Green AI

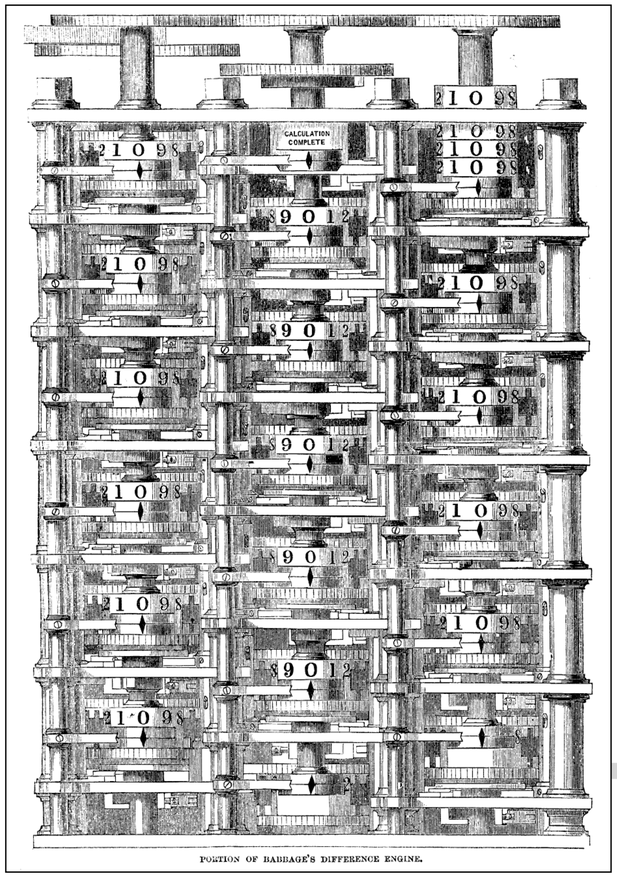
Making Machine Learning Environmentally Sustainable

Charles Humble, Conissaunce Limited
@charleshumble

AI Development

- Most software develops in parallel with improvements in hardware, and this has been largely true in AI

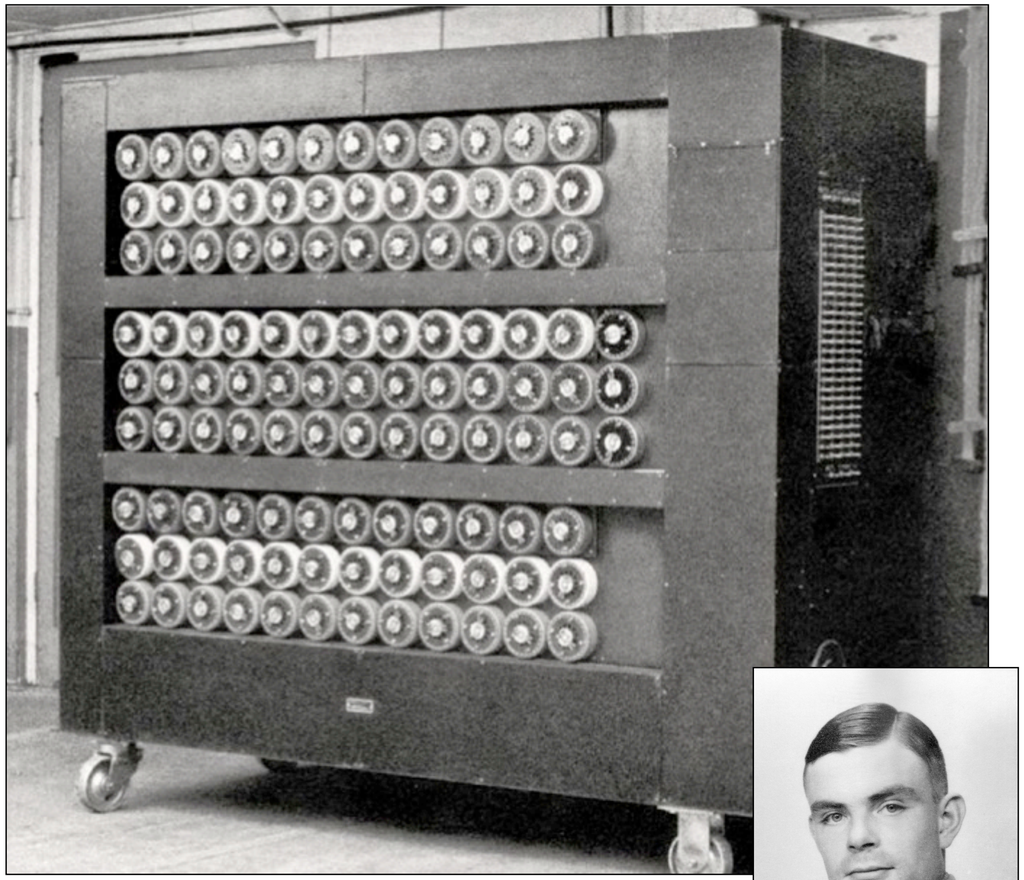
Mechanical AI



PORTION OF BABBAGE'S DIFFERENCE ENGINE.



Difference Engine
Charles Babbage - 1822

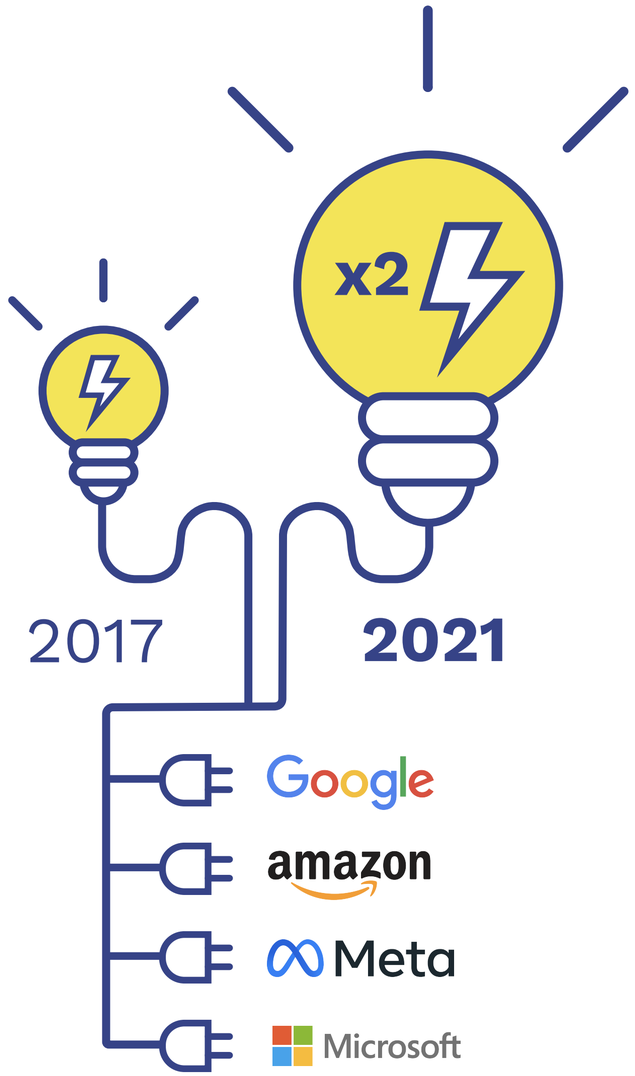


Bombe
AlanTuring - 1940

AI Development

- Most software develops in parallel with improvements in hardware, and this has been largely true in AI

