

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



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



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**.



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.



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



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



China Data Lab

<http://chinadatalab.net>



HOME

People

Resources

Partners

Events

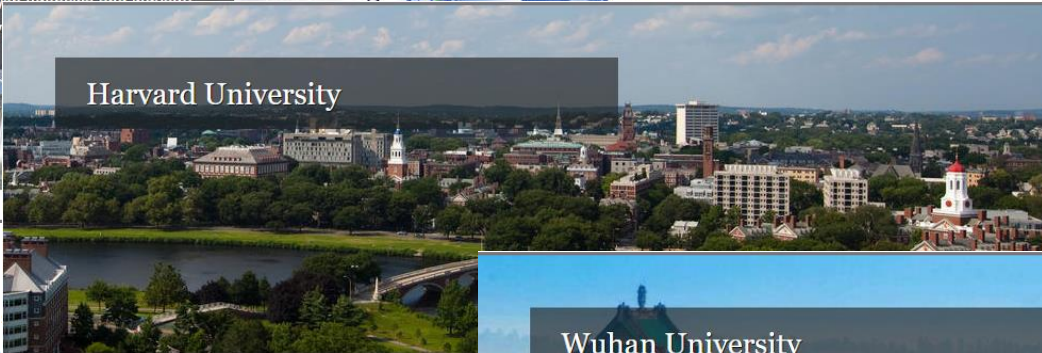
About

Spatiotemporal Thinking, Computing, and Applications (STC)

The center is dedicated to collaborate with agencies and industry to conduct leading spatiotemporal innovation.



Harvard University



LAB NEW

China Data Lab

Friday, May 10, 2019

NSF I/UCRC Spatiotemporal Innovation Center Held

Friday, June 21, 2019

Wuhan University



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

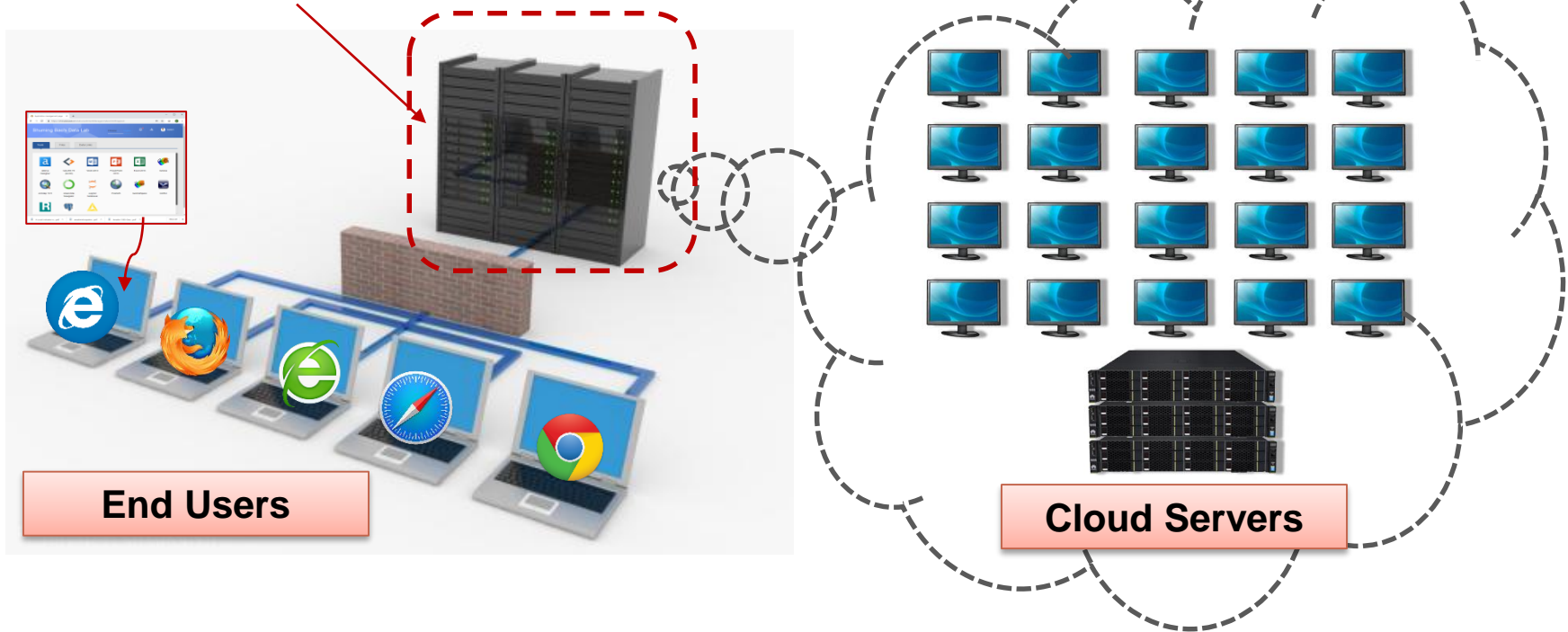
Outline

- ① About China Data Lab (CDL)
- ② **CDL Platform**
- ③ Data Management
- ④ Plans

