China Data Lab

HARVARD UNIVERSITY

Introduction to China Data Lab (CDL)

https://chinadatalab.net

Presenter: Wendy Guan & Tao Hu

Date: 2019-09-24

Outline

① About China Data Lab (CDL)

② CDL Platform

③ Data Management

④ Plans

About CDL

❑ Establishment

China Data Lab

A cloud-based geospatial data analysis platform for geospatial data gathering, management, analysis, visualization, and sharing.

Sponsored by the Spatiotemporal Innovation Center of the NSF Industry-University Cooperative Research Centers (I/UCRC) Program

CGA

The Center for Geographical Analysis (CGA) at Harvard University. Its core mission is to support research and teaching in all disciplines across Harvard University with emerging geospatial technologies.

CDI

The China Data Institute, a Michigan based not-for-profit organization. It aims to promote the use and sharing of China data; support quantitative research on China in social science, digital humanity and other research subjects.

GCSS

The GeoComputation Center for Social Science at Wuhan University. It promotes the scientific research on the theory and method of spatial data in scientific research, personnel training, international cooperation and social practice.

Partners for Data, Tools and Case Studies

esri China BEIJING THE SCIENCE OF WHERE · CSMAR · ACMR 北京华通人商用信息有限公司 · GAUSS · Knowledge@Share 智享

Goal and Objectives

Build a resources center for the spatial study of China and training

Objectives:

- **A data center** for China studies based on cloud
- **A development center** for data case studies on China
- **A research center** for collaborations on China studies
- **A training center** for China studies, including theory, methodology, technology, data and applications for research and teaching

projects.iq.harvard.edu/chinadatalab

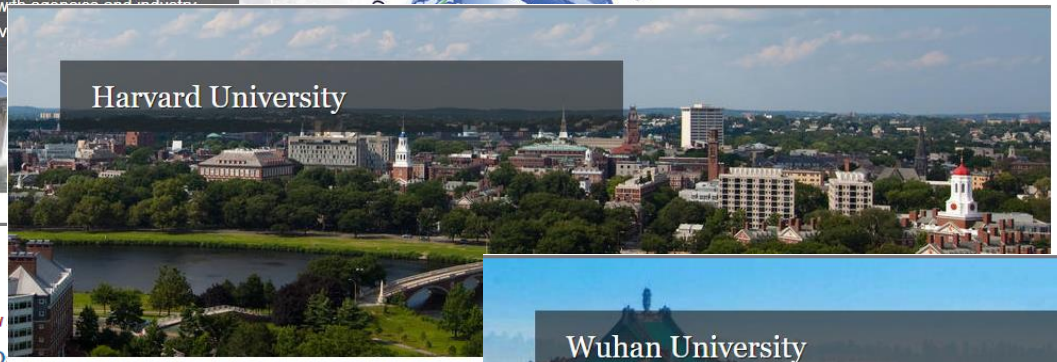
HARVARD UNIVERSITY

HARVARD.EDU

China Data Lab

http://chinadatalab.net

HOME | People | Resources | Partners | Events | About



LAB NEW

China D

Friday, Ma

NSF I/UCRC Spatiotemporal Innovation Center Held

Friday, June 21, 2019

Advisory Committee

Jason Ur, **Committee Chair**
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
Professor, Department of Earth System Science
Dean, School of Science
Tsinghua University

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

China Data Lab

http://chinadatalab.net

Alteryx

GAUSS

GeoDa

GWR4

ArcMap

Jupyter

RAnalyticFlow

KNIMEAnalytics

Outline

① About China Data Lab (CDL)

