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[illegible]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。

Death

Severe cases

Total Infected

Total Recovered

确诊[illegible]例中，南昌市228例、新[illegible]129例、上饶市123例、九江市117[illegible]宜春市106例、赣州[illegible]例、抚州市72例、萍乡市33例、吉安[illegible]潭市18例、景德[illegible]1例[illegible]，南昌[illegible]44例、上饶市35例、宜春市24例、九[illegible]抚州市22例、赣州[illegible]10例[illegible]德镇市3[illegible]。

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

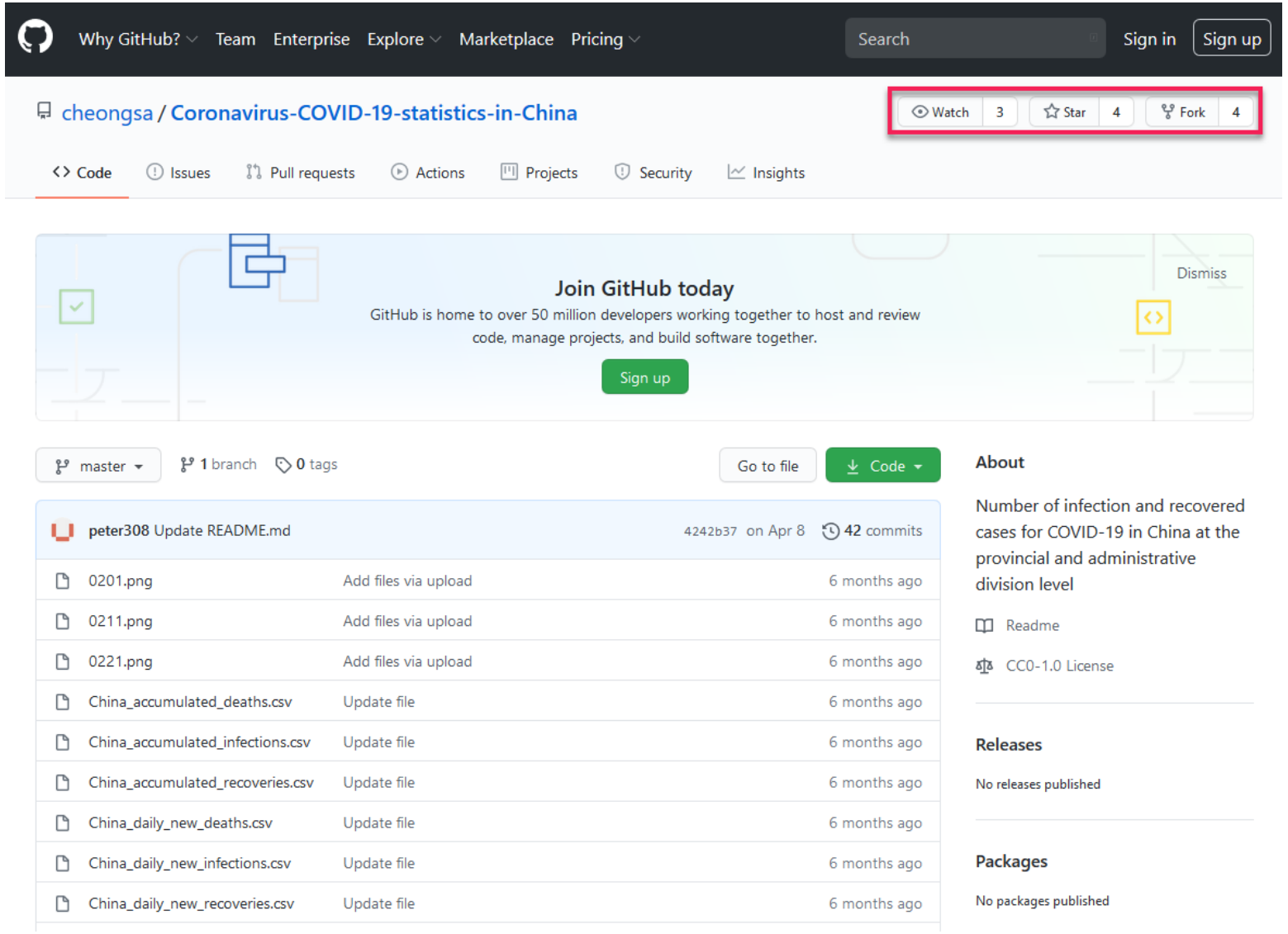
COVID-19 data

- 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 Cou	Provincial-le	地级行政区或县	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 Muni	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 Muni	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)

