China Data Institute

Center for Geographic Analysis
Harvard University

社会地理计算联合研究中心
Geocomputation Center for Social Science

“Resources for COVID-19”

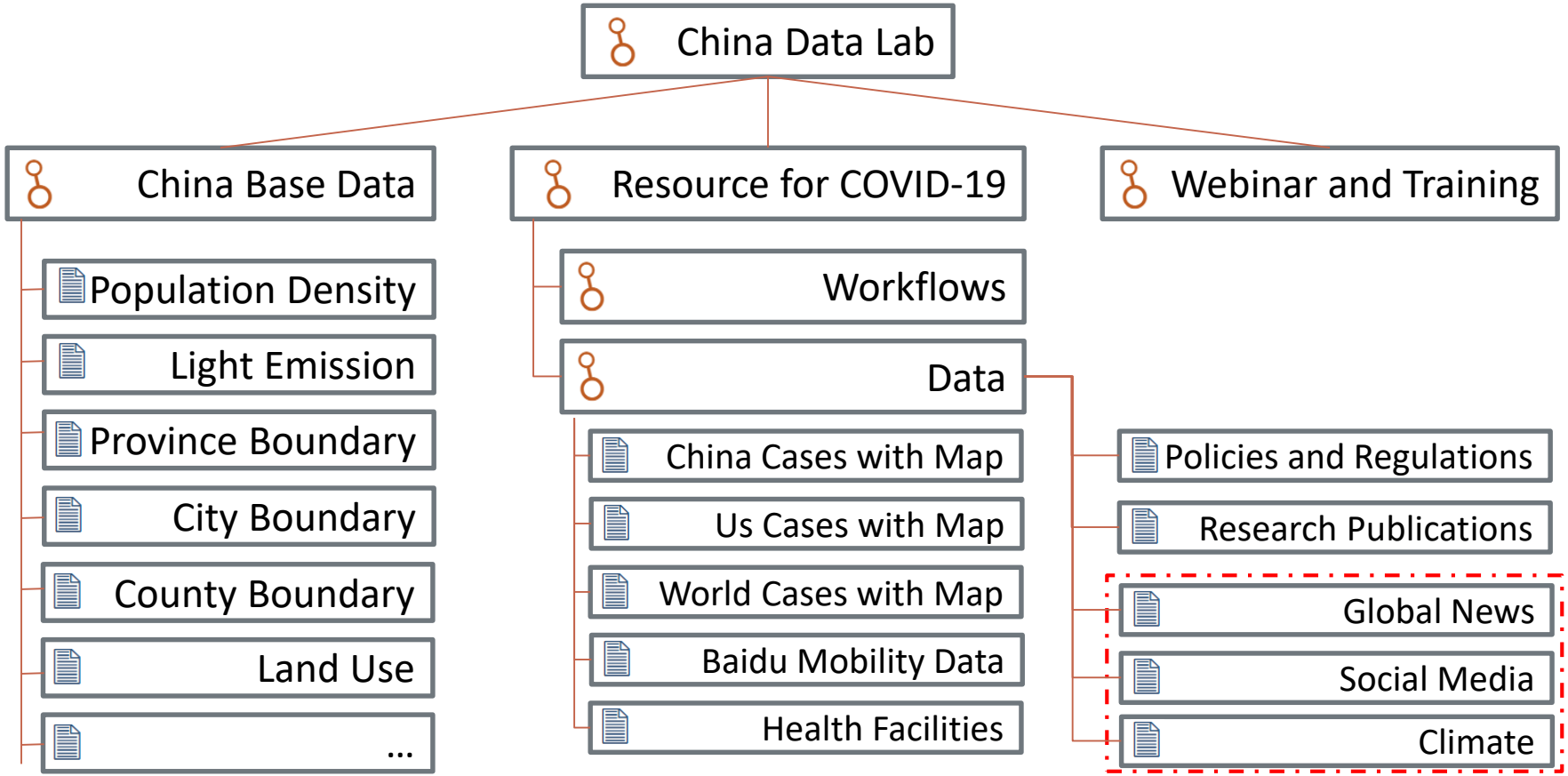
Introduction to the Virus Data

Tao Hu, Harvard University
Lingbo Liu, Hongqiang Liu and Jing Du, Wuhan University

Outline

- ❑ Data Sources
- ❑ Data Processing
- ❑ Data Integration
- ❑ Data Depository
- ❑ Data Cloud
- ❑ Data Workflows
- ❑ Sample Works for Data Analysis
- ❑ Discussion

“Resources for COVID-19” at Harvard Dataverse



Data Sources

- Base data:
  - China Data Institute, China Province and City Census Data with Comparable GIS Maps for Census 2000 and 2010.
  - China Data Institute, US State and County Census Data with Comparable GIS Maps for Census 1970, 1980, 1990, 2000 and 2010.
- Virus data:

Region	Source	Link	Start Date	Update
China	DXY	https://github.com/BlankerL/DXY-COVID-19-Data	1/14/2020	3/18/2020
US	USA FACTS The New York Times	https://github.com/nytimes/covid-19-data; https://usafacts.org/	1/22/2020	3/29/2020
Italy	Department of Civil Protection	https://github.com/pcm-dpc/COVID-19	1/22/2020	3/28/2020
Global	John Hopkins University	https://github.com/CSSEGISandData/COVID-19	1/22/2020	3/28/2020

Data Classification

China

- China virus data by administrative units:
  - Province
  - City
- China virus data by categories:
  - Confirmed cases
  - Death cases
  - Recovered cases

USA

- US virus data by administrative units:
  - State
  - County
- US virus data by categories:
  - Confirmed cases
  - Death cases
  - Recovered cases

Global

Country virus data by categories:

- Confirmed cases
- Death cases
- Recovered cases

Data Processing and Integration: China

Base maps:

1. Merge the urban districts of Shanghai, Beijing, Chongqing and Tianjin of the prefecture city map
2. Recode those merged districts of Shanghai, Beijing, Chongqing and Tianjin of the city map

OBJECTI	GbCity	City_CH	City_EN
1	1101	北京市辖区	Beijing (Districts)
2	1102	北京市辖县	Beijing (Counties)
3	1201	天津市辖区	Tianjin (Districts)
4	1202	天津市辖县	Tianjin (Counties)
5	1301	石家庄市	Shijiazhuang
6	1302	唐山市	Tangshan
7	1303	秦皇岛市	Qinhuangdao
8	1304	邯郸市	Handan
9	1305	邢台市	Xingtai
10	1306	保定市	Baoding
11	1307	张家口市	Zhangjiakou
12	1308	承德市	Chengde
13	1309	沧州市	Cangzhou
14	1310	廊坊市	Langfang
15	1311	衡水市	Hengshui
16	1401	太原市	Taiyuan
17	1402	大同市	Datong
18	1403	阳泉市	Yangquan
19	1404	长治市	Changzhi

Virus data tables:

1. Assign GB code as in the base map for each city in the virus data table
2. Match those cities in the virus data table to those cities on the base map
3. Merge those county level cities or special regions in the virus table to match the city base map

Map merging:

1. Merge the virus data table with the province map by GB code
2. Merge the virus data table with the prefecture city by GB code

