Curs 06 – Analiza datelor în lumea reală

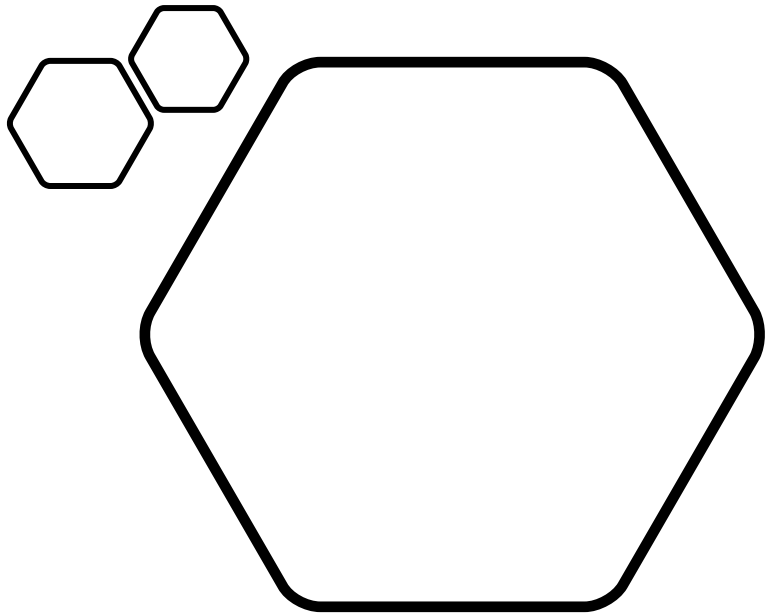
Obiective ale tehnicilor de analiză - recapitulare

- Explorarea datelor
  - Frecvențe, medii și corelații
- Măsurare: analiza factorială / reducerea dimensionalității
  - Măsurarea constructelor latente prin indicatori vizibili
  - Identificarea dimensiunilor personalității pe baza acțiunilor și preferințelor
- Clasificare: analiza cluster
  - Clasificarea indivizilor în tipuri: „Spune-mi cu cin’ te-nsoțești, ca să-ți spun cine ești”
- Explicare prin factori externi: analiza de regresie
- Explicare prin contagiune: analiza de rețea
- Extrapolare: seriile de timp
  - Explicarea acțiunilor prin factori externi și patternuri temporale
  - Tendințe inerțiale, sezoniere, variații aleatorii

Structura cursului

1. Why?
2. Cauzalitate
3. Măsurare
4. Modelare și eșantionare
5. Tehnici de analiză
6. Analiza datelor în RL
   - Studiu de caz – relații de cuplu
   - Studiu de caz – cauzalitate
   - Studiu de caz – măsurători și erori
   - Părtiniri în analiza datelor
7. Programare și ML
8. Why Privacy? (16.11.2023)
9. Anonimizarea, de-identificarea si pseudonimizarea datelor
10. Homomorphic Encryption. PIR
11. Differential Privacy
12. Membership inference
13. Federated Architecture. Multi-Party Computation
14. Zero proof. Blockchain architectures

Studiu de caz



Tehnici de analiză a datelor

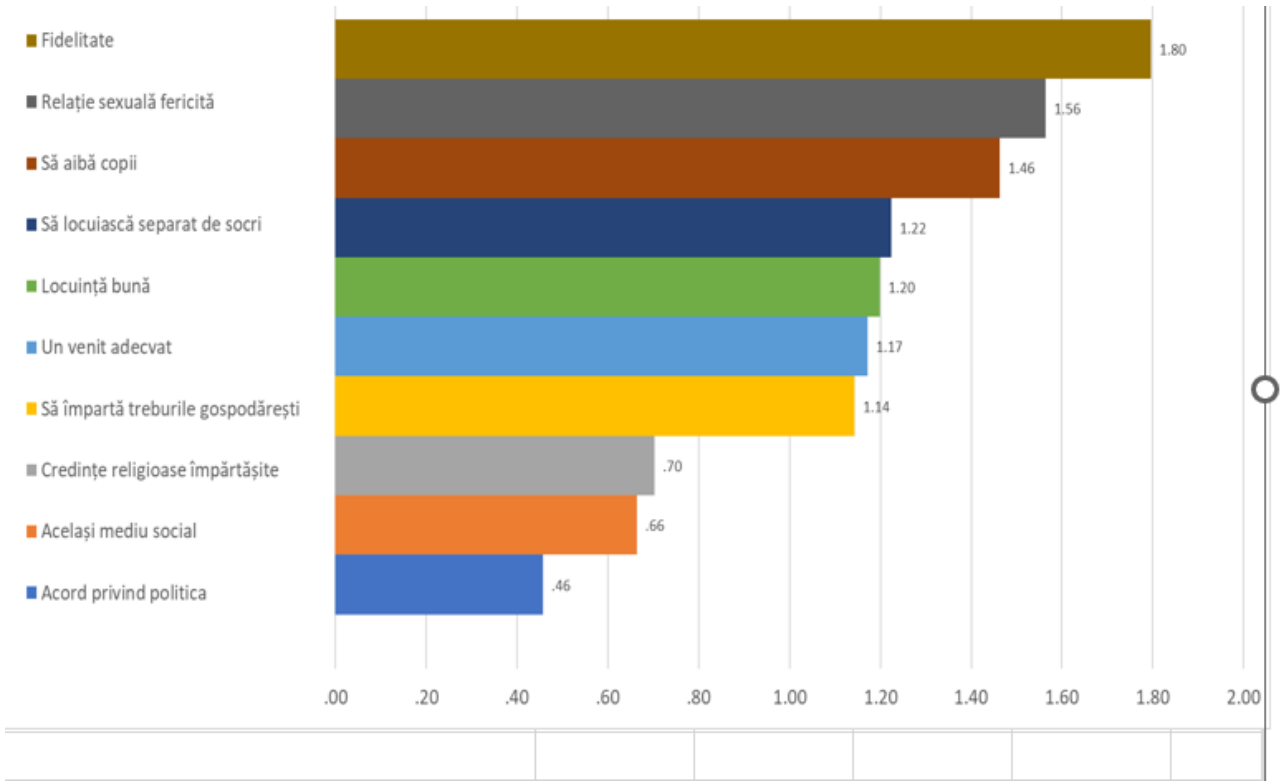
Condiții ale unei relații de cuplu de succes

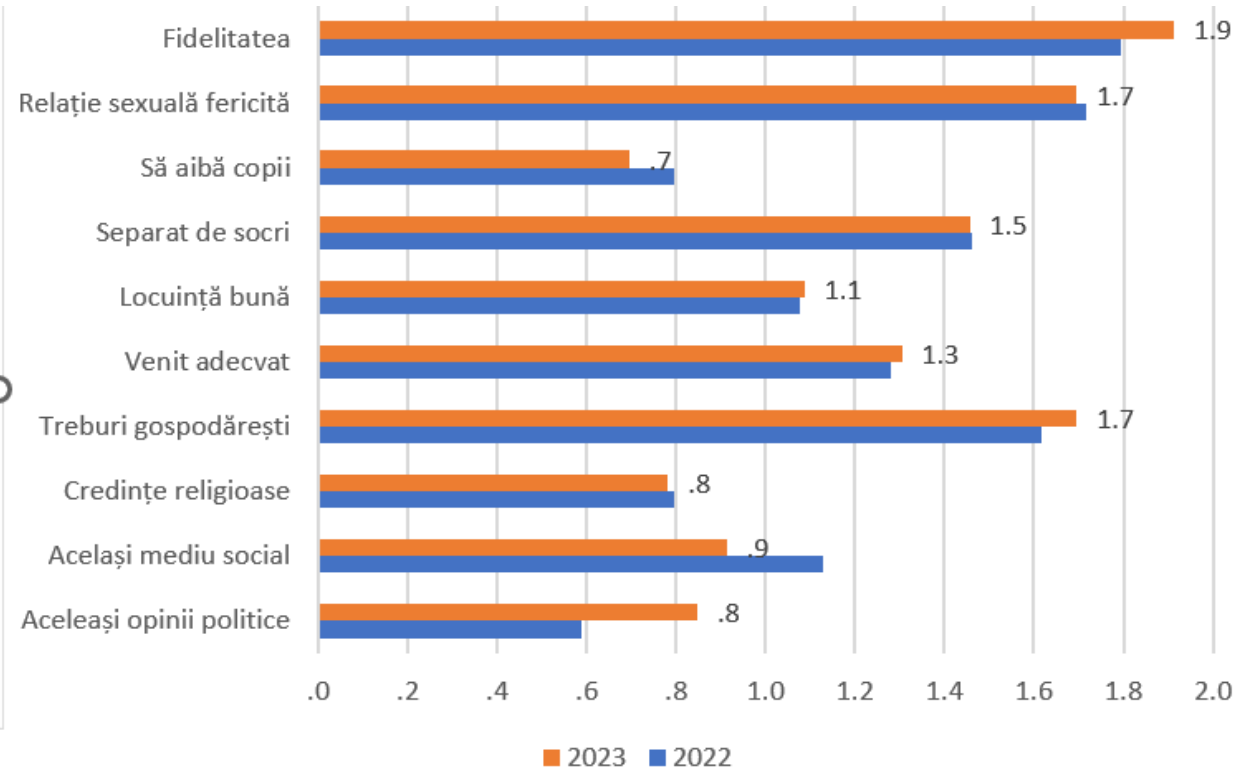
Studiu de caz

Aici sunt câteva aspecte despre care unii oameni cred că sunt importante pentru o căsătorie de succes. Te rog, pentru fiecare, spune-mi cum crezi că este, pentru succesul unei căsătorii: (2=foarte important, 1=destul de important, 0=nu prea important)

- Fidelitatea
- Partenerii să fie din același mediu social
- Să aibă aceeași religie
- Să aibă o locuință bună
- Să aibă un venit adecvat
- Să fie de acord în chestiunile politice
- Să trăiască separat de socri
- Să aibă o relație sexuală fericită
- Să împartă treburile domestice
- Să aibă copii

Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia





Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia

Analiză participanților la cursul de PR ce au completat chestionarul, 2023 și 2022

Tabelul de corelații - EVS

	Fidelitate	Copii	Venit	Casă	Același mediu	Aceeași religie	Acord în politică	Locuiesc singuri	Relație sexuală	Împart treburile
Fidelitate	1	.223	.079	.088	.058	.160	.028	-0.021	.041	.085
Copii	.223	1	.138	.177	.102	.192	.058	.034	.111	.175
Venit	.079	.138	1	.454	.298	.146	.143	.090	.134	.096
Casă	.088	.177	.454	1	.252	.191	.186	.116	.144	.185
Același mediu	.058	.102	.298	.252	1	.355	.301	.079	.027	.063
Aceeași religie	.160	.192	.146	.191	.355	1	.316	-.013	-.027	.074
Acord în politică	.028	.058	.143	.186	.301	.316	1	.065	.051	.122
Locuiesc singuri	-.021	.034	.090	.116	.079	-.013*	.065	1	.247	.137
Relație sexuală	.041	.111	.134	.144	.027	-.027	.051	.247	1	.233
Împart treburile	.085	.175	.096	.185	.063	.074	.122	.137	.233	1

Coeficienți de corelație Bravais-Pearson

Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia

Tabelul de corelații - ACS

	Fidelitate	Copii	Venit	Casă	Locuiesc singuri	Relație sexuală	Același mediu	Aceeași religie	Acord politic	Împart treburile
Fidelitate	1	.118	.025	.079	.177	-.096	.160	.212	.106	-.005
Copii	.118	1	-.178	-.166	-.318**	-.164	.067	.257*	-.134	-.044
Venit	.025	-.178	1	.639**	.173	.100	.161	.034	.020	.018
Casă	.079	-.166	.639**	1	.263*	.074	.084	.165	.107	.069
Locuiesc singuri	.177	-.318**	.173	.263*	1	.088	.016	-.104	.012	.009
Relație sexuală	-.096	-.164	.100	.074	.088	1	.059	-.080	.008	.067
Același mediu	.160	.067	.161	.084	.016	.059	1	.176	.064	.013
Aceeași religie	.212	.257*	.034	.165	-.104	-.080	.176	1	.261*	-.061
Acord politic	.106	-.134	.020	.107	.012	.008	.064	.261*	1	.145
Împart treburile	-.005	-.044	.018	.069	.009	.067	.013	-.061	.145	1

Analiza participanților la cursul de PR ce au completat chestionarul

Analiza factorială EVS

Patru factori selectați pentru 10 itemi: cum îi interpretăm?

