Four Takeaways on the Race to Amass Data for A.I.

nytimes.com/2024/04/06/technology/ai-data-tech-takeaways.html

Online data has long been a valuable commodity. For years, Meta and Google have used data to target their online advertising. Netflix and Spotify have used it to recommend more movies and music. Political candidates have turned to data to learn which groups of voters to train their sights on.

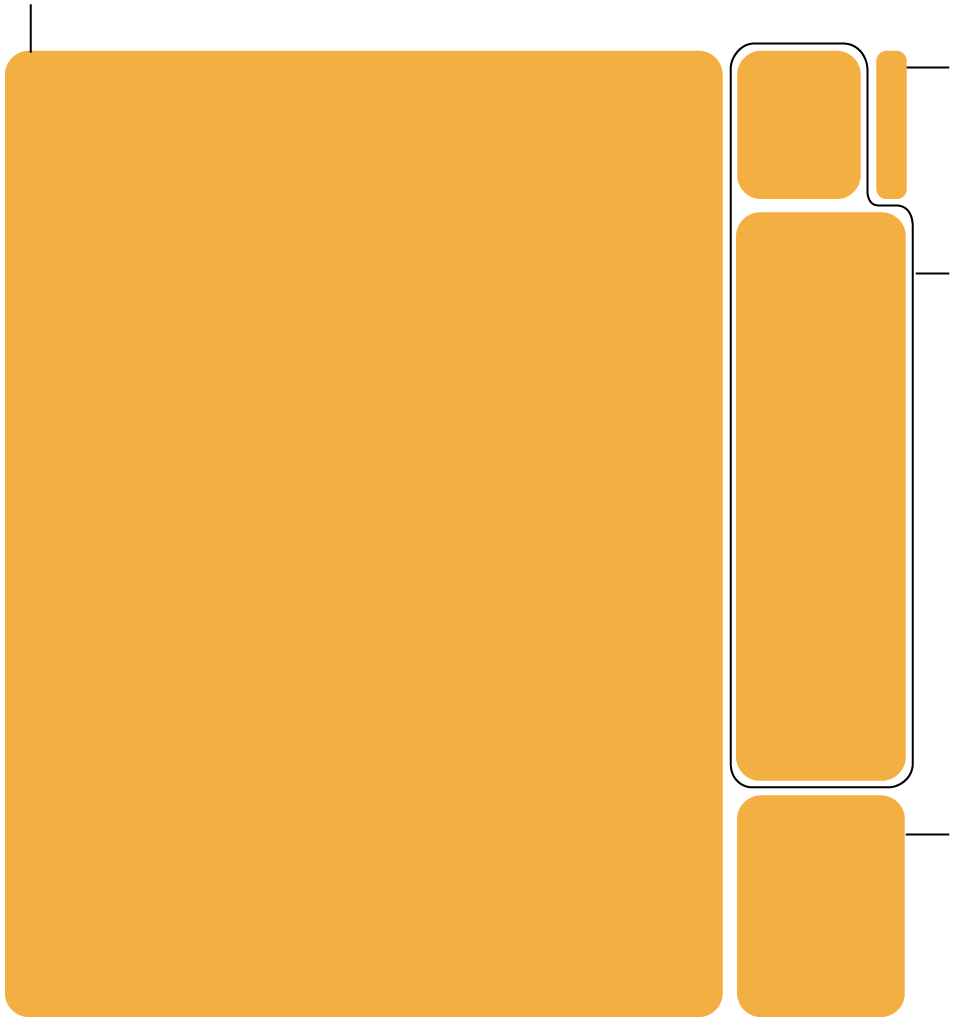
Over the last 18 months, it has become increasingly clear that digital data is also crucial in the development of artificial intelligence. Here's what to know.

The more data, the better.

The success of A.I. depends on data. That's because A.I. models become more accurate and more humanlike with more data.

In the same way that a student learns by reading more books, essays and other information, large language models — the systems that are the basis of chatbots — also become more accurate and more powerful if they are fed more data.

Some large language models, such as OpenAI's GPT-3, released in 2020, were trained on hundreds of billions of "tokens," which are essentially words or pieces of words. More recent large language models were trained on more than three trillion tokens.



Common Crawl

Text from web pages collected since 2007.

Wikipedia

(3 billion tokens)

English-language

Wikipedia pages.

12

billion

Books 1 and Books 2

OpenAI has not explained the contents of these datasets. They are believed to contain text from millions of published books.

55 billion

410 billion tokens

WebText2

Web pages linked from Reddit that received three or more upvotes – an indication of approval from users.

19 billion

Online data is a precious and finite resource.

Tech companies are using up publicly available online data to develop their A.I. models, faster than new data is being produced. According to one prediction, high-quality digital data will be exhausted by 2026.

Tech companies are going to great lengths to obtain more data.

In the race for more data, OpenAI, Google and Meta are turning to new tools, changing their terms of service and engaging in internal debates.

At OpenAI, researchers created a program in 2021 that converted the audio of YouTube videos into text and then fed the transcripts into one of its A.I. models, going against YouTube's terms of service, people with knowledge of the matter said.

(The New York Times has sued OpenAI and Microsoft for using copyrighted news articles without permission for A.I. development. OpenAI and Microsoft have said they used news articles in transformative ways that did not violate copyright law.)

Google, which owns YouTube, also used YouTube data to develop its A.I. models, wading into a legal gray area of copyright, people with knowledge of the action said. And Google revised its privacy policy last year so it could use publicly available material to develop more of its A.I. products.

At Meta, executives and lawyers last year debated how to get more data for A.I. development and discussed buying a major publisher like Simon & Schuster. In private meetings, they weighed the possibility of putting copyrighted works into their A.I. model, even if it meant they would be sued later, according to recordings of the meetings, which were obtained by The Times.

One solution may be ‘synthetic’ data.

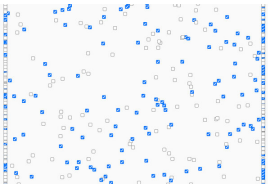
OpenAI, Google and other companies are exploring using their A.I. to create more data. The result would be what is known as “synthetic” data. The idea is that A.I. models generate new text that can then be used to build better A.I.

Synthetic data is risky because A.I. models can make errors. Relying on such data can compound those mistakes.

The A.I. Race

Inside the A.I. Arms Race That Changed Silicon Valley Forever
Dec. 5, 2023

Ego, Fear and Money: How the A.I. Fuse Was Lit
Dec. 3, 2023



The Era of Borderless Data Is Ending
May 23, 2022

Explore Our Coverage of Artificial Intelligence

News and Analysis

- U.S. clinics are starting to offer patients a new service: having their mammograms read not just by a radiologist, but also by an A.I. model.

- OpenAI unveiled Voice Engine, an A.I. technology that can recreate a person's voice from a 15-second recording.
- Amazon said it had added $2.75 billion to its investment in Anthropic, an A.I. start-up that competes with companies like OpenAI and Google.

The Age of A.I.

- Teen girls are confronting an epidemic of deepfake nudes in schools across the United States, as middle and high school students have used A.I. to fabricate explicit images of female classmates.
- A.I. is peering into restaurant garbage pails and crunching grocery-store data to try to figure out how to send less uneaten food into dumpsters.
- David Autor, an M.I.T. economist and tech skeptic, argues that A.I. is fundamentally different from past waves of computerization.
- Economists doubt that A.I. is already visible in productivity data. Big companies, however, talk often about adopting it to improve efficiency.
- The Caribbean island Anguilla made $32 million last year, more than 10& of its G.D.P., from companies registering web addresses that end in .ai.
- When it comes to the A.I. that powers chatbots, China trails the United States. But when it comes to producing the scientists behind a new generation of humanoid technologies, China is pulling ahead.

Related Content