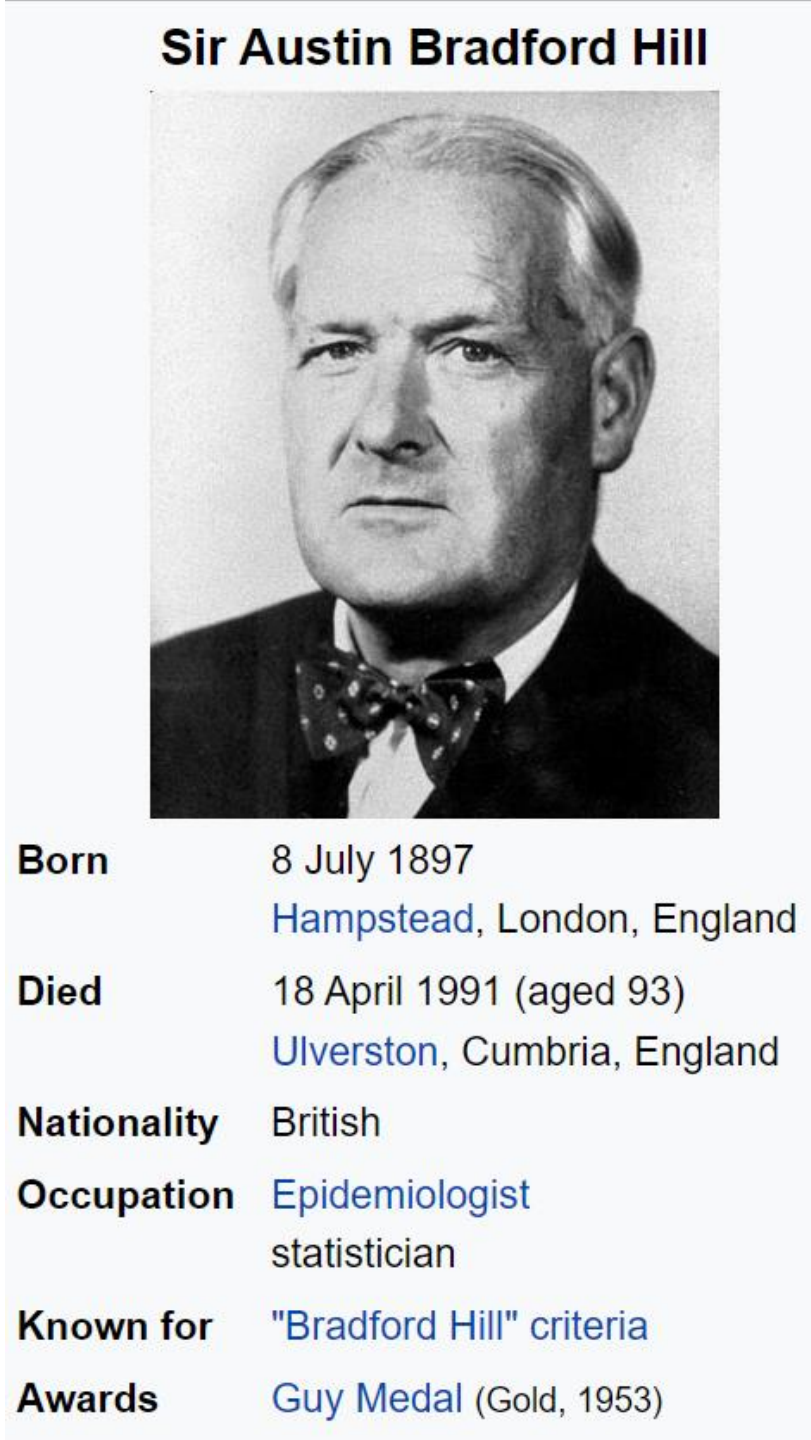
Austin Bradford Hill, “The Environment and Disease: Association or Causation?,” *Proceedings of the Royal Society of Medicine*, 58 (1965), 295-300.

Criteriile cauzalității: Austin Bradford Hill (1965)



1. **Strength** (effect size): A small association does not mean that there is not a causal effect, though the larger the association, the more likely that it is causal.
2. **Consistency** (reproducibility): Consistent findings observed by different persons in different places with different samples strengthens the likelihood of an effect.
3. **Specificity**: Causation is likely if there is a very specific population at a specific site and disease with no other likely explanation. The more specific an association between a factor and an effect is, the bigger the probability of a causal relationship.[1]
4. **Temporality**: The effect has to occur after the cause (and if there is an expected delay between the cause and expected effect, then the effect must occur after that delay).
5. **Biological gradient** (dose–response relationship): Greater exposure should generally lead to greater incidence of the effect. However, in some cases, the mere presence of the factor can trigger the effect. In other cases, an inverse proportion is observed: greater exposure leads to lower incidence.[1]
6. **Plausibility**: A plausible mechanism between cause and effect is helpful (but Hill noted that knowledge of the mechanism is limited by current knowledge).
7. **Coherence**: Coherence between epidemiological and laboratory findings increases the likelihood of an effect. However, Hill noted that "lack of such [laboratory] evidence cannot nullify the epidemiological effect on associations".
8. **Experiment**: "Occasionally it is possible to appeal to experimental evidence".
9. **Analogy**: The use of analogies or similarities between the observed association and any other associations.

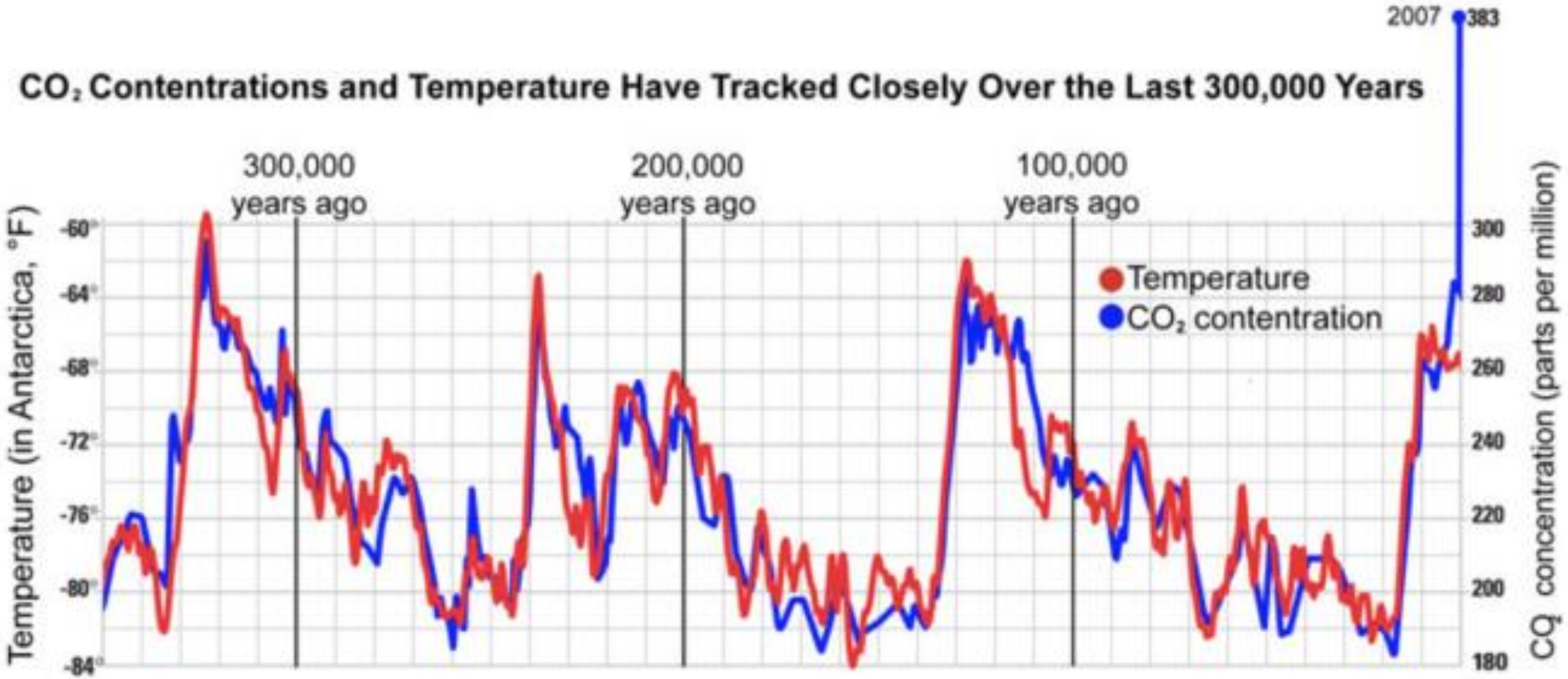
Wikipedia



Sir Austin Bradford Hill

Born	8 July 1897 Hampstead, London, England
Died	18 April 1991 (aged 93) Ulverston, Cumbria, England
Nationality	British
Occupation	Epidemiologist statistician
Known for	"Bradford Hill" criteria
Awards	Guy Medal (Gold, 1953)