Fiecare factor este o nouă variabilă

Fiecare respondent are o valoare pentru fiecare factor

Similaritate | **Intimitate** | **Familia tradițională** | **Banii NU contează**

	Factor			
	1	2	3	4
Fidelitate	-.002	-.070	.766	.005
Copii	.016	.118	.701	-.100
Venit	-.022	-.024	.003	-.875
Locuință bună	.038	.081	.077	-.771
Același mediu	.619	-.058	-.080	-.316
Aceeași religie	.737	-.121	.244	.028
Acord în politică	.792	.140	-.113	.078
Locuiesc singuri	.053	.680	-.224	-.035
Relație sexuală	-.122	.726	.053	-.087
Împart treburile	.095	.606	.270	.068

Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia

Analiza factorială ACS

Patru factori selectați pentru 10 itemi
- conform eficienței explicative

Explică în total **58%** din varianța datelor

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.008	20.081	20.081	2.008	20.081	20.081
2	1.629	16.294	36.375	1.629	16.294	36.375
3	1.160	11.600	47.975	1.160	11.600	47.975
4	1.084	10.839	58.814	1.084	10.839	58.814
5	.990	9.903	68.718			
6	.917	9.170	77.888			
7	.788	7.883	85.770			
8	.604	6.037	91.808			
9	.497	4.969	96.776			
10	.322	3.224	100.000			

Extraction Method: Principal Component Analysis.

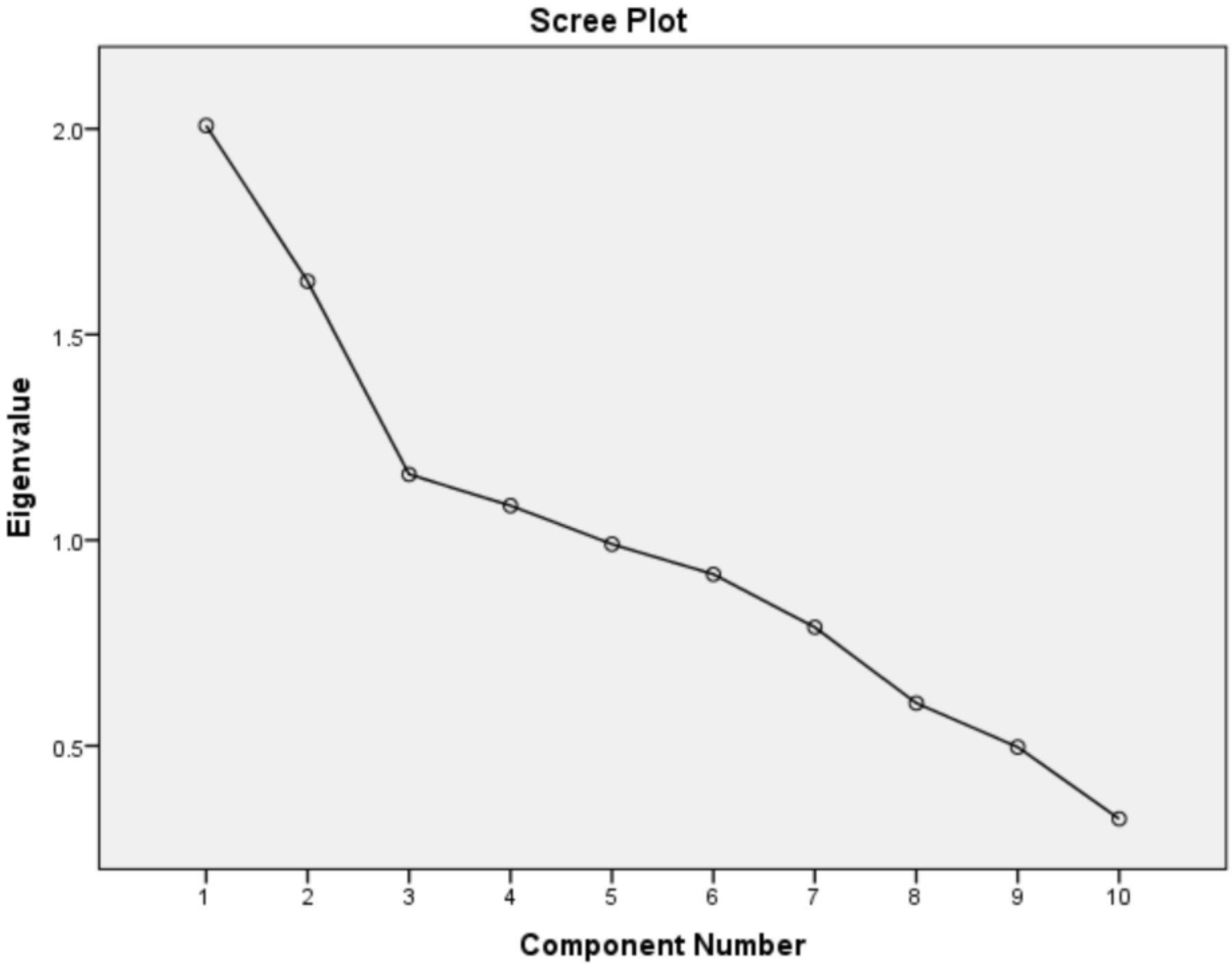
Analiza participanților la cursul de PR ce au completat chestionarul

Analiza factorială ACS

Testul grohotișului

Trei factori sunt cei mai relevanți

Explică în total **48%** din varianța datelor



Analiza factorială ACS

Interpretarea factorilor

Ce corelează pozitiv?

Ce corelează negativ?

Intimitate fără copii

Familia tradițională

Prietenie & banii nu contează

Confort & sexualitate

	Factor			
	1	2	3	4
Fidelitate	.188	.521	-.012	-.598
Copii	-.456	.558	-.253	.216
Separat de socri	.540	-.240	-.038	-.607
Venit adecvat	.768	.016	-.319	.307
Locuință bună	.811	.107	-.199	.190
Același mediu	.250	.423	-.062	.148
Aceeași religie	.107	.771	.072	.107
Acord politic	.255	.329	.706	-.046
Împart treburi	.143	-.057	.637	.203
Relație sexuală	.256	-.304	.200	.323

Analiza participanților la cursul de PR ce au completat chestionarul

Regresie pentru factorii ACS

Cum explică genul, vârsta, statutul marital și religiozitatea valorile studenților?

Predictori socio-demografici:	**Intimitate fără copii**	**Familia tradițională**	**Prietenie & banii nu contează**
	Beta	**Beta**	**Beta**
Gen feminin	.319	.203	.100
Varsta	-.411	.016	-.075
Vreodata casatorit	.239	.356	-.283
Religiozitate	-.106	.088	-.077
R Square (% varianță a efectului explicată de predictori)	.192	.190	.126

Analiza participanților la cursul de PR ce au completat chestionarul

Analiza cluster EVS

Cuplu cu copii, familie extinsă

Cuplu cu copii, familie nucleară

Cuplu fără copii, familie nucleară

Totul contează

	Cluster			
	1	2	3	4
Fidelitate	1.81	1.86	1.58	1.91
Copii	1.42	1.87	.64	1.77
Relație sexuală	1.25	1.79	1.58	1.69
Locuință bună	.92	1.25	1.00	1.64
Venit	.90	1.15	1.05	1.60
Împart treburile	.85	1.46	.86	1.40
Aceeași religie	.63	.31	.30	1.49
Același mediu	.41	.31	.55	1.41
Locuiesc singuri	.32	1.71	1.65	1.41
Acord în politică	.30	.25	.33	.96

Analiză European Values Study, 1981-2010
România, Italia, Franța, Germania, Suedia

Analiza cluster ACS

Interpretarea clusterelor în populația de studenți ACS

Totul contează

Intimitate seculară anti-materialistă

Intimitate seculară fără copii

	Cluster		
	1 24 cazuri	**2** 29 cazuri	**3** 32 cazuri
Fidelitate	2.0	1.8	1.9
Relație sexuală	1.6	1.7	1.8
Împart treburi domestice	1.6	1.6	1.8
Acord politic	1.2	**.4**	**.7**
Aceeași religie	1.6	**.4**	**.5**
Locuință bună	1.3	**.4**	1.6
Venit adecvat	1.3	1.0	1.6
Separat de socri	1.2	1.2	1.9
Același mediu	1.1	.9	1.0
Copii	1.1	1.1	**.2**

Analiza participanților la cursul de PR ce au completat chestionarul

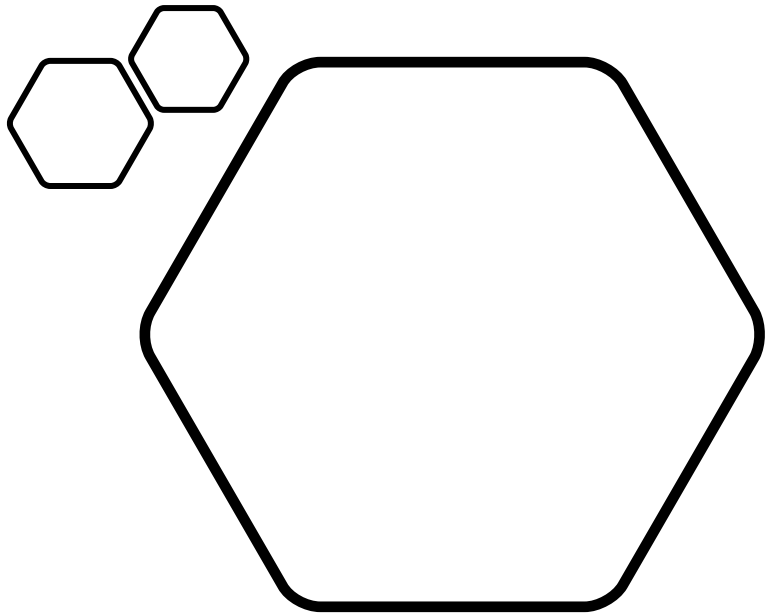
Analiza cluster ACS

Profilul socio-demografic al clusterelor

	Cluster „Totul contează”	Cluster „Intimitate seculară anti-materialistă”	Cluster „Intimitate seculară fără copii”
	Medie	Medie	Medie
Gen feminin	.50	.48	.63
Vârsta	27.75	28.31	22.69
Vreodată căsătorit	.17	.17	.13
Religiozitate	3.08	2.66	1.94

