An In-Depth Look at Baidu's AI Aspirations

JULIA LI

lixing08@baidu.com

NEWSHA ARDALANI

newsha@baidu.com

Baidu百度

百度一下

apollo

DUEROS

Baidu大脑
Baidu Brain

Baidu百度

AI & HPC

Make communication easier

- Speech Recognition
- Text-to-Speech Synthesis
- Simultaneous Translation
- Language Model

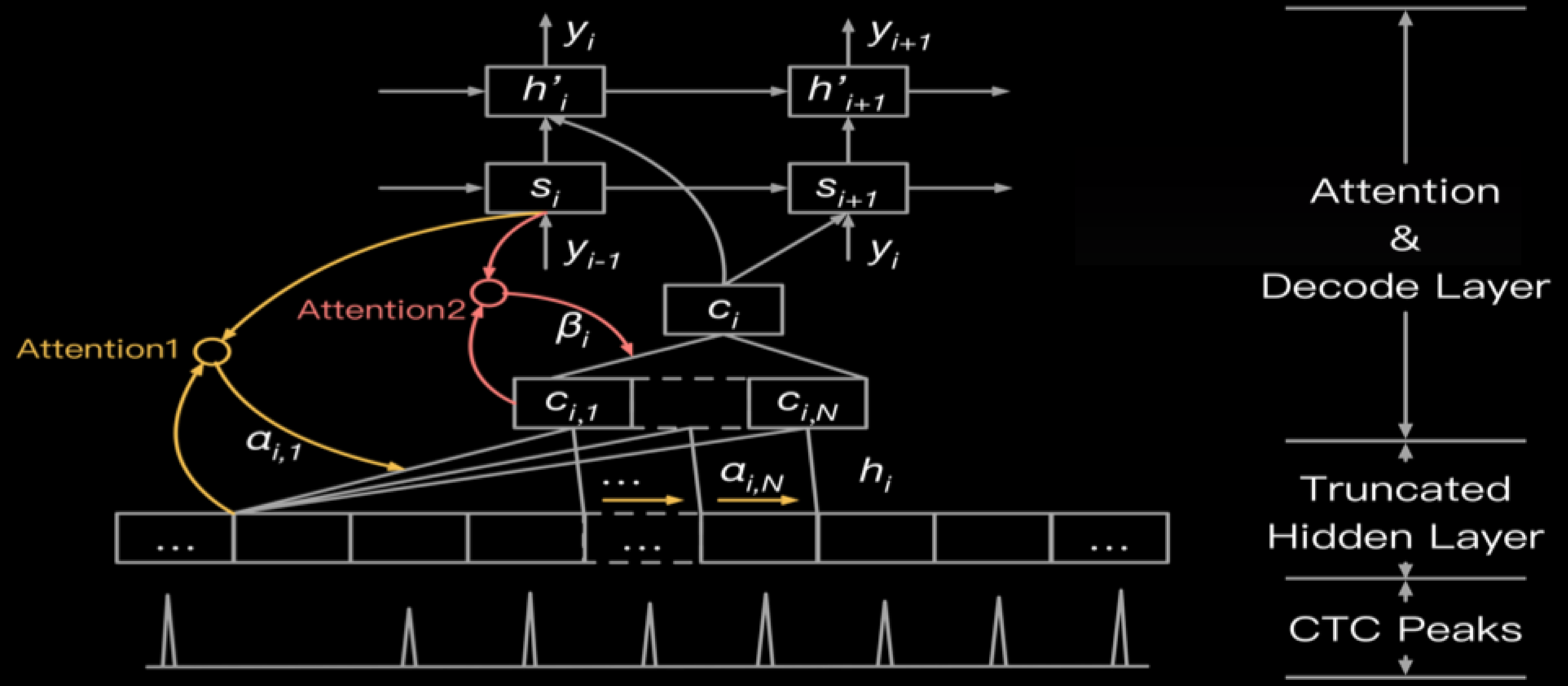
Make AI faster

- High Performance Computing

Speech Recognition Model — SMLTA

Features

- Streaming
- Multi-layer Neural Network
- Large Data - Code Switching



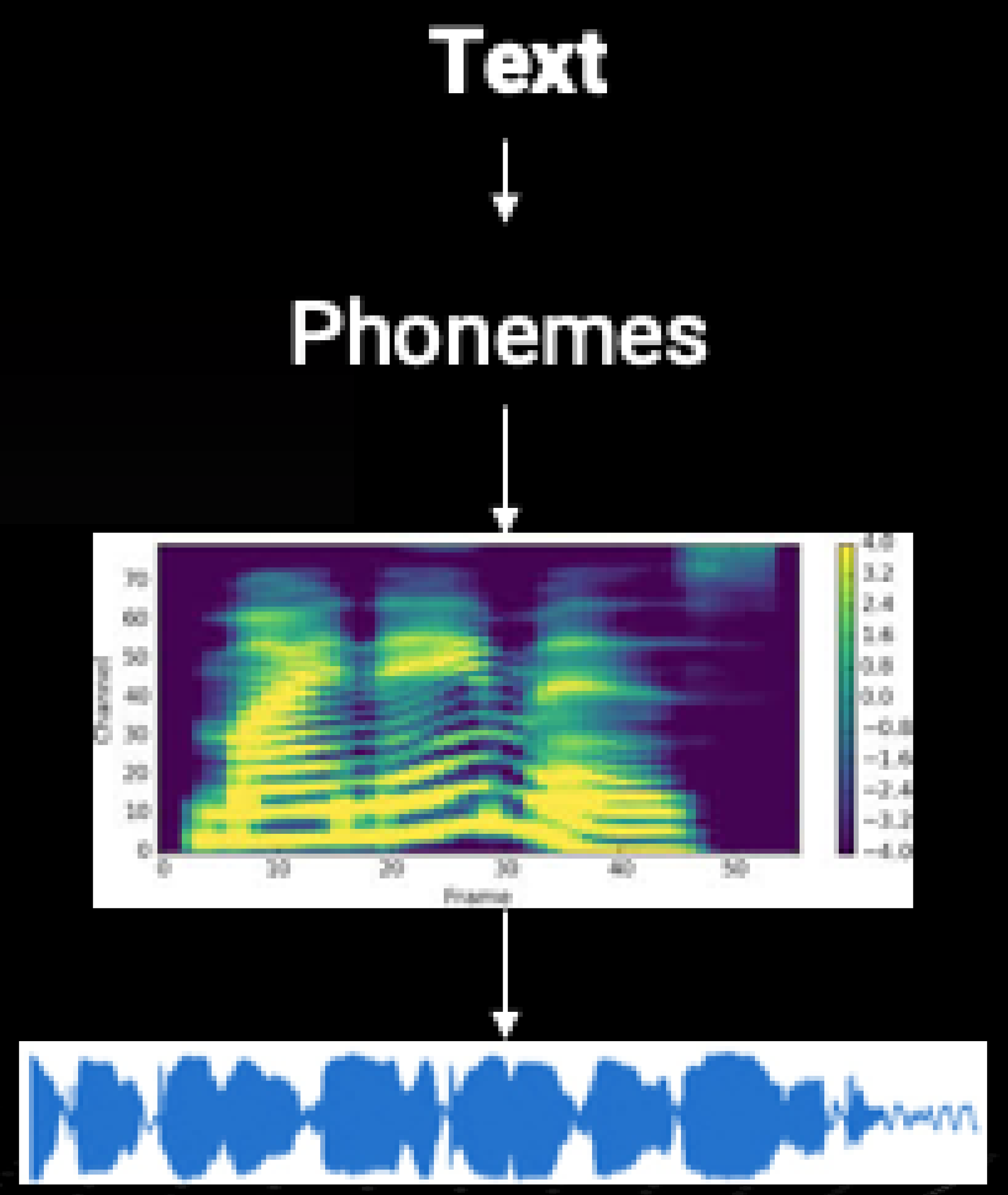
Tech blog: research.baidu.com/Blog/index-view?id=109

