新冠病毒数据分析系列公益讲座 COVID-19 Data Analysis Webinars

第三讲、疫情数据工作流建模分析

主讲人：胡涛，哈佛大学

主持人：杨超伟，乔治梅森大学

3. COVID-19 Data Analysis with Workflows

Speaker: Tao Hu, Harvard University

Chair: Chaowei Yang, George Mason University

北京时间 2020年5月15日上午9-10点

注意事项

- 欢迎各位来到新冠病毒疫情数据分析讲座！
- **请将各自的话筒静音，以免影响录音质量**
- 欢迎随时Chat提交问题，由主讲人最后统一回答
- 如有问题,可以联系office@chinadatacenter.net
- 下次讲座预告：

讲座题目：疫情数据空间显示分析

时间：5月22日上午9-10点,

主讲人：关蔚禾, 哈佛大学；付品德, 美国环境系统研究所（ESRI）

点评人：隋殿志, 阿肯色大学

主持人：林晖，江西师范大学地理与环境学院

合办单位：江西师范大学地理与环境学院

议题 Topics

1. 项目背景 Background
2. 数据来源 Data Sources
3. 工作流软件介绍 Introduction to Workflow Tool
4. 疫情数据分析工作流 Data Analysis with Workflows
5. 工作流共享平台 Cloud for Workflow Sharing
6. 致谢 Acknowledgement

1 背景: 研究项目Resources for COVID-19

项目背景： “新冠病毒疫情资源与全球研究”

https://projects.iq.harvard.edu/chinadatalab/resources-covid-19

项目目标: to provide an information infrastructure for the global study of novel coronavirus (COVID-19)

项目任务：Objectives:

- 疫情研究数据长久**保存**：To establish a **permanent collection** for the COVID-19 study at local, regional and global levels with information collected and integrated from different sources
- 提供方法与技术**支撑**：To facilitate the quantitative research on spatial spread and impacts of COVID-19 study with advanced **methodology and technology**
- 推动研究**合作**：To promote collaborative research in the COVID-19 study with the **cloud platform**
- 促进教学研究**应用**：To promote the use and sharing of data sources in **teaching and research** with DataVerse/WorldMap/CDL
- **团队**能力建设：To build the **capacity** for future collaborative projects

1 背景：研究团队/Research Teams

哈佛大学地理分析中心，美国

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.

中国数据研究所，美国

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.

武汉大学社会地理计算中心，中国

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.

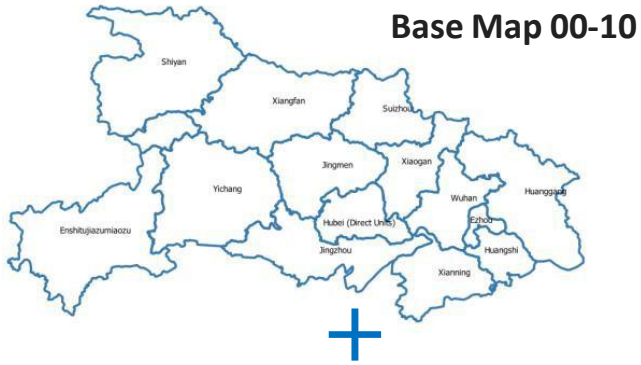
RMDS实验室，美国

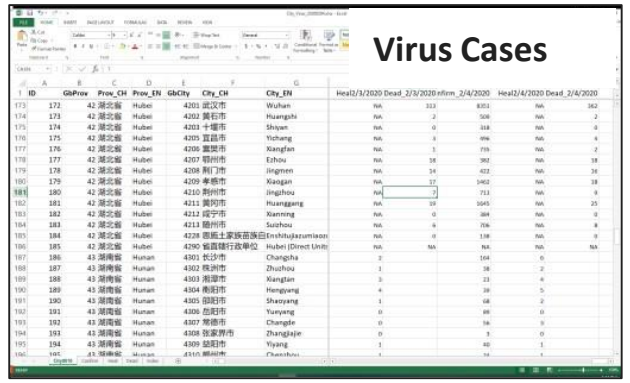
RMDS

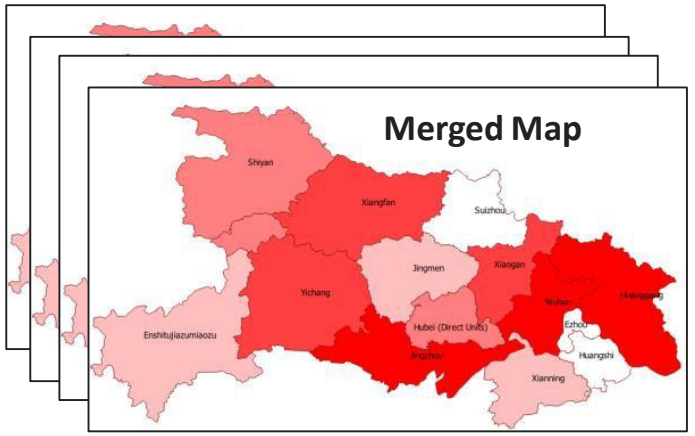
The RMDS Lab is designed to serve a global community of companies and data scientists, empowering them to reap the benefits of big data, artificial intelligence, and machine learning.

2 数据来源/Data Sources

- **基础数据 / Base data**
  - Base maps (province, prefecture and county)
  - Census data (population census and economic census)
  - Statistics (province, prefecture and county)
  - Others
- **病例数据 / Virus data**
  - Virus reports from gov't (country, province and city)
- **迁徙数据 / Migration data**
  - Daily floating population data (Baidu)
  - Migration data from census (NBS) 国家统计局
- **医疗机构数据 / Health facility data**
  - POI data (高德地图)
  - Economic Census (NBS)
- **社交媒体数据 / Social media data**
  - GDELT (Global Database of Events, Language, and Tone)
  - Twitter
- **论文 / Papers**
- **报告 / Reports**
- **政策法规 / Policies and regulations**







2 数据来源：基础数据

中国数据

- **Government Statistics 政府统计**
  - Provincial Statistics (1949 -)
  - City Statistics (1996 -)
  - County Statistics (1997 -)
- **Population Census 人口普查**
  - Census 1953
  - Census 1964
  - Census 1982
  - Census 1990
  - Census 2000/2010 (province, city, county, township, GRID)
- **Economic Census 经济普查**
  - 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 organizations 基本单位
- **Geography and Environment 地理与环境**
  - Land Use data 土地利用
  - Night-Time data 夜光强度

美国数据

- **Population Census** (State, County, Tract, Block, Metropolitan, CCD, Place)
  - Census 1970
  - Census 1980
  - Census 1990
  - Census 2000
  - Census 2010
  - American Community Survey

Economic Census (State, Metropolitan, County)

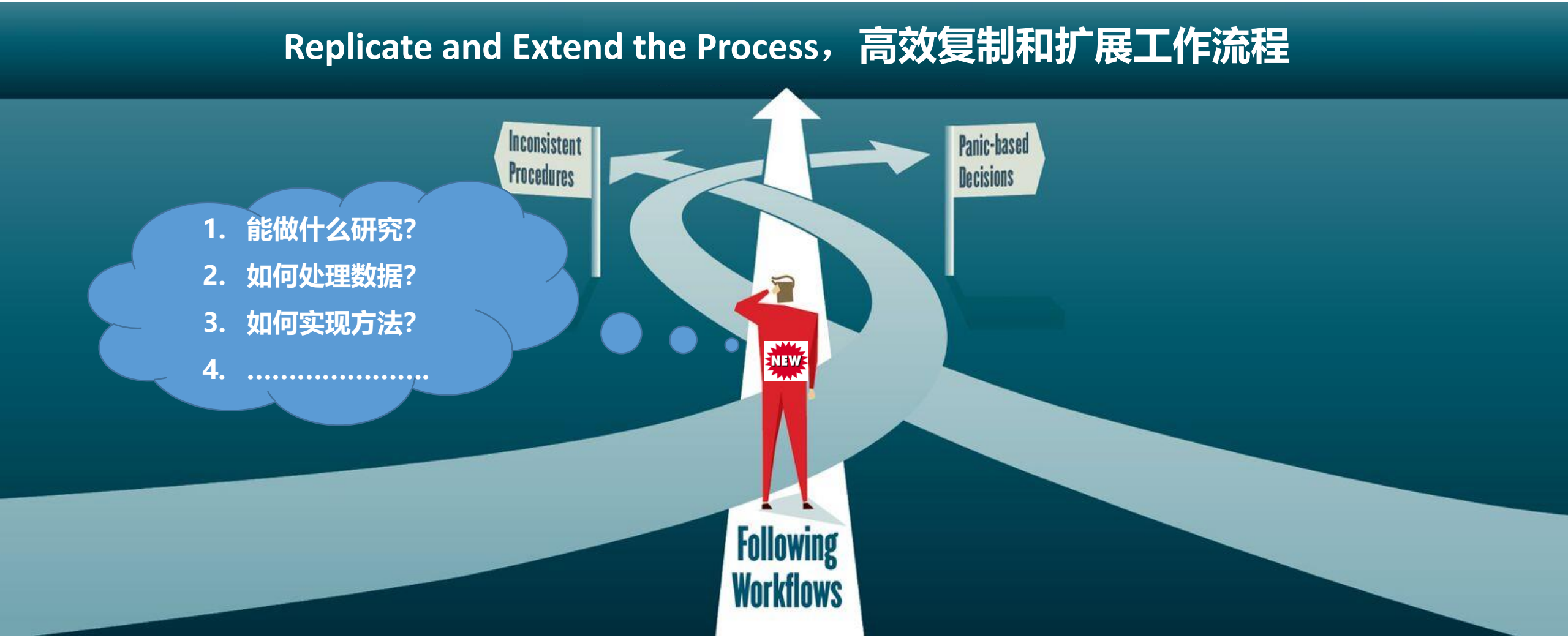
- **Business Data** (State, Metropolitan, County, Tract, CCD, Place)

- **Establishments** (individual business and organizations)
  - Firm level records (more than 13 million records with 85 fields for each business)

2 主要数据集 Data Collection

序号	数据项	Data Sets
2.1	疫情病例数据	Coronavirus cases data
2.2	人口迁徙数据	Population mobility data
2.3	医疗机构数据	Health facilities data
2.4	行为轨迹数据	Trace data
2.5	航线航班数据	Flight data
2.6	社交媒体数据	Social media data
2.7	新闻政策数据	News and Policy data
2.8	高铁班次数据	High-speed train data
2.9	气象气候数据	Meteorological data

3.1 为什么使用工作流? 可复制、可重复、可扩展



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

3.2 工作流软件: 工作流管理系统

Workflow (Definition): the sequence of industrial, administrative, or other processes through which a piece of work passes from initiation to completion. https://www.lexico.com/en/definition/workflow

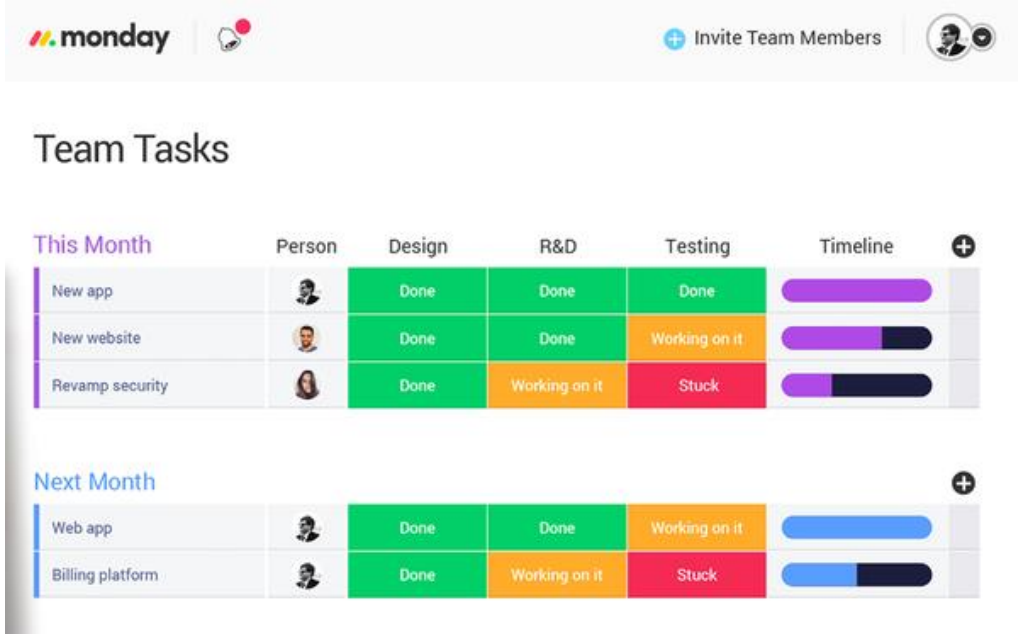
❑ COVID-19 Test Workflow

COVID-19测试工作流

https://www.mdpi.com/2075-4418/10/3/165/htm

❑ Administrative Workflow

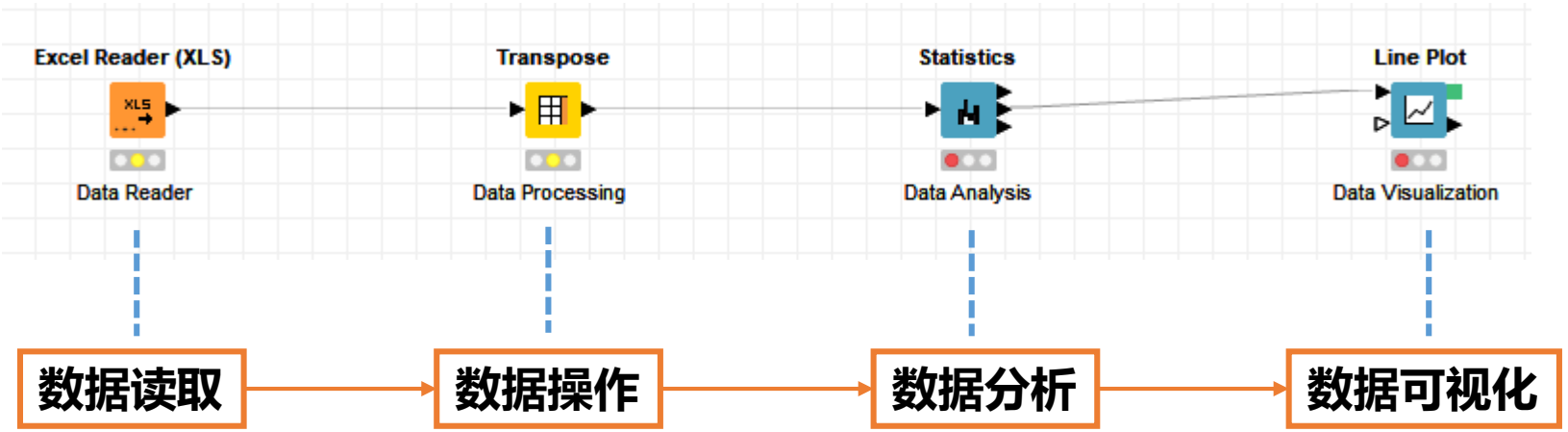
办公协同工作流



https://monday.com/

3.2 工作流软件：工作流数据分析

❑ Data Science Workflow 数据科学工作流



- **低门槛**：利用极少或者无需编程工作即可形成一套完整的数据分析流程；
- **易操作**：通过拖拽的形式选择工作节点，通过连接线实现数据互通；
- **组件模块化**：每个分析操作节点模块化，可完整分析独立运行；
- **可复制性**：更换输入数据源即可轻松复制数据处理，分析和可视化流程；
- **可扩展性**：修改、删除或增加分析操作节点可轻松实现工作流的扩展性

