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:

Data, Tools, and Applications

http://chinadatalab.org

http://chinadatalab.net

Outline

- Challenges and Motivations
- Platform, Data and Tools
- Data Access Demo
- Workflow-based Case Studies
- User Account Application

About China Data Lab

❑ 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 · GAUSS · Knowledge@Share

China Data Lab

https://projects.iq.harvard.edu/chinadatalab

HARVARD UNIVERSITY

HARVARD.EDU

http://chinadatalab.net

China Data Lab

HOME | People | Resources | Partners | Events | About

LAB NEWS

China Data Lab (CDL) Established
Friday, May 17, 2019

NSF I/UCRC Spatiotemporal Innovation Center Held the 9th Semi-Annual Industrial Advisory Board Meeting
Friday, June 21, 2019

More ▸

UPCOMING EVENTS

2019 JUL 09
Training Workshop on "Spatial Data Lab"
(All day)

2019 JUL 07
2019 International Workshop on Geocomputation for Social Sciences (The 3rd Call for Papers)
Sun Jul 7 (All day) to Tue Jul 9 (All day)

Challenges

❑ Data Sharing

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

❑ Tool Sharing

- Licensed and free tools
- Integrated environment for tools for data
- Maintenance and updates

❑ Results Sharing

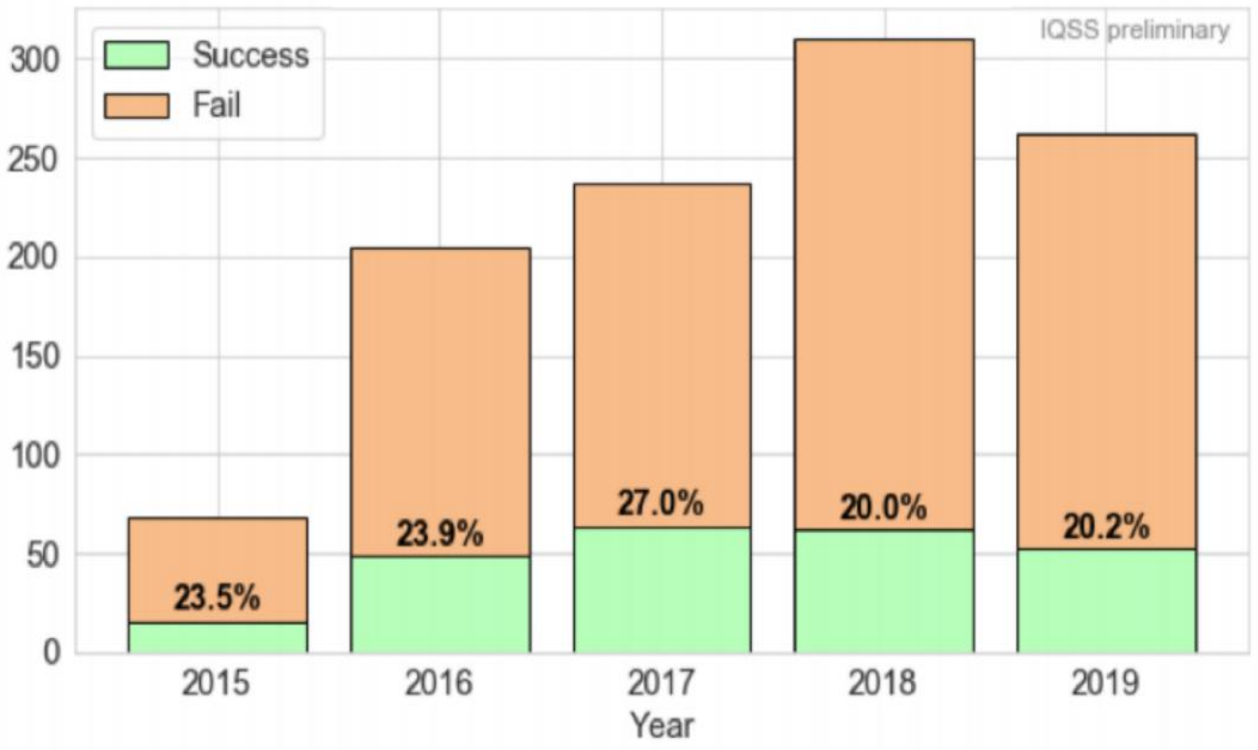
- Research (**reproducible, replicable, and generalizable**)
- Teaching (students with different interests and skills)

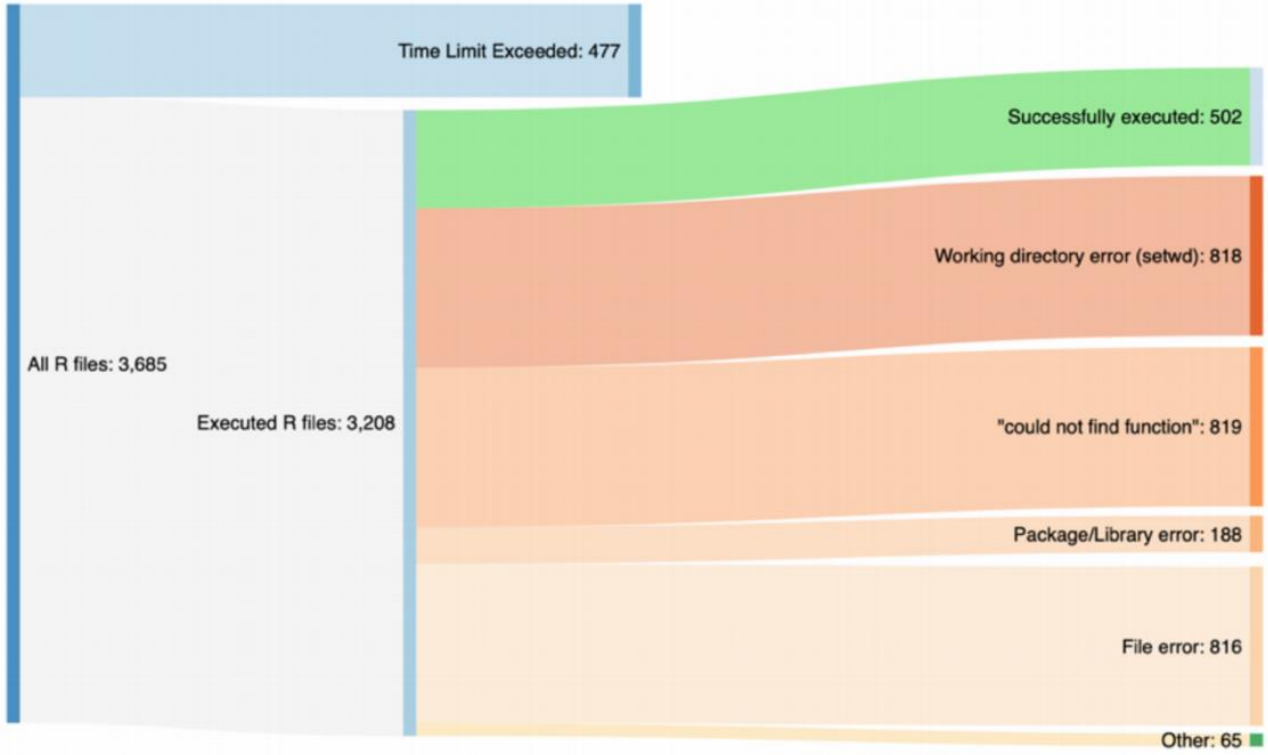
Challenges

The Dataverse® Project

Re-execution of R code in datasets published on Harvard Dataverse

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





source: https://grmds.s3.us-east-2.amazonaws.com/2019-12/Danny%20Liu%20%26%20Alex%20Liu%20Keynote%20.pdf

Goal and Objectives

Goal: to build ECOSYSTEM for reproducible, replicable, and expandable research

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 theory, methodology, technology, data and workflow case studies for research and teaching

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

Resources

http://chinadatalab.net





GAUSS

GeoDa

GWR4

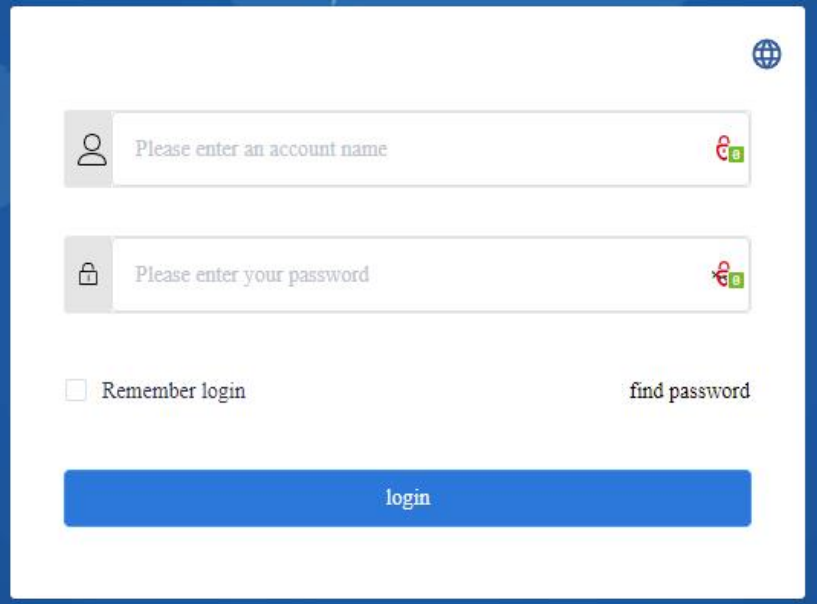

ArcMap


Jupyter


RAnalyticFlow

KNIMEAnalytics

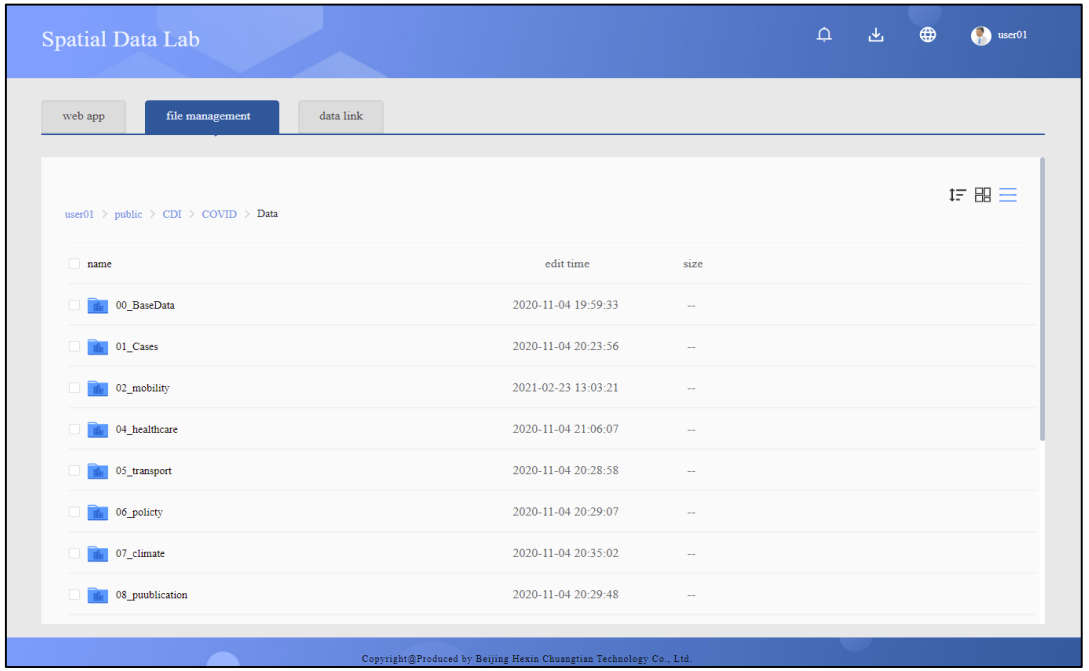
Spatial Data Lab



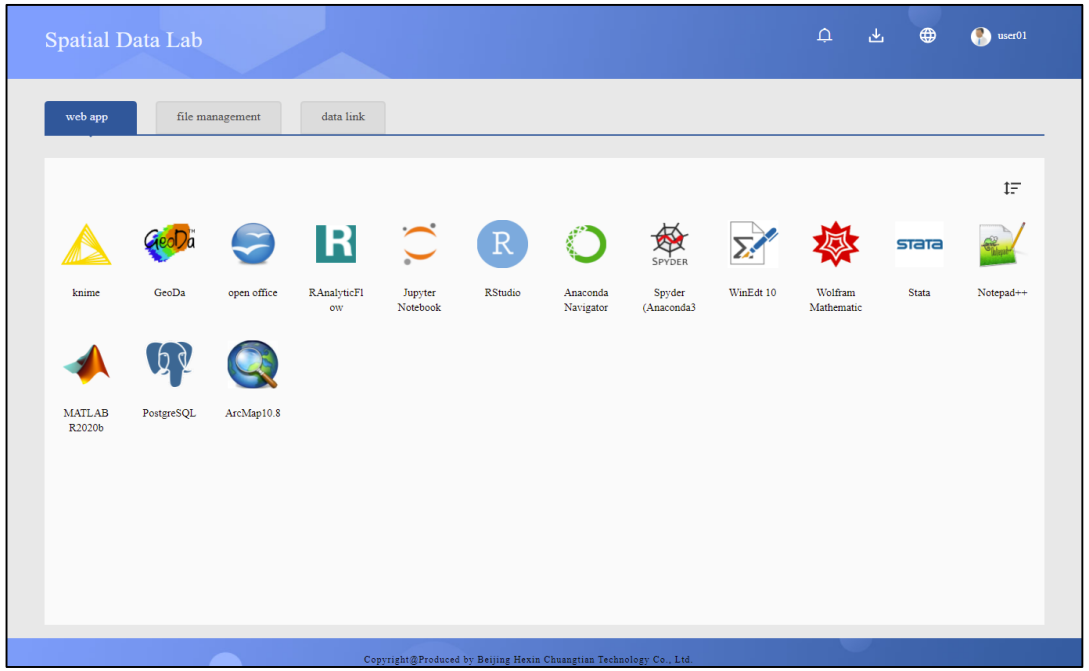
http://chinadatalab.org

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

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

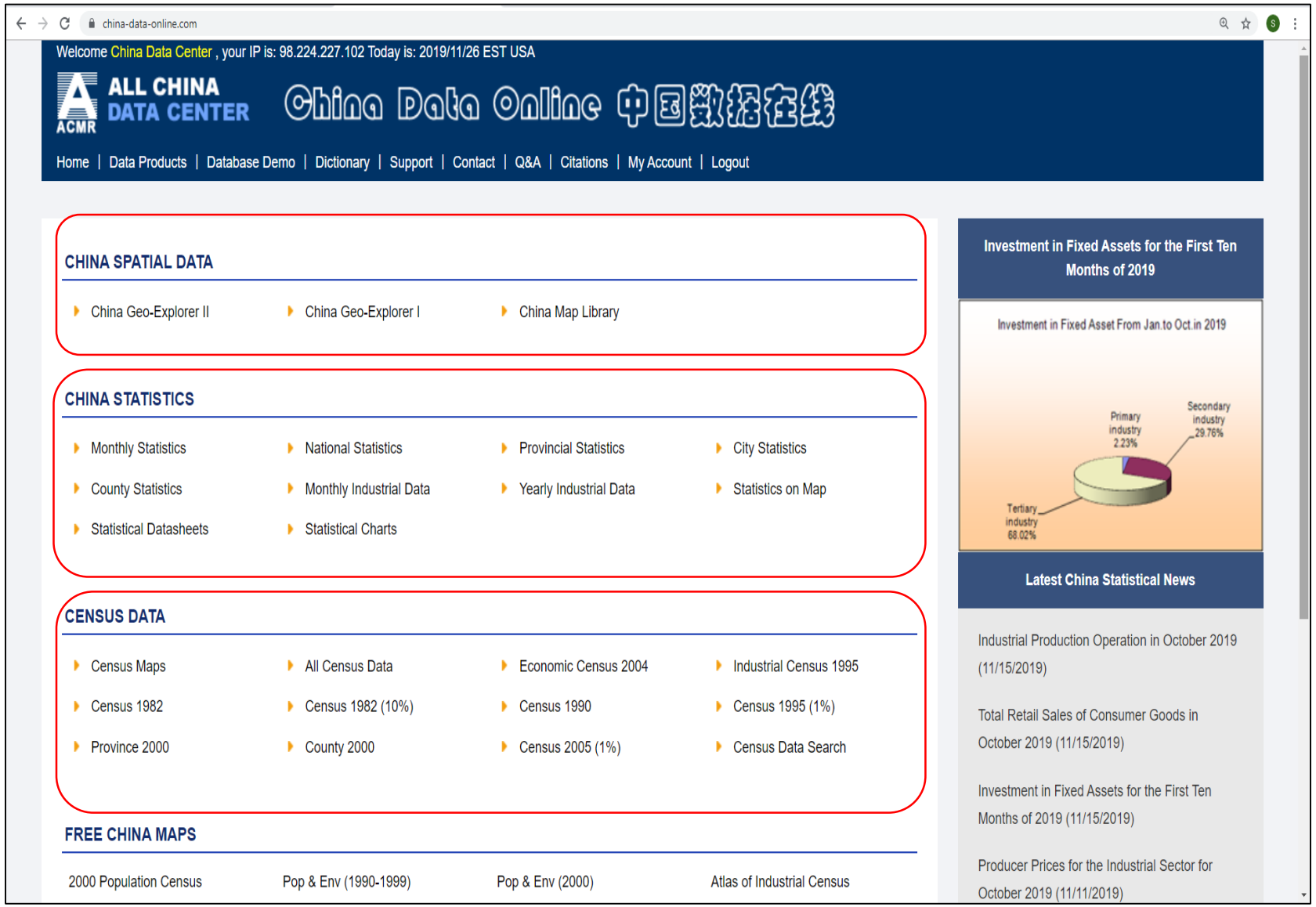
Authentic, Unique, Comprehensive, & Curated China Data

Statistical Database:

- Monthly Statistics
- National Statistics
- Provincial Statistics
- City Statistics
- County Statistics
- Monthly Industrial Data
- Yearly Industrial Data
- Statistics on Map
- Statistical Yearbooks

Census Database:

- Population Census 1982
- Population Census 1990
- Population Survey 1995, 2005
- Province Census 2000
- County Census 2000
- Economic Census 2004



Authentic, Unique, Comprehensive, & Curated China Data

Monthly Training Webinars on Research Data: Sources, Tools and Applications

1. September 24, 2020, **Understanding the Gov't Statistics**
2. October 22, 2020, **Understanding the Population Census and Demographic Changes**
3. November 19, 2020, **Understanding the Economic Census and Industrial Changes**
4. December 17, 2020, **Understanding the Administrative Maps and Regional Geography**
5. January 21, 2021, **Spatial Study of Innovation with Patent Data**
6. February 18, 2021, **Spatial Study of Health with Statistics, Census and GIS Data**
7. March 18, 2021, **Spatial Study of Environment with Statistics, Census and GIS Data**
8. April 22, 2021, **Understanding the Urban Development**
9. May 20, 2021, **Understanding the Rural Development**
10. June 17, 2021, **Understanding the Mobility in Geography**
11. July 15, 2021, **Understanding the Culture in Geography**
12. August 12, 2021, **New Development and Directions: Data, Tools and Applications**

Registration

IncoPat Patent Data

incoPat | Global Patent Database | 可信好用的全球专利数据库

Patent	Patent Family	Full-text Translation in Chinese
137,941,704	75,090,774	36,557,787

Country	Onsite Service for Chinese Cities	Users	User Renewal
120	26	40000	99%

The Grand Champion of "Competition of Intellectual Property Tools"

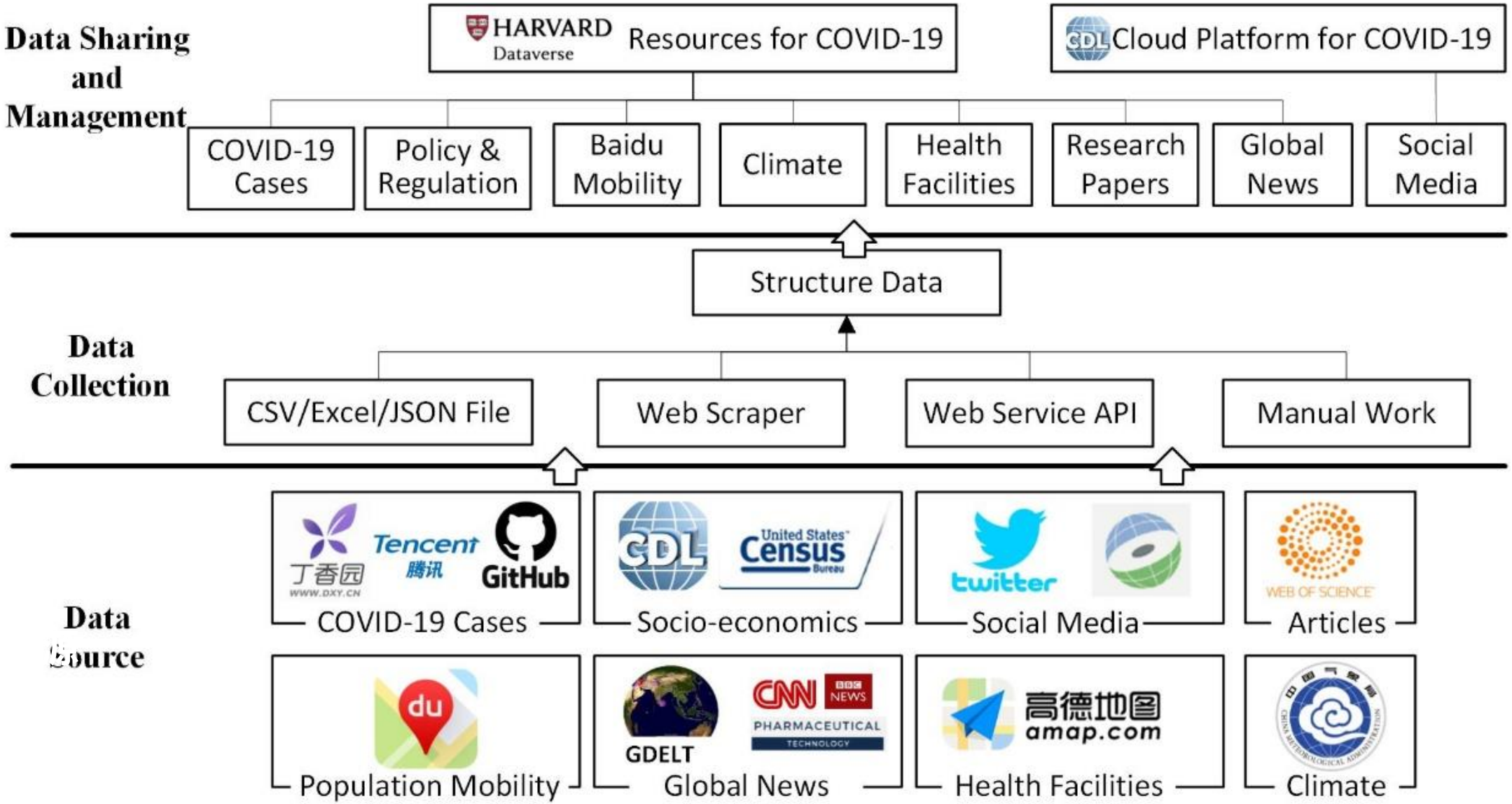
"Mobile Intelligent Terminal 2017(METIS) Award"

"A Leading Enterprise in IP SaaS Service Industry of The Year Award"

IncoPat China Patent Data

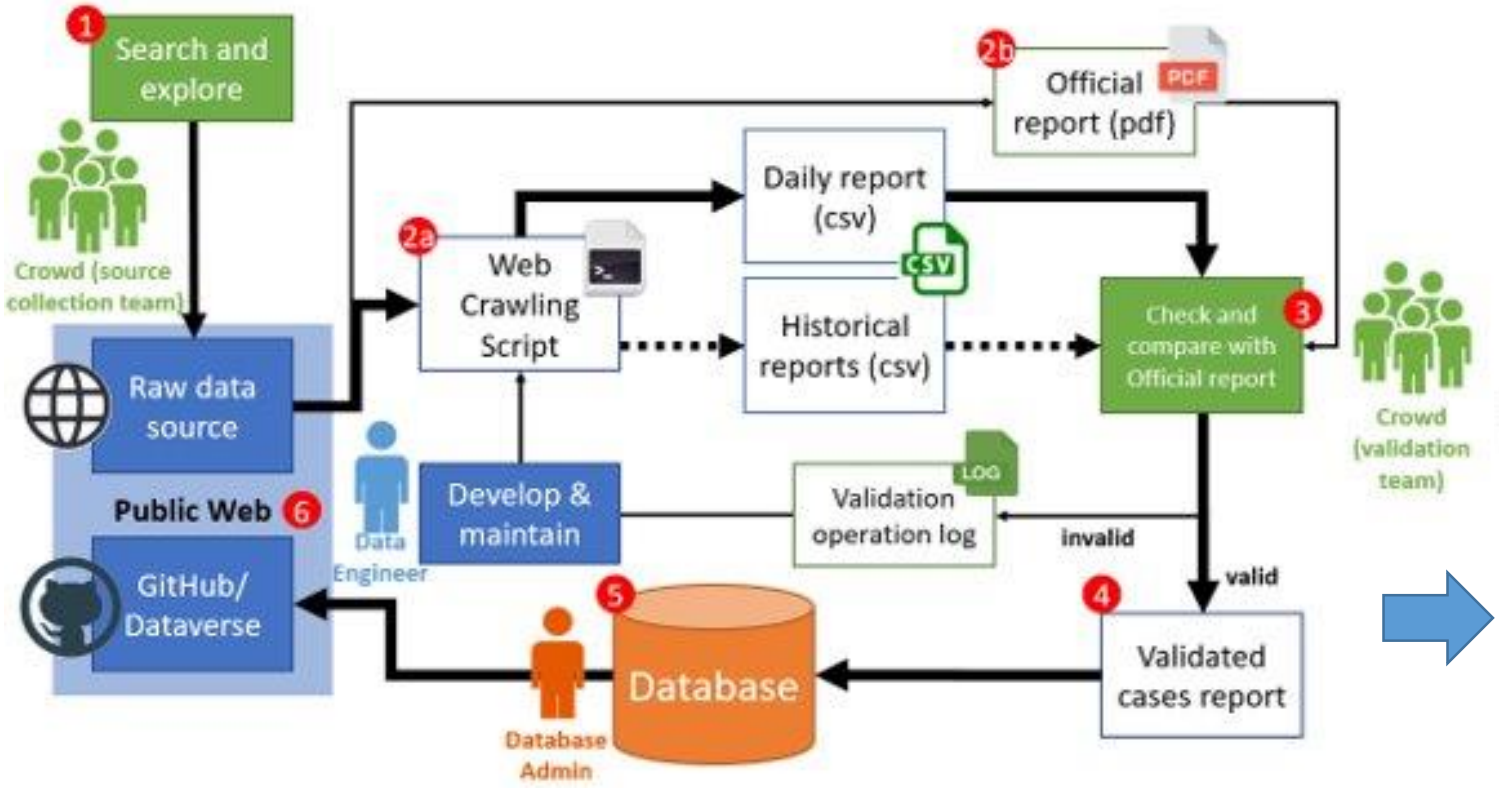
字段名	Field	字段名	Field	字段名	Field
标题（O）	Tio	公开（公告）日	Publication Date	引证类别(backward)	Citation Origin Code
摘要（O）	Abo	转让执行日	Assignment ExecutionDate	X引证文献	ct-X
专利价值度	Vlstar	专利有效性	status	家族引证次数(backward)	Family Citation Number of Times
申请人(原始)	AP-OR	当前法律状态	CN status lite	家族被引证次数（forward）	Family Citation-Forward Number of Times
发明(设计)人(原始)	in-or	法律状态文字信息	Legal Free Text	非专利引证	Citation-Nonpatent
代理人姓名	Attorney	引证信息（backward）	Citation	优先权日	Priority Date
代理机构	Agency	被引证信息（forward）	Citation-Forward	申请人省市	Applicant province
IPC分类	IPC	引证次数（backward）	Citation Number of Times	中国省市代码	PC-CN
IPC主分类	IPC Main	被引证次数 （forward）	Citation-Forward Number of Times	中国申请人省市代码	AP-PC
洛迦诺分类	Locarno Classification	家族引证（backward）	Family Citation	中国申请人地市	City
CPC分类	CPC	家族被引证（forward）	Family Citation-Forward	中国申请人区县	County
申请人地址	Applicant Address Information	引证申请人（backward）	Citation Applicant	受让人	Assignee
申请号	Application Number	被引证申请人（forward）	Citation-Forward Applicant	转让人	Assignor
优先权号	Priority Number	家族引证申请人(backward)	Family Citation Applicant	行业门类	Bclas1
公开类型	Publication Type	家族被引证申请人（forward）	Family Citation-Forward Applicant	行业大类	Bclas2
简单同族	MainFamily	引证号码(backward)	Citation Number	行业中类	Bclas3
扩展同族	CompleteFamily	被引证号码（forward）	Citation-Forward Number	公开(公告)号	Publication Number
同族国家	Fa-Country	引证国别(backward)	Citation Authority	授权公告日	Grant Date
申请日	Application Date	被引证国别（forward）	Citation-Forward Authority	首项权利要求	First Claim

COVID-19 Data Sources



COVID-19 Data Sources

Data collecting, processing, validating and sharing



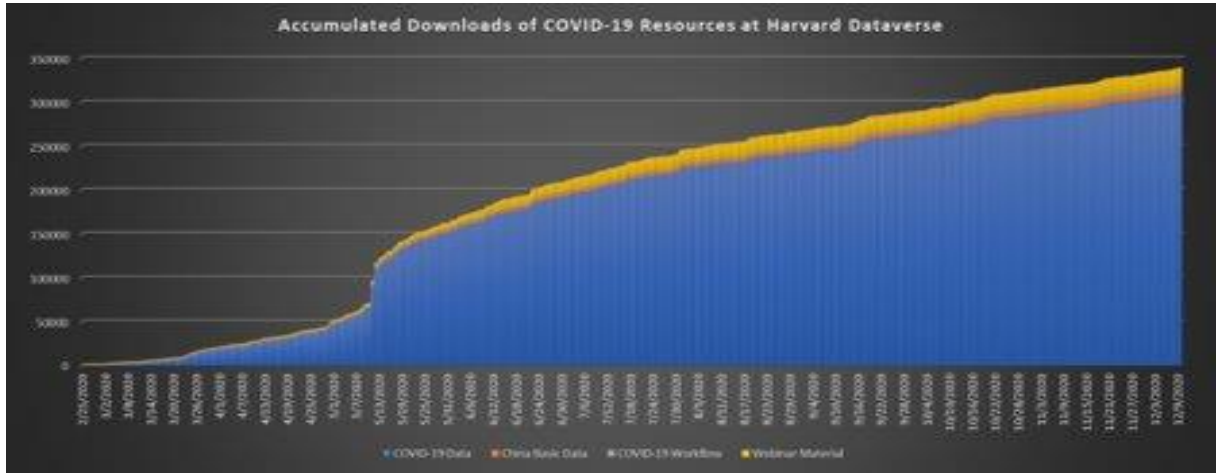
Map-based Data and standard association

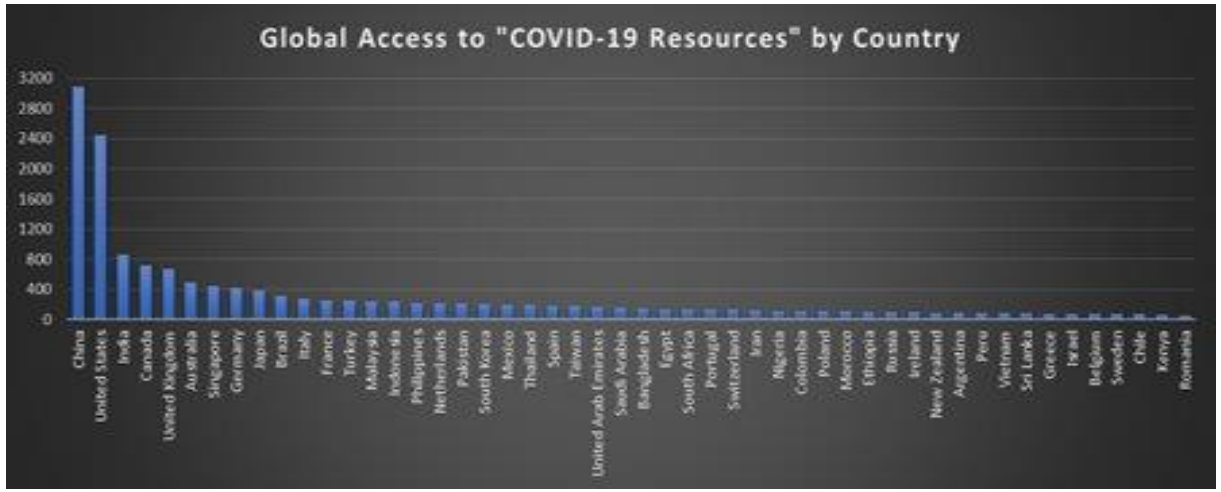
COVID-19 Datasets

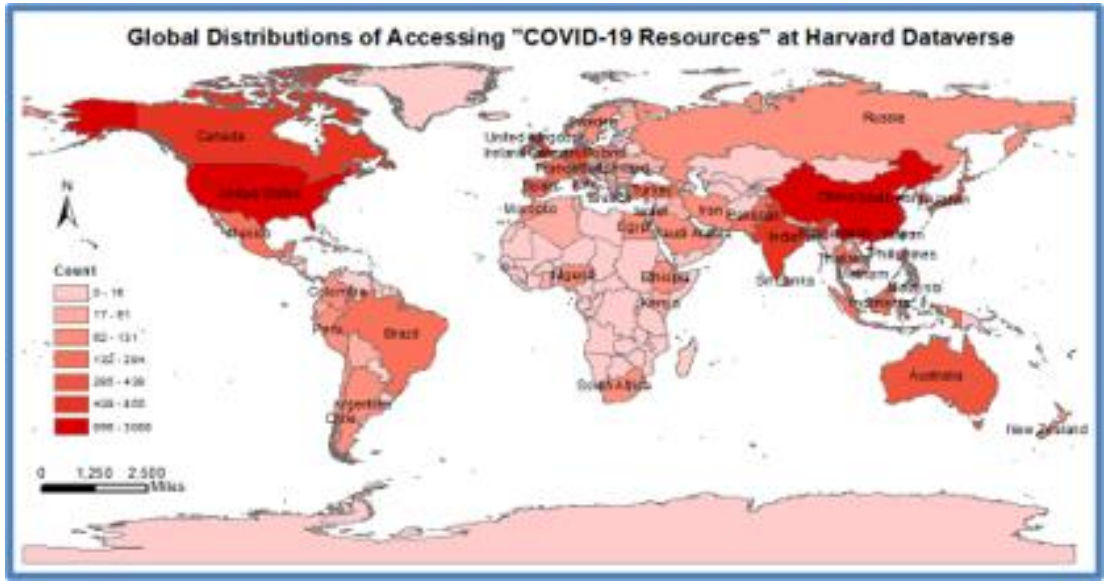
序号	数据项	Data Sets	Availability
1	全球疫情数据	Coronavirus cases data	Harvard Dataverse
2	人口流动数据	Population mobility data	Harvard Dataverse
3	医疗机构数据	Health facilities data	Harvard Dataverse/SDL
4	行为轨迹数据	Trace data	SDL
5	航线航班数据	Flight data	SDL
6	高铁班次数据	High-speed train data	SDL
7	全球新闻数据	Global News data	SDL
8	社交媒体数据	Social media data	SDL/CGA
9	政策数据	Policy Data	Harvard Dataverse
10	气象气候数据	Meteorological data	Harvard Dataverse
11	社会经济数据	Socioeconomic Data	Harvard Dataverse
12	疫苗数据	Vaccine Data	Harvard Dataverse

COVID-19 Datasets

Global users from **150+** countries downloaded COVID-19 datasets, publications, source code for over **400,000** times as of Mar. 1, 2021







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.

