Introduction to
the Spatial Data Lab (SDL) Project

Wendy Guan

Nov. 7th, 2023

Geospatial & Temporal Data Are Everywhere

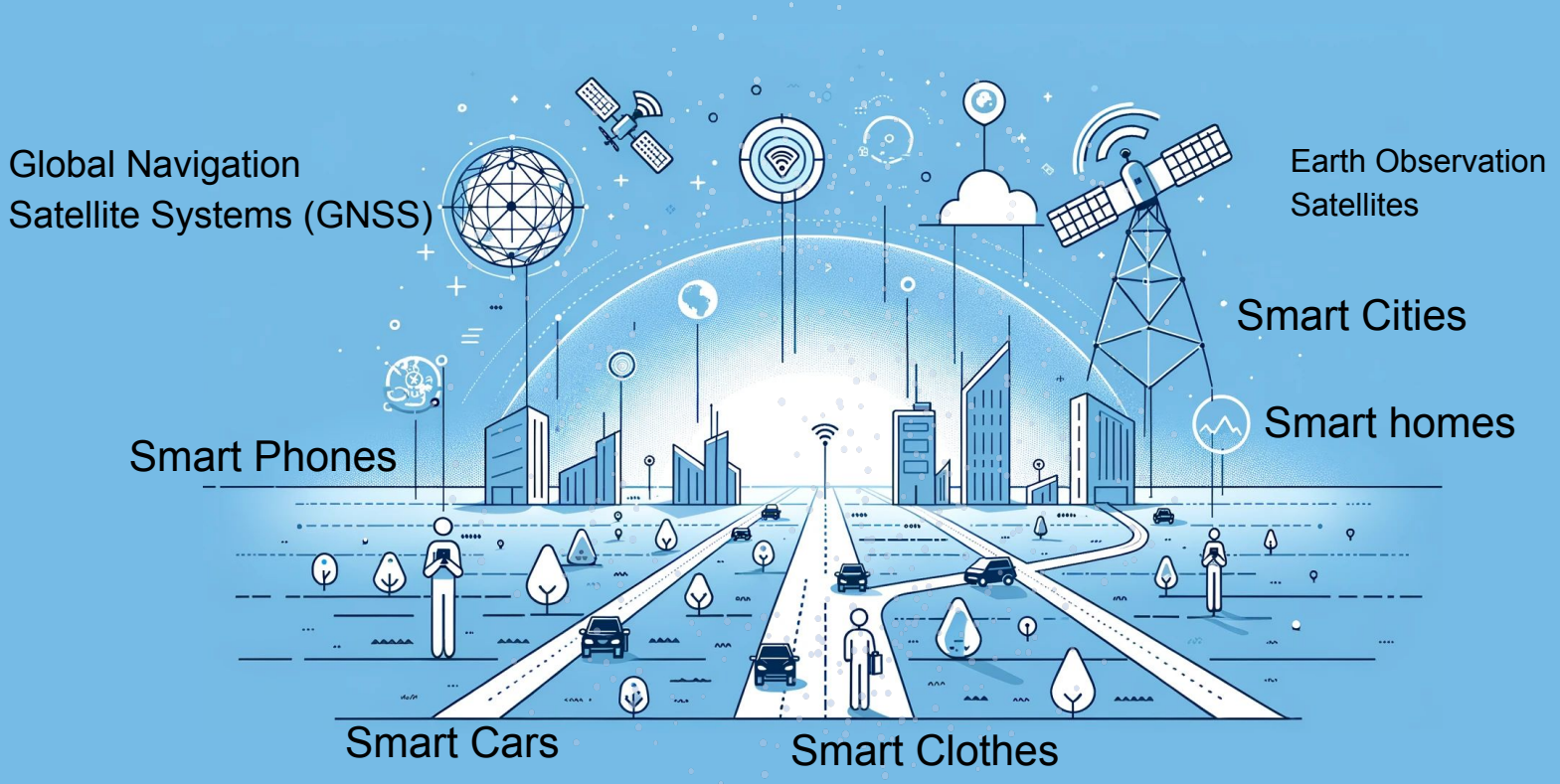


Image: Generated by DALL·E3

Challenges in Spatiotemporal Research

- Data Sharing
  - Licensed data
  - Restricted data
  - Sensitive data
  - Large size data
  - Different Resources
- Tool Sharing
  - Licensed and free tools
  - Integrated environment for tools and data
  - Management (maintenance and updates)
- Research Sharing
  - Research (reproducible, replicable, generalizable)
  - Training (students with different interests & skills)
  - Applications (efficient, effective, and expandable)

