Curs 02 – Cauzalitate

Co-incidență, asociere, corelație

Atribuiri robuste: criteriile lui Hill

Manipulări

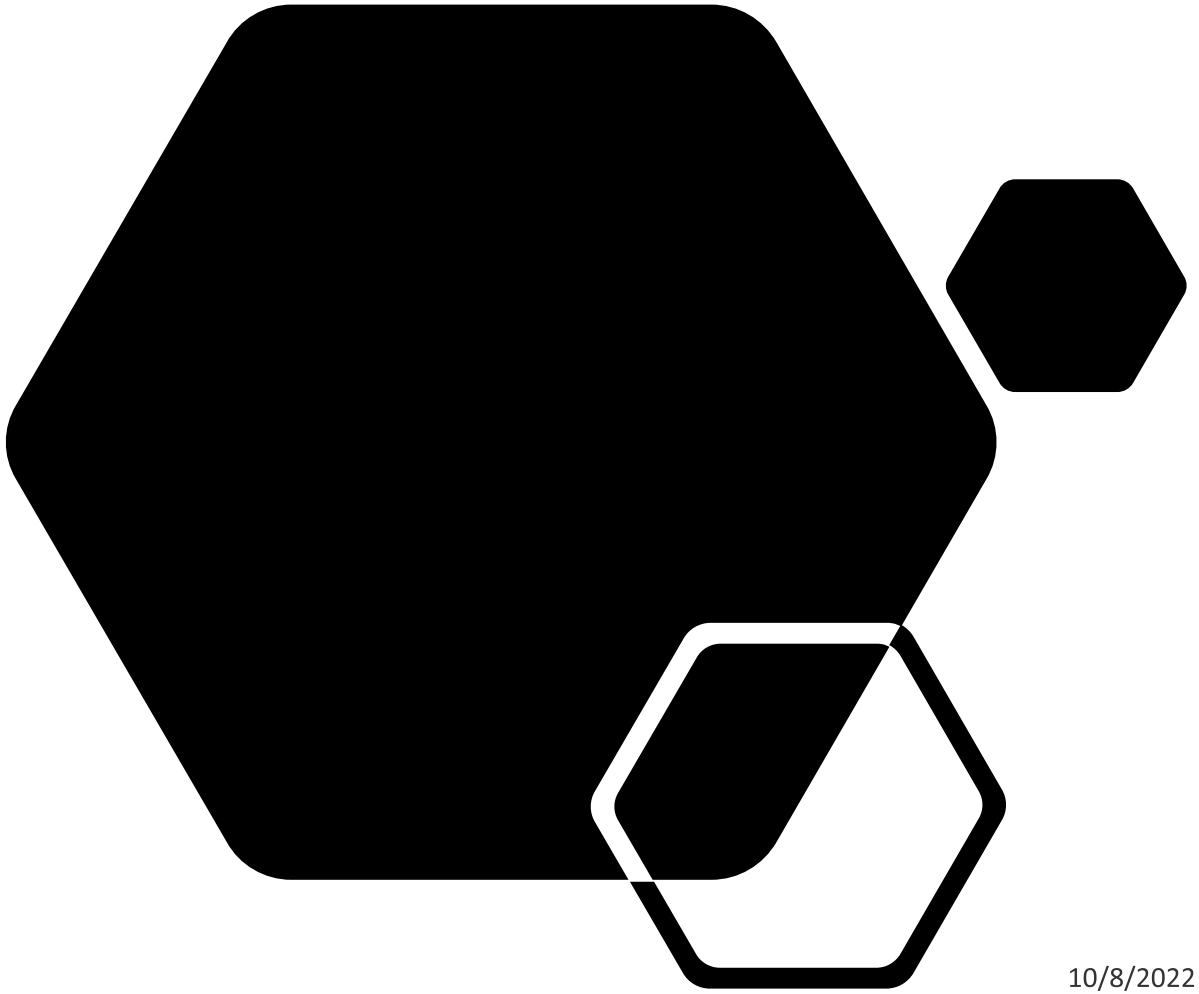
10/8/2022

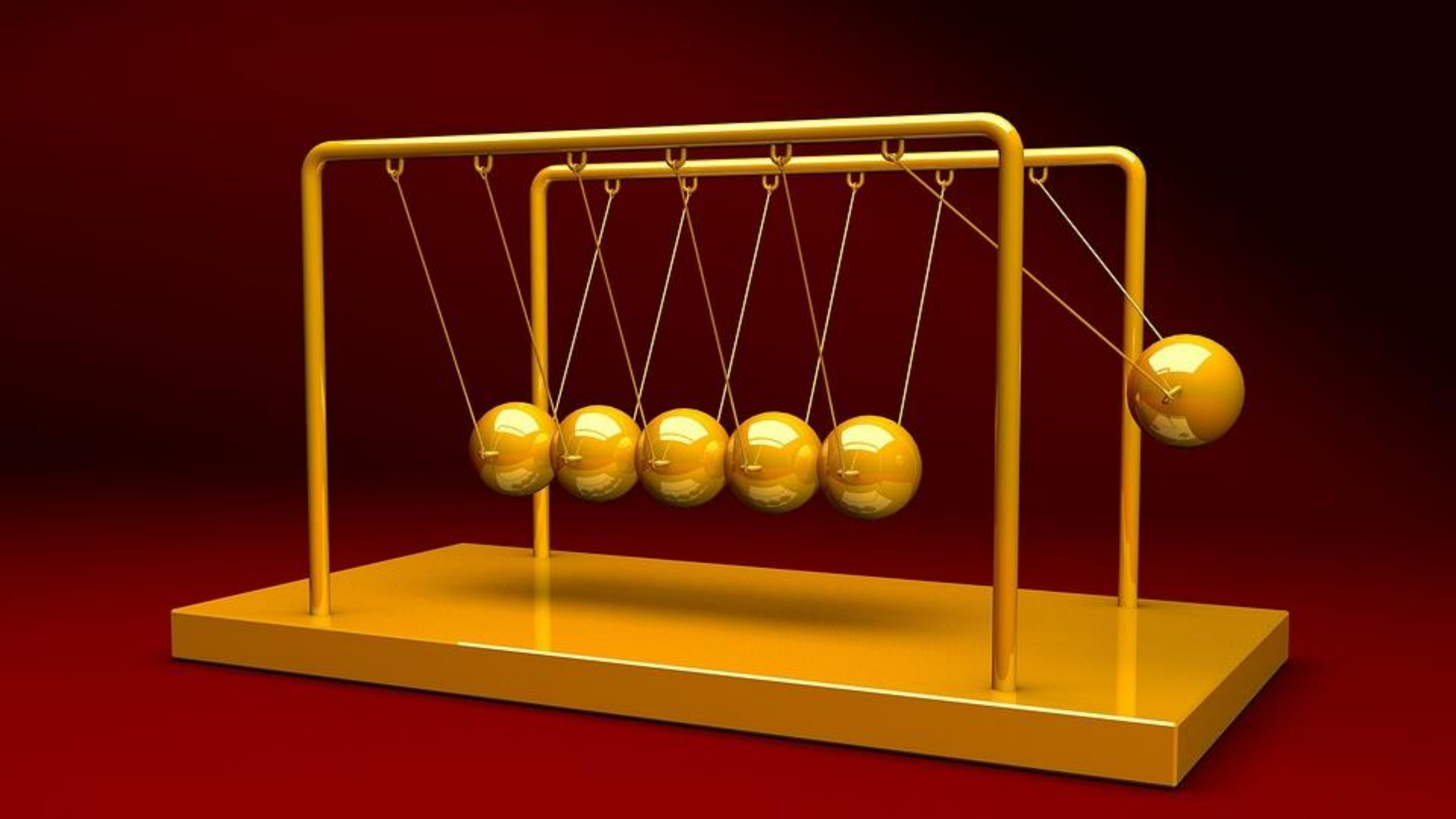
Structura cursului

1. Why?
2. Cauzalitate
   - Co-incidență, asociere, corelație
   - Atribuiri robuste: criteriile lui Hill
   - Manipulări
3. Măsurare
4. Modelare și eșantionare
5. Tehnici de analiză
6. Predicție
7. Programare și ML
8. ML și Deep Learning
9. Producția
10. Why Privacy?
11. Privacy Preserving Algorithms
12. Privacy Architectures and Federated Learning

Co-incidență, asociere, corelație

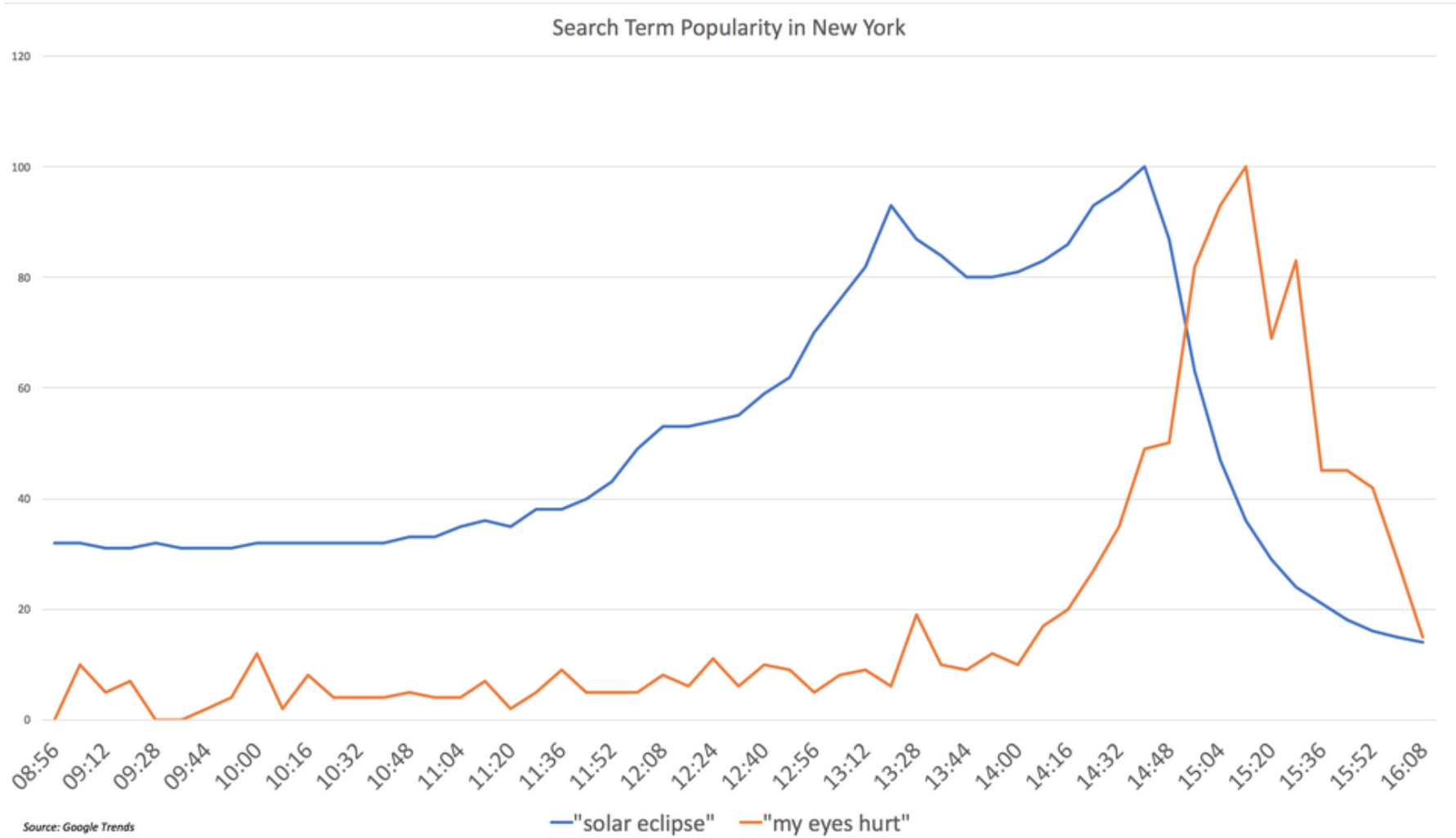
Atribuiri cauzale





Co-incidență

2017



Sursa

Asociere 2011

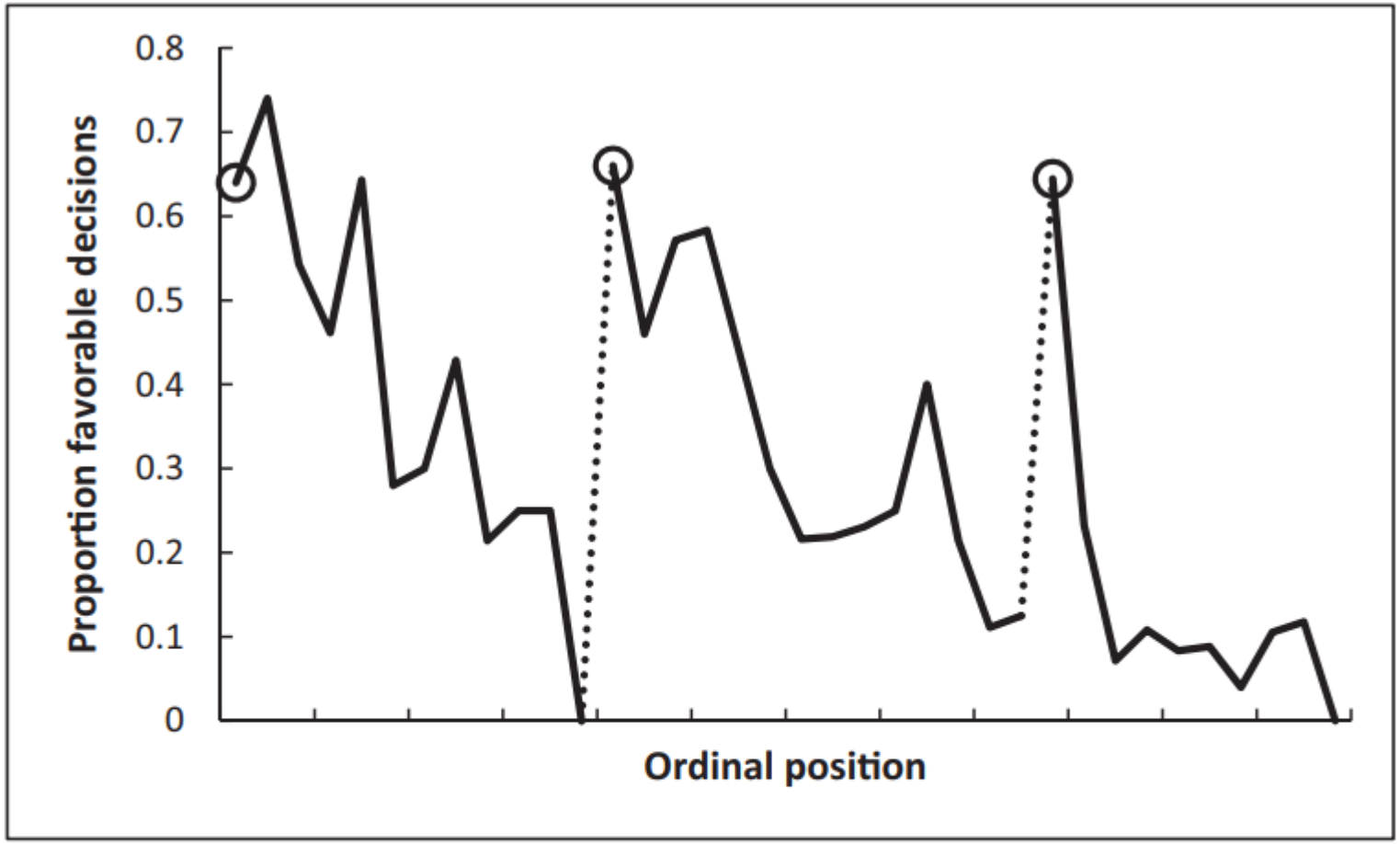
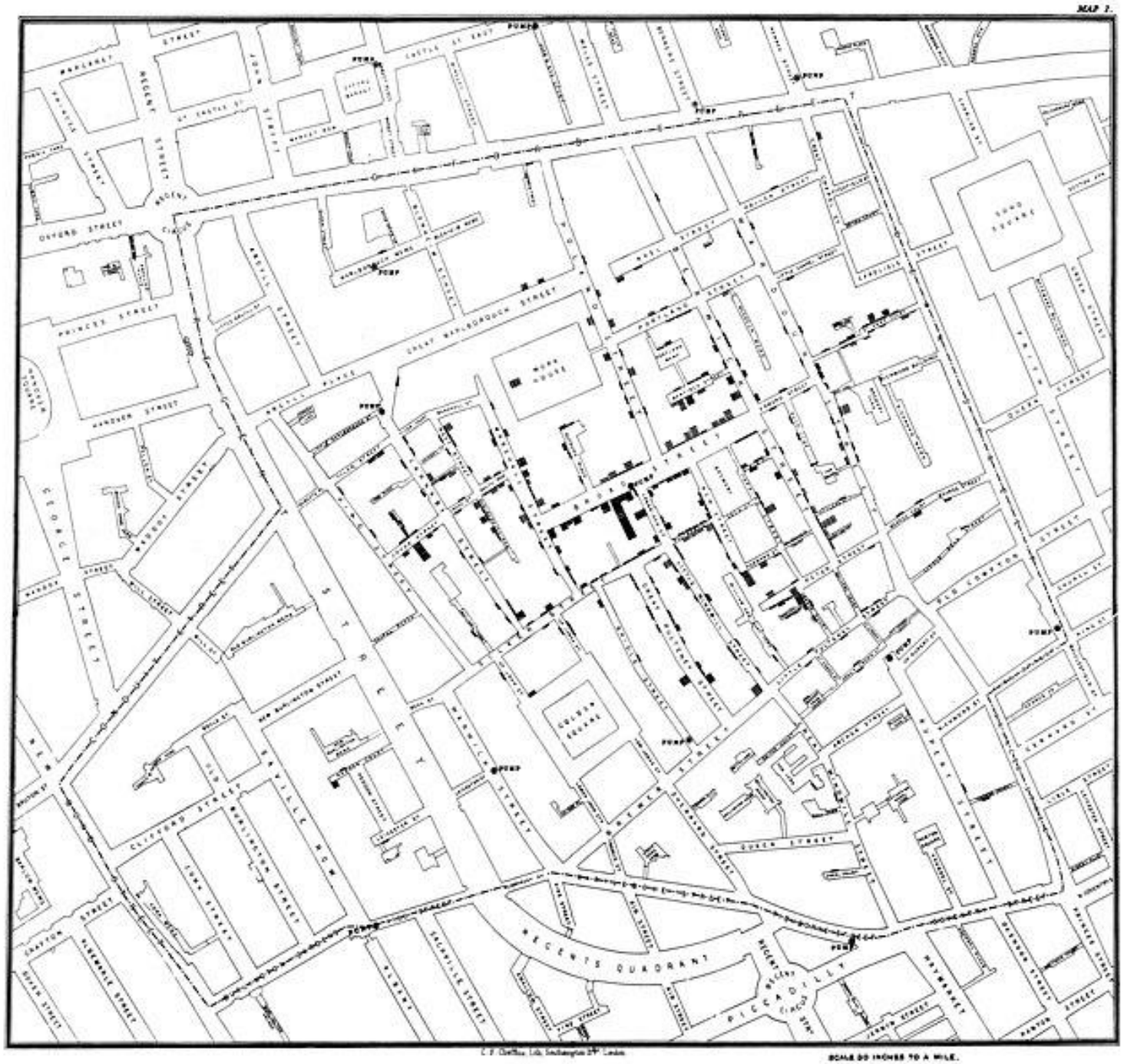


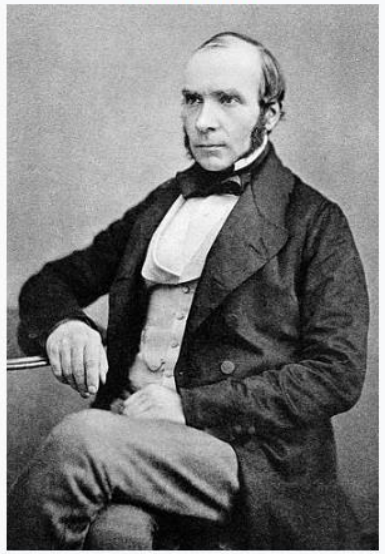
Fig. 1. Proportion of rulings in favor of the prisoners by ordinal position. Circled points indicate the first decision in each of the three decision sessions; tick marks on *x* axis denote every third case; dotted line denotes food break. Because unequal session lengths resulted in a low number of cases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

Sursa

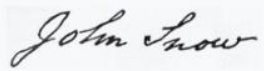
Asociere **1854**



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	

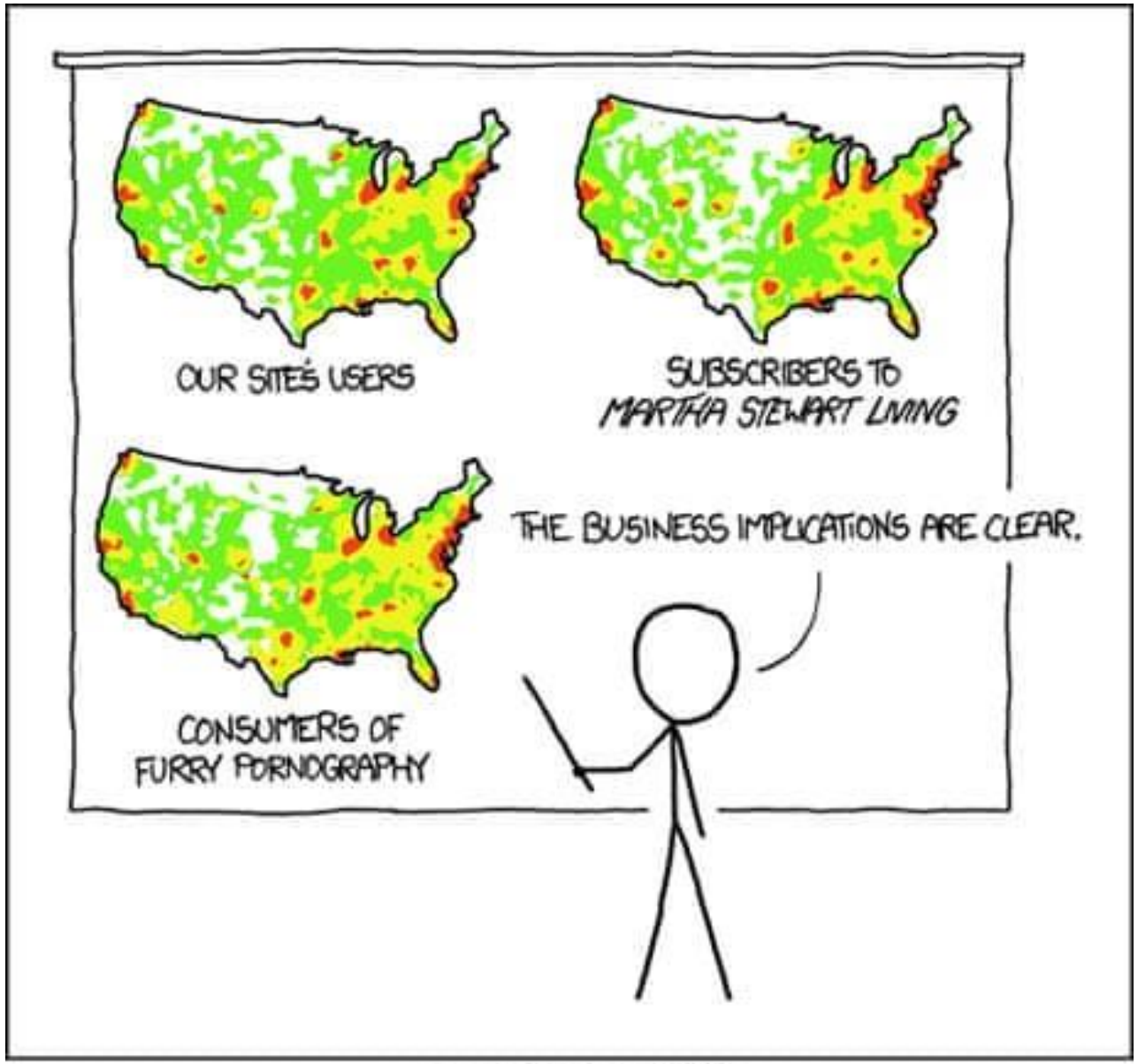


[Sursa]

„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.”

