

$$f = G \frac{m_1 m_2}{r^2}$$

# Data Engineering – class 4

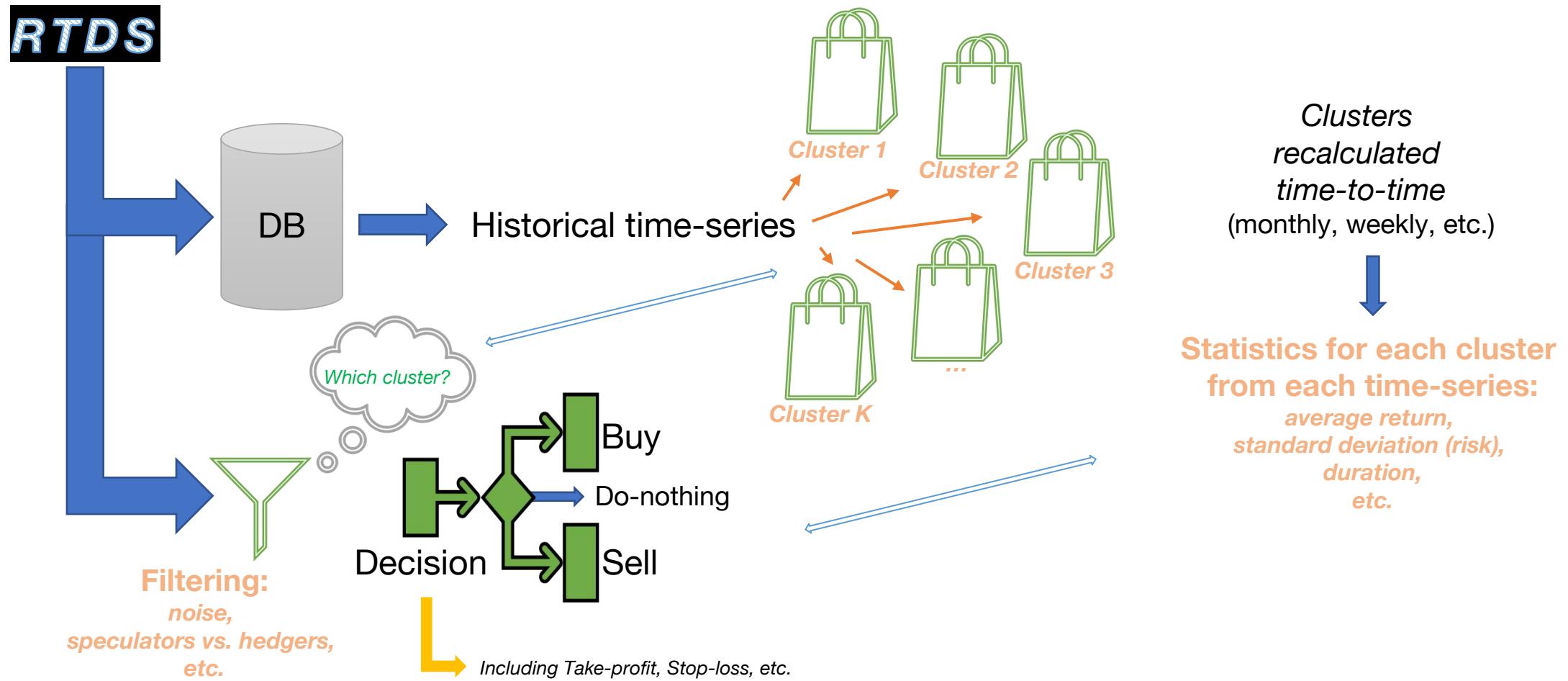
## Endogenous and exogenous

data

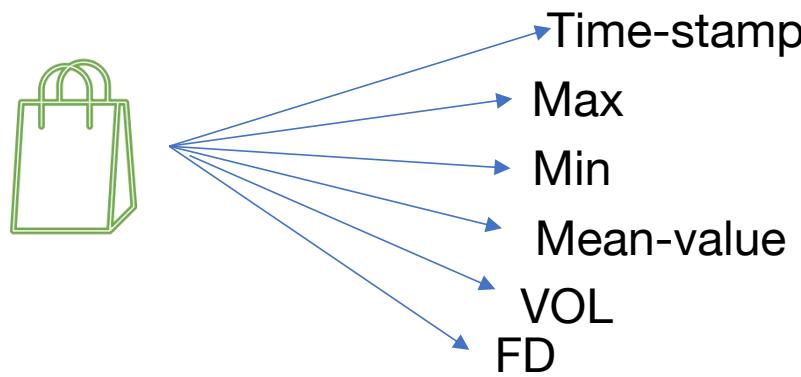
Carlos J. De Oliveira  
[\(cnd279@nyu.edu\)](mailto:cnd279@nyu.edu)

Fall 2023

# Real-Time Data Systems (RTDS)



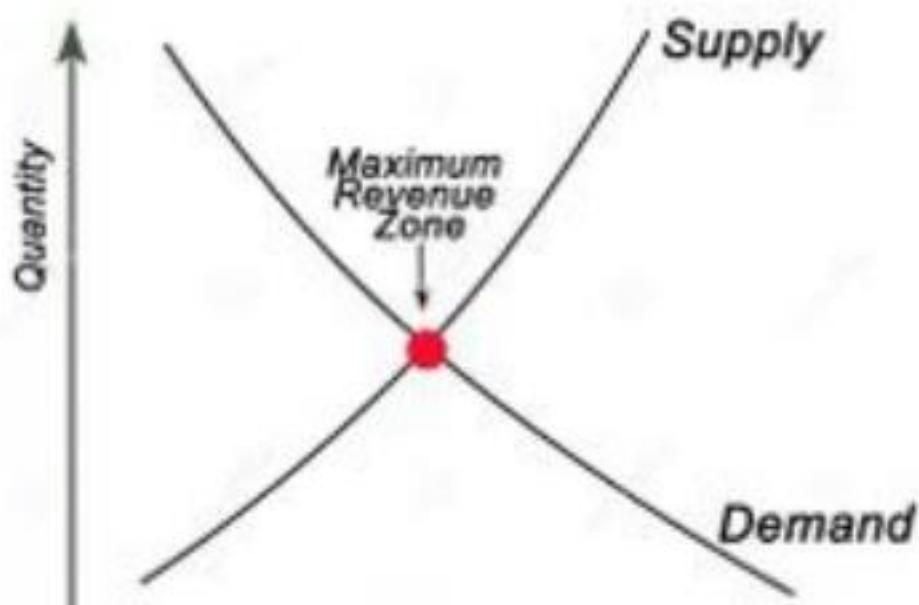
# Clusters' behavior



- Endogenous → utility function of economic agents
- Exogenous → for example, macro-economic data
  - Will approach the inflation problem
  - Covid (started 2 years ago)

