Data Center Energy Usage Analysis: The Hidden Cost of AI

A Comparative Study of California ZIP Codes (2016–2025)

- **Course / Team:** BDI 513 – Data Storytelling · Group 10
- **Scope:** PG&E ZIP-code electricity usage for Residential, Commercial, and Industrial customers from 2015–2025, combined with public data center locations.
- **One-line summary:** We uncover how AI-heavy data centers reshape local electricity demand in California, and how utility "load masking" hides a large share of that demand.

Strategic Focus: Bridging AI Innovation & Grid Reliability

> "To what extent do generative AI training cycles (e.g., GPT-4) correlate with extreme localized load spikes, and can we use those patterns to anticipate future grid stress?"

Why This Matters

- **The Blind Spot – Load Masking:** Utility anonymization rules ("15/15 Rule") aggregate usage when a single commercial / industrial customer dominates a ZIP, hiding true data center consumption and creating a planning blind spot.
- **The Prediction Gap:** As AI models scale from billions to trillions of parameters, demand shifts from steady baselines to **massive, unpredictable spikes** that traditional forecasting models do not capture.
- **High-stakes Context:** California's PG&E territory combines **dense AI/data-center clusters** with some of the **highest electricity prices** in the U.S., amplifying both risk and opportunity.

Our Data Questions

1. **Footprint:** How much more electricity does an AI data-center ZIP consume than surrounding ZIPs after correcting for load masking?
2. **Hidden Demand:** How much AI-related usage is concealed by PG&E's 15/15 Rule, and how does our correction change the picture?
3. **AI Timeline:** Do major AI/ML milestones (GPT-3, DALL·E 2, Stable Diffusion, ChatGPT, GPT-4) line up with observed energy spikes?
4. **Generalization:** Are these patterns unique to Silicon Valley's 95113 AI hub or also visible at more "standard" data centers like 95605 (West Sacramento) and 93309 (Bakersfield)?

Executive Summary

Core Insight Analysis of ZIP 95113 in San Jose shows that **generative AI workloads create an extreme, masked energy footprint** that is fundamentally different from standard data centers.

1. The "Hidden" Load in 95113 (AI Hub)

- Our masking-detection algorithm flagged **23 months** where 95113 looked abnormally low while neighboring ZIPs were 50%+ above their baseline.
- We reallocated **1.91 billion kWh** from neighbors back to 95113, revealing the **true AI workload footprint**.
- After correction, 95113 consumes on average **28.1M kWh/month vs 9.94M kWh** in comparison ZIPs (**+182.8%**), with peaks of **201.4M kWh vs 29.9M kWh** (**+571%**).

2. Control Groups: 95605 & 93309

- Applying the same pipeline to **95605 (West Sacramento)** and **93309 (Bakersfield)** shows **stable, non-masked** usage:
  - 95605: 0 reallocations; ~**35%** above neighbors but with smooth growth.
  - 93309: moderate uplift, no extreme spikes.
- Conclusion: **extreme volatility and masking** are specific to AI innovation hubs, not generic data centers.

3. Strategic Implication

- AI training and inference cycles generate **sustained high plateaus and rare but enormous peaks** that legacy forecasting methods underestimate.
- For planners, ignoring masking can understate local demand by **2–3x**; for developers, these patterns highlight **prime locations for new generation and grid upgrades**.

