

Machine Learning Carbon Footprints

Peter Henderson

We're marching towards trillion parameter models being common

SWITCH TRANSFORMERS: SCALING TO TRILLION
PARAMETER MODELS WITH SIMPLE AND EFFICIENT
SPARSITY

GPT-3 Scared You? Meet Wu Dao 2.0: A Monster of 1.75 Trillion Parameters

Wu Dao 2.0 is 10x larger than GPT-3. Imagine what it can do.

Efficient Large-Scale Language Model Training on GPU Clusters

Deepak Narayanan^{†‡}, Mohammad Shoeybi[†], Jared Casper[†], Patrick LeGresley[†],
Mostofa Patwary[†], Vijay Korthikanti[†], Dmitri Vainbrand[†], Prethvi Kashinkunti[†],
Julie Bernauer[†], Bryan Catanzaro[†], Amar Phanishayee^{*}, Matei Zaharia[‡]
[†]NVIDIA [‡]Stanford University ^{*}Microsoft Research

ZeRO-Infinity and DeepSpeed: Unlocking unprecedented model scale for deep learning training

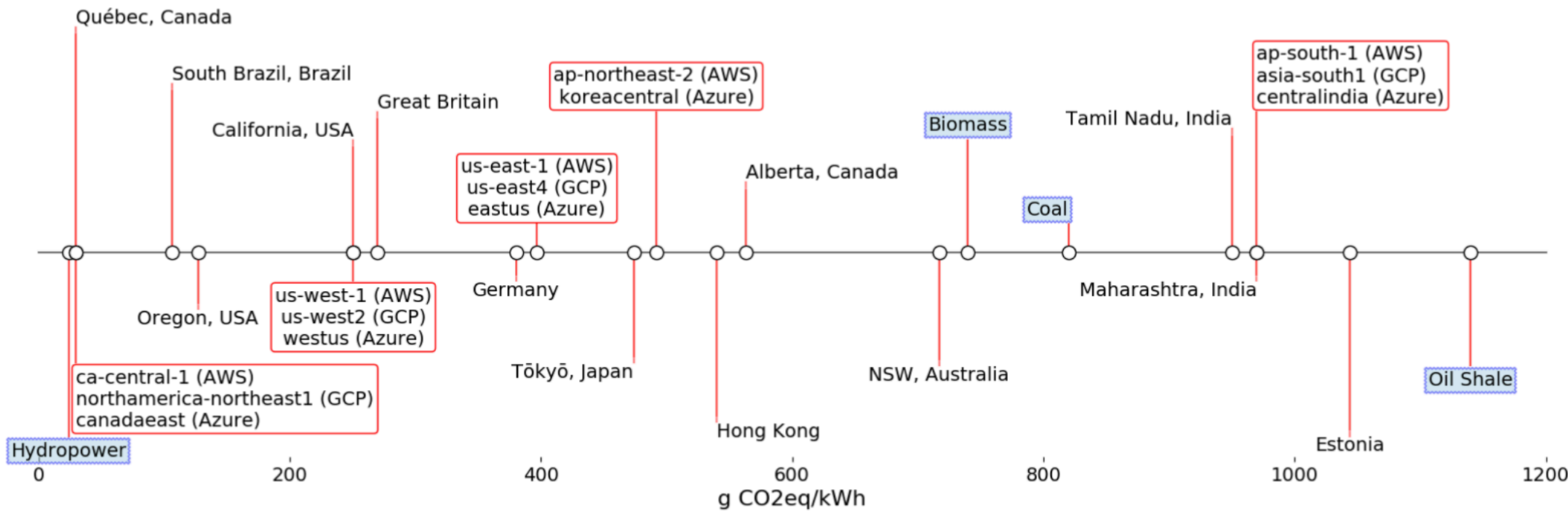
- Offering the system capability to train a model with over 30 trillion parameters on 512 NVIDIA V100 Tensor Core GPUs, 50x larger than state of the art.

Will this have an impact on the environment? Maybe, it depends where you run the model.

Consumption	CO ₂ e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO₂ emissions from training common NLP models, compared to familiar consumption.¹

(Strubell et al., 2019)



(Henderson et al., 2020)

But we don't really know how bad the problem is.

Anecdotally, we tried to audit NeurIPS conference papers and could not even get enough data to estimate carbon emissions for even a small subsample.

