Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning

Pablo Villalobos*, Jaime Sevilla*†, Lennart Heim*§, Tamay Besiroglu*‡, Marius Hobbhahn *¶, Anson Ho*

***Abstract*—We analyze the growth of dataset sizes used in machine learning for natural language processing and computer vision, and extrapolate these using two methods; using the historical growth rate and estimating the compute-optimal dataset size for future predicted compute budgets. We investigate the growth in data usage by estimating the total stock of unlabeled data available on the internet over the coming decades. Our analysis indicates that the stock of high-quality language data will be exhausted soon; likely before 2026. By contrast, the stock of low-quality language data and image data will be exhausted only much later; between 2030 and 2050 (for low-quality language) and between 2030 and 2060 (for images). Our work suggests that the current trend of ever-growing ML models that rely on enormous datasets might slow down if data efficiency is not drastically improved or new sources of data become available.**

Key Takeaways

- We project the growth of training datasets for vision and language models using both the historical growth rate and the compute-optimal dataset size given current scaling laws and existing compute availability estimates (Section III-A).
- We also project the growth in the total stock of unlabeled data, including high-quality language data (Section III-B).
- Language datasets have grown exponentially by more than 50% per year, and contain up to 2e12 words as of October 2022. (section IV-A)
- The stock of language data currently grows by 7% yearly, but our model predicts a slowdown to 1% by 2100. This stock is currently between 7e13 and 7e16 words, which is 1.5 to 4.5 orders of magnitude larger than the largest datasets used today (Section IV-B1).
- Based on these trends, we will likely run out of language data between 2030 and 2050 (Section IV-D).
- However, language models are usually trained on high-quality data. The stock of high-quality language data is between 4.6e12 and 1.7e13 words, which is less than one order of magnitude larger than the largest datasets (Section IV-B2).
- We are within one order of magnitude of exhausting high-quality data, and this will likely happen between 2023 and 2027 (Section IV-D).
- Projecting the future growth of image datasets is less obvious than for language, because the historical trend stopped in the past four years[1]. However, the growth rate seems likely to be around 18% to 31% per year. The current largest dataset is 3e9 images (Section IV-A).
- The stock of vision data currently grows by 8% yearly, but will eventually slow down to 1% by 2100. It is currently between 8.11e12 and 2.3e13 images – three to four orders of magnitude larger than the largest datasets used today (Section IV-C).
- Projecting these trends highlights that we will likely run out of vision data between 2030 to 2070 (Section IV-D).

*Epoch, †University of Aberdeen, ‡MIT Computer Science & Artificial Intelligence Laboratory, §Centre for the Governance of AI, ¶University of Tübingen

[1]New models appeared which use much more data than what was the case in the previous years, see [1].

I. Introduction

Training data is one of the three main factors that determine the performance of Machine Learning (ML) models, together with algorithms and compute. Current understanding of scaling laws suggests that future ML capabilities will strongly depend on the availability of large amounts of data to train large models [2, 3].

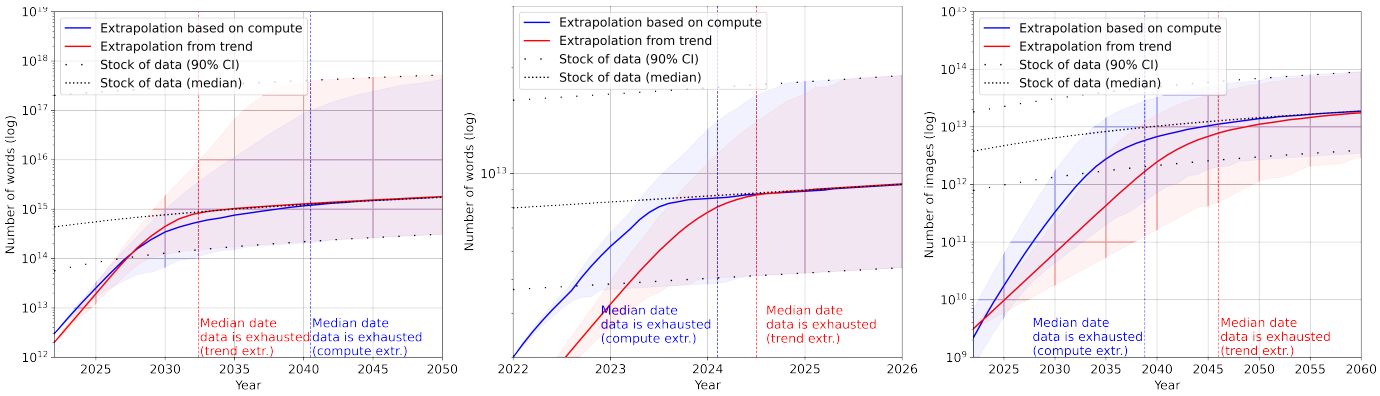
Previous work compiled a database of more than 200 training datasets used in ML models [1] and estimated historical rates of growth in dataset size for vision and language models.

To learn about the limits of this trend, we developed probabilistic models to estimate the total amount of image and language data that will be available between 2022 and 2100. Based on our dataset size trend projections, we then estimated the limit of these trends due to the exhaustion of available data.

