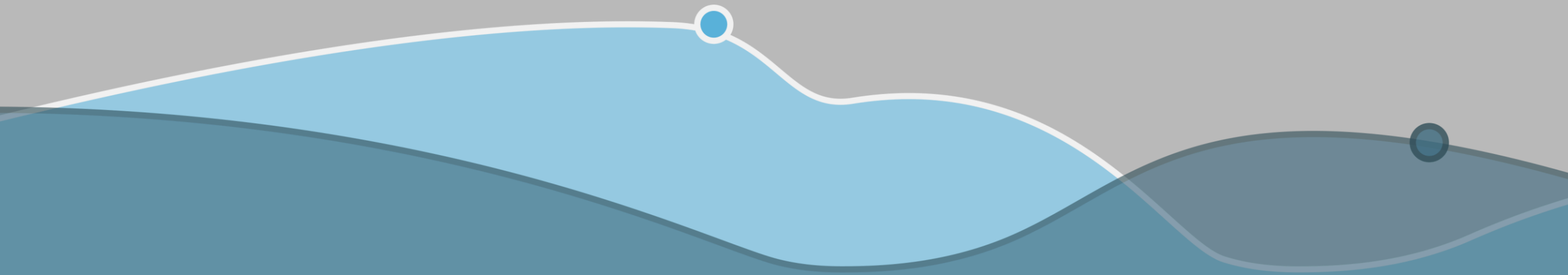
Microsoft

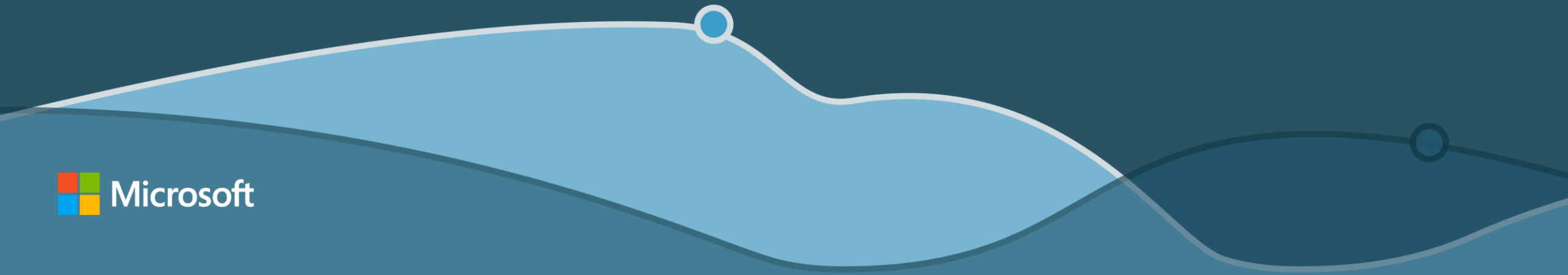
# Cortana Analytics Workshop

## Sept 10 – 11, 2015 • MSCC



# Deep Neural Networks

Misha Bilenko — Principal Researcher
Alexey Kamenev — Senior Research Software Engineer
Ye Xing — Senior Data Scientist



Microsoft

# Why deep neural networks?

## State-of-the-art accuracy for vision and speech

Active research in natural language processing: machine translation, text similarity, etc

## Trained models can be used as featurizers

Pre-trained “competition-grade” models can be reused as general representations for all learners

