

Large-Scale Deep Learning With TensorFlow

Jeff Dean
Google Brain team
g.co/brain

In collaboration with **many** other people at Google

We can now store and perform computation on large datasets



We can now store and perform computation on large datasets



But what we really want is not just raw data,
but computer systems that **understand** this data



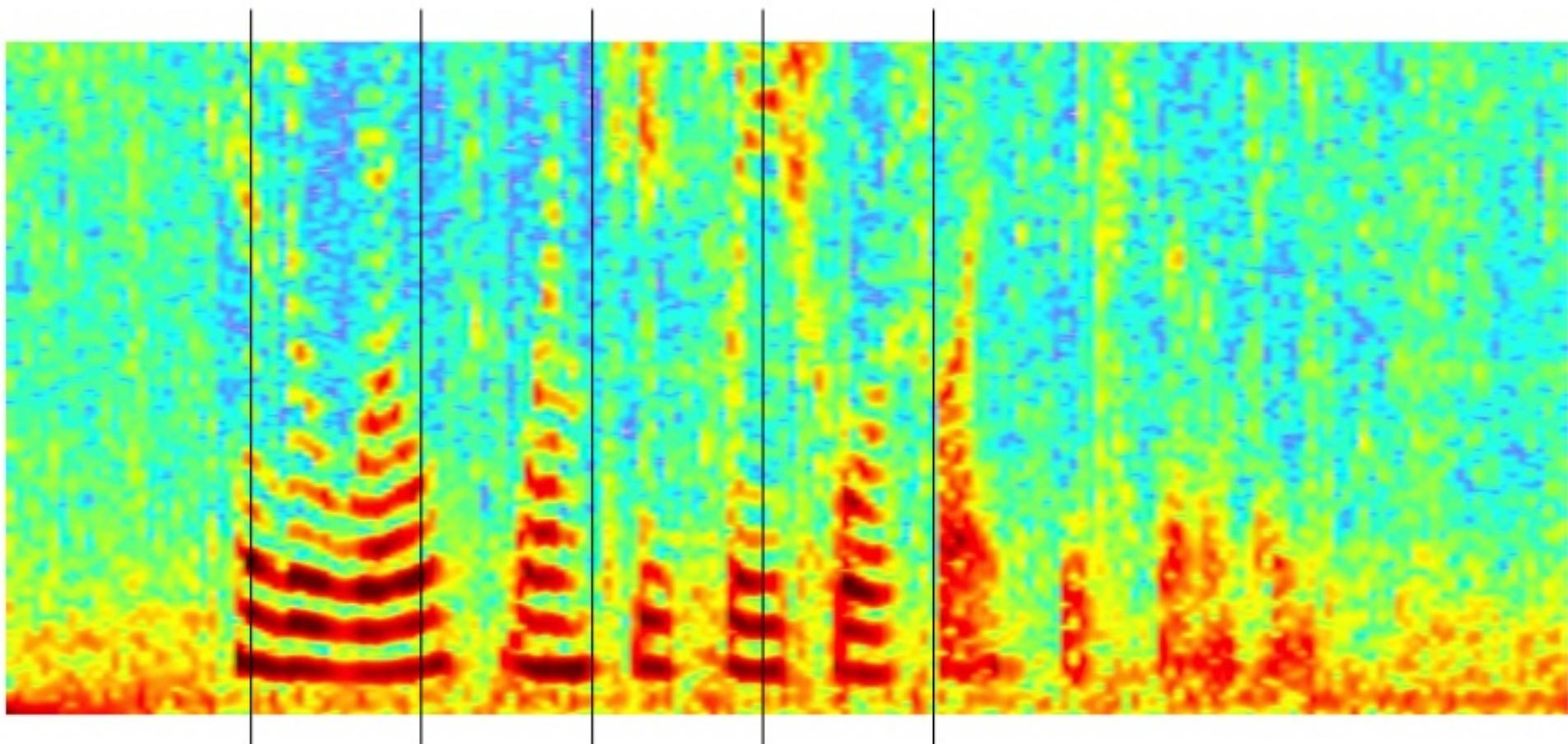
What do I mean by understanding?



What do I mean by understanding?

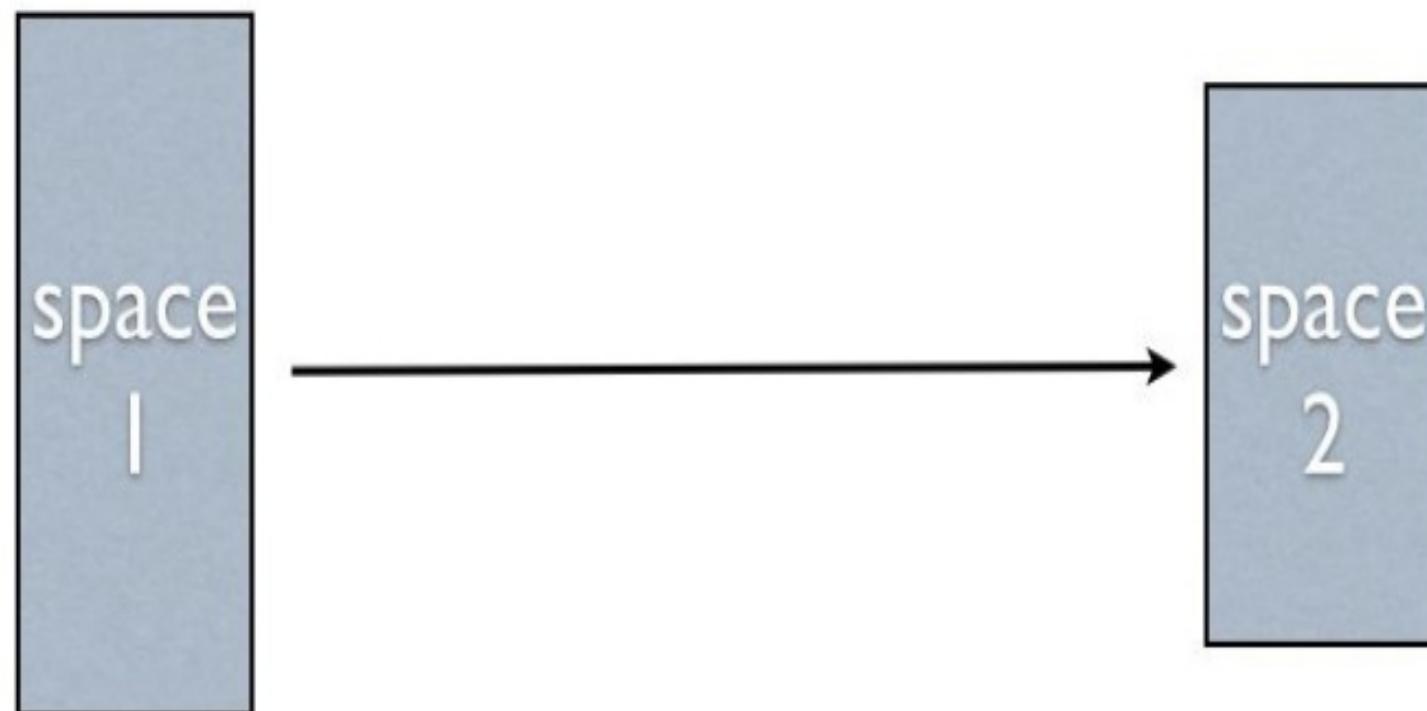


What do I mean by understanding?



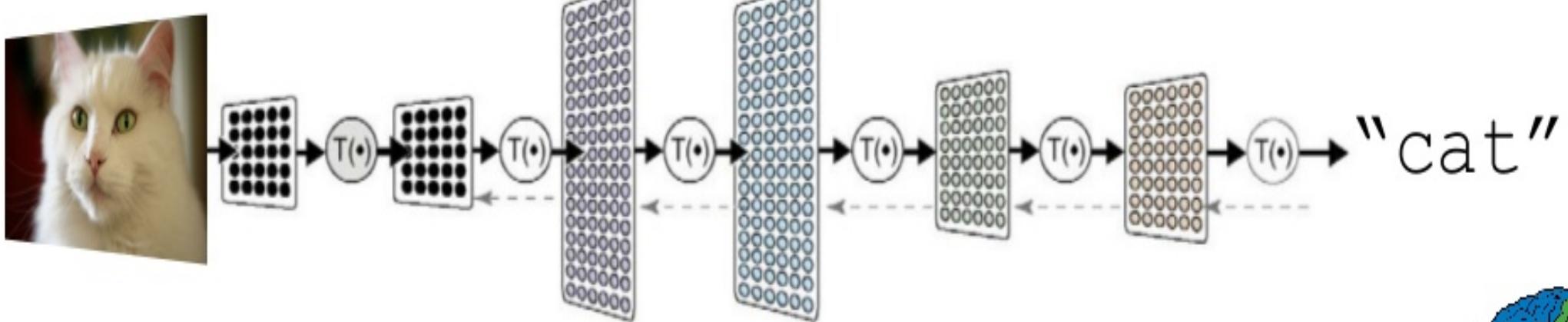
Neural Networks

- Learn a complicated function from data



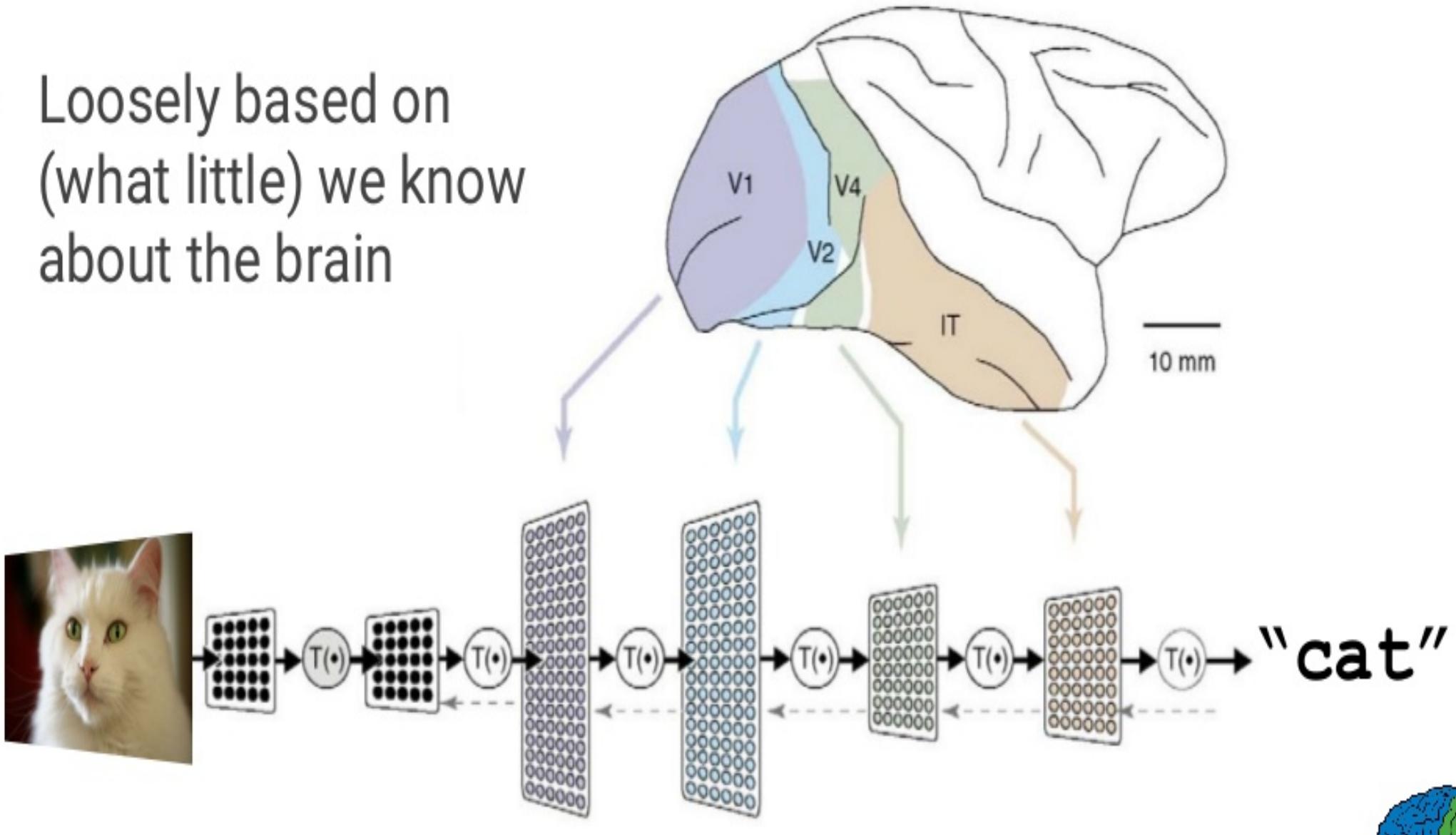
What is Deep Learning?

- A powerful class of machine learning model
- Modern reincarnation of artificial neural networks
- Collection of simple, trainable mathematical functions

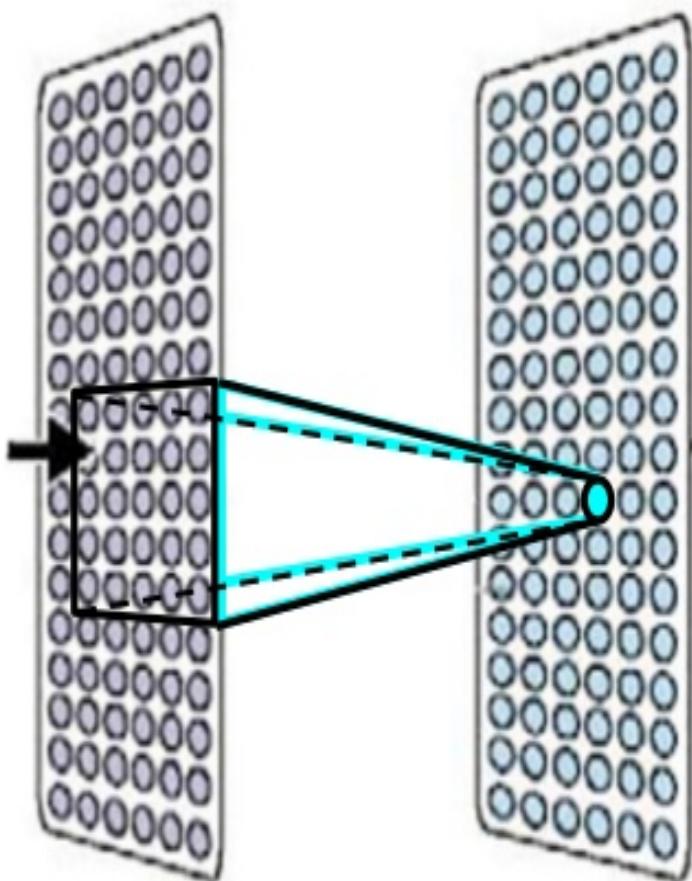


What is Deep Learning?