II. Previous work

Stock of data: There have been several estimates of the size of the internet and the total amount of information available [4, 5, 6]. However, in recent years, these types of reports have not provided breakdowns of different data modalities (for example into the number of images, videos, or blog posts), and instead aggregated all data modalities into a single number in bytes [7].

Data bottleneck in ML: In [8], the author estimated the stock of high-quality data and used the scaling laws [3] to predict that the stock of data is not enough to scale language models more than 1.6x the size of DeepMind's Chinchilla language model [3] using compute-optimal scaling. We improve this analysis by creating explicit models of dataset size growth and more detailed estimations of the stock of data over time, which allows us to predict the date that datasets will become as large as the total stock of data.



(a) Projections for low-quality language data

(b) Projections for high-quality language data

(c) Projections for vision data

Fig. 1: Projections of data usage. Each graph shows two extrapolations of data usage, one from past trends and one from compute availability estimations plus scaling laws. Both projections are constrained to be lower than the estimated data stock. In all three cases, this constraint causes a slowdown in data usage growth.

III. Methods

A. *Projecting growth in training dataset sizes*

Previous work compiled historical trends of dataset size work for different application domains[2] [1].

Our definition of *dataset size* is the number of unique datapoints on which the model is trained. The definition of "datapoint" is different for each domain. In particular, for language data we define a datapoint as a word, and for image data we define a datapoint as an image. Additional details on this choice of dataset size metric can be found in [1].

Using the historical trend, together with the size of the largest datasets used to date, we can estimate the future evolution of dataset sizes. However, this projection naively assumes that the past trend will be sustained indefinitely. In reality, there are constraints on the amount of data that a model can be trained on. One of the most important constraints is compute availability. This is because increasing the amount of training data for a given model requires additional compute, and the amount of compute that can be used is limited by the supply of hardware and the cost of buying or renting that hardware.

To account for this constraint, we make another projection, based on compute availability and the compute-optimal dataset size. Scaling laws can be used to predict the optimal balance of model size and dataset size for a given compute budget (measured in FLOP) [2, 3]. Concretely, the optimal dataset size is proportional to the square root of the compute budget ($D \propto \sqrt{C}$).

Previous work [9], projected the available compute to the largest training into the future[3]. We use those projections to estimate the optimal training dataset size that will be achievable in each future year.

B. *Estimating data accumulation rates*

In recent years, unsupervised learning has successfully created foundation models that can be fine-tuned for several tasks using small amounts of labeled data and large amounts of unlabeled data. In addition, unsupervised models have also proved able to generate valuable pseudo-labels for unlabeled data [10]. For these reasons, we will focus on the stock and accumulation rates of unlabeled data, even if the amount of labeled data is much lower[4].

Before delving into the details, let us consider a theoretical framework of what we expect the data accumulation rate to look like. The vast majority of data is user-generated and is stored in social media platforms, blogs, forums, etc. There are three factors that determine how much content is produced in a given period: human population, internet penetration rate, and the average amount of data produced by each internet user.

Human population has been extensively studied so we use the standard United Nations projections [11]. Internet penetration (the percentage of the population who uses the Internet) grows as an S-curve from 0% in 1990 to 50% in 2018 to over 60% today [12]. We model this as a sigmoid function of time and fit it to the data in [12].

The average amount of data produced by users changes over geography and time according to internet usage trends, and is not easy to analyze[5]. For simplicity, let us assume the average amount of data produced by users is constant over time.

This model of Internet population (the number of Internet users) closely matches the historical number of Internet users

[2]The domains that were included were vision, language, recommendation, speech, drawing, and games. However, there is only significant data for vision and language.

[3]Note that this projection has a wide range of uncertainty and includes scenarios in which spending on compute grows orders of magnitude over current levels, up to 1% of GDP.

[4]Note that while transfer learning vastly reduces the need for labeled data, it does not eliminate it. In addition, labeled data is usually much harder to acquire than unlabeled data. Therefore, labeled data might turn out to be a bottleneck even though it is required in smaller quantities.

[5]Doing so would require taking into account the effects of culture, demography and socioeconomic development in different countries and times, which is out of the scope of this paper.

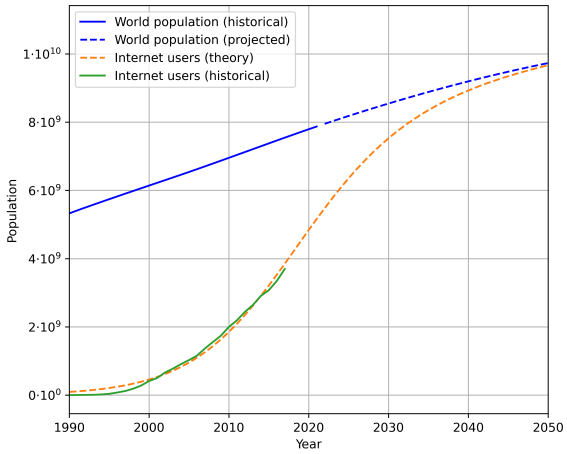


Fig. 2: Real and projected evolution of internet users.

(Figure 2). To test its ability to predict Internet data generation, we conducted an empirical test by fitting this model to Reddit submission data. This model fit the data better than both exponential and sigmoid models (see Appendix C).