河南省：济源市

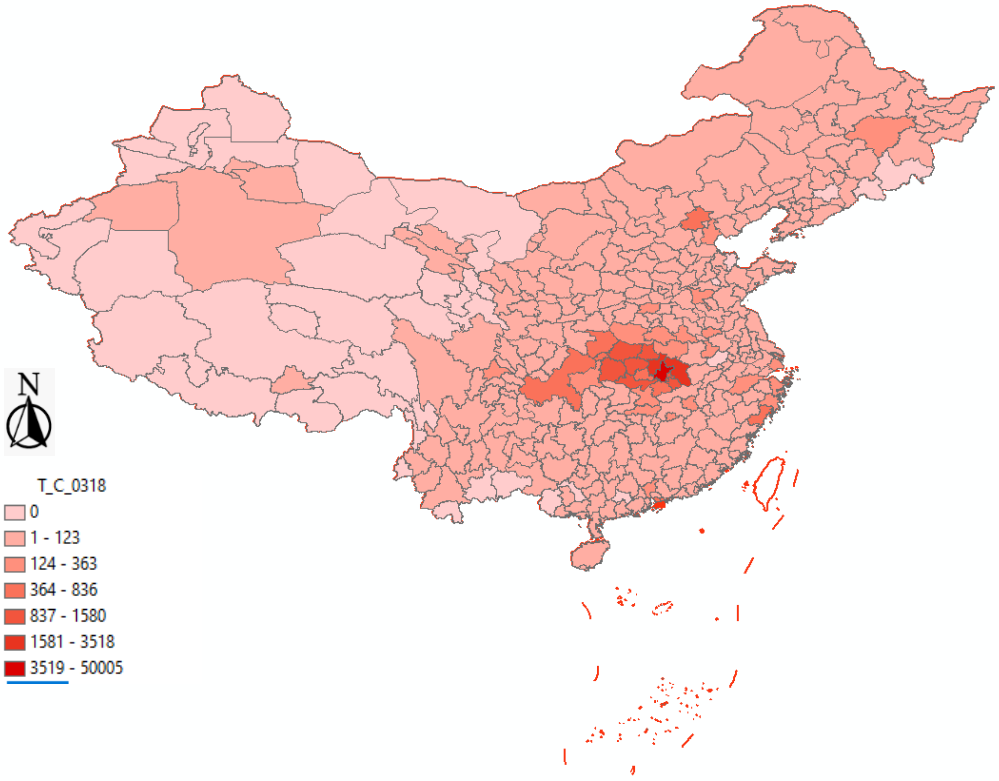
湖北省：仙桃市、潜江市、天门市、神农架林区

海南省：五指山市、琼海市、文昌市、万宁市、东方市、屯昌县、定安县、澄迈县、临高县、保亭黎族苗族自治县、琼中黎族苗族自治县、白沙黎族自治县、陵水黎族自治县、昌江黎族自治县、乐东黎族自治县

新疆维吾尔自治区：石河子市、阿拉尔市、图木舒克市、五家渠市、北屯市、铁门关市、双河市、可克达拉市、昆玉市、胡杨河市

Sample Map for the Virus Data of China

The Choropleth Map of Confirmed Cases as of March 27, 2020



Data Processing and Integration: US

- Retrieve the virus data by states from https://usafacts.org/

```
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
```

→ Data Transformation →

A	B	C	D	E	F
FIPS	state	1/21/2020	1/22/2020	1/23/2020	1/24/2020
53	Washington	1	1	1	1
17	Illinois	0	0	0	1
6	California	0	0	0	0
4	Arizona	0	0	0	0
25	Massachusetts	0	0	0	0
55	Wisconsin	0	0	0	0
48	Texas	0	0	0	0
31	Nebraska	0	0	0	0

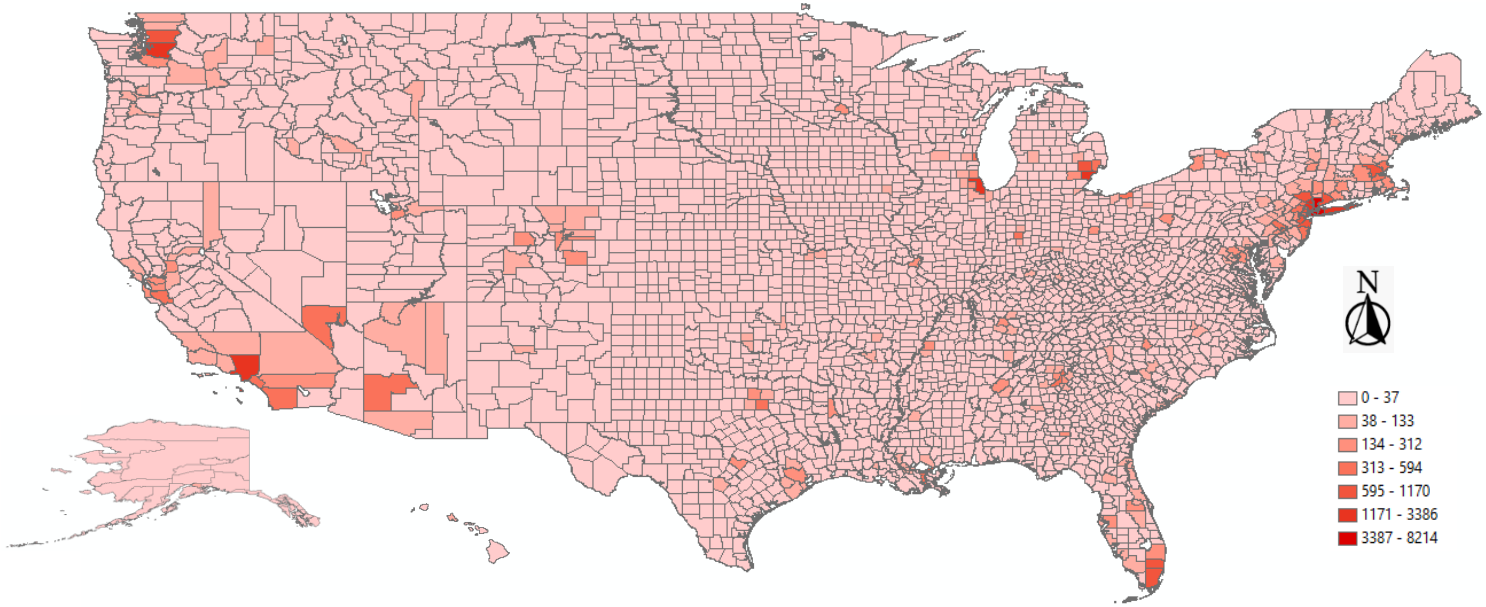
- Retrieve the virus data by county

Website → Monitor Request → Parse JSON → Write CSV

- Merge the virus data with the US Base map by state and county based on the FIPS code

Sample Map for the Virus Data of US

The Choropleth Map of Confirmed Cases as of March 27, 2020



Data Processing: Italy

Italy Virus Data

- Region
- Province

A	B	C	D	E	F	G
code_region	name_reg	code_province	name_province	province_abbr	2/24/2020	2/25/2020
13	Abruzzo	69	Chieti	CH	0	0
13	Abruzzo	66	L'Aquila	AQ	0	0
13	Abruzzo	68	Pescara	PE	0	0
13	Abruzzo	67	Teramo	TE	0	0
13	Abruzzo	979	In fase di definizione/aggiornamen		0	0
17	Basilicata	77	Matera	MT	0	0
17	Basilicata	76	Potenza	PZ	0	0
17	Basilicata	980	In fase di definizione/aggiornamen		0	0
4	P.A. Bolza	21	Bolzano	BZ	0	1
4	P.A. Bolza	981	In fase di definizione/ag			
18	Calabria	79	Catanzaro	CZ		
18	Calabria	78	Cosenza	CS		