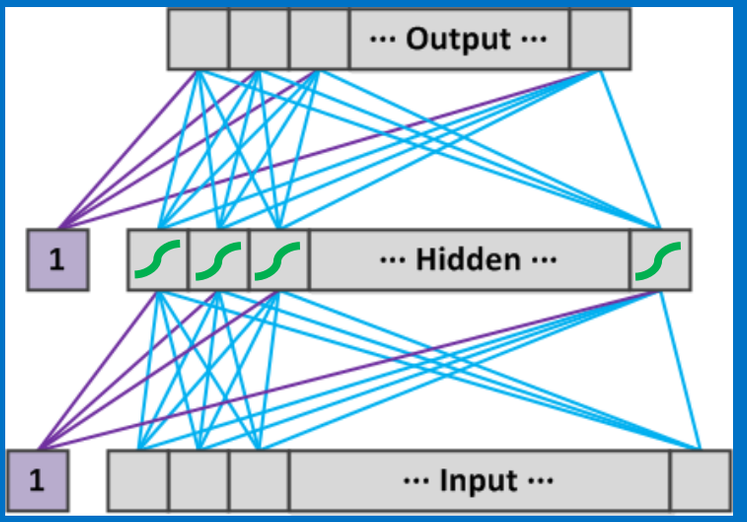
## Neural Nets are a strong learner

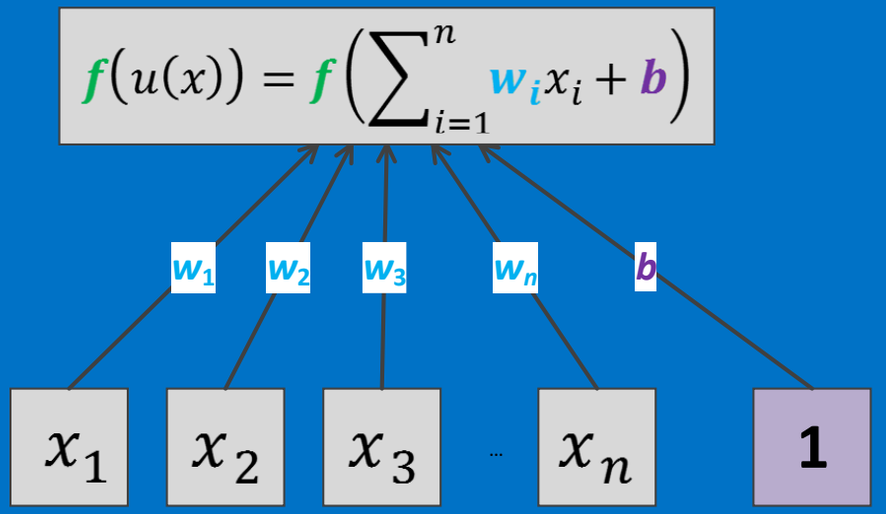
- Workhorse non-linear trainer alongside boosted trees/random forests
- With sufficient training data, additional layers add generalization capacility

# Neural Nets: Gentle Refresher

- "Layers" of transformed linear models
  - Input layer: features
  - Hidden layers: linear models followed by non-linearity
  - Output layer: predictions
- Hidden layer node values
  - **Weighted** sum of source nodes + **bias**
  - Output is transformed via an activation function ***f***
- Formally:

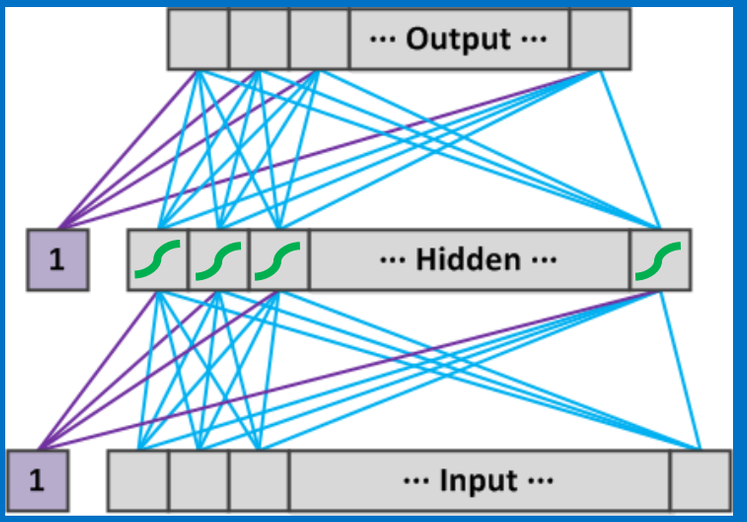
$$\overrightarrow{hidden} = f_1(W_1 \overrightarrow{input} + \overrightarrow{b_1})$$
$$\overrightarrow{output} = f_2(W_2 \overrightarrow{hidden} + \overrightarrow{b_2})$$

