

Journalismus & Data Literacy

Daten verstehen, einordnen, verarbeiten und präsentieren

Simon Huwiler, Nikolai Thelitz

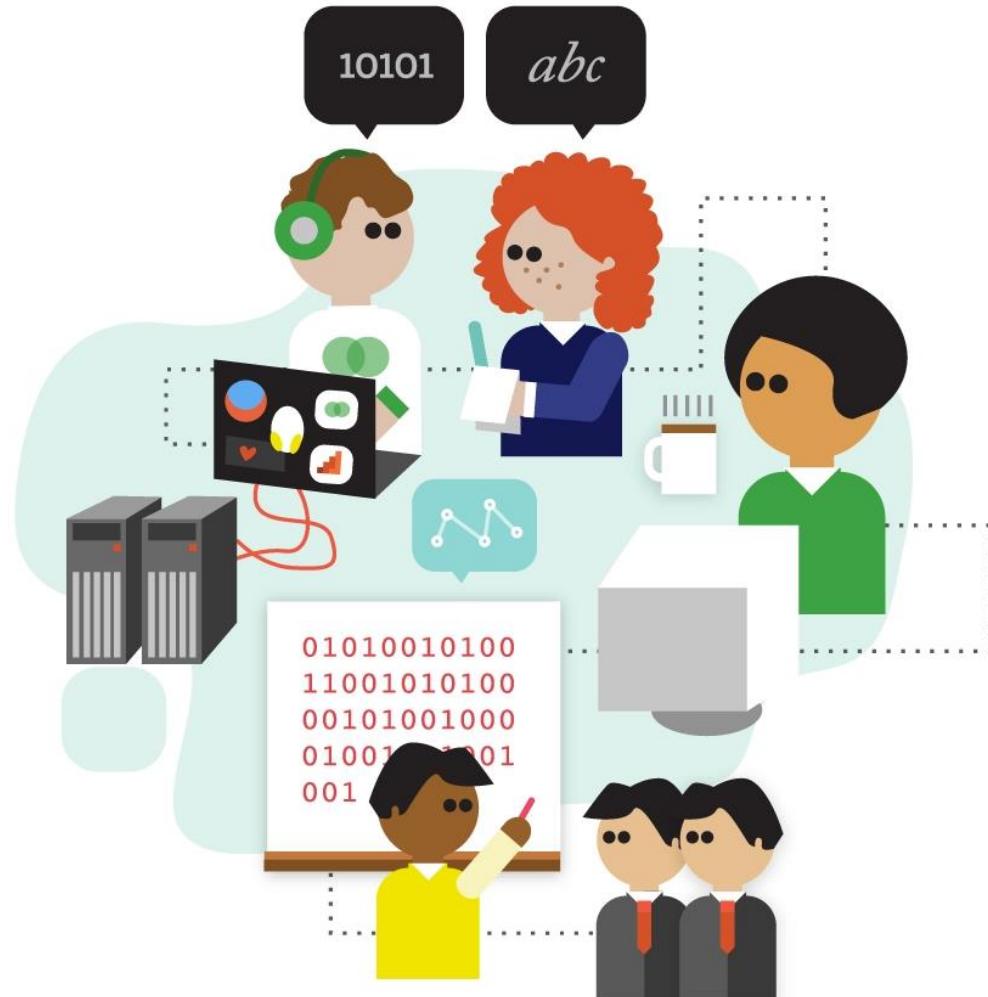
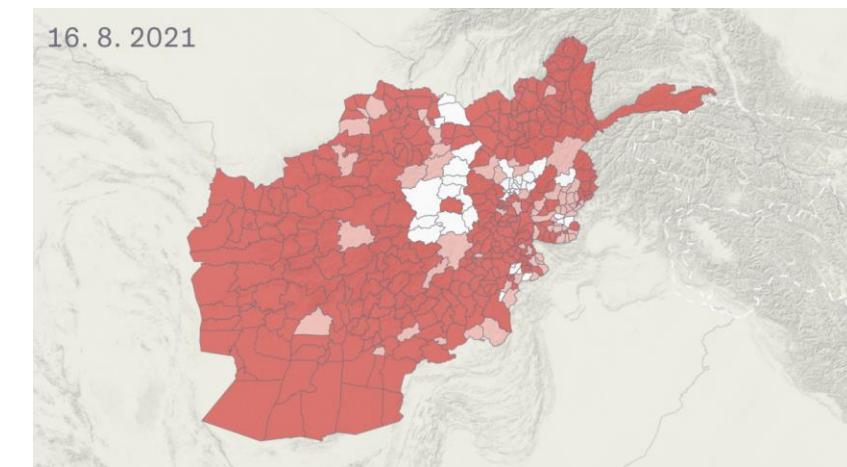
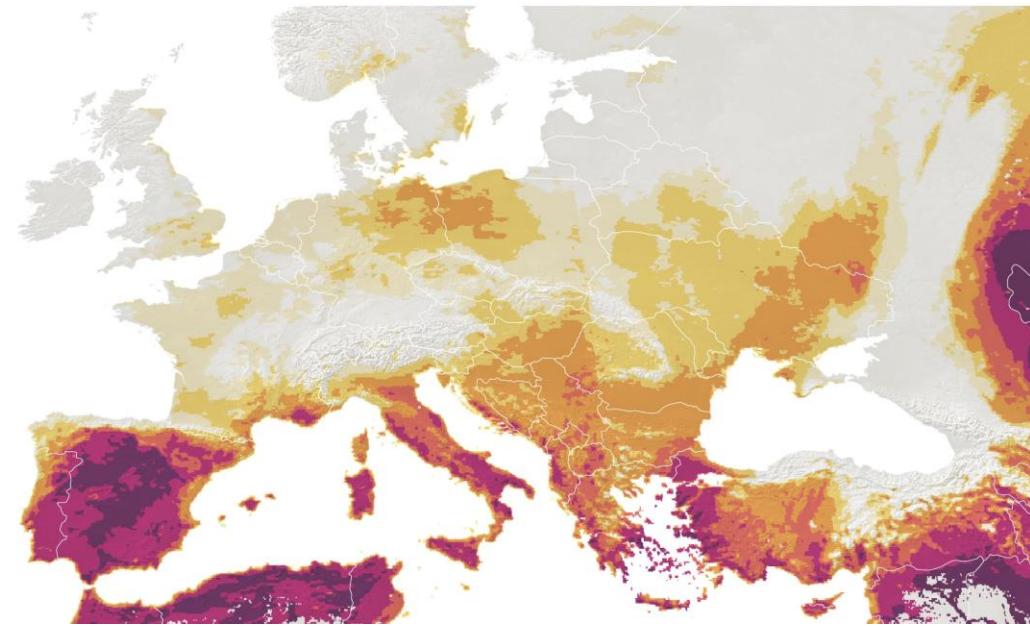
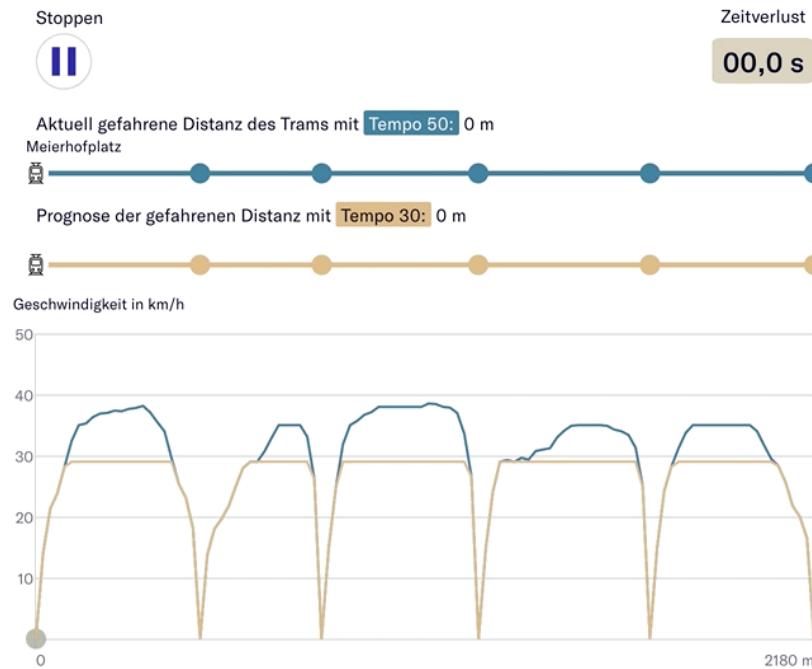
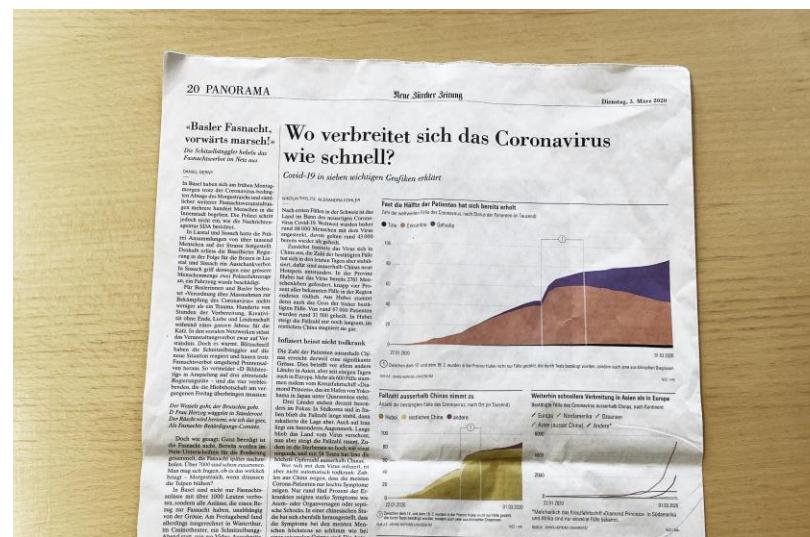
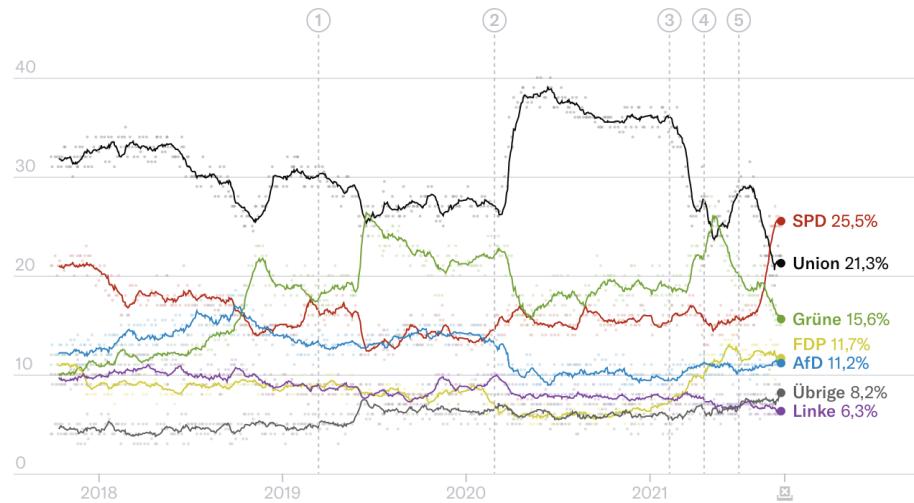


Bild: Kate Hudson



https://github.com/nzzdev/ddj_sfgz

Was wir bei der NZZ so machen:



Inhalt dieses Kursblocks

- **Teil 1: Grundlagen des datenbasierten Arbeitens**
 - Übung: Thema wählen, Hypothese finden
- **Teil 2: Wie man Datensätze findet und verarbeitet**
 - Übung: Recherche, Datensatz aufbereiten
- **Teil 3: Die gute Grafik**
 - Übung: Datensatz visualisieren, Verschiedene Tools ausprobieren
- **Teil 4: Wie man seine Erkenntnisse kommuniziert**
 - Übung: Fertigstellen der eigenen Recherche
 - Präsentation

Euer Projekt in diesem Kursblock:

- Eine Grafik
- Eine kurze Beschreibung dazu, was gezeigt wird
- Eine Präsentation
- Ihr versucht, mit eurem Werk eine Fragestellung zu beantworten
- Themen- Softwarewahl frei

Bewertung:

- Fragestellung
- Recherche
- Analyse
- Präsentation

Leistungsnachweis in Zweiergruppen

- These aufstellen
zb: «Frauen verdienten auch 2020 weniger als Männer»
- Daten suchen, bewerten und auswerten
- Grafik (Visualisierung) erstellen
- Kurzen, journalistischen Text dazu schreiben (1000 – 2000 Zeichen)
- präsentieren
- Thema und Wahl der Tools ist frei
- Die These kann auch nicht beantwortet werden – scheitern erlaubt

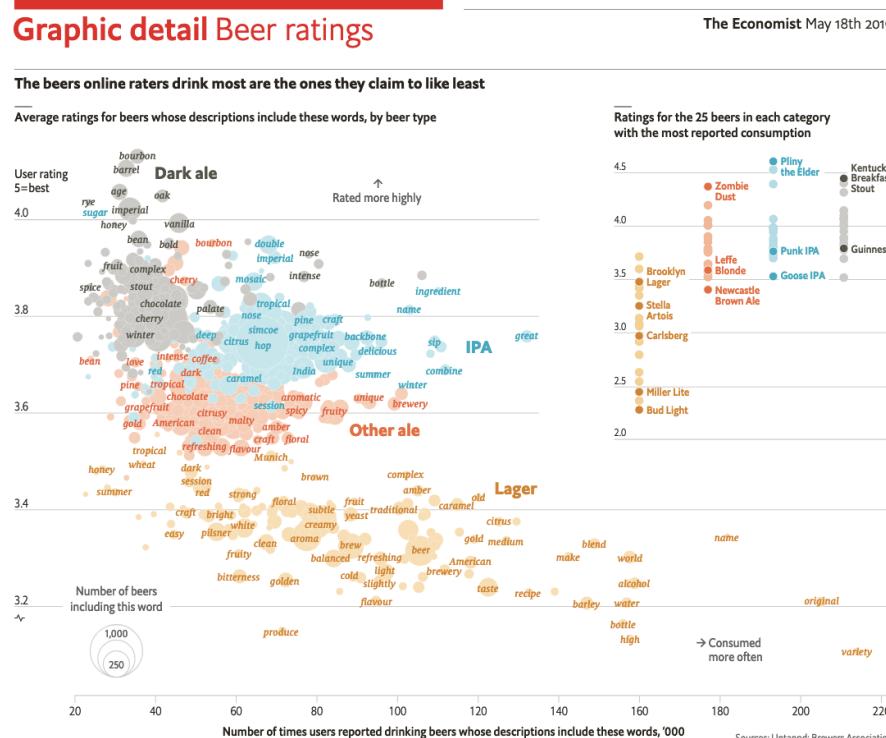
Präsentation sollte enthalten

- Was war unsere These?
- Wie sind wir vorgegangen?
- Haben wir Daten gefunden? Wie gut sind sie?
- Zu welchem Ergebnis sind wir gekommen?
- Wieso haben wir uns für diese Visualisierung entschieden?

Bewertet wird

- Prozess: Themenfindung, Schwierigkeiten bewältigen, Lösungen finden
- Qualität des Produkts
 - Visualisierung: Aussagekraft, Quellen, etc. enthalten?
 - Text: Interpretation der Ergebnisse
- Präsentation

Graphic detail Beer ratings



Familiarity Fosters contempt

Beer snobs guzzle lagers they claim to dislike. How long can that last?

CARLSBERG, A DANISH brewery, used to boast that its lager was "probably the best beer in the world". No longer. In March it began selling a new pilsner—a pale, Czech-style lager—after admitting that drinkers had soured on its original recipe.

Data from Untappd, a beer-rating site with 7m (mostly American) users, confirm that pontificating pint-swillers turn their noses up at mass-market lager. Among the 5,000 beers its users reported drinking most often, lagers—made with "bottom-fermenting" yeast, which yields a light-bodied, mild brew—are rated 3.29 out of 5 on average. The rest get an average of 3.69.

