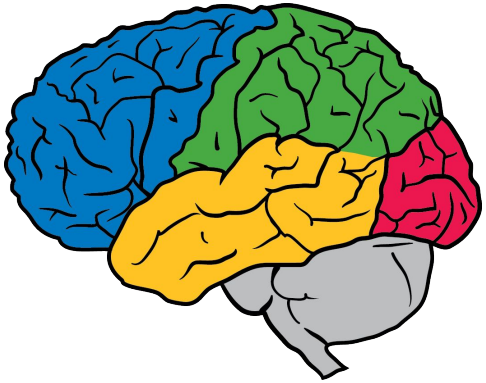


Large-Scale Deep Learning With TensorFlow

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

Google™

MapReduce

Spark

Google Cloud BigTable

We can now store and perform computation on large datasets

Google™ MapReduce

Spark

Google Cloud BigTable

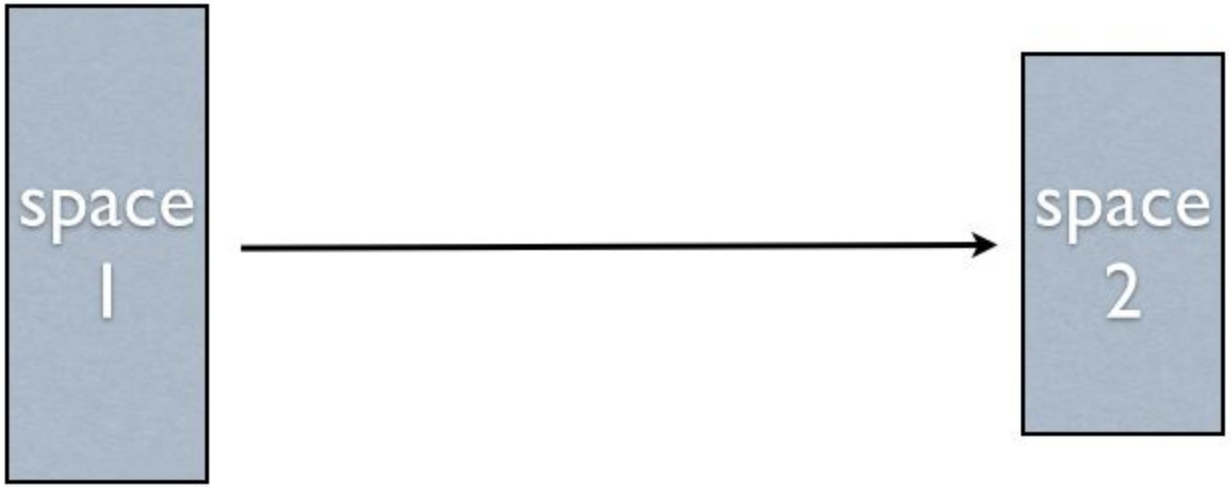
But what we really want is not just raw data,
but computer systems that **understand** this data

What do I mean by understanding?

What do I mean by understanding?

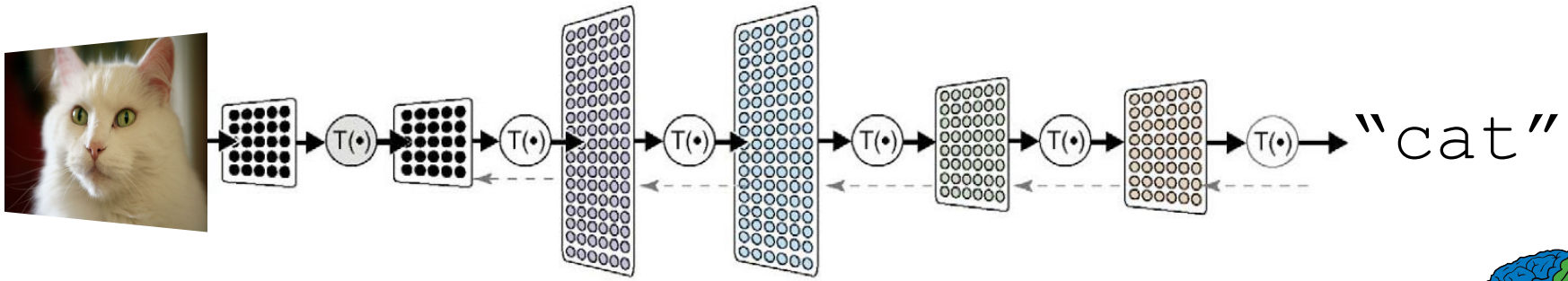
Neural Networks

- Learn a complicated function from data



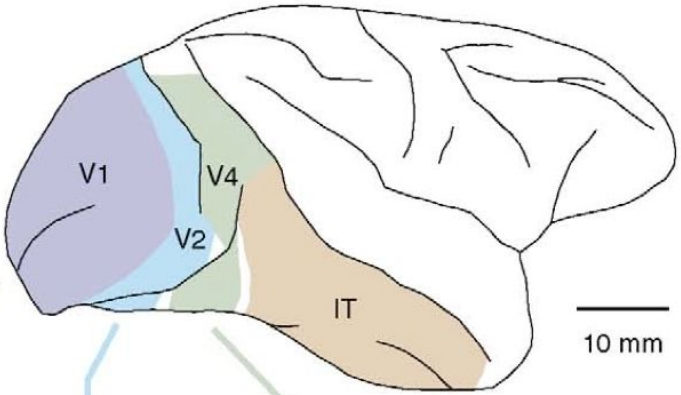
What is Deep Learning?

