EN. 601.467/667
Introduction to Human Language Technology
Deep Learning I

Shinji Watanabe

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Today's agenda

- Big impact in HLT technologies including speech recognition
  - Before/after deep neural network
- Brief introduction of deep neural network
- Why deep neural network after 2010?
- Success in the other HLT areas (image and text processing)

Today’s agenda

- Big impact in HLT technologies including speech recognition
  - Before/after deep neural network
- Brief introduction of deep neural network
- Why deep neural network after 2010?
- Success in the other HLT areas (image and text processing)

No math, no theory, based on my personal experience

Short bio

- Research interests
  - Automatic speech recognition (ASR), speech enhancement, application of machine learning to speech processing
- Around 20 years of ASR experience since 2001

Automatic speech recognition?

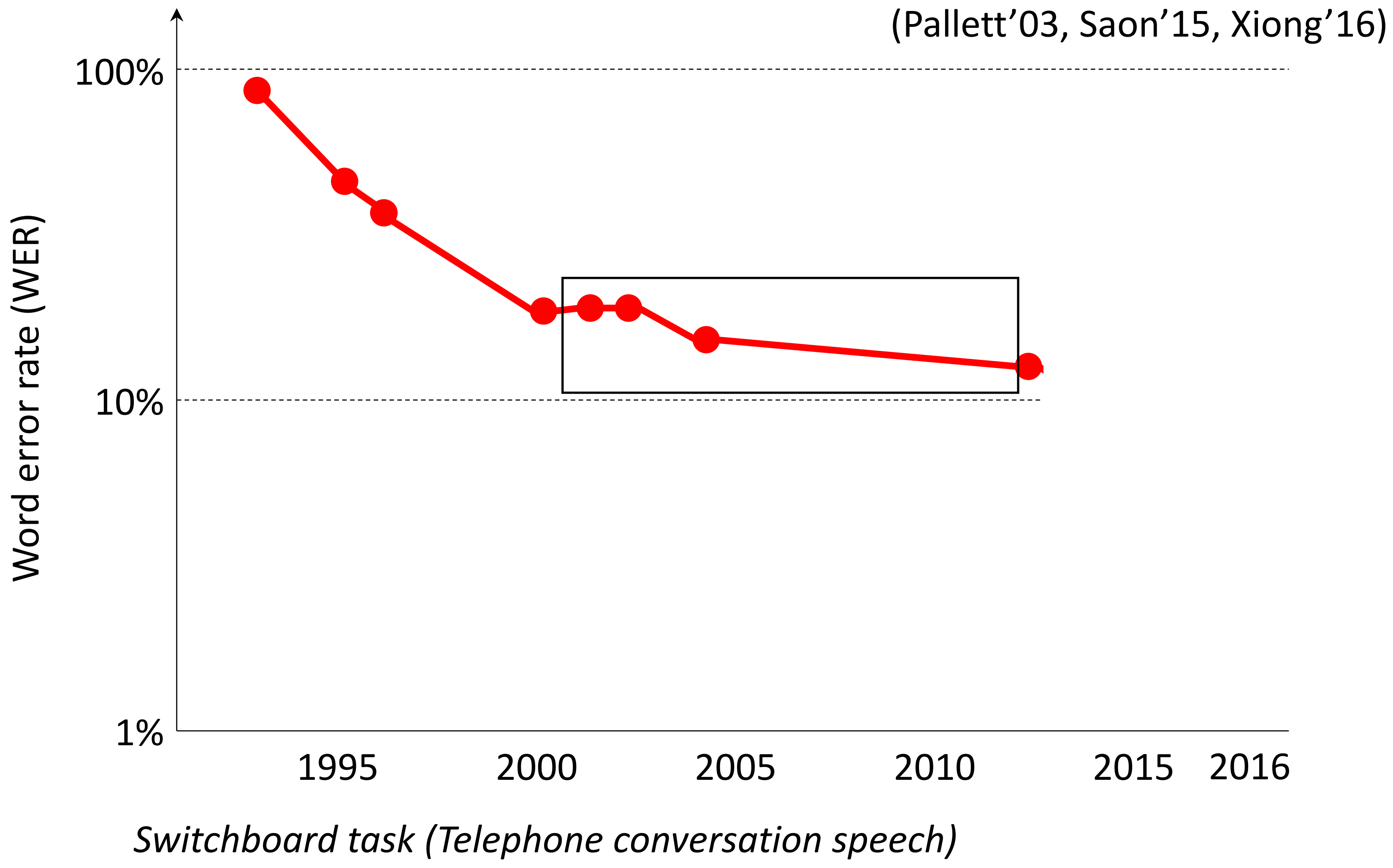
Speech recognition evaluation metric

- Word error rate (WER)
  - Using edit distance word-by-word:

```
Reference)
I want to go to the Johns Hopkins campus
Recognition result)
I want to go to the 10 top kids campus
```

- # insertion errors = 1, # substitution errors = 2, # of deletion errors = 0 ➡ Edit distance = 3
- Word error rate (%): Edit distance (=3) / # reference words (=9) * 100 = 33.3%
- How to compute WERs for languages that do not have word boundaries?
  - Chunking or using character error rate

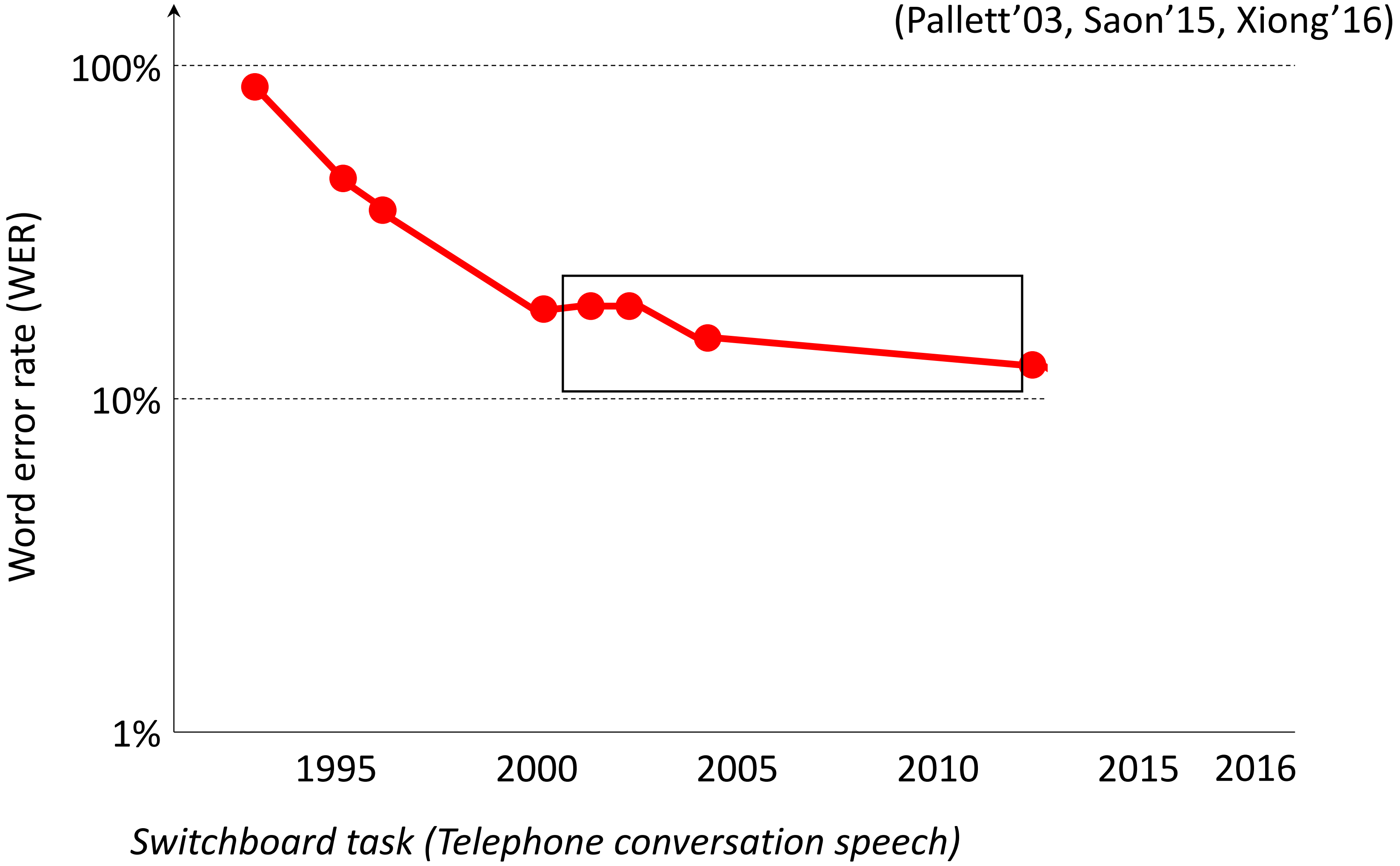
2001: when I started speech recognition….