Analiza participanților la cursul de PR ce au completat chestionarul

Studiu de caz

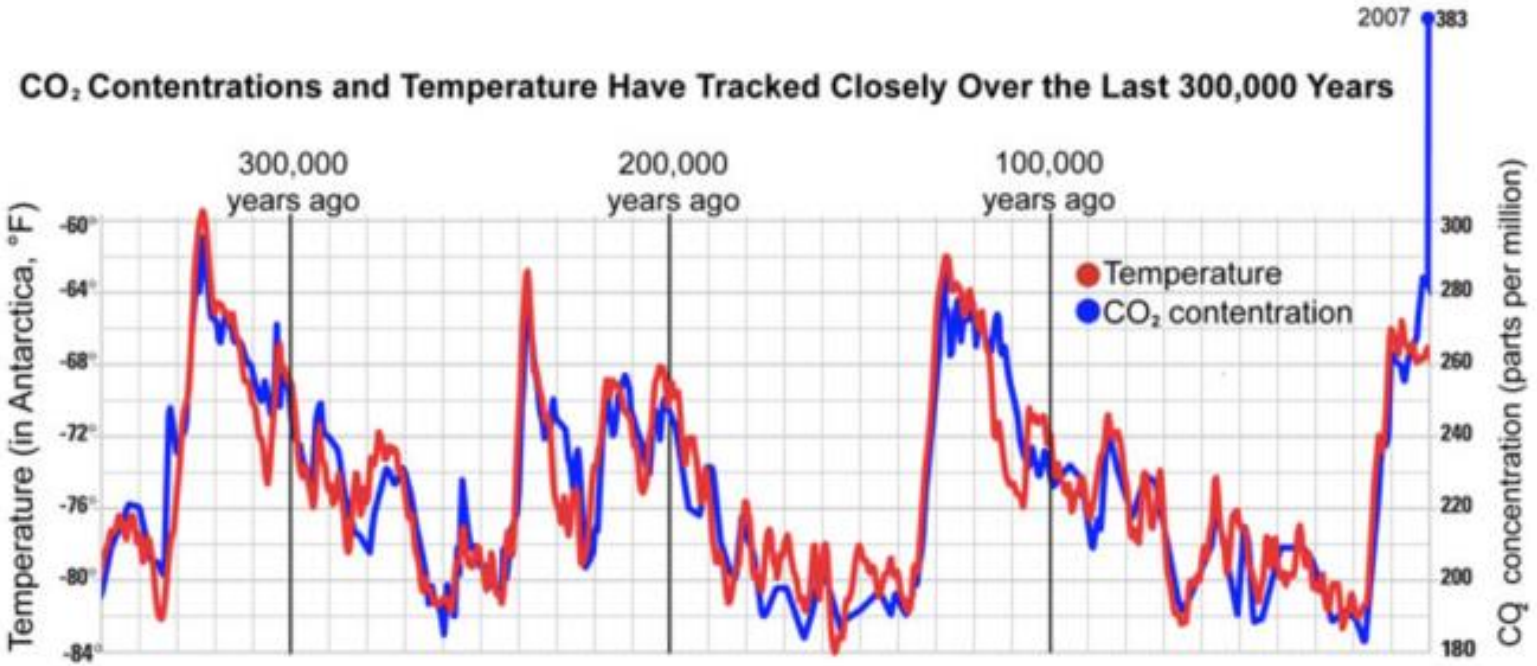


Stabilirea cauzalității:

CO_2 vs. încălzirea globală

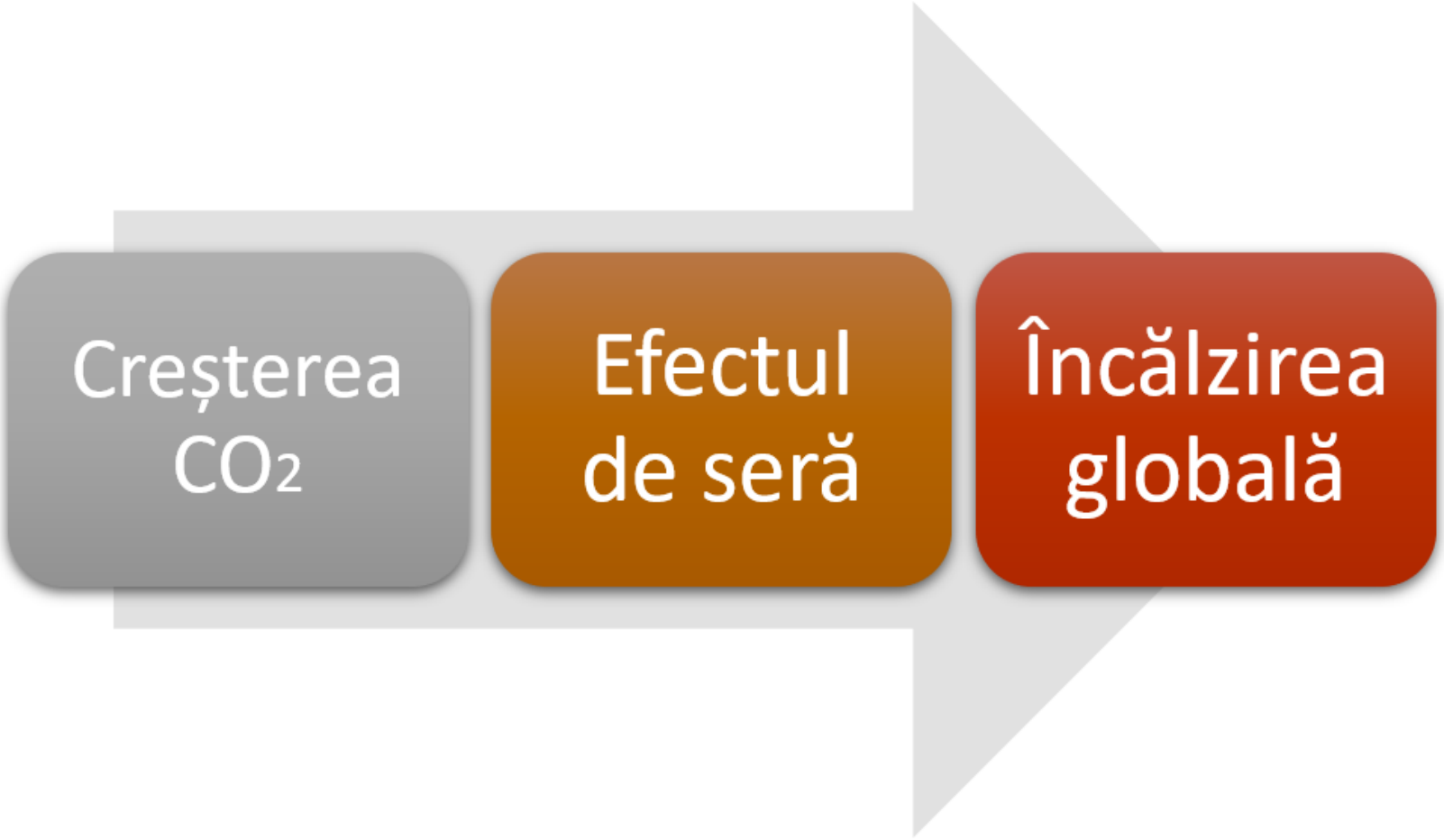
Creșterea CO2

Încălzirea planetei



Sursa

Creșterea CO_2 creează încălzire globală...



... Și încălzirea planetei creează mai mult CO_2

Argument

- Încălzirea planetei este de fapt sursa creșterii CO_2, nu invers
- „O serie de evoluții științifice recente care pun la îndoială relația cauzală modernă, $CO_2 \rightarrow T$ ”
  - Un prim studiu publicat în 1990 în *Nature* și care a inițiat actuala schimbare de paradigmă
  - Două studii recente, 2020 și 2023

HotNews.ro
STIRI NON-STOP

HOME ECONOMIE SPORT LIFE SCIENCE REVISTA PRESEI DOSARE JURIDICE

Război în Ucraina International Politic Educație Sănătate Minte&Corp Opinii Copii&Părinți Administrație

OPINII Joi, 12 Octombrie 2023, 08:03

Opinie Despre ouă și găini, temperaturi și CO2 și legături cauzale în procesele climatice contemporane

Constantin Crânganu • Contributors.ro

share

Cu două mii de ani în urmă, Plutarh a formulat o întrebare care definește dilema cauzalității: Ce a fost mai întâi, oul sau găina? Secole la rând, răspunsurile au variat în funcție de cei care le-au dat: filosofi, preoți, biologi, paleontologi ș.a. În cazul unor evenimente discrete, formularea unui răspuns trebuie să respecte precedența temporală a cauzei față efect. Aceeași cerință se aplică (cu modificări specifice) și în cazul comparării unor procese, deterministe ori stocastice.

Cum evaluăm argumentul?

- Care este **stadiul actual al cunoașterii** privind relația CO_2 – încălzire globală?
- Știința este un **ecosistem**
  - Apar noi metodologii, modele
  - Testate prin multe studii
  - Care au credibilitate diferențiată
- Stadiul actual e sintetizat ținând cont de **toate studiile** și **credibilitatea lor**

Articolele invocate

- [2] Kuo, C., Lindberg, C., and Thomson, D., 1990, *Coherence established between atmospheric carbon dioxide and global temperature*, *Nature*, vol. 343, pp. 709–714, https://doi.org/10.1038/343709a0
  - Jurnal cu credibilitate mare, publicație veche
- [3] Koutsoyiannis, D., and Kundzewicz, Z.W., 2020, *Atmospheric temperature and CO_2: Hen-or-egg causality?* Sci, 2 (4), 83, https://doi.org/10.3390/sci2040083
  - Jurnal cu credibilitate mică, publicație nouă
- [5] Koutsoyiannis, D., Onof, O., Zbigniew W. Kundzewicz, Z. W., and Christofides, A., 2023, *On Hens, Eggs, Temperatures and CO_2: Causal Links in Earth's Atmosphere*, Sci. 5 (3), 10.3390/sci5030035.
  - Jurnal cu credibilitate mică, publicație nouă

CITESCORE 3.1

Sci

Submit to *Sci*

Review for *Sci*

Share

Journal Menu

- *Sci* Home
- Aims & Scope
- Editorial Board
- Topical Advisory Panel
- Instructions for Authors
- Special Issues

Sci

Sci is an international, peer-reviewed, open access journal on all research fields published quarterly online by MDPI.

- **Open Access** — free for readers, with **article processing charges (APC)** paid by authors or their institutions.
- **High Visibility:** indexed within **Scopus**, and **other databases**.
- **Journal Rank:** CiteScore - Q2 (*Multidisciplinary*)
- **Rapid Publication:** manuscripts are peer-reviewed and a first decision is provided to authors approximately 38.1 days after submission; acceptance to publication is undertaken in 6.8 days (median values for papers published in this journal in the first half of 2023).
- **Recognition of Reviewers:** reviewers who provide timely, thorough peer-review reports receive vouchers entitling them to a discount on the APC of their next publication in any MDPI journal, in appreciation of the work done.

Imprint Information | Journal Flyer | Open Access | **ISSN: 2413-4155**

E-Mail Alert

Add your e-mail address to receive forthcoming issues of this journal:

Enter Your E-Mail Address...

Subscribe

News

19 October 2023

Open Access Week 2023 – the Global Drive to Open Continues

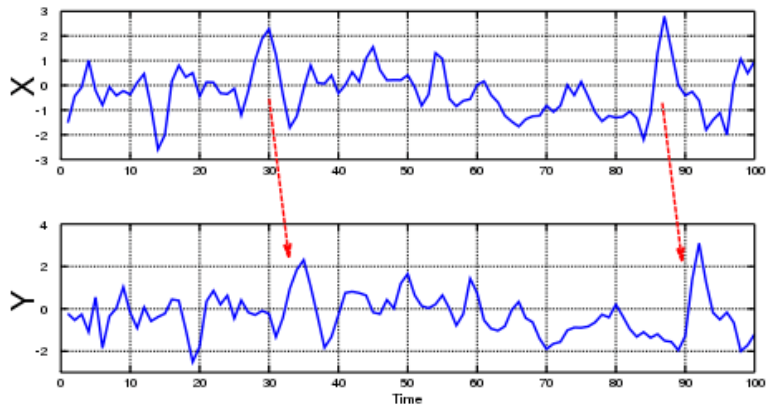
Granger causality

13 languages

Article Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

The **Granger causality test** is a statistical hypothesis test for determining whether one time series is useful in forecasting another, first proposed in 1969.[1] Ordinarily, regressions reflect "mere" correlations, but Clive Granger argued that causality in economics could be tested for by measuring the ability to predict the future values of a time series using prior values of another time series. Since the question of "true causality" is deeply philosophical, and because of the post hoc ergo propter hoc fallacy of assuming that one thing preceding another can be used as a proof of causation, econometricians assert that the Granger test finds only "predictive causality".[2] Using the term "causality" alone is a misnomer, as Granger-causality is better described as "precedence",[3] or, as Granger himself later claimed in 1977, "temporally related".[4] Rather than testing whether *X causes* Y, the Granger causality tests whether X *forecasts Y*.[5]



When time series *X* Granger-causes time series *Y*, the patterns in *X* are approximately repeated in *Y* after some time lag (two examples are indicated with arrows). Thus, past values of *X* can be used for the prediction of future values of *Y*.

