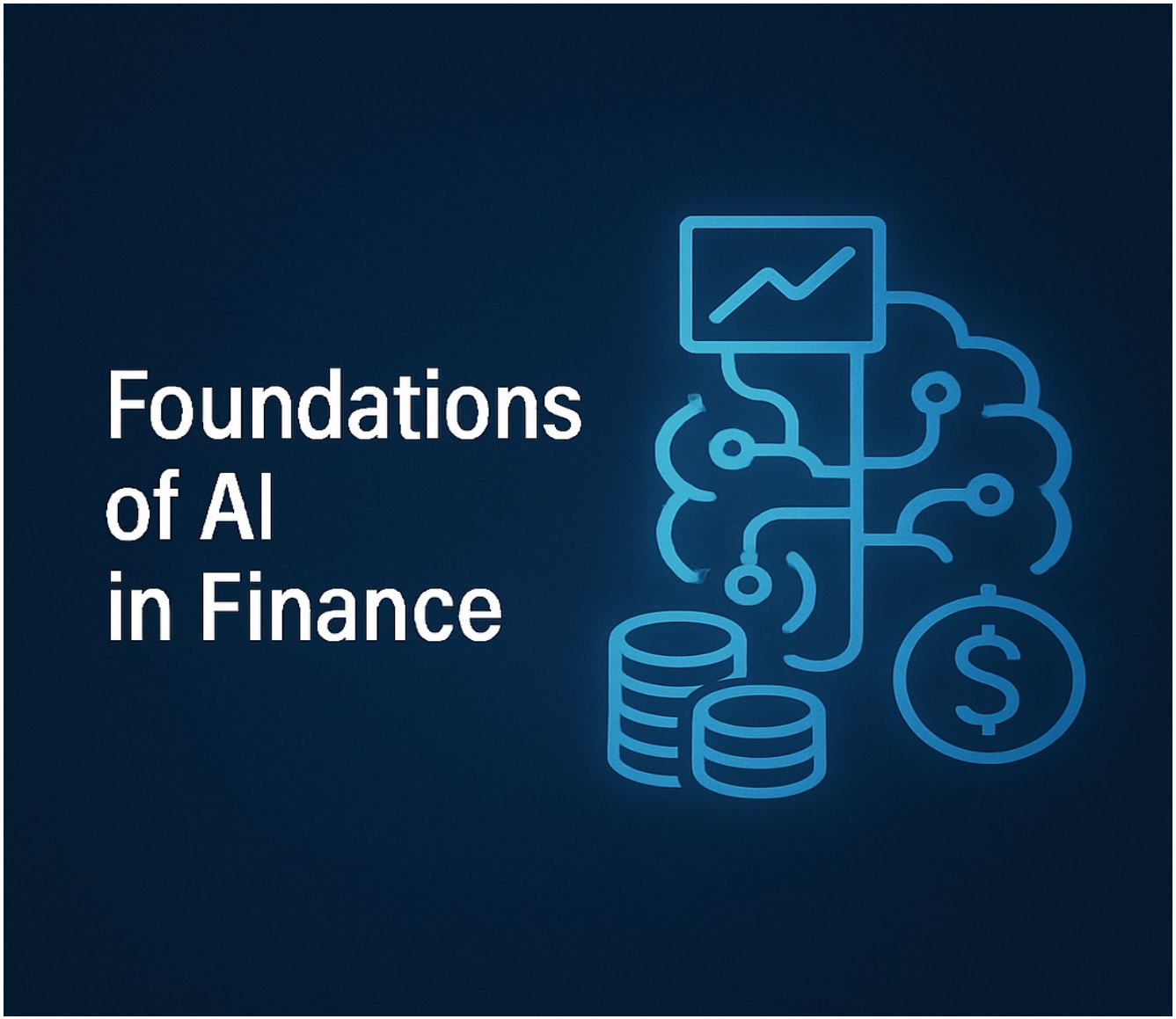
AI in Finance

Arpit Gupta (NYU Stern)
Spring 2026
Session 1



Course Highlights

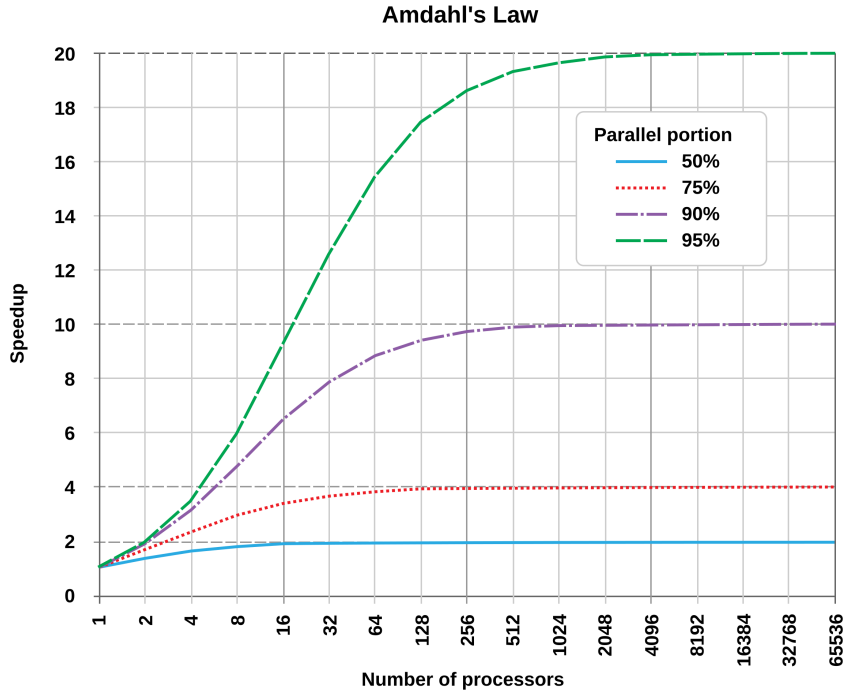
Key Themes

1. Fix the Slow Part
2. Turn Insight into Action
3. Follow the Price

1. Fix the Slow Part

- AI can only move so fast as the slowest part of the system it's in
- This means it's critical to identify the true bottleneck; whether that is in the data pipelines, internal controls, incentive structures, or regulation
- **Amdahl's Law**: "the overall performance improvement gained by optimizing a single part of a system is limited by the fraction of time that the improved part is actually used"
- This means the right unit of analysis isn't the AI tool itself, but the full system it operates in: Tasks -> jobs -> organizations -> markets
- So what?
  - We need to map where the actual bottlenecks and adoption frictions arise within systems
  - Ask: who owns the biding constraint, and is it something AI can even fix?

System Speedup is Limited by How Large the Task is

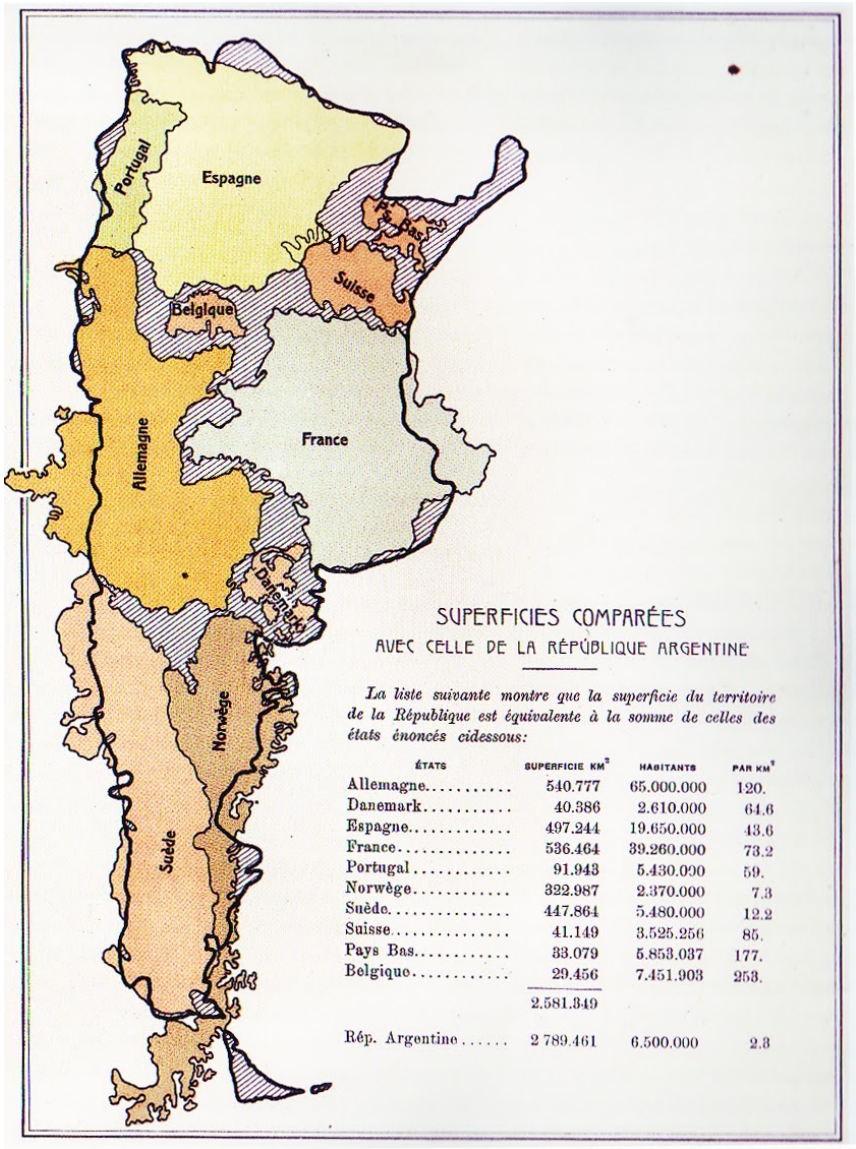


2. Turn Insight into Action

- The core of Gen AI is building **compressed maps**, or lower-dimensional representations of the world
  - We do this to simplify the world into a quantitative object that we can make sense of
  - This is harder to do, but potentially more valuable, the more complex is reality
- The value of this comes when we move to **causal reasoning** and **action** and being explicit about costs and risks
  - Going beyond model evaluation as accuracy to consider:
    - Cost-latency-quality tradeoffs
    - Costs of inaction vs costs of flawed action

