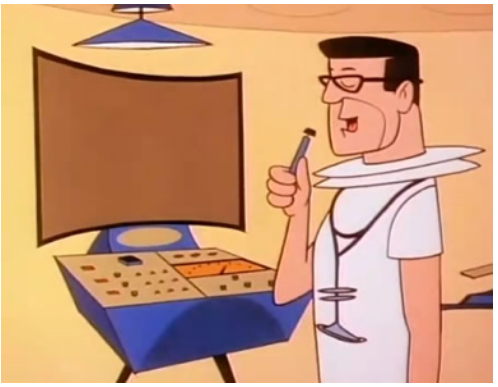
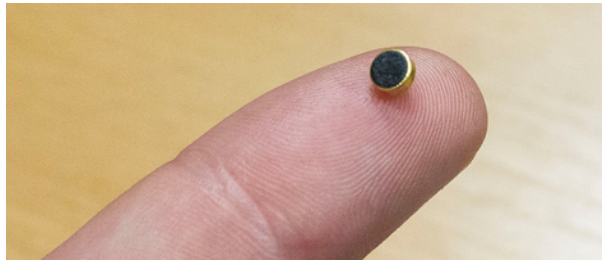
Deep Speech 2: End-to-End Speech Recognition in English and Mandarin

Baidu Silicon Valley AI Lab, Baidu Speech Technology Group [Presented by Eric Battenberg]

Speech Interfaces: A New/Old Input Paradigm

- Speaking is human
- Speed: speak 150 wpm vs type 40 wpm
- Hands-free
- Skip the confusing menus
- Small footprint
- The huge difference between 95% vs 99% accuracy