3.2 工作流软件比较

工作流软件对比

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

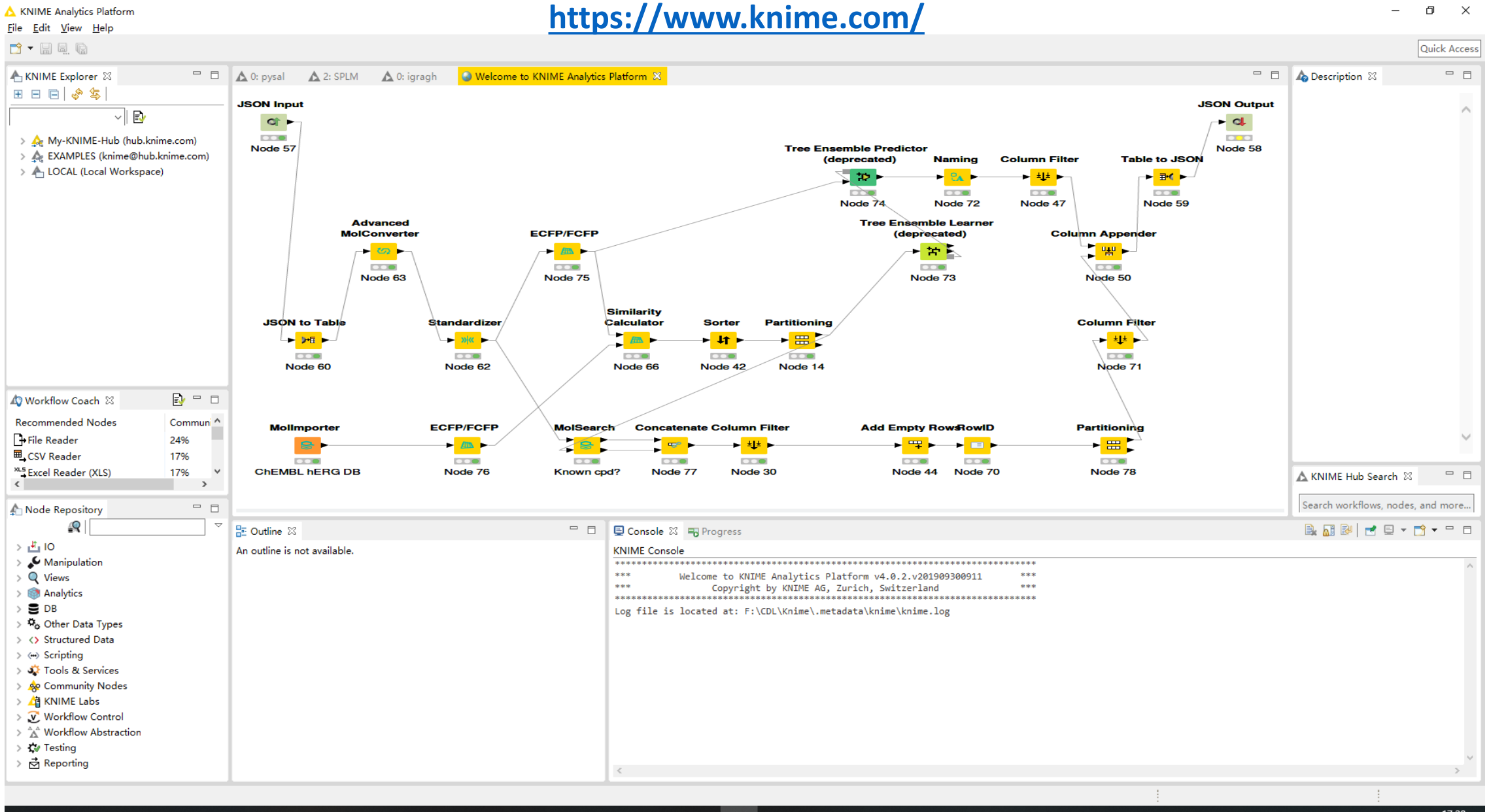
3.2 工作流软件比较

工作流软件对比

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	✔✔✔	✔	✔✔✔	✔✔✔	✔✔	✔✔✔	✔✔✔	✔✔

3.3 数据分析工作流软件 KNIME （免费）

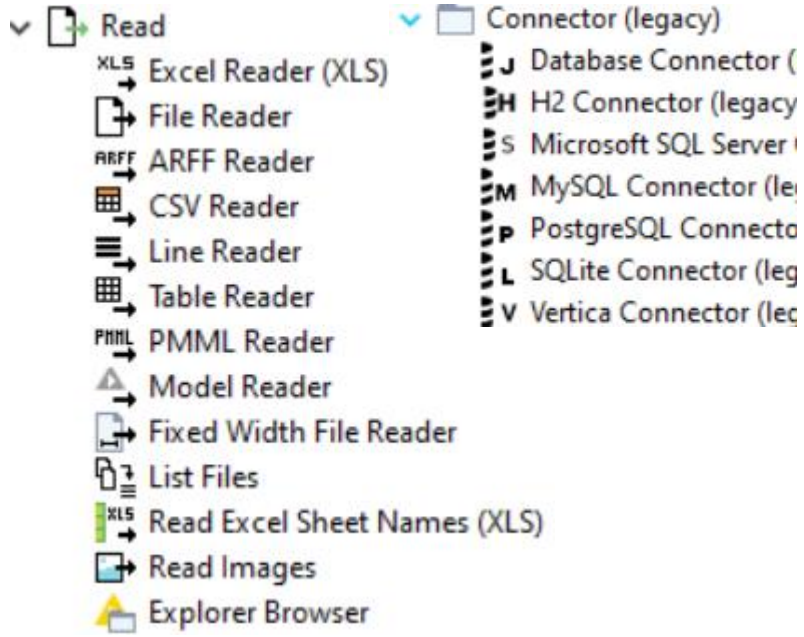
https://www.knime.com/



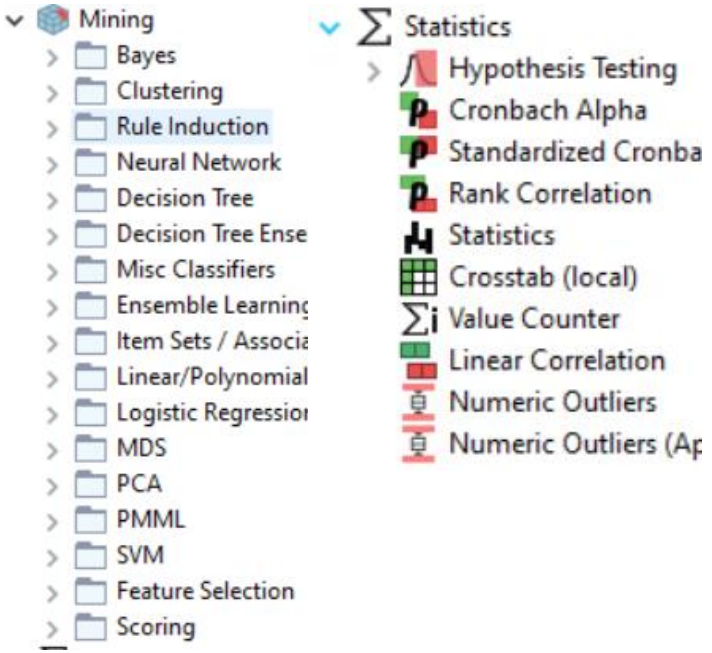
3.3 数据分析工作流软件 KNIME

❑ 数据分析节点 Node Repository

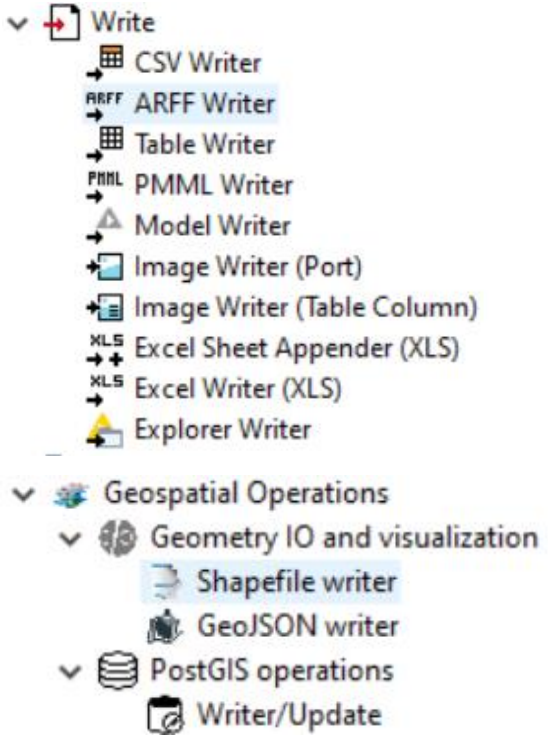
Input输入



Analysis分析



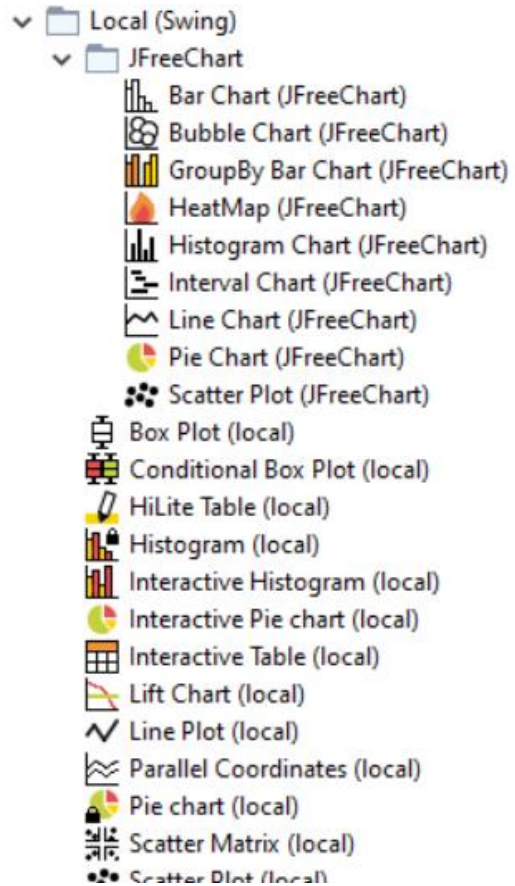
Output输出



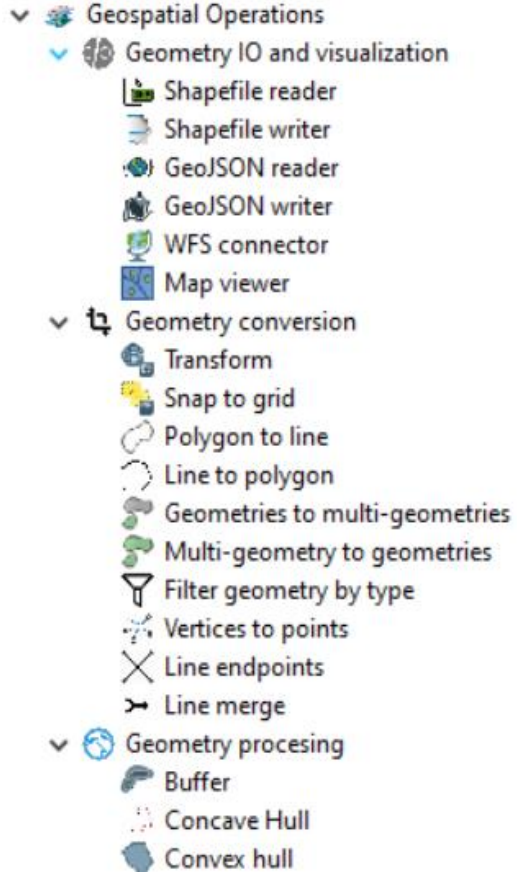
3.3 数据分析工作流软件 KNIME

❑ 其它节点 Other Nodes

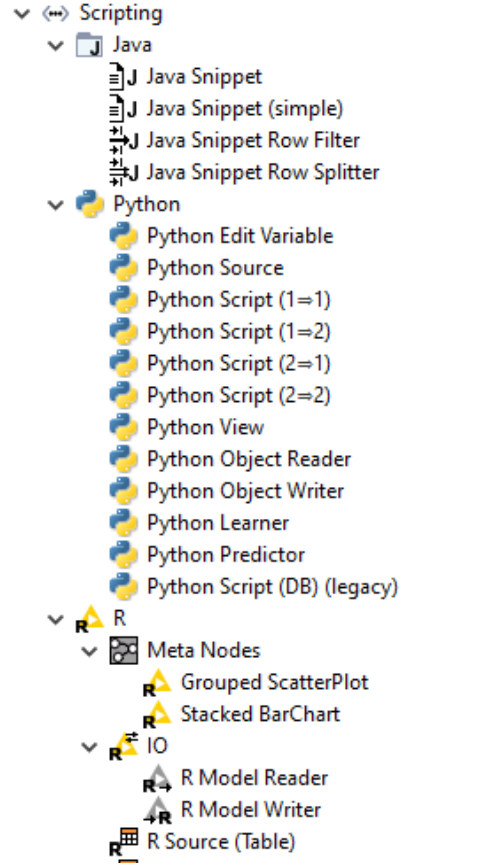
Visualization可视化



Spatial Operations



Scripting脚本



3.4 数据分析工作流：目标

- **Improve the efficiency of data analysis** 提高数据分析效率
- **Cultivate professional data analysis** 培养专业的数据分析流程
- **Promote knowledge and information sharing** 促进知识和信息共享
- **Promote data applications in research and education** 促进数据在学术和教育领域的广泛应用

3.4 数据分析工作流：报告

- ❑Title
- ❑Key words
- ❑Abstract
- ❑Background (hypothesis, debates from literature)
- ❑Data
- ❑Methodology
- ❑Results of the Analysis
- ❑Conclusions and Discussions
- ❑Acknowledgements
- ❑References
- ❑Diagram of the workflow for data analysis

4 基于工作流的疫情数据分析

- 疫情数据空间可视化分析
- 人口迁徙与疫情扩散分析
- 基于工作流的气象数据流程化处理
- 基于社交媒体数据的疫情分析
- 疫情研究文献分析

4.1 疫情数据空间可视化分析：数据来源

❑研究目标

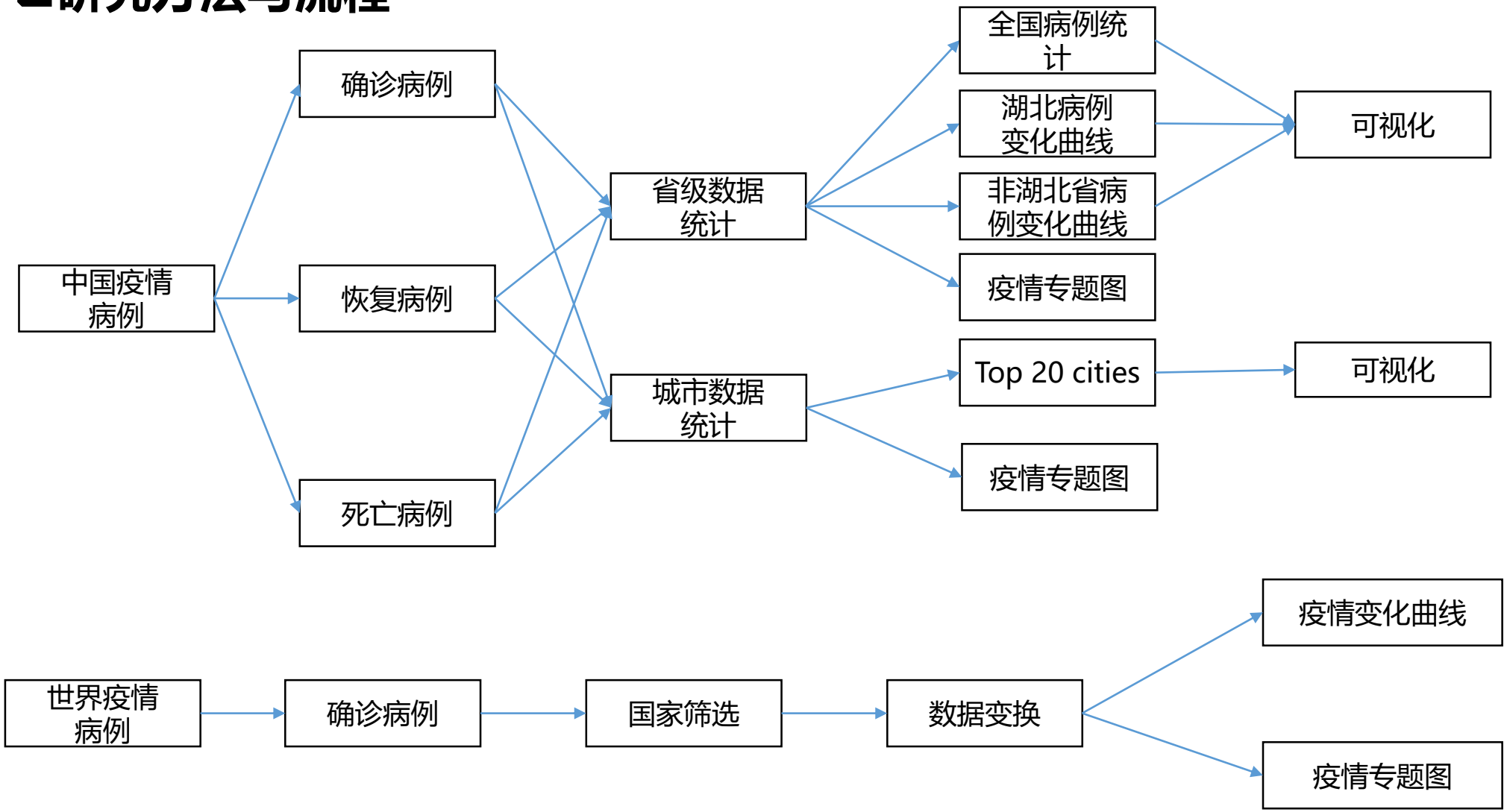
分别以中国和全球疫情数据为例，将疫情数据与基础地图数据关联整合，通过曲线、直方图、地图等可视化形式，全面且直观发现各地区疫情特征。

