E9 205 Machine Learning for Signal Processing

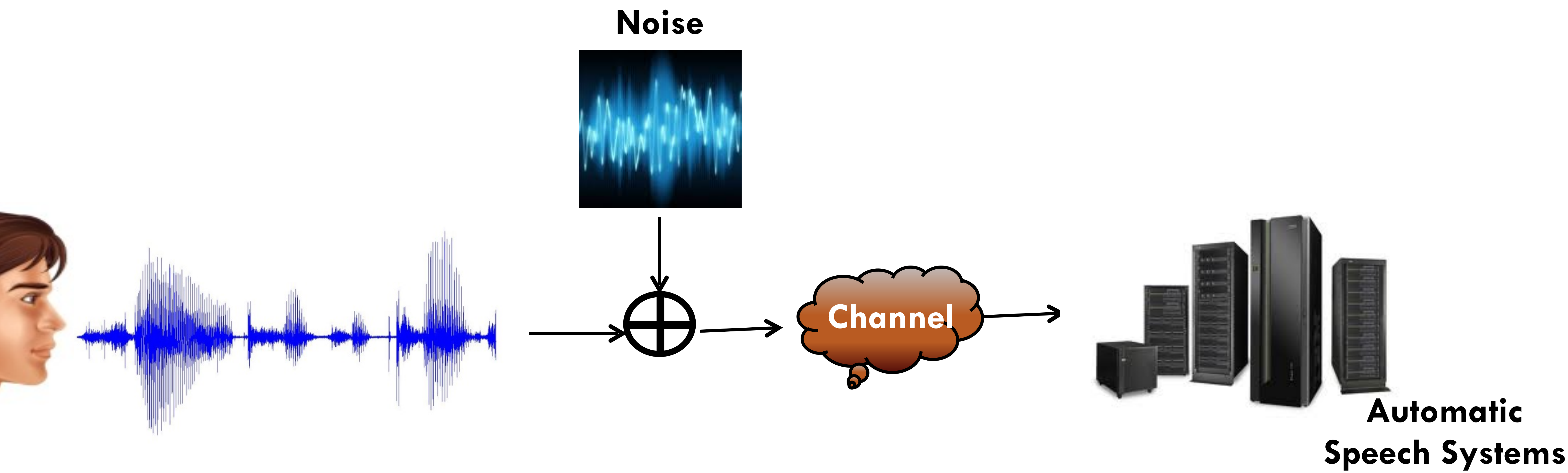
Deep Learning for Audio and Vision

20-11-2019

Established 1911

INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

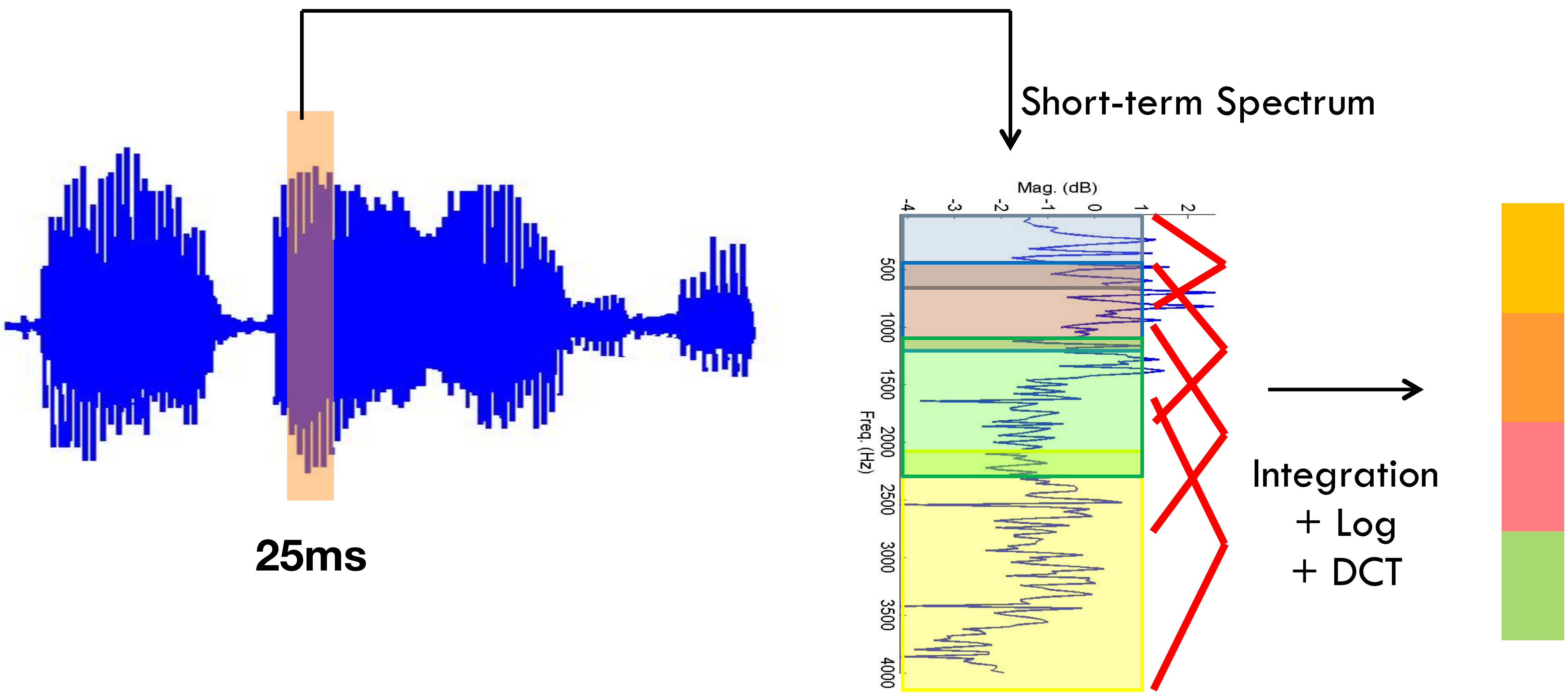
Speech Recognition



Courtesy – Google Images

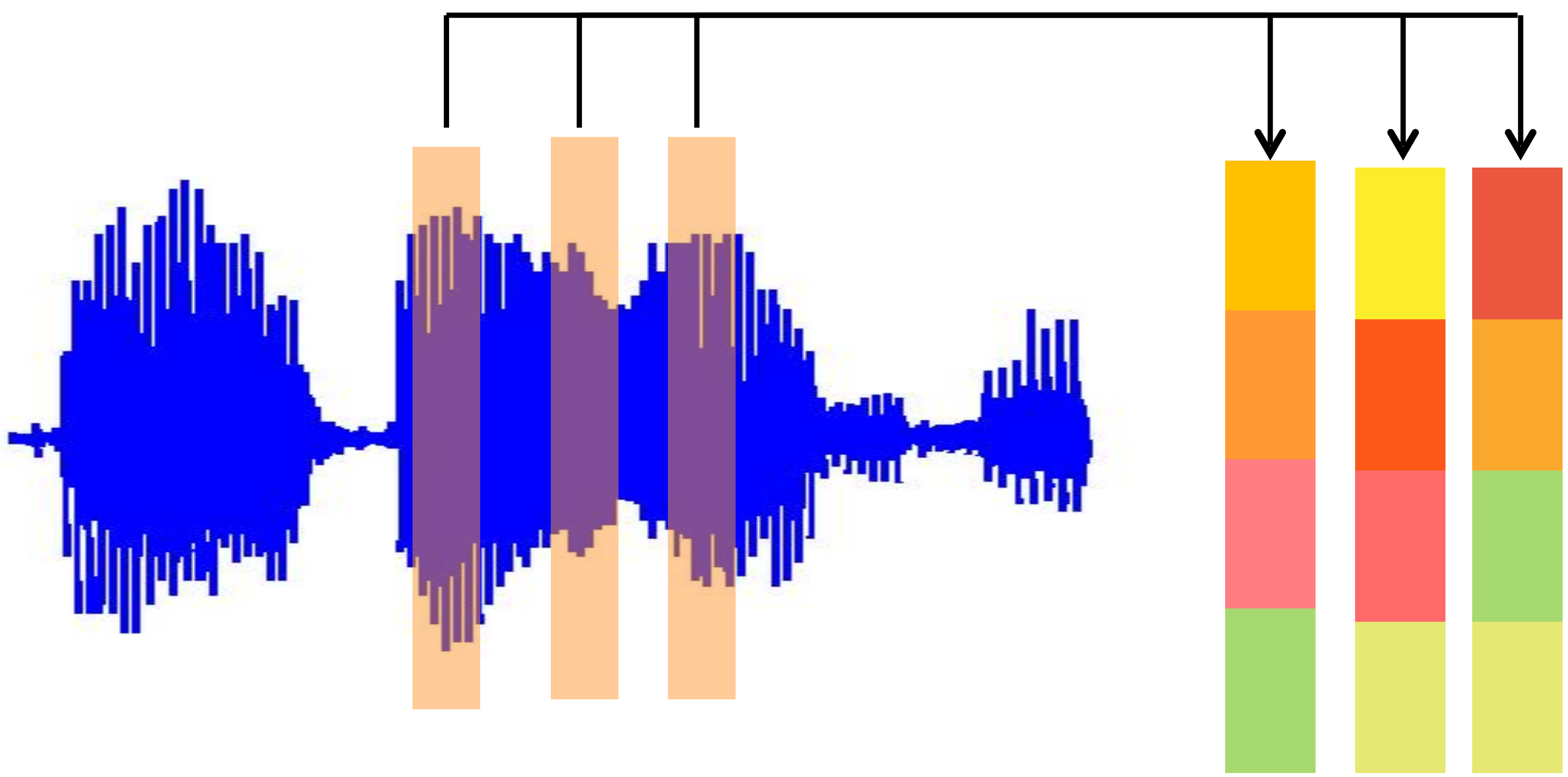
Signal Modeling

- Short-term spectra integrated in mel frequency bands followed by log compression + DCT – mel frequency cepstral coefficients (MFCC) [Davis and Mermelstein, 1979].



