



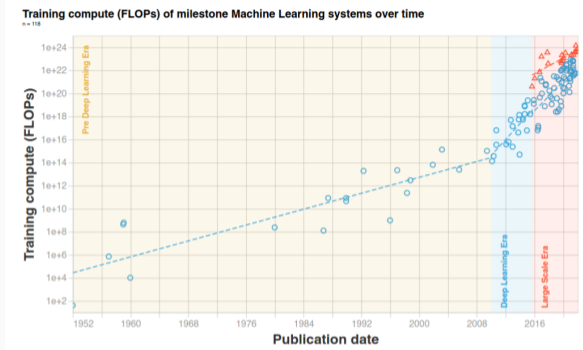
Evolution of the environmental damages of machine learning training over time

Clément Morand & Anne-Laure Ligozat & Aurélie Névéol

November 21, 2024

Université Paris-Saclay, Laboratoire Interdisciplinaire des Sciences du Numérique, CNRS

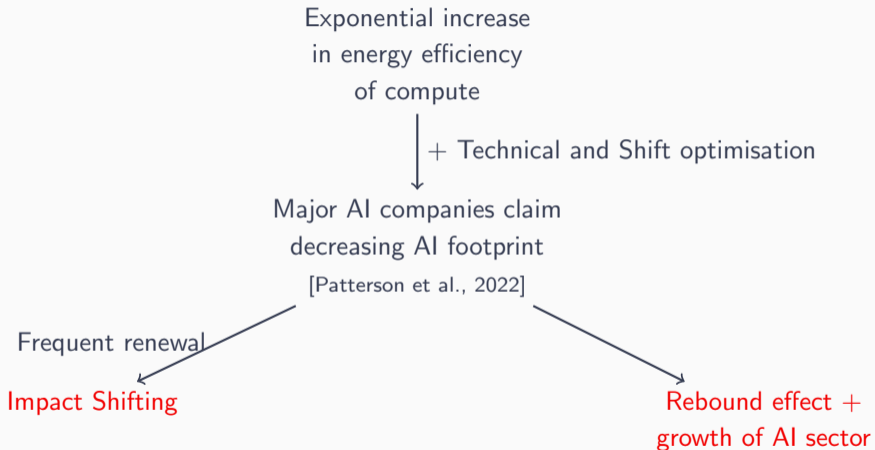
Machine Learning requires an ever-increasing amount of compute



[Sevilla et al., 2022]

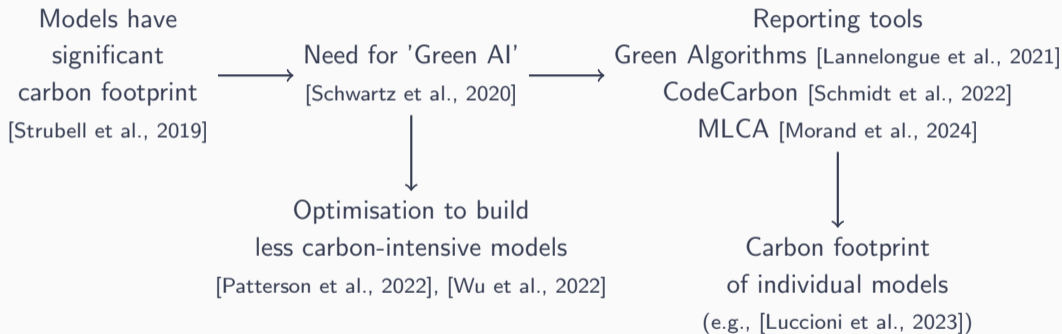
- 10 - 15 % of Google's energy consumption [Patterson et al., 2022]
- Important emissions from energy consumption : 552 tCO₂e to train GPT-3 once and 38 tCO₂e for BLOOM [Luccioni et al., 2023]

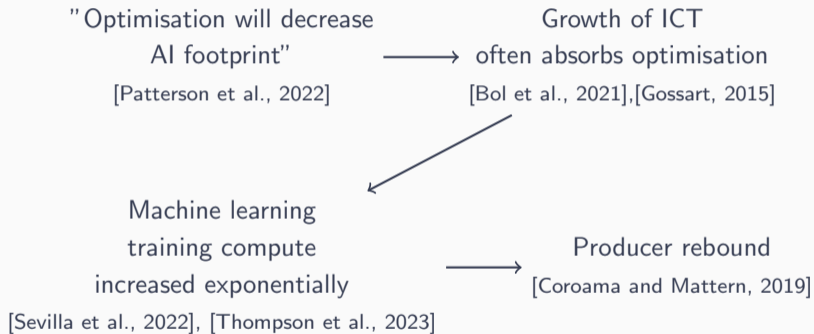
Numerous optimisations: how are the impacts of compute evolving?