Researchers don't report: location, gpu, time of experiments, or enough information to make informed policy recommendations.

Companies are less transparent, but this is improving.

Steps forward

1. **Reduce** unnecessary energy usage (don't rely on carbon offsets)
2. **Understand** the harm-benefit trade-offs of using a large model
3. **Report** metrics necessary for users to make informed decisions about which model to run and deploy.

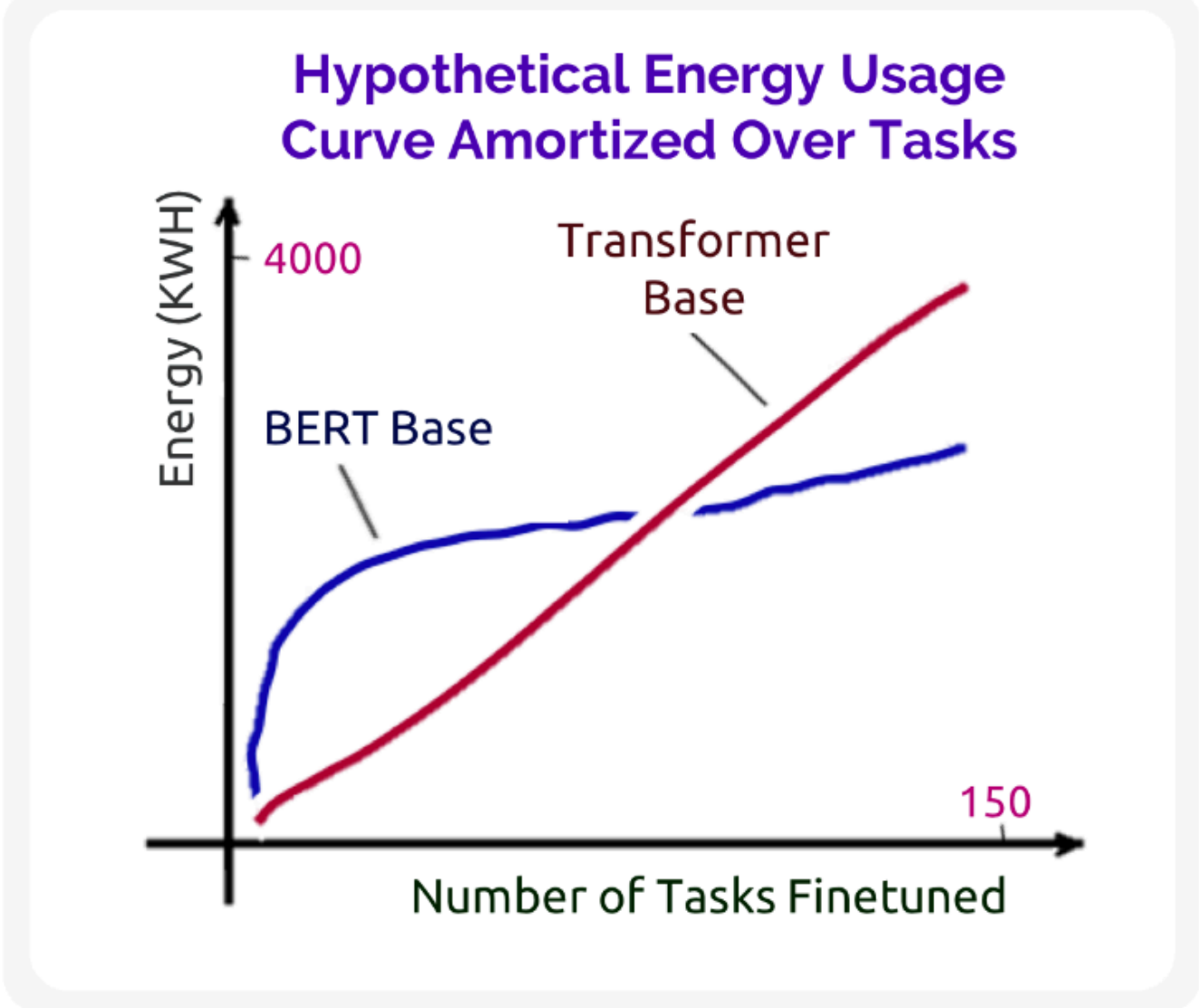
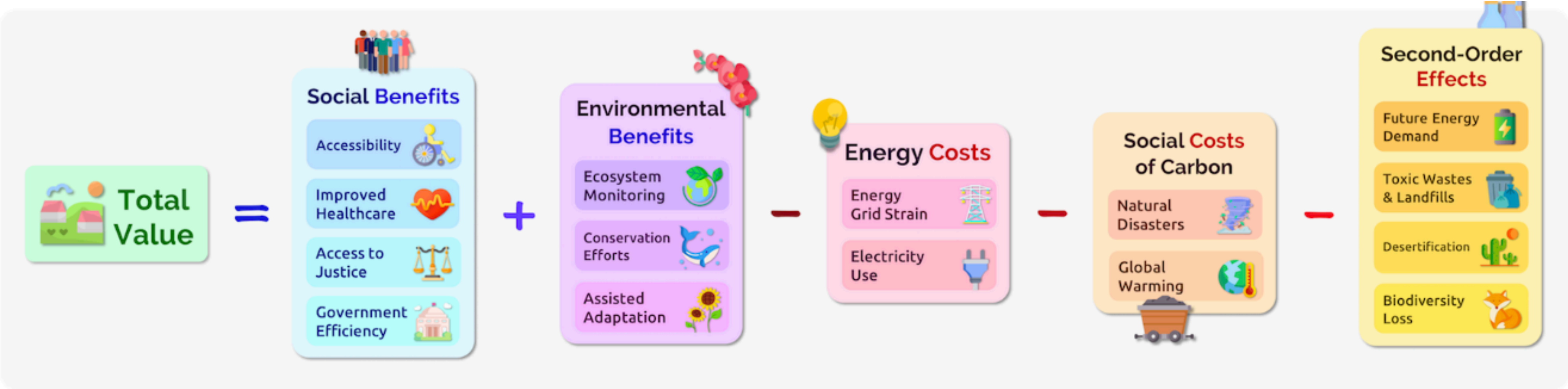
Mitigation

