Google

Compressing Search with Language Models

Tommy Mulc & Jennifer Steele

October 2024

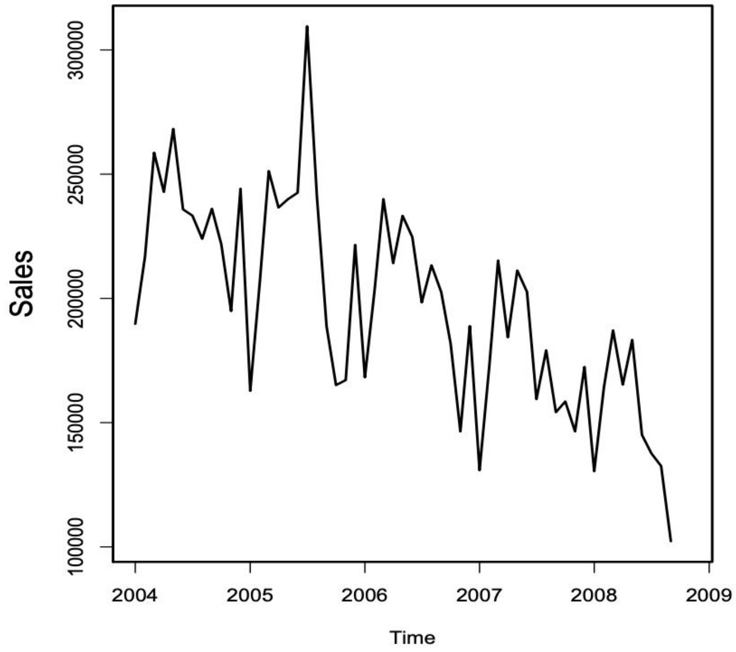
Agenda

Google

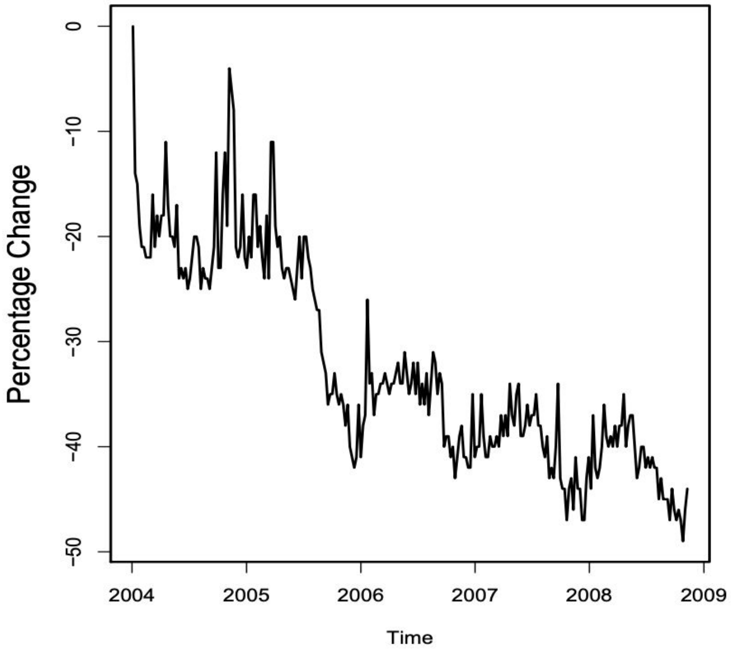
01

Modeling
With Search Data

Predicting the Present with Google Trends [1]



