12-3-2020

Data Science in the Time of COVID-19

Tony Breitzman
Rowan University, breitzman@rowan.edu

Data Science in the Time of COVID-19

IEEE WIE Philadelphia Region 2
December 3, 2020

Anthony Breitzman, PhD
Associate Professor of Computer Science
Data Science Program Coordinator
Rowan University

Rowan University

Introduction

- Since the start of the Pandemic it seems that everyone is an expert…

Source: Reddit.com

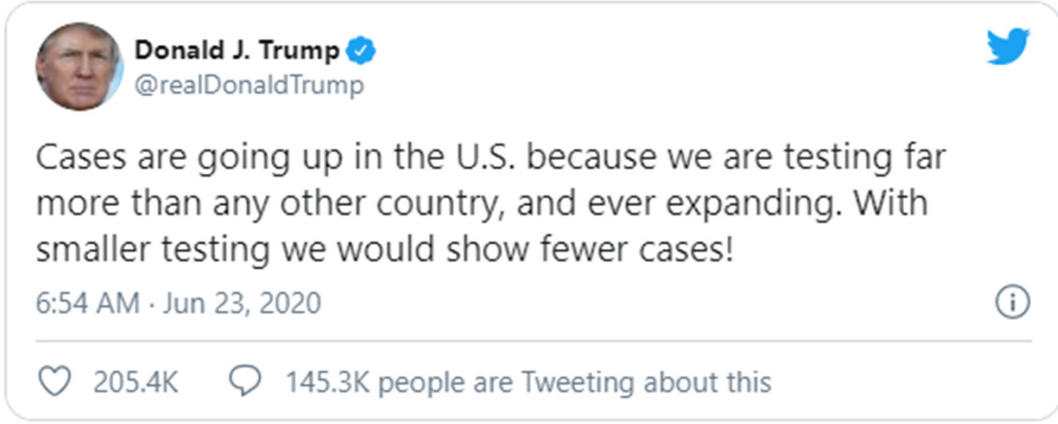
Introduction (2)

- Just as everyone on Facebook is an epidemiologist…
- It seems everyone is a Data Scientist.

- Tonight we'll talk about actual contributions from Data Scientists
- But first we'll look at some bad Data Science and do some mythbusting

Overview

- Part 1: Mythbusting
  - Bad visualizations
  - Mythbusting with good visualizations
- Part 2: Good news
  - How Data Science is helping



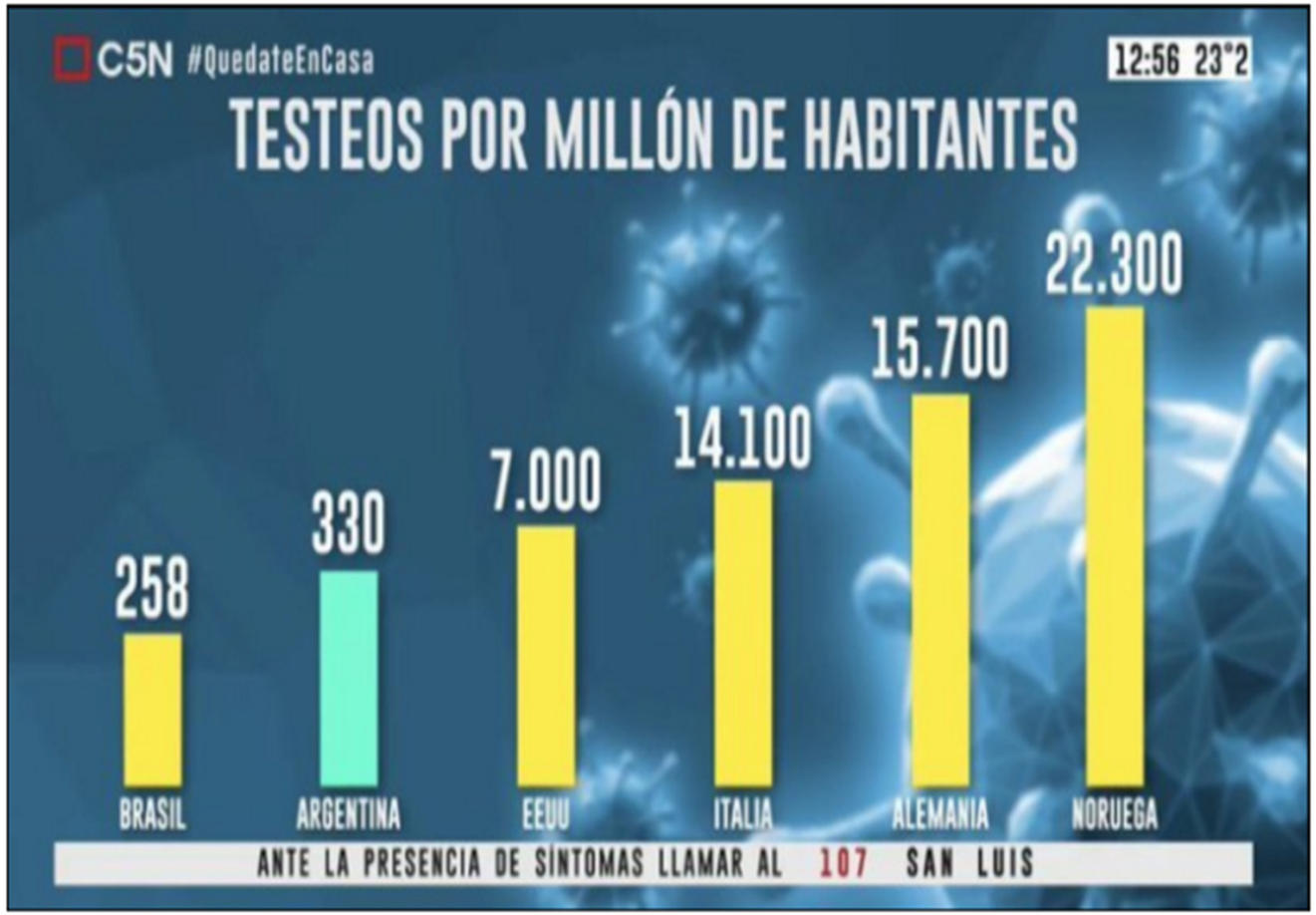
Part 1a: Bad Visualizations

- "There are three kinds of lies: lies, damned lies, and statistics.“ Origin unknown. Popularized by Mark Twain

- New variation: “There are lies, damned lies and bad visualizations”

- Corollary: “Numbers don’t lie, people with graphs do.”

Look! Argentina is doing a great job on testing!

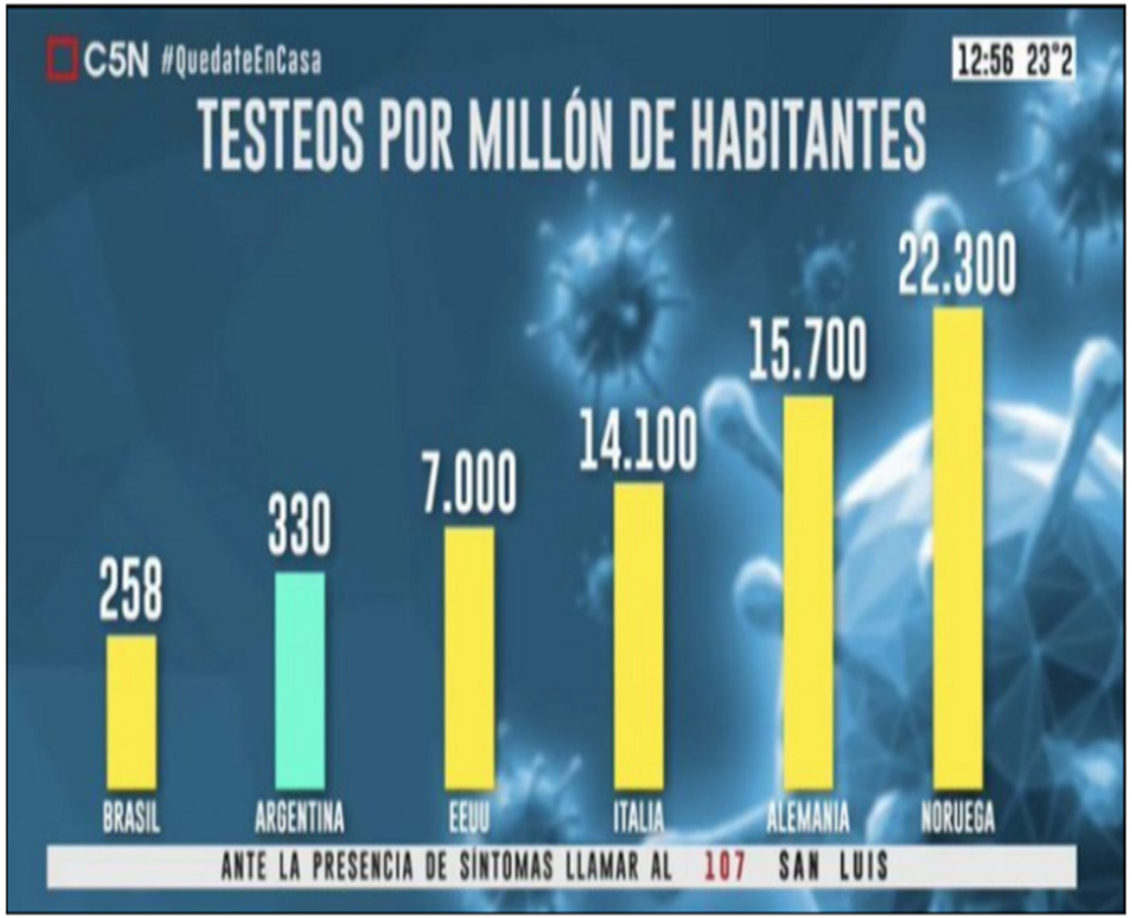


Argentinian TV channel C5N manipulating the y-axis to hide the terrible number of COVID-19 test. Source: Reddit. Original Source: C5N

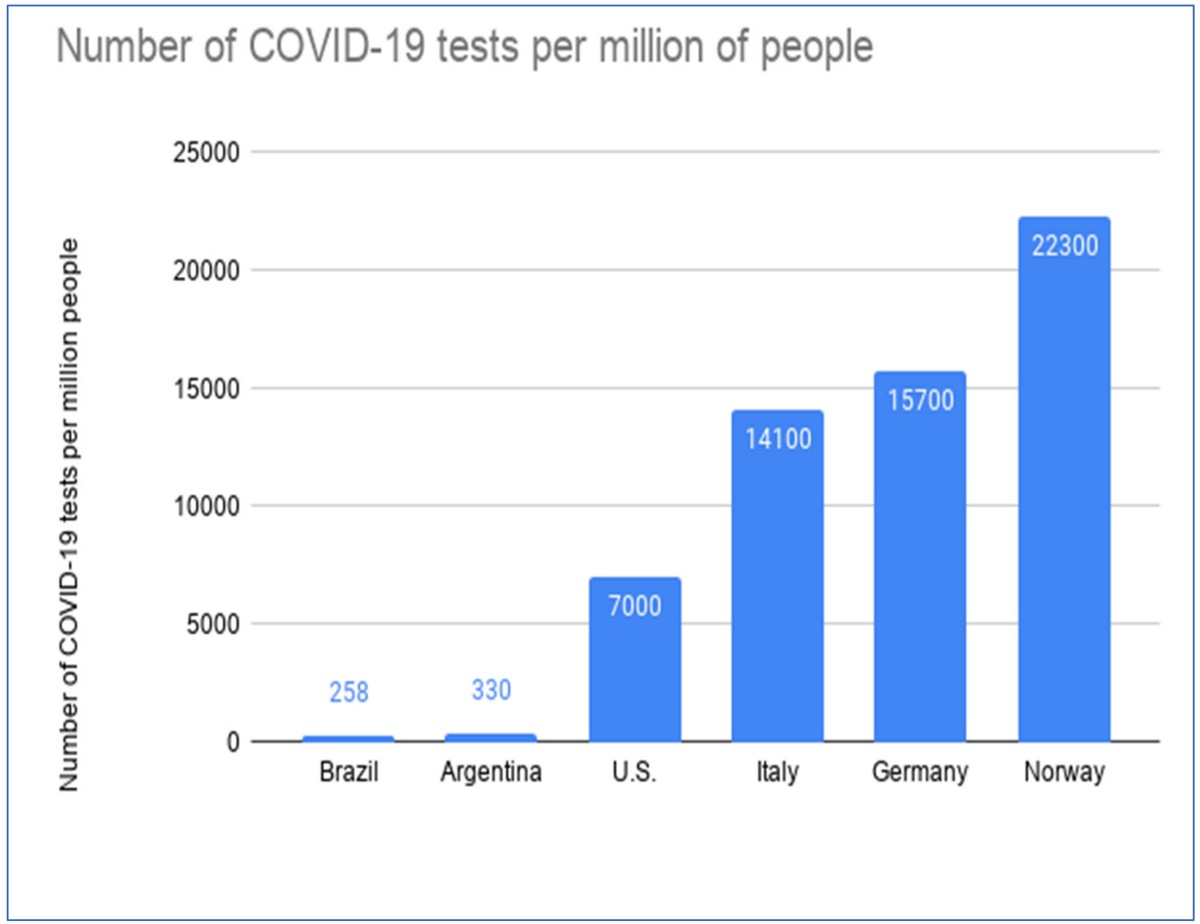
Reprinted from: https://towardsdatascience.com/stopping-covid-19-with-misleading-graphs-6812a61a57c9

- From Nikita Kotsehub in Towards Data Science
- Note EEUU is United States, Alemania is Germany, and Noruega is Norway.
- This purports to show that Argentina has done almost as much testing as the US and Italy
- It's nice that they normalized the data (tests per million population) but…
- It's not obvious from the bars that Argentina's numbers are only 4.7% of the US and 2.3% of Italy

Proper graph on the right…



Argentinian TV channel C5N manipulating the y-axis to hide the terrible number of COVID-19 test. Source: Reddit. Original Source: C5N



Reprinted from: https://towardsdatascience.com/stopping-covid-19-with-misleading-graphs-6812a61a57c9

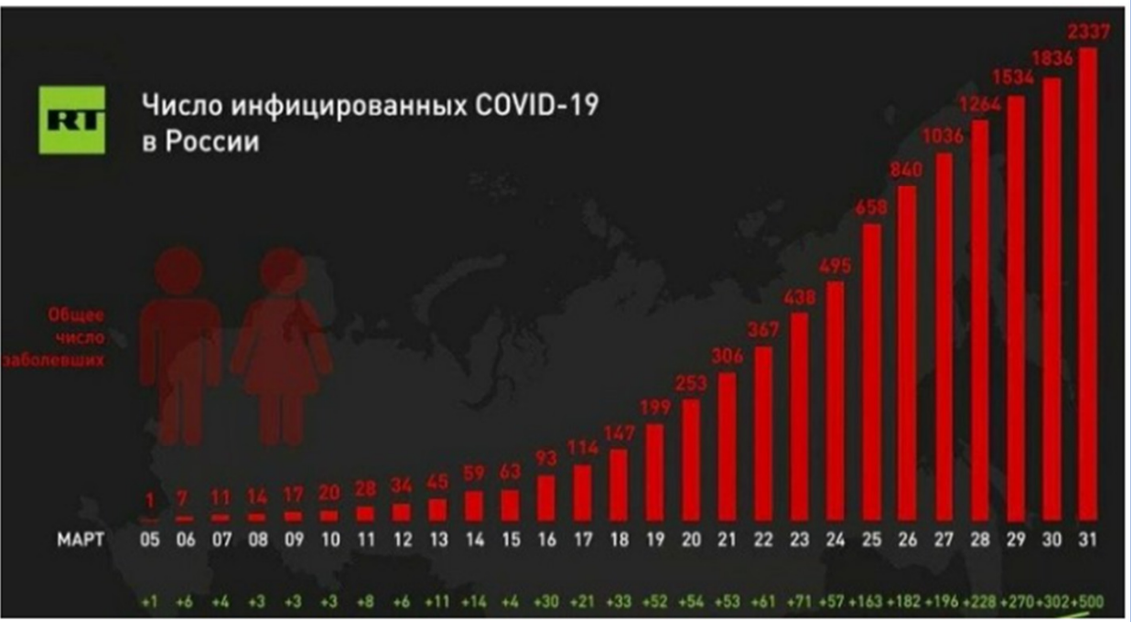
Where do we even start with this one?

