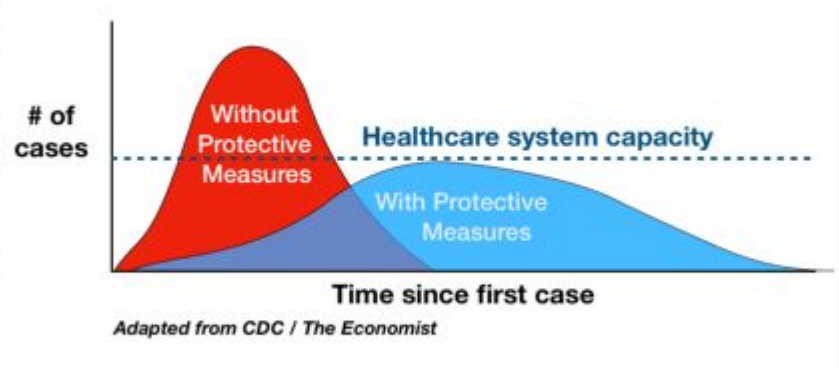
BEST PRACTICES FOR DATA VISUALIZATION

Jess Cohen-Tanugi, Visualization Specialist, Harvard Library

The chart that went around the world



Graph by Drew Harris.

HOW TO DESIGN EFFECTIVE VISUALIZATIONS

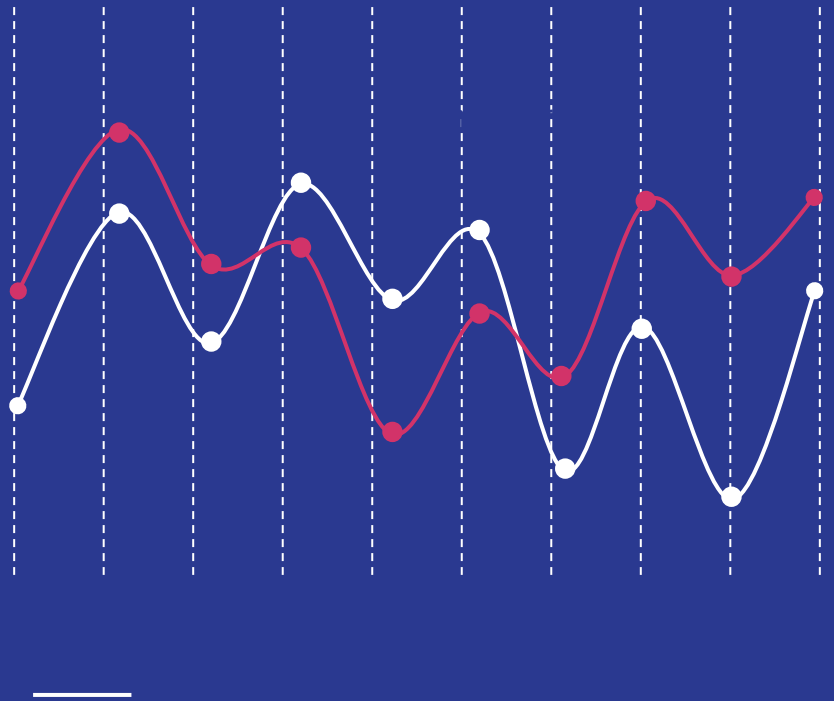
Have effective data in a good format.

University Name	State	2019	2018	2017
Princeton University	NJ	1	1	1
Harvard University	MA	2	2	2
Yale University	CT	3	3	3
University of Chicago	IL	3	3	3
Columbia University	NY	3	5	5
Massachusetts Institut	MA	3	5	7
Stanford University	CA	7	5	5
University of Pennsylva	PA	8	8	8
Duke University	NC	8	9	8
Johns Hopkins Univers	MD	10	11	10
Northwestern Univers	IL	10	11	12
California Institute of	CA	12	10	12
Dartmouth College	NH	12	11	11
Brown University	RI	14	14	14
Vanderbilt University	TN	14	14	15
Cornell University	NY	16	14	15

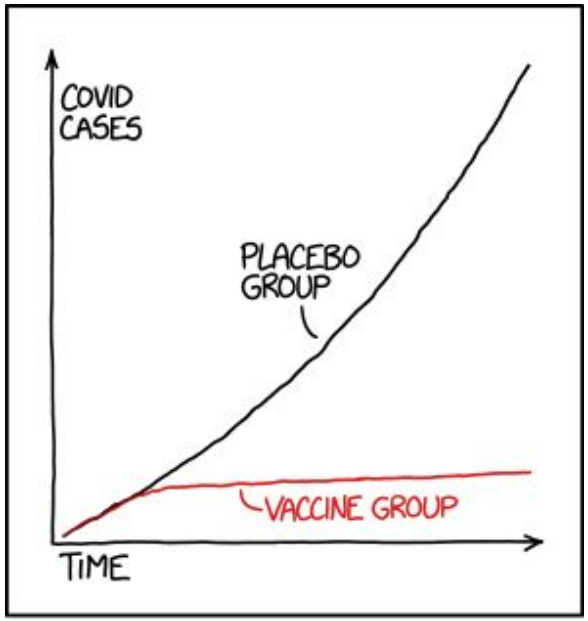
University Name	State	Year	Rank
Columbia University	NY	1988	18
Columbia University	NY	1989	8
Columbia University	NY	1990	11
Columbia University	NY	1991	10
Columbia University	NY	1992	9
Columbia University	NY	1993	10
Columbia University	NY	1994	11
Columbia University	NY	1995	9
Columbia University	NY	1996	15
Columbia University	NY	1997	11
Columbia University	NY	1998	9
Columbia University	NY	1999	10
Columbia University	NY	2000	10
Columbia University	NY	2001	10
Columbia University	NY	2002	9

Price	Living.Area	Bathrooms	Bedrooms	
142212	1982	1	3	0	2	133
134865	1676	1.5	3	1	0.38	14
118007	1694	2	3	1	0.96	15
138297	1800	1	2	2	0.48	49
129470	2088	1	3	1	1.84	29
206512	1456	2	3	0	0.98	10
50709	960	1.5	2	0	NA	12
108794	1464	1	2	0	0.11	87
68353	1216	1	2	0	0.61	101
123266	1632	1.5	3	0	0.23	14
309808	2270	2.5	3	2	4.05	9
157946	1804	2.5	3	1	0.43	0
80248	1600	1.5	3	0	0.36	16
135708	1460	2	2	0	0.18	17
173723	1548	2	3	1	0.36	0
140510	1590	2.5	3	1	0.42	0
122221	1170	1.5	4	0	3	26
151917	1510	2.5	3	1	0.39	0
235105	2299	2.5	4	1	0.8	6
259999	2577	2.5	4	1	0.77	1
211517	2328	2.5	4	1	0.85	10
102068	1172	2.5	3	1	0.85	73
128440	1554	1.5	3	0	4.87	103
115659	1242	2	3	1	0.72	30
145583	1376	2	3	1	0.46	25
116289	1107	1	3	1	0.46	43
238792	2250	2.5	4	1	2.48	10
221925	2472	2	4	0	0.62	183
310696	2843	2.5	4	1	0.71	5
139079	1400	1.5	3	1	1	35
109578	1342	1	2	1	0.57	41
65325	813	1	2	0	0.39	68
89893	1480	2.5	2	1	0.29	14
87588	1392	1	2	0	0.24	17
132311	1512	1.5	3	1	0.25	13
131411	1512	1.5	3	1	0.05	13
158863	1696	2.5	3	1	0.71	5
130490	1595	2	3	1	0.22	6
88207	1480	1.5	3	0	0.07	14
178767	2291	2.5	4	1	0.1	0
148246	1391	2	3	0	0.04	12
205073	2501	2.5	4	0	0.07	0
185323	1662	2	3	0	0.5	0
71904	957	1	3	0	0.19	45
82556	1480	1.5	3	1	0.1	14

Explore your data first!



Choose a message, and compose a visualization.



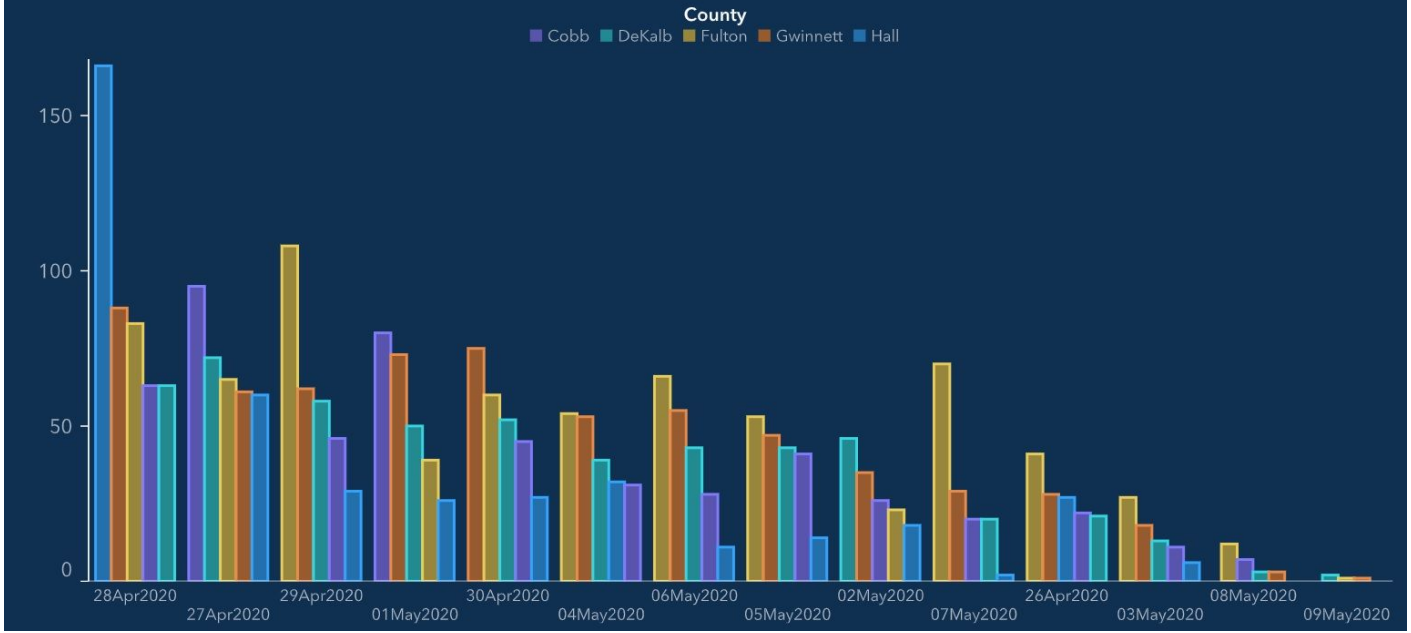
STATISTICS TIP: ALWAYS TRY TO GET DATA THAT'S GOOD ENOUGH THAT YOU DON'T NEED TO DO STATISTICS ON IT

Your visualization is probably part of a narrative, and will have a message you can summarize in words.

You can (and probably will) tell many different stories with the same data.

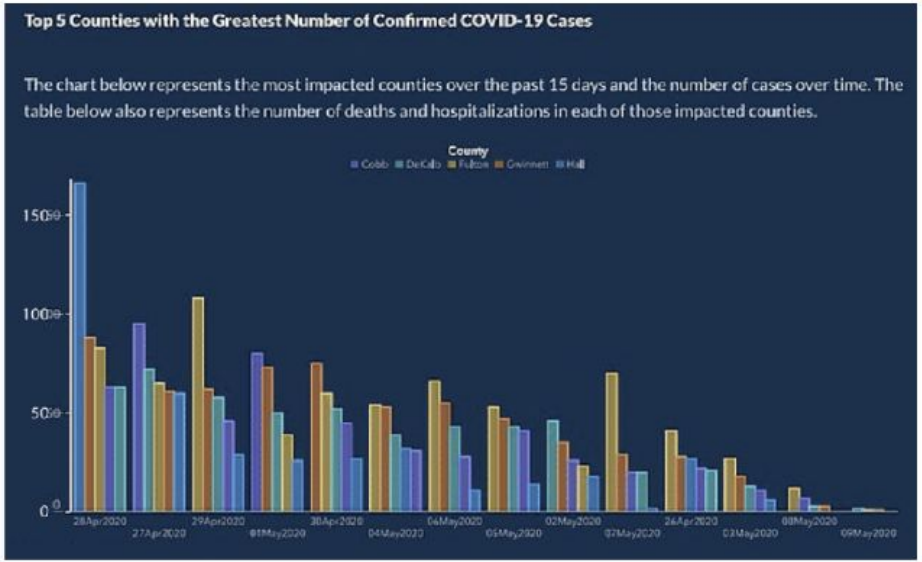
Top 5 Counties with the Greatest Number of Confirmed COVID-19 Cases

The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.

