Workflow-based Intelligent Data Analysis

Tao Hu

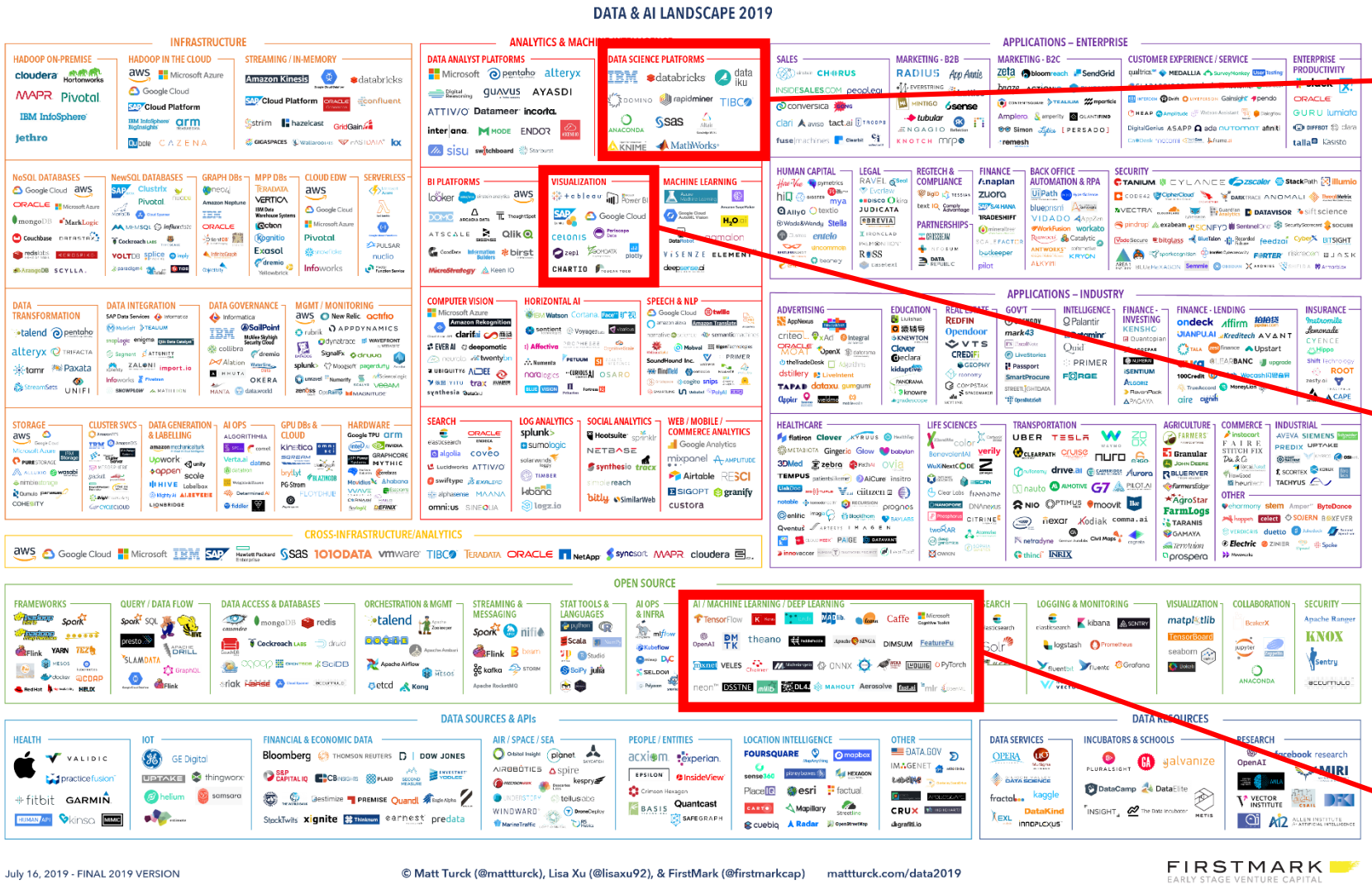
Center for Geographic Analysis, Harvard University

Topics

- Introduction on Workflow Tools
- Case Study on Spatiotemporal Data Visualization
- Case Studies on COVID-19 Pandemics
  - Human Mobilities and Pandemics Outbreak
  - Text mining-based Social Media Data Analysis
- Case Studies on Regional Science
  - Spatiotemporal hierarchy of regional inequality of China
  - Unconditional convergence of Chinese provinces
- Summary

WHY USING WORKFLOW?

DATA & AI LANDSRAPE 2019



http://mattturck.com/wp-content/uploads/2019/07/2019_Matt_Turck_Big_Data_Landscape_Final_Fullsize.png

DATA SCIENCE PLATFORMS

IBM, databricks, dataiku, DOMINO, rapidminer, TIBCO, ANACONDA, SAS, Altair Knowledge Works, KNIME, MathWorks

VISUALIZATION

tableau, Microsoft Power BI, SAP Lumira, Google Cloud, Periscope Data, celonis, zepl, ZOOMDATA, plotly, CHARTIO, TOUCAN TOCO

AI / MACHINE LEARNING / DEEP LEARNING

TensorFlow, Keras, torch, MADlib, learn, Caffe, Microsoft Cognitive Toolkit, OpenAI, DMTK, theano, PaddlePaddle, Apache SINGA, DIMSUM, FeatureFu, mxnet, VELES, Chainer, Michelangelo, ONNX, WEKA, LOOPS, PyTorch, neon, DSSTNE, mllib, DL4J, MAHOUT, Aerosolve, fast.ai, mlr, OpenML

WHY USING WORKFLOW?

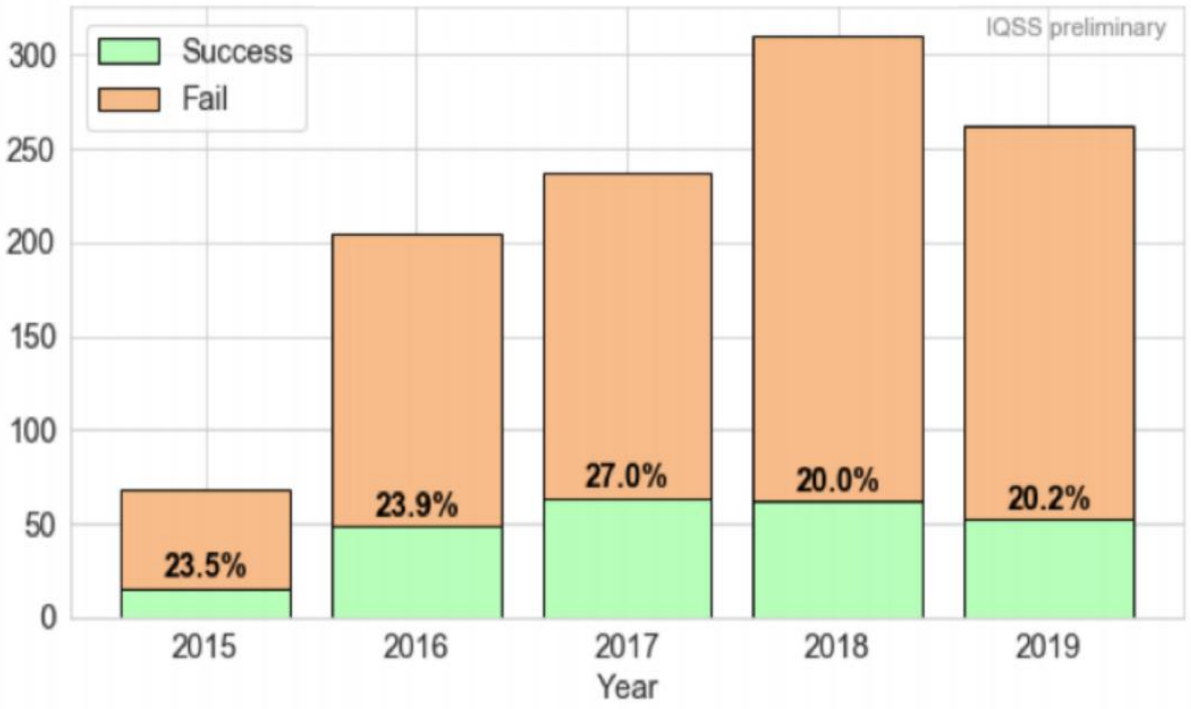
Replicable, Reproducible, and Expandable

https://www.synecticsglobal.com/en/resources-en/blogs/171-the-benefits-of-workflows-critical-decision-making-through-integrated-surveillance

WHY USING WORKFLOW?

The Dataverse® Project

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

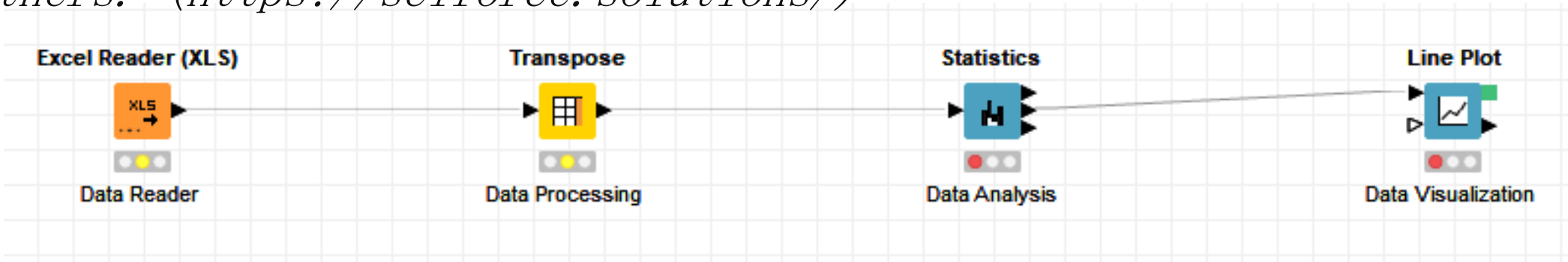


About ***84%*** re-executions are failed!!!

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

WHAT IS WORKFLOW?

A typical workflow for a data science project, including ***data preparation*** *(extraction, cleaning, and understanding),* ***analysis*** *(modeling),* ***reflection*** *(finding new paths), and* ***communication*** *of the results to others. (https://sciforce.solutions/)*



- **Easy to Operate**
- **Without Programming Knowledge**
- **Modulization**

- **Sharable**
- **Reproducible**
- **Replicable**
- **Expandable**

Existing Workflow Tools

Name	Category	Free	Open Source	Education Licence	Popularity	Open source community	Latest version	Website
KNIME analytics	***Data analytics***	✔	✔	✔	✔✔	✔✔	***4.0.2 (2019)***	***https://www.knime.com***
Talend Open Studio	Data analytics	✔	✔	✔	✔✔✔	✔✔	7.3 (2019)	https://www.talend.com/products/talend-open-studio/
Pentaho Kettle	ETL	✔	✔	✔	✔✔	✔✔✔	8.3 (2019)	https://community.hitachivantara.com/s/article/data-integration-kettle
GeoKettle	Spatial ETL	✔	✔	✔			2.5 (2015)	http://www.spatialytics.org/
RapidMiner Studio	Data analytics	Limited		✔	✔✔✔		9.4 (2019)	https://rapidminer.com/
Tableau Prep	ETL			✔	✔✔✔		2019.3.2 (2019)	https://www.tableau.com/products/prep
Alteryx Designer	Data analytics			✔	✔✔✔		2019.3 (2019)	https://www.alteryx.com/
Dataiku DSS	Data analytics	Limited		✔	✔✔		5.1(2019)	https://www.dataiku.com/

Existing Workflow Tools

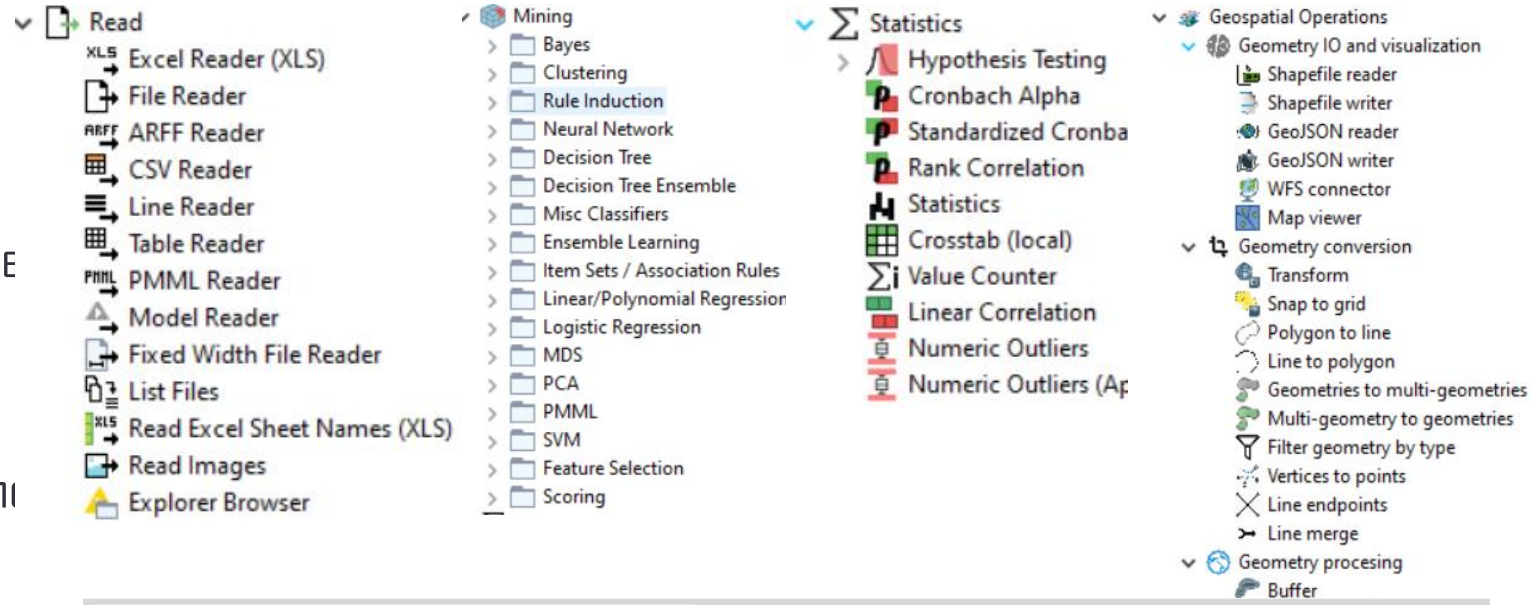
Name	Data Input/Output	Data Manipulation	Spatial Processing	Regression Analysis	Machine Learning	R	Python	Charts	Reporting
KNIME analytics	***Tabular data, shapefile***	✔✔	✔	✔✔✔	✔✔✔	✔✔	✔✔	✔✔	✔✔
Talend Open Studio	Tabular data, shapefile	✔✔✔	✔	✔	✔✔✔	✔	✔	✔✔	✔✔
Pentaho Kettle	Tabular data, shapefile	✔✔✔	✔	✔	✔	✔	✔	✔✔	✔✔
GeoKettle	Tabular data, shapefile	✔	✔✔✔					✔	✔
RapidMiner Studio	Tabular data, shapefile	✔✔	✔	✔✔✔	✔✔✔	✔✔	✔✔	✔✔✔	✔✔
Tableau Prep	Tabular data	✔✔						✔	
Alteryx Designer	Tabular data, shapefile	✔✔✔	✔✔✔	✔✔	✔✔	✔✔	✔✔	✔✔	✔✔
Dataiku DSS	Tabular data, shapefile	✔✔✔	✔	✔✔✔	✔✔✔	✔✔	✔✔✔	✔✔✔	✔✔

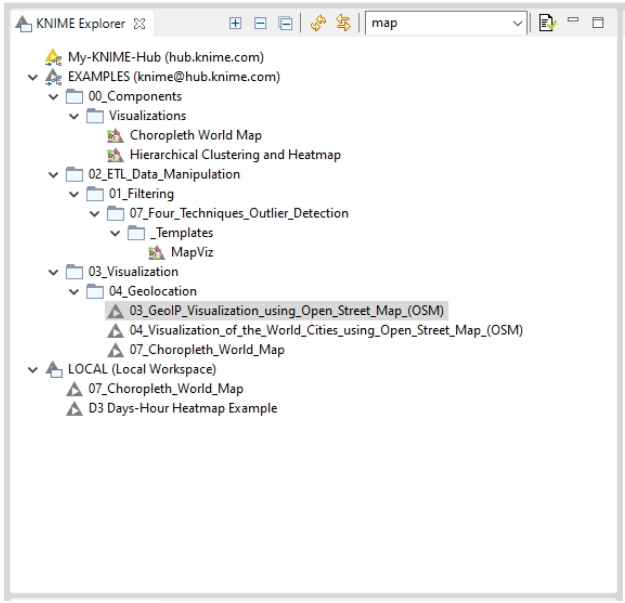
Introduction on Workflow Tool KNIME

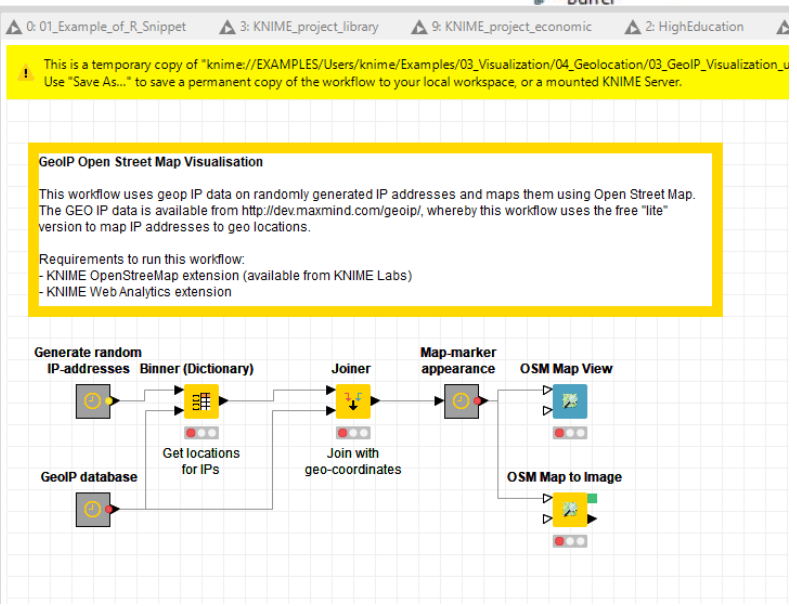
Open for Innovation
KNIME

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
- Have a decent size community that supports Q&A.