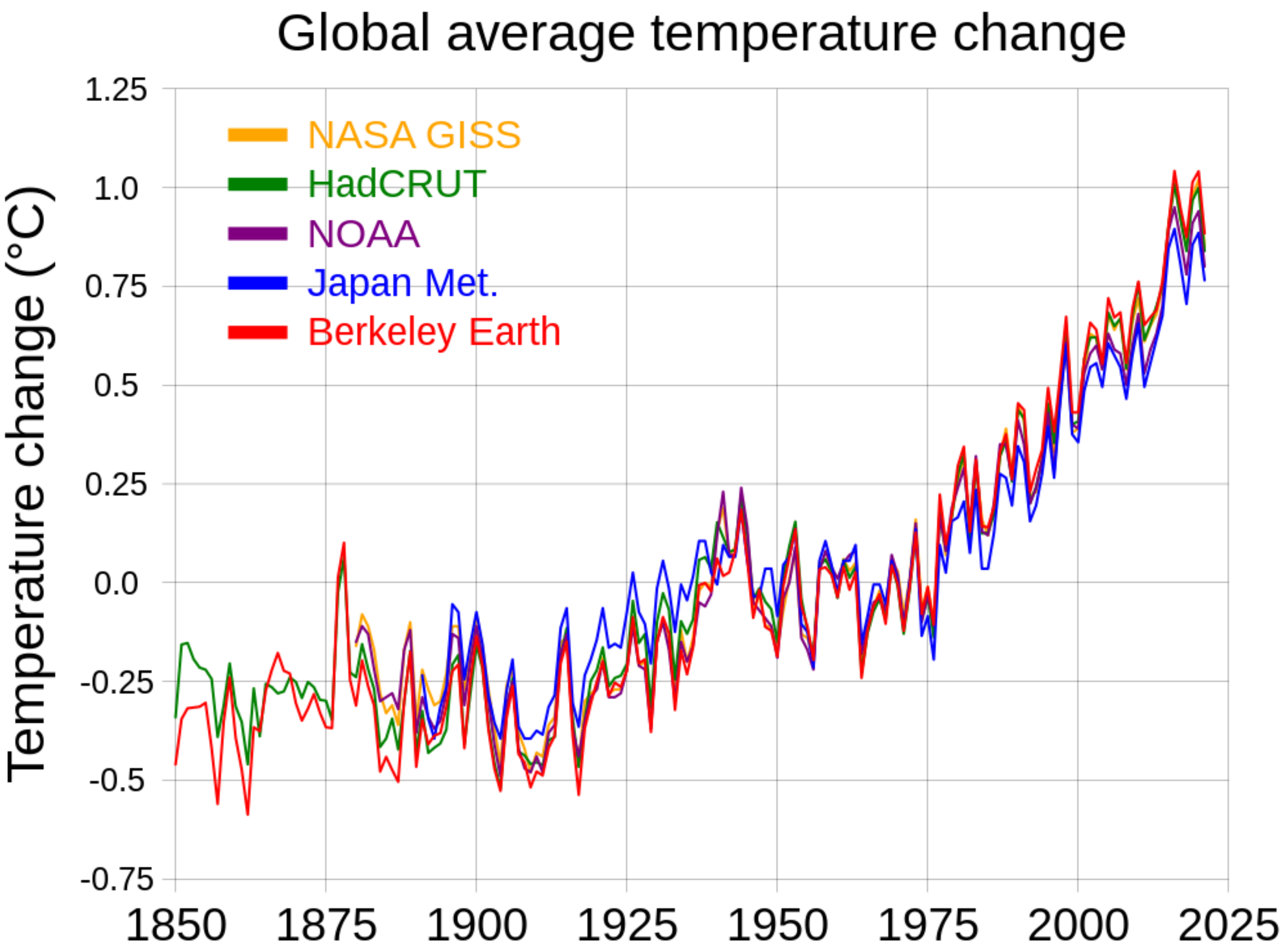
Criteriul 1: Intensitatea corelației



Sursa

Miller

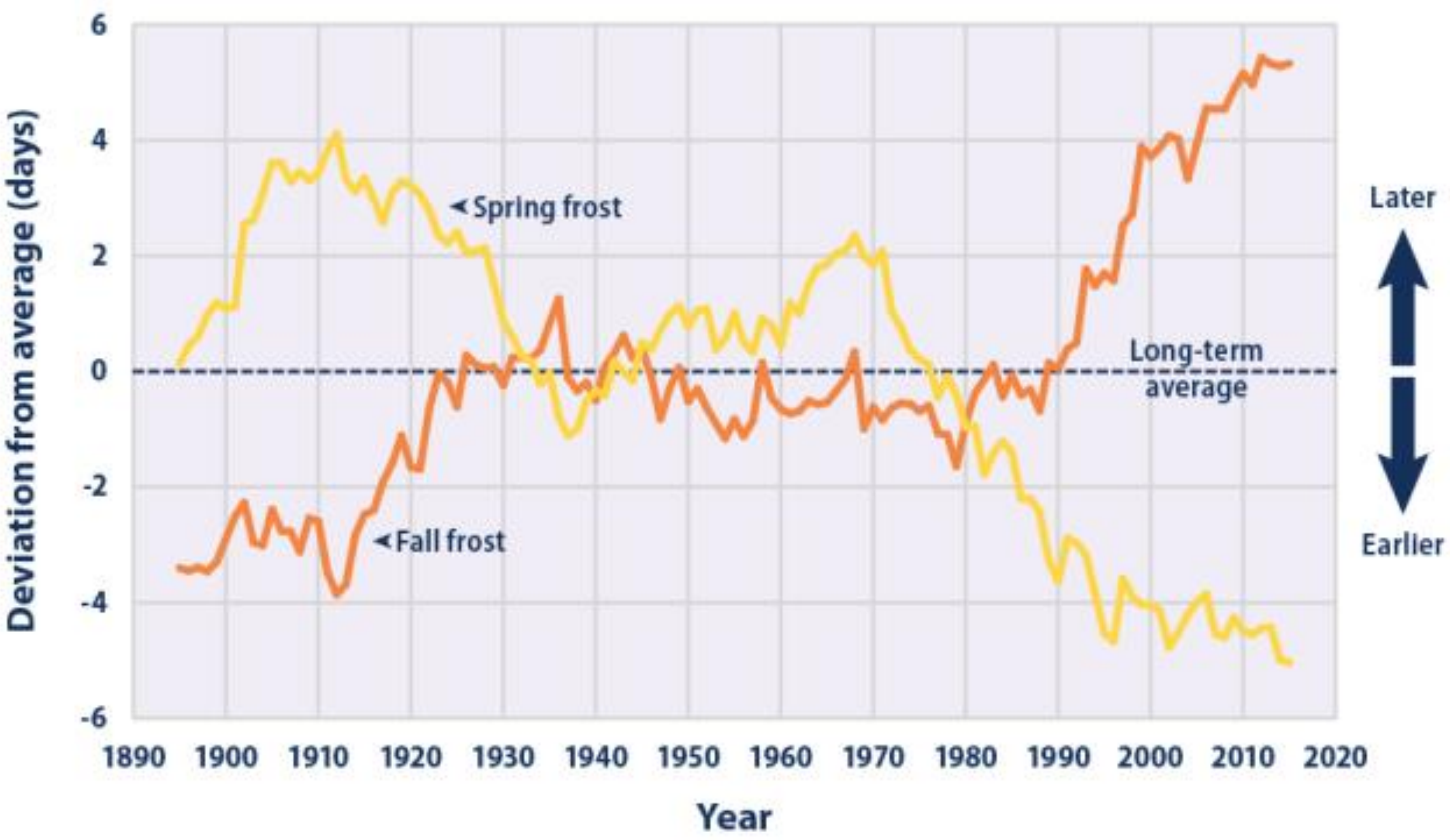
Criteriul 2: Consistența între surse diferite