- Loosely based on (what little) we know about the brain



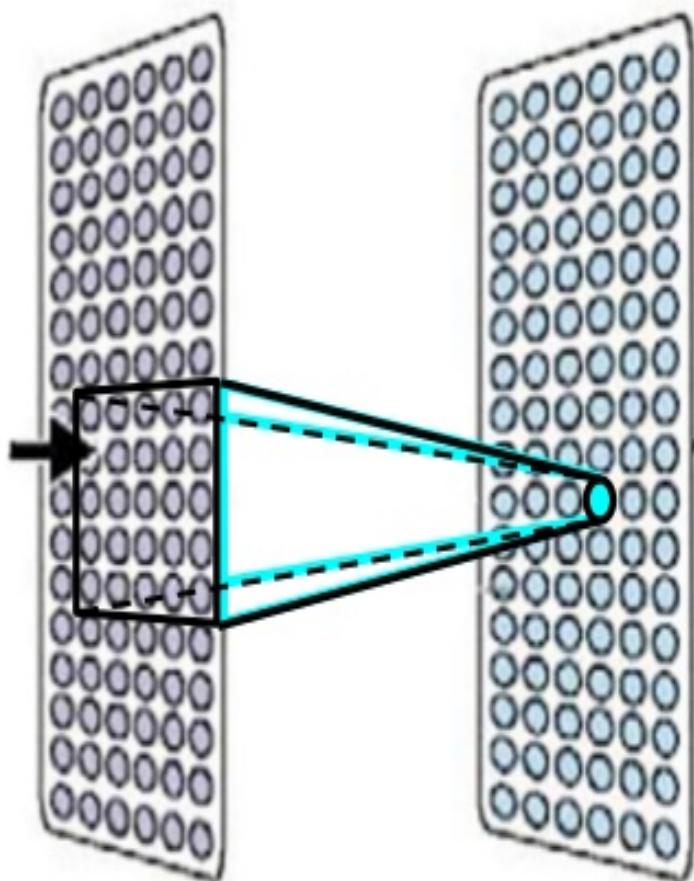
What is Deep Learning?



Commonalities with real brains:

- Each neuron is connected to a small subset of other neurons.
- Based on what it sees, it decides what it wants to say.
- Neurons learn to cooperate to accomplish the task.

What is Deep Learning?



Each neuron implements a relatively simple mathematical function.

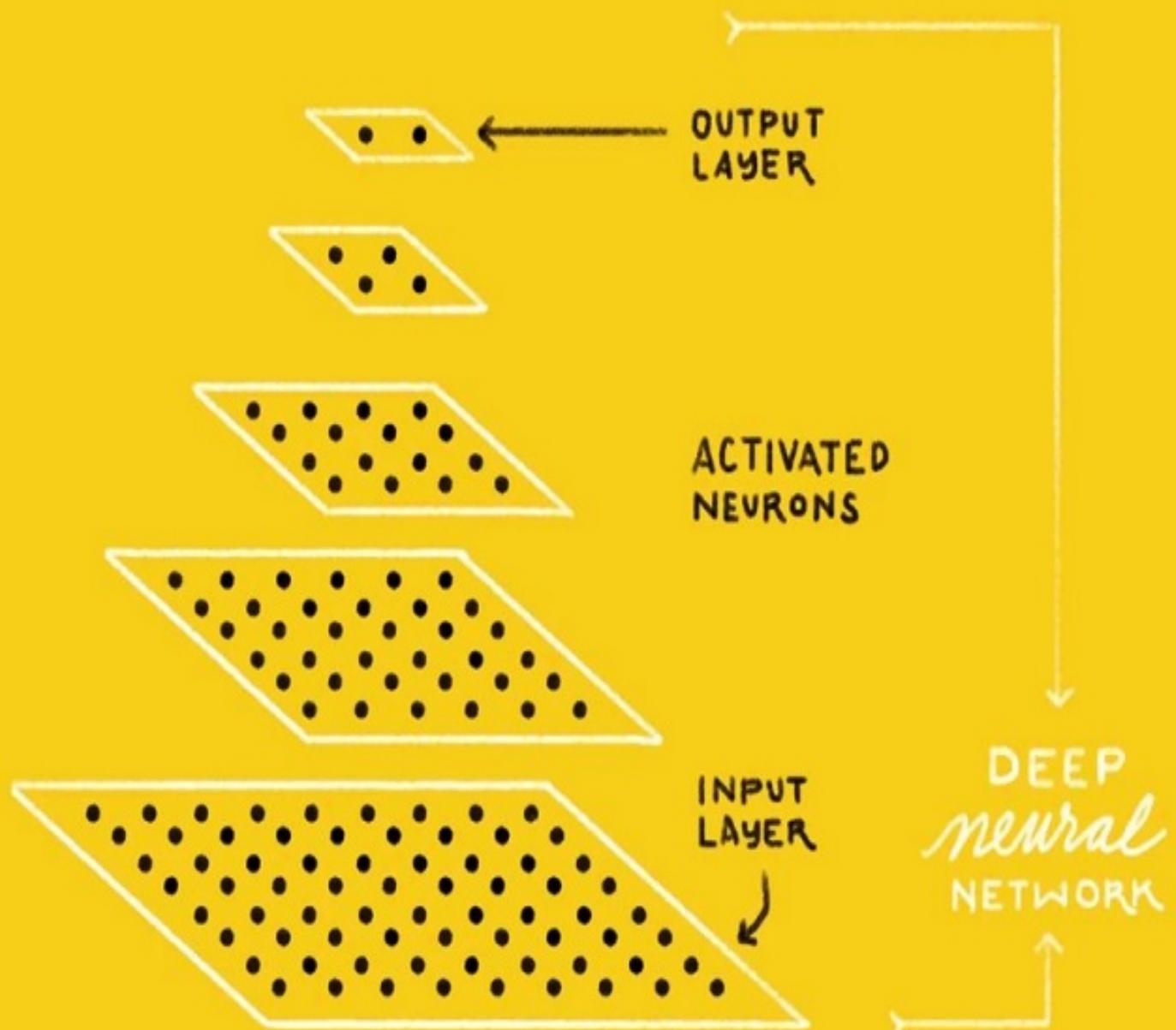
$$y = g(\vec{w} \cdot \vec{x} + b)$$

But the composition of 10^6 - 10^9 such functions is surprisingly powerful.

IS THIS A
CAT or DOG?



CAT DOG



Important Property of Neural Networks

Results get better with

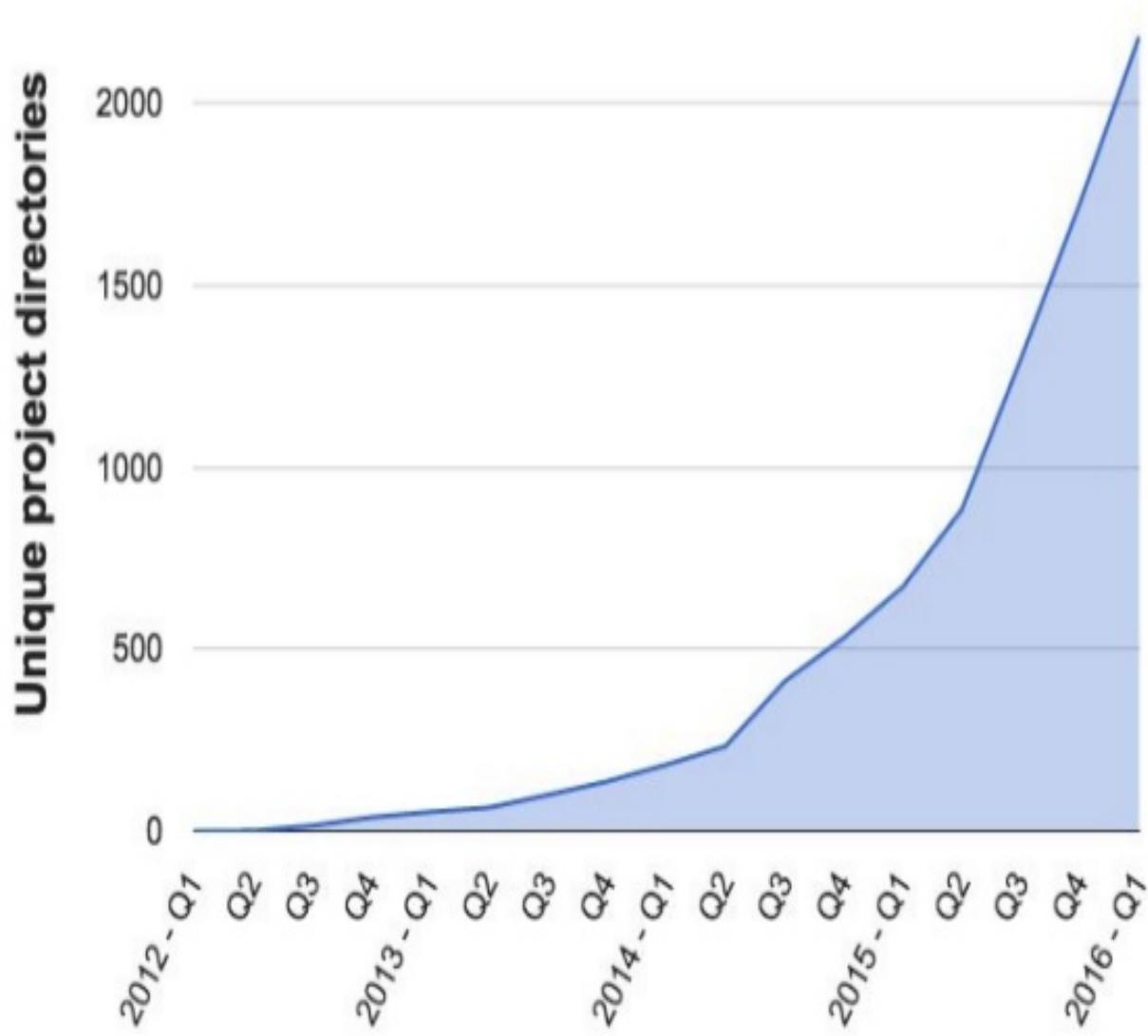
**more data +
bigger models +
more computation**

(Better algorithms, new insights and improved
techniques always help, too!)



Growing Use of Deep Learning at Google

of directories containing model description files



Across many products/areas:

- Android
- Apps
- drug discovery
- Gmail
- Image understanding
- Maps
- Natural language understanding
- Photos
- Robotics research
- Speech
- Translation
- YouTube
- ... many others ...