- There has been a shift in the last decade or so with the introduction of LLMs, particularly following the development of GPT (generative pre-trained transformer) in 2012



Combined electricity use by Amazon, Microsoft, Google, and Meta more than doubled between 2017 and 2021, rising to around **72 TWh** *in 2021*

— *International Energy Agency*

https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks#tracking

AI's Hidden Environmental Cost Rises

- Microsoft reported in May of 2024 that its total carbon emissions have risen nearly 30% since 2020, primarily due to the construction of data centres to meet its push into AI

https://www.microsoft.com/en-us/corporate-responsibility/sustainability/report

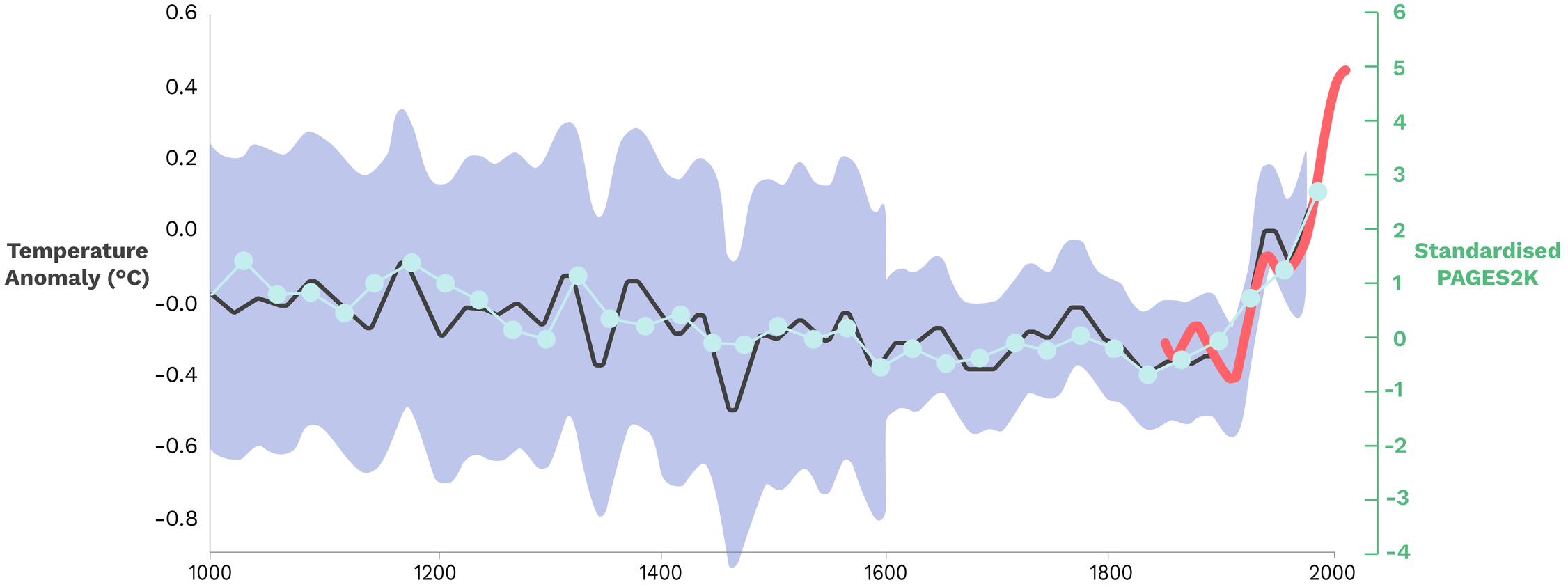
AI's Hidden Environmental Cost Rises

- Microsoft reported in May of 2024 that its total carbon emissions have risen nearly 30% since 2020, primarily due to the construction of data centres to meet its push into AI

- Google’s emissions have surged nearly 50% compared to 2019. They increased 13% year-on-year in 2023 alone, according to their own report. The company attributed the emissions spike to an increase in data centre energy consumption and supply chain emissions driven by artificial intelligence

https://www.gstatic.com/gumdrop/sustainability/google-2024-environmental-report.pdf

The Hockey Stick



https://www.epa.gov/climatechange-science/causes-climate-change

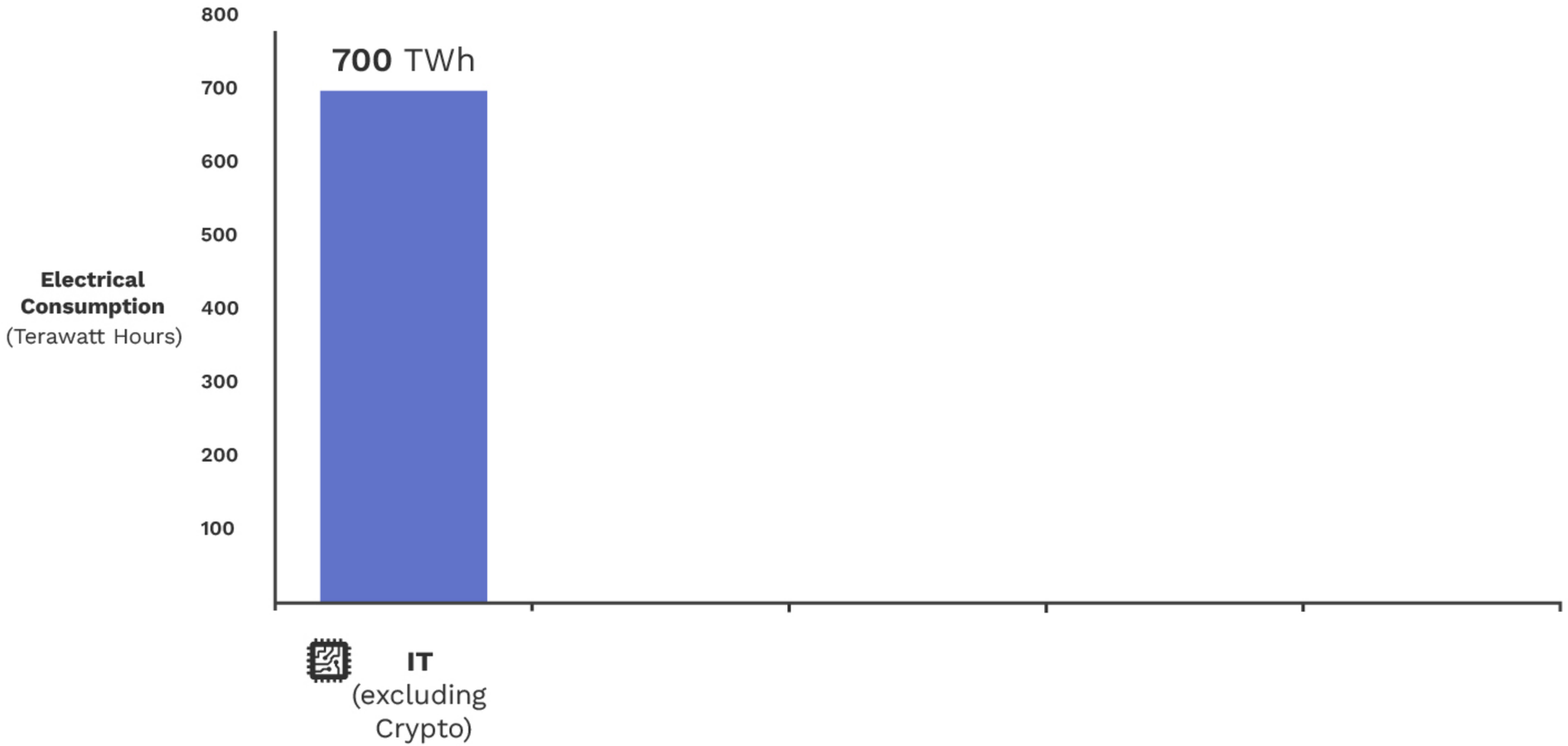
How Much Carbon is IT Responsible For?

