# Deep Learning with Python & TensorFlow

PyCon SG 2016

#pyconsg

# Ian Lewis

## Developer Advocate - Google Cloud Platform

Tokyo, Japan

+Ian Lewis
@IanMLewis

# PyCon JP 2016

Everyone's different, all are wonderful.

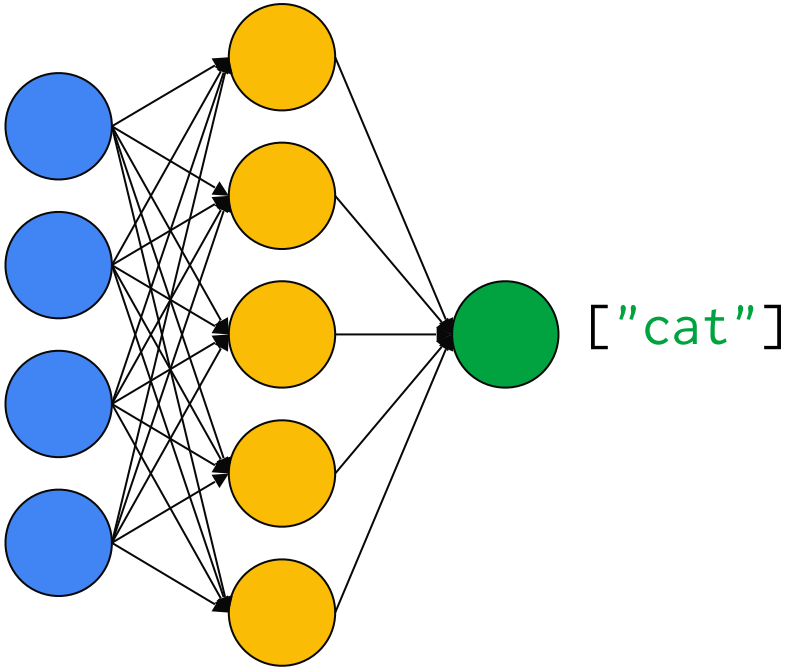

python™

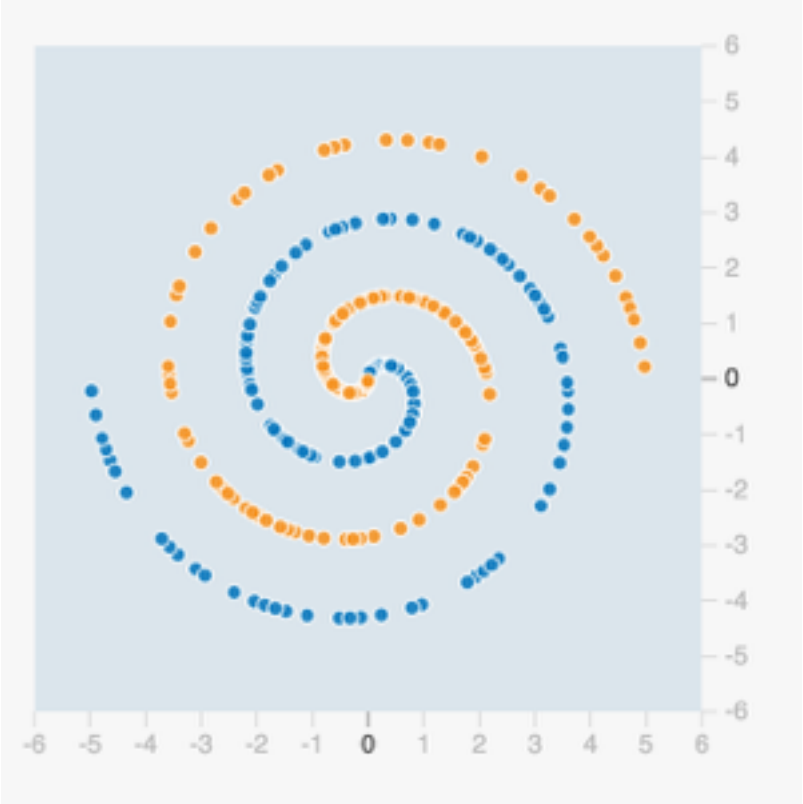

GO!
GO!

# Deep Learning 101

Input Hidden Output(label)