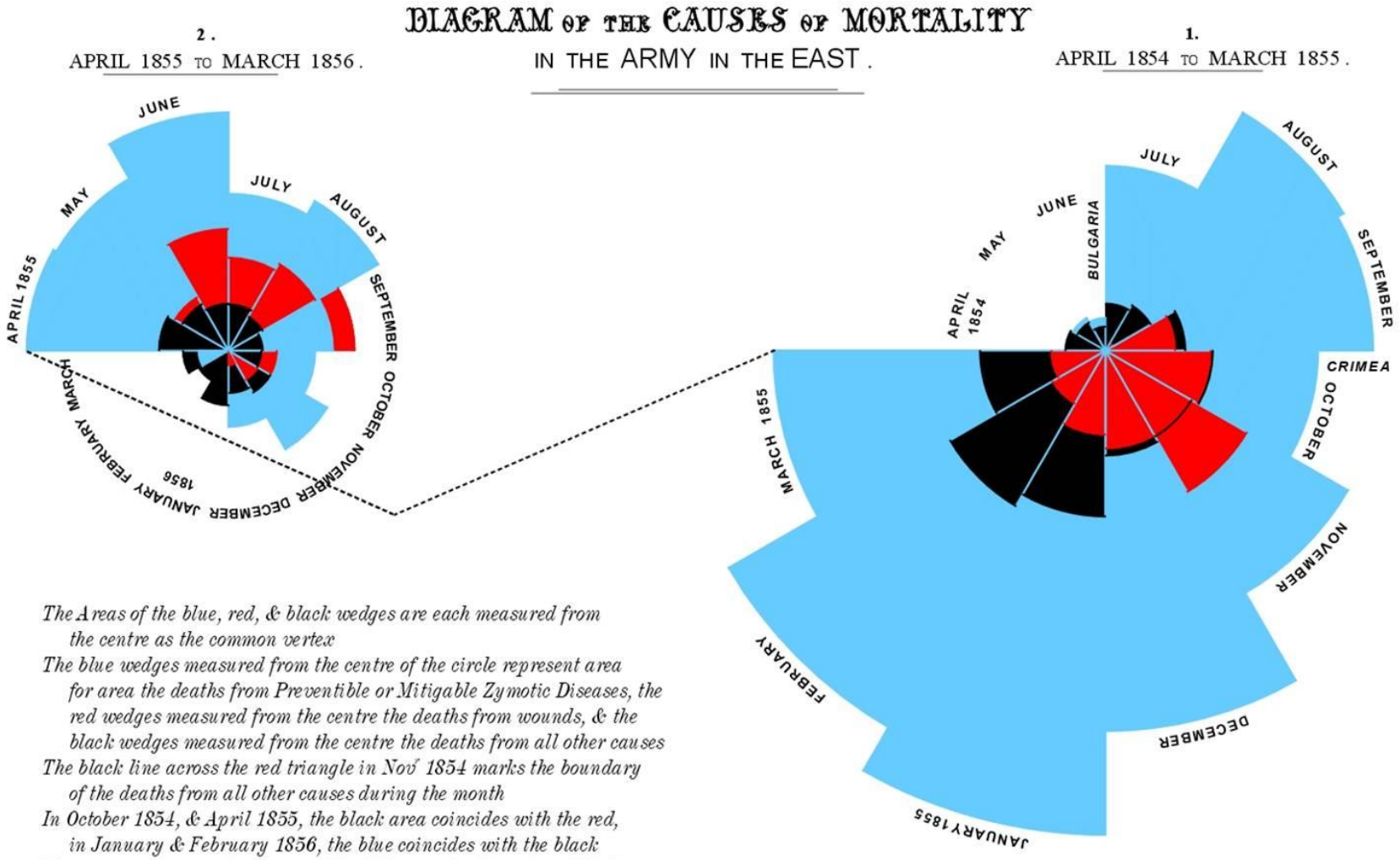
Sursă

Discuție in The Guardian



PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

Sursă



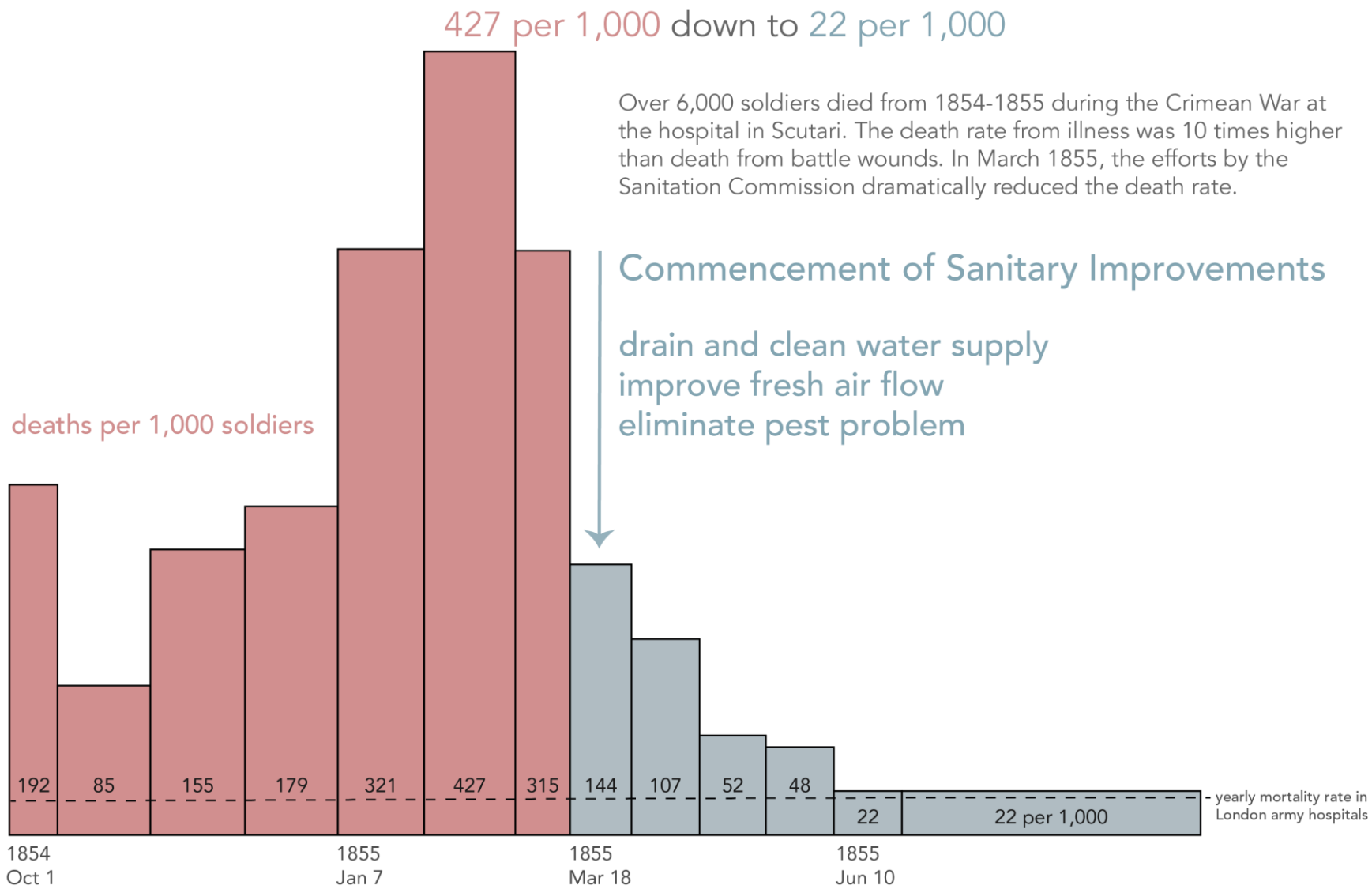
The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic Diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes
The black line across the red triangle in Nov 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



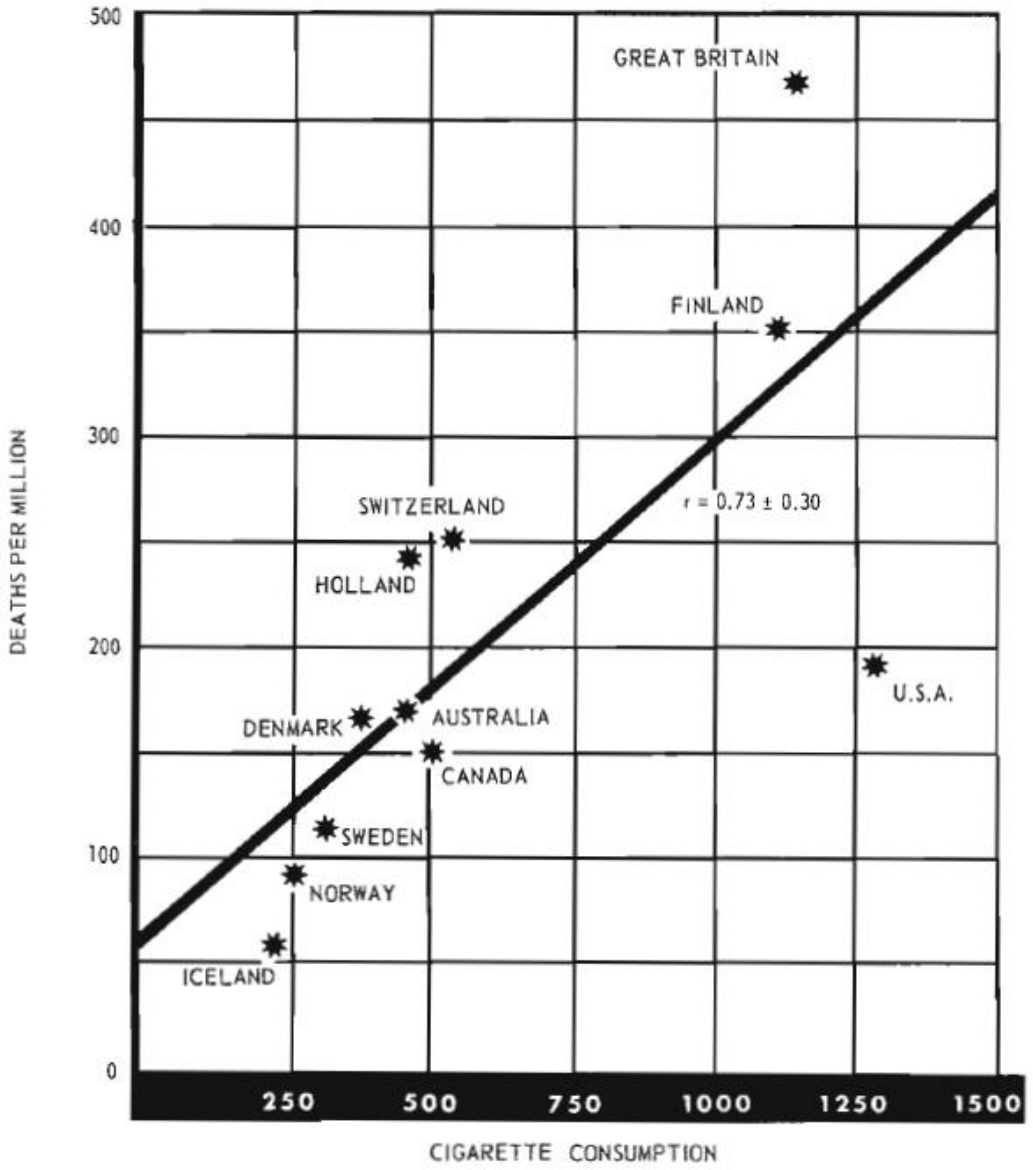
A redesign of Florence Nightingale's Rose Chart created by Jeffrey A. Shaffer | DataPlusScience.com

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

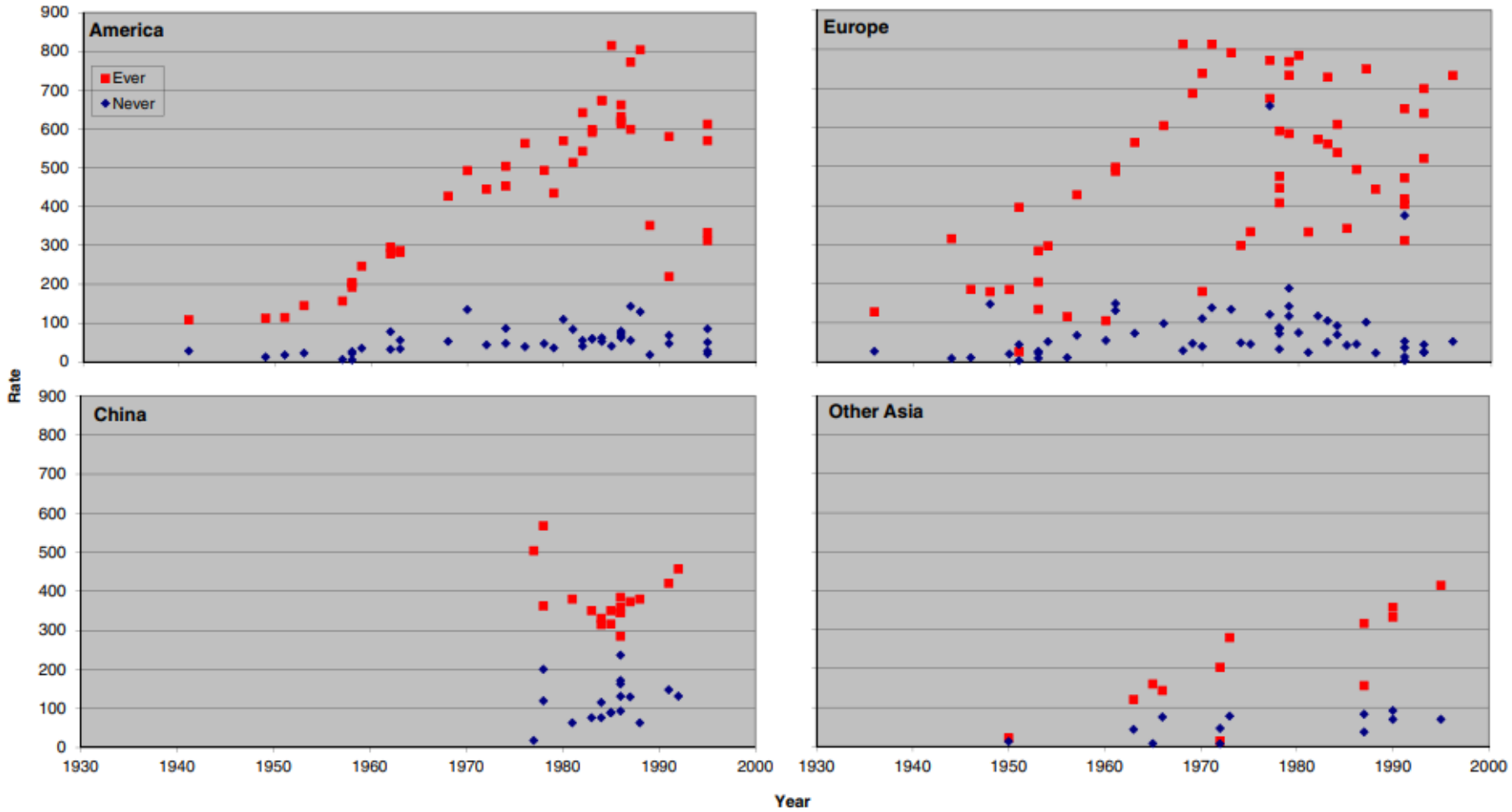
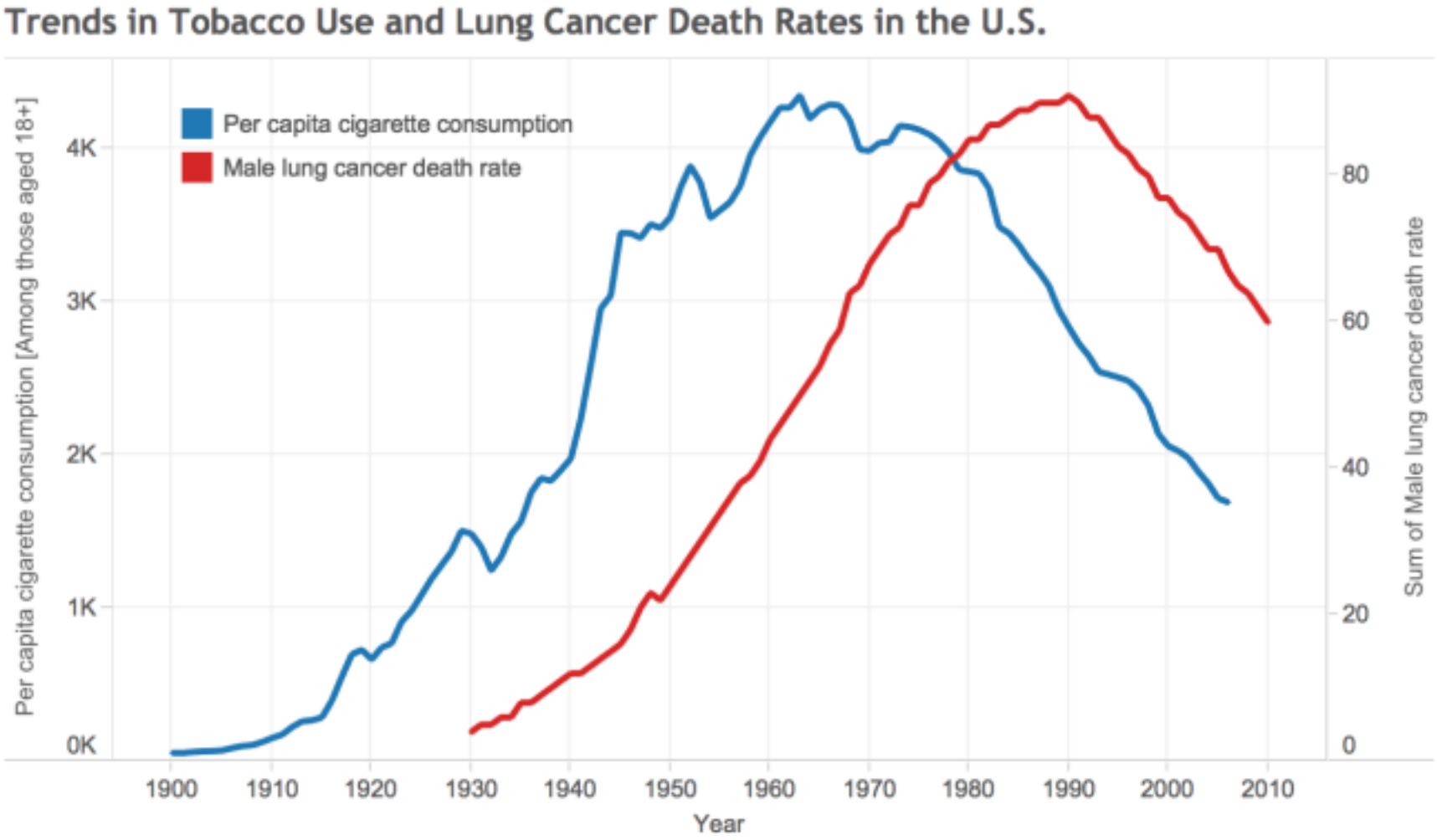


Figure 1 Scatter plot of lung cancer rates in males for never and ever smokers. Table 4 presents indirect estimates of mortality rates (per 100,000 per year) by smoking habit for all lung cancer in males. The individual study estimates for never smokers (blue diamonds) and ever smokers (red squares) are plotted against the midpoint year of the study, with separate plots shown for America (Canada, US and South/Central America), Europe (UK, Scandinavia, West Europe and East Europe), China (including Hong Kong), and Other Asia (Japan, South Korea, Singapore and Thailand).

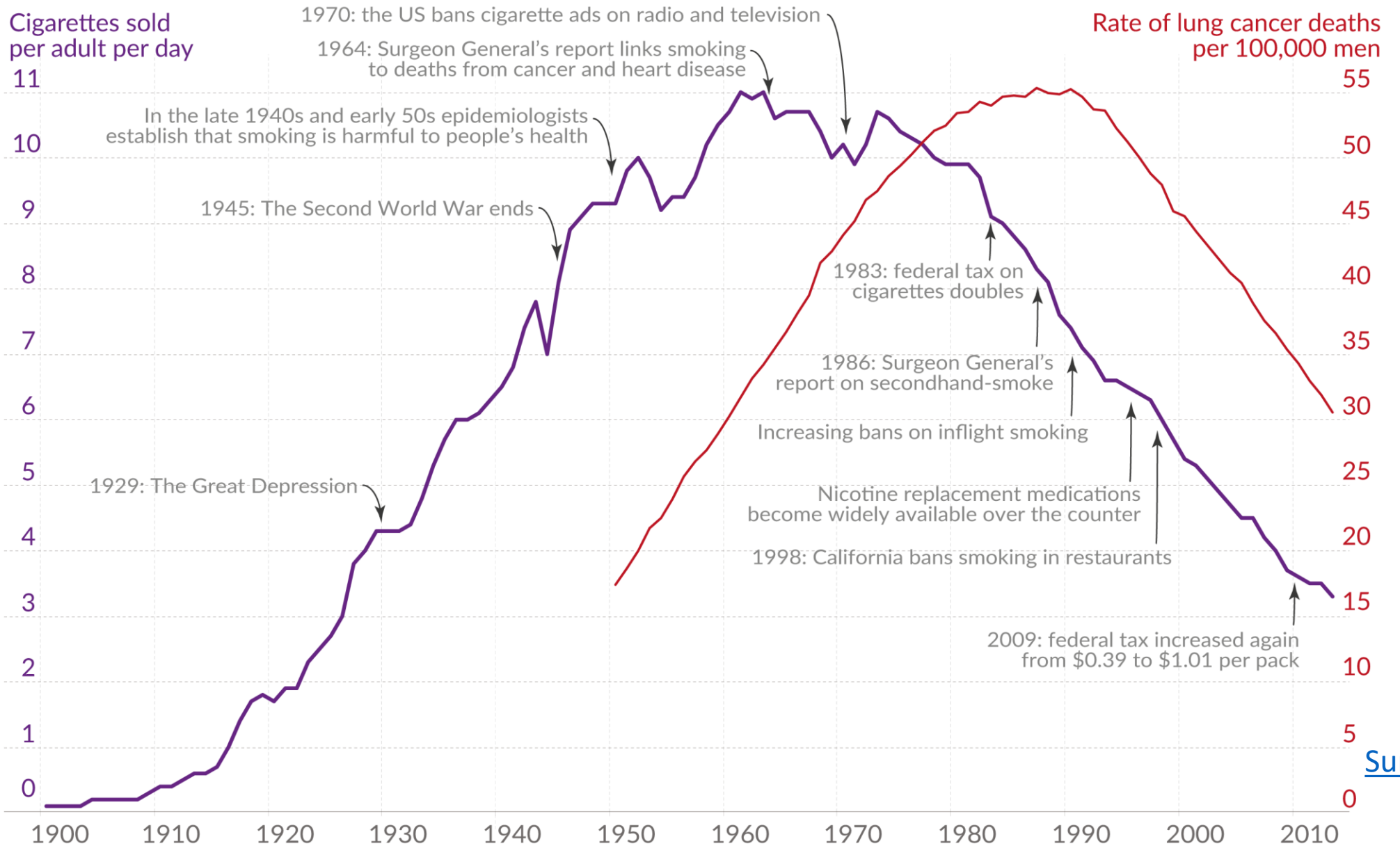
Sursa


Trends in Tobacco Use and Lung Cancer Death Rates in the U.S.
Per capita cigarette consumption
Male lung cancer death rate
Per capita cigarette consumption [Among those aged 18+]
Sum of Male lung cancer death rate
4K
3K
2K
1K
0K
80
60
40
20
0
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
Year
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



Sursa

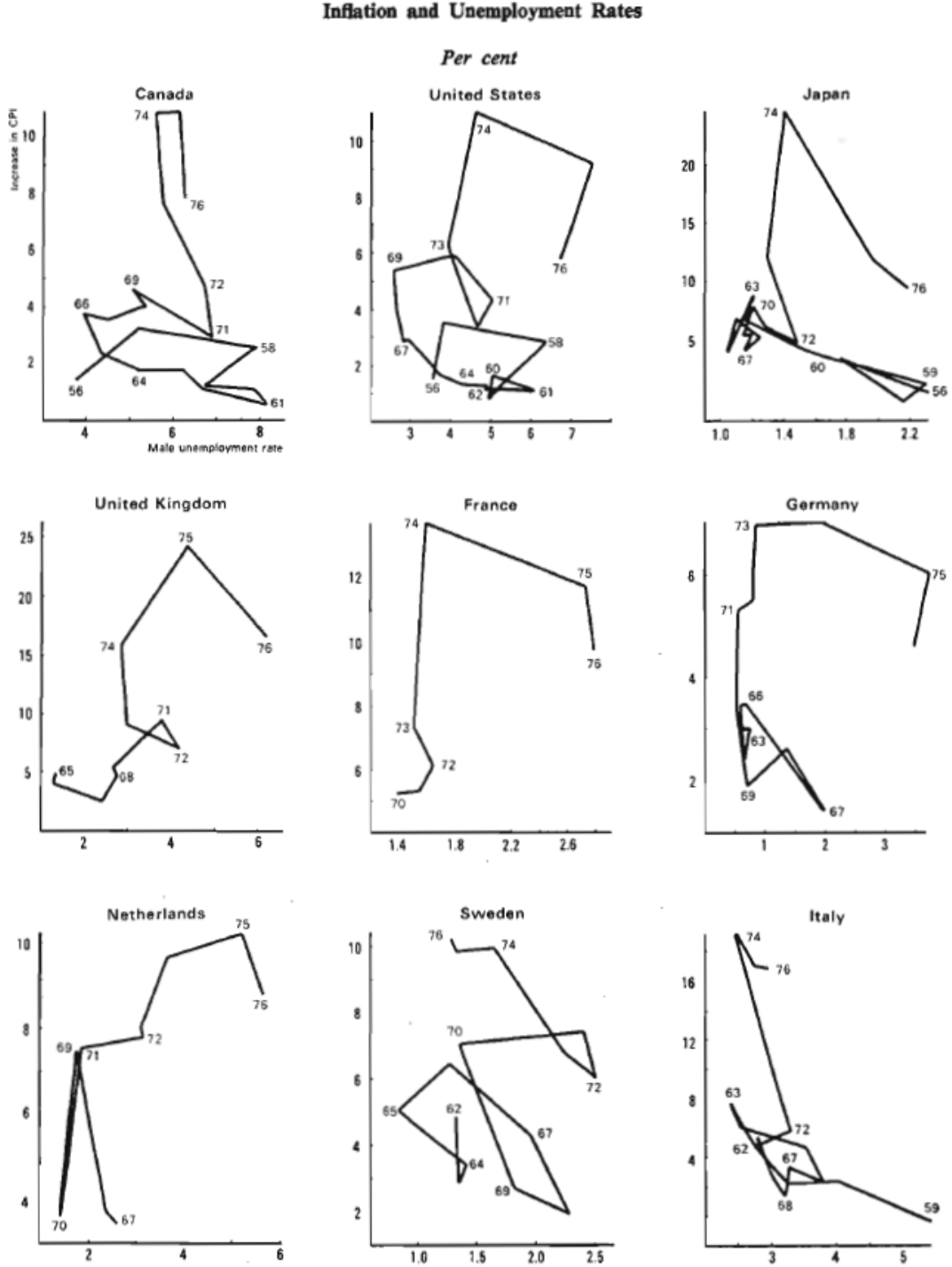
Data sources: International Smoking Statistics (2017); WHO Cancer Mortality Database (IARC). The death rate from lung-cancer is age-standardized.

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

Licensed under CC-BY by the author Max Roser.