❑研究数据

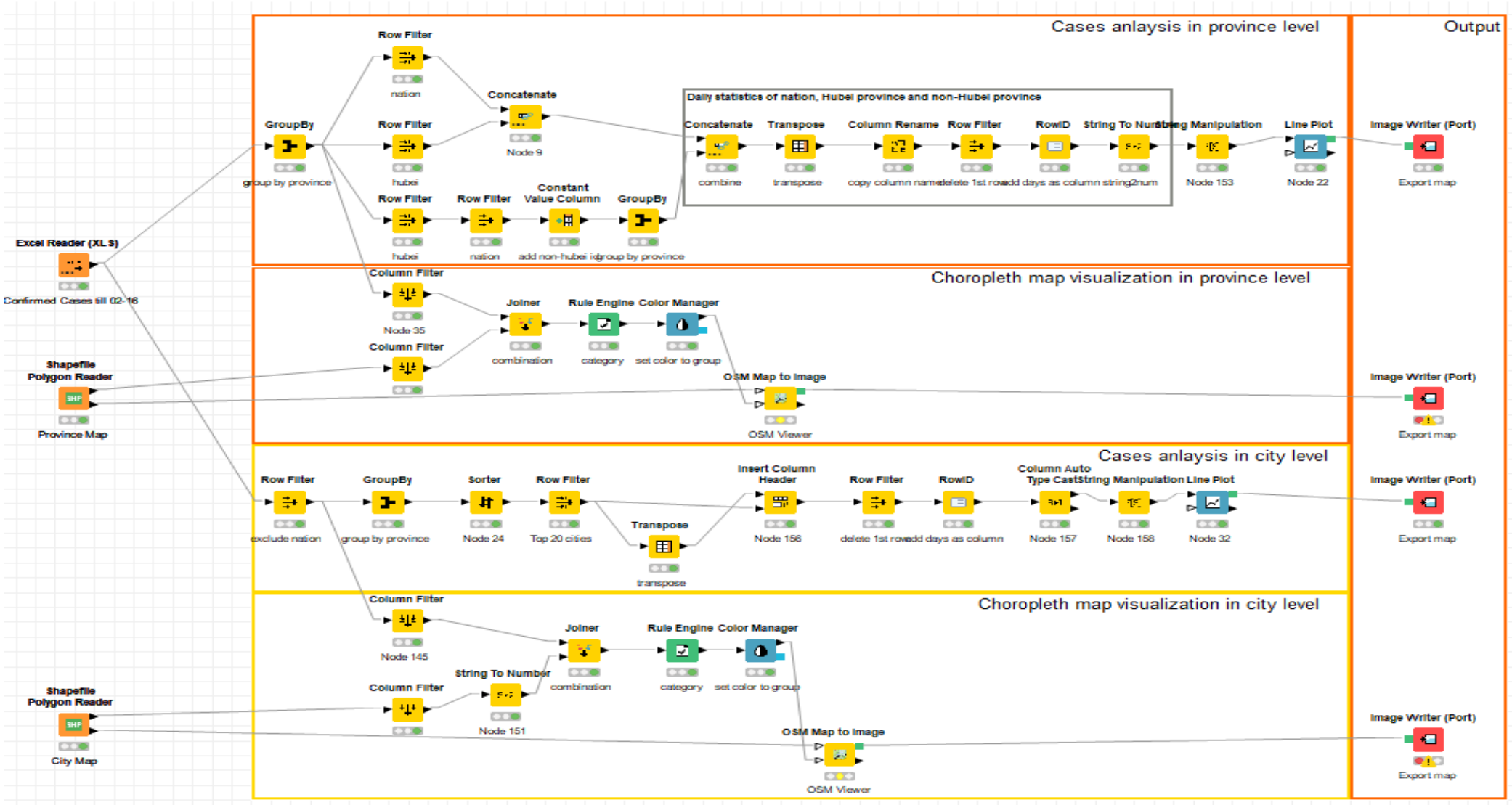
- **中国疫情**
  - **省疫情数据 Province daily cases**
  - **省基础底图 Province basemap**
  - **城市疫情数据 City daily cases**
  - **城市基础底图 City basemap**
- **全球疫情**
  - **全球疫情数据 Global daily cases**
  - **全球基础础底图 Global basemap**