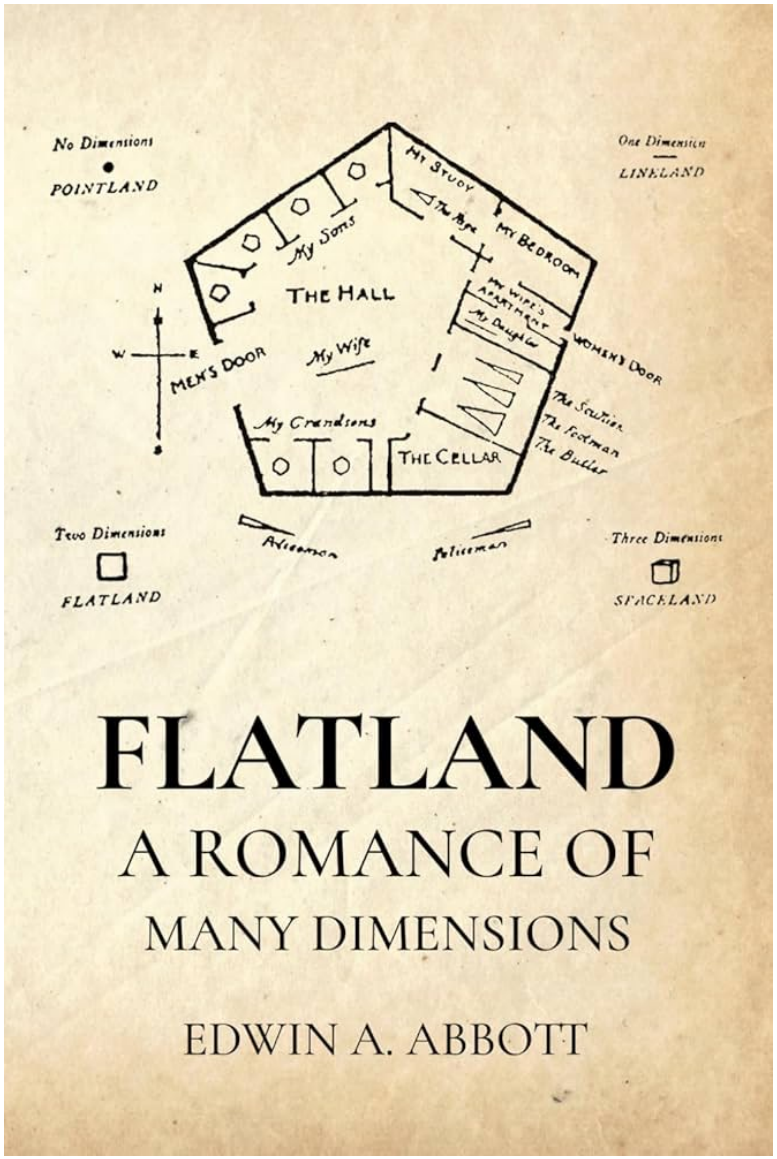
Representational Maps

"In this empire, the art of cartography was taken to such a peak of perfection that the map of a single province took up an entire city and the map of the empire, an entire province. In time, these oversize maps outlived their usefulness and the college of cartographers drew a map of the empire equal in format to the empire itself, coinciding with it point by point. The following generations, less obsessed with the study of cartography, decided that this overblown map was useless and somewhat impiously abandoned it to the tender mercies of the sun and seasons. There are still some remains of this map in the western desert, though in very poor shape, the abode of beasts and beggars. No other traces of the geographical disciplines are to be seen throughout the land."
— Borges, *A Universal History of Infamy*

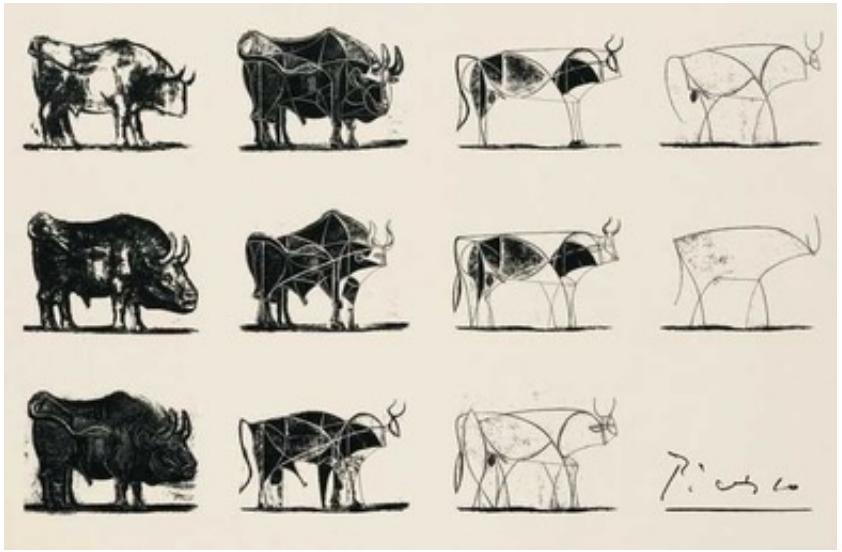


Dimensional Representation

- “Either this is madness or it is Hell.” “It is neither,” calmly replied the voice of the Sphere, “it is Knowledge; it is Three Dimensions: open your eye once again and try to look steadily.”
  — Edwin A. Abbott, Flatland: A Romance of Many Dimensions

