MACSS 30112 Winter 2025
Project Report 1
Group Chick-Fil-A
Charlotte Li, Baihui Wang & Anqi Wei

Research proposal: Investigating the Global Spatial Distribution of Open-Source Artificial Intelligence Development

1. Project Description

Open-source artificial intelligence (AI) has become a driving force in technological advancement, fostering innovation through collaborative development. Platforms like GitHub have facilitated remote cooperation among developers and transformed digital outputs into a crucial component of collective knowledge. While the digital nature of open-source AI theoretically enables global accessibility, empirical observations indicate a pronounced geographical clustering of AI developers in specific regions, such as California in the United States, the Yangtze River Delta and Pearl River Delta in China, and other emerging AI hubs.

This study aims to systematically examine the global spatial distribution of open-source AI development, addressing three key questions: (1) Which cities and regions contribute the most to AI-related open-source projects? (2) To what extent do economic development and the density of higher education institutions explain these geographic patterns? (3) Is AI open-source development more geographically decentralized compared to AI industry clusters, such as AI companies and R&D centers?

This project expands upon our initial proposal, which initially focused solely on the United States. Preliminary data scraping from GitHub revealed that contributors are distributed across various global cities, prompting us to adopt a broader, international approach. By analyzing city-level data before aggregating at the national and regional levels, we aim to provide a comprehensive understanding of AI development's spatial distribution.

From a social science perspective, this study examines the intersection of technological innovation, economic geography, and knowledge diffusion. Open-source AI development plays a crucial role in democratizing access to cutting-edge technology, yet its concentration in certain regions raises questions about digital inequality and the accessibility of AI expertise. Understanding the socioeconomic and institutional factors driving AI clustering can inform policies that promote more equitable technological development globally.

This research builds upon existing work in economic geography and the sociology of technology, particularly studies on the unequal spatial distribution of open-source software production (Wachs, 2022). Prior research highlights the significant economic

value of OSS, demonstrating that open-source software contributes to both the supply side (cost efficiency in software development) and the demand side (widespread usage across industries) (Harvard Business School, 2023). Moreover, OSS has been shown to fuel entrepreneurial growth, as businesses and individual developers strategically leverage open-source platforms to drive innovation and competitiveness (Harvard Business School, 2023). However, access to these benefits remains geographically concentrated. Economic geography research suggests that regional innovation systems, characterized by venture capital availability, high-tech firms, and leading universities, create environments where OSS development flourishes, such as Silicon Valley and China's tech hubs. The role of higher education institutions is also critical, as they serve as hubs for knowledge diffusion and skill development, further reinforcing geographic disparities in OSS contributions (Harvard Business School, 2023). By applying computational methods to analyze spatial patterns in AI open-source participation, our study bridges traditional social science theories with data-driven insights, offering empirical evidence on the socioeconomic determinants of Open-source Artificial intelligence activity.

To conduct this study, we define and operationalize the following key concepts:

1. **To measure the AI Open-Source Development Activity**, we will scrape AI-related repositories, ranking them by the number of stars they have received, which serves as a proxy for community recognition and project impact. For each repository, we will extract key metadata, including the number of contributors, commit frequency, and the geographic locations of contributors (as provided in their GitHub profiles). This data will be stored in a structured CSV file for analysis.
2. **Geographic Clustering**: we will use spatial concentration metrics such as the Gini coefficient and location quotient (LQ) to determine the extent of AI development centralization.
3. **Regional Economic Development**: we will look into GDP per capita, median income, and investment in research and development (R&D) at the city or regional level.
4. **Higher Education Density**: Operationalized as the number of universities and AI-related research institutions per capita in a given region, supplemented by university rankings and AI-related publication output.
5. **AI Industry Presence**: Defined by the number of AI startups, corporate R&D centers, and venture capital investment in AI within a given geographic area.

We expect to find that AI open-source development follows a semi-concentrated pattern, meaning that while digital collaboration enables some degree of decentralization, the most active contributors remain clustered in economically developed regions with strong higher education institutions. Specifically:

- **H1:** Cities with higher GDP per capita and stronger higher education infrastructure will have significantly higher levels of AI-related open-source activity.
- **H2:** The geographic distribution of AI open-source development will be less concentrated than that of AI corporate R&D centers but will still exhibit clustering in major innovation hubs.
- **H3:** Regions with strong AI industry presence (e.g., AI startups and venture capital investment) will also have a high concentration of open-source AI contributions, suggesting a reinforcing relationship between corporate AI development and open-source collaboration.

