China Data Institute

Center for Geographic Analysis
Harvard University

NSF Spatiotemporal Innovation Center

Geo-computation Center for Social Sciences
Wuhan University

Future Data Lab

Introduction to Spatial Data Lab

- Building a Cyberinfrastructure for Spatial and Global Studies

Dr. Wendy Guan, Harvard University
Dr. Shuming Bao, Future Data Lab

Topics

- Background of the Spatial Data Lab project
- Challenges and Motivations
- About Spatial Data Lab
- Sponsored Research Activities
- User Guide and Application

Background of the Spatial Data Lab Project

- It is a project in the NSF **IUCRC** Spatiotemporal Innovation Center (**STC**)
- It is sponsored by **China Data Institute** and **Future Data Lab**, two Michigan-based non-profit organizations
- It is conducted at the Center for Geographic Analysis (**CGA**) at Harvard University
- It is a **collaboration** with many contributing organizations and scholars **world-wide**.

IUCRC: Industry–University Cooperative Research Centers

- IUCRC is a program under the National Science Foundation (NSF)
- It enables industrially-relevant, pre-competitive research via a multi-member, sustained partnerships among industry, academe, and government.
- The program was initiated in 1973.
- There are 75 active centers in 2021 (http://www.iucrc.org/centers), several dozen in the planning stage, and 100+ graduated in the past decades.

Why IUCRC

- Contributing to **the nation's research enterprise** by developing long-term partnerships among industry, academe, and government;
- **Leveraging NSF funds with industry** to support graduate students performing industrially relevant pre-competitive research;
- Expanding the innovation capacity of our nation's competitive workforce through **partnerships between industries and universities**; and
- Encouraging the nation's research enterprise to remain **competitive through active engagement** with academic and industrial leaders throughout the world.

Spatiotemporal Innovation Center

- This center's full name is I/UCRC for Spatiotemporal Thinking, Computing, and Applications (STC)
- It is one of the IUCRCs, formed in 2013.
- STC's research focus is on spatiotemporal principles, methodologies and applications.
- The STC completed its first phase (2013-2018), currently is in the second phase (2019-2024).

STC's Vision

Build a national and international spatiotemporal infrastructure to advance --

- **human intelligence** through spatiotemporal thinking;
- **computer software and tools** through spatiotemporal computing;
- **human capability** of responding to deep scientific questions and grand engineering challenges through spatiotemporal applications and education.

Sites and Principle Investigators

- George Mason University (phase I and II)
  - Chaowei (Phil) Yang, Professor of the Department of Geography and Geoinformation Science
- Harvard University (phase I and II)
  - S. V. Subramanian, Professor of Population Health and Geography, Faculty Chair of the Center for Geographic Analysis
- University of California at Santa Barbara (phase I)
  - Keith Clarke, Professor of the Department of Geography

Current and Past Members

- CDI (2019-present)
- DOS (2016-present)
- East View Geospatial (2013-2014)
- FDL (2020-present)
- NGA (2015-2016)
- OmniSci (MapD) (2016-present)
- RMDS (2019-present)
- SASMG (2016-2017)
- STIS (2015-2016)
- USGS CEGIS (2014-2016)
- ZJSG (2017-2019)
- FGDC (2013-2014)
- Harris (2014-2015)
- Microsoft (2013-2014)
- NASA Goddard (2013-present)
- NASA NCCS (2013-present)
- NASA PDCO (2016-present)
- NASG (2013-2018)
- NGCC (2013-2017)
- NOAA (2016-present)
- Northrop Grumman (2014-2016)
- UN (2014-2016)
- Siemens (2016-2018)

Research Projects at Harvard Site (some w/ GMU)

- STC-14-00 Advancing spatiotemporal studies to enable 21st century sciences and applications
- GMU-14-03 Developing a big spatiotemporal data computing platform
- H-14-02 Developing a place name extraction methodology for historic maps
- H-14-03 Development and application of ontologies for NHD and related TNM data layers
- STC-15-02 **Mapping Secondary Cities for Resiliency and Emergency Preparedness**
- H-17-01 **Evaluating OmniSci, Open Source GPU-powered SQL Database**
- H-19-01 **Developing an online spatial data sharing and management platform**
- H-19-06 Elevating Research Excellence with Data Repository and AI Ecosystem
- STC-20-02 **Spatiotemporal Analytics of COVID-19's Second-Order Impacts on Global Vulnerable Urban Areas**
- H-21-01 **Enabling replicable spatiotemporal research with virtual spatial data lab**

Evaluating OmniSci, Open Source GPU-powered SQL Database

- OmniSci is a **massively accelerated analytics and data Science** system. It leverages the full power and parallelism of modern hardware, both CPU and GPU, to explore and extract insight from your largest datasets without lag.
- This project installed OmniSci and PostGIS as **Community Apps on Harvard's HPC Cluster**, and applied it in many big spatial data analytics.
- One application is the computation of **partisan segregation** for 180 million US voters: identify 1000 nearest neighbours and compute distances for **180 billion voter pairs**.

Article on Nature (3/8/2021)

nature human behaviour

ARTICLES

The measurement of partisan sorting for 180 million voters

Jacob R. Brown and Ryan D. Enos

Segregation across social groups is an enduring feature of nearly all human societies and is associated with numerous social maladies. In many countries, reports of growing geographic political polarization raise concerns about the stability of democratic governance. Here, using advances in spatial data computation, we measure individual partisan segregation by calculating the local residential segregation of every registered voter in the United States, creating a spatially weighted measure for more than 180 million individuals. With these data, we present evidence of extensive partisan segregation in the country. A large proportion of voters live with virtually no exposure to voters from the other party in their residential environment. Such high levels of partisan isolation can be found across a range of places and densities and are distinct from racial and ethnic segregation. Moreover, Democrats and Republicans living in the same city, or even the same neighbourhood, are segregated by party.

Segregation between human social groups is associated with a range of profoundly negative outcomes, including intergroup conflict, prejudice, inefficient resource allocation, poor democratic governance and other socially deleterious effects[1–4]. Segregation is also implicated in topics of intense interest across the social sciences, including interpersonal contact and intergroup relations[5,6], the bridging nature of social networks[7–9], poverty[10,11] and political representation[12,13]. Drawing on these associations and using aggregate data, popular and scholarly accounts of politics in the United States—and, increasingly, other Western democracies—describe stark partisan segregation, with members of different political parties living separate lives, resulting in partisan rancour and threatening the functions of the democracy[14–16]. Yet, despite the association between segregation and important outcomes, and the claims of increasing partisan segregation, the measurement of segregation among partisans, as with the measurement of segregation for most social groups, is severely limited: researchers must usually rely on data aggregations that do not include the actual locations of individuals, and thus measurements are limited to summaries across large geographical areas, and the experience of individual exposure across groups is masked.

These high levels of partisan isolation have several important implications. In the United States, political party affiliation is considered a social identity, analogous to race or religion[17], and is a powerful predictor of a range of attitudes and behaviours[18], including behaviours outside of the explicitly political realm[19–21]. Because partisanship is correlated with political ideology and other attitudes and behaviours, the extent of a voter's partisan isolation is likely to affect their exposure to individuals different from themselves and to competing sociopolitical viewpoints, thus affecting a range of important outcomes. Cross-group exposure can be consequential for the shaping of intergroup attitudes and behaviours[1], including the prejudicial attitudes that are levelled across parties in the United States[22].