<http://tensorflow.org/>

and

<https://github.com/tensorflow/tensorflow>

Open, standard software for
general machine learning

Great for Deep Learning in
particular

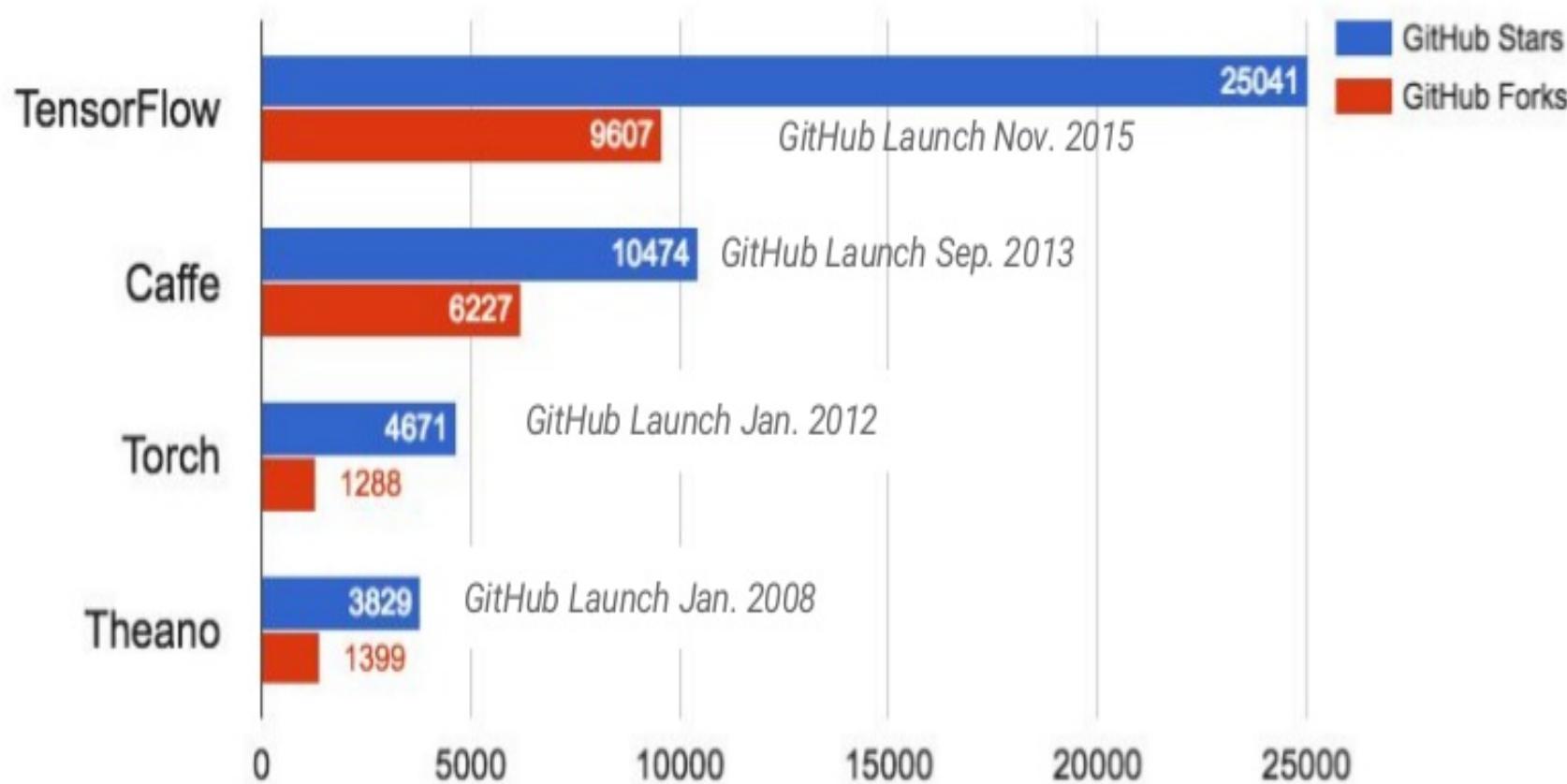
First released Nov 2015

Apache 2.0 license

Strong External Adoption



Adoption of Deep Learning Tools on GitHub



50,000+ binary installs in 72 hours, 500,000+ since Nov, 2015

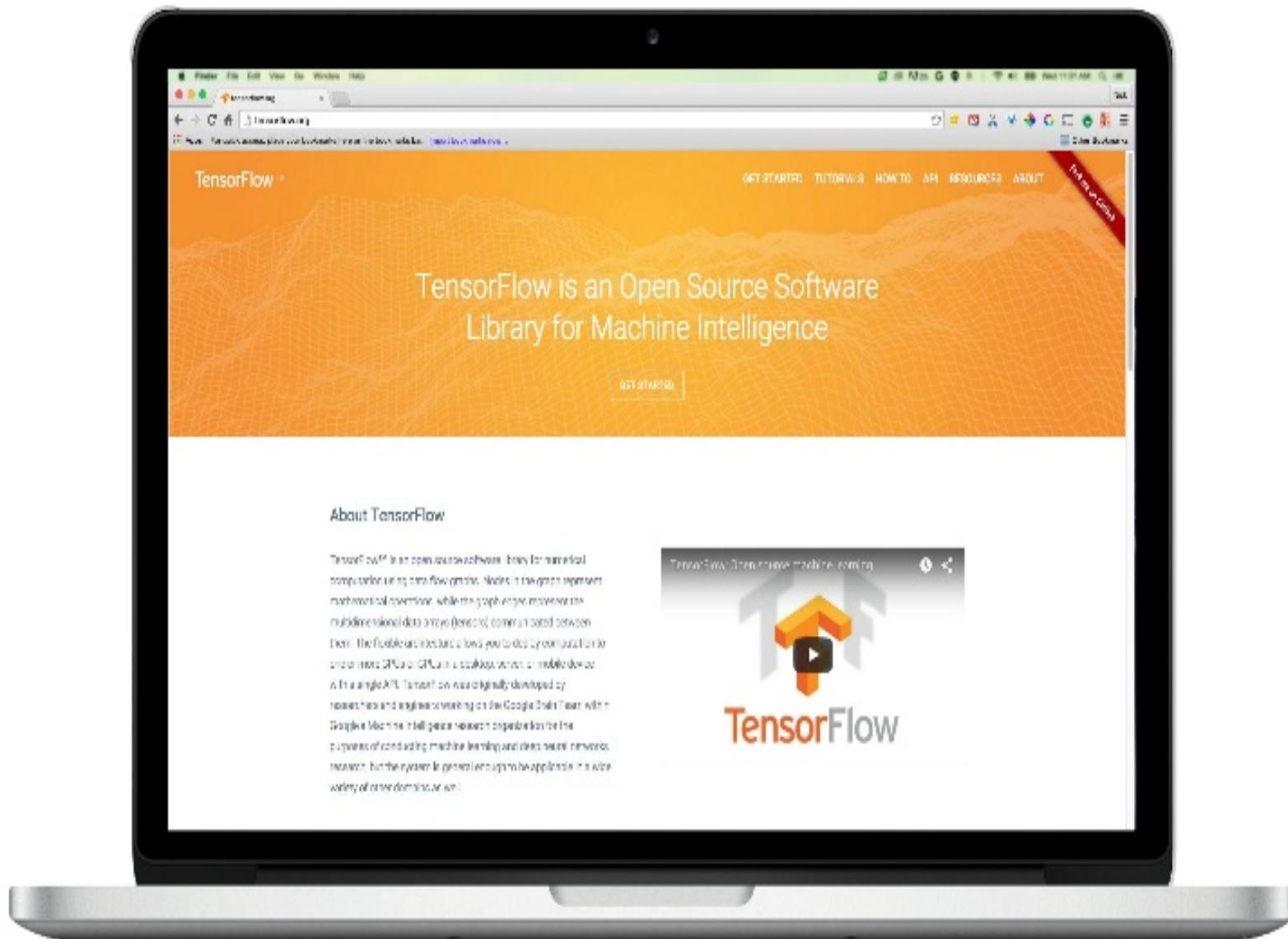
Most forks of any GitHub repo in 2015, despite only being available starting in Nov, 2015



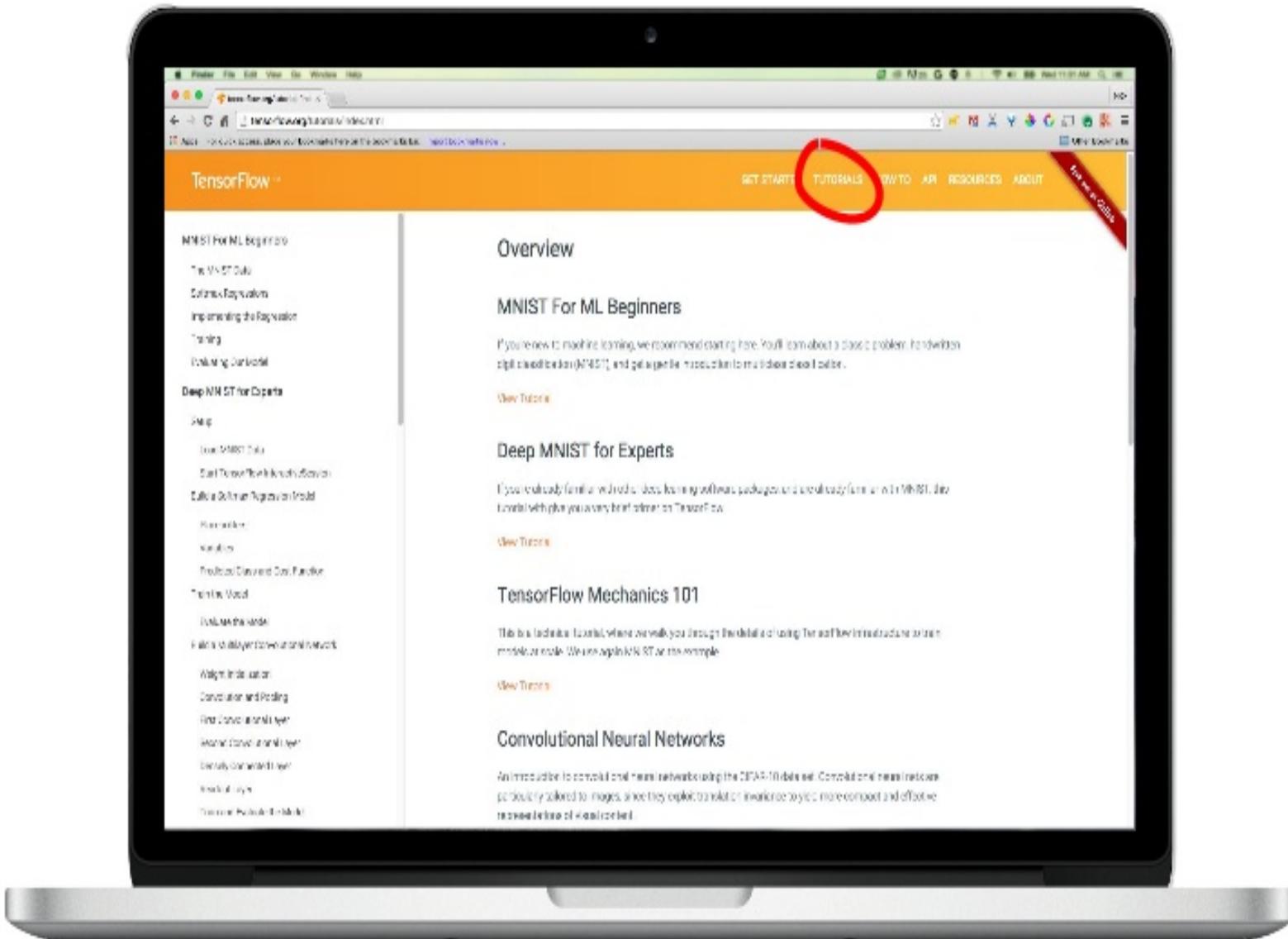
Motivations

- DistBelief (our 1st system) was scalable and good for production deployment, but not as flexible as we wanted for research purposes
- Better understanding of problem space allowed us to make some dramatic simplifications
- Define a standard way of expressing machine learning ideas and computations
- Short circuit the MapReduce/Hadoop inefficiency

<http://tensorflow.org/>

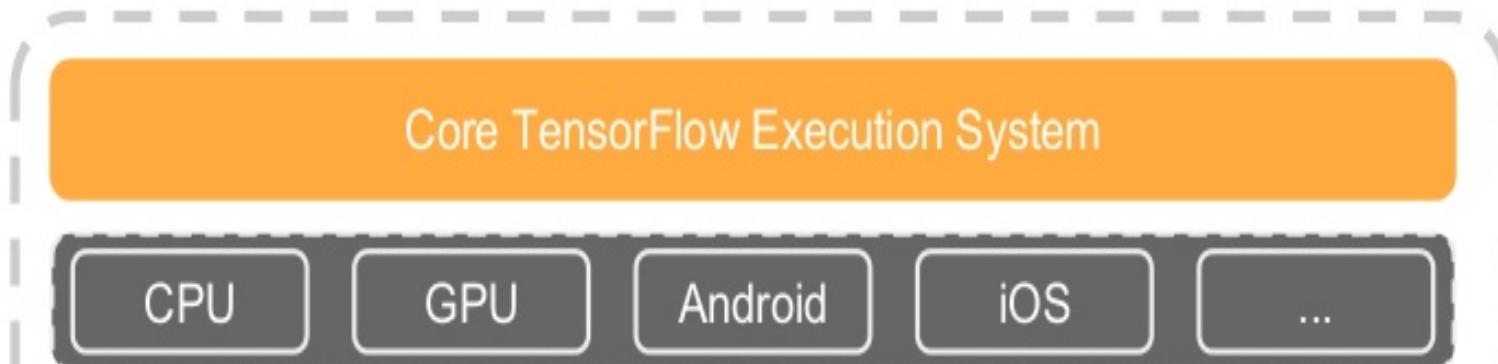


http://tensorflow.org/



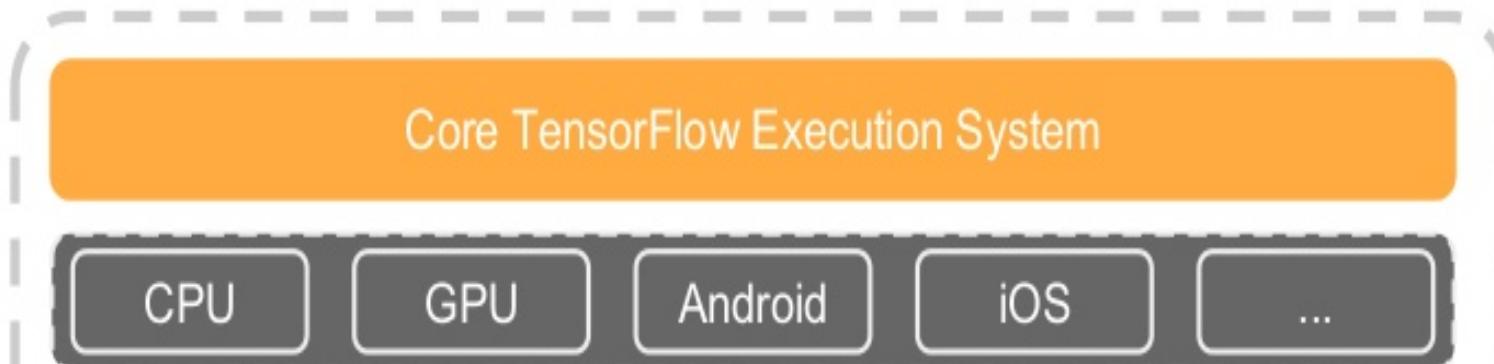
TensorFlow: Expressing High-Level ML Computations

- Core in C++
 - Very low overhead



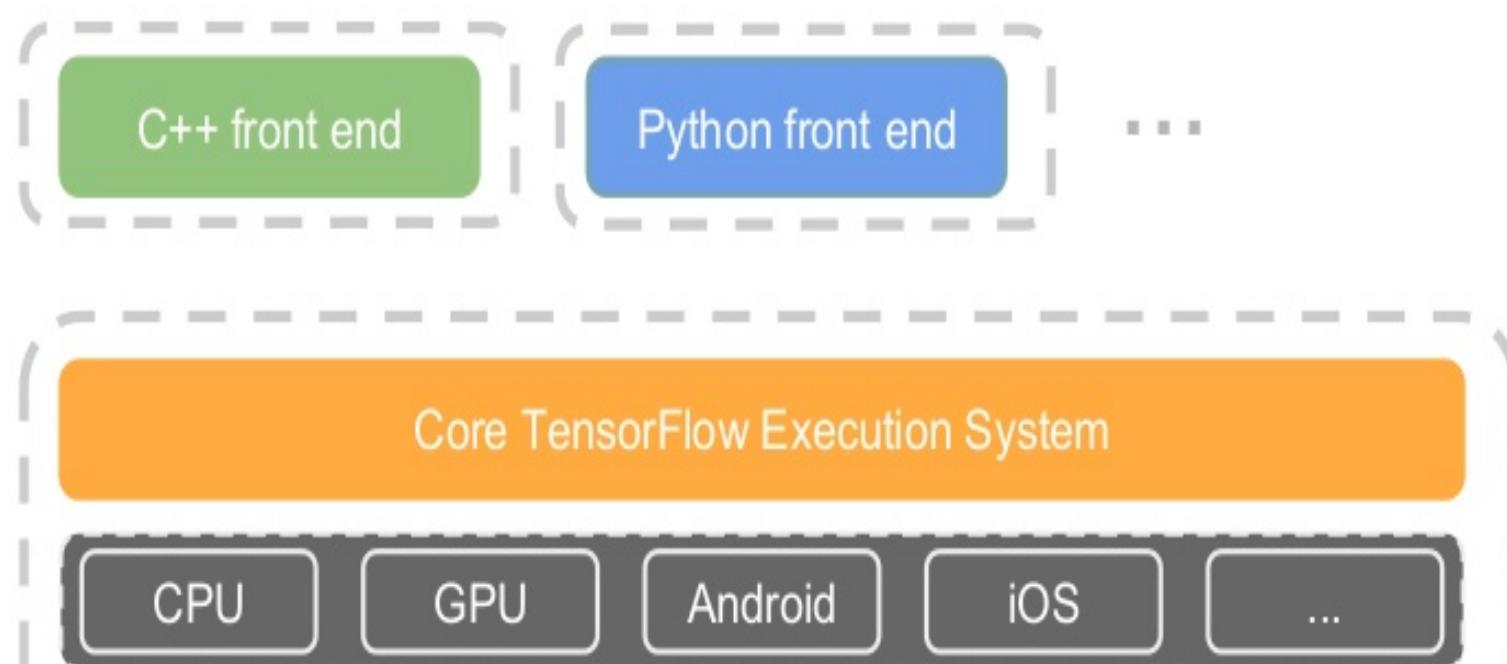
TensorFlow: Expressing High-Level ML Computations