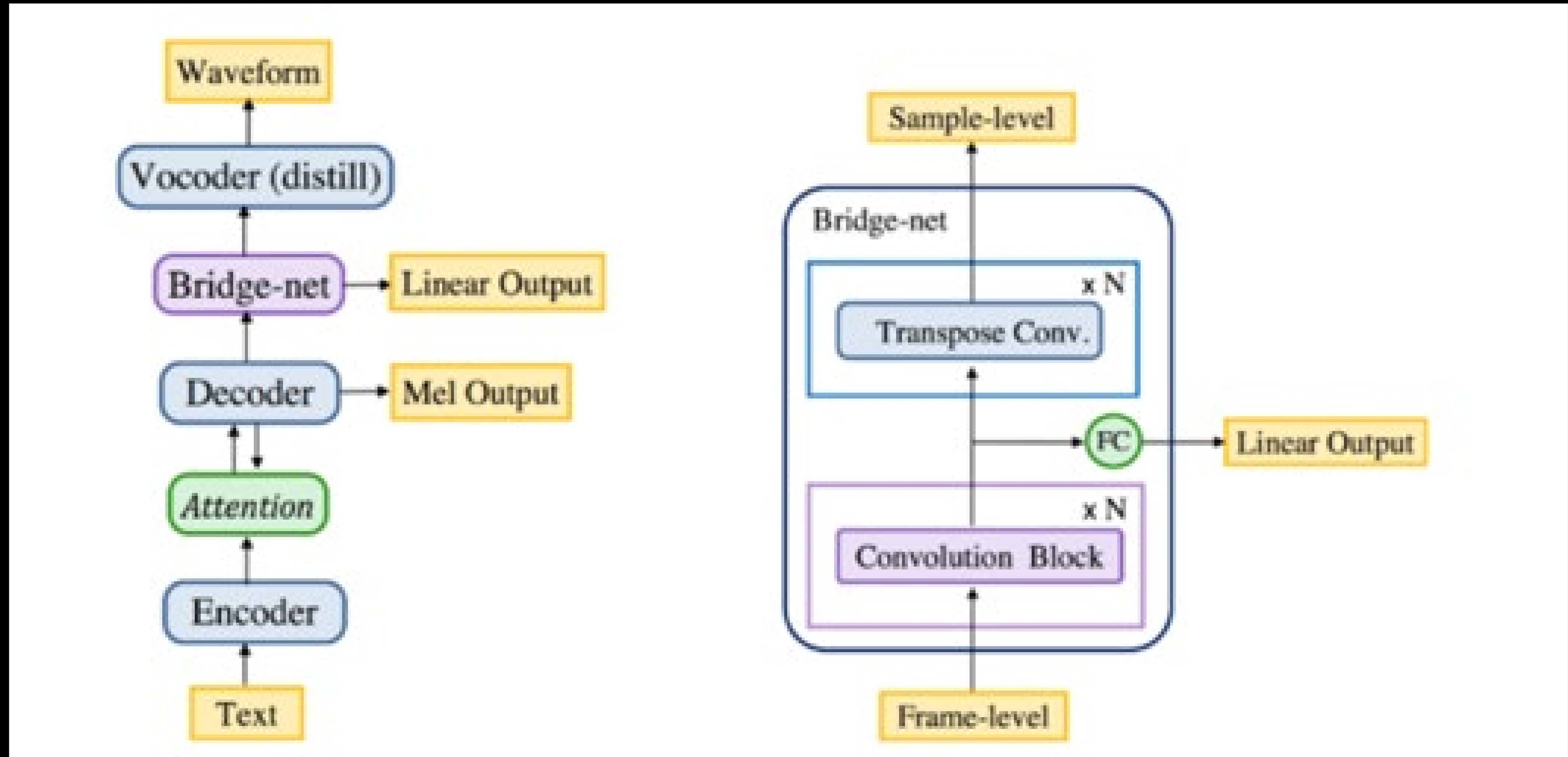
Baidu 百度

Text-to-Speech Synthesis (TTS)

TTS Model — Clarinet

A Fully End-To-End Neural Network Model

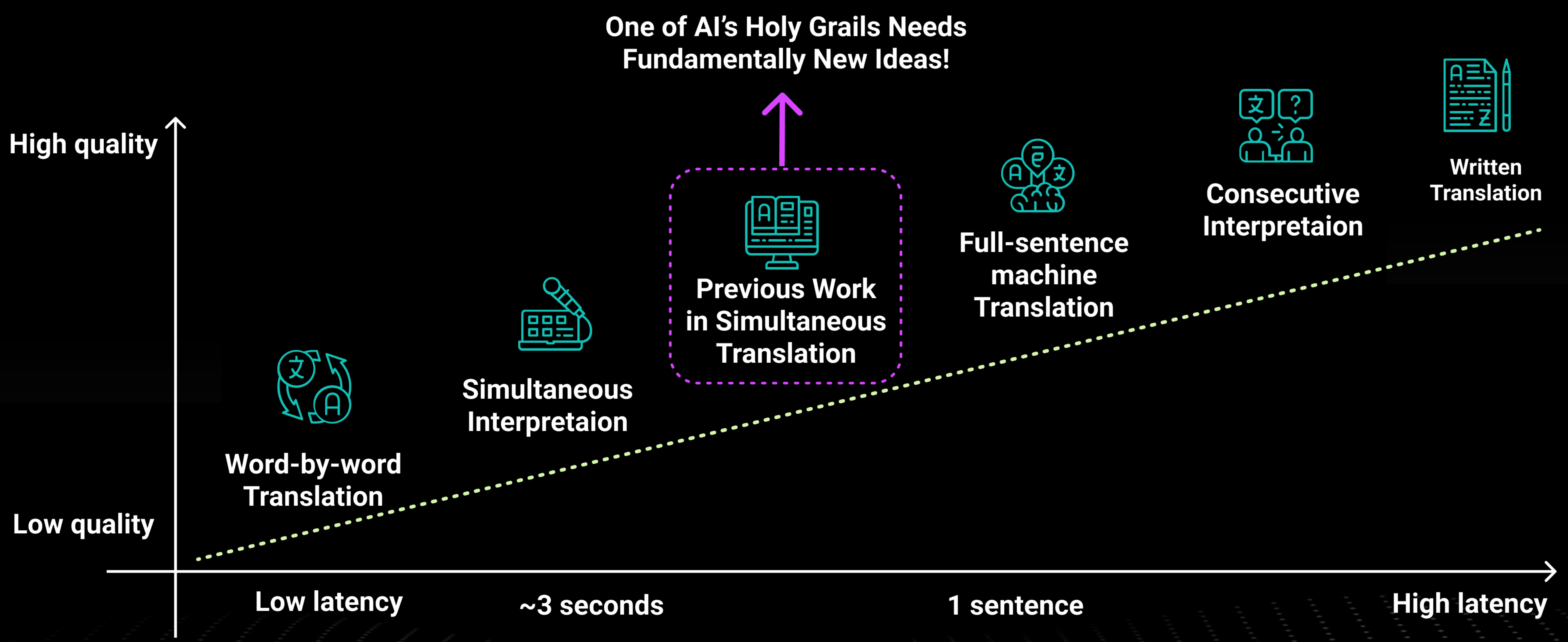




(a) Text-to-wave architecture (b) Bridge-net

Paper: https://arxiv.org/pdf/1807.07281.pdf
Audio Samples: https://clarinet-demo.github.io/