# Inflation

*Understanding Laws of Supply & Demand*



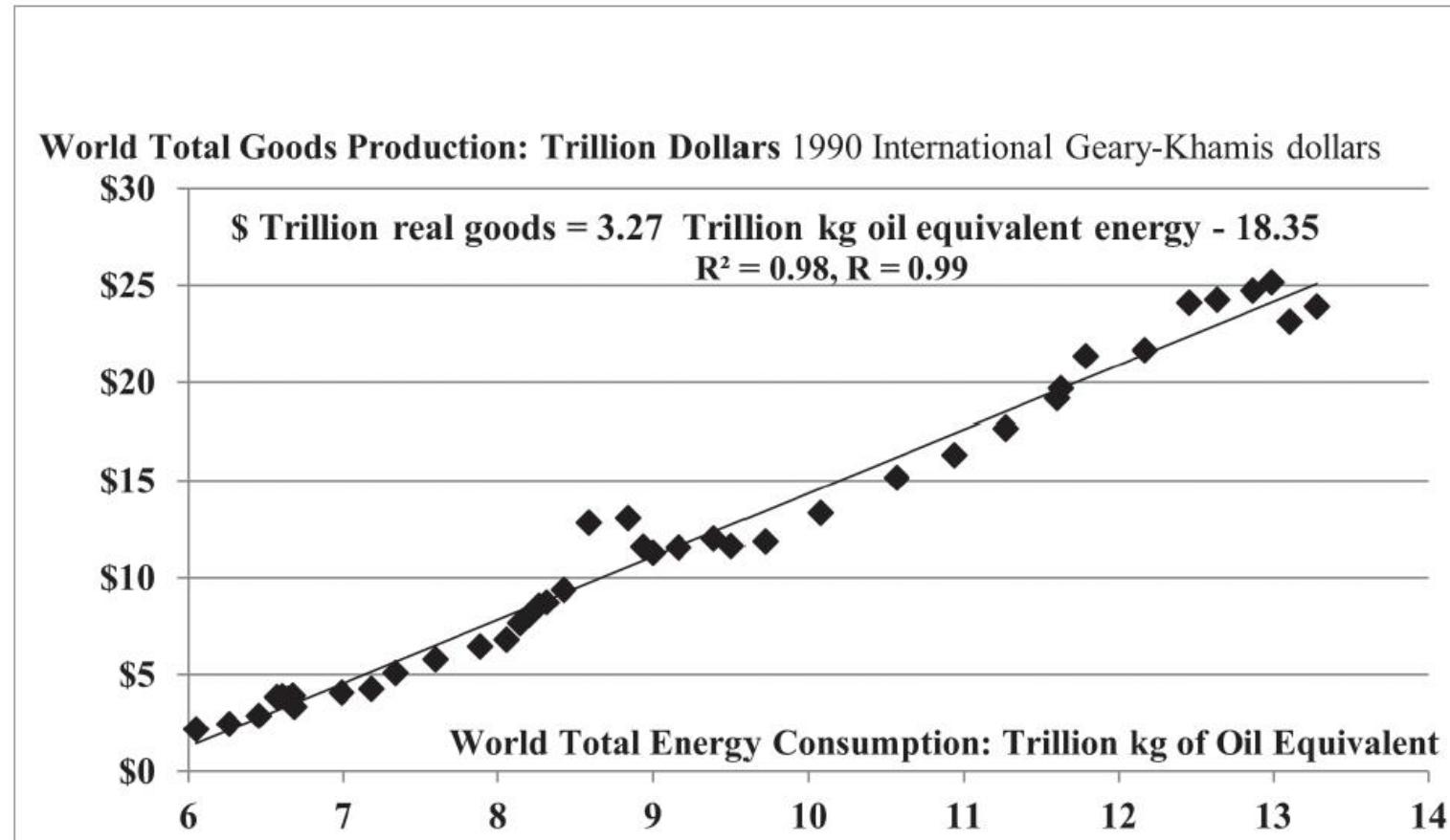
---

## Supply Inflation

---

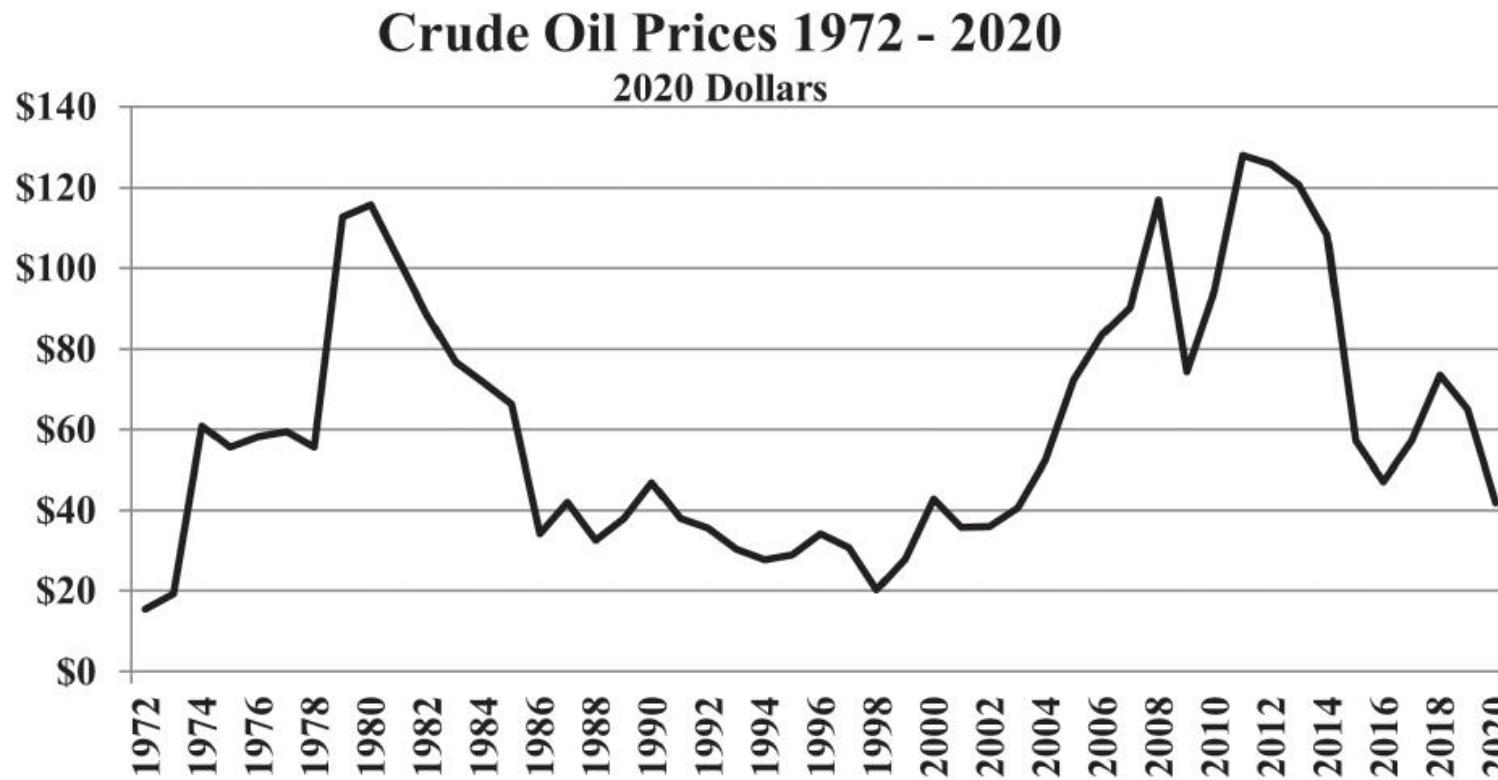
## Demand Inflation

# Industrial production vs. Oil prices (1976 – 2016)



Real Goods production and Energy Consumption:  
British Petroleum, "BP Statistical Review of World Energy, June 2017," 2018. <http://www.bp.com/statisticalreview>  
"Manufacturing Value Added," *The Global Economy*, December 3, 2022. [https://www.theglobaleconomy.com/rankings/manufacturing\\_value\\_added/](https://www.theglobaleconomy.com/rankings/manufacturing_value_added/)  
Angus Maddison, *Historical Statistics for the World Economy.xls*, 2007. [http://www.ggdc.net/maddison/historical\\_statistics/horizontal-file\\_03-2007.xls](http://www.ggdc.net/maddison/historical_statistics/horizontal-file_03-2007.xls)  
"Field Listing – GDP – composition by sector," *The World Factbook*, CIA, 2012. <https://web.archive.org/web/20070613004816/https://www.cia.gov/library/publications/the-world-factbook/fields/2012.html>  
"Gross Domestic Product for United States," Archival Economic Data, FRED, St. Louis Fed, December 3, 2022. <https://fred.stlouisfed.org/series/MKTGDPUSA646NWDB>  
"Real GDP at Constant National Prices for Indonesia," Archival Economic Data, FRED, St. Louis Fed, December 3, 2022. <https://alfred.stlouisfed.org/series?seid=RGDPNAIDA666NRUG>  
"Gross Domestic Product for Germany," Archival Economic Data, FRED, St. Louis Fed, December 3, 2022. <https://alfred.stlouisfed.org/series?seid=MKTGDPDEA646NWDB>  
Martin Neil Baily and Barry R. Bosworth, "U.S. Manufacturing: Understanding Its Past and Its Potential Future," *Journal of Economic Perspectives* 28, no. 1 (Winter 2014): 3–26.

# Oil prices



What is the correlation between Oil Prices and Oil Production?

Why *white noise* prevails?