- Core in C++
 - Very low overhead
- Different front ends for specifying/driving the computation
 - Python and C++ today, easy to add more



TensorFlow: Expressing High-Level ML Computations

- Core in C++
 - Very low overhead
- Different front ends for specifying/driving the computation
 - Python and C++ today, easy to add more



Example TensorFlow fragment

- Build a graph computing a neural net inference.

```
import tensorflow as tf
from tensorflow.examples.tutorials.mnist import input_data

mnist = input_data.read_data_sets('MNIST_data', one_hot=True)
x = tf.placeholder("float", shape=[None, 784])
W = tf.Variable(tf.zeros([784,10]))
b = tf.Variable(tf.zeros([10]))
y = tf.nn.softmax(tf.matmul(x, W) + b)
```

Python API for Machine Learning

- Automatically add ops to calculate symbolic gradients of variables w.r.t. loss function.
- Apply these gradients with an optimization algorithm

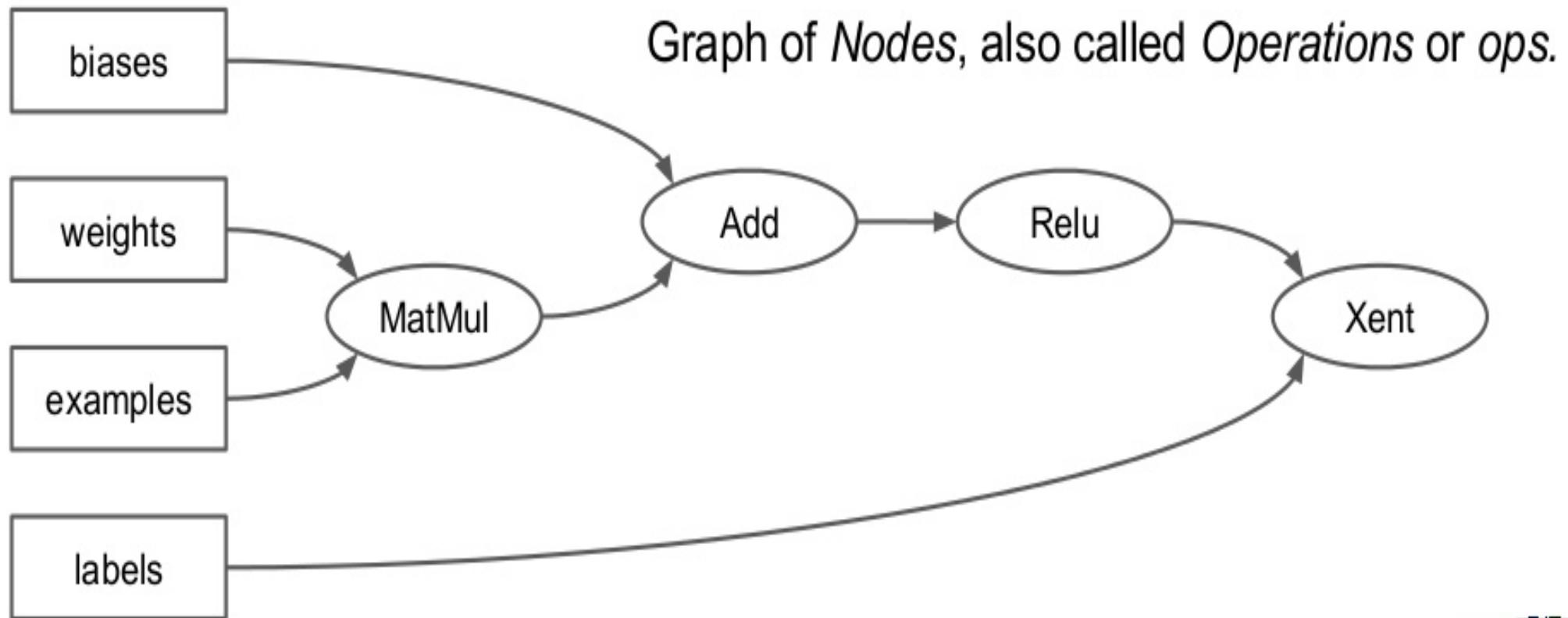
```
y_ = tf.placeholder(tf.float32, [None, 10])
cross_entropy = -tf.reduce_sum(y_*tf.log(y))
opt = tf.train.GradientDescentOptimizer(0.01)
train_op = opt.minimize(cross_entropy)
```

Python API for Machine Learning

- Launch the graph and run the training ops in a loop

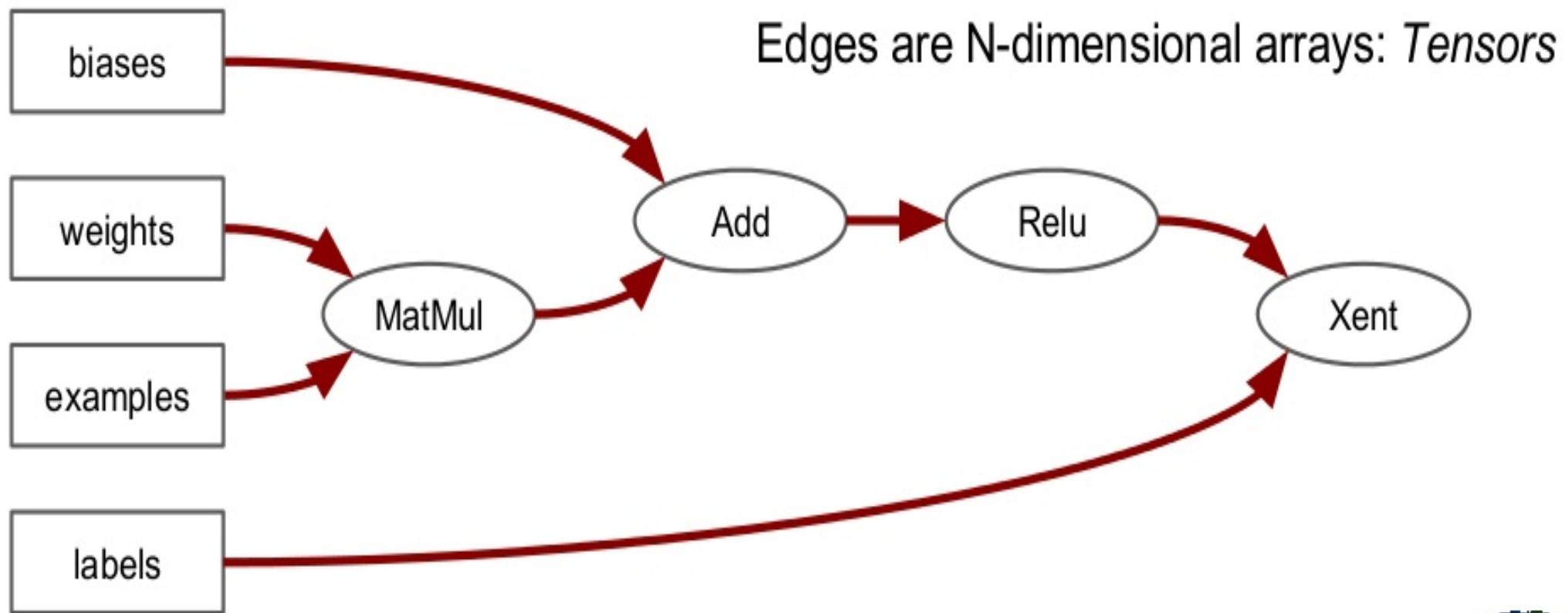
```
init = tf.initialize_all_variables()
sess = tf.Session()
sess.run(init)
for i in range(1000):
    batch_xs, batch_ys = mnist.train.next_batch(100)
    sess.run(train_step, feed_dict={x: batch_xs, y_: batch_ys})
```

