Data in the AI Age: From abundance to sudden scarcity

ABRAHAM THOMAS

is venture partner at Merak Ventures.

We are living through a data explosion. When Steve Jobs released the Apple II in 1977, a terabyte of memory occupied a large warehouse and cost over $100 million. Today, you can buy a 1TB hard drive smaller than your palm for under $100. That's a million-fold reduction in price and size: the effect of Moore's Law, compounding over 50 years.

When storage is cheap, more data gets stored. The most successful businesses in the world are built around this simple fact. Facebook, YouTube and Instagram rely on cheap plentiful storage to host the user-generated content that attracts eyeballs and advertisers. Amazon, Uber and Zepto collect data on customer preferences to relentlessly optimize their products. Nvidia, Apple and Microsoft build the hardware and software infrastructure that powers these loops. Data rules the world.

Then came artificial intelligence (AI), and the world changed. Software and data are two sides of the same coin. Software is useless without data to apply itself to, and data is worthless without software to interpret it. Economists call these "perfectly complementary inputs": the abundance of one increases demand for the other. The data explosion over the past two decades has turbo-charged the usage and value of software. Generative AI flips this equation. Language models like GPT, Llama and Bard are voracious consumers of data; they also make software dramatically easier to create. As a result, we suddenly find ourselves in a world where computation is abundant and data is the (relatively) scarce resource.

This benefits companies that already own data. But what's valuable to AI is subtly different from what was valuable in the past. There are at least five different flavours of data value that AI unlocks.

In a world of competing models trained on the same few data-sets, unique data is more precious than ever. Also valuable is latent data: information that hasn't been used effectively in the past, but can be now, thanks to AI. Domain-specific custom data is critical to unlocking tangible business value from AI. Data of exceptional accuracy, known as golden data, is always in high demand. And finally, some data-sets hold value not in themselves, but in how they unlock the value of other data assets; such catalyst data can be enormously lucrative. The very best data assets are the new AI gold mines. But there are also big opportunities for picks and shovels to mine this gold: tools to build new data assets, connect these assets to AI infrastructure, extract latent data using AI, and train models on unique, small or custom data. There's also an entire commercial framework waiting to be built around these tools: new pricing and usage models, data valuation and validation, data compliance and rights, data marketplaces and commons.

The story doesn't end there. Generative models don't just consume data, they produce it. As a result, we're entering a world of unlimited content. With the proliferation of spam bots, image generators, hallucinations, deepfakes and content factories, it's hard to tell counterfeit apart from real. Who created this content; can you prove they created it; are they who they say they are; is what they created 'good'; and does it match what I need or want? We'll see the emergence of tools to answer each of these questions, building a 'confidence chain' comprising signatures, provenance, identity, quality and curation.

Many of these will be mediated by software agents. In place of a human-in-the-loop approach to improve software processes, we'll increasingly see software-in-the-loop to streamline human processes. This manifests as AI co-pilots and research assistants, AI tutors and curators, and a plethora of AI productivity apps.

QUICK READ

As large language AI models are voracious consumers of data and also ease software creation, we now find that computation is abundant and data is the relatively scarce resource.

Artificial scarcity is also a possibility as players might seek to capture the gains, although the bigger question is what happens to human work. We must explore AI to find out.

The economist William Jevons observed a paradox in the coal industry. Even though individual coal plants got more efficient over time, the quantity of coal used by industry went up. Efficiency lowered the price of coal energy, leading to more demand for it.

Something similar is happening with data and AI. It's not just that data and software reinforce each other in a productivity flywheel. It's not just that generative models produce and consume both data and code. It's that the price of 'smart computation' has fallen and the consequence is that there will be a lot more smart computation in the world.

So, where are the new areas of scarcity? The hardware that powers computation and data is one candidate: as the latter grows exponentially, the former cannot keep up. Persistent chip shortages are a symptom of this. Energy is another candidate, for similar reasons. More generally, as technologies built on 'bits' become ever more powerful and ubiquitous, technology built on 'atoms' becomes ever more valuable.

Artificial scarcity is another possibility. Society may benefit from abundance, but individuals and corporations may seek to selfishly constrain or capture the gains from ubiquitous, cheap, powerful data and computation.

Finally and most provocatively, what happens to human beings in this brave new world? Are we a scarce and valuable resource, and if so why? For that nebulous entity we call 'creativity' or for our ability to accomplish physical tasks? Will AI augment human capacity or automate it away? I believe in abundance and I'm optimistic; the only way to find out is to go exploring!