Cases by Province

Florence	Firenze
Forli' - Cesena	ForlÃ¬-Cesena
Mantua	Mantova
Monza and Brianza	Monza e della Brianza
Padua	Padova
Pesaro E Urbino	Pesaro e Urbino
Reggio Di Calabria	Reggio di Calabria
Reggio Nell'Emilia	Reggio nell'Emilia
Syracuse	Siracusa

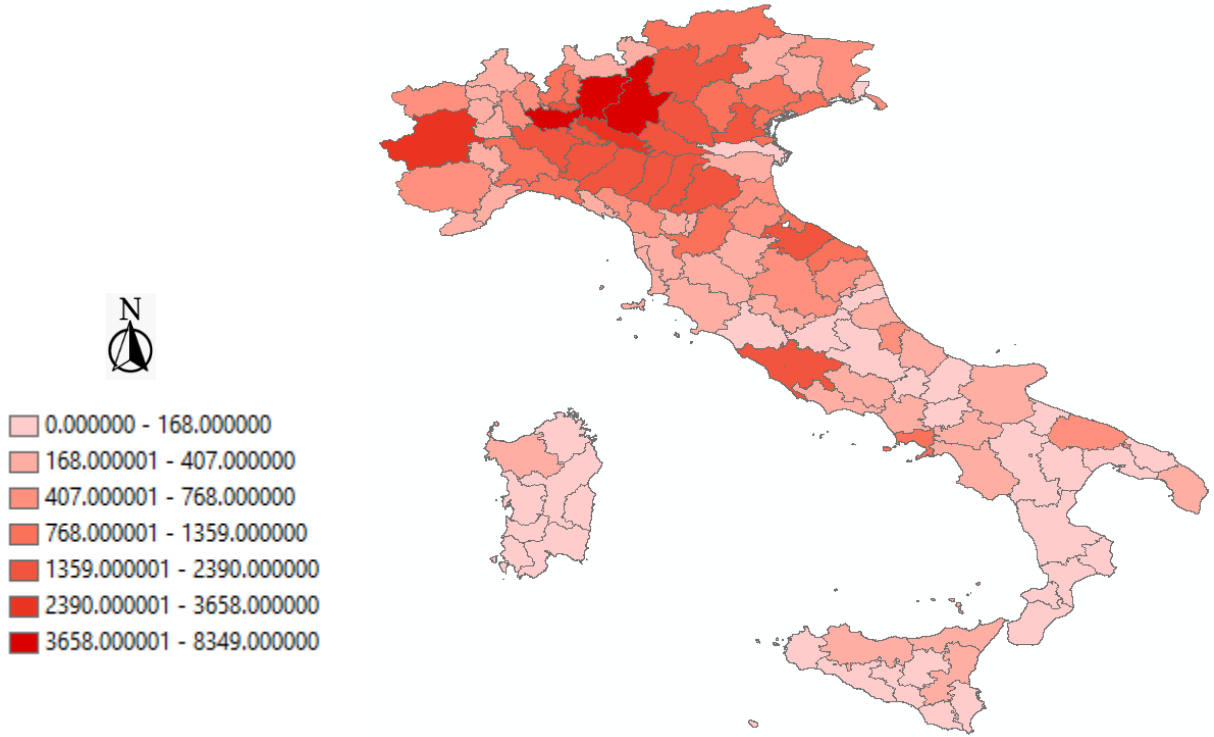
Province Mapping List

ID_1	NAME_1	ID_2	NAME_2
16	Toscana	92	Livorno
16	Toscana	93	Lucca
16	Toscana	94	Massa Carrara
16	Toscana	95	Pisa
16	Toscana	96	Pistoia
16	Toscana	97	Prato
16	Toscana	98	Siena
17	Trentino-Alto A	99	Bolzano
17	Trentino-Alto A	100	Trento
18	Umbria	101	Perugia
18	Umbria	102	Terni
19	Valle d'Aosta	103	Aosta
20	Veneto	104	Belluno

Province Basemap

Sample Map for the Virus Data of Italy

The Choropleth Map of Confirmed Cases as of March 28, 2020



Data Processing and Integration: World

Global Virus Data

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

Global Cases

Shape *	OBJECTID	CNTRY_NAME
Polygon	1	Aruba
Polygon	2	Antigua and Barbuda
Polygon	3	Afghanistan
Polygon	4	Algeria
Polygon	5	Azerbaijan
Polygon	6	Albania
Polygon		Armenia
Polygon	8	Andorra
Polygon	9	Angola
Polygon	10	American Samoa
Polygon	11	Argentina
ygon	12	Australia
ygon	13	Austria
ygon	14	Anguilla

Map provided by ArcGIS

case_country	country_arcgis
Bahamas	Bahamas, The
Belarus	Byelarus
Democratic Republic of Congo	Congo
Faeroe Islands	Faroe Islands
Gambia	Gambia, The
Cote d'Ivoire	Ivory Coast
Isle of Man	Man, Isle of
Myanmar	Myanmar (Burma)
Saint Kitts and Nevis	St. Kitts and Nevis
Saint Lucia	St. Lucia
Saint Vincent and the Grenadines	St. Vincent and the G
Tanzania	Tanzania, United Rep
United States Virgin Islands	Virgin Islands
Palestine	West Bank

