Big Data Final Project

•••

Group 3 Immortals:

Hongyu Zhai, Yuhan Chen, Sifan Chen

Research Questions

- Which countries/regions are slow/fast to take actions?
- Are there any country/region that is ignoring the rising numbers?
- Are there any country/region being extra cautious?
- What patterns can we find when examining the data?
    - Do countries/regions with low medical resources tend to take more stringent actions?
    - Do countries/regions with high population density tend to take more stringent actions?
    - Do countries/regions with higher percentage of elder people tend to take more stringent actions?
- What about states? Can we find similar patterns in the state level?

Government Responses Data

- Country Level
  - Oxford University - Coronavirus Government Responses Tracker: https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker
- State Level
  - Kaiser Family Foundations - State Data and Policy Actions to Address Coronavirus: https://www.kff.org/health-costs/issue-brief/state-data-and-policy-actions-to-address-coronavirus/

Medical Resources Data

- Country Level
  - World Bank - Hospital Beds (per 1,000 people): https://data.worldbank.org/indicator/SH.MED.BEDS.ZS
  - World Bank - Physicians (per 1,000 people): https://data.worldbank.org/indicator/SH.MED.PHYS.ZS
  - World Bank - Nurses (per 1,000 people): https://data.worldbank.org/indicator/SH.MED.NUMW.P3
  - World Bank - Percentage of Ages 65+: https://data.worldbank.org/indicator/SP.POP.65UP.TO.ZS
- State Level (summing all county level data)
  - Kaiser Health News - Hospital by County: https://khn.org/news/as-coronavirus-spreads-widely-millions-of-older-americans-live-in-counties-with-no-icu-beds
  - Kaiser Health News - ICU Beds by County: https://khn.org/wp-content/uploads/sites/2/2020/03/KHN-ICU-bed-county-analysis_2.zip

Step 1: Preparing the Datasets

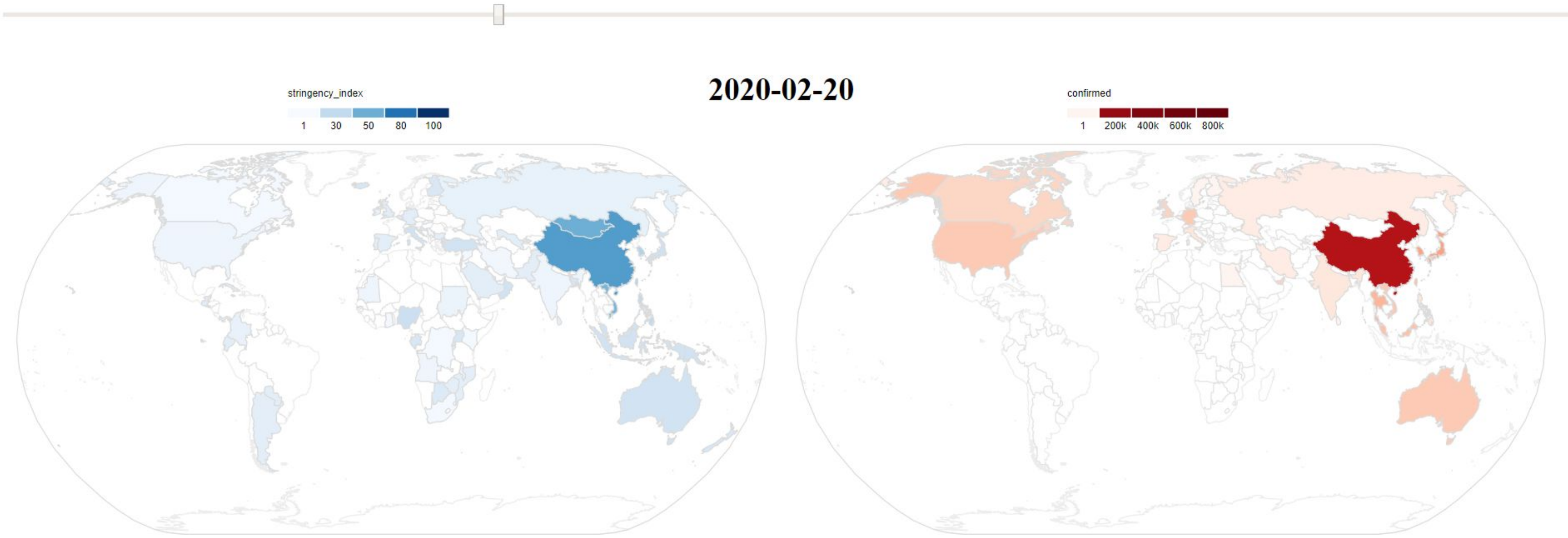
- Download the datasets from the internet.
  - Keep the original filename
  - One folder for each data source
  - Record the date we retrieved the dataset, with the URL to that link (if available).
- Perform necessary cleanup steps
  - Remove the header/footnotes from the table.
  - Every steps detailed in a Jupyter Notebook