Source

Primul articol invocat de Crânganu, publicat în Nature, 1990:

Coherence established between atmospheric carbon dioxide and global temperature

Cynthia Kuo, Craig Lindberg & David J. Thomson

Mathematical Sciences Research Center, AT&T Bell Labs, Murray Hill, New Jersey 07974, USA

The hypothesis that the increase in atmospheric carbon dioxide is related to observable changes in the climate is tested using modern methods of time-series analysis. The results confirm that average global temperature is increasing, and that temperature and atmospheric carbon dioxide are significantly correlated over the past thirty years. Changes in carbon dioxide content lag those in temperature by five months.

From atmospheric chemistry, global temperature depends nonlinearly on CO_2 concentration (T. Graedel, personal communication). The procedure used here implicitly uses a linear dependence. Bispectral estimates provide evidence of quadratic terms, although the shortness of these series makes this difficult to quantify. Also, except for the solar modulation of the temperature series near 1 cycle yr^{-1}, we have ignored the cyclostationary properties of these two series (that is, the statistics of these series vary periodically).

Un articol omis de Crânganu, publicat în Nature – Scientific Reports, 2016

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

On the causal structure between CO_2 and global temperature

Adolf Stips[1], Diego Macias[1], Clare Coughlan[1], Elisa Garcia-Gorriz[1] & X. San Liang[2]

Received: 29 June 2015
Accepted: 27 January 2016
Published: 22 February 2016

We use a newly developed technique that is based on the information flow concept to investigate the causal structure between the global radiative forcing and the annual global mean surface temperature anomalies (GMTA) since 1850. Our study unambiguously shows one-way causality between the total Greenhouse Gases and GMTA. Specifically, it is confirmed that the former, especially CO_2, are the main causal drivers of the recent warming. A significant but smaller information flow comes from aerosol direct and indirect forcing, and on short time periods, volcanic forcings. In contrast the causality contribution from natural forcings (solar irradiance and volcanic forcing) to the long term trend is not significant. The spatial explicit analysis reveals that the anthropogenic forcing fingerprint is significantly regionally varying in both hemispheres. On paleoclimate time scales, however, the cause-effect direction is reversed: temperature changes cause subsequent CO_2/CH_4 changes.

Source

Alt articol omis de Crânganu, publicat în Nature – Scientific Reports, 2022

- Sinteză a mai multor studii recente
- Majoritatea studiilor susțin pentru perioada modernă relația dominantă:

  CO2 → temperatură

- Sinteza critică studiul Koutsoyiannis (2020)

nature › scientific reports › articles › article

Article | Open access | Published: 19 August 2022

Compression complexity with ordinal patterns for robust causal inference in irregularly sampled time series

Aditi Kathpalia, Pouya Manshour & Milan Paluš

Scientific Reports **12**, Article number: 14170 (2022) | Cite this article

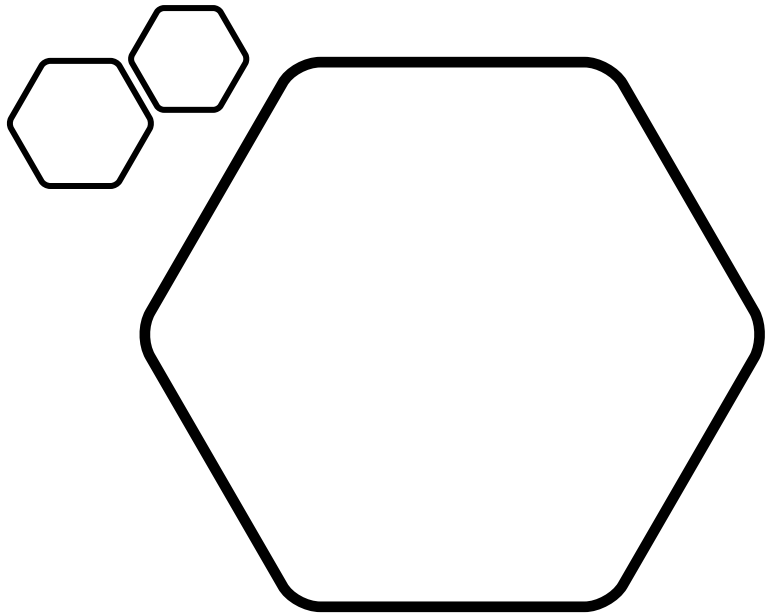
1030 Accesses | **2** Citations | Metrics

For example, the relationship between CO_2 concentrations and temperature of the atmosphere has been studied from the mid 1800s[83,84], beginning when a strong link between the two was recognized. Relatively recently, with causal inference tools available, a number of studies have begun to look at the directionality of relationship between the two on different temporal scales. To mention a few findings, Kodra et al.[85] found that CO_2 Granger causes temperature. Their analysis was based on data taken from 1860 to 2008. Atanassio[86] found a clear evidence of GC from CO_2 to temperature using lag-augmented Wald test, for a similar time range. On the other hand, Stern and Kaufmann[87] found bidirectional GC between the two, again for a similar time range. Kang and Larsson[88] also find bidirectional causation between the two using GC, however, by using data from ice cores for the last 800,000 years. Many of these latter studies criticize the former. Also, the drawbacks of one or more of these studies are explicitly mentioned in Refs.[87,89,90] and highlight the issues with the data and/ or the methodology employed. Other than GC and its extensions, a couple of other measures have also been used to study CO_2-T relationship. Stips et al.[91] have applied a measure called Liang's Information flow on CO_2-T recordings, both on recent (1850–2005) and paleoclimate (800 ka ice core reconstructions) time-scales. The study finds unidirectional causation from $CO_2 \rightarrow T$ on the recent time-scale and from $T \rightarrow CO_2$ on the paleoclimatic scale.

They have also analysed the CH_4-T relationship and found T to drive CH_4 on the paleoclimate scale. This study has been criticized by Goulet et al.[92]. They show that an assumption of 'linearity' made by Liang's information flow is nearly always rejected by the data. Convergent cross mapping, which is applied to the 800 ka recordings in another study, finds a bidirectional causal influence between both CO_2 - T and CH_4-T[93]. Another recent study, that infers causation using lagged cross-correlations between monthly CO_2 and temperature, taken from the period 1980–2019, has found a bidirectional relationship on the recent monthly scale, with the dominant influence being from $T \rightarrow CO_2$[94]. In the light of the limitations of CCM[95,96], especially for irregularly sampled or missing data[42], and of the widely known pitfalls of correlation coefficient[97], it is difficult to rely on the inferences of the latter two studies.

Source

Studiu de caz

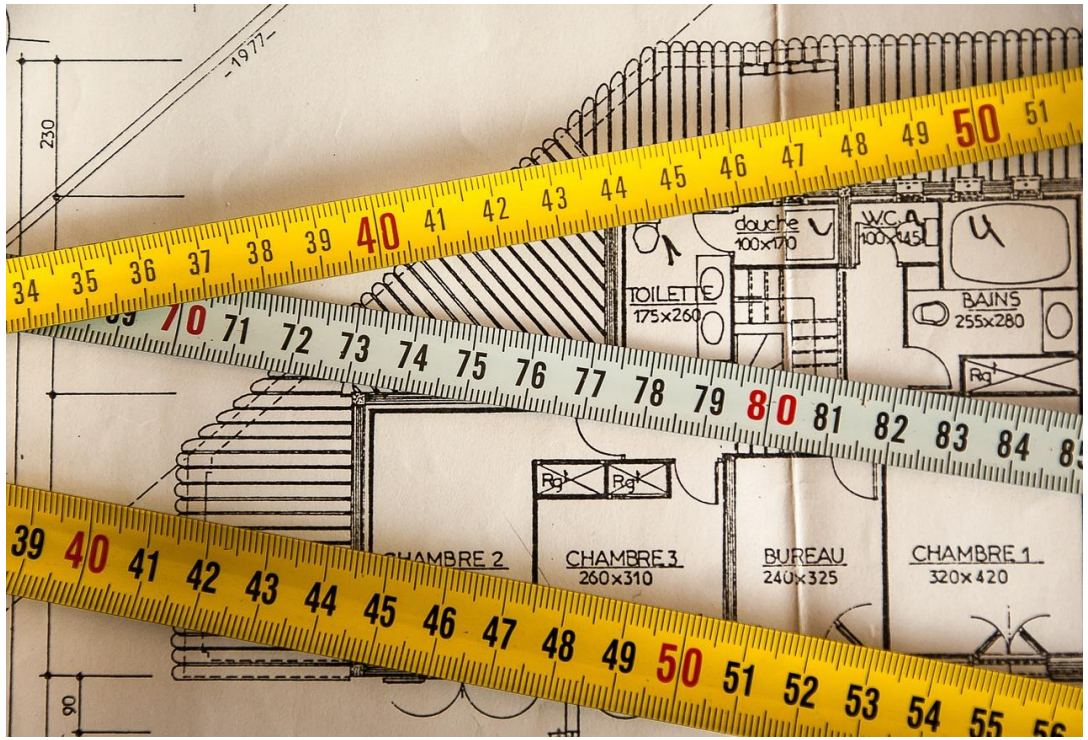


Evaluarea politicilor de lockdown

Trei măsurători ale impactului COVID19

Măsurători ale mortalității

- **Fatalitatea**: decese cauzate de COVID19 / persoane îmbolnăvite
- **Mortalitatea**: decese cauzate de COVID19 / populație totală
- **Mortalitatea excesivă**: diferența dintre mortalitatea reală și cea prezisă (media ultimilor 3 sau 5 ani) – pe o perioadă dată



Erori și limitări ale măsurătorilor

Măsurători ale mortalității

- **Fatalitatea**: decese cauzate de COVID19 / persoane îmbolnăvite
  - Erori [1] [2] + aleatorii
- **Mortalitatea**: decese cauzate de COVID19 / populație totală
  - Erori [1] [3] + aleatorii
- **Mortalitatea excesivă**: diferența dintre mortalitatea reală și cea prezisă (media ultimilor 3 sau 5 ani) – pe o perioadă dată
  - Obs. [4]

Erori sistematice în compararea țărilor

[1] Atribuire: a muri **cu / de** COVID19

[2] Câți oameni s-au îmbolnăvit de fapt de COVID19?