4.1 疫情数据空间可视化分析：分析流程

❑研究方法与流程

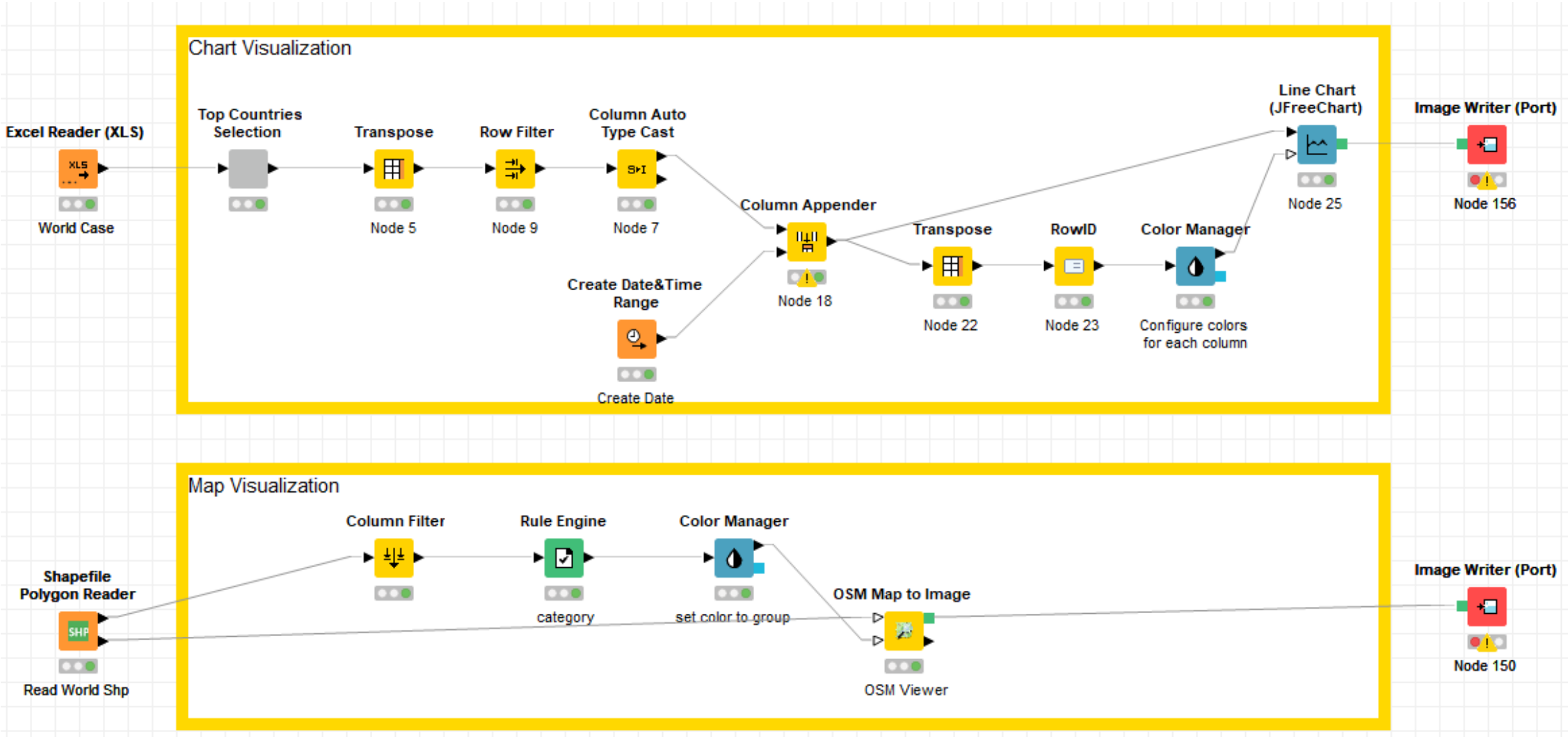


4.1 疫情数据空间可视化分析：工作流实现



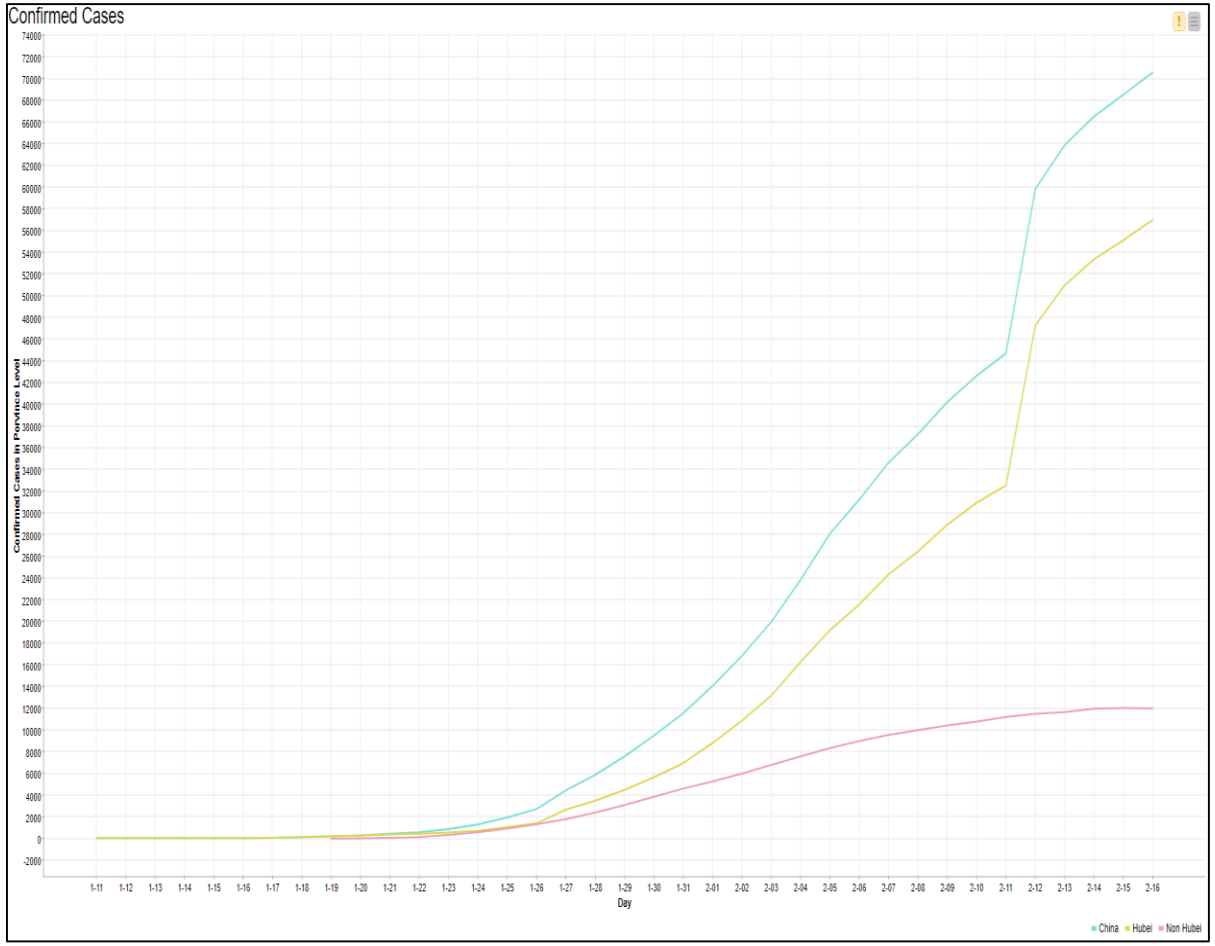
4.1 疫情数据空间可视化分析：地图分析工作流

❑基于工作流的方法实现

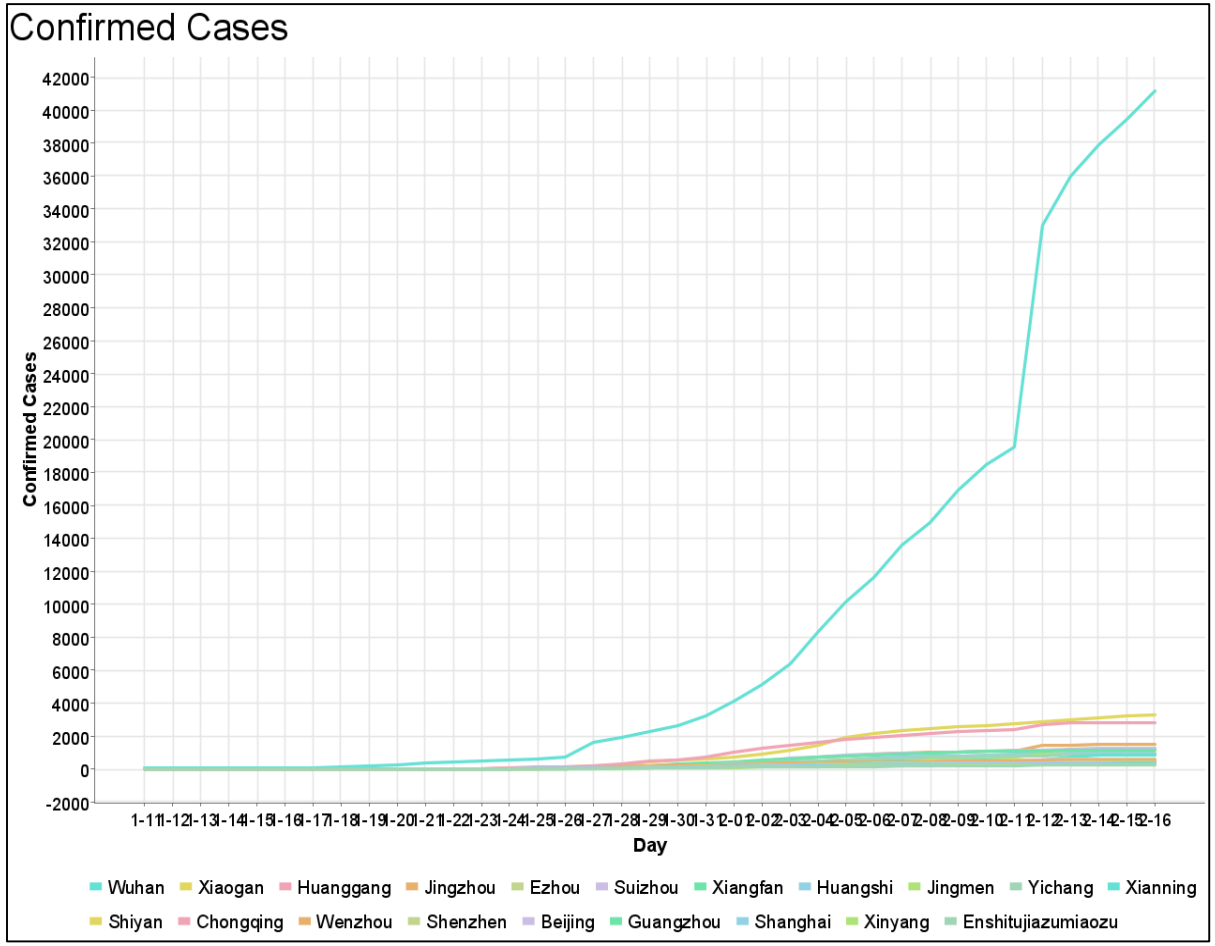


4.1 疫情数据空间可视化分析：统计图分析

全国、湖北省及非湖北省确诊病例变化（2020.2.16）



武汉市及其他前20名城市确诊病例变化（2020.2.16）



4.1 疫情数据空间可视化分析：地图分析

确诊病例专题图（2020.2.16）

确诊病例专题图（2020.2.16）

4.1 疫情数据空间可视化分析：数据共享

❑疫情数据与数据分析工作流共享

- Website: http://chinadatalab.org
- World Cases File Folder: Public/Covid/Workflows/03_world_cases_analysis
  - Workflow File: covid19_worldcase_analysis.knwf
  - Presentation file: 03_case_study_Analysis and spatial visualization of COVID-19 statistics.pptx
  - Input Data:
    - ✓ Public/Covid/Workflows/03_world_cases_anlaysis/data/Countries-Confirmed.xlsx
    - ✓ Public/Covid/Workflows/03_world_cases_anlaysis/data/Countries-ConfirmedV1.shp
- China Cases File Folder: Public/Covid/Workflows/00_China_cases_stats
  - Workflow File: covid19_stats_confirmed.knwf
  - Presentation file: 03_case_study_Analysis and spatial visualization of COVID-19 statistics.pptx
  - Input Data:
    - ✓ Public/Covid/Workflows/ 00_China_cases_stats/data/City_Virus_20200216_V2.xls
    - ✓ Public/Covid/Workflows/03_world_cases_anlaysis/data/map/XXX.shp

4.2 人口迁徙与疫情扩散分析：数据来源

❑研究目标

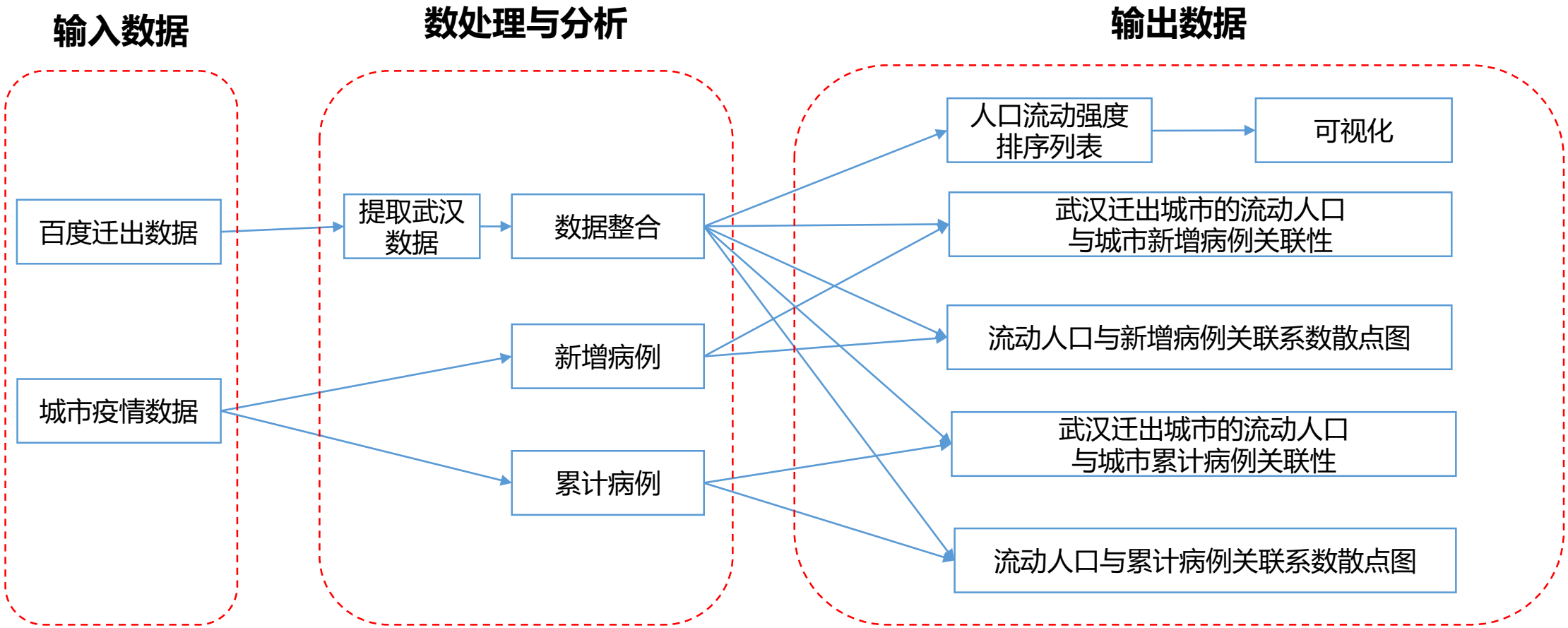
以中国疫情数据为例，基于百度人口迁徙数据集，分析武汉迁出城市的迁徙强度与各大城市疫情爆发的关联情况，具体包括：