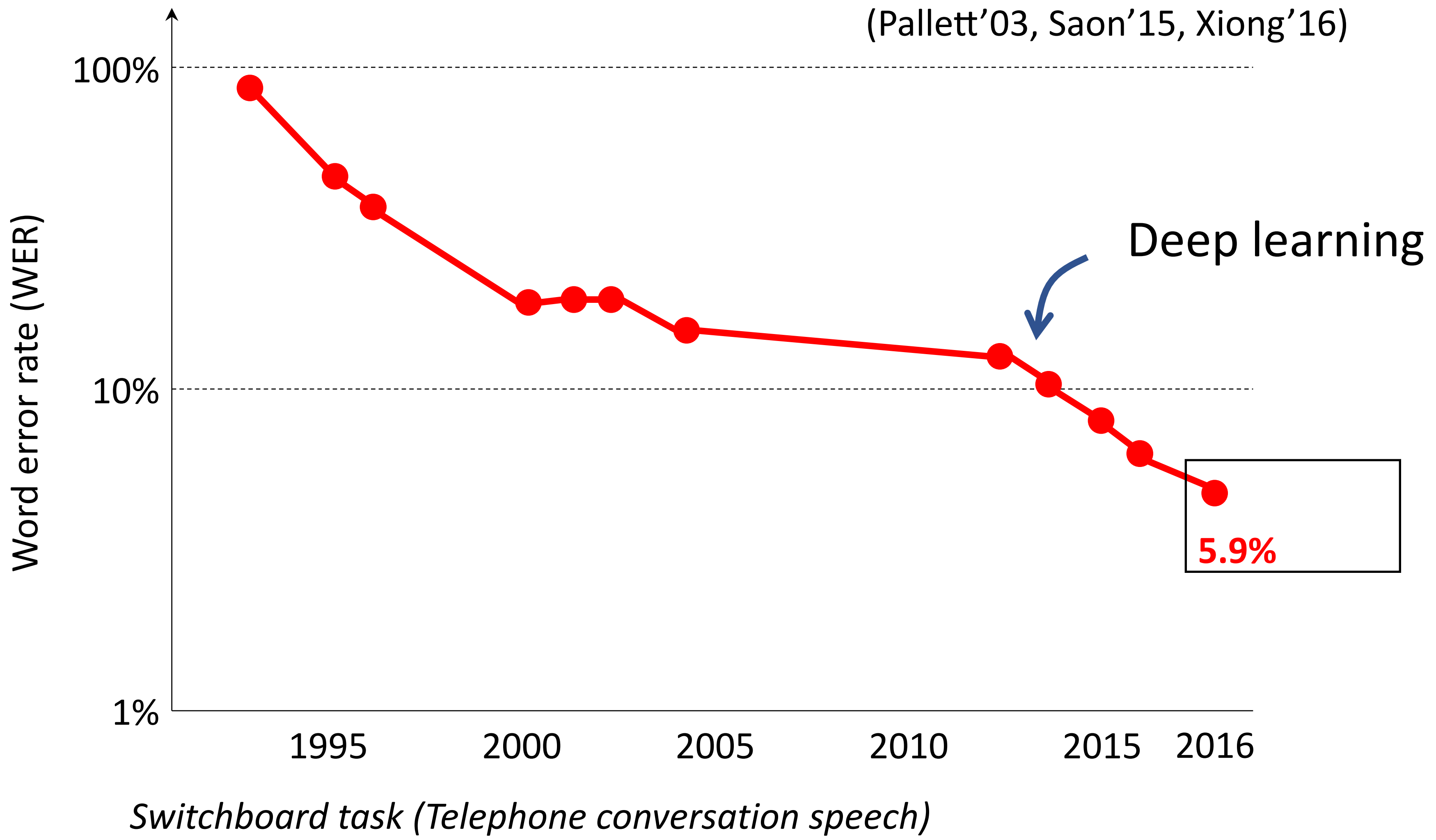
Really bad age….

- No application
- No breakthrough technologies
- Everyone outside speech research criticized it…
- General people don't know "what is speech recognition"

2001: when I started speech recognition….



Now we are at


(Pallett'03, Saon'15, Xiong'16)
100%
10%
1%
Word error rate (WER)
Deep learning
5.9%
1995
2000
2005
2010
2015
2016
Switchboard task (Telephone conversation speech)

Everything was changed

- No application
- No breakthrough technologies
- Everyone outside speech research criticized it…
- General people don’t know “what is speech recognition”

Everything was changed

- ~~No application~~ **voice search, smart speakers**
- No breakthrough technologies
- Everyone outside speech research criticized it…
- General people don’t know “what is speech recognition”

Everything was changed

- ~~No application~~ **voice search, smart speakers**
- ~~No breakthrough technologies~~ **deep neural network**
- Everyone outside speech research criticized it…
- General people don’t know “what is speech recognition”

Everything was changed

- ~~No application~~ **voice search, smart speakers**
- ~~No breakthrough technologies~~ **deep neural network**
- ~~Everyone outside speech research criticized it…~~ **many people outside speech research know/respect it**
- General people don't know "what is speech recognition"

Everything was changed

- ~~No application~~ **voice search, smart speakers**
- ~~No breakthrough technologies~~ **deep neural network**
- ~~Everyone outside speech research criticized it…~~ **many people outside speech research know/respect it**
- Ge~~neral people don’t know “what is speech recognition”~~ **now my kids know what I’m doing**

Today's agenda

- Big impact in HLT technologies including speech recognition
  - Before/after deep neural network
- Brief introduction of deep neural network
- Why deep neural network after 2010?
- Success in the other HLT areas (image and text processing)

Supervised training

- Classification

Please list an HLT related example of these problems

- Binary classification

- Regression

Supervised training

- Classification (mostly in HLT applications)
  - Phoneme or word recognition given speech
  - Word prediction given history (language modeling)
  - Image classification
- Binary classification
  - Special case of classification problems (only two cases)
  - Segmentation
- Regression
  - Predict a clean speech spectrogram from a noisy speech spectrogram

Learning based (data-driven) HLT techniques

- How to formulate our problems with simple classification problems
  - Pairs of supervised data (X, Y)
  - X: input
  - Y: output
- Then, we can train a classifier from the supervised data
  - $p_\theta(Y|X)$ where θ is a set of model parameters

Speech recognition as a classification problem

Word (Y)	**Speech sample (X)**
Yes	
No	
one	
four	

- Use probability distribution $p(Y|X)$, and pick up the most possible word
- Note that this is a very simple problem. The real speech recognition needs to consider a sequential problem

Google speech command database https://ai.googleblog.com/2017/08/launching-speech-commands-dataset.html

Language model as a classification problem

- “I want to go to my office”: correct sentence

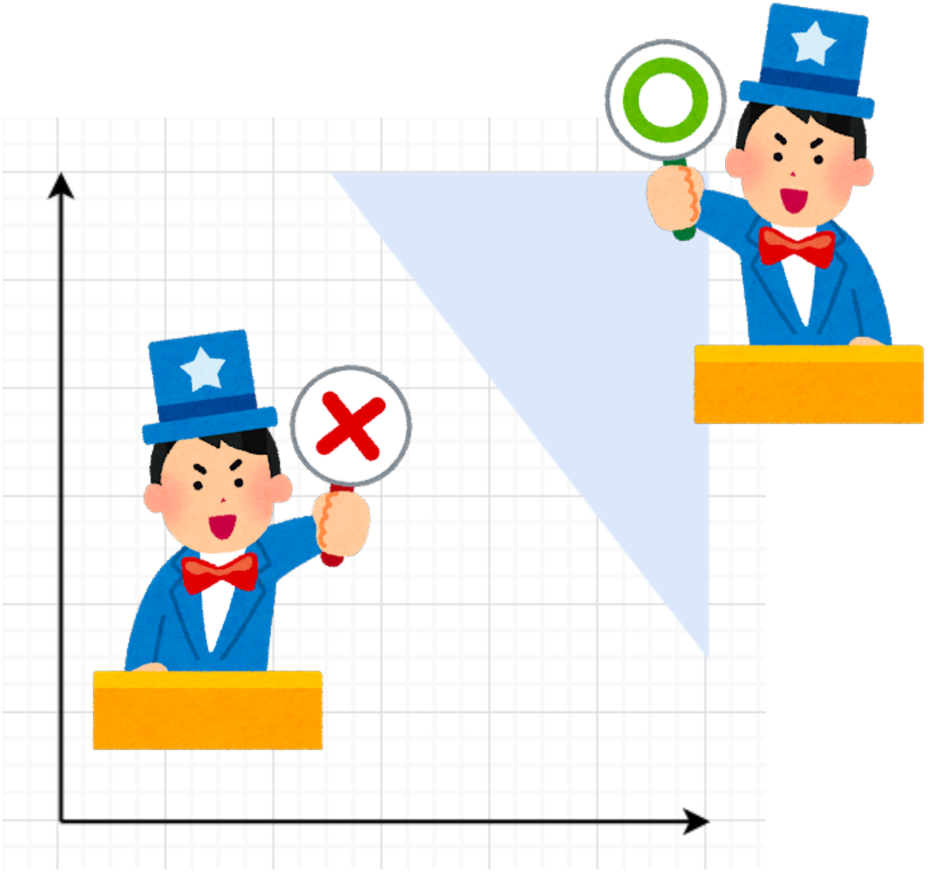
- “I want to go to my [redacted]”: language modeling is to predict [redacted] given history

- Classification problem
  - X: “I want to go to my”
  - Y: “office”
  - Use probability distribution $p(Y|X)$, and pick up the most possible word

