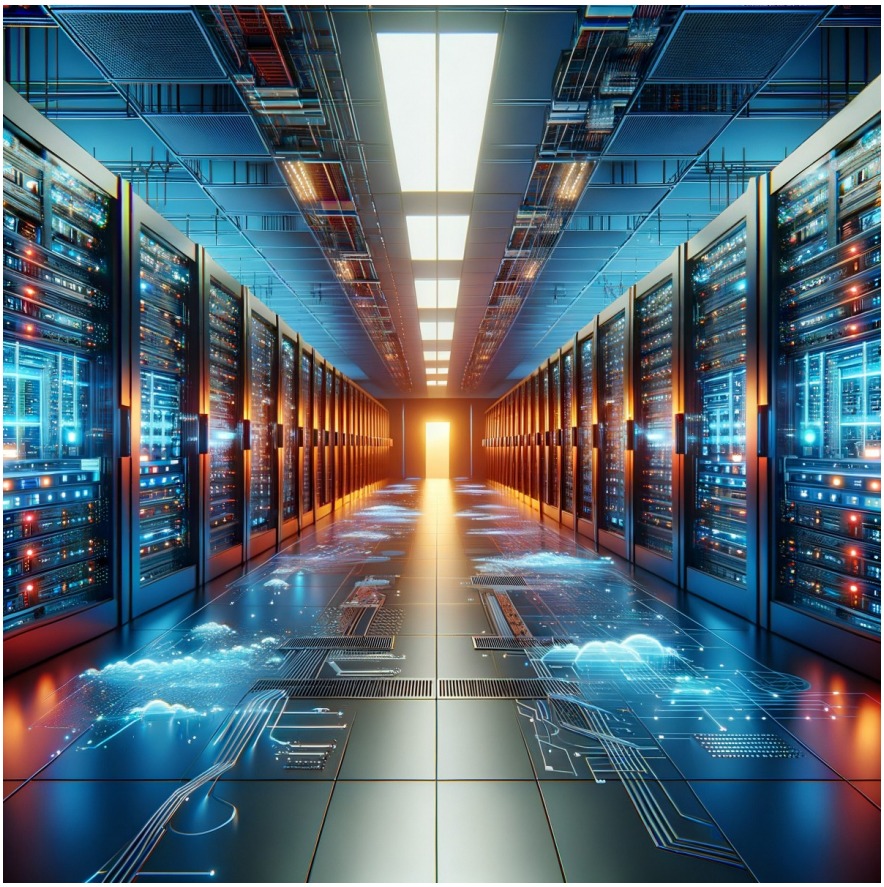
Data is the new gold for AI development

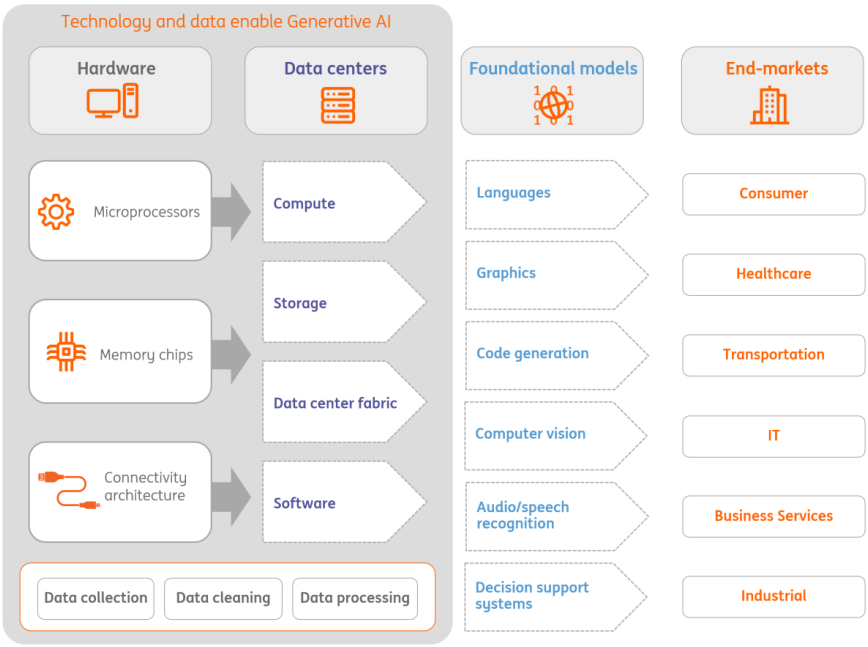
The latest AI models require spectacularly large datasets. In this article, we discuss leading data sources as well as the challenges that come with it. We expect that the owners of private data will build proprietary solutions, opening a new source of revenue. Nevertheless, more clarity is needed around copyright and the rights of content creator



The foundational models need a lot of data to be trained. Companies can use public data to train their models, such as the data that is available on the open internet. Moreover, the quest for the best foundational model may also result in a search for the best data. The AI developments are therefore an opportunity for companies that own, collect, and manage data. Furthermore, solutions that can prove the authenticity of content (or the person we interact with) will also

become much more important.

Data is an important input for the AI value chain



Source: ING

Use of data is not without challenges

A feature that has enabled the creation of ever larger models is the availability of relatively good quality data sets that are free to use to train AI models. Although there is quite a lot of training data available, the latest AI models are characterised by billions of training tokens, and therefore require (extremely) large data sets, which are not readily available. As noted in a research paper from Google, “we find that for compute-optimal training, the model size and the number of training tokens should be scaled equally: for every doubling of model size the number of training tokens should also be doubled”. The table below shows the increase in size of the training dataset for some models from OpenAI.

Newer models from OpenAI require larger training datasets

Model name	Size training dataset (words / datapoints)
GPT-2 (1.5B)	3.00E+09
GPT-3 175B (davinci)	3.74E+11
GPT-4	4.90E+12

Source: Epoch

Nevertheless, so far, there has been limited disclosure as to the exact data that has been used for

the training of some frequently used models. Also, there is litigation on the use of data that is copyright protected, such as newspaper articles and books. Furthermore, many potential business users of AI technology are reluctant to send their customer data to a public tool, because of potential confidentiality issues. These limitations create an opportunity for the companies that own private data to develop their own services. Or they could sell this data to the large technology platforms if it follows from regulations that the large technology platforms cannot use data sources that are copyright protected.

Datasets are being built to train AI models

A great effort has been made to build datasets, also for scientific purposes. There are quite a few sources that are widely used such as: Common Crawl, Colossal Clean Crawled Corpus (C4), The Pile, BookCorpus and Open Super-large Crawled Aggregated coRpus (OSCAR, based on Common Crawl). Datasets sometimes consist of multiple sources, such as books and crawled web pages or combinations of them. Other sources are, for example, Wikipedia, social media sites, (X, Reddit), software repositories (GitHub), and scientific journals.

Before data can be used, the data has to be cleaned to omit inappropriate words, discriminatory language as well as incomplete sentences. With the cleaning process comes the deletion of a substantial amount of valuable training data. Also, some sources are large but contain vast amounts of text that seems similar, such as some stories in the data included within BookCorpus. There is also a risk that training data has been generated by AI. This may create complications, which need a solution (such as a watermark so that the data can be removed more easily).

Content creators are dependent on copyright protection

The availability of training data is challenged by copyright protection. Many models have been trained with data that may be subject to rights protecting the creators of texts. The most famous example of a rights owner starting litigation is the New York Times. The newspaper has argued in court that its articles cannot be used to train AI. Other interest groups are the Rights Alliance, from Denmark. They have argued that Books3 infringed copyright laws. Moreover, in the USA, 10,000 authors, united in the Authors Guild have written an open letter, asking for financial compensation.

There is, therefore, still debate on whether technology companies can use published materials for training their models, as there is much debate about the interpretation of copyright protection and the fair use of materials. In this context, it is relevant to mention that many authors today mention (in electronic form) that their text cannot be used for the training of AI. Also, work is underway to create a better version of The Pile, dubbed The Pile v2, which aims to improve the data quality over the previous version. Also, the way the data will be cleaned may be improved as companies are more experienced today doing this in an effective way.

Proprietary data becomes ever more valuable

The training requirements make data a valuable input. Companies also collect data when their applications are used to be able to refine their models. This is something to consider when using applications because the input one gives can contain confidential (or private) information. There are therefore substantial risks involved when transmitting confidential (customer) data to applications which are managed outside a company. Because of privacy regulations and reputational risks, companies often take the utmost care to protect their data, which may require more tailored solutions in the form of private AI applications.

Also, proprietary datasets may represent significant value. Monetising owned data sources through AI tools can constitute one of the new business models enabled by AI. The first solutions that come to mind are in the scientific publishing space, or owners of databases with large repositories of programming code.

Although it is slightly out of the scope of this article, we would like to mention that there are still challenges with the data generated by AI. Is the data copyright protected without significant human involvement? How can the problem of hallucinations be reduced? And will tools solve, at some point, the problem of generating data that is false or contains discriminatory language? We do not even discuss the problems arising from deep fake images. And can we prevent AI-generated data from being used for further training? Including a (mandatory) watermark in AI-generated content would seem to be necessary to solve some of these challenges. Nevertheless, detecting AI-generated content is also a nice opportunity for new software solutions.

More clarity is needed on the ability to use data for training purposes

Over time, we expect to get more clarity on the balance between the rights of content creators and the desire by the technology companies to use data. This is also a task for regulators. We think that a guiding principle should be that content creators must have the choice of whether their data can be used for the training of generative AI models. Nevertheless, the quality of data sets will likely improve over time, while new business models around the management of data will arise.

Author

Jan Frederik Slijkerman
Sector Strategist, TMT
jan.frederik.slijkerman@ing.com

discussed herein should contact ING Financial Markets LLC, which is a member of the NYSE, FINRA and SIPC and part of ING, and which has accepted responsibility for the distribution of this report in the United States under applicable requirements.

Additional information is available on request. For more information about ING Group, please visit http://www.ing.com.