② CDL Platform

③ Data Management

④ Plans

CDL Platform Framework

China Data Lab Platform

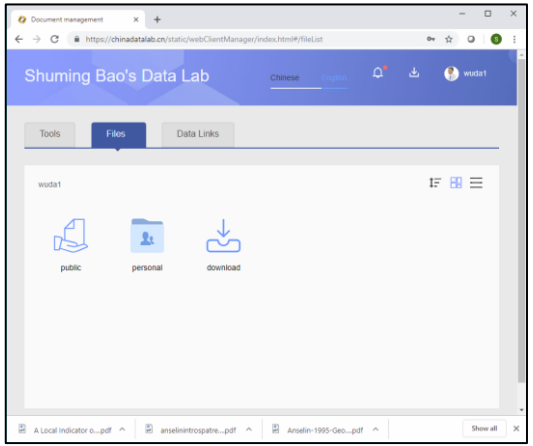
End Users

Cloud Servers

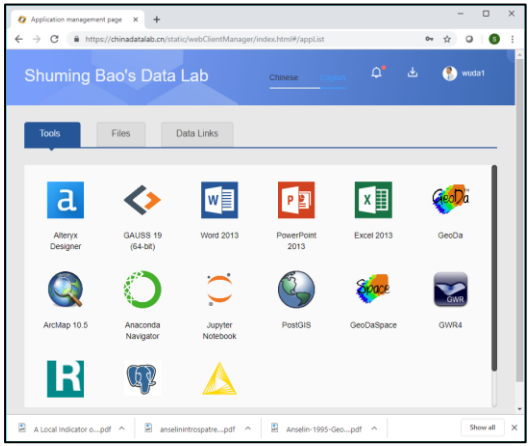
Cloud Platform Features

- ❑ 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

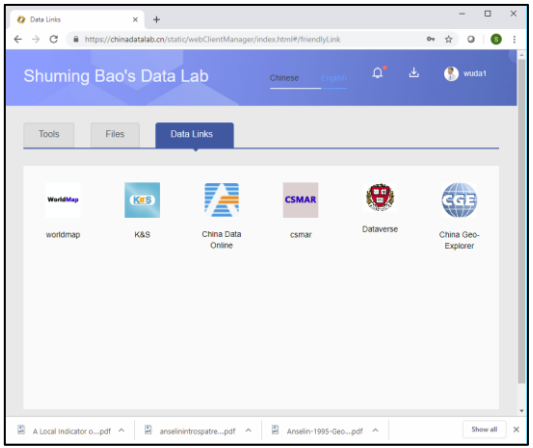
Personal & Shared Data



Tools

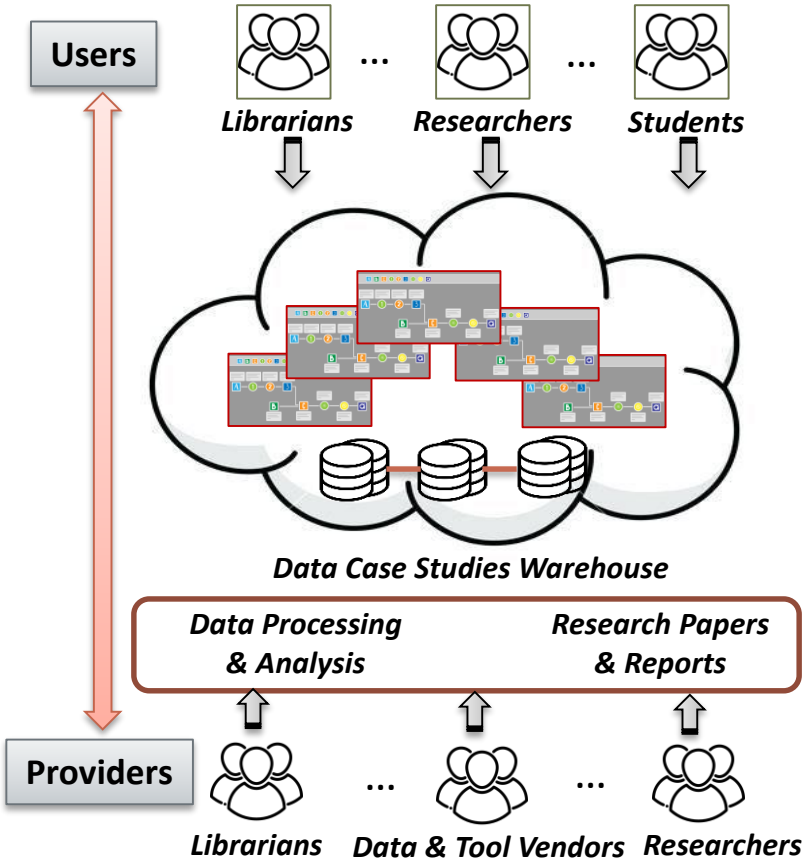


Links



CDL Platform Features

- ❑ Cloud-based virtual desktop (easy to access)
- ❑ Customized software / tool installation
- ❑ Efficient workflow design and implementation
- ❑ Large scale data management (to do)
- ❑ Varieties of resources sharing (data and case studies)



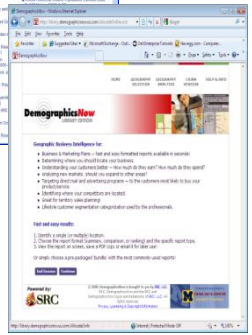
Authentic, Unique, Comprehensive, & Curated China Data