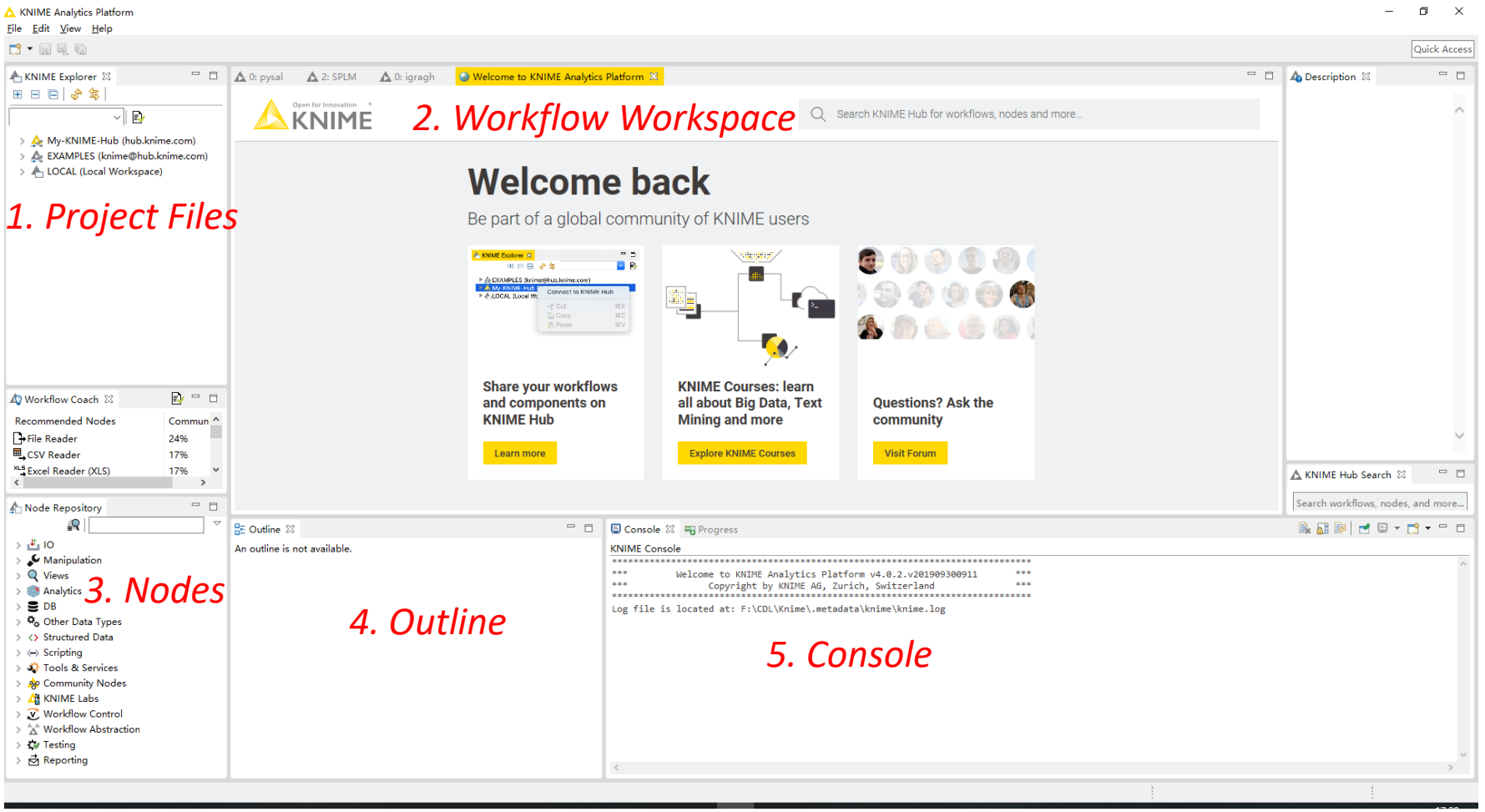






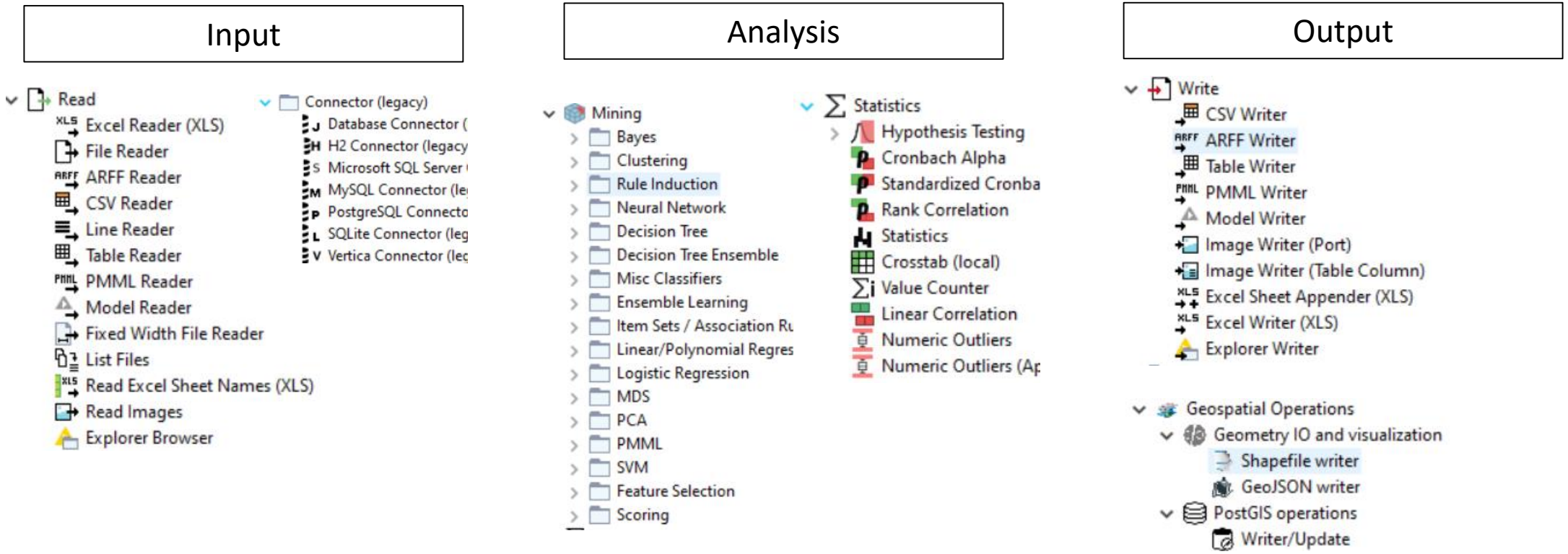
Introduction on Workflow Tool KNIME

- ❑ **Download KNIME** https://www.knime.com/downloads/download-knime
- ❑ **KNIME Interface**



Introduction on Workflow Tool KNIME

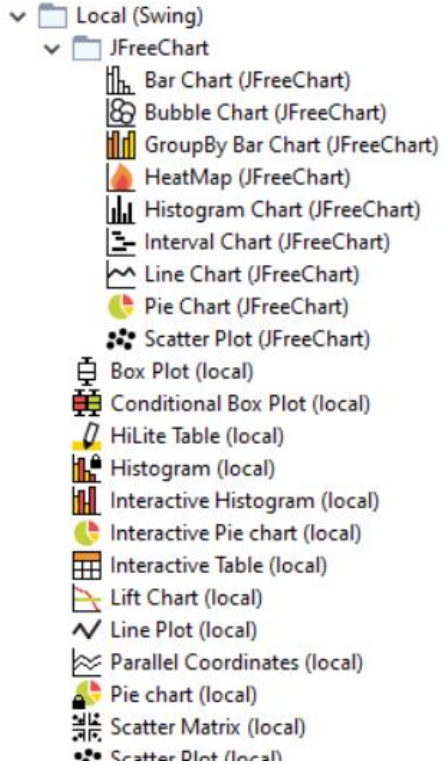
❑ Node Repository



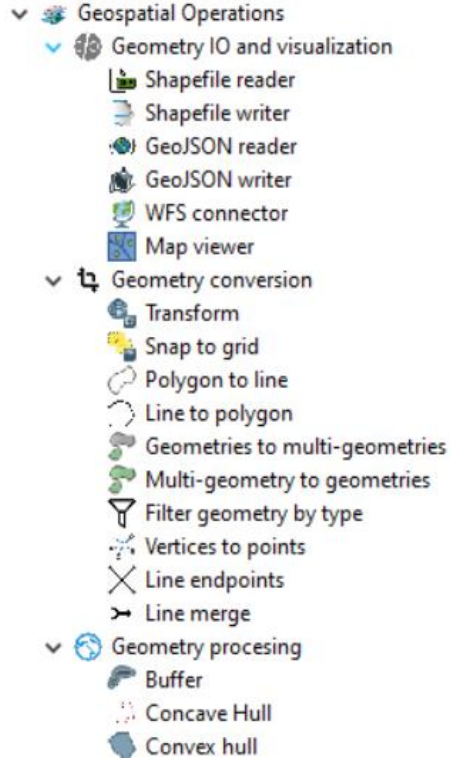
Introduction on Workflow Tool KNIME

❑ Other Nodes

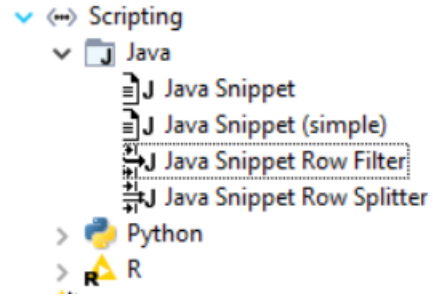
Visualization



Geospatial Operations



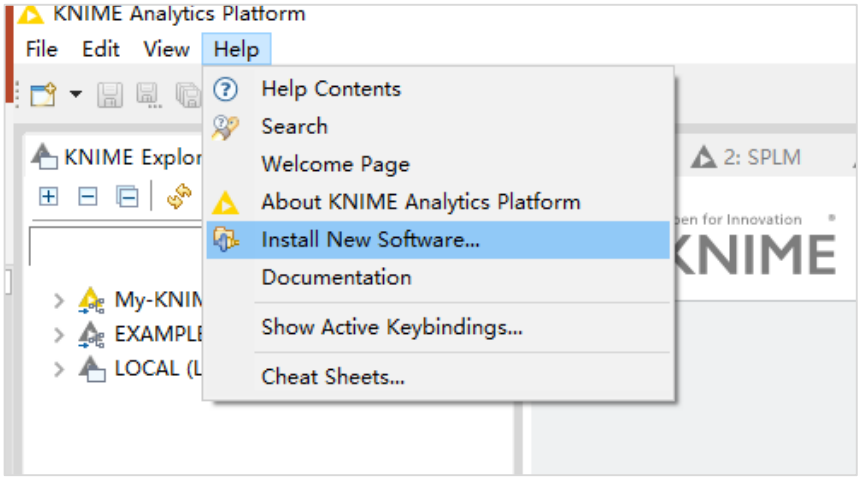
Scripting



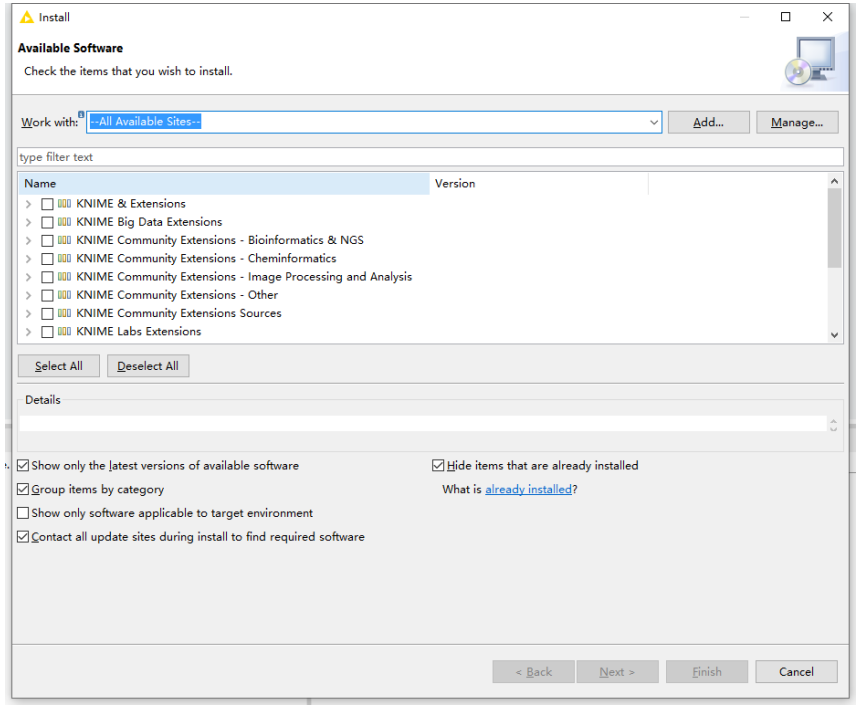
Introduction on Workflow Tool KNIME

❑ Install Extensions (R/Python)

