Green NLP

Fixing our environmental impact

Leon Derczynski

GPT3

Power consumption

- Using V100S PCIe GPU hardware
- Tensor performance of 130 TFLOPS
- Microsoft DC uses 12.5% power on cooling
- Compute time is 27955.84 days
- Power usage is 188701.92 kWh
- I.e. 0.19 gigawatt-hours
    - 20 minutes of a Barsebäck reactor
    - **23 200 kg coal burned**
    - 84 738 kg CO2 equivalent

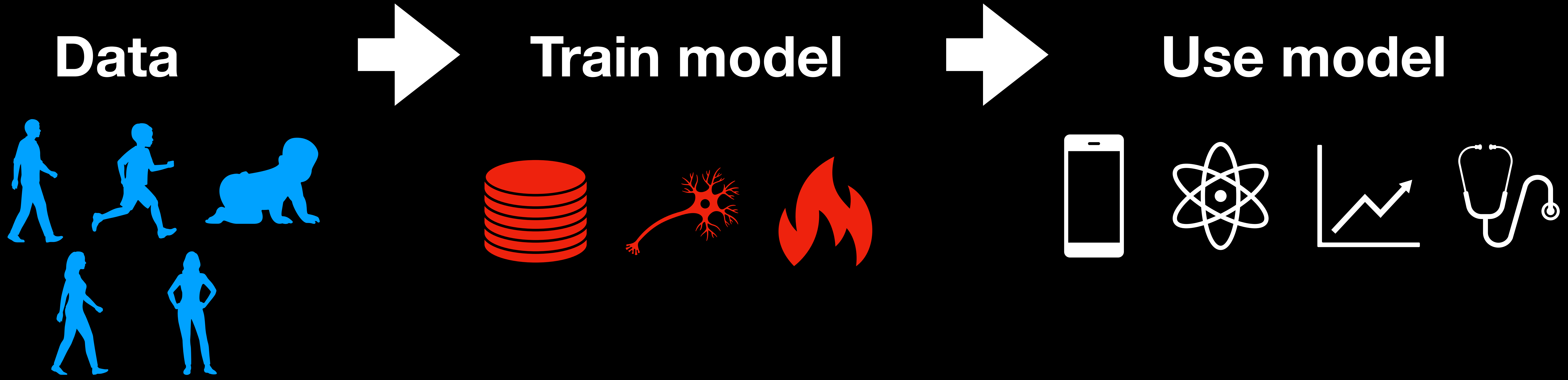
"Tack…, ni svenska vakttorn. Med plutonium…, tvingar vi danskarna ner på knä.”

Source: https://arxiv.org/pdf/2007.03051.pdf,

Energy is expensive

- All energy use warms the earth
    - Thermodynamics is a bitch
- Most electrical energy generation generates CO_2
    - Direct: burning makes CO_2
    - Indirect: concrete makes CO_2
    - Implicit: wind backup is often fossil
- Even if you like to 💩 on your doorstep:
    - Energy use means higher cost

Machine learning paradigm



Machine learning paradigm

Data: energy costs

- Who put the data together?
- How long did that take?
- Where & how is it hosted?
- How much does it cost to store?
- How much energy does it take to load?