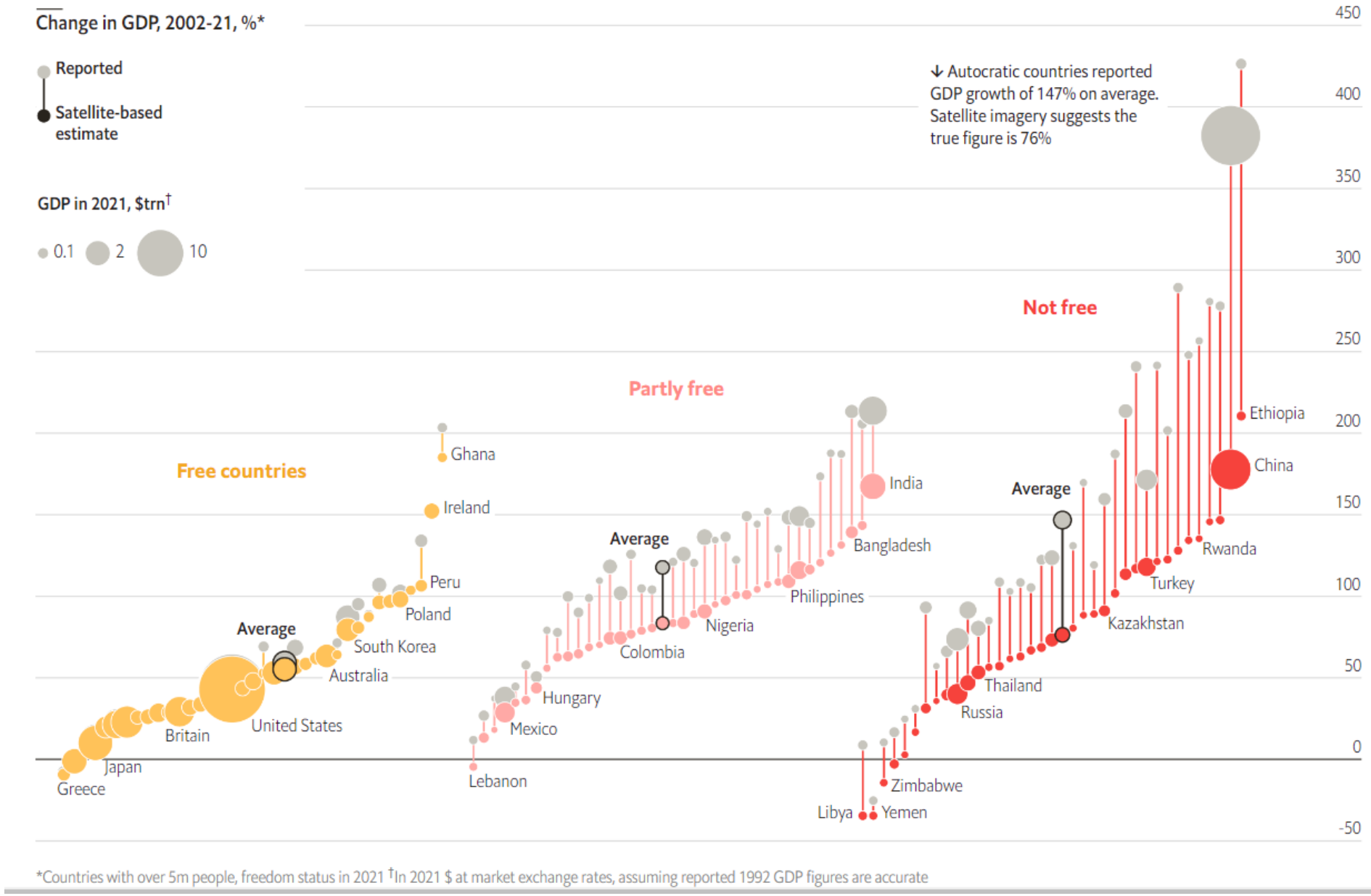
Lipsa corelației

1977



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

Sursă:
Tufte, 2001



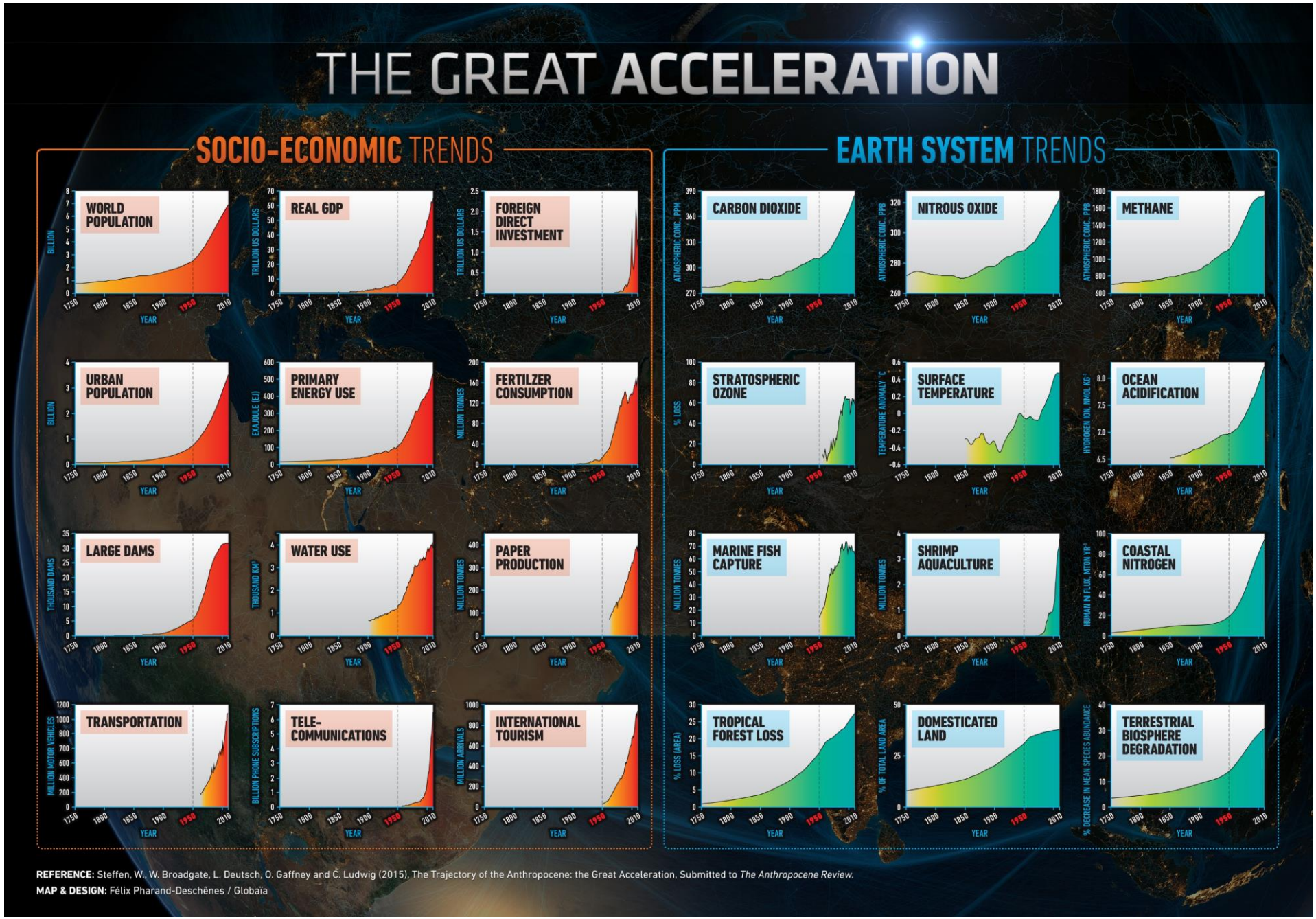
Sursa:
The Economist

Antropocen și climă

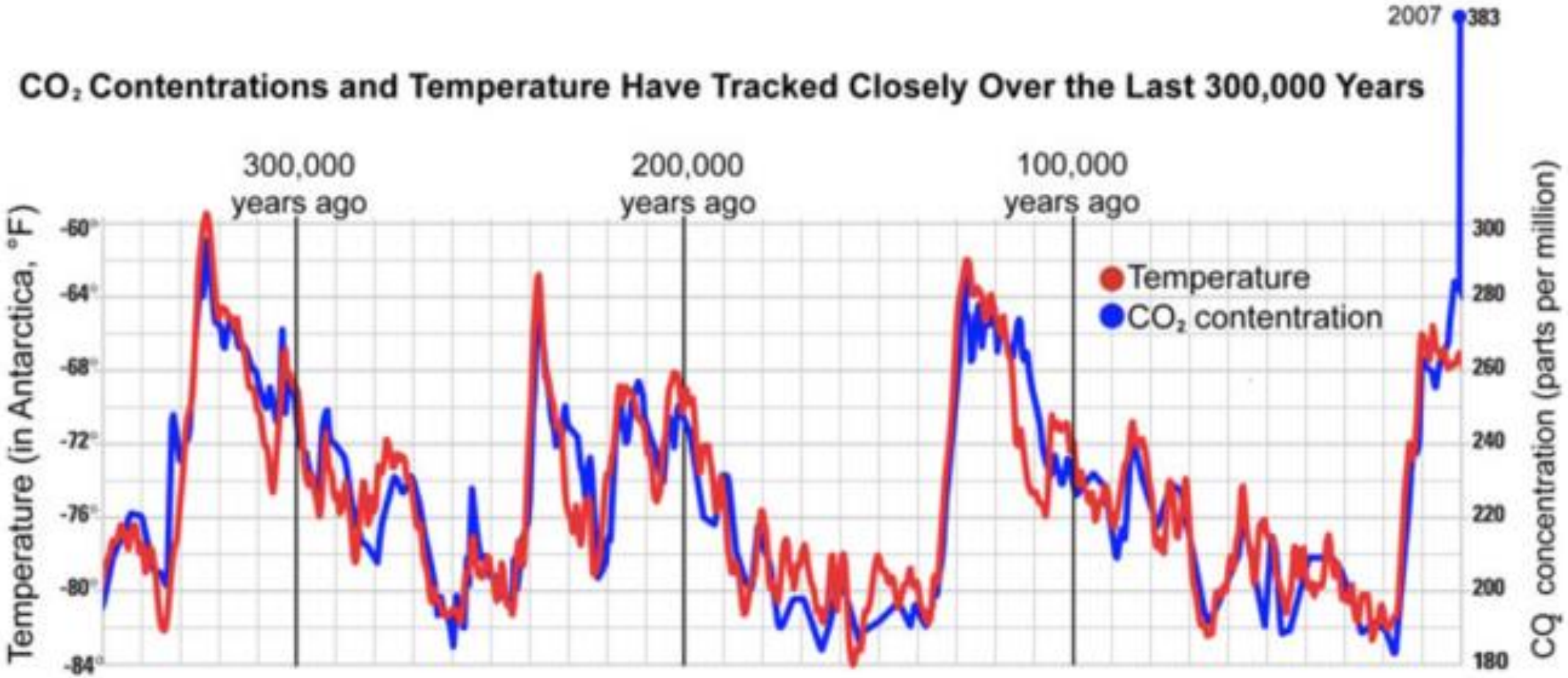
Schimbarea climatică în antropocen

- Tendințe: „Marea Accelerare”
- Efectul de seră: corelația concentrației CO2 cu temperatura medie
- Tendințe: Curba Keeling
  - Importanța scalei: de la liniar la exponențial și vertical
  - „Hockey stick”: Excepționalitatea fenomenului
- Vizualizări: benzile încălzirii (Warming Stripes)
  - Reprezentare numerică / cromatică
  - Reprezentare liniară / circulară

2015

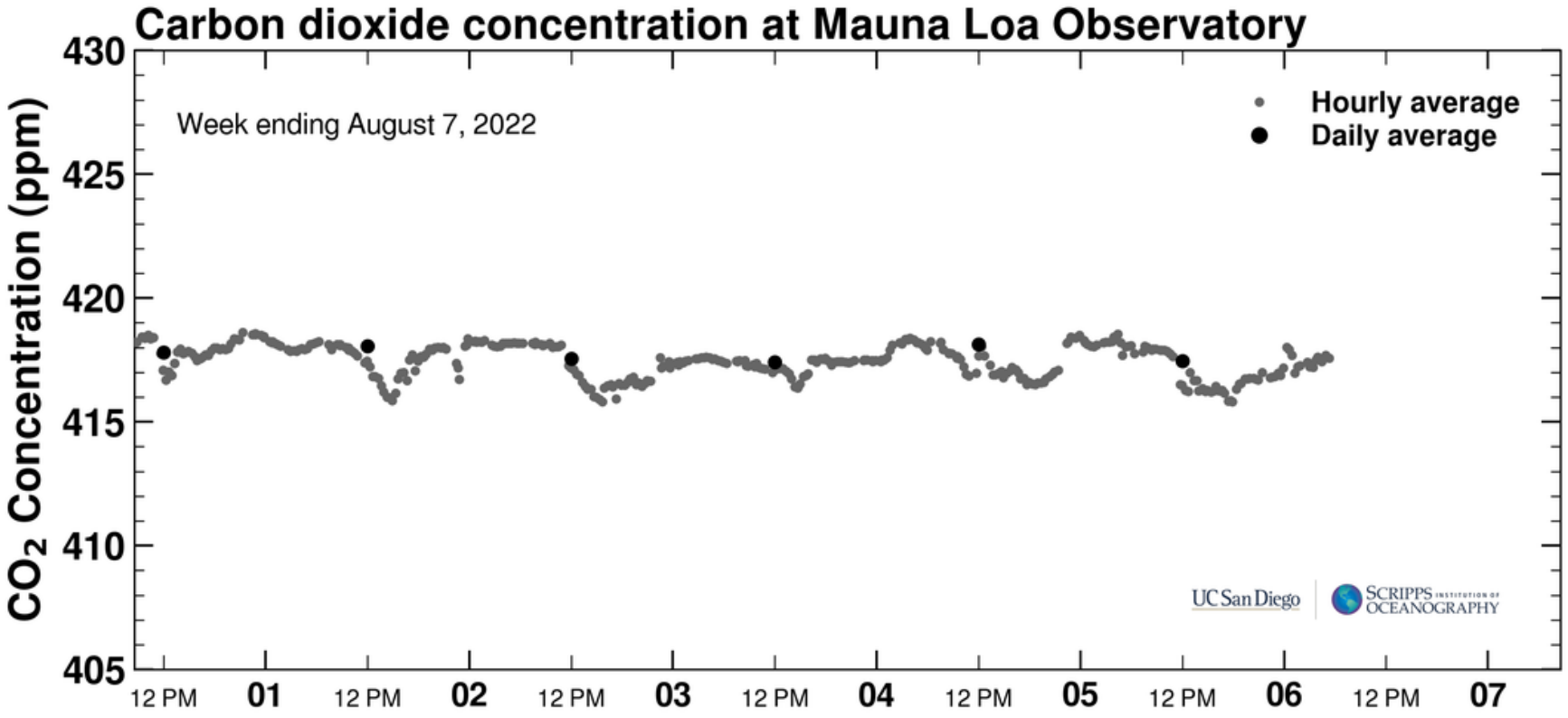

THE GREAT ACCELERATION
SOCIO-ECONOMIC TRENDS
WORLD POPULATION
BILLION
REAL GDP
TRILLION US DOLLARS
FOREIGN DIRECT INVESTMENT
TRILLION US DOLLARS
URBAN POPULATION
BILLION
PRIMARY ENERGY USE
EXAJOULE (EJ)
FERTILZER CONSUMPTION
MILLION TONNES
LARGE DAMS
THOUSAND DAMS
WATER USE
THOUSAND KM³
PAPER PRODUCTION
MILLION TONNES
TRANSPORTATION
MILLION MOTOR VEHICLES
TELE-COMMUNICATIONS
BILLION PHONE SUBSCRIPTIONS
INTERNATIONAL TOURISM
MILLION ARRIVALS
YEAR
EARTH SYSTEM TRENDS
CARBON DIOXIDE
ATMOSPHERIC CONC., PPM
NITROUS OXIDE
ATMOSPHERIC CONC., PPB
METHANE
ATMOSPHERIC CONC., PPB
STRATOSPHERIC OZONE
% LOSS
SURFACE TEMPERATURE
TEMPERATURE ANOMALY °C
OCEAN ACIDIFICATION
HYDROGEN ION, NMOL KG⁻¹
MARINE FISH CAPTURE
MILLION TONNES
SHRIMP AQUACULTURE
MILLION TONNES
COASTAL NITROGEN
HUMAN N FLUX, MTON YR⁻¹
TROPICAL FOREST LOSS
% LOSS (AREA)
DOMESTICATED LAND
% OF TOTAL LAND AREA
TERRESTRIAL BIOSPHERE DEGRADATION
% DECREASE IN MEAN SPECIES ABUNDANCE
YEAR
REFERENCE: Steffen, W., W. Broadgate, L. Deutsch, O. Gaffney and C. Ludwig (2015), The Trajectory of the Anthropocene: the Great Acceleration, Submitted to The Anthropocene Review.
MAP & DESIGN: Félix Pharand-Deschênes / Globaïa

Sursa


CO_2 Contentrations and Temperature Have Tracked Closely Over the Last 300,000 Years
300,000 years ago
200,000 years ago
100,000 years ago
2007
383
Temperature
CO_2 contentration
Temperature (in Antarctica, °F)
-60°
-64°
-68°
-72°
-76°
-80°
-84°
CO2 concentration (parts per million)
300
280
260
240
220
200
180

Sursa

2022

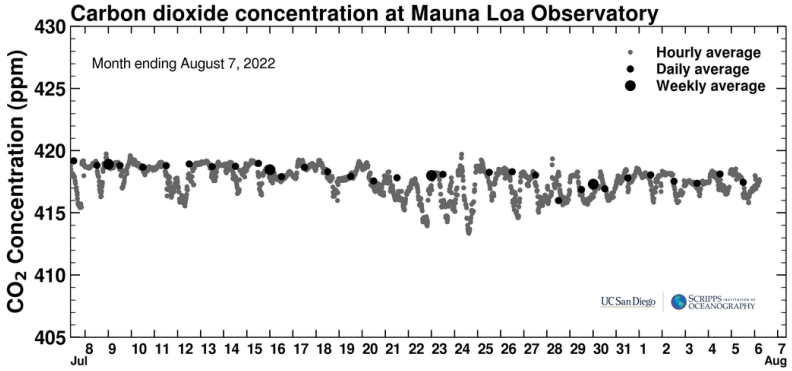


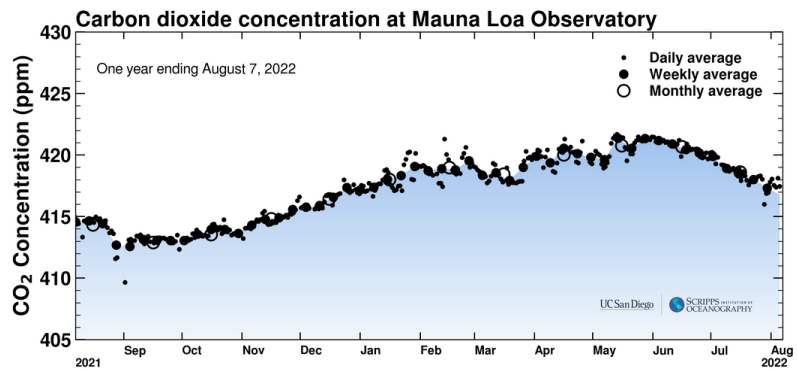
Charles David Keeling

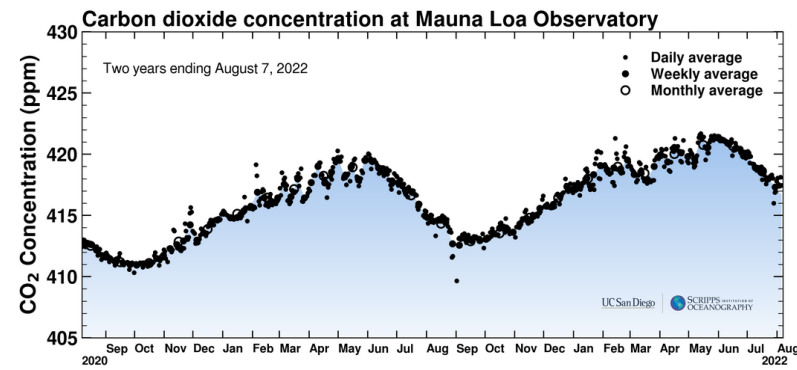
Keeling receives the Medal of Science in 2001

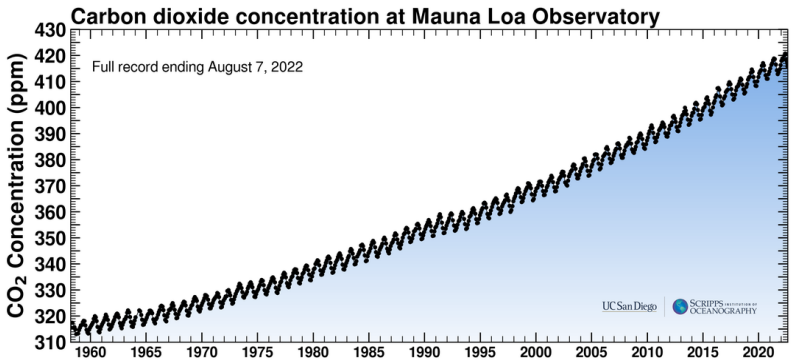
Born	April 20, 1928
Died	June 20, 2005 (aged 77)
Citizenship	American
Alma mater	University of Illinois Northwestern University
Known for	Keeling Curve
Awards	Second Half Century Award, American Meteorological Society (1981) Maurice Ewing Medal (1991) Blue Planet Prize (1993) National Medal of Science (2002) Tyler Prize for Environmental Achievement (2005)

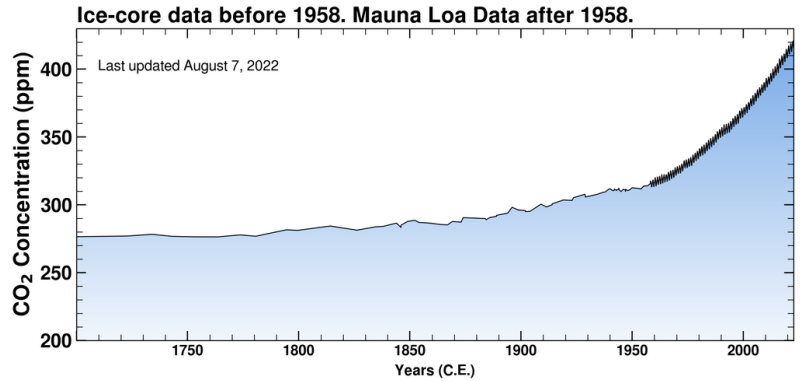
Sursă

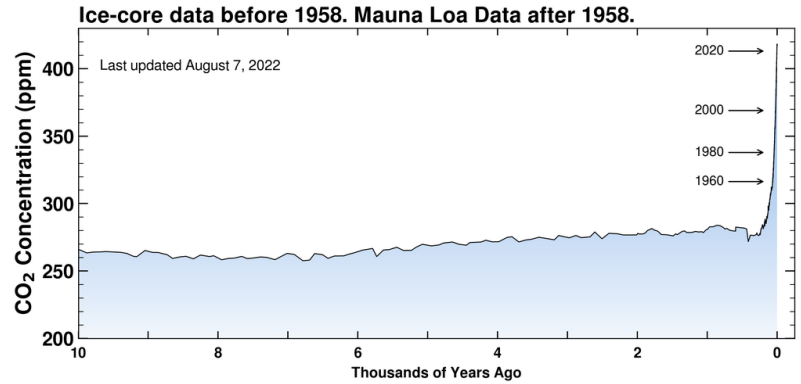

Carbon dioxide concentration at Mauna Loa Observatory
Month ending August 7, 2022
Hourly average
Daily average
Weekly average
CO2 Concentration (ppm)
UC San Diego
SCRIPPS INSTITUTION OF OCEANOGRAPHY

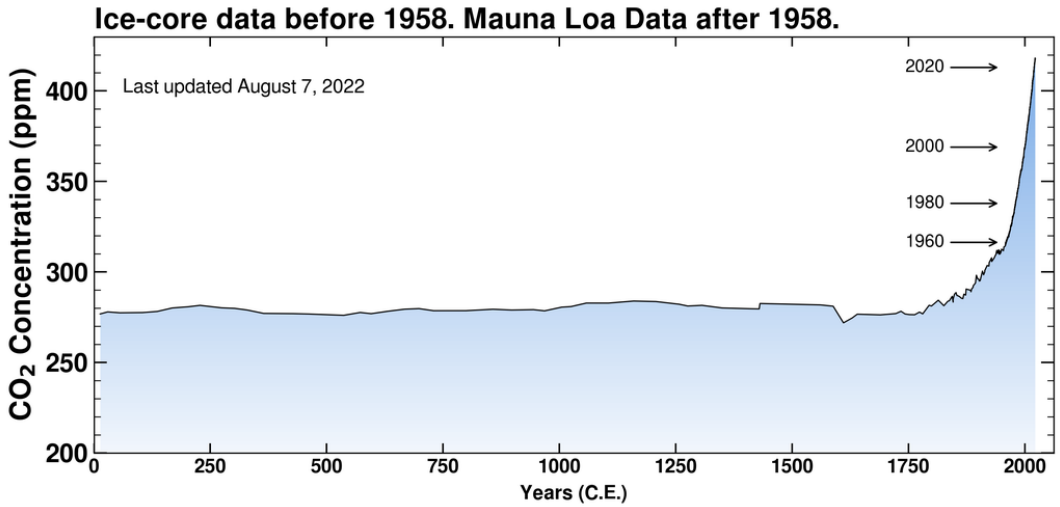

Carbon dioxide concentration at Mauna Loa Observatory
One year ending August 7, 2022
Daily average
Weekly average
Monthly average
CO2 Concentration (ppm)
UC San Diego
SCRIPPS INSTITUTION OF OCEANOGRAPHY

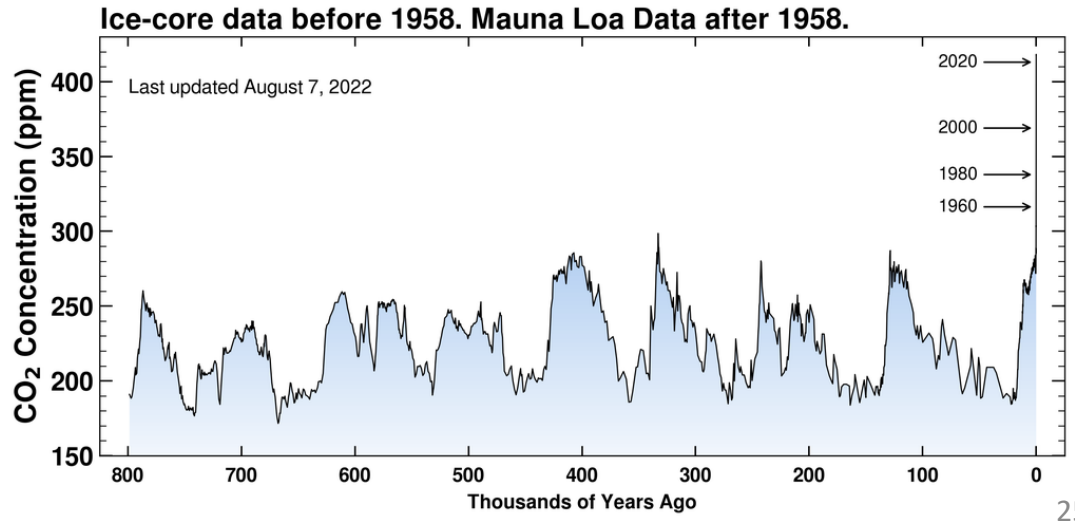

Carbon dioxide concentration at Mauna Loa Observatory
Two years ending August 7, 2022
Daily average
Weekly average
Monthly average
CO2 Concentration (ppm)
UC San Diego
SCRIPPS INSTITUTION OF OCEANOGRAPHY


Carbon dioxide concentration at Mauna Loa Observatory
Full record ending August 7, 2022
CO2 Concentration (ppm)
UC San Diego
SCRIPPS INSTITUTION OF OCEANOGRAPHY

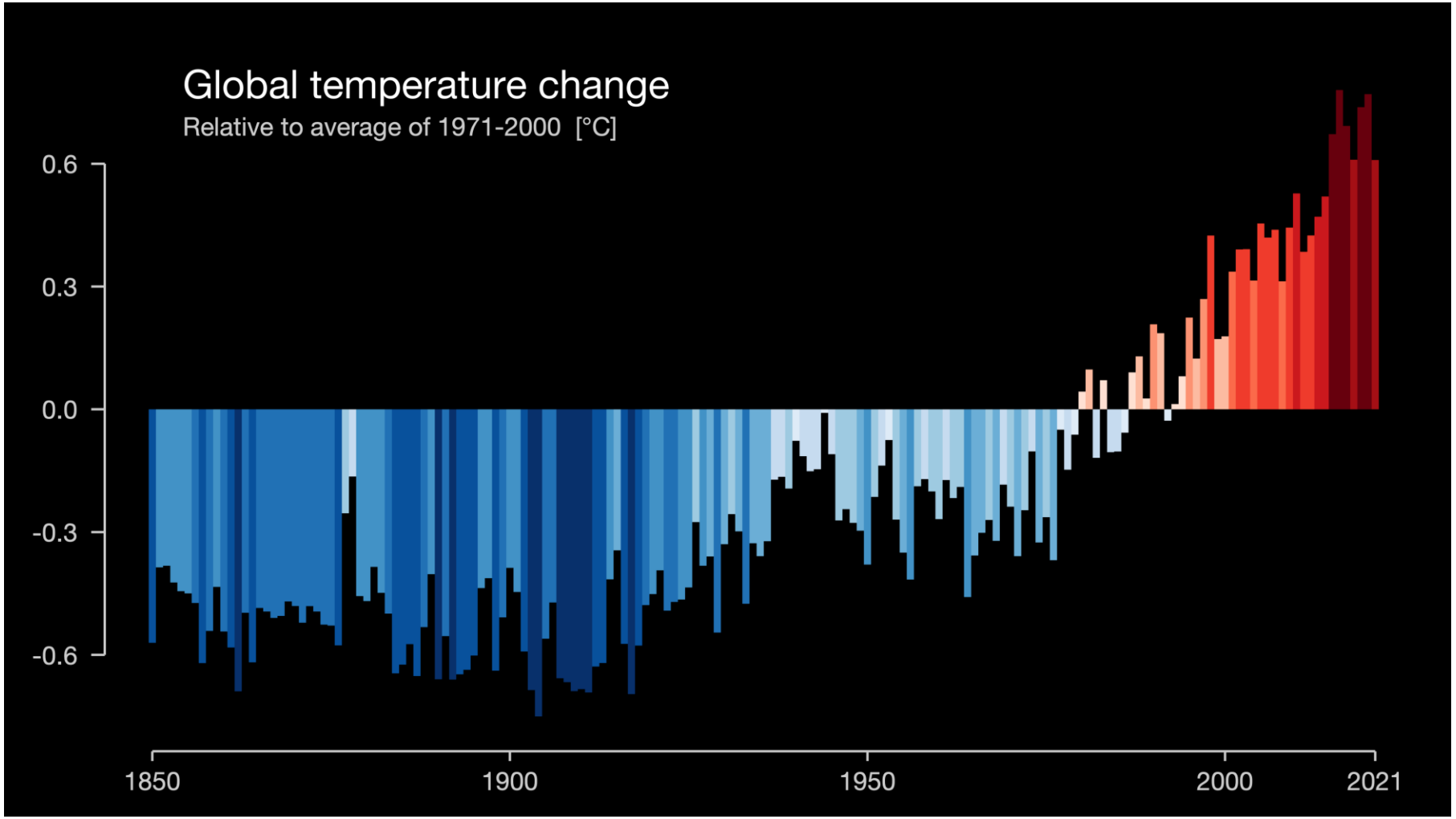

Ice-core data before 1958. Mauna Loa Data after 1958.
Last updated August 7, 2022
CO2 Concentration (ppm)
Years (C.E.)


