



NANYANG  
TECHNOLOGICAL  
UNIVERSITY  
SINGAPORE

# COVID-19 statistics in China

Assoc Prof Cheong Siew Ann

Dr Liu Wenyuan

SPMS, NTU



# Where the story began

- Around late Feb 2020, COVID-19 at its peak in China
- Wanted to study its propagation, but no suitable open dataset available
- Preprints do not share datasets, difficult/impossible to reproduce their results
- Decided to build our own dataset, and share with community

# COVID-19 data

- Daily reports from China government websites, extracted relevant information manually

## 2020年2月16日江西省新型冠状病毒肺炎疫情情况

发布时间：2020-02-17

【字体：大 中 小】 打印本

Date

New Infected

New Recovered

New Infected (divisions)

2020年2月16日0-24时，江西省报告新型冠状病毒肺炎新增确诊病例5例，新增治愈出院病例36例。

新增确诊病例中，宜春市2例、南昌市1例、景德镇市1例、萍乡市1例。新增治愈出院病例中，南昌市14例、新余市7例、宜春市4例、上饶市4例、景德镇市2例、抚州市2例、萍乡市1例、赣州市1例、吉安市1例。

截至2月16日24时，江西省现有确诊病例654例。累计报告新型冠状病毒肺炎确诊病例930例，其中治愈出院病例275例，死亡病例1例。现有重症病例38例。现有疑似病例19例。

确诊病例中，南昌市228例、新余市129例、上饶市123例、九江市117例、宜春市106例、赣州市90例、抚州市72例、萍乡市33例、吉安18例、鹰潭市18例、景德镇市11例，南昌1544例、上饶市35例、宜春市24例、九江10例、抚州市22例、赣州10例、景德镇市3例。

Death

Severe cases

Total Infected

Total Recovered

目前追踪到密切接触者24775人，解除医学观察20491人，尚有4284人正在接受医学观察。

疑似病例19例，其中九江市7例、鹰潭市5例、抚州市4例、宜春市1例、上饶市1例、吉安市1例。

江西省卫生健康委员会

2020年2月17日

# COVID-19 data

- Daily statistics of confirmed cases (new and cumulative), recoveries (new and cumulative) and deaths (new and cumulative) at city level

	A	B	C	D	E	F	G	H	I	J	K	L
1	Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8	Column9	Column10	Column11	Column12
2	Prefectural level or County	Provincial-level	地级行政区或县级行政区	省级行政区	2020-01-20	2020-01-21	2020-01-22	2020-01-23	2020-01-24	2020-01-25	2020-01-26	2020-01-27
3	Beijing Municipality	Beijing Municipality	北京市	北京市	3.0	5.0	4.0	12.0	10.0	15.0	17.0	12.0
4	Tianjin Municipality	Tianjin Municipality	天津市	天津市		2.0	2.0	1.0	3.0	2.0	4.0	9.0
5	Tangshan	Hebei	唐山市	河北省			0.0	0.0	0.0	0.0	0.0	1.0
6	Cangzhou	Hebei	沧州市	河北省			1.0	0.0	1.0	0.0	0.0	3.0
7	Zhangjiakou	Hebei	张家口市	河北省			0.0	0.0	0.0	0.0	0.0	0.0
8	Baoding	Hebei	保定市	河北省			0.0	0.0	1.0	2.0	0.0	0.0
9	Handan	Hebei	邯郸市	河北省			0.0	0.0	0.0	2.0	0.0	2.0
10	Langfang	Hebei	廊坊市	河北省			0.0	0.0	0.0	0.0	2.0	4.0
11	Shijiazhuang	Hebei	石家庄市	河北省			1.0	0.0	3.0	1.0	2.0	2.0
12	Xingtai	Hebei	邢台市	河北省			0.0	0.0	0.0	0.0	1.0	1.0
13	Qinhuangdao	Hebei	秦皇岛市	河北省			0.0	0.0	0.0	0.0	0.0	0.0
14	Hengshui	Hebei	衡水市	河北省			0.0	0.0	0.0	0.0	0.0	2.0
15	Chengde	Hebei	承德市	河北省			0.0	0.0	1.0	0.0	0.0	0.0
16	Taiyuan	Shanxi	太原市	山西省			1.0	0.0	1.0	0.0	0.0	1.0
17	Datong	Shanxi	大同市	山西省			0.0	0.0	1.0	0.0	0.0	0.0
18	Yangquan	Shanxi	阳泉市	山西省			0.0	0.0	0.0	1.0	0.0	0.0
19	Changzhi	Shanxi	长治市	山西省			0.0	0.0	1.0	0.0	0.0	0.0
20	Jincheng	Shanxi	晋城市	山西省			0.0	0.0	0.0	0.0	0.0	0.0
21	Shuozhou	Shanxi	朔州市	山西省			0.0	0.0	0.0	0.0	2.0	0.0
22	Jinzhong	Shanxi	晋中市	山西省			0.0	0.0	0.0	2.0	0.0	3.0
23	Xinzhou	Shanxi	忻州市	山西省			0.0	0.0	0.0	0.0	0.0	0.0