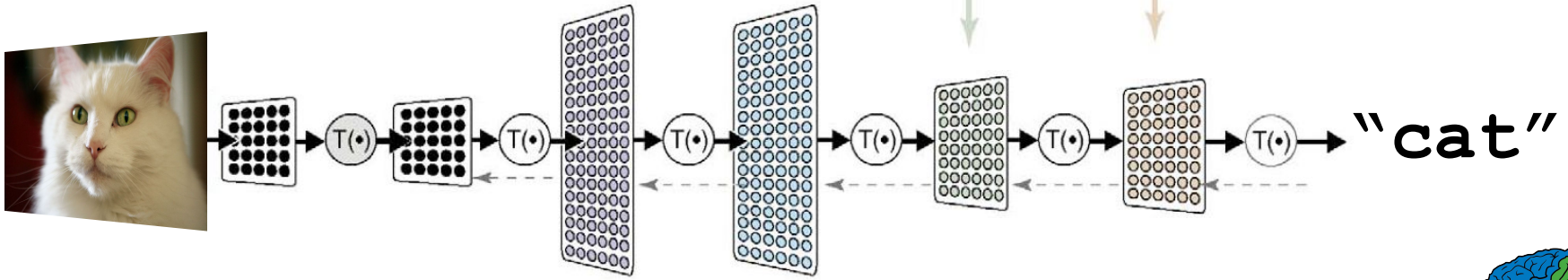
- A powerful class of machine learning model
- Modern reincarnation of artificial neural networks
- Collection of simple, trainable mathematical functions



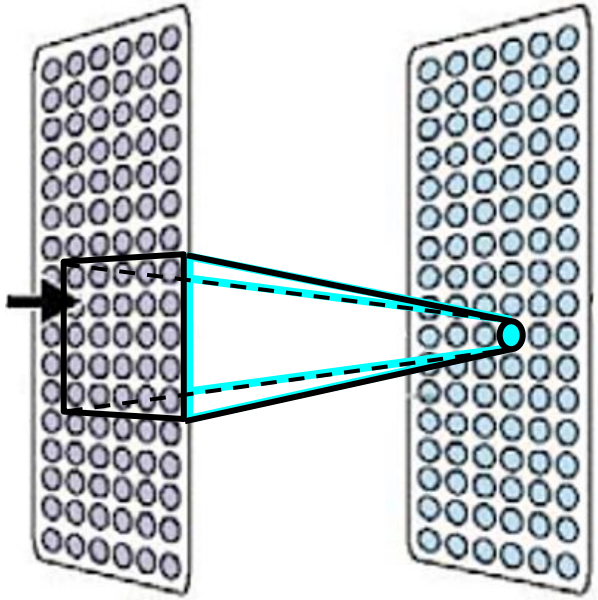
What is Deep Learning?

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





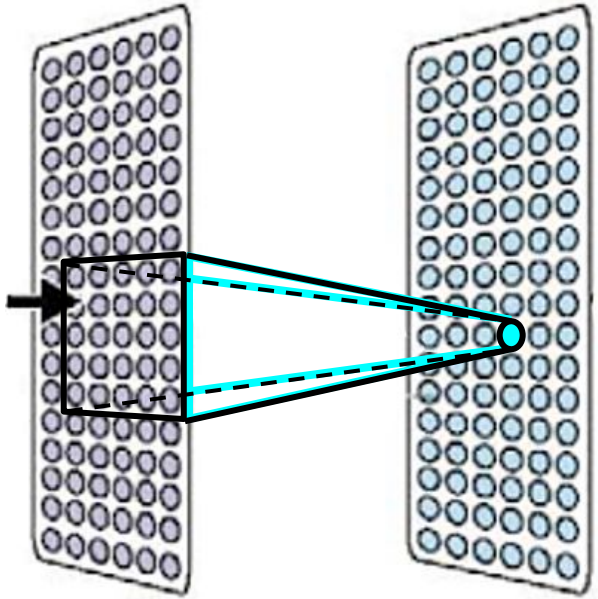
What is Deep Learning?



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

OUTPUT LAYER

ACTIVATED NEURONS

INPUT LAYER

DEEP *neural* NETWORK

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!)

Aside

Many of the techniques that are successful now were developed 20-30 years ago

What changed? We now have:

sufficient computational resources
large enough interesting datasets

Use of large-scale parallelism lets us look ahead many generations of hardware improvements, as well

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

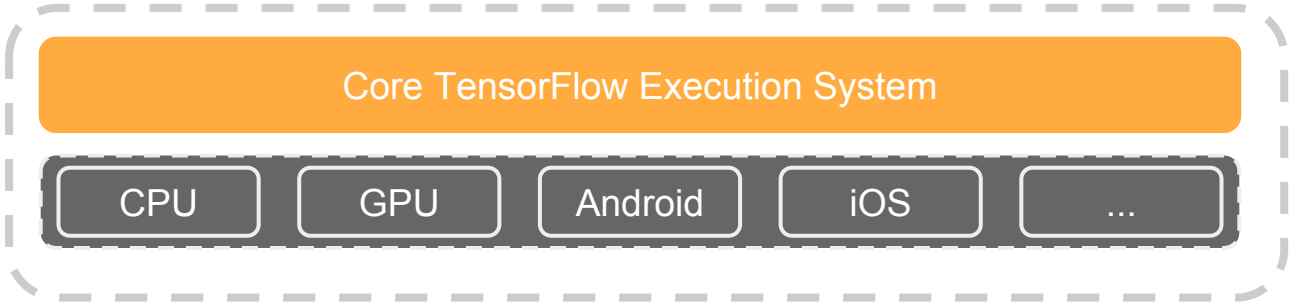
Apache 2.0 license

Motivations

- DistBelief (our 1st system) was the first scalable deep learning system, but not as flexible as we wanted for research purposes
- Better understanding of problem space allowed us to make some dramatic simplifications
- Define the industrial standard for machine learning
- Short circuit the MapReduce/Hadoop inefficiency

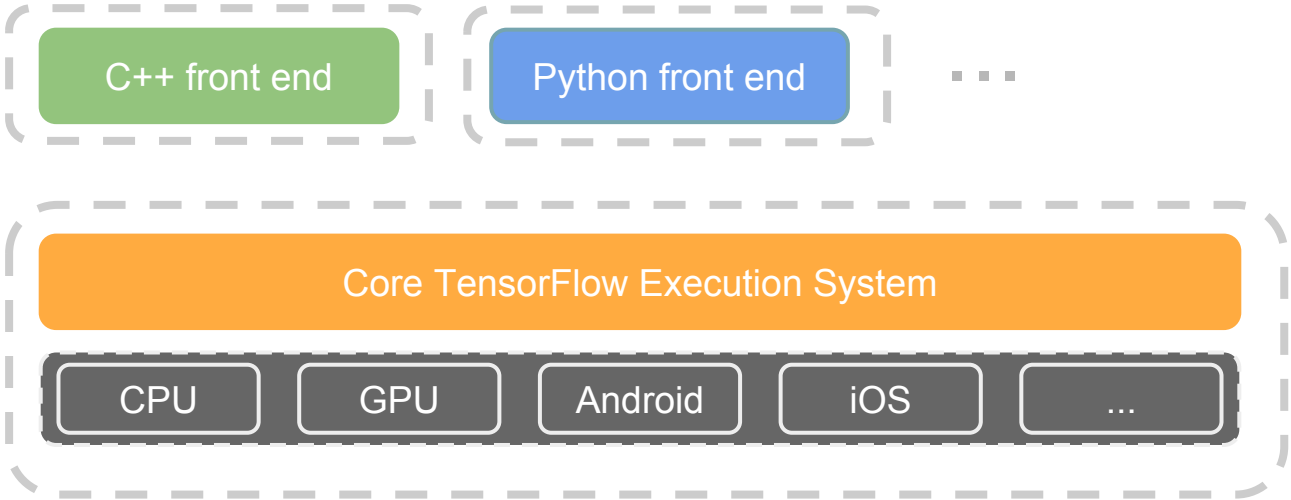
TensorFlow: Expressing High-Level ML Computations