CDL Platform Framework



China Data Lab Platform



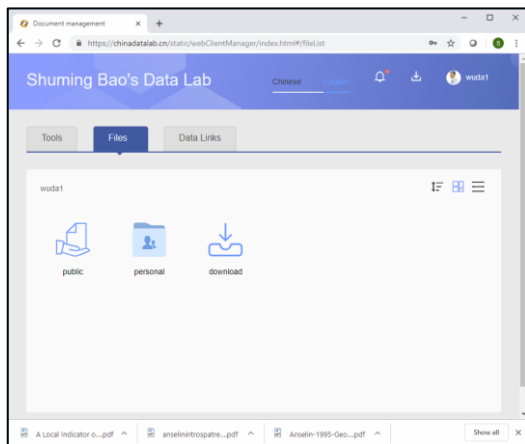
End Users

Cloud Servers

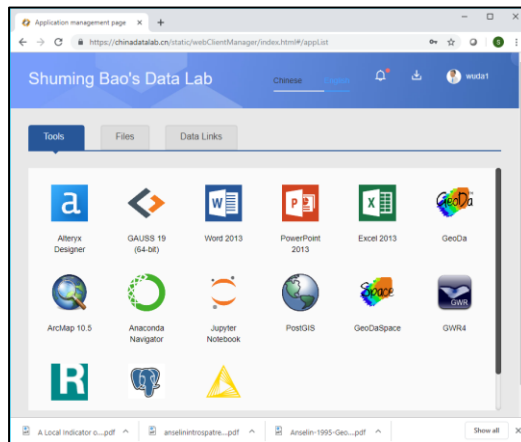
Cloud Platform Features

- ❑ 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

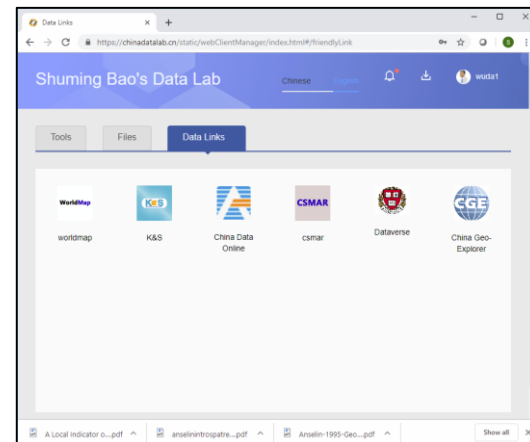
Personal & Shared Data



Tools

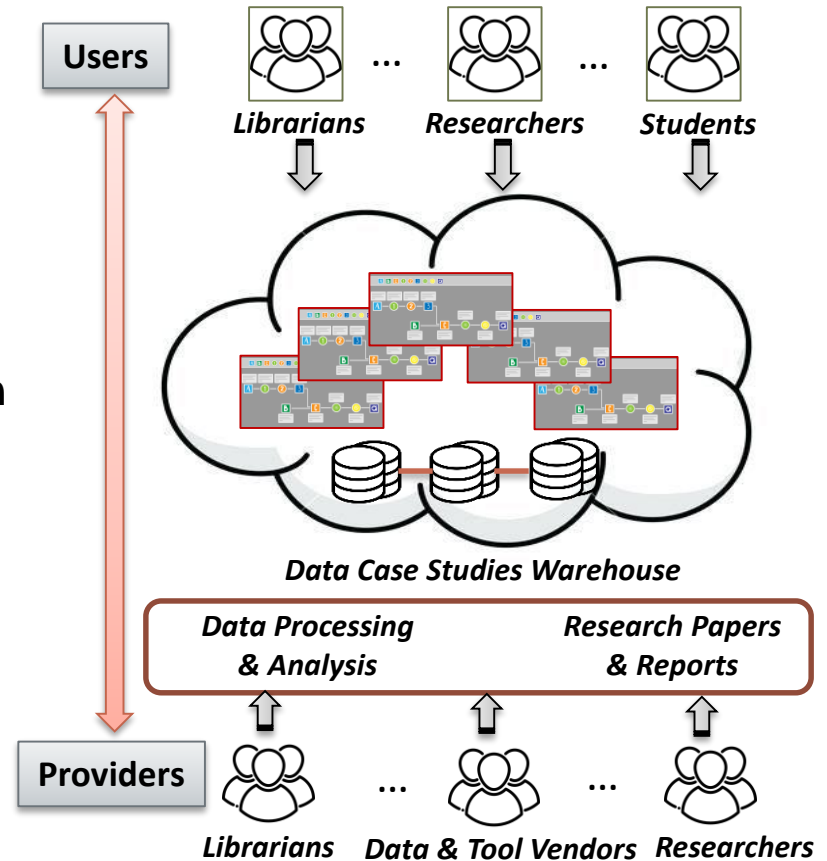


Links



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)