The Dataverse® Project

About ***74%*** re-executions failed

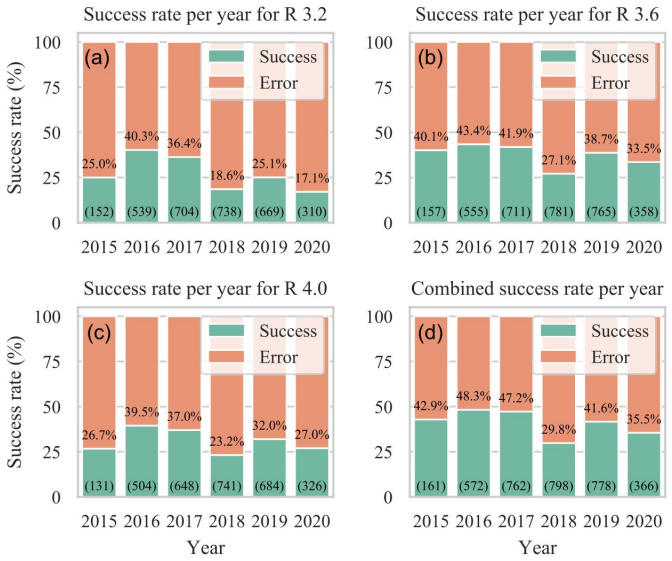


Fig. 11 Re-execution success rates per year per R software version.

Image source:
https://www.nature.com/articles/s41597-022-01143-6.pdf?origin=ppub

About the Spatial Data Lab Project

GOAL: Build an innovative platform for effective and collaborative spatiotemporal research

Project in — NSF IUCRC Spatiotemporal Innovation Center(STC)

Sponsored by — **Future Data Lab**

Open for Innovation KNIME

Conducted at — Center for Geographic Analysis, Harvard University

Supported by — George Mason University

Objectives

- Promote integrated spatiotemporal data services
- Develop tools for spatiotemporal data analysis
- Build workflow-based spatiotemporal case studies
- Deliver training programs for spatiotemporal data science
- Cultivate collaborative spatiotemporal research communities
- Grow a repository with spatiotemporal data products, tool packages, research case studies, and training materials

KNIME as the Technical Foundation for Spatiotemporal Data Science

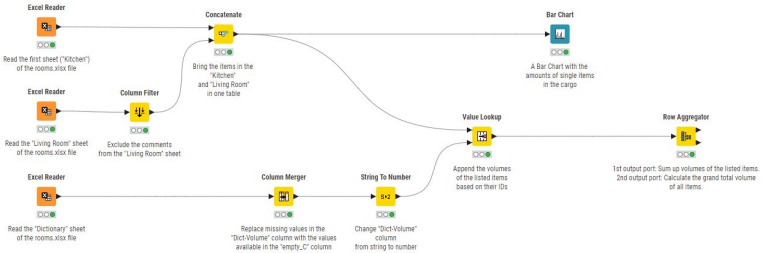
OPEN
Open source

End To End
IO, Manipulation, Visualization, Mining, Reporting

Visual
Codeless 2000+ Visual and workflow-based Programming

Coding
Python
R
Java

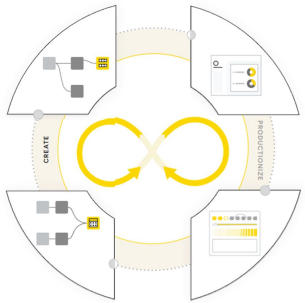
Cloud
WebPortal
Integrated software



Open for Innovation
KNIME

Model & Visualize
Choose from a complete range of analytic techniques.

