Implementation Plan:
The Social Impacts of the U.S. Data Center Boom

Detailed Project Implementation Plan

November 23, 2025

1 Project Structure and Phases

The project will be implemented in four main phases:

1. **Phase 1: Conceptual design and data scoping**
2. **Phase 2: Data acquisition and construction of the state–year panel**
3. **Phase 3: Analysis and modeling**
4. **Phase 4: Visualization, interpretation, and reporting**

Each phase is described in detail below, with explicit tasks and outputs.

2 Phase 1: Conceptual Design and Scoping

2.1 Refine Research Questions and Theory

- Finalize the set of research questions:
  - RQ1: Uneven data center growth across states.
  - RQ2: Data centers and electricity use/prices.
  - RQ3: Data centers, emissions, and air quality.
  - RQ4: Data centers and water stress.
  - RQ5: Data centers and housing outcomes.
  - RQ6: Typologies of states via clustering.
- Clarify the theoretical lenses:
  - Critical infrastructure and socio-technical systems.
  - Environmental justice (distribution of burdens/benefits).
  - Political economy of digital infrastructure and AI.
- Map out hypothesized causal chains, e.g.:

Digital & AI growth → Data center expansion → Electricity demand → Prices & emissions → Community i

2.2 Define Unit of Analysis and Time Window

- Unit: U.S. state (50 states + District of Columbia).
- Time frame: six most recent years with consistent data (e.g. 2017–2022).
- Dataset size: 51 units × 6 years = 306 panel observations.

3 Phase 2: Data Acquisition and Construction

This is the core technical phase, focused on building the integrated dataset.

3.1 Step 2.1: Create Panel Skeleton

- Construct a base table with variables:
  - `state`, `state_fips`, `year`.
- Enumerate all combinations of state and year for 2017–2022.
- This table is the merge "backbone" for all subsequent joins.

3.2 Step 2.2: Data Center and Digital Sector Data

BLS QCEW – NAICS 518210

- Download state-level QCEW data for NAICS 518210 for the period 2017–2022.
- Key variables:
  - `dc_emp`: annual average employment.
  - `dc_estab`: number of establishments.
  - `dc_total_wages`: total annual wages.
  - `dc_avg_wage`: derived as total wages divided by employment.
- Similarly download QCEW totals for all industries to construct:
  - `total_emp`: total private employment.
  - `dc_emp_share` = `dc_emp / total_emp`.

Data Center Footprint – Open Atlas

- Download the open data center atlas or similar geospatial dataset with facility points and building footprints.
- Aggregate by state:
  - `dc_facilities_count`: number of data center facilities.
  - `dc_total_sqft`: sum of data center floor area.
- Merge state land area (from Census Gazetteer) to compute:

  - `dc_sqft_per_capita`: footprint per resident.
  - `dc_land_share`: footprint as a fraction of land area.
- Treat these footprint indicators as static across the 2017–2022 panel.

3.3 Step 2.3: Electricity Sales and Prices (EIA)

- Download state-level electricity sales and average price data from EIA for 2017–2022.
- Extract variables:
  - `elec_sales_total_mwh`: total sales (all sectors).
  - `elec_sales_comm_mwh`: commercial-sector sales.
  - `elec_sales_ind_mwh`: industrial-sector sales.
  - `price_all_cents_kwh`: average retail price, all sectors.
  - `price_comm_cents_kwh`: commercial price.
  - `price_ind_cents_kwh`: industrial price.
- Later, compute per-capita metrics after merging population data:
  - `elec_sales_pc_mwh` = `elec_sales_total_mwh / pop`.

3.4 Step 2.4: Emissions (EIA SEDS)

- Download state-level energy-related CO_2 emissions from the State Energy Data System.
- Focus on variables:
  - `co2_power_mt`: power-sector CO_2 (million metric tons).
  - Optionally `co2_total_mt`: total CO_2 from all sectors.
- Compute per-capita measures:
  - `co2_power_pc_tons` = `co2_power_mt` $\times 10^6 /$ `pop`.

3.5 Step 2.5: Water Use and Water Stress

USGS Water Use

- Download state-level water-use data (e.g. 2015 and 2020).
- Extract public supply withdrawals:
  - `public_withdrawals_mgd`: million gallons per day.
- Compute per-capita withdrawals:
  - `public_withdrawals_pc_gpd` = `public_withdrawals_mgd` $\times 10^6 /$ `pop`.
- Use 2015 values for earlier years and 2020 for later years, or treat the average as a static indicator.