- $P(Y|X)$ is obtained by a deep neural network

Binary classification

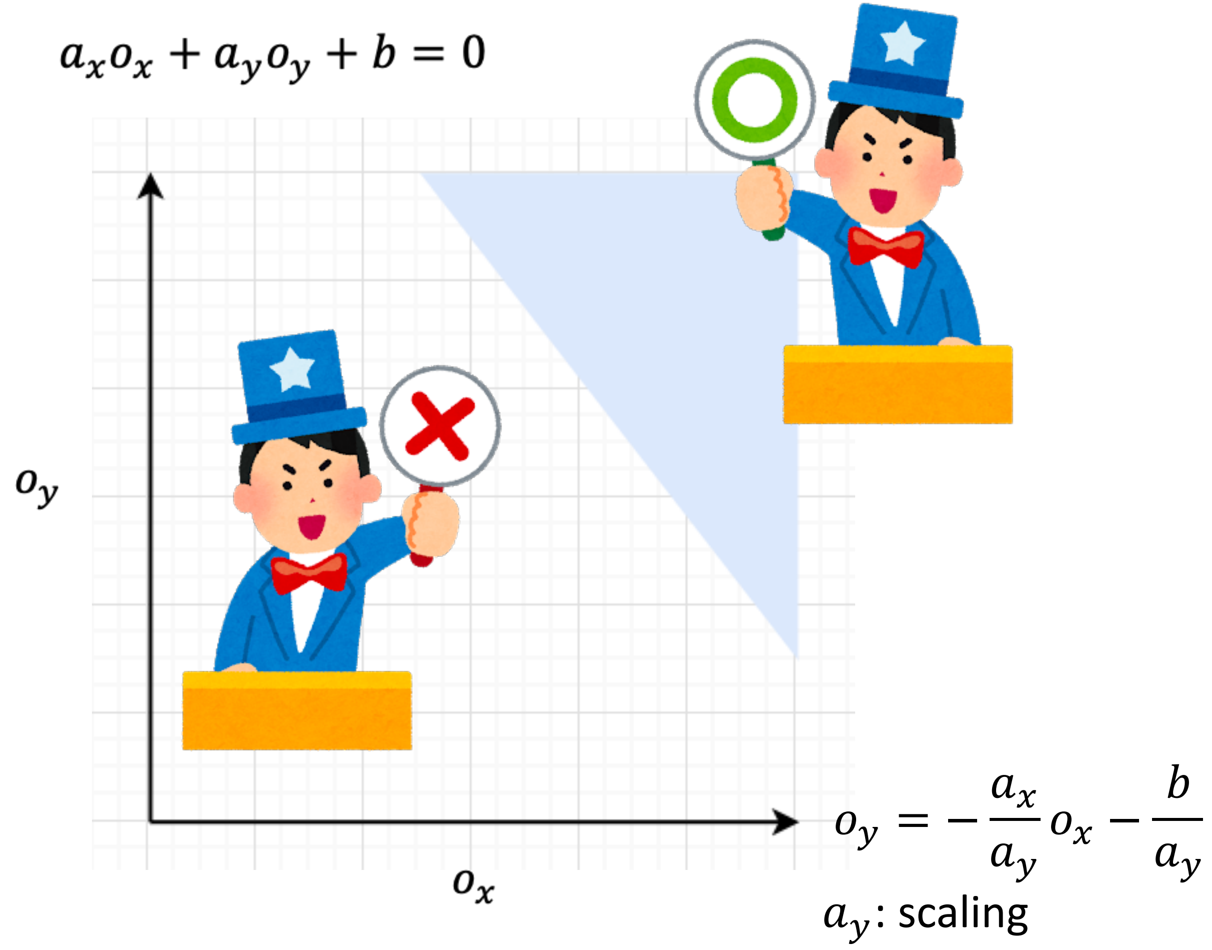
- Two category classification problems



Binary classification
(we will have more discussion later)

- We can use a linear classifier
  - ◯: $a_x o_x + a_y o_y + b > 0$
  - ×: $a_x o_x + a_y o_y + b < 0$
- We can also make a probability with the sigmoid function $\sigma(\quad)$
  - $p(\circ | o_x, o_y) = \sigma(a_x o_x + a_y o_y + b)$
  - $p(\times | o_x, o_y) = 1 - \sigma(a_x o_x + a_y o_y + b)$
- Sigmoid function

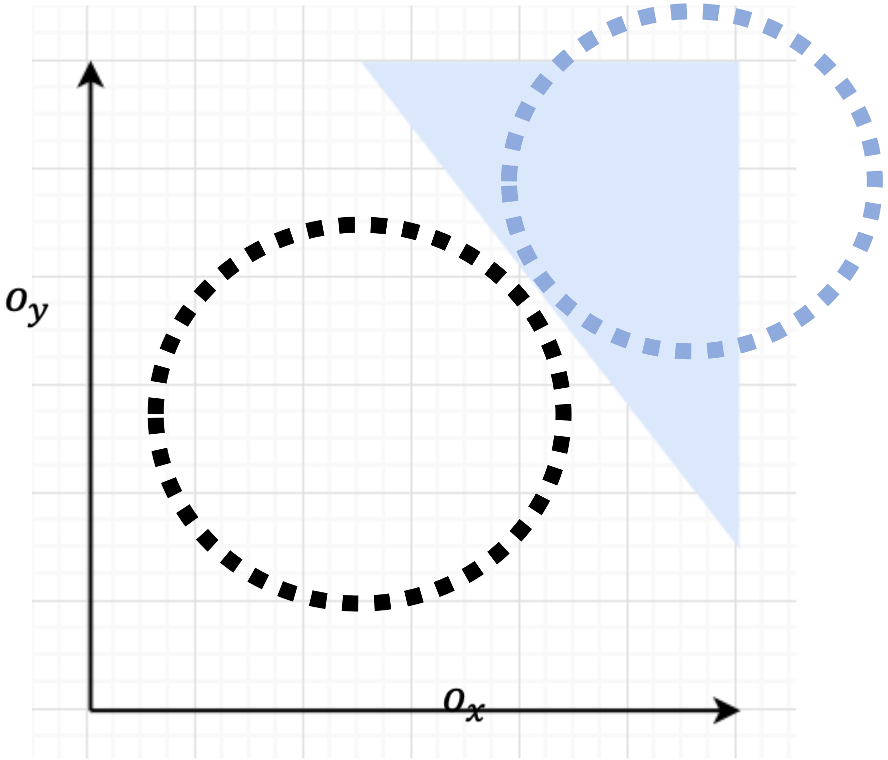
$$\sigma(x) = \frac{1}{1 + e^{-x}}$$



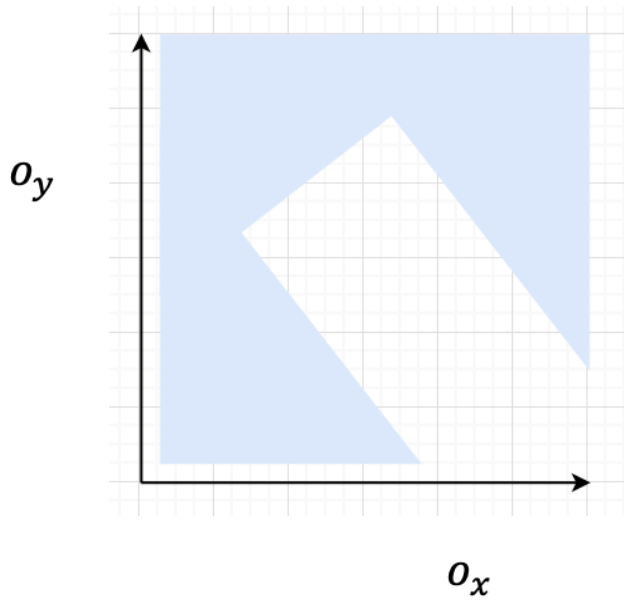
from http://cs.jhu.edu/~kevinduh/a/deep2014/140114-ResearchSeminar.pdf

Binary classification

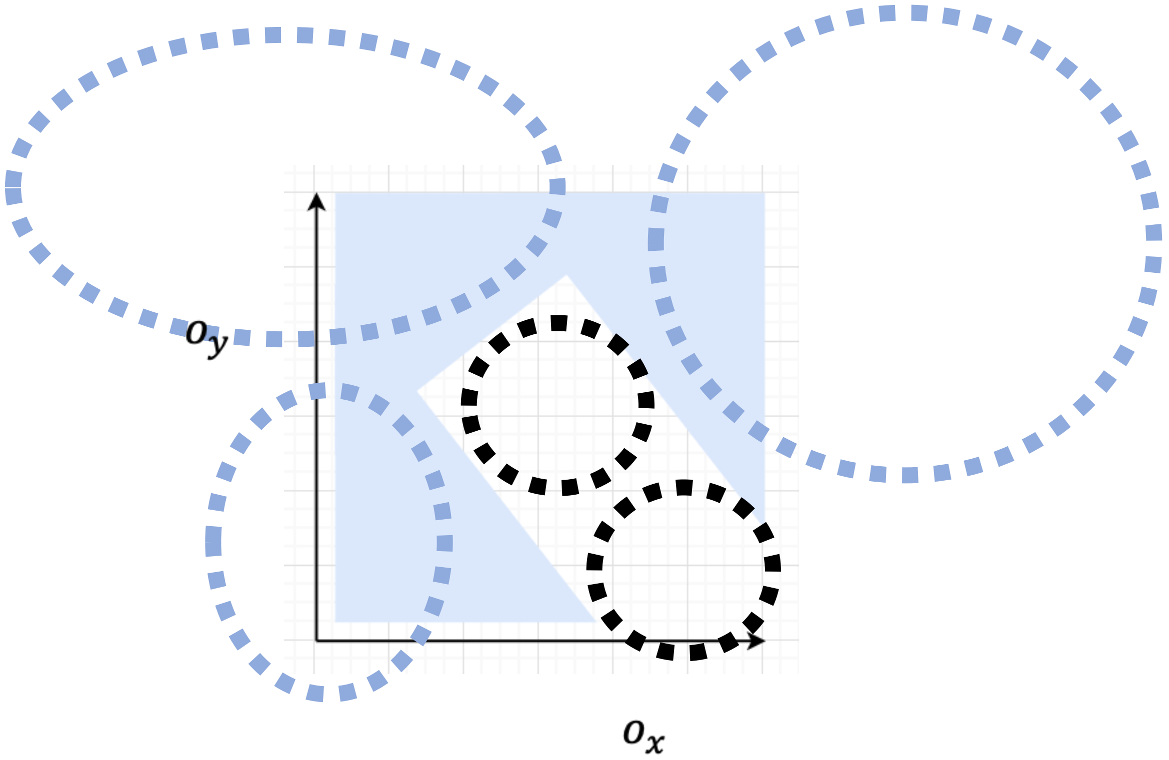
- We can use a GMM (although not suitable) to model the data distribution
  - $p\left(o_x, o_y \middle| \circ\right) = \sum_k \omega_k \, N(\boldsymbol{o}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
  - $p\left(o_x, o_y \middle| \times\right) = \sum_k \omega'_k \, N\left(\boldsymbol{o}\middle|\boldsymbol{\mu}'_k, \boldsymbol{\Sigma}'_k\right)$
- We can build a classifier from the Bayes theorem (speech recognition before 2010)
  - $p\left(\circ \middle| o_x, o_y\right) = \frac{p\left(o_x, o_y \middle| \circ\right)}{p\left(o_x, o_y \middle| \circ\right) + p\left(o_x, o_y \middle| \times\right)}$
  - $p\left(\times \middle| o_x, o_y\right) = \frac{p\left(o_x, o_y \middle| \times\right)}{p\left(o_x, o_y \middle| \circ\right) + p\left(o_x, o_y \middle| \times\right)}$



