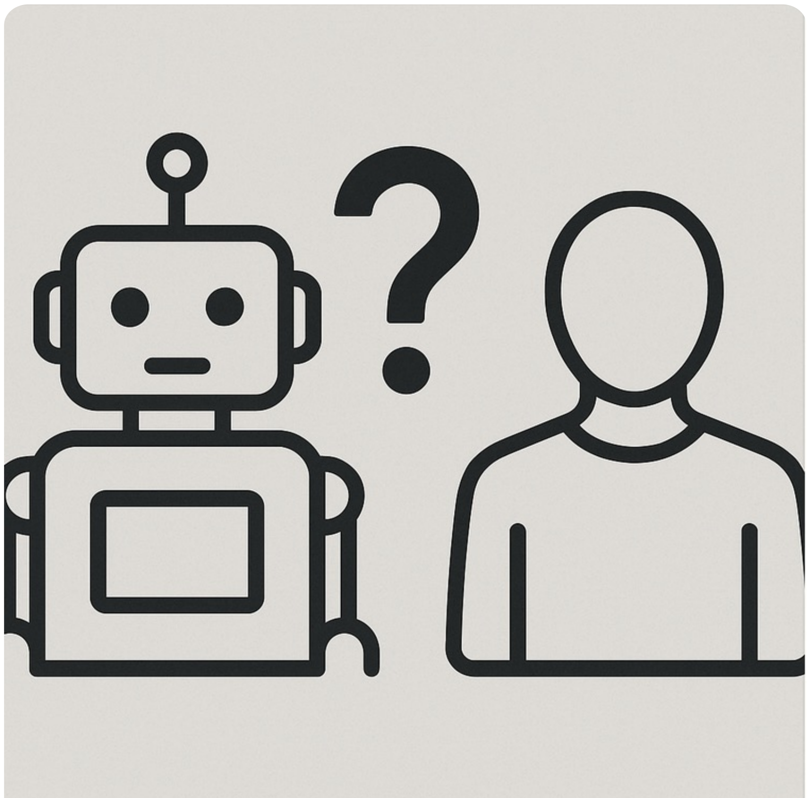
HDSI | Harvard Data Science Initiative

AI: Solution or Obstacle for Climate Action?

- Francesca Dominici, PhD
- Professor of Biostatistics, Population Health and Data Science
- Harvard T.H. Chan School of Public Health
- Director of the Harvard Data Science Initiative

Outline



- The trillion-dollar, highly political scientific questions
- AI as a promising solution
- My journey
- Geo-AI for Air Pollution Exposure Estimation
- The first foundation model for climate adaptation
- Responsible AI
- The Environmental Impact of AI
- Conclusions

AI as Eutopia or Dystopia?