- Core in C++
  - Very low overhead

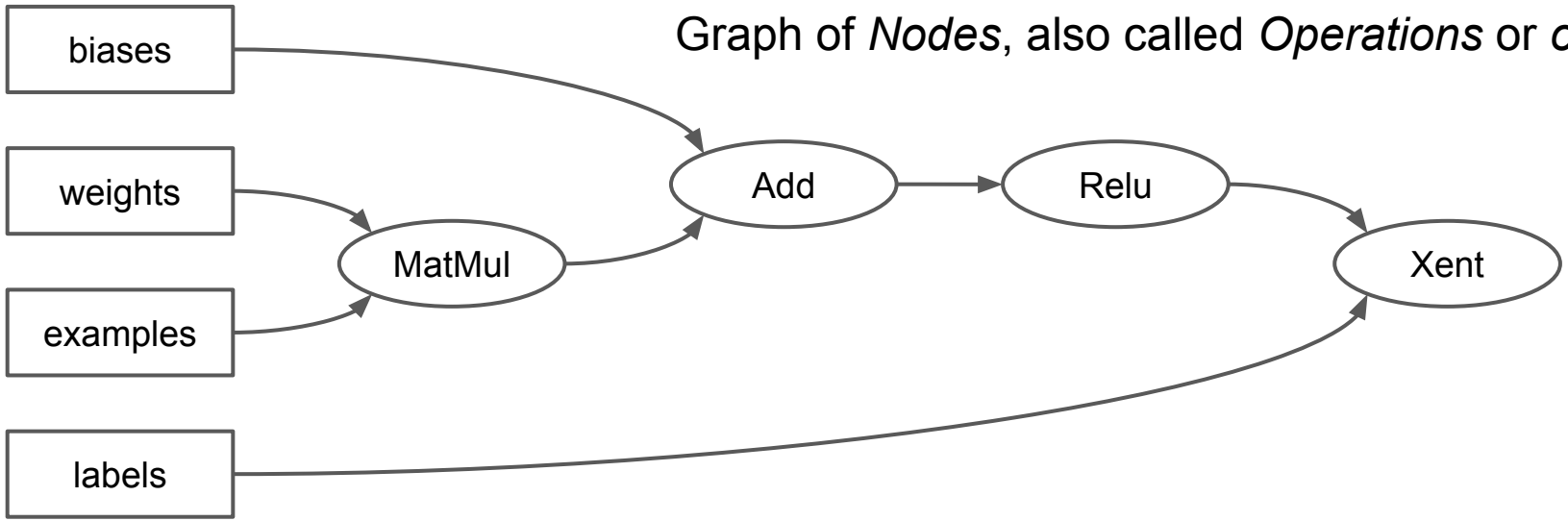


TensorFlow: Expressing High-Level ML Computations

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



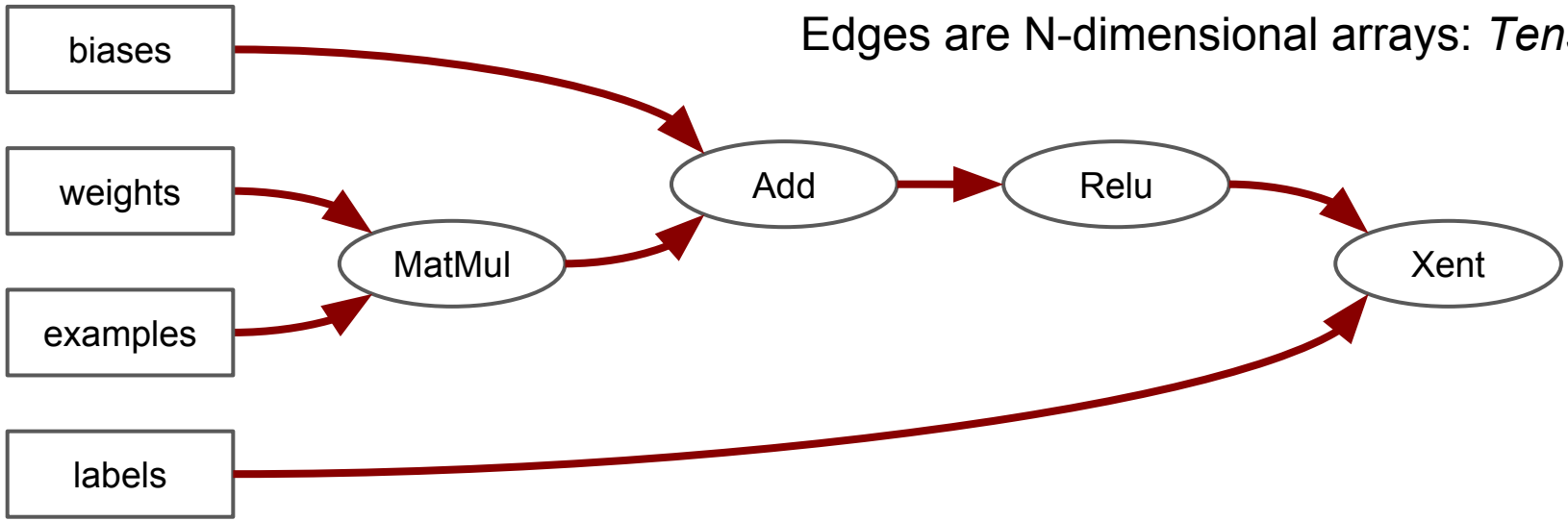
Computation is a dataflow graph



Graph of *Nodes*, also called *Operations* or *ops.*

Computation is a dataflow graph

with tensors



Edges are N-dimensional arrays: *Tensors*

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)
```

Symbolic Differentiation

- 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)
```