Per capita tons CO2e / year:
India - 1.9

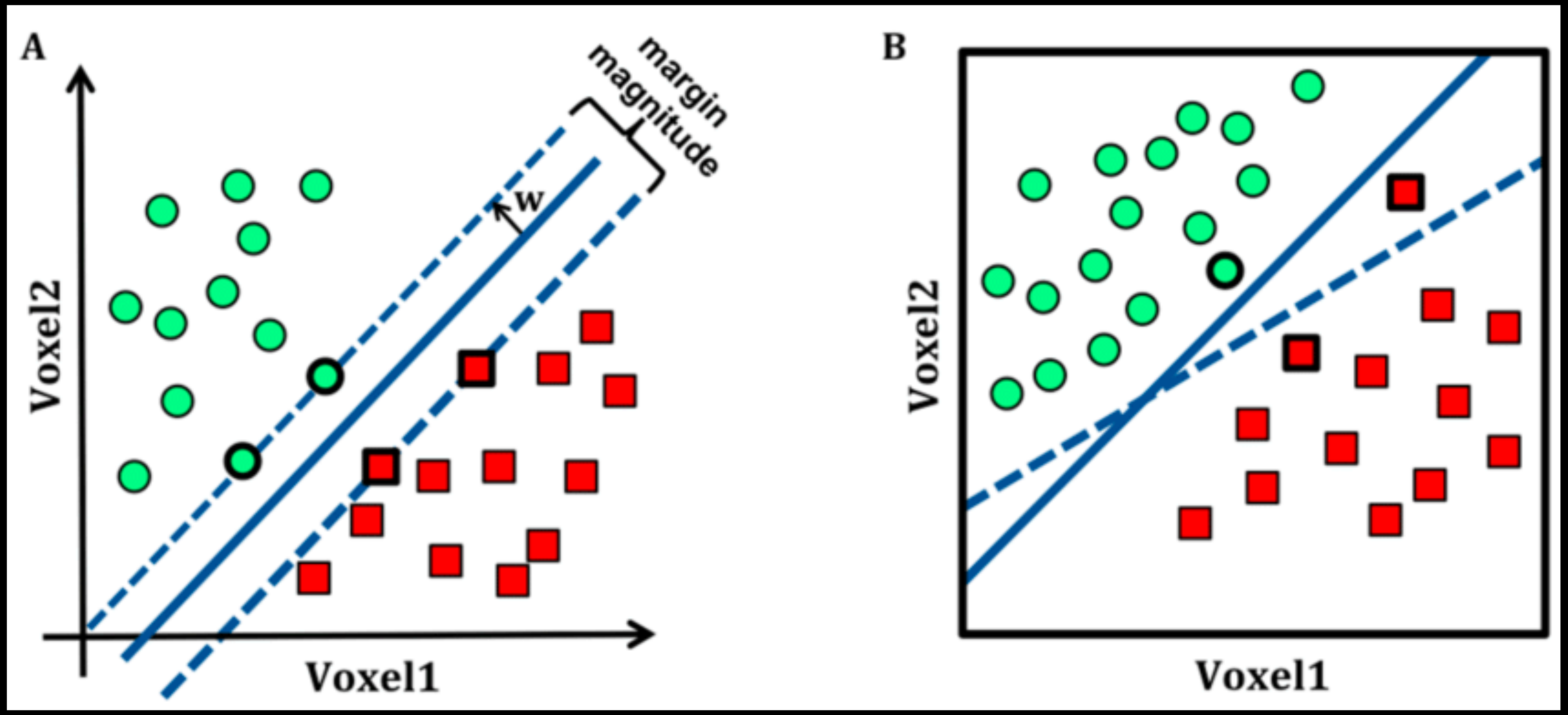
Per capita tons CO2e / year:
Denmark - 7.2

Per capita tons CO2e / year:
USA - 16.1

Making the data green

Solutions & trade-offs

- Select the data points carefully
- Repeating the same thing isn't useful!
- Points close to a decision boundary are the most useful



