Explainable Financial Forecasting

Kevin Cushing, Anik Dey, and
Aarav Sane

PURDUE
UNIVERSITY

Automating a Financial Analyst

Task/Domain

- Identify investment opportunities
  - Research companies
  - Analyze past performance
  - Forecast potential earnings

Problem

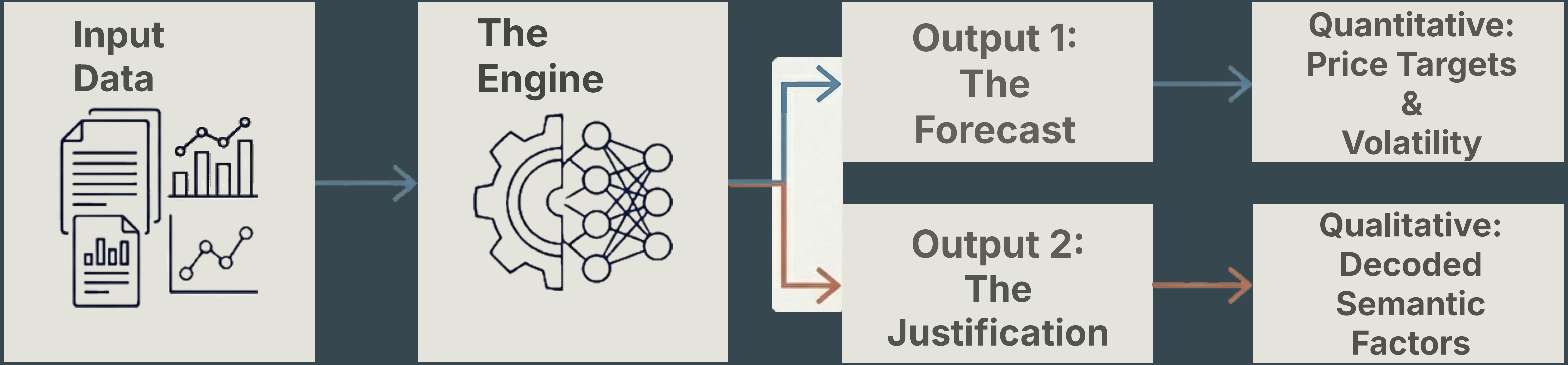
- Complex models are black-boxes
- Lack justifications for decisions/predictions

Requirement

- Leverage Multi-modal Data
  - Time-series
  - Text
- More Faithful Projections
  - Accurately predict prices/volatility
  - Justify these predictions

PURDUE UNIVERSITY

Process Overview



Data

Global Factors

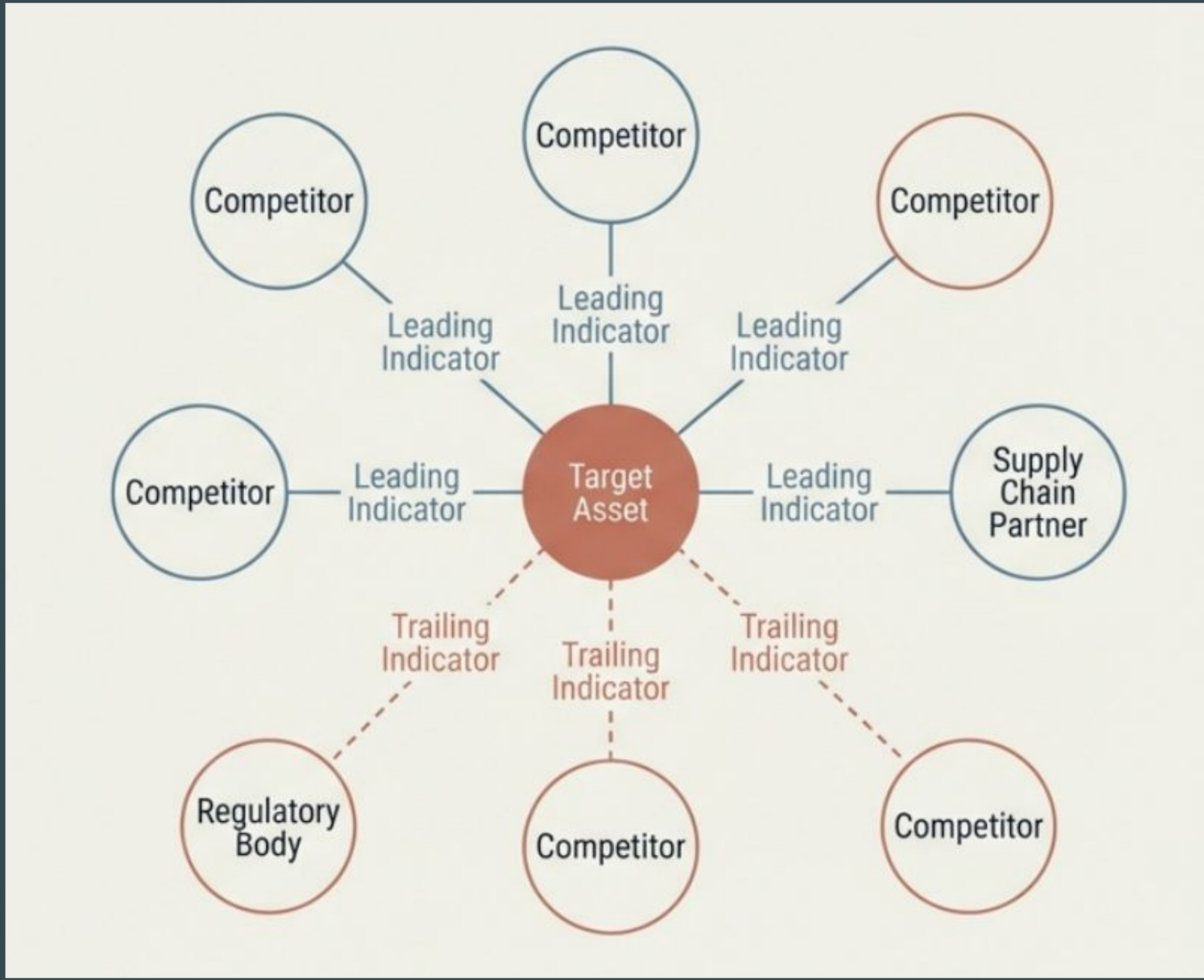
- fed rates
- market indices
- government policies

