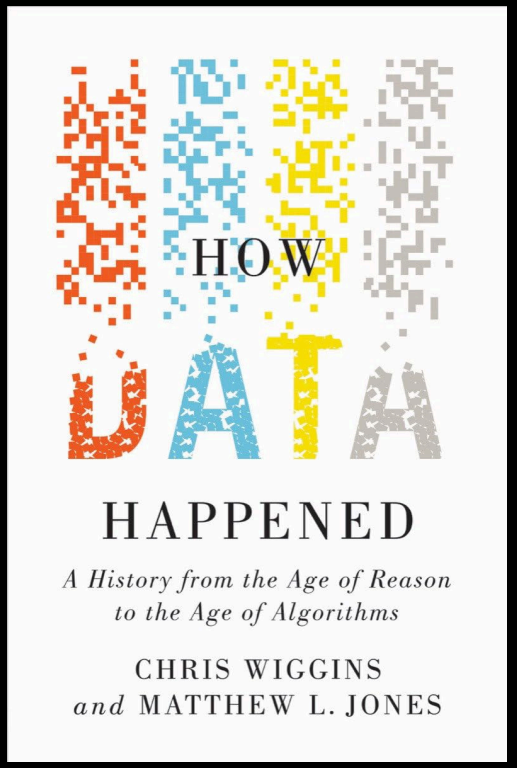
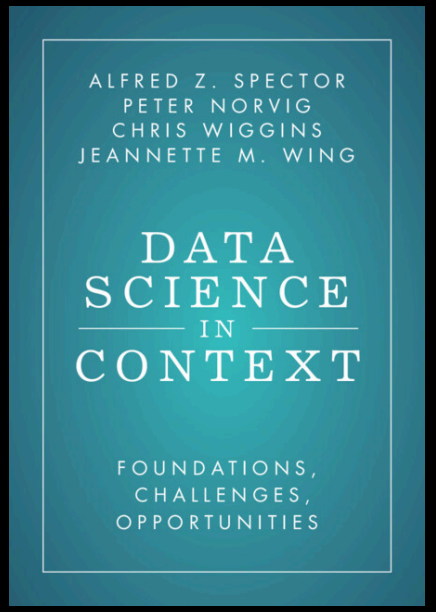
data: past, present, and future

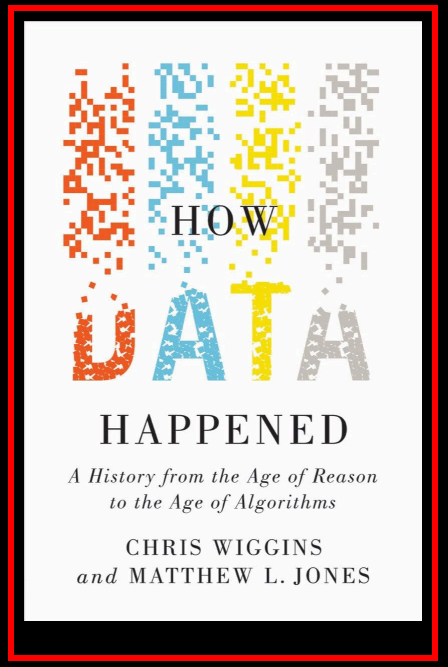
SFI ACtioN Annual Risk Meeting
2025-11-06

chris.wiggins@columbia.edu

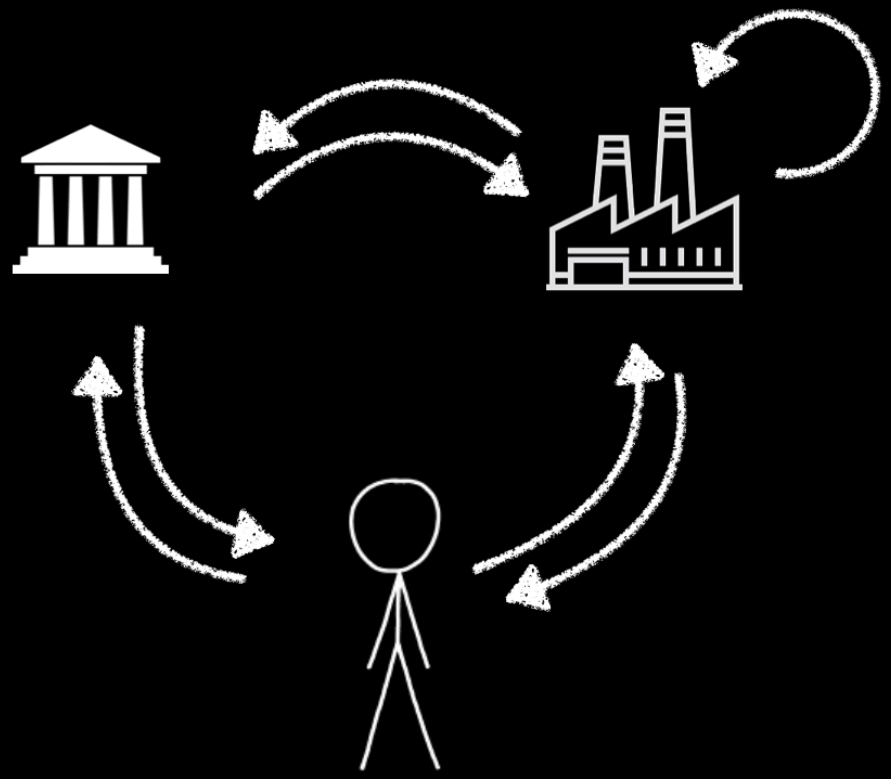


self-intro


ALFRED Z. SPECTOR
PETER NORVIG
CHRIS WIGGINS
JEANNETTE M. WING
DATA SCIENCE IN CONTEXT
FOUNDATIONS, CHALLENGES, OPPORTUNITIES