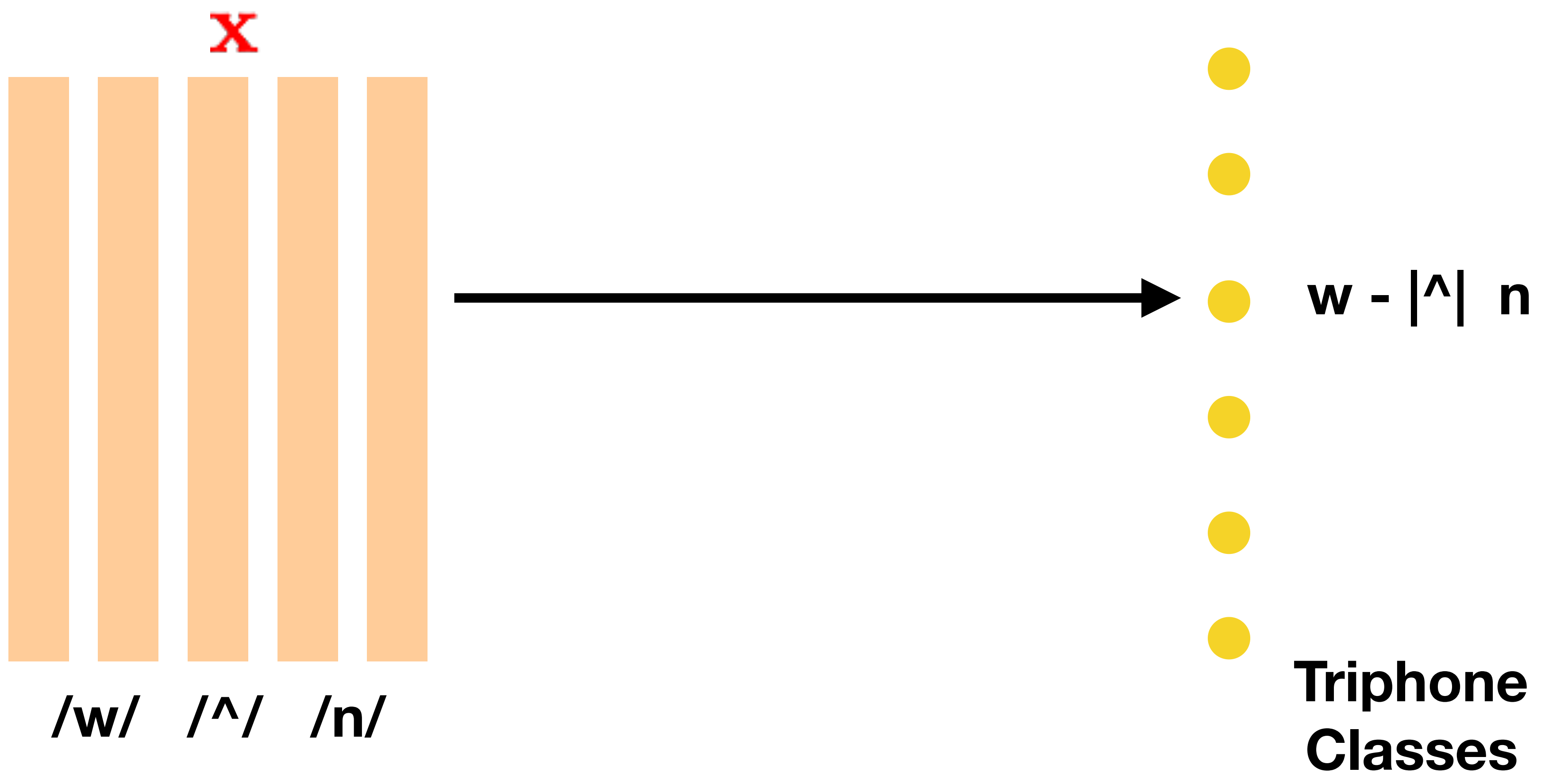
Mel Frequency Cepstral Coefficients

- MFCC processing repeated for every short-term frame yielding a sequence of features. Typically 25ms frames with 10ms hop in time.



Speech Recognition

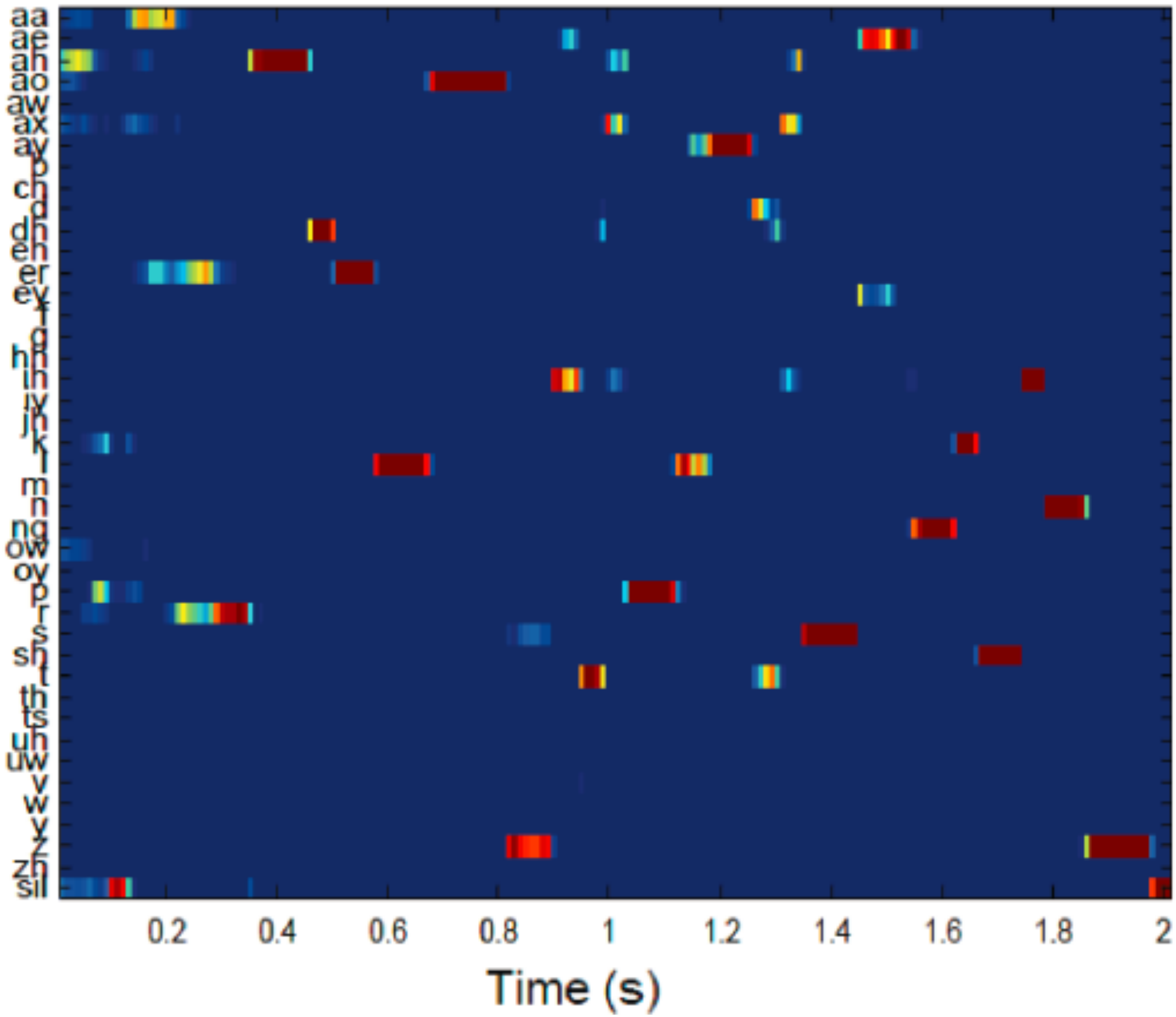
- Map the features to phone class. Using phone labelled data.

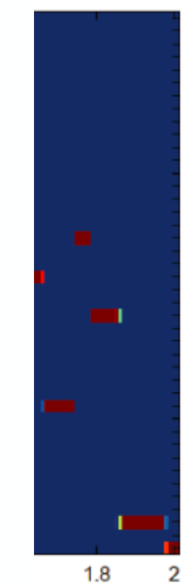


- Classical machine learning - train a classifier on speech training data that maps to the target phoneme class.