Moreover, the lagers online raters like most don't taste like lager. When grouped by the words in Untappd descriptions (many copied from labels), the best-rated

terms are ones mostly used for ale, such as "tropical" and "dark". Yet despite such poor reviews, the specific beers Untappd users say they drink most often are lagers. Why?

One explanation is fragmentation. Though reported consumption tends to be higher for individual lagers than for ales, there are far more ales than lagers. As a result, ales account for 73% of drinking of the 5,000 leading beers recorded on Untappd.

But crowd-sourced data are a poor measure of overall demand. According to IWSR, a research firm, Americans buy six times as much mass-market lager as craft beer.

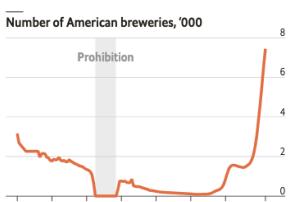
Most drinkers are not beer snobs, and

even ale devotees might secretly enjoy a frosty lager on a hot day. And most importantly, lagers dominate supply chains. Craft ales abound at organic grocers and hipster bars; Carlsberg (rated 2.96) and Budweiser (2.54) are everywhere.

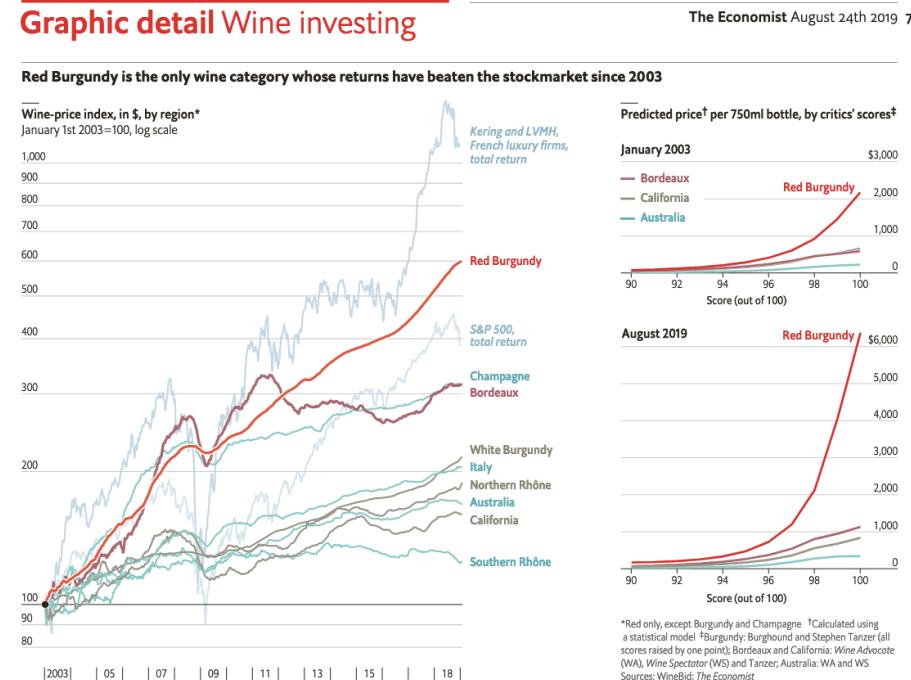
Low costs originally gave lager its distribution advantage. Its cold fermentation translates well to large batches, and using fewer hops saves money. In the 19th century these economies of scale let big firms flood America with watery lager. Prohibition reinforced this pattern: most craft houses closed shop for good, while large producers resumed brewing afterwards.

In recent years the market as a whole has inched closer to Untappd users' preferences. In 2010–18 American consumption of mass-market lager fell by 12.5%, while that of craft beer doubled—even though craft costs 67% more than lager on average.

Unfortunately for the beer industry, it sells so much lager that this switch has hurt it. Real revenues in America are down 9% since 2010. Giants like Carlsberg face an extra obstacle. Even if they launch or buy a rich, craft-style ale, snobs may shun it because it was made by a behemoth. ■



Graphic detail Wine investing



A cellar's market

Want a top-performing liquid asset?
Try Pinot Noir

WINE COLLECTORS like to proclaim that "all roads lead to Burgundy." They often wince at the plonk they drank when starting their hobby. In America and Australia, a common entry point is local "fruit bombs": heavy, alcoholic wines that taste of plum or blackberry, bear the vanilla or mocha imprint of oak barrels; and should be drunk within a few years of bottling.

As oenophiles gain experience, they start seeking reds to have with, say, chicken as well as steak. That leads to lower-octane French options: Cabernet Sauvignon from Bordeaux rather than Napa; Rhône Syrah instead of Barossa Shiraz. But once you value complexity and finesse over power, your vinous destination is pre-ordained.

Encyclopaedic wine knowledge is most precious in Burgundy. The French region is split into hundreds of named vineyards. In turn, myriad producers own specific rows within each vineyard, from which they all

make unique wines. This yields thousands of distinct pairings, each consisting of a few thousand bottles at most.

Moreover, red Burgundy is made from Pinot Noir, a grape with a maddening ageing pattern. After a few years of storage, it tends to "shut down" and lose flavour. The best wines blossom after a few decades, but many never "wake up" from their slumber.

In the past, Burgundy's complexity and small output relegated it to a market niche. A decade ago, Bordeaux—which makes fewer distinct wines in larger batches—became popular in Asia, and prices soared. But the bubble burst in 2012, when China's government began to frown on lavish gifts.

As tastes moved on from commodified Bordeaux, mastery of Burgundy became seen as the test of connoisseurship, both in Asia and the West. But the region's vast array of wines—including trophies as scarce as 300 bottles a year—makes reliable pricing data hard to find. Among the hundreds of fine red Burgundies, Liv-ex, a marketplace, includes just 11 in its regional index.

To create a sturdier measure, WineBid, the biggest online wine auctioneer, kindly gave us a full sales record for every wine sold at least ten times on its site since 2003. The data contain 1.6m lots, covering 33,000 wines. We built portfolios of 50–500 of the

most expensive unique labels (one vintage of one wine) from each region. We then estimated the returns for each portfolio, before storage and transaction costs.

Collectors who have drunk most of their Pinot already may need another glass after seeing the results. By the end of 2018, red Burgundy had returned 497%, versus 279% for the S&P 500. (Our index does not extend to 2019, since many of the wines it contains have not been traded this year.) The index has also been less volatile than stocks are, though this may be an artefact of how it is calculated: no one knows what each wine would have sold for in the crash of 2008–09. Bordeaux and Champagne rose by 214% in 2003–18; everywhere else did worse.

It is hard to fathom how Burgundy can maintain such appreciation. Many people can buy a \$300 bottle. But at \$3,000, the market depends on the whims of the rich. Even if prices keep rising, the best-performing stocks tend to beat their vinous peers. For example, Kering and LVMH—luxury conglomerates whose owners have bought Burgundy vineyards—returned 958% in 2003–18. And with dividend yields over 2% in recent years, they have paid enough income for a *grand cru* bottle, too. The best way to make money in Burgundy is probably making wine, not buying it. ■

Was bringt Datenkompetenz?



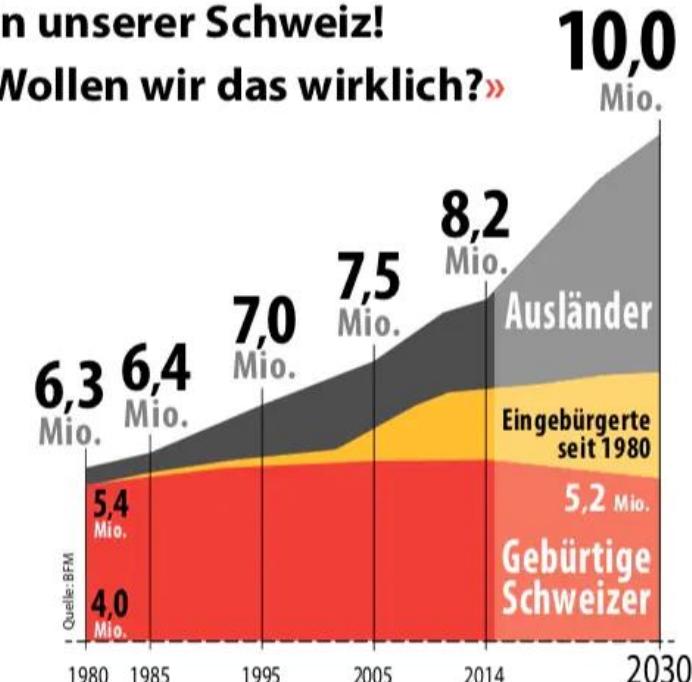
Toni Brunner,
Parteipräsident
SVP Schweiz

Liebe 20-Minuten-Leserinnen und Leser

Jährlich wächst unser Land wegen der ungesteuerten Zuwanderung um eine Stadt St. Gallen (+80'000 Personen). Wenn das so weiter geht, wird die Schweiz bereits im Jahr 2030 über 10 Millionen Einwohner zählen, die Hälfte davon Ausländer und Eingebürgerte. Die Folgen: verbaute Landschaften, überfüllte Züge, Stau auf den Strassen, teure Wohnungen und Häuser, Kulturwandel am Arbeitsplatz, immer mehr Kosten (Arbeitslosenkasse, Sozialhilfe usw.). Die SVP ist die einzige Partei, die hier Einhalt gebietet. Die SVP will, dass wir die Zuwanderung wieder selber steuern. Nur so können wir dafür sorgen, dass die Bevölkerung in der Schweiz genügend Arbeitsplätze findet. Am 18. Oktober können Sie dabei mithelfen. Bewahren Sie unser Land. Wählen Sie SVP. Herzlichen Dank.

Ihr Toni Brunner

**«10 Millionen Einwohner
in unserer Schweiz!
Wollen wir das wirklich?»**

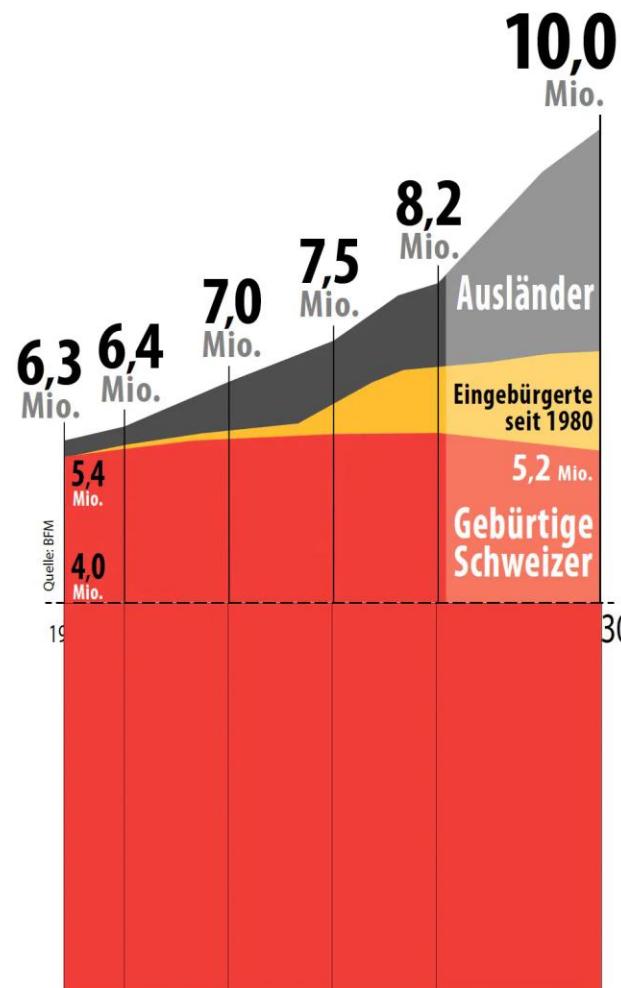


Anzeige
Jetzt wählen!

**Stellen Sie bei den Wahlen am
18. Okt. die Weichen richtig! Danke.**

Was bringt Datenkompetenz?

Korrekte Darstellung:



Liebe 20-Minuten-Leserinnen und Leser

Jährlich wächst unser Land wegen der ungesteuerten Zuwendung um eine Stadt St. Gallen (+80'000 Personen). Wenn das so weiter geht, wird die Schweiz bereits im Jahr 2030 über 10 Millionen Einwohner zählen, die Hälfte davon Ausländer und Eingebürgerte. Die Folgen: verbaute Landschaften, überfüllte Züge, Stau auf den Strassen, teure Wohnungen und Häuser, Kulturwandel am Arbeitsplatz, immer mehr Kosten (Arbeitslosenkasse, Sozialhilfe usw.). Die SVP ist die einzige Partei, die hier Einhalt gebietet. Die SVP will, dass wir die Zuwendung wieder selber steuern. Nur so können wir dafür sorgen, dass die Bevölkerung in der Schweiz genügend Arbeitsplätze findet. Am 18. Oktober können Sie dabei mithelfen. Bewahren Sie unser Land. Wählen Sie SVP. Herzlichen Dank.

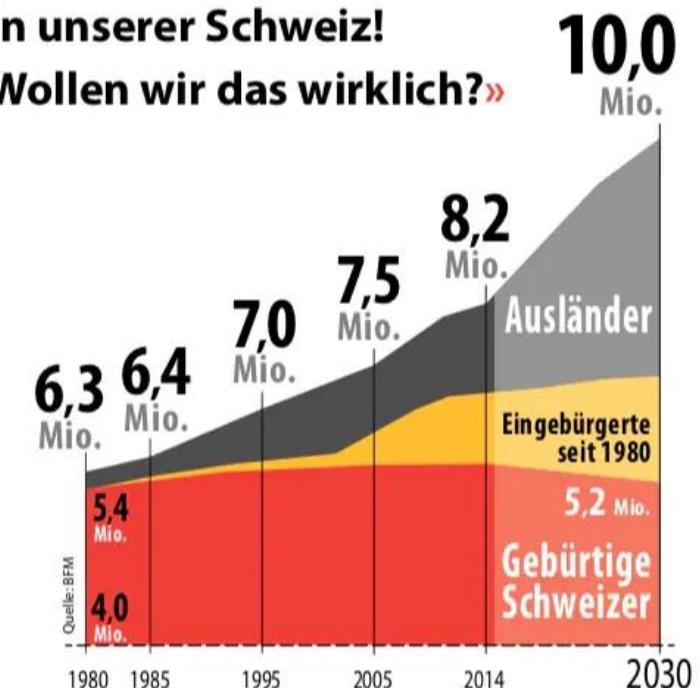
Ihr Toni Brunner

Bild: SVP/Blick.ch

Anzeige

Jetzt wählen!

«10 Millionen Einwohner in unserer Schweiz! Wollen wir das wirklich?»



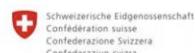
Zunahme der Bevölkerung in der Schweiz von 1985–2030

Stellen Sie bei den Wahlen am 18. Okt. die Weichen richtig! Danke.

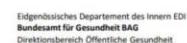
Was bringt Datenkompetenz?