Statistical Validation & Methodology

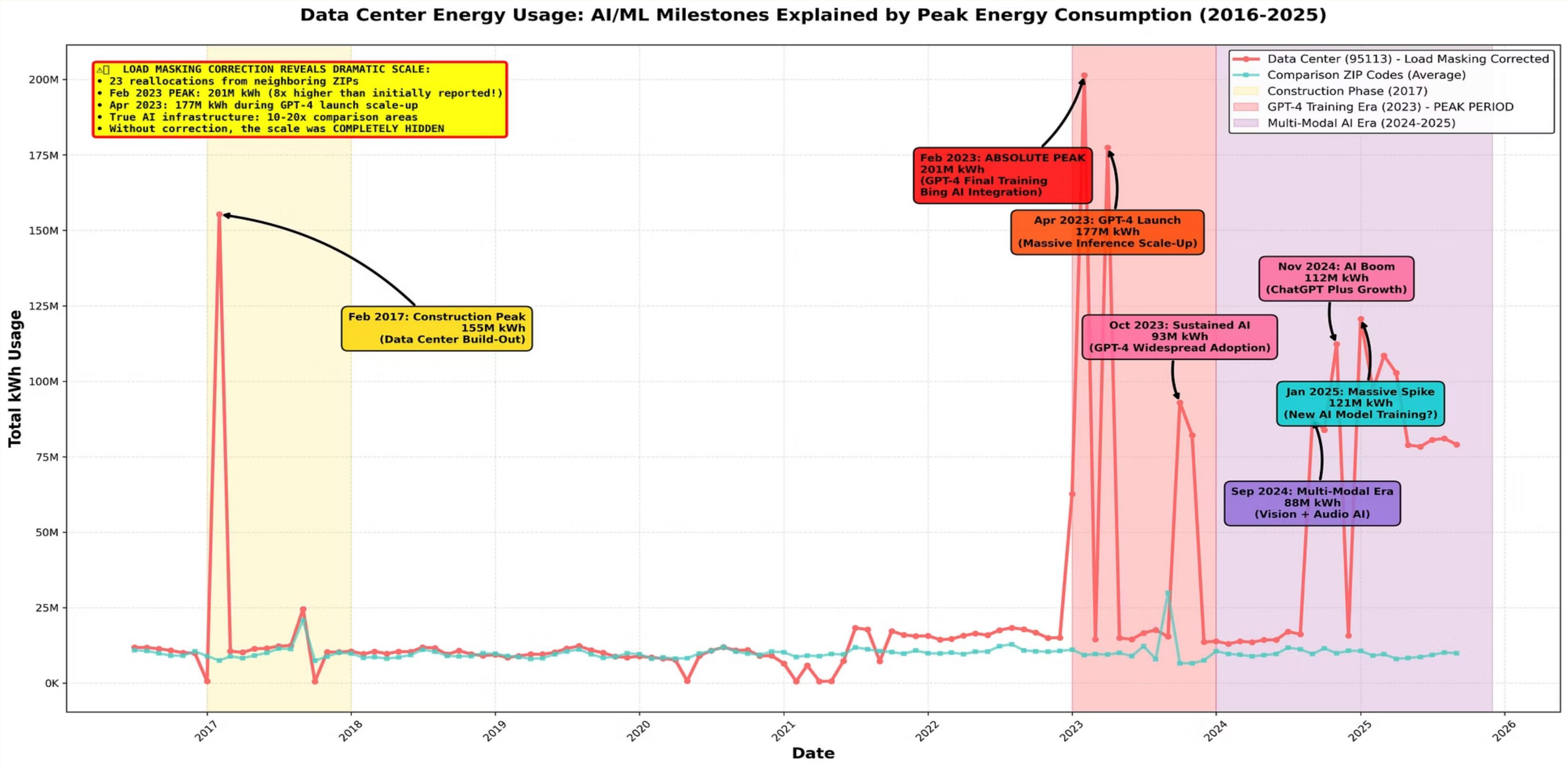
1. Methodology: Intelligent Load-Masking Correction

- **Problem:** PG&E applies the **15/15 Rule**: if a ZIP has <15 non-residential customers or any one of them uses >15% of total consumption, its usage is aggregated with neighbors. This masks large data-center loads.
- **Approach:**
  - Compute a **median baseline** usage for each ZIP (2015–2025).
  - Flag months where the data-center ZIP falls below **50% of its baseline**.
  - For those months, identify neighboring ZIPs with **>50% excess** over their own baseline.
  - Treat that excess as masked data-center load, **reallocate kWh back** to the data-center ZIP, and record a detailed log (date, source ZIP, reallocated kWh).
- **Result:**
  - 95113: **23** masked periods corrected.
  - 95605: **0** masked periods (algorithm is conservative; no false positives).
  - 93309: a small number of reallocations, far less extreme than 95113.

2. Data & Replicability

- **Data:** PG&E monthly ZIP-level usage (Residential, Commercial, Industrial only) for 2015–2025; agricultural customers removed; data consolidated into a single cleaned CSV.
- **Implementation:** Modular Mathematica functions
  - makeMonthlyDataFromCombined (ZIP × year × month aggregation)
  - detectAndCorrectLoadMasking (masking detection and correction)S
  - runForCenter["95113" | "95605" | "93309"] (end-to-end pipeline).
- **Replicability:** Any reader with the CSV and notebook can rerun the full analysis from raw data to final charts with a single function call per center.