The trillion-dollar, highly political scientific questions

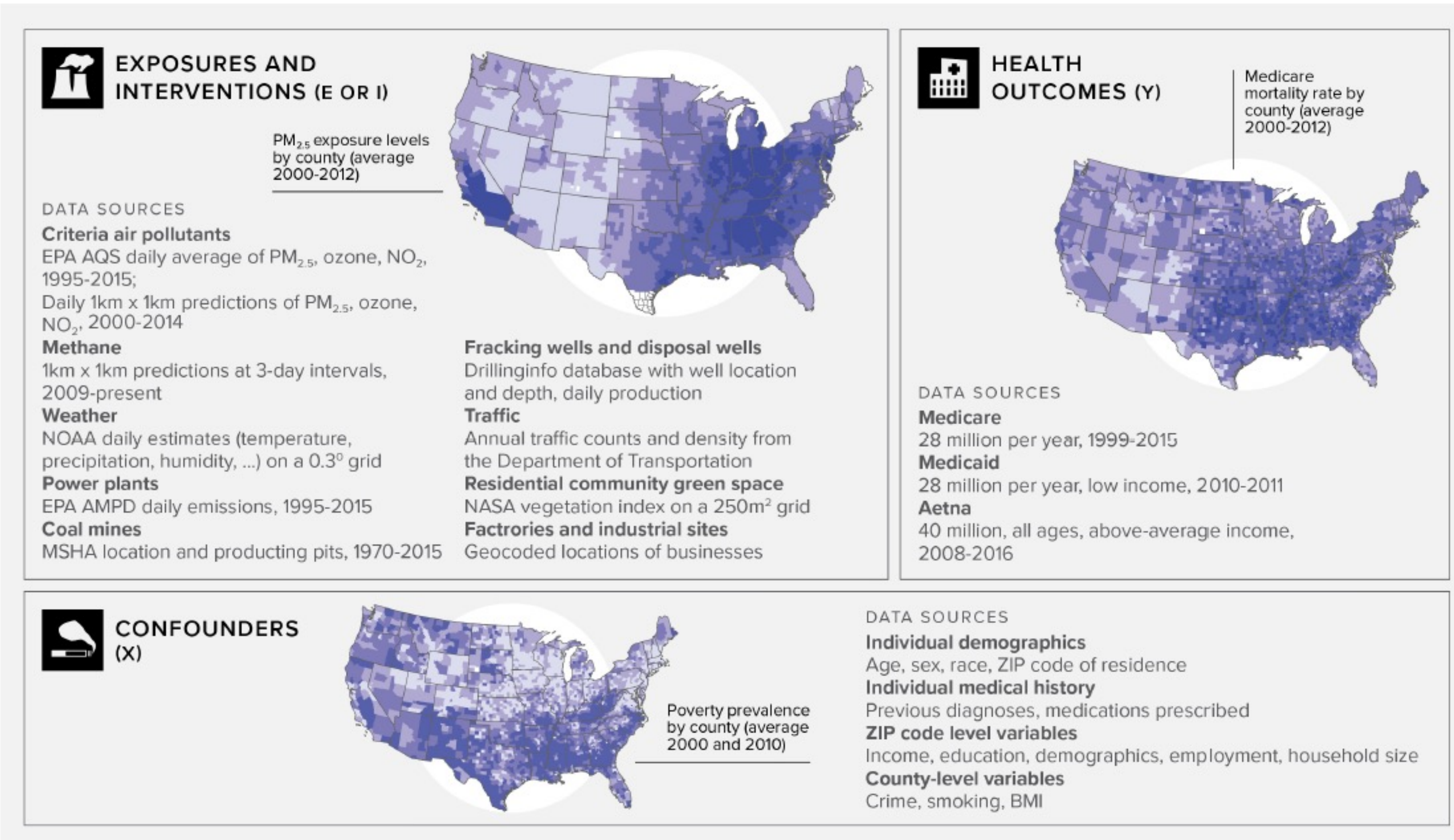
- Does exposure to fine particulate matter, even at low levels, cause an increase in hospitalizations?
- Is air pollution from coal-fired power plants more toxic than air pollution from other sources?

DATA

Data integration of over 20 government data repositories

- All Medicare participants (n=67,682,479) in the continental United States from 2000 to 2021
- Outcomes: all-cause mortality and cause specific hospitalization
- Individual level information: date of death, age of entry, year of entry, sex, race, whether eligible for Medicaid (proxy for SES)
- Zip code of residence and other covariates

Causal Reasoning AI for Policy Decisions (9 TB of data)



Seing

Doing

Imagining

The NEW ENGLAND JOURNAL of MEDICINE

Air Pollution and Mortality at the Intersection of Race and Social Class

Kevin P. Josey, Ph.D., Scott W. Delaney, Sc.D., J.D., Xiao Wu, Ph.D., Rachel C. Nethery, Ph.D., Priyanka DeSouza, Ph.D., Danielle Braun, Ph.D., and Francesca Dominici, Ph.D.

Table 1. Characteristics of the Medicare Cohort, 2000 through 2016.*

Characteristic	Full Cohort†	Black Persons		White Persons	
		Higher Income‡	Low Income§	Higher Income‡	Low Income§
Persons — no. (% of full cohort)	73,129,782 (100)	4,872,714 (6.7)	1,671,776 (2.3)	56,422,414 (77.2)	4,989,457 (6.8)
Person-yr — no. (% of total person-yr)	623,042,512 (100)	37,862,780 (6.1)	14,886,928 (2.4)	483,479,863 (77.6)	48,247,908 (7.7)
Deaths — no. (% of total deaths)	29,467,648 (100)	1,488,555 (5.1)	1,154,227 (3.9)	20,773,208 (70.5)	4,769,240 (16.2)
Median follow-up time — yr	8.0	7.0	8.0	8.0	8.0
Age at entry — %					
65–74 yr	80.6	86.2	77.4	80.4	72.7
75–84 yr	14.8	10.7	15.6	15.3	17.2
85–94 yr	4.2	2.5	6.2	4.0	9.0
≥95 yr	0.4	0.6	0.8	0.3	1.1
Female sex — %	55.4	54.9	68.1	54.3	68.0
Medicaid eligible — %	11.6	0	100	0	100

Lowering exposure from 12 to 9 unit → 5 % mortality reduction among Black Americans; 2.5% mortality reduction among White Americans

NEW RESEARCH

Cleaner Air Helps Everyone. It Helps Black Communities a Lot.

A new study quantified the benefits of pollution reduction in terms of race and class.