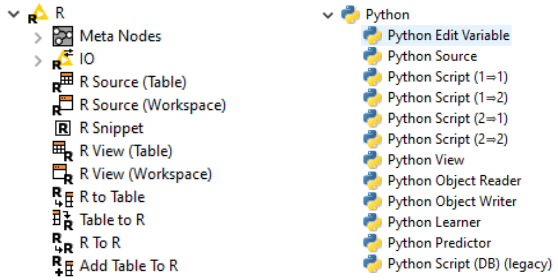
1. Help->Install New Software



2. Type Filter Text (R/Python)



3. Install and Restart KNIME. You will see new nodes in the Node Repository

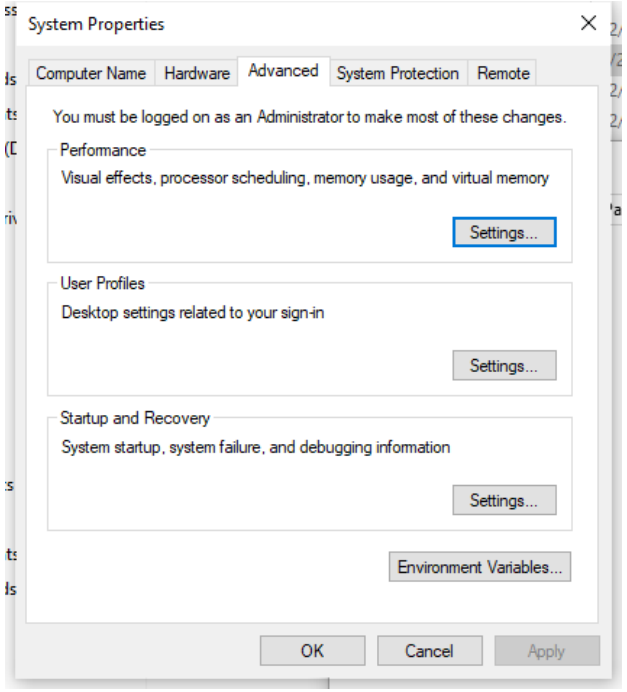


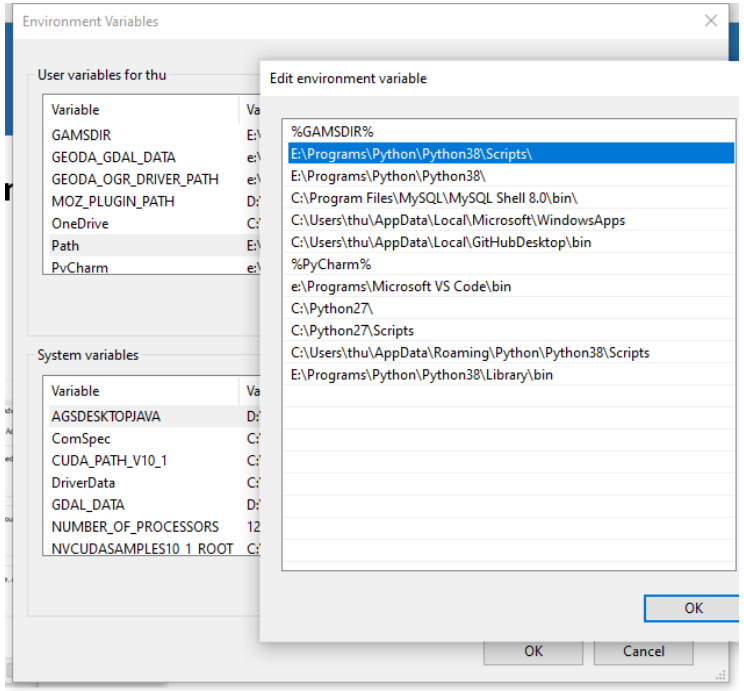
Introduction on Workflow Tool KNIME

❑ Environment Settings for R/Python

1. Install Anaconda or PYTHON (>3.7 is preferred)

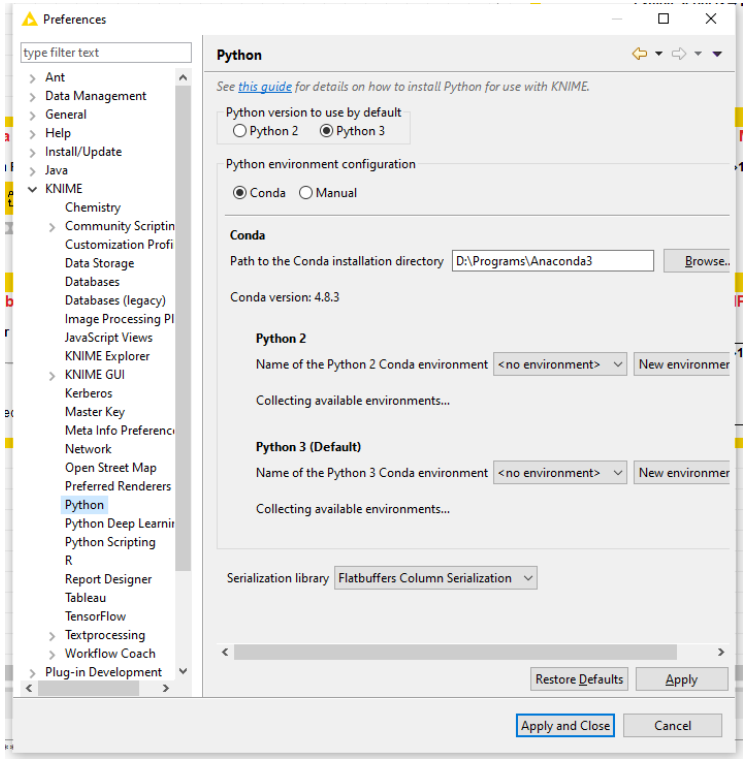
2. Set Path in the OS





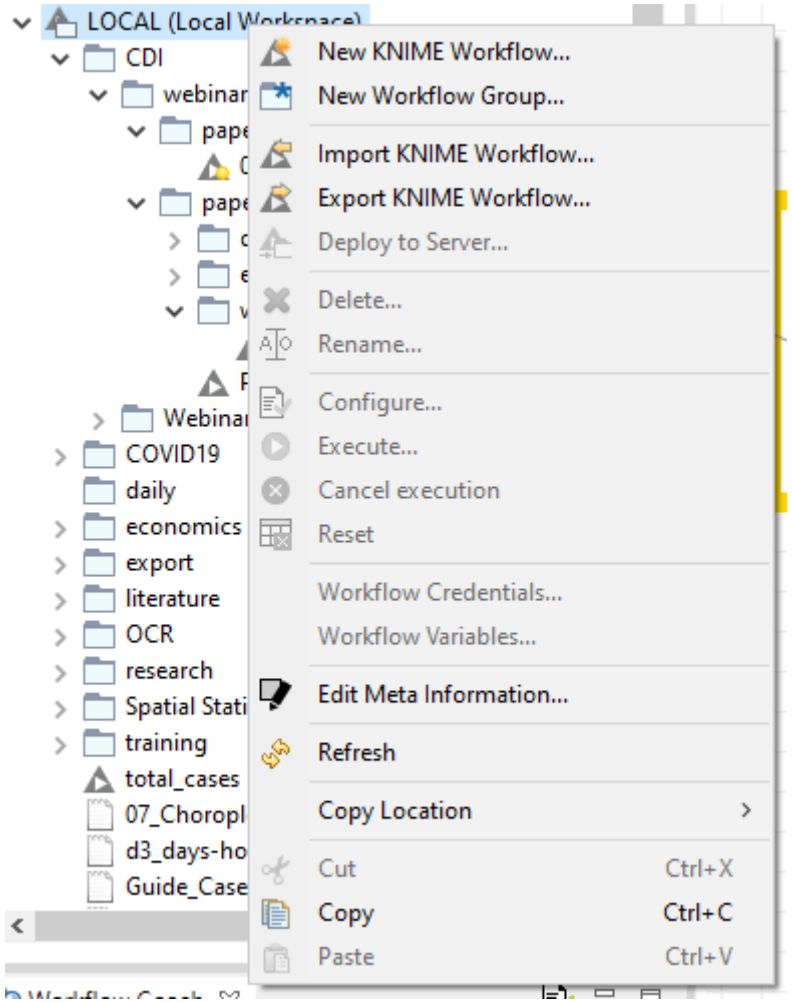
3. Set Path in the OS

File-> Preference

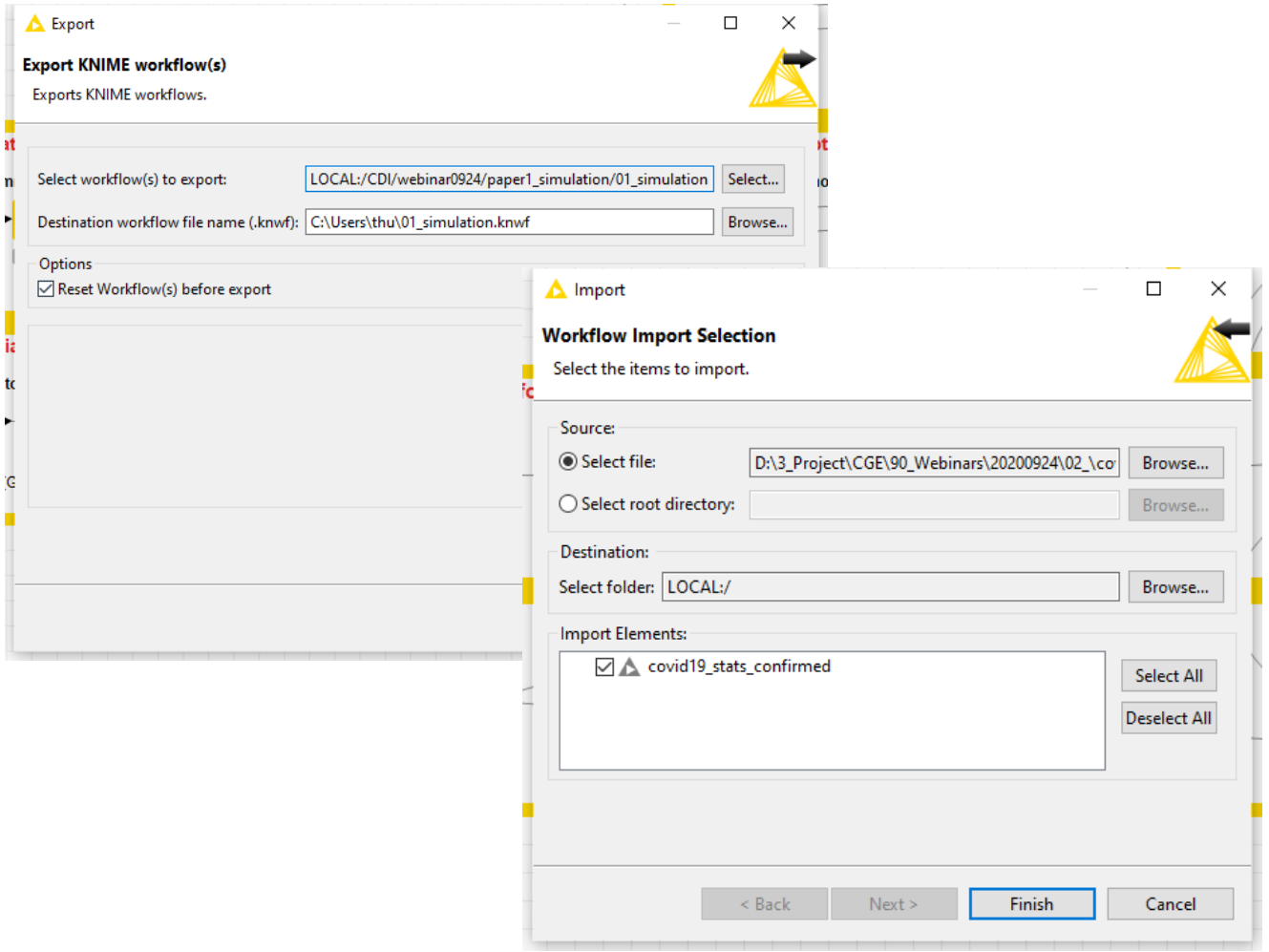


Introduction on Workflow Tool KNIME

❑ Create a workflow



❑ Import/Export an existing workflow



Outline

- Introduction on Workflow Tools
- Case Study on Spatiotemporal Data Visualization
- Case Studies on COVID-19 Pandemics
  - Human Mobilities and Pandemics Outbreak
  - Text mining-based Social Media Data Analysis
- Case Studies on Regional Science
  - Spatiotemporal hierarchy of regional inequality of China
  - Unconditional convergence of Chinese provinces
- Summary

Case Study One: Spatiotemporal Data Visualization

Objectives

- Develop a prototype for spatiotemporal data visualization with dynamic maps
- Build a workflow for efficient China Data Online data visualization
- Make the data analysis and visualization replicable, reproducible and expandable
- Lower the barrier for spatiotemporal data visualization

Data Sources

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

Data

- Yearly Population index by province of Mainland China (1990-2018) [Link]
- Province boundary basemap of Mainland China [Link]

Case Study One: Spatiotemporal Data Visualization

❑ Methodology

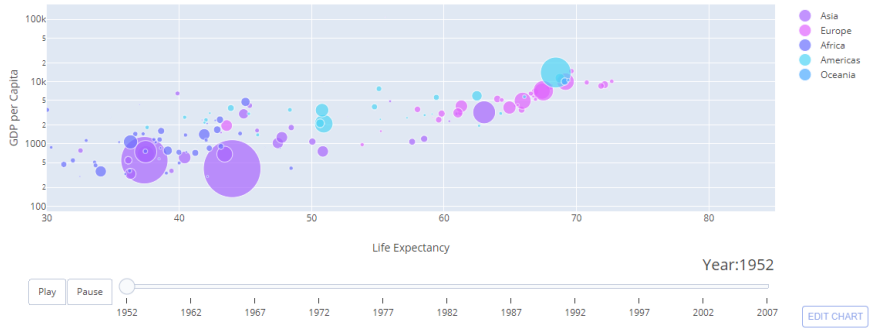
Plotly Python Open Source Graphing Library https://plotly.com/python/

Mapbox Choropleth

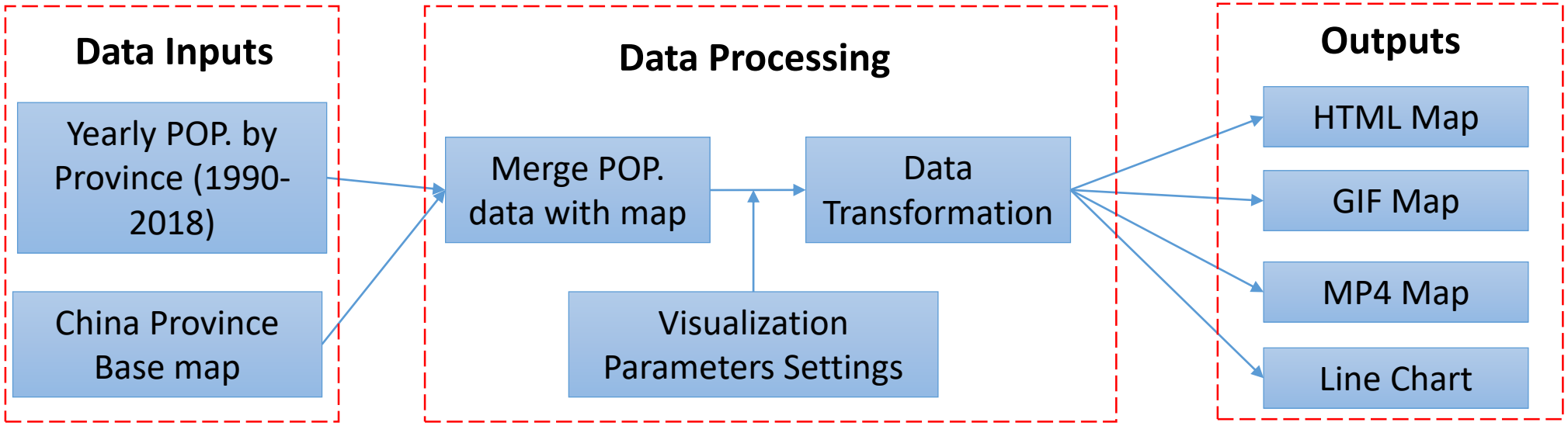
Lines on Mapbox

Filled Area on Maps

Bubble Maps

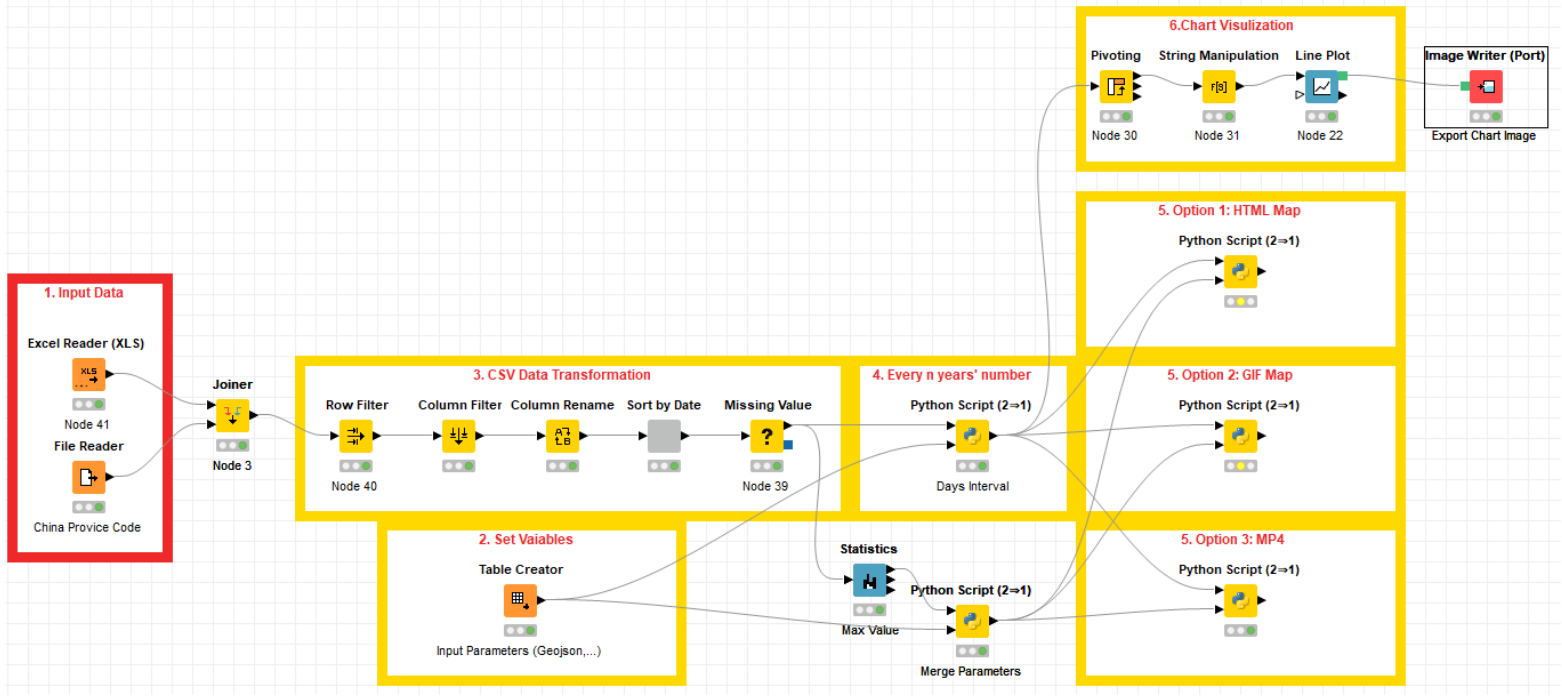


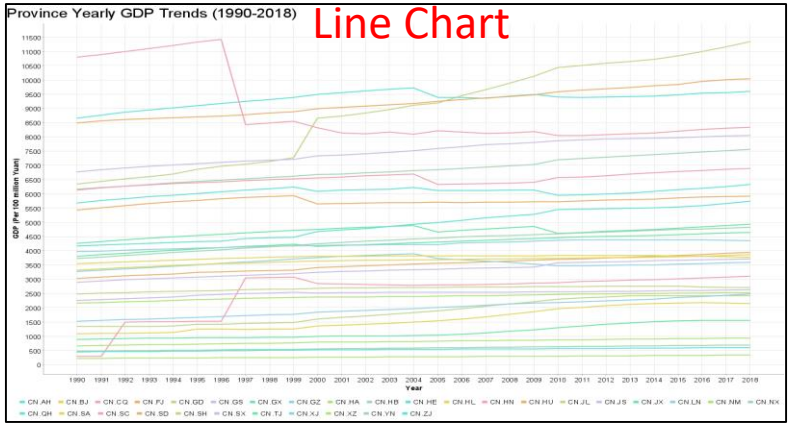
❑ Flowchart

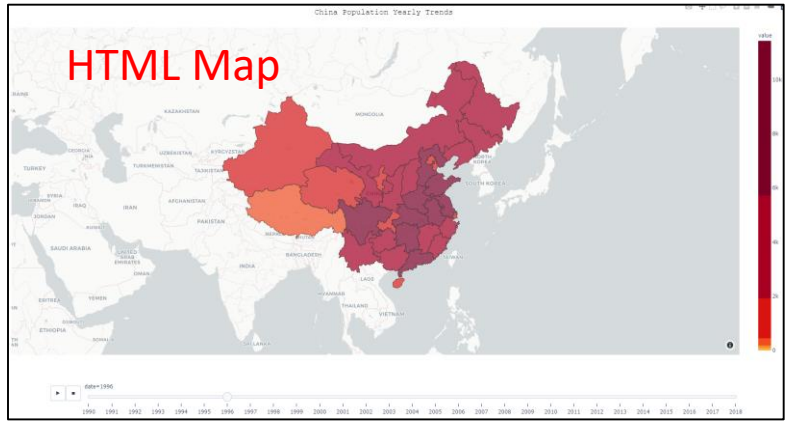


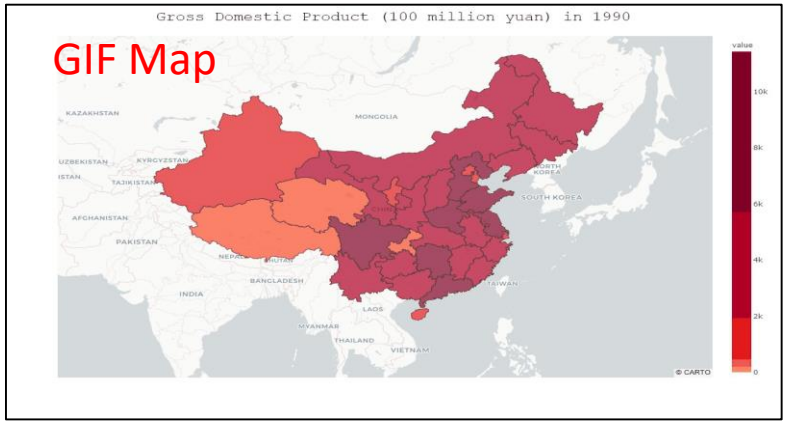
Case Study One: Spatiotemporal Data Visualization