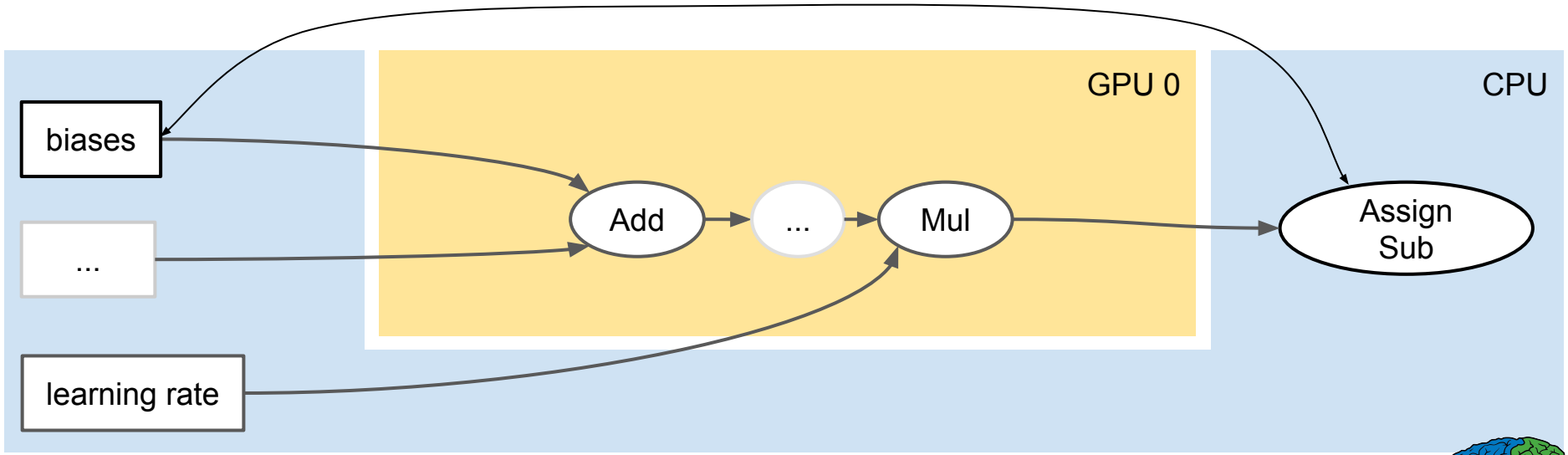
Define graph and then execute it repeatedly