Share full article

St. James, La., one of several Mississippi River towns dotted by chemical plants and oil refineries. William Widmer for The New York Times

The NEW ENGLAND JOURNAL of MEDICINE

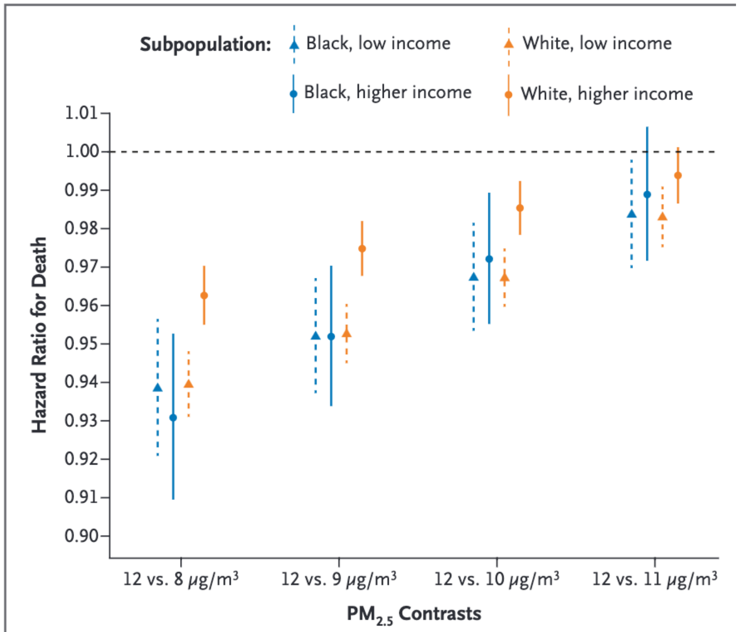


Figure 4. Differences in Mortality with Decreasing $PM_{2.5}$ Exposure among Marginalized Subpopulations.

Shown are point estimates and 95% confidence intervals of the hazard ratio for death comparing different levels of annual average $PM_{2.5}$ exposure (12 μg per cubic meter vs. 11, 10, 9, or 8 μg per cubic meter) on average for subpopulations defined in selected ways. Low income was defined as dual eligibility for both Medicare and Medicaid. Confidence intervals were not adjusted for multiplicity; therefore, they should not be used in place of hypothesis testing.

Doing

How my lab has impacted this decision

Data Science

April 2023

The NEW ENGLAND JOURNAL of MEDICINE

SPECIAL ARTICLE | VOL. 388 NO. 15, APR 13, 2023

Air Pollution and Mortality at the Intersection of Race and Social Class

K.P. Josey and Others | N Engl J Med 2023; 388:1396-1404

In this large study, the mortality benefits of reducing levels of fine particulate matter air pollution were greater for low-income and higher-income Black persons and for low-income White persons than for higher-income White persons.

Nov 2023

Science

RESEARCH ARTICLE

POLLUTION

Mortality risk from United States coal electricity generation

Policy

Dec 2023, **COP28 Opening day, Kerry talk about my study: Climate Crisis is a Health Crisis.**

A lot of seeing and doing

Impact

February 2024

Biden Administration Moves to Tighten Limits on Deadly Air Pollution

A new rule would, for the first time in a decade, reduce emissions of soot that disproportionately harm communities of color.

Give this article 95

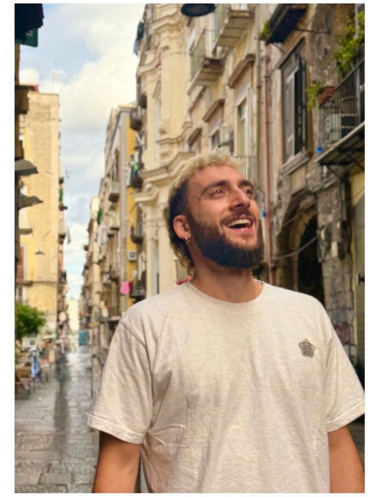
- ✓ Cleaner Air
- ✓ Lives Saved
- ✓ Less GHG

The potential of AI for Climate Adaptation

- Climate change brings more extreme weather, wildfires, and shifting disease patterns
  - Understanding and mitigating health impacts is *complicated* – e.g. heatwaves affecting vulnerable people, wildfire smoke causing respiratory illnesses
- **AI's Promise: AI can analyze *unprecedentedly massive multimodal* data to find *generalizable* patterns and make predictions *more accurately* than traditional methods**
  - **This can inform early warnings and adaptive responses (e.g. alerting hospitals of an incoming heat-related patient surge)**

