Deep Learning scaling is predictable (empirically)

Greg Diamos

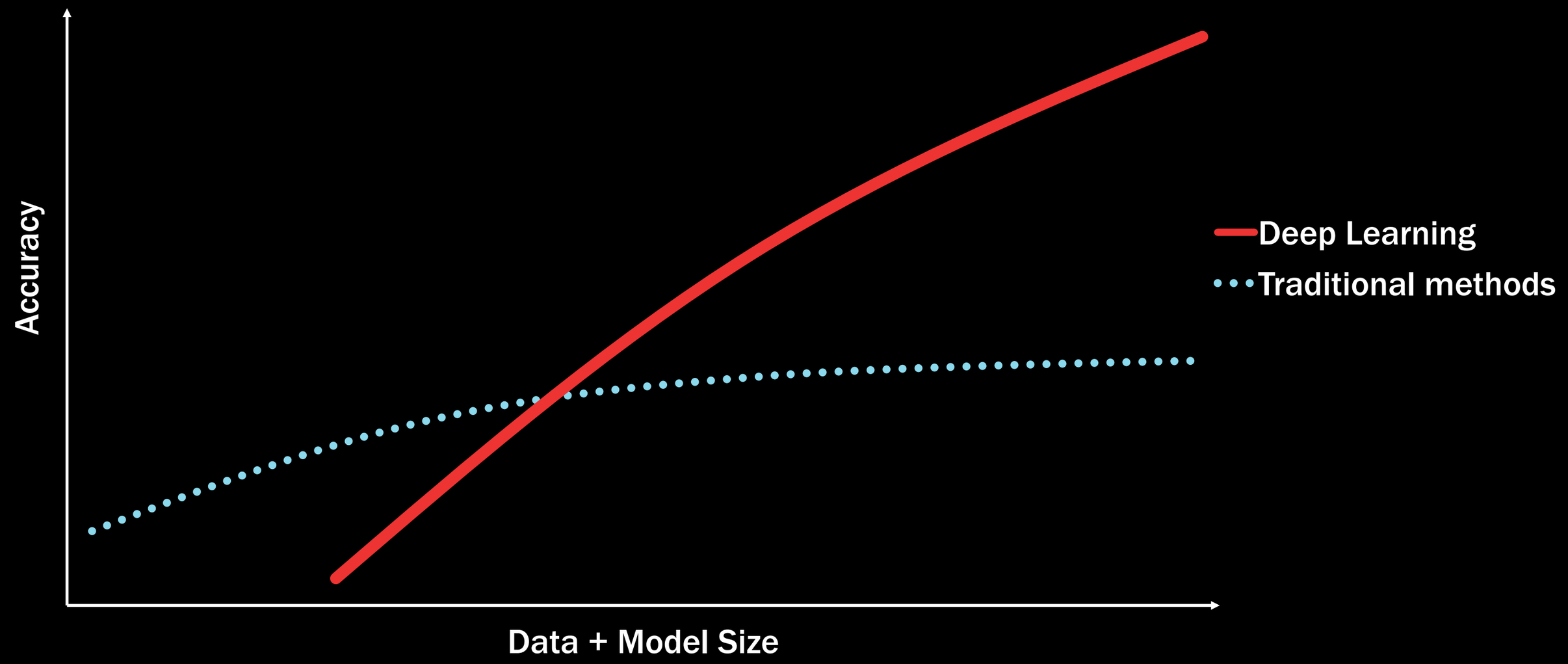
December 9, 2017

Baidu Research | Silicon Valley AI Lab

AI

- AI is like electricity

Deep Learning scales



Why?

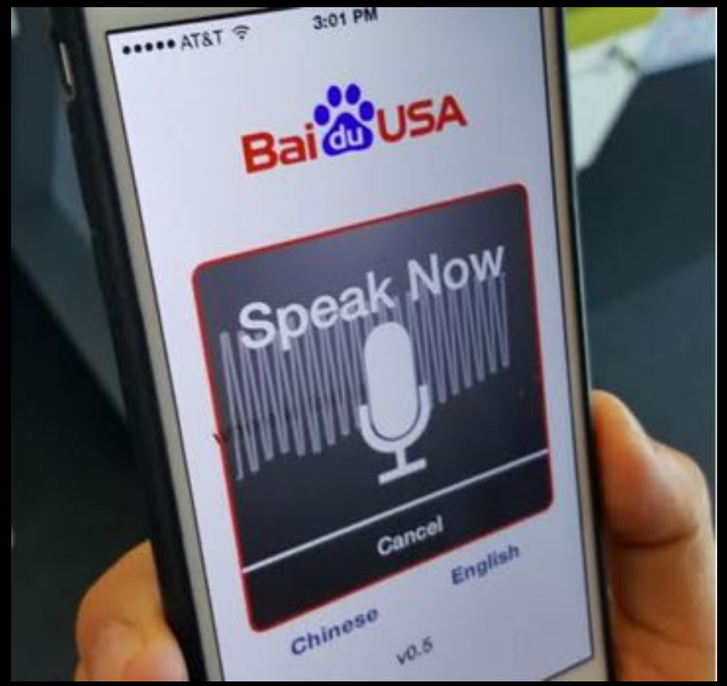
- Why do deep neural networks scale so well?
- How much data do we need?
- How fast do computers need to be?

This talk: looking deeper

SVAIL's ASIMOV supercomputer

- We used a 11 PFLOP/s GPU supercomputer to study deep learning scaling
- 1500 GPUs
- 2 months training time
- **This experiment would cost over $2 million USD if performed on AWS**