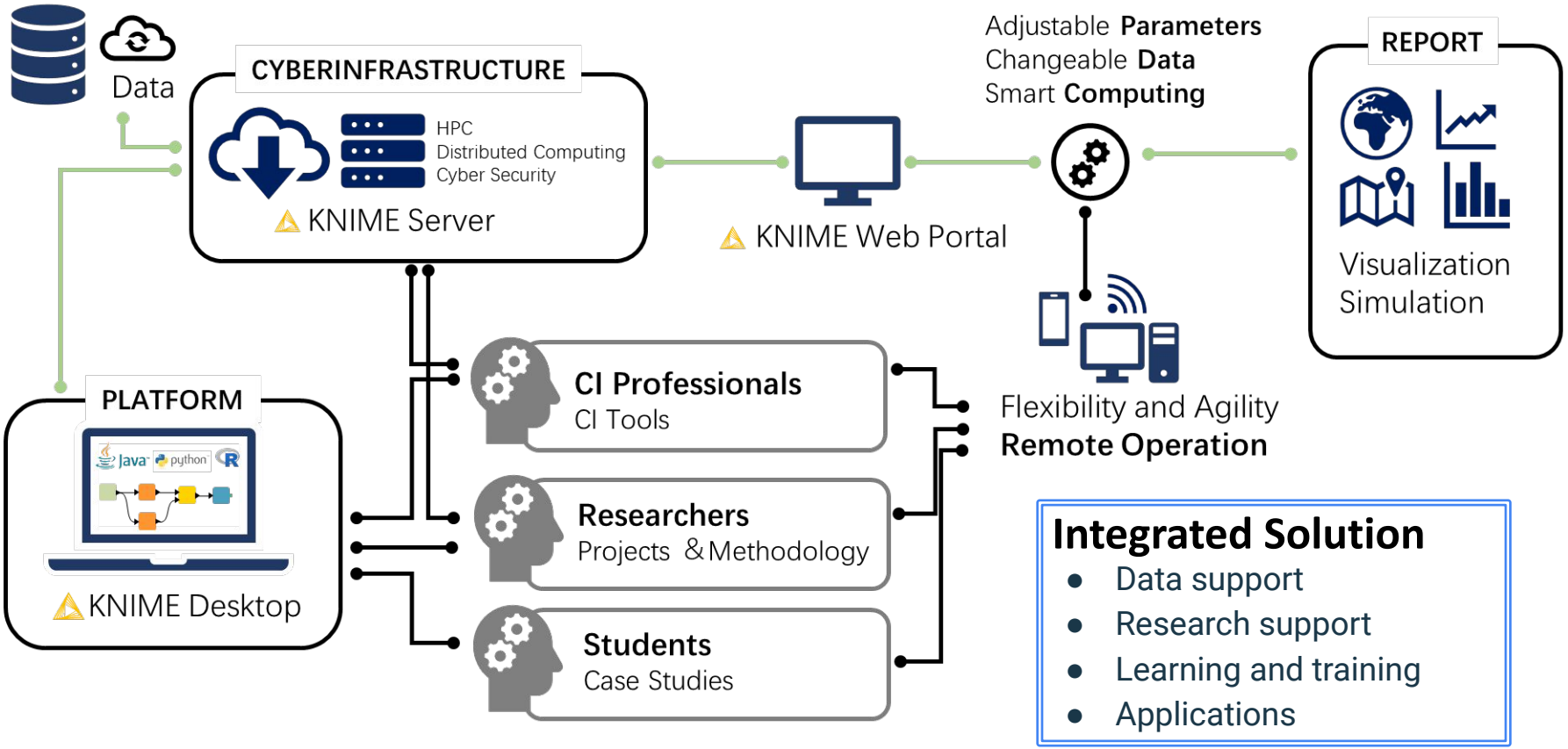
Blend & Transform
Access any data type from any source with 300+ connectors.



Deploy & Monitor
Securely deploy data science solutions with a standard process.

