China Data Institute

Center for Geographic Analysis
Harvard University

NSF Spatiotemporal Innovation Center

Geo-computation Center for Social Sciences
Wuhan University

Future Data Lab

Replicable and Reproducible Research with Workflow

Dr. Tao Hu, Harvard University
Dr. Hanchen Yu, Arizona State University
Vincent Chen, RMDS

06/22/2021

Agenda

- Background
- Data Access with Workflow
- Case Study with Workflow
- Discussion

Motivation

REPRODUCIBILITY, REPLICABILITY, AND GENERALIZATION IN THE SOCIAL, BEHAVIORAL, AND ECONOMIC SCIENCES

REPRODUCIBILITY: the ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator.

REPLICABILITY: the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected.

GENERALIZABILITY: whether the results of a study apply in other contexts or populations that differ from the original one

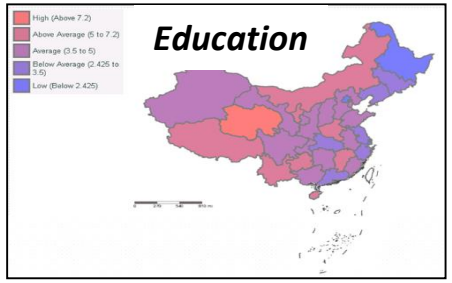
Report of the Subcommittee on Replicability in Science of the SBE Advisory Committee to the National Science Foundation
13 May 2015 Presentation
at SBE AC Spring Meeting by K. Bollen.

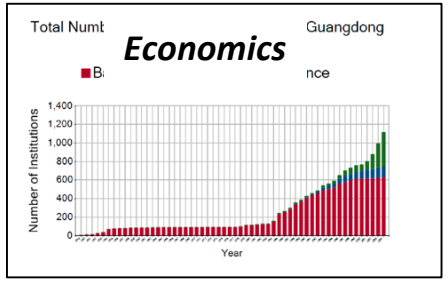
Challenges

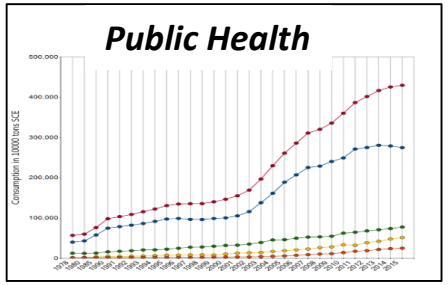
A survey conducted by ***Nature*** found that more than **70%** of researchers have failed to reproduce another scientist's studies, and more than **50%** have failed to reproduce their own experiments (Baker, 2016).

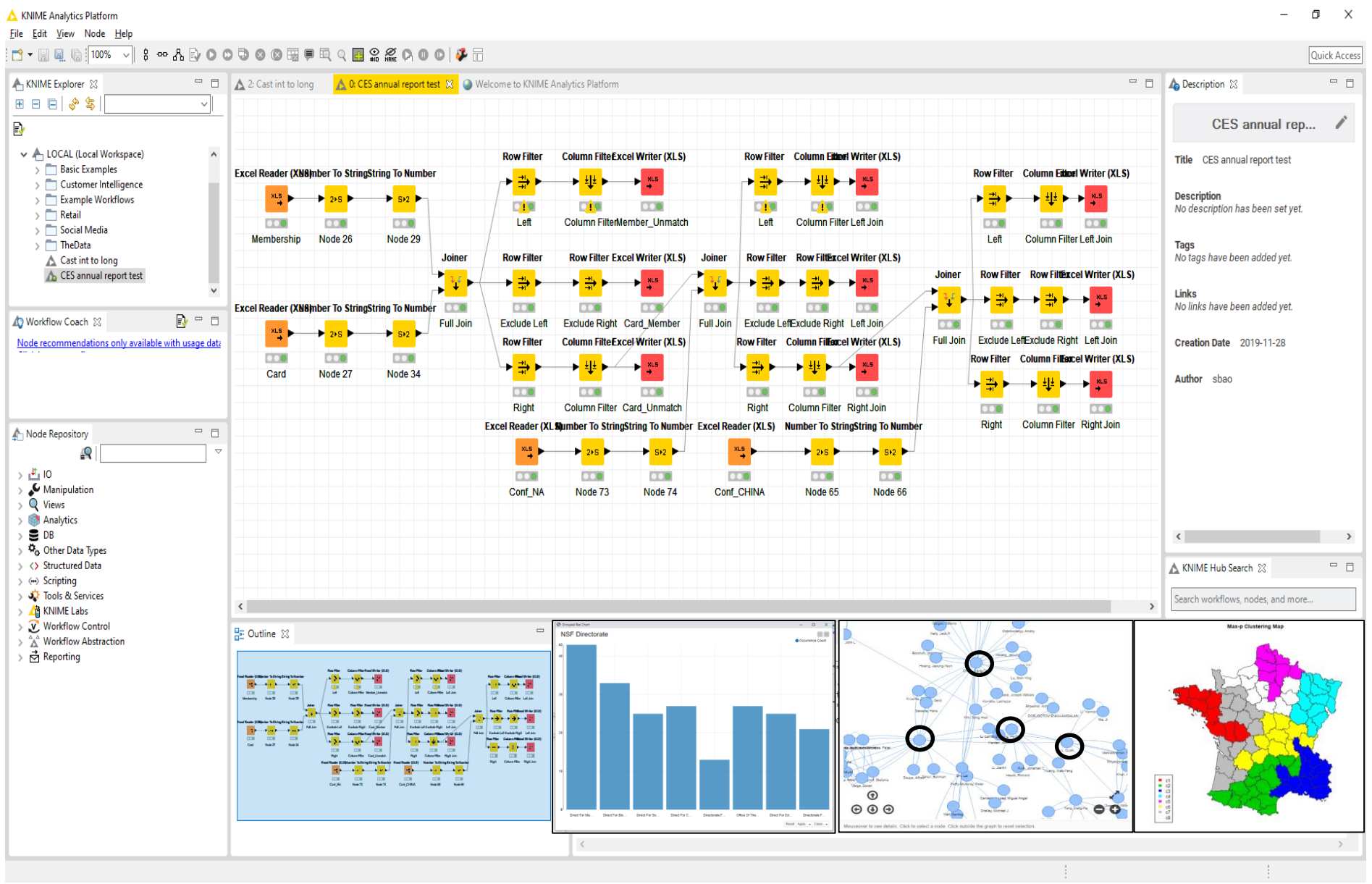
- ❑ **Unavailable dataset**
- ❑ **Different environment**
  - Operation system
  - Computing environment
  - Software
  - Library versions
- ❑ **Lacks documentation**