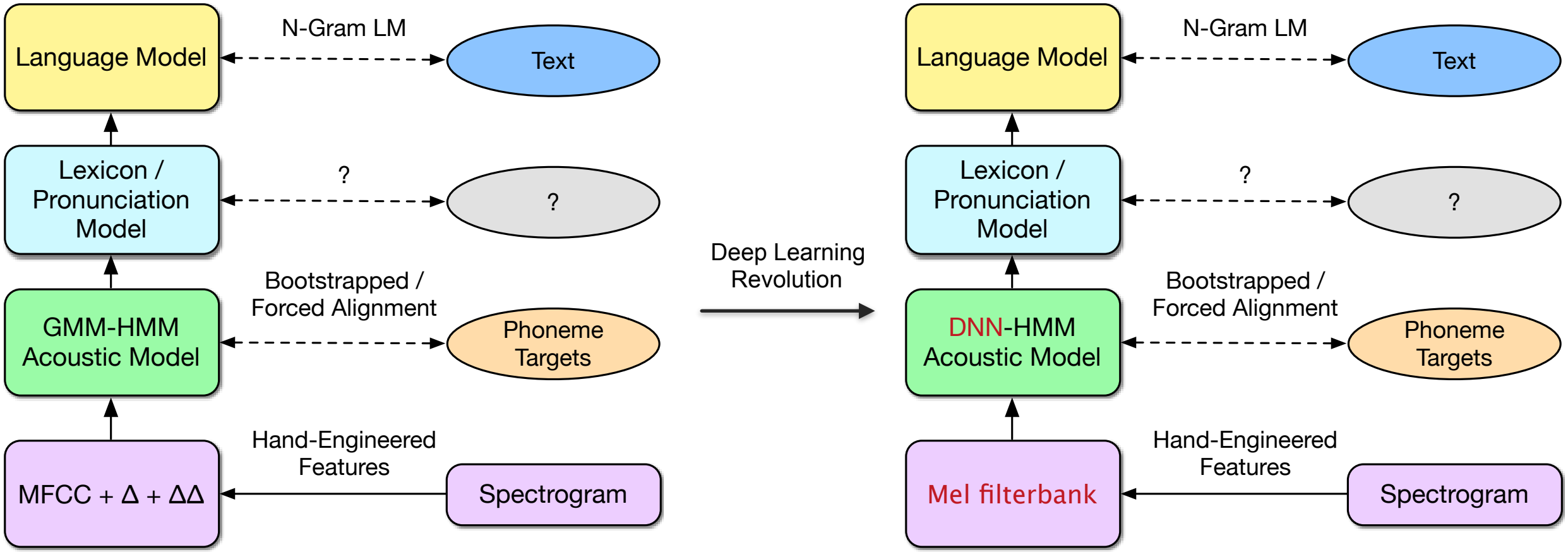


The Deep Learning Bet

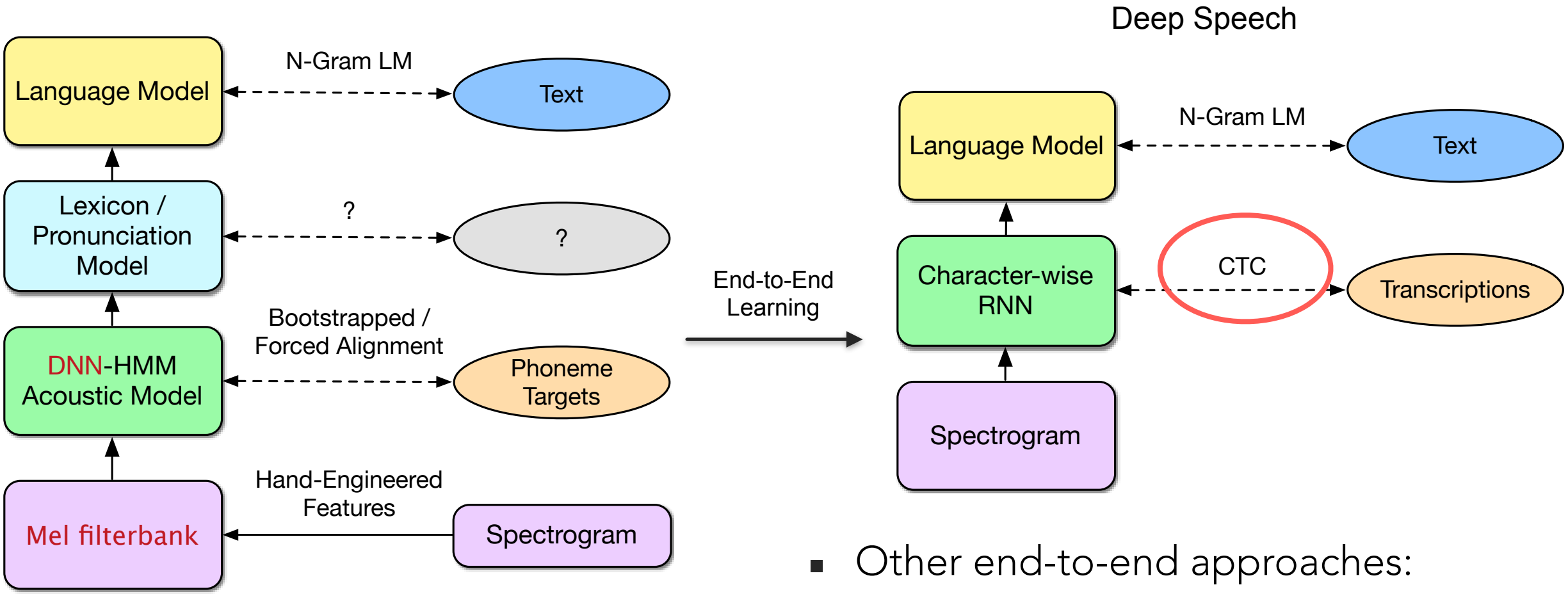
- SVAIL is an AI lab, not a speech lab.
- Instead of more domain expertise...
- Scale the model, scale the data
  - -> Improved performance.
- -> Superhuman Chinese speech recognition

SVAIL
Silicon Valley Artificial Intelligence Lab

The Evolution of Speech Recognition Systems



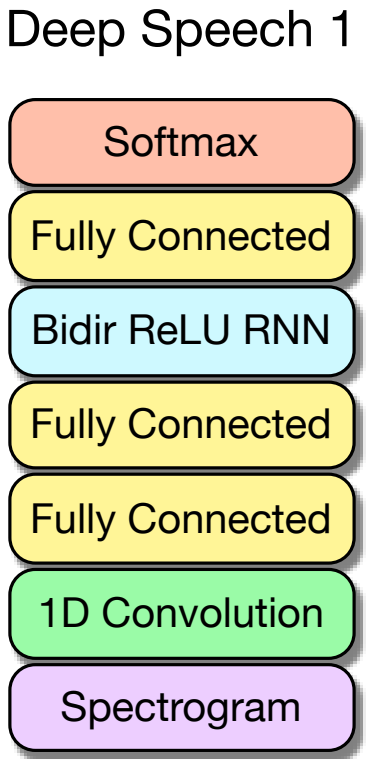
The Evolution of Speech Recognition Systems