Isolated partisan environments may also affect behaviour

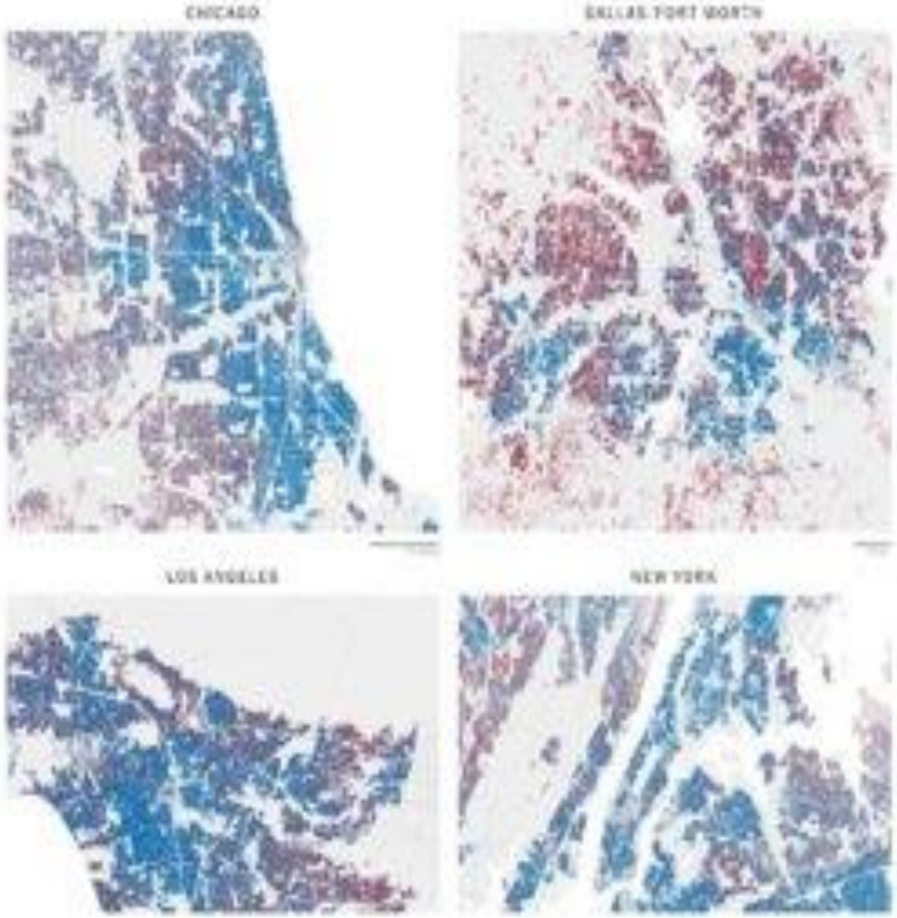
https://www.nature.com/articles/s41562-021-01066-z.epdf

Article on New York Times (3/17/2021)

The New York Times

The Upshot

A Close-Up Picture of Partisan Segregation, Among 180 Million Voters



https://www.nytimes.com/interactive/2021/03/17/upshot/partisan-segregation-maps.html

Mapping Secondary Cities for Resiliency and Emergency Preparedness

- What are Secondary Cities?
  - Secondary Cities are non-primary cities, characterized by population size, function and/or economic status. They are urban centers of governance, logistics, and production. They are are often **data poor** and **under-resourced**.
- Global initiative to address critical geospatial data needs of secondary cities
  - **Emergency preparedness**, **human security**, **resilience**
- Data generation and sharing
  - **Open data access**
- Applied Geography - Human geography thematic areas

* Status as of 2019. In 2021 all cities are graduated.

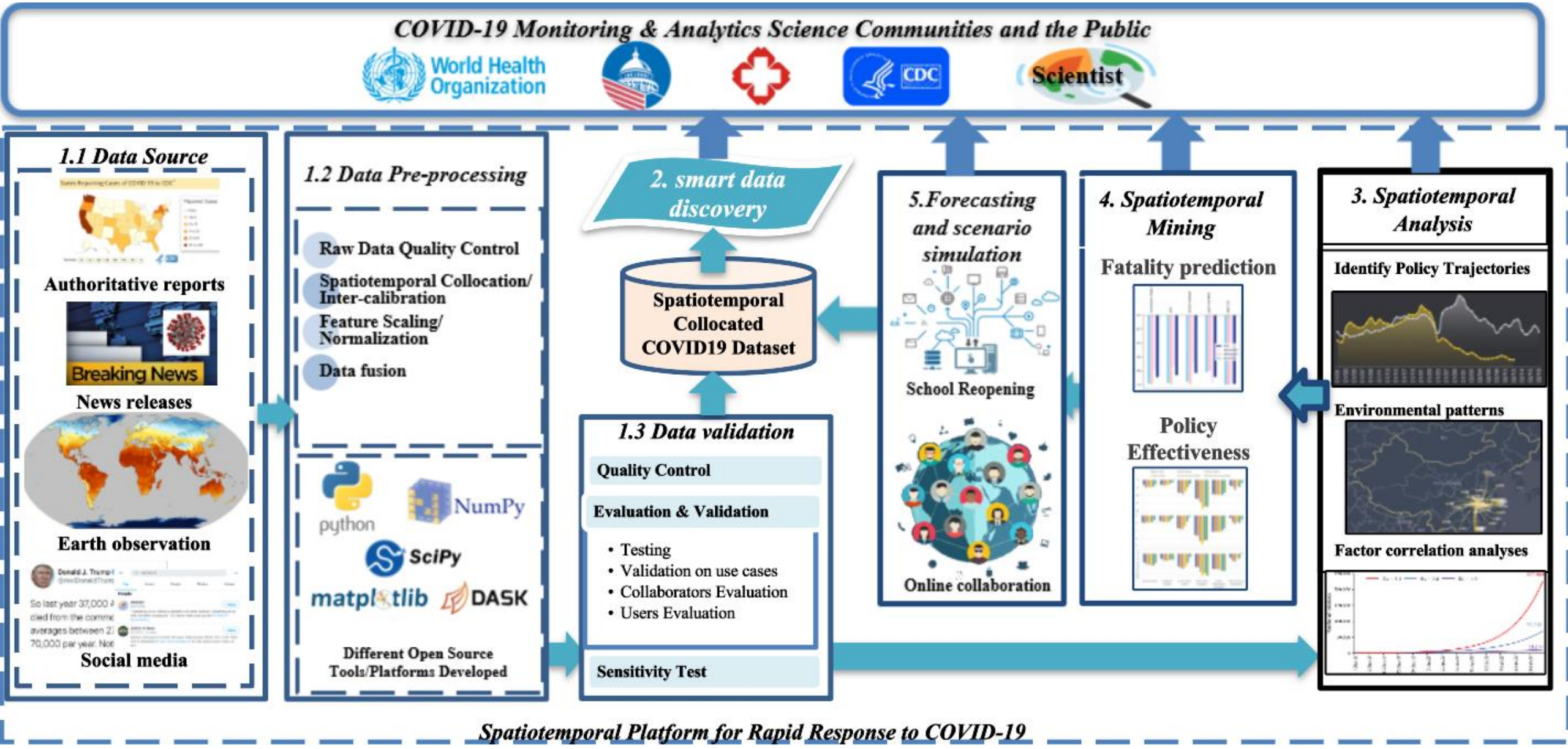
Spatiotemporal Analytics of COVID-19's Second-Order Impacts on Global Vulnerable Urban Areas

- The Cities' COVID Mitigation Mapping program (C2M2) builds on global networks of geospatial experts to analyze **second-order impacts of COVID-19**.
- The goal of this program is to increase the capacity to understand the **distribution and gaps in resources** available to **vulnerable populations in urban communities** through geospatial analysis.
- This program has three **regional hubs** in Africa, Asia, and Latin America, where select local organizations and universities, with regional geospatial and community development expertise, identify and work with various **local project partners** to develop and guide C2M2 projects in each region.

3 Hubs 10 Countries 12 Cities

Image Source: https://mapgive.state.gov/c2m2/

Spatiotemporal Rapid Response to COVID-19



Demands for Cyberinfrastructure for Spatial Studies

- Allowing online spatial data sharing in secured environment
- Enabling replicable spatiotemporal research with advanced tools
- Facilitating research collaborations within team members
- Contributing to Spatiotemporal Rapid Response to COVID-19

Challenges for Spatial and Global Studies

❑ Data Sharing

- Licensed data
- Restricted data
- Sensitive data
- Large size data
- Different Resources

❑ Tool Sharing

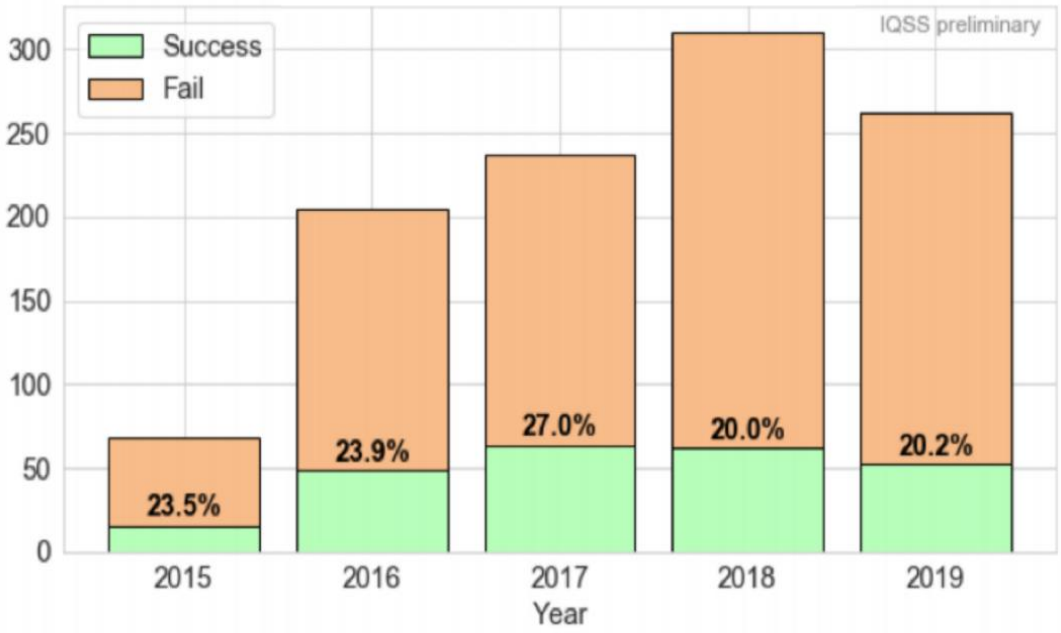
- Licensed and free tools
- Integrated environment for tools and data
- Management (maintenance and updates)

❑ Research Sharing

- Research (reproducible, replicable, and generalizable)
- Training (students with different interests and skills)
- Applications (efficient, effective, and expandable)

The Dataverse® Project

About ***84%*** re-executions are failed



About Spatial Data Lab

As a part of the NSF Spatiotemporal Innovation Center project, the Spatial Data Lab is co-sponsored by the Center for Geographical Analysis at Harvard University, the China Data Institute, and the Future Data Lab.

Goal: build a cyberinfrastructure for effective and efficient spatial and global studies

Objectives:

- ❑ **data sharing** for spatial data studies on the cloud
- ❑ **tool sharing** for quantitative analysis
- ❑ **research sharing** for workflow-based case studies
- ❑ **education and training** on data, methodology, technology, and case studies

Main tasks:

- ❑ spatial data services
- ❑ development of spatial data analysis platform and tools
- ❑ development of workflow based case studies for spatiotemporal analysis
- ❑ spatial data science training

Advisory Committee

S (Subu) V Subramanian, **Committee Chair**
Professor of Population Health and Geography
Harvard University

Jason Ur
Professor of Archaeology
Director of the Center for Geographic Analysis Harvard University

Peter K. Bol
Charles H Carswell Professor
Dept of East Asian Languages and Civilizations
Harvard University

Luc Anselin
Professor of Sociology
Director, Center for Spatial Data Science
University of Chicago

Daniel Sui
Distinguished Professor of Geography
Vice Chancellor for Research and Innovation University of Arkansas

Peng Gong
Chair Professor of Global Sustainability
Vice-President and Pro-Vice-Chancellor, University of Hong Kong

Yasheng Huang
Epoch Foundation Professor of International Management
Professor of Global Economics and Management, Massachusetts Institute of Technology

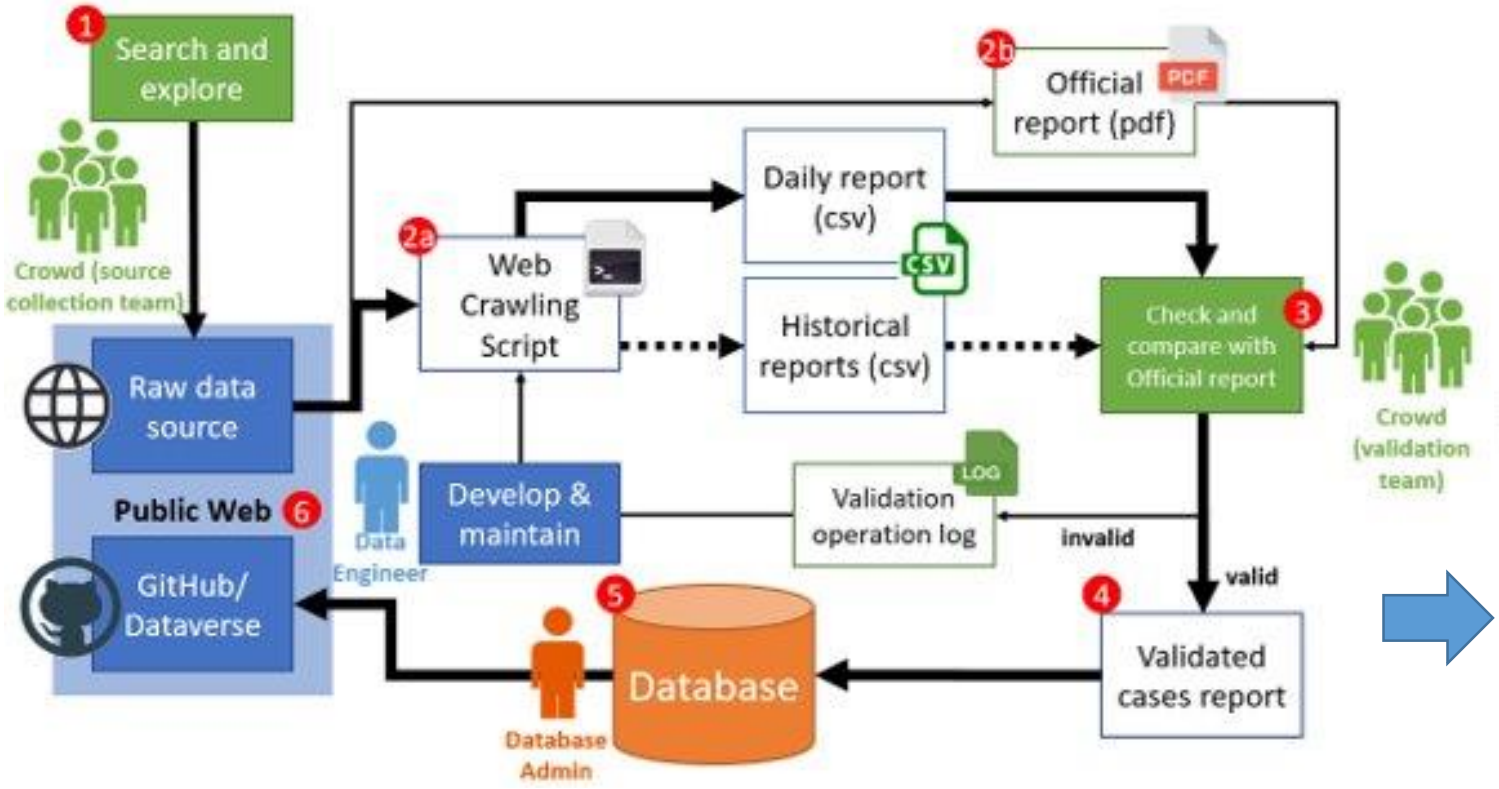
Gary King
Weatherhead University Professor
Director of the Institute for Quantitative Social Science
Harvard University

Peter X Zhou
Director and Assistant University Librarian
C.V. Starr East Asian Library
University of California, Berkeley

Pinde Fu
Platform Engineering Team Lead, ESRI
Adjunct Faculty at University of Redlands and Harvard Extension School

COVID-19 Data Collections

Data collecting, processing, validating and sharing



Map-based Data and standard association

Base Data Collections

- **Government Statistics**
  - Provincial Statistics (1949 -)
  - City Statistics (1996 -)
  - County Statistics (1997 -)
- **Population Census**
  - Census 1953
  - Census 1964
  - Census 1982
  - Census 1990
  - Census 2000/2010 (province, city, county, township, GRID)
- **Economic Census**
  - Industrial Census 1995 (province, city, county, ZIP)
  - Basic Unit Census 2001 (province, city, county, ZIP)
  - Economic Census 2004/2008 (province, city, county, ZIP)
- **Establishments** (more than 7 millions companies and organizations)
- **Geography and Environment**
  - Land Use data
  - Night-Time data

Domain-specific Data Collections

COVID-19 Data Collections:

1. US COVID-19 Daily Cases with Basemap
2. US Metropolitan COVID-19 Daily Cases with Basemap
3. China COVID-19 Daily Cases with Basemap
4. import Cases from forigen countries
5. World COVID-19 Daily Cases with Basemap
6. Community Cases from Tecent
7. Trajactory Data from Tecent
8. China COVID-19 events timeline
9. Global COVID-19 Events Timeline
10. US Vaccine distribution
11. Global Vaccine distribution

Mobility Data Collections:

1. Descartes Lab Mobility Report
2. Mobility Metrics and Social Distancing Index
3. Google Community Mobility Reports
4. Apple Mobility Reports
5. Foursquare Community Mobility Data
6. Safegraph Foot Traffic Data
7. Human OD (Origion-Destination) Flow
9. Baidu Mobility Data

Transportation Data Collections:

1. Flight
2. Train

POI Data Collections:

1. Hospital POI in China
2. Hospital POI in US

Social Media Data:

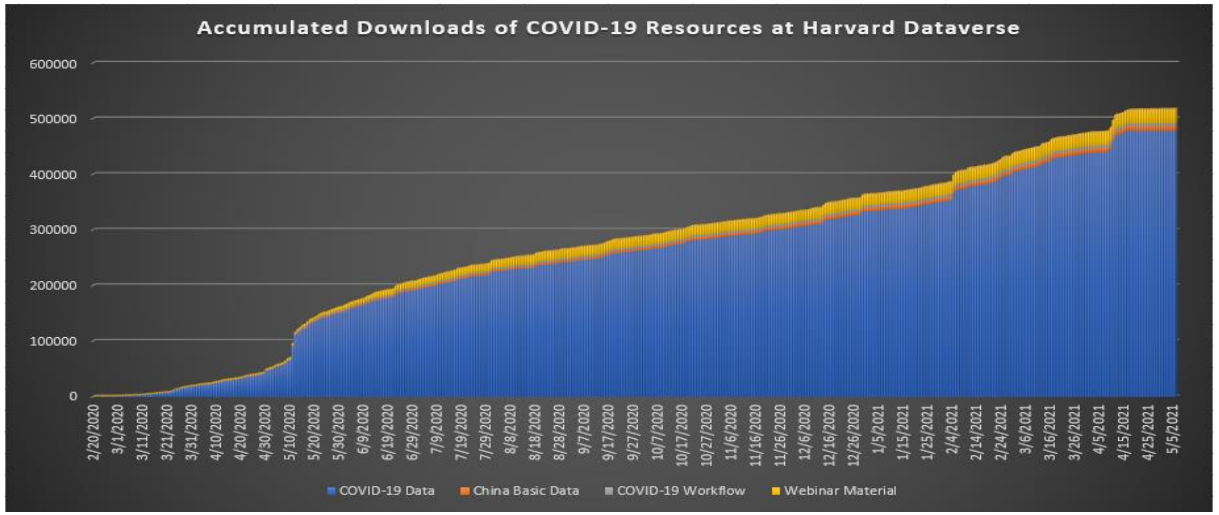
1. GDELT
2. Twitter(eng)

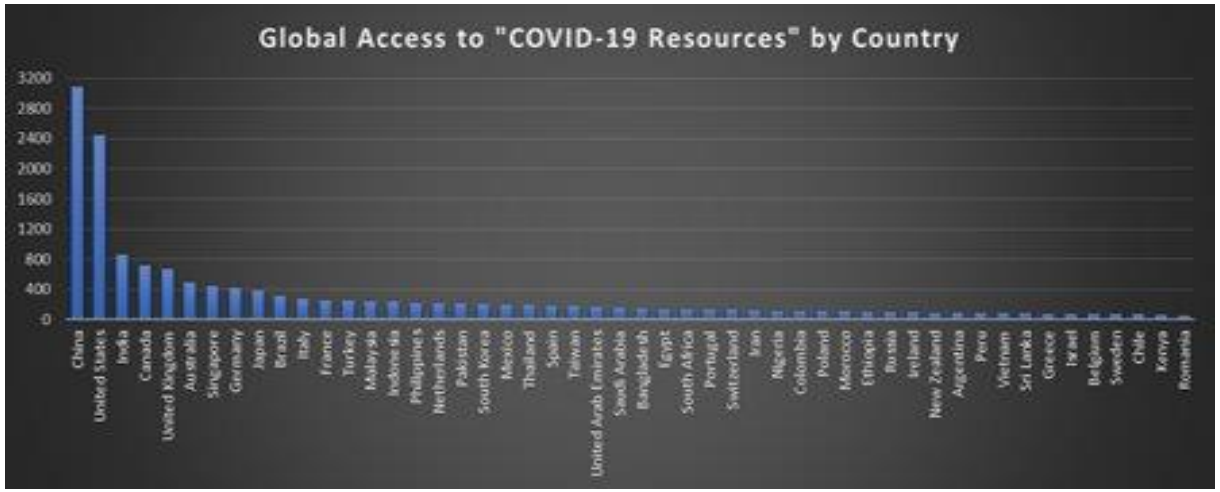
Other Data Collections:

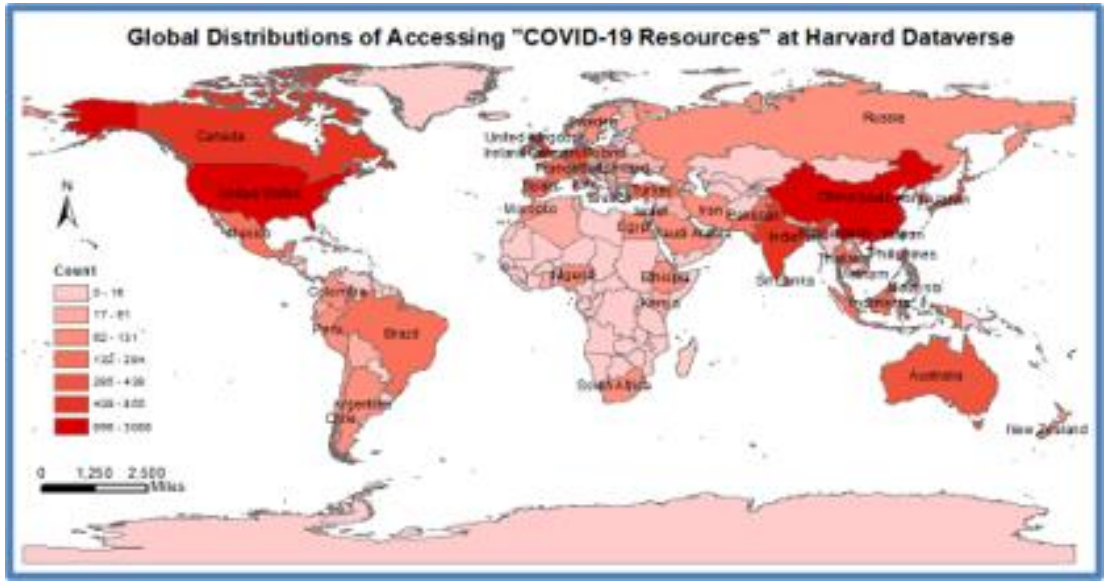
1. City-level Air Quailty Data with China Map
2. City-level Meterlogical Data with China Map
3. Daily Meteorological Data in China
4. City-level CO2 data
5. City-level PM2.5 data
6. China Patent Data
7. US Patent Data
8. Innovation Index, 1990_2019
9. Province Culture, 2004-2018

Datasets for Download on Harvard Dataverse

Global users from **150+** countries with about **600,000+** downloads since Jan 2020.





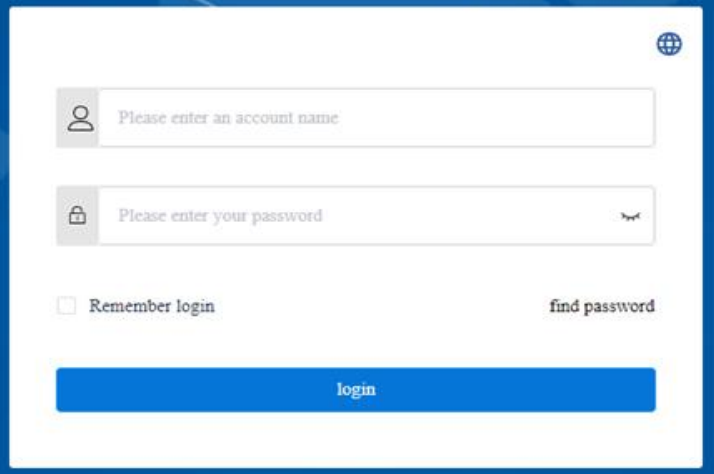


Our Datasets have been cited by many publications and world-wide organizations, including Domino Data Lab; UCGIS; Emory University Libraries, The World Bank/IMF Library, George Washington University Library, NTU Library, and so on.

http://chinadatalab.org

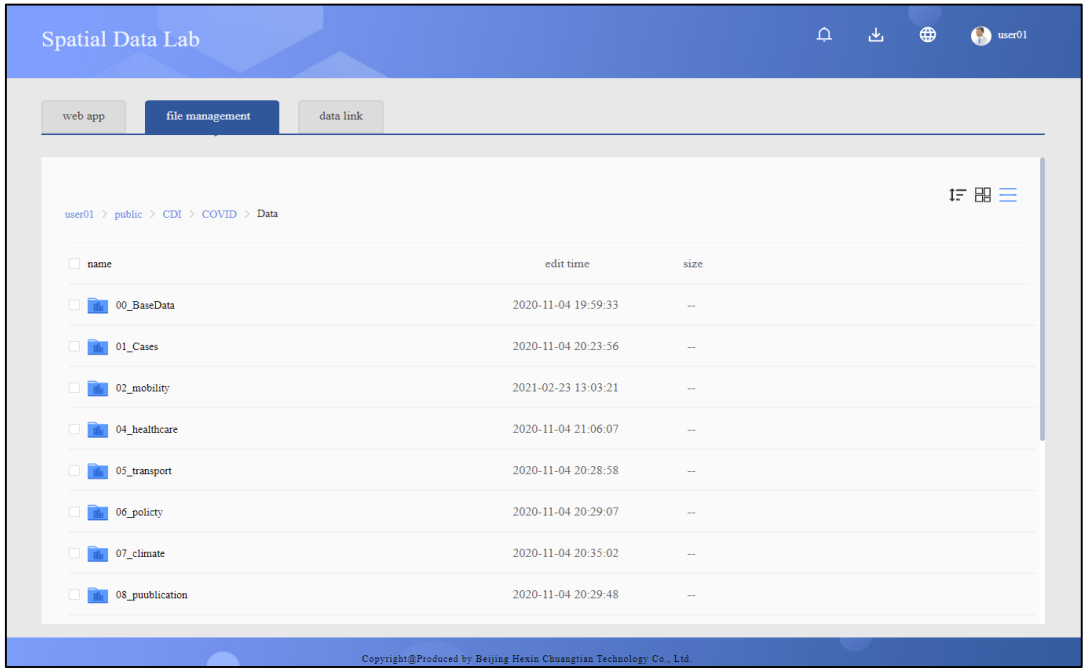
cdl-win02.rc.fas.harvard.edu/static/web/web/index.html#/login

Data Research and Teaching Cloud Platform System

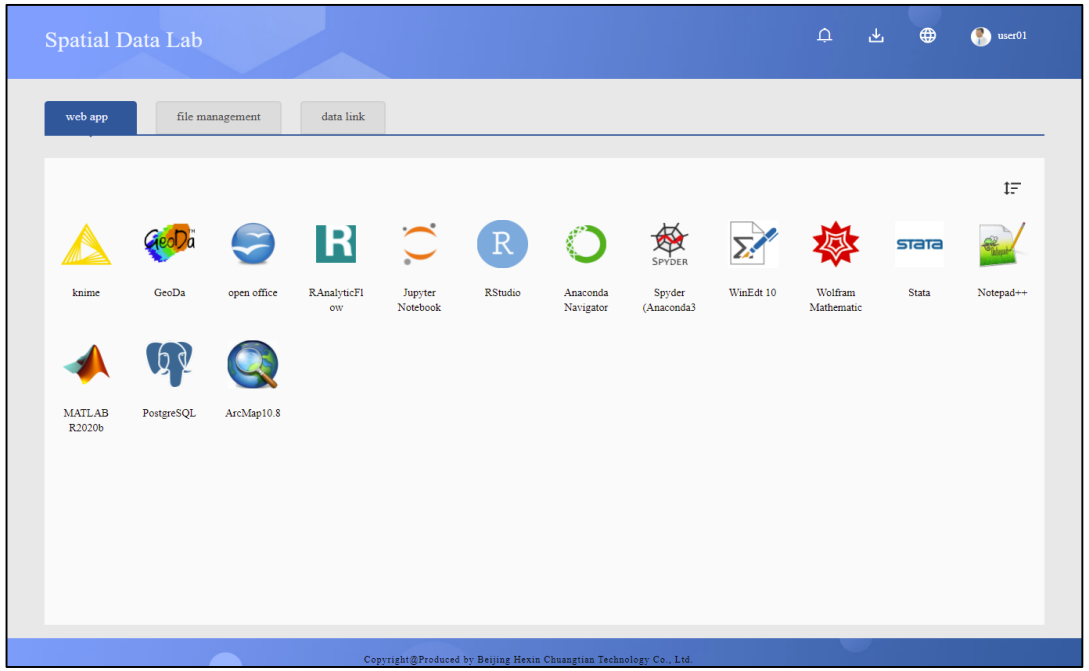


Spatial Data Lab Platform

- ❑ Data available only on the cloud (users can upload own data)
- ❑ Tools available on the cloud
- ❑ All computation are on the cloud (the results can be downloadable)
- ❑ No maintenance required for end users



Personal & Shared Data



Tools

Tools on the Cloud

❑ Office tools

open office

Notepad++

WinEdt 10

❑ Spatial Analysis Tools

GWR4

GeoDa

ArcMap10.8

❑ Statistical Tools

MATLAB R2020b

Stata

Wolfram Mathematica 12.2

❑ Programming tools

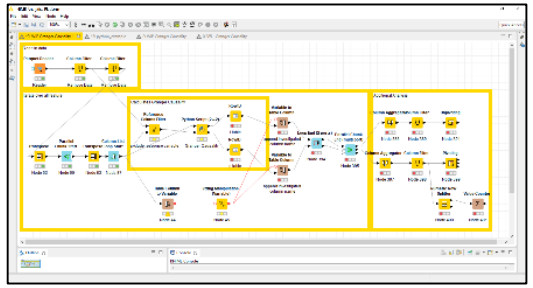
Rstudio

Anaconda Navigator

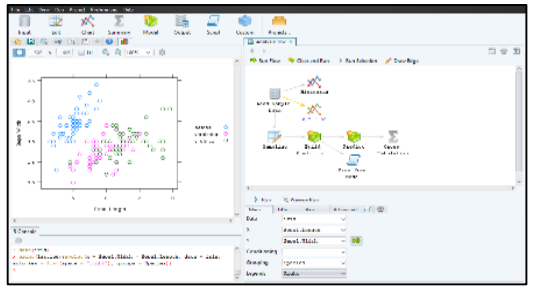
Spyder

❑ Workflow tools

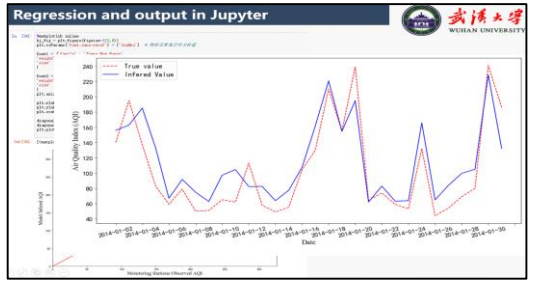
Open for Innovation KNIME



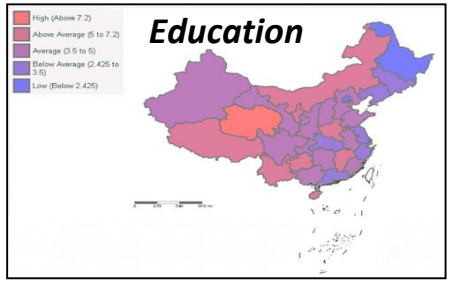
AnalyticFlow

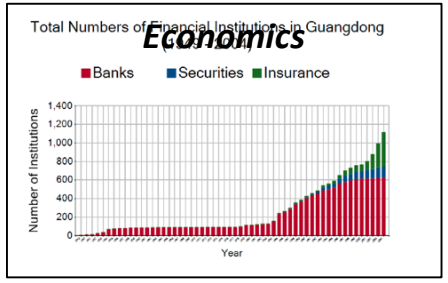


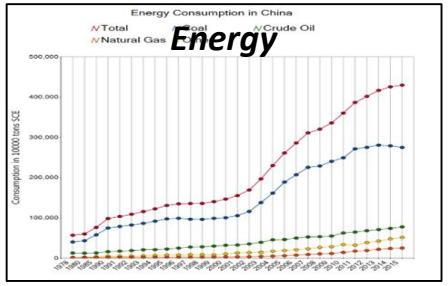
Jupyter Notebook

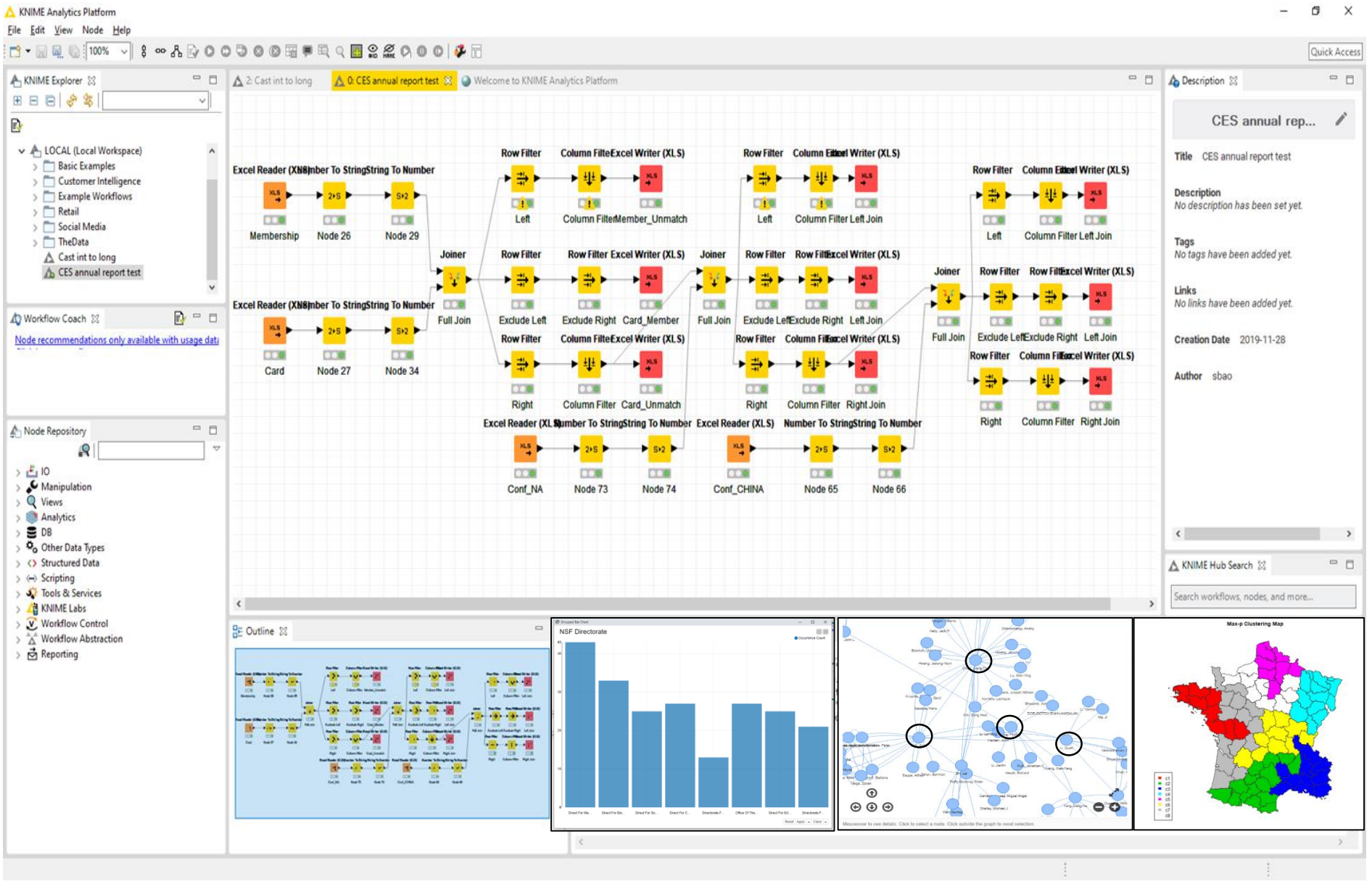


Replicable, Reproducible and Expandable Data Analysis









Sponsored Research

Research groups

- Data collection
- Mobility study
- Patent study
- COVID-19 modeling
- Vaccine study
- Workflow development

Harvard student CVT groups

- Literature Review on COVID twitter and GDELT Research
- Twitter Data Processing & Twitter Data Analysis
- GDELT (Global News) Analysis and Visualization
- Geotagged Social Media-based Vaccination Analysis
- COVID-19 Prediction Models Review
- COVID-19 Impact on Other Causes of Death
- Who got vaccinated?
- Assessing household preparedness for Covid-19 in Bangladesh
- COVID-19 Impact on Schools across the United States

Sponsored Training

About *85* webinars with over *11,000* participants around the world

- **Webinar series on Covid-19 studies** （28）
  - COVID-19 data analysis （12）
  - COVID-19 modeling （8）
  - COVID-19 Impact Analysis （8）
- **Webinar series on data, tools and literature studies** (26)
- **Webinar series on “Spatiotemporal Study of Urban Dynamics”,** jointly with CPGIS （7）
- **Monthly training webinars on research data**, Jointly with CDI （12）
- **Monthly training webinars on workflow based data analysis**, jointly with RMDS Lab （12）

Sponsored Case Studies for Workflow Based Data Analysis

Developed more than 60 workflow based case studies for spatiotemporal research

Workflows for spatiotemporal studies

- Test the impacts of religious diversity on regional development in China
- Spatial panel data analysis in regional development
- Spatiotemporal Pattern Analysis of Higher Education
- Changes in human induced turbidity in Poyang Lake based on remote sensing data
- Air quality analysis with social media data
- Space-time analysis of high education in China
- Spatial factor analysis of road network based on the traffic accidents
- Industrial Co-agglomeration Analysis based on Machine Learning Algorithm

Workflows for COVID-19 Studies

- Dynamic Map Visualization on Time-series COVID-19 Data
- Correlation Analysis between COVID-19 Cases and Human Motilities
- Text Mining on COVID-19 related Social Media Data
- Literature Review on COVID-19 Publications
- Environment Data Processing
- Time-varying SIR Model, provided by University of Michigan
- London School of Hygiene and Tropical Medicine
- DELPHI Epidemiological Case Predictions
- An SEIR Infectious Disease Model with Testing and Conditional Quarantine

Workflows for Data Access

- Database Access and Data Analysis on China Data Online Data
- Database Access and Data Analysis on Patent Data
- Dataverse Data Access and Analysis
- Human Mobility Data Access and Analysis via HTTP
- Github Data Access and Analysis

Workflows for Monthly Training Webinars

- Statistical Data Analysis with Workflows
- Analysis of Population Census Data & Demographic Change
- Analysis of Economic Census Data & Industrial Change
- The Integration of Data and Maps for Spatial Analysis
- Spatial Analysis of Patent Data
- Spatial Analysis of Health with Statistics, Census and GIS Data
- Spatial Analysis of Environment with Statistics, Census and GIS Data
- Spatiotemporal Analysis of Urban Development
- Spatiotemporal Analysis of Rural Development
- Human Mobility in Space and Time

Sponsored Publications

- Chintala, Srikar, Ritvik Dutta, and Doron Tadmor, 2021. COVID-19 spatiotemporal research with workflow-based data analysis. In Infect Genet Evol. 2021 Mar; 88: doi: 10.1016/j.meegid.2020.104701.
- Tadmor, Doron, 2020. COVID-19 Rapid Response Research with Workflow-based Data Analysis. http://dx.doi.org/10.2139/ssrn.3643621.
- Liu, Lingbo,Tao Hu,Shuming Bao,Hao Wu, Zhenghong Peng and Ru Wang, 2021. The Spatiotemporal Interaction Effect of COVID-19 Transmission in the United States. In ISPRS Int. J. Geo-Inf. 2021, 10(6), 387; https://doi.org/10.3390/ijgi10060387.
- Li, Ruiyun, and etc. 2020. Global COVID-19 pandemic demands joint interventions for the suppression of future waves. In PNAS, October 20, 2020. 117 (42) 26151-26157; https://doi.org/10.1073/pnas.2012002117.
- Wang, P., Ren, H., Zhu, X., Fu, X., Liu, H., & Hu, T. (2021). Spatiotemporal characteristics and factor analysis of SARS-CoV-2 infections among healthcare workers in Wuhan, China. Journal of Hospital Infection, 110, 172-177.
- Wang, P., Zhu, X., Guo, W., Ren, H., & Hu, T. (2021). Spatiotemporal Differences of COVID-19 Infection Among Healthcare Workers and Patients in China From January to March 2020. IEEE Access, 9, 28646-28657.
- Peng, Z., Ao, S., Liu, L., Bao, S., Hu, T., Wu, H., & Wang, R. (2021). Estimating Unreported COVID-19 Cases with a Time-Varying SIR Regression Model. International Journal of Environmental Research and Public Health, 18(3), 1090.
- Zhang, M., Wang, S., Hu, T., Fu, X., Wang, X., Hu, Y., . & Bao, S. (2021). Human mobility and COVID-19 transmission: a systematic review and future directions. medRxiv.
- Zhang, M., Wang, S., Hu, T., . & Bao, S. (2021). Human mobility and COVID-19 transmission: a systematic review and future directions. Annals of GIS. (Under Review)
- Tao Hu, Wendy Guan, et al. & Shuming Bao*. Building an Open Resources Repository for COVID-19 Research. Data and Information Management, 2020, 4(3), 130-147.
- Yang, C., Sha, D., Liu, Q., Li, Y., Lan, H., Guan, W. W., . & Wang, Z. (2020). Taking the pulse of COVID-19: A spatiotemporal perspective. International Journal of Digital Earth, 13(10), 1186-1211.
- Hu, T., Yue, H., Wang, C., She, B., Ye, X., Liu, R., . & Bao, S. (2020). Racial segregation, testing sites access, and COVID-19 incidence rate in Massachusetts, USA. medRxiv.
- Siqin Wang, Yan Liu, Tao Hu. Examining the change of human mobility adherent to social restriction policies and its effect on COVID-19 cases in Australia. International Journal of Environmental Research and Public Health.
- Sha, D., Liu, Y, Liu, Q., Li, Y., Tian, Y., Beaini, F., Zhong, C., Hu, T., Wang, Z., Lan, H., Zhou, Y., Zhang, Z. and Yang, C., 2020. A Spatiotemporal Viral Cases Data Collection for COVID-19 Rapid Response. Big Earth Data.
- Yang, C., Sha, D., Liu, Q., Li, Y., Lan, H., Guan, W. W., Hu, T., . & Wang, Z. (2020). Taking the pulse of COVID-19: A spatiotemporal perspective. International Journal of Digital Earth, 13:10, 1186-1211.
- Liu, Q., Sha, D., Liu, W., Houser, P., Zhang, L., Hou, R., . Hu, T., & 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), 1576.
- Hu, T., & Zhang, Y. (2020). A spatial–temporal network analysis of patent transfers from US universities to firms. Scientometrics, 1-28.
- Hu, T., She, B., Duan, L., Yue, H., & Clunis, J. (2019). A Systematic Spatial and Temporal Sentiment Analysis on Geo-Tweets. IEEE Access, 8, 8658-8667.
- Hu, T., Wang, S., She, B., Zhang, M., Huang, X., Cui, Y. & Li, Z. (2021). Human Mobility Data in the COVID-19 Pandemic: Characteristics, Applications, and Challenges. International Journal of Digital Earth. (Under Review)

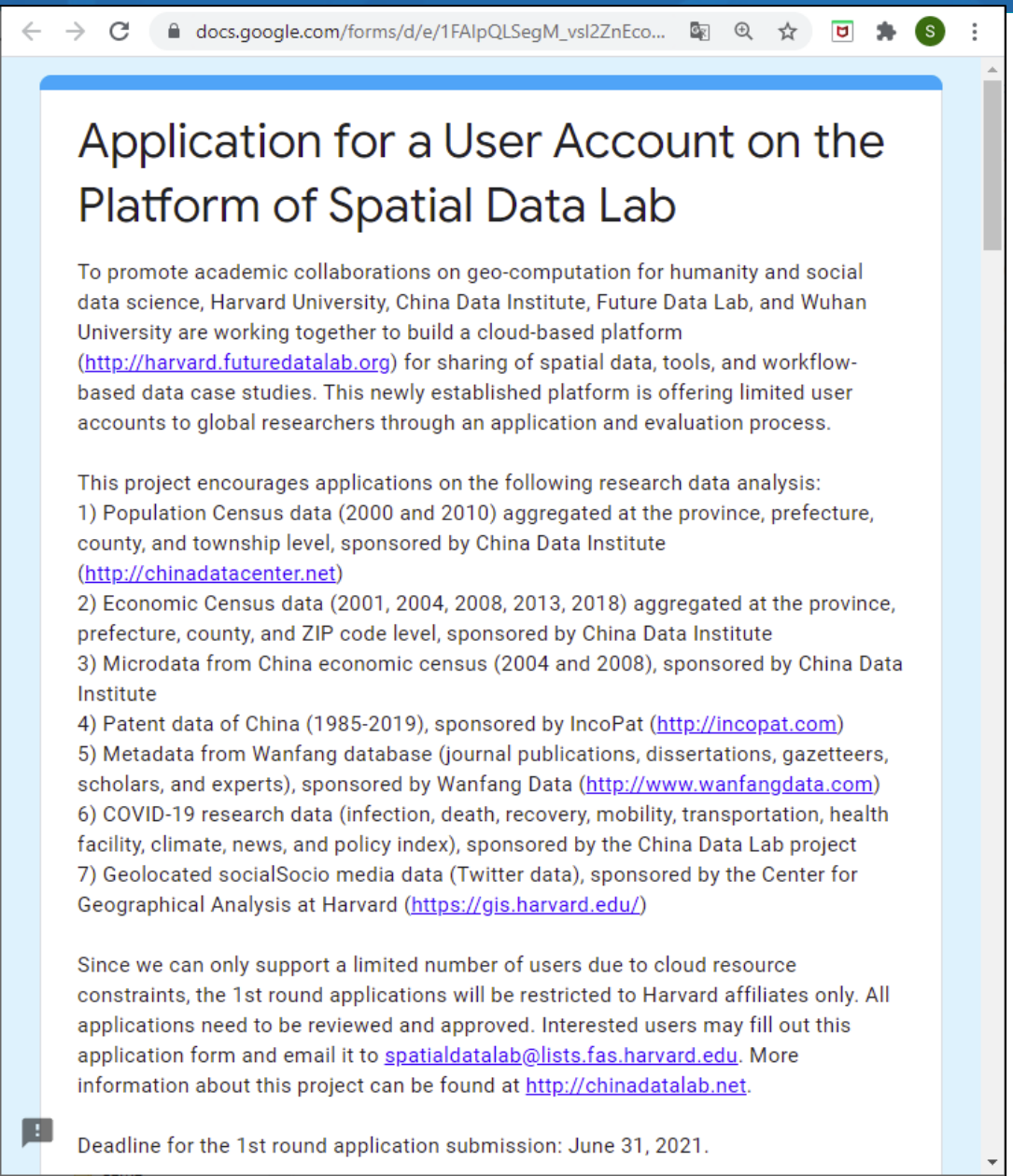
Participants

- NSF Spatiotemporal Innovation Center
- Geocomputation Center for Social Sciences, Wuhan University
- Dataverse team, IQSS, Harvard University
- Center for Geographic Analysis, Harvard University
- China Data Institute
- Future Data Lab
- RMDS Lab
- IncoPat, Inc
- Vesystem Inc.
- All China Marketing Research Inc.

Open Opportunities for Global Collaboration

- ❑ Data collection and service
- ❑ Tool development
- ❑ Workflow based case studies development
- ❑ Training on spatial data analysis
- ❑ Research applications

User Guide and Application for Spatial Data Lab

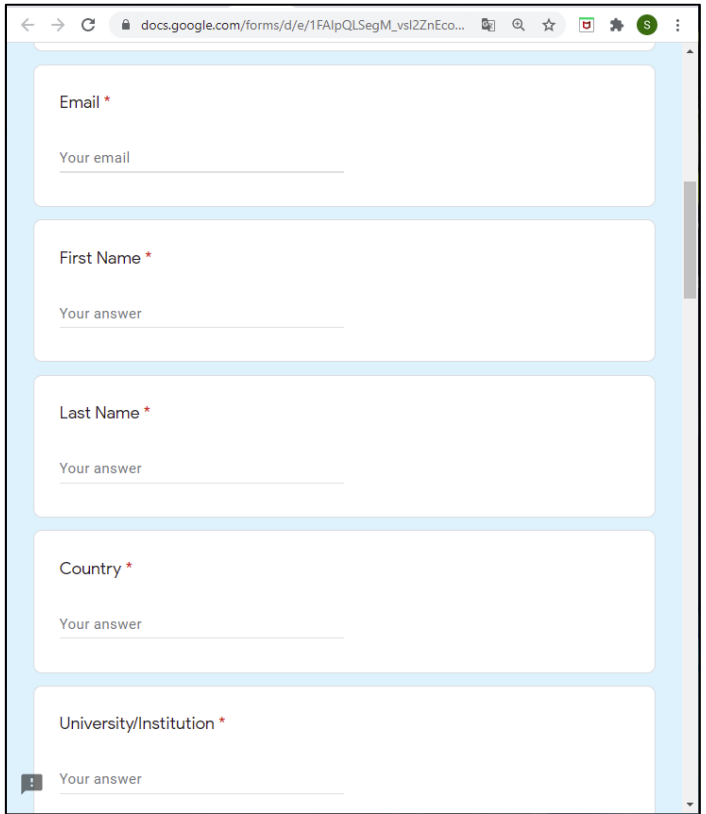


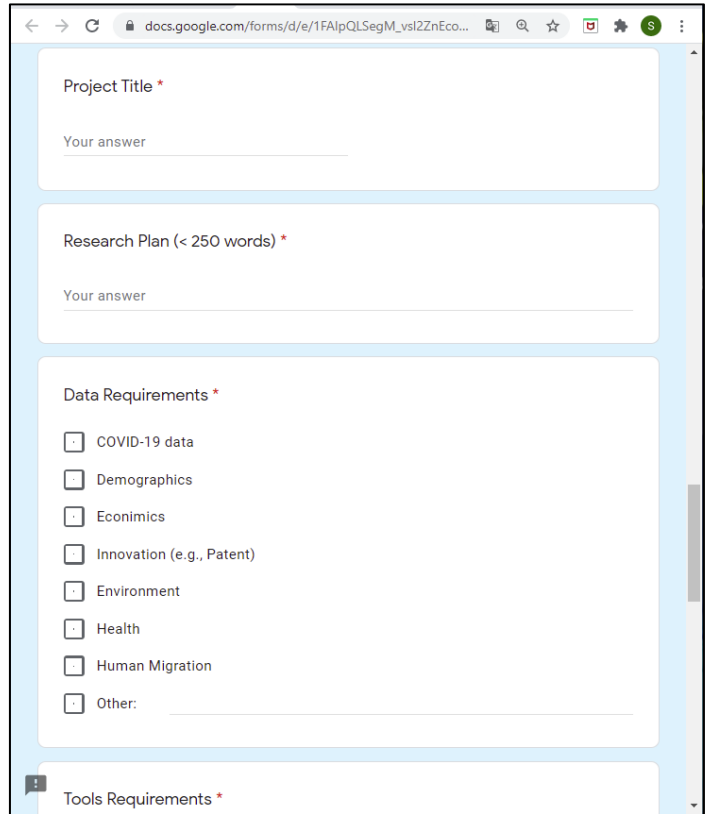
☐ User Guide for Spatial Data Lab

https://docs.google.com/document/d/1xFeBGh325AZk11d-XoXP8lE_5Hk1-RGM/edit

☐ Online Application Form

https://docs.google.com/forms/d/e/1FAIpQLSegM_vsl2ZnEcoGSv5CrjDrRqtDfo-zsuJ19RJyPPmQIgLrOw/viewform





Web Site & Contact

Spatial Data Lab

http://spatialdatalab.org

Contact

Spatialdatalab@lists.fas.harvard.edu