# DE: *Real* reality vs. *Virtual* reality

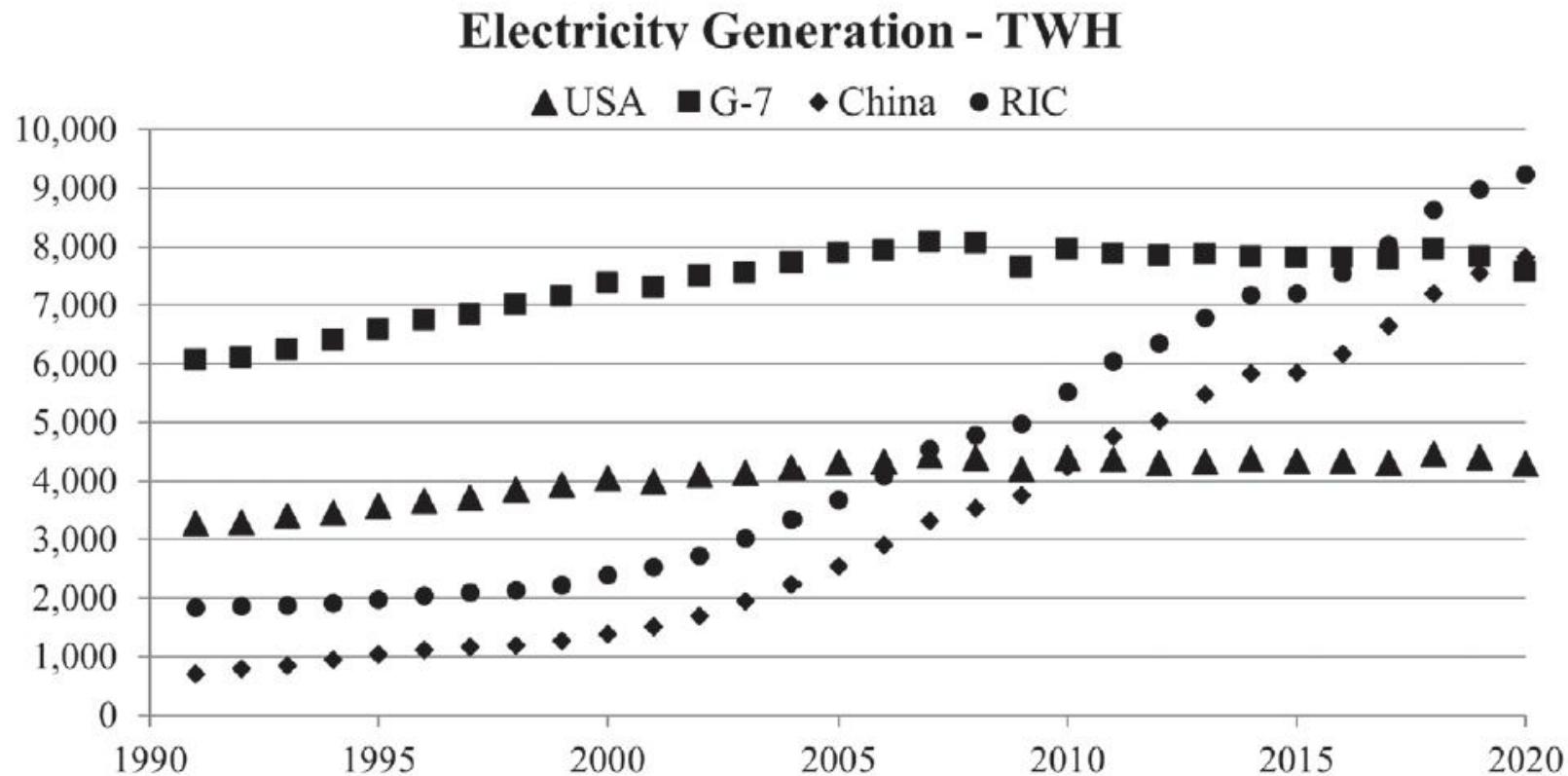
Hedge Funds Vanguard, BlackRock, State Street & T. Rowe Price Sample of Corporate Media Ownership		
Media Corporation	%	Media Outlets
Newscorp	66%	Wall Street Journal, New York Post, Sky News, Times UK, The Sun, Fox News
Comcast	26%	NBC News, Sky News
Warner Bros.	25%	CNN
Alphabet (Google)	27%	YouTube
Twitter	27%	Twitter
Meta	26%	Facebook
Netflix	26%	Netflix

stragetical level

# X-Changes

- Vanguard, Blackrock, State Street, and T. Rowe Price own 28.9% of ICE.
- Vanguard, Blackrock, and State Street own 25.5% of CME.
  - White noise (again)?

