Carbon Footprint of Machine Learning

David Patterson, Google and UC Berkeley
September 2022

Based on the following paper:

The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink

IEEE Computer, July 2022

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang,
Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, & Jeff Dean

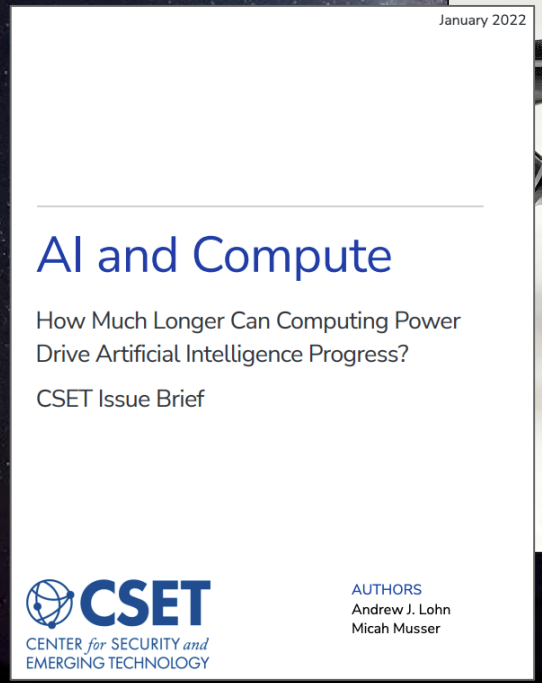
Malthusian Predictions about ML Training

- Environmental cost to improve ML task (2024)?*
  "The answers are grim: Training such a model would cost US $100 billion and would produce as much carbon emissions as New York City does in a month. And if we estimate the computational burden of a 1 percent error rate, the results are considerably worse."

  Thompson, N.C., et al., October 2021.
  Deep Learning's Diminishing Returns: The Cost of Improvement is Becoming Unsustainable, *IEEE Spectrum*

- *"In fact, by 2026, the training cost of the largest AI model predicted by the compute demand trend line would cost more than the total U.S. GDP."*
  [$20T]

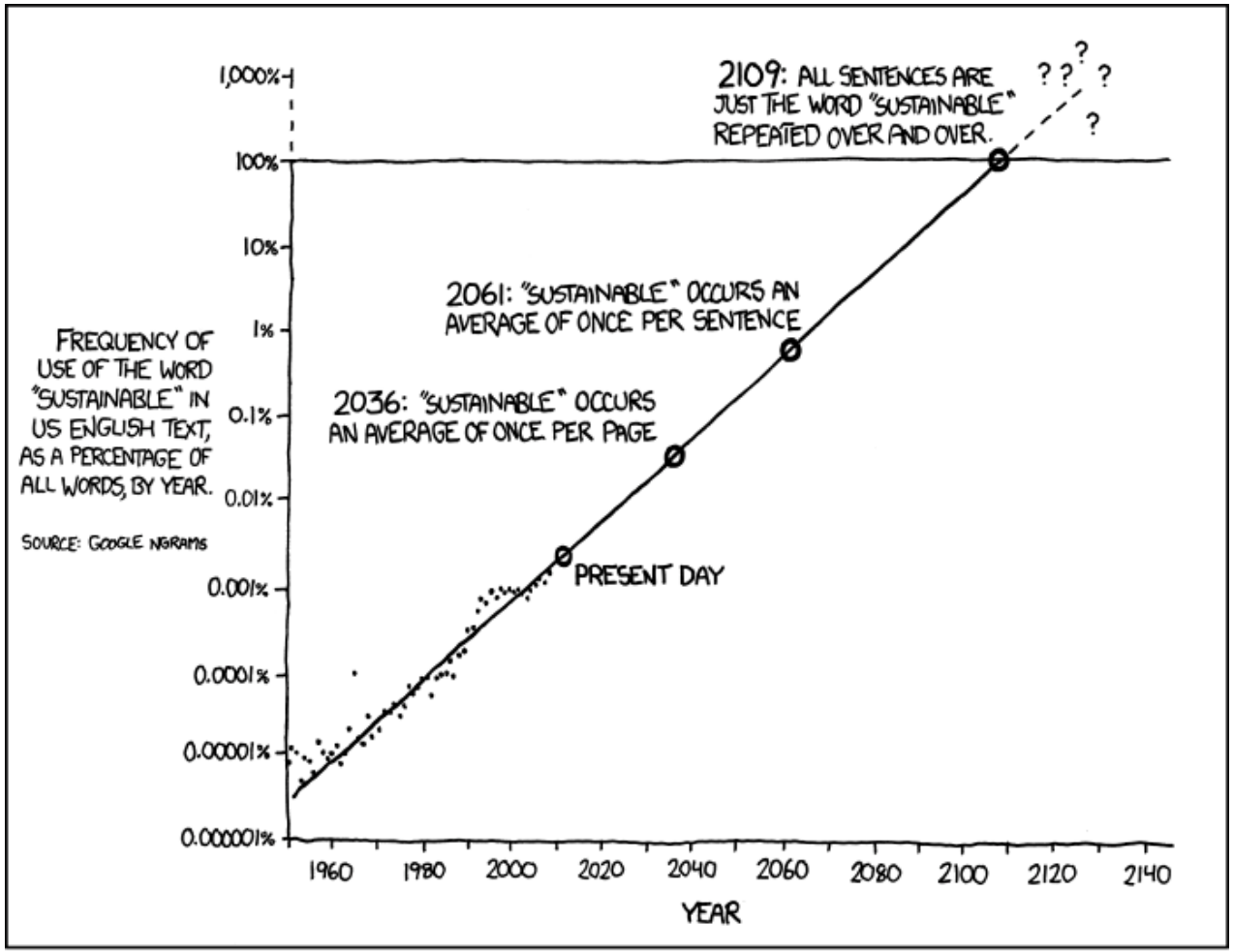
  Lohn, J. and Musser, M., January 2022.
  AI and Compute —How Much Longer Can Computing Power Drive Artificial Intelligence Progress?
  Center for Security and Emerging Technology