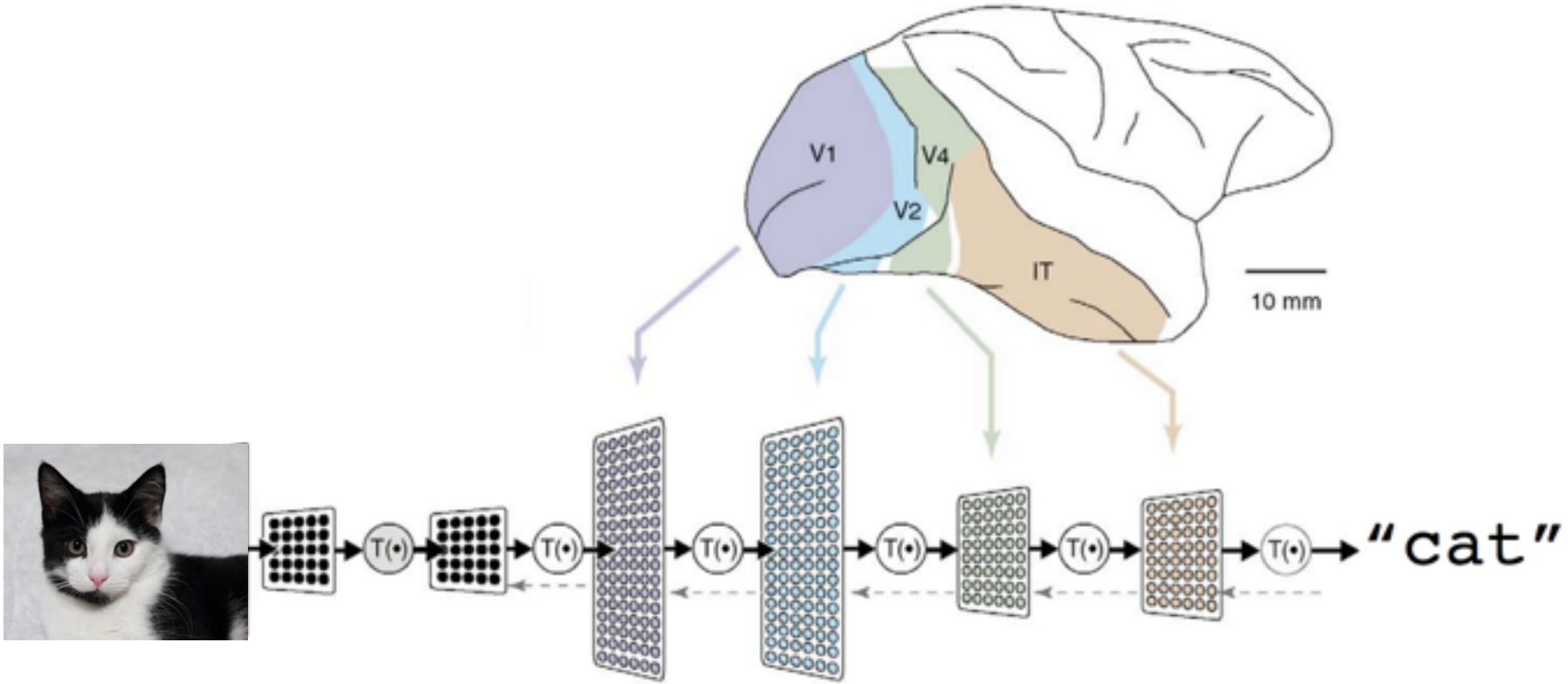
pixels(  )

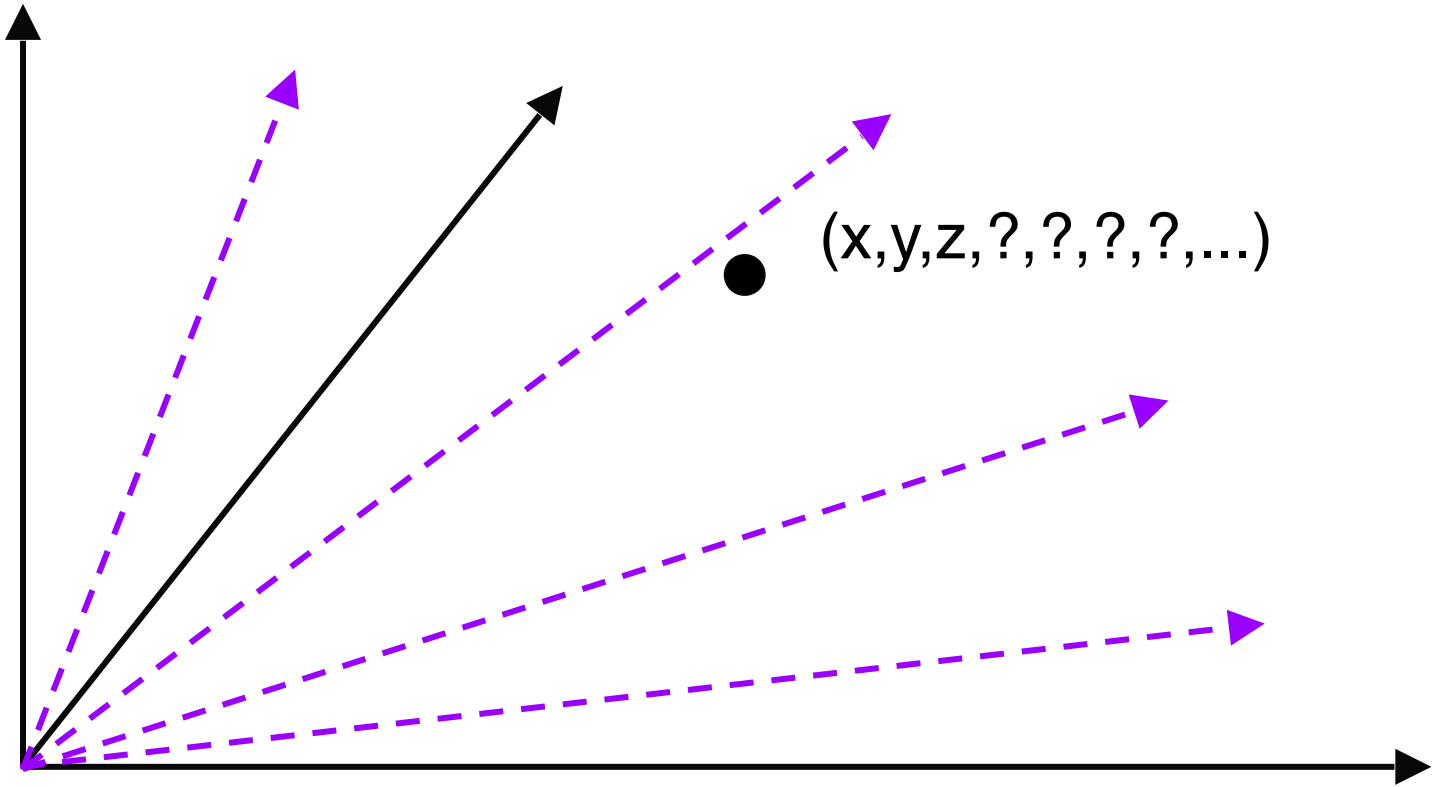





How do you classify these data points?