Green **Defaults:** Using 8-bit optimizers by default, distilled models, etc.

Green **Information:** Flagging the green options.

Green **Badges:** Rewarding green alternatives.

Understanding trade-offs

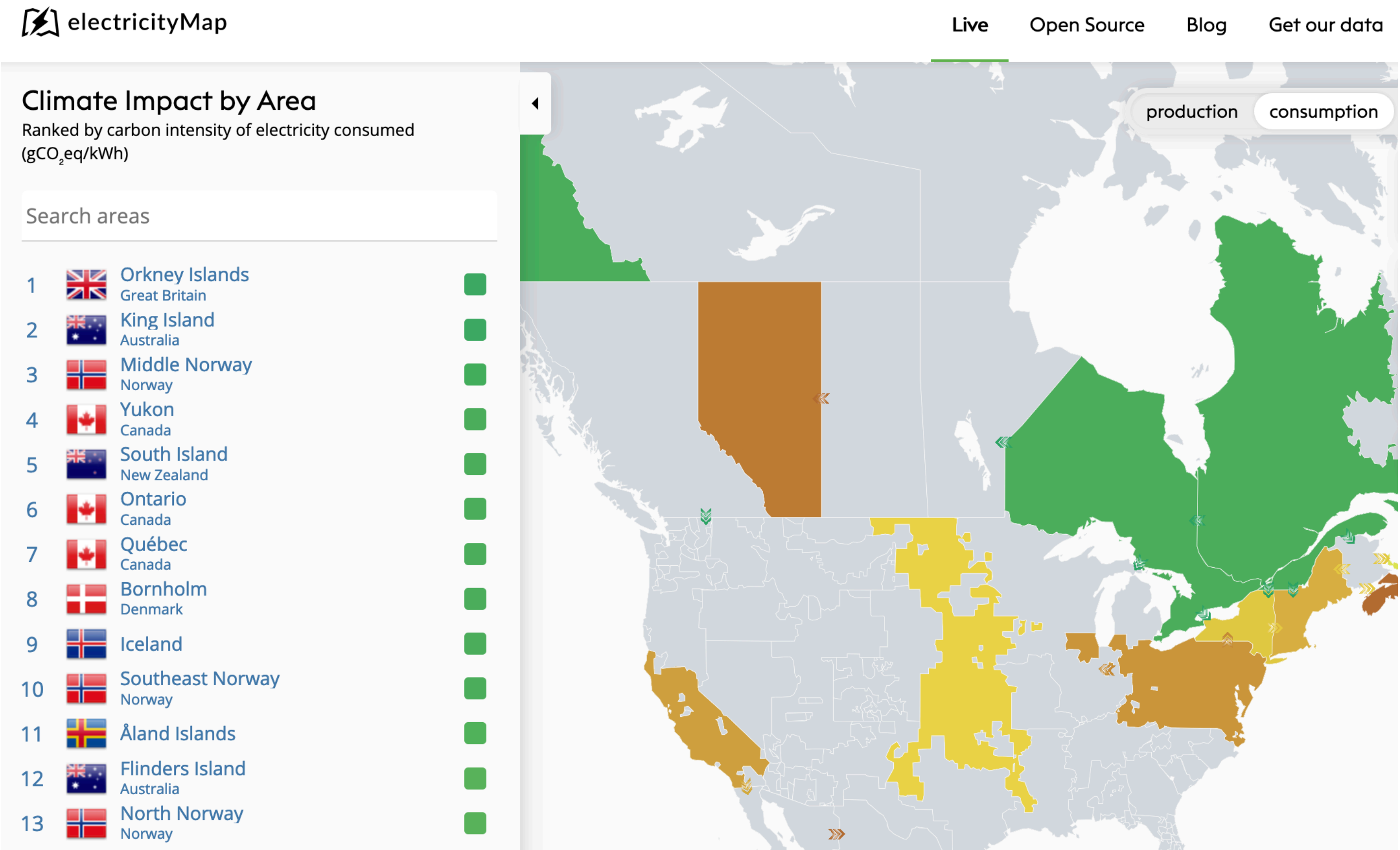


(Bommasani et al., 2021)

Reporting

Cannot understand trade-offs if can't make estimates.


Reporting



Some open source tools try to compensate through realtime calculations, but often patchy. Would be better if all energy grids/cloud providers reported live carbon intensity.

Reporting


Some progress!

 Google Cloud

Google Cloud Region Picker

This tool helps you pick a Google Cloud region considering carbon footprint, price and latency.

Optimize for

 Lower carbon footprint [?]


Not important

Important

\$ Lower price [?]

Not important

Important


 Lower latency [?]

Not important

Important




Recommended regions

1.



europa-north1

Hamina, Finland


  

\$

\$




\$

2.



northamerica-northeast1

Montréal, Canada


  

\$

\$




\$

3.



us-central1

Iowa, USA

\$

\$

\$

Reporting

Some progress!

Breakend / experiment-impact-tracker <small>Public</small>	Unwatch ▾	7	Star	182	Fork	16
mlco2 / codecarbon <small>Public</small>	Watch ▾	16	Star	244	Fork	35
lfwa / carbontracker <small>Public</small>	Watch ▾	8	Star	154	Fork	6
epfl-iglobalhealth / cumulator <small>Public</small>	Watch ▾	3	Star	10	Fork	0

Many tools to track carbon intensity of experiments.

Regulation?