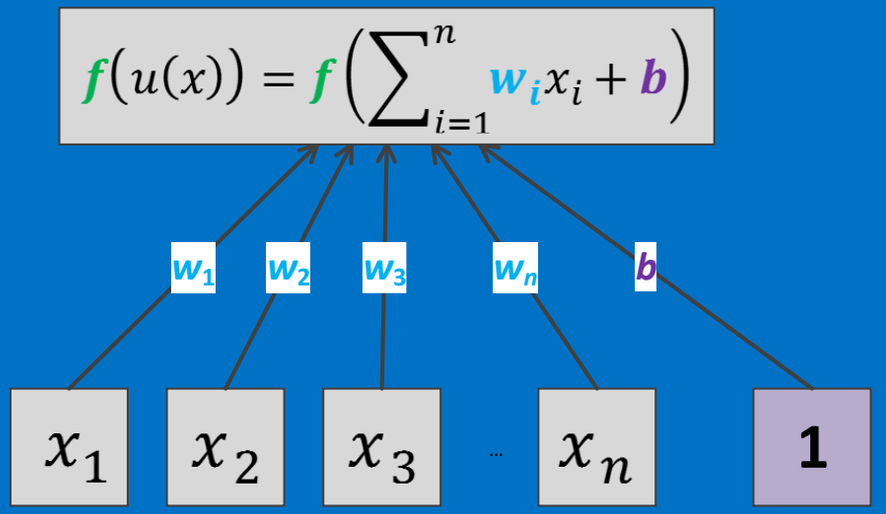


# Training Neural Nets

- Goal: good weights and biases $W$
- Minimize cost function
  - Error on training set $L(W)$
  - Model regularization
- Iteratively update weights in the "direction" of reducing the cost function → reducing error



# Training neural networks

## Many different types of nets

DNN, CNN, RNN, LSTM, DSSM, …

## Q: Can we use just one net for all tasks?

## A: No – NFL theorem [D. Wolpert, W. Macready; 1996, 1997]

No Free Lunch theorem: any two optimization algorithms are equivalent when their performance is averaged across all possible problems

No universal machine learning algorithm that performs best on all tasks ☹

## The tools are important!

Cortana Analytics covers both training and prediction

# Why now?

## Neural nets have been around for decades

Backpropagation in 1970s, Convolutional nets in late 1980s etc

## Two main factors:

## Large, real-world (and free!) datasets

ImageNet, SVHN, MIT Places etc

## Advances in hardware

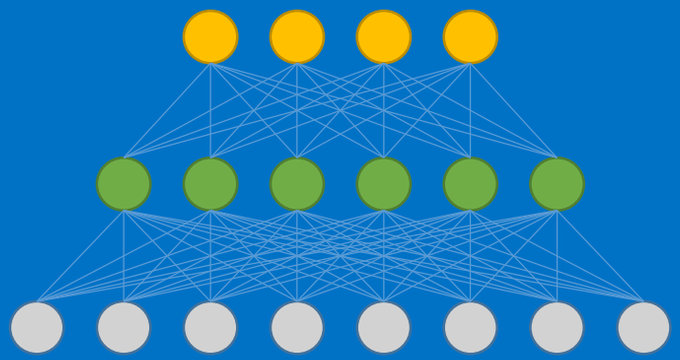
# Net#: Topology and Model Language

- Flexible language for neural network topology
  - Used for training and/or models
  - Simple, readable, general, open
  - Not too image/speech/text specific but a good fit for any of these tasks
  - Syntax is similar to C# (lambdas, consts etc.)
- Core connectivity patterns (“connection bundles”)
  - Full, Filtered (sparse), Convolutional, Pooling, Response normalization
  - Weight sharing among bundles (for RNNs)
  - Expandable
- Core activation functions
  - Sigmoid, tanh, ReLU, Soft ReLU, abs, sqr, expandable
- Model conversion from other frameworks (e.g. Caffe)

# Net#: basic topologies

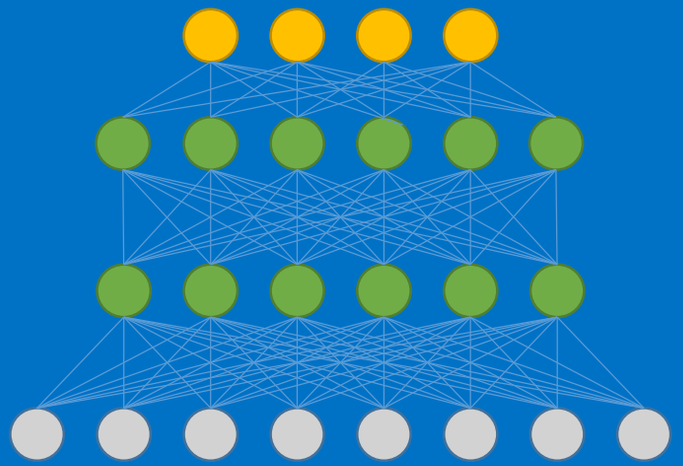
1 hidden layer, fully connected net:

```
input Picture [28,28];
hidden H [200] from Picture all;
output Result [10] softmax from H all;
```