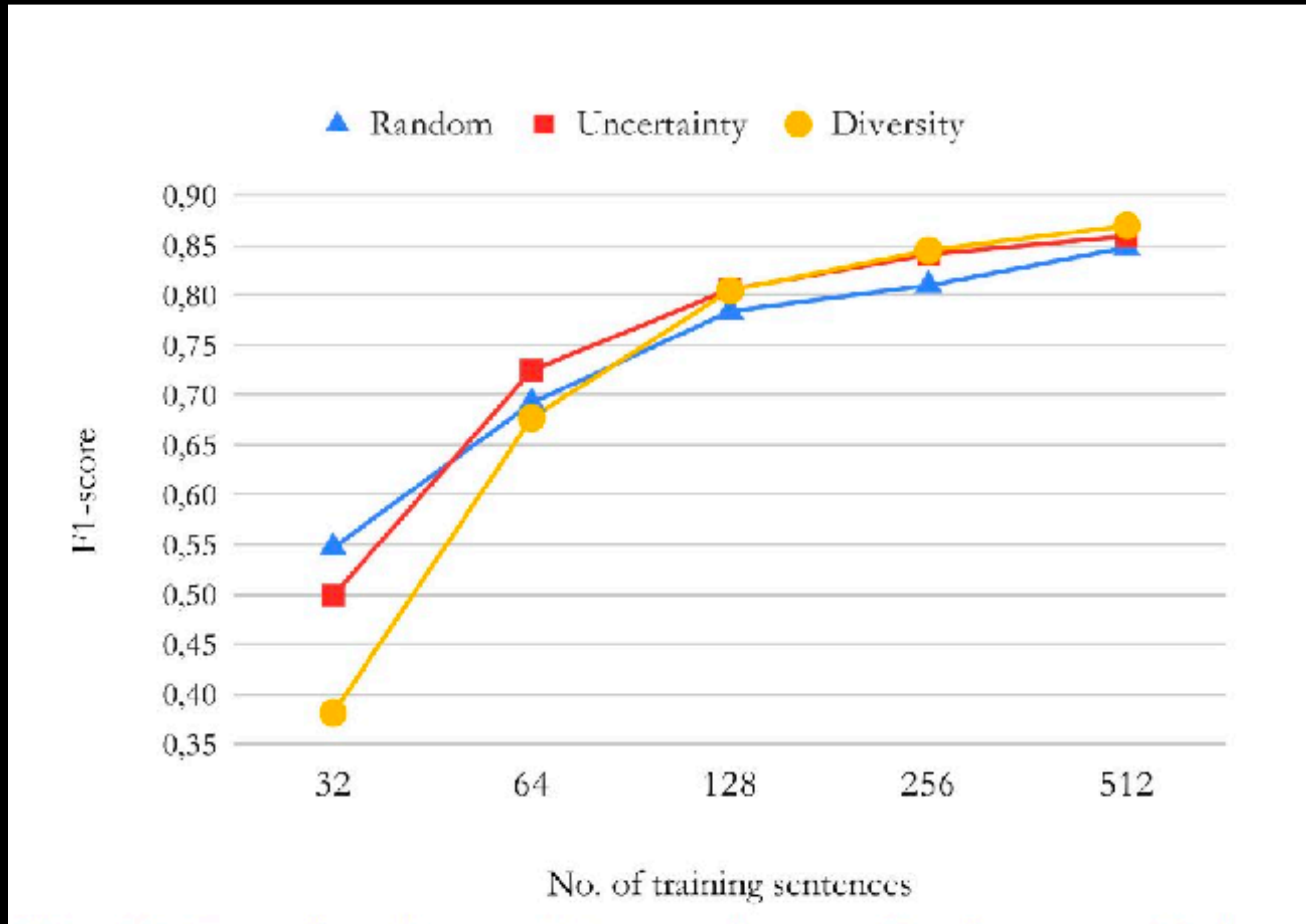


Figure 15: F1-score for each query strategy according to number of sentences trained on

Wonsild & Møller 2020

- Selecting which data to train from: "active learning"
- Impact: less data; fewer cycles

Making the model green

Integer maths

- Integers: whole numbers
- Floats: all numbers
- A 3GHz CPU does 3 billion steps per second
  - But limited capacity for floating point calculations!