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

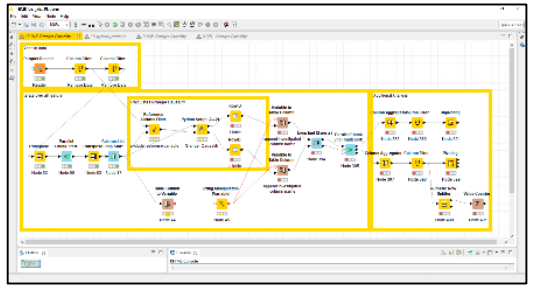
Rstudio

Anaconda Navigator

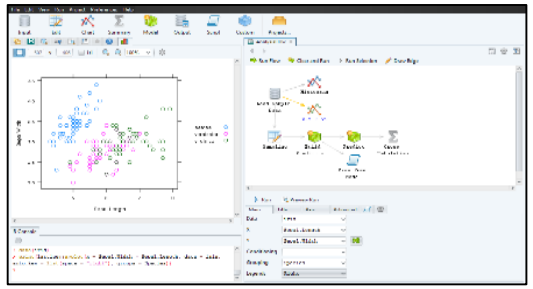
Spyder

❑ Workflow tools

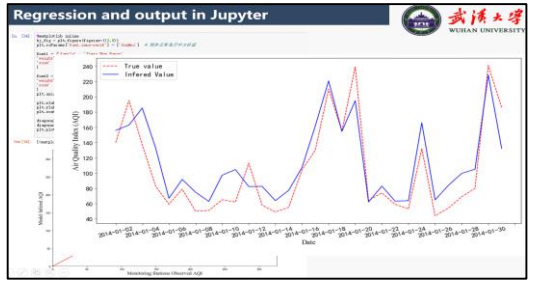
Open for Innovation KNIME



AnalyticFlow



Jupyter Notebook

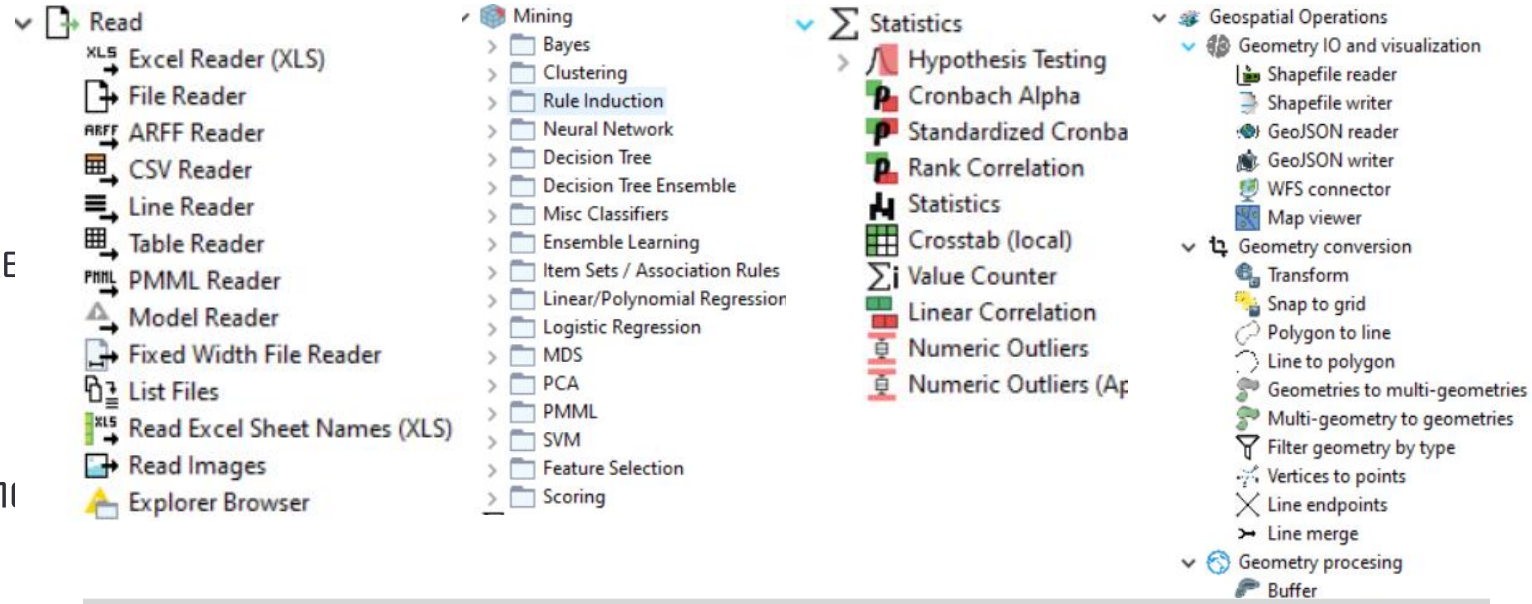


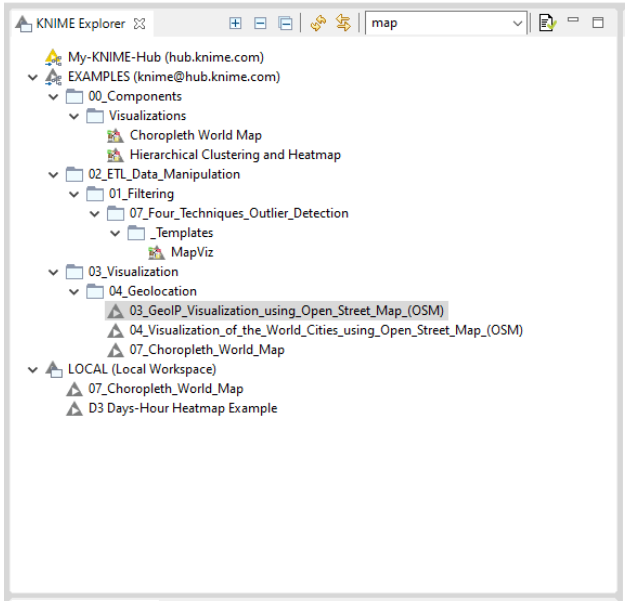
Tools on the Cloud

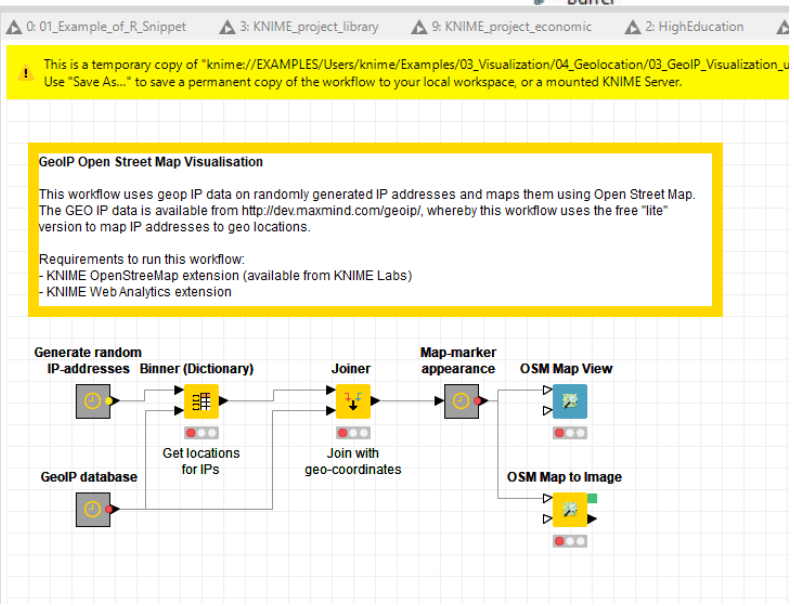
Open for Innovation
KNIME

https://www.knime.com/

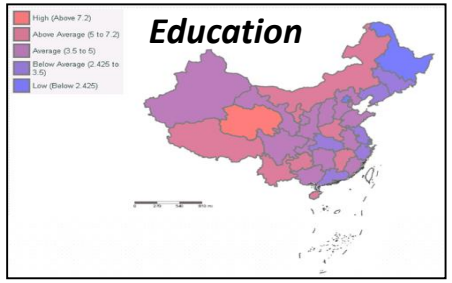
- Workflow is displayed as connected nodes which make it easy to troubleshoot and visualize
- Easy to use without much knowledge of coding
- Great extensions for data preprocessing, analysis, and visualization
- Connection to other languages, such as JS, R, Python, etc.
- Open-source
- Cross platform interoperability
- Has a decent size community that supports Q&A.

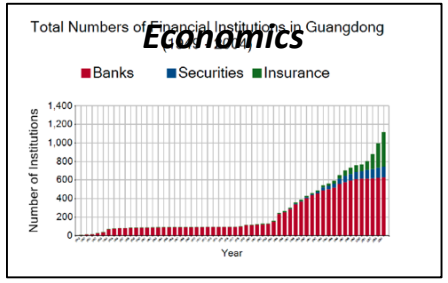


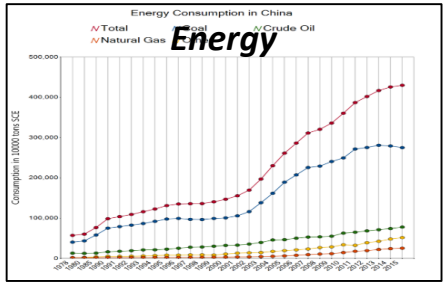


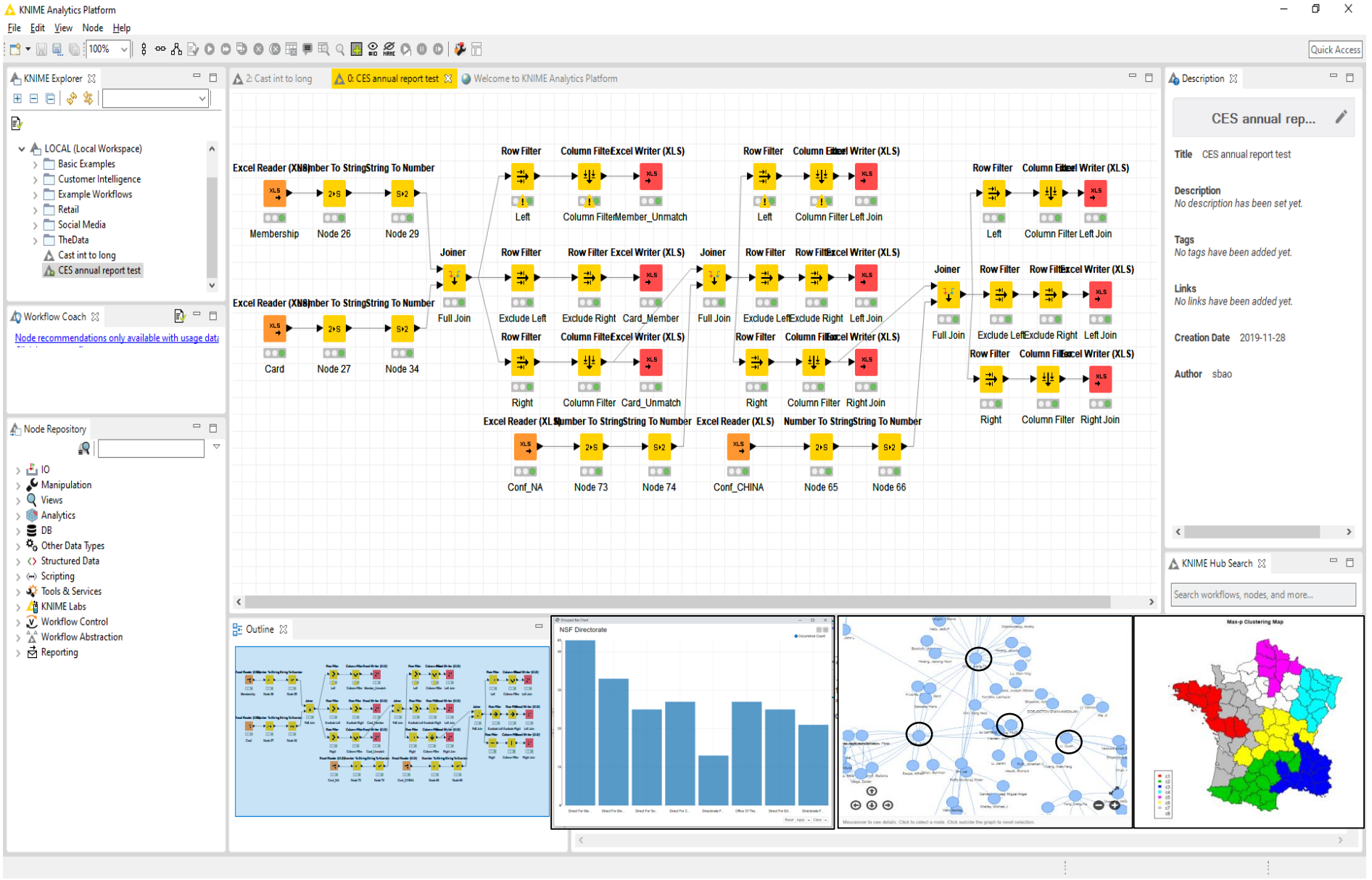


Replicable, Reproducible and Expandable Data Analysis





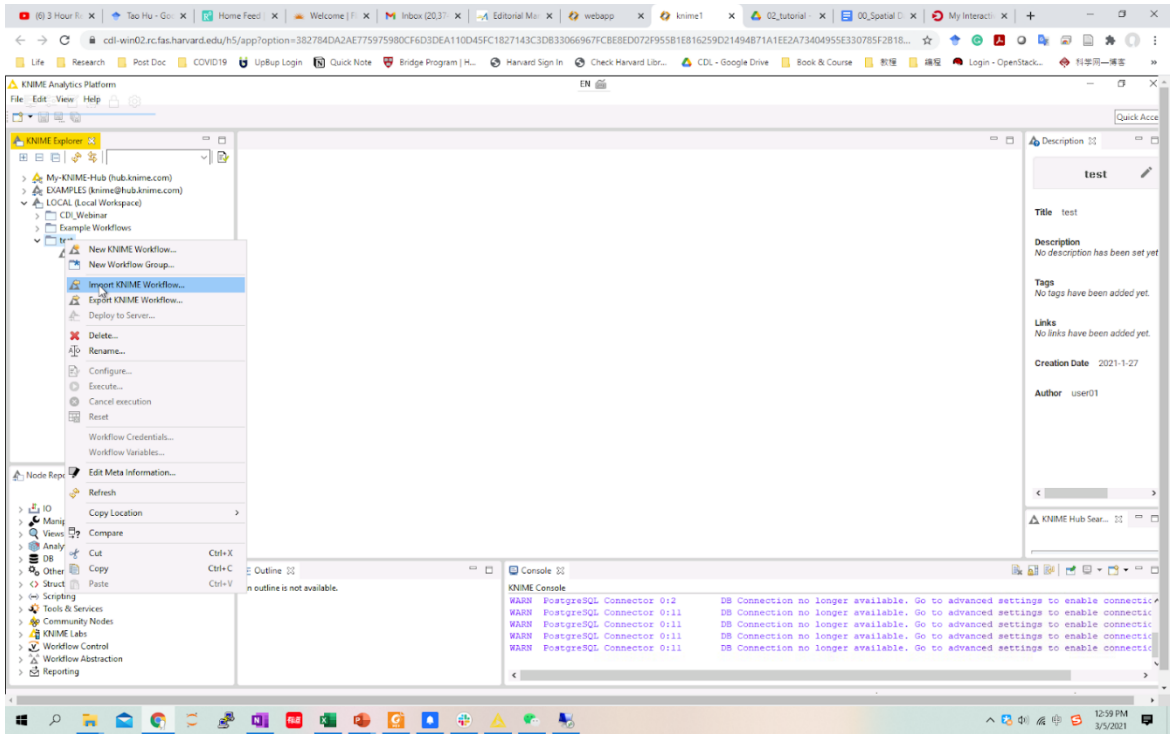


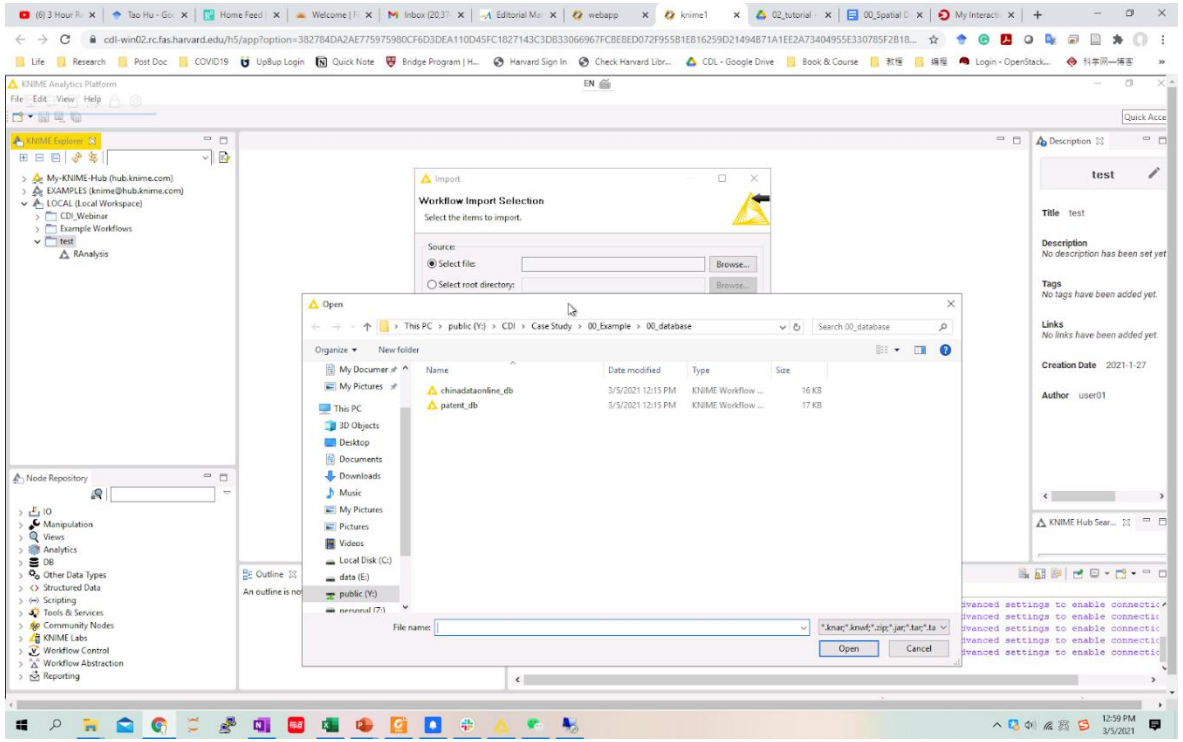


Data Access on the Platform

1. Access Cross-Platform Census Data from China Data Online

1) Import Workflow from Public Folder **Y:\CDI\Case Study\00_Example\00_database\chinadataonline_db**