Computation is a dataflow graph



Computation is a dataflow graph

with tensors



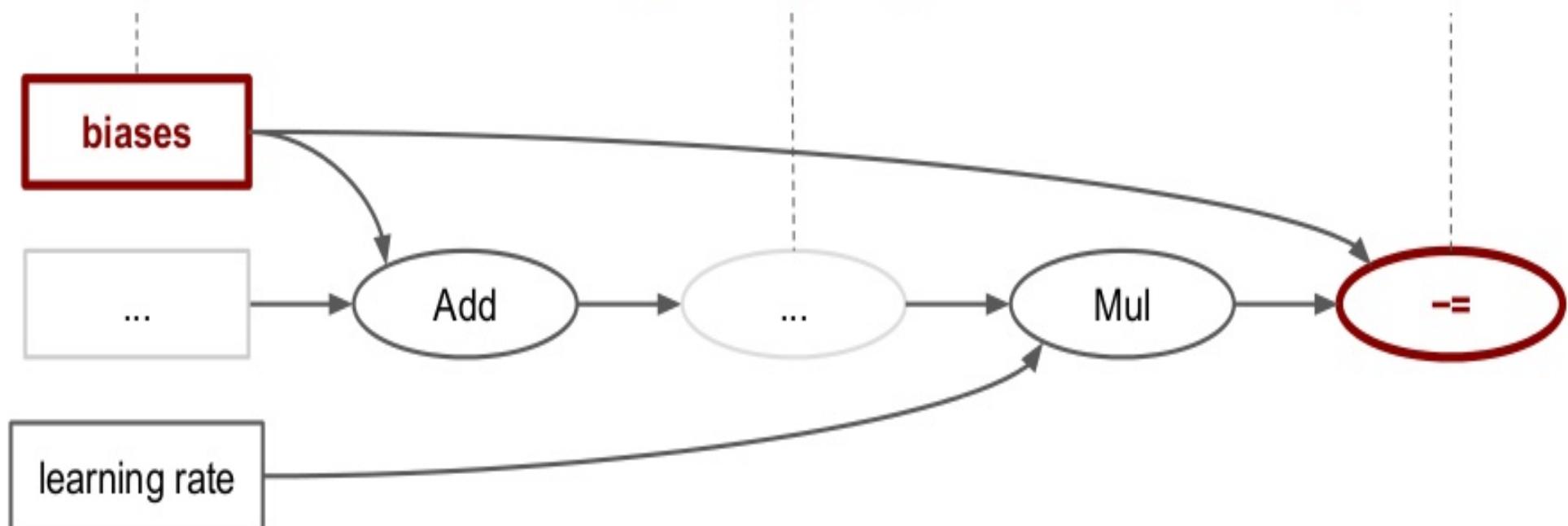
Computation is a dataflow graph

with state

'Biases' is a variable

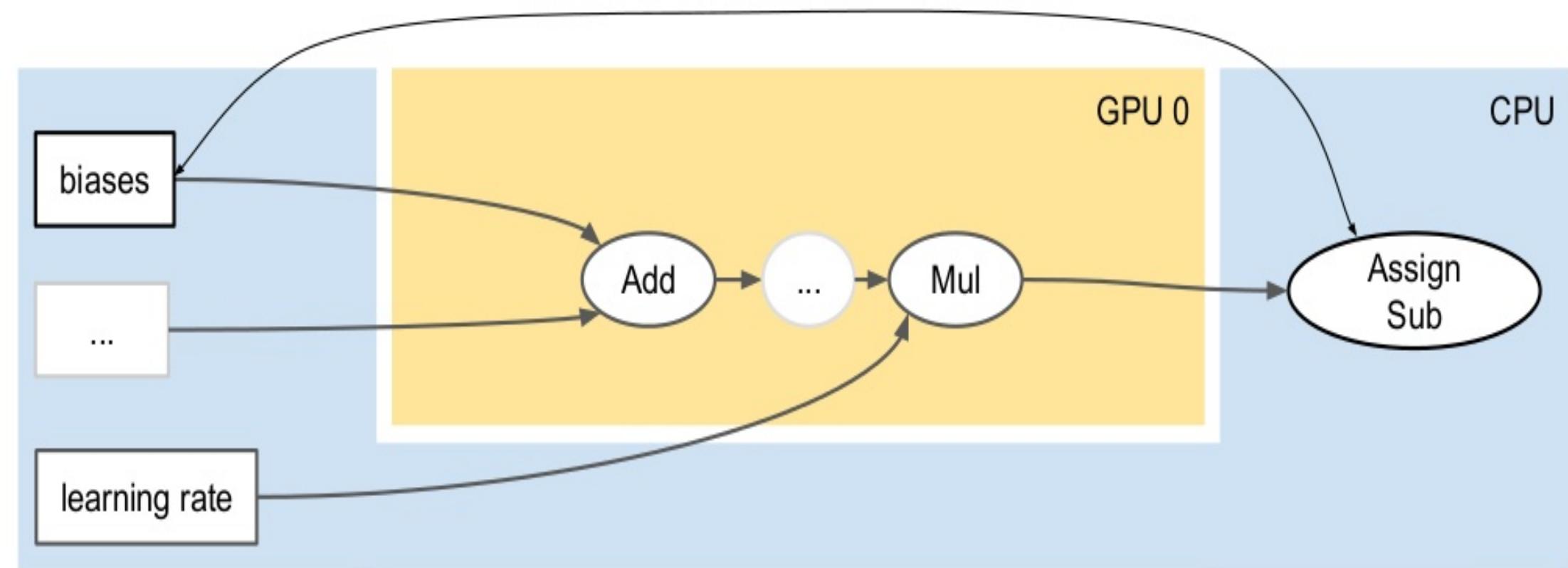
Some ops compute gradients

`-=` updates biases



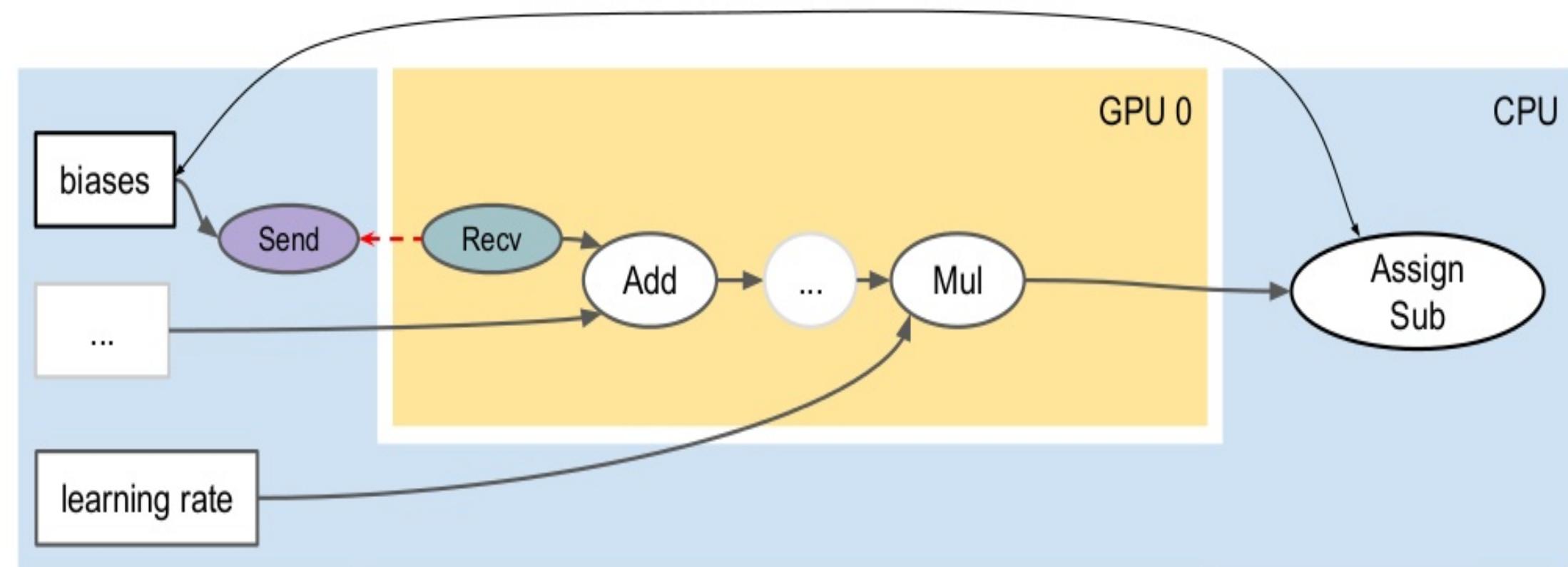
Computation is a dataflow graph

distributed



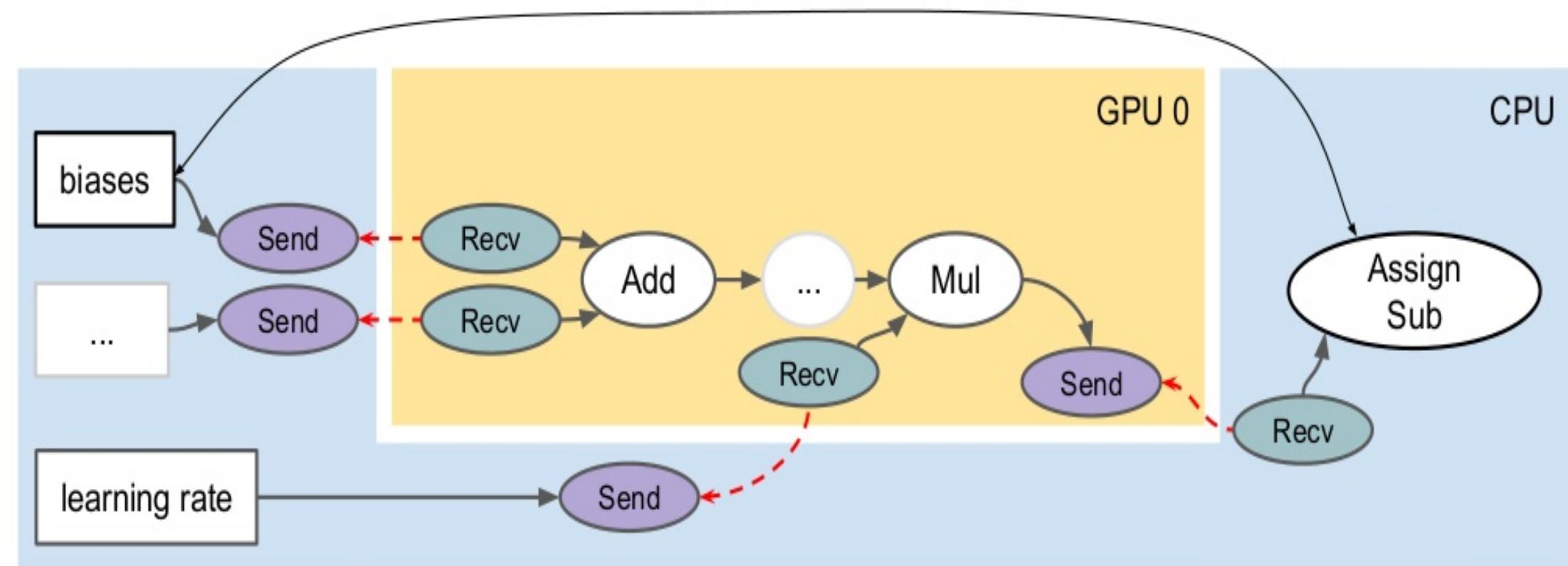
Assign Devices to Ops

- TensorFlow inserts Send/Recv Ops to transport tensors across devices
- Recv ops pull data from Send ops



Assign Devices to Ops

- TensorFlow inserts Send/Recv Ops to transport tensors across devices
- Recv ops pull data from Send ops



Automatically Runs on Variety of Platforms

phones



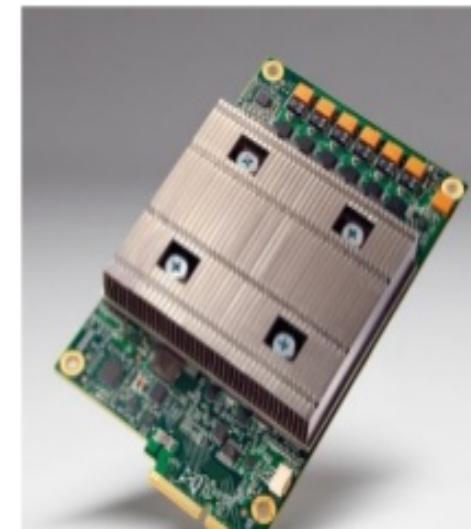
distributed systems of 100s
of machines and/or GPU cards



single machines (CPU and/or GPUs) ...



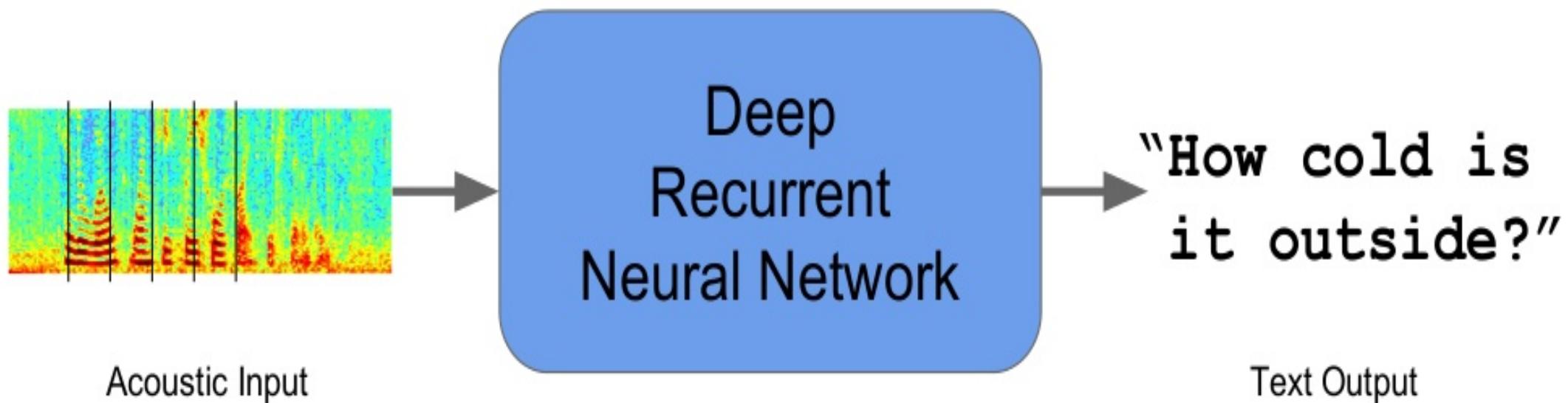
custom ML hardware



What are some ways that
deep learning is having
a significant impact at Google?

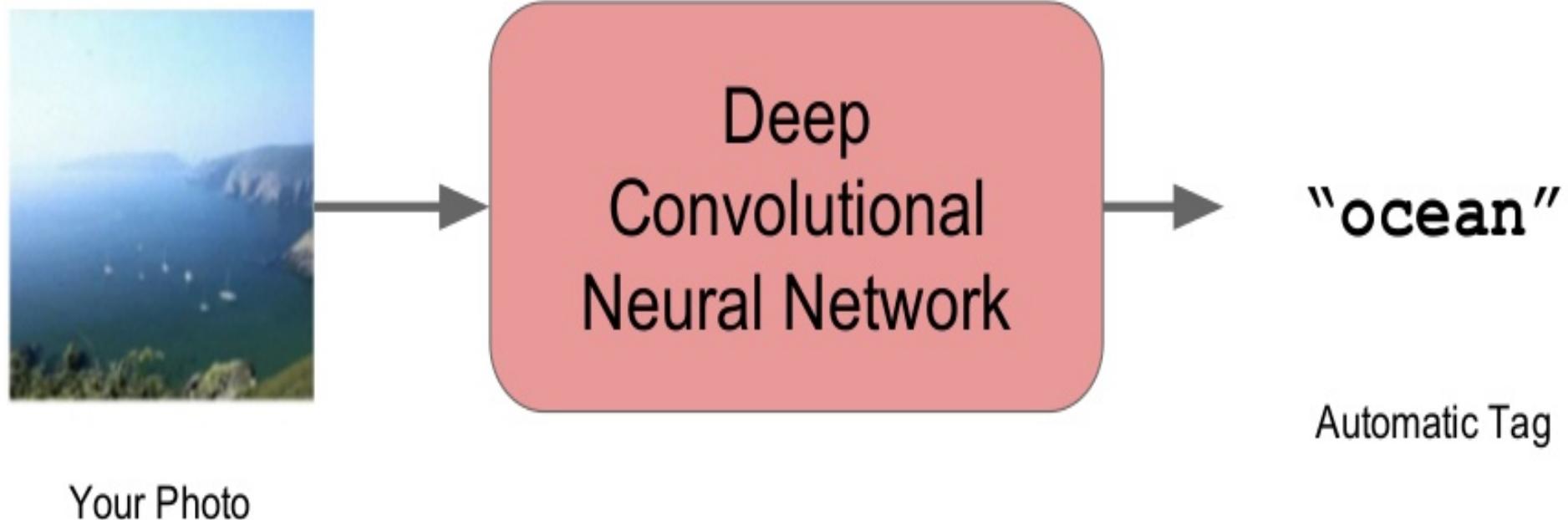


Speech Recognition



Reduced word errors by more than 30%

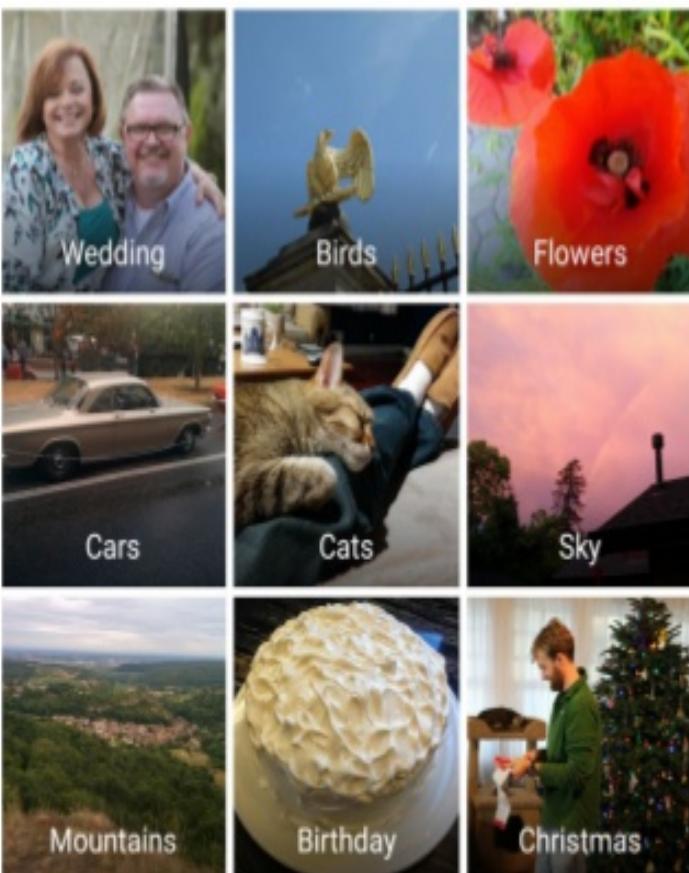
Google Photos Search



Search personal photos without tags.

