AI for 100 million people with deep learning

Adam Coates

@adampaulcoates

Baidu Research

Silicon Valley AI Lab

Silicon Valley AI Lab

- **Mission:** Develop hard AI technologies that let us have significant impact on hundreds of millions of users.

➢ Choose problems that significantly affect large numbers of people.

AI for 100 million people

- First goal: speech recognition everywhere.

If you're connecting to internet for first time in 2017,
you're likely using a mobile device.

Speech will transform mobile device interfaces.

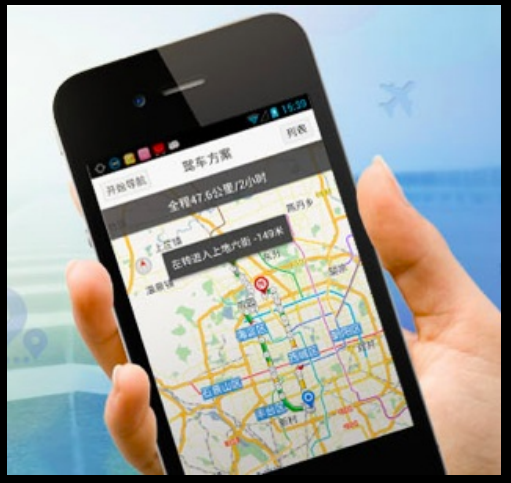
AI for 100 million people

- First goal: speech recognition everywhere.

Captioning

Cars / Hands-free interfaces

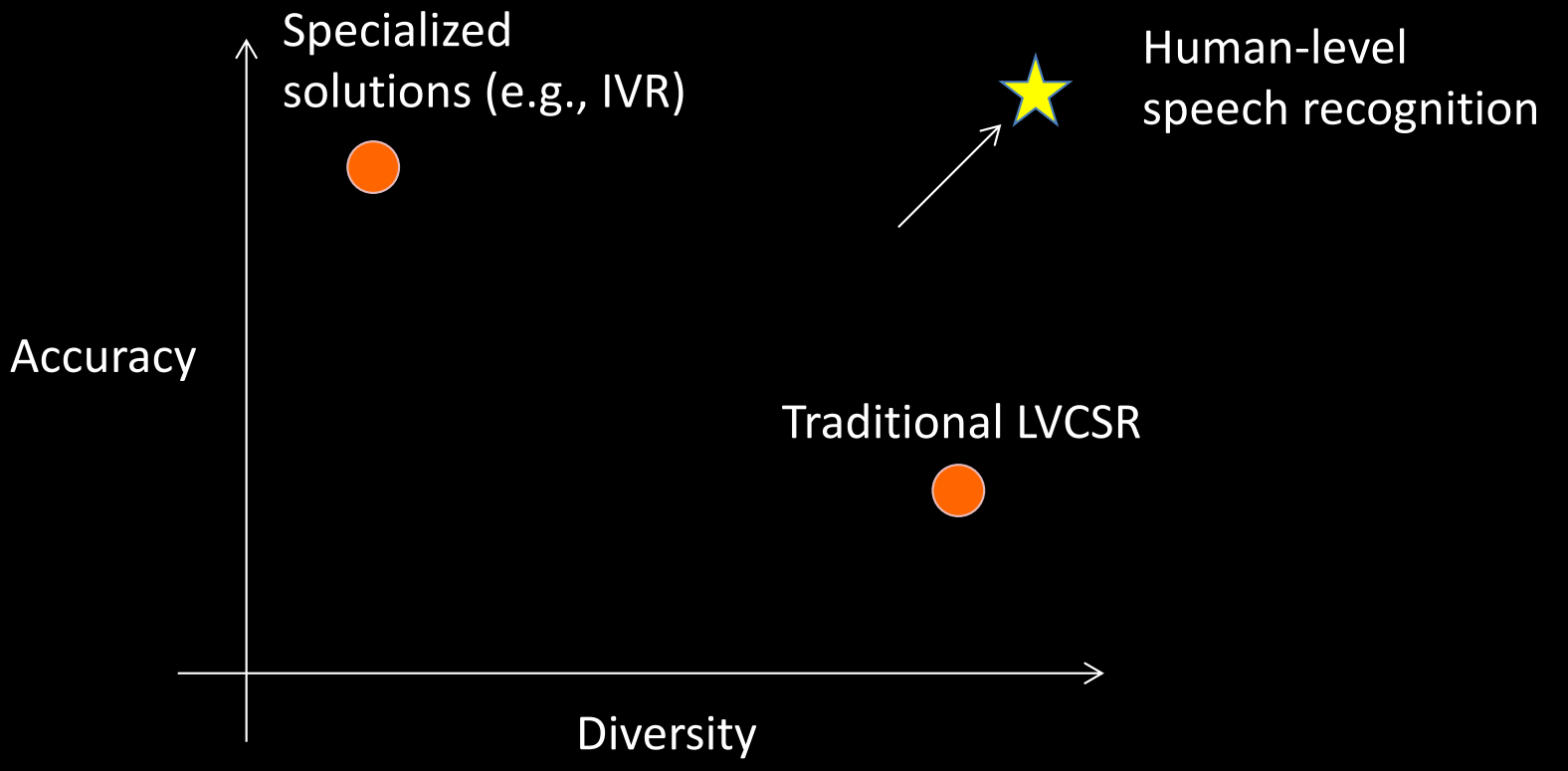
Home devices



Mobile devices

AI for 100 million people

- First goal: speech recognition everywhere.



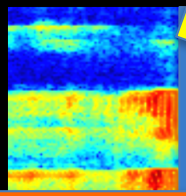
Speech recognition: Traditional ASR

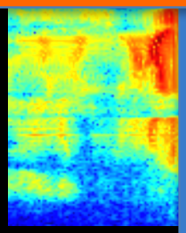
- Traditional speech systems built on standard ML + engineering practices.

Compute features

phonemes

Merge with pronunciation and language data.