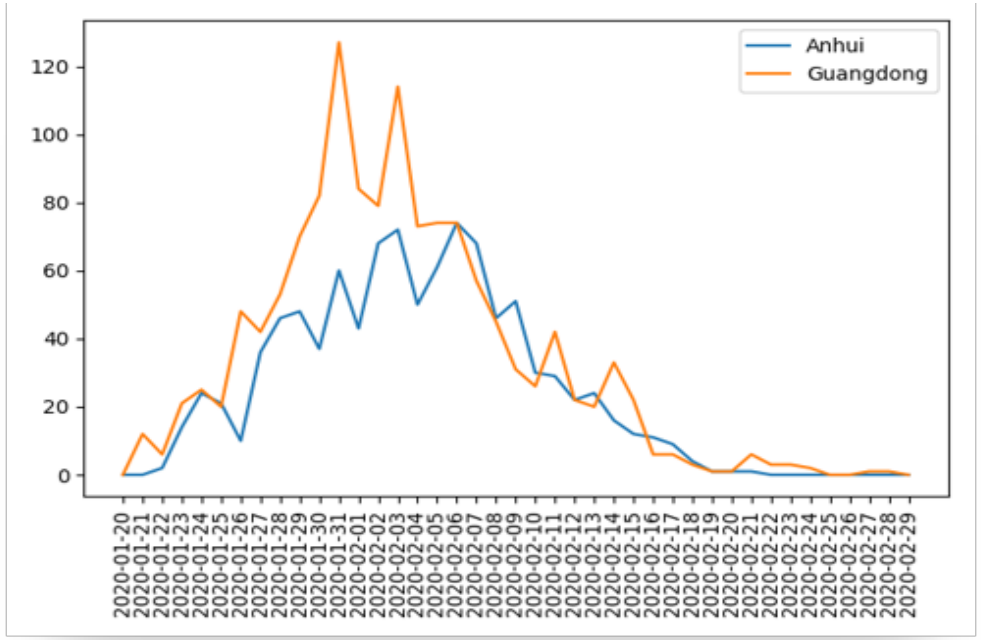


https://github.com/cheongsa/Coronavirus-COVID-19-statistics-in-China

Our dataset

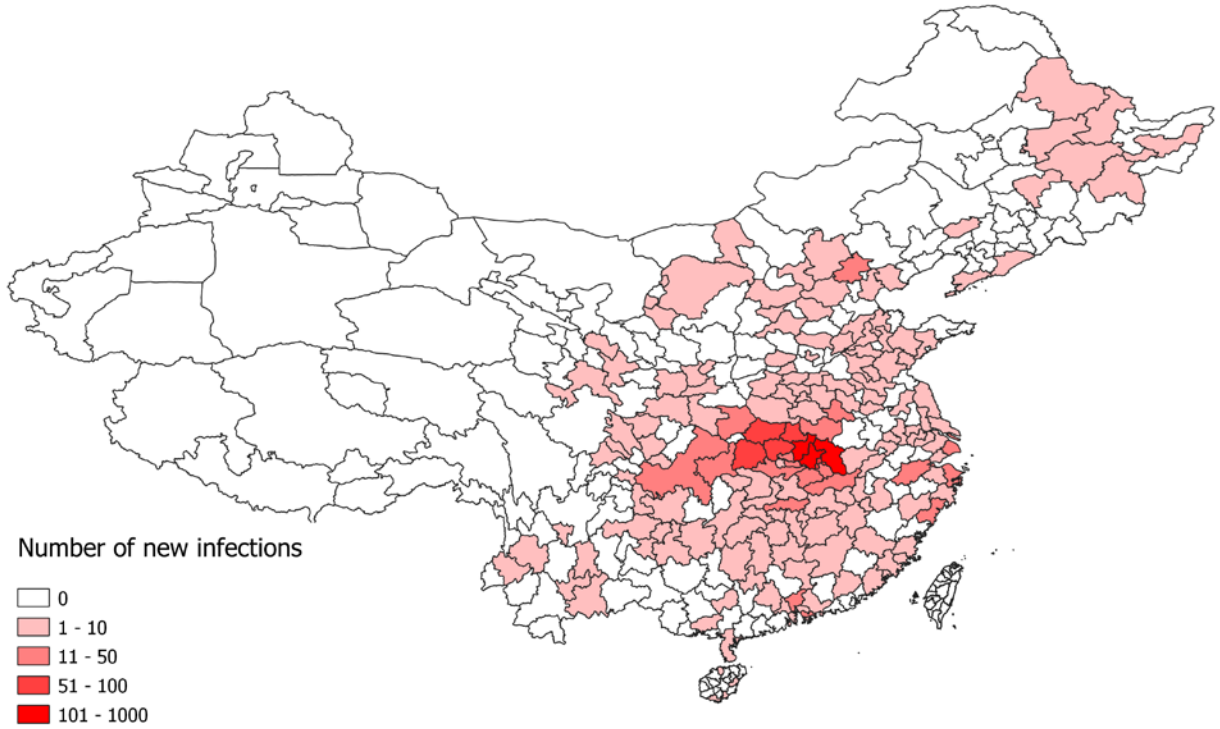
- GitHub used for code versioning and sharing

```
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()
```