- IMUL (integer multiplication)
  - 1 standard operation
- FMUL (floating point multiplication)
  - 4 floating-point operations

☆ Low resolution is best resolution ☆

Activation functions

ReLU vs tanh

```
relu: push eax
      rol eax, 1
      xor eax, eax
      and eax, 1
      pop ebx
      imul eax, ebx
      ret
```

(a) ReLU in x86-like code, with EAX holding a 32-bit float on entry. No floating point stack required; the function is applied bitwise with no branching. Grey instructions take one micro-op. Timings from Fog (2019).

```
tanh: fst dword [tmp1]
      call exp
      fst dword [tmp2]
      fld dword [tmp1]
      fchs
      call exp
      fst dword [tmp1]
      fld dword [tmp2]
      fsubr
      fld dword [tmp2]
      fld dword [tmp1]
      fadd
      fdiv
      ret
exp:  fldl2e
      fmulp st1,st0
      fld1
      fscale
      fxch
      fld1
      fxch
      fprem
      f2xm1
      faddp st1,st0
      fmulp st1,st0
      ret
```

(b) tanh in x86-like code; floating-point operations here begin 'f', which need FPUs and have higher execution times. Red instructions take more than ten micro-ops.

Figure 1: x86 style versions of ReLU vs. tanh.