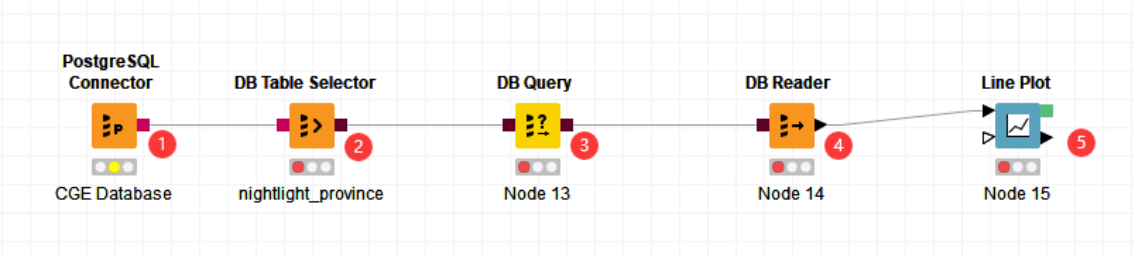




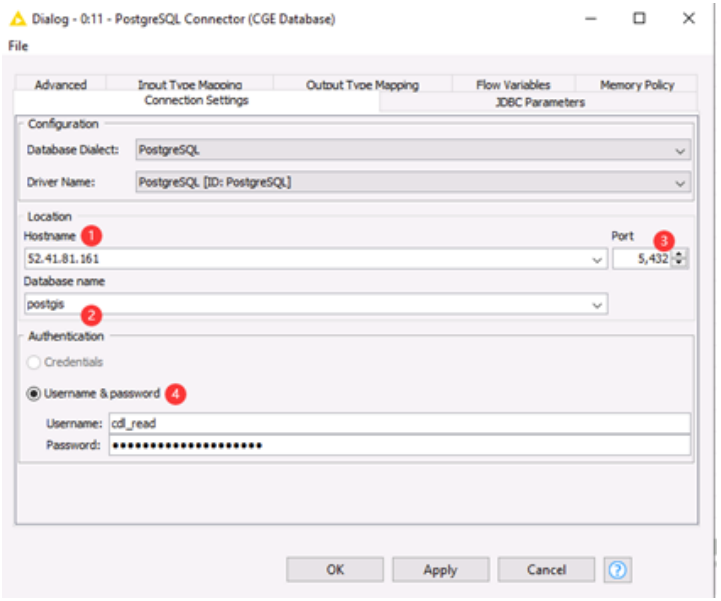
Data Access on the Platform

1. Access Cross-Platform Census Data from China Data Online

2) The imported workflow is shown as below. There are five components: PostgreSQL database settings; DB Table selector; DB Query; read data from DB; line plot visualization.



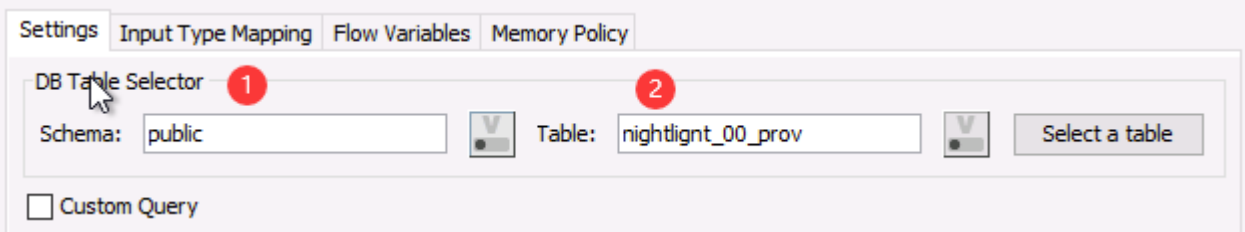
3) The imported workflow is shown as below. There are five components: PostgreSQL database settings; DB Table selector; DB Query; read data from DB; line plot visualization.



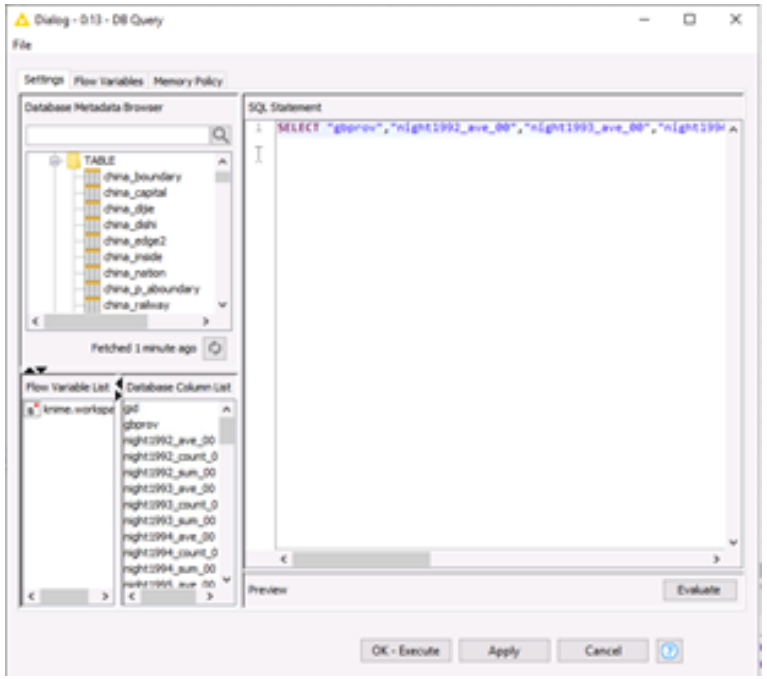
Data Access on the Platform

1. Access Cross-Platform Census Data from China Data Online

4) Select table in the node ‘DB Table Selector’



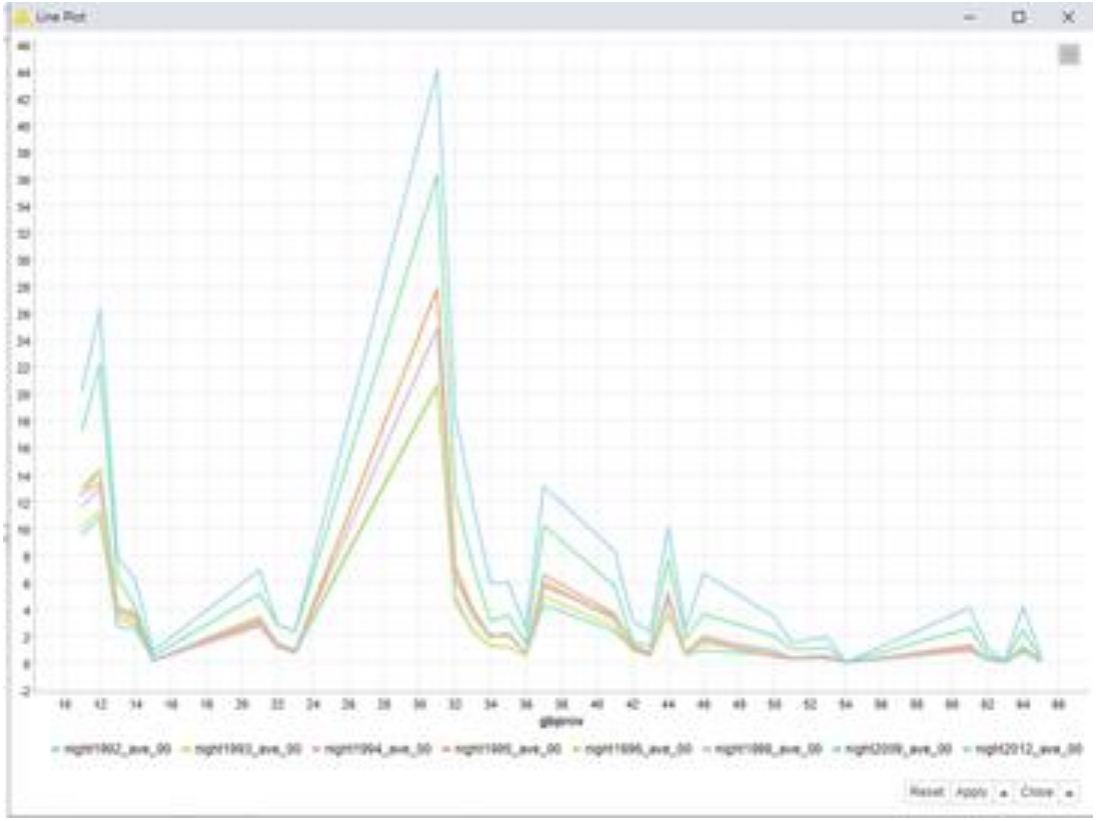
5) Create SQL in the node ‘DB Query’, e.g., “select "gbprov", "night1992_ave_00", "night1993_ave_00" from #table# AS "table"”



Data Access on the Platform

1. Access Cross-Platform Census Data from China Data Online

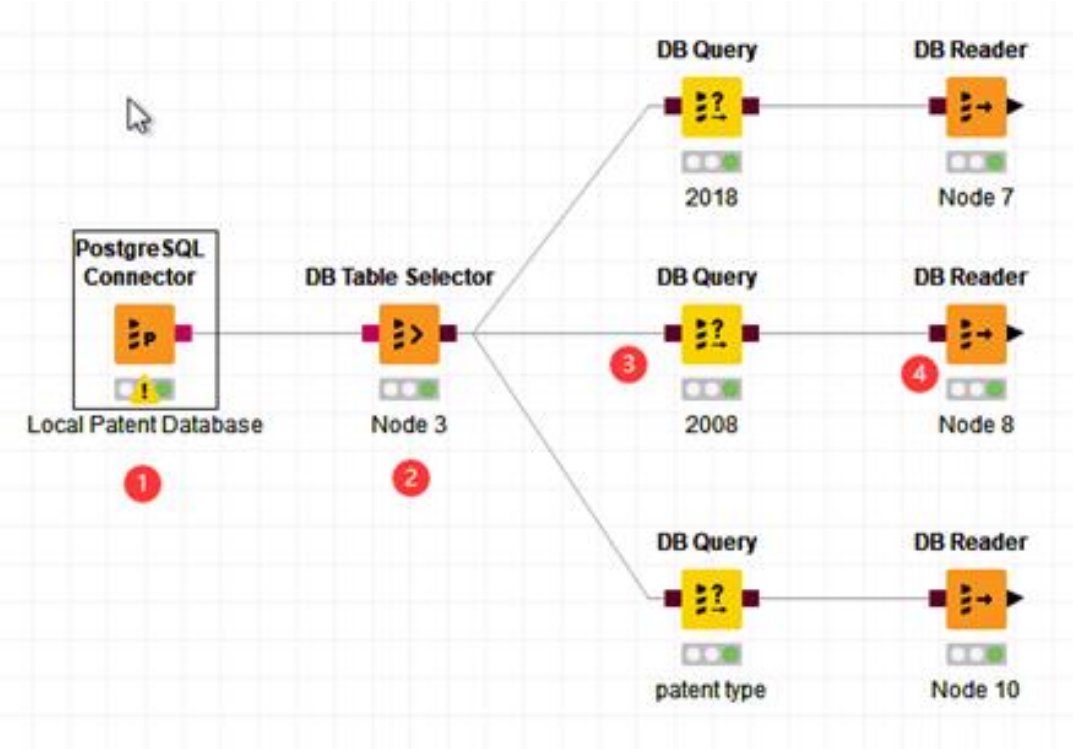
6) Load data by executing node ‘DB Reader’
7) Right click the node and select ‘execute’. After it is completed, right click the node and select ‘Interactive View: Line Plot’.



Data Access on the Platform

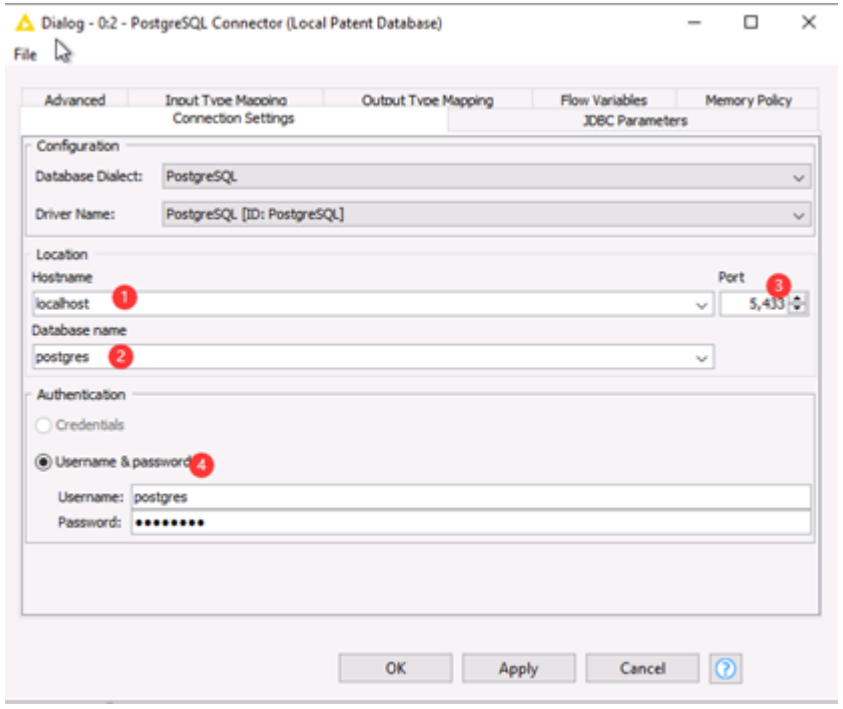
2. Access Large-scale Patent Data from Platform Database

1) Import Workflow from Public Folder **Y:\CDI\Case Study\00_Example\00_database\patent_db**
2) The imported workflow is shown as below. There are five components: PostgreSQL database settings; DB Table selector; DB Query; read data from DB to get results.

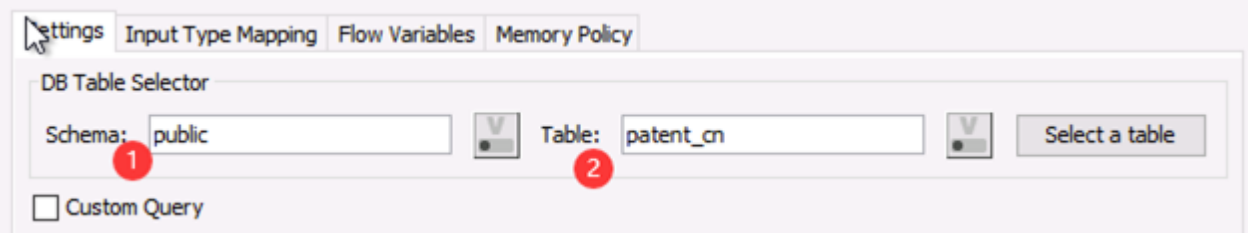


Data Access on the Platform

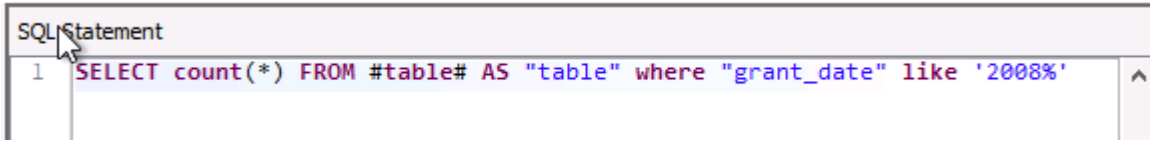
3) For PostgreSQL database settings, apply the default values. You can right click on the node and select ‘configure’ to check the settings.



