Comment

https://doi.org/10.1038/s41928-023-00985-1

Solving the big computing problems in the twenty-first century

In the format provided by the authors and unedited

1. List of data points in Fig. 1

Model Name	Year	Compute (FLOP)	FLOP/W of GPU	Power (W)	Energy (TWh)
Dropout	2012.5	1.81E+17	1.56E+10	1.16E+07	3.22E-09
Visualizing and Understanding Conv Nets	2013.8	5.36E+17	1.83E+10	2.93E+07	8.15E-09
DQN	2013.9	2.33E+15	1.83E+10	1.28E+05	3.55E-11
seq2seq	2014.75	8.04E+18	1.83E+10	4.40E+08	1.22E-07
VGG	2014.7	1.04E+19	1.83E+10	5.68E+08	1.58E-07
DeepSpeech2	2015.85	2.59E+19	1.83E+10	1.42E+09	3.94E-07
Xception	2016.7	4.32E+20	1.83E+10	2.37E+10	6.57E-06
Neural Architecture Search	2016.8	1.90E+21	1.83E+10	1.04E+11	2.89E-05
Neural Machine Translation	2016.71	6.83E+21	1.83E+10	3.74E+11	1.04E-04
Alex Net	2012.3	5.01E+17	1.56E+10	3.20E+07	8.90E-09
AlphaGoZero	2017.8	1.57E+23	4.97E+10	3.17E+12	8.79E-04
AlphaZero	2017.9	3.72E+22	4.97E+10	7.48E+11	2.08E-04
GPT 3	2020.5	3.14E+23	4.97E+10	6.33E+12	1.76E-03
Gopher 280B	2021.95	6.31E+23	6.15E+11	1.03E+12	2.85E-04
GPT 2	2019.2	2.49E+21	4.97E+10	5.01E+10	1.39E-05
Megatron Turing 530B	2020.9	1.35E+24	4.97E+10	2.72E+13	7.55E-03
Megatron NLG	2019.7	9.10E+21	4.97E+10	1.83E+11	5.09E-05
Pathways Language Model	2022.25	2.50E+24	6.15E+11	4.07E+12	1.13E-03

Supplementary Table 1: Model and GPU dataset found in Supplementary Ref. 1.

2. Does AI represent all of computing?

AI represents the currently dominant computing market. Past and future computing approaches will certainly not embody the flavor of AI we presently witness. However, trends in AI (such as the economics-driven growth over the past decade) are indicative of the appetite for computing in the economic world. The trends we note here (e.g., a doubling time in performance of 1.3 to 3 years) held true for the early days of computing as well (e.g., Moore's-law-driven hardware advances and more recent architecture-driven performance advances). Quantum, reversible, analog, optical computing, etc., will become dominant in the years to come, and the types of problems addressed by them will also evolve, but all such computing approaches strive to maintain the high demands of the performance improvement trends that have created a robust market demand.

3. Computing trends indicating economics-limited infrastructure development

The initial 6 years of AI (post the 2012 AlexNet resurgence) saw an exponential growth that was limited only by improvements to model quality and their convergence accuracy.[2] In other words, the infrastructure required to support models under research already existed for the initial duration of AI research, and the costs of setting up of such infrastructure was not a significant factor in pursuing more complex models. Since around 2018, this trend has witnessed fluctuations and saturation, which were limited by economics. As an example, the most expensive published model (GPT-3) required over US$ 10 million to train once to perform one task (natural language processing). It is estimated that GPT-4's cost was over US$ 100 million. This cost is now a major factor in funding of research into newer models, which was not much of a concern roughly 5 years ago. We certainly do not envision any hard constraints posed by costs, and the costs of future single-task computing infrastructure (for AI or other computing approaches) may very well far exceed the present limits we see today. However, we suggest that such costs will be driven by economic needs and critical commercial (or defense) applications, and not merely due to academic or research interests. It is worth considering two trends – AI (along with training) will be increasingly moved to personalized devices as energy efficiency allows, and we are already witnessing fluctuations and saturation in AI model sizes. These trends mean that there will be empirical limits on the average expenditure on a single AI model, which we predict to be US$ 10 billion in today's value (based on the fact that today's largest corporations are valued at US$ 3 trillion). Of course, there will be outliers to these trends in the form of unusually large models being trained for critical commercial or societal needs, which will likely be achieved via conglomerations of massive corporations or international governmental collaborations.

4. The calculations behind the three exemplary problems

For each of the three problems we consider, we provide details on the assumptions and sources used to identify the computing energy per year for today, presented in Fig. 1. While the simpler problems (such as weather modeling) are well defined, the larger problems (such as human evolutionary simulation) have not been defined in great detail, and in many cases, the parameters going into such problems are unknown. As such, the estimates for large-scale problems beyond our present reach are bound to have large error bars (by several orders of magnitude, as described in our examples below). Nonetheless, these calculations, along with our projections, give us an idea of the human generation that will witness solutions to these problems.

4.1. Planetary weather modeling

There are firm calculations of the compute operations required to model Earth's climate with prediction durations of up to two weeks. DeBenedictis[3] and Malone et al.[4] have established the factors by which computing operations must increase on top of present-day climate modeling computers. Every additional feature on top of present-day capabilities (e.g., increased spatial resolution, inclusion of stratospheric climate, inclusion of biogeochemistry, accounting for complex reactions, corrections for drift over time, etc.) will require a computing operations to increase by a specific factor. As a result, performing reasonably accurate predictions in planetary climate, extending to two weeks with finer time steps (~minutes), would incur a compute factor of 10^{10}-10^{12} over present-day computers. We use the Japanese Earth Simulator Project as a baseline for today's climate prediction computing,[5] which performs $>10^{13}$ FLOP/s. Thus, an extrapolation with the aforementioned compute factors results in a range of $>10^{23}$ FLOP/s. We use the minimum of this range as the compute required for planetary weather modeling. We then calculate the total operations for running such a system for one year continuously (to obtain a yearly compute and energy cost), and multiply the resulting quantity with the energy efficiency of a widely available advanced GPU (NVIDIA A100), which is about 6.24×10^{11} FLOP/J. This process results in the total energy required for one year of continuous planetary weather forecasting, which is the starting point for our predictions in Fig. 1.

4.2. Brain-scale modeling

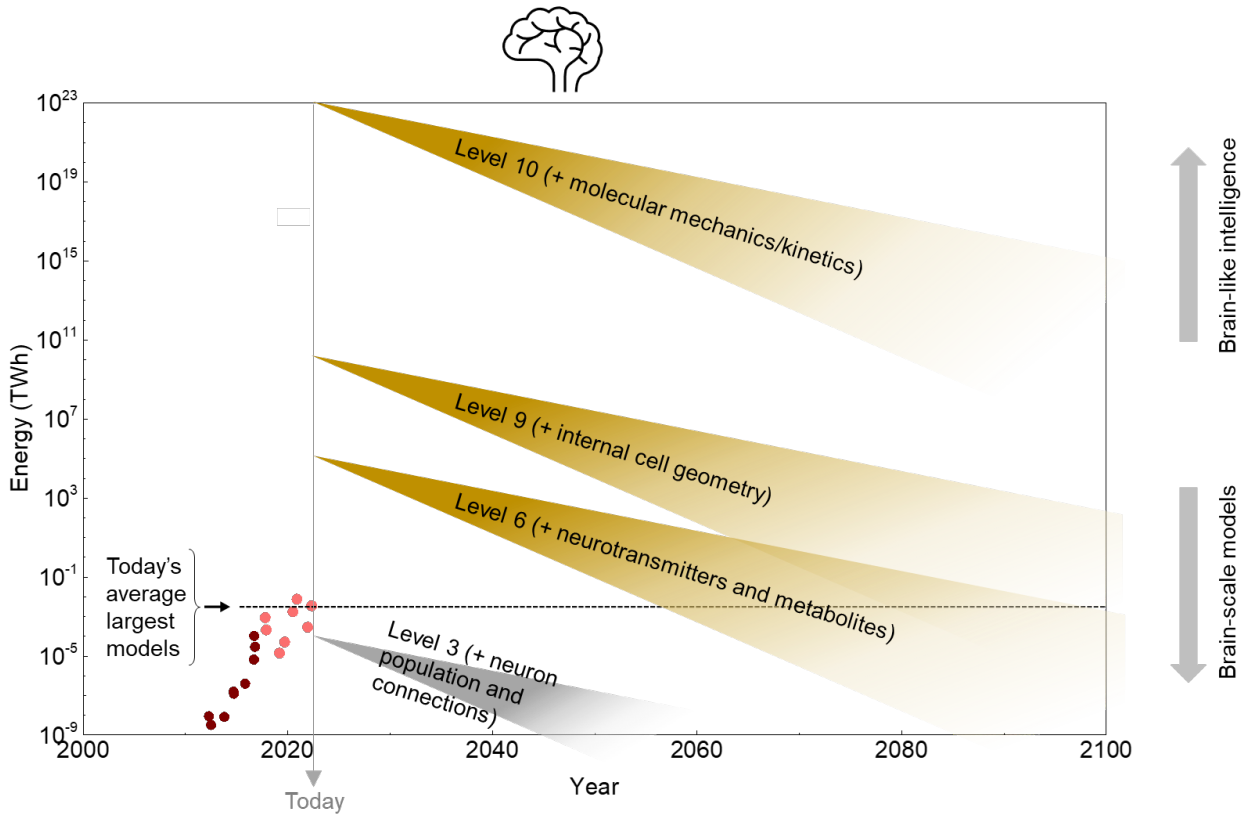
Brain-like intelligence is likely the prime motivator for the world of AI, which, after all, originated from early efforts in the 1940s in modeling neurons using electrical circuits.[6,7,8] The level of computing required to achieve brain-like intelligence has been debated fiercely.

