The Social Impacts of the U.S. Data Center Boom: A Social Data Science Project Summary

Project Proposal Summary

November 23, 2025

1 Overview

This project examines the rise of data centers in the United States and their social impacts, focusing on energy systems, environmental outcomes, and community-level effects such as housing costs. Data centers form the physical backbone of the digital economy and recent artificial intelligence (AI) growth. Their rapid expansion has raised concerns about soaring electricity demand, greenhouse gas emissions, water use, and pressures on local infrastructure.

We will construct a multi-year, state-level dataset for all U.S. states and the District of Columbia, covering at least six recent years. The dataset will integrate indicators of data center growth, energy use, emissions, water stress, air quality, and socio-economic conditions. Using this dataset, we will apply social data science methods—including descriptive analysis, panel regression, and clustering—and develop Tableau dashboards to tell a coherent story about who gains and who bears the costs of the data center boom.

2 Research Motivation and SDG Alignment

Data centers are highly concentrated in particular regions and increasingly dominate new electricity demand. At the same time, local communities and regulators face trade-offs: data centers can bring jobs and tax revenues, but they also require large amounts of electricity and, often, water for cooling. These demands may contribute to higher energy prices, increased emissions, and environmental stress, especially in already vulnerable areas.

The project is aligned with the following Sustainable Development Goals (SDGs):

- **SDG 7: Affordable and Clean Energy** – Investigating how large digital loads affect electricity consumption and prices, and how these impacts may vary across states.
- **SDG 13: Climate Action** – Assessing the relationship between data center intensity and power-sector CO_2 emissions and air quality.
- **Related SDGs** – SDG 8 (Decent Work and Economic Growth) through employment and wages; SDG 9 (Industry, Innovation and Infrastructure) through digital infrastructure; and SDG 11 (Sustainable Cities and Communities) via housing and environmental stress.

3 Research Questions and Hypotheses

We will address the following main research questions:

- **RQ1 (Uneven growth):** How uneven is data center growth across U.S. states over the last six years?
- **RQ2 (Energy and prices):** Are states with faster data center growth experiencing larger increases in per-capita electricity consumption and higher electricity prices?
- **RQ3 (Emissions and air quality):** Is data center intensity associated with higher per-capita CO_2 emissions from the electric power sector and worse air quality?
- **RQ4 (Water and stress):** Do data center–intensive states with high baseline water stress face particularly strong energy and emissions impacts?
- **RQ5 (Housing and communities):** Is data center intensity associated with rising housing costs, after controlling for income and unemployment?
- **RQ6 (Typologies):** Can we identify clusters of states with distinct "data center–impact" profiles (e.g., hyperscale hubs vs. low-digital states)?

Guided by theory and prior evidence, we hypothesize that:

- Higher data center intensity is positively associated with per-capita electricity sales and power-sector emissions.
- In states with limited clean energy and high baseline water stress, the link between data center intensity and emissions is stronger.
- Data center hubs may experience more rapid increases in rents and home values, especially in already tight housing markets.
- The employment share of data center-related industries remains relatively small, suggesting that energy and environmental costs may be more diffuse than direct job benefits.

4 Dataset Design

4.1 Unit of Analysis and Time Frame

The unit of analysis is the **state–year**. We will construct a balanced panel dataset with:

- All 50 U.S. states plus the District of Columbia.
- A time span of **six recent years** (e.g. 2019–2024).

This yields at least 306 instances (51×6), satisfying the requirement of more than 300 observations.

4.2 Data Sources

The dataset will be constructed by merging open, downloadable data from multiple official sources:

- **Data center and digital sector:** State-level employment, wages, and establishments in NAICS 518210 (Data Processing, Hosting, and Related Services) from the BLS Quarterly Census of Employment and Wages (QCEW); geocoded data center locations and floor area from an open data center atlas.
- **Energy and prices:** Electricity sales and average retail prices by state and sector (residential, commercial, industrial) from U.S. Energy Information Administration (EIA) state electricity profiles.
- **Emissions:** Power-sector CO_2 emissions and related metrics from EIA's State Energy Data System (SEDS).
- **Water and water stress:** State-level public water withdrawals from U.S. Geological Survey (USGS) water-use datasets; baseline water stress indices from the World Resources Institute's Aqueduct project.
- **Air quality:** State-level aggregations of $PM_{2.5}$ and related air quality metrics derived from U.S. Environmental Protection Agency (EPA) Air Quality System (AQS) data.
- **Socioeconomic and housing:** Median household income, population, median gross rent, median home value, and rent burden from the American Community Survey (ACS); state unemployment rates from BLS Local Area Unemployment Statistics (LAUS).

4.3 Attributes

The combined dataset will include well over fifteen attributes, such as:

- Data center employment, establishments, and wages; employment share and facilities per capita.
- Electricity sales per capita and average prices across sectors.
- Power-sector CO_2 emissions per capita.
- Baseline water stress and public water withdrawals per capita.
- Air quality indicators (e.g., mean $PM_{2.5}$).
- Population, median income, unemployment, median rent, median home value, rent burden.

This structure provides a rich basis for exploring linkages between data centers and social outcomes.

5 Methods and Analytical Strategy

5.1 Descriptive Analysis

We will begin with descriptive statistics and exploratory data analysis to understand the distribution and trends of key variables, including:

- Growth in data center employment and footprint by state.
- Trends in per-capita electricity sales, prices, emissions, and housing costs.
- Correlations among data center intensity, energy use, emissions, water stress, and socioeconomic indicators.

5.2 Regression Analysis

Next, we will estimate panel regression models to assess associations between data center intensity and various outcomes, controlling for other factors:

- Models relating data center jobs per capita to per-capita electricity sales and average retail electricity prices.
- Models relating data center intensity to per-capita power-sector CO_2 emissions and air quality metrics.
- Models relating data center intensity to housing outcomes (median rent, median home value), with controls for income and unemployment.
- Interaction terms between data center intensity and baseline water stress to test whether impacts are stronger in water-stressed states.

Fixed effects for states and years will help control for time-invariant differences across states and common national shocks.

5.3 Clustering and Typologies

We will apply clustering methods such as k-means or hierarchical clustering to create typologies of states based on:

- Data center intensity (employment and facilities).
- Per-capita electricity use and prices.
- Emissions, water stress, and housing costs.

This will help identify groups such as "data center hubs under stress", "energy-intensive but not data-heavy", and more balanced states.

6 Visualization and Storytelling with Tableau

Using Tableau Desktop, we will build dashboards to communicate our findings:

- Maps of data center intensity and data center facility footprints by state.
- Time-series comparisons of electricity use and prices in high-intensity versus low-intensity states.
- Scatterplots of data center intensity versus electricity use, prices, emissions, and housing outcomes.
- Bivariate maps and interaction plots showing where high data center intensity overlaps with high baseline water stress.
- Cluster maps that present typologies of states and their associated social impacts.

The visualizations will be designed to tell a clear story about how digital infrastructure growth is linked to energy, environmental, and community outcomes.

7 Expected Contributions

The project will contribute:

- An integrated, open-data-based panel dataset on data centers and social impacts at the state level.
- Empirical evidence on how data center growth is associated with electricity consumption, prices, emissions, and housing.
- A typology of state experiences that highlights inequality and environmental justice concerns.
- Interactive visualizations for non-technical audiences and policymakers.

Overall, the project aims to provide a rigorous, data-driven assessment of the social implications of the U.S. data center boom and to demonstrate the power of social data science in addressing contemporary infrastructure questions.