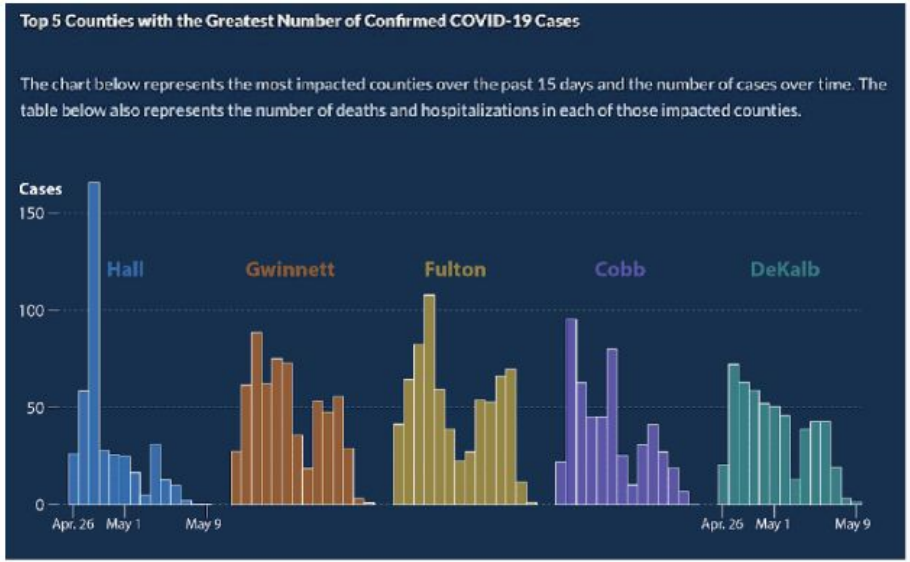


Georgia Department of Health

ORIGINAL



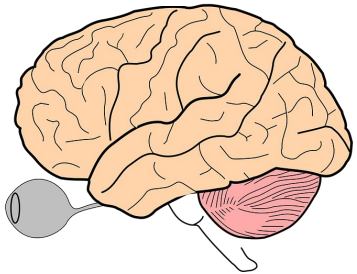
MAKEOVER



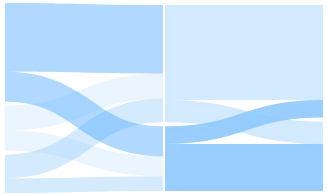
Fixed graph by Alberto Cairo: https://ijnet.org/en/story/tips-visualizing-covid-19-data

PSYCHOLOGY OF VISUAL PROCESSING

How does the brain process visual information?



+

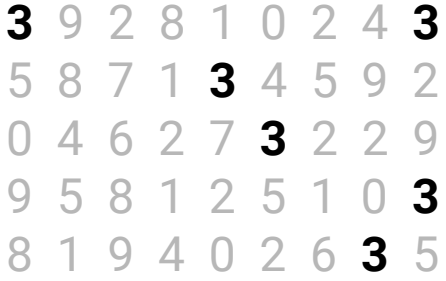


How many 3s can you find?

3 9 2 8 1 0 2 4 3
5 8 7 1 3 4 5 9 2
0 4 6 2 7 3 2 2 9
9 4 8 2 7 5 1 0 3
8 1 9 4 0 2 6 3 5

How many 3s can you find?

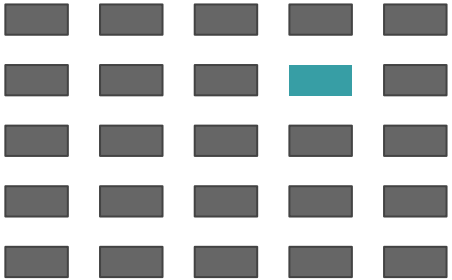
3 9 2 8 1 0 2 4 3
5 8 7 1 3 4 5 9 2
0 4 6 2 7 3 2 2 9
9 4 8 2 7 5 1 0 3
8 1 9 4 0 2 6 3 5

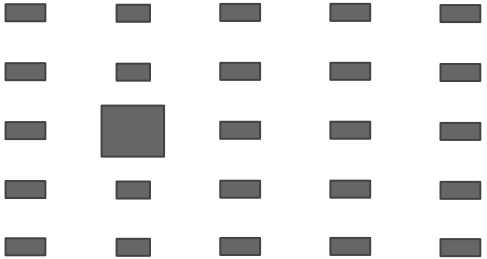


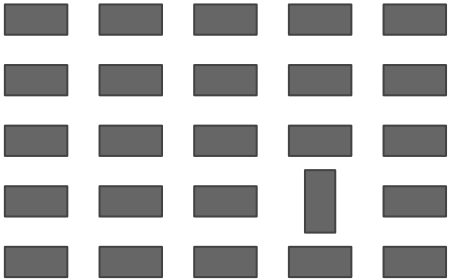
Preattentive visual cues

3 9 2 8 1 0 2 4 3
5 8 7 1 3 4 5 9 2
0 4 6 2 7 3 2 2 9
9 4 8 2 7 5 1 0 3
8 1 9 4 0 2 6 3 5

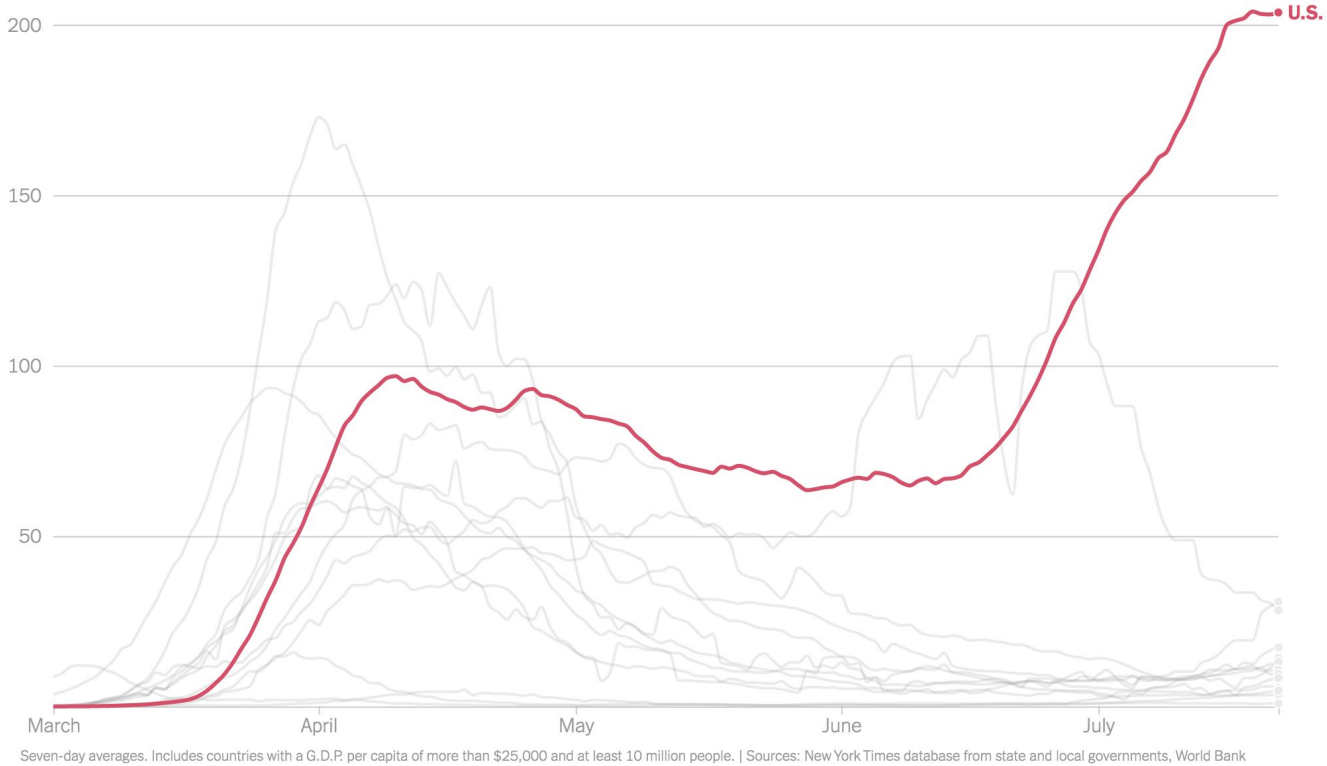
3 9 2 8 1 0 2 4 **3**
5 8 7 1 **3** 4 5 9 2
0 4 6 2 7 **3** 2 2 9
9 5 8 1 2 5 1 0 **3**
8 1 9 4 0 2 6 **3** 5





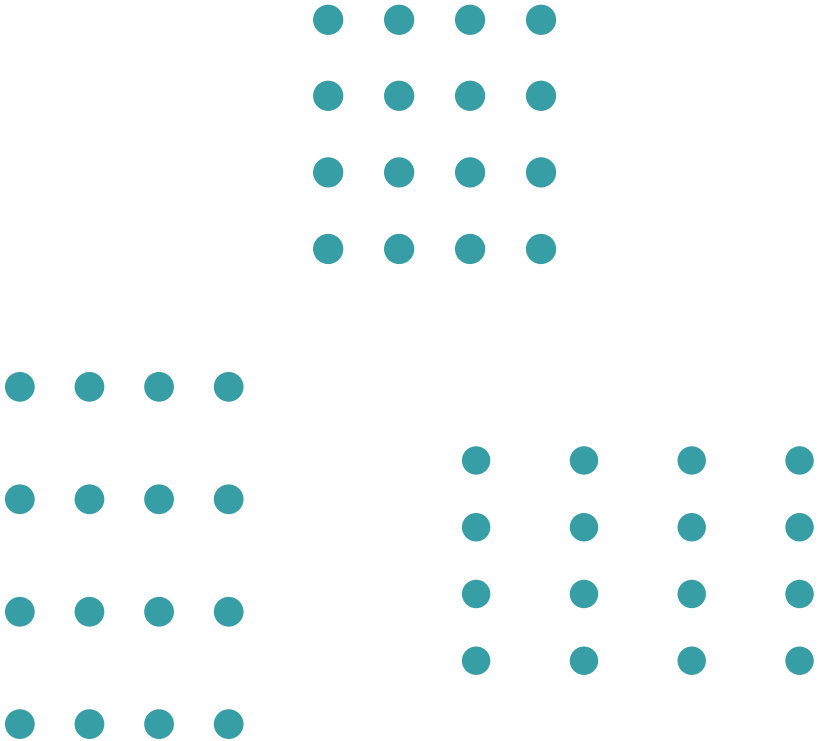


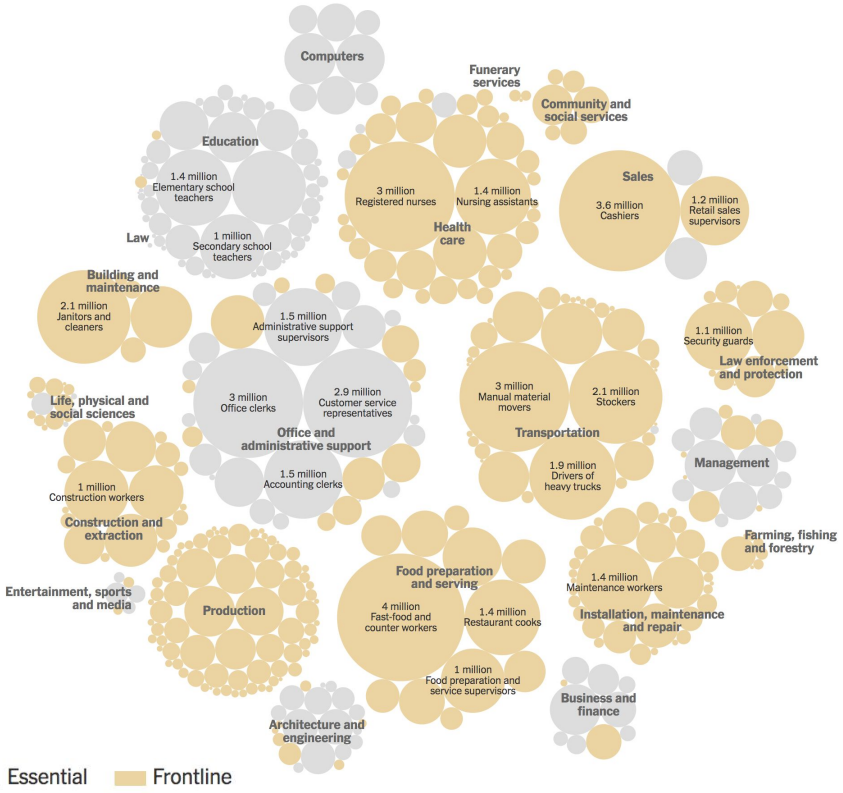
Example



Source: New York Times. https://www.nytimes.com/interactive/2020/07/23/us/coronavirus-hotspots-countries.html