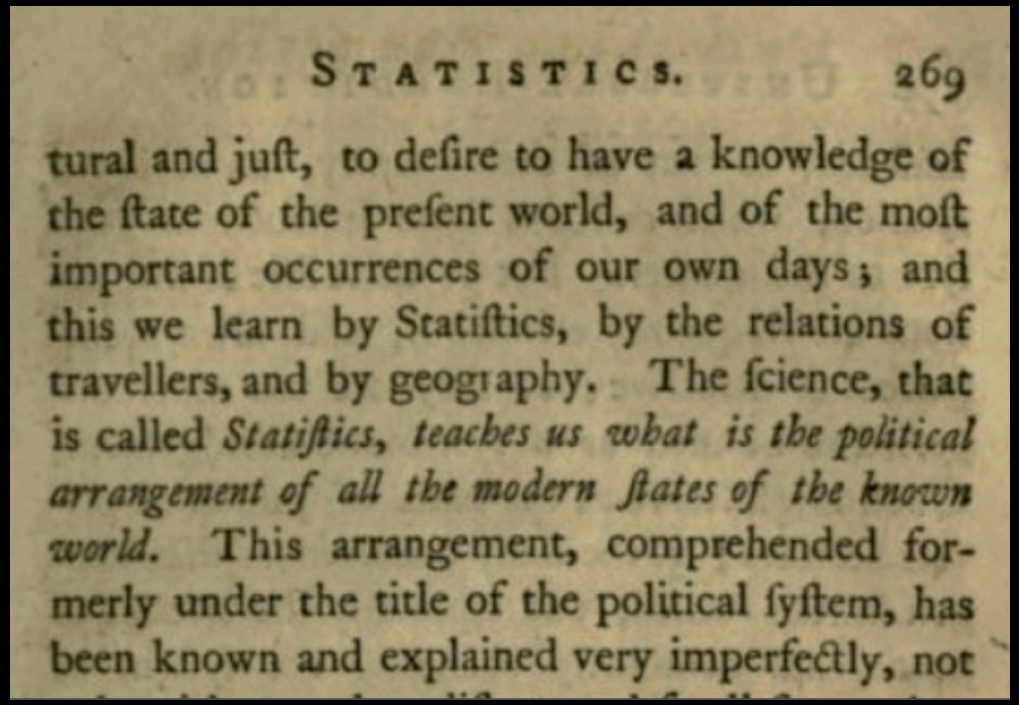

HOW DATA HAPPENED
A History from the Age of Reason to the Age of Algorithms
CHRIS WIGGINS and MATTHEW L. JONES

> "national standing depends very much on what technologies a nation commands"
> ... an unstable 3-player game



Part 1: past: data & power

state power & "vulgar statistics" (1806)



STATISTICS. 269

tural and juſt, to deſire to have a knowledge of the ſtate of the preſent world, and of the moſt important occurrences of our own days; and this we learn by Statiſtics, by the relations of travellers, and by geography. The ſcience, that is called *Statiſtics, teaches us what is the political arrangement of all the modern ſtates of the known world.* This arrangement, comprehended formerly under the title of the political ſyſtem, has been known and explained very imperfectly, not

"höhere"/"edlere" (higher)/(nobler) vs "gemeine" (common/vulgar statistics) practiced by "Zahlenknechte" (number servants) and "Tabellenfabrikanten" (table manufacturers).

corp power & beer (Guinness, 1886 IPO)

- Gosset ("Student") wields data as technology controlled by a corp, not the state
- competitive, proprietary advantage

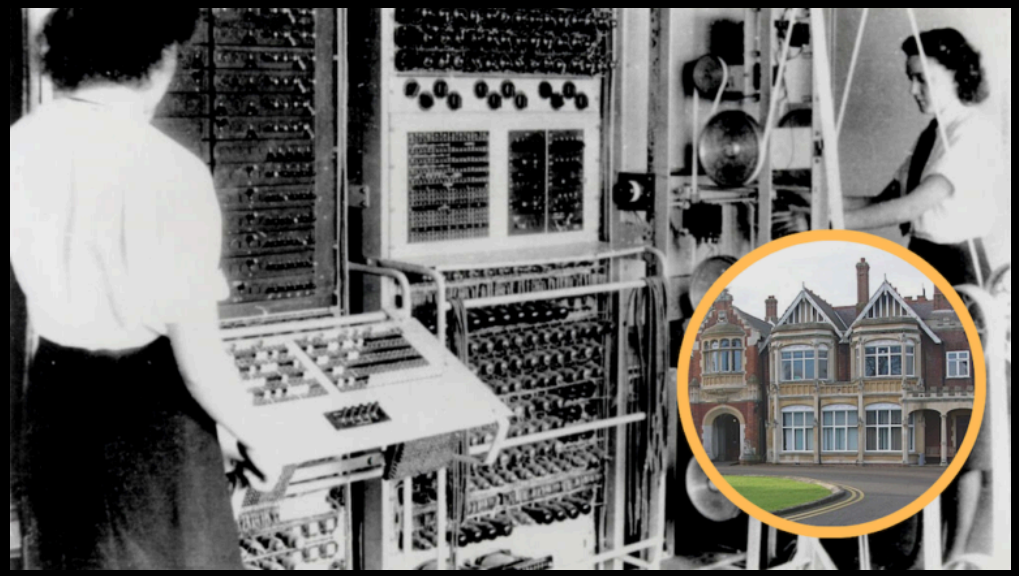
1920s: state power

> "economy was based on agriculture and manufacturing"

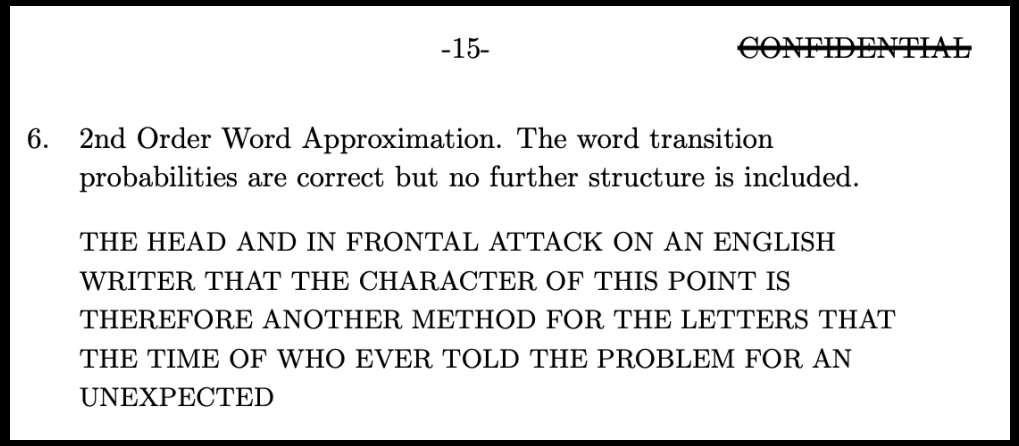
- **United Kingdom - R.A. Fisher:** @ Rothamsted Experimental Station
- **Poland - Jerzy Neyman:** @ Worked on agricultural statistics for Polish state
- NB: these two would go on to fight for 50+ years about truth & data