Tradeoff between Latency and Quality

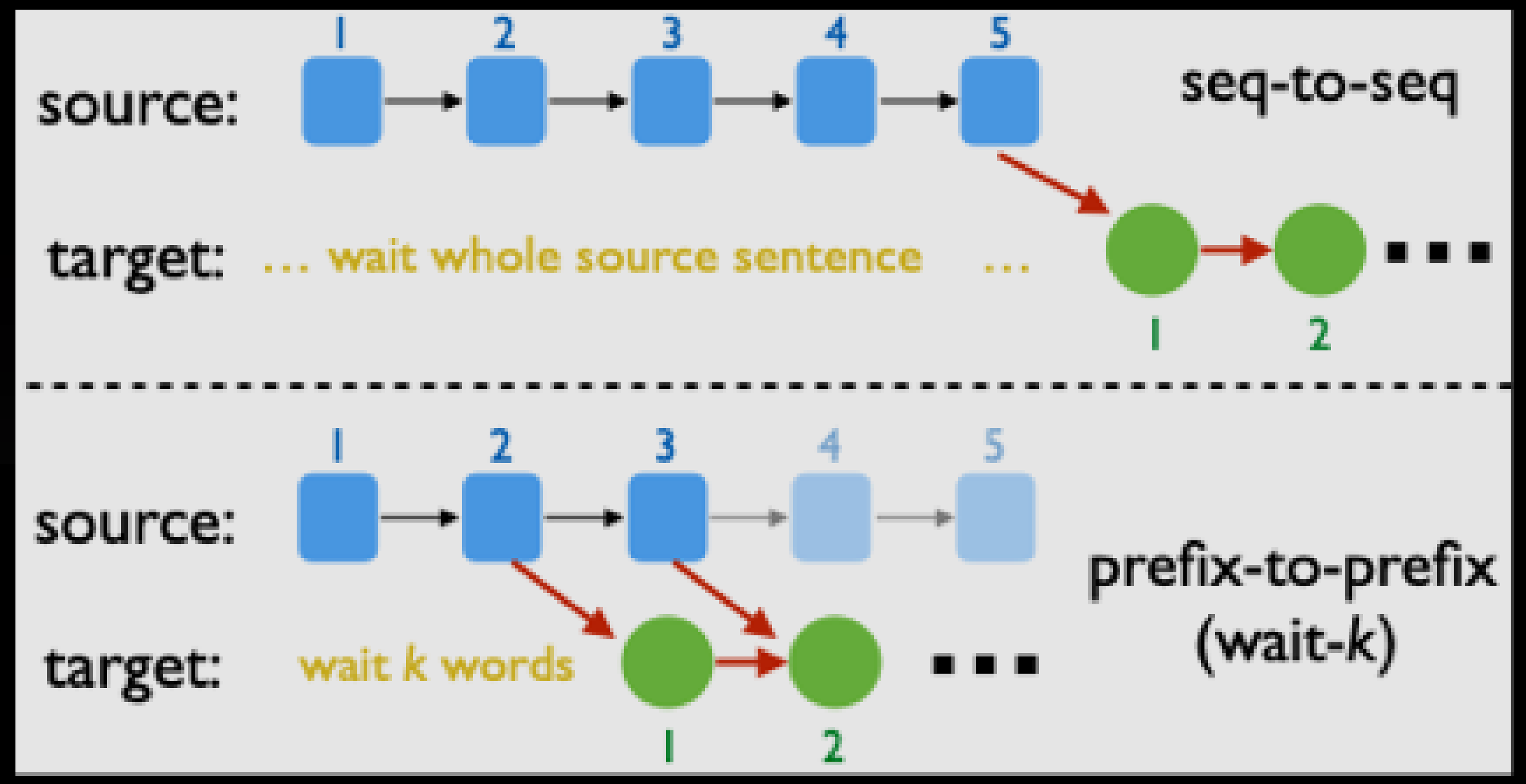


Don't want to be the last one to applause!

Simultaneous Translation Model — STACL

- A prefix-to-prefix framework
- Controlable latency

$p(y_i \mid x_1 \ldots x_n, y_1 \ldots y_{i-1})$



$p(y_i \mid x_1 \ldots x_{i+k-1}, y_1 \ldots y_{i-1})$

Paper: https://arxiv.org/abs/1810.08398

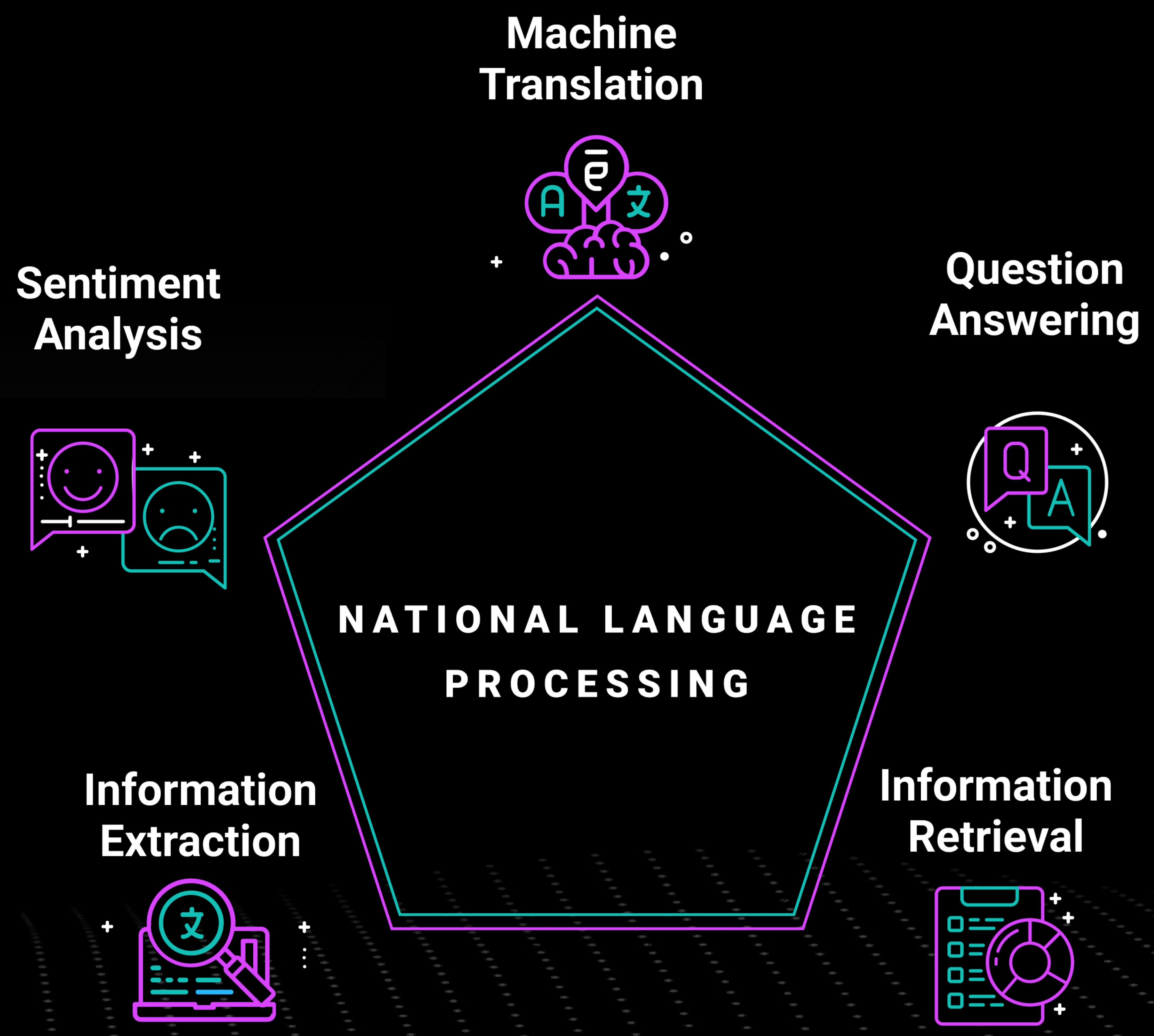
Natural Language Processing

Challenge

- NLP is a diversified field with many distinct tasks
- Shortage of training data