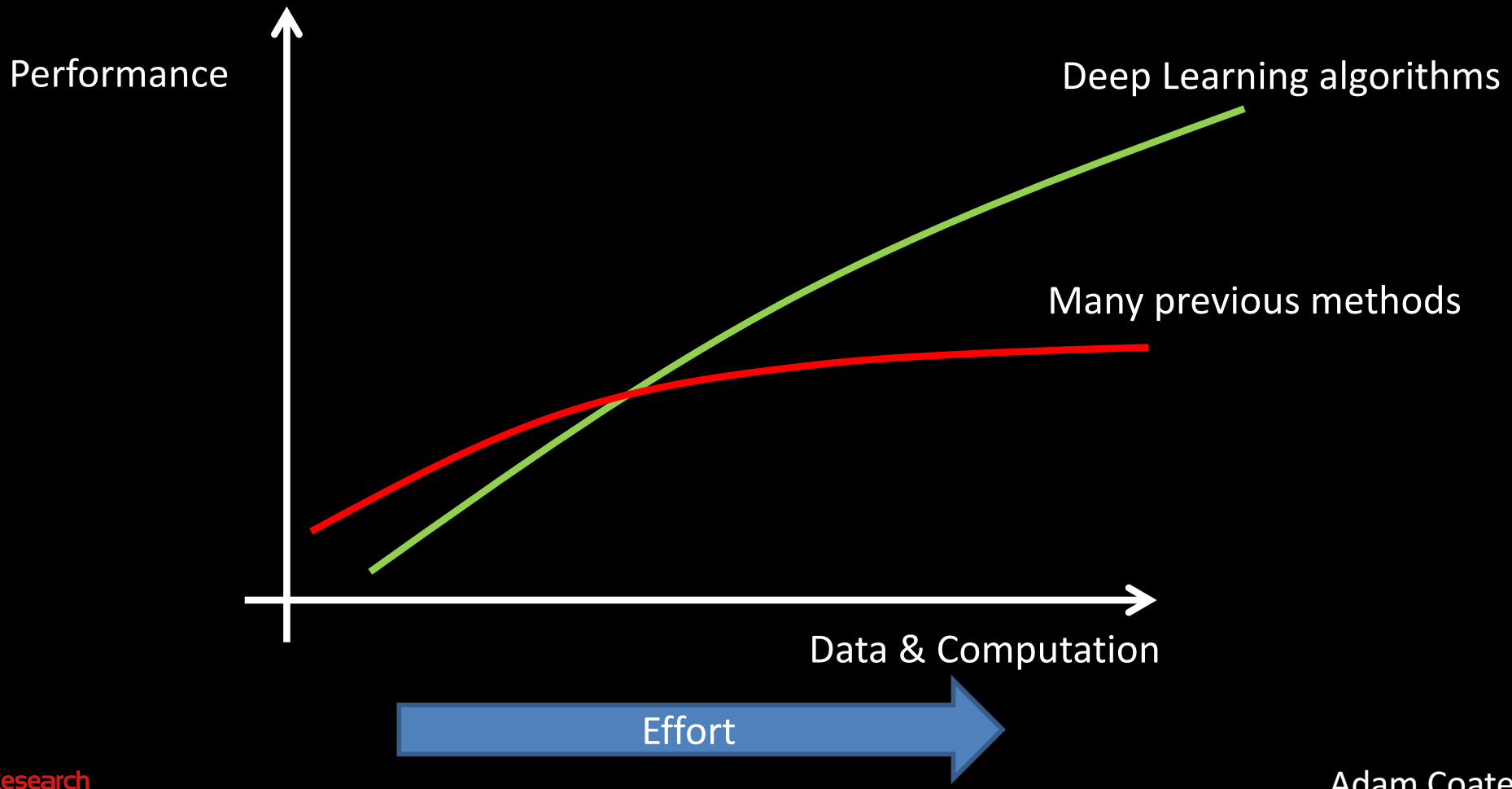


Some applications can be solved this way.
But it's hard to scale our own cleverness.

"What time is it in Beijing?"

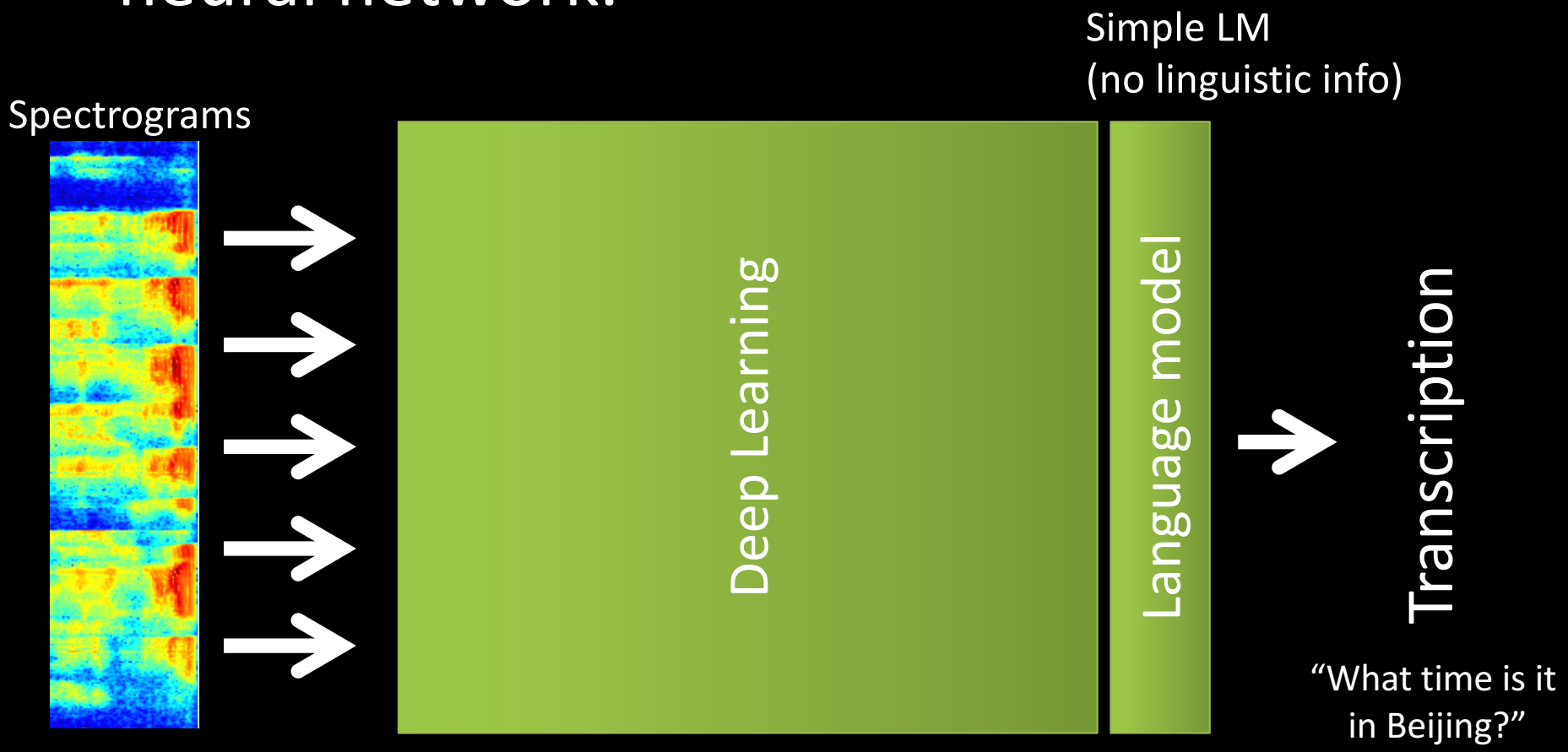
Deep learning

Major advantage of deep learning: scalability.