Beruflich

- Datenkompetenz ist gefragt: In öffentlichen Stellen, in Unternehmen, in der öffentlichen Debatte
- Die Kompetenz ist noch nicht überall vorhanden



Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra



Eidgenössisches Departement des Innern EDI
Bundesamt für Gesundheit BAG
Direktionsbereich Öffentliche Gesundheit

Doses de vaccin (Pfizer/BioNTech) et (Moderna) administrées à ce jour : total au 24.1.2021

	Total CH	Arc lémanique			Plateau				Nord-ouest de la Suisse				Zurich
		GE	VD	VS	FR	NE	JU	BE	SO	BS	BL	AG	
Doses livrées	535'115	29'220	41'475	35'230	16'740	12'390	5'895	55'250	16'740	23'700	23'295	38'700	78'935
Nombre total de doses administrées à ce jour	197'368	11'811	15'847	10'210	6'084	2'585	2'032	14'308	10'234	10'890	8'315	12'973	25'528
Doses administrées pour 100 habitants	2.29	2.34	1.97	2.95	1.89	1.46	2.76	1.38	3.72	5.56	2.87	1.89	1.66

	Nord-est de la Suisse					Suisse centrale					Grisons / Tessin		FL		
	SH	TG	AR	AI	SG	GL	SZ	ZG	LU	NW	OW	UR	GR	TI	
Doses livrées	6'195	16'635	4'520	1'375	27'885	2'950	10'815	7'770	24'705	4'120	3'045	4'020	13'290	27'145	3'075
Nombre total de doses administrées à ce jour	3'982	4'072	2'349	917	8'870	1'540	3'250	6'599	11'241	3'218	1'866	1'361	4'316	11'815	1'155
Doses administrées pour 100 habitants	4.84	1.46	4.24	5.69	1.74	3.79	2.03	5.17	2.72	7.47	4.92	3.71	2.17	3.36	

- Berufliche Weiterbildung: Je akademischer, desto wichtiger wird Datenkompetenz

Privat

- Datenkompetenz hilft, das SVP-Extrablatt kritisch zu prüfen
- ...die Welt zu verstehen
- ...eigenständig recherchieren zu können anstatt sich auf andere verlassen zu müssen
- ...gut informiert den Abstimmungszettel auszufüllen, finanzielle Entscheidungen zu treffen, und so weiter

Was ist Datenjournalismus?



Klassische Reporterarbeit:

- Wichtige Fragen identifizieren und diesen nachgehen
- Mit Leuten telefonieren, die Sachen besser wissen als man selbst
- Die besten Datenquellen identifizieren
- Entscheiden: Was ist wichtig, was nicht?
- Schreiben!

 A screenshot of a computer monitor showing a code editor with several lines of programming code. The code is written in a language like Ruby or Python, dealing with account and application credentials and environment endpoints. The background of the slide features a faint watermark of the 123RF logo.

Programmieren:

- Daten beschaffen, formatieren, analysieren, Modelle rechnen, Daten weiterschicken
- Applikationen und Darstellungen fürs Web programmieren
- Schnell etwas in Excel zusammenrechnen
- Explorative Analyse mit Grafiken

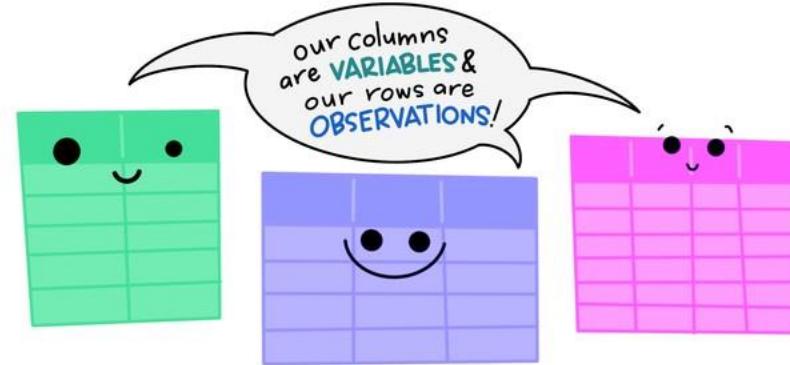


Präsentation

- Die Daten möglichst effizient kommunizieren
- Die wichtigsten Daten auswählen
- Die beste Darstellungsform finden, aus kommunikativer und ästhetischer Sicht
- Methodikentscheidefällen: Was ist in welchem Fall angemessen?

Grundlagen: Der gute Datensatz

The standard structure of tidy data means that
“tidy datasets are all alike...”



“...but every messy dataset is
messy in its own way.”

—HADLEY WICKHAM

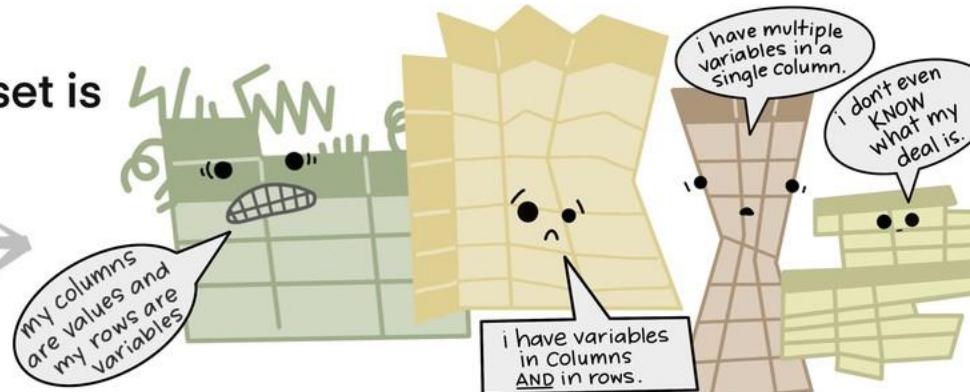


Bild: Allison Horst

Grundsatz:

- Ein gleichmässiges Raster, die erste Zeile sind die Spaltennahmen
- Jeden Spalte ist eine Variable, jede Zeile ist eine Beobachtung

Vorteile

- Keine Umstände mit «cleaning», man kann gleich zur Analyse schreiten
- Vorteil: Aggregation ist einfach, schnell, sauber
- Filter sind schnell angewendet

Nachteile

- Übersichtlichkeit
- Grafiktools wollen teils Datenreihen

Automatisches Speichern AUS ⌂ ⌄ ... WPP2019_POP_F01_1_TOTAL_POPULATION_BOTH_SEXES

Start Einfügen Zeichnen Seitenlayout Formeln Daten Überprüfen Ansicht Sie wünschen Freigeben Kommentare

Einfügen Arial 9 A A Standard Bedingte Formatierung Einfügen Einfügen v Als Tabelle formatieren Löschchen v Zellenformatvorlagen v Bearbeiten Daten analysieren Vertraulichkeit

A1 x ✓ fx |

A B C D E F G H I J K L M N

United Nations Population Division Department of Economic and Social Affairs

World Population Prospects 2019
File POP/1-1: Total population (both sexes combined) by region, subregion and country, annually for 1950-2100 (thousands)
Estimates, 1950 - 2020
POP/DB/WPP/Rev.2019/POP/F01-1
© August 2019 by United Nations, made available under a Creative Commons license CC BY 3.0 IGO: <http://creativecommons.org/licenses/by/3.0/igo/>
Suggested citation: United Nations, Department of Economic and Social Affairs, Population Division (2019). World Population Prospects 2019, Online Edition. Rev. 1.

Total population, both sexes combined, as of 1 July (thousands)

Index	Variant	Region, subregion, country or area *	Notes	Country code	Type	Parent code	1950	1951	1952	1953	1954	1955	1956
1	Estimates	WORLD		900	World	0	2 536 431	2 584 034	2 630 862	2 677 609	2 724 847	2 773 020	2 822 44
2	Estimates	UN development groups	a	1803	Label/Separator	900
3	Estimates	More developed regions	b	901	Development Group	1803	814 819	824 004	833 720	843 788	854 060	864 430	874 82
4	Estimates	Less developed regions	c	902	Development Group	1803	1 721 612	1 760 031	1 797 142	1 833 822	1 870 786	1 908 590	1 947 62
5	Estimates	Least developed countries	d	941	Development Group	902	195 428	199 180	203 015	206 986	211 133	215 486	220 06
6	Estimates	Less developed regions, excluding least developed countries	e	934	Development Group	902	1 526 184	1 560 850	1 594 126	1 626 836	1 659 653	1 693 104	1 727 55
7	Estimates	Less developed regions, excluding China	f	948	Development Group	1803	1 157 420	1 179 933	1 203 963	1 229 440	1 256 303	1 284 497	1 313 97
8	Estimates	Land-locked Developing Countries (LLDC)	g	1636	Special other	1803	103 803	105 870	108 079	110 423	112 894	115 488	118 20
9	Estimates	Small Island Developing States (SIDS)	h	1637	Special other	1803	23 771	24 209	24 685	25 187	25 710	26 249	26 80
10	Estimates	World Bank income groups	i	1802	Label/Separator	900
11	Estimates	High-income countries	j	1503	Income Group	1802	694 989	703 004	711 534	720 436	729 596	738 929	748 37
12	Estimates	Middle-income countries	k	1517	Income Group	1802	1 703 596	1 741 086	1 777 129	1 812 536	1 847 973	1 883 962	1 920 87
13	Estimates	Upper-middle-income countries	l	1502	Income Group	1517	938 931	962 816	984 350	1 004 408	1 023 703	1 042 796	1 062 08
14	Estimates	Lower-middle-income countries	m	1501	Income Group	1517	764 666	778 270	792 779	808 129	824 269	841 166	858 79
15	Estimates	Low-income countries	n	1500	Income Group	1802	137 042	139 119	141 352	143 769	146 387	149 214	152 24
16	Estimates	No income group available	o	1518	Income Group	1802	804	826	847	868	891	915	94
17	Estimates	Geographic regions	p	1840	Label/Separator	900
18	Estimates	Africa	q	903	Region	1840	227 794	232 328	237 097	242 092	247 311	252 749	258 40
19	Estimates	Asia	r	935	Region	1840	1 404 909	1 435 819	1 464 834	1 492 895	1 520 768	1 549 042	1 578 12
20	Estimates	Europe	s	908	Region	1840	549 329	554 325	559 694	565 282	570 970	576 679	582 37

ESTIMATES MEDIUM VARIANT HIGH VARIANT LOW VARIANT CONSTANT-FERTILITY INSTANT-REPLACEMENT MOMENTUM ZERO-MIGRATION +

Bereit 125%

WPP2019_TotalPopulationBySex.csv

Add Column Delete Column Add Row Delete Row Convert...

LocID	Location	VarID	Variant	Time	MidPeriod	PopMale	PopFemale	PopTotal	PopDensity
4	Afghanistan	2	Medium	1950	1950.5	4099.243	3652.874	7752.117	11.874
4	Afghanistan	2	Medium	1951	1951.5	4134.756	3705.395	7840.151	12.009
4	Afghanistan	2	Medium	1952	1952.5	4174.45	3761.546	7935.996	12.156
4	Afghanistan	2	Medium	1953	1953.5	4218.336	3821.348	8039.684	12.315
4	Afghanistan	2	Medium	1954	1954.5	4266.484	3884.832	8151.316	12.486
4	Afghanistan	2	Medium	1955	1955.5	4318.945	3952.047	8270.992	12.669
4	Afghanistan	2	Medium	1956	1956.5	4375.8	4023.073	8398.873	12.865
4	Afghanistan	2	Medium	1957	1957.5	4437.157	4098	8535.157	13.073
4	Afghanistan	2	Medium	1958	1958.5	4503.156	4176.941	8680.097	13.295
4	Afghanistan	2	Medium	1959	1959.5	4573.914	4260.033	8833.947	13.531
4	Afghanistan	2	Medium	1960	1960.5	4649.573	4347.394	8996.967	13.781
4	Afghanistan	2	Medium	1961	1961.5	4730.25	4439.156	9169.406	14.045
4	Afghanistan	2	Medium	1962	1962.5	4816.05	4535.392	9351.442	14.324
4	Afghanistan	2	Medium	1963	1963.5	4907.03	4636.17	9543.2	14.618
4	Afghanistan	2	Medium	1964	1964.5	5003.245	4741.527	9744.772	14.926
4	Afghanistan	2	Medium	1965	1965.5	5104.765	4851.553	9956.318	15.25
4	Afghanistan	2	Medium	1966	1966.5	5210.122	4964.718	10174.84	15.585
4	Afghanistan	2	Medium	1967	1967.5	5319.123	5080.813	10399.936	15.93
4	Afghanistan	2	Medium	1968	1968.5	5434.458	5202.606	10637.064	16.293
4	Afghanistan	2	Medium	1969	1969.5	5559.836	5333.936	10893.772	16.686
4	Afghanistan	2	Medium	1970	1970.5	5697.024	5476.63	11173.654	17.115
4	Afghanistan	2	Medium	1971	1971.5	5845.351	5630.099	11475.45	17.577
4	Afghanistan	2	Medium	1972	1972.5	6000.895	5790.327	11791.222	18.061
4	Afghanistan	2	Medium	1973	1973.5	6157.843	5951.12	12108.963	18.548
4	Afghanistan	2	Medium	1974	1974.5	6308.583	6104.377	12412.96	19.013
4	Afghanistan	2	Medium	1975	1975.5	6446.273	6242.891	12689.164	19.436
4	Afghanistan	2	Medium	1976	1976.5	6573.732	6369.361	12943.093	19.825
4	Afghanistan	2	Medium	1977	1977.5	6689.144	6482.15	13171.294	20.175
4	Afghanistan	2	Medium	1978	1978.5	6776.023	6565.176	13341.199	20.435
4	Afghanistan	2	Medium	1979	1979.5	6813.205	6597.855	13411.06	20.542
4	Afghanistan	2	Medium	1980	1980.5	6788.273	6568.227	13356.5	20.458
4	Afghanistan	2	Medium	1981	1981.5	6698.73	6472.949	13171.679	20.175
4	Afghanistan	2	Medium	1982	1982.5	6557.673	6324.845	12882.518	19.732

Encoding Unicode (UTF-8)

Separator , ; → |

Decimal . ,

Quote " " \ " none

Header

Das Seniorenrennen – verschiedene Datentypen

Name	Geburtstag	Zeit	Note	Qualifizierung
Urs	1963-03-22	13.45	genügend	TRUE
Peter	1960-03-10	14.33	genügend	TRUE
Hans	1960-09-02	13.56	genügend	TRUE
Marlis	1959-09-26	19.22	ungenügend	FALSE
Anna	1957-02-07	13.31	gut	TRUE
Brigitte	1954-06-22	12.09	sehr gut	TRUE

Variablen können verschiedene Formate annehmen:

- Text («String»)
- Zahl (Dezimalzahl)
- Datum
- Boolean
- ...

...und verschiedene Skalenniveaus aufweisen:

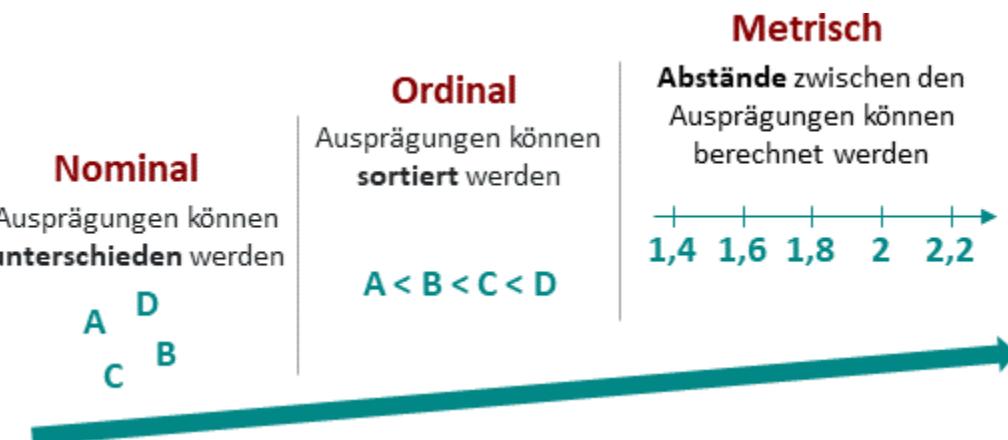


Bild: Datatab.de

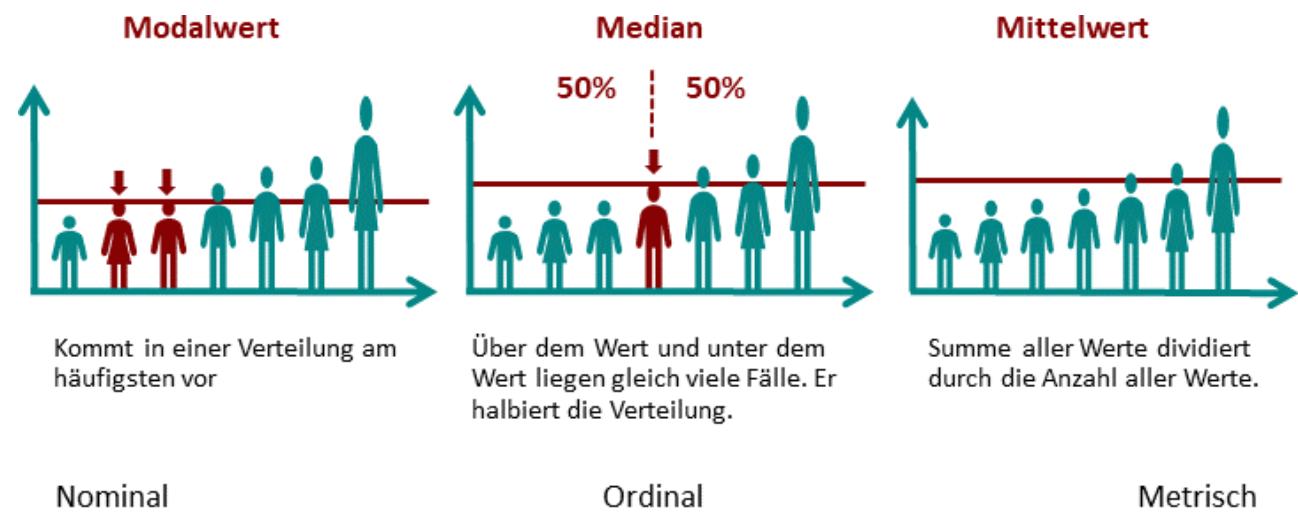
Aggregation: Wie kann ich Daten zusammenfassen?