- **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

China Geo-Explorer

Statistics

Census



GIS

Data

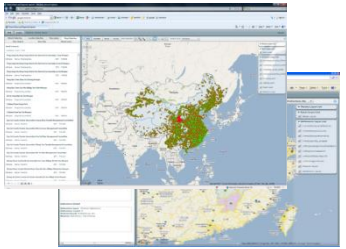
Output

Charts

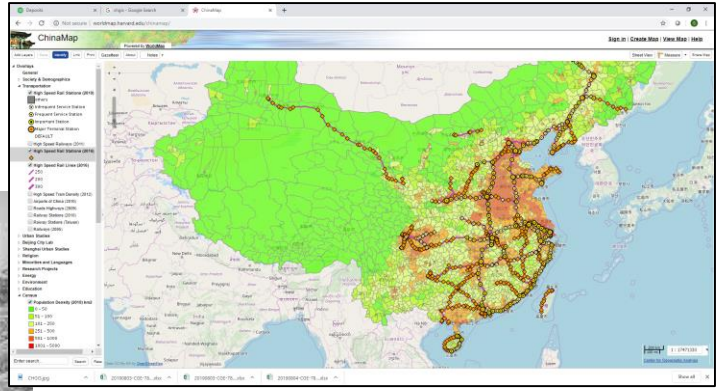
Tables

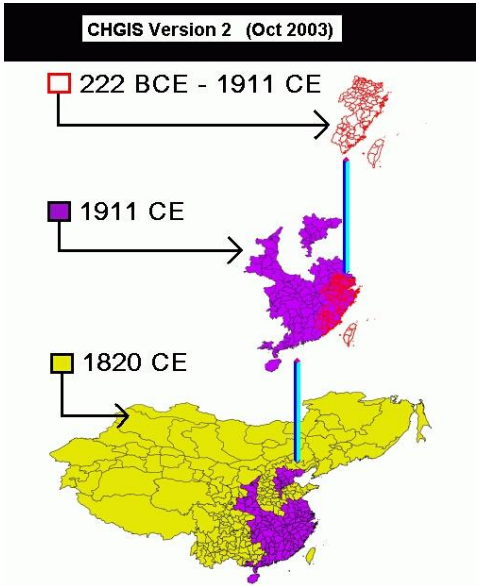
Reports

Maps



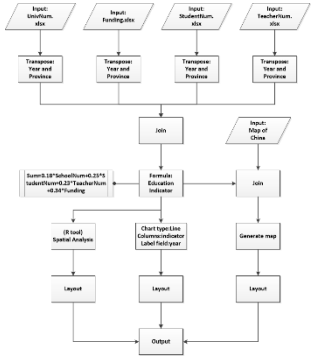
CHGIS: GIS Data for Historical Studies of China



