

From Persisting to Predicting

Pioneering Analytics for a Rapidly Growing Company

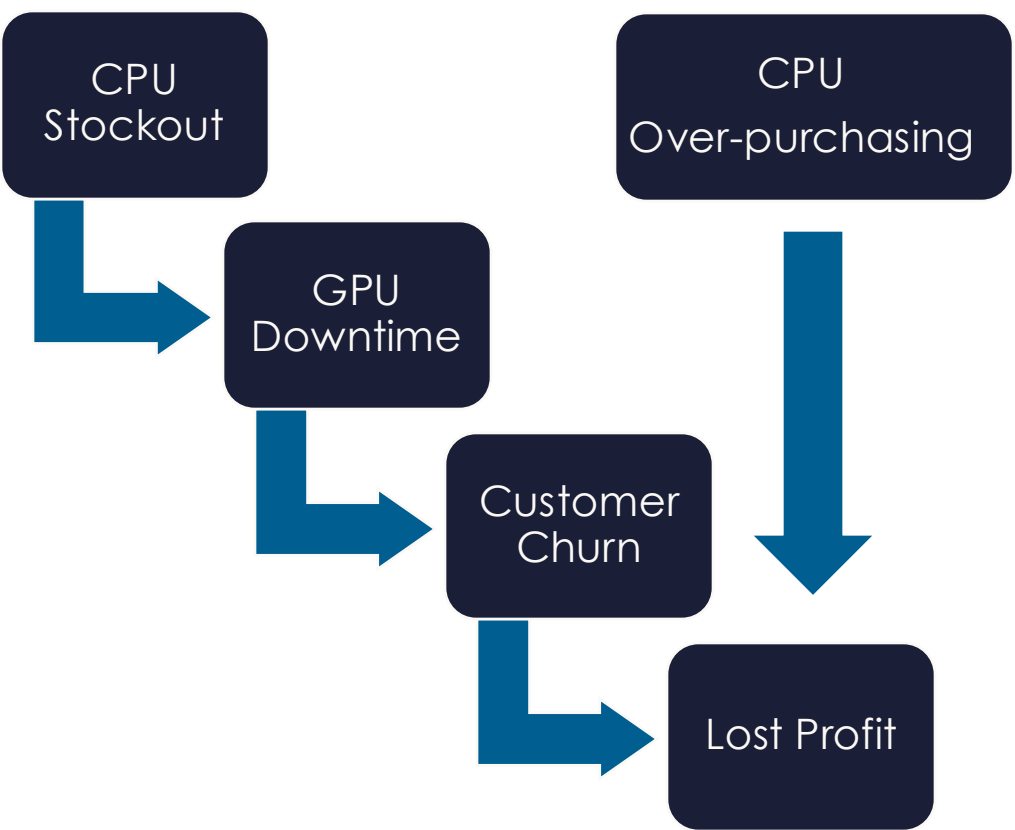
CoreWeave Team: Anna Dinh, Luis Malicay, Michael Rizzo, Wester Schoonenberg, Sekou Sako
MIT Faculty Advisor: James Butler