Getting more difficult with the GMM classifier or linear classifier

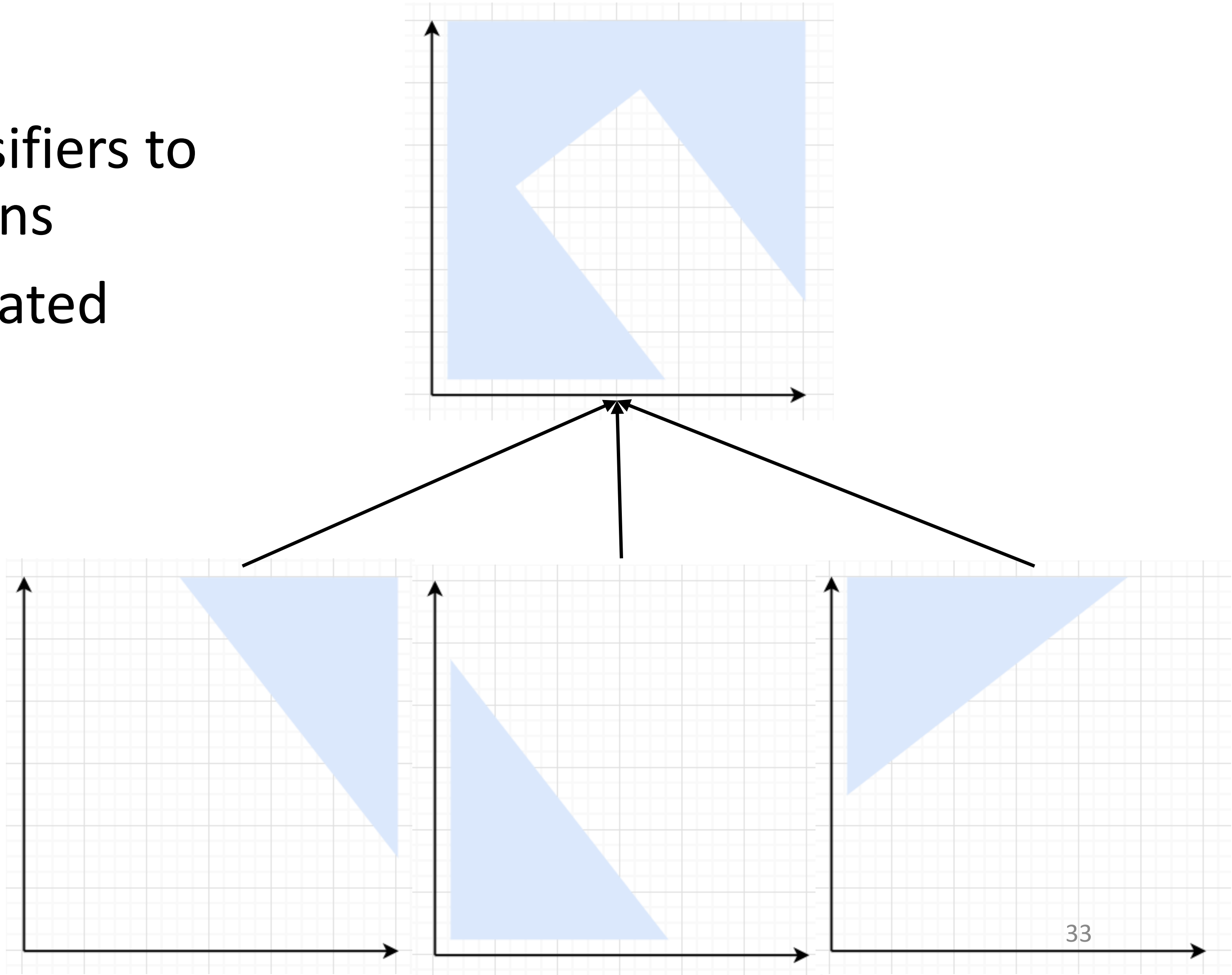


Getting more difficult with the GMM classifier or linear classifier



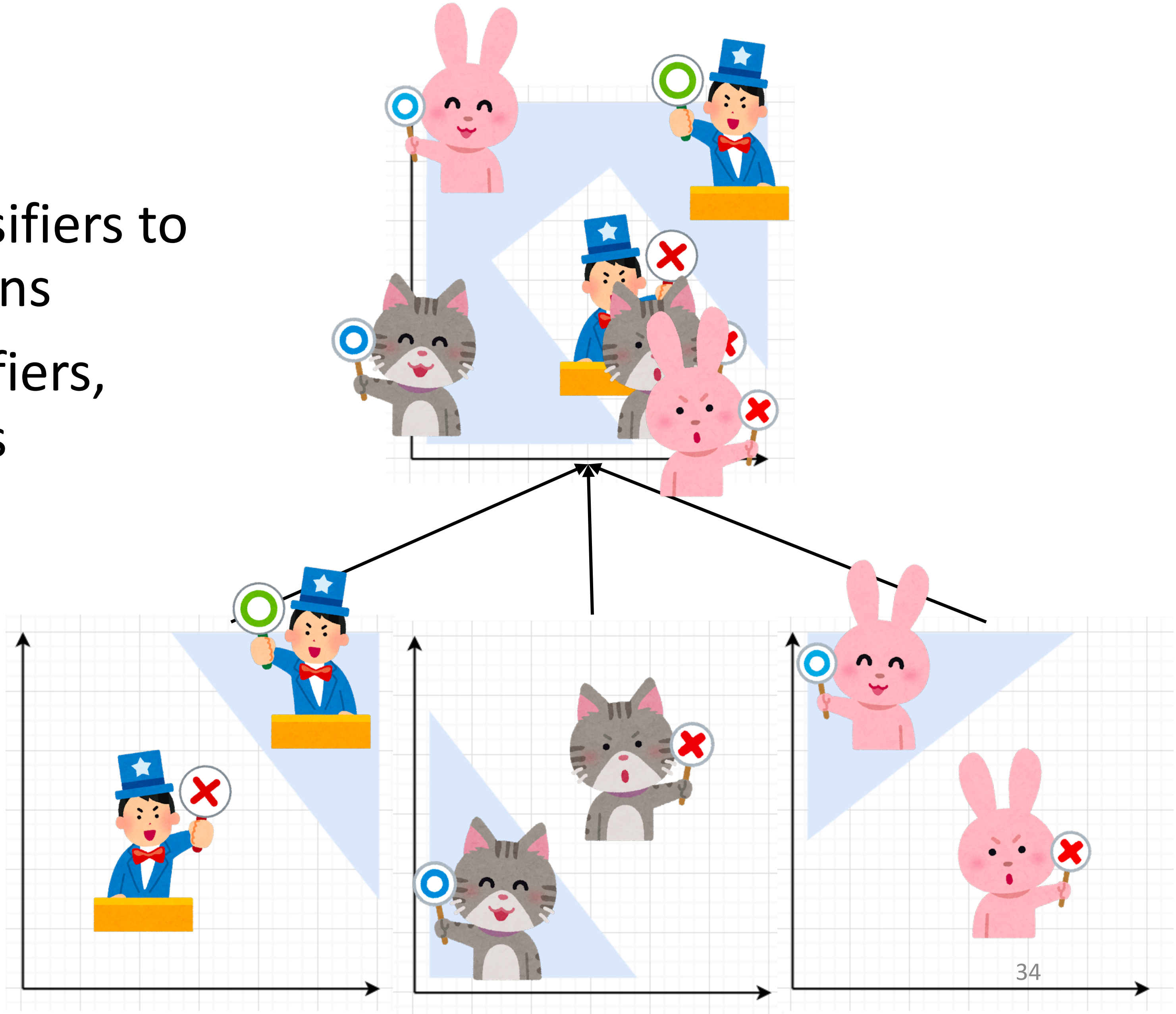
Neural network

- Combination of linear classifiers to classify complicated patterns
- More layers, more complicated patterns

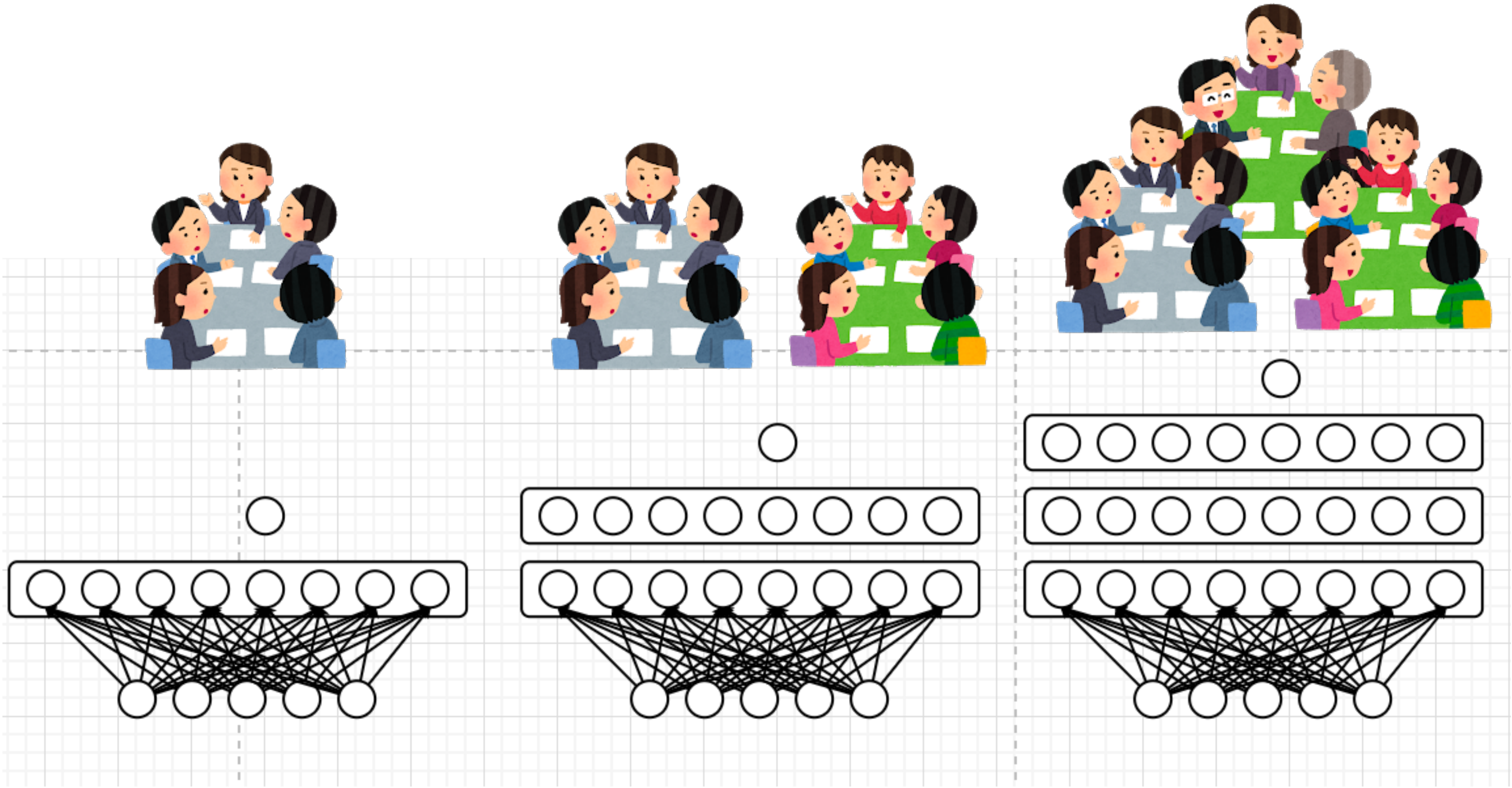


Neural network

- Combination of linear classifiers to classify complicated patterns
- More layers or more classifiers, more complicated patterns