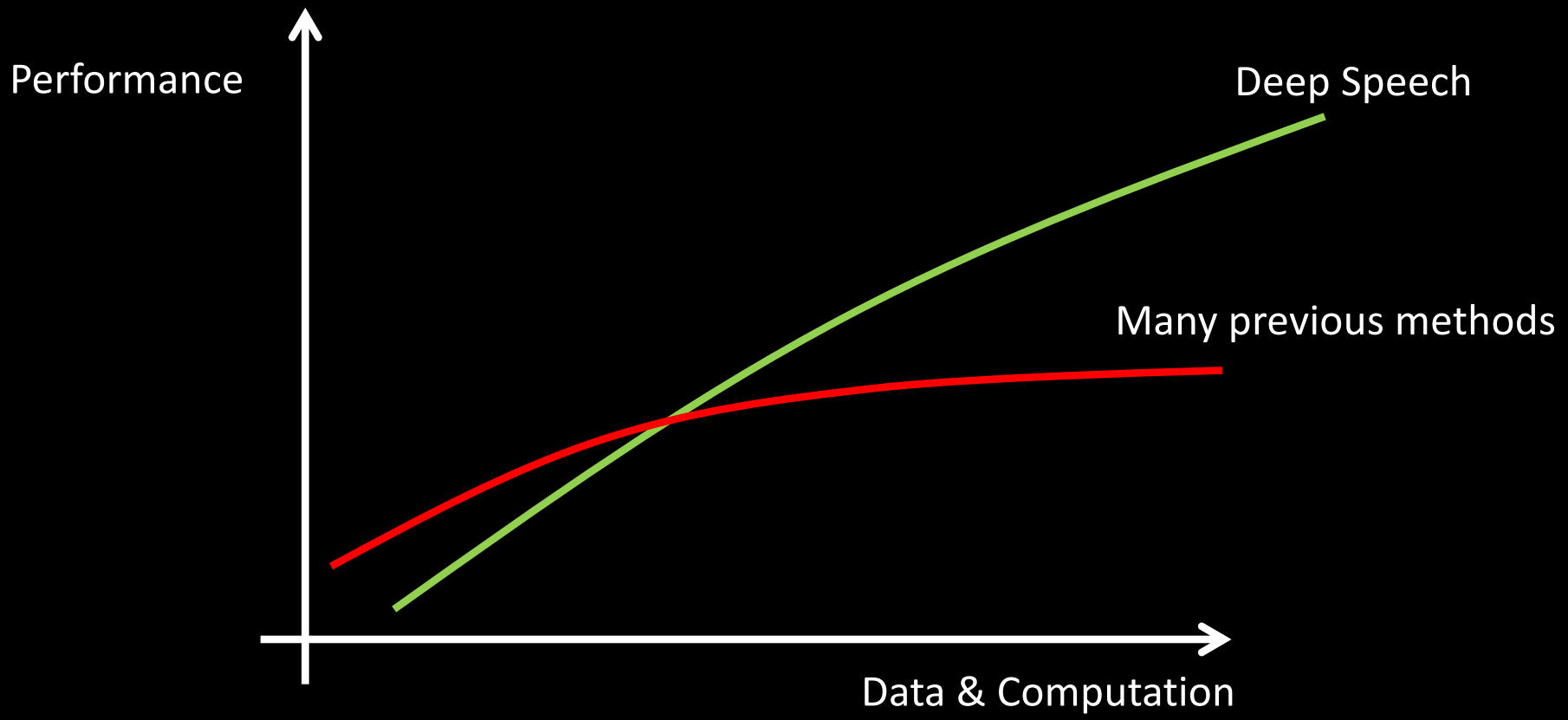
Speech recognition with deep learning

- Replace most of speech system with large neural network.



“Deep Speech”

- Pour effort into data + computation.
  - Try to catch up to human accuracy by scaling.

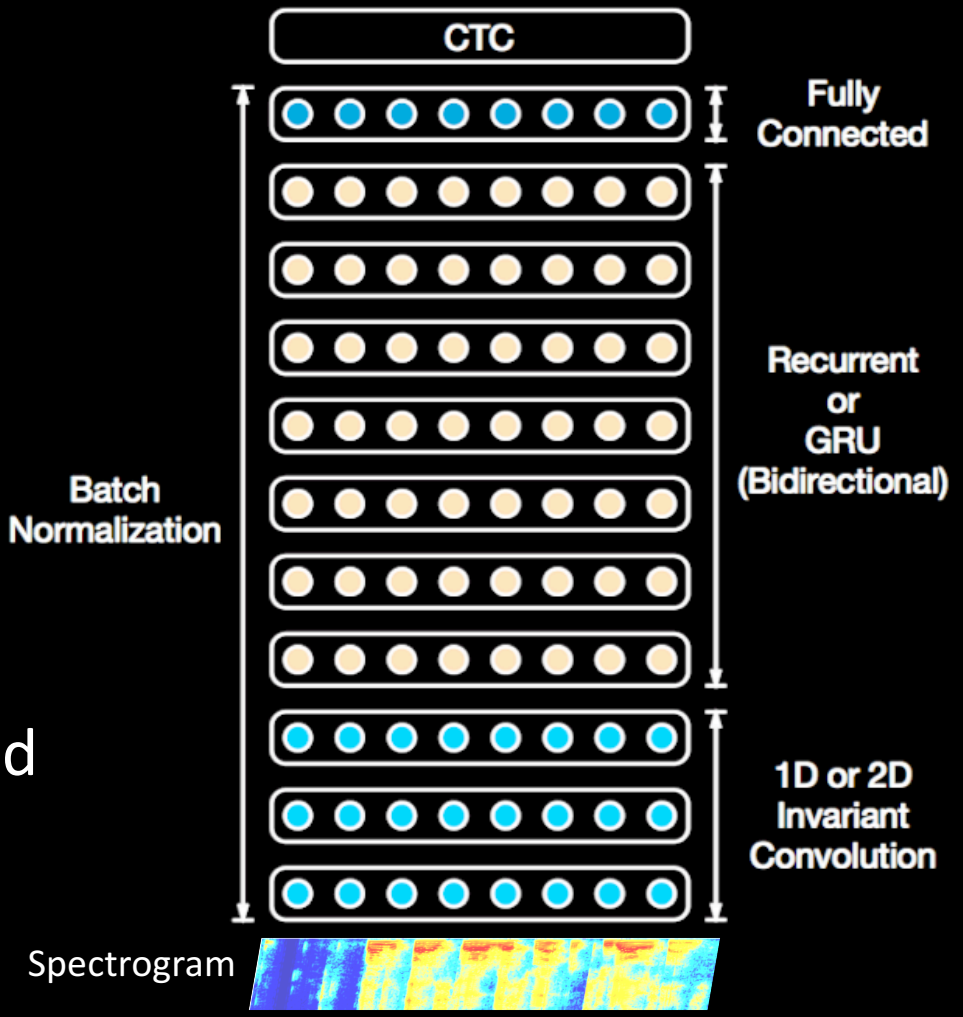


Deep Speech

- Train giant neural networks to predict characters from audio.
  - Train “end to end” [e.g., Graves et al., 2006]

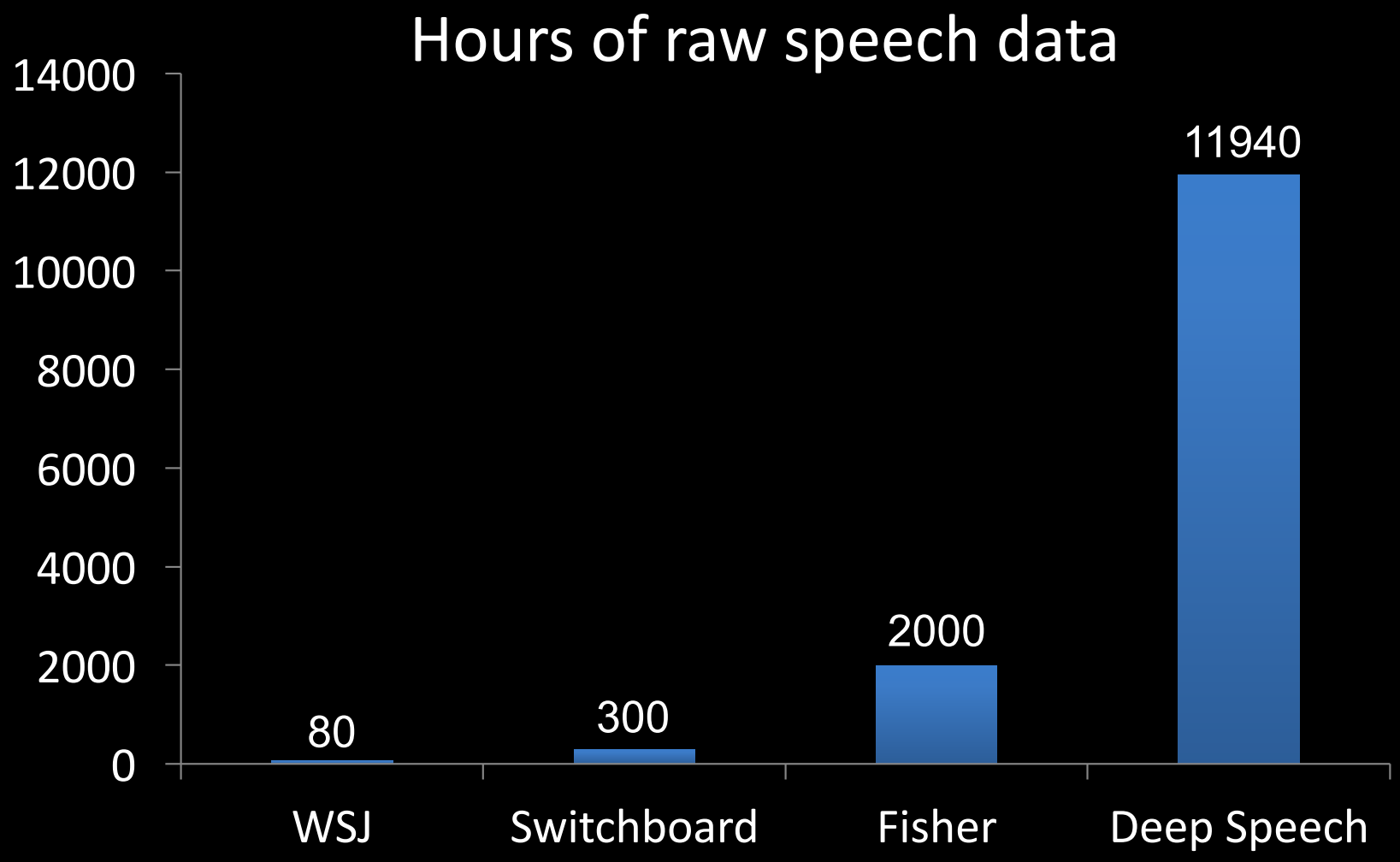
➢ Hard part is training at scale and searching for best model.