Metrisch skaliert:

- Durchschnitt: Wie lange haben die Senioren im Durchschnitt für 100 Meter gebraucht? 60-er vs 50-er Jahrgänge?
- Median: Welches war der «mittlere Senior»?
- Summe: Wie lange sind sie alle zusammen gerannt?
- Rollender Durchschnitt, Rollender Median (Nur bei Zeitreihen)
- Min/Max Wer war der langsamste?
- Quantile: Wo ist die Grenze der 25% schnellsten Senioren?
- Standardabweichung: In welches Intervall fallen 68% oder 95% der Beobachtungen?
- Relationen: Wie viel Prozent schneller war die Schnellste als die Langsamste?

Ordinal oder Nominal

- Moduswert: Welcher Wert kommt am häufigsten vor? Was ist die häufigste Note unter den Senioren?
- Zählen: Wie viel % der Senioren haben sich qualifiziert?
- Wer hat die beste Note?



Median vs. Mittelwert

Mittelwert

- + Weit verbreitet, verständlich
- + Gut geeignet für Vergleiche zwischen Gruppen
- nicht robust gegenüber Ausreissern und ungleichen Verteilungen

Median

- + Robust gegenüber Ausreissern und ungleichen Verteilungen
- + Gut geeignet um den «typischen» Wert für eine Gruppe zu ermitteln
- + Muss nur ordinal skaliert sein (zB: welches ist die mittlere Bewertung eines Seniors?)
- Nicht so verbreitet, muss u.U. erklärt werden

Dreisatz- und Prozentrechnen

Dreisatz:

$$a \rightarrow b$$

$$c \rightarrow x$$

Formel: $c * b / a = x$

8 Mio \rightarrow 100%

1 Mio \rightarrow X

8 Mio \rightarrow 500

100'000 \rightarrow X

Achtung bei Prozenten

- Was ist 100%?
- Steigt um 200% \rightarrow Verdreifachung
- Bei Veränderungen immer von **Prozentpunkten** sprechen.

Bsp: In Portugal sind 80 Prozent der Einwohner vollständig gegen das Coronavirus geimpft, das sind 25 Prozentpunkte mehr als in der Schweiz (55%).

55% \rightarrow 100%

80% \rightarrow 145%

(... es sind rund 45 Prozent mehr)

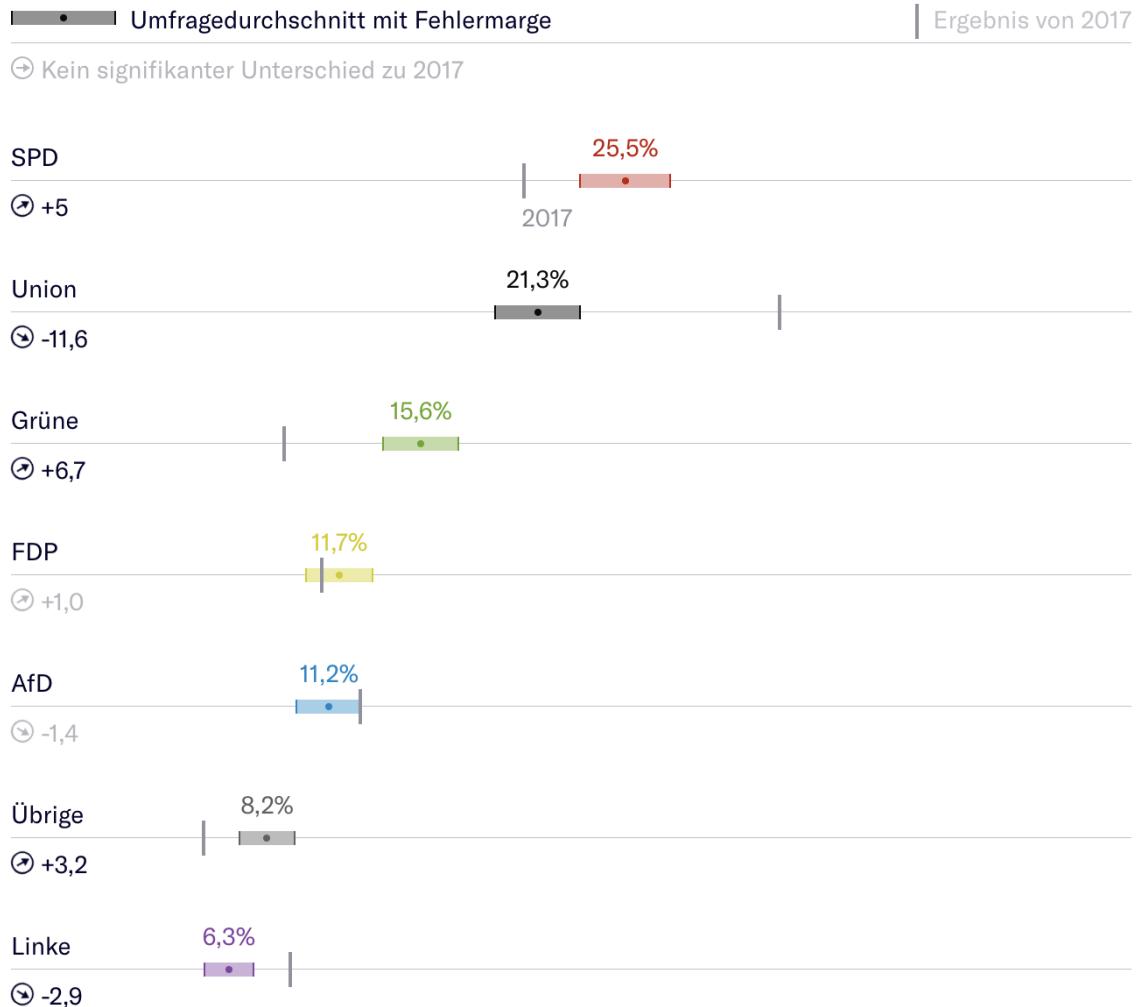
Statistische Unsicherheit

- Jede Umfrage ist mit einer Unsicherheitsmarge behaftet
- Meist zwischen 2 und 3 Prozent
- Erst wenn die Unterschiede *statistisch signifikant* sind, lässt sich eine klare Aussage treffen.
- Die Unsicherheitsbereiche werden bei Untergruppen grösser
- Die Methodik wird undurchsichtiger
- Mehr Umfragen würden helfen



Grüne legen gegenüber 2017 deutlich zu, die Union stürzt ab

Derzeitige Wahlabsicht der Befragten (in Prozent), verglichen mit dem Wahlresultat von 2017



Stand: 21. 9. 2021.

Quellen: [Wahlrecht.de](#), eigene Berechnungen

NZZ / Visuals & Editorial Tech

Statistische Unsicherheit

Abo Nachwahlbefragung zum CO₂-Gesetz

Ausgerechnet die Jungen sagten Nein

Keine Altersgruppe lehnte das CO₂-Gesetz und die beiden Agrarinitiativen so klar ab wie die 18- bis 34-Jährigen. Das wird zum Problem für die Klimastreikbewegung.



Luca De Carli

Publiziert: 14.06.2021, 15:22



287 Kommentare



Die Jugend hat das CO₂-Gesetz nicht zu Fall gebracht – und die FDP hat ein Problem

Die offizielle Nachwahlbefragung zu den Abstimmungsvorlagen des 13. Juni zeigt vor allem etwas sehr deutlich: Noch selten hat eine Abstimmung auf dem Land so viele Menschen an die Urnen gebracht wie die Agrarinitiativen.

Christina Neuhaus

50 Kommentare →

30.07.2021, 10.00 Uhr



Hören



Merken



Drucken



Teilen

Haben Sie das revidierte CO₂-Gesetz angenommen?

Total 16'249 gewichtete Antworten der Tamedia-Umfrage zur Abstimmung vom 13. Juni 2021, in Prozent

Stimmentscheid nach Alter



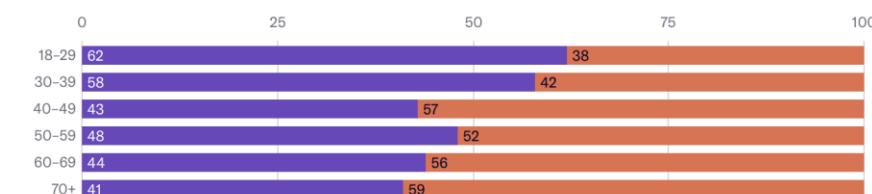
Umfragetage: 11.–13. Juni 2021. Maximaler Stichprobenfehlerbereich: 3 Prozentpunkte. Die Umfrage wurde in Zusammenarbeit mit der Leewas GmbH der Politologen Lucas Leemann und Fabio Wasserfallen durchgeführt. Weitere Informationen unter www.tamedia.ch.

Grafik: mat • Quelle: «20 Minuten»/Tamedia-Onlinenumfrage • Daten herunterladen

Die Vox-Befragung zeigte: je älter die Stimmbürgerinnen und Stimmbürger, desto mehr Nein-Anteile

Angaben in Prozent

● Ja ● Nein



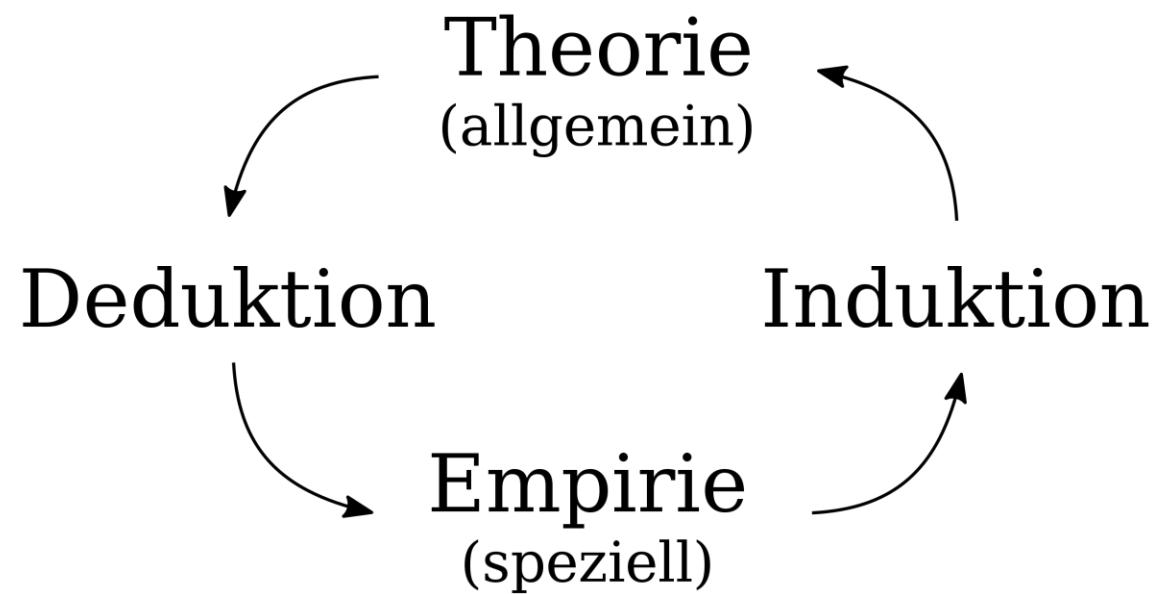
N = 3000. Nach jeder eidgenössischen Abstimmung führt GFS Bern im Auftrag der Bundeskanzlei eine repräsentative Umfrage unter 3000 zufällig ausgewählten Stimmberechtigten durch.

Quelle: GFS.Bern

NZZ / bsk.

Wie erlangen wir Wissen?

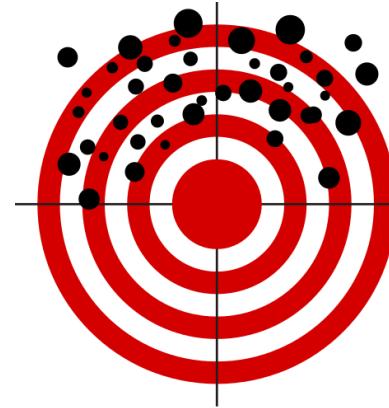
- Wir haben eine **Theorie**
- Wir stellen eine konkrete Fragestellung auf, wir haben eine **Hypothese** (und eine Nullhypothese)
- Wir **testen** diese **Hypothese**
 - **Qualitative** Methoden: Interviews, Gruppendiskussionen...
 - **Quantitative** Methoden: Umfragen, Statistische Analysen, Experimente
- Wir erlangen **Resultate**, anhand derer sich die Hypothesen **falsifizieren** lassen
- Unsere Arbeit wird von unseren Kollegen bewertet
- Sie wird Teil der wissenschaftlichen Literatur, wird zitiert, kritisiert etc...



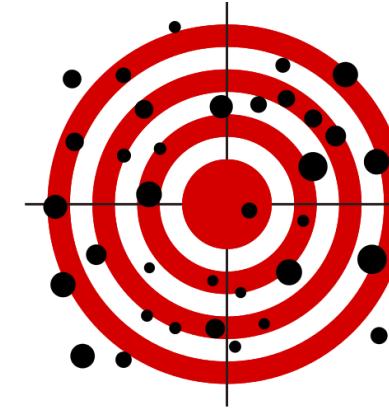
Was ist ein guter Test?

Gütekriterien eines Tests

- **Validität:** Messen wir das, was wir messen wollen?
- **Reliabilität:** Liefert unser Messinstrument zuverlässige Resultate?
- **Objektivität:** Hat die Forscherin einen Einfluss auf den Test?