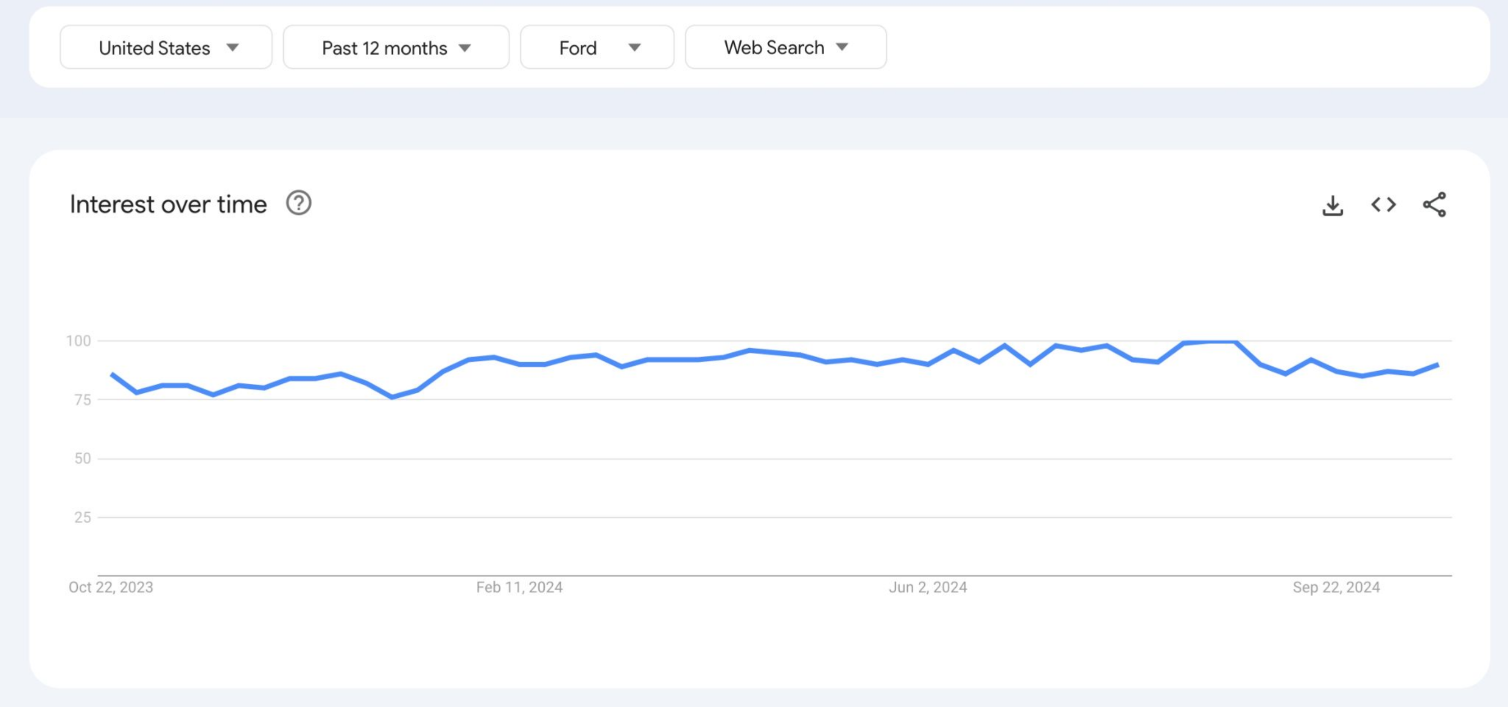
(a) Ford Monthly Sales



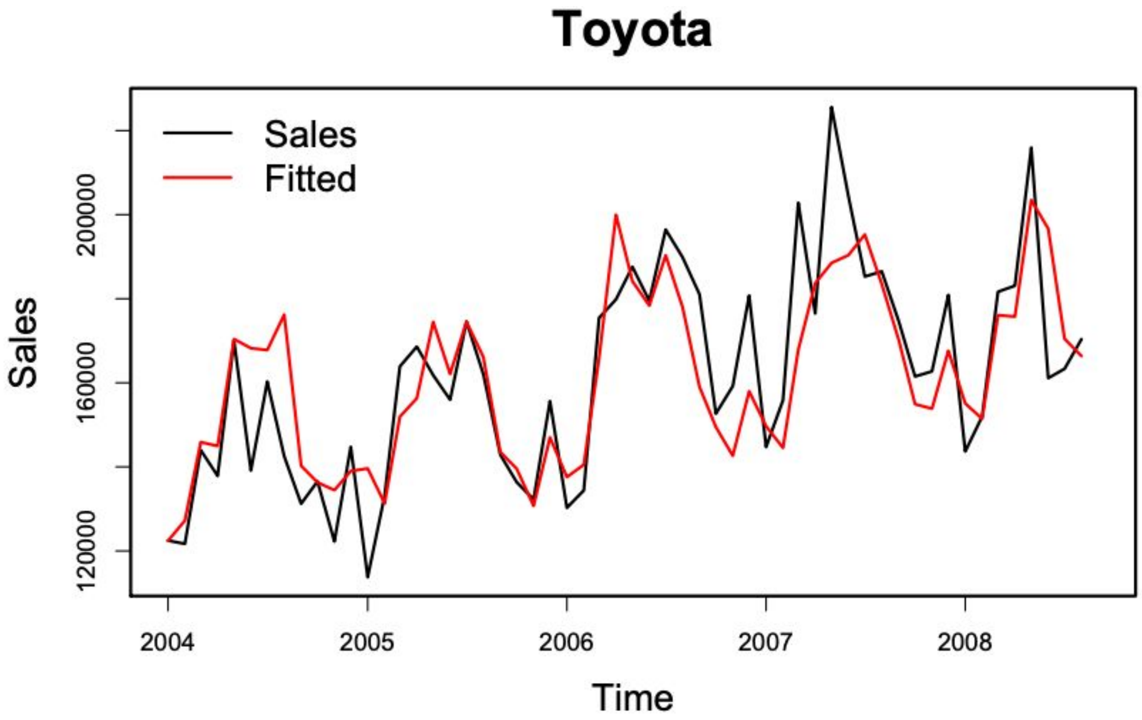
(b) Ford Google Trends

*Great recent lecture here.

Google Trends Data



Predicting the Present with Google Trends



Take Away

Search is a proxy for the marketplace

Useful in many cases where a leading indicator is needed. Search contains the intent behind the queries that often predict the measured variable.

Google

02

Embeddings

Quantifying Search with Embeddings

Search Data

Search Logs*

Term	Volume
craigslist used cars for sale	10
for sale cars craigslist	1500
...	...
awd vs. 4wd	250

* Data is mocked for demonstration purposes.

Embeddings

Features that the model uses to predict the variable of interest.

01

Individual Terms

Simplest way to group terms is to treat each unique term as its own category

02

Categories

Using classification systems, you can group terms together which allow you to expand the total number of included terms.

03

Language Models

Our method uses language models, which are continuous mappings of text to a learned embedding space. The models are pretrained on giant text corpora and capture the meaning of text without relying on human classification boundaries.