Reproducibility, Replicability, and Expansion with Workflow







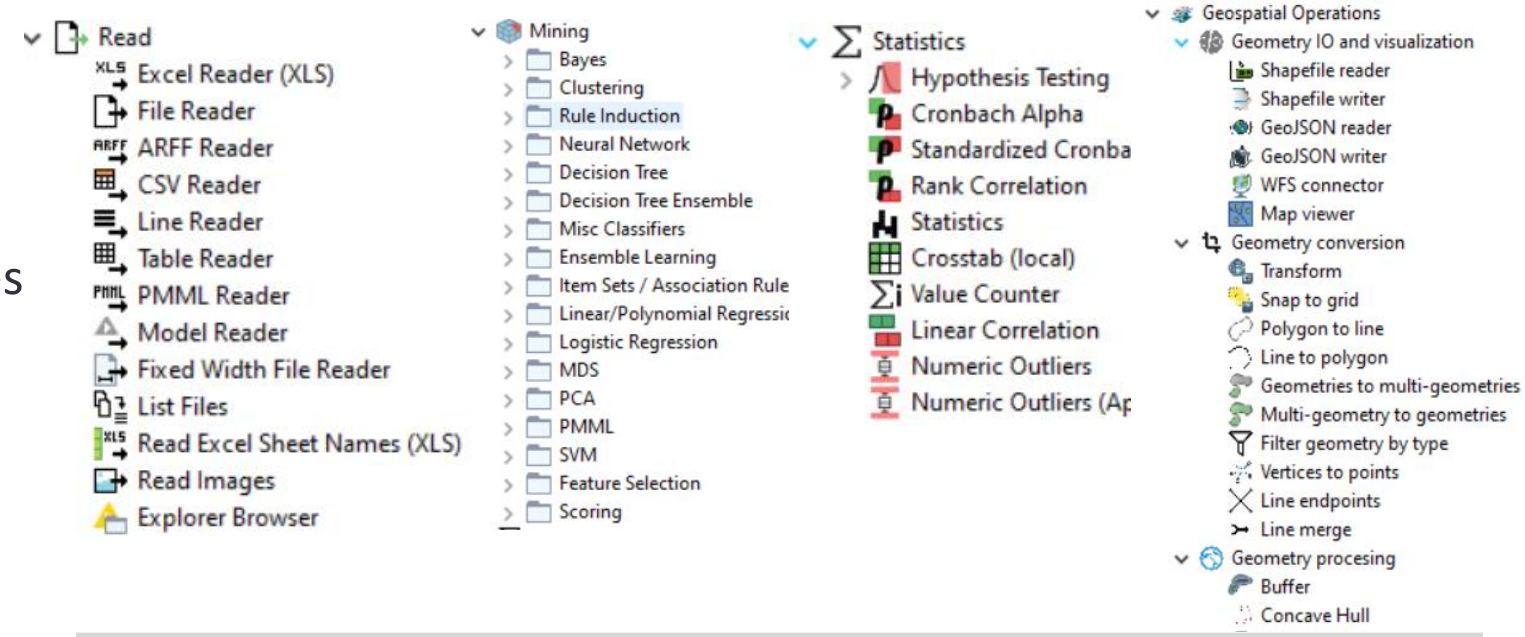


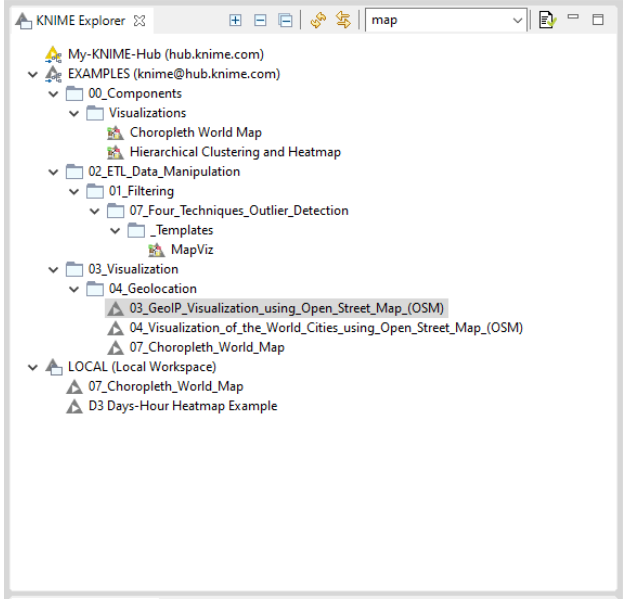
Data Science Workflow Tool KINME

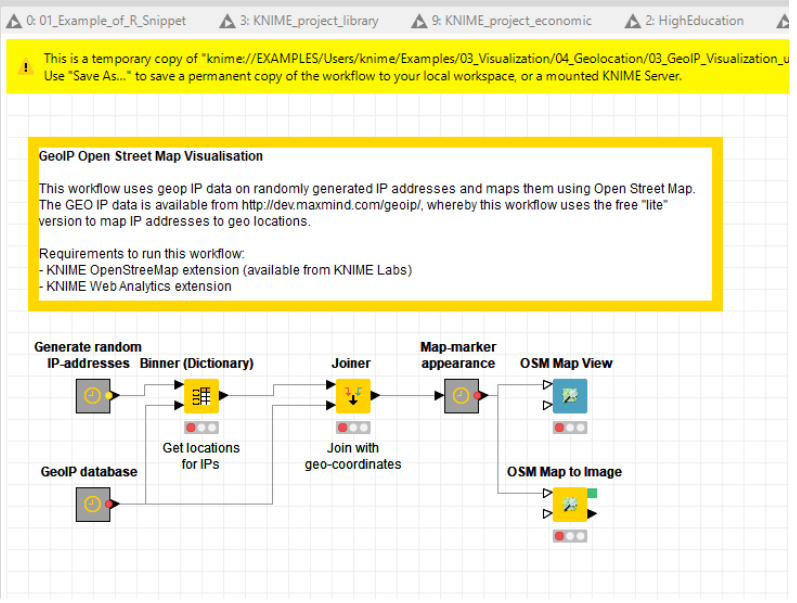
Open for Innovation
KNIME

https://www.knime.com/

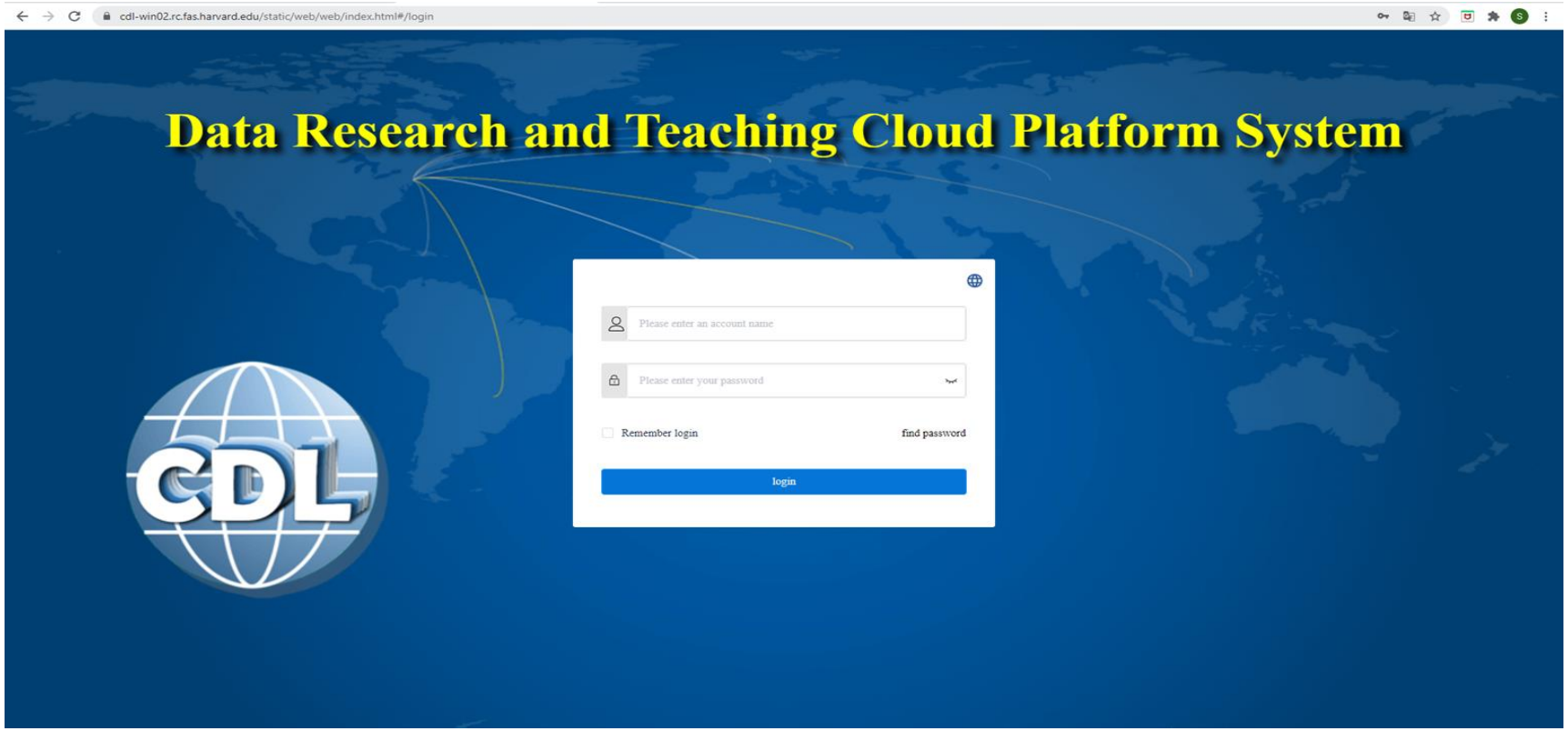
- Workflow is displayed as connected nodes which makes it easy to troubleshoot and visualize
- Easy to use without much knowledge of coding
- Great extensions for data preprocessing, analysis, and visualization
- Connection to other languages, such as JS, R, Python, etc.
- Open-source
- Cross platform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

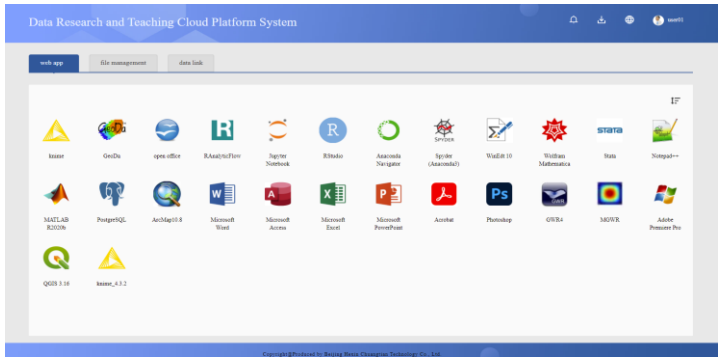


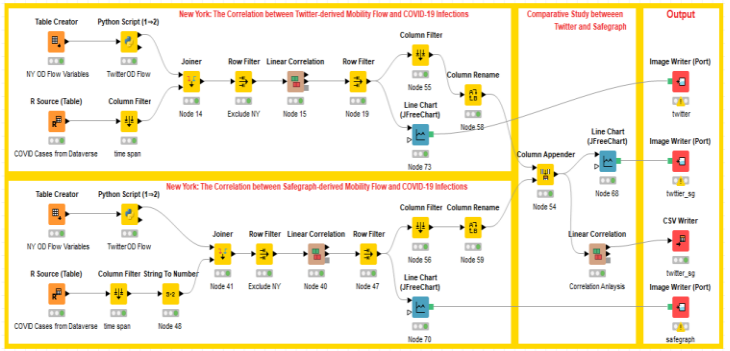




Spatial Data Lab







- ❑ Integrated hundreds of datasets with thousands of variables
- ❑ Developed more than 60 workflow-based case studies for spatiotemporal research
- ❑ Comprehensive documents for easy reproducibility, and replicability

Data Access with Workflows

- ❑ China Data Online platform
- ❑ Database (Postgres)