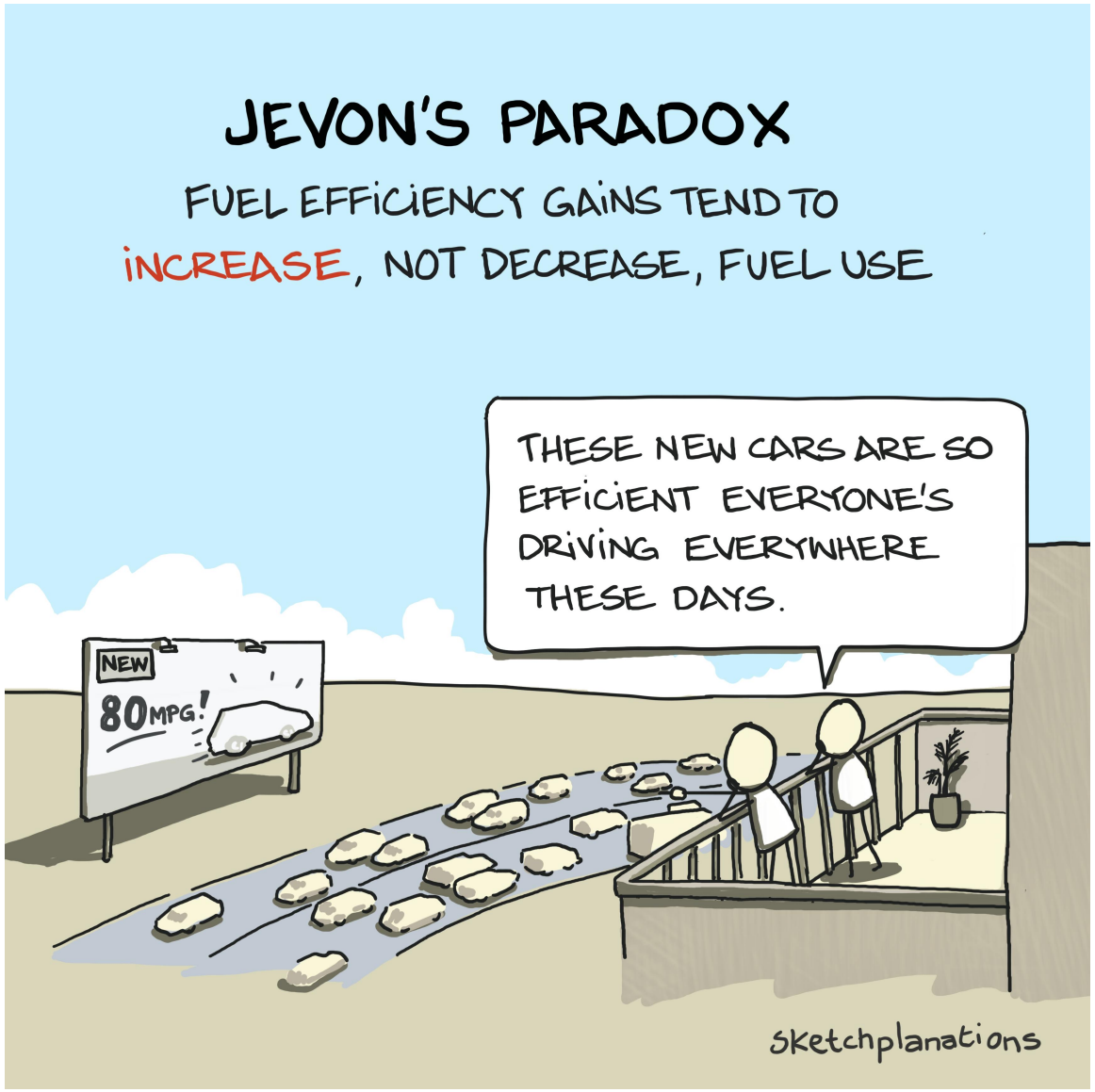


Other Abstractions



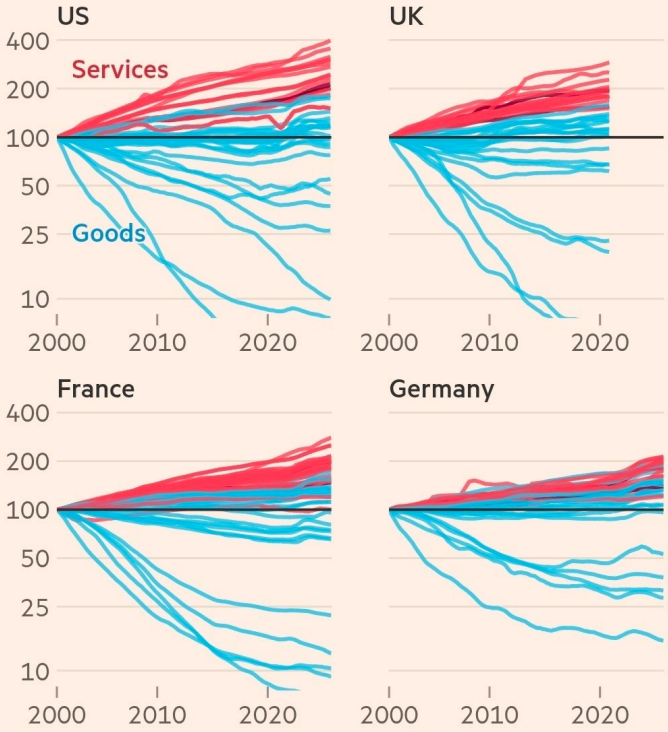
3. Follow the Price

- Automation drives **price adjustment**, as costs with near-zero marginal costs get cheap
  - This drives down the cost of substitutes, and drives up the cost of (scarce) complements
- Therefore **elasticity** and **complementarity** determine how jobs are redesigned and where rent flows along the model chain
  - When demand is elastic, automation **expands** the pie (**Jevon's paradox** — greater efficiency can increase consumption)
  - When demand is inelastic, AI can **substitute**


JEVON'S PARADOX
FUEL EFFICIENCY GAINS TEND TO
INCREASE, NOT DECREASE, FUEL USE
THESE NEW CARS ARE SO EFFICIENT EVERYONE'S DRIVING EVERYWHERE THESE DAYS.
NEW
80MPG!
sketchplanations

Costs of labour-intensive services have soared in high-income countries, while tradable goods have plummeted in price

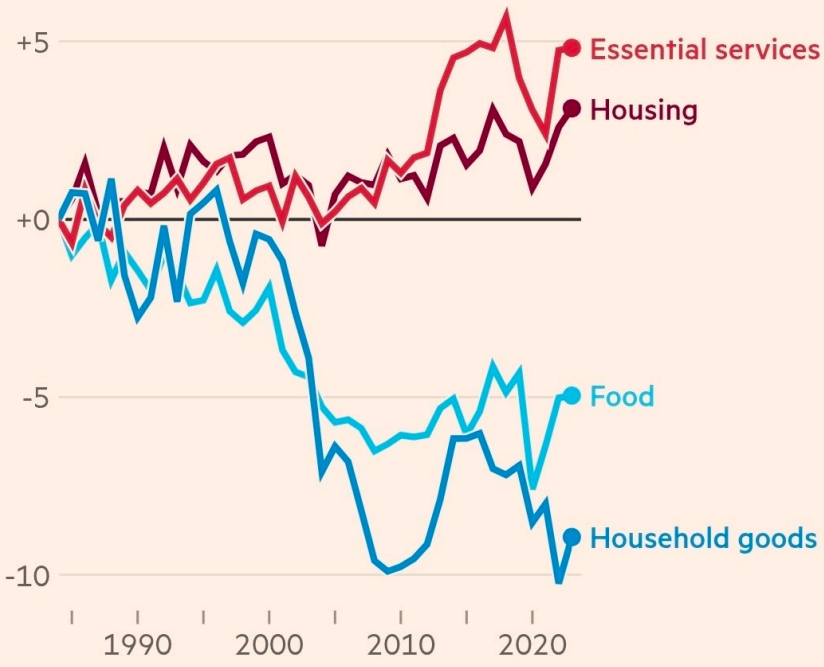
Relative change in price for different items (100 = price in year 2000). Select a line to see details



Source: FT analysis of BLS CPI and Eurostat Harmonised Indices of Consumer Prices
FT graphic: John Burn-Murdoch / @jburnmurdoch
©FT

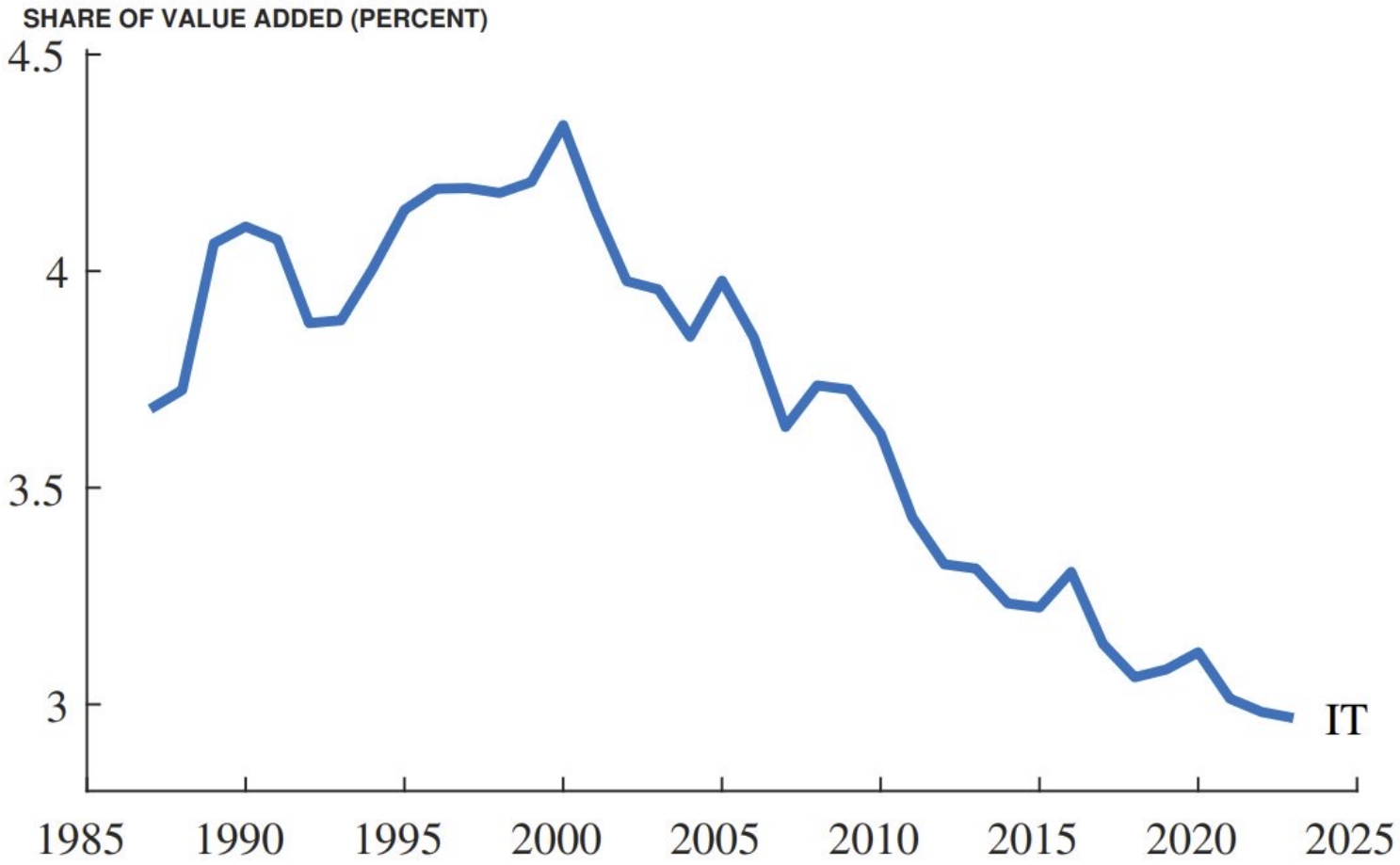
Housing, childcare and healthcare are squeezing incomes, but this has been more than offset by falling costs of goods

Percentage-point change in share of median US household income spent on different categories



Source: FT analysis of BLS Consumer Expenditure Survey
FT graphic: John Burn-Murdoch / @jburnmurdoch
©FT

Figure 2: The Share of Factor Income Paid to Computers



Jones Tonetti 2026 “Past Automation and Future A.I.: How Weak Links Tame the Growth Explosion”

What's the Takeaway?