* The ML task is object recognition using the Imagenet benchmark to reduce the error rate for an ML task* to a 5% from 11.5% today.

Google

https://xkcd.com/1007/

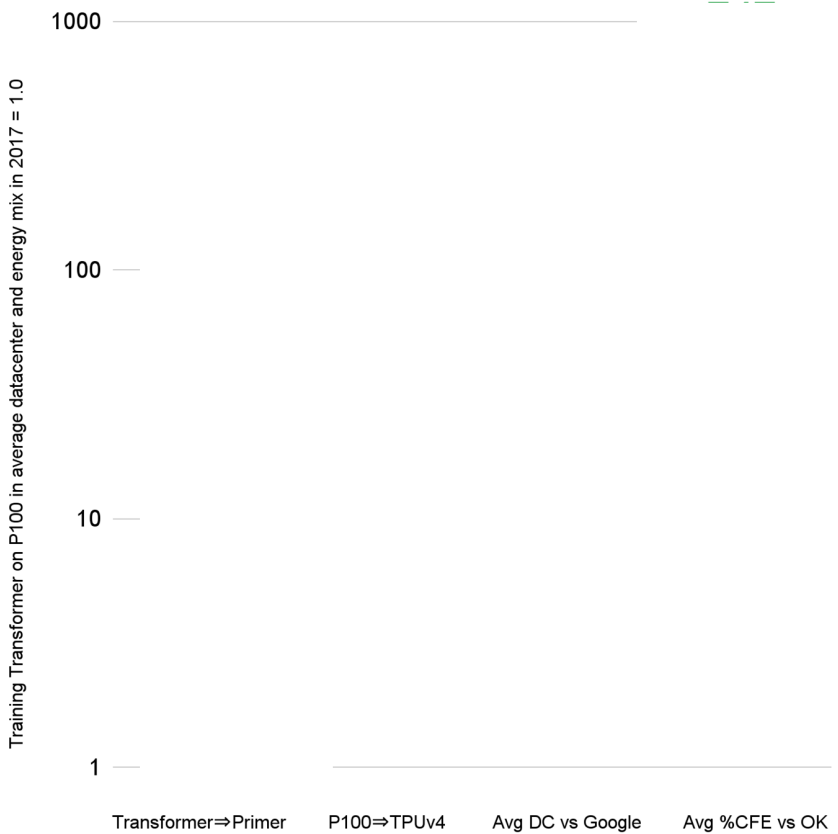


THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.

Good News #1: Reduce energy 100X, CO2e 1000X

Energy efficiency in ML can be improved by 4 (multiplicative) best practices "**4Ms of ML Energy Efficiency**"