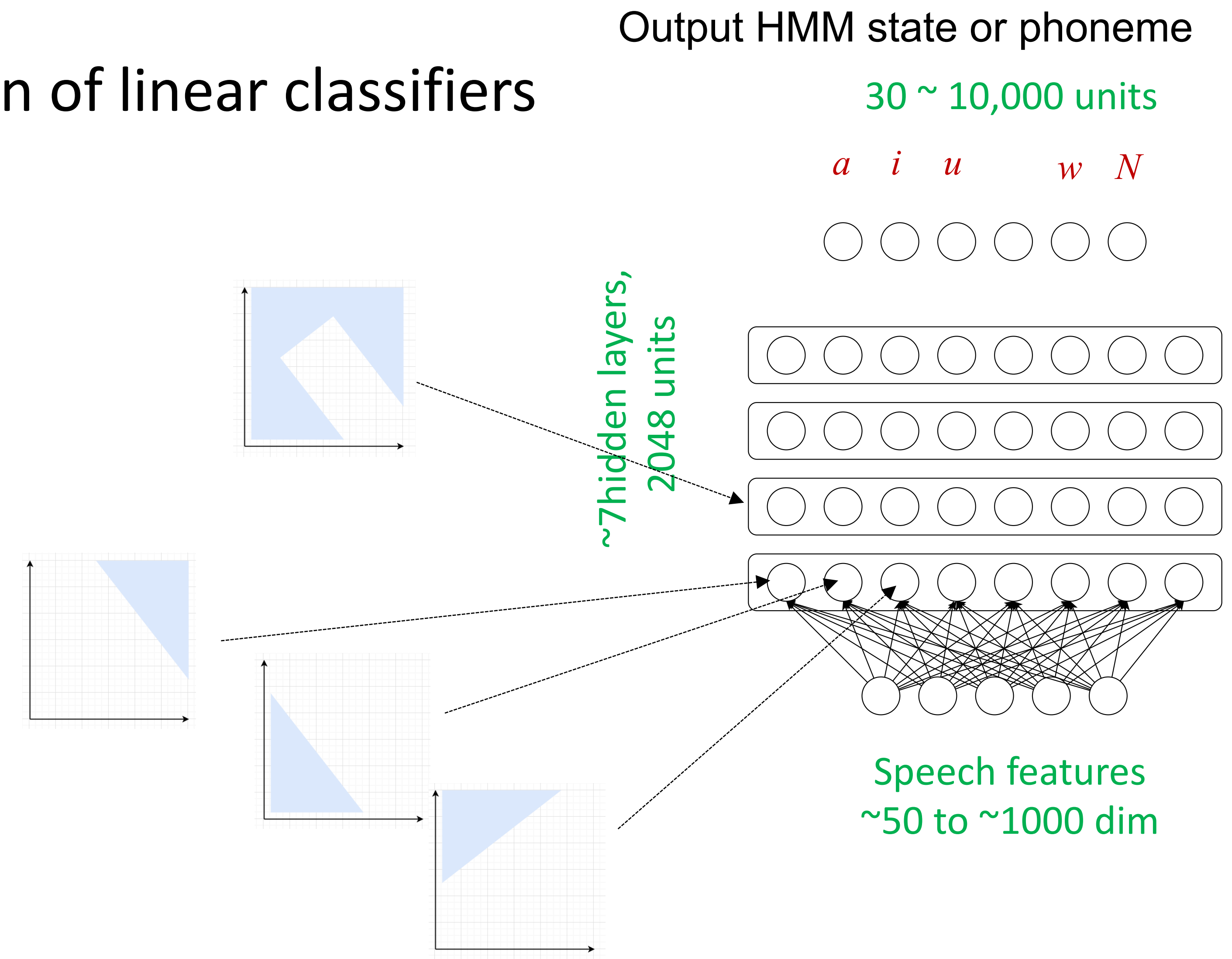


Going deeper-> more accurate



Neural network used in speech recognition

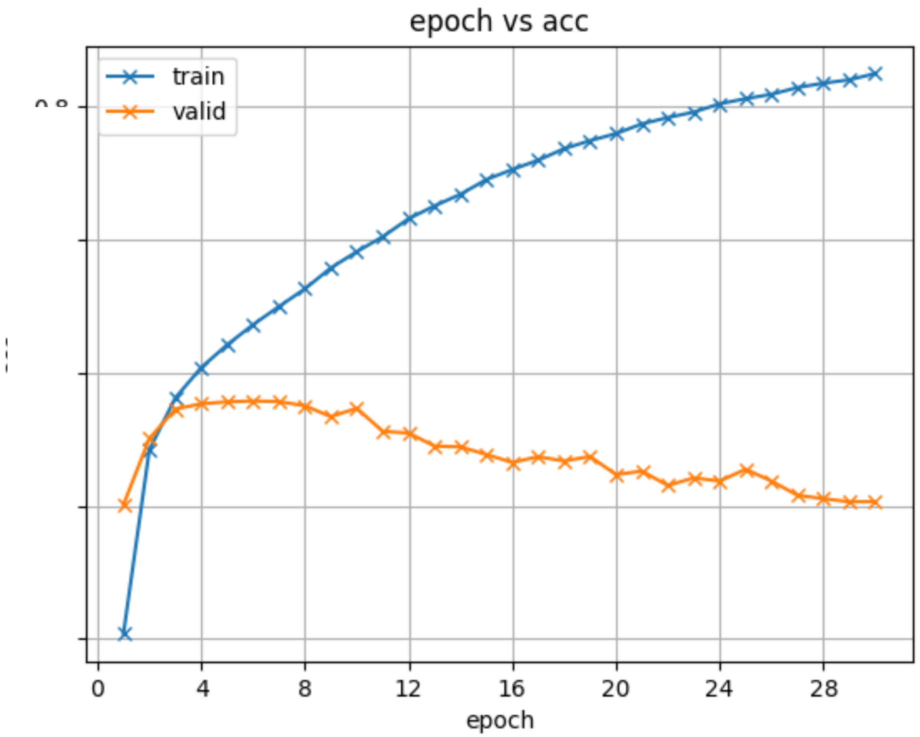
- Very large combination of linear classifiers

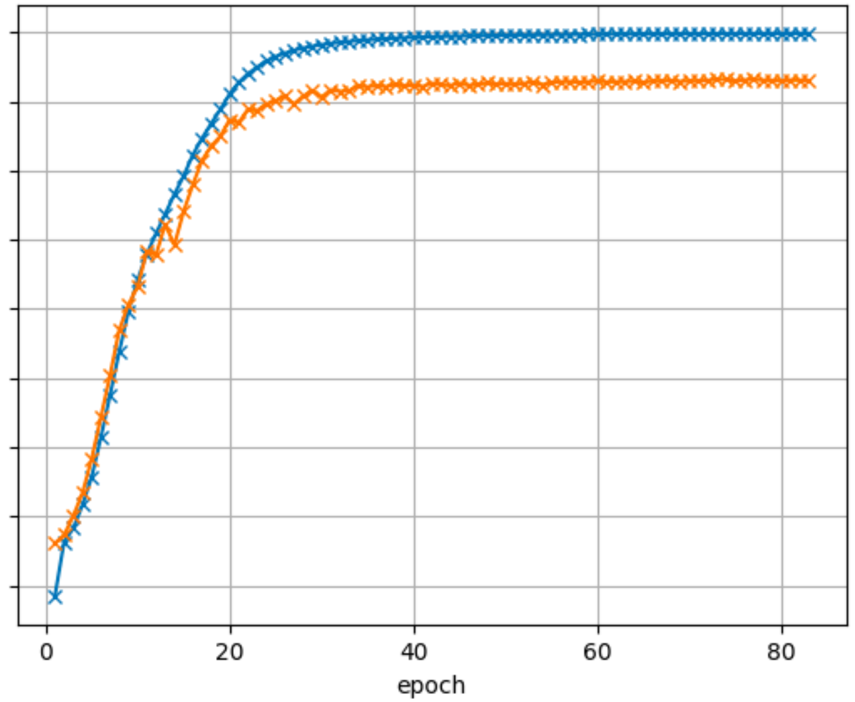


Difficulties of training

Which one is better?