Industry Factors

- industry indices
- supplying/consuming companies
- competitors/partners

Local Factors

- trade volumes/prices
- earnings reports/quarterly filings
- press releases
- relevant news articles
- sentiment analysis from social media

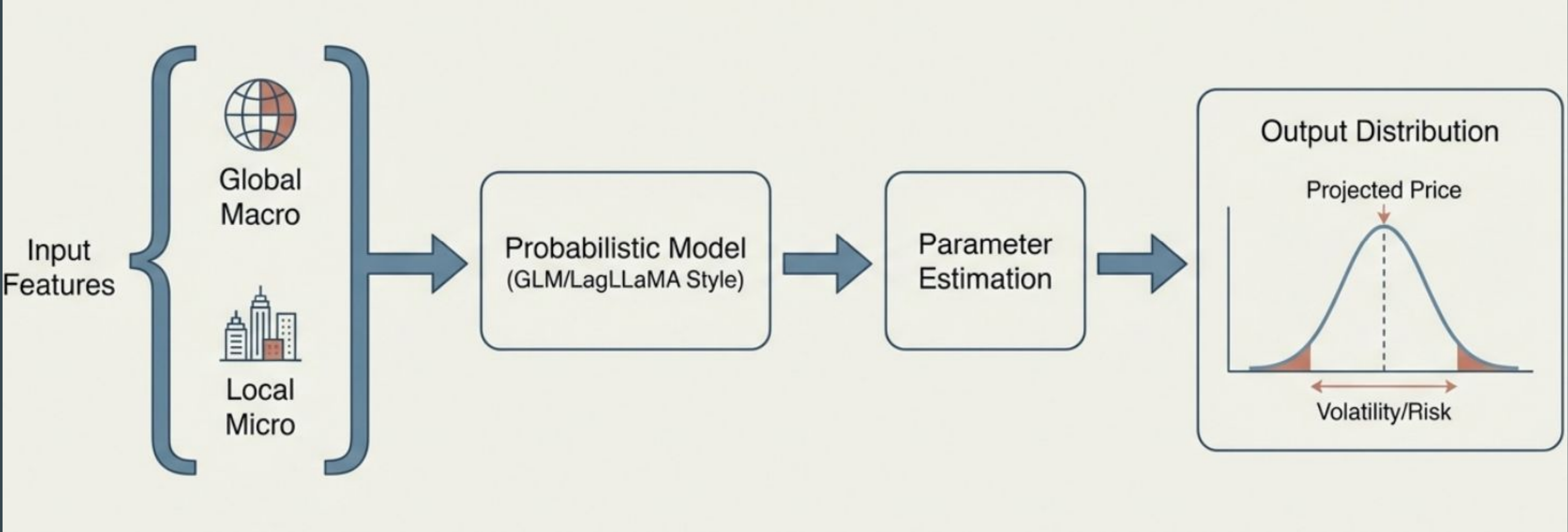


PURDUE UNIVERSITY

Potential Data Sources

- CRSP, Yahoo Finance – historical price data
- SEC Edgar – corporate filings
- Google Search API (or GDELT) – news articles
- Twitter/Reddit API – open-forum comments
- Seeking Alpha – earnings call transcripts