Neural Network can **find a way** to solve the problem

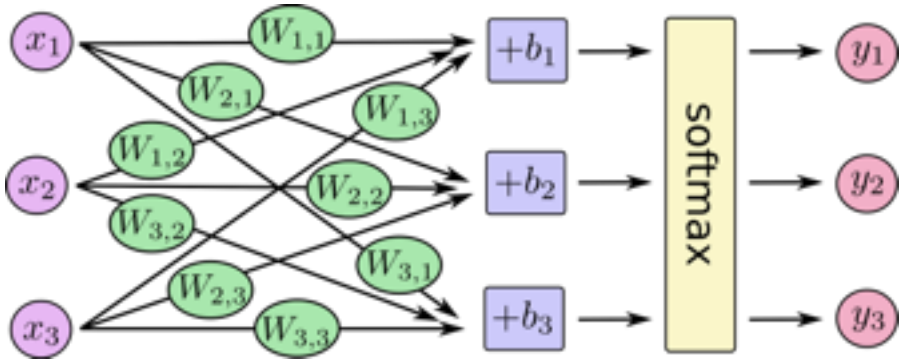

V1
V4
V2
IT
10 mm
T(•)
T(•)
T(•)
T(•)
T(•)
T(•)
"cat"


(x,y,z,?,?,?,?,...)

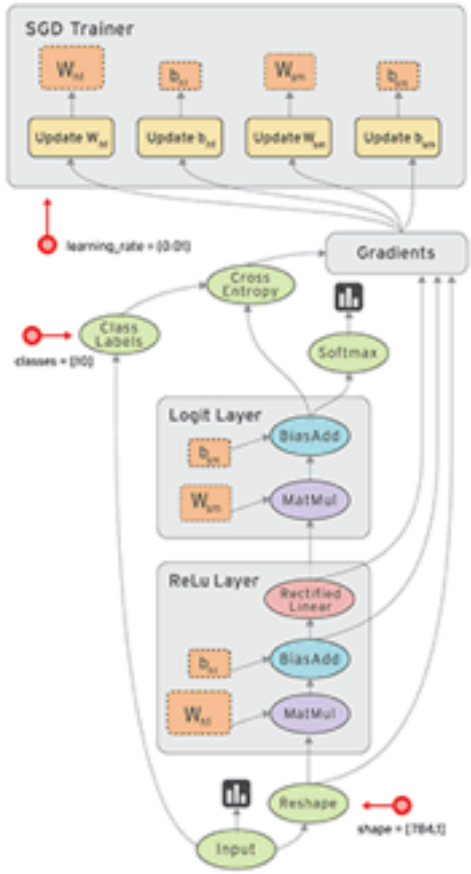
v[x] => vector

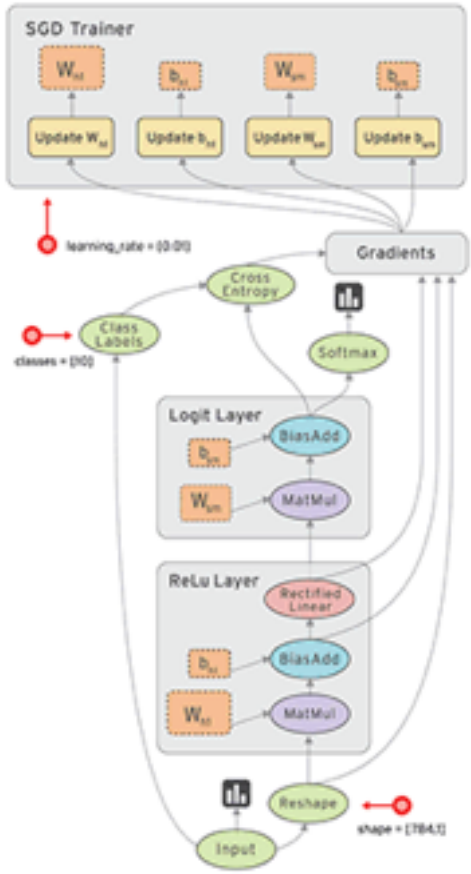
m[x][y][z] => matrix

t[x][y][z][?][?]... => tensor

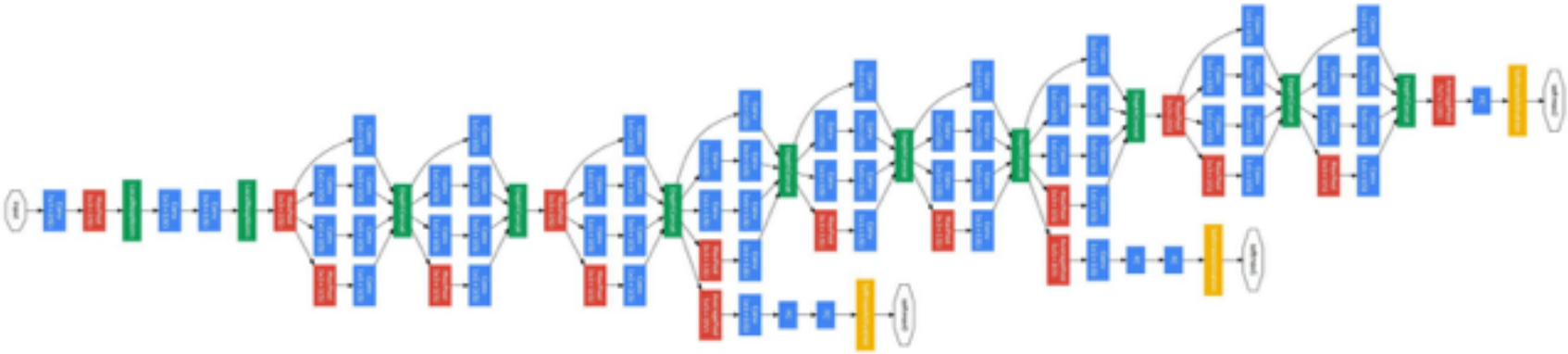



$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \text{softmax} \begin{pmatrix} W_{1,1}x_1 + W_{1,2}x_2 + W_{1,3}x_3 + b_1 \\ W_{2,1}x_1 + W_{2,2}x_2 + W_{2,3}x_3 + b_2 \\ W_{3,1}x_1 + W_{3,2}x_2 + W_{3,3}x_3 + b_3 \end{pmatrix}$$

