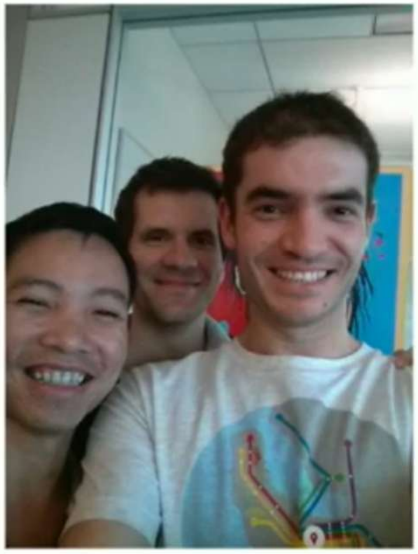
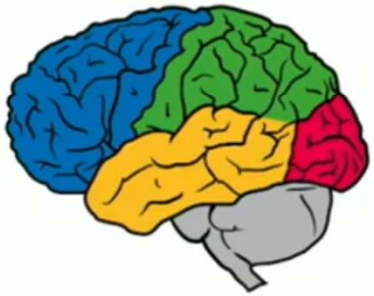
Sequence to sequence learning with neural networks: what a decade 😅

2014

Sequence to Sequence Learning with Neural Networks



Ilya Sutskever
Oriol Vinyals
Quoc Le



Google Brain

2024

What we did

- Autoregressive model trained on text
- Large neural network
- Large dataset

What we got right

“The Deep Learning Hypothesis”

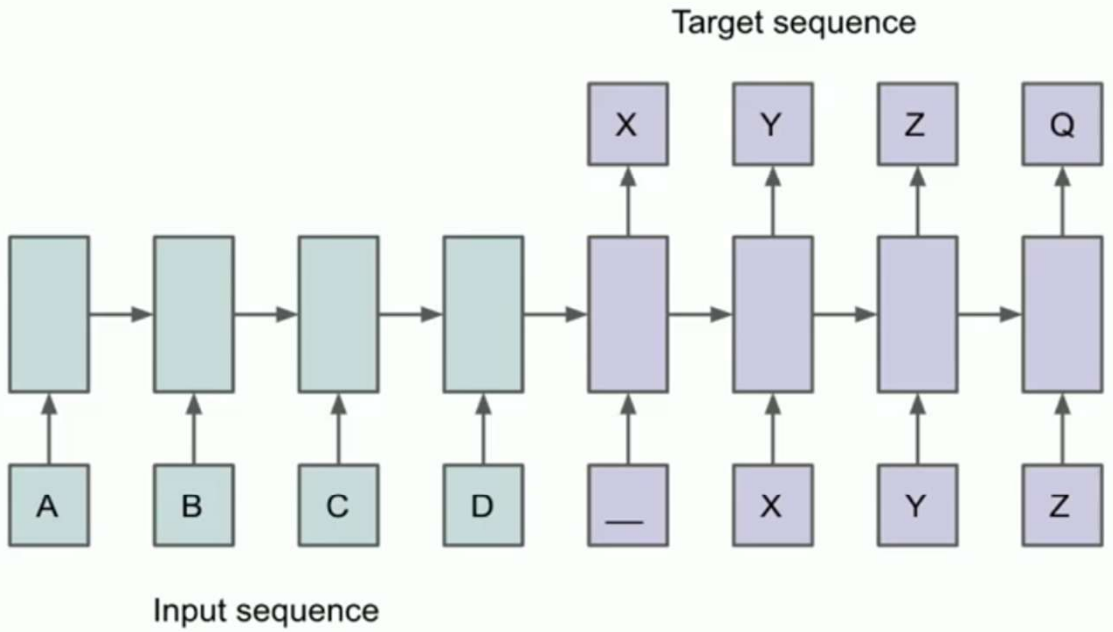
- Human perception is fast
  - Neurons fire at most 100 times a second
  - Humans solve perception in 0.1 seconds

  → our neurons fire 10 times, at most

Anything a human can do in 0.1 seconds, a big 10-layer neural network can do, too!