Consume & Interact
Go enterprise-scale, with cloud-native architecture.

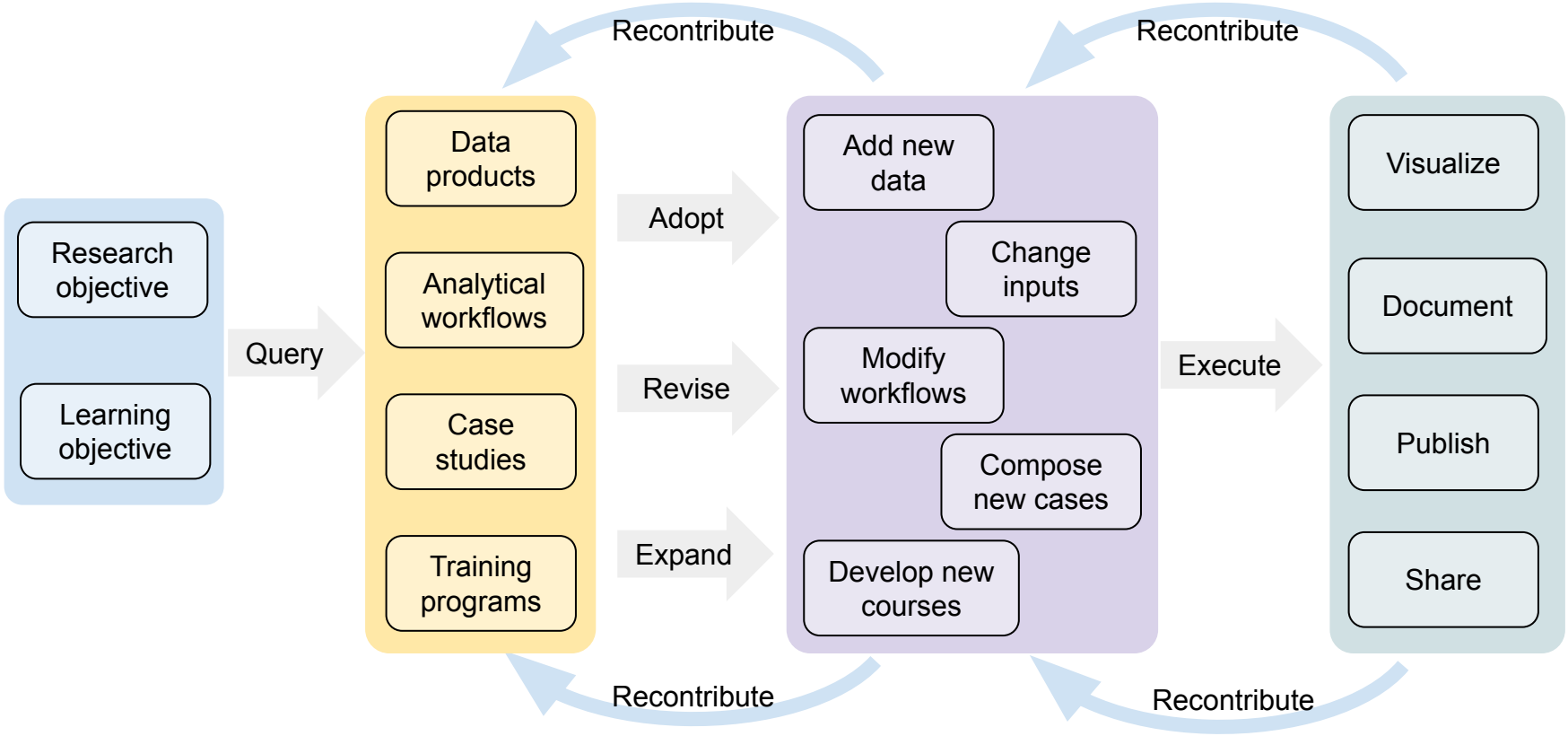
The KNIME-based SDL Architecture Overview