- 分析武汉迁出城市的流动人口
- 研究从武汉迁出城市的流动人口与城市新增病例关联性
- 研究从武汉迁出城市的流动人口与城市累计病例相关性

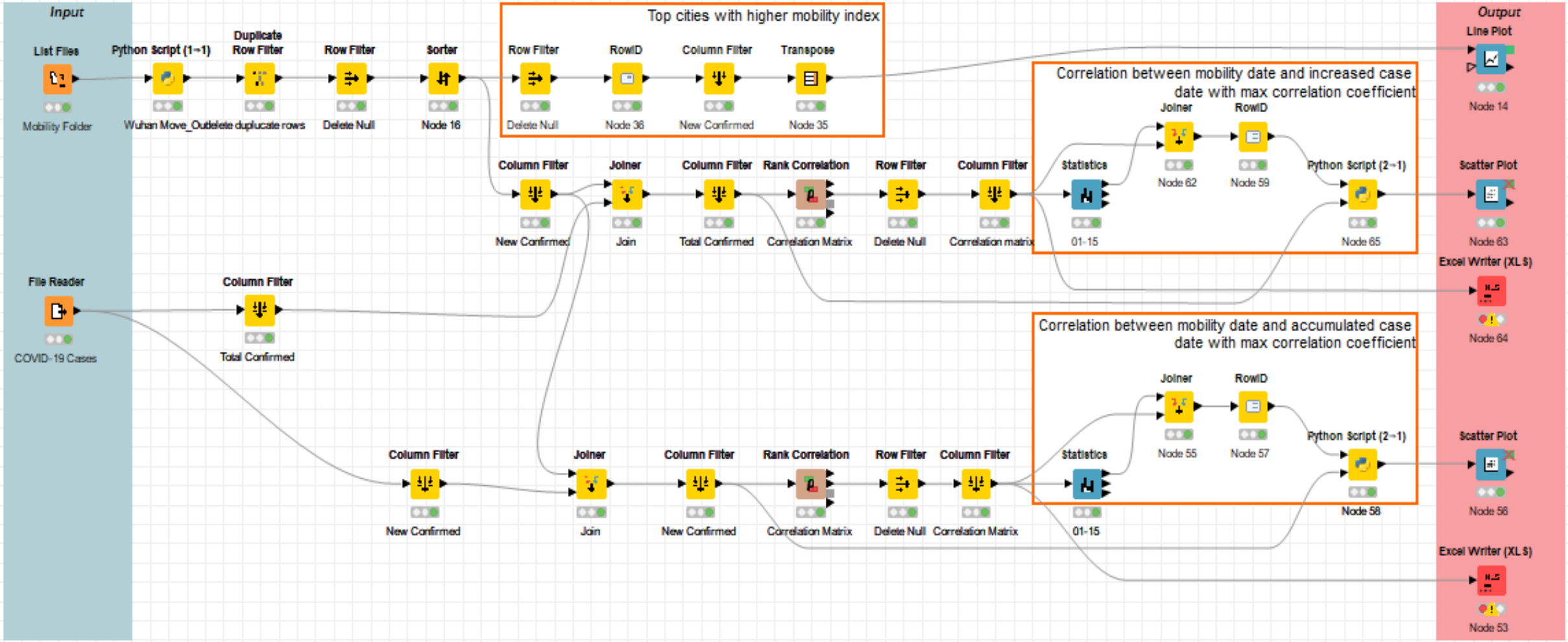
❑研究数据

- **百度城市间迁徙数据**
- **中国城市疫情时序数据**

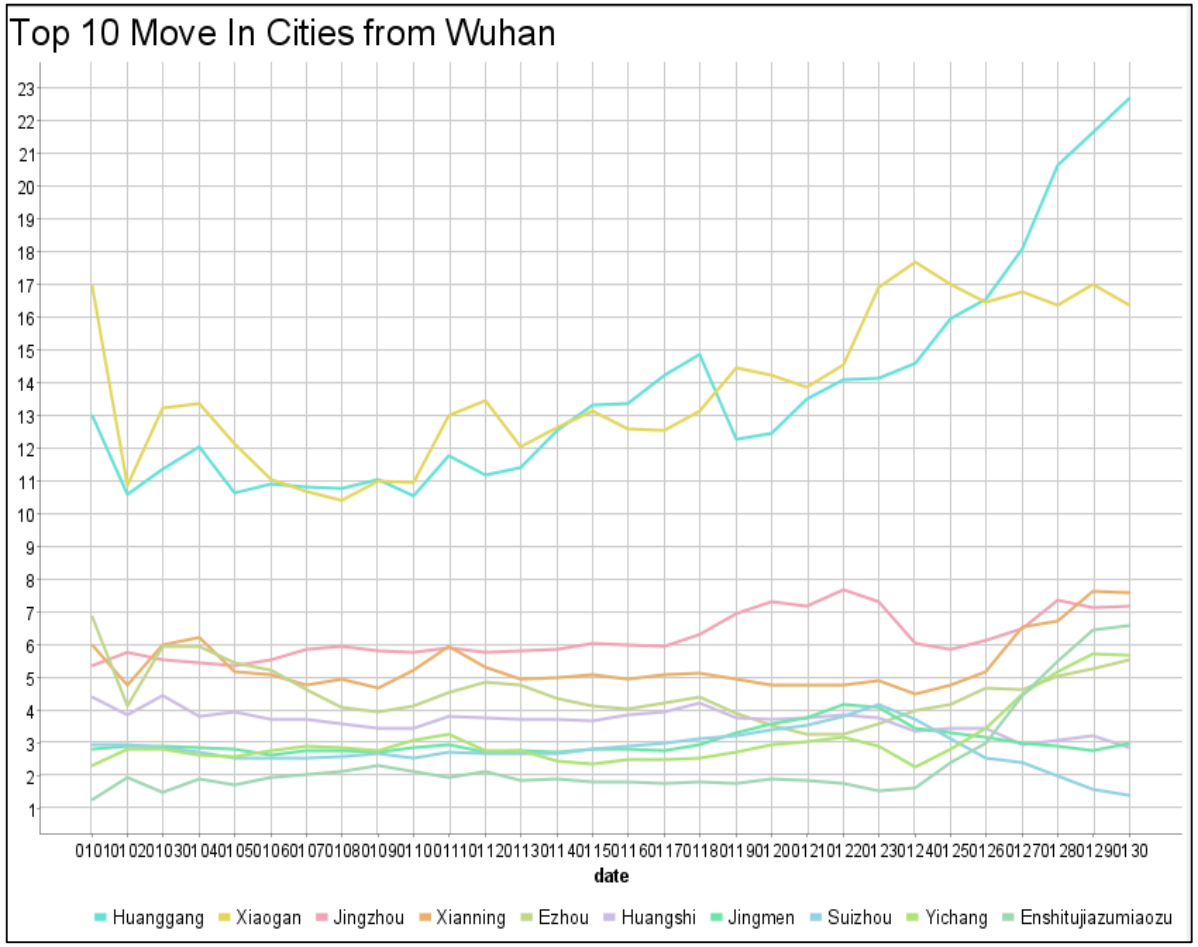
4.2 人口迁徙与疫情扩散分析：分析流程

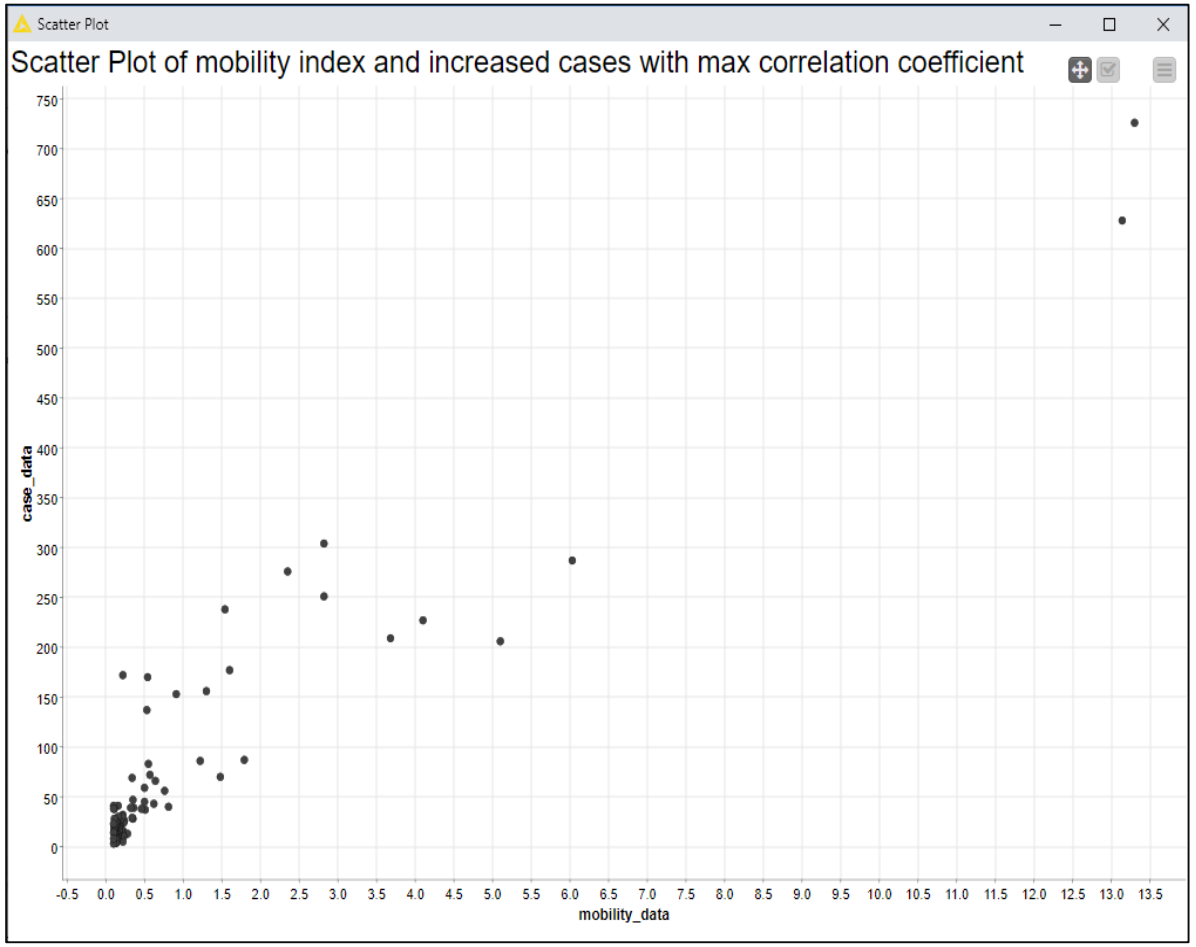


4.2 人口迁徙与疫情扩散分析：工作流实现



4.2 人口迁徙与疫情扩散分析：统计图分析



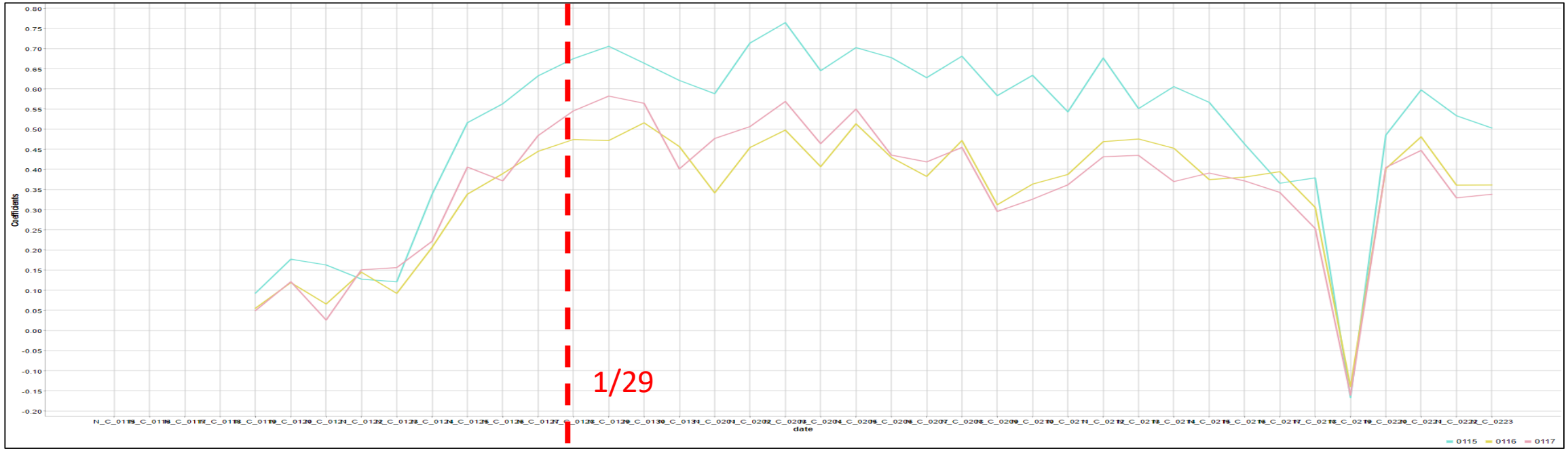


4.2 人口迁徙与疫情扩散分析：相关分析

The correlation coefficients of mobile population from Wuhan and new confirmed cases of receiving 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...	-0.1667905256373...	0.072990881810008...	0.158934565045 41...	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...	0.069277309558 45...	0.069345600961 31...	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...	0.115428852863 83...	0.087050785118 83...	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...	0.046034847695 71...	0.104188970691 83...	-0.11702202047283...
N_C_0122	0.1274614414397117	0.14467257377933...	0.15072987521105438	0.08750596616854...	0.06168724411958...	0.13536779869815832	0.141359 1005569477	0.14849664034433752	0.1855545891955449	-0.0032648634103...	0.080739374181 6597	-0.09567585246904...	-0.06730277191853...
N_C_0123	0.12094027626190487	0.092031778712 50...	0.15613638069644778	0.176418133487 57...	0.11319320082774...	0.06952461789157581	0.07098477547948898	0.06188941958765155	0.007802619799675...	0.00184096833775...	0.032948095118 96...	0.060750661692 14...	0.136484224141464
N_C_0124	0.3382522832151056	0.2063297917117872	0.221540172904987	0.254075299773 10...	0.22346310420578...	0.18931427572820073	0.2259128070428507	0.2648022482432988	0.1908725519165474	0.20145608636506...	0.144195707356 2417	0.259670595610 49...	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...	0.216990266451 83...	0.222517262980 74...	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...	0.198829474540 62...	0.236443826117 1659	0.07557627725924...
N_C_0127	0.6320335646965454	0.4447103307613721	0.48392575589501274	0.4988687834107277	0.45146484678060...	0.45190821274552606	0.45332674670930134	0.44628966096227657	0.38734619534135833	0.3513025081982499	0.311076976449 58...	0.263250998227 3468	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	0.290560969929 6254	0.265022642442 3889	0.22146623199184...
N_C_0129	0.7056550361654592	0.47150182912957...	0.5819758343791975	0.5595910700875395	0.4243739432444717	0.4259968187488185	0.39328827440018577	0.44221375006415103	0.3456106031332887	0.3226042858810695	0.250603150054 3518	0.313612230372 17...	0.1359589130886309
N_C_0130	0.663659197157656	0.5156421923375882	0.5641315184121813	0.5353552391826458	0.4581138752332417	0.45751184331291317	0.35652106136518785	0.3847422518048375	0.36223627649623646	0.4296029203464742	0.311427445992 4828	0.319498894321 18...	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	0.202128258822 85...	0.165072804006 48...	0.19899287772210...

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



4.2 人口迁徙与疫情扩散分析：数据共享

❑疫情数据与数据分析工作流共享

- 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

4.3 基于工作流的气象数据流程化处理：数据来源

❑研究目标

以中国气象数据为例，将从气象局获取的城市级空气质量、天气数据与中国基础底图数据匹配，自动化读取、转换和导出结构化数据，实现处理过程的可复制与可扩展。

❑研究数据

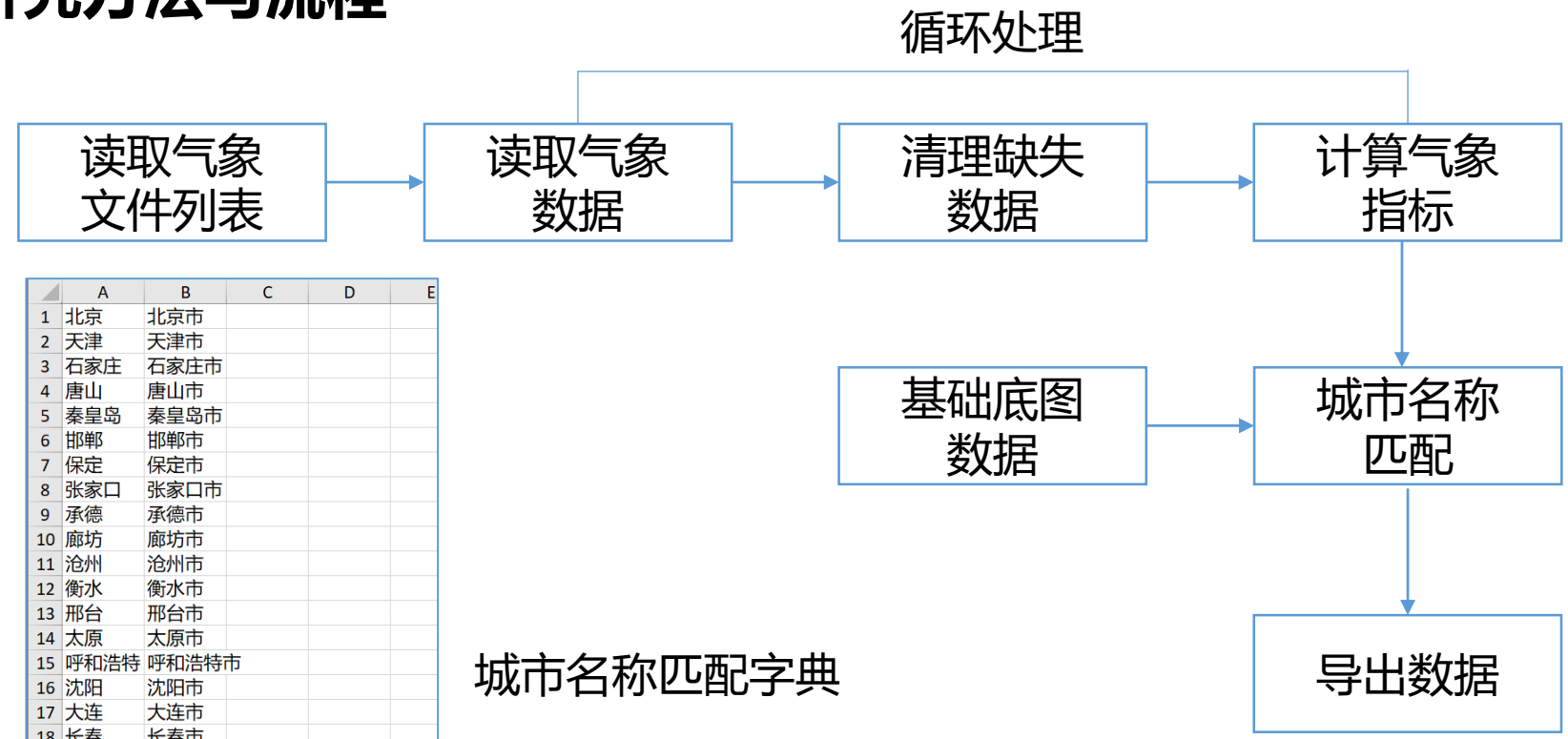
- **气象数据**
  - **数据源**: https://quotsoft.net/air
  - **China Air Quality Data**

    PM2.5 ($\mu g/m^3$), PM10 ($\mu g/m^3$), SO2 ($\mu g/m^3$), NO2 ($\mu g/m^3$), O3 ($\mu g/m^3$), CO (mg/m^3) and AQI
  - **China Meteorological Data**

    温度，气压，湿度，风向，水平气流等
- **基础底图数据**
  - **数据源**：China Data Institute https://chinadatacenter.net/Default.aspx
  - **中国城市基础底图数据**

4.3 基于工作流的气象数据流程化处理：处理流程

❑研究方法与流程

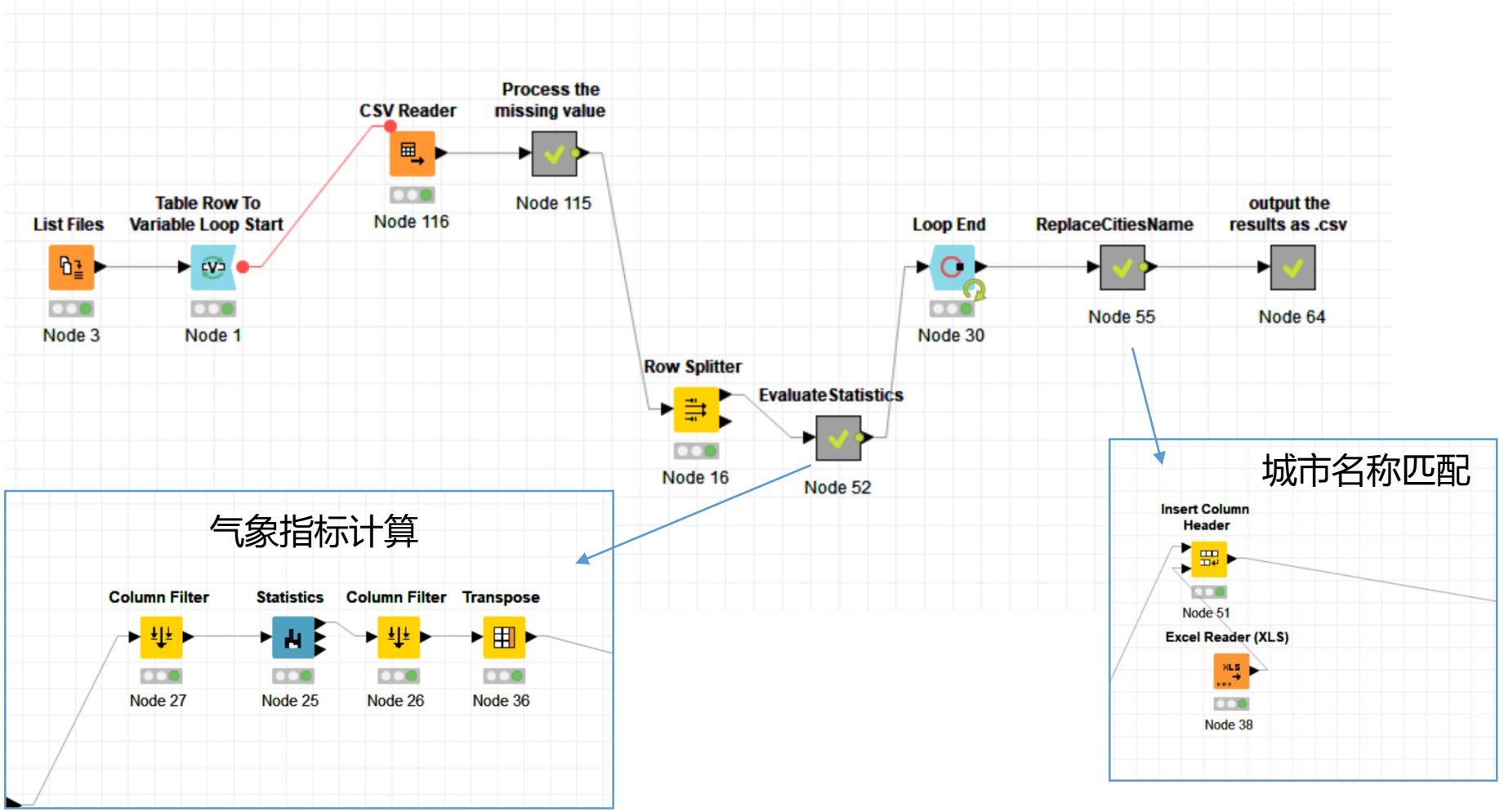


	A	B	C	D	E
1	北京	北京市			
2	天津	天津市			
3	石家庄	石家庄市			
4	唐山	唐山市			
5	秦皇岛	秦皇岛市			
6	邯郸	邯郸市			
7	保定	保定市			
8	张家口	张家口市			
9	承德	承德市			
10	廊坊	廊坊市			
11	沧州	沧州市			
12	衡水	衡水市			
13	邢台	邢台市			
14	太原	太原市			
15	呼和浩特	呼和浩特市			
16	沈阳	沈阳市			
17	大连	大连市			
18	长春	长春市			
19	哈尔滨	哈尔滨市			
20	上海	上海市			
21	南京	南京市			
22	苏州	苏州市			
23	南通	南通市			
24	连云港	连云港市			
25	徐州	徐州市			
26	扬州	扬州市			
27	无锡	无锡市			
28	常州	常州市			
29	镇江	镇江市			
30	泰州	泰州市			
31	淮安	淮阴市			
32	盐城	盐城市			