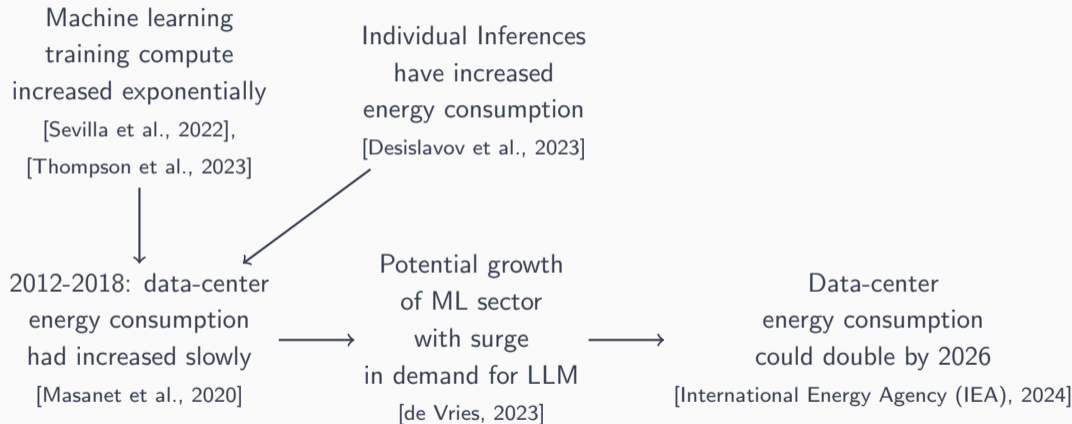
State of the Art

Green AI, optimisations and individual model assessment





Studies at scale of part of AI sector



Methodology

Gathering information on Graphics cards for Machine Learning

TECHPOWERUP



TechPowerUp
GPU database

Wikipedia list of
NVIDIA graphics cards

Other sources
(e.g., Google documentation)

NVIDIA Workstation graphics cards
between 2013 & 2023

specific cards (e.g., Google's TPU)

- 167 cards models
- 74 cross-validated (44%)
- NVIDIA datasheets when diverging

- *Thermal Design Power* (TDP)
- GPU die area and technological node
- memory type and size
- compute power
- release date

EpochAI Notable ML systems dataset [Epoch AI, 2022]

Models that have advanced the state of the art, had a large influence in the field's history, or had a large impact within the world.¹

Required information to estimate the environmental damages of model training:

- training duration
- training hardware
- electricity source

¹<https://epochai.org/data/notable-ai-models-documentation>

If training duration and number of cards are available

- 107 models (13% of entries)
- $GPU\ hours = training\ duration \times \#cards$
- most reliable estimate as it uses information directly from papers presenting models

If training duration and number of cards are available

- 107 models (13% of entries)
- $GPU\ hours = training\ duration \times \#cards$
- most reliable estimate as it uses information directly from papers presenting models

If Training hardware and number of FLOP during training are available

- 93 other models ($\sim 25\%$ of entries in total)
- $GPU\ hours = \frac{\#FLOPS}{peak\ performance}$
- linear regression to predict performance ratio when both estimates are available (87 observations)
- predicts $\sim 27\%$ constant performance ratio

values consistent with hyper-scaler datacenters

- 2 CPU per server plus:
 - NVIDIA workstation cards: 4 graphics cards
 - NVIDIA non-workstation cards: 2 graphics cards
 - non-NVIDIA cards: manufacturer documentation for the number of cards
- Lack of information → **Memory not accounted for**, source of under-estimation
- 3 year server duration based on graphics card lifespan [Ostrouchov et al., 2020]
- Information from META: near optimal 1.1 PUE, average utilization of 50% [Wu et al., 2022]
- Supposed 100% hardware usage during training.