To date, the largest efforts in brain-scale modeling have used spiking neural network frameworks with additional complexity to capture synapse dynamics with four different molecular structures. A model simulating 68 billion neurons and 5.6 trillion synapses for 1 minute was run on Japan's

K supercomputer (10.51 PFLOP/s, 12.6 MW) over 10 hours. AI models that aim towards developing a general intelligence utilize simple models of synapses and neurons with lower efficiency – however the most computationally demanding part is training the network. It is an open question if generalized human-level intelligence will emerge from brain-scale AI models.[9] Thus, there are no firm claims on the compute required to emulate brain-like intelligence, let alone consensus. Thus, it is unclear if brain-scale modeling will lead to full brain-like intelligence. Therefore, we refrain from using brain-like intelligence as a goal, but we instead aim at brain-scale modeling, and identify a compute quantity that uses rigorous estimates.

Sandberg and Bostrom[11,12] have laid out 11 levels of details in various brain-scale models. The most basic ones account for only the number of neurons, while the more complex ones account for neurotransmitters, molecular structures, etc. The most complex ones account for molecular and quantum kinetics as well. It is not known if such quantum processes are required in a model to capture the phenomenological behaviors of a brain, the answer to which may very well depend on the behaviors we seek to simulate.

Here we used the estimate of 10^{25} FLOP/s of compute (Level 6 as laid out by Table 9 of Supplementary Ref. 12) for per second brain emulation which accounts for most known physiological parameters, with the exception of cellular structures and quantum/molecular kinetics.[11,12] From this estimate, we follow a process similar to planetary weather modeling to arrive at a yearly energy estimate.



Supplementary Fig. 1. A plot of energy against years, similar to Fig. 1 (main text), with wedge-shaped projections corresponding to various levels of brain-scale modeling, as laid out by Sandberg and Bostrom[11,12]*. We have already achieved Level 3 models. In Fig. 1 (main text)*

4.3. Human evolutionary simulation

Human evolutionary simulation aims to capture the behaviors of individual humans, interactions among them, environmental interactions and the resulting evolutionary process. In short, this problem is a primer for recreating the entire ecosystem in a simulation, and is often discussed in the context of whether we are living in a simulation. Notwithstanding the speculative context in which this problem is discussed, we consider its most basic version. The first ingredient required is to create phenomenological brain-scale simulations. Second, it is important to consider the nature of the interactions among individuals, their environmental conditions, hazards, etc. Prior studies have predicted the computational requirements of performing such simulations, of course, with varying projections. [11-14] From these studies, we choose the minimum projected computational requirements being employed to simulate evolution of 50,000 individuals for 50 years, which resulted in 10^{28} FLOP/s. Thus, our estimate is the bare minimum compute needed to even address a problem of this type, likely at a very small scale. However, achieving the level of compute we have estimated for the simplest form of this problem is a necessary first step in interplanetary multi-generational missions and colonies. Similar to the two other problems above, we converted FLOP/s to an yearly energy estimate in Fig. 1.

5. A performance doubling time of 1.3 to 3 years

We estimated an optimistic performance doubling time of 1.3 years, following the Huang's law, which accounts for both hardware and non-hardware advances. There are many challenges to maintaining this trend, including the obvious physical limits in shrinking devices. Other challenges include having to invent new forms of computing, novel algorithms, architectures, software stack, etc. As noted in the main text, significant algorithmic advances are not gradual, and are difficult to predict. Further, many algorithmic innovations may not be relevant to artificial intelligence. Thus, maintaining this trend is our optimistic projection, while this is by no means a limit to any new ideas that might break new ground (e.g., quantum computing for specific problems such as cryptography).

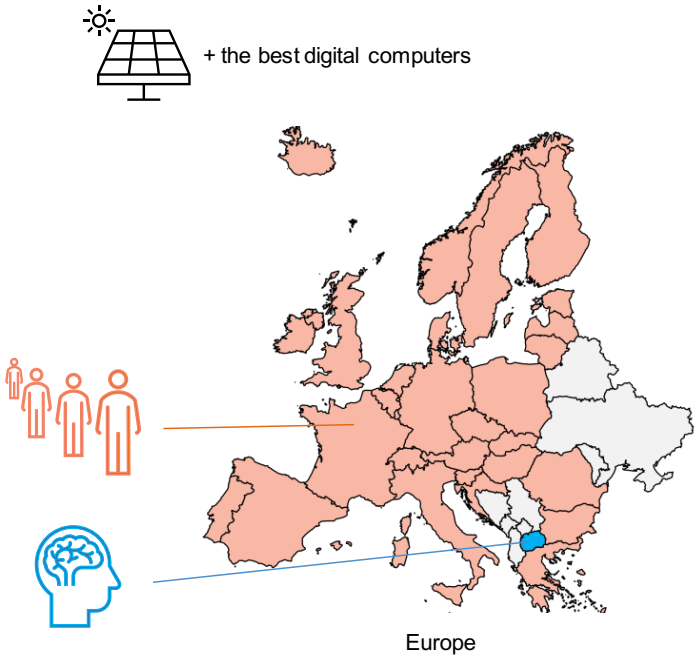
As it has been generally identified, overhauls in computing require a co-design of the entire computing stack from materials to software. However, once the materials hit their physical limits (e.g., at a node size of 1 nm, with present node sizes at 5 nm), we will be more constrained in our design efforts. In our pessimistic projections, we assume no non-hardware advances. Thus, we estimate a pessimistic performance doubling time of 3 years, which is an adverse approximation of the slowest that Koomey's law[15] has been (~2.6 years doubling time in the early 2010s). We increased the doubling time from 2.6 to 3 years to account for the adverse effects of cluster-scale computing compared to chip-scale computing (e.g., due to bottlenecks in inter-chip data movement in server racks).

6. The limits of digital computers

The Landauer limit is generally accepted as a fundamental limit to the minimum energy required to process (flip) a bit of information, which is in the order of kT (k: Boltzmann constant; T: temperature).[16] For each of the three problems, we calculate the energy required to solve them when the supporting infrastructure is operated at the Landauer's limit. We then report this energy as a band in the wedge-shaped projections. Every FLOP takes up to 10^6 bit flips, assuming high-precision representation, which can be reduced by modulating the speed required and the complexity of the operation involved.[17] Here we use this high estimate for bit flips per FLOP because most analyses on the various problems we considered here assume double precision representation of quantities. Further, operating at the Landauer's limit means that we consume kT amount of energy to manipulate a bit, but we certainly will be spending a significant part of the energy in overcoming heating and parasitic losses. To account for such losses, we included a factor of 15 in our calculations.

7. The calculations behind solar panel coverage to solve the two larger problems

We considered 400 kWh/m^2/year as the energy efficiency of solar panels, which is typical for European countries. This quantity could range between 200 and 2500 kWh/m^2/year depending on the region of the world and other engineering factors. These quantities can be found in several publicly available databases.[18] We converted the total energy required to solve the two hardest problems in our list (i.e., human evolutionary simulation and brain-scale modeling) to area of solar panel coverage by dividing the energy required (over a year) to solve the problems by solar panel efficiency. Particularly, the solar panel coverage illustrated in Supplementary Fig. 2 was made using the energy quantities at the digital limit (the interface between digital and post-digital approaches).



Supplementary Fig. 2: *Illustration of the solar panel coverage required to solve the two larger problems in Fig. 1 by operating at the digital limit. Even with the best digital computers, we need to cover most of Europe with solar panels to address the simplest forms of human evolutionary simulation.*

8. The era of memristor-based AI

Over the last five years, research into two-terminal memories, which are generally known as memristors, has progressed to manufacturable chip-scale demonstrations of AI functions. Such memristor AI chips promise very significant (by up to 5 orders of magnitude) performance improvements (in terms of energy and speed) relative to GPUs of similar scales.[19] The performance improvements are mainly due to the massive parallelism in matrix multiplication operations enabled by memristor crossbars. This thrust in research has been accompanied by several small start-up companies and large corporations pursuing mass-manufactured memristor-based AI chips.[20-24] Thus, memristor-based AI that can outperform GPUs is already a reality, though at smaller production volumes. The trend is clear – we anticipate production volumes and commercial adoption to ramp up over the next five years.[19] For a few decades, such memristor-based AI chips will likely be the backbone of many types of AI, which heavily leverage matrix multiplications as their workhorse. The end of the era of memristor-based AI, which is a classical computing technique, is likely going to be motivated by the limitations in 3D densities of memristor chips (i.e., limitations in size and density scaling), and will likely be succeeded by non-classical approaches such as quantum computing.

Supplementary Information References

1. J. Sevilla, P. Villalobos, J.F. Cerón, et al., "Parameter, Compute and Data Trends in Machine Learning" URL: https://docs.google.com/spreadsheets/d/1AAIebjNsnJj_uKALHbXNfn3_YsT6sHXtCU0q7OIPuc4/edit#gid=1503579905 (2022). Accessed Jul 8, 2022.
2. J. Sevilla, L. Heim, A. Ho, et al., "*Compute trends across three eras of machine learning*," arXiv preprint, arXiv:2202.05924 (2022).
3. E. P. DeBenedictis, "*Reversible logic for supercomputing*," in Proceedings of the 2nd Conference on Computing Frontiers, 391 (2005).
4. Malone, Robert C., et al. *"High-end computing in climate modeling."* contribution to SCaLeS report (2004).
5. URL: https://en.wikipedia.org/wiki/Earth_Simulator. Accessed Jul 8, 2022.
6. W.S. McCulloch, W. Pitts, *"A logical calculus of the ideas immanent in nervous activity,"* Bull. Math. Biophys. 5, 115–133 (1943).
7. W. Pitts, *"Some observations on the simple neuron circuit,"* Bull. Math. Biophys. 4, 121–129 (1942).
8. A.L. Hodgkin, A.F. Huxley, *"A quantitative description of membrane current and its application to conduction and excitation in nerve,"* J. Physiol. 117, 500–544 (1952).
9. H. Yamaura, I. Jun, and T. Yamazaki, *"Simulation of a human-scale cerebellar network model on the K computer,"* Frontiers in Neuroinformatics 14, 16 (2020).
10. J. Tuszynski, *"The dynamics of c-termini of microtubules in dendrites: A possible clue for the role of neural cytoskeleton in the functioning of the brain,"* Journal of Geoethical Nanotechnology 1 (2006).
11. A. Sandberg, *"Feasibility of whole brain emulation,"* Philosophy and theory of artificial intelligence. 251-264 (2013).
12. A. Sandberg and N. Bostrom, *"Whole Brain Emulation-A Roadmap. The Future of Humanity Institute Technical Report 3,"* Oxford University, URL: http://www.fhi.ox.ac.uk/wp-content/uploads/brain-emulation-roadmap-report1.pdf 7, 251-264. (2008).
13. N. Bostrom, *"Are we living in a computer simulation?"* The Philosophical Quarterly 53, 243 (2003).
14. R. Yampolskiy, "*Why We Do Not Evolve Software? Analysis of Evolutionary Algorithms*," Evolutionary Bioiformatics, 14 (2018)
15. URL: https://en.wikipedia.org/wiki/Koomey%27s_law. Accessed Jul 8, 2022.
16. S. Kumar, "*Fundamental limits to Moore's law*," arXiv preprint, arXiv:1511.05956 (2015).
17. URL: https://www.lesswrong.com/posts/N7KYWJPmyzB6bJSYT/the-next-ai-winter-will-be-due-to-energy-costs-1. Accessed Jul 8, 2022.
18. URL: https://photovoltaic-software.com/principle-ressources/solar-radiation-databases. Accessed Jul 8, 2022.

19. J. D. Kendall and S. Kumar, "*The building blocks of a brain-inspired computer*," Applied Physics Reviews 7, 011305 (2020).
20. URL: https://www.eenewseurope.com/en/analog-in-memory-ai-processor-startup-uses-memristors/
21. URL: https://www.eetimes.com/startup-beats-hp-hynix-to-memristor-learning/
22. URL: https://www.reuters.com/article/chips-rain-neuromorphics-funding-idCNL1N2UC1YA
23. URL: https://www.eetimes.com/reram-research-improves-independent-ai-learning/
24. URL: https://www.businesswire.com/news/home/20190219005358/en/AI-Innovators-Join-Forces-in-Consortium-for-Development-and-Commercialization-of-Best-in-Class-AI-Computing-Platform