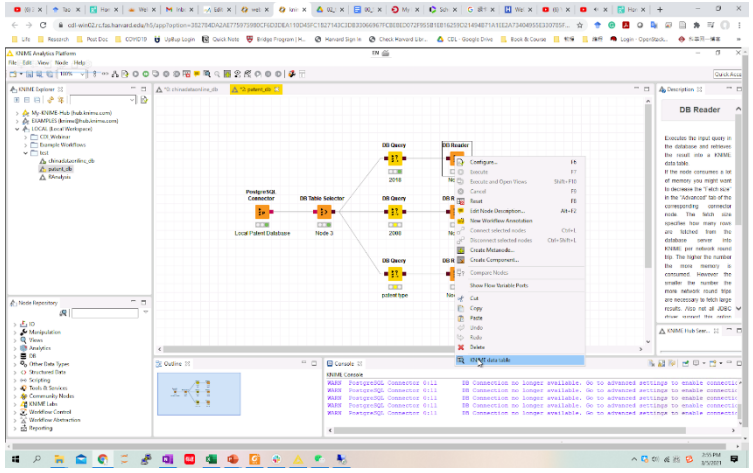
4) Select table in the node ‘DB Table Selector’



5) Create SQL in the node ‘DB Query’ shown as below



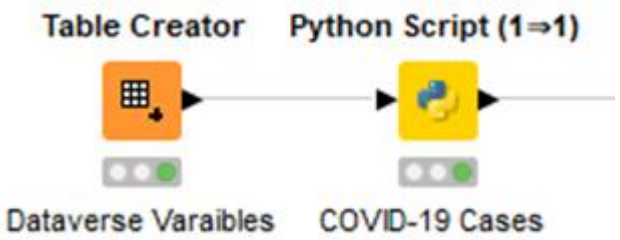
6) Check result from node ‘DB Reader’ by right clicking ‘table’



Data Access on the Platform

3. Access Public COVID-19 Data from Harvard Dataverse

1) Import workflow from public case study folder Y:\CDI\Case Study\00_Example\01_dataverse\dataverse_api.knwf

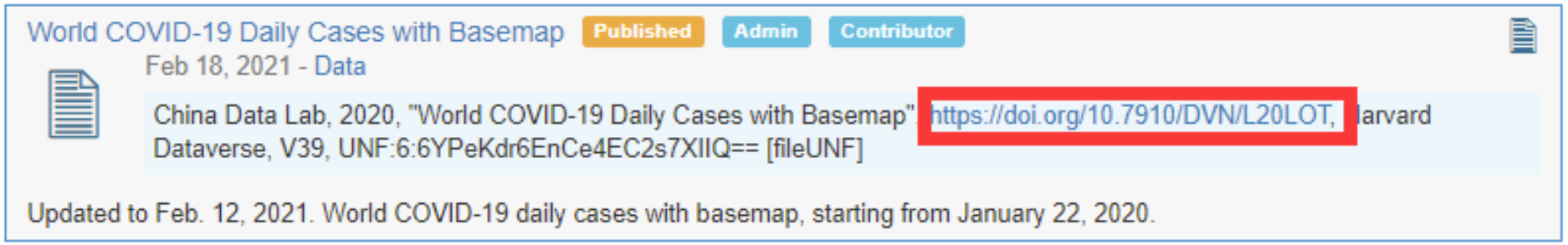


2) Set variables in the 1st node ‘Table Creator’. Right click node and select ‘configure’ and fill out the information in each parameter.

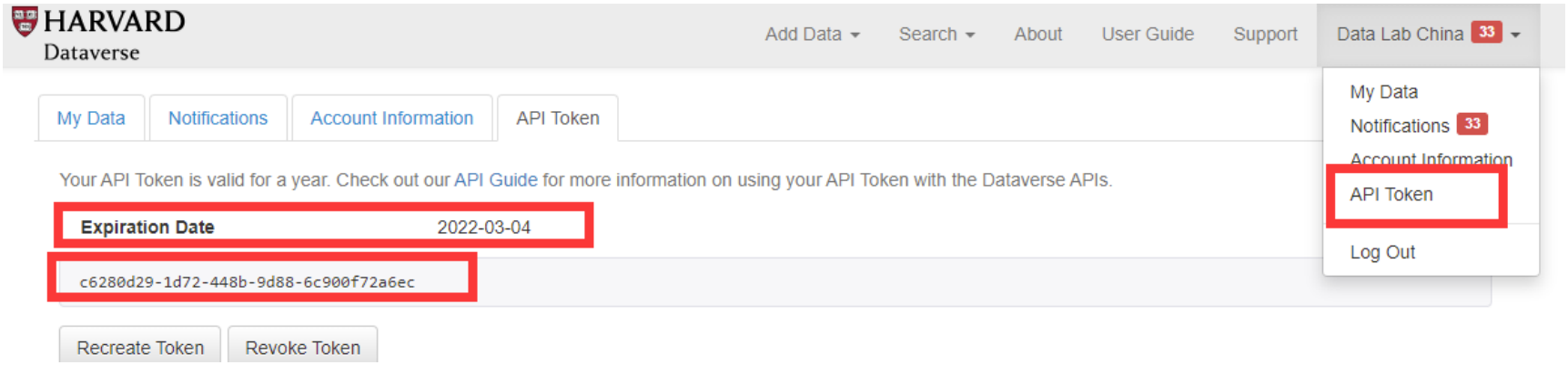
doi	api_token	file_id
7910/DVN/L20LOT	b7918464-48b7-47da-88d8-749a95ffde12	4411772

Data Access on the Platform

2.1) Doi: dataset doi number shown in the datasets metadata on Harvard Dataverse shown as below. After you go to https://dataverse.harvard.edu/dataverse/2019ncov, datasets will be listed and the DOI number is shown in each dataset.

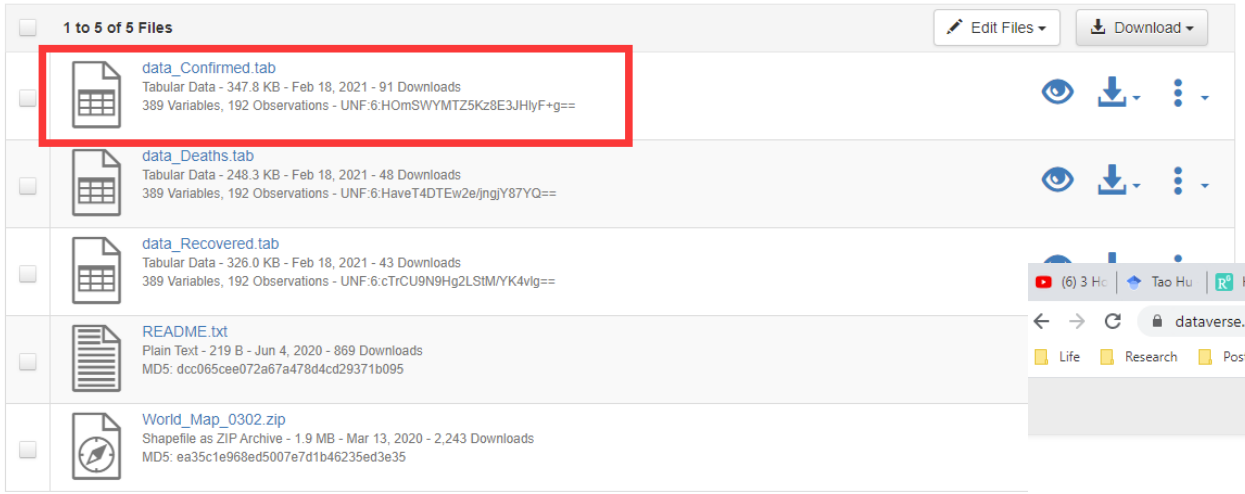


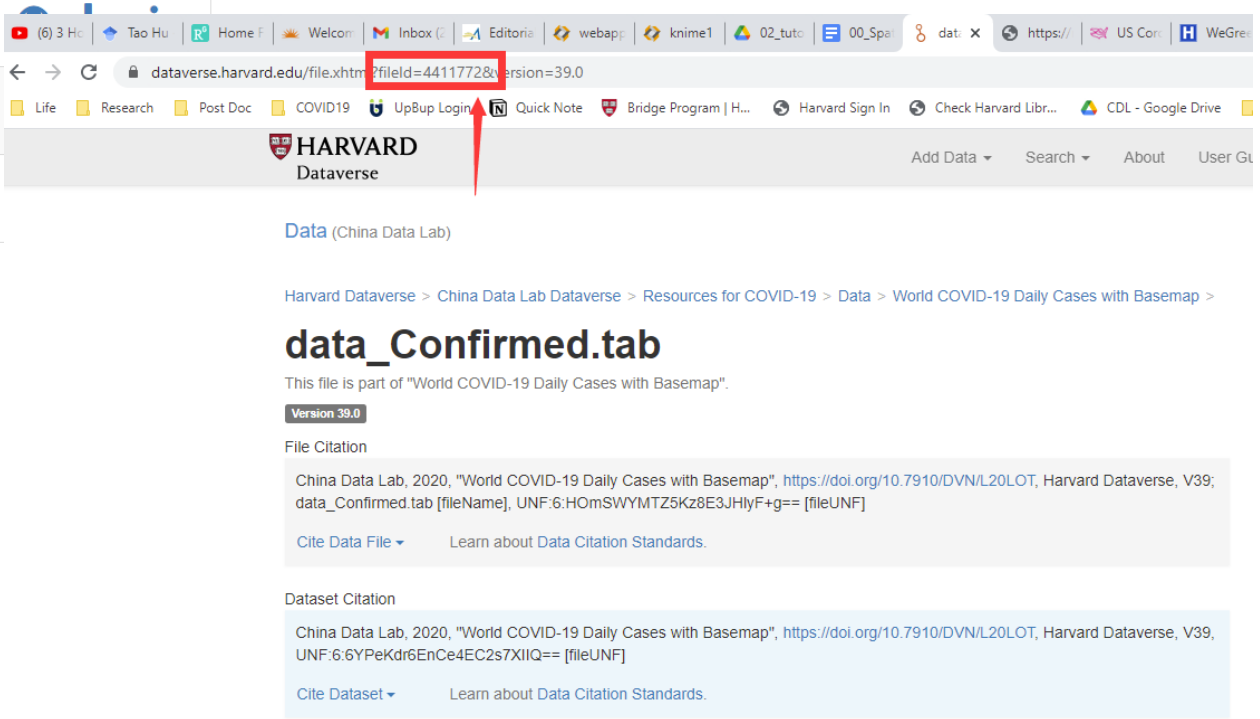
2.2) Api_token: **c6280d29-1d72-448b-9d88-6c900f72a6ec** is the default value provided by SDL and it will be updated weekly. Users can generate new api_token after logging into Harvard Dataverse. Notice that the Token has an **expiration date**.



Data Access on the Platform

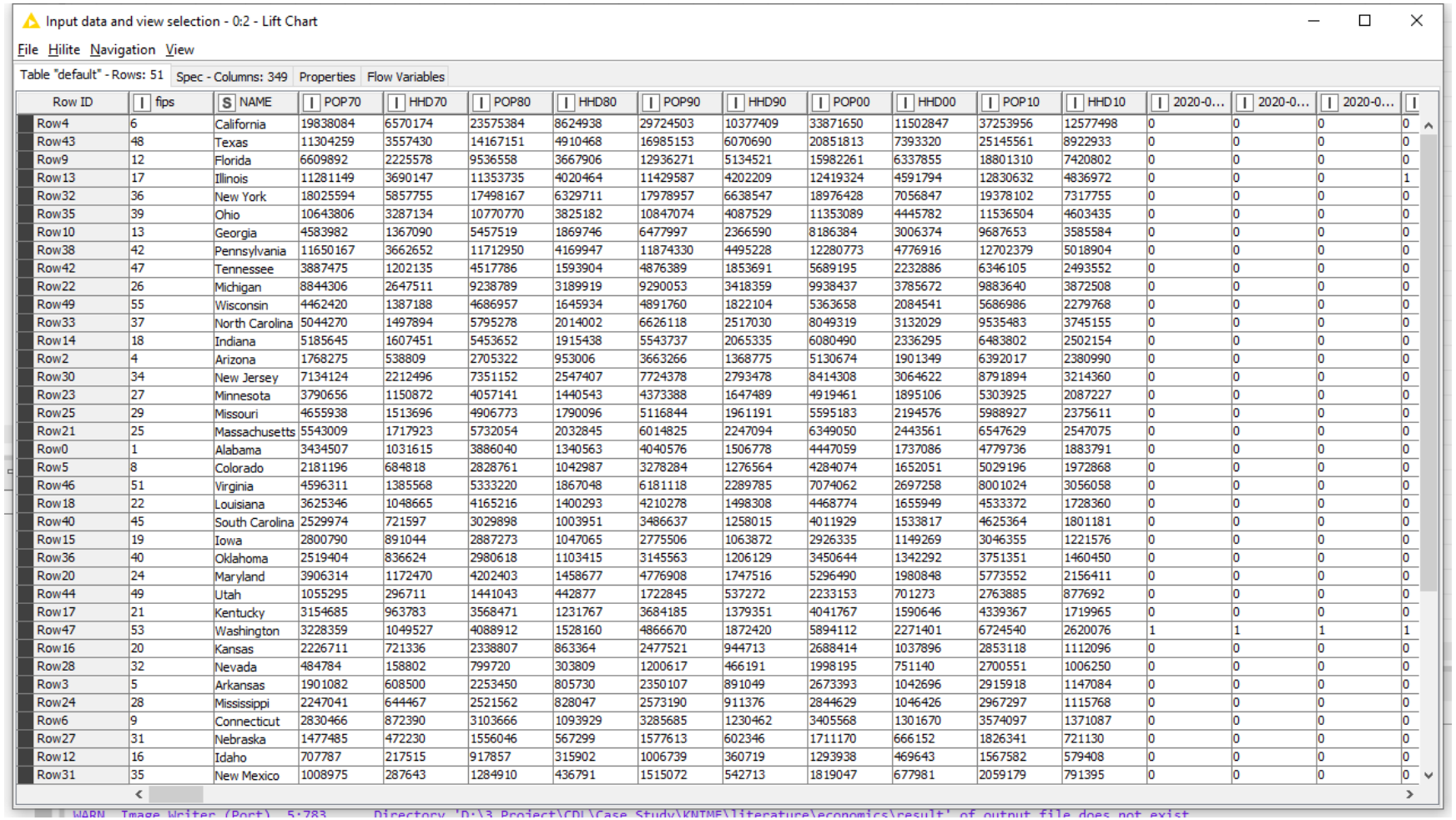
2.3) File_id: Go to the data file which you want to use and click the file. In the new web page, you will find the file id in the link. Please follow the steps shown below.





Data Access on the Platform

3) Right click the 2nd node ‘Python Script’ and select ‘table’ after it is finished. Users will see the results shown as below.



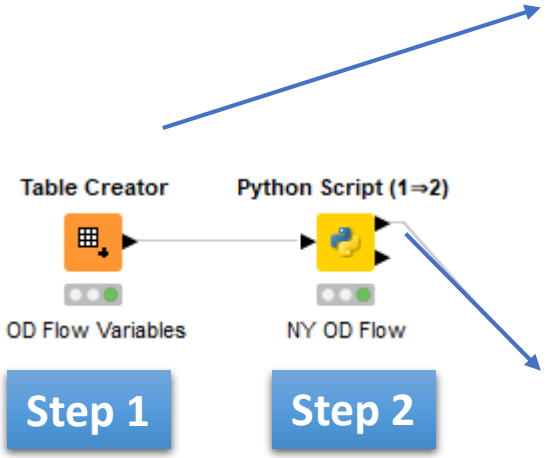
Row ID	fips	NAME	POP70	HHD70	POP80	HHD80	POP90	HHD90	POP00	HHD00	POP10	HHD10	2020-0...	2020-0...	2020-0...	
Row4	6	California	19838084	6570174	23575384	8624938	29724503	10377409	33871650	11502847	37253956	12577498	0	0	0	0
Row43	48	Texas	11304259	3557430	14167151	4910468	16985153	6070690	20851813	7393320	25145561	8922933	0	0	0	0
Row9	12	Florida	6609892	2225578	9536558	3667906	12936271	5134521	15982261	6337855	18801310	7420802	0	0	0	0
Row13	17	Illinois	11281149	3690147	11353735	4020464	11429587	4202209	12419324	4591794	12830632	4836972	0	0	0	1
Row32	36	New York	18025594	5857755	17498167	6329711	17978957	6638547	18976428	7056847	19378102	7317755	0	0	0	0
Row35	39	Ohio	10643806	3287134	10770770	3825182	10847074	4087529	11353089	4445782	11536504	4603435	0	0	0	0
Row10	13	Georgia	4583982	1367090	5457519	1869746	6477997	2366590	8186384	3006374	9687653	3585584	0	0	0	0
Row38	42	Pennsylvania	11650167	3662652	11712950	4169947	11874330	4495228	12280773	4776916	12702379	5018904	0	0	0	0
Row42	47	Tennessee	3887475	1202135	4517786	1593904	4876389	1853691	5689195	2232886	6346105	2493552	0	0	0	0
Row22	26	Michigan	8844306	2647511	9238789	3189919	9290053	3418359	9938437	3785672	9883640	3872508	0	0	0	0
Row49	55	Wisconsin	4462420	1387188	4686957	1645934	4891760	1822104	5363658	2084541	5686986	2279768	0	0	0	0
Row33	37	North Carolina	5044270	1497894	5795278	2014002	6626118	2517030	8049319	3132029	9535483	3745155	0	0	0	0
Row14	18	Indiana	5185645	1607451	5453652	1915438	5543737	2065335	6080490	2336295	6483802	2502154	0	0	0	0
Row2	4	Arizona	1768275	538809	2705322	953006	3663266	1368775	5130674	1901349	6392017	2380990	0	0	0	0
Row30	34	New Jersey	7134124	2212496	7351152	2547407	7724378	2793478	8414308	3064622	8791894	3214360	0	0	0	0
Row23	27	Minnesota	3790656	1150872	4057141	1440543	4373388	1647489	4919461	1895106	5303925	2087227	0	0	0	0
Row25	29	Missouri	4655938	1513696	4906773	1790096	5116844	1961191	5595183	2194576	5988927	2375611	0	0	0	0
Row21	25	Massachusetts	5543009	1717923	5732054	2032845	6014825	2247094	6349050	2443561	6547629	2547075	0	0	0	0
Row0	1	Alabama	3434507	1031615	3886040	1340563	4040576	1506778	4447059	1737086	4779736	1883791	0	0	0	0
Row5	8	Colorado	2181196	684818	2828761	1042987	3278284	1276564	4284074	1652051	5029196	1972868	0	0	0	0
Row46	51	Virginia	4596311	1385568	5333220	1867048	6181118	2289785	7074062	2697258	8001024	3056058	0	0	0	0
Row18	22	Louisiana	3625346	1048665	4165216	1400293	4210278	1498308	4468774	1655949	4533372	1728360	0	0	0	0
Row40	45	South Carolina	2529974	721597	3029898	1003951	3486637	1258015	4011929	1533817	4625364	1801181	0	0	0	0
Row15	19	Iowa	2800790	891044	2887273	1047065	2775506	1063872	2926335	1149269	3046355	1221576	0	0	0	0
Row36	40	Oklahoma	2519404	836624	2980618	1103415	3145563	1206129	3450644	1342292	3751351	1460450	0	0	0	0
Row20	24	Maryland	3906314	1172470	4202403	1458677	4776908	1747516	5296490	1980848	5773552	2156411	0	0	0	0
Row44	49	Utah	1055295	296711	1441043	442877	1722845	537272	2233153	701273	2763885	877692	0	0	0	0
Row17	21	Kentucky	3154685	963783	3568471	1231767	3684185	1379351	4041767	1590646	4339367	1719965	0	0	0	0
Row47	53	Washington	3228359	1049527	4088912	1528160	4866670	1872420	5894112	2271401	6724540	2620076	1	1	1	1
Row16	20	Kansas	2226711	721336	2338807	863364	2477521	944713	2688414	1037896	2853118	1112096	0	0	0	0
Row28	32	Nevada	484784	158802	799720	303809	1200617	466191	1998195	751140	2700551	1006250	0	0	0	0
Row3	5	Arkansas	1901082	608500	2253450	805730	2350107	891049	2673393	1042696	2915918	1147084	0	0	0	0
Row24	28	Mississippi	2247041	644467	2521562	828047	2573190	911376	2844629	1046426	2967297	1115768	0	0	0	0
Row6	9	Connecticut	2830466	872390	3103666	1093929	3285685	1230462	3405568	1301670	3574097	1371087	0	0	0	0
Row27	31	Nebraska	1477485	472230	1556046	567299	1577613	602346	1711170	666152	1826341	721130	0	0	0	0
Row12	16	Idaho	707787	217515	917857	315902	1006739	360719	1293938	469643	1567582	579408	0	0	0	0
Row31	35	New Mexico	1008975	287643	1284910	436791	1515072	542713	1819047	677981	2059179	791395	0	0	0	0