Traditional Embeddings

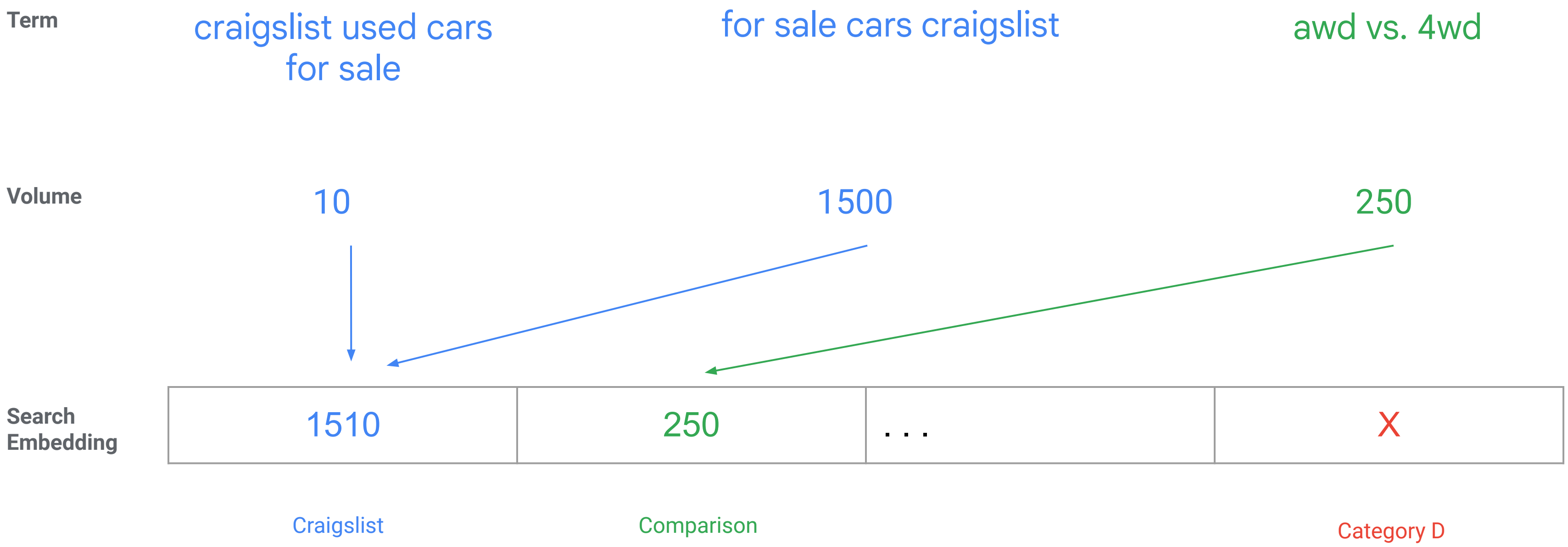
Individual Queries [2]

Term	craigslist used cars for sale	for sale cars craigslist		awd vs. 4wd
Volume	10	1500		250
Search Embedding	10	1500	. . .	250
	1	2		D

D is the number of unique terms to keep in your embedding

Traditional Embeddings

Classification [1]



D is the number of categories

Embeddings as Tables: Individual Terms

Previous methods use discrete embeddings. Individual terms have one-hot rows.

Term	Dim 0	Dim 1	…	Dim D
craigslist used cars for sale	**1**	0	…	0
for sale cars craigslist	0	**1**	…	0
…				
awd vs. 4wd	0	0	…	**1**

Embeddings as Tables: Categorical

Previous methods use discrete embeddings. Queries are mapped to K-hot rows via classification systems.

Term	Dim 0	Dim 1	…	Dim D
craigslist used cars for sale	**1**	0	…	0
for sale cars craigslist	**1**	0	…	0
…				
awd vs. 4wd	0	**1**	…	0