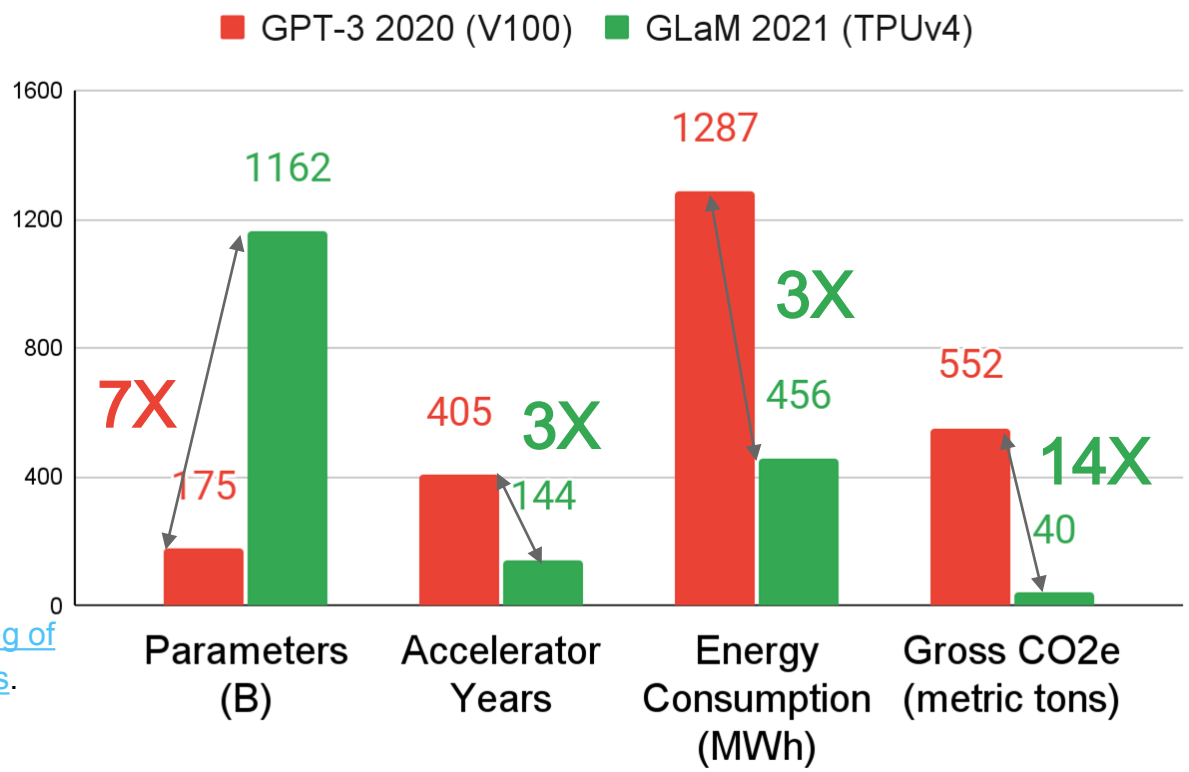
1. Model. Transformer (2017) to Primer (2021) is 4x
2. Machine. P100 (2017) to TPUv4 (2021) is 14x
3. Mechanization (datacenter efficiency). PUE from global average to Google average is 1.4x
4. Maps (geographic location, energy source). Avg %Carbon Free Energy (2017) to Google OK%CFE is 9x (2021)



Thanks to Cliff Young for 4M mnemonic!

4Ms for NLP: GLaM (TPUv4, Google Oklahoma datacenter, 2021) vs GPT-3 (V100 GPU, Microsoft datacenter, 2020)

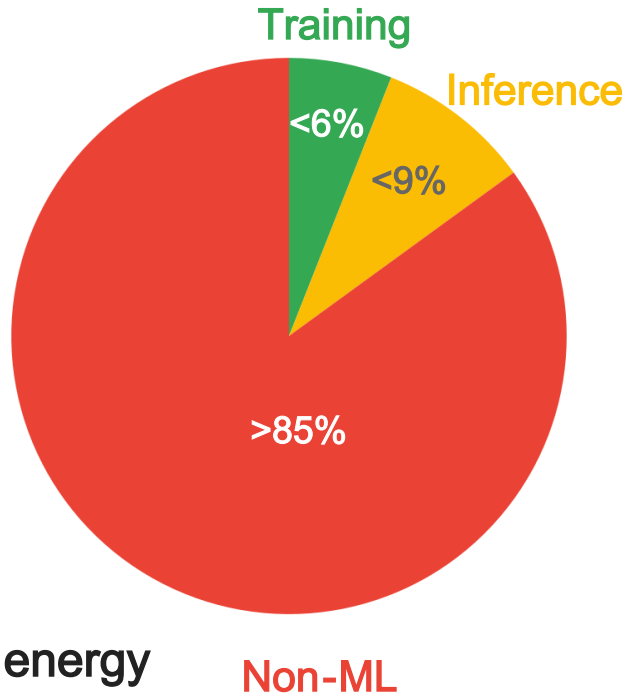
- 18 months after GPT-3
- GLaM has *better accuracy* for same tasks as GPT-3
- **7X** more parameters
- Mixture of experts: **8% parameters**/token
- **3X** less time, energy
- **14X** less CO_2e



Du, N., et al 2021. GLaM: Efficient Scaling of Language Models with Mixture-of-Experts. arXiv preprint arXiv:2112.06905.