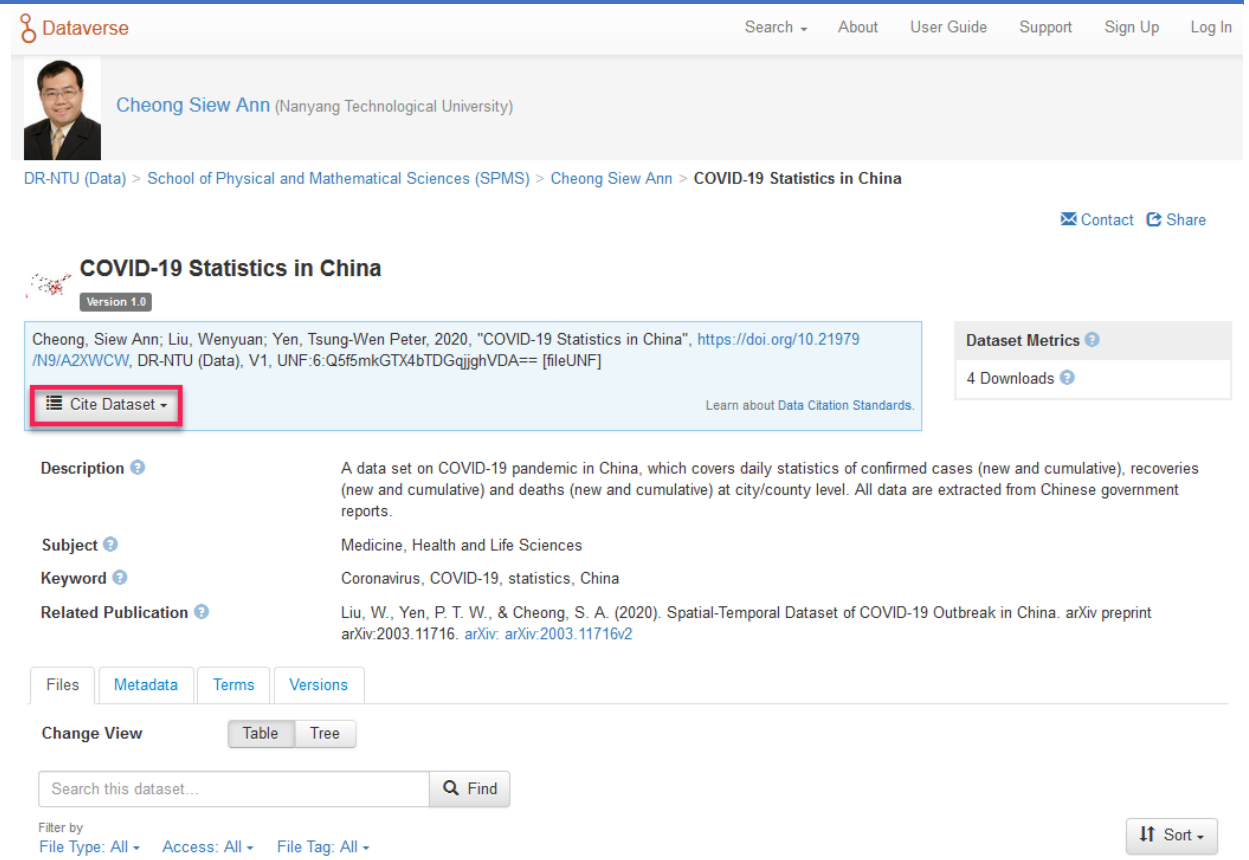
Our dataset

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



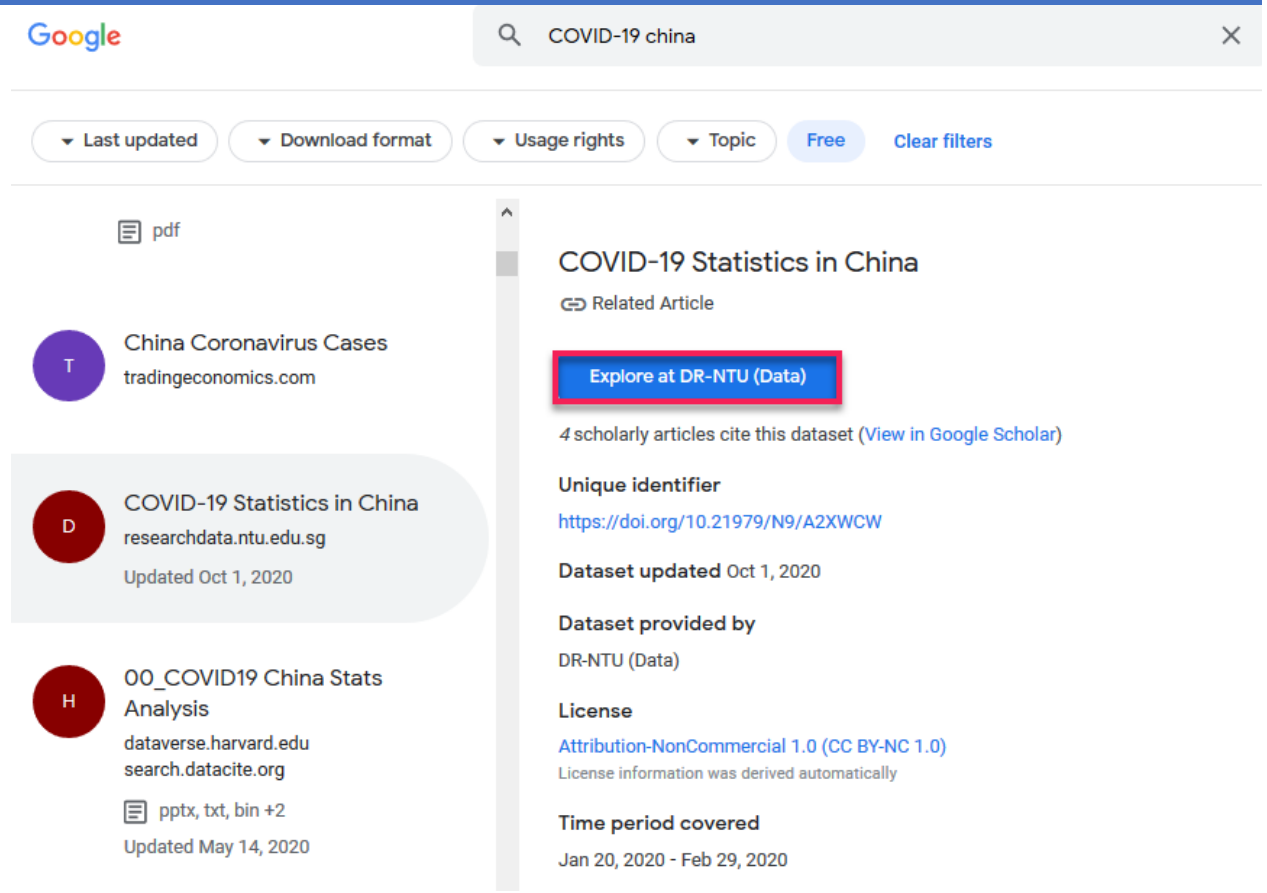
Publishing our dataset on DR-NTU (Data)

- 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"
- DataCite DOI required for inclusion in Web of Science Data Citation Index
- **Cite Dataset** feature facilitates proper data citation and attribution



Publishing our dataset on DR-NTU (Data)

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



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