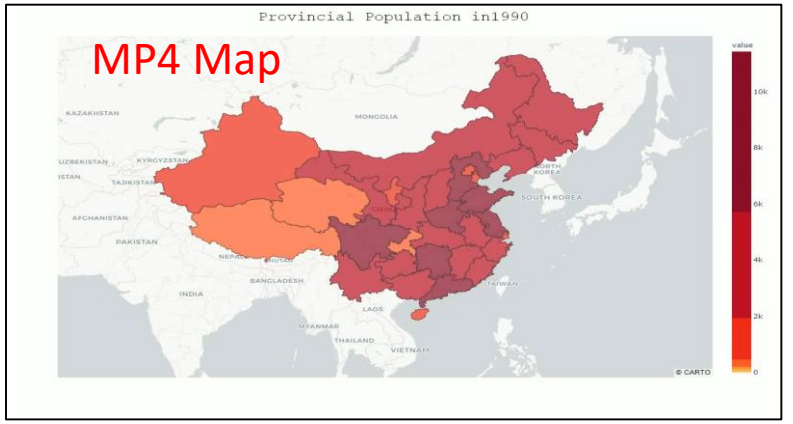
❑ Implementation by KNIME











Case Study One: Spatiotemporal Data Visualization

❑ 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 (in building)
Step 4: **Import Knime workflow** file (pop_maps_china.knwf)
Step 5: **Launch the workflow** by double clicking the imported workflow
Step 6: **Configure "Input Data"** by selecting "China_Province.csv" and "8.POP.xlsx"
Step 7: **Configure "Set Variables"**

- Select Geojson file 'China_admin_1.json' path;
- Set Output html, GIF, and MP4 file path;
- Set Option: 1->html; 2->GIF; 3->mp4
- Set Title: name of the html file
- Set Days interval: default 3;

Step 8: Configure path in "Image Write" node
Step 9: Click **Run** function from the top menu
Step 10: **Display the outputs**:

- HTML Map
- GIF Map
- MP4 Map
- Chart

*** Please install plotly, imageio, and opencv python library before running the workflow
> pip install plotly, plotly-orca
> pip install imageio
> pip install opencv-python

Outline

- Introduction on Workflow Tools
- Case Study on Spatiotemporal Data Visualization
- Case Studies on COVID-19 Pandemics
  - Human Mobilities and Pandemics Outbreak
  - Text mining-based Social Media Data Analysis
- Case Studies on Regional Science
  - Spatiotemporal hierarchy of regional inequality of China
  - Unconditional convergence of Chinese provinces
- Summary

Human Mobilities and Pandemics Outbreak

Objectives

- To explore the human mobilities from Wuhan before and during the pandemic
- To estimate the correlation between daily COVID-19 cases and human mobilities in the target cities
- To estimate the correlation between accumulated COVID-19 cases and human mobilities in the target cities

Data Sources

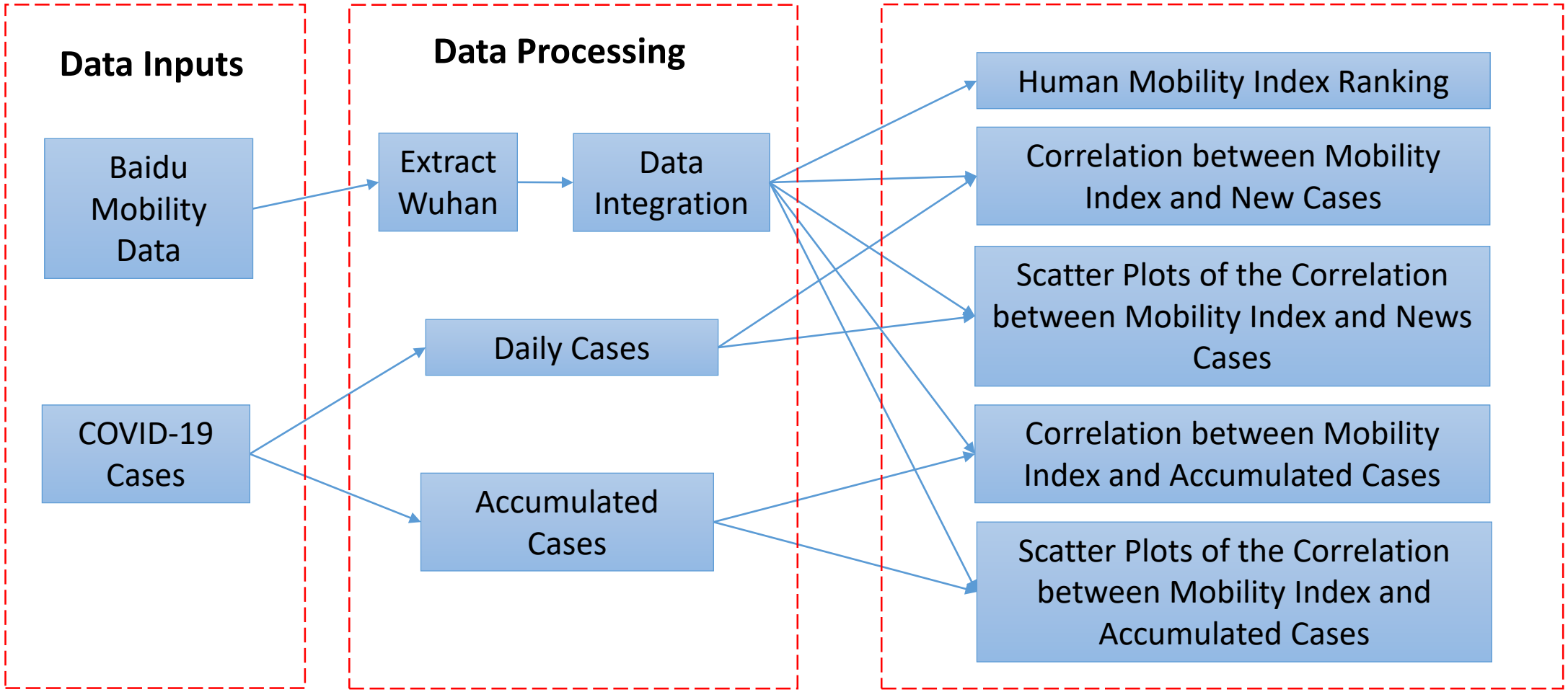
- China Data Lab at Harvard Dataverse (http://china-data-online.com)

Data

- City-level COVID-19 cases in China [Link]
- Baidu Mobility Index in China [Link]

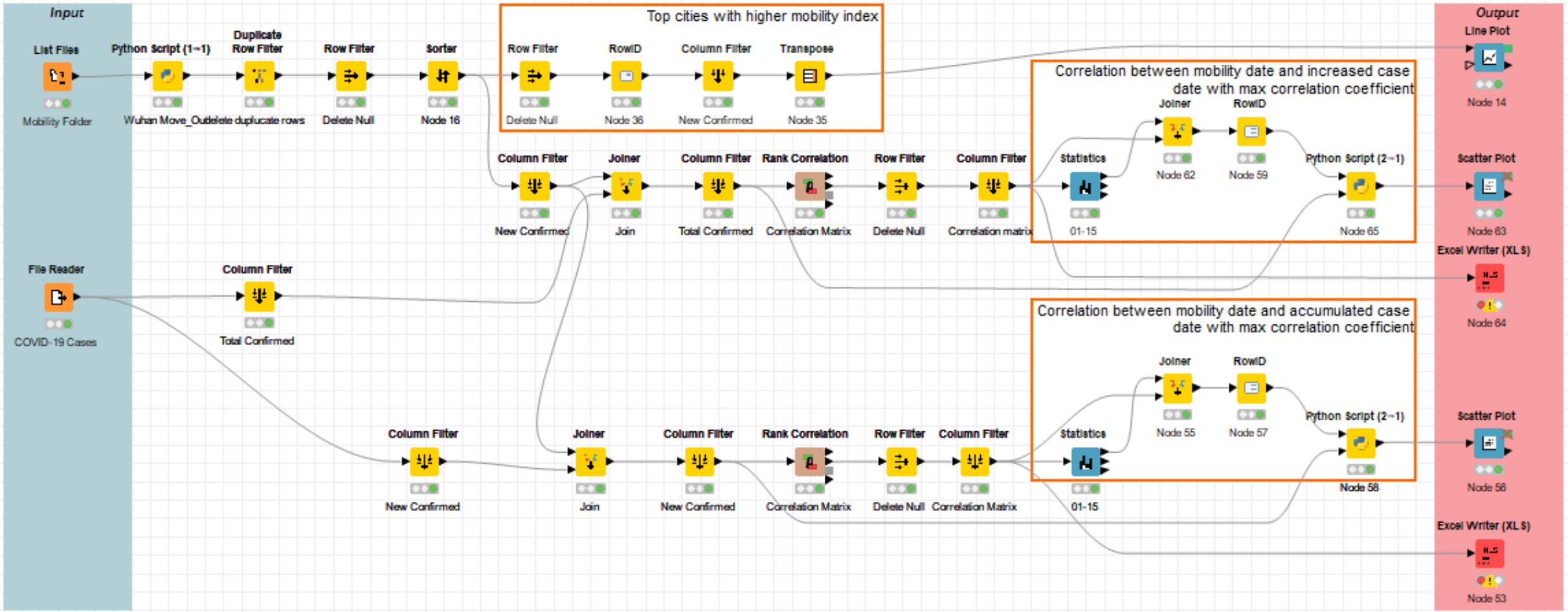
Human Mobilities and Pandemics Outbreak

❑Flowchart



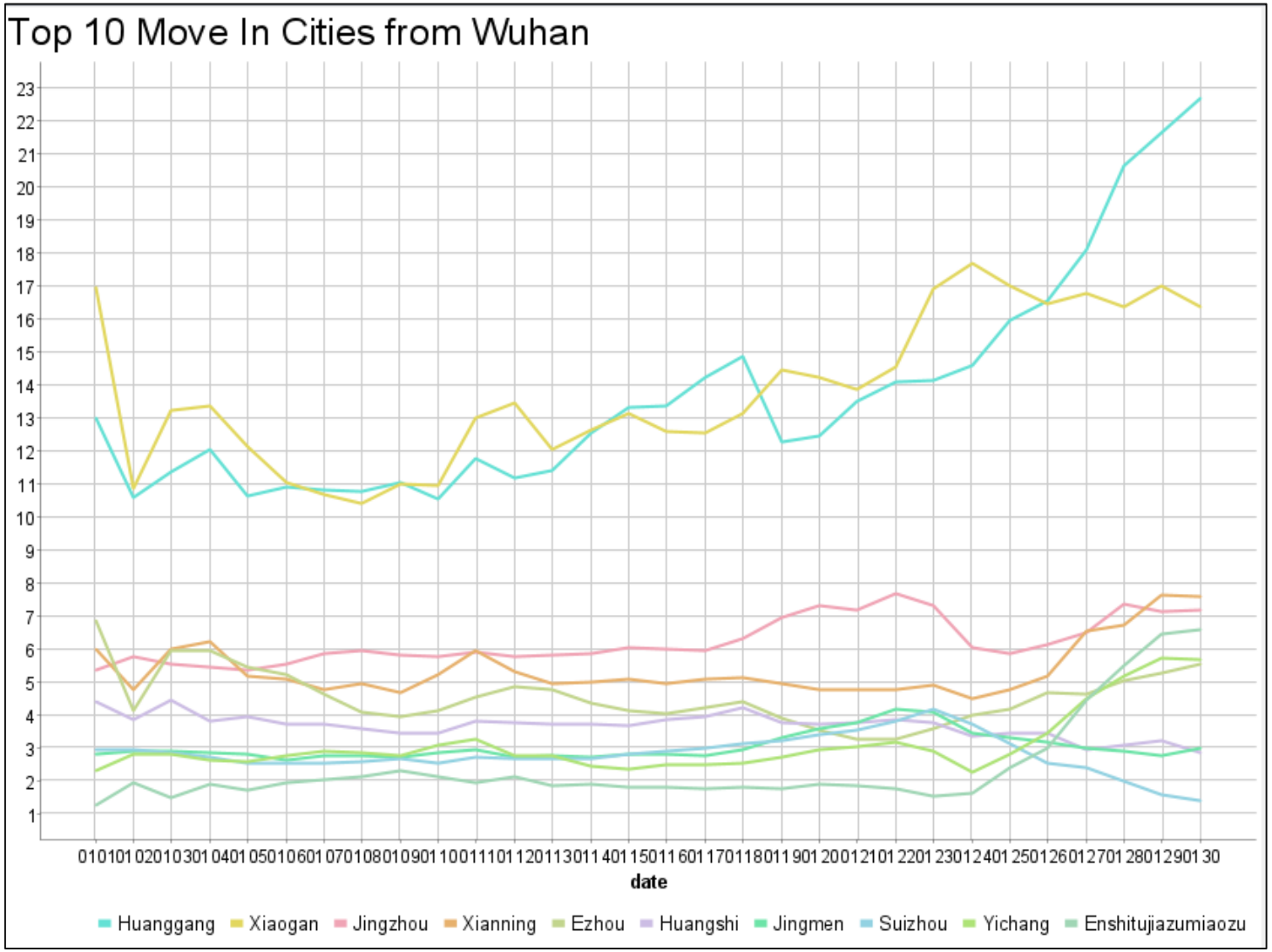
Human Mobilities and Pandemics Outbreak

Workflow Implementation by KNIME



Human Mobilities and Pandemics Outbreak

❑Top 10 Move in Cities from Wuhan



Human Mobilities and Pandemics Outbreak

❑Correlations Estimate

The correlation coefficients of mobile population from Wuhan and new confirmed cases of target cities

Row ID	0115	0116	0117	0118	0119	0120	0121	0122	0123	0124	0125	0126	0127
0130	-0.02477357698848...	0.10339122415866...	0.022502615888261...	0.07011226031854...	-0.0162608704266...	-0.01829622057184...	-0.12295371154331...	-0.12601619834551...	-0.18582431425223...	-0.16679052563733...	0.07299088181008...	0.15893456504541...	0.17786521323510...
N_C_0119	0.09285768234338783	0.05532438345064...	0.049394291348066...	0.04544643326322...	0.03757406173858...	0.03556924649178405	0.0335995368211173	0.041533242925616...	0.04349267815368113	0.08105422321951...	0.06927730955845...	0.06934560096131...	0.05369591723504...
N_C_0120	0.17692706665821062	0.11881930561105...	0.12118921530686376	0.2359964520287542	0.08004795832572...	0.07213140539570778	0.05886706968284388	0.05285287976107428	0.05342001228113258	0.12986572862979...	0.11542885286383...	0.08705078511883...	0.04887681786013...
N_C_0121	0.16264240802893054	0.06564241546648...	0.02589802025697616	0.15898929379992...	0.04347719329990...	0.09679612722764427	0.06793038326579716	0.049674149141471...	-0.03301008066604...	0.07012663726719...	0.04603484769571...	0.10418897069183...	-0.11702202047283...
N_C_0122	0.1274614414397117	0.14467257377933...	0.15072987521105438	0.08750596616854...	0.06168724411958...	0.13536779869815832	0.1413591005569477	0.14849664034433752	0.1855545891955449	-0.0032648634103...	0.0807393741816597	-0.09567585246904...	-0.06730277191853...
N_C_0123	0.12094027626190487	0.09203177871250...	0.15613638069644778	0.17641813348757...	0.11319320082774...	0.06952461789157581	0.07098477547948898	0.06188941958765155	0.007802619799675...	0.00184096833775...	0.03294809511896...	0.06075066169214...	0.136484224141464
N_C_0124	0.3382522832151056	0.2063297917117872	0.221540172904987	0.25407529977310...	0.22346310420578...	0.18931427572820073	0.2259128070428507	0.2648022482432988	0.1908725519165474	0.20145608636506...	0.1441957073562417	0.25967059561049...	0.18771930582918...
N_C_0125	0.5160158717058445	0.33856886265275...	0.40579511600065676	0.36146324176851...	0.332907867803027	0.35014556123631796	0.29374187306388605	0.3038995560118951	0.32493391925628046	0.22730425318917...	0.21699026645183...	0.22251726298074...	0.10267663601009...
N_C_0126	0.5628925230371874	0.38939441391661...	0.3715145283159348	0.34797566780061...	0.3733560714476319	0.37227176161862774	0.41141064162439067	0.4152788019135796	0.31924403766198584	0.37765333088975...	0.19882947454062...	0.2364438261171659	0.07557627725924...
N_C_0127	0.6320335646965454	0.4447103307613721	0.48392575589501274	0.4988687834107277	0.45146484678060...	0.4519082127455260 6	0.45332674670930134	0.44628966096227657	0.38734619534135833	0.3513025081982499	0.31107697644958...	0.2632509982273468	0.12829698459810...
N_C_0128	0.6748836253324684	0.47419462011400...	0.5452078175650954	0.4356508461530392	0.43866997062605...	0.44131731495307336	0.3217150436679323	0.3425118496920029	0.4305759331880146	0.4099289944636582	0.2905609699296254	0.2650226424423889	0.22146623199184...
N_C_0129	0.7056550361654592	0.47150182912957...	0.5819758343791975	0.5595910700875395	0.4243739432444717	0.425996818748818 5	0.39328827440018577	0.4422137500641510 3	0.3456106031332887	0.3226042858810695	0.2506031500543518	0.31361223037217...	0.1359589130886309
N_C_0130	0.663659197157656	0.5156421923375882	0.5641315184121813	0.5353552391826458	0.4581138752332417	0.4575118433129131 7	0.3565210613651878 5	0.384742251804837 5	0.3622362764923646	0.4296029203464742	0.3114274459924828	0.31949889432118...	0.13983279655735...
N_C_0131	0.6208432399753827	0.45654543243666...	0.40108503743589907	0.528902944967682	0.3928494161898712	0.36468261233572336	0.33683362301759817	0.3181811476120938	0.2642007434340939	0.2483943255392632	0.20212825882285...	0.16507280400648...	0.19899287772210...