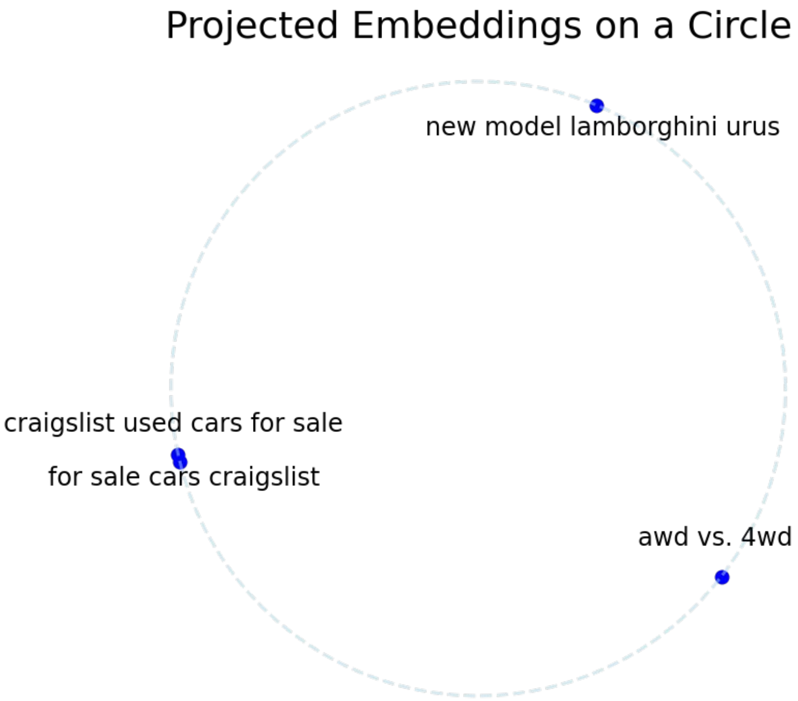
Embeddings as Tables: Language Models

Our method uses LMs to map terms to D-dimensional vectors

Term	Dim 0	Dim 1	…	Dim D
craigslist used cars for sale	0.0411	0.0265	…	0.0131
for sale cars craigslist	0.0217	0.0333	…	0.0201
…				
awd vs. 4wd	0.0086	0.0553	…	0.035

Embeddings Properties

- Synonyms are close together in embeddings spaces
- Not reliant on human classification systems
- Continuous embeddings capture rich information
- Can potentially leverage improvements made from existing language models



Google

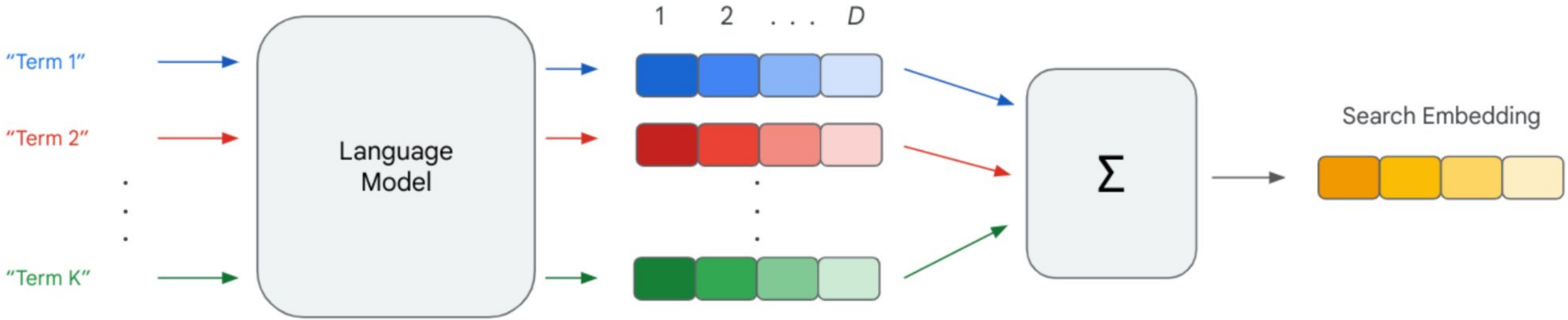
03

Our Approach

SLaM & CoSMo

SLaM Compression

Search Language Model Compression



SLaM Compression

CoSMo

Using a pretrained language model, SLaM compression embeds each term into a fixed-length vector before aggregating all the terms using a reduce operation.

The resulting data is called the "search embedding." It's a highly compressed summary of search that has been shown to be highly predictive.

A custom neural network, CoSMo, uses only the search embedding to learn the important aspects of search data.

Reduce operation is the sum in in the standard setup.

