Collecting and Visualizing COVID-19 Case Count Data from Multiple Open Sources

Ryan Hafen

@hafenstats

https://slides.com/hafen/covid19-casecounts

Epidemic Intelligence from Open Sources (EIOS)

https://www.who.int/eios

- EIOS is a collaboration between various public health stakeholders around the globe, led by WHO
- Mission is early detection, verification and assessment of public health risks and threats using open source information
- Aimed at consolidating a wide array of endeavors and platforms to build a strong public health intelligence (PHI) community supported by robust, harmonized and standardized PHI systems and frameworks across organizations and jurisdictions

COVID-19 Case Counts

- Confirmed cases and deaths at different levels of geographic resolution, as provided by health departments, ministries of health, etc.
- EIOS aims to provide analysts with the ability to:
  - Quickly understand trajectories of counts and related statistics at different levels of geography
  - Observe discrepancies between different data sources
- Case count considerations
  - Methods for counting vary by health care system
  - Level of testing varies geographically and over time

Case Count Sources

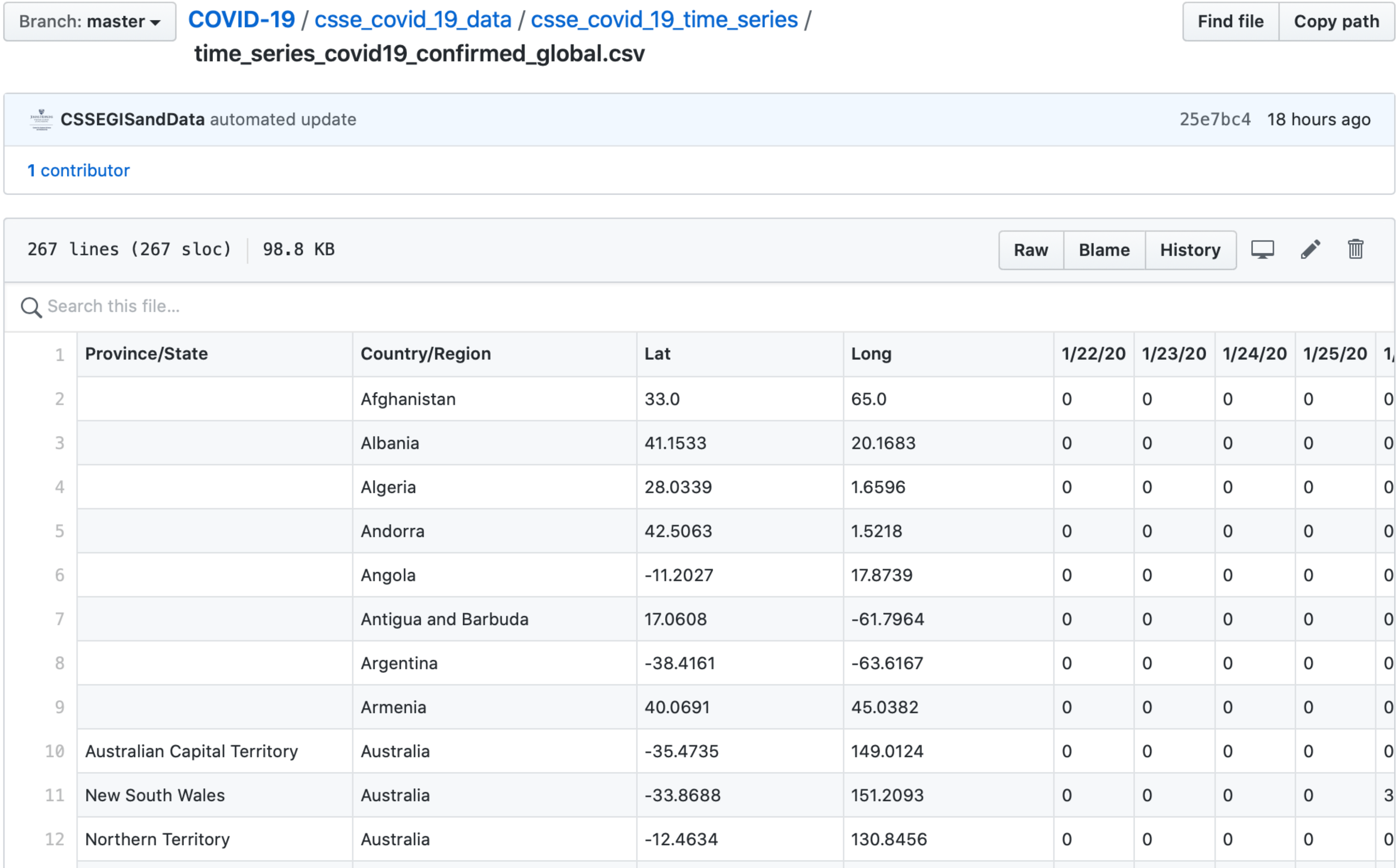
- Global (country-level)
  - Johns Hopkins CSSE
  - European CDC
  - WHO
  - Worldometer
  - Others
- United States (state and county-level)
  - Johns Hopkins CSSE
  - New York Times
  - USA FACTS
  - Others

Challenges of Data Standards in Open Data Communities

- Often not much thought is given to standards
- When it is, everyone has a different idea of "standard"
- Often little incentive to adhere to someone else's standard

It's hard to expect strict adherence to a standard for a given type of data, but ideally we would all adhere to some **best practices**

Example of Bearable Practices - JHU



COVID-19 / csse_covid_19_data / csse_covid_19_time_series / time_series_covid19_confirmed_global.csv

CSSEGISandData automated update 25e7bc4 18 hours ago

1 contributor

267 lines (267 sloc) | 98.8 KB

Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/
	Afghanistan	33.0	65.0	0	0	0	0	0
	Albania	41.1533	20.1683	0	0	0	0	0
	Algeria	28.0339	1.6596	0	0	0	0	0
	Andorra	42.5063	1.5218	0	0	0	0	0
	Angola	-11.2027	17.8739	0	0	0	0	0
	Antigua and Barbuda	17.0608	-61.7964	0	0	0	0	0
	Argentina	-38.4161	-63.6167	0	0	0	0	0
	Armenia	40.0691	45.0382	0	0	0	0	0
Australian Capital Territory	Australia	-35.4735	149.0124	0	0	0	0	0
New South Wales	Australia	-33.8688	151.2093	0	0	0	0	3
Northern Territory	Australia	-12.4634	130.8456	0	0	0	0	0

