Large Scale Deep Learning

Jeff Dean
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

...

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

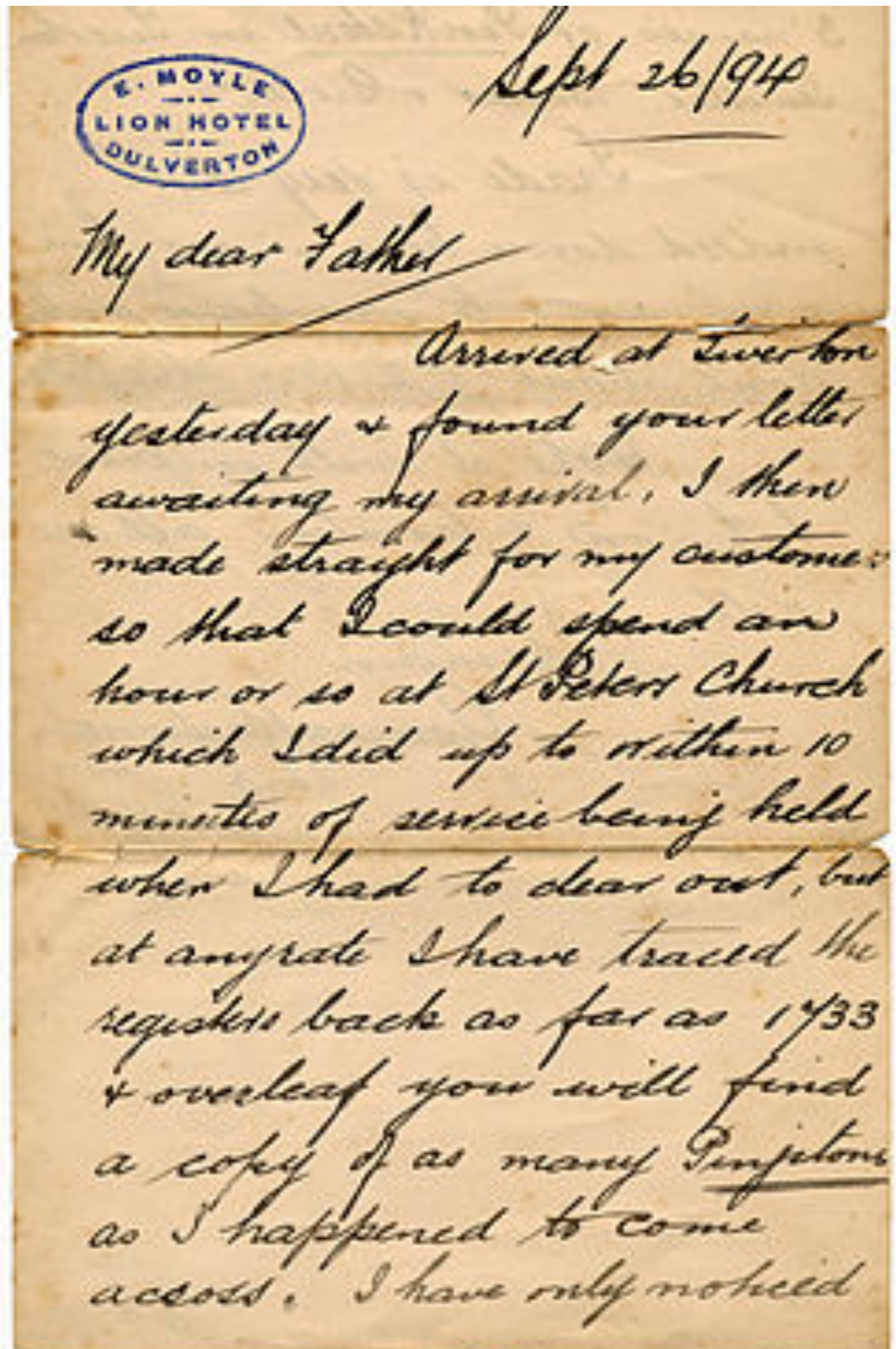
- Text: trillions of words of English + other languages
- Visual: billions of images and videos
- Audio: thousands of hours of speech per day
- User activity: queries, result page clicks, map requests, etc.
- Knowledge graph: billions of labelled relation triples
- ...

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 Pimpstone as I happened to come across. I have only noticed

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

Textual understanding

"This movie should have NEVER been made. From the poorly done animation, to the beyond bad acting. I am not sure at what point the people behind this movie said "Ok, looks good! Lets do it!" I was in awe of how truly horrid this movie was."