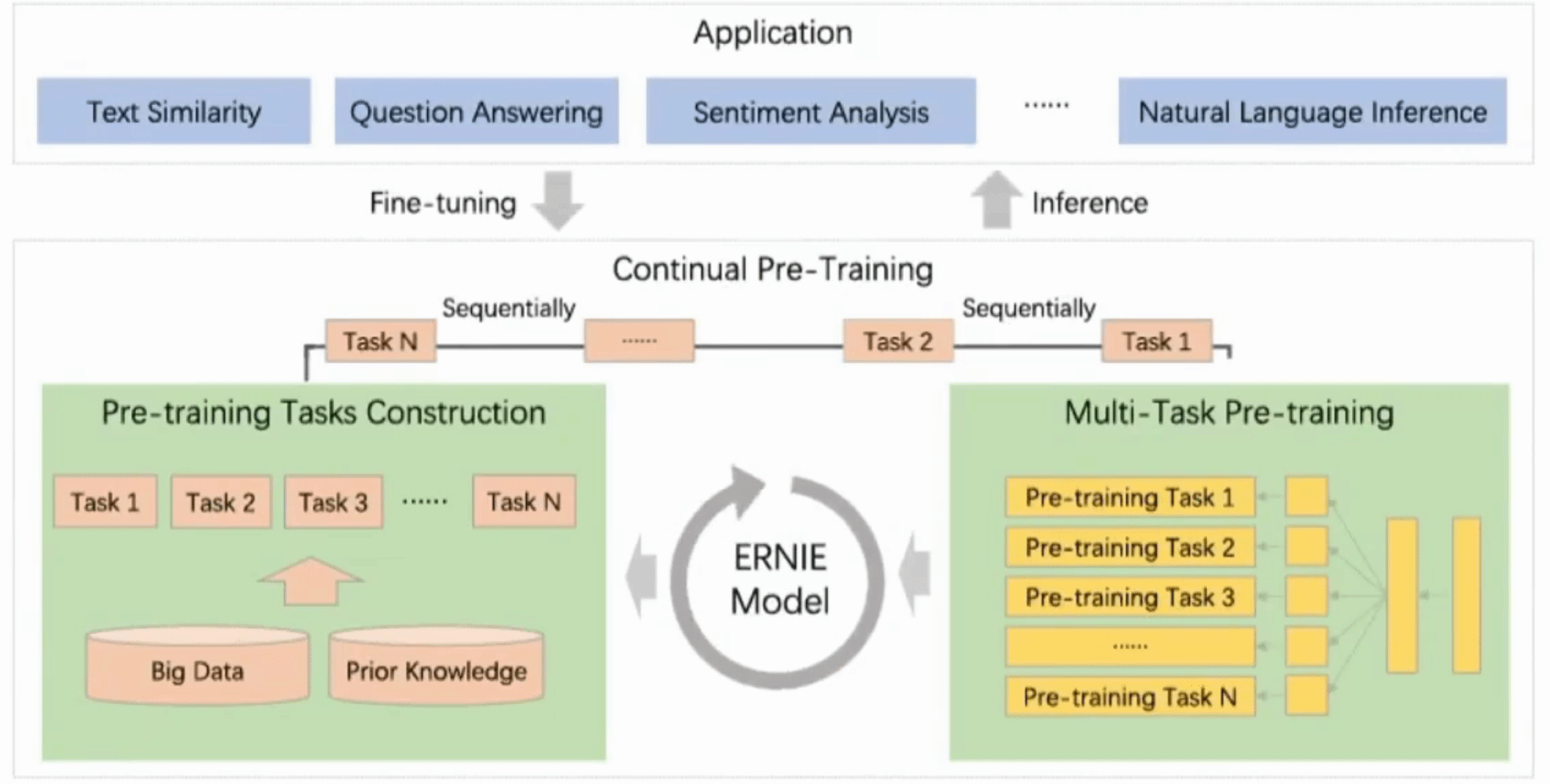
New Trend

- Pre-training + Fine-tuning framework
  - Pre-training(using the enormous amount of unannotated text data)
  - Fine-tuning(using small-data NLP tasks in resulting in substantial accuracy improvements)



Language Model — ERNIE 2.0

- Inspired by BERT
- Incorporate more information
  - Named entities
  - Semantic closeness
  - Sentence order or discourse relations
- Design a continual pre-training framework for language understanding



Paper: https://arxiv.org/abs/1907.12412

Baidu 百度

Bigger Model is Better?

Model	Hidden size	Layer	Parameters
BERT-base	768	12	110M
BERT-large	1024	24	340M
GPT2-large	1024	24	1.5B
Megatron	1024	72	8.3B
T5	E1024 D1024	E24 D24	11B

"BERT was performed on 16 Cloud TPUs (64 TPU chips total). Each pretraining took 4 days to complete".

Paper: https://arxiv.org/pdf/1810.04805.pdf

"T5: TPU pods are multi-rack ML supercomputers that contain 1,024 TPU v3 chips connected via a high-speed 2D mesh interconnect with supporting CPU host machines."

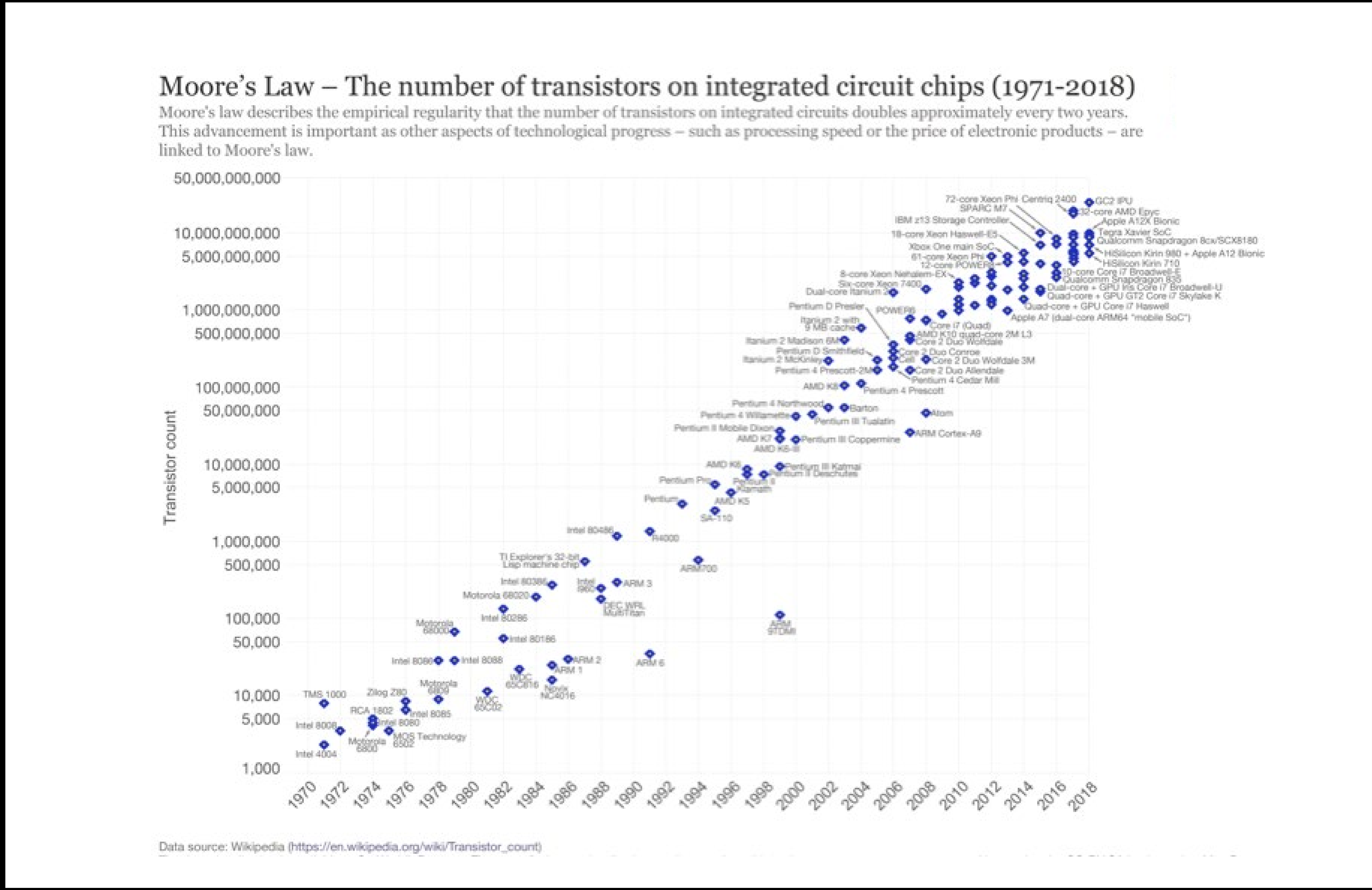
Paper :https://arxiv.org/pdf/1910.10683.pdf.

Bai du 百度

Why Bigger is Better?

What Are the Implications for System Community?

Moore's Law

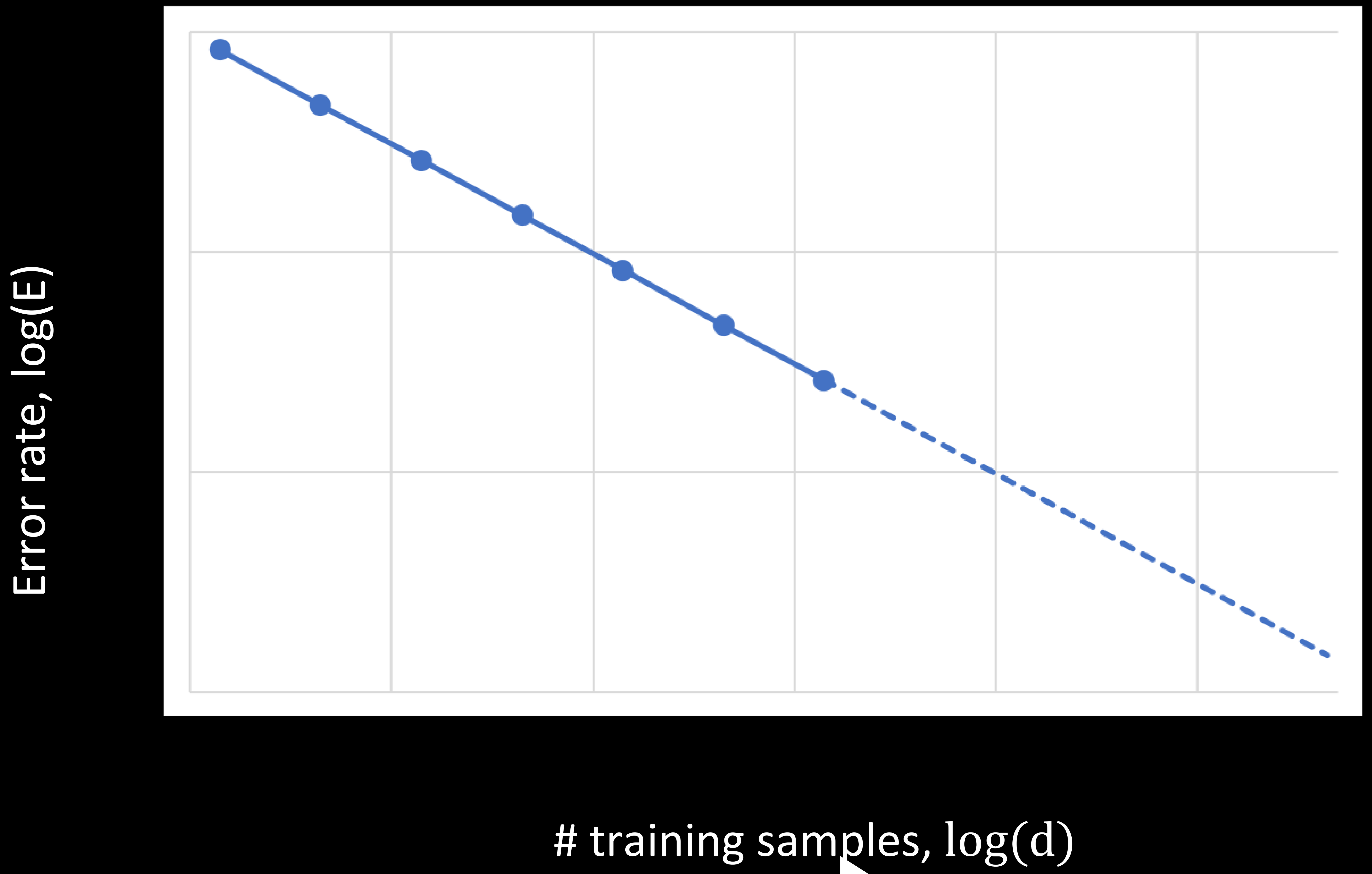

Moore's Law – The number of transistors on integrated circuit chips (1971-2018)
Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years.
This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are linked to Moore's law.
Transistor count
50,000,000,000
10,000,000,000
5,000,000,000
1,000,000,000
500,000,000
100,000,000
50,000,000
10,000,000
5,000,000
1,000,000
500,000
100,000
50,000
10,000
5,000
1,000
1970 1972 1974 1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018
Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)

What does it mean to be “Better”?

- Better Accuracy
- Faster Training

Better as Better Accuracy

Better

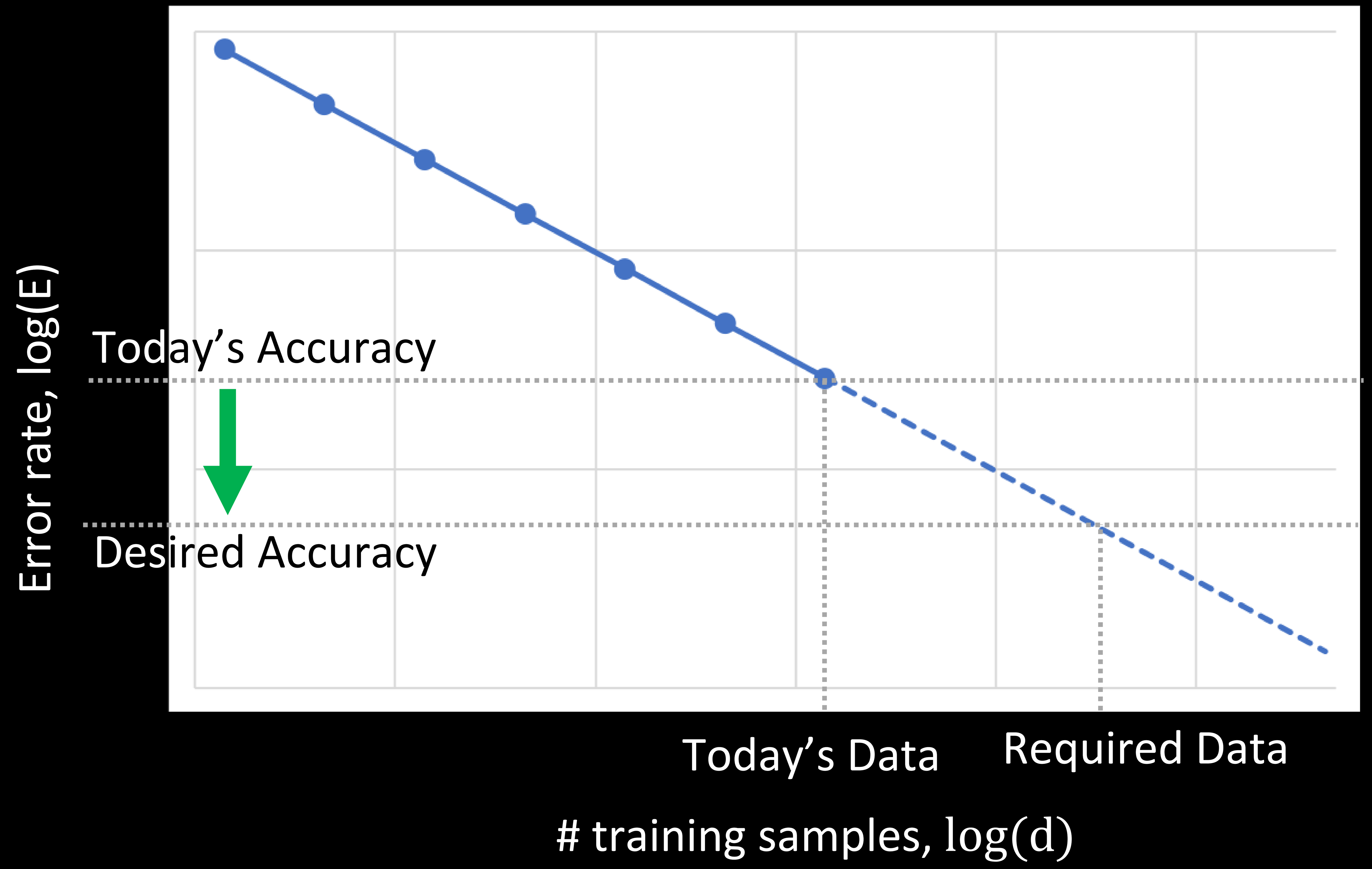


Bigger

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., ... & Zhou, Y. (2017).
Deep learning scaling is predictable, empirically. arXiv preprint arXiv:1712.00409.

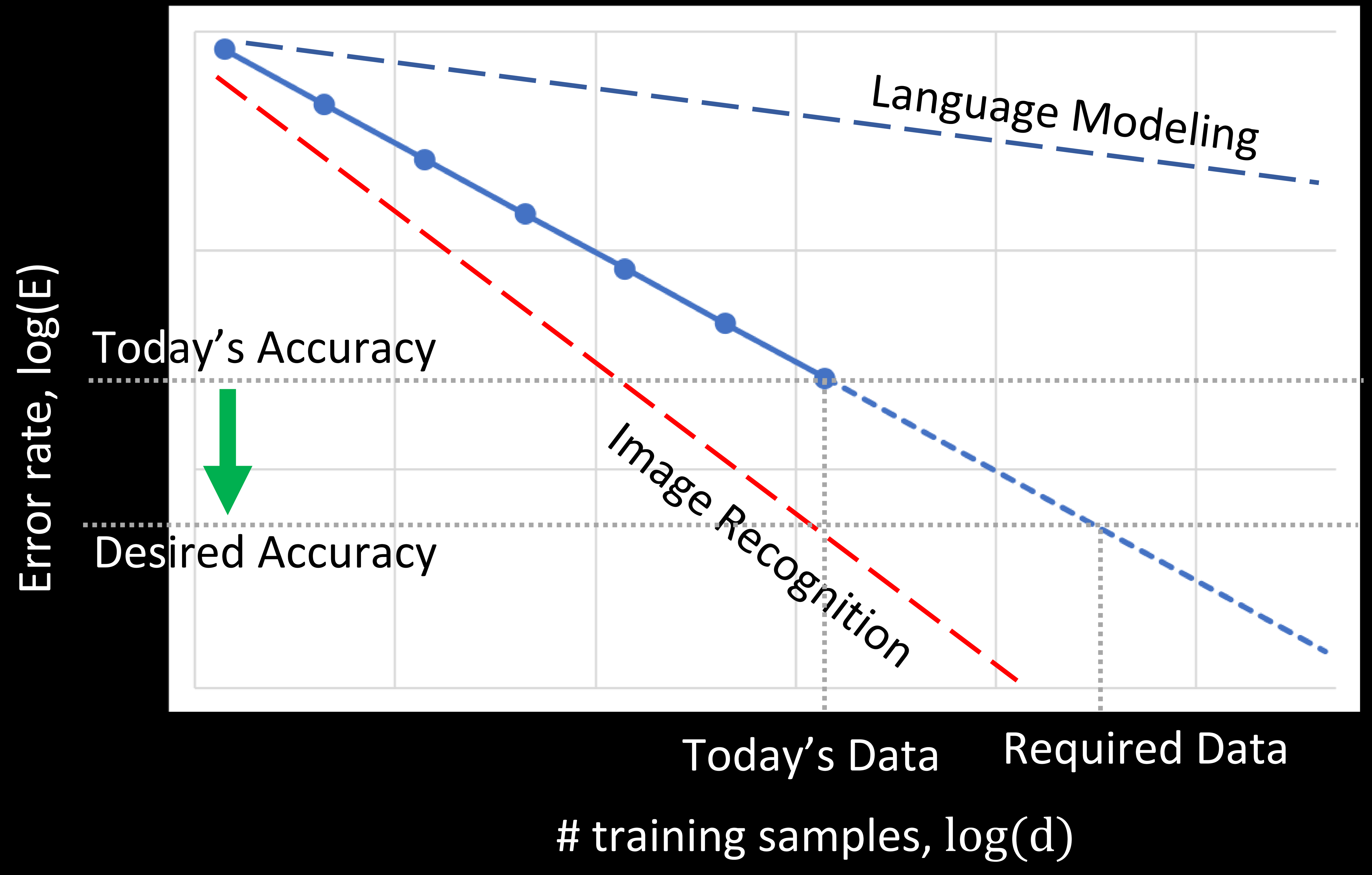
Baidu 百度

Better as Better Accuracy



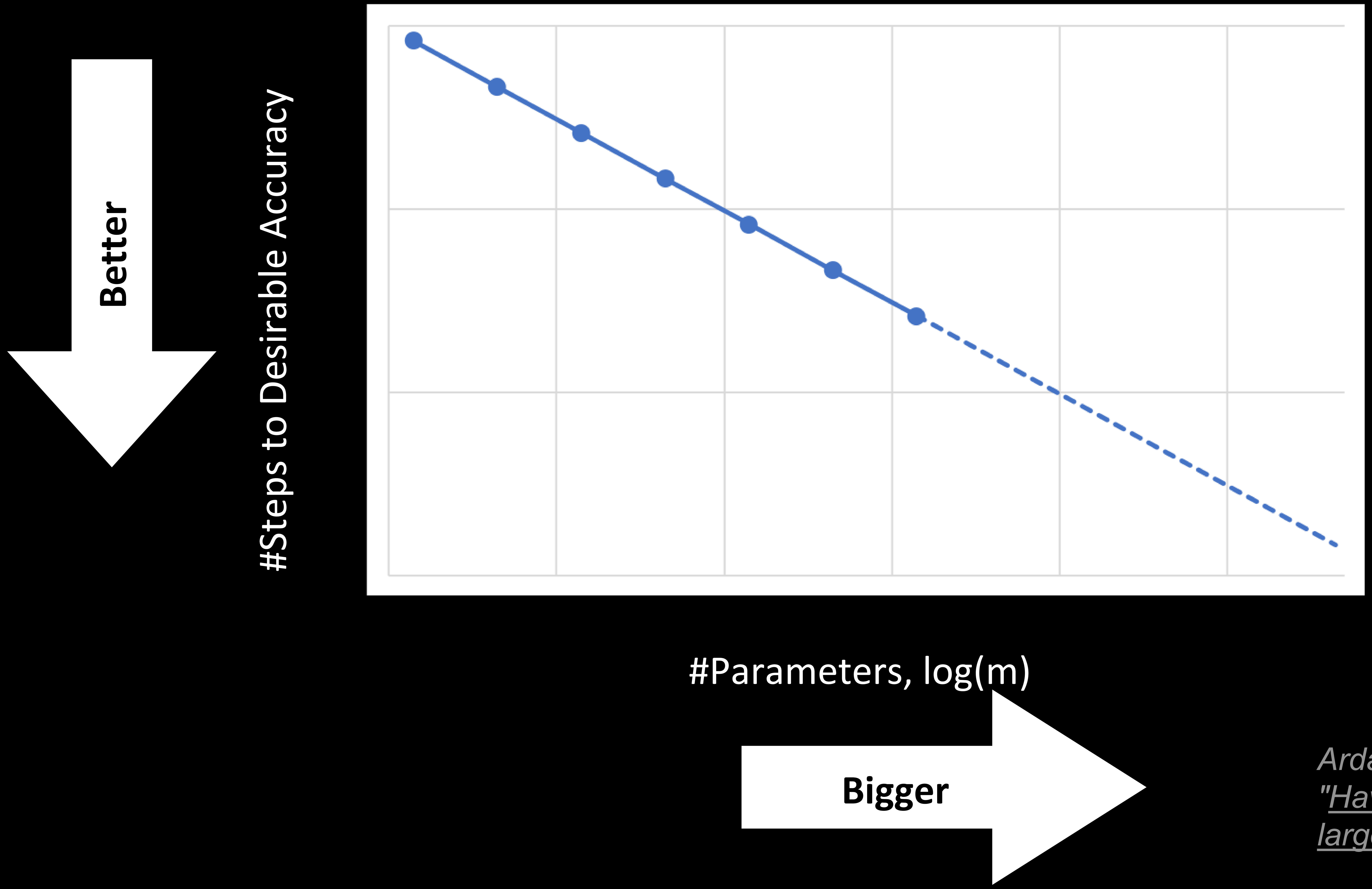
Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., ... & Zhou, Y. (2017).
Deep learning scaling is predictable, empirically. arXiv preprint arXiv:1712.00409.

Better as Better Accuracy



Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., ... & Zhou, Y. (2017). Deep learning scaling is predictable, empirically. arXiv preprint arXiv:1712.00409.

Better as Faster Training



Ardalani, N., Hestness, J., and Gregory Diamos. "Have a larger cake and eat it faster too: A guideline to train larger models faster." (SysML 2018).

Baidu 百度

Models and data are growing so fast in size…

Models and data are growing so fast in size…

Memory capacity/chip can grow only so much…

Models and data are growing so fast in size…

Memory capacity/chip can grow only so much…

Break the model and data into smaller chunks

Models and data are growing so fast in size…

Memory capacity/chip can grow only so much…

Break the model and data into smaller chunks

We need to exploit all forms of parallelism

Data Parallelism

Model Parallelism

Pipeline Parallelism

Hybrid Parallelism

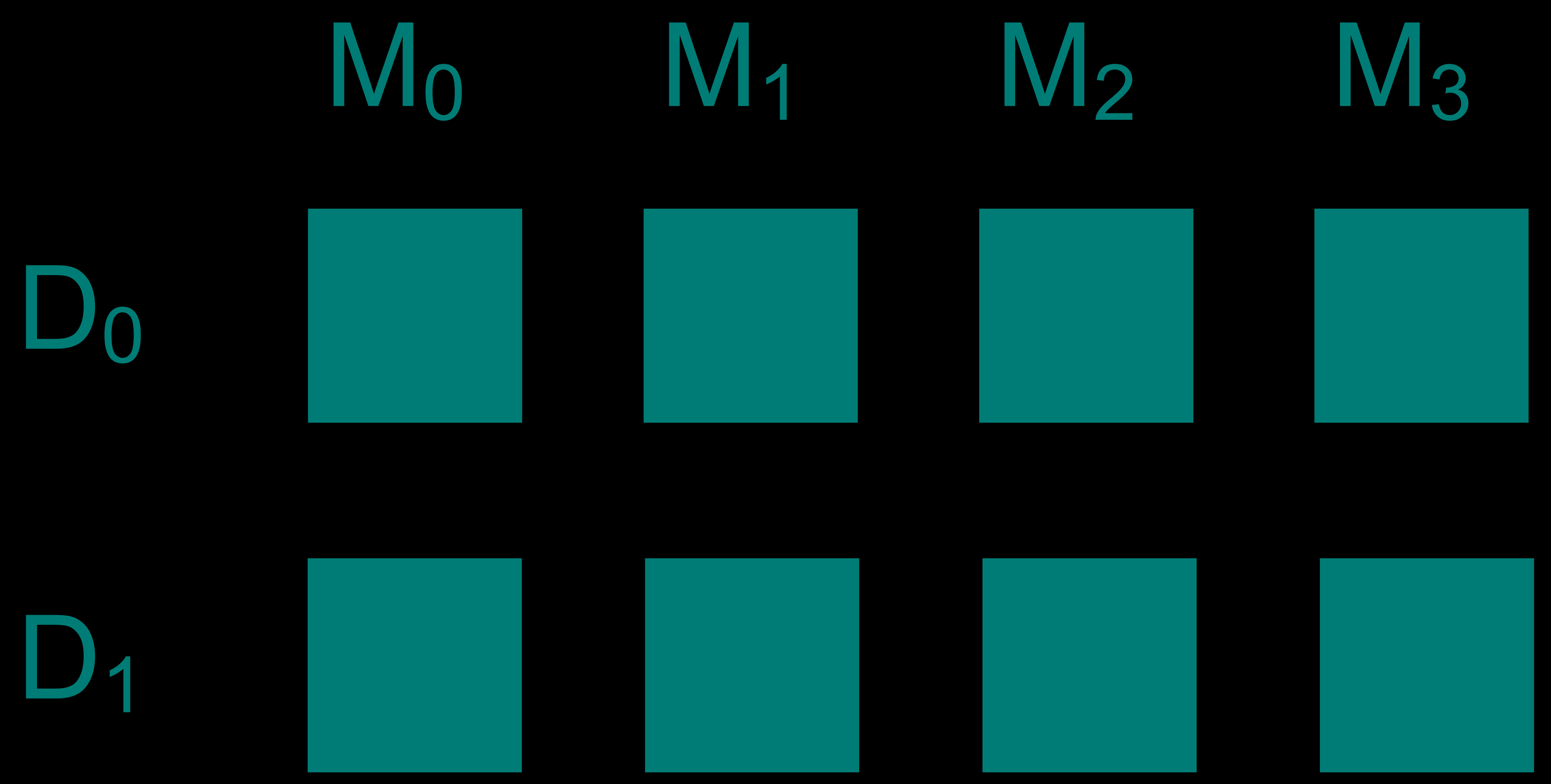
…

How to find a good parallelism strategy?