Google Photos Search

Things



Google my photos of siamese cats

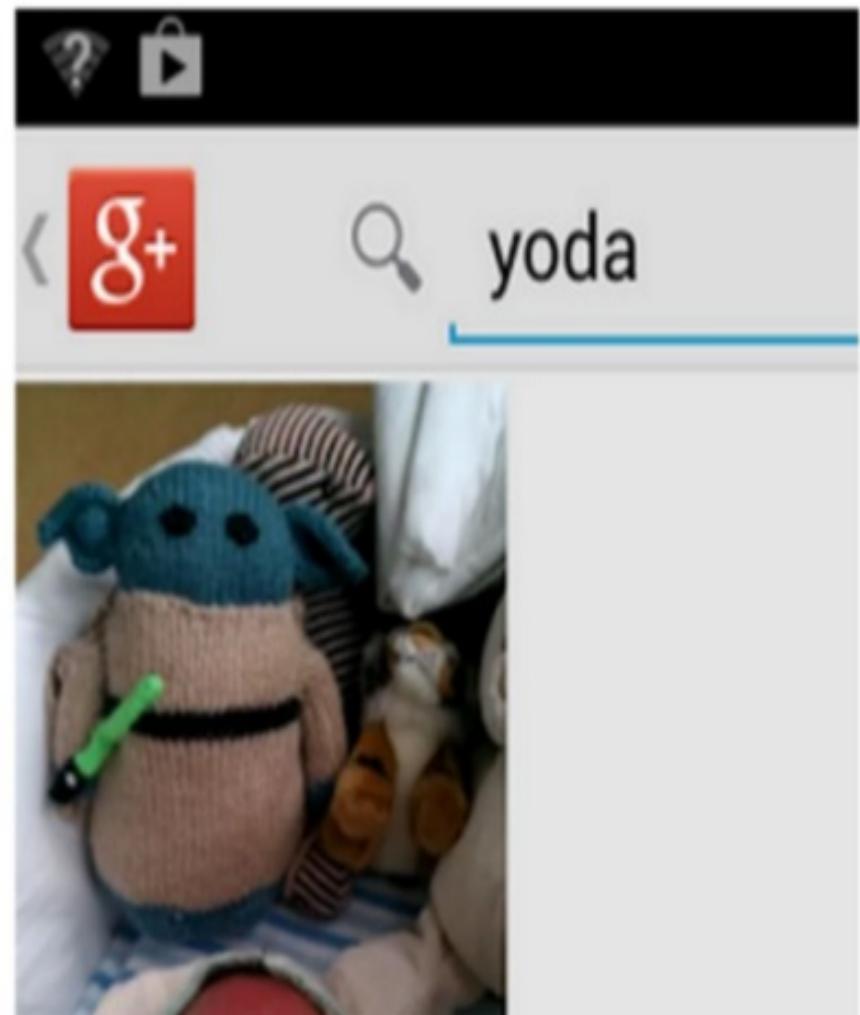
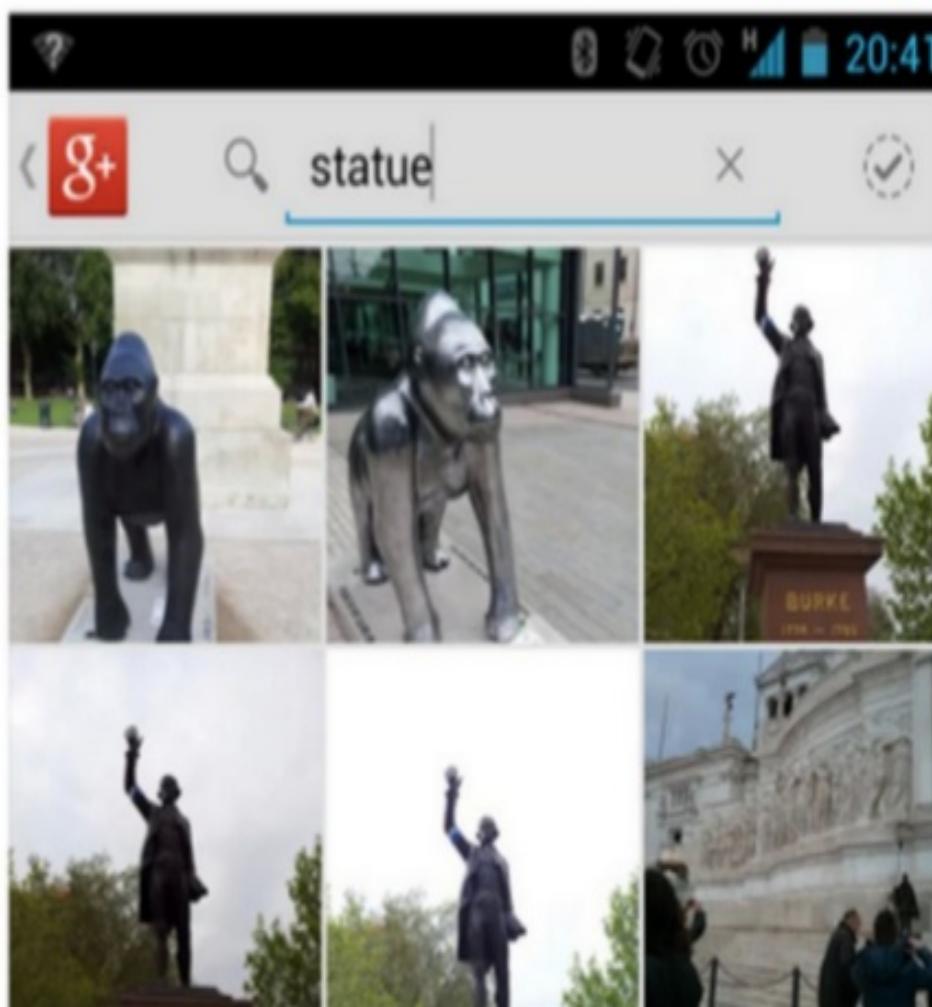
Web Images Shopping Videos More ▾

Your photos
Only you can see these results

The search results page shows a header with the Google logo and search bar, followed by a navigation bar with 'Web' selected. Below this is a section titled 'Your photos' with the subtext 'Only you can see these results'. The main area contains a 3x4 grid of 12 thumbnail images, each depicting a Siamese cat in different scenarios: sitting on a person's lap, being petted, sleeping, and interacting with other cats.

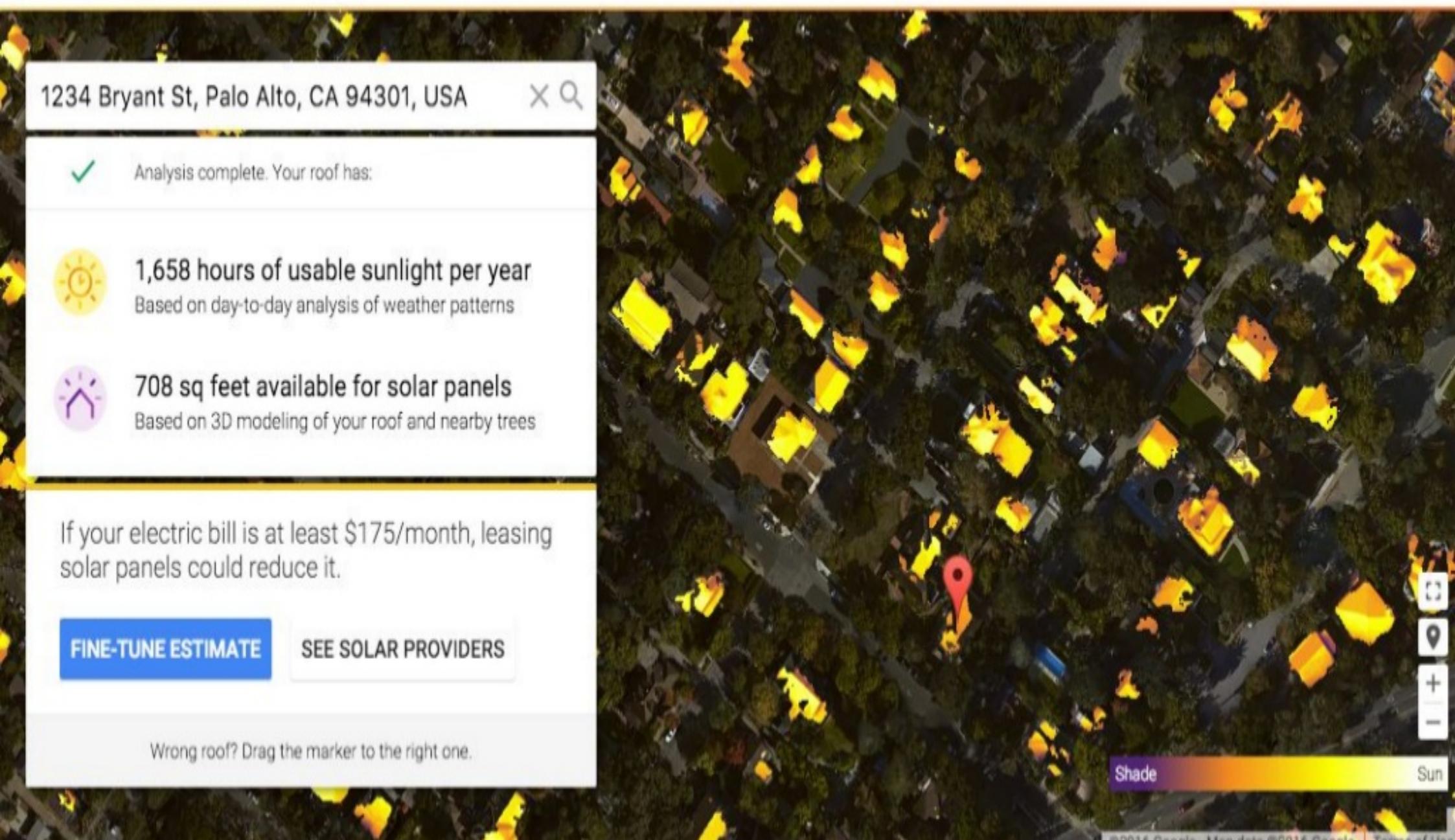
Google Photos Search

"Wow. The new Google photo search is a bit insane. I didn't tag those... :)"



Google Project Sunroof

www.google.com/sunroof



“Seeing” Go

Google's AI just cracked the game that supposedly no computer could beat

By Mike Murphy | January 27, 2016

BBC News Sport Weather iPlayer TV Radio More Search Find local news

Home UK World Business Politics Tech Science Health Education Entertainment & Arts More

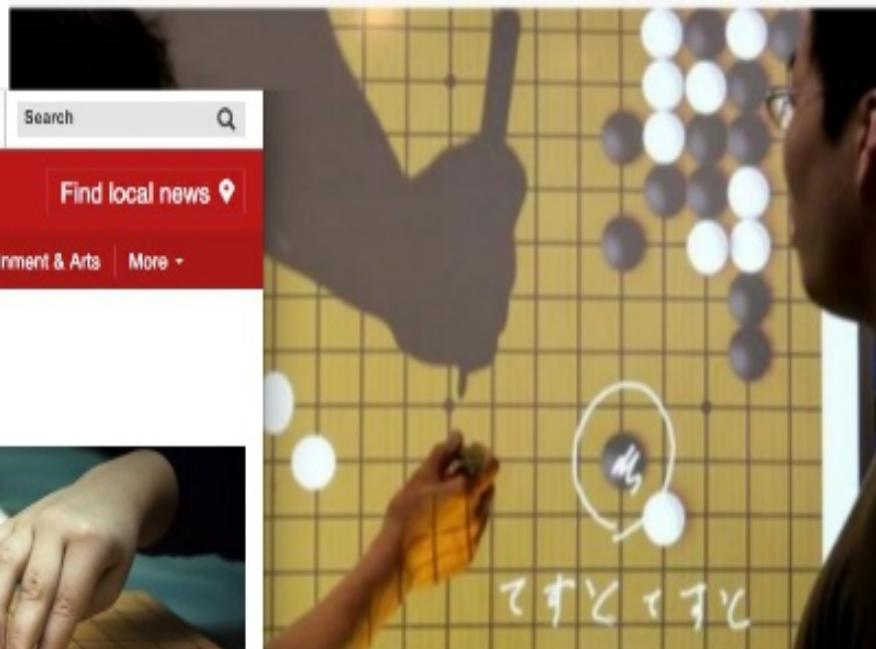
Technology

Google achieves AI 'breakthrough' at Go

An artificial intelligence program developed by Google beats Europe's top player at the ancient Chinese game of Go, about a decade earlier than expected.

27 January 2016 | Technology

How did they do it? What is the game Go? Facebook trains AI to beat humans at Go



Kiyoshi Ota)

Only started to encroach on activities we previously thought only the brilliant human brain could handle. In 1997, a computer beat Grand Master Garry Kasparov at chess, and in 2011 IBM's Watson beat former human winners at the quiz game *Jeopardy*. But the ancient board game Go has long been one of the major goals of artificial intelligence research. It's understood to be one of the most difficult games for computers to handle due to the sheer number of possible moves a player can make at any given point.



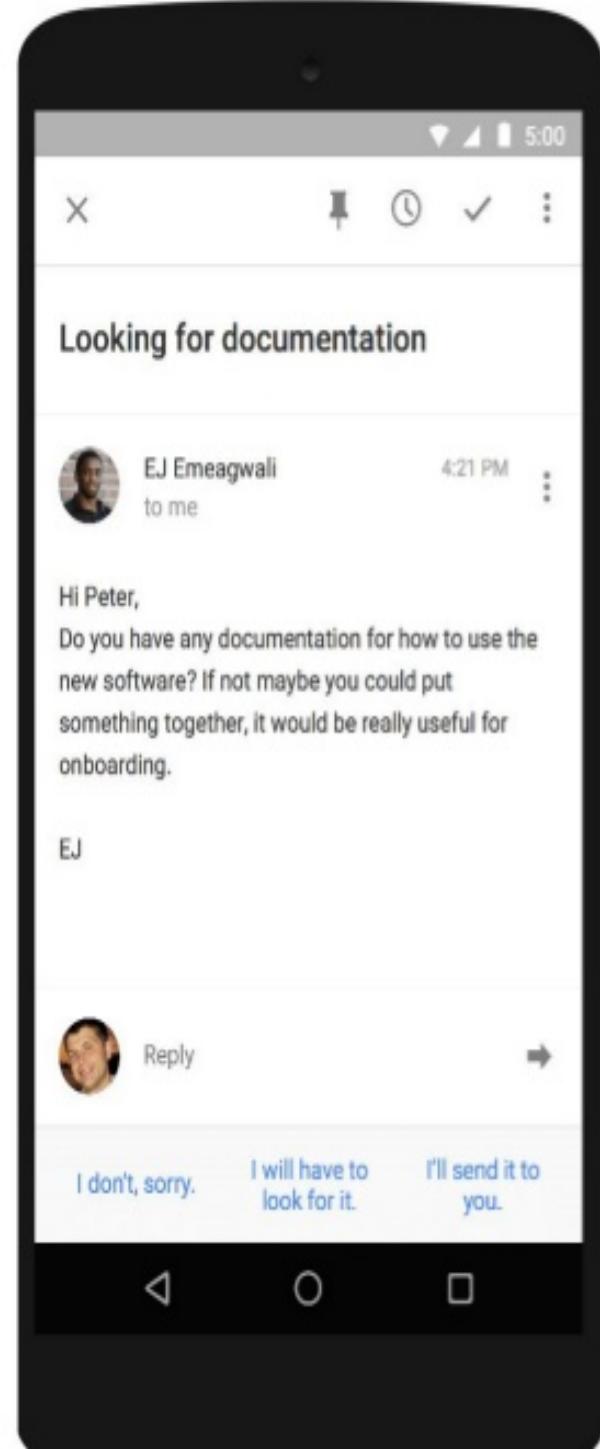


Smart Reply

April 1, 2009: April Fool's Day joke

Nov 5, 2015: Launched Real Product

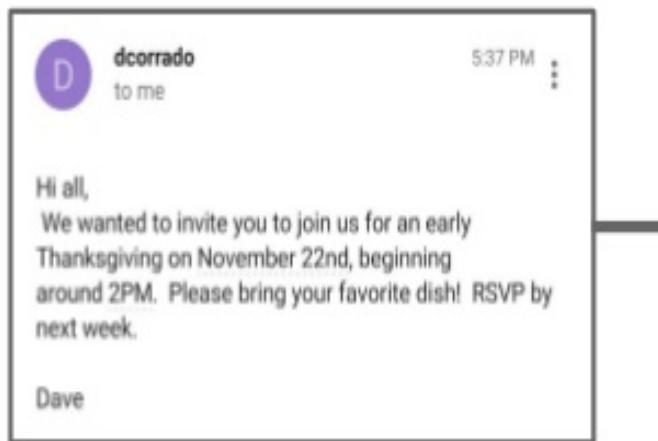
Feb 1, 2016: >10% of mobile Inbox replies