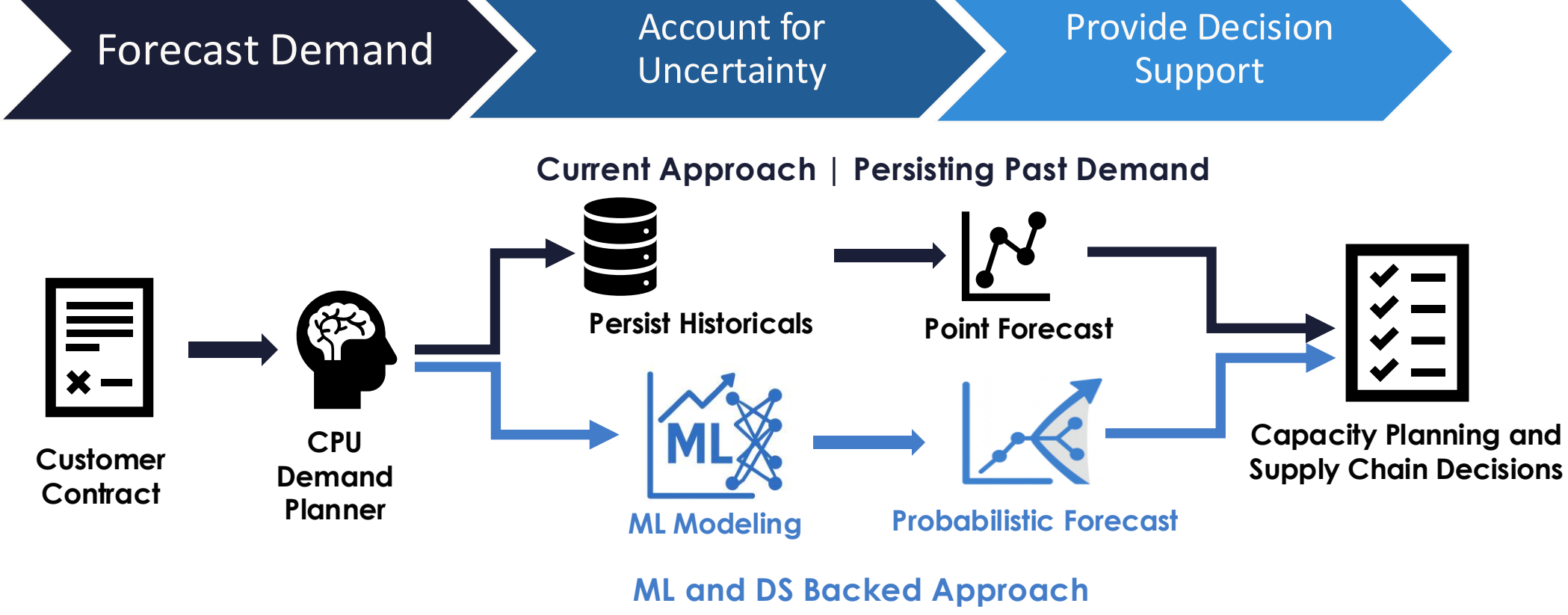
Shane Epstein-Petrullo

Problem Statement

- The Problem:** While GPU capacity is contractually tracked due to its high value, **CPU nodes are necessary, but their demand is hidden by customer workloads**, leaving **no reliable way to forecast CPU demand**—a critical blind spot for capacity planners.
- In AI workloads, **GPUs do the heavy lifting**, but **CPUs coordinate the playbook**—managing data flow, scheduling, and orchestration.
- It is vital to **reduce Capital Expenditure** and **accurately forecast CPU demand** in order to have space for **high-value GPUs**.



Objective



"We currently have **no way to account for uncertainty**. . . If we **can** account for uncertainty, we can decrease the risk of stockouts and I'll have a reasonable buffer with a data-driven decision support tool."