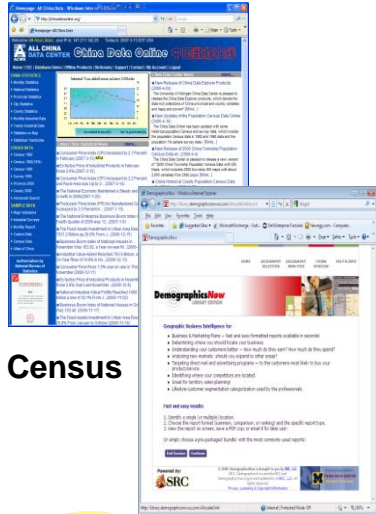
Authentic, Unique, Comprehensive, & Curated China Data



- **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



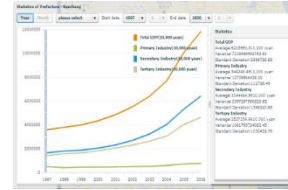
GIS

Data



Output

Charts

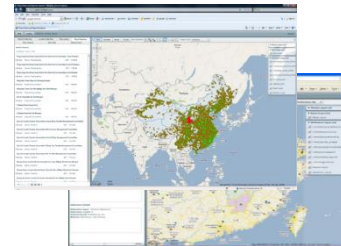


Tables

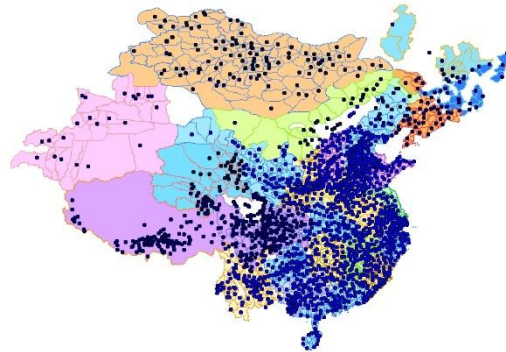
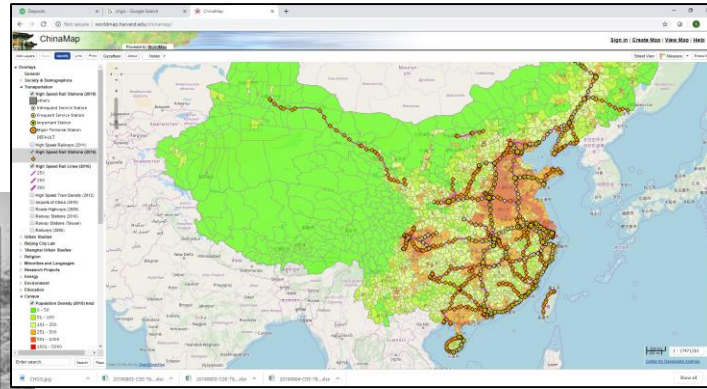
This figure shows a screenshot of a 'Demographic Summary Report' table. The table has columns for 'Year', 'Total', 'Male', and 'Female'. The data is presented in a grid format with alternating row colors. Below the table, there is a bar chart showing the distribution of the population by gender (Male and Female) across different age groups (0-14, 15-64, 65+).

Reports

Maps

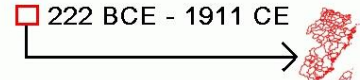


CHGIS: GIS Data for Historical Studies of China

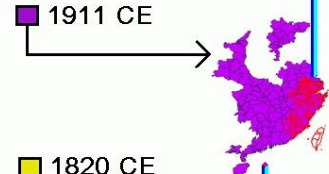


CHGIS Version 2 (Oct 2003)

222 BCE - 1911 CE



1911 CE

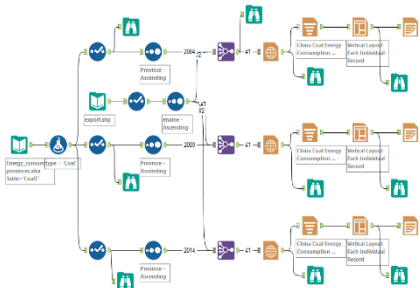
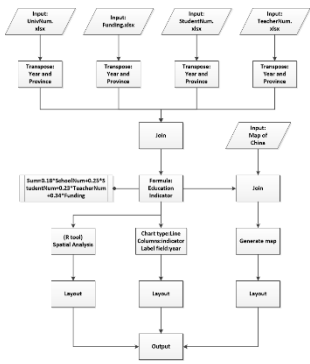