- Rata testării e variabilă

[3] Populațiile diferă prin factorii de risc ai mortalității

- Vârstă, obezitate, diabet

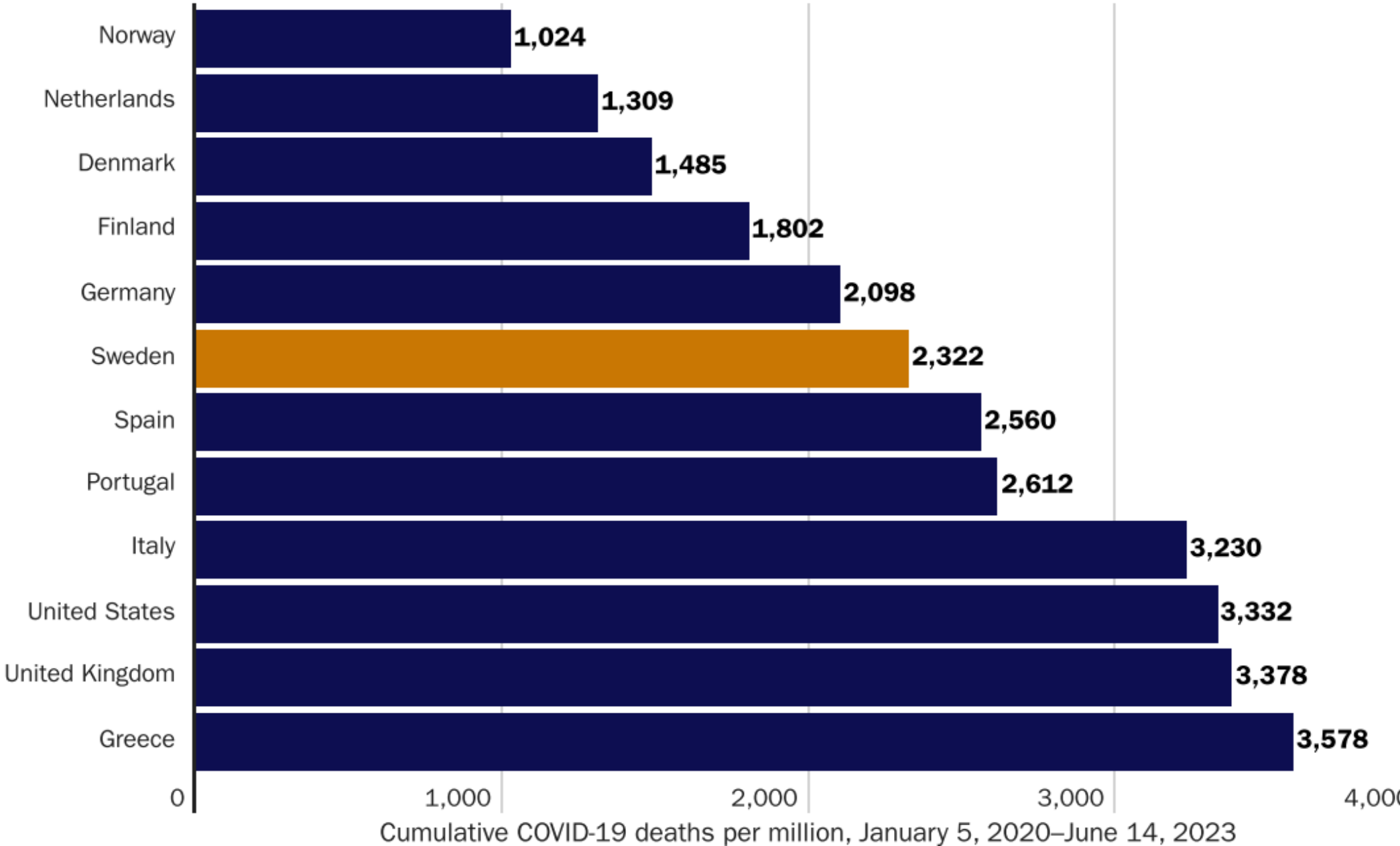
[4] Mortalitatea excesivă captează toate schimbările perioadei – nu doar boala

Mortality

Cummulative

Figure 1

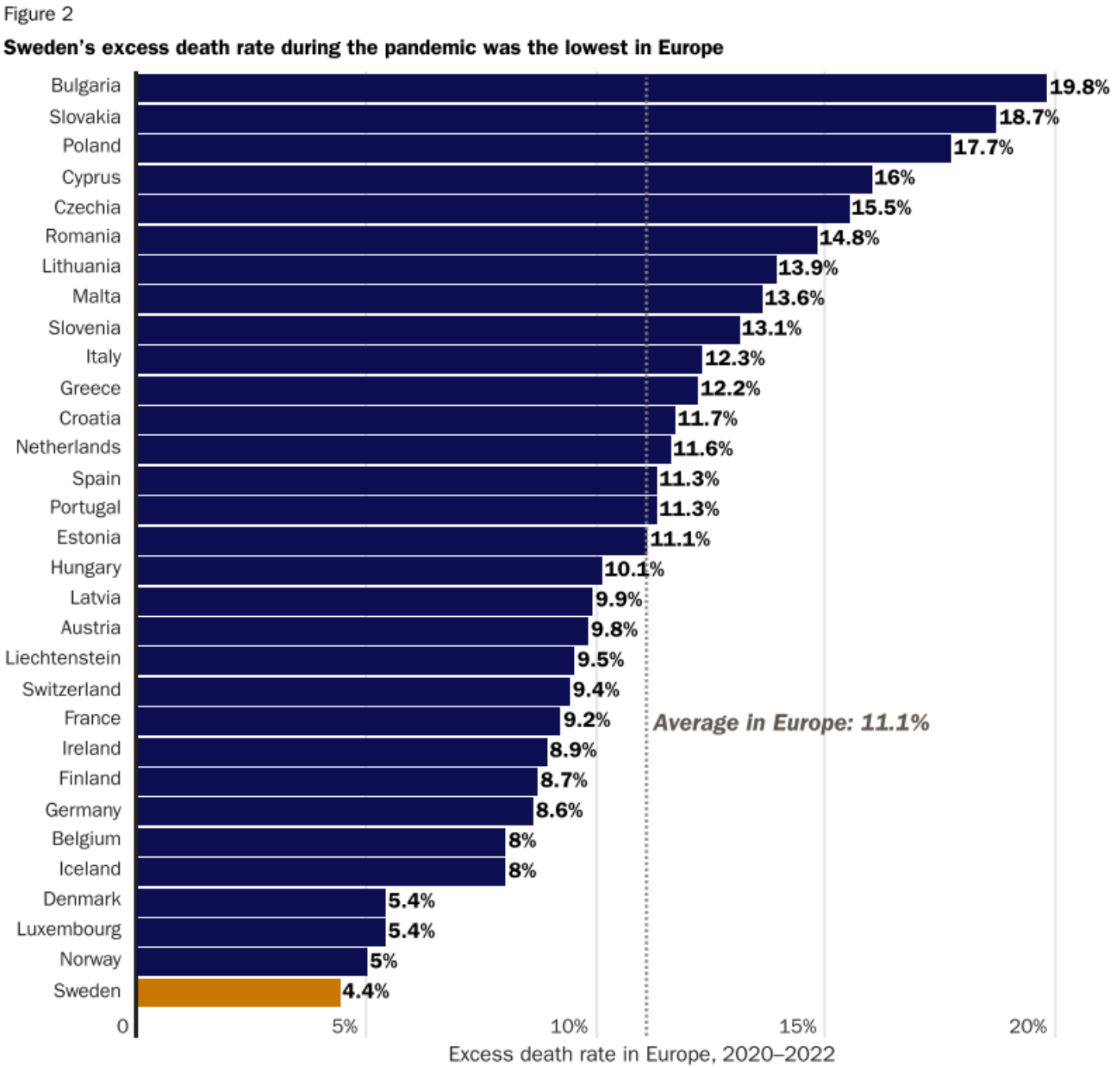
Sweden's COVID-19 death rate is not an outlier



Excess mortality

Cummulative

vs. previous
3 years



Source: Statistics Sweden, quoted in Therese Bergstedt, "Anders Tegnell: 'gillar inte ordet 'revansch,'" *Svenska Dagbladet (Stockholm)*, March 4, 2023.

Sursa

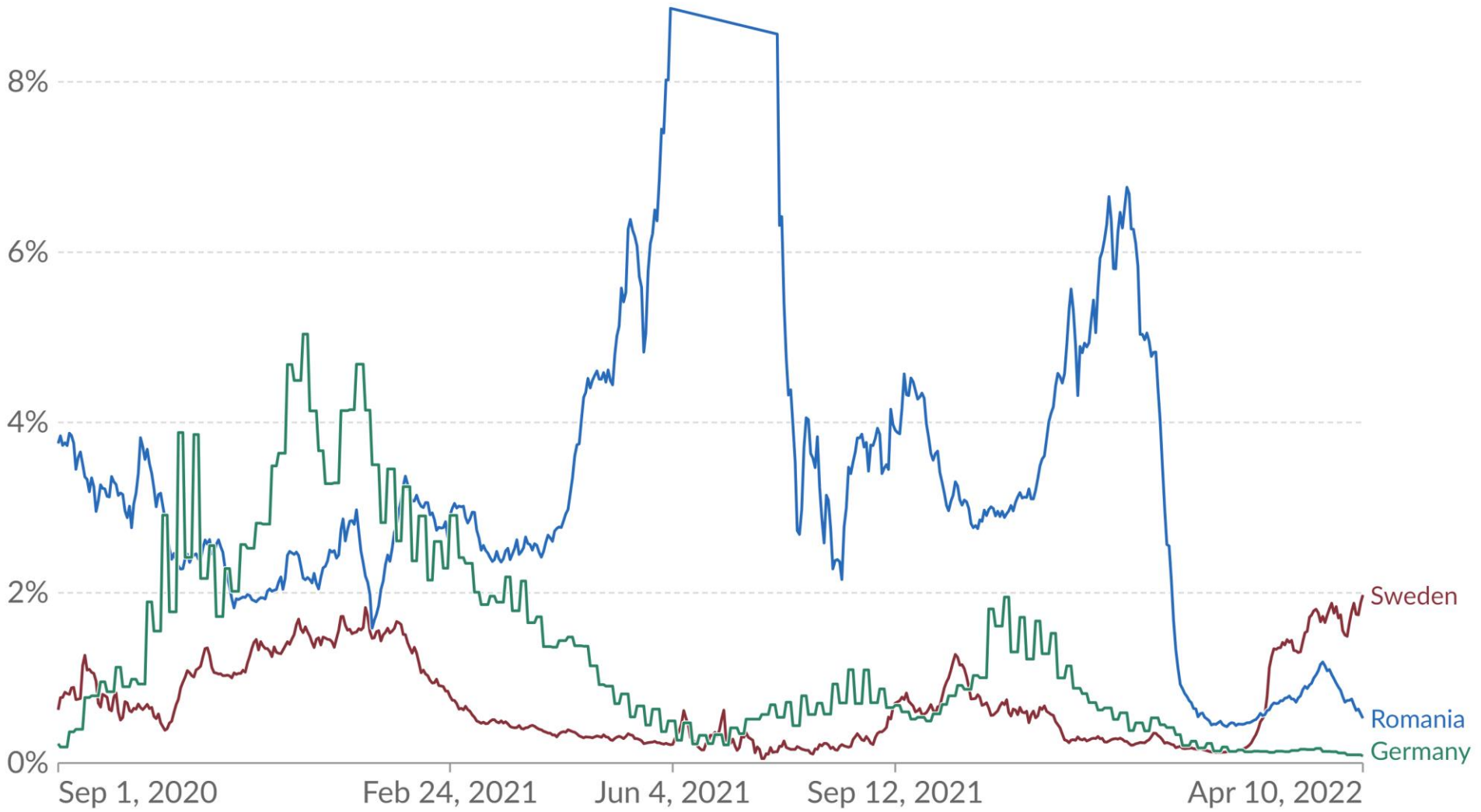
Fatality

Weekly

Moving-average case fatality rate of COVID-19

The case fatality rate (CFR) is the ratio between confirmed deaths and confirmed cases. Our rolling-average CFR is calculated as the ratio between the 7-day average number of deaths and the 7-day average number of cases 10 days earlier.

