Deep Speech: Scaling up end-to-end speech recognition

Awni Hannun

Outline

- State of Speech Recognition
- Overview: Deep Learning
- Deep Speech
- Next Steps

Outline

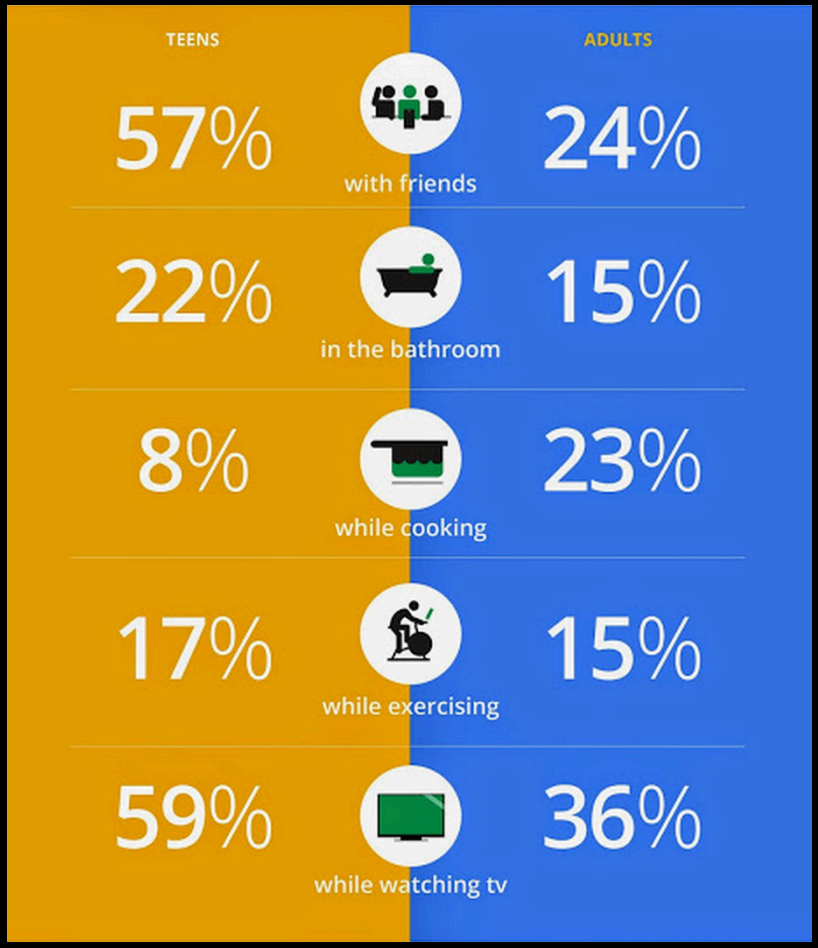
- State of Speech Recognition
- Overview: Deep Learning
- Deep Speech
- Next Steps

State of Speech

Where does ASR not work well?

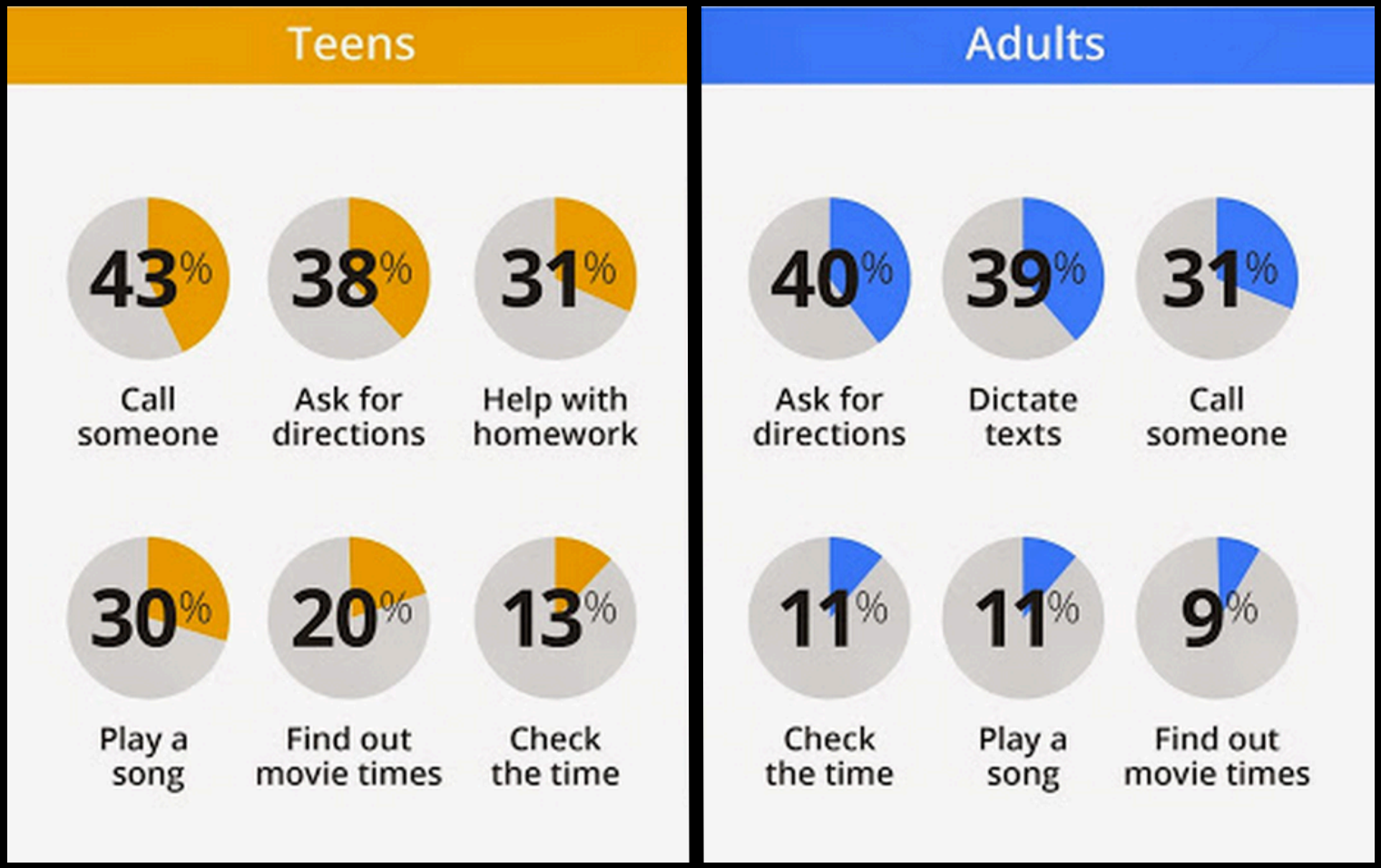
- Low signal-noise ratio
- Speaker variability (e.g. accents)
- Natural/conversational speech

State of Speech



Source : http://googleblog.blogspot.com/2014/10/omg-mobile-voice-survey-reveals-teens.html

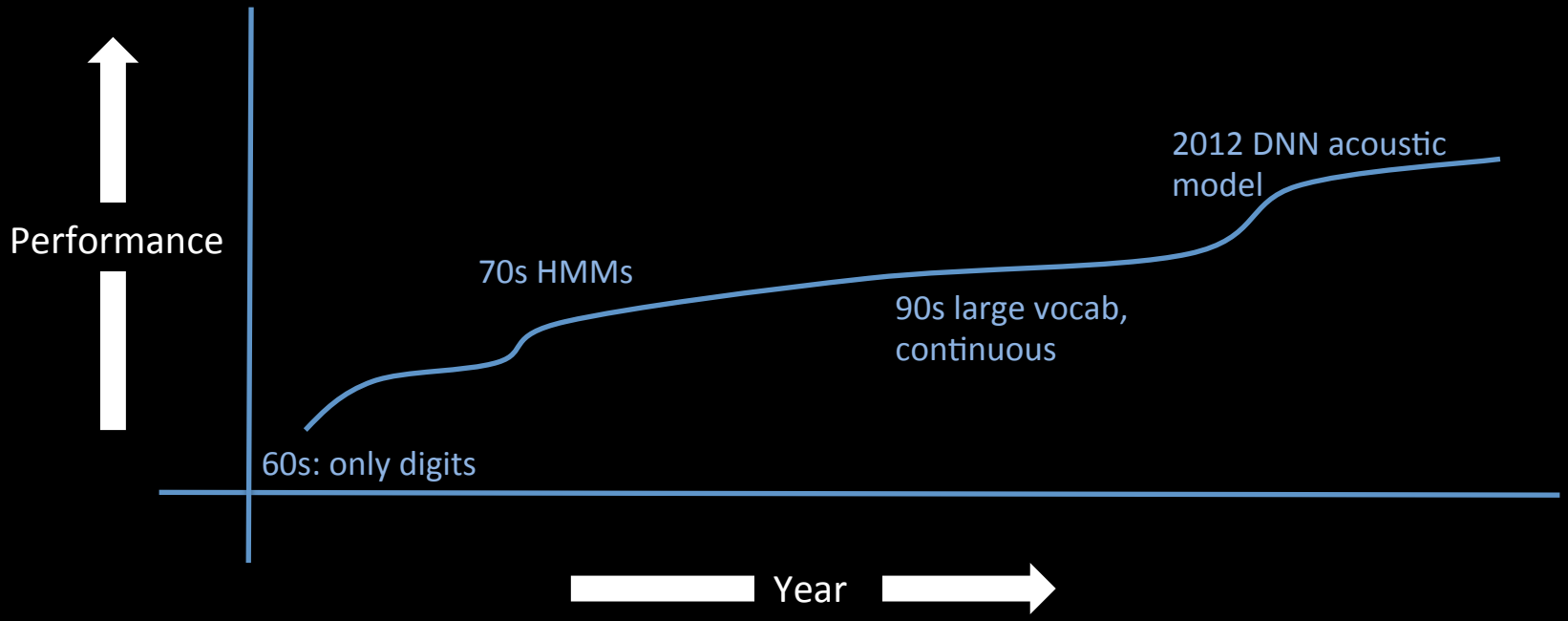
State of Speech



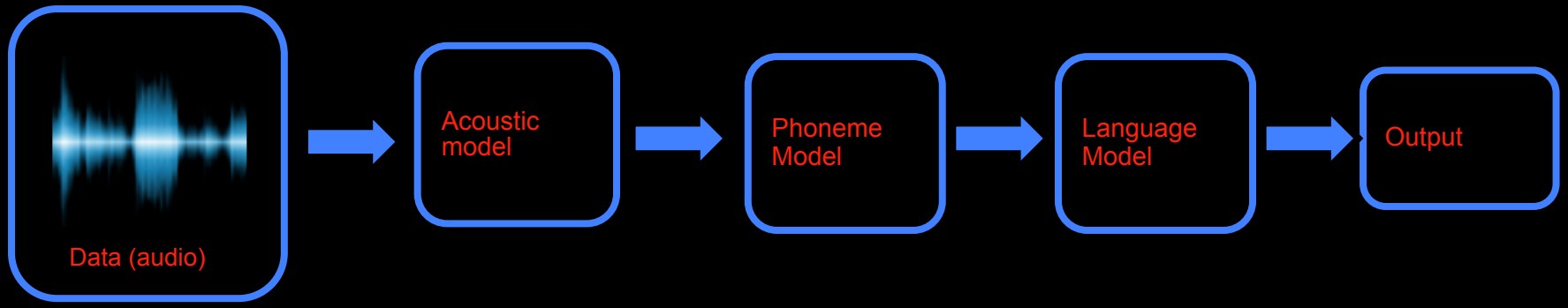
Source : http://googleblog.blogspot.com/2014/10/omg-mobile-voice-survey-reveals-teens.html

State of Speech

History of Speech Recognition



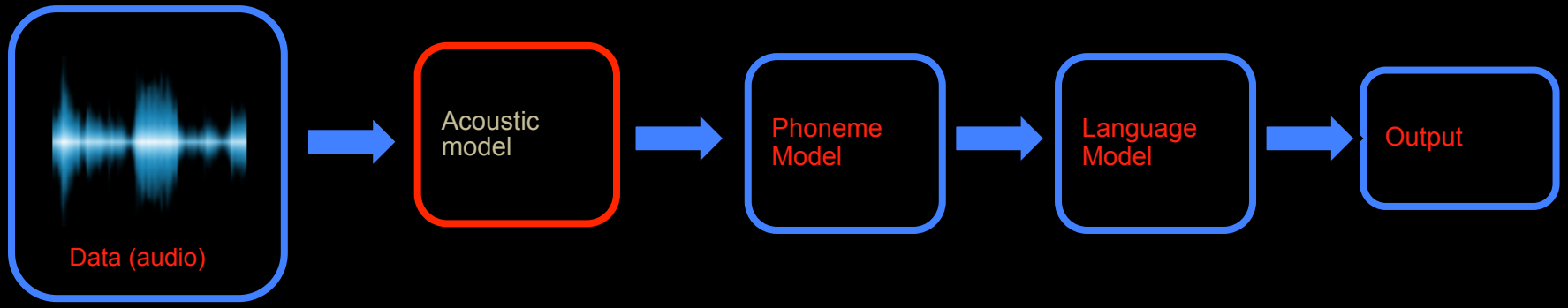
Speech Recognition Today



ðə kwɪk brawn
fɑks dʒəmps
ovər ðə lezi dɒg.