Not clear if needed to restrict scale of compute.

Need to understand scope of problem via reporting, maybe some regulatory effort here.

Much of the problem can be mitigated by moving jobs to green regions.

Solve the energy grid problem and carbon footprints of ML go away.

Regulation?



The screenshot shows the California Energy Commission website. At the top is the logo and a search bar with the placeholder text "Enter keywords, e.g. Tracking Progress". Below the logo is a navigation menu with links: HOME, PROCEEDINGS, RULES AND REGULATIONS, PROGRAMS AND TOPICS, FUNDING, DATA AND REPORTS, and View All. A breadcrumb trail below the menu reads: Home > Resource > Publications > A Plug-Loads Game Changer: Computer Gaming Energy Efficiency without Performance Compromise. The main content area features the title "A Plug-Loads Game Changer: Computer Gaming Energy Efficiency without Performance Compromise" in large, bold, dark blue text. To the right of the title is a green sidebar with the heading "PUBLICATIONS" and four links: "Energy Commission Publications", "Latest Publications", "Energy Research and Development Reports", and "Transportation Reports".

A Plug-Loads Game Changer: Computer Gaming Energy Efficiency without Performance Compromise

PUBLICATIONS

- Energy Commission Publications
- Latest Publications
- Energy Research and Development Reports
- Transportation Reports

Green defaults as regulation.
California bans low-efficiency idle modes in gaming GPUs.