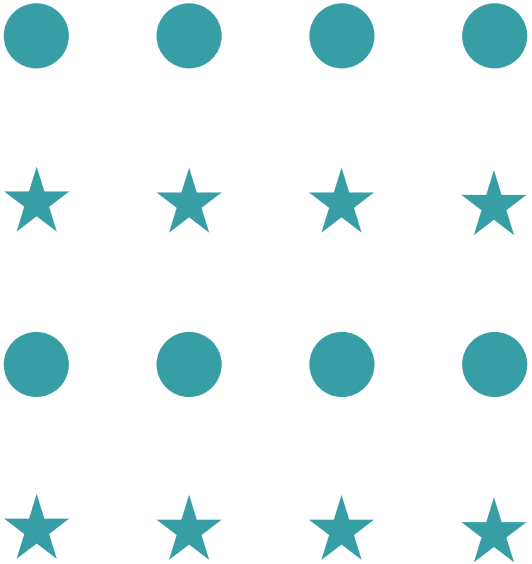
Gestalt principles: proximity

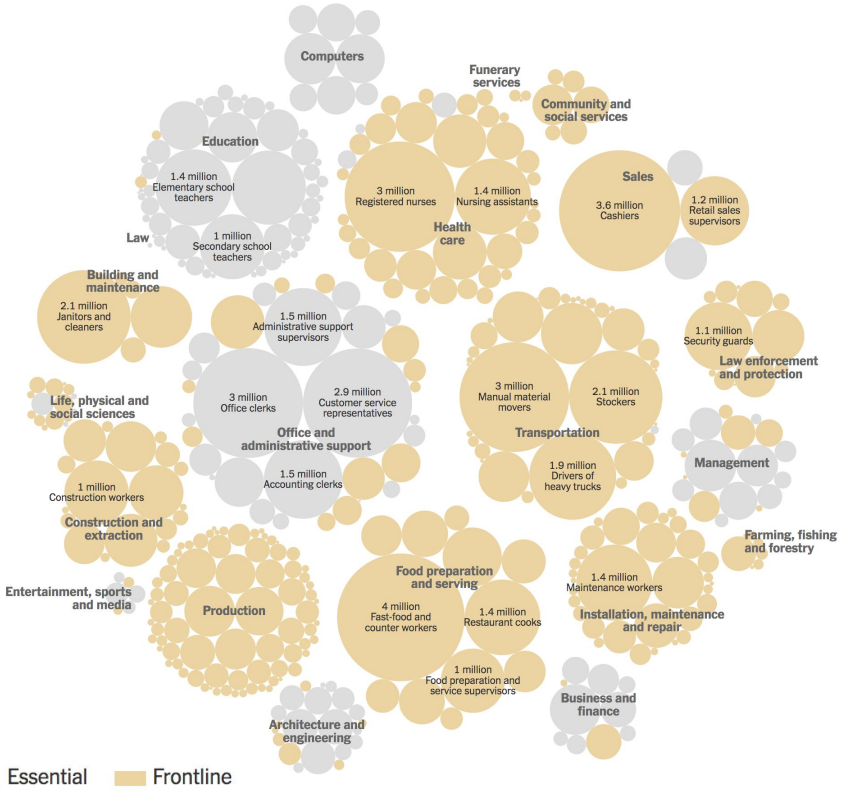




Source: New York Times. https://www.nytimes.com/2020/12/05/health/covid-vaccine-first.html

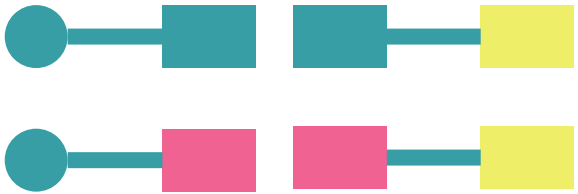
Gestalt principles: similarity





Source: New York Times. https://www.nytimes.com/2020/12/05/health/covid-vaccine-first.html

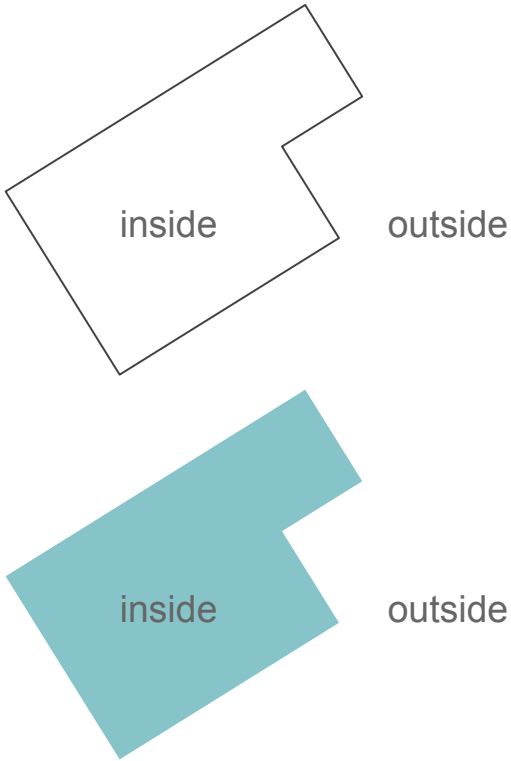
Gestalt principles: connectedness

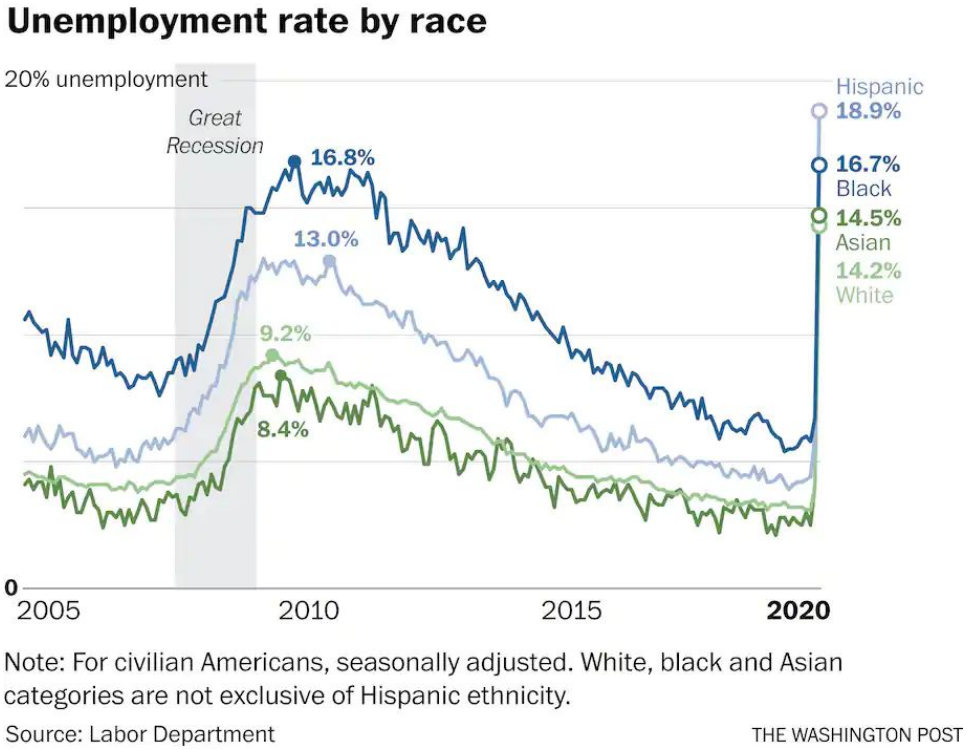


Top how, where, what and why internet searches for the week of April 5 to 11, for 2019 against 2020.

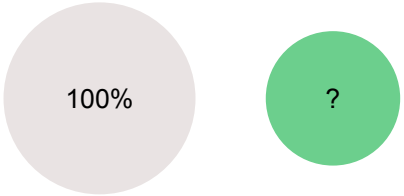
Source: Washington Post. https://flowingdata.com/2020/04/20/change-in-google-searches-since-the-virus/

Gestalt principles: common region

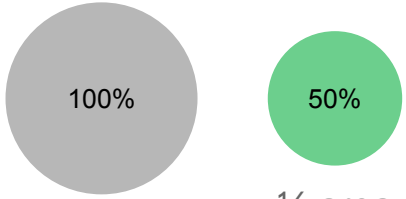




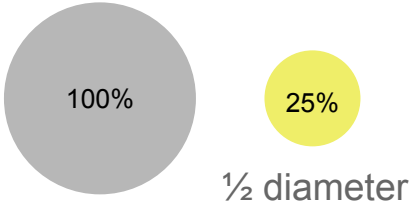
Source: Washington Post. https://www.washingtonpost.com/business/2020/05/08/april-2020-jobs-report/


100%
?

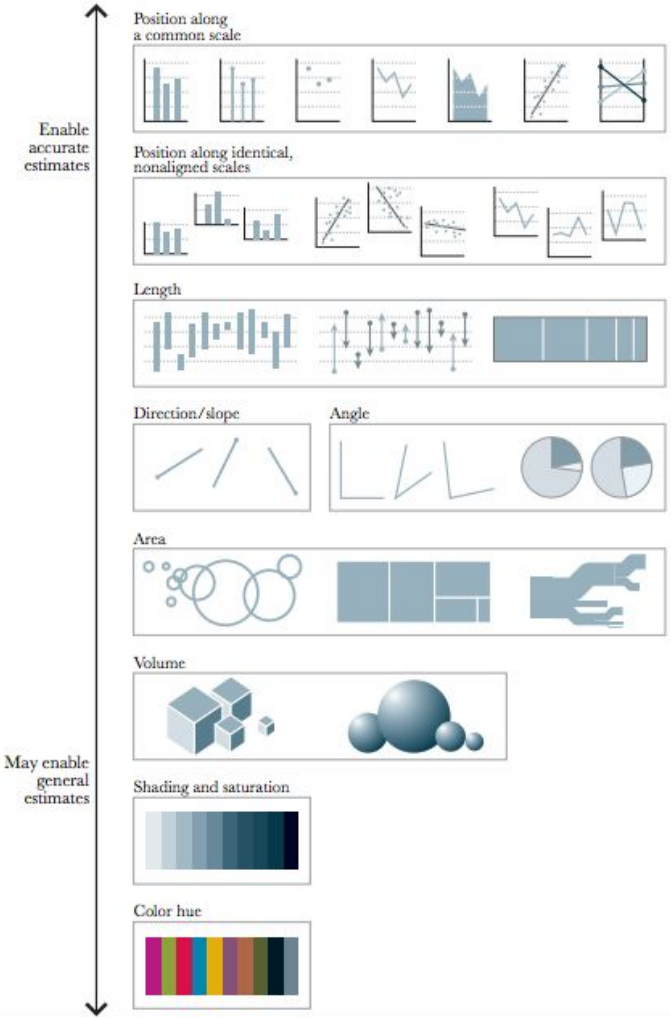

100%
?