Example of Bearable Practices

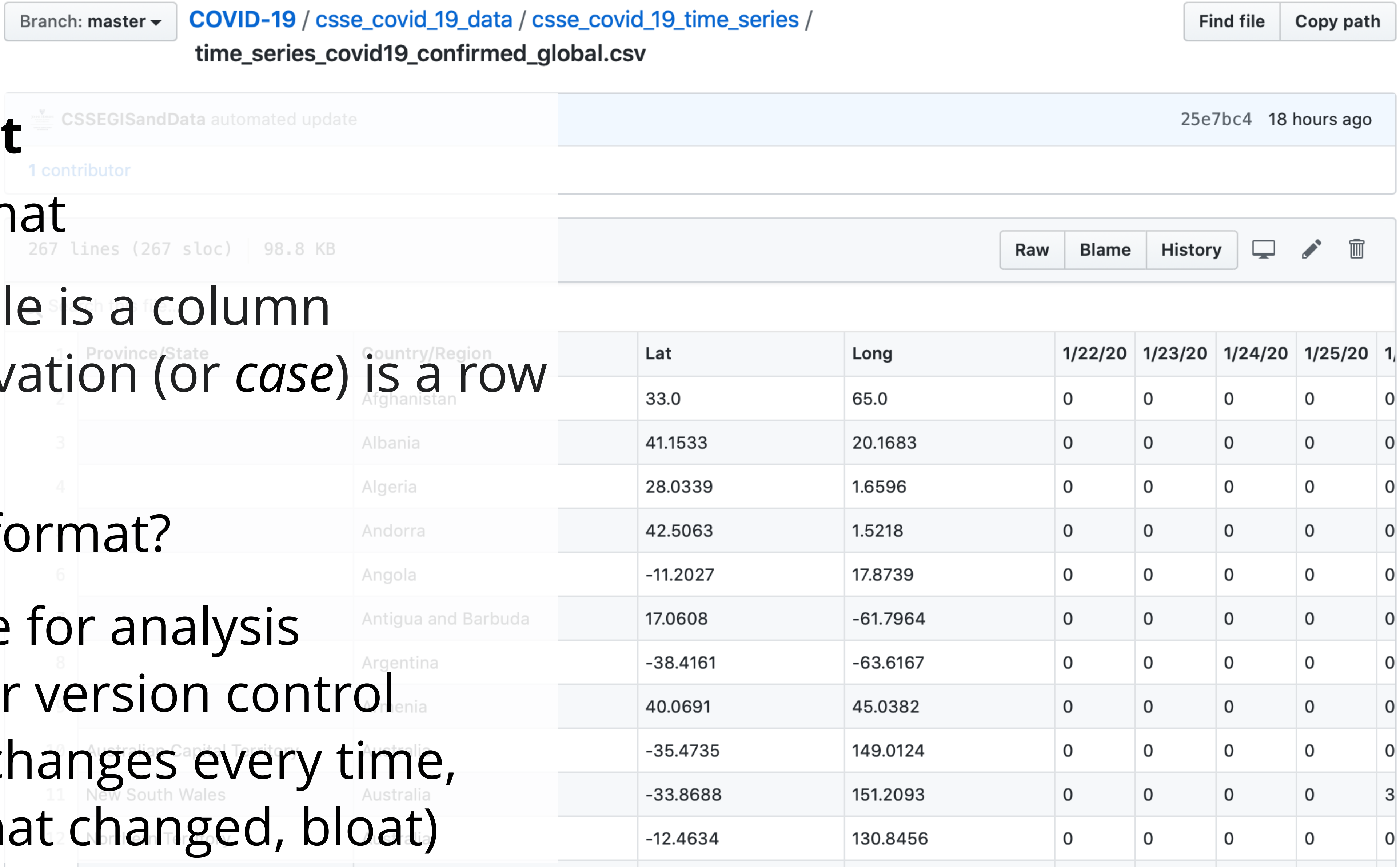
1. Wide format

Prefer *tidy* format

- Each variable is a column
- Each observation (or *case*) is a row

Why not wide format?

- Not suitable for analysis
- Not ideal for version control (every line changes every time, can't tell what changed, bloat)



COVID-19 / csse_covid_19_data / csse_covid_19_time_series / time_series_covid19_confirmed_global.csv

CSSEGISandData automated update 25e7bc4 18 hours ago

1 contributor

267 lines (267 sloc) 98.8 KB

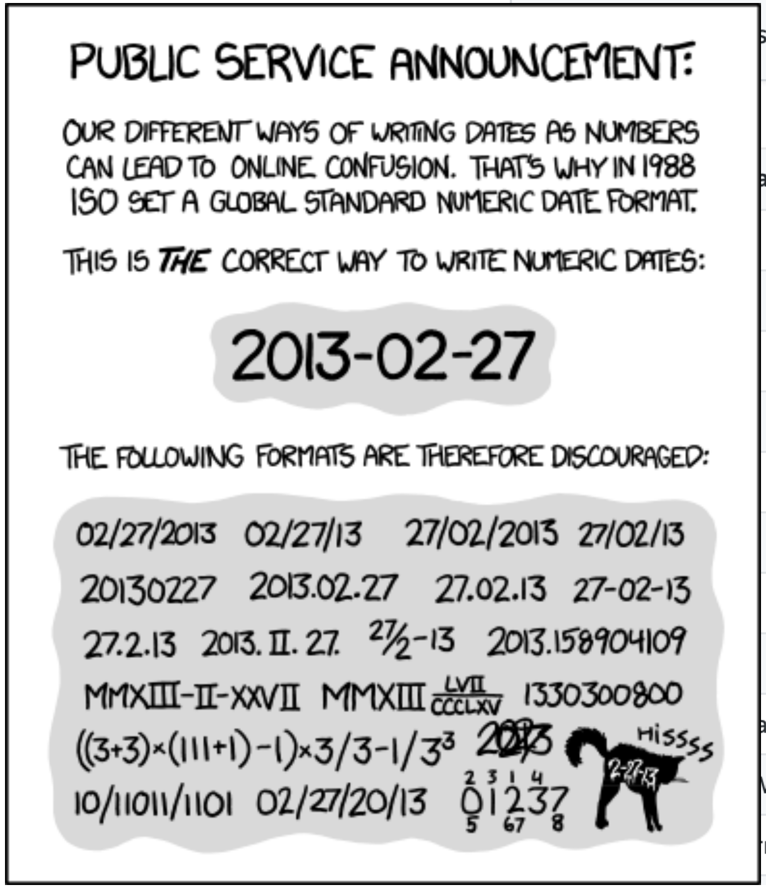
Search this file...

Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/
	Afghanistan	33.0	65.0	0	0	0	0	0
	Albania	41.1533	20.1683	0	0	0	0	0
	Algeria	28.0339	1.6596	0	0	0	0	0
	Andorra	42.5063	1.5218	0	0	0	0	0
	Angola	-11.2027	17.8739	0	0	0	0	0
	Antigua and Barbuda	17.0608	-61.7964	0	0	0	0	0
	Argentina	-38.4161	-63.6167	0	0	0	0	0
	Armenia	40.0691	45.0382	0	0	0	0	0
Australian Capital Territory	Australia	-35.4735	149.0124	0	0	0	0	0
New South Wales	Australia	-33.8688	151.2093	0	0	0	0	3
Northern Territory	Australia	-12.4634	130.8456	0	0	0	0	0