    https://github.com/iamzhaihy/BD2020-Final-Project/blob/master/Data%20Processing.ipynb

Step 2: Visualization and Exploring

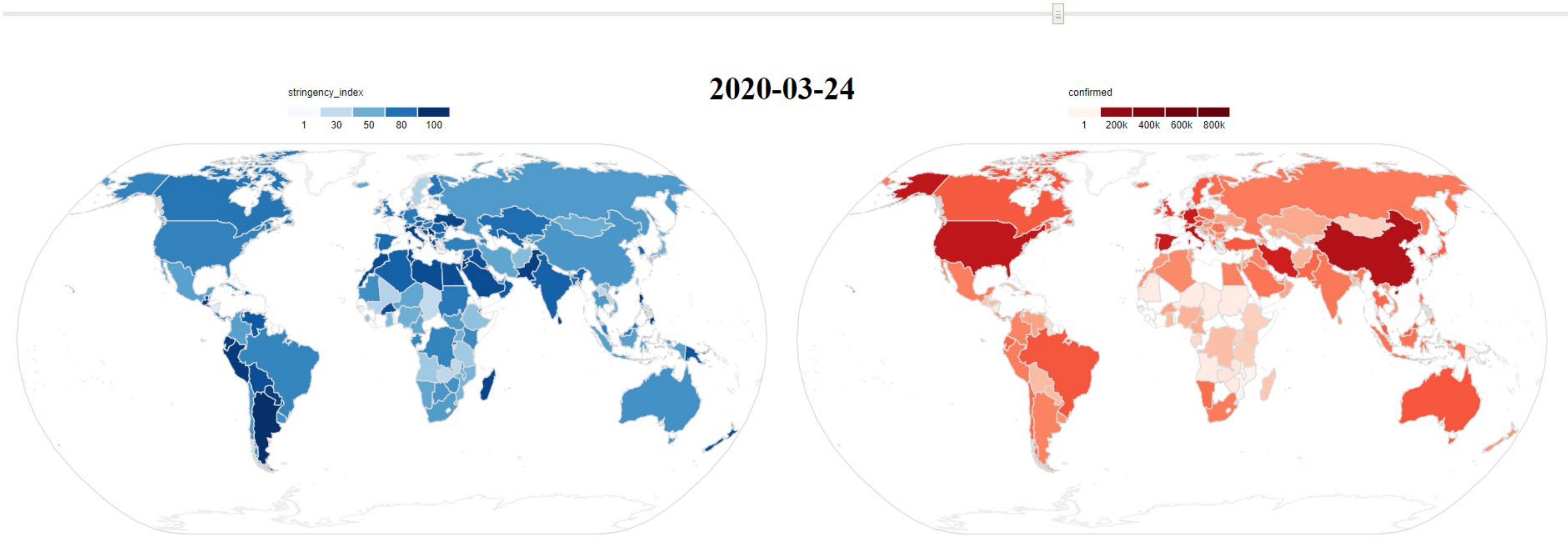
- We made 2 interactive visualizations: one for the countries and one for the states.
- Each visualization shows the stringency_index (a number we used to measure the stringency of government responses) and the number of confirmed cases.
- By using the slider on top, we are able to see those numbers for each day, and observe how things change.
- Let's do a quick live demo https://iamzhaihy.github.io/BD2020-Final-Project