Sursa

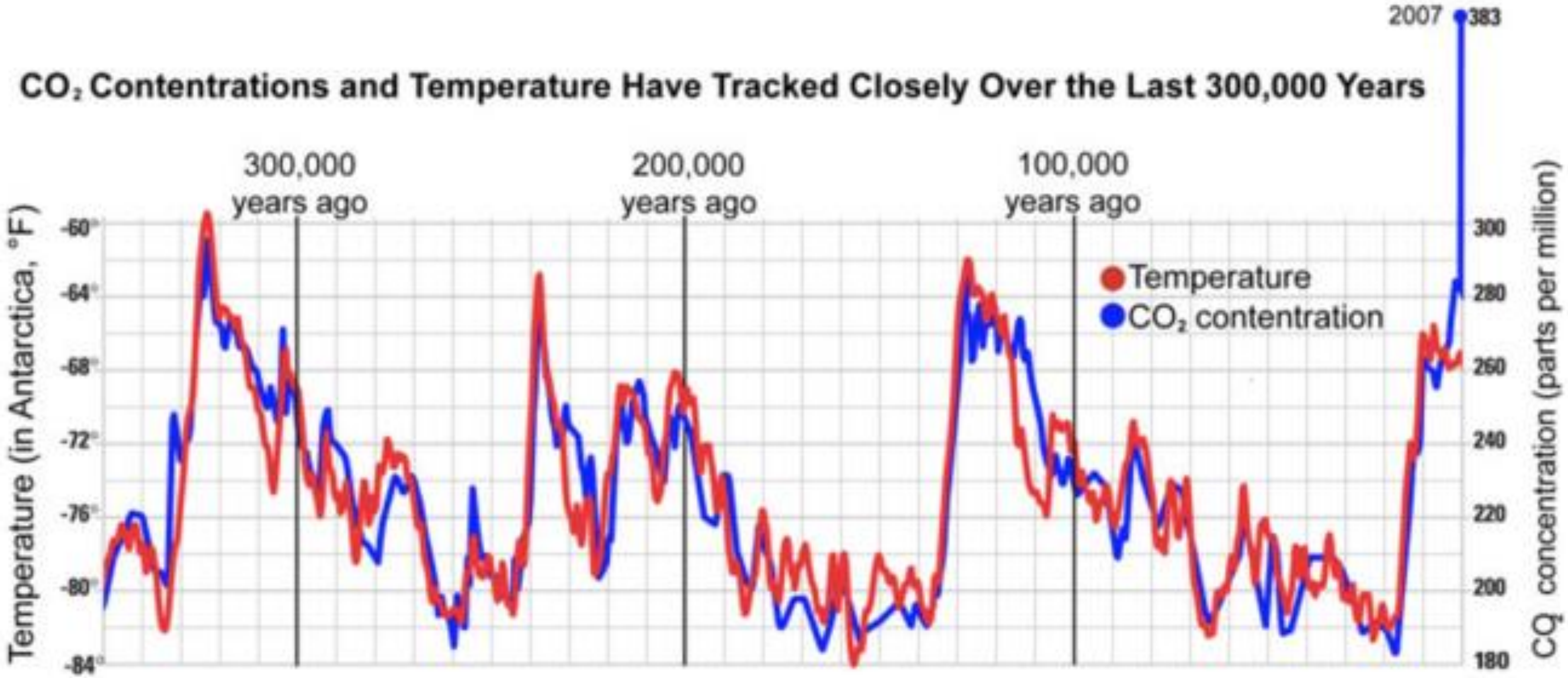
Miller

Criteriul 2: O altă sursă de date: sezonul agricol în SUA



Miller

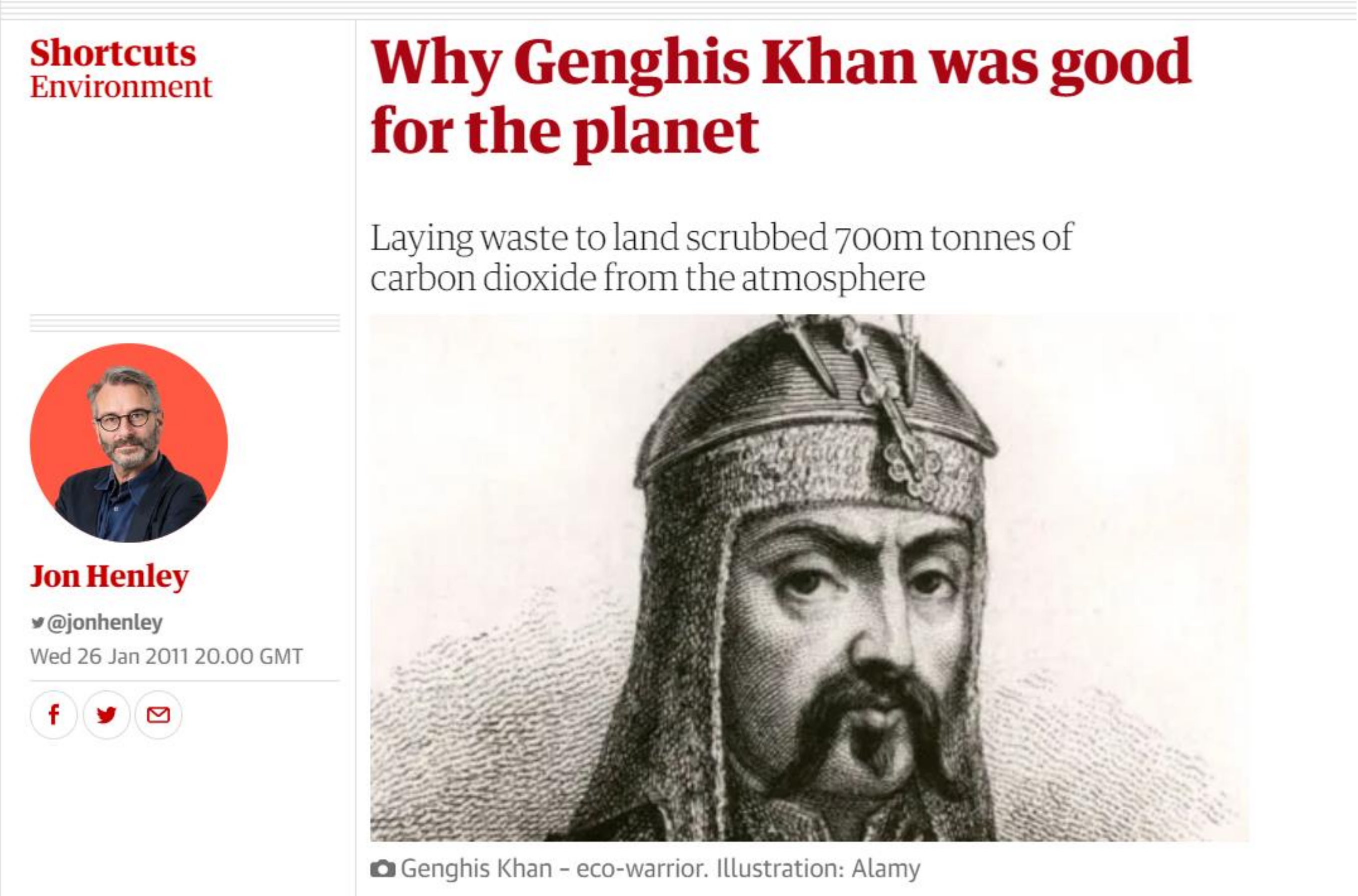
Criteriul 3: Specificitatea pentru o perioadă



Miller

Criteriul 8: Dovezi experimentale

„His empire lasted a century and a half and eventually covered nearly a quarter of the earth's surface. His murderous Mongol armies were responsible for the massacre of as many as 40 million people. Even today, his name remains a byword for brutality and terror. But boy, was Genghis green.”



Shortcuts
Environment

Why Genghis Khan was good for the planet

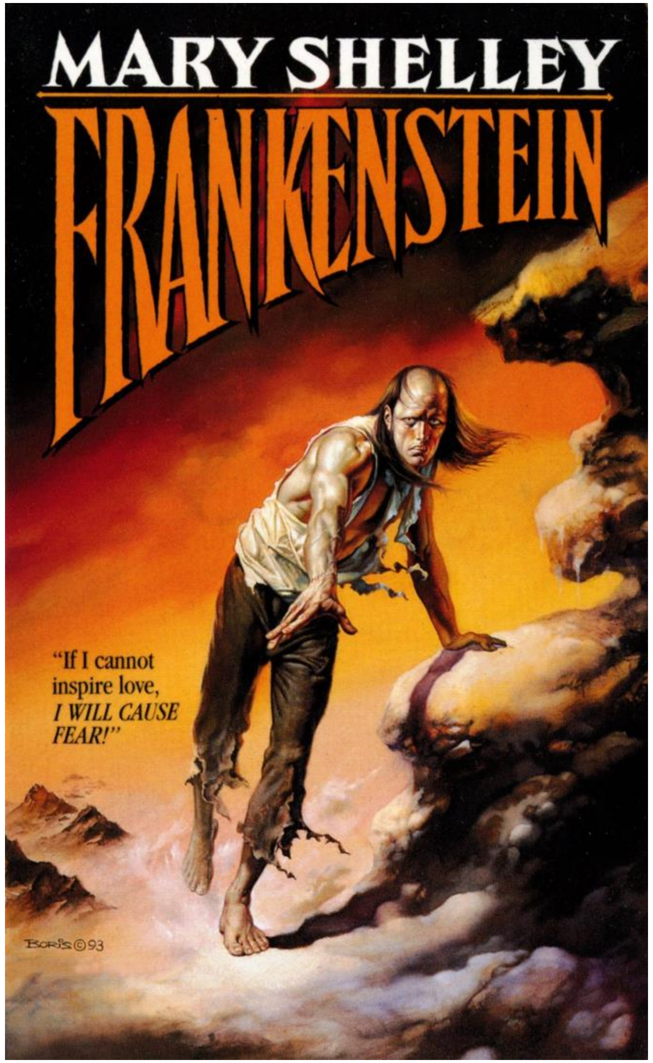
Laying waste to land scrubbed 700m tonnes of carbon dioxide from the atmosphere

Jon Henley
@jonhenley
Wed 26 Jan 2011 20.00 GMT

Genghis Khan - eco-warrior. Illustration: Alamy

Sursa

Criteriul 8: Erupția din 1815 (Tambora, Indonezia)



Draisine, 1817

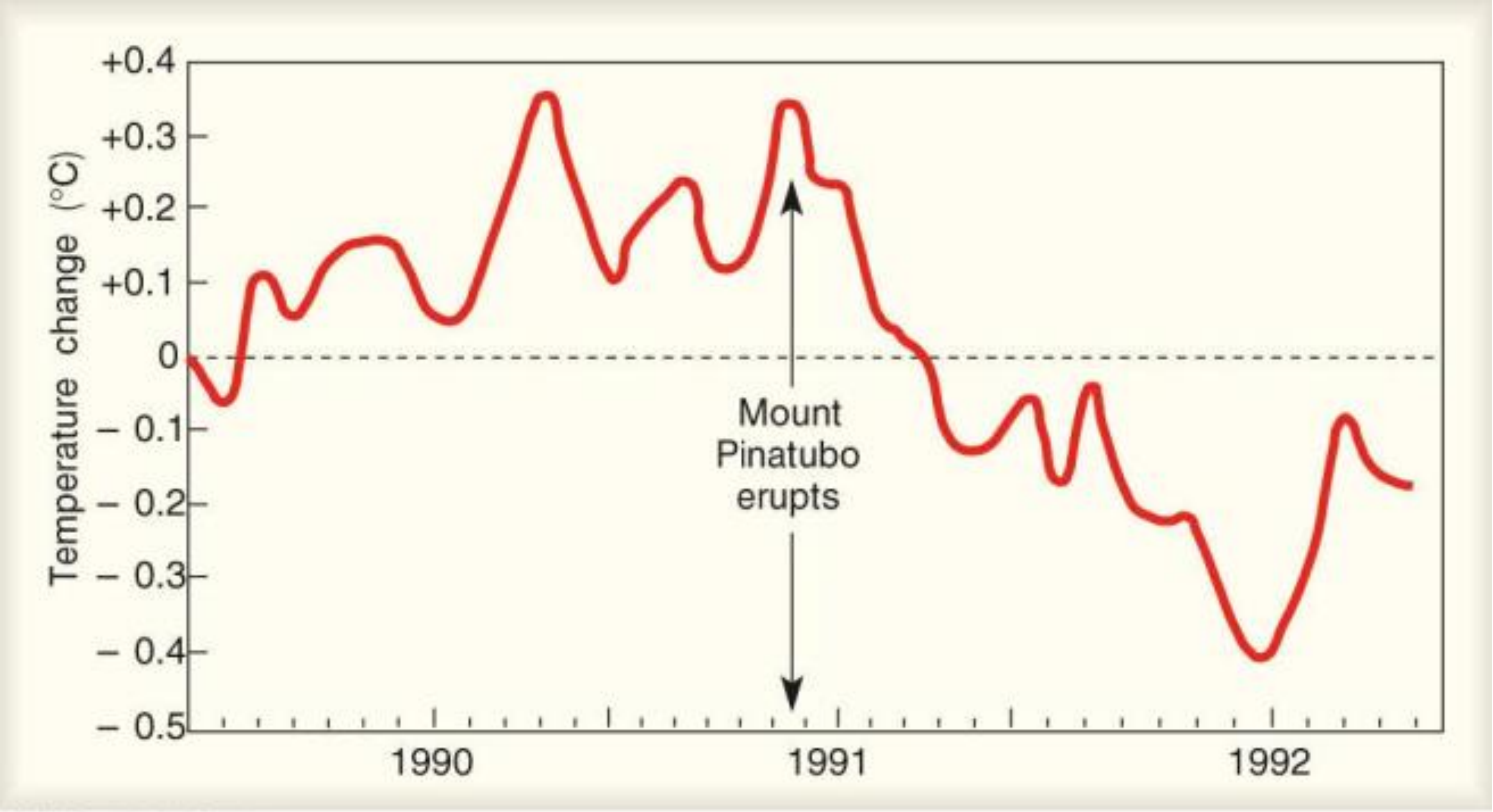
Criteriul 8: 1815 - anul fără vară

- „April 5, 1815: Tambora eruption (Indonesia)
- In parts of North America and Europe temperatures dropped by more than eighteen degrees Fahrenheit.
- There was snow in New England in July, and dark rain clouds swept over Europe throughout the summer months.
- Food shortages compounded those already in place in the wake of the Napoleonic wars.
- Record numbers of people starved to death in Paris in 1816, in what would be Europe's last major subsistence crisis."