Database

- ❑ External Platforms with Rest API
- ❑ Harvard Dataverse

Web API

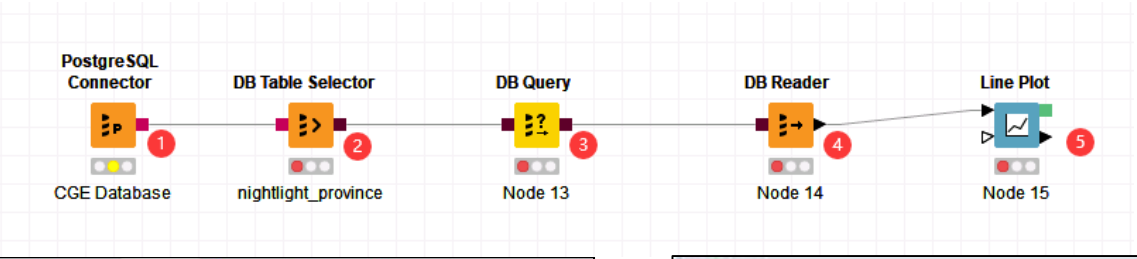
- ❑ Data on GitHub

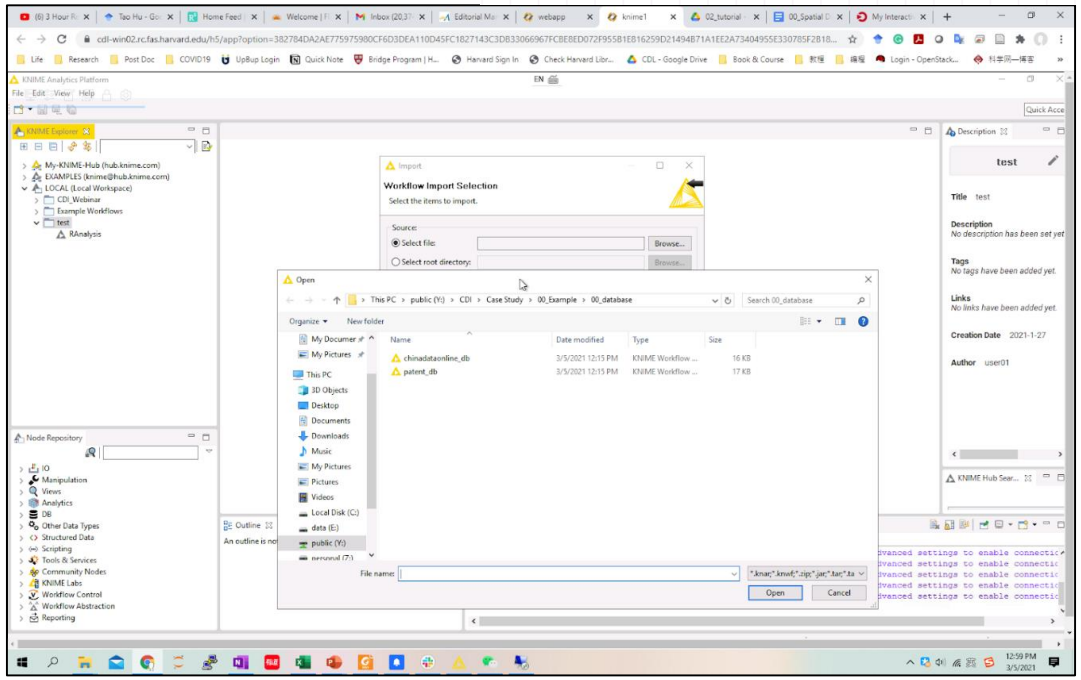
Open Data Repository

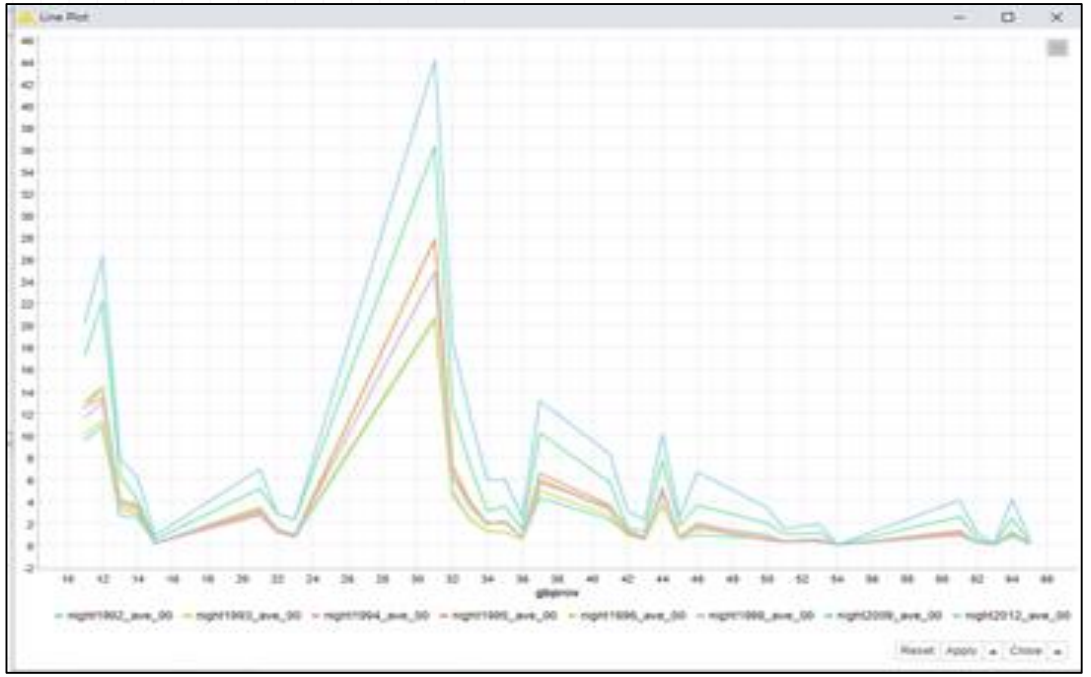
Data Access with Workflows

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

Ex: Import Workflow from Public Folder **\\CDI\Case Study\00_Example\00_database\chinadataonline_db**



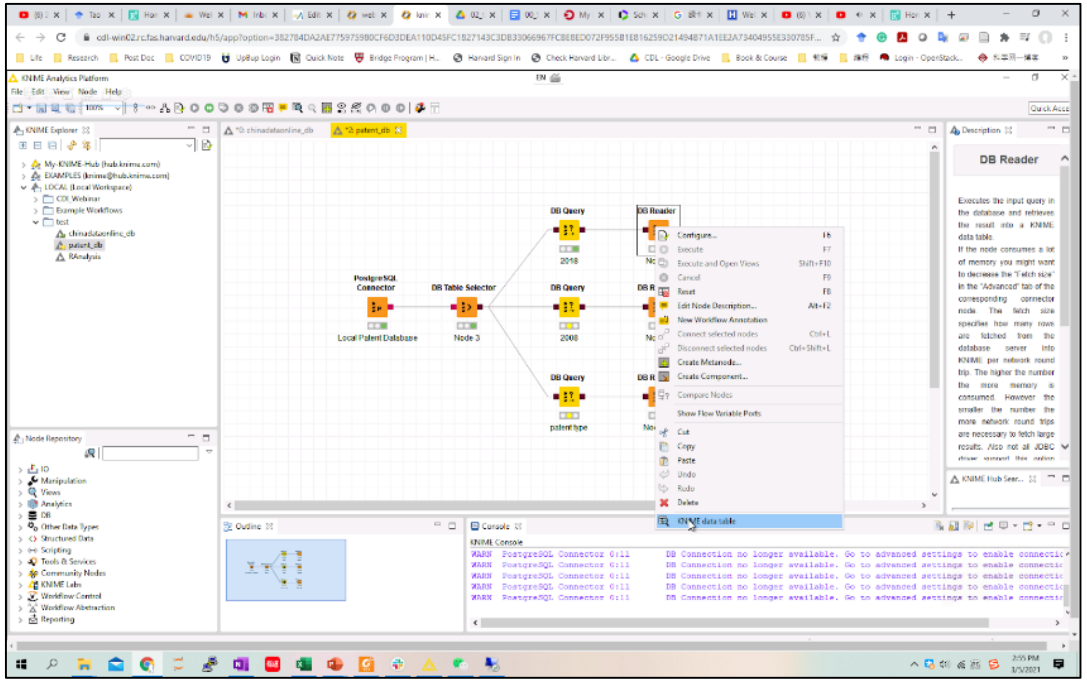




Data Access with Workflows

2. Access Large-scale Patent Data from Platform Database

1) Import Workflow from Public Folder **\\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.

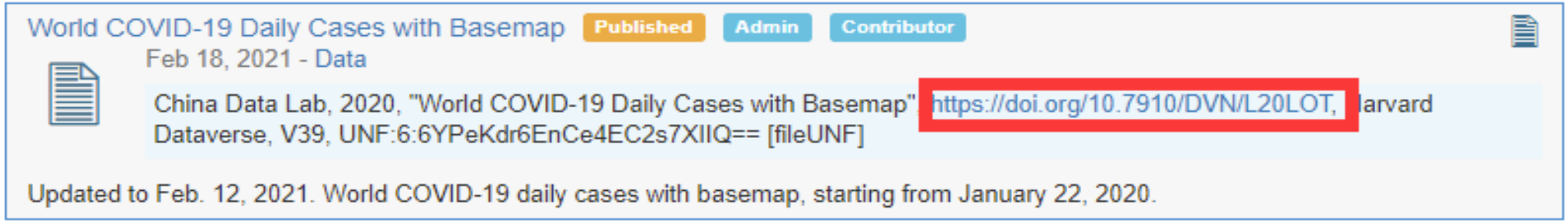


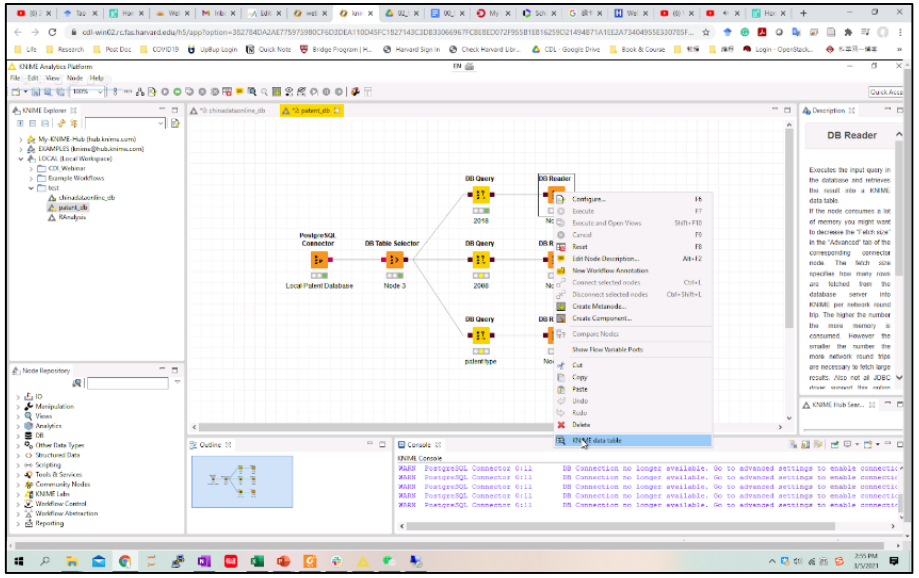
字段名	Field	字段名	Field	字段名	Field
标题（O）	Tio	公开（公告）日	Publication Date	引证类别(backward)	Citation Origin Code
摘要（O）	Abo	转让执行日	Assignment ExecutionDate	X引证文献	ct-X
专利价值度	Vlstar	专利有效性	status	家族引证次数(backward)	Family Citation Number of Times
申请人(原始)	AP-OR	当前法律状态	CN status lite	家族被引证次数（forward）	Family Citation-Forward 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