# Electricity generation



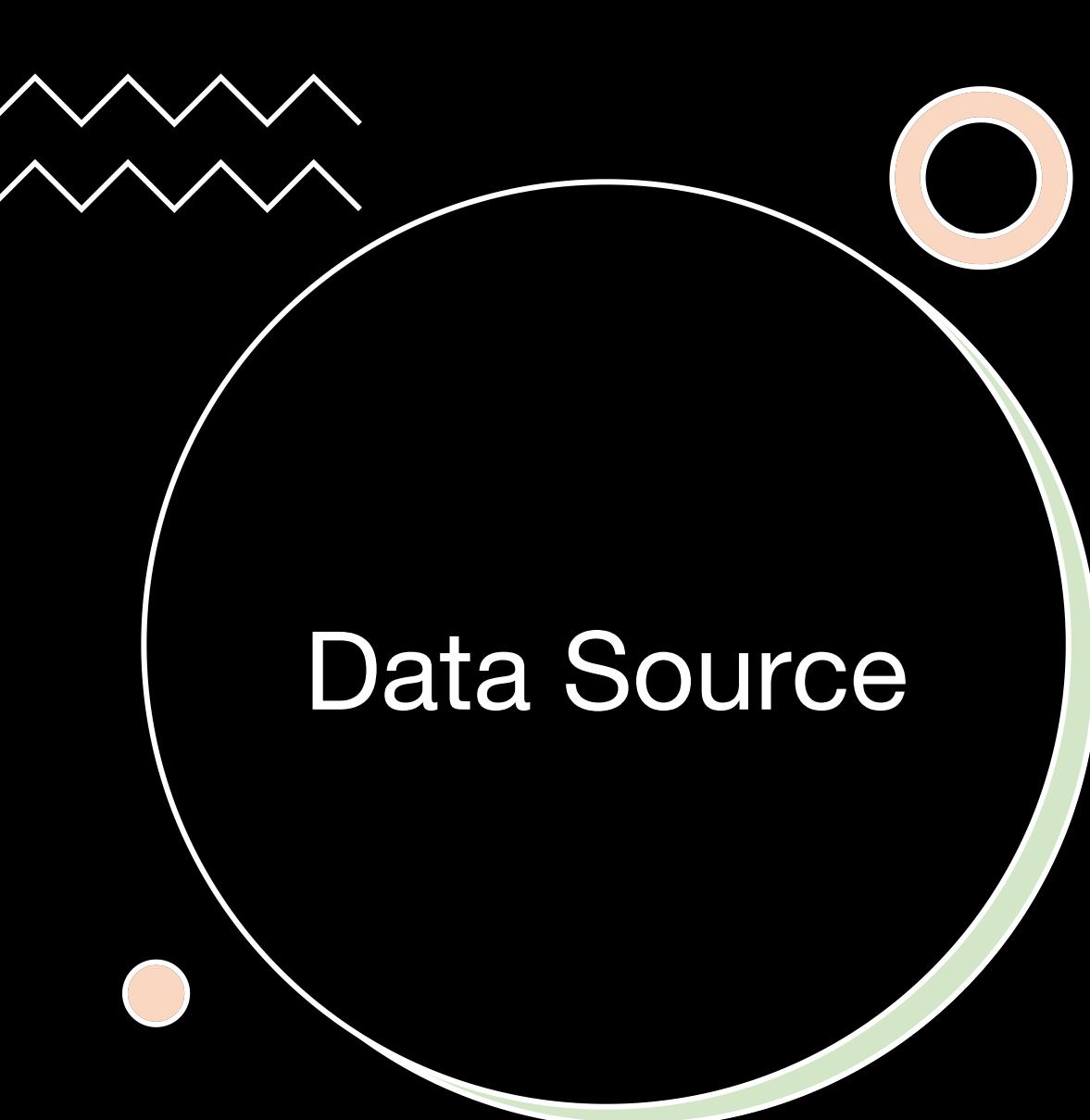




Elastic vs. Inelastic

**Today:**  
**Supply inflation dominance**  
**Mainly on inelastic products (energy  
& food)**

**Do interest rate hikes solve the inflation problem?**



# Data Source

- <https://www.bls.gov/charts/consumer-price-index/consumer-price-index-by-category-line-chart.htm>
- <https://fred.stlouisfed.org/#>