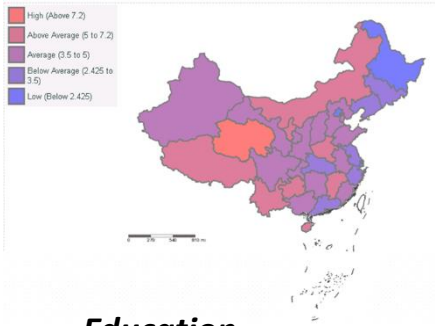


CDL for Reproducible, Replicable, Generalizable Research

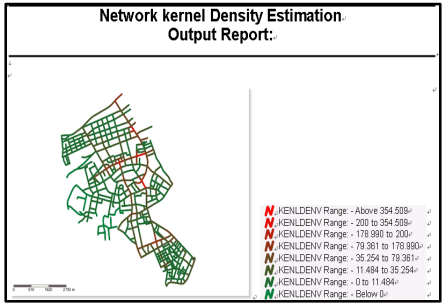
❑ Workflow Data Analysis



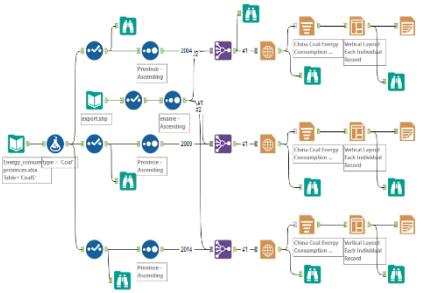
Environment

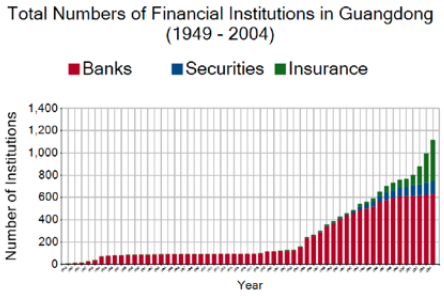


Education

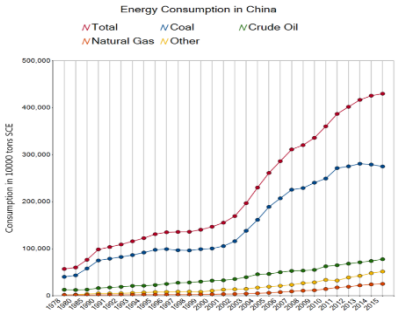


Transportation

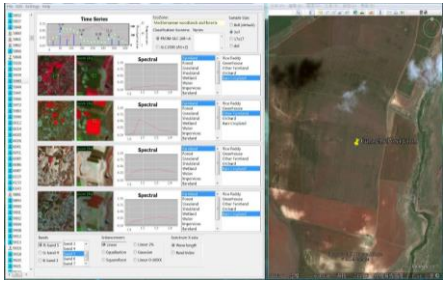




Economics

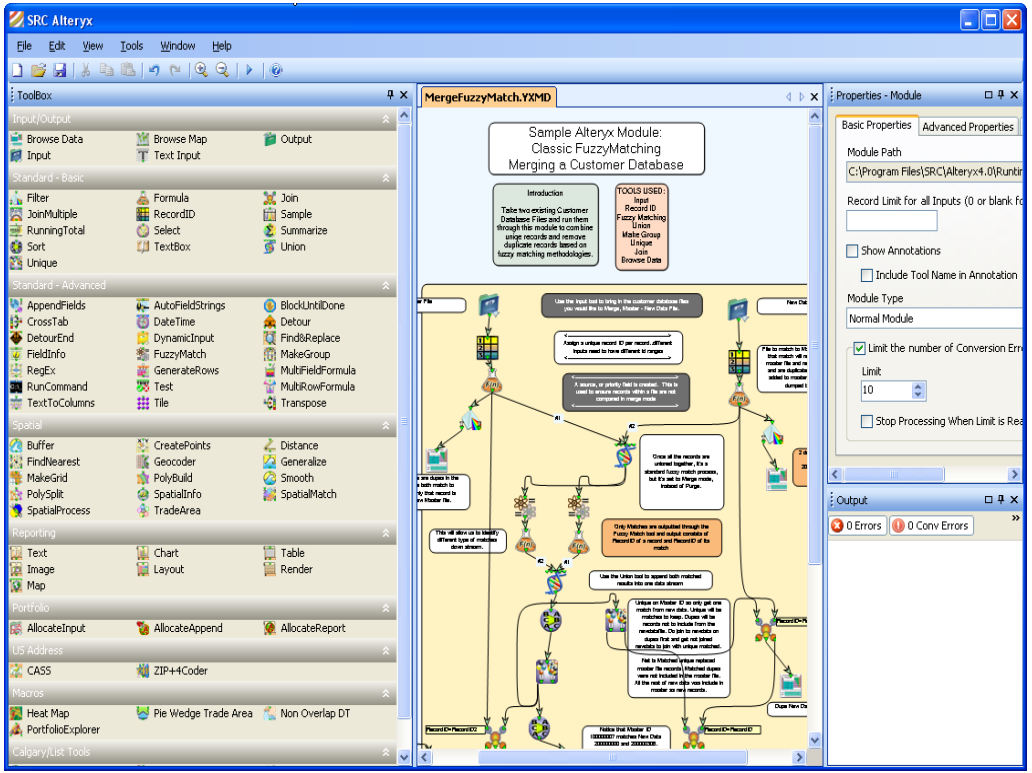


Energy

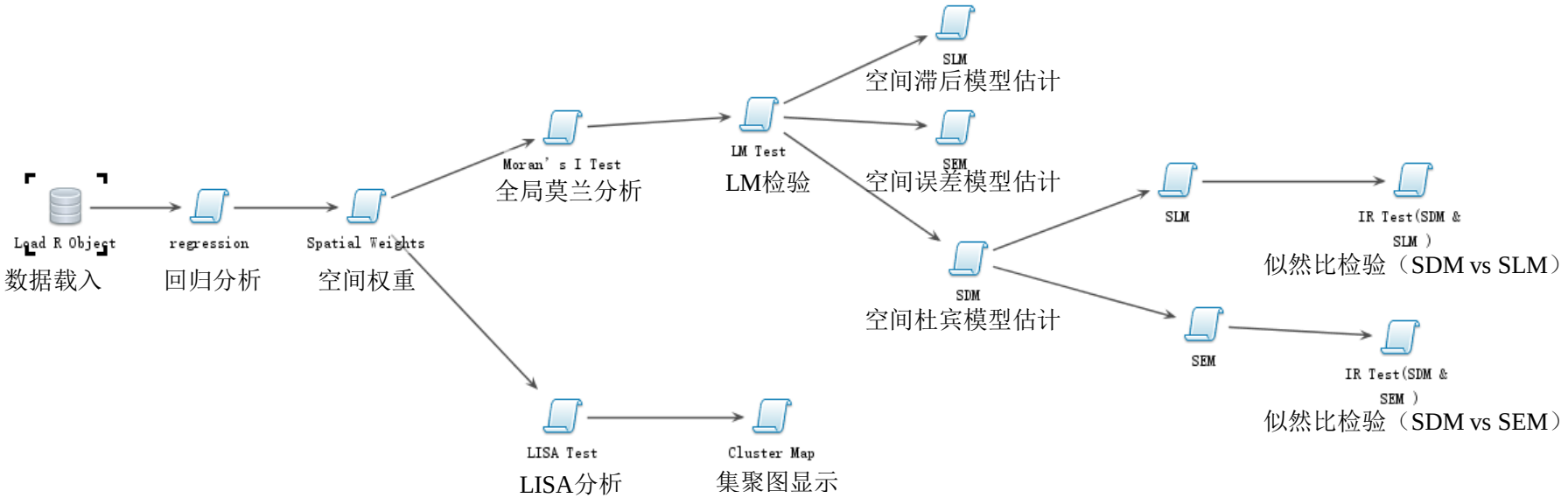


Land Use

Workflow Data Analysis with Alteryx



Workflow Data Analysis with R AnalyticFlow



Liner regression estimates

lm(formula = CRIME ~ INC + HOVAL, data = columbus)

Residuals:

Min	1Q	Median	3Q	Max
-34.418	-6.388	-1.580	9.052	28.649

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	68.6190	4.7355	14.490	< 2e-16	***
INC	-1.5973	0.3341	-4.780	1.83e-05	***
HOVAL	-0.2739	0.1032	-2.654	0.0109	*

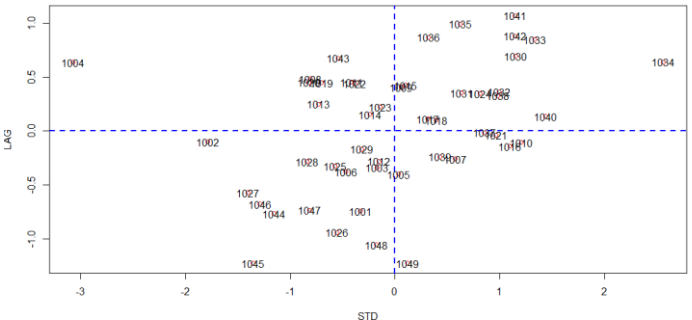
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.43 on 46 degrees of freedom
Multiple R-squared: 0.5524, Adjusted R-squared: 0.5329
F-statistic: 28.39 on 2 and 46 DF, p-value: 9.341e-09

Liner regression estimates

	Ii	E.Ii	Var.Ii	Z.Ii	Pr(z > 0)
1005	-0.0127	-0.0208	0.4615	0.0120	0.4952
1001	0.2501	-0.0208	0.3014	0.4935	0.3108
1006	0.1818	-0.0208	0.2213	0.4308	0.3333
1002	0.1811	-0.0208	0.2213	0.4292	0.3339
1007	-0.1517	-0.0208	0.1184	-0.3803	0.6481
1008	-0.4008	-0.0208	0.4615	-0.5593	0.7120
1004	-2.0018	-0.0208	0.2213	-4.2108	0.9999
1003	0.0588	-0.0208	0.1413	0.2118	0.4161
1018	0.0407	-0.0208	0.1013	0.1935	0.4233
1010	-0.1277	-0.0208	0.2213	-0.2272	0.5899
1038	0.3280	-0.0208	0.1733	0.8381	0.2010

.....