1820 CE

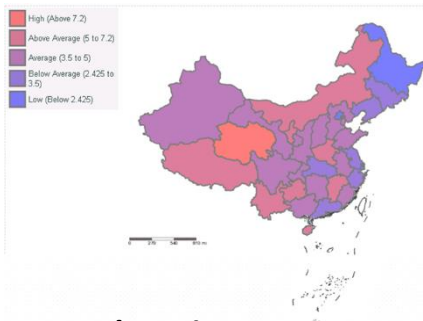


CDL for Reproducible, Replicable, Generalizable Research

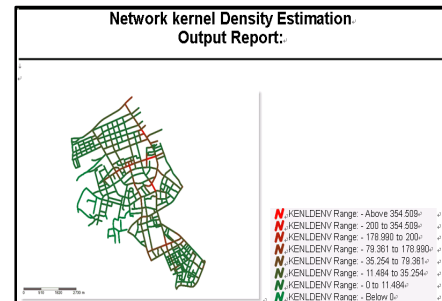
Workflow Data Analysis



Environment

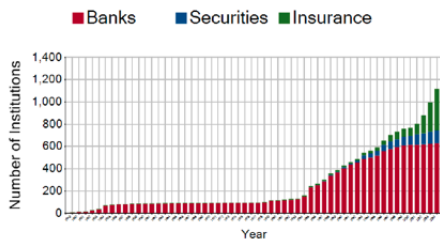


Education

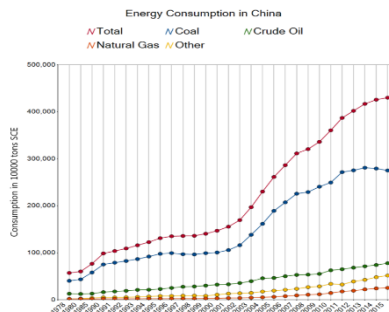


Transportation

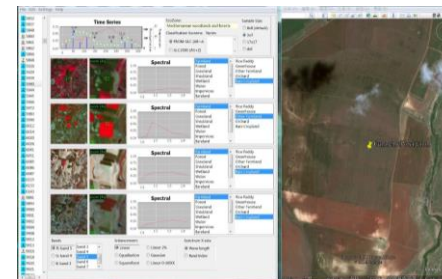
Total Numbers of Financial Institutions in Guangdong (1949 - 2004)



