*Syllabus*

**Geographic Data Science: *Geospatial Big Data Analytics and Applications***

**Course Objective:**

This course empowers students with Data Science approaches, leveraging Geographic Big Data and data analytics to address urban environmental and societal challenges. Through hands-on experience with diverse real-world datasets, students will develop computational skills in harnessing the power of geographic Big Data, including but not limited to, geographic data management, geospatial analysis, text analytics, and spatial data mining and machine learning techniques.

**Datasets Used in This Course:**

Geo-located Social Media Data (Geo-located tweets)

* A collection of geo-located Twitter data with worldwide coverage (2014–2022).
* 2014 data will be used for course demonstrations. Students have access to all datasets for project work and exploring topics of interest.

Mobile Device Location Data (SafeGraph)

* Mobile device location data covering the United States (2018-2021)
* POI (Point-of-Interest) data showing visitor origins and visitation patterns

Remote Sensing Imagery and Geo-tagged Photographs

* High-resolution satellite imagery data from SpaceNet and NOAA for building footprint detections
* Geo-tagged photographs from GeoGraph with user-rated scenic scores (Image segmentation and scenic perception/aesthetics)

Spatial Networks

* OpenStreetMap (Street Networks)
* Spatial Networks of Interactions (e.g., human mobility/activity, movement/information/interaction flows, and physical/virtual connectivity)

**Course Schedule**

**Week 1: Introduction to Geographic Data Science (GDS)**

* The new paradigm of geographic Big Data and Geographic Data Science.
* Overview of geographic data science and its applications.
* Understanding social and environmental challenges.
* The role of Big Data in addressing urban and population dynamics.
* Different types of geographic Big Data sources.

**Week 2: Data Management and Computing Resources**

**Week 3: Data Management and Computing Resources**

*Computing Environment*:

* Introduction to the state-of-art computational platforms/frameworks, such as high-performance computing (HPC) and distributed computing environments for GDS (introduction of **CyberGIS**).
* Demonstration of access and basic operations in a HPC environment (e.g., supercomputing resources provided by the ACCESS (<https://access-ci.org>) resources or GW-offered computing clusters).

*Data Management*:

* Data types and formats (Tabular data, Shapefile, (geo)JSON, plain text, etc.).
* Geographic datasets: Rater vs Vector
  + Spatial indexing for rapid spatial search (Coding: GeoPandas and Spark operations)
  + Intro to spatial databases
* NoSQL databases (e.g., MongoDB, demonstration).
* Introduction to Apache Hadoop and Spark for Big Data query and processing.
* (Labs: Accessing HPC/cloud, MongoDB data storage, indexing, and query)

**Week 4**: **Geographic Big Data Analytics**

* Common geospatial operations (Nearest Neighbor search, point-in-polygon, spatial join)
* Software (ArcGIS and QGIS), computing tools/frameworks (Python and related packages), and visualization tools/frameworks (web-based map interfaces)
* Examples/Case studies:
  + Geo-located tweets data pre-processing, e.g., extracting geo-locations of each tweet and other useful fields of information
  + Using geographical boundaries to extract data records fall in (e.g., SafeGraph data and geo-tagged tweets)
  + Illustrate the processes with Python scripts (available in Jupiter notebook)
* (Labs: Run Python scripts to extract geo-locations from geo-located tweets, creating Shapefile to visualize in QGIS/ArcGIS, and as overlays on OpenStreetMap)

**Week 5**: Text Analytics in Geographical Contexts

Social media data analysis with geo-located tweets

* Topic modeling (word cloud, TF/IDF, BERTopic)
* Classification (machine learning approaches: Scikit-learn based)
* Sentiment analysis
* (Labs: Apply lexicon-based sentiment analysis on geo-located tweets and visualize the results)

**Week 6**: **Spatial Network Modeling and Analysis**

Network Modeling:

* Spatial networks
* Real-world physical networks (constructing street networks as graphs)
* Real-world virtual networks (constructing a network of flows as graphs, e.g., movements captured by geo-located tweets and SafeGraph)

Network Analysis:

* Network centrality measures and applications
* Community detection concepts and methods
* (Labs: constructing spatial networks from OSM street networks, measure network centrality, and visualize the results)

**Week 7:** Mid-term Project Proposal

Students will submit an initial project proposal outlining their research question and proposed analytical methods.