- When we want to improve the workflow for an entire system, we need to track the relevant frictions and pain points
- The decision-making process should start with the AI simplification, and
  - Fix the bottleneck
  - Simplify the world to make decisions
  - Automate the easy stuff and focus value-add on harder to replace substitutes

The Road to Gen AI

Statistical Science
2001, Vol. 16, No. 3, 199–231

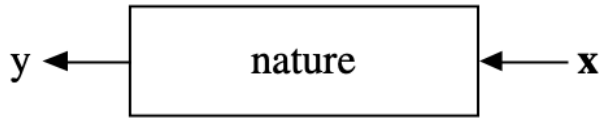
Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:



There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

There are two different approaches toward these goals:

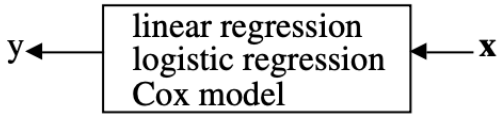
The Two Cultures

The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

$$\text{response variables} = f(\text{predictor variables, random noise, parameters})$$

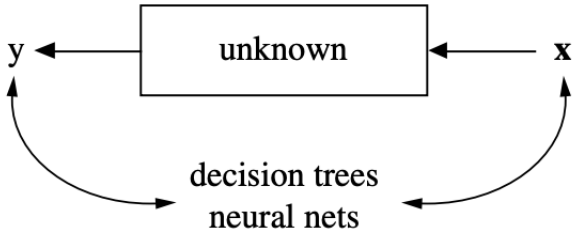
The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on $\mathbf{x}$ to predict the responses $\mathbf{y}$. Their black box looks like this:



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

The Bitter Lesson

Rich Sutton

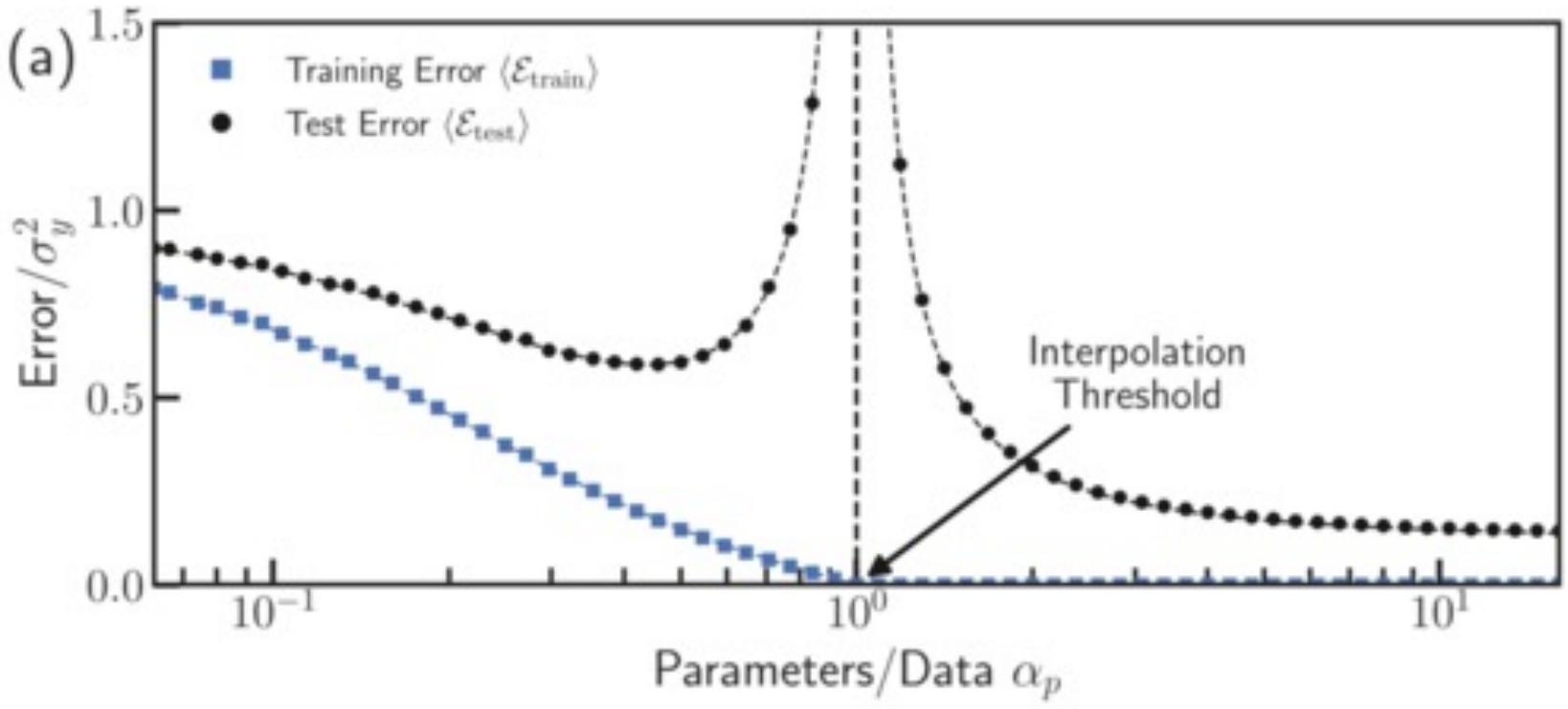
March 13, 2019

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant (in which case leveraging human knowledge would be one of the only ways to improve performance) but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available. Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation. These two need not run counter to each other, but in practice they tend to. Time spent on one is time not spent on the other. There are psychological commitments to investment in one approach or the other. And the human-knowledge approach tends to complicate methods in ways that make them less suited to taking advantage of general methods leveraging computation. There were many examples of AI researchers' belated learning of this bitter lesson, and it is instructive to review some of the most prominent.