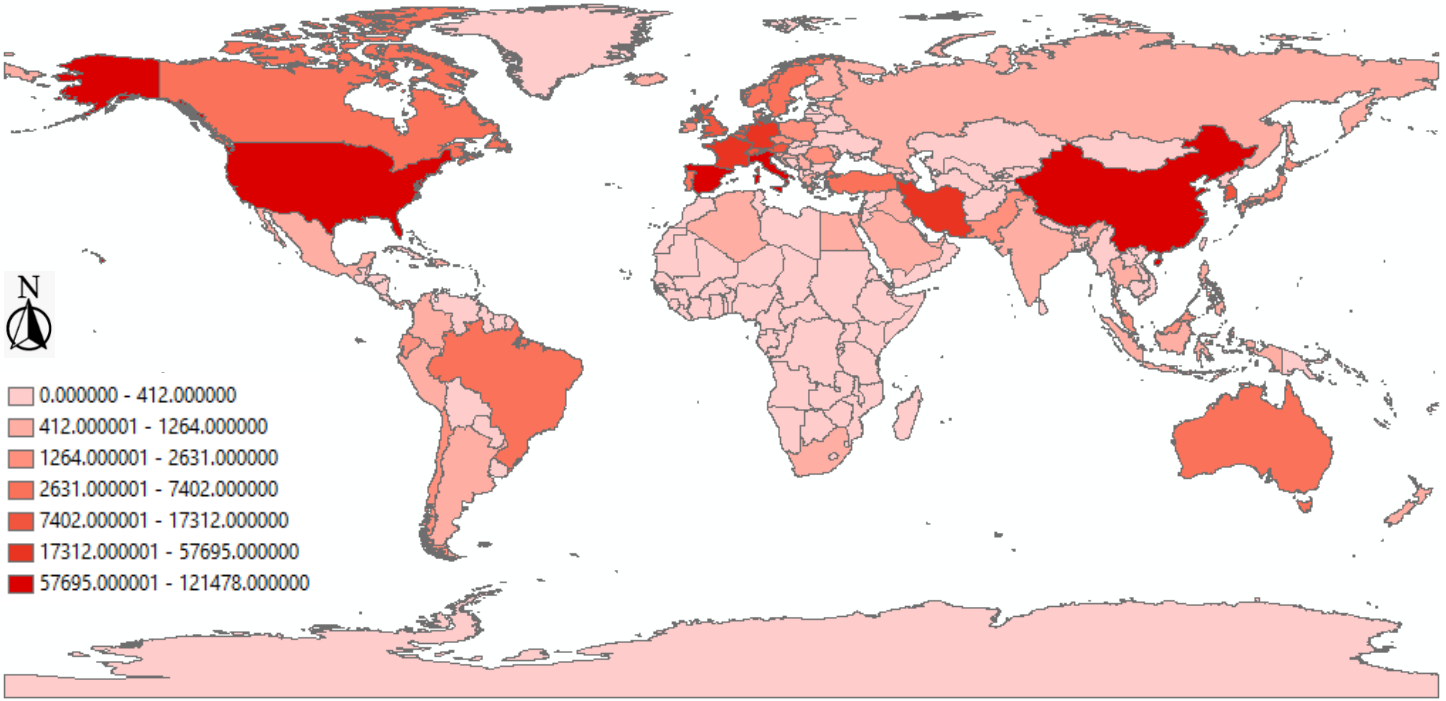
Country Mapping List

Kosovo, Vatican, Timor not found

https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv

Sample Map of Virus Data by Country

The Choropleth Map of Confirmed Cases as of March 28, 2020



The List of Base Data

Data Title	File Name	File Type	File Path
Prefecture County Census Data with 2000-2010 GIS Maps	2000-2010CountyCensus.shp	Shape File	public\Covid\Data\BaseData\China
Prefecture City Census Data with 2000-2010 GIS Maps	2000-2010PrefectureCensus.shp	Shape File	public\Covid\Data\BaseData\China
Prefecture Province Census Data with 2000-2010 GIS Maps	2000-2010ProvCensus.shp	Shape File	public\Covid\Data\BaseData\China
Base map, County	County0010.shp	Shape File	public\Covid\Data\BaseData\China
Base map, Prefecture City	City0010.shp	Shape File	public\Covid\Data\BaseData\China
Base map, Province	Prov0010.shp	Shape File	public\Covid\Data\BaseData\China
Base map, Country Boundary	chinaln.shp	Shape File	public\Covid\Data\BaseData\China
Index File for Province Census Data with 2000-2010 GIS Maps	List Maps three categories Prov 00-10.xlsx	Excel File	public\Covid\Data\BaseData\China
Index File for Prefecture Census Data with 2000-2010 GIS Maps	List Maps three categories City 00-10.xlsx	Excel File	public\Covid\Data\BaseData\China
Index File for County Census Data with 2000-2010 GIS Maps	List Maps three categories County 00-10.xlsx	Excel File	public\Covid\Data\BaseData\China
Base map, US States	ST CARTO.shp	Shape File	public\Covid\Data\COVID-19 Cases\US
Base map, US Counties	CO CARTO.shp	Shape File	public\Covid\Data\COVID-19 Cases\US
Base map, Countries	Countries.shp	Shape File	public\Covid\Data\COVID-19 Cases\World

Base maps

- China: 2000-2010 comparable census maps (province, city and county/urban district)
- USA: 1970-2010 comparable census maps (state, county)
- World: World map by country

The List of Virus Data

Country	Data Title	File Name	File Type	File Path
China	City map with virus cases	City Map 0115-0308.csv	CSV file	public\Covid\Data\COVID-19 Cases\China
China	Province map with virus cases	Prov Map 0115-0308.csv	CSV file	public\Covid\Data\COVID-19 Cases\China
China	Individual virus cases	腾讯看点-疫情小区0316V1.1.xlsx	Excel File	public\Covid\Data\COVID-19 Cases\China\Communities
China	Tracks of individual patients	轨迹数据0316V1.0.xlsx	Excel File	public\Covid\Data\COVID-19 Cases\China\trajectories
US	Confirmed cases by US county	US County Confirm Cases 0311.csv	CSV file	public\Covid\Data\COVID-19 Cases\US
US	Recovered cases by US county	US County Recover Cases 0311.csv	CSV file	public\Covid\Data\COVID-19 Cases\US
US	Death cases by US county	US County Death Cases 0311.csv	CSV file	public\Covid\Data\COVID-19 Cases\US
US	Confirmed cases by US states	US State Confirm Cases 0311.csv	CSV file	public\Covid\Data\COVID-19 Cases\US
US	Recovered cases by US statss	US State Recover Cases 0311.csv	CSV file	public\Covid\Data\COVID-19 Cases\US
US	Death cases by US states	US State Death Cases 0311.csv	CSV file	public\Covid\Data\COVID-19 Cases\US
World	Virus by Countries0302.xlsx	Virus Cases by Countries	Excel File	public\Covid\Data\COVID-19 Cases\World

Note: The virus data aggregated by city in China has been adjusted to the 2000-2010 comparable city map

Data Depository: dataverse.harvard.edu

The Dataverse® Project

Open source research data repository software

Researchers

Enjoy full control over your data. Receive *web visibility, academic credit,* and *increased citation counts.* A personal dataverse is easy to set up, allows you to display your data on your personal website, can be branded uniquely as your research program, makes your data more discoverable to the research community, and satisfies data management plans. Want to set up your personal dataverse?

Journals

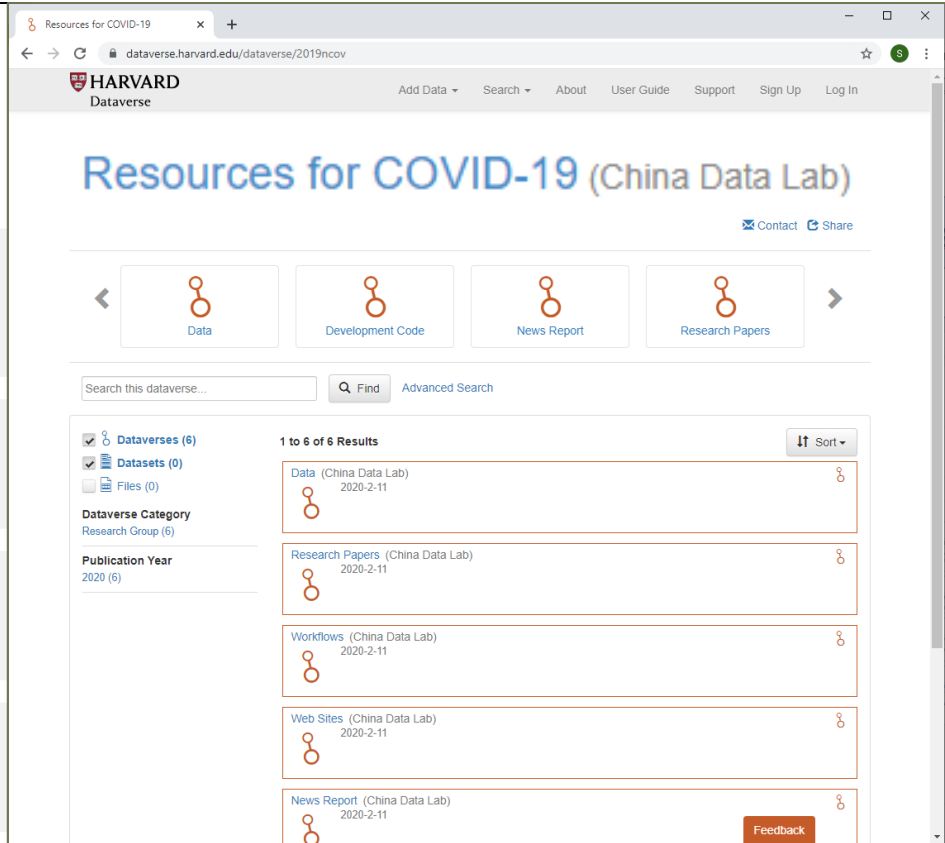
Seamlessly manage the submission, review, and publication of data associated with published articles. Establish an *unbreakable link* between *articles in your journal* and *associated data.* Participate in the open data movement by using Dataverse as part of your journal data policy or list of repository recommendations. Want to find out more about journal dataverses?