Integrated Solution

- Data support
- Research support
- Learning and training
- Applications

The Re-engineered Research Lifecycle in SDL



SDL Development: Geospatial Analytics Extension For KNIME

Latest Version 1.2.0

Spatial IO

Spatial Calculation

Spatial Manipulation

Spatial Transformation

Spatial Conversion

Spatial Visualization

Exploratory Spatial Data Analysis

Spatial Modelling

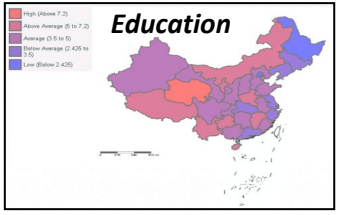
Location Analysis

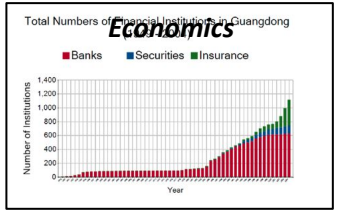
Spatial Network

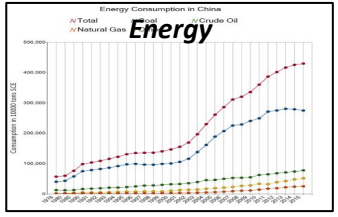
Open Dataset

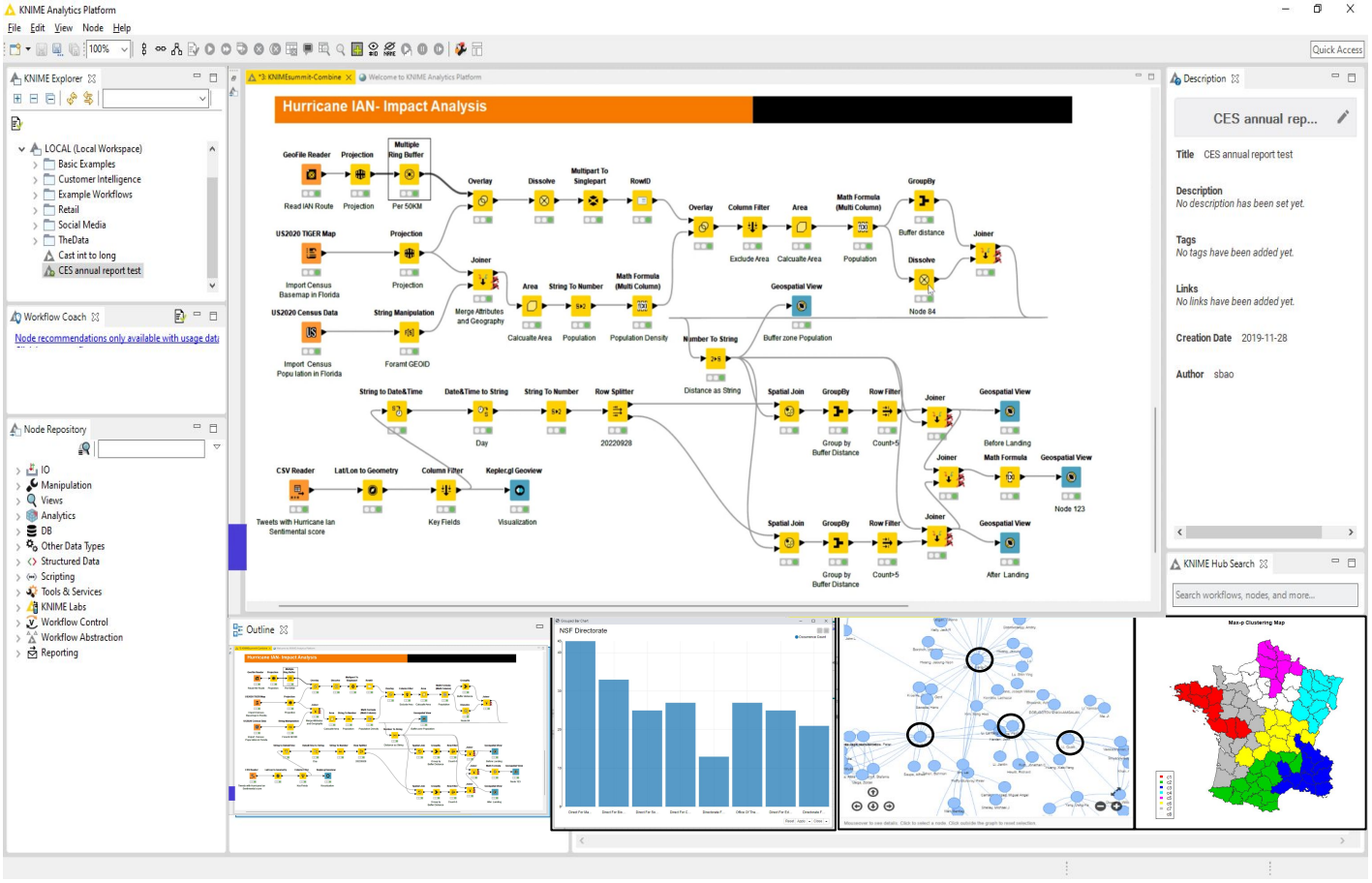
Spatial Clustering

Replicable, Reproducible and Expandable Data Analysis









SDL Development: Domain-specific Data Collections

COVID-19 Data Collections:

1. US COVID-19 Daily Cases with Basemap
2. US Metropolitan COVID-19 Daily Cases with Basemap
3. China COVID-19 Daily Cases with Basemap
4. import Cases from forigen countries
5. World COVID-19 Daily Cases with Basemap
6. Community Cases from Tecent
7. Trajactory Data from Tecent
8. China COVID-19 events timeline
9. Global COVID-19 Events Timeline
10. US Vaccine distribution
11. Global Vaccine distribution

Mobility Data Collections:

1. Descartes Lab Mobility Report
2. Mobility Metrics and Social Distancing Index
3. Google Community Mobility Reports
4. Apple Mobility Reports
5. Foursquare Community Mobility Data
6. Safegraph Foot Traffic Data
7. Human OD (Origion-Destination) Flow
9. Baidu Mobility Data

Transportation Data Collections:

1. Flight
2. Train

POI Data Collections:

1. Hospital POI in China
2. Hospital POI in US

