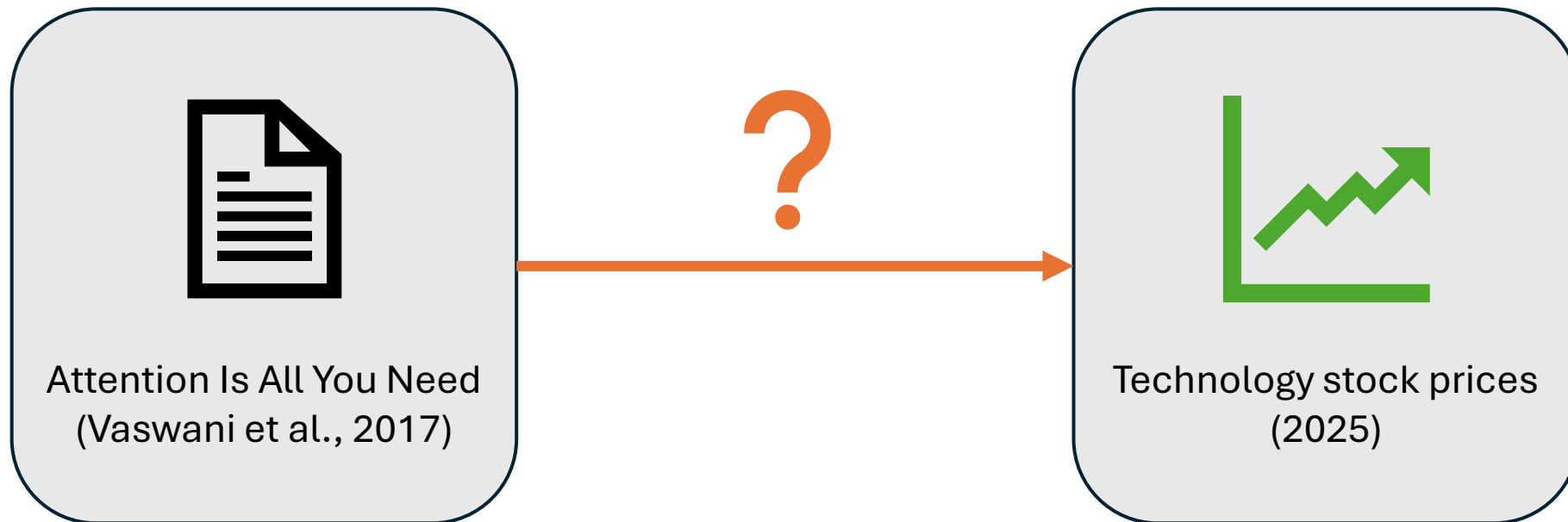
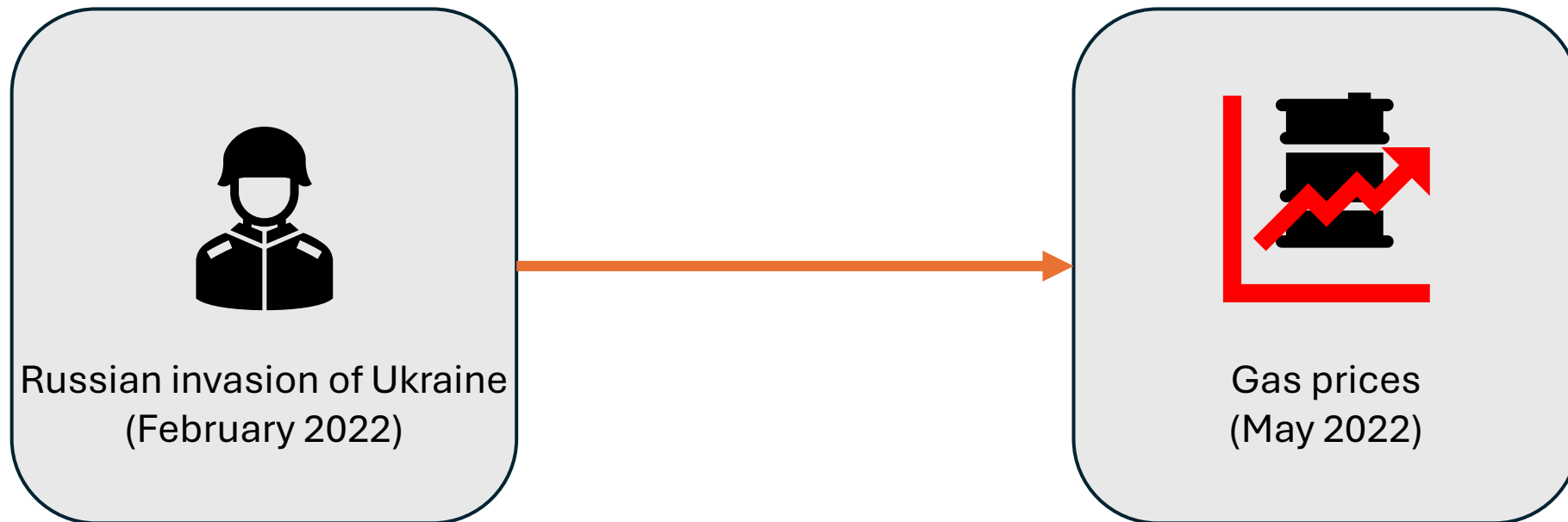