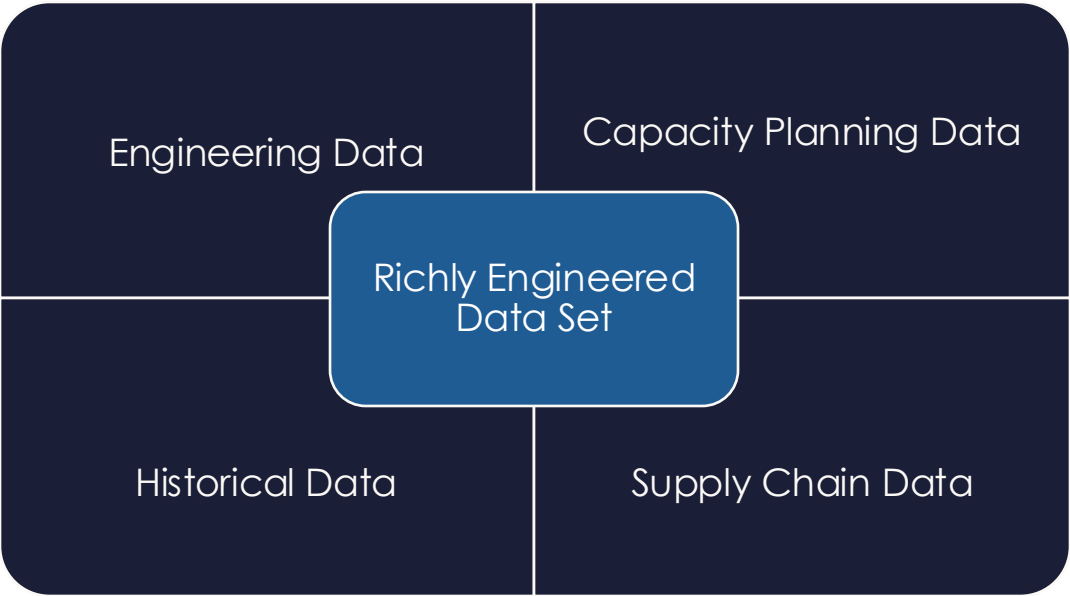
Head CPU Planner, CoreWeave

Challenges and Approach