Ice-core data before 1958. Mauna Loa Data after 1958.
Last updated August 7, 2022
2020
2000
1980
1960
CO2 Concentration (ppm)
Thousands of Years Ago


Ice-core data before 1958. Mauna Loa Data after 1958.
Last updated August 7, 2022
2020
2000
1980
1960
CO2 Concentration (ppm)
Years (C.E.)


Ice-core data before 1958. Mauna Loa Data after 1958.
Last updated August 7, 2022
2020
2000
1980
1960
CO2 Concentration (ppm)
Thousands of Years Ago

Sursă

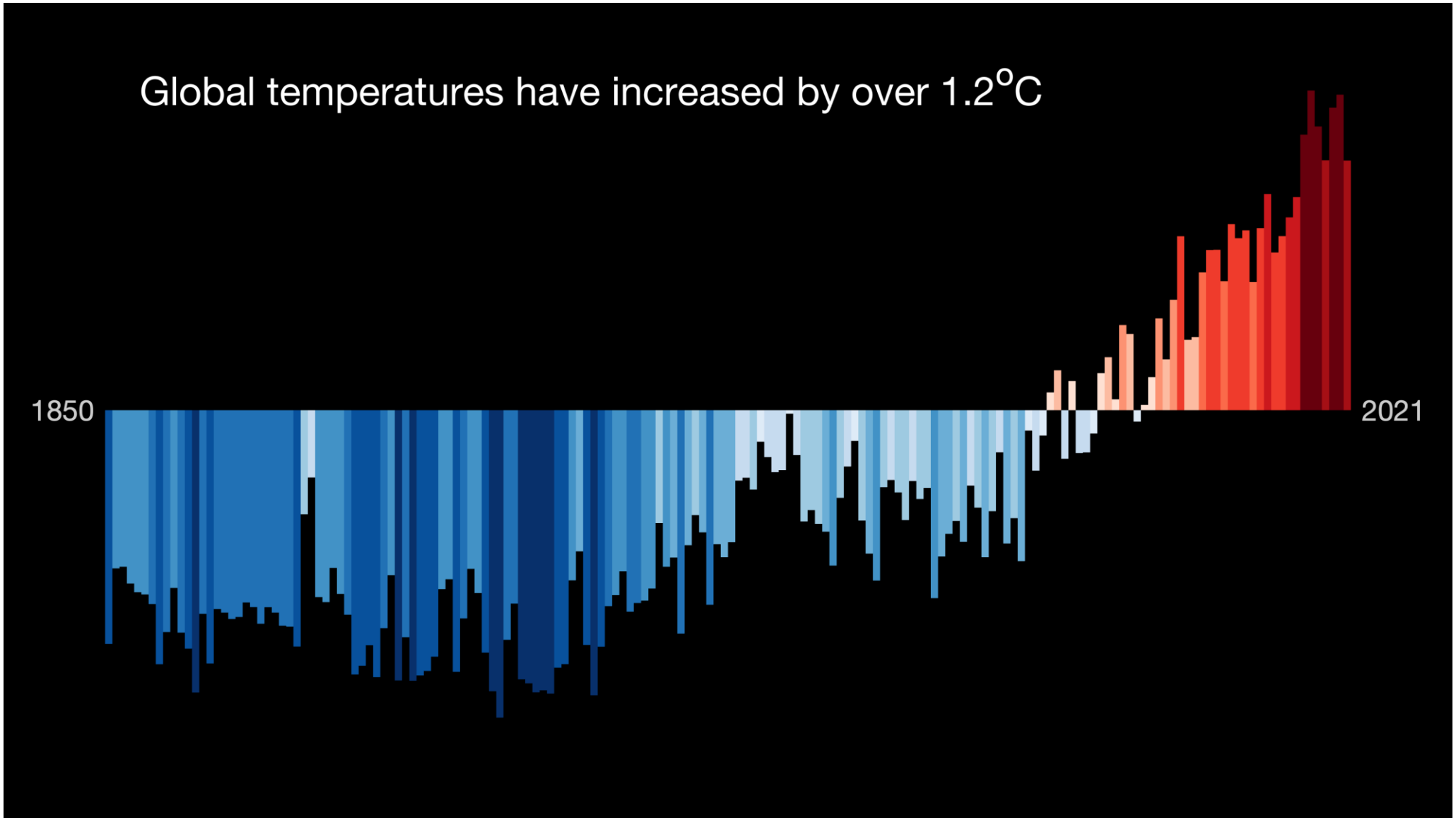


[Sursă]

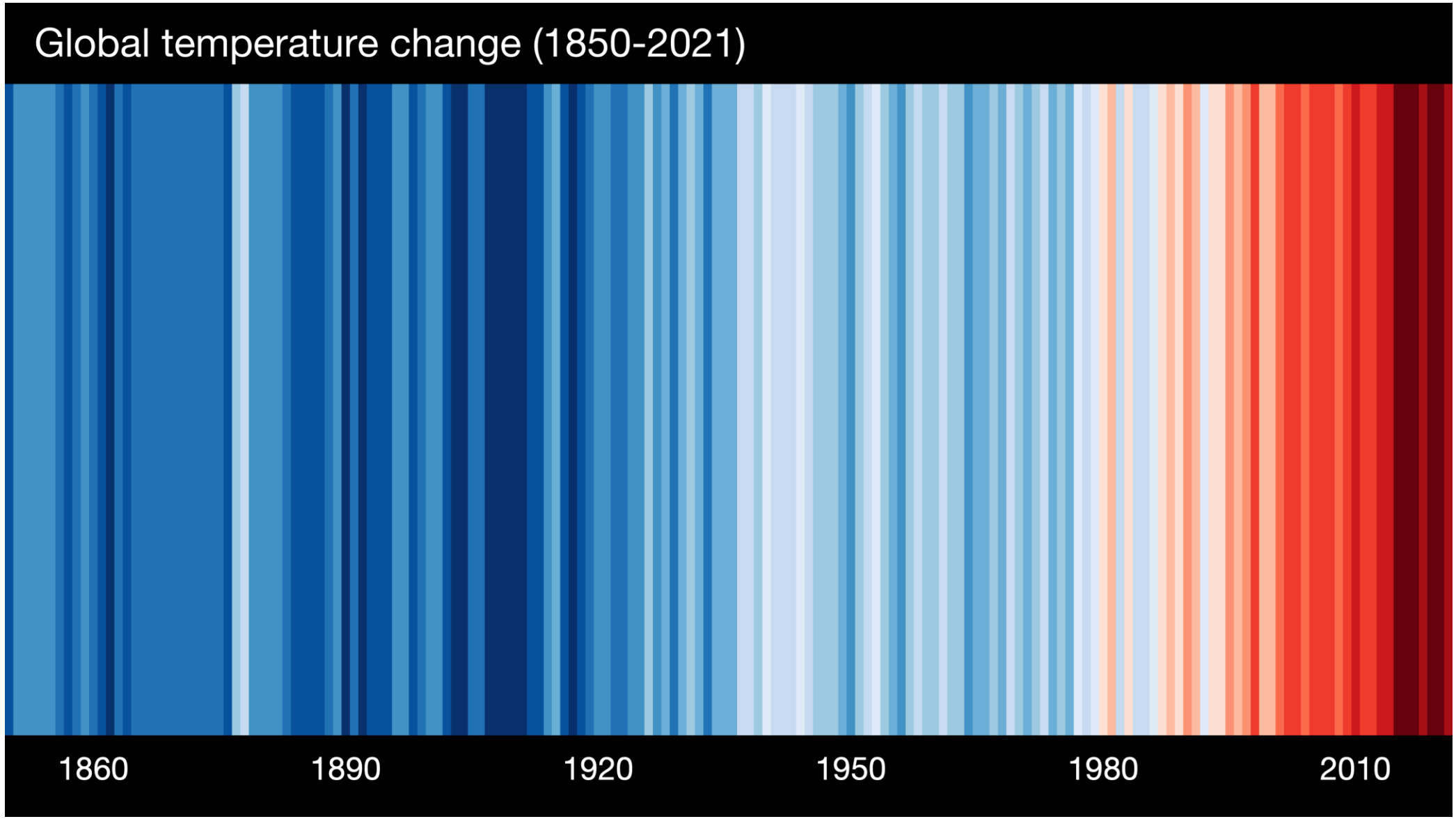
Ed Hawkins
MBE

Hawkins speaking at COP 21 (Paris, 2015)

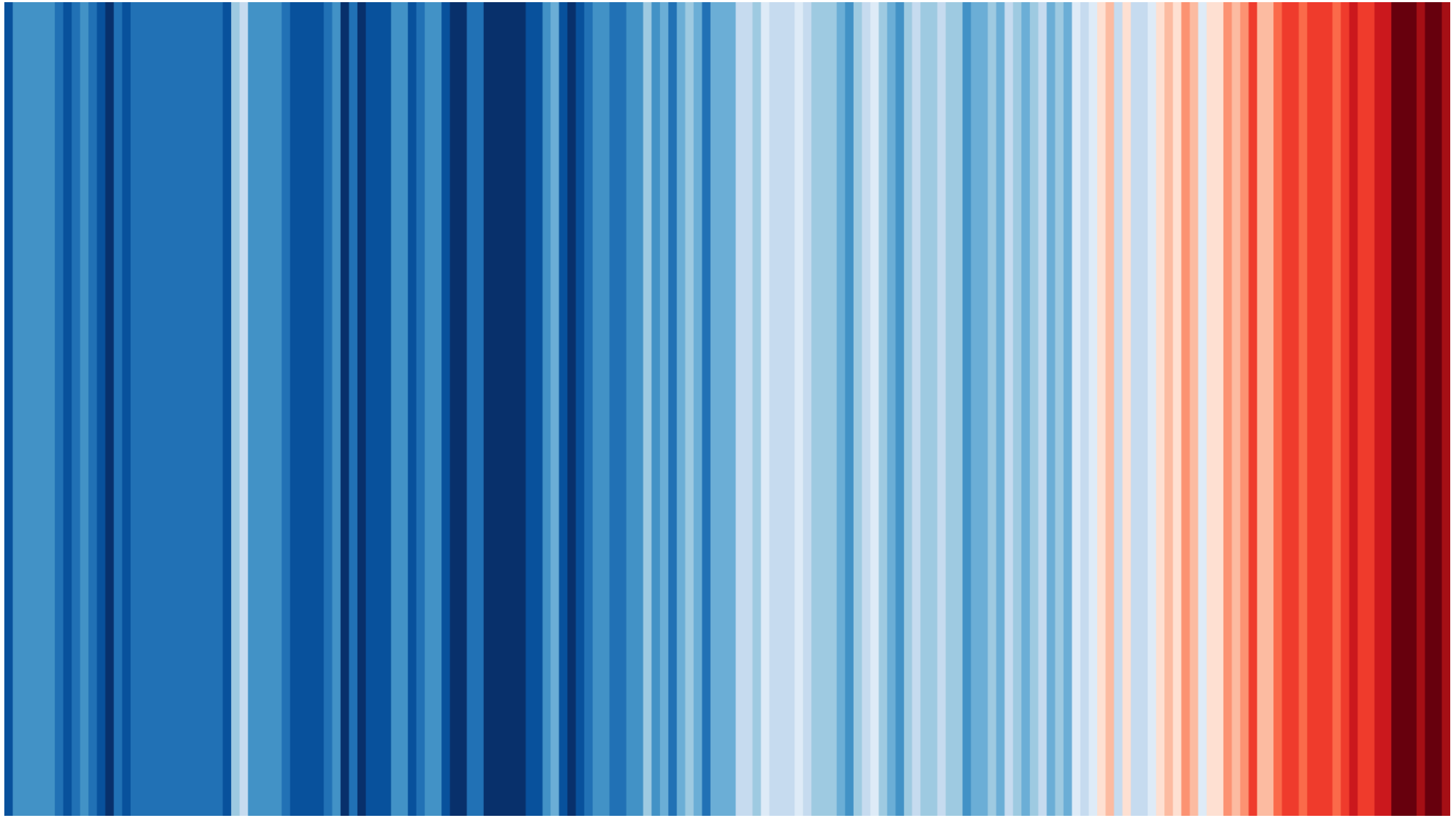
Nationality	British
Known for	Global warming data visualization
Scientific career	
Fields	Climatology, data visualization
Institutions	University of Reading
Website	www.climate-lab-book.ac.uk



Sursă

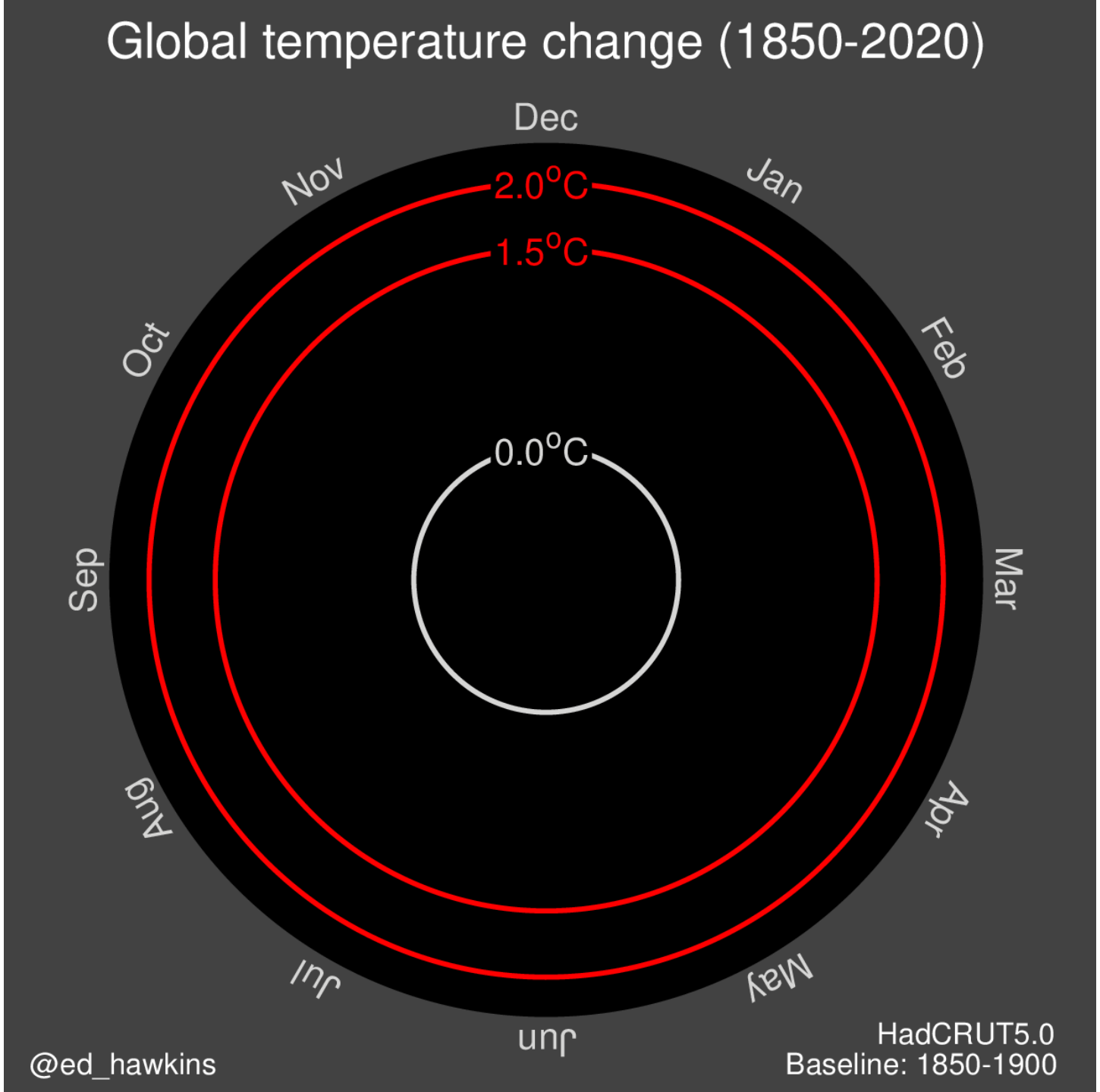

Global temperature change (1850-2021)
1860
1890
1920
1950
1980
2010

Sursă



Sursă

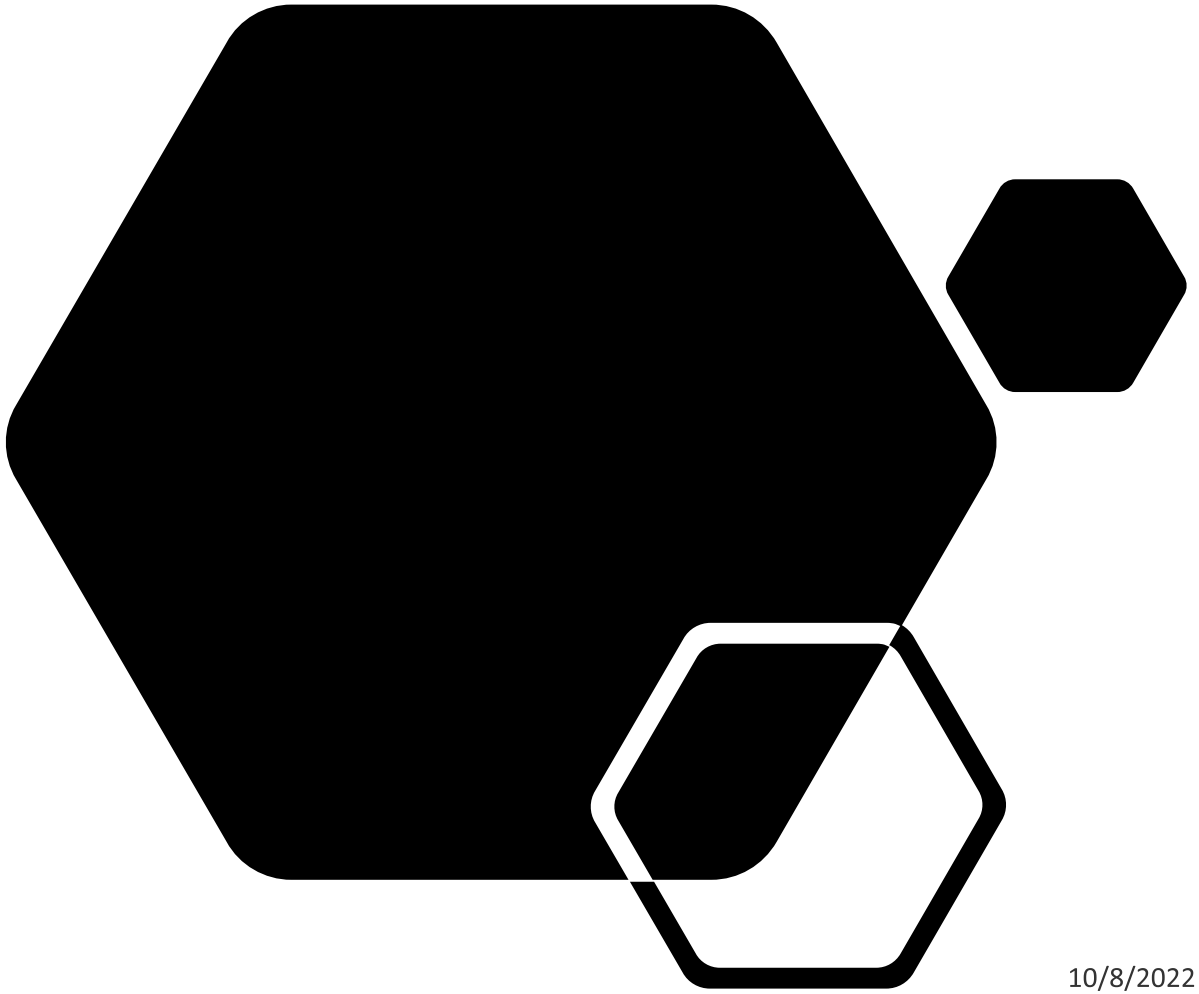
2021


Global temperature change (1850-2020)
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
2.0°C
1.5°C
0.0°C
@ed_hawkins
HadCRUT5.0
Baseline: 1850-1900

[Sursă]

Atribuiri cauzale **robuste**

Criteriile lui Hill



Criteriile cauzalității @ Austin Bradford Hill (1965)

- *Strength*
- *Consistency*
- *Specificity*
- *Temporality*
- *Biological gradient*
- *Plausibility*
- *Coherence*
- *Experiment*
- *Analogy*

- Intensitatea asocierii
- Consistență între surse diferite
- Specificitate
- Temporalitate (înainte / după)
- Doză / răspuns
- Plauzibilitate
- Coerență experiment - epidemiologie
- Dovezi experimentale
- 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)

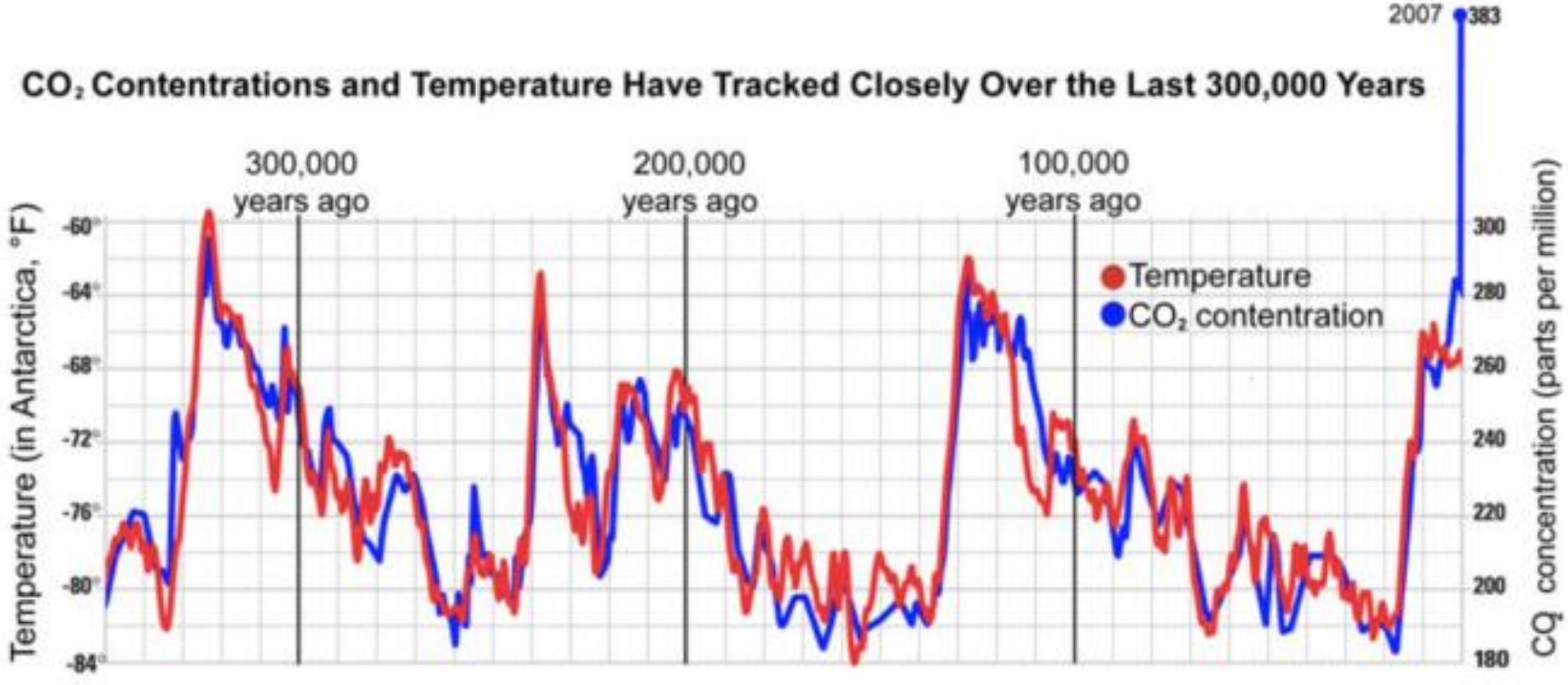
1. **Strength** (effect size): A small association does not mean that there is not a causal effect, though the larger the association, the more likely that it is causal.
2. **Consistency** (reproducibility): Consistent findings observed by different persons in different places with different samples 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)