> national standing depends [on] technologies

dawn of data science & digital compute



- Colossus computer invented for cryptanalysis - fundamentally a data science problem, solved by the state
- **Enigma:** corp technology advancing national standing
- from **Bletchley Park** to **Bell Labs** data @ center of state to corporate transition tying technology and national standing



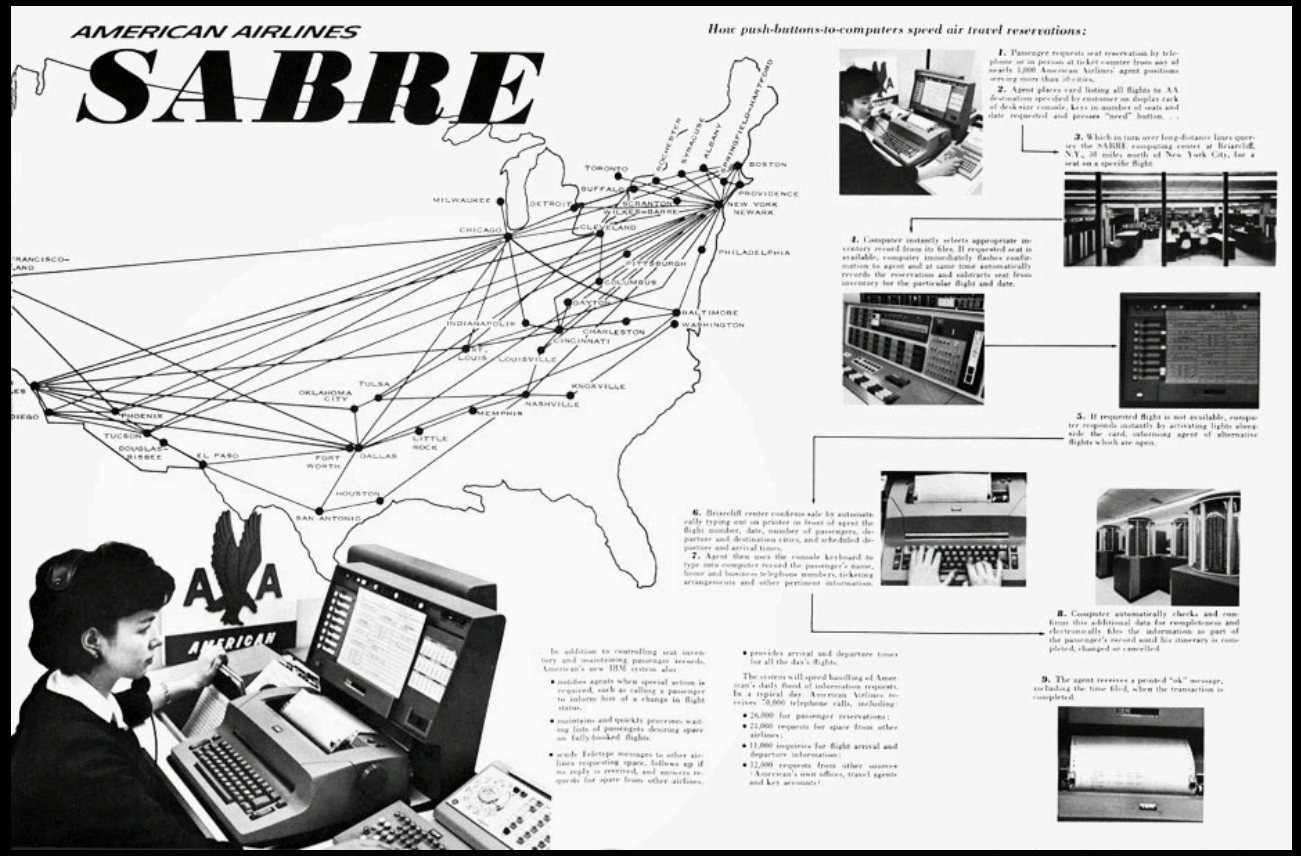
-15- ~~CONFIDENTIAL~~

6. 2nd Order Word Approximation. The word transition probabilities are correct but no further structure is included.

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

- @ ATT/Bell, Shannon's 1945 "Mathematical Theory of Cryptography" : **small language model**
- .com and .mil, national standing and technology, long before "military-industrial complex" (1961)
- cf. Mark V. Shaney (1984) natural language prank / program

SAGE (Semi-Automatic Ground Environment, 1952-1955) → *SABRE* (Semi-Automated Business Research Environment, 1960-present)



Part 2: present: scarcity, data, tokens

The Economic Principle

Mankiw's First Principle of Economics: (including *data economics*)

> "People face trade-offs"

When one thing becomes abundant, something else becomes scarce.

information glut->attention economy

> "In an information-rich world, the wealth of *information* means a dearth of something else: a *scarcity* of whatever it is that information consumes. What information consumes is rather obvious: it consumes the *attention* of its recipients. Simon (1971)

> "Compared with monetary transactions, attention transactions on the Web will be far more numerous."
> ..."inequality" between "stars and fans" Goldhaber (1997)