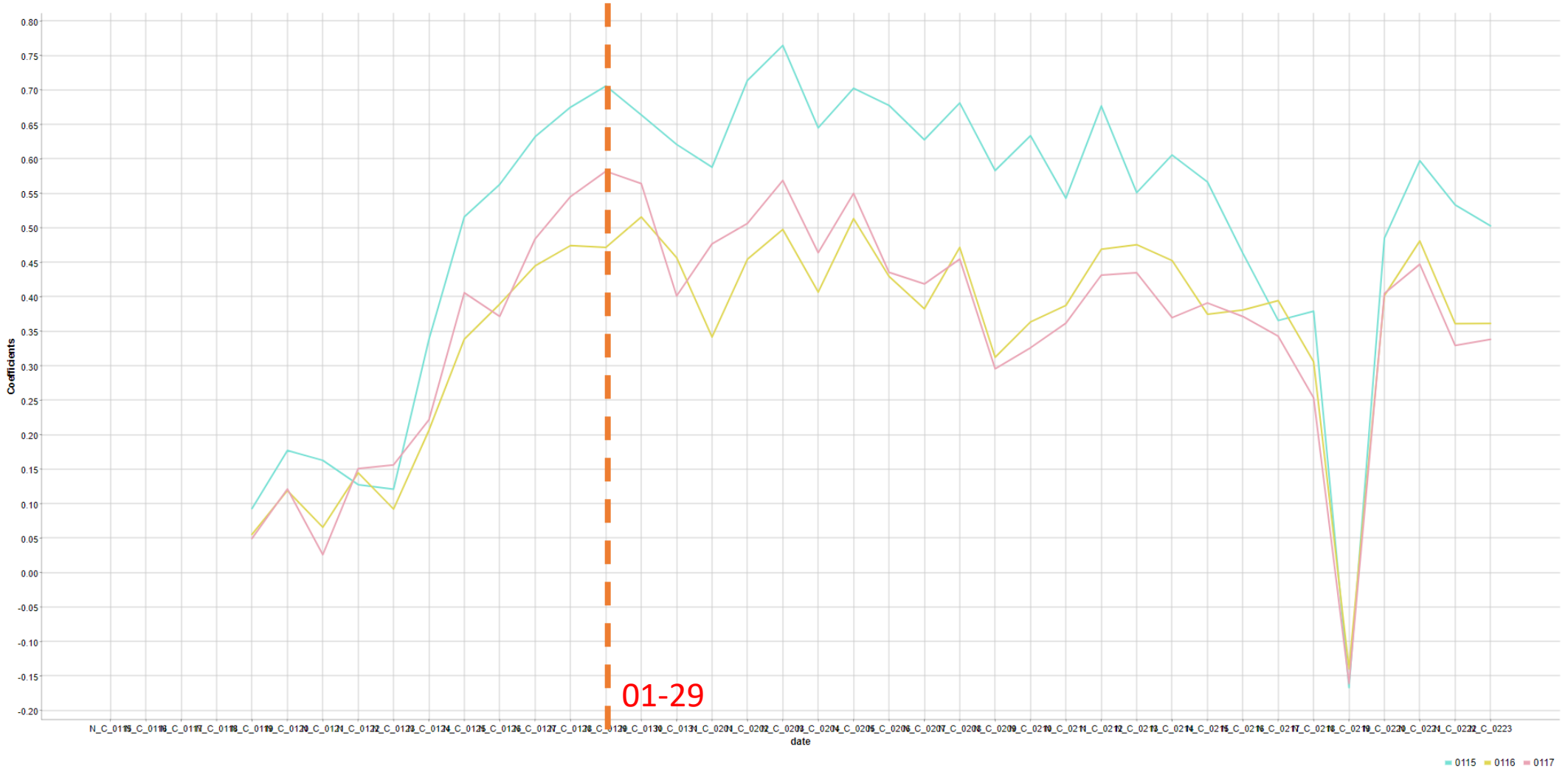
The correlation coefficients of mobile population from Wuhan and accumulated confirmed cases of target cities

Row ID	0115	0116	0117	0118	0119	0120	0121	0122	0123	0124	0125	0126	0127
T_C_0119	0.09285768234338...	0.05532438345064...	0.049394291348066...	0.045446433263226...	0.03757406173858...	0.03556924649178...	0.0335995368211...	0.041533242925616...	0.04349267815368...	0.08105422321951...	0.06927730955845...	0.06934560096131...	0.053695917235048...
T_C_0120	0.17692706665821...	0.11881930561105...	0.12118921530686376	0.2359964520287542	0.08004795832572...	0.07213140539570...	0.0588670696828...	0.05285287976107428	0.05342001228113...	0.12986572862979...	0.11542885286383...	0.08705078511883...	0.04887681786013008
T_C_0121	0.20607434958923...	0.0991410680273447	0.05859136248420173	0.2060804953725627	0.07214144998727...	0.12713585497778...	0.0976134143147...	0.07830397279762698	-0.00697457978318...	0.1032839225981965	0.07539756299799...	0.1323203593787733	-0.09884827930535317
T_C_0122	0.1180211994541 4...	0.1136965359312927	0.1147837447763988	0.18158574677714923	0.06927532183566...	0.1336497872956413	0.1369365771734...	0.13425230513446057	0.12520829447206...	0.0108712542939 1...	0.07768244329450...	-0.07862474897263...	-0.0920284027024515
T_C_0123	0.1540953100263 2...	0.1303512550518 0...	0.16799074400920025	0.16477343764123042	0.10798805042872...	0.1455969294147547	0.1311173654486...	0.12151573141691216	0.10753497220707...	0.0127750033987 6...	0.05550315727924...	0.01255986904876...	0.08598230067010071
T_C_0124	0.2781500562638 8...	0.2128512864736 9...	0.2560725738195611	0.29940267178037244	0.2033605158961123	0.2056150636920 7...	0.2192349128160...	0.24114289872576114	0.1816326442953 5...	0.1304604558731 5...	0.1148815339730 5...	0.1524546109977 9...	0.13024971742245947
T_C_0125	0.5120308784850189	0.3903995202544 0...	0.43880433005377295	0.47587835724058186	0.3764554654899543	0.3881892373839523	0.3620460273108...	0.3851371972799914	0.3439472485387 7...	0.2760329071062812	0.2403501906976 1...	0.2488363762331 9...	0.18788870256188914
T_C_0126	0.6215102895009349	0.4808960345528963	0.5202030987141322	0.5302992117208377	0.4633466042858716	0.46484378067476...	0.4391003044114...	0.4533281231197825	0.3900214187670024	0.3381933502512953	0.2512256017222285	0.2874456207350953	0.140087663723648
T_C_0127	0.6559803244669664	0.4920878258897 2...	0.5324869392320378	0.5613955836855012	0.4811089374312 0...	0.4835947865166392	0.4645626049462...	0.4737342886113417	0.4058564669997724	0.3600369952402858	0.2766673850248282	0.2912668736316 9...	0.13540619308751373
T_C_0128	0.6949468070274818	0.4947421495564 6...	0.5538442136914342	0.5393156383468932	0.4987137202253619	0.5109222339487671	0.4491534430360...	0.4620218547018626 3	0.4355891906497 3...	0.3744521470358096	0.2801947376454534	0.3022338240782 5...	0.16322404845538183
T_C_0129	0.7198909402557722	0.5045909506790396	0.5888412791388346	0.5642222091166793	0.4873632141255376	0.4963197041694 2...	0.4379561904069...	0.4717450126814305	0.4247746405256 1...	0.3666601906781 7...	0.2631712144418 4...	0.3170471428818 9...	0.163515831141515
T_C_0130	0.7164880137379463	0.5066780557161399	0.5936447291599082	0.5640319611373578	0.4821625458116816	0.491619575655582	0.4193168130126...	0.456240387400816	0.4197735204091 6...	0.3828372216701183	0.2764522262258592	0.3319125883636172	0.1523885071328371
T_C_0131	0.7461267565265629	0.5360045435904246	0.5941511965694835	0.6103085657311889	0.4984116378732 1...	0.4955067022124929	0.4282408913487...	0.4430814257195369	0.4027150568577605	0.3569749932833 8...	0.2723473036710191	0.300244131537024	0.1512426765588442

Human Mobilities and Pandemics Outbreak

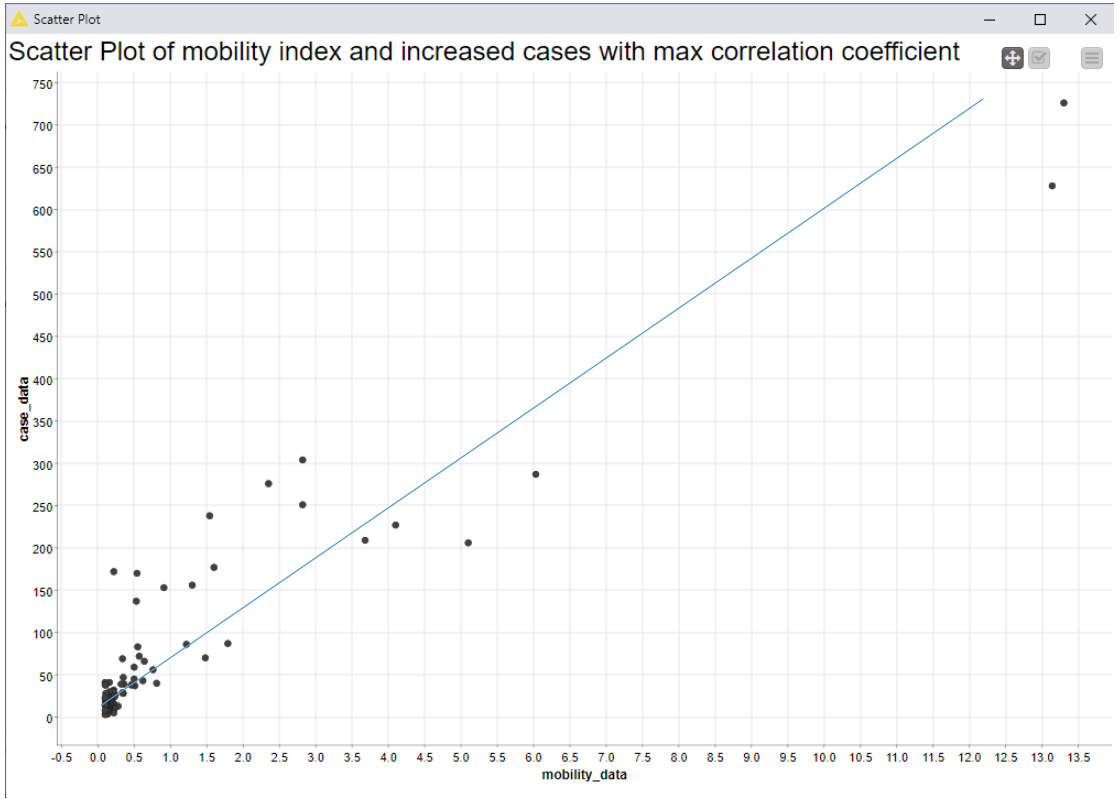
❑Correlation Coefficients Visualization

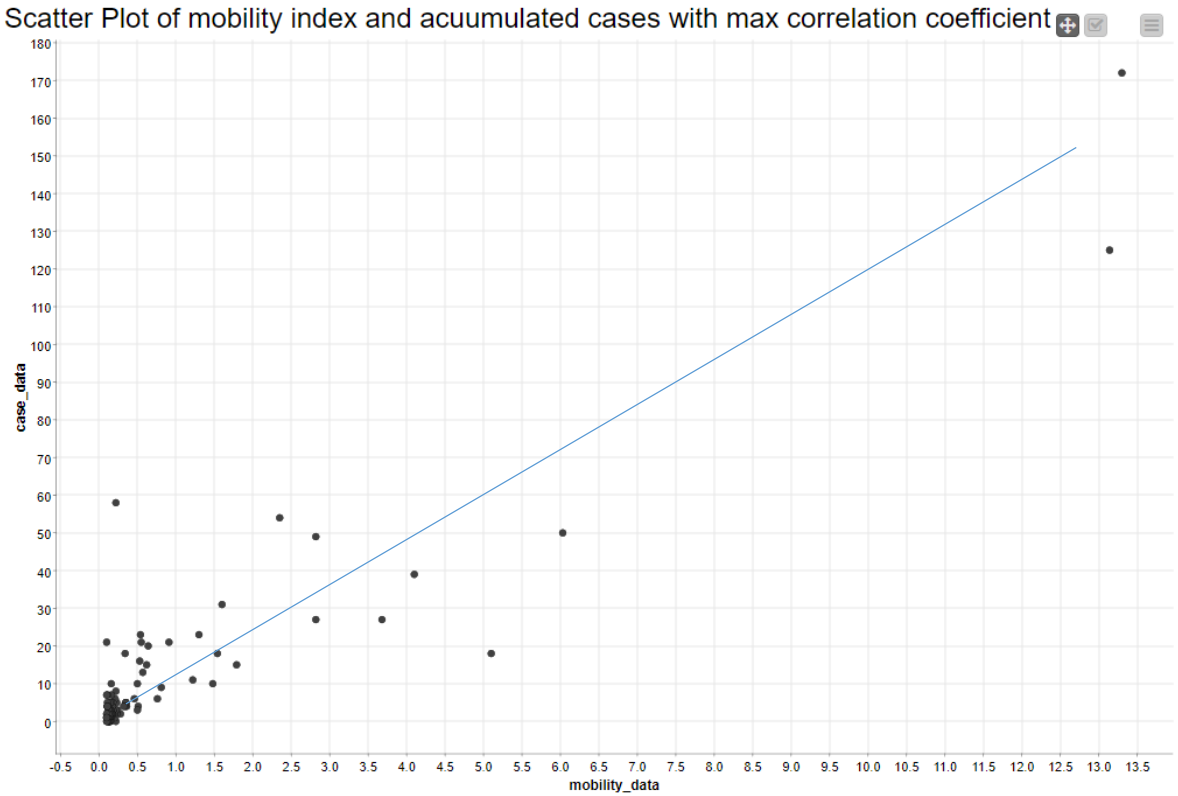
The line charts of correlation coefficients of mobile population from Wuhan and new confirmed cases of target cities



Human Mobilities and Pandemics Outbreak

Mobility index values in January 15 and increased (accumulated) cases in January 29 with max correlation coefficient





Human Mobilities and Pandemics Outbreak

❑ Steps for Running the Workflow

- Website: http://chinadatalab.org
- File Folder: Public/Covid/Workflows/05_social_media
  - Workflow File: social_media.knwf
  - Presentation file: Introductions to Social Media Data_EN.pptx
  - Input Data:
    - ✓ Public/Covid/Workflows/04_mobility_anlasis/data/City_Confirmed_Map_0115_0318.csv
    - ✓ Public/Covid/Workflows/04_mobility_anlasis/data/OUT/
  - Output Data:
    - ✓ maxtrix1.csv: correlation matrix between daily mobility data and daily increased cases
    - ✓ matrix2.csv: correlation matrix between daily mobility data and daily accumulated cases

Text mining-based Social Media Data Analysis

Objectives

- To explore the posted COVID-19 tweets in the United States
- To investigate the hashtag topics in the COVID-19 tweets
- To investigate the topics in the COVID-19 tweets content
- To compare the topics differences before and after March 17th, 2020

Data Sources

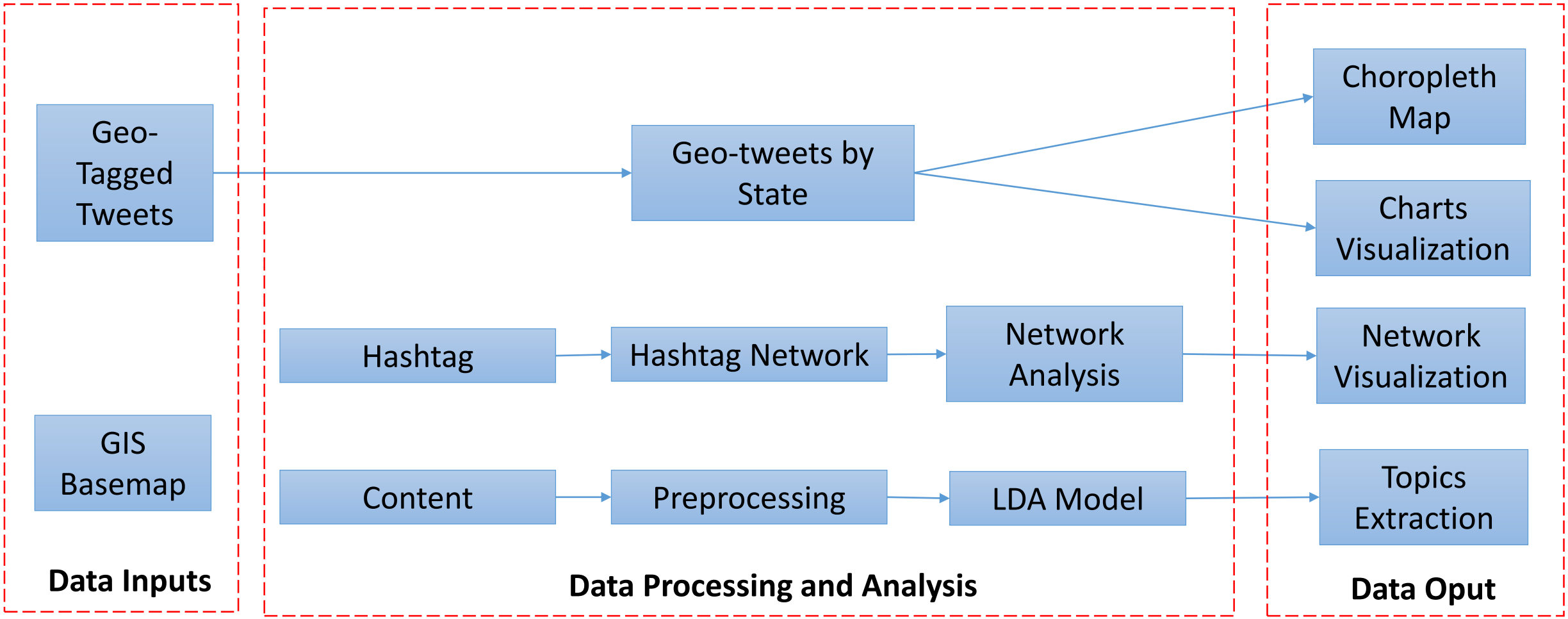
- Twitter

Data

- Geo-Tagged Tweets collected by the Center for Geographic Analysis at Harvard University [Link]