Institutions

Establish a research data management solution for your community. Federate with a growing list of Dataverse repositories worldwide for increased discoverability of your community's data. Participate in the drive to set norms for sharing, preserving, citing, exploring, and analyzing research data. Want to install a Dataverse repository?

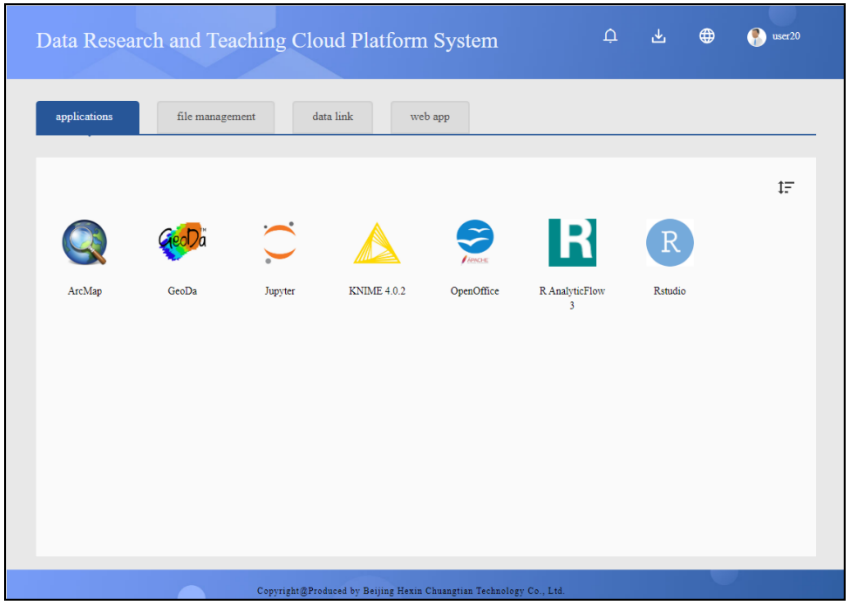
Developers

Participate in a vibrant and growing community that is helping to drive the norms for sharing, preserving, citing, exploring, and analyzing research data. Contribute code extensions, documentation, testing, and/or standards. *Integrate research analysis, visualization* and *exploration tools,* or other research and data archival systems with Dataverse. Want to contribute?