2 hidden layers, fully connected net:

```
input Picture [28,28];
hidden H1 [200] tanh from Picture all;
hidden H2 [200] tanh from H1 all;
output Result [10] softmax from H2 all;
```



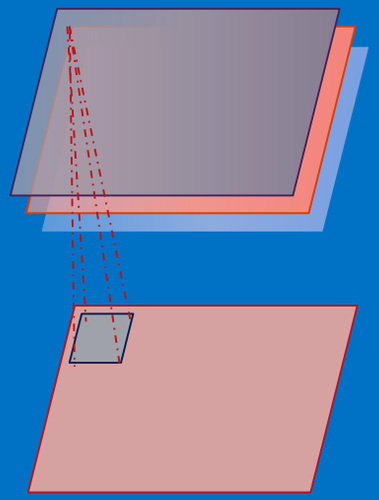
# Net# - advanced topologies I

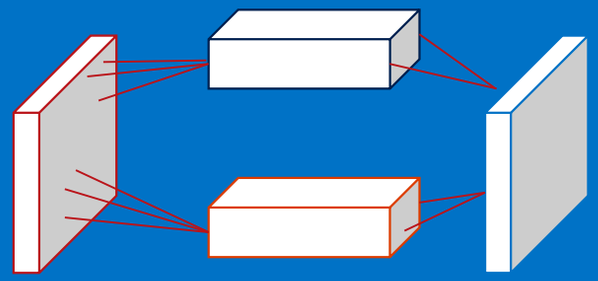
Convolutional layer:

```
const { T = true; F = false; }

input data [3 * 224 * 224];

hidden conv1 [64, 224, 224] rlinear
    from data convolve
    {
        InputShape  = [3, 224, 224];
        KernelShape = [3, 3, 3];
        Padding     = [F, T, T];
        MapCount    = 64;
    }
```





Max pooling layer:

```
hidden pool1 [96, 28, 28]
    from rnorm1 max pool
    {
        InputShape  = [96, 56, 56];
        // Overlapped pooling
        KernelShape = [1, 3, 3];
        Stride      = [1, 2, 2];
        Padding     = [F, T, T];
    }
```

Multiple bundles:

```
// Two input layers.
input Picture [28, 28];
input Metadata [100];

hidden H1 [200] from Picture all;

// H2 connected both to H1 and Metadata layers using different connection types
hidden H2 [200] {
  from H1 all;                              // This is a fully-connected bundle.
  from Metadata where (s, d) => s < 50; // This is a filtered (sparse) bundle.
}

output Result [10] softmax from H2 all;
```