Economics



Energy



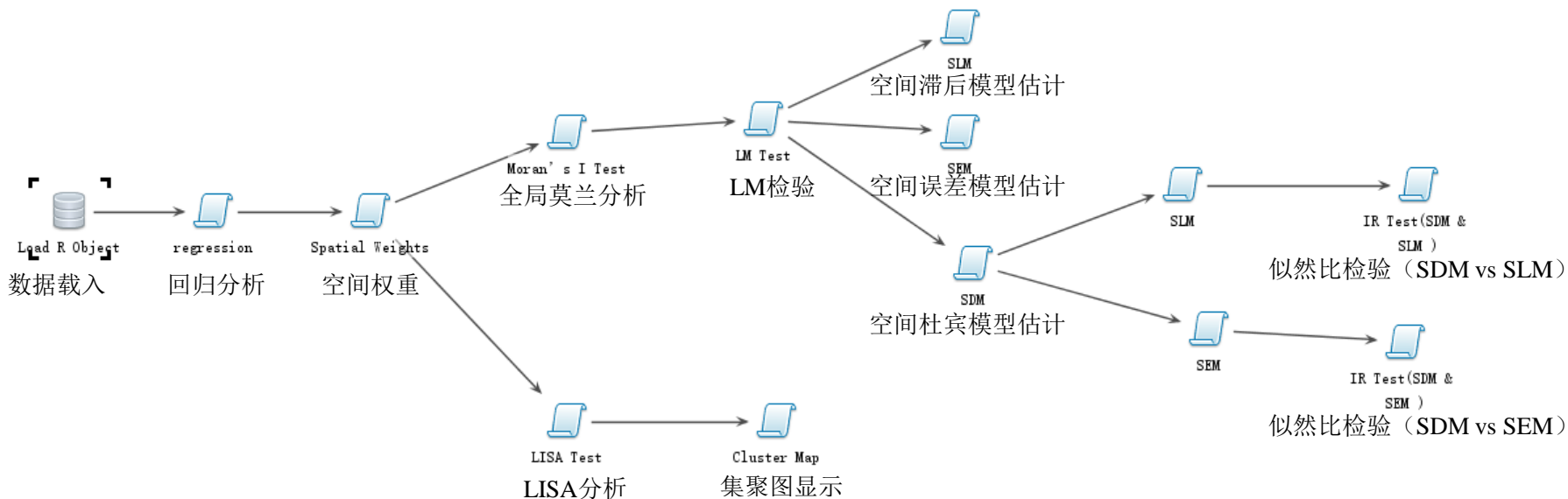
Land Use

Workflow Data Analysis with Alteryx

The screenshot displays the Alteryx software interface. The central workspace shows a workflow titled "MergeFuzzyMatch.YXMD". A text box in the center reads: "Sample Alteryx Module: Classic Fuzzy Matching Merging a Customer Database". Below this, an "Introduction" box states: "Take two existing Customer Database Files and use them through this module to combine using records and remove duplicate records based on fuzzy matching methodologies." A "TOOLS USED" box lists: "Input Record ID", "Fuzzy Matching", "Union", "Multi-Group", "Linkage", "Join", and "Browse Data". The workflow diagram includes several tool icons connected by arrows, with callouts explaining steps like "Use the Input tool to bring in the customer database that you need for Merge, Union, and New Data File.", "Assign a unique record ID per record. Record IDs need to have different lengths.", "File to match to ID that matches it in a master file and use duplicate filter to remove duplicates.", "Close all the records are merged together. It's a standard fuzzy match process, but it's not a merge process, it's a fuzzy match.", "This will filter up to 100 different types of records that are merged.", "Use the Union tool to support both matched results from one side stream.", "Linkage on Master ID to only get one match from one side. Linkage will be necessary to filter. Check off the necessary tool to include from the master side. This tool is necessary to ensure that we get one record to match to the other side.", "This is the final output report. It's a standard fuzzy match process, but it's not a merge process, it's a fuzzy match.", "Filter the Master ID to ensure that we get one record to match to the other side.", "Filter the Master ID to ensure that we get one record to match to the other side." The right-hand pane shows the "Properties - Module" window with fields for "Module Path", "Record Limit", "Show Annotations", "Include Tool Name in Annotation", "Module Type", "Normal Module", "Limit" (set to 10), and "Stop Processing When Limit is Reached".

This section is a collage of Alteryx interface elements. At the top, there are three overlapping "SRCAlteryx Beta Wizard - Kiddie_Kandid" windows. The first shows date selection for "Start Date" (March, 2008) and "End Date" (April, 2008). The second shows "Please Select Children's Age" with a dropdown menu. The third shows "Please Select Store" with a dropdown menu. Below these are two data visualization outputs. The first is a "2007 Summary of Total Sales" for "PetStores, Inc." featuring a bar chart and a "Summary Table" with columns: NAME, ADDRESS, CITY, STATE, ZIP, and AVE. The second is "Pins and Feathers Pet Shoppe: 2007 Sales Figures" featuring a bar chart and a map showing store locations. A third visualization, "Mutt Hut: 2007 Sales Figures", also features a bar chart and a map.