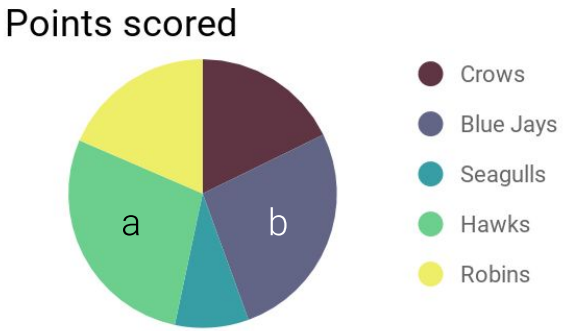


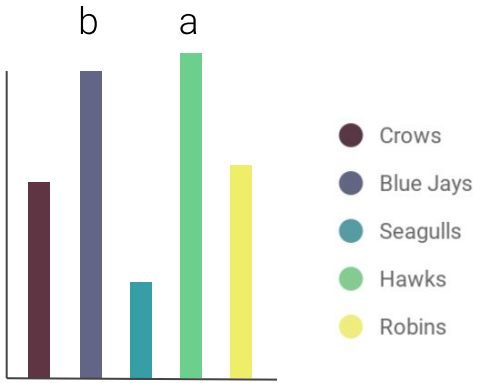
½ area



½ diameter

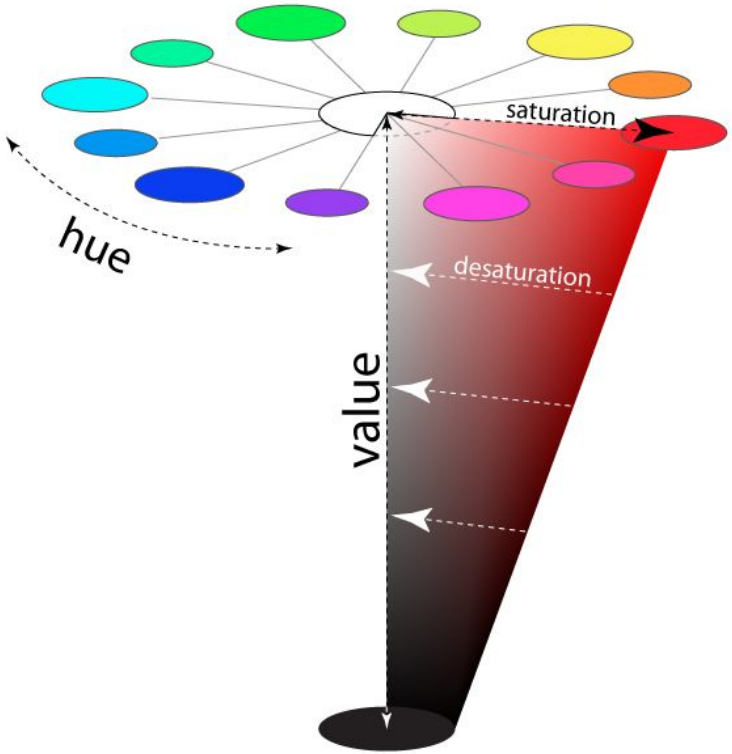






Estimates visual from Cairo, Alberto. *The Functional Art: An Introduction to Information Graphics and Visualization.* Berkeley, CA: New Riders, 2012.

Choose an appropriate color palette



For discrete or **qualitative** data:

Ex. “apples, bananas,” or “USA, UK, France.”

For continuous or **sequential** data:

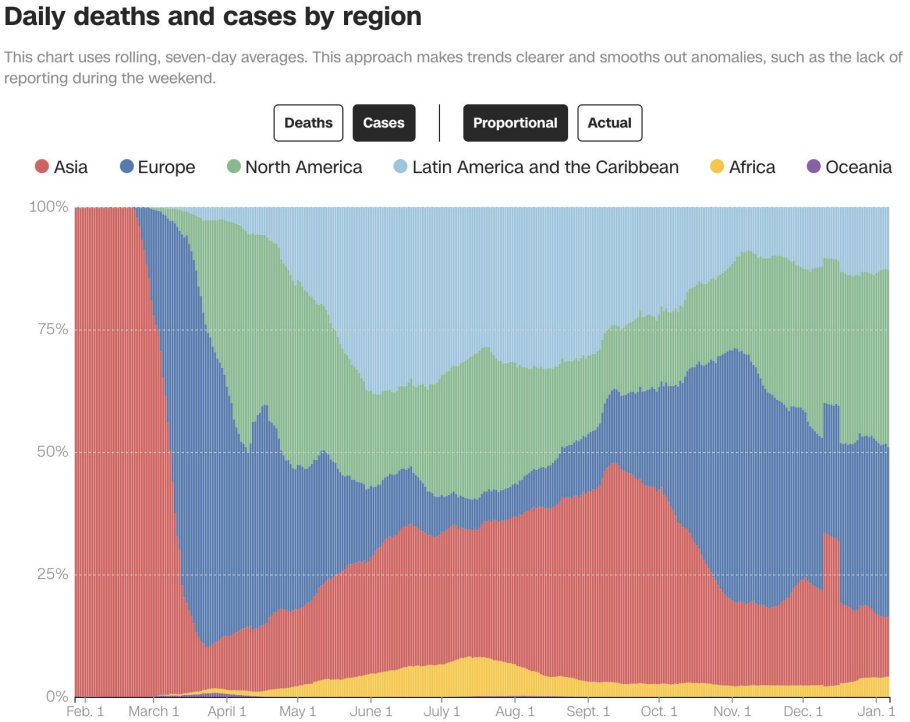
Ex. income, age data.

For **continuous** data diverging from a midpoint:

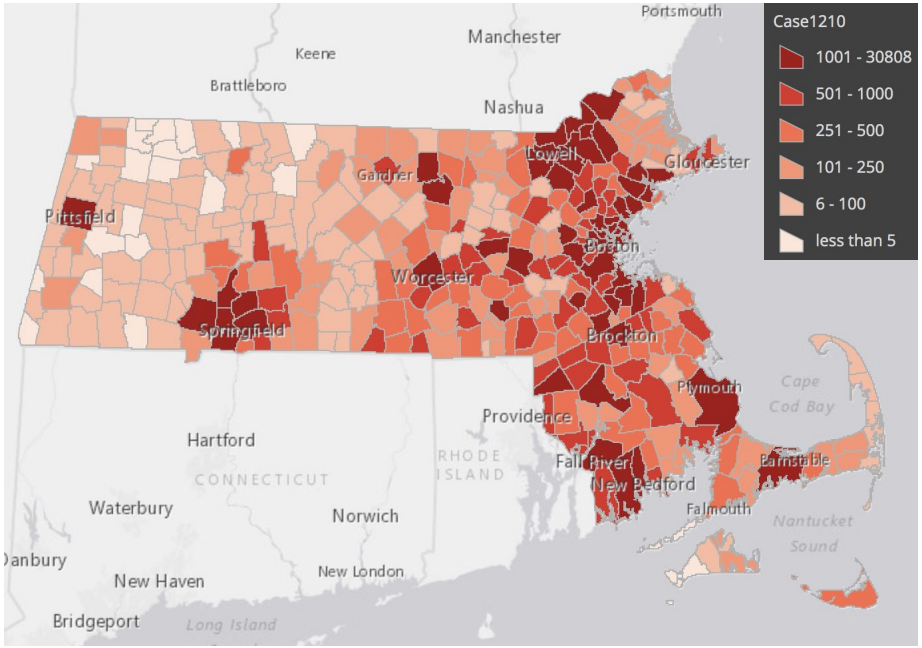
Ex. temperature or stock data.

Palettes from Ware, Colin. *Information Visualization*, Third Ed. San Francisco, CA: Morgan Kaufmann, 2012.

Examples:

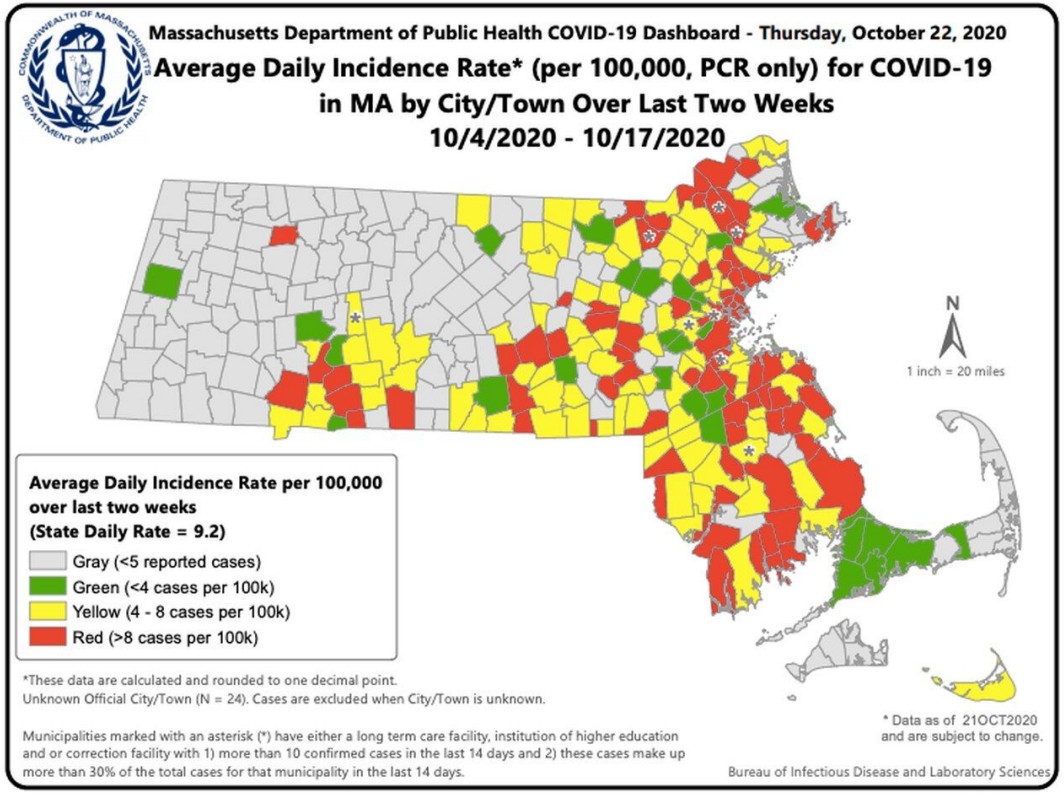


https://www.cnn.com/interactive/2020/health/coronavirus-maps-and-cases/

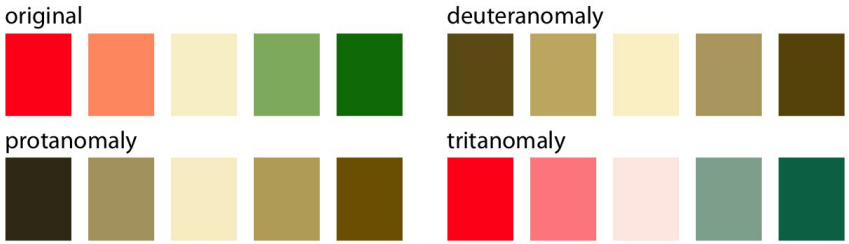


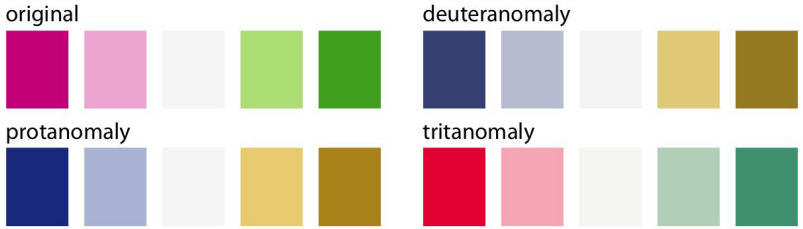
Boston University School of Public Health.
https://bucas.maps.arcgis.com/apps/MapSeries/index.html?appid=e820a92d6bbc4c9099c59494a4e9367a

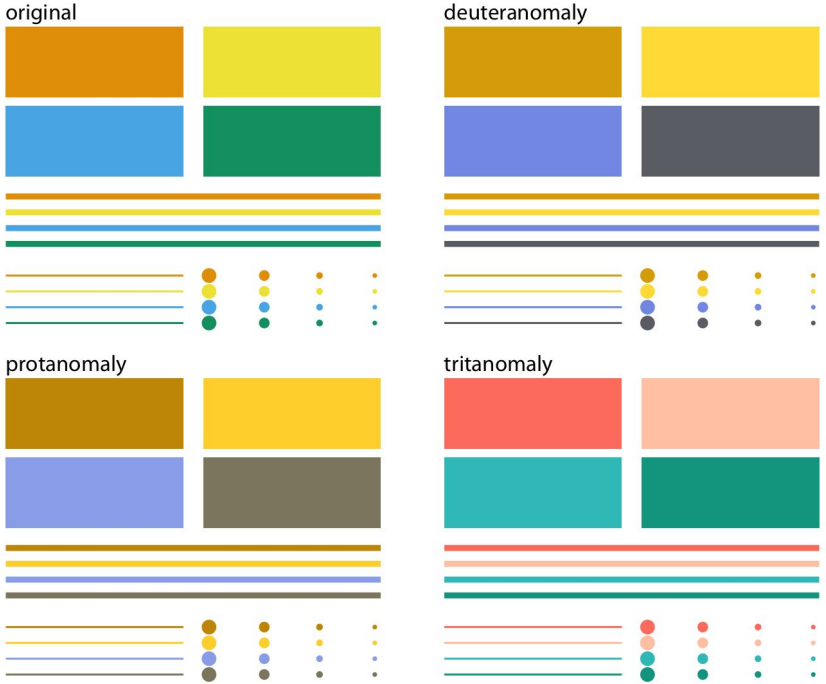
What not to do



Color blindness







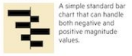
CHOOSING THE RIGHT CHART TYPE

CHOOSING THE RIGHT CHART TYPE

Deviation

Emphasise variations (+/-) from a fixed reference point. Typically the reference point is zero but it can also be a target or a long-term average. Can also be used to show sentiment (positive/neutral/negative).

