新冠病毒数据分析系列公益讲座 COVID-19 Data Analysis Webinars

第五讲、疫情数据时空特性分析

杨超伟，乔治梅森大学地理和地球科学院教授

点评人：龚健雅院士，武汉大学遥感信息工程学院院长

主持人：张学良，上海财经大学长三角与长江经济带研究院执行院长

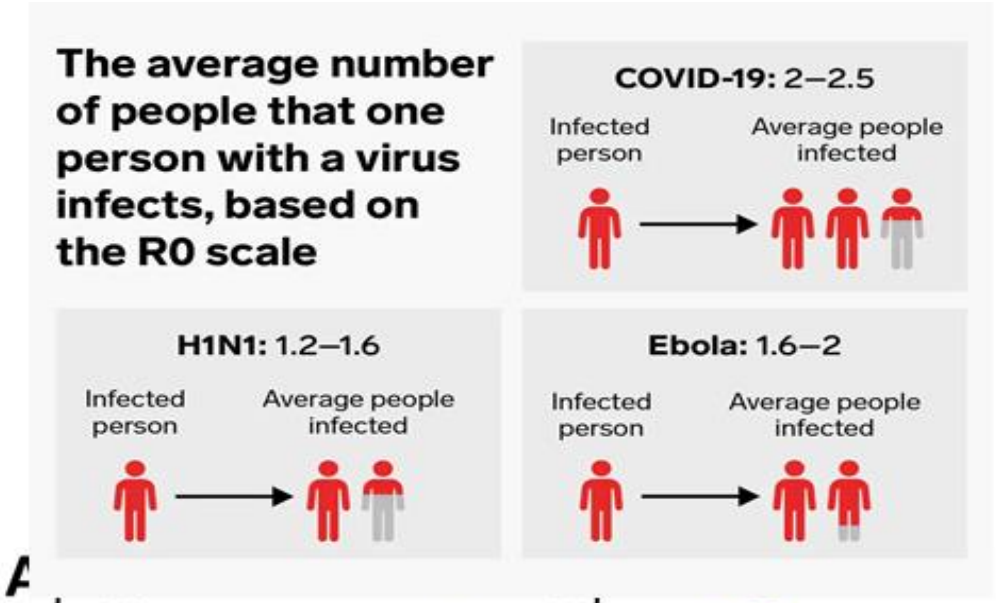
合办单位：上海财经大学长三角与长江经济带研究院

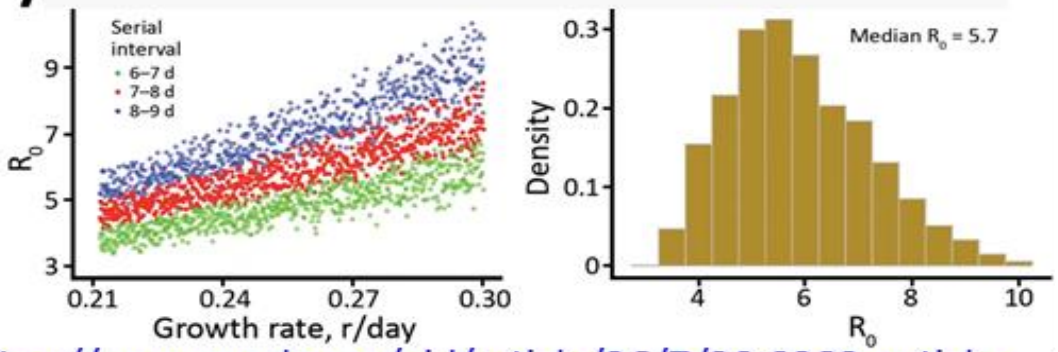
北京时间 2020年5月29日上午9-10点

Outline：提纲

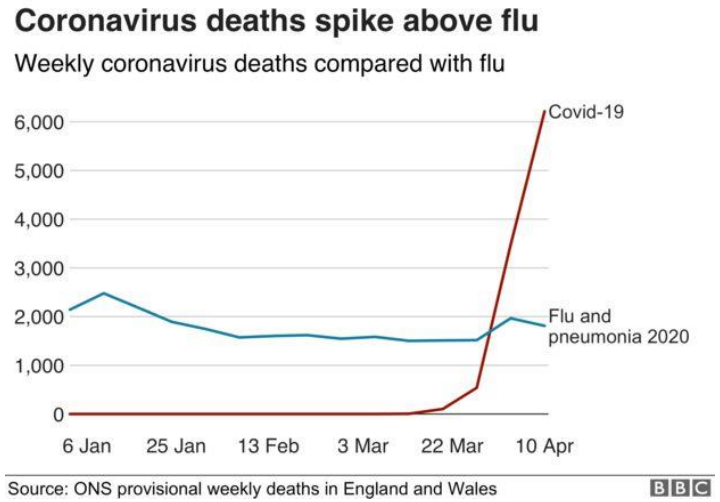
1. Introduction & Background： 研究背景
2. Computing Infrastructure： 计算平台
3. Data Collection & Sharing： 数据收集与共享
4. Spatiotemporal Pattern Analytics： 时空特性分析
5. Summary： 总结
6. Acknowledgements： 致谢

新型冠状病毒梦魇

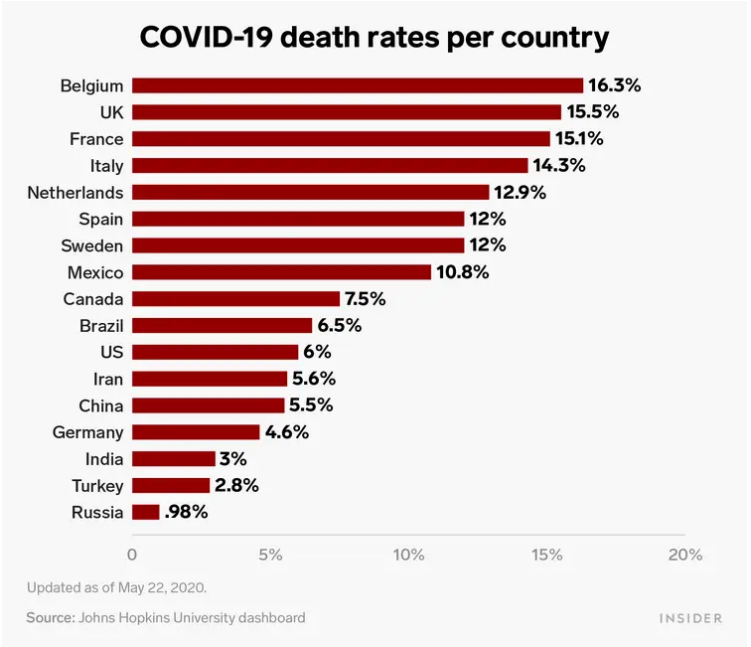




https://wwwnc.cdc.gov/eid/article/26/7/20-0282_article



https://www.bbc.com/news/health-52361519



时空快速响应

- STC 任务：建立国家范围乃至国际范围的时空架构
  - 时空思维-> 人类智慧
  - 时空计算-> 计算工具
  - **时空应用-> 国家、国际范围的重要应用: 新型冠状病毒时空研究（covid-19）**
- IAB成员支持通过的快速响应项目
  - 组成国际工作小组
  - IAB 支持
  - NSF快速响应项目
  - 亚马逊云计算支持
  - 国内外的合作

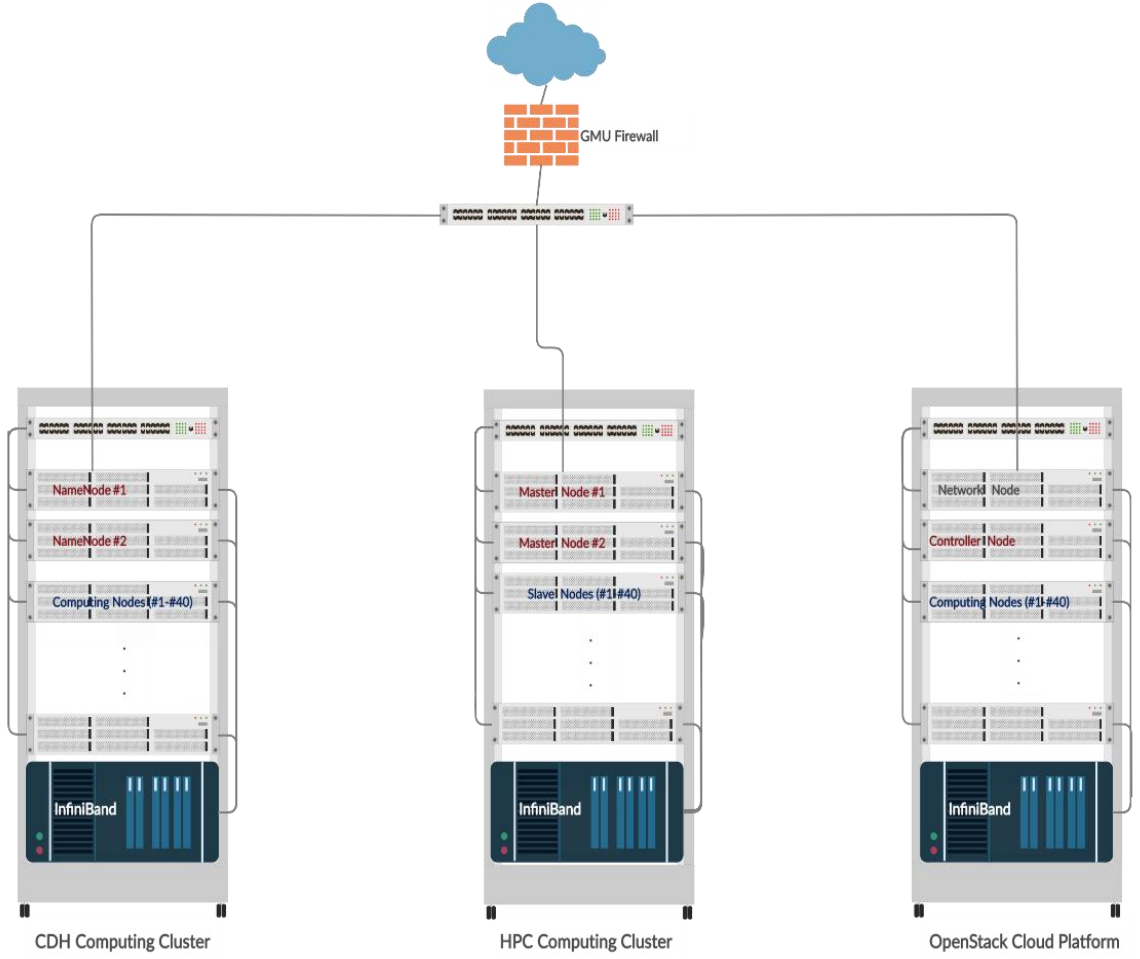
我们应对快速响应的措施

- 近似实时的数据和信息采集
- 提供收集数据的便捷访问
- 整合中心前7年时空分析工具集
- 测试整合来自我们团队的分析方法及开源产品，以及
- 利用弹性时空计算来支持数据分析处理，从而快速生成支持决策和应对的必要信息

Outline：提纲

1. Introduction & Background： 研究背景
2. Computing Infrastructure：计算平台
3. Data Collection & Sharing：数据收集与共享
4. Spatiotemporal Pattern Analytics：时空特性分析
5. Summary：总结
6. Acknowledgements： 致谢