Lagrange multiplier diagnostics for spatial dependence

model: lm(formula = CRIME ~ INC + HOVAL, data = columbus)
weights: col.listw

LMerr = 4.6111, df = 1, p-value = 0.03177

Lagrange multiplier diagnostics for spatial dependence

model: lm(formula = CRIME ~ INC + HOVAL, data = columbus)
weights: col.listw

RLMerr = 0.033514, df = 1, p-value = 0.8547

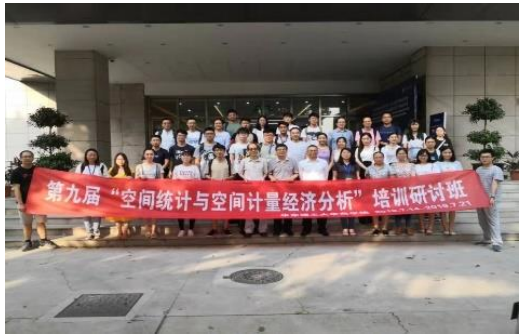
CDL Platform for Training

Three training workshops were held in Wuhan, Shanghai and Beijing using the CDL platform with over 200 scholars and students from different institutions.

- Spatial panel data analysis in regional development with R AnalyticFlow
- Changes in human induced turbidity in Poyang Lake based on remote sensing data
- Air quality analysis with social media data with Jupyter
- Space-time analysis of high education in China with Alteryx
- Spatial factor analysis of road network based on the traffic accidents with Alteryx

Training Workshop on "Spatial Statistics and Spatial Econometric Analysis"

Shanghai



Training Workshop on "Spatial Data Lab"

Wuhan

Training Workshop on "Spatial Econometrics"

Beijing

Outline

① About China Data Lab (CDL)

② CDL Platform

③ Data Management

④ Plans

The Synergy of Dataverse & WorldMap & CDL

Integration, Searching, Sharing, Visualization, Execution

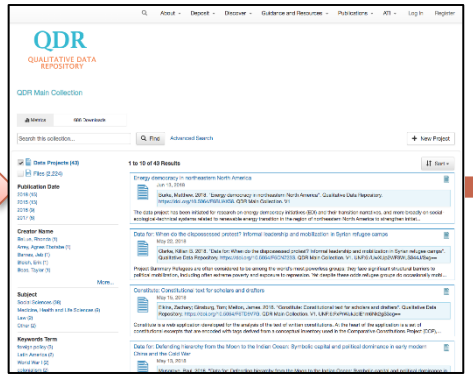
- ❑ Manage spatial and non-spatial data
- ❑ Provide integrated environment for research and training
- ❑ Ensure data security and privacy

CDL Data

- Census Data
- Statistics
- Map Data
- Case Studies
- ...