- Notice the Y-Axis increases by 30-30-30-10-30-30-30-50-10-50-50-50
- Oddly the 246 is closer to 240 then 250
- The step between 112 and 129 (17) is the same as the step between 192 and 246 (54)

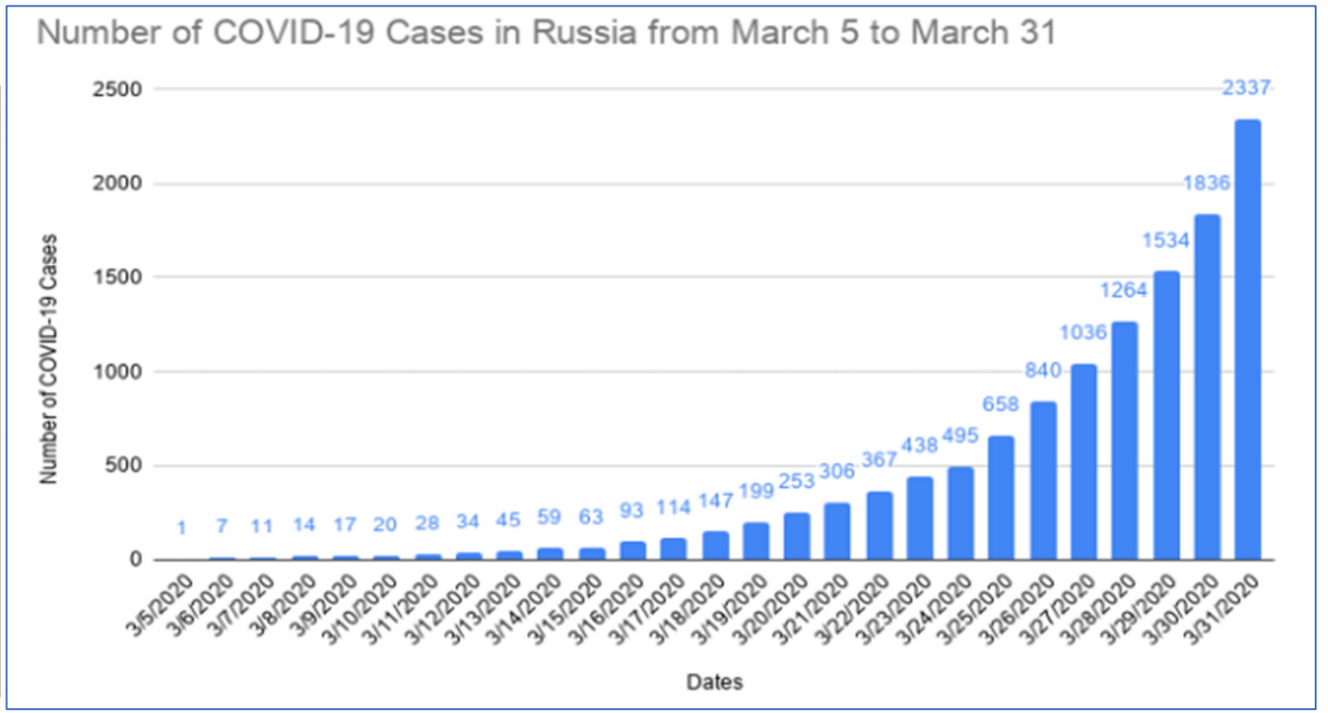
- I don't know if this was intentionally misleading or just incompetence
- The narrative at the time was "it's time to open the economy." Increasing the Y-Axis tick marks at the end of the month would help that narrative.

Source: Fox 31 (KDVR), Denver CO. (Widely discussed on Reddit)

Here's a Russian TV Channel Showing how to Flatten the curve!



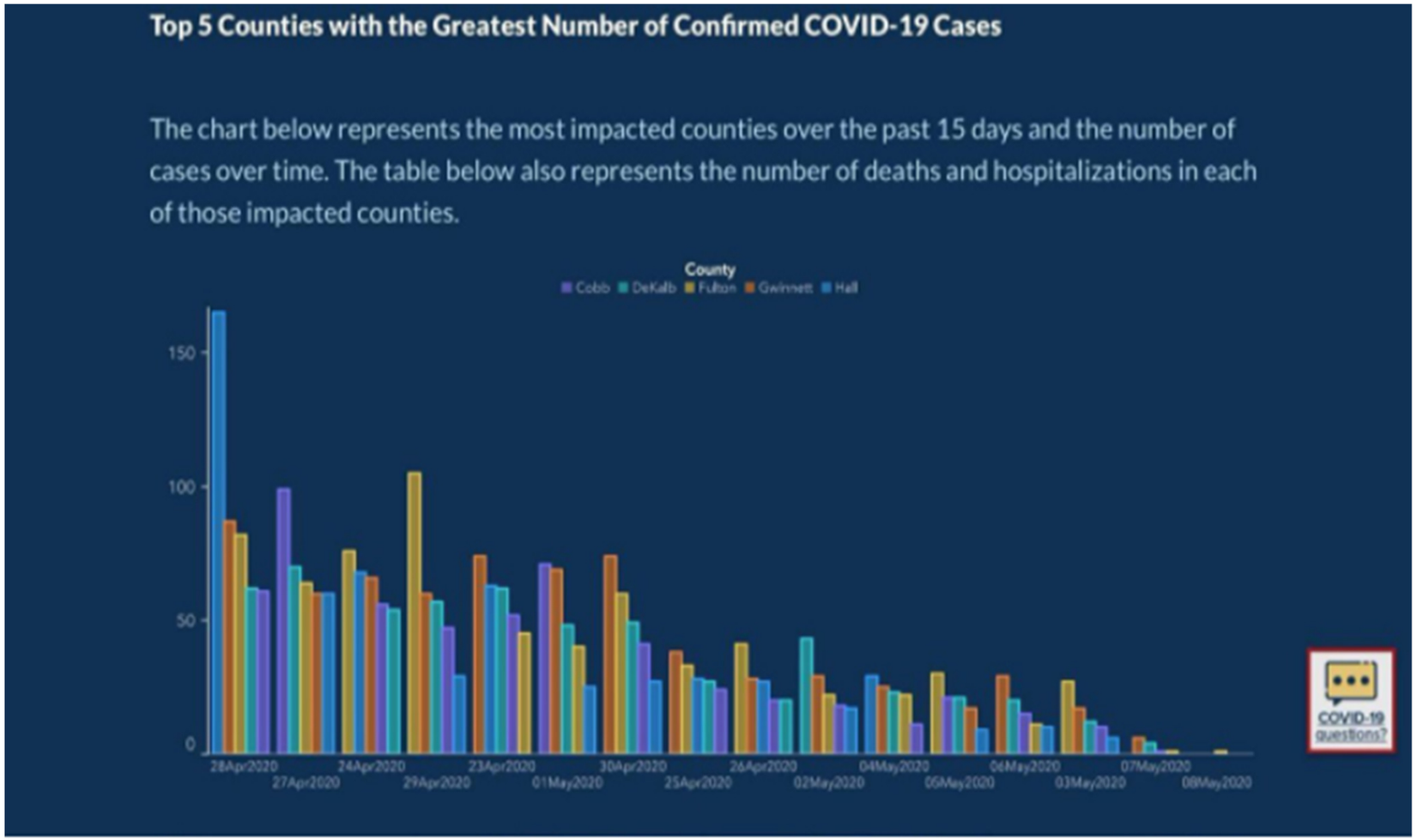
The number of COVID-19 cases in Russia from March 5 to March 31. Source: Reddit. Original Source: Russia Today



Reprinted from: https://towardsdatascience.com/stopping-covid-19-with-misleading-graphs-6812a61a57c9

- Another example from Nikita Kotsehub
- There seems to be an epidemic of Y-Axis manipulation
- In the next slide we'll see the state of Georgia get creative with the X-Axis

This was a much-discussed example on Twitter

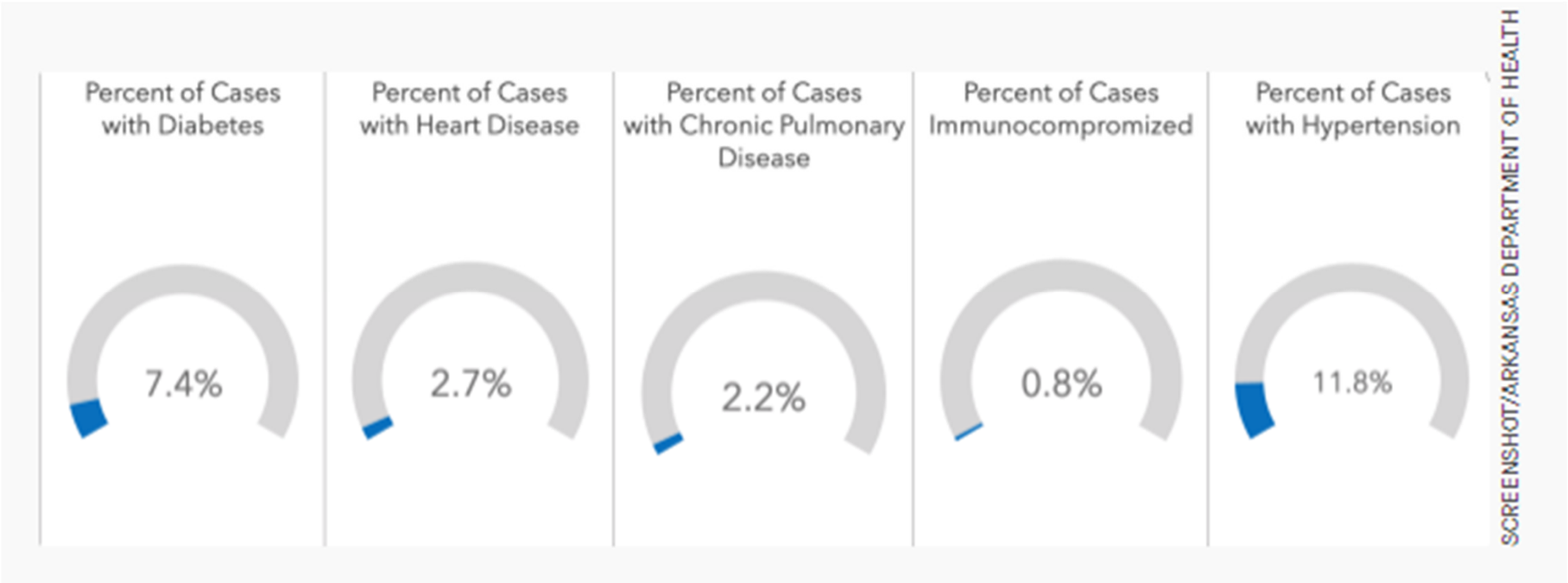


A screenshot of the chart published on Georgia's Covid-19 dashboard in May that falsely showed a decrease in the state's infections—by rearranging the order of the dates at the bottom.

Reprinted from: https://www.atlantamagazine.com/great-reads/behind-georgias-covid-19-dashboard-disaster/
Original Source: Georgia Department of Public Health

- "Only in Brian Kemp's Georgia is the first Thursday in May followed immediately by the last Sunday in April," a Washington Post columnist quipped.
- Pete Corson of the Atlanta Journal Constitution tweeted that the graphic had been "the subject of much head scratching" at his publication.
- In a response to Corson, Kemp's director of communications, Candice Broce, implied the health department was to blame: "The graph was supposed to be helpful," she tweeted, "but was met with such intense scorn that I, for one, will never encourage DPH to use anything but chronological order on the x axis moving forward."

This one is not intentionally misleading. It's just plain dumb

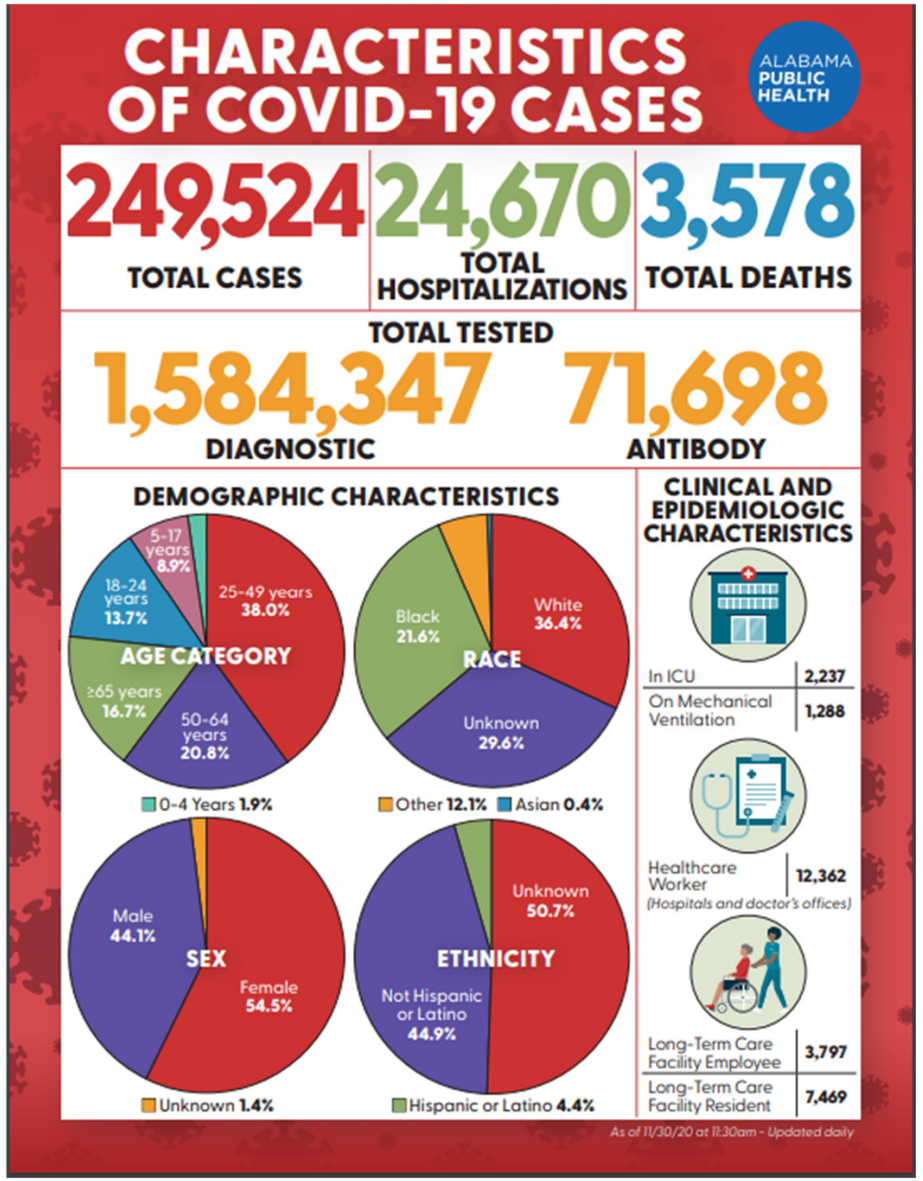


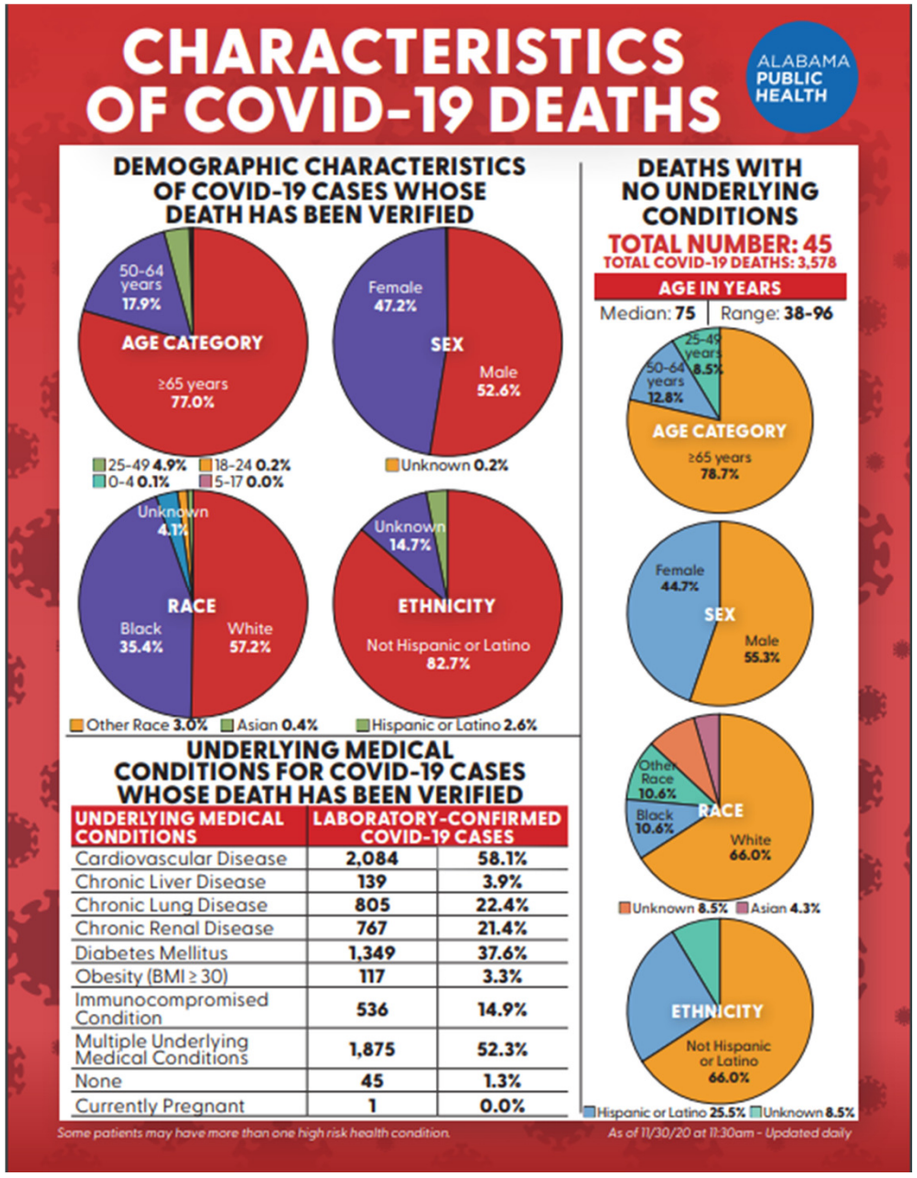
Reprinted from: https://qz.com/1872980/how-bad-covid-19-data-visualizations-mislead-the-public/

Original Source: Arkansas Department of Health

- This example is from Katherine Ellen Foley at QZ.com
- Like a lot of websites this dashboard is automatically updated daily using software called ArcGIS
- It's well-designed software so what could go wrong?
- Without providing any context we can only assume that these pre-existing conditions are not a big deal.
- Arkansas has had 157,000 cases. Hypertension is a huge contributor to COVID complications. We see no indication that 17,000+ people are at risk for complications due to hypertension

Alabama puts out this 2-page Scoreboard every day!