# Our dataset

- Cleaned data and organized into 6 csv files
- Published dataset on GitHub
- Attention tracked (i.e. Watch, Star, Fork)

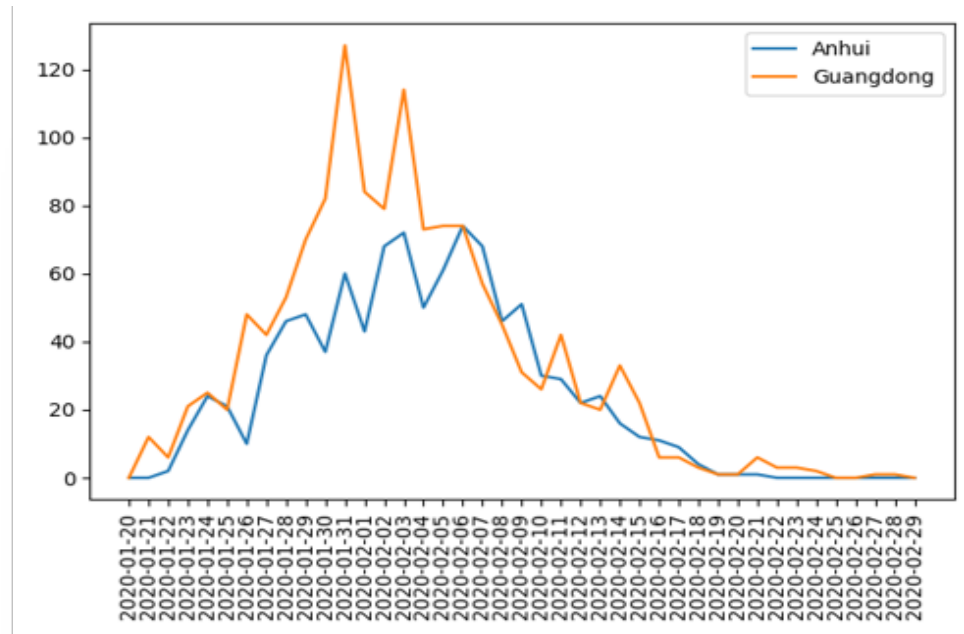
The screenshot shows the GitHub repository page for 'cheongsa / Coronavirus-COVID-19-statistics-in-China'. The repository is highlighted with a red box, showing 3 Watchers, 4 Stars, and 4 Forks. The repository is currently on the 'master' branch with 1 branch and 0 tags. The commit history shows a recent update to the README.md file by peter308 on April 8, with 42 commits. The file list includes 0201.png, 0211.png, 0221.png, and six CSV files: China\_accumulated\_deaths.csv, China\_accumulated\_infections.csv, China\_accumulated\_recoveries.csv, China\_daily\_new\_deaths.csv, China\_daily\_new\_infections.csv, and China\_daily\_new\_recoveries.csv. The right sidebar shows the 'About' section with a description of the dataset, a README link, and a CC0-1.0 License. The 'Releases' and 'Packages' sections indicate no releases or packages published.