Baseline Water Stress (Aqueduct)

- Download baseline annual water stress indices for U.S. administrative units.
- Aggregate to the state level if necessary.
- Add:
  - `baseline_water_stress`: continuous index, or categorized (low/medium/high).
  - Optionally `baseline_water_depletion`.
- Treat these as static over the 2017–2022 period.

3.6 Step 2.6: Air Quality (EPA AQS)

- Download pre-generated daily $PM_{2.5}$ (and optionally ozone) files for 2017–2022.
- For each monitor:
  - Compute annual average $PM_{2.5}$.
  - Count days above the standard threshold (e.g. 12 $\mu g/m^3$).
- Aggregate to state-year level:
  - `pm25_mean_ugm3`: mean of monitor averages (or population-weighted).
  - `pm25_days_above_12`: average days above 12 $\mu g/m^3$.

3.7 Step 2.7: Socioeconomic and Housing Indicators

ACS State-Level Indicators

- For each year 2017–2022, collect:
  - `pop`: total population.
  - `median_income`: median household income.
  - `median_gross_rent`: median gross rent.
  - `median_home_value`: median owner-occupied home value.
  - `rent_burden_share`: % renters spending $\geq 30\%$ of income on rent.

Unemployment and Broader Digital Sector

- From BLS LAUS:
  - `unemployment_rate`: state annual unemployment rate.
- From QCEW (Information sector, NAICS 51):
  - `info_emp_share`: information-sector employment as a share of total employment.

3.8 Step 2.8: Merging and Cleaning

- Sequentially merge all datasets onto the `state--year` skeleton using consistent state codes and years.
- Ensure each merge preserves the panel structure (no duplicate state–year combinations).
- Handle missing values due to:
  - Data suppression (e.g. small QCEW cells) by interpolation or exclusion when rare.
  - Infrequent measurement (e.g. water-use years) by carrying forward/back or averaging.
- Standardize key continuous variables (e.g. z-scores) for clustering:
  - `dc_emp_per_10k_pop`, `elec_sales_pc_mwh`, `price_all_cents_kwh`, `co2_power_pc_tons`, `median_gross_rent`, etc.

4 Phase 3: Analysis and Modeling

4.1 Step 3.1: Descriptive Statistics and EDA

- Compute summary statistics (mean, median, standard deviation) for all key variables.
- Produce correlation matrices to explore relationships among:
  - Data center intensity measures.
  - Electricity use and prices.
  - Emissions and air quality.
  - Water stress and withdrawals.
  - Income, unemployment, housing variables.
- Identify outliers and interesting cases (e.g. states with very high data center intensity and high water stress).

4.2 Step 3.2: Panel Regression Models

Model Family 1: Data Centers, Electricity Use, and Prices

Specification For state s and year t:

$$\begin{aligned}\text{elec\_sales\_pc\_mwh}_{st} = \alpha_s + \gamma_t + \beta_1 \text{dc\_emp\_per\_10k\_pop}_{st} + \beta_2 \log(\text{median\_income}_{st}) + \beta_3 \text{info\_emp\_share}_{st} \\ + \beta_4 \text{unemployment\_rate}_{st} + \varepsilon_{st},\end{aligned}$$

where α_s are state fixed effects and γ_t are year fixed effects.

A similar model will be estimated for average retail price, $\text{price\_all\_cents\_kwh}_{st}$, as the dependent variable.

Interpretation

- β_1: association between data center employment intensity and electricity use or prices per capita, controlling for income, digital sector size, and unemployment.
- Hypothesis: $\beta_1 > 0$ for both electricity use and prices.

Model Family 2: Emissions and Air Quality

Specification

$$\begin{aligned}\text{co2\_power\_pc\_tons}_{st} = \alpha_s + \gamma_t + \delta_1 \text{dc\_emp\_per\_10k\_pop}_{st} + \delta_2 \text{elec\_sales\_pc\_mwh}_{st} \\ + \delta_3 \text{baseline\_water\_stress}_s + \delta_4 \big(\text{dc\_emp\_per\_10k\_pop}_{st} \times \text{baseline\_water\_stress}_s\big) + u_{st}.\end{aligned}$$

Interpretation

- δ_1: direct association between data center intensity and power-sector CO_2 per capita.
- δ_4: whether this association is stronger in high water-stress states (interaction).
- Hypotheses: $\delta_1 > 0$ and $\delta_4 > 0$.

Similarly, we can estimate models with air quality measures (e.g. $PM_{2.5}$) as outcomes.