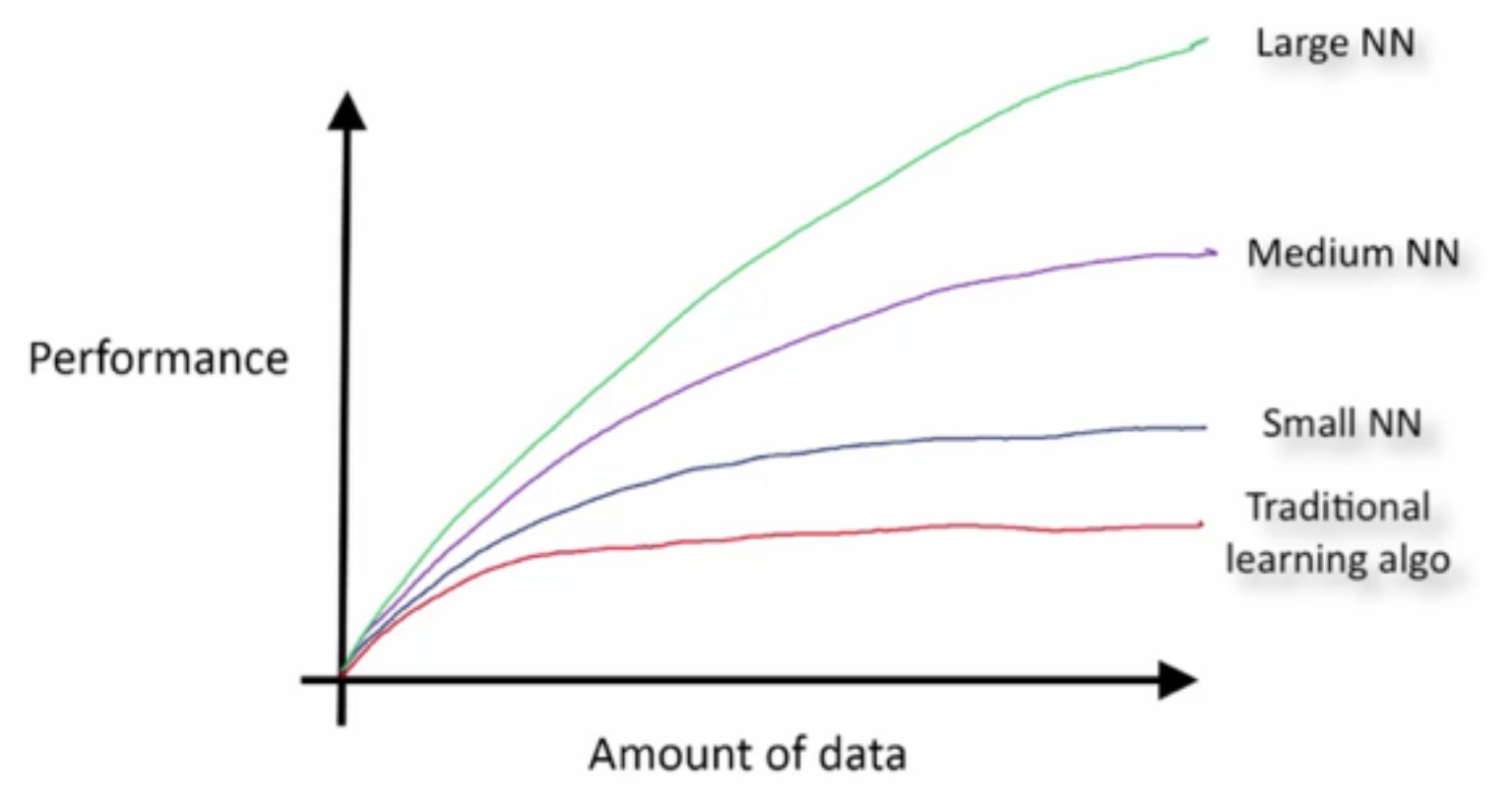

SGD Trainer
Update W_sl
Update b_sl
Update W_sm
Update b_sm
learning_rate = [0.01]
Gradients
Cross Entropy
Class Labels
classes = [10]
Softmax
Logit Layer
BiasAdd
MatMul
ReLu Layer
Rectified Linear
BiasAdd
MatMul
Reshape
shape = [784,1]
Input


SGD Trainer
Update W_sl
Update b_sl
Update W_sm
Update b_sm
learning_rate = [0.01]
Gradients
Cross Entropy
Class Labels
classes = [10]
Softmax
Logit Layer
BiasAdd
MatMul
ReLu Layer
Rectified Linear
BiasAdd
MatMul
Reshape
shape = [784,1]
Input

# Breakthroughs



# The Inception model (GoogLeNet, 2015)




From: Andrew Ng


DNN = a large **matrix** ops

a few GPUs >> CPU

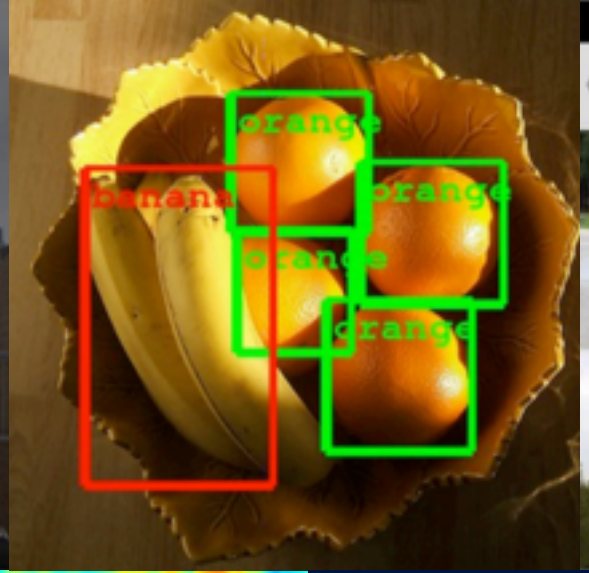
(but it still takes **hours/days** to train)