<https://github.com/cheongsa/Coronavirus-COVID-19-statistics-in-China>

# Our dataset

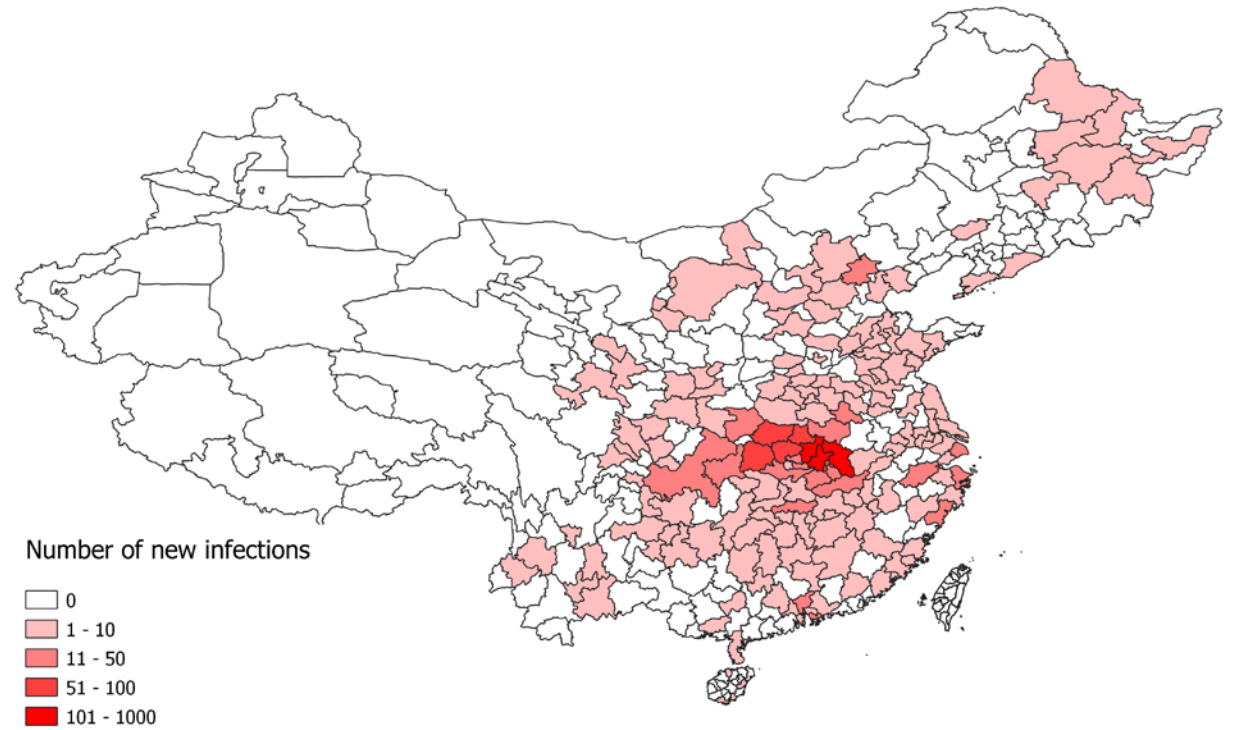
- GitHub used for code versioning and sharing

```
province_data = new_infections.groupby('Provincial-level regions').sum().reset_index()
dates = province_data.columns[1:].tolist()
Anhui = province_data[province_data['Provincial-level regions'] == "Anhui"]
Guangdong = province_data[province_data['Provincial-level regions'] == "Guangdong"]
fig, ax = plt.subplots()
ax.plot(dates, Anhui.iloc[0][1:].tolist(), label="Anhui")
ax.plot(dates, Guangdong.iloc[0][1:].tolist(), label="Guangdong")
ax.set_xticklabels(dates, rotation='vertical')
plt.legend()
```