By testing these hypotheses, this study aims to contribute empirical evidence on the spatial dynamics of AI innovation, bridging insights from computational social science, economic geography, and technology policy.

2. Data Sources

Data Source	Scraping/Downloading	Timeframe	Data Volume	Reliability & Validity Issues	Additional Info
GitHub AI Repositories: https://github.com/topics/ai	Scraping	Current (live data)	Top 100 most-starred AI repositories, 1,514 unique contributors with location data (where available)	- Users may not always provide accurate location data. - Data may be skewed toward popular repositories. Solution: Cross-check locations with other sources; clean data for consistency.	Focus on repository stars as a popularity metric; filter for active contributors
World Higher Education	Downloading	Latest available data	21,000+ accredited higher	-Variability in data quality	Authoritative global database of

Database (WHED): https://www.whed.net/home.php			education institutions across 180+ countries, ignoring countries with insufficient data	across countries. -potential outdated records. Solution: cross-reference with national education databases where possible.	accredited higher education institutions.
World Bank World Development Indicators: https://databank.worldbank.org/	Downloading	Latest available datasets (e.g., 2022-2024)	Data for Covers GDP, population, and economic metrics for 200+ countries; thousands of data points spanning multiple indicator	- Data collection methods vary by country. -Time lags in reporting. Solution: compare with IMF and national statistics where discrepancies arise.	Comprehensive source for global GDP and macroeconomic indicators.
Stanford AI Index Report: https://aiindex.stanford.edu/report/	Downloading	Latest report (e.g., 2024)	Report format with quantitative data on AI publications, investments, and global trends across dozens of countries	- Aggregated from multiple sources, so methodology varies. - May not include all OSS contributions. Solution: Use AI	Useful for benchmarking findings from GitHub data.

				Index insights for contextual analysis, not raw data.	
World Bank Internet Usage Data: https://data.worldbank.org/indicator/IT.NET.USER.ZS	Downloading	Latest report (e.g., 2020-2024)	Internet penetration rates globally, annually updated	Differences in data reporting standards; addressed by using consistent country-level definitions	Critical control variable for tech accessibility and OSS participation

3. Data Cleaning and Wrangling

To ensure the reliability and consistency of our analysis, we will undertake a structured data preparation process to clean, standardize, and integrate multiple data sources. Given our study's focus on the global participation in open-source AI projects, we will apply the following steps:

a. **Extracting AI Open-Source Participation Data**
   - Scrape AI-related repositories from GitHub, ranking them by star count as a proxy for project impact.
   - Collect metadata, including:
     - Repository name, creation date, total star count, number of contributors, and commit frequency.
     - Contributor profile information, such as geographical location and organizational affiliation (if available).
   - Convert all contributor locations into city, country format, removing missing or ambiguous entries.
   - Aggregate repository and contributor data at both city and country levels, forming the basis for calculating the AI Open-Source Participation Index (AIOSPI).

b. **Standardizing Geographic Data**
   - Extract and clean location data from GitHub user profiles.
   - Use geocoding APIs (Google Maps API, OpenStreetMap) to normalize city and country names, ensuring consistency in location-based analysis.

- Aggregate contributors by city and country, enabling computation of:
    - Total number of contributors per region.
    - Total and average star count per contributor.
    - Per capita participation rates (contributors and stars per population).

c. Merging with External Data Sources

To examine the relationship between AI open-source participation and socioeconomic factors, we will integrate the following datasets:

- Economic Data: 2023 GDP per country (sourced from the World Bank, IMF, or OECD).
- Higher Education Data: Number of universities per country (sourced from the QS World University Rankings or national education databases).
- Digital Infrastructure Data (Control Variable): Internet penetration rate (percentage of the population using the Internet), to account for regional disparities in online accessibility.