What we got right: Autoregressive models

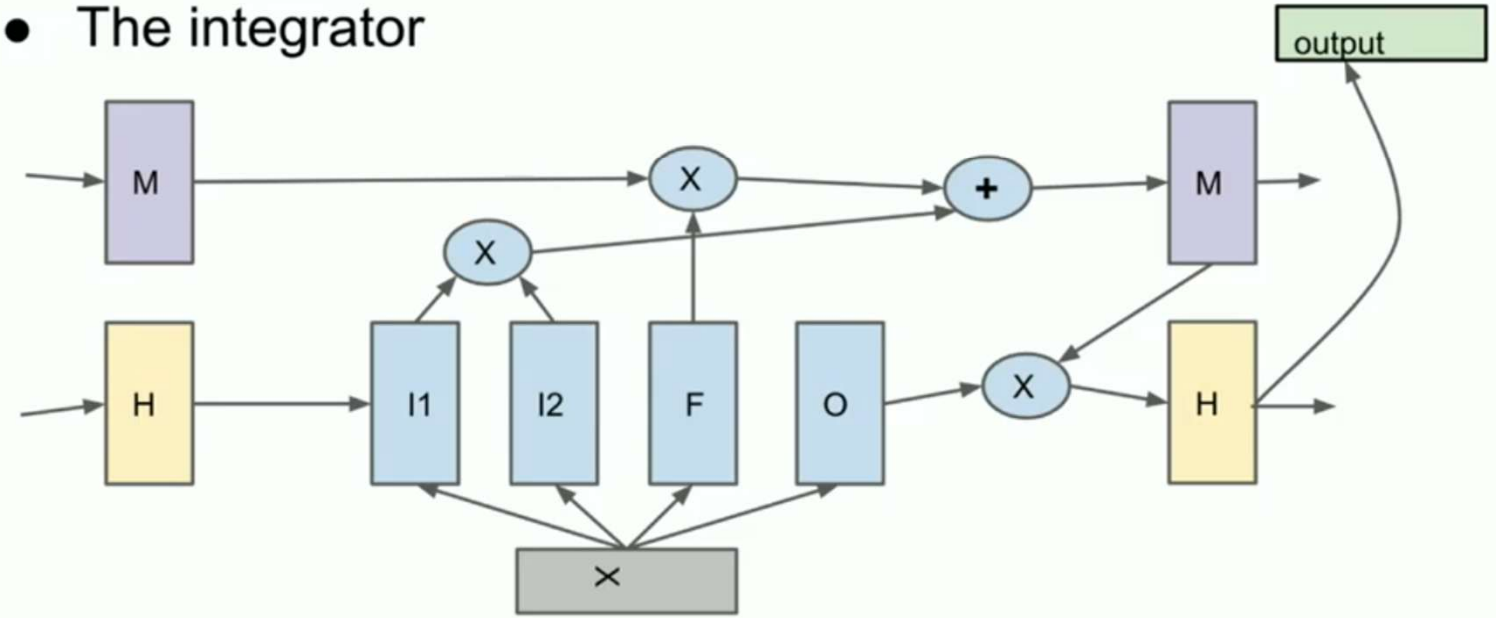
Our main idea



What we got wrong: the LSTM

The heart of the LSTM

- The integrator



Early distributed training

Parallelization

- Use an 8 GPU machine
- One layer per GPU, softmax for remaining GPUs
- **3.5x speedup** over a single GPU
- **8x more RAM**

- Model can be run on a single K40

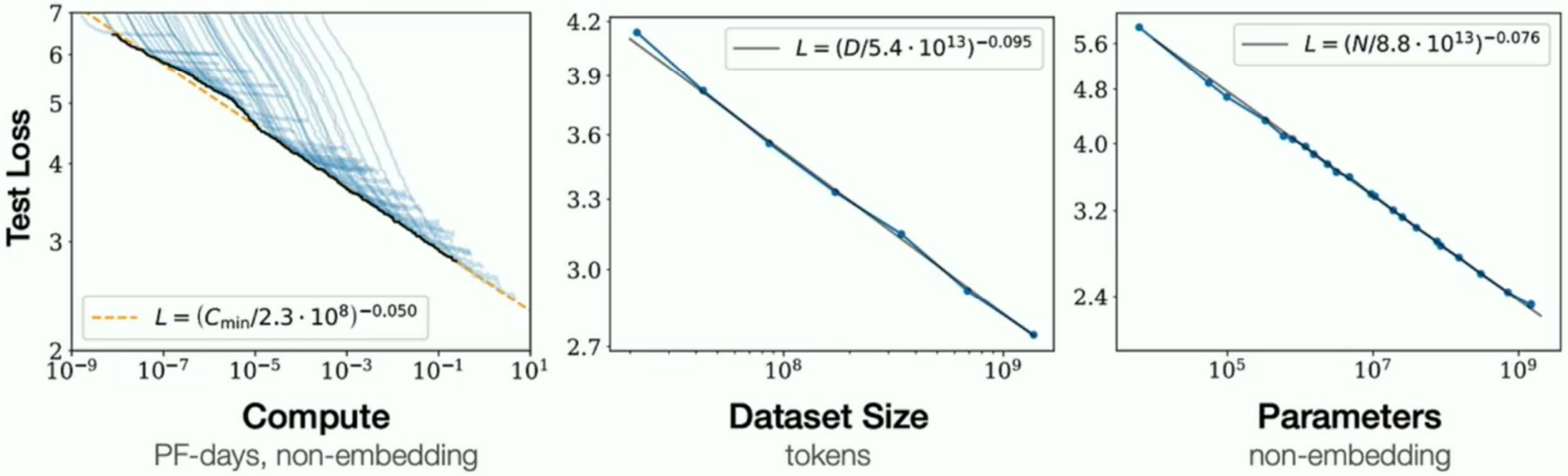
What we got right: early scaling hypothesis

Conclusions

- If you have a large big dataset
- And you train a very big neural network
- Then success is guaranteed!