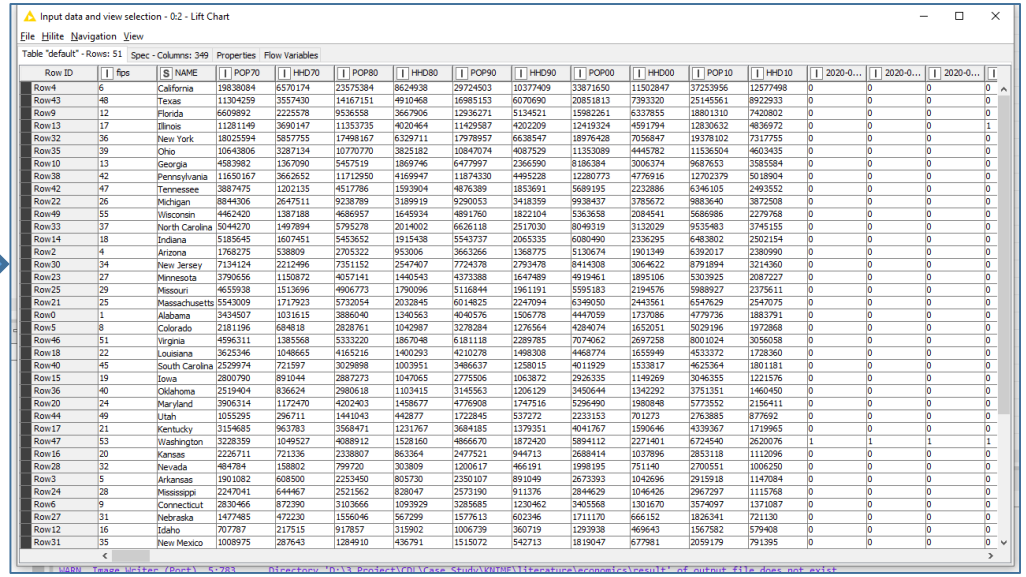
Data Access with Workflows

3. Access Public COVID-19 Data from Harvard Dataverse

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

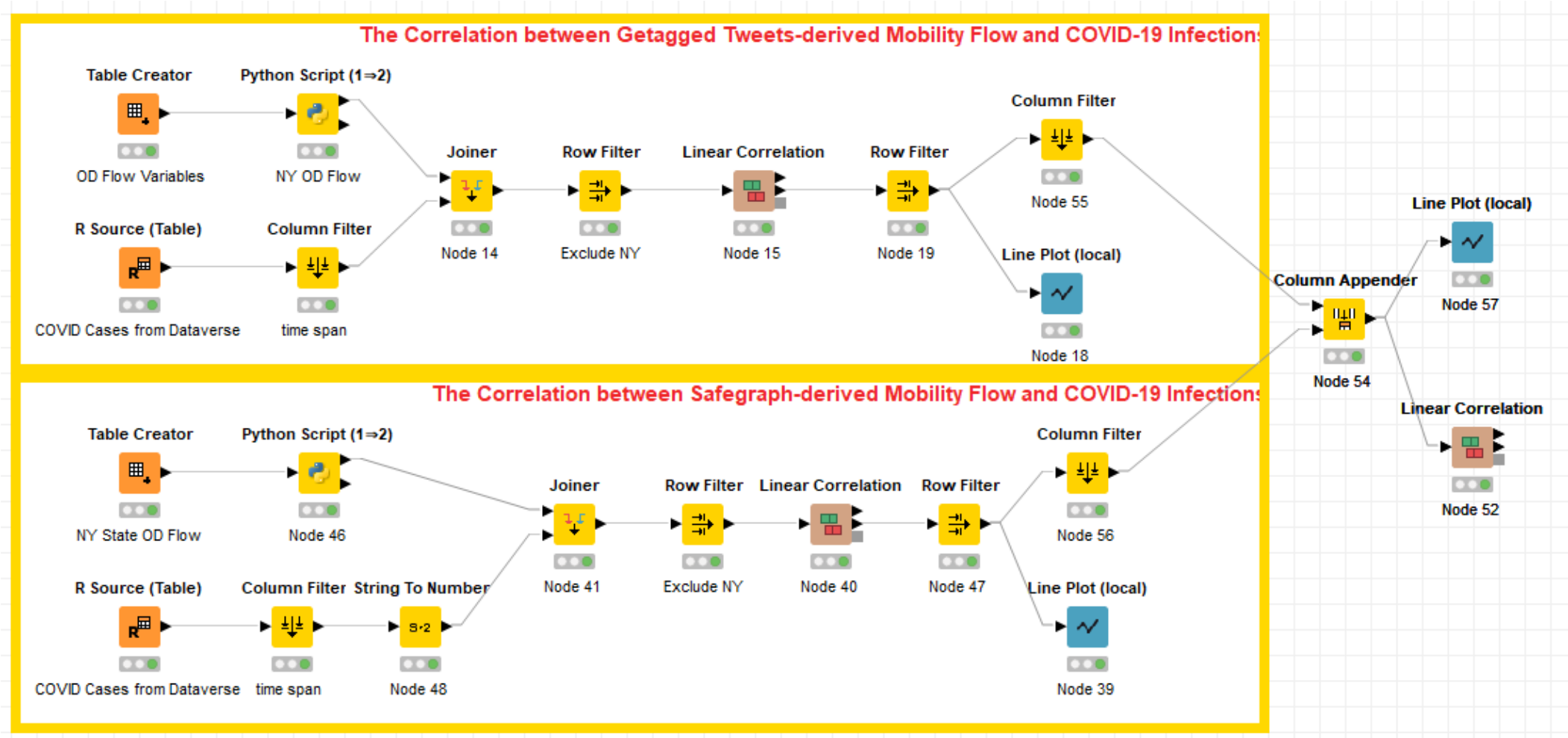






Data Access with Workflows

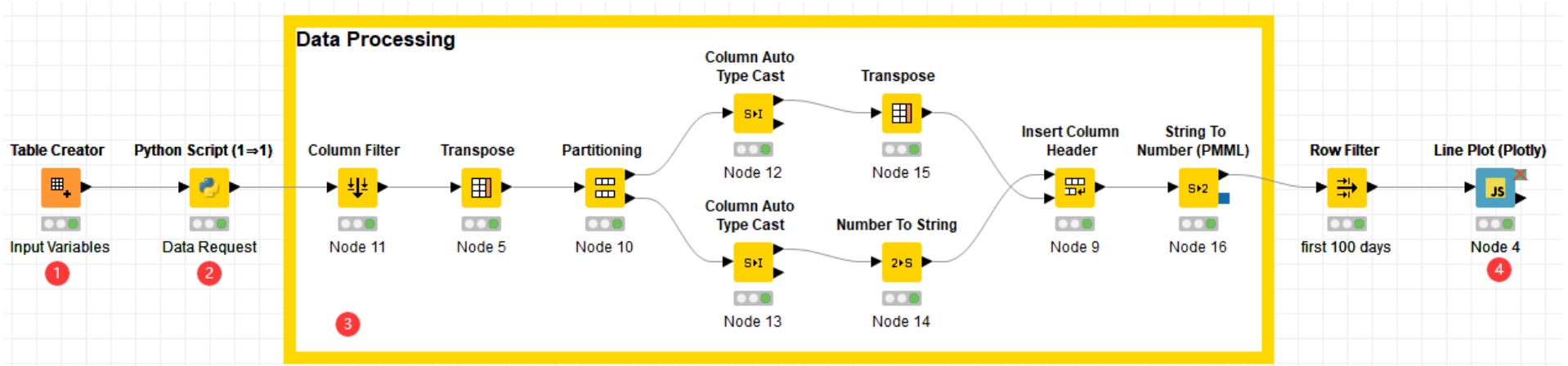
4. Access External Human Mobility Index Data via Rest API