- „**Mary Shelley traveled to Geneva in April 1816**, accompanied by her half sister, Claire Clairmont, and her lover, Percy Bysshe Shelley.
- **Barely five hundred kilometers north of Byron's villa, Baron Karl Drais**, a student of mathematics and a self-styled inventor, was witnessing first-hand the effects of the shift in climate.
- Crops, which had had no chance to recover after being ransacked by Napoleon's armies, failed across the area, and the subsequent oat shortage led to the starvation of humans and livestock alike.
- Historians agree that it was **this sudden dearth of horses that led Drais to turn to turn to man-powered transport."**

Source

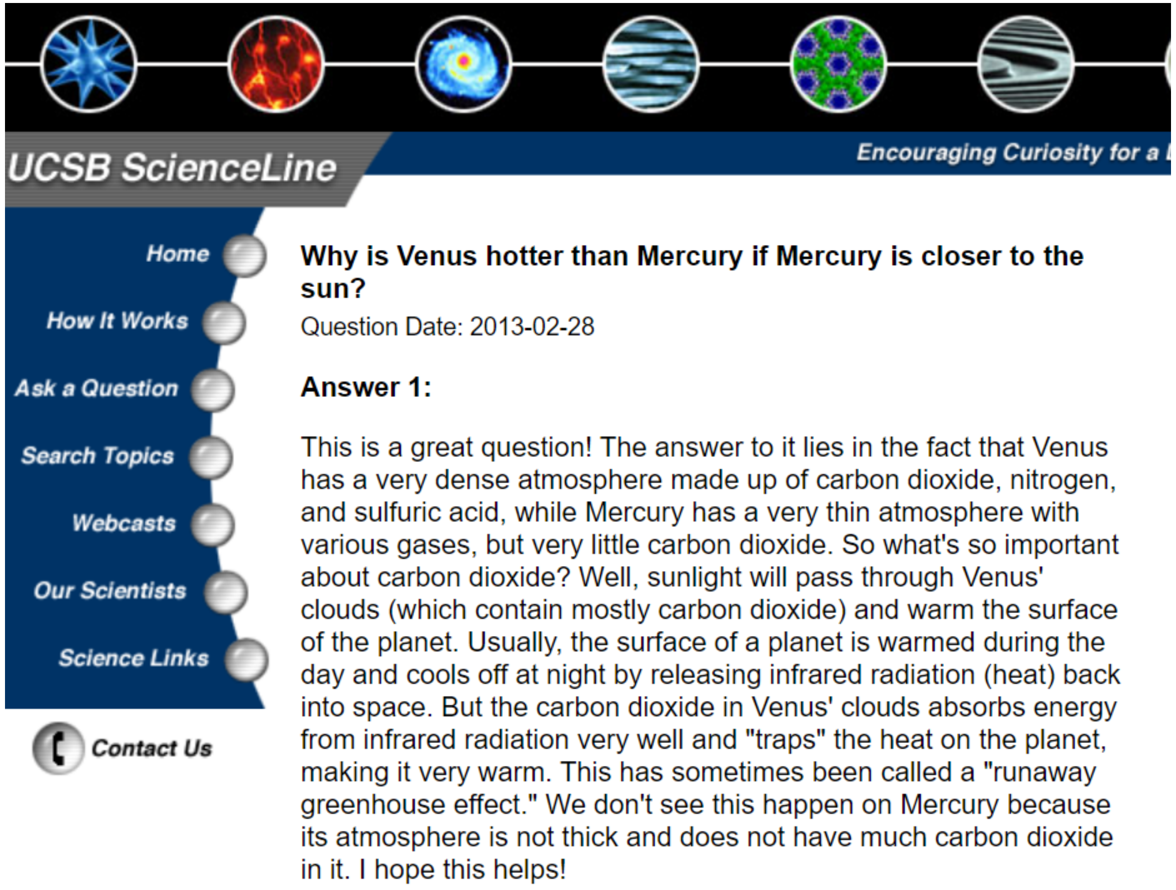
Criteriul 8 – Erupția din 1991 (Pinatubo, Filipine)



© 2007 Thomson Higher Education

Miller

Criteriul 9: Analogia



Why is Venus hotter than Mercury if Mercury is closer to the sun?

Question Date: 2013-02-28

Answer 1:

This is a great question! The answer to it lies in the fact that Venus has a very dense atmosphere made up of carbon dioxide, nitrogen, and sulfuric acid, while Mercury has a very thin atmosphere with various gases, but very little carbon dioxide. So what's so important about carbon dioxide? Well, sunlight will pass through Venus' clouds (which contain mostly carbon dioxide) and warm the surface of the planet. Usually, the surface of a planet is warmed during the day and cools off at night by releasing infrared radiation (heat) back into space. But the carbon dioxide in Venus' clouds absorbs energy from infrared radiation very well and "traps" the heat on the planet, making it very warm. This has sometimes been called a "runaway greenhouse effect." We don't see this happen on Mercury because its atmosphere is not thick and does not have much carbon dioxide in it. I hope this helps!