# Our dataset

- Received many enquiries from social media (3,333 views on LinkedIn)
- Many requests on ResearchGate





# Publishing our dataset on DR-NTU (Data)

- With the help from Library, we published our dataset and documentation on [DR-NTU \(Data\)](#)
- This gives our data a DOI “[doi:10.21979/N9/A2XWCW](https://doi.org/10.21979/N9/A2XWCW)”
- DataCite DOI required for inclusion in Web of Science Data Citation Index
- **Cite Dataset** feature facilitates proper data citation and attribution

The screenshot shows the Dataverse interface for a dataset titled "COVID-19 Statistics in China". The page is for user Cheong Siew Ann from Nanyang Technological University. The breadcrumb trail is: DR-NTU (Data) > School of Physical and Mathematical Sciences (SPMS) > Cheong Siew Ann > COVID-19 Statistics in China. There are links for "Contact" and "Share". The dataset version is 1.0. The citation information is: Cheong, Siew Ann; Liu, Wenyuan; Yen, Tsung-Wen Peter, 2020, "COVID-19 Statistics in China", <https://doi.org/10.21979/N9/A2XWCW>, DR-NTU (Data), V1, UNF:6:Q5f5mkGTx4bTDGqjjhVDA== [fileUNF]. A "Cite Dataset" button is highlighted with a red box. The "Dataset Metrics" section shows 4 Downloads. The description states: "A data set on COVID-19 pandemic in China, which covers daily statistics of confirmed cases (new and cumulative), recoveries (new and cumulative) and deaths (new and cumulative) at city/county level. All data are extracted from Chinese government reports." The subject is "Medicine, Health and Life Sciences" and the keyword is "Coronavirus, COVID-19, statistics, China". The related publication is: Liu, W., Yen, P. T. W., & Cheong, S. A. (2020). Spatial-Temporal Dataset of COVID-19 Outbreak in China. arXiv preprint arXiv:2003.11716. arXiv: arXiv:2003.11716v2. The page has tabs for "Files", "Metadata", "Terms", and "Versions". There are "Change View" buttons for "Table" and "Tree". A search bar is present with the text "Search this dataset..." and a "Find" button. At the bottom, there are filters for "File Type: All", "Access: All", and "File Tag: All", along with a "Sort" button.

# Publishing our dataset on DR-NTU (Data)

- Discoverability of published dataset in [Google Dataset Search](#)
- Easily accessible to research community

The screenshot shows a Google search for "COVID-19 china" on the Google Dataset Search platform. The search results are filtered by "Free" and "Clear filters". The top result is "COVID-19 Statistics in China" by researchdata.ntu.edu.sg, updated on Oct 1, 2020. A red box highlights the "Explore at DR-NTU (Data)" button. Other results include "China Coronavirus Cases" by tradingeconomics.com and "00\_COVID19 China Stats Analysis" by dataverse.harvard.edu.

Google

COVID-19 china

Last updated Download format Usage rights Topic Free Clear filters

pdf

China Coronavirus Cases tradingeconomics.com

COVID-19 Statistics in China researchdata.ntu.edu.sg Updated Oct 1, 2020

00\_COVID19 China Stats Analysis dataverse.harvard.edu search.datacite.org Updated May 14, 2020

pptx, txt, bin +2

COVID-19 Statistics in China

Related Article

Explore at DR-NTU (Data)

4 scholarly articles cite this dataset (View in Google Scholar)

Unique identifier <https://doi.org/10.21979/N9/A2XWCW>

Dataset updated Oct 1, 2020

Dataset provided by DR-NTU (Data)

License Attribution-NonCommercial 1.0 (CC BY-NC 1.0) License information was derived automatically

Time period covered Jan 20, 2020 - Feb 29, 2020



# Conclusion

---

- Data sharing lowers information barriers and speeds up research
- Data sharing encourages other researchers to follow your work
- Publish your dataset with DOI can make it easy for others to access your dataset