Back to Speech Recognition

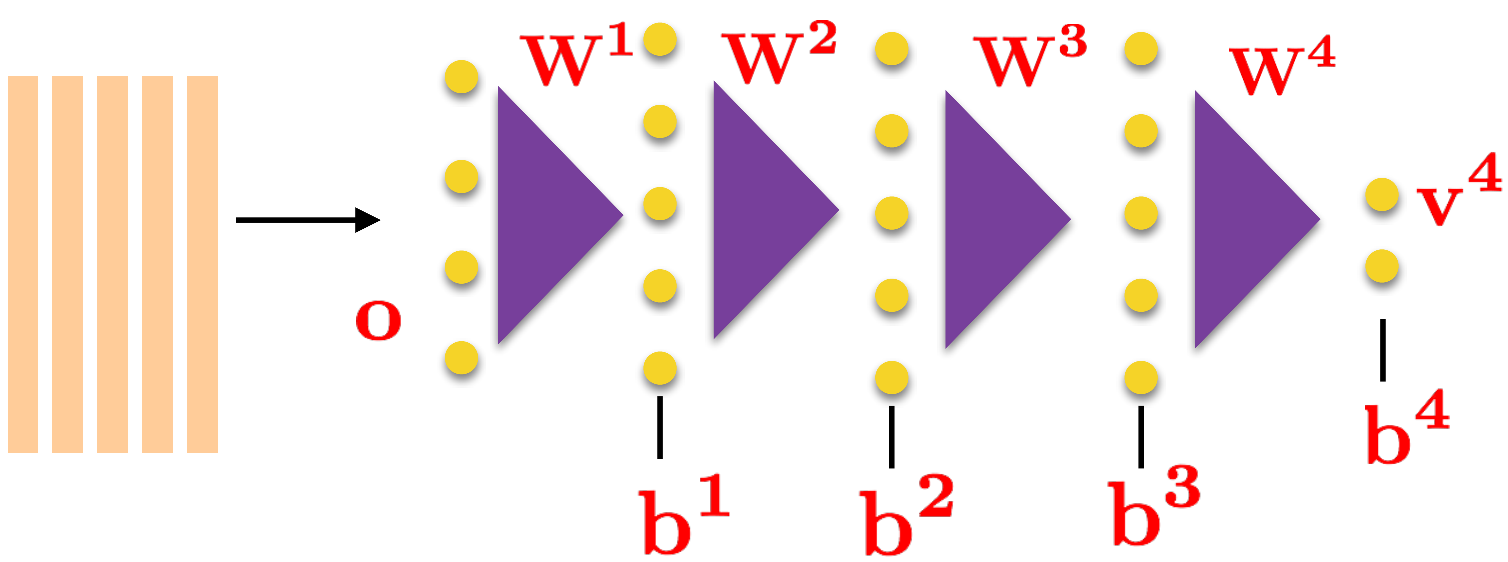
Ma

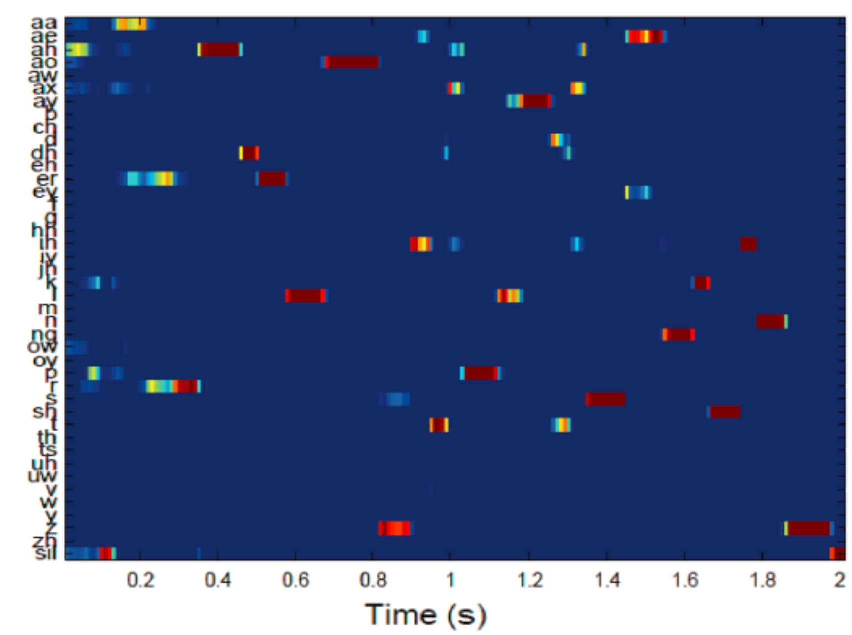




Back to Speech Recognition

Mapping Speech Features to Phonemes to words



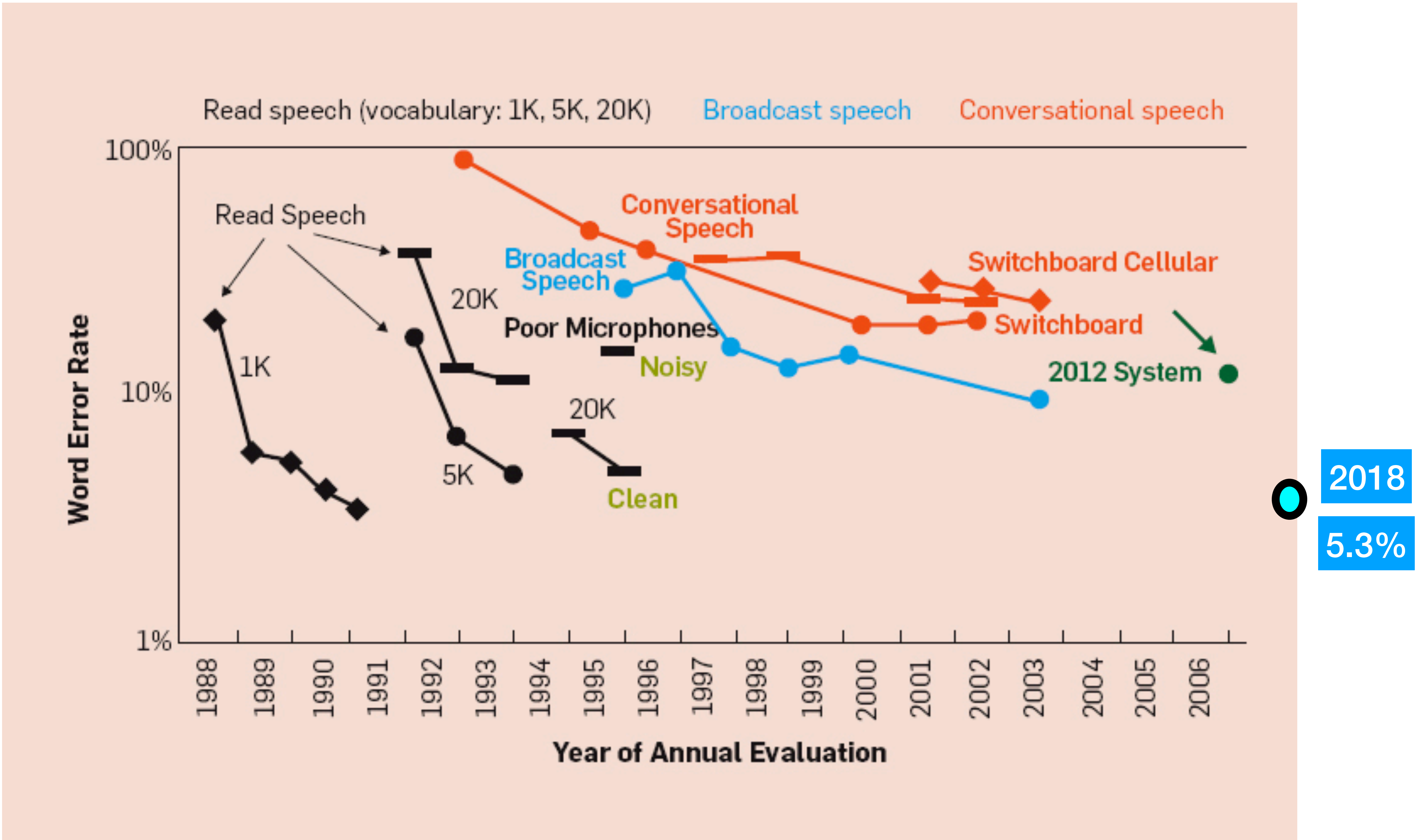


Language Model

[Dictionary of Words
Pronunciation Model
Word Syntax]

Decoded Text

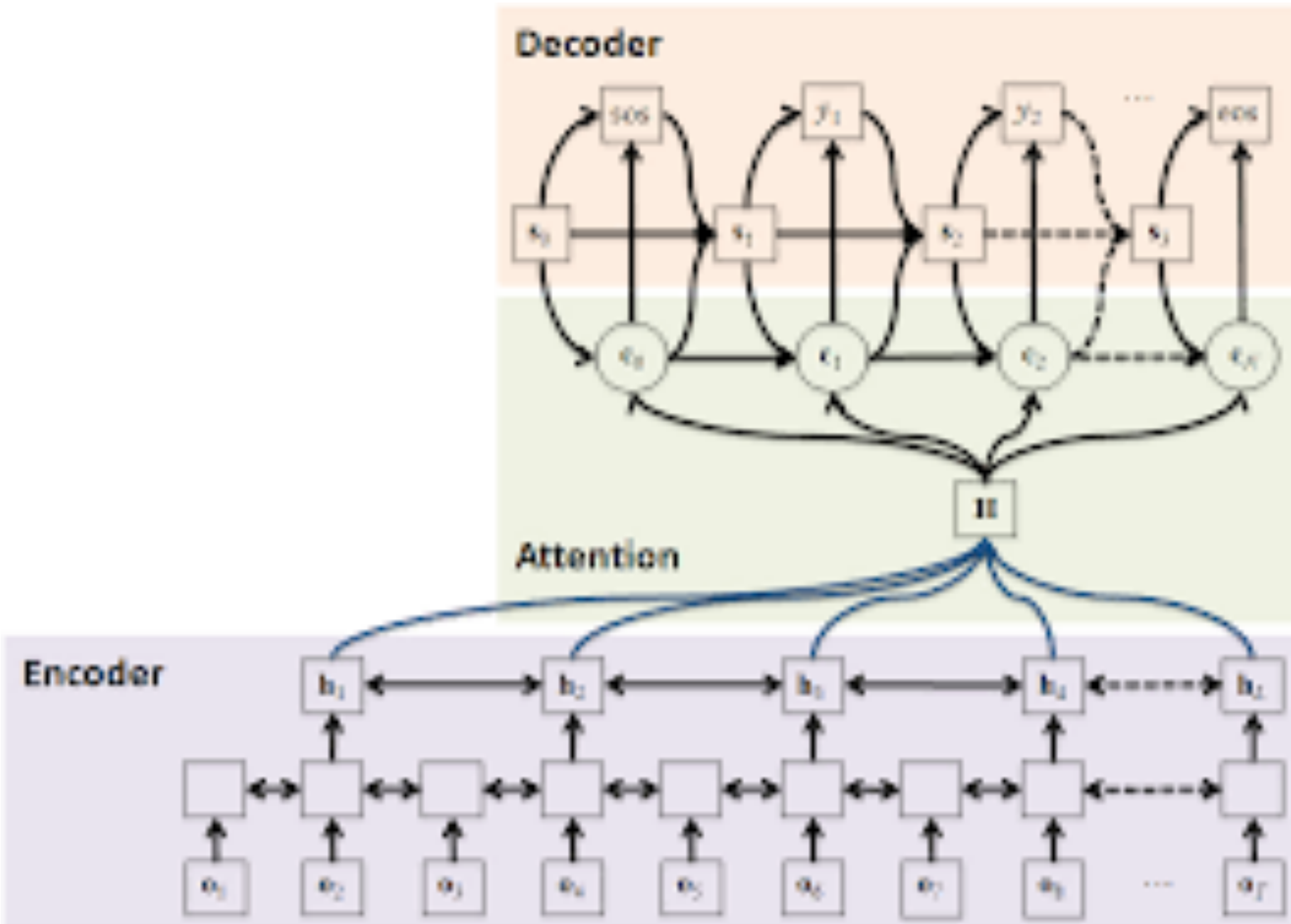
State of Progress



Claims of human parity using BLSTM based Models !!!