Workflow Data Analysis with R AnalyticFlow



```

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

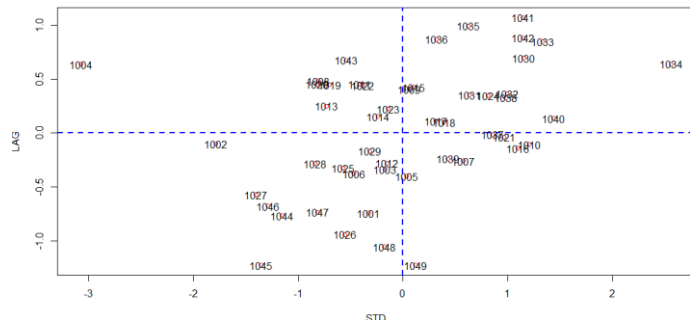
Residuals:
    Min       1Q   Median       3Q      Max
-34.418 -6.388 -1.580  9.052 28.649

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
            Hi      E(Li)   Var(Li)   Z(Li)  Pr(z > |Li|)
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

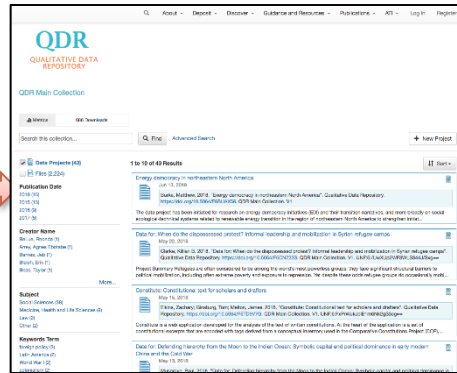
The Synergy of DataVerse & WorldMap & CDL

Integration, Searching, Sharing, Visualization, Execution

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

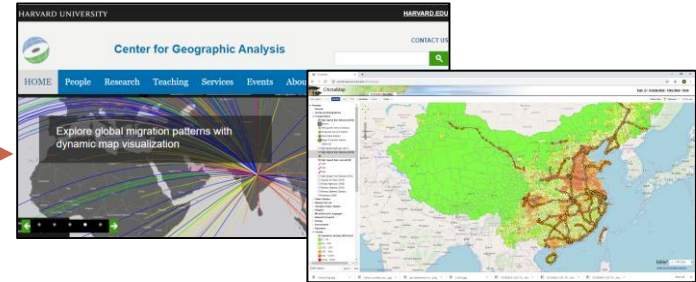
DataVerse

Data Management



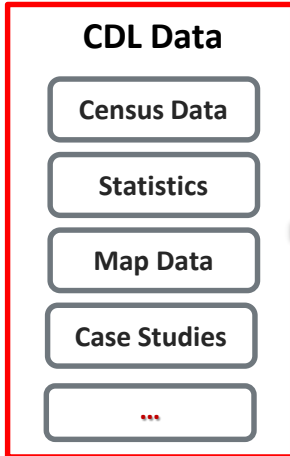
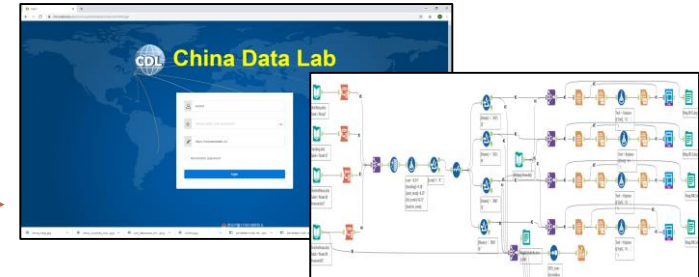
WorldMap

Data Visualization



China Data Lab

Workflow Data Analysis



DataVerse: An Open Tool for Data and Metadata Management



Open source research data repository software



Researchers

Enjoy full control over your data. Receive *web visibility*, *academic credit*, and *increased citation counts*. A personal dataverse is easy to set up, allows you to display your data on your personal website, can be branded uniquely as your research program, makes your data more discoverable in the research community, and satisfies data management plans. [Want to set up your personal dataverse?](#)



Journals

Seamlessly manage the submission, review, and publication of data associated with published 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 policies and of repository recommendations. [Want to find out more about journal dataverses?](#)



Institutions

Establish a research data management solution for your community. Federate with a growing network of Dataverse repositories worldwide for increased discoverability of your community's data. Participate in the drive to set norms for sharing, preserving, citing, exploring, and analyzing research data. [Want to install a Dataverse repository?](#)



Developers

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

Data Citation with Persistent Identifier (DOI)

Data Files

Metadata

Data Licenses, User Agreements

Dataset Versions

Clinical Illness and Outcomes in Patients with Ebola in Sierra Leone Version 1.0

Schiffelin, John; Shaffer, Jeffrey; Goba, Augustine; Gballe, Michael; Gire, Stephen; Colubri, Andres; Swafflon, Rachel; Kamah, Larasana; Mogbo, Alex; Momoh, Mambo; Fula, Mohammed; Moses, Lina; Brown, Bethany; Anderson, Kristian; Wrinicki, Sarah; Schaffner, Stephen; Park, Daniel; Yozwiak, Nathan; Jiang, Pan-Pan; Karibo, David; Jalloh, Simbirie; Fonnle, Mbaku; Sinnah, Vandi; French, Isaac; Kovoma, Alico; Kamara, Fatima; Tucker, Veronica; Konuwa, Edwin; Seilu, Josephine; Mustapha, Ibrahim; Foday, Momoh; Yillah, Mohamed; Kamah, Franklyn; Saffa, Sidiki; Massally, James; Bolein, Matt; Branco, Luis; Vandi, Mohammed; Grant, Donald; Hippo, Christian; Gesso, Sahn; Fletcher, Thomas; Fowler, Robert; Baurich, Daniel; Saberi, Paride; Khan, Humar; Garry, Robert, 