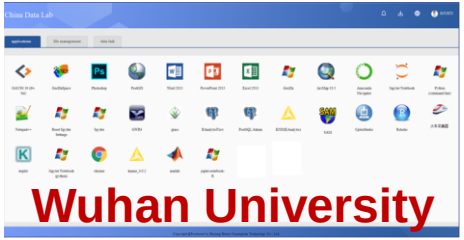
Data Cloud

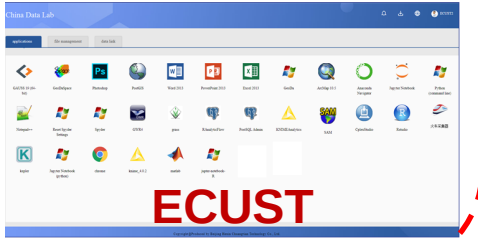
CDL in Harvard



http://harvard.chinadatalab.org

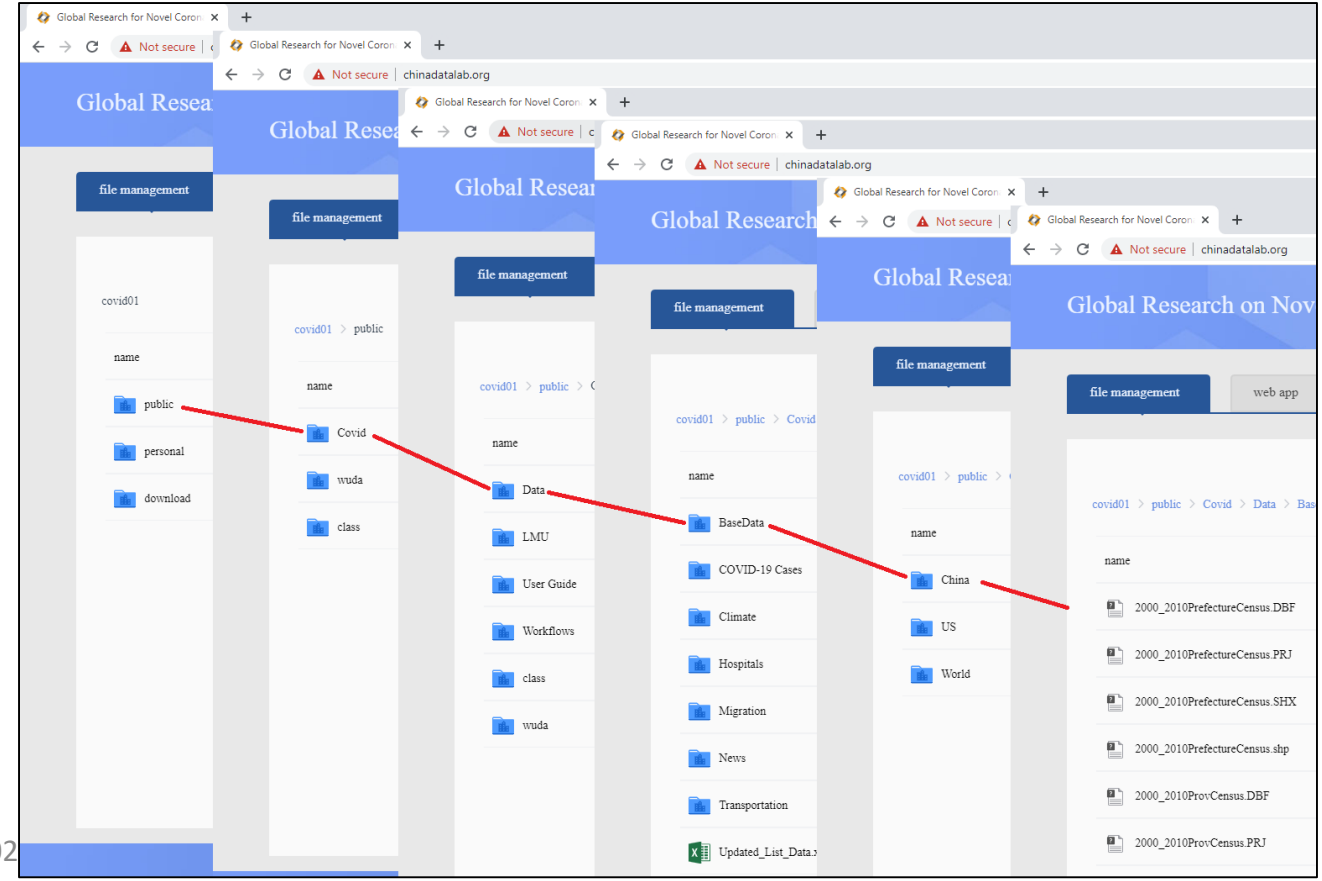
CDL in China





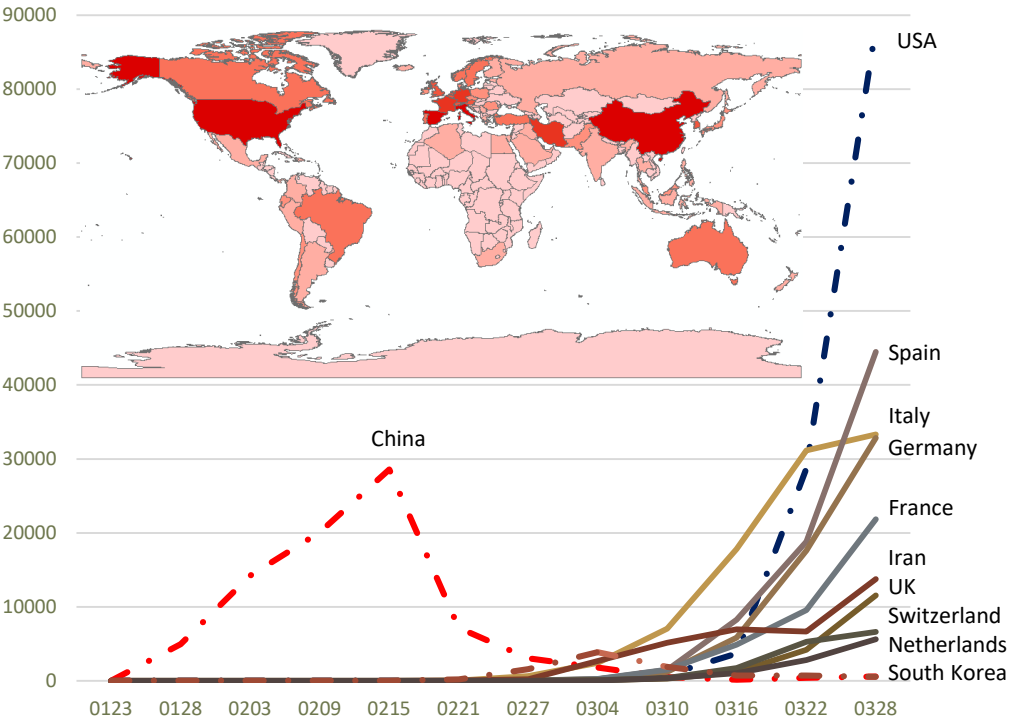
http://chinadatalab.org

The Hierarchical Data Structure

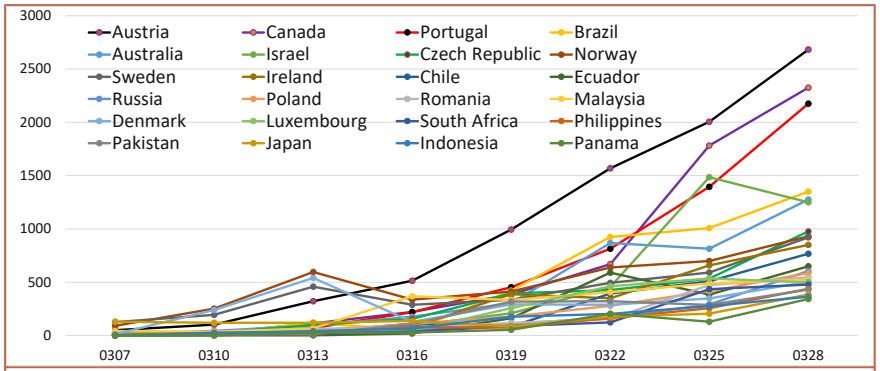


Sample Work of Virus Data Analysis

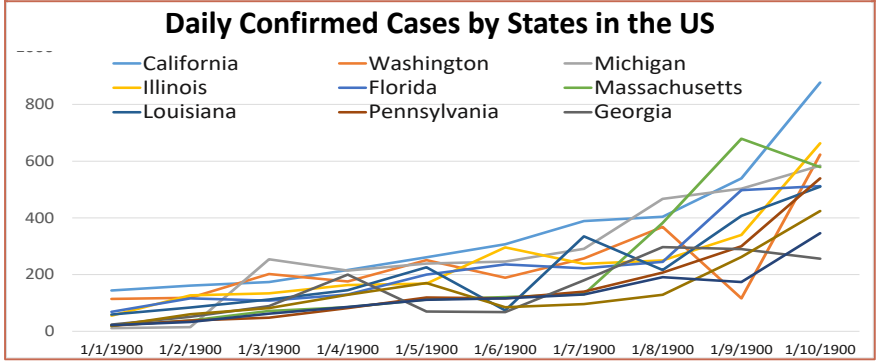
Countries with Mostly Confirmed Cases (6-day average)



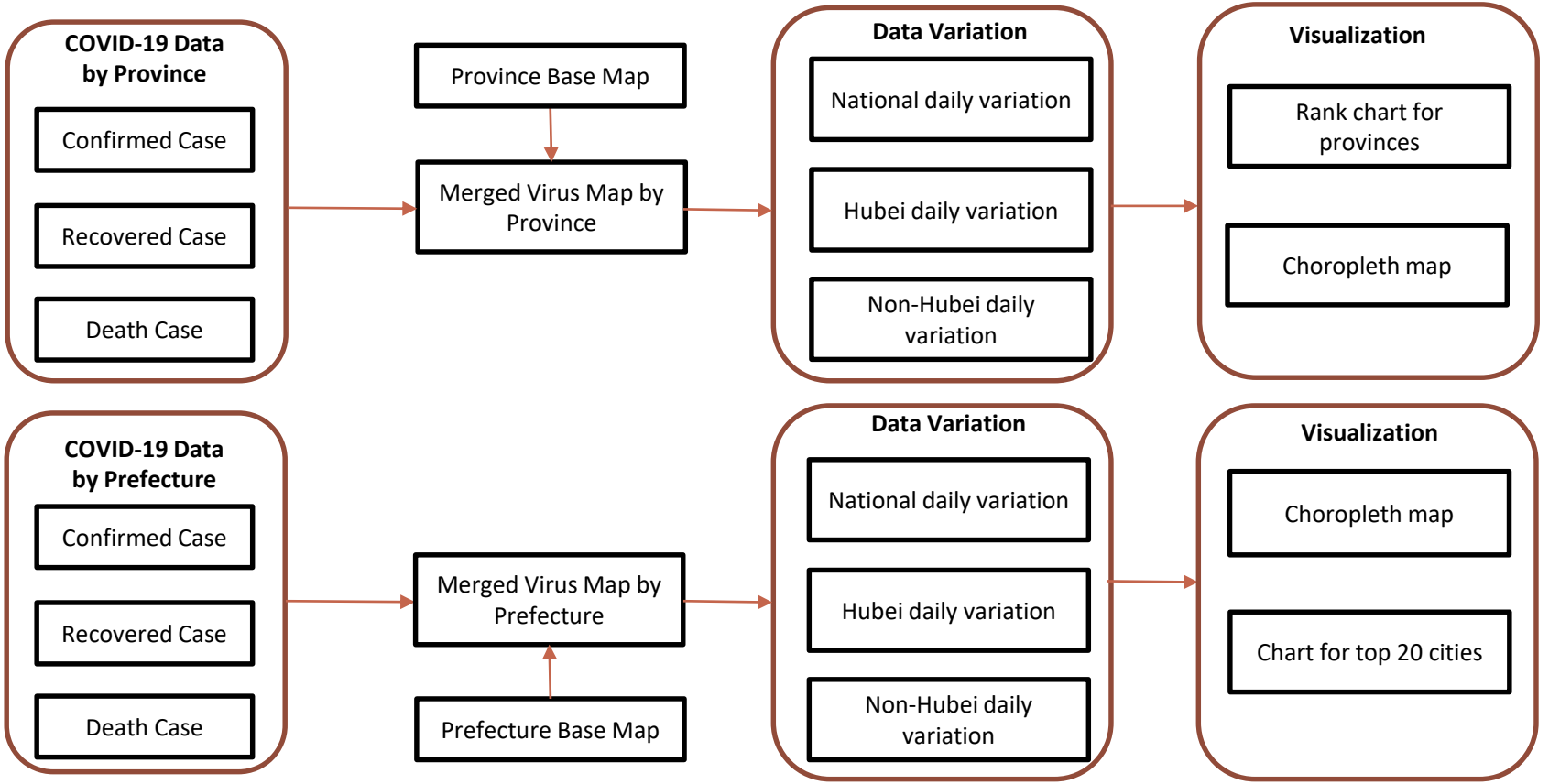
Confirmed Cases by Other Countries (3-day average)