General Machine Learning Approaches

- Learning by labeled example: *supervised learning*
  - e.g. An email spam detector
  - amazingly effective if you have lots of examples
- Discovering patterns: *unsupervised learning*
  - e.g. data clustering
  - difficult in practice, but useful if you lack labeled examples
- Feedback right/wrong: *reinforcement learning*
  - e.g. learning to play chess by winning or losing
  - works well in some domains, becoming more important

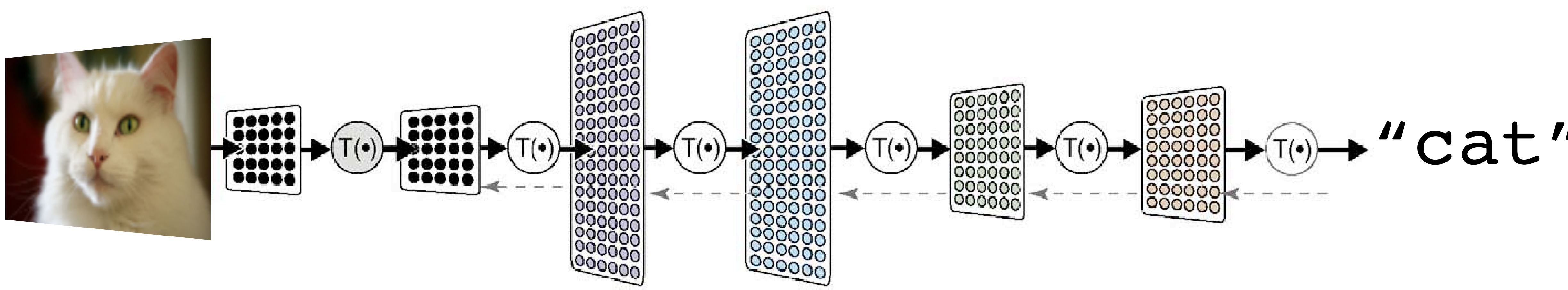
Machine Learning

- For many of these problems, we have lots of data

- Want techniques that minimize software engineering effort
  - simple algorithms, teach computer how to learn from data
  - don't spend time hand-engineering algorithms or high-level features from the raw data

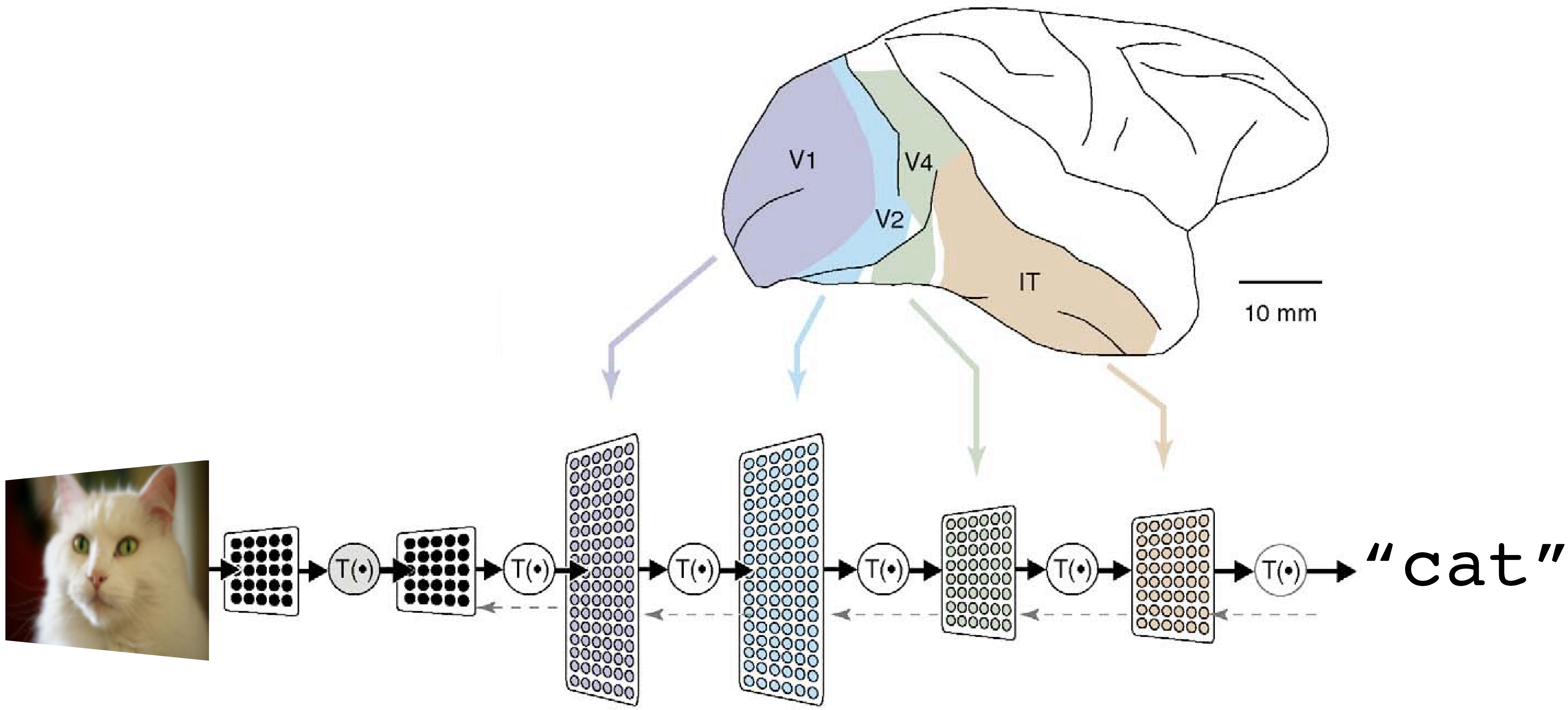
What is Deep Learning?