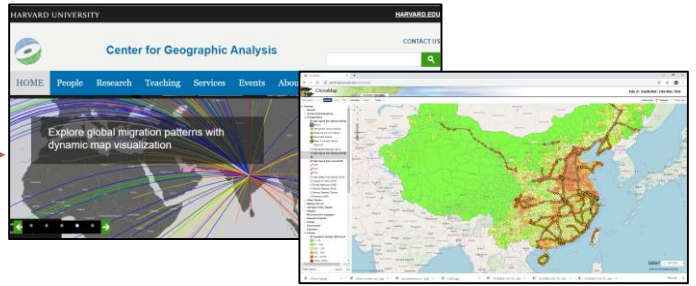
DataVerse

Data Management



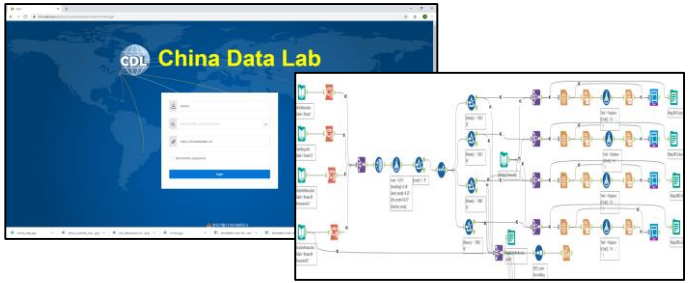
WorldMap

Data Visualization



China Data Lab

Workflow Data Analysis



DataVerse: An Open Tool for Data and Metadata Management

The Dataverse® Project

Open source research data repository software

Researchers

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

Journals

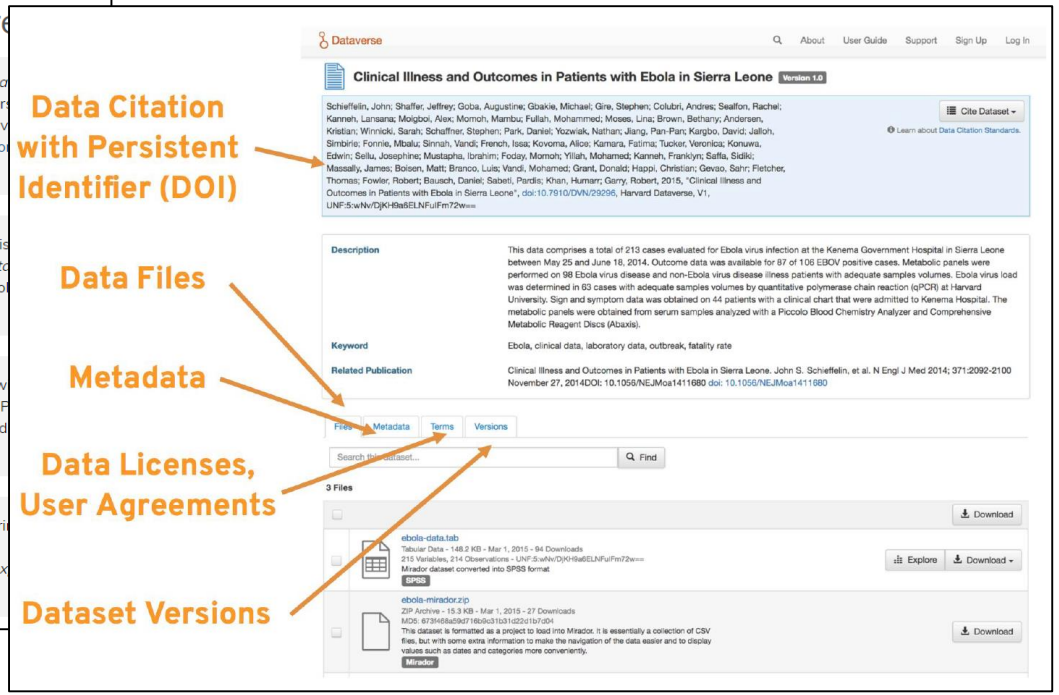
Seamlessly manage the submission, review, and publication of data associated with publis articles. Establish an *unbreakable link* between *articles in your journal* and *associated data* Participate in the open data movement by using Dataverse as part of your journal data pol of repository recommendations.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

Institutions

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

Developers

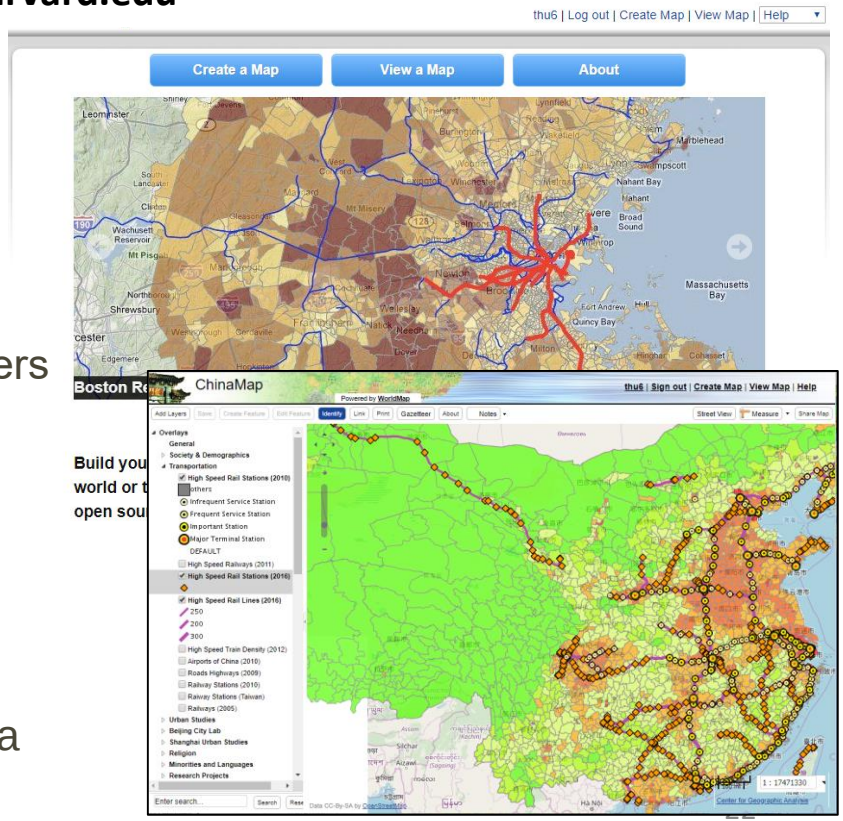
Participate in a vibrant and growing community that is helping to drive the norms for shari preserving, citing, exploring, and analyzing research data. Contribute code extensions, documentation, testing, and/or standards. *Integrate research analysis, visualization* and *ex tools,* or other research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