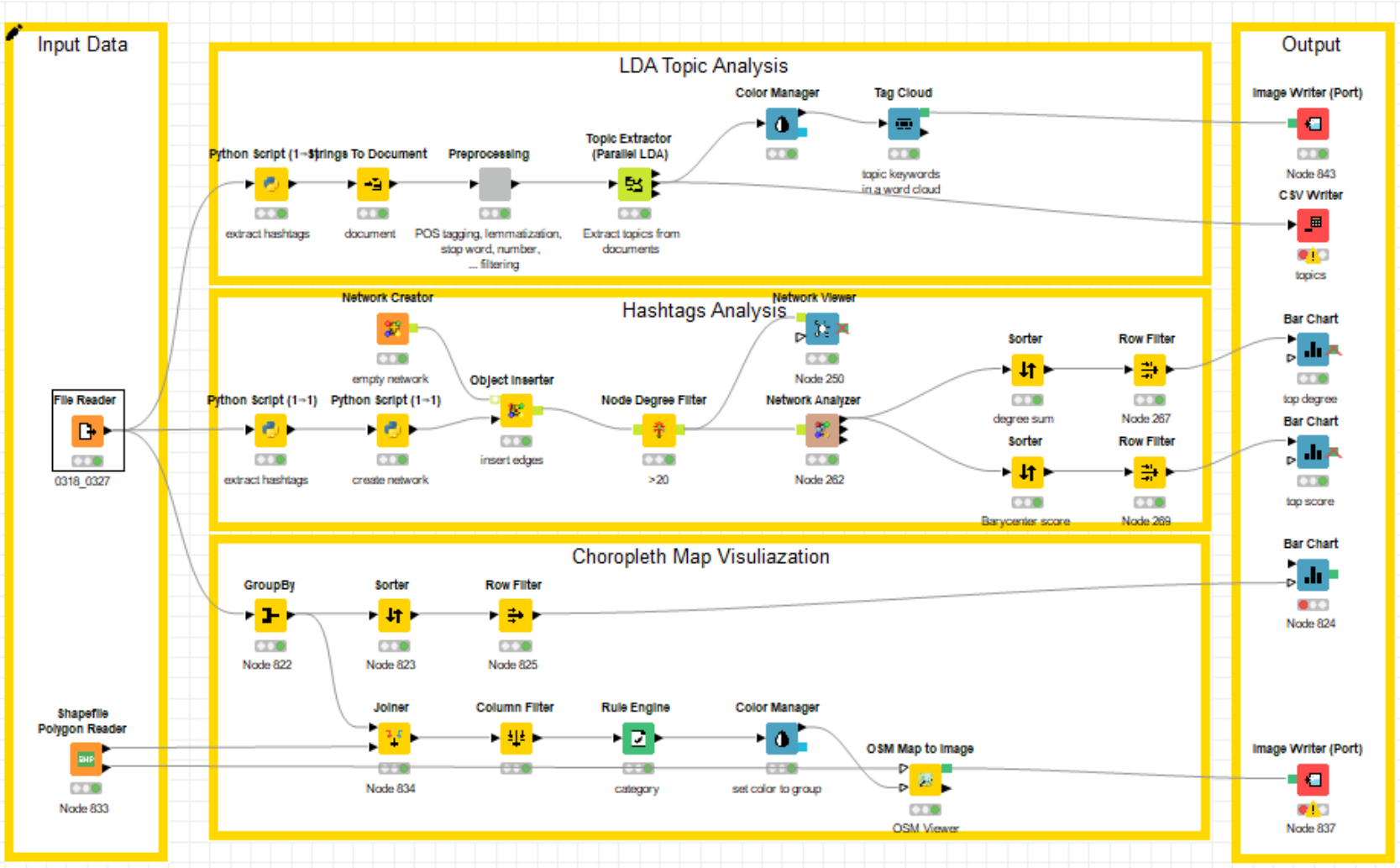
Text mining-based Social Media Data Analysis

❑Flowchart

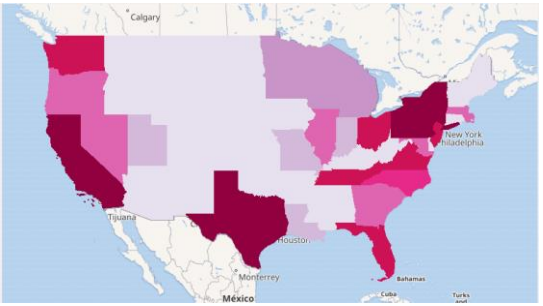


Text mining-based Social Media Data Analysis

❑Workflow Implementation

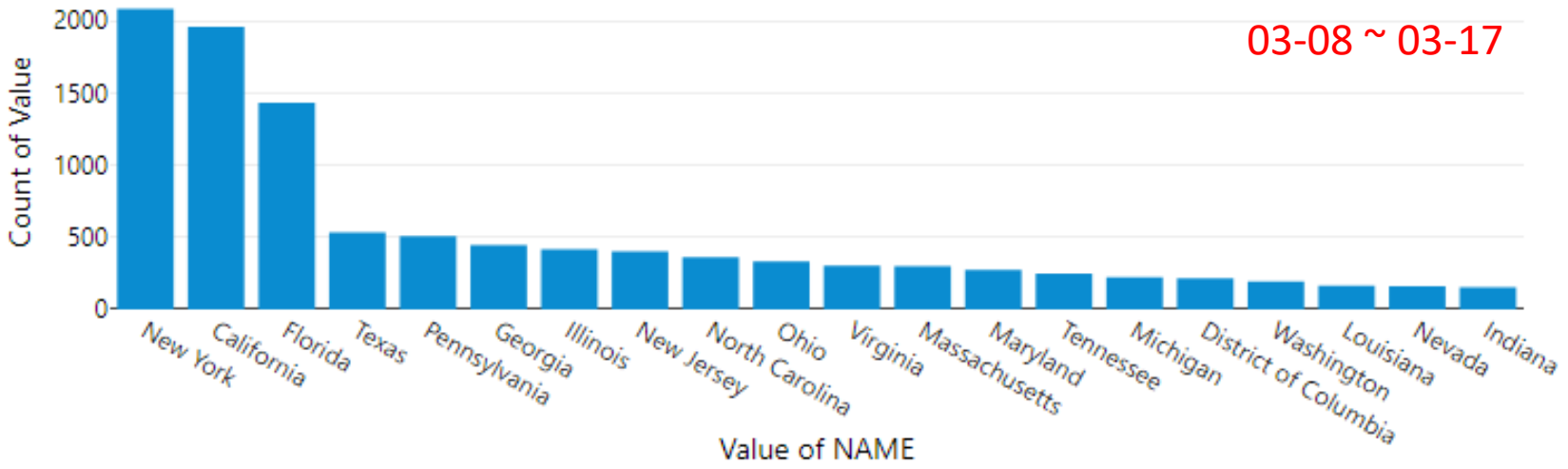


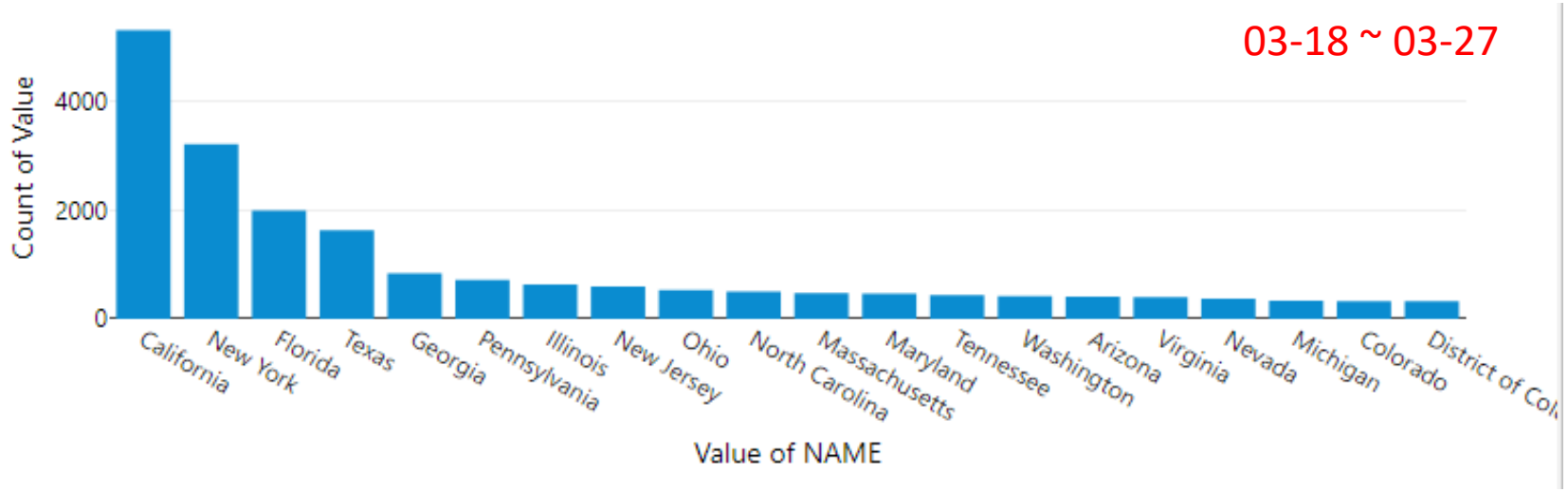
Workflow in KNIME



Text mining-based Social Media Data Analysis

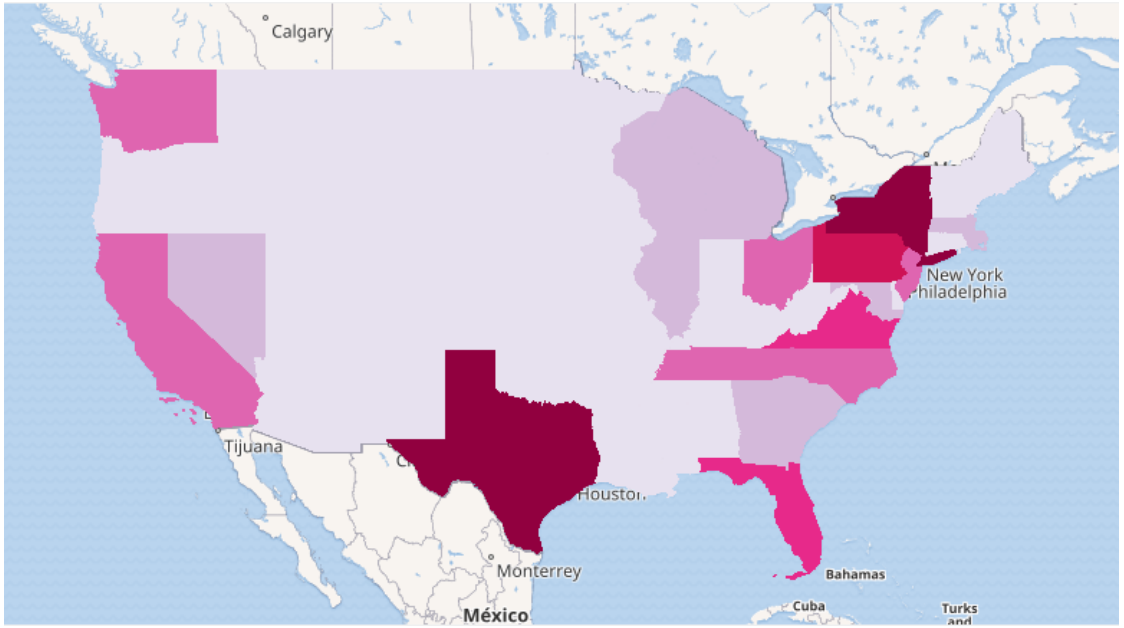
❑Results—Top 20 States with the most posted COVID-19 Tweets



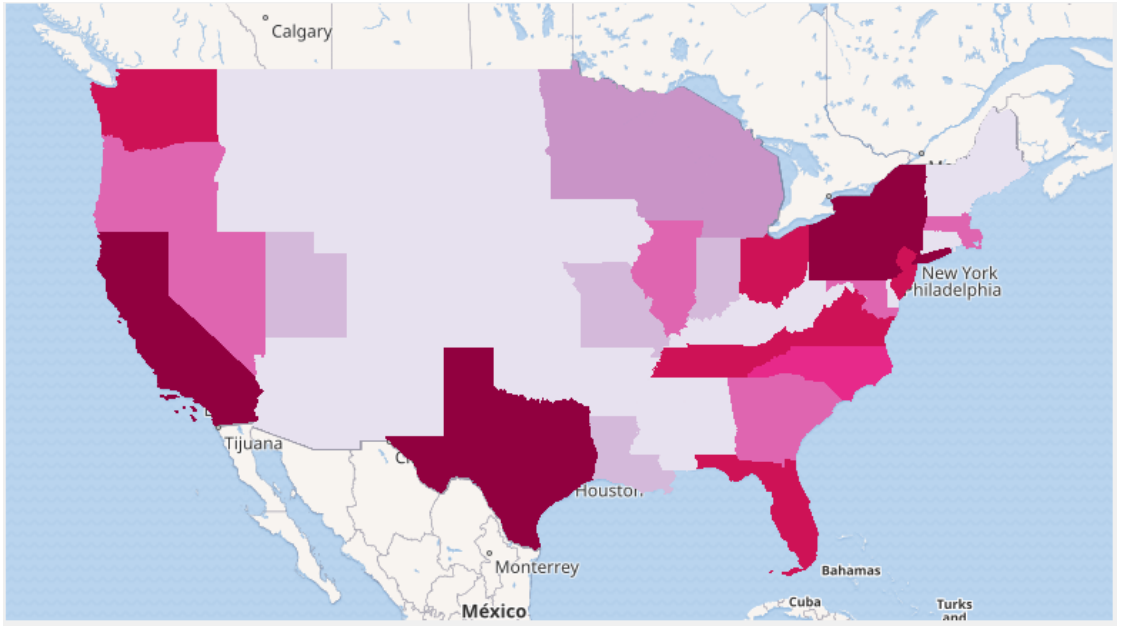


Text mining-based Social Media Data Analysis

❑Choropleth Map Visualization



03/08/2020 ~ 03/17/2020



03/18/2020 ~ 03/27/2020

Text mining-based Social Media Data Analysis

❑Hashtag Analysis

#coronavirus	3290
#covid_19	1209
#Corona	1137
#covid19	768
#Coronavirus	474
#CoronaVirus	400
#pandemic	384
#Socialdistancing	330
#Nyc	326
#covid	309

#quarantine	305
#washyourhands	300
#virus	275
#toiletpaper	247
#COVID19	238
#Corona	185
#coronavir	180
#2020	174
#love	153
#miami	141

03/08/2020 ~ 03/17/2020

#coronavirus	5428.0
#covid_19	4615.0
#covid19	2505.0
#corona	2452.0
#quarantine	1957.0
#socialdistancing	1539.0
#Stayhome	1064.0
#pandemic	839.0
#Coronavirus	829.0
#COVID19	700.0

#CoronaVirus	610
#covid	517
#nyc	489
#virus	438
#2020	432
#staysafe	411
#love	379
#coronavirusmemes	329
#Covid19	327
#shelterinplace	307

03/18/2020 ~ 03/27/2020

Text mining-based Social Media Data Analysis

❑Topic Analysis (LDA Model)

Topic 1	Coronavirus, que, corona, Georgia, con, del,Covid19,atlanta,por,Para
Topic 2	Coronavirus,pandemic,covid19,Corona,people,covid-19,day,mask,virus,Time
Topic 3	Coronavirus,corona,florida,day,covid19 Beach, virus, miami, time, Happy
Topic 4	covid-19,coronavirus,due,update,close Spread,concern,pandemic,health,Covid19
Topic 5	Coronavirus,corona,covid19,York,link,Bio,job, vegas,las,Virus
Topic 6	Los,angeles,county,course,altitude,groundspeed,squawk,n951la,n954la,N955la
Topic 7	…

Topic 1	coronavirus, corona, covid19, quarantine, day, pandemic, time, virus, florida, socialdistancing
Topic 2	los,angeles,county,course,altitude,groundspeed ,squawk,n958la,california
Topic 3	Coronavirus,covid19,florida,corona,Que,del,Miami,para,quarantine,Covid2019
Topic 4	Coronavirus,covid19,Corona,day,covid-19,Park,pandemic,South,Carolina,Washington
Topic 5	Covid19,coronavirus,Quarantine,corona,Socialdistancing,Day,stay,Stayhome,pandemic,Social
Topic 6	Coronavirus,san,Francisco,covid-19,update California,pandemic,test,county,Covid19
Topic 7	…

Text mining-based Social Media Data Analysis

❑ Steps for Running the Workflow

- ❑ Website: http://chinadatalab.org
- ❑ File Folder: Public/Covid/Workflows/05_social_media
  - Workflow File: social_media.knwf
  - Presentation file: Introductions to Social Media Data_EN.pptx
  - Input Data:
    - Public/Covid/Workflows/05_social_media/data/sample.csv
    - Public/Covid/Workflows/05_social_media/data/us_state_2017.shp
  - Output Data:
    - hashtags.csv: network analysis results of hashtags
    - topics.csv: topics analysis results

Outline

- Introduction on Workflow Tools
- Case Study on Spatiotemporal Data Visualization
- Case Studies on COVID-19 Pandemics
  - Human Mobilities and Pandemics Outbreak
  - Text mining-based Social Media Data Analysis
- Case Studies on Regional Science
  - Spatiotemporal hierarchy of regional inequality of China
  - Unconditional convergence of Chinese provinces
- Summary

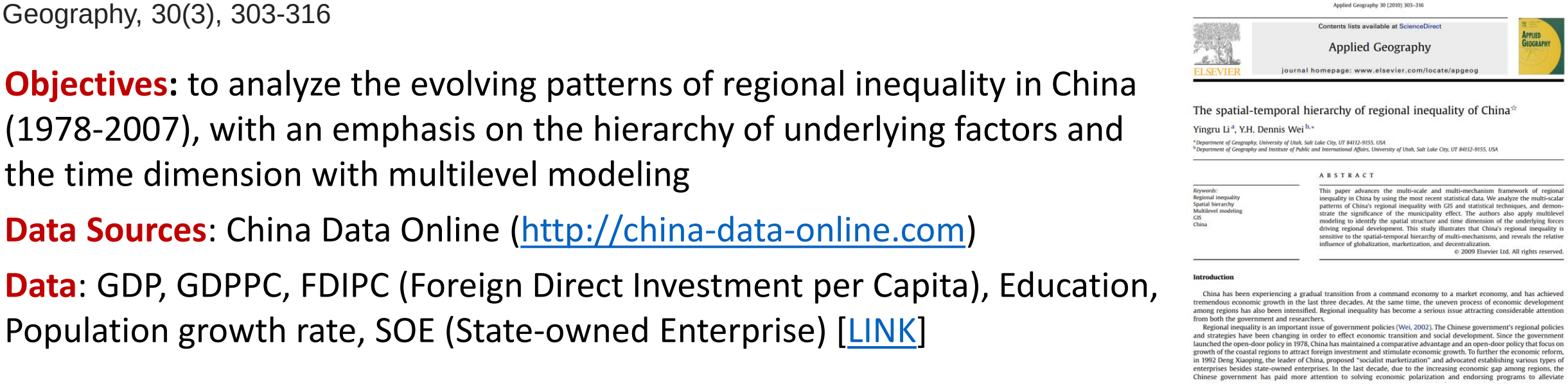
Case Study One: Regional inequality of China

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

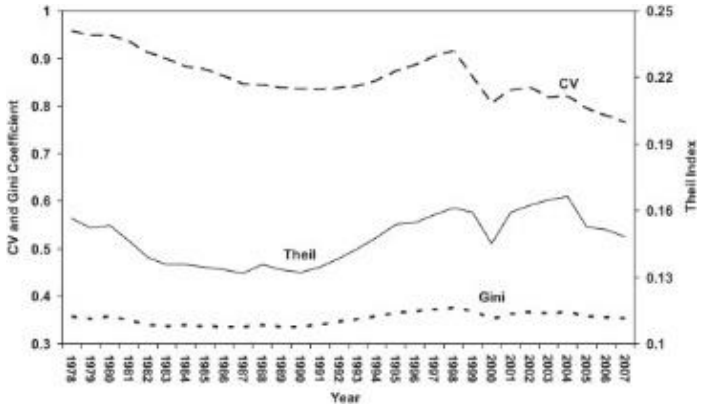
This paper advances the multi-scale and multi-mechanism framework of regional inequality in China by using the most recent statistical data. We analyze the multi-scalar patterns of China's regional inequality with GIS and statistical techniques, and demonstrate the significance of the municipality effect. 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.