# Representation learning

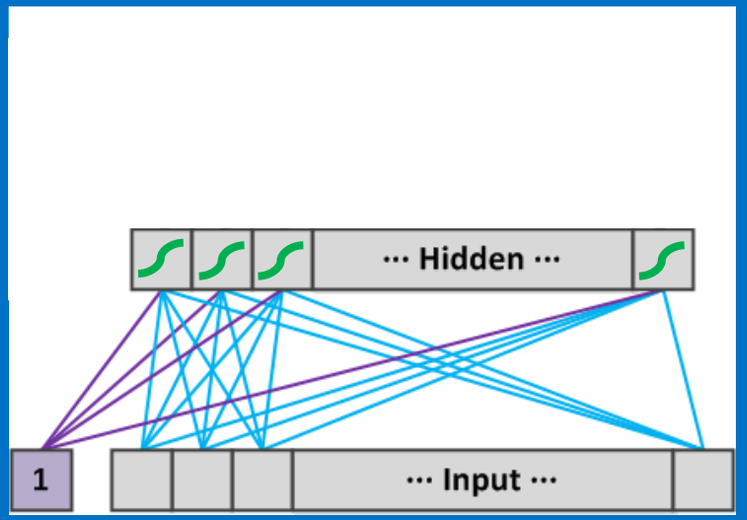
## Deep neural nets learn features

Many other ML algorithms require manual feature engineering

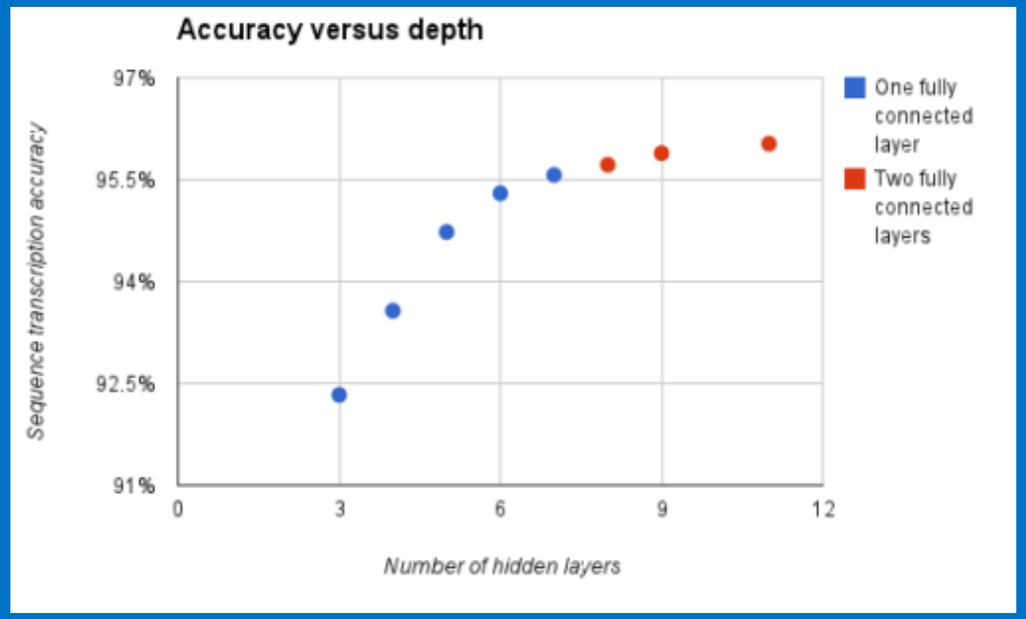
## Features are reusable

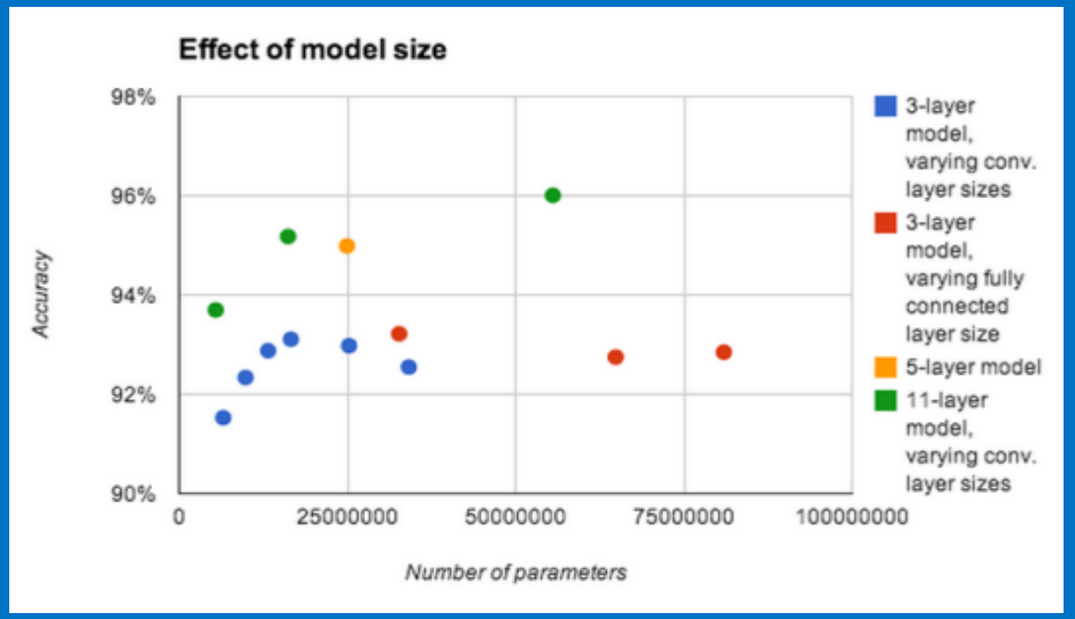
Other applications like dimensionality reduction, similarity search