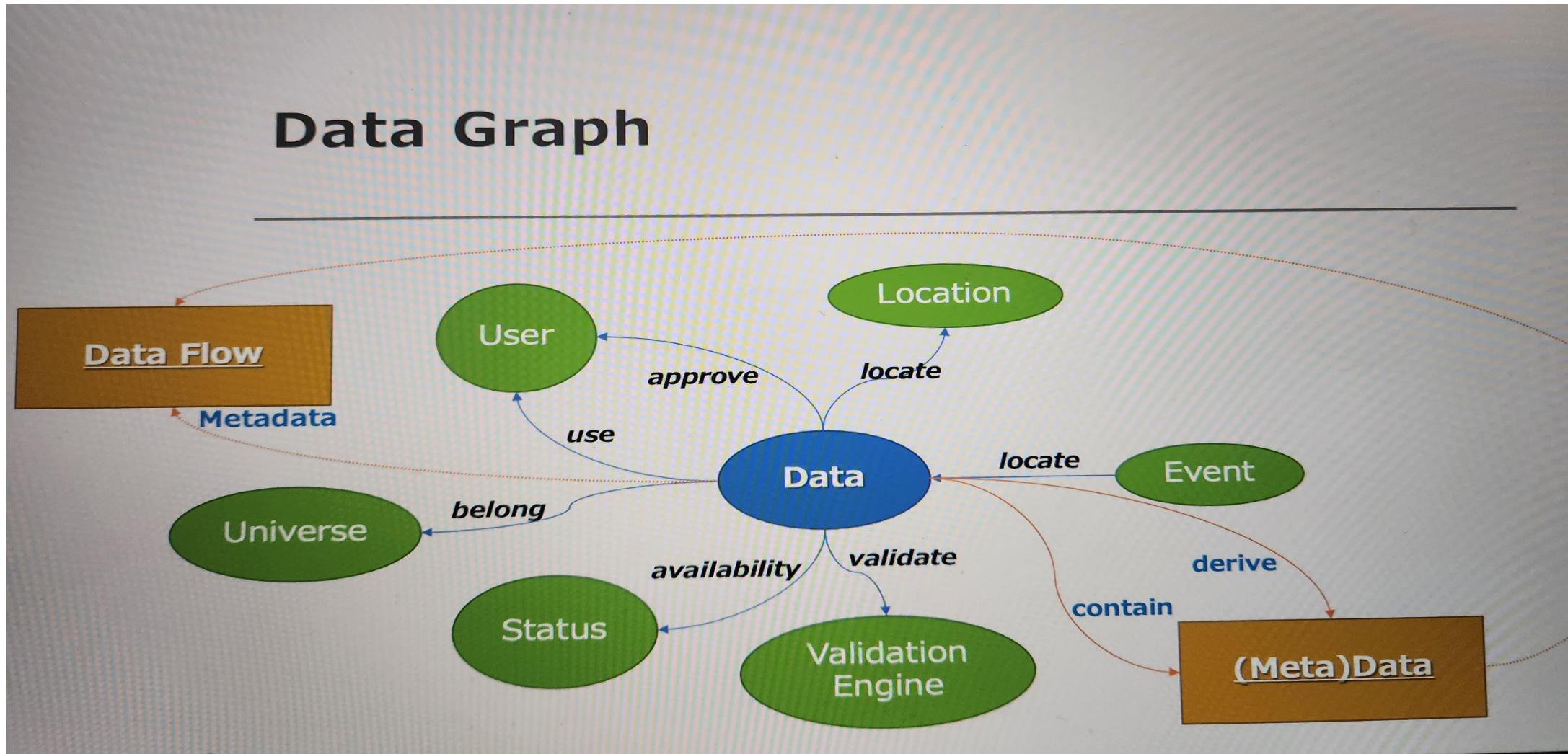


# Example (inflation)

U.S. BUREAU OF LABOR STATISTICS 12-month percentage change, Consumer Price Index, selected categories, not seasonally adjusted															<input type="text"/> Search BLS.gov		
Month	All items	Food	Food away from home	Energy	Gasoline (all types)	Electricity	Natural gas (piped)	All items less food and energy	Commodities less food and energy commodities	Apparel	New vehicles	Medical care commodities	Services less energy services	Shelter	Medical care services	Education and communication	
	1.2%	1.9%	1.6%	2.6%	-12.3%	-17.3%	0.0%	-22.5%	2.5%	-0.9%	-2.1%	-1.8%	3.6%	3.9%	4.0%	5.0%	2.5%
May 2002	1.2%	1.9%	1.6%	2.6%	-12.3%	-17.3%	0.0%	-22.5%	2.5%	-0.9%	-2.1%	-1.8%	3.6%	3.9%	4.0%	5.0%	2.5%
June 2002	1.1%	1.6%	1.0%	2.6%	-11.1%	-15.4%	-3.1%	-17.8%	2.3%	-1.0%	-2.9%	-1.8%	3.3%	3.7%	3.7%	4.8%	2.4%
July 2002	1.5%	1.4%	0.7%	2.5%	-5.2%	-3.7%	-3.9%	-13.9%	2.2%	-1.3%	-3.2%	-1.8%	3.6%	3.8%	3.7%	5.3%	2.7%
Aug 2002	1.8%	1.2%	0.4%	2.3%	-2.8%	-0.2%	-3.2%	-10.5%	2.4%	-0.7%	-1.7%	-1.6%	3.5%	3.7%	3.6%	5.1%	2.9%
Sept 2002	1.5%	1.3%	0.5%	2.3%	-4.8%	-7.3%	-2.7%	-4.8%	2.2%	-1.1%	-1.7%	-1.1%	3.3%	3.6%	3.6%	5.1%	2.7%
Oct 2002	0.0%	0.0%	0.1%	0.0%	0.0%	7.0%	0.0%	4.4%	0.0%	1.0%	0.1%	1.1%	0.0%	0.7%	0.6%	5.1%	0.1%



# Data graphs and metadata



# **Modeling (non-exhaustive) Pros vs. Cons**

---

## **Pros:**

- More (and better) data – think about recessions
- Process memory – think about Markov Chains
- Life-cycle – think about Google projects

## **Cons:**

- Expensive training process
- Annotation dependent
- New architecture

# Metadata example

---



Data: 32 entries

Max error →  $\pm 1$  min  
Avg error → zero!



# Data dynamics example

---

(Non-)Transitory Inflation Equilibrium

Inelastic Inflation dominance

Wrong T10 pricing?

**New!!** → *Productivity vs. Inflation targeting*

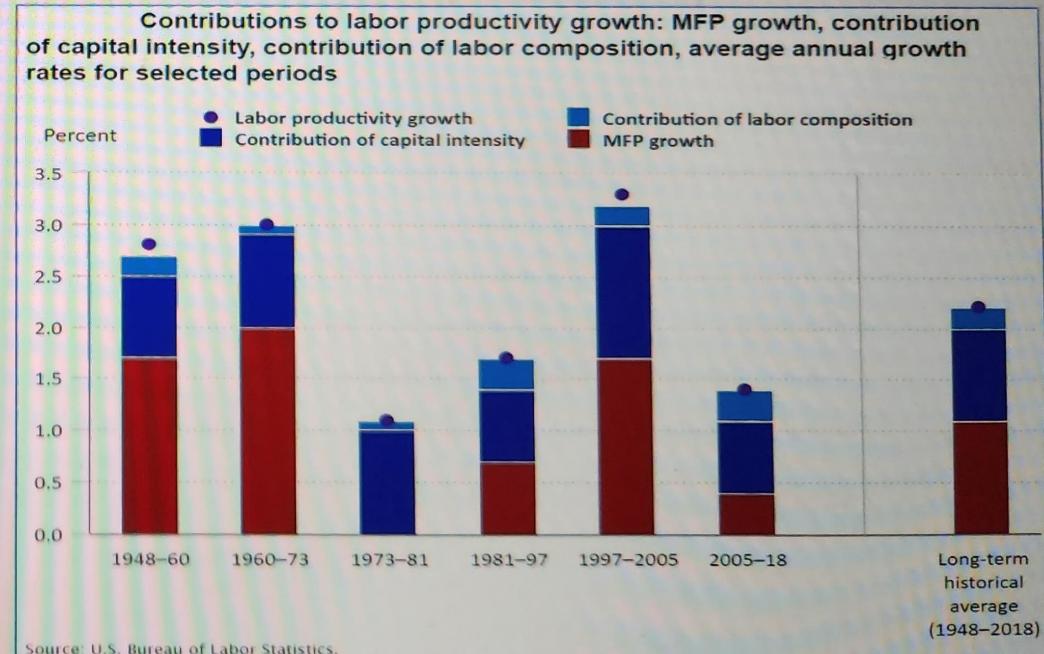
# Analysis of the US productivity slowdown

Productivity growth = multifactor productivity (MFP) growth + contribution of capital intensity + contribution of labor composition  
(by definition)

MFP growth → tech advances in production, improvements of managerial efficiency, etc., not accounted for by the growth of capital and labor

Contribution of capital intensity → reflects businesses' decision-making process between hiring more workers and purchasing more or higher-quality equipment, or of substituting equipment for workers or vice versa

Contribution of labor composition → when firms hire more workers with higher skills and more experience or lay off workers with lower skills or less experience, or when labor costs rise relative to capital costs, the contribution of labor composition to labor productivity growth increases