**Week 8**: **Image-based Data Analysis**

* Satellite image processing and analysis (Python with RasterIO)
* Building detection with Neural Network models (Hugging face models)
* Image segmentation (Segment Anything Model (SAM) from Meta)
* (Labs: perform image segmentation on satellite images from NOAA and run building footprint detection using model from Hugging face and ArcGIS)

**Week 9: Exploratory Spatial Data Analysis (ESDA)**

**Week 9**: Spatial Data Mining and Machine Learning - Part 1

* Exploratory Spatial Data Analysis (ESDA) using Python (GeoPandas, PySAL).
* Spatial Autocorrelation:
  + Moran’s I and Local Indicators of Spatial Association (LISA).
* Geospatial queries, spatial joins, buffering, overlay, and proximity analysis

**Week 10**: Spatial Data Mining and Machine Learning - Part 2

* Spatial Machine Learning Fundamentals:
  + Traditional vs. spatial machine learning methods.
  + Supervised and unsupervised learning using spatial predictors.
* Spatial clustering methods (e.g., DBSCAN and HDBSCAN)
* Introduction to spatially explicit neural networks and deep learning.

**Week 11**: Geographic Data Science for Social Good – Part 1

Understanding Human Interactions with Urban Space:

* Mobility patterns, daily activities, population distribution.

Geospatial Data Processing:

* Pinpointing social media text records (geotagged tweets) to spatial units (e.g., buildings, addresses, census tracts, counties, states).
* Using Python and GeoPandas for geospatial operations on large-scale datasets.

Geographic Data Fusion Integrative Analysis:

* Combining social media data, remote sensing imagery, and infrastructure datasets (e.g., subway access, EV charging stations, points of interest).

**Week 12**: Geographic Data Science for Social Good – Part 2

Exploring Social and Environmental Challenges:

* Evaluating environmental equity and exposure
* Case studies on environmental justice and public health (pollution exposure with SafeGraph data, environmental perception using GeoGraph data.)

Student-Led Discussions:

* Encouraging students to propose and discuss their own research topics related to social and environmental challenges.

Case Study Demonstration:

* Mining social media data to evaluate public perception (sentiment/emotion) of metro transportation in Washington, D.C., using techniques covered in previous weeks.

**Week 14**: Final Project Presentations

Students will work on a data-driven research project applying computational social science methods to analyze a social or environmental issue of their choice. The project will involve data acquisition, processing, analysis, and visualization, culminating in a final presentation.

**Grading:**

* Labs/Homework Assignments (weekly labs): 40%
* Mid-term Project Proposal: 15%
* Final Project and Presentation: 35%
* Participation/attendance and Student-led Discussions: 10%

|  |  |  |
| --- | --- | --- |
| Score | Grade |  |
| >= 90 | A (Excellent) | 4 |
| >= 85 && < 90 | A- (Intermediate grade) | 3.5 |
| >= 80 && < 85 | B (Good) |  |
| >= 75 && < 80 | B- () |  |
|  |  |  |

**Course Materials:**

All lecture slides and notes are available in the GitHub repository.

Recommended reading materials:

Yin, Junjun, Guangqing Chi, and Bin Jiang. "Unlocking the Secrets of Scenic Beauty: A Quantitative Analysis of Object Variety and Connections in Scenic Images." Journal of the Royal Society Interface (2025)

Yin, Junjun, Matthew Brooks, Donghui Wang, and Guangqing Chi. "Characterizing climate change sentiments in Alaska on social media." Digital Geography and Society 8 (2025): 100110.

Yin, Junjun, and Guangqing Chi. "A tale of three cities: uncovering human-urban interactions with geographic-context aware social media data." Urban Informatics 1, no. 1 (2022): 20.

Yin, Junjun, Yizhao Gao, and Guangqing Chi. "An evaluation of geo-located Twitter data for measuring human migration." International Journal of Geographical Information Science 36, no. 9 (2022): 1830-1852.

Alba, Charles, Bing Pan, Junjun Yin, William L. Rice, Prasenjit Mitra, Michael S. Lin, and Yun Liang. "COVID-19’s impact on visitation behavior to US national parks from communities of color: Evidence from mobile phone data." Scientific reports 12, no. 1 (2022): 13398.

Yin, Junjun, and Guangqing Chi. "Characterizing people’s daily activity patterns in the urban environment: A mobility network approach with geographic context-aware twitter data." Annals of the American Association of Geographers 111, no. 7 (2021): 1967-1987.

Yin, Junjun, Aiman Soliman, Dandong Yin, and Shaowen Wang. "Depicting urban boundaries from a mobility network of spatial interactions: A case study of Great Britain with geo-located Twitter data." International Journal of Geographical Information Science 31, no. 7 (2017): 1293-1313.

Jiang, Bin, Junjun Yin, and Sijian Zhao. "Characterizing the human mobility pattern in a large street network." Physical Review E—Statistical, Nonlinear, and Soft Matter Physics 80, no. 2 (2009): 021136.