A PERSONAL OPEN-DATA PROJECT · KYUMIN KIM

# Where should the next data center go?

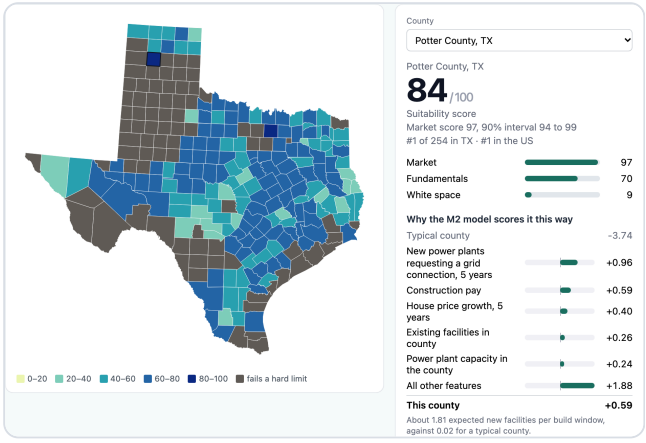
A county-level look at data center siting, built only from public data. The project learns from where data centers were actually built, combines that with the conditions that make a site work (power, land, hazards, water, labor, policy and local acceptance), and shows the result on an interactive map with the reasons behind every score.

**3,108**
counties scored, contiguous US

**40+**
public datasets combined

**~80**
conditions measured per county

**2**
segments, scored separately




The interactive map: Texas, hyperscale segment. Each county is coloured by its score; grey counties fail a deal-breaker. The card explains one county and the conditions that pushed its score up or down.

## Why I built this

Following the news about the data center boom, I kept wondering how companies decide where to build. Power and connectivity come first, but a site also has to work for the people around it: labor markets, land and housing costs, natural hazards, water, local politics, and whether a community welcomes a new facility or pushes back with a moratorium.

I could not find an open, transparent way to look at all of this together, so I set out to see how far a light project could get with machine learning and openly available data alone.

## Two segments, because they want different places

### A · Hyperscale and AI

Campuses of 50 MW or more, or any site run by a hyperscaler (Amazon, Google, Microsoft, Meta, Apple, Oracle). Driven by power availability, land and grid capacity.

### B · Colocation and enterprise

Smaller, multi-tenant or corporate facilities. Driven by customers, fibre, skilled labour and metro access.

## What it suggests

### Open data goes far

Trained only on 2016 to 2020 builds, the model placed 45 to 50% of the facilities built in 2021 to 2026 in its top 10% of counties.

### Power dominates

New generation waiting to connect and nearby plant capacity are the strongest signals, ahead of land or labor costs.

### White space exists

Some counties have strong fundamentals but few facilities so far: places the market has not reached yet.

### Local acceptance is the gap

Moratoria and community opposition shape real projects, but public data on them is thin.

HOW IT WORKS

# From public data to a score you can question

**1 · Collect**
40+ public sources, each cached and versioned with its licence.

**2 · Measure**
About 80 measures per county, for 2015, 2020 and today.

**3 · Learn**
A model learns which conditions came before new builds.

**4 · Score**
Four 0 to 100 scores and deal-breaker checks.

**5 · Explain**
Map, rankings and a county card showing why.

## What data goes in

| THEME | WHAT IS MEASURED | MAIN SOURCES |
|---|---|---|
| **Facilities** | Location, size (MW), operator, status and year of known data centers and projects | Compute Atlas, OpenStreetMap |
| **Power** | Industrial price, outages, local generation, distance to power lines and substations, projects waiting to connect to the grid | EIA-860 and 861, LBNL Queued Up, OpenStreetMap power grid, NREL |
| **Grid access today** | Expected years to get power, how crowded the grid is, special rates for very large users | LBNL queue (38,000 requests), utility tariff filings, OpenEI URDB |
| **Hazards** | Natural hazard risk, flood zones, earthquakes, wildfire, landslides, sinkhole-prone ground | FEMA NRI and flood maps, USGS, USFS |
| **Economy and land** | Population, income, IT and construction labour and wages, GDP, home prices, rents, farmland value | Census ACS, BLS, BEA, FHFA, USDA |
| **Climate and water** | How much cooling the climate needs, summer heat, water use and water scarcity | NOAA, USGS, WRI Aqueduct |
| **Connectivity and policy** | Fibre coverage, roads, airports, air quality status, sales tax breaks, local bans on new projects | FCC, Census TIGER, FAA, EPA, state statutes |