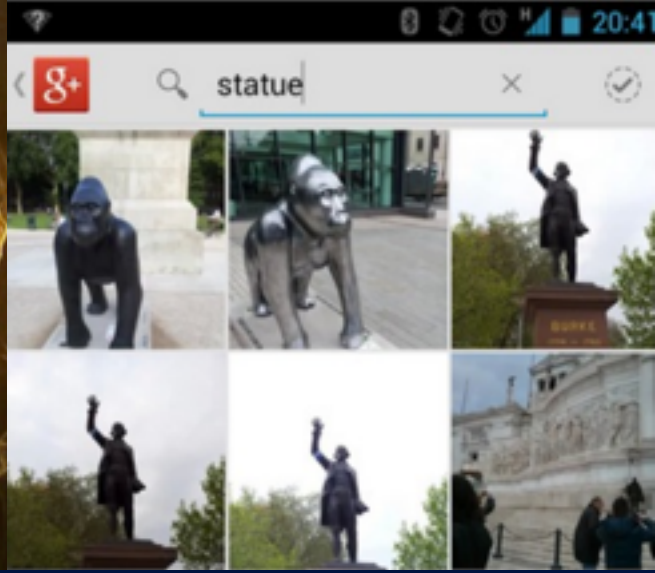
a supercomputer >> a few GPUs

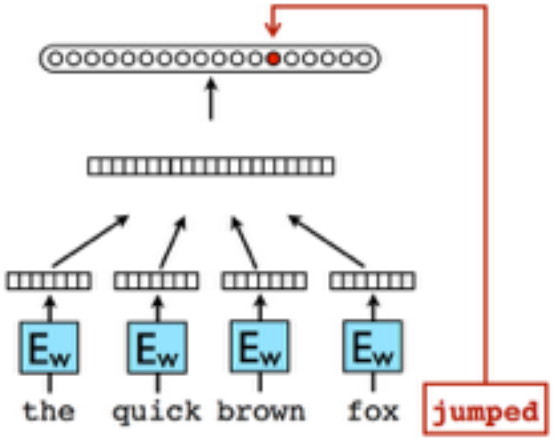
(but you don't have a supercomputer)