- 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.



What is Deep Learning?

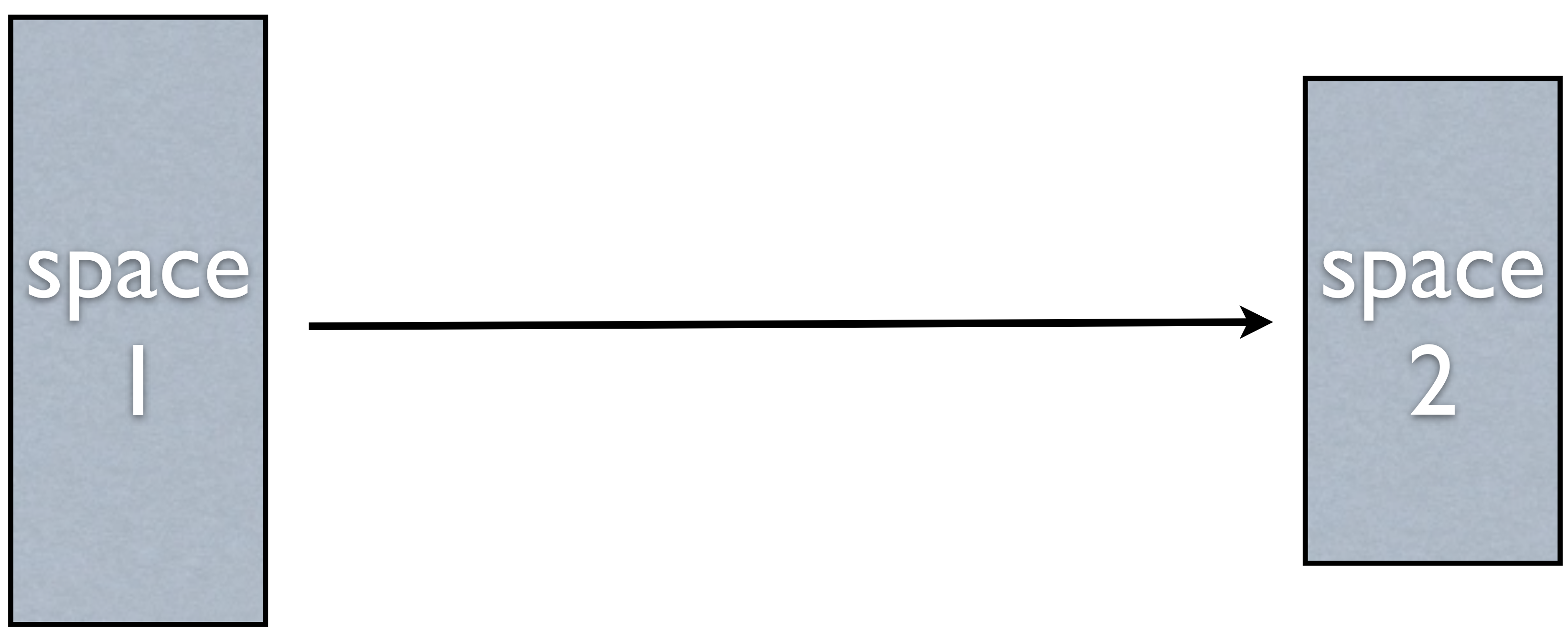
- Loosely inspired by what (little) we know about the biological brain.
- Higher layers form higher levels of abstraction



Google

Neural Networks

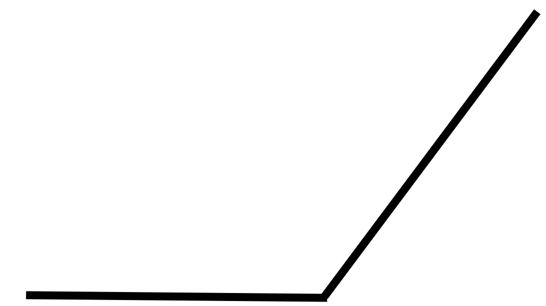
- Learn a complicated function from data



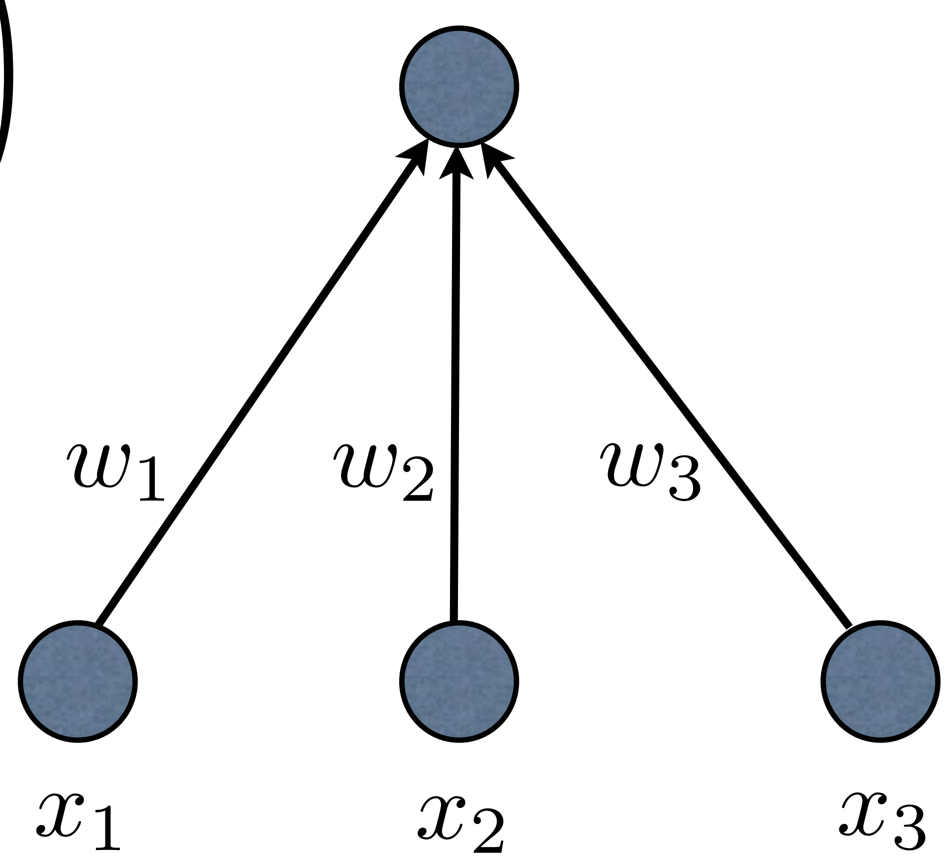
The Neuron

- Different weights compute different functions

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

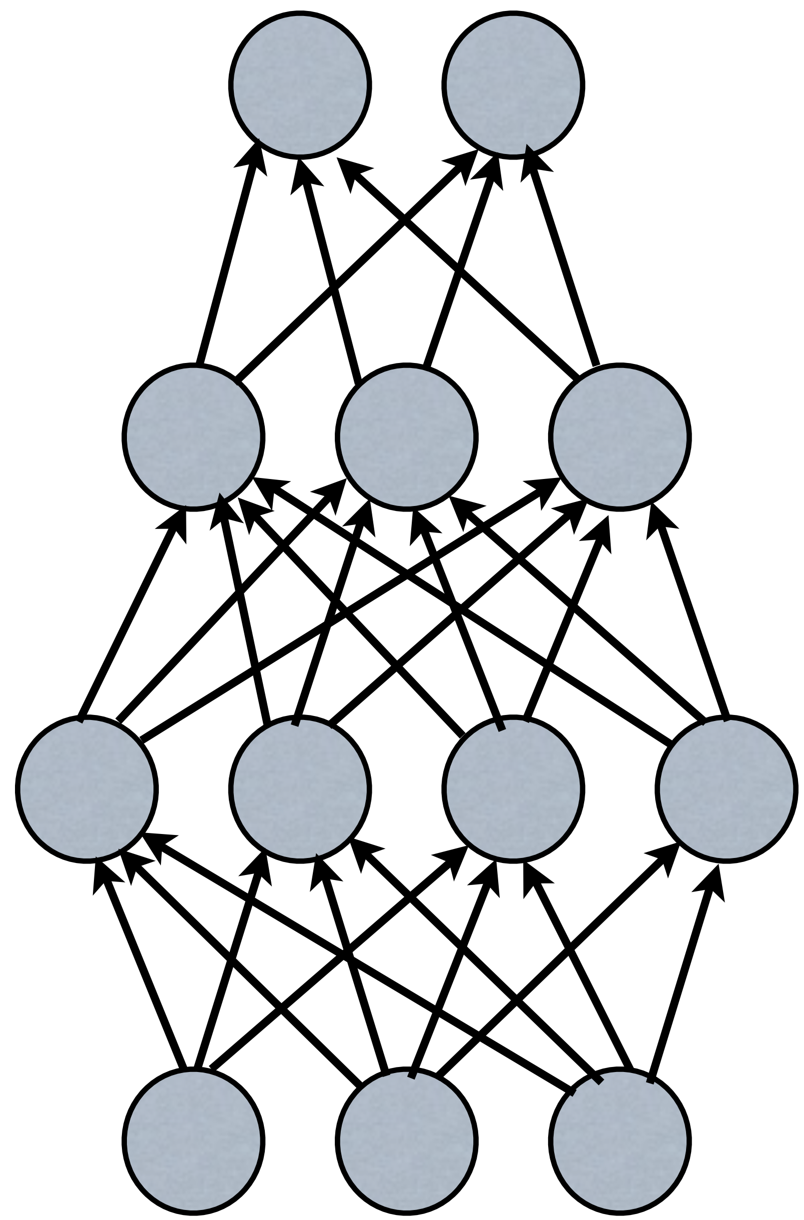


$$F(x) = \max(0, x)$$

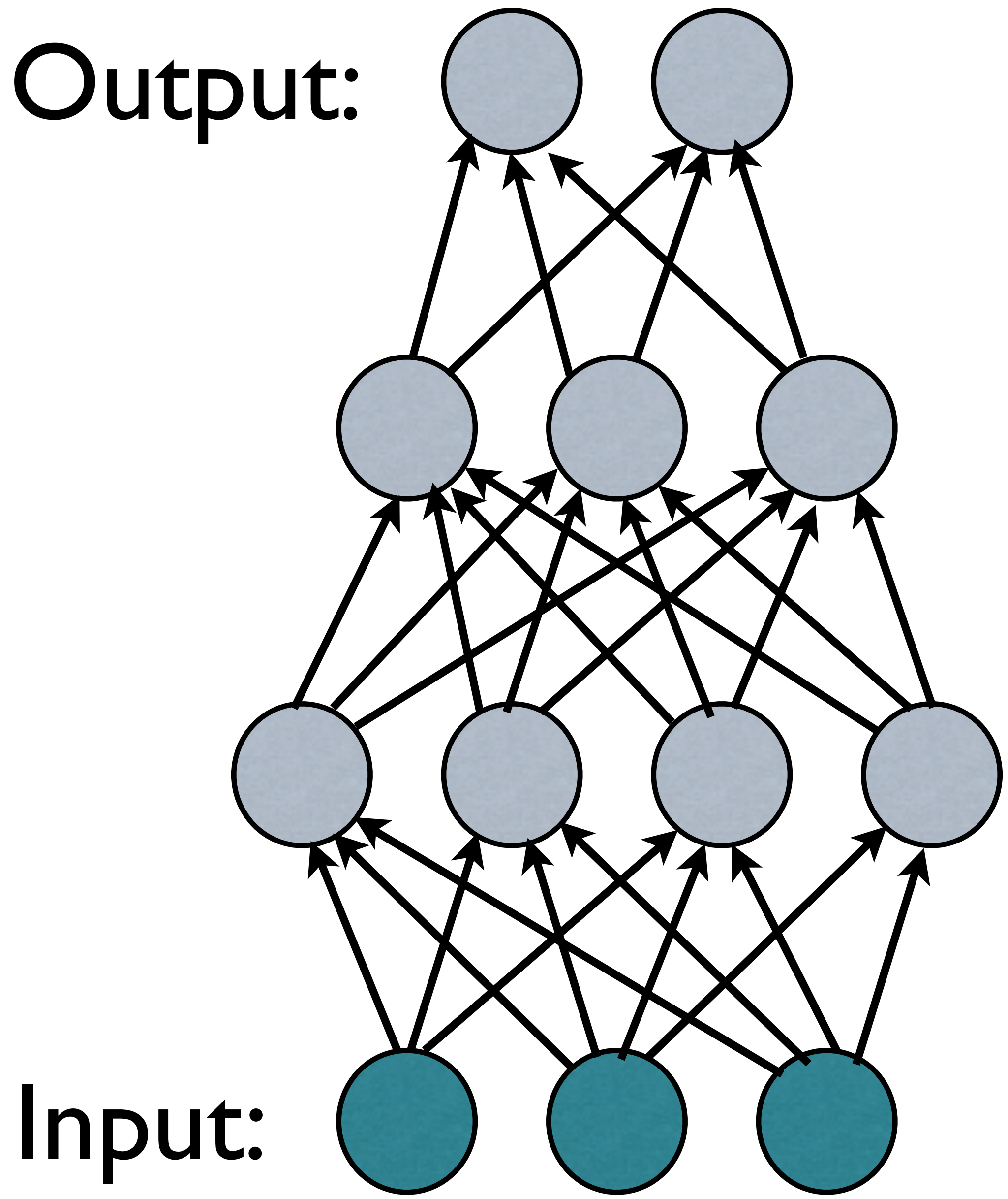


Neural Networks

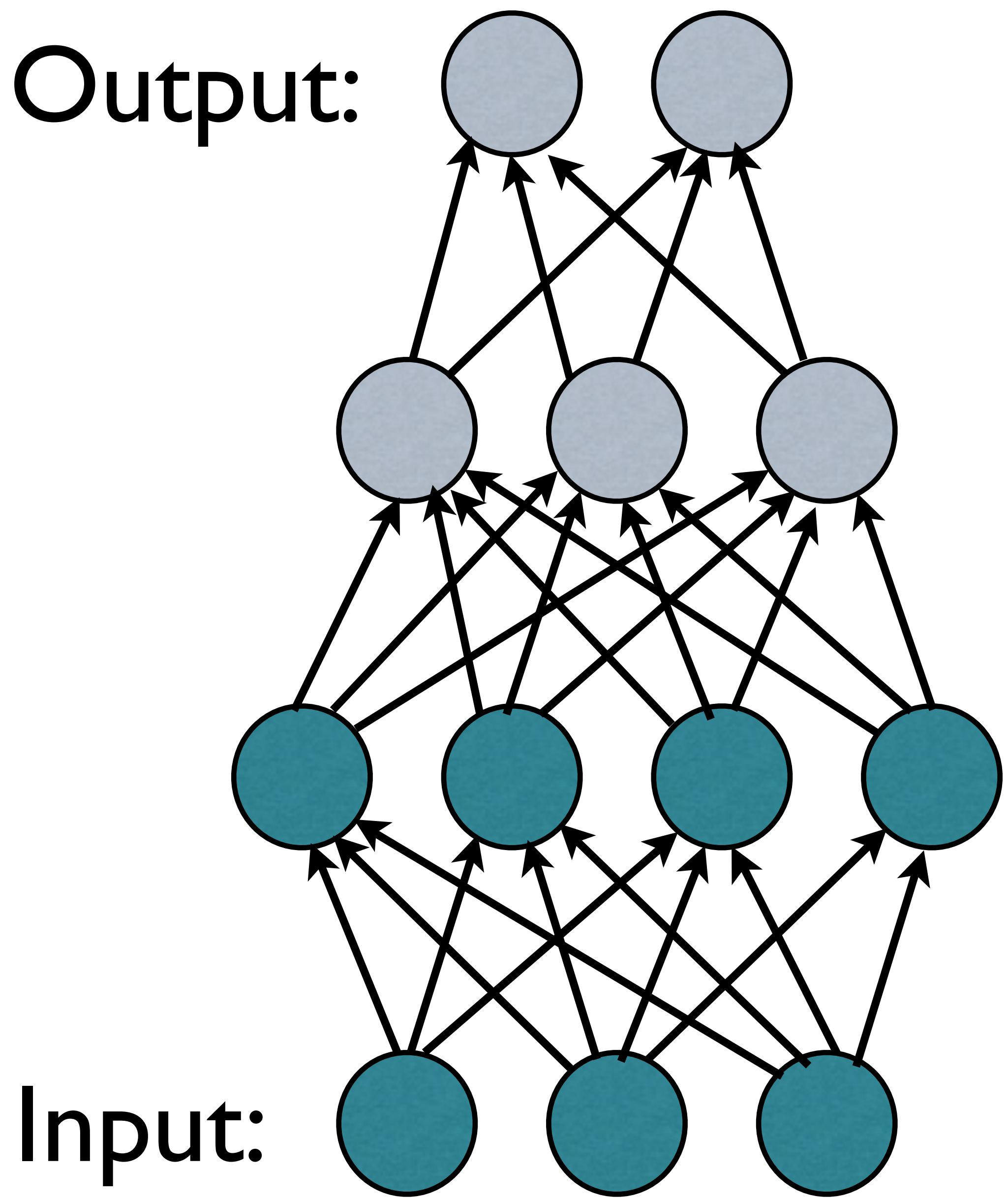
- Different weights compute different functions



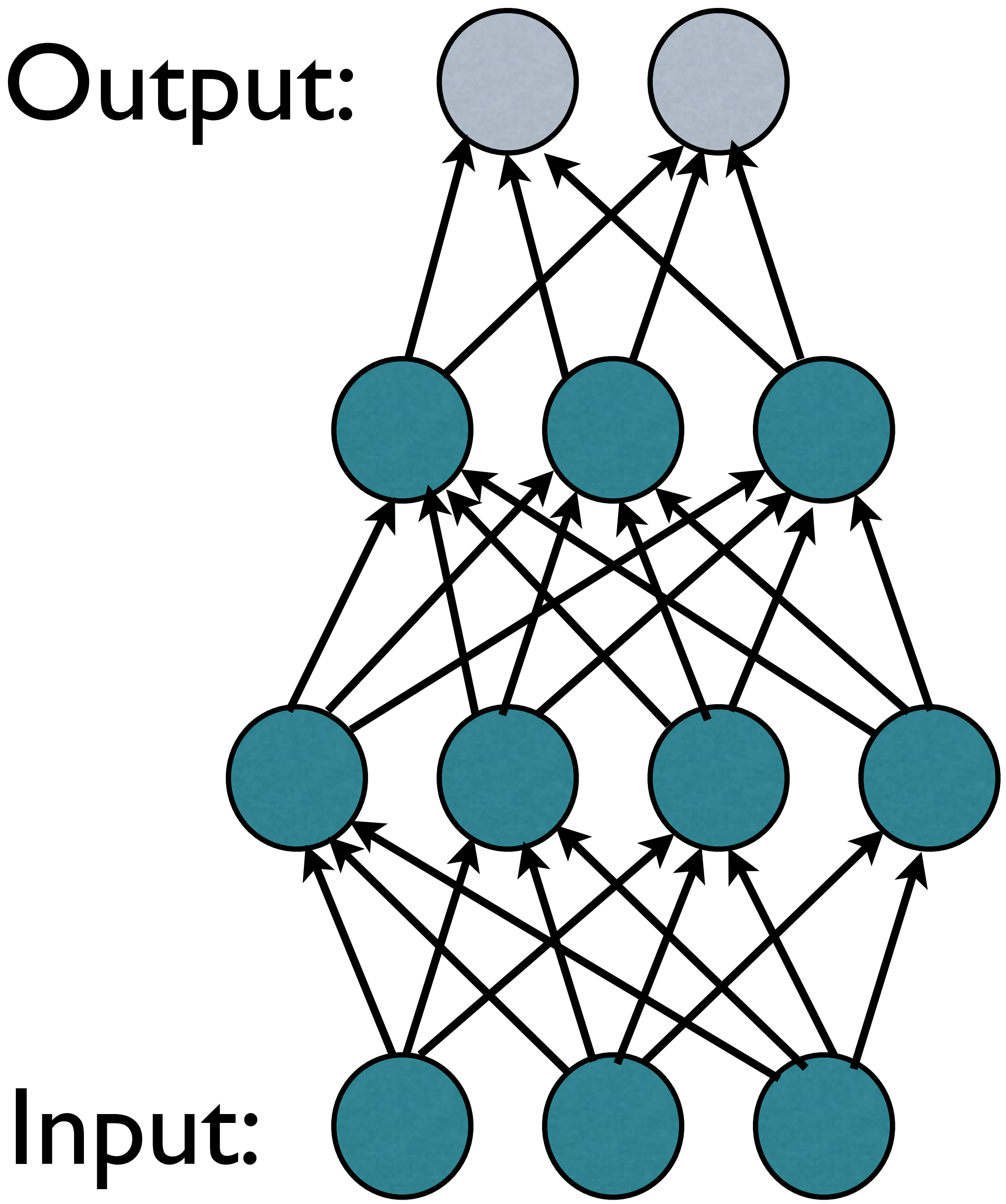
Neural Networks



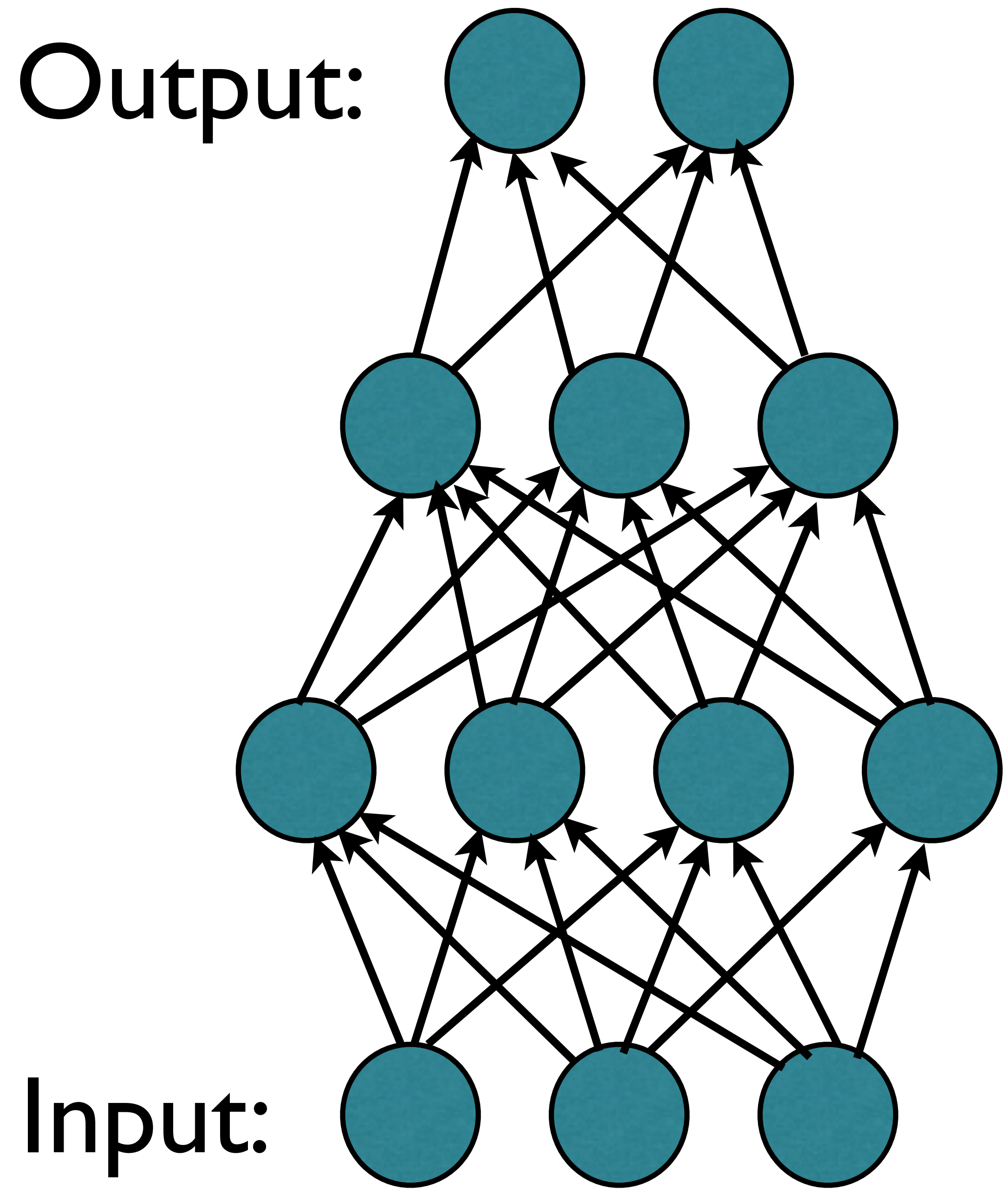
Neural Networks



Neural Networks



Neural Networks



Learning Algorithm