WorldMap for Spatial Data Depository & Visualization

http://worldmap.harvard.edu

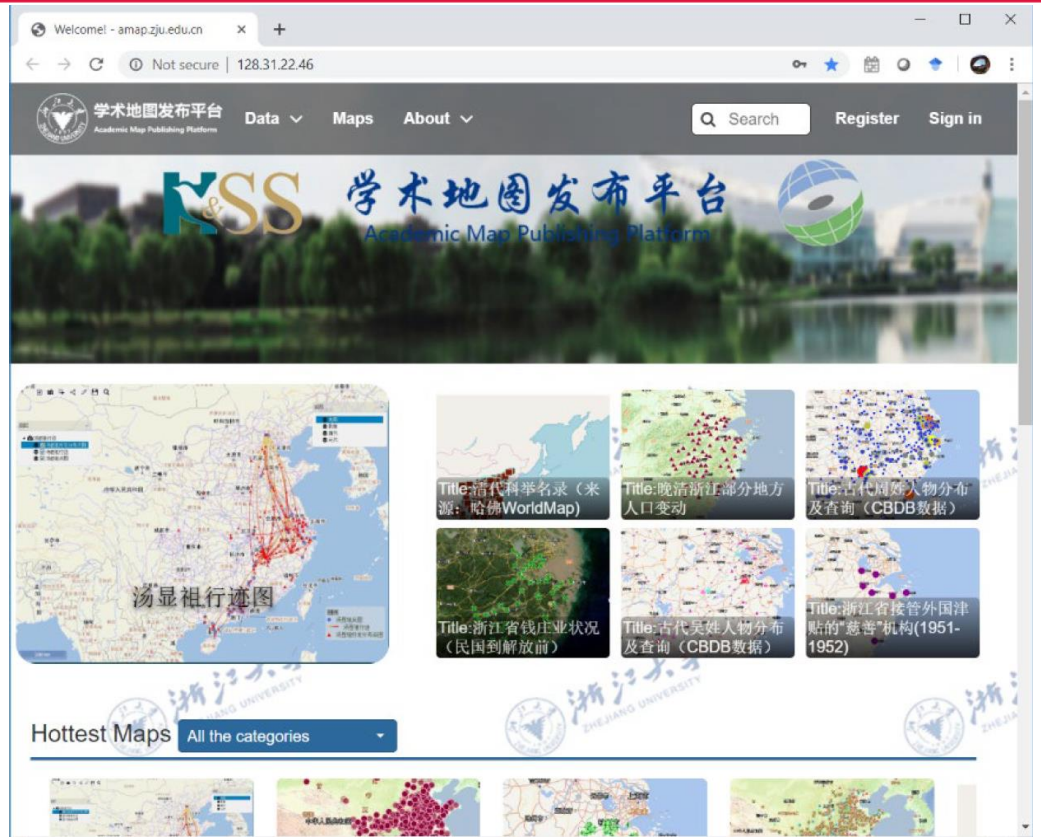
- **Allow Users to**
  - **Organize** one's own mapping datasets online
  - **Discover** other people's (public) data
  - **Mashup / Overlay** one's data with data of others
  - **Create** maps with complex symbology
  - **Collaborate** by letting several people edit the same map, comment on maps and rank data
  - **Publish** maps and data to the world or to just a few collaborators



CGA: China Academic Map Publishing Platform

- Zhejiang University, China
- http://amap.zju.edu.cn/
- Over 300 maps

  - 汤显祖行迹图
  - 全宋文作者分布
  - 《四库全书总目提要》定位查询
  - 浙江家谱目录查询
  - 清代妇女作者分布图
  - 第一批汉文珍贵古籍名录定位查询
  - 唐寅（1470-1523）行迹图
  - 出版家张元济行迹图（1867-1959）
  - 袁宏道（1568-1610）行迹图
  - More…



CGA: China Academic Map Publishing Platform

哈佛燕京学社社友（HYI ALUMNI)分布及查询（截

Geographic Distribution of Harvard-Yenching Institute Alumnus

Info | Share | Ratings | Comments | Favorite

Title	哈佛燕京学社社友（HYI ALUMNI)分布及查询（截止2017年）
License	Not Specified
Publication Date	July 9, 2018, 11:10 a.m.
Category	人文历史
Regions	Global
Owner	camp2018

曾国藩（1811-1872）行迹图

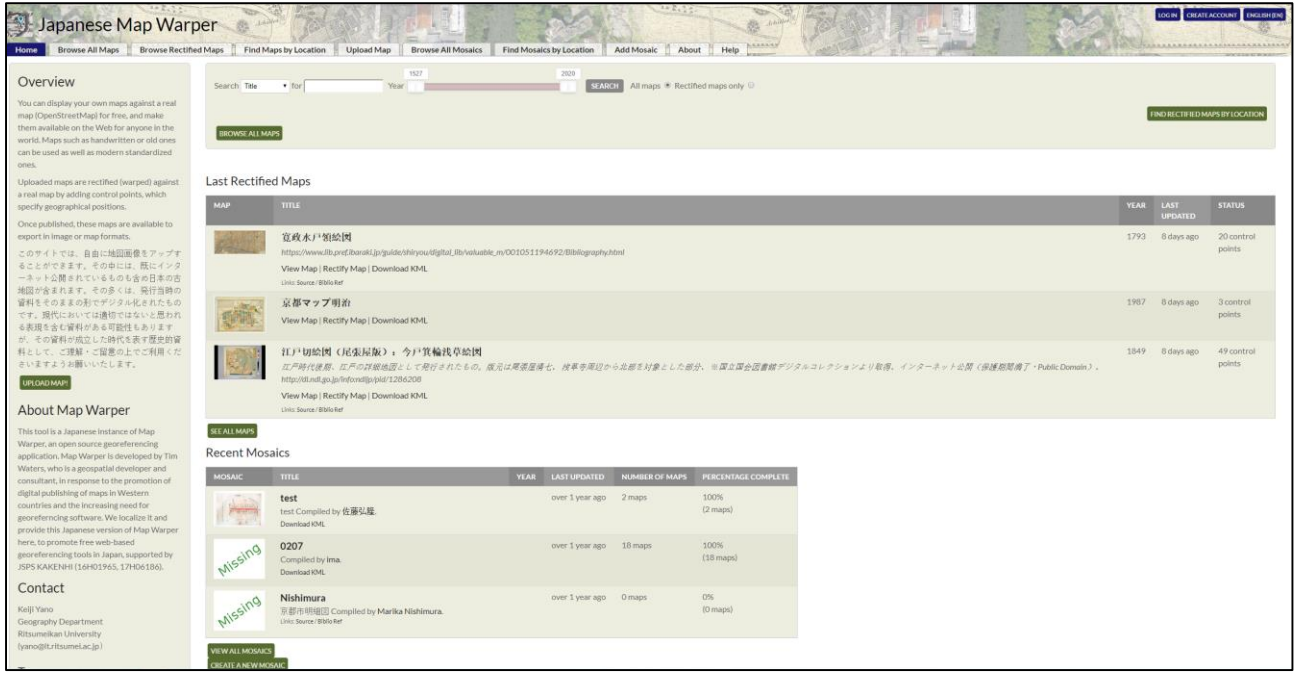
Guofan Zeng's Geographic Trajectories in his life