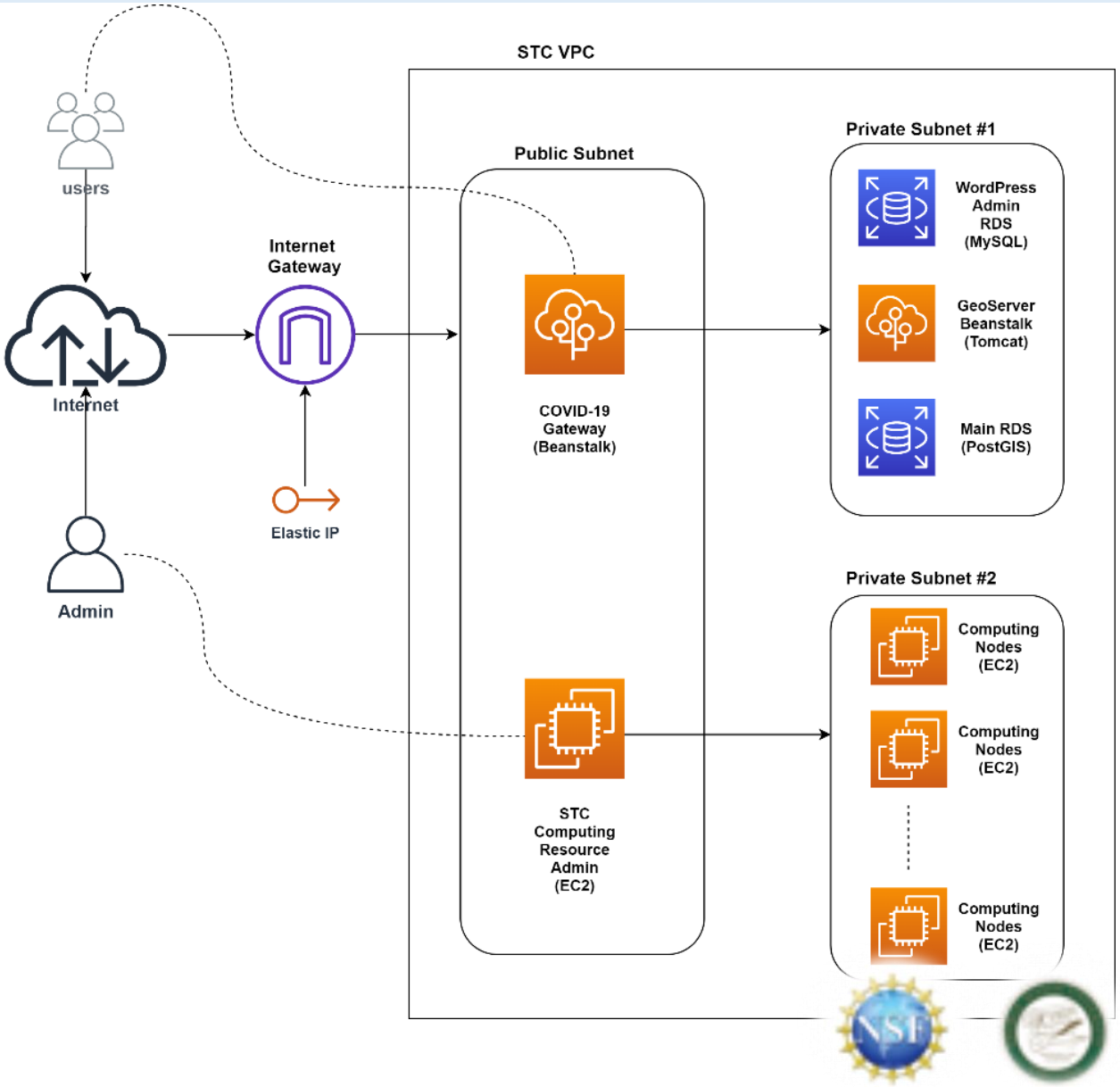
本地计算资源及架构



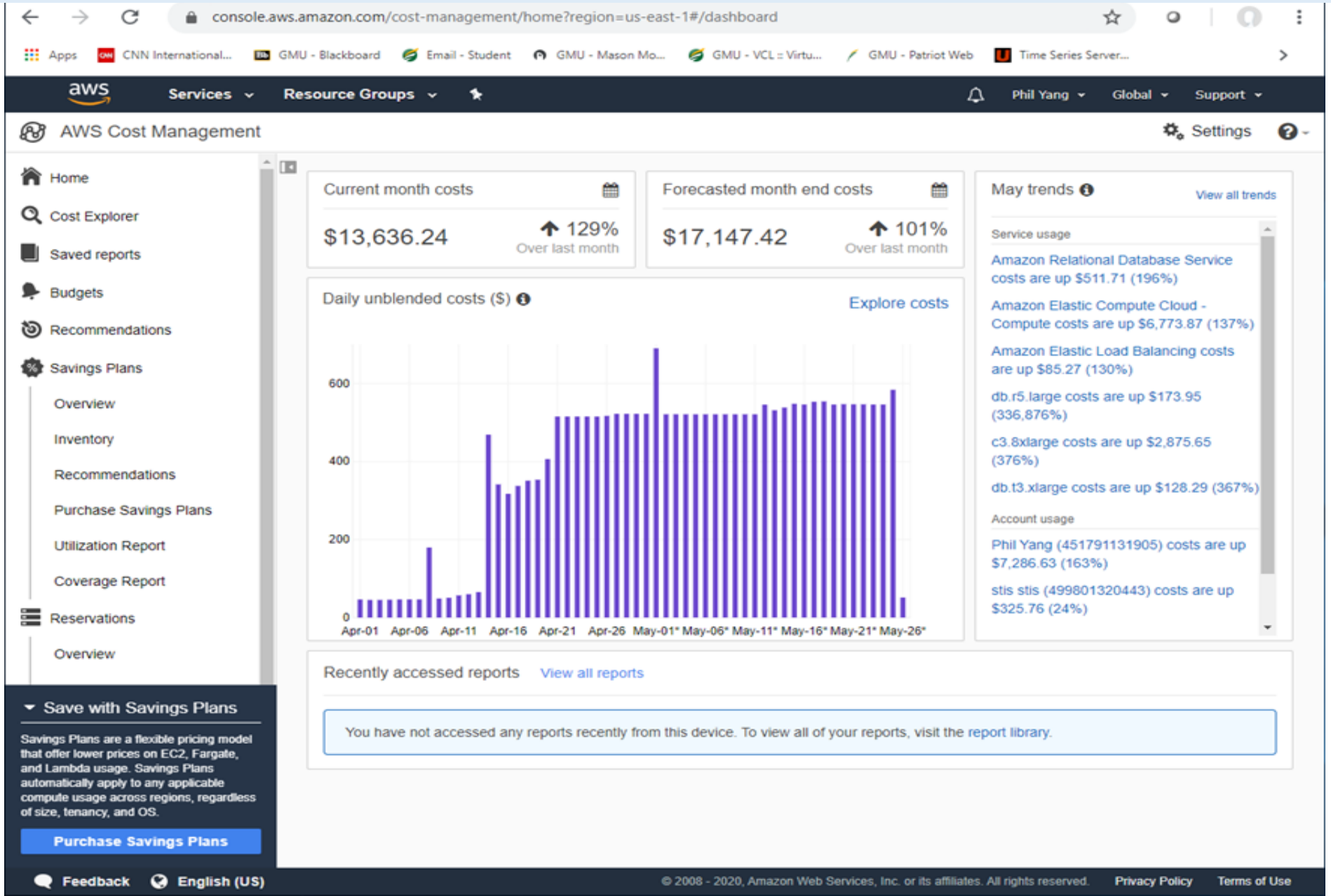
- 3个应对不同任务独立设计的计算集群总共拥有 3024 个CPU core：
  - 基于Hadoop分布式文件系统的云计算平台 (CDH)
  - 基于MPI的高性能计算集群 (HPC)
  - 基于Openstack的云共享弹性平台 (Cloud)

基于亚马逊云的高可用性应用

- 通过Beanstalk 和 RDS构建在亚马逊云上的**高可用**的covid-19 门户网站
- 应用**弹性计算**资源支持covid-19相关研究. 例如夜光遥感观测分析
- 构建了**HDFS和YARN**的弹性云计算环境以支持基于内存的分布式计算(Spark)



亚马逊公有云使用情况


console.aws.amazon.com/cost-management/home?region=us-east-1#/dashboard
Apps
CNN International...
GMU - Blackboard
Email - Student
GMU - Mason Mo...
GMU - VCL :: Virtu...
GMU - Patriot Web
Time Series Server...
aws
Services
Resource Groups
Phil Yang
Global
Support
AWS Cost Management
Settings
Home
Cost Explorer
Saved reports
Budgets
Recommendations
Savings Plans
Overview
Inventory
Recommendations
Purchase Savings Plans
Utilization Report
Coverage Report
Reservations
Overview
Current month costs
$13,636.24
129%
Over last month
Forecasted month end costs
$17,147.42
101%
Over last month
May trends
View all trends
Service usage
Amazon Relational Database Service costs are up $511.71 (196%)
Amazon Elastic Compute Cloud - Compute costs are up $6,773.87 (137%)
Amazon Elastic Load Balancing costs are up $85.27 (130%)
db.r5.large costs are up $173.95 (336,876%)
c3.8xlarge costs are up $2,875.65 (376%)
db.t3.xlarge costs are up $128.29 (367%)
Account usage
Phil Yang (451791131905) costs are up $7,286.63 (163%)
stis stis (499801320443) costs are up $325.76 (24%)
Daily unblended costs ($)
Explore costs
600
400
200
0
Apr-01 Apr-06 Apr-11 Apr-16 Apr-21 Apr-26 May-01* May-06* May-11* May-16* May-21* May-26*
Recently accessed reports
View all reports
You have not accessed any reports recently from this device. To view all of your reports, visit the report library.
Save with Savings Plans
Savings Plans are a flexible pricing model that offer lower prices on EC2, Fargate, and Lambda usage. Savings Plans automatically apply to any applicable compute usage across regions, regardless of size, tenancy, and OS.
Purchase Savings Plans
Feedback
English (US)
© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

Outline：提纲

1. Introduction & Background： 研究背景
2. Computing Infrastructure：计算平台
3. Data Collection & Sharing：数据收集与共享
4. Spatiotemporal Pattern Analytics：时空特性分析
5. Summary：总结
6. Acknowledgements： 致谢

COVID-19数据的挑战与目标

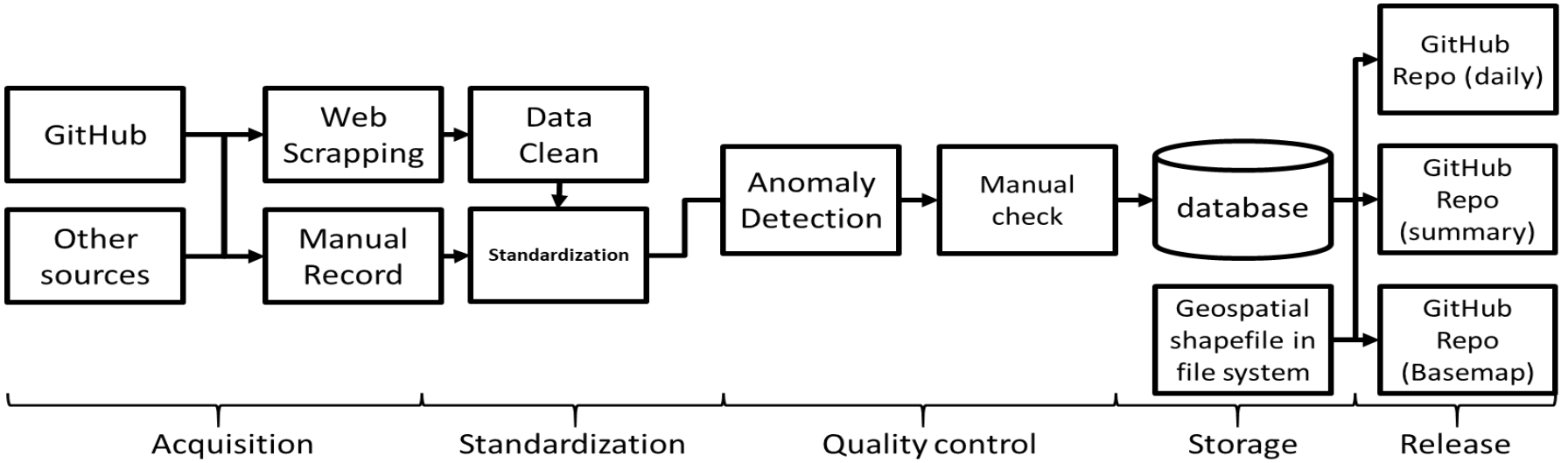
❑挑战

- 数据来自于很多机构组织，存储和发布非常分散
- 没有统一的数据标准，异质数据难以对比和相互融合
- 数据在时间上的持续性不好，长期稳定的有效数据难以获得

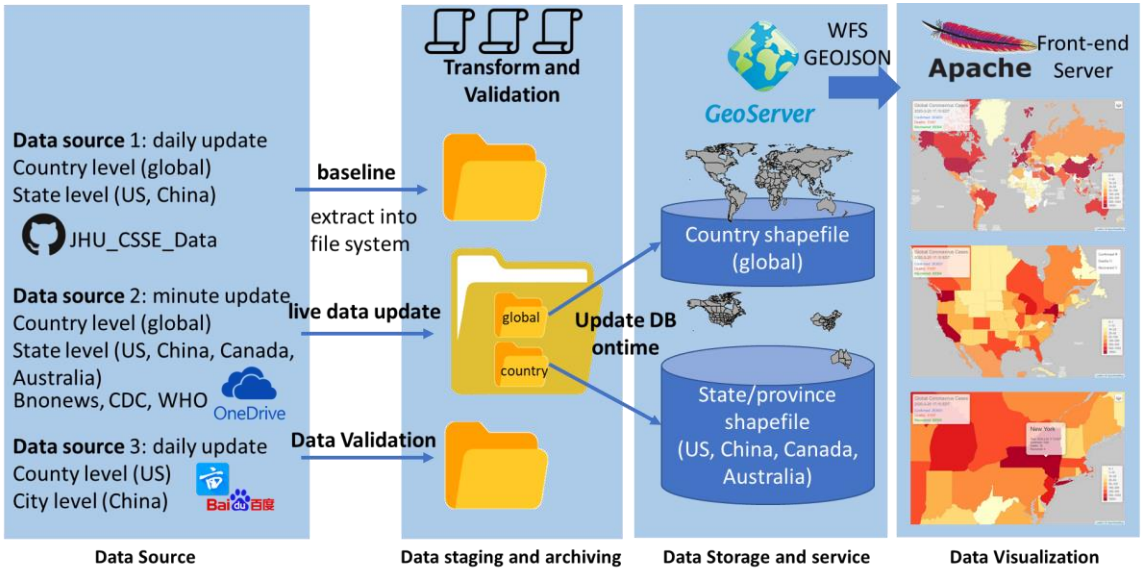
❑目标

- 做高质量的时空疫情数据集
- 整合多源数据
- 建立在不同空间尺度下的统一数据标准
- 收集长期有效的数据
- 以供疫情相关的联合交叉研究

数据工作流程

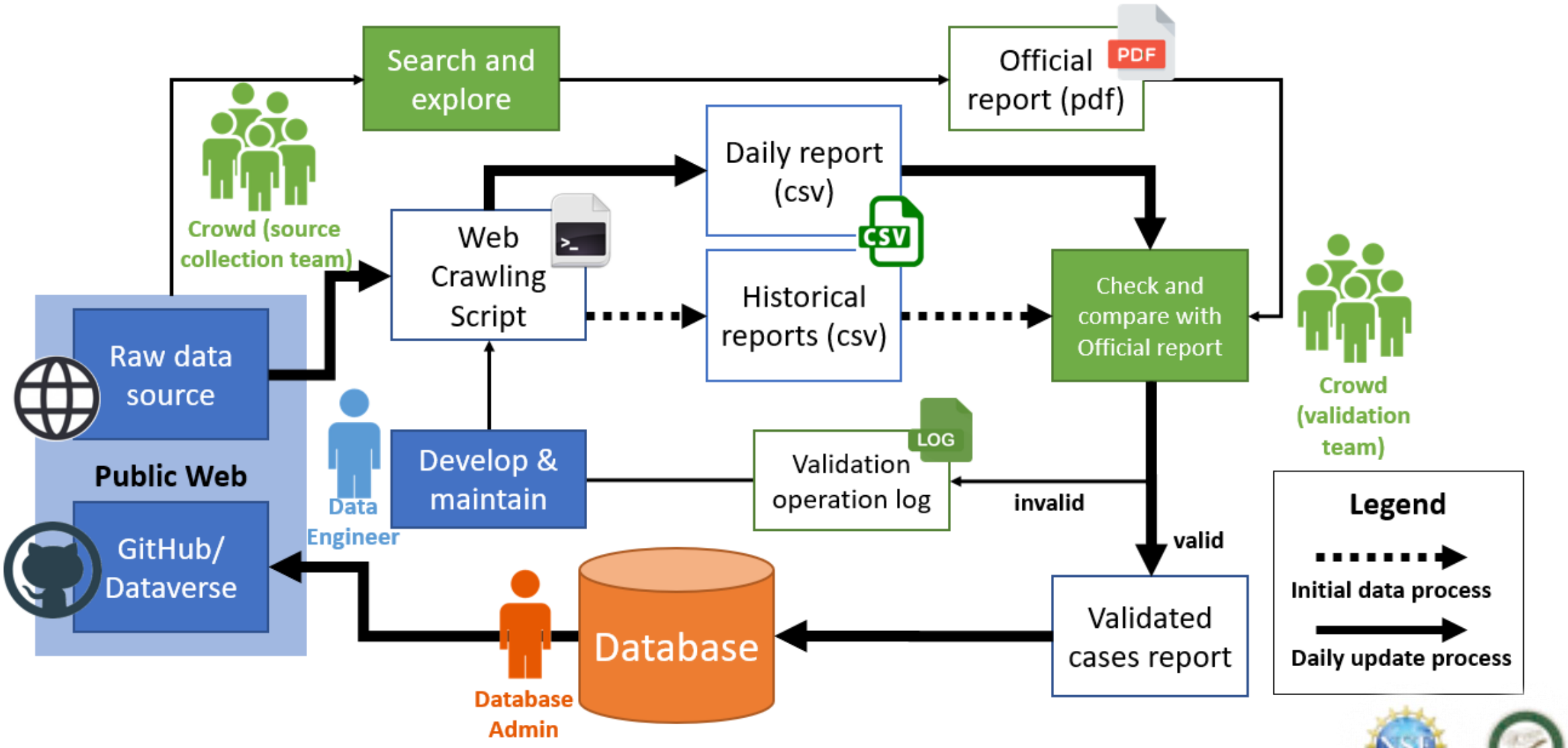


疫情数据处理流程



应用：实时疫情数据展示

基于众包的数据质量控制



GitHub

- GitHub库地址：https://github.com/stccenter/COVID-19-Data
  - 按全球/国家组织的病例数据
  - 每个国家：提供每日病例数据和病例时间序列数据

Overall data sources by Country

source data source source validaiton source

Country / Region	Continent	Admin level	DataSource
Global	Global	0	source Johns Hopkins CSSE source WHO
United States	North America	1 , 2	source Johns Hopkins CSSE source USAFACTS source CDC.gov
China	Asia	1	source Ding Xiang Yuan source NHC
Canada	North America	1	source Johns Hopkins CSSE source canada.ca
Australia	Oceania	1	source Johns Hopkins CSSE source health.gov.au
Italy	Europe	1	source protezionecivile.it source world0meters
Germany	Europe	1	source covid19-eu-data source rki.de
Austria	Europe	1	source sozialministerium.at source sozialministerium.at
Brazil	South America	1	source covid.saude.gov.br source covid.saude.gov.br
Chile	South America	1	source Covid-19 Latinoamérica source minsal.cl
Japan	Asia	1	source covid-2019.live source stopcovid.jp
Russia	Europe	1	source yandex.ru source стопкоронавирус.рф
South Africa	Africa	1	source NICD source health.gov.za source statssa.gov.za
Malaysia	Asia	1	source covid-19-malaysia source covid-19-infosihat.gov.my
Denmark	Europe	1	source covid-19-ssi.dk source coronatracker.com
Finland	Europe	1	source arcgis.com source thl.fi
Greece	Europe	1	source covid19.gov.gr source eody.gov.gr
Hungary	Europe	1	source koronavirus.gov.hu source abouthungary.hu
Croatia	Europe	1	source koronavirus.hr source world0meters
Iceland	Europe	1	source wikipedia.org source covid.is
Slovakia	Europe	1	source wikipedia.org source korona.gov.sk

Daily data

Daily data provides automatically updated information of COVID-19 cases, and related attributes daily.

Attribute Name	Description	Format	Example
date	The date representing the current day in which the data represents. UTC time is used for this dataset, all values will calculated before the end of UTC time of the date.	Date (YYYY/MM/DD) in UTC	2020/04/09
country_name	Name of the country.	string	United States
iso3	3 digit ISO country codes.	varchar(3)	USA
admin1_name	The name for admin 1 level.	string	Virginia
hasc1	This will represent the Hierarchical administrative subdivision codes (HASC) for admin 1 level.	string	US.VA (for Virginia, United States)
local_id1	This will represent the ID for specific admin 1 level. ID that represents the country's admin 1 level	string	VA (for Virginia, United States)
confirmed	The number of confirmed cases.	integer	777
death	The number of death cases.	integer	19
recovered	The number of recovered cases. (might be null for admin 2 level)	integer	null
Miscellaneous	Other data attributed to our dataset.	TBD	TBD

Summary data

Summary data records the COVID-19 cases, and related attributes, to show the timeline of cases.