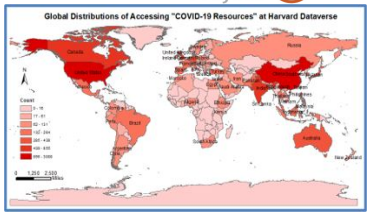
Social Media Data:

1. GDELT
2. Twitter(eng)

Other Data Collections:

1. City-level Air Quailty Data with China Map
2. City-level Meterlogical Data with China Map
3. Daily Meteorological Data in China
4. City-level CO2 data
5. City-level PM2.5 data
6. China Patent Data
7. US Patent Data
8. Innovation Index, 1990_2019
9. Province Culture, 2004-2018

The Dataverse® Project



Case Studies for Workflow-Based Data Analysis

Developed more than 70 workflow-based case studies for spatiotemporal research

Workflows for spatiotemporal studies

- Test the impacts of religious diversity on regional development in China
- Spatial panel data analysis in regional development
- Spatiotemporal Pattern Analysis of Higher Education
- Changes in human induced turbidity in Poyang Lake based on remote sensing data
- Air quality analysis with social media data
- Space-time analysis of high education in China
- Spatial factor analysis of road network based on the traffic accidents
- Industrial Co-agglomeration Analysis based on Machine Learning Algorithm

Workflows for COVID-19 Studies

- Dynamic Map Visualization on Time-series COVID-19 Data
- Correlation Analysis between COVID-19 Cases and Human Motilities
- Text Mining on COVID-19 related Social Media Data
- Literature Review on COVID-19 Publications
- Environment Data Processing
- Time-varying SIR Model, provided by University of Michigan
- London School of Hygiene and Tropical Medicine
- DELPHI Epidemiological Case Predictions
- An SEIR Infectious Disease Model with Testing and Conditional Quarantine