ClimaCare: A Foundation Model for Healthy Climate Adaptation

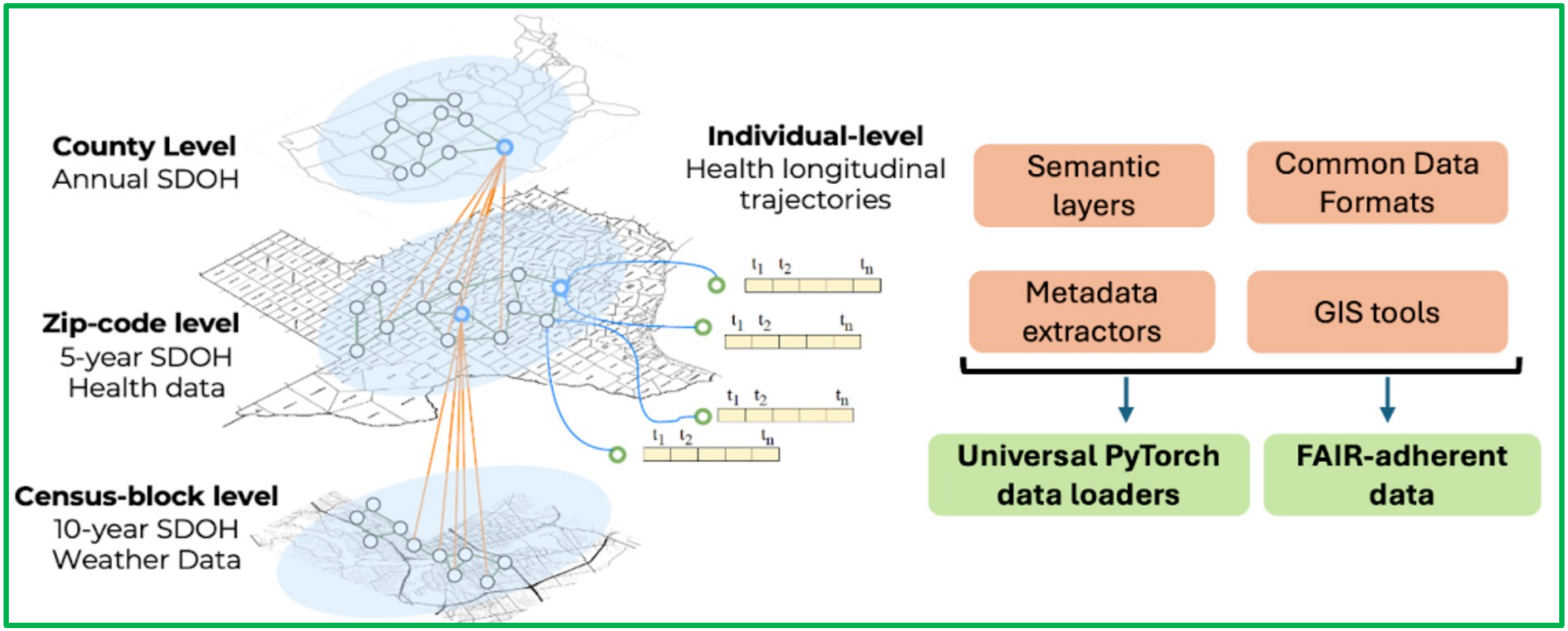
Claudio Battiloro[*,†], James Kitch[*,†], Bret Nestor[*,†], Mauricio Tec[†], Michelle Audirac, Danielle Braun, Francesca Dominici



Claudio Battiloro

1. Pre-trained on the entire US health care system x environmental data x societal data
2. It produces unified embeddings that capture the complex spatiotemporal relationships between climate stressors, socioeconomic variables, and health outcomes.
3. We evaluate the model on benchmark downstream tasks, i.e., health outcomes interpolation, extrapolation, downscaling, and forecasting
4. We implement "what-if" scenario forecasting for climate adaptation using synthetic ground-truth data to validate counterfactual predictions when any input exposure is altered.

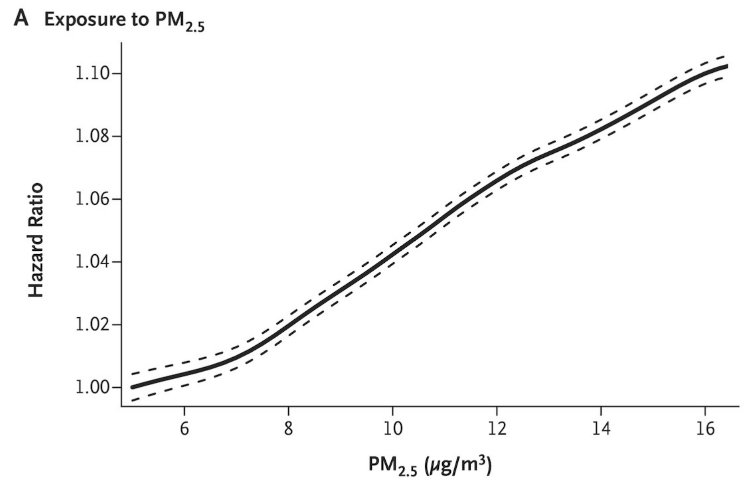
Towards a One-of-a-Kind geo-AI for Healthy Climate Adaptation