➢ **Need data + computing power.**



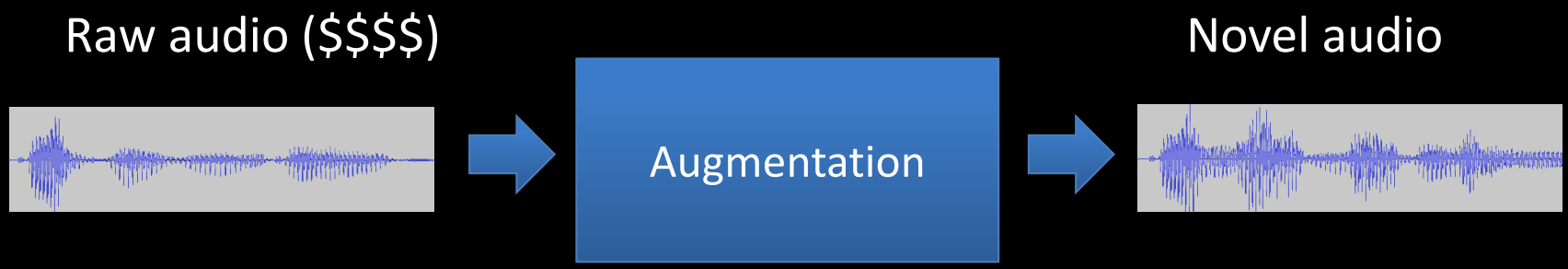
Raw Training Data

We need *a lot of* data for end to end DL systems: use read speech.



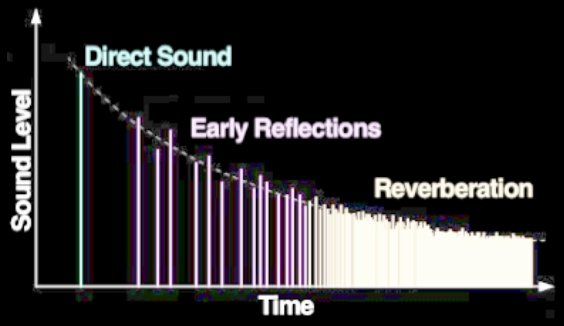
Data augmentation

- Many forms of distortion that model should be robust to:
  - Reverb, noise, far field effects, echo, compression artifacts, changes in tempo.
- Learn to be robust by training from data with distortions!
  - Easier to engineer data pipeline than to engineer recognition pipeline.



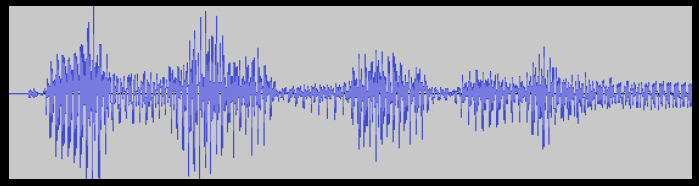
Example: far field

Speech



Impulse Response

* =



Augmented Far-field speech

Compare: Real Far-field speech

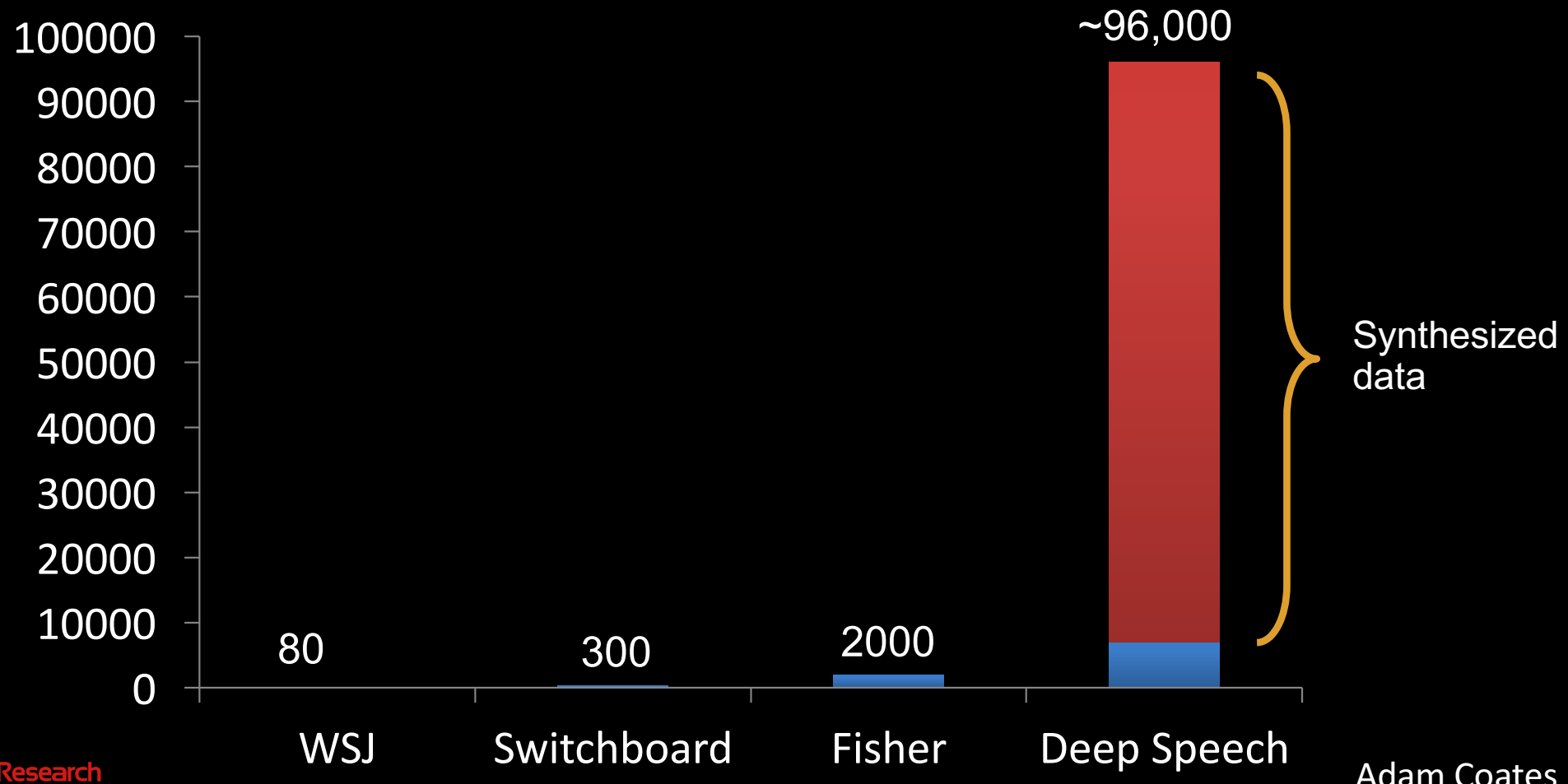
Reduces errors on far-field data by 15%-20%.
Relies on model search + large-scale training.