- The IEA suggests that estimated global data centre electricity consumption in 2022 was 240-340 TWh, accounting for around 1-1.3% of all global electricity demand

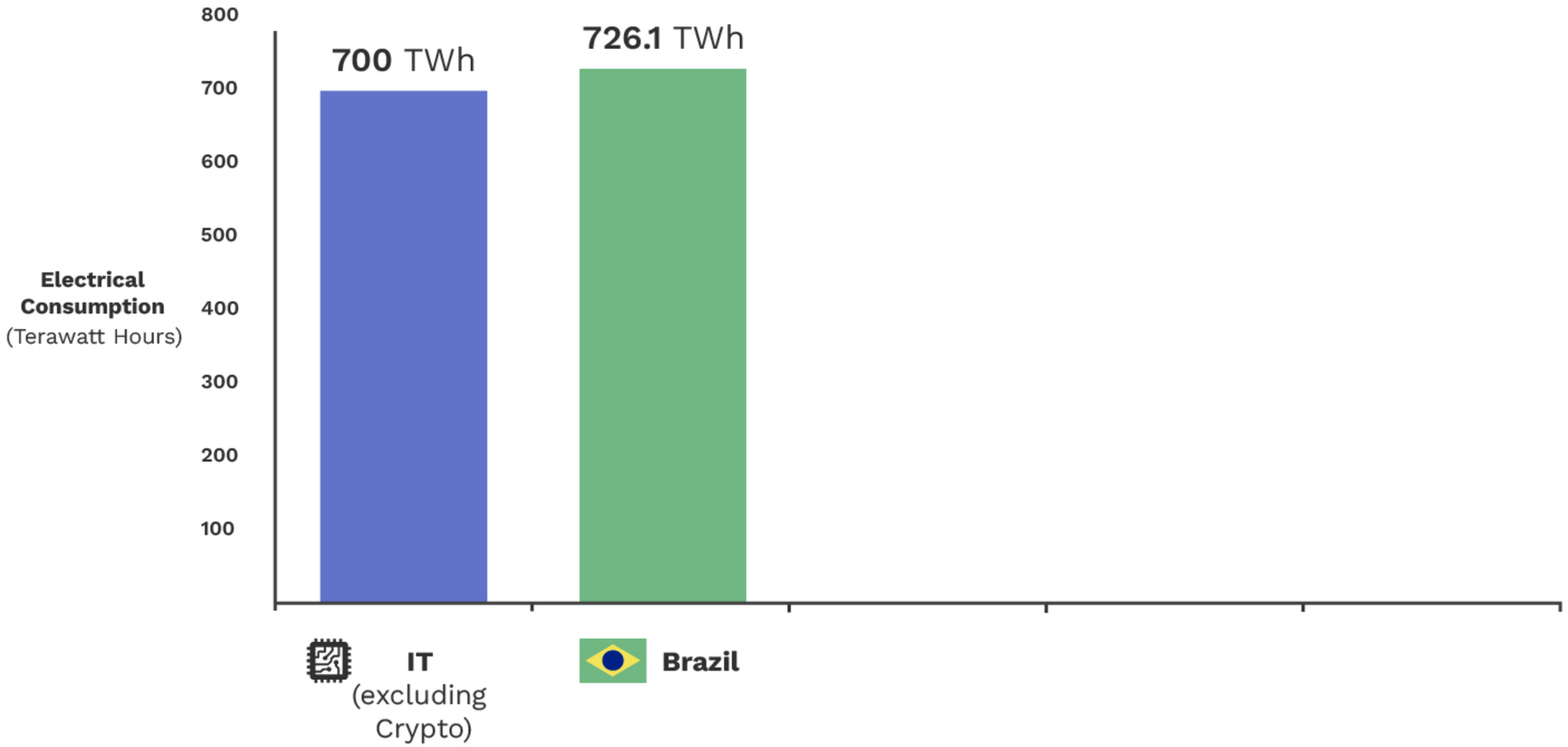
- That figure excludes data transmission networks, which more-or-less double this figure, adding an estimated 260-360 TWh in the same period, or another 1-1.5% of global electricity use

- It also excludes the energy used for cryptocurrency mining, which was estimated to be around 110 TWh in 2022, a further 0.4% of annual global electricity demand

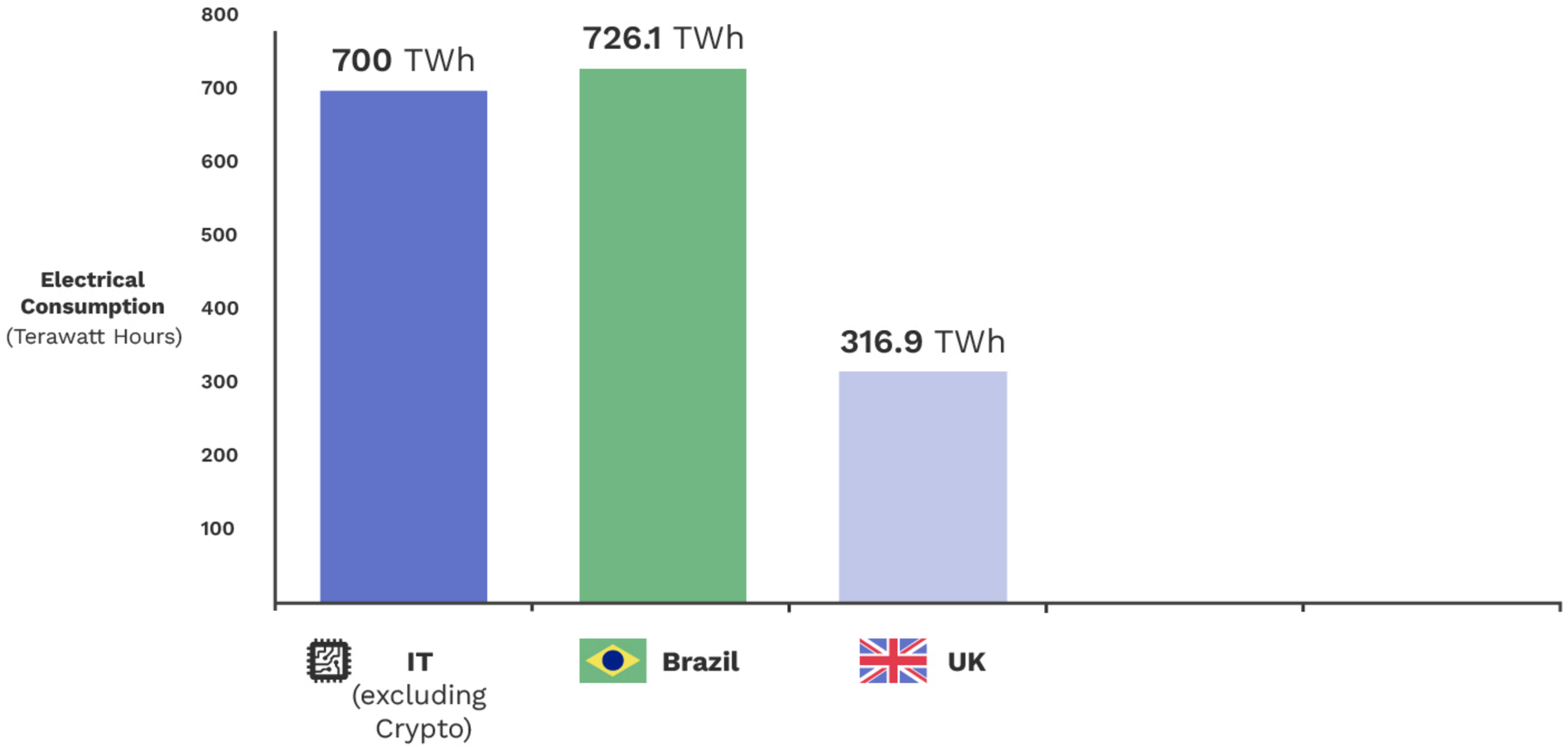
How Much Carbon is IT Responsible For?