Example FT uses
Trade surplus/deficit, climate change

Diverging bar

A simple standard bar chart that can handle both negative and positive magnitude values.

Diverging stacked bar

Perfect for presenting survey results which involve sentiment (eg disagree/neutral/agree).

Spine

Splits a single value into two contrasting components (eg male/female).

Surplus/deficit filled line

The shaded area of these charts allows a balance to be shown – either against a baseline or between two series.

Correlation

Show the relationship between two or more variables. Be mindful that, unless you tell them otherwise, many readers will assume the relationships you show them to be causal (ie, one causes the other).

Example FT uses
Inflation and unemployment, income and life expectancy

Scatterplot

The standard way to show the relationship between two continuous variables, each of which has its own axis.

Column + line timeline

A good way of showing the relationship between an amount (columns) and a rate (line).

Connected scatterplot

Usually used to show how the relationship between 2 variables has changed over time.

Bubble

Like a scatterplot, but adds additional detail by sizing the circles according to a third variable.

XY heatmap

A good way of showing the patterns between 2 categories of data, less effective at showing fine differences in amounts.

Ranking

Use where an item's position in an ordered list is more important than its absolute or relative value. Don't be afraid to highlight the points of interest.

Example FT uses
Wealth, deprivation, league tables, constituency election results

Ordered bar

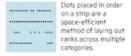
Standard bar charts display the ranks of values much more easily when sorted into order.

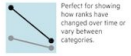
Ordered column

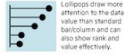
See above.

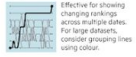
Ordered proportional symbol

Use when there are big variations between values and/or seeing fine differences between data is not so important.

Dot strip plot

Dots placed in order on a strip are a space-efficient method of laying out ranks across multiple categories.

Slope

Perfect for showing how ranks have changed over time or vary between categories.

Lollipop

Lollipops draw more attention to the data value than standard bar/column and can also show rank and value effectively.

Bump

Effective for showing changing rankings across multiple dates. For large datasets, consider grouping lines using colour.

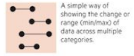
Distribution

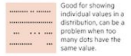
Show values in a dataset and how often they occur. The shape (or 'skew') of a distribution can be a memorable way of highlighting the lack of uniformity or equality in the data.

Example FT uses
Income distribution, population (age/sex) distribution, revealing inequality

Histogram

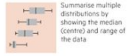
The standard way to show a statistical distribution - keep the gaps between columns small to highlight the 'shape' of the data.

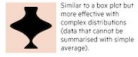
Dot plot

A simple way of showing the change or range (min/max) of data across multiple categories.

Dot strip plot

Good for showing individual values in a distribution, can be a problem when too many dots have the same value.

Barcode plot

Like dot strip plots, good for displaying all the data in a table, they work best when highlighting individual values.

Boxplot

Summarise multiple distributions by showing the median (centre) and range of the data

Violin plot

Similar to a box plot but more effective with complex distributions (data that cannot be summarised with simple average).

Population pyramid

A standard way for showing the age and sex breakdown of a population distribution, effectively, back to back histograms.

Cumulative curve

A good way of showing how unequal a distribution is: y axis is always cumulative frequency, x axis is always a measure.

Change over Time

Give emphasis to changing trends. These can be short (intra-day) movements or extended series traversing decades or centuries. Choosing the correct time period is important to provide suitable context for the reader.

Example FT uses
Share price movements, economic time series, sectoral changes in a market

Line

The standard way to show a changing time series. If data are irregular, consider markers to represent data points.

Column

Columns work well for showing change over time – but usually best with only one series of data at a time.

Column + line timeline

A good way of showing the relationship over time between an amount (columns) and a rate (line).

Slope

Good for showing changing data as long as the data can be simplified into 2 or 3 points without missing a key part of story.

Area chart

Use with care – these are good at showing changes to total, but seeing change in components can be very difficult.

Candlestick

Usually focused on day-to-day activity, these charts show opening/closing and high/low points of each day.

Fan chart (projections)

Use to show the uncertainty in future data points - usually this grows the further forward to projection.

Connected scatterplot

A good way of showing changing data for two variables whenever there is a relatively clear pattern of progression.

Magnitude

Show size comparisons. These can be relative (just being able to see larger/bigger) or absolute (need to see fine differences). Usually these show a 'counted' number (for example, barrels, dollars or people) rather than a calculated rate or per cent.

Example FT uses
Commodity production, market capitalisation, volumes in general

Column

The standard way to compare the size of things. Must always start at 0 on the axis.

Bar

See above. Good when the data are not time series and labels have long category names.

Paired column

As per standard column but allows for multiple series. Can become tricky to read with more than 2 series.

Paired bar

See above.

Marimekko

A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Proportional symbol

Use when there are big variations between values and/or seeing fine differences between data is not so important.

Isotype (pictogram)

Excellent solution in some instances – use only with whole numbers (do not slice off an arm to represent a decimal).

Lollipop

Lollipop charts draw more attention to the data value than standard bar/column – does not have to start at zero (but preferable).

Part-to-whole

Show how a single entity can be broken down into its component elements. If the reader's interest is solely in the size of the components, consider a magnitude-type chart instead.

Example FT uses
Fiscal budgets, company structures, national election results

Stacked column/bar

A simple way of showing part-to-whole relationships but can be difficult to read with more than a few components.

Marimekko

A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Pie

A common way of showing part-to-whole data – but be aware that it's difficult to accurately compare the size of the segments.

Donut

Similar to a pie chart – but the centre can be a good way of making space to include more information about the data (eg total).

Treemap

Use for hierarchical part-to-whole relationships; can be difficult to read when there are many small segments.

Voronoi

A way of turning points into areas – any point within each area is closer to the central point than any other centroid.

Arc

A hemicycle, often used for visualising parliamentary composition by number of seats.

Gridplot

Good for showing % information, they work best when used on small numbers and work well in small multiple layout form.

Spatial

Aside from locator maps only used when precise locations or geographical patterns in data are more important to the reader than anything else.

Example FT uses
Population density, natural resource locations, natural disaster risk/impact, catchment areas, variation in election results

Basic choropleth (rate/ratio)

The standard approach for putting data on a map – should always be rates rather than totals, and use a sensible base geography.

Proportional symbol (count/magnitude)

Use for totals rather than rates – be wary that small differences in data will be hard to see.

Flow map

For showing unambiguous movement across a map.

Contour map

For showing areas of equal value on a map. Can use deviation colour schemes for showing +/- values

Equalised cartogram

Converting each unit on a map to a regular and equally-sized shape – good for representing voting regions with equal value.

Scaled cartogram (value)

Stretching and shrinking a map so that each area is sized according to a particular value.

Dot density

Used to show the location of individual events/locations – make sure to annotate any patterns the reader should see.

Heat map

Grid-based data values mapped with an intensity colour scale. As choropleth map – but not snapped to an admin/political unit.

Flow

Show the reader volumes or intensity of movement between two or more states or conditions. These might be logical sequences or geographical locations.

Example FT uses
Movement of funds, trade, migrants, lawsuits, information, relationship graphs.

Sankey

Shows changes in flows from one condition to at least one other; good for tracing the eventual outcome of a complex process.

Waterfall

Designed to show the sequencing of data through a flow process, typically budgets. Can include +/- components.

Chord

A complex but powerful diagram which can illustrate 2-way flows (and net winner) in a matrix.

Network

Used for showing the strength and inter-connectedness of relationships of varying types.

Visual vocabulary

ft.com/vocabulary

Deviation

Emphasise variations (+/-) from a fixed reference point. Typically the reference point is zero but it can also be a target or a long-term average. Can also be used to show sentiment (positive/neutral/negative).

Example FT uses
Trade surplus/deficit, climate change

Diverging bar

A simple standard bar chart that can handle both negative and positive magnitude values.

Diverging stacked bar

Perfect for presenting survey results which involve sentiment (eg disagree/neutral/agree).

Spine

Splits a single value into two contrasting components (eg male/female).

Surplus/deficit filled line

The shaded area of these charts allows a balance to be shown – either against a baseline or between two series.

Correlation

Show the relationship between two or more variables. Be mindful that, unless you tell them otherwise, many readers will assume the relationships you show them to be causal (i.e. one causes the other).

Example FT uses
Inflation and unemployment, income and life expectancy

Scatterplot

The standard way to show the relationship between two continuous variables, each of which has its own axis.

Column + line timeline

A good way of showing the relationship between an amount (columns) and a rate (line).

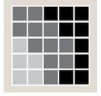
Connected scatterplot

Usually used to show how the relationship between 2 variables has changed over time.

Bubble

Like a scatterplot, but adds additional detail by sizing the circles according to a third variable.

XY heatmap



A good way of showing the patterns between 2 categories of data, less effective at showing fine differences in amounts.

Ranking

Use where an item's position in an ordered list is more important than its absolute or relative value. Don't be afraid to highlight the points of interest.

Example FT uses
Wealth, deprivation, league tables, constituency election results

Ordered bar

Standard bar charts display the ranks of values much more easily when sorted into order.

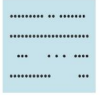
Ordered column

See above.

Ordered proportional symbol

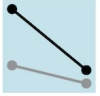
Use when there are big variations between values and/or seeing fine differences between data is not so important.

Dot strip plot



Dots placed in order on a strip are a space-efficient method of laying out ranks across multiple categories.

Slope



Perfect for showing how ranks have changed over time or vary between categories.

Distribution

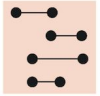
Show values in a dataset and how often they occur. The shape (or 'skew') of a distribution can be a memorable way of highlighting the lack of uniformity or equality in the data.

Example FT uses
Income distribution, population (age/sex) distribution, revealing inequality

Histogram

The standard way to show a statistical distribution - keep the gaps between columns small to highlight the 'shape' of the data.

Dot plot



A simple way of showing the change or range (min/max) of data across multiple categories.

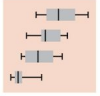
Dot strip plot

Good for showing individual values in a distribution, can be a problem when too many dots have the same value.

Barcode plot

Like dot strip plots, good for displaying all the data in a table, they work best when highlighting individual values.

Boxplot



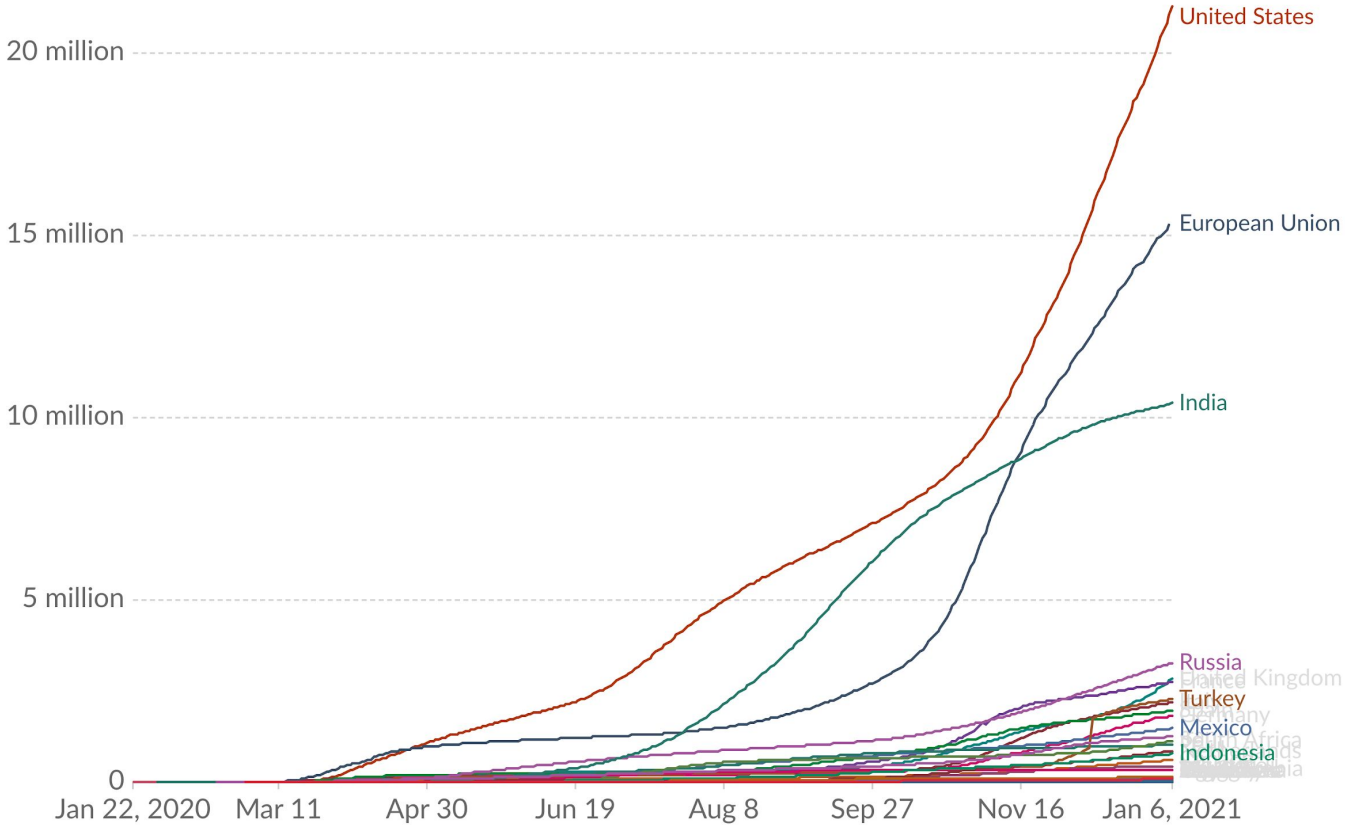
Summarise multiple distributions by showing the median (centre) and range of the data

COVID VISUALIZATIONS: A TIMELINE

Cumulative confirmed COVID-19 cases

The number of confirmed cases is lower than the number of actual cases; the main reason for that is limited testing.