城市名称匹配字典

4.3 基于工作流的气象数据流程化处理：工作流实现

❑基于工作流的方法实现



	A	B	C	D	E
1	北京	北京市			
2	天津	天津市			
3	石家庄	石家庄市			
4	唐山	唐山市			
5	秦皇岛	秦皇岛市			
6	邯郸	邯郸市			
7	保定	保定市			
8	张家口	张家口市			
9	承德	承德市			
10	廊坊	廊坊市			
11	沧州	沧州市			
12	衡水	衡水市			
13	邢台	邢台市			
14	太原	太原市			
15	呼和浩特	呼和浩特市			
16	沈阳	沈阳市			
17	大连	大连市			
18	长春	长春市			
19	哈尔滨	哈尔滨市			
20	上海	上海市			
21	南京	南京市			
22	苏州	苏州市			
23	南通	南通市			
24	连云港	连云港市			
25	徐州	徐州市			
26	扬州	扬州市			
27	无锡	无锡市			
28	常州	常州市			
29	镇江	镇江市			
30	泰州	泰州市			
31	淮安	淮阴市			
32	盐城	盐城市			

4.3 基于工作流的气象数据流程化处理：处理结果

❑分析与处理结果

- AQI.csv
- CO.csv
- NO2.csv
- O3.csv
- PM2.5.csv
- PM10.csv
- SO2.csv

City_CH	City_EN	FID	Field_1	GbCity	GbProv	Prov_CH	Prov_EN	AQI_Mean_0101	AQI_Mean_0102
北京市	Beijing	0	24	1100	11	北京市	Beijing	51.95833333	70.375
天津市	Tianjin	1	245	1200	12	天津市	Tianjin	81.5	111.0833333
石家庄市	Shijiazhua	2	228	1301	13	河北省	Hebei	97.25	134.125
唐山市	Tangshan	3	244	1302	13	河北省	Hebei	97	113.2916667
秦皇岛市	Qinhuang	4	201	1303	13	河北省	Hebei	62.95833333	106.7083333
邯郸市	Handan	5	86	1304	13	河北省	Hebei	76.70833333	122.0833333
邢台市	Xingtai	6	282	1305	13	河北省	Hebei	83.79166667	116.3333333
保定市	Baoding	7	21	1306	13	河北省	Hebei	79.20833333	113.4166667
张家口市	Zhangjiak	8	318	1307	13	河北省	Hebei	50.29166667	45.54166667
承德市	Chengde	9	40	1308	13	河北省	Hebei	77	101.7083333
沧州市	Cangzhou	10	30	1309	13	河北省	Hebei	87	160.7083333
廊坊市	Langfang	11	143	1310	13	河北省	Hebei	65.41666667	89.375
衡水市	Hengshui	12	98	1311	13	河北省	Hebei	78.45833333	132.375
太原市	Taiyuan	13	241	1401	14	山西省	Shanxi	107.9583333	155.875
大同市	Datong	14	49	1402	14	山西省	Shanxi	111.6666667	123.5416667
阳泉市	Yangquan	15	296	1403	14	山西省	Shanxi	87	106.6666667
长治市	Changzhi	16	323	1404	14	山西省	Shanxi	77.95833333	119.2916667
晋城市	Jincheng	17	128	1405	14	山西省	Shanxi	79.20833333	116.125
朔州市	Shuozhou	18	231	1406	14	山西省	Shanxi	172.7916667	168.5
晋中市	Jinzhong	19	129	1407	14	山西省	Shanxi	100.9583333	132.2083333

Max_value	Min_value	Mean_value	Std_value

4.3 基于工作流的气象数据流程化处理：数据共享

❑气象数据与数据分析工作流共享

- Website: http://chinadatalab.org
- File Folder: Public/Covid/Workflows/80_weather_data
  - Workflow File: covid_weather.knwf
  - Presentation file: Introduction_meteorological_weather_data_0511.pptx
  - Input Data:
    - ✓ Public/Covid/Workflows/80_weather_data/data
    - ✓ Public/Covid/Workflows/80_weather_data/data/city_dict_air.xlsx
  - Output Data:
    - ✓ AQI_max.csv
    - ✓ AQI_mean.csv
    - ✓ AQI_min.csv
    - ✓ AQI_std.csv

4.4 基于社交媒体数据的疫情分析：数据来源

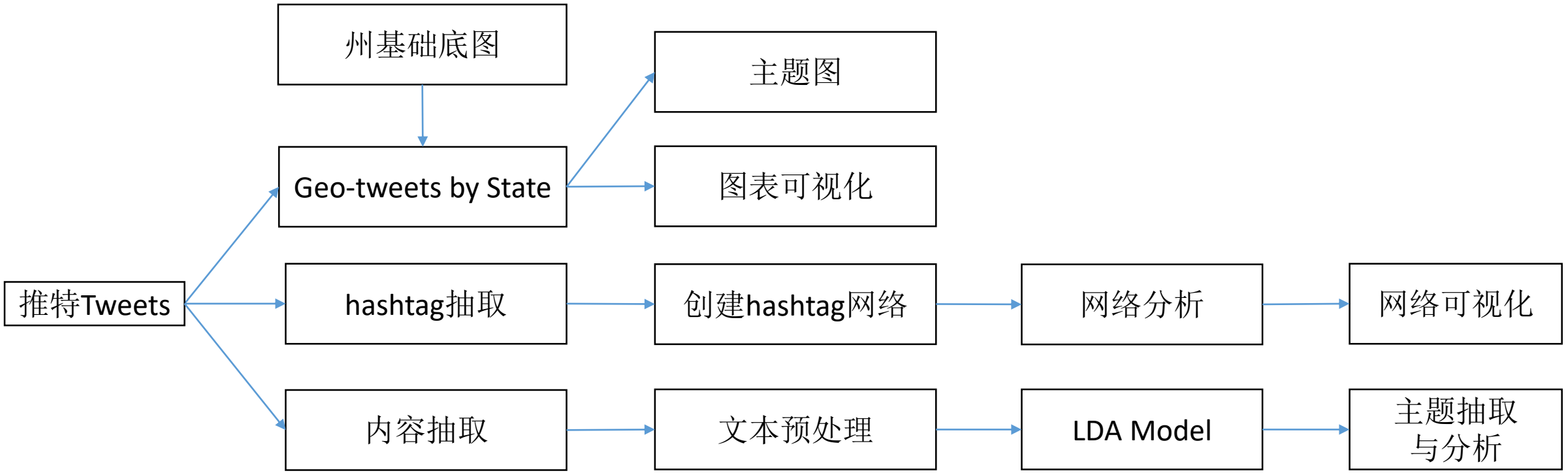
❑研究目标

- 探索美国各州发布COVID-19相关的Tweets相关情况
- 分析美国公众发布COVID-19相关的Tweets中主要讨论的hashtag主题是哪些
- 分析美国公众发布COVID-19相关的Tweets中主要讨论的话题是哪些
- 对比3月17号前后上述分析内容的差别（Coronavirus spreads to all 50 US states on March 17, 2020)