Attribute Name	Description	Format	Example
country_name	Name of the country.	string	"US"
iso3	3 digit ISO country codes.	varchar(3)	USA
admin1_name	The name for admin 1 level.	string	State for USA
date	The date representing the current day in which the data represents. UTC time is used for this dataset, all values will calculated before the end of UTC time of the date.	UTC	YYYY/MM/DD

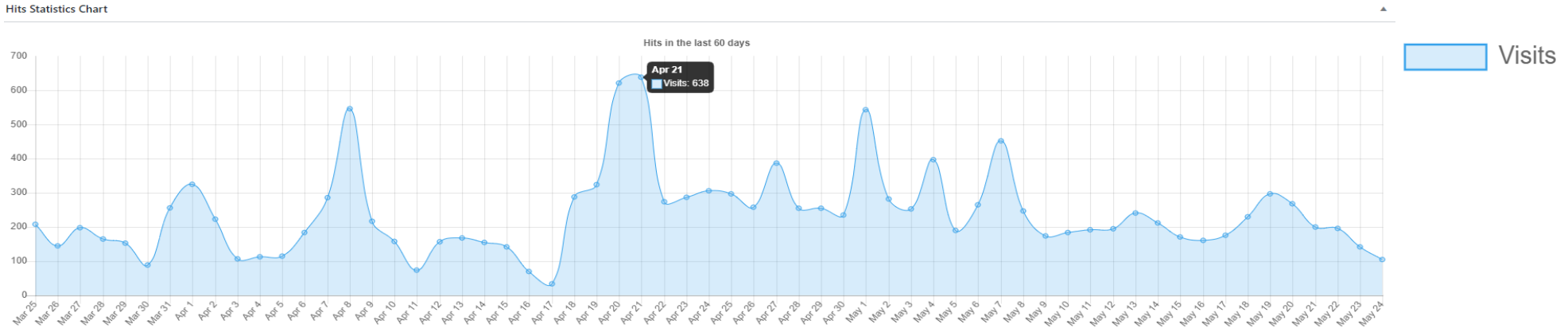
新冠数据、信息、知识门户（Gateway）

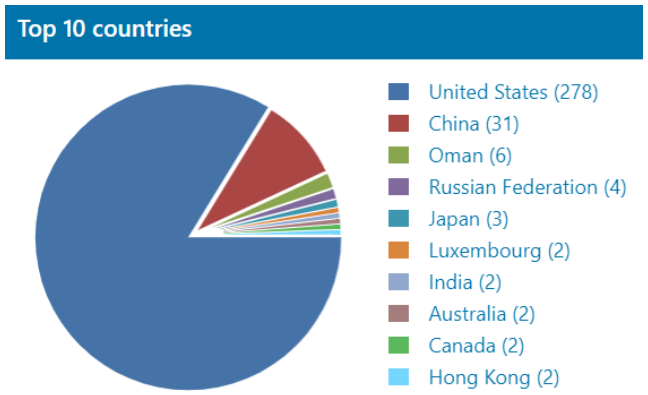
https://covid-19.stcenter.net/

COVID-19 Gateway

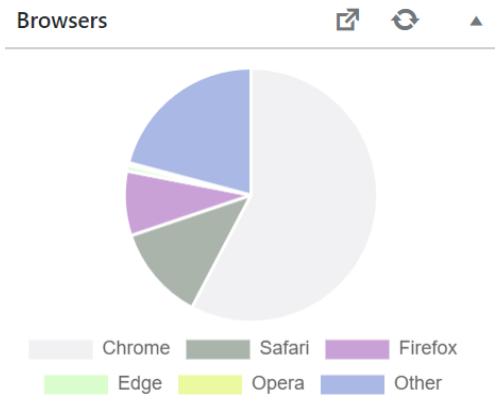
https://www.stcenter.net/

Index	value
Visitors in 30 days	596
Visit in 30 days	7492
Total visit from start	16766





Top 10 countries for visitors



Top browsers for visitors

Top Pages

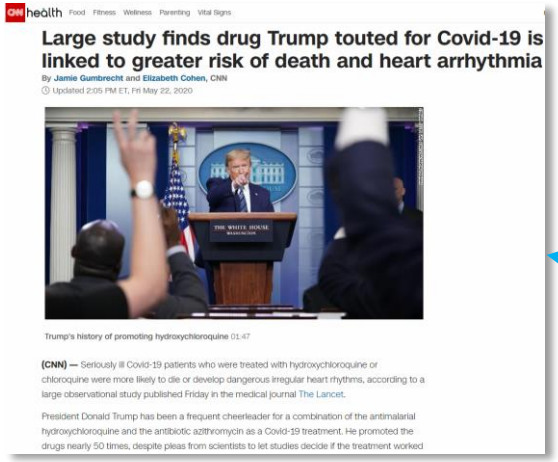
ID	Title	Link	Visits
1	Home Page	/	5,620
2	Data Collection	/index.php/data-access/	501
3	COVID-19 Live Map	/index.php/covid19-livemap/	371
4	Publication Analytics	/index.php/publication-analysis/	275
5	News Timeline	/index.php/chronicle/	234
6	Methods and Tools	/index.php/methods/	194
7	Home Page	/index.php	157
8	Task Forces	/index.php/task-forces/	105
9	Architecture	/index.php/architecture/	99
10	-	/index.php/login/	62

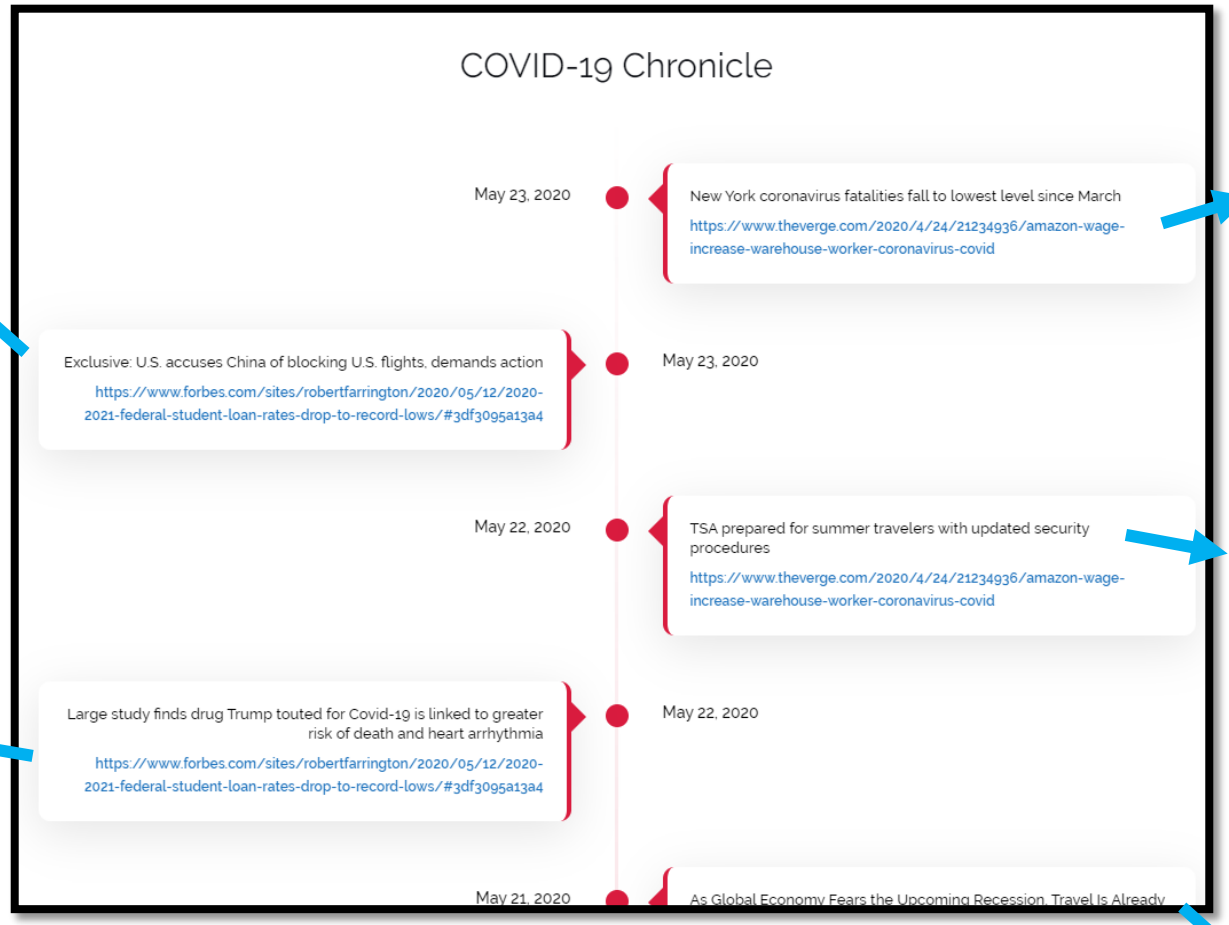
Top visit pages in last 30 days

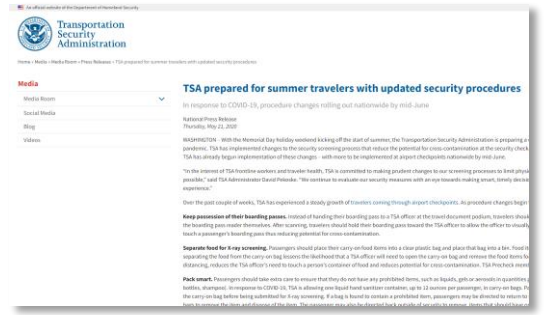
Outline：提纲

1. Introduction & Background: 研究背景
2. Computing Infrastructure: 计算平台
3. Data Collection & Sharing: 数据收集与共享
4. Spatiotemporal Pattern Analytics: 时空特性分析
5. Summary: 总结
6. Acknowledgements: 致谢

疫情大事记

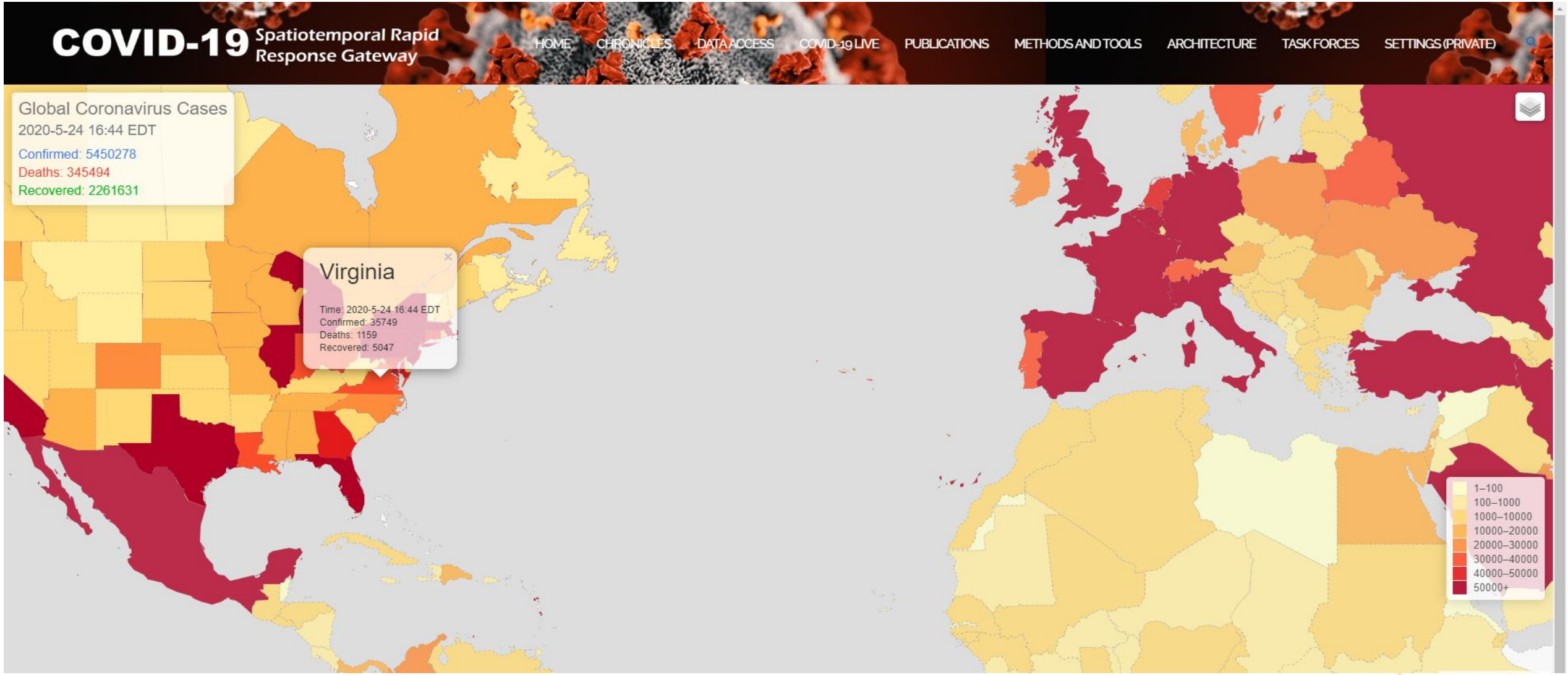






News Source:
CNN, CNBC, FOX News, New York Times, Los Angeles Times, Washington Post, The Hill, Reuters, Washington Business Journal, Bloomberg, etc.

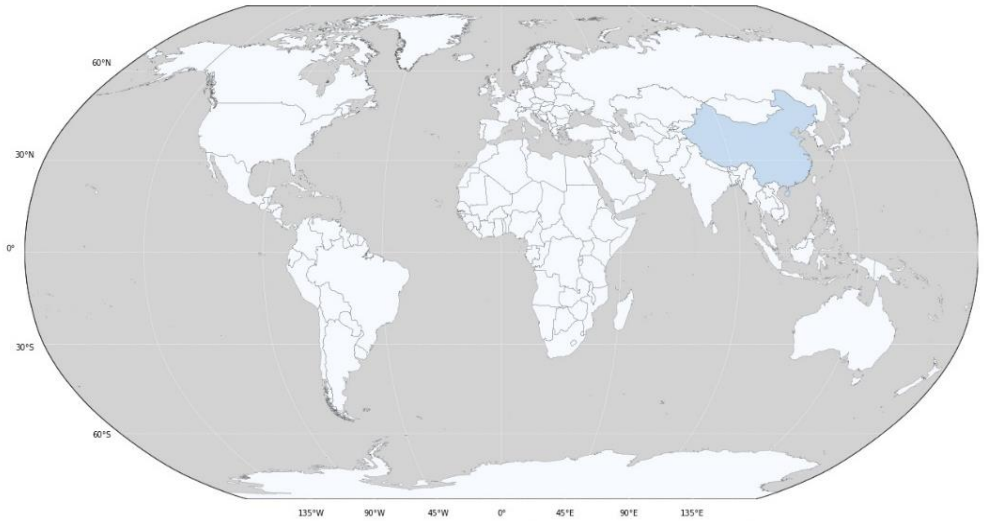
当前全球疫情状态



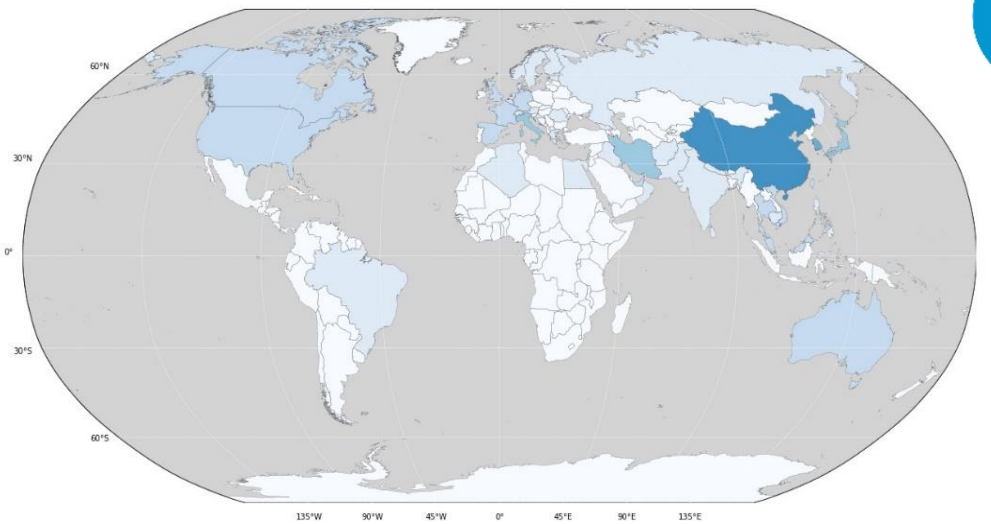
https://covid-19.stcenter.net/index.php/covid19-livemap/

病毒是如何在全球传播的？

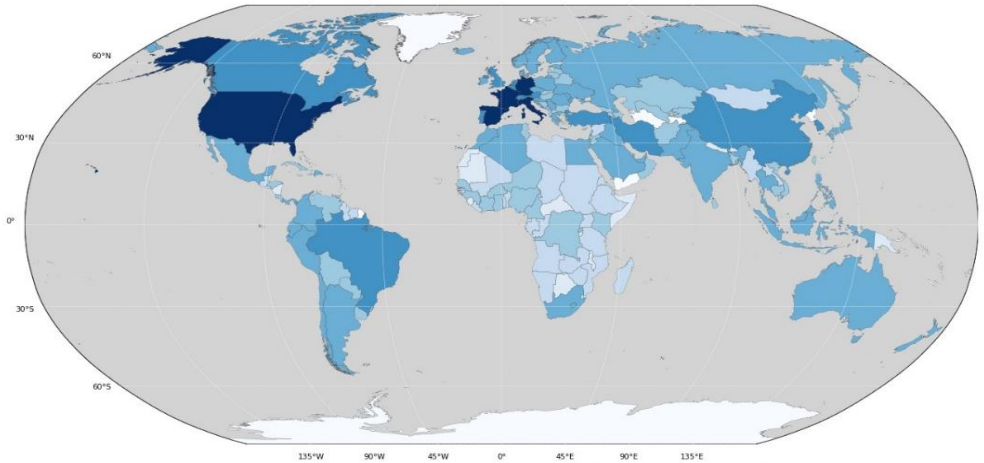
01/17/2020



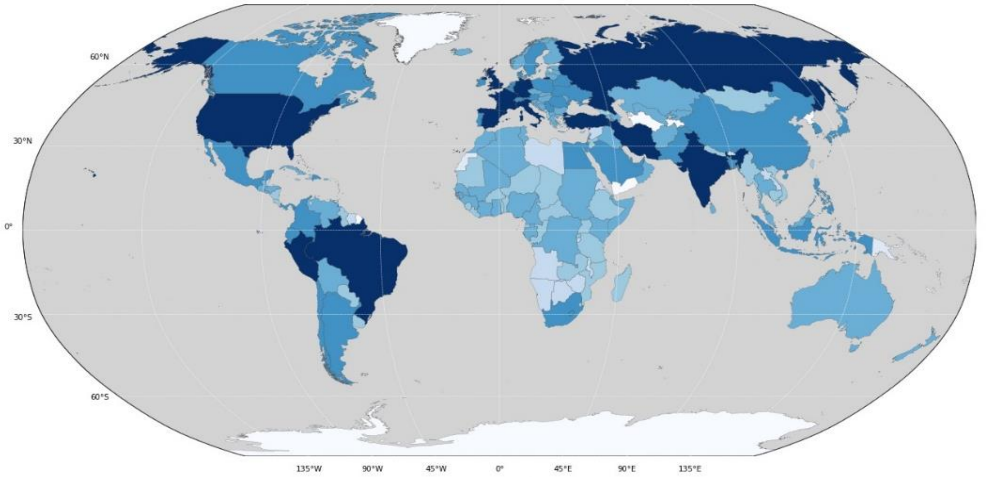
02/27/2020



04/07/2020



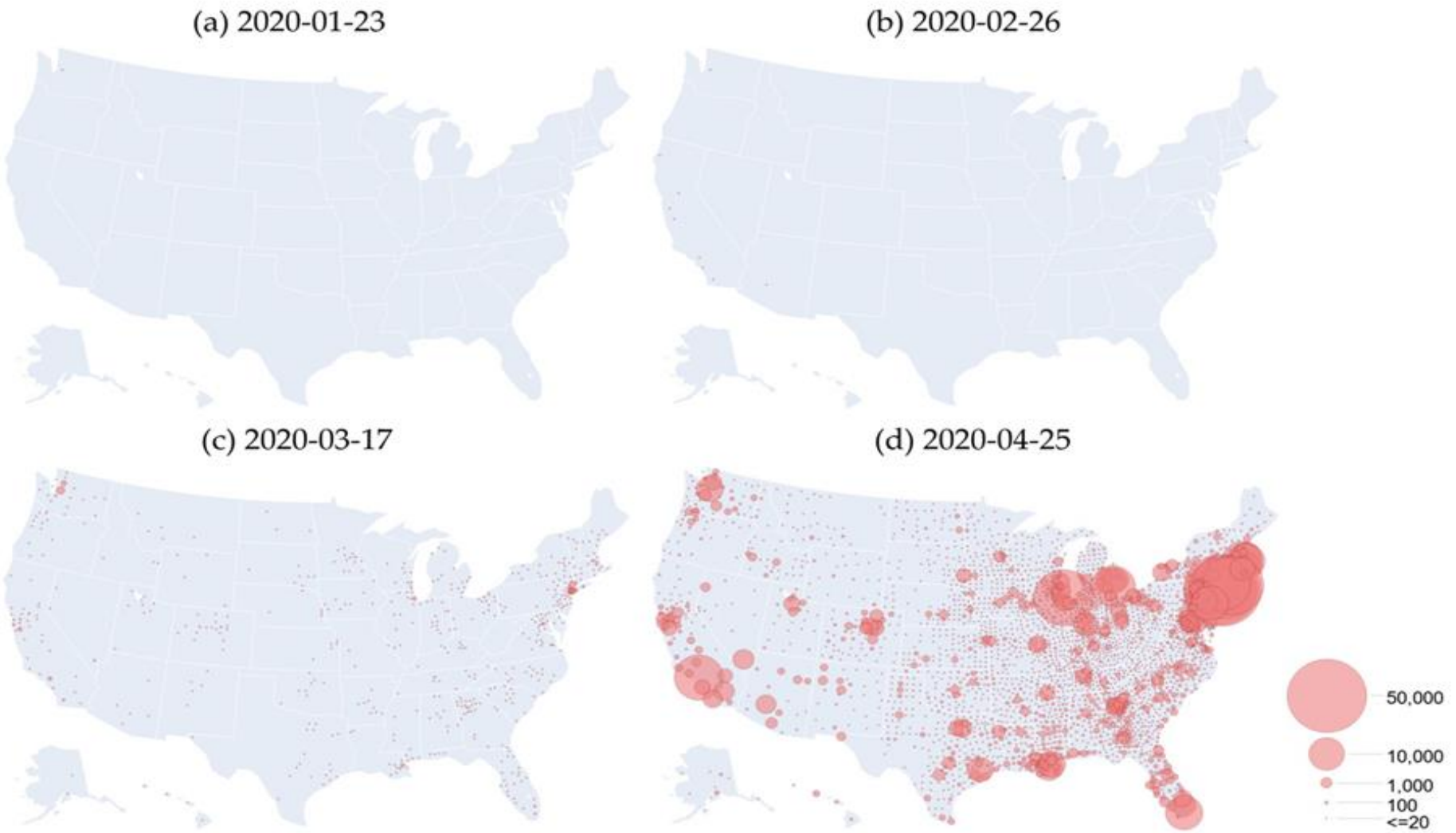
05/23/2020



0 1 10 100 1,000 10,000 100,000 2,000,000

Data Source: https://github.com/stccenter/COVID-19-Data/tree/master/Global

病毒是如何在美国传播的？



政府与组织对疫情的措施

许多国家都采取了**边境管制**和**封锁措施**，以控制COVID-19的扩散

- **政策严厉指数:**

  NSF 时空创新中心从全球资源收集，处理和可视化的数据

- **全球政策严厉指数变化：**

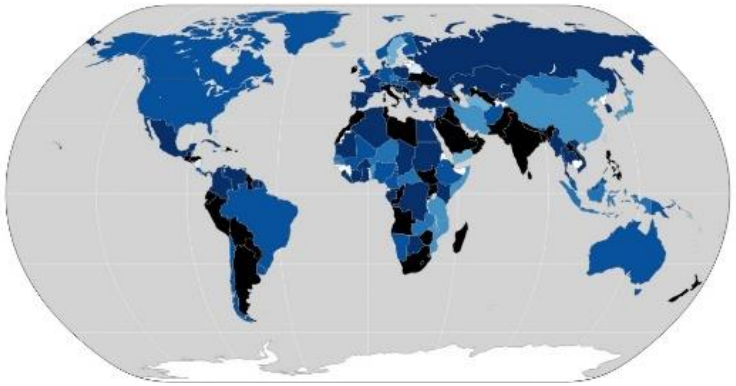
  一月至四月的全球政策严厉指数正在增加，并且与疫情传播密切相关

- **美国政策严厉指数变化：**

  美国政策严厉指数从三月初到四月下旬所有增长，并且与疫情传播密切相关

不同国家的政策严格程度

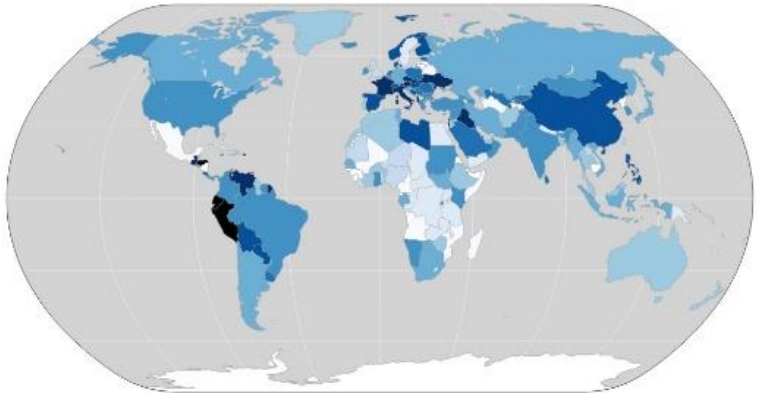
01/01/2020



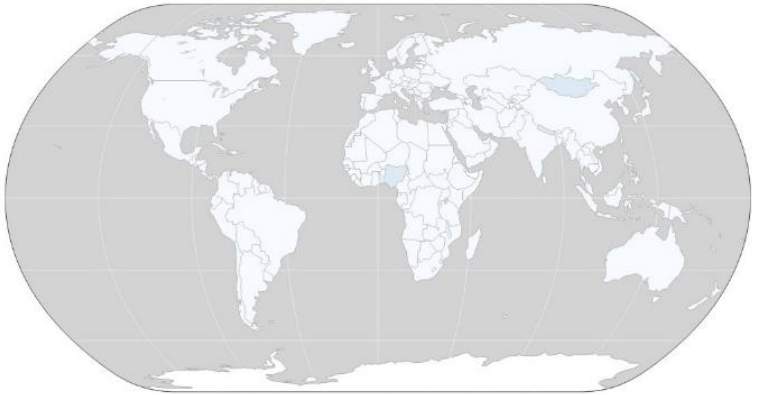
01/17/2020

02/17/2020

03/17/2020

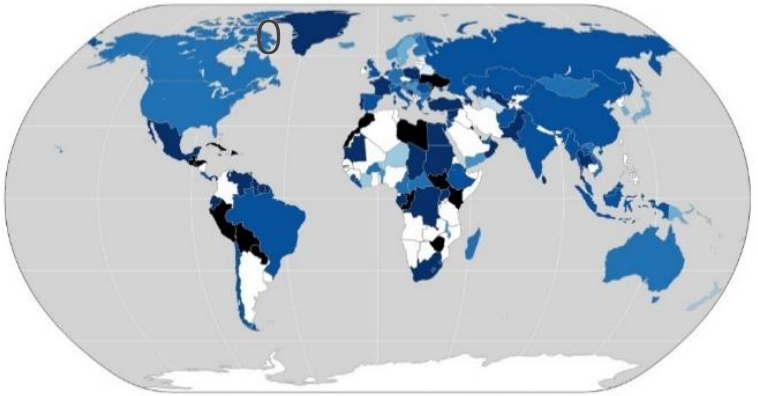


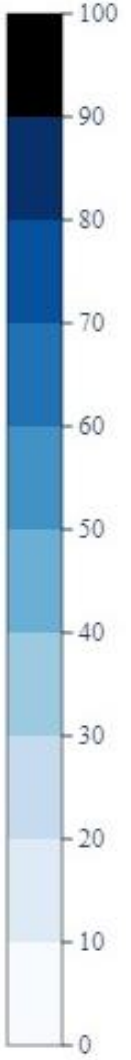
04/17/2020



05/17/202

0

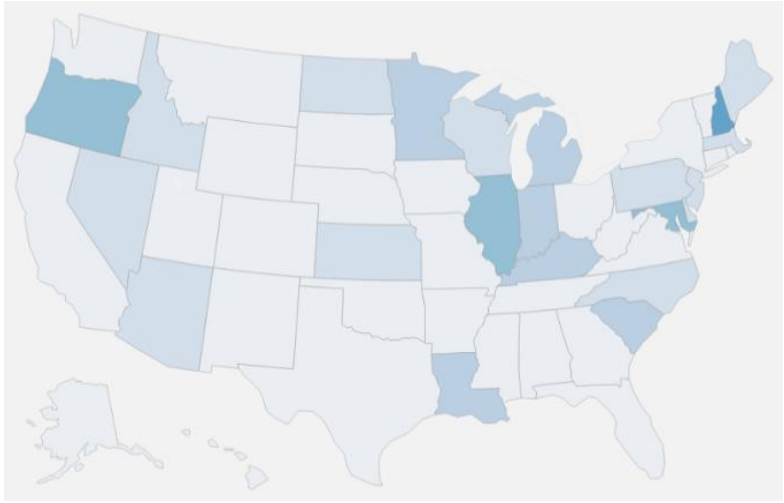




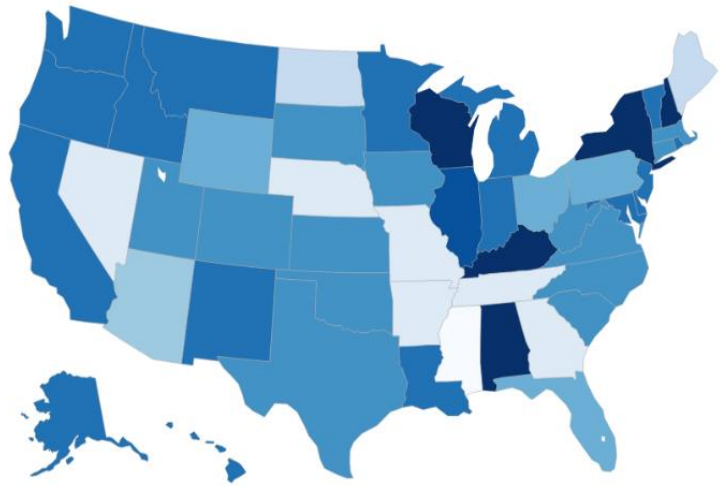
不同州的政策严格程度

03/05/2020

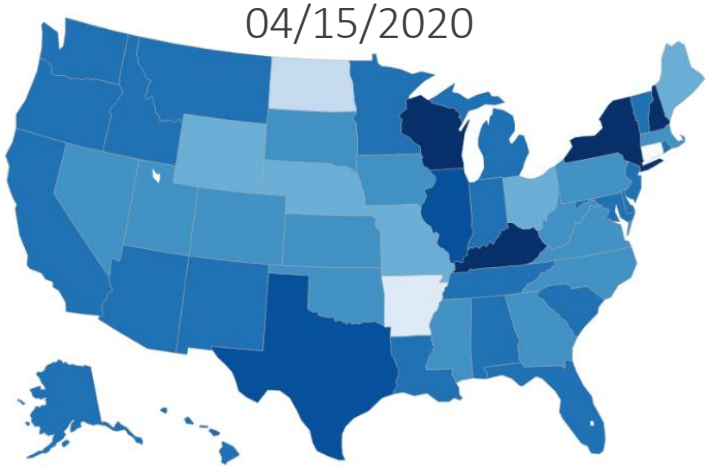
03/15/2020



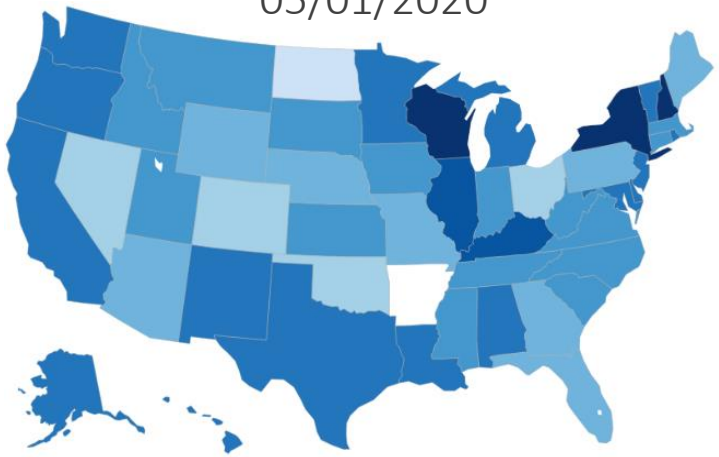
03/30/2020



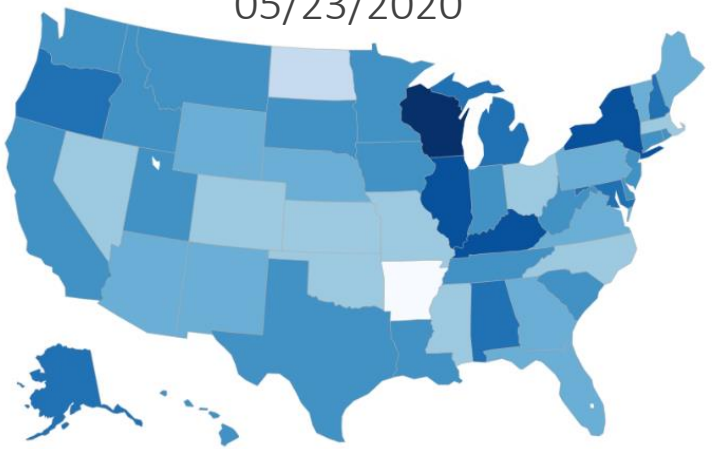
04/15/2020

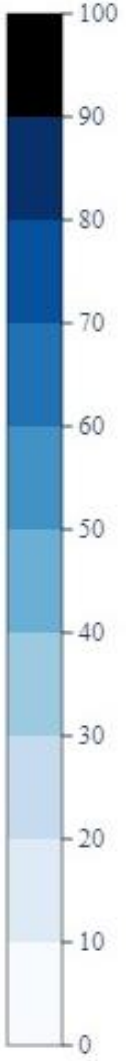


05/01/2020



05/23/2020





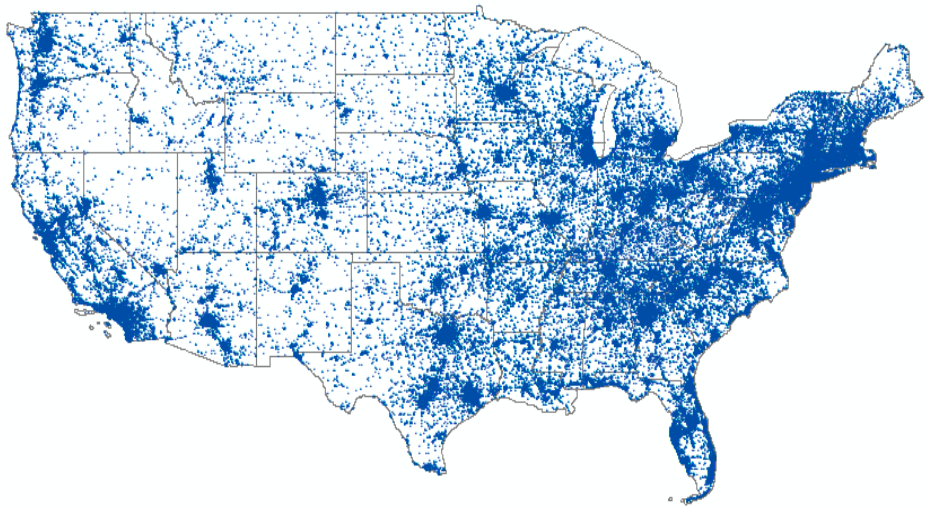
Data Source: https://github.com/stccenter/COVID-19-Data/tree/master/US

疫情如何影响人员流动？

人员流动影响疫情扩散

❑**社交媒体数据：**带地理坐标的推文

从Twitter收集的带GPS地理位置的推文



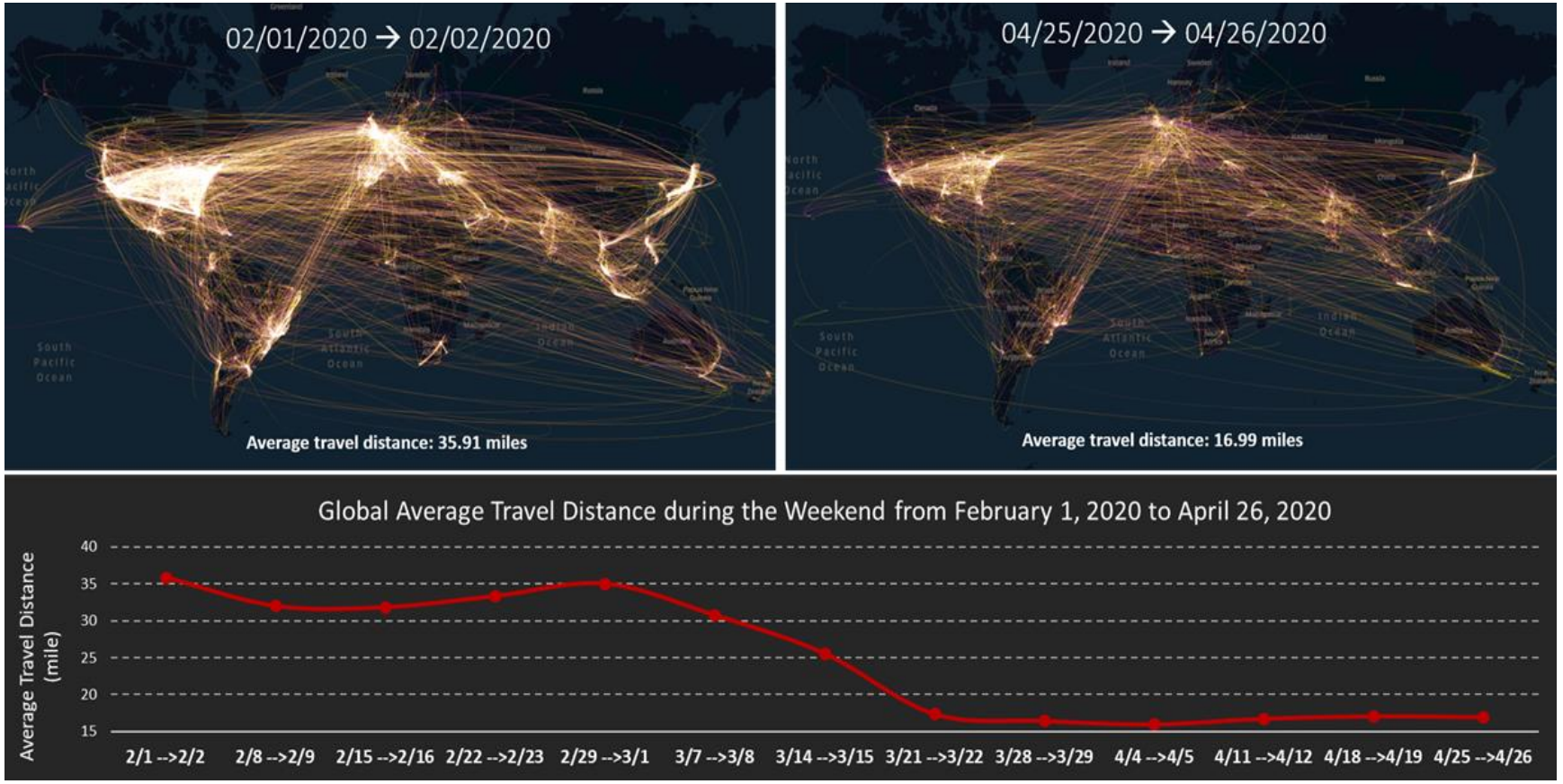
Geo-tweets in March

❑**手机数据:** 社区流动性报告

由Google收集和汇总，数据记录了不同类别地点（例如零售和娱乐场所，杂货店和药房，公园，公交车站，工作场所和住宅）的人员流动随时间变化的趋势

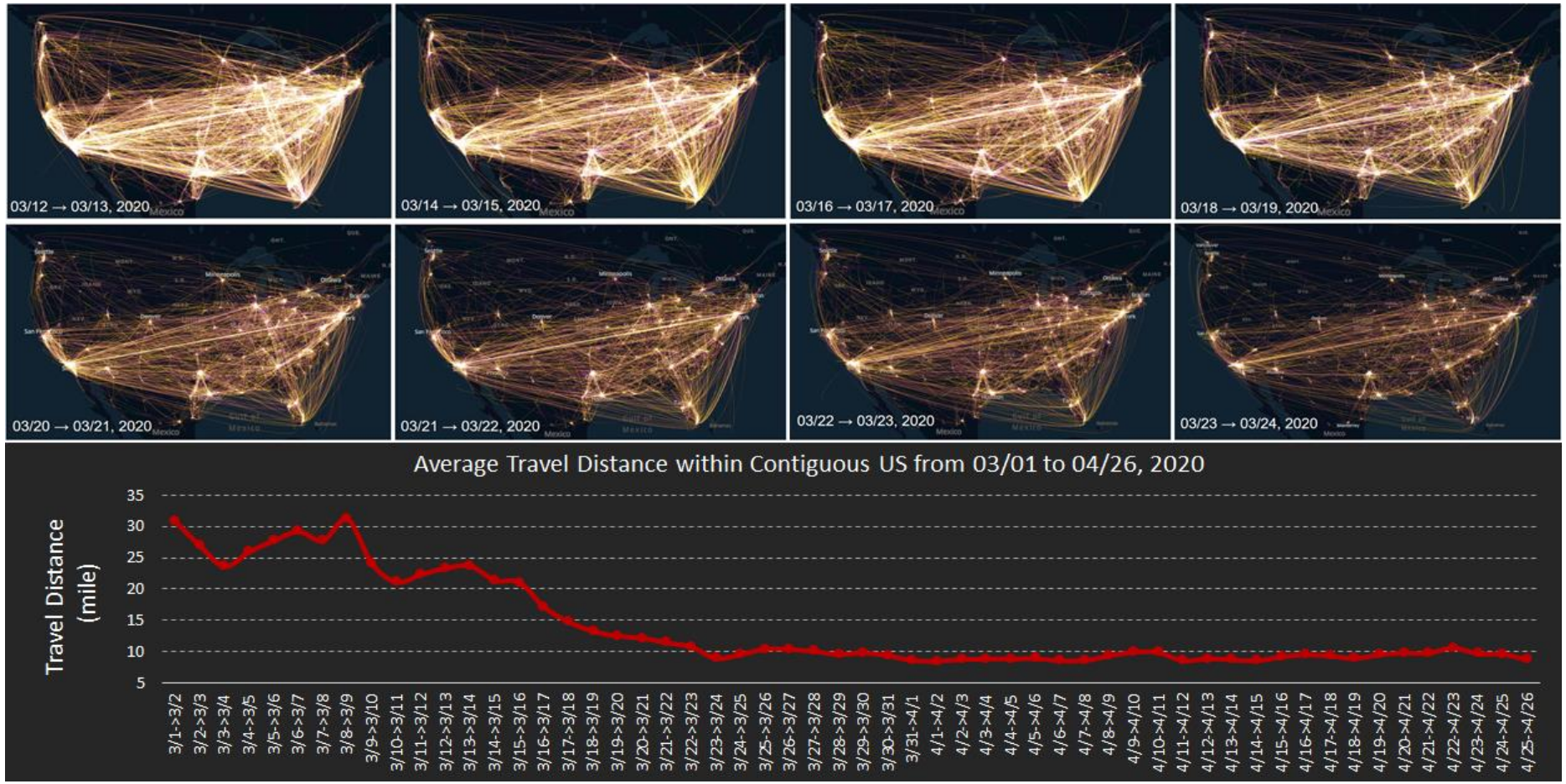
全球范围内的人员流动变化

- **Geotagged Tweets**: global average travel distance during the weekend from 02/01/2020 to 04/26/2020



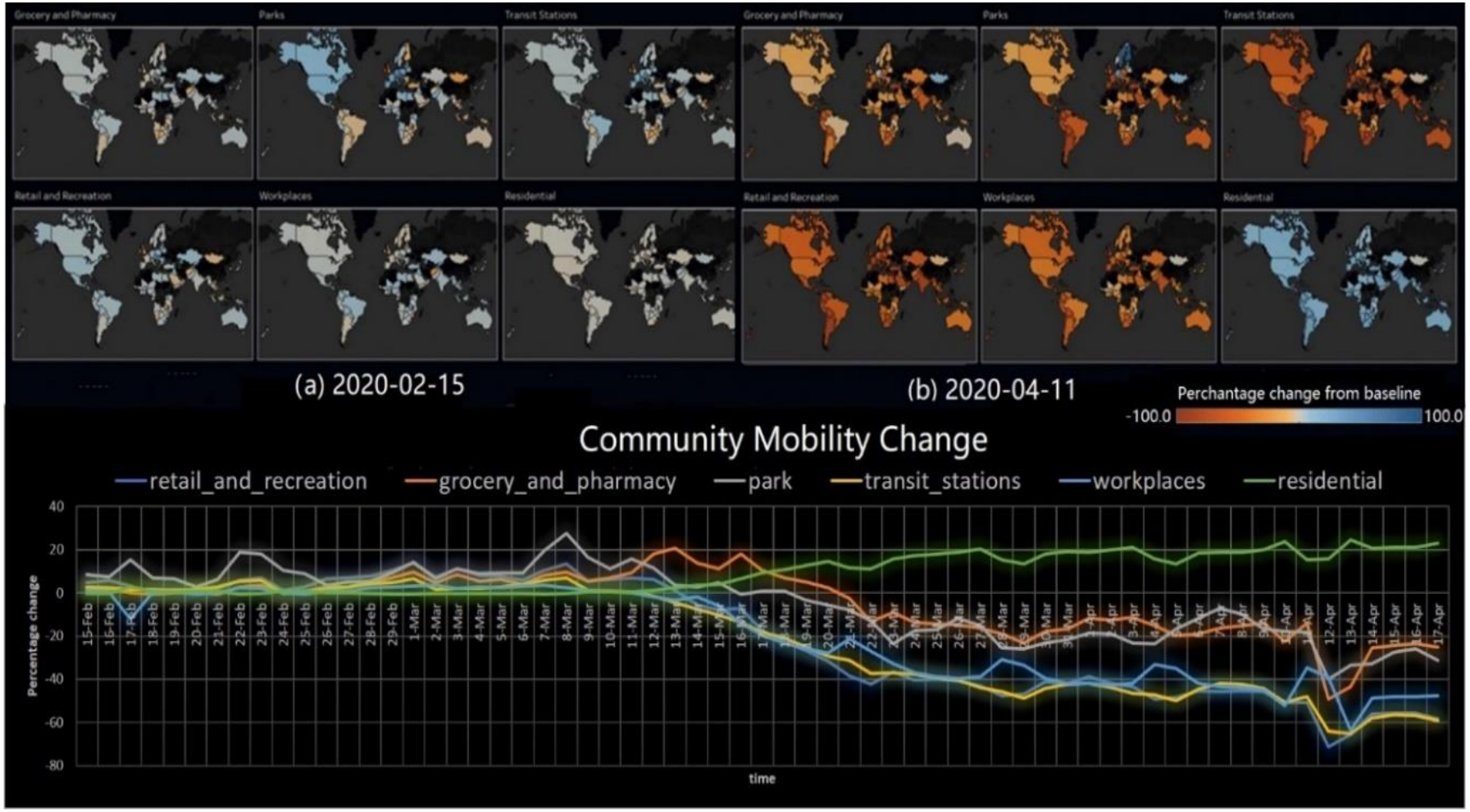
美国人员流动变化

- **Geotagged Tweets:** Population movements before and after the lockdown in the contiguous US.



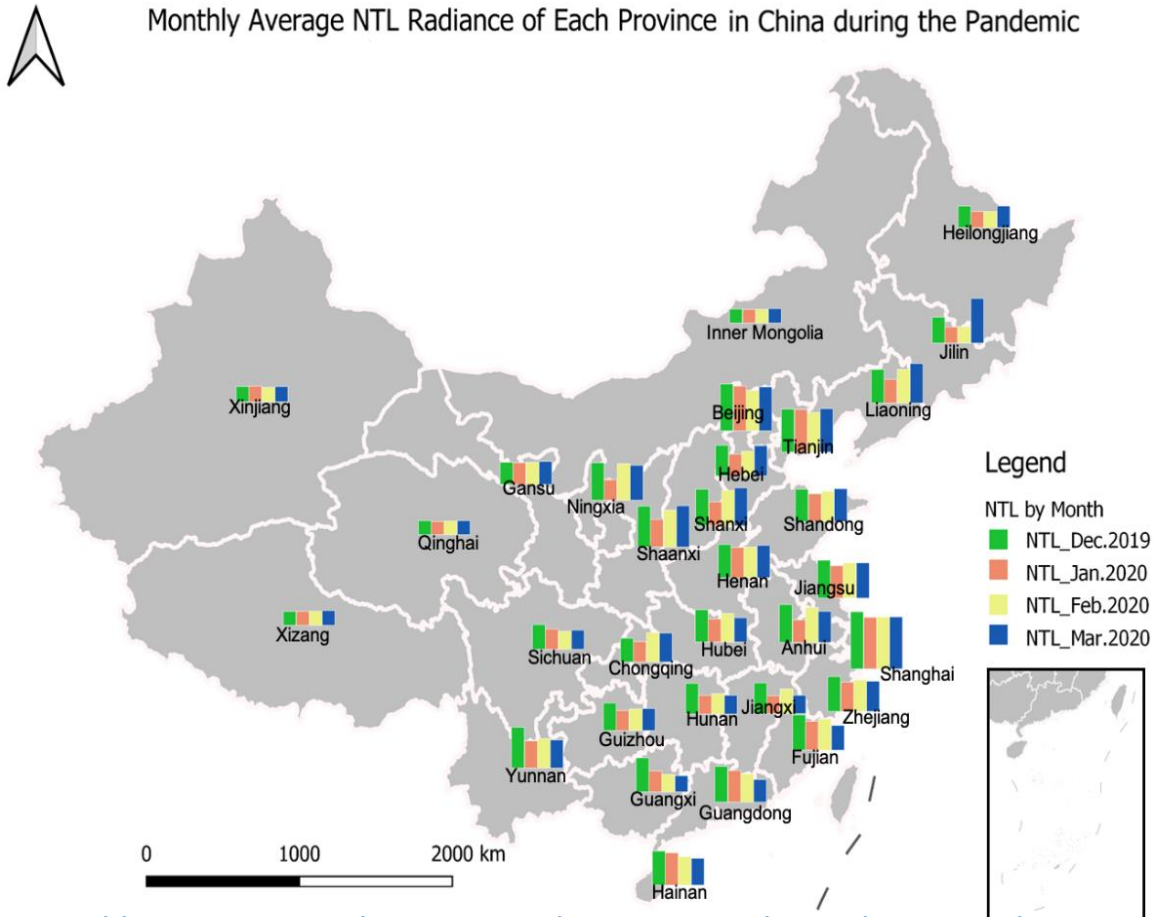
社区活动的变化

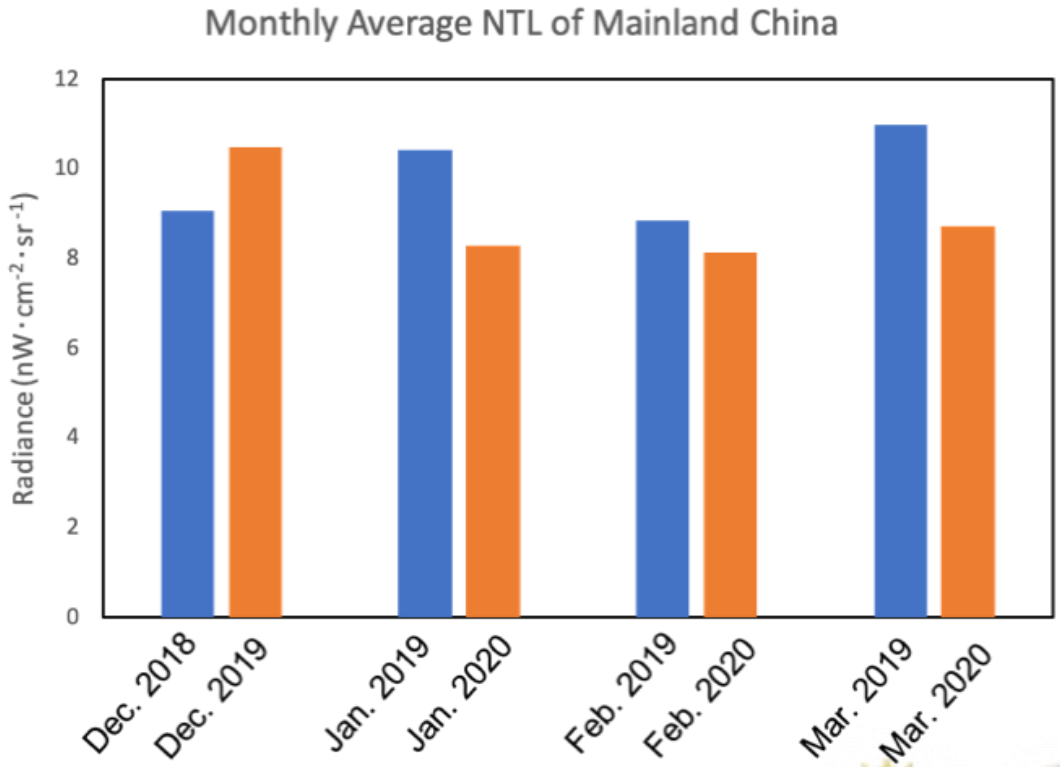
- **Mobile Phone data:** community movement change during the outbreak of COVID-19 across different categories of places



光污染的变化

1. 2019年12月至2020年2月，由于疫情蔓延和隔离政策，夜间光照辐射强度呈下降趋势，并且随着各个地方的重新开放，在2020年3月有小幅回升
2. 2020年初的平均夜间光照强度低于2019年同时期。

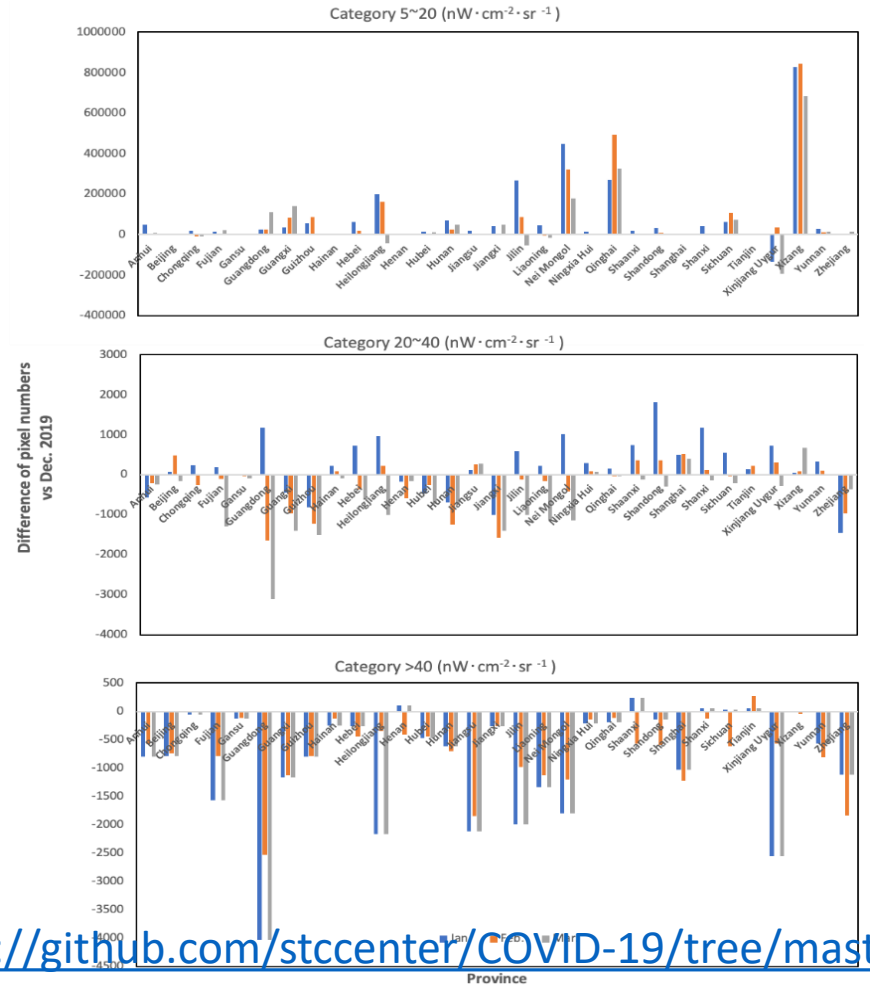


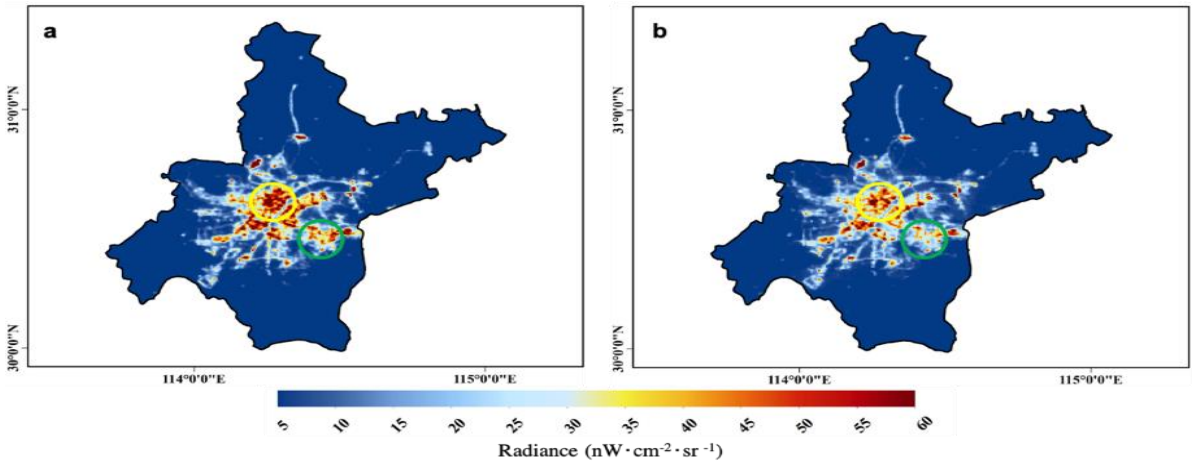


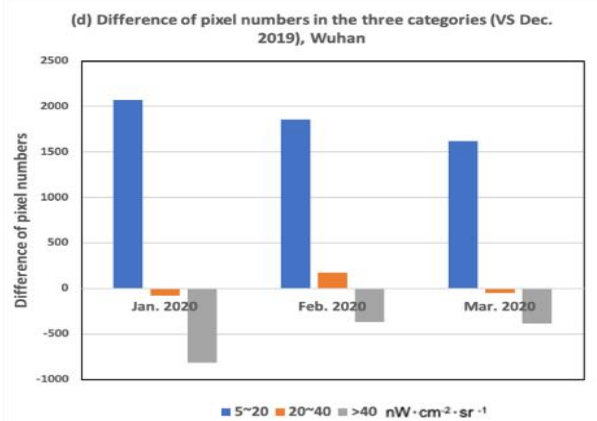
https://github.com/stccenter/COVID-19/tree/master/analysis/nightlight

光污染的变化

- 不同区域的夜间光照像素个数呈不同变化趋势：居民区像素个数在疫情期间与之前相比有所增加，商业中心像素个数有所减少；
- 武汉市夜间光照在新冠病毒暴发期间与之前相比强度较弱。





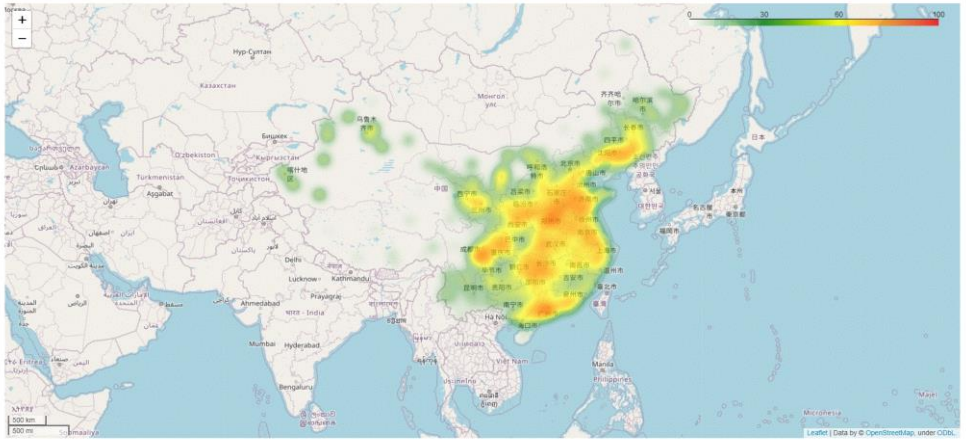


https://github.com/stccenter/COVID-19/tree/master/analysis/nightlight

空气质量的变化

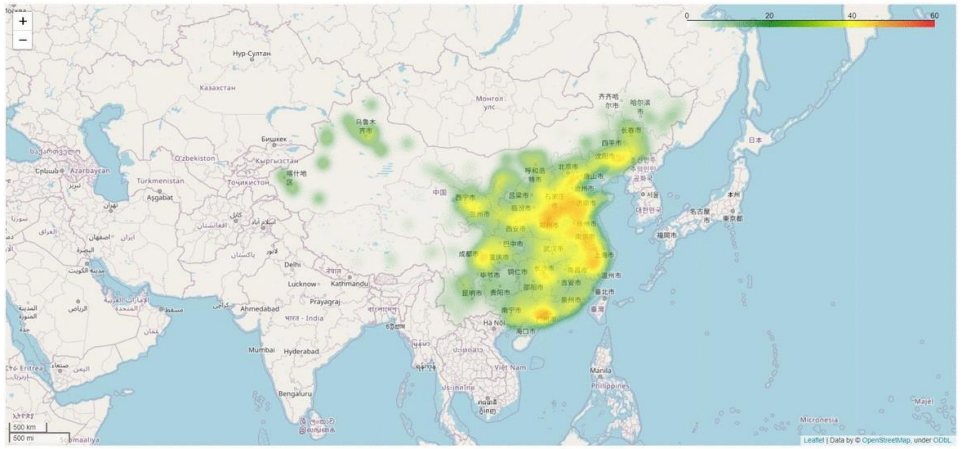
- 空气污染指数以及二氧化氮浓度在疫情期间与之前和之后相比均有所减弱；
- 日均空气污染指数在疫情期间呈下降趋势

Heatmap of the Air Quality Index on Jan. 01, 2020



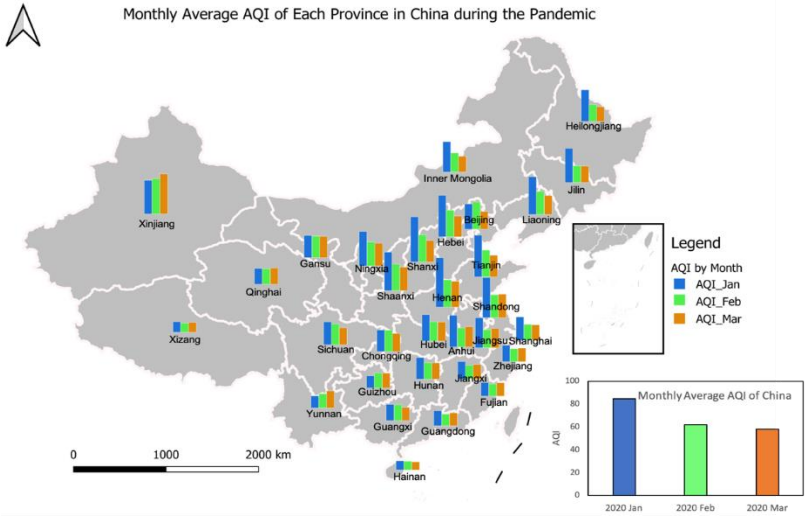
Data and visual analytics provided by Qian Liu and Zhiran Zhang, NSF Spatiotemporal Innovation Center.

Heatmap of the NO2 Emission on Jan. 01, 2020

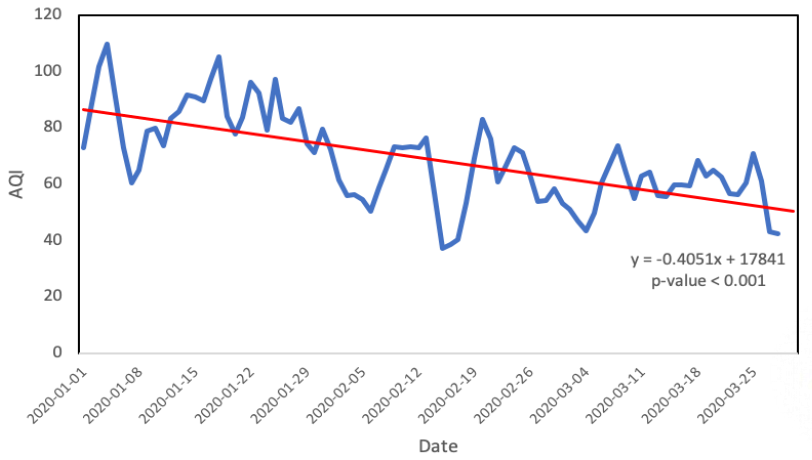


Data and visual analytics provided by Qian Liu and Zhiran Zhang, NSF Spatiotemporal Innovation Center.

Monthly Average AQI of Each Province in China during the Pandemic



China AQI trend before and during the COVID-19



https://github.com/stccenter/COVID-19/tree/master/analysis

新冠病毒研究的进展

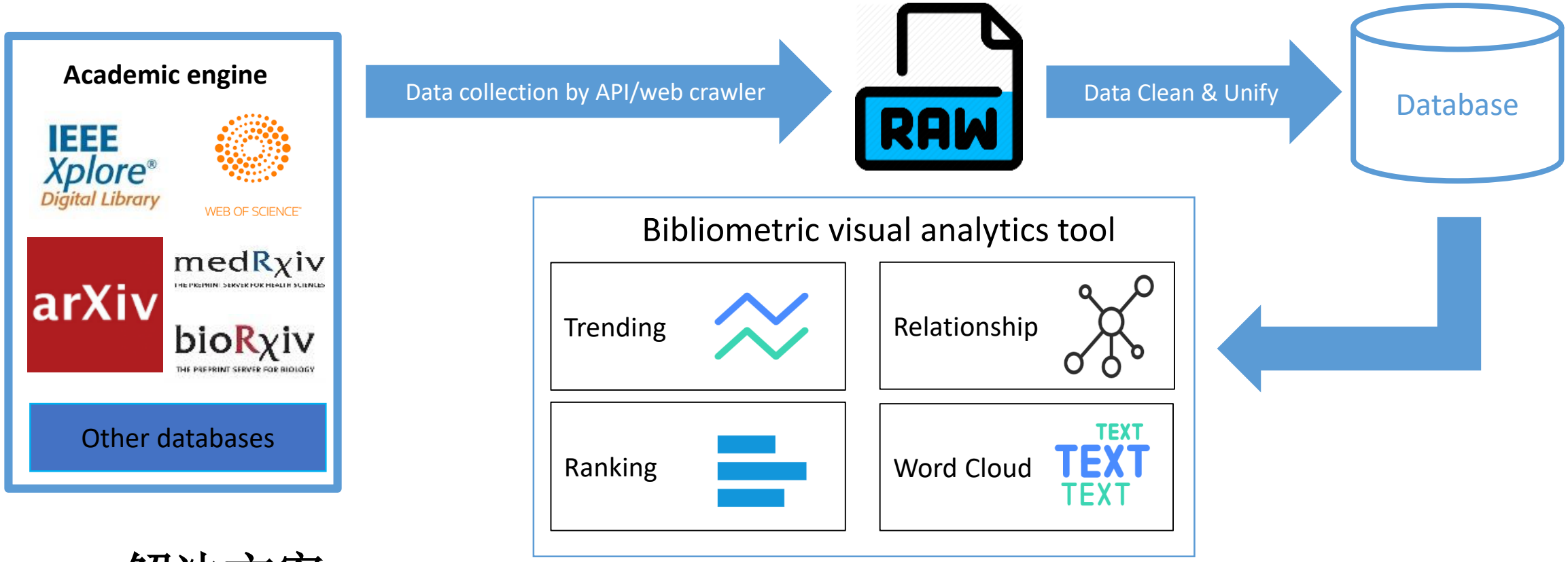
问题

- 大量COVID-19相关的文献
- 相关文献分散在各个学术数据库里
- 每个数据库的数据标准不同
- 整合同领域的所有文献相对困难
- 还未有很好的可视化工具来展示某个领域所有文献的趋势和之间的关系等

目标

- 从各个数据源收集文献数据
- 统一数据标准，存入数据库
- 开发可视化工具来展示文献之间的内在逻辑结构

新冠病毒研究知识库



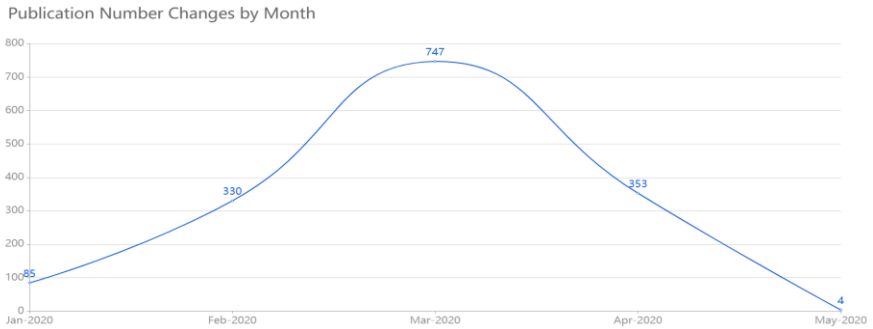
解决方案

- 从各个学术数据库收集并整合文献数据
- 清理数据，并对数据进行统一化
- 储存进数据库
- 提供文献分析模块来可视化文献的变化与关联
- 用更直观的图表展示文献的核心信息

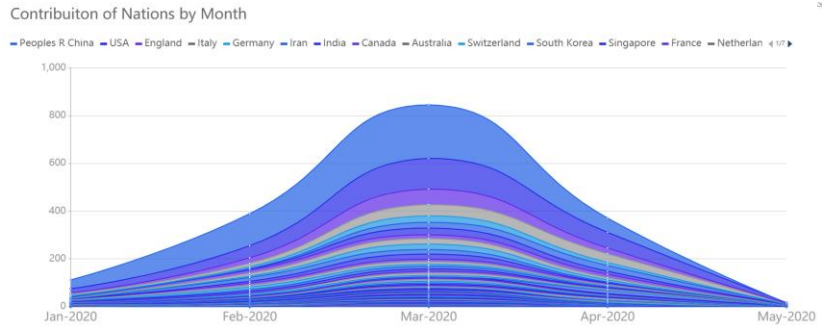
新冠病毒研究的现状

https://www.stcenter.net/

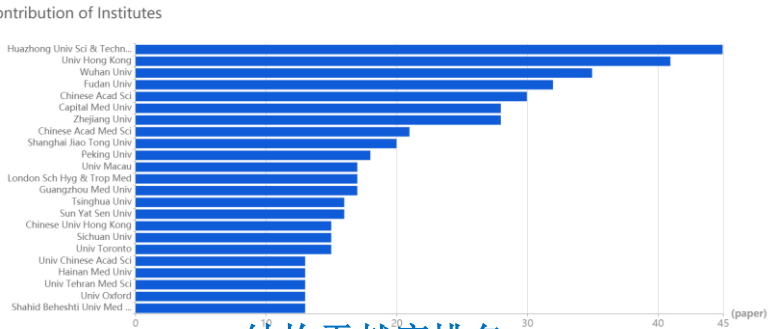
- Search
- Trend
  - Publication
  - Nation
  - Institute
- Relationship
  - Nations
  - Institutes
  - Authors
- Ranking
  - Keyword
  - Institute
- References



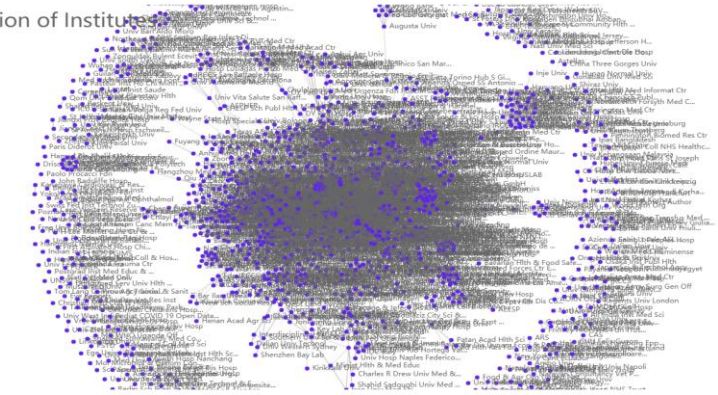
文章发布数量趋势



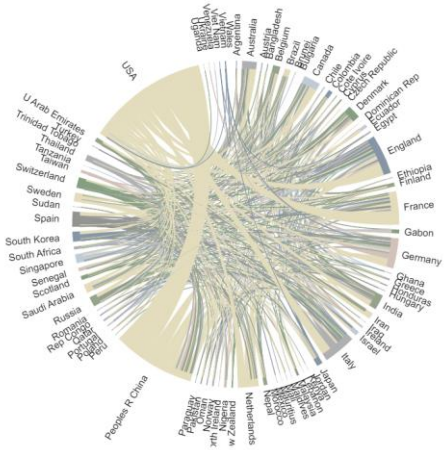
各国文章数趋势



结构贡献度排名



研究者与机构的合作情况



国家间的合作情况

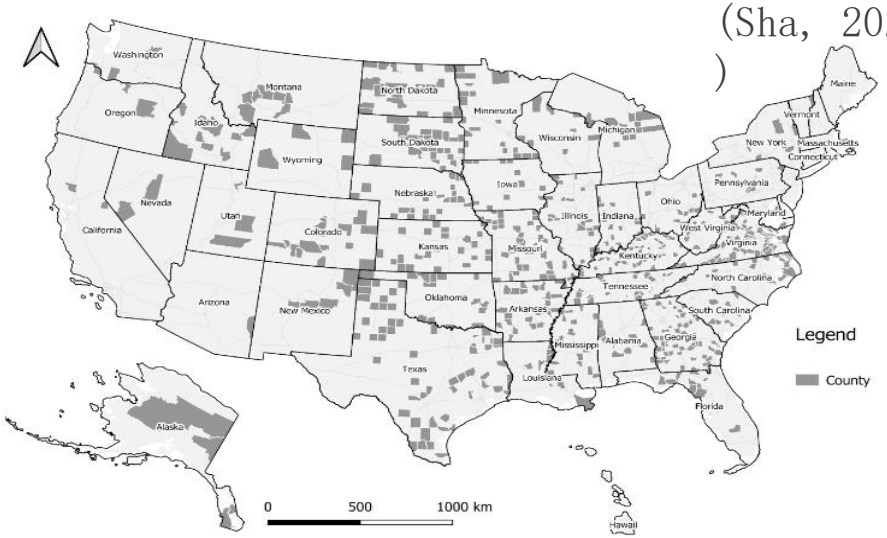
关键词词云分析

https://covid-19.stcenter.net/index.php/publication-analysis/

美国的公卫系统承载力如何？

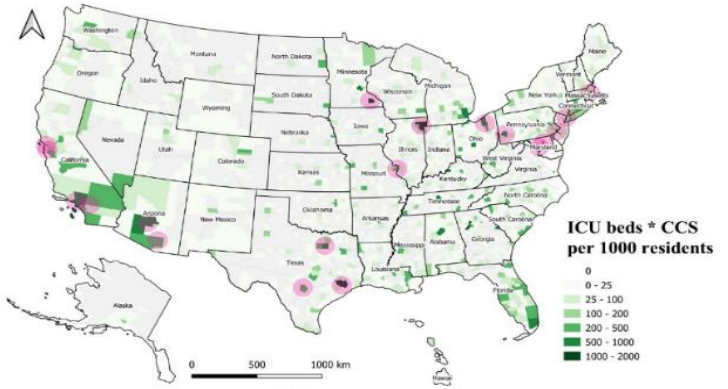
通过对基本医疗软硬件病床和医护人员的数据整理，我们发现医疗资源在空间上分布不均衡，集中在重要的医疗中心，大多乡村地区的医疗资源不足。

图右上：可以看到有671个县是没有病床或者对于新冠肺炎治疗的医疗专家的。

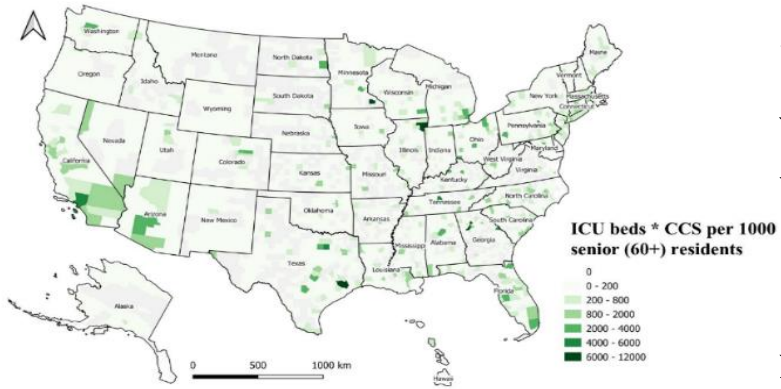


(Sha, 2020.)

The 671 counties without licensed beds or CCS.

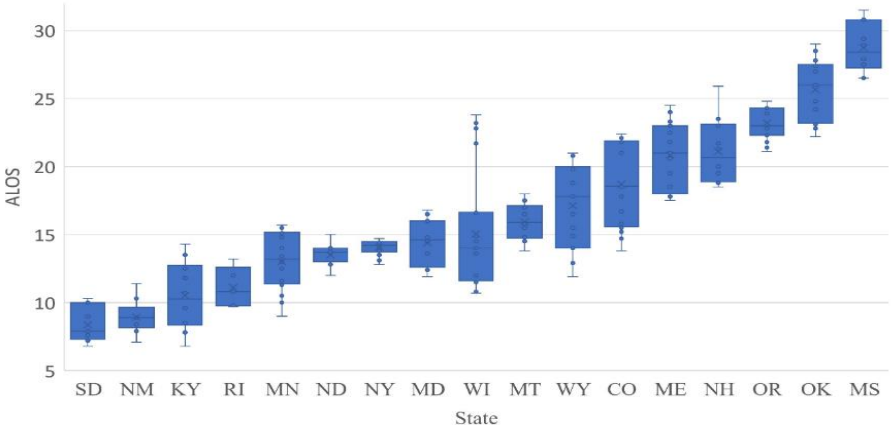


(a)



(b)

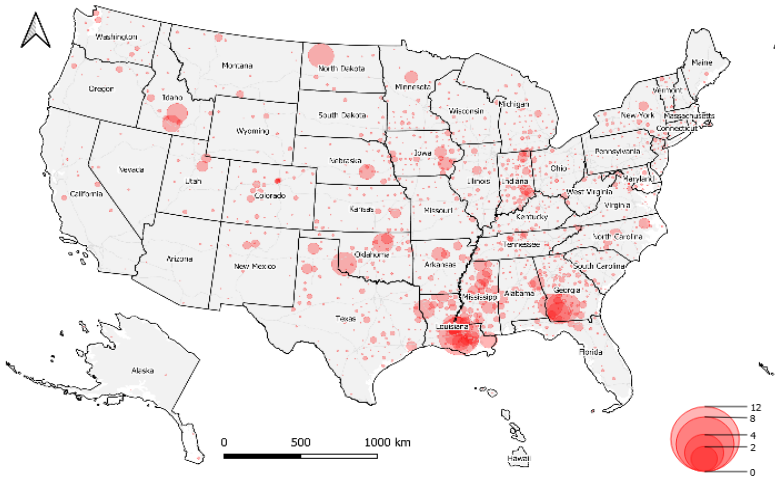
各州的治疗水平也明显不同，根据19个有数据披露的州，发现住院的病人平均在医院的时间长度也不同，纽约州平均14天出院，而经济水平较差的密西西比州需要住院28天。



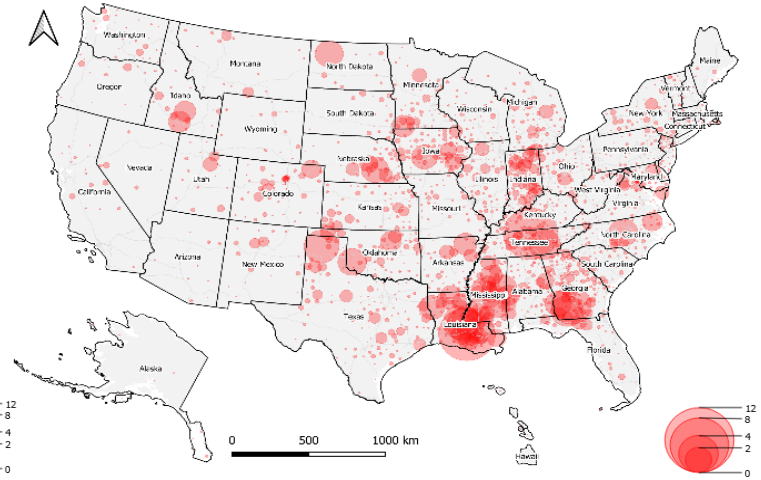
Box-plot of hospitalized ALOS among 19 states

美国的公卫系统承载力如何？

(a) March 15, 2020



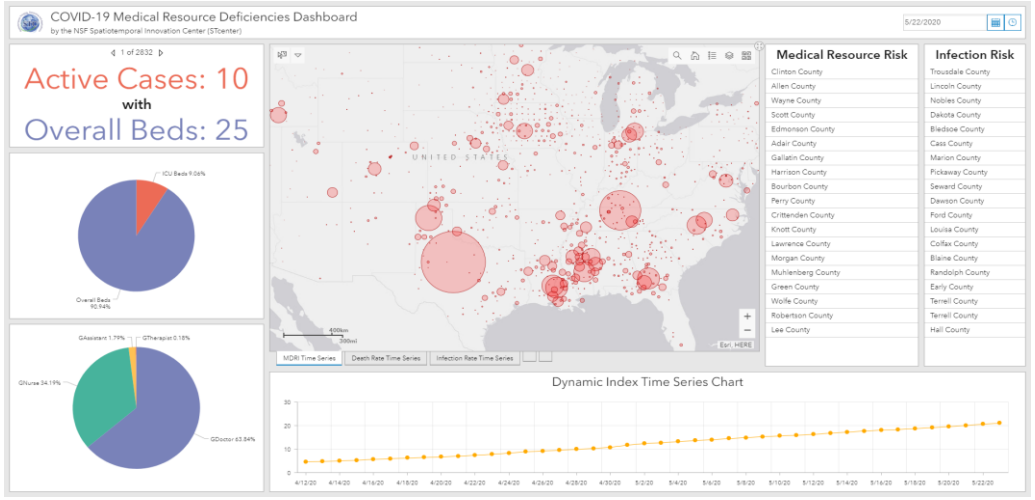
(b) April 15, 2020



(c) May 1, 2020

General MRDI trend.

$$MRDI = \frac{N_c - N_{death}}{N_{licbed} \cdot N_{CCS}}$$



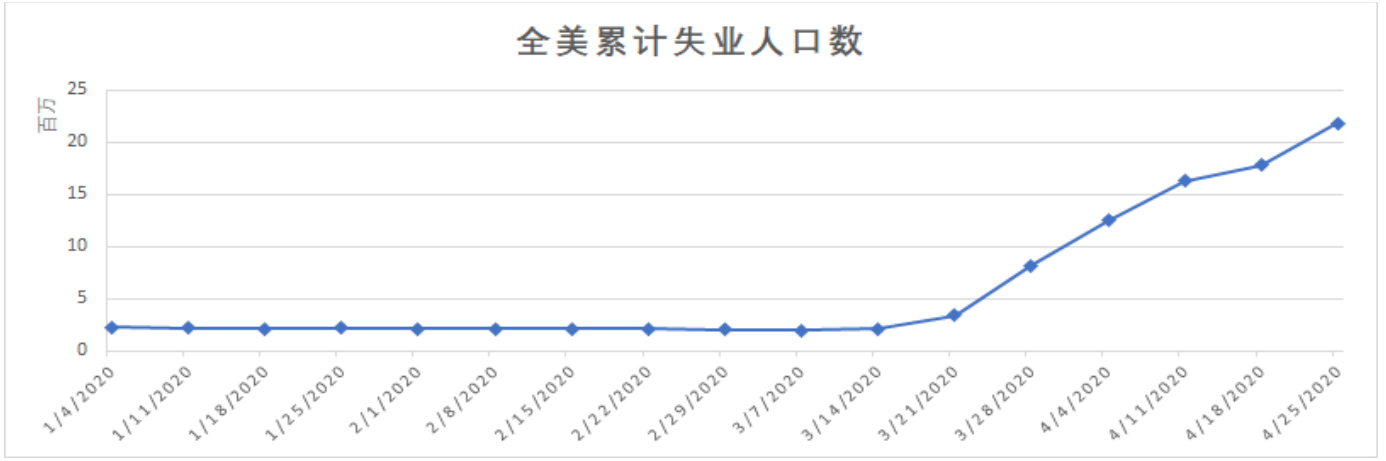
http://mrd-dashboard.stcenter.net/

Spatiotemporal visualization interface based on ArcGIS Dashboards.

引入动态的每日确诊死亡人数，用医疗资源数相除，得到医疗资源动态承载力指数。可见医疗资源不发达地区的严重程度强于纽约加州等疫情大爆发的区域。

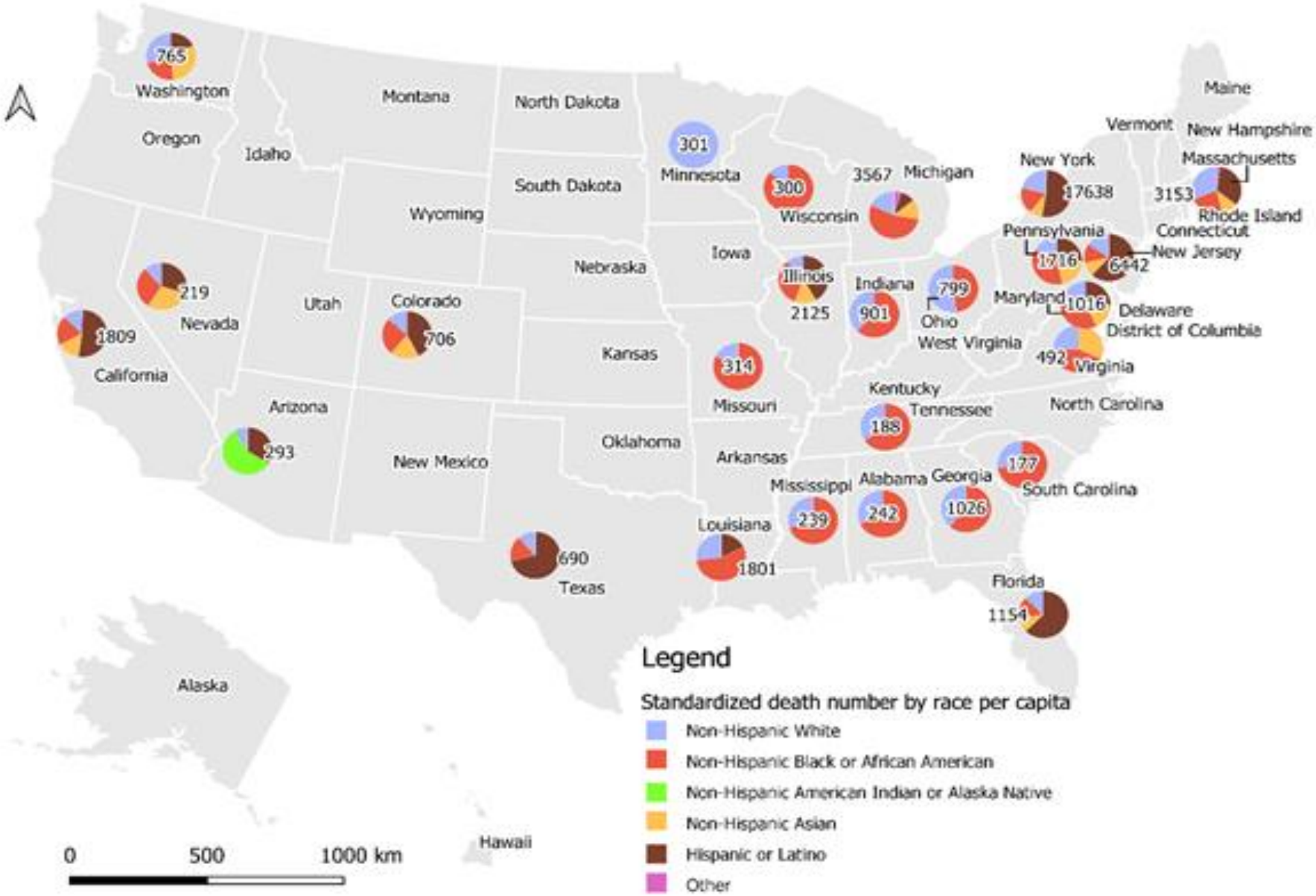
疫情对经济有什么影响？

- 全球股票市场呈现的时空模式与病毒传播高度相关



- 美国累积失业人口变化

新冠病毒是否针对特定的人群?



各州每个种族死亡数与人口数的比值，可见不同族裔受到影响显著不同。

Sha et al., 2020

新冠病毒是否针对特定的人群?

Variable	Coefficient	Std. Error	Probability
Elderly Rate	-12.6535	5.29928	0.02364
Educational Attainment	-9.96245	5.2992	0.05011
Poverty Rate	17.0477	5.99688	0.00447
Income	0.3127	0.137774	0.02320
Spatial Lag	0.591188	0.0487105	0.00000
Constant	4.13026	5.59176	0.00233
R-squared	0.577802		

以麻州的城市为例，利用空间滞后回归模型，分析社会经济指标与患病率的关系。结果表明：老龄率和接受高等教育的比例与患病率存在负相关性，贫困率与患病率存在较强正相关性。

Outline：提纲

1. Introduction & Background: 研究背景
2. Computing Infrastructure: 计算平台
3. Data Collection & Sharing: 数据收集与共享
4. Spatiotemporal Pattern Analytics: 时空特性分析
5. Summary: 总结
6. Acknowledgements: 致谢

总结：时空特性

1. 疾病传播的方式。
2. 社会经济差距和脆弱性。
3. 流动性与社会距离。
4. ICU和呼吸机等医疗保健设施的可用性。
5. 从暴发城市和地区的相似气候带（从武汉、伊朗、新加坡、巴西、印度、俄罗斯、意大利到西雅图和纽约）来看，气候与疫情爆发有一定联系、但并不密切。
6. 经济问题，例如消费者需求变化、供应链中断、（失业）就业模式。
7. 政策、新闻或社会情绪变化与疫情爆发严重程度之间的关系是病毒长期控制和约束以及人类社会设法与病毒共存的关键。
8. 平衡开放数据和隐私保护策略。

总结：反思与观点

1. 应对政策和行政措施
2. 地球上的生命
3. 自然和人力改造的平衡
4. 政治、人道和对全球紧急情况的快速反应
5. 开放的世界和地球村

下一步研究计划

- 联合研究、数据共享、有针对性提供服务
- 继续跟踪、分析和识别问题
- 扩展到西班牙裔和非裔美国人等社区，提高他们的防范能力
- 调查大流行后的应变能力
- 重新思考时空研究如何帮助改善对全国乃至全球紧急情况的反应

Outline：提纲

1. Introduction & Background: 研究背景
2. Computing Infrastructure: 计算平台
3. Data Collection & Sharing: 数据收集与共享
4. Spatiotemporal Pattern Analytics: 时空特性分析
5. Summary: 总结
6. Acknowledgements: 致谢

致谢/Acknowledgements

- 与哈佛大学CGA Dr. Wendy Guan， Dr. Tao Hu 紧密合作
- NSF I / UCRC项目提供了中心和快速响应帮助
- 亚马逊提供了计算信用额度
- 各个工作小组的大力支持
- Harvard CVT（冠状病毒可视化小组）

致谢/Acknowledgements

- Team Members: Dexuan Sha, Yun Li, Qian Liu, Zifu Wang, Hai Lan, Yifie Tian, Joy Zhou, Fayez Beani, Kyla Carte, Zoe Xu, Shiyang Ruan, etc.
- Our current and past members provides foundational support, NASA Goddard, NCCS, USGS, NASG, NGCC, Harris, Northrop Grumman, Microsoft, USDA, NOAA, UN, State Dept., Eastview Geospatial, OminiSci, RMDS Inc., CDI and the institutional support from GMU, Harvard, UCSB.
- We give our special thanks to our NSF program project directors, Rita Rodriguez, Dmitri Perkins, Behrooz Shirazi, our evaluators Donald Price & David Meyer, IAB chairs Lynn Usery (past chair) and Myra Bambacus, and Daniel Duffy
- All our partners

参考文献/References

1. Yang, C., Sha, D., Liu, Q., Li, Y., Lan, H., Guan, W.W., Hu, T., Li, Z., Zhang, Z., Thompson, J.H. and Wang, Z., etc. 2020. Taking the pulse of COVID-19: A spatiotemporal perspective. arXiv preprint arXiv:2005.04224.
2. Liu, Q., Sha, D., Liu, W., Houser, P., Zhang, L., Hou, R., Lan, H., Flynn, C., Lu, M., Hu, T. and Yang, C., 2020. Spatiotemporal Patterns of COVID-19 Impact on Human Activities and Environment in Mainland China Using Nighttime Light and Air Quality Data. Remote Sensing, 12(10), p.1576.
3. Sha, D., Miao, X., Lan, H., Stewart, K., Ruan, S., Tian, Y., Tian, Y., Yang, C., 2020. Spatiotemporal Analysis of Medical Resource Deficiencies in the U.S. under COVID-19 Pandemic. DOI: 10.13140/RG.2.2.20238.38721
4. Hu, T., Guan, W.W., Zhu, X., Shao, Y., Liu, L., Du, J., Liu, H., Zhou, H., Wang, J., She, B. and Zhang, L., 2020. Building an Open Resources Repository for COVID-19 Research.
5. Yang, C.; Clarke, K.; Shekhar, S.; Tao, C.V. Big Spatiotemporal Data Analytics: a research and innovation frontier. Taylor & Francis: 2019.
6. Yang, C.; Huang, Q.; Li, Z.; Liu, K.; Hu, F. Big Data and cloud computing: innovation opportunities and challenges. International Journal of Digital Earth 2017, 10, 13-53.
7. Yu, M.; Bambacus, M.; Cervone, G.; Clarke, K.; Duffy, D.; Huang, Q.; Li, J.; Li, W.; Li, Z.; Liu, Q. Spatiotemporal event detection: a review. International Journal of Digital Earth 2020, 1-27.
8. Yang, C.; Wu, H.; Huang, Q.; Li, Z.; Li, J. Using spatial principles to optimize distributed computing for enabling the physical science discoveries. Proceedings of the National Academy of Sciences 2011, 108, 5498-5503.

Thank you: 谢谢

Question & Comments?

Points of contact:

Phil Yang, cyang3@gmu.edu

https://www.stcenter.net/

加入疫情研究群与注册系列讲座

加入疫情研究群

在线系列讲座注册