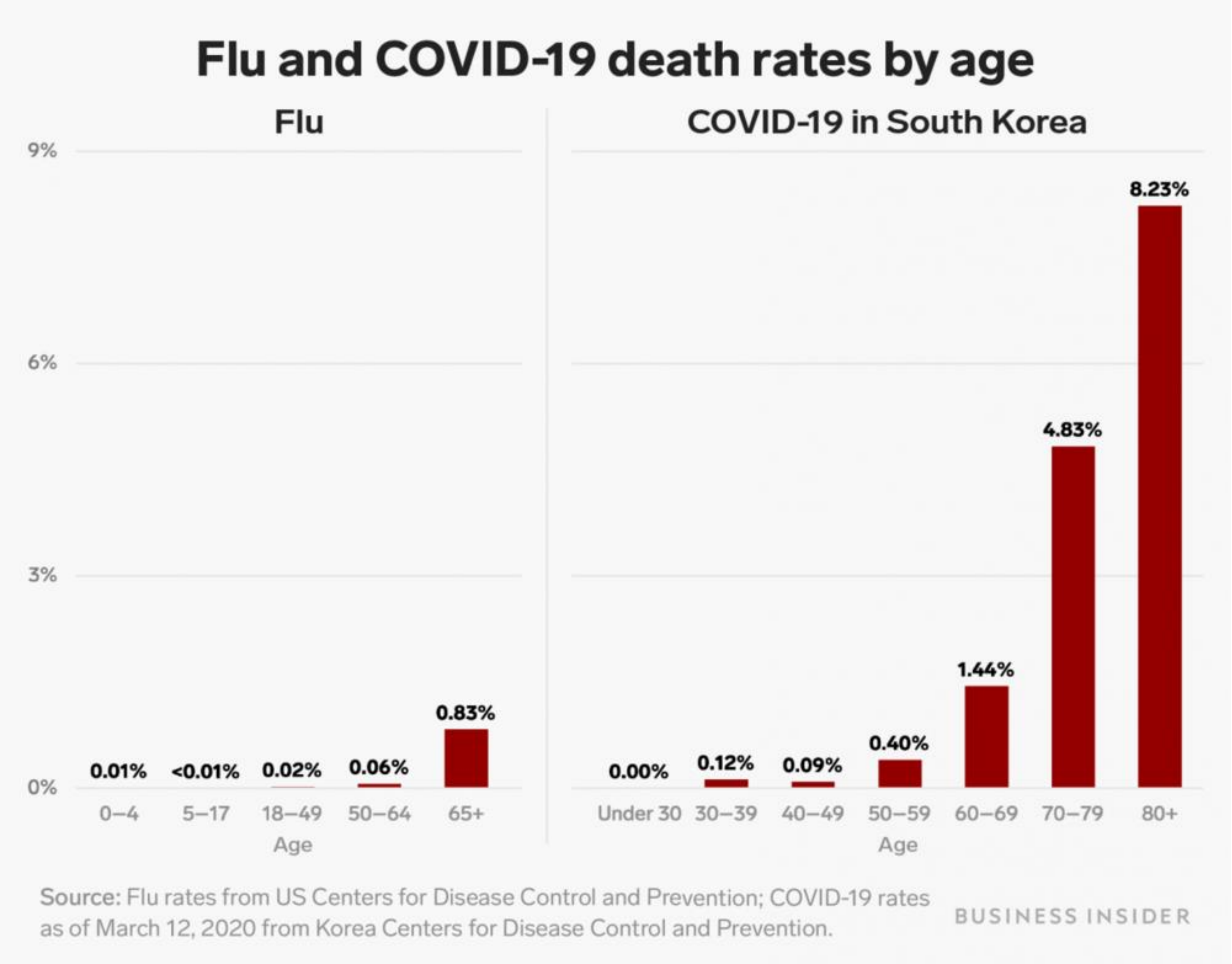
Miller

Manipularea prin date

Vizualizări exagerate

Vizualizări ciuntite

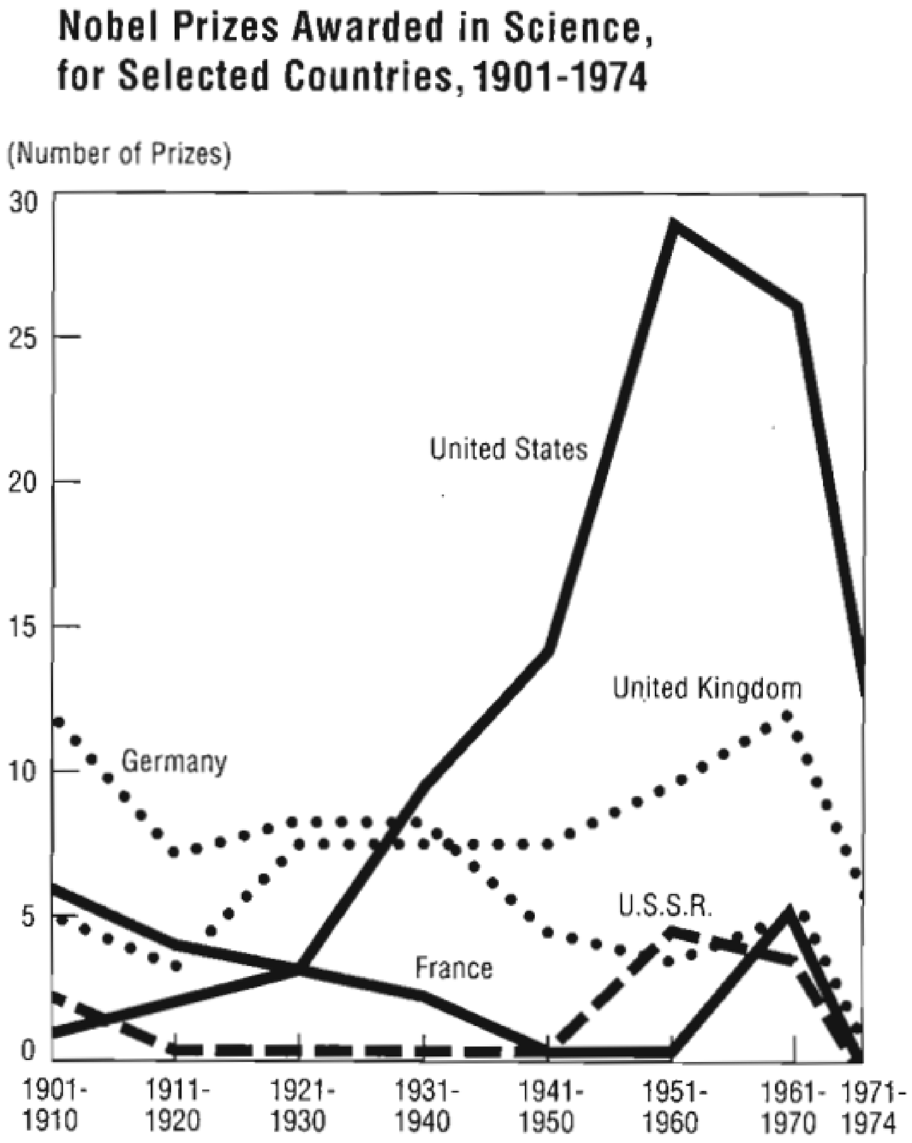
Axe diferite



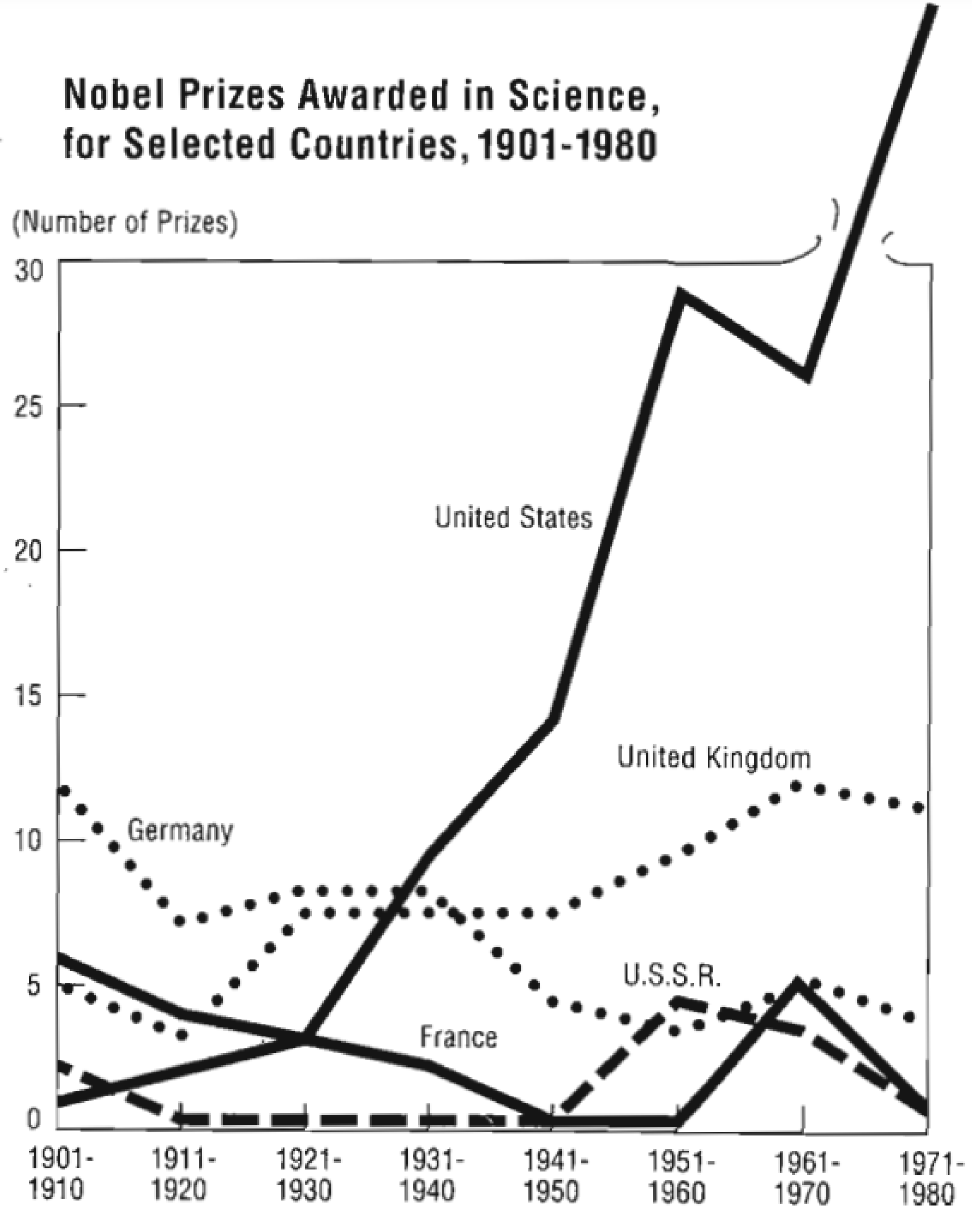
Source: Flu rates from US Centers for Disease Control and Prevention; COVID-19 rates as of March 12, 2020 from Korea Centers for Disease Control and Prevention.

Sursa

Grafic „ciuntit” la dreapta



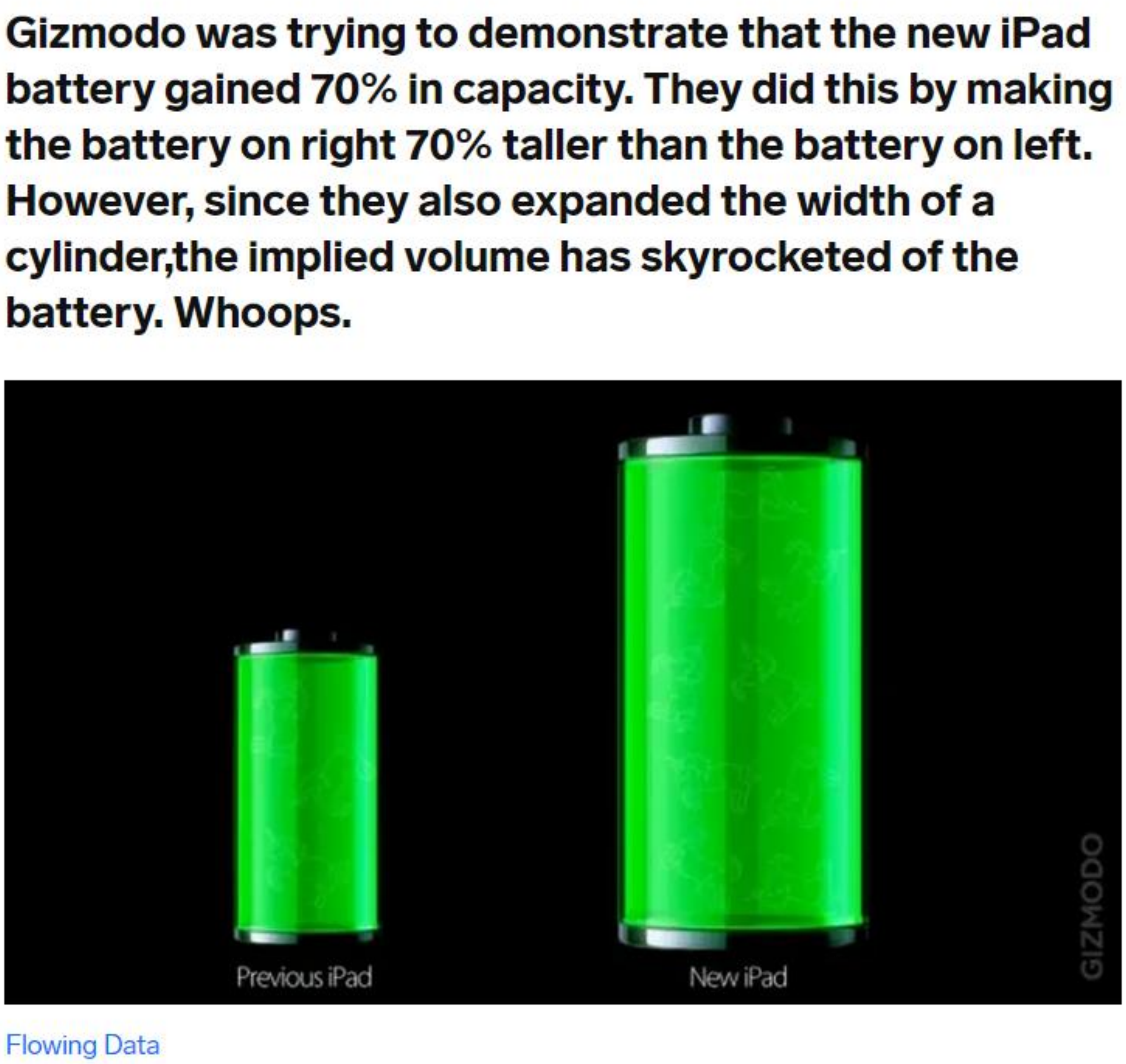
Nobel Prizes Awarded in Science, for Selected Countries, 1901-1974



Nobel Prizes Awarded in Science, for Selected Countries, 1901-1980

National Science Foundation, *Science Indicators, 1974* (Washington, D.C., 1976), p. 15.