- Other end-to-end approaches:
  - e.g., Attention

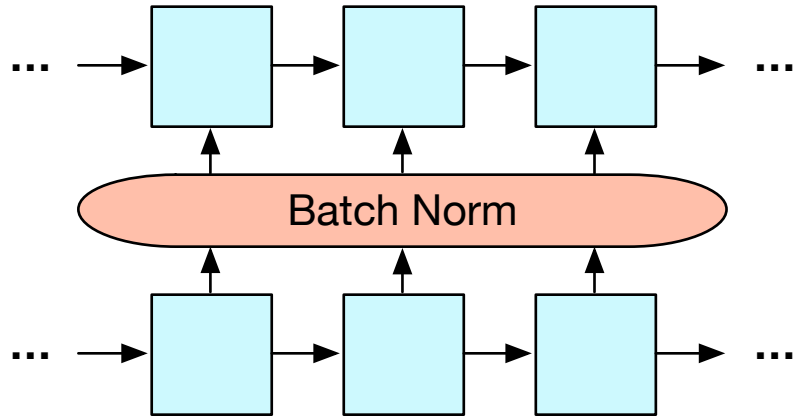
Deep Speech 1 to Deep Speech 2

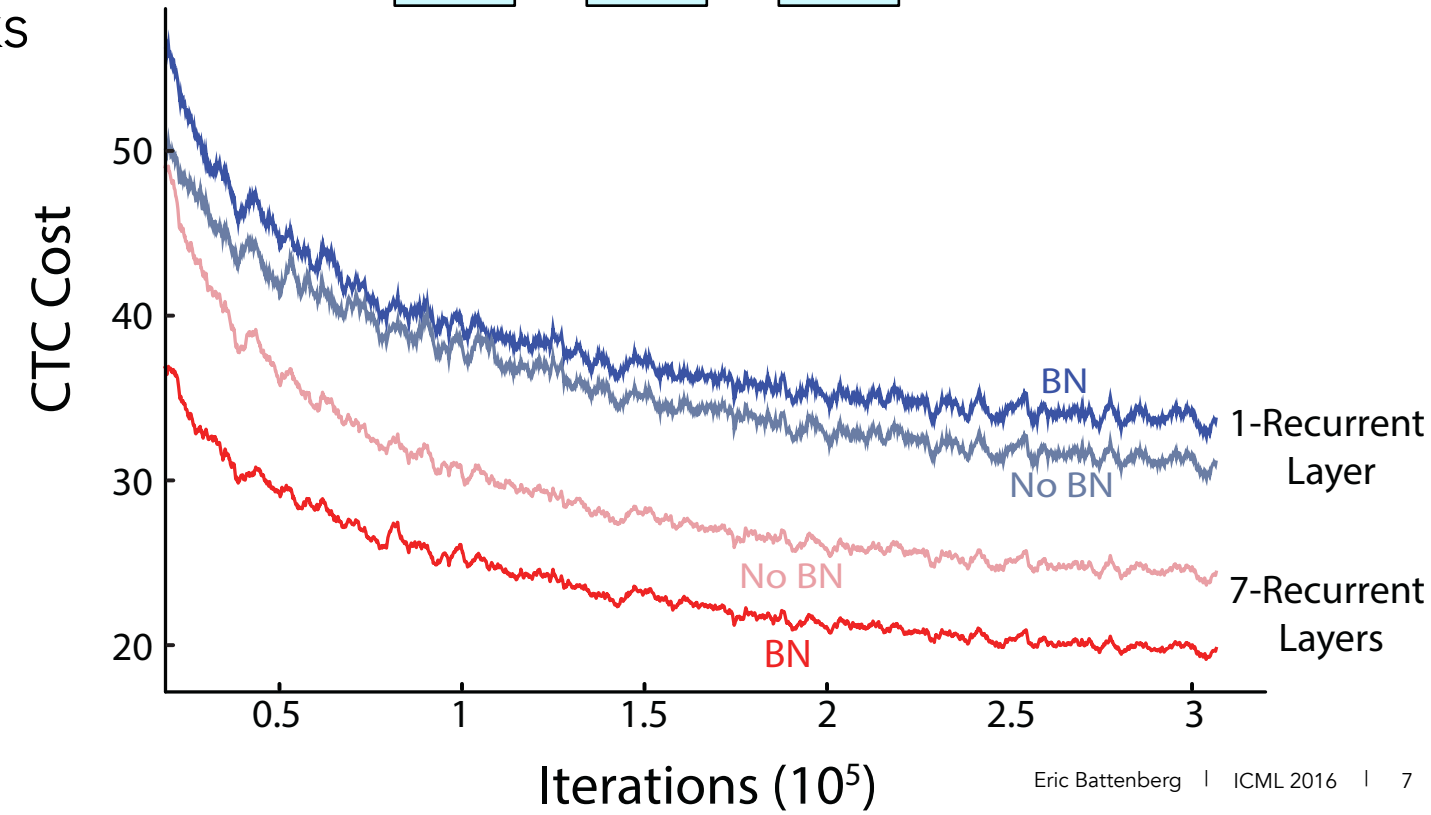
- Multiple layers of 2D convolution
- Up to 7 recurrent layers
- Trained with Batch Norm
- + Lots of systems expertise



Batch Normalization For RNNs

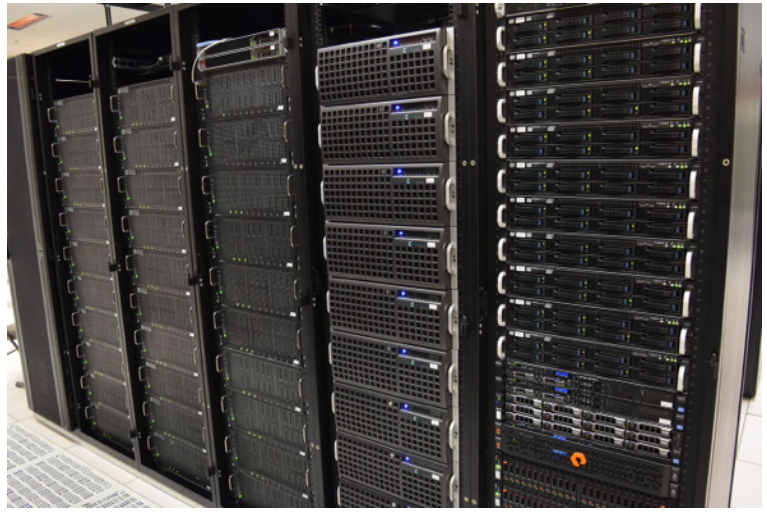
- Sequence-wise Batch Norm on upward connections
- More effective in deeper networks
- Speeds up training
- Improves test accuracy