Predicting Near-Term Stock Price Movements From Scientific Literature Trends

Motivation: Scientific literature “oversees the horizon” of discoveries.
Can emerging trends in the scientific literature predict shifts in market sectors?



Consideration: the effect of scientific trends takes time to unfold.
As opposed to political trends for example, scientific trends have a bigger lag



Question: what does a “scientific trend” even mean?

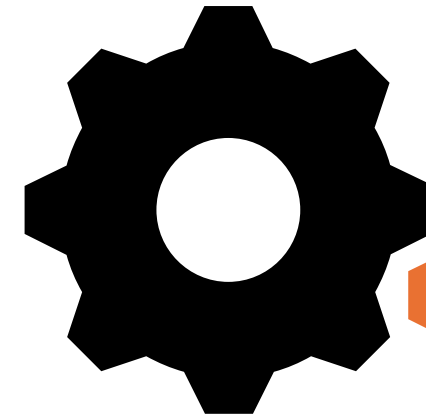
Quantifying the term “trend” is crucial to extract relevant features for prediction



noun

1. a general direction in which something is developing or changing.

Scientific field growth



Topic growth

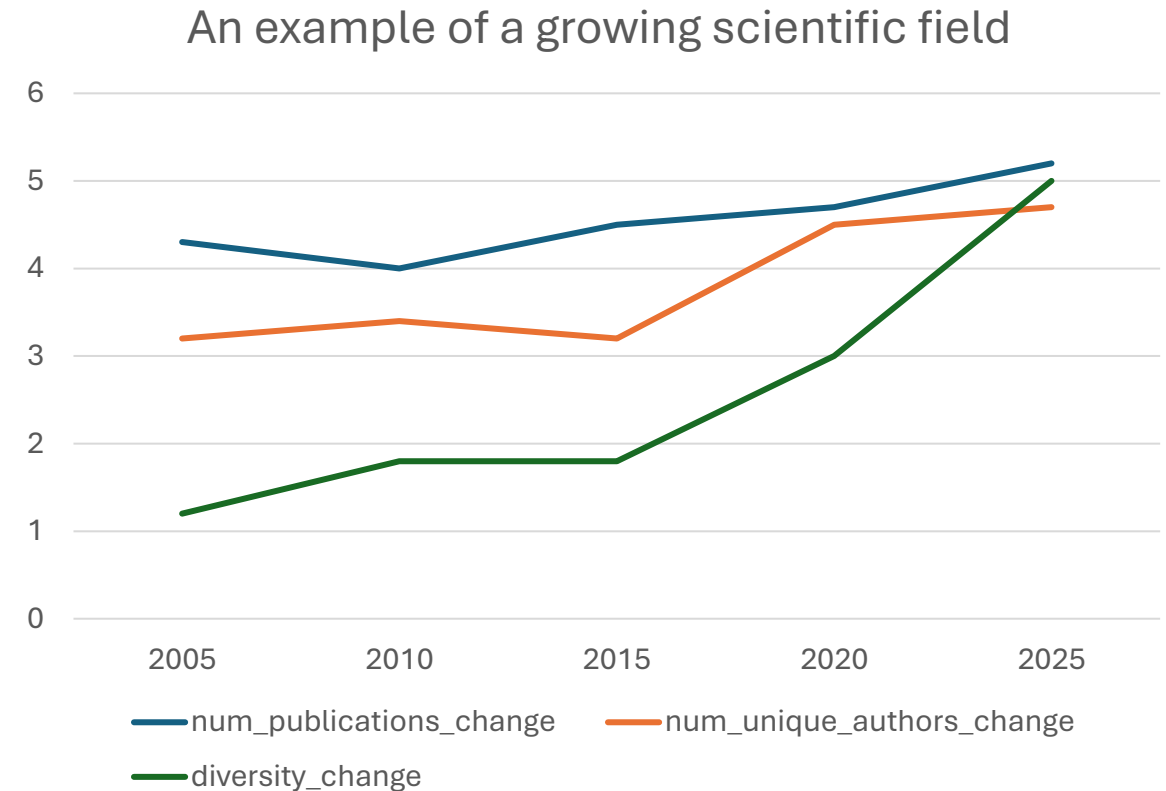


Aim 1: measure scientific field growth

Promising measures: # publications, # unique authors, and “diversity” per field

- **Number of publications:** volume of papers being published in a scientific field suggests its growth
- **Unique authors:** indicates that more researchers are publishing in a scientific field
- **Diversity:** a field’s applications are spreading into others (e.g., a computer science paper with a finance application)

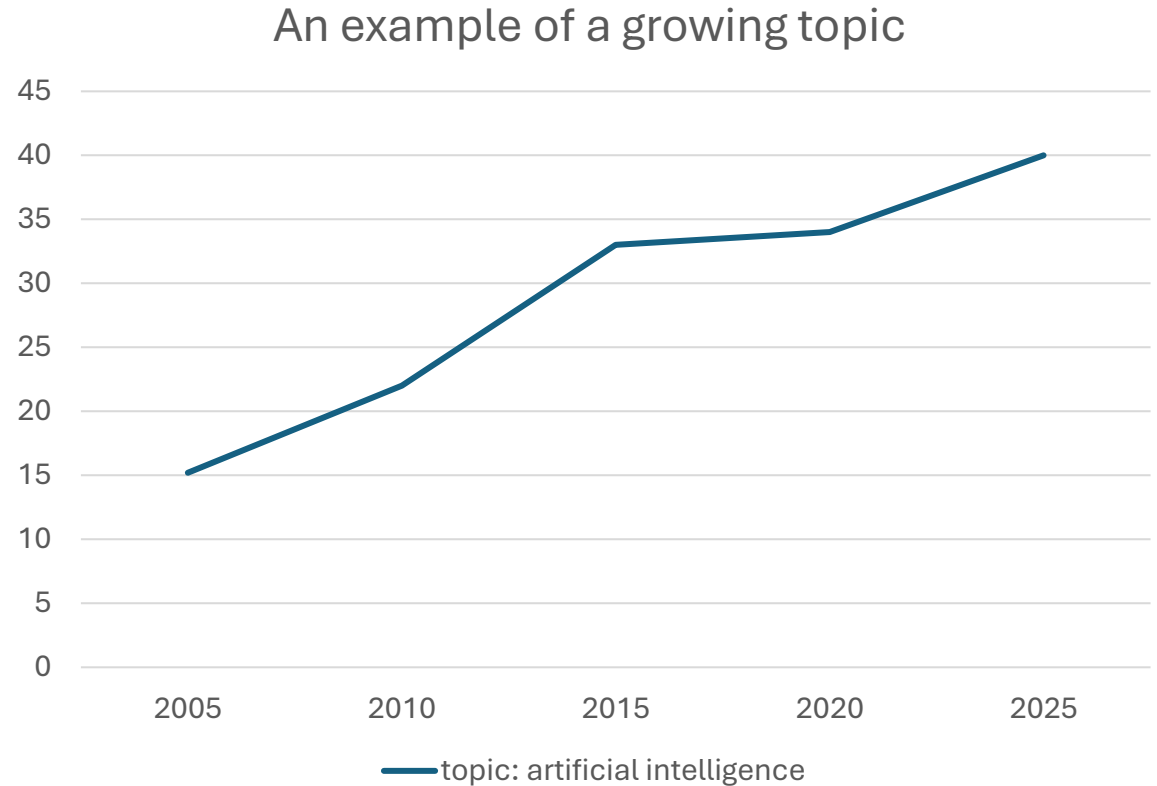
Measuring growth by field enables analyzing each field’s impact on specific market sectors (e.g., CS may influence tech, biology may not)