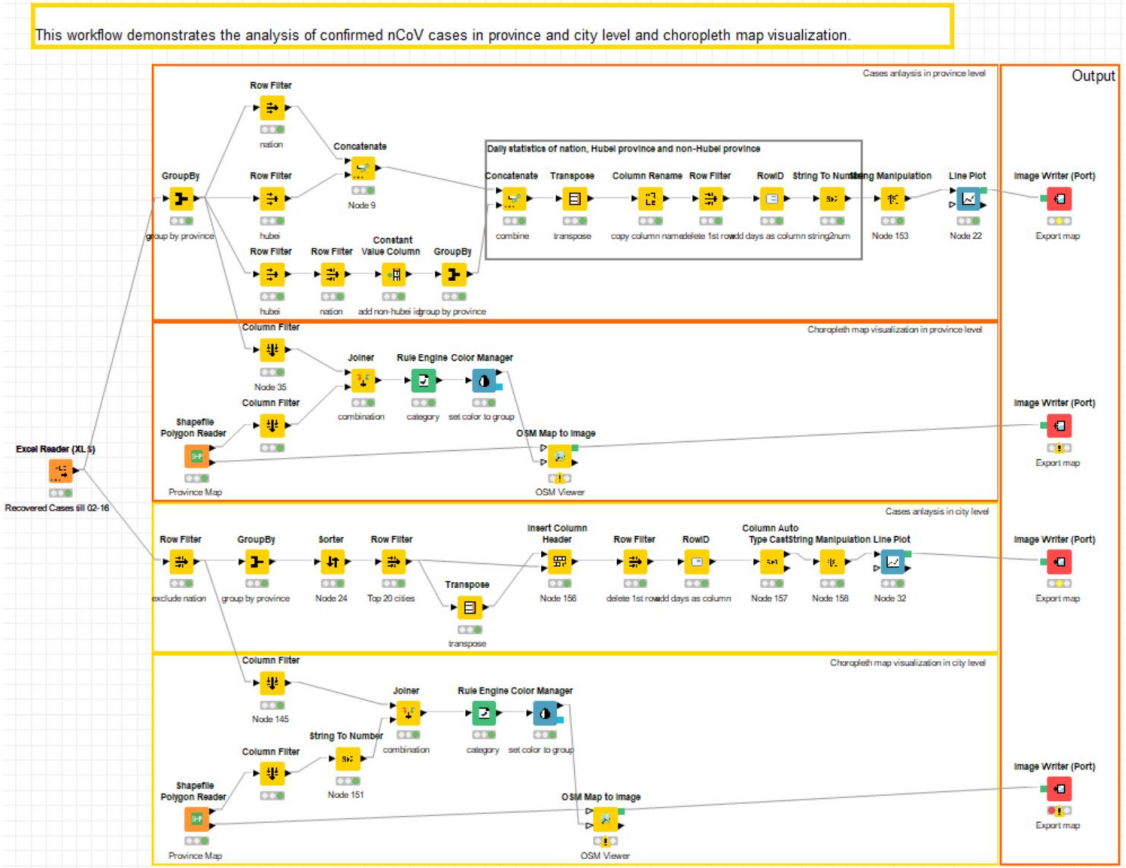
Daily Confirmed Cases by States in the US



The Flowchart for Virus Data Integration & Visualization



The Workflow for Prefecture City Data Analysis

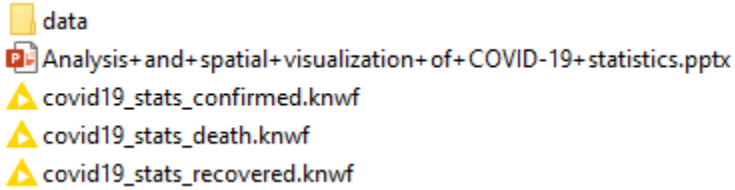


Workflow in KNIME

Open for Innovation
KNIME

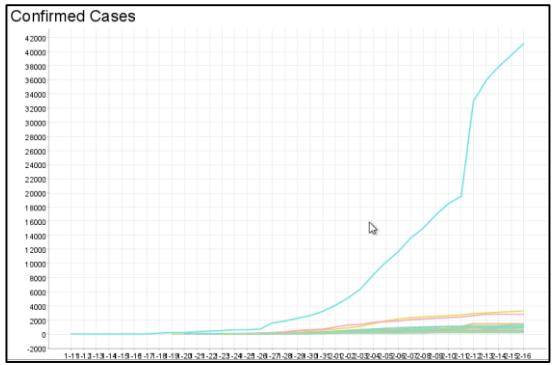
The Case Study of China Virus Data Analysis

- Workflow and data for downloading from DataVerse: https://doi.org/10.7910/DVN/FWOPW2
- Workflow and data on the cloud: http://chinadatalab.org
  - File folder: Public/Covid/Workflows/00_cases_stats
  - Workflow and data Files:

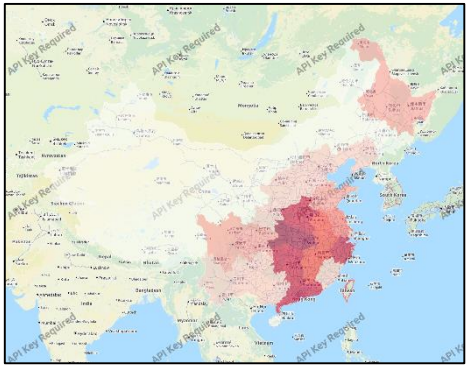


- Outputs from the workflow:

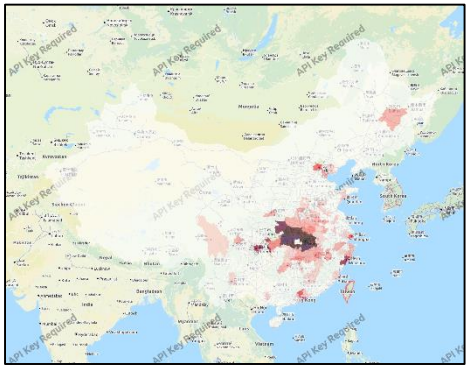
Confirmed Virus Cases by Province



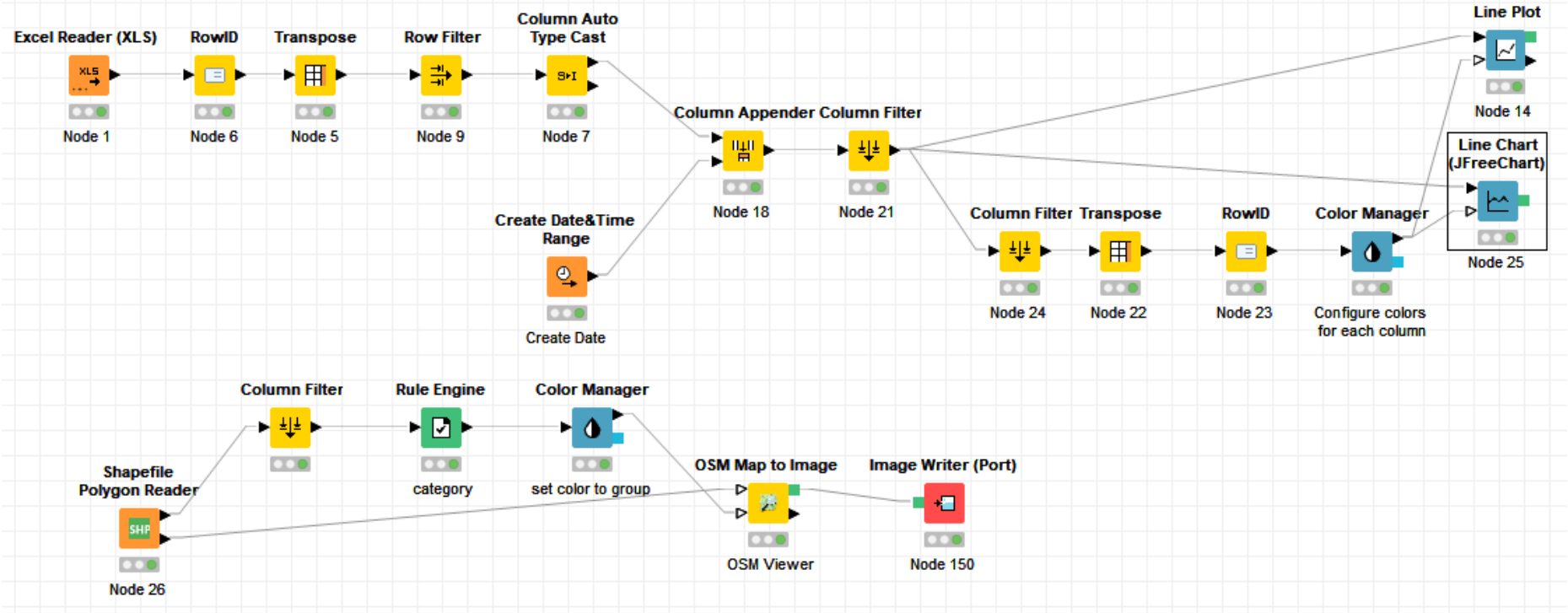
Confirmed Virus Cases by Province



Confirmed Virus Cases by City

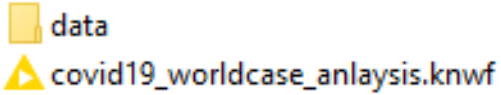


The Workflow for Global Virus Data Analysis



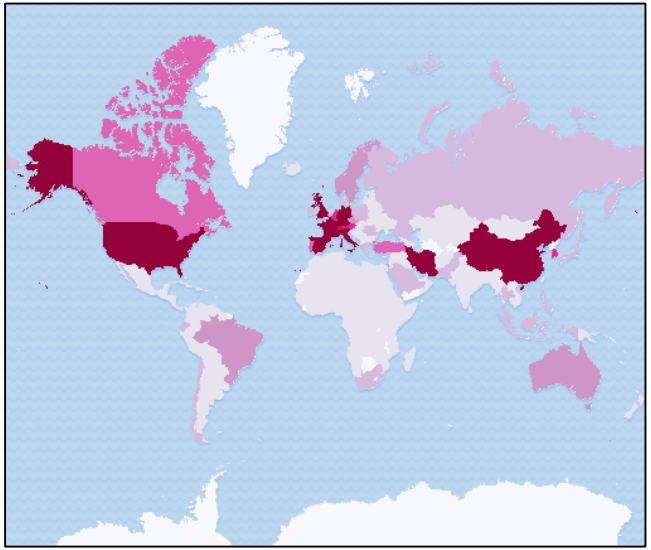
The Case Study of Global Virus Data Analysis

- Website: http://chinadatalab.org
- File folder: Public/Covid/Workflows/03_world_cases_anlaysis
- Workflow and data Files:

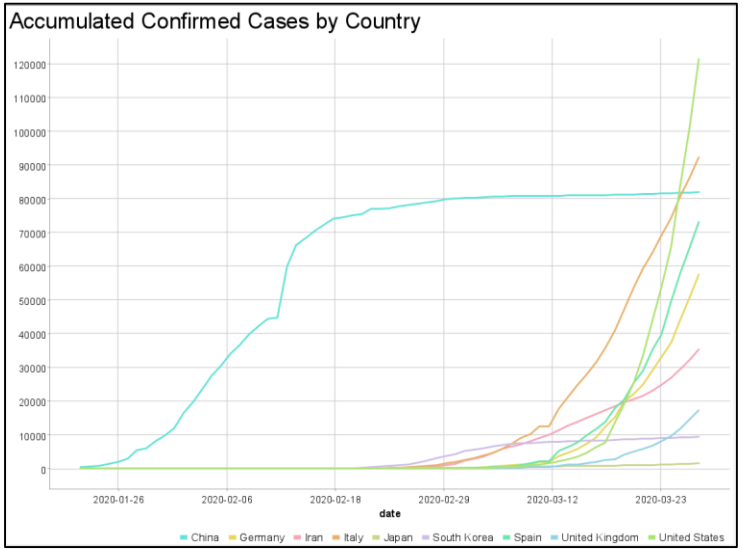


- Outputs:

Global Distribution of Confirmed Virus Cases



Accumulated Confirmed Virus Cases by Country



Discussion

Possible impacts of the following factors on the data accuracy:

- Changes in official standards for virus tests
- Changes in policies and regulations
- Changes in local leaderships

See "Policies and Regulations Timeline", https://doi.org/10.7910/DVN/OAM2JK

Possible impacts of the following factors on the spatio-temporal patterns:

- Economic linkages
- Migration
- Transportation
- Health facility
- Spatial distribution of population and business

Websites and Contact

Resources for Coronavirus Study

http://chinadatalab.net

The Cloud for COVID-19 Study

Http://chinadatalab.org

Webinars on Virus Study

https://dataverse.harvard.edu/dataverse/cdl_training

covid-2019@umich.edu