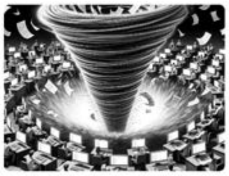
Data Science Is an Ecosystem

This Pub is a **Commentary** on



Data Science at the Singularity

by David Donoho

Published on Jan 29, 2024 · hdsr.mitpress.mit.edu

Data science institutes have sprouted at many universities in the last decade, fertilized by the massive data sets generated by the exponential rise of computing power and computer memory. At UC San Diego, the new Halıcıoğlu Data Science Institute[1] went from zero to 63 faculty in 6 years. It was recently promoted from institute to school status alongside the School of Medicine, the School of Engineering, the School of Biological Sciences, and the School of Physical Sciences. The rapid rise of data science has indeed been singular. David Donoho's article on "Data Science at the Singularity" insightfully analyzes this recent explosion. My goal here is to put data science into a broader context.

Data Science Is Evolving

What Donoho (2024) describes in data science is what in nature would be considered an ecosystem, a complex web of interacting species and resources. Ecosystems thrive through mutually beneficial relationships that include a balance of cooperation and competition, not unlike the ecosystem of companies, universities unlike the ecosystem of companies, universities, and the computational resources that are available to today's AI community.[2] Donoho has identified several features of this ecosystem that have made it thrive, such as the frictionless reproducibility emerging from the free exchange of programs, data, and benchmarks that vastly accelerate progress. Data science research can expand synergistically on a solid mathematical foundation.

Biological evolution advances in spurts when new capabilities emerge that give rise to new radiations of species. For example, the Cambrian explosion more than 500 million years ago created most of today's animal phyla, including the earliest chordates, our distant ancestors. This 'explosion' lasted around 25 million years, a blink in the billions of years of life on our planet. Singularities are not the ends of evolution but steps along the way. Just as evolution has gone from explosion to explosion —punctuated equilibria—science has progressed with sequences of breakthroughs over hundreds of years. After each paradigm shift, there was a long period of consolidation and incremental improvements (Kuhn, 1962). Typically, after a major innovation in nature, great diversity is whittled down over time by the survival of the fittest to fewer and stronger branches. The resulting stable equilibrium prevents further major advances until an environmental crisis occurs. Is this how data science will evolve?

Machine Learning and Modern AI

Brain architectures inspired the first Neural Information Processing Systems (NeurIPS) meeting in 1987, focusing on learning from data and attended by researchers in diverse disciplines, including neuroscience, cognitive science, computer science, statistics, information theory, control theory, and electrical engineering (Breiman, 2001).[3] Developing a common mathematical language that united these disciplines took a generation. A machine learning ecosystem evolved. Over several decades, advances were made, generating new tools and insights into data. These advances included support vector machines, graphical models, variational methods, and independent component analysis. The remarkable capabilities of recent large language models (LLMs) were an overnight success that took 40 years of slow but steady progress by researchers from many scientific, engineering, and mathematical disciplines. Hybrid vigor is another concept from biology that may underlie the astonishing recent advances in machine learning.

Learning was not a central part of traditional AI. The goal of AI in the 20th century was to program intelligence, and different programs were needed for sensory, motor, and planning modules. For example, the goal of the vision module was to create an internal model of the external world (Churchland et al., 1994). Writing a vision program proved to be more difficult than anyone imagined. But is vision an end in itself? Vision facilitates motor interactions with the world. Recent research has shown more motor feedback to the visual cortex than feedforward visual signals (Musall et al., 2019). These pre-movement motor signals predict self-generated visual signals, freeing up limited feedforward bandwidth for unpredicted visual input (Li et al., 2023). There is no good reason why nature should have confused us with all these details. Nonetheless, the early groundwork laid by AI set the stage, raising expectations for what later became possible only with much more computer power and data in the 21st century.

The Future of AI

Artificial general intelligence (AGI) is a holy grail for AI. The degree to which generative LLMs exhibit such general intelligence is much debated. Disagreements about whether LLMs 'understand' what they generate remind me of debates about 'life' a hundred years ago: What is the difference between living matter and inanimate matter? Vitalists believed that life was a nonphysical 'vital force' infused in us but not in rocks. This debate was not a fruitful way to make progress. The discovery of DNA clarified many issues and led to a productive singularity that revolutionized biology. Today's debates are the equivalent of debates about life, with AGI akin to the 'vital force.' The direction machine learning is taking today may lead to a new conceptual framework that is as fundamental to AI as the architecture of DNA is to biology. Now is an excellent time to reassess old AI concepts in the light of new evidence from machine learning.

My perspective aligns with Donoho's view that what is emerging is not an "AI singularity." Being objective during an explosion is difficult, but Donoho's manifesto rings true. Geometry, calculus, and probability are foundational areas of science and engineering. We are witnessing the birth of new mathematics based on geometry, calculus, and probability in high-dimensional spaces (Sejnowski, 2020). LLMs are the equivalent of the cathedrals built in the Middle Ages by trial and error (Figure 1). As LLMs inspire new mathematics, a new conceptual framework will emerge; their progeny will be skyscrapers that will reify intelligence.

Disclosure Statement

Terrence J. Sejnowski has no financial or non-financial disclosures to share for this article.

Figure 1. Built during medieval France, construction of the Notre Dame cathedral began in 1163 and was largely completed by 1260, though it was modified in succeeding centuries (Wikipedia, Notre-Dame de Paris, CC0 1.0 Universal Public Domain)

References

Breiman, L. (2001) Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Donoho, D. (2024). Data science at the singularity. *Harvard Data Science Review*, *6*(1). https://doi.org/10.1162/99608f92.b91339ef

Churchland, P. S., Ramachandran, V. S., &

Sejnowski, T. J. (1994). A critique of pure vision. In C. Koch & J. Davis (Eds.), *Large-scale neuronal theories of the brain* (pp. 23–60). MIT Press.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Li, J. S., Sarma, A. A., Sejnowski, T. J., & Doyle, J. C. (2023). Internal feedback in the cortical perception-action loop enables fast and accurate behavior. *Proceedings of the National Academy of Sciences*, *120*(39), Article e2300445120. https://doi.org/10.1073/pnas.2300445120

Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S., & Churchland, A. K. (2019). Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience*, *22*(10), 1677–1686. https://doi.org/10.1038/s41593-019-0502-4

Sejnowski, T. J. (2020). The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences*, *117*(48), 30033–30038. https://doi.org/10.1073/pnas.1907373117

Connections

1 of 15 children and siblings

FOOTNOTES • 3

COMMENTS • 0

? Login to discuss

No comments here

Why not start the discussion?

READ NEXT • On Emergent Limits to Knowledge—Or, How to Trust the Robot Researchers: A Pocket Guide