Baidu Research

[Billy Jun, Rewon Child, Sanjeev Satheesh]

Augmented dataset

- Augmentation greatly expands available data.
  - Trained models have heard *decades* of unique audio.



Compute

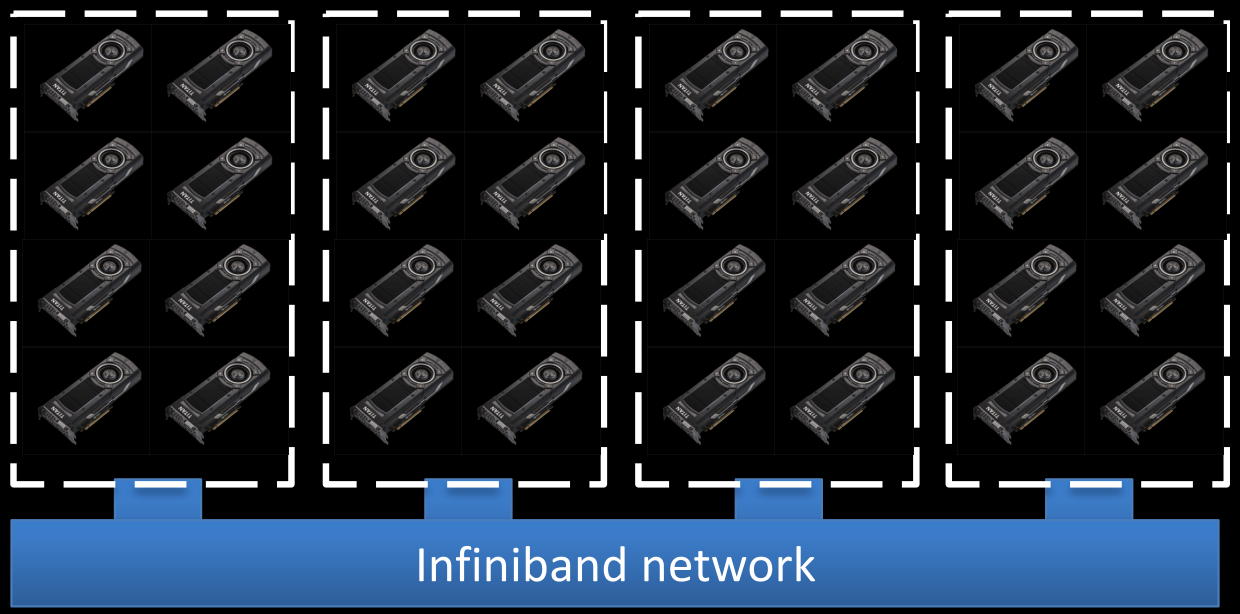
1 experiment requires >10,000,000,000,000,000,000 FLOPs
(10s of ExaFLOPs)

Titan X GPU
~6 TeraFLOPs

“Speed of light” = 3-6 weeks on 1 GPU

Compute

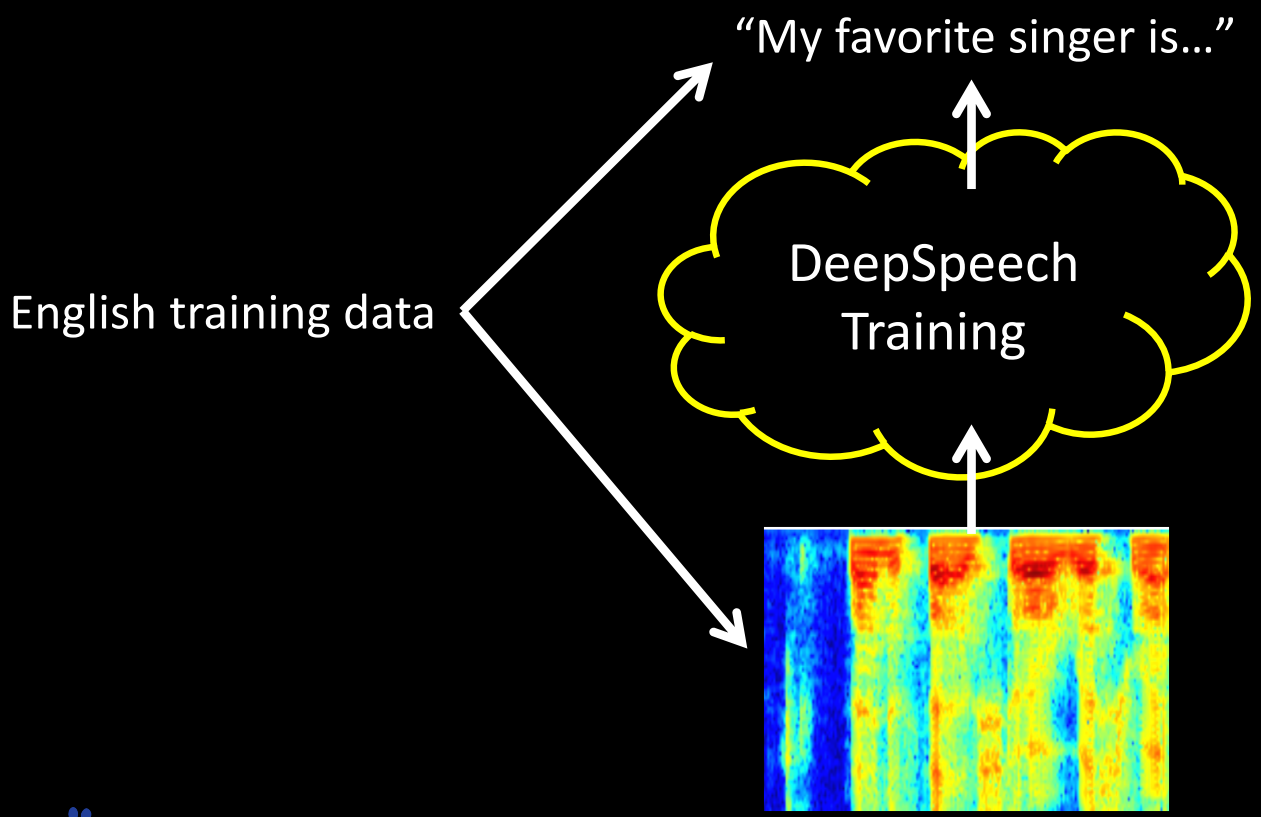
Scale out to large numbers of GPUs (e.g., 8 – 64)



- Cut experiment times to ~3-5 days.
  - Achieve ~50% of peak FLOPs on 8+ GPUs.
  - Comparable to supercomputing workloads.