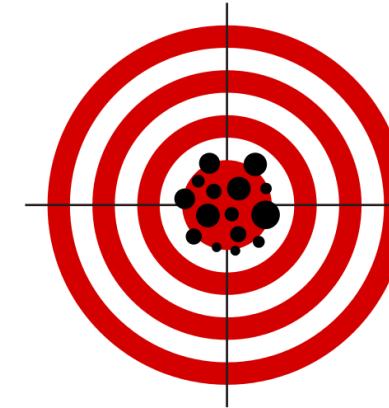
Unreliable & Unvalid



Unreliable, But Valid

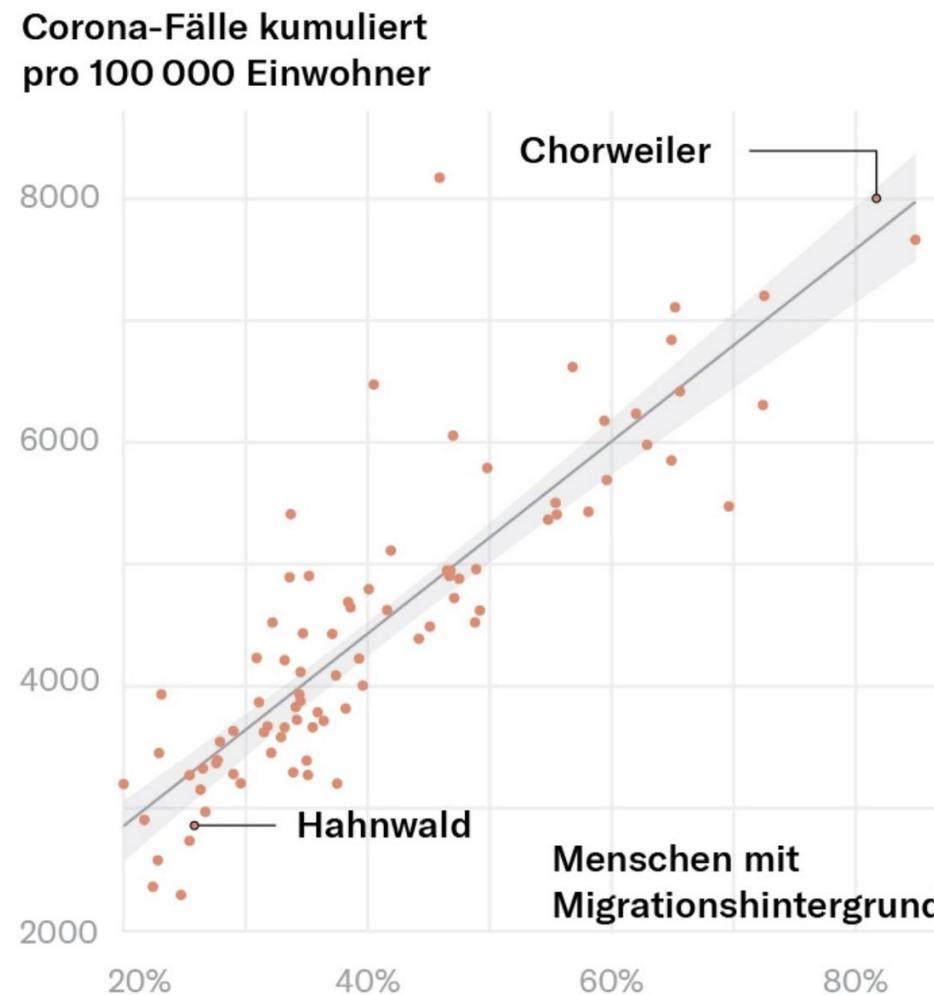


Reliable, Not Valid



Both Reliable & Valid

Der Klassiker: Korrelation



Stand Corona-Fallzahlen: 28. 4. 2021.

NZZ / sih., eck., nth.

Call:

```
lm(formula = Coronafälle_kumulativ_pro_100000_Einwohner ~ Arbeitslosenquote +
  Mieten_in_Euro_pro_Quadratmeter + Abiturientenquote + Migrationshintergrund +
  Haushaltsdichte_pro_Hektar + AfD_2017, data = koeln)
```

Residuals:

Min	1Q	Median	3Q	Max
-1127.0	-318.2	-126.7	191.0	3146.7

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1479.259	1439.305	1.028	0.307
Arbeitslosenquote	39.827	43.398	0.918	0.362
Mieten_in_Euro_pro_Quadratmeter	-1.820	140.903	-0.013	0.990
Abiturientenquote	-2.534	9.505	-0.267	0.790
Migrationshintergrund	58.943	12.114	4.866	0.00000572 ***
Haushaltsdichte_pro_Hektar	1.536	5.273	0.291	0.772
AfD_2017	44.914	59.389	0.756	0.452

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

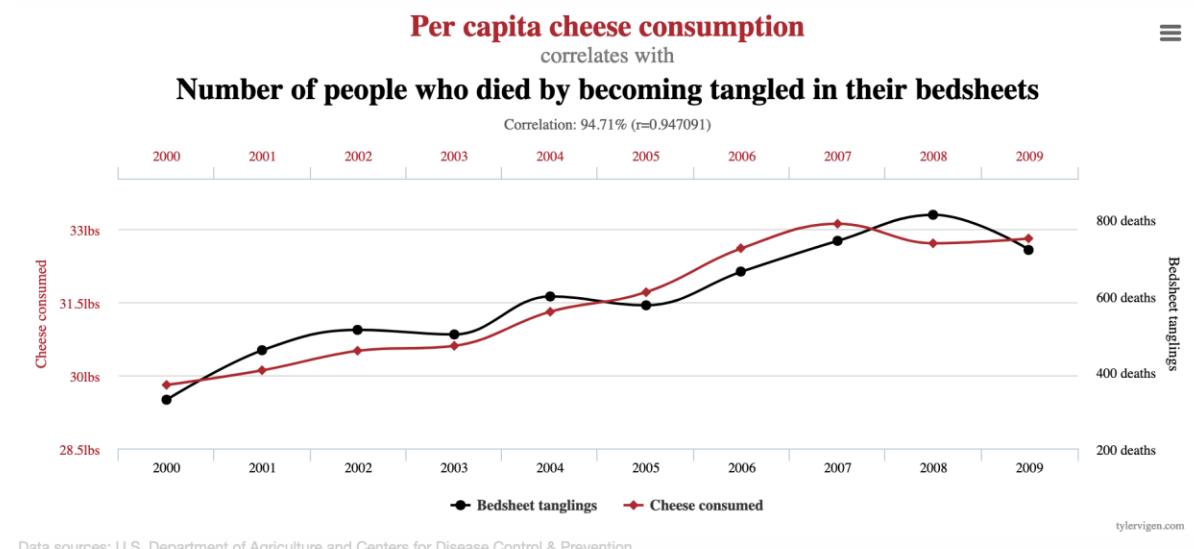
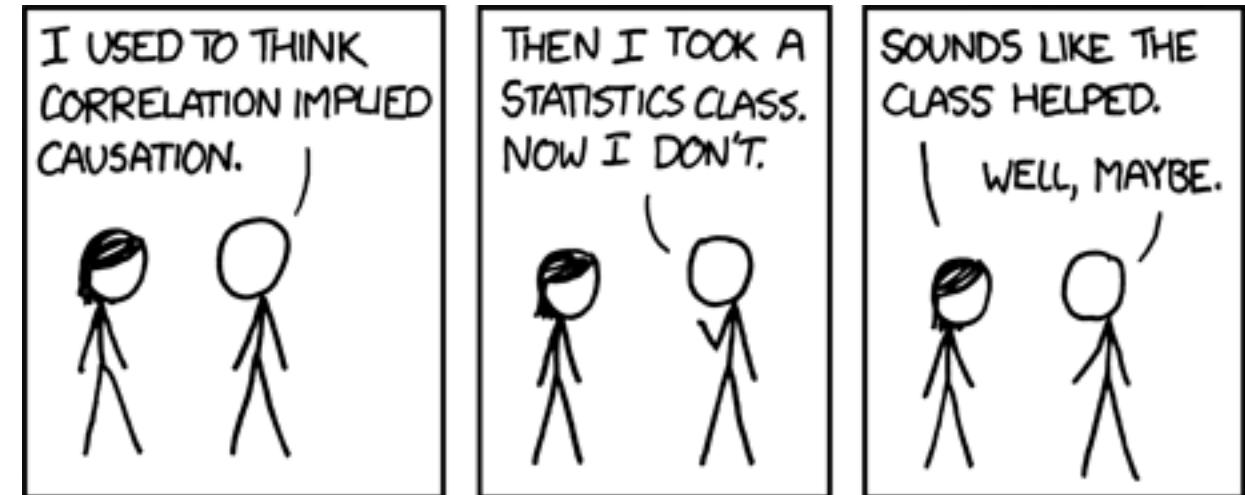
Residual standard error: 647.7 on 79 degrees of freedom

Multiple R-squared: 0.7713, Adjusted R-squared: 0.7539

F-statistic: 44.41 on 6 and 79 DF, p-value: < 0.000000000000022

Korrelation ist nicht Kausalität

- Ohne Experimente ist es nicht möglich, eine Kausalität zu beweisen
- Zusammenhänge können immer zufällig sein
- Es können immer Drittvariablen im Spiel sein, die man in seinem Modell nicht berechnet hat
- Es kann eine Kausalitätskette vorliegen
- Es können Interaktionseffekte vorliegen
- Die Richtung der Kausalität könnte umgekehrt sein als vermutet



Studien und Befragungen richtig interpretieren

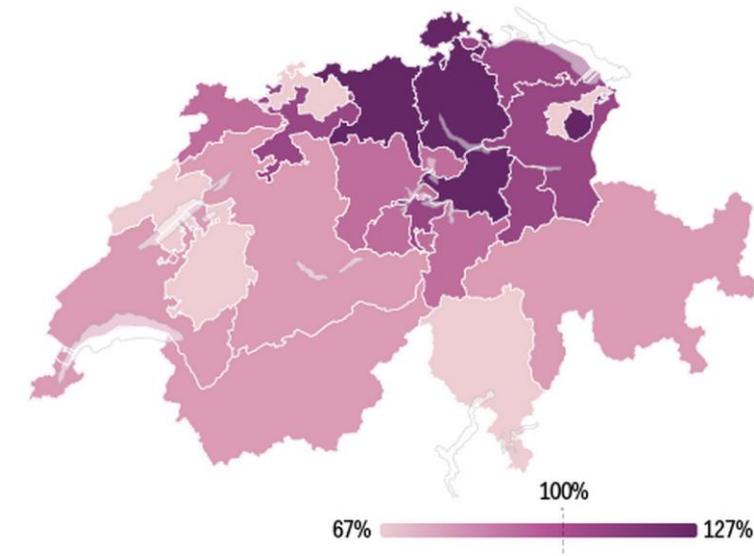
- Wer hat die Studie in wessen Auftrag erstellt?
 - Geht es nur darum, Aufmerksamkeit zu erhaschen?
 - Geht es um wirtschaftliche/politische Interessen?
- Ist die Studie in einer wissenschaftlichen Publikation mit Peer Review erschienen?
- Wird über eine neue Studie auch in anderen Medien berichtet?
- Wird ein Link zur Originalstudie, zum Datensatz angeboten?
- **Eine Studie macht noch keinen Beweis**

VERKAUFS-STATISTIK

Aktualisiert 25. März 2017, 19:22

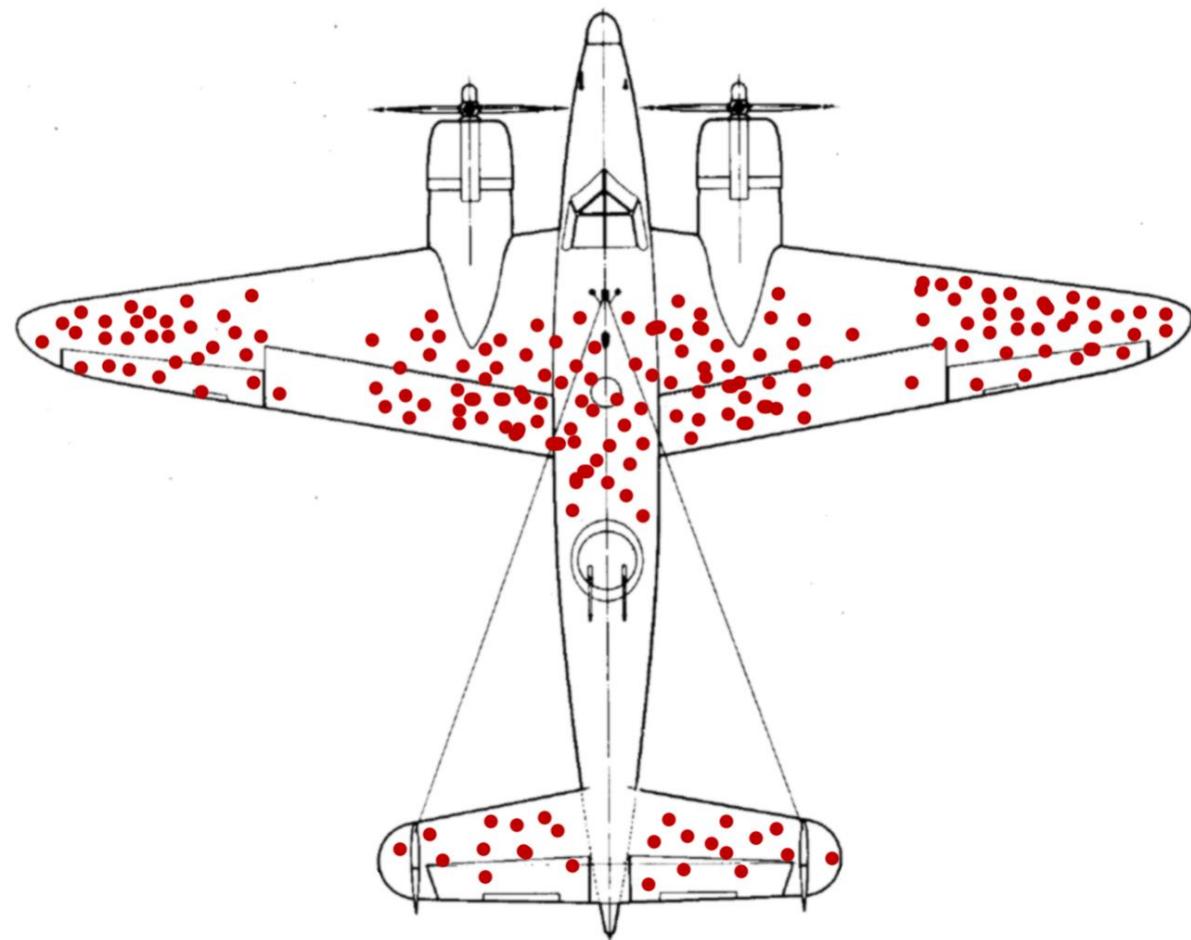
Hier verläuft der Sextoy-Grabен

«Fifty Shades Darker» sorgt für einen Sexspielzeug-Boom bei Jungen. Wer auf dem Land wohnt, bestellt aber eher online, um nicht dem Nachbarn zu begegnen.



«Bias» – Wo man sich überall irren kann

- **Confirmation Bias:** Ich nehme Fakten stärker wahr, die meine Sicht bestätigen
- **Publication Bias:** Wissenschaftler publizieren eher Resultate, welche einen Effekt zeigen.
- **Survivorship Bias:** Es wird sich nur ein Subset aller Fälle angeschaut, wenn man eigentlich alle Fälle anschauen sollte
- **Regression to the mean:** Interventionen nach Extremereignissen werden überschätzt



Wie wir bei der NZZ aus Daten Geschichten machen

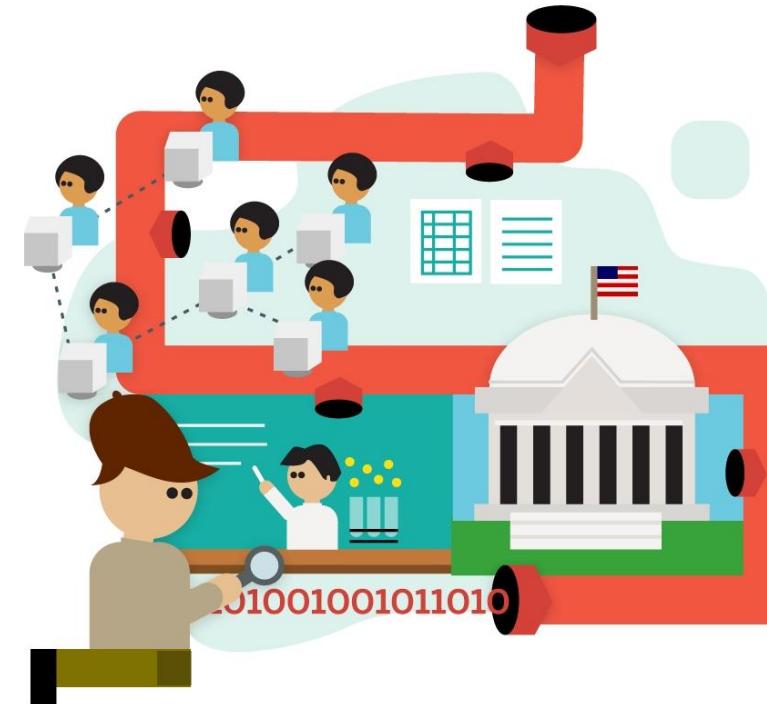
- Am Anfang steht entweder eine **Fragestellung** oder ein **unerforschter Datensatz**

Recherche von Daten: Was gibt es?