*Reduced responsiveness to productivity gains at the firm level !!*

# A note on MFP

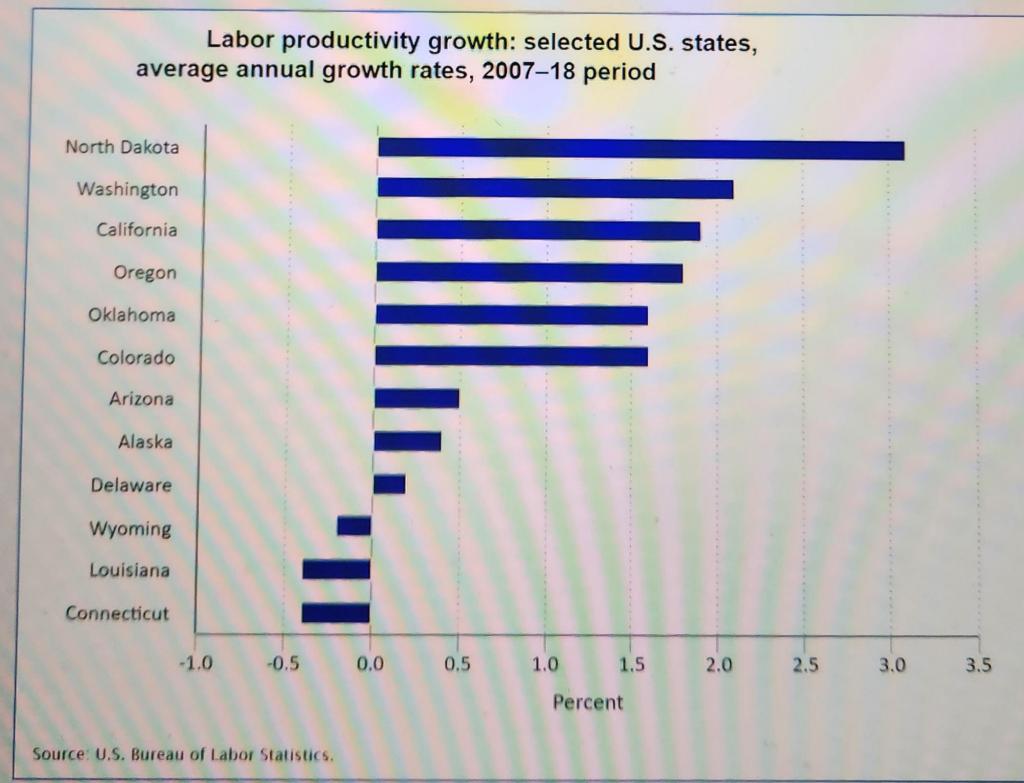
---

From a data-wide perspective, we can identify several plausible explanations for the slowdown in MFP growth, including

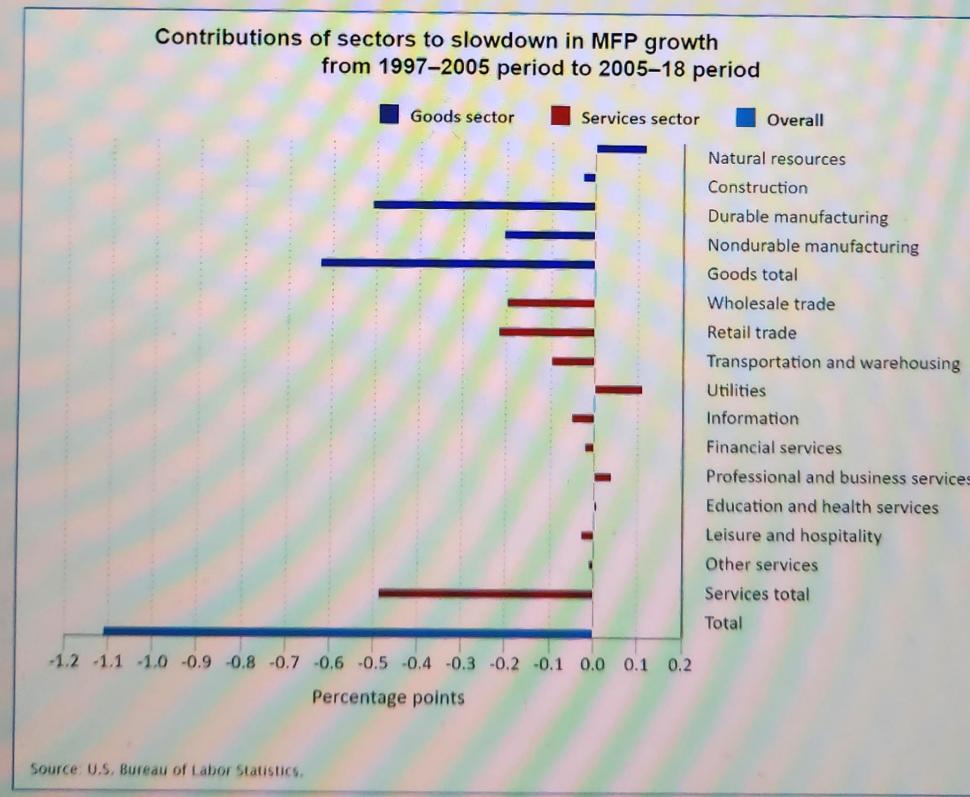
- Declining rates of productivity-enhancing job reallocation,
- Rising market power and industry concentration,
- Greater restraints on competition,
- Growing income inequality,
- The drag from the global financial crisis and its weak recovery,
- Diminishing returns to innovation relative to that of the late 19<sup>th</sup> and early to mid-20<sup>th</sup> centuries