Good News #2: ML at Google <15% overall energy

- ML energy use April 2019, 2020, 2021
- Almost all ML training and most inference run on TPUs and GPUs
  - For CPU inference, Google-Wide Profiling to measure libraries used for ML inference
- Each year for past 3 years, ML portion of Google energy use (research, development, production) between 10% and 15%
  - Overall energy use grows annually with usage, but ML% is stable
- 3/5 for inference, 2/5 for training/year
- **DNNs were 70% - 80% FLOPs yet 10% – 15% energy**
  - Lesson 6: It's the memory, stupid (not the FLOPs)
  - Lesson 8: Logic, Wires, SRAM, & DRAM improve unequally
  - CO_2e if replaced TPUs with CPUs of equivalent FLOPS? (50X as many?)



Climate change is one of our most important challenges

- **But must get numbers right to ensure work on biggest challenges**
  - No good way to amend published papers with faulty numbers
  - Check your results with original authors before publishing?

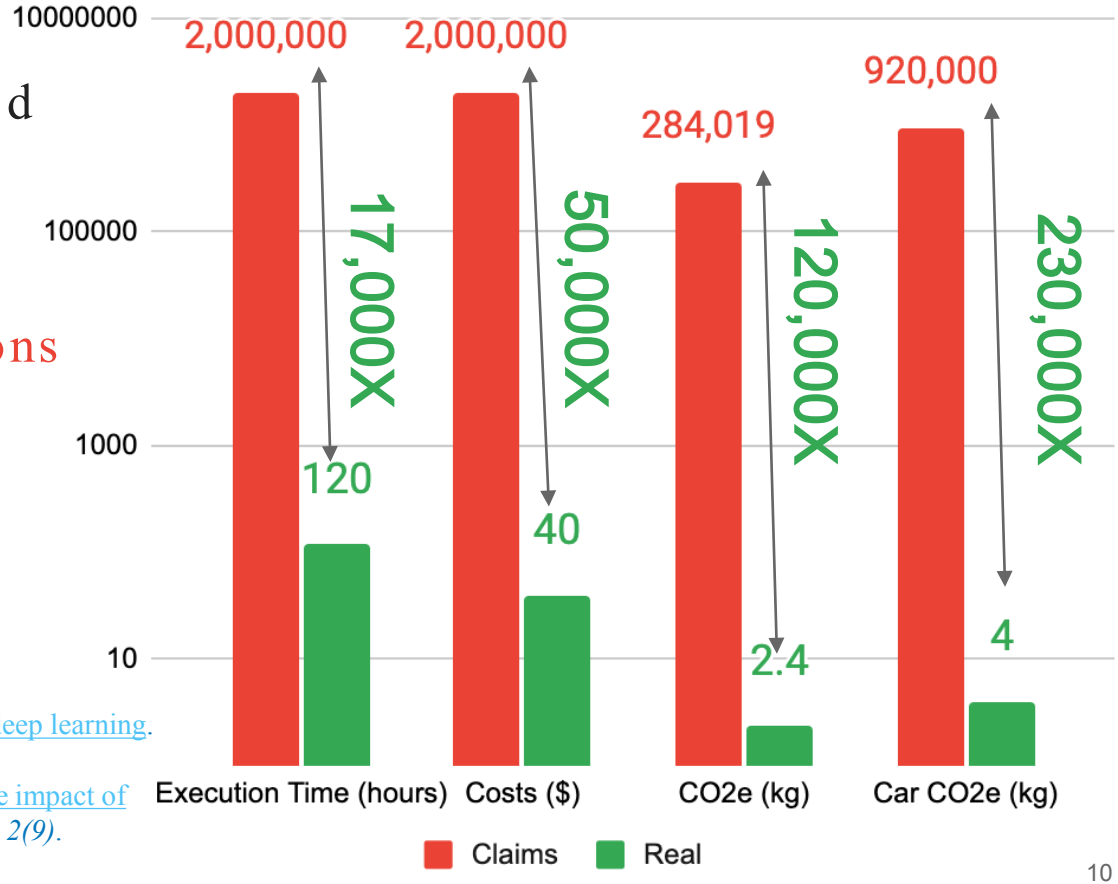
Good News #3: Dire ML estimates were faulty

- Concerns rightly raised about CO_2e of ML
- [So19] NAS for Evolved Transformer didn't include emissions
- [Str19] estimated emissions of this Neural Architecture Search (NAS)
  - Cited ~1500 times
  - Used P100 vs TPUv2, US averages vs Google DC: 5X too high for NAS
  - + Used full model vs small proxy for search: 19X ⇒ 88X too high for NAS
- Some papers citing [Str19] confused NAS with Training cost
  - NAS done once per problem domain+architectural search space
  - NAS emissions ~1000x training emissions of DNN model found in search

[Str19] Strubell, E., Ganesh, A. and McCallum, A., June 2019. Energy and policy considerations for deep learning in NLP. *Annual Meeting of the Association for Computational Linguistics*.
[So19] So, D., Le, Q. and Liang, C., 2019. The Evolved Transformer. In International Conference on Machine Learning (ICML).