C. *High-quality data*

We have developed a model for the accumulation rate of user-generated content. However, for language data, this kind of content tends to be of lower quality than more specialized language data like books or scientific papers. Models trained on the latter kind of data perform better [13], so it is common practice to use it for training language models [14, 15, 3]. We have little insight on data quality for image models nor how to identify high-quality image data[6], so we will focus on language in this section.

Because of our limited insights and unawareness of research into the tradeoffs involved in using high- versus low-quality data, we provide estimates and growth projections for both high- and low-quality data separately. To identify high-quality date we defer to the expertise of practitioners and look at the composition of the datasets used to train large language models. The most common sources in these datasets are books, news articles, scientific papers, Wikipedia, and filtered web content[7][8].

A common property of these sources is that they contain data that has passed usefulness or quality filters. For example, in the case of news, scientific articles, or open-source code projects, the usefulness filter is imposed by professional standards (like peer review). In the case of Wikipedia, the filter is standing the test of time in a community of dedicated editors. In the case of filtered web content, the filter is receiving positive engagement from many users. While imperfect, this property can help us identify additional sources of high-quality data, so we will use it as our working definition of high-quality data.

Some high-quality data, such as filtered web content and Wikipedia, is generated by dedicated internet contributors. This means we can use the same model developed for general user-generated content.

However, other sources of high-quality data are generated by subject matter experts (such as scientists, authors, and open-source developers). In this case, the generation rate is not determined by human population or internet penetration but by the size of the economy and the share of the economy devoted to creative sectors (like science and art).

OECD countries have spent roughly 2% of their GDP on R&D over the past 20 years [17]. This number is increasing slowly, but we will assume it is mostly constant. So the data accumulation rate should be roughly proportional to the size of the world economy, which grows around 4% each year. This prediction is consistent with the observed growth in scientific publications [18].

We estimate the proportion of these two classes of data (dedicated contributors and professionals) in high-quality datasets by looking at existing datasets and classifying each of their subcomponents into a class.

D. *Limitations*

There are a number of reasons why our estimates of the growth rate of dataset sizes might be incorrect:

- There might be less need for data in the future to achieve equivalent levels of performance. This is particularly likely since there have previously been large increases in data efficiency in other domains[19, 8].
- Compute availability might grow slower than expected for a number of reasons, including technical obstacles to efficiency increases, supply chain disruptions, or reduced willingness to spend.
- Current scaling laws could be wrong, as has happened in the past[9]. Even if there is no additional increase in data efficiency, perhaps there are better ways of scaling that use less data.
- Multimodal models might prove to perform better than models with a single modality via transfer learning. This would effectively increase the data stock to encompass a combination of the stocks of all data modalities.

In addition, there are some limitations in our estimates of the stock of data:

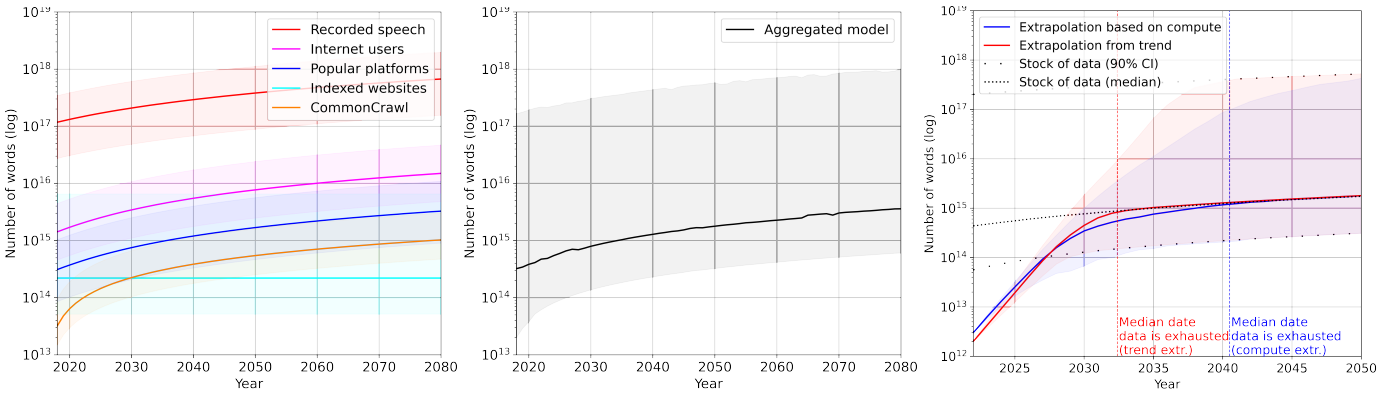
- The use of synthetic data could make the stock of data virtually infinite. We are uncertain about the usefulness and cost of synthetic data for training.

[6]Other than very crude metrics like image resolution. For example, comparing robustness across distributional shifts of image-text models trained on different commonly-used datasets shows that there is no single dataset that induces better robustness across all shifts [16]

[7]Filtered web content is regular web content selected using a proxy measure of quality, like the number of upvotes of a link shared in Reddit. The MassiveWeb and WebText datasets were built in this way.

[8]Other common sources are GitHub (for code), subtitles and transcriptions of educational videos, podcasts or parliamentary sessions, and emails.

[9]In [2] the authors recommended increasing the training dataset size fivefold for each tenfold increase in compute. In the more recent [3], they revisit the problem and recommend instead increasing the training dataset size threefold for each tenfold increase in compute.



(a) Comparison of the different data stock models.

