The Knowledge Graph for Macroeconomic Analysis with Alternative Big Data

Yucheng Yang

Princeton University

Joint with Yue Pang (PKU), Guanhua Huang (USTC) and Weinan E (PKU & Princeton)

2022 Monash-Warwick-Zurich Text-as-Data Workshop

Motivation

- Traditional macro models only have a handful of variable inputs.
- Big data and machine learning allow us to study macro with many more variables, especially new alternative variables.
- Two approaches:
    1. Put many variables in statistical models directly, w/o understanding their relationships.
    2. Study one or few alternative variables in a time.
- Need new knowledge system on relations among traditional and new economic variables to design model inputs systematically.
- This paper: build a knowledge graph (KG) of relations among traditional and alternative data variables, using text data of research papers/reports.

Introduction: Knowledge Graph

- Knowledge graph: knowledge base that uses graph topology to represent interlinked descriptions of entities.
- Basic elements: "RDF triple" with form {subject, predicate, object}.
- Prominent application: Google Search.

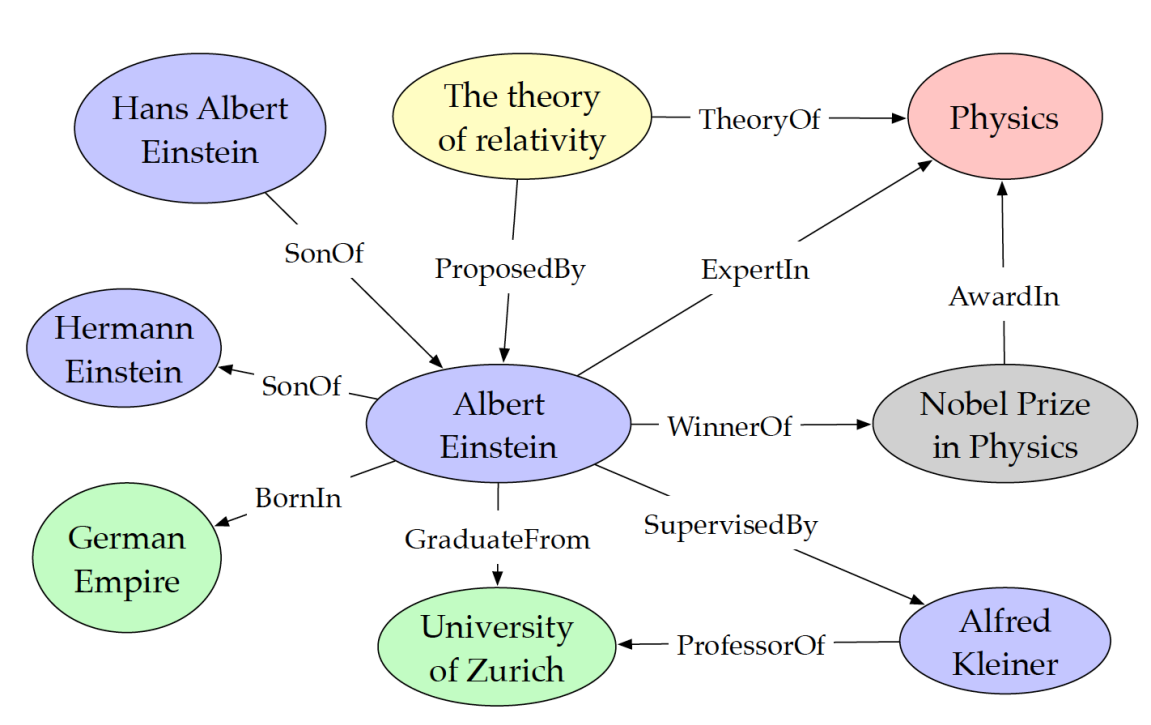
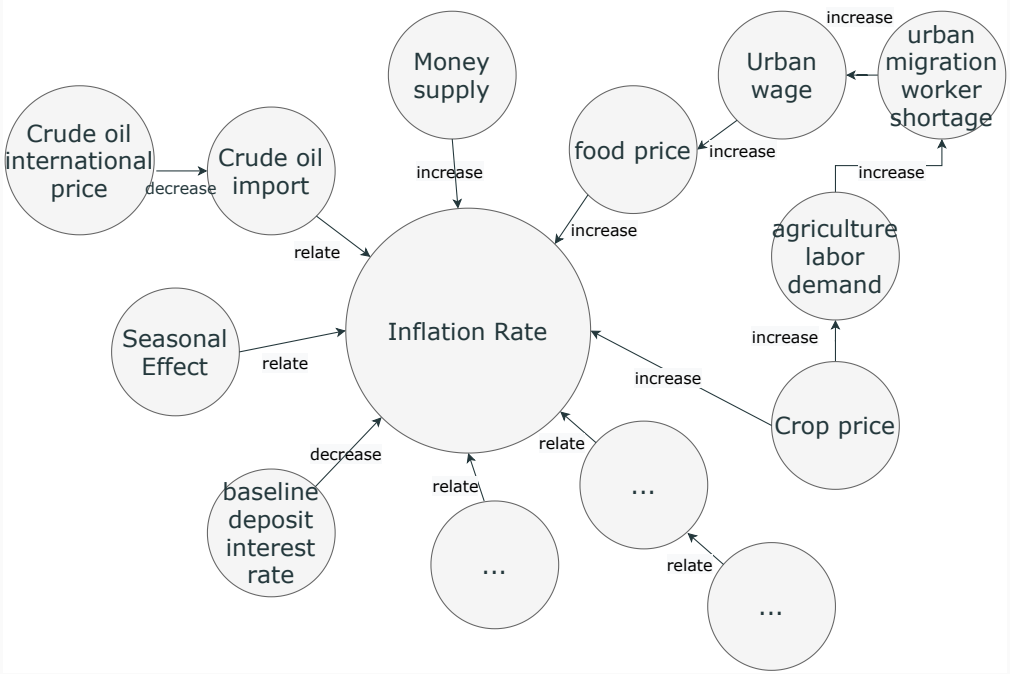


Figure 1: Example of Knowledge Graph on Einstein (Ji et al., 2020)

This Paper



- We build knowledge graph (KG) of linkages between traditional economic variables and alternative data variables.
- Extract "RDF triples" from text data of research paper/reports.
- We apply KG of economic variables for variable selection in economic forecasting.

Literature

- Big data and machine learning in macroeconomics:
  1. McCracken and Ng (2016): FRED-MD/QD.
  2. Chetty et al. (2020): private sector data to study impact of COVID-19.
  3. Stock and Watson (2016), Coulombe et al (2019): factor and ML models.
  4. Han, Yang, and E (2021): deep learning to solve high dim macro models.

  Our work: a new knowledge system to study big data in macro.
- Knowledge graph of scientific research: Luan et al (2018), Tshitoyan et al (2019). Our work: the first knowledge graph in economics.
- Text as data in economics: Gentzkow et al. (2019), Ash and Hansen (2022). Our work: extract knowledge and reasoning, beyond sentiment and topics.
- Variable selection and model reduction in economic forecasting:
  1. Tibshirani (1996), Zou and Hastie (2005): shrinkage methods.
  2. Giannone et al. (2008), Stock and Watson (2016): factor models.
  3. Goodfellow et al. (2016): autoencoder with deep learning.

  Our work: knowledge graph based variable selection.

Data and Methodology for Knowledge Graph Construction

Textual Data for Knowledge Graph Construction

- Data: industry macro research reports from China.
    1. Focus on analyzing or forecasting dynamics of aggregate variables, and always clearly state what variables are studied in each report.
    2. Mostly adopt the narrative approach (Shiller, 2017), clearly state the logic chains of their analysis in narrative language, rather than in theoretical or quantitative models.
    3. Freely available and can be downloaded massively from the WIND database.

Construction of Knowledge Graph: An Example

A research report paragraph that studies the dynamics of inflation rate in China:

> "A long-term systematic migrant worker shortage began to appear in the Chinese migrant labor market around 2005, which greatly increased the growth rate of migrant workers' wages, resulted in the increase of food prices, and pushed up the increase in consumer price index, making the average level of inflation probably 100 to 200 basis points higher."

We hope to extract all the economic variables and relation among those variables, store them in *RDF triples* of {variable 1, relation, variable 2} format.

Construction of Knowledge Graph: An Example (Ct'd)

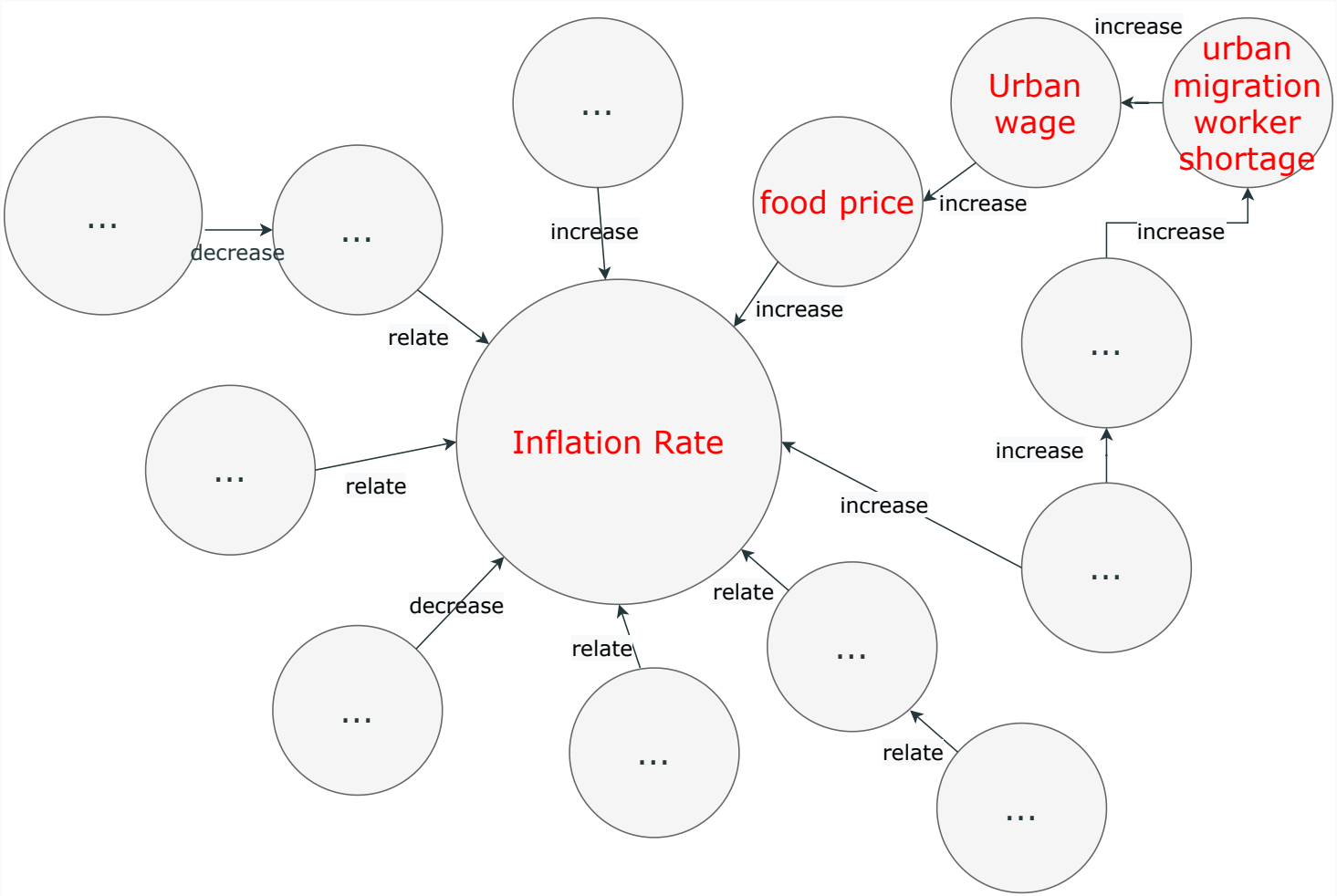
"A long-term systematic **migrant worker shortage** began to appear in the Chinese migrant labor market around 2005, which greatly **increased** the **growth rate of migrant workers' wages**, **resulted in the increase** of **food prices**, and **pushed up** the increase in **consumer price index**, **making** the average level of **inflation** probably 100 to 200 basis points **higher**."

RDF triples of {variable 1, relation, variable 2} format:

- {migrant worker shortage, increase, growth rate of migrant workers' wages}
- {growth rate of migrant workers' wages, resulted in the increase, food prices}
- {food prices, push up, consumer price index}
- {food prices, make higher, inflation}

Construction of Knowledge Graph: An Example (Ct'd)

After removing duplicates, we get:



General Procedure

Step 1. Make a list of aggregate variables of interest, together with their variants. E.g. Inflation (CPI).

Step 2. Find all these aggregate variables and their variants in the documents with string matching.

Step 3. For each aggregate variable detected in the documents, find all the other variables around it, as well as the relation among aggregate variables and other variables.

Step 4. Represent all the variables and relations extracted with the *RDF triple* structure.

Step 5. Drop all the duplicates and build the knowledge graph.

Main Challenges: Named Entity Recognition

Named entity recognition (NER) for economic variables is very hard, since they are mostly multi-token entities with complicated semantic patterns.

- Examples: "migration worker shortage", "growth rate of migration workers' wages", "processing firm registrations in China", "leverage rate of local government financing vehicles".

We develop an active learning algorithm with human involvement to extract variable entities and relation keywords from the textual data.

- Active learning: when labeling is expensive, need an algorithm to decide which data we want the human editors to label and the model to learn from.

Active Learning with Human Involvement

Step 1. Initial set of economic variables and initial set of relation keywords.

Step 2. Use current set of economic variables to train a language model to predict whether a phrase is a variable. Expand the variable set.

Step 3. Find sentences containing many variables, but few relation keywords. Then use human editors to find the relation keywords in these sentences. Expand the set of relation keywords.

Step 4. Find sentences containing relation keywords, but few variables. Then use human editors to find the variables in those sentences. Repeat Step 2 to expand the variable set.

Step 5. Repeat Steps 3 and 4, until convergence.

Construction of Knowledge Graph: Results

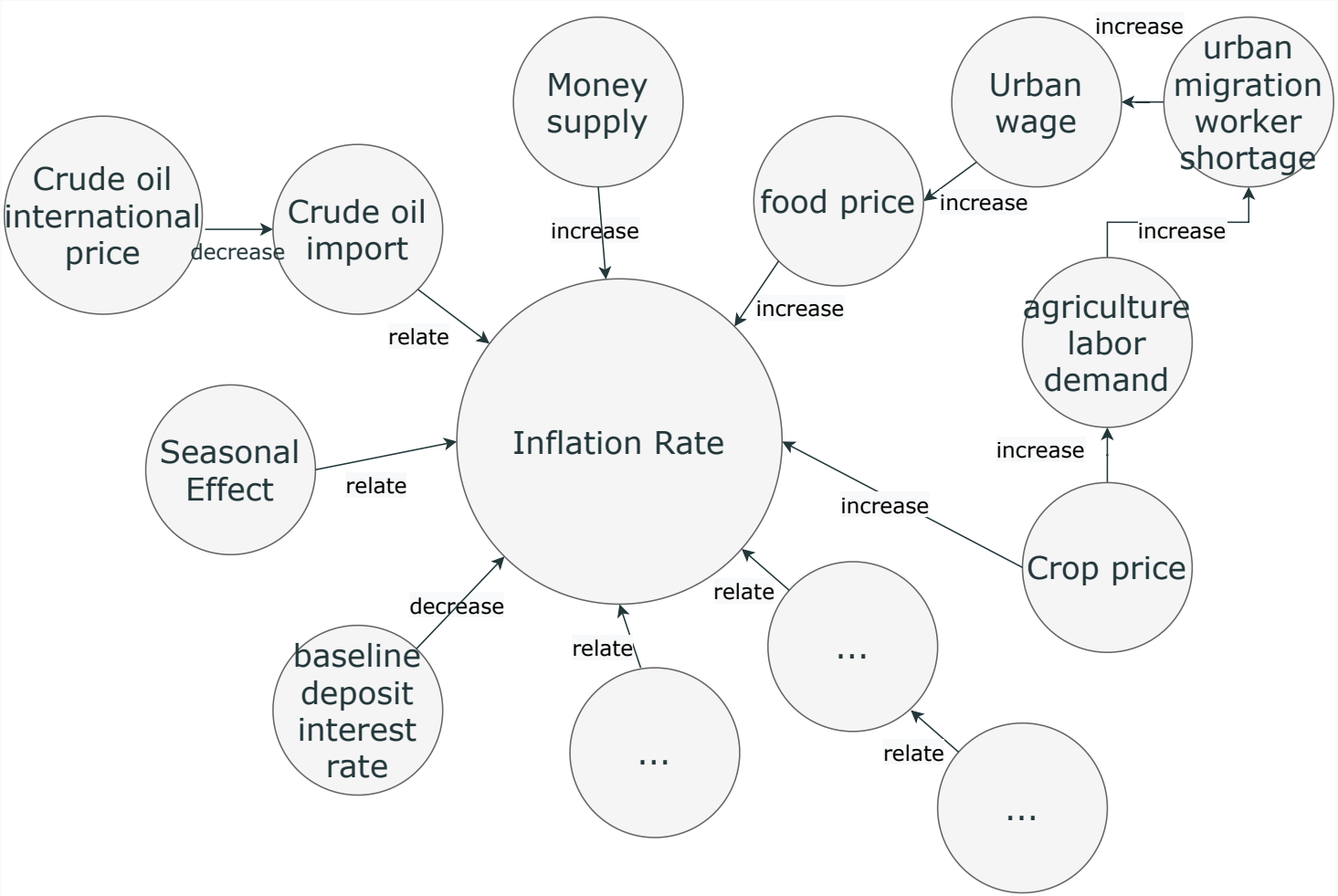


Figure 2: Example of Knowledge Graph on Inflation

Construction of Knowledge Graph: Some Remarks

1. Link traditional variables of interest to other traditional variables or alternative variables.
2. Relation belong to three classes: increase (positive relation), decrease (negative relation), relate (neutral relation).
3. Linkages can be one layer or multiple layers.
4. Can put time stamps, institution stamps, etc. on the variables and linkages in the graph.
5. The algorithm can be applied to other textual data: NBER working papers, leading institutional forecasters' reports.

Construction of Knowledge Graph: Results

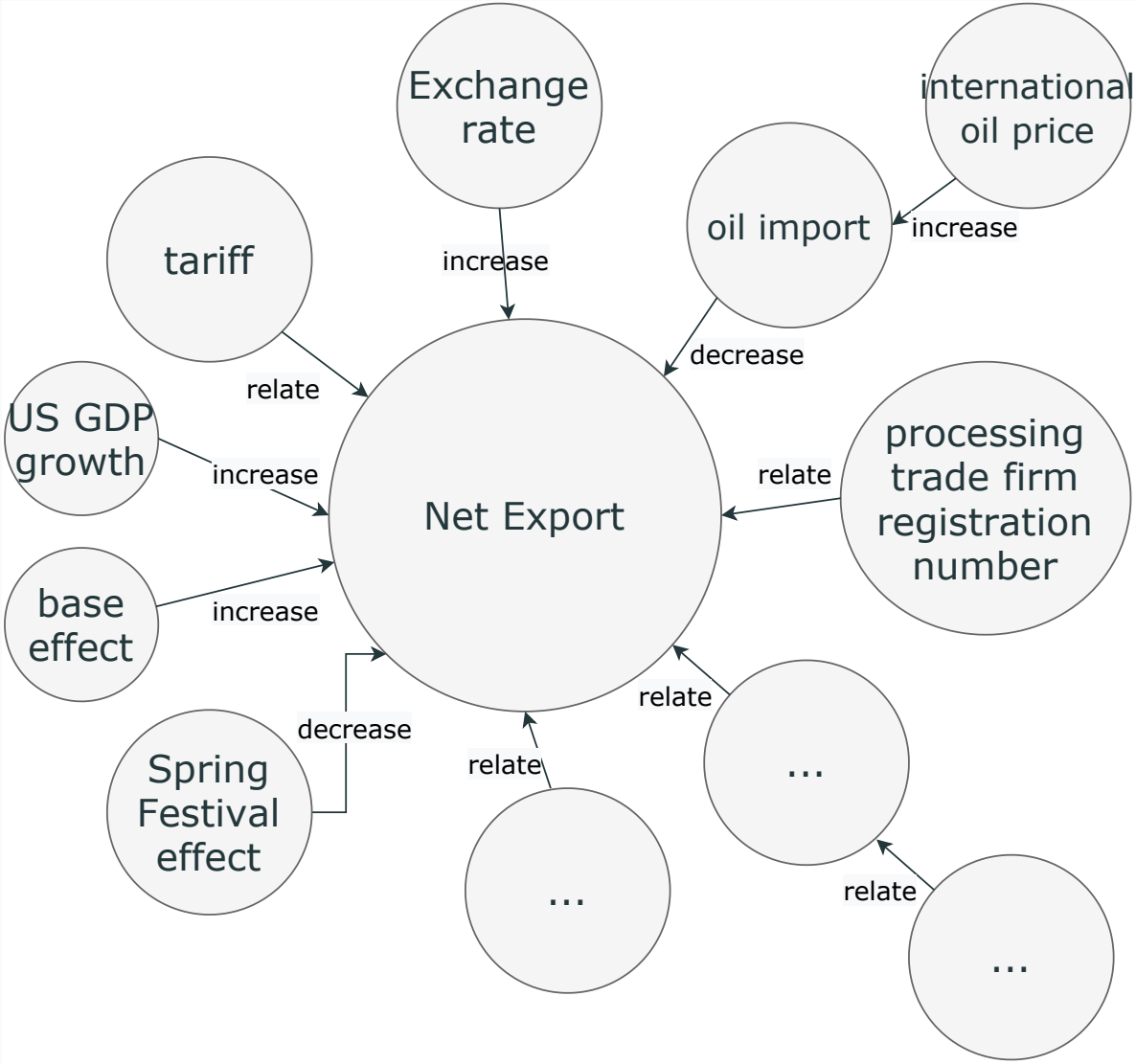


Figure 3: Example of Knowledge Graph on Export

Applications of Knowledge Graph

Application of Knowledge Graph: Economic Forecasting with Many Inputs

We forecast China's monthly *inflation rate* and *nominal investment* time series from April 1996 to June 2019.

For each $i = 1, 2, ..., 12$, we hope to forecast y_{t+i} in i months ahead, with input variables from the past three months $\{\mathbf{X}_s\}_{s=t-3}^{t}$:

$$y_{t+i} = f(\{\mathbf{X}_s\}_{s=t-3}^{t})$$

- Baseline model: standard time series (Higgins and Zha, 2015) as model inputs + statistical method (Lasso).
- KG-based model: model inputs guided by the knowledge graph + Lasso.

Robustness: different lags, replace Lasso with random forest, gradient boosting, etc.

Baseline vs KG: Model Inputs for Inflation Forecasting

- Baseline model: 12 standard time series ($\times$ lags).
  - Real GDP, Nominal Investment, Nominal Consumption, M2, Nominal Imports, Nominal Exports, 7-Day Repo, Benchmark 1-year Deposit Rate, Nominal GDP, GDP Deflator, CPI, Investment Price.
- KG-based model: guided by KG, collect as many as possible (25$\times$ lags).
  - CPI, GDP, benchmark 1-year deposit interest rate, benchmark 1-year loan interest rate, nationwide fiscal expenditure, urbanization rate, central government fiscal expenditure, share of manufacturing output in GDP, urban unemployment rate, worldwide GDP growth rate, M1 money supply, M2 money supply, USD/RMB exchange rate, crude oil production, raw coal production, copper production, raw coal production, Non-ferrous metal production, OPEC Basket Price, crude oil import amount, raw coal import amount, copper import amount, steel import amount, Spring Festival dummy, National Day Festival dummy.

Application of Knowledge Graph: Inflation Forecasting

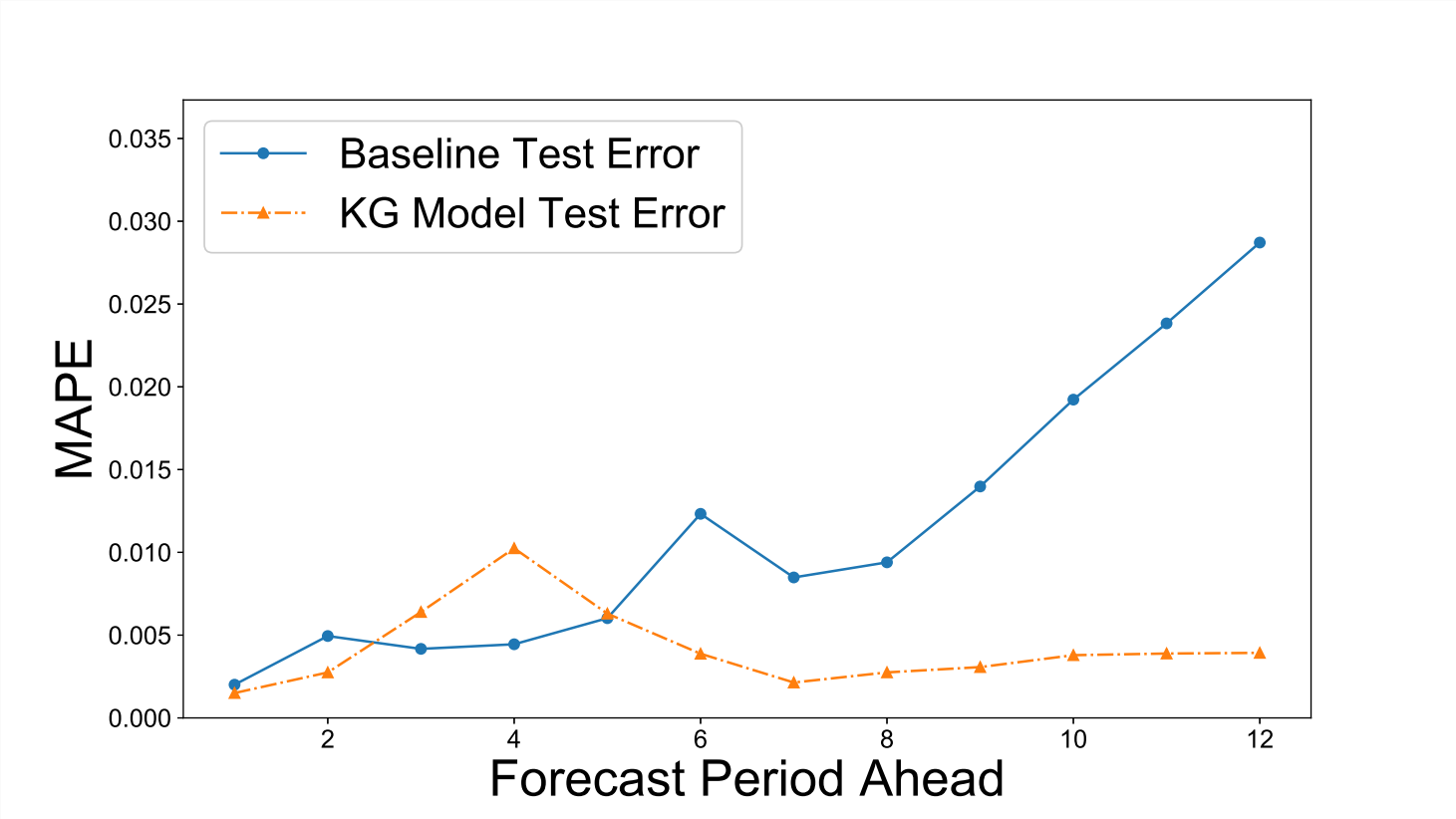


Figure 4: Inflation Forecasting Errors: Baseline vs KG-Based Model

Comparisons are significant under Diebold-Mariano test.

Application of Knowledge Graph: Takeaways

- **Short run**: forecast errors for both models are comparable.
- **Long run**: baseline model gets worse, while the KG-based model achieves a stable and much higher accuracy.
- **Test of comparison**: comparisons are significant under the Diebold-Mariano test.
- **Variable Importance**: baseline model entirely on lagged CPI and GDP Deflator, KG model on lagged CPI and commodity output.
- "**Guaranteed interpretation**": all variables in the KG-based model come from some economic logic narratives.
- "Short term forecasts rely on statistics, long term on logic." KG could better capture underlying logic of the economy than statistical methods on big data.

Application of Knowledge Graph: Investment Forecasting

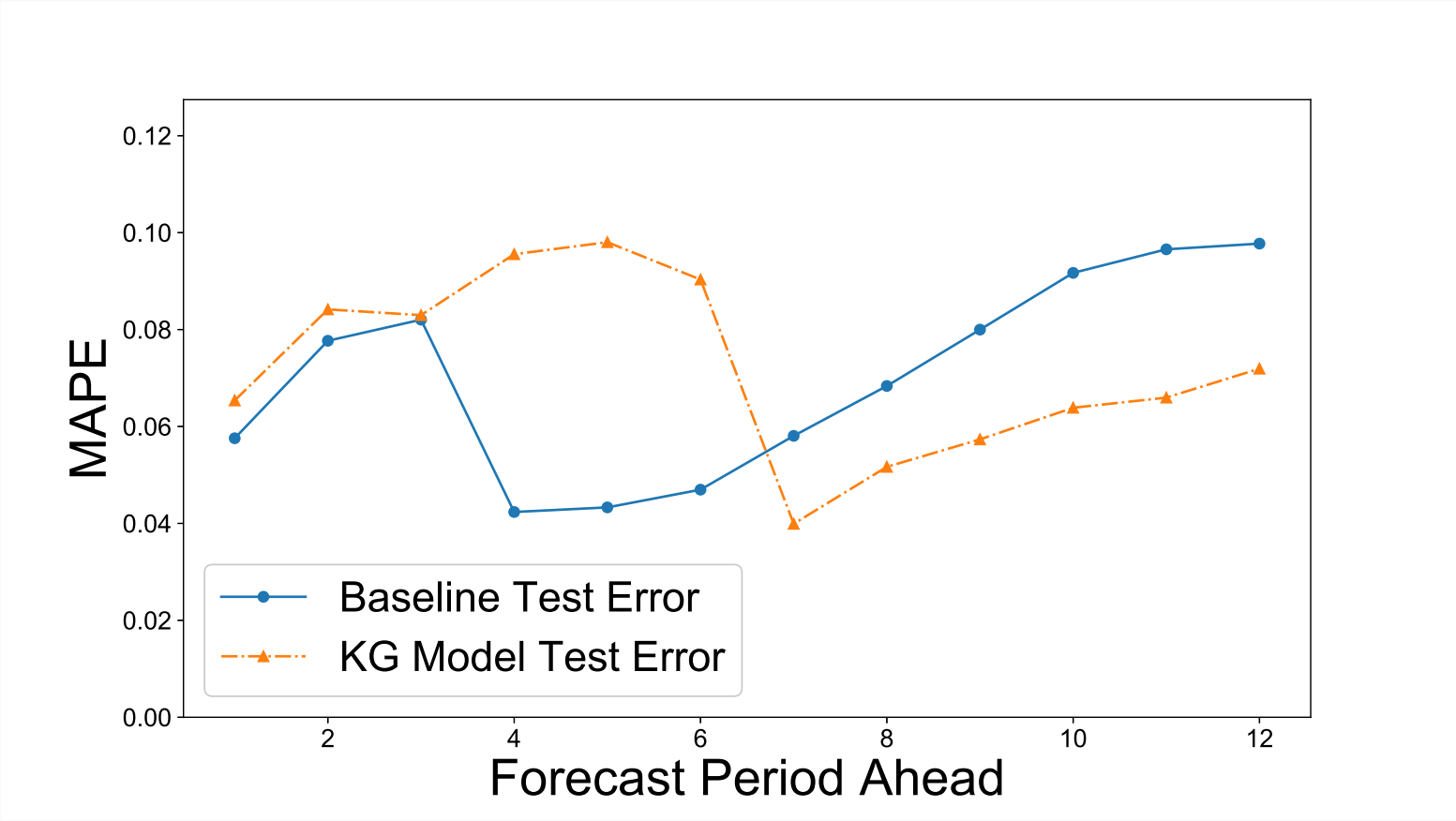


Figure 5: Investment Forecasting Errors: Baseline vs KG-Based Model

Conclusion

- Introduce the concept of knowledge graph to study macroeconomics with (alternative) big data more systematically.
- Develop an active learning NLP algorithm to build KG of relations among traditional and alternative variables from textual data.
- Many exciting applications of new knowledge system! Example: variable selection in economic forecasting.
- Part of our broader agenda of macroeconomics in the age of big data and machine learning.

Link to the Paper

https://ssrn.com/abstract=3707964

Email: yuchengy@princeton.edu

Thanks for your attention!