# Did the productivity slowdown progress differently?

---



# Industry-level analysis of the U.S. labor productivity



# **Bottom line**

---

Data about data might be relevant

Data about changing data might be very relevant

Sometimes (always?) data about data are more relevant than data

And flows about metadata controls the all chain



# How to deal with this gigantic amount of data?

## **future of coding with AI?**

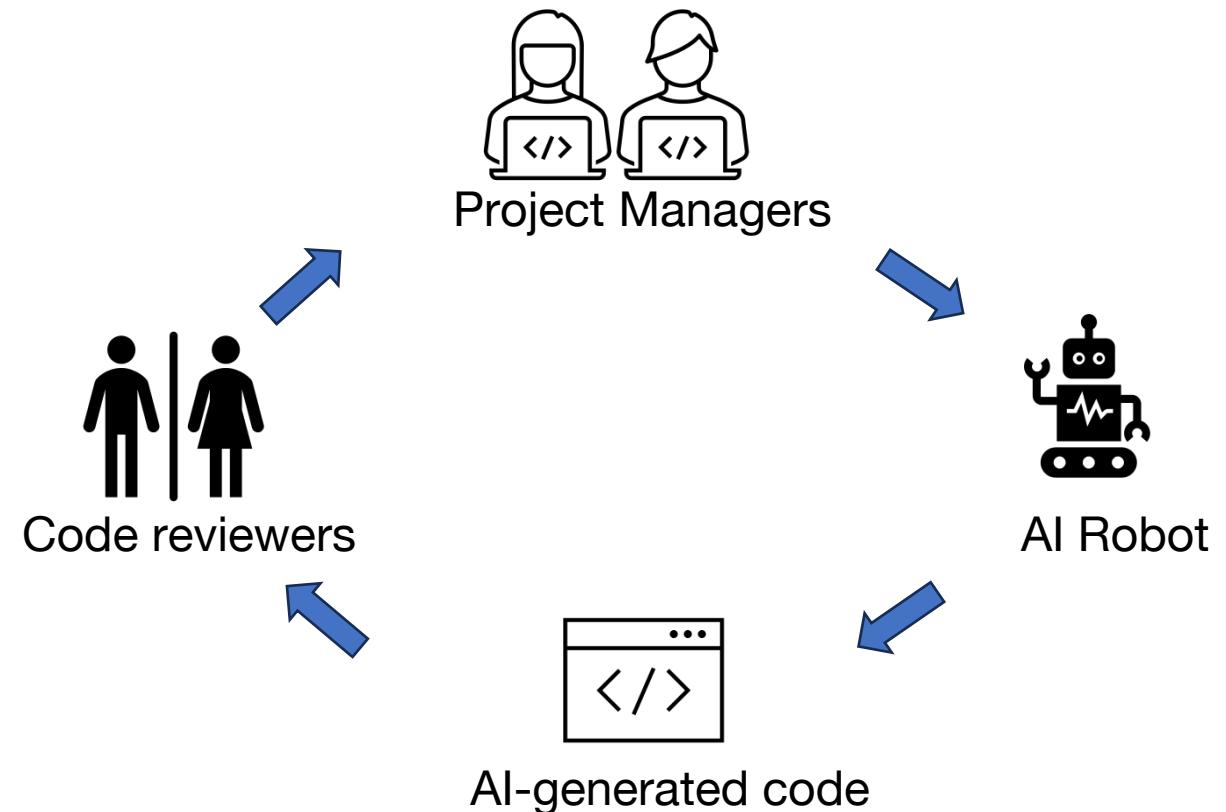
LLMs author big chunks of complex code  
humans rarely write any code themselves from scratch  
programmers use AI as coding partner & technical advisor  
... and novices use AI to create attacks & malware

## **no, this is not the future!**

this is the present (at least for some programmers)

## **a more radical future?**

LLMs replace (most or all) programmers



# Cybersecurity

The screenshot shows a news article from WIRED. At the top, there's a navigation bar with the WIRED logo and categories: BACKCHANNEL, BUSINESS, CULTURE, GEAR, IDEAS, SCIENCE, SECURITY. Below that is a sub-navigation bar with LILY HAY NEWMAN, SECURITY, and DEC 10, 2021 2:54 PM. The main title is 'The Internet Is on Fire'. A subtitle below it reads: 'A vulnerability in the Log4j logging framework has security teams scrambling to put in a fix.' A large quote in the middle of the page says: 'affected 93% of enterprise clouds, 40% of businesses companies didn't know they were even using Log4J Log4J incorporated eval, the most dangerous function'

The screenshot shows a preprint on arXiv. The header includes the arXiv logo, a search bar, and links for Help | Advance. The category is Computer Science > Machine Learning. It was submitted on 14 Apr 2022. The title is 'Planting Undetectable Backdoors in Machine Learning Models' by Shafi Goldwasser, Michael P. Kim, Vinod Vaikuntanathan, Or Zamir. The abstract discusses how a malicious learner can plant an undetectable backdoor into a classifier without being detected by a computationally-bounded observer. It details two frameworks for planting undetectable backdoors with incomparable guarantees. The text also covers how to plant a backdoor in any model using digital signature schemes and how to insert undetectable backdoors in models trained using Random Fourier Features (RFF) or Random ReLU networks. The construction guarantees that given black-box access to the original model and the backdoored version, it is computationally infeasible to find even a single input where they differ. The paper also addresses the related issue of robustness to adversarial examples.

# Why LLMs are imperfect DE (and DS) coders?

DN

can you write a javascript function that takes an integer and turns it into a string, so if you gave it 1000 it would return "one thousand"?

```
> numberToWords(101)  
"one hundred one"  
> numberToWords(100)  
"one hundred"  
> numberToWords(99)  
"ninety nine"  
> numberToWords(43)  
"forty three"  
> numberToWords(9)  
"nine"  
> numberToWords(1001)  
"one thousand hundred one"  
> numberToWords(113)  
"one hundred eleven"
```

## solution lacks nice structure

160 = one hundred and sixty

1600 = sixteen hundred

## when you can't guess, you have to reason

LLM executes in constant time, so can't solve halting problem  
so will never be able to do arbitrary program reasoning



```
const units = ['', 'one', 'two', 'three', 'four', 'five', 'six',  
const teens = ['', 'eleven', 'twelve', 'thirteen', 'fourteen', 'fif  
const tens = ['', '', 'twenty', 'thirty', 'forty', 'fifty', 'sixty',  
const thousands = ['', 'thousand'];  
  
function numberToWords(num) {  
    if (num === 0) return 'zero';  
    if (num === 10) return 'ten';  
  
    let words = '';  
    let digit, i = 0;  
  
    while (num > 0) {  
        if (i === 0) {  
            digit = num % 10;  
            words = units[digit];  
        } else if (i === 1) {  
            digit = num % 10;  
            if (digit === 1) {  
                words = teens[parseInt(num % 100) - 10];  
            } else {  
                words = tens[digit] + ' ' + words;  
            }  
        } else if (i === 2) {  
            digit = num % 10;  
            words = units[digit] + ' hundred ' + words;  
        } else if (i === 3) {  
            digit = num % 10;  
            words = units[digit] + ' ' + thousands[i] + ' ' + words;  
        }  
        num = Math.floor(num / 10);  
        i++;  
    }  
  
    return words.trim();  
}
```

- 1)buy 20% not functing
- 2)if Price>Avg(Price of window 1) then buy 30%(40%)
- 3)if Price > Avg(price of window1&2) then buy(30%+remaining W2) (40%)
- 4)same as W3
- 5)sell everything
- simulator will do the above
- Long EURUSD
- (ask price[buy price]/ bid price[sell price]) ask>bid
- 100 units long buy 100 dollars
- 1. replace the number(the number of translaction is always = 100 dollars)