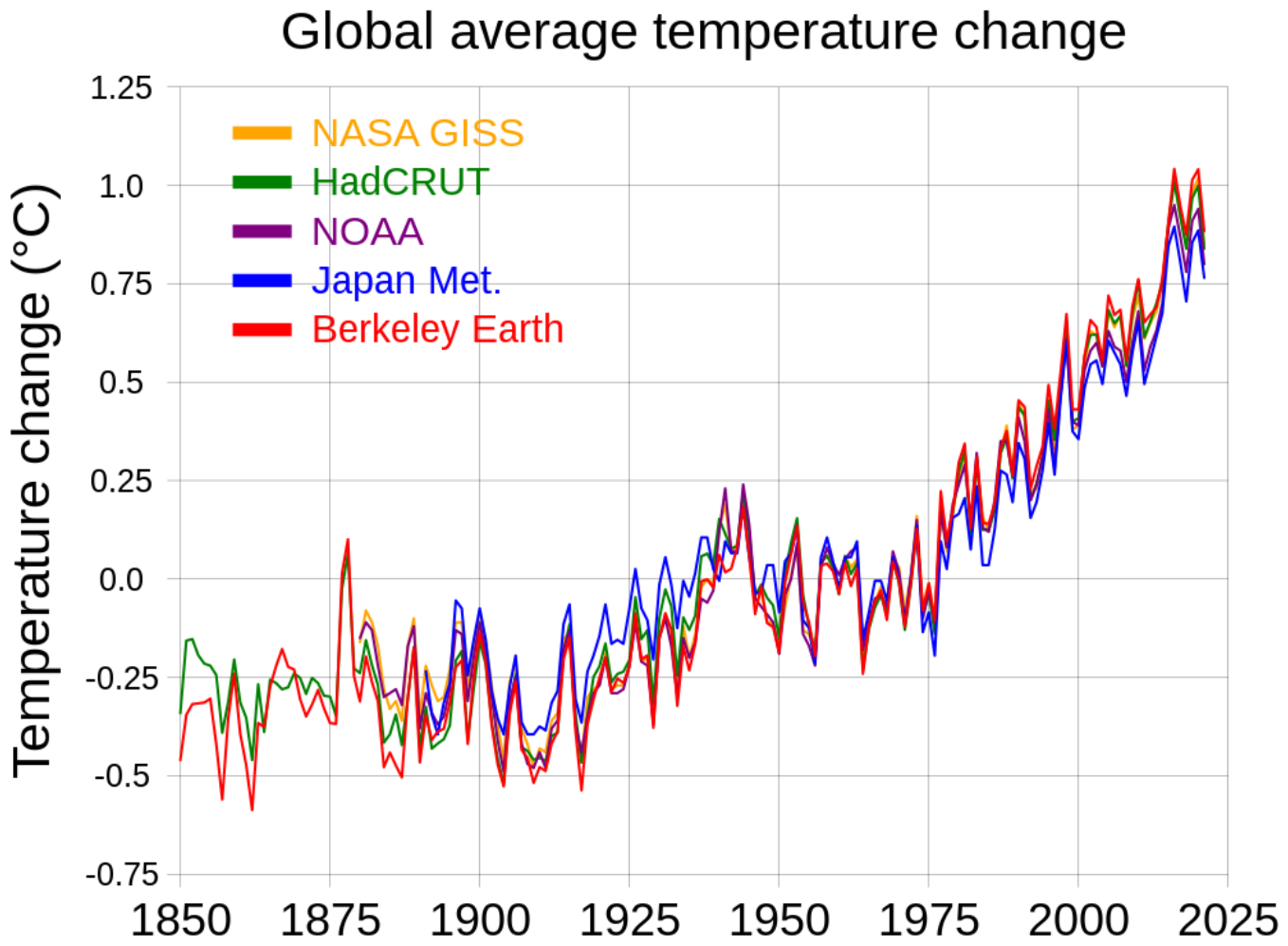
Criteriul 1: Intensitatea corelației



Sursa

Miller

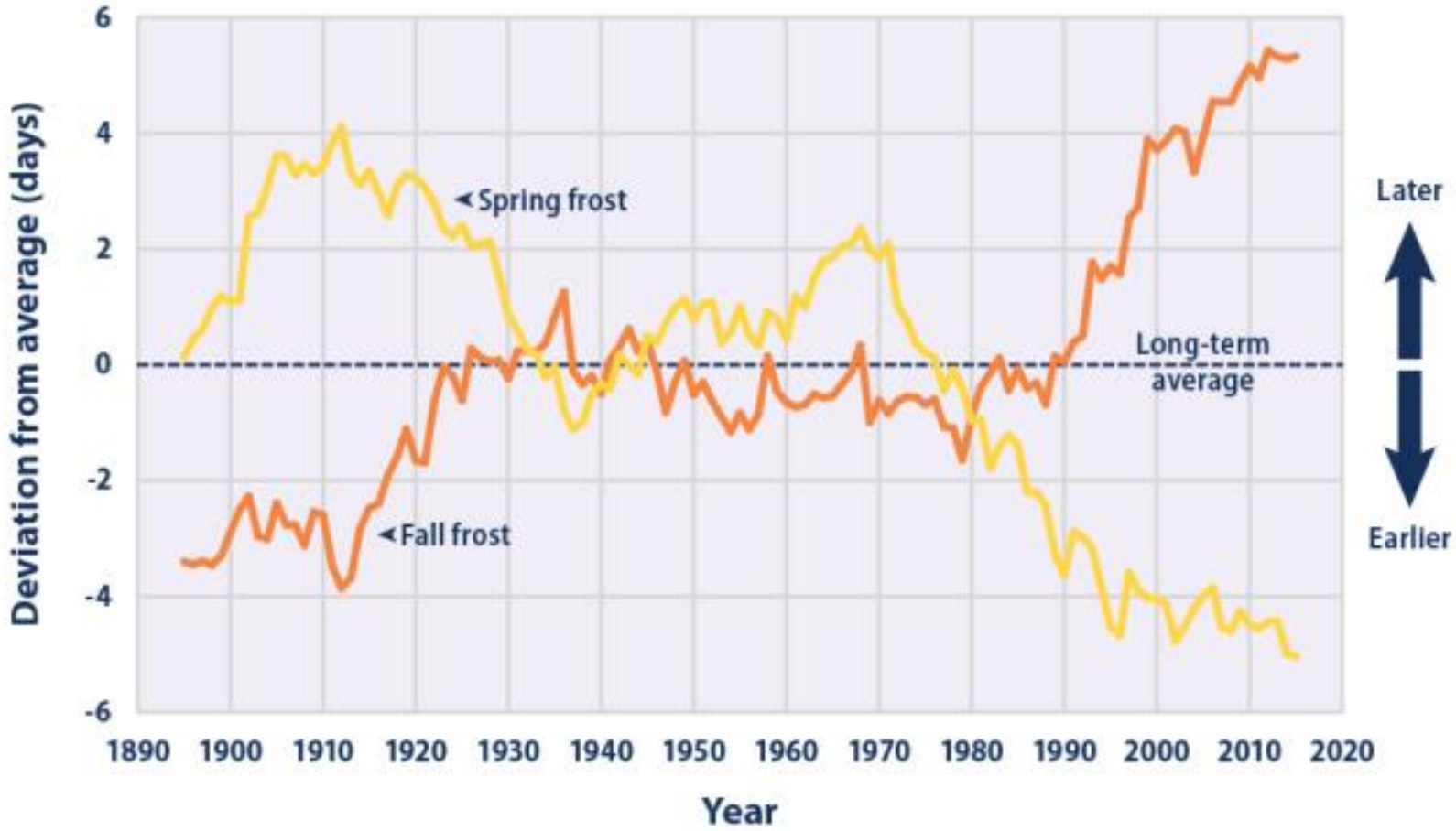
Criteriul 2: Consistența între surse diferite



Sursa

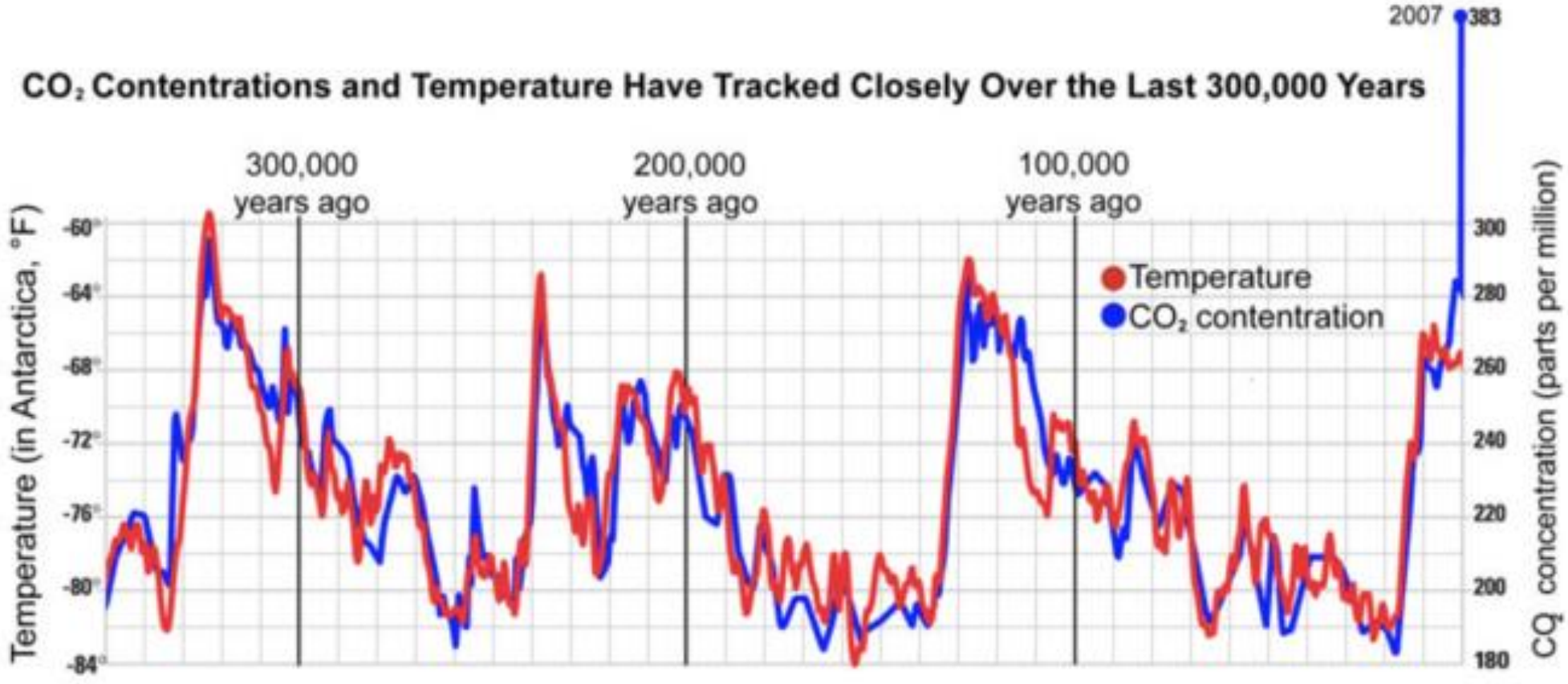
Miller

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



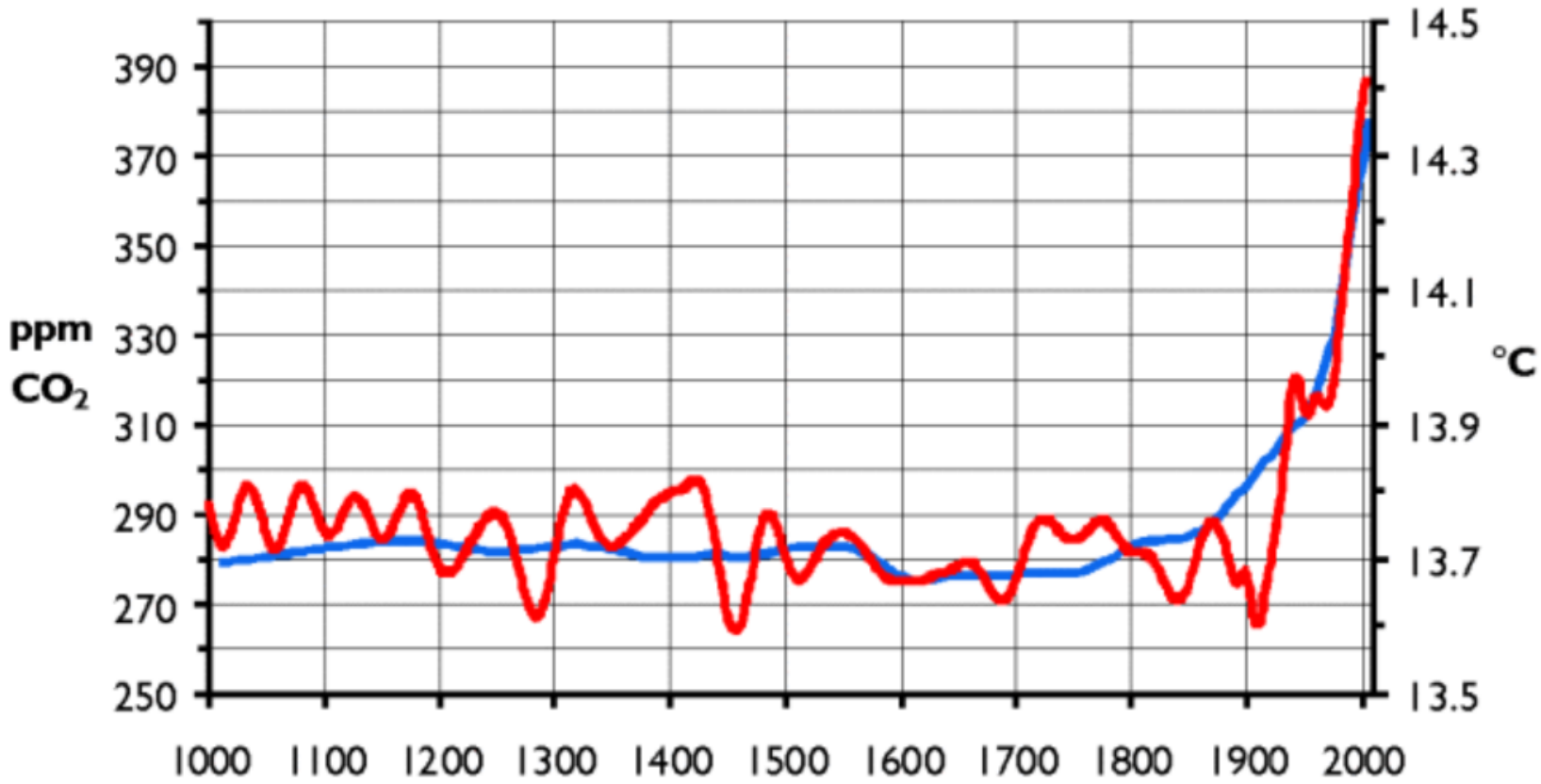
Miller

Criteriul 3: Specificitatea pentru o perioadă

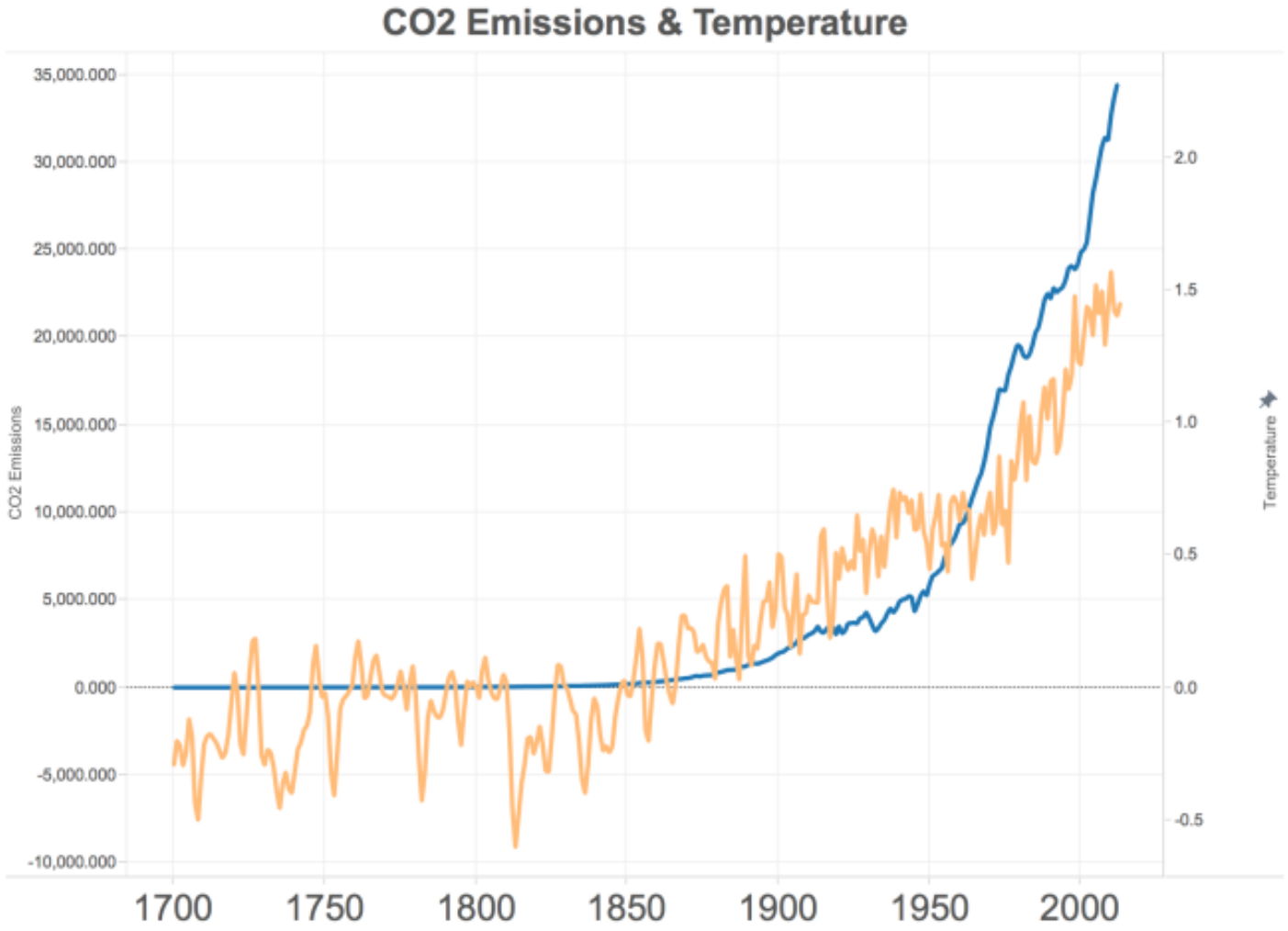


Miller

Criteriul 3 – specificitate pentru sec. 20-21



Criteriul 4: Temporalitate (înainte / după)



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

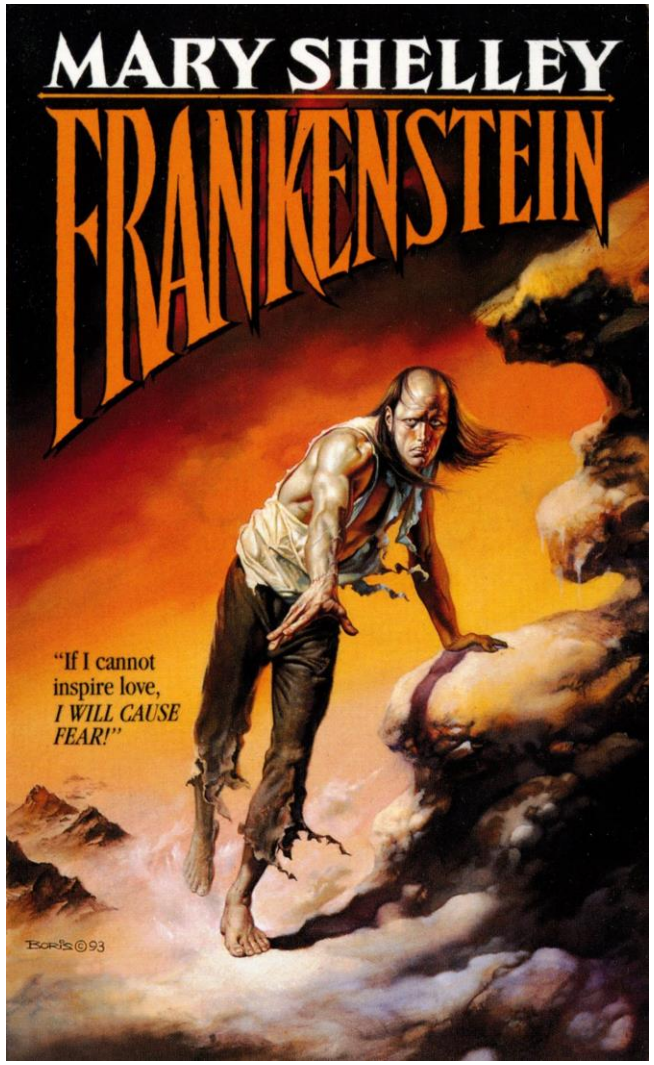
@jonhenley

Wed 26 Jan 2011 20.00 GMT

Genghis Khan – eco-warrior. Illustration: Alamy

Sursa

Criteriul 8: Erupția din 1815 (Tambora, Indonezia)



Draisine, 1817

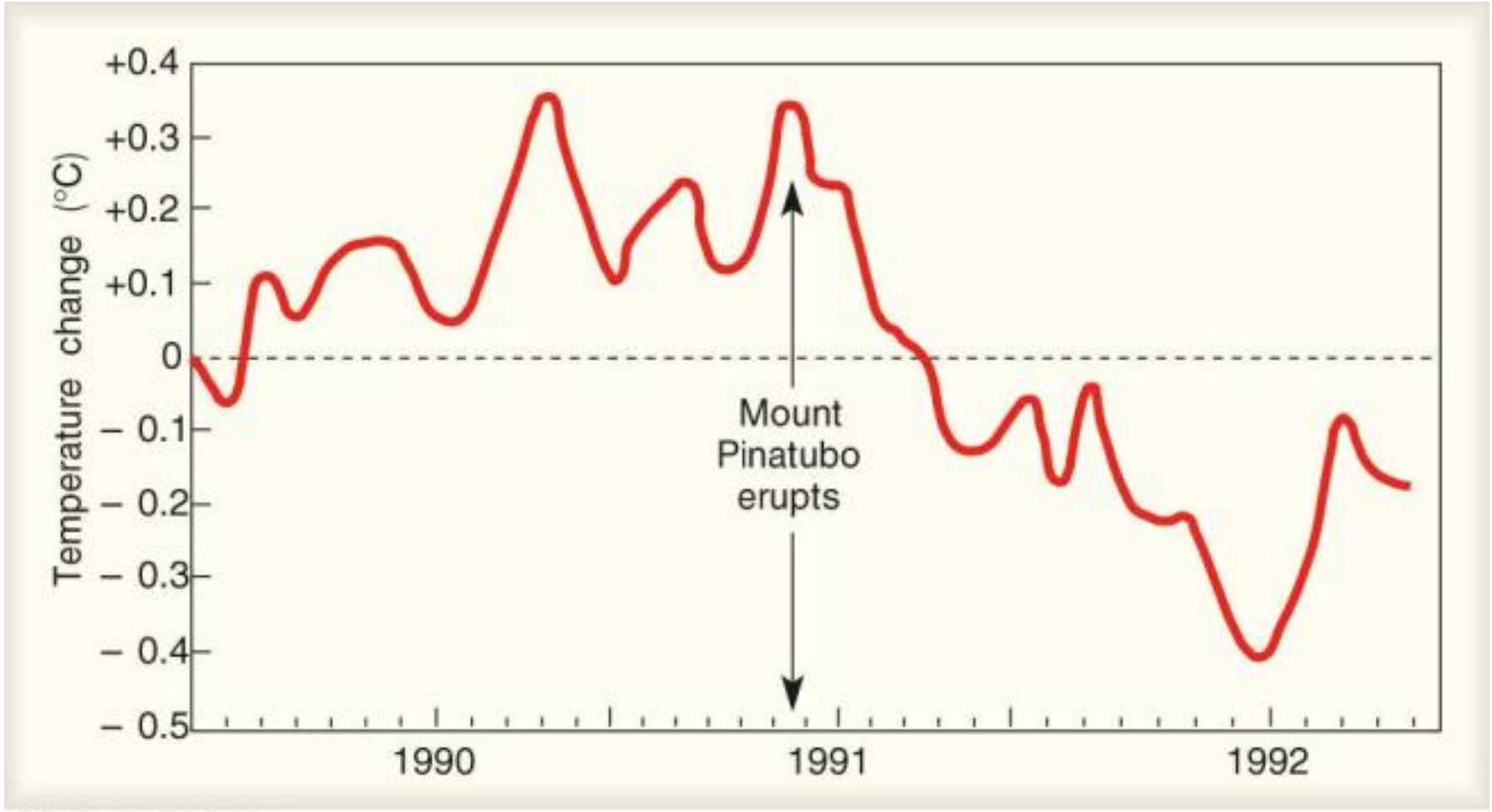
Criteriul 8: 1815 - anul fără vară

- „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)



Miller

Criteriul 9: Analogia

UCSB ScienceLine

Encouraging Curiosity for a

Home
How It Works
Ask a Question
Search Topics
Webcasts
Our Scientists
Science Links
Contact Us

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!