“The quick brown fox jumps over the lazy dog”

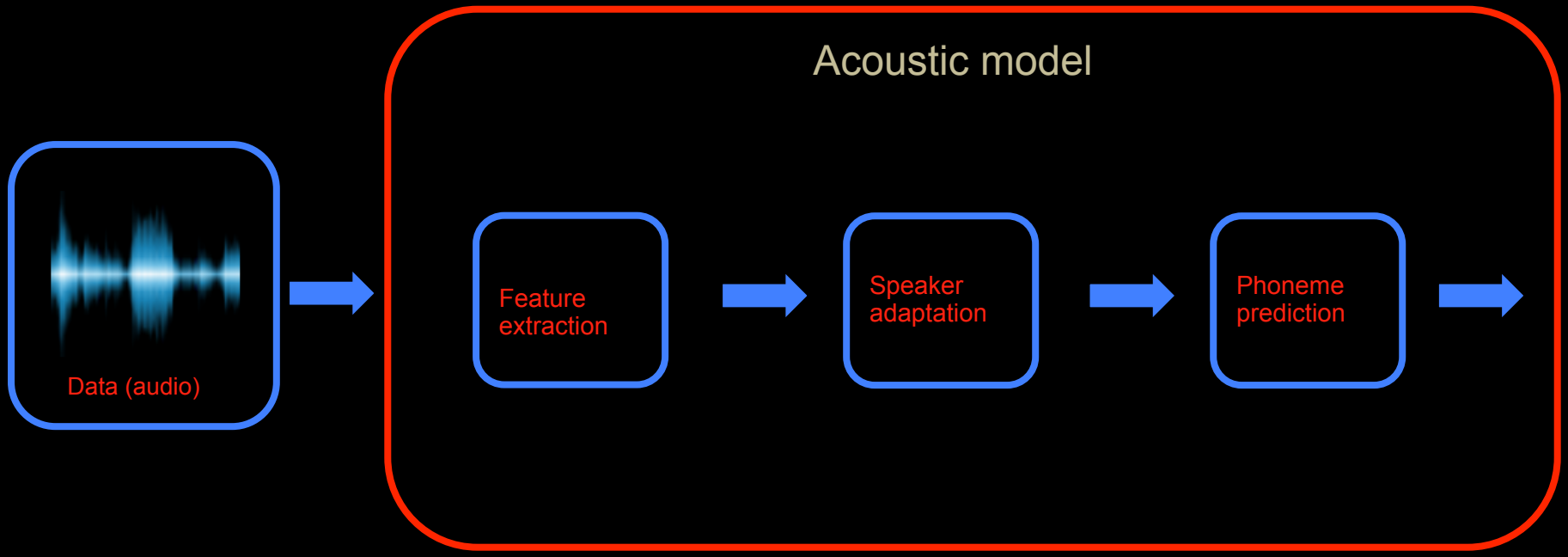
Speech Recognition Today



ðə kwɪk brawn
fɑks dʒəmps
ovər ðə lezi dɒg.

“The quick brown fox jumps over the lazy dog”

Speech Recognition Today



Outline

- State of Speech Recognition
- Overview: Deep Learning
- Deep Speech
- Next Steps

Aside: Supervised Learning

Goal: Learn to recognize a coffee mug

Aside: Supervised Learning

Coffee mug

Coffee mug

Coffee mug

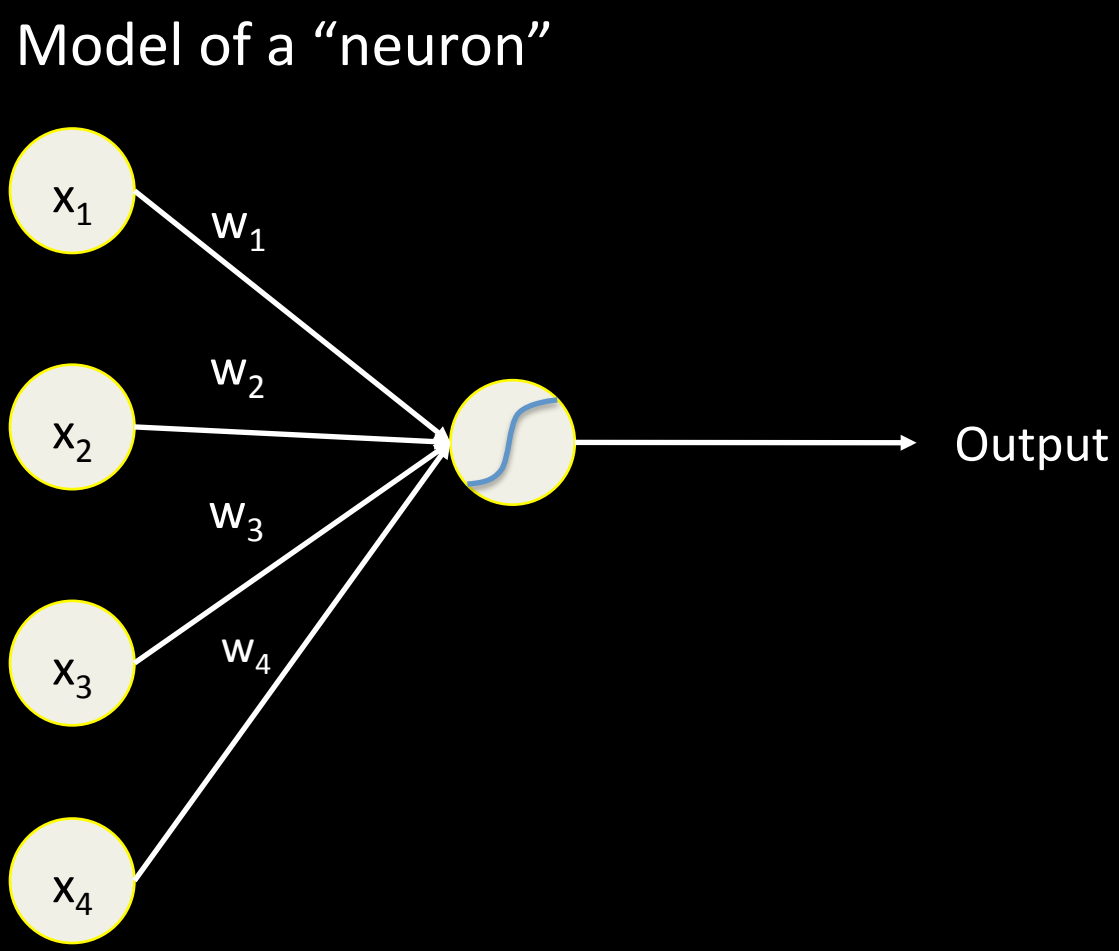
Coffee mug

Coffee mug

Coffee mug

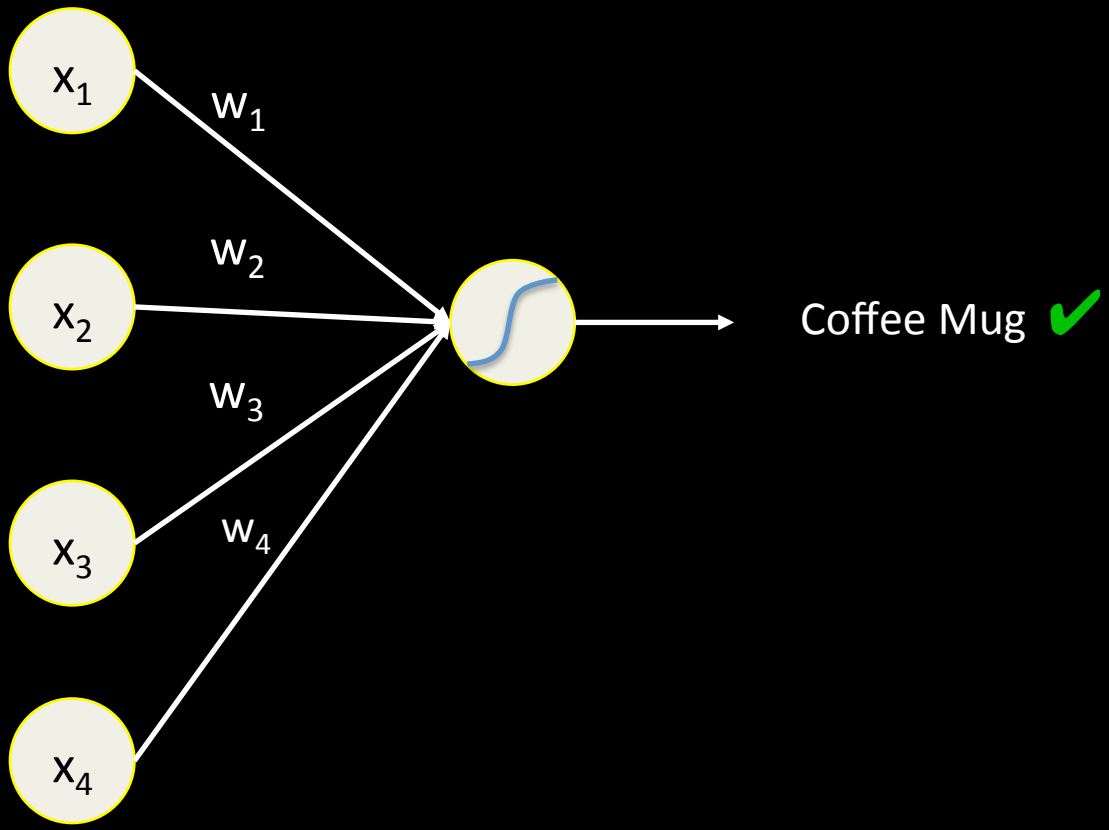
Testing: What is this?

Deep Learning: Overview



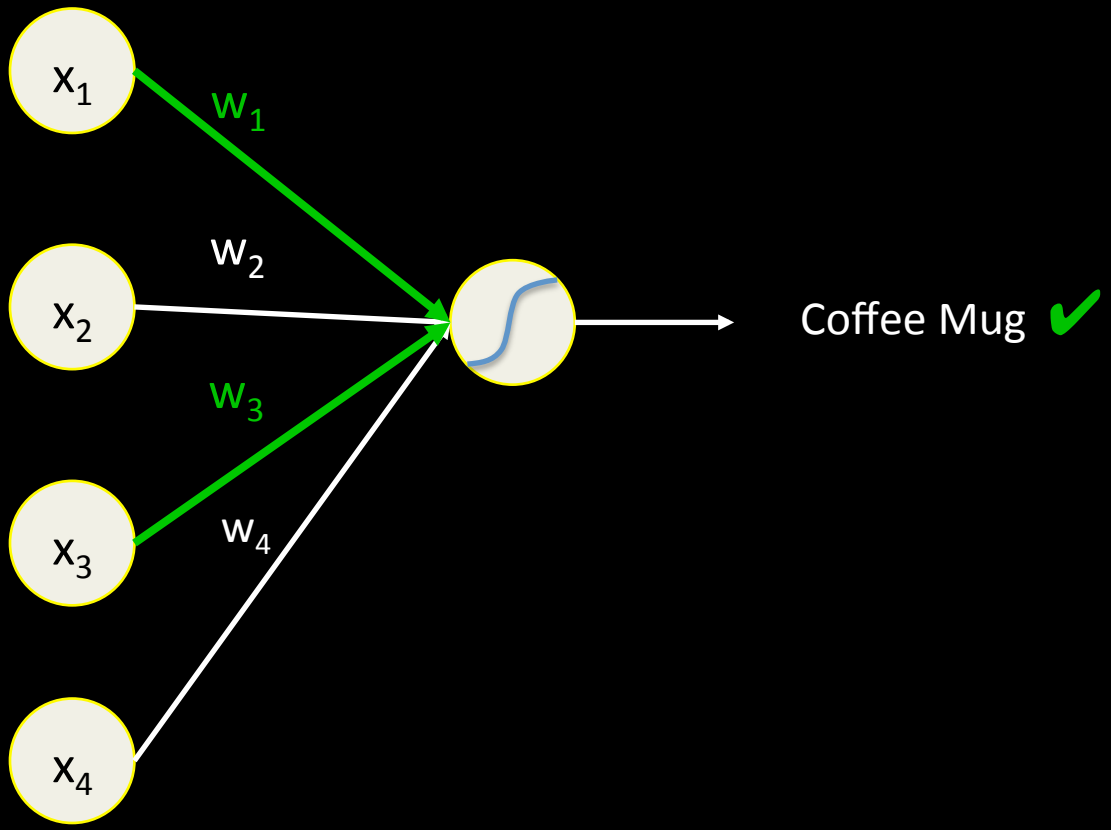
Deep Learning: Overview

Learning with a “Neuron”