- Offene Daten, Daten auf Anfrage
- Scraping, Unstrukturierte Daten
- Geodaten

Analyse: Was machen wir damit?

- Verfügbare Daten einfach präsentieren
- «summary statistics»
- Modelle und Analysen rechnen



Wie wir bei der NZZ aus Daten Geschichten machen

Einschätzung

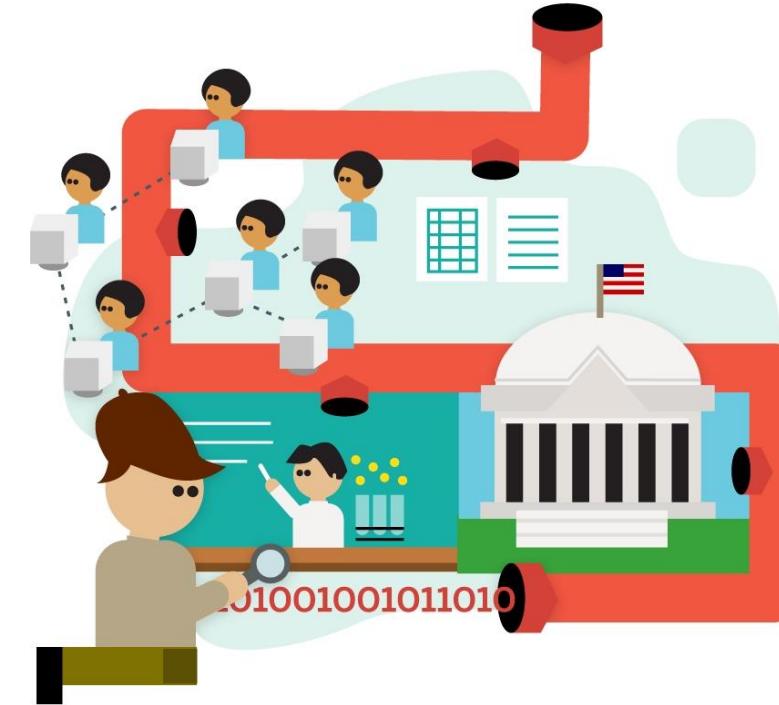
- Was sagen die Daten aus? Wird unsere Hypothese bestätigt (wenn wir eine hatten)?
- Welche gesellschaftlich relevanten Aspekte stehen mit der Geschichte im Zusammenhang?

Qualitative Recherche

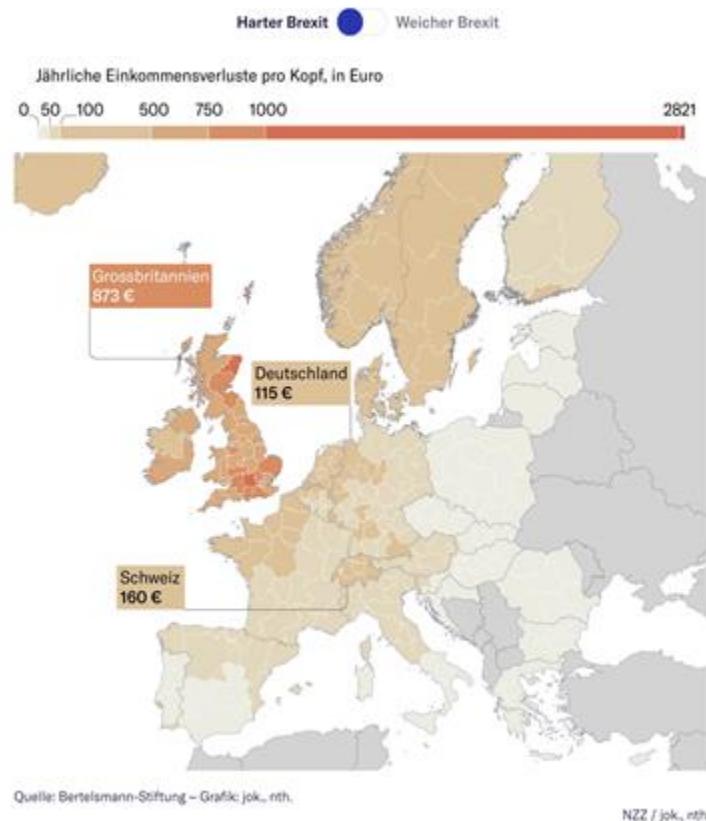
- Sprechen die Daten für sich selbst? Oder muss ein Experte sie einschätzen?
- Muss jemand zu den Daten Stellung nehmen können?

Verpackung

- Wie erzählen wir die Geschichte spannend? Welche Grafiken erzählen die Geschichte am besten?
- Wie verpacken wir das Ganze möglichst attraktiv?



«Schnelle» Storys



Je reicher das Land, desto mehr Rechte haben Frauen

Zusammenhang zwischen Wohlstandsniveau und Gleichstellung (Indexwert, 0-100)



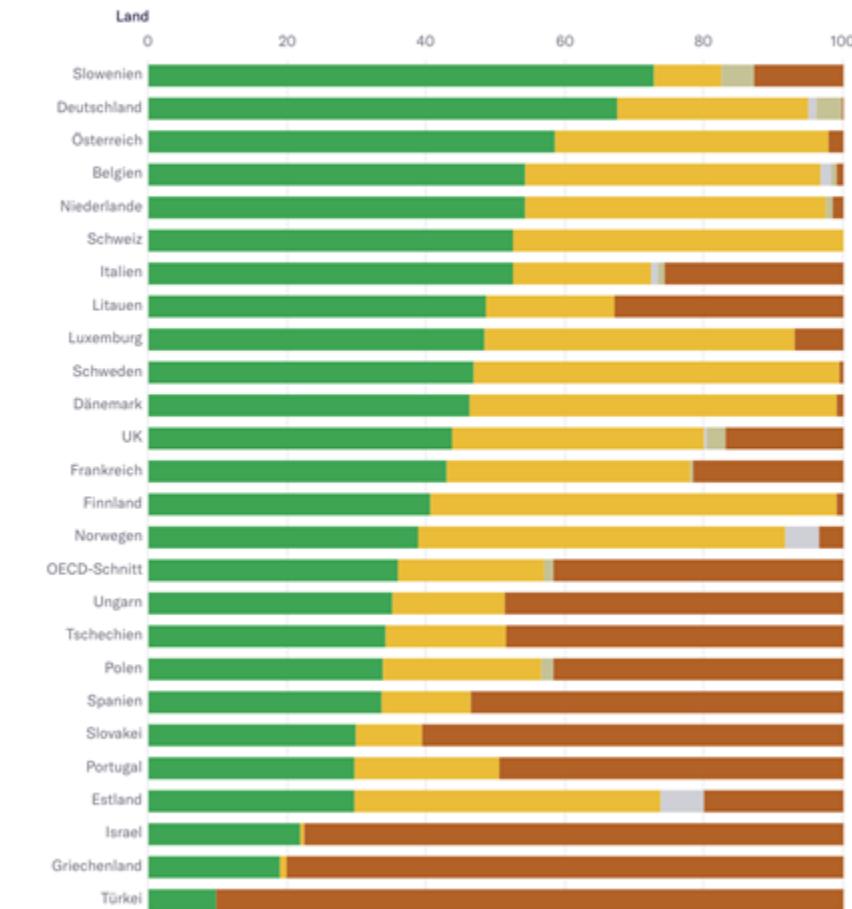
Die Trendlinie stellt statistische Erwartungswerte (Loess-Modell) für die Gleichstellung, basierend auf dem Wohlstandsniveau, dar.

Quellen: [Equal Measures 2030](#), [Weltbank](#)

Slowenien ist in Sachen Recycling Spitze

Methoden der Abfallverarbeitung, in Prozent der Gesamtmenge

● Recycling & Kompost ● Verbrennung mit Energiegewinnung ● Anderes
● Verbrennung ohne Energiegewinnung ● Müllhalde



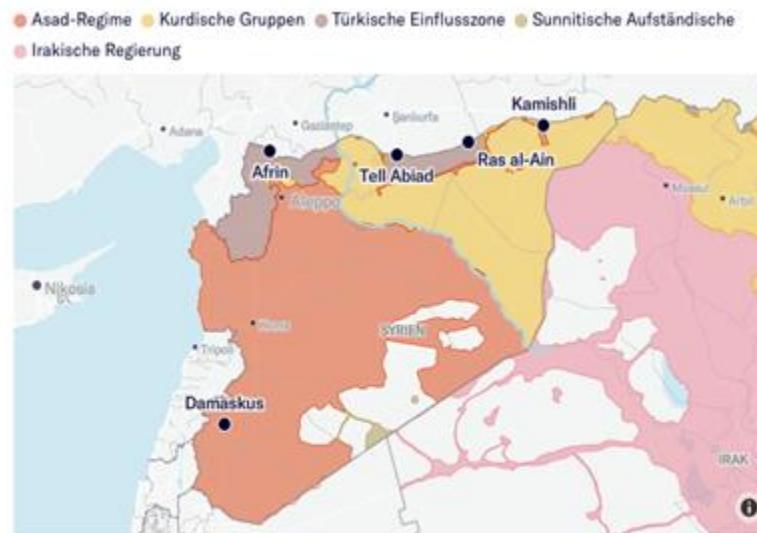
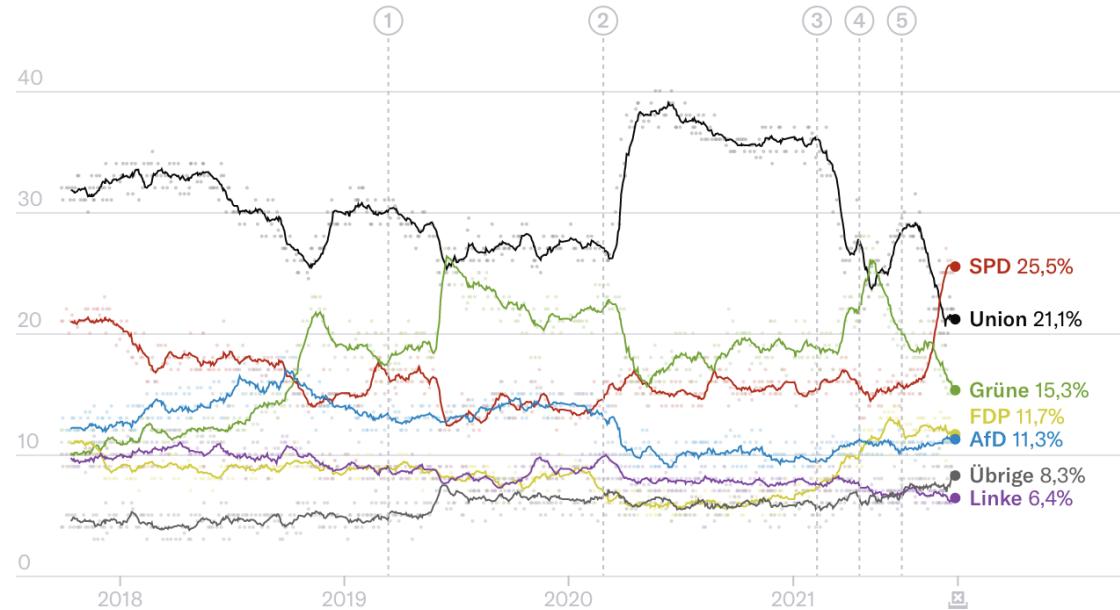
Es wurden nur OECD-Länder mit vollständigen und aktuellen Daten berücksichtigt.

Quelle: [OECD](#)

NZZ / nth.

NZZ / nth., jok.

Wiederkehrende Aktualität

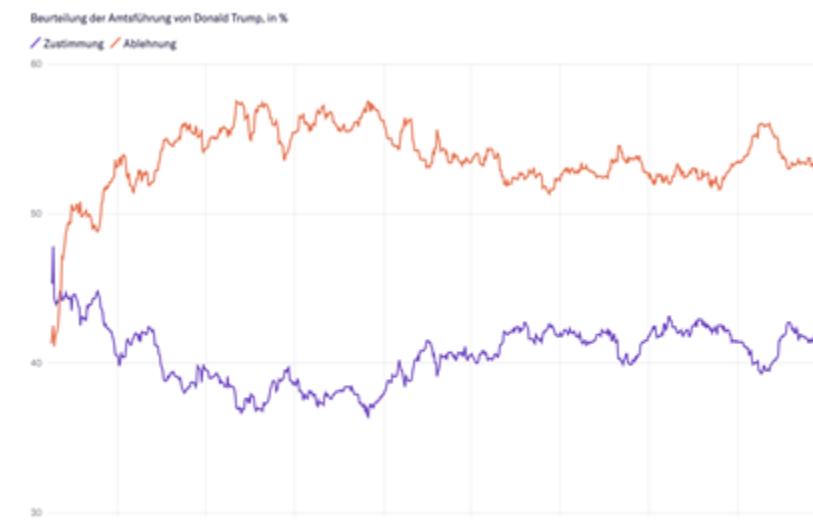


NZZ / iro.

```

15 # read in the data
16 datafile<-read.csv("https://projects.fivethirtyeight.com/trump-approval-data/approval_topline.csv", header=T, sep=",")
17
18 #if fetching the file fails, try downloading it manually from here https://projects.fivethirtyeight.com/trump-approval-ratings/
19 #and load it from the project folder
20 #datafile<-read.csv(paste0(getwd(),"/data/approval_topline.csv"), header=T, sep=",")
21
22
23 # get overview of data
24 summary(datafile)
25 str(datafile)
26 head(datafile)
27 tail(datafile)
28
29 #####COMPUTE#####
30
31 allpolls <- filter(datafile, subgroup == "All polls") #get rid of subsets (all adults, all voters)
32
33 allpolls_small <- allpolls[,c(3,4,7)] #drop columns we don't need (min+max estimates)
34
35 colnames(allpolls_small) <- c("Datum", "Zustimmung", "Ablehnung")
36
37 allpolls_small$Zustimmung <- round(allpolls_small$Zustimmung, digits = 1)
38 allpolls_small$Ablehnung <- round(allpolls_small$Ablehnung, digits = 1)
39
40 allpolls_small

```



Die Kurven zeigen ein gewichtetes Mittel aller Meinungsumfragen. Umfragen sind immer mit Unsicherheit behaftet. Die Fehlermarge variiert je nach Zeitpunkt, liegt aber jeweils bei rund +/- 5 Prozentpunkten. Stand: 27. März 2019. - Quelle: [FiveThirtyEight](https://fivethirtyeight.com) - Grafik: dca

Wiederkehrende Aktualität

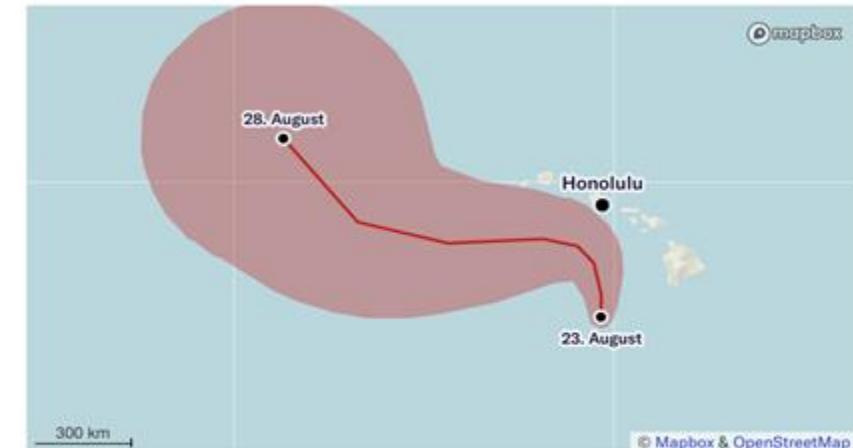
Wahrscheinlichkeit, dass Windgeschwindigkeiten von mehr als 63 km/h, also tropische Stürme, auftreten

● >90% ● ● ● ● ● ● ● ● ● ● <5% ↗ Verlauf (Prognose: gepunktet)

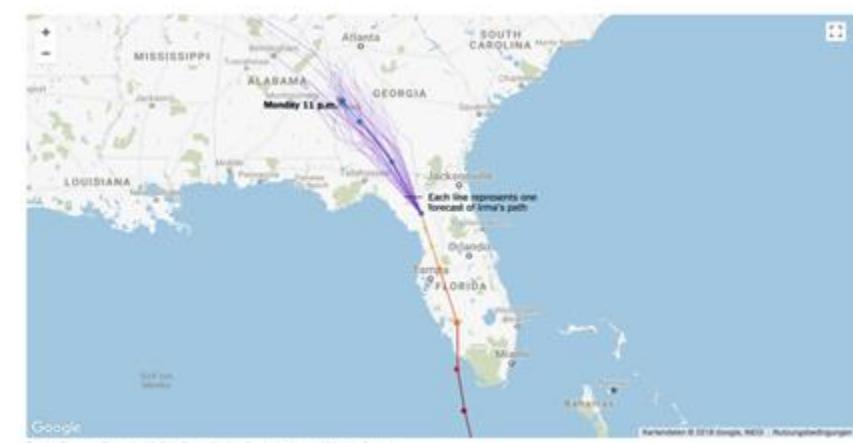


Stand: 24. 8. 2018. – Quelle: [National Hurricane Center / National Oceanic and Atmospheric Administration](#) – Grafik: mjk., awi.