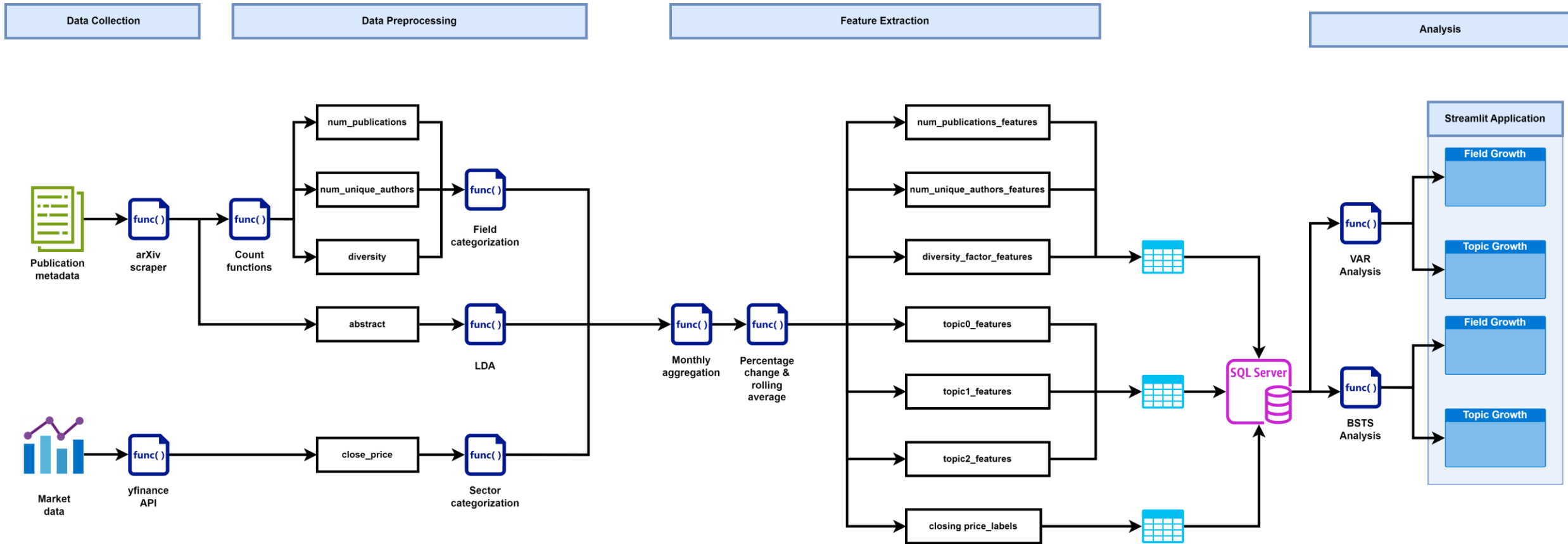
Aim 2: measure topic growth

It could be measured by identifying the most prevalent latent topics in the literature

- **Topic prevalence:** if it's prevalent across the scientific literature, it's *trendy*
- **Topic growth:** tracks emerging themes across scientific fields (e.g., AI arising in both CS and ethics).