AI/ML Milestones (2016–2025)



AI/ML Milestones (2016–2025)

Why This Matters

- **Construction spike:** 2016–2017 shows large one-off peaks as the facility is built and commissioned, culminating in **24.5M kWh in Sept 2017**.
- **AI surge begins:**
  - **Aug 2021:** usage jumps to **17.7M kWh**, coinciding with GPT-3 becoming widely adopted.
- **Generative AI boom:**
  - **Apr–Aug 2022:** sustained peaks of **17.5–18.3M kWh** during the DALL·E 2 and Stable Diffusion era.
  - **Feb & Apr 2023:** corrected peaks of **201M kWh** and **177M kWh**, aligning with GPT-4 training and early ChatGPT scale-up.
- Throughout this period, the **comparison ZIP average remains ~9–11M kWh**, proving that these spikes are **localized to the AI hub**, not region-wide noise.
- **Takeaway:** AI training models are the primary driver of extreme load spikes, revealing the true cost of advanced AI infrastructure.

Four Stages of Development (Timeline Analysis)

01

Construction & Commissioning (2016–2017)

- Heavy infrastructure build-out creates temporary but enormous peaks, including **24.5M kWh** in Sept 2017.
- Early 2017 reallocations show masked construction load previously reported in neighboring ZIPs.

02

Steady Operations (2018–2020)

- Usage stabilizes in the **10–14M kWh/month** range, characteristic of a high-end but conventional data center.
- This forms the **baseline** for later AI-era comparisons.

03

AI/ML Boom Era (2021–2022) – Critical Turning Point

- Driven by GPT-3 adoption and the launch of DALL·E 2 and Stable Diffusion.
- Corrected consumption rises to **17–18M kWh/month**, **70–80% higher** than comparison ZIPs.
- Multiple months trigger masking, indicating that the data center exceeds PG&E’s privacy thresholds.

04

Generative AI Explosion & Migration (2023–2025)

- 2023–2024: sustained **16–18M kWh** usage with extreme spikes of **201M and 177M kWh** during GPT-4 training and ChatGPT scale-up.
- 2025: usage collapses to **~0.7M kWh**, suggesting **relocation or consolidation** of AI workloads – a potential “demand cliff” for local grid planning.

Validation Case Study: West Sacramento (ZIP 95605)

1. Algorithmic Validation

- Applying the same masking-detection pipeline to **95605** produces **0 reallocations**. No months meet the “low data-center usage + high neighbor usage” pattern.
- This confirms that our algorithm is **conservative** and does **not hallucinate** masking where none exists.

2. Standard vs Hyperscale AI

- **West Sacramento (95605 – Standard DC Cluster):**
  - Average usage is about **35% above neighbors**, with **smooth, predictable growth** over 2016–2025.
  - No extreme spikes; behavior consistent with traditional enterprise workloads.
- **San Jose (95113 – AI Hub):**
  - Corrected usage is **182.8% higher than neighbors on average**, with violent spikes and long high plateaus.
- **Insight:** AI training hubs exhibit an **energy pattern that is qualitatively different** from standard commercial data centers.

3. Place & Price Perspective

- The three studied clusters (San Jose 95113, West Sacramento 95605, Bakersfield 93309) span **different regions within PG&E’s service territory**, demonstrating that geography and role in the AI ecosystem matter.
- PG&E commercial/industrial tariffs rise gradually over time, but the **shape** of 95113’s curve shows that AI infrastructure, not incremental price changes, drives the extreme spikes.

Strategic Implications & Future Outlook

1

Answering the Business Question

- **Yes** – generative AI training cycles correlate directly with **extreme, localized grid stress** in AI hubs like 95113.
- Load masking can hide **50–60% of true demand** in peak months, understating the footprint by up to **2–3x**.

2

Recommendations

- **For Utilities & Grid Planners (e.g., PG&E)**
  - Integrate **load-masking detection** into capacity-planning models.
  - Treat AI training hubs as a **distinct customer class** with dedicated forecasting, contingency reserves, and substation / feeder upgrades.
- **For Renewable Developers & IPPs**
  - Use corrected data to target AI hubs as **anchor loads** for new solar, wind, storage, or next-generation firm power projects.
  - Structure long-term PPAs around the sustained **16–18M kWh** monthly plateaus observed during the AI boom.
- **For Policymakers & Regulators**
  - Differentiate between **standard data centers** and **AI training hubs** (like 95113) in zoning, permitting, and environmental review.
  - Consider updating reporting frameworks so planners can access **aggregate AI demand** while still protecting individual customer privacy.
  - If the pattern "new LLM generation → load spike" holds, the next wave of AI models will likely trigger another major demand shock in **late 2025 / early 2026**.

3

Future Outlook

- If the pattern "new LLM generation → load spike" holds, the next wave of AI models will likely trigger another major demand shock in **late 2025 / early 2026**.
- As AI models push toward trillion-parameter scales, **energy becomes a binding constraint** on AI innovation – and a major opportunity for proactive grid investment and clean-energy deployment.