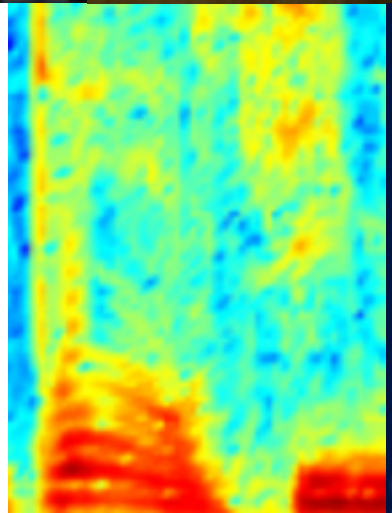
You need **Distributed Training**


CIANO MOTOR ENGINEERS
BMW, MINI & TOYOTA


orange
banana
orange
orange
orange

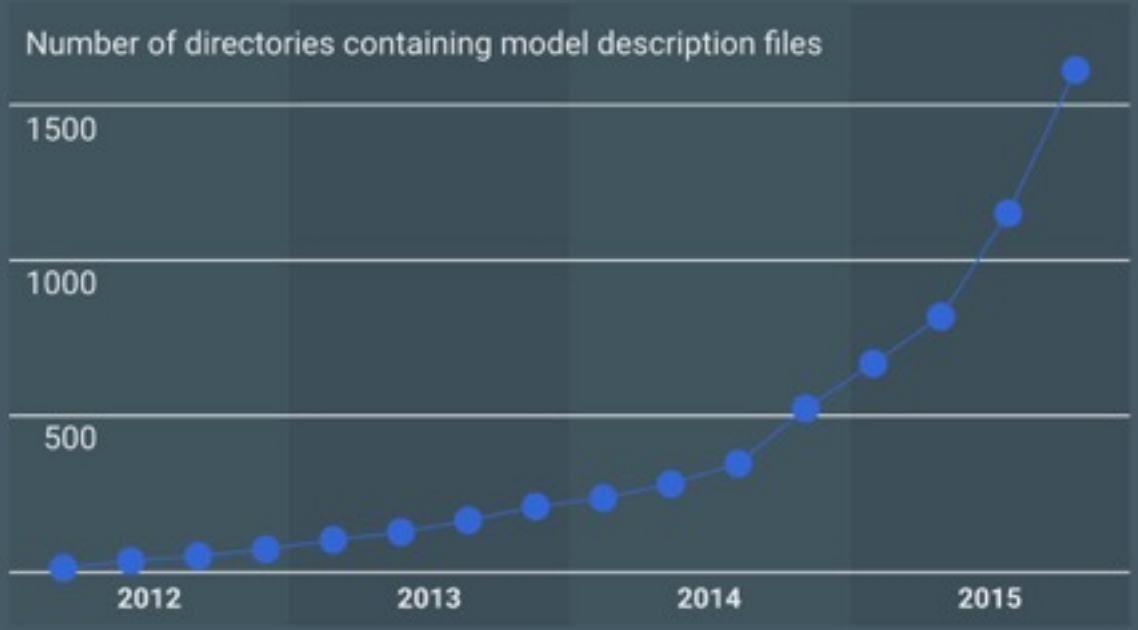

20:41
statue


Ew
Ew
Ew
Ew
the
quick
brown
fox
jumped



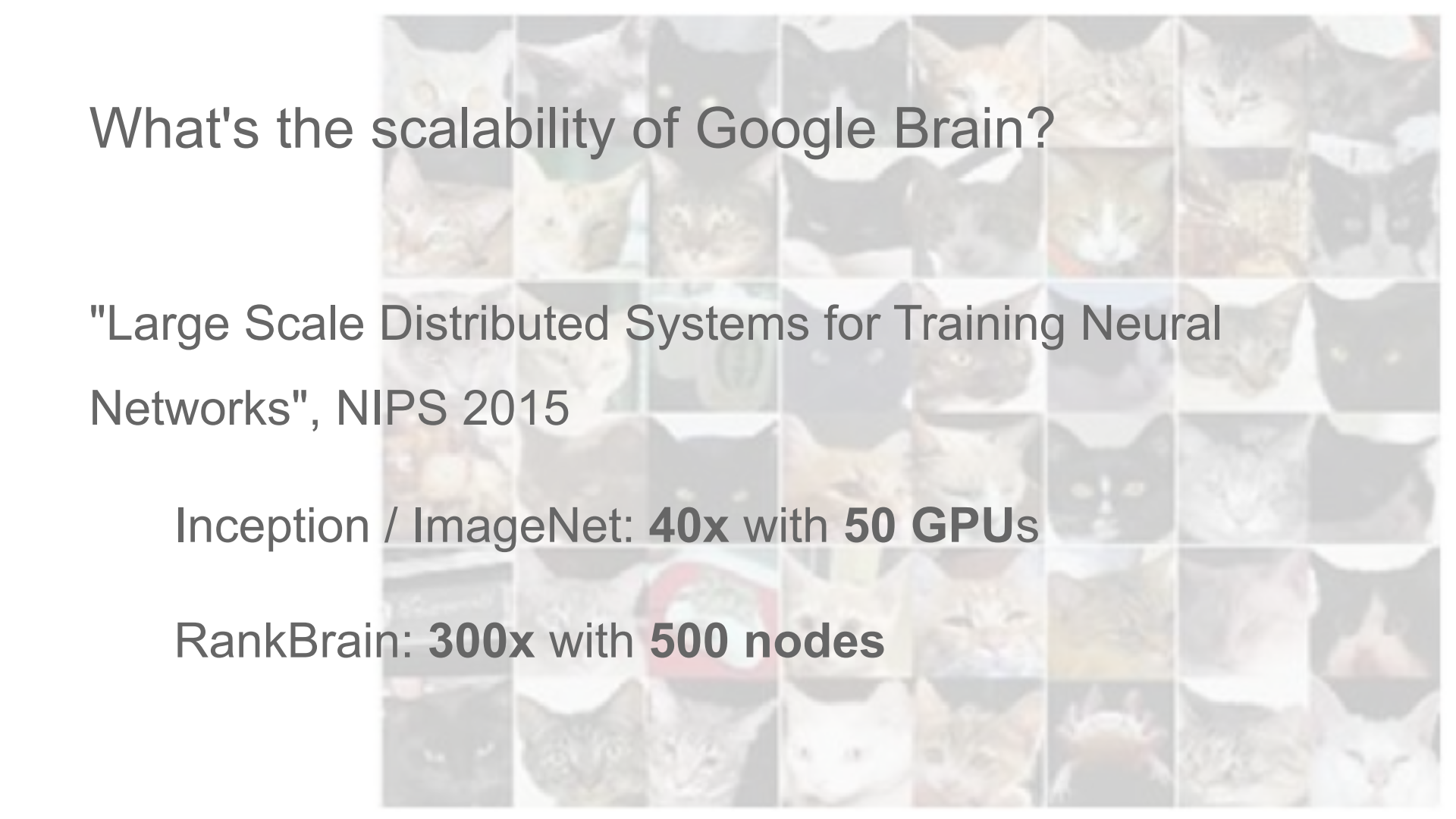

KBA
Google DeepMind
Challenge Match
AlphaGo
ALPHAGO
00:05:30
LEE SEDOL
00:28:28

# Growing use of deep learning at Google




Across many areas

- AlphaGo
- Apps
- Maps
- Photos
- Gmail
- Speech
- Android
- YouTube
- Translation
- Robotics Research
- Image Understanding
- Natural Language Understanding
- Drug Discovery



# What's the scalability of Google Brain?

"Large Scale Distributed Systems for Training Neural Networks", NIPS 2015

- Inception / ImageNet: **40x** with **50 GPU**s
- RankBrain: **300x** with **500 nodes**

# TensorFlow

# What is Tensorflow?

Google's **open source** library for machine intelligence

**tensorflow.org** launched in Nov 2015

The second generation

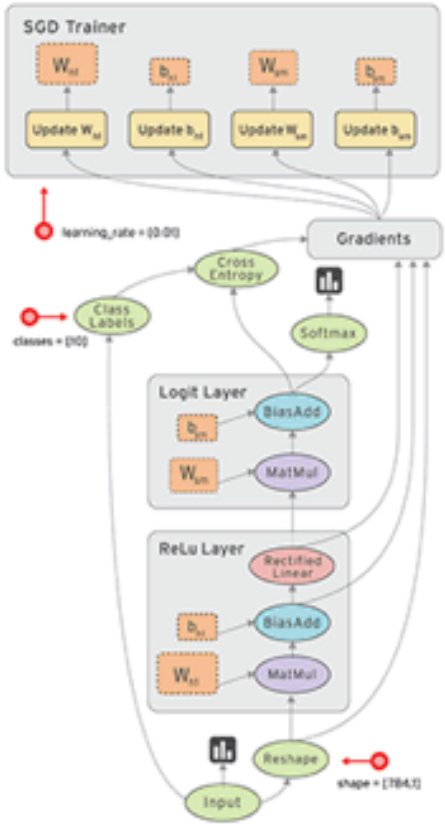
Used by many production ML projects

# TensorFlow

Operates over **tensors**: *n-dimensional arrays*

Using a **flow graph**: *data flow computation framework*

- Flexible, intuitive construction
- automatic differentiation
- Support for threads, queues, and asynchronous computation; distributed runtime
- Train on CPUs, GPUs
- Run wherever you like