Electricity source and modeling carbon intensity optimisation

Use the carbon intensity of the country of the ML system producer
If multiple countries are involved, all are considered to create a value interval

Electricity source and modeling carbon intensity optimisation

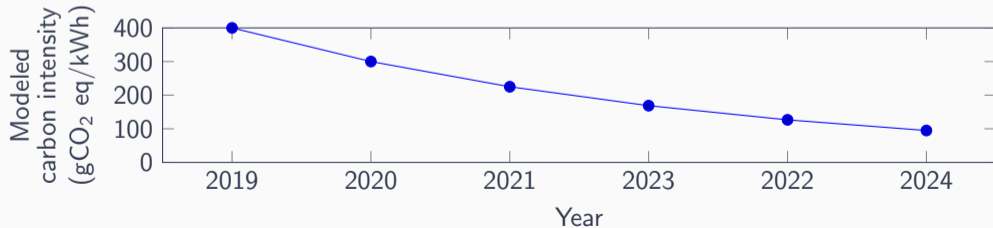
Use the carbon intensity of the country of the ML system producer

If multiple countries are involved, all are considered to create a value interval

Modeling strategies for reducing the environmental impact of energy usage

- Aims at accounting for compute location shifting and investment for de-carbonizing data-center electricity sources
- Continuous reduction of the carbon intensity of up to 25% per year starting in 2019.

Example (Modeled evolution of the carbon intensity of the USA electricity mix:)



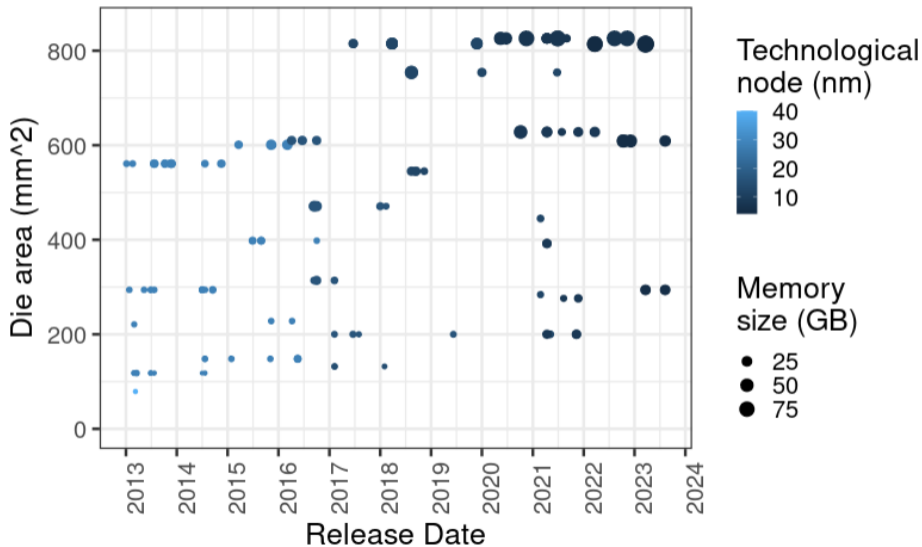
Bottom-up approach to evaluate hardware production and usage based on hardware characteristics and information about training process

Assesses:

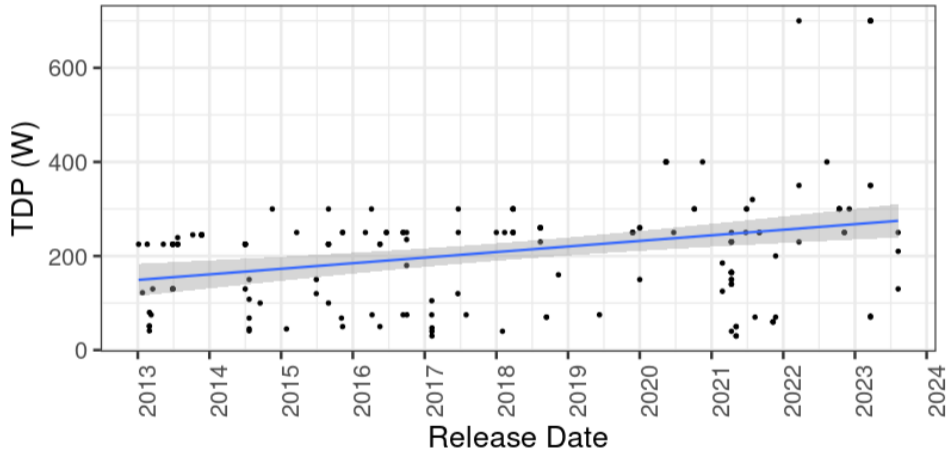
- Carbon footprint through *Global Warming Potential* (GWP100, expressed in kgCO₂ eq)
- Metallic resource depletion through *Abiotic Resource Depletion* (ADP, expressed in kgSb eq)