## Methods

The core prediction is **how many new data centers a county is likely to attract in the next five or so years**, for each segment. Count models (Poisson objective) learned from two real build waves: what was built in 2016 to 2020 given conditions in 2015, and in 2021 to 2026 given conditions in 2020. I compared ten model families (penalized regressions, random forest, XGBoost, CatBoost, a neural net and LightGBM); LightGBM, tuned with Optuna, came out best.

Every score comes with its reasons. SHAP values show which conditions pushed a county up or down, and the market score carries a 90% interval.

### Four scores, 0 to 100

- **Market**: where the model expects new builds next.
- **Fundamentals**: how good the place is on its conditions alone. Users can change how much each condition counts; the starting values come from the model.
- **Suitability**: a blend of the two, with a slider for the mix.
- **White space**: good conditions, but few facilities built or planned nearby.

## Power and grid layer

### Time to power

A survival model of how long 38,000 past grid-connection requests actually took estimates how many years a project waits for power in each county, with a likely range.

### Grid congestion

Planned data center demand within 100 km, compared with nearby power plants, new supply likely to come online and the number of high-voltage lines.

### Rates for very large users

55 utility rate plans for big users, read from public filings: minimum size, contract length, minimum monthly bill and deposits. They feed a long-run power cost.

DOES IT WORK?

# Tested the way it would be used

THE KEY TEST: PREDICT THE FUTURE FROM THE PAST

I trained the model only on builds from 2016 to 2020, ranked every county using 2015 data, and then checked what was actually built in 2021 to 2026. Choosing counties at random, the top 10% would hold about 10% of new capacity.

**32%**
of new hyperscale and AI megawatts were in the model's top 10% of counties, about **3 times** random (45% of the facilities).

**57%**
of new colocation and enterprise megawatts were in its top 10%, about **6 times** random (50% of the facilities).

### Checks behind the numbers

- **Tested on unseen regions.** Whole regions (200 km blocks) are held back while the model learns, then used to test it (spatial cross-validation).
- **Fair model choice.** The four strongest methods (LightGBM, XGBoost, random forest, CatBoost) were each tuned with the same effort and compared on the same test. LightGBM came first every time.
- **Measured progress.** Since the first version, prediction error on unseen regions fell 17% (hyperscale) and 7% (colocation).
- **Every decision written down.** Each modelling choice, default and known weakness is recorded with its reason.

### Design choices

- **Deal-breakers** grey out counties instead of hiding them, with the reason: no high-voltage line (230 kV) within 25 km, strong earthquake risk, 30% or more in flood zones, severe water shortage, a ban on new projects, and, optionally, a long wait for power.
- **Responsible use.** Race and ethnicity are never model inputs; they are used only to check for bias. Community opposition is shown as concerns to address, never as a way to find "least resistant" places.
- **Open data.** Every source is public, and each one's licence is recorded.

## Limits

- **County level, not parcel level.** It narrows the search to regions; it does not pick sites.
- **Public data only.** Commercial facility databases would sharpen the colocation picture.
- **Estimates where data is thin.** There is no public data on how long data centers wait for power, so time to power is estimated from power plants' waits instead. Most rate terms were read from filings and are marked low confidence.
- **Ranks, not forecasts of totals.** Scores say where activity is likelier, not how many megawatts the country will build.

## Explore it

### Interactive map

A lightweight version runs in the browser at **www.kyuminkim.com/projects/dc-site**.

### Fuller version

A fuller version with more detail (cost calculator, county comparison, facility map) is available on request.

### Get in touch

Questions, ideas or collaboration: **kyumin.kim.econ@gmail.com**

A personal project by Kyumin Kim, built with Lorenzo Swank · November 2026. Facilities from Compute Atlas (CC BY 4.0) and OpenStreetMap contributors (ODbL; used as model input only). Interconnection data from LBNL Queued Up and GridTracker (CC BY 4.0). Scores are a screening aid, not investment advice.