Google

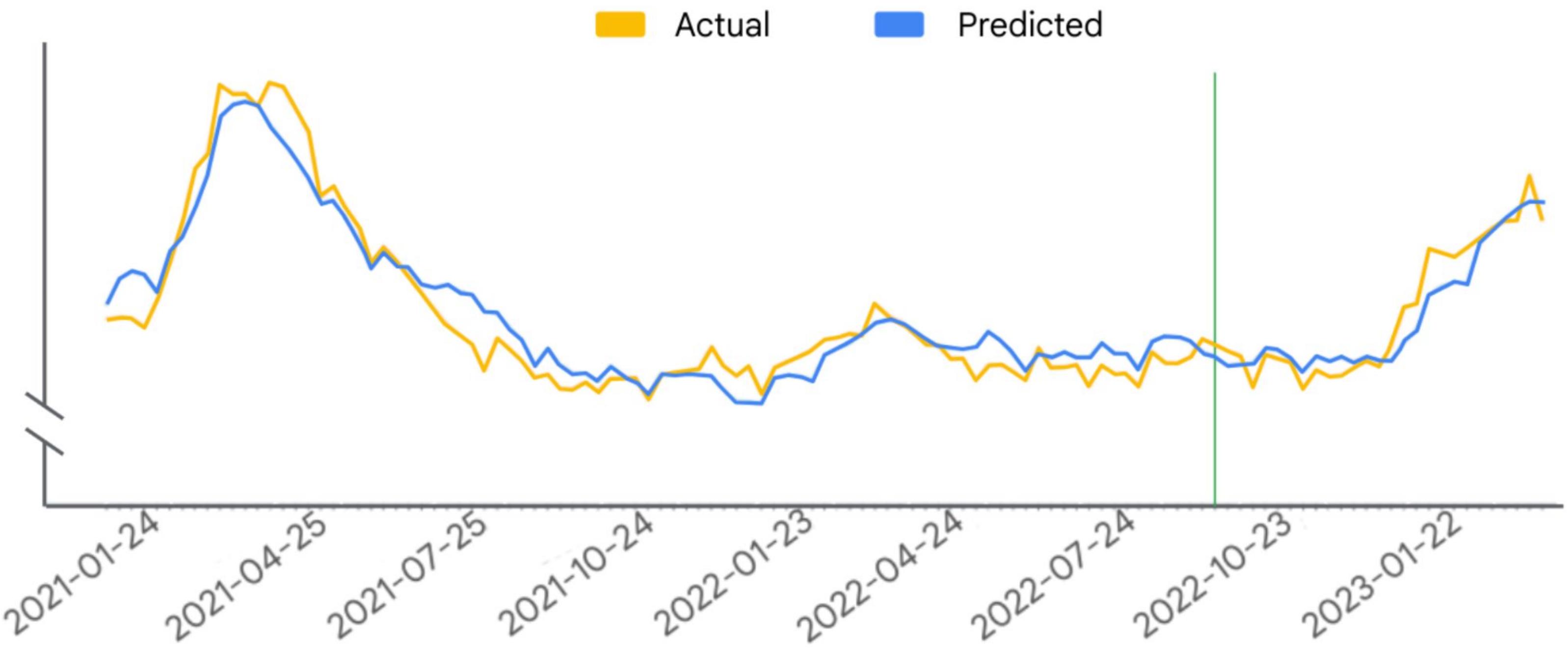
04

Results
Auto & Flu

Model

Automotive Results

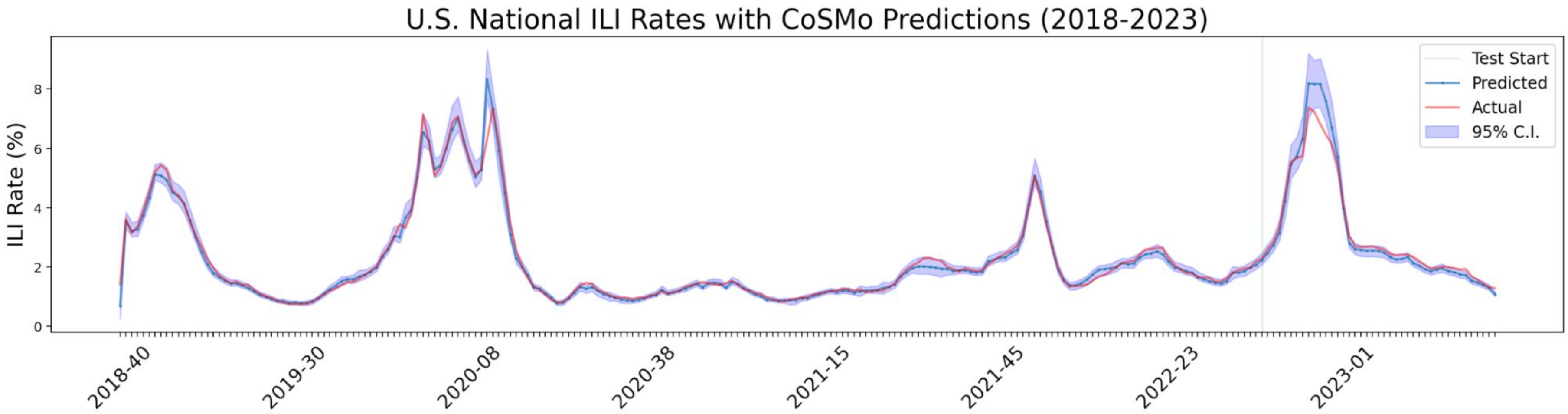
Test period after the green line



Model is **not** autoregressive.

Flu Results