Sursa: Tufte 2001, p. 60

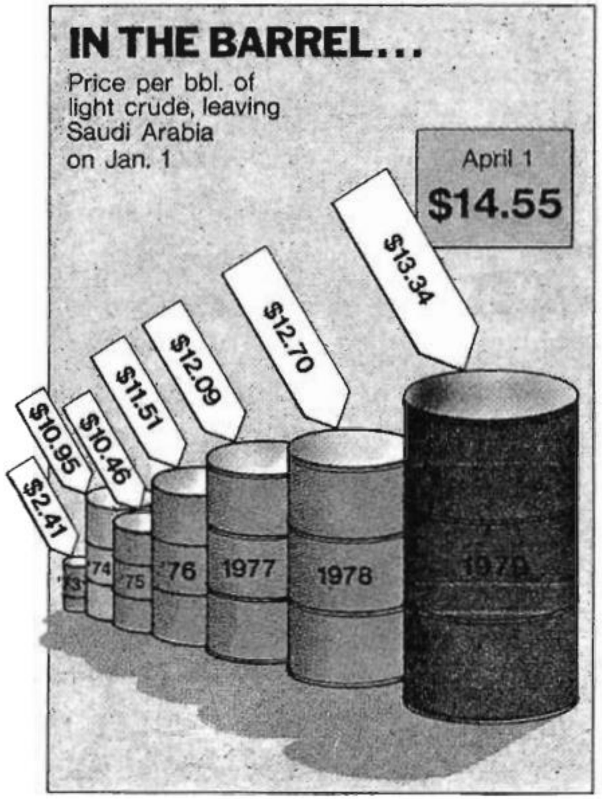


Gizmodo was trying to demonstrate that the new iPad battery gained 70% in capacity. They did this by making the battery on right 70% taller than the battery on left. However, since they also expanded the width of a cylinder,the implied volume has skyrocketed of the battery. Whoops.

Flowing Data

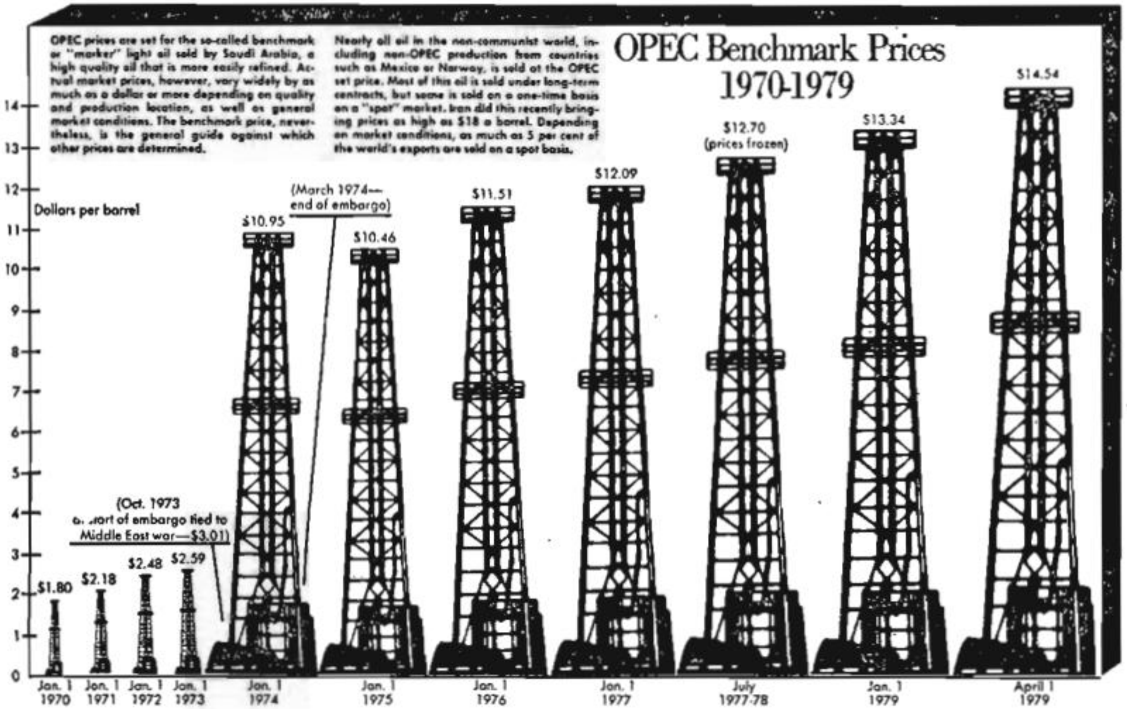
Sursa

Design variation infected similar graphics in other publications. Here an increase of 454 percent is depicted as an increase of 4,280 percent, for a Lie Factor of 9.4:



Time, April 9, 1979, p. 57.

And an increase of 708 percent is shown as 6,700 percent, for a Lie Factor of 9.5:

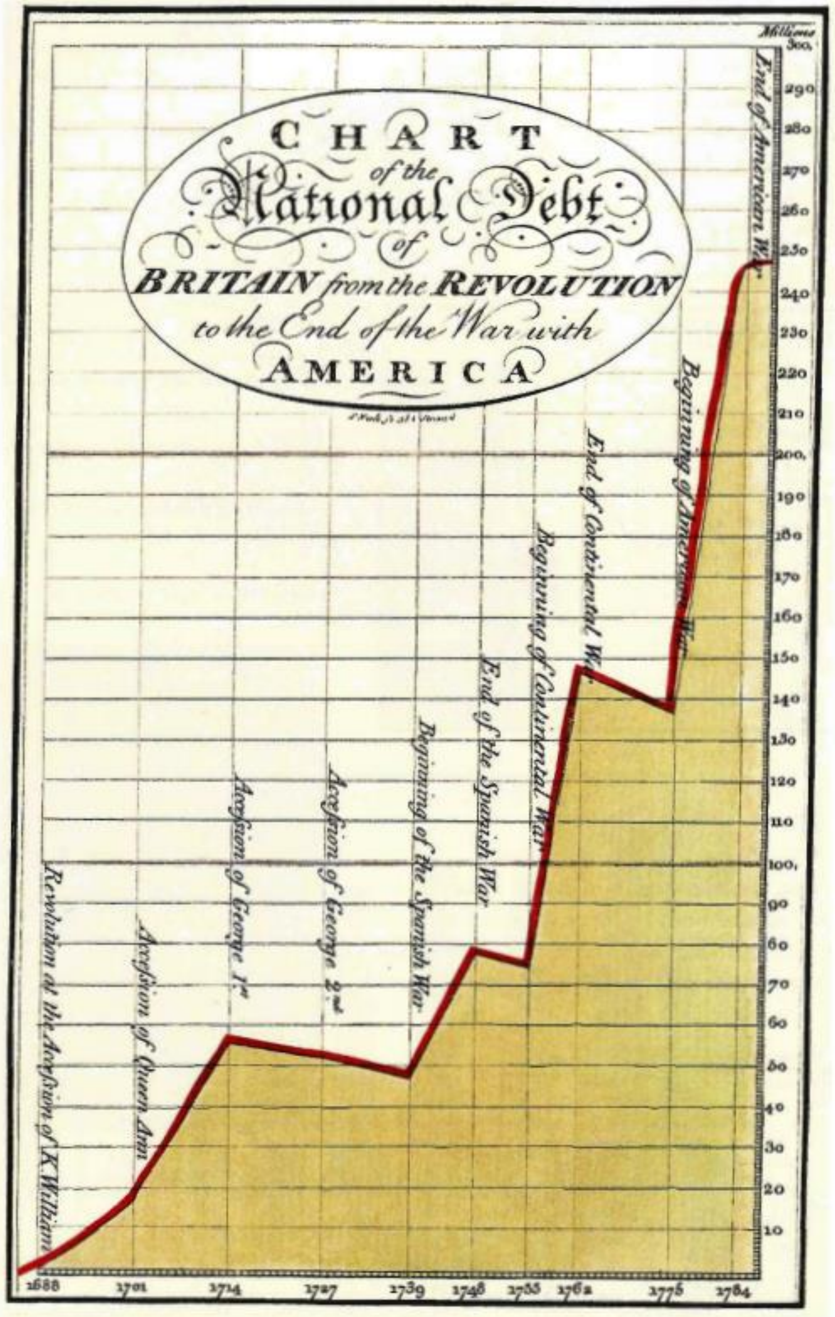

OPEC prices are set for the so-called benchmark or "marker" light oil sold by Saudi Arabia, a high quality oil that is more easily refined. Actual market prices, however, vary widely by as much as a dollar or more depending on quality and production location, as well as general market conditions. The benchmark price, nevertheless, is the general guide against which other prices are determined.

Nearly all oil in the non-communist world, including non-OPEC production from countries such as Mexico or Norway, is sold at the OPEC set price. Most of this oil is sold under long-term contracts, but some is sold on a one-time basis on a "spot" market. Iran did this recently bringing prices as high as $18 a barrel. Depending on market conditions, as much as 5 per cent of the world's exports are sold on a spot basis.