- Blue: accuracy of training data (higher is better)
- Orange: accuracy of validation data (higher is better)





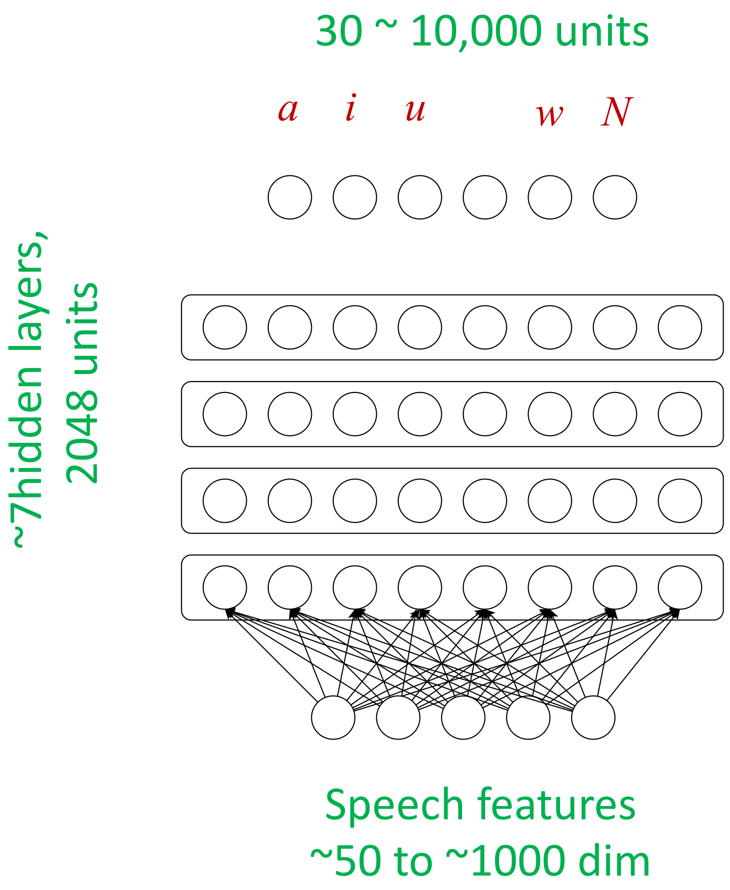
Today’s agenda

- Big impact in HLT technologies including speech recognition
  - Before/after deep neural network
- Brief introduction of deep neural network
- Why deep neural network after 2010?
- Success in the other HLT areas (image and text processing)

Why neural network was not focused

1. Very difficult to train
   - Batch? On-line? Mini-batch?
   - Stochastic gradient decent
     - Learning rate? Scheduling?
   - What kind of topologies?
   - Large computational cost

2. The amount of training data is very critical
3. CPU -> GPU



Before deep learning (2002 – 2009)

- Success of neural networks was very old period
- People believed that GMM was better
- But very small gain from standard GMMs

from https://en.wikipedia.org/wiki/Geoffrey_Hinton

When I noticed deep learning (2010)

- A. Mohamed, G. E. Dahl, and G. E. Hinton, "Deep belief networks for phone recognition," in NIPS Workshop on Deep Learning for Speech Recognition and Related Applications, 2009.

Table 4: *Reported results on TIMIT core test set*

Method	PER
Stochastic Segmental Models [28]	36%
Conditional Random Field [29]	34.8%
Large-Margin GMM [30]	33%
CD-HMM [4]	27.3%
Augmented conditional Random Fields [4]	26.6%
Recurrent Neural Nets [31]	26.1%
Bayesian Triphone HMM [32]	25.6%
Monophone HTMs [33]	24.8%
Heterogeneous Classifiers [34]	24.40%
Deep Belief Networks(DBNs) (this work)	**23.0%**

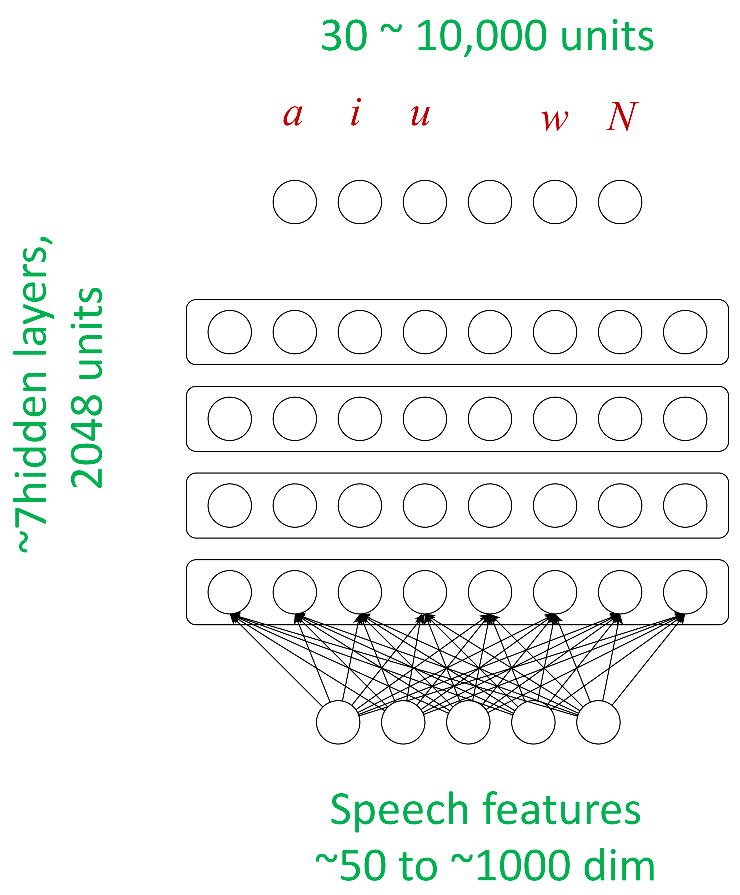
- Using deep belief network as **pre-training**
- **Fine-tuning** deep neural network

→ Provides stable estimation

- This still did not fully convince me (I introduced it at NTT's reading group)

Pre-training and fine-tuning

- First train neural network like parameters with deep belief network or autoencoder
- Then, using deep neural network training



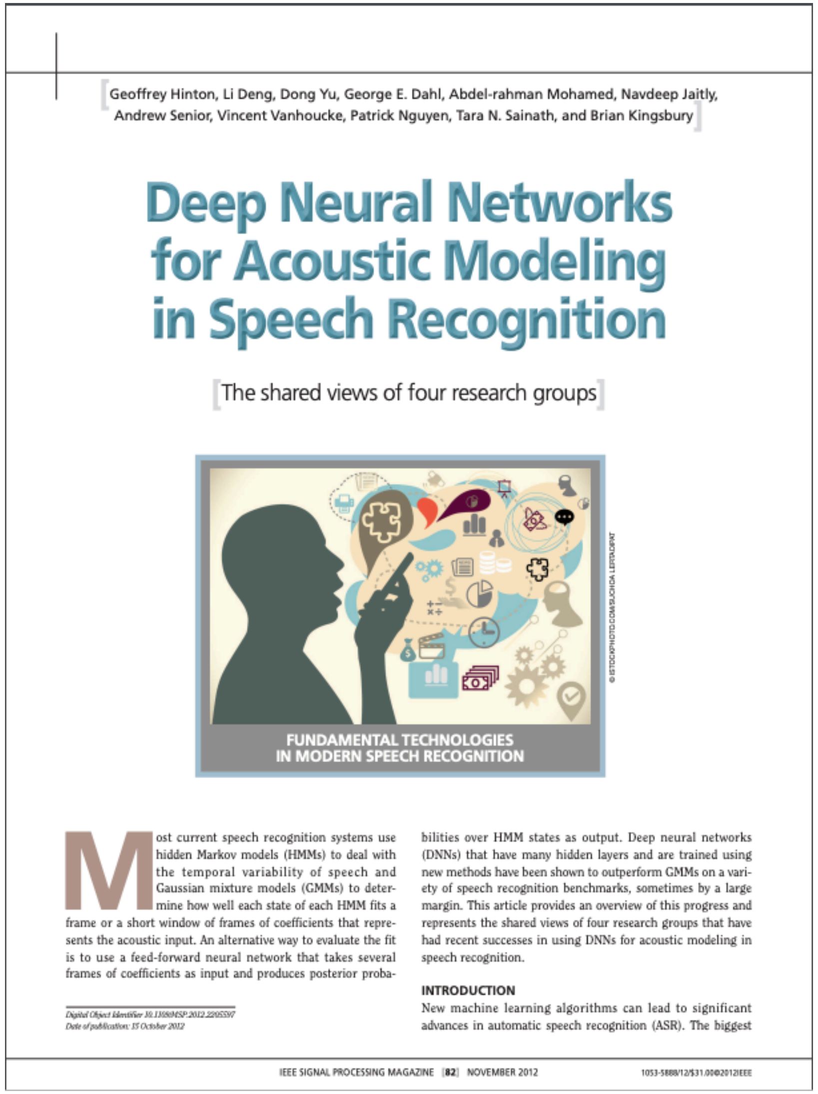
Interspeech 2011 at Florence

- The following three papers convinced me
  - Feature extraction: Valente, Fabio / Magimai-Doss, Mathew / Wang, Wen (2011): "Analysis and comparison of recent MLP features for LVCSR systems", In *INTERSPEECH-2011*, 1245-1248.
  - Acoustic model: Seide, Frank / Li, Gang / Yu, Dong (2011): "Conversational speech transcription using context-dependent deep neural networks", In INTERSPEECH-2011, 437-440.
  - Language model: Mikolov, Tomáš / Deoras, Anoop / Kombrink, Stefan / Burget, Lukáš / Černocký, Jan (2011): "Empirical evaluation and combination of advanced language modeling techniques", In INTERSPEECH-2011, 605-608.
- I discussed this potential to my NLP folks in NTT but they did not believe it (SVM, log linear model)

Late 2012

- My first deep learning (Kaldi nnet)
  - Kaldi started to support DNN since 2012 (mainly developed by Karel Vesely)
    - Deep belief network based pre-training
    - Feed forward neural network
    - Sequence-discriminative training

	Hub5 ‘00 (SWB)	WSJ
GMM	18.6	5.6
DNN	14.2	3.6
DNN with sequence-discriminative training	12.6	3.2



Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury

Deep Neural Networks for Acoustic Modeling in Speech Recognition

The shared views of four research groups

FUNDAMENTAL TECHNOLOGIES IN MODERN SPEECH RECOGNITION

Most current speech recognition systems use hidden Markov models (HMMs) to deal with the temporal variability of speech and Gaussian mixture models (GMMs) to determine how well each state of each HMM fits a frame or a short window of frames of coefficients that represents the acoustic input. An alternative way to evaluate the fit is to use a feed-forward neural network that takes several frames of coefficients as input and produces posterior probabilities over HMM states as output. Deep neural networks (DNNs) that have many hidden layers and are trained using new methods have been shown to outperform GMMs on a variety of speech recognition benchmarks, sometimes by a large margin. This article provides an overview of this progress and represents the shared views of four research groups that have had recent successes in using DNNs for acoustic modeling in speech recognition.

Digital Object Identifier 10.1109/MSP.2012.2205597
Date of publication: 15 October 2012

INTRODUCTION

New machine learning algorithms can lead to significant advances in automatic speech recognition (ASR). The biggest

1053-5888/12/$31.00©2012IEEE

Build speech recognition with public tools and resources

- TED-LIUM (~100 hours)
- LIBRISPEECH (~1000 hours)

- We can build Kald+DNN+TED-LIUM to make a English speech recognition system by using one machine (GPU + many core machines)
- Before this, it’s only realized by a big company.