Good News #3: Dire ML estimates were faulty

- Claims that training Evolved Transformer took:
  2M GPU hours*,
  Cost $millions*,
  CO_2e = 5X lifetime emissions of a car**
- Right numbers:
  120 TPUv2 hours,
  Cost $40,
  0.00004 car emissions



* Thompson, N.C., et al., 2020. The computational limits of deep learning. arXiv:2007.05558.
** Freitag, C., et al, 2021. The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. *Patterns 2(9)*.

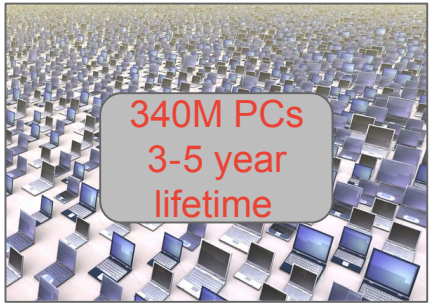
Recommendations for ML Research and ML Practice

- **Model**: ML researchers keep developing more efficient ML models: 2x–4x
  - Challenge: Also publish energy consumption and carbon footprint of model to
    - Foster competition beyond ML quality e.g., speed, emissions
    - Ensure accurate accounting of their work (external estimates were off 100x–100,000x)
- **Machine**: Build faster, more efficient ML HW (e.g., A100 GPU, TPU v4): 2x–4x
  - Challenge: How to do lifecycle costs (Scope 3), not just operational costs (Scope 2)
- **Mechanization** : Data center operators publish data center efficiency (PUE): 1.4x
  - Challenge: Also publish %carbon free of energy supply and PUE per location
- **Map**: ML practitioners use greenest data centers per region, often in Cloud: 5x-10x
  - Practice: Increase carbon free energy per location (2 in Europe, 3 in US ~90% carbon free energy)
- Co-optimize 4Ms to realize the amazing potential of ML to positively impact many fields in a sustainable manner

Backup Slides

Get numbers right to ensure working on the actual biggest information technology challenge

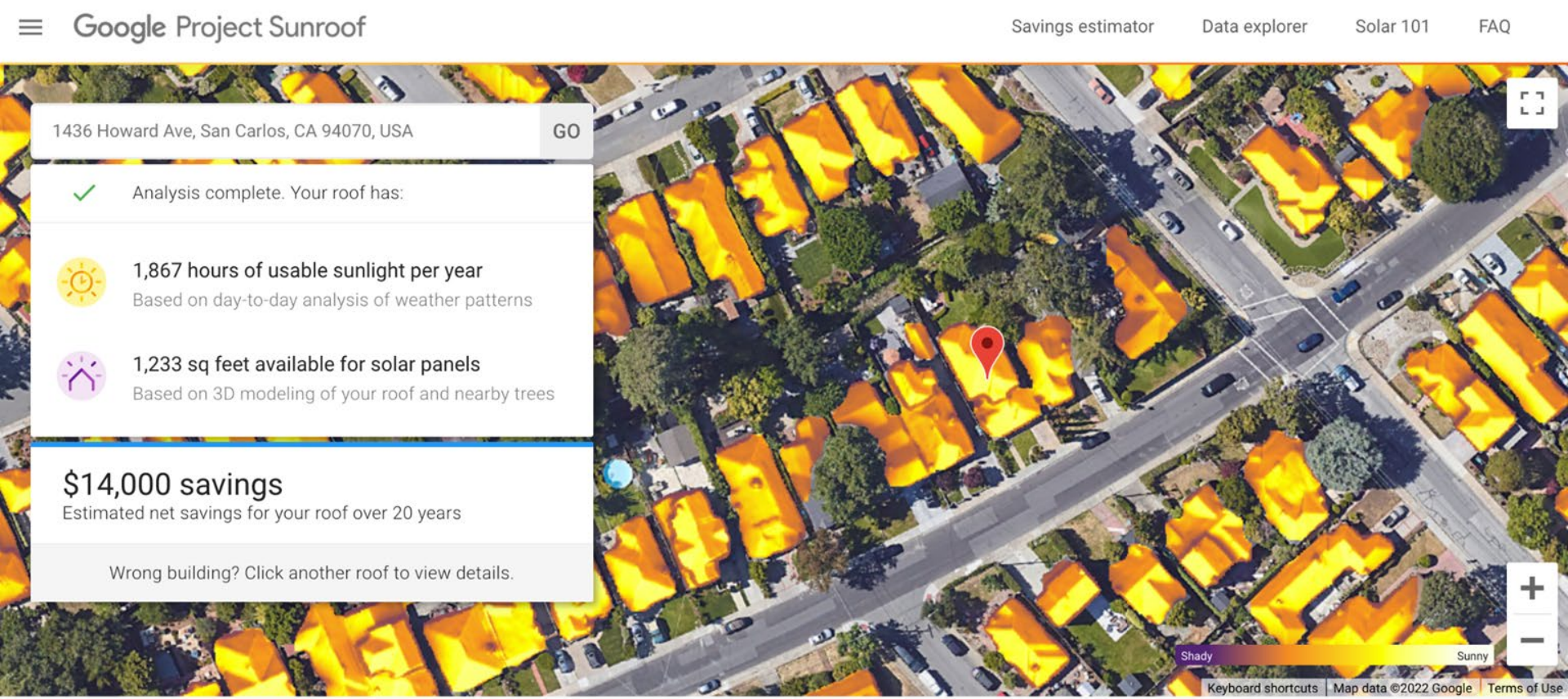
- **Within IT, more likely climate challenge is lifecycle/embodied cost of manufacturing computing equipment of all types/sizes vs operational cost of ML training**