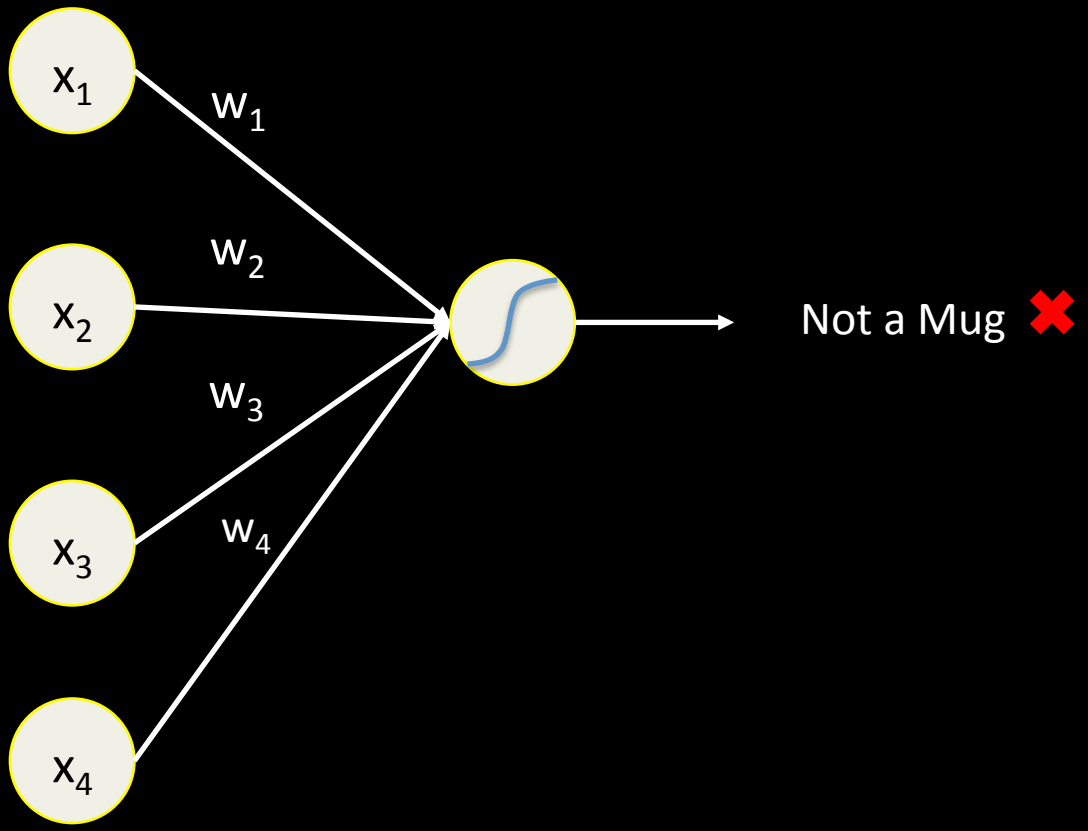
Deep Learning: Overview

Learning with a “Neuron”



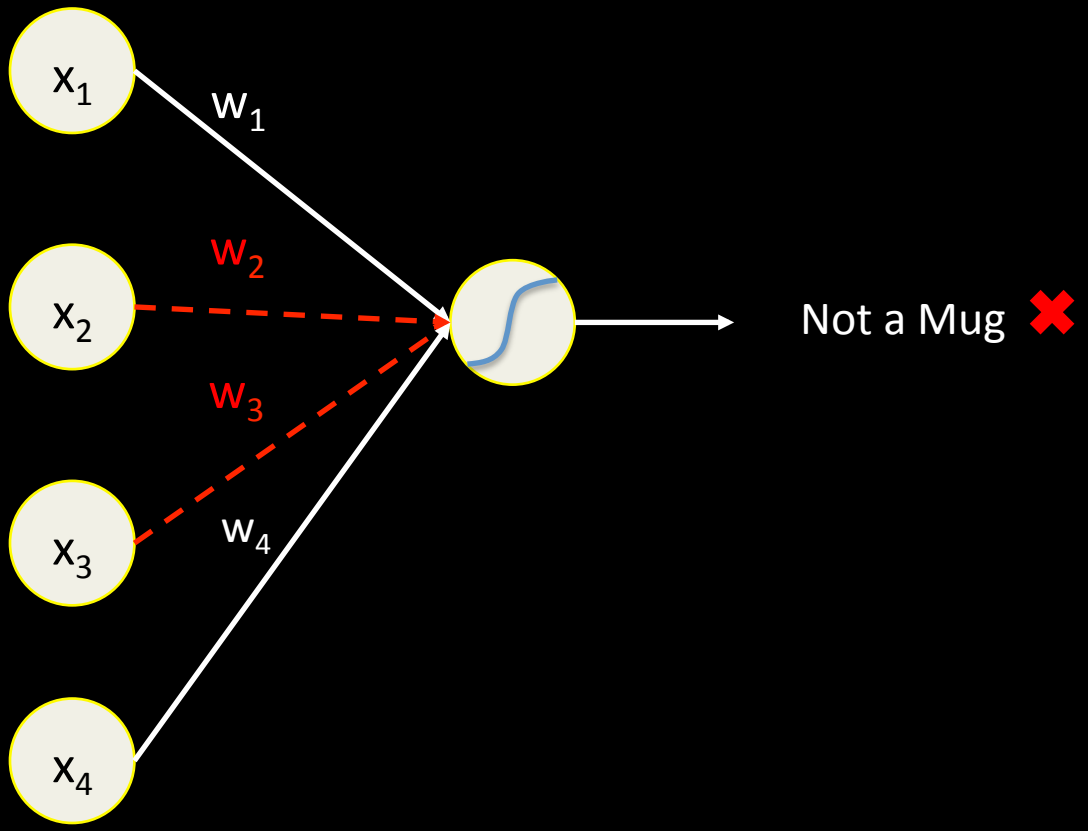
Deep Learning: Overview

Learning with a “Neuron”



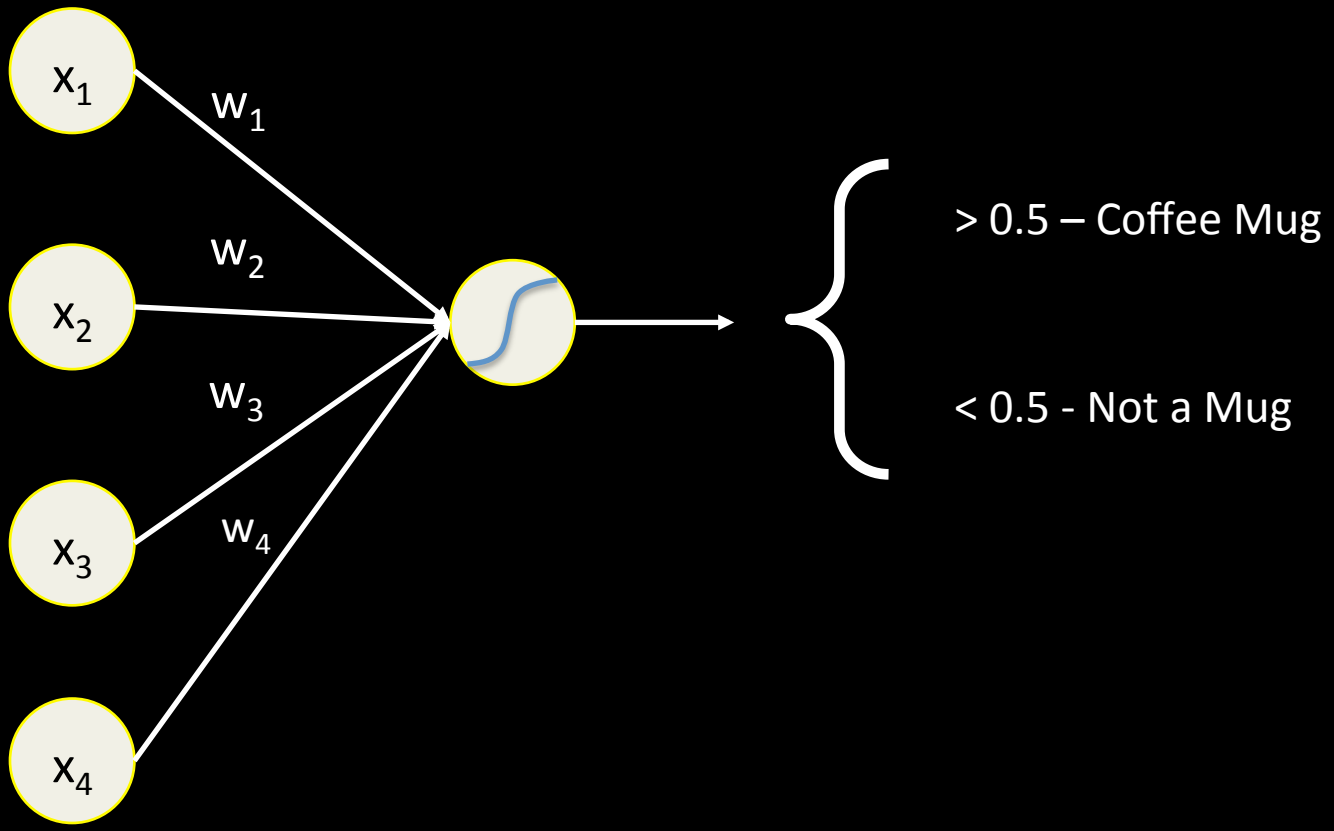
Deep Learning: Overview

Learning with a “Neuron”



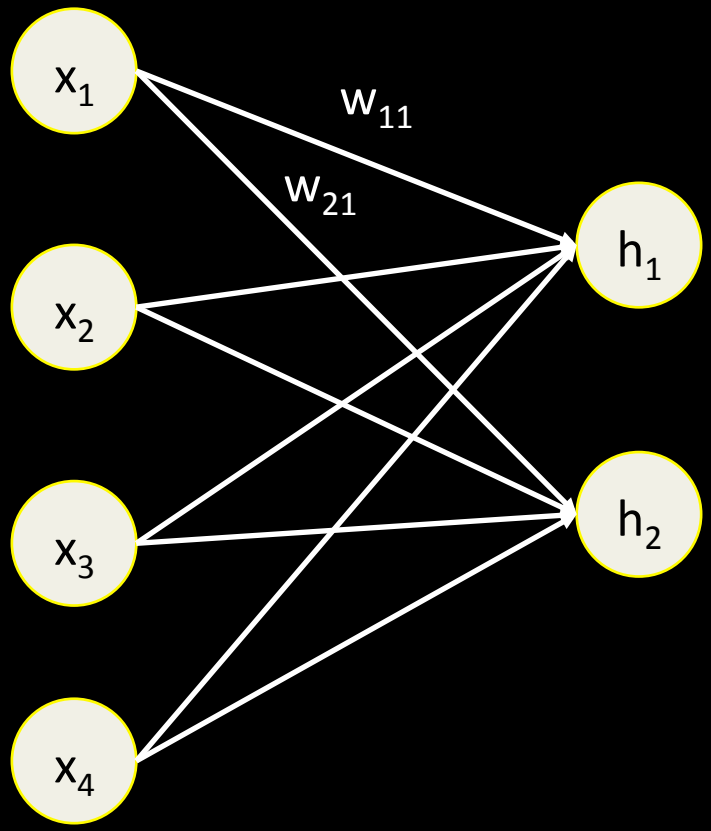
Deep Learning: Overview

Prediction with a "Neuron"