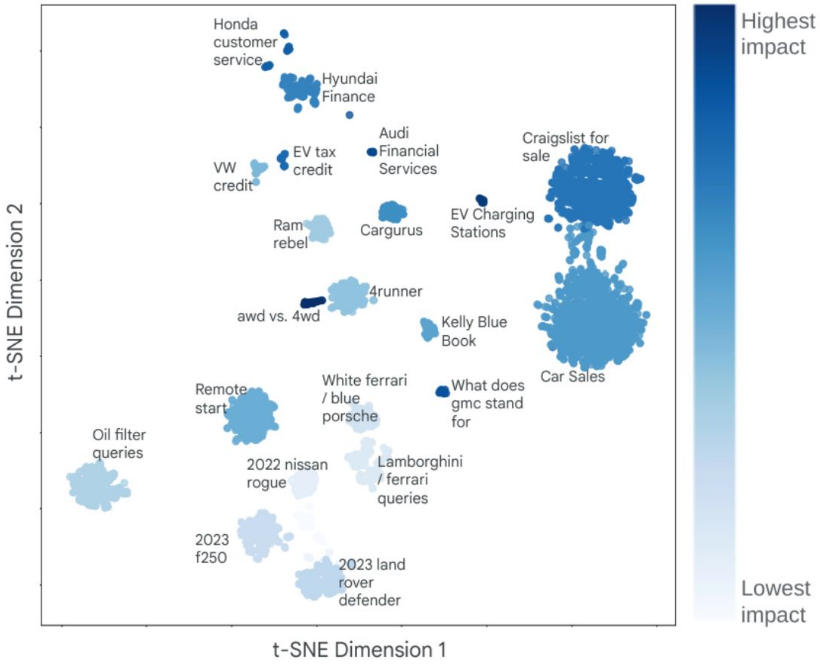
Test period after the green line



Model is **not** autoregressive.

Model Interpretation

Automotive Model

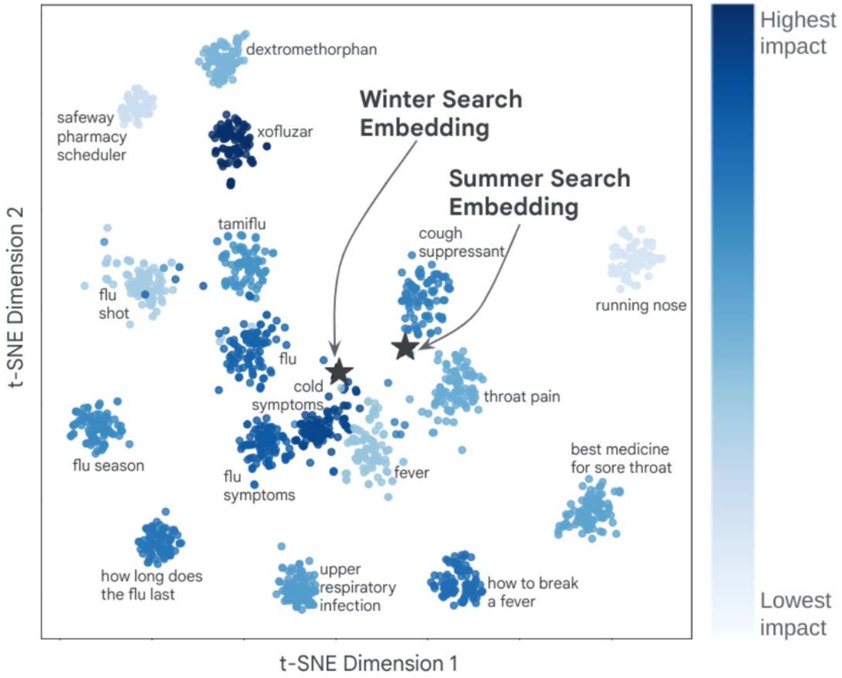


Only a sample of the query clusters are displayed here..

