

# **Project ideas**

**Some hints to start with...**

# Dataset 1: human mobility data

## HumMob Challenge 2023

- <https://connection.mit.edu/humob-challenge-2023>
- <https://zenodo.org/record/8111993>
- Paper: <https://arxiv.org/abs/2307.03401>
- Access is granted upon request
- Questions: analyze individual trajectories, predict human movements, analyze POIs

# Dataset 2: collective attention

## Wikipedia pageview data

- Data: <https://dumps.wikimedia.org/other/analytics/>
- Pageview data of all Wikipedia pages, disaggregated by project and by country (only recent data)
- Questions: how do collective attention to public events differ across countries/languages? Is collective attention driven by news, or other?

# Dataset 3: social networks

## Long-ties social network data

- <https://data.humdata.org/dataset/long-ties-data>
- Paper: <https://www.pnas.org/doi/10.1073/pnas.2211062120>
- The data measures the frequency of long ties in different geographic locations of the US and Mexico. Long ties, the edges in the connectivity network without any mutual contact, have been shown to play an important role in access to economic opportunities.
- Questions: are long ties related to social mobility? Can they explain health behaviours like vaccination adoption or hesitancy?

# Dataset 4: social mobility and inequalities

## Opportunity insights

- Data: <https://opportunityinsights.org/data/>
- High resolution data on: education, social mobility, income, racial disparities, and many other socio-economic indicators in the US at different spatial resolutions:
  - State, county, zip code
- Questions: do these quantities correlate with long ties data? Do they explain other social outcomes?

# Dataset 5: social networks

## High-resolution proximity networks

- Large scale datasets of individual proximity and mobility, collected with wearable sensors and mobile phones:
  - Copenhagen Network Study: [https://figshare.com/articles/dataset/The\\_Copenhagen\\_Networks\\_Study\\_interaction\\_data/7267433/1](https://figshare.com/articles/dataset/The_Copenhagen_Networks_Study_interaction_data/7267433/1)
  - SocioPatterns: <http://www.sociopatterns.org/datasets/>
- Paper: <https://www.nature.com/articles/s41597-019-0325-x>
- Papers: <http://www.sociopatterns.org/publications/>
- Questions: comparison of network properties between settings, analysis of contagion models on different datasets, creating agent-based models with empirical data.

# Dataset 6: News articles

## GDELT

- Large-scale collection of news articles published on the Web
- Article links, entities cited, topics, sentiment, etc
- Possibility to retrieve also article text
- <https://www.gdeltproject.org/data.html>
- (Suggestion: access through Google Cloud Platform)
- Questions: media coverage of specific events, relationship (networks) between topics, framing of events, ...

# Dataset 7: Open-source developer community

## GitHub

- Large-scale collection of GitHub EVENTS
- Event types ranging from commits, forks, pushed etc.
- Easy download via wget @ <https://www.gharchive.org/>
- Questions: temporal dynamics of repositories, resolution of issues, collaboration, communities of open-source developers ...