Moving to End-to-End

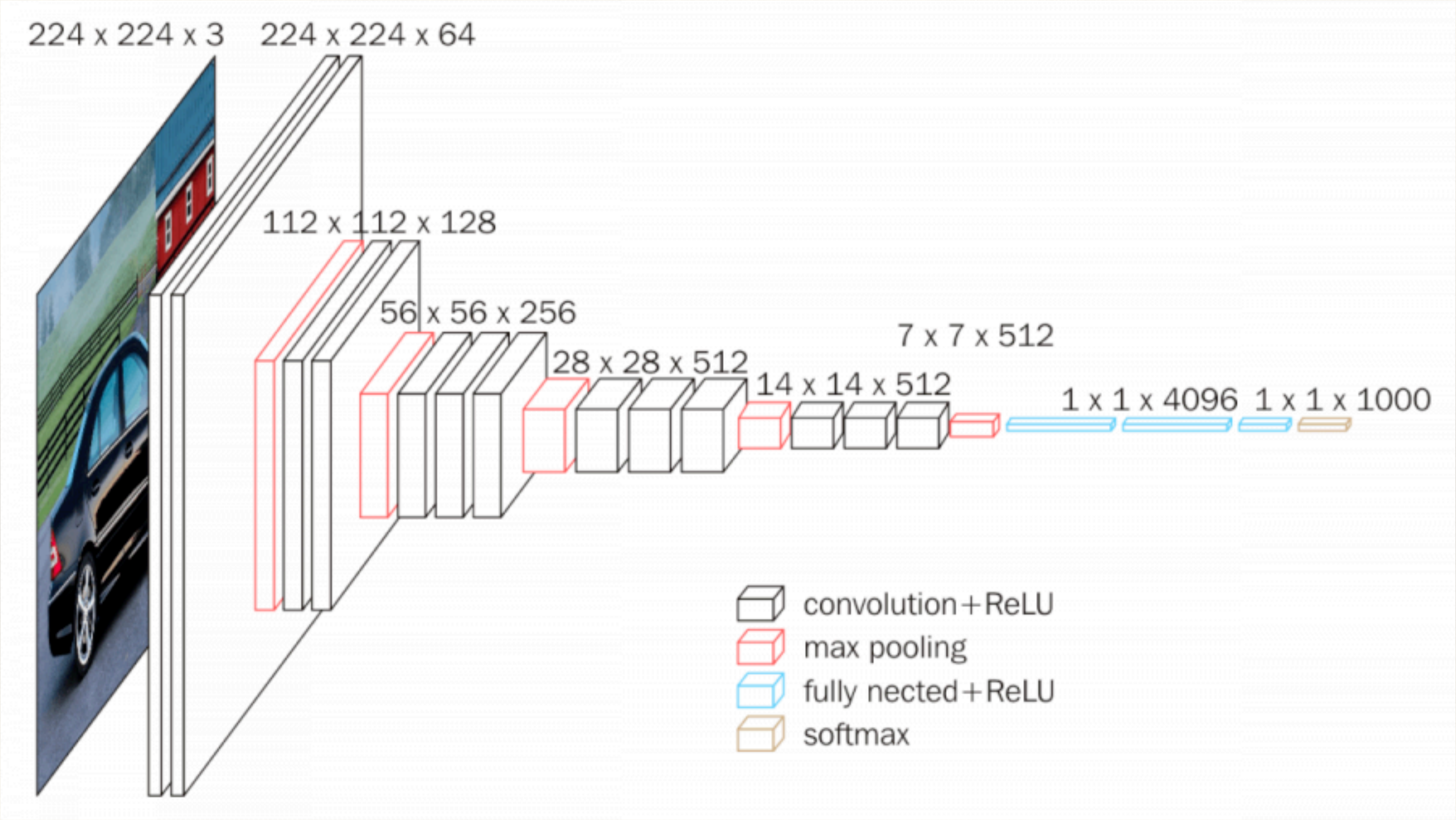
Text Output



Audio Features

Image Processing

Visual Graphics Group Network



ImageNet Task

1000 images in each of 1000 categories. In all, there are roughly 1.2 million training images, 50,000 validation images, and 150,000 testing images. ImageNet consists of variable-resolution images. Therefore, the images have been down-sampled to a fixed resolution of 224×224.

Method	top-1 val. error (%)	top-5 val. error (%)	top-5 test error (%)
VGG (2 nets, multi-crop & dense eval.)	**23.7**	**6.8**	**6.8**
VGG (1 net, multi-crop & dense eval.)	24.4	7.1	7.0
VGG (ILSVRC submission, 7 nets, dense eval.)	24.7	7.5	7.3
GoogLeNet (Szegedy et al., 2014) (1 net)	-	7.9	
GoogLeNet (Szegedy et al., 2014) (7 nets)	-	**6.7**	
MSRA (He et al., 2014) (11 nets)	-	-	8.1
MSRA (He et al., 2014) (1 net)	27.9	9.1	9.1
Clarifai (Russakovsky et al., 2014) (multiple nets)	-	-	11.7
Clarifai (Russakovsky et al., 2014) (1 net)	-	-	12.5
Zeiler & Fergus (Zeiler & Fergus, 2013) (6 nets)	36.0	14.7	14.8
Zeiler & Fergus (Zeiler & Fergus, 2013) (1 net)	37.5	16.0	16.1
OverFeat (Sermanet et al., 2014) (7 nets)	34.0	13.2	13.6
OverFeat (Sermanet et al., 2014) (1 net)	35.7	14.2	-
Krizhevsky et al. (Krizhevsky et al., 2012) (5 nets)	38.1	16.4	16.4
Krizhevsky et al. (Krizhevsky et al., 2012) (1 net)	40.7	18.2	-

Can we go deeper

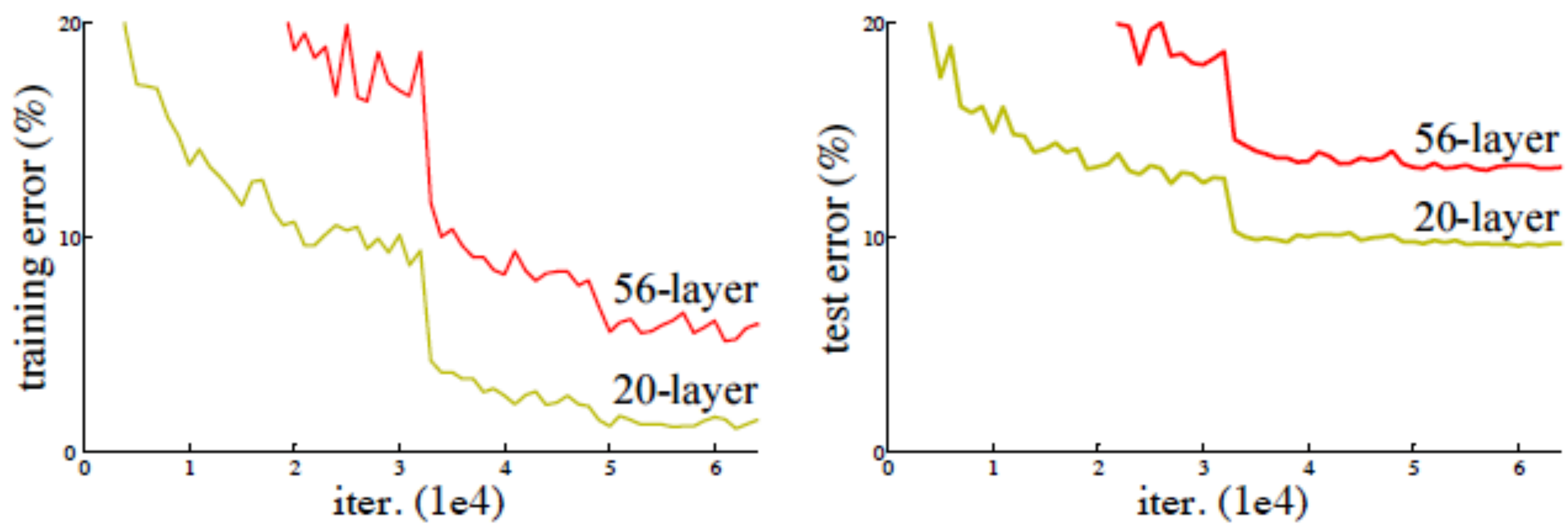


Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer "plain" networks. The deeper network has higher training error, and thus test error. Similar phenomena on ImageNet is presented in Fig. 4.