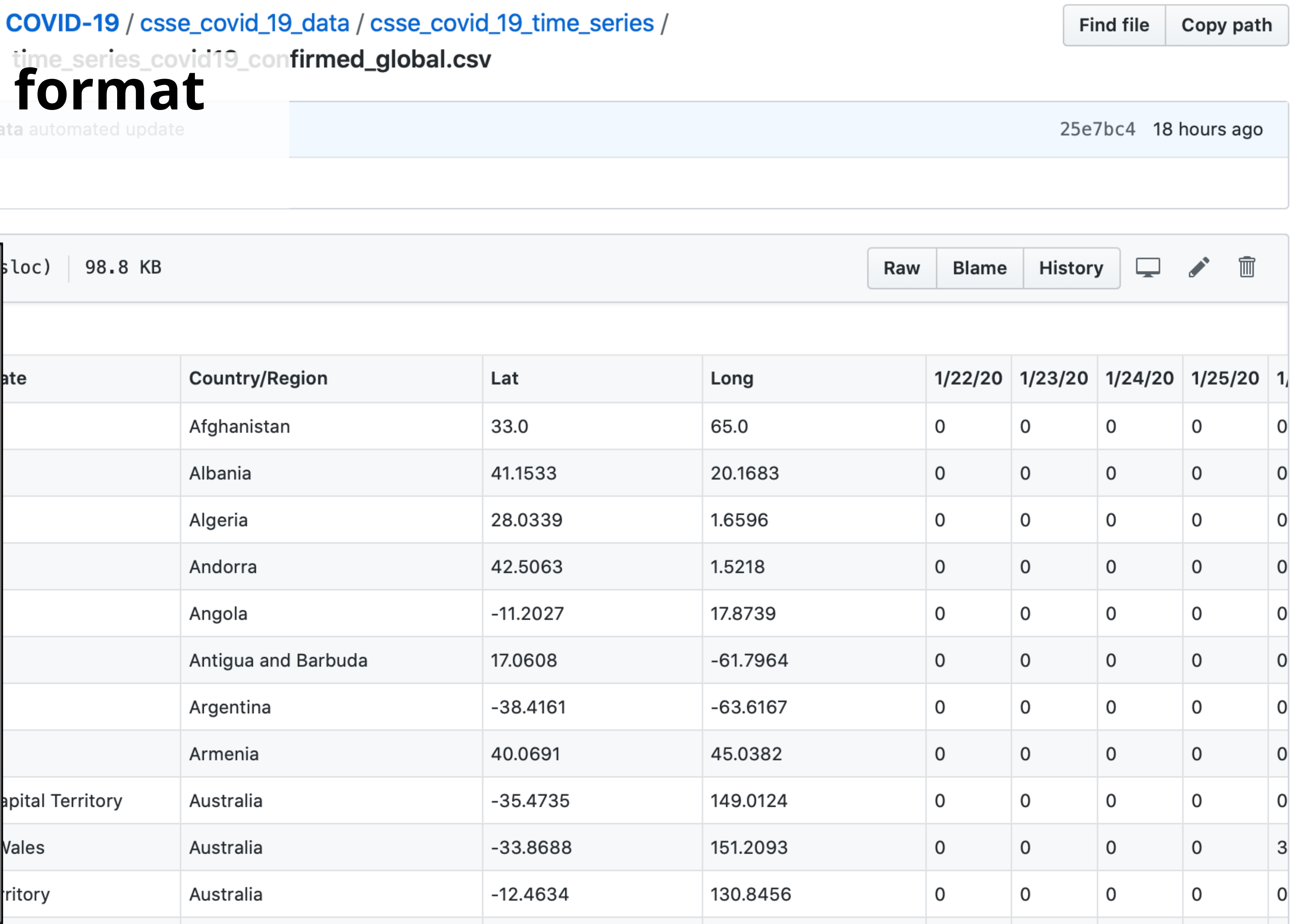
Example of Bearable Practices

2. Non-standard date format

Use ISO 8601



PUBLIC SERVICE ANNOUNCEMENT:
OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.
THIS IS THE CORRECT WAY TO WRITE NUMERIC DATES:
2013-02-27
THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:
02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013. II. 27. 27/2-13 2013.158904109
MMXIII-II-XXVII MMXIII LVII/CCCLXV 1330300800
((3+3)×(111+1)-1)×3/3-1/3³ 2013
HISSSSS
2-27-13
10/11011/1101 02/27/20/13 01237



COVID-19 / csse_covid_19_data / csse_covid_19_time_series / time_series_covid19_confirmed_global.csv

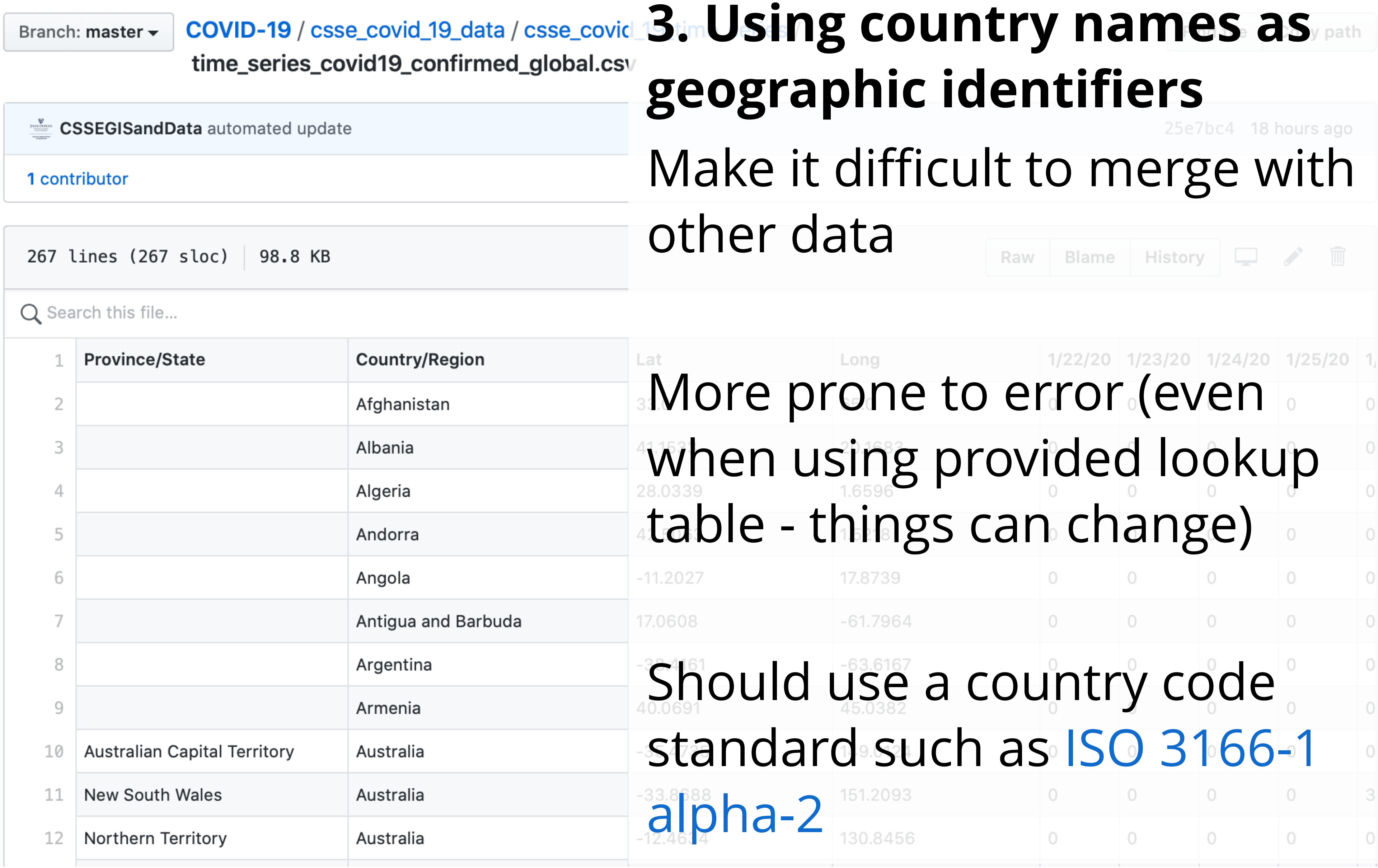
CSSEGISandData automated update 25e7bc4 18 hours ago