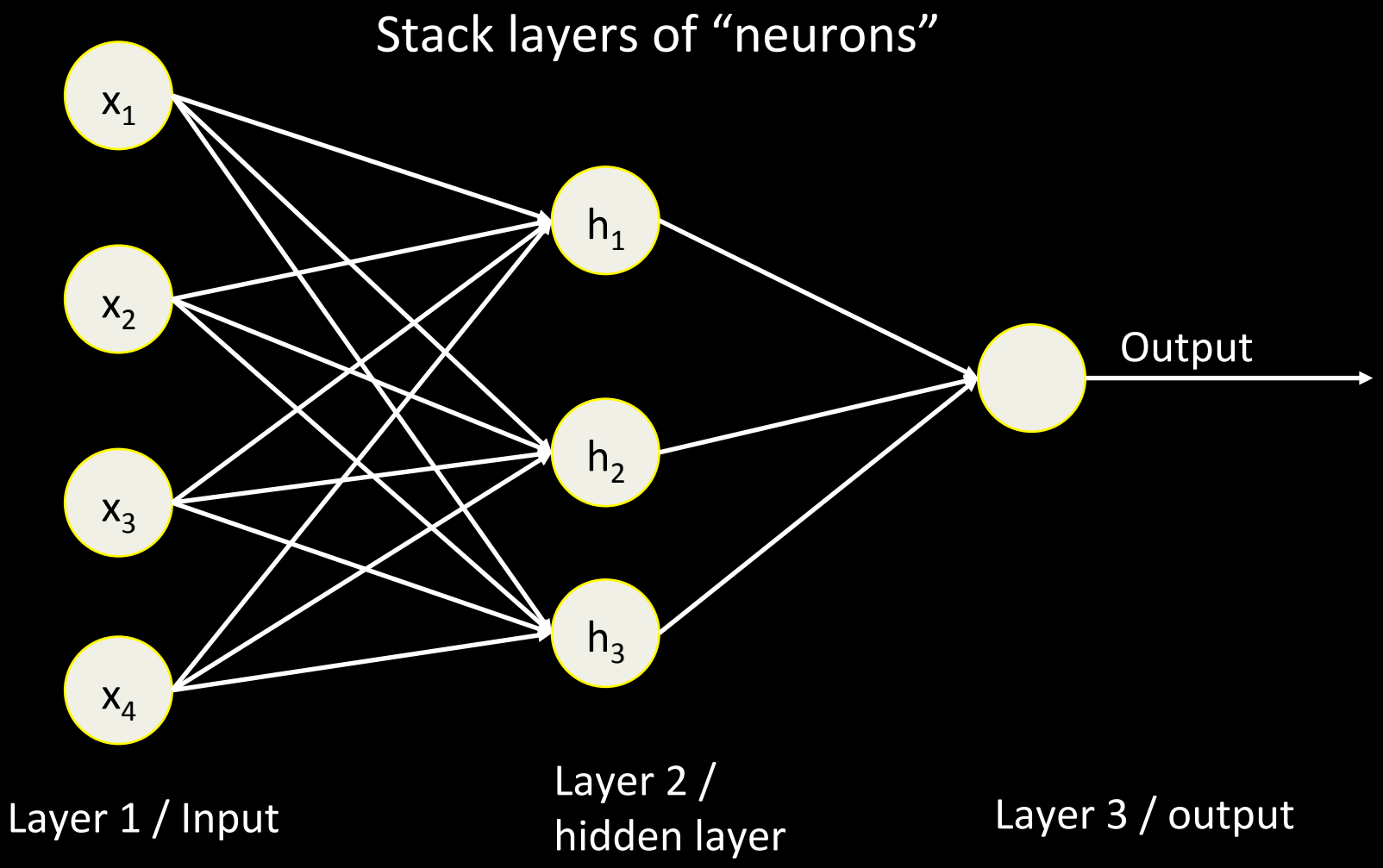


Deep Learning: Overview

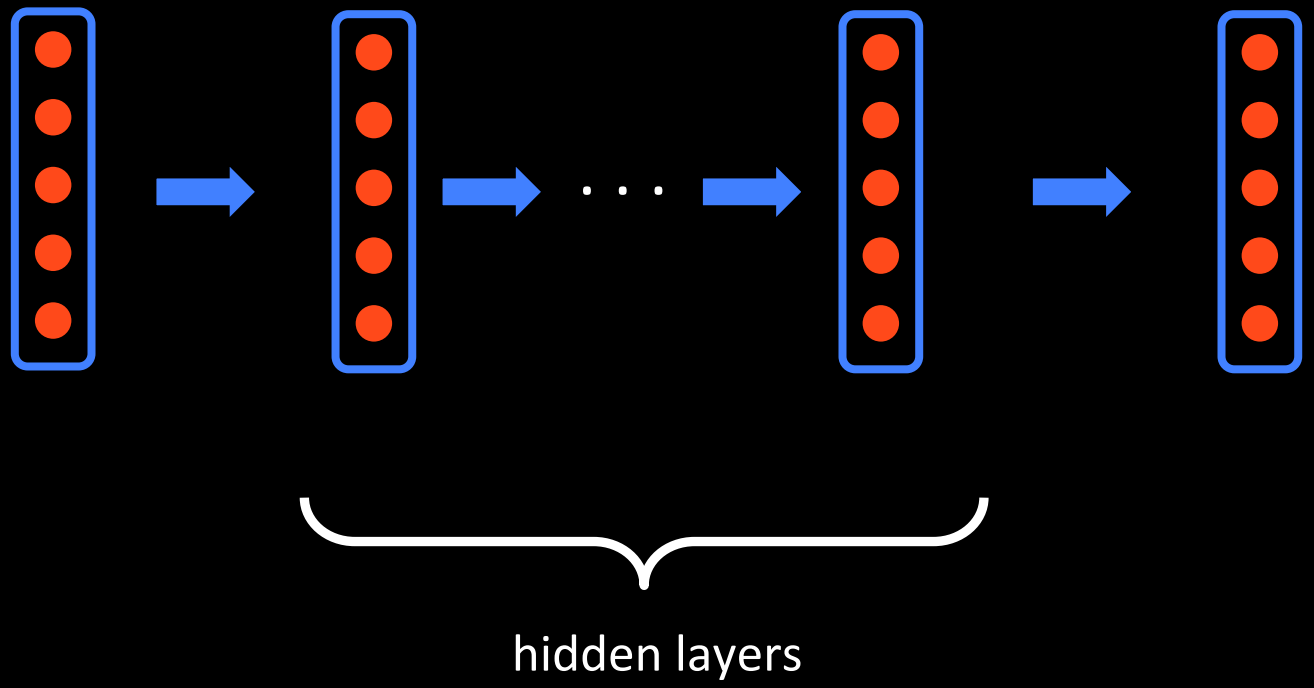
Wire many “neurons” together



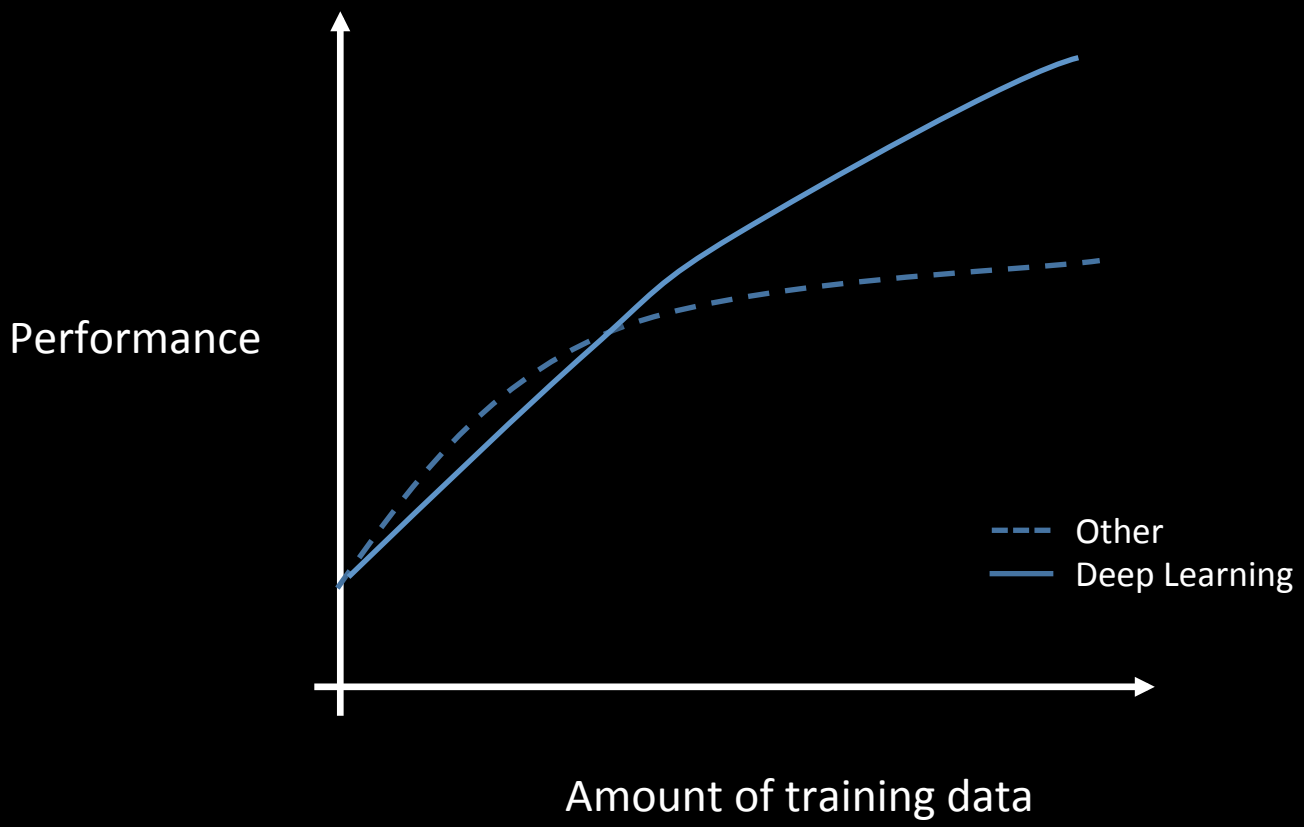
Deep Learning: Overview



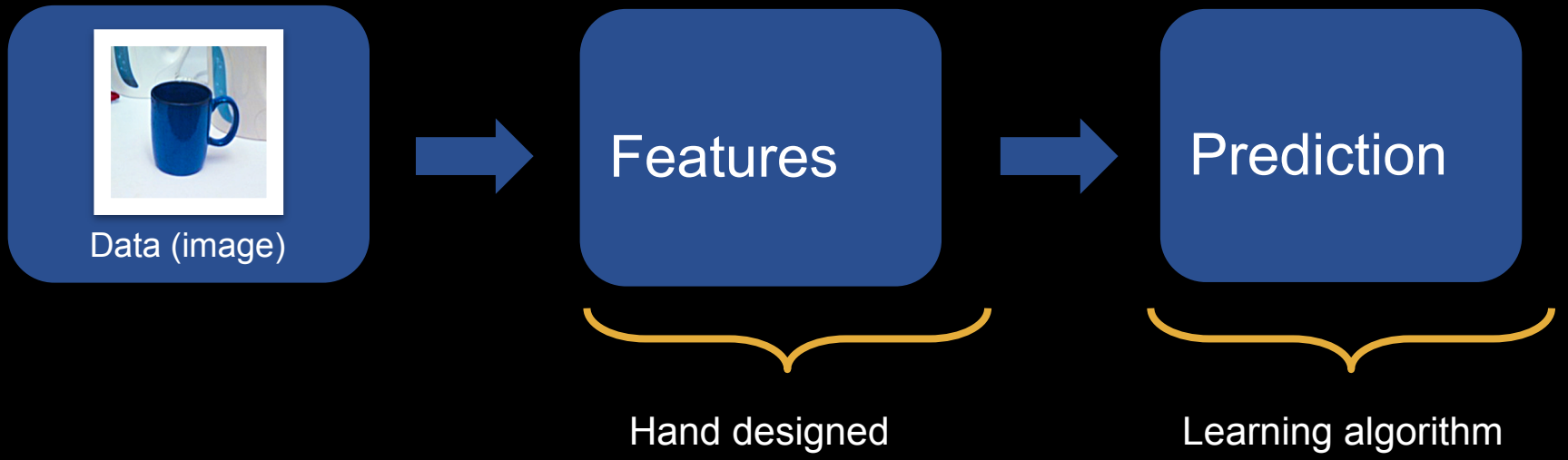
Deep Learning: Overview



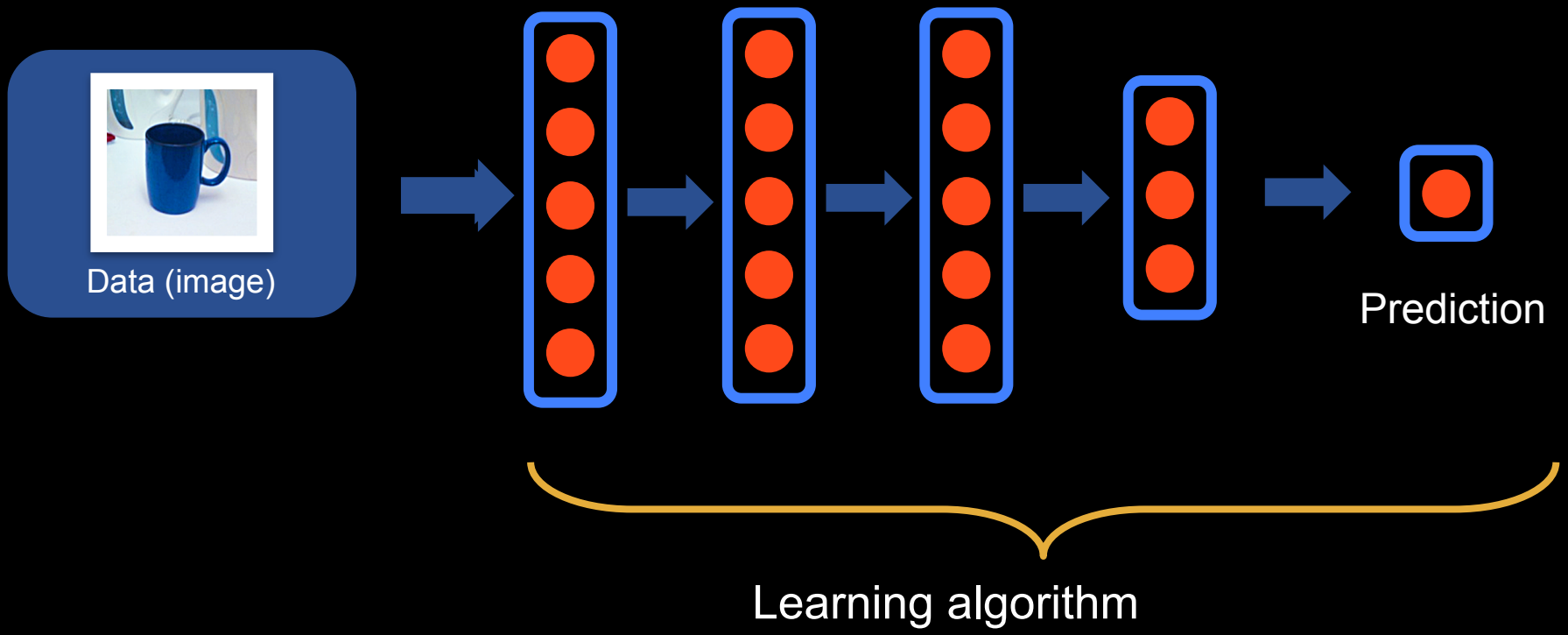
Why Deep Learning?



Computer vision: Hand designed features

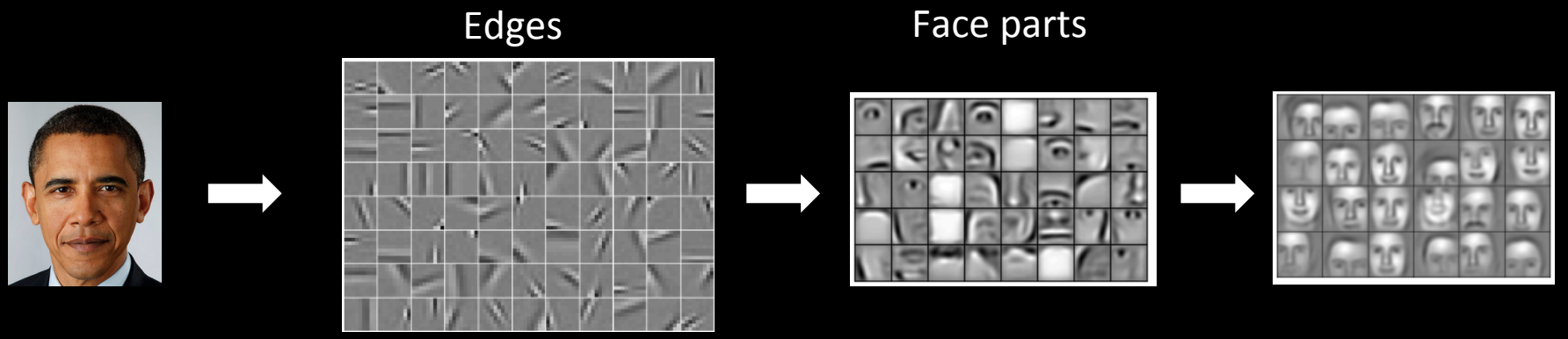


Computer vision: Hand designed features



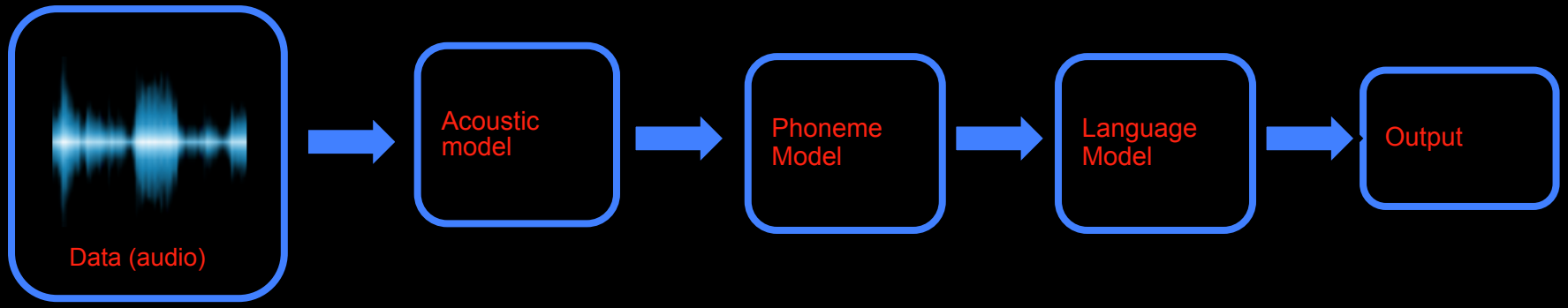
Deep Learning: Computer Vision

Computer Vision – feature learning



Images from: Lee, H. et al. *Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks*. Communications of the ACM, vol. 54, no. 10, pp. 95-103, 2011

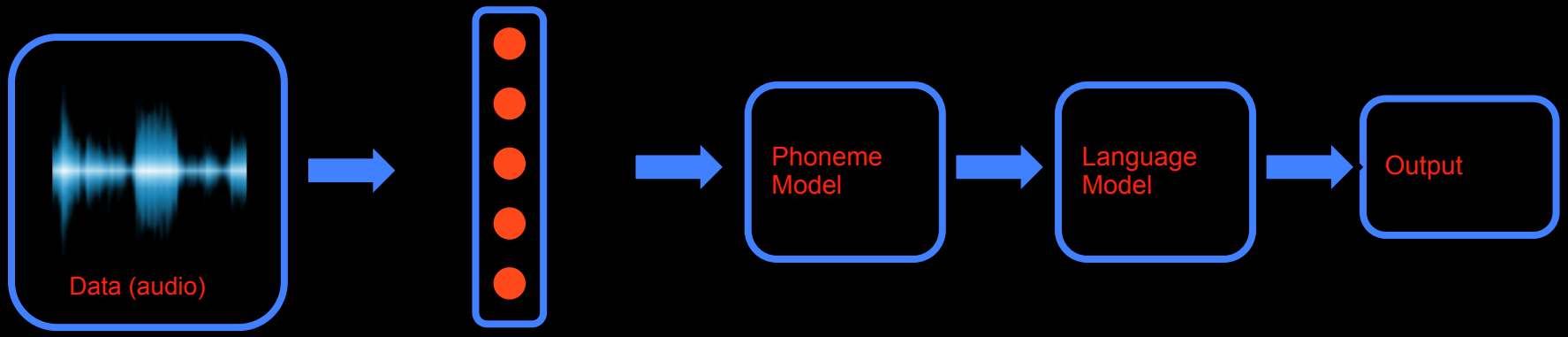
Deep Learning: Speech Recognition



ðə kwɪk brawn
fɑks dʒəmps
ovər ðə lezi dɒg.

“The quick brown fox jumps over the lazy dog”

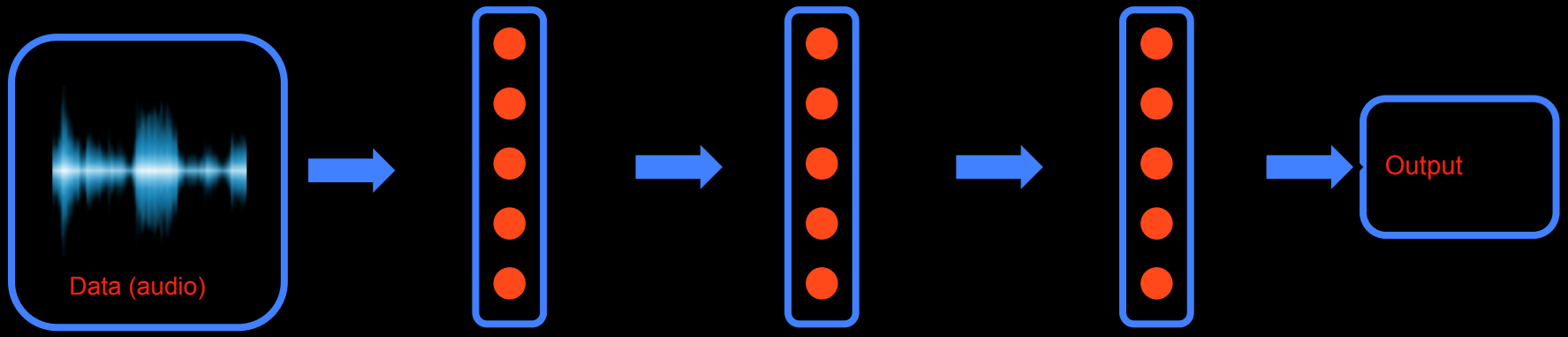
Deep Learning: Speech Recognition



ðə kwɪk brawn
fɑks dʒəmps
ovər ðə lezi dɒg.

“The quick brown fox jumps over the lazy dog”

Deep Learning: Speech Recognition



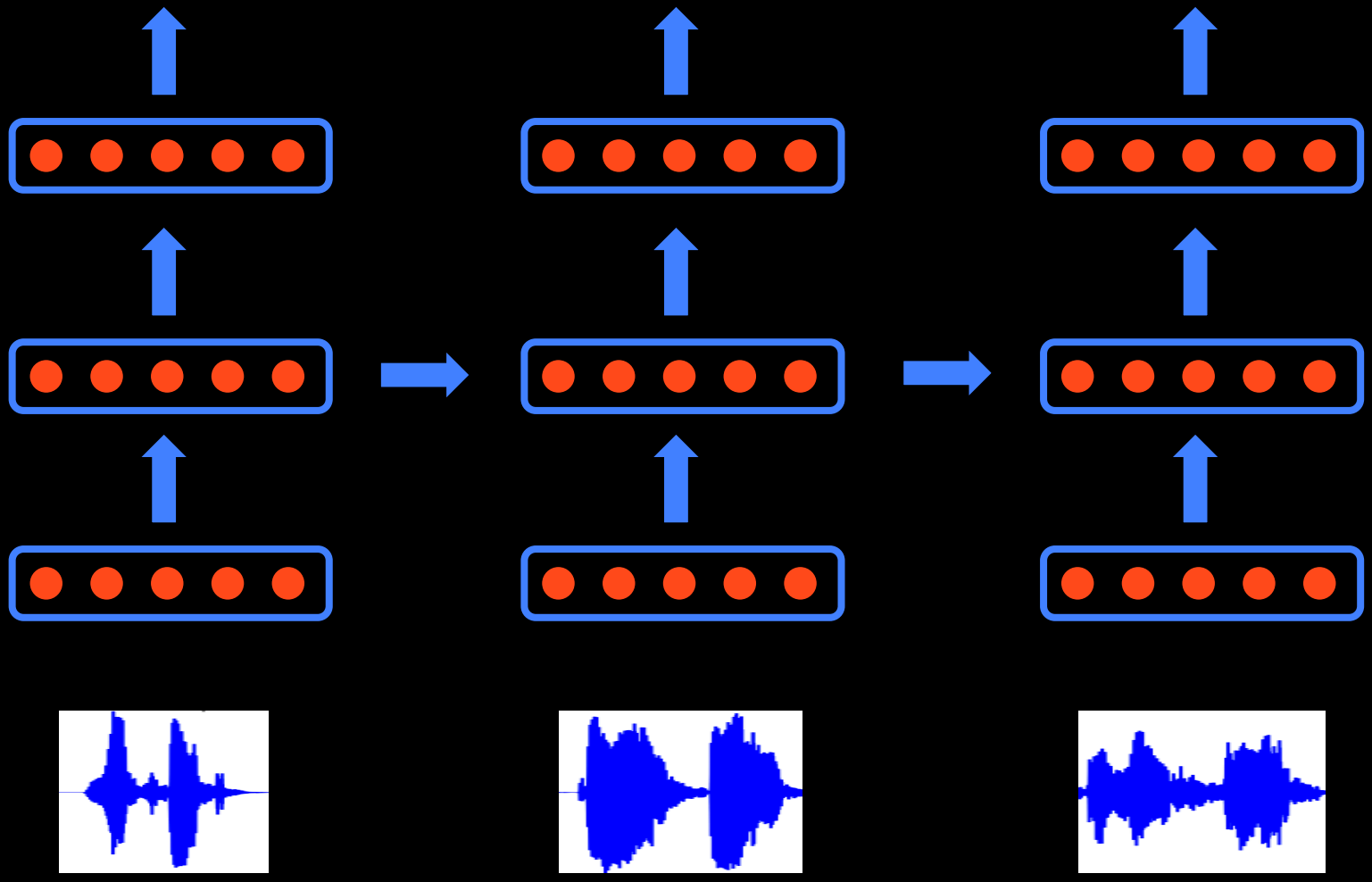
ðə kwık brawn
fɑks dʒəmps
ovər ðə lezi dɒg.

“The quick brown fox jumps over the lazy dog”

Deep Learning: Time series

- Variable Length problem
  - Solution 1: Make everything the same length
  - Solution 2: A model which can handle variable length inputs

Deep Learning: Recurrent Neural Network



Deep Learning - Challenges

Data

- Supervised (labeled data)
  - Real
  - Synthesized
- Unsupervised (unlabeled data)

Deep Learning - Challenges

Data – Real

- Benchmarks

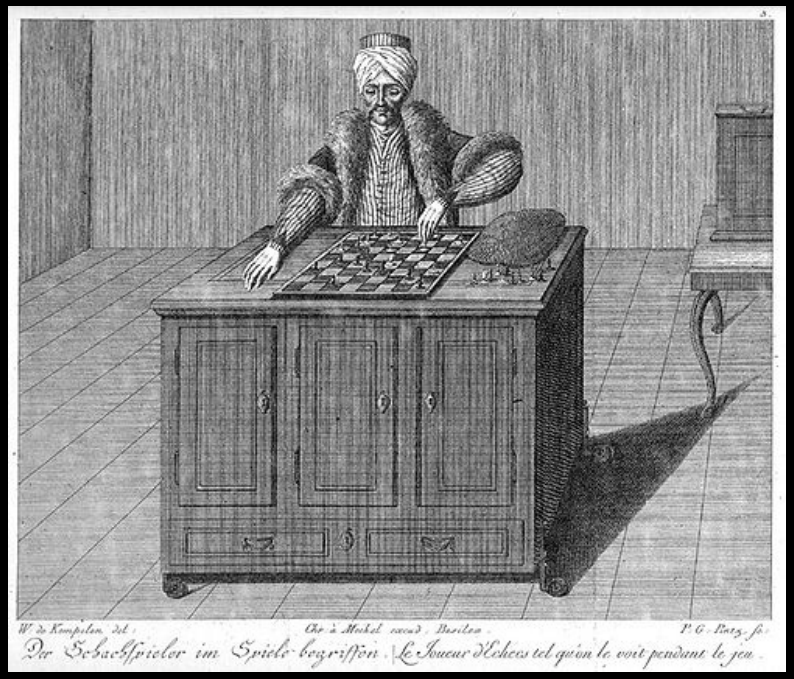
  >14 million images

- Big companies (Baidu, Amazon, Google, Microsoft, …)

- Mechanical Turk

Deep Learning - Challenges

Mechanical Turk



Deep Learning - Challenges

Data – Synthetic

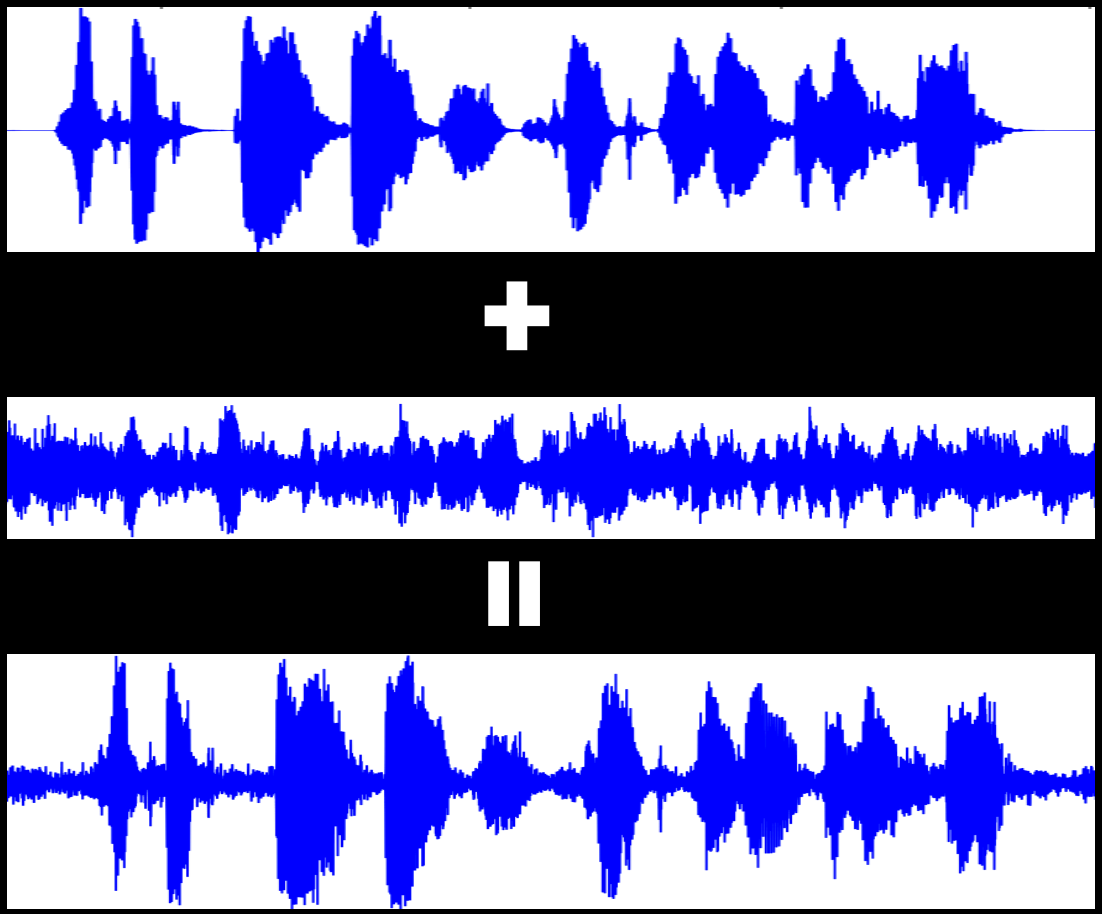
House

Translated House

Reflected House

Rotated house

Deep Learning - Challenges



Speech

Noise

Noisy Speech

Deep Learning - Challenges

Scale

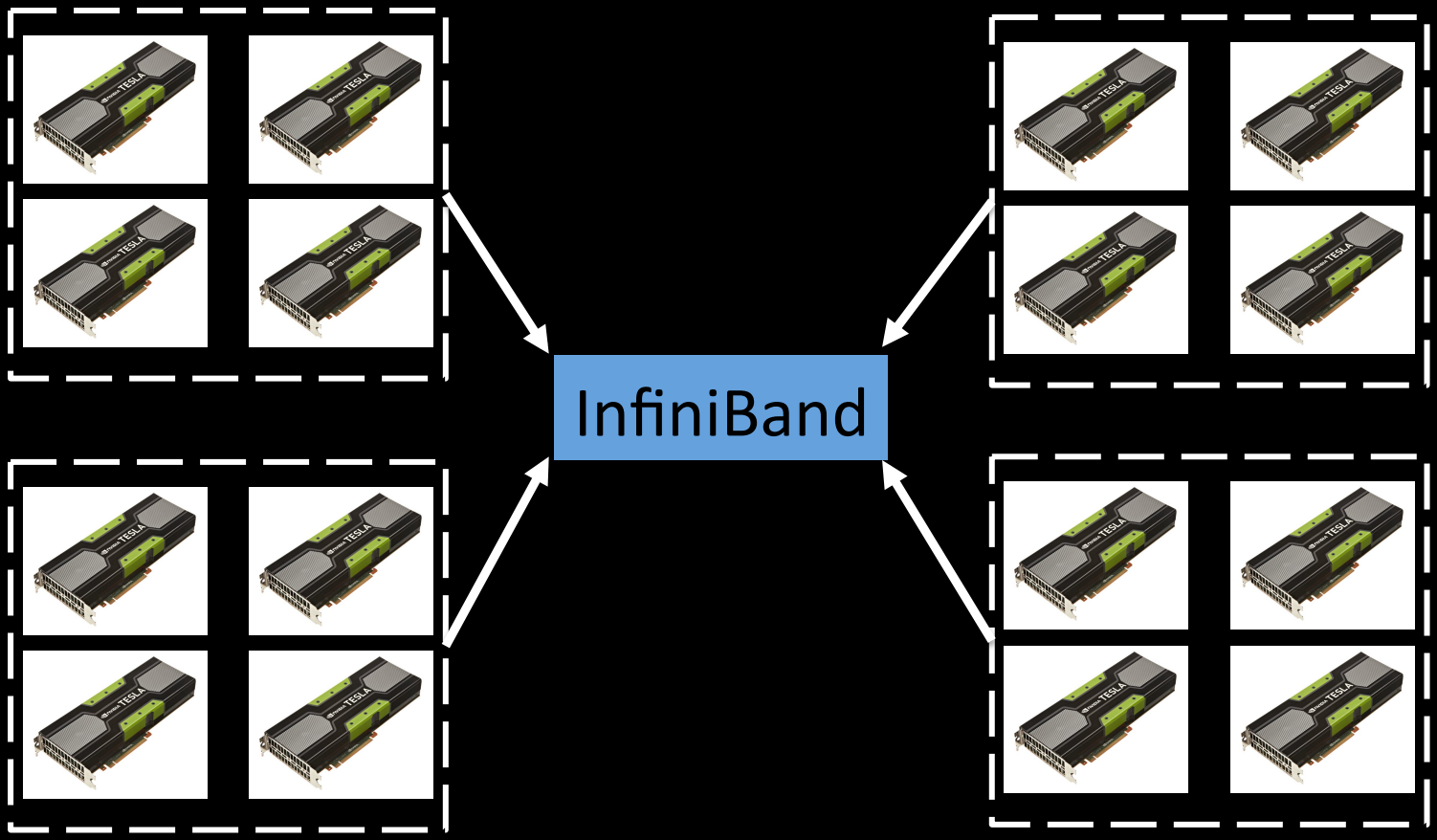
- More data ⟶ Longer training time
- Bigger model ⟶ Longer training time

Deep Learning - Challenges

CPU
~100 Giga flops

GPU
~5 Tera flops

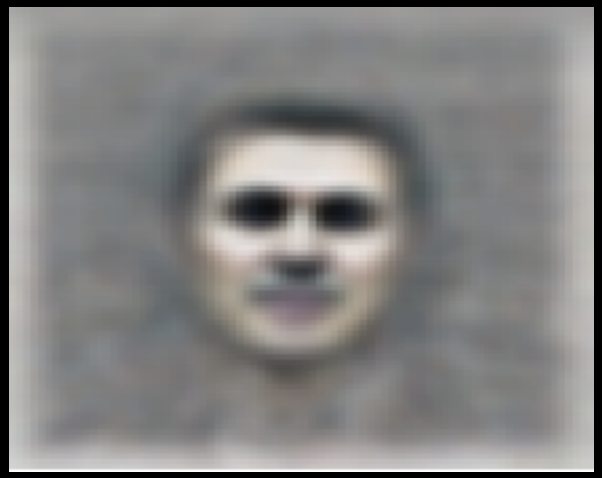
Deep Learning - Challenges



Deep Learning - Challenges

Google Brain

- 1+ billion parameters
- Many billions of connections
- 16 thousand CPU cores



Source: Le, Q. et. al., Building high-level features using large scale unsupervised learning. ICML, 2012.