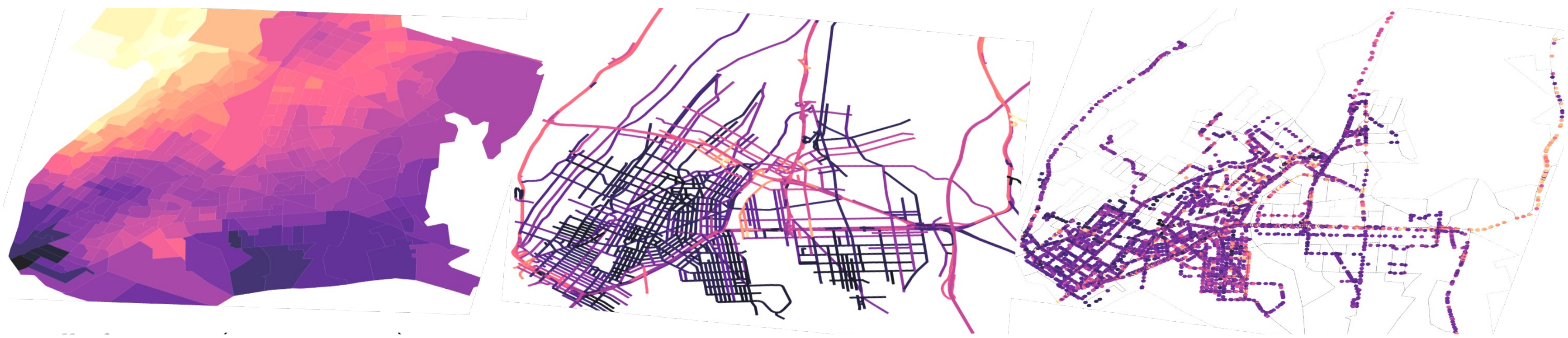


ClimaCare: Downstream Tasks

- **Spatiotemporal Downstream Tasks:**
  - Spatial Interpolation
  - Spatial Extrapolation
  - Forecasting

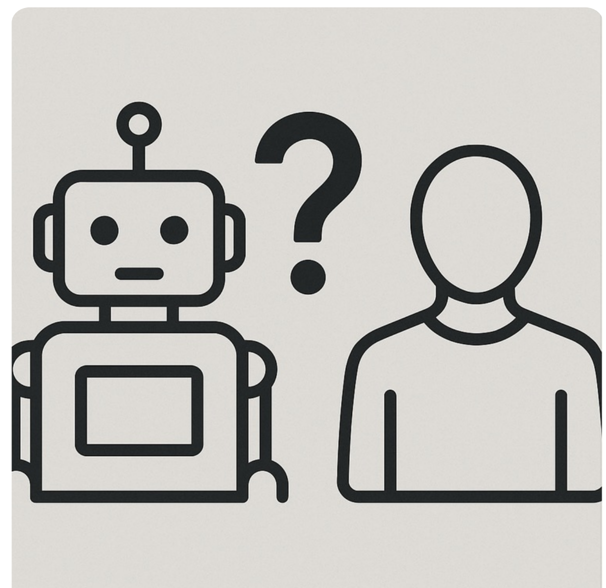
What-If Downstream Tasks:
ERC Estimation
Enhanced Causal Inference





2050: Let’s imagine a world where we have solved causal reasoning

- AI agents are everywhere
- Is this a better future?




At Amazon's Biggest Data Center, Everything Is Supersized for A.I.
On 1,200 acres of cornfield in Indiana, Amazon is building one of the largest computers ever for work with Anthropic, an artificial intelligence start-up.