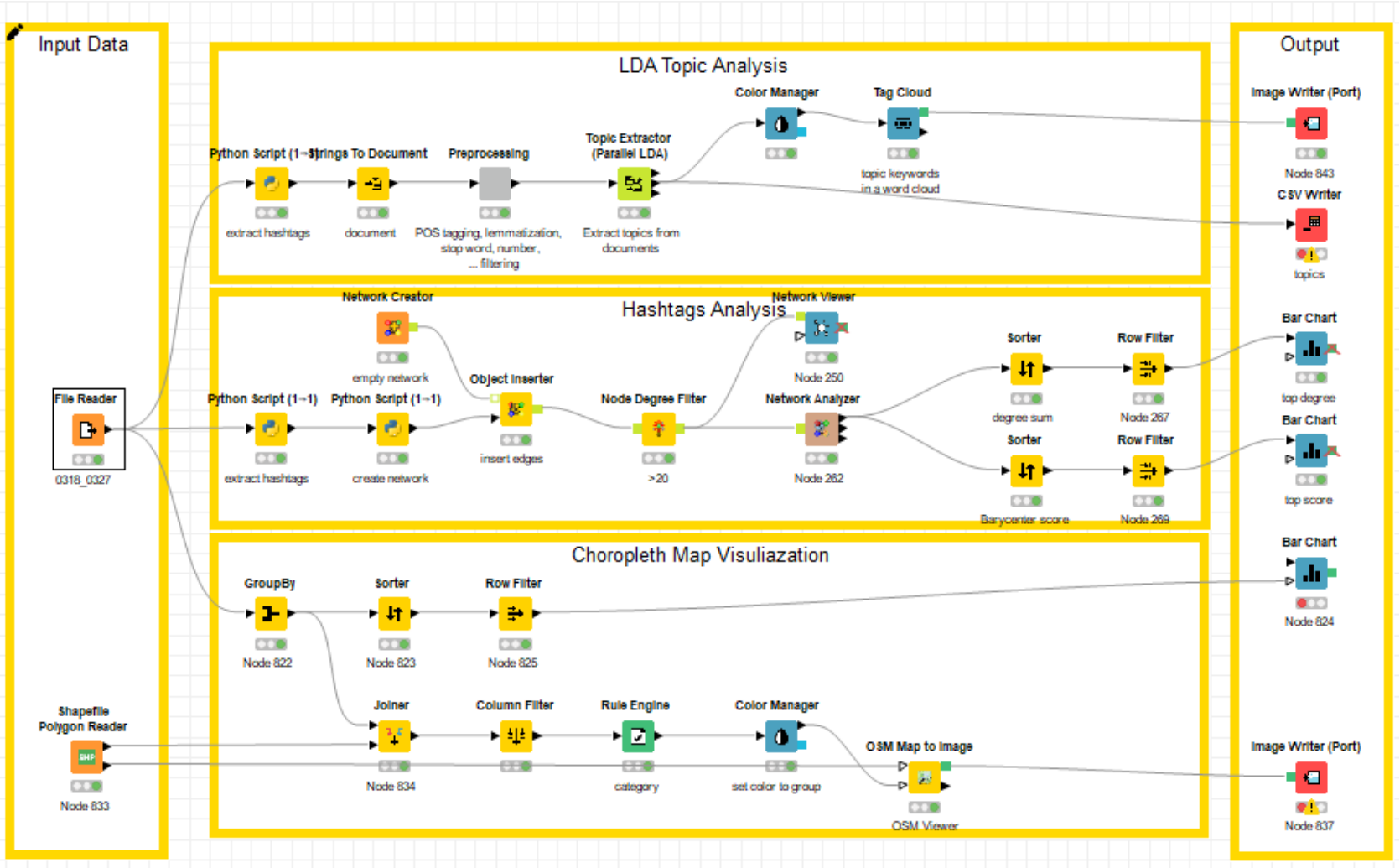
❑研究数据

- Geo-tweets data located at United States in March

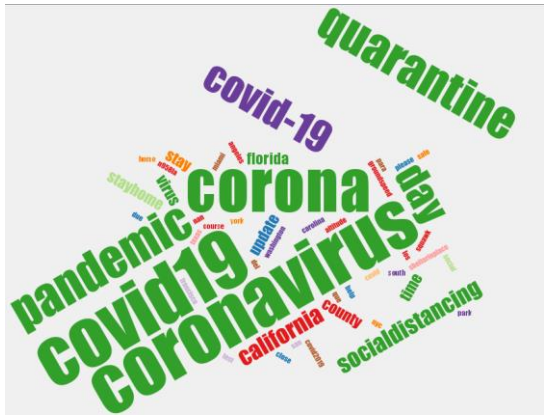
4.4 基于社交媒体数据的疫情分析：分析流程

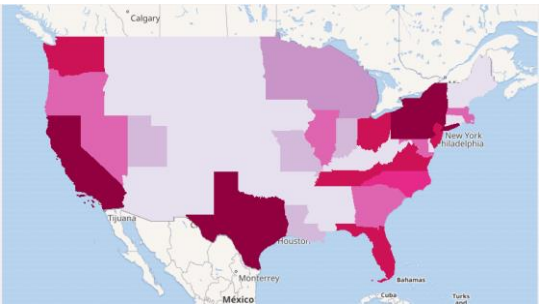


4.4 基于社交媒体数据的疫情分析：工作流实现



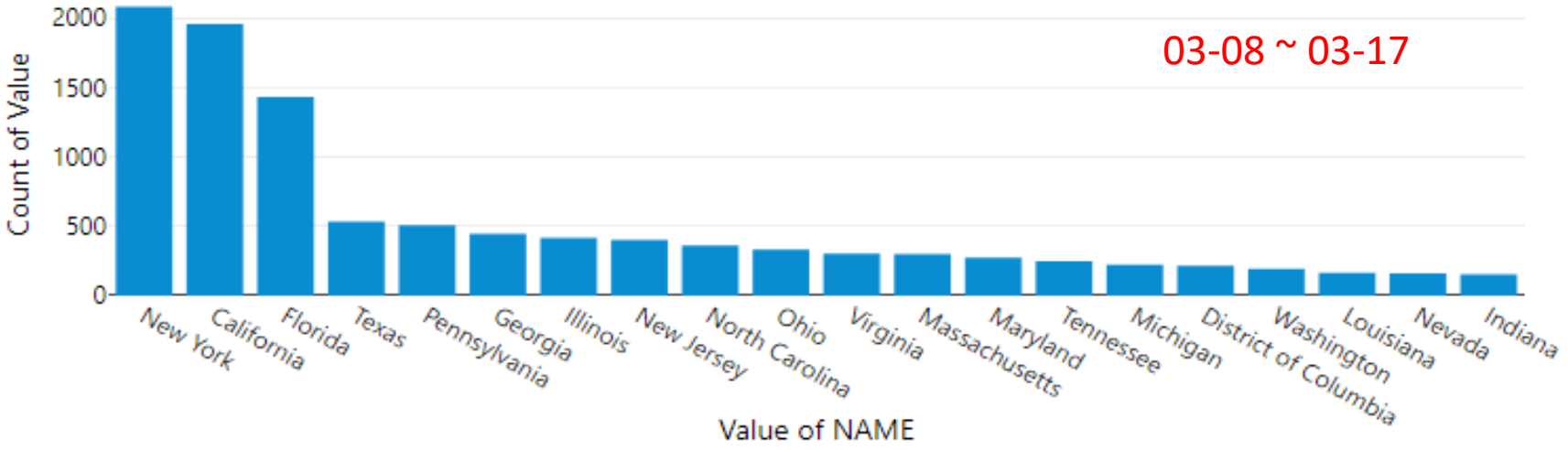
Workflow in KNIME

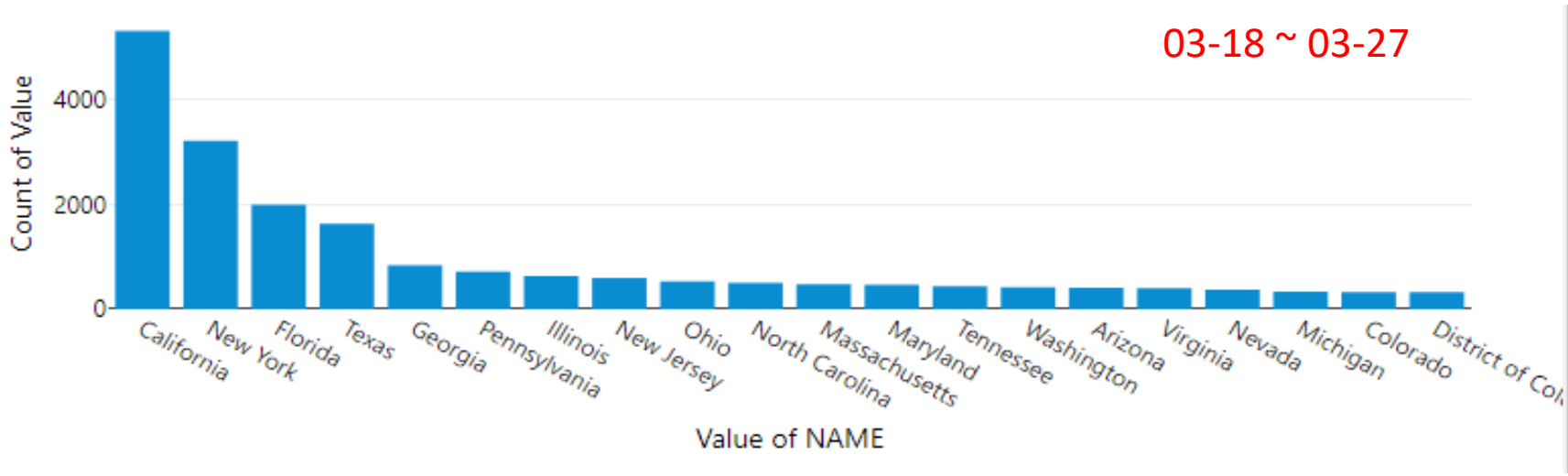




4.4 基于社交媒体数据的疫情分析：分析结果

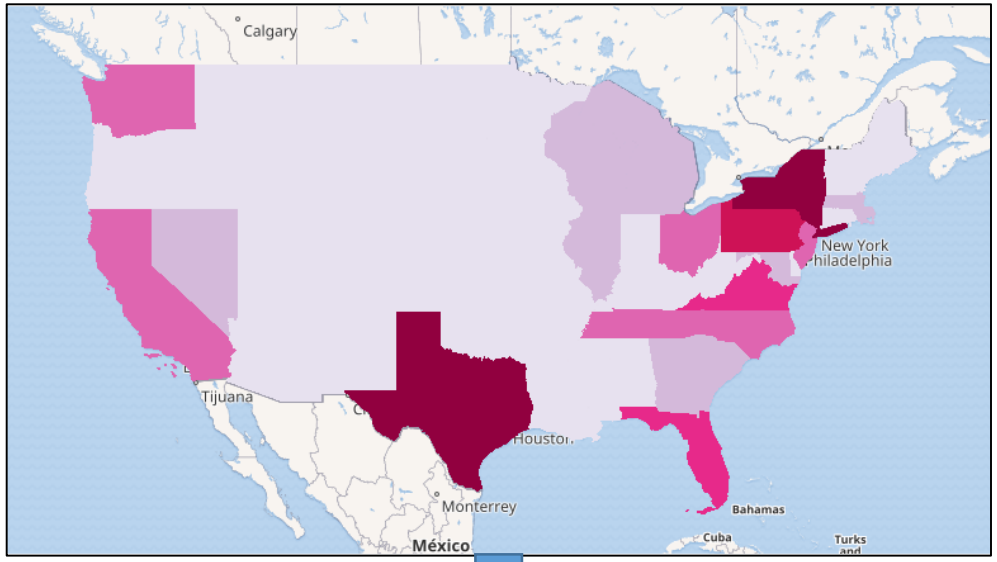
前20州确诊人数排名

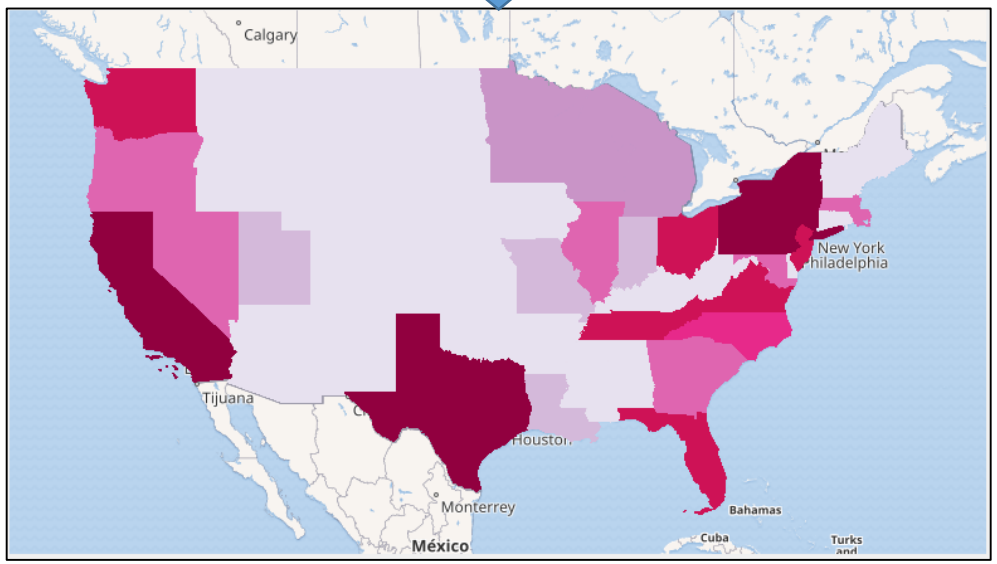




4.4 基于社交媒体数据的疫情分析：地图分析

03/08/2020 ~ 03/17/2020





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

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

03/18/2020 ~ 03/27/2020

4.4 基于文本挖掘方法的社交媒体数据分析: 数据共享

❑数据与工作流共享

- ❑ 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

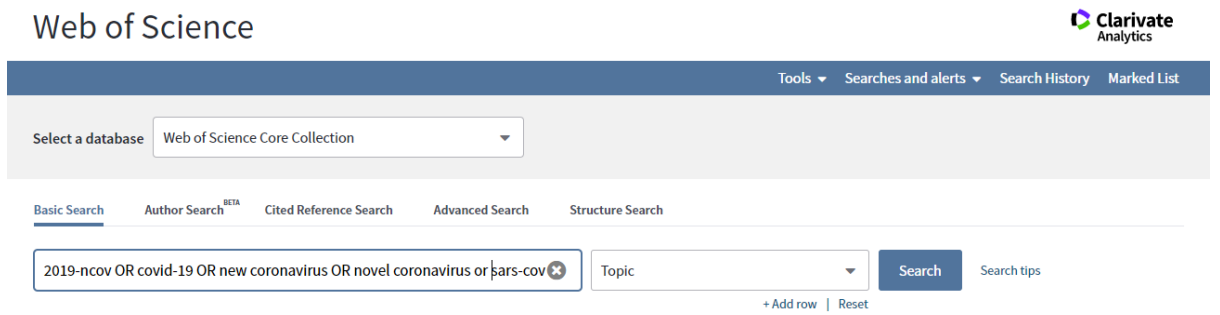
4.5 疫情研究文献分析：数据来源

❑研究目标

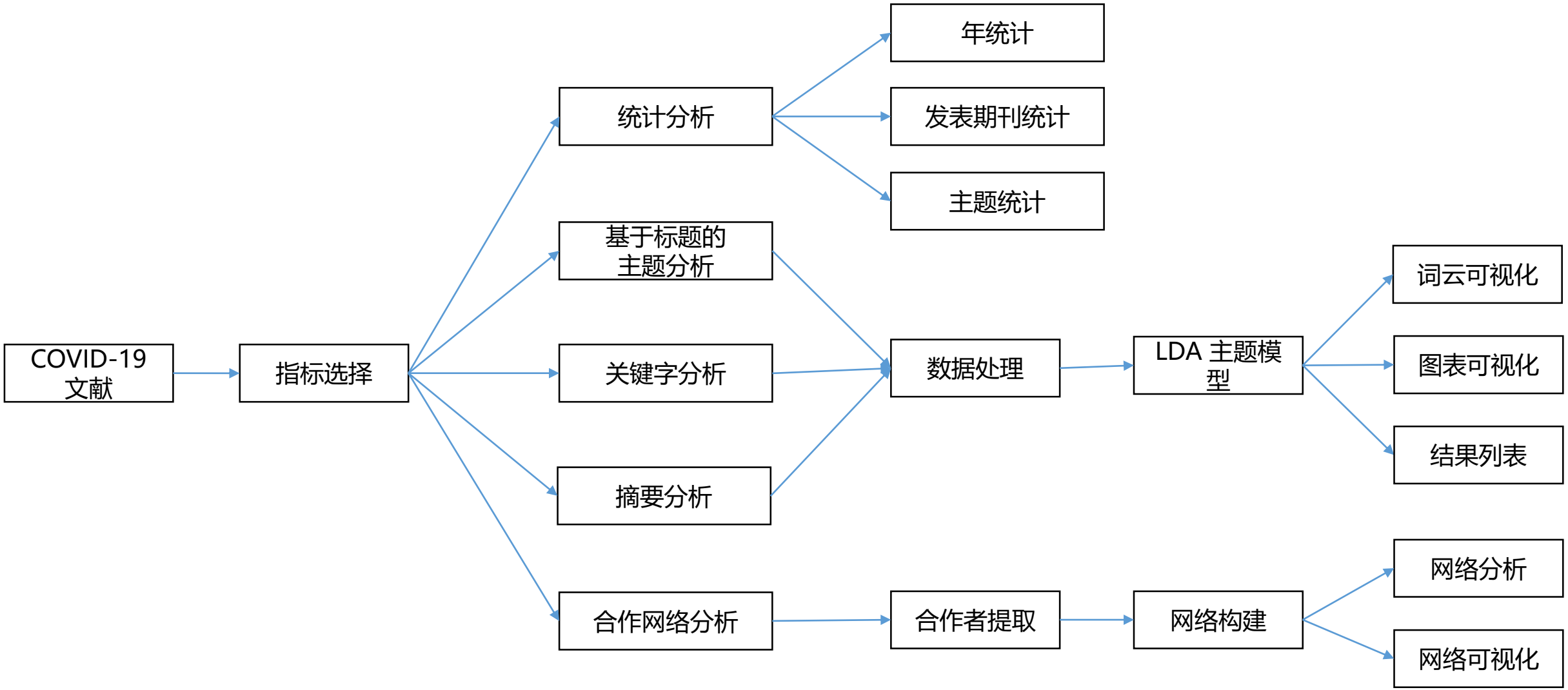
- 探索COVID-19最新文献主要发表在哪些期刊，涉及哪些研究领域
- 分析COVID-19文献的标题主要涉及哪些研究内容
- 分析COVID-19文献的摘要主要讨论哪些研究内容
- 分析COVID-19合作者之间的复杂网络

❑研究数据

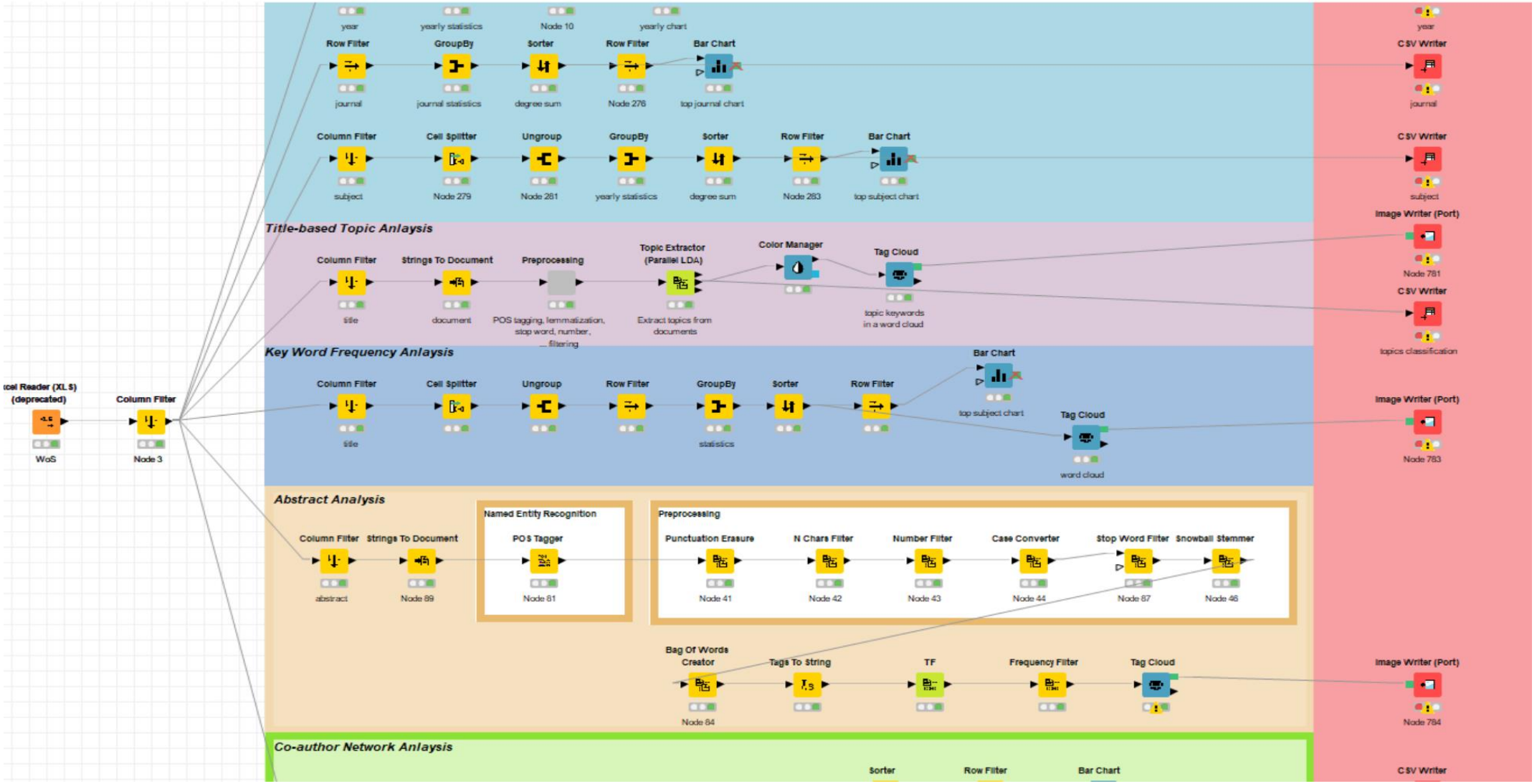
- Web of Science　www.webofknowledge.com
- Key words: "2019-ncov'' OR “covid-19” OR “new coronavirus” OR “novel coronavirus” or “sars-cov-2”.
- Results: author, title, abstract, journal, pub date, subject…