Discussion: Datacenter energy consumption

- Worry that growth of cloud means explosion of energy use
- End users purchasing fewer servers for on premise datacenters, instead computing more in cloud
  - Cloud is greener: Lower PUE, not idle burning power, …
- *Science* paper*: global datacenter energy consumption increased by only 6% vs 2010, despite computing capacity increasing by 550% from 2010-2018
- Only 15%-20% workloads moved to the cloud** ⇒ still plenty of headroom for Cloud growth to replace inefficient on-premise datacenters

* Masanet, E., Shehabi, A., Lei, N., Smith, S. and Koomey, J., 2020. Recalibrating global datacenter energy-use estimates. *Science*, 367(6481), pp.984-986.
Koomey, J. and Masanet, E., 2021. Does not compute: Avoiding pitfalls assessing the Internet's energy and carbon impacts. *Joule*, 5(7), pp.1625-1628.
** Evans, B. 2021, Amazon Shocker: CEO Jassy Says Cloud Less than 5% of All IT Spending, https://cloudwars.co/amazon/amazon-shocker-ceo-jassy-cloud-less-than-5-percent-it-spending/

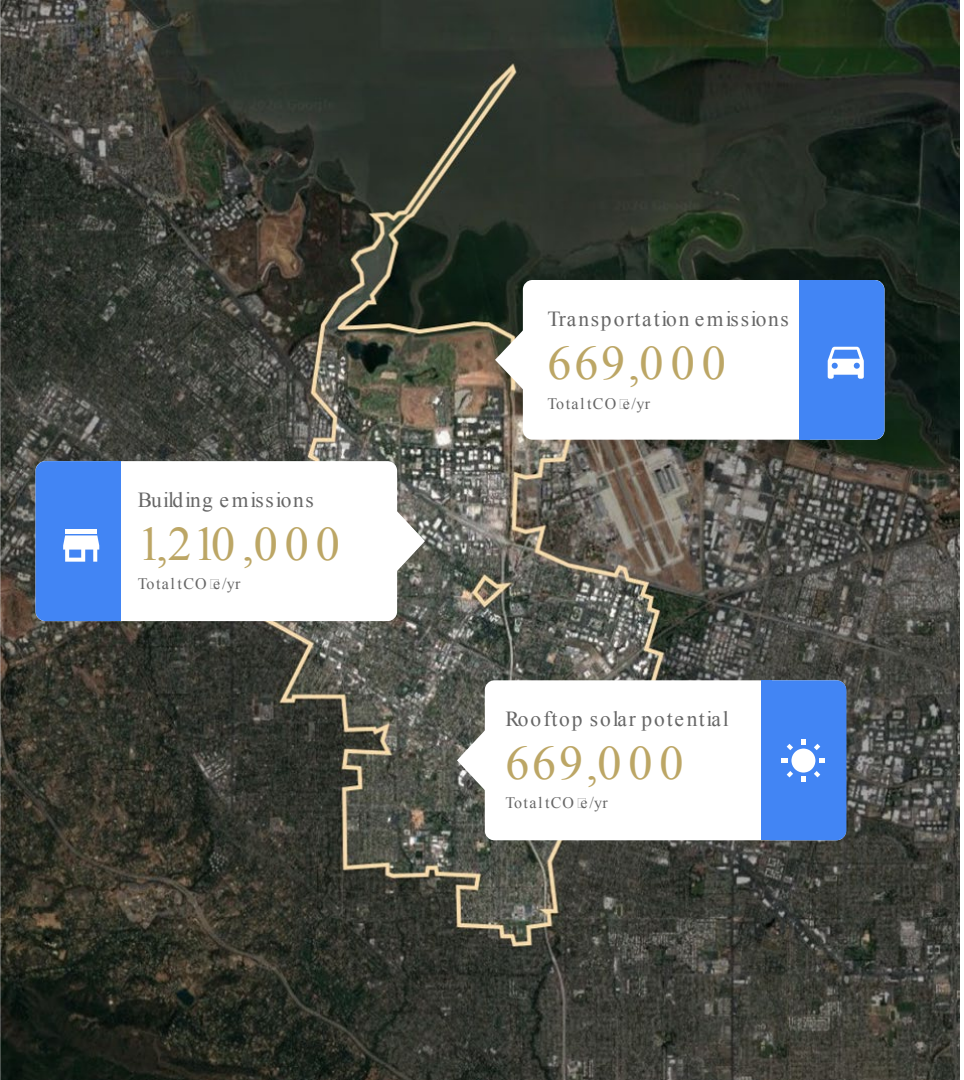


sunroof.withgoogle.com : >170M rooftops mapped w/ solar data across 21,500 cities

Fine-tune your information to find out how much you could save.

From Jeff Dean Keynote "Sustainable Computation and Machine Learning Platforms at Google", MIT Climate Implications of Computing & Communications Workshop, 3/3/22

Google



Environmental Insights Explorer

Helping cities make meaningful progress toward reducing carbon emissions by using Google Maps data

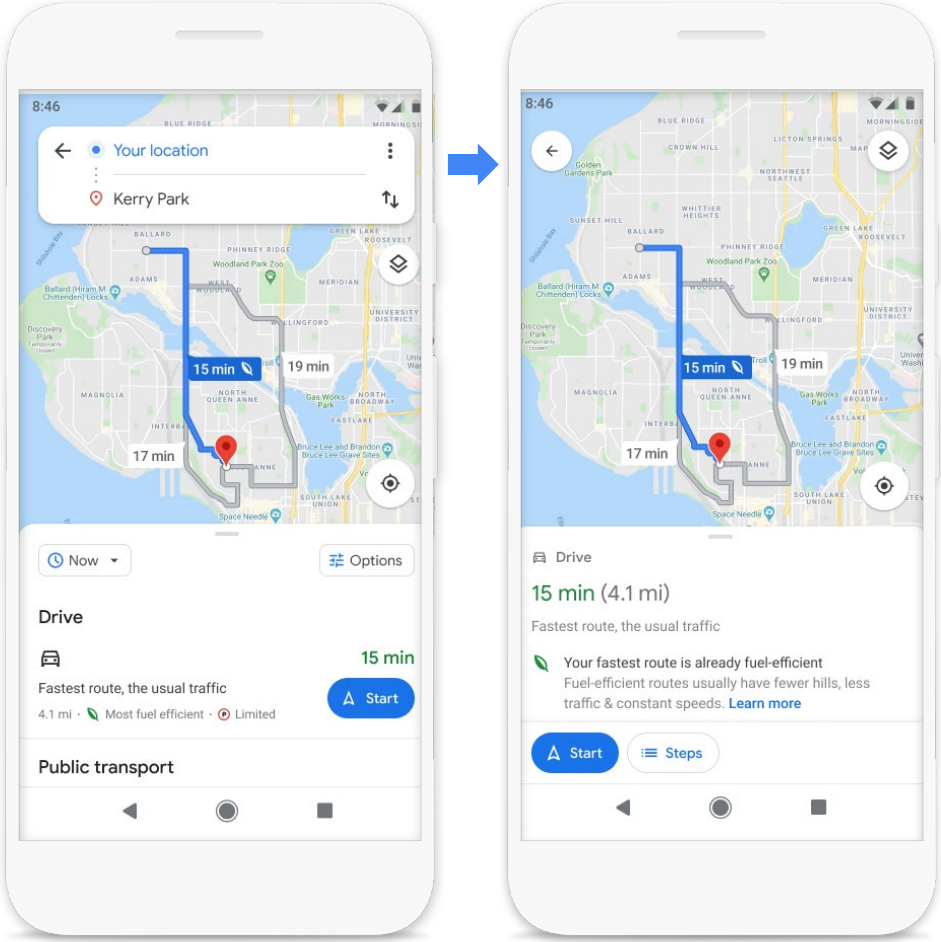
400 cities using Environmental Insights Explorer today