Forecasting



Unlike standard regression models that output a single point, this architecture predicts parameter values for probability distributions. It integrates methods for explaining projections directly within the distribution parameters, offering a view of both "what could happen" and "how likely it is

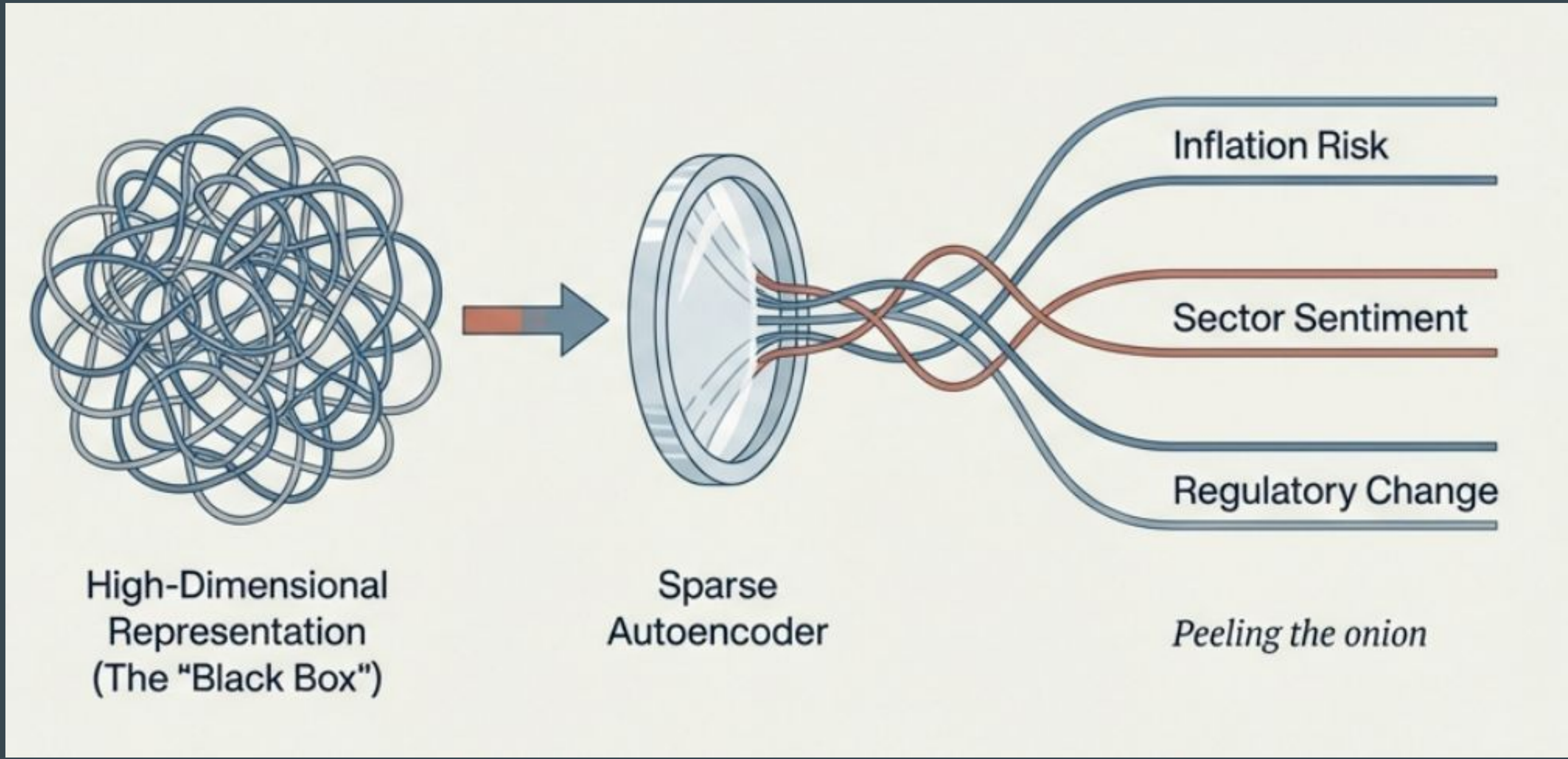
Justification

Existing approaches:

- LIME
- SHAP
- conditional effect analysis
- counterfactual reasoning

Probable approach:

- Sparse Autoencoders (SAEs)
- learn mapping from neuron activations to sparse vector of financial concepts
- explain projections by which "concepts" are activated



Why Models Are Hard to Interpret

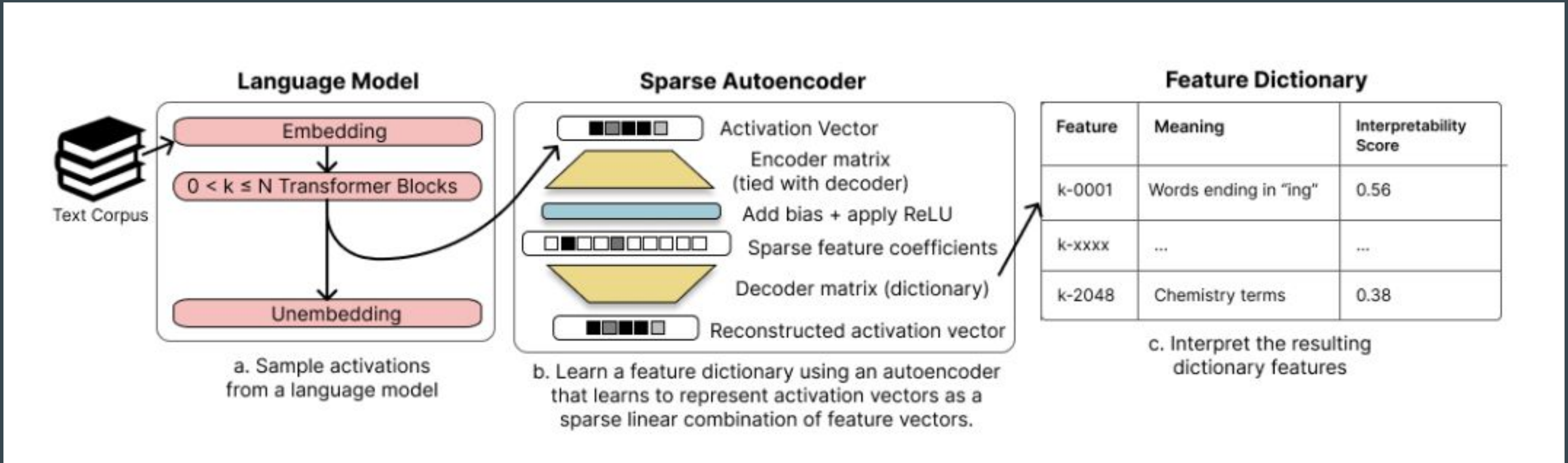
Polysemanticity:

- A single neuron often activates for multiple unrelated concepts. Example: one neuron lights up for both dogs and the color yellow.
- Makes it difficult to find human-understandable explanations.

Superposition:

- The network stores more features than neurons.
- Encodes features as combinations of neurons (directions in activation space).
- Each neuron contributes to many features.

SAE Methodology (Figure)



SAE Methodology

Goal:
Extract meaningful, human-interpretable features from tangled activations of a model (NN/Transformer).

Architecture:
Input x → Encoder (Mx + b) → ReLU → Sparse code c → Decoder ($M^T c$) → Reconstructed vector x̂

$c = ReLU(Mx + b)$

$\hat{x} = M^T c = \Sigma c_i f_i$

Loss Function: $L(x) = ||x - \hat{x}||^2{}_2 + \alpha||c||_1$

• Reconstruction term: keeps x̂ close to x.

• Sparsity term: ensures only a few features activate.

Key Idea:
By enforcing sparsity, the autoencoder learns disentangled directions, taking features out of superposition.

PURDUE UNIVERSITY

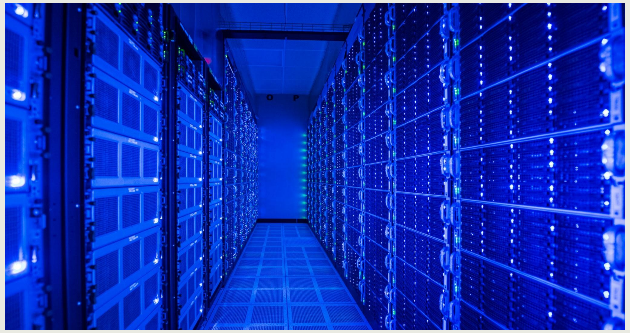
Challenges

Data

- Finding/creating a dataset
- Combining stock prices and text data from news sources

Training

- Large models require significant resources.



Mapping Concepts

- SAEs are mostly used for LLMs/LVLMs
- Can they map to concepts from time-series data?

PURDUE UNIVERSITY

Questions/Thoughts?

PURDUE UNIVERSITY

Thank You