2015, 'Clinical Illness and Outcomes in Patients with Ebola in Sierra Leone', doi:10.7910/DVN/29296, Harvard Dataverse, V1, UNF:5:wNw/DjKH9aELNFuFm72w==

Description
This data comprises a total of 213 cases evaluated for Ebola virus infection at the Kenema Government Hospital in Sierra Leone between May 25 and June 10, 2014. Outcome data was available for 97 of 100 EBOV positive cases. Metabolic panels were performed on 98 Ebola virus disease and non-Ebola virus disease illness patients with adequate samples volumes. Ebola virus load was determined in 63 cases with adequate samples volumes by quantitative polymerase chain reaction (qPCR) at Harvard University. Sign and symptom data was obtained on 44 patients with a clinical chart that were admitted to Kenema Hospital. The metabolic panels were obtained from serum samples analyzed with a Piccolo Blood Chemistry Analyzer and Comprehensive Metabolic Reagent Discs (Abaxis).

Keyword
Ebola, clinical data, laboratory data, outbreak, fatality rate

Related Publication
Clinical Illness and Outcomes in Patients with Ebola in Sierra Leone. John S. Schiffelin, et al. N Engl J Med 2014; 371:2092-2100 November 27, 2014;DOI:10.1056/NEJMoA1411680 doi:10.1056/NEJMoA1411680

Files Metadata Terms Versions

Search this dataset... Find

3 Files

ebola-data.lib
Tabular Data - 148.2 KB - Mar 1, 2015 - 94 Downloads
215 Variables, 214 Observations - UNF:5:wNw/DjKH9aELNFuFm72w==
Mirador dataset converted into SPSS format
SPSS

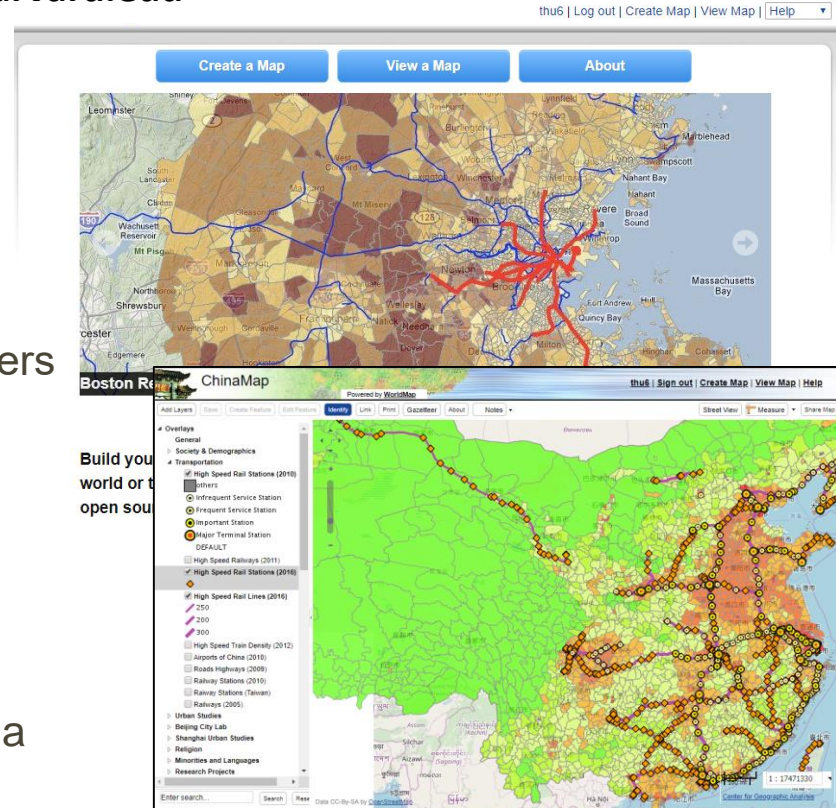
ebola-mirador.zip
ZIP Archive - 15.3 KB - Mar 1, 2015 - 27 Downloads
MDS: 673948ba9c9716b0c31031d221b7d04
The dataset is formatted as a project to load into Mirador. It is essentially a collection of CSV files, but with some extra information to make the navigation of the data easier and to display values such as dates and categories more conveniently.
Mirador

WorldMap for Spatial Data Depository & Visualization

<http://worldmap.harvard.edu>

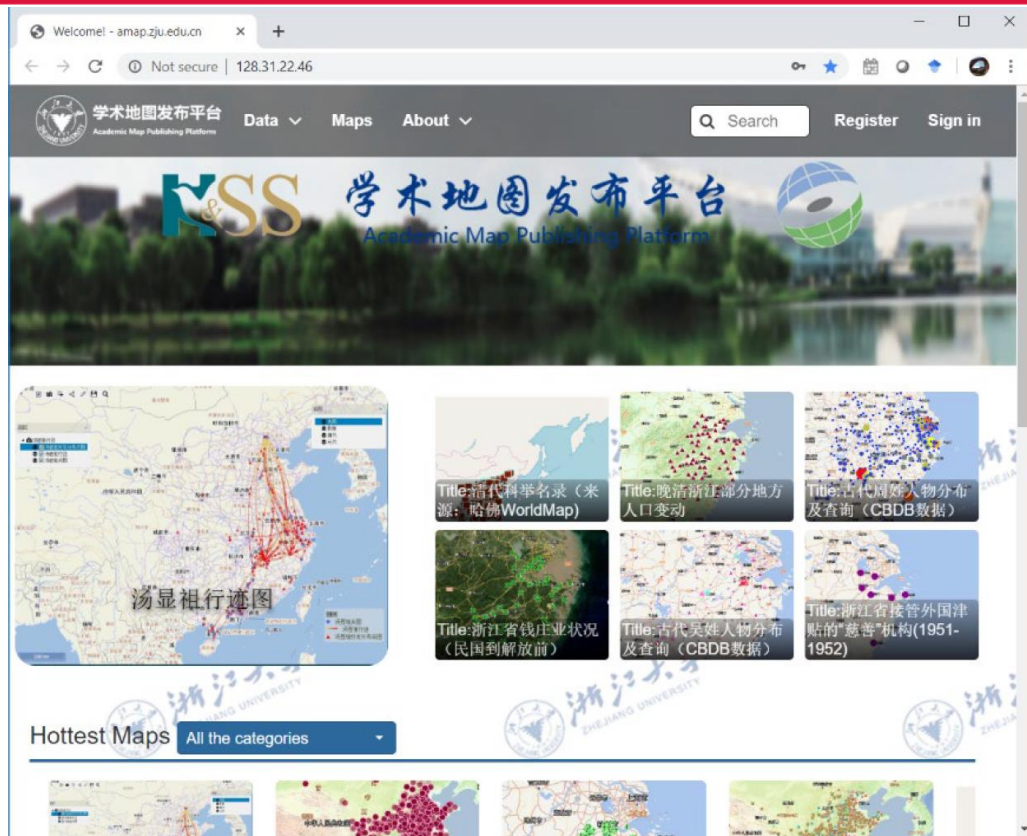
❑ Allow Users to

- ❖ **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

哈佛燕京学社校友 (HYI ALUMNI) 分布及查询 (截

Geographic Distribution of Harvard-Yenching Institute Alumnus



Info Share Ratings Comments Favorite

Title 哈佛燕京学社校友 (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/>

Japanese Map Warper