Deep Residual Learning for Image Recognition

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun
Microsoft Research
{kahe, v-xiangz, v-shren, jiansun}@microsoft.com

Residual Blocks

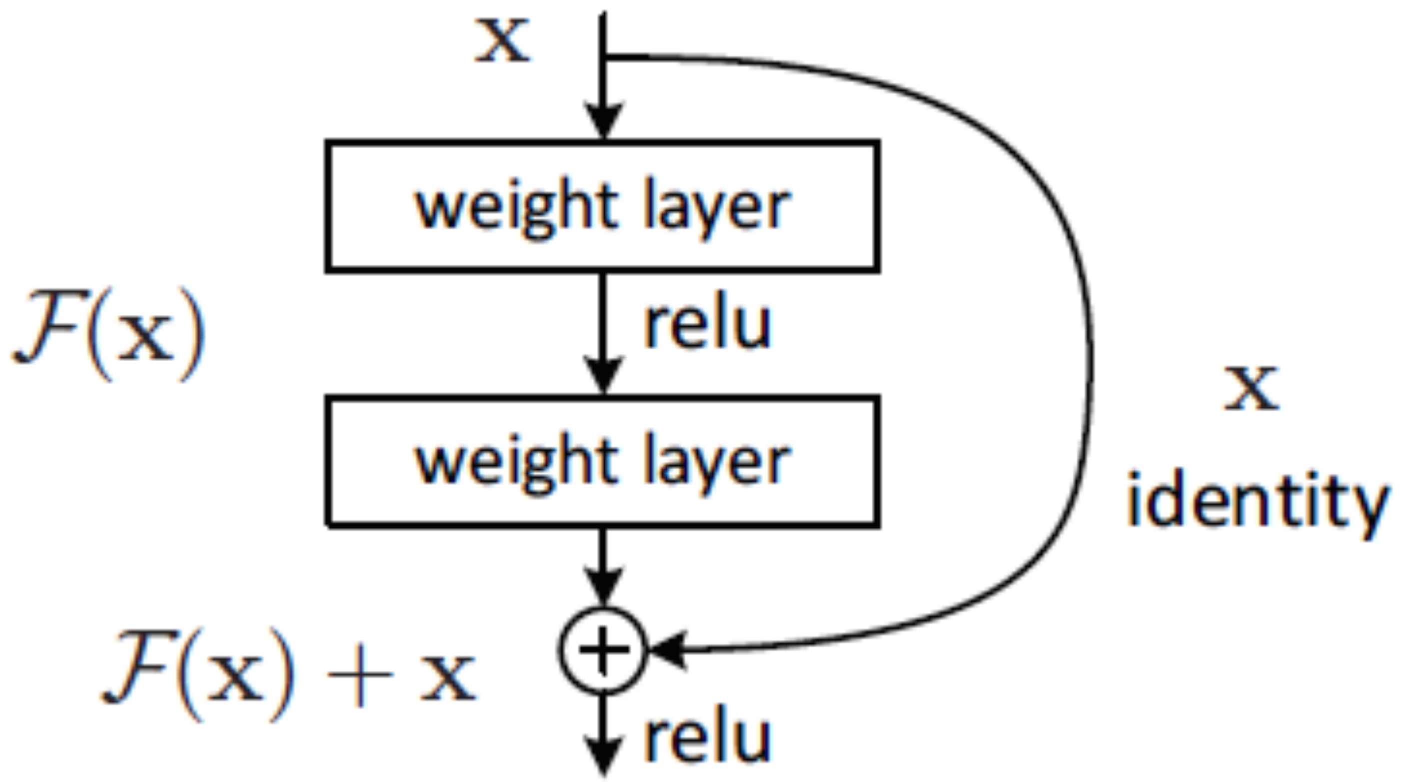
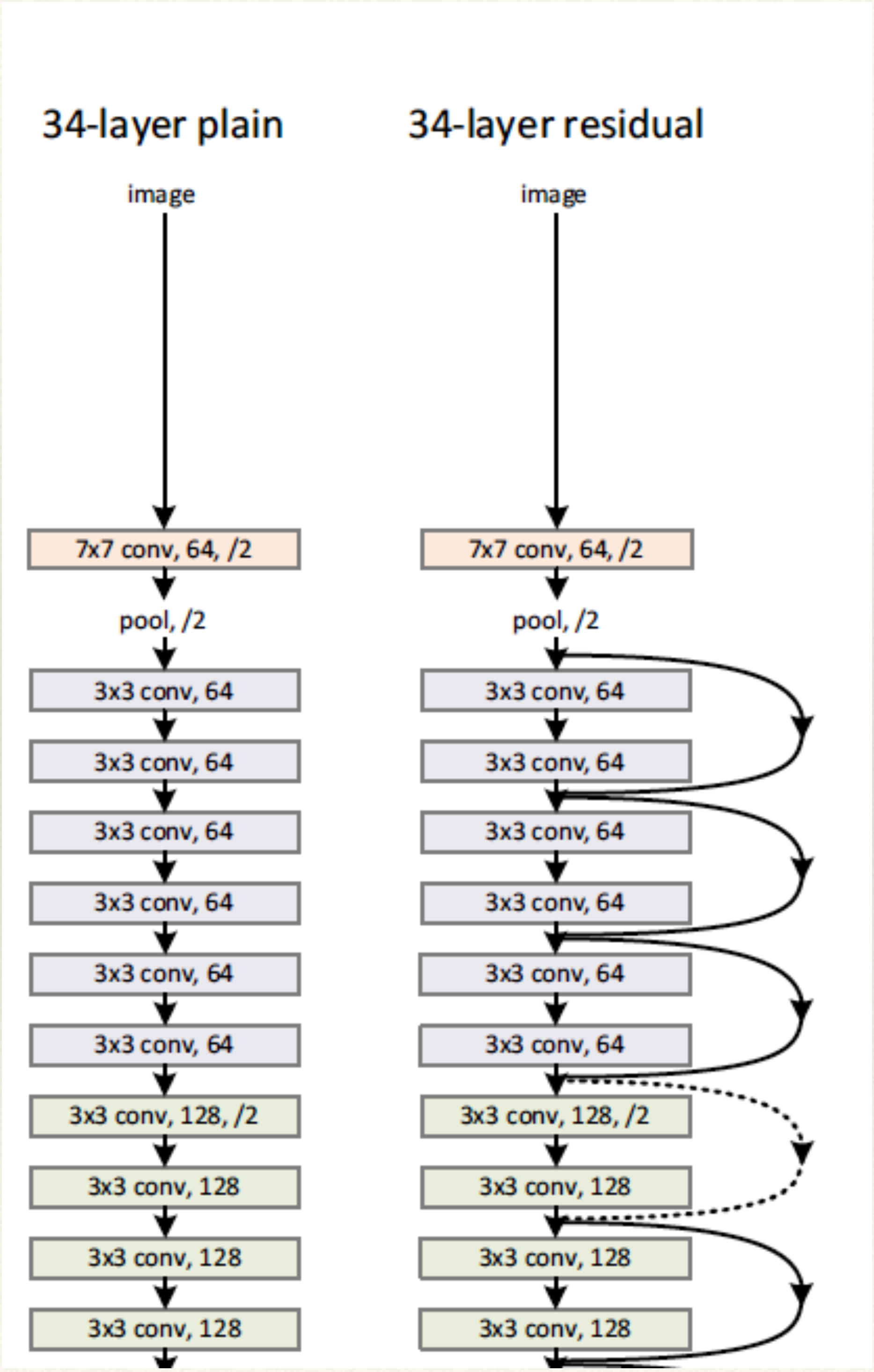


Figure 2. Residual learning: a building block.

Deep Networks with Residual Blocks



Deep Networks with Residual Blocks

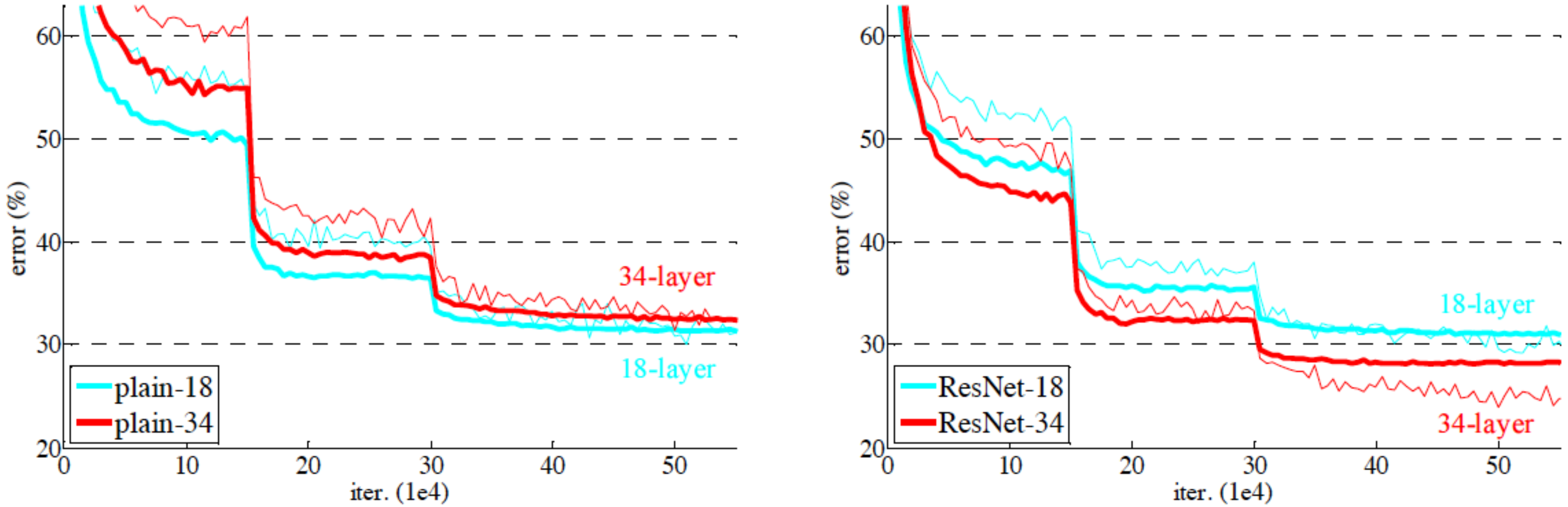


Figure 4. Training on **ImageNet**. Thin curves denote training error, and bold curves denote validation error of the center crops. Left: plain networks of 18 and 34 layers. Right: ResNets of 18 and 34 layers. In this plot, the residual networks have no extra parameter compared to their plain counterparts.

Results with ResNet

	plain	ResNet
18 layers	27.94	27.88
34 layers	28.54	**25.03**

Table 2. Top-1 error (%, 10-crop testing) on ImageNet validation. Here the ResNets have no extra parameter compared to their plain counterparts. Fig. 4 shows the training procedures.