403 Hyperscale data centers and 3318 energy supplier power plants in the US (May 2024 to April 2025)

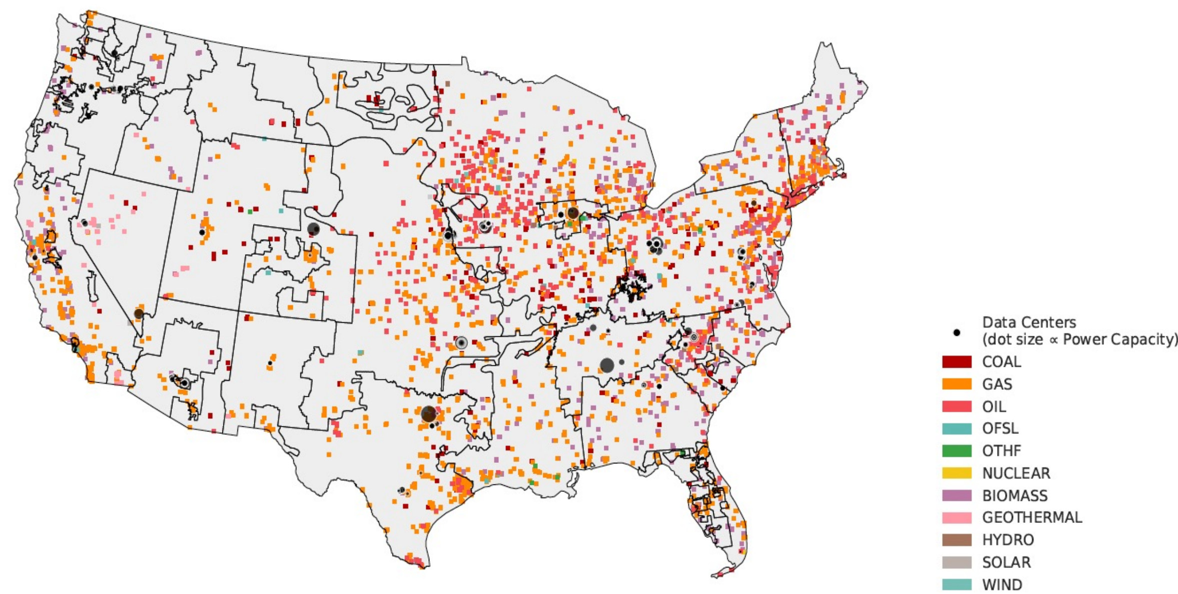


Fig. 1. Geographic distribution of hyperscale data centers and power plants in the contiguous US, overlaid with balancing authority regions. This figure shows the 403 hyperscale data centers and 3,318 operational power plants included in our analysis for the study period from May 2024 to April 2025. The map is displayed at the balancing authority (BA) level, representing regions where electricity supply and demand are managed in real time. The size of each hyperscale data center marker is proportional to its power capacity, while power plants are colored by their primary fuel type.

Scientific questions

1. What are the electricity consumption, sources, and attributable CO_2 emissions of those 403 data centers?

2. What is the fuel mix of the power plants supplying electricity to data centers?

3. Which states have the highest CO_2 emissions attributable to data centers?

Hint: With a data pipeline that can answer those questions, we make informed decisions, such as: Where should I place a data center? Where should I intervene on the power grid? How can we decarbonize this sector?

Materials and Methods

01

Dataset Compilation and Validation

Our analysis began with compiling and validating a dataset of 403 HDCs.

This was achieved using private data providers, web scraping, and satellite imagery.

02

Power Capacity Estimation

Missing power capacities were estimated using a Gradient Boosted Regression Tree model.

03

Electricity Consumption Calculation

Annual electricity consumption was calculated by multiplying the power capacity, hours per year, and a utilization rate of 66.3%.

The utilization rate was determined empirically.

04

CO_2 Emissions Estimation

Each HDC was assigned to its balancing authority and corresponding power plants.

CO_2 emissions were estimated using an energy generation-weighted model and EPA emission factors.

05

Carbon Intensity Computation

Carbon intensity was computed as grams of CO_2 emitted per kilowatt-hour consumed.

This computation was performed at various geographic levels.