Source: Derczynski 2020. https://arxiv.org/pdf/2006.07237v1.pdf

tl;dr: fancy maths is super slow. shocking but true

note - the `call` commands are branches. So tanh is even slower than it looks

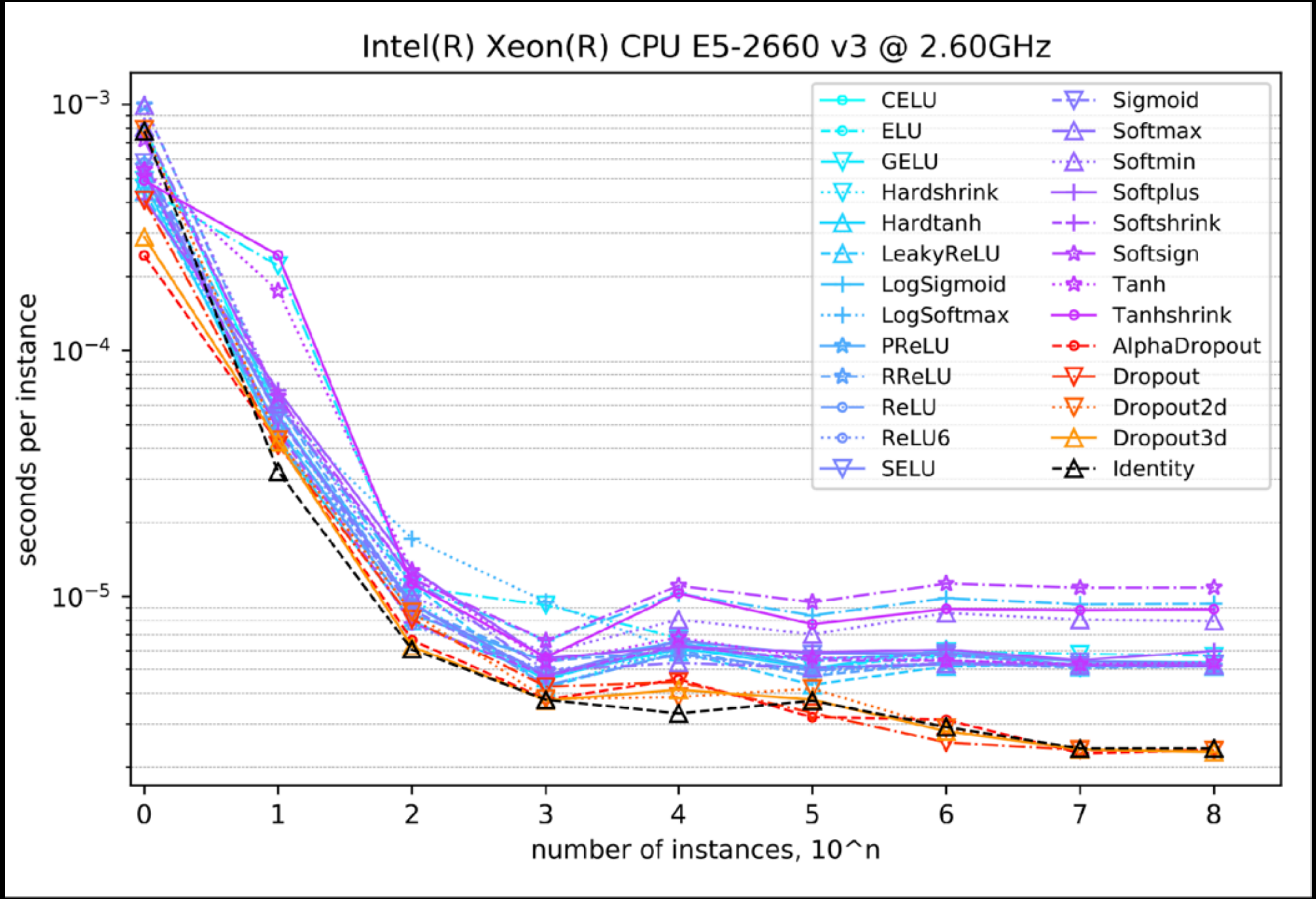
Activation functions

Does it make a difference?

Yes, plenty of difference

The slowest function takes around 8x as long as the fastest

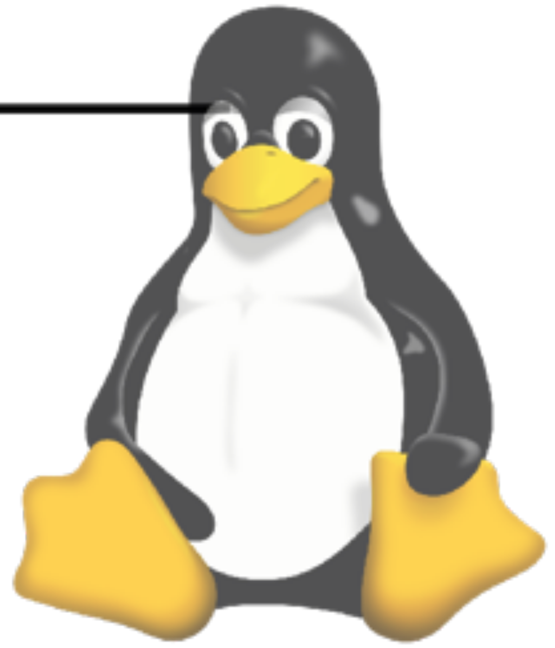
Why burn that energy?



Data Driven Climate Action

Tomorrow builds tech that empowers people and organisations to understand and reduce their carbon footprint.

Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models



Lasse F. Wolff Anthony* [1] **Benjamin Kanding*** [1] **Raghavendra Selvan** [1]

Abstract

Deep learning (DL) can achieve impressive results across a wide variety of tasks, but this often comes at the cost of training models for extensive periods on specialized hardware accelerators. This energy-intensive workload has seen immense growth in recent years. Machine learning (ML) may become a significant contributor to climate change if this exponential trend continues. If practitioners are aware of their energy and carbon footprint, then they may actively take steps to reduce it whenever possible. In this work, we

ized hardware accelerators such as graphics processing units (GPUs). From 2012 to 2018 the compute needed for DL grew 300000-fold (Amodei & Hernandez, 2018).

This immense growth in required compute has a high energy demand, which in turn increases the demand for energy production. In 2010 energy production was responsible for approximately 35% of total anthropogenic greenhouse gas (GHG) emissions (Bruckner et al., 2014). Should this exponential trend in DL compute continue then machine learning (ML) may become a significant contributor to climate change.