Model Family 3: Housing and Community Outcomes

Specification

$$\begin{aligned}\text{median\_gross\_rent}_{st} = \alpha_s + \gamma_t + \theta_1 \text{dc\_emp\_per\_10k\_pop}_{st} + \theta_2 \log(\text{median\_income}_{st}) \\ + \theta_3 \text{unemployment\_rate}_{st} + v_{st}.\end{aligned}$$

A parallel model will use `median_home_value` as the dependent variable.

Interpretation

- θ_1: association between data center intensity and housing costs, conditional on income and unemployment.
- Hypothesis: $\theta_1 > 0$ in data center hub states.

Robustness and Diagnostics

- Check multicollinearity among predictors (e.g. variance inflation factors).
- Test alternative measures of data center intensity (e.g. employment share vs. facilities count).
- Consider simple lagged models (data center intensity at $t-1$ predicting outcomes at t).

4.3 Step 3.3: Clustering and Typology

- Construct a state-level feature vector (using mean or final-year values) with:
  - `dc_emp_per_10k_pop`, `dc_facilities_count`
  - `elec_sales_pc_mwh`, `price_all_cents_kwh`
  - `co2_power_pc_tons`, `baseline_water_stress`
  - `median_gross_rent`, `median_income`
- Standardize these variables to have mean zero and unit variance.
- Apply k-means clustering (e.g. $k = 3$ or $k = 4$) and compare inertia and silhouette scores to choose k.
- Interpret clusters with descriptive tables and maps, e.g.:
  - Cluster A: "data center hubs under stress"
  - Cluster B: "energy-intensive, low-digital"
  - Cluster C: "low data center, moderate everything"
  - Cluster D: "rich digital states with high housing costs"

5 Phase 4: Visualization and Reporting

5.1 Step 4.1: Tableau Dashboards

Design and implement a set of Tableau dashboards corresponding to the main themes:

Dashboard 1: Geography of Data Centers

- Choropleth map: data center employment per 10,000 residents by state and year.
- Bar chart: top 10 states by data center intensity for a selected year.
- Interactive filters: year slider, ability to highlight specific states.

Dashboard 2: Energy and Emissions

- Line chart: per-capita electricity sales for selected high- and low-intensity states.
- Scatterplot: data center intensity vs. electricity use/prices, with trend lines.
- Scatterplot: data center intensity vs. CO_2 per capita.

Dashboard 3: Water Stress and Air Quality

- Map with dual encoding: baseline water stress and data center intensity.
- Scatter with color by water stress category, showing interaction patterns with emissions or $PM_{2.5}$.

Dashboard 4: Housing and Clusters

- Scatterplot: data center intensity vs. rent or home values (and their growth).
- Cluster map: states colored by cluster membership.
- Tooltip: cluster mean characteristics (energy, emissions, housing).

5.2 Step 4.2: Interpretation and Narrative

- Synthesize regression and clustering results into a narrative:
  - Where are data centers expanding the fastest?
  - How is this linked to energy use, prices, and emissions?
  - Which states face "double burdens" (e.g. high data center intensity and high water stress)?
  - How do housing markets respond in data center hubs?
- Connect findings to SDG 7 and SDG 13, and discuss implications for:
  - Energy and climate policy.
  - Water-resource planning.
  - Land use and housing policy.
  - Environmental justice and equity.

5.3 Step 4.3: Limitations and Future Work

- Discuss limitations:
  - NAICS 518210 is a proxy for data centers and includes other data-processing services.
  - State-level data may mask within-state spatial inequalities.
  - Observational design limits causal claims.
  - Water-use data are infrequent; water stress indices are static.
- Suggest extensions:
  - County- or metro-level analysis.
  - Incorporation of detailed tax incentive and land-use data.
  - Micro-level analysis of employment types and wage distribution.

6 Indicative Timeline

Assuming a semester-scale project, an example timeline:

- Weeks 1–2: Finalize conceptual design and data source list.
- Weeks 3–5: Acquire and clean all datasets; construct the state–year panel.

- Weeks 6–8: Conduct descriptive analysis and initial regression models.
- Weeks 9–10: Refine models, run robustness checks, and perform clustering.
- Weeks 11–12: Build Tableau dashboards and draft narrative.
- Weeks 13–14: Finalize report, presentation, and visualizations.

7 Summary

This implementation plan provides a step-by-step roadmap for building a rich, integrated dataset on data centers and social impacts, applying robust social data science methods, and delivering clear visual and written outputs. Following this plan will ensure that the project:

- Meets the requirement for more than 300 instances and 15 attributes.
- Uses openly available datasets from credible official sources.
- Produces both quantitative results and compelling visual storytelling.
- Addresses an important, timely topic at the intersection of digital infrastructure, energy, environment, and communities.