As inputs to other ML algorithms (e.g. logistic regression)



# Why deep learning?




[Goodfellow et al, 2013]

# How deep?

Depends on the problem, for example, ImageNet: around 20 layers

# Net Topology Demo

Alexey Kamenev

# ImageNet


## Large image dataset:

14M+ images
Images are photographs of different sizes

## Organized into WordNet-based hierarchy

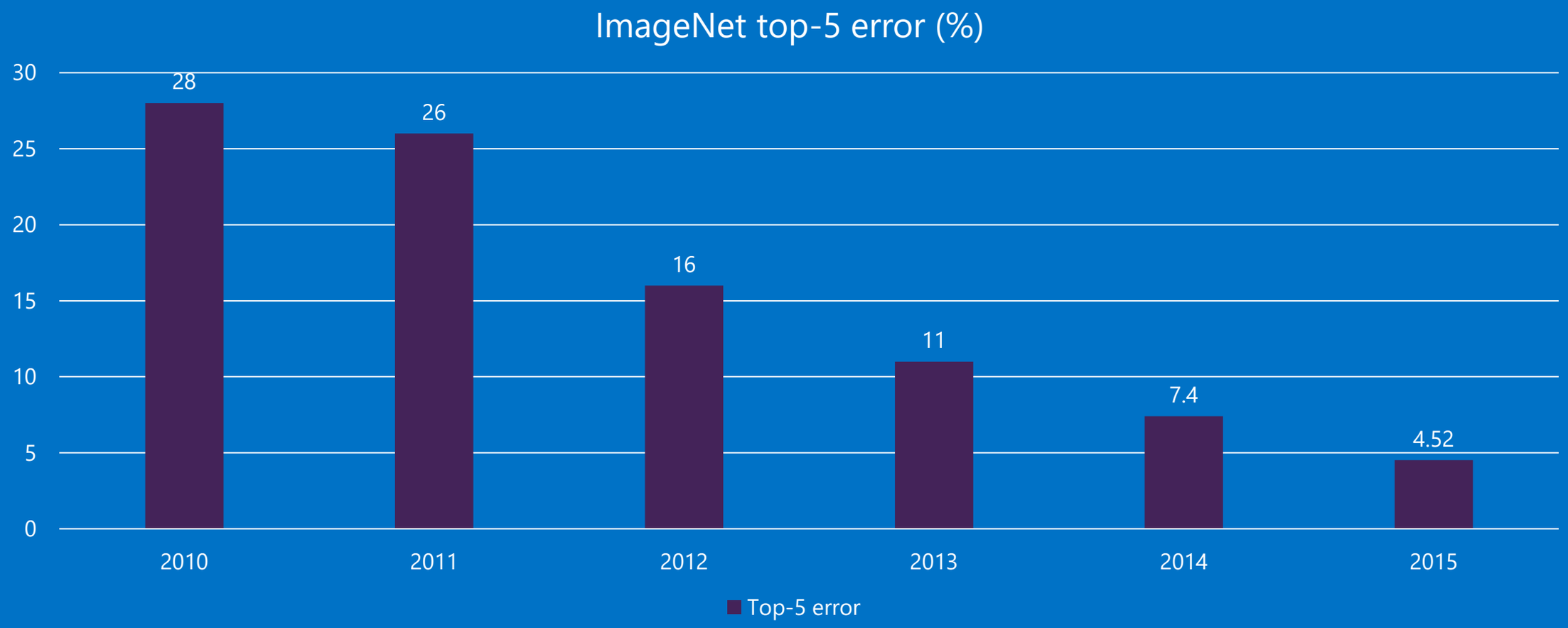
21,841 synonym sets (synsets), such as:

- n02882647: bowling pin, pin
- n00449517: auto racing, car racing
- n10698368: television reporter, television newscaster, TV reporter, TV newsman