- 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})
```

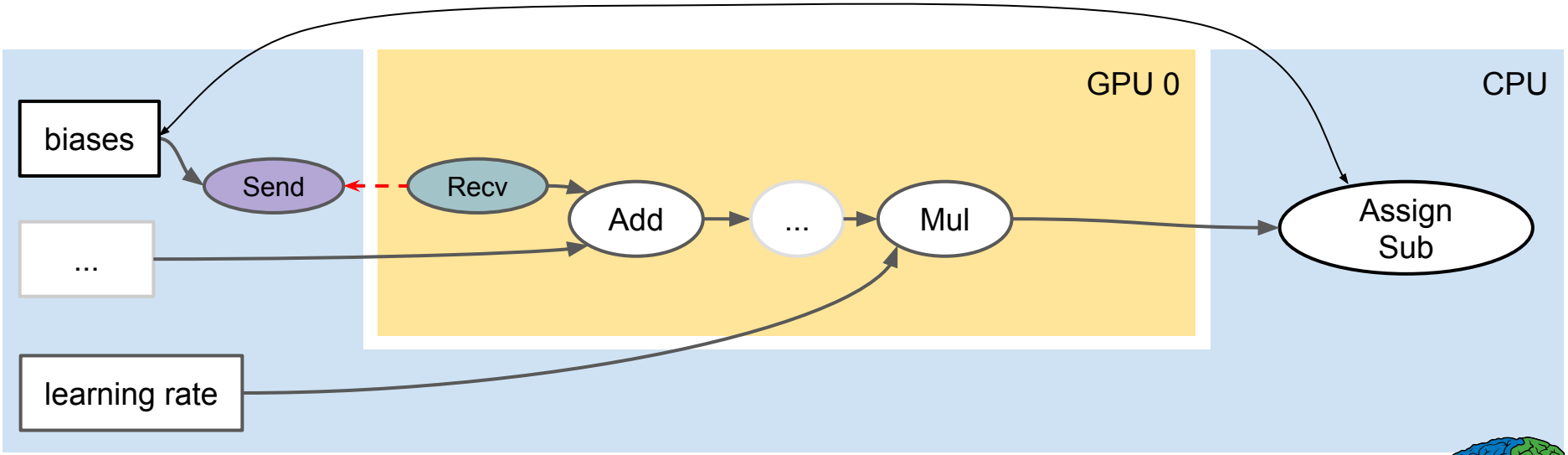
Computation is a dataflow graph

distributed



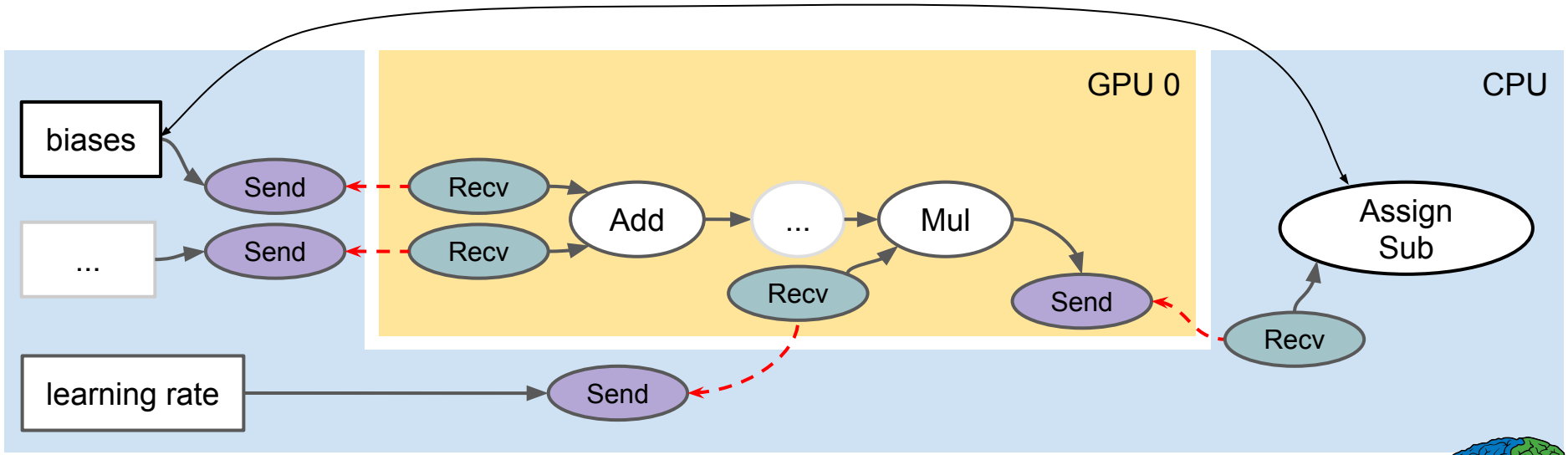
Assign *Devices* to Ops

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



Assign *Devices* to Ops

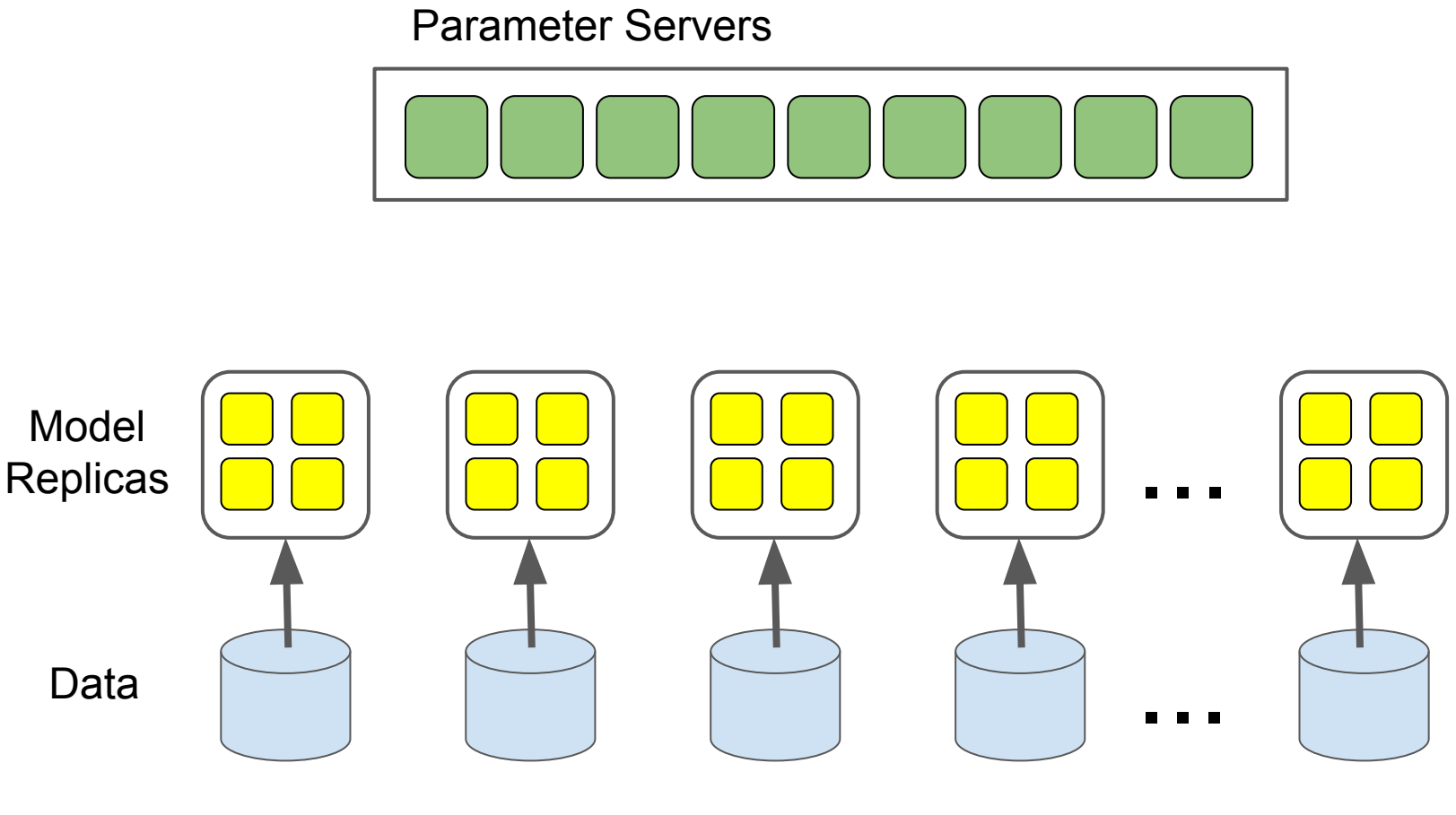
- TensorFlow inserts *Send/Recv* Ops to transport tensors across devices
- *Recv* ops pull data from *Send* ops



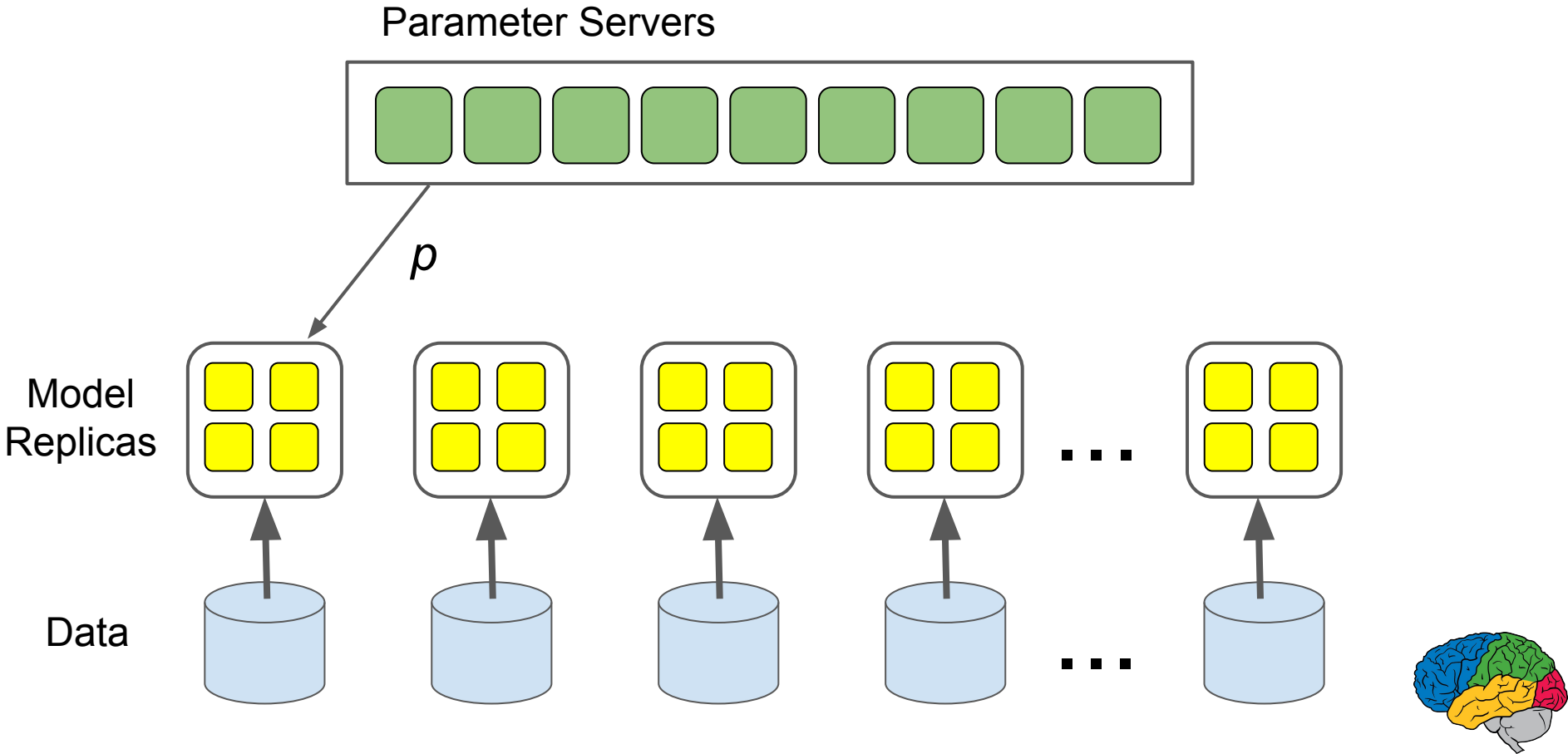
Experiment Turnaround Time and Research Productivity

- **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

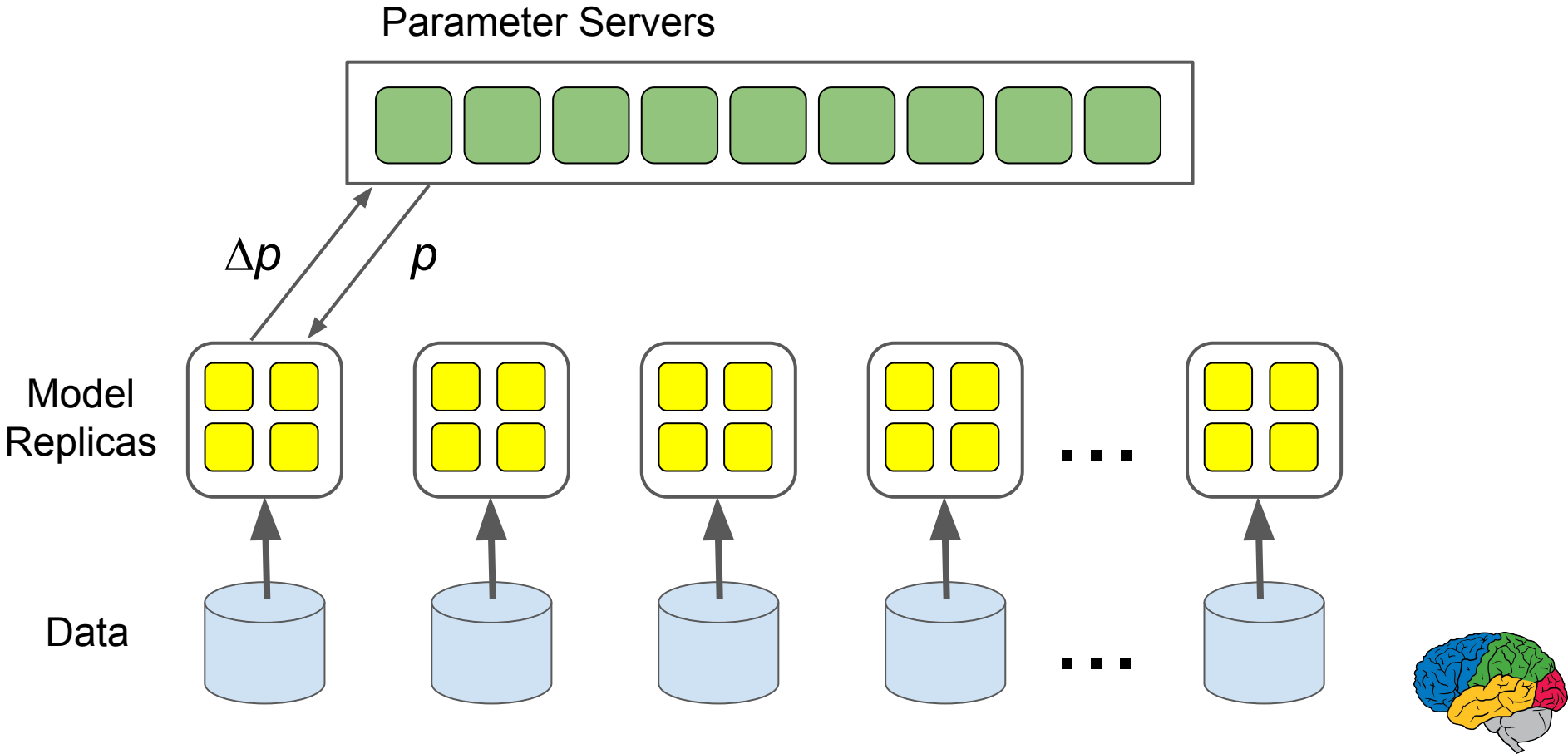
Data Parallelism



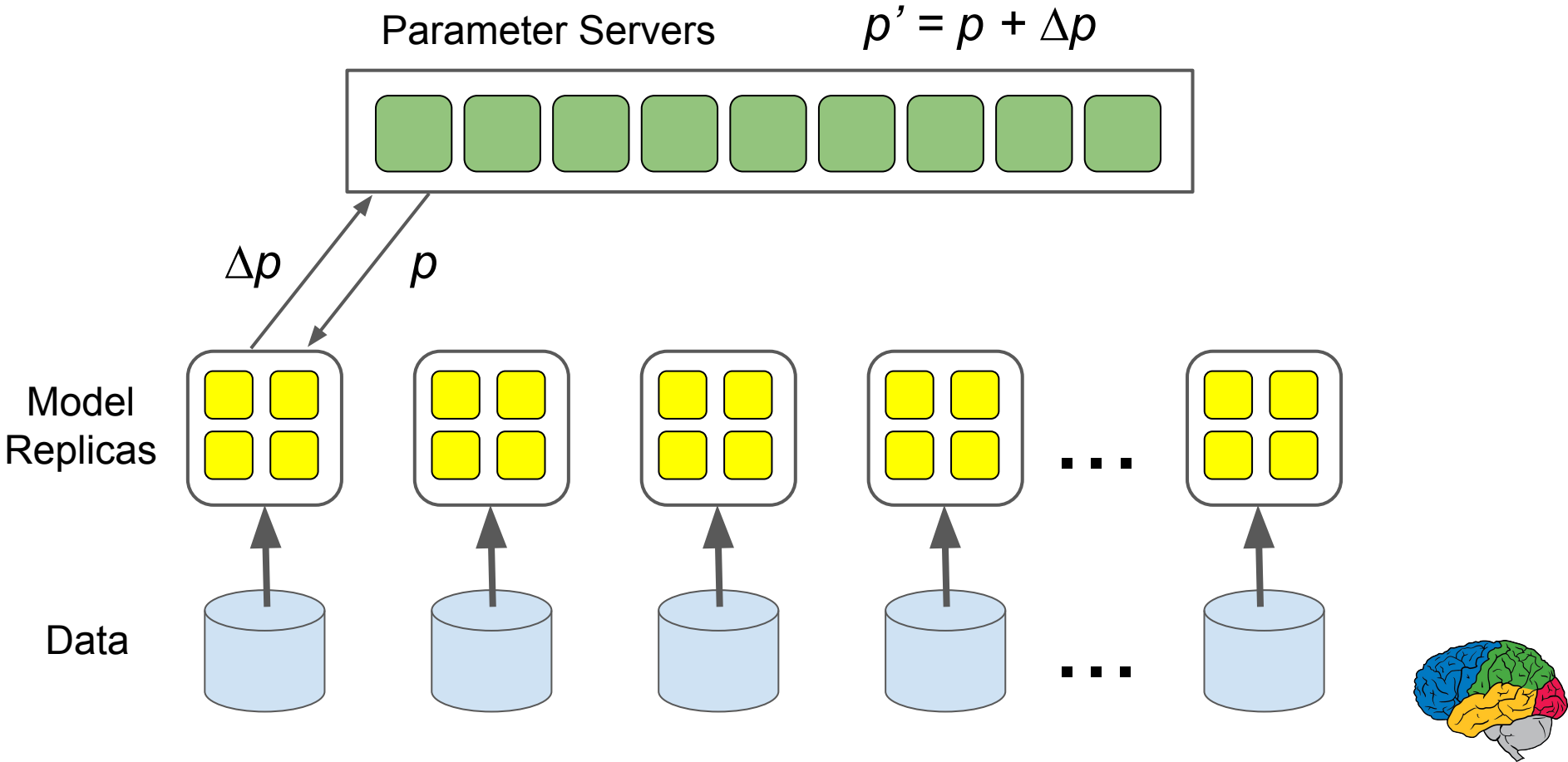
Data Parallelism



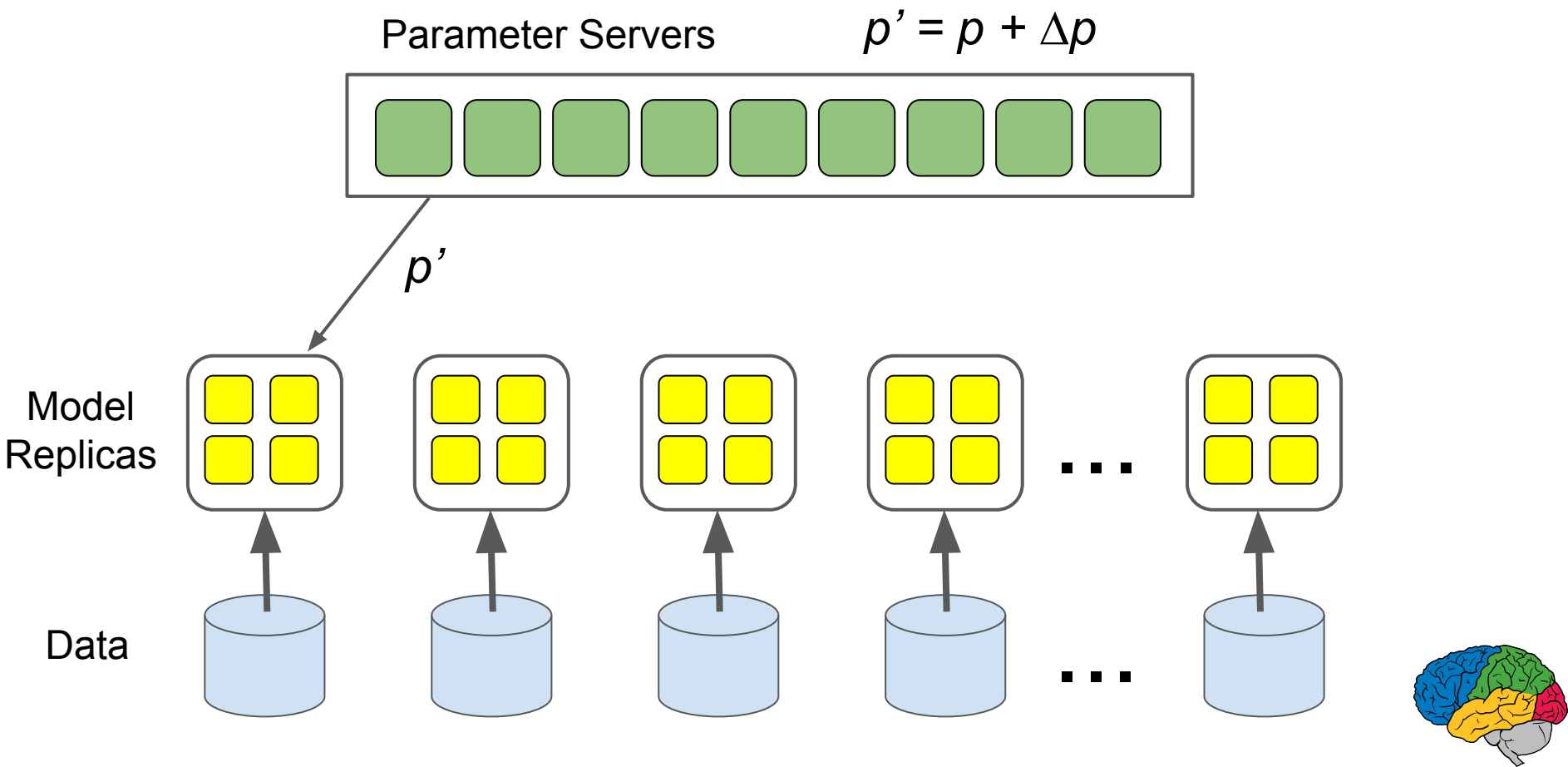
Data Parallelism