Same things happened in *computer vision*

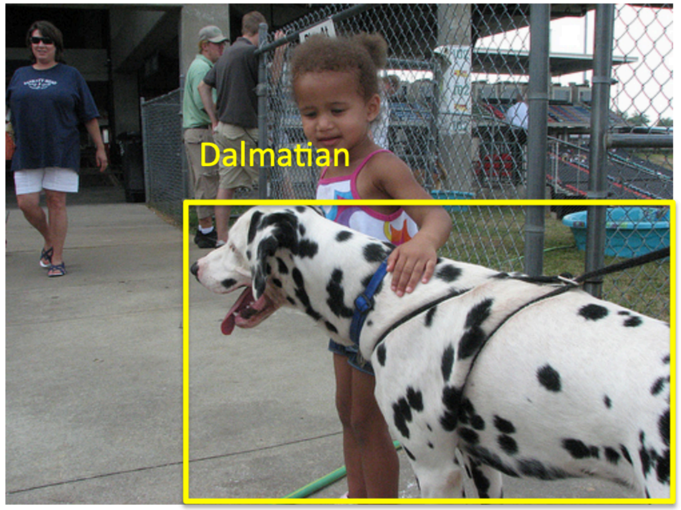
from https://en.wikipedia.org/wiki/Geoffrey_Hinton

ImageNet challenge (Large scale data)

IMAGENET Large Scale Visual Recognition Challenge (ILSVRC) 2010-2012

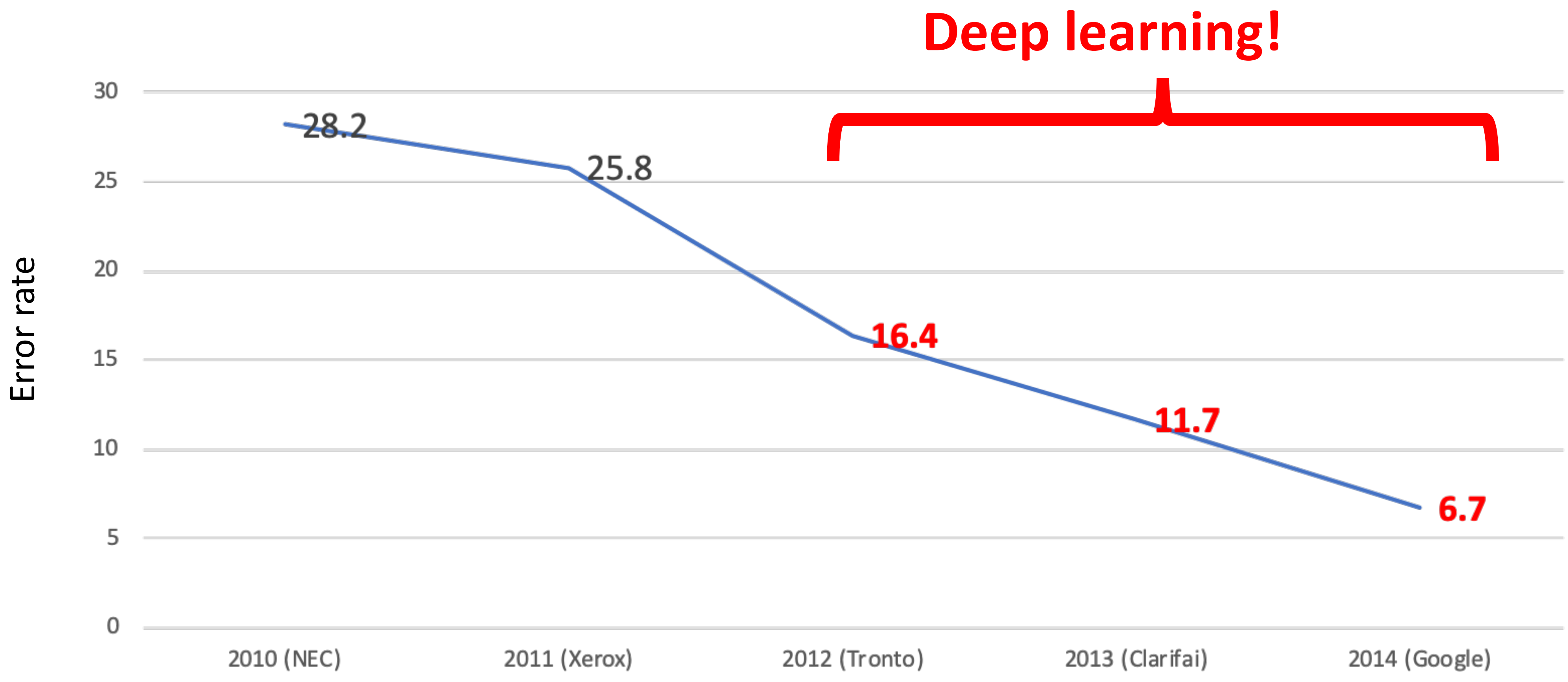
~~20 object classes~~ ~~22,591 images~~

1000 object classes **1,431,167 images**



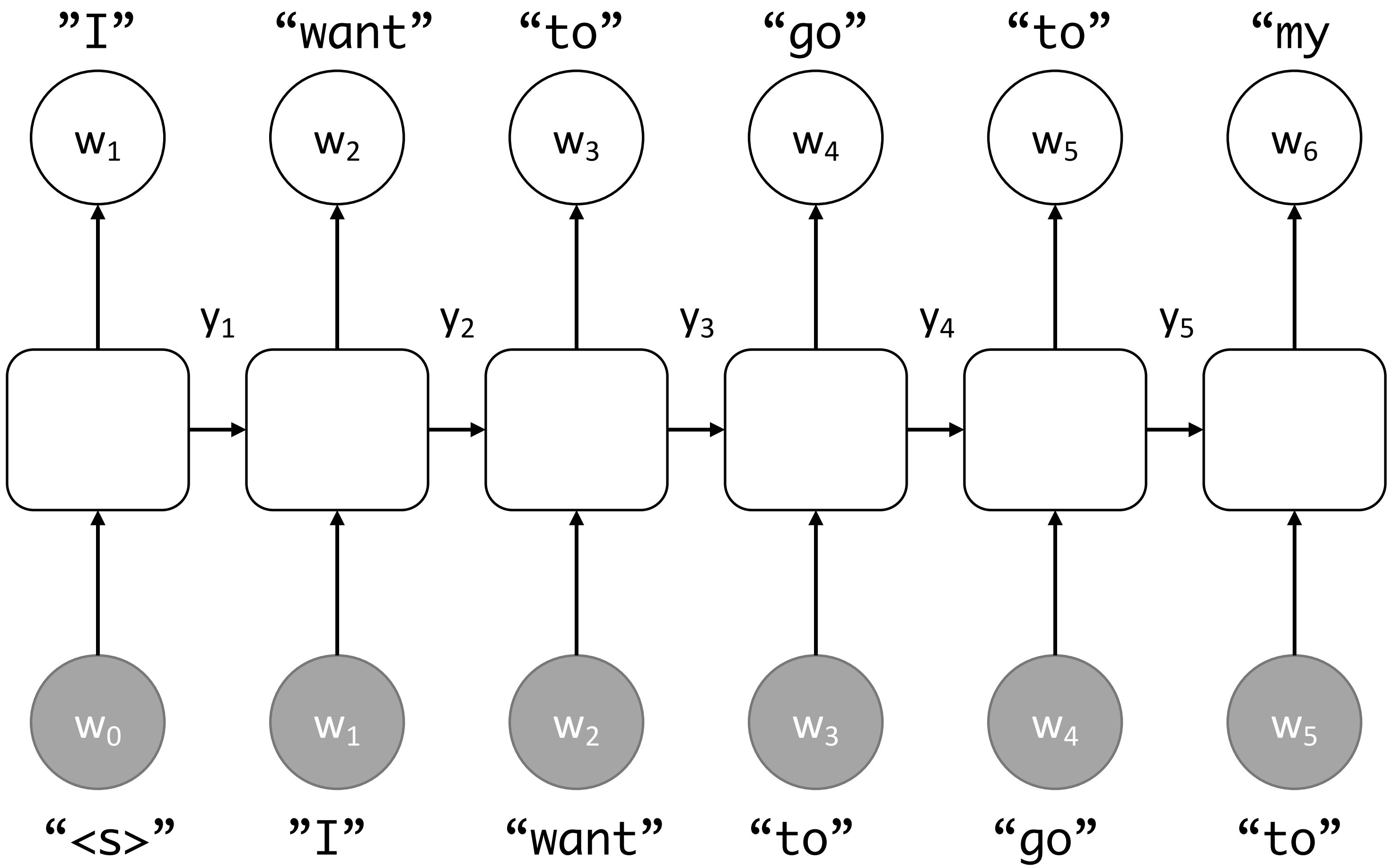
L. Fei-Fei and O. Russakovsky, **Analysis of Large-Scale Visual Recognition**, *Bay Area Vision Meeting, October, 2013*