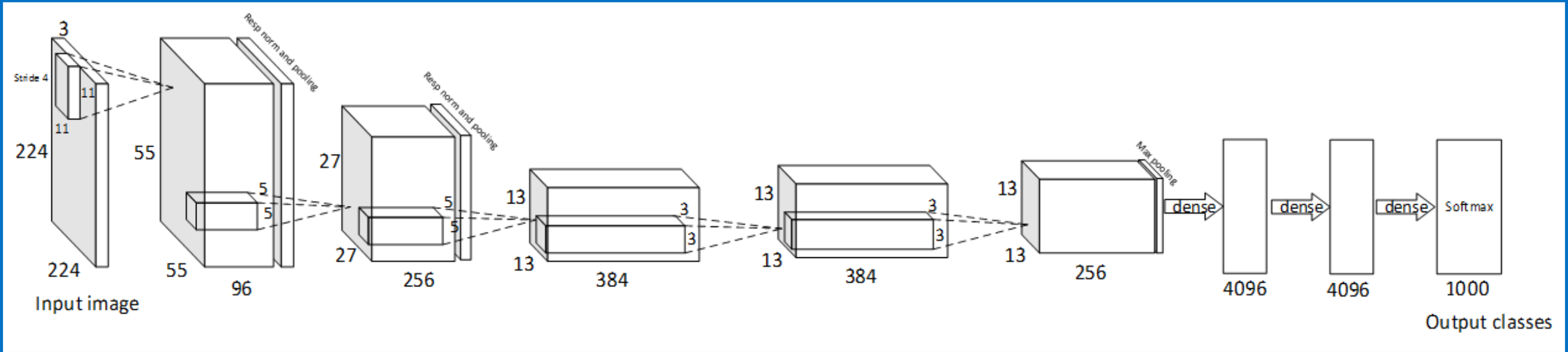
## Smaller 1000 classes dataset is used in competitions

Classes are mutually exclusive
Error is measured as top-5 error:
Take top 5 predictions, if sample's label is within these 5 classes, then count it as correct labeling case