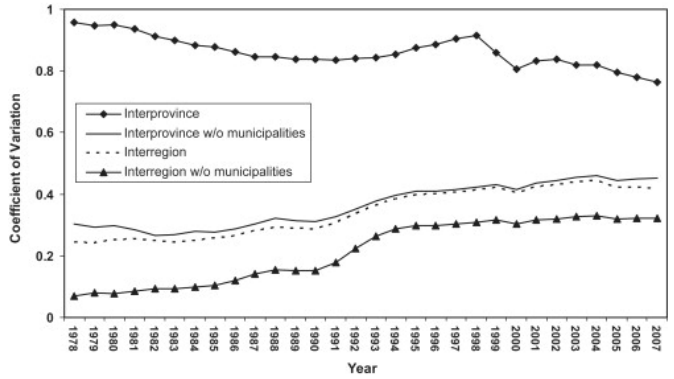
© 2009 Elsevier Ltd. All rights reserved.

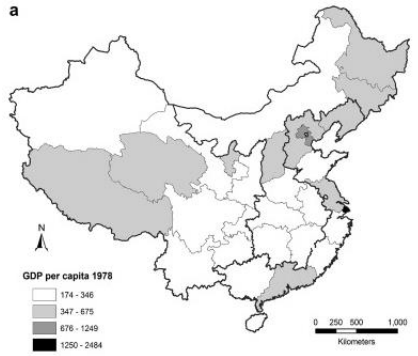
Introduction

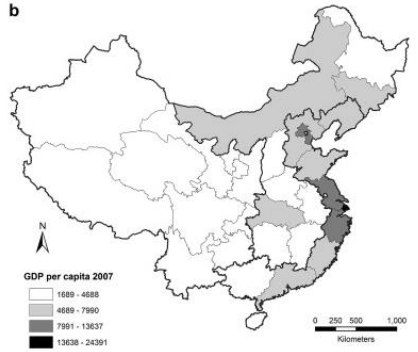
China has been experiencing a gradual transition from a command economy to a market economy, and has achieved tremendous economic growth in the last three decades. At the same time, the uneven process of economic development among regions has also been intensified. Regional inequality has become a serious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









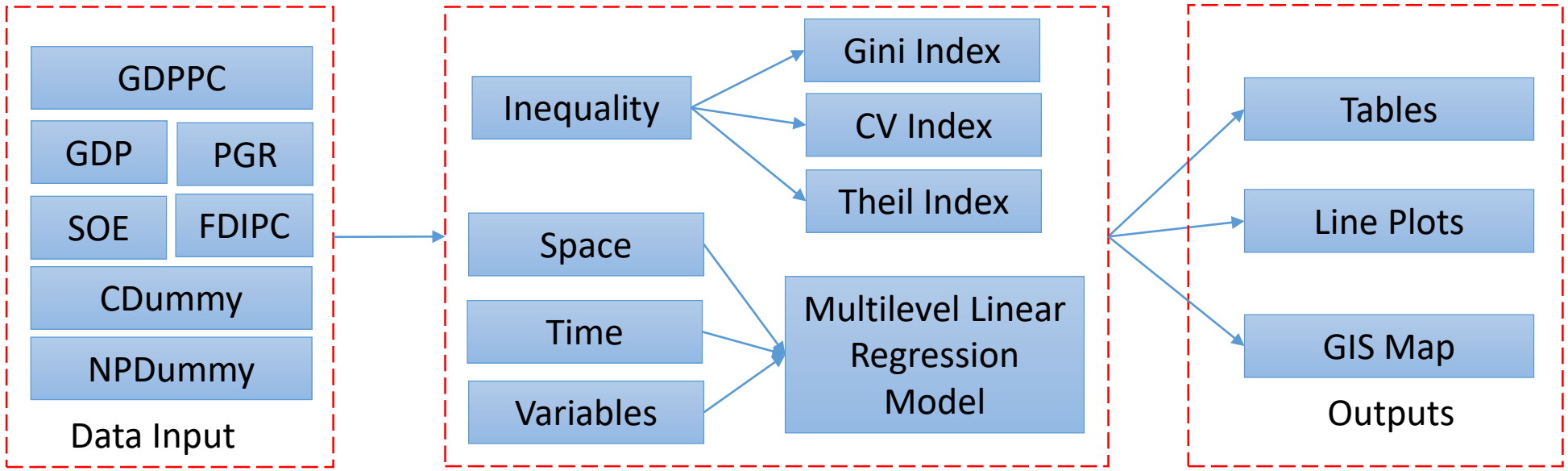
Case Study One: Regional inequality of China

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

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

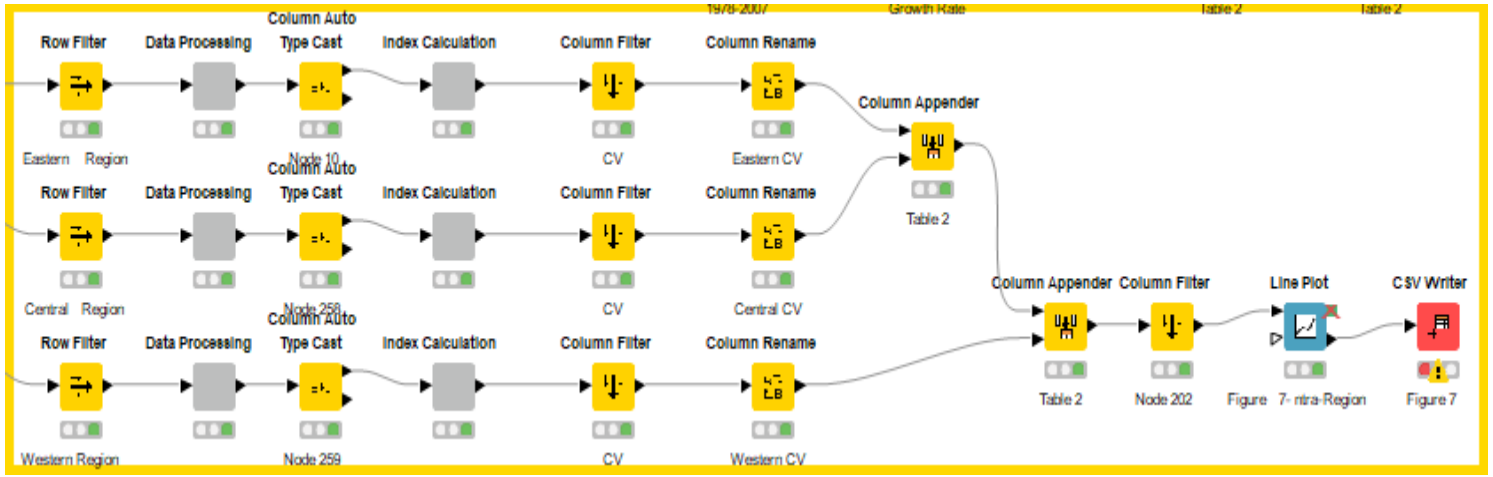
single-level (province), two-level (region and province) and three-level (time, region, and province)

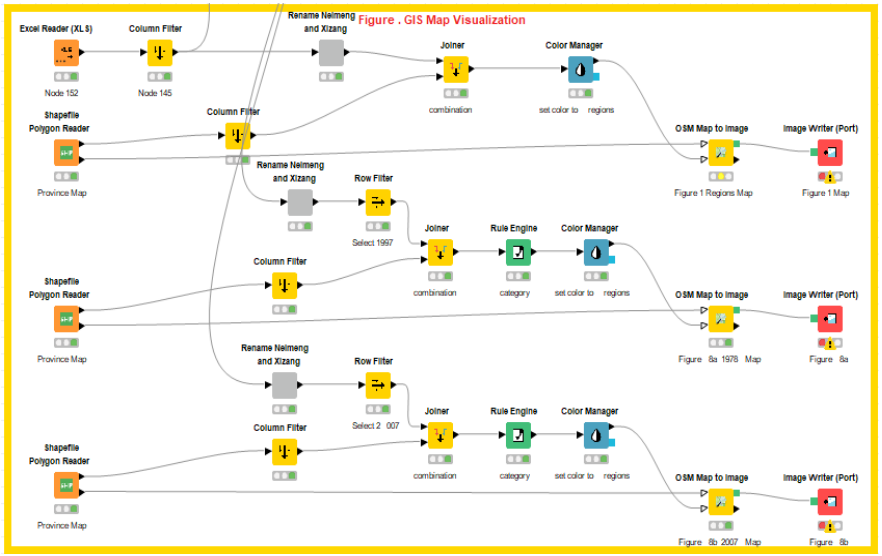
- **Flowchart**

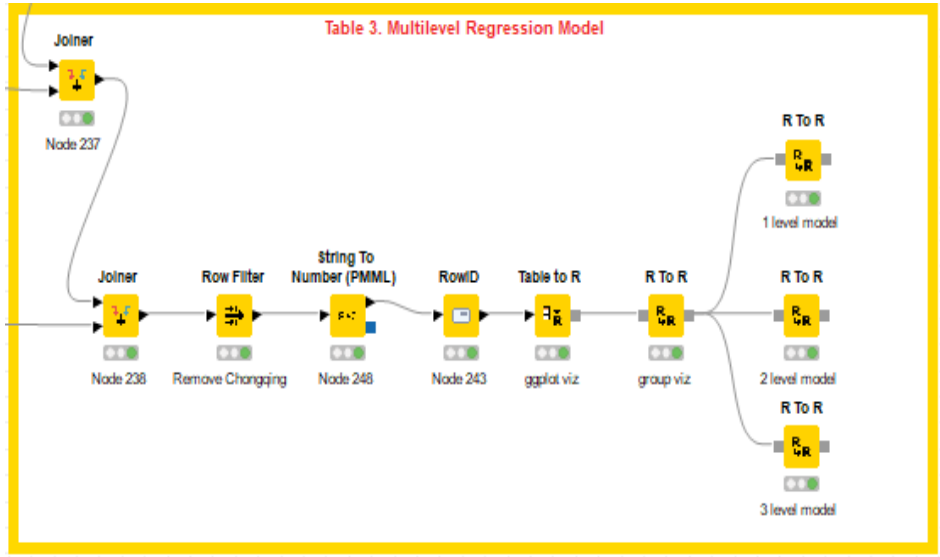


Case Study One: Regional inequality of China

The Workflow Implementation by KNIME

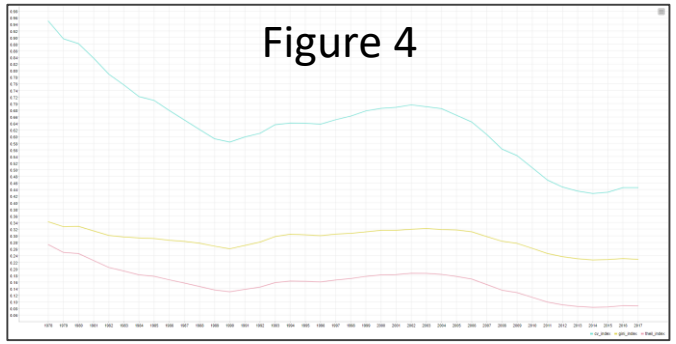


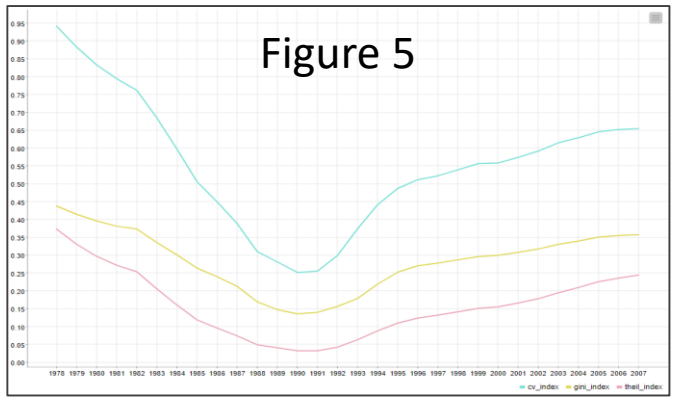


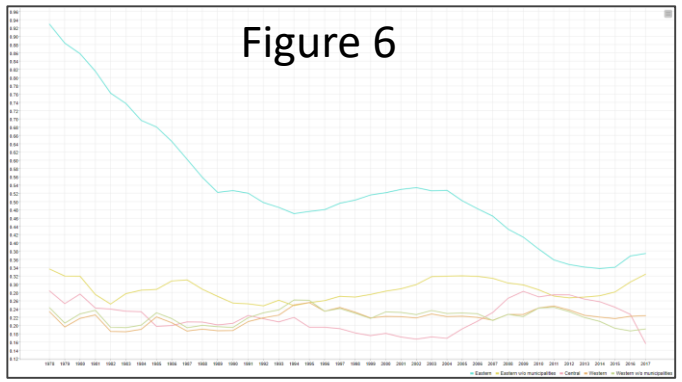


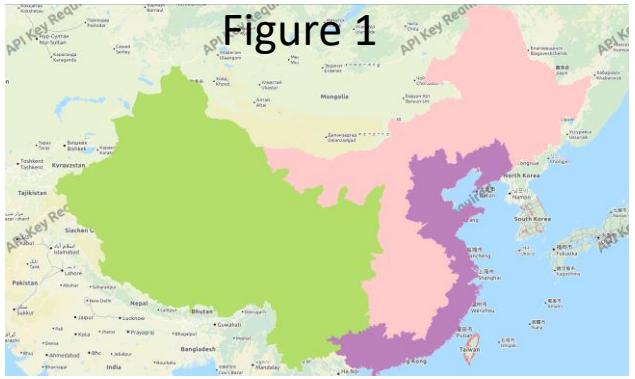
Case Study One: Regional inequality of China