Google Research Blog
- Nov 2015

Incoming Email

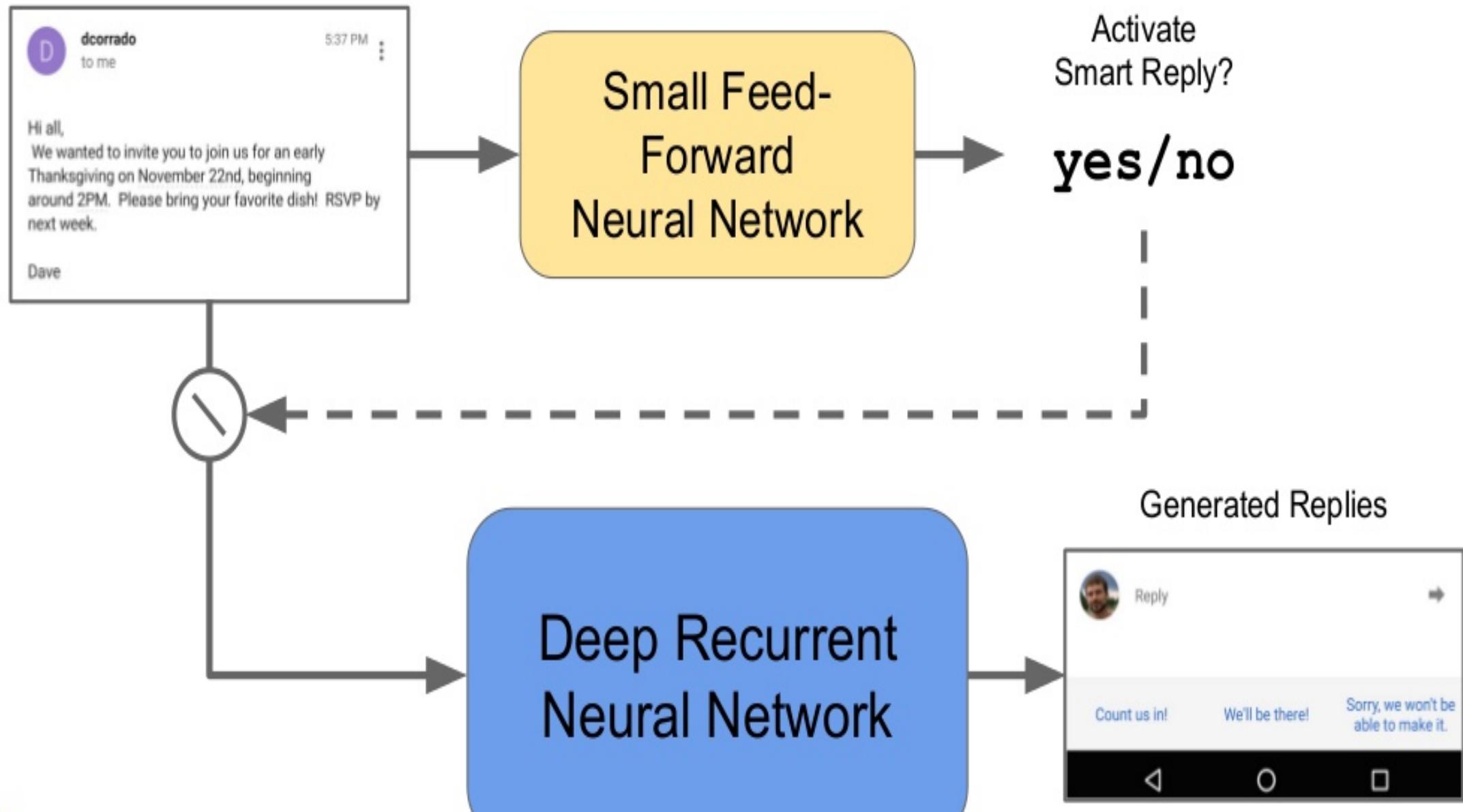
Smart Reply



Activate
Smart Reply?
yes/no

Incoming Email

Smart Reply



Combined Vision + Translation

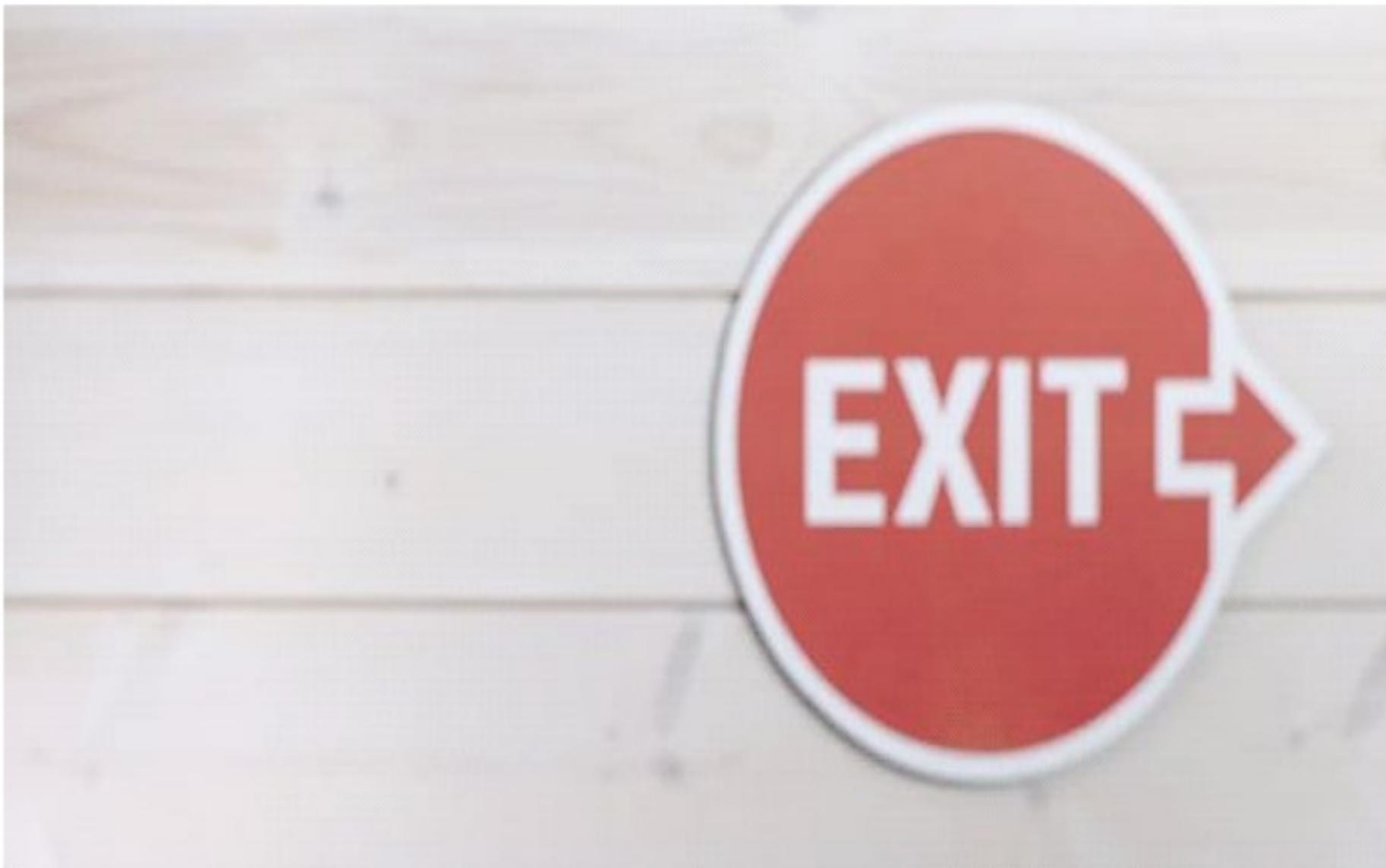


Image Captions Research



Human: A young girl asleep on the sofa cuddling a stuffed bear.

Model: A close up of a child holding a stuffed animal.

Model: A baby is asleep next to a teddy bear.



A man holding a tennis racquet
on a tennis court.



A group of young people



Two pizzas sitting on top
of a stove top oven



A man flying through the air



Experiment Turnaround Time and Research Productivity

- Minutes, Hours:
 - **Interactive research! Instant gratification!**
- 1-4 days
 - Tolerable
 - Interactivity replaced by running many experiments in parallel
- 1-4 weeks
 - High value experiments only
 - Progress stalls
- >1 month
 - Don't even try