Our World in Data



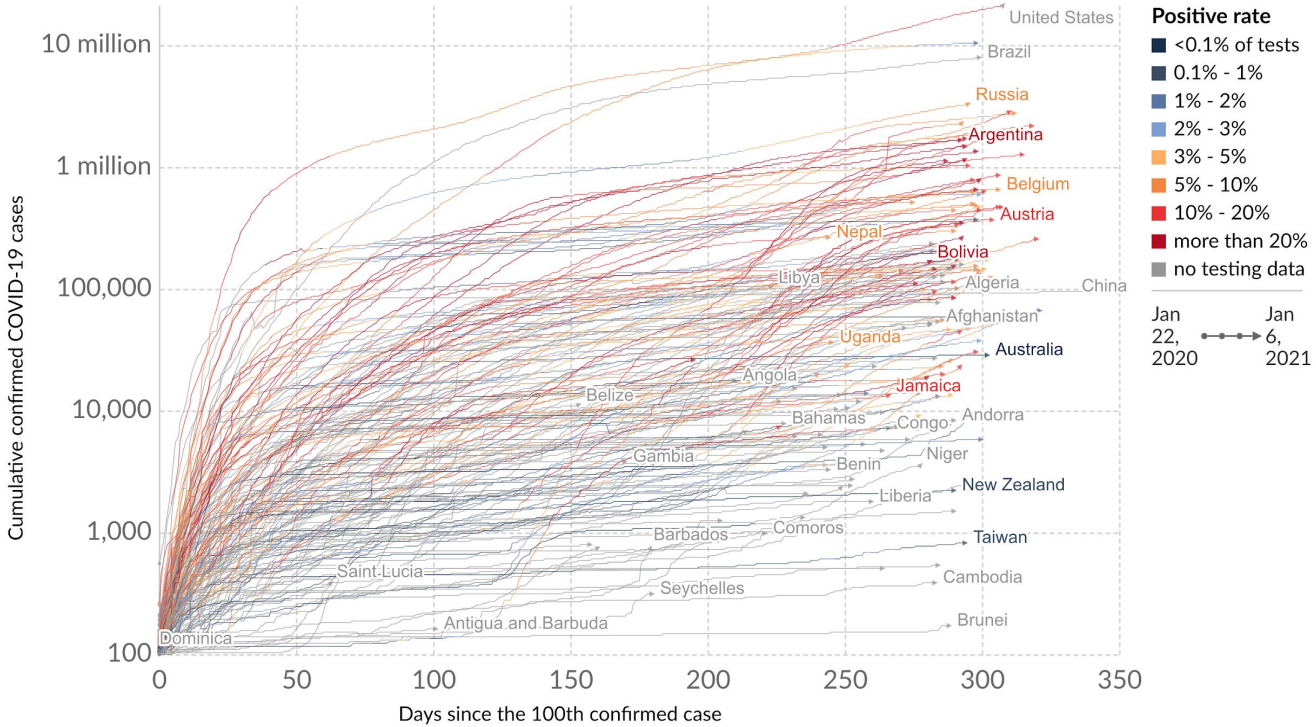
Source: Johns Hopkins University CSSE COVID-19 Data – Last updated 7 January, 06:07 (London time)

Created with Our World in Data: https://ourworldindata.org/covid-cases#what-is-the-total-number-of-confirmed-cases

Cumulative confirmed COVID-19 cases

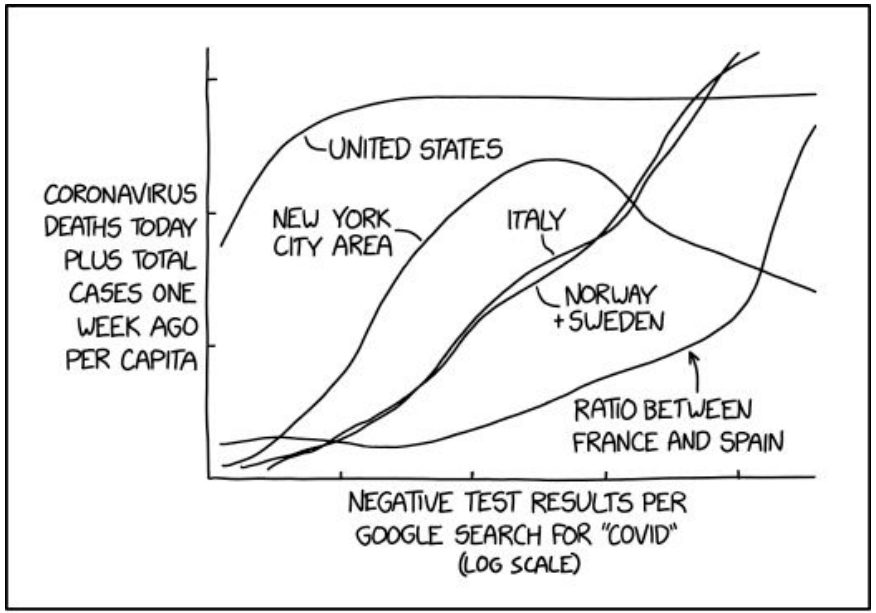
The number of confirmed cases is lower than the number of actual cases; the main reason for that is limited testing.

Our World in Data



Source: Johns Hopkins University CSSE COVID-19 Data – Last updated 7 January, 06:07 (London time), Official data collated by Our World in Data

Graphs got a little complicated

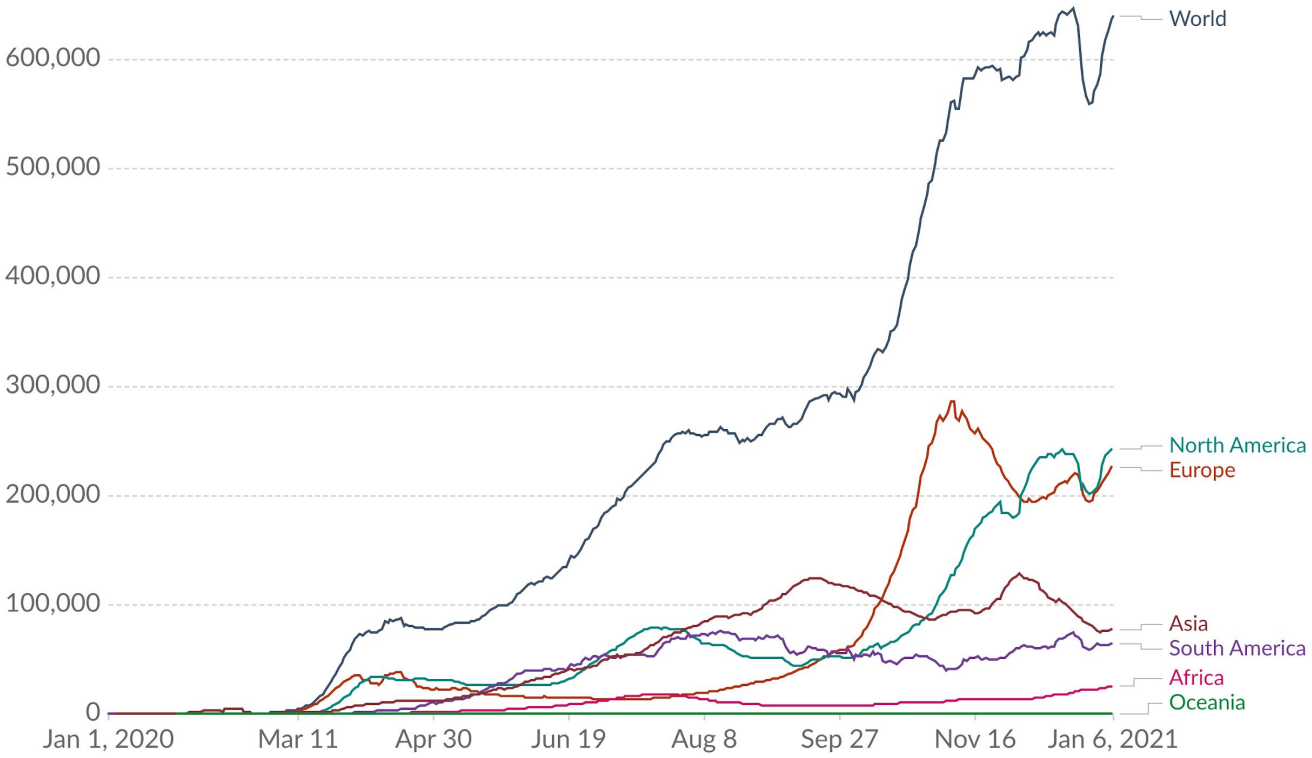


I'M A HUGE FAN OF WEIRD GRAPHS, BUT EVEN I ADMIT SOME OF THESE CORONAVIRUS CHARTS ARE LESS THAN HELPFUL.

Daily new confirmed COVID-19 cases

Shown is the rolling 7-day average. The number of confirmed cases is lower than the number of actual cases; the main reason for that is limited testing.