Source: https://www.alabamapublichealth.gov/covid19/assets/cov-al-cases-113020.pdf

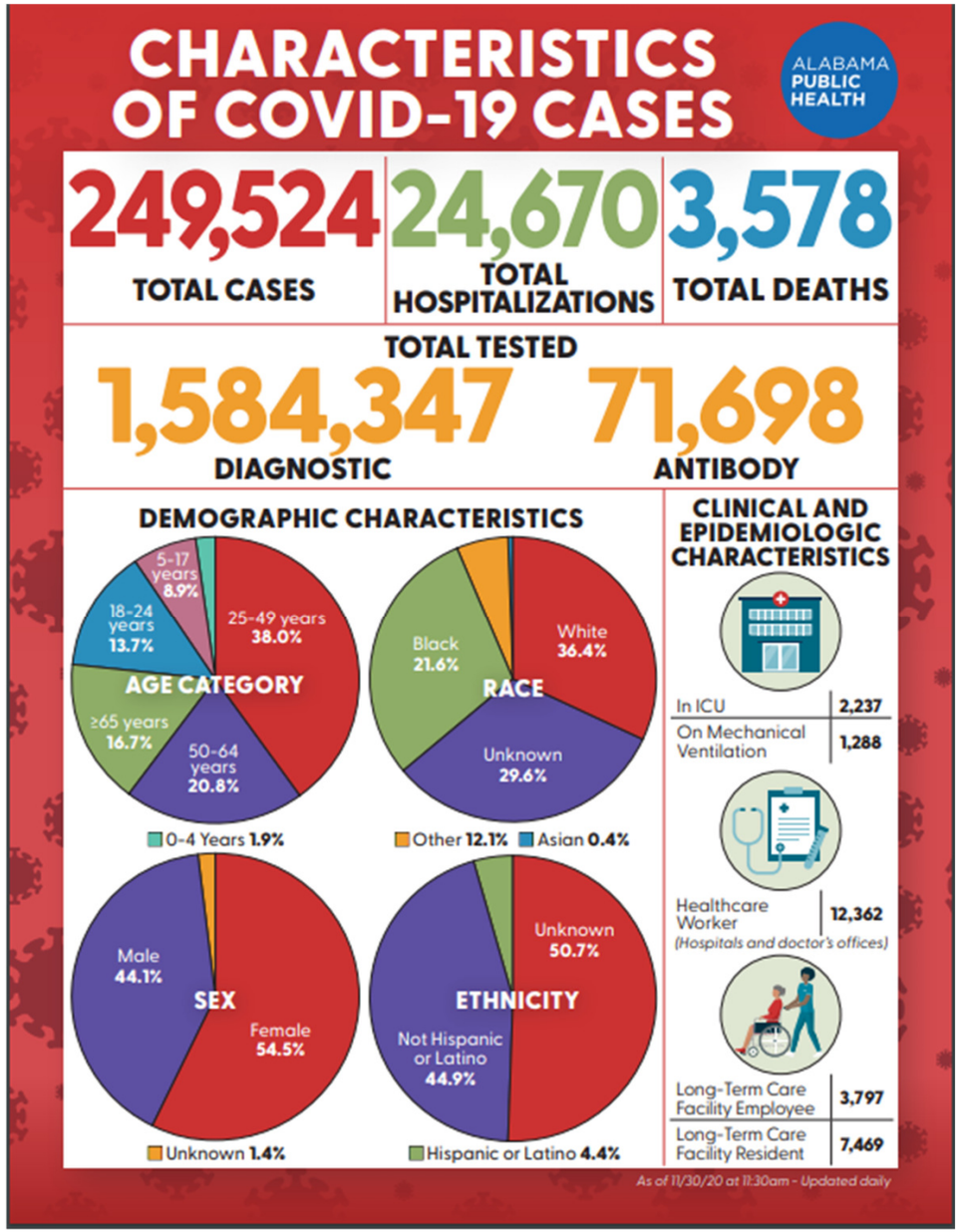
Another example from Katherine Ellen Foley at QZ.com

Surely, I'm not going to pick on Alabama for keeping its citizens up to date. Right? Wrong!

What is the point of updating the numbers daily if we're never going to show a trend!

Continued on next slide

Alabama puts out this 2-page Scoreboard every day!



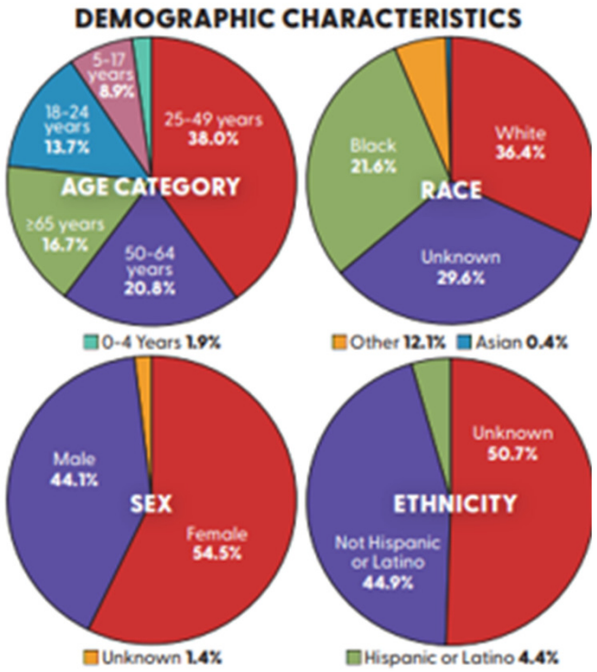
Source: https://www.alabamapublichealth.gov/covid19/assets/cov-al-cases-113020.pdf

A pie chart has no place in Data Science/Data Visualization

- It doesn't show trends
- Humans don't read them well
- Pie charts usually only work well when you have 3-4 slices, and you wish to show that one slice is very small relative to other slices

Almost without exception, if you are tempted to use a pie-chart, a bar chart will be a better choice

Same Pies followed by a well-designed table from NY



Pie subset from:
https://www.alabamapublichealth.gov/covid19/asses/cov-al-cases-113020.pdf

Fatalities by Race/Ethnicity *Data is preliminary. With 99% reporting, below is the breakdown for NYS excluding NYC. With 63% reporting, below is the breakdown for NYC as provided by NYCDOHMH.*

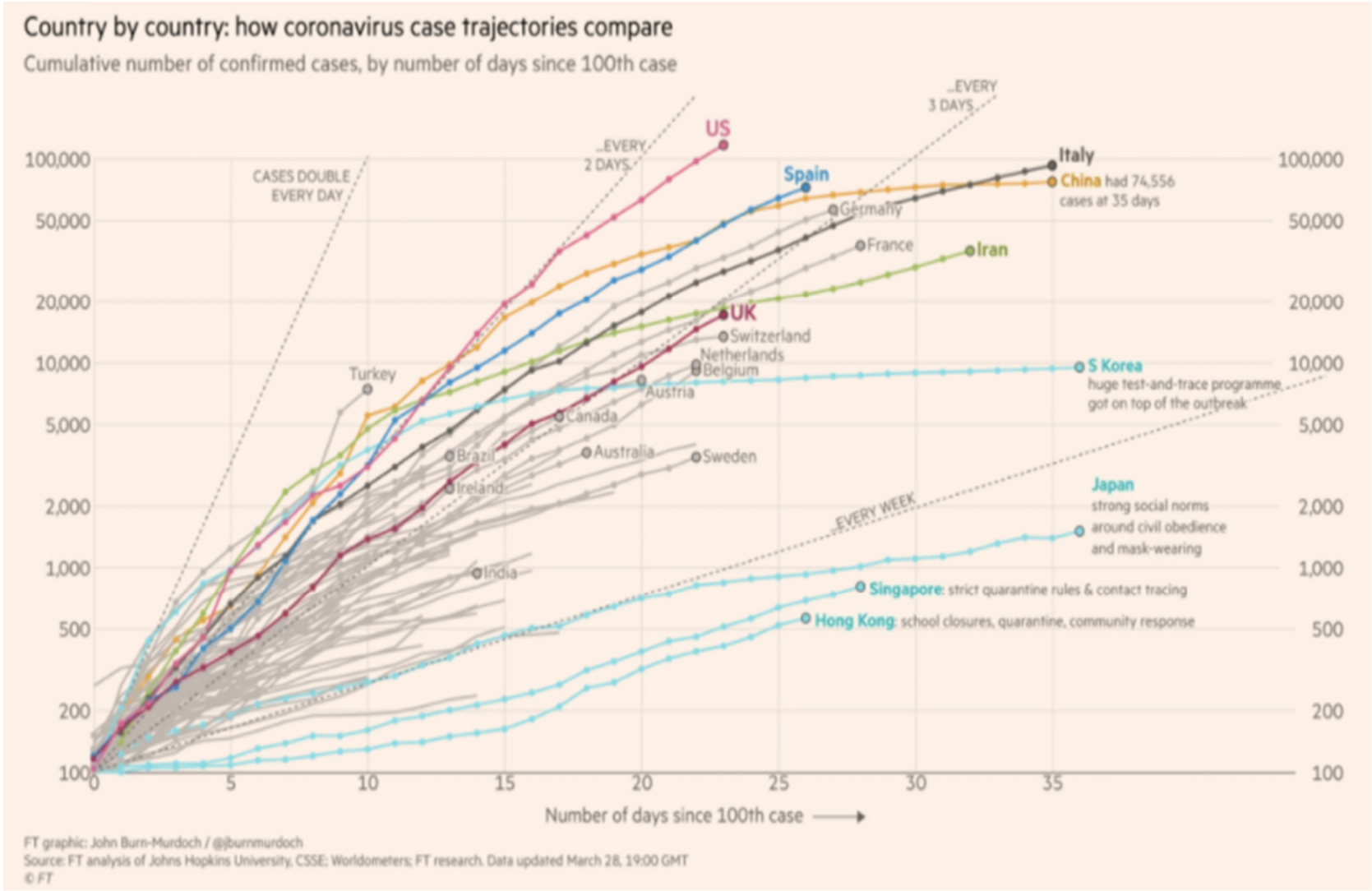
Click to see NYS excl. NYC age-adjusted rate

Race/Ethnicity	NYC	NYS Excl. NYC
Hispanic	34% (29% of population)	14% (12% of population)
Black	28% (22% of population)	17% (9% of population)
White	27% (32% of population)	61% (74% of population)
Asian	7% (14% of population)	4% (4% of population)
Other	4% (3% of population)	4% (1% of population)

Reprinted from: https://qz.com/1872980/how-bad-covid-19-data-visualizations-mislead-the-public/
Original Source: New York Department of Health

- Sometimes a well-designed table is better than a graph
- The table doesn't show trends (but they have trend plots elsewhere on the NYDH website. Alabama does not.)
- The table does provide context which the Pies do not. For example 21.6% of the cases are Black, but what percentage of Alabama's population is black? Can't tell from the Pie chart!
- Example from Katherine Ellen Foley at QZ.com

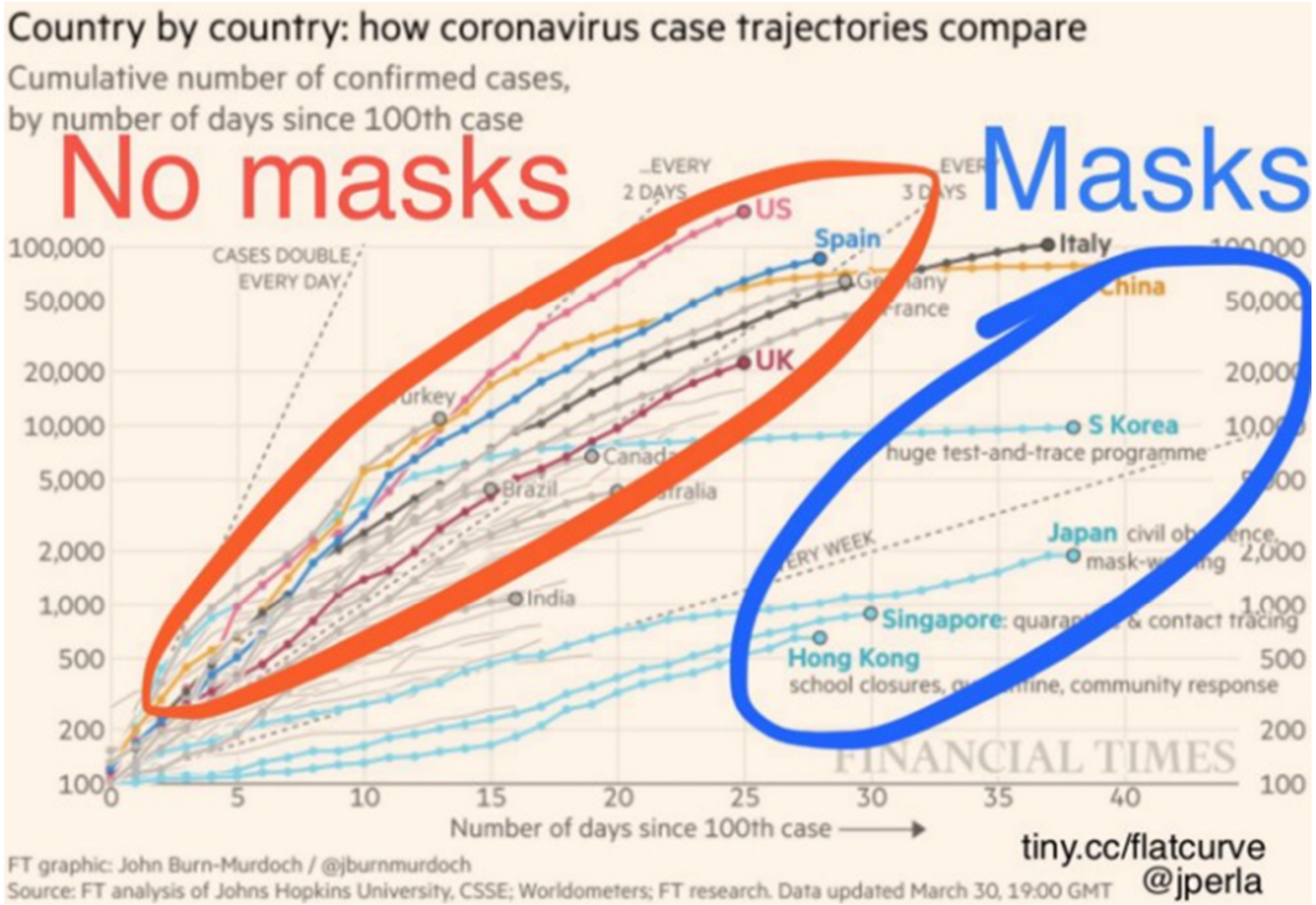
One of the best visualizations devised during COVID-19



Reprinted from: John Burn-Murdoch, Financial Times

- Uses a logarithmic scale which is reasonable for exponential growth
- Starts at days since 100th case so countries that have their first case at different times can be compared
- Contains lines for cases doubling every day, 2-days, week, etc. so that you can compare your country's slope to those guideposts and predict your growth rate
- Many data scientists (including me) have adopted the format
- If this is such a great visualization, then why is it in the section on bad visualizations? See next slide.