All these accounts of oil prices made a second error, by showing the price of oil in inflated (current) dollars. The 1972 dollar was worth much more than the 1979 dollar. Thus in sweeping from

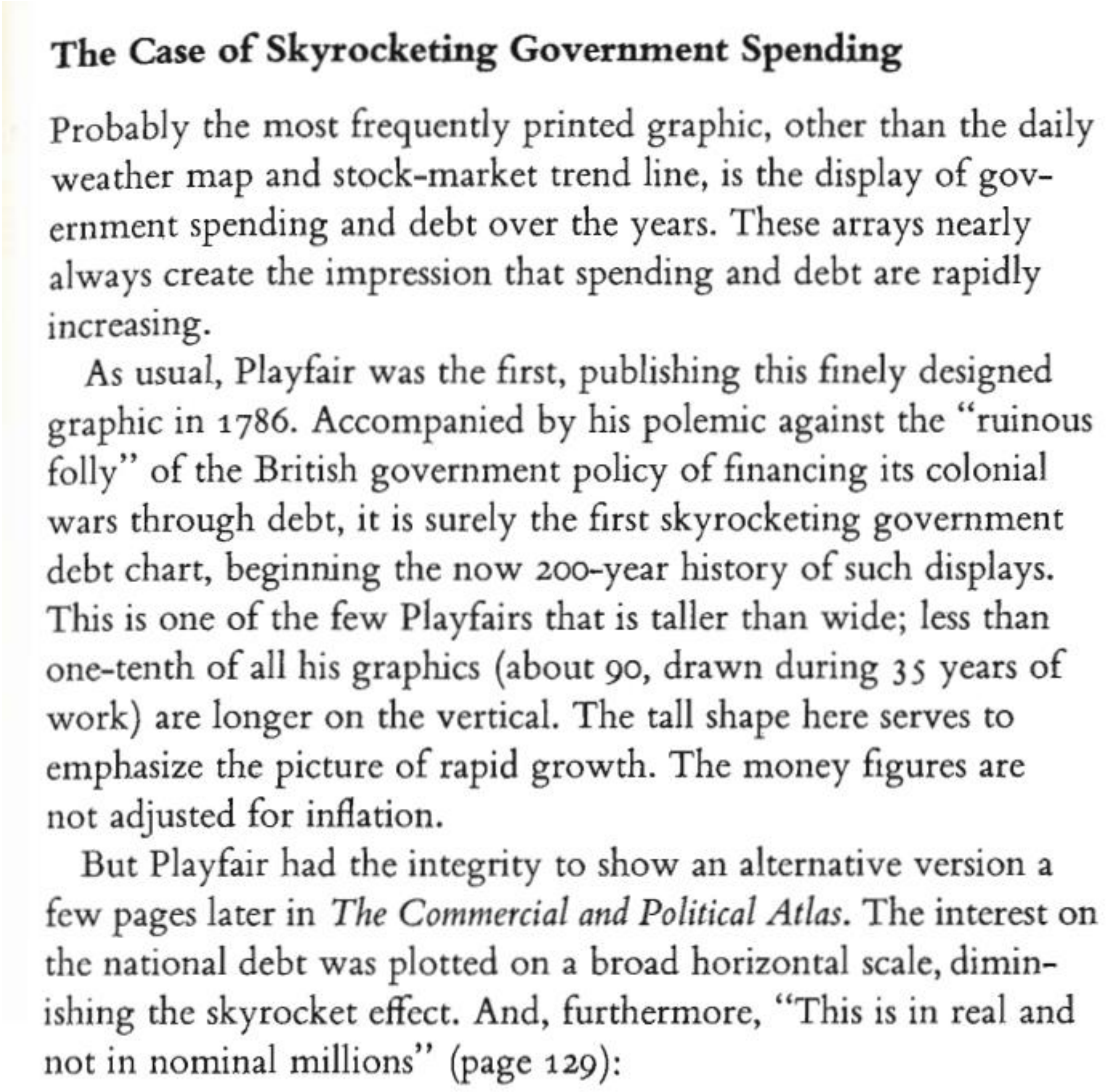
Washington Post, March 28, 1979, p. A-18.

Sursa: Tufte 2001, p. 62



The Divisions at the Bottom are Years, & those on the Right hand Money.

Comprimarea axei X produce o impresie de accelerare



The Case of Skyrocketing Government Spending

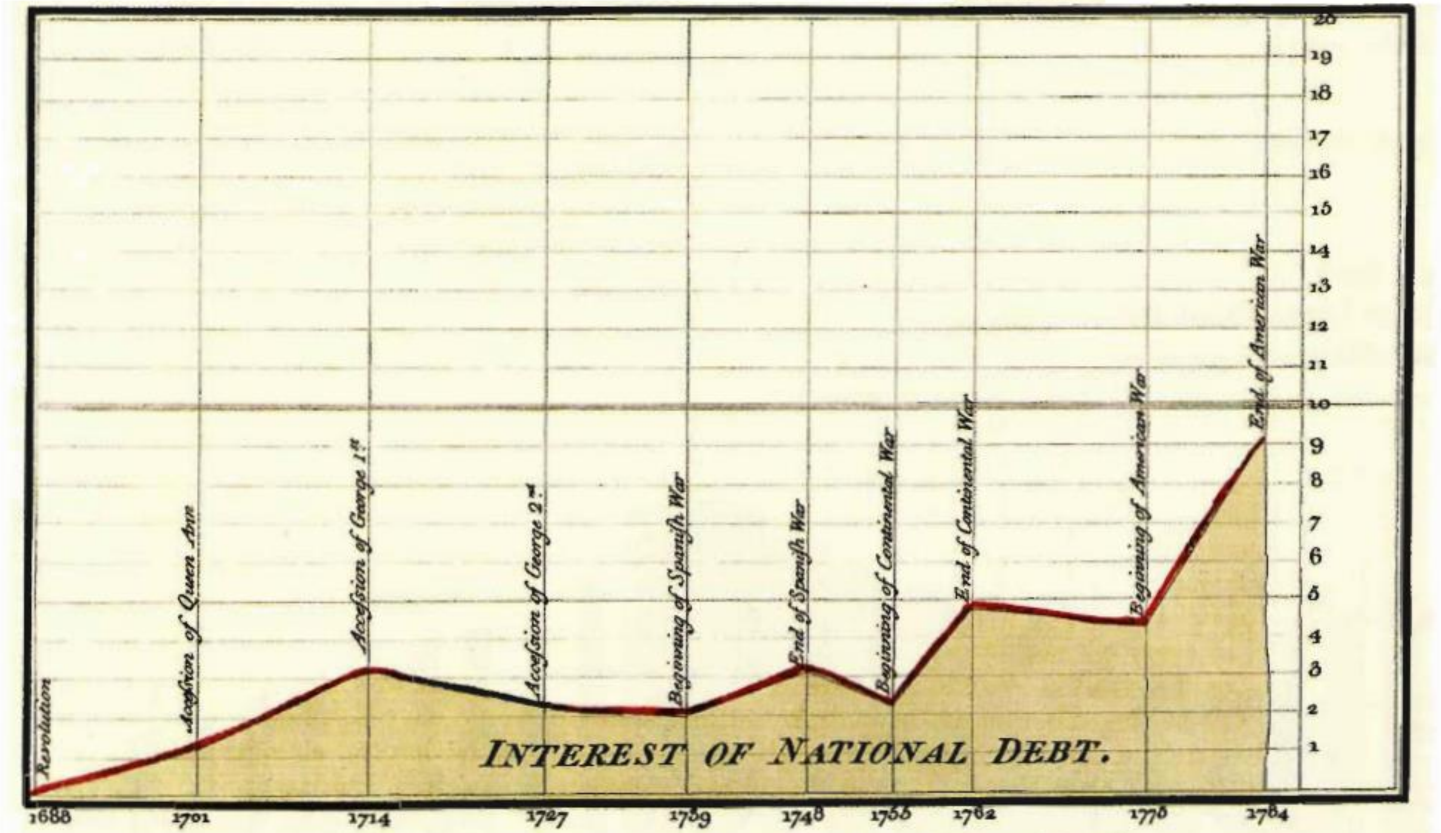
Probably the most frequently printed graphic, other than the daily weather map and stock-market trend line, is the display of government spending and debt over the years. These arrays nearly always create the impression that spending and debt are rapidly increasing.

As usual, Playfair was the first, publishing this finely designed graphic in 1786. Accompanied by his polemic against the "ruinous folly" of the British government policy of financing its colonial wars through debt, it is surely the first skyrocketing government debt chart, beginning the now 200-year history of such displays. This is one of the few Playfairs that is taller than wide; less than one-tenth of all his graphics (about 90, drawn during 35 years of work) are longer on the vertical. The tall shape here serves to emphasize the picture of rapid growth. The money figures are not adjusted for inflation.

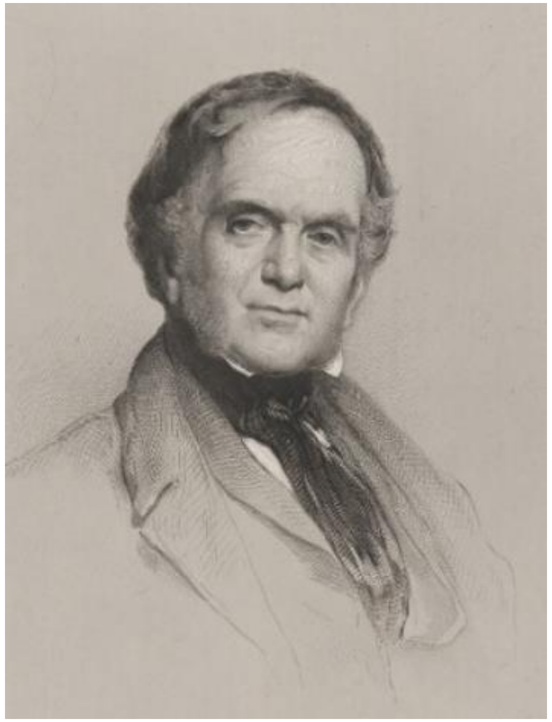
But Playfair had the integrity to show an alternative version a few pages later in *The Commercial and Political Atlas*. The interest on the national debt was plotted on a broad horizontal scale, diminishing the skyrocket effect. And, furthermore, "This is in real and not in nominal millions" (page 129):

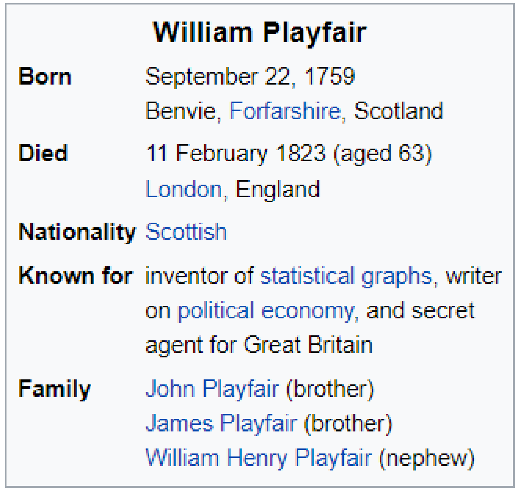
Sursa: Tufte 2001, p. 64

Interest of the NATIONAL DEBT from the Revolution.