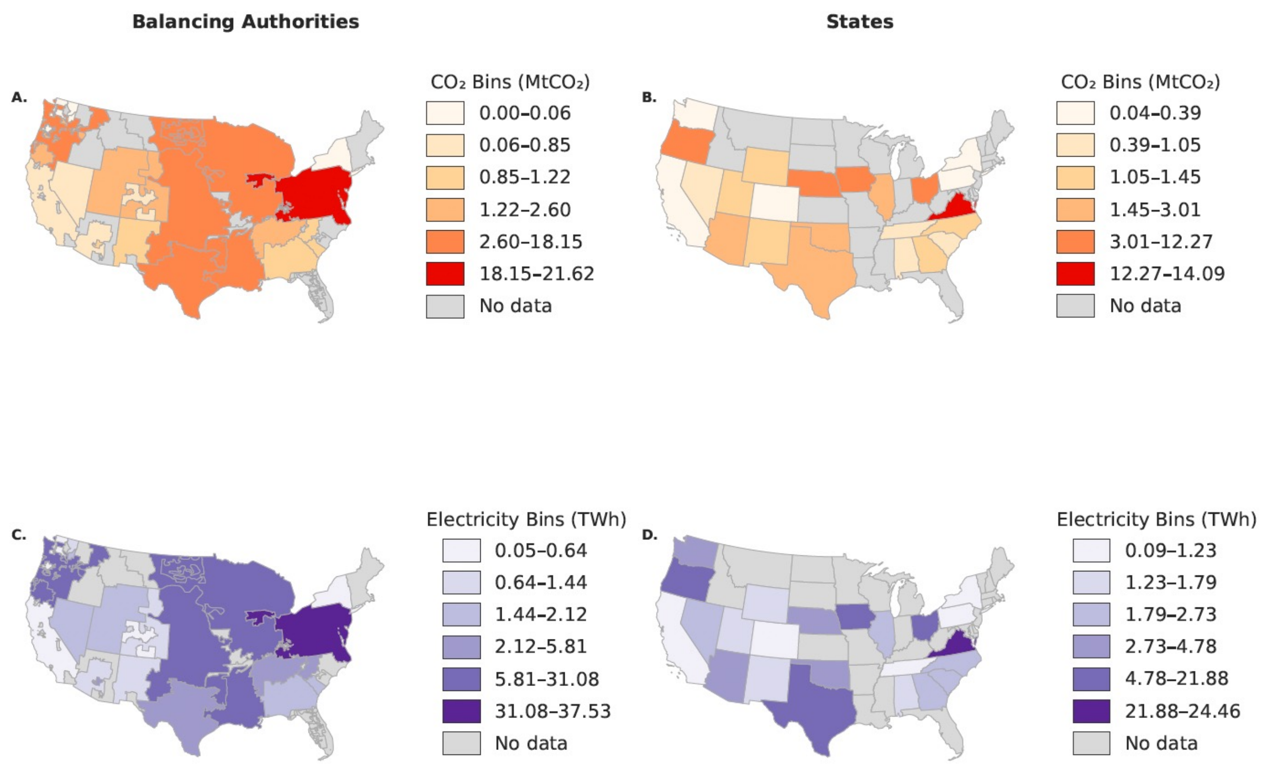


Fig. 2. Hyperscale Data Center electricity consumption and CO_2 emissions. (Left column, A and C) The balancing authority (BA) region in which a hyperscale data center is located determines the mix of power plants that supply its electricity and thus its attributable emissions. See fig.S.4.1 for BA regions and corresponding names. (Right column, B and D) Maps at the state level show electricity consumption and emissions for which the hyperscale data centers within the state are responsible for. Color bins represent percentile-based ranges: 0–20%, 20–40%, 40–60%, 60–80%, 80–99%, and 99–100%.