Our World in Data



Data source:

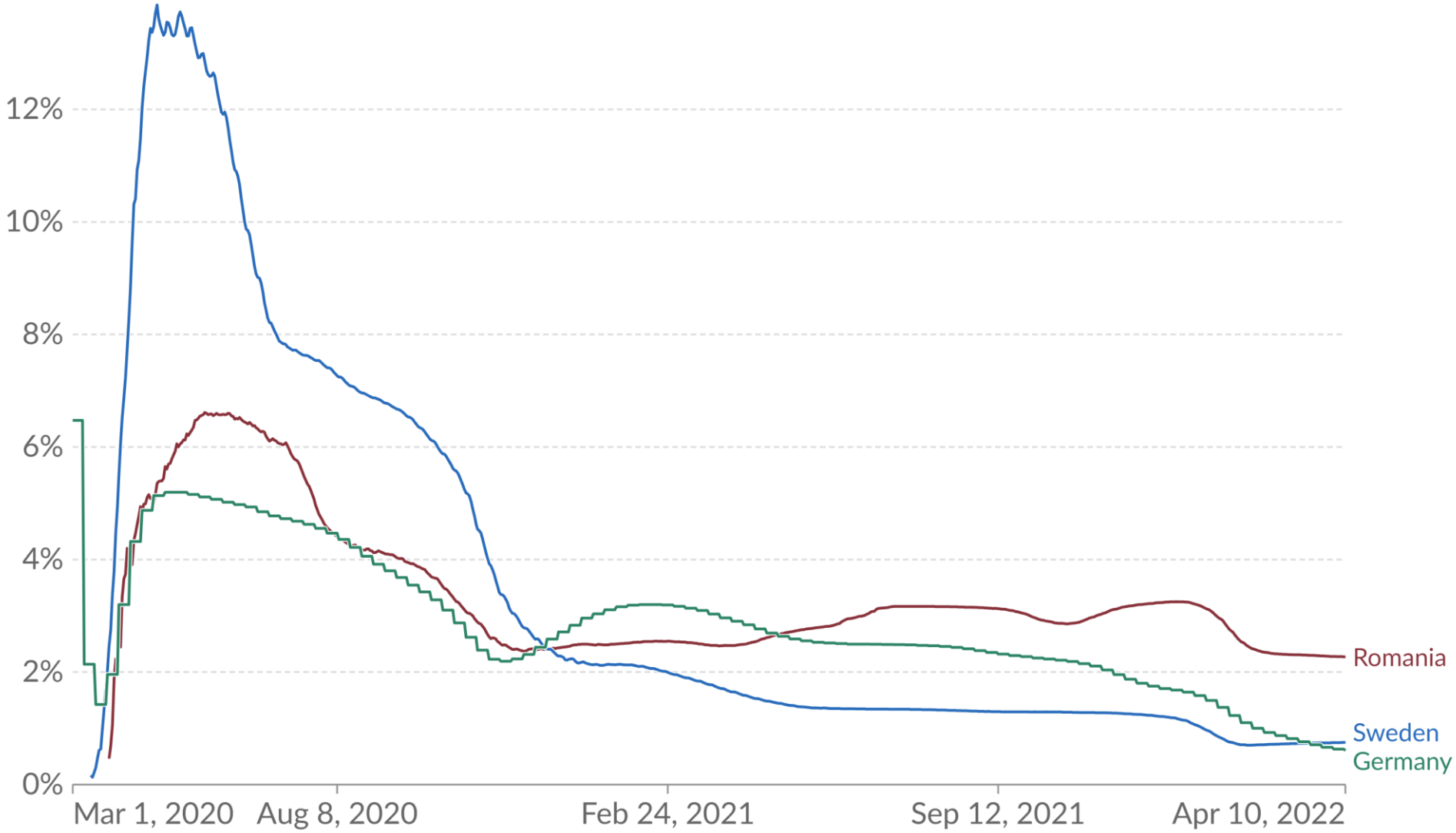
Fatality

Cumulative

Case fatality rate of COVID-19

The case fatality rate (CFR) is the ratio between confirmed deaths and confirmed cases. The CFR can be a poor measure of the mortality risk of the disease. We explain this in detail at https://OurWorldInData.org/mortality-risk-covid

Our World in Data



Data source:

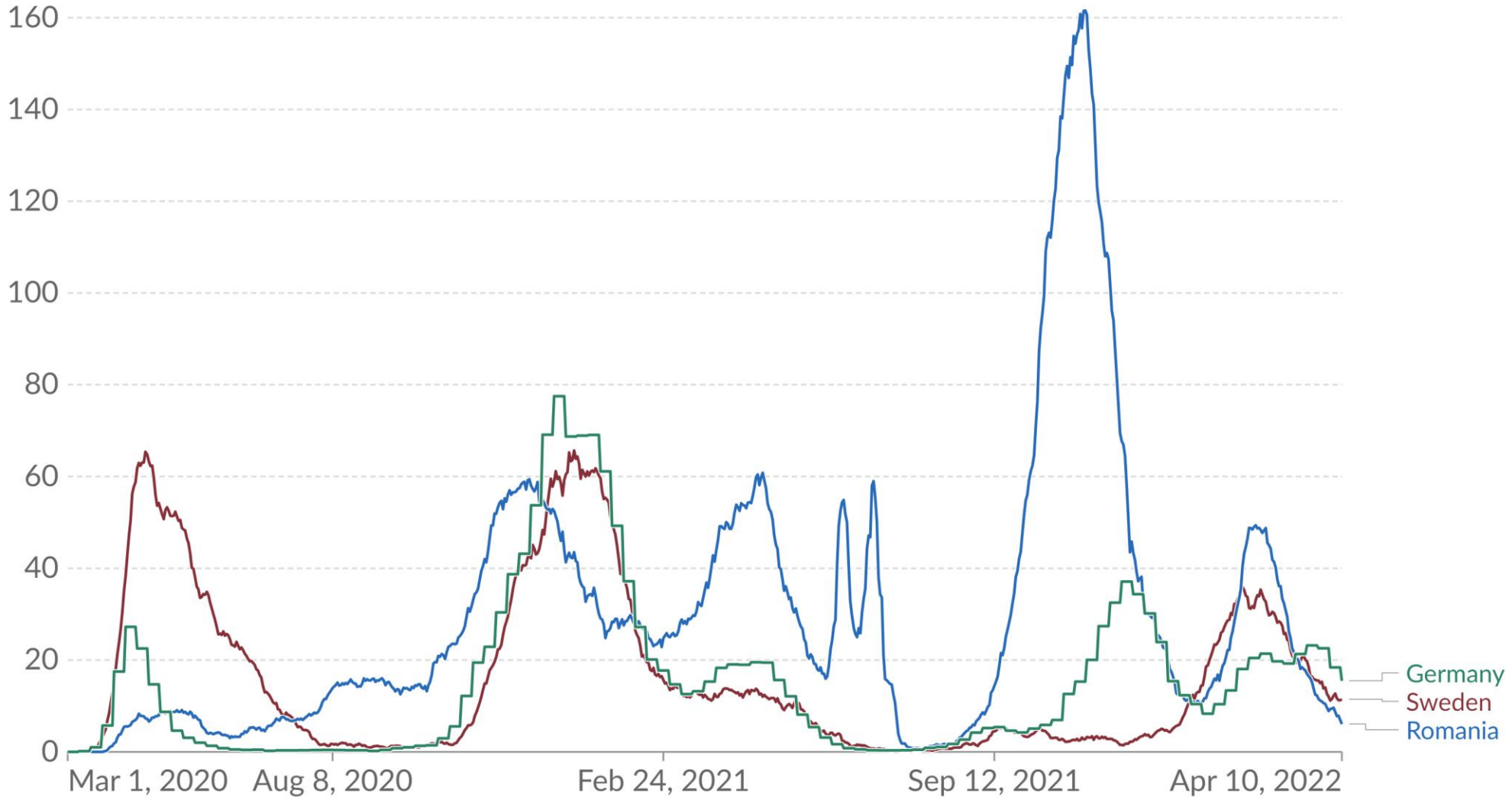
Mortality

Weekly

Weekly confirmed COVID-19 deaths per million people

Our World in Data

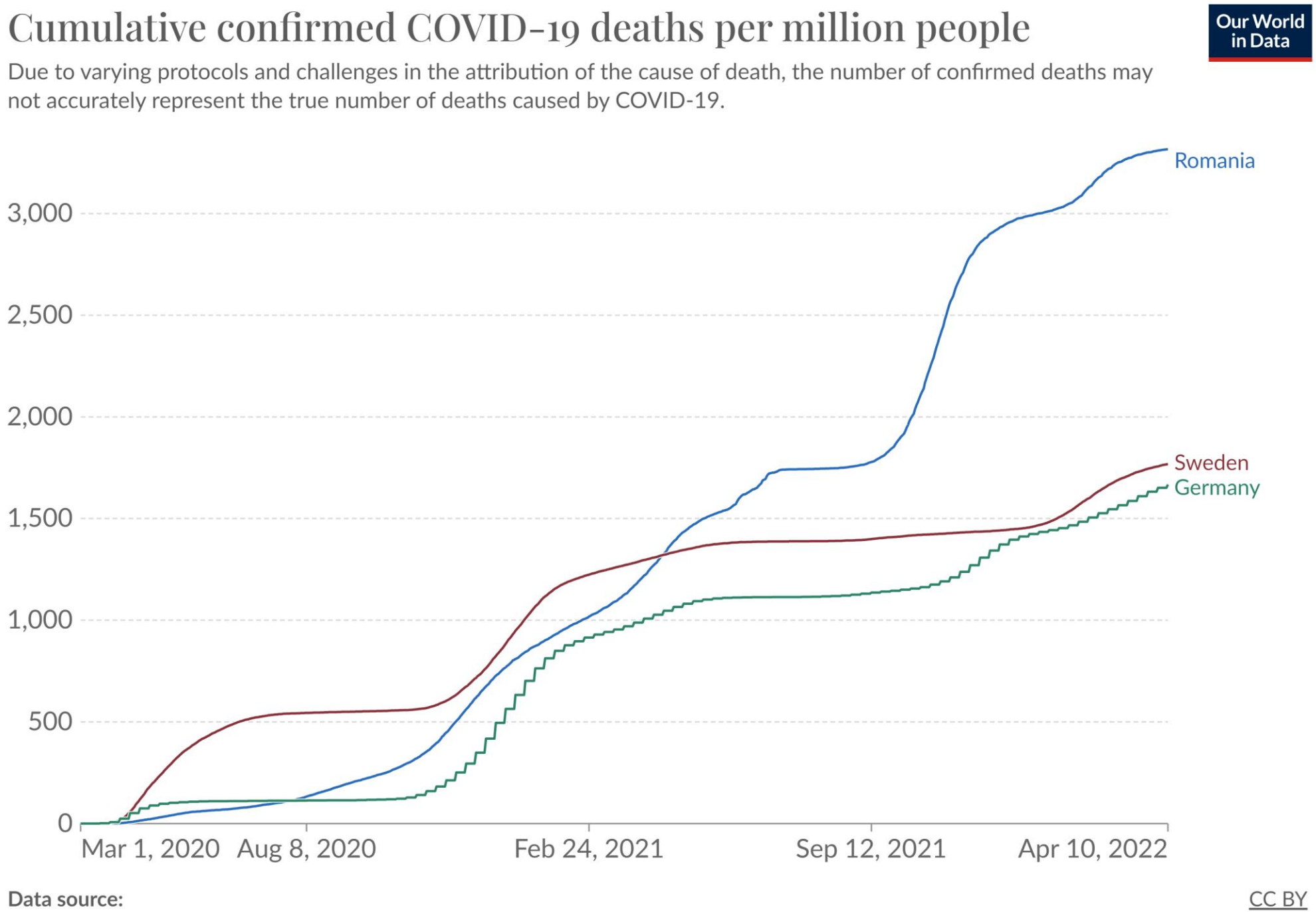
Weekly confirmed deaths refer to the cumulative number of confirmed deaths over the previous week. Due to varying protocols and challenges in the attribution of the cause of death, the number of confirmed deaths may not accurately represent the true number of deaths caused by COVID-19.