Data Access on the Platform

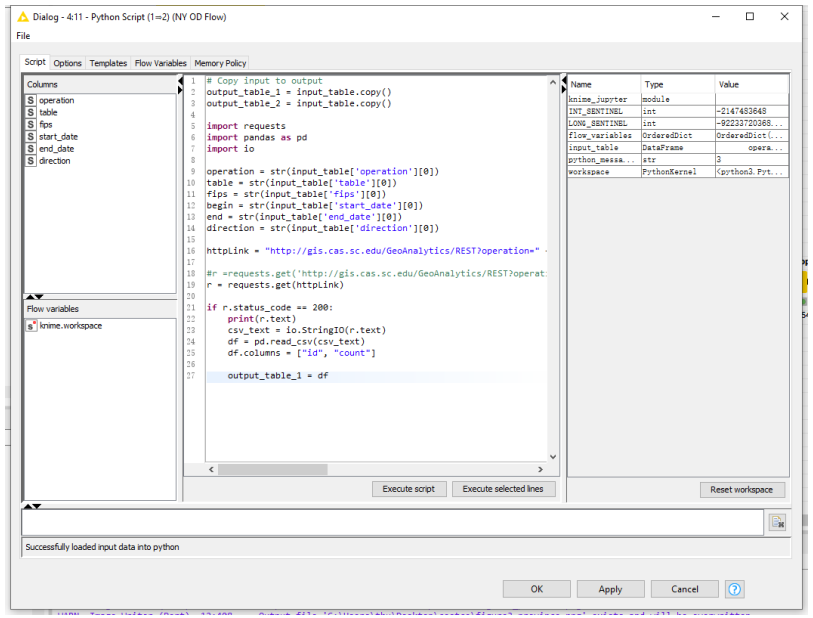
4. Access External Human Mobility Index Data via Rest API

http://gis.cas.sc.edu/GeoAnalytics/REST?operation=od_by_fips&table=twitter_od_2020_county&fips=12086&begin=01/01/2020&end=01/10/2020&direction=both

- operation
- table
- fips
- direction
- data_type
- bbox
- begin
- end



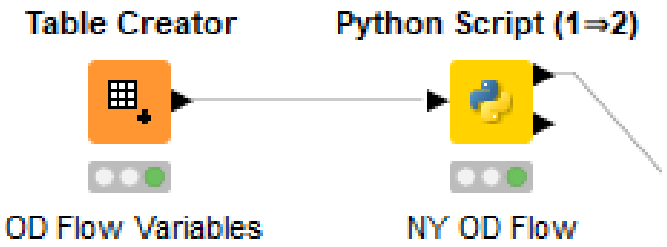
operation	table	fips	start_date	end_date	direction
od_by_fips	twitter_od_2...	New York	03/1/2020	03/21/2020	o_to_d



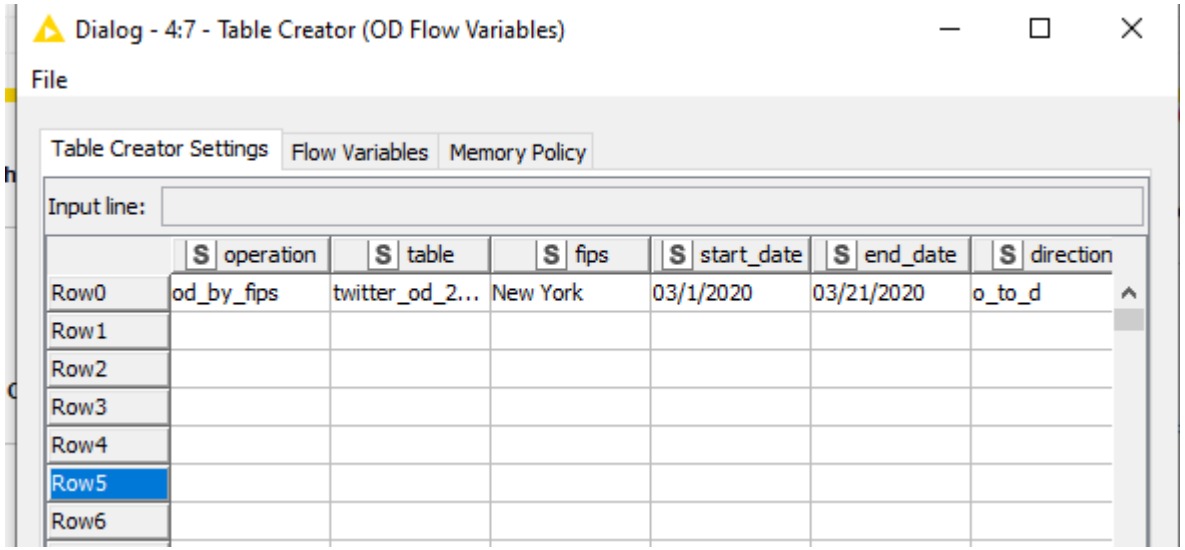
Row ID	id	count
Row11	Idaho	3
Row12	Michigan	141
Row13	Georgia	260
Row14	Massachusetts	461
Row15	South Dakota	12
Row16	Kentucky	56
Row17	Montana	7
Row18	Iowa	18
Row19	Oregon	43
Row20	Minnesota	38
Row21	California	902
Row22	Illinois	253
Row23	Vermont	162
Row24	Arizona	84
Row25	Delaware	64
Row26	Nevada	101
Row27	Tennessee	122
Row28	Pennsylvania	1039
Row29	Nebraska	20
Row30	Connecticut	987
Row31	West Virginia	40
Row32	Alabama	28
Row33	Texas	356
Row34	Arkansas	23
Row35	Maryland	265
Row36	Louisiana	95

Data Access on the Platform

1) Import workflow from public case study folder Y:\CDI\Case Study\00_Example\02_Rest_API\geotweets_test.knwf



2) In the 1st node ‘Table Creator’, fill out values for each variable. Detailed description for the variables can be found at shared google doc.



Data Access on the Platform

3) Right click the node ‘Python Script’ and click Table 1. The data will be shown as below.

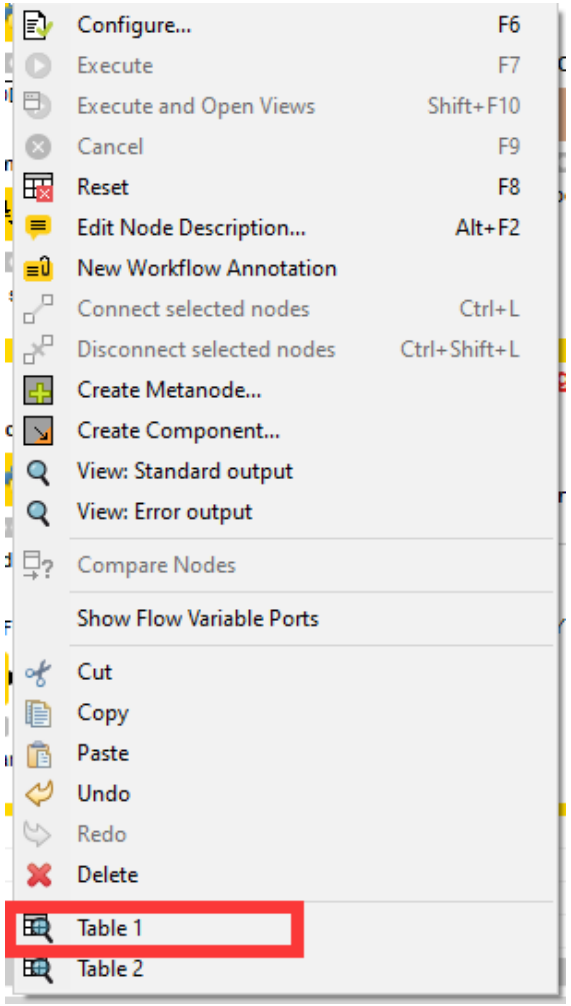


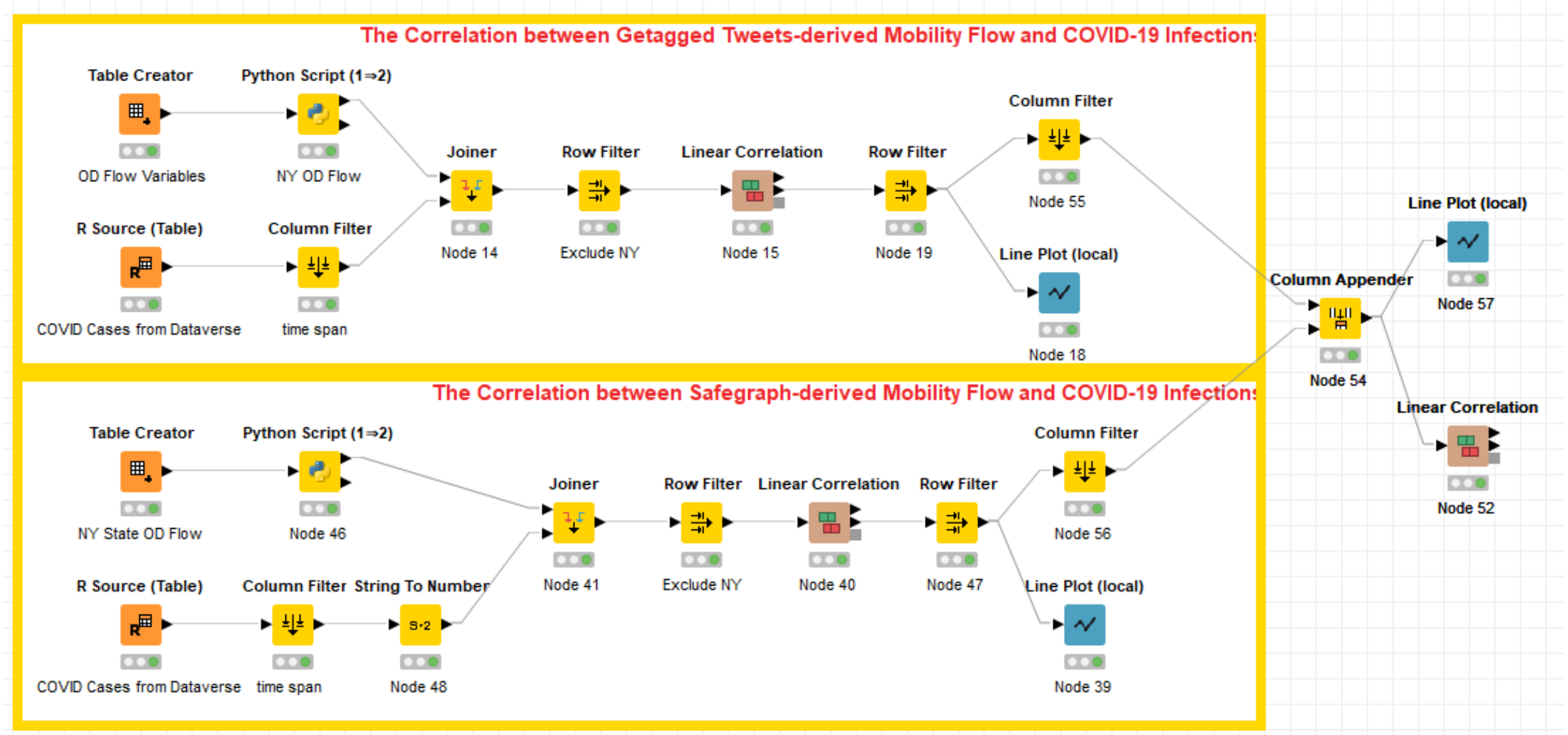
Table 1 - 4:11 - Python Script (1⇒2) (NY OD Flow

File Hilite Navigation View

Table "default" - Rows: 48 | Spec - Columns: 2 | Properti

Row ID	S id	I count
Row0	Oklahoma	34
Row1	Wyoming	7
Row2	North Dakota	1
Row3	Rhode Island	77
Row4	Florida	580
Row5	Maine	25
Row6	District of Columbia	234
Row7	North Carolina	279
Row8	Indiana	80
Row9	New Hampshire	42
Row10	Colorado	86
Row11	Idaho	3
Row12	Michigan	141
Row13	Georgia	260
Row14	Massachusetts	461
Row15	South Dakota	12
Row16	Kentucky	56
Row17	Montana	7
Row18	Iowa	18
Row19	Oregon	43

Data Access Demo

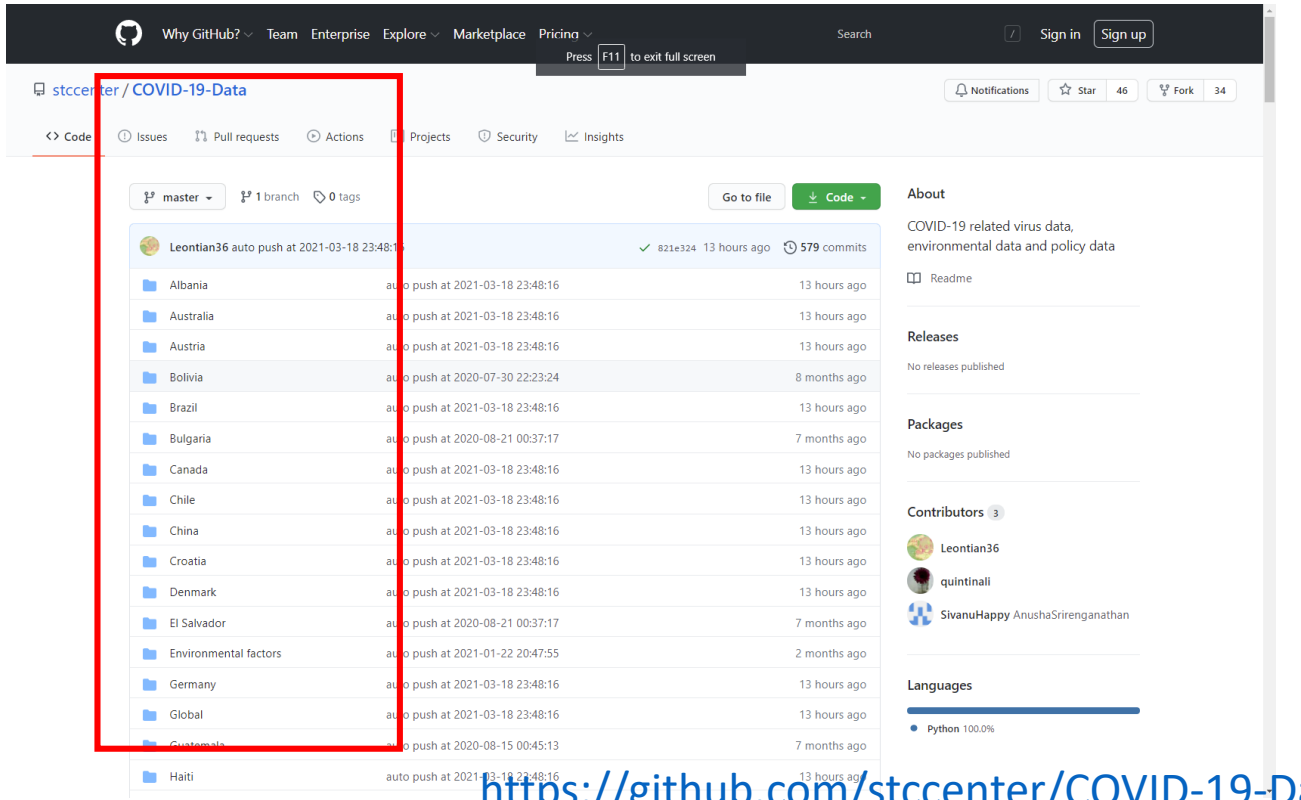


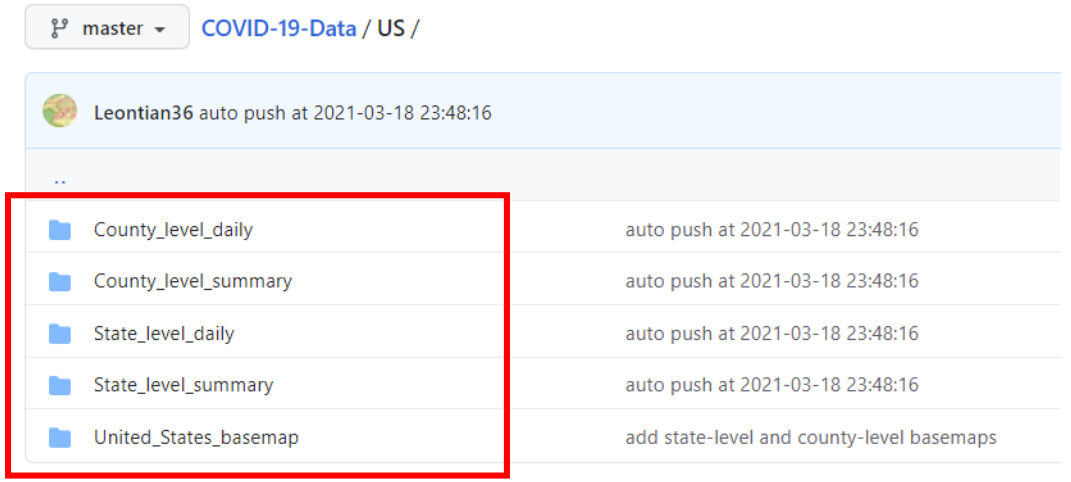
Data Access on the Platform

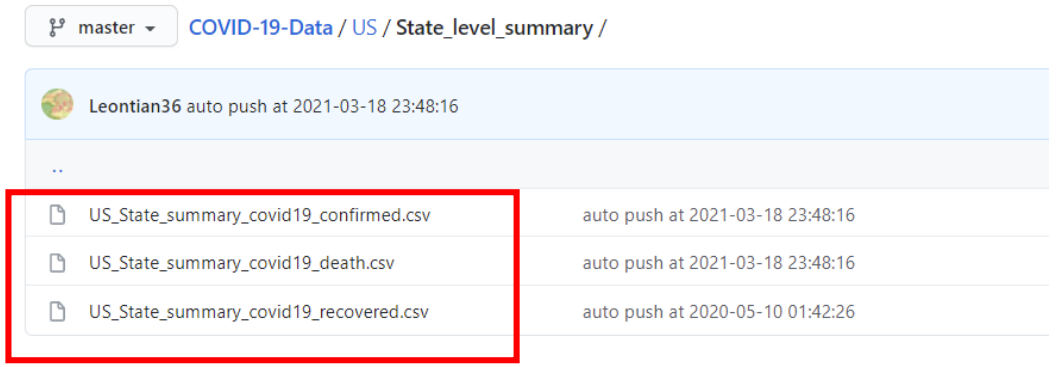
5. Access External Data via GitHub API

https://github.com/

GitHub, is a provider of Internet hosting for software development and version control using Git. It offers the distributed version control and source code/data management functionality of Git.