# ImageNet Competition Results
## aka "The Chart you'll see in every DNN talk"



# AlexNet (simplified)

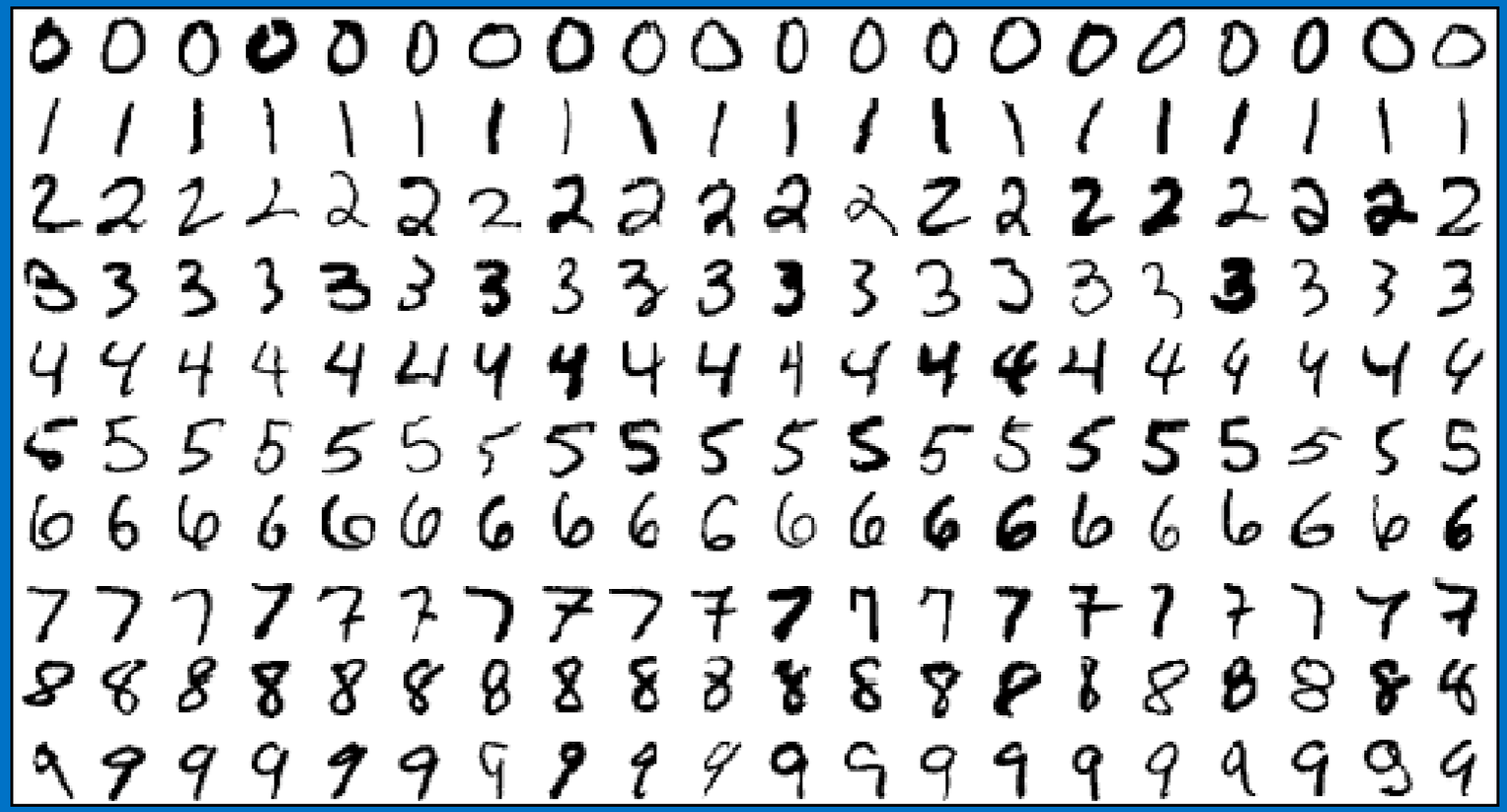



[Krizhevsky, Sutskever,Hinton, 2012]

# Neural Nets in Azure ML Demo

Ye Xing

# MNIST dataset



# Azure Machine Learning Service

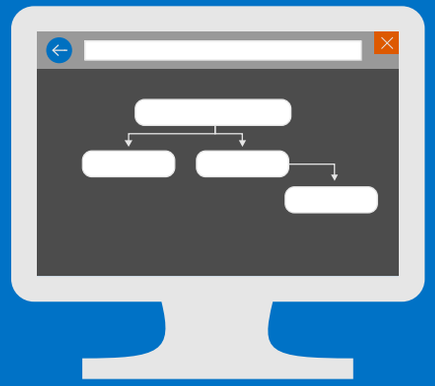
Data -> Predictive model -> Operational web API in minutes

Data

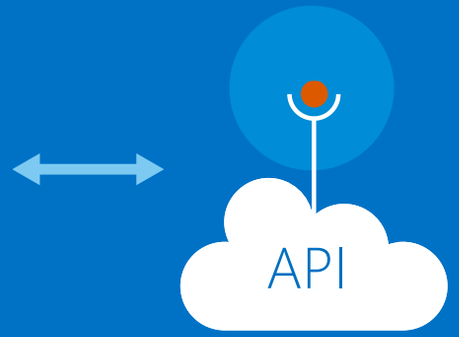
Blobs and Tables

Hadoop (HDInsight)

Relational DB (Azure SQL DB)



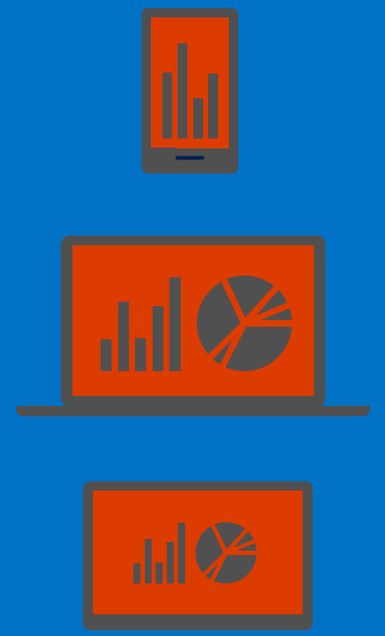
Integrated development environment for Machine Learning




Model is now a web service that is callable

Monetize the API through our marketplace

Clients



# Conclusion

- Deep neural network is state-of-art with applications in image, speech and natural language processing.
    - No manual feature engineering work is needed
    - High accuracy
- Azure ML provides a cloud-based platform for deep neural network with an easy deployment of web service
- Cortana analytics suite provides E2E analytics pipeline.

Microsoft