Results

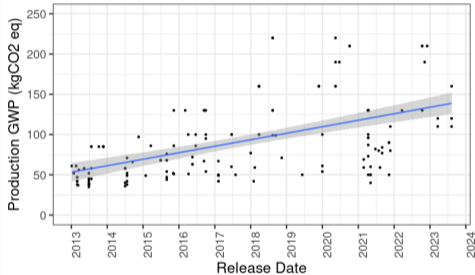
Evolution of the characteristics of NVIDIA workstation graphics cards



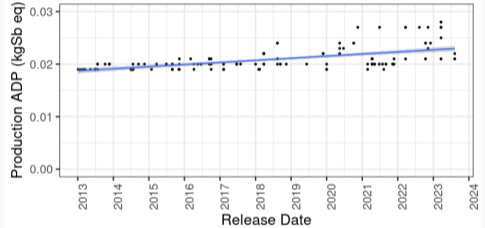
Energy efficiency to scale-up compute



Increase in the environmental damages of produced graphics cards

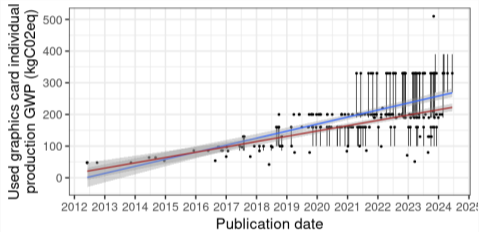


Carbon Footprint

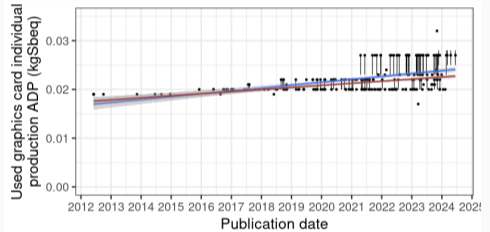


Metallic resource depletion

Increase in the environmental damages of graphics cards used

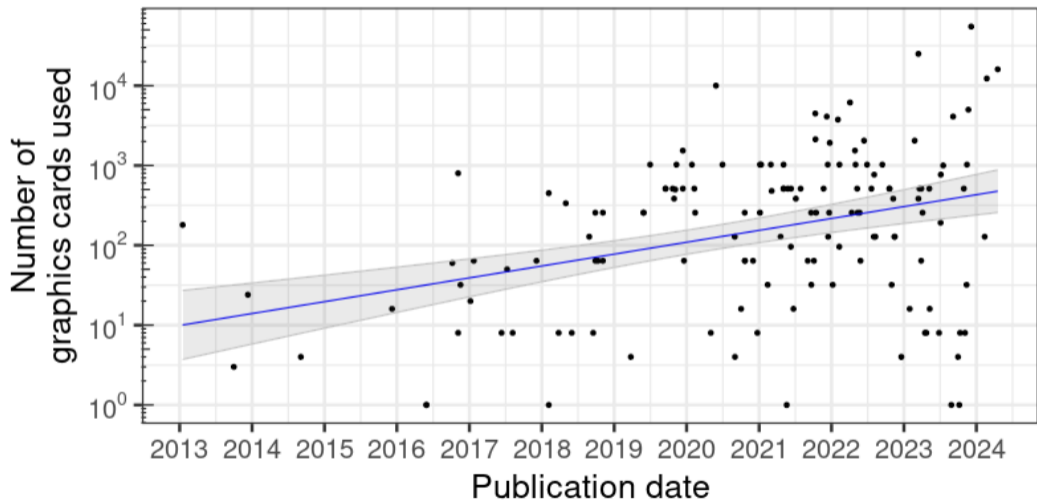


Carbon Footprint

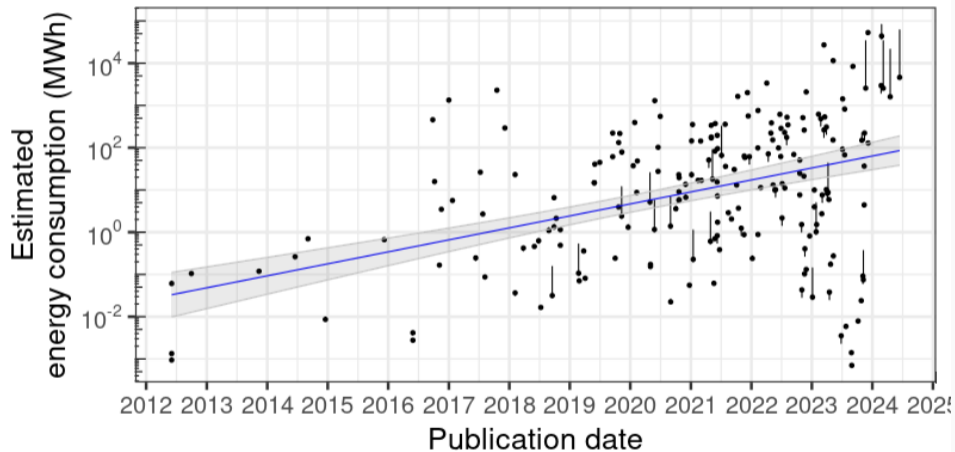


Metallic resource depletion

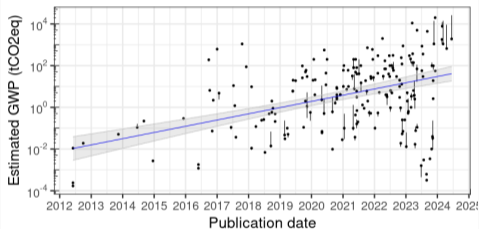
Large increase in the number of cards to train models



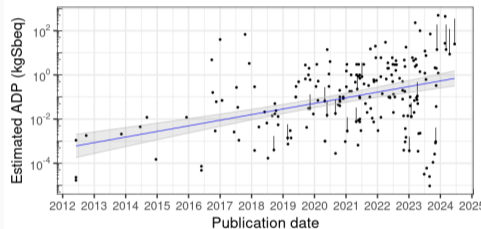
Exponential increase in the energy consumption of models training