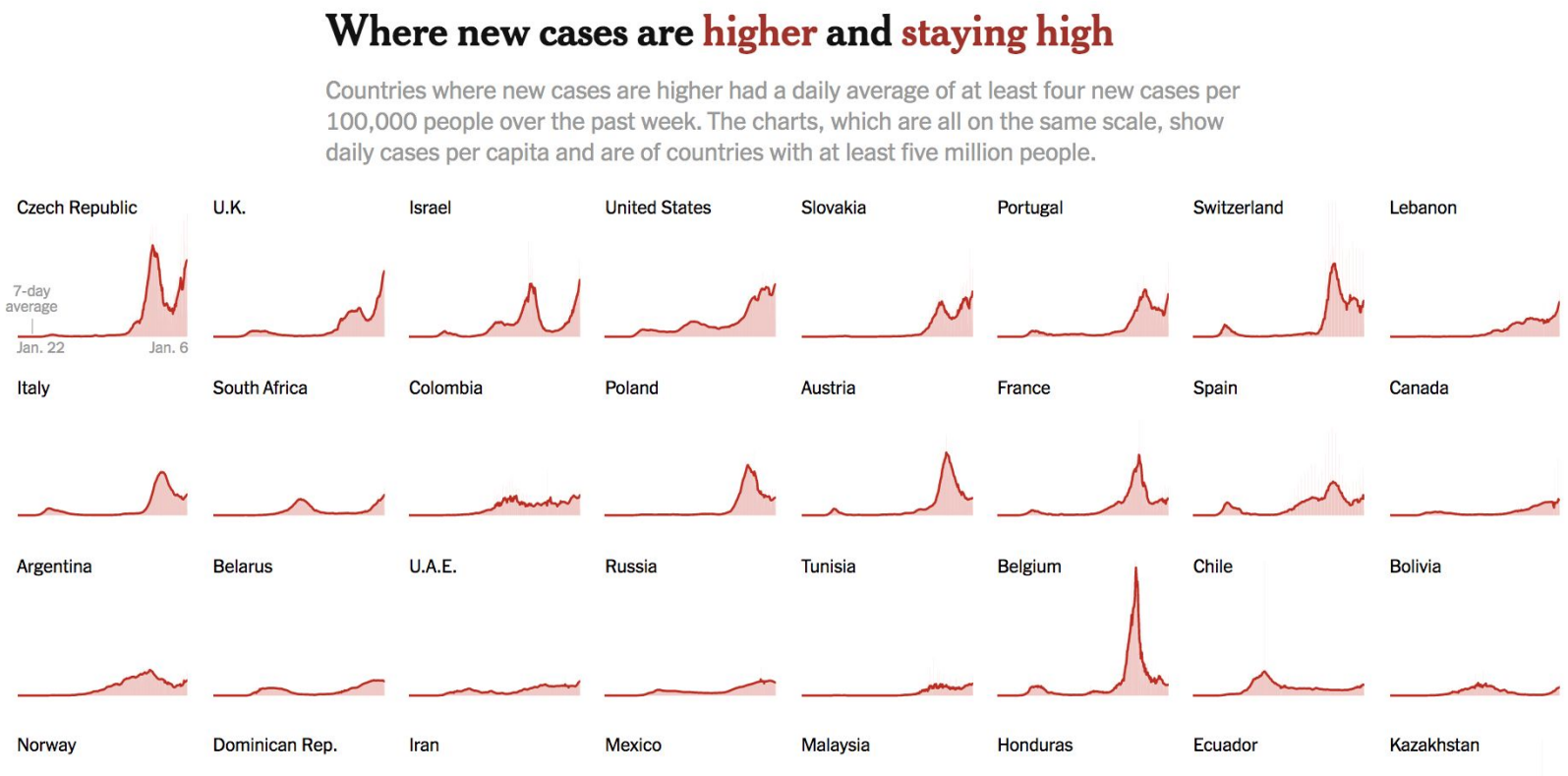
Our World in Data

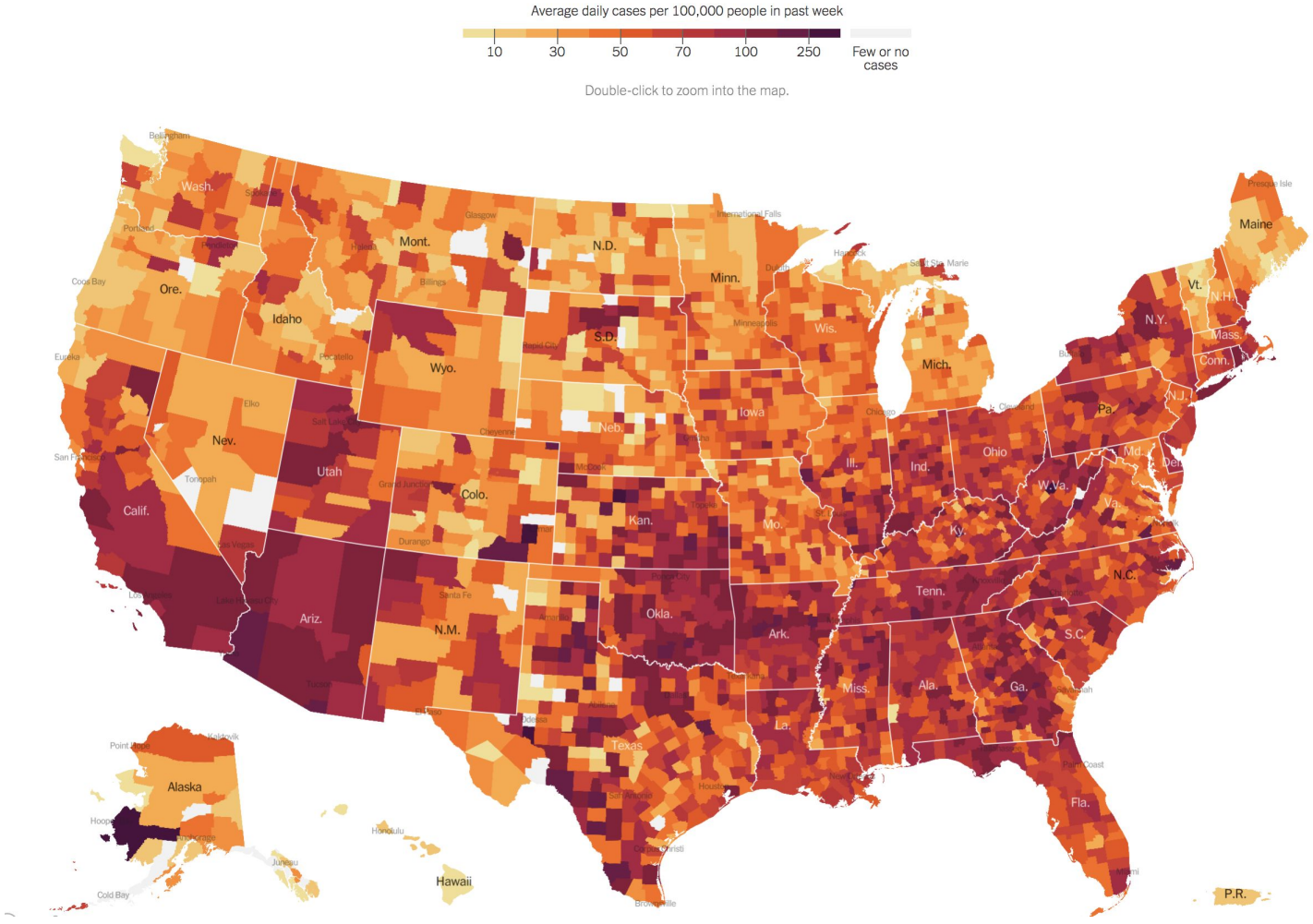


Source: Johns Hopkins University CSSE COVID-19 Data – Last updated 7 January, 06:07 (London time)

Created with Our World in Data: https://ourworldindata.org/covid-cases#what-is-the-total-number-of-confirmed-cases

Small multiples became very, very popular





New York Times. https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html

Where new cases are higher and staying high

States where new cases are higher had a daily average of at least 15 new cases per 100,000 people over the past week. Charts show daily cases per capita and are on the same scale. Tap a state to see detailed map page.

Cases and deaths by state and county

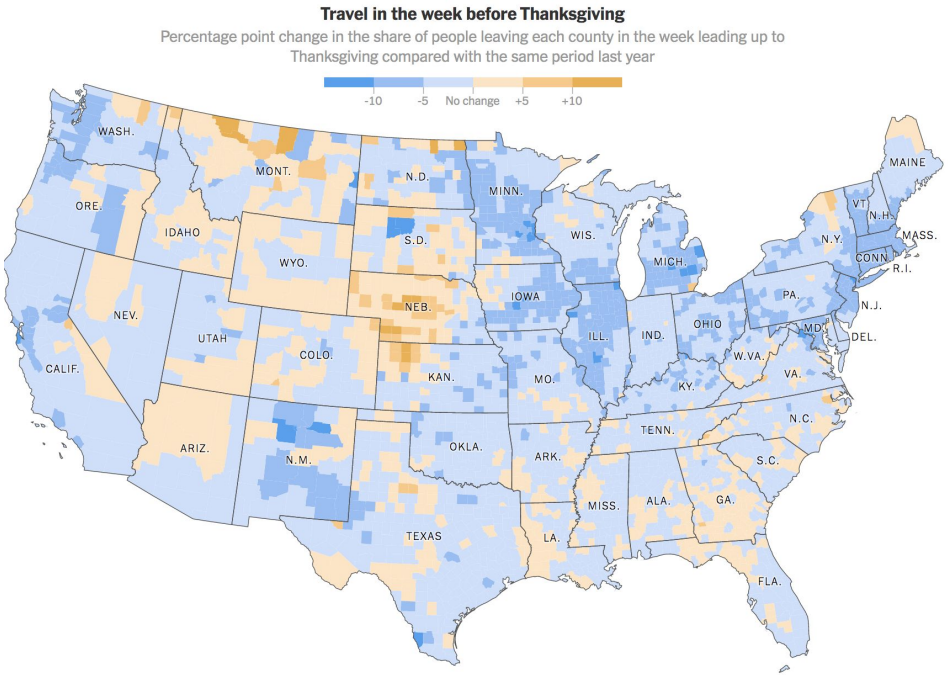
This table is sorted by places with the most cases per 100,000 residents in the last seven days. Charts are colored to reveal when outbreaks emerged.

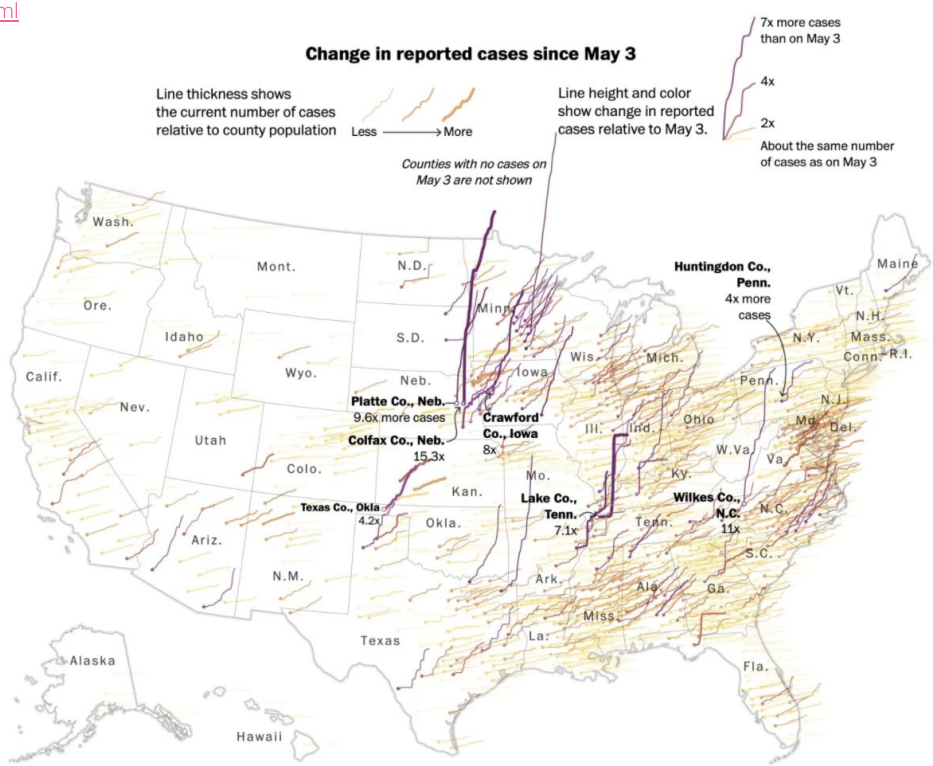
Cases | Deaths | Search counties

	TOTAL CASES	PER 100,000	DAILY AVG. IN LAST 7 DAYS	▼ PER 100,000	WEEKLY CASES PER CAPITA (FEWER – MORE)
+ Arizona MAP »	578,623	7,950	9,037	124	March 1 – Jan. 6
+ Rhode Island MAP »	95,463	9,011	1,073	101	
+ California MAP »	2,548,494	6,450	38,998	99	
+ Arkansas MAP »	242,593	8,039	2,880	95	
+ Utah MAP »	292,720	9,131	2,969	93	
+ Tennessee MAP »	610,347	8,937	6,270	92	
+ Oklahoma MAP »	311,573	7,874	3,506	89	
+ West Virginia MAP »	94,678	5,283	1,493	83	
+ South Carolina MAP »	333,235	6,472	4,251	83	
+ Alabama MAP »	384,184	7,835	3,909	80	