Googles aim to help more than 500 cities reduce an aggregate of 1 gigaton of carbon emissions annually by 2030

From Jeff Dean Keynote "Sustainable Computation and Machine Learning Platforms at Google", MIT Climate Implications of Computing & Communications Workshop, 3/3/22

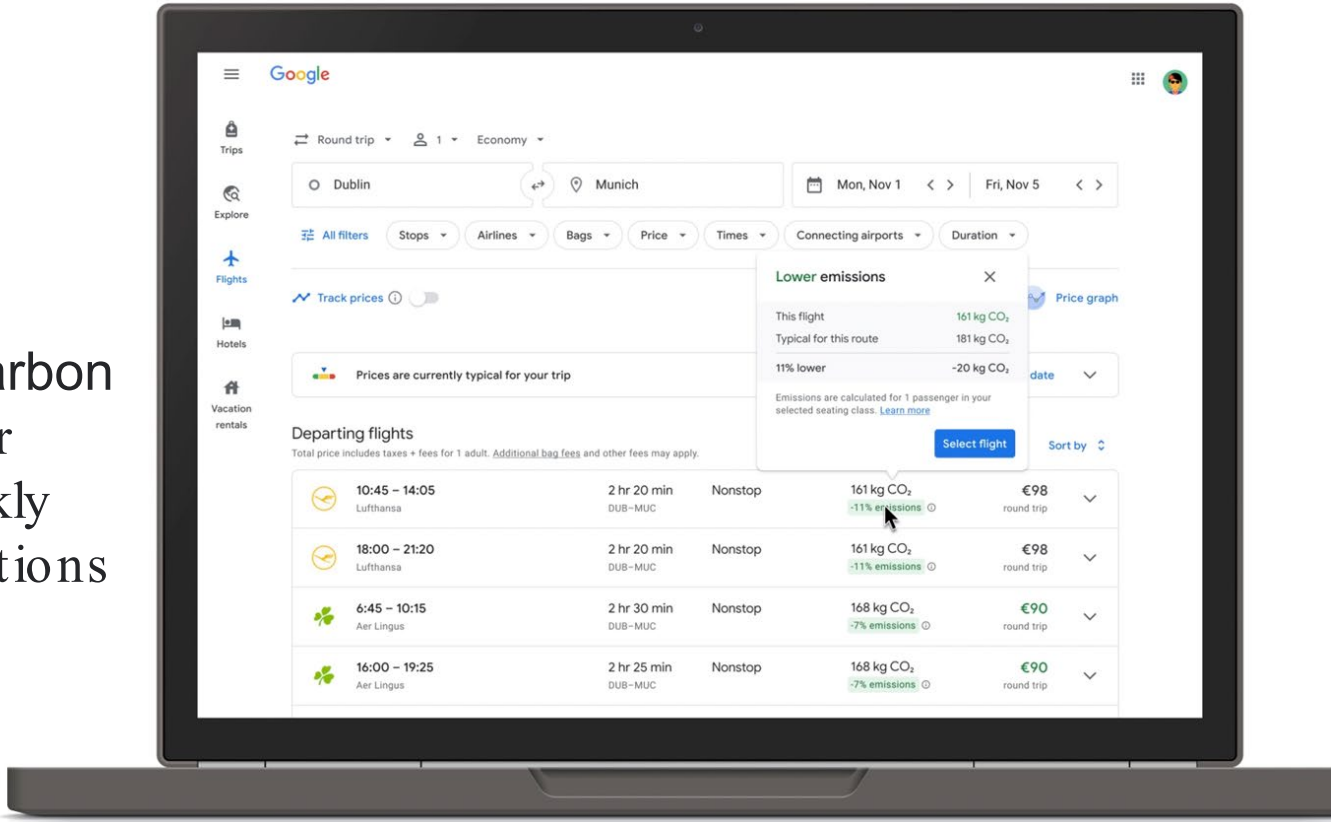
Find more eco-friendly options to get around with Google Maps

e.g. 1 billion km of transit results on Google Maps per day , helping to limit carbon emissions by giving people access to mass transit options, bike routes, and traffic information.



From Jeff Dean Keynote "Sustainable Computation and Machine Learning Platforms at Google", MIT Climate Implications of Computing & Communications Workshop, 3/3/22

See the associated carbon emissions per seat for every flight, and quickly find lower-carbon options on Google Flights



From Jeff Dean Keynote "Sustainable Computation and Machine Learning Platforms at Google", MIT Climate Implications of Computing & Communications Workshop, 3/3/22

Google