System Architecture



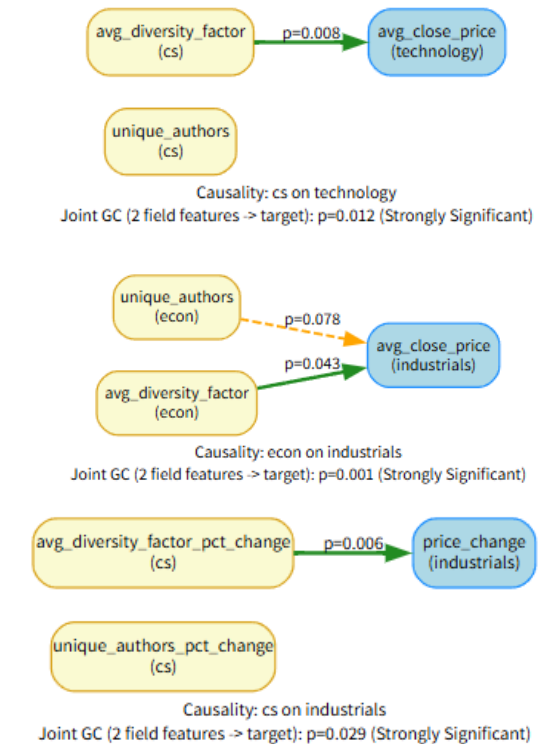
Data Sources and Assumptions

- **Six scientific fields:** computer science, mathematics, physics, economics, electrical engineering & systems science, quantitative biology
- **Five market sectors:** technology, financials, healthcare, energy, industrials
- **Four feature types:** monthly values, monthly percentage changes, rolling average (period = 6 months), rolling growth rate (period = 6 months)
- **Two analyses:** Vector Autoregression (VAR), and Bayesian Structural Time Series (BSTS)

Key Findings (Field Analysis)

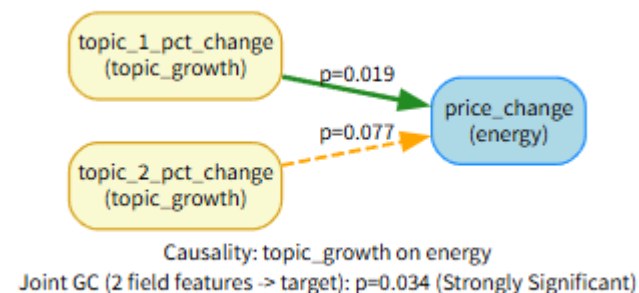
- Scientific field growth Granger-caused changes in their related sectors
 - CS and Tech, Economics and Industrials
- Rolling average and growth rate models underperformed or diverged
 - Likely cause: smoothing over already stable series “oversmoothened” the series
- Diversity was a key predictor across multiple fields

Granger Causality Network

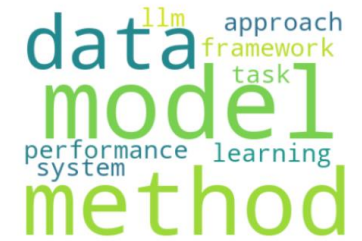


Key Findings (Topic Analysis)

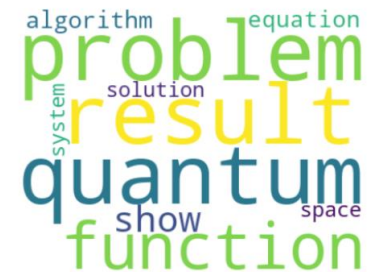
- Topic 1 growth Granger-caused monthly averages in 5 of 6 sectors
 - Stronger effects observed with monthly percentage change data
- Topic growth Granger-caused changes in their related sectors
 - Energy topic and the energy sector
- Rolling average and growth rate models did not provide useful insight