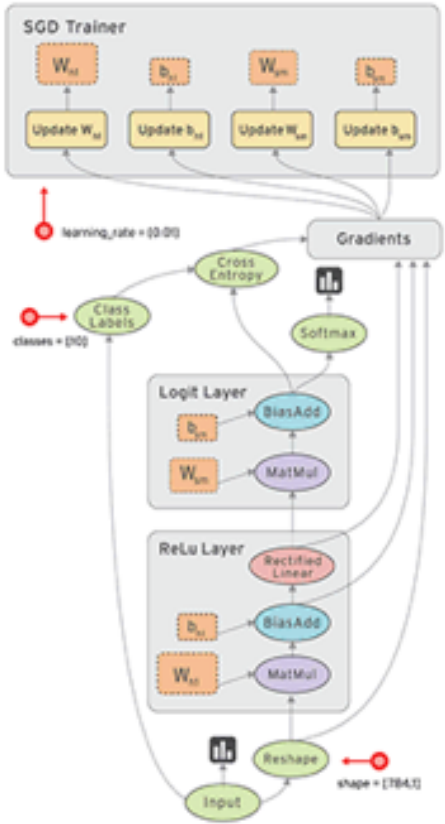


# TensorFlow

Operates over **tensors**: *n-dimensional arrays*

Using a **flow graph**: *data flow computation framework*

- Flexible, intuitive construction
- automatic differentiation
- Support for threads, queues, and asynchronous computation; distributed runtime
- Train on CPUs, GPUs
- Run wherever you like



# Core TensorFlow data structures and concepts...

- **Graph**: A TensorFlow computation, represented as a dataflow graph.
  - collection of ops that may be executed together as a group
- **Operation**: a graph node that performs computation on tensors
- **Tensor**: a handle to one of the outputs of an Operation
  - provides a means of computing the value in a TensorFlow Session.

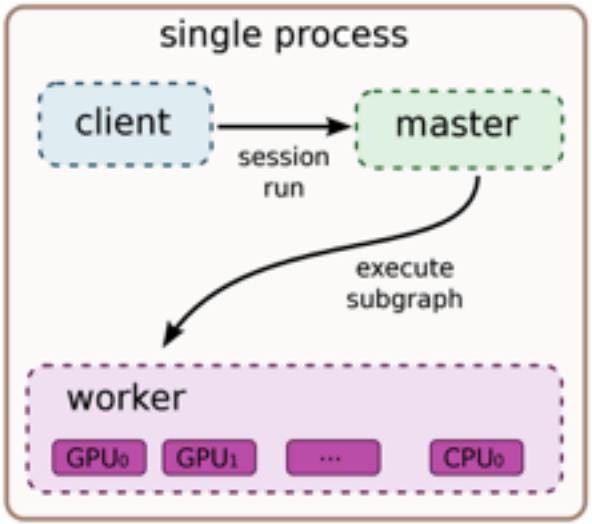
- **Constants**
- **Placeholders**: must be fed with data on execution
- **Variables**: a modifiable tensor that lives in TensorFlow's graph of interacting operations.
- **Session**: encapsulates the environment in which Operation objects are executed, and Tensor objects are evaluated.

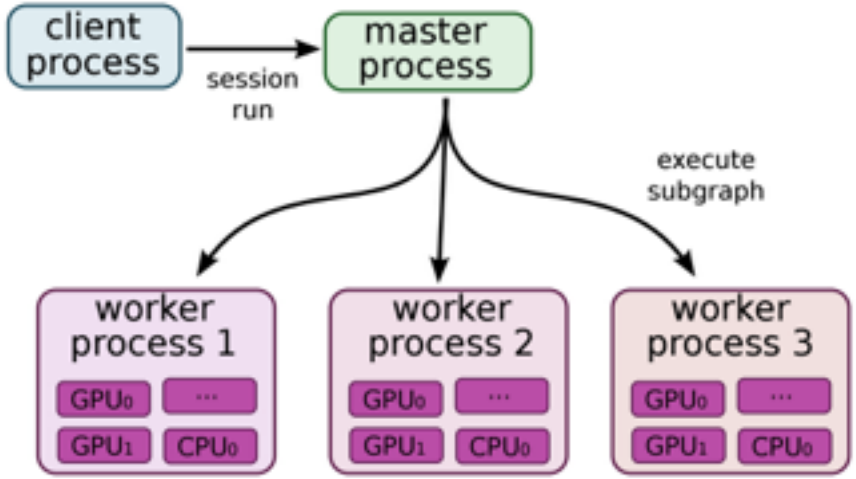
# Operations

| *Category* | *Examples* |
| --- | --- |
| Element-wise math ops | **Add**, Sub, **Mul**, Div, Exp, Log, Greater, Less… |
| Array ops | **Concat**, Slice, **Split**, Constant, Rank, **Shape…** |
| Matrix ops | **MatMul**, MatrixInverse, MatrixDeterminant… |
| Stateful ops | **Variable**, Assign, AssignAdd... |
| NN building blocks | **SoftMax**, Sigmoid, **ReLU**, **Convolution2D…** |
| Checkpointing ops | **Save**, **Restore** |
| Queue & synch ops | Enqueue, Dequeue, MutexAcquire… |
| Control flow ops | Merge, Switch, Enter, Leave... |