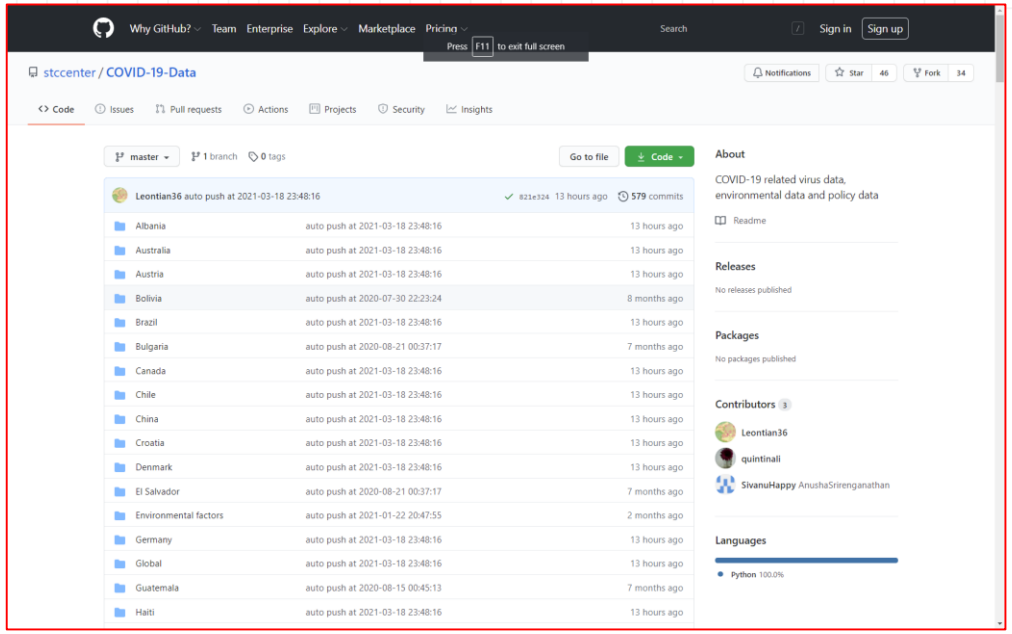


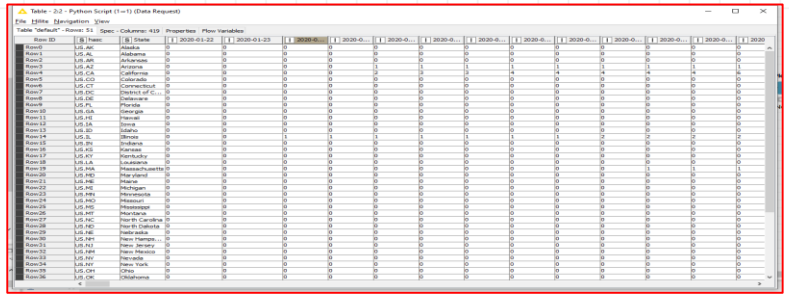
Data Access with Workflows

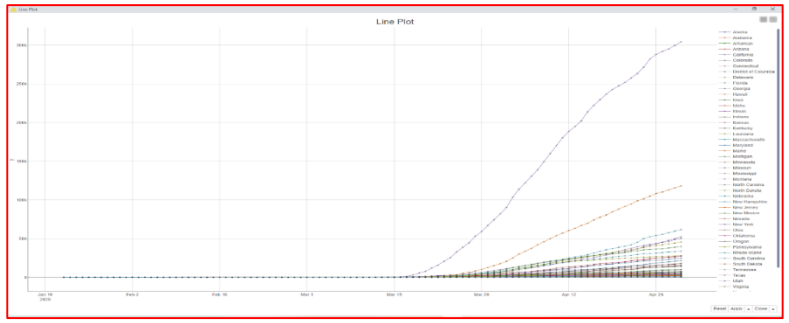
5. Access External Data via GitHub API

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









Case Study 1: Air Pollution 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

- Harvard Dataverse (https://doi.org/10.7910/DVN/XETLSS)

Data

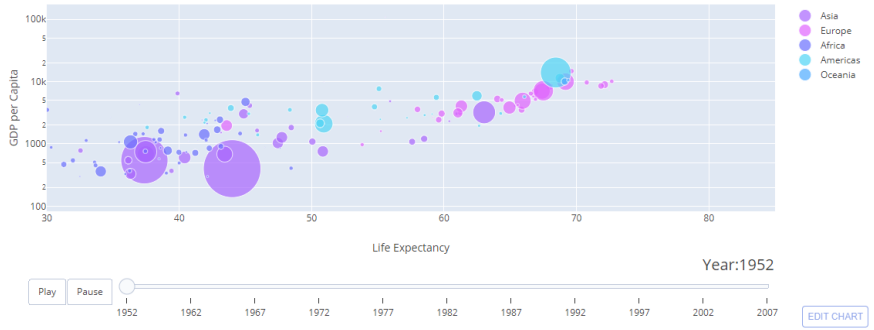
- Daily PM2.5 emissions by China cities (Jan-Feb, 2020)
- Province boundary basemap of Mainland China

Method and Flowchart

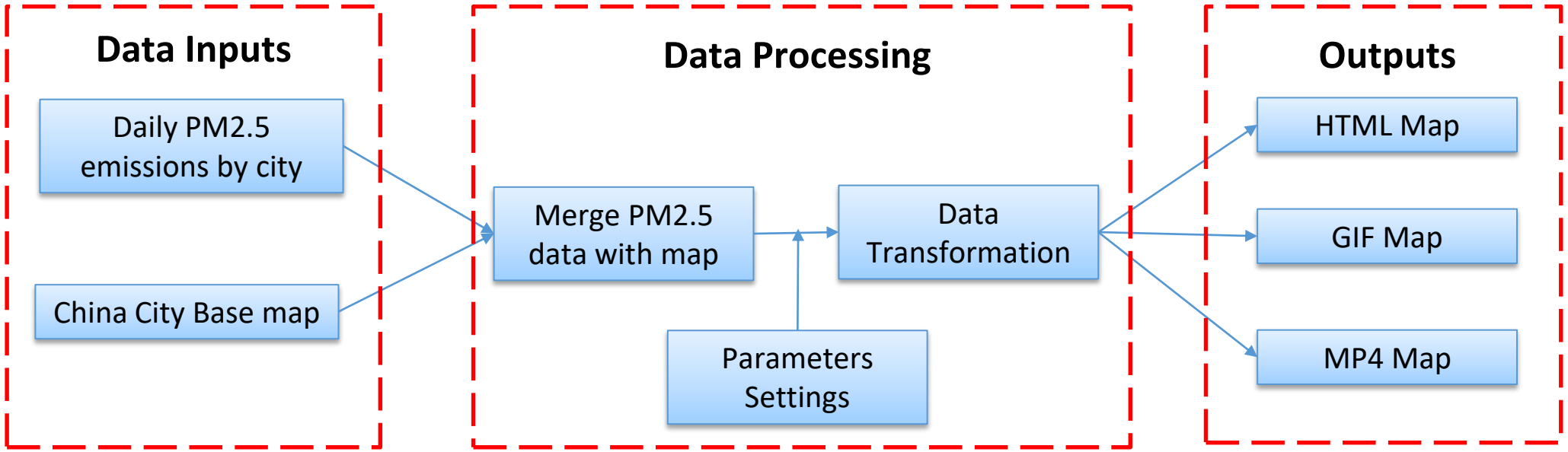
Methodology

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

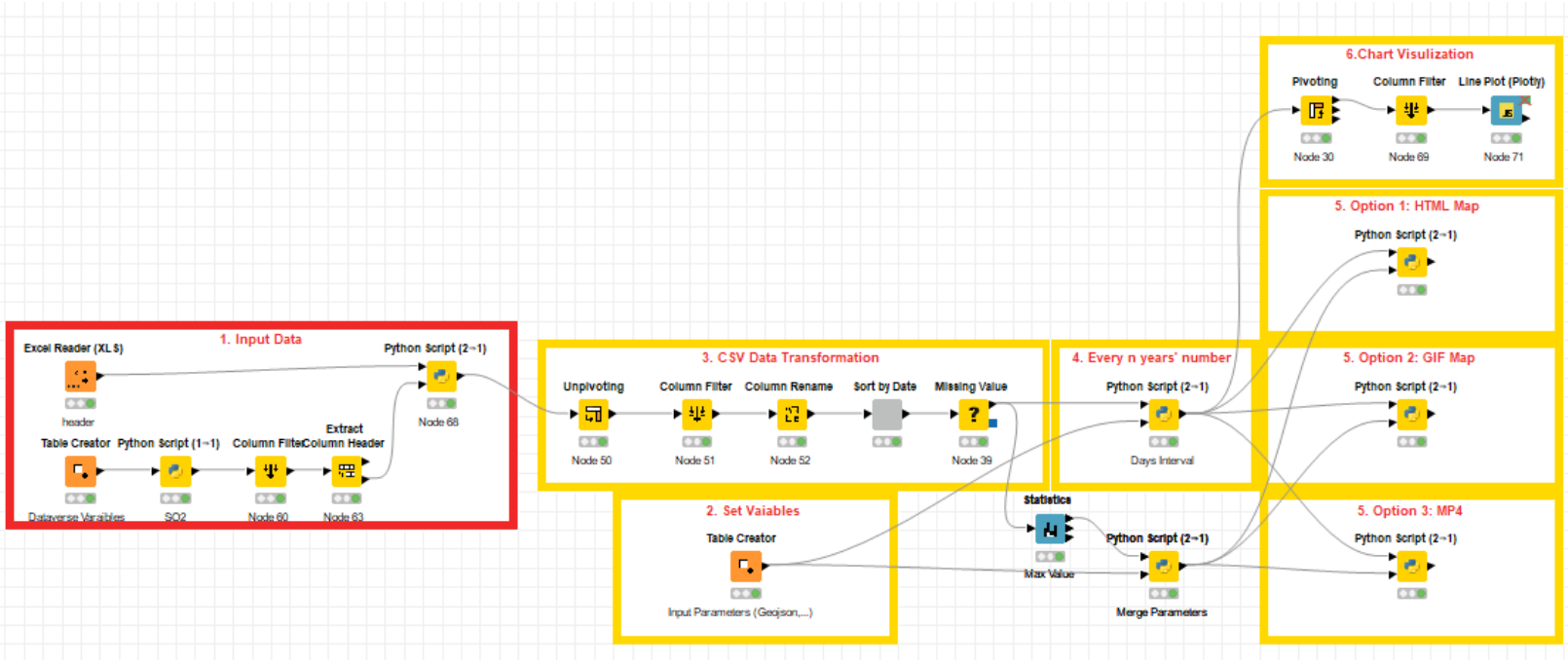
Mapbox Choropleth | Lines on Mapbox | Filled Area on Maps | Bubble Maps