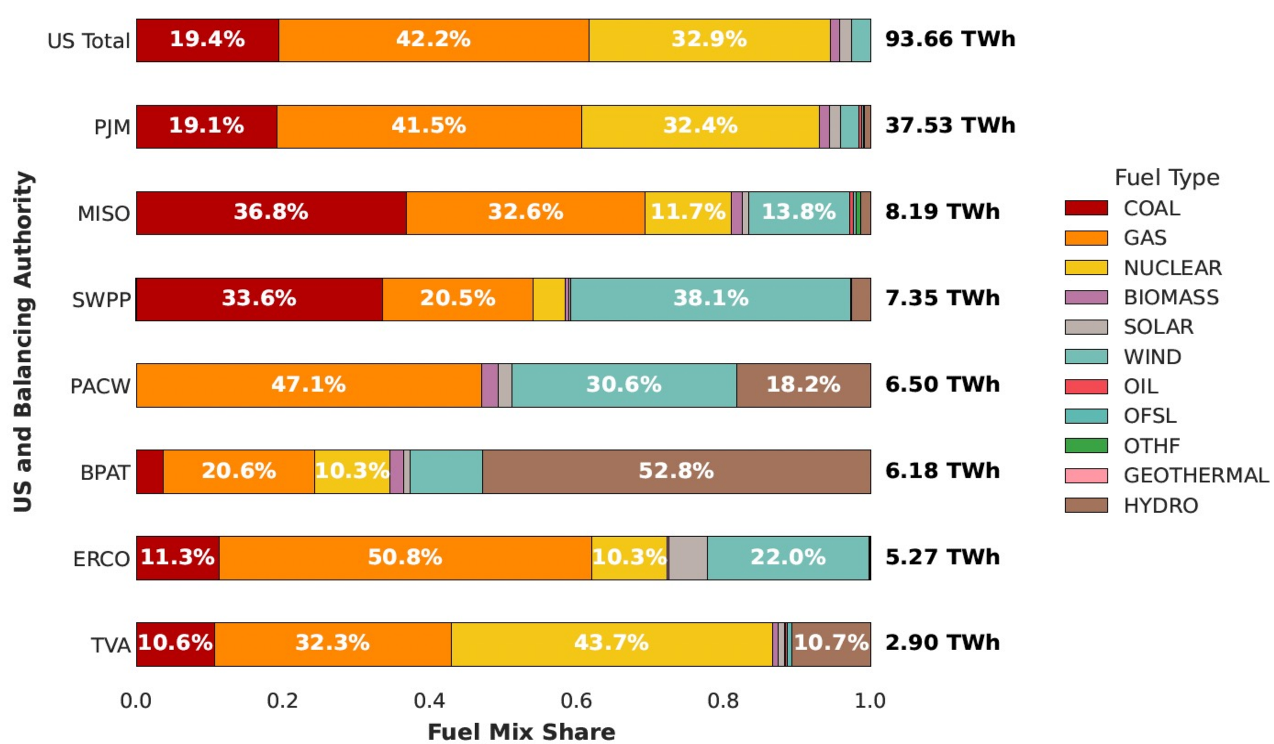


Fig. 4. Fuel mix of power plants supplying electricity for hyperscale US data centers. The top bar represents the distribution of fuel types used by the power plants supplying electricity for hyperscale US data centers in our study. The bottom bars show the largest balancing authorities ranked by aggregated power capacity of hyperscale data centers (shown on the vertical axis), and the amount of electricity produced per fuel type. See fig.S.4.1 for BA regions and corresponding names.

Carbon Emissions Attributable to Hyperscale Data Centers

Total CO2 Emissions from HDCs

52.69M

The total CO2 emissions attributable to the 403 hyperscale data centers (HDCs) amounted to 52.69 million metric tons.

Proportion of US Carbon Emissions

1.10%

This represents approximately 1.10% of the total US carbon emissions from electricity consumption in 2023.

Increase Since 2018

5x

This is more than five times the total emissions reported for HDCs in 2018.

Highest Emissions by State

24.46M

Virginia had the highest CO2 emissions attributable to HDCs, amounting to 24.46 million metric tons.

Significant State Contributions

5.82M

Ohio followed with 5.82 million metric tons of CO2 emissions attributable to HDCs.

- 52.69 M represents the annual CO_2 emissions of a major U.S. city or a sizable portion of the U.S. aviation industry.

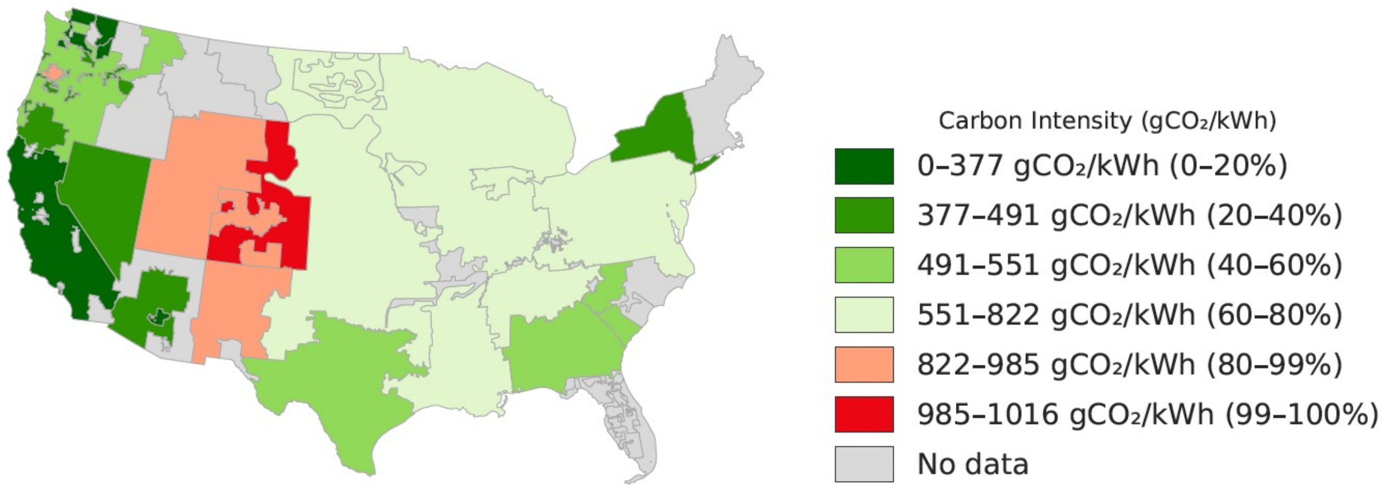


Fig. 3. Carbon intensities of electricity consumption for hyperscale US data centers by balancing authority. Carbon intensity is defined as the amount of carbon dioxide emissions produced per unit of electricity generated, or consumed, and is expressed in units such as grams of CO_2 per kilowatt-hour (gCO_2/kWh) for electricity generation. The figure shows HDCs' carbon intensity for electricity consumption at the balancing authority level, in grams of CO_2 per kWh. Color bins represent percentile-based ranges: 0–20%, 20–40%, 40–60%, 60–80%, 80–99%, and 99–100%.

AI as Eutopia or Dystopia?

The EPA's 2024 regulatory analysis projects that new standards for coal (and some new gas) power plants will cut about 55 million metric tons per year

The total CO_2 emissions attributable to the 403 hyperscale data centers (HDCs) amounted to 52.69 million metric tons.

.