Application domains



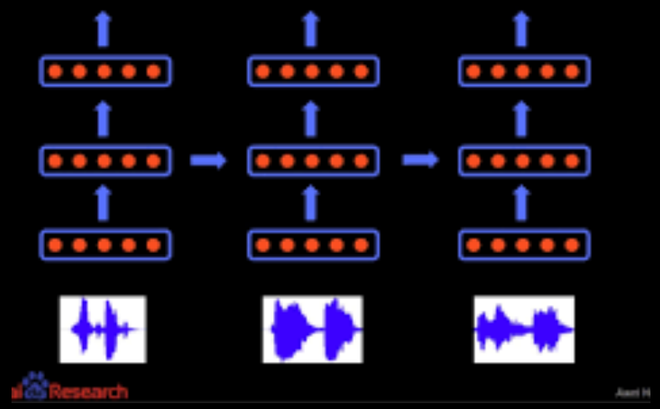
Speech Recognition

Computer Vision

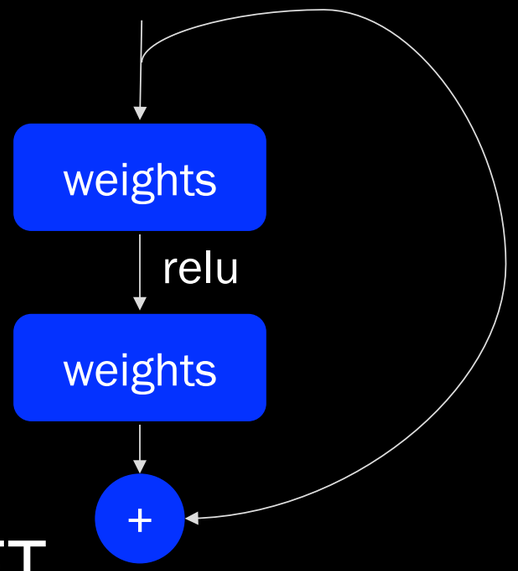
Speech Synthesis

Natural Language Understanding

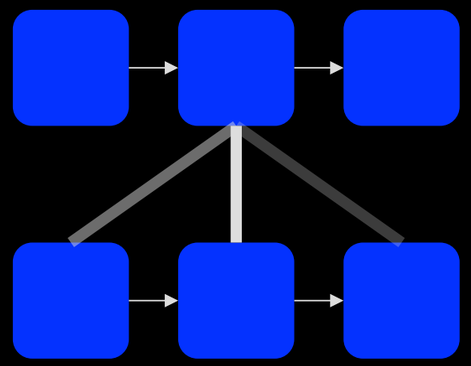
State of the art neural nets



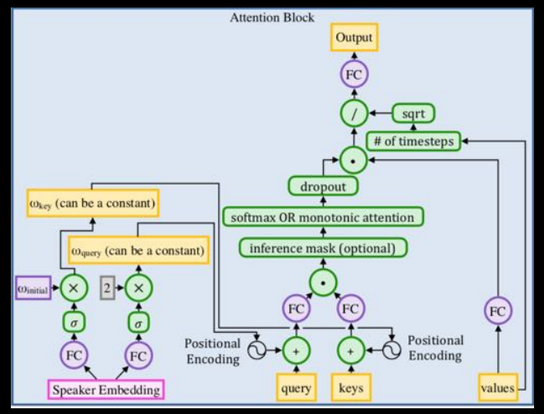
CONV + RNN



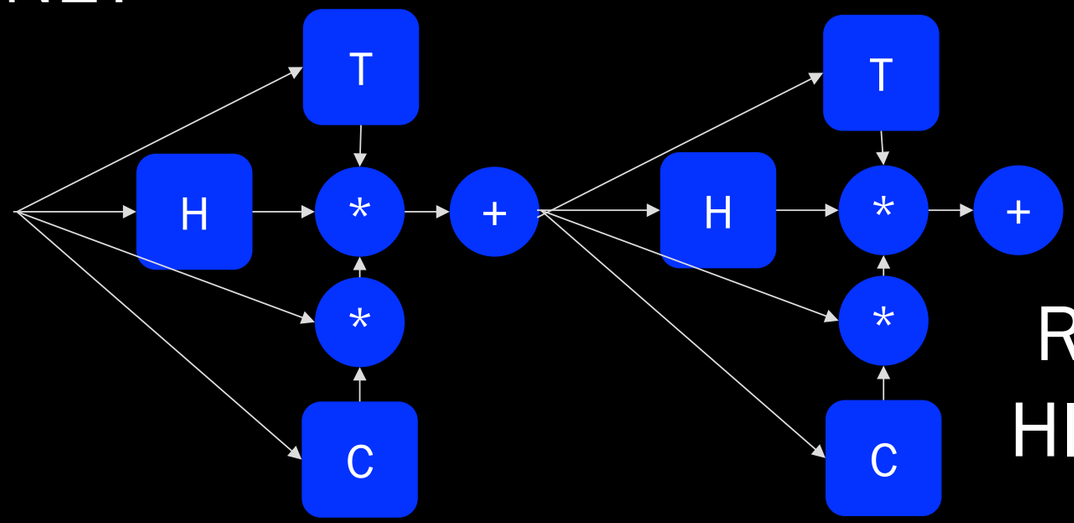
RESNET



RNN + ATTENTION

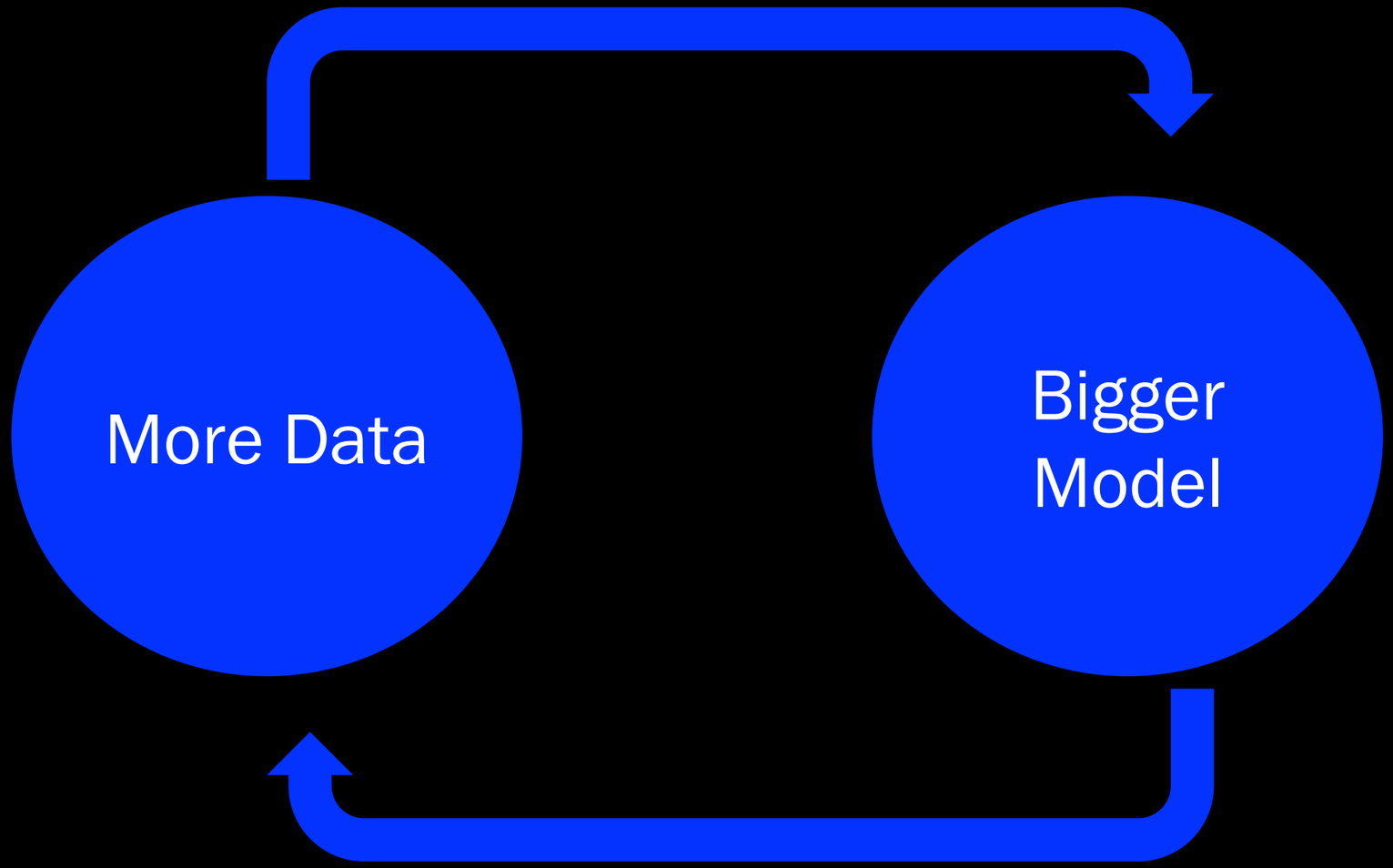


SPRECTRA NET



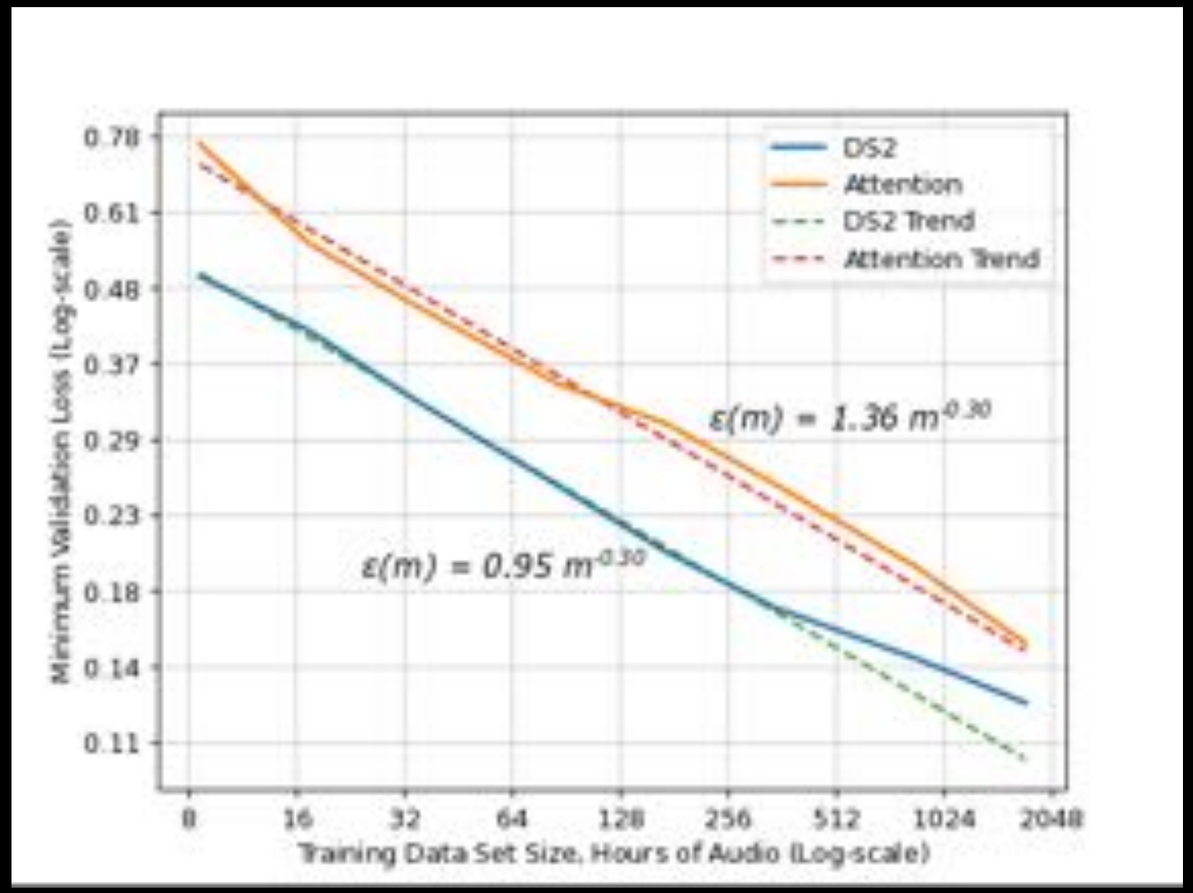
RECURRENT HIGHWAY NET

Methodology

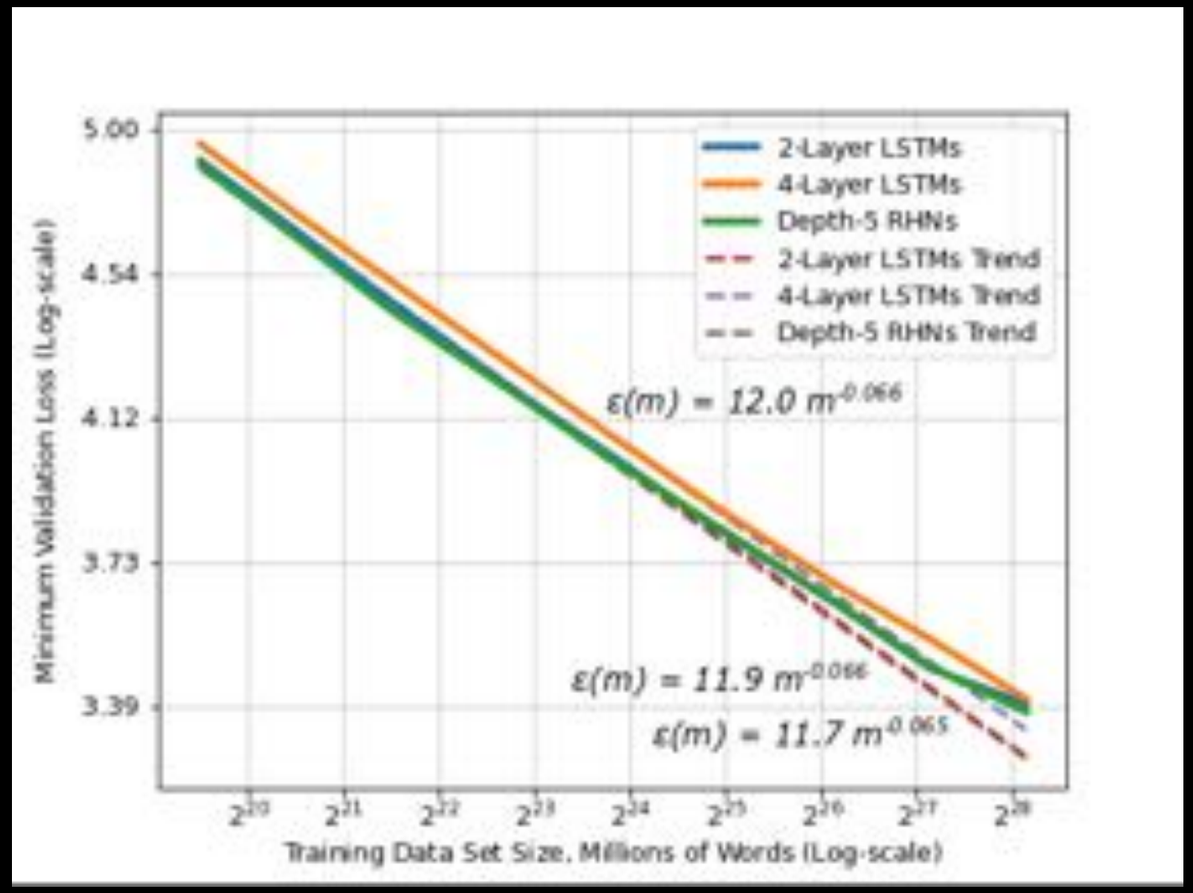


Generalization error scaling

Generalization error scaling



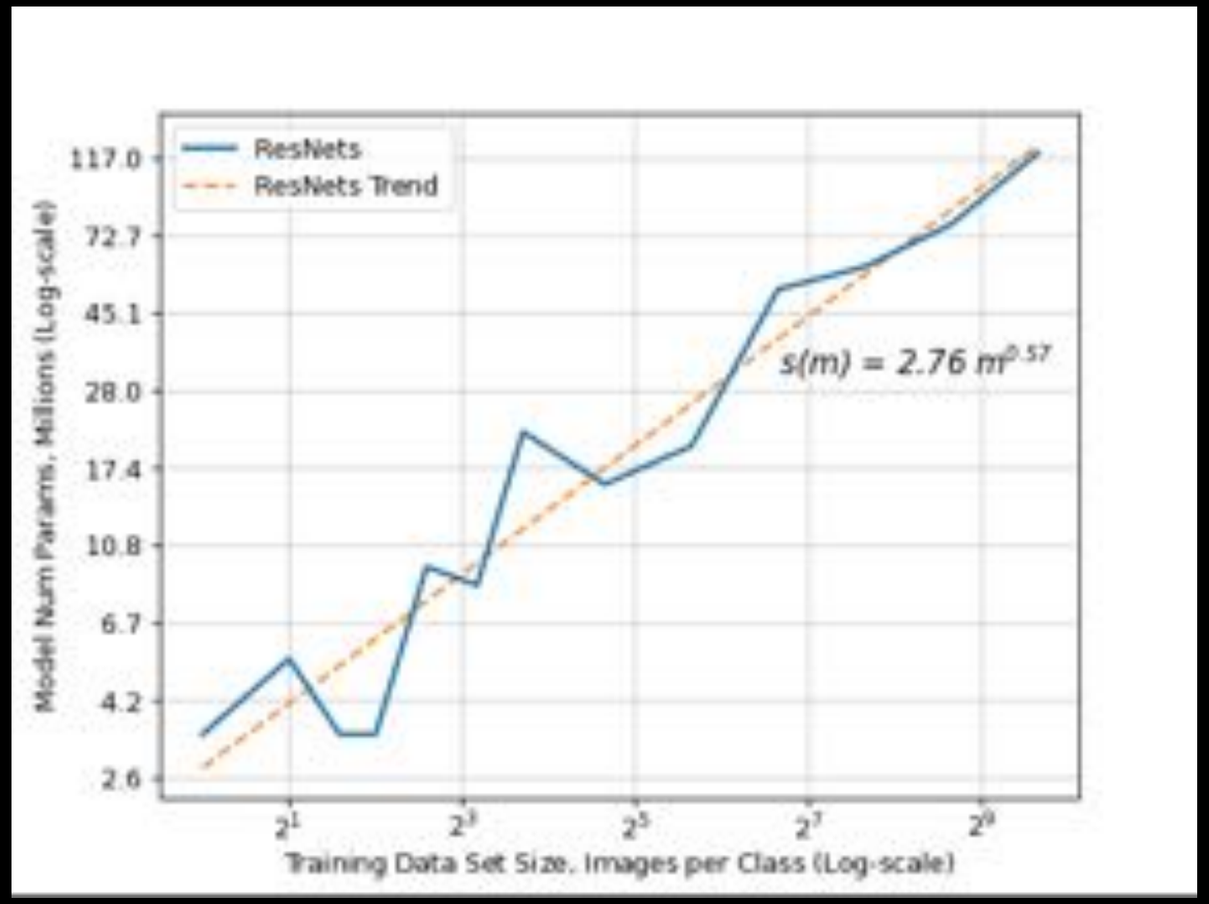
Deep Speech



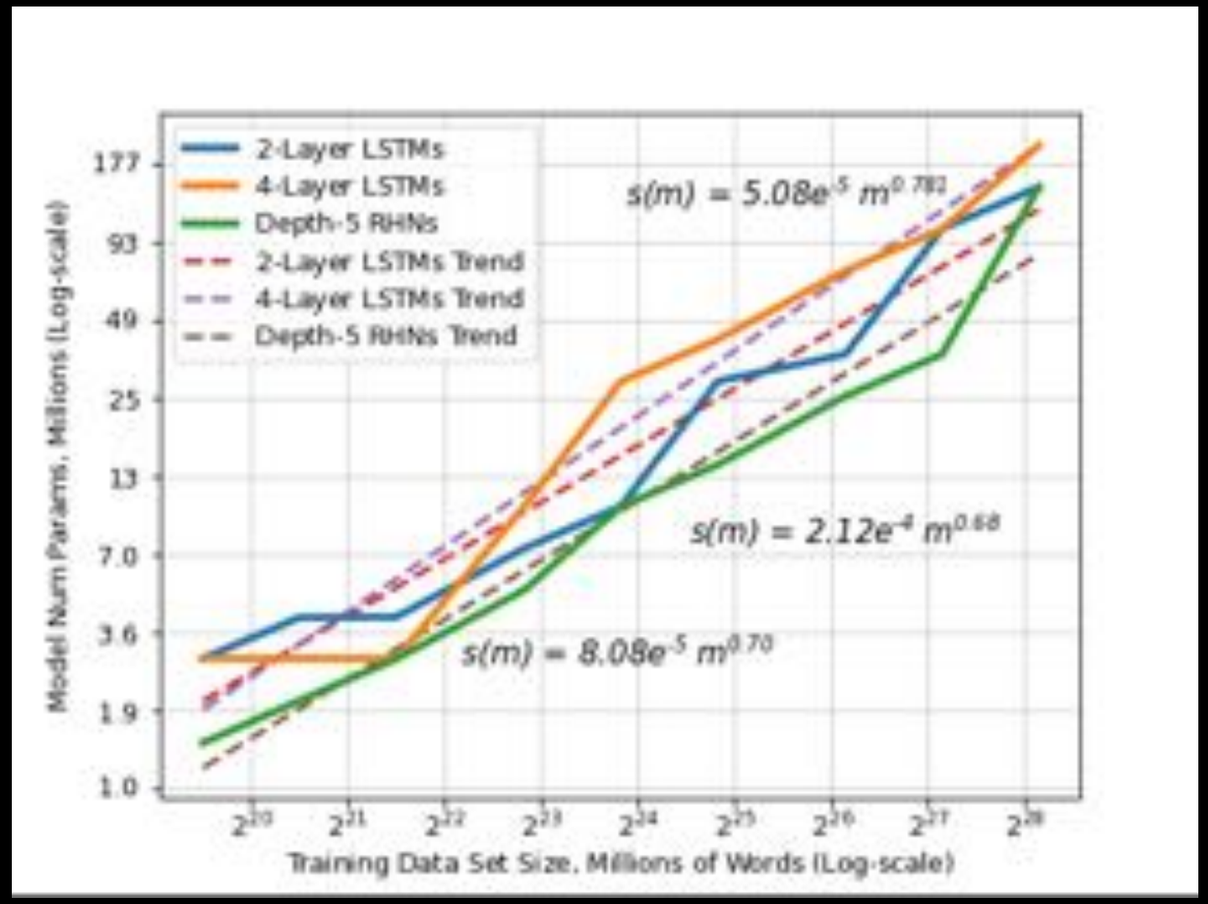
Neural Language Model

Model size scaling data

Model size scaling data



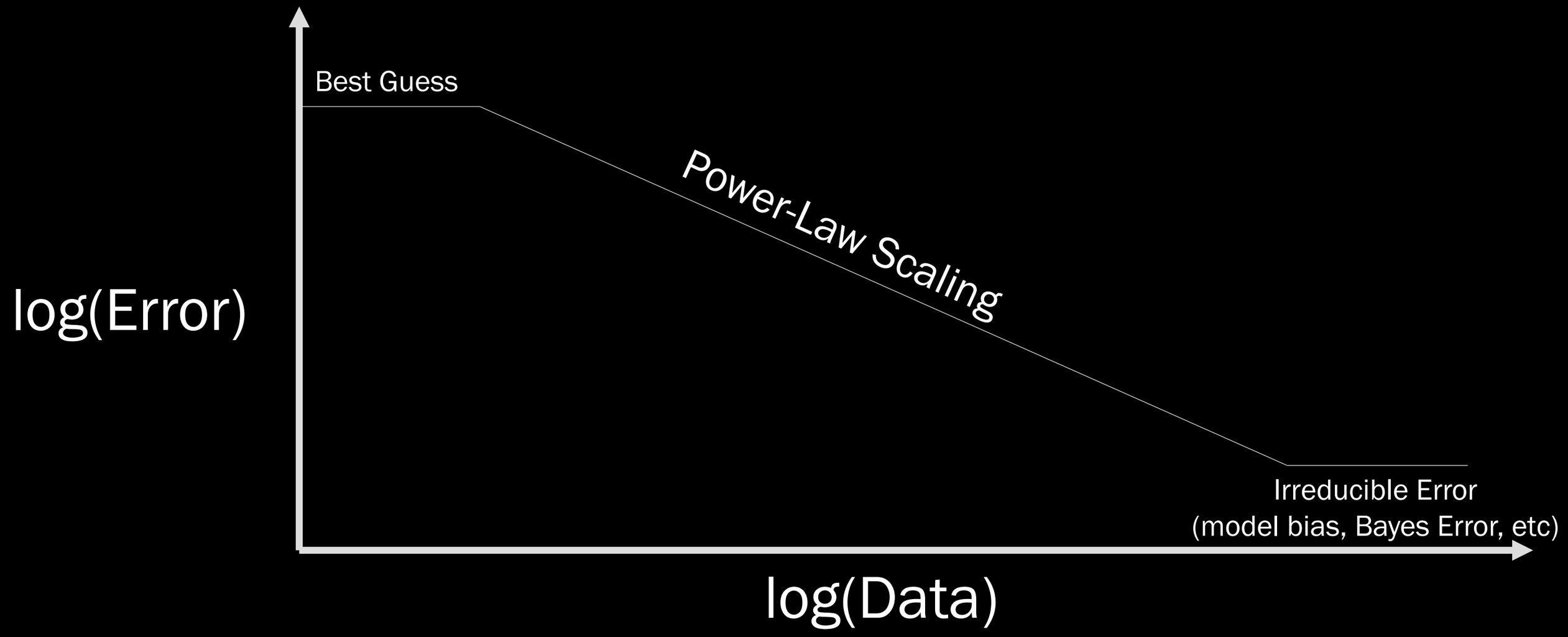
Resnet50 Object Detection



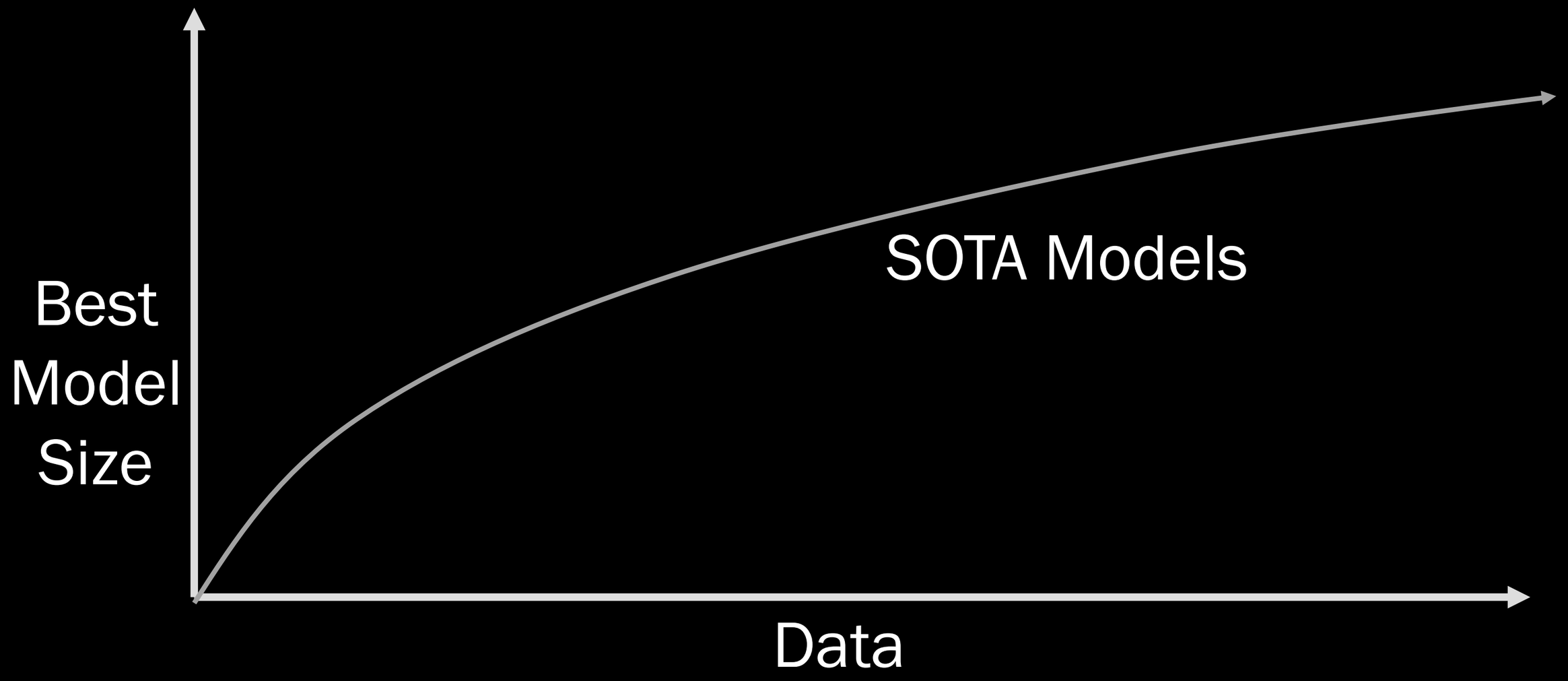
Neural Language Model

What do you think?

We find: generalization error scaling consistently follows a power-law



We find: model size scales sublinearly



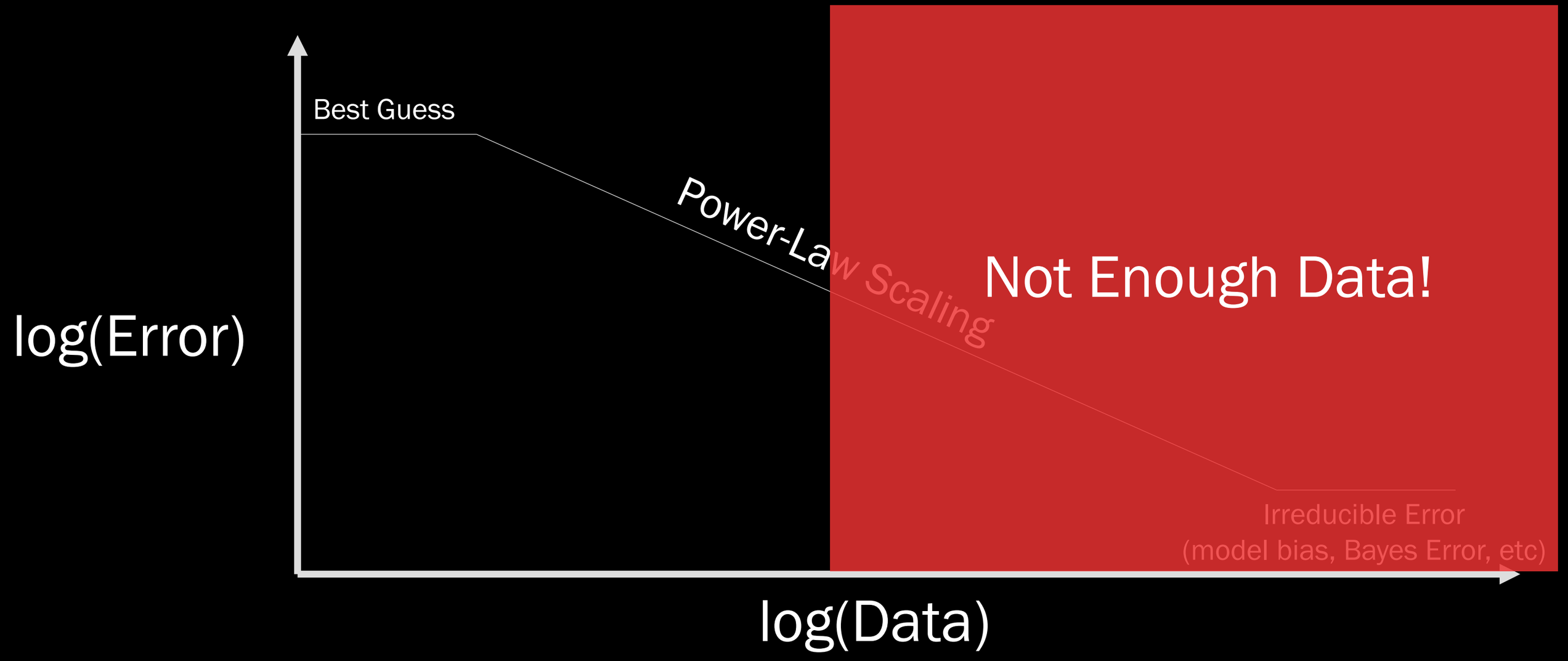
Acknowledgements

Baidu Research | Silicon Valley AI Lab

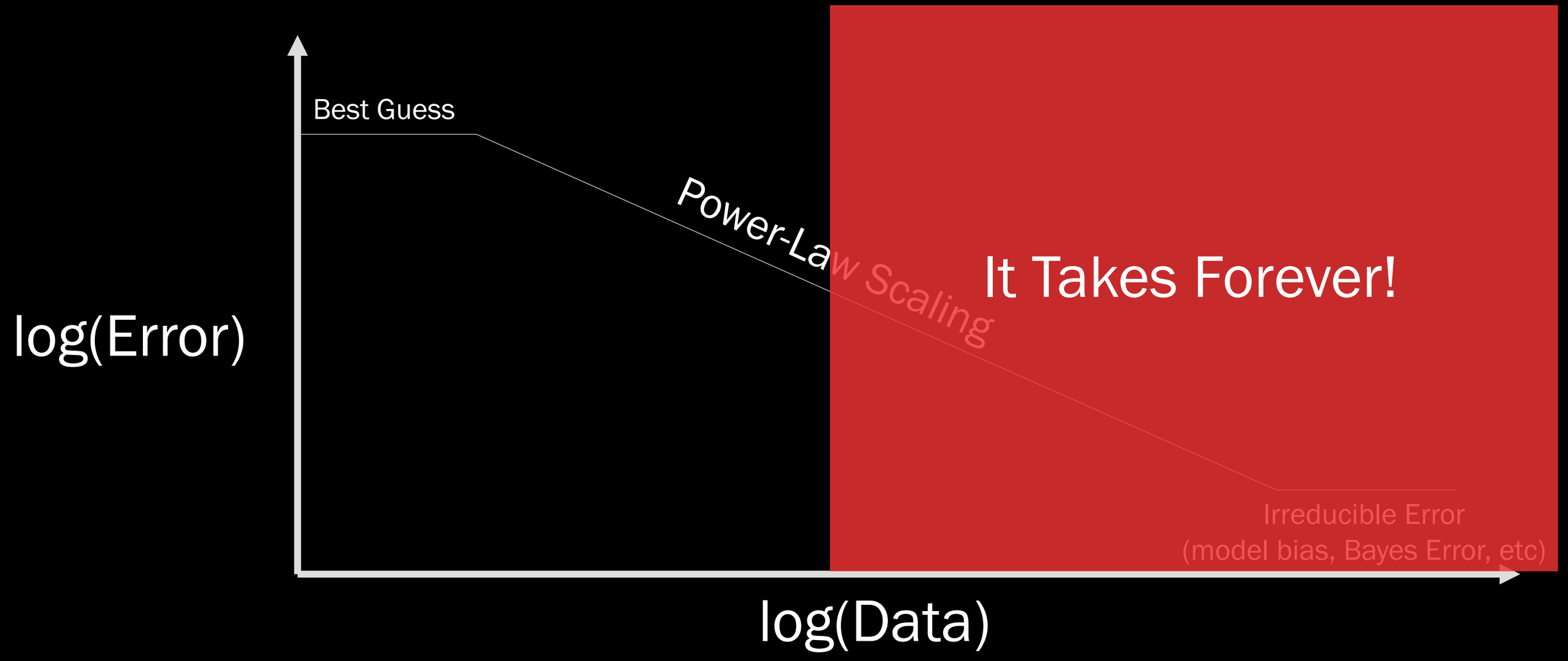
The Deep Learning Recipe

Baidu Research | Silicon Valley AI Lab

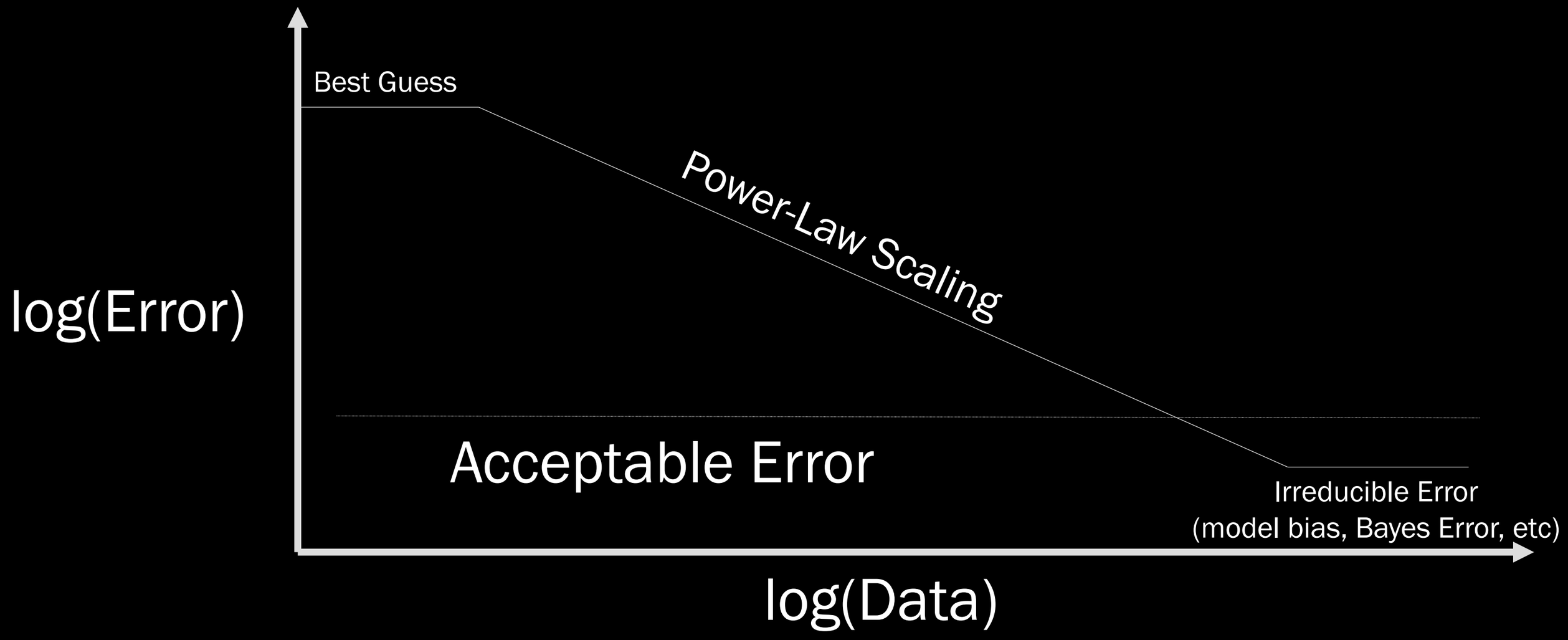
Data-limited problems



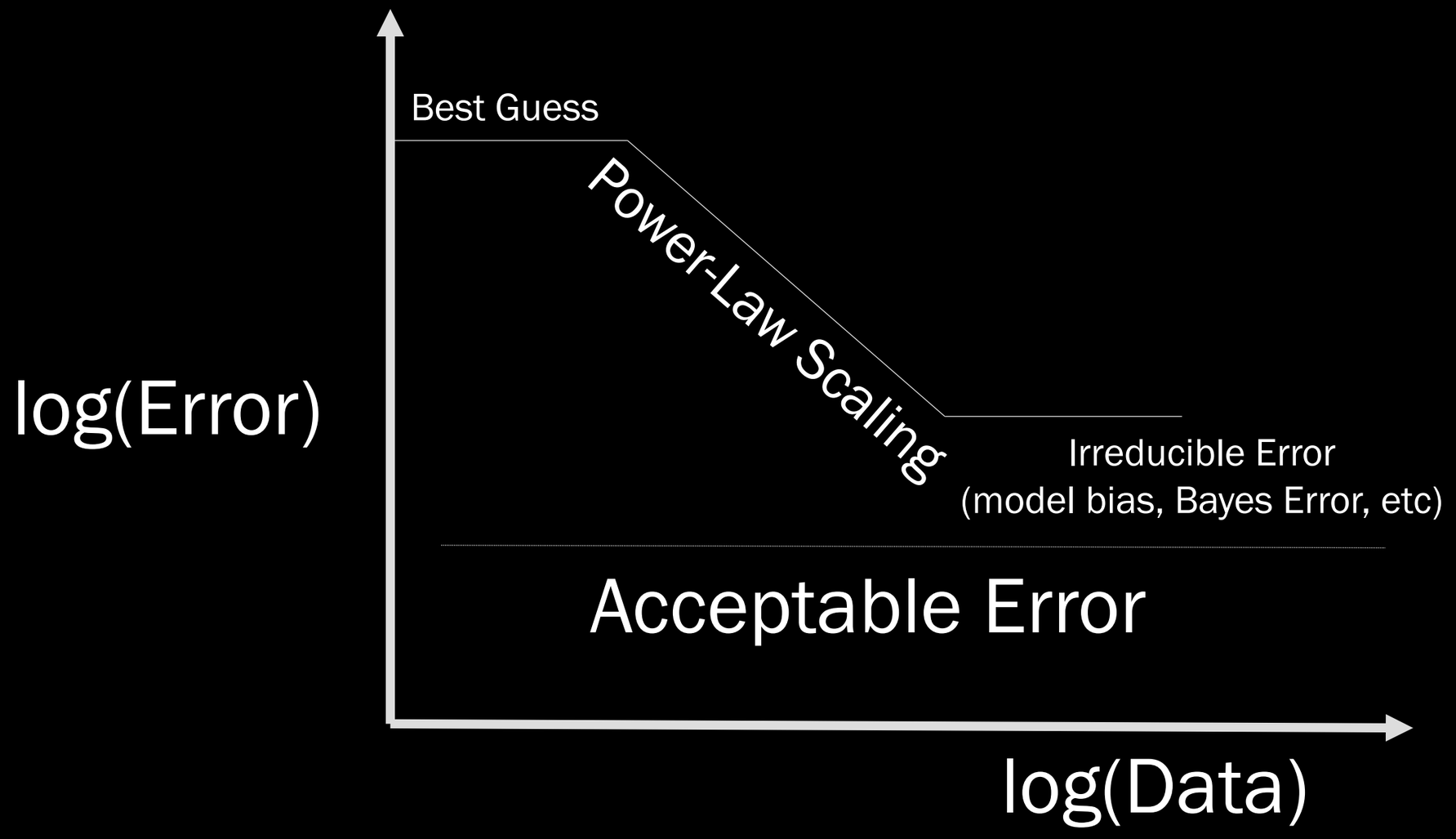
Compute-limited problems

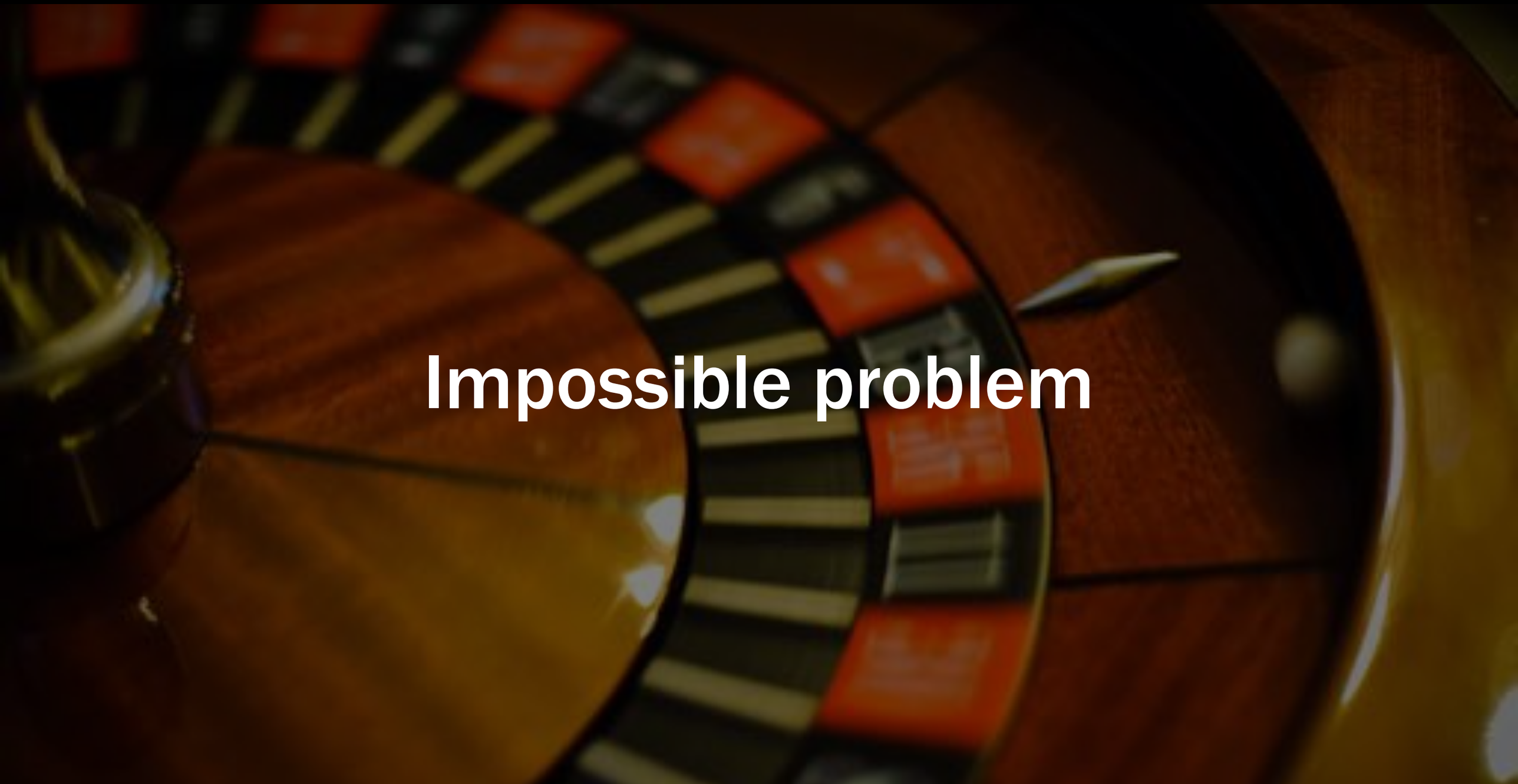


Solved problems



Impossible problems





Impossible problem

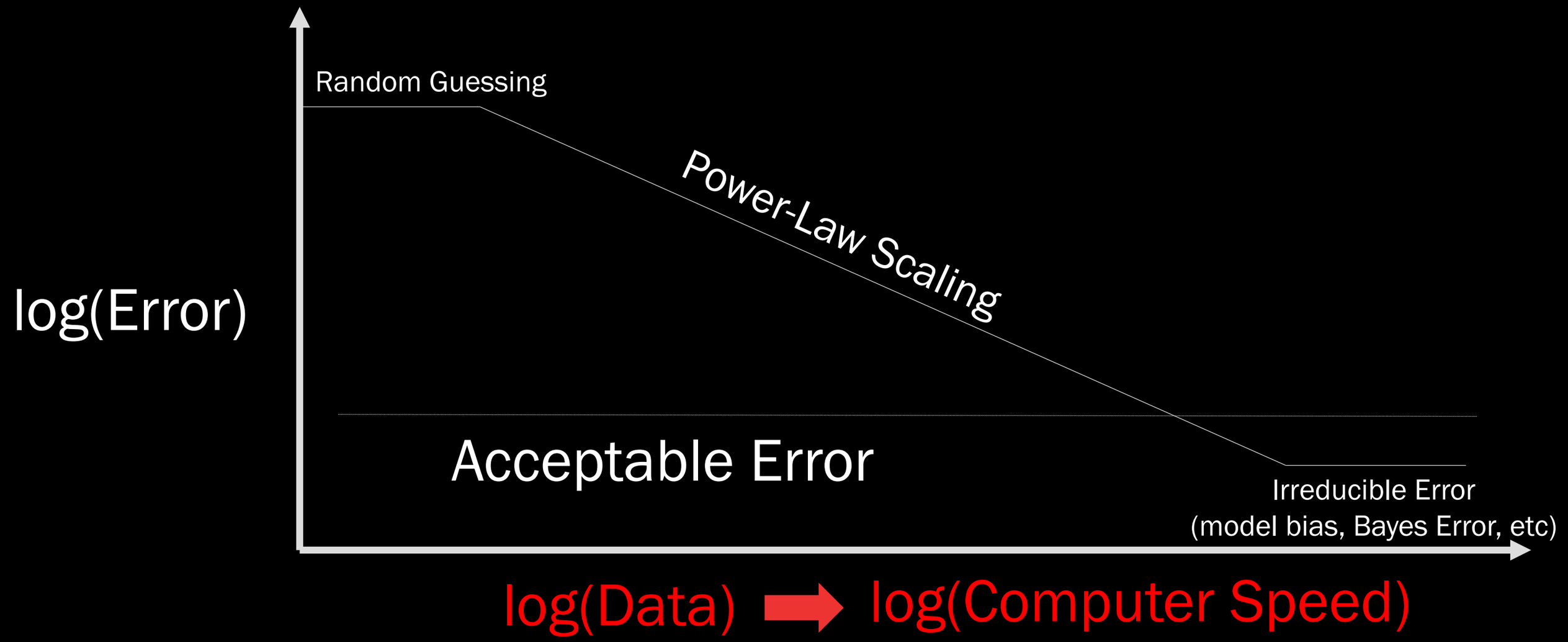
Implications

#1: Data is extremely valuable

- If all you need is scale, then we should invest in data

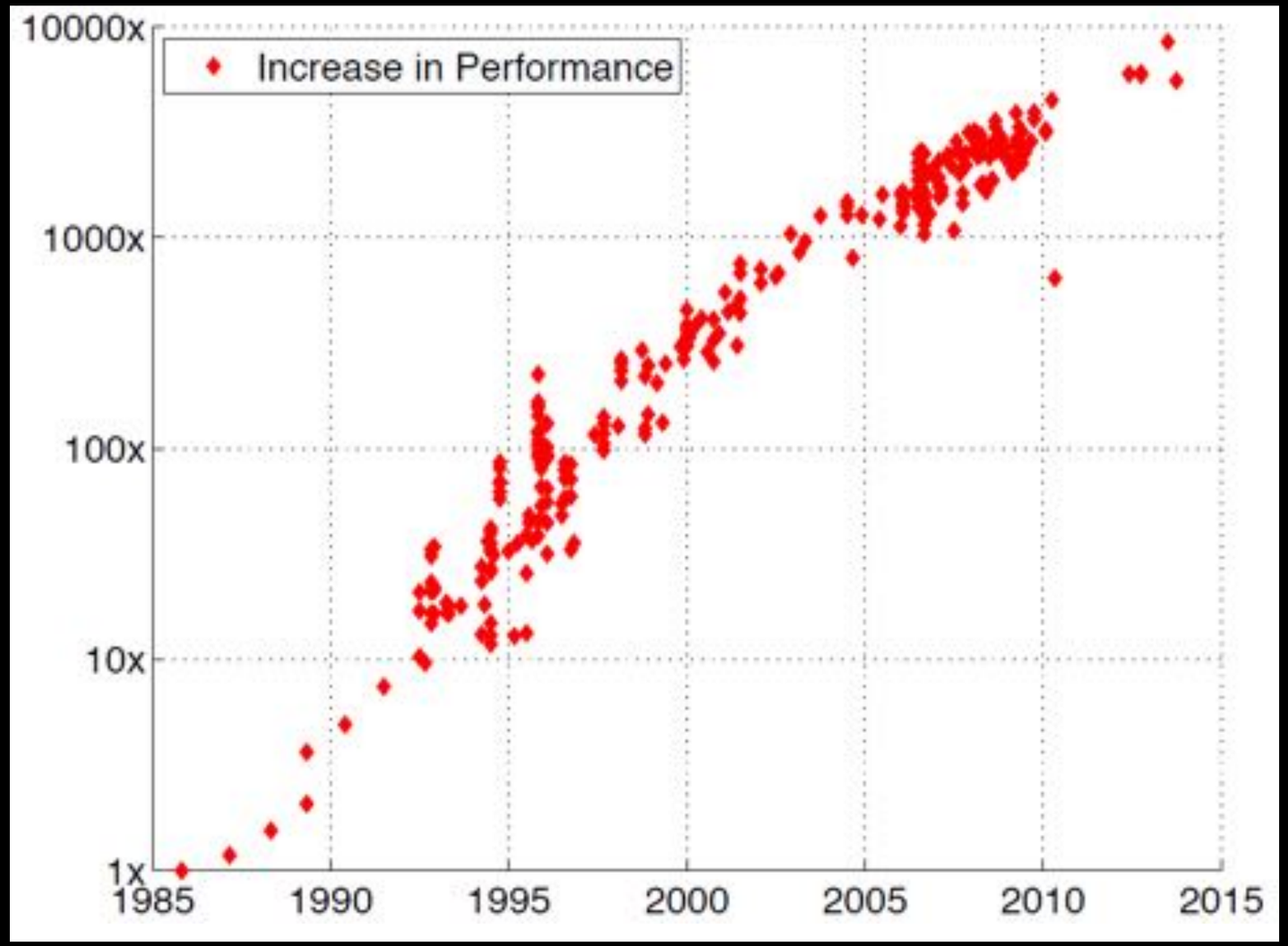
- How can we reduce the cost to collect and label data?

#2: Achievable error follows Moore's Law



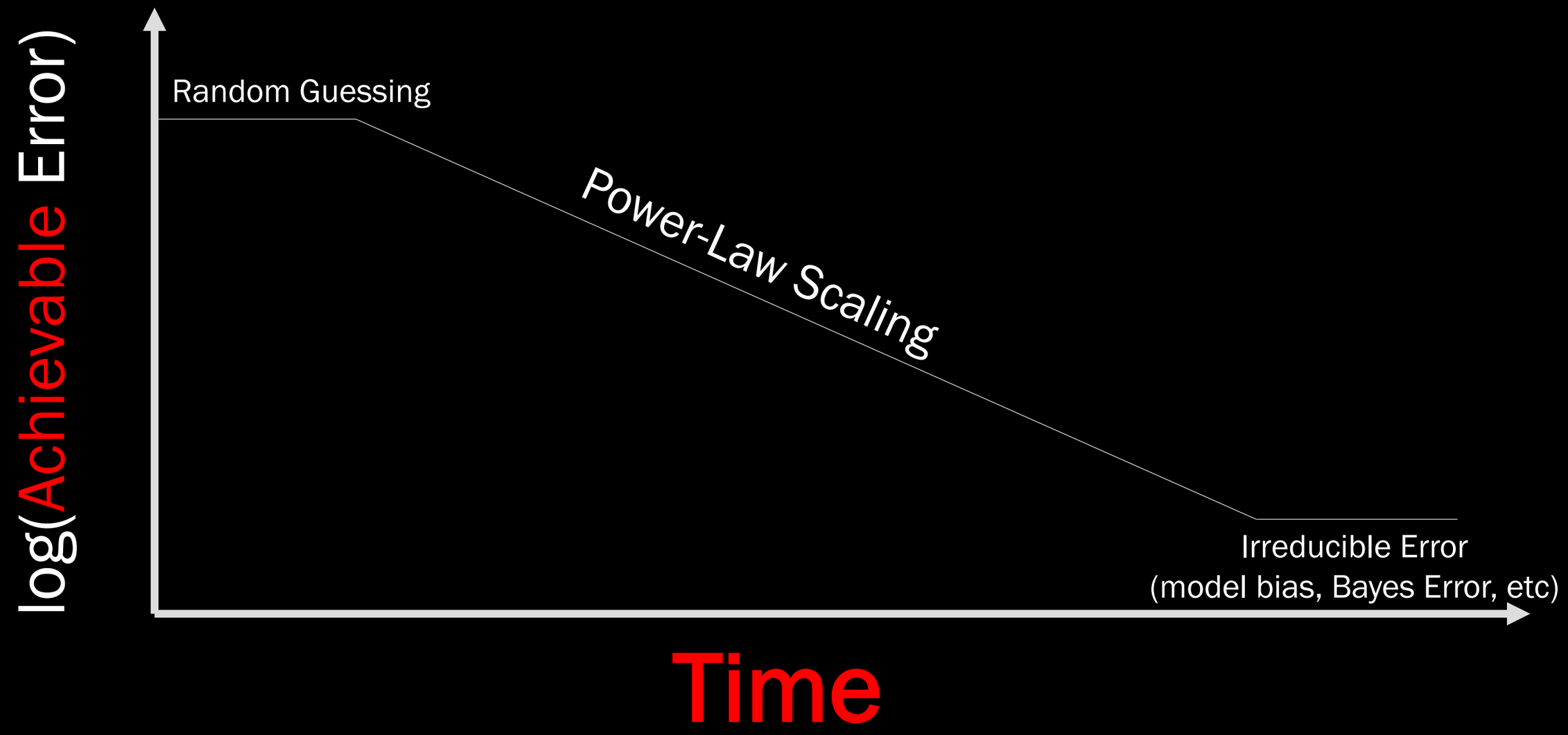
#2: Achievable error follows Moore's Law

log(Computer Speed)



Time

#2: Achievable error follows Moore's Law



#3: Requirements are predictable

- We can now predict
  - How much data we need
  - How fast computers need to be

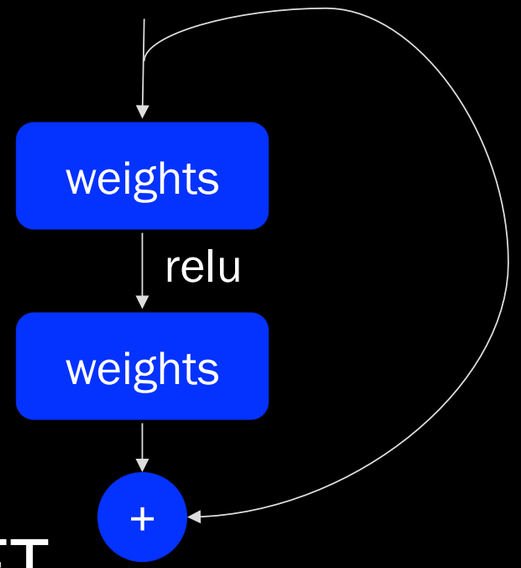
#4: Model architecture search

- Search may be feasible in the small data regime
  - if architecture affects the intercept, not the slope
- Caveats:
  - variance
  - models with different irreducible error

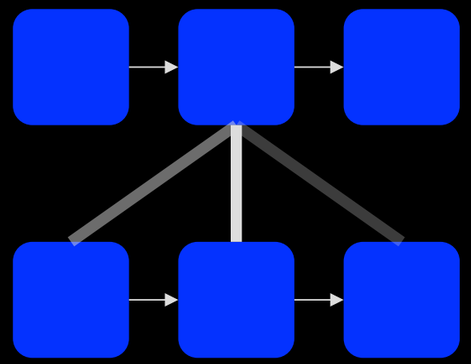
We need you!

Baidu Research | Silicon Valley AI Lab

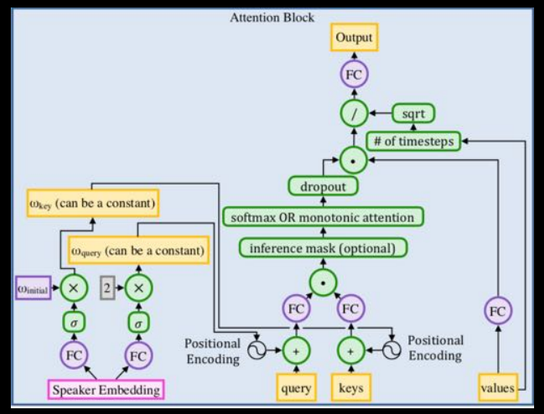
Reproduce our work



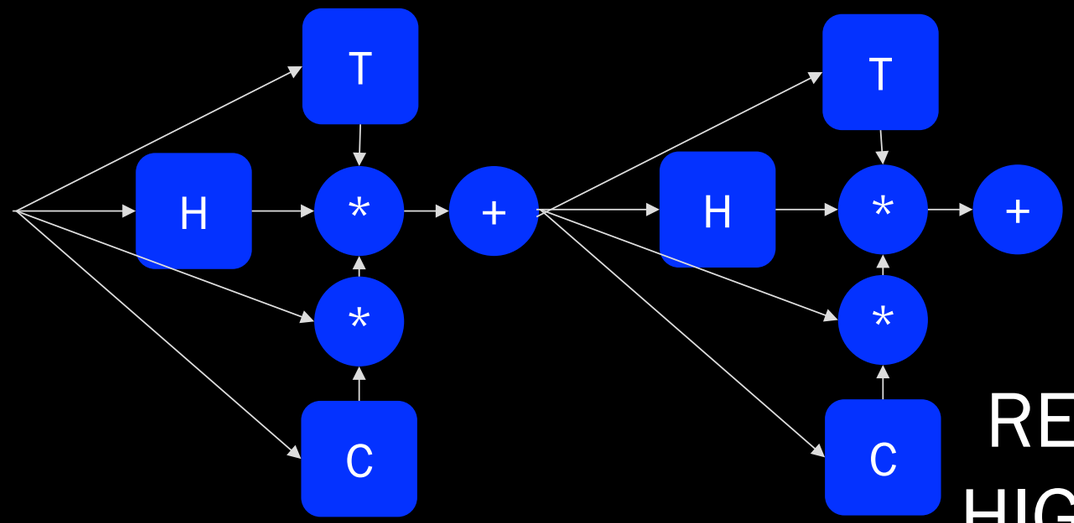
RESNET



RNN + ATTENTION



SPRECTRA NET



RECURRENT HIGHWAY NET

?

Build AI Data Centers

AI Node
1x
2017

AI Data Center
10,000x-100,000x
2025

Improved AI Chips
10x-100x
2025

Join Us!

- http://bit.ly/join-svail

Deep Learning scaling is predictable (empirically)

http://research.baidu.com/deep-learning-scaling-predictable-empirically/
https://arxiv.org/abs/1712.00409

Greg Diamos

December 9, 2017

Baidu Research | Silicon Valley AI Lab