1 contributor

sloc) | 98.8 KB

ate	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/
	Afghanistan	33.0	65.0	0	0	0	0	0
	Albania	41.1533	20.1683	0	0	0	0	0
	Algeria	28.0339	1.6596	0	0	0	0	0
	Andorra	42.5063	1.5218	0	0	0	0	0
	Angola	-11.2027	17.8739	0	0	0	0	0
	Antigua and Barbuda	17.0608	-61.7964	0	0	0	0	0
	Argentina	-38.4161	-63.6167	0	0	0	0	0
	Armenia	40.0691	45.0382	0	0	0	0	0
apital Territory	Australia	-35.4735	149.0124	0	0	0	0	0
Wales	Australia	-33.8688	151.2093	0	0	0	0	3
ritory	Australia	-12.4634	130.8456	0	0	0	0	0

Example of Bearable Practices



COVID-19 / csse_covid_19_data / csse_covid_19_time_series / time_series_covid19_confirmed_global.csv

CSSEGISandData automated update   25e7bc4 18 hours ago

1 contributor

267 lines (267 sloc) | 98.8 KB

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/
1	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/
2		Afghanistan	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0
3		Albania	[illegible]	20.1683	[illegible]	[illegible]	[illegible]	[illegible]	0
4		Algeria	28.0339	1.6596	0	0	0	0	0
5		Andorra	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0
6		Angola	-11.2027	17.8739	0	0	0	0	0
7		Antigua and Barbuda	17.0608	-61.7964	0	0	0	0	0
8		Argentina	[illegible]	-63.6167	0	0	0	0	0
9		Armenia	40.0691	45.0382	0	[illegible]	0	0	0
10	Australian Capital Territory	Australia	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0
11	New South Wales	Australia	-33.8688	151.2093	0	0	0	0	3
12	Northern Territory	Australia	[illegible]	130.8456	0	0	0	0	0

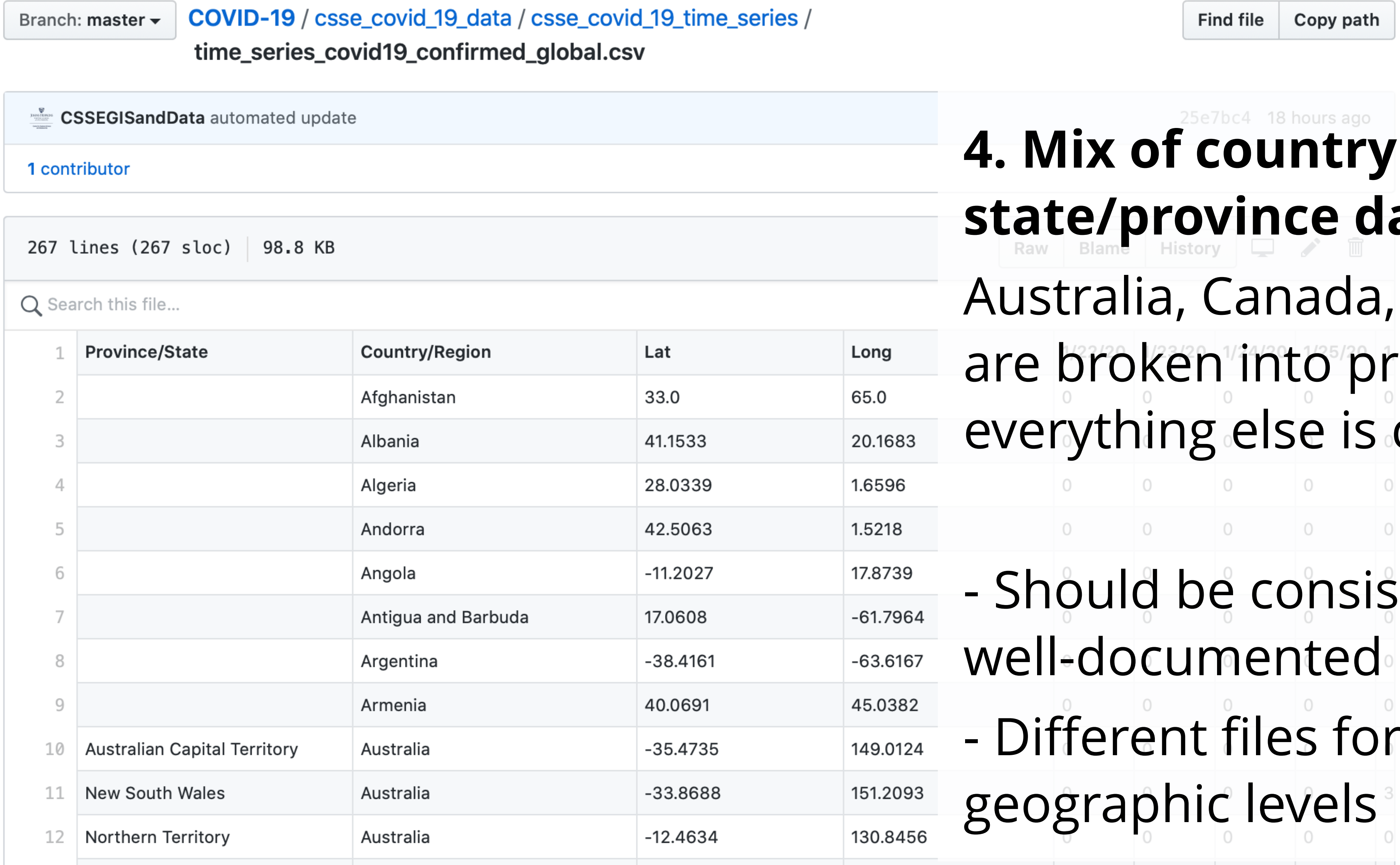
3. Using country names as geographic identifiers

Make it difficult to merge with other data

More prone to error (even when using provided lookup table - things can change)

Should use a country code standard such as ISO 3166-1 alpha-2

Example of Bearable Practices



COVID-19 / csse_covid_19_data / csse_covid_19_time_series / time_series_covid19_confirmed_global.csv

CSSEGISandData automated update 25e7bc4 18 hours ago

1 contributor

267 lines (267 sloc) | 98.8 KB

	Province/State	Country/Region	Lat	Long
1	Province/State	Country/Region	Lat	Long
2		Afghanistan	33.0	65.0
3		Albania	41.1533	20.1683
4		Algeria	28.0339	1.6596
5		Andorra	42.5063	1.5218
6		Angola	-11.2027	17.8739
7		Antigua and Barbuda	17.0608	-61.7964
8		Argentina	-38.4161	-63.6167
9		Armenia	40.0691	45.0382
10	Australian Capital Territory	Australia	-35.4735	149.0124
11	New South Wales	Australia	-33.8688	151.2093
12	Northern Territory	Australia	-12.4634	130.8456

4. Mix of country and state/province data

Australia, Canada, and China are broken into provinces while everything else is country-level

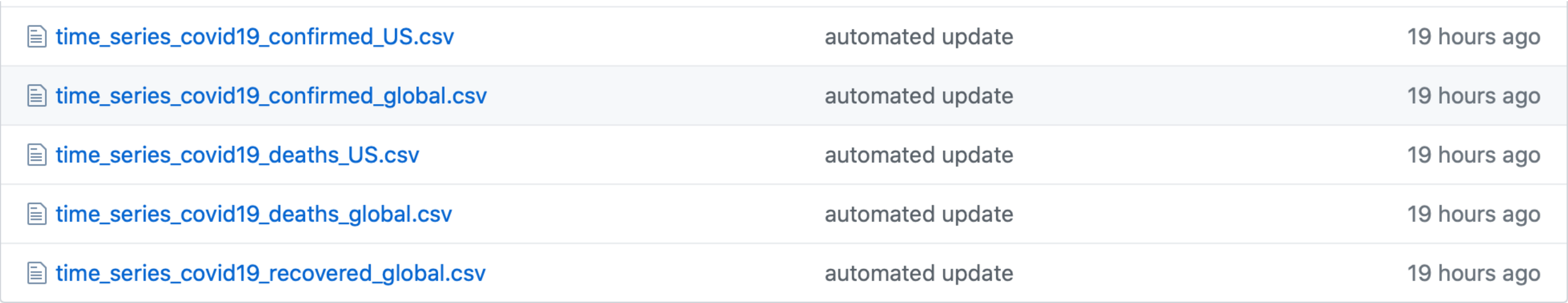
- Should be consistent and well-documented
- Different files for different geographic levels