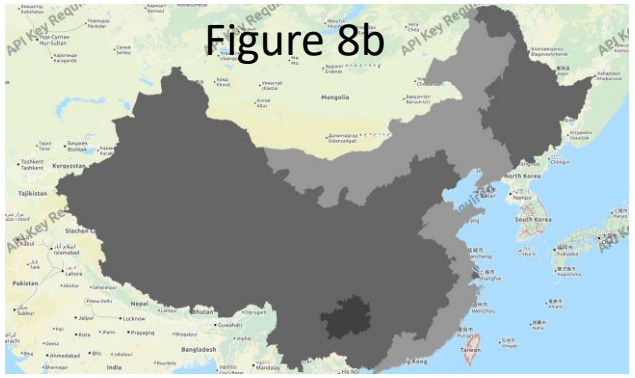
- **Outputs**












```
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 One: Regional inequality of China

❑ Steps for Running the Workflow

*** Please install R extension
> install.package('lme4')
> install.package('lmetest')
> install.package('matrix')

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 Two: Unconditional Convergence of China

Wu, M. L. (2020). Unconditional convergence of Chinese provinces (1952-2017): Some statistical analysis results. Journal of Economics and International Finance, 12(2), 57-64.

Objectives: to test the economic convergence hypothesis of 31 mainland Chinese provinces over the period from 1952 to 2017

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

Data: GDP, GDPPC [LINK]

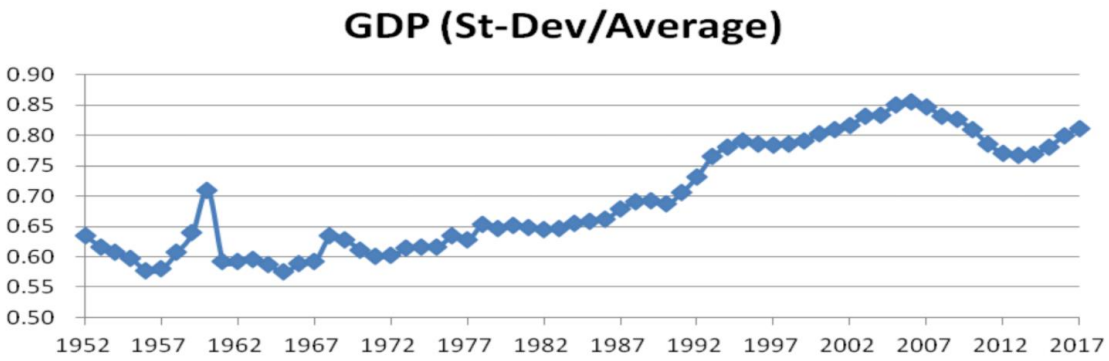


Figure 2. Relative variation of 31 Chinese provinces' GDP.

Vol. 12(2), pp. 57-64, April-June 2020
DOI: 10.5897/JEIF2020.1034
Article Number: 703812363445
ISSN 2006-9812
Copyright©2020
Author(s) retain the copyright of this article
http://www.academicjournals.org/JEIF

ACADEMIC JOURNALS

Journal of Economics and International Finance

Full Length Research

Unconditional convergence of Chinese provinces (1952-2017): Some statistical analysis results

Ming-Lu Wu

Division of Business and Management, United International College (UIC), Beijing Normal University-Hong Kong Baptist University, 2000 Jintong Road, Tangjiawan, Zhuhai, Guangdong 519085, P. R. China.

Received 25 February, 2020; Accepted 27 March, 2020

This research tests the economic convergence hypothesis of 31 inland Chinese provinces over the period from 1952 to 2017. Regression and descriptive analysis methods are adopted to study the economic convergence among these Chinese provinces in terms of GDP growth and per-capita GDP growth. The research results show that GDP growth does not exhibit a tendency of convergence, rejecting the absolute convergence hypothesis among the Chinese provinces. But per-capita GDP growth does suggest convergence, especially after China's economic reform from 1978 to 2017, supporting the relative convergence hypothesis among the Chinese provinces. Practical and policy implications are provided based on the research results.

Key words: Economic convergence, regression, descriptive analysis, GDP, per-capita GDP, China.

INTRODUCTION AND RESEARCH BACKGROUND

The idea of economic convergence

As directly implied by the assumption of diminishing same exogenous parameters of economic growth across regions, their initial levels of per capita income (or equivalently, per capita capital stock) are not

Table 1a. Test of absolute convergence from 1952 to 2017.

ln(GDP2017/GDP1952) = *a* + *b**ln(GDP1952)				
Parameter	**Estimate**	**Standard error**	***t*-value**	***p*-value**
Intercept (*a*)	7.4489	0.2534	29.3958	0.0000
Slope (*b*)	-0.0969	0.0882	-1.0986	**0.2813**
R-Square	0.0414			

Case Study Three: Unconditional Convergence of China

- **Methodology**: regression model of growth rate was used but for all 31 mainland provinces in China

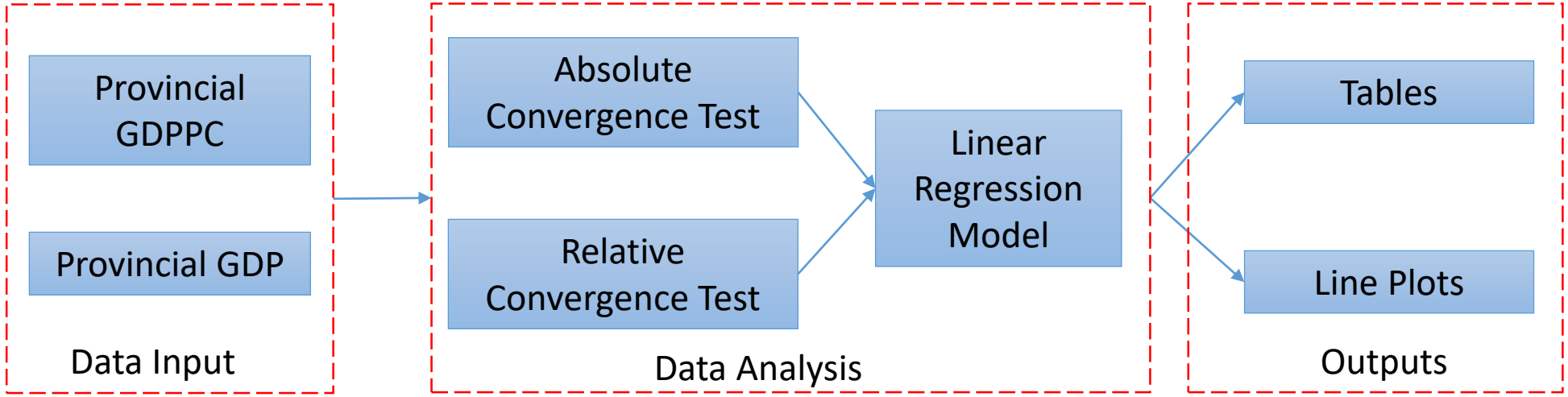
$$\ln(GDP2017/GDP1978) = a + b*\ln(GDP1978)$$

$$\ln(GDP2017/GDP1952) = a + b*\ln(GDP1952)$$

$$\ln[(\text{GDP per capita 2017})/(\text{GDP per capita 1952})] = a + b*\ln(\text{GDP per capita 1952})$$

$$\ln[(\text{GDP per capita 2017})/(\text{GDP per capita 1978})] = a + b*\ln(\text{GDP per capita 1978})$$

- **Flowchart**



Case Study Two: Unconditional Convergence of China

Linear Regression Tables (1a, 1b, 2a, 2b)

Table 1A - Test of Absolute Convergence from 1952 to 2017

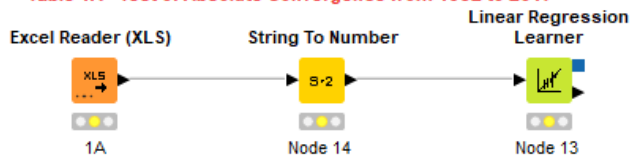


Table 1B - Test of Absolute Convergence from 1978 to 2017

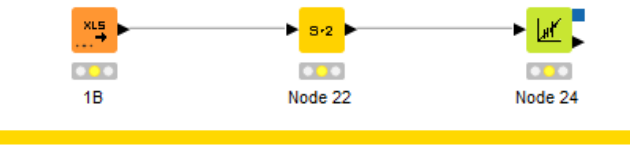


Table 2A - Test of Relative Convergence from 1952 to 2017

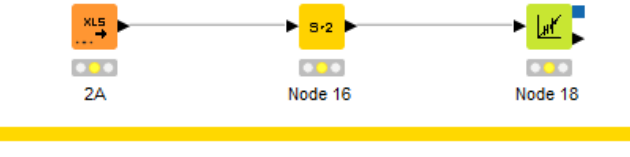
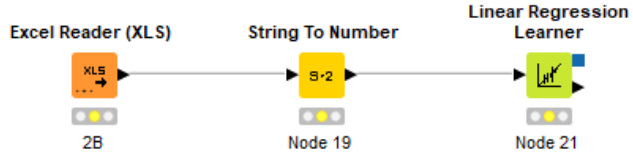


Table 2B - Test of Relative Convergence from 1978 to 2017



Figures 2, 3 (St-Dev/Average) 4, 5 (Relative to Bejing, St-Dev/Average)

Figure 4 - GDP, Relative to Beijing (St-Dev)

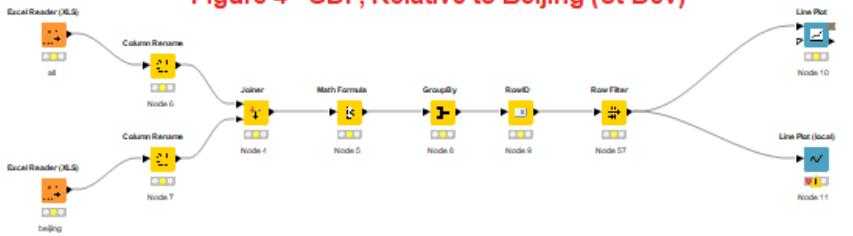


Figure 5 - GDP Per Capita, Relative to Beijing (St-Dev)

Figure 2 - GDP (St-Dev)

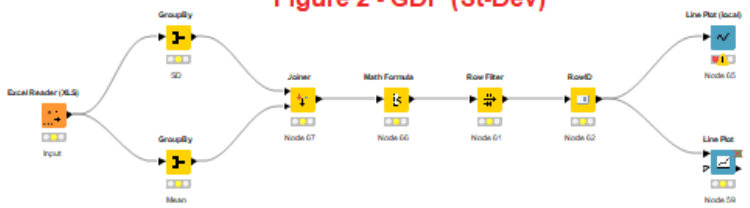
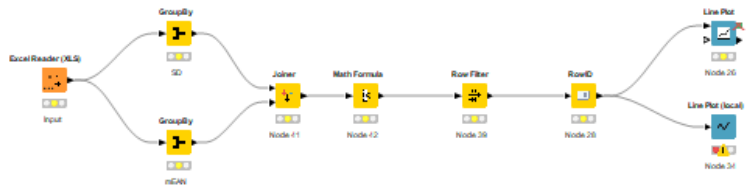
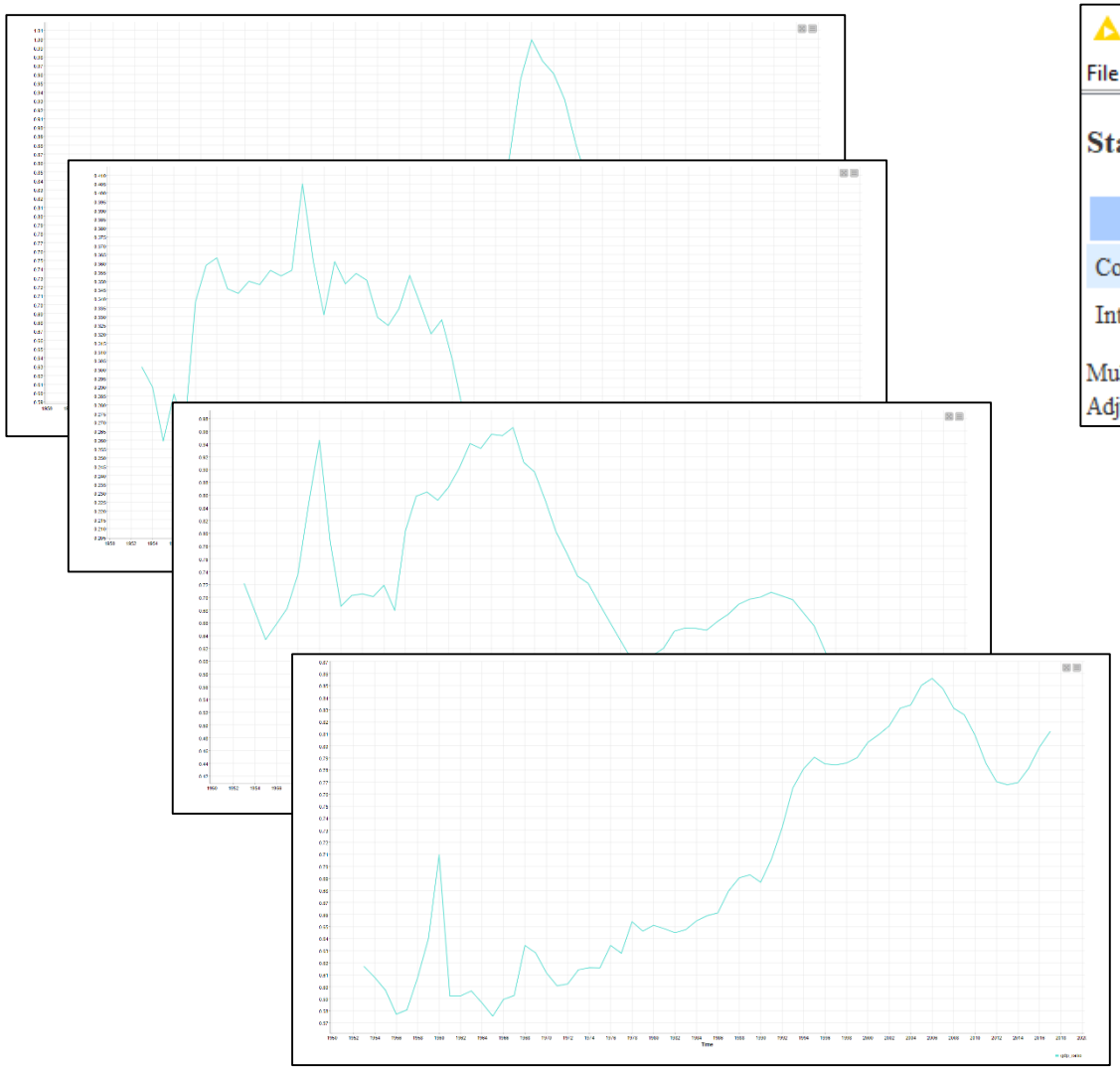


Figure 3 - GDP Per Capita (St-Dev)



Case Study Two: Unconditional convergence of China

Relative variation and over Beijing of 31 Chinese provinces' GDP and GDP per capita



Test of absolute and relative convergence from 1952/1978 to 2017 using GDP and GDP per capita

Linear Regression Result View - 3:13 - Linear Regression Learner

File

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
Col4				
Intercept				

Multiple R-Sc
Adjusted R-S

Linear Regression Result View - 3:24 - Linear Regression Learner

File

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
Col4				
Intercept				

Multiple R-Squ
Adjusted R-Squ

Linear Regression Result View - 3:18 - Linear Regression Learner

File

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
Col4				
Intercept				

Multiple R-Sq
Adjusted R-Sc

Linear Regression Result View - 3:21 - Linear Regression Learner

File

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
Col4	-0.5311	0.103	-5.1579	1.80E-5
Intercept	8.1605	0.6133	13.3059	1.25E-13

Multiple R-Squared: 0.4872
Adjusted R-Squared: 0.4689

Case Study Two: Unconditional Convergence of China

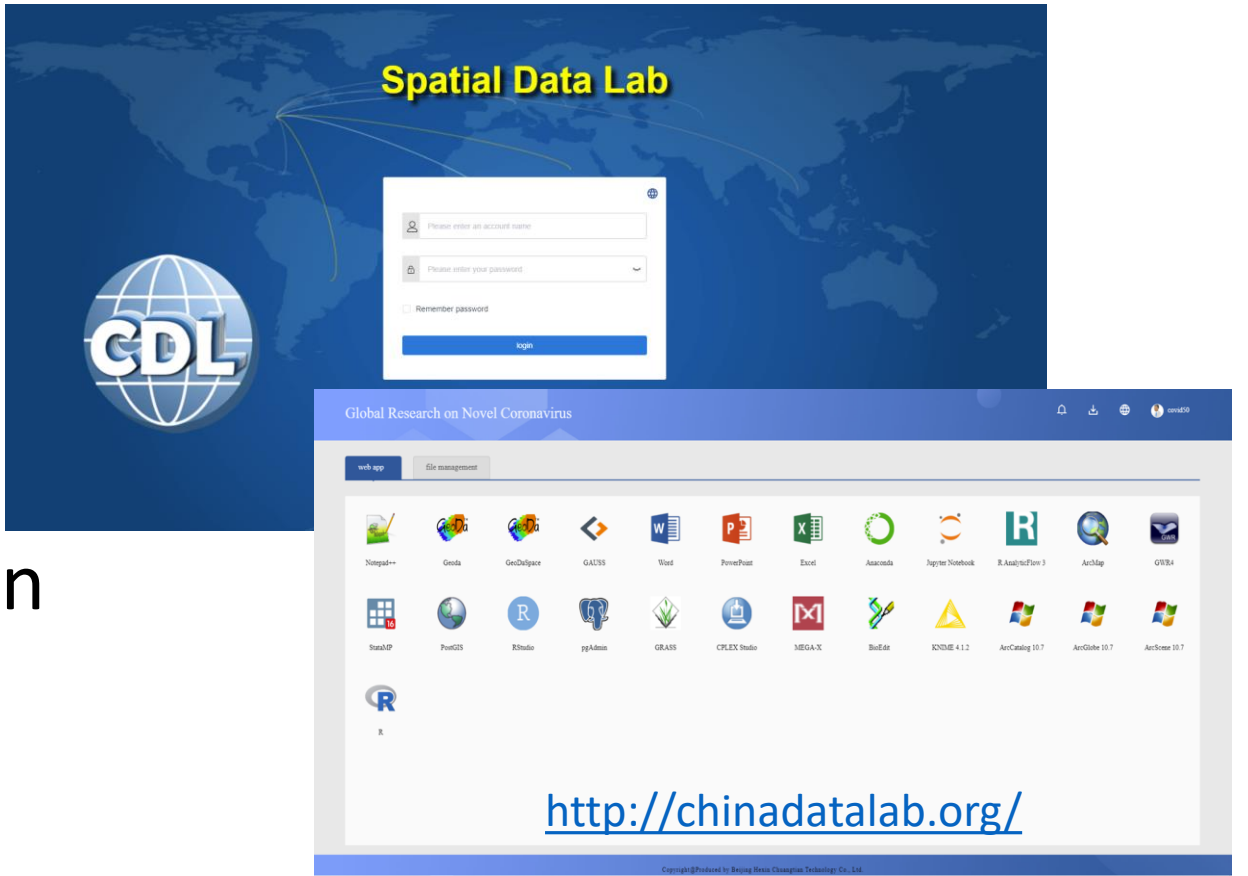
❑ Steps for Running the Workflow

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 3: **Import KNIME workflow** file (Convergence of Chinese Provinces Workflow.knwf) from the File menu
Step 4: **Launch the workflow** by double clicking the imported workflow
Step 5: **Configure "Input Datasets"** by Changing Location
Step 6: Click **Run** function from the top menu
Step 7: **Display the outputs**:

- **2 Interactive View: Linear Regression** for Test of absolute convergence from 1952 and 1978 to 2017 using GDP
- **2 Interactive View: Linear Regression** for Test of absolute convergence from 1952 and 1978 to 2017 using GDP per capita
- **2 Interactive View: Line plot** for Relative variation of 31 Chinese provinces' GDP and GDP per capita
- **2 Interactive View: Line plot** for Variation of 31 Chinese provinces' GDP and GDP per capita over Beijing's

Summary

- IBM predicted that the demand for data scientists will increase by 28 percent by 2020
- Everyone can be a data scientist
- Data Science Ecosystem
  - Sharing Data
  - Sharing Tools
  - Sharing Process
  - Sharing Knowledge
  - Intelligent Recommendation



Web Sites

China Data Online

http://china-data-online.com

China Data Center

http://chinadatacenter.net

China Data Lab

http://chinadatalab.net

office@chinadatacenter.net