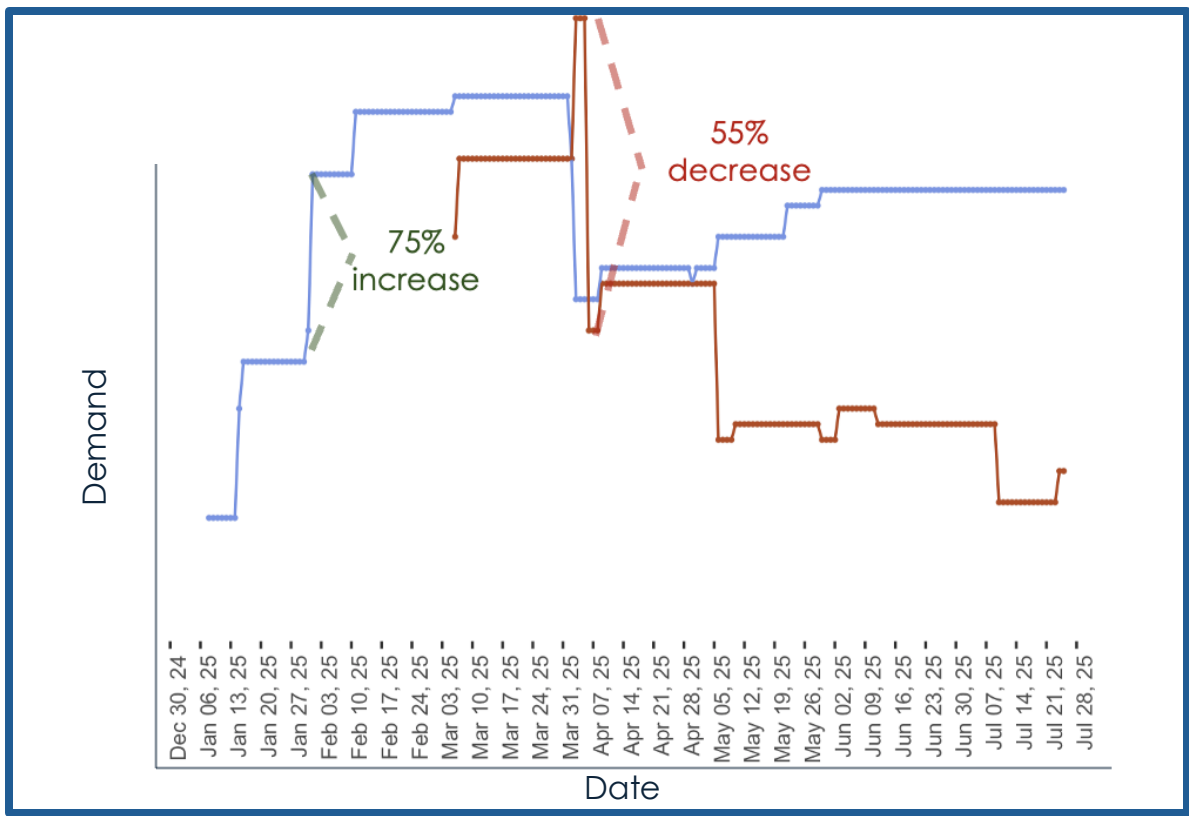
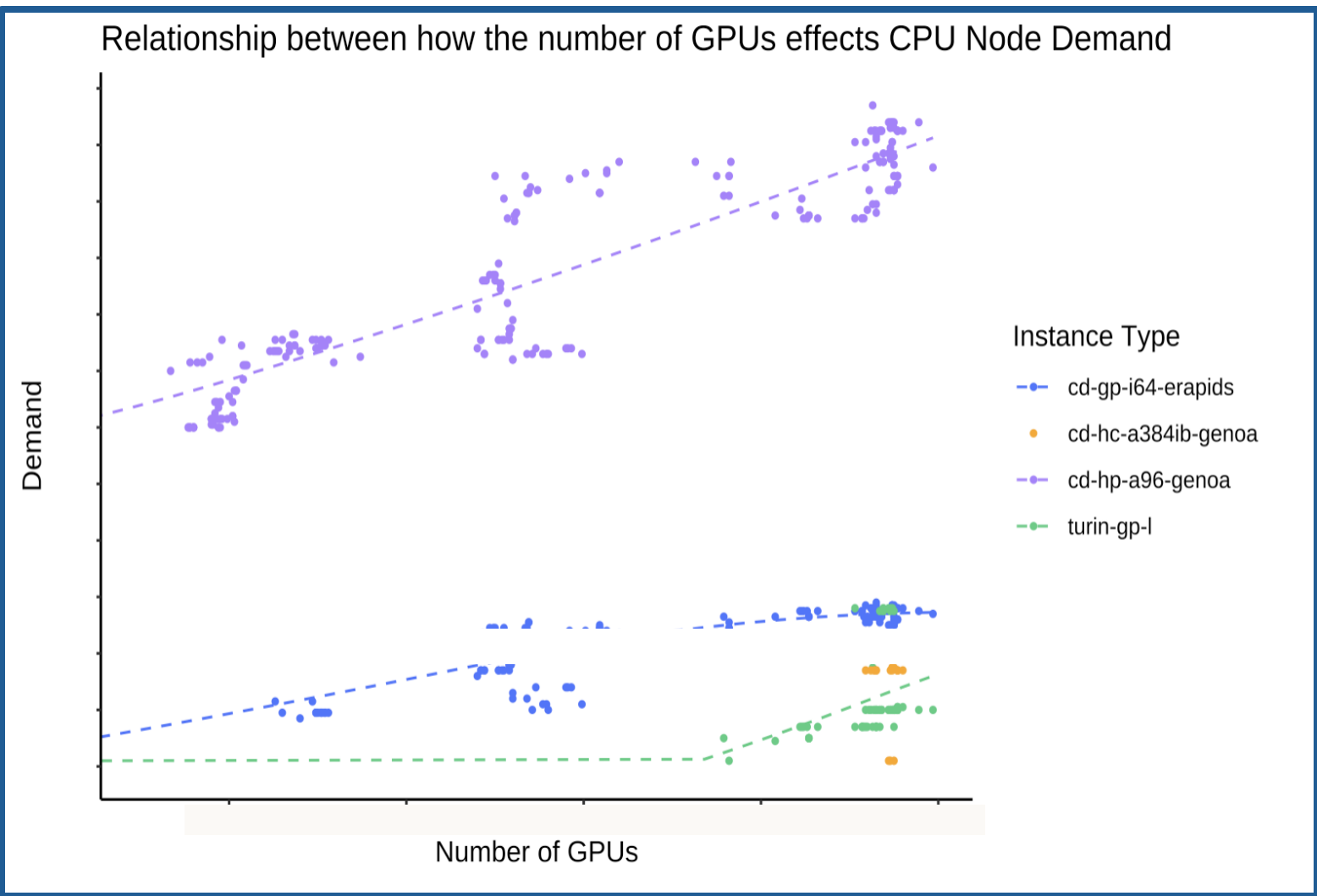
Intermittent Demand & Regime Shifts