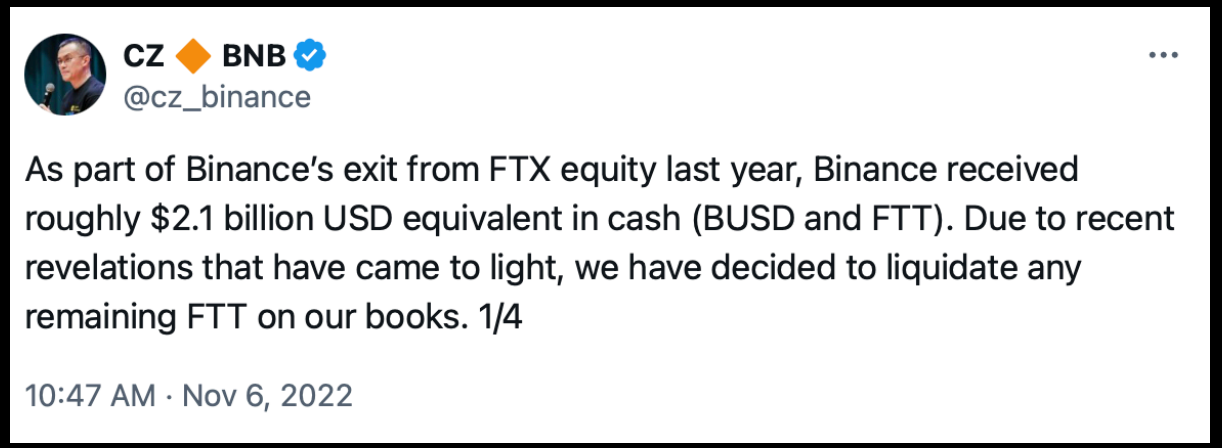
Web2.0, UGC, and attention, money, data

User-generated content platforms (Twitter, Reddit, Wikipedia; YouTube) created:

- The star/fan monetization of attention
- **also:** massive corpus of UGC esp. text, perfect for training statistical models.

November 2022: "technology disrupts and re-orders the economy"

November 2022: "technology disrupts and re-orders the economy"