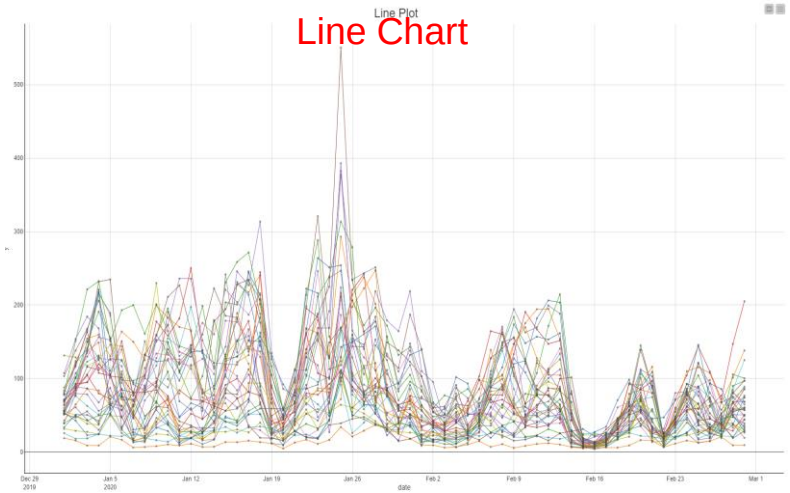


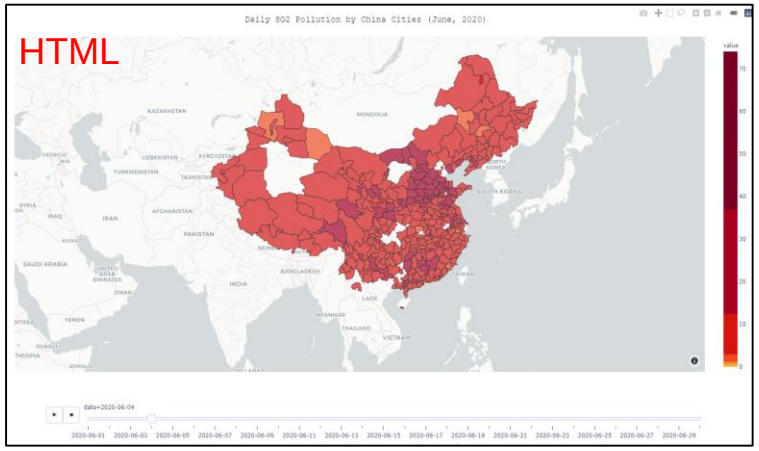
Flowchart

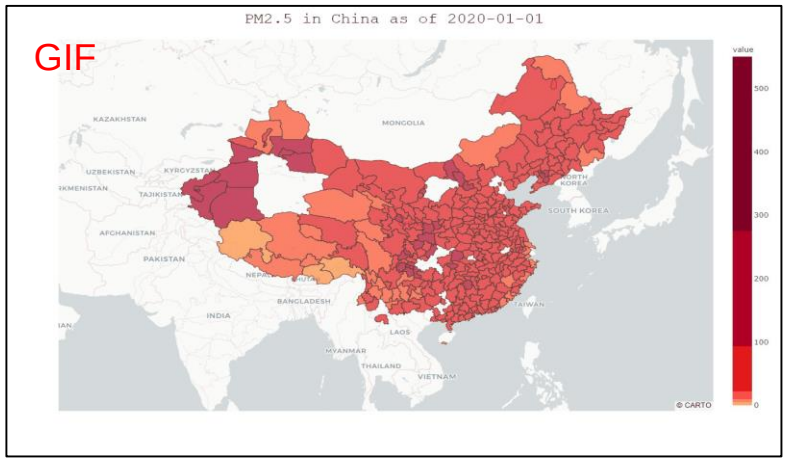


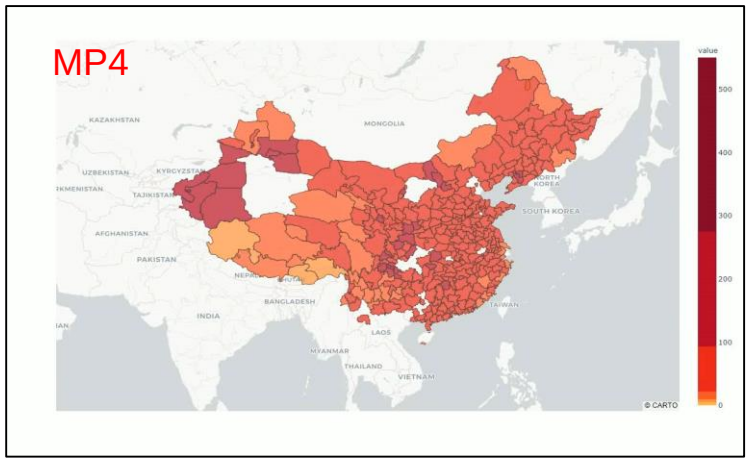
Workflow Implementation by KNIME











Steps for Running the Workflow

Step 1: **Download data** from \01_dynamic_maps\01_data\
Step 2: **Download workflow** from \01_dynamic_maps\02_workflow\environment_dynamic_map.knwf
Step 3: **Open KNIME** from local PC or China Data Lab Cloud Platform (in building)
Step 4: **Import KNIME workflow** file (environment_dynamic_map.knwf)
Step 5: **Launch the workflow** by double clicking the imported workflow
Step 6: **Configure "Input Data"**
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 Anaconda as the Python environment and set Python path in the KNIME preference
*** install python libraries following commands below

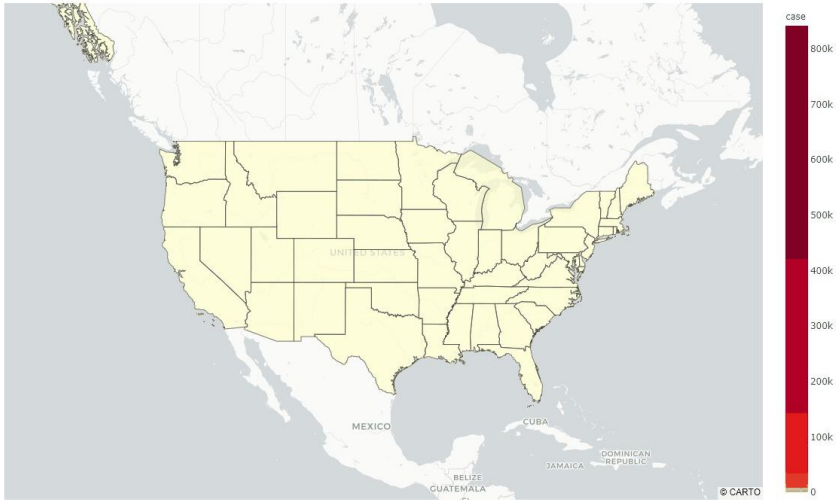
- conda install –c plotly plotly
- conda install –c plotly plotly-orca
- conda install –c plotly python-kaleido
- conda install -c conda-forge imageio
- conda install -c anaconda opencv

*** Please set Chrome as default browser
*** Please install suggested extensions when opening workflow

Summary

https://covid-19.stcenter.net/index.php/global-countries/

2020-01-22



Infection, Genetics and Evolution 88 (2021) 104701

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Short communication

COVID-19 spatiotemporal research with workflow-based data analysis

Srikar Chintala [a,*,1], Ritvik Dutta [b,1], Doron Tadmor [c,1]

[a] University Preparatory Academy, San Jose, CA 95125, USA
[b] Monta Vista High School, Cupertino, CA 95014, USA
[c] Imperial College Business School, Imperial College London, London SW7 2BU, UK

ARTICLE INFO

Keywords:
COVID-19
Coronavirus
Dynamic maps
KNIME
Spatial visualization
Visual analysis

ABSTRACT

Given the pertinence and acceleration of the spread of COVID-19, there is an increased need for the replicability of data models to verify the veracity of models and visualize important data. Most of these visualizations lack reproducibility, credibility, or accuracy, and are static, which makes it difficult to analyze the spread over time. Furthermore, most current visualizations depicting the spread of COVID-19 are at a global or country level, meaning there is a dearth of regional analysis within a country. Keeping these issues in mind, a replicable, efficient, and simple method to generate regional COVID-19 visualizations mapped with time was created by using the KNIME software, an open-source data analytics platform that can create user-friendly applications or workflows. For this analysis, Albania, Sweden, Ukraine, Denmark, Russia, India, and Australia were closely observed. Among the maps generated for the aforementioned countries, it was noticed that regions with a high population or high population density were often the epicenters within their respective country. The regions caused the virus to spread to their neighboring regions: kickstarting the "domino effect", leading to the infection of another region until the country is overwhelmed with cases—what we call a proximity trend. These dynamic maps are crucial to fighting the pandemic because they can provide insight as to how COVID-19 spreads by providing researchers or officials with an accurate and insightful tool to aid their analysis. By being able to visualize the spread, health and government officials can dive deeper to identify the sources of transmission and attempt to stop or reverse them accordingly.

COVID-19 Rapid Response Research with Workflow-based Data Analysis

Doron Tadmor

Imperial College London

doron.tadmor20@imperial.ac.uk

[1] Chintala, S., Dutta, R., & Tadmor, D. (2021). COVID-19 spatiotemporal research with workflow-based data analysis. Infection, Genetics and Evolution, 88, 104701.
[2] Tadmor, D. (2020). COVID-19 Rapid Response Research with Workflow-based Data Analysis. Available at SSRN 3643621.

Case Study 2: Air pollution and COVID-19 mortality

Wu, X., Nethery, R. C., …, & Dominici, F. (2020). Air pollution and COVID-19 mortality in the United States: Strengths and limitations of an ecological regression analysis. Science Advances, 6(45)

Objective: Assessing whether long-term exposure to air pollution increases the severity of COVID-19 health outcomes, including death, is an important public health objective.

Data Sources: Harvard Dataverse, NASA, Census Bureau

Data: COVID-19 Cases, PM 2.5, Population density, income, percent of owner-occupied housing, adult population, age distribution, black residents, et al.