- Erratic Demand Data with 3-6 Month Span**
- Dips and Peaks are **NOT** outages. These can be explained.
- Historical data alone is not enough.**

The Data Matures with the Company



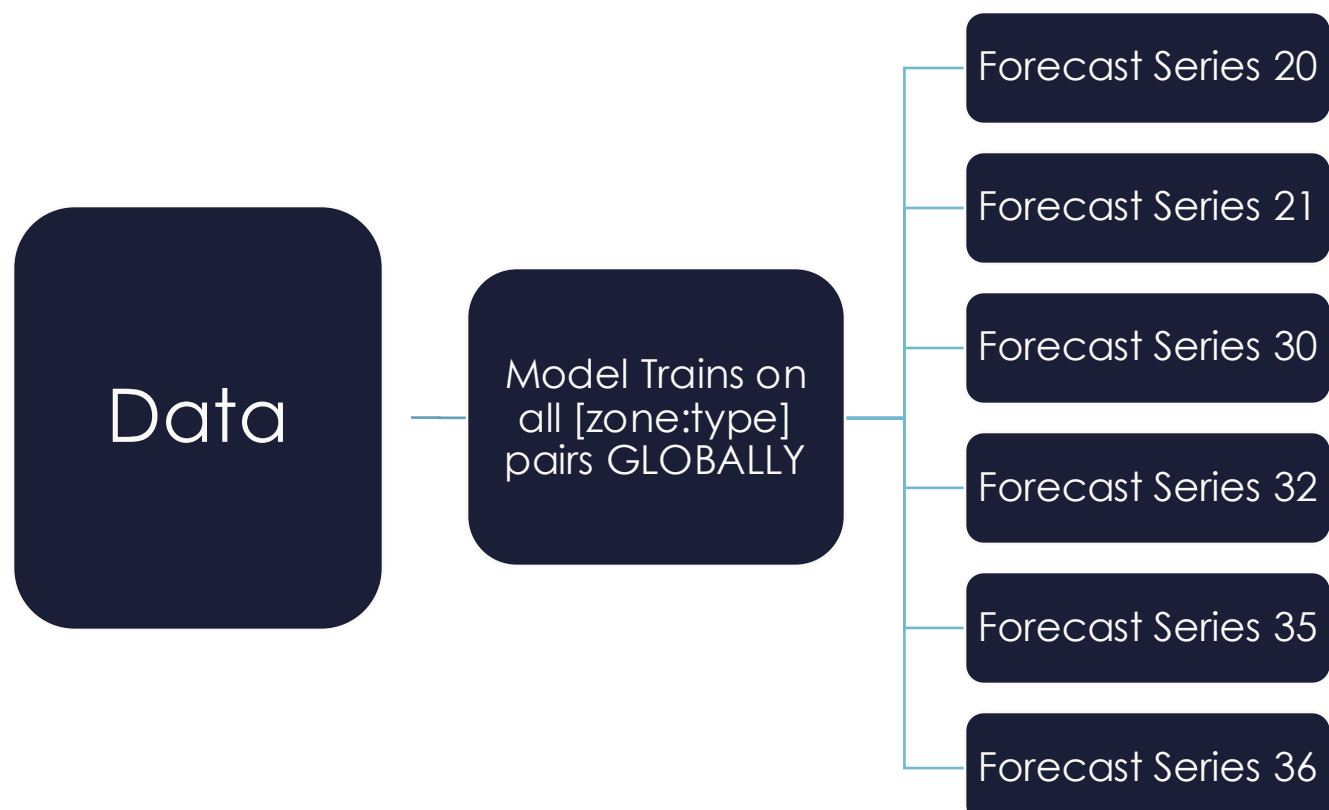
Exploratory Data Analysis



252 unique time series of **Varying Lengths**, Start, and End Dates.