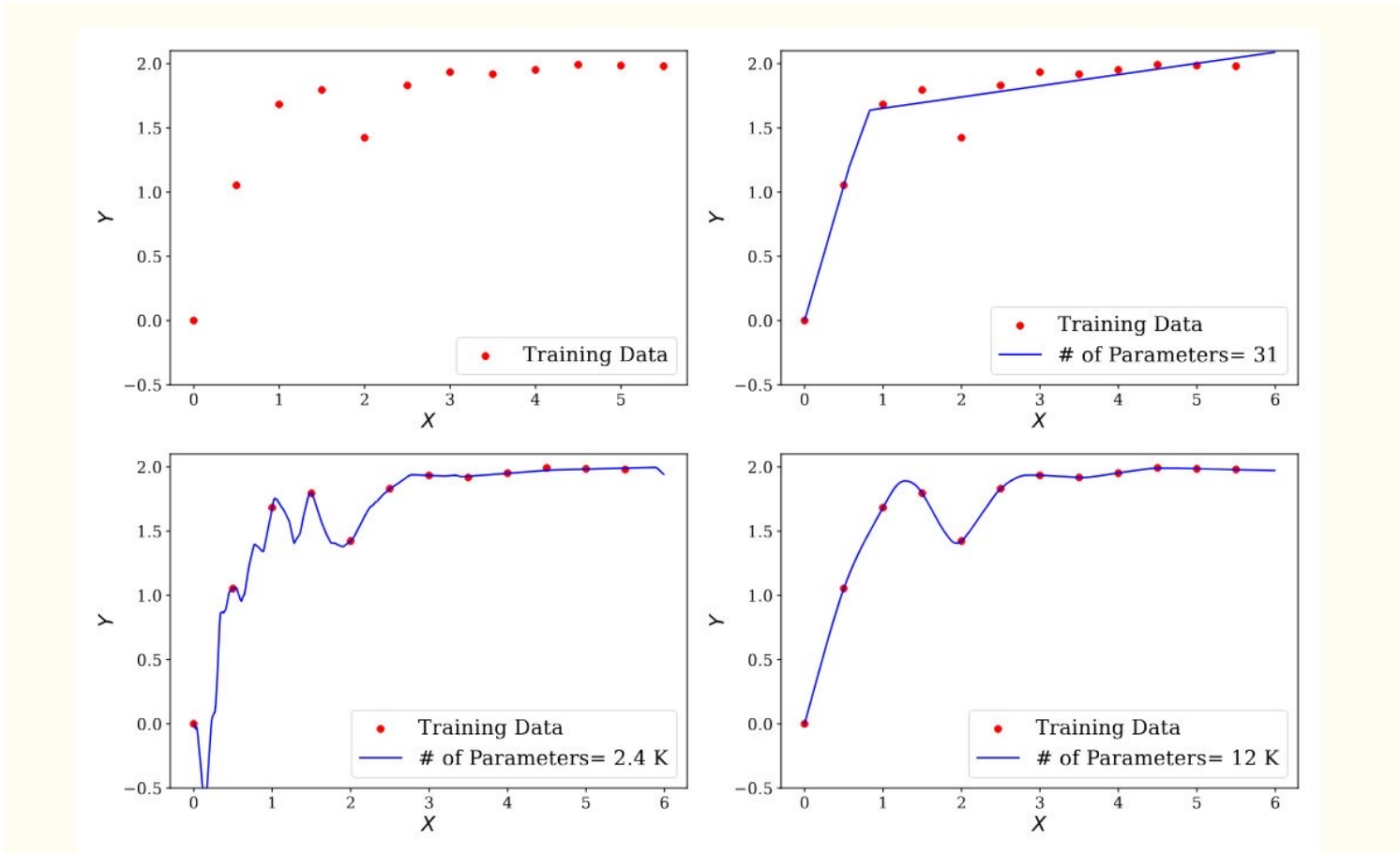
Prediction vs Causality

- Traditional approaches have favored **causality** and **legibility**
  - I.e., you first write down a concrete model you have in mind (i.e., CAPM)
  - Then you test the model on data, taking into account biases
    - "Gold Standard" evidence is RCTs, i.e., an A/B Test
  - Intuitively, we favor **simple rules**
    - These are likely to be robust out of sample (avoid overfitting)
    - Easy to explain (to regulators, customers, etc.)
- New AI approaches primarily focus on **scaling** and **prediction**
  - Learn patterns from huge amounts of data
  - Take advantage of cheaper computational cost
  - Has worked surprisingly well, at the cost of "Black Box" complexity

Double Descent



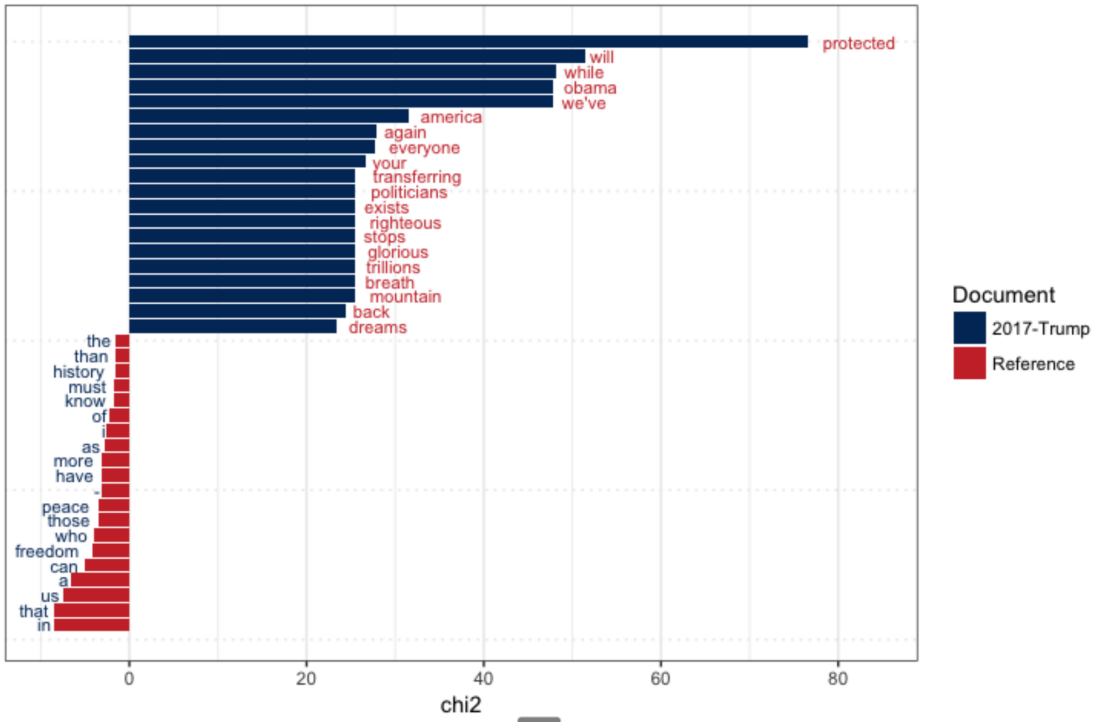
Double Descent



First Generation: Text Dictionaries

- Classify individual words into categories
  - I.e., **self** = *I, me, my, mine, myself*
  - Importantly, create positive and negative word lists to enable sentiment analysis
- Harvard IV-4 dictionary: one historically used classification for financial and accounting texts
- A typical dictionary may have ~182 categories
- Can be used for **bag of words**: count word frequencies, impute sentiment to text
  - I.e., earnings transcript was positive, so buy the stock

Example: Relative Frequency of Words



Problem: Financial Contexts!

Panel B: Fin-Neg

Full 10-K Document				MD&A Subsection			
Word in H4N-Inf	Word	% of Total Fin-Neg Word Count	Cumulative %	Word in H4N-Inf	Word	% of Total Fin-Neg Word Count	Cumulative %
✓	LOSS	9.73%	9.73%	✓	LOSS	9.51%	9.51%
✓	LOSSES	5.67%	15.40%	✓	LOSSES	7.58%	17.10%
	CLAIMS	3.15%	18.55%	✓	IMPAIRMENT	4.71%	21.81%
✓	IMPAIRMENT	3.04%	21.59%		RESTRUCTURING	2.93%	24.74%
✓	AGAINST	2.58%	24.17%	✓	DECLINE	2.89%	27.62%
✓	ADVERSE	2.44%	26.61%		CLAIMS	2.71%	30.33%
	RESTATED	2.09%	28.70%	✓	ADVERSE	2.44%	32.77%
✓	ADVERSELY	1.75%	30.45%	✓	AGAINST	2.01%	34.78%
	RESTRUCTURING	1.72%	32.17%	✓	ADVERSELY	1.94%	36.72%
	LITIGATION	1.67%	33.83%		LITIGATION	1.67%	38.40%
	DISCONTINUED	1.57%	35.40%		CRITICAL	1.63%	40.03%
	TERMINATION	1.35%	36.75%		DISCONTINUED	1.62%	41.64%
✓	DECLINE	1.19%	37.93%	✓	*DECLINED*	1.30%	42.94%
✓	CLOSING	1.08%	39.01%		TERMINATION	1.06%	44.00%
✓	FAILURE	0.97%	39.98%	✓	NEGATIVE	0.96%	44.96%
	UNABLE	0.84%	40.82%	✓	FAILURE	0.93%	45.89%
✓	*DAMAGES*	0.82%	41.64%		UNABLE	0.91%	46.80%

Loughran McDonald 2011

When is a Liability not a Liability?
Textual Analysis, Dictionaries, and 10-Ks

Journal of Finance, forthcoming

TIM LOUGHRAN and BILL MCDONALD

ABSTRACT

Previous research uses negative word counts to measure the tone of a text. We show that word lists developed for other disciplines misclassify common words in financial text. In a large sample of 10-Ks during 1994 to 2008, almost three-fourths of the words identified as negative by the widely used Harvard Dictionary are words typically not considered negative in financial contexts. We develop an alternative negative word list, along with five other word lists, that better reflect tone in financial text. We link the word lists to 10-K filing returns, trading volume, return volatility, fraud, material weakness, and unexpected earnings.

Key words: Textual analysis; Harvard Dictionary; negative word counts; term weighting.

- Existing dictionary approach assigns typical finance words in 10-Ks (“taxes” “liabilities”) as negative, despite context-dependence in Finance (i.e., “liabilities went down”)
- Develop a new finance-specific dictionary

Next Iteration: n-grams

The colour of finance words ☆

Diego García [a], Xiaowen Hu [b], Maximilian Rohrer [c]

Show more

https://doi.org/10.1016/j.jfineco.2022.11.006 ↗ Get rights and conten

Abstract

Our paper relies on stock price reactions to colour words, in order to provide new dictionaries of positive and negative words in a finance context. We extend the machine learning algorithm of Taddy (2013), adding a cross-validation layer to avoid over-fitting. In head-to-head comparisons, our dictionaries outperform the standard bag-of-words approach (Loughran and McDonald, 2011) when predicting stock price movements out-of-sample. By comparing their composition, word-by-word, our method refines and expands the sentiment dictionaries in the literature. The breadth of our dictionaries and their ability to disambiguate words using bigrams both help to colour finance discourse better.

Table 6: Disambiguating the token "demand"

This table presents a subset of the bigrams associated with the token "demand," a total of 228 unique bigrams (using our dtm with 2^{16} terms). The column "Term" lists the bigram. The column "Freq in %" is the relative counts of the bigram out of all the bigrams that contain "demand," i.e. 5.1% of the times "demand" is written, it is within the bigram "strong demand". Tokens marked with a positive/negative sign (blue/red) denote LM positive/negative words.