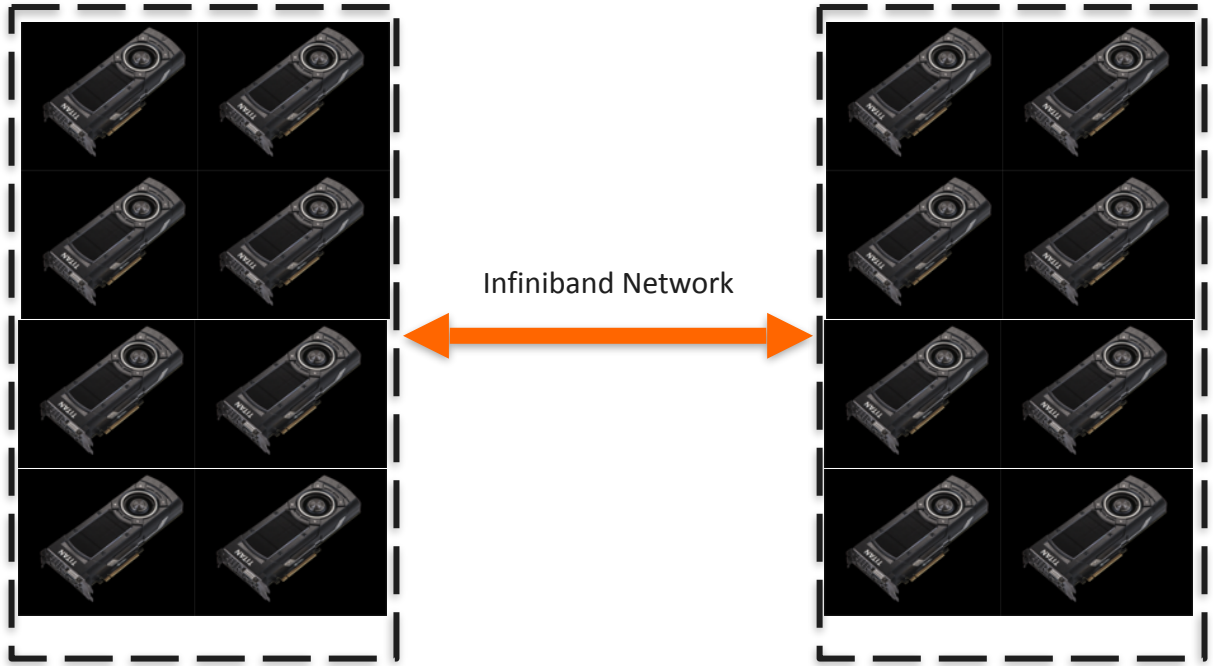
Scaling Up Deep Speech 2

- Tens of exaflops required to train model
- Cluster with 8 TitanX GPUs per node
  - Partition minibatch across GPUs
  - Synchronous SGD



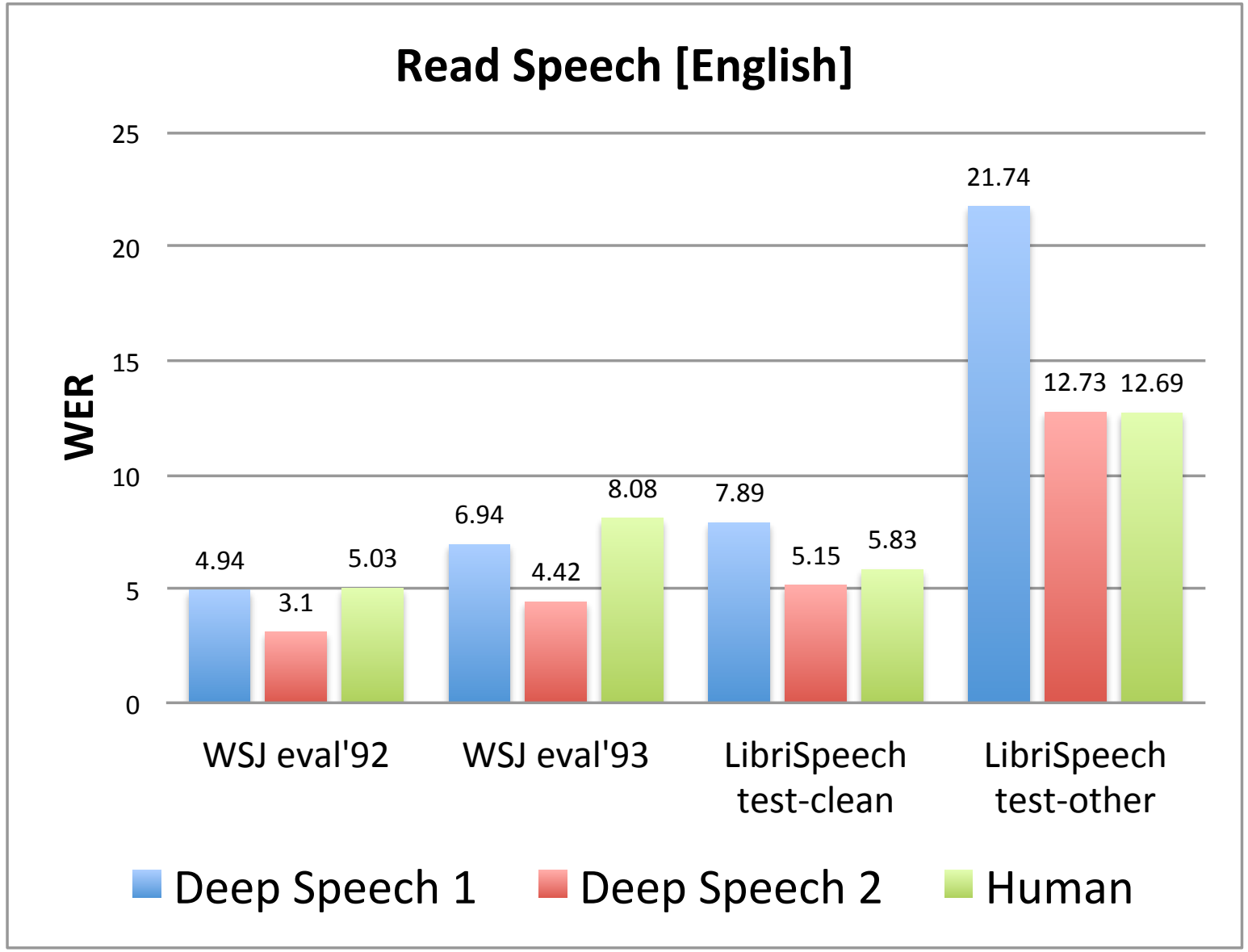
Scaling Up Deep Speech 2

- Custom All-Reduce code
  - 4x-21x speedup over OpenMPI's
- Fast GPU CTC (Warp-CTC)
  - Reduced training time by 10%-20%
- Fastest available kernels
- Overall, sustained 45% peak performance on each node.