Workflows for Data Access

- Database Access and Data Analysis on China Data Online Data
- Database Access and Data Analysis on Patent Data
- Dataverse Data Access and Analysis
- Human Mobility Data Access and Analysis via HTTP
- Github Data Access and Analysis

Workflows for Monthly Training Webinars

- Statistical Data Analysis with Workflows
- Analysis of Population Census Data & Demographic Change
- Analysis of Economic Census Data & Industrial Change
- The Integration of Data and Maps for Spatial Analysis
- Spatial Analysis of Patent Data
- Spatial Analysis of Health with Statistics, Census and GIS Data
- Spatial Analysis of Environment with Statistics, Census and GIS Data
- Spatiotemporal Analysis of Urban Development
- Spatiotemporal Analysis of Rural Development
- Human Mobility in Space and Time

Public Webinars and Training Workshops

About *90* webinars with over *12,000* participants around the world.

- Webinar series on Covid-19 studies (28)
  - COVID-19 data analysis (12)
  - COVID-19 modeling (8)
  - COVID-19 Impact Analysis (8)
- Webinar series on data, tools and literature studies (26)
- Webinar series on “Spatiotemporal Study of Urban Dynamics”, jointly with CPGIS (7)
- Monthly training webinars on research data, jointly with FDL & CDI (12)
- Monthly training webinars on workflow based data analysis, jointly with RMDS Lab (12)
- Monthly webinars on spatiotemporal Innovation
- Spatiotemporal Innovation Workshop (4)

See the latest list of all SDL events

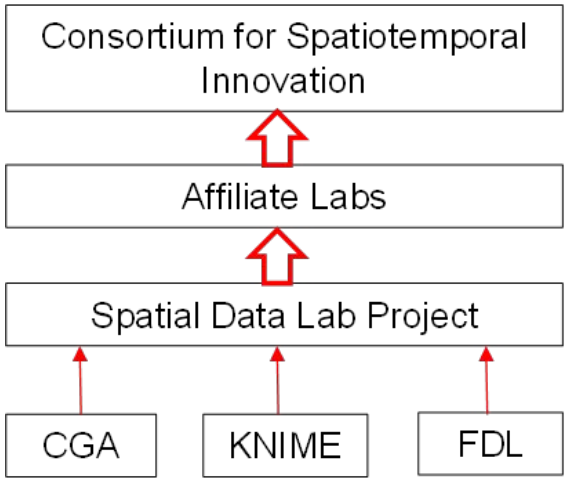
Building a Global Network of Affiliated Labs

Tasks for the affiliate labs:

- Collect and share research data
- Participate in the development and test of Geospatial Analytics (Knime nodes) and case studies
- Participate in the development of new case studies (5-10 each year)
- Organize or participate in public webinars or training workshops (3-5 each year)
- Apply SDL technology and case studies to research and teaching

Support from the SDL:

- Leadership and technical training
- Shared case study database
- Research data support
- Technical support
- Promotion of the achievements by the affiliate labs
- Reorganization of the achievements by the affiliate labs



The List of Current Affiliated Labs

- ❑ Spatial Data Lab for Big Data
- ❑ Healthy Urban Renewal Lab
- ❑ Spatial Data Lab for Regional Policy and Future Work
- ❑ Criminal Justice Data Analytics Lab
- ❑ Environmental, Health, and Development Lab
- ❑ Spatial Data Lab for Urban Informatics
- ❑ Spatial Data Lab for Healthcare System
- ❑ Spatial Data Lab for Geocomputation
- ❑ Spatial Social Science Research Lab
- ❑ Urban Inequality Lab
- ❑ Space Studies and Prospective Laboratory

Sponsors and Contributing Open Source Projects

NSF IUCRC
Spatiotemporal Innovation Center(STC)
https://www.stcenter.net/

National Science Foundation

Future Data Lab

Open for Innovation
KNIME

Open Street Map

kepler.gl

GeoPandas

Leaflet

The Python Spatial Analysis Library for open source, cross platform Geospatial Data Science

pulp