SCIENCE ADVANCES | RESEARCH ARTICLE

CORONAVIRUS

Air pollution and COVID-19 mortality in the United States: Strengths and limitations of an ecological regression analysis

X. Wu[1]†, R. C. Nethery[1]*†, M. B. Sabath[1], D. Braun[1,2], F. Dominici[1]*

Copyright © 2020 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

Assessing whether long-term exposure to air pollution increases the severity of COVID-19 health outcomes, including death, is an important public health objective. Limitations in COVID-19 data availability and quality remain obstacles to conducting conclusive studies on this topic. At present, publicly available COVID-19 outcome data for representative populations are available only as area-level counts. Therefore, studies of long-term exposure to air pollution and COVID-19 outcomes using these data must use an ecological regression analysis, which precludes controlling for individual-level COVID-19 risk factors. We describe these challenges in the context of one of the first preliminary investigations of this question in the United States, where we found that higher historical $PM_{2.5}$ exposures are positively associated with higher county-level COVID-19 mortality rates after accounting for many area-level confounders. Motivated by this study, we lay the groundwork for future research on this important topic, describe the challenges, and outline promising directions and opportunities.

INTRODUCTION

The suddenness and global scope of the coronavirus disease 2019 (COVID-19) pandemic have raised urgent questions that require coordinated investigation to slow the disease's devastation. A critically important public health objective is to identify key modifiable environmental factors that may contribute to the severity of health outcomes [e.g., intensive care unit (ICU) hospitalization and death] among individuals with COVID-19. Numerous scientific studies reviewed by the U.S. Environmental Protection Agency (EPA) have linked fine particles ($PM_{2.5}$; particles with diameter, ≤ 2.5 μm) to a variety of adverse health events (1) including death (2). It has been hypothesized that because long-term exposure to $PM_{2.5}$ adversely affects the respiratory and cardiovascular systems and increases mortality risk (3–5), it may also exacerbate the severity of COVID-19 symptoms and worsen the prognosis of this disease (6).

Epidemiological studies to estimate the association between long-term exposure to air pollution and COVID-19 hospitalization and death is a rapidly expanding area of research that is attracting attention around the world. Two studies have been published using data from European countries (7, 8), and many more are available as preprints. However, because of the unprecedented nature of the pandemic, researchers face serious challenges when conducting these studies. One key challenge is that, to our knowledge, individual-level data on COVID-19 health outcomes for large, representative populations are not publicly available or accessible to the scientific community. Therefore, studies of long-term exposure to air pollution and COVID-19 outcomes must use ecological regression analyses of air pollution and COVID-19 health outcomes and describe additional challenges related to evolving data quality, statistical modeling, and control of measured and unmeasured confounding, paving the way for future research on this topic. We discuss these challenges and illustrate them in the context of a specific study, in which we investigated the impact of long-term $PM_{2.5}$ exposure on COVID-19 mortality rates in 3089 counties in the United States, covering 98% of the population.

Illustration of an ecological regression analysis of historical exposure to $PM_{2.5}$ and COVID-19 mortality rate

We begin by describing how to conduct an ecological regression analysis in this setting. COVID-19 death counts (a total of 116,747 deaths) were obtained from the Johns Hopkins University Coronavirus Resource Center and were cumulative up to 18 June 2020. We used data from 3089 counties, of which 1244 (40.3%) had reported zero COVID-19 deaths at the time of our analysis. Daily $PM_{2.5}$ concentrations were estimated across the United States on a 0.01° × 0.01° grid for the period 2000–2016 using well-validated atmospheric chemistry and machine learning models (9). We used zonal statistics to aggregate $PM_{2.5}$ concentration estimates to the county level and then averaged across the period 2000–2016 to perform health outcome analyses. Figure 1 illustrates the spatial variation in 2000–2016 average (hereafter referred to as "long-term average") $PM_{2.5}$ concentrations and COVID-19 mortality rates (per 1 million population) by county.

Downloaded from http://advances.sciencemag.org/ on February 12, 2021

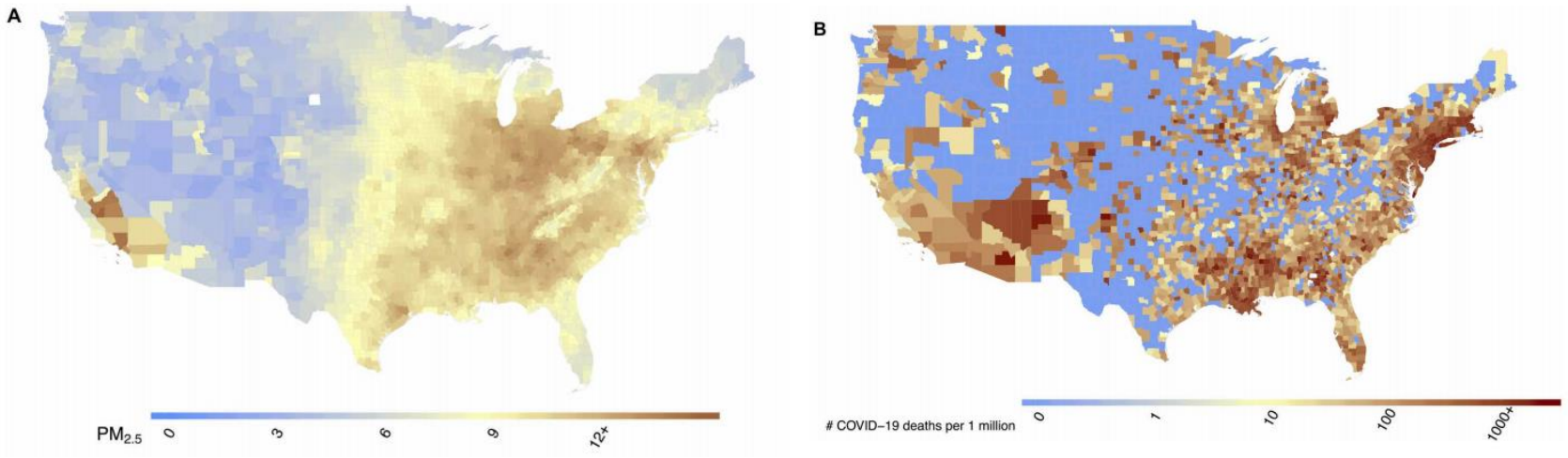


Table S2: Characteristics of the study cohort up to and including June 18, 2020, mean (standard deviation).

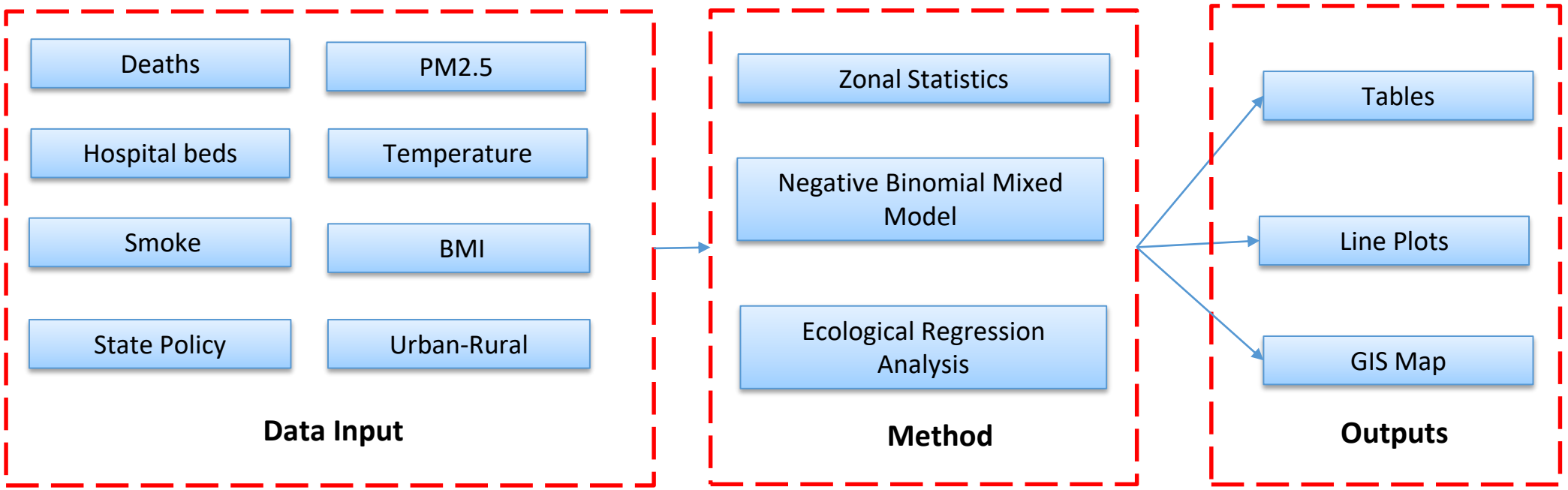
	Total (3,087 counties)	$PM_{2.5}$< 8 $\mu g/m^3$ (1,217 counties)	$PM_{2.5}$≥ 8 $\mu g/m^3$ (1,870 counties)
COVID-19 death rate (per 100,000)	15.5 (31.2)	7.0 (19.7)	21.0 (35.7)
Average $PM_{2.5}$($\mu g/m^3$)	8.4 (2.5)	5.7 (1.4)	10.1 (1.2)
Rate of hospital beds (per 100,000)	300.8 (428.4)	338.8 (548.2)	276.1 (325.3)
Days since first case	76.9 (21.1)	68.2 (27.8)	82.6 (12.2)
Days since stay-at-home order	59.4 (38.1)	52.4 (41.6)	63.9 (34.9)
% Smokers	17.4 (3.5)	15.8 (3.1)	18.5 (3.4)
% Obese	32.9 (5.4)	31.2 (5.1)	34.0 (5.3)
% In poverty	10.5 (5.9)	9.6 (5.8)	11.0 (5.9)
% Less than high school education	21.3 (10.7)	16.6 (8.8)	24.4 (10.6)
% Owner-occupied housing	75.0 (8.3)	76.4 (7.6)	74.1 (8.6)
% Hispanic	7.5 (12.3)	9.6 (13.6)	6.1 (11.1)
% Black	8.0 (14.1)	1.0 (1.9)	12.6 (16.5)
% ≥65 years of age	16.0 (4.1)	17.4 (4.6)	15.0 (3.4)
% 45-64 years of age	26.4 (3.0)	26.9 (3.8)	26.1 (2.4)
% 15-44 years of age	37.6 (6.5)	35.2 (8.2)	39.2 (4.5)
Population density (person/km^2)	149.5 (699.2)	46.1 (146.9)	216.9 (884.4)
Median household income ($1,000)	49.3 (13.4)	50.6 (11.1)	48.4 (14.6)
Median house value ($1,000)	135.7 (89.9)	140.4 (87.8)	132.7 (91.1)
Average summer temperature (°F)	86.0 (5.7)	83.7 (6.7)	87.4 (4.4)
Average winter temperature (°F)	45.1 (11.9)	39.4 (11.4)	48.7 (10.7)
Average summer relative humidity (%)	89 (9.6)	83.2 (11.5)	92.8 (5.5)
Average winter relative humidity (%)	87.5 (4.8)	88.0 (5.6)	87.2 (4.1)