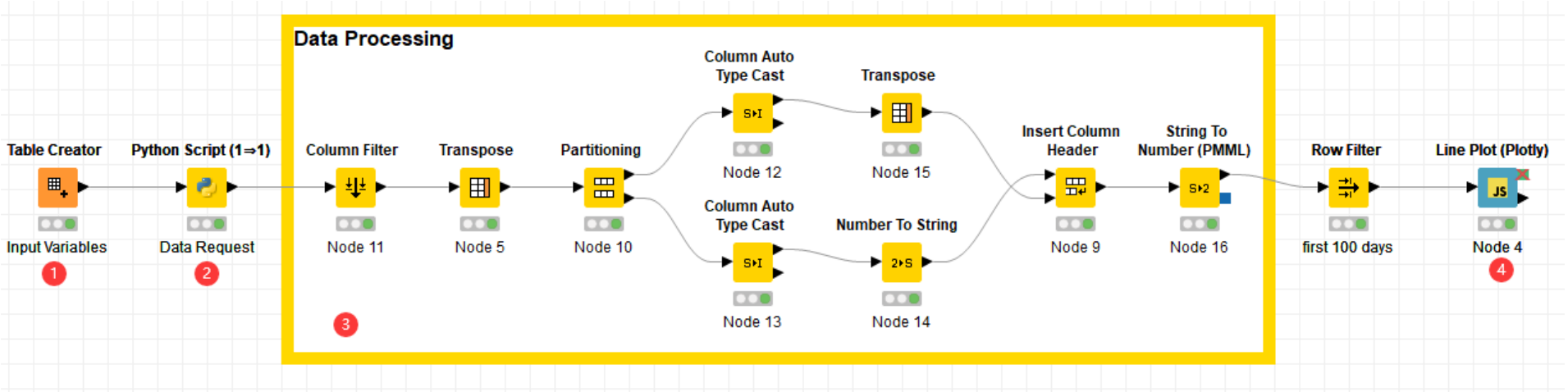


https://github.com/stccenter/COVID-19-Data

Data Access on the Platform

5. Access External Data via GitHub API

1) Import workflow from public case study folder Y:\CDI\Case Study\00_Example\02_Rest_API\github_test.knwf

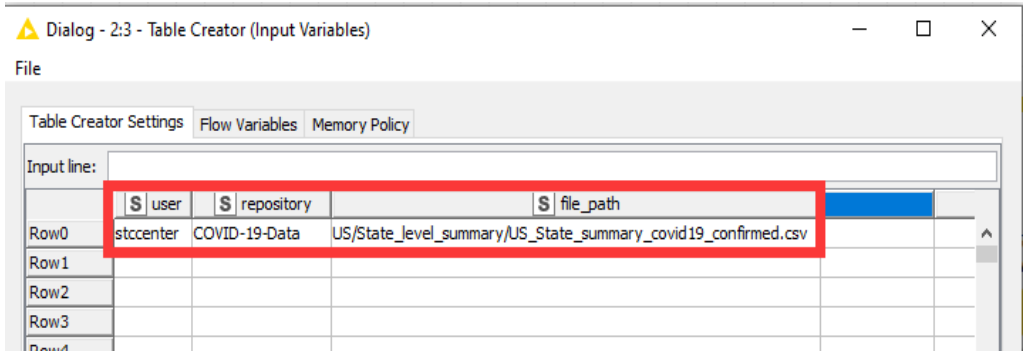


2) Set variables in the first node

a) User: GitHub username. The default value is **stccenter**.

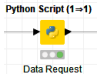
b) Repository: repository name is the user's github. The default value is COVID-19-Data, which is published by stccenter.

c) File_path: the path in the COVID-19-Data repository. The default value is US/State_level_summary/US_State_summary_covid19_confirmed.csv

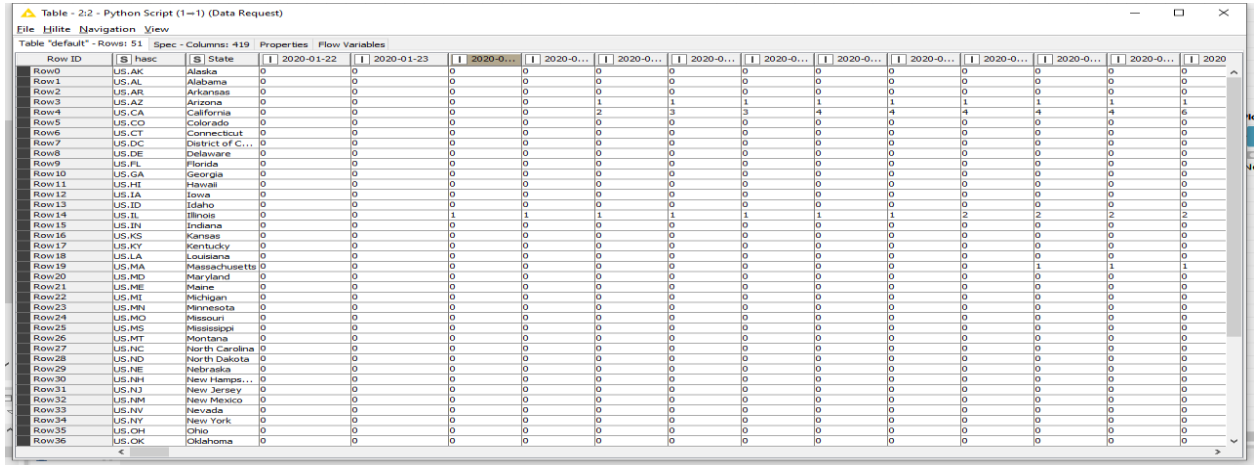


Data Access on the Platform

5. Access External Data via GitHub API

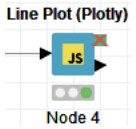
3) Check the output in the 2nd node 

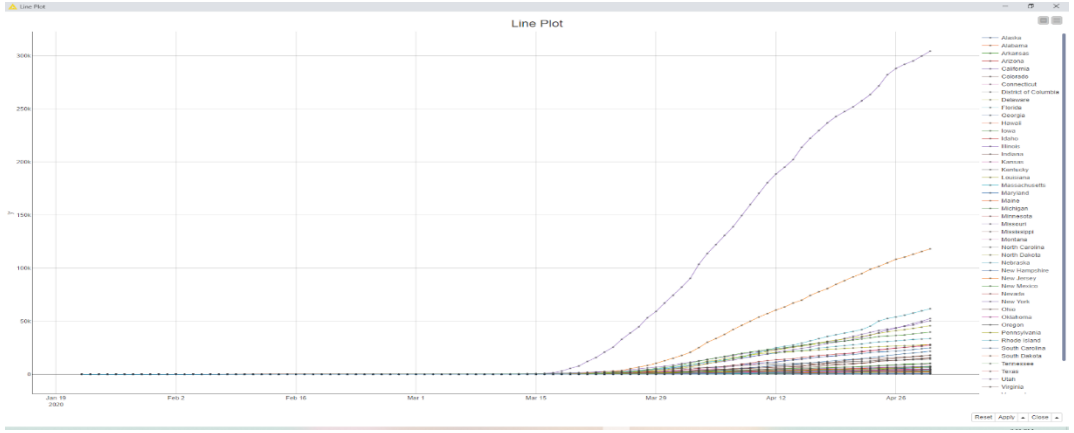
by right click and select table.



Row ID	hasc	State	2020-01-22	2020-01-23	2020-0...	2020-0...	2020-0...	2020-0...	2020-0...	2020-0...	2020-0...	2020-0...	2020-0...	2020-0...	2020-0...	2020
Row0	US.AK	Alaska	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row1	US.AL	Alabama	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row2	US.AR	Arkansas	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row3	US.AZ	Arizona	0	0	0	1	1	1	1	1	1	1	1	1	1	1
Row4	US.CA	California	0	0	0	2	3	3	4	4	4	4	4	4	4	6
Row5	US.CO	Colorado	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row6	US.CT	Connecticut	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row7	US.DC	District of C...	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row8	US.DE	Delaware	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row9	US.FL	Florida	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row10	US.GA	Georgia	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row11	US.HI	Hawaii	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row12	US.IA	Iowa	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row13	US.ID	Idaho	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row14	US.IL	Illinois	0	0	1	1	1	1	1	1	1	2	2	2	2	2
Row15	US.IN	Indiana	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row16	US.KS	Kansas	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row17	US.KY	Kentucky	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18	US.LA	Louisiana	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row19	US.MA	Massachusetts	0	0	0	0	0	0	0	0	0	1	1	1	1	1
Row20	US.MD	Maryland	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row21	US.ME	Maine	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row22	US.MI	Michigan	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row23	US.MN	Minnesota	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row24	US.MO	Missouri	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row25	US.MS	Mississippi	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row26	US.MT	Montana	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row27	US.NC	North Carolina	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row28	US.ND	North Dakota	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row29	US.NE	Nebraska	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row30	US.NH	New Hamps...	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row31	US.NJ	New Jersey	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row32	US.NM	New Mexico	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row33	US.NV	Nevada	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row34	US.NY	New York	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row35	US.OH	Ohio	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row36	US.OK	Oklahoma	0	0	0	0	0	0	0	0	0	0	0	0	0	0

4) Click run button  in the toolbar and check the visualization from Node 4





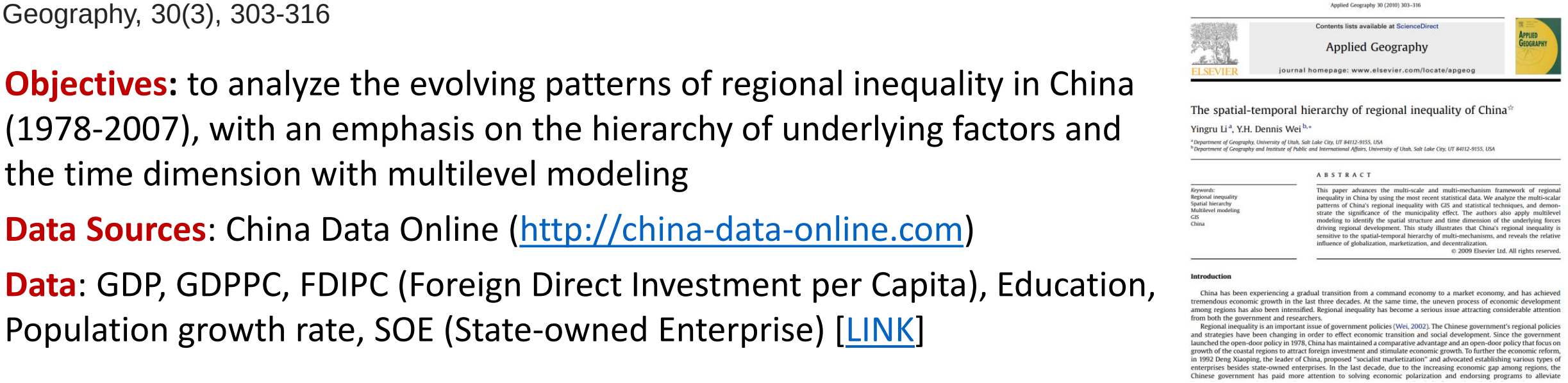
Case Study: Regional inequality of China

Li, Y., & Wei, Y. D. (2010). The spatial-temporal hierarchy of regional inequality of China. Applied Geography, 30(3), 303-316

Objectives: to analyze the evolving patterns of regional inequality in China (1978-2007), with an emphasis on the hierarchy of underlying factors and the time dimension with multilevel modeling

Data Sources: China Data Online (http://china-data-online.com)

Data: GDP, GDPPC, FDIPC (Foreign Direct Investment per Capita), Education, Population growth rate, SOE (State-owned Enterprise) [LINK]



Applied Geography 30 (2010) 303–316

Contents lists available at ScienceDirect

Applied Geography

journal homepage: www.elsevier.com/locate/apgeog

The spatial-temporal hierarchy of regional inequality of China☆

Yingru Li[a], Y.H. Dennis Wei[b,*]

[a] Department of Geography, University of Utah, Salt Lake City, UT 84112-9155, USA
[b] Department of Geography and Institute of Public and International Affairs, University of Utah, Salt Lake City, UT 84112-9155, USA

ABSTRACT

Keywords:
Regional inequality
Spatial hierarchy
Multilevel modeling
GIS
China

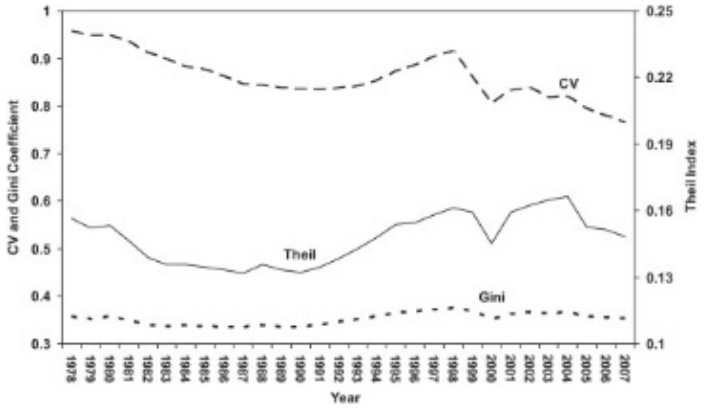
This paper advances the multi-scale and multi-mechanism framework of regional inequality in China by using the most recent statistical data. We analyze the multi-scalar patterns of China's regional inequality with GIS and statistical techniques, and demonstrate the significance of the municipality effect. The authors also apply multilevel modeling to identify the spatial structure and time dimension of the underlying forces driving regional development. This study illustrates that China's regional inequality is sensitive to the spatial-temporal hierarchy of multi-mechanisms, and reveals the relative influence of globalization, marketization, and decentralization.

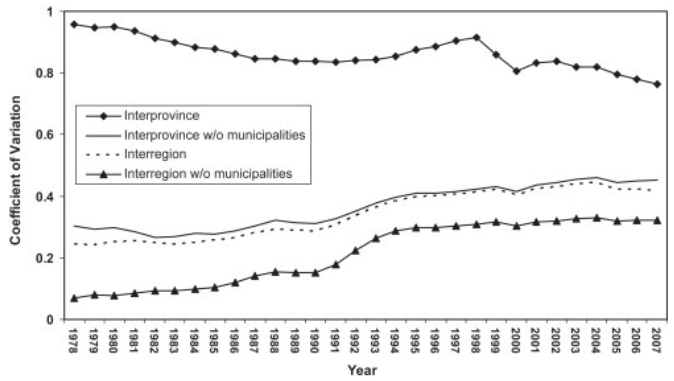
© 2009 Elsevier Ltd. All rights reserved.

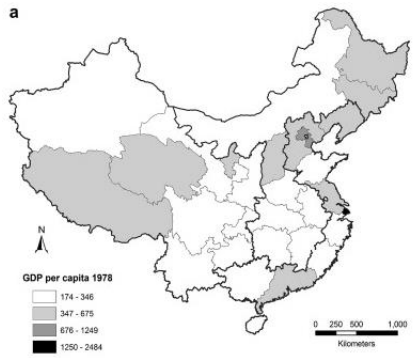
Introduction

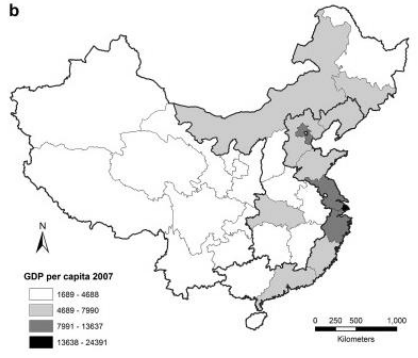
China has been experiencing a gradual transition from a command economy to a market economy, and has achieved tremendous economic growth in the last three decades. At the same time, the uneven process of economic development among regions has also been intensified. Regional inequality has become a serious issue attracting considerable attention from both the government and researchers.