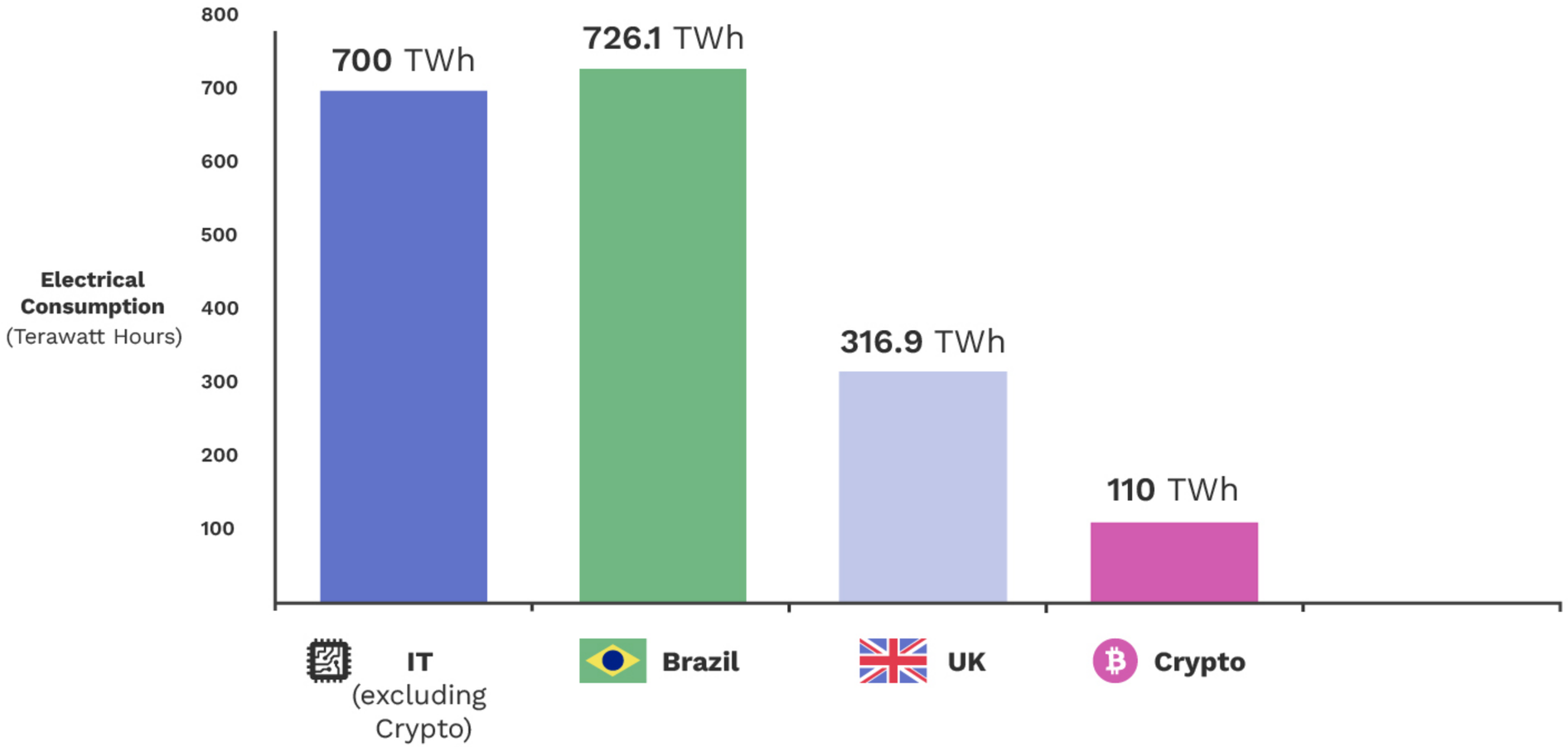
How Much Carbon is IT Responsible For?



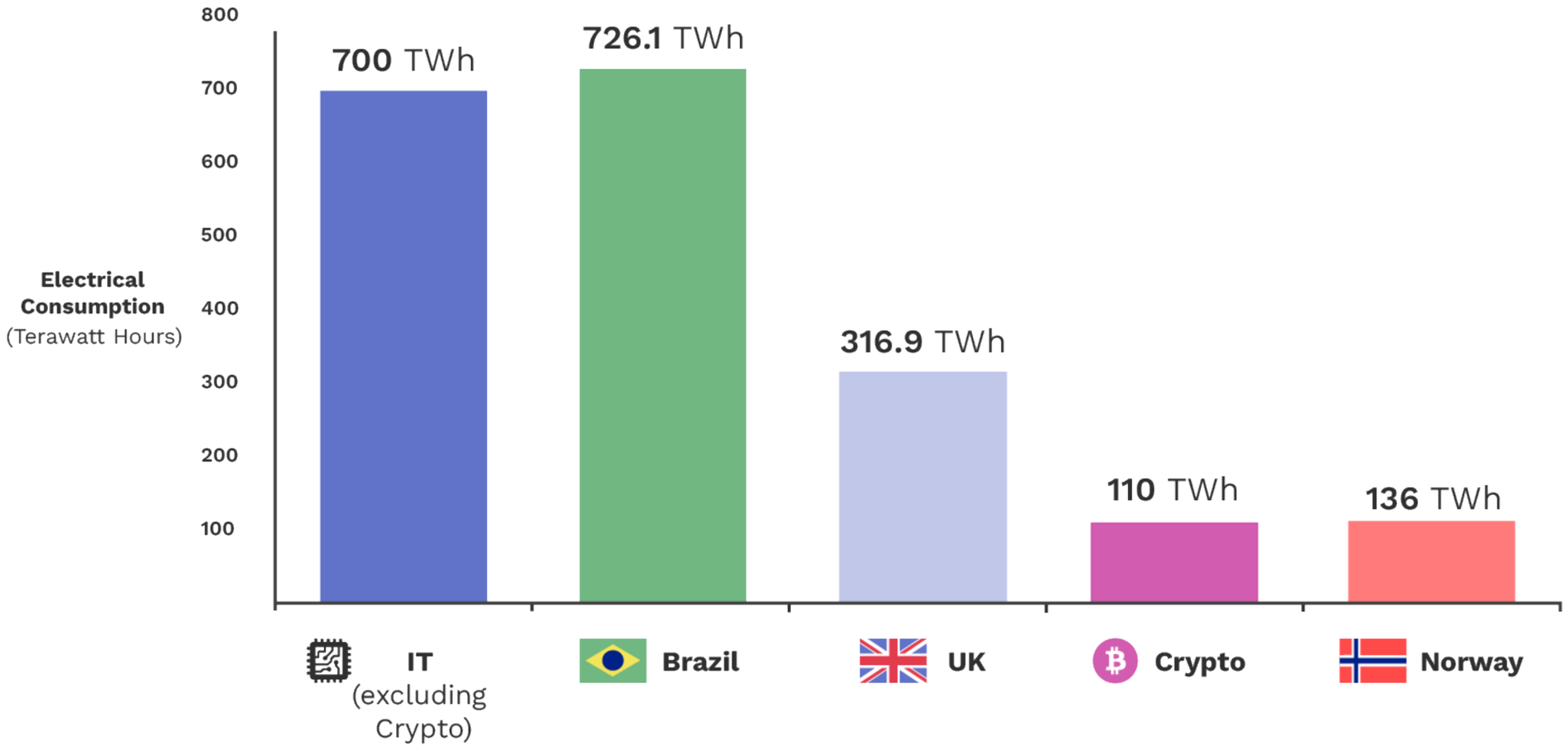
How Much Carbon is IT Responsible For?



How Much Carbon is IT Responsible For?



How Much Carbon is IT Responsible For?



How Much Carbon is IT Responsible For?

- If you look at academic studies, estimates for ICT are around 2-4% of global emissions

How Much Carbon is IT Responsible For?

- If you look at academic studies, estimates for ICT are around 2-4% of global emissions
    - By comparison aviation is about 3%



How Much Carbon is IT Responsible For?

- Globally, data centre electricity consumption has grown by around 12% per year since 2017, more than four times faster than the rate of total electricity consumption

- The IEA's latest projections see electricity consumption from data centres roughly doubling from 485 TWh in 2025 to 950 TWh in 2030, accounting for around 3% of global electricity demand by that date

Embodied Carbon vs. Direct Emissions

- For IT we can sub-divide carbon emissions into direct emissions - the ones from our electricity use, and embodied carbon - the carbon used in the manufacture, transpiration and eventual destruction of our hardware
- For end-user devices - laptops, mobile phones and the like - their embodied carbon absolutely dwarfs their direct carbon
- But with servers and GPUs it isn’t quite as straightforward, because efficiency gains in some cases offset embodied carbon costs

Project Planning



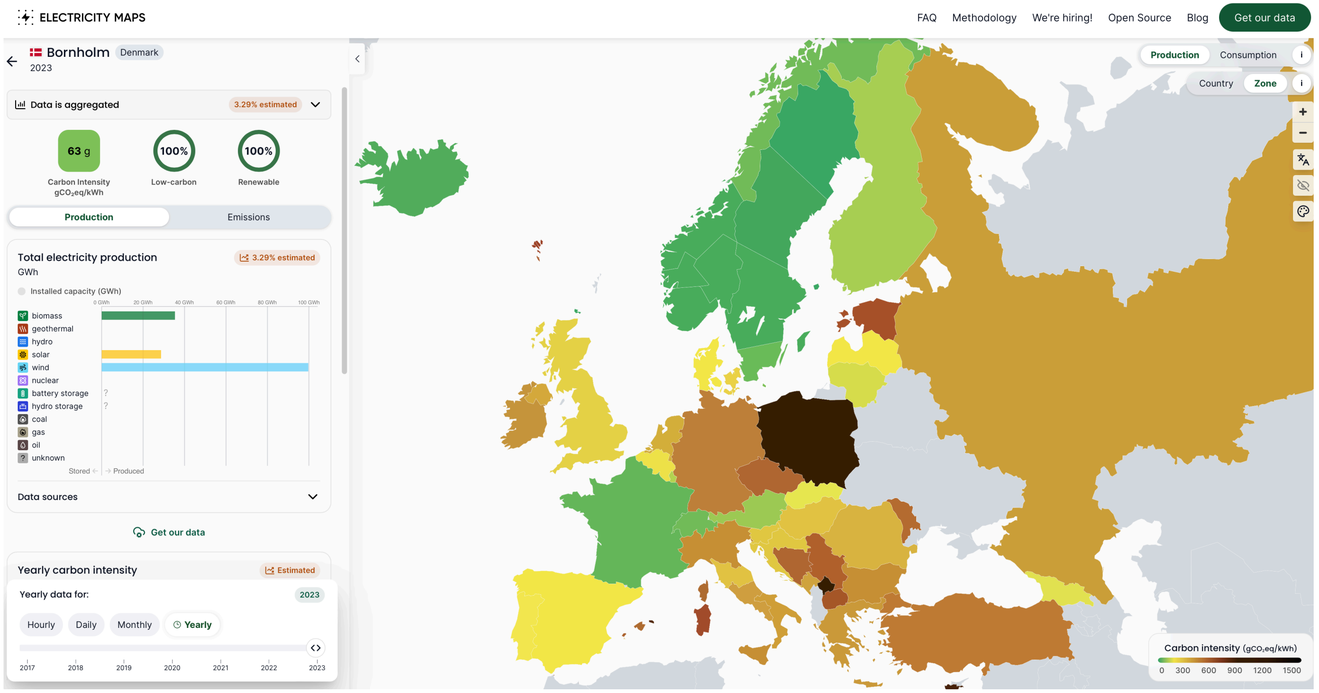
- Do you really need it?
- Do you need to train a new model from scratch or can you use a pre-built one?
  - Think about model choice. Meta has stated that it “developed OPT-175B with energy efficiency in mind by successfully training a model of this size using only 1/7th of the carbon footprint as that of GPT-3”
- If you can relax part of your SLA does that allow you to run your workloads in a greener location?

Location, Location, Location

- Location matters because moving to carbon-free electricity sources is a slow process and, inevitably, different locations will get there at different times

Location, Location, Location

- Location matters because moving to carbon-free electricity sources is a slow process and, inevitably, different locations will get there at different times
- Using a tool like Electricity Maps allows you to identify locations that are using renewables and/or nuclear energy



https://app.electricitymaps.com/map

Project Planning



- Do you really need it?
- Do you need to train a new model from scratch or can you use a pre-built one?
- If you can relax part of your SLA does that allow you to run your workloads in a greener location?
- **Can you use demand shaping?**

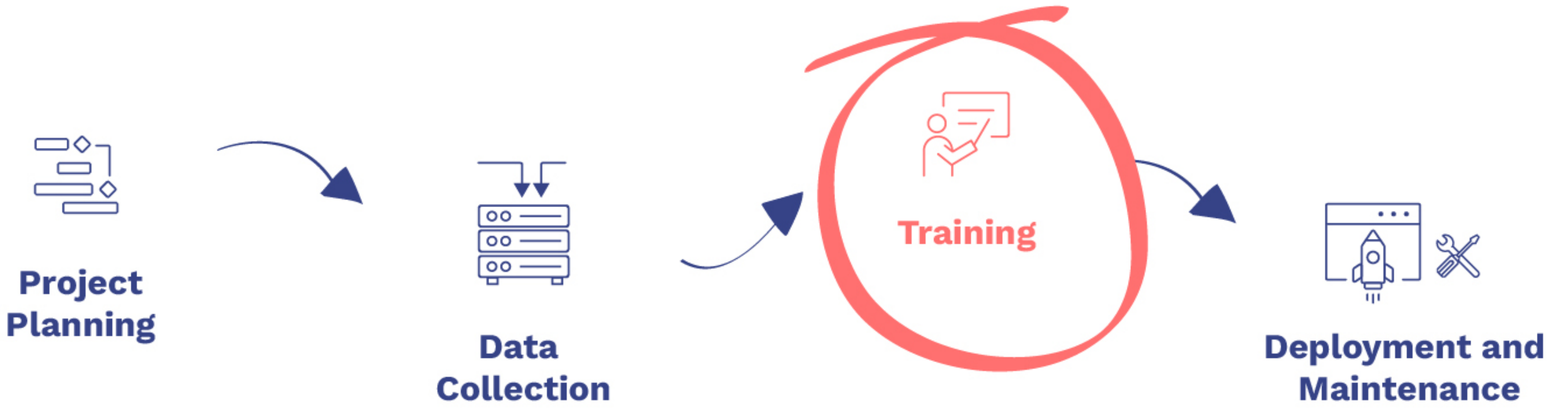
https://www.conissaunce.com/demand-shifting-and-shaping.html

Data Collection



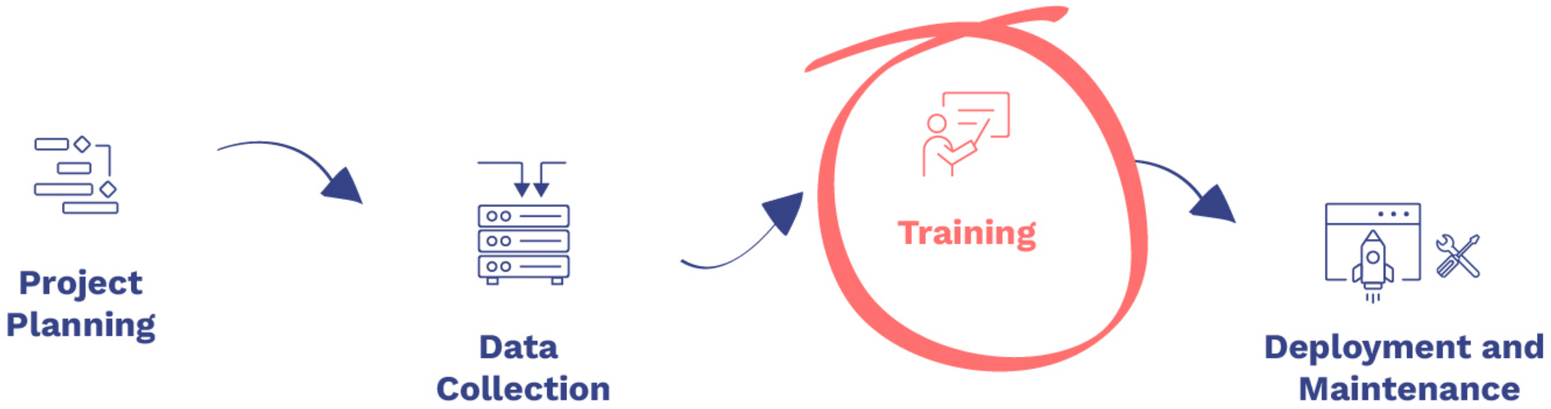
- How much data do you actually need?
- Are there open source data sets you can use? Hugging Face has over 300,000 and Kaggle over 430,000 data sets available
- Does data collection need to happen on demand? If not, consider demand shifting as one way to make use of when and where there is green energy available

Training



- For any work that isn’t particularly latency sensitive, such as training a machine learning (ML) model, it’s smart to do it in a region with lower carbon intensity and at times when you have access to the greenest power

Training



- For any work that isn't particularly latency sensitive, such as training a machine learning (ML) model, it's smart to do it in a region with lower carbon intensity and at times when you have access to the greenest power
- Researchers from University College Dublin have found that practicing time-shifting methodologies for ML models can reduce software-related carbon emissions between 45% and 99%

https://ieeexplore.ieee.org/document/6128960

Training

- Federated learning, despite being slower to converge, can be a greener technology than training centralised in data centres, especially for smaller data sets or less complex models
- Training on the edge could also be greener in some cases

https://thenewstack.io/ai-at-the-edge-federated-learning-for-greater-performance/

Size Matters

- By shrinking the model size, it is possible to speed up training time as well as increase the resource efficiency of training
- Shrinking model sizes is an ongoing research area, with several initiatives exploring topics like pruning, distillation, and quantization as means of compression

Distillation

- The basic process of distillation is that you capture a set of good results using a larger model, then use the stored completions to evaluate the performance of both the larger model and a smaller one to establish a baseline

- Amazon claims that distilled models in Amazon Bedrock are up to 500% faster and up to 75% less expensive than original models

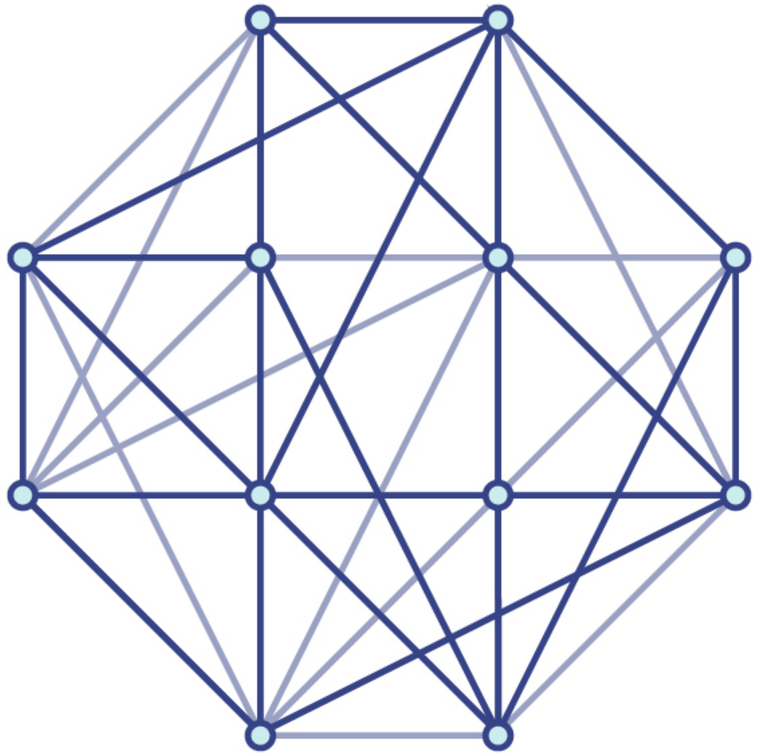
Quantization

- Quantization is the process of reducing the precision of a digital signal, typically from a higher-precision format to a lower-precision format
- Within LLMSs the process can be used to convert weights and activation values of high-precision data, usually 32-bit floating point (FP32) or 16-bit floating point (FP16), to lower-precision data, like 8-bit integer (INT8)
- Google has released AQT for tensor operation quantization in JAX

https://github.com/google/aqt

AI Pruning

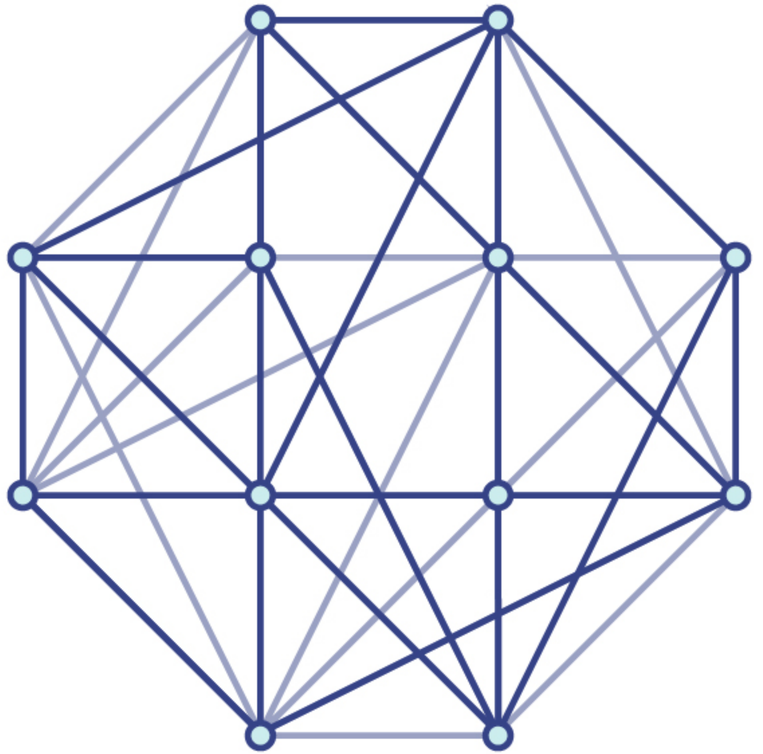
- In mammals a biological process of synaptic pruning takes place in the brain during development



Original Network

AI Pruning

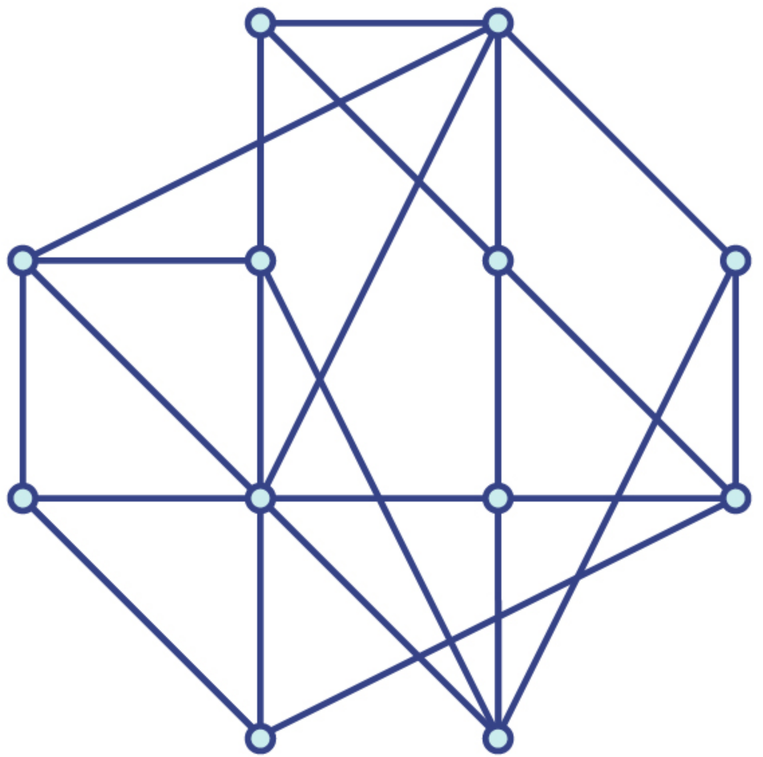
- In mammals a biological process of synaptic pruning takes place in the brain during development
- In AI, pruning is the practice of removing parameters (which may entail removing individual parameters, or parameters in groups such as by neurons) from an existing artificial neural network



Original Network

AI Pruning

- In mammals a biological process of synaptic pruning takes place in the brain during development
- In AI, pruning is the practice of removing parameters (which may entail removing individual parameters, or parameters in groups such as by neurons) from an existing artificial neural network



Pruned Network

Deployment and Maintenance



- For production workloads, deployment and maintenance may very well be where the most carbon is spent
- Quantization, distillation, and pruning can all be applied post-training to decrease the size of the model used when inferencing
- Another interesting technique is speculative decoding

Speculative Decoding

- Speculative decoding works in a similar manner to branch prediction in modern pipelined CPUs
- The goal is to increase concurrency by computing several tokens in parallel
- The technique can reduce the inference times for LLMs significantly

https://research.google/blog/looking-back-at-speculative-decoding/

Speculative Decoding

- Speculative decoding works in a similar manner to branch prediction in modern pipelined CPUs
- The goal is to increase concurrency by computing several tokens in parallel
- The technique can reduce the inference times for LLMs significantly
- Google recently released a set of tools for Gemma called Multi-Token Prediction (MTP) Drafters, which use a specialised speculative decoding architecture. The company claims they will deliver “up to a 3x speedup without any degradation in output quality or reasoning logic”.

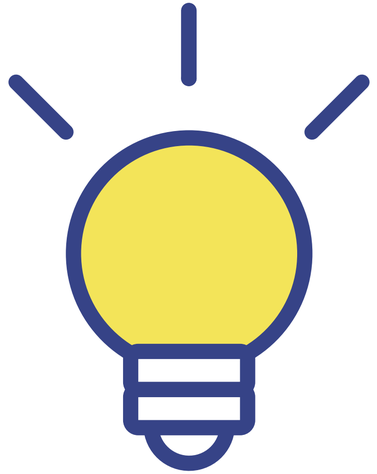
https://blog.google/innovation-and-ai/technology/developers-tools/multi-token-prediction-gemma-4/

Other Tactics

- Semantic caching is widely used with LLMs to speed up response times and reduce API cost
- Knowledge base and prompt compression tools can be used to aid token minimisation - see for example Headroom and Kompress Base

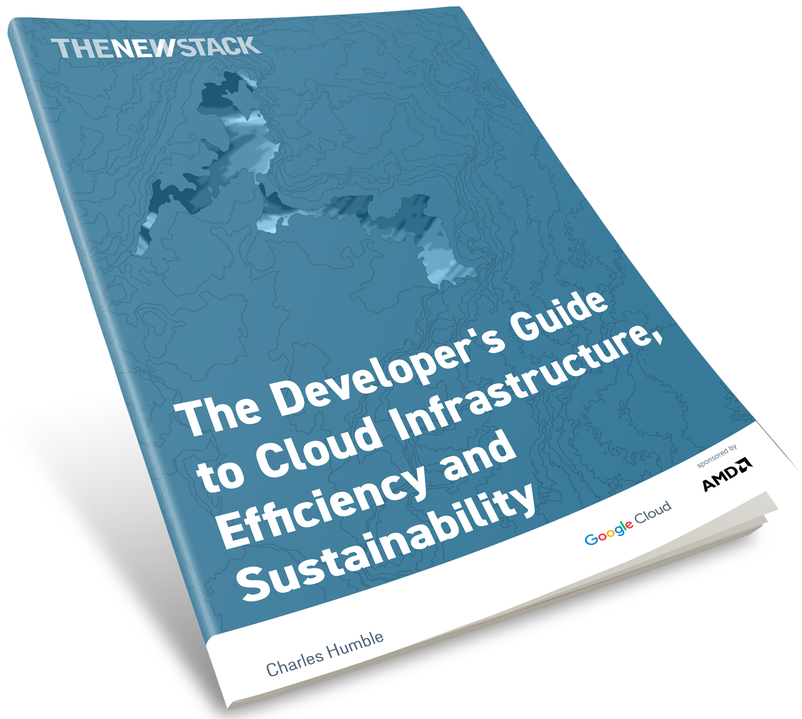
4 Steps to Make All Computing Greener

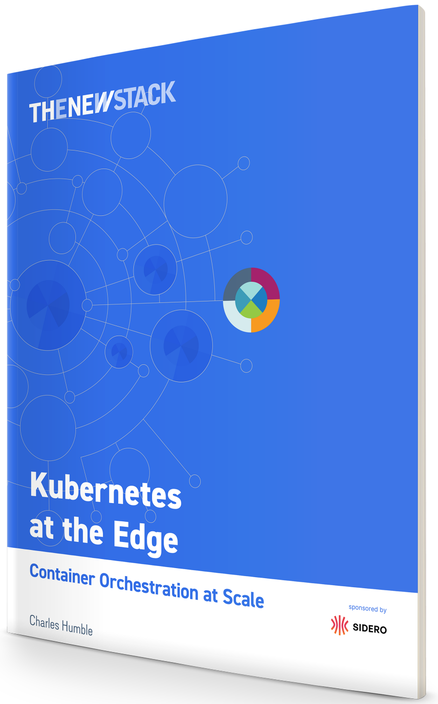
- Use the smallest hardware configuration that can safely execute the job
- Run compute in areas where low carbon electricity is abundant and where there are credible plans to make the grid even cleaner
- Use cloud services from cloud providers that have data centres in green locations and provide good tooling to help reduce your footprint
- Optimise the execution time of jobs to further reduce the footprint

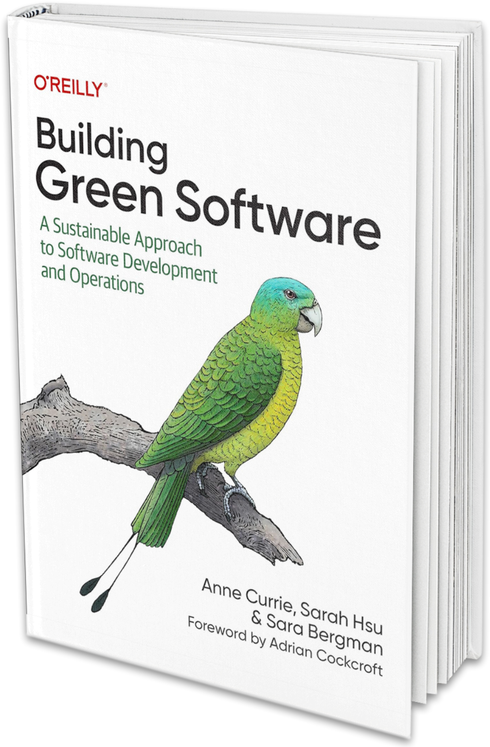


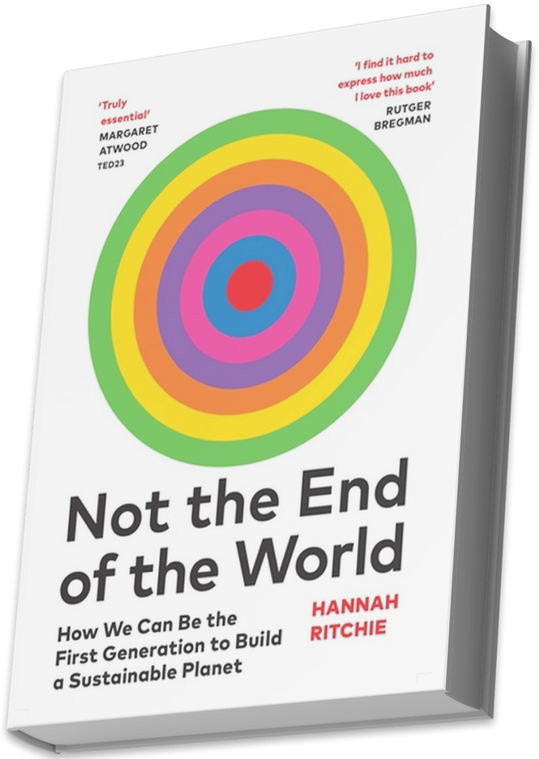
I believe *that sustainability should join cost, performance, security, regulatory concerns, and reliability as one of the top-level considerations for your computing workloads*

Recommended Reading









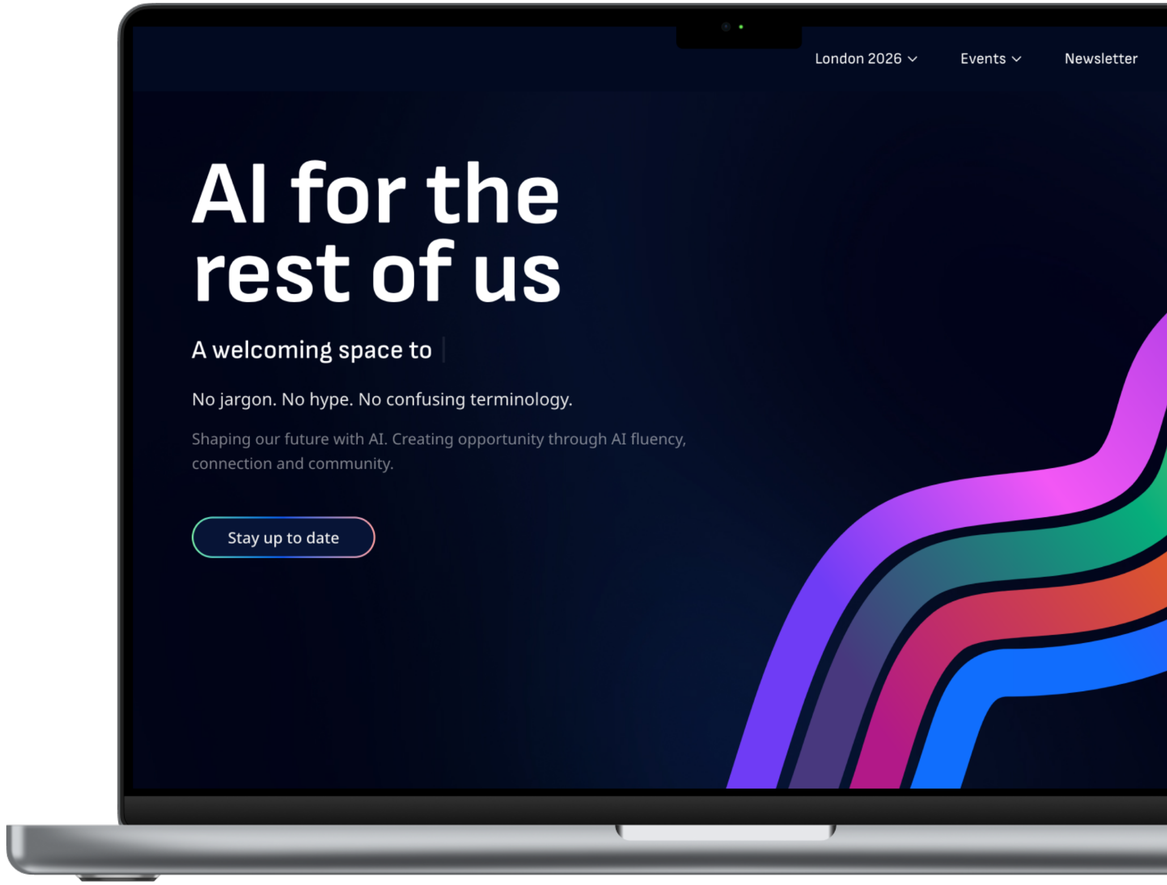
AI for the rest of us

Every Sunday @hannah-foxwell and I publish the newsletter

AI for the rest of us

Real insights, zero jargon

Sign up for
the newsletter



Thanks for listening

Charles Humble, Conissaunce Limited
@charleshumble



GRAB SLIDE DECK HERE