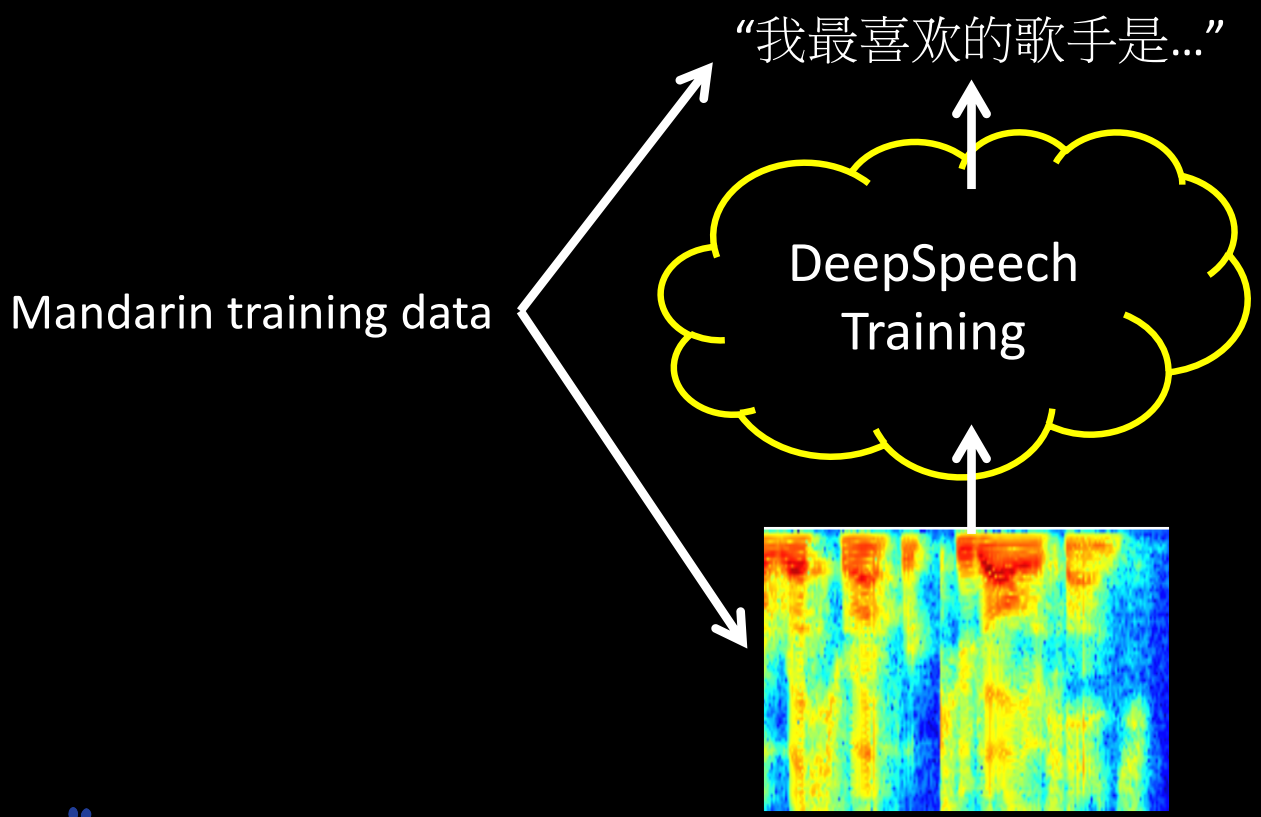
Deep Speech for Mandarin

- Deep Speech is driven by data.
  - Mandarin is very different from English.
    - "Tonal", thousands of characters



Deep Speech for Mandarin

- Deep Speech is driven by data.
  - Mandarin is very different from English.
    - "Tonal", thousands of characters



Deep Speech for Mandarin

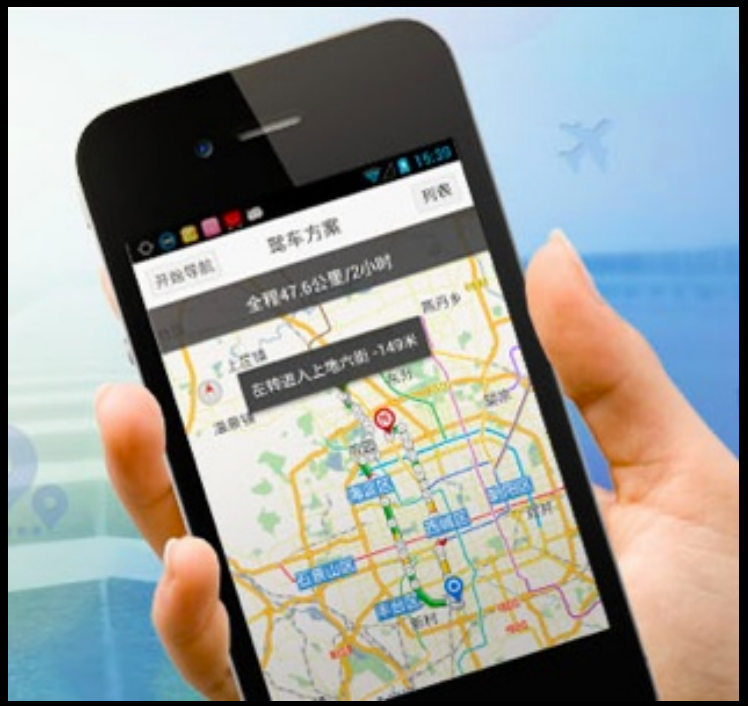
- With a few changes, a single algorithm can learn an entirely new language.
  - Competitive with committee of native speakers for short audio clips.

➢ Learns hybrid speech (e.g., famous people, iphones):

我最喜欢的歌手是Angelababy

Making devices easier and more efficient

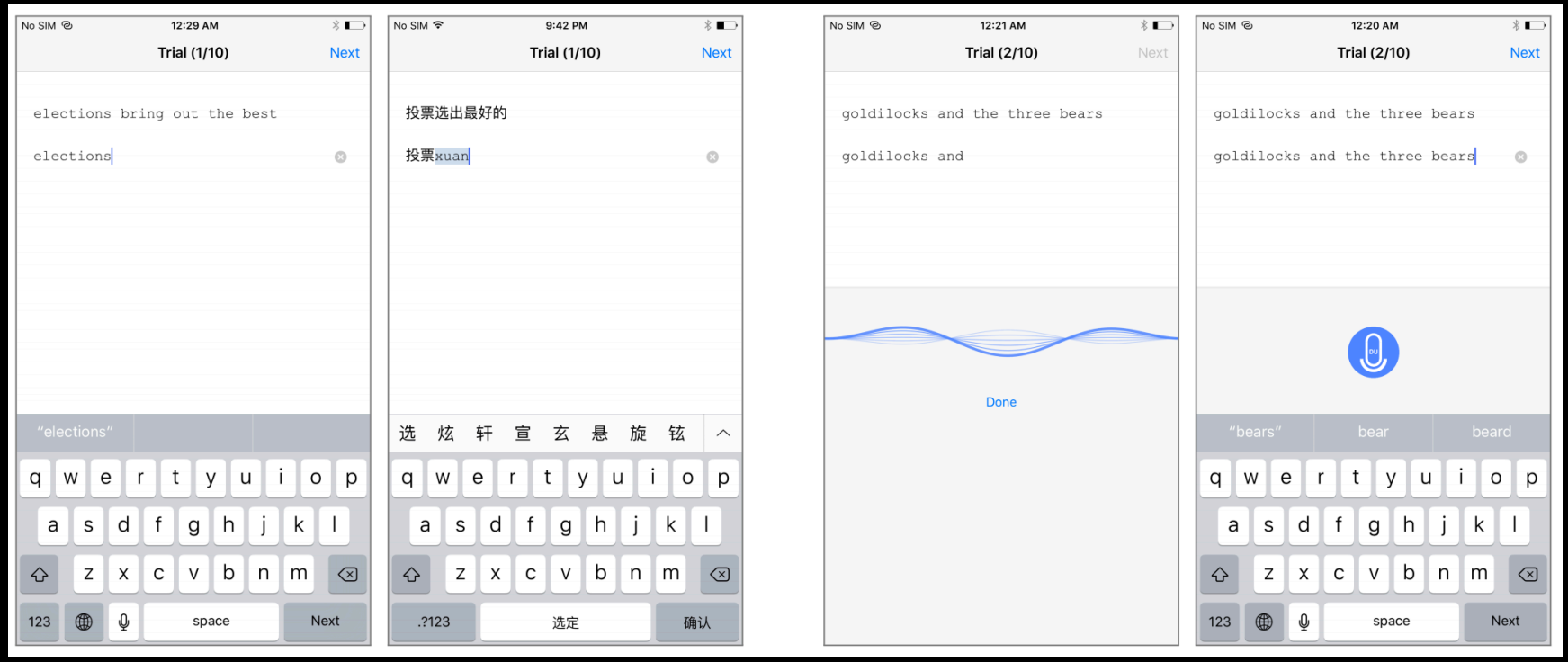
- Does speech make a difference? YES!