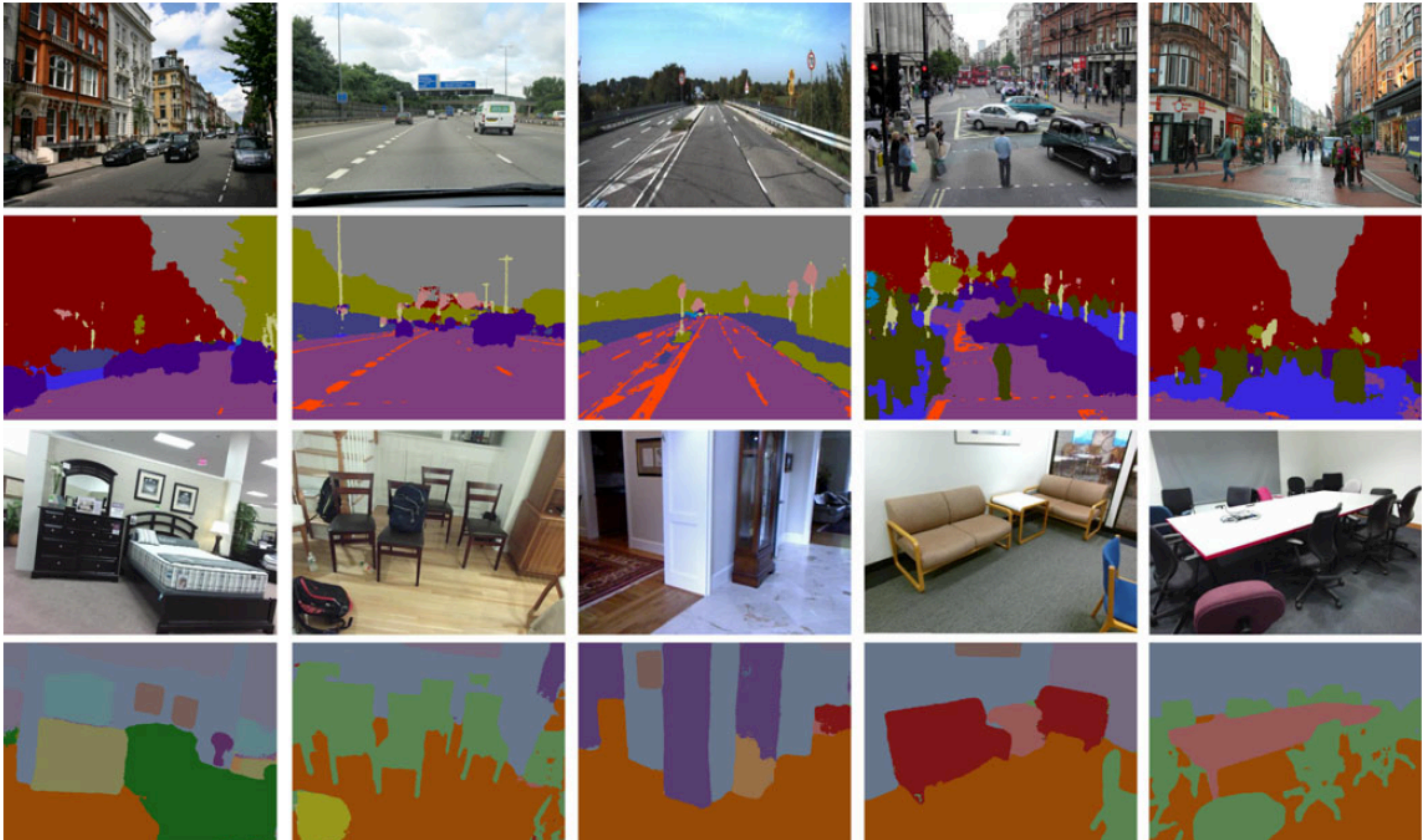
[illegible]	[illegible]	[illegible]
ResNet-50	22.85	6.71
ResNet-101	21.75	6.05
ResNet-152	**21.43**	**5.71**

Image Segmentation

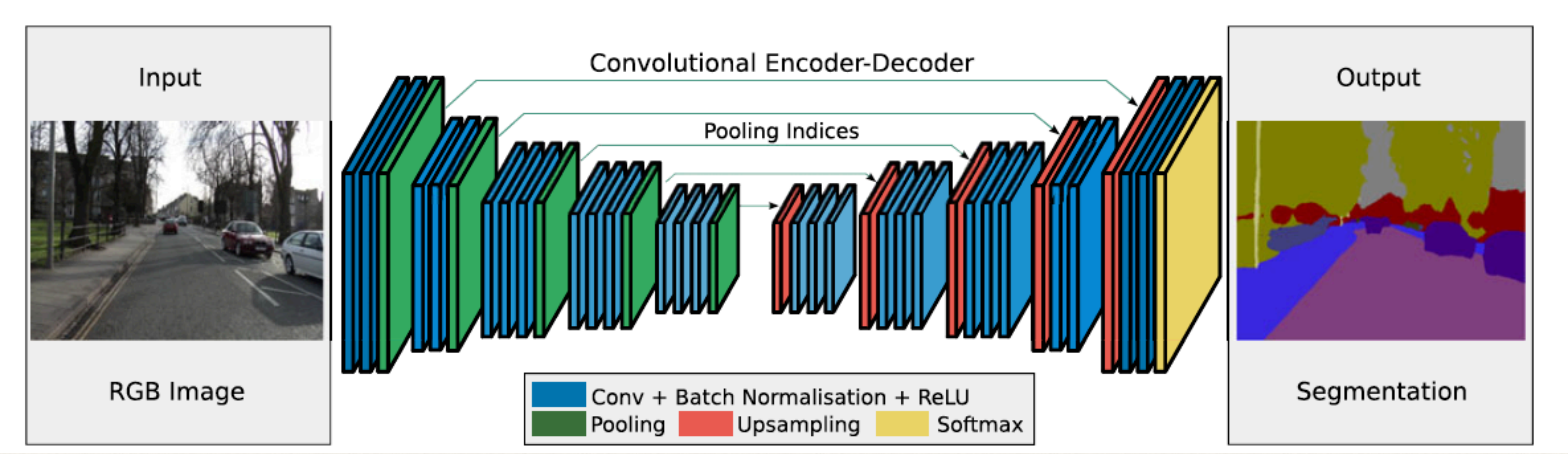
SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation

Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, *Senior Member, IEEE*

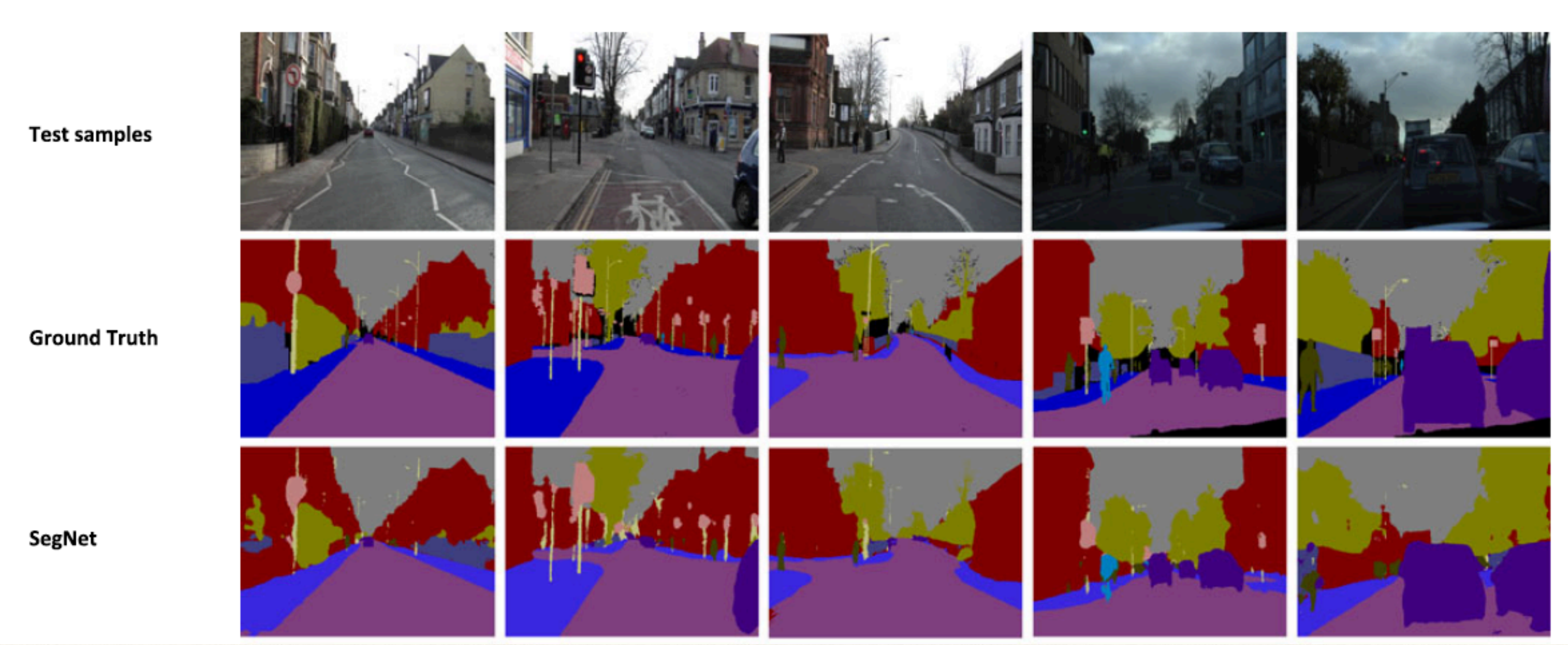
The Problem of Segmentation



SegNet Architecture

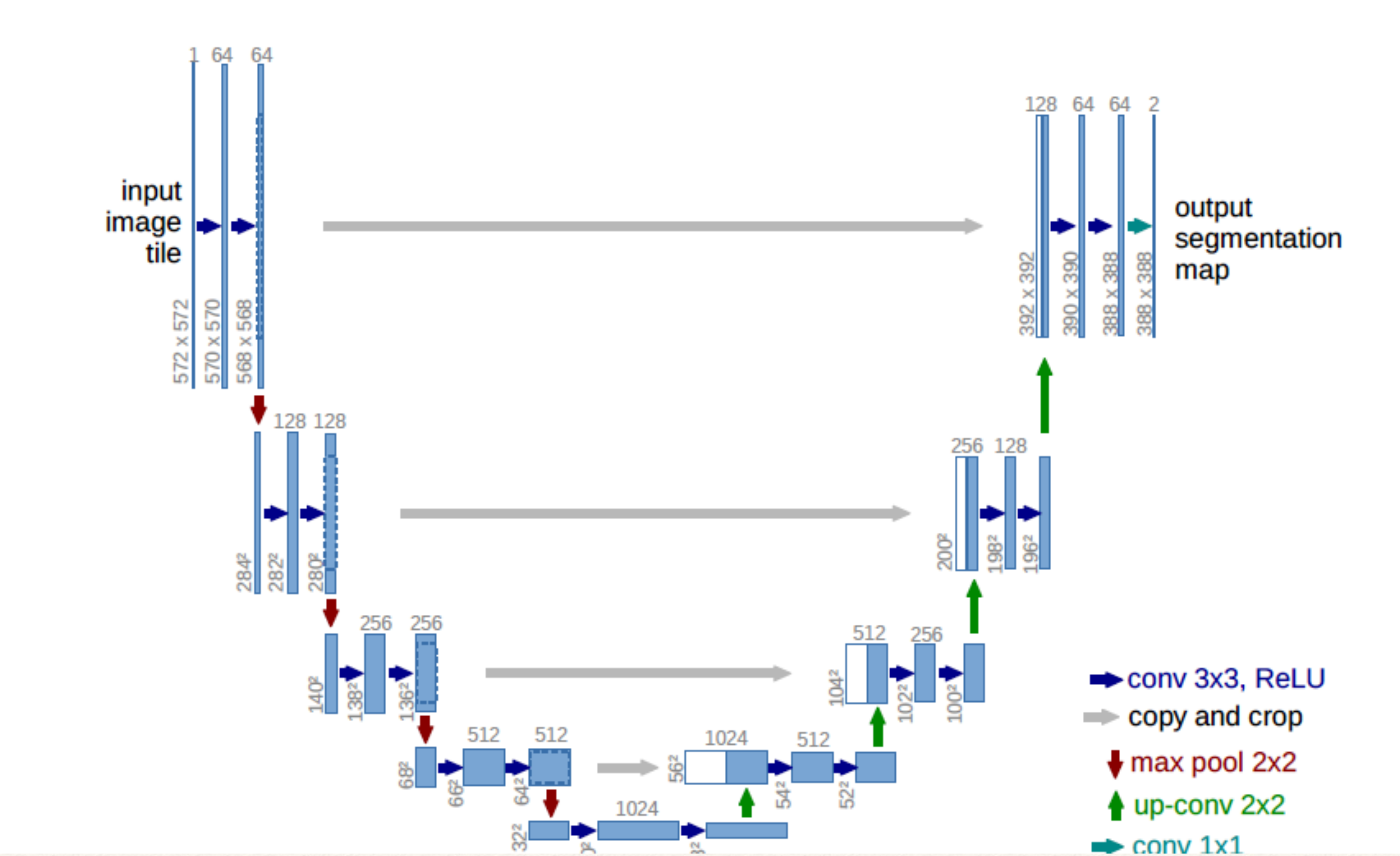


Results from Segnet



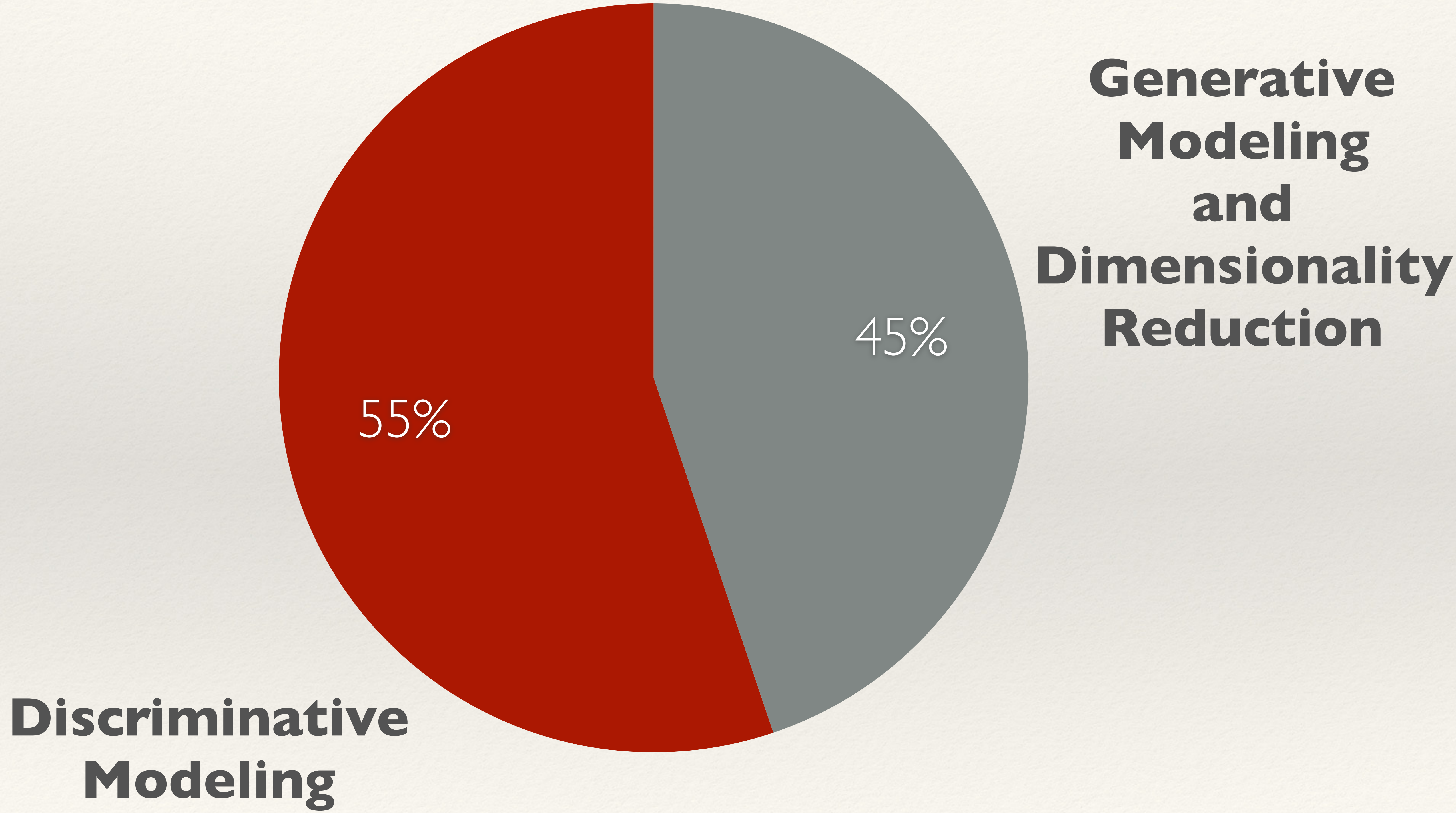
U-Net: Convolutional Networks for Biomedical Image Segmentation