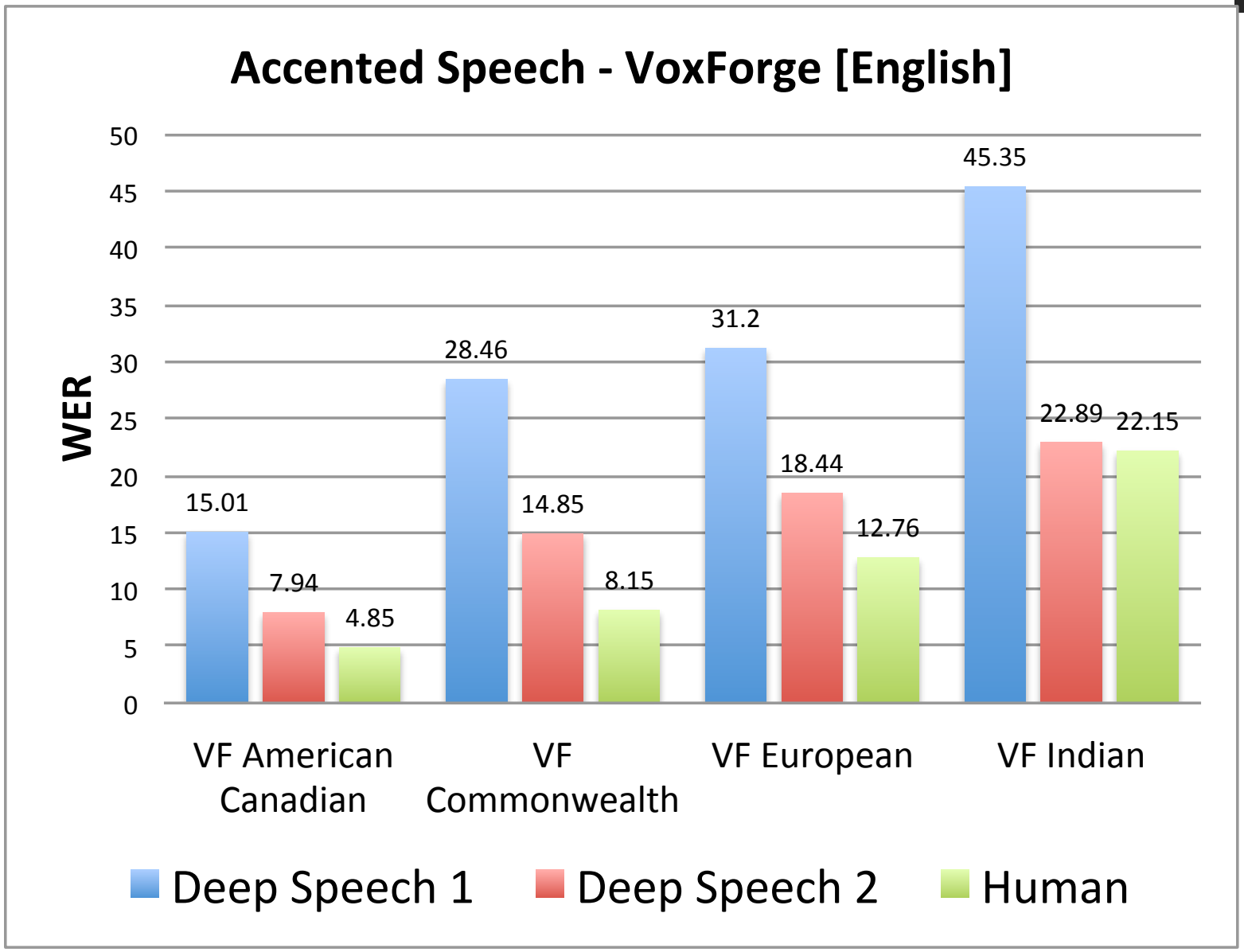
English Results

- English training data:
  - 11,940 hours
  - 8 million utterances
  - ...and growing everyday