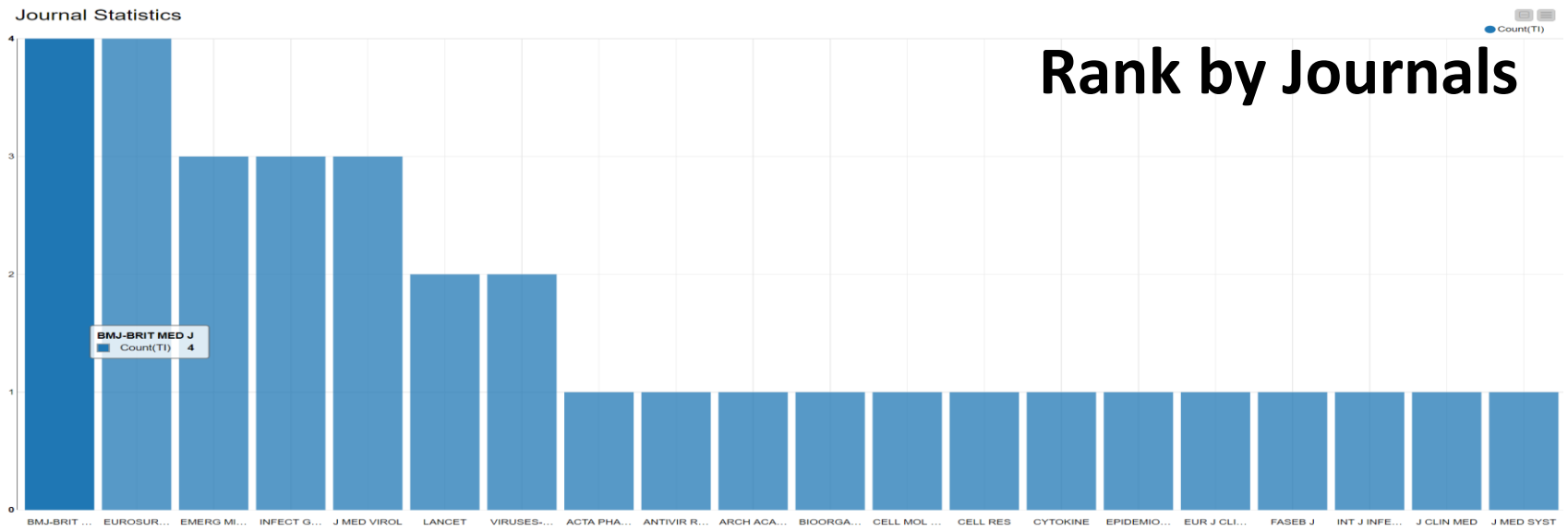
4.5 疫情研究文献分析：分析流程



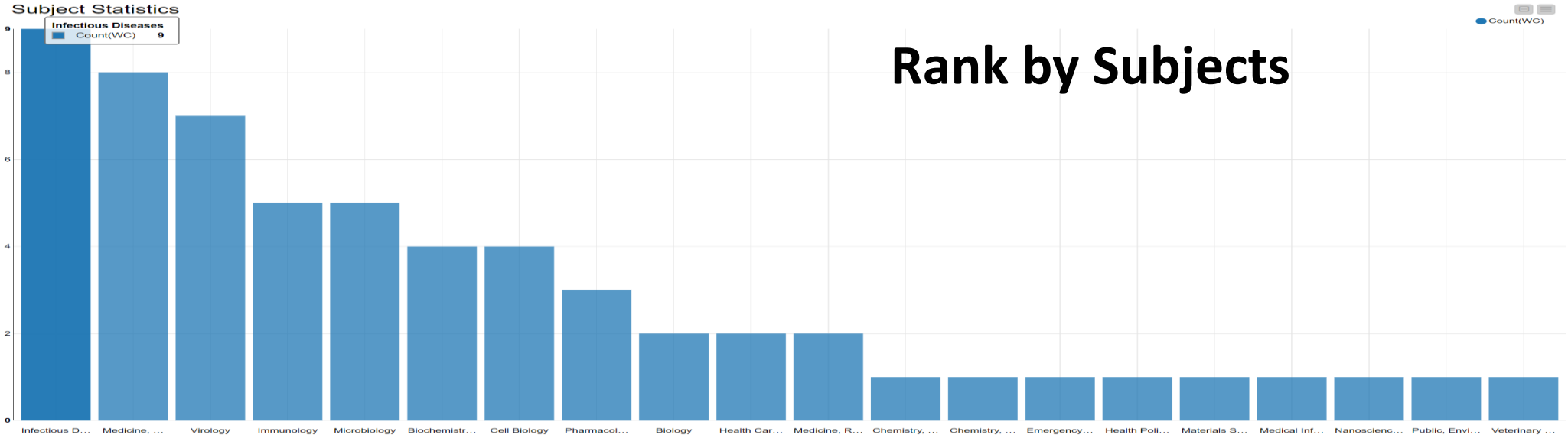
4.5 疫情研究文献分析：工作流实现



4.5 疫情研究文献分析：分析结果

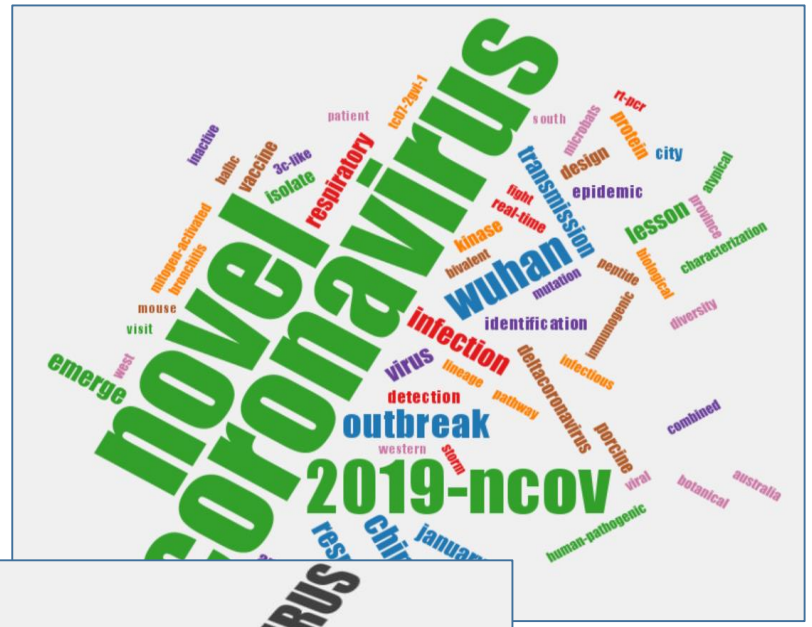


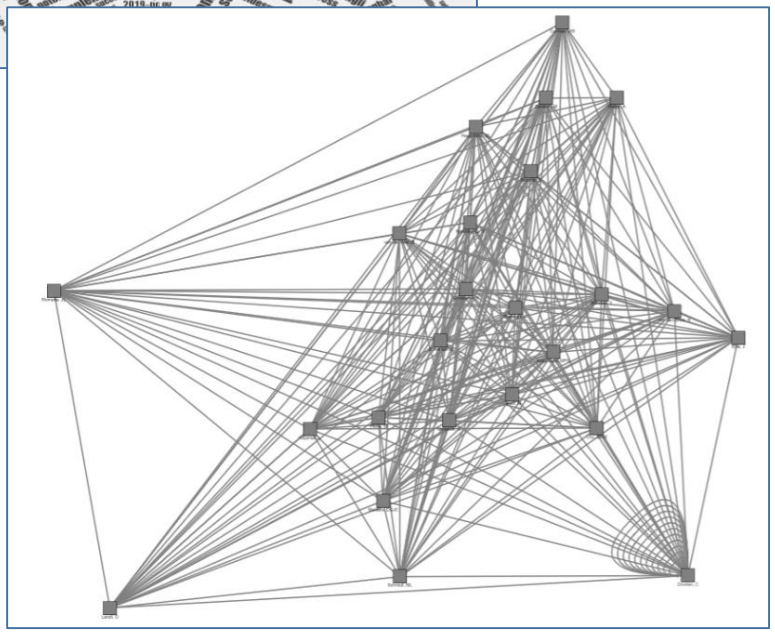
Rank by Journals



Rank by Subjects

4.5 疫情研究文献分析: 词云分析与合作网络分析





4.5 疫情研究文献分析：数据共享

- Website: http://chinadatalab.org
- File Folder: Public/Covid/Workflows/01_wos_publication
  - Workflow File: covid19_wos_pub.knwf
  - Presentation file: Literature+analysis+on+COVID-19.pptx
  - Input Data:
    - Public/Covid/Workflows/ 01_wos_publication /data/wos_pub_ncov.xls
  - Output Data:
    - Journal statistics.jpg:
    - Subject statistics.jpg
    - Keywords_cloudtag.jpg
    - Abstract_cloudtag.jpg
    - Co-author network indicators.csv
    - Title_lda_topics.csv
    - Abstract_lda_topics.csv

5 疫情数据分析工作流共享平台：DataVerse&云平台

https://dataverse.harvard.edu/dataverse/2019ncov

http://chinadatalab.org

5 疫情数据分析工作流调用：DataVerse=>云平台

The Dataverse Project

Open source research data repository software

Enjoy full control over your data. Receive *web visibility, academic credit,* and *increased citation counts*. A personal dataverse is easy to set up, allows you to display your data on your personal website, can be branded uniquely as your research program, makes your data more discoverable to

COVID19 Literature Analysis

Version 2.0

China Data Lab, 2020, "COVID19 Literature Analysis", https://doi.org/10.7910/DVN/3EBJD2, Harvard Dataverse, V2

Cite Dataset

Dataset Metric

Description This dataset stores workflows related to literature analysis. (2020-02-20)

Subject Earth and Environmental Sciences; Computer and Information Science; Social Scien

Keyword workflow

Files | Metadata | Terms | Versions

Search this dataset... Find

Filter by

File Type: All Access: All

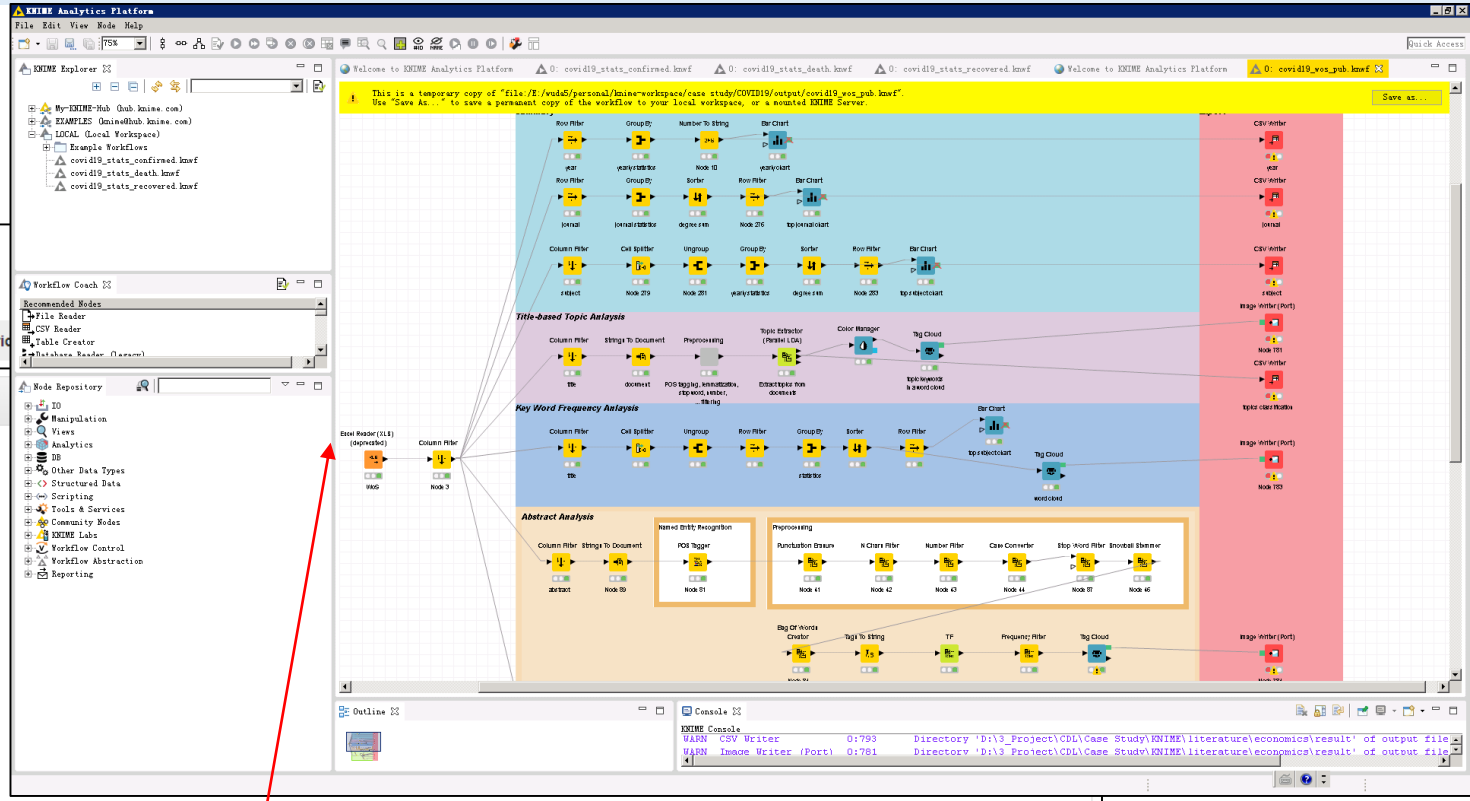
1 to 4 of 4 Files

00_ppt_Literature analysis on COVID-19.pptx
MS Powerpoint - 2.2 MB - Feb 20, 2020 - 243 Downloads
MD5: 88ebbd5858bcd480d0717abaff6c2be2
Case study PPT

01_data_wos_pub_ncov.xls
MS Excel Spreadsheet - 135.0 KB - Feb 20, 2020 - 168 Downloads
MD5: f6b594795c9a935b5abf0e633f7e5649

02_workflow_covid19_wos_pub.knwf
Unknown - 170.0 KB - Feb 20, 2020 - 89 Downloads
MD5: 268c2c5d36ae3c3f419b4682f15fe8fb
This workflow demonstrates the most recent publications related to 2019 nCoV. The acc
veappc://88D2AA413AED9040FAE30B4D4FBE0916B56E161735A76FC2032F86F8E3

README.txt
Plain Text - 186 B - Mar 24, 2020 - 83 Downloads
MD5: d2f50a9ecde6bce3f9bbbad015372ea8



File Metadata

Preview

File Persistent ID

Download URL Use the Download URL in a Wget command or a download manager to avoid interrupted downloads, time outs or other failures.
User Guide - Downloading via URL

`https://dataverse.harvard.edu/api/access/datafile/:persistentId?persistentId=doi:10.7910/DVN/3EBJD2/EZLA2L`

MD5 268c2c5d36ae3c3f419b4682f15fe8fb

Publication Date 2020-02-20

Size 170.0 KB

Type Unknown

Description This workflow demonstrates the most recent publications related to 2019 nCoV. The access link is veappc://88D2AA413AED9040FAE30B4D4FBE0916B56E161735A76FC2032F86F8E33C3585A38FAF311C47CCD3473BF961D7(

Deposit Date 2020-02-20

参与人员/Participants

Tao Hu[1], Wendy Guan[1], Xinyan Zhu[2,3], Yuanzheng Shao[2,3], Shuming Bao[4], Lingbo Liu[5], Jing Du[2], Hongqiang Liu[6], Huan Zhou[7], Jialei Wang[7], Bing She[8], Luyao Zhang[9], Zhibin Li[1,10], Peixiao Wang[11], Yicheng Tang[12], Ruizhi Hou[9], Yun Li[13], Dexuan Sha[13], Yifan Yang[14], Ben Lewis[1], Devika Kakkar[1]

[1] 哈佛大学 Center for Geographic Analysis, Harvard University, Cambridge, MA 02138, USA;
[2] 武汉大学 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan, China;
[3] 武汉大学 Collaborative Innovation Center of Geospatial Technology, Wuhan University, 129 Luoyu Road, Wuhan, China;
[4] 中国数据研究所 China Data Institute, University of Michigan, Ann Arbor, MI, US;
[5] 武汉大学 School of Urban Design, Wuhan University, No.8 East lake South Road, Wuchang District, Wuhan, Hubei, China;
[6] 山东科技大学 College of Geomatics, Shandong University of Science and Technology, Qingdao, China;
[7] 武汉大学 School of Geodesy and Geomatics, Wuhan University, 129 Luoyu Rd. Wuhan, Hubei, China;
[8] 密西根大学 University of Michigan, Ann Arbor, MI 48106;
[9] 华东师范大学 East China Normal University, 3663 Zhongshanbei Road, Putuo District, Shanghai, China;
[10] 北京大学 School of Government, Peking University, Beijing, China;
[11] 福州大学 The Academy of Digital China, Fuzhou University, Fuzhou, China;
[12] 合肥工业大学 School of Management, Hefei University of Technology, Hefei, China;
[13] 乔治梅森大学 Department of Geography and GeoInformation Science, George Mason University, Fairfax, VA, USA;
[14] 加州大学圣地亚哥分校 University of California, San Diego, 9500 Gilman Dr, La Jolla, CA, USA;

致谢/Acknowledgement

Geocomputation Center for Social Sciences, Wuhan University
武汉大学社会地理计算中心

Center for Geographic Analysis, Harvard University
哈佛大学地理分析中心

RMDS Lab 研究方法与数据科学实验室

China Data Institute 中国数据研究所

Amazon.com, Inc 亚马逊

MicroSoft China 微软中国

Vesystem Inc.和信创天科技有限公司

Knowledge Sharing Inc.百智享科技有限公司

All China Marketing Research Inc. 华通人市场研究有限公司

联系方式/Contact

Resources for Coronavirus Study

http://chinadatalab.net

The Cloud for COVID-19 Study

Http://chinadatalab.org

spatialdatalab@lists.fas.harvard.edu

加入疫情研究群与注册系列讲座

加入疫情研究群

在线系列讲座注册