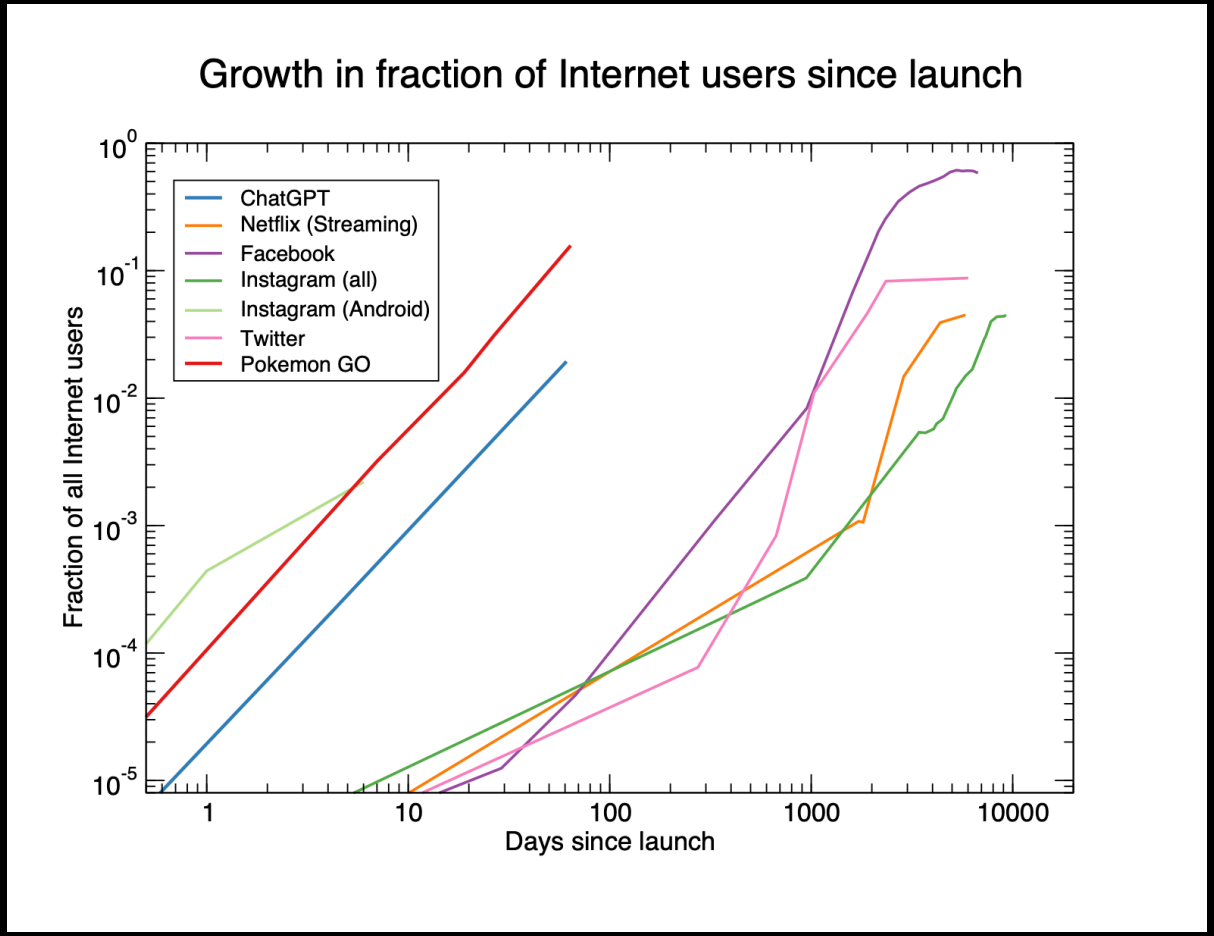


CZ 🔶 BNB ✔
@cz_binance

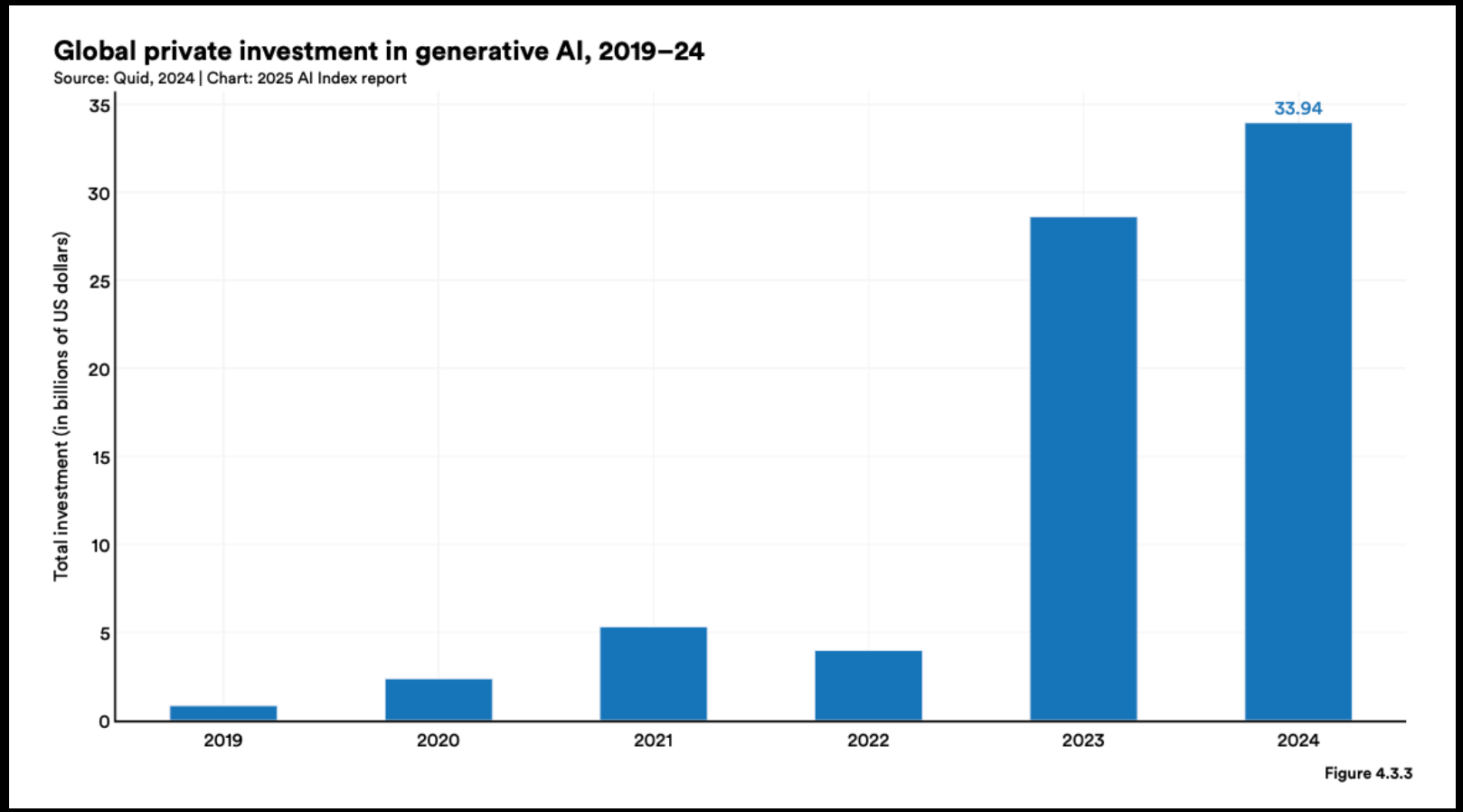
As part of Binance's exit from FTX equity last year, Binance received roughly $2.1 billion USD equivalent in cash (BUSD and FTT). Due to recent revelations that have came to light, we have decided to liquidate any remaining FTT on our books. 1/4

10:47 AM · Nov 6, 2022

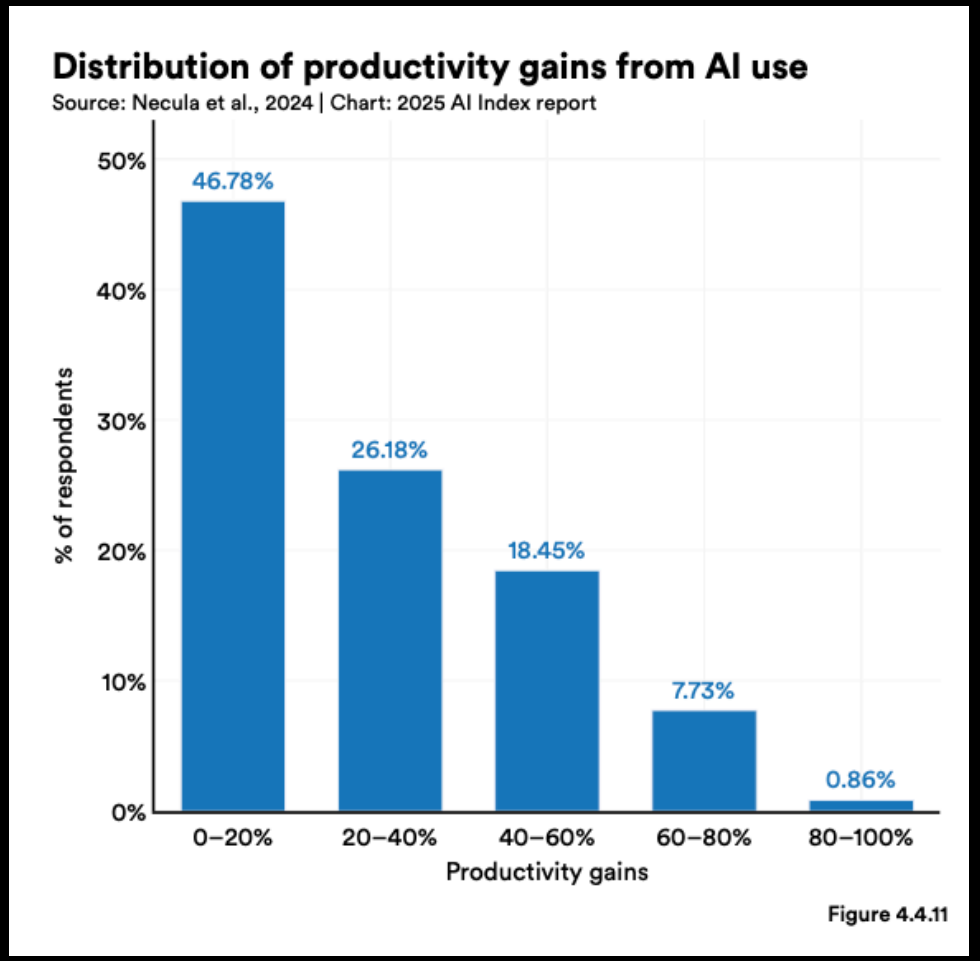
November 2022: "technology disrupts and re-orders the economy"