Data source:

Mortality

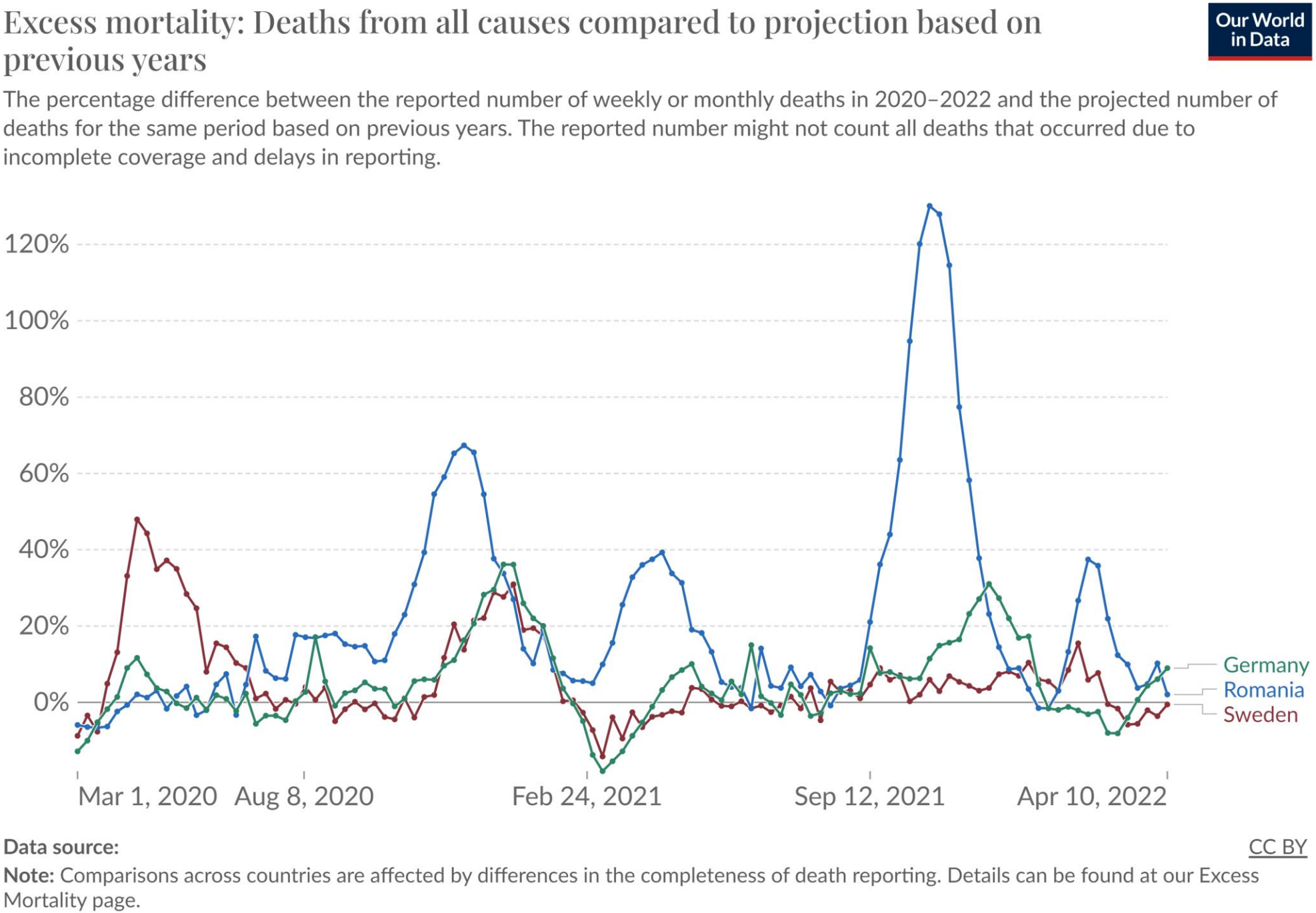
Cummulative



Excess mortality

Weekly

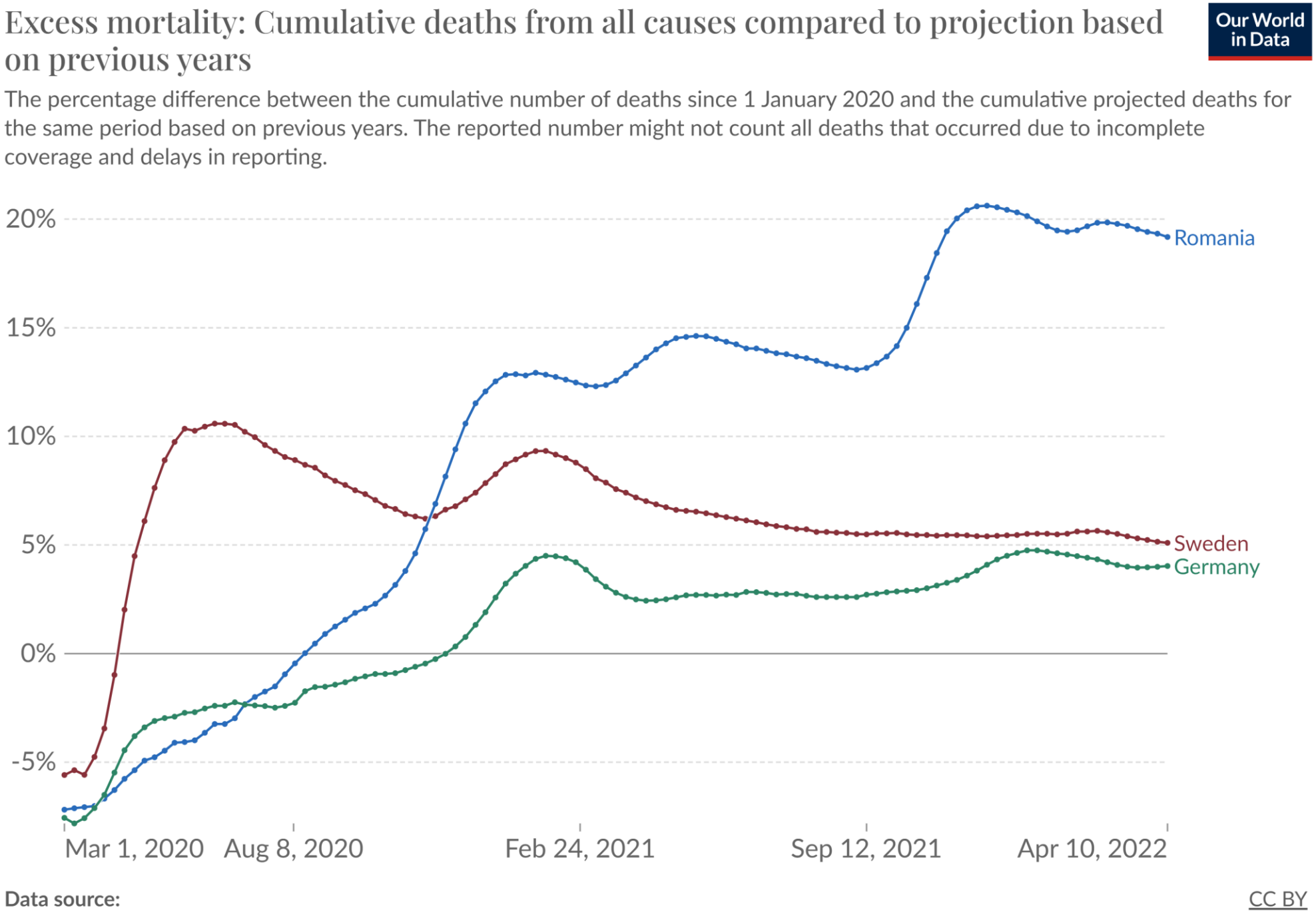
vs. previous 5 years



Excess mortality

Cummulative

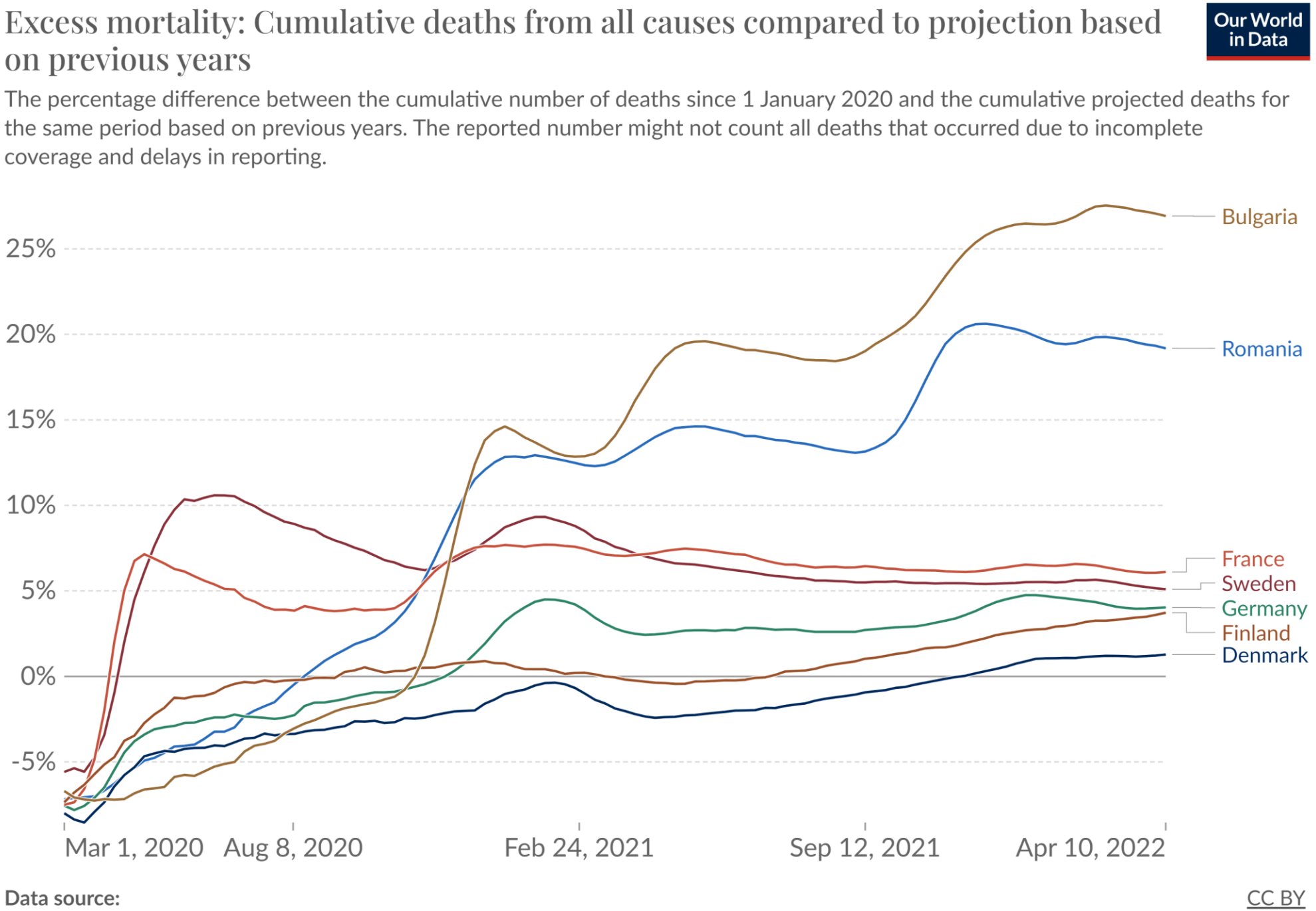
vs. previous 5 years



Excess mortality

Cummulative

vs. previous 5 years



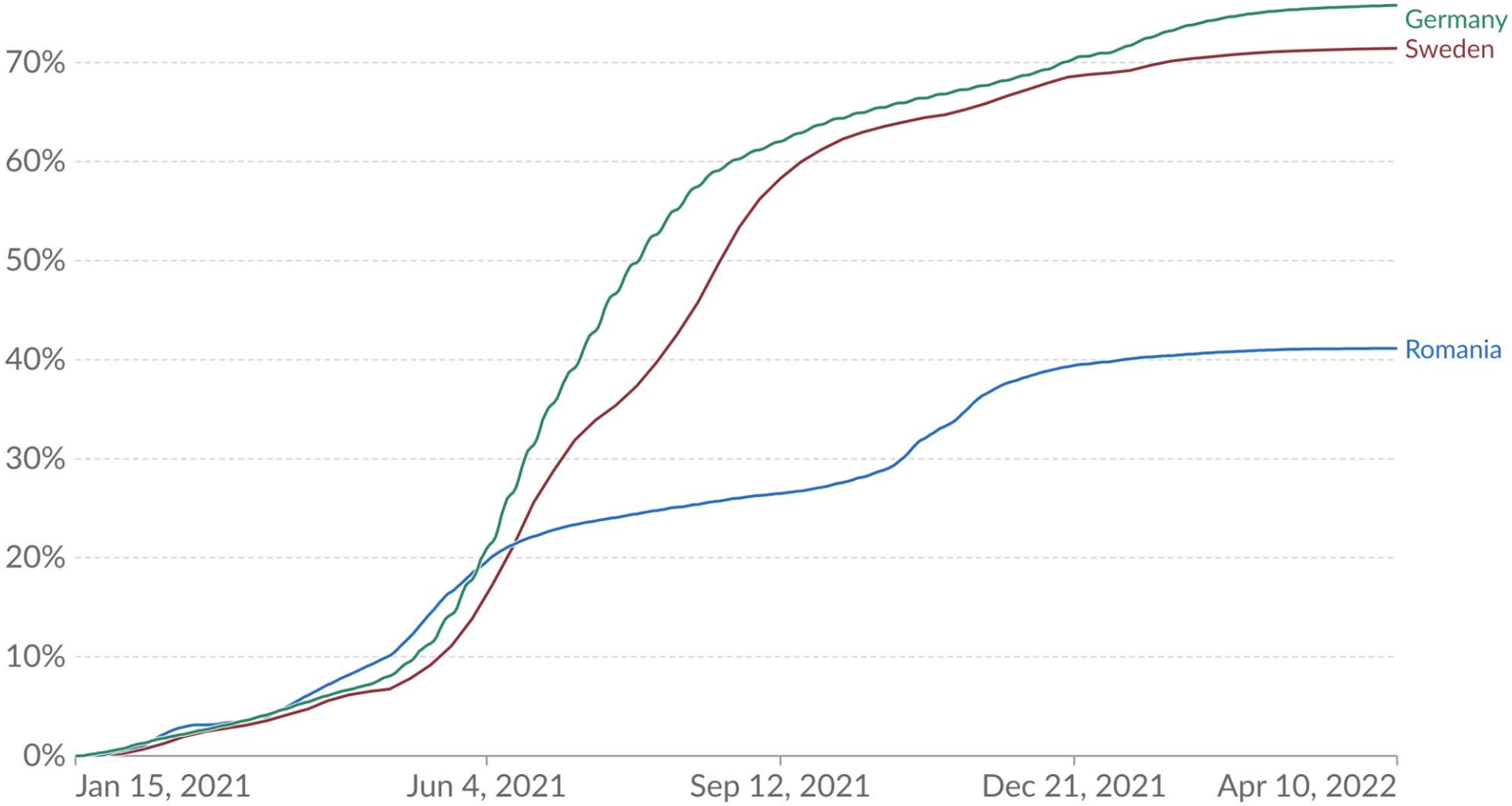
People fully vaccinated

Cummulative

Share of people who completed the initial COVID-19 vaccination protocol

Total number of people who received all doses prescribed by the initial vaccination protocol, divided by the total population of the country.

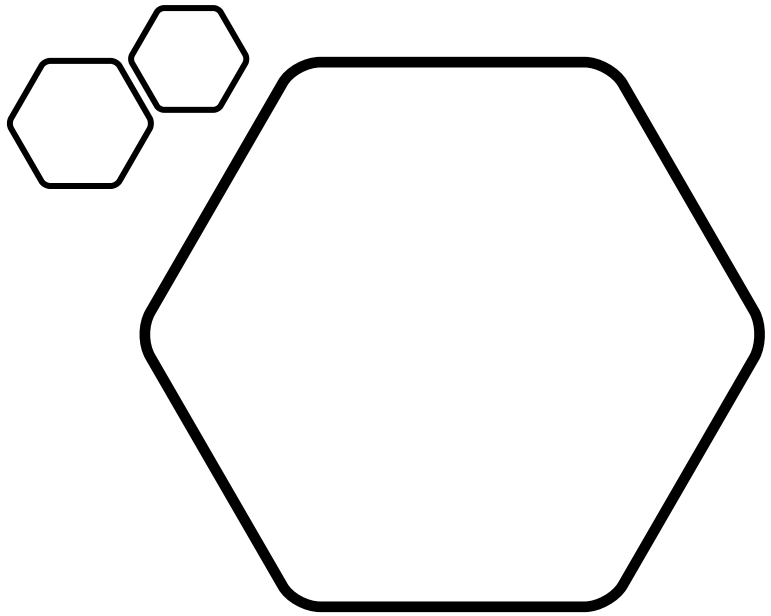
Our World in Data



Data source:

Note: Alternative definitions of a full vaccination, e.g. having been infected with SARS-CoV-2 and having 1 dose of a 2-dose protocol, are ignored to maximize comparability between countries.

Studiu de caz



Tipuri de părtiniri în Analiza Datelor

- Reporting bias – what becomes data
- Selection bias – how we sample data
- Confirmation bias – how we interpret data

- Automation bias

Reporting bias: what events become data?

- Reporting bias occurs when the frequency of events captured in a data set does not reflect their real-world frequency.
- People tend to focus on circumstances that are unusual or memorable, assuming that the ordinary can "go without saying."

Source: Google crash-course on ML

Selection bias: how do we sample the events?

- **Coverage** bias: The sample omits some subsets of events
- **Non-response** bias (or participation bias): The sample only includes *agreeable* individuals - who agree to respond
- **Sampling** bias: Proper randomization is not used during sample selection

Source: Google crash-course on ML

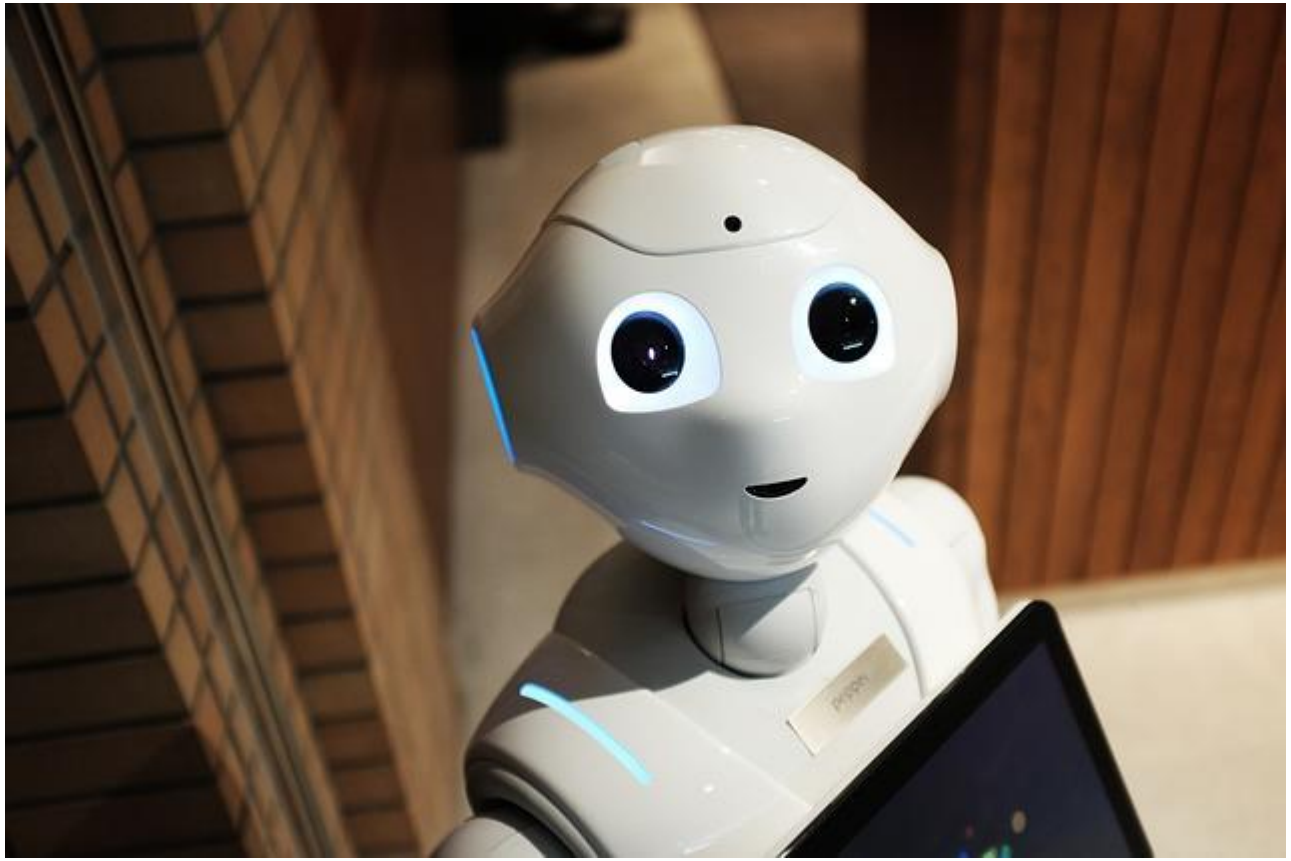
Confirmation bias: interpretation vs. expectation

- In **confirmation bias** model builders unconsciously process data in ways that affirm preexisting beliefs
- In some cases, a model builder may actually keep training a model until it produces a result that aligns with their original hypothesis; this is called **experimenter's bias**.

Source: Google crash-course on ML

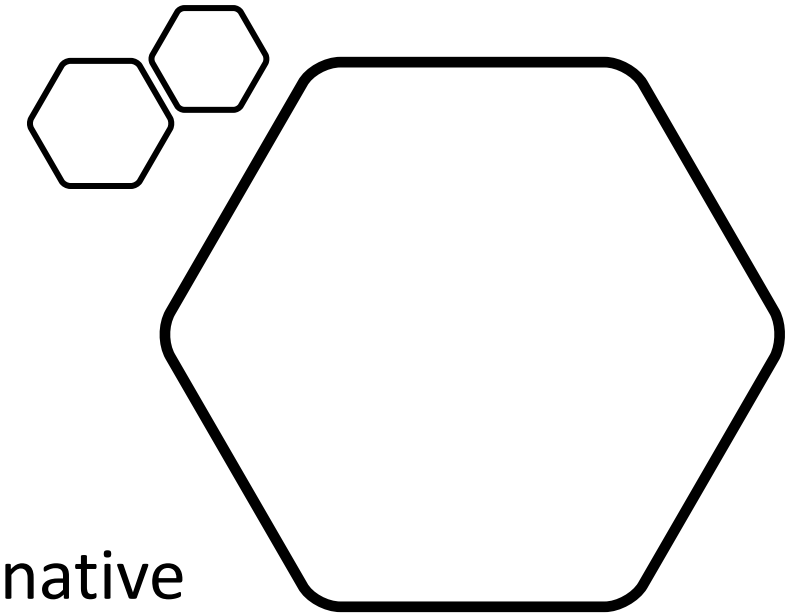
Automation bias: whose analysis do we trust?

- Automation bias is a tendency to favor results generated by automated systems over those generated by non-automated systems
- Perceived “objectivity” of technology



Source: Google crash-course on ML

Concluzii: provocări ale analizei datelor în RL



- Identificarea metodei potrivite de analiză
- Erori sistematice diferite ale măsurătorilor alternative
- Importanța înțelegerii stadiului cunoașterii
  - Cunoașterea este o activitate colaborativă de scală largă
- Patru tipuri de părtiniri în formularea, selecția și analiza datelor