Model

Model Interpretation

Flu Model



Term	Score Percentile
efluenza	0.04
flu center	0.09
uniflu tablets	0.17
summer flu ohio	0.26
flu current status	0.86
flu vactination	0.95
flu cdc recommendations	1.30
flu & pneumonia	1.74
benzonate	38.51
do 5 year olds get immunizations	60.76
kansas vaccine locations	86.79
does putting vicks on your chest help	86.78
nasal spray toddlers	95.48
runny nose puffy eyes	99.90
post nasal drip spitting	99.90

Only a sample of the query clusters are displayed here..

Thank You!

References

[1]Hal R. Varian and Hyunyoung Choi. 2009. Predicting the Present with Google Trends. ssrn.com (2009). http://dx.doi.org/10.2139/ssrn.1659302

[2] J. Ginsberg, M. Mohebbi, and R. et al. Patel. 2009. Detecting influenza epidemics using search engine query data. Nature 457 (2009), 1012–1014

[3] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019. Multilingual universal sentence encoder for semantic retrieval. arXiv preprint arXiv:1907.04307 (2019).

[4] CDC. 2024. https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html

[5] N. Woloszko. 2020. Tracking activity in real time with Google Trends. OECD Economics Department Working Papers 1634 (2020). https://dx.doi.org/10.1787/ 6b9c7518-en

Appendix

Auto Results

Baselines

Frequency	Embedding	Model	Test $R^2 \uparrow$	Test MAPE (%) $\downarrow$
Weekly	Categorical	Lasso	0.5869	10.90
Weekly	Categorical	CoSMo	0.5381	10.85
Weekly	SLaM	CoSMo	**0.7486**	**7.12**
Monthly	SLaM	CoSMo	**0.9065**	**3.03**

Flu Results

Baselines

	Test MAPE(%) ↓	Test r ↑
Logistic Regression	24.9 ± 0.1	.98
MLP	7.3 ± 1.5	.99
Google Flu Trends [17]	[9.5 - 33.1]	[.66 - .97]
Elastic Net [17]	[9.8 - 15.1]	[.92 - .99]
Guassian Process [17]	[9.4 - 14.6]	[.94 - .99]
AR [17]	[6.7 - 14.3]	[.88 - .98]
AR+Google Flu Trends [17]	[6.2 - 12.5]	[.88 - .99]
AR+Elastic Net [17]	[5.1 - 8.7]	[.93 - ≈ 1]
AR+Guassian Process [17]	[5.0 - 8.6]	[.93 - ≈ 1]
CoSMo (Ours)	**5.5 ± 0.4**	**.99**
CoSMo (Ours, Test selection)	**3.9 ± 0.1**	**≈1**

Flu Results

LMs

	Test MAPE(%)↓	Test r ↑
MLSE (baseline)	5.46 ± 0.43	.9933 ± .0005
sT5 Base	6.51 ± 0.13	.9906 ± .0016
sT5 Large	6.51 ± 0.97	.9894 ± .0049
MLSE (English only)	9.11 ± 0.99	.9902 ± .0014
sT5 Base (English only)	7.69 ± 1.29	.9846 ± .0013
sT5 Large (English only)	7.22 ± 0.69	.9878 ± .0040

Flu Results

Other Aggregation Techniques

Aggregation Method	Test MAPE (%)	Test r
Summation (1)	5.46 ± 0.43	$.9933 \pm .0005$
Marginal Distribution	6.86 ± 0.35	$.9952 \pm .0004$

CoSMo

Constrained Search Model