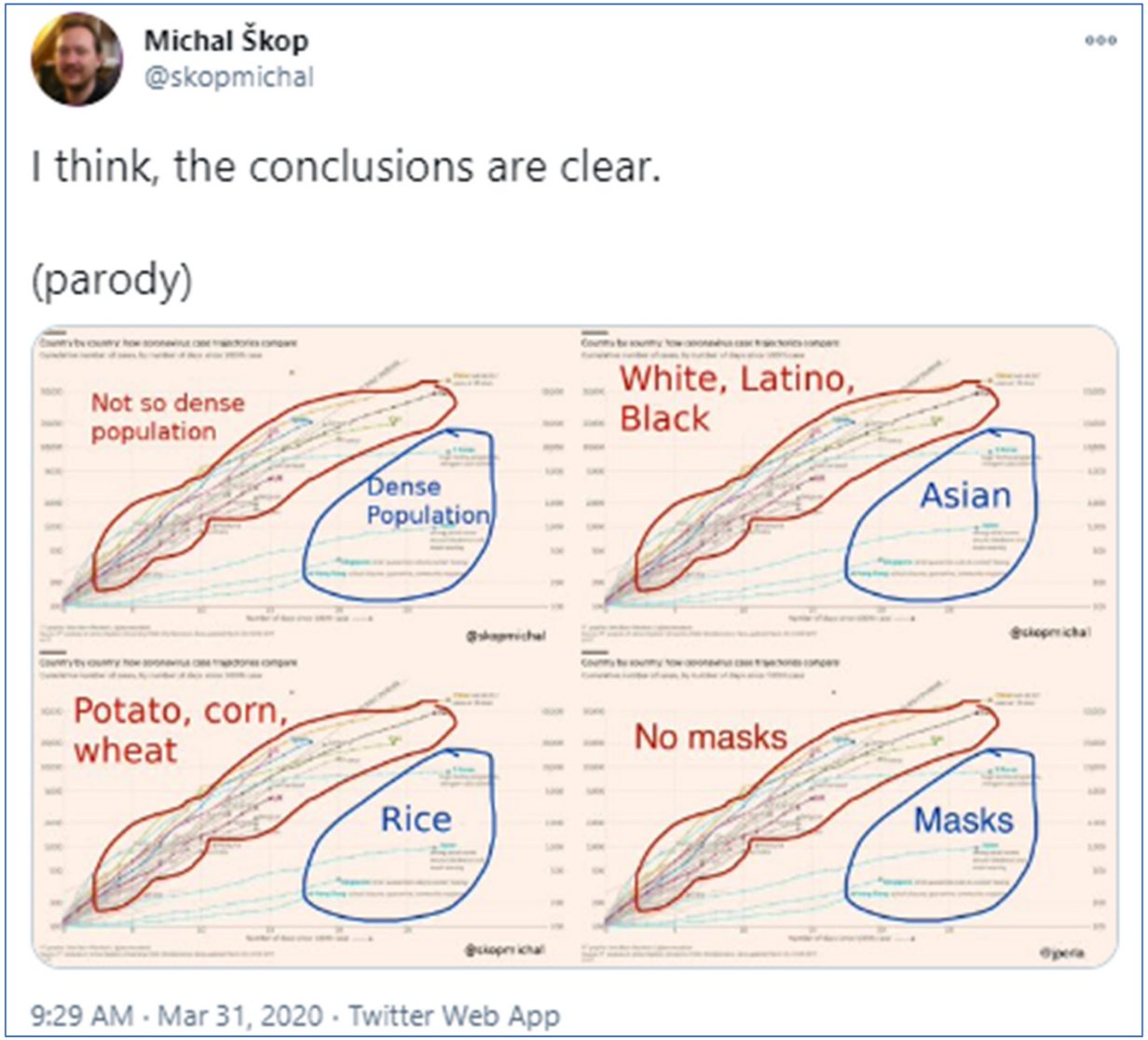
I actually have mixed feelings on this one….

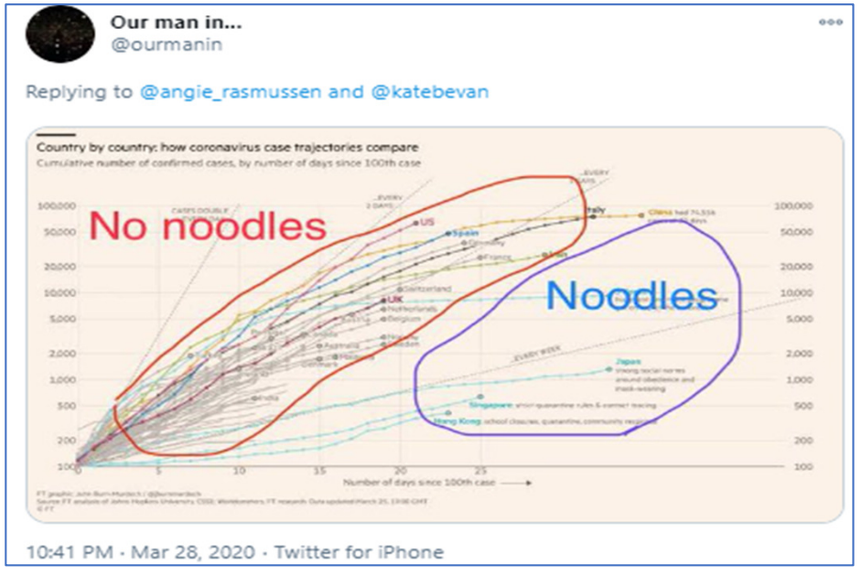


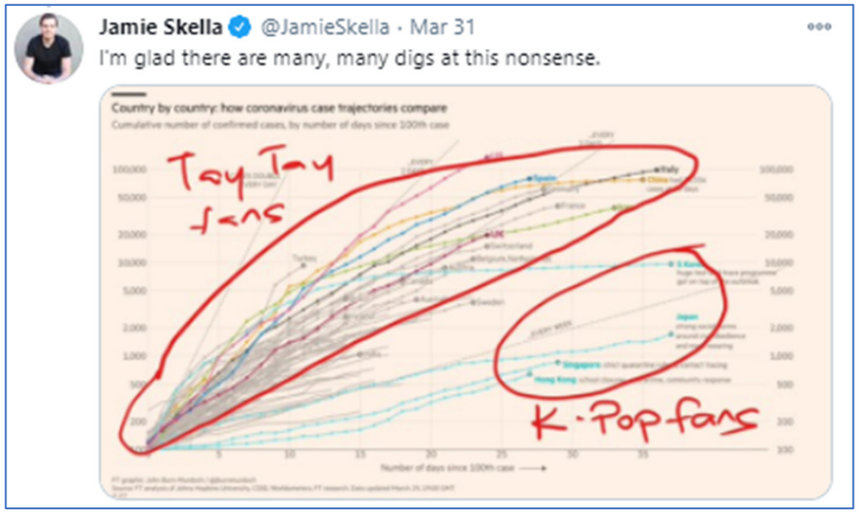
Reprinted from: https://medium.com/nightingale/bad-data-visualization-in-the-time-of-covid-19-5a9f8198ce3e
Original source: https://medium.com/@jperla/how-to-change-global-policy-when-you-are-not-a-billionaire-4ef05aa357c5

- **On the one hand…**
- This is a very effective message. It was retweeted thousands of times and helped change public policy in many cities, states, countries…
- **On the other hand…**
- It goes against one of the great principles of Data Science and Statistics.
- Everyone repeat after me: "Correlation does not equal causation!"
- Example from Danny D. Leybzon Medium.com

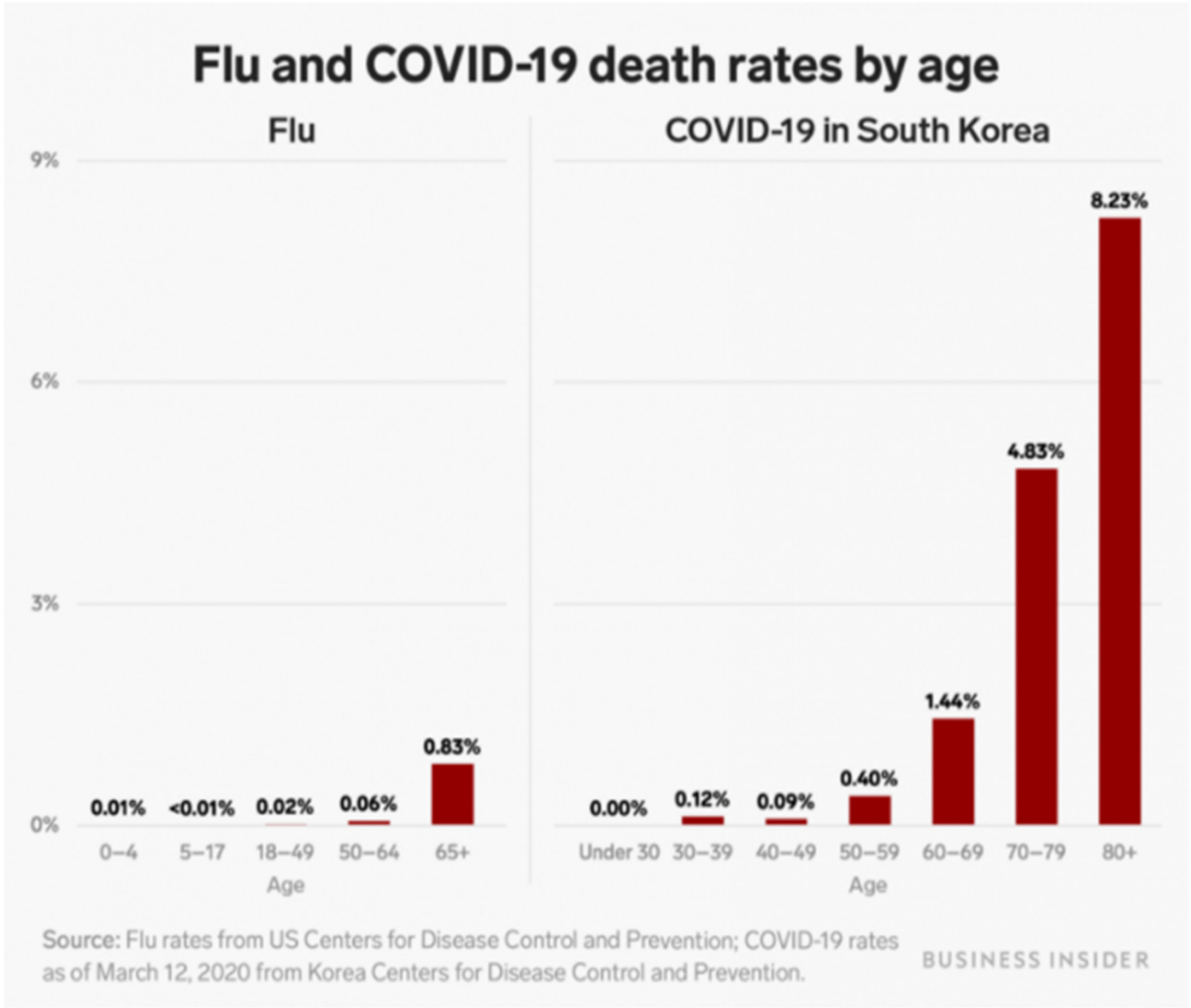
Of course it led to many memes on Twitter







I could show bad visualizations all night, but we'll do just one more…

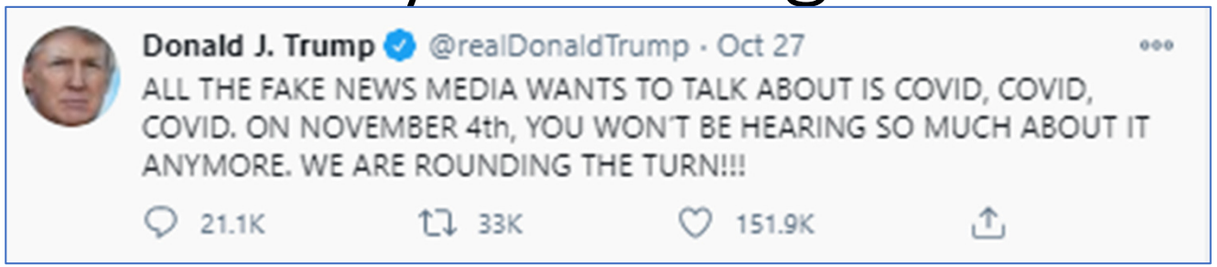


Reprinted from: https://analythical.com/blog/covid19-in-charts
Original Source: Business Insider

- This one was pointed out by Stephen Tracy at Analytical.com
- It's terrible for many reasons
  - First it compares US Flu rates to S.Korean COVID rates. You mean to tell me they couldn't find Korean Flu rates or US COVID rates?
  - Next it uses different groupings (e.g. 65+ versus 60-69, 70-79, 80+)
  - How do we know that there aren't 10 times as many people 60-69 as 70-79 and 80+? That would make the grouped bar on the right under 2 percent for 60+
- This is not just comparing apples and oranges but more like apples and watermelons

Part 1b: Mythbusting with Good visualizations

Mythbusting: We'll start with an easy one!



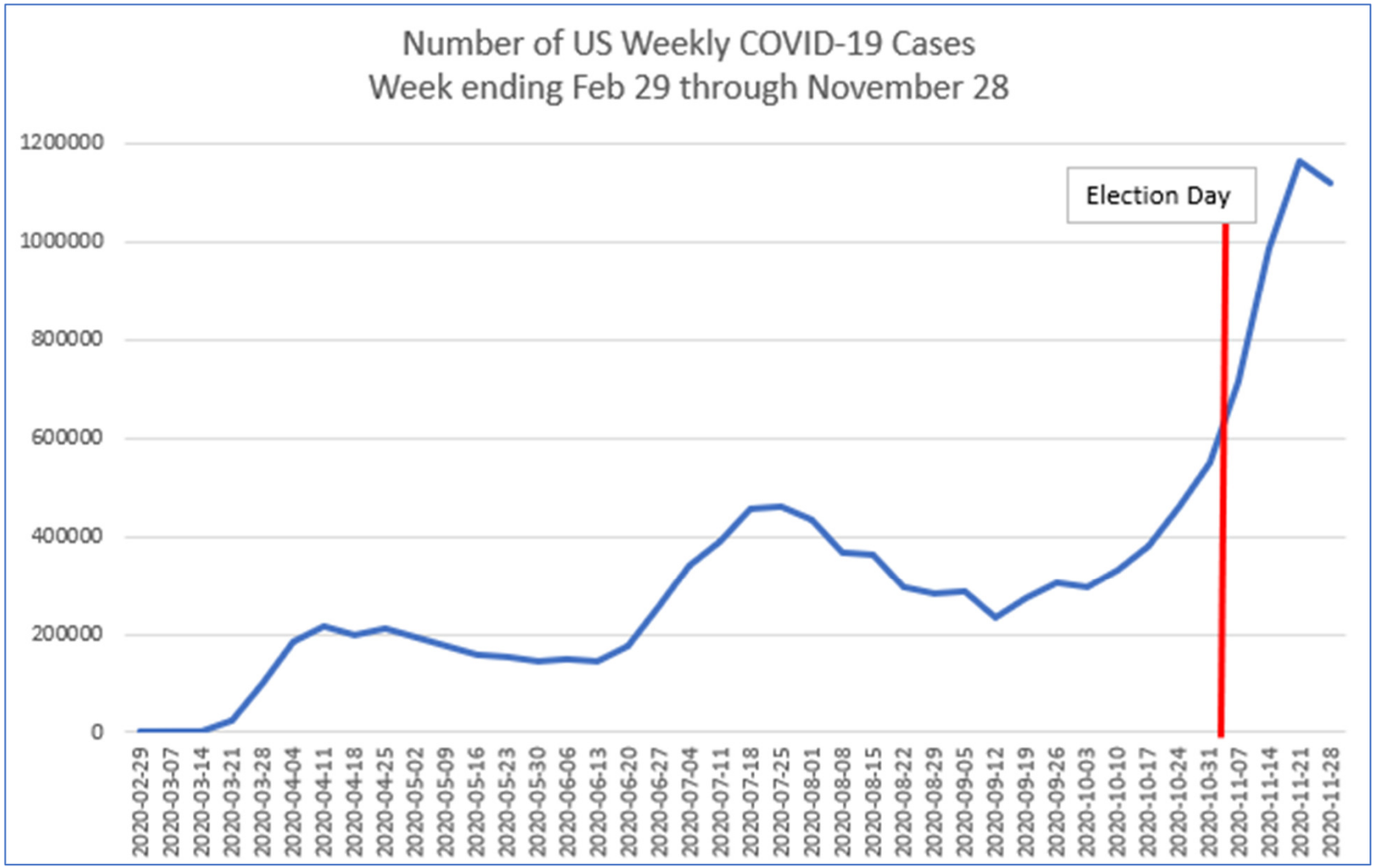


Chart Source: Breitzman 12/2/2020
Data Source: https://covidtracking.com/data/api

Mythbusting: Rounding the Corner

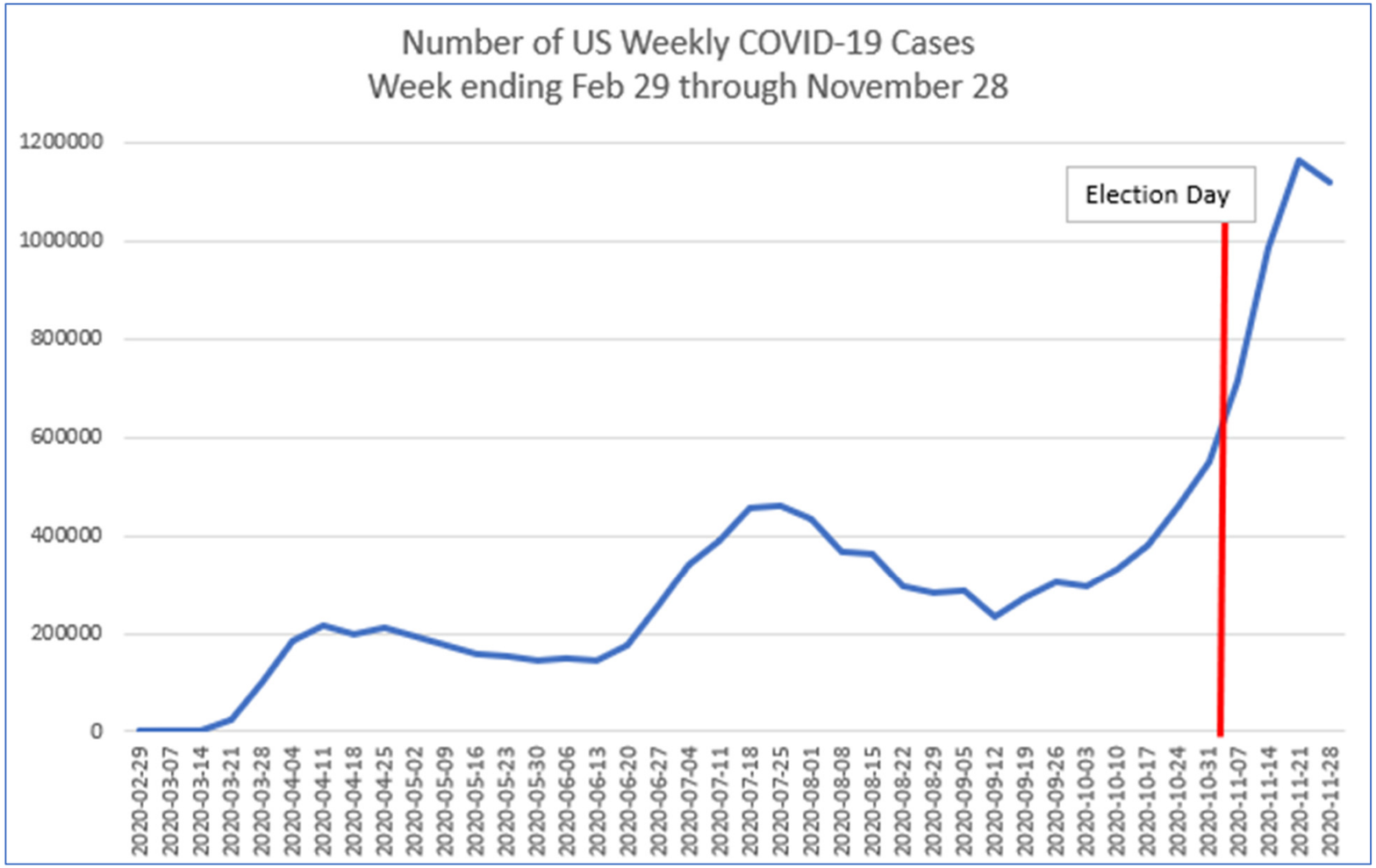


Chart Source: Breitzman 12/2/2020
Data Source: https://covidtracking.com/data/api

I don't want to make this all about the President, but we have speeches in which he said 'we are rounding the turn' or 'rounding the corner' on:

9/4/2020
9/10/2020
9/12/2020
9/13/2020
9/15/2020
9/16/2020
9/19/2020
9/21/2020
10/15/2020
10/17/2020
10/19/2020
10/21/2020
10/23/2020
10/26/2020

Mythbusting: We're in a 3rd wave.