Deep Learning - Challenges

- Google Brain
  - Billion parameters
  - Millions of images
  - 16 Thousand CPUs ($Millions)
  - Several days to train
- Coates et al. (COTS HPC)
  - Billion(s) parameters
  - Millions of images
  - 12 GPUs ($Thousands)
  - Several days to train

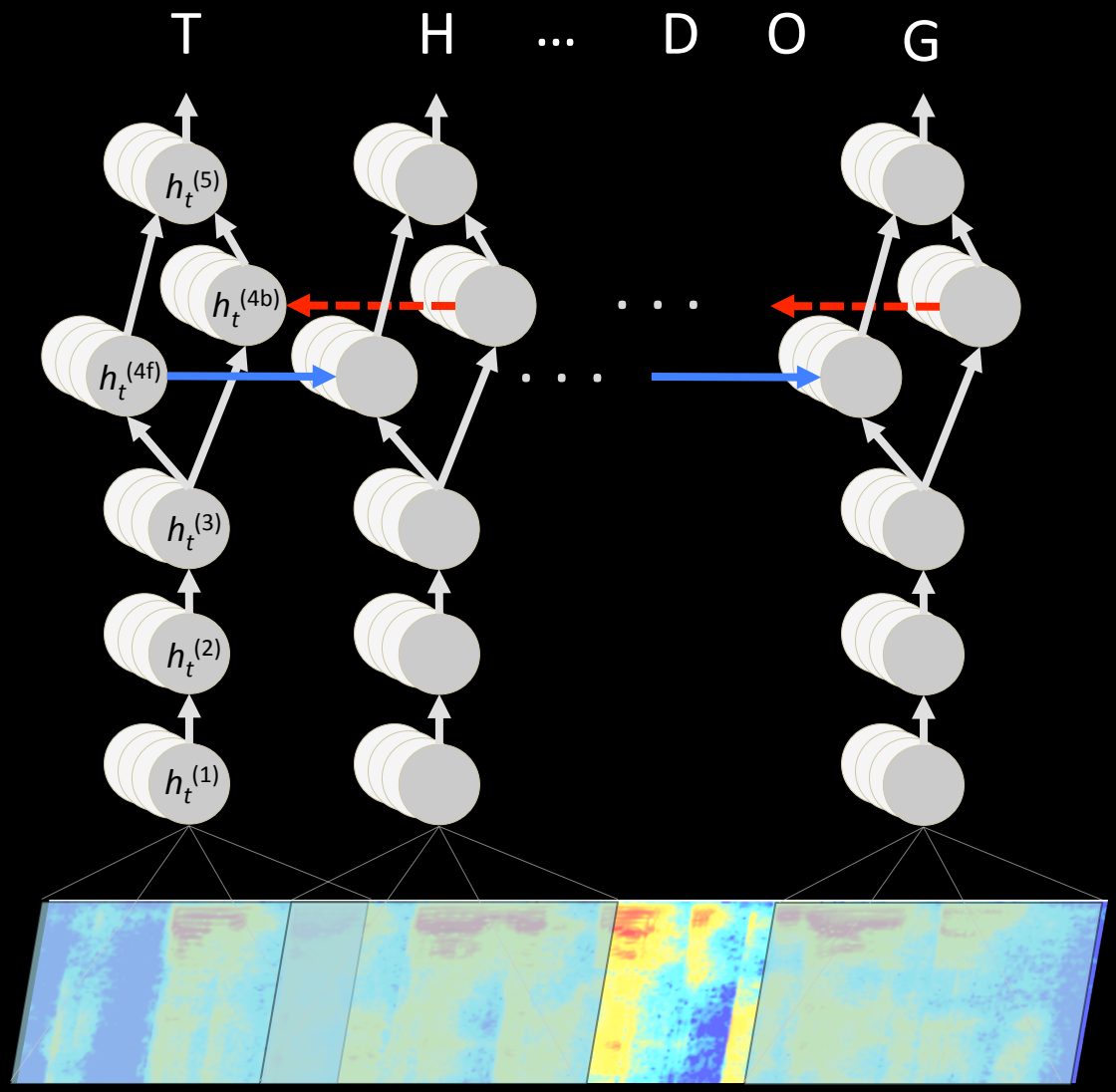
Outline

- State of Speech Recognition
- Overview: Deep Learning
- Deep Speech
- Next Steps

Deep Speech – Key ingredients

- Data
  - No alignment needed, using objective from [Graves, Fernandez, Gomez and Schmidhuber, 2006]
- Computation (GPU)

Deep Speech



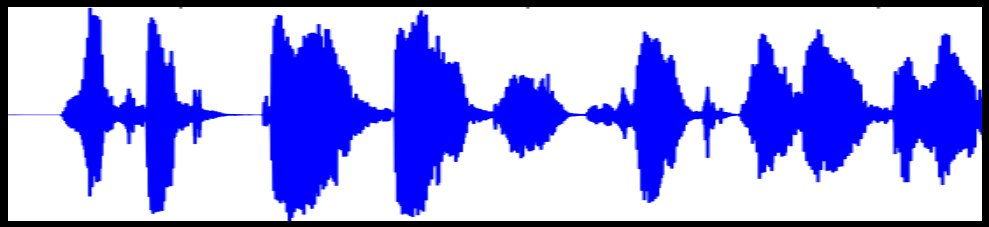
Deep Speech

No alignment needed!

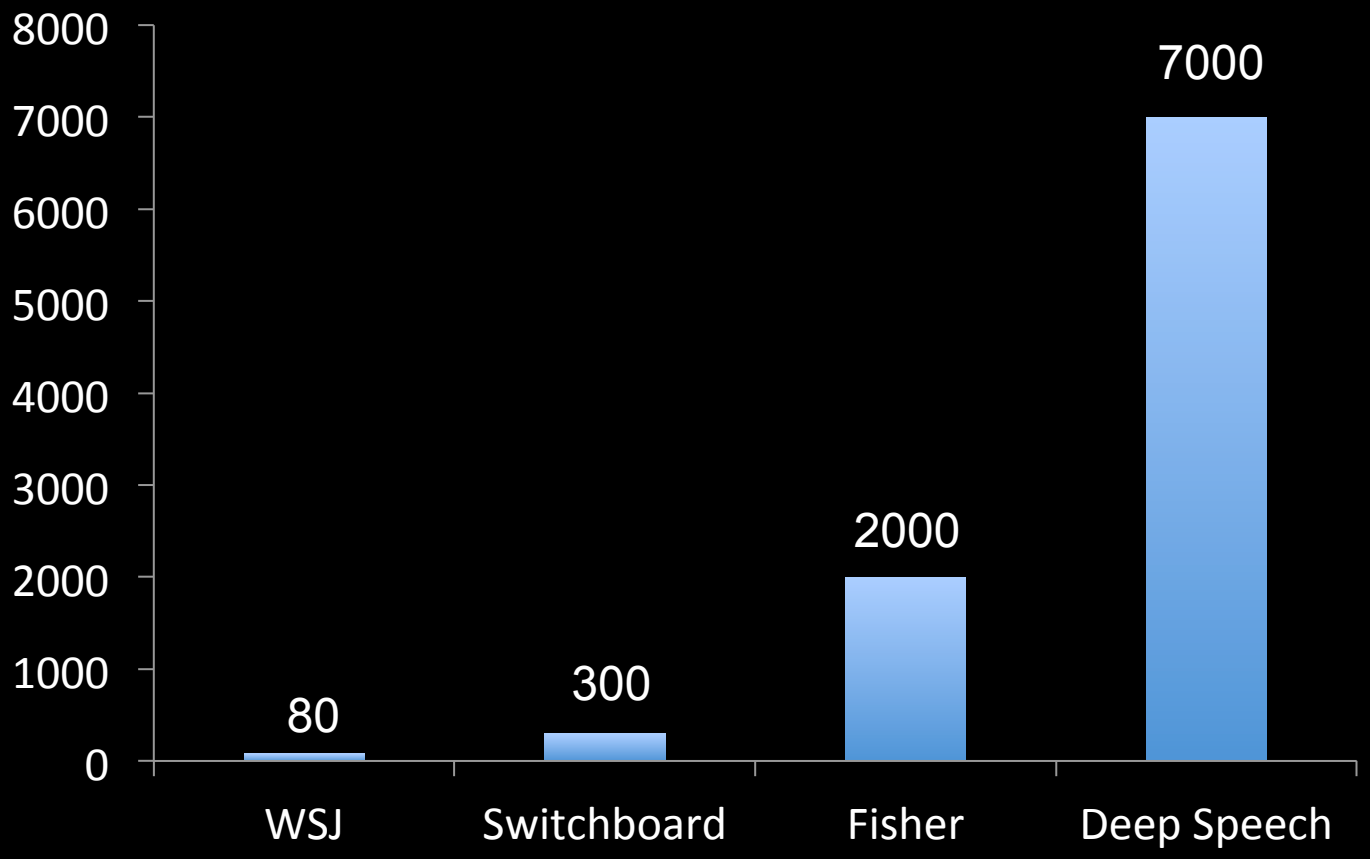
_ _ T H _ _ _ _ _ E _ _ – _ C _ _ A A A _ _ T T _ _ –