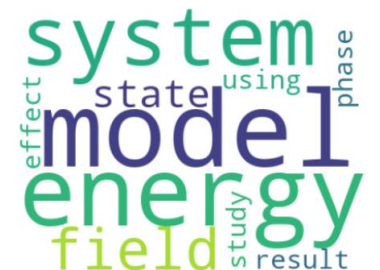
Word Cloud for Topic 0



Word Cloud for Topic 1



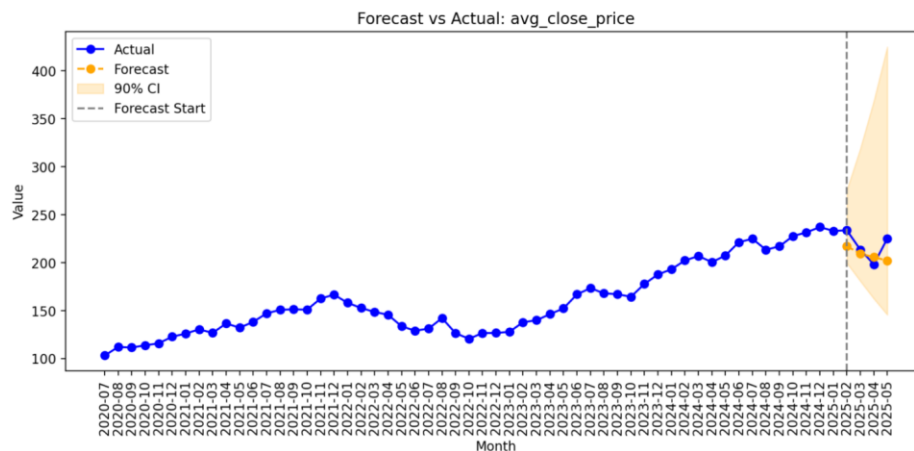
Word Cloud for Topic 2



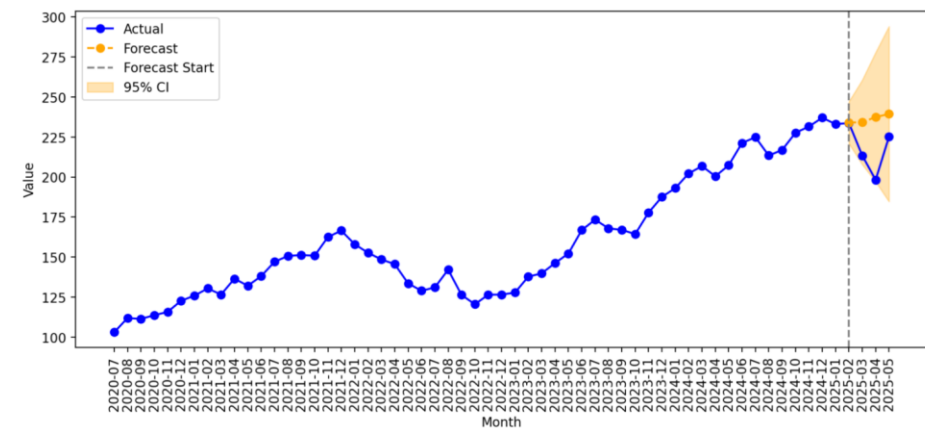
Forecasts

- Forecast window: 2025-02 to 2025-05
 - *A peculiar* time to forecast market prices
- BSTS forecasted better than VAR but displayed larger confidence intervals on average
 - Likely cause: small data sample

BSTS (CS on Tech)



VAR (CS on Tech)



Future Work

- **Data Collection:**

- More data allows for investigating longer-term dependencies.
- Expand ability to scrape from other archives and APIs (e.g., Wiley)

- **Feature Extraction:**

- *Not all papers were born equal*: investigate highly cited papers / journals of high impact
- Sentiment analysis: is a popular topic's sentiment a reliable predictor?
- Newly emerging topics (a topic with a rapidly increasing growth)

Future Work (Cont.)

- **Analysis:**
 - More data would allow for modeling using DNNs like LSTM and CNN
 - Rolling Forecasts: forecasts over small windows are highly subject to noise (e.g., the 2025-04 dip due to Liberation day tariffs)
- **Interactive Web App:**
 - Use Flask/Django to enable more interactive features and customizability
- **Optimizations:**
 - Faster function implementations, better documentation, LDA & BSTS fine-tuning
 - Model containerization to facilitate deployment and iterative retraining
- **Other Directions:**
 - Influence of political agreement and alliances (e.g., will the recent AI agreements between USA and GCC boost scientific literature growth which is expected to in turn have an influence on market price movements?)