investment vs time

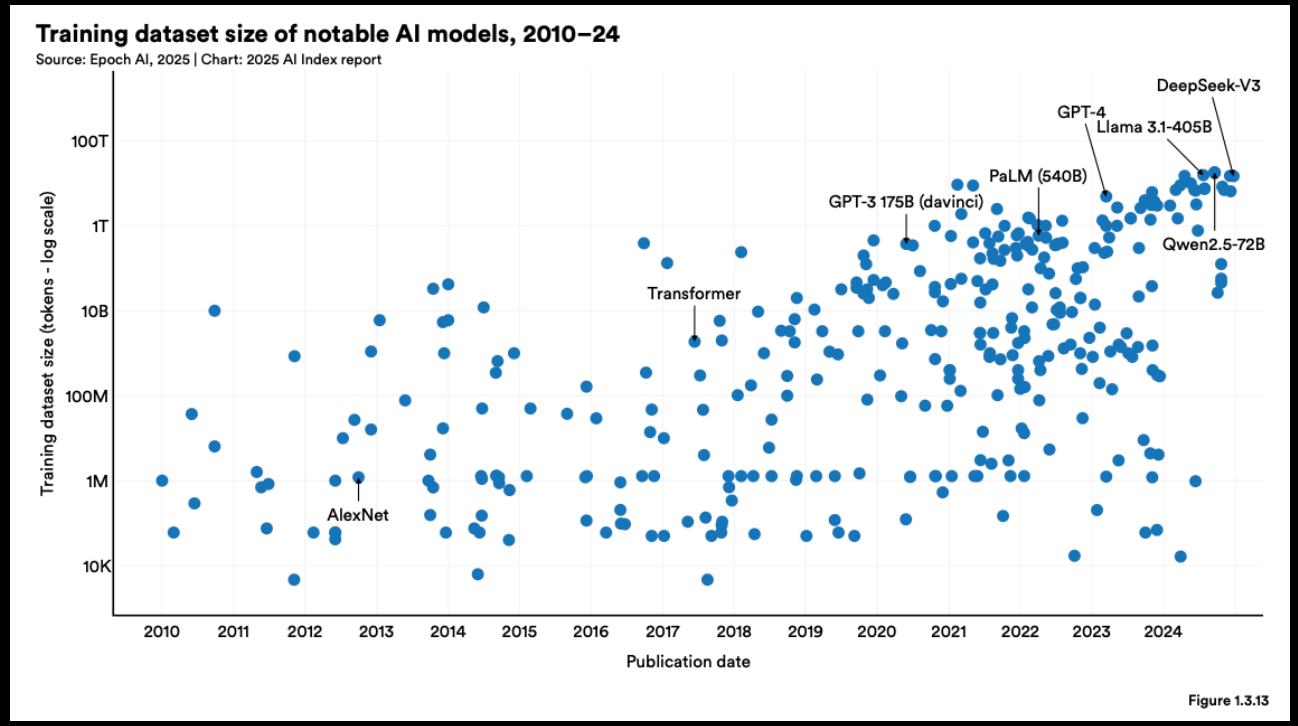


peak of inflated expectations



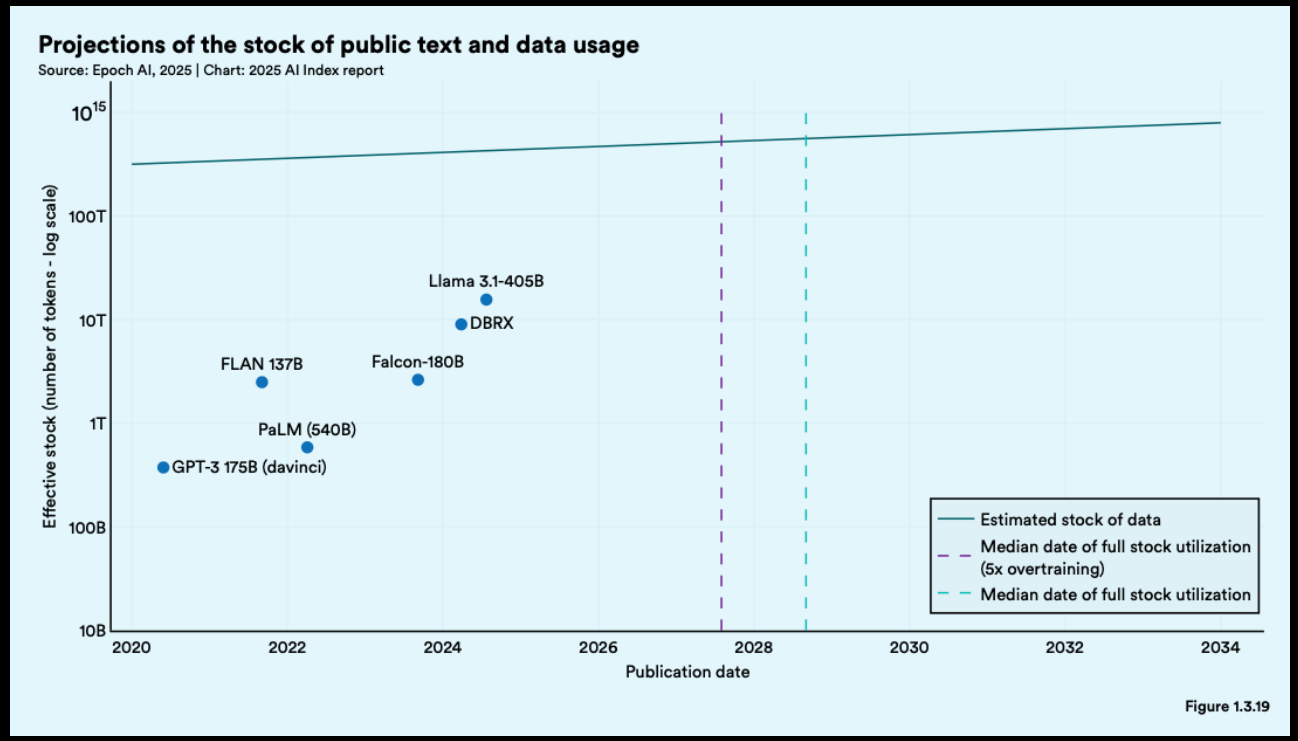
how long before crash?