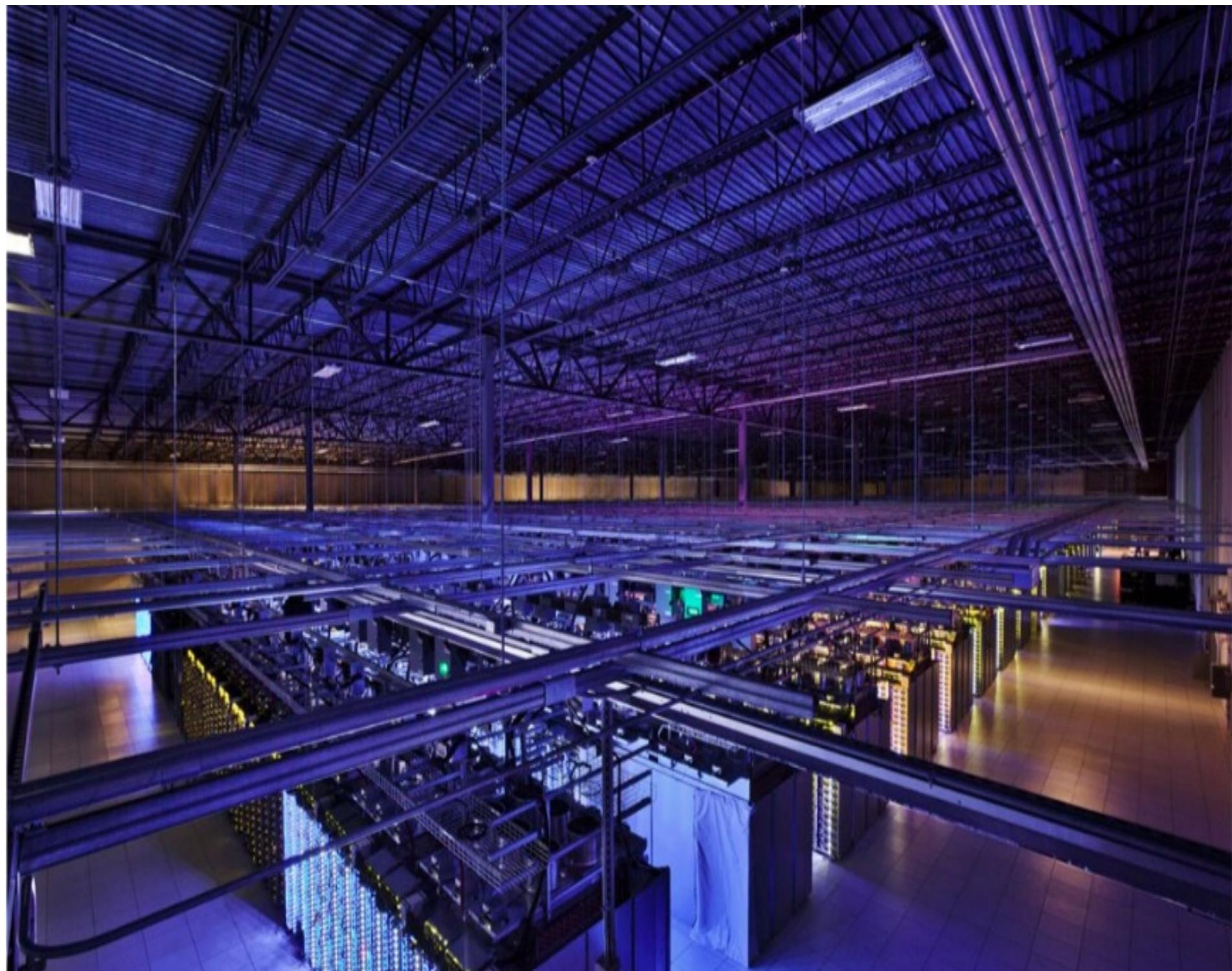


Image Model Training Time

Precision @ 1

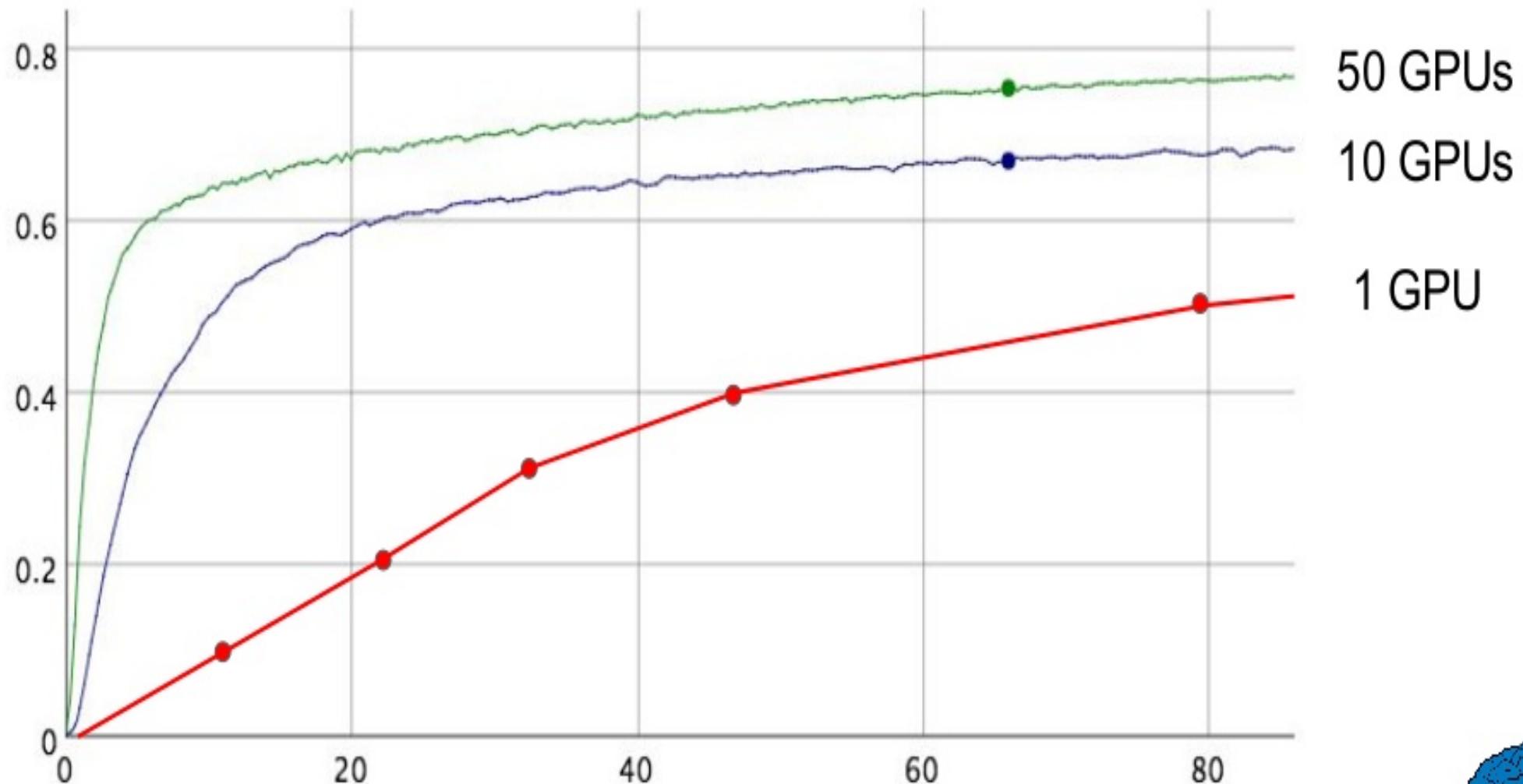
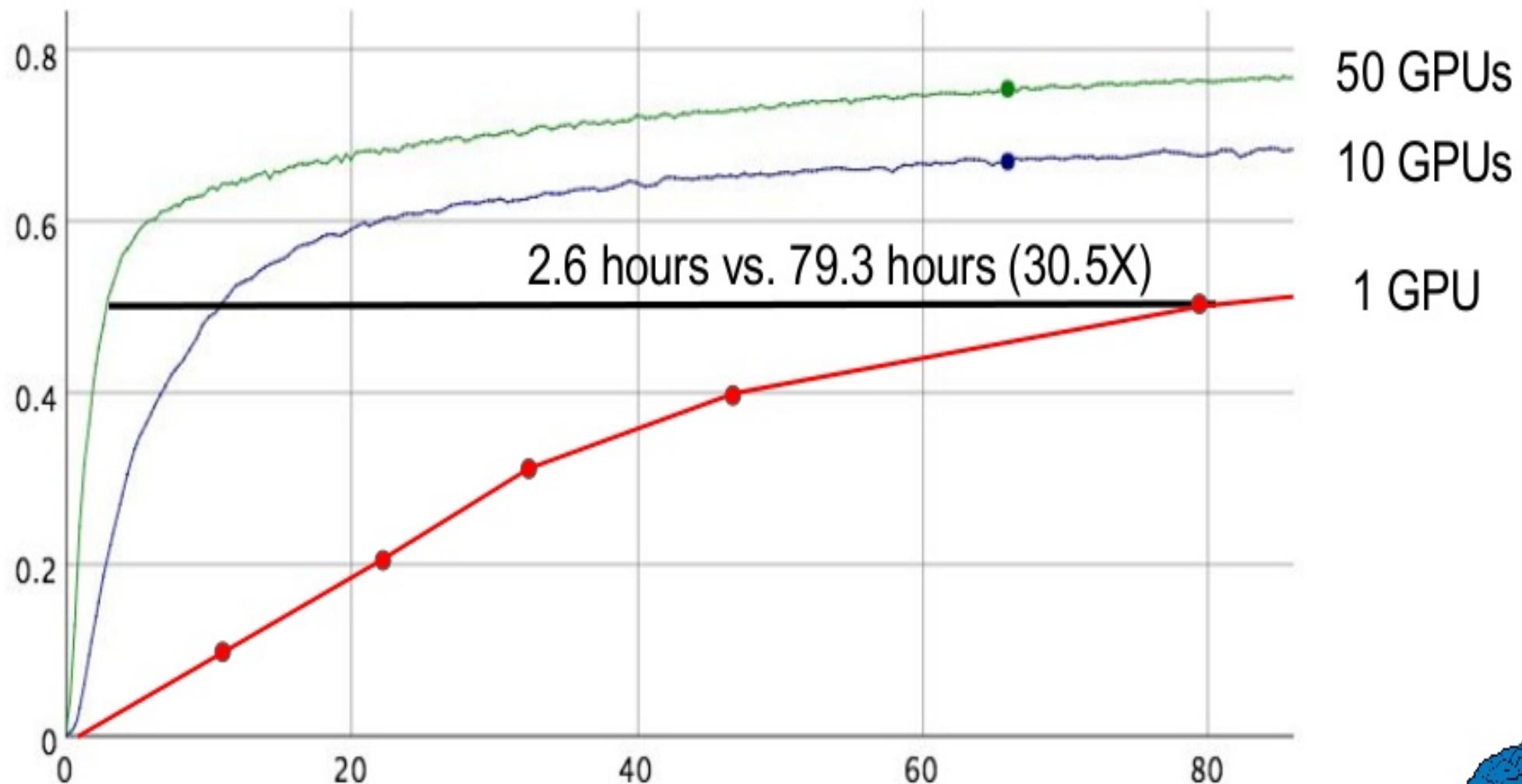


Image Model Training Time

Precision @ 1



How Can You Get Started with Machine Learning?

Four ways, with varying complexity:

- (1) Use a Cloud-based API (Vision, Speech, etc.)
- (2) Run your own pretrained model
- (3) Use an existing model architecture, and retrain it or fine tune on your dataset
- (4) Develop your own machine learning models for new problems

More
flexible,
but more
effort
required



(1) Use Cloud-based APIs



GOOGLE TRANSLATE API

Dynamically translate between thousands of available language pairs

cloud.google.com/translate



CLOUD SPEECH API

ALPHA

Speech to text conversion powered by machine learning

cloud.google.com/speech



CLOUD VISION API

Derive insight from images with our powerful Cloud Vision API

cloud.google.com/vision

CLOUD TEXT API

ALPHA

Use Cloud Text API for sentiment analysis and entity recognition in a piece of text

cloud.google.com/text

(1) Use Cloud-based APIs



GOOGLE TRANSLATE API

Dynamically translate between thousands of available language pairs

cloud.google.com/translate



CLOUD SPEECH API

ALPHA

Speech to text conversion powered by machine learning

cloud.google.com/speech



CLOUD VISION API

Derive insight from images with our powerful Cloud Vision API

cloud.google.com/vision

CLOUD TEXT API

ALPHA

Use Cloud Text API for sentiment analysis and entity recognition in a piece of text

cloud.google.com/text

Google Cloud Vision API

<https://cloud.google.com/vision/>



"running", "score": 0.99803412,
"marathon", "score": 0.99482006



"joyLikelihood": "VERY_LIKELY"

"description": "ABIERTO\\n",
"local": "es"

(2) Using a Pre-trained Image Model yourself with TensorFlow

www.tensorflow.org/tutorials/image_recognition/index.html

TensorFlow™

GET STARTED TUTORIALS HOW TO API RESOURCES ABOUT

Find me on GitHub

Version: v0.8

Usage with Python API

`classify_image.py` downloads the trained model from `tensorflow.org` when the program is run for the first time. You'll need about 200M of free space available on your hard disk.

The following instructions assume you installed TensorFlow from a PIP package and that your terminal resides in the TensorFlow root directory.

```
cd tensorflow/models/image/imagenet  
python classify_image.py
```

Deep MNIST for Experts

Setup

The above command will classify a supplied image of a panda bear.

Load MNIST Data

Start TensorFlow InteractiveSession

Build a Softmax Regression Model

Placeholders

Variables

Predicted Class and Cost Function

Train the Model

Evaluate the Model

Build a Multilayer Convolutional



If the model runs correctly, the script will produce the following output:

```
giant panda, panda, panda bear, coon bear, Ailuropoda melanoleuca (score = 0.88493)  
indri, indris, Indri indri, Indri brevicaudatus (score = 0.00878)  
lesser panda, red panda, panda, bear cat, cat bear, Ailurus fulgens (score = 0.00317)  
custard apple (score = 0.00149)
```

For training your own models (3 & 4), two choices:

Run open-source release on your own physical machines or virtual machines in a cloud hosting environment

or



CLOUD MACHINE LEARNING ALPHA

Machine Learning on any data, of any size

cloud.google.com/ml

(3) Training a Model on Your Own Image Data

www.tensorflow.org/versions/master/how_tos/image_retraining/index.html

TensorFlow™

GET STARTED TUTORIALS HOW TO API RESOURCES ABOUT

Fork me on GitHub

Version: master

Variables: Creation,
Initialization, Saving, and
Loading

Creation

Device placement

Initialization

Initialization from another
Variable

Custom Initialization

Saving and Restoring

Checkpoint Files

Saving Variables

Restoring Variables

How to Retrain Inception's Final Layer for New Categories

Modern object recognition models have millions of parameters and can take weeks to fully train. Transfer learning is a technique that shortcuts a lot of this work by taking a fully-trained model for a set of categories like ImageNet, and retrains from the existing weights for new classes. In this example we'll be retraining the final layer from scratch, while leaving all the others untouched. For more information on the approach you can see [this paper on Decaf](#).

Though it's not as good as a full training run, this is surprisingly effective for many applications, and can be run in as little as thirty minutes on a laptop, without requiring a GPU. This tutorial will show you how to run the example script on your own images, and will explain some of the options you have to help control the training process.

Contents

(4) Develop your own machine learning models

https://www.tensorflow.org/versions/master/get_started/basic_usage.html

TensorFlow™

GET STARTED

Overview

TensorFlow is a programming system in which you represent computations as graphs. Nodes in the graph are called ops (short for operations). An op takes zero or more Tensors, performs some computation, and produces zero or more Tensors. A Tensor is a typed multi-dimensional array. For example, you can represent a mini-batch of images as a 4-D array of floating point numbers with dimensions [batch, height, width, channels].

A TensorFlow graph is a description of computations. To compute anything, a graph must be launched in a Session. A Session places the graph ops onto Devices, such as CPUs or GPUs, and provides methods to execute them. These methods return tensors produced by ops as `numpy ndarray` objects in Python, and as `tensorflow::Tensor` instances in C and C++.

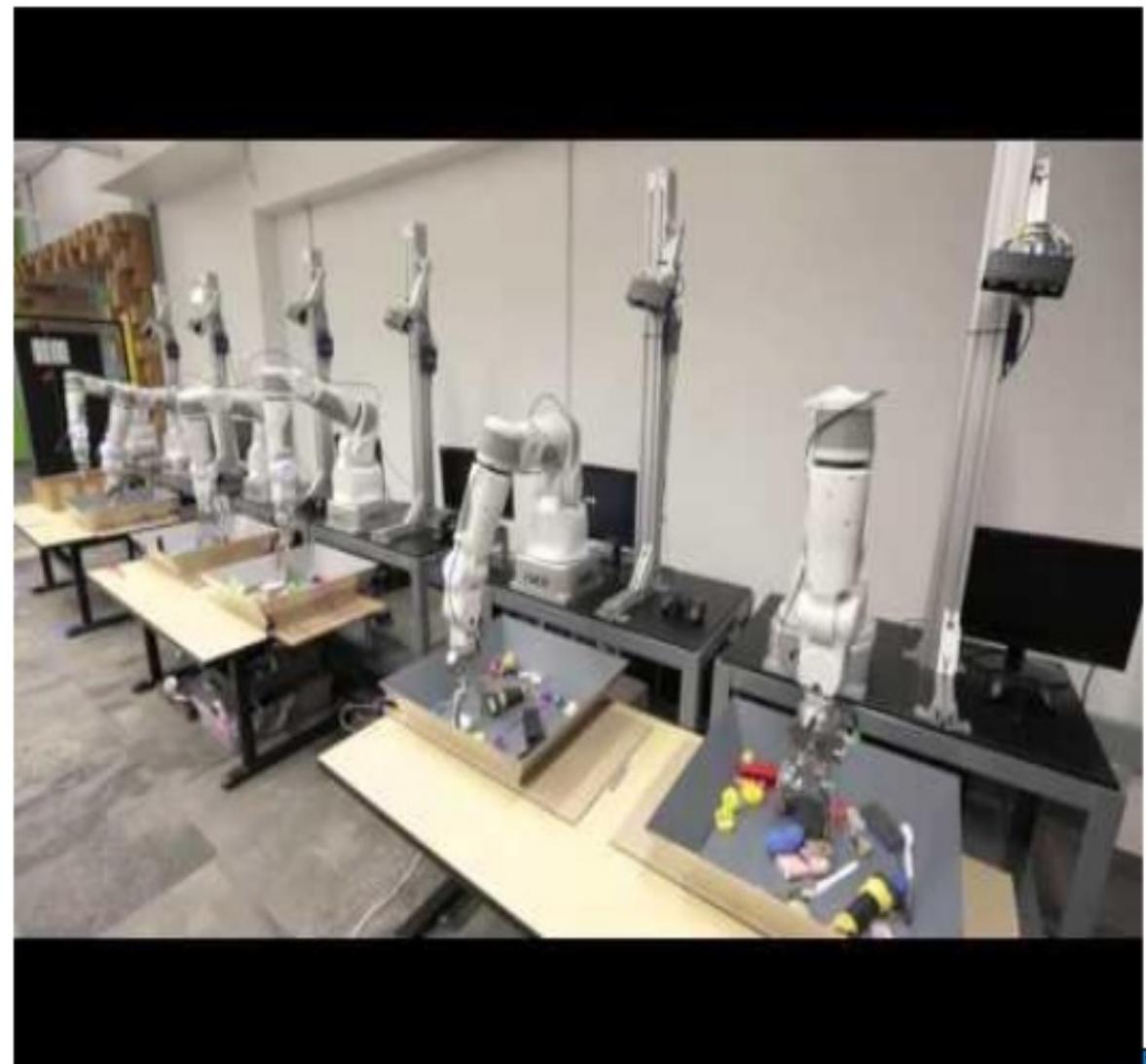
The computation graph

TensorFlow programs are usually structured into a construction phase, that assembles a graph, and an execution

Example: Combining Vision with Robotics

“Deep Learning for Robots: Learning from Large-Scale Interaction”, Google Research Blog, March, 2016

“Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection”,
Sergey Levine, Peter Pastor, Alex Krizhevsky, & Deirdre Quillen,
Arxiv, arxiv.org/abs/1603.02199



<https://www.youtube.com/watch?v=iaF43Ze1oel>



What Does the Future Hold?

Deep learning usage will continue to grow and accelerate:

- Across more and more fields and problems:
 - robotics, self-driving vehicles, ...
 - health care
 - video understanding
 - dialogue systems
 - personal assistance
 - ...



Conclusions

Deep neural networks are making significant strides in understanding:

In speech, vision, language, search, ...

If you're not considering how to use deep neural nets to solve your vision or understanding problems, **you almost certainly should be**



Further Reading

- Dean, et al., *Large Scale Distributed Deep Networks*, NIPS 2012, research.google.com/archive/large_deep_networks_nips2012.html.
- TensorFlow white paper, tensorflow.org/whitepaper2015.pdf (clickable links in bibliography)
- TensorFlow: A System for Large-Scale Machine Learning, <http://arxiv.org/abs/1605.08695>

tensorflow.org

research.google.com/people/jeff

We're Hiring! See g.co/brain