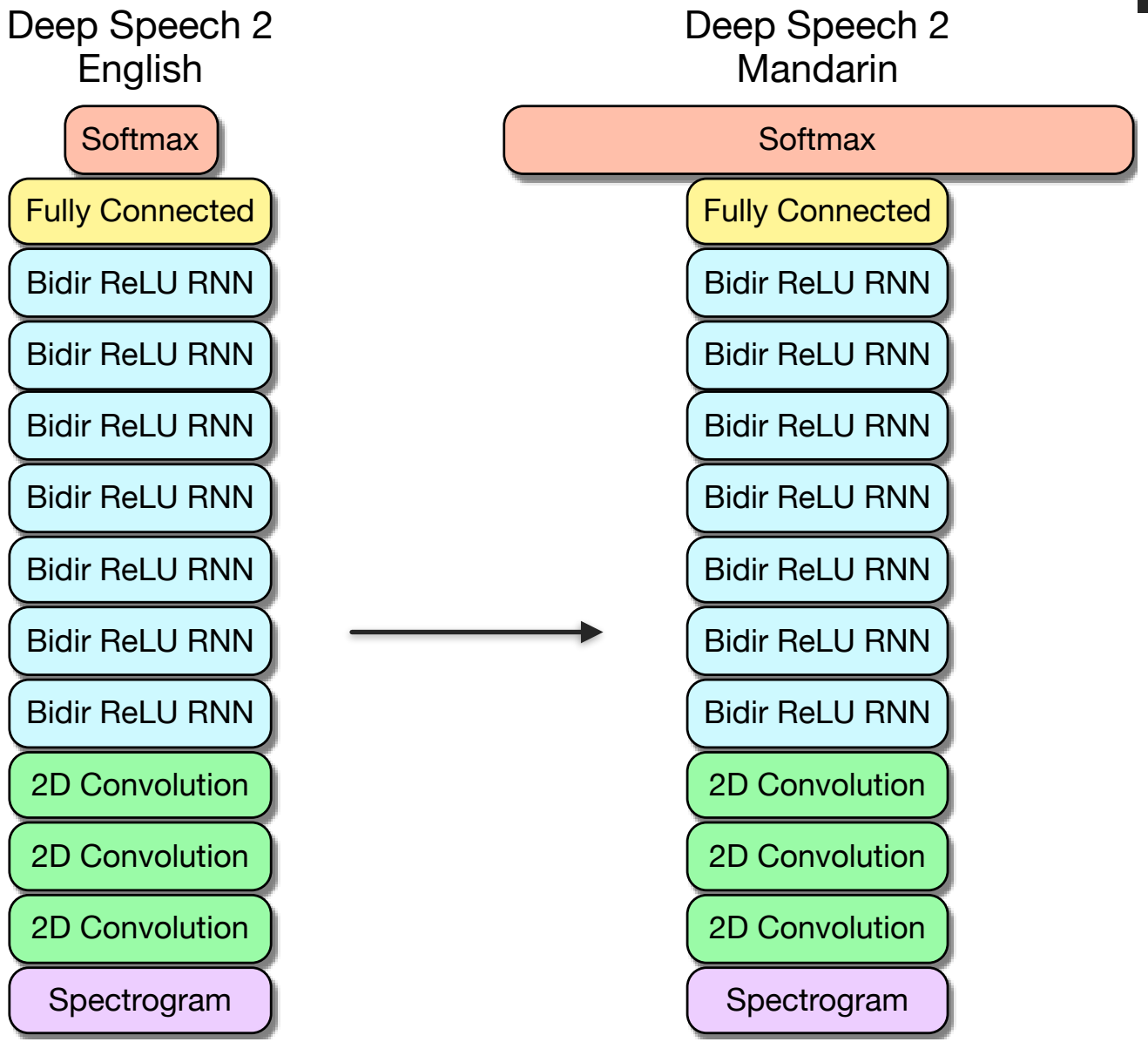
English Results

- English training data:
  - 11,940 hours
  - 8 million utterances
  - ...and growing everyday



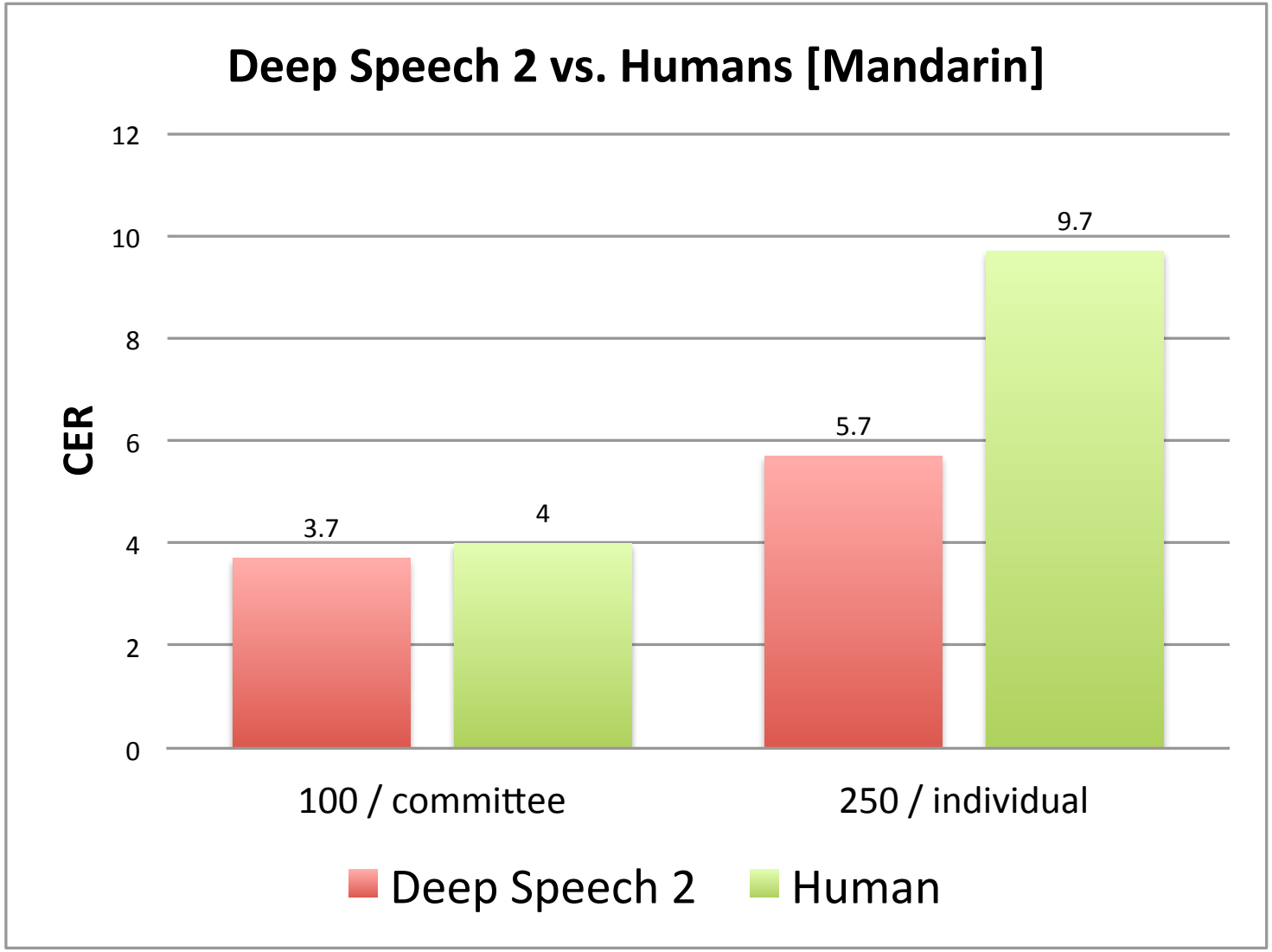
Porting to Mandarin

- Increase softmax size:
  - from 29 to ~6000
- Feed in Mandarin data
- Tweak some hyperparameters