Example of Bearable Practices

5. Three files for three variables (cases, deaths, recovered)

These need to be joined to get an analysis dataset

All variables would ideally be in one file, one column per variable - back to tidy data principles

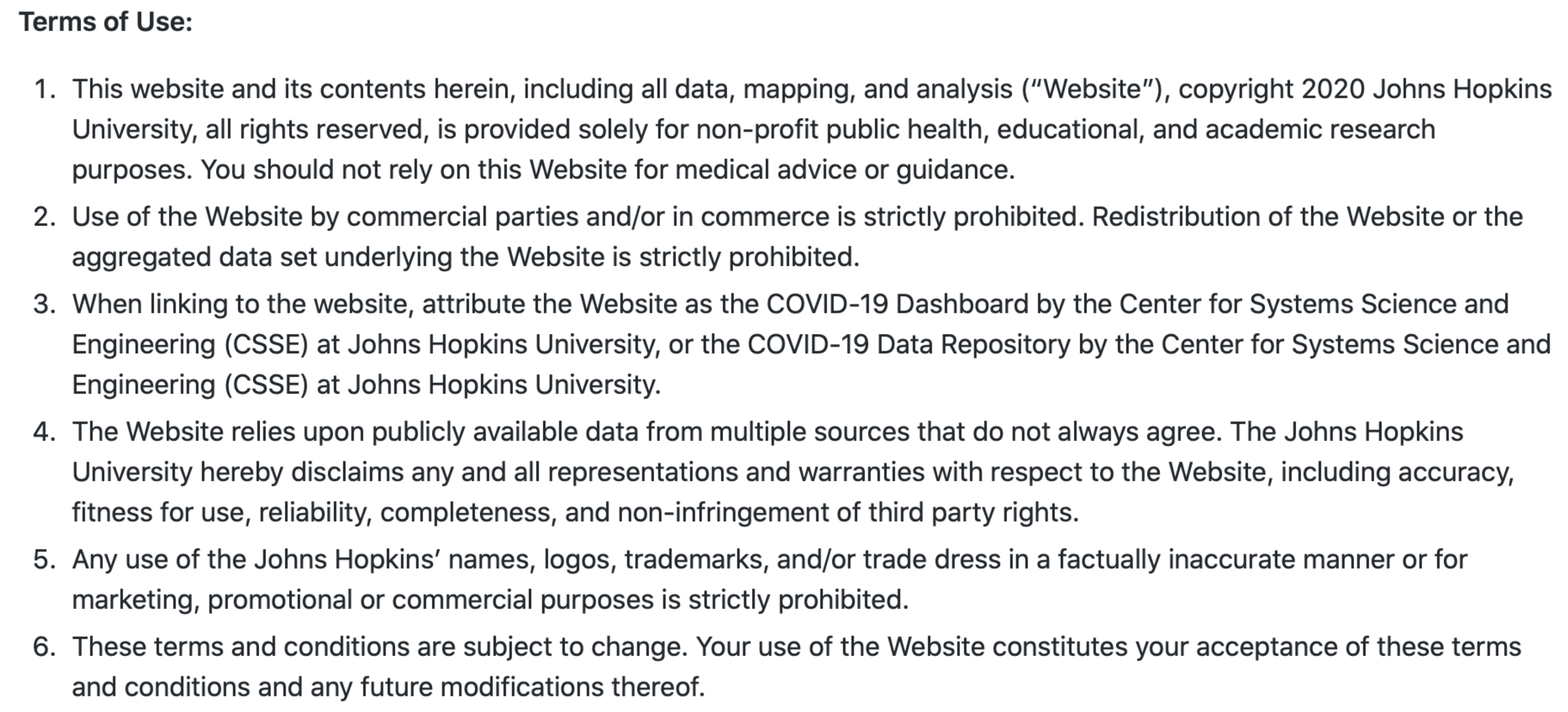


time_series_covid19_confirmed_US.csv	automated update	19 hours ago
time_series_covid19_confirmed_global.csv	automated update	19 hours ago
time_series_covid19_deaths_US.csv	automated update	19 hours ago
time_series_covid19_deaths_global.csv	automated update	19 hours ago
time_series_covid19_recovered_global.csv	automated update	19 hours ago

Example of Bearable Practices

6. Ambiguous terms of use and no standard open license

- Non-standard and too-restrictive terms can impede the progress of science
- Ideally for open data should use a standard license such as Creative Commo

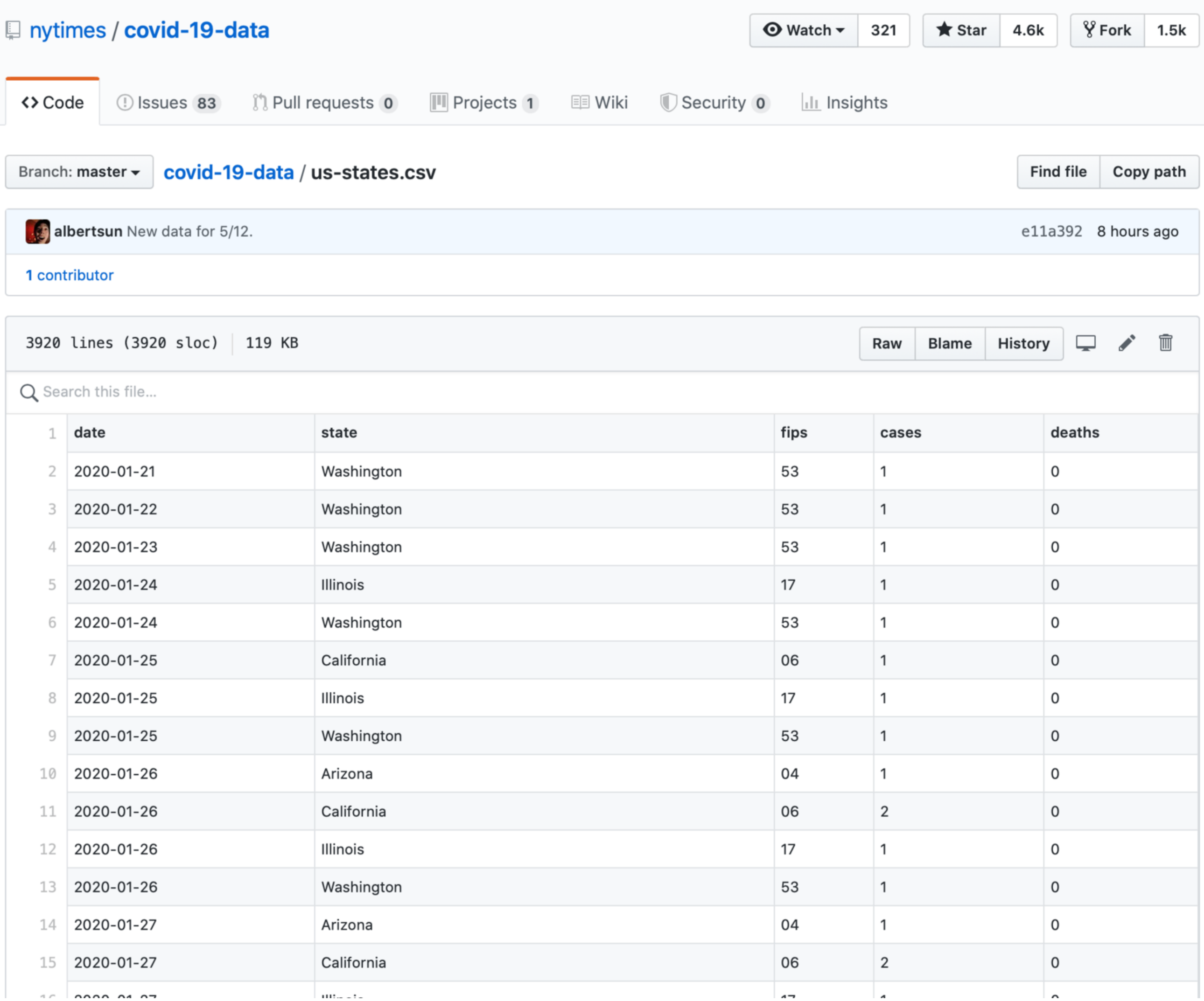


Terms of Use:

1. This website and its contents herein, including all data, mapping, and analysis (“Website”), copyright 2020 Johns Hopkins University, all rights reserved, is provided solely for non-profit public health, educational, and academic research purposes. You should not rely on this Website for medical advice or guidance.
2. Use of the Website by commercial parties and/or in commerce is strictly prohibited. Redistribution of the Website or the aggregated data set underlying the Website is strictly prohibited.
3. When linking to the website, attribute the Website as the COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, or the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.
4. The Website relies upon publicly available data from multiple sources that do not always agree. The Johns Hopkins University hereby disclaims any and all representations and warranties with respect to the Website, including accuracy, fitness for use, reliability, completeness, and non-infringement of third party rights.
5. Any use of the Johns Hopkins’ names, logos, trademarks, and/or trade dress in a factually inaccurate manner or for marketing, promotional or commercial purposes is strictly prohibited.
6. These terms and conditions are subject to change. Your use of the Website constitutes your acceptance of these terms and conditions and any future modifications thereof.

Example of Best Practices - New York Times

- Tidy format
- ISO 8601 date format
- Standard geocodes for admin 1 and 2 data (FIPS)
- State and county-level data are in separate files
- License is co-extensive with the Creative Commons Attribution-NonCommercial 4.0 International license



covid-19-data / us-states.csv

albertsun New data for 5/12. e11a392 8 hours ago

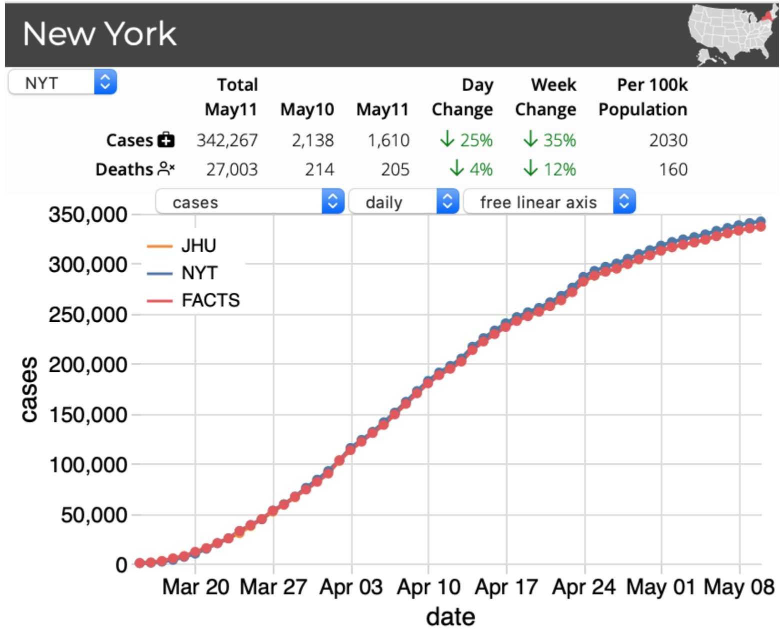
3920 lines (3920 sloc) 119 KB

date	state	fips	cases	deaths
2020-01-21	Washington	53	1	0
2020-01-22	Washington	53	1	0
2020-01-23	Washington	53	1	0
2020-01-24	Illinois	17	1	0
2020-01-24	Washington	53	1	0
2020-01-25	California	06	1	0
2020-01-25	Illinois	17	1	0
2020-01-25	Washington	53	1	0
2020-01-26	Arizona	04	1	0
2020-01-26	California	06	2	0
2020-01-26	Illinois	17	1	0
2020-01-26	Washington	53	1	0
2020-01-27	Arizona	04	1	0
2020-01-27	California	06	2	0

Building a Tool for Case Count Data

- Pull country-level case counts every 5 minutes from the following sources
  - WHO
  - JHU
  - ECDC
  - Worldometer
- Roll up counts to WHO Region, continent, and global levels
- Compute statistics of interest for each geographic entity
  - Day-to-day and week-to-week change in new cases / deaths
  - Case fatality rate (# deaths / # of cases)*
  - Attack rate (# cases / population)
  - Etc.

*Does not take time to onset of death into account

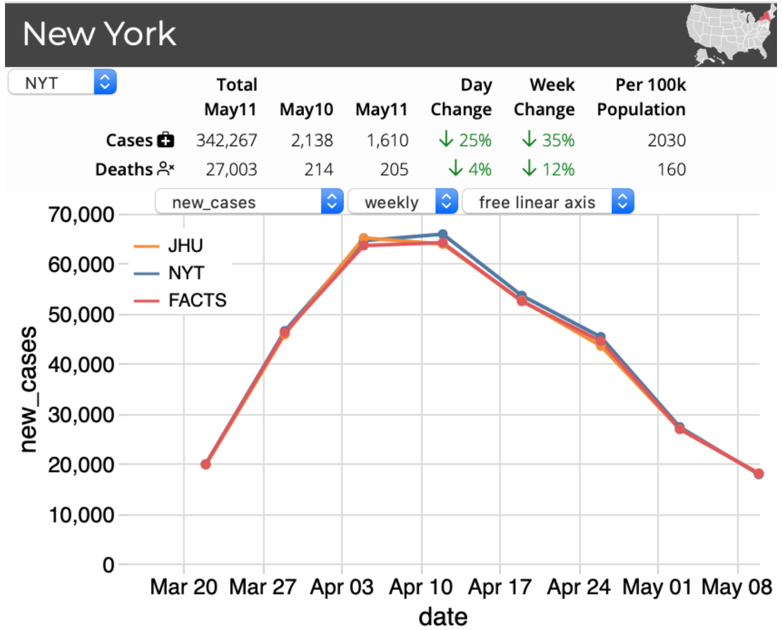


New York

NYT

	Total May11	May10	May11	Day Change	Week Change	Per 100k Population
Cases	342,267	2,138	1,610	↓ 25%	↓ 35%	2030
Deaths	27,003	214	205	↓ 4%	↓ 12%	160

cases | daily | free linear axis

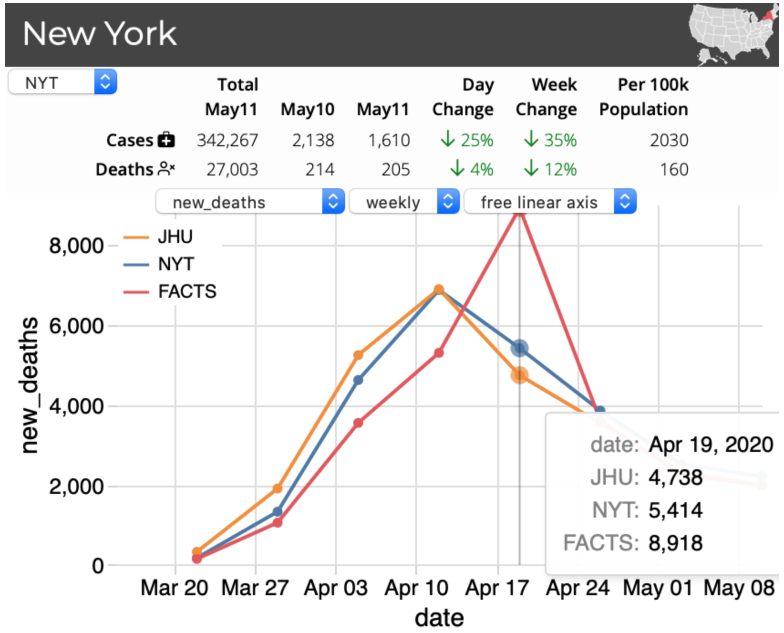


New York

NYT

	Total May11	May10	May11	Day Change	Week Change	Per 100k Population
Cases	342,267	2,138	1,610	↓ 25%	↓ 35%	2030
Deaths	27,003	214	205	↓ 4%	↓ 12%	160

new_cases | weekly | free linear axis



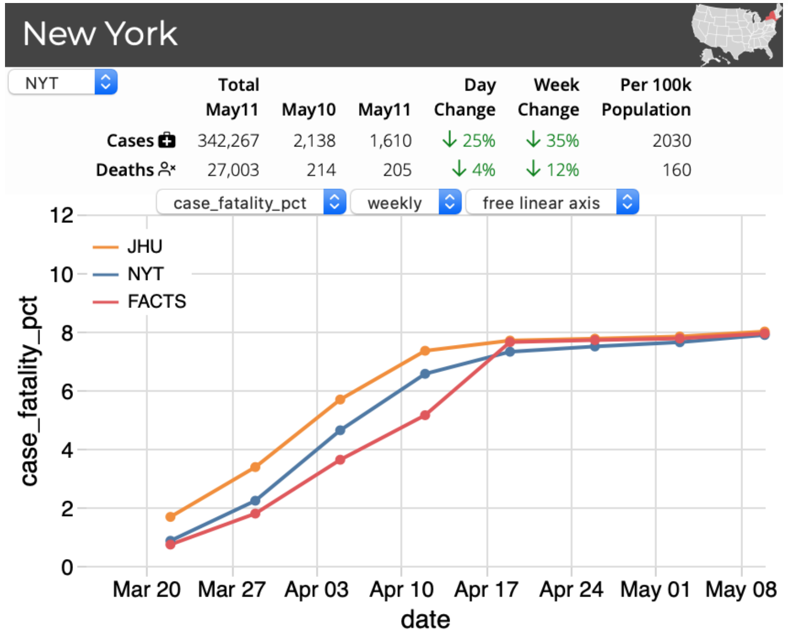
New York

NYT

	Total May11	May10	May11	Day Change	Week Change	Per 100k Population
Cases	342,267	2,138	1,610	↓ 25%	↓ 35%	2030
Deaths	27,003	214	205	↓ 4%	↓ 12%	160

new_deaths | weekly | free linear axis

date: Apr 19, 2020
JHU: 4,738
NYT: 5,414
FACTS: 8,918



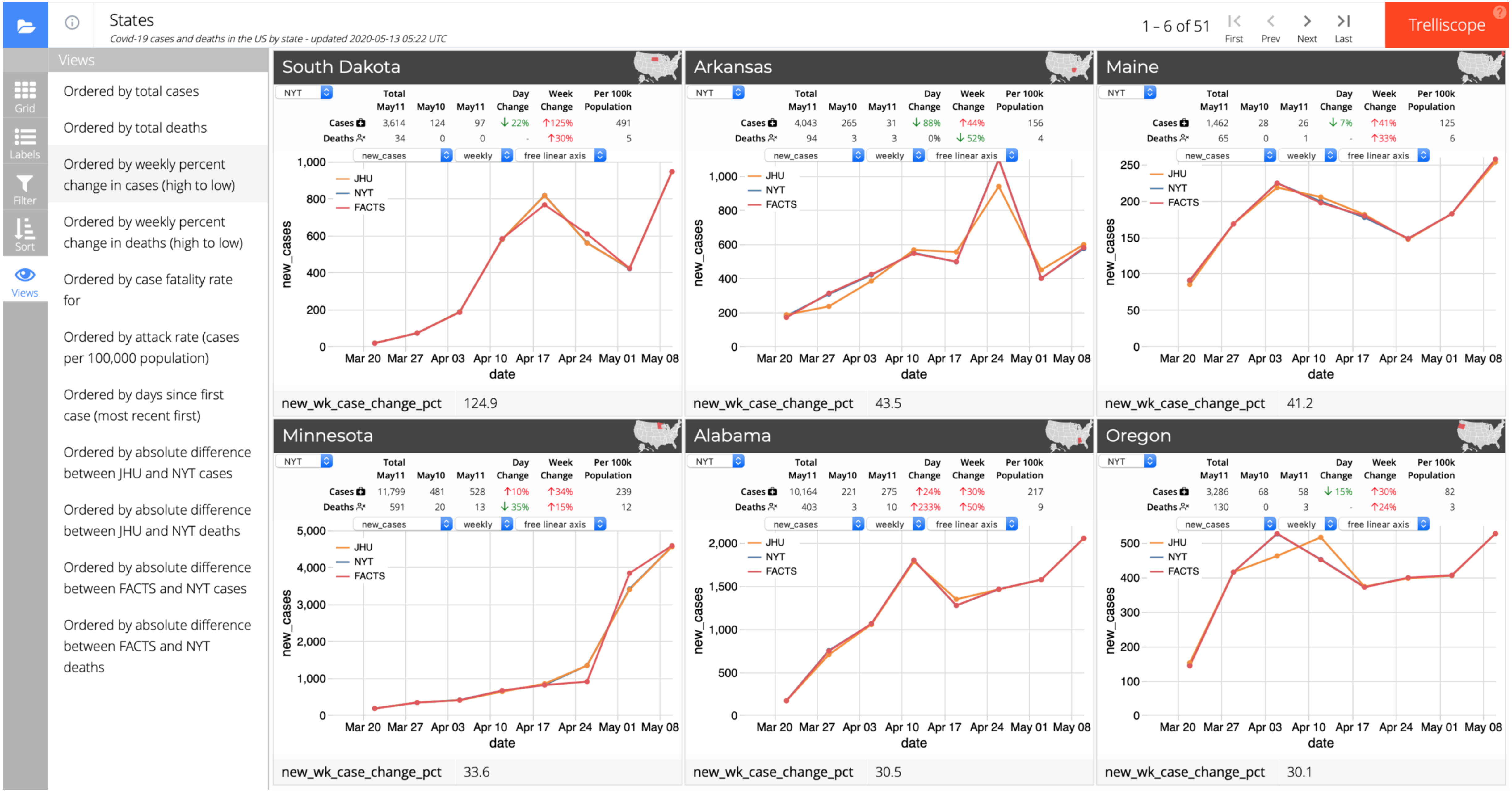
New York

NYT

	Total May11	May10	May11	Day Change	Week Change	Per 100k Population
Cases	342,267	2,138	1,610	↓ 25%	↓ 35%	2030
Deaths	27,003	214	205	↓ 4%	↓ 12%	160

case_fatality_pct | weekly | free linear axis

Provide a set of visualizations for each geographic entity for the user to interact with



States

Covid-19 cases and deaths in the US by state - updated 2020-05-13 05:22 UTC

Views

- Ordered by total cases
- Ordered by total deaths
- Ordered by weekly percent change in cases (high to low)
- Ordered by weekly percent change in deaths (high to low)
- Ordered by case fatality rate for
- Ordered by attack rate (cases per 100,000 population)
- Ordered by days since first case (most recent first)
- Ordered by absolute difference between JHU and NYT cases
- Ordered by absolute difference between JHU and NYT deaths
- Ordered by absolute difference between FACTS and NYT cases
- Ordered by absolute difference between FACTS and NYT deaths

South Dakota

	Total			Day	Week	Per 100k
	May11	May10	May11	Change	Change	Population
Cases	3,614	124	97	↓22%	↑125%	491
Deaths	34	0	0	-	↑30%	5

new_wk_case_change_pct 124.9

Arkansas

	Total			Day	Week	Per 100k
	May11	May10	May11	Change	Change	Population
Cases	4,043	265	31	↓88%	↑44%	156
Deaths	94	3	3	0%	↓52%	4

new_wk_case_change_pct 43.5

Maine

	Total			Day	Week	Per 100k
	May11	May10	May11	Change	Change	Population
Cases	1,462	28	26	↓7%	↑41%	125
Deaths	65	0	1	-	↑33%	6

new_wk_case_change_pct 41.2

Minnesota

	Total			Day	Week	Per 100k
	May11	May10	May11	Change	Change	Population
Cases	11,799	481	528	↑10%	↑34%	239
Deaths	591	20	13	↓35%	↑15%	12

new_wk_case_change_pct 33.6

Alabama

	Total			Day	Week	Per 100k
	May11	May10	May11	Change	Change	Population
Cases	10,164	221	275	↑24%	↑30%	217
Deaths	403	3	10	↑233%	↑50%	9

new_wk_case_change_pct 30.5

Oregon

	Total			Day	Week	Per 100k
	May11	May10	May11	Change	Change	Population
Cases	3,286	68	58	↓15%	↑30%	82
Deaths	130	0	3	-	↑24%	3

new_wk_case_change_pct 30.1

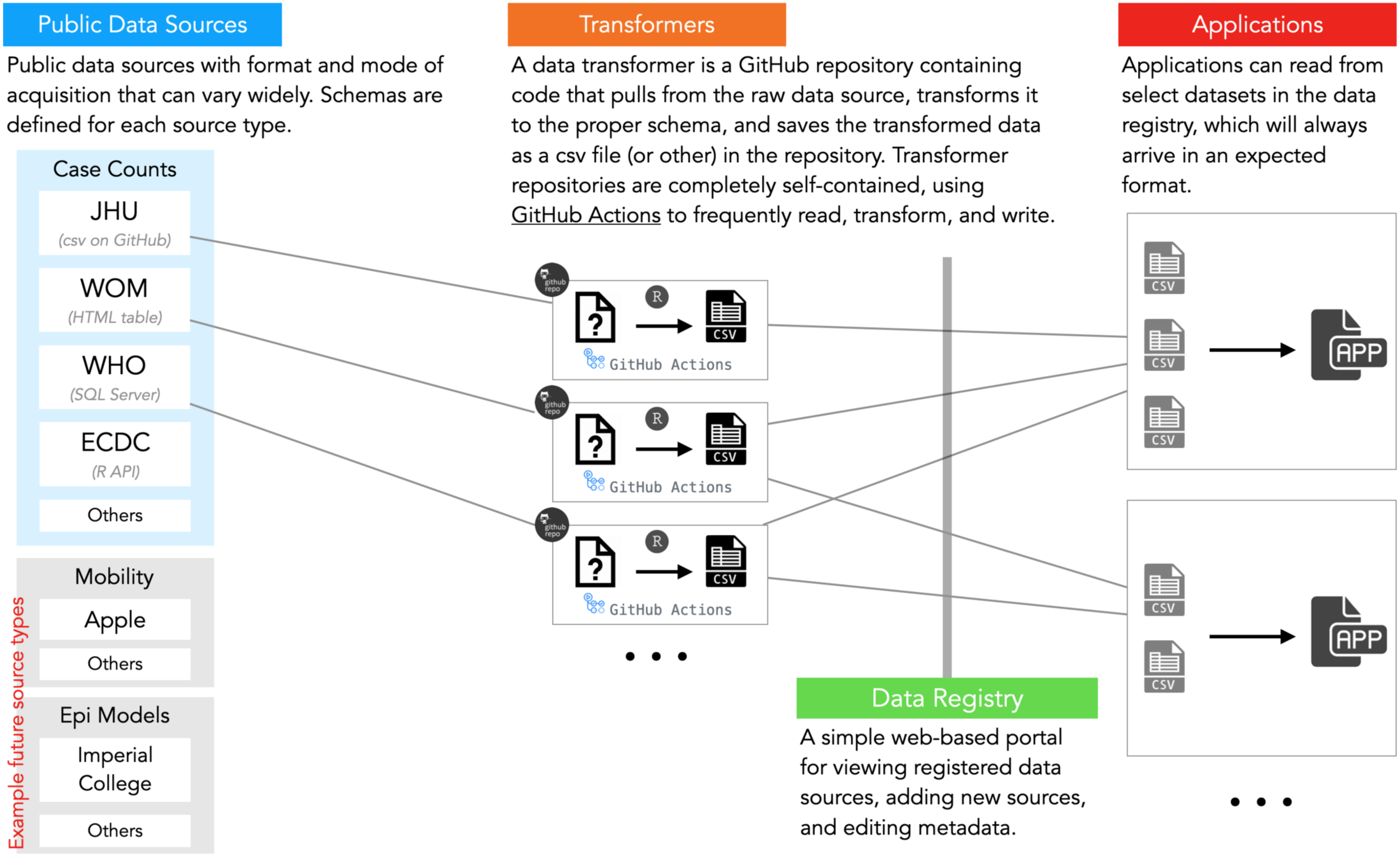
With Trelliscope these can be navigated interactively https://covid19-us-casecounts.netlify.app

COVID-19 Data Registry

Efforts exist for pulling multiple sources of COVID-19 data together, e.g.

- coronadatascraper
- #data4covid19

We are working toward a set of data registry tools that enable the open data community to register datasets in a way that conforms to standards but doesn't require the original data provider to change the way they are publishing their data



Public Data Sources

Public data sources with format and mode of acquisition that can vary widely. Schemas are defined for each source type.

Transformers

A data transformer is a GitHub repository containing code that pulls from the raw data source, transforms it to the proper schema, and saves the transformed data as a csv file (or other) in the repository. Transformer repositories are completely self-contained, using GitHub Actions to frequently read, transform, and write.

Applications

Applications can read from select datasets in the data registry, which will always arrive in an expected format.

Data Registry

A simple web-based portal for viewing registered data sources, adding new sources, and editing metadata.

Potential Future Work

- Standard schemas and transformers for new data types
  - Mobility data
  - Administrative statistics (capacity, vulnerability, demographics, etc.)
  - Models (IHME, Imperial College, Amherst, etc.)
- Augmenting interfaces to incorporate this information in insightful ways

Thank You

rhafen@gmail.com

@hafenstats

https://slides.com/hafen/covid19-casecounts