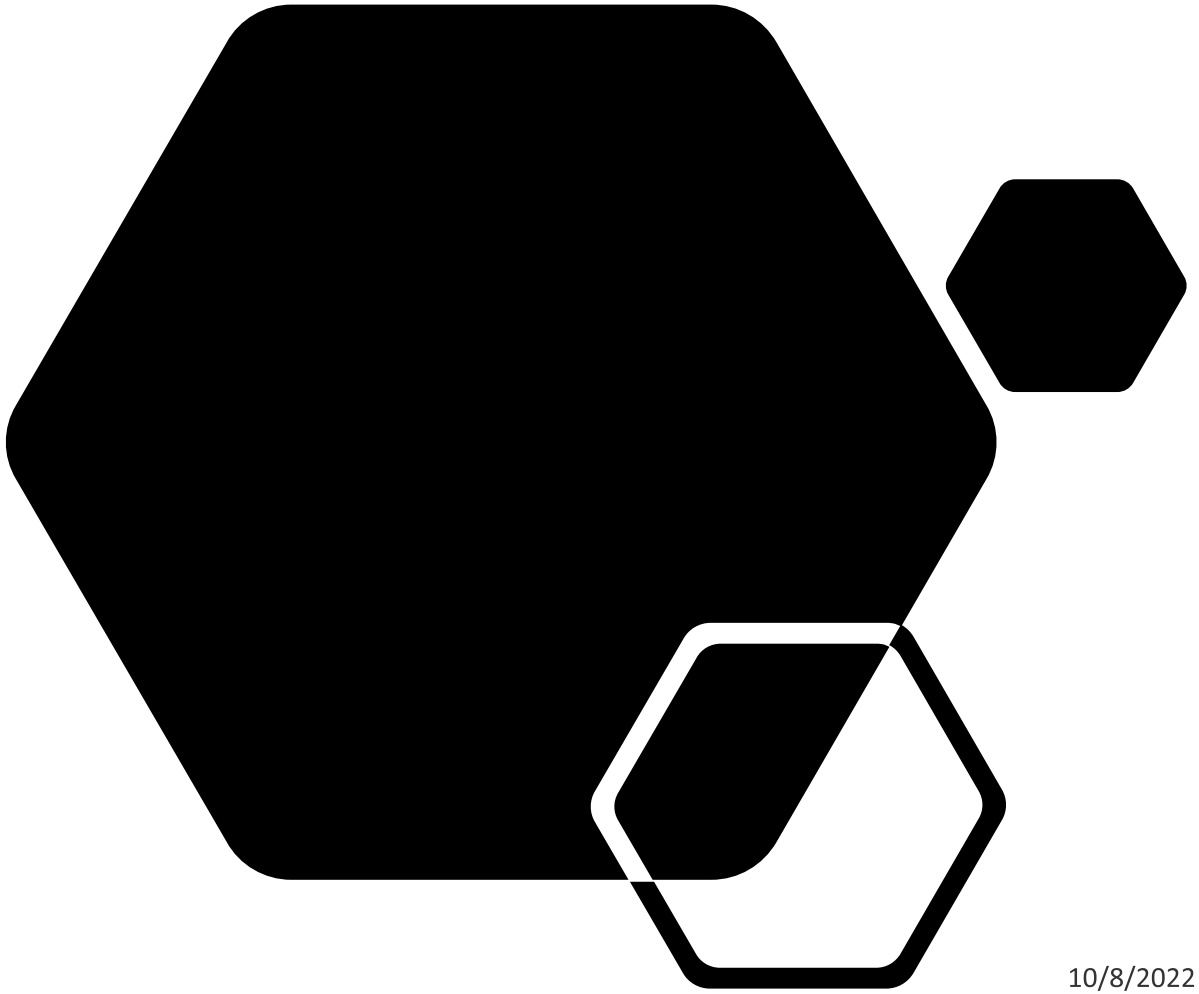
Miller

Manipularea prin date

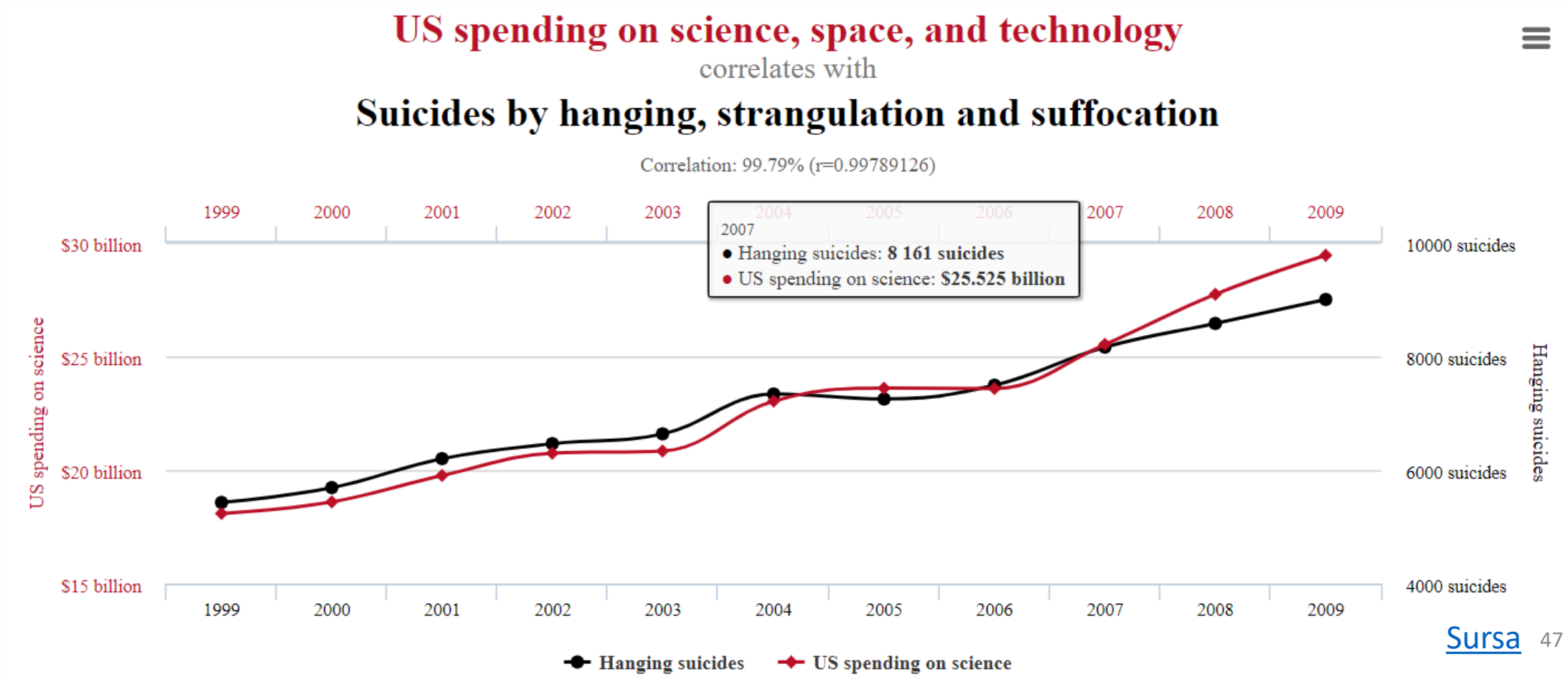
Co-incidență vs. cauzalitate

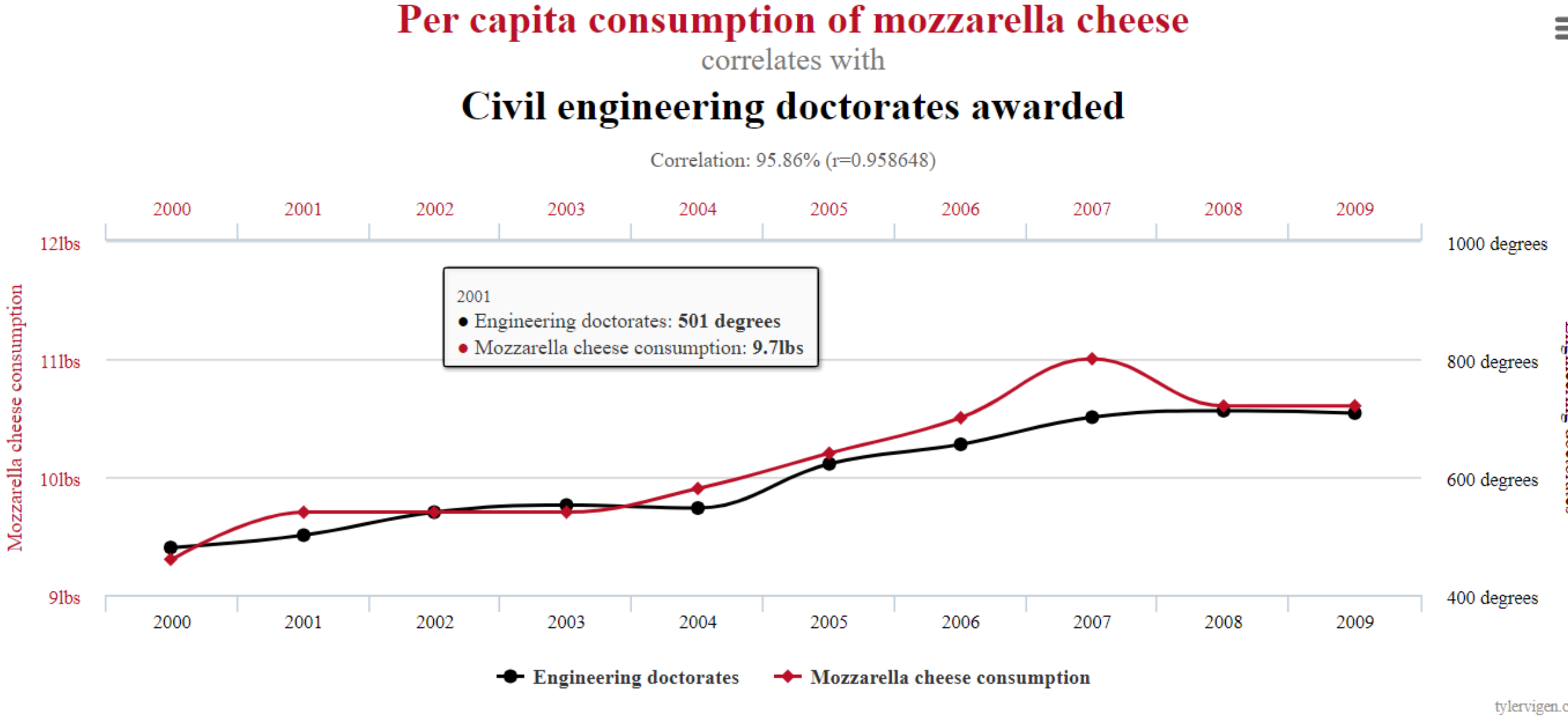
Vizualizări ciuntite

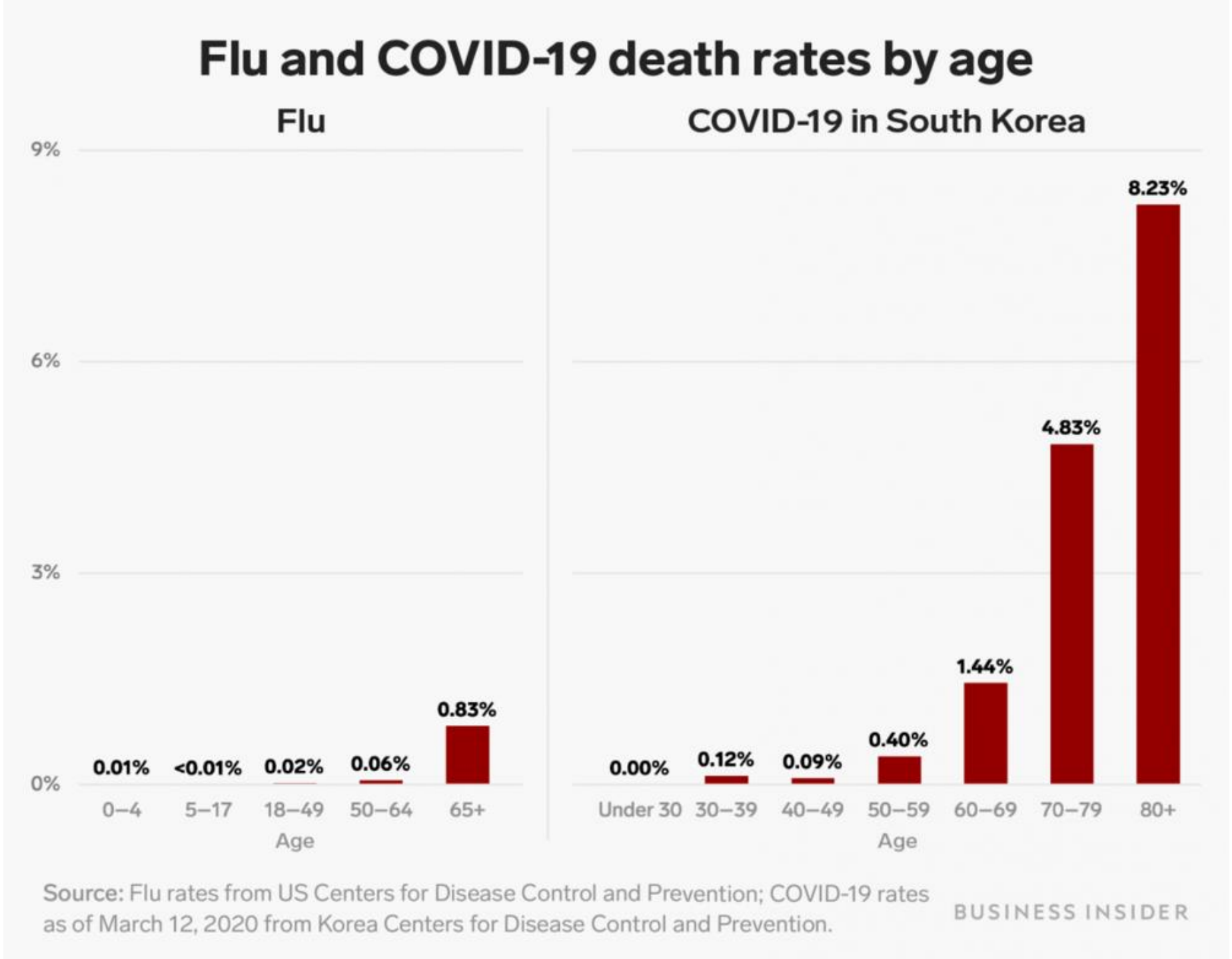
Vizualizări exagerate



Spurious correlations

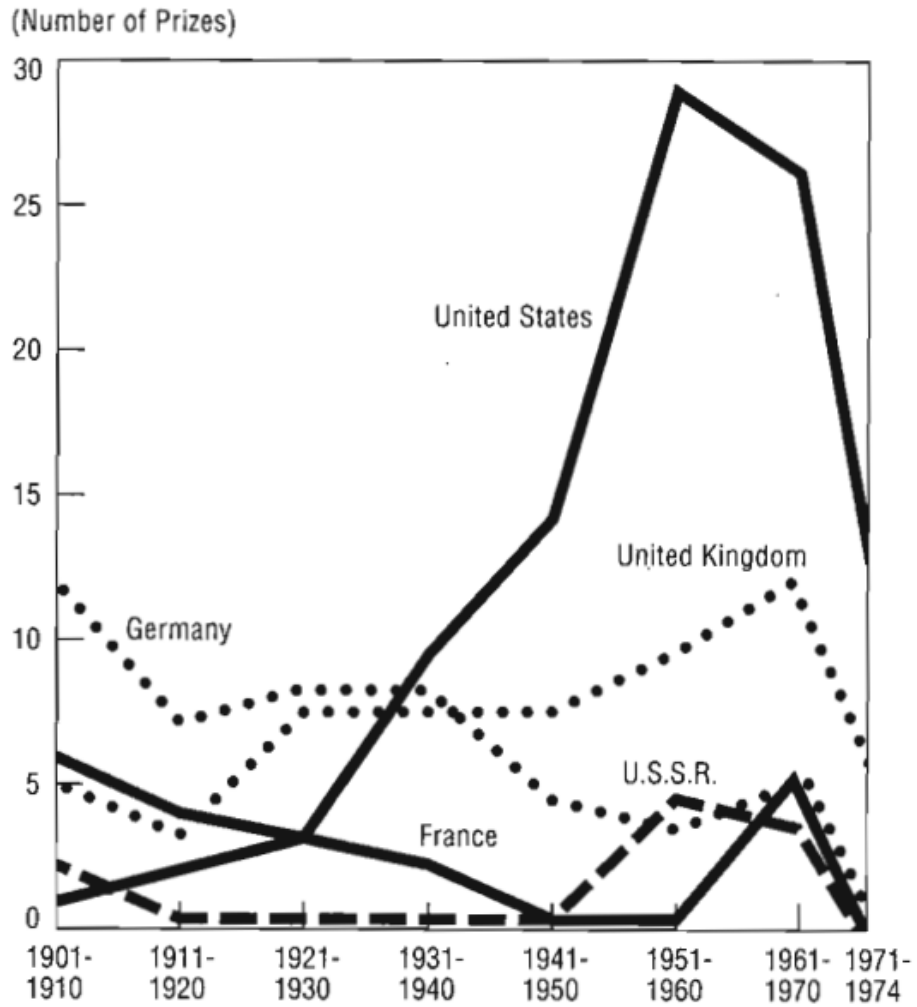



Per capita consumption of mozzarella cheese
correlates with
Civil engineering doctorates awarded
Correlation: 95.86% (r=0.958648)
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
12lbs
11lbs
10lbs
9lbs
1000 degrees
800 degrees
600 degrees
400 degrees
Mozzarella cheese consumption
Engineering doctorates
2001
Engineering doctorates: 501 degrees
Mozzarella cheese consumption: 9.7lbs
Engineering doctorates
Mozzarella cheese consumption
tylervigen.co
Data sources: U.S. Department of Agriculture and National Science Foundation

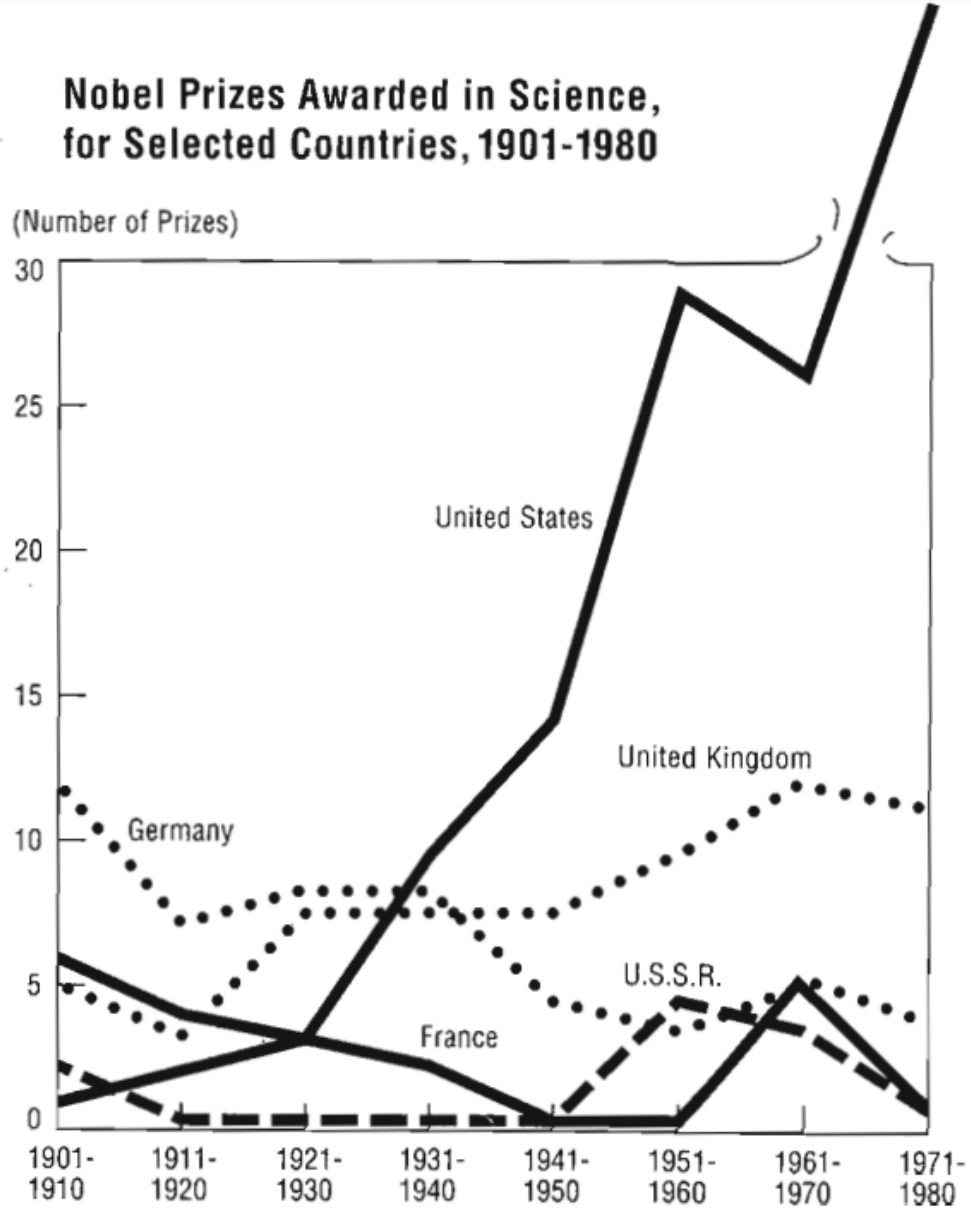

Flu and COVID-19 death rates by age
Flu
COVID-19 in South Korea
9%
6%
3%
0%
0.01%
<0.01%
0.02%
0.06%
0.83%
0–4
5–17
18–49
50–64
65+
Age
0.00%
0.12%
0.09%
0.40%
1.44%
4.83%
8.23%
Under 30
30–39
40–49
50–59
60–69
70–79
80+
Age
Source: Flu rates from US Centers for Disease Control and Prevention; COVID-19 rates as of March 12, 2020 from Korea Centers for Disease Control and Prevention.
BUSINESS INSIDER

Sursa

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



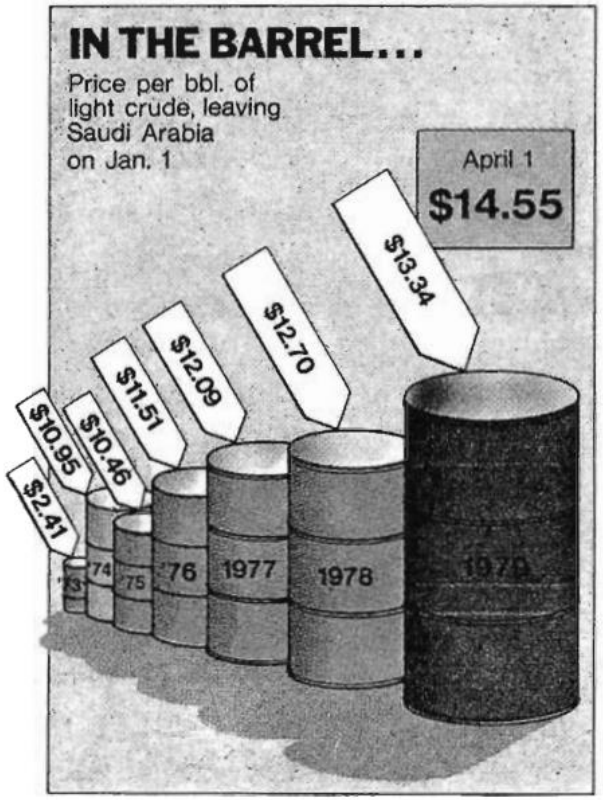
Nobel Prizes Awarded in Science, for Selected Countries, 1901-1980



Sursa: Tufte 2001, p. 60

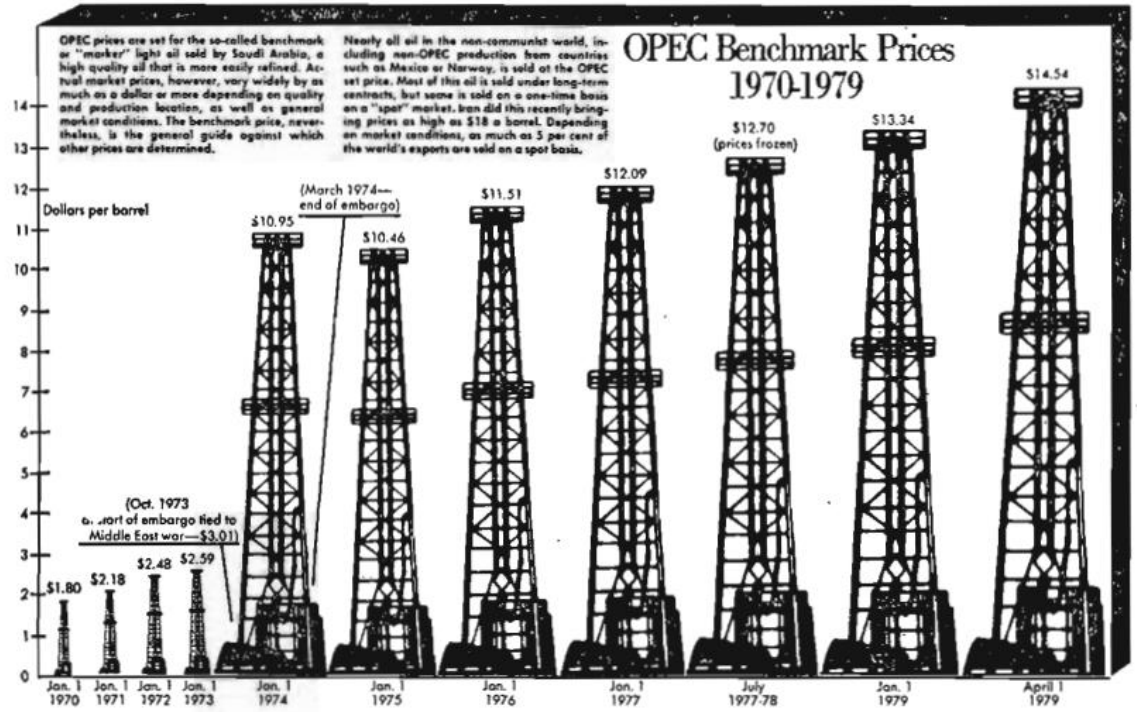
National Science Foundation, *Science Indicators, 1974* (Washington, D.C., 1976), p. 15.