- Current Practice: Hire Expert Programmers

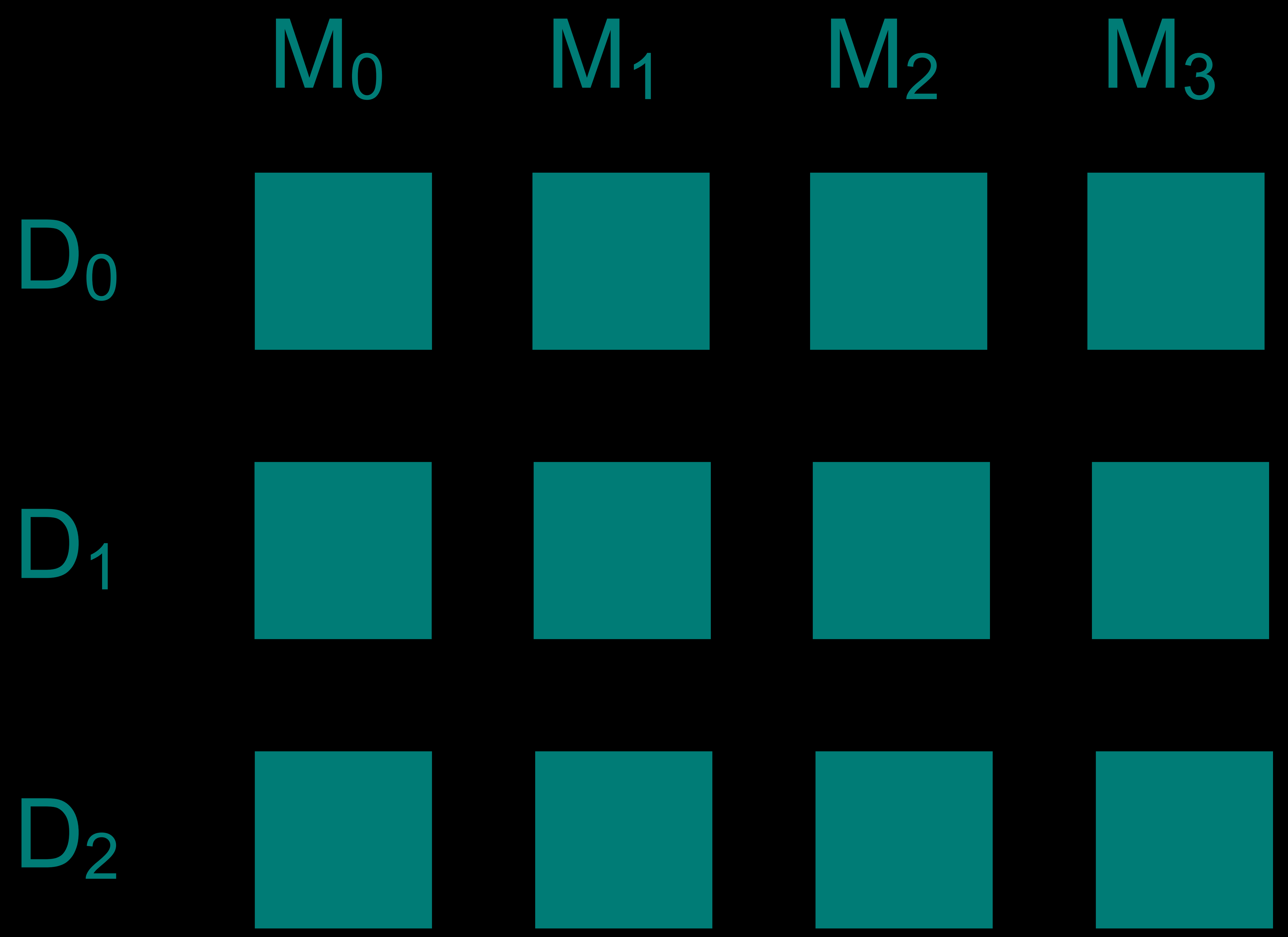
How to find a good parallelism strategy?

- Current Practice: Hire Expert Programmers



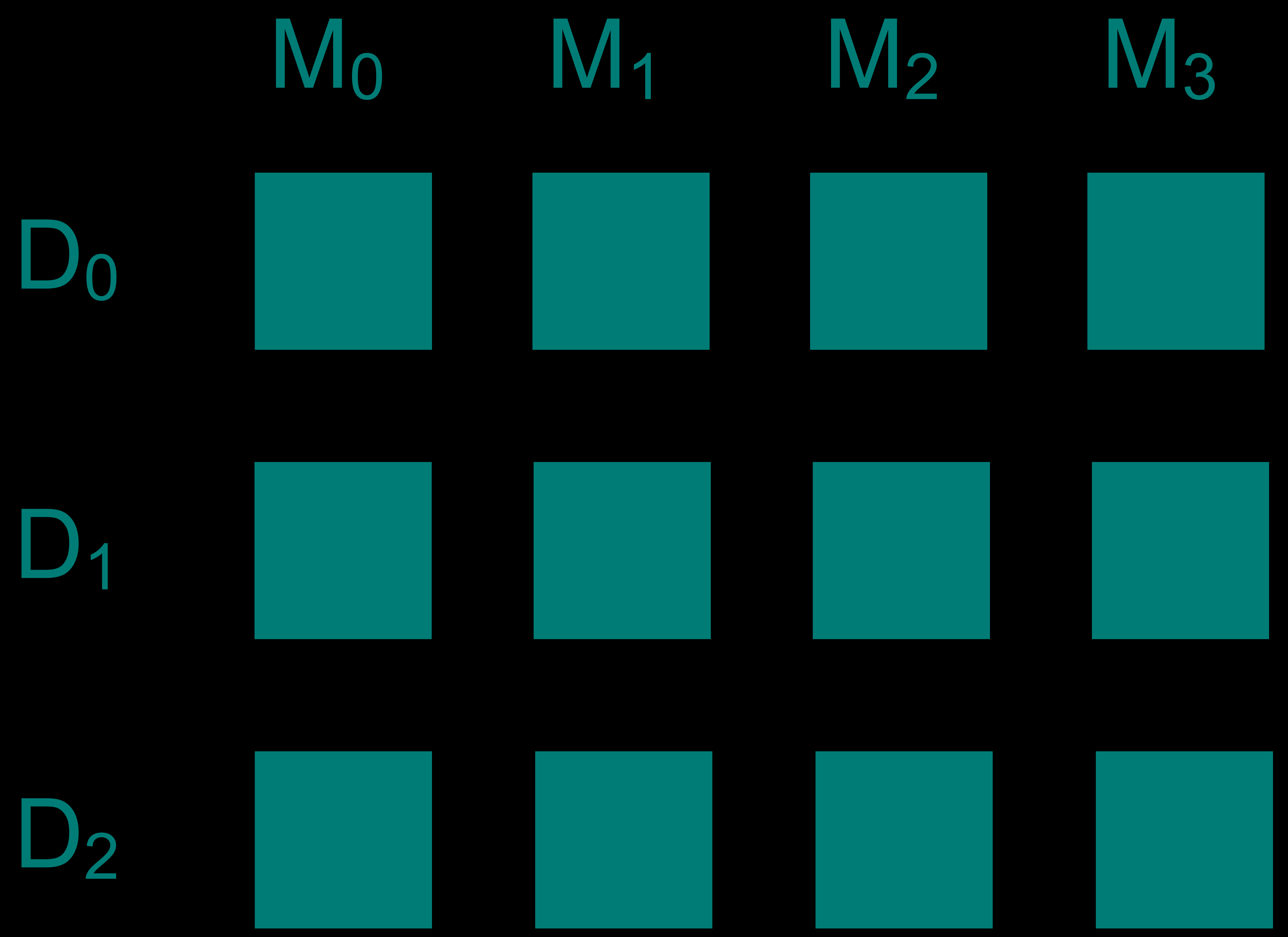
How to find a good parallelism strategy?

- Current Practice: Hire Expert Programmers



How to find a good parallelism strategy?

- Current Practice: Hire Expert Programmers
- Cutting edge: Reinforcement Learning, Dynamic Programming



GOOD NEWS

Best mapping/Best timing

BAD NEWS

System Under-utlization

Solution?

Co-design

Parallelism Strategy & Hardware Accelerator

Baidu 百度

Conclusion

GOOD NEWS

Bigger is Better

BAD NEWS

Memory is Bottleneck
Systems Underutilization

WHAT CAN WE DO?

Co-design
AI & HPC System

Bai du 百度

Cloud AI Computing Platform KongMing Architecture

Application	Speech/Images/NLP/Recommendation			
Framework	PaddlePaddle			
Cluster	K8S cloud native	Smart Scheduling	Elastic provision	
Performance Optimization	Compute optimization	IO optimiaztion	Communication optimzation	
System Design	AI computer X–MAN	High performance storage pool	High speed interconnect	
Chip	CPU	GPU	ASIC	**Baidu Kunlun**

- Highly Scalable
- HW/SW Co-design
- Heterogeneous Computing


Baidu百度
Thank you for listening!
百度一下