d. Cleaning and Formatting Data

- Handling Missing Values: Use interpolation or regional averages where necessary to fill gaps in contributor location data.
- Standardizing Population Data: Extract country and city population figures from UN population estimates and national census reports to compute per capita metrics.
- Converting Categorical Data: Transform country names, repository tags, and university names into structured variables for consistent aggregation.

e. Generating Summary Statistics and Initial Visualizations

- Compute summary statistics (mean, median, standard deviation, min/max) for both dependent and independent variables.
- Rank countries and cities by AIOSPI and contributor density to identify global AI hubs.
- Create heatmaps and bubble charts using GeoPandas to visualize geographic clustering of AI contributors.
- Plot a social network graph using NetworkX to map collaboration patterns among the top 10 AI repositories.

f. Preparing for Regression Analysis

- Ensure all variables are in the appropriate format for Ordinary Least Squares (OLS) regression in Python.
- Conduct multicollinearity checks to verify that independent variables (GDP, university density, digital infrastructure) are not overly correlated.
- Generate scatter plots and correlation matrices to visually assess relationships between AIOSPI and economic/education variables.

4. Data Analysis and Visualization

In this study, we will focus on each country's participation in open-source AI projects. Specifically:

a. **Dependent Variable: AI Open-Source Participation Index (AIOSPI)**:

We will first gather the geographical locations of contributors and the star counts of the repositories they work on. Based on this, we will calculate the "number of contributors in a given region" as well as the "total star count and average star count per contributor." Next, we will derive per capita contributors/stars for each region based on its population. We will observe the results from two calculation methods in our regression analysis:

- **Geographic AIOSPI**: AIOSPI = 0.5 × (number of contributors) + 0.5 × (total star count)
- **Per Capita AIOSPI**: AIOSPI = 0.5 × (number of contributors / population) + 0.5 × (total star count / population)

These figures allow us to assess the relative level of activity in open-source AI across different regions.

b. **Independent Variables**:
   - **Regional economic level** (proxied by each country's GDP in 2023)
   - **Per capita number of universities** ("number of universities in 2023" / "local population in 2023")

We will conduct regression analysis to examine whether these external resources significantly influence the level of open-source AI participation.

In terms of the analysis process, we will perform data visualization:

1) Calculate the mean, median, standard deviation, maximum, and minimum values of the dependent and independent variables, rank countries accordingly, and also identify the highest-density cities.

2) Determine the global distribution of AI contributors, and use **geopandas** to create heat maps or bubble charts at both global and more granular levels (e.g., metropolitan areas). Through varying color intensities or classified color schemes, we will visualize the per capita distribution of AI repository contributors and national rankings, revealing which areas show high-density clusters and which fall behind.

3) Draw a social network graph using **networkx**. Specifically, we will plot the collaboration network (contributors from different cities working in the same GitHub repository) among builders of the top 10 most significant repositories, to uncover how knowledge spills over and how collaboration takes place across regions.

We will then employ multiple linear regression (OLS) in **Python**, controlling for **Digital Infrastructure** (e.g., Individuals using the Internet [% of population]) to test whether the independent variables correlate significantly and positively with the dependent variables. If the results indicate that areas with higher GDP or greater university density exhibit more robust open-source contributions, this would confirm a coupling between resource concentration and technological innovation. Conversely, if the relationship is not significant, it may suggest that open-source platforms can mitigate the constraints that traditional resource distribution imposes on the spread of cutting-edge technologies.

We will also select the top three cities with the highest contributor density for further scrutiny, examining whether their high levels of activity stem predominantly from economic and academic advantages or whether specific policies and international collaborations also play a substantial role.

Through this multi-layered quantitative and visualization-based analysis, coupled with in-depth exploration of representative cases, we aim to systematically unveil the global spatial distribution of AI open-source activity and the economic, educational, and industrial factors underpinning it. This study not only continues the scholarly tradition of economic geography and the sociology of technology but also offers new empirical insights for discussions on the digital divide, technology policy, and global innovation cooperation.

5. Responsibilities

- Charlotte Li: Assist with geographic location matching, data regression analysis, and visualization code
- Baihui Wang: Improved the geographic data scraping, lead data cleaning and preposscing efforts, and contribute to regression analysis and visualization.
- Anqi Wei: complete the final steps of debugging the current version of the code (web-scraping), revise the README, and finalize the project report