Info | Share | Ratings | Comments | Favorite

Title	曾国藩（1811-1872）行迹图
License	Not Specified
Publication Date	July 5, 2018, 12:28 p.m.
Category	人文历史
Regions	Global
Owner	3150105755

CGA: An Open Tool for Processing & Sharing Scanned Maps

- http://mapwarper.h-gis.jp/



Outline

① About China Data Lab (CDL)

② CDL Platform

③ Data Management

④ Plans

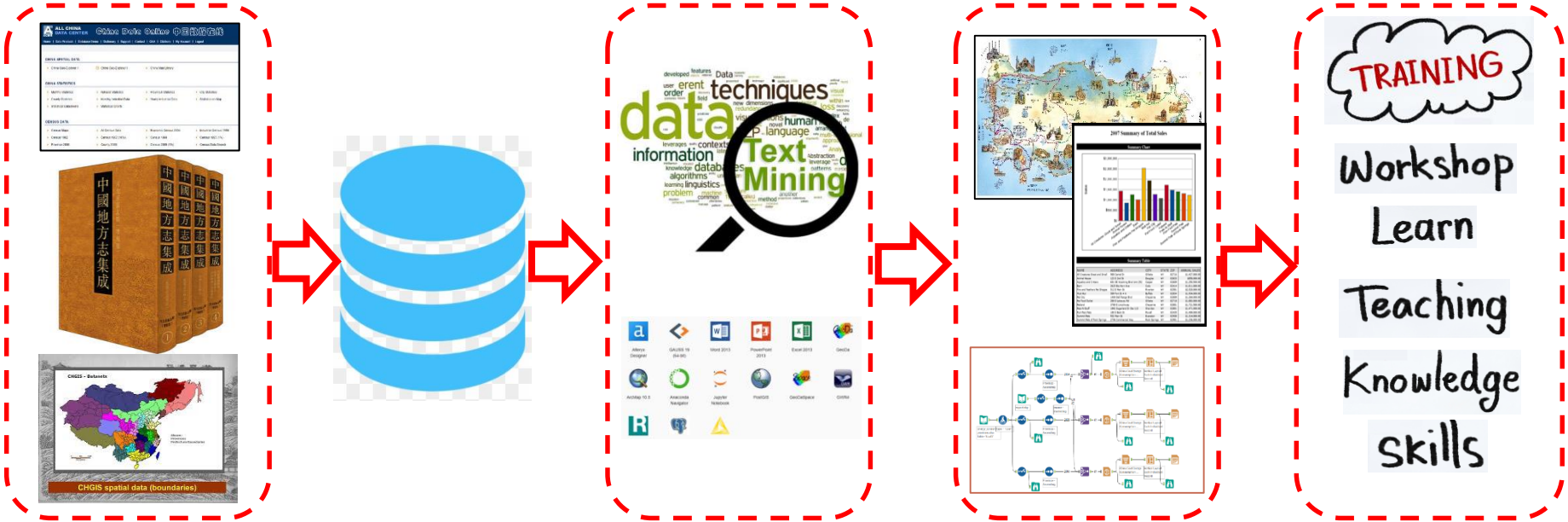
New Challenges for Libraries

- **How to facilitate the online sharing of research data (library, faculty and data vendors)**
  - Spatial & non-spatial data
  - Current & historical data
  - Numeric and text data
- **How to facilitate research and collaboration**
  - Shared data access restricted by research team members
  - Shared tools for online data analysis
- **How to facilitate training and education**
  - Replicable and reproducible data analysis
  - Quick learning for students with different backgrounds

China Data Lab for Digital Humanities

- **Build a connection between library, faculty and students**

  Build a joint team, including librarians, faculty and students, with the infrastructural and technical support from the CDL and libraries, for research and training on digital humanity and social sciences.



CDL for Digital Humanities: Global Partners

- Yenching Library (US)
- Center for Geographical Analysis (US)
- Max-Planck Institute (German)
- Wuhan University (China)
- Zhejiang University (China)
- ….

Related Web Sites

China Data Lab

http://chinadatalab.net

China Data Lab on the Cloud

http://chinadatalab.cn

Contacts:

- Dr. Wendy Guan
- wguan@cga.harvard.edu
- (617) 496-6102

- Dr. Shuming Bao
- office@chinadatacenter.net
- (734) 274-2819