Home Browse All Maps Browse Rectified Maps Find Maps by Location Upload Map Browse All Mosaics Find Mosaics by Location Add Mosaic About Help

Overview

You can display your own maps against a real map (OpenStreetMap) for free, and make them available on the Web for anyone in the world. Maps such as hand-written or old ones can be used as well as modern standardized ones.

Uploaded maps are rectified (warped) against a real map by adding control points, which specify geographical positions.

Once published, these maps are available to export in image or map formats.

このサイトでは、自由に地図画像をアップすることが出来ます。その中には、既にインターネット公開されているものも含め日本の各地図が含まれます。その中には、発行当時の資料をそのままの形でデジタル化したものもあります。現代においては適当ではないと思われる表現を含む資料がある可能性もあります。その資料が成立した時を基とする史料資料として、ご閲覧・ご閲覧の上でご利用くださいますようお願いいたします。

UPLOADED MAPS

Last Rectified Maps

MAP	TITLE	YEAR	LAST UPDATED	STATUS
	寛政水戸船絵図 https://www.lib.pref.ibaraki.jp/guide/hh/you/digital_lib/wakuiki_m/001051194692/00lib/glyphy.html View Map Rectify Map Download KMZ Like Source: /lib/ol/er	1793	0 days ago	20 control points
	京都マツ明治 View Map Rectify Map Download KMZ	1907	0 days ago	3 control points
	江戸切絵図 (尾張版)・今川宣輪棧草絵図 江戸時代後期、江戸の浮城地区として行われたもの。版元は尾張屋七七。複製中周辺から北部を対象とした部分。©国立国会図書館デジタルコレクションより転載。インターネット公開 (著作権関係なく・Public Domain)。 http://dl.ndl.go.jp/in/fo/m/001/048/1286208 View Map Rectify Map Download KMZ Like Source: /lib/ol/er	1849	0 days ago	49 control points

SEE ALL MAPS

Recent Mosaics

MOSAIC	TITLE	YEAR	LAST UPDATED	NUMBER OF MAPS	PERCENTAGE COMPLETE
	test test: Compiled by 佐藤弘毅 Download KMZ		over 1 year ago	2 maps (2 maps)	100%
	0207 Compiled by ima. Download KMZ		over 1 year ago	10 maps (10 maps)	100%
	Nishimura 赤松市明徳園: Compiled by Marika Nishimura. Like Source: /lib/ol/er		over 1 year ago	0 maps (0 maps)	0%

VIEW ALL MOSAICS
CREATE A NEW MOSAIC

Japanese Map Warper

LOGIN CREATE ACCOUNT ENGLISH (EN)

Search: Title Year SEARCH All maps Rectified maps only

BROWSE ALL MAPS

FIND RECTIFIED MAPS BY LOCATION

日本版 Map Warper

LOGIN アカウント作成 日本語 (JA)

ホーム すべての地図を見る 地図をアップロードする すべてのモザイクを見る

ホーム > 地図

整形済みの地図を探す。

地図を動かしたり、拡大してみてください - 表示領域に応じた地図とリストが表示されます。

1527 年

場所を探す

見つかった 1547 地図。表示中 1 -

20

- [天保地誌(改正日本国) 1837 地図を禁く
- 六日軍軍金田 1883 地図を禁く
- 【越川】帝國軍 地図を禁く
- 【名野】帝國軍 地図を禁く
- 【筑前】二十万分之一地勢圖 1930 地図を禁く
- 【羽博】二十万分之一地勢圖 1930 地図を禁く
- 【天保】二十万分之一地勢圖 1930 地図を禁く

Japanese Map Warper

HOME 全ての地図を見る 地図をアップロードする 全てのモザイクを見る

ホーム > 地図

整形済みの地図を探す。

地図を動かしたり、拡大してみてください - 表示領域に応じた地図とリストが表示されます。

1527 年

場所を探す

見つかった 1547 地図。表示中 1 -

20

- [天保地誌(改正日本国) 1837 地図を禁く
- 六日軍軍金田 1883 地図を禁く
- 【越川】帝國軍 地図を禁く
- 【名野】帝國軍 地図を禁く
- 【筑前】二十万分之一地勢圖 1930 地図を禁く
- 【羽博】二十万分之一地勢圖 1930 地図を禁く
- 【天保】二十万分之一地勢圖 1930 地図を禁く

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 humanities 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