_ T _ _ _ H _ _ E E _ _ – _ C _ _ A A _ _ T _ _ _ _ –

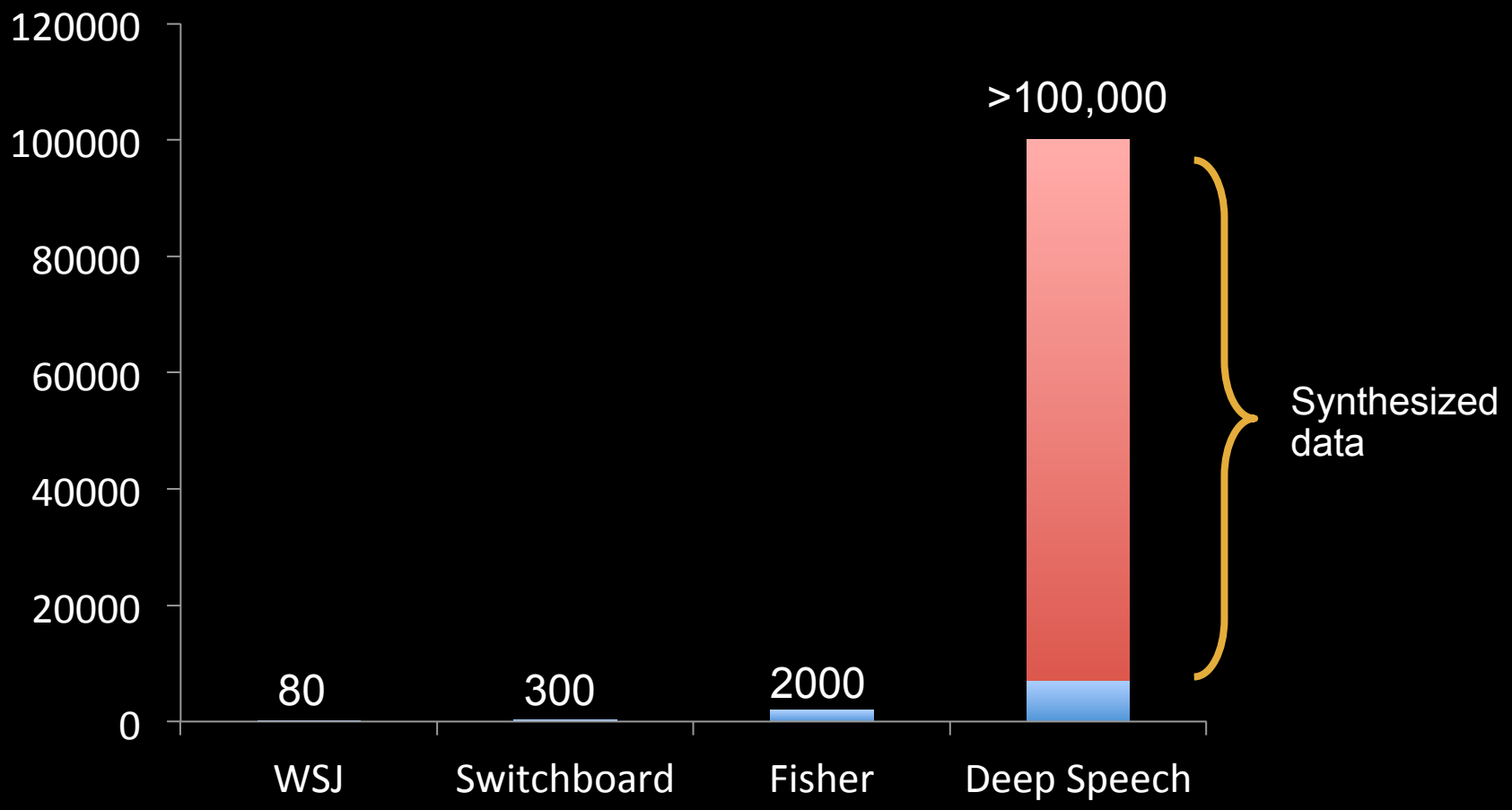
THE—CAT—



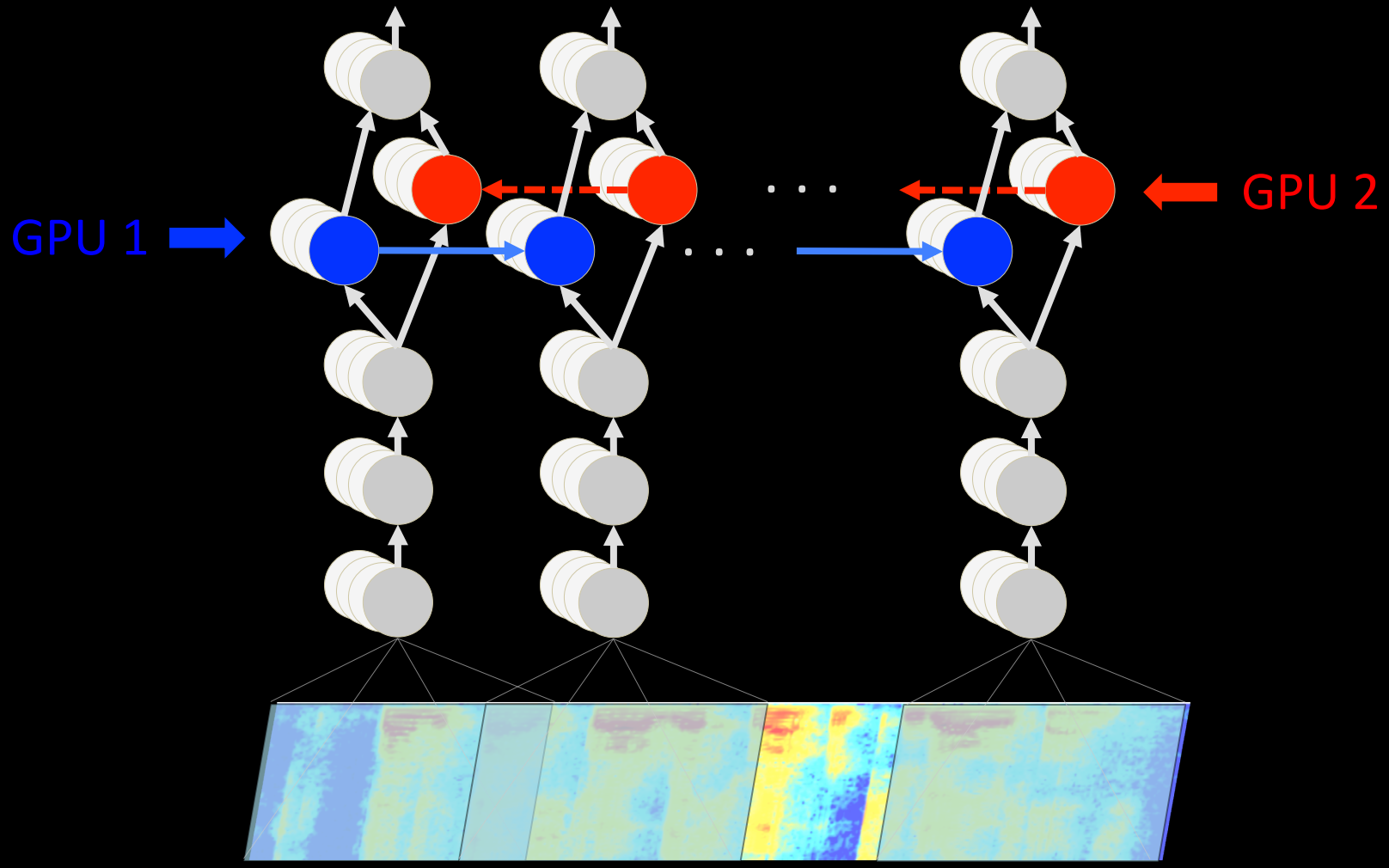
Deep Speech - Hours of speech data



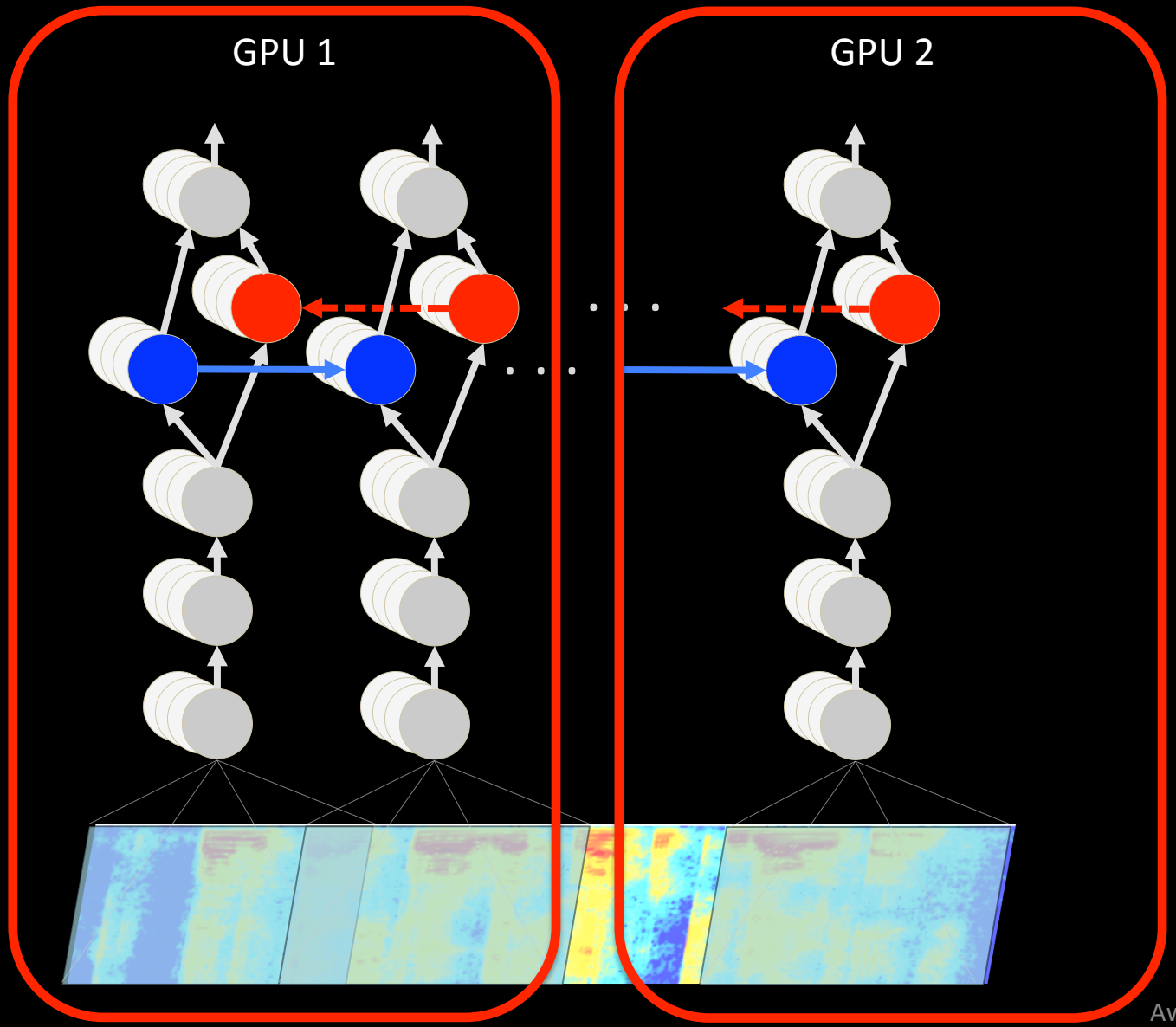
Deep Speech - Hours of speech data



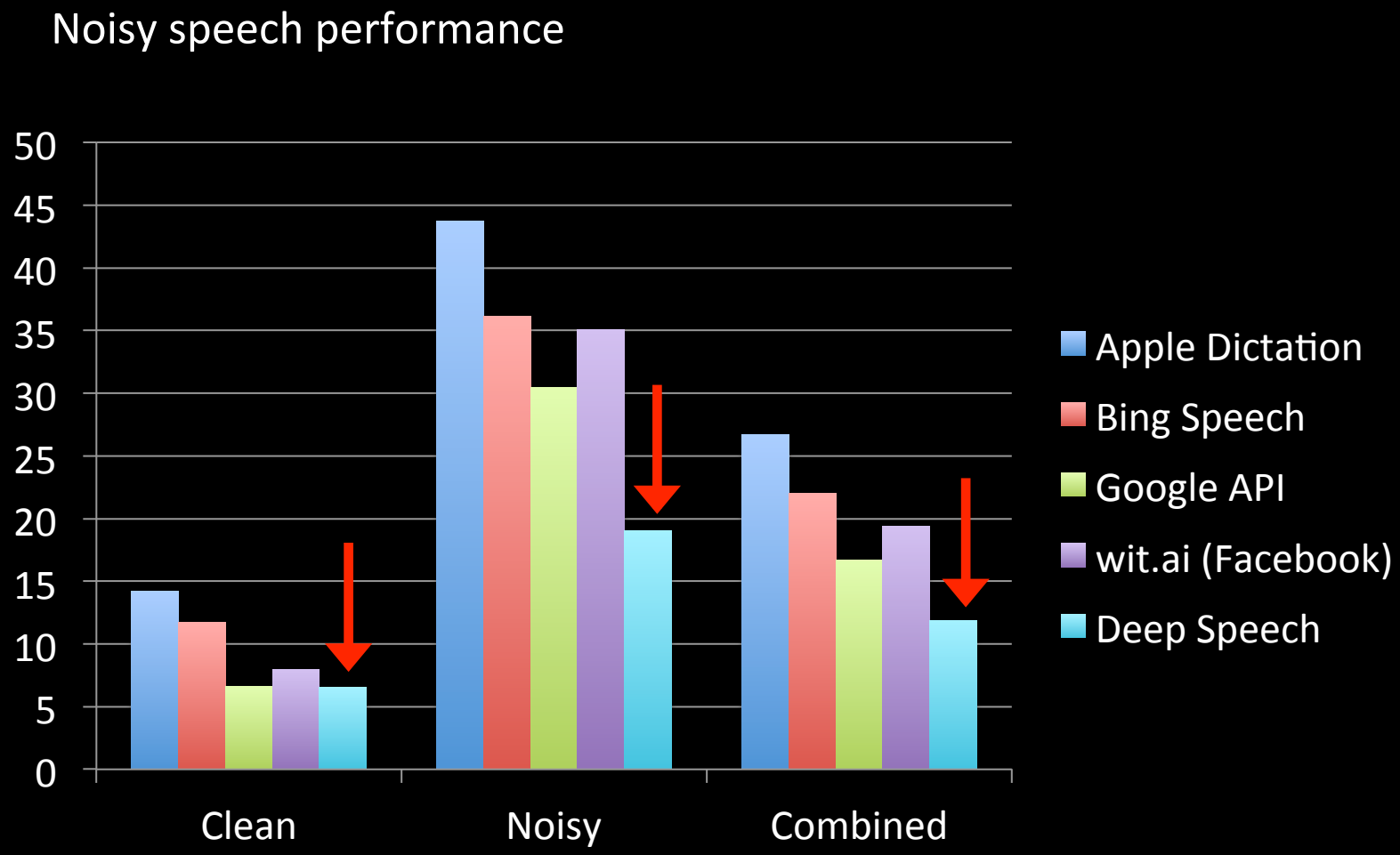
Deep Speech – GPU scaling



Deep Speech – GPU scaling



Deep Speech


Noisy speech performance
50
45
40
35
30
25
20
15
10
5
0
Clean
Noisy
Combined
Apple Dictation
Bing Speech
Google API
wit.ai (Facebook)
Deep Speech

Outline

- State of Speech Recognition
- Overview: Deep Learning
- Deep Speech
- Next Steps

Next Steps

- What will it take to have machines transcribe as well as humans?
  - More data & bigger models
  - Algorithmic innovations (?)
  - Unsupervised learning

Learn More

- Machine learning - CS229 (or ML Coursera class)
- Stanford UFLDL tutorial - http://deeplearning.stanford.edu/tutorial/
- Reading
  - http://ufldl.stanford.edu/wiki/index.php/UFLDL_Recommended_Readings
  - http://deeplearning.net/tutorial/
  - Deep Speech – bit.ly/deepspeech

Baidu Research

- Follow our tech blog @ usa.baidu.com/index.php/category/baidutechblog/

- Jobs @ Baidu usa.baidu.com/index.php/jobs/

Work with …

Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, Andrew Y. Ng …