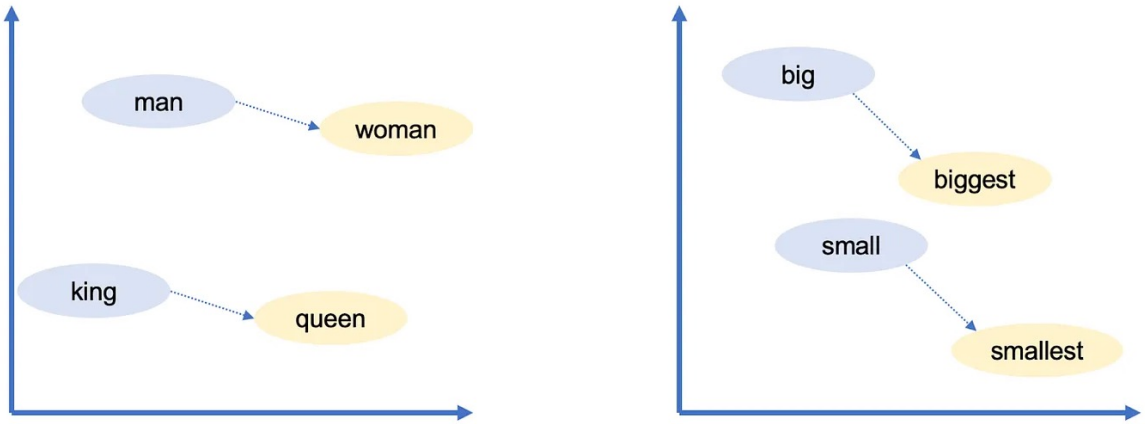
Positive bigrams		**Neutral bigrams**		**Negative bigrams**	
Term	Freq in %	Term	Freq in %	Term	Freq in %
strong^{+}demand	5.1	market demand	2.6	supply demand	1.8
demand products	2.1	customer demand	2.5	lower demand	1.0
increased demand	2.0	demand environment	1.8	demand response	1.0
growing demand	1.0	consumer demand	1.5	demand side	0.8
see demand	1.0	meet demand	1.1	demand market	0.6
demand across	0.9	demand new	0.9	weak^{-}demand	0.5
increase demand	0.9	demand growth	0.9	demand well	0.4
good^{+}demand	0.9	overall demand	0.9	reduced demand	0.4
up demand	0.7	more demand	0.9	weaker^{-}demand	0.4
demand services	0.7	global demand	0.9	current demand	0.4
demand seeing	0.6	increasing demand	0.8	demand generation	0.3
solid demand	0.6	demand trends	0.8	demand drivers	0.3
loan demand	0.5	high demand	0.8	seasonal demand	0.3

Embeddings turn Words to Geometry

- We can extend the idea of an n-gram to a **vector** ("word vector")
- Imagine an object with as many entries as there are words in the English language.
- A simple **embedding** (map from corpus of text to a vector) would count the frequency of each word
- More complicated embeddings are smaller, and preserve meaning so words closer in "embedding space" mean similar things

Embedding Geometry

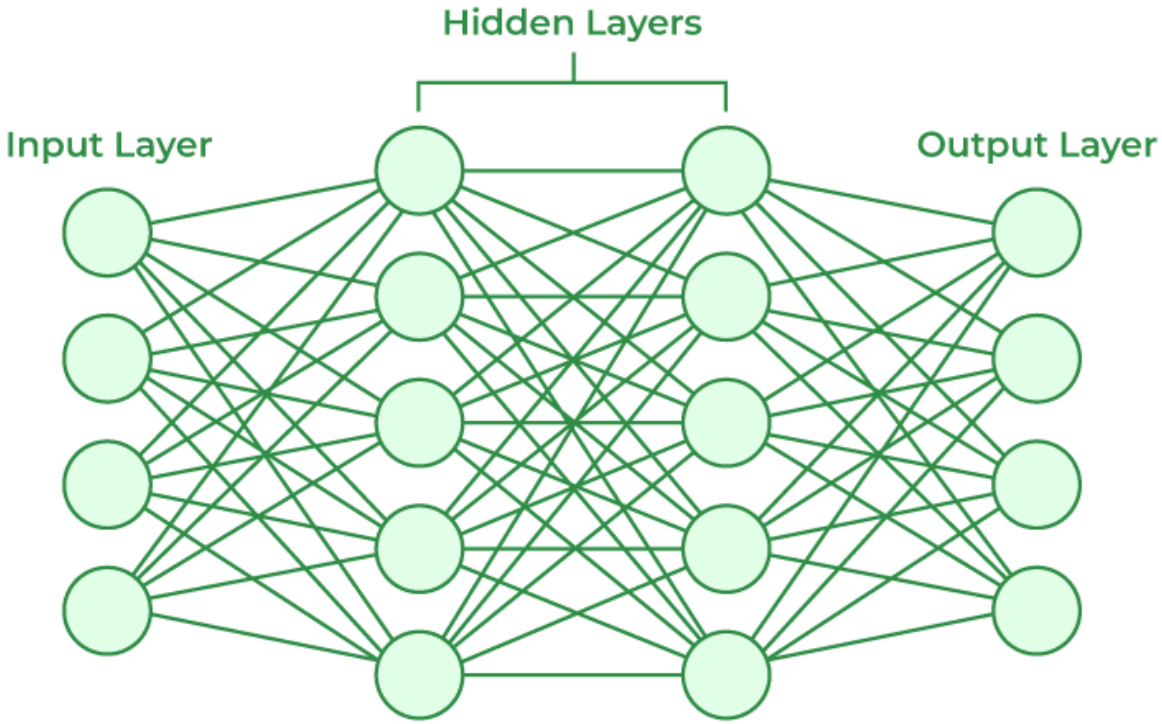
Google's word vectors had another intriguing property: you could "reason" about words using vector arithmetic. For example, Google researchers took the vector for **biggest**, subtracted **big**, and added **small**. The word closest to the resulting vector was **smallest**.



Contextual Embeddings

- Many words are ambiguous (”bank” in river vs. finance)
- Key breakthrough was representing the **same word** with a **different embedding** depending on the context (contextual embeddings)

Neural Net



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

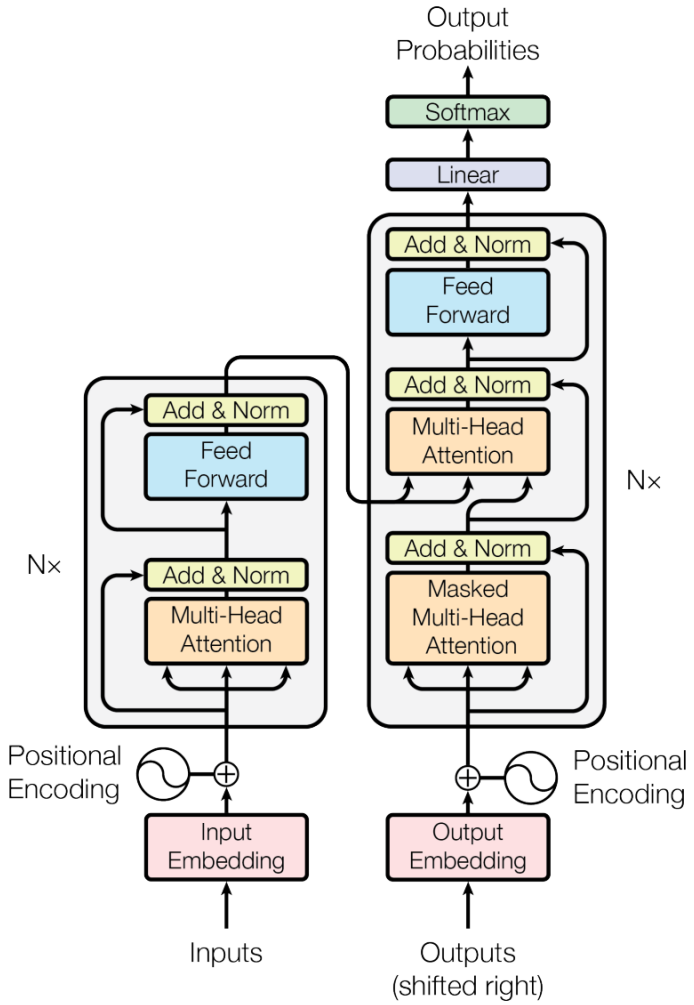
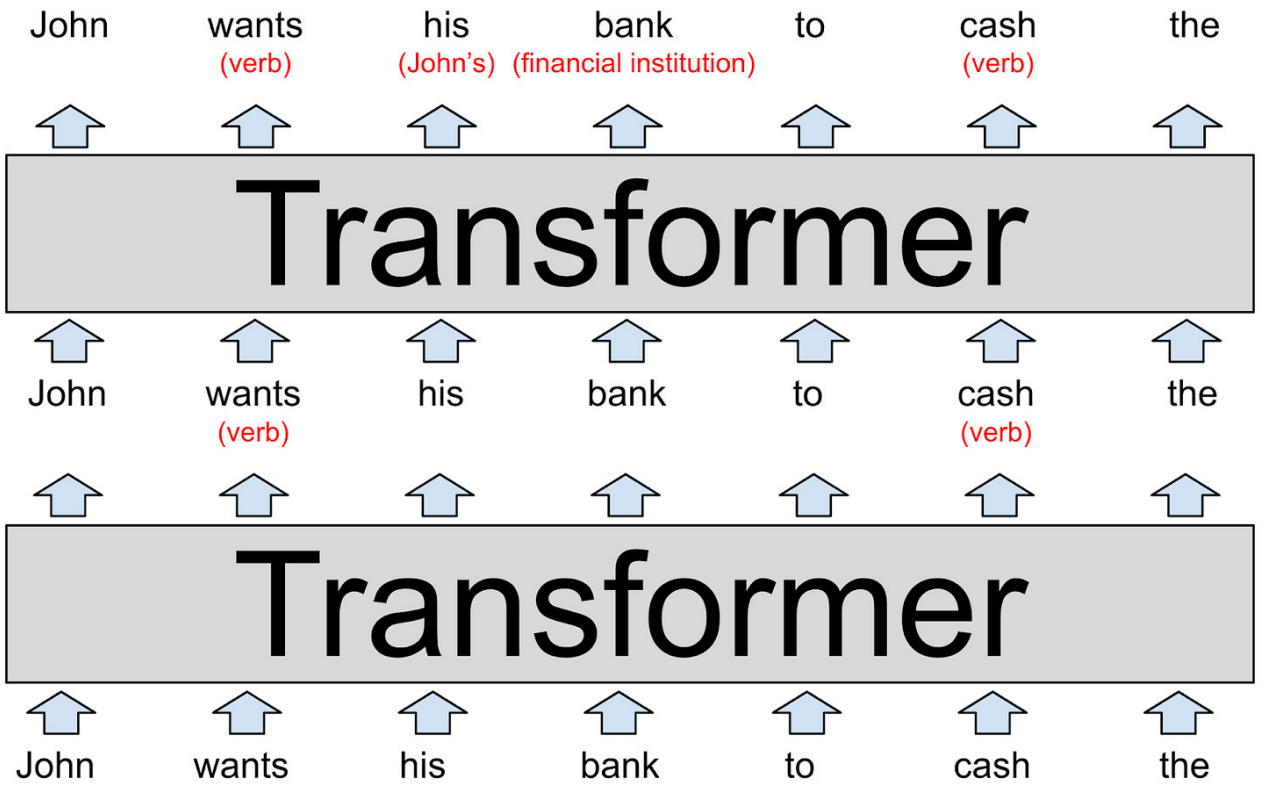