log(tokens) vs time



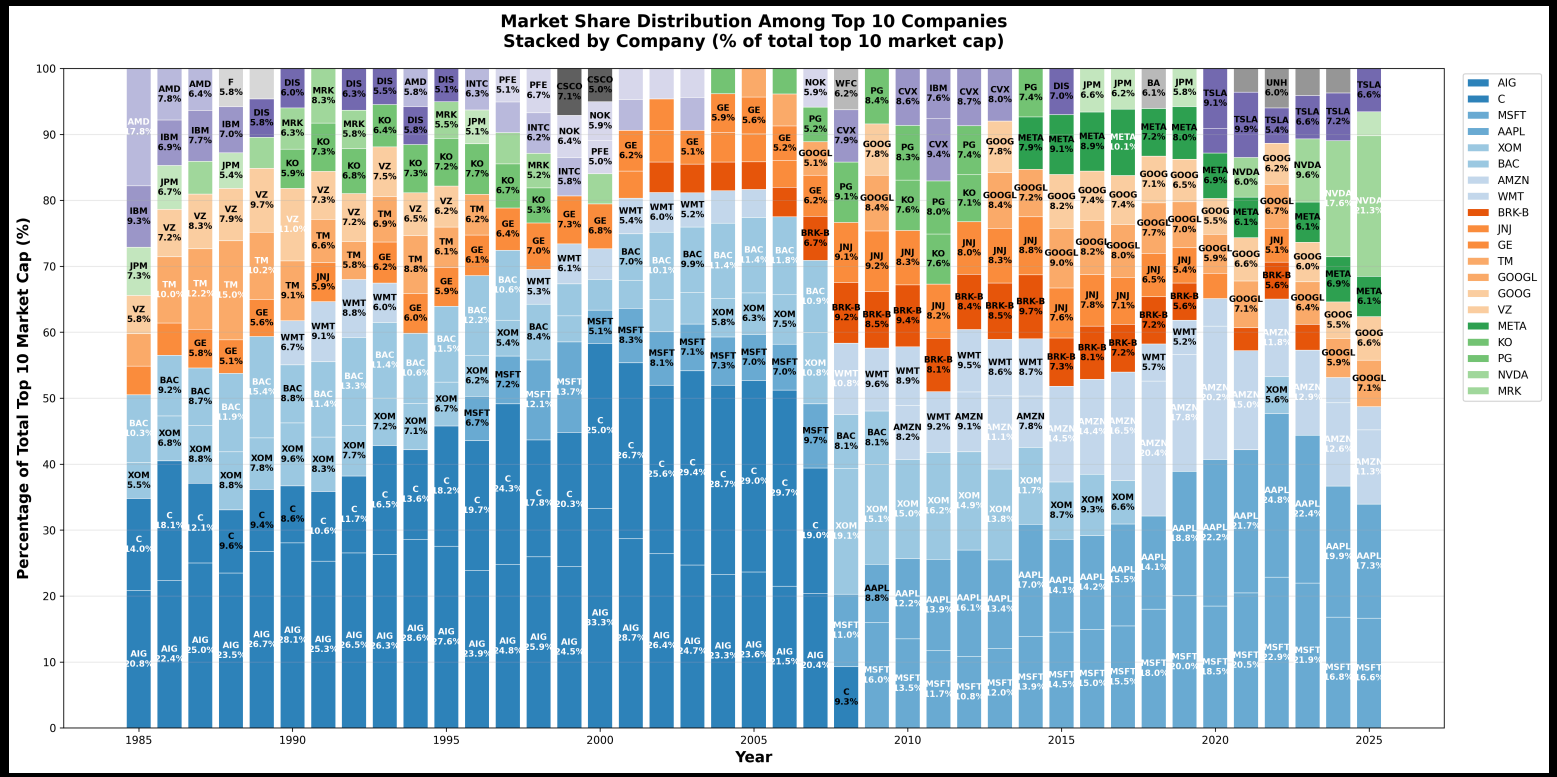
"Will we run out of data? Limits of LLM scaling based on human-generated data" (Villalobos et al. https://arxiv.org/pdf/2211.04325)

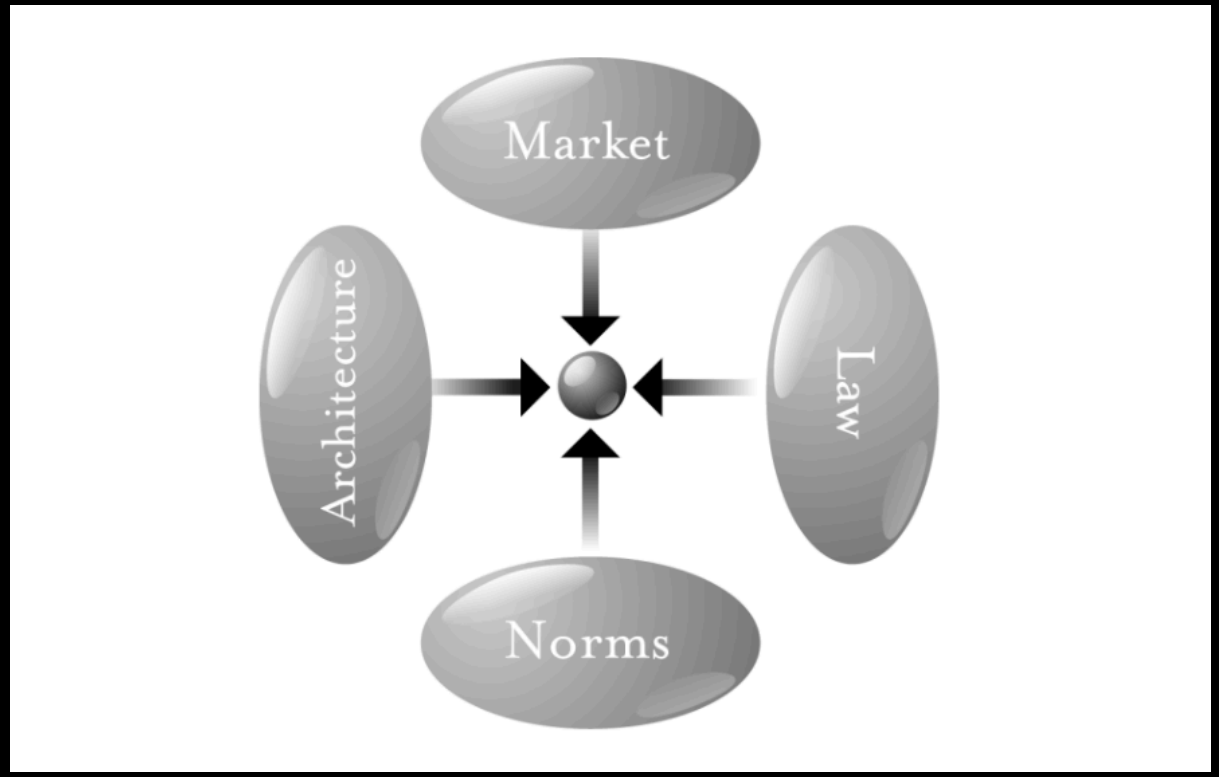
log(tokens) vs time, WWW-scale shown



"Will we run out of data? Limits of LLM scaling based on human-generated data" (Villalobos et al. https://arxiv.org/pdf/2211.04325)