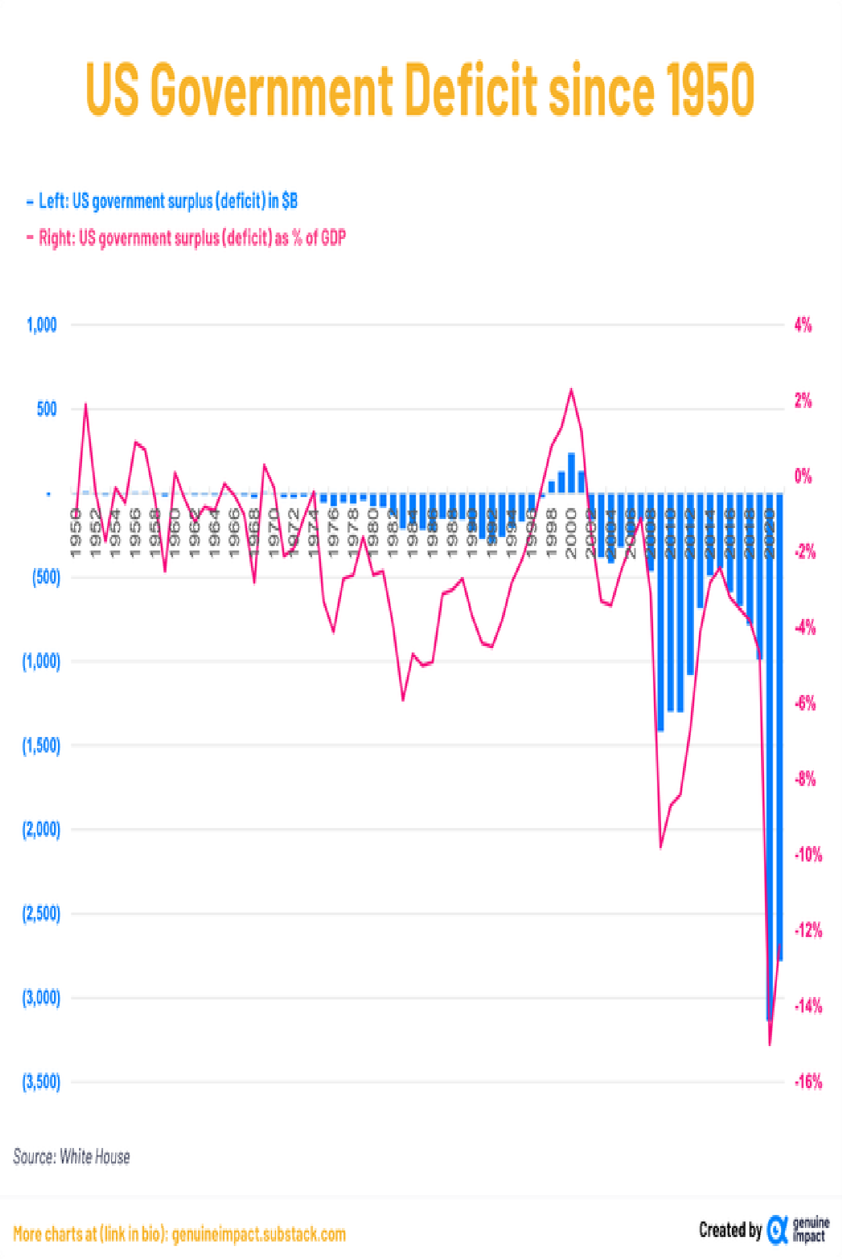
The Bottom line is Years, those on the Right hand Millions of Pounds.

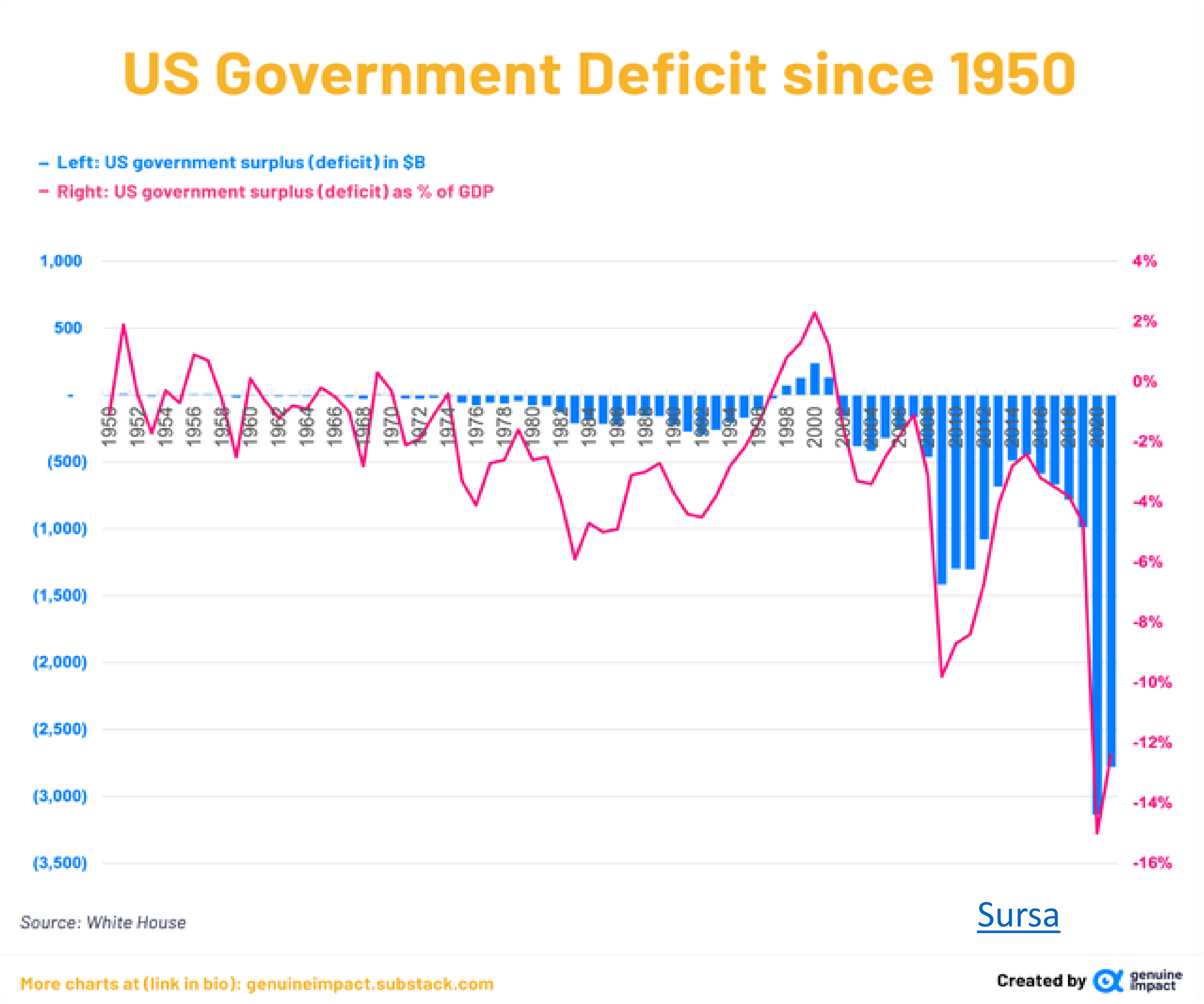




William Playfair

Born	September 22, 1759 Benvie, Forfarshire, Scotland
Died	11 February 1823 (aged 63) London, England
Nationality	Scottish
Known for	inventor of statistical graphs, writer on political economy, and secret agent for Great Britain
Family	John Playfair (brother) James Playfair (brother) William Henry Playfair (nephew)


US Government Deficit since 1950
- Left: US government surplus (deficit) in $B
- Right: US government surplus (deficit) as % of GDP
1,000
500
-
(500)
(1,000)
(1,500)
(2,000)
(2,500)
(3,000)
(3,500)
4%
2%
0%
-2%
-4%
-6%
-8%
-10%
-12%
-14%
-16%
1950
1952
1954
1956
1958
1960
1962
1964
1966
1968
1970
1972
1974
1976
1978
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
Source: White House
More charts at (link in bio): genuineimpact.substack.com
Created by genuine impact


US Government Deficit since 1950
- Left: US government surplus (deficit) in $B
- Right: US government surplus (deficit) as % of GDP
1,000
500
-
(500)
(1,000)
(1,500)
(2,000)
(2,500)
(3,000)
(3,500)
4%
2%
0%
-2%
-4%
-6%
-8%
-10%
-12%
-14%
-16%
1950
1952
1954
1956
1958
1960
1962
1964
1966
1968
1970
1972
1974
1976
1978
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
Source: White House
More charts at (link in bio): genuineimpact.substack.com
Created by genuine impact

Sursa

Concluzii

- Cauzalitatea este atribuită prin...
  - Co-incidență
  - Asociere și corelație (co-incidențe repetate)
- O atribuire robustă ține cont de criteriile lui Hill
  - Exemplu: efectul de seră și încălzirea globală
- Prezentarea datelor poate fi manipulativă
  - Implicarea unei cauzalități iluzorii
  - Implicarea unui exces sau deficit exagerat

Referințe

1. Hugh Small, Florence Nightingale's forgotten legacy: Public Health laws. Personal blog
2. The Economist. Worth a Thousand Words. 2003 [Site]
3. Edward R. Tufte, The graphical display of quantitative information. Graphics Press. 2001
4. Edward R. Tufte, Beautiful evidence. Graphics Press, 2006.
5. Betsy Mason, The Underappreciated Man Behind the “Best Graphic Ever Produced”, National Geographic, 2017
6. Topi Tjukanov. Notable people. 2022
7. Neil Halloran. The Fallen of WW2. 2015
8. Zach Gemignani. 20 Best examples of charts and graphics.
9. Tableau. Data is beautiful: 10 of the best data visualisation examples from history to today
10. Gapminder
11. Walt Hickey. The 27 worst charts of all time. Business Insider, 2013.
12. Seth Miller. What climate skeptics taught me about global warming. Medium, 2017
13. Phillips DP, Barker GE. A July spike in fatal medication errors: a possible effect of new medical residents. Journal of general internal medicine. 2010 Aug;25(8):774-9.