Data Parallelism



Data Parallelism



General Computations

Although we originally built TensorFlow for our uses around deep neural networks, it’s actually quite flexible

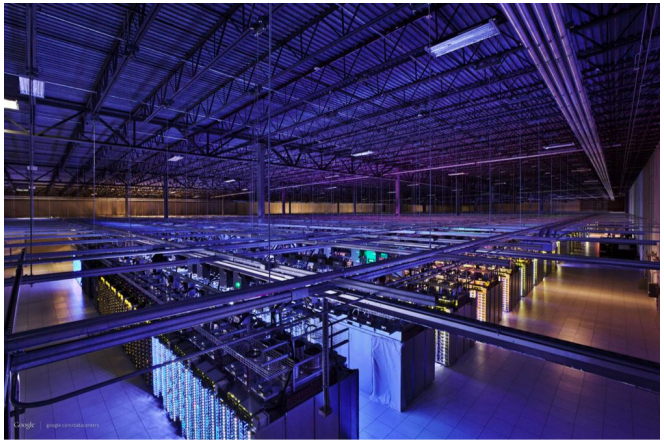
Wide variety of machine learning and other kinds of numeric computations easily expressible in the computation graph model

Runs on Variety of Platforms

phones

single machines (CPU and/or GPUs) …

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



custom ML hardware

Fathom

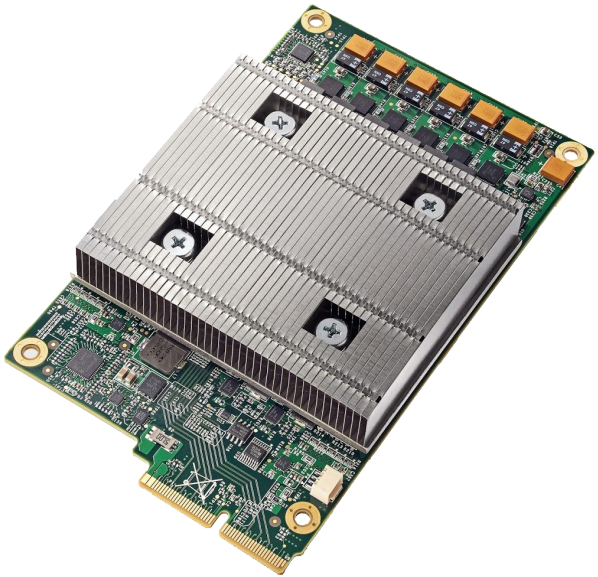
Trend: Much More Heterogeneous hardware

General purpose CPU performance scaling has slowed significantly

Specialization of hardware for certain workloads will be more important

Tensor Processing Unit

Custom machine learning ASIC