Regional inequality is an important issue of government policies (Wei, 2002). The Chinese government's regional policies and strategies have been changing in order to effect economic transition and social development. Since the government launched the open-door policy in 1978, China has maintained a comparative advantage and an open-door policy that focus on growth of the coastal regions to attract foreign investment and stimulate economic growth. To further the economic reform, in 1992 Deng Xiaoping, the leader of China, proposed "socialist marketization" and advocated establishing various types of enterprises besides state-owned enterprises. In the last decade, due to the increasing economic gap among regions, the Chinese government has paid more attention to solving economic polarization and endorsing programs to alleviate









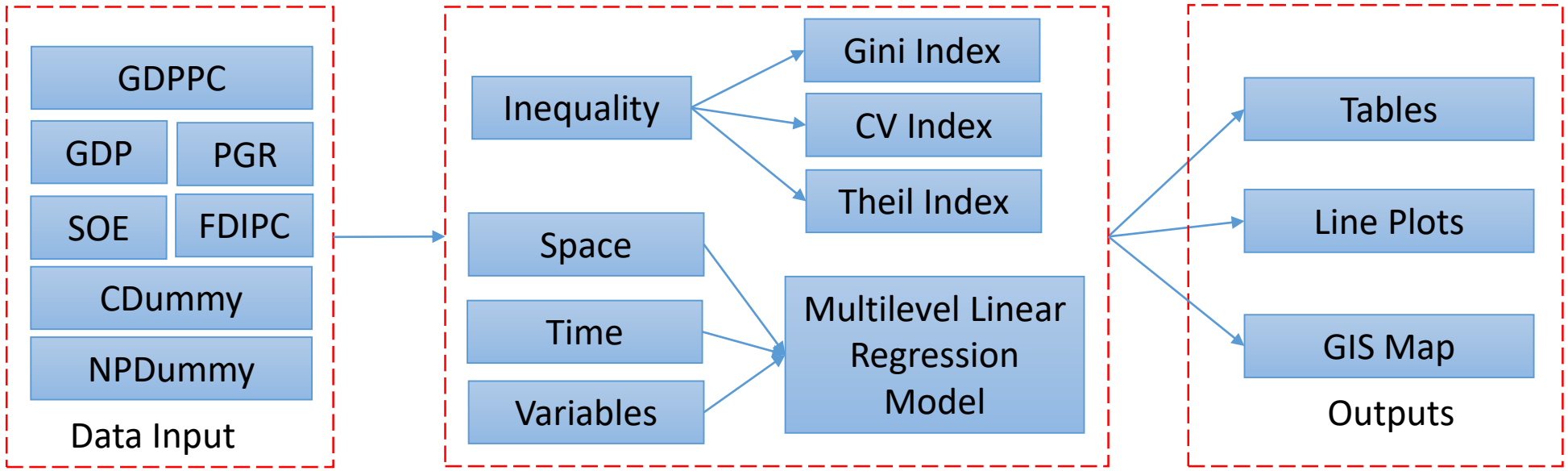
Case Study: Regional inequality of China

- **Methodology:** to understand China's regional inequality, **multilevel regression modeling** is applied to examine the underlining mechanism

$$y_{ijt} = \beta_0 + \beta_1 x_{ijt} + u_t + r_{jt} + e_{ijt}$$

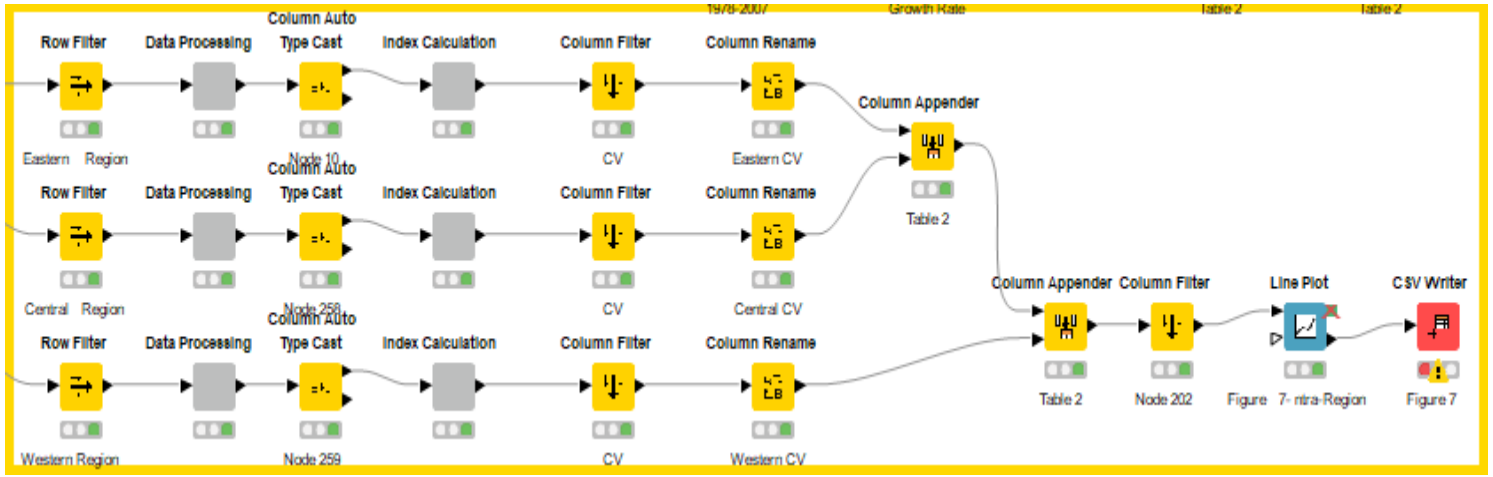
single-level (province), two-level (region and province) and three-level (time, region, and province)

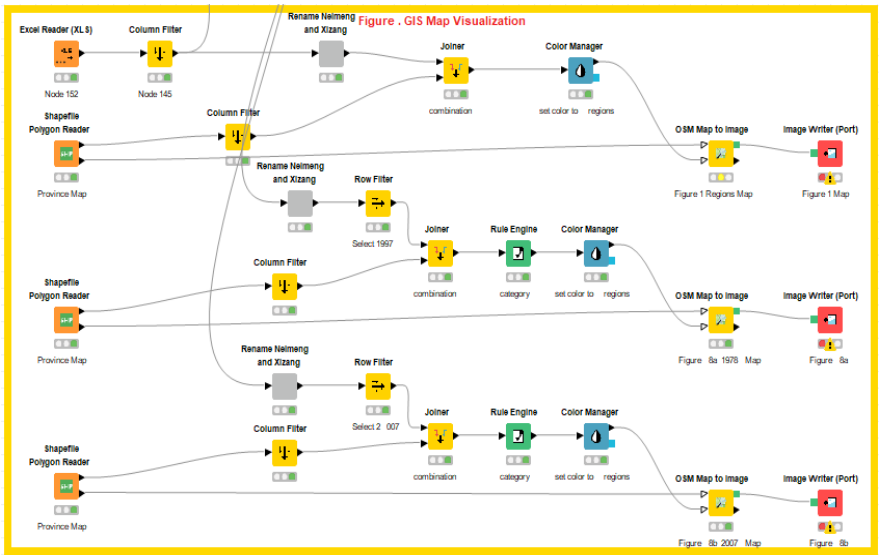
- **Flowchart**

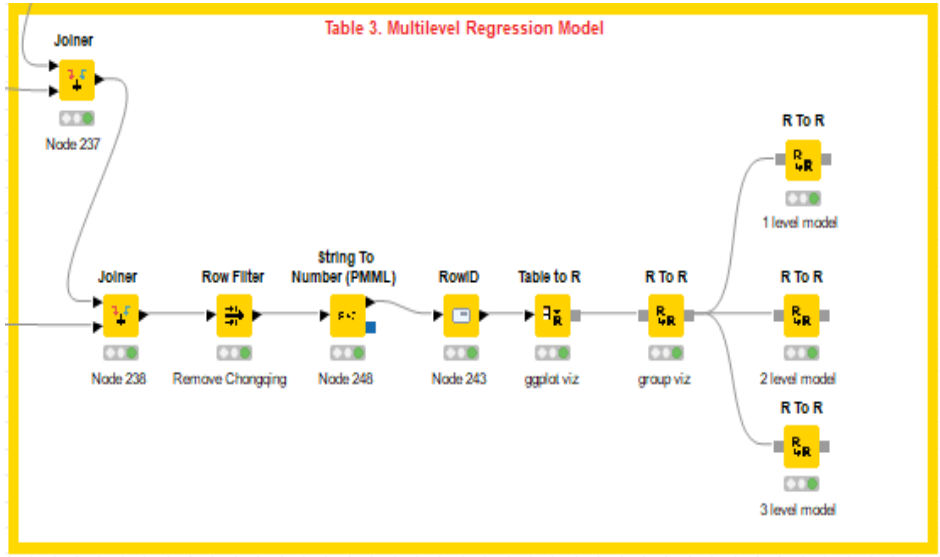


Case Study: Regional inequality of China

The Workflow Implementation by KNIME







Case Study: Regional inequality of China

Figure 4

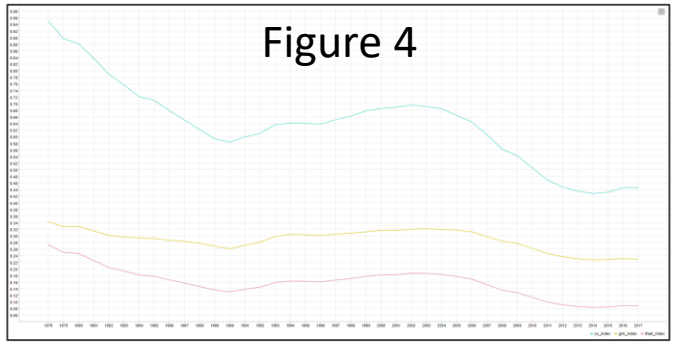


Figure 5

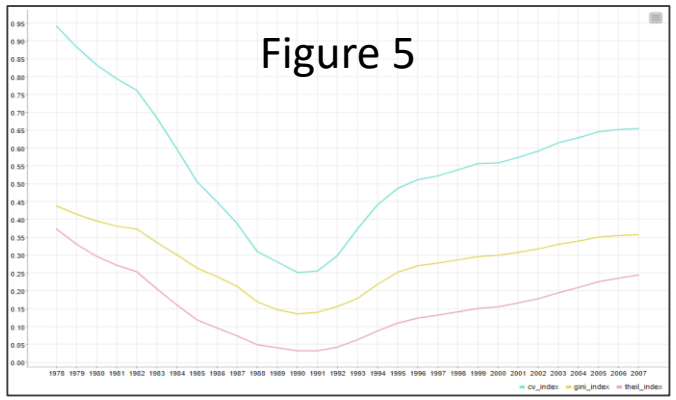


Figure 6

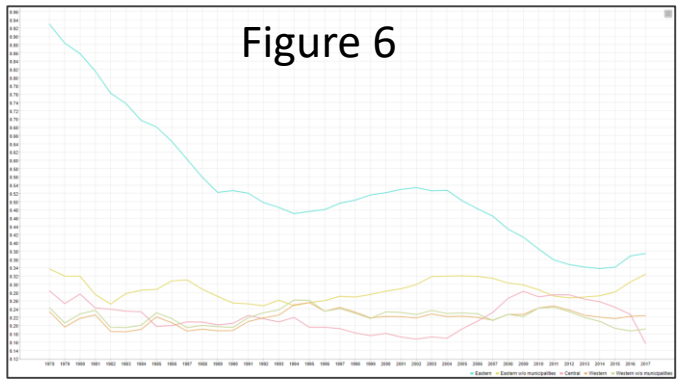


Figure 1

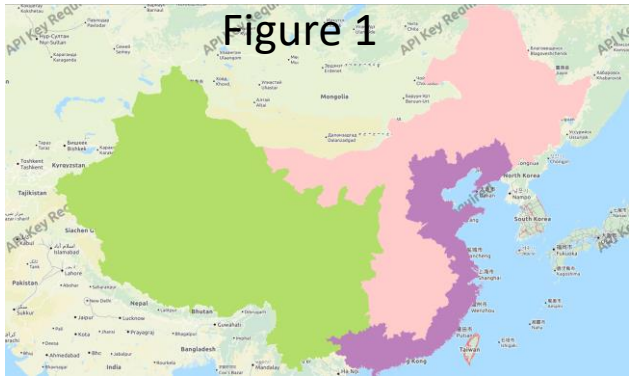
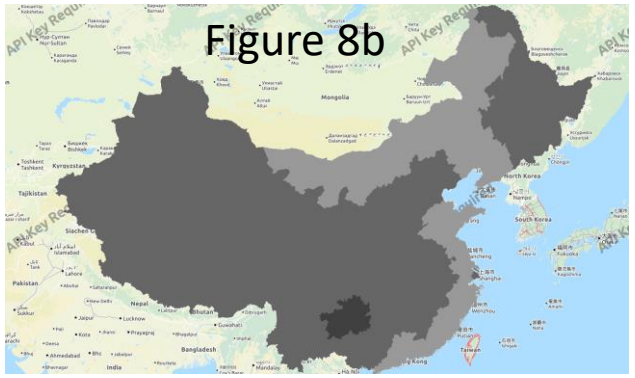


Figure 8a

Figure 8b



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9061.972   1636.329   5.538 1.42e-07 ***
FDIPC         101.175      5.755  17.581  < 2e-16 ***
SOE            -9.227      2.841  -3.248  0.00145 **
EDU            82.131     13.637   6.023 1.38e-08 ***
POPGR        -425.832    101.141  -4.210 4.49e-05 ***
Cdummy      -1515.530    935.249  -1.620  0.10734
NPDummy       627.368    906.484   0.692  0.49000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

One level model

```
Fixed effects:
             Estimate Std. Error       df t value Pr(>|t|)
(Intercept)  9061.972   1636.329  143.000   5.538 1.42e-07 ***
FDIPC         101.175      5.755  143.000  17.581  < 2e-16 ***
SOE            -9.227      2.841  143.000  -3.248  0.00145 **
EDU            82.131     13.637  143.000   6.023 1.38e-08 ***
POPGR        -425.832    101.141  143.000  -4.210 4.49e-05 ***
Cdummy      -1515.530    935.249  143.000  -1.620  0.10734
NPDummy       627.368    906.484  143.000   0.692  0.49000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Two level model

```
Fixed effects:
             Estimate Std. Error       df t value Pr(>|t|)
(Intercept) 7872.8626 2315.5891   9.4136   3.400  0.00738 **
FDIPC         97.1562    5.1140 139.4599  18.998  < 2e-16 ***
SOE           -0.1297    2.9538 142.9073  -0.044  0.96504
EDU           10.6793   16.6574 141.7802   0.641  0.52248
POPGR       -245.4551  106.5454 142.6835  -2.304  0.02268 *
Cdummy      -651.8903  830.7197 139.7167  -0.785  0.43394
NPDummy       136.9396  800.8813 139.3536   0.171  0.86448
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Three level model

Case Study: Regional inequality of China

❑ Steps for Running the Workflow

Step 1: **Download data** from Google Drive data folder
Step 2: **Download workflow** from Google Drive workflow folder
Step 3: **Open KNIME** from local PC or China Data Lab Cloud Platform
Step 4: **Import KNIME workflow** file (wei2010.knwf)
Step 5: **Configure "Input Data"** for each table and figure
Step 6: Click **Run** function from the top menu
Step 7: **Display the outputs**:

- **Table 1** for inequality index from 1978 to 2007
- **Table 2** for growth rate of the provinces and regions of China from 1978-2007
- **Table 3** for multi-level regression model from 1978 to 2007
- **Image View** for GIS map visualization (Figure 1, 8a and 8b)
- **Table View** for Interregional inequality of GDP per capita 1978–2007 (Figure 5)
- **Image View** for Coefficient of variation (CV), Gini coefficient, and Theil index (Figure 6)
- **Image View** for Inequalities of intra region (CV) (Figure 7)

Case Study: Regional inequality of China

❑ Summary

- The original study results are mostly replicated using the workflow tool KNIME
- It is found that regional inequality at different geographical scales has shown various patterns, which is influenced greatly by the four municipalities.
- Globalization is the dominant mechanism causing regional inequality, since the important driving force of economic growth, the FDI, is extremely unevenly distributed among the three regions.

❑ Limitation

- The paper was published ten years ago, so several original data are difficult to get.
- The data source for 1978 constant prices is not mentioned in the manuscript, thus some GDP and GDPPC related results are not consistent.

User Account Application

Google Form

Project Title *

Short answer text

Research Plan (< 250 words) *

Long answer text

Data Requirements *

- [ ] COVID-19 data
- [ ] Demographics
- [ ] Econimics
- [ ] Innovation (e.g., Patent)
- [ ] Environment
- [ ] Health
- [ ] Human Migration
- [ ] Other...

Expected Outcomes *

- ( ) Publications
- ( ) Conference Presentations
- ( ) Workflows
- ( ) New Datasets
- ( ) Models
- ( ) Visualization
- ( ) Other...

Web Sites

China Data Lab

http://chinadatalab.net

China Data Online

http://china-data-online.com

Contact

office@chinadatacenter.net

Training Webinars on Workflow-based Data Analysis

- Co-sponsored by RMDS Lab and Future Data Lab

- 10/1/2020 Statistical Data Analysis With Workflows
- 11/6/2020 Analysis of Population Census Data & Demographic Change
- 12/3/2020 Analysis of Economic Census Data & Industrial Change
- 1/7/2020 The Integration of Data and Maps for Spatial Analysis
- 1/28/2021 Spatial Analysis of Patent Data
- 2/25/2021 Spatial Analysis of Health with Statistics, Census and GIS Data
- 3/25/2021 Spatial Analysis of Environment with Statistics, Census and GIS Data
- 4/29/2021 Spatiotemporal Analysis of Urban Development
- 5/27/2021 Spatiotemporal Analysis of Rural Development
- 6/24/2021 Human Mobility in Space and Time
- 7/22/2021 Spatiotemporal Analysis of Culture and Society
- 8/19/2021 Future Data Analysis: New Development and Directions

Training Webinars on Workflow-based Data Analysis

Registration: https://www.eventbrite.com/e/monthly-data-science-training-webinars-analytical-workflows-tickets-121260468325?aff=ebdssbonlinesearch