Too Much of Data

Notes based on 'Data Deluge' article by Prof Gautam Desiraju

4 min read · 12 hours ago

Yogesh Haribhau Kulkarni (PhD)

Listen Share ••• More

Painting by J. M. W. Turner (Source)

Data has become the most powerful resource of the modern age. It has been called the new oil, but in some ways, it is even more valuable. Oil is finite. Data is constantly created, copied, refined, and reused. Modern artificial intelligence systems live and breathe data. Without massive, high-quality datasets, today's AI simply would not exist.

In the technology world, a popular saying goes: whoever controls data and chips controls AI. This is not an exaggeration. Large language models, recommendation engines, protein structure predictors, and autonomous systems all depend on huge volumes of structured and unstructured data. The explosive growth of AI over the last decade was made possible by what can only be described as a global data deluge.

But abundance has a dark side

Researchers Gautam R Desiraju and Deekhit Bhattacharya raise a troubling possibility: too much data may be damping our capacity for original thinking. This sounds counterintuitive at first. After all, more data should mean better insights. But science is not only about accumulation. It is about synthesis, intuition, and creative leaps.

Today’s scientific ecosystem is highly fragmented. Specialization has created thousands of micro-disciplines, each with its own datasets, jargon, tools, and publishing norms. These datasets often cannot “talk” to each other. Cross-disciplinary analysis is rare because it is slow, hard to fund, and difficult to publish.

The publishing system itself has become part of the problem. Peer review was designed for a slower, smaller scientific world. It now operates under extreme pressure. Editors struggle to find qualified reviewers. Reviewers are unpaid, overworked, and often anonymous. There is no universal system to score or rank referees based on quality or integrity.

The results are visible. In 2023, more than 10,000 scientific papers were retracted worldwide, many due to fabricated, manipulated, or unreliable data. Journals are overwhelmed, yet most still rely on manual processes that were designed decades ago.

AI could be the solution. Or the accelerant.

Machine learning models and large language models can scan millions of papers, detect patterns, and summarize complex topics in seconds. Tools similar to AlphaFold have already solved problems that were once thought to be nearly impossible. Used well, AI can increase scientific productivity and lower the cognitive burden on researchers.

But AI is not neutral. It amplifies what it sees. If the underlying data is biased, flawed, or incomplete, AI scales those flaws. A single incorrect machine-generated summary can be cited by hundreds of future papers. Over time, errors become embedded in the scientific canon.

Bias is another serious risk. Bias can enter through dataset selection, algorithmic assumptions, institutional priorities, and even geopolitical power structures. These biases are often invisible at the user level, yet they shape outcomes in powerful ways.

The structure of scientific papers itself may change. Instead of long narrative articles, future "papers" may be structured datasets and code, exposed through APIs. Readers may not read a paper in the same way. An AI system could restate the same research in different styles, levels of complexity, and even emotional tones, depending on the reader.

This sounds efficient, but it raises deep questions. Who should be credited? Who is accountable for errors? If authors become anonymous data contributors, what happens to motivation, funding, and career progression?

Ownership of AI-generated data adds another layer of complexity. AI systems and their outputs are not owned by humanity as a whole. They are controlled by a small number of institutions and corporations. Legal disputes around data usage, such as (this article), show how fragile and contested this space is.

So what can be done?

Some promising directions are already visible. Preprint culture allows early, open sharing of research. Decentralized and community-driven peer review could introduce transparency. Smart contracts and content escrows could make reviewer compensation fairer. Public reputation systems could rank referees and score papers for trustworthiness.

We may also see the rise of "replicability indexes", where claims are automatically scored based on how often they can be reproduced. Papers could gain or lose credibility over time, instead of being treated as permanently valid.

The uncomfortable truth is this: the traditional peer-reviewed journal is no longer enough. It is too slow, too fragile, and too disconnected from the reality of AI-driven science.

The future of science will not be human-only or AI-only. It will be a careful, continuously evolving partnership. The real challenge is not whether we should use AI in science. We already do. The real challenge is designing systems that can harness AI's speed and scale without sacrificing truth, trust, and creativity.

If we fail to redesign this system urgently, we may face a world where knowledge grows faster, but understanding becomes weaker.

Data Artificial Intelligence Ideas Creativity Publications On Medium

Following

Published in Technology Hits

4K followers · Last published 10 hours ago

We cover important, high-impact, informative, engaging stories on all aspects of technology. Subscribe to our content marketing strategy newsletter: https://drmehmetyildiz.substack.com/ Apply via: https://digitialmehmet.com/contact External: https://illumination-curated.com

Edit profile

Written by Yogesh Haribhau Kulkarni (PhD)

1.8K followers · 2.1K following

PhD in Geometric Modeling | Google Developer Expert (Machine Learning) | Top Writer 3x (Medium) | More at https://www.linkedin.com/in/yogeshkulkarni/