jupyter

# Distributed Training with TensorFlow



# Distributed Training with TensorFlow

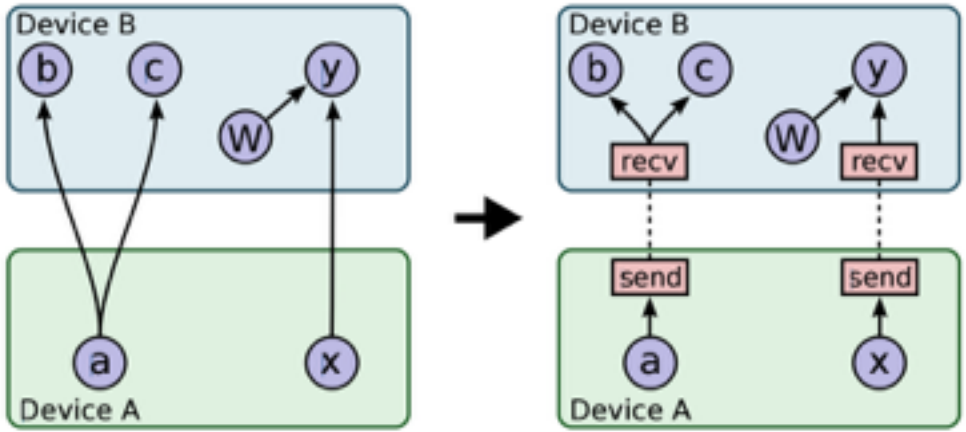
CPU/GPU scheduling

Communications

- Local, RPC, RDMA
- 32/16/8 bit quantization

Cost-based optimization

Fault tolerance



# Distributed Training

| Model Parallelism | |
|---|---|
| Sub-Graph | ● Allows fine grained application of parallelism to slow graph components<br>● Larger more complex graph |
| Full Graph | ● Code is more similar to single process models<br>● Not necessarily as performant (large models) |

| Data Parallelism | |
|---|---|
| Synchronous | ● Prevents workers from “Falling behind”<br>● Workers progress at the speed of the slowest worker |
| Asynchronous | ● Workers advance as fast as they can<br>● Can result in runs that aren’t reproducible or difficult to debug behavior (large models) |

# Cloud Machine Learning (Cloud ML)

**Fully managed, distributed training** and **prediction** for custom **TensorFlow** graph

Supports **Regression** and **Classification** initially

Integrated with **Cloud Dataflow** and **Cloud Datalab**

**Limited Preview** - cloud.google.com/**ml**

# Cloud ML

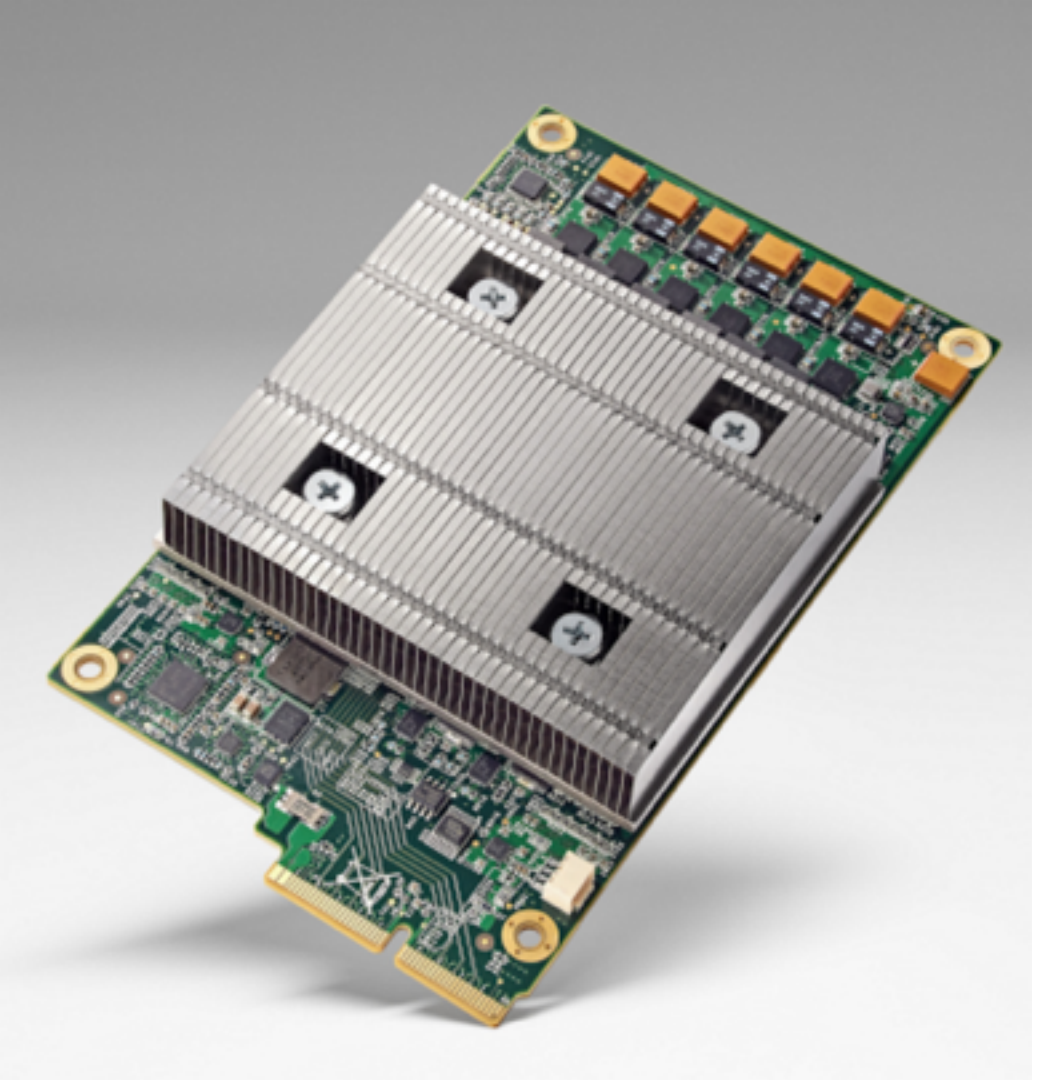
Jeff Dean's keynote: YouTube video

Define a custom **TensorFlow** graph

Training at local: **8.3 hours** w/ 1 node

Training at cloud: **32 min** w/ **20 nodes** (**15x** faster)

Prediction at cloud at **300 reqs / sec**



# Tensor Processing Unit

ASIC for TensorFlow

Designed by Google

10x better perf / watt

8 bit quantization

# Thank You

https://www.tensorflow.org/

https://cloud.google.com/ml/

http://bit.ly/tensorflow-workshop

Ian Lewis
@IanMLewis