Exponential increase in the environmental damages of models training

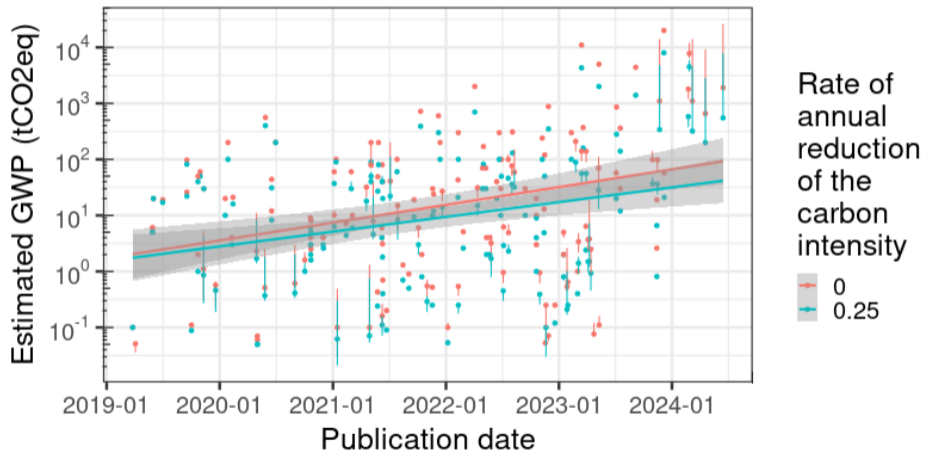


Carbon footprint



Metallic resource depletion

Greener energy cannot void carbon footprint of models training






Conclusion





Current impact reduction strategies alone cannot curb the growth in the environmental impacts of AI.



- Impacts are partly shifting to the production phase
- Increase in the environmental damages of producing graphics cards
- Optimizations have served scaling-up and not scaling down
- Growth paradigm for machine learning models translates into an exponential growth of the energy consumption and environmental damages of models training
- Need to combine impact reduction strategies with broader reflection on the place and role of AI in a sustainable society.




References



-  Bol, D., Pirson, T., and Dekimpe, R. (2021).
Moore's law and ICT innovation in the anthropocene.
In *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 19–24.
-  Coroama, V. C. and Mattern, F. (2019).
Digital rebound - why digitalization will not redeem us our environmental sins.
In Wolff, A., editor, *Proceedings of the 6th International Conference on ICT for Sustainability, ICT4S 2019, Lappeenranta, Finland, June 10-14, 2019*, volume 2382 of *CEUR Workshop Proceedings*. CEUR-WS.org.
-  de Vries, A. (2023).
The growing energy footprint of artificial intelligence.
Joule, 7(10):2191–2194.


-  Desislavov, R., Martínez-Plumed, F., and Hernández-Orallo, J. (2023).
Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning.
Sustainable Computing: Informatics and Systems, 38:100857.
-  Epoch AI (2022).
Parameter, compute and data trends in machine learning.
Accessed: 2024-07-19.
-  Gossart, C. (2015).
Rebound effects and ict: A review of the literature.
In Hilty, L. M. and Aebischer, B., editors, *ICT Innovations for Sustainability*, pages 435–448, Cham. Springer International Publishing.
-  International Energy Agency (IEA) (2024).
Electricity 2024.
Licence: CC BY 4.0.

-  Lannelongue, L., Grealey, J., and Inouye, M. (2021).
Green algorithms: Quantifying the carbon footprint of computation.
Advanced Science, 8(12):2100707.
-  Luccioni, A. S., Viguier, S., and Ligozat, A.-L. (2023).
Estimating the carbon footprint of BLOOM, a 176b parameter language model.
Journal of Machine Learning Research, 24(253):1–15.
-  Masanet, E., Shehabi, A., Lei, N., Smith, S., and Koomey, J. (2020).
Recalibrating global data center energy-use estimates.
Science, 367(6481):984–986.
-  Morand, C., Névéol, A., and Ligozat, A.-L. (2024).
MLCA: a tool for Machine Learning Life Cycle Assessment.
In *2024 International Conference on ICT for Sustainability (ICT4S)*, Stockholm, Sweden.

-  Ostrouchov, G., Maxwell, D., Ashraf, R. A., Engelmann, C., Shankar, M., and Rogers, J. H. (2020).
Gpu lifetimes on titan supercomputer: Survival analysis and reliability.
In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–14.
-  Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D. R., Texier, M., and Dean, J. (2022).
The carbon footprint of machine learning training will plateau, then shrink.
Computer, 55(7):18–28.

-  Schmidt, V., Goyal-Kamal, Courty, B., Feld, B., Amine, S., kngoyal, Zhao, F., Joshi, A., Luccioni, S., Léval, M., Bogroff, A., de Lavoreille, H., Laskaris, N., Connell, L., Wang, Z., Saboni, A., Catovic, A., Blank, D., Stechly, M., alencon, JPW, Books, M., Swadik, S., M., H., Coutarel, M., Pollard, M., McCarthy, C., Husom, E. J., Vicente, F., and Tae, J. (2022).
mlco2/codecarbon: v2.1.4.
-  Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. (2020).
Green AI.
Commun. ACM, 63(12):54–63.
-  Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., and Villalobos, P. (2022).
Compute trends across three eras of machine learning.
In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

-  Strubell, E., Ganesh, A., and McCallum, A. (2019).
Energy and policy considerations for deep learning in NLP.
In Korhonen, A., Traum, D. R., and Màrquez, L., editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3645–3650. Association for Computational Linguistics.
-  Thompson, N., Greenewald, K., Lee, K., and Manso, G. F. (2023).
The Computational Limits of Deep Learning.
In *Ninth Computing within Limits 2023*. LIMITS.
<https://limits.pubpub.org/pub/wm1lwjce>.

 Wu, C., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Behram, F. A., Huang, J., Bai, C., Gschwind, M., Gupta, A., Ott, M., Melnikov, A., Candido, S., Brooks, D., Chauhan, G., Lee, B., Lee, H. S., Akyildiz, B., Balandat, M., Spisak, J., Jain, R., Rabbat, M., and Hazelwood, K. M. (2022).

Sustainable AI: environmental implications, challenges and opportunities.

In Marculescu, D., Chi, Y., and Wu, C., editors, *Proceedings of Machine Learning and Systems 2022, MLSys 2022, Santa Clara, CA, USA, August 29 - September 1, 2022*.
mlsys.org.