(b) Aggregated data stock model.

(c) Projection of historical trend of training dataset sizes and available data stock.

Fig. 3: Models of low-quality language data.

- Big economic shifts might significantly impact the production of data. For example, large-scale adoption of self-driving cars would result in an unprecedented amount of road video recordings.
- Similarly, actors with big budgets (such as governments or large corporations) might be able to increase the production of data with enough spending, especially in the case of high-quality data for niche domains. Some possibilities are widespread screen recording or mass surveillance.
- We might find better ways to extract high-quality data from low-quality sources, for example, by coming up with robust automatic quality metrics.

IV. Analysis

A. *Trends in dataset size*

Previous work [1] identified the historical rate of growth for training datasets in different domains. Since there is only significant data for the language and vision domains, we will limit our analysis to those two domains. The trends are summarized in Table I.

Domain	Doubling time median and CI (months)		Largest training dataset (datapoints)
Language	15.8	[11.2; 20.9]	2e12
Vision	41.5	[30.4; 48.3]	3e9

TABLE I: Trends in training dataset size for vision and language models.

B. *Language*

1) *Low-quality data*

We have used five different models to estimate the amount of data and the accumulation rate. Table II summarizes these different models, which are further illustrated in Figure 3a and explained in more detail in Appendix A. The aggregated model finds an estimated current total stock between 6.9e13 and 7.1e16 words, and current growth between 6.41% and 17.49% per year.

Note that the high end of this estimate comes from two highly theoretical models that we trust the least. The way we interpret this range is: 1e14 words is what is very likely available to single, well-funded actors such as Google; 1e15 words is what is available to the combined group of all major actors (all tech companies); 1e16 words is what humanity might be able to collectively produce with a worldwide, multiyear effort, employing practices such as recording all text messages, phone calls and video meetings, practices which are currently very far outside the Overton window.

Using the aggregated data stock model as an upper bound for scaling datasets, we project the size of training datasets and find that it grows rapidly until it exhausts the stock of data. After this point, growth slows down substantially (Figure 3c).

Model	Stock of data (#words)	Growth rate
Recorded speech	1.46e17	5.2%
	[3.41e16; 4.28e17]	[4.95%; 5.2%]
Internet users	2.01e15	8.14%
	[6.47e14; 6.28e15]	[7.89%; 8.14%]
Popular platforms	4.41e14	8.14%
	[1.21e14; 1.46e15]	[7.89%; 8.14%]
CommonCrawl	9.62e13	16.68%
	[4.45e13; 2.84e14]	[16.41%; 16.68%]
Indexed websites	2.21e14	NA
	[5.16e13; 6.53e15]	
Aggregated model	**7.41e14**	**7.15%**
	[6.85e13; 7.13e16]	**[6.41%; 17.49%]**

TABLE II: Median and 90% CI of estimates of high-quality language data accumulation.

2) *High-quality data*

We studied high-quality data by looking at the composition of several high-quality datasets and determining how much each component can be scaled. We considered three datasets: The Pile [13], MassiveText [3], and the PaLM pretraining dataset [15].

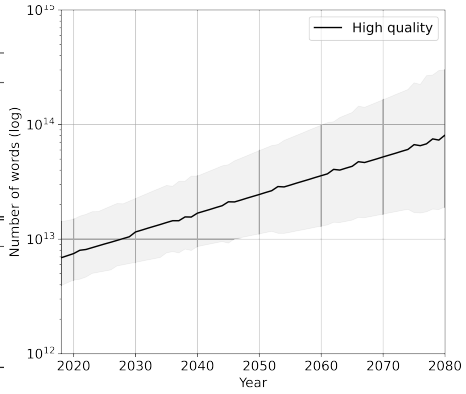
From these, we can see that high-quality datasets are usually

Component	Raw Size
Pile-CC	227.12 GiB
PubMed Central	90.27 GiB
Books3†	100.96 GiB
OpenWebText2	62.77 GiB
ArXiv	56.21 GiB
Github	95.16 GiB
FreeLaw	51.15 GiB
Stack Exchange	32.20 GiB
USPTO Backgrounds	22.90 GiB
PubMed Abstracts	19.26 GiB
Gutenberg (PG-19)†	10.88 GiB
OpenSubtitles†	12.98 GiB

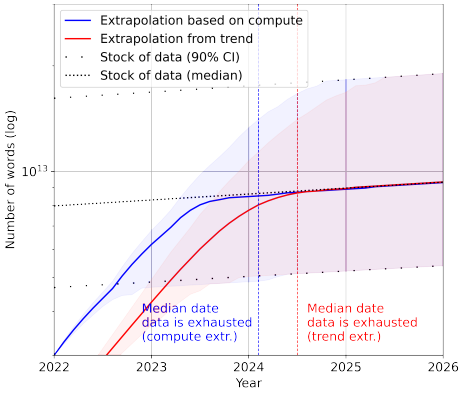
Data source
Social media conversations
Filtered webpages
Books (English)
GitHub (code)
Wikipedia (multilingual)
News (English)

	Disk Size	Documents
MassiveWeb	1.9 TB	604M
Books	2.1 TB	4M
C4	0.75 TB	361M
News	2.7 TB	1.1B
GitHub	3.1 TB	142M
Wikipedia	0.001 TB	6M

(a) Composition of high-quality datasets: The Pile (left), PaLM (top-right), MassiveText (bottom-right)



(b) Data stock model.



(c) Projection of historical trend of training dataset sizes and available data stock.

Fig. 4: Models of high-quality language data.

composed of 50% scraped user-generated content (Pile-CC, OpenWebText2, social media conversations, filtered webpages, MassiveWeb, C4), 15-20% books, 10-20% scientific papers, <10% code and <10% news. In addition, they all incorporate known small very-high-quality datasets like Wikipedia (Figure 4a).

We estimated the amount of available text in digitized books, public GitHub repositories, and scientific papers. Assuming all of these form between 30% to 50% of a hypothetical high-quality dataset, we could reach **9e12** [4.6e12; 1.7e13] words. We assume the amount of high-quality data grows at 4-5% per year in line with the world economy, as explained in the Introduction (see Figure 4b). Details of the model can be found in Appendix A.

Projecting the growth of language datasets using the high-quality stock instead of the low-quality stock as an upper bound, we find the same pattern of slowdown, with the distinction that the slowdown happens much earlier, before 2026 (Figure 4c).

C. *Vision*

We used two different estimates for vision: one produced by Rise Above Research [20], and one using the combined images and videos posted to the most popular social media platforms. The aggregated model shows there are between 8.11e12 and 2.3e13 images on the internet today, with the current yearly growth rate around 8%. The models are summarized in Table III and Figure 5a.

Using the aggregated data stock model as an upper bound for scaling datasets, we project the size of training datasets from both the historical trend and the compute-optimal extrapolation. The historical projection is very uncertain since we do not yet know if the recent high outliers indicate a new higher-growth trend. The compute projection is also more uncertain than the corresponding projection for language because we do not have a great understanding of scaling laws for vision[10].

[10]This is because images can have different resolutions, so image tokenization is more variable than text tokenization.

Similarly to the case of language, dataset sizes grow exponentially until reaching the size of the data stock, at which point they revert to a much slower rate of growth (Figure 5c).

We do not understand the impact of data quality for unlabeled vision data and how to distinguish high-quality data, so we did not attempt to estimate it.

Model	Stock of data (#images)	Growth rate
Popular platforms	9.48e12	8.14%
	[5.02e12 ; 1.79e13]	[7.89% ; 8.14%]
External estimate	1.28e12	8.14%
	[6.5e11 ; 2.58e12]	[7.89% ; 8.14%]
Aggregated model	**4.36e12**	**8.14%**
	[8.11e12 ; 2.3e13]	**[7.89% ; 8.14%]**

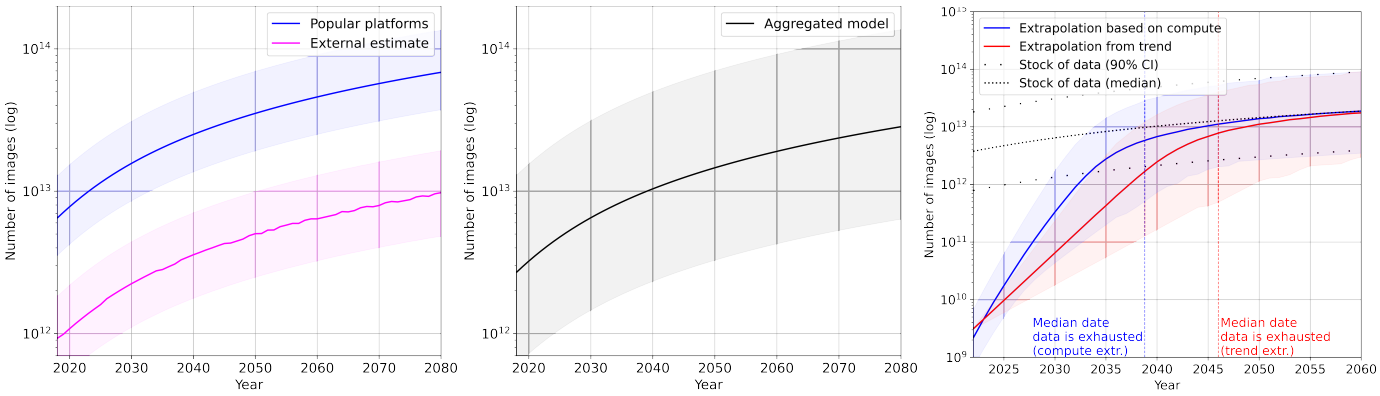
TABLE III: Summary of estimates of image data accumulation. The bottom row contains an aggregate of all the models.

D. *Will data become a bottleneck?*

So far we have found that data stocks grow much slower than training dataset sizes (see Figures 3c, 4c, and 5c). This means that exhausting our data stocks is inevitable if current trends continue. In addition, the high-quality data stock is much smaller than the low-quality stock. The two dataset size projections, based on historical trends and compute availability extrapolations, are very similar in the first years, but later diverge.

We computed the probability that exhaustion will happen each year for each of our projections of data stock and dataset size (Figure 6). While there is significant uncertainty in the exhaustion dates for low-quality language and vision stocks, it seems unlikely that it will happen before 2030 or after 2060. However, the high-quality language stock will almost surely be exhausted before 2027 if current trends continue. The quantiles for these distributions are shown in Table IV.

[2023.55, 2024.5, 2025.75]



(a) Comparison of data stock models.

(b) Aggregated data stock model.

(c) Projection of historical trend of training dataset sizes and available data stock.

Fig. 5: Models of vision data.

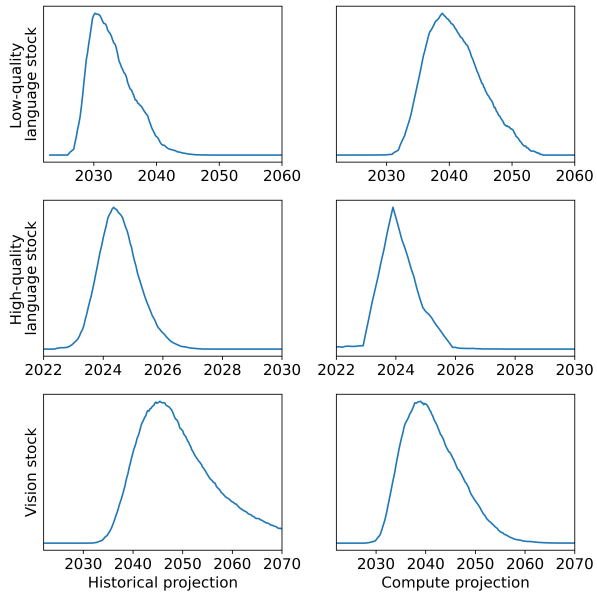


Fig. 6: Distribution of exhaustion dates for each intersection of the data availability trend and data consumptiin trend. Note that the time scale is different for each kind of data.

	Historical projection	Compute projection
Low-quality language stock	**2032.4** [2028.4 ; 2039.2]	**2040.5** [2034.6 ; 2048.9]
High-quality language stock	**2024.5** [2023.55 ; 2025.75]	**2024.1** [2023.2 ; 2025.3]
Vision stock	**2046** [2037 ; 2062.8]	**2038.8** [2032.0 ; 2049.8]

TABLE IV: Median and 90% CI of exhaustion year for each of the intersections.

V. Discussion

Scaling laws for language models indicate that scaling is dependent on the amount of available data [3, 8]. Under this view, around half of the improvement in language models over the past four years comes from training them on more data. Without further room to scale datasets, this would lead to a slowdown in AI progress.

The accumulation rate of data for both language and vision models is much slower than the growth in dataset size that we have observed so far, both historically and taking compute constraints into account. As a consequence, we might be headed for a bottleneck in training data. This would happen between 2030 and 2040 for language models and between 2030 and 2060 for image models (Figure 6).

This is particularly true for high-quality language data, which seems likely to be exhausted by 2027. It is unclear whether large enough datasets can substitute for poor data quality, but even if this is the case, it would not be enough to completely avoid the slowdown, since our ability to scale training datasets is also limited by compute availability.

Given these projections, it might be tempting to conclude that a slowdown is inevitable. However, we have significant reasons to believe that our models are not adequately capturing the evolution of ML progress (see **Limitations**).

In particular, the future evolution of data efficiency and the impact of data quality on performance are crucial to predict future data requirements. Unfortunately, our understanding of these variables is insufficient to provide detailed forecasts. Future work could try to incorporate these considerations into the analysis.

VI. Conclusion

We have projected the growth of training dataset sizes and data stocks. Data stocks grow at a much slower pace than dataset sizes, so if current trends continue, datasets will eventually stop growing due to data exhaustion. Our models show this is likely to happen between 2030 and 2040 for language data, and between 2030 and 2060 for vision data. In addition, high-quality language data will be exhausted by 2026.

If our assumptions are correct, data will become the main bottleneck for scaling ML models, and we might see a slowdown in AI progress as a result. However, as outlined, there

are multiple reasons to doubt that these trends will continue as projected, such as the possibility of algorithmic innovations in data efficiency.

Acknowledgments

Thanks to Ben Cottier, Michael Aird, David Atkinson, Matthew Barnett and Keith Wynroe for their comments on earlier drafts of this article. Also thanks to Adam Papineau for reviewing and editing.

References

[1] P. Villalobos and A. Ho, "Trends in training dataset sizes," https://epochai.org/blog/trends-in-training-dataset-sizes, 2022, accessed: 2022-09-27.

[2] J. Kaplan *et al.*, "Scaling laws for neural language models," 2020.

[3] J. Hoffmann *et al.*, "Training compute-optimal large language models," 2022.

[4] K. Coffman and A. Odlyzko, "The size and growth rate of the internet," 1998.

[5] B. Murray H. and A. Moore, "Sizing the internet," Cyveillance, Tech. Rep., 7 2000.

[6] P. Lyman and H.R. Varian, "How much information," 2003.

[7] D. Reinsel, J. Gantz, and J. Rydning, "The digitization of the world from edge to core," International Data Corporation, Tech. Rep., 11 2018.

[8] nostalgebraist, "chinchilla's wild implications," 2022.

[9] L.H. Tamay Besiroglu and J. Sevilla, "Projecting compute trends in machine learning," https://epochai.org/blog/projecting-compute-trends, 2022, accessed: 2022-09-27.

[10] H. Pham *et al.*, "Meta pseudo labels," 2020.

[11] "World population prospects 2022, online edition," 2022.

[12] H. Ritchie and M. Roser, "Technology adoption," *Our World in Data*, 2017, https://ourworldindata.org/technology-adoption.

[13] L. Gao *et al.*, "The pile: An 800gb dataset of diverse text for language modeling," 2021.

[14] N. Du *et al.*, "Glam: Efficient scaling of language models with mixture-of-experts," 2021.

[15] A. Chowdhery *et al.*, "Palm: Scaling language modeling with pathways," 2022.

[16] T. Nguyen *et al.*, "Quality not quantity: On the interaction between dataset design and robustness of clip," 2022.

[17] OECD, "Gross domestic spending on R&D (indicator)," 2022.

[18] "Publications output: U.s. trends and international comparisons," 2019.

[19] W. Ye *et al.*, "Mastering atari games with limited data," 2021.

[20] E. Lee, "2021 worldwide image capture forecast: 2020 – 2025," 2021.

Appendix

A. *Appendix A: Models of language data accumulation*

In this Appendix we provide a brief overview of the models and the sources for their parameters. The full code is available in https://github.com/epoch-research/data-stock.

1) Total recorded human speech

Consider the amount of raw speech generated by humans. Assume each person pronounces or writes between 5k and 20k words per day.[11] Supoosing that between 0.5% and 50% of those words are digitally recorded[12], then the production of words per person-year would be between 160k and 2.6M. Multiplying this by the number of person-years (the integral of the population size) over a given period we can get the total production of words.

2) Total text produced by internet users

As described in the introduction, the amount of text uploaded to the internet can be modeled as a product of three factors:

- Human population
- Internet penetration
- Amount of text generated in a year by the average internet user

Assume the third factor is constant over time and its value is between 10k and 100k words per year.[13] Then we can fit a sigmoid to internet penetration data, and use existing human population projections as approximations for the first two factors. The product of those three is the total stock.

3) Popular platforms

In an unspecified recent year, around 500 hours of video were uploaded to YouTube each minute.[14] Assuming between 5% and 50% of those contain speech, and 9k words per hour,[15] we get that between 130B and 1.3T words are uploaded to YouTube per year.

Since 2012 there have been around 500M tweets per day.[16] Assuming between 10 and 50 words per tweet, Twitter produces 2-20T words each year.

Several sources estimate the number of daily published blog posts to be around 7.5M.[17] Assuming an average blog post

[11] This corresponds to talking for 30 minutes and 2 hours at 150wpm, respectively. This study (https://doi.org/10.1126/science.1139940) found an average of 16000 words per day, with a standard deviation of 7300.

[12] That is, they are processed by a digital device, be it in a phone conversation, text messages or video meetings.

[13] Professional writers can write a book of 100k words in half a year, and sending 5 text messages per day with ten words each already produces almost 20k words.

[14] This is an old figure from a few years ago that YouTube published in their blog. Can also be found on Statista. Plausibly higher nowadays, maybe up to 800h.

[15] 150 words per minute for 60 minutes.

[16] Source: https://www.internetlivestats.com/twitter-statistics/. Outdated, could be a bit higher today

[17] This number is repeated over the internet, and it's consistent with a single-digit percentage of the world population posting one blog post a week. An example source is https://www.websiterating.com/research/internet-statistics-facts/#chapter-3.

length of 100-1000 words,[18] in a single year between 0.2T and 2T words are produced.

Adding all of those, and assuming they account for between 5% and 40% of the whole internet, between 8.9T and 110T words are produced per year. Multiplying this by the number of internet users normalized to 2022 and integrating, we get the stock at any given year.

4) Indexed websites

The indexed web size has been roughly constant in size at tens of billions of websites[19]. Assuming between 500 and 50k words per website, this means the indexed web contains between 1e13 and 1e15 words.

5) CommonCrawl

CommonCrawl releases monthly crawls which for the past four years have contained 5-10TB of compressed plaintext.[20] Half of the URLs in each monthly crawl are visited for the first time, so let's assume around half of the plaintext is new data. Assuming a compression rate for English between 30% and 85%, the new total yearly amount of uncompressed text is 160TB. At 200M words per GB of text, CommonCrawl produces between 8.3e12 and 4.4e13 words per year.

6) High-quality data

a) Code

In 2020, GitHub archived 21TB of public repositories in the Arctic Vault. Software Heritage has archived 13B source code files. Assuming each of them is 5KB,[21] this gives 65TB of code. So the size of GitHub in 2022 is reasonably between 21TB and 65TB.

The size of a git repository includes all the blobs used for storing the history of the repository. This storage overhead ranges from 100% for a one-year-old repo with <100 commits up to 600% for a very old repository like the Linux kernel. In the MassiveText dataset, the overhead is 3.1TB / 844GB - 1 = 267%. This means that approximately 1/4 of the size is actual source code.

It's not clear how much code constitutes a "word," but to be consistent with natural language we'll take a TB of code to be 200B "words." Note that tokenizers usually have worse compression rates for code than natural text,[22] so the size in tokens will be larger than in "words." In total this is between 320B and 4.2T words.

b) Papers

In 2014, there were an estimated 114M English scientific articles on the web.[23] The Web of Science lists 82M papers in total, with 55M published before 2014. Assuming the coverage of Web of Science is constant over time, the total number of papers has increased by 50% since.[24] This means today there are up to 114M*1.5 = 170M papers.