U-net

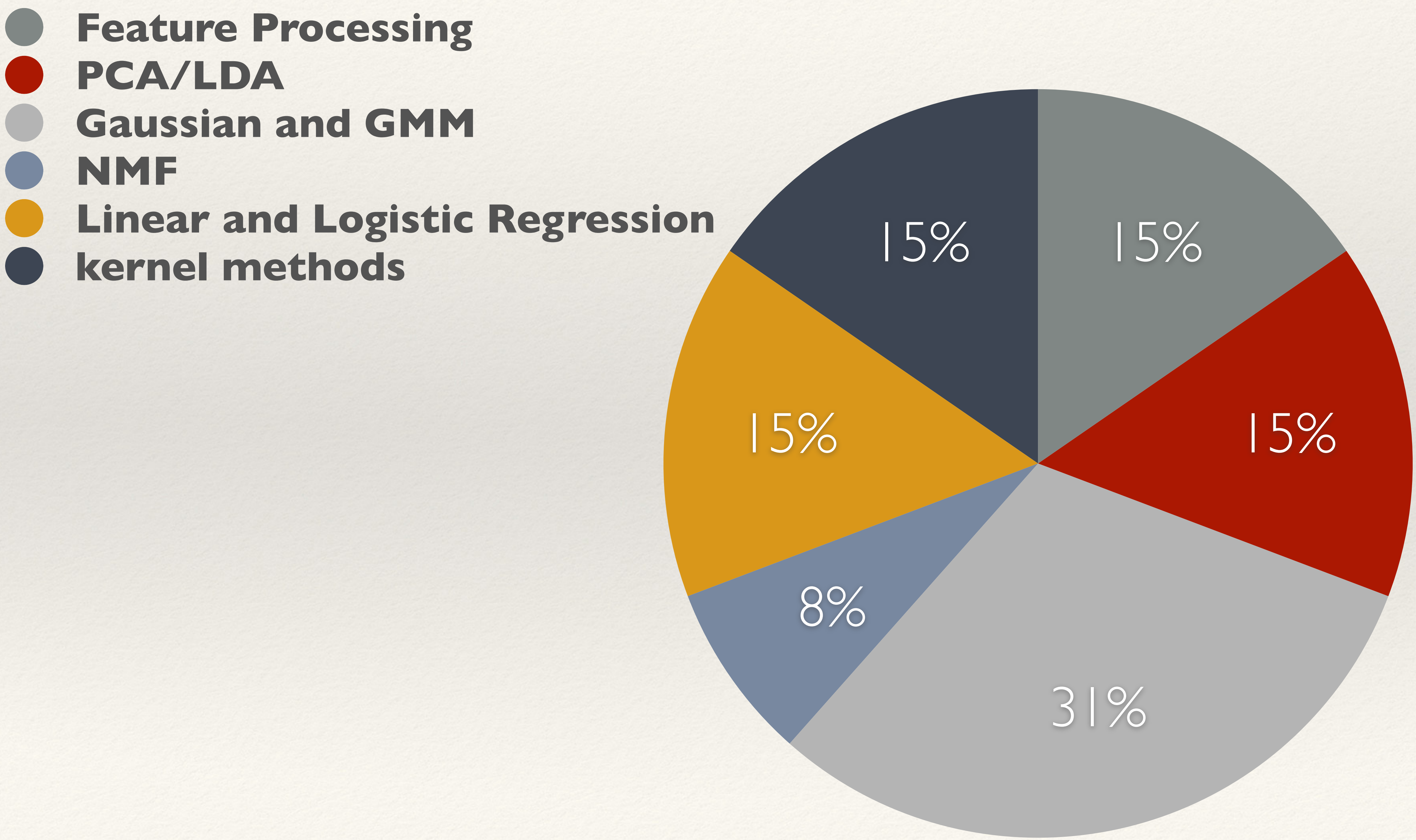

input
image
tile
output
segmentation
map
1
64
64
572 x 572
570 x 570
568 x 568
128
128
284²
282²
280²
256
256
140²
138²
136²
512
512
68²
66²
64²
1024
32²
1024
512
56²
54²
52²
512
256
104²
102²
100²
256
128
200²
198²
196²
128
64
64
2
392 x 392
390 x 390
388 x 388
388 x 388
conv 3x3, ReLU
copy and crop
max pool 2x2
up-conv 2x2
conv 1x1

Summary of the Course

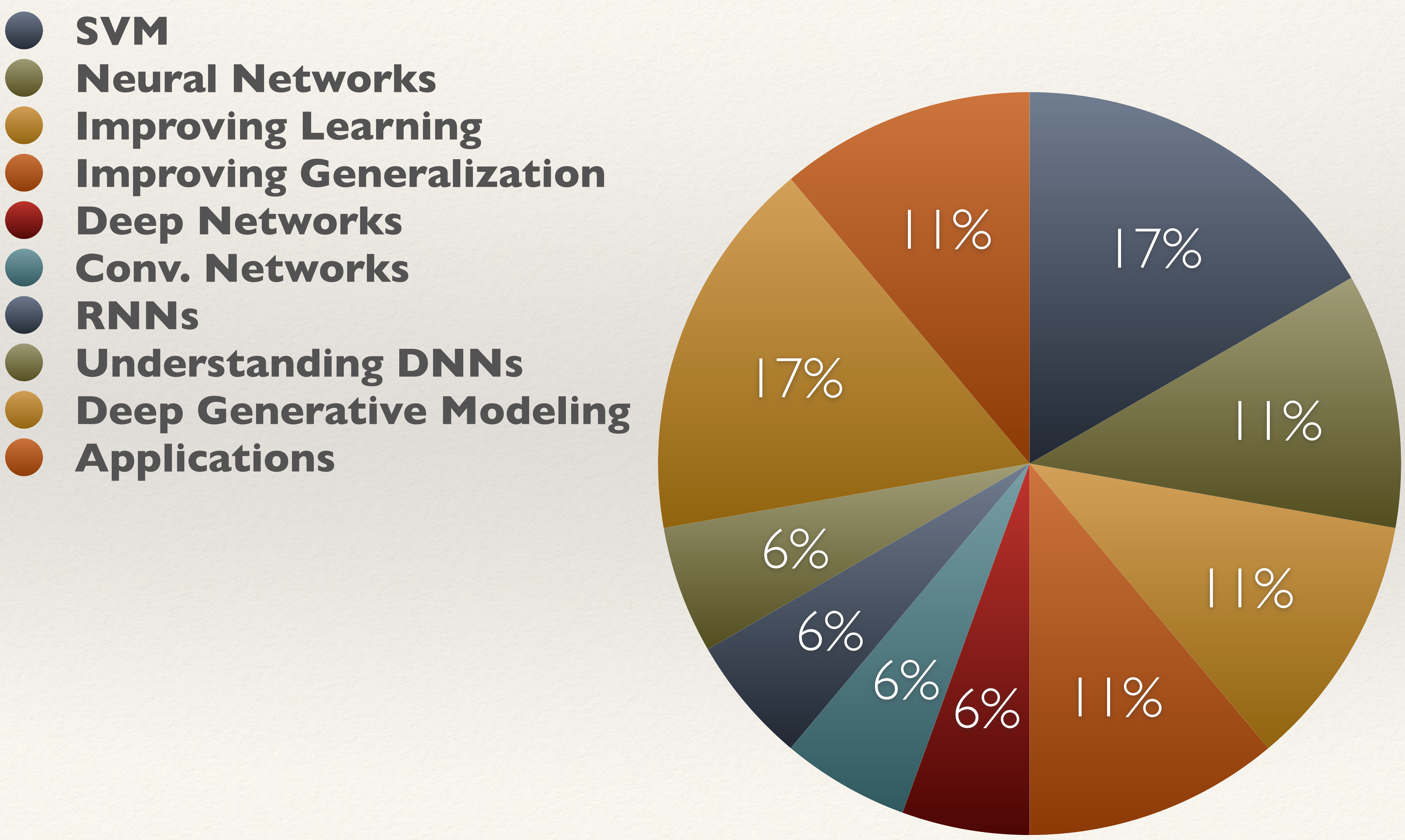
Distribution Pie Chart



Generative Modeling and Dimensionality Reduction



Discriminative Modeling



When we started …

Dates of Various Rituals

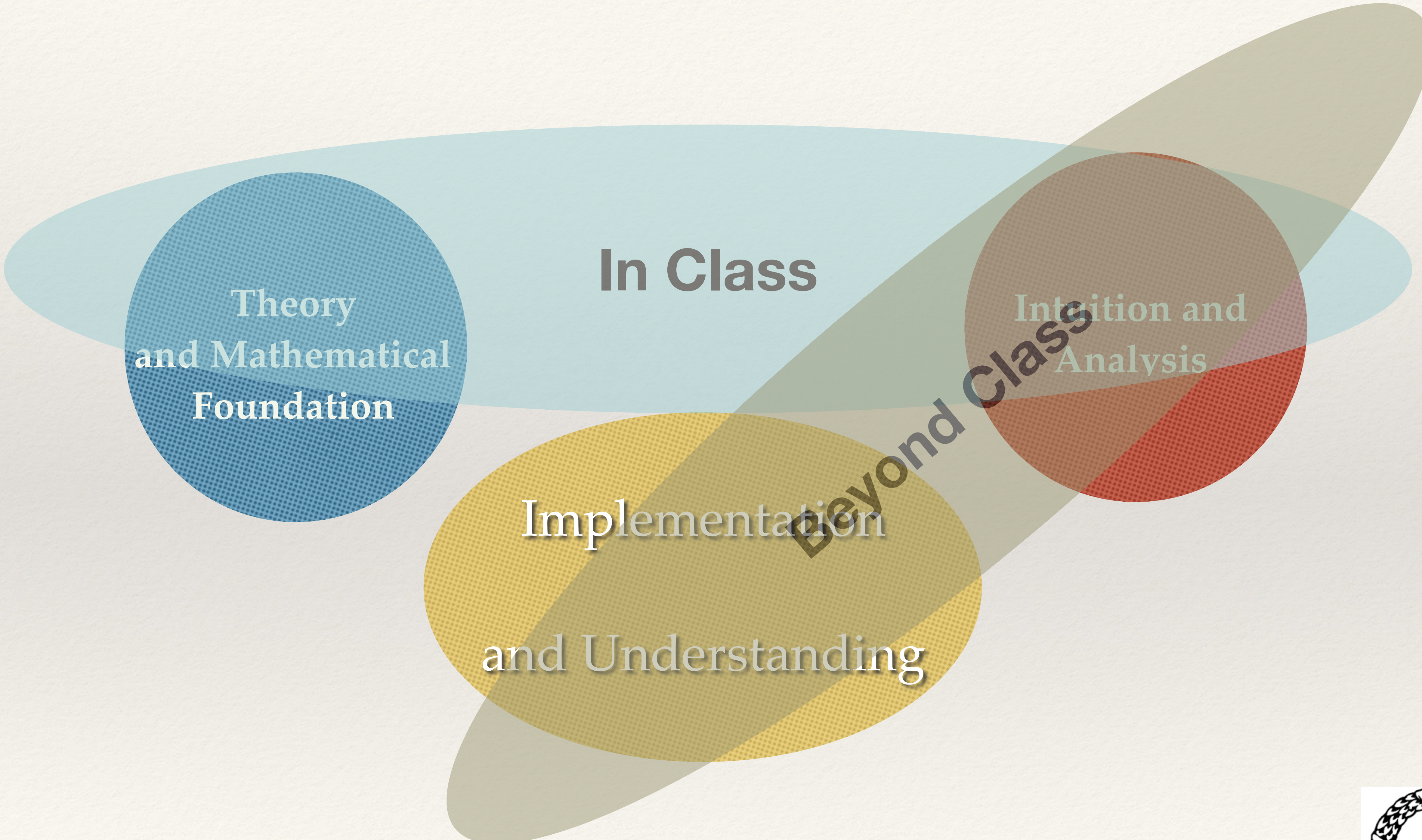
- 5 Assignments spread over 3 months (roughly one assignment every two weeks).
- September 1st week - project topic announcements.
- September 3rd week - 1st Midterm
- September 4th week - project topic and team finalization and proposal submission. [1 and 2 person teams].
- October 1st week - Project Proposal
- October 3rd week - 2nd MidTerm
- November 1st week - Project MidTerm Presentations.
- **December 1st week - Final Exams**
- **December 2nd week - Project Final Presentations.**

Established 1911

INDIAN INSTITUTE OF SCIENCE

भारतीय विज्ञान संस्थान

Content Delivery



Established
1911

INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान