Large Scale Deep Learning

Jeff Dean
Google Senior Fellow

Google

Joint work with **many** colleagues at Google

How Can We Build More Intelligent Computer Systems?

Need to perceive and understand the world

Basic speech and vision capabilities

Language understanding

User behavior prediction

Ability to interact with environment

...

How can we do this?

- Cannot write algorithms for each task we want to accomplish separately
- Need to write general algorithms that learn from observations

Can we build systems that:

- Generate understanding from raw data
- Solve difficult problems to improve Google's products
- Minimize software engineering effort
- Advance state of the art in what is possible

Plenty of Data

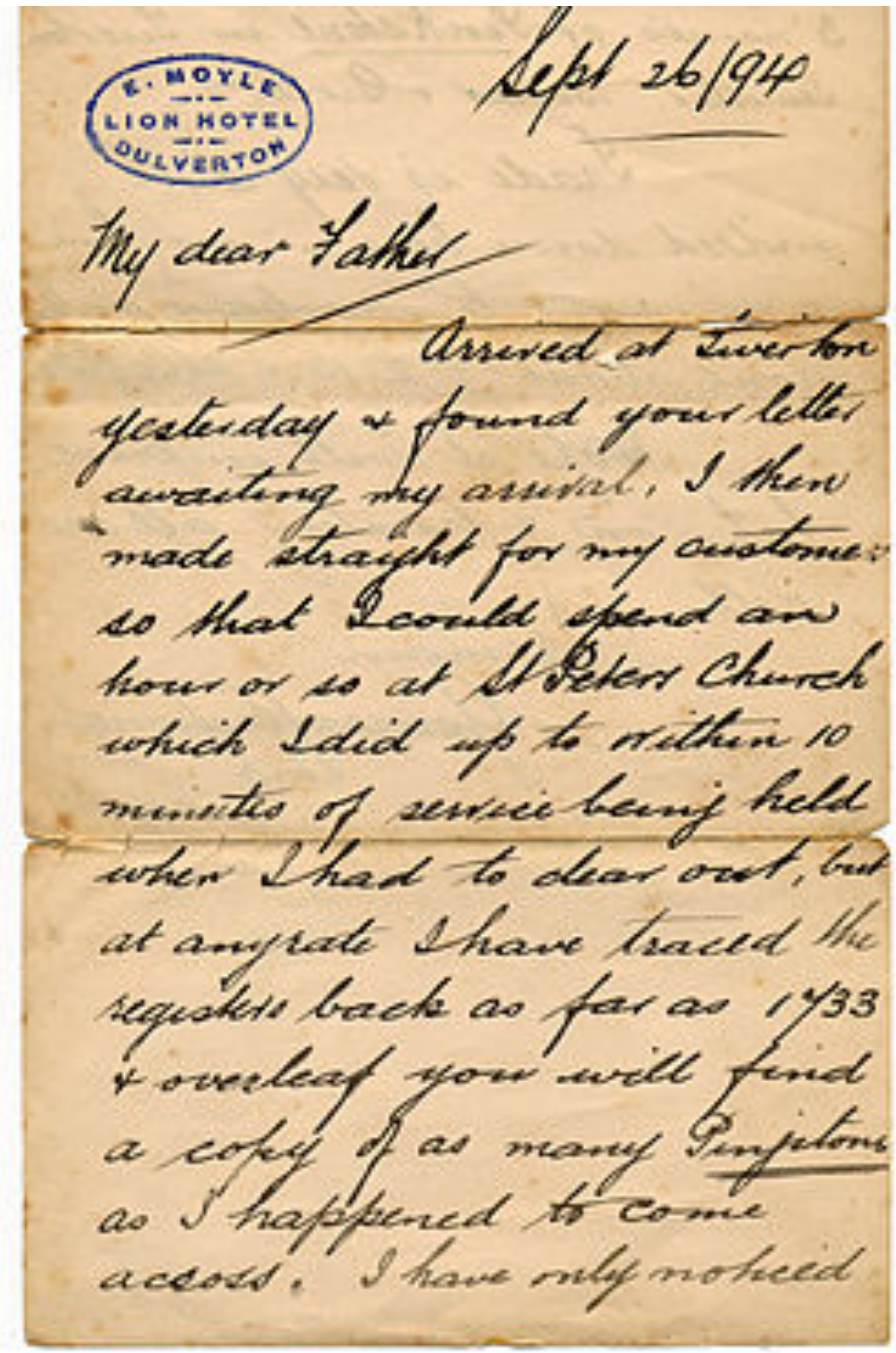
- Text: trillions of words of English + other languages
- Visual: billions of images and videos
- Audio: thousands of hours of speech per day
- User activity: queries, result page clicks, map requests, etc.
- Knowledge graph: billions of labelled relation triples
- ...

Image Models

What are these numbers?

What are all these words?

How about these words?


E. MOYLE
LION HOTEL
DULVERTON

Sept 26/94

My dear Father

Arrived at Tiverton yesterday & found your letter awaiting my arrival. I then made straight for my customer so that I could spend an hour or so at St Peters Church which I did up to within 10 minutes of service being held when I had to clear out, but at anyrate I have traced the registers back as far as 1733 & overleaf you will find a copy of as many Pimpstones as I happened to come across. I have only noticed

เป็นมนุษย์สุดประเสริฐเลิศคุณค่า
กว่าบรรดาฝูงสัตว์เดรัจฉาน
จงฝ่าฟันพัฒนาวิชาการ
อย่าล้างผลาญฤๅเข่นฆ่าบีฑาใคร
ไม่ถือโทษโกรธแช่งซัดฮึดฮัดด่า
หัดอภัยเหมือนกีฬาอัชฌาสัย
ปฏิบัติประพฤติกฎกำหนดใจ
พูดจาให้จ๊ะ ๆ จ๋า ๆ น่าฟังเอยฯ

Textual understanding

Understand difference between:

Idea 1

I was given a card by her in the garden.
or
She gave me a card in the garden.

vs.

Idea 2

I gave her a card in the garden.
or
In the garden, I gave her a card.

National Academy of Engineering
Grand Challenges for 21st Century

Make solar energy economical

Provide energy from fusion

Develop carbon sequestration methods

Manage the nitrogen cycle

Provide access to clean water

Restore/improve urban infrastructure

Advance health informatics

Engineer better medicines

Reverse-engineer the brain

Prevent nuclear terror

Secure cyberspace

Enhance virtual reality

Advance personalized learning

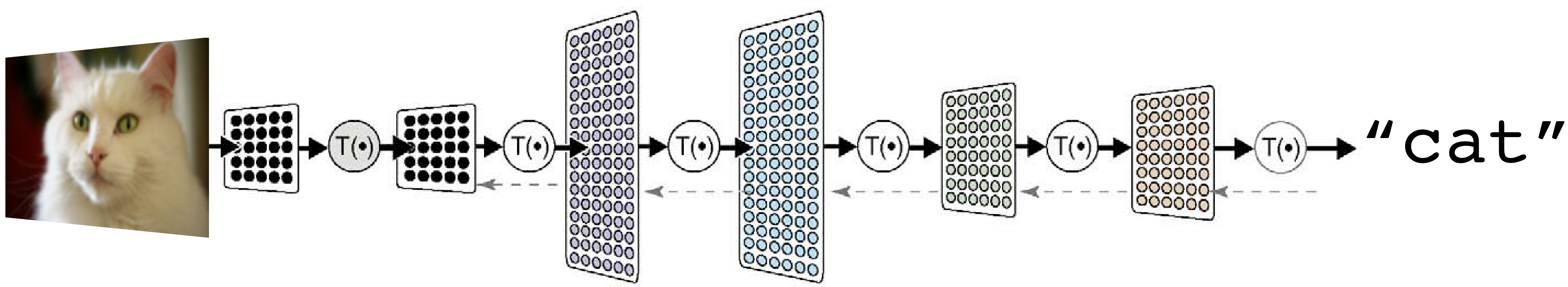
Engineer tools of scientific discovery

http://www.engineeringchallenges.org/cms/challenges.aspx

Google

What is Deep Learning?

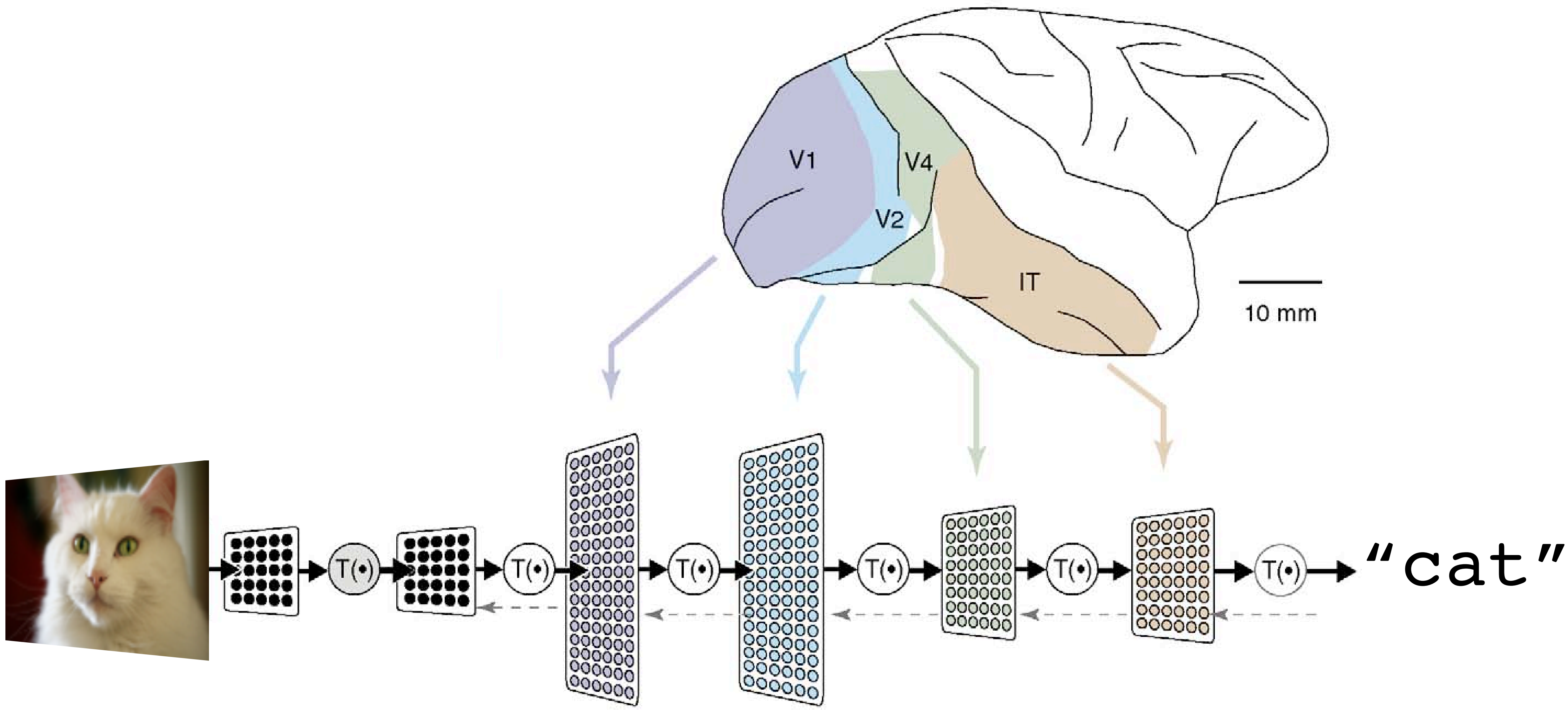
- The modern reincarnation of Artificial Neural Networks from the 1980s and 90s.
- A collection of simple trainable mathematical units, which collaborate to compute a complicated function.
- Compatible with supervised, unsupervised, and reinforcement learning.



Google

What is Deep Learning?

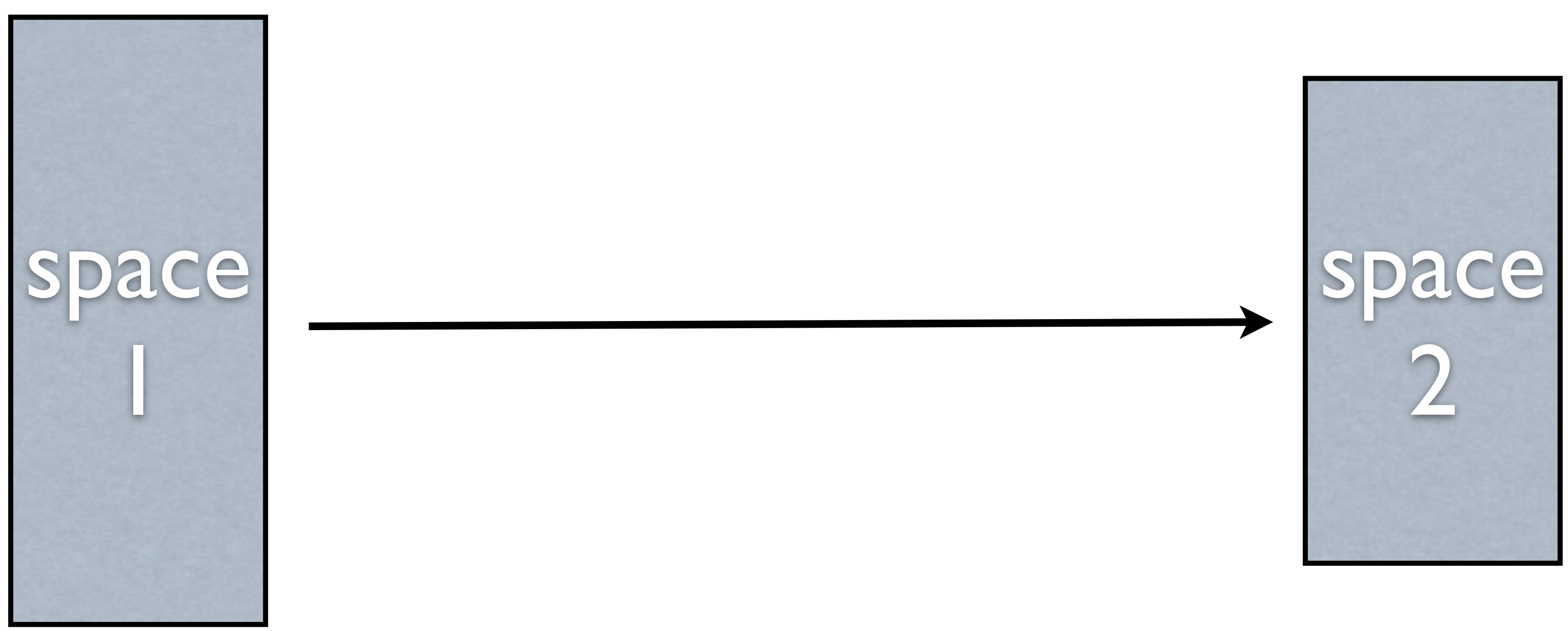
- Loosely inspired by what (little) we know about the biological brain.
- Higher layers form higher levels of abstraction



Google

Neural Networks

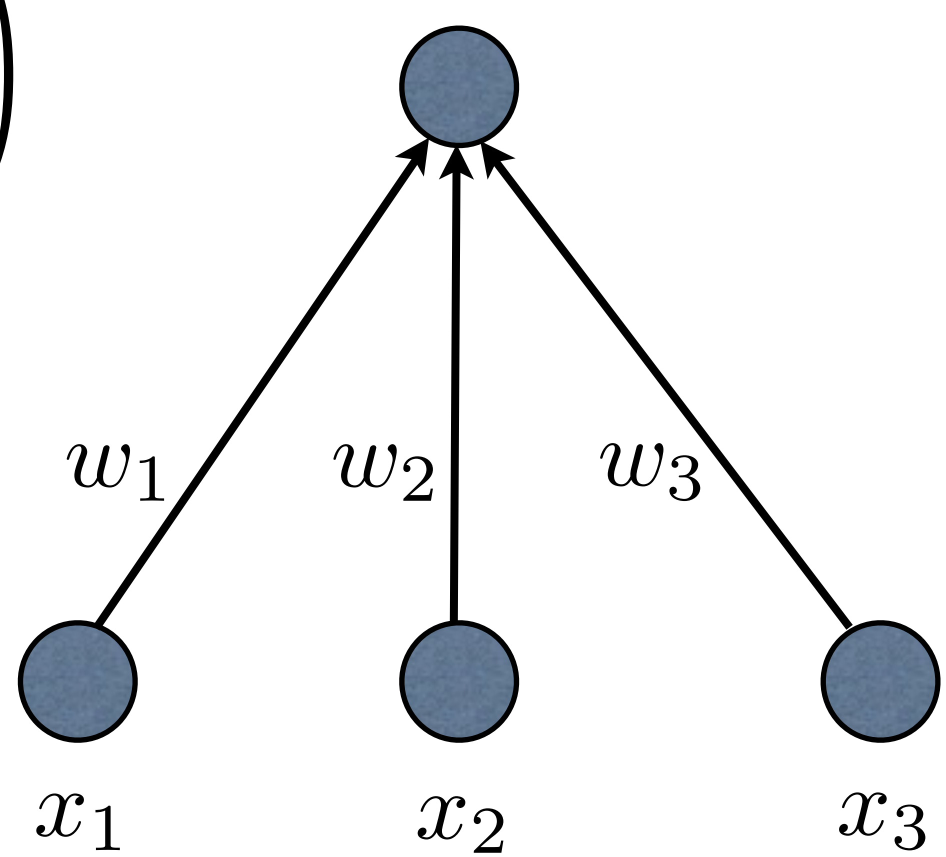
- Learn a complicated function from data

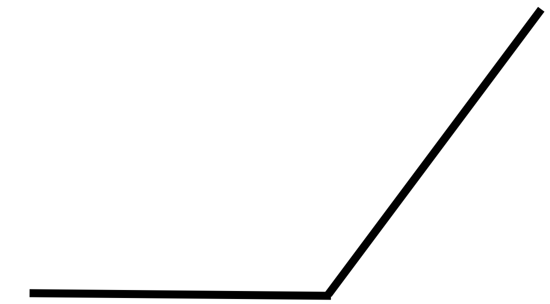


The Neuron

- Different weights compute different functions

$$y_i = F\left(\sum_i w_i x_i\right)$$

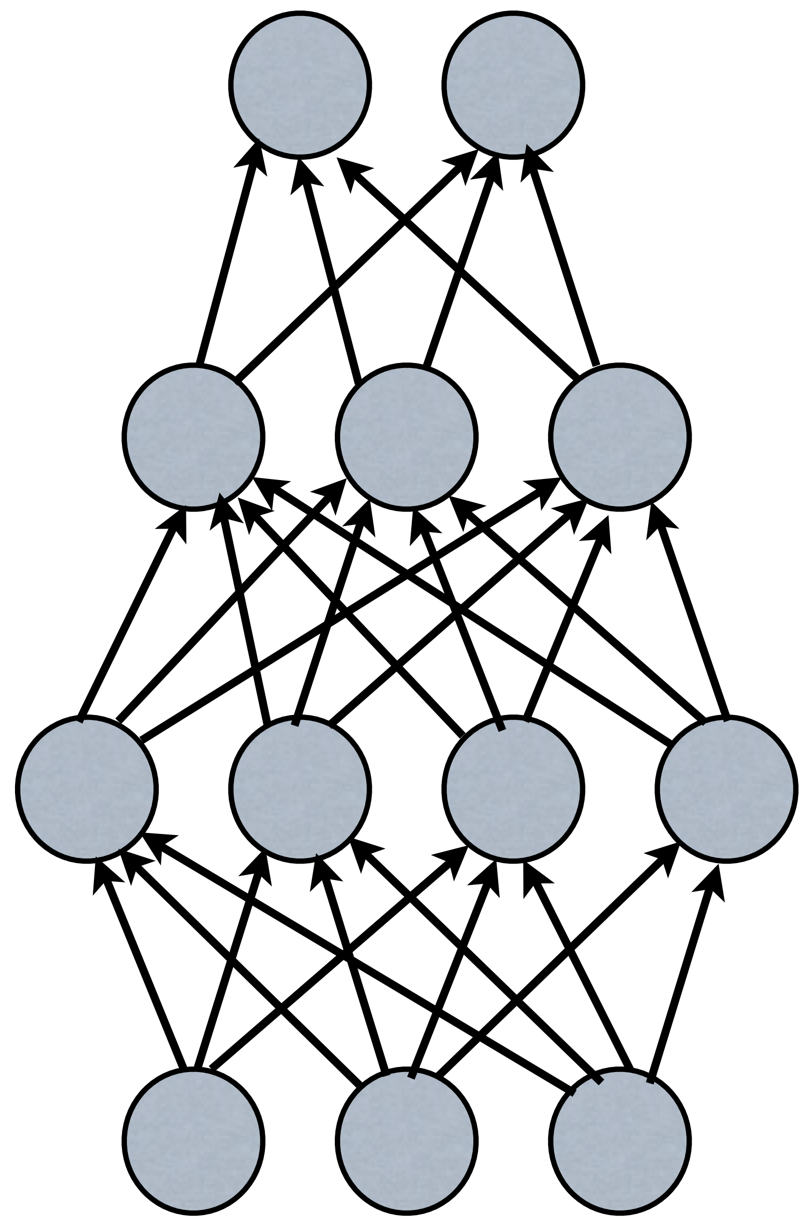




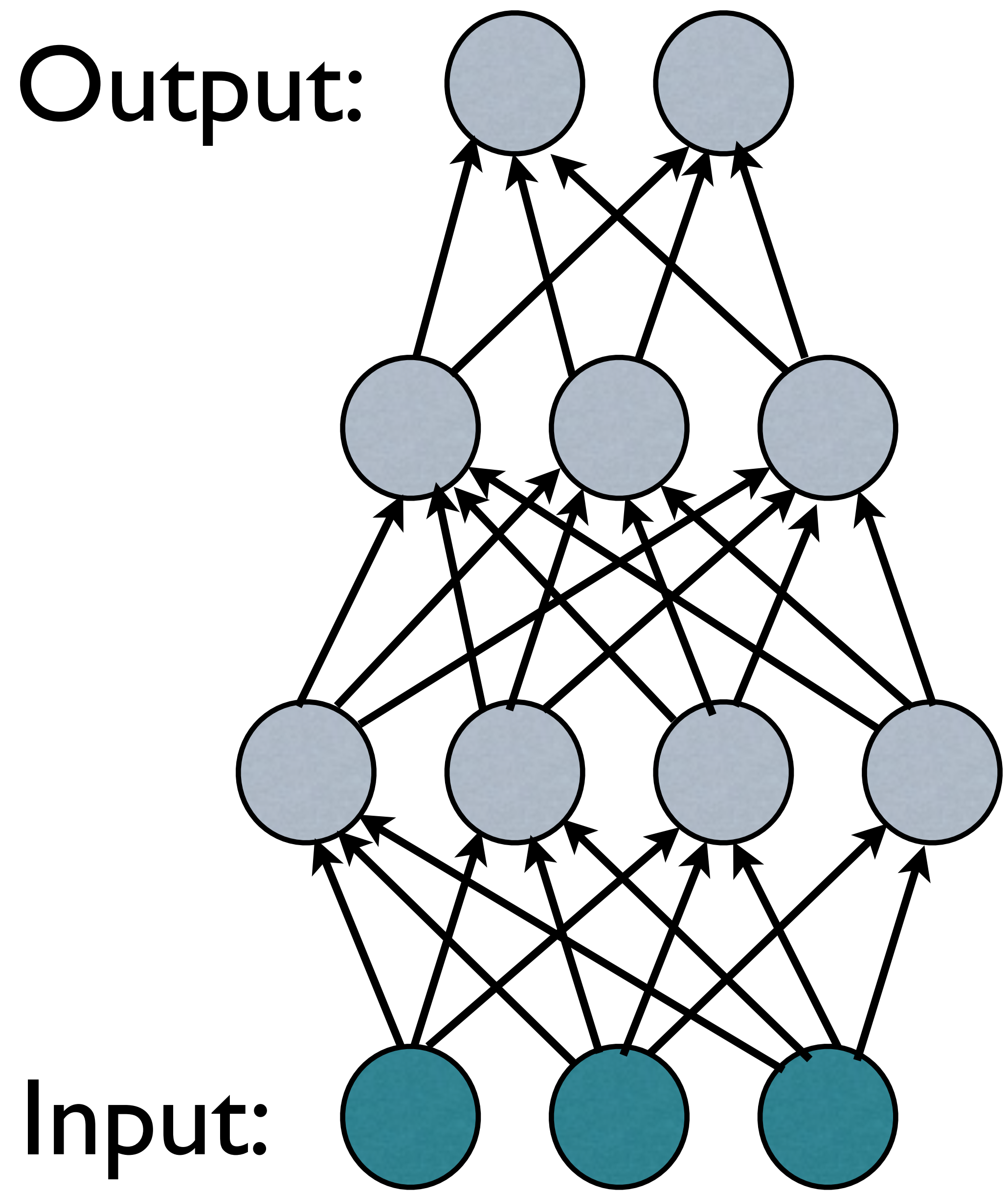
$F(x) = \max(0, x)$

Neural Networks

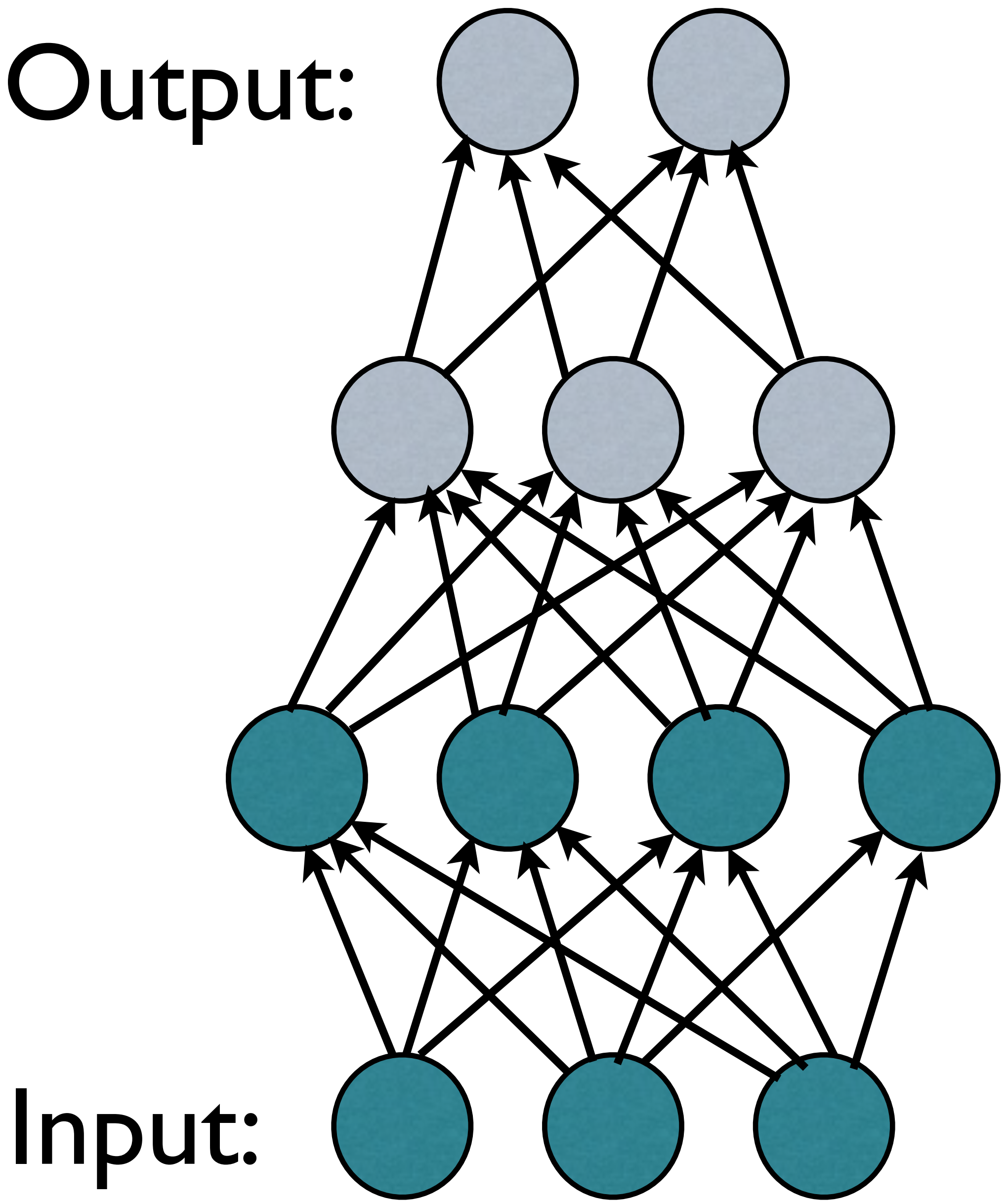
- Different weights compute different functions



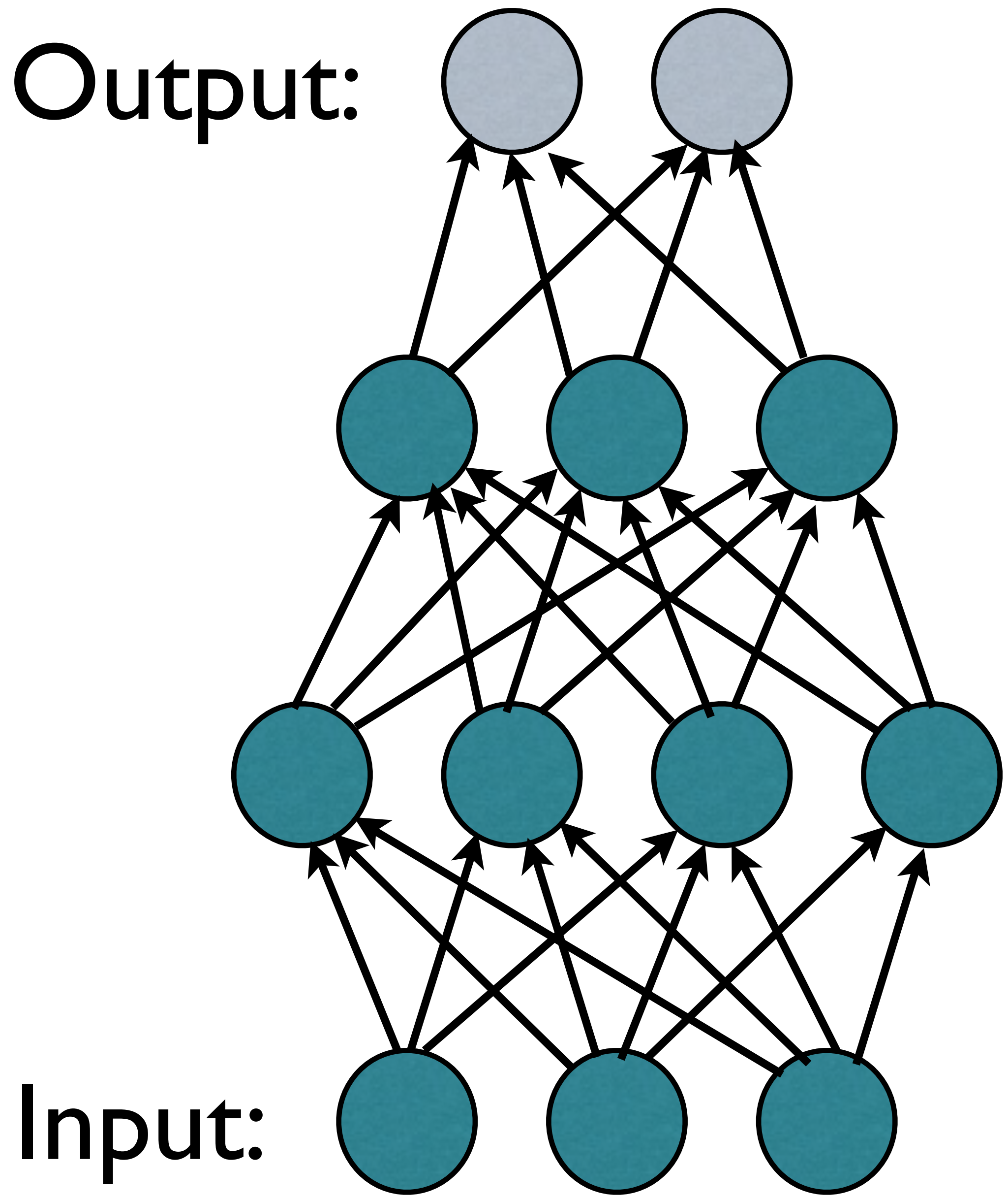
Neural Networks



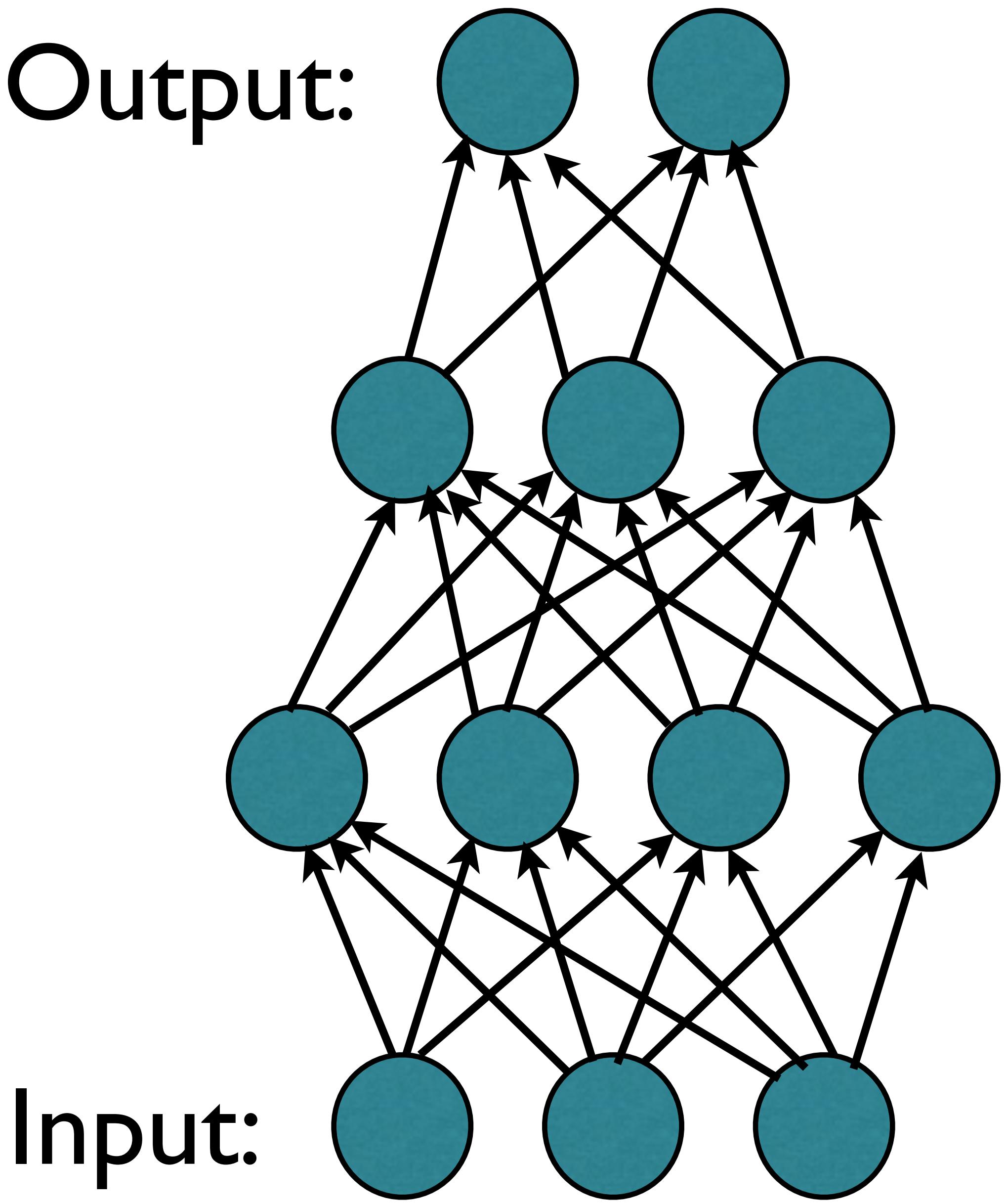
Neural Networks



Neural Networks



Neural Networks



Learning Algorithm

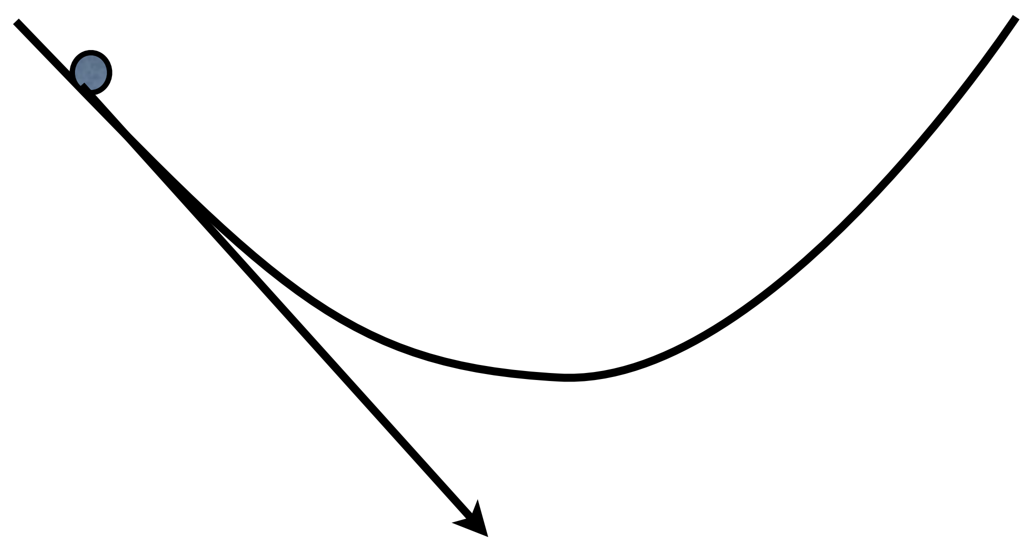
- **while** not done
  - pick a random training case **(x, y)**
  - run neural network on input **x**
  - modify connections to make prediction closer to **y**

Learning Algorithm

- **while** not done
  - pick a random training case **(x, y)**
  - run neural network on input **x**
  - <u>modify connection weights to make prediction closer to y</u>

How to modify connections?

- Follow the gradient of the error w.r.t. the connections



Gradient points in direction of improvement

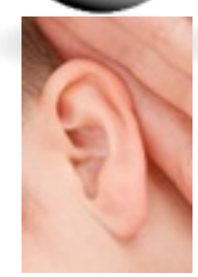
One Simple Scalability Aid

- Previous algorithm was a bit of a lie. We don't do one (x,y) example at a time.
- Rather, we do "mini-batches" of, say, 32 to 1024 different (x,y) pairs at a time, and average the gradient for all of these examples
- Turns matrix-vector operations into matrix-matrix operations
  - Nicely suited for GPUs

What can neural nets compute?

- Human perception is very fast (0.1 second)
  - Recognize objects ("see")

  - Recognize speech ("hear")

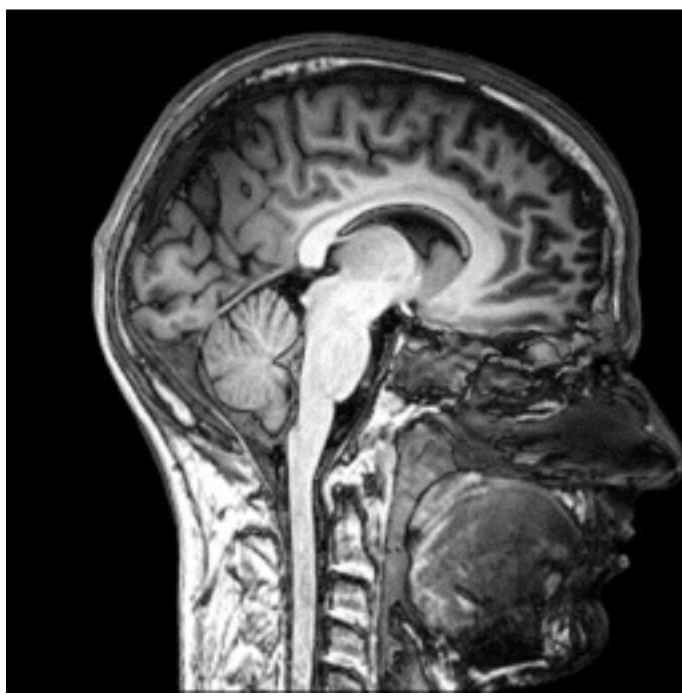


  - Recognize emotion

  - Instantly see how to solve some problems
  - And many more!

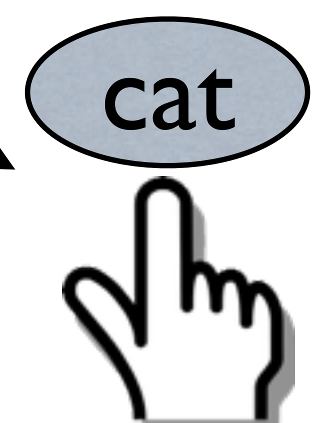
Why do neural networks work?

0.1 sec: neurons can fire only 10 times!



see image

click if cat



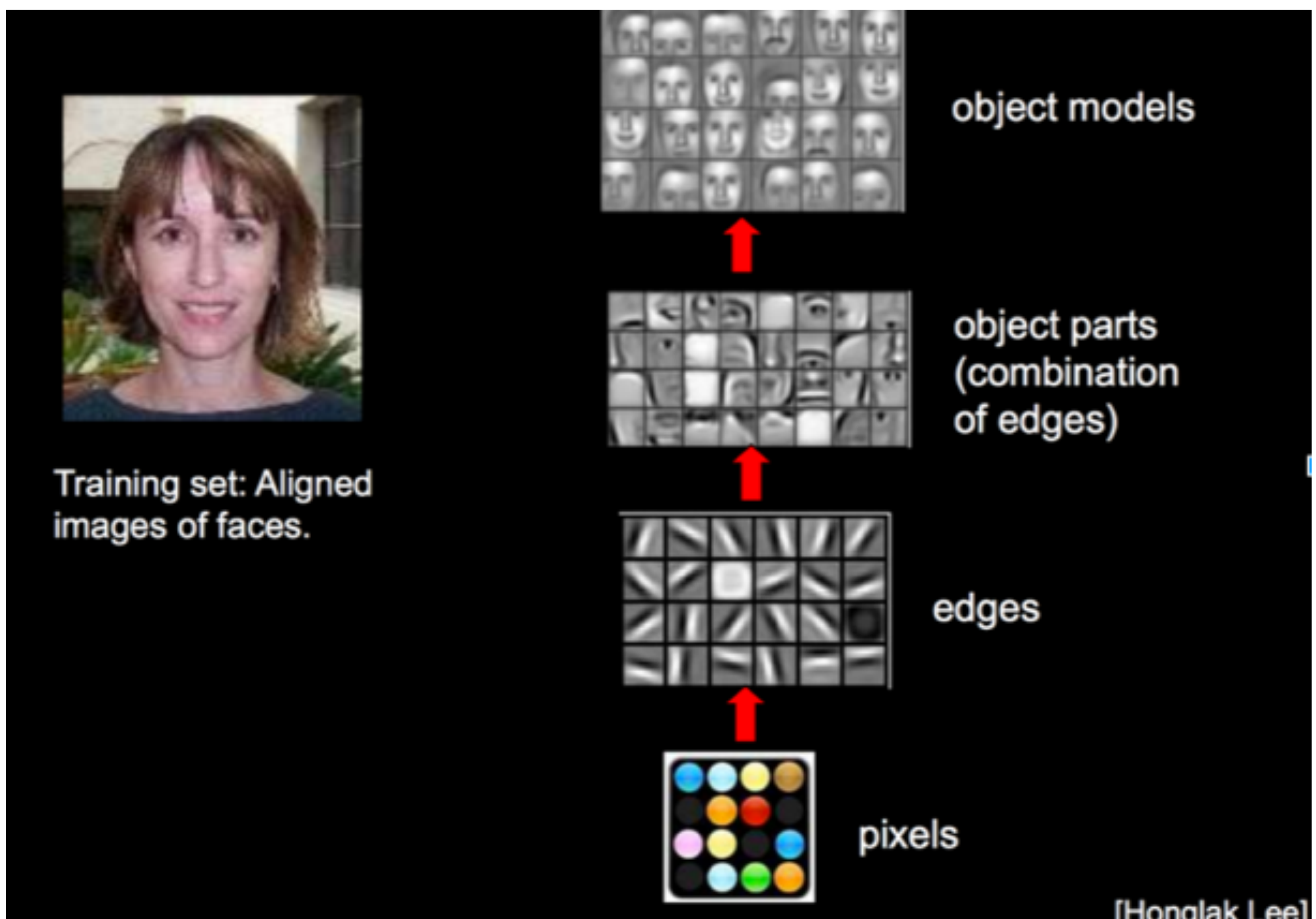
Why do neural networks work?

- **Anything humans can do in 0.1 sec, the right big 10-layer network can do too**

Important Properties of Deep Neural Networks

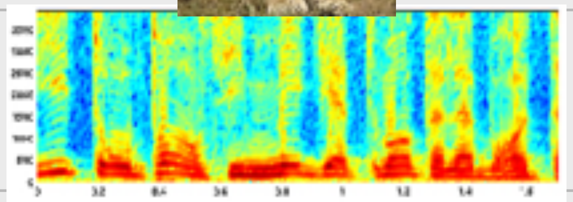
- **Automatic**: features developed as part of learning process
  - very good at learning from raw data (pixels, audio waveforms, etc.)
- **Hierarchical**: complex features built from simple features

Together: Amazing pattern recognition ability



Google

Functions Artificial Neural Nets Can Learn

Input	Output
Pixels: 	"lion"
Audio: 	"see at tuhl res taur aun ts"
<query, doc>	P(click on doc)
"Hello, how are you?"	"Bonjour, comment allez-vous?"
Pixels: 	"A close up of a small child holding a stuffed animal"

Research Objective: Make It Simple!

- Internal software framework usable by anyone at Google
  - Enable both research as well as training/use of models for products
  - Many dozens of production launches of neural nets for real problems
- Allows neural architectures and training procedures to be easily described
- Handles fault tolerance, recovery, parallelization, etc. with just a few simple hints from the user

Dean, *et al.* , *Large Scale Distributed Deep Networks,* NIPS, 2012.

Research Objective: Make It Simple!

```
m = Model(num_partitions=4)
input = m.ImageInput("/dir/myimages", rows=256, cols=256)
hidden = m.NeuralRELULayer(input, 2000)
sm = m.Softmax(hidden, num_classes=10, labels=input.labels)
```

Dean, *et al.* , *Large Scale Distributed Deep Networks,* NIPS, 2012.

Time for Training & Its Effect on Research

Minutes, Hours:

- Interactive research! Instant gratification!
- Parameter exploration.

1-4 Days:

- Tolerable.
- Interactivity replaced by parallelization of experiments.

1-4 Weeks:

- High value experiments only.
- Progress stalls.

> 1 Month:

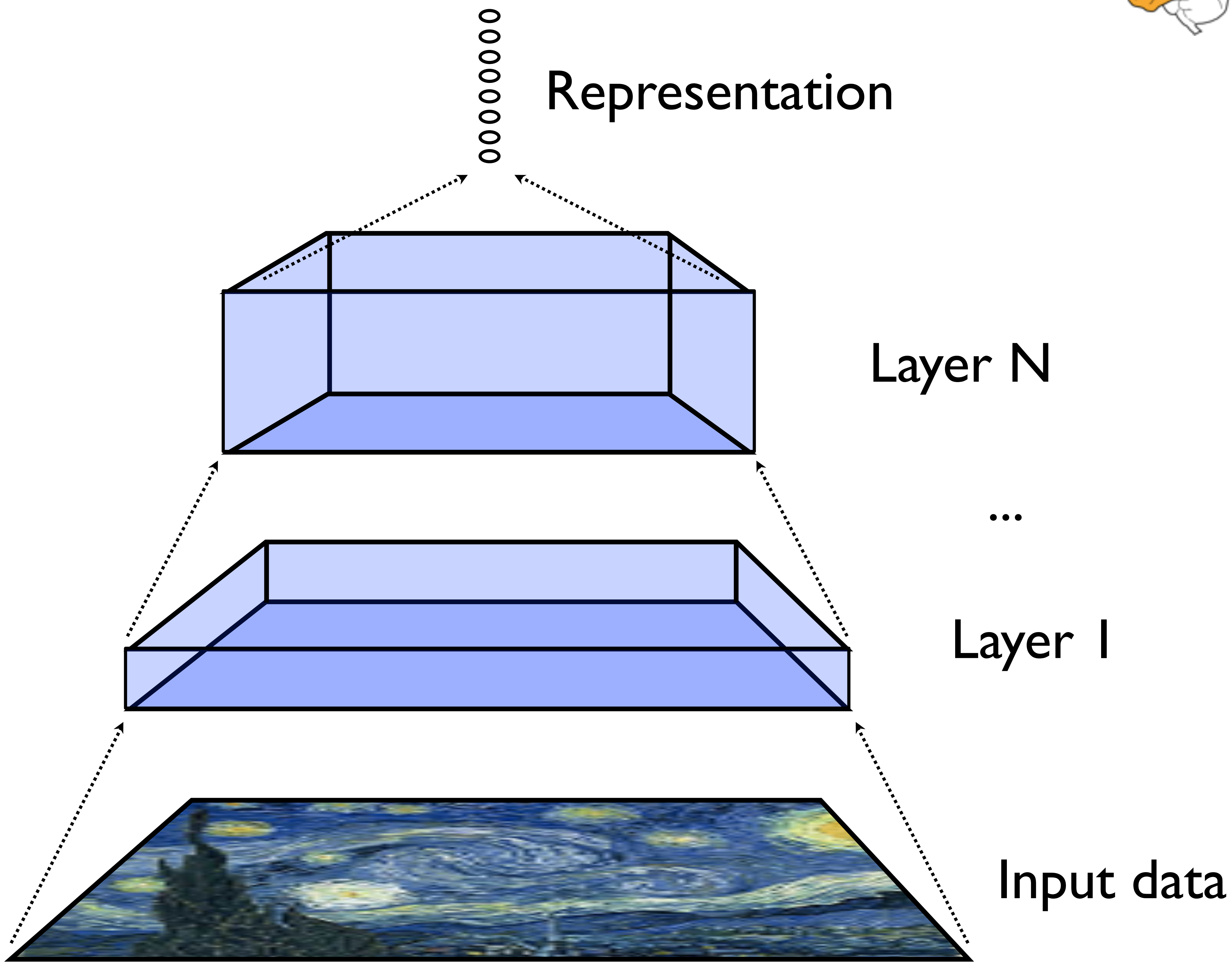
- Don't even try.

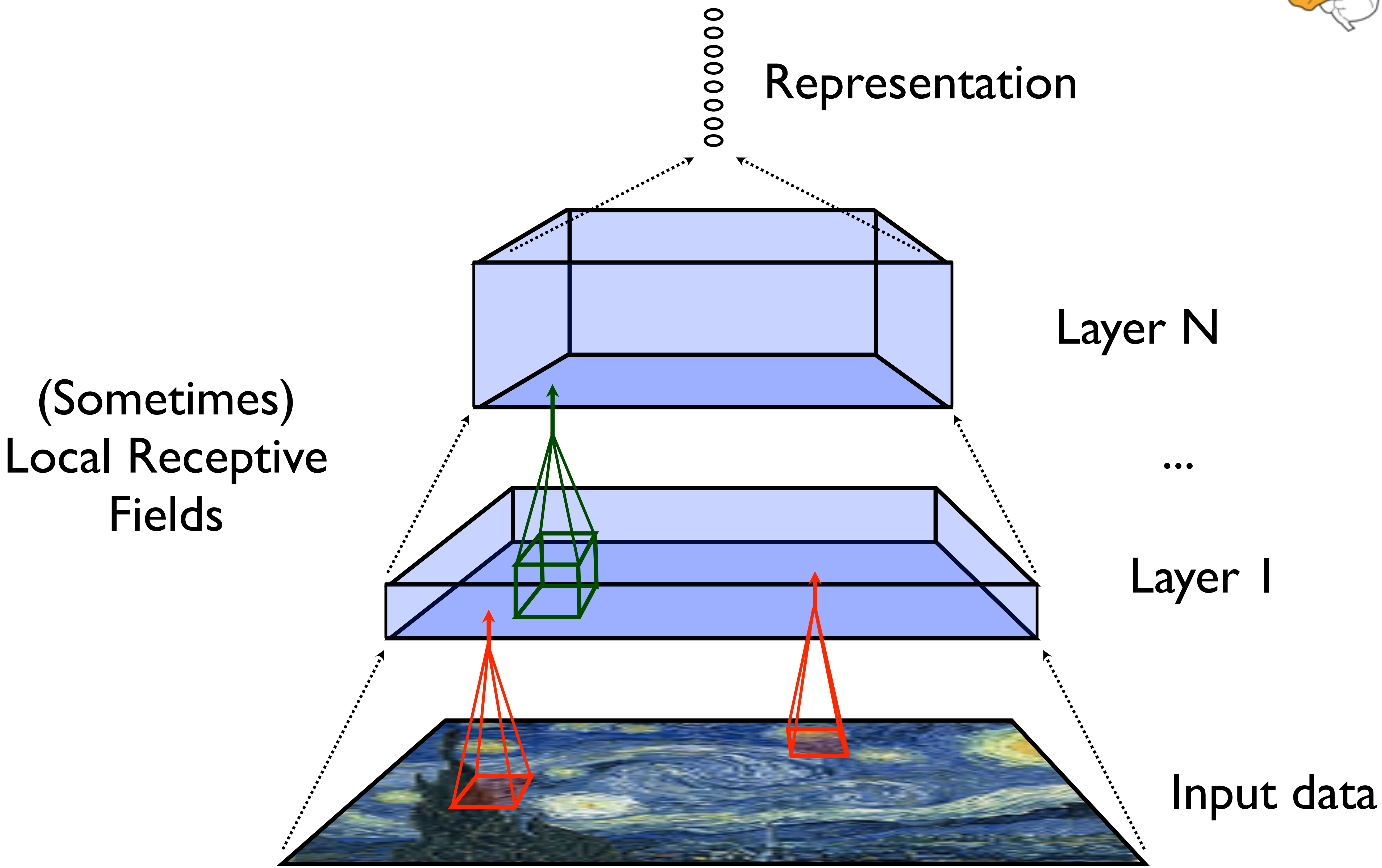
Train in a day what takes a single GPU card 6 weeks

How Can We Train Big Nets Quickly?

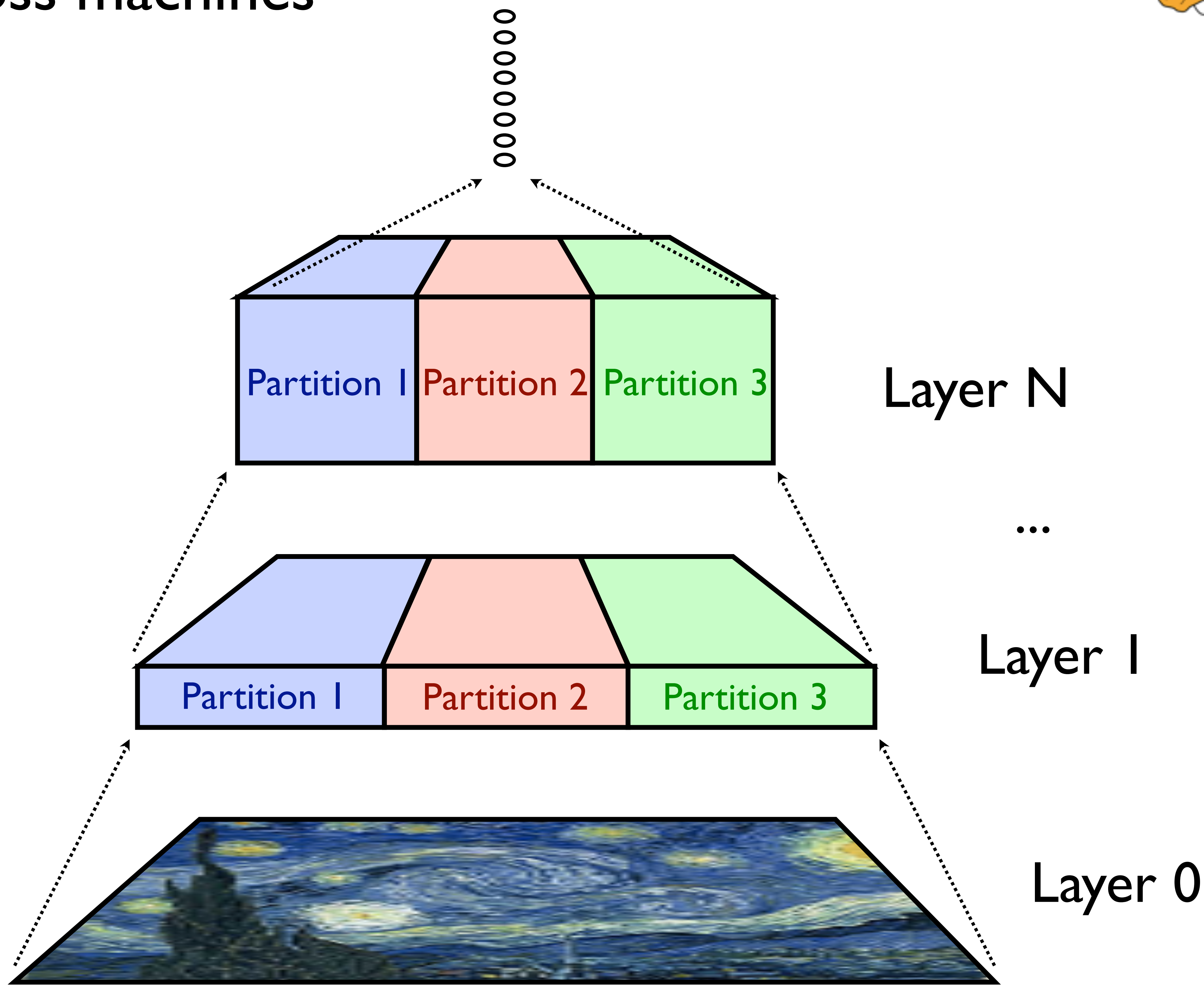
- Exploit many kinds of parallelism

- Model parallelism
- Data parallelism
- (Plus running many simultaneous experiments on top of these approaches)


Representation
Layer N
...
Layer 1
Input data

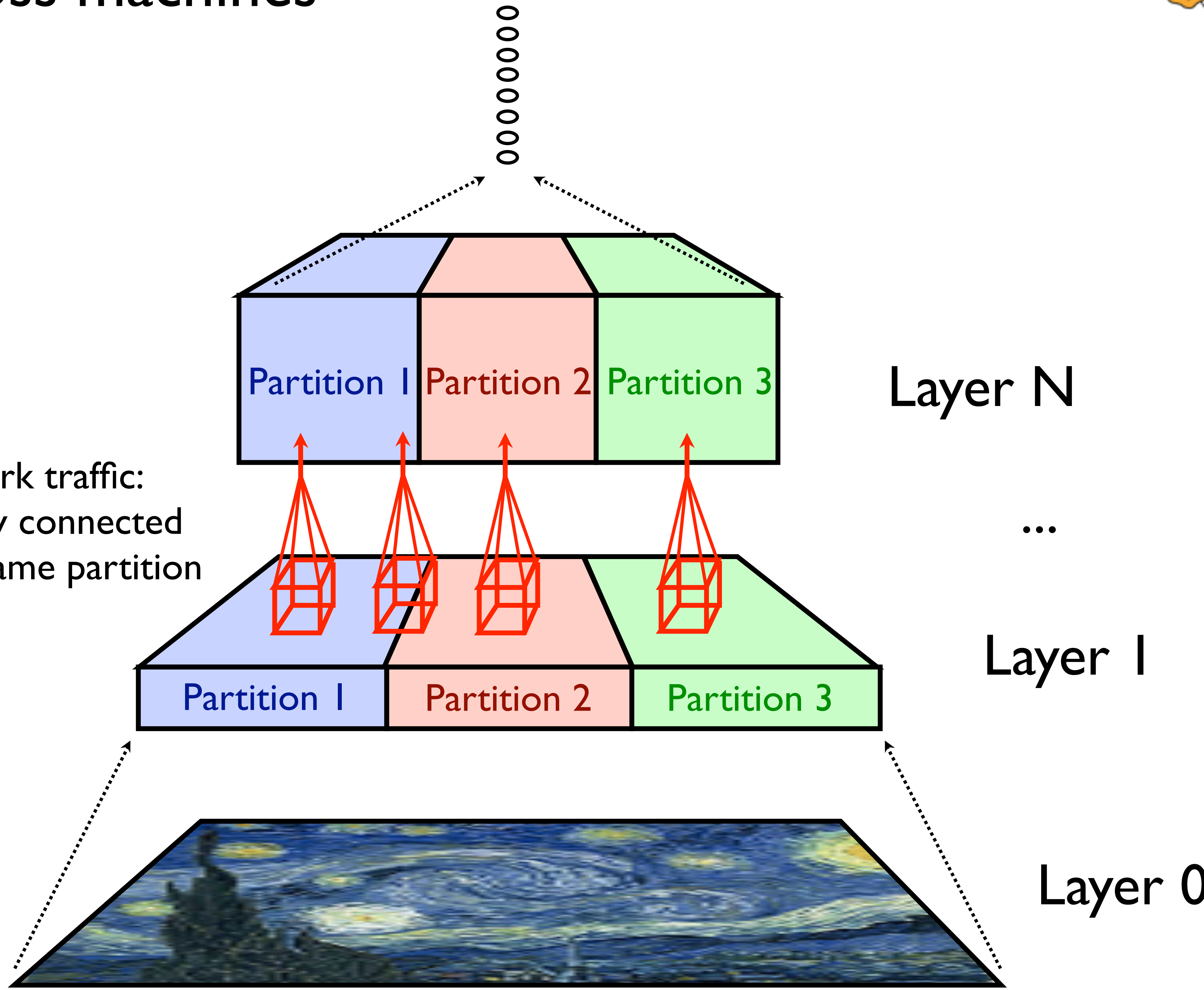

Representation
Layer N
...
Layer 1
Input data
(Sometimes)
Local Receptive
Fields

Model Parallelism: Partition model across GPUs and/or across machines



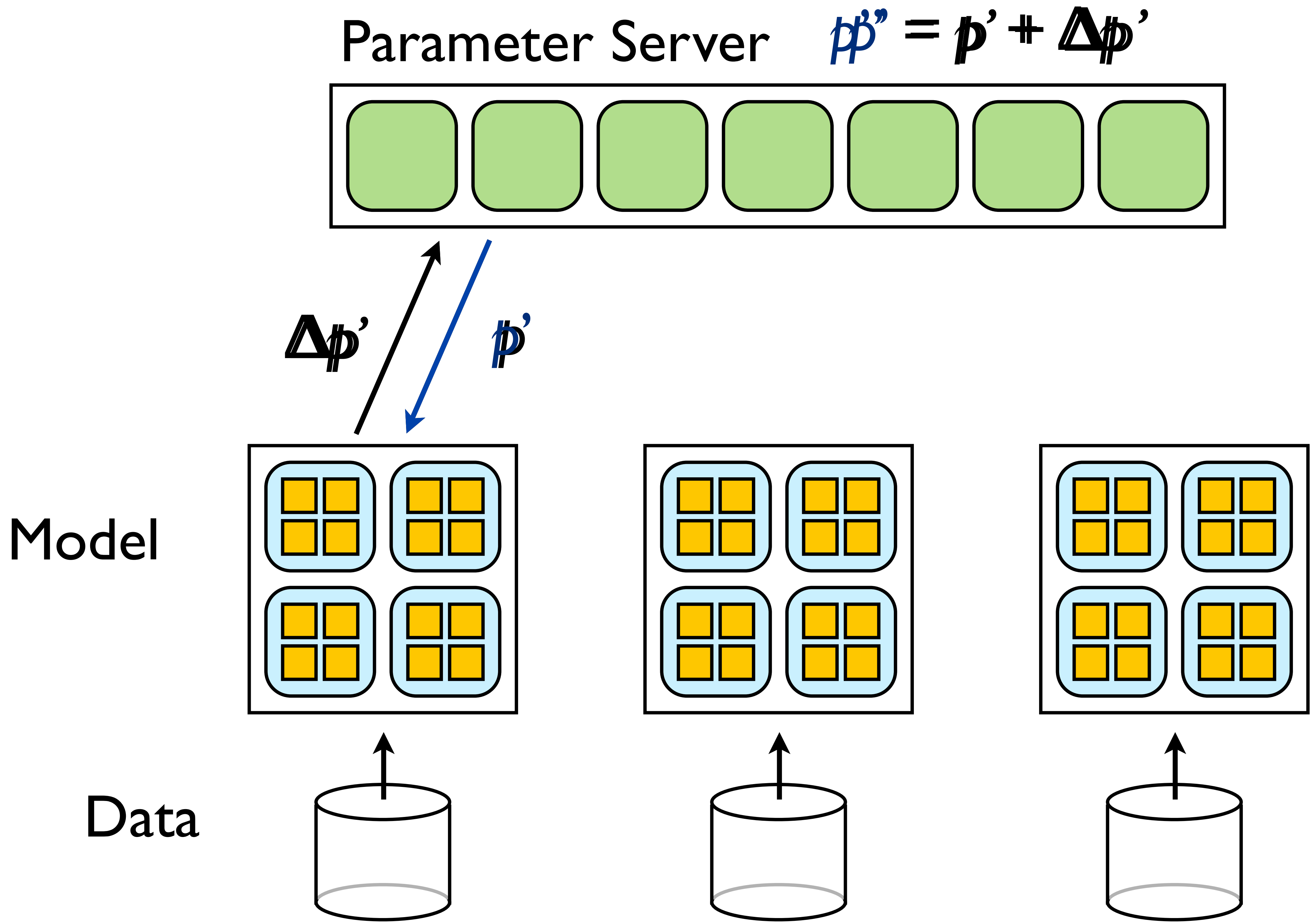
Model Parallelism: Partition model across GPUs and/or across machines

Minimal network traffic:
The most densely connected
areas are on the same partition



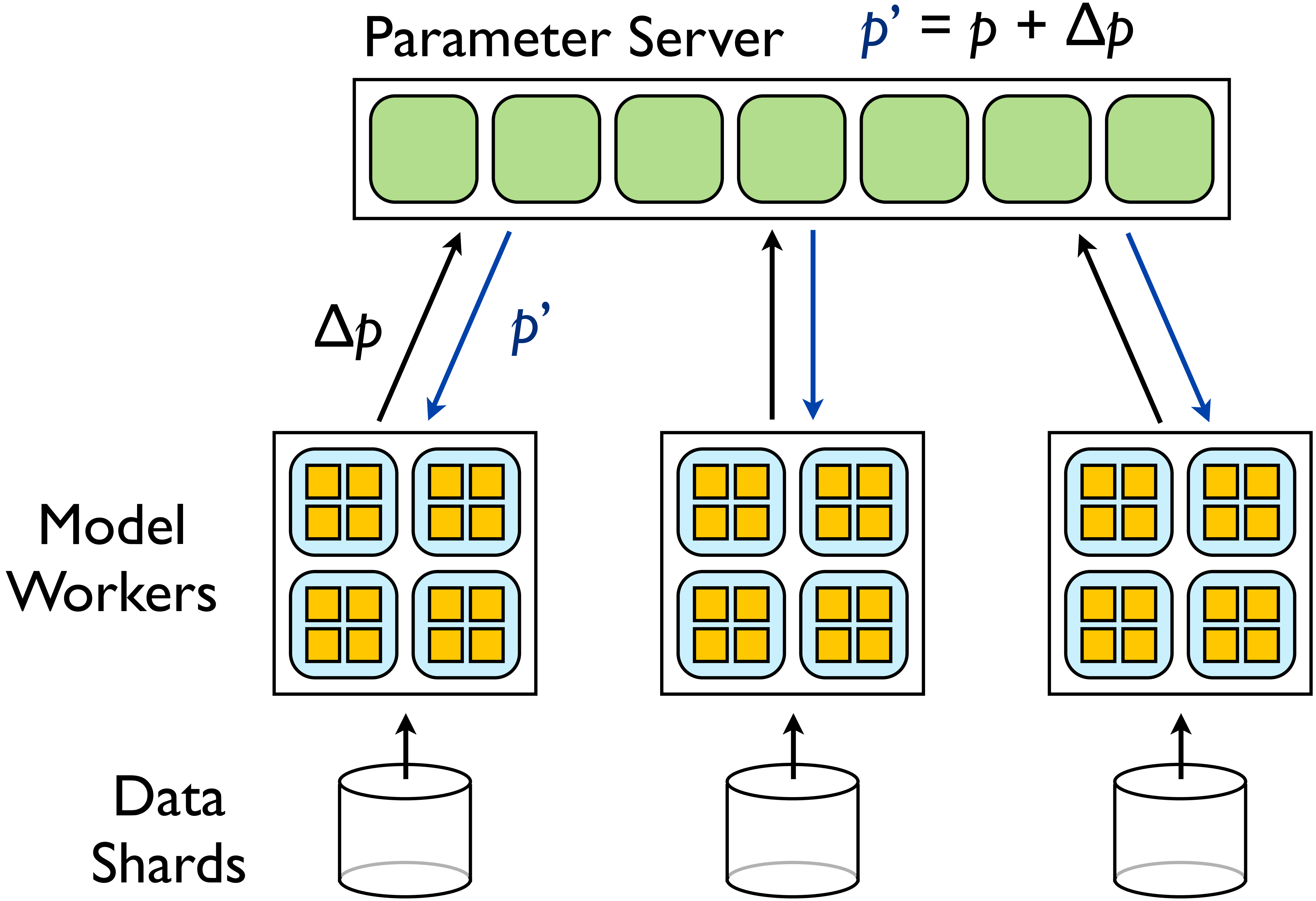
Regularly use models that are spread across dozens of machines

Data Parallelism:
Asynchronous Distributed Stochastic Gradient Descent



Data Parallelism:

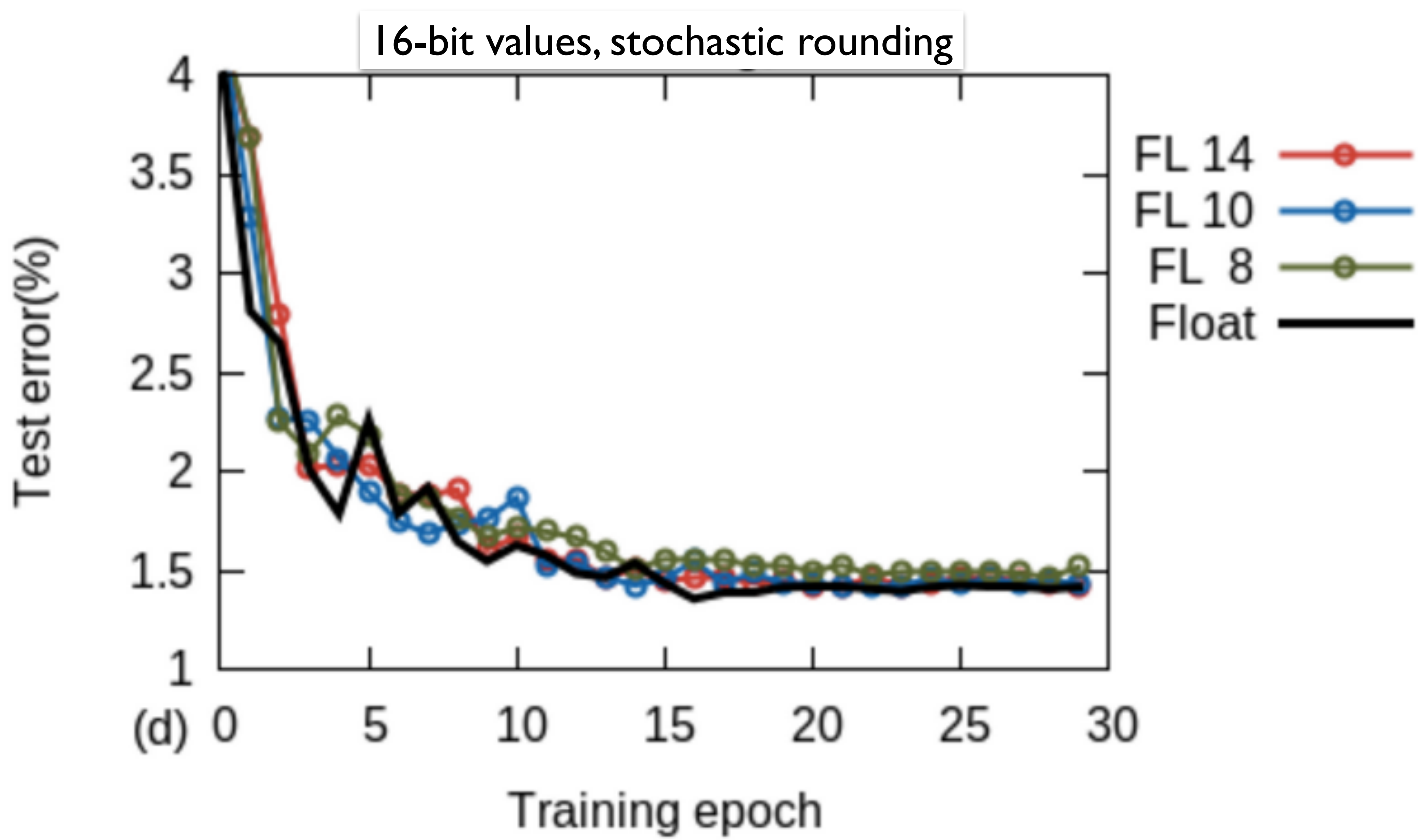
Asynchronous Distributed Stochastic Gradient Descent



Regularly use hundreds of model replicas
(each of which might be dozens of machines)

Other Scalability Aids

- Neural nets very tolerant of reduced precision arithmetic
  - e.g. chop 32-bit floats to 16 bits for network transfers
- Can even use 16-bit arithmetic. From Arxiv paper *Deep Learning with Limited Numerical Precision*, Gupta *et al.* (IBM):



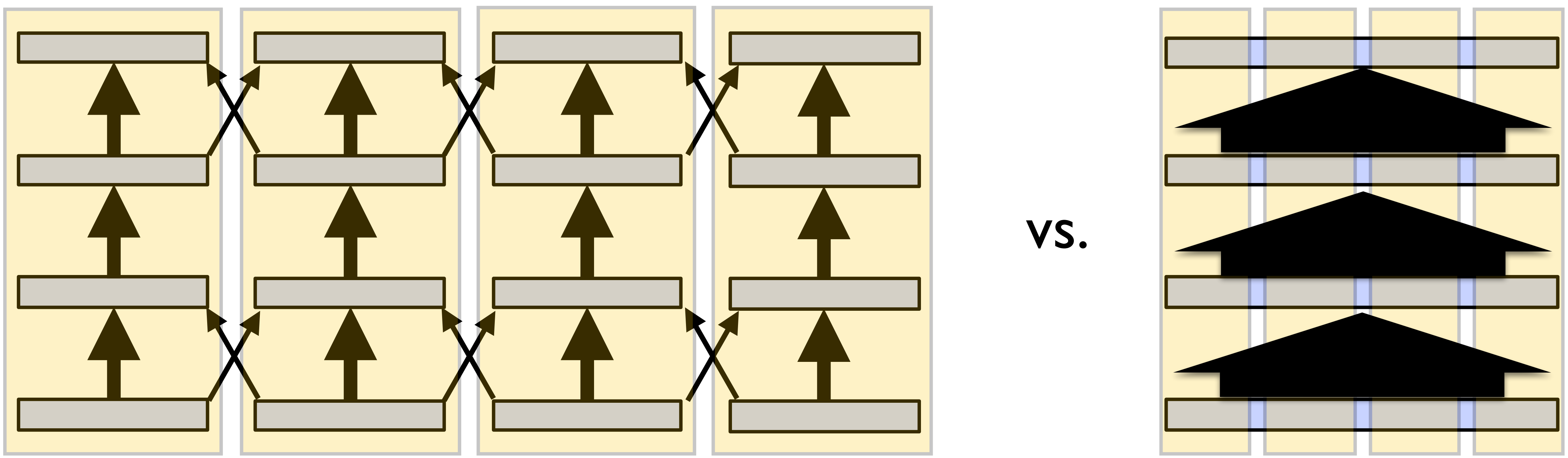
Other Scalability Aids

- ReLU activation functions produce “true zero values”
  - these are quite compressible

- “Concurrent steps”: pipelining overlaps computation and communication

Other Scalability Aids

- Use model connectivity structures that are adapted to the underlying communication channel capacities



Deep Learning @ Google

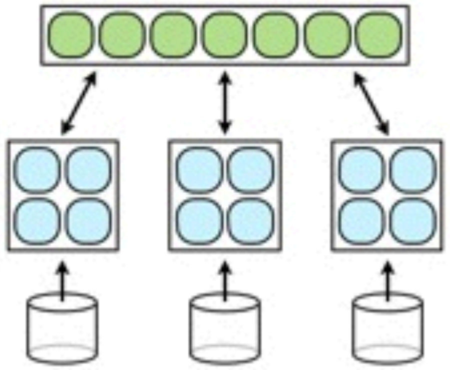
- Google has invested decades of person-years of systems engineers and artificial intelligence researchers in building the state-of-the-art infrastructure.
- We often leverage thousands of CPUs and GPUs to learn from billions of data samples in parallel.
  Dean *et al.*, Large Scale Distributed Deep Networks. NIPS 2012
- We publish frequently, and often place first in academic challenges in image recognition, speech recognition, etc.
  Szegedy *et al.*, GoogLeNet: Going Deeper with Convolutions. ILSVRC 2014
- Extensive and accelerating experience in using deep learning in real products: 47 production launches in the last 2 years.
  e.g. Photo search, Android speech recognition, StreetView, Ads placement...

Widely Applicable

Some areas we've published in:

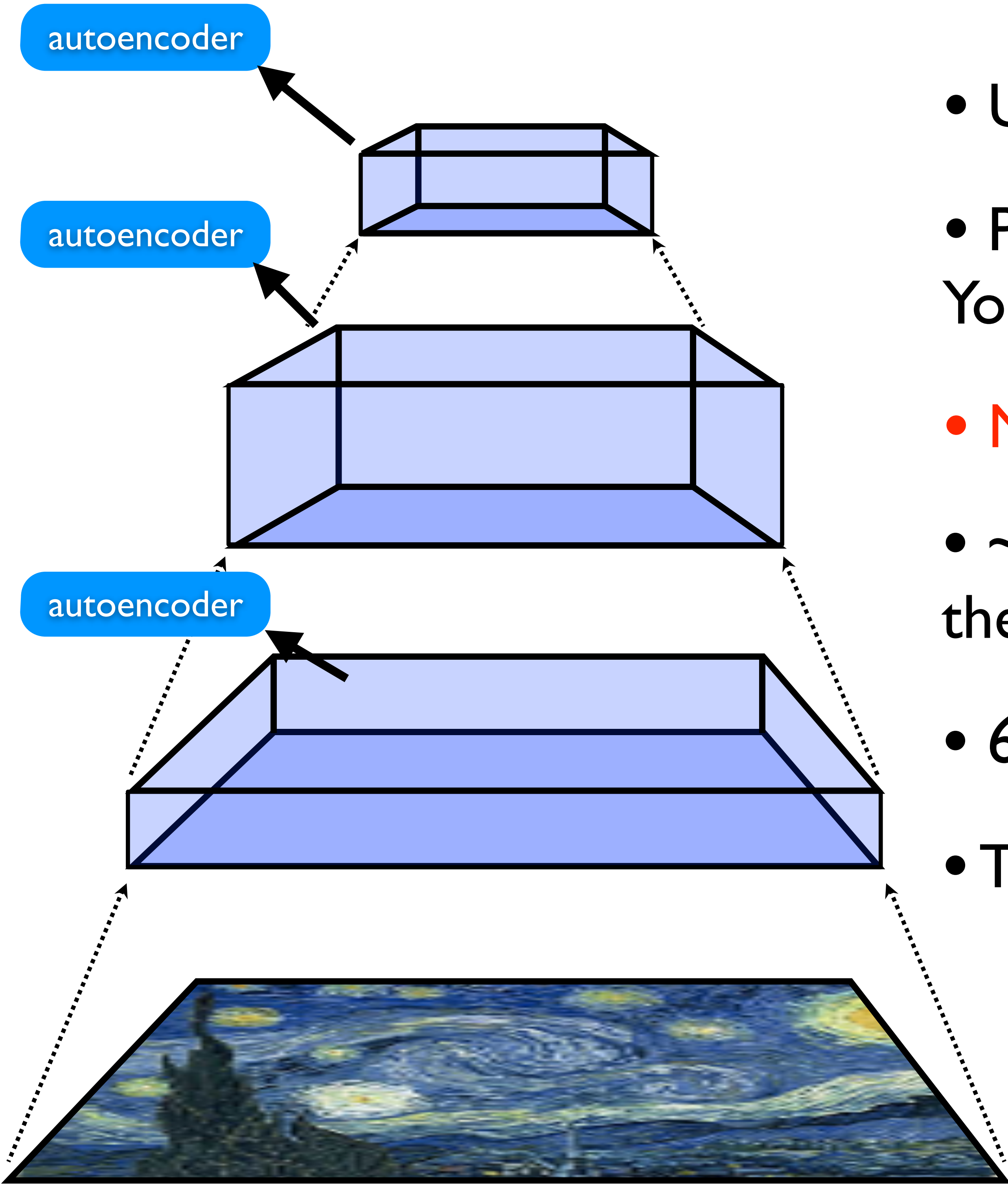
- Distributed training of large neural nets (Dean *et al.*, NIPS 2012)
- Object recognition in images (Erhan *et al.*, 2014)
- Object category discovery in video (Le *et al.*, ICML 2012)
- Speech recognition (Vanhoucke *et al.*, NIPS Workshop 2011)
- Annotating images with text (Vinyals *et al.*, arXiv 2014)
- OCR: reading text from images (Goodfellow *et al.*, ICLR 2014)
- Natural language understanding (Mikolov *et al.*, NIPS 2013)
- Machine translation (Sutskever *et al.*, NIPS 2014)
- Online advertising (Corrado *et al.*, ICML Workship 2012)

Google



Applications

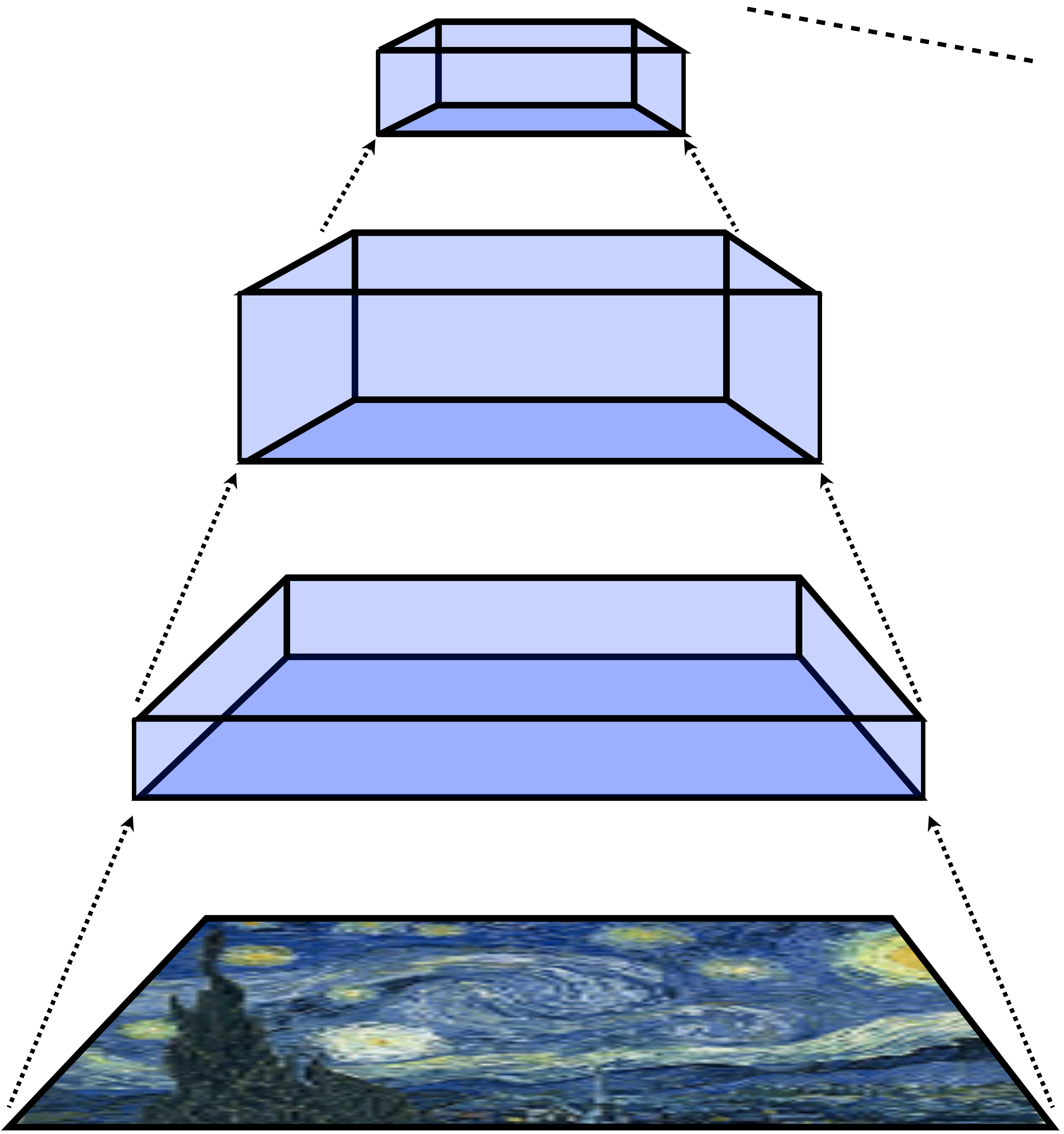
Newborn Baby+YouTube = ???



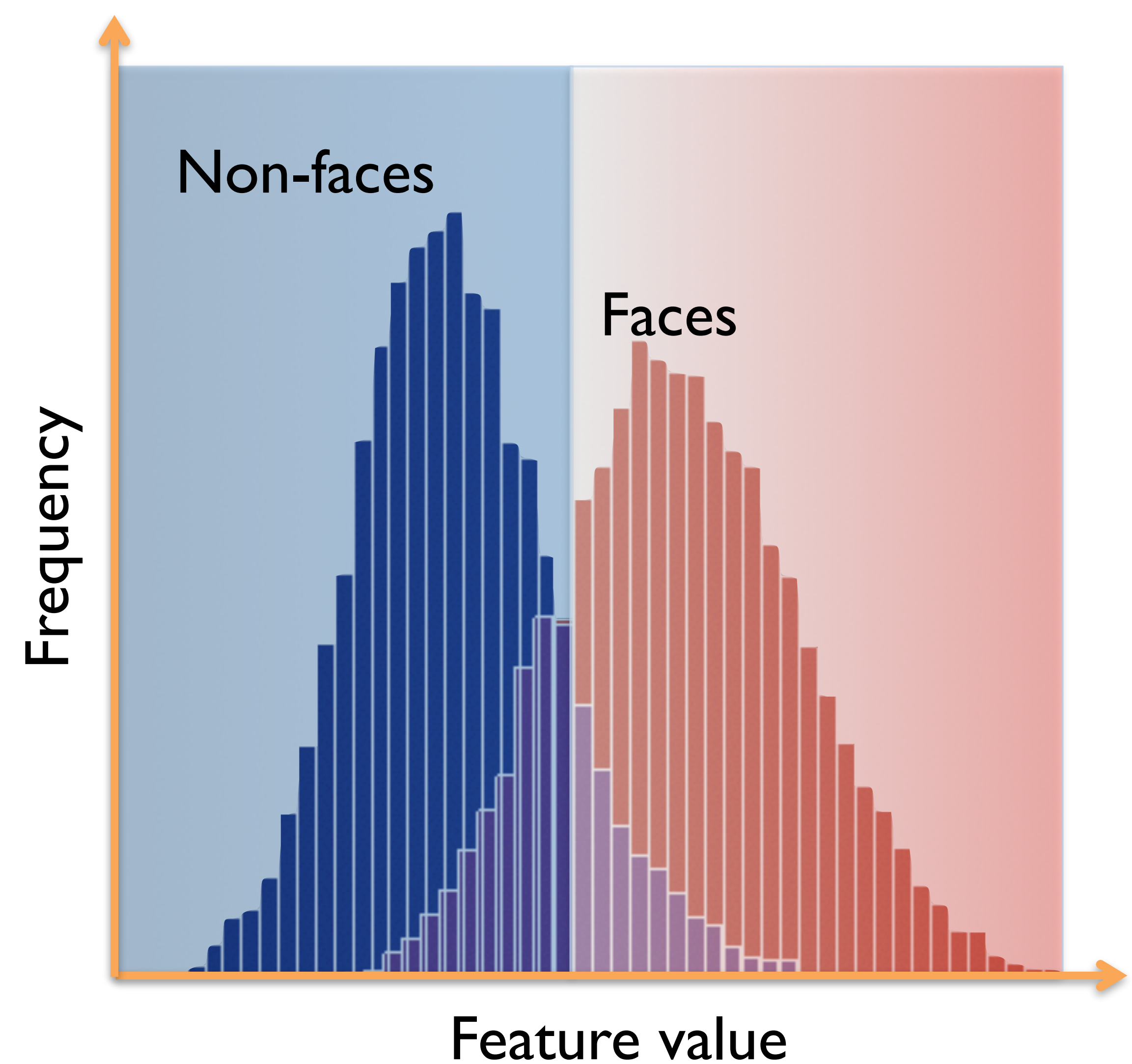
- Unsupervised training
- Pick one frame from each of 10 million YouTube videos & train model
- No labels!
- ~50x larger than largest deep network in the literature at the time
- 60,000 neurons at top level
- Trained on 16,000 CPU cores for 1 week

•Le, Ranzato, Monga, Devin, Chen, Corrado, Dean, & Ng. *Building High-Level Features Using Large Scale Unsupervised Learning*, ICML 2012.

Newborn Baby+YouTube = ???



Top level neurons seem to discover high-level concepts. For example, one neuron is a decent face detector:



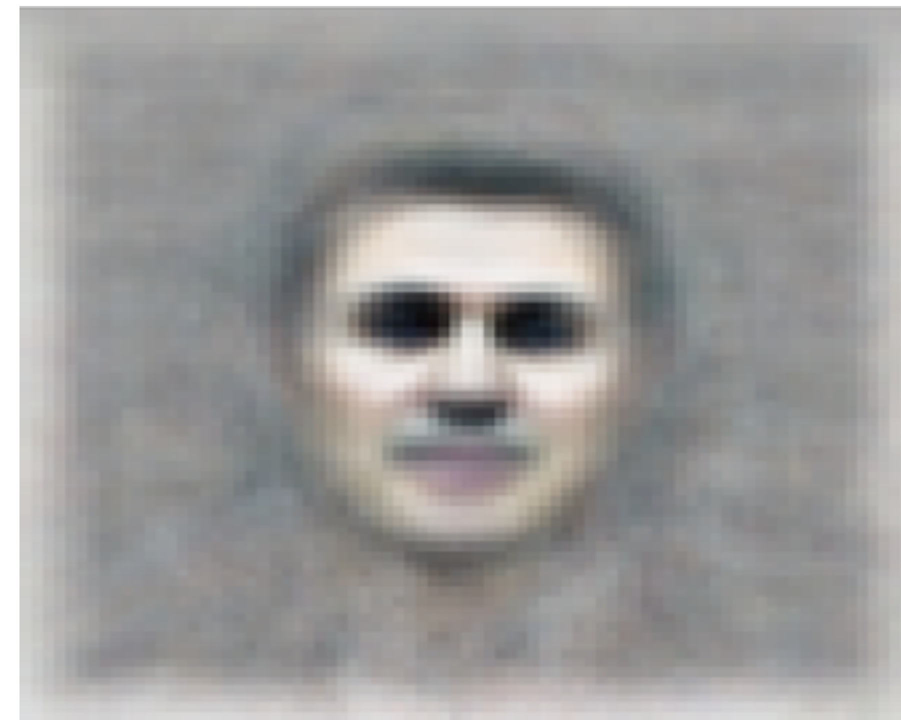
- Le, Ranzato, Monga, Devin, Chen, Corrado, Dean, & Ng. *Building High-Level Features Using Large Scale Unsupervised Learning*, ICML 2012.

Purely Unsupervised Feature Learning in Images

Most face-selective neuron

Top 48 stimuli from the test set

Optimal stimulus
by numerical optimization

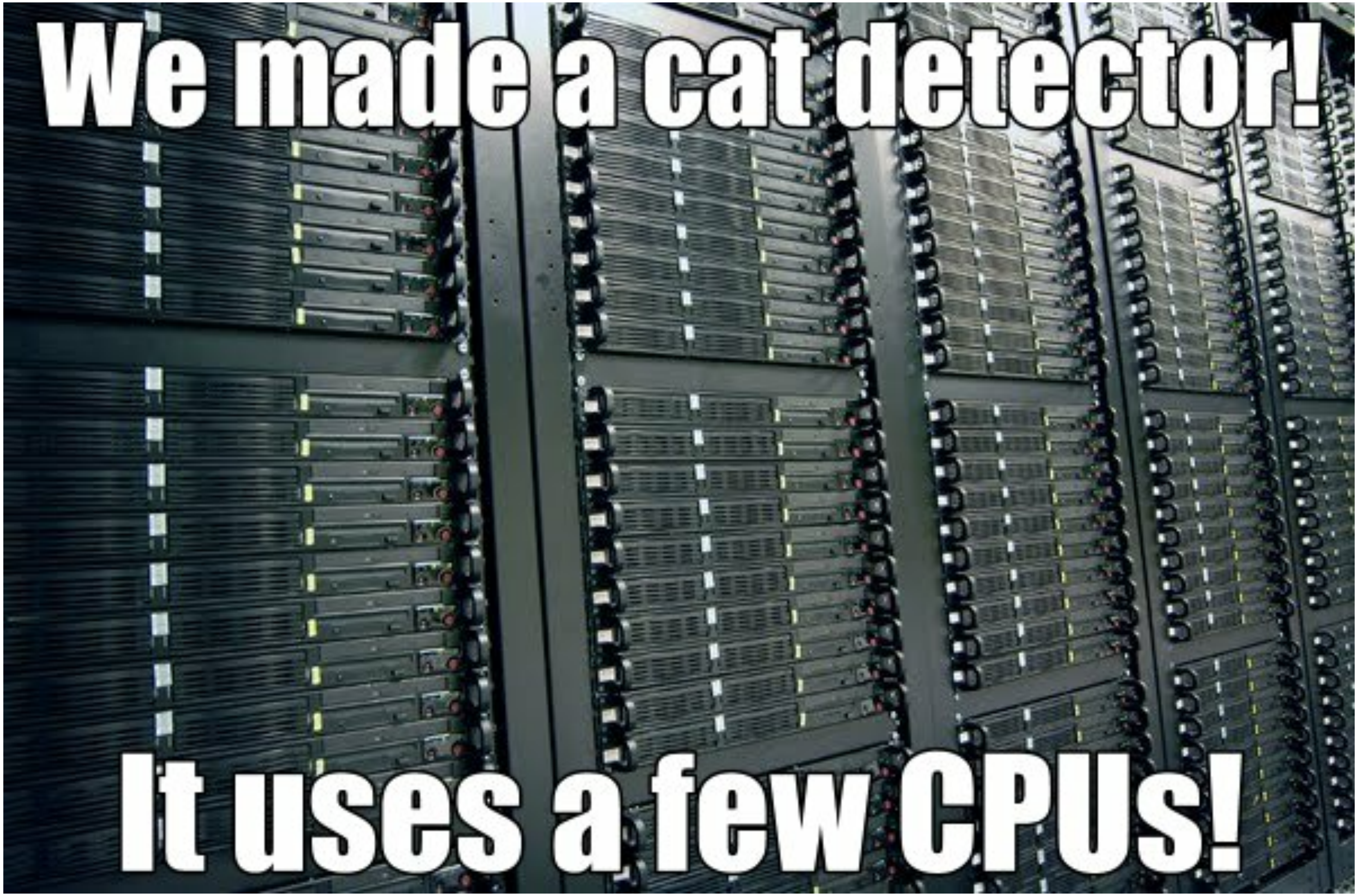


Purely Unsupervised Feature Learning in Images

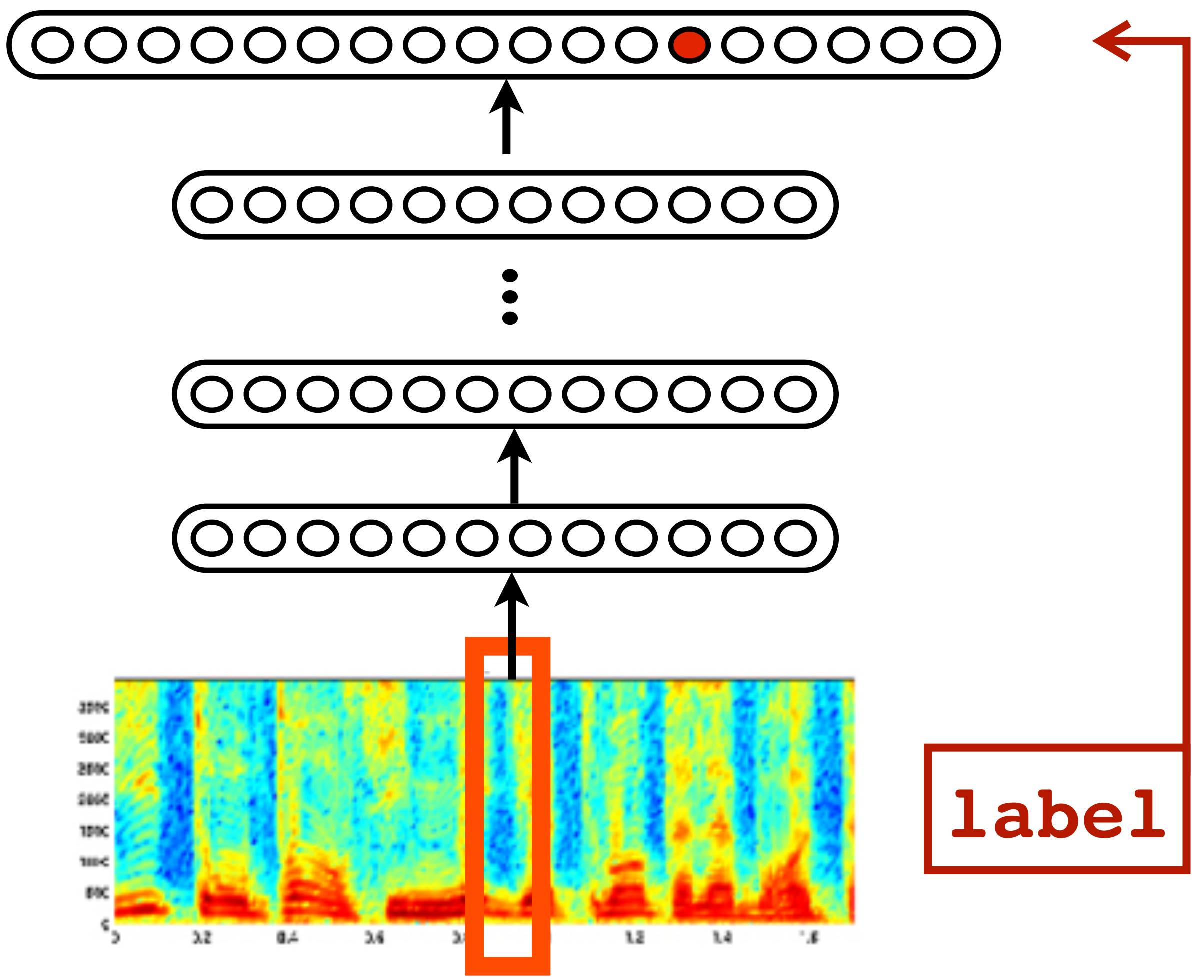
It is YouTube... We also have a cat neuron!

Top stimuli from the test set

Optimal stimulus


We made a cat detector!
It uses a few CPUs!

Acoustic Modeling for Speech Recognition



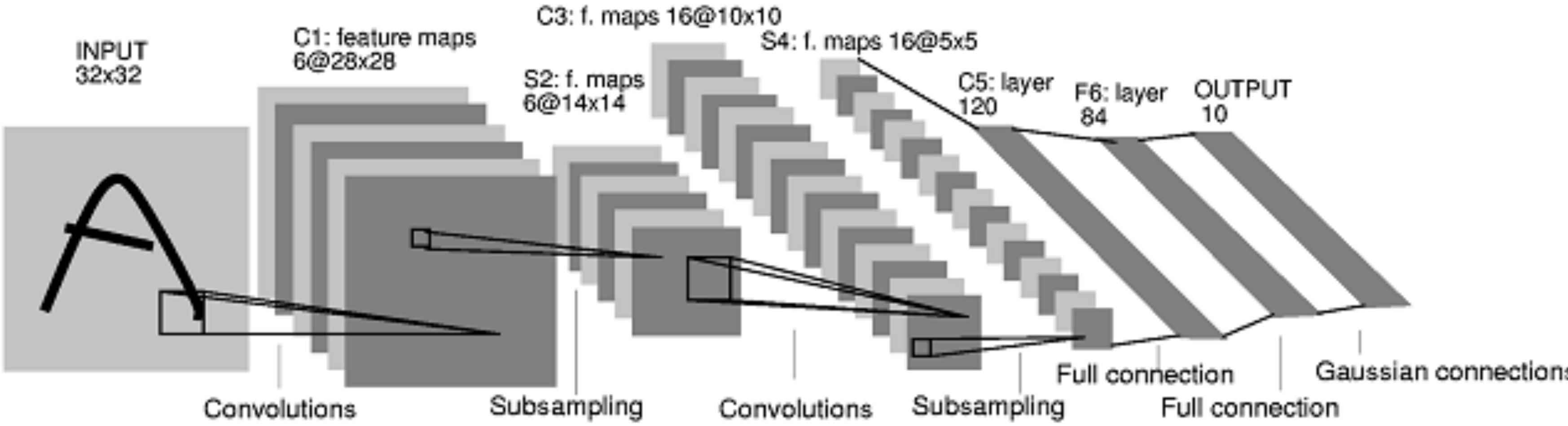
Close collaboration with Google Speech team

Trained in <5 days on cluster of 800 machines

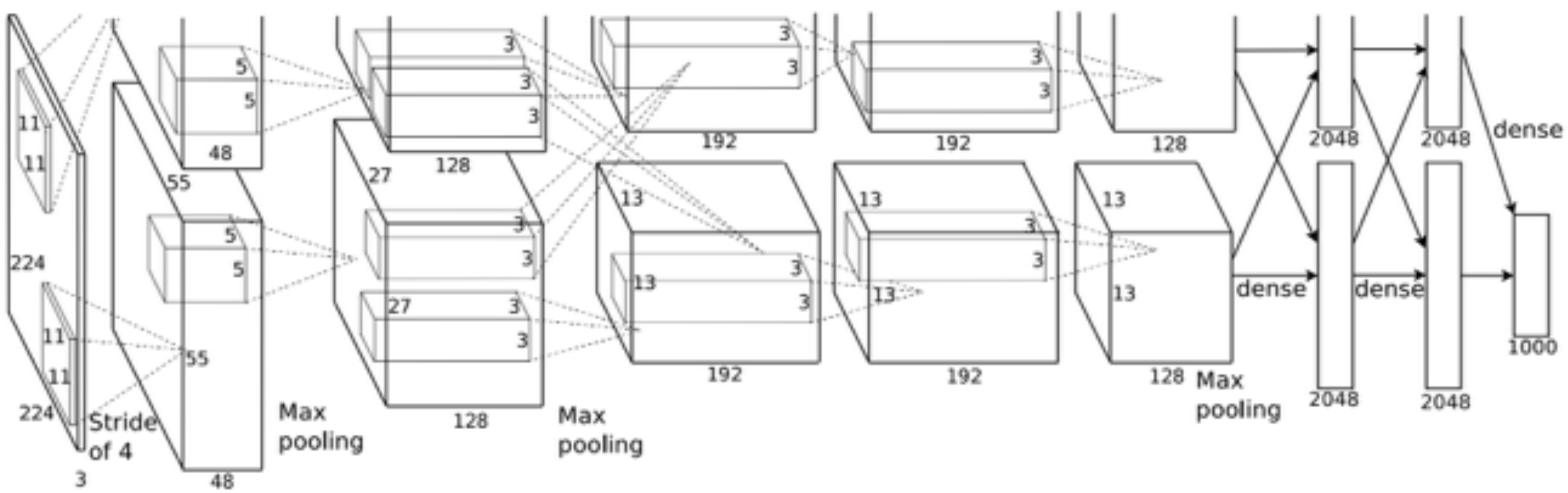
30% reduction in Word Error Rate for English
("biggest single improvement in 20 years of speech research")

Launched in 2012 at time of Jellybean release of Android

Convolutional Models for Image Classification



LeCun *et al.*, 1989



Krizhevsky *et al.*, NIPS 2012

2012 ImageNet winner: **16.4%** top-5 error rate

GoogLeNet

2014 ImageNet winner:
6.66% top-5 error rate

Improvement Happening Rapidly

	Top 5 error
Imagenet 2011 winner (not CNN)	25.7%
Imagenet 2012 winner	16.4% (Krizhesvky et al.)
Imagenet 2013 winner	11.7% (Zeiler/Clarifai)
Imagenet 2014 winner	6.7% (GoogLeNet)
Human: Andrej Karpathy	5.1%
Baidu Arxiv paper: 3 Jan ‘15	6.0%
MS Research Arxiv paper: 6 Feb ‘15	4.9%
Google Arxiv paper: 2 Mar ‘15	4.8%

Good Fine-grained Classification

“hibiscus”

“dahlia”

Good Generalization

Both recognized as a "meal"

Sensible Errors

“snake”

“dog”

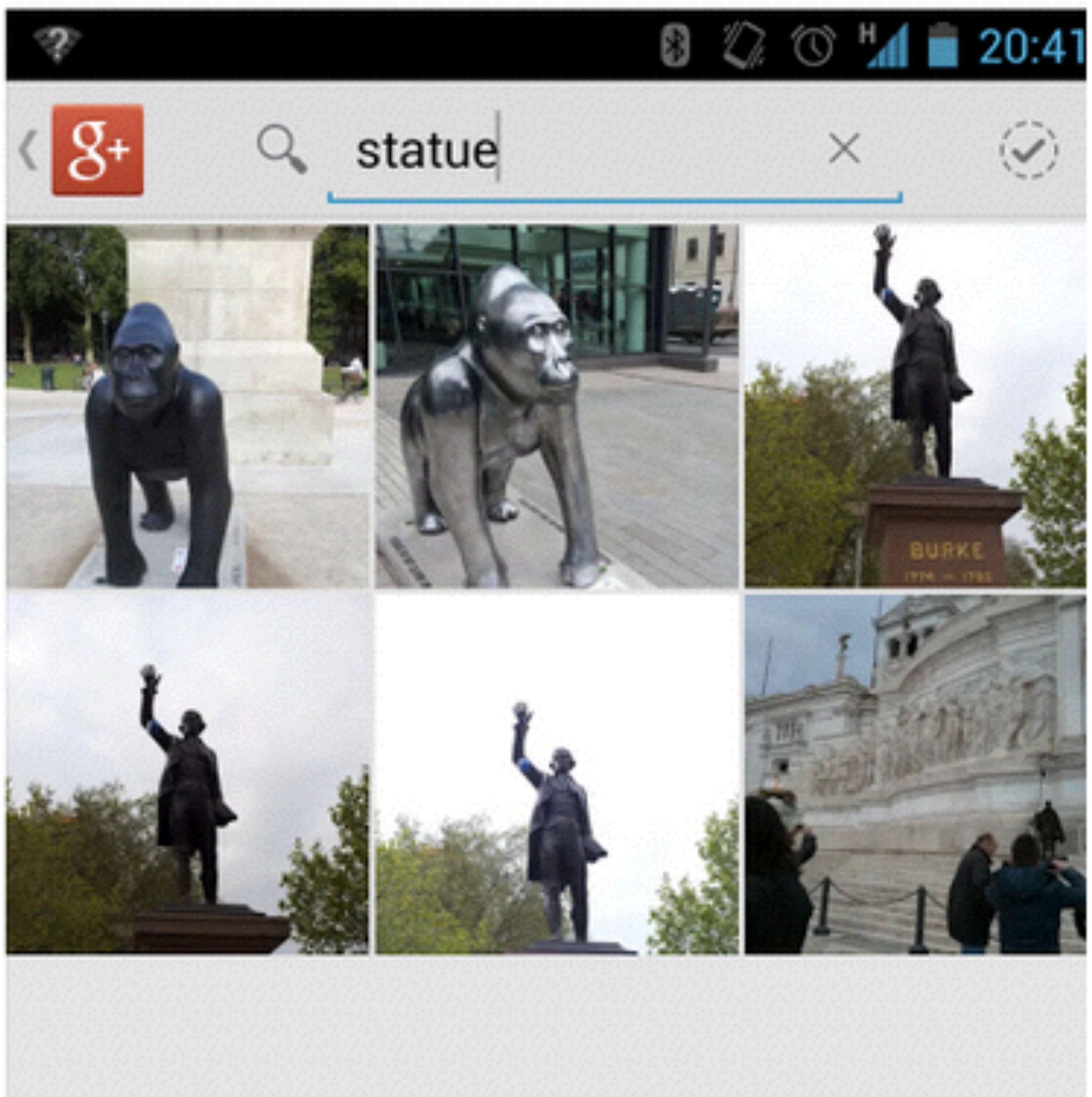
Works in practice

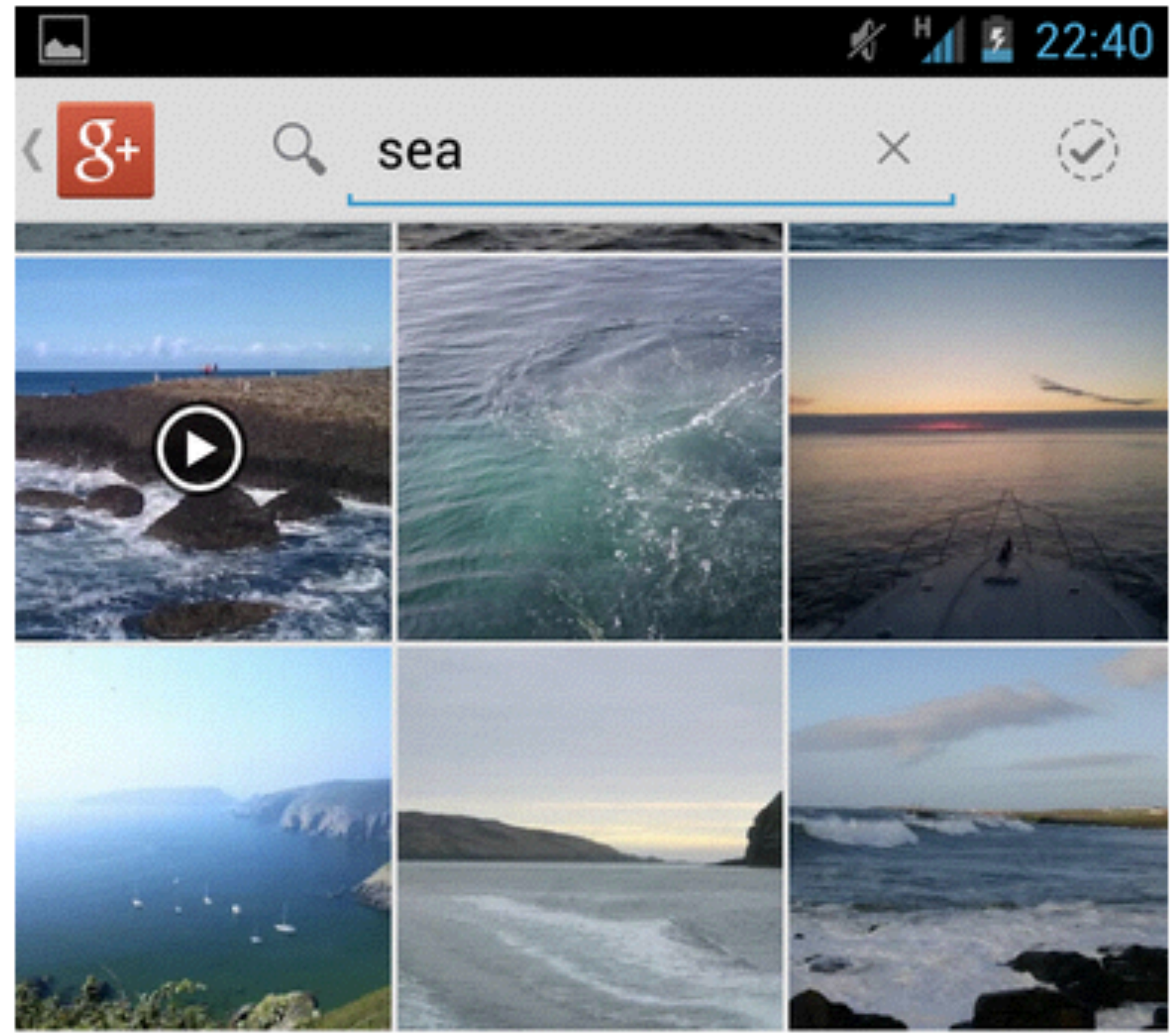
for real users.

Wow.

The new Google plus photo search is a bit insane.

I didn't tag those... :)

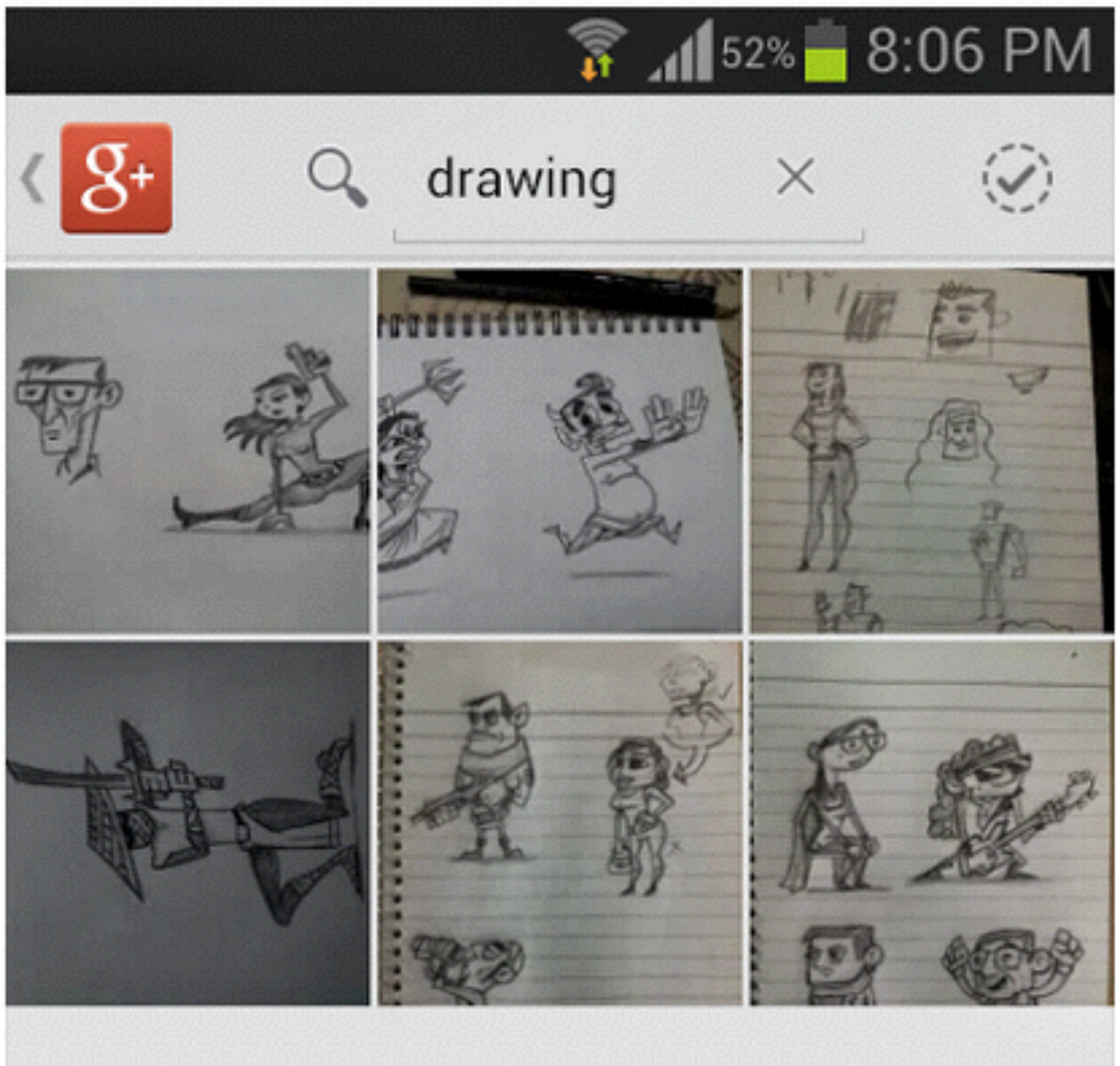


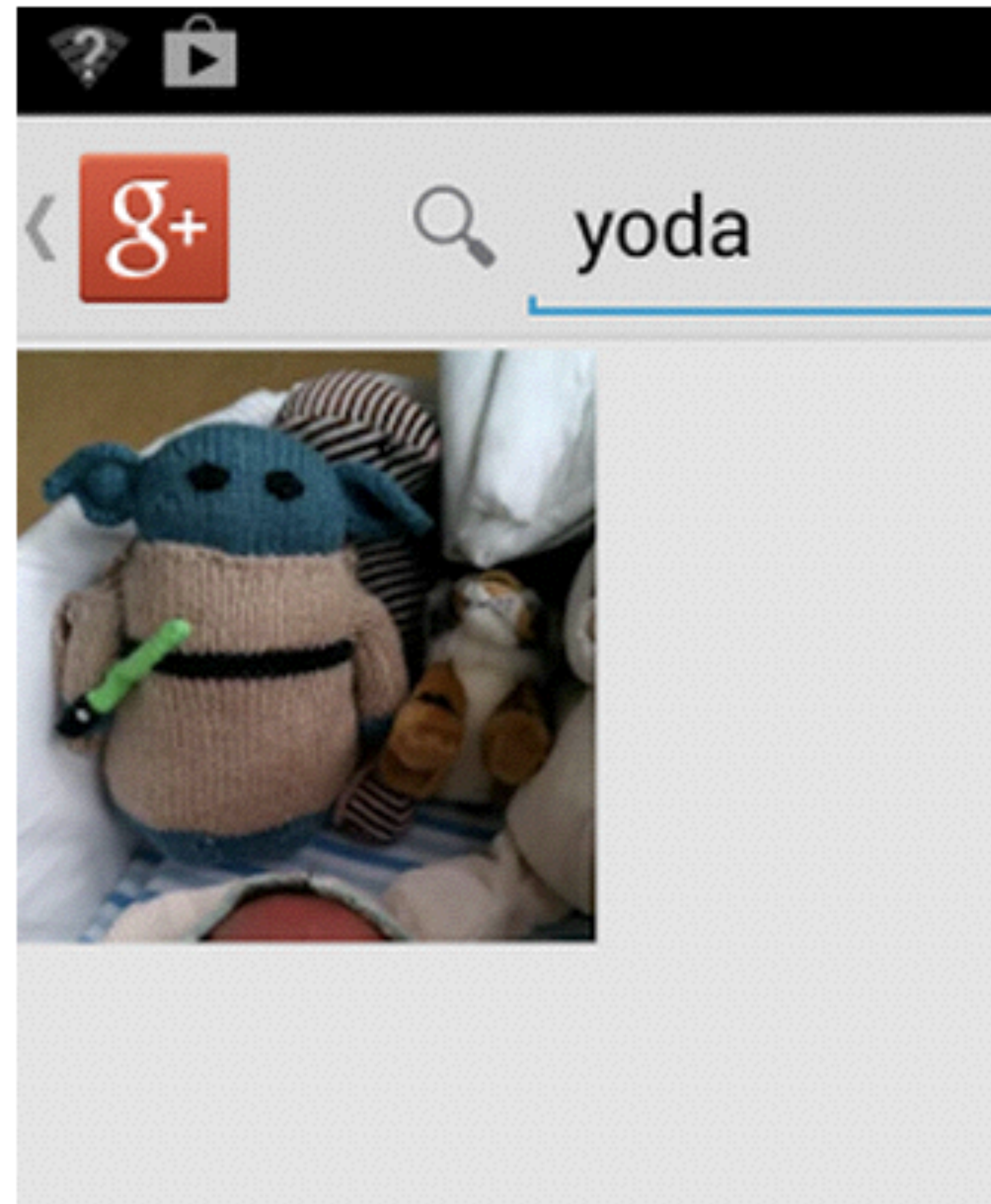


Works in practice

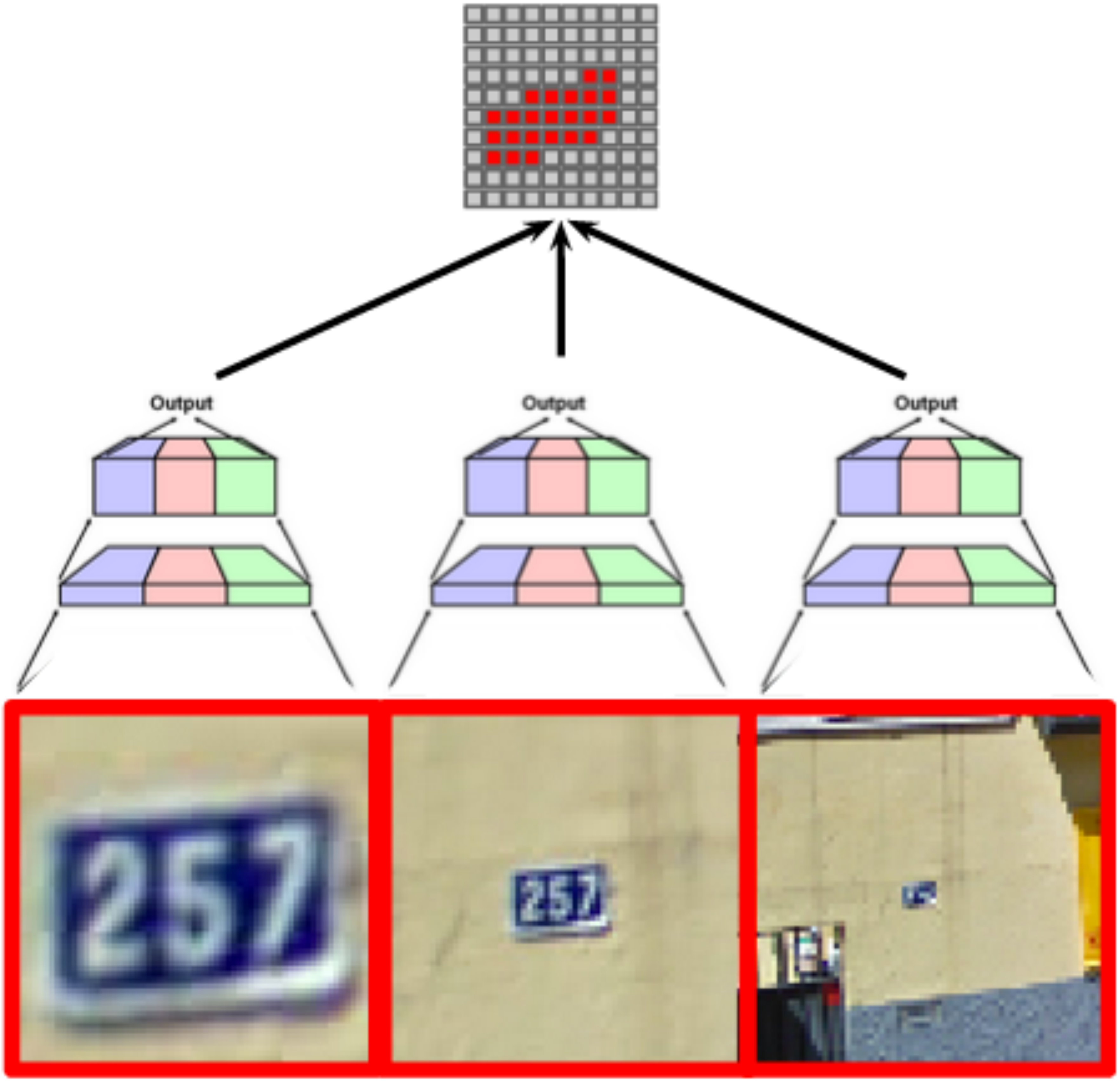
for real users.

Google Plus photo search is awesome. Searched with keyword 'Drawing' to find all my scribbles at once :D





Text?

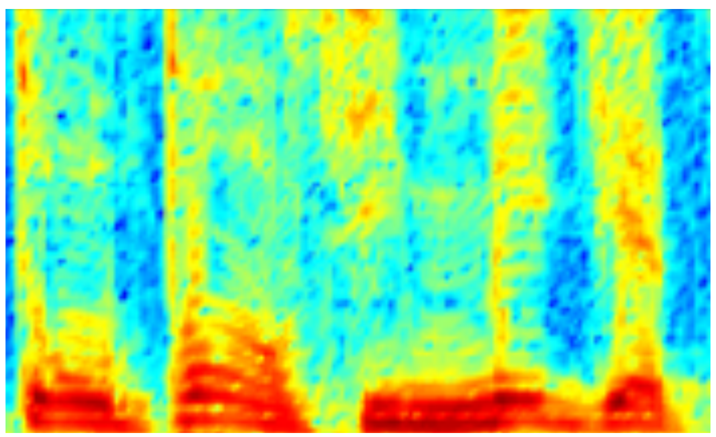


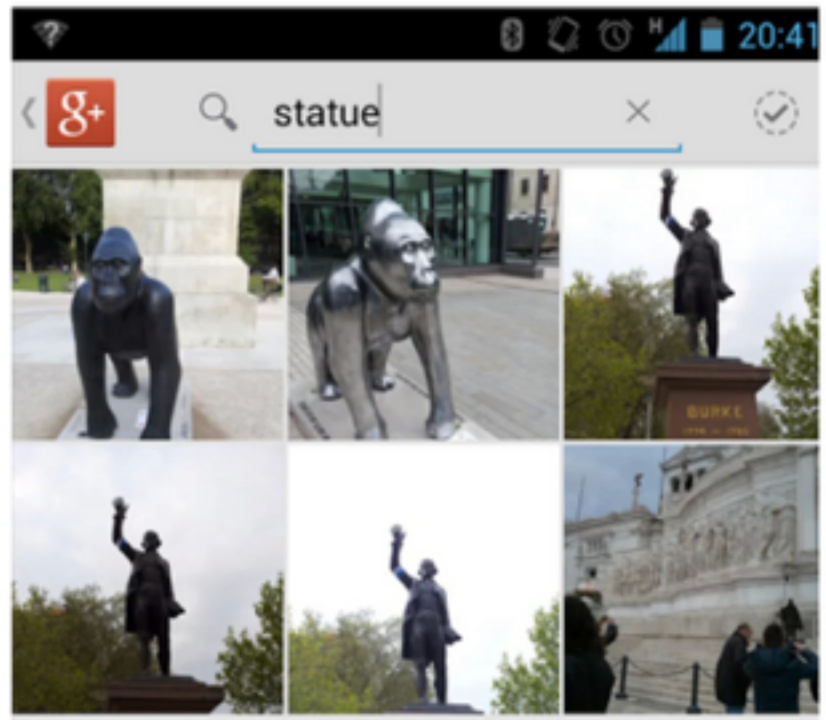
Work by Matt Zeiler (summer intern), Julian Ibarz and Jeff Dean


ASIAWIDE TRAVEL
Tel: (02) 9745 3355 1ST Floor, 240 BURWOOD RD
aria's Bakery Inn 超羣餅屋
Maria's Bakery Inn 超羣餅屋

Corner
Cubbyhouse
THUMP
www.thumphq.com
CIANO MOTOR ENGINEERS
CME
MECHANICAL REPAIRS TO ALL MAKES AND MODELS
Specialising In BMW, MINI & TOYOTA
8 REGATTA ROAD FIVE DOCK 9745 3173
88
LATEST DIAGNOSTIC EQUIPMENT • REGO INSPECTIONS
NEW CAR/LOGBOOK SERVICING • BRAKES • CLUTCHES
STEERING • SUSPENSION • TYRES • WHEEL ALIGNMENTS
RADIATORS • AIR CONDITIONING • EFI TUNING
FUEL INJECTION SERVICING • BATTERIES • AUTO ELECTRICAL
Factory Trained Technicians

Deep neural networks have proven themselves across a range of supervised learning tasks involve dense input features.



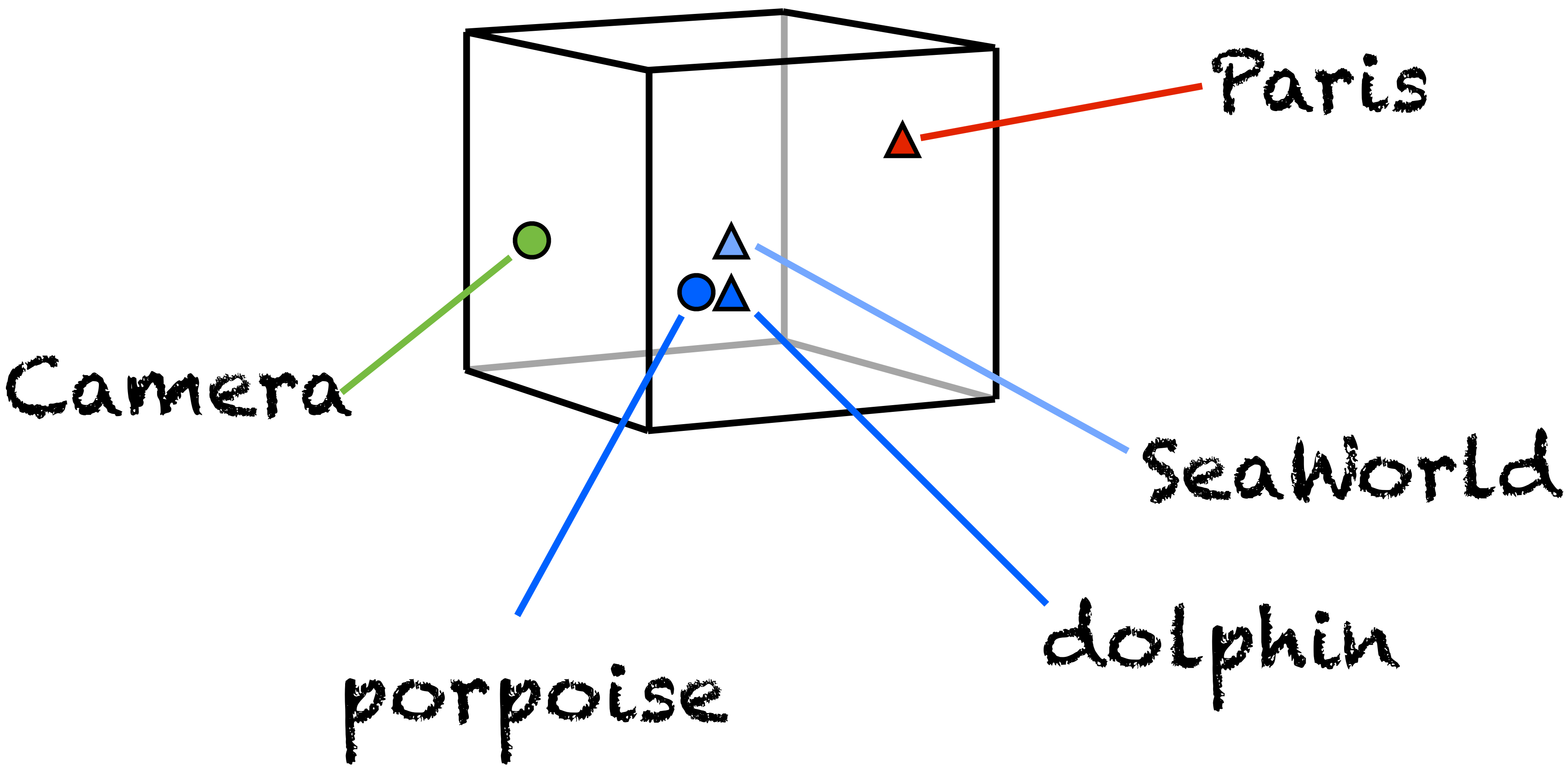


What about domains with sparse input data?

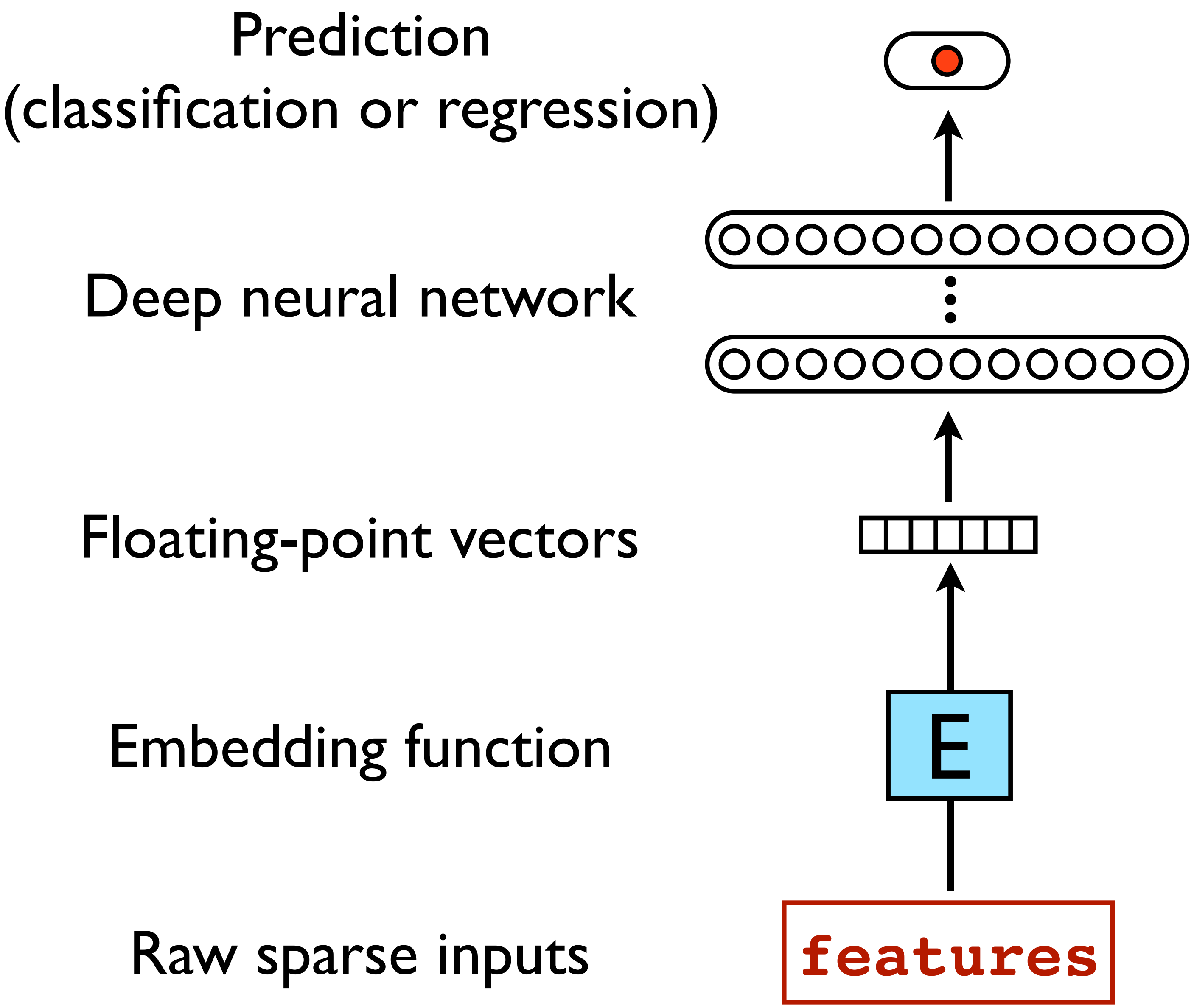
How can DNNs possibly deal with sparse data?

Answer: Embeddings

~1000-D joint embedding space

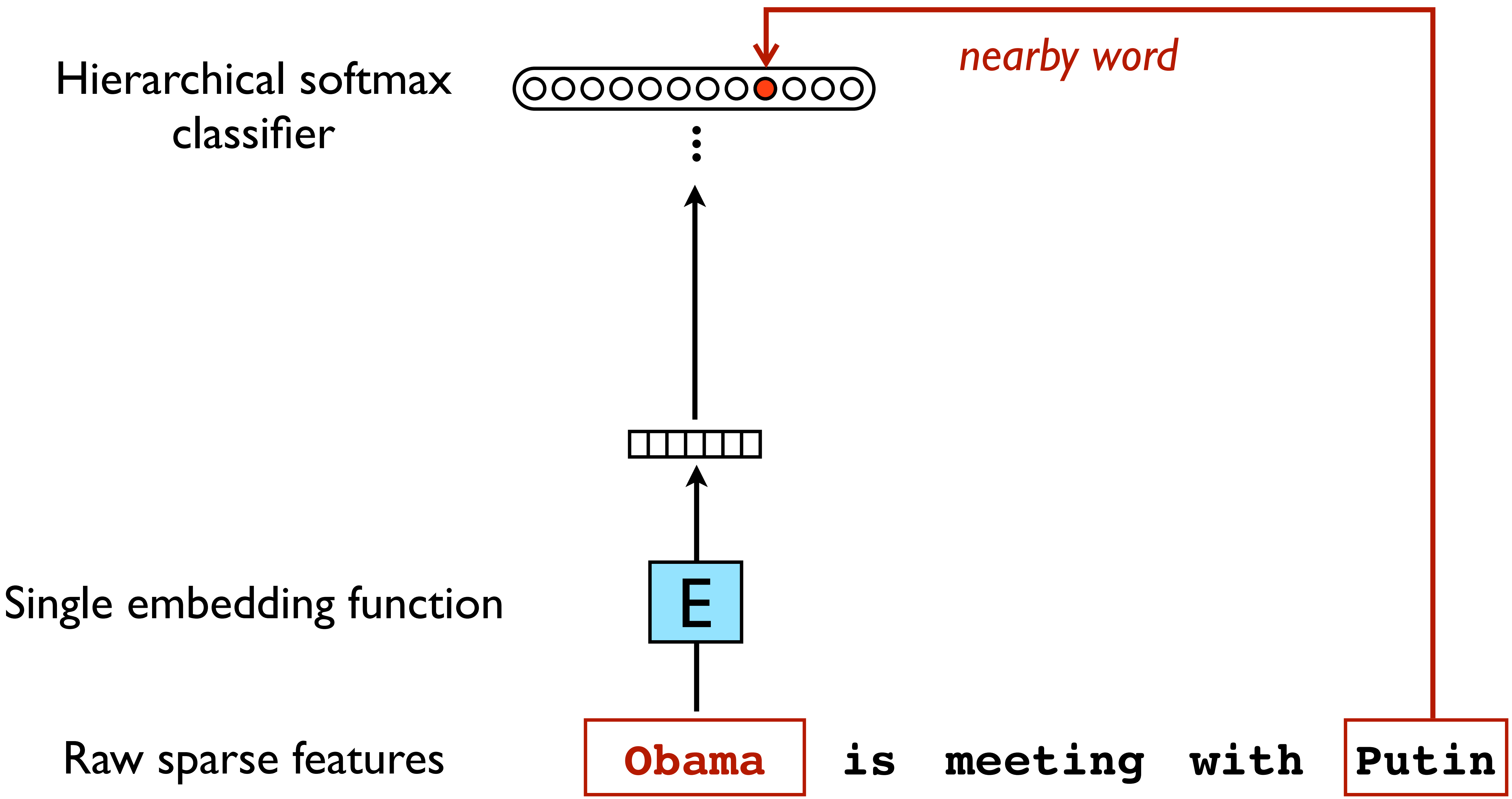


How Can We Learn the Embeddings?



How Can We Learn the Embeddings?

Skipgram Text Model

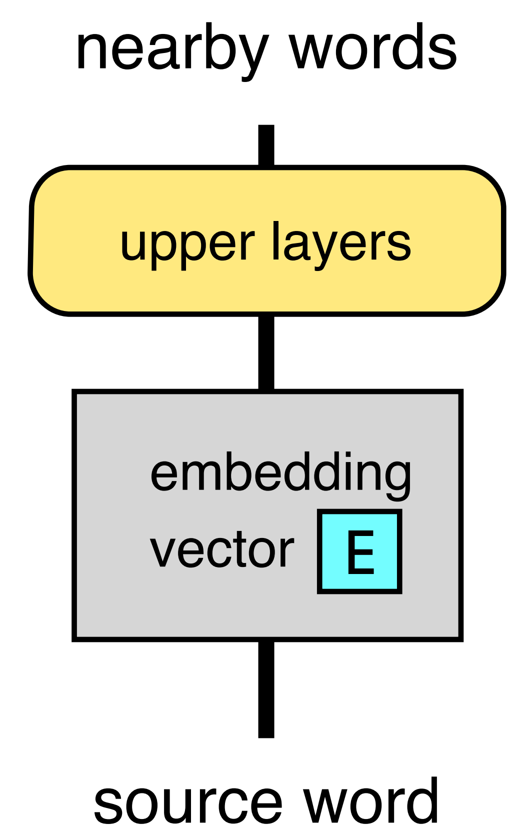


Mikolov, Chen, Corrado and Dean. *Efficient Estimation of Word Representations in Vector Space*, http://arxiv.org/abs/1301.3781.

Nearest neighbors in language embeddings space are closely related semantically.

- Trained skip-gram model on Wikipedia corpus.

tiger_shark	**car**	**new_york**
bull_shark	cars	new_york_city
blacktip_shark	muscle_car	brooklyn
shark	sports_car	long_island
oceanic_whitetip_shark	compact_car	syracuse
sandbar_shark	autocar	manhattan
dusky_shark	automobile	washington
blue_shark	pickup_truck	bronx
requiem_shark	racing_car	yonkers
great_white_shark	passenger_car	poughkeepsie
lemon_shark	dealership	new_york_state



* 5.7M docs, 5.4B terms, 155K unique terms, 500-D embeddings

Solving Analogies

- Embedding vectors trained for the language modeling task have very interesting properties (especially the skip-gram model).

E(*hotter*) - E(*hot*) ≈ E(*bigger*) - E(*big*)

E(*Rome*) - E(*Italy*) ≈ E(*Berlin*) - E(*Germany*)

Solving Analogies

- Embedding vectors trained for the language modeling task have very interesting properties (especially the skip-gram model).

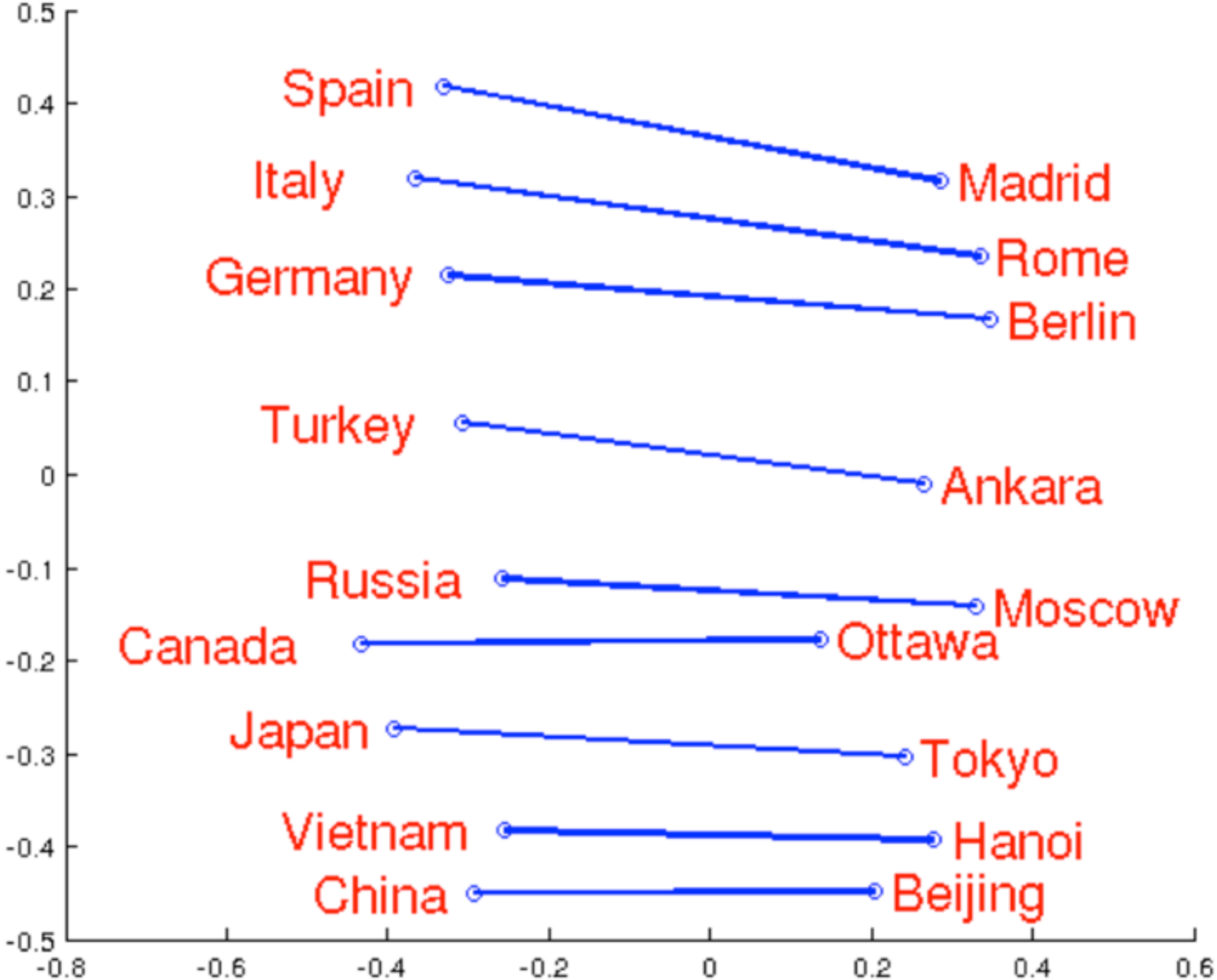
E(*hotter*) - E(*hot*) + E(*big*) ≈ E(*bigger*)

E(*Rome*) - E(*Italy*) + E(*Germany*) ≈ E(*Berlin*)

Skip-gram model w/ 640 dimensions trained on 6B words of news text achieves 57% accuracy for analogy-solving test set.

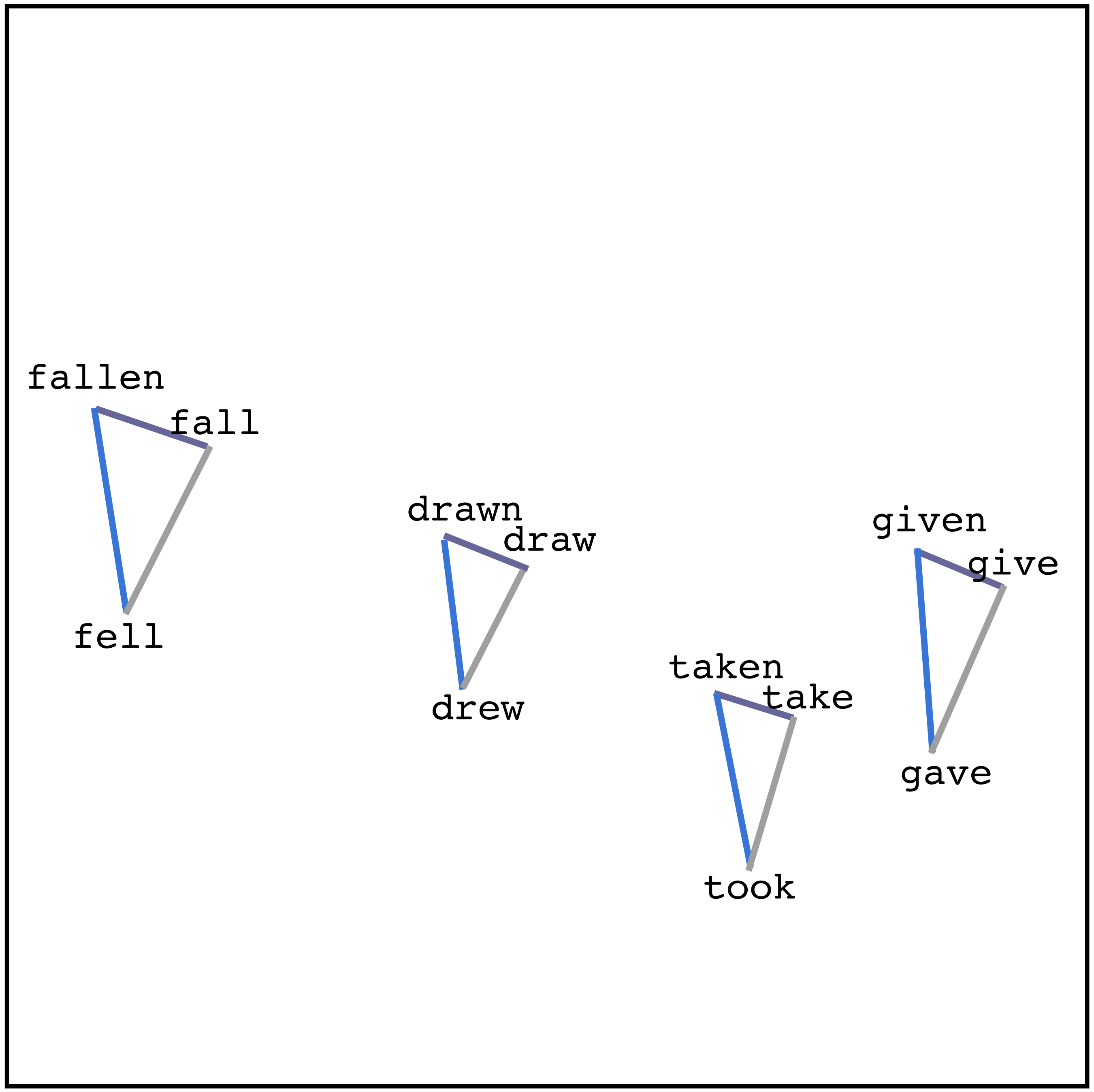
Visualizing the Embedding Space

Projected down from 640 dimensions to 2 dimensions via Principal Components Analysis (PCA)



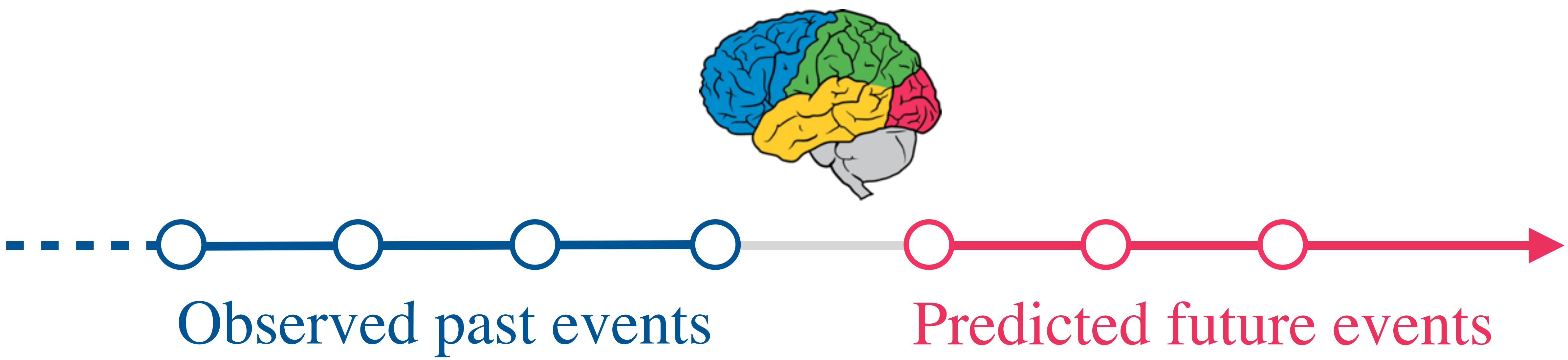
Embeddings are Powerful

Projected down from 640 dimensions to 2 dimensions via Principal Components Analysis (PCA)



Sequence Prediction

Given a sequence of events so far, guess what will follow.

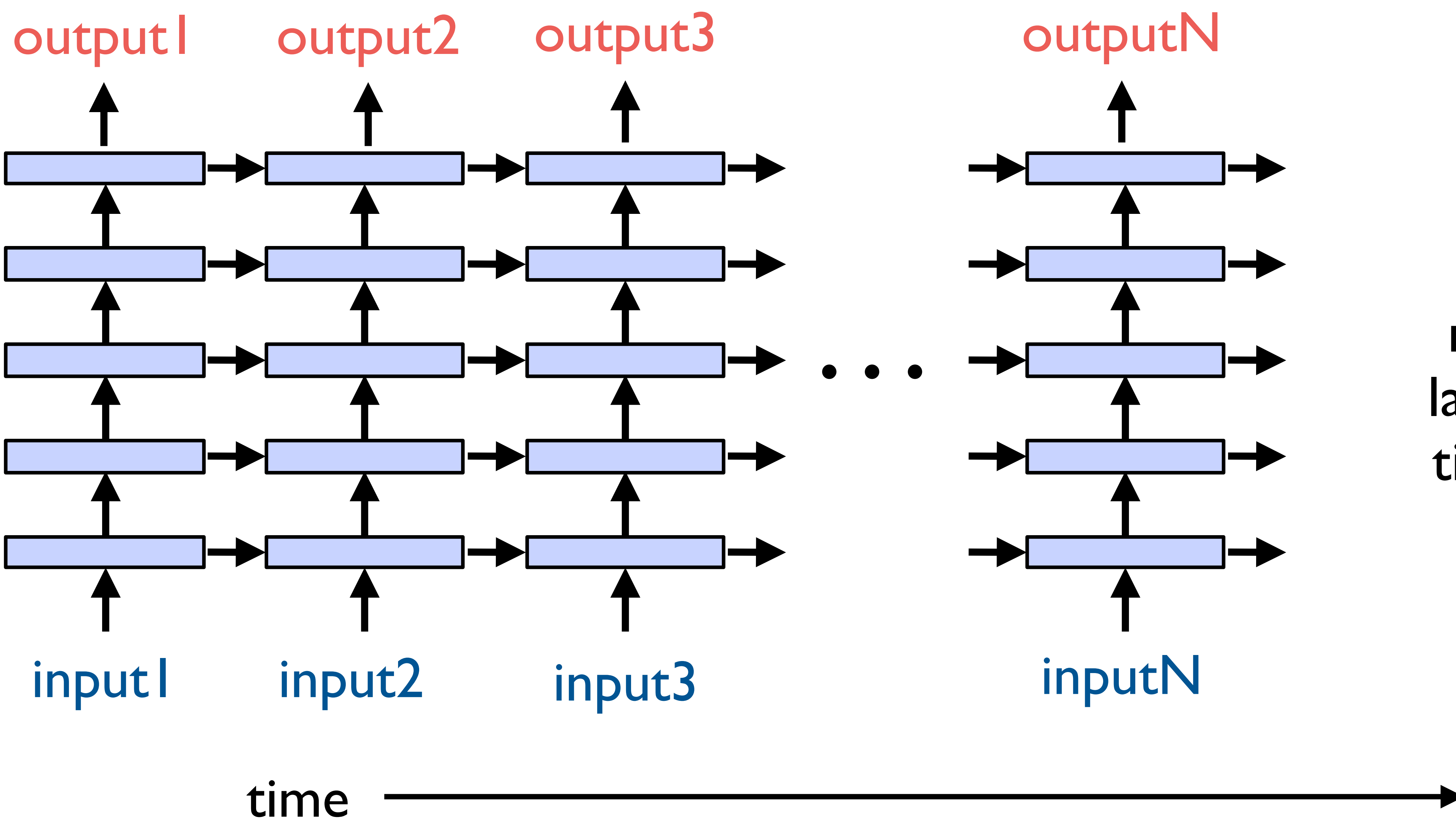


Seems simple. But a surprisingly broad problem framing.

Ilya Sutskever, Oriol Vinyals, Quoc Le.
Sequence to Sequence Learning with Neural Networks. NIPS, 2014.

Sequence Prediction Model

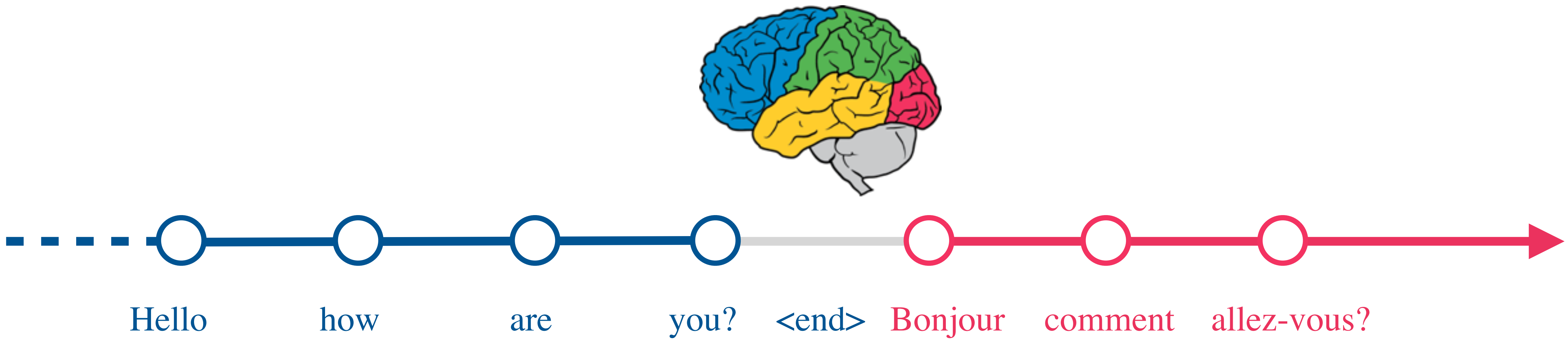
Recurrent neural network (LSTM)



deep LSTM: multiple layers per time step

Ilya Sutskever, Oriol Vinyals, Quoc Le.
Sequence to Sequence Learning with Neural Networks. NIPS, 2014.

Translation with Sequence Prediction



Approach gives state-of-the-art results on public WMT translation task/ dataset

Ilya Sutskever, Oriol Vinyals, Quoc Le.
Sequence to Sequence Learning with Neural Networks. NIPS, 2014.

Google

Conversation with Sequence Prediction

Given a conversation with computer tech support so far, model can complete the tech support rep's sentence.

Customer: Hi.
TechSupport: Hi, this is Andrew from Techstop Connect, how can I help?
Customer: I cannot connect to VPN.
TechSupport: When did...

... you last successfully connect to VPN?

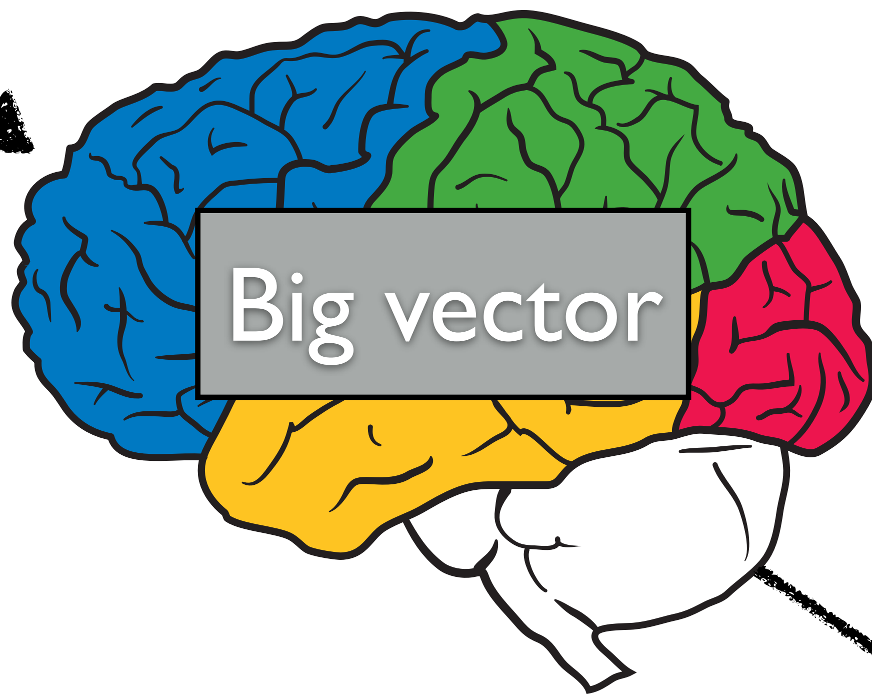
Predictions reflect situational context.

Ilya Sutskever, Oriol Vinyals, Quoc Le.
Sequence to Sequence Learning with Neural Networks. NIPS, 2014.

Google

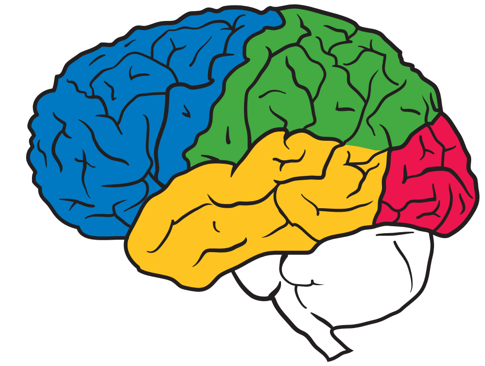
Example of LSTM-based representation: Machine Translation

Input: "Cogito ergo sum"

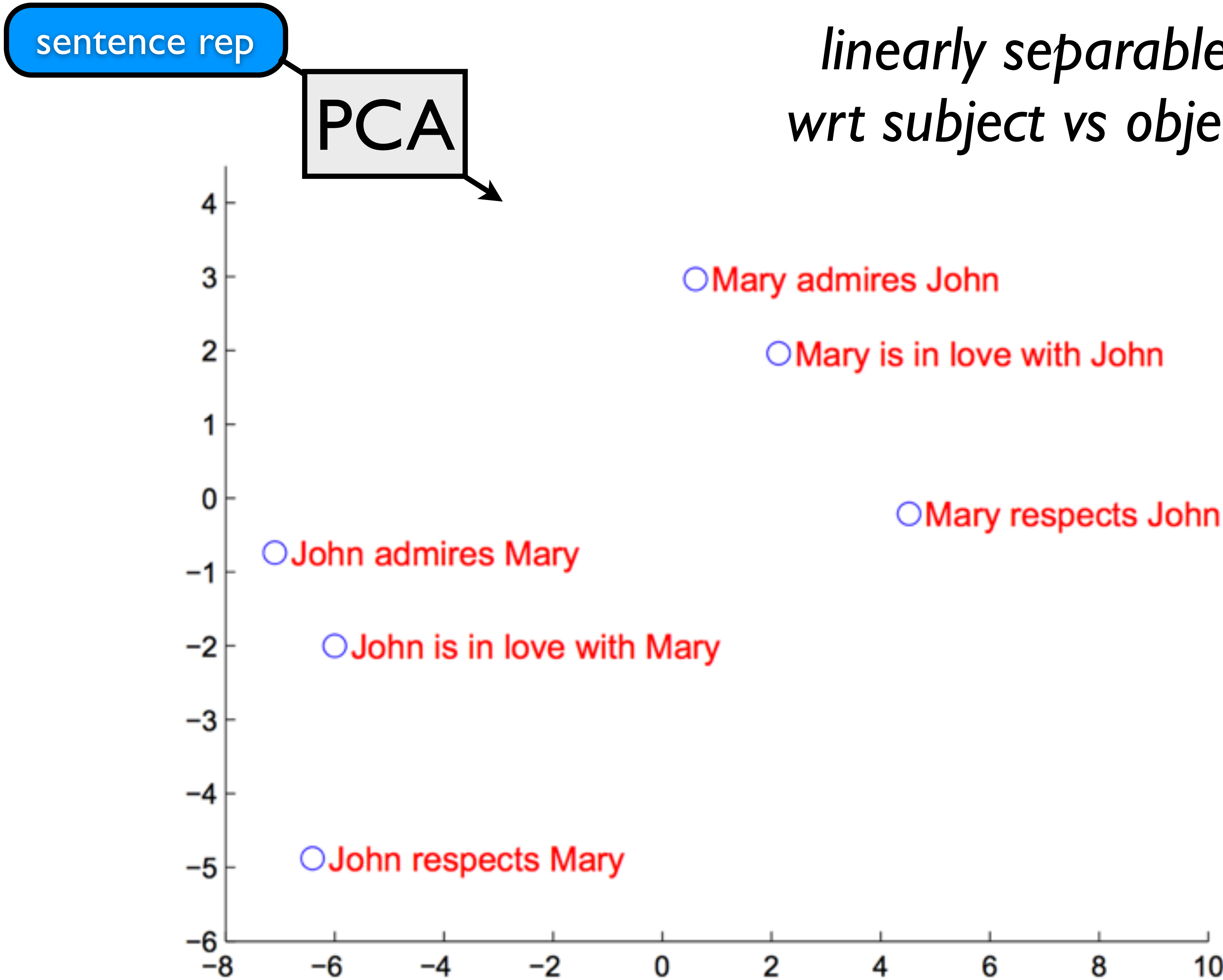


Output: "I think, therefore I am!"

LSTM for End to End Translation

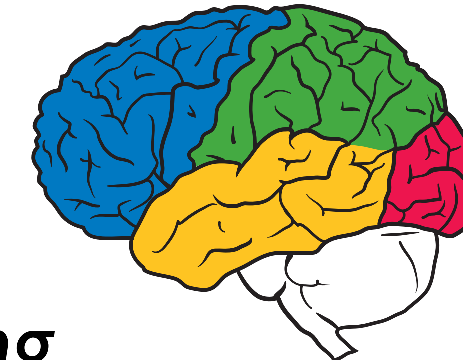


linearly separable
wrt subject vs object



Google

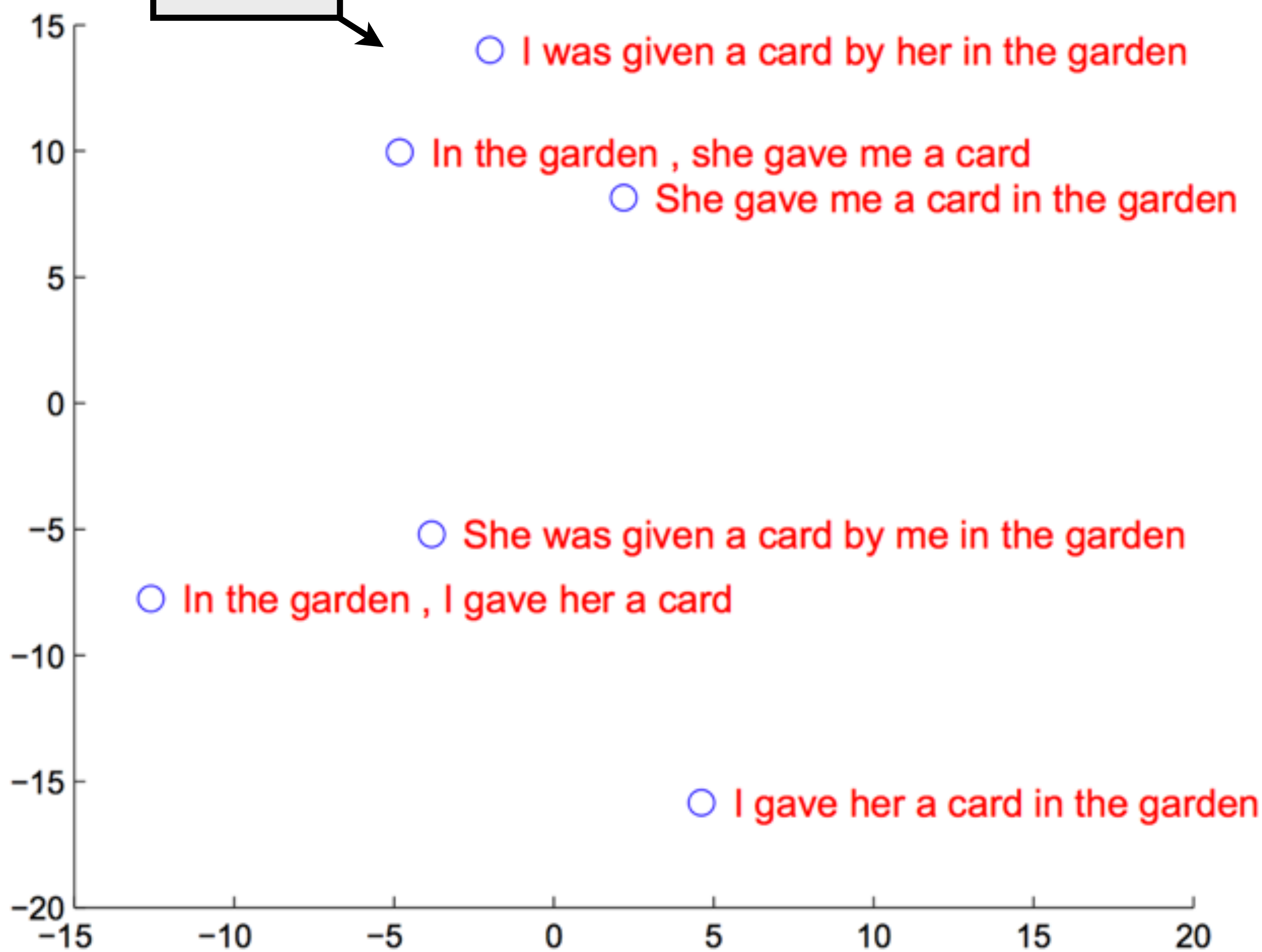
LSTM for End to End Translation



mostly invariant to paraphasing

sentence rep

PCA



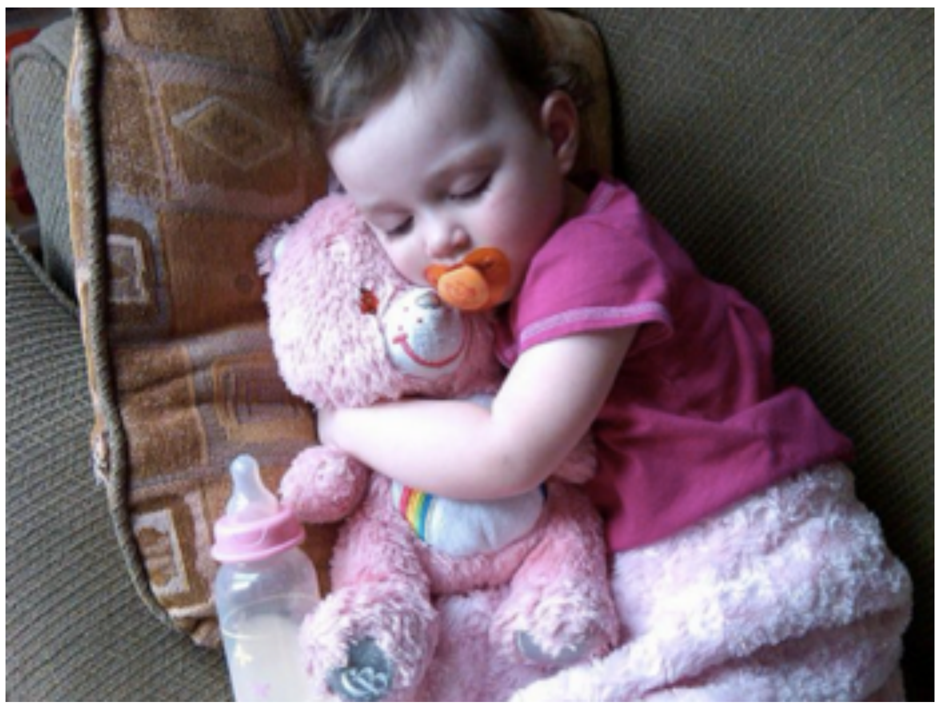
Combining modalities
e.g. vision and language

Google

Captions with Sequence Prediction

Initial state can also come from non-sequence data.

Given a photograph, generate a caption.



Two captions Brain suggests:

"A close up of a child holding a stuffed animal."

"A baby is asleep next to a teddy bear."

This example highlights that these models:

(1) Can handle very complex inputs, and inputs other than text.

(2) Works even in settings with multiple plausible future sequences.

New York Times article: http://nyti.ms/11eGpAW

Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan. Show and Tell: A Neural Image Caption Generator. arXiv, 2014.

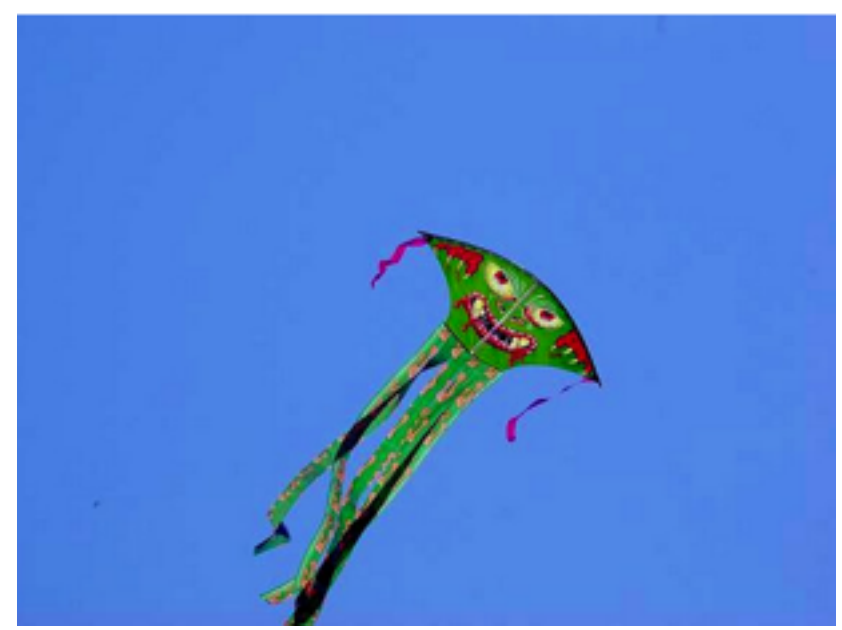
Captions with Sequence Prediction

Given a photograph, we can automatically generate a text caption.

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

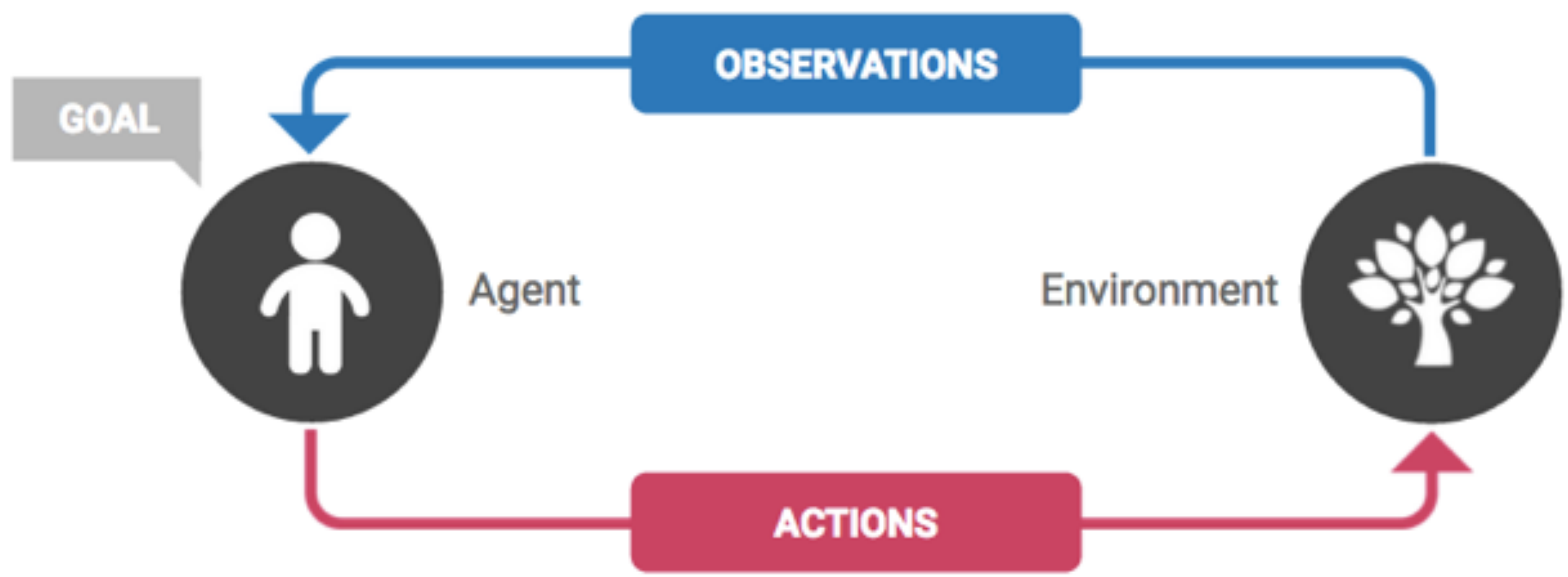


A man flying through the air while riding a snowboard

Learning to Play Atari

- Work done in Google's DeepMind research group in London
- Cover article in *Nature* a couple of weeks ago

- Instead of just classifying, learn to take actions in some environment, and learn from observing the results of those actions



Deep Networks and Reinforcement Learning

Learn automatically from raw inputs, not pre-programmed

Atari 2600 games used as proving ground:

– Agents just get raw pixels+score as inputs (~30k inputs)

– Wired up to action buttons but NOT told what they do

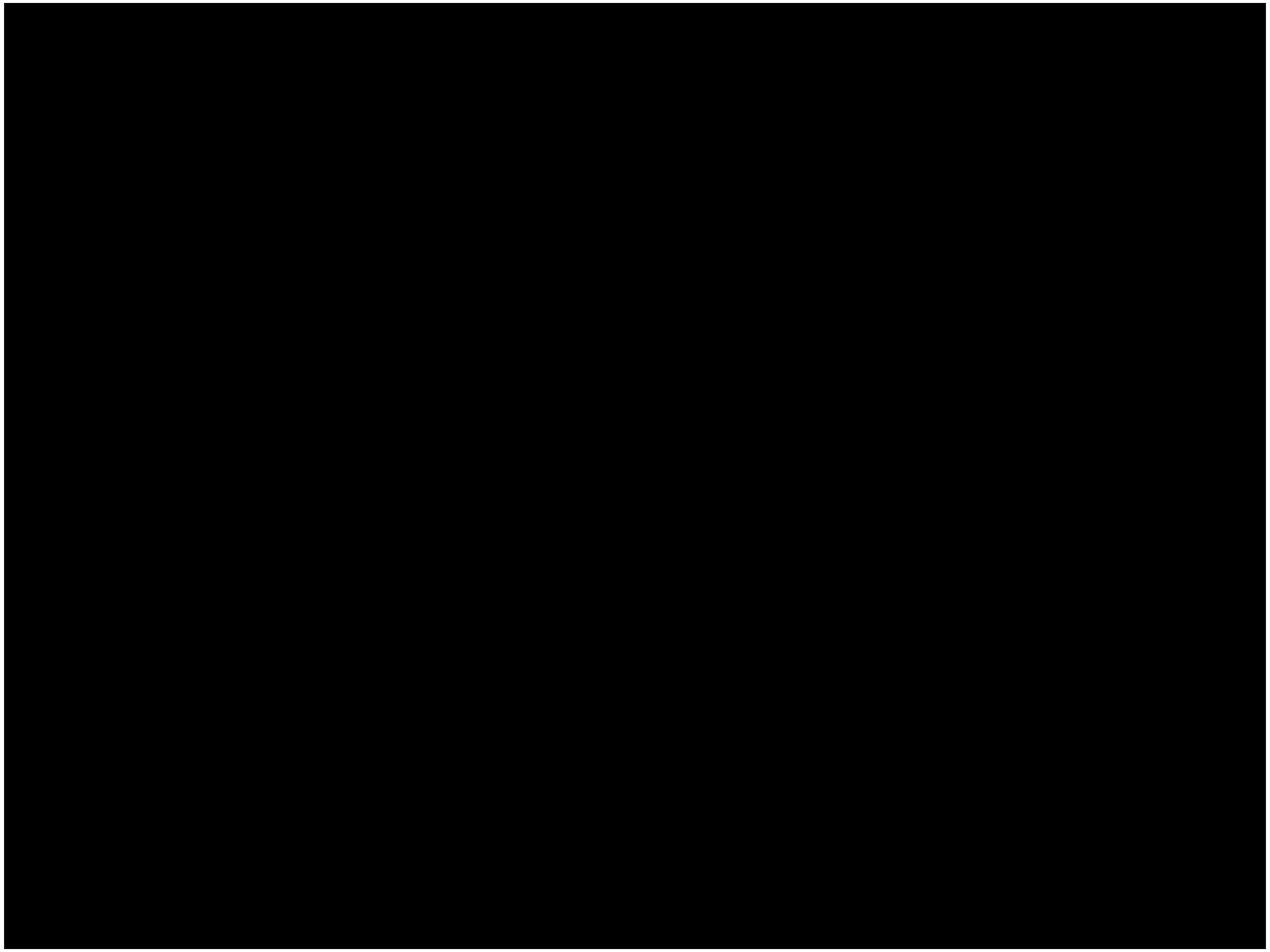
– Goal is simply to maximize the score

Everything learnt from scratch, ZERO prior knowledge

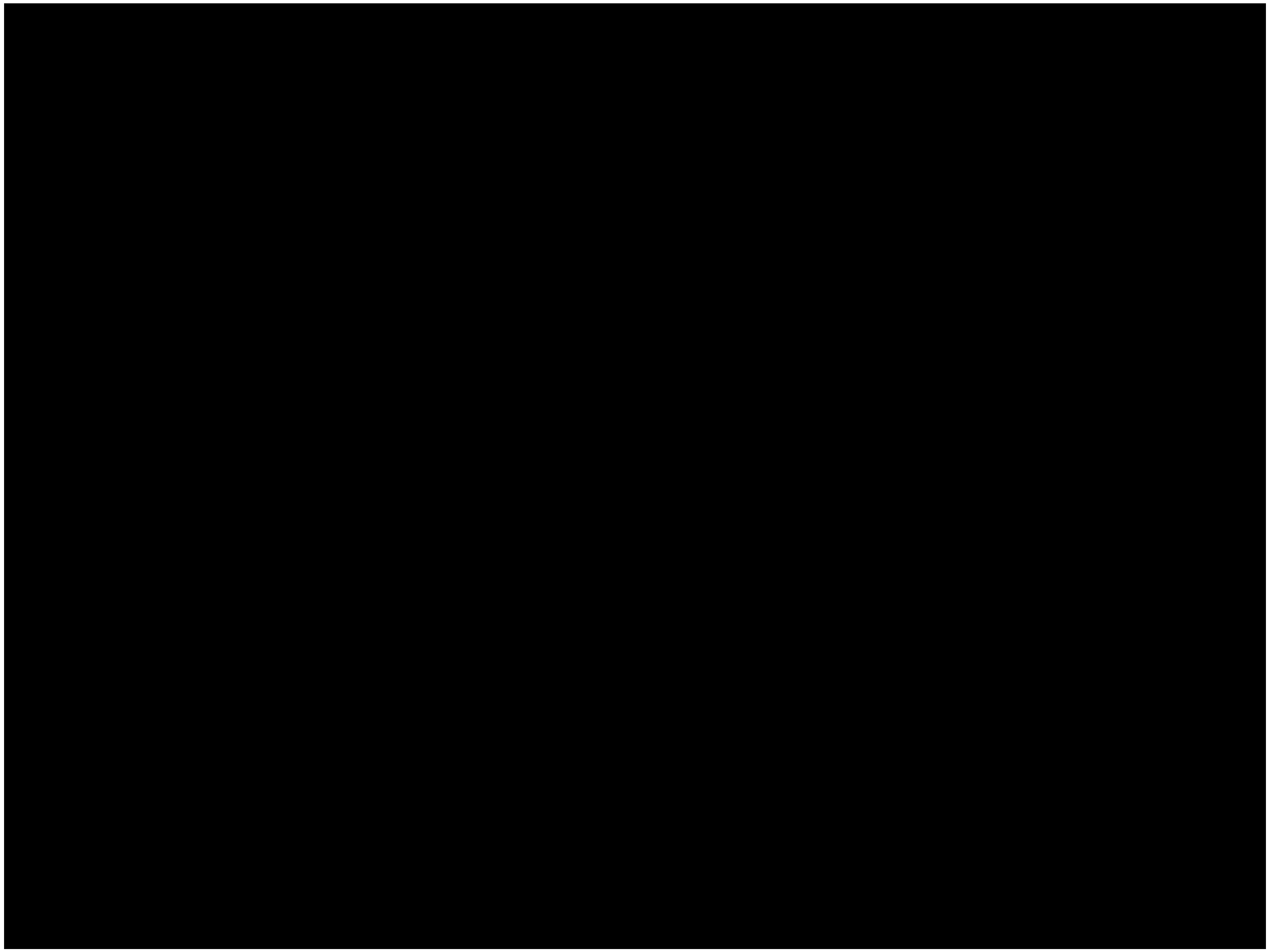
Single system has to master 100s of different games

Space Invaders

Breakout



General Atari Player



Conclusions

- Deep neural networks are very effective for wide range of tasks
  - By using **several kinds of parallelism,** we can quickly train very large and effective deep neural models on very large datasets
  - **Automatically build high-level representations** to solve desired tasks
  - By using **embeddings**, can work with sparse data
  - **Reinforcement learning** can be used to teach agents to **perform complex tasks** that are learned from scratch
  - **Effective in many domains:** speech, vision, language modeling, user prediction, language understanding, translation, advertising, …

An important tool in building intelligent systems.

Google

Questions?

Joint work with many collaborators! Further reading:

- Le, Ranzato, Monga, Devin, Chen, Corrado, Dean, & Ng. *Building High-Level Features Using Large Scale Unsupervised Learning*, ICML 2012.
- Dean, *et al.* , *Large Scale Distributed Deep Networks,* NIPS, 2012.
- Mikolov, Sutskever, Chen, Corrado and Dean. *Distributed Representations of Words and Phrases and their Compositionality,* http://arxiv.org/abs/1310.4546. NIPS, 2013.
- Zeiler, Ranzato, Monga, Mao, Yang, Le, Nguyen, Senior, Vanhoucke, Dean, Hinton. *On Rectified Units for Speech Processing*. ICASSP 2013.
- Heigold, Vanhoucke, Senior, Nguyen, Ranzato, Devin and Dean. *Multilingual Acoustic Models using Distributed Deep Neural Networks*, ICASSP 2013.
- Sutskever, Vinyals, and Le. *Sequence to Sequence Learning with Neural Networks*, http://arxiv.org/abs/1409.3215. NIPS, 2014.
- Vinyals, Toshev, Bengio, and Erhan. *Show and Tell: A Neural Image Caption Generator.* http://arxiv.org/abs/1411.4555

We're having lots of fun and we're hiring!

g.co/ml-jobs

Google