● Wahrscheinlich betroffenes Gebiet (*) ↗ Prognose des Verlaufs bis 28. August



(*) Stand: 24. 8. 2018. Zahlen aus der Vergangenheit legen nahe, dass das Zentrum des Sturms mit einer Wahrscheinlichkeit von 66% in diesem Gebiet liegen wird. Das heisst: Das Risiko ist auch in Gebieten ausserhalb des Kegels hoch. – Quelle: [National Hurricane Center / National Oceanic and Atmospheric Administration](#) – Grafik: mjk.

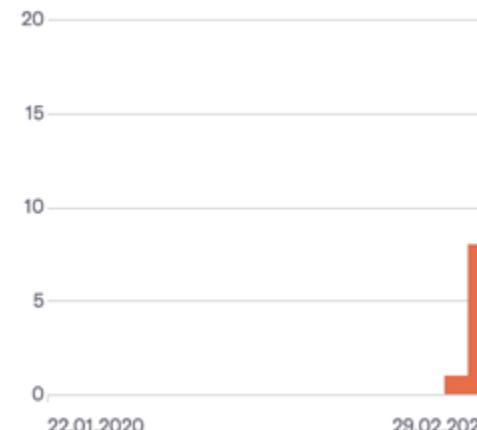


Hurrikan-Visualisierung der [«New York Times»](#), 5. 9. 2017.

Corona...

Zahl der Fälle des Coronavirus in der Schweiz, nach Status der Patienten

● Tote ● Erkrankte ● Geheilte



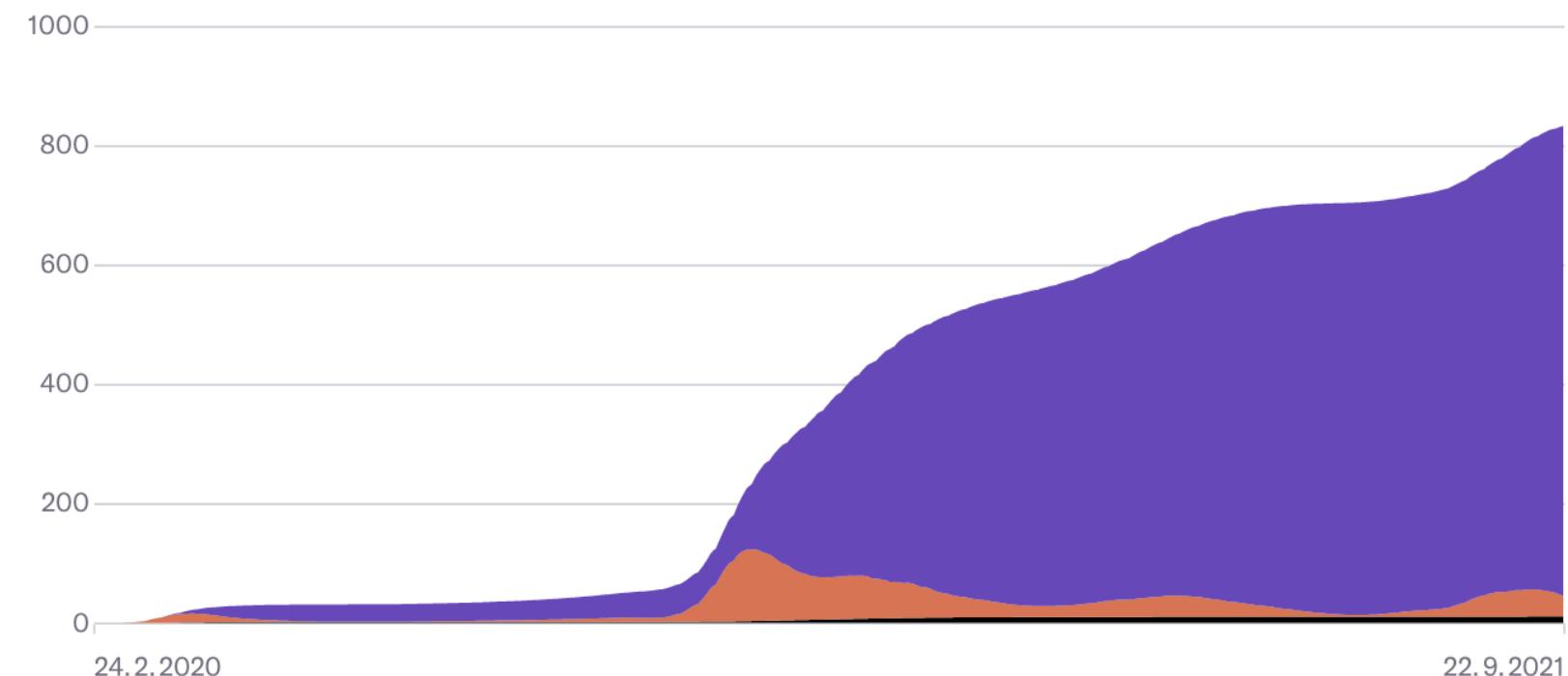
Quelle: [Johns-Hopkins-Universität](#)

NZZ / nth.

Über 831 000 bestätigte Infektionen in der Schweiz

Bestätigte Coronavirus-Fälle in der Schweiz und in Liechtenstein (in Tausend)

● Tote ● gegenwärtig Infizierte ● Genesene (Schätzung)



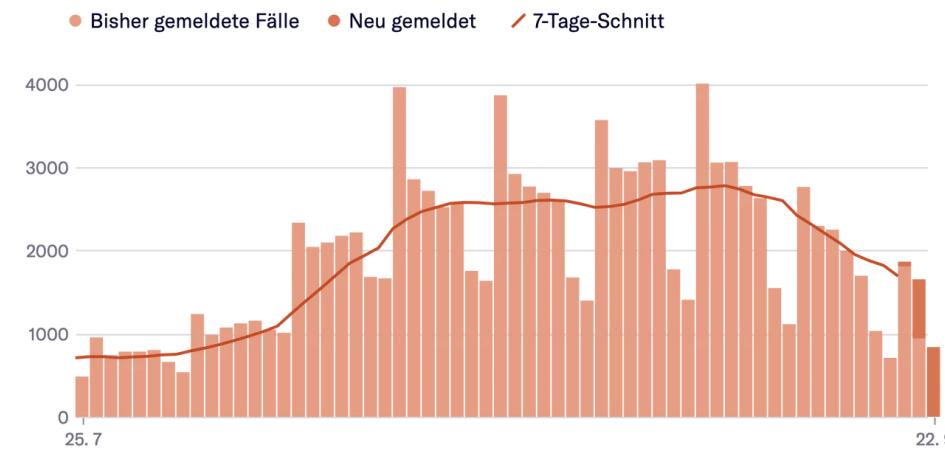
Die Zahlen der letzten zwei Tage sind als provisorisch zu betrachten, da die Meldungen teilweise verzögert eintreffen.
Die Zahl der Genesenen basiert auf einer Schätzung, siehe Quellen.

Quellen: [BAG](#), [eigene Berechnungen](#)

NZZ / nth.

Die Fallzahlen sinken wieder

Bestätigte neue Coronavirus-Fälle pro Tag in der Schweiz und in Liechtenstein, nach Testdatum

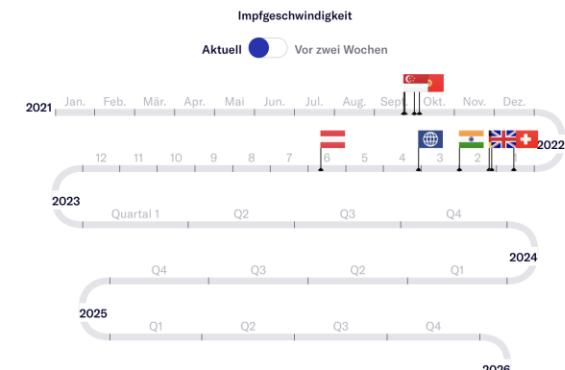
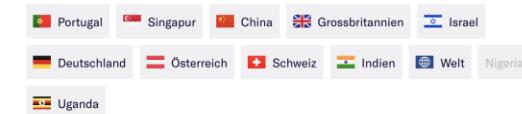


Diese Zahlen beziehen sich nur auf Fälle, die durch einen Test bestätigt wurden. Die Zahl der effektiven Neuinfektionen in der Bevölkerung ist nicht bekannt, dürfte aber weit höher liegen.

Quelle: BAG

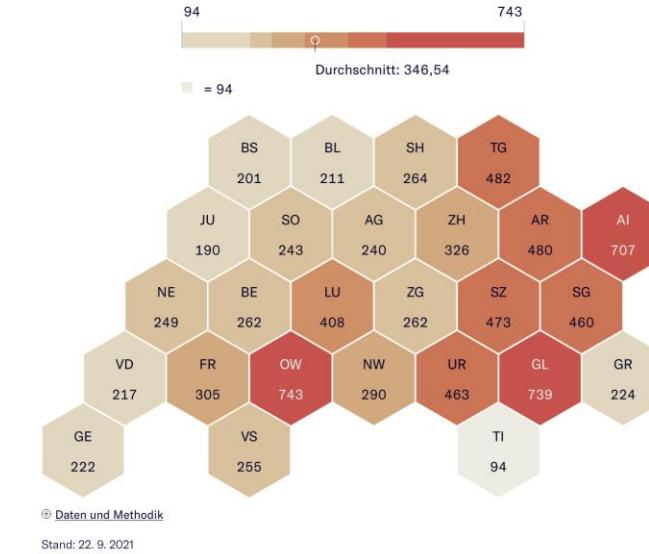
Wie bald die Länder eine Impfquote von 80 Prozent erreichen

Anhand der bisherigen Impfgeschwindigkeit* geschätzter Zeitpunkt der Erreichung einer Impfquote von 80 Prozent, ausgewählte Länder



Obwalden zählt pro Kopf am meisten Neuinfektionen

Sars-CoV-2-Fälle pro 100 000 Einwohner in den letzten zwei Wochen (Inzidenz), nach Kanton



Die Zahl der neuen Impfungen sinkt nach kurzem Anstieg wieder

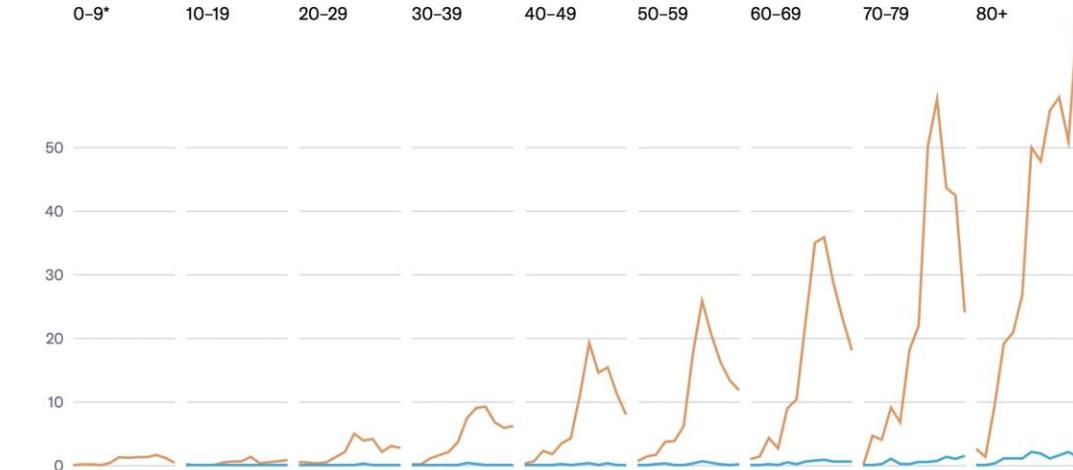
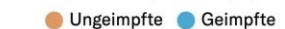
7-Tage-Schnitt der Zahl der täglich verabreichten Impfdosen in der Schweiz und in Liechtenstein



Quelle: BAG

In allen Altersgruppen landen Ungeimpfte deutlich häufiger im Spital

Hospitalisationen pro 100 000 geimpfte beziehungsweise ungeimpfte Einwohner in der Schweiz und Liechtenstein, nach Altersgruppe und Kalenderwoche



Automatisierung von Updates

✓ Update Corona Charts	Update Corona Charts #3084: Scheduled	📅 26 minutes ago	...
✓ Update Corona Charts	Update Corona Charts #3083: Scheduled	📅 1 hour ago	...
✓ Update Corona Charts	Update Corona Charts #3082: Scheduled	📅 1 hour ago	...
✓ Update Corona Charts	Update Corona Charts #3081: Scheduled	📅 2 hours ago	...
✓ Update Corona Charts	Update Corona Charts #3080: Scheduled	📅 2 hours ago	...
✓ Update Corona Charts	Update Corona Charts #3079: Scheduled	📅 3 hours ago	...
✓ Update Corona Charts	Update Corona Charts #3078: Scheduled	📅 3 hours ago	...
✓ Update Corona Charts	Update Corona Charts #3077: Scheduled	📅 18 hours ago	...