Visualization 1



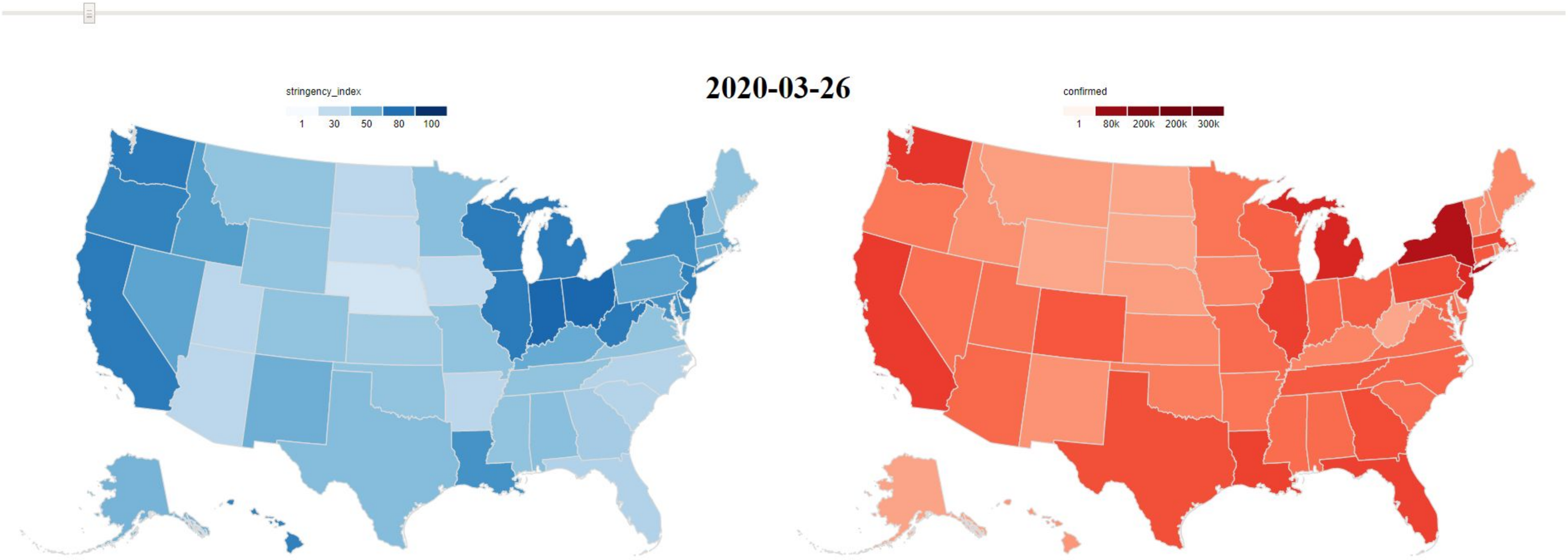
2020-02-20

Visualization 1

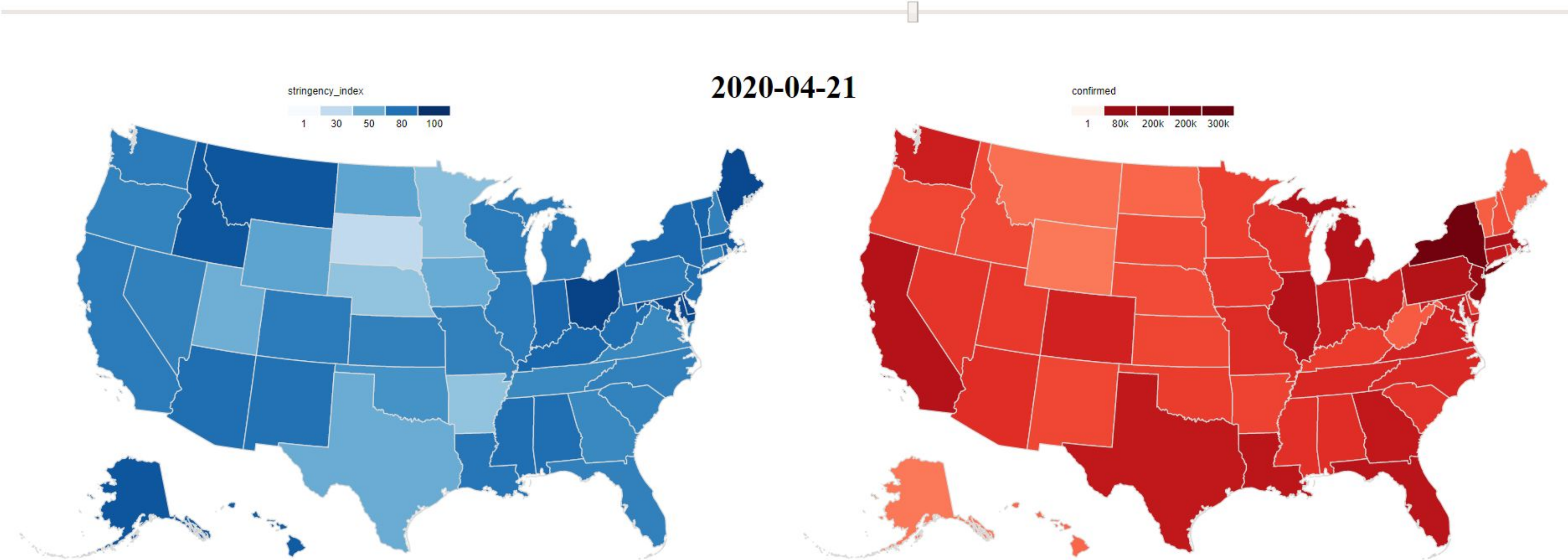


2020-03-24

Visualization 2



Visualization 2



2020-04-21

Step 3: Try to Make Sense of the Data

- Compute correlations between stringency_index and other indicators
    - Generally show weak positive relationship
    - Due to the complex nature of governments, simple correlations cannot tell us much.
    - More details can be found in the project report.
- Observe and try to find the pattern
    - For example, almost all red states are taking less stringent actions
    - One interesting exception is Ohio (fast and stringent).
- It is hard to find a strongly correlated indicator.
    - Many factors are affecting the decisions simultaneously.

Challenges

- Diversity of the datasets
    - Different sources
    - Different granularity
    - Different column names
    - Different keys (country names, country codes, etc.)
- How we dealt with it
    - Manual adjustments
    - Study the datasets and try to find an ideal key to perform join

Challenges

- Lack of state level data
    - Hard to find daily policy changes
    - Sources are diverse and chaotic
    - Need to compute stringency_index by ourselves
- How we dealt with it
    - Utilize the snapshots on archive.org
    - Manual adjustments using information on Wikipedia
    - Came up with our own encoding rules to compute stringency_index for states

Challenges

- Visualizing the data
    - Need GeoJSON data.
    - Static visualization shows too little.
    - Extreme values cause trouble for charts (stretched).
- How we dealt with it
    - Use D3 to make interactive visualizations.
        - Viewers can filter what information to be drawn.
    - Study the materials and find the right scaling function.

Limitations

- stringency_index simplifies things, but also hides nuances.
    - Only one number is used, so details are lost.
    - We cannot answer the questions like: which countries took most extreme actions to restrict international travel.
- stringency_index, as the name suggests, only measures the stringency.
- We do not have daily data on state actions
    - No convenient way to collect.
    - Too labor-intensive for three people.
    - Lack of data means we might miss some important changes.

Limitations

- The indicators we collected are somewhat outdated
  - Data for some countries/regions are last updated more than 10 years ago.
  - We collected and used the best data we can find. JHU and other institutes also rely on the same data.