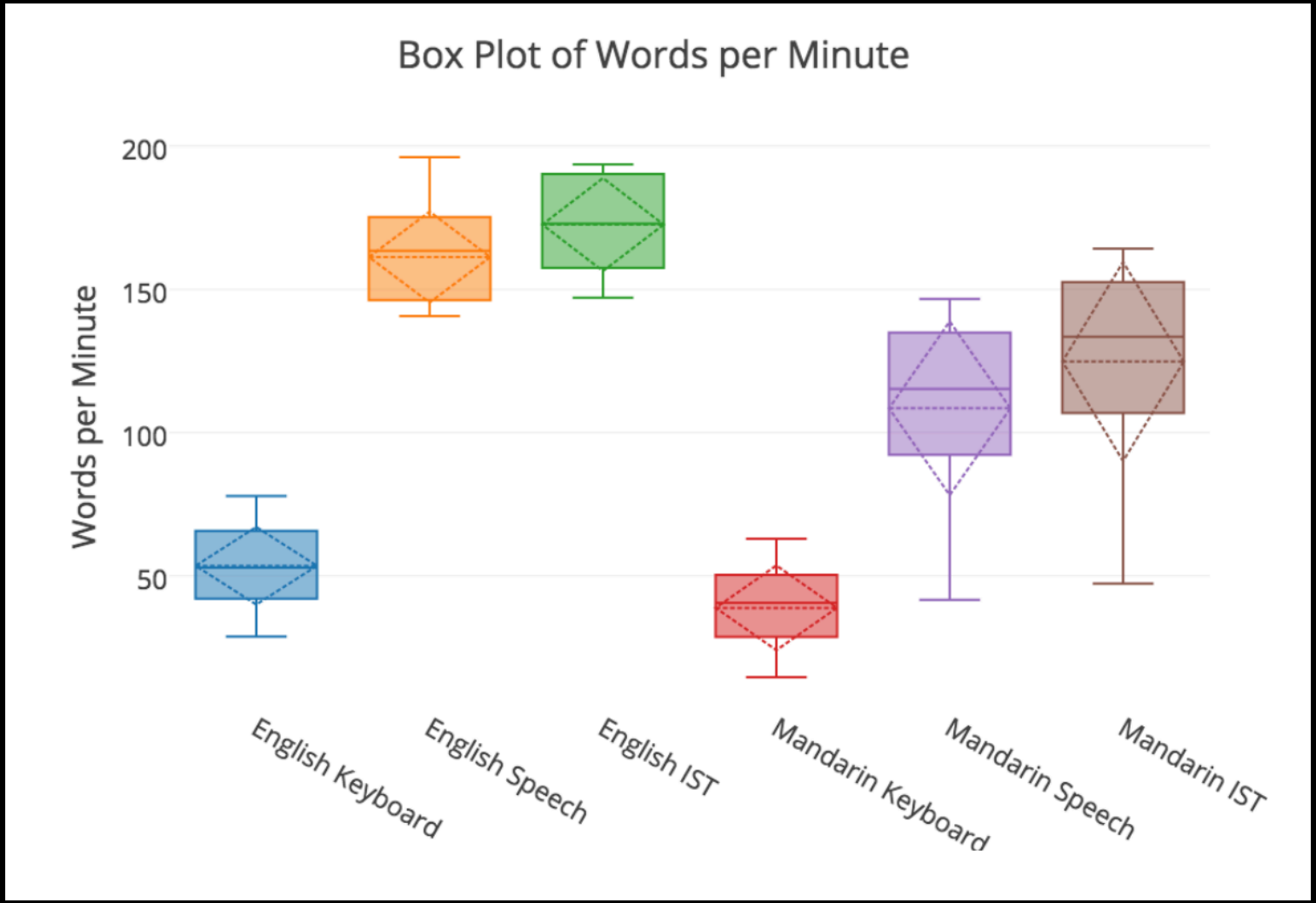
Comparing speech with keyboard input

- Compare user performance/experience for speech vs. traditional keyboard.

[Ruan et al., arxiv.org/abs/1608.07323]