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:



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

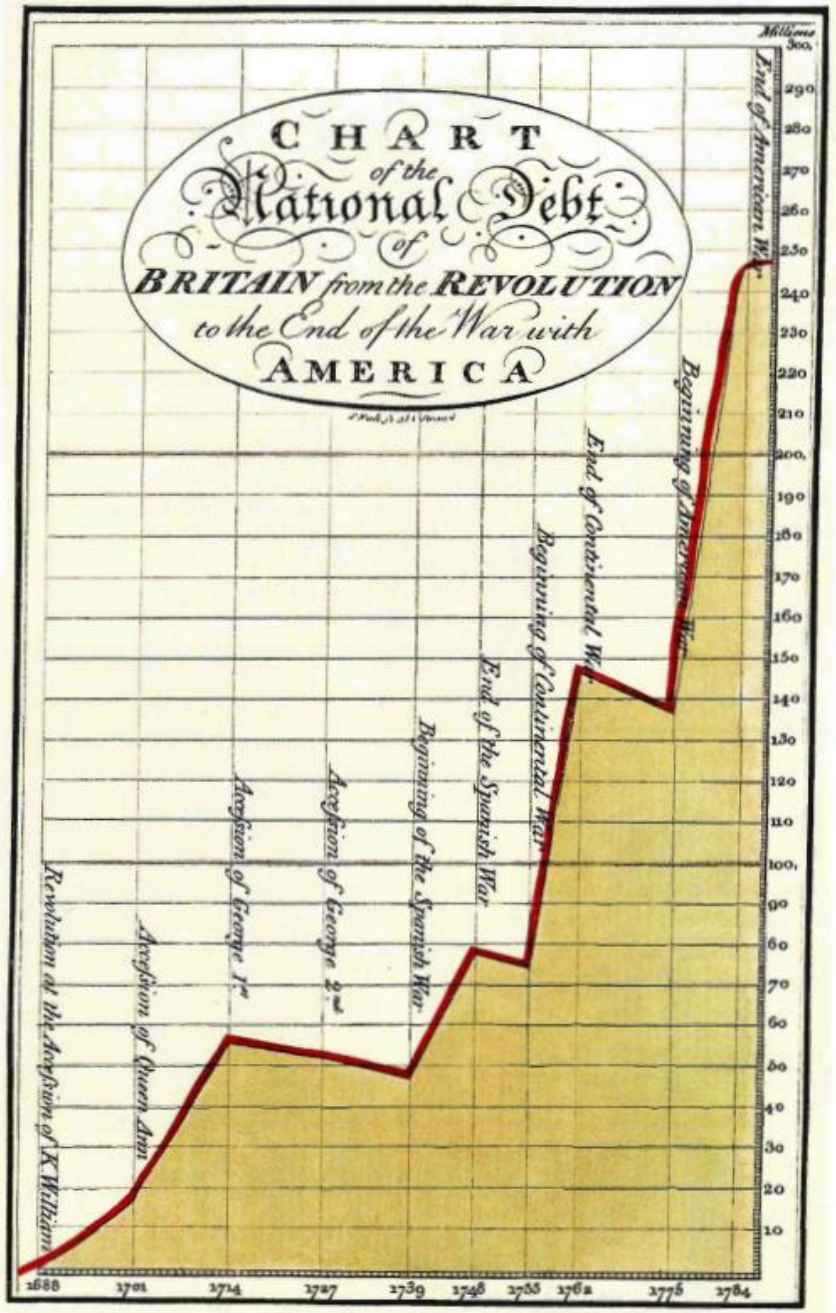
Washington Post, March 28, 1979, p. A-18.

Sursa: Tufte 2001, p. 62

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



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

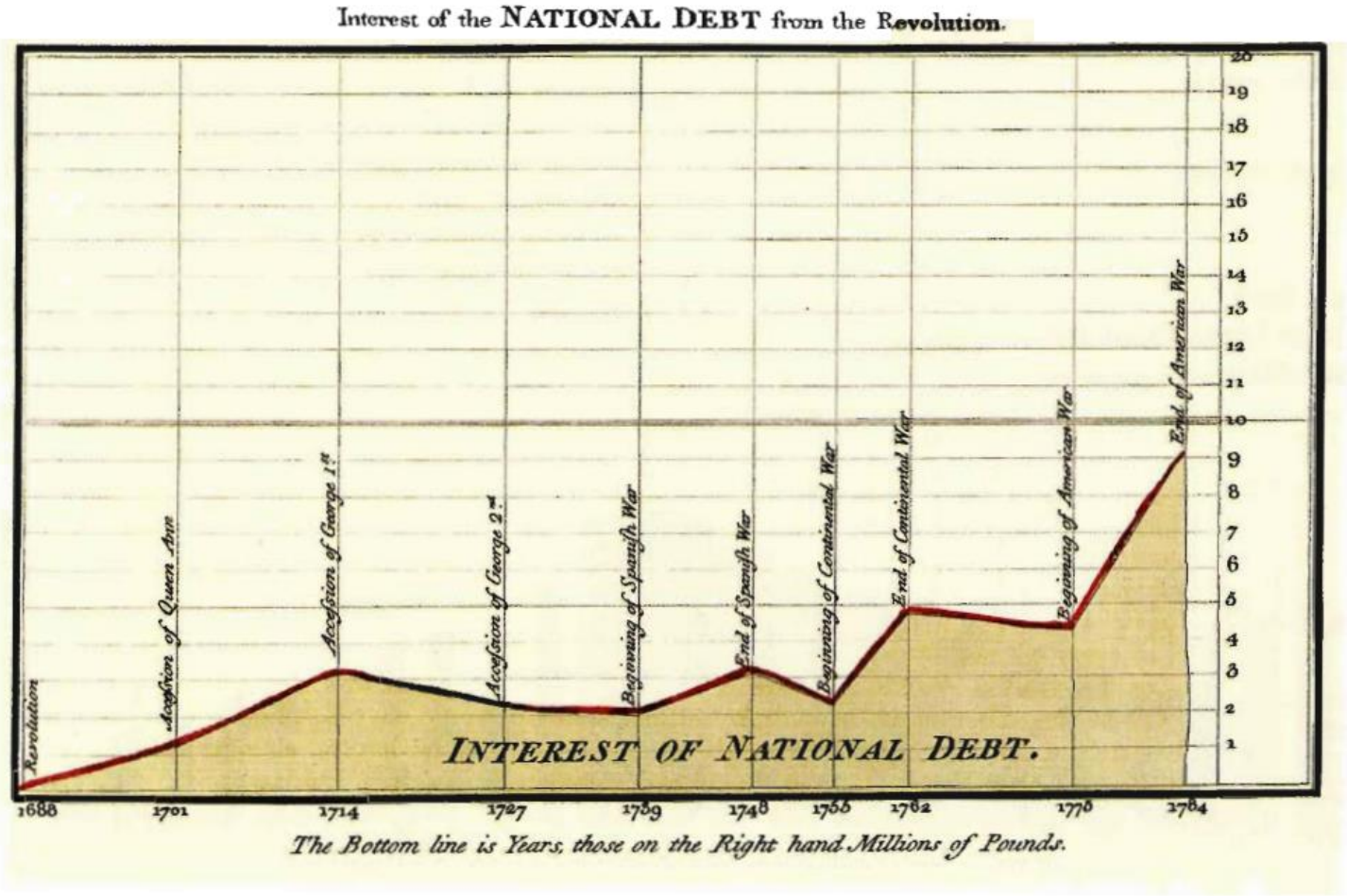
The Case of Skyrocketing Government Spending

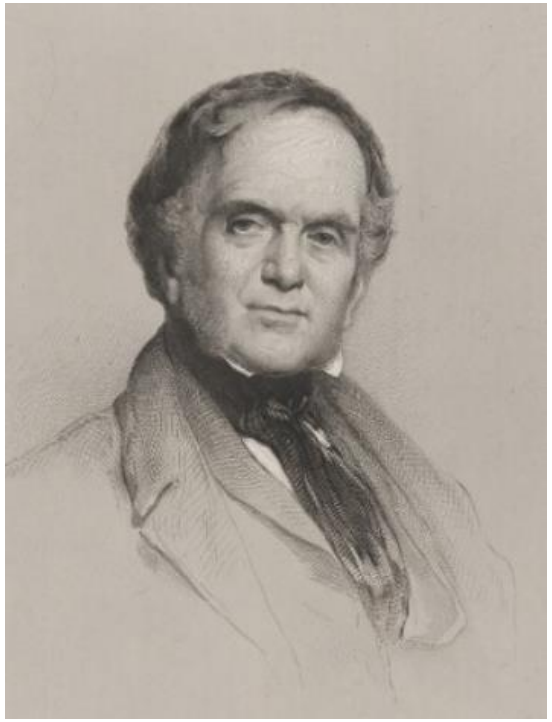
Probably the most frequently printed graphic, other than the daily weather map and stock-market trend line, is the display of government spending and debt over the years. These arrays nearly always create the impression that spending and debt are rapidly increasing.

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):

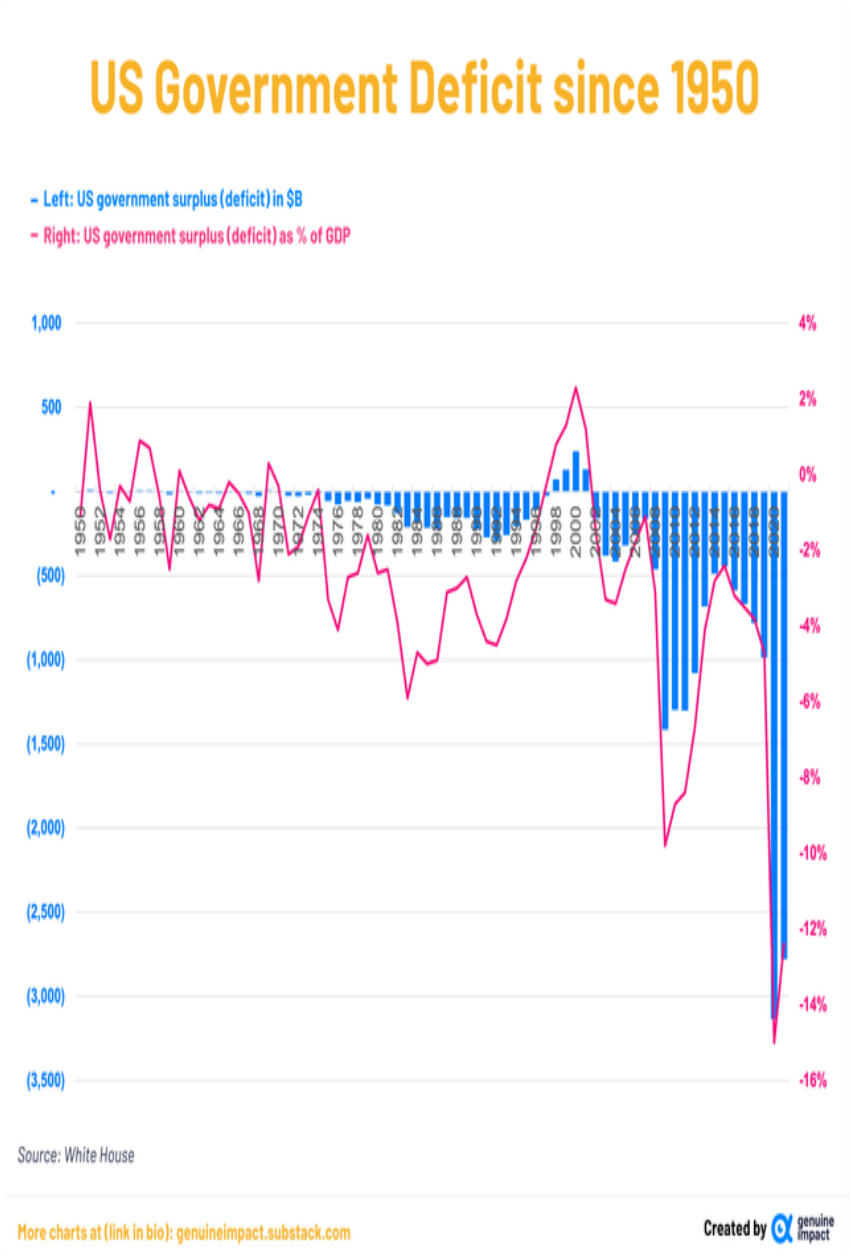
Sursa: Tufte 2001, p. 64

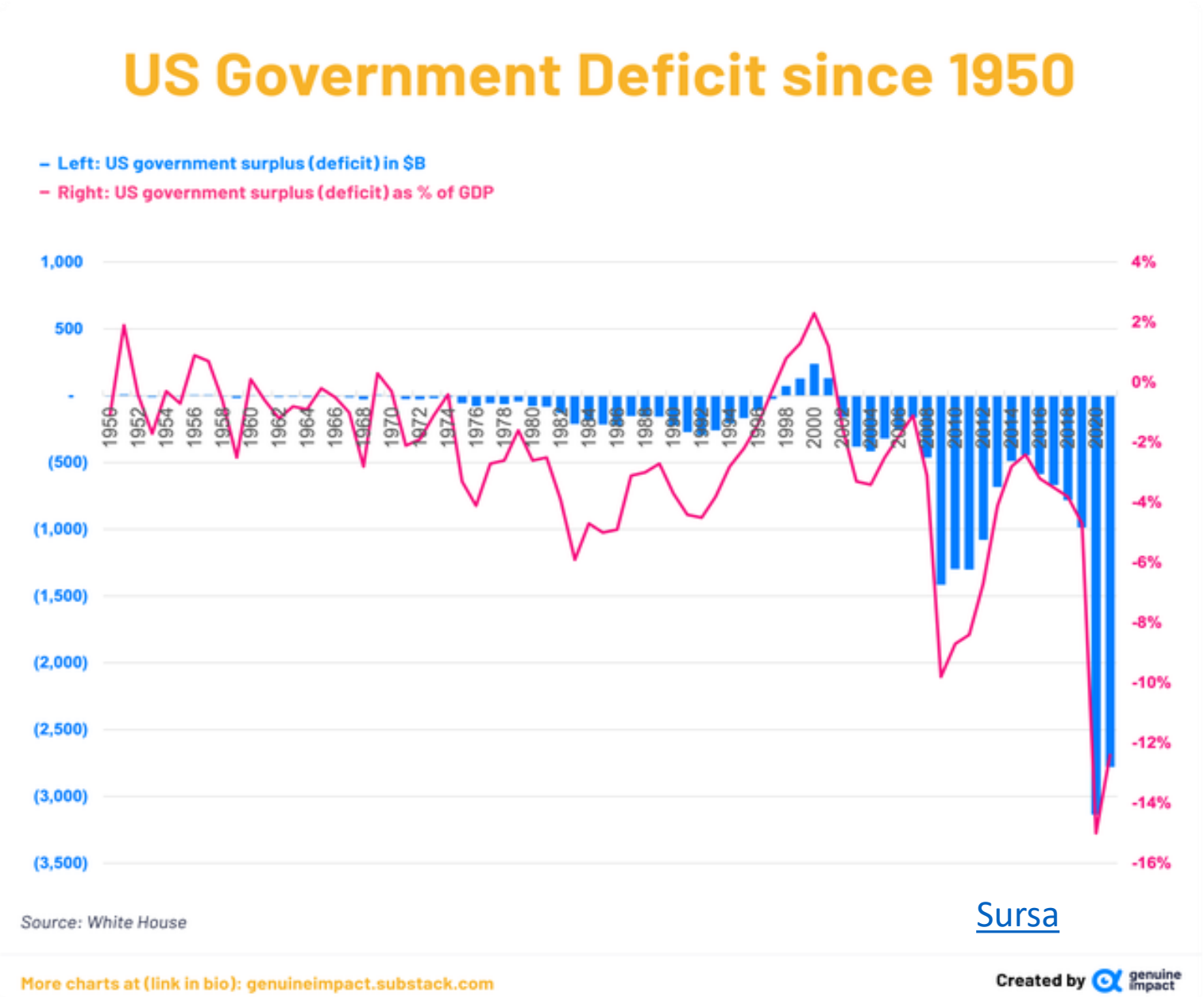




William Playfair

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)

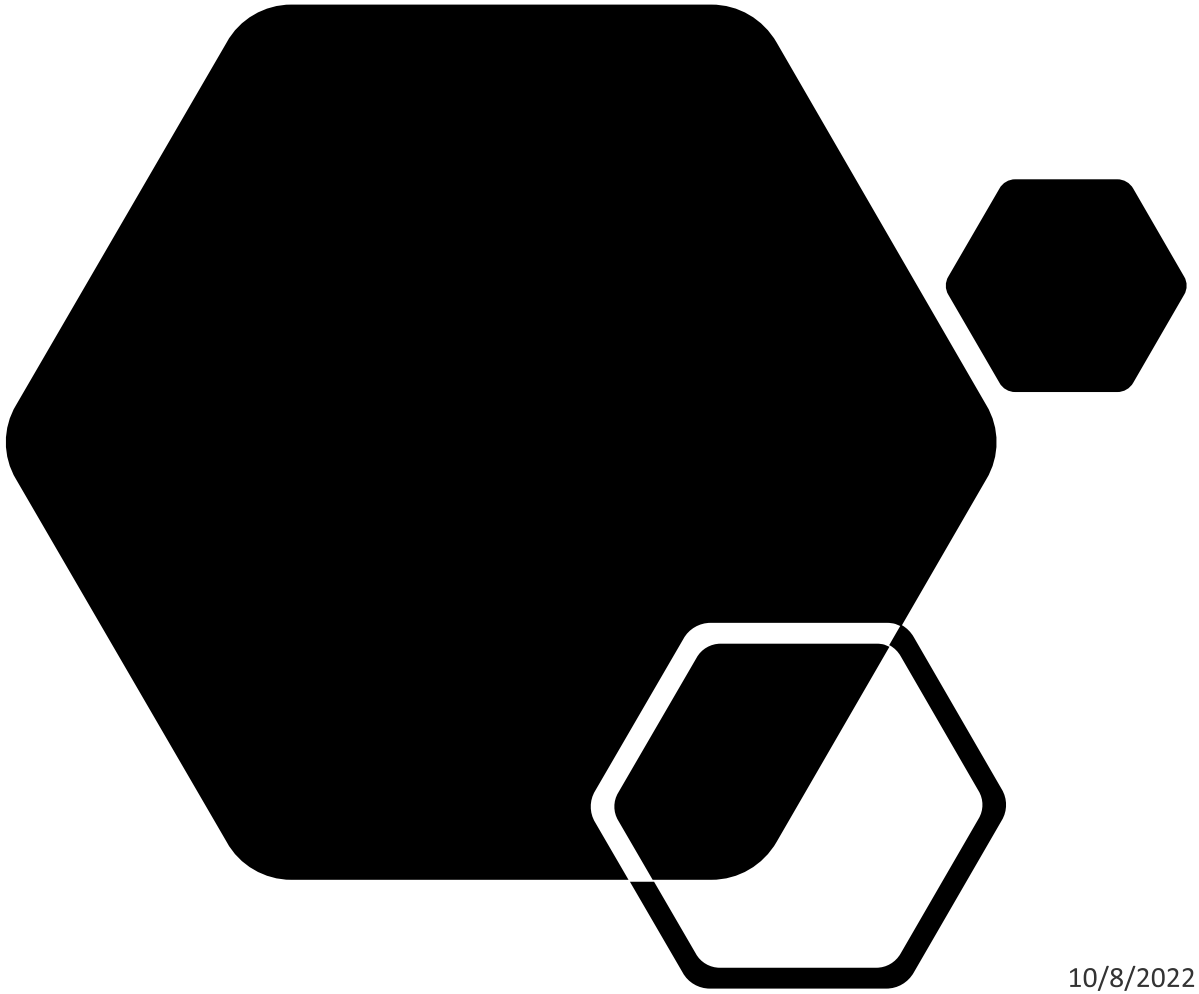

US Government Deficit since 1950
– Left: US government surplus (deficit) in $B
– Right: US government surplus (deficit) as % of GDP
Source: White House
More charts at (link in bio): genuineimpact.substack.com
Created by genuine impact


US Government Deficit since 1950
– Left: US government surplus (deficit) in $B
– Right: US government surplus (deficit) as % of GDP
Source: White House
More charts at (link in bio): genuineimpact.substack.com
Created by genuine impact

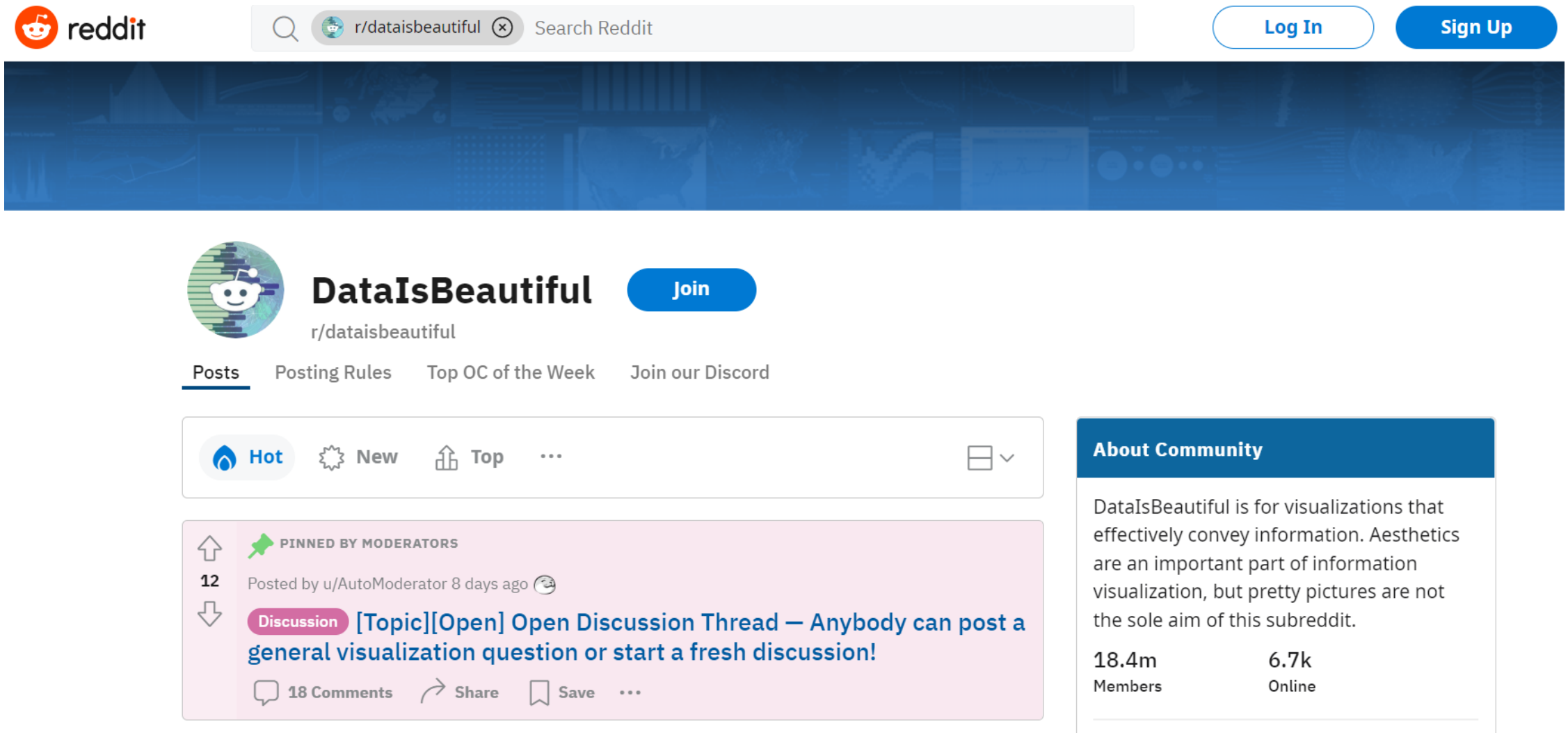
Sursa

Toate’s vechi și nouă toate

Colecții de vizualizări de date



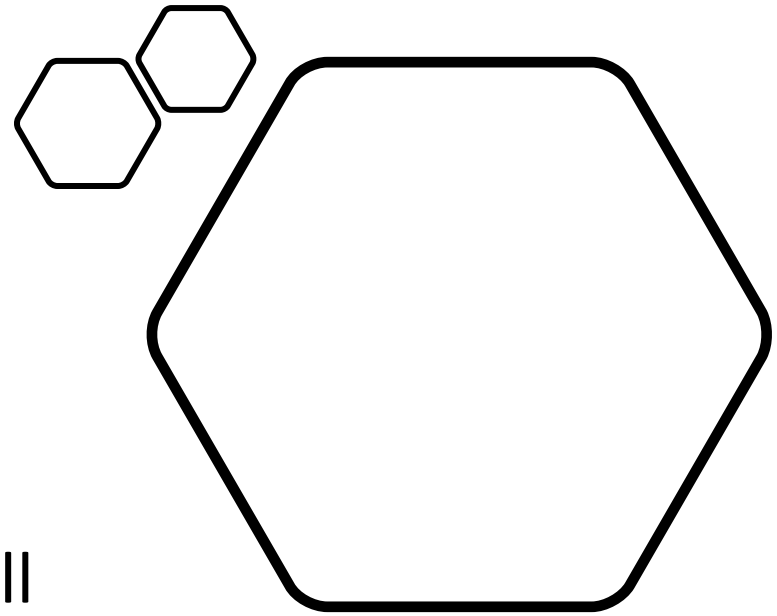
Click pe [Launch Interactive](#)!



Reddit – Data Is Beautiful

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.