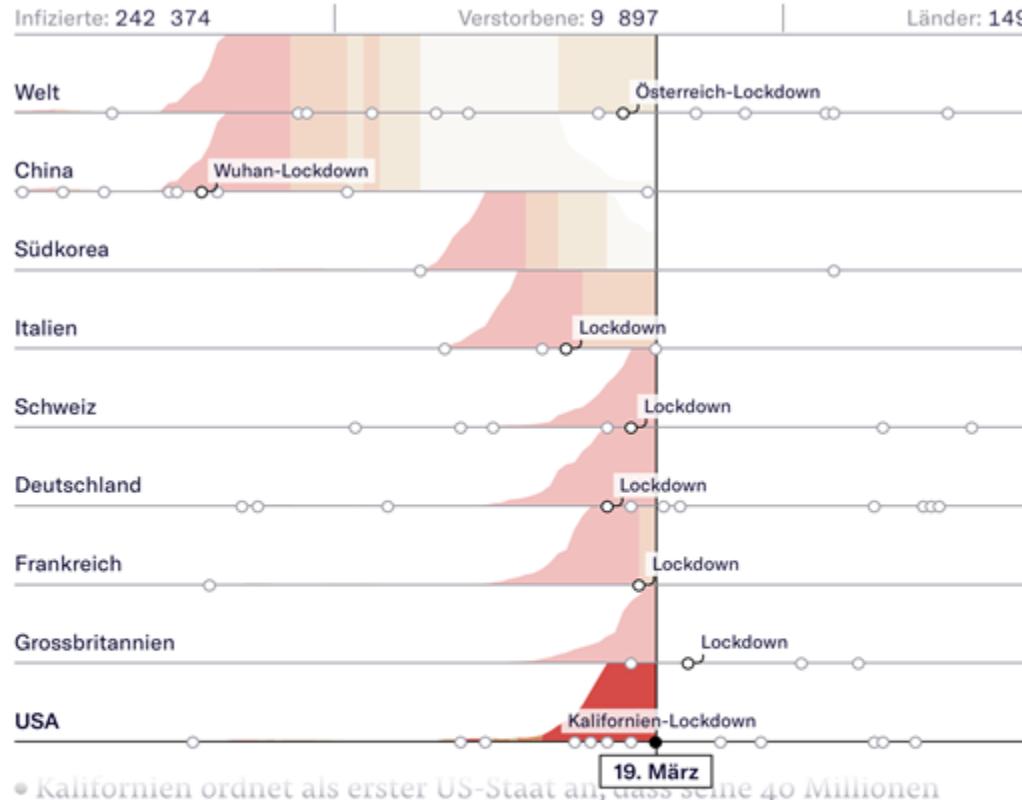
```

build
succeeded 23 minutes ago in 3m 43s

> ✓ Set up job
> ✓ Run actions/checkout@v2
> ✓ Setup R
> ✓ Install system dependencies
> ✓ Cache R packages
> ✓ Restore R packages
> ✓ Run R scripts
> ✓ Setup python
> ✓ Get pip cache dir
> ✓ Cache python packages
> ✓ Install python packages
> ✓ Run python scripts
> ✓ Setup Node.js
> ✓ Install node packages
> ✓ Run Q cli
> ✓ Post Setup Node.js
> ✓ Post Cache python packages
> ✓ Post Cache R packages
> ✓ Post Run actions/checkout@v2
> ✓ Complete job

```

Spin-offs und Meta-Stories



Die perfekte Corona-Grafik gibt es nicht

In der Corona-Krise sind Datenvisualisierungen so wichtig wie selten zuvor. Die richtige Darstellungsform für die Entwicklung der Pandemie zu finden, ist jedoch gar nicht so einfach.

Nikolai Thelitz, Kaspar Manz
08.04.2020, 14.21 Uhr

Hören Merken Drucken Teilen

Wie und mit welchen Daten wir die weltweite Ausbreitung des Coronavirus zeigen

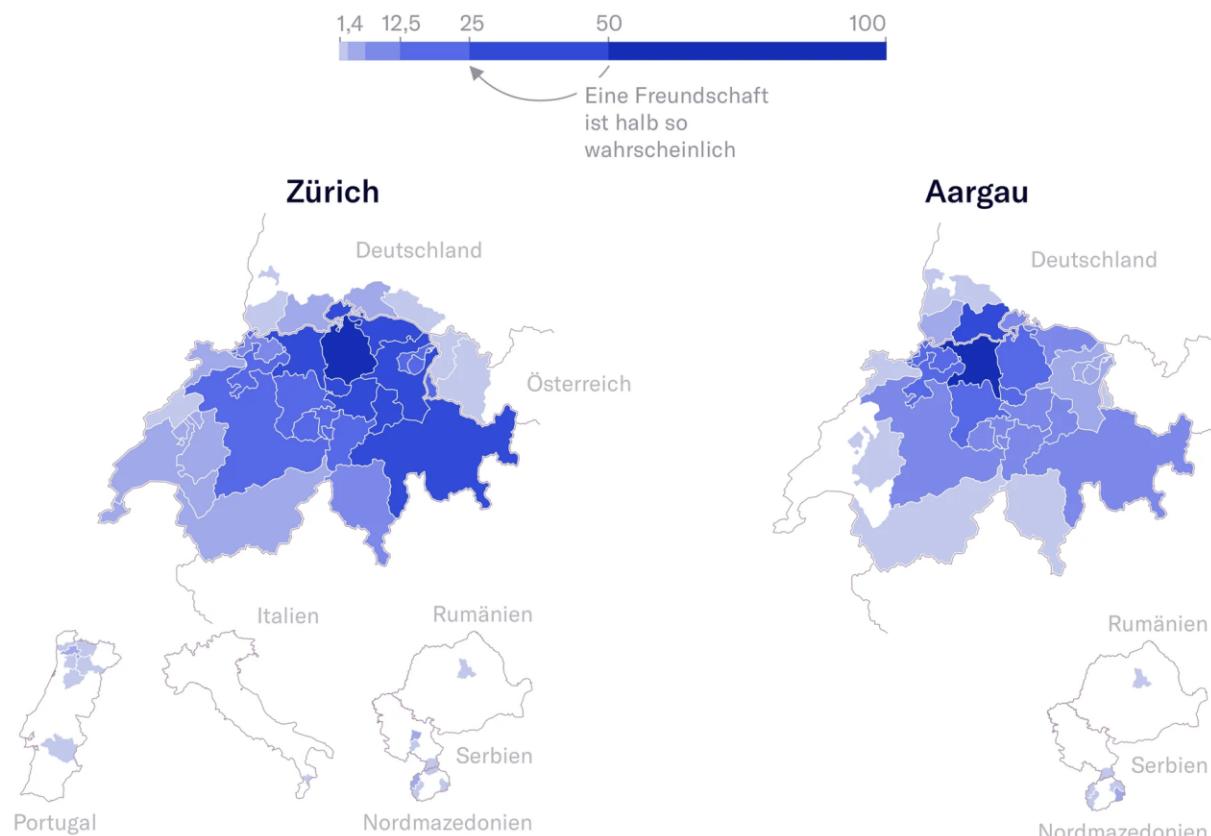
Die grafische Darstellung der Coronavirus-Zahlen birgt einige Herausforderungen. Welche Daten nutzt die NZZ, um die Ausbreitung des Coronavirus zu analysieren und zu zeigen?

Alexandra Kohler, Nikolai Thelitz,
Barnaby Skinner
04.06.2020, 15.35 Uhr

Hören Merken Drucken Teilen

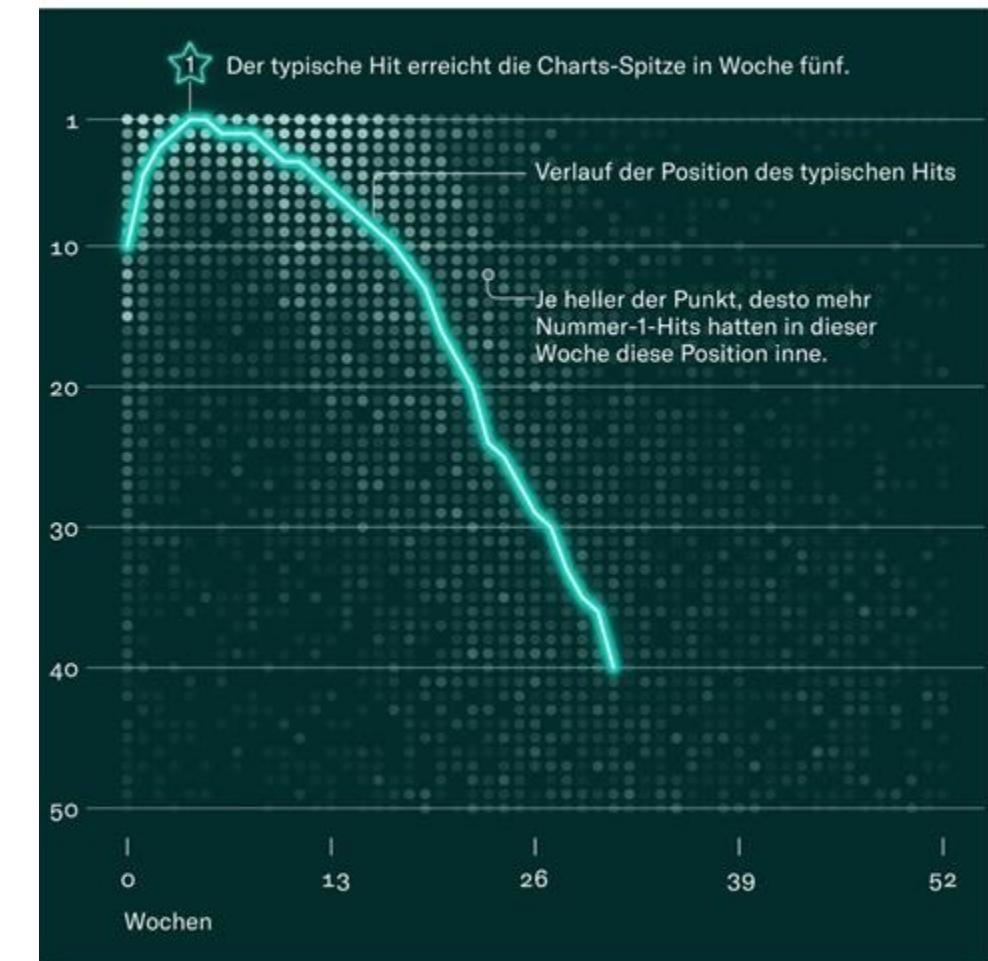
Grössere Recherchen

Die Deutschschweiz: Zürich ist mit der ganzen Schweiz befreundet, das Aargau mit den deutschen Nachbarn



Zwei Wochen auf Platz 1

Mittlere Position eines Nummer-1-Hits in der Schweizer Hitparade, in Wochen seit der ersten Charts-Platzierung

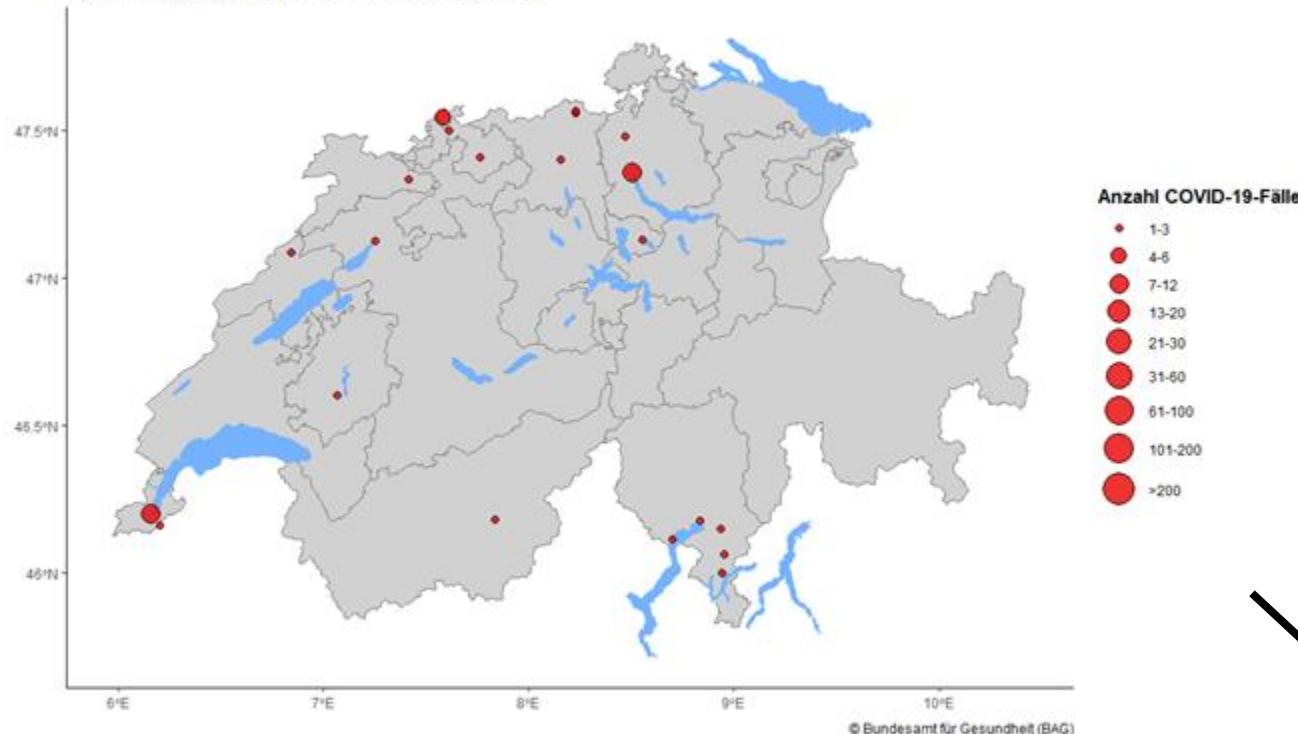


Der Medianwert wird nur bis zur 31. Woche angegeben, da nach diesem Zeitpunkt mehr als drei Viertel aller Nummer-1-Hits aus der Hitparade verschwunden sind.
Quelle: Hitparade.ch / eigene Berechnungen

NZZ / nth..jok.

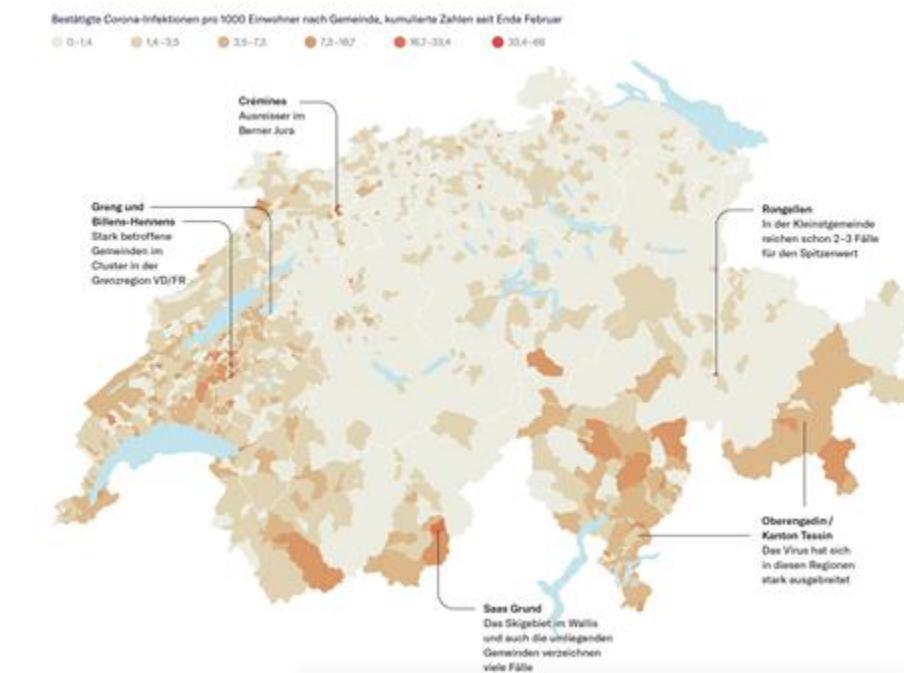
Grössere Recherchen

Laborbestätigte COVID-19-Fälle für die Woche 9



Mit anderen Worten: Das neue Kompetenzzentrum für Datenwissenschaften des BfS funktioniert als Schnittstelle nicht nur für den Datentausch innerhalb der Verwaltung, sondern auch zur Öffentlichkeit oder zur Privatwirtschaft?

Kuonen: Ich möchte hier die Daten-Diskussion abbrechen. Es geht um den Einsatz innovativer datenwissenschaftlicher Methoden. Daten sind das Mittel zum Zweck. Um nochmals Ihr Beispiel zu nennen: Das BAG muss die Daten gar nicht schicken, die bleiben schön bei ihm. Es geht darum, das Amt mithilfe datenwissenschaftlicher Ansätze darin zu befähigen, die entwickelten Algorithmen auf seine Daten in seinem Haus anzuwenden.



Übung: Thema finden und Forschungsfrage formulieren

Tipps:

- Google Scholar für wissensch. Studien
- Statista ist verboten ;)
- In Medienberichten nach der Quelle suchen und diese dann googlen