This can be mitigated by exploring how to improve energy efficiency in DL. Moreover, if practitioners are

Carbon intensity

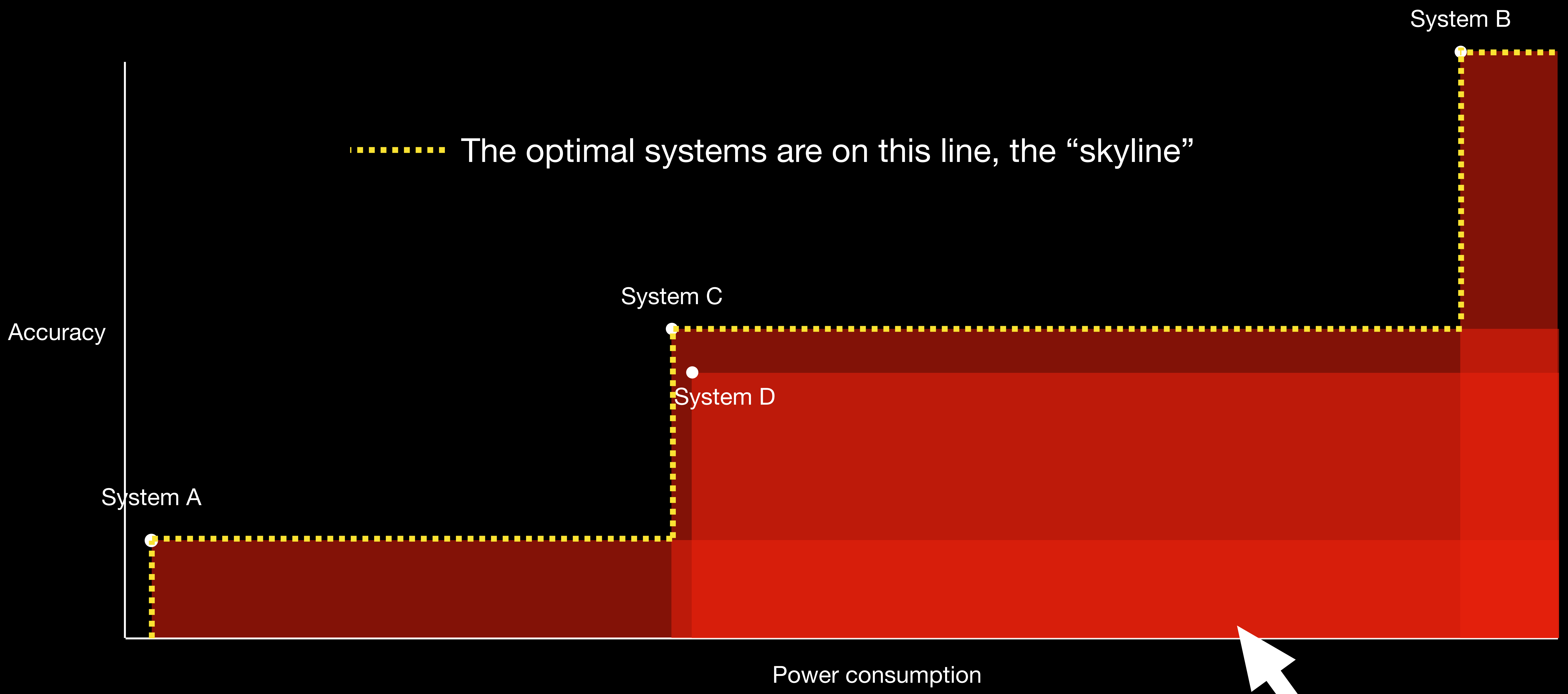
How can you minimise this?

Best place for serious data science is in a datacenter (e.g. via Colab)

1. Microsoft Azure: Carbon neutral since 2012, leaders in good PUE
2. Google Cloud: 100% renewable (this doesn’t mean zero-CO2e)
3. Amazon AWS: committed to catch up with Google by 2030; currently has reached 50%

Coal is cheap partly because nobody wants it

Mapping power : accuracy tradeoff



Pareto / Skyline curve

Green AI: takeaway points

1. Tiny models are best
2. Some model types consume less power than others
3. Use only the useful training data examples - ignore repeats
4. Run your code at low-carbon times

Thank you!