„Efectul de iulie” în erorile fatale de medicație

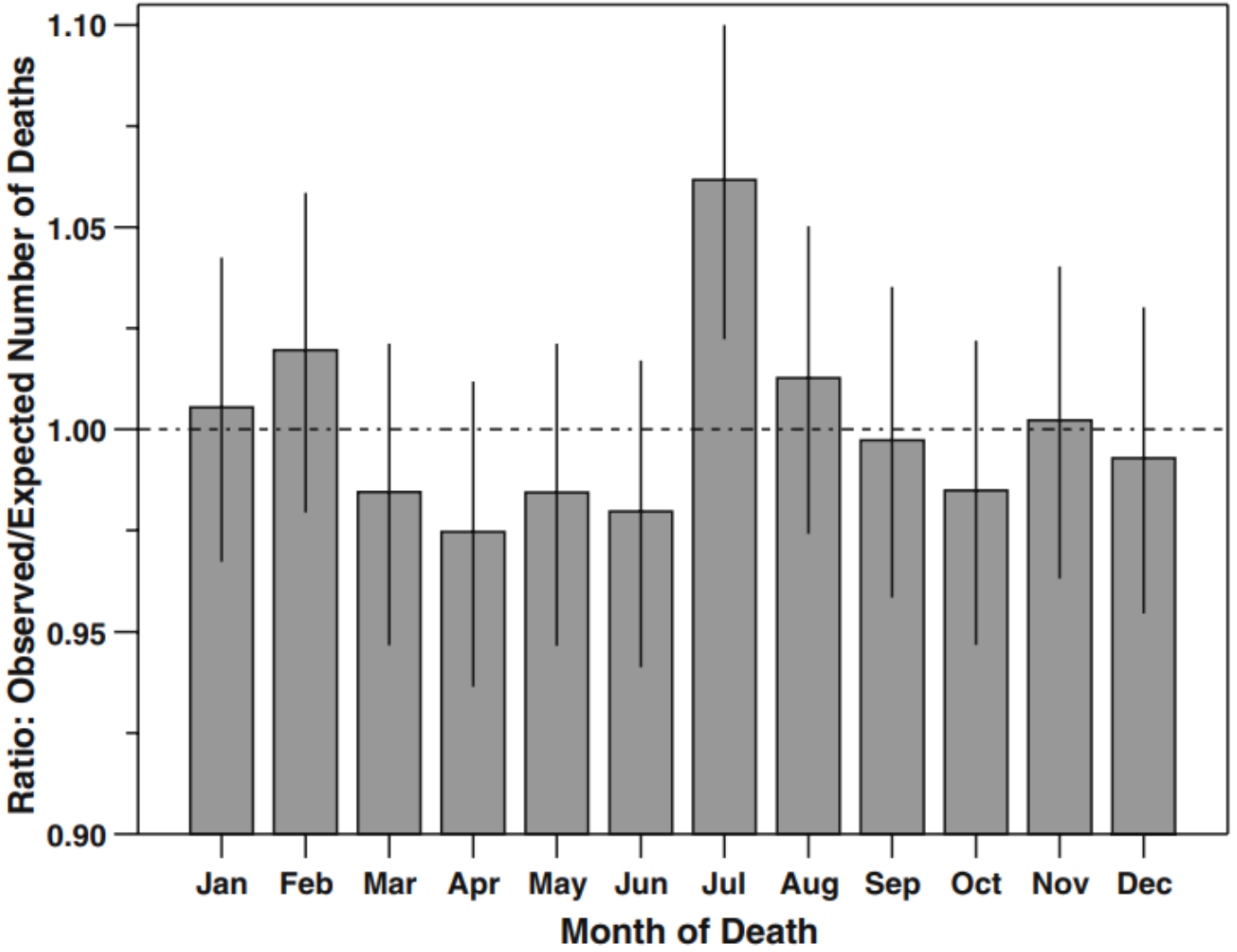


Figure 1. Ratio of observed to expected deaths for inpatient medication errors by month, United States, 1979-2006 (with 95% confidence intervals). Unless otherwise noted, error bars in Figure 1 and in subsequent figures were determined using a poisson approximation.[46]

Sursa

Criteriile lui Hill

Each July thousands begin medical residencies and acquire increased responsibility for patient care. ... these new medical residents may produce errors and worsen patient outcomes—the so-called “July Effect”	Patern sistematic (surse multiple de măsurare / observație)
Inside medical institutions, in counties containing teaching hospitals, fatal medication errors spiked by 10% in July and in no other month	Intensitate Temporalitate (înainte / după)
There was no July spike in counties without teaching hospitals. These findings held only for medication errors, not for other causes of death.	Specificitate Plauzibilitate
The greater the concentration of teaching hospitals in a region, the greater the July spike (r=.80; P=.005).	Doză / răspuns

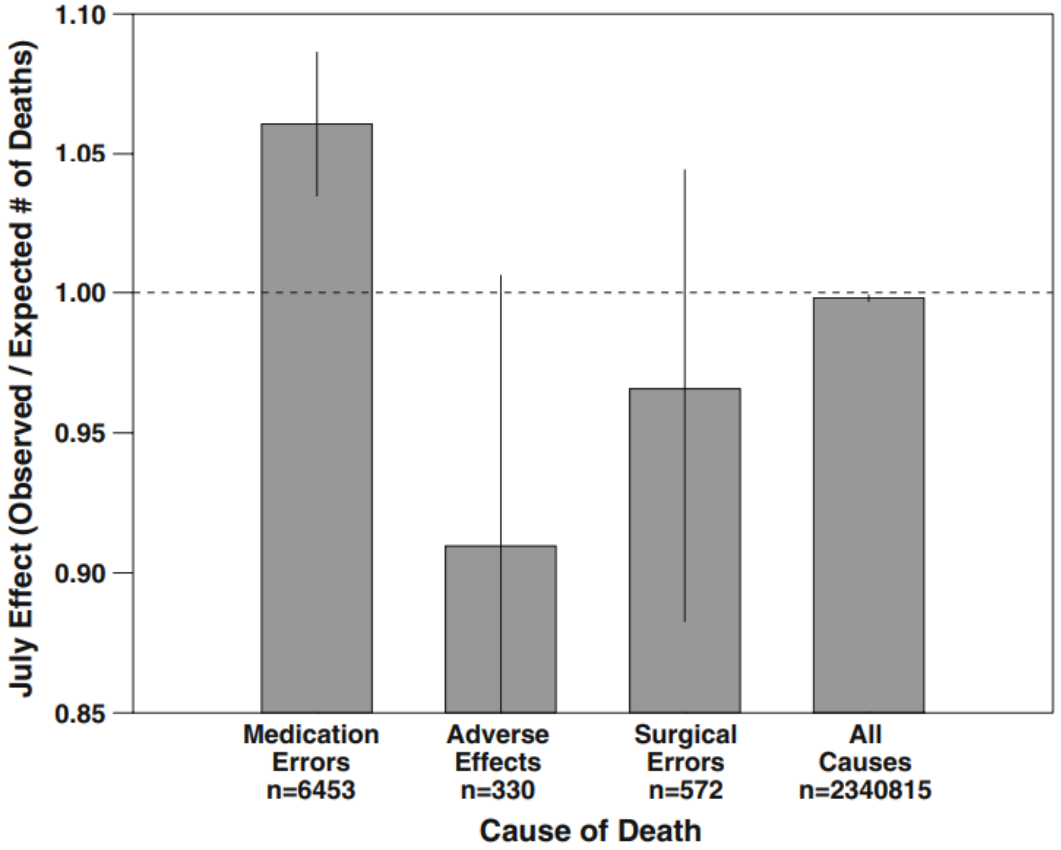
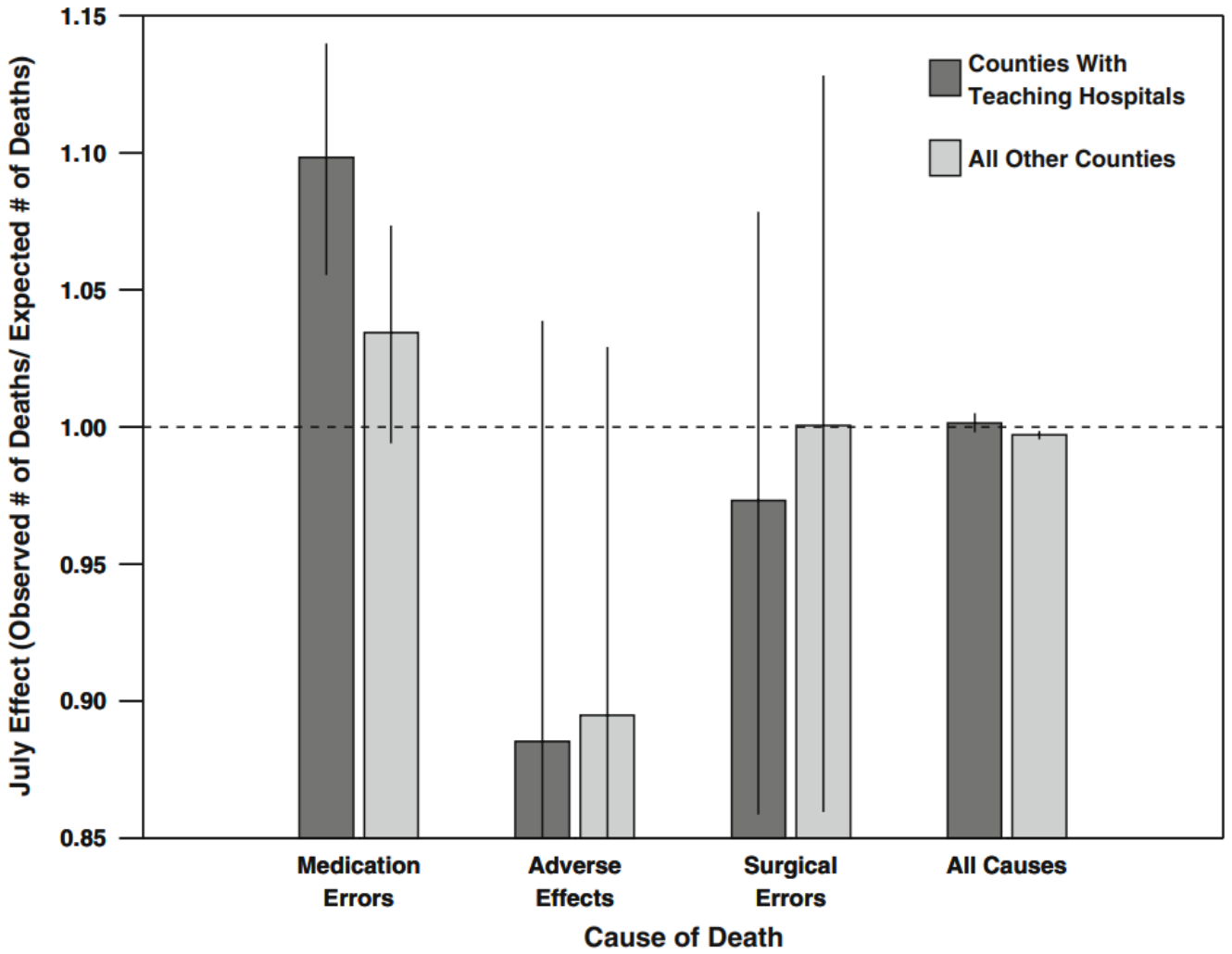
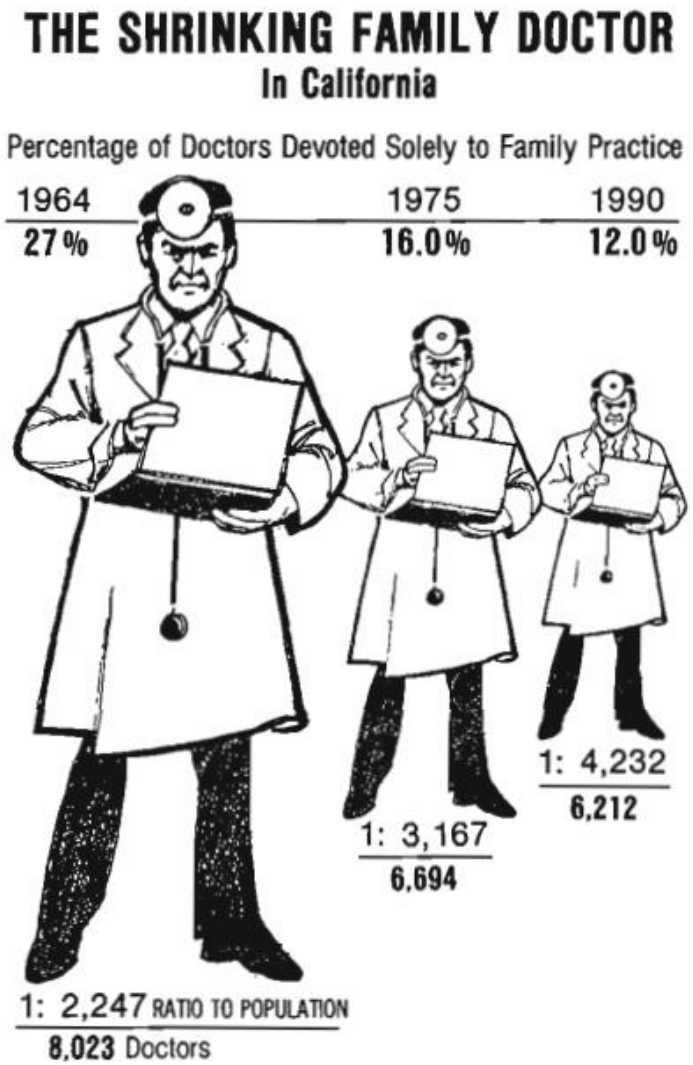


Figure 3. July effect for fatal medication errors and for comparison causes of death inside medical institutions, United States, 1979-2006 (with 95% confidence intervals).



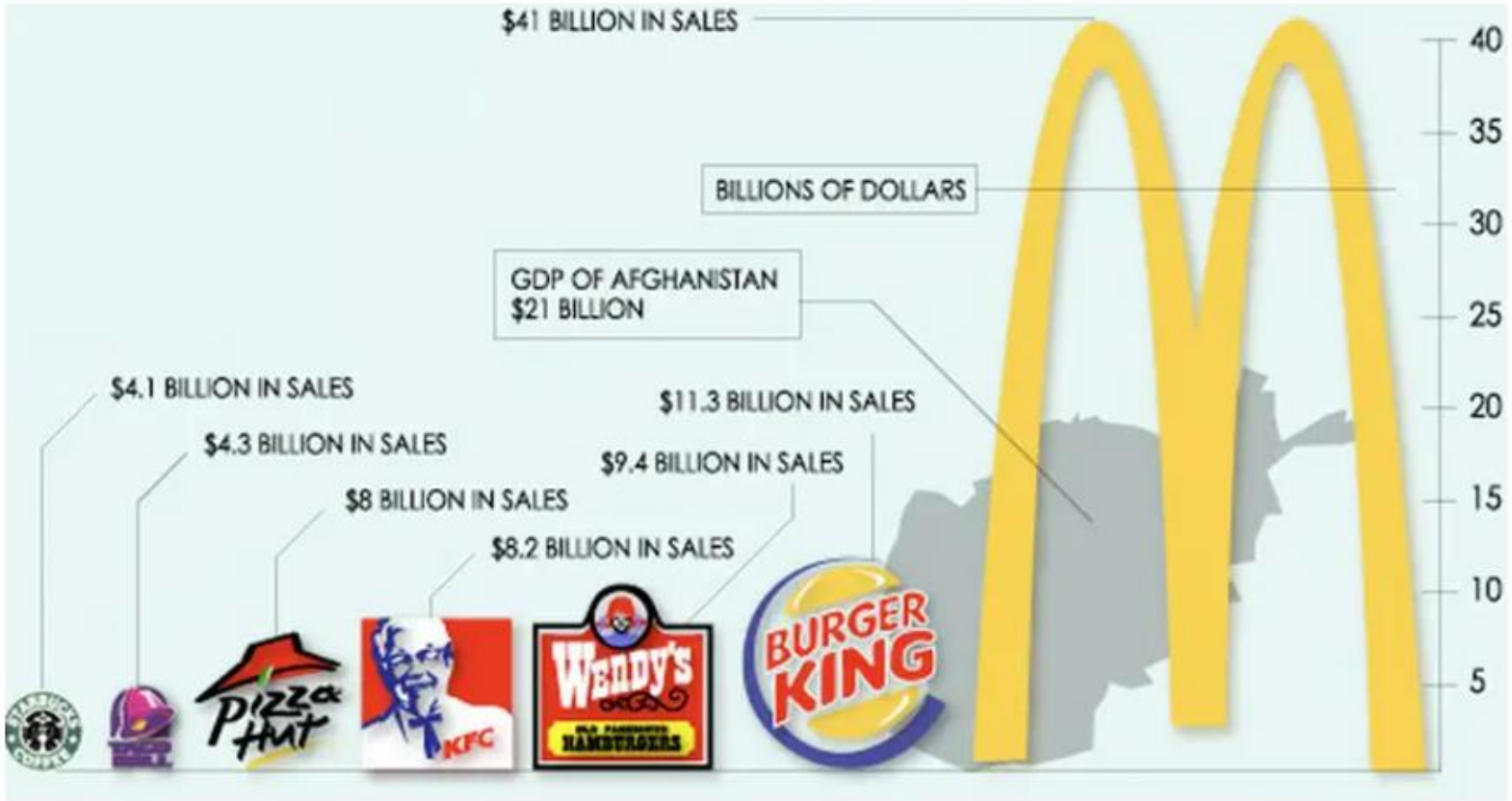
And here is the incredible shrinking doctor, with a Lie Factor of 2.8, not counting the additional exaggeration from the overlaid perspective and the incorrect horizontal spacing of the data:



Sursa: Tufte 2001, p. 69

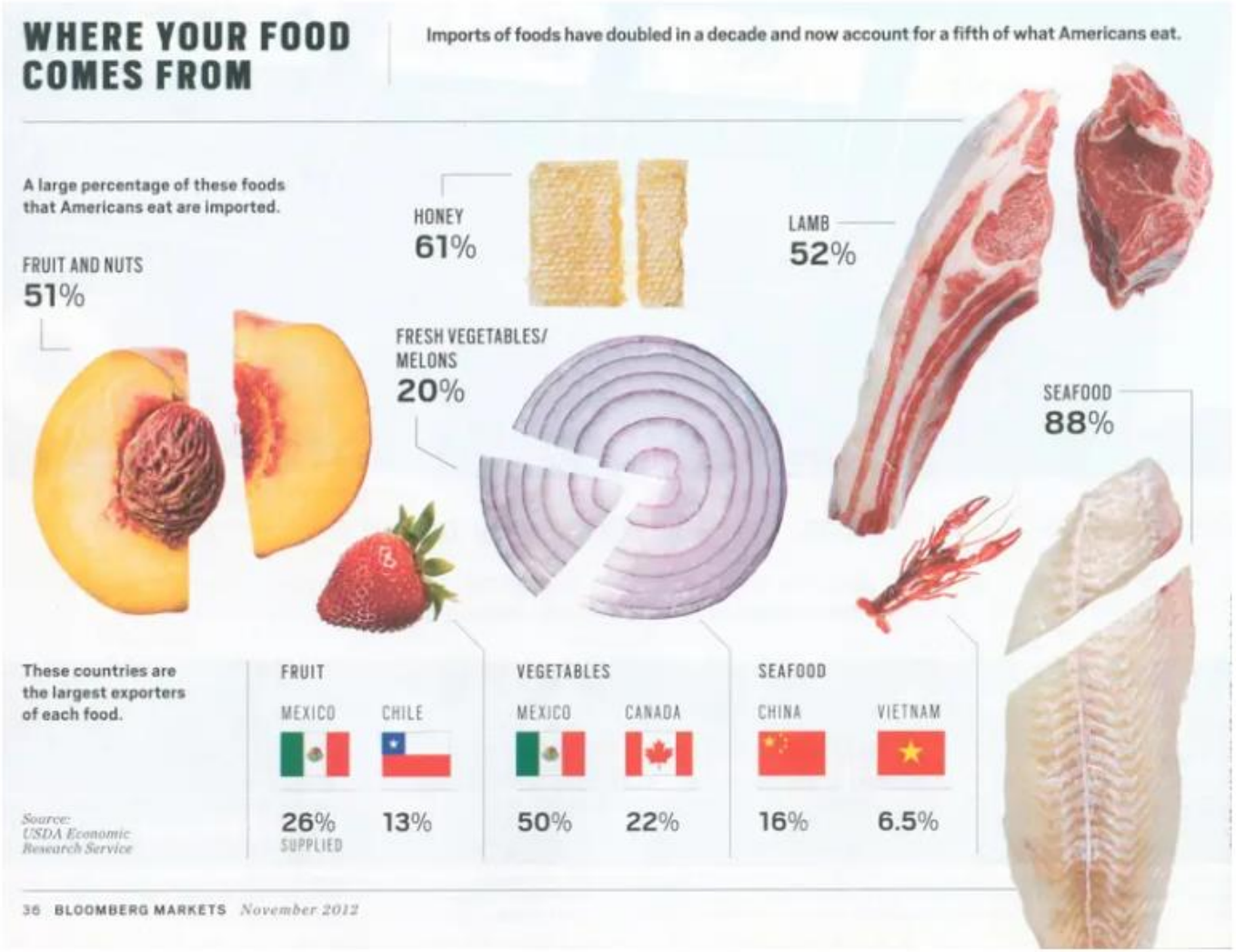
Los Angeles Times, August 5, 1979, p. 3.

Burger King has 3 times as much in sales than Starbucks. It makes sense that it's three times taller. But the fact that its area is nine times that of Starbucks shows why this chart exemplifies everything that is wrong with charts that try to incorporate cutesy graphics.



Sursa

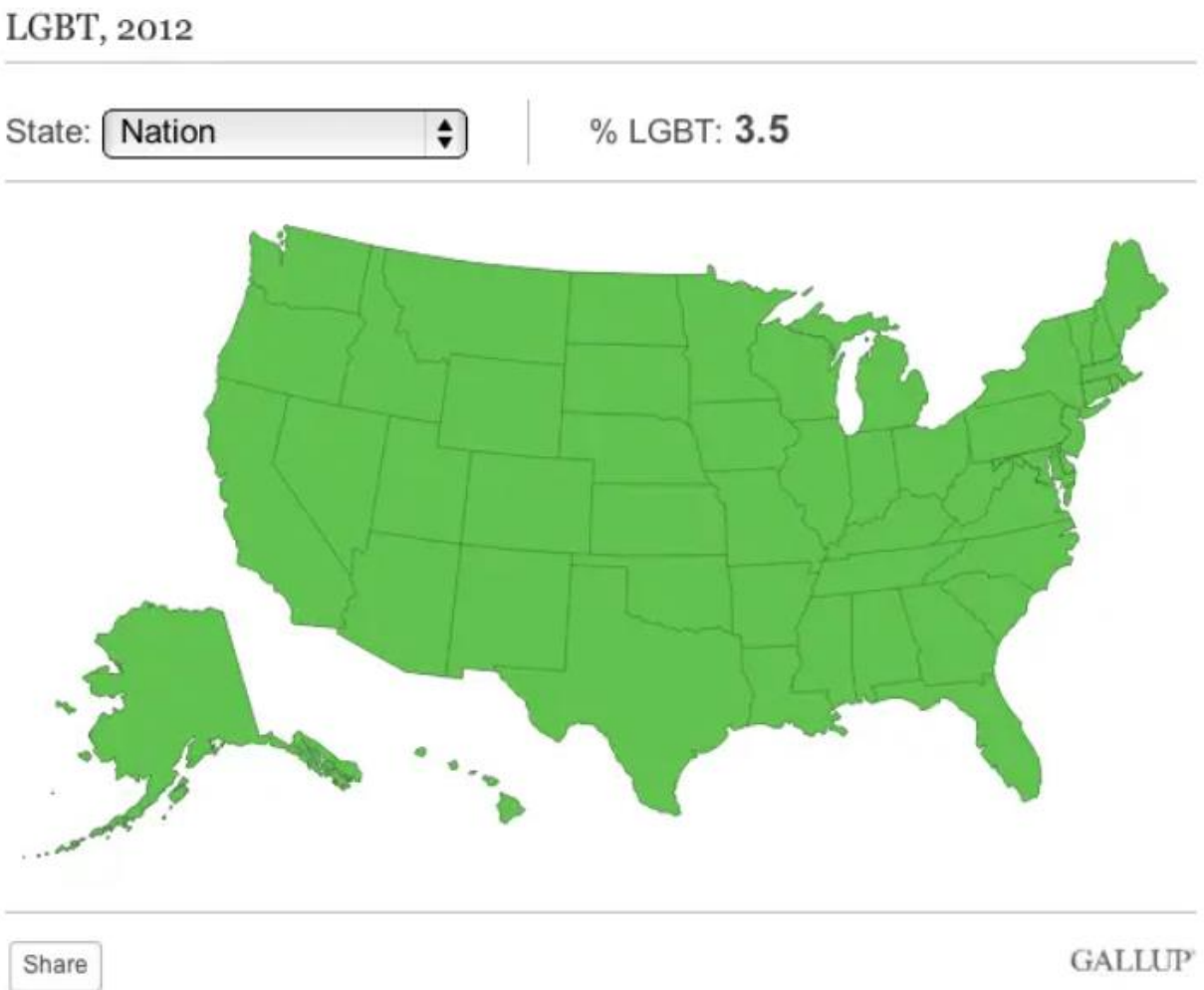
There's a lot going on with this Bloomberg chart. That doesn't seem like an evenly cut lamb chop, and while I'm not a biologist I have a strong feeling an onion is not a melon.



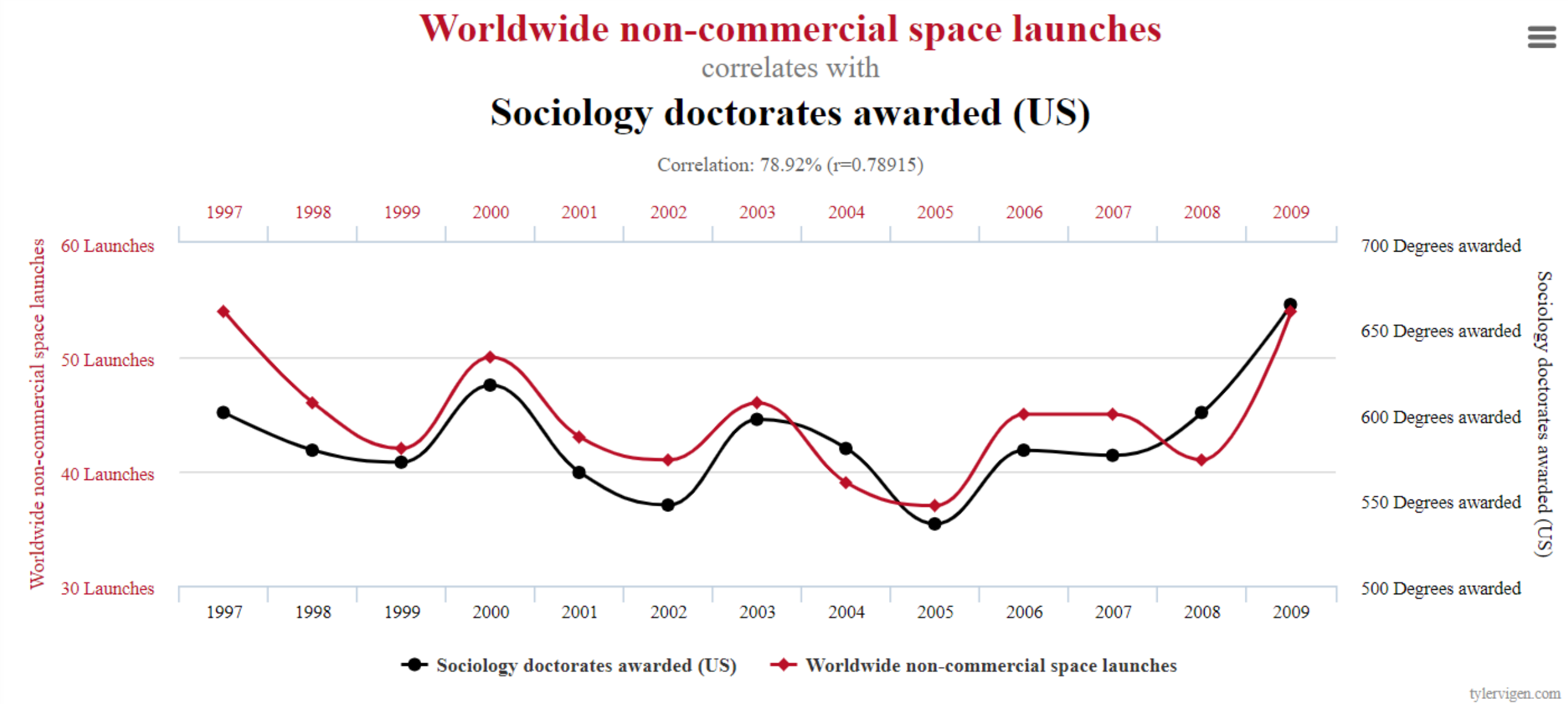
JunkCharts

Sursa

Gallup polled to find out the percent of LGBT individuals in each state. Unsurprisingly, all of the states were rather close — from 1.7% in North Dakota to 5.1% in Hawaii — but their coloration system means that the whole map is green. Whoops.



Sursa



Data sources: Federal Aviation Administration and National Science Foundation