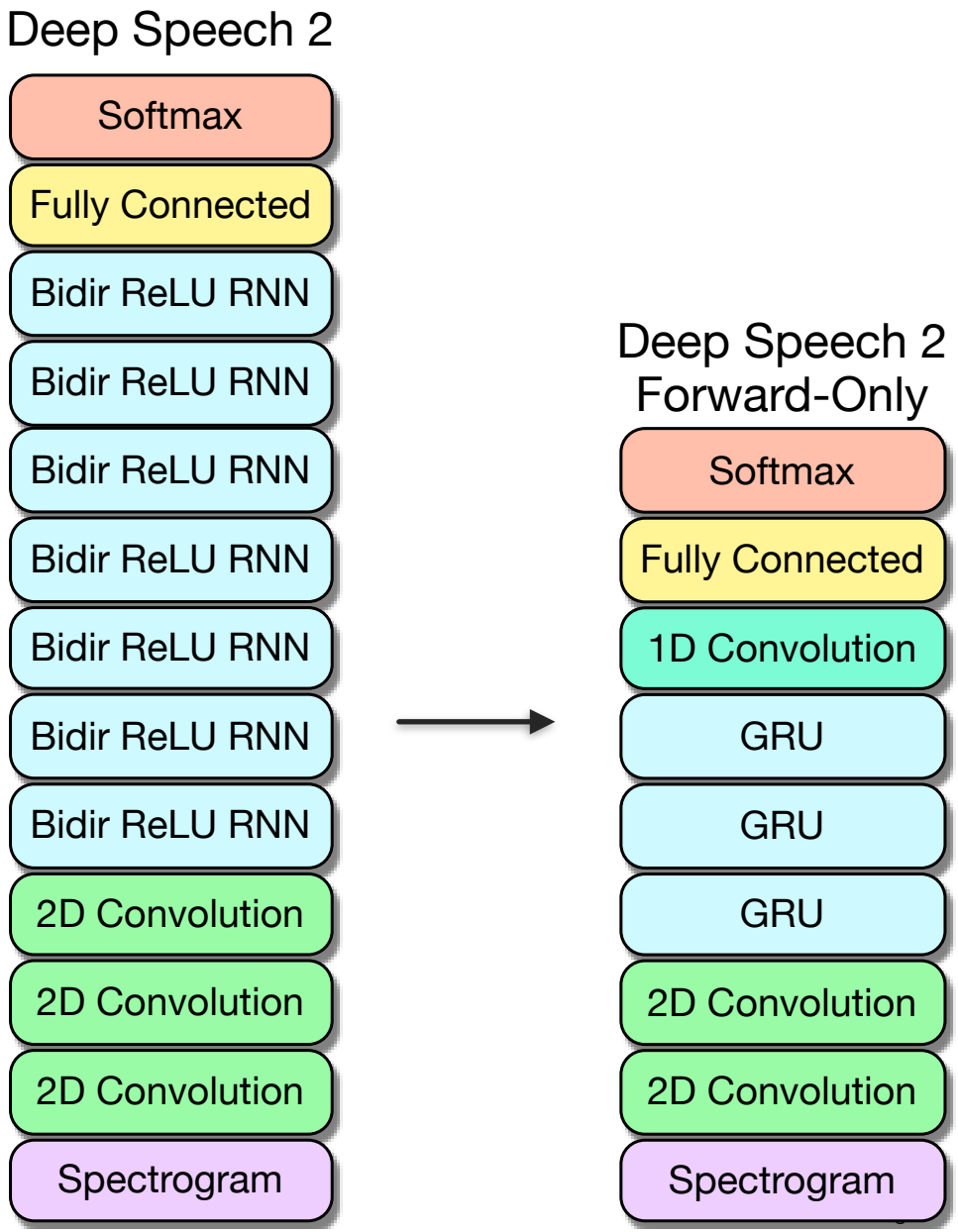
Mandarin Results

- Mandarin training data:
  - 9,400 hours
  - 11 million utterances
  - More diverse than English data
- (...growing everyday)