Figure 1: The Transformer - model architecture.

Large Language Models (LLMs)



What's Going on Here?

- An LLM starts with an **input** "John wants his bank to cash the"
- These words, transformed into a vector, are fed into the first **transformer**
  - A transformer is a type of neural net based on an **attention mechanism**
  - A **neural net** is a network of nodes, inspired by the brain, which translates some input into output, as a function of (possibly hidden) layers
  - **Attention** = retrieving information from the prompt
  - **Feed forward** step processes attention gathered, to form higher level predictions

Continuing the Example

- The first transformer layer classifies “wants” and “cash” as verbs
  - These are new “hidden state” passed to the next layer
- A second transformer clarifies that “bank” is a financial institution (not a river), and “his” refers to John
- To do this, the transformers do two important things:
  - The **attention** mechanism “looks around” (”attends to”) other words with relevant context. Decides what matters to understand each word.
  - The **feed-forward** step “thinks” about information gathered, and predicts the next word
- Key innovation is this can be done in **parallel** enabling scale
- GPT-4 has ~1.8b parameters across 120 layers
  - Earlier layers process words and grammar; later layers phrases; deep layers abstract concepts

BERT

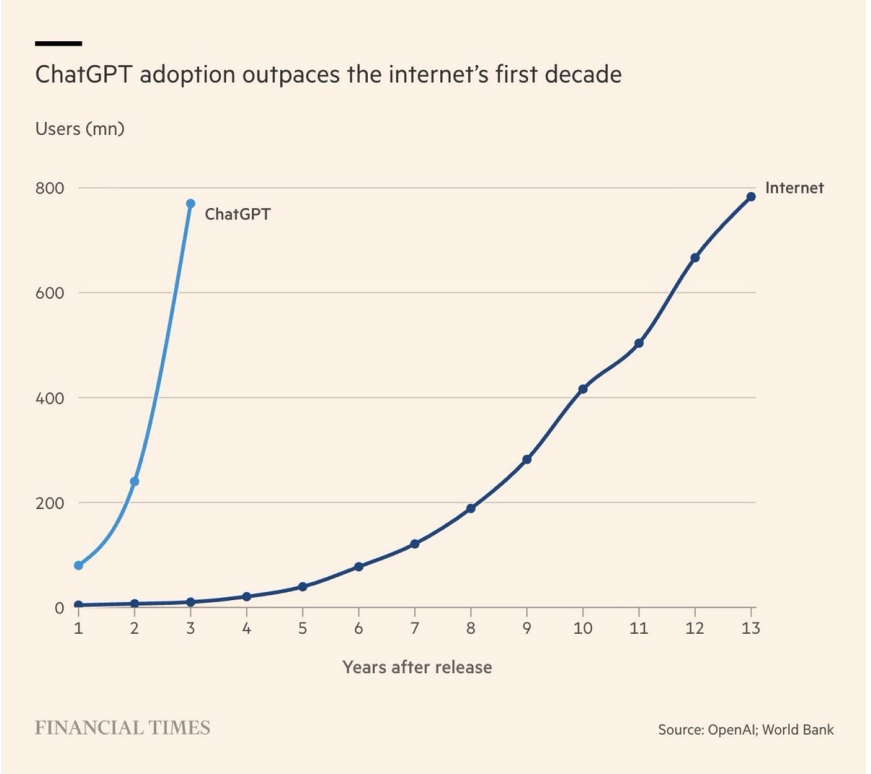
- Trained transformer model, trained on billions of words
  - Ie, “masked” words to predict and verify
- Financial version: FinBERT

FinBERT: Financial Sentiment Analysis with Pre-trained Language Models

Dogu Tan Araci
dogu.araci@student.uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Large Language Models

- GPT-style models have seen **enormous** adoption growth
- **Generative** models which create and process new information (as opposed to existing classification)
- Easy to use interface; bringing AI to the masses



Training

- LLMs train on ordinary text without labels by repeatedly engaging in **next token prediction**
  - Token: “Every moment is a beginning” -> [“Every,” “mo”, “ment”, “is”, “a”, “begin”, “ning”]
- Learning happens when a forward prediction is compared to reality, which adjusts a weight (backprop)
- More parameters + more data + more compute -> better performance, according to existing scaling laws
- Next token prediction is simple, but at scale leads to powerful representations, while still leaving us with something hard to interpret