ImageNet challenge
AlexNet, GoogLeNet, VGG, ResNet, …



Same things happened in *text processing*

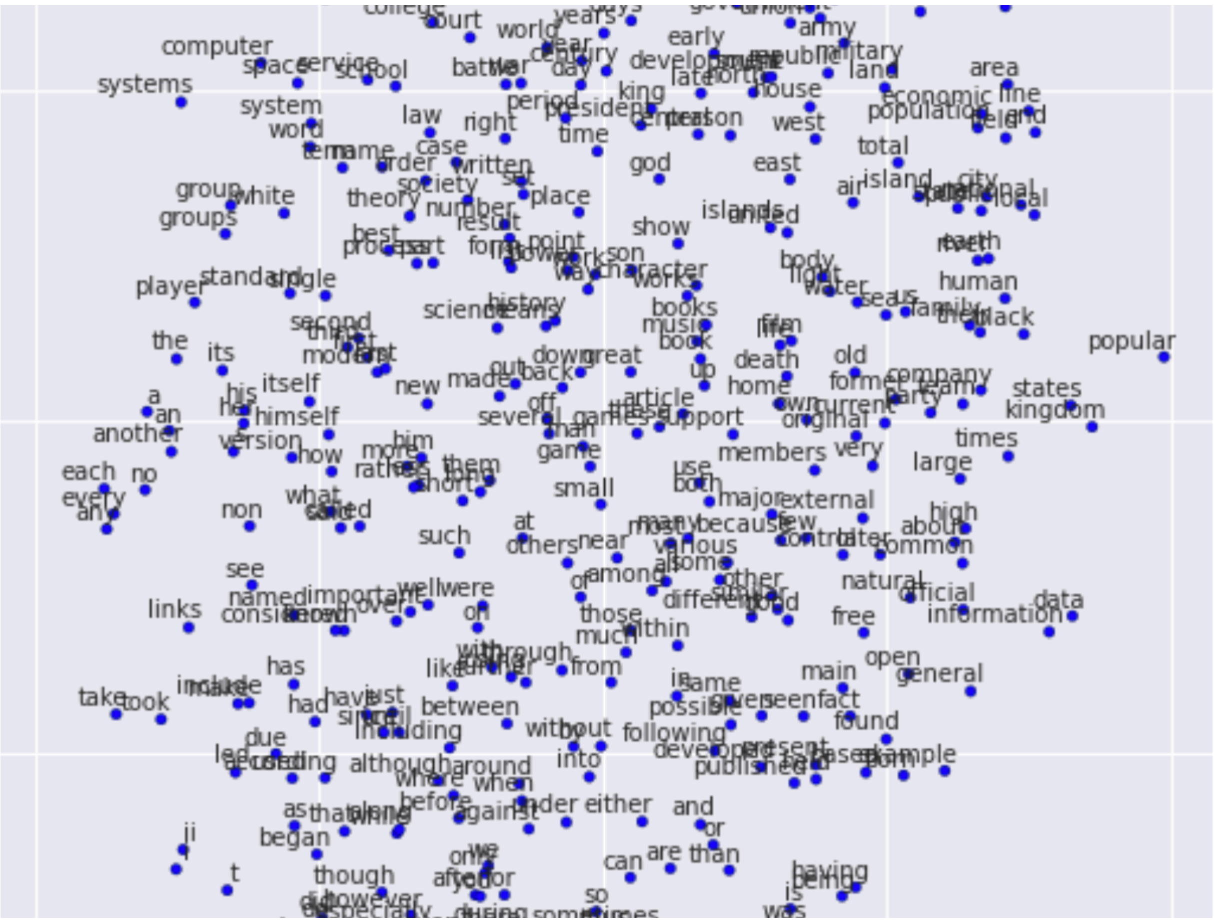
- Recurrent neural network language model (RNNLM) [Mikolov+ (2010)]



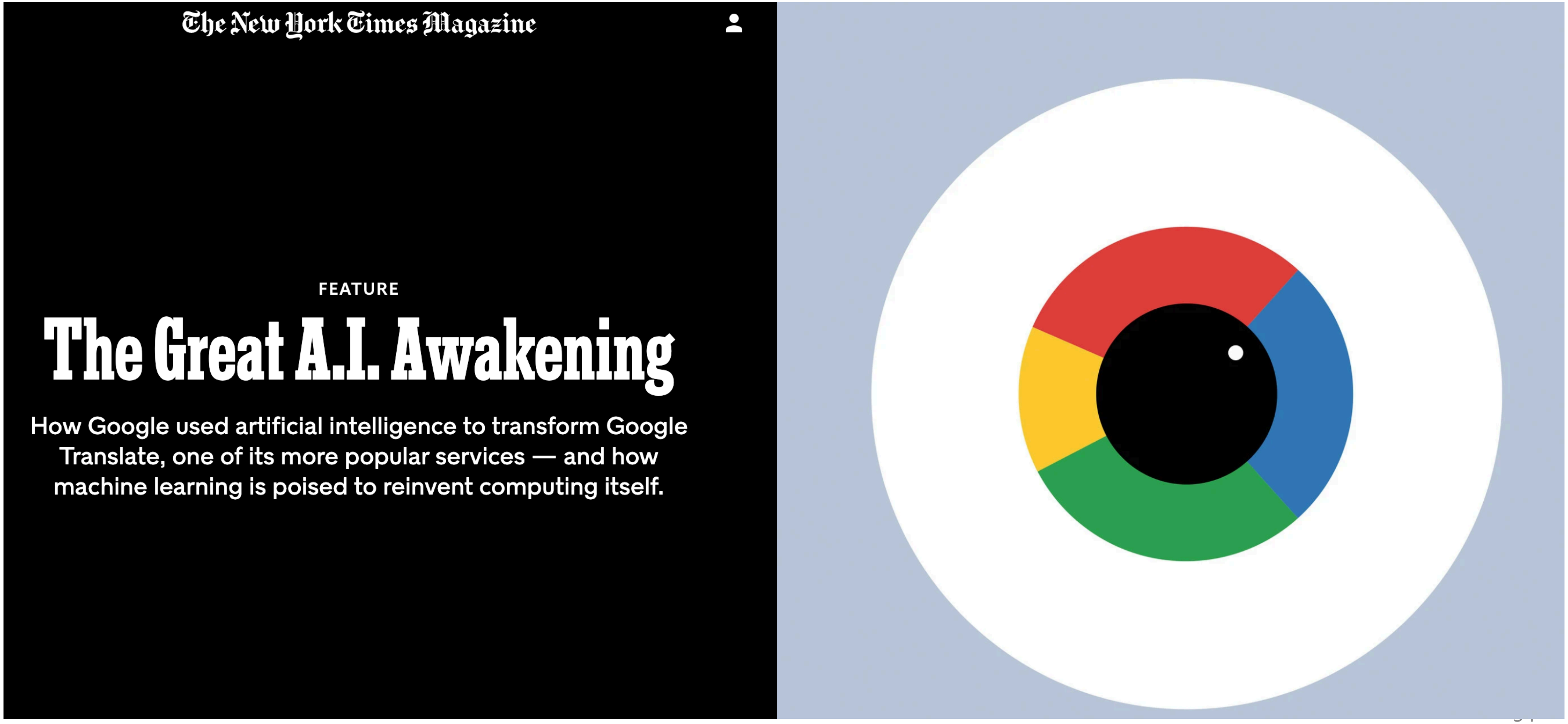
	Perplexity
N-gram (conventional)	336
RNNLM	**156**

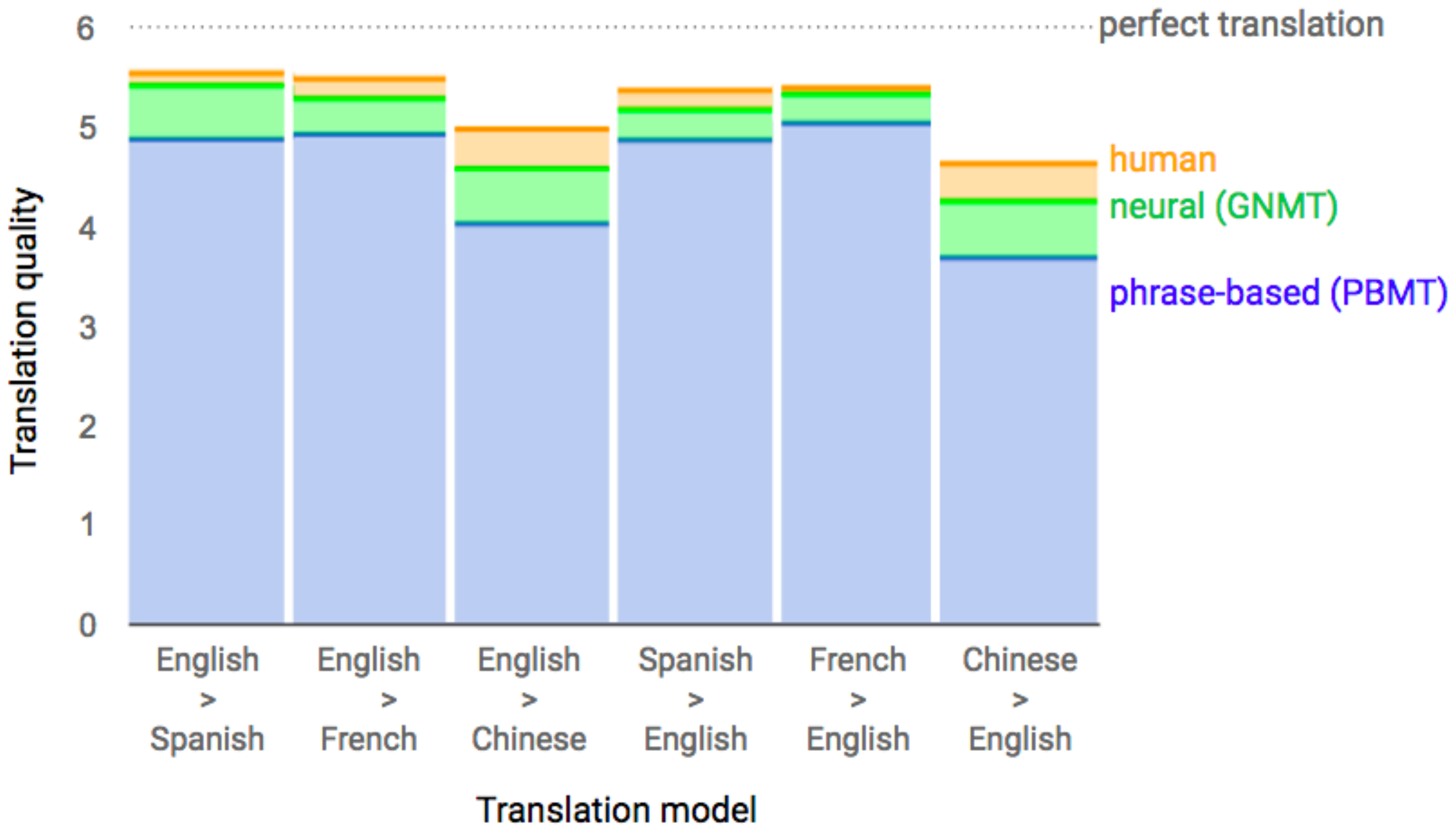
Word embedding example

https://www.tensorflow.org/tutorials/word2vec



Neural machine translation
(New York Times, December 2016)





from https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html

Deep neural network toolkit (2013-)

- Theano
- Caffe
- Torch
- CNTK
- Chainer
- Keras

Later

- Tensorflow
- **PyTorch (used in this course)**
- MxNet
- Etc.

Summary

- Before 2011
  - GMM/HMM, limitation of the performance, bit boring
  - This is because they are linear models…
- After 2011
  - DNN/HMM
  - Toolkit
  - Public large data
  - GPU
  - NLP, image/vision were also moved to DNN
  - Always something exciting

Any questions?