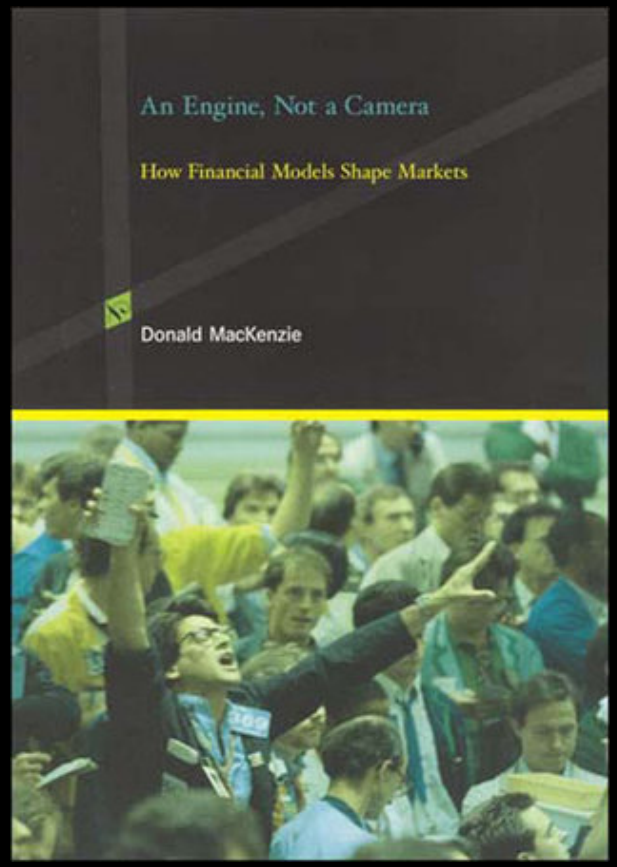
> "technologies...change almost annually" (cf. bit.ly/MarketViz)





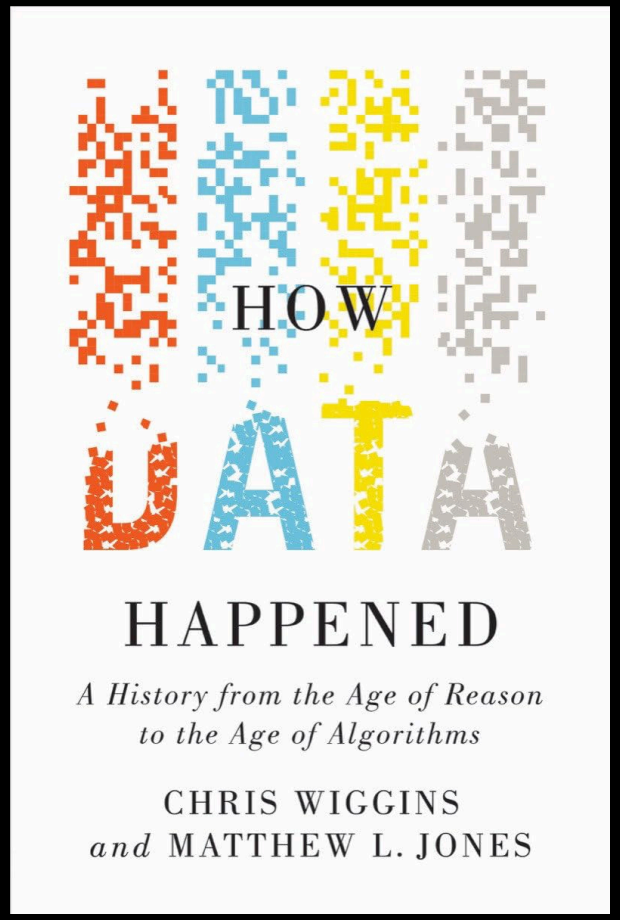
Lessig 1998, 1999, 2006: modes, dynamics (inc. people power)

> "Solutions cause problems which call for further solutions"



- run out of 'net -> synthetic data (img, text, vid...) -> model collapse
- custom LLMs for specific industries-> data poisoning
- price "truth" vs. "truthiness"?
- unlike SAGE → SABRE (.mil->.com) this is .mil->.com->.mil

thank you



Human activity generates data.

Human thought generates tokens.

chris.wiggins@columbia.edu

References

Banko, Michele and Eric Brill. "Scaling to Very Very Large Corpora for Natural Language Disambiguation." *ACL* (2001).

Brin, Sergey, Rajeev Motwani, Lawrence Page, and Terry Winograd. "What can you do with a Web in your Pocket?" *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* (1998).

Butler, Smedley D. *War is a Racket*. 1935.

Goldhaber, Michael H. "The Attention Economy and the Net." *First Monday* 2 (1997).

Halevy, Alon, Peter Norvig, and Fernando Pereira. "The Unreasonable Effectiveness of Data." *IEEE Intelligent Systems* 24.2 (2009): 8-12.

References (continued)

Lessig, Lawrence. *Code and Other Laws of Cyberspace*. 1999.

Simon, Herbert A. "Designing Organizations for an Information-Rich World." In *Computers, Communication, and the Public Interest*, edited by Martin Greenberger. Baltimore, MD: The Johns Hopkins Press, 1971.

Wiggins, Chris and Matthew L. Jones. *How Data Happened: A History from the Age of Reason to the Age of Algorithms*. W. W. Norton & Company, 2023.