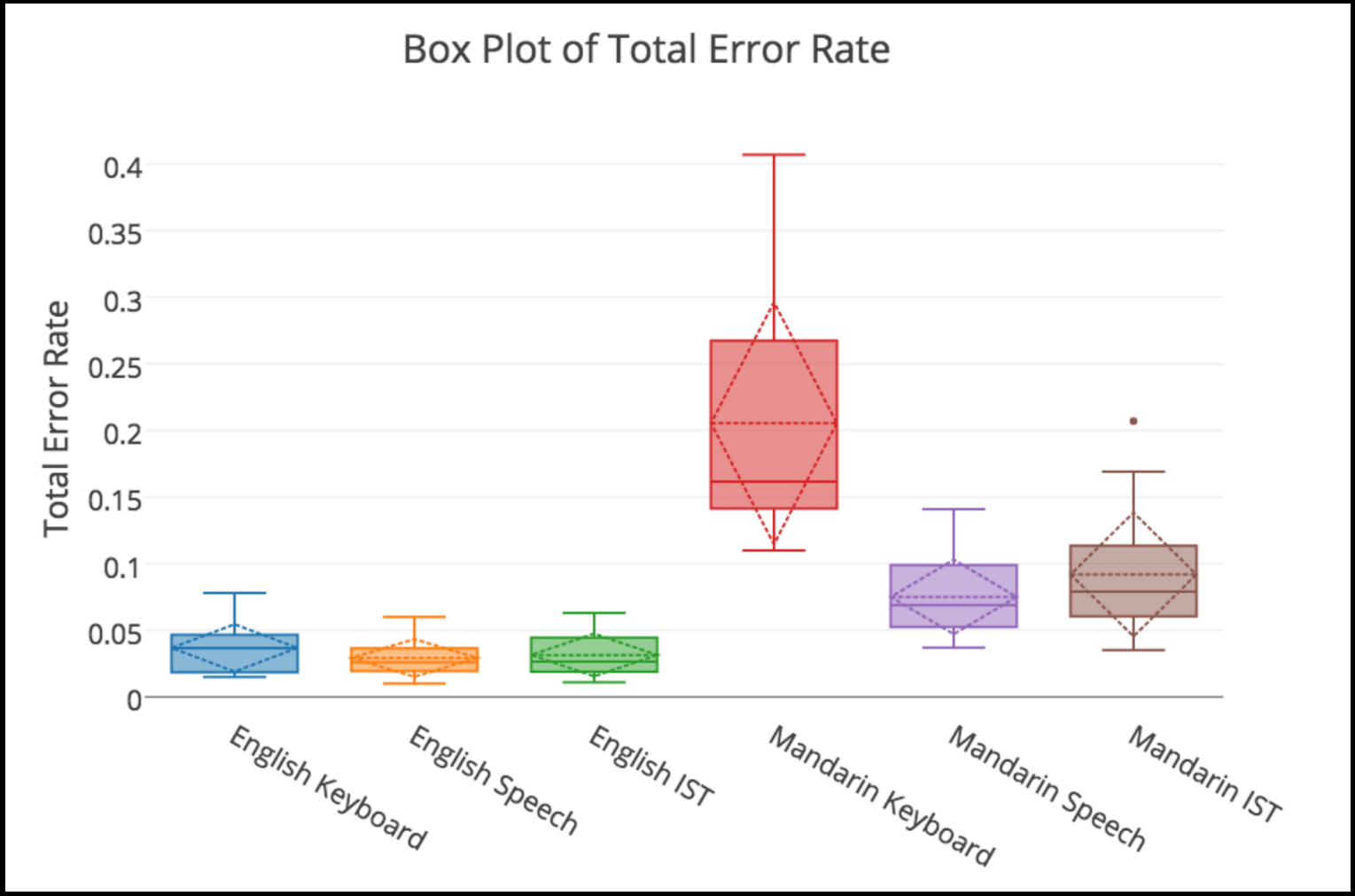


Speech is 3x faster than typing



Ruan et al., arxiv.org/abs/1608.07323

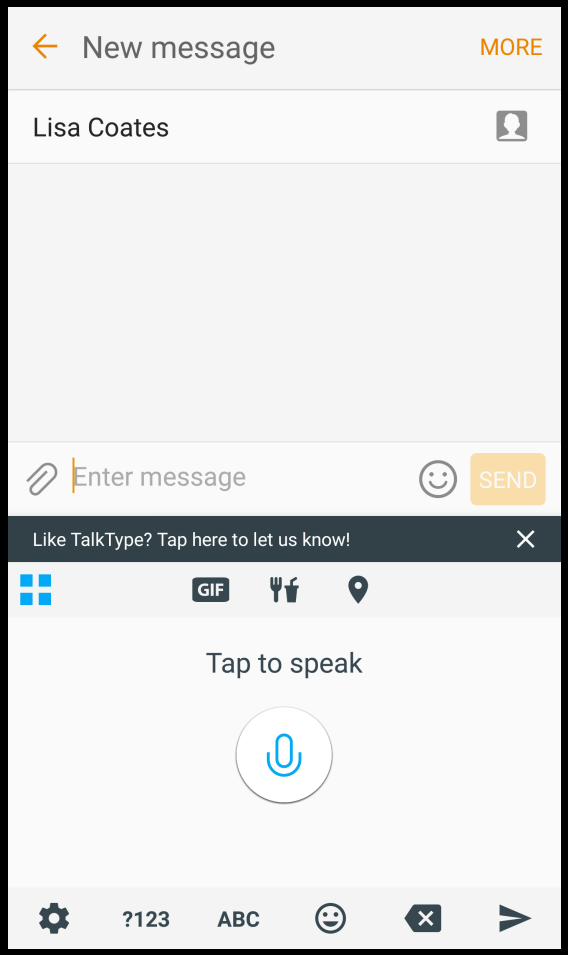
…and more accurate.

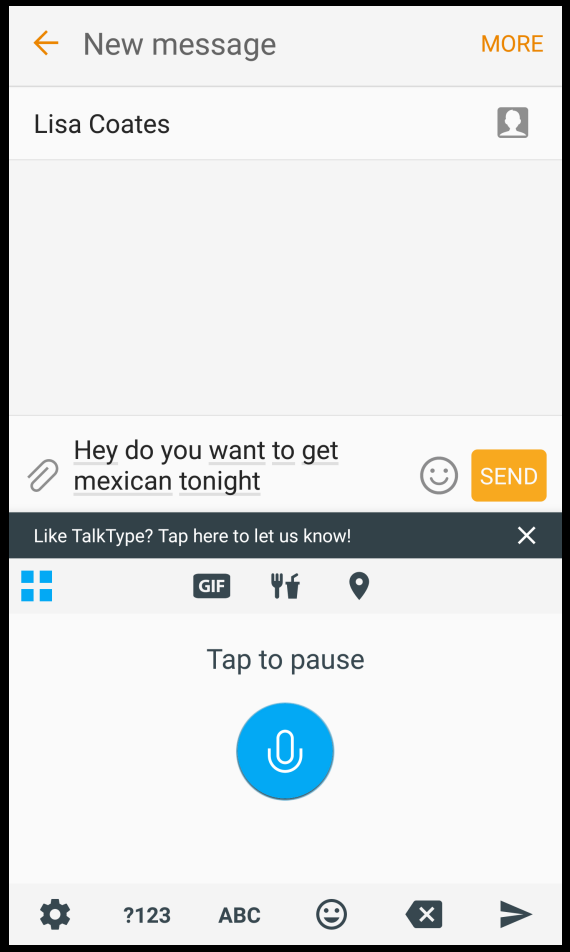


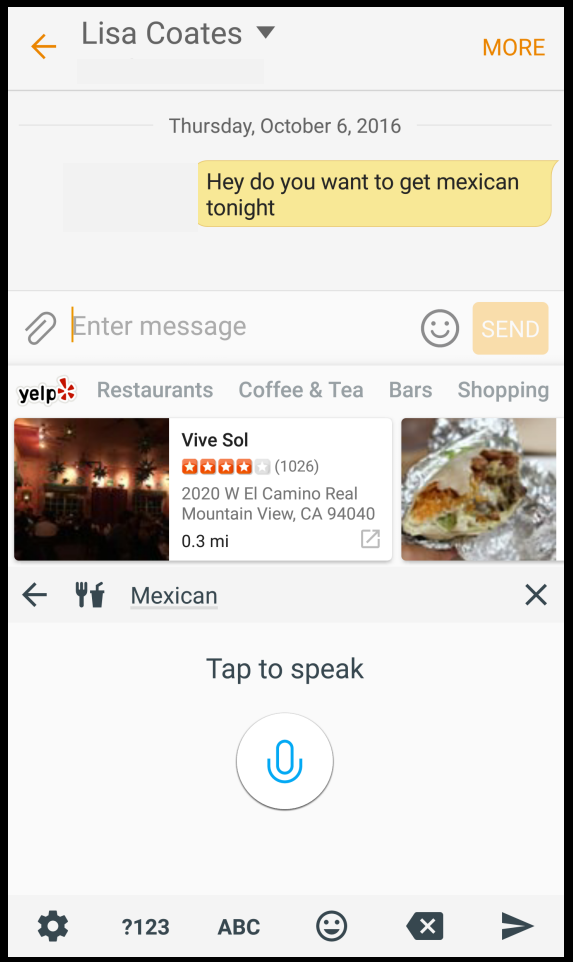
Ruan et al., arxiv.org/abs/1608.07323

TalkType – voice-centric keyboard for Android

- Opportunity to rethink product experiences around speech and AI.







talktype.baidu.com [Nicky Chan, Bijit Halder, Kenny Liou, Thuan Nguyen, Nina Wei]

AI for 100 million people

- Deep Learning is closing gap with humans on speech, through scalability.
  - Still more to do; but it keeps getting better.
- Speech already enabling proliferation of new AI products.
  - Let's make them work for everyone.

Thank you!

Adam Coates


adamcoates@baidu.com


@adampaulcoates

If you want to help bring AI to 100s of millions of people, come talk to us!

http://research.baidu.com