Methodology

Methodology:

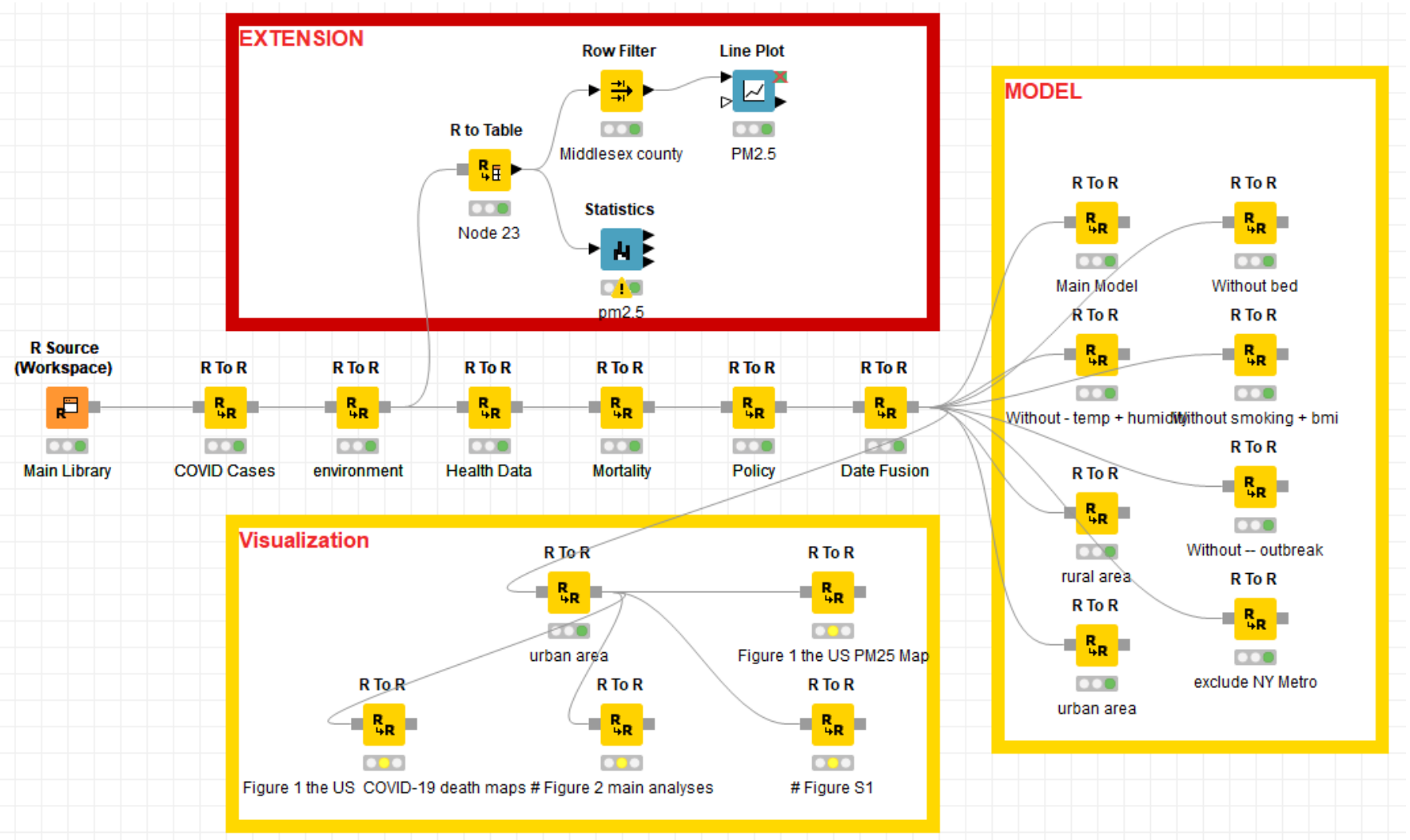
- Employ ***Ecological Regression Analysis*** to investigate relation between air pollution and COVID-19 health outcomes
- Fit a **negative binomial mixed model** using COVID-19 mortality rates as the outcome
- ***Generalized linear mixed-effects model (GLMM)***

Flowchart

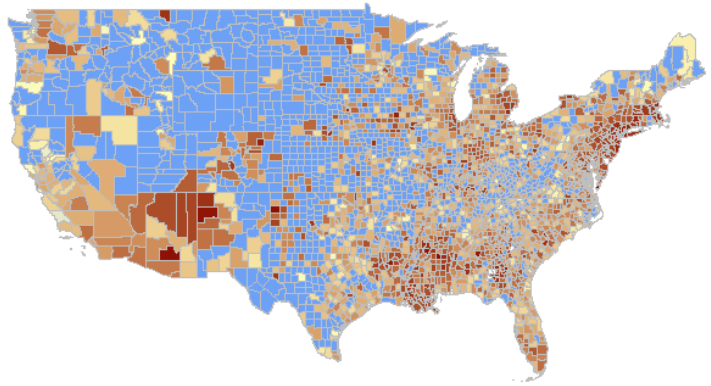


Workflow Implementation

Workflow Implementation




```
Fixed effects:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -10.217335   0.250012 -40.867  < 2e-16 ***
mean_pm25                           0.106715   0.026763   3.987 6.68e-05 ***
factor(q_popdensity)2              -0.099611   0.122614  -0.812 0.416565
factor(q_popdensity)3              -0.091375   0.125733  -0.727 0.467384
factor(q_popdensity)4              -0.303044   0.131418  -2.306 0.021113 *
factor(q_popdensity)5              -0.081058   0.149414  -0.543 0.587470
scale(poverty)                      0.039139   0.037745   1.037 0.299760
scale(log(medianhousevalue))        0.120289   0.068393   1.759 0.078611 .
scale(log(medhouseholdincome))      0.174529   0.066201   2.636 0.008381 **
scale(pct_owner_occ)                0.111521   0.038323   2.910 0.003614 **
scale(education)                    0.184048   0.045915   4.008 6.11e-05 ***
scale(pct_blk)                      0.399574   0.039466  10.125  < 2e-16 ***
scale(hispanic)                     0.056267   0.046403   1.213 0.225291
scale(older_pecent)                 0.044078   0.057518   0.766 0.443476
scale(prime_pecent)                -0.255398   0.074339  -3.436 0.000591 ***
scale(mid_pecent)                  -0.276175   0.055160  -5.007 5.53e-07 ***
scale(date_since_social)            0.150374   0.131378   1.145 0.252380
scale(date_since)                   0.873910   0.078822  11.087  < 2e-16 ***
scale(beds/population)             -0.001939   0.036004  -0.054 0.957052
scale(obese)                       -0.034140   0.035629  -0.958 0.337957
scale(smoke)                        0.121470   0.063175   1.923 0.054510 .
scale(mean_summer_temp)             0.101337   0.081507   1.243 0.213760
scale(mean_winter_temp)            -0.148755   0.111845  -1.330 0.183515
scale(mean_summer_rm)              -0.071627   0.079634  -0.899 0.368413
scale(mean_winter_rm)              -0.035043   0.052008  -0.674 0.500434
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Summary

Findings

- The higher historical PM2.5 exposures are positively associated with higher county-level COVID-19 mortality rates after accounting for many area-level confounders.
- It also found that population density, days since the first COVID-19 case was reported, median household income, percent of owner-occupied housing, percent of the adult population with less than high school education, age distribution, and percent of Black residents are important predictors of the COVID-19 mortality rate in the model.

Discussion

- Integrate more datasets from multiple data sources
- Develop more case studies with workflow tools, including KNIME, R AnalyticalFlow, Jupyter Notebook (Python/R)
- Collaborate with global scholars to replicate publications
- Build a case study repository for easy access and query
- Organize training webinars for students, researchers, and educators to promote cross-disciplinary reproducible, replicable, and generalizable research

Acknowledgement

Research Data Group:

- Cynthia Baiqing Zhang, Central Washington University
- Fuyu Frank Xu, Maine University
- Jianxin Liu, UC San Diego
- Jiayi Liu, Shanghai University of Finance and Economics
- Manni Zhang, Purdue University
- Pengcheng Wu, Jiujiang College
- Shane Fan, UC Berkeley
- Sijia Yuen, ACMR
- Tao Hu, Harvard University
- Xiaokang Fu, Wuhan University
- Yaohong Ma, Wuhan University
- Ru Wang, Wuhan University
- Yuting Zhou, Northeast Normal University

Harvard CVT Student Group:

- Akhil Kumar
- Jacy Zeng
- Ritvik Dutta
- Shayaan Kaleem
- Srikar Chintala
- Zachary Huang
- Doron Tadmor
- David Liu

Thank you !

Questions/Comments?

Contact Email: spatialdatalab@lists.fas.harvard.edu