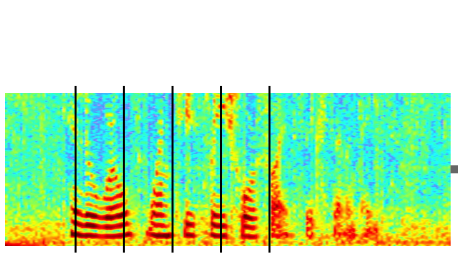
In production use for >16 months: used on every search query, used for AlphaGo match, ...

See Google Cloud Platform blog: Google supercharges machine learning tasks with TPU custom chip, by Norm Jouppi, May, 2016

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

All of these examples implemented using TensorFlow or our predecessor system

Speech Recognition



Acoustic Input

Deep
Recurrent
Neural Network

"How cold is
it outside?"

Text Output

Reduced word errors by more than 30%

Google Research Blog - August 2012, August 2015

Google Photos Search

Your Photo

Deep Convolutional Neural Network

"ocean"

Automatic Tag

Search personal photos without tags.

Google Research Blog - June 2013

Image Captions Research

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

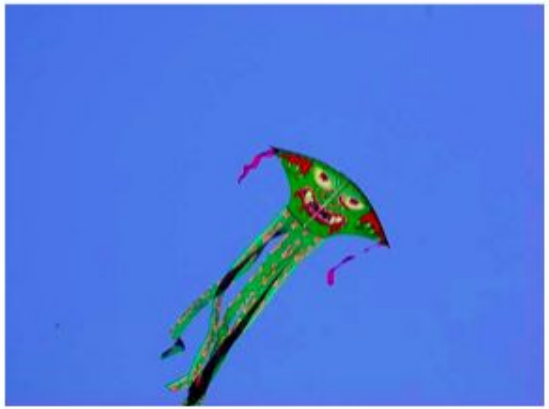
Model: A close up of a child holding a stuffed animal.

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

A man holding a tennis racquet on a tennis court.

Two pizzas sitting on top of a stove top oven

A group of young people playing a game of Frisbee



A man flying through the air while riding a snowboard

Deep neural networks are making significant strides in understanding:
In speech, vision, language, search, robotics, …

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

Conclusion

- TensorFlow is a general, flexible system for Large-scale machine learning.
- New problems introduced by TensorFlow
  - Model placement
  - High speed network
- Machine learning in network control and troubleshooting

Thanks