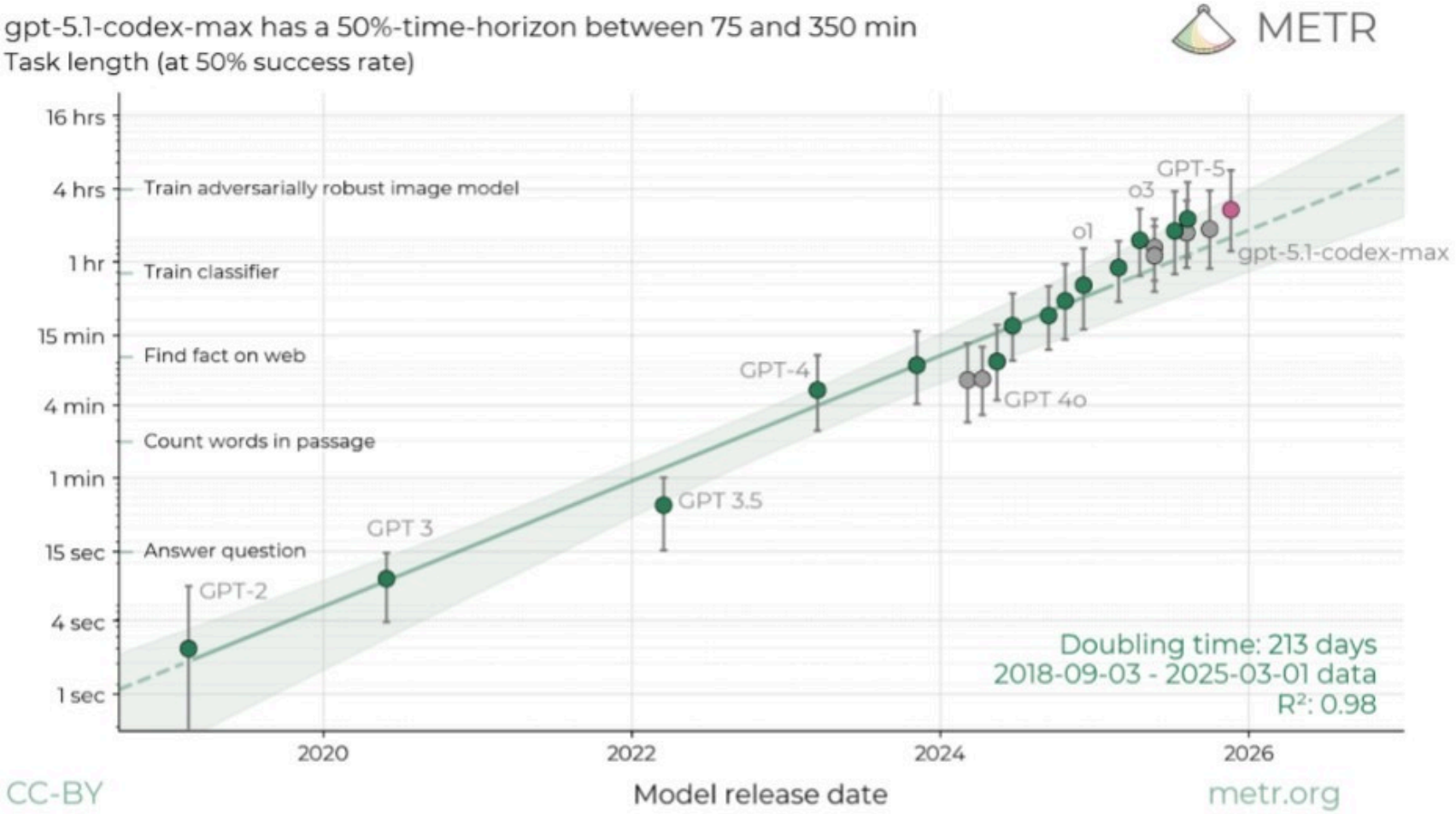

gpt-5.1-codex-max has a 50%-time-horizon between 75 and 350 min
Task length (at 50% success rate)
METR
16 hrs
4 hrs
1 hr
15 min
4 min
1 min
15 sec
4 sec
1 sec
Train adversarially robust image model
Train classifier
Find fact on web
Count words in passage
Answer question
GPT-2
GPT 3
GPT 3.5
GPT-4
GPT 4o
o1
o3
GPT-5
gpt-5.1-codex-max
Doubling time: 213 days
2018-09-03 - 2025-03-01 data
R²: 0.98
2020
2022
2024
2026
Model release date
CC-BY
metr.org

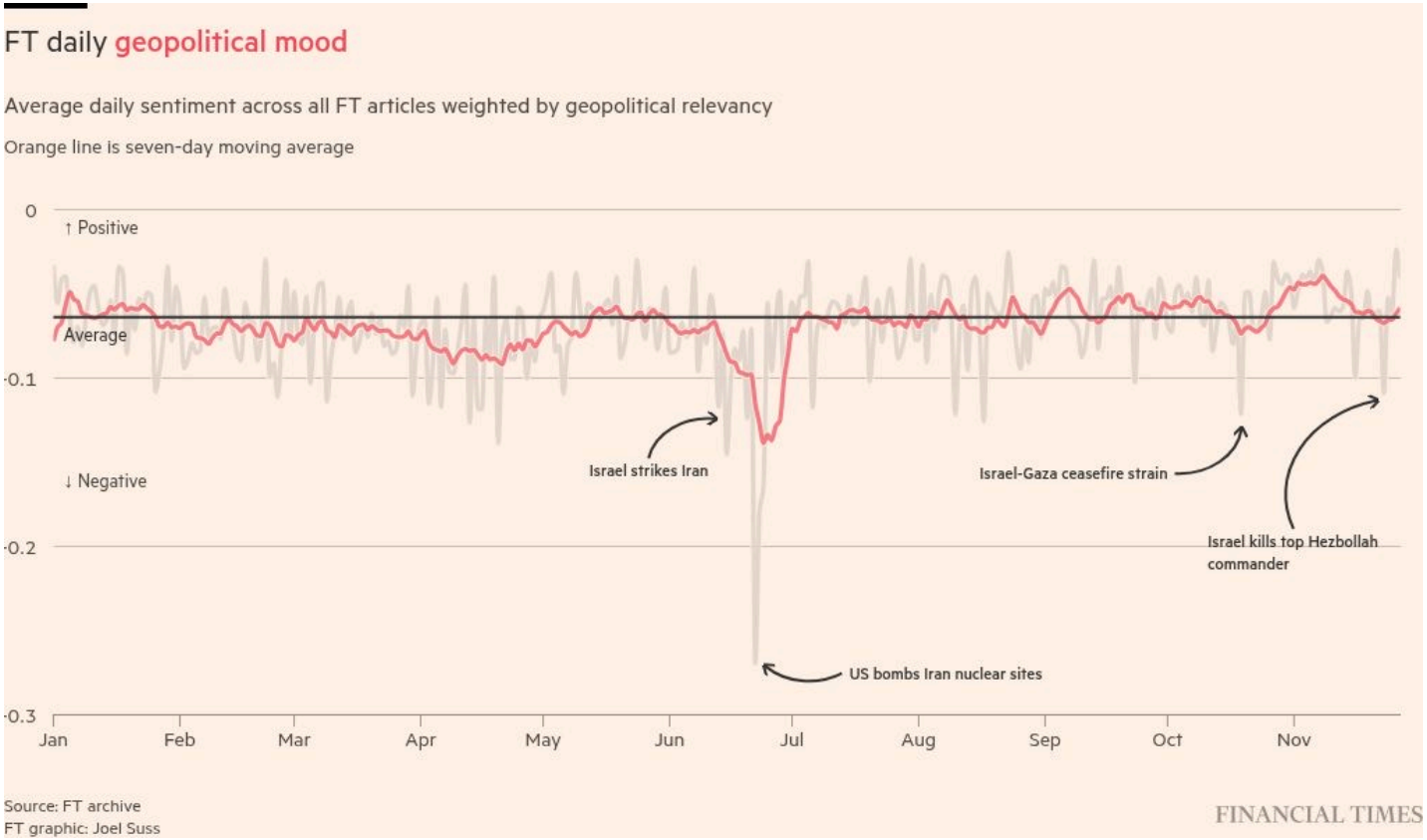
Figure 15

Note that Lower Dimensional Representations Need not be “Accurate”

(a) World model **(b)** World model with noise **(c)** Transformer

Figure 3: Reconstructed maps of Manhattan from sequences produced by three models: the true world model (left), the true world model corrupted with noise (middle), and a transformer trained on random walks (right). Edges exit nodes in their specified cardinal direction. In the zoomed-in images, edges belonging to the true graph are black and false edges added by the reconstruction algorithm are red. We host interactive reconstructed maps from transformers at the following links: shortest paths, noisy shortest paths, and random walks.

Example: Geopolitical Risk

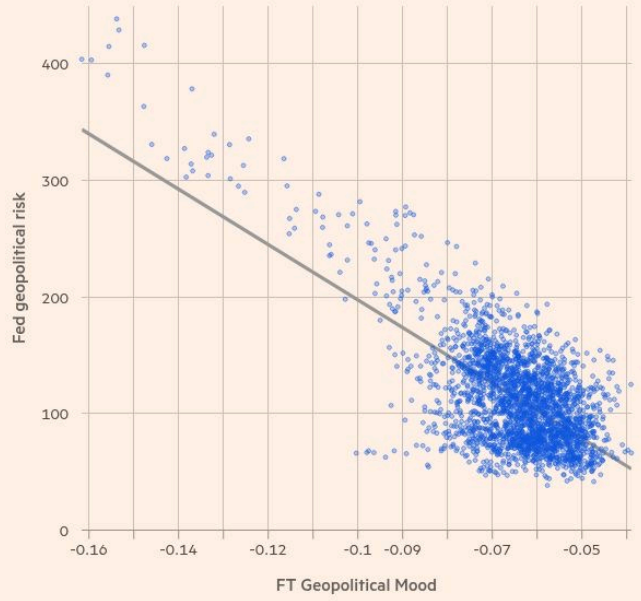


Compare to Baseline Bag of Words

Similarity Across Measures

The FT 'Geopolitical Mood' index is strongly correlated with Fed measure of geopolitical risk

Both measures are rolling seven-day average, data from 2019-2025



FINANCIAL TIMES

Source: FT archive, Caldara and Iacoviello (2022), "Measuring Geopolitical Risk" •
R-squared = 0.452, r = -0.67
FT graphic: Joel Suss

But Can Capture Broader Content

The FT's geopolitical mood index captures far more relevant content than the Fed's dictionary approach

All articles weighted above 0.5 for geopolitical content. Dots are embeddings projected on to two-dimensional space (articles that are closer together are closer in topic)

Darker blue dots are articles not flagged by GPR dictionary; red dots are those flagged

FINANCIAL TIMES

Source: FT archive • All articles are from 2025. Dimensionality of embedding vectors reduced using UMAP algorithm
FT graphic: Joel Suss

Review

- LLMs work by predicting next word (token) across massive amounts of texts. This works surprisingly well in producing general capabilities.
- In the background, the model converts words, sentences, and documents into numerical representations (embeddings) which capture meaning and relationships
- They can work to reason, write code, translate, summarize long documents
- But they also hallucinate (confidently say things which are not true), they are brittle (small changes in prompting produce very different results), and they suffer from a lack of fundamental “understanding”