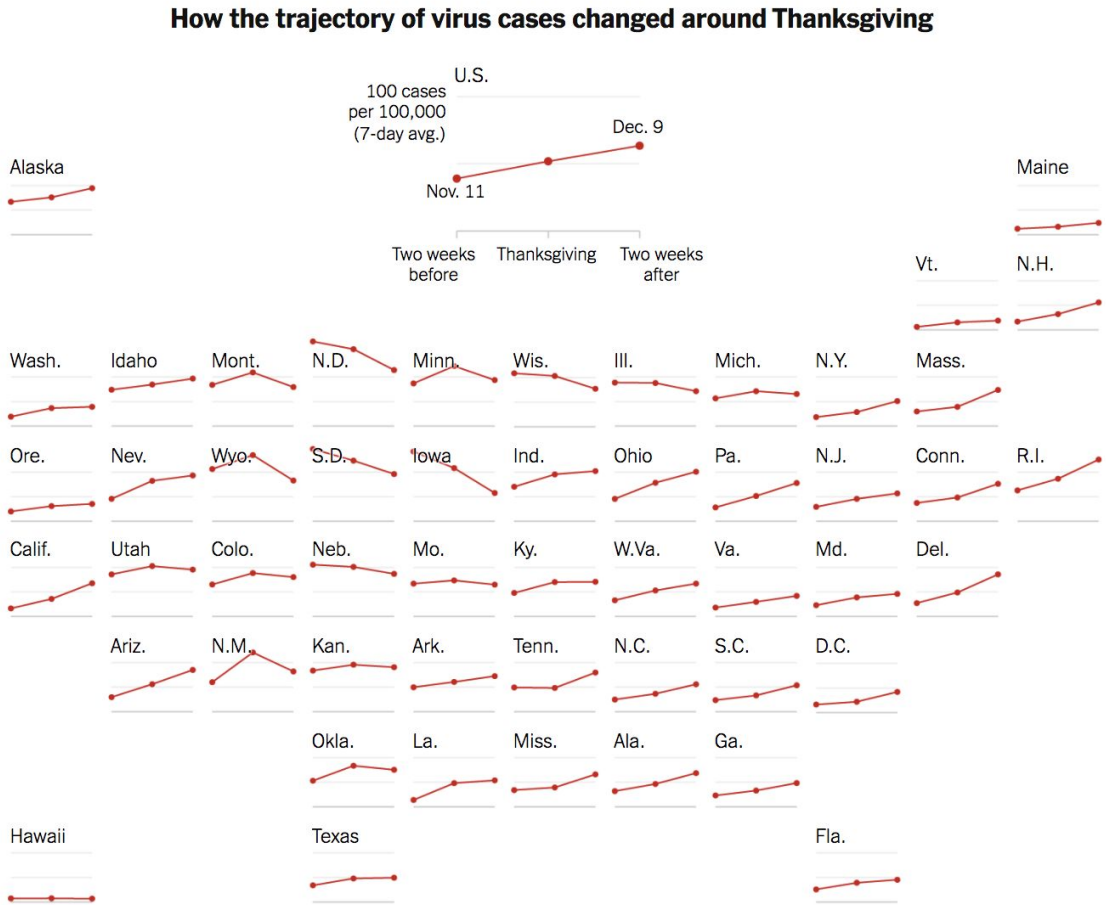
Creative maps

New York Times. https://www.nytimes.com/interactive/2020/12/20/us/covid-thanksgiving-effect.html

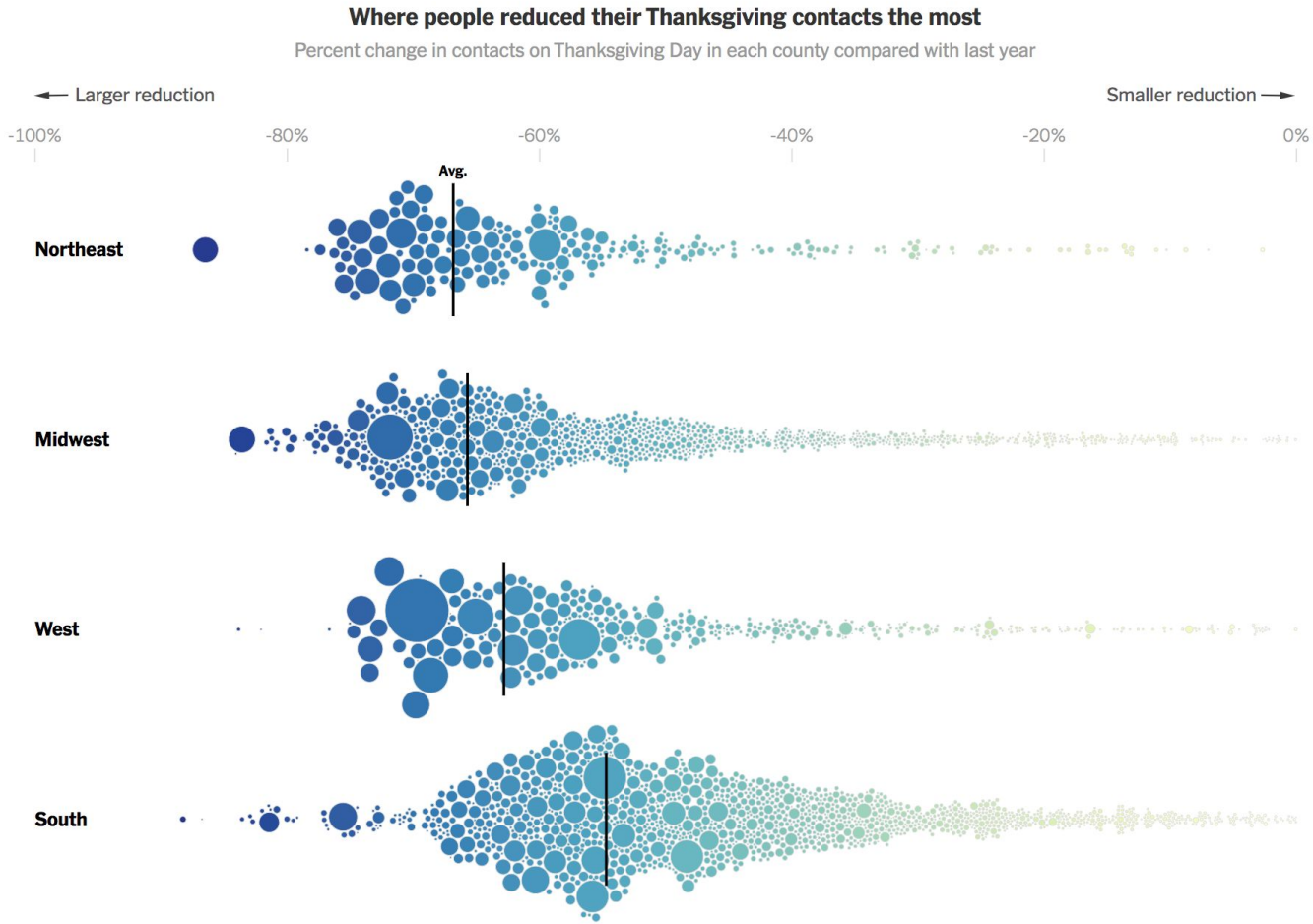




Washington Post. https://www.washingtonpost.com/nation/2020/05/24/coronavirus-rural-america-outbreaks/



New York Times. https://www.nytimes.com/interactive/2020/12/20/us/covid-thanksgiving-effect.html



New York Times. https://www.nytimes.com/interactive/2020/12/20/us/covid-thanksgiving-effect.html

FINALLY: ANNOTATE AND SOURCE YOUR VISUALIZATIONS

Annotate, annotate, annotate.

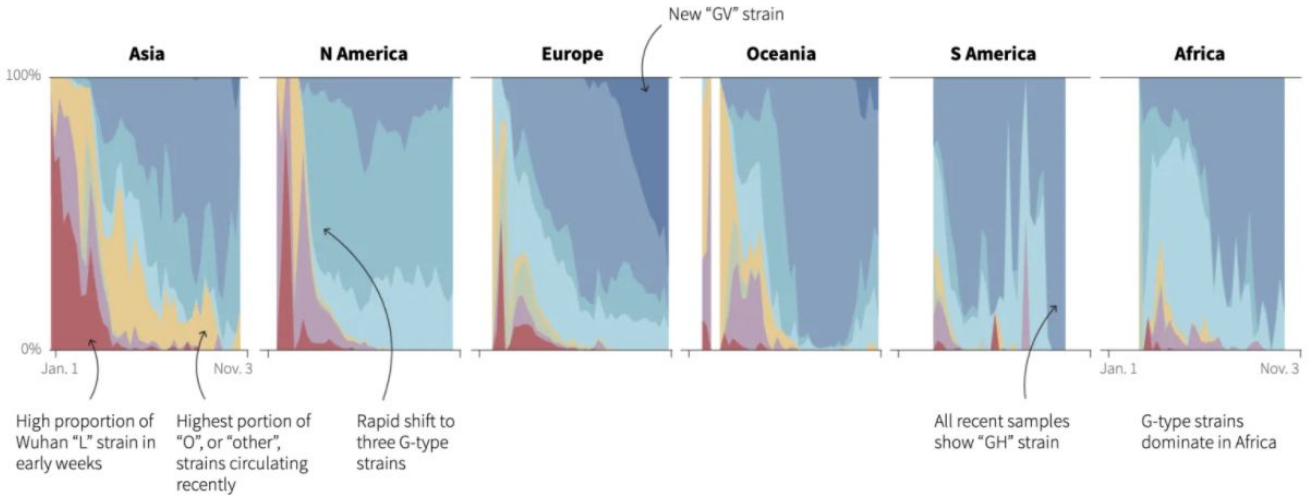
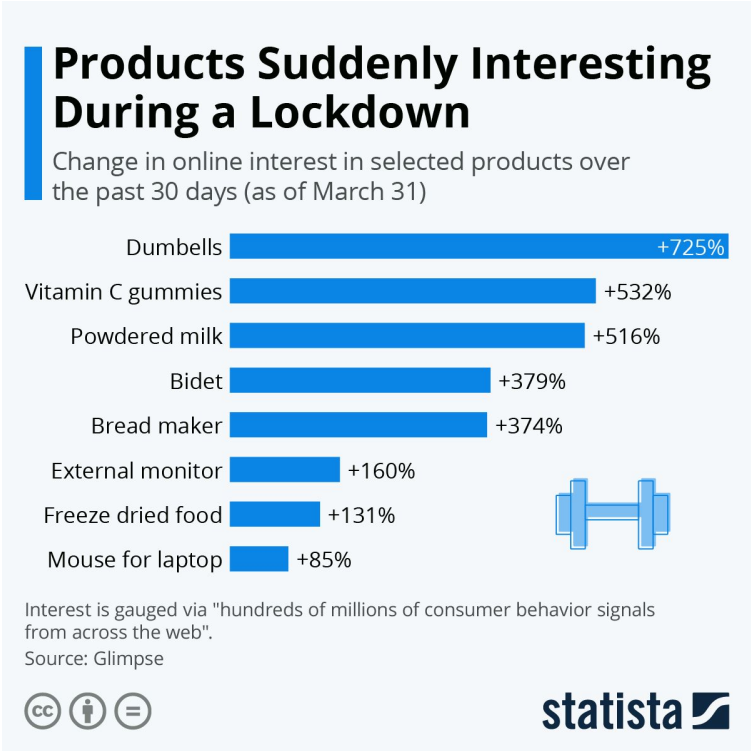


Image by Reuters: https://graphics.reuters.com/HEALTH-CORONAVIRUS/EVOLUTION/yxmpjqkdzvr/

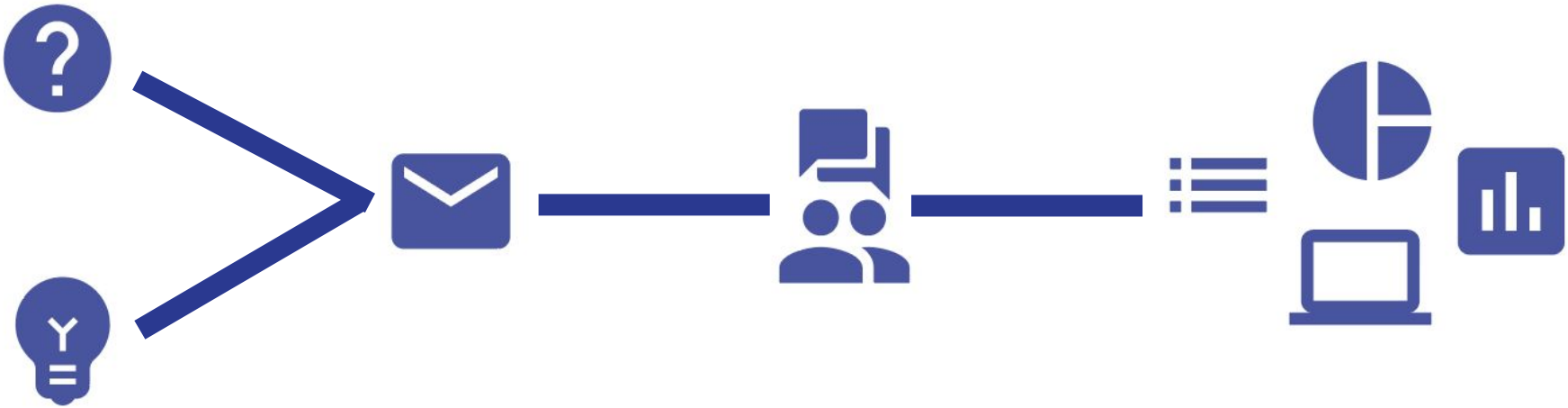
Let's explore a master case of this:
https://www.nytimes.com/interactive/2020/03/15/business/economy/coronavirus-worker-risk.html

Don’t forget to include source and attribution preferences.



By Statista. License: CC-BY-SA.

Ask for help early and often.



Jess Cohen-Tanugi
Visualization Specialist
jessica_cohen-tanugi@harvard.edu

Hugh Truslow
Head, Social Sciences and Visualization Services for Academic Programs
truslow@fas.harvard.edu