The age of Pre-Training

- GPT-2 [Radford et al., 2019]
- GPT-3 [Brown et al., 2020]
- Scaling laws [Kaplan et al. 2020]



Pre-training as we know it will end

Compute is growing:

- Better hardware
- Better algorithms
- Larger clusters

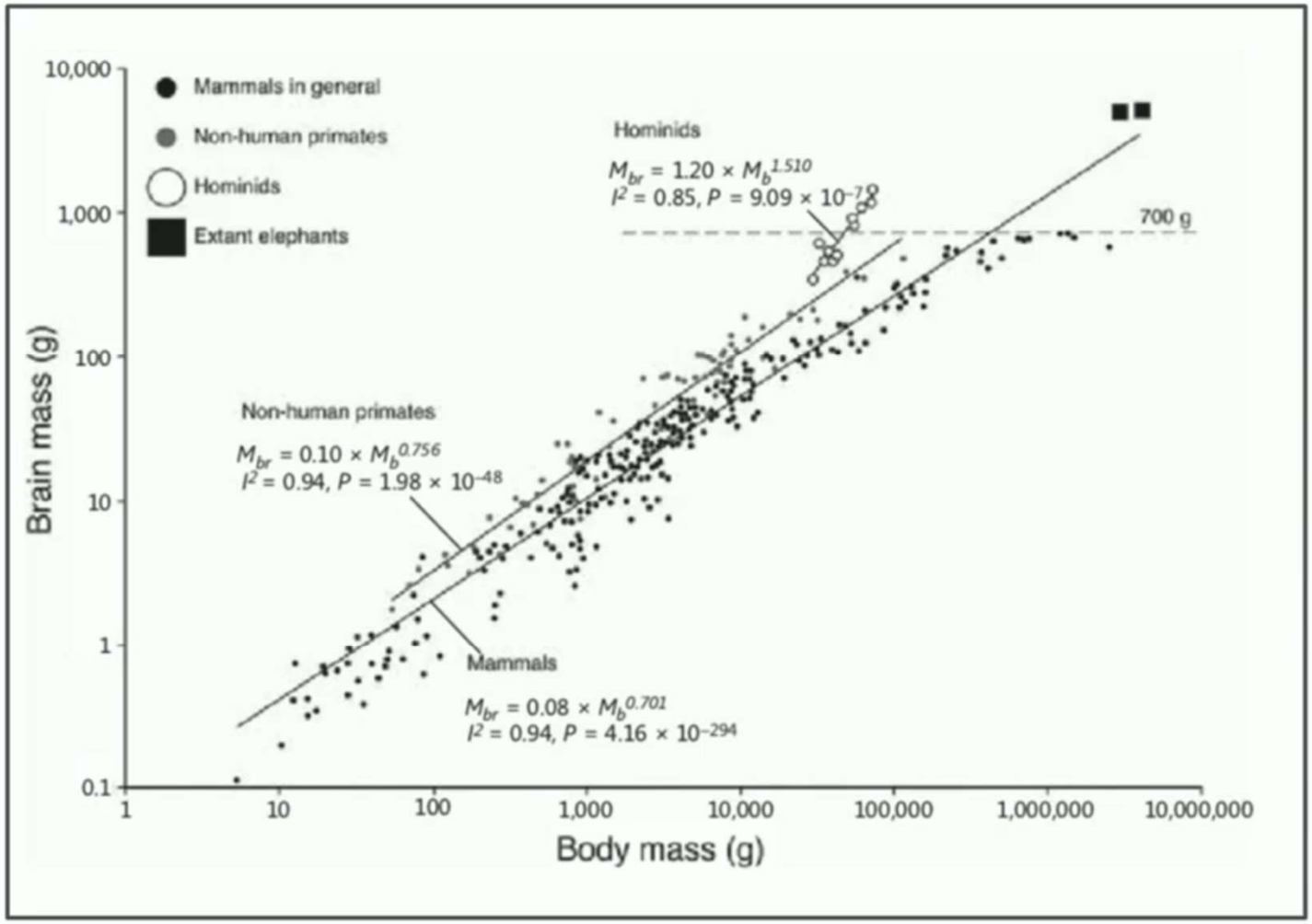
Data is not growing:

- We have but one internet
- **The fossil fuel of AI**

What comes next?

- “Agents”??
- “Synthetic data”
- Inference time compute ~ O1

What comes next? Example from nature



From: The evolutions of large brain size in mammals: the 'over-700-gram club quartet'

[Manger, et al, 2013]

What comes next? The long term

Superintelligence

- Agentic
- Reasons
- Understands
- Is self aware

THE END