- Or is it a second wave?
- Or is it still the first wave?
- Or is a third surge in the first wave?
- I've seen headlines for all.
- There is certainly evidence for 3 surges; Let's explore…

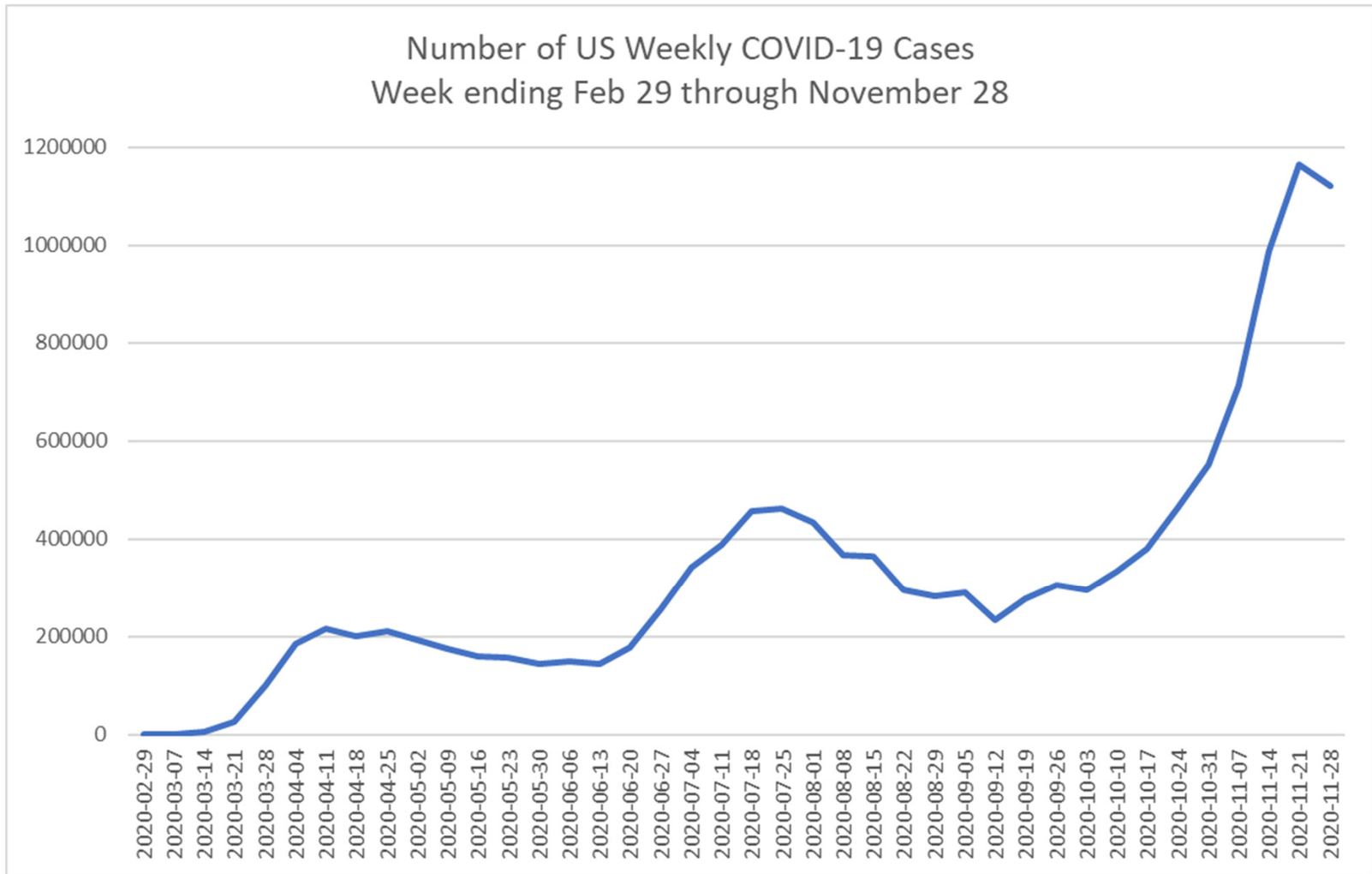
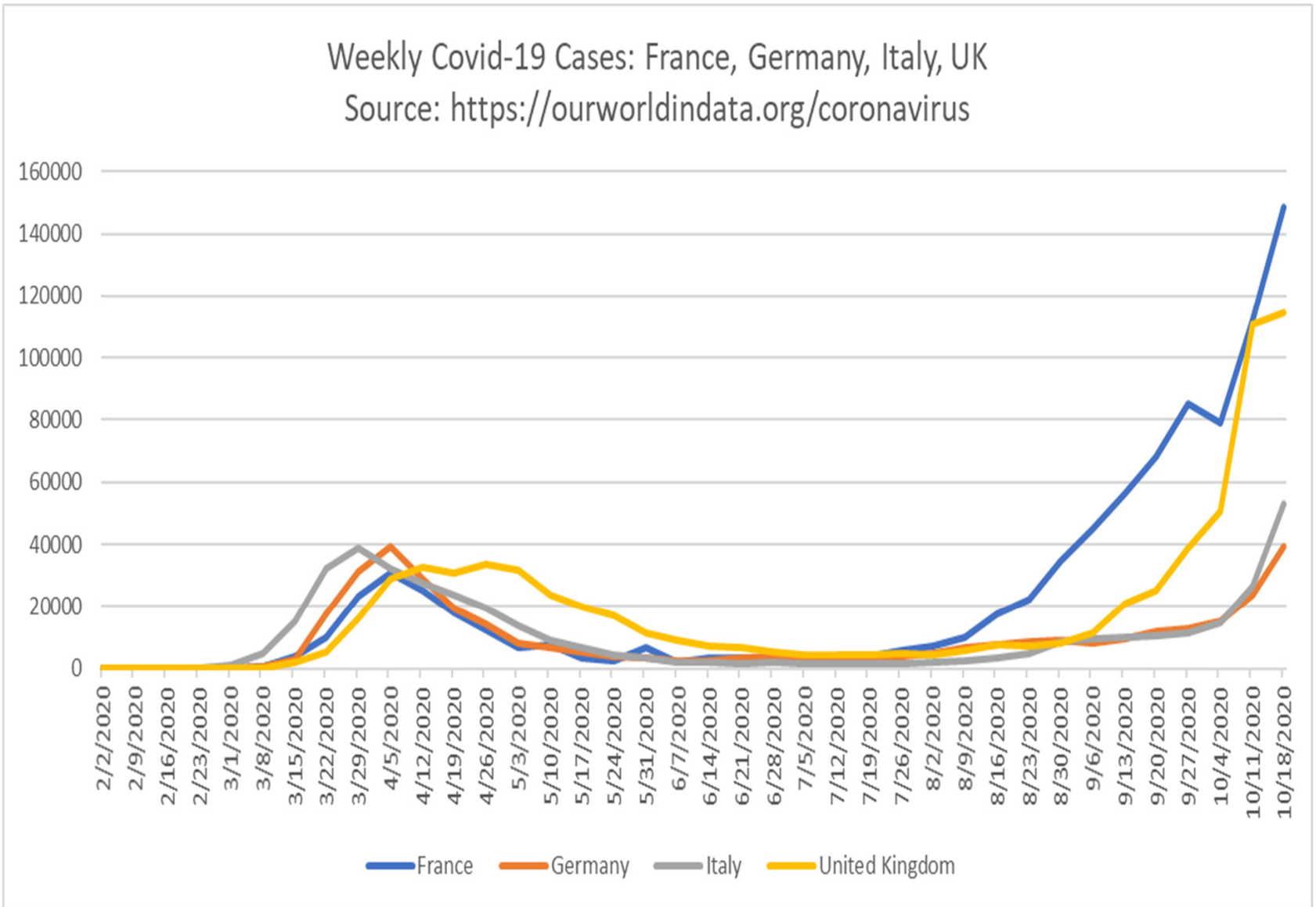


Chart Source: Breitzman 12/2/2020
Data Source: https://covidtracking.com/data/api

Just to clarify: This is what a second wave looks like



- Europe is in a second wave
  - First wave
  - Big lull where cases go down to almost zero
  - Second wave

- If we go back to the US chart on previous slide we see that there was never a lull where cases went near zero

Sometimes it can be instructive to look at a bad visualization!

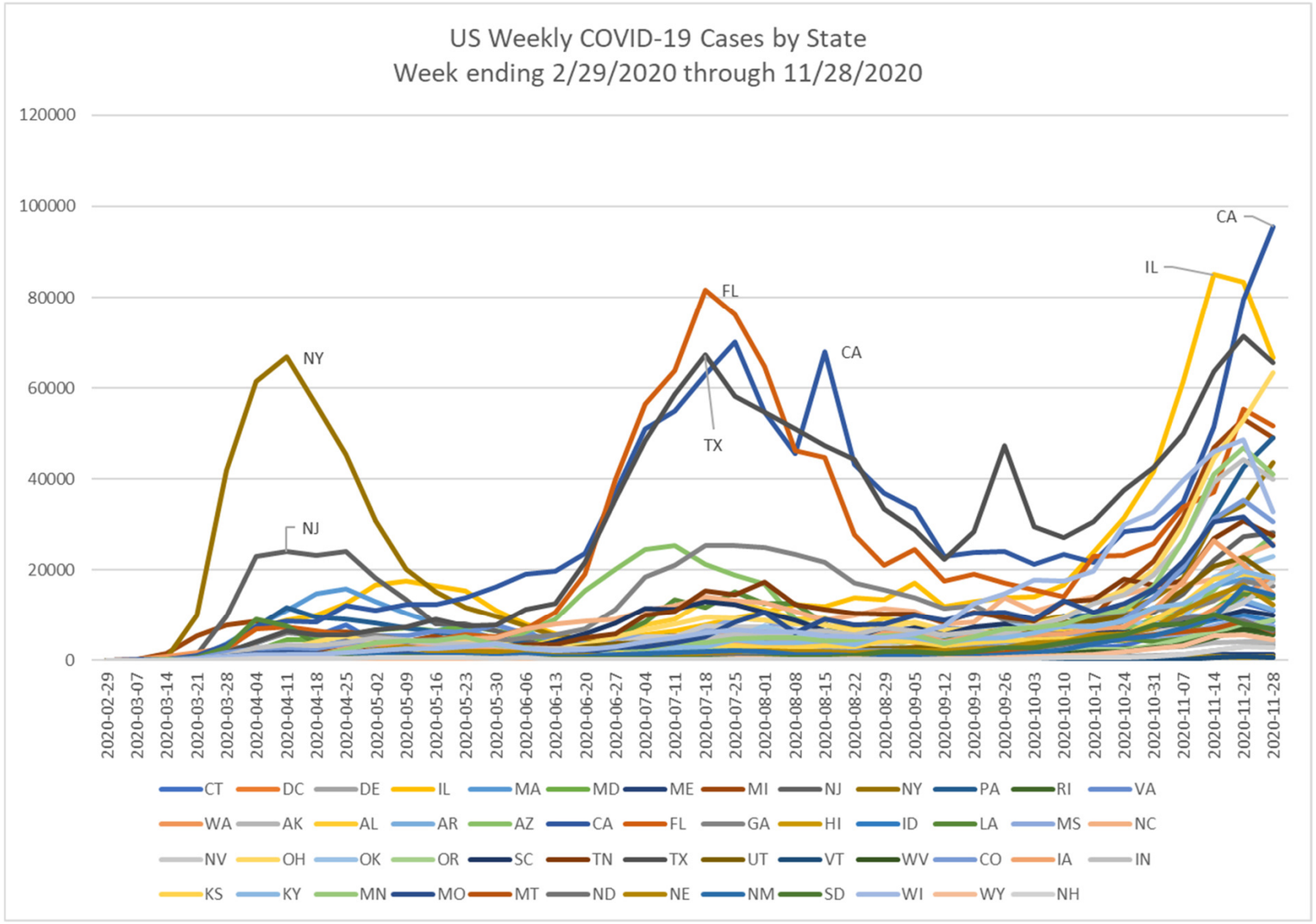


Chart Source: Breitzman 12/2/2020
Data Source: https://covidtracking.com/data/api

- I won't pretend that this is a good visualization, but it is instructive
- We see the states driving the individual surges
- Northeast was behind the first surge
- FL, TX, CA behind the second
- Several states didn't have a big surge until September

Previous Chart Split into Three…

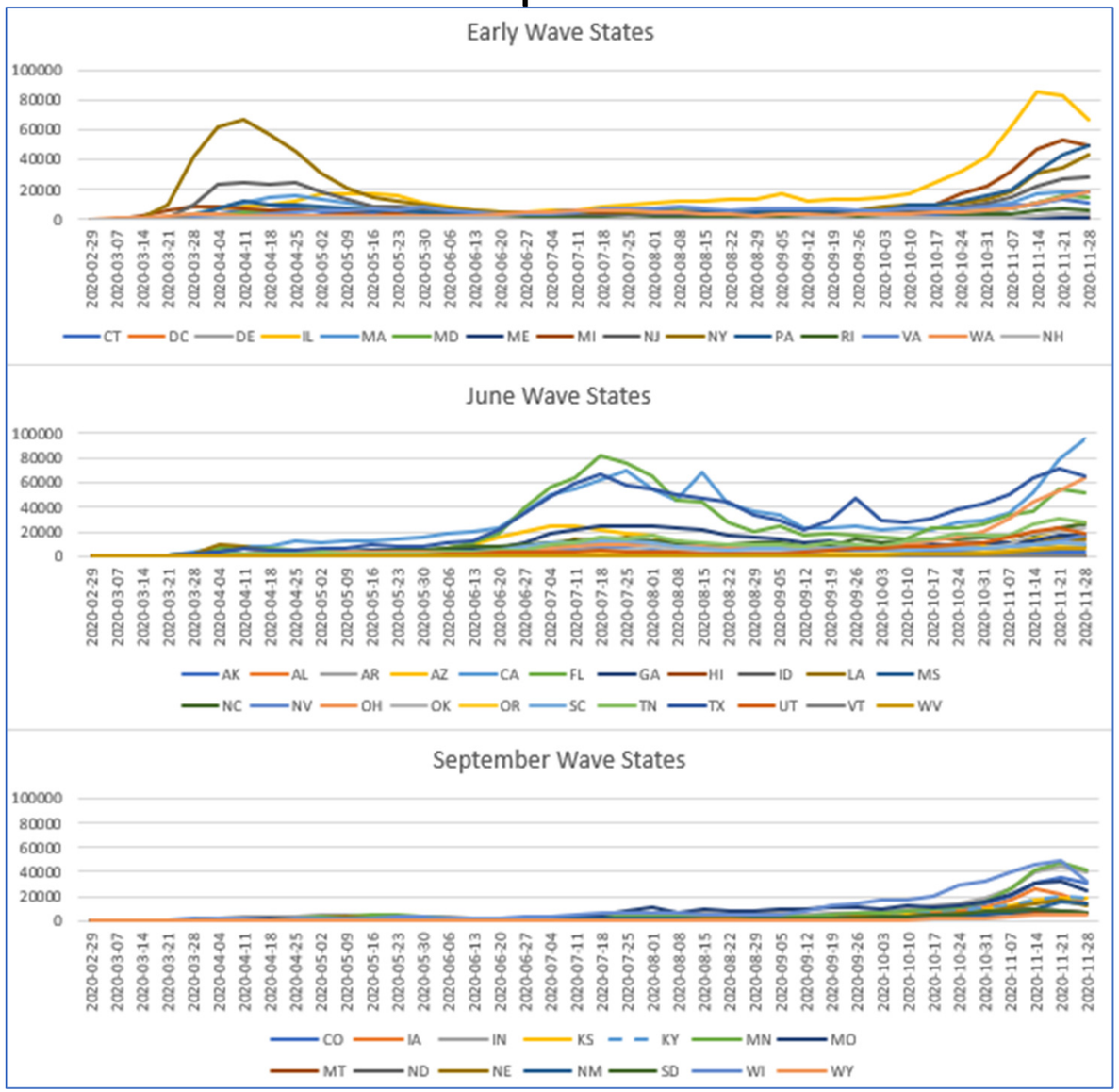


Chart Source: Breitzman 12/2/2020
Data Source: https://covidtracking.com/data/api

- This visualization is not much better, but it gives some clarity
- The Northeast is in a second wave (it had the lull, like the countries in Europe)
- California, Florida, Texas, etc. had a late wave that drove the second ‘surge’
- Other states didn’t start to get significant numbers of cases until September

A Cleaned-up Visualization

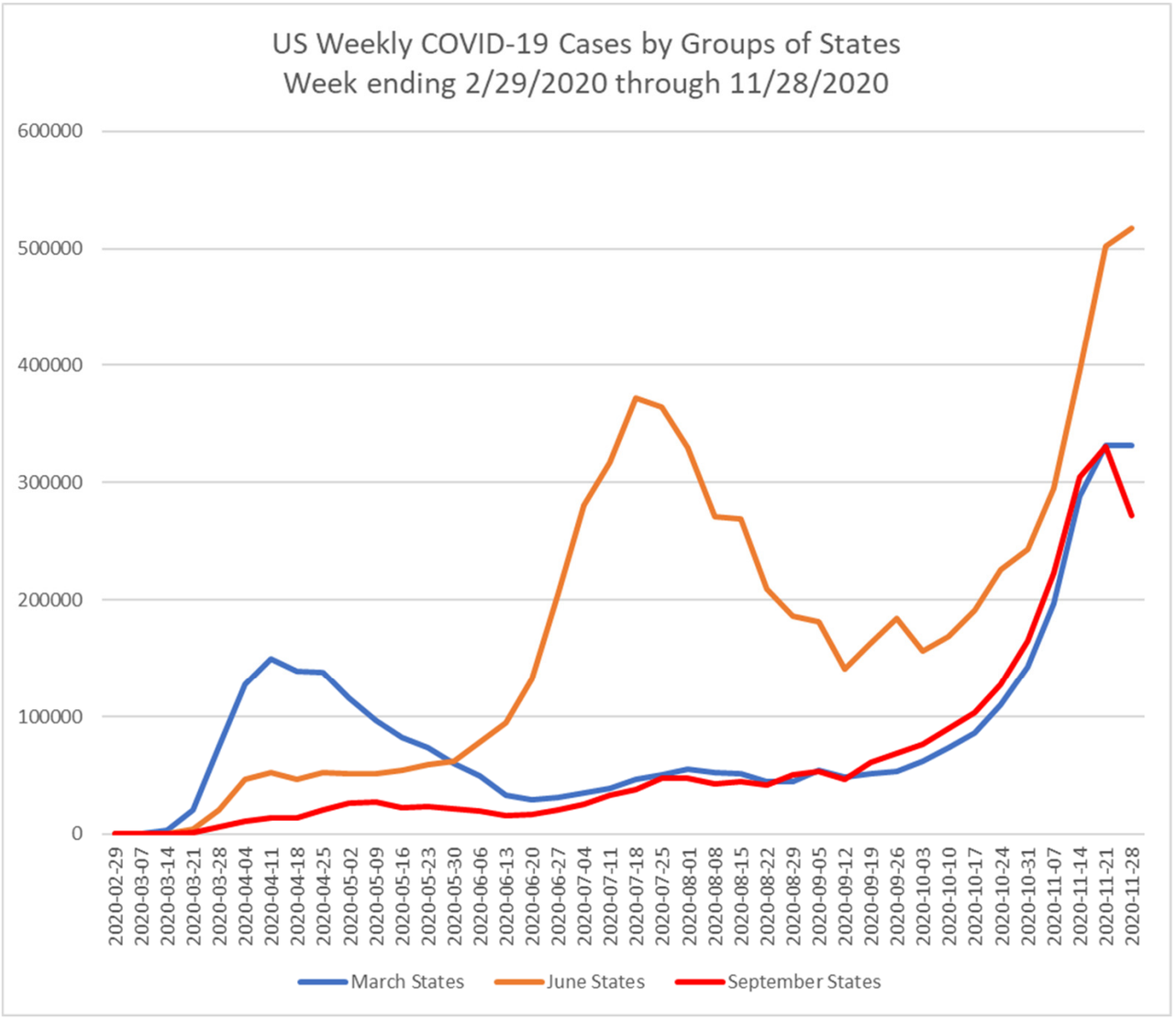
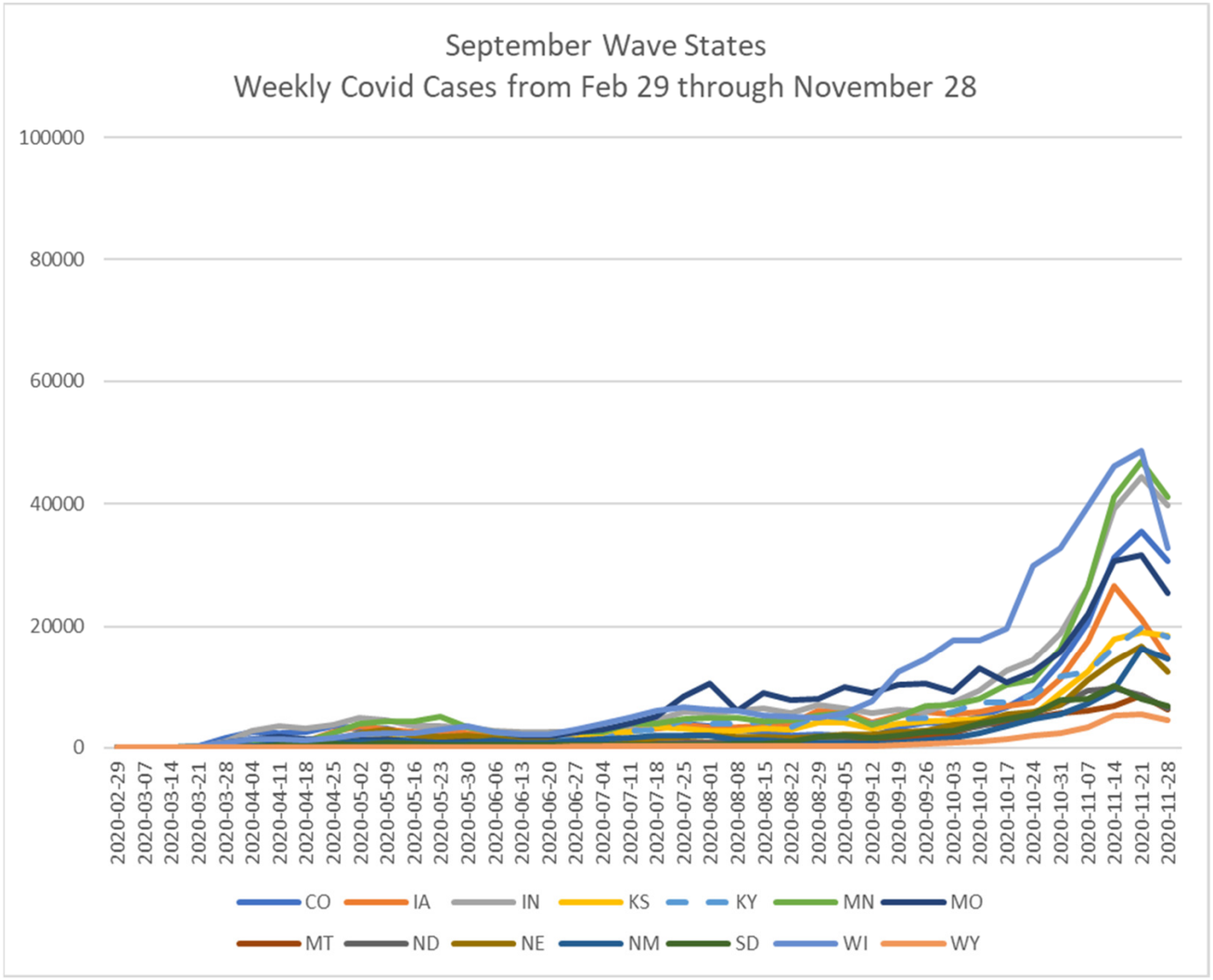


Chart Source: Breitzman 12/2/2020
Data Source: https://covidtracking.com/data/api

- If we stack these 3 lines on top of each other we get the whole US chart that looks like 3 surges
- The 3 alleged surges are really different regions having their different peaks
- The current time period looks really bad, because all 3 regions are having a ‘surge’ at the same time in November. (It just happens to be the first for some regions, and the second in others)

Another item worth noting



- This chart showing the states having a late first wave suggests that people from certain rural states are not irrational!

- It's no wonder that states like Kentucky didn't want the economy closed in March and April. They didn't start seeing a lot of cases until recently.

Chart Source: Breitzman 12/2/2020
Data Source: https://covidtracking.com/data/api

Another item worth noting

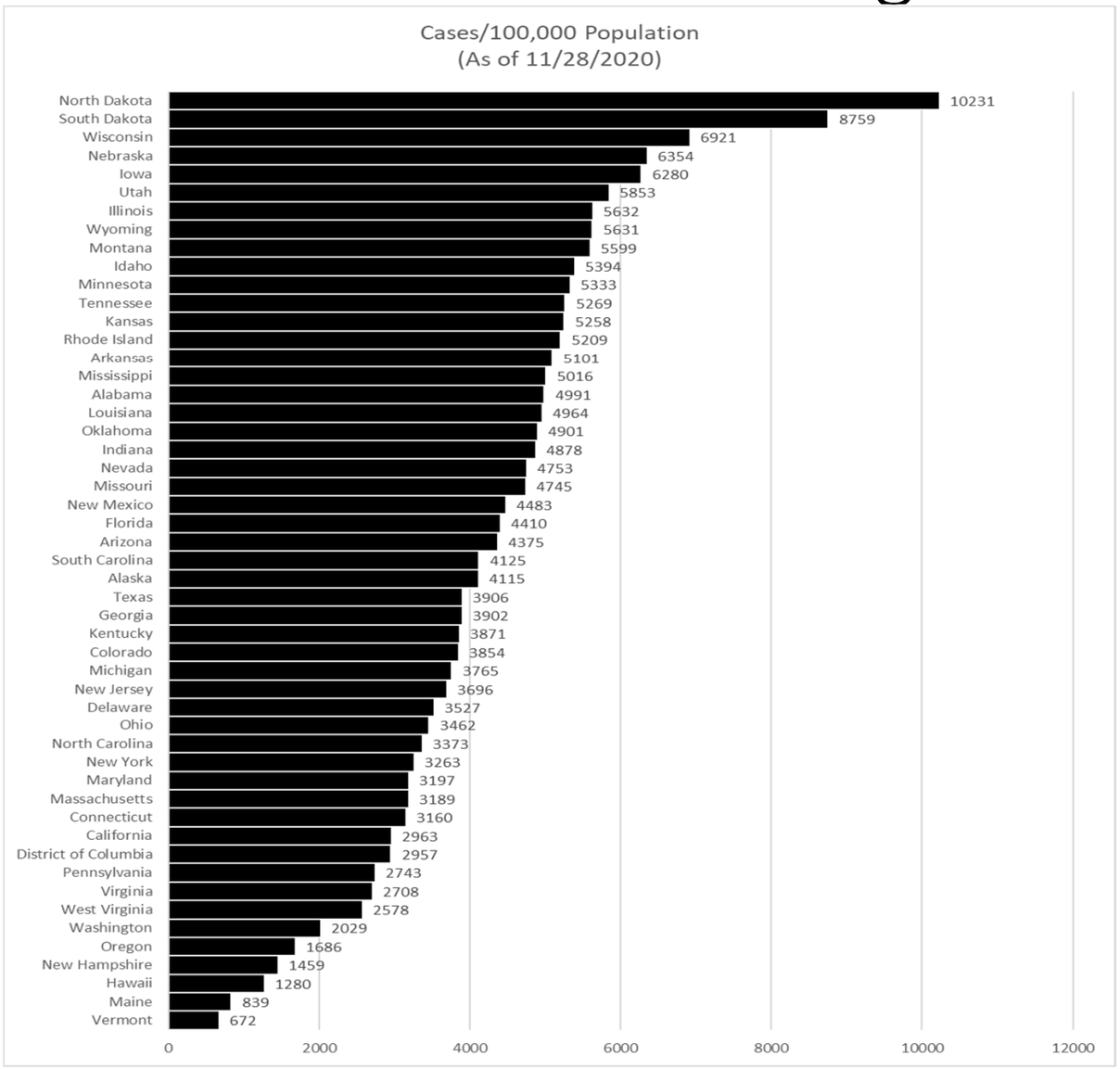


Chart Source: Breitzman 12/2/2020
Data Source: https://covidtracking.com/data/api Plus US Census for Population Data

- The previous charts were not normalized by Population
- We add this chart where we look at cumulative cases per 100,000 residents

Just for fun we repeat the last slide with colors based on the Presidential Election

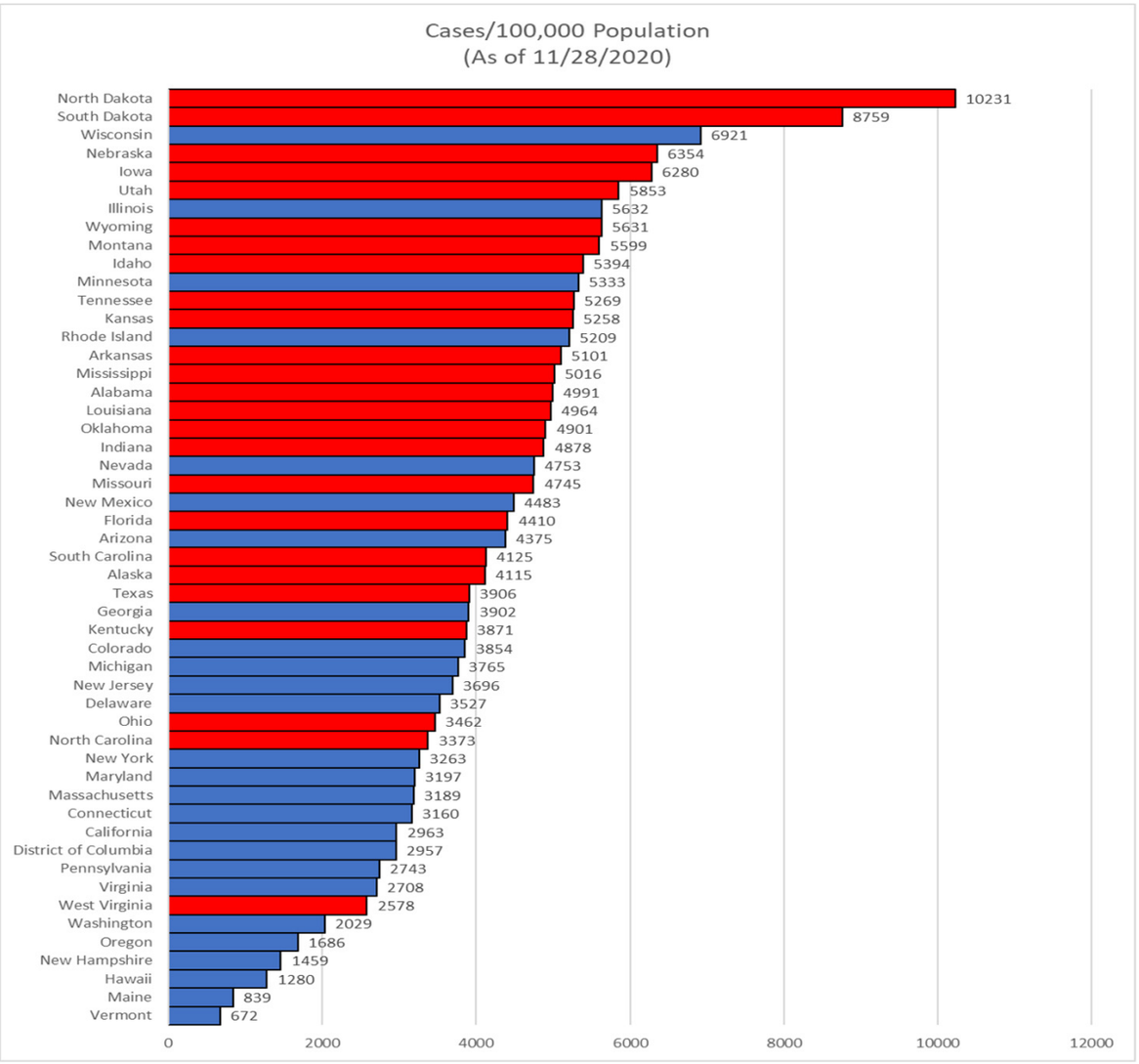


Chart Source: Breitzman 12/2/2020
Data Source: https://covidtracking.com/data/api Plus US Census for Population Data

- I know correlation doesn't equal causation, but red states tend to be more resistant to masks and social distancing
- This is not quite fair since many of the states should be colored purple
- California for example has more Republican voters than any other state including Texas (they just happen to be outnumbered by Democrats in the state)

Back to Mythbusting: More testing leads to more cases.

- This one can be busted in multiple ways
- Notice if we plot cases versus tests for All US that we see testing rising even in times of declining cases
- This is not as clean as it could be because of the multiple surge artifacts discussed previously
- We'll make it clearer by using NY as an example (next slide)

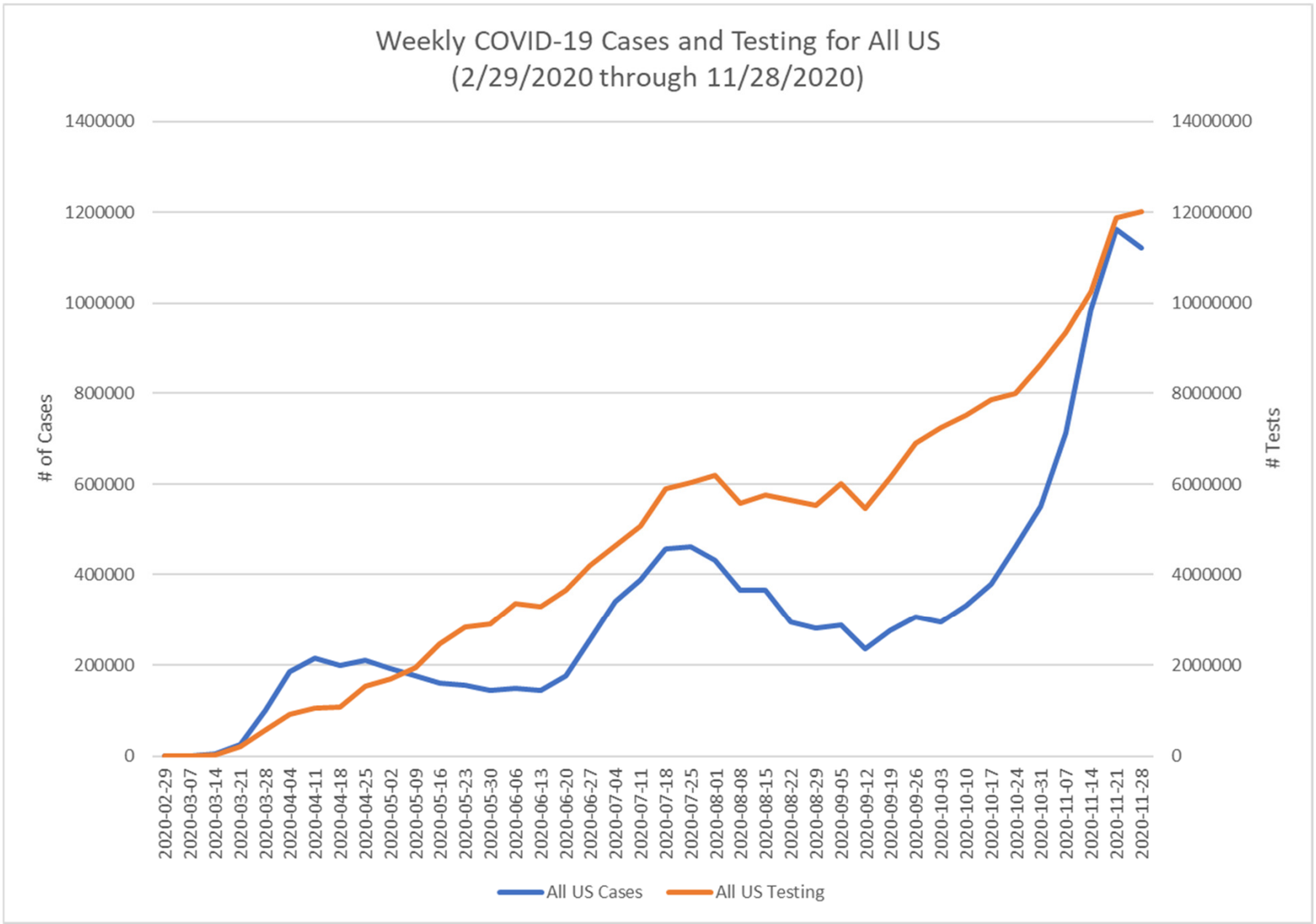


Chart Source: Breitzman 12/2/2020
Data Source: https://covidtracking.com/data/api

Mythbusting: More testing leads to more cases.

- NY is instructive (other states are similar)
- Note that testing continued to rise even through months of declining cases

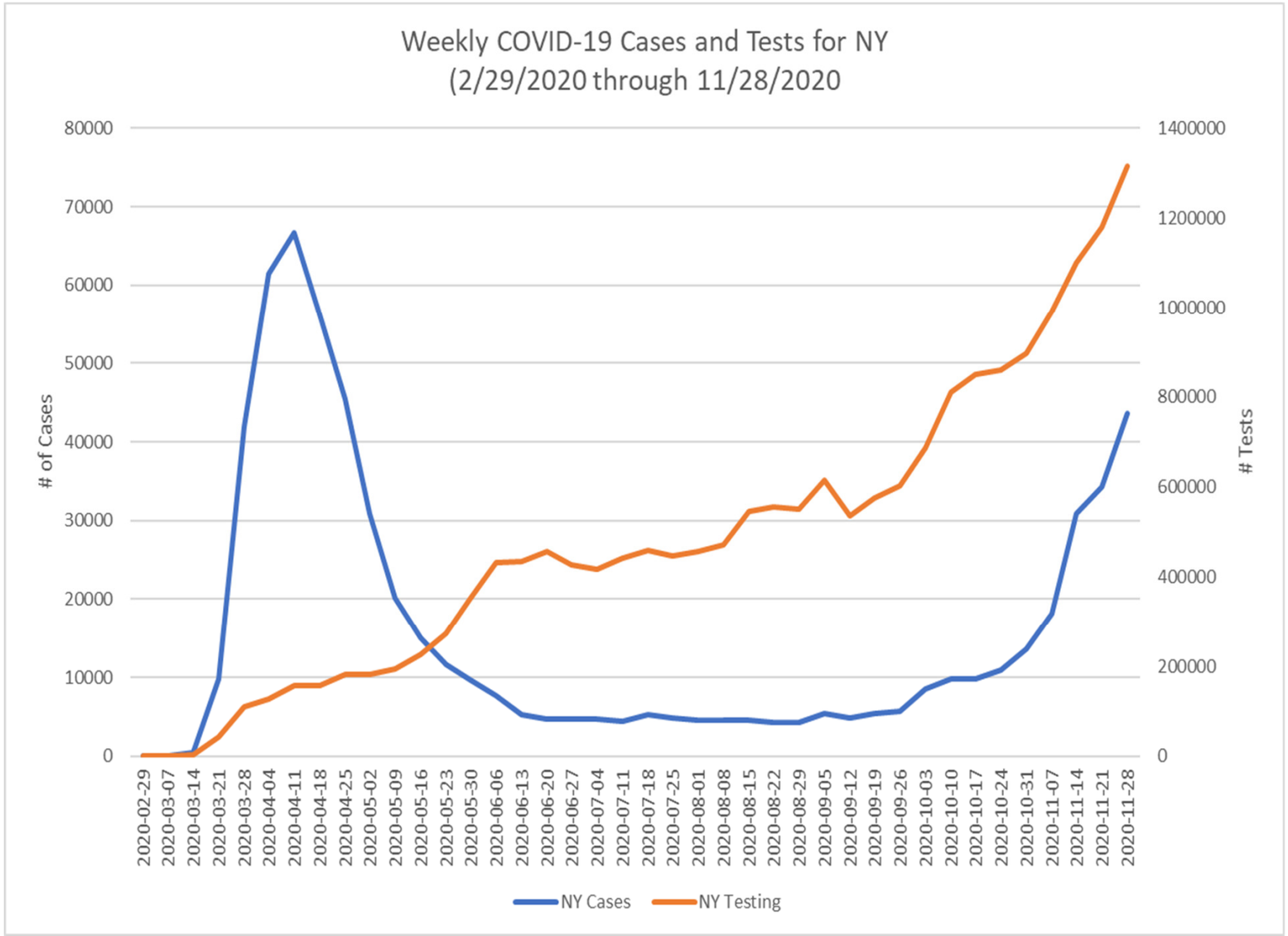
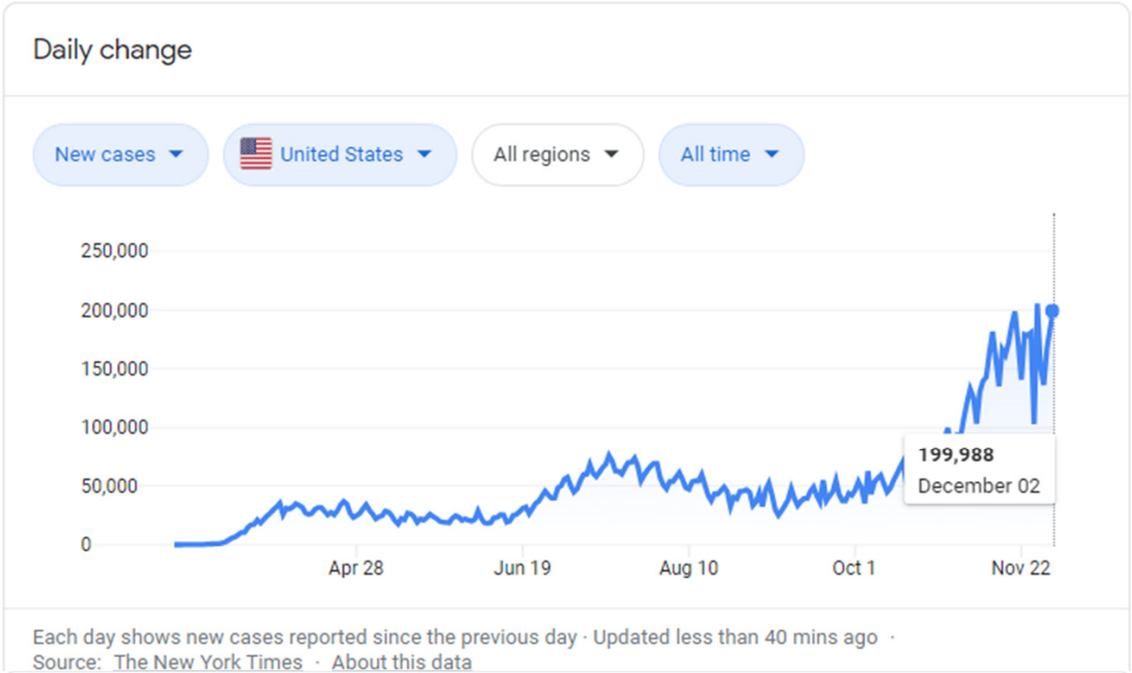


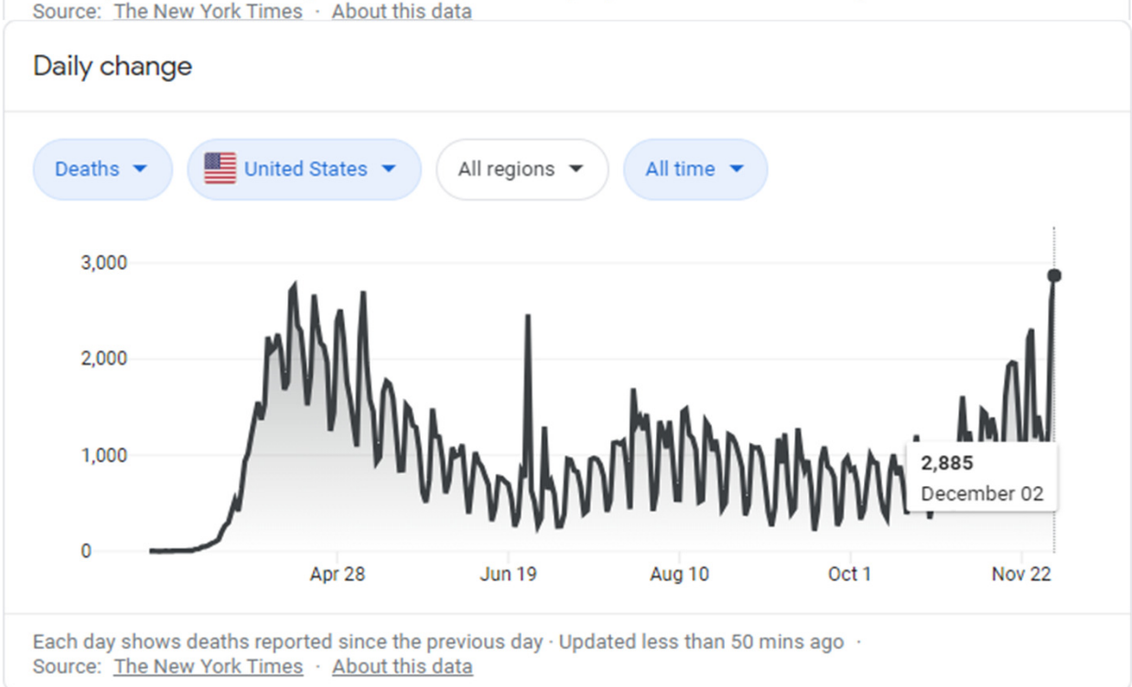
Chart Source: Breitzman 12/2/2020
Data Source: https://covidtracking.com/data/api

Mythbusting: More testing leads to more cases.

- There are other ways to bust this myth, but we don’t have time:
  - Other countries showed that increased testing led to decreased cases
  - Noting that hospitalizations rise when cases rise in each state and people don’t go to the hospital because of a positive test, but because they are sick
  - Noting that the positive test rate goes down when testing is increased
  - Etc. Etc.

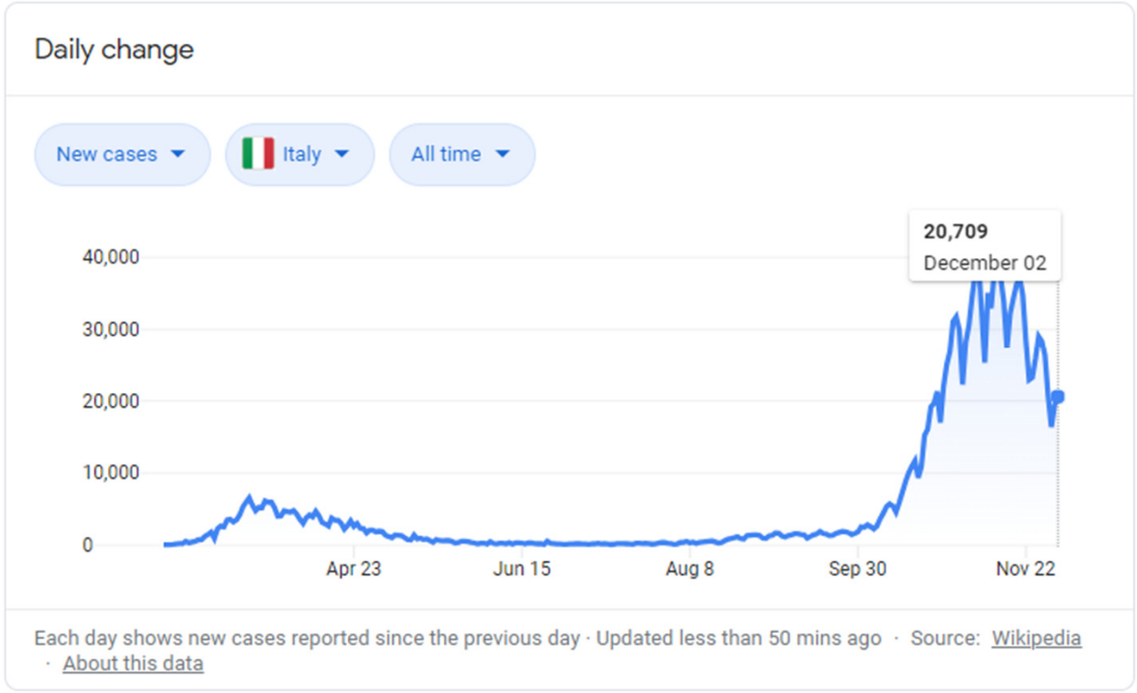
Part 2: Good News and How Data Scientists are Contributing

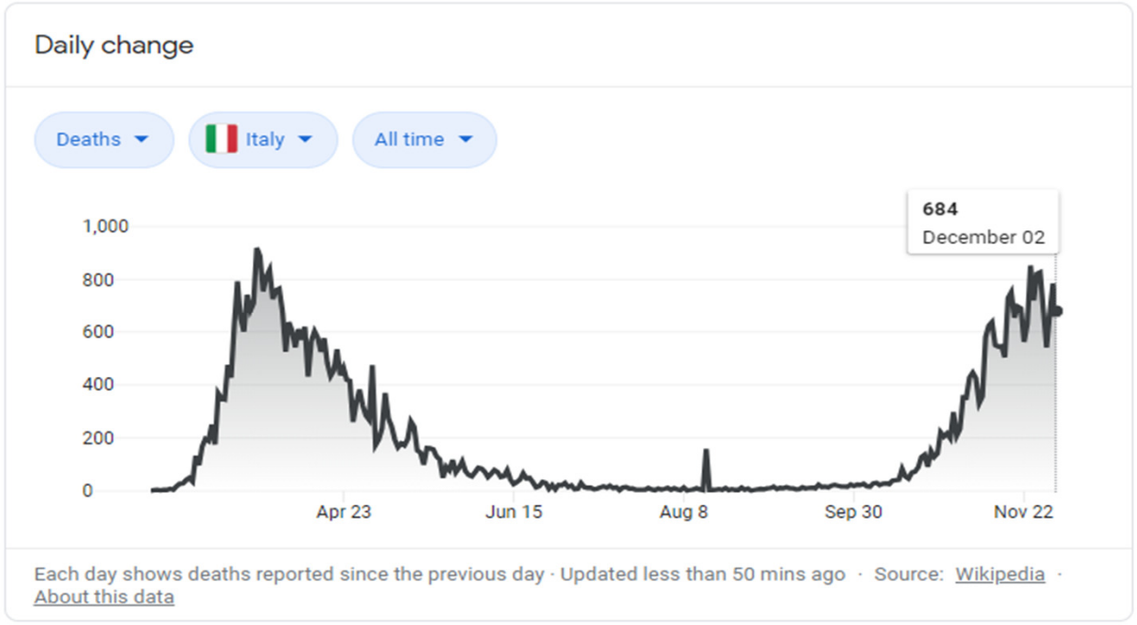




Good news (sort of)

- We have gotten better at treating COVID-19
- While the number of cases is 4 times higher than in April, the number of deaths is roughly equal (i.e Mortality rate is lower)
- While almost 3,000 deaths per day is an appalling number, it would be 4 times as bad if best practices hadn't improved (e.g., reposition patients to avoid ventilators, improved medicine cocktails, etc.)



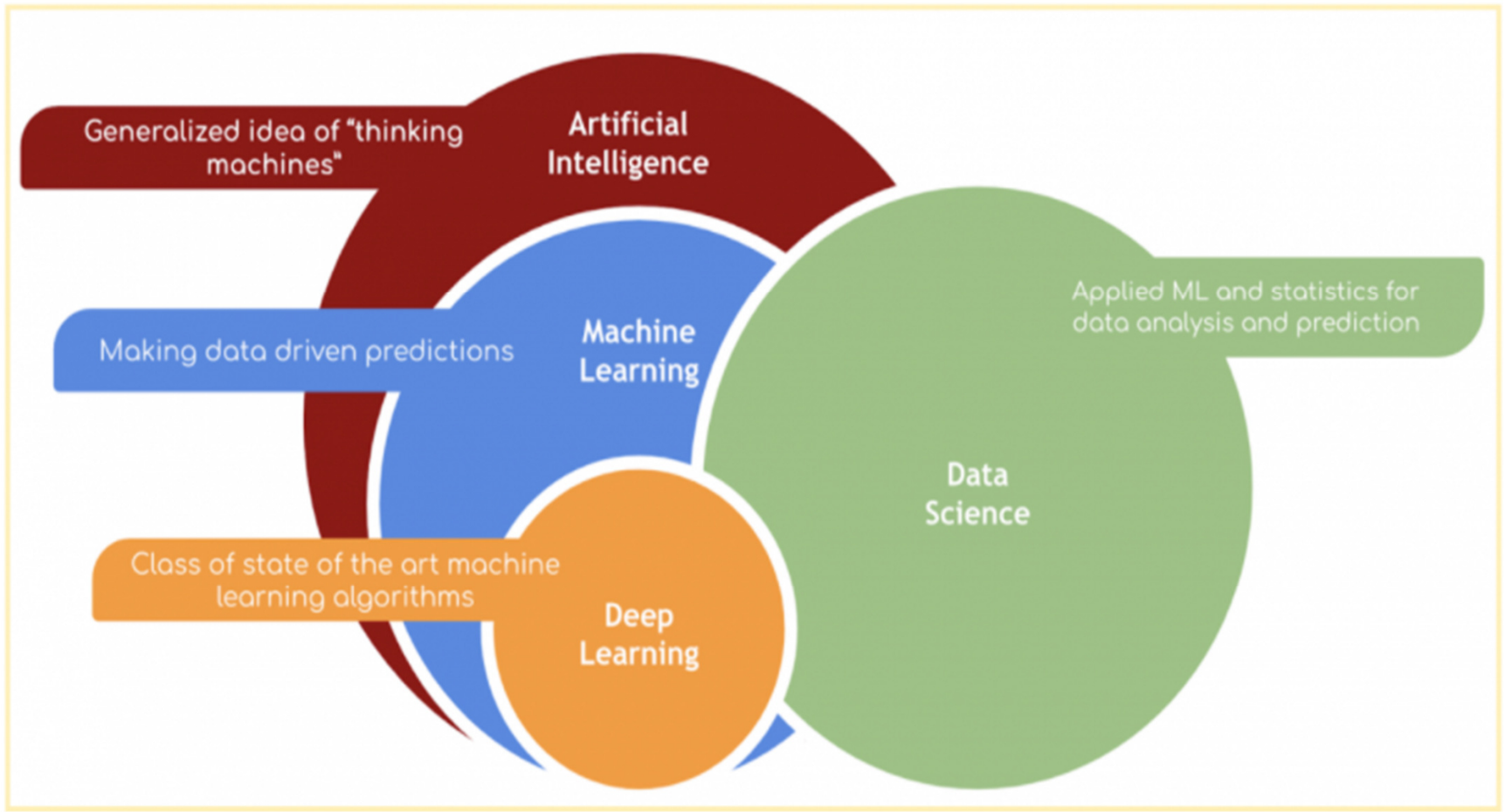


Good news (sort of)

- Same is true worldwide
- Italy for example was on the news every night in April for its high mortality rate

- During the second wave which had 4 times as many cases, the mortality rate was considerably reduced

For this Next Part We Need to Define Data Science more Broadly



How AI, Machine Learning, Deep Learning and Data science intersect.

Source: https://www.turnitin.com/blog/artificial-intelligence-and-machine-learning-at-turnitin
Inspired by: https://www.aitrends.com/data-science/machine-learning-engineer-vs-data-scientist-who-does-what/

- Data science goes beyond visualization and data analysis
- My type of data science is essentially Applied Machine Learning where we build predictive models using Machine Learning
- Machine Learning is subset of Artificial Intelligence
- Deep Learning is a subset of Machine Learning
- We'll take examples from all of these areas

Example 1: Data Mining Suggests COVID Stability

- The Global Initiative on Sharing All Influenza Data (GISAID) a German based public private partnership has collected 85,000 COVID genome sequences by August 24

- Professor Sergei Pond took part in a big-data analysis of the database and concluded that there are an extremely small number of mutations between March and August.

- “Given that most mutations have no effect and often aren’t transmitted, this should make it relatively easier to create an effective vaccine,” said Pond

- “If there had been a lot of change since then [March], we might have to worry that those early vaccine designes, as well as potential treatments and diagnostics, might not be as effective. But that is not the case.”

- This may suggest that we won’t need a changing vaccine each year as we do for Flu

Source: Temple University College of Science and Technology: Outlook (Fall 2020)

Example 2: Data Mining Leads to COVID 19 Breakthrough

- A supercomputer-powered genetic study of COVID-19 patients has spawned a possible breakthrough into how the novel coronavirus causes disease—and points toward new potential therapies to treat its worst symptoms. The supercomputer crunched data sets representing some 17,000 genetic samples and compared them to some 40,000 genes.
- The genetic data mining research predicts a hyperabundance of bradykinin in a coronavirus patient's body at the points of infection, which can have well-known and sometimes deadly consequences. Extreme bradykinin levels in various organs can lead to dry coughs, myalgia, fatigue, nausea, vomiting, diarrhea, anorexia, headaches, decreased cognitive function, arrhythmia and sudden cardiac death. All of which have been associated with various manifestations of COVID-19.
- Lung-fluid samples from COVID-19 patients consistently revealed overexpression of genes that produce bradykinin, while also underexpressing genes that would inhibit or break down bradykinin.
- The bradykinin genetic discovery points to potential therapies like icatibant, danazol, stanozolol, ecallantide, berinert, cinryze and haegarda, all of whose predicted effect is to reduce bradykinin levels in a patient. Even Vitamin D, whose observed deficiency in COVID-19 patients is also explained by the group's research, could play a role in future COVID-19 therapies.
- **None of these, it's important to stress, has yet been tested in clinical trials. But the data-miner's job is to find potential life-saving patterns that lead to discoveries. Not to test them!**

Source: https://spectrum.ieee.org/the-human-os/computing/hardware/has-the-summit-supercomputer-cracked-the-covid-code

Example 3: Machine learning translation of COVID-19 Research Papers

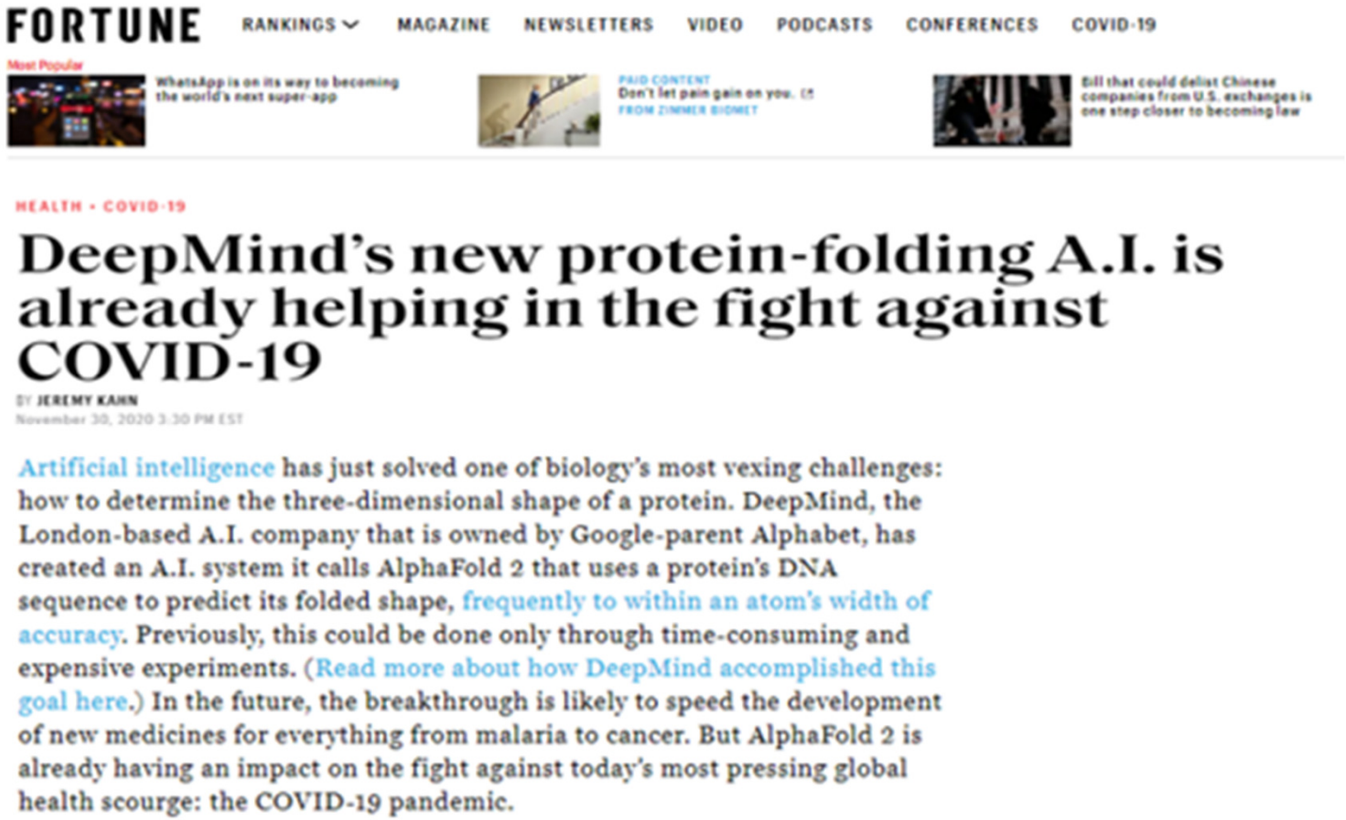
- Through a project known as the Translation Initiative for Covid-19 (TICO-19). Translations Without Borders is working with researchers at Carnegie Mellon and a who's who of major tech companies including Microsoft, Google, Facebook, and Amazon (with the notable exception of Apple) to translate Covid-related materials in 36 languages through these companies' networks of translators (and on their dimes).

- The next stage will be to repurpose this newly translated material as training data—the massive amounts of text and recordings needed in each language as raw materials for tools like machine translation and automatic speech recognition

- Publishers have made a large number of COVID research articles freely available, however many need to be translated into different languages to be useful. This project aims to make that easier

Source: Gretchen McCulloch, Wired, https://www.wired.com/story/covid-language-translation-problem/

Example 4: Deepmind and Protein Folding



FORTUNE RANKINGS MAGAZINE NEWSLETTERS VIDEO PODCASTS CONFERENCES COVID-19

Most Popular

WhatsApp is on its way to becoming the world's next super-app

PAID CONTENT
Don't let pain gain on you.
FROM ZIMMER BIOMET

Bill that could delist Chinese companies from U.S. exchanges is one step closer to becoming law

HEALTH • COVID-19

DeepMind's new protein-folding A.I. is already helping in the fight against COVID-19

BY JEREMY KAHN
November 30, 2020 3:30 PM EST

Artificial intelligence has just solved one of biology's most vexing challenges: how to determine the three-dimensional shape of a protein. DeepMind, the London-based A.I. company that is owned by Google-parent Alphabet, has created an A.I. system it calls AlphaFold 2 that uses a protein's DNA sequence to predict its folded shape, frequently to within an atom's width of accuracy. Previously, this could be done only through time-consuming and expensive experiments. (Read more about how DeepMind accomplished this goal here.) In the future, the breakthrough is likely to speed the development of new medicines for everything from malaria to cancer. But AlphaFold 2 is already having an impact on the fight against today's most pressing global health scourge: the COVID-19 pandemic.

Source: https://fortune.com/2020/11/30/covid-protein-folding-deepmind-ai/

- Professor Venki Ramakrishnan, Nobel Laureate and president of the Royal Society, said: "This computational work represents a stunning advance on the protein-folding problem, a 50-year-old grand challenge in biology.
- "It has occurred decades before many people in the field would have predicted.
- "It will be exciting to see the many ways in which it will fundamentally change biological research."
- Source: https://www.independent.co.uk/life-style/gadgets-and-tech/protein-folding-ai-deepmind-google-cancer-covid-b1764008.html

HEALTH IT ANALYTICS
xtelligent HEALTHCARE MEDIA

Google Gives $8.5M to Fund COVID-19 Data Analytics, AI Projects

The donation will support 31 organizations around the world in using artificial intelligence and data analytics to better respond to COVID-19.

Source: https://healthitanalytics.com/news/google-gives-8.5m-to-fund-covid-19-data-analytics-ai-projects

By Jessica Kent

September 14, 2020 - Google.org **is donating** more than $8.5 million to 31 universities, nonprofits, and other academic institutions that are using artificial intelligence and data analytics to combat COVID-19.

For more coronavirus updates, visit our resource page, updated twice daily by Xtelligent Healthcare Media.

The funding is part of Google.org's $100 million commitment to COVID-19 relief and focuses on four key areas where new information and action is needed to mitigate the impact of the pandemic.

These areas will include projects centered around monitoring and **forecasting disease spread**, which will lead to a better understanding of where the virus is likely to spike.

"Understanding the spread of COVID-19 is critical to informing public health decisions and lessening its impact on communities," Mollie Javerbaum, program manager of Google.org, and Meghan Houghton, university relations program manager wrote in a recent blog.

"We're supporting the development of data platforms to help model disease and projects that explore the use of diverse public datasets to more accurately predict the spread of the virus."

Organizations conducting projects in this area include Carnegie Mellon University, where researchers will inform public health officials with interactive maps that display real-time COVID-19 data from multiple sources.

Additionally, a team from Boston Children's Hospital, Oxford University, and Northeastern University will build a platform to support accurate public health data for researchers, public health officials, and citizens.

The grants will also fund projects that aim to improve **health equity** and minimize secondary effects of the pandemic.

Summary

- In this brief talk we've shown how data analysis can be used to convey trends and bust myths
- We've also shown multiple ways it can be used badly
- Finally, we showed just a few of the many ways that data science, machine learning, and AI are being used to help in the fight against COVID-19

Thank you! Any Questions?

Anthony Breitzman, PhD

Associate Professor of Computer Science

Data Science Program Coordinator - Rowan University

Director of Research – 1790 Analytics, LLC

Breitzman@rowan.edu

Rowan University