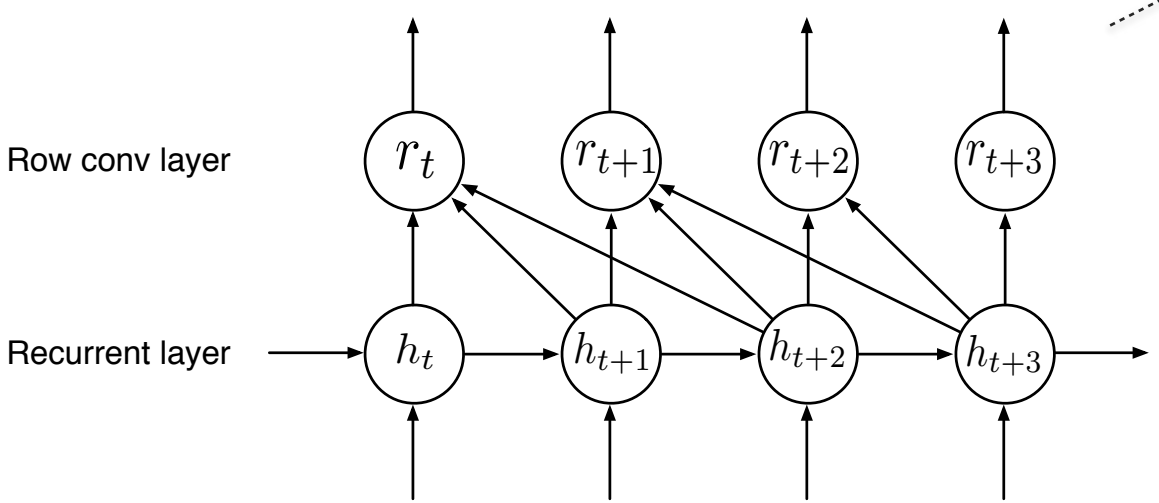
To Production-Ready Models

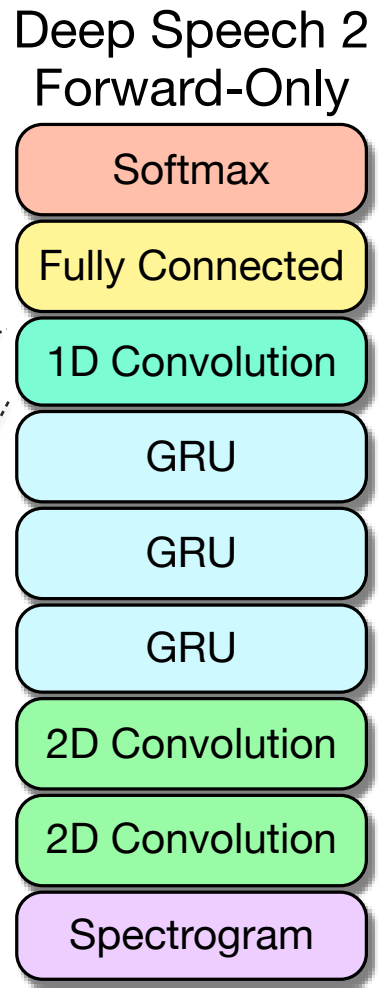
- Production models require causality
- Unidirectional recurrent layers
- GRUs instead of ReLU RNNs
- 1D lookahead convolution



Lookahead Convolutions

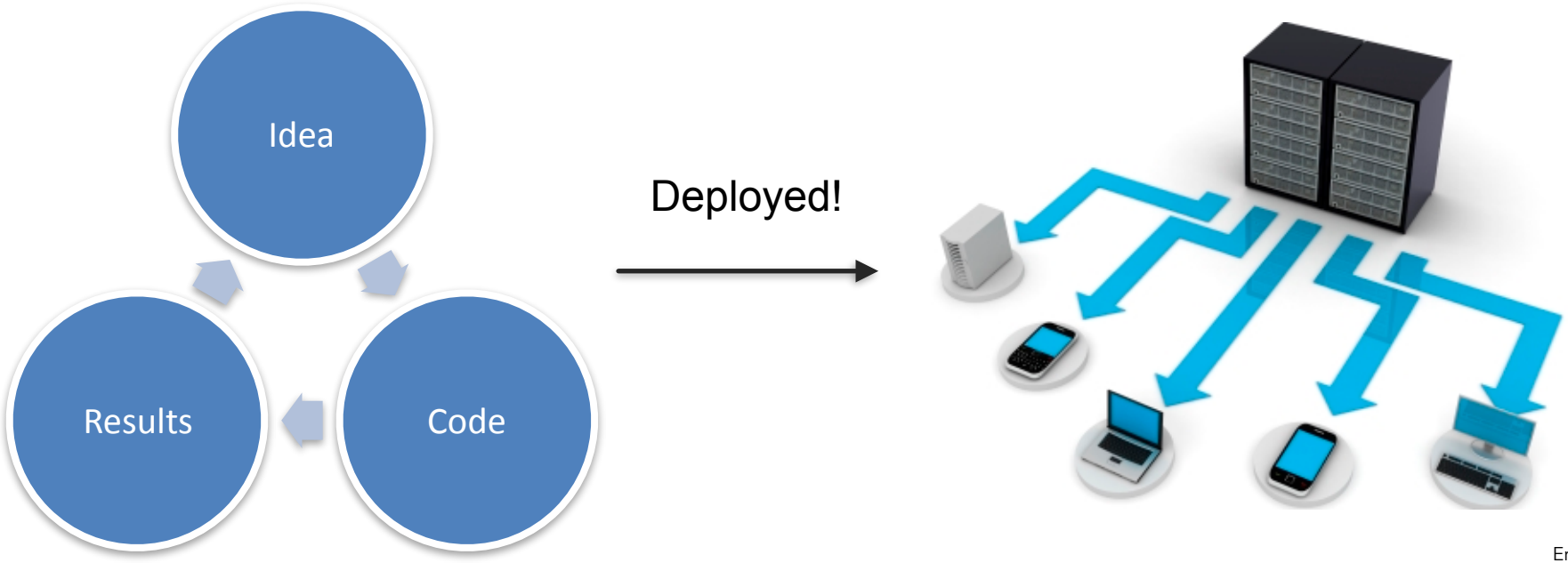
- Some future context is useful.
- Difficult to force network to delay predictions.
- We use 1D convolution after recurrent layers
- Within 5% relative performance of bidirectional models





Research to Deployment

- Deep Speech 2 is now in production!
- Baidu Silicon Valley AI Lab + Baidu Speech Technology Group
- From research idea to deployed product in 2 years.
- Combined efforts of 69 researchers and engineers



Conclusion

- The story of Deep Speech 2
  - End-to-end speech recognition
  - Scaled up
  - Ported to Mandarin
  - Deployed to production

- Deep Speech 3, coming to a conference near you in 2017!