Modeling Architecture

Global Models + Feature Engineering = Cross Learning



Cardinal Rules:

- No Clairvoyance
- Start Linear, then Advance

Methodology and Process

Multi-Step Forecasting

Recursive Forecast

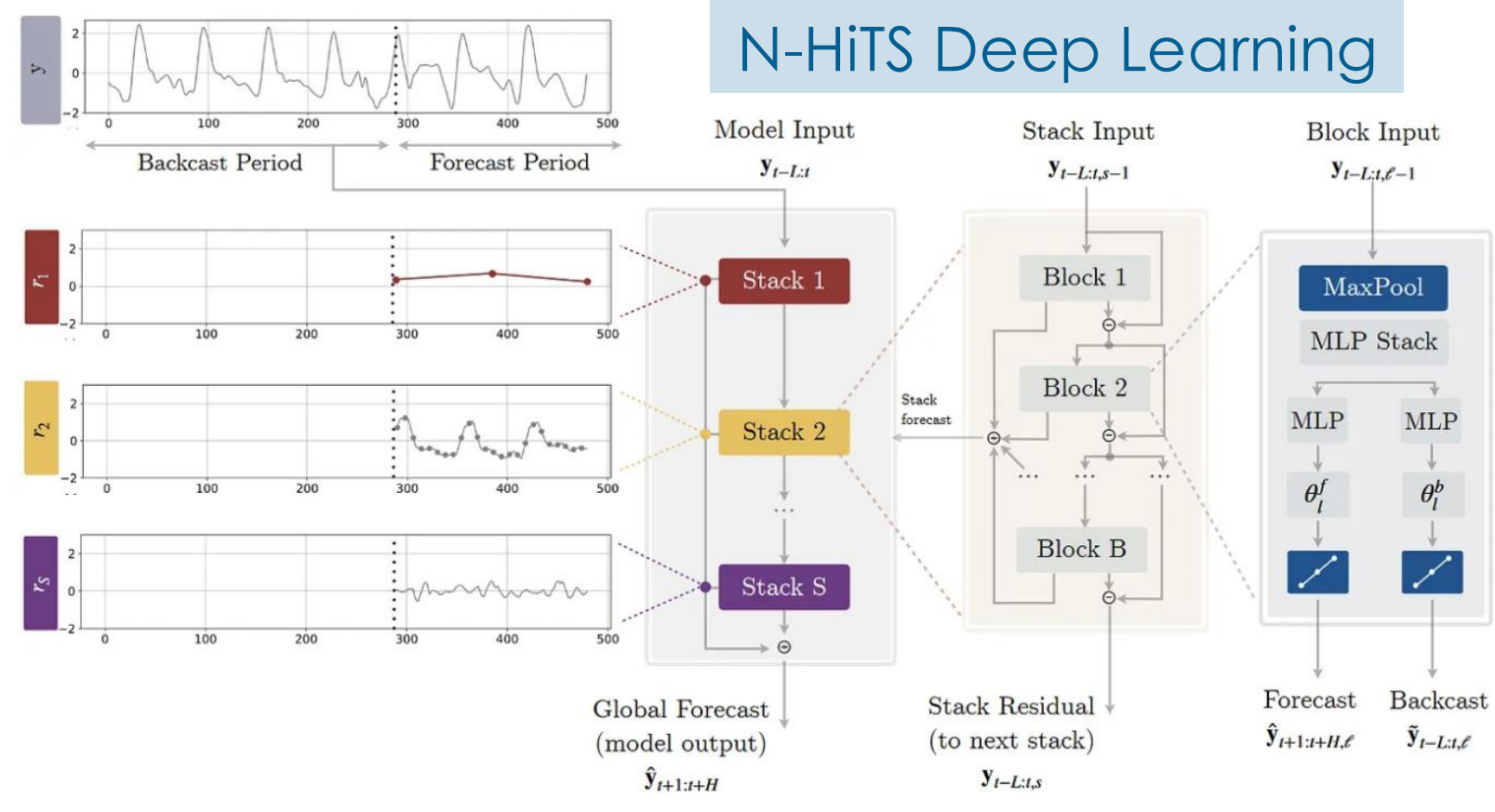
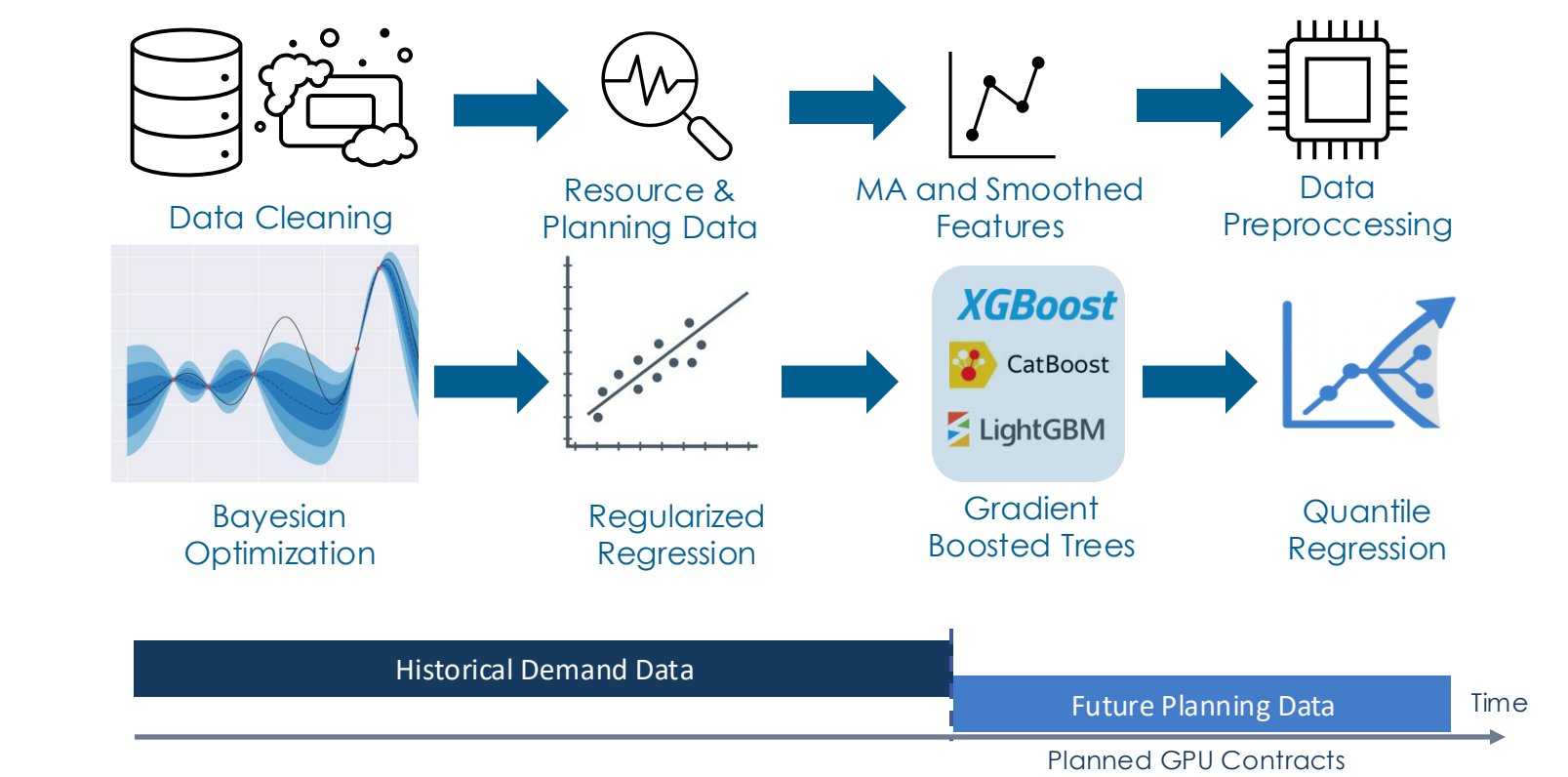
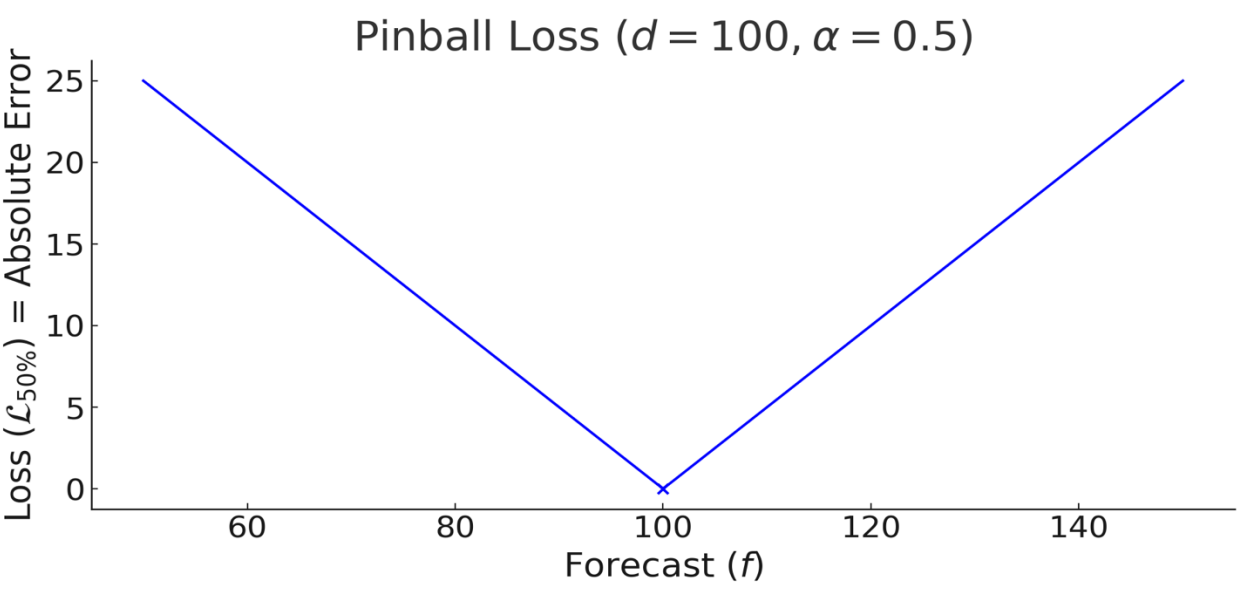
$$\hat{y}_{t+h} = f(\hat{y}_{t+h-1}, \hat{y}_{t+h-2}, \dots, \hat{y}_{t+h-w}, \hat{x}_{t+h-1}, \hat{x}_{t+h-2}, \dots)$$

Day 30 is predicted using day 29's prediction

Direct Forecast

Fit models for every step in the horizon: $t, (t+1), \dots, (t+h)$

Probabilistic > Point



Business Impact and Implementation

\$2.6M

Reduction in Capital Expenditure across instance types in the CKS Fleet

~50%

Reduction in stockouts due to improved forecasts

48%

Reduction in forecasting error (MAPE) compared to baseline



- Next 2 Months: Model in production**
- Legacy Baseline vs. ML Models**
- Performance Evaluated in Real-Time**
- Update Parameters & Manage Models**