Average paper length is 6k words,[25] so we get between 600B and 1T words.

c) Books

1M books were published yearly as of 1996 (How much information, 1996). At 4% yearly growth, for 25 years, this is 2.6M published yearly today. But I don't trust this estimate. Other unreliable Internet sources claim between 500k and 4M books published per year.

It's not clear which fraction of published books is digitized. E-book sales are around 10% those of print books,[26] so this seems like a reasonable guess. Another source estimated that there were 3.4M ebooks and 48.5M print books on Amazon a few years ago.[27] So the fraction of digitized books is probably between 2% and 20%.

Using the estimate of yearly publishing and assuming exponential economic growth, the stock of books is the yearly published books divided by the logarithm of the growth rate.

We also have a direct estimate of the stock in 2022: there are 12M ebooks in Amazon kindle,[28] and the Internet Archive has 20M books in its digital library.[29] So the number of ebooks is probably between 10M and 30M.

Taking the average of these two estimates and a length of 100k words per book, we get between 620B and 1.8T words.

d) Total

Assuming books, papers and code represent between 30% and 50% of high-quality data (as found in current datasets) and exponential growth of 4%, we get the estimate of the stock of high-quality data.

Appendix B: Models of vision data accumulation

In this Appendix we provide a brief overview of the models and the sources for their parameters. The full code is available in https://github.com/epoch-research/data-stock.

7) Popular platforms

If we count one second of YouTube video as a single image and assume 10% to 50% of hours contain usable images (not a still background, etc), then the number of images uploaded to YouTube in 2022 is between 120B and 620B.[30]

[18] CommonCrawl documents have around this length, this can be seen in the 'Mean Document Size' column on Table 1 of the The Pile paper. It is also easy to find claims on the internet that the optimal blog post size to get more engagement is 1000 words.

[19] According to estimates from https://www.worldwidewebsize.com/

[20] Example: https://commoncrawl.org/2022/07/june-july-2022-crawl-archive-now-available/

[21] 5KB is the average for GitHub, this can be seen from The Pile and MassiveText.

[22] 50% vs 80% in the case of the Gopher tokenizer.

[23] Source: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0093949

[24] This translates to a yearly growth of 4.5%, which matches our expectations.

[25] From the Pile paper. Mean document length is 30KB, so 30*200 words.

[26] Source: https://www.cnbc.com/2019/09/19/physical-books-still-outsell-e-books-and-heres-why.html

[27] Source: https://justpublishingadvice.com/how-many-kindle-ebooks-are-there

[28] Source: https://justpublishingadvice.com/how-many-kindle-ebooks-are-there/#2022_Update_and_new_methodology

[29] Source: https://archive.org/details/texts

[30] Using the figure of 500h of video uploaded per minute found in the popular platforms model for language data.

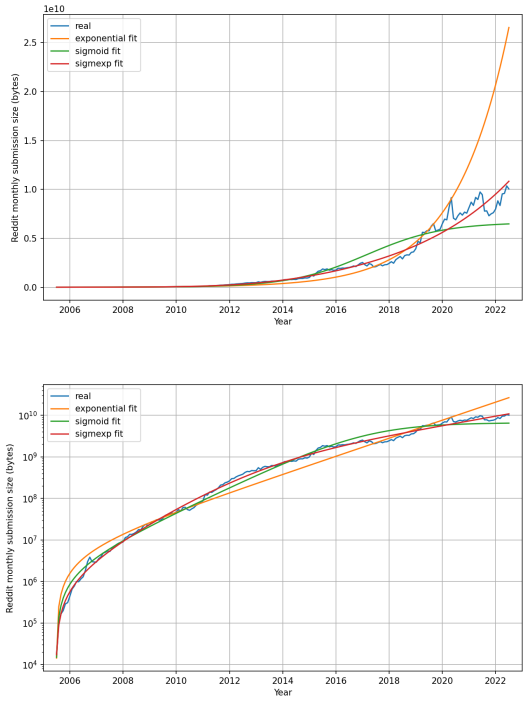


Fig. 7: Monthly user submissions to Reddit, in linear scale(up) and log scale(down). While the three functions appear to fit the data well in the log scale, the linear plot shows that the sigm*exp function predicts much better the recent years.

In 2015, the number of pictures shared daily on Instagram + Snapchat + WhatsApp + Facebook was around 3.2B.[31] Extrapolating that to the current year, the number is probably between 5B and 20B. A single image might be shared on average between 5 and 15 times, so dividing by that we get between 170B and 1T images per year published on social media.

Taking the sum of those two estimates and assuming the yearly production grows in proportion to the internet population, we get the estimate of the total stock.

8) External estimate

This estimate was produced by Rise Apart Research. The number of images in 2022 is between 5e10 and 2e11.[32] We extrapolate this stock by assuming it is proportional to the number of person-years on the internet.

B. Appendix C: Test of the theoretical growth model

We check our theoretical model of data accumulation rates developed in the Methods section on Reddit submission data. The combined sigmoid times exponential function fits the monthly submission history better than either sigmoid or exponential functions alone (Figure 7).

[31] Page 90 of this report: https://www.kleinerperkins.com/perspectives/2016-internet-trends-report/.

[32] Numbers found here: https://photutorial.com/photos-statistics/, primary source here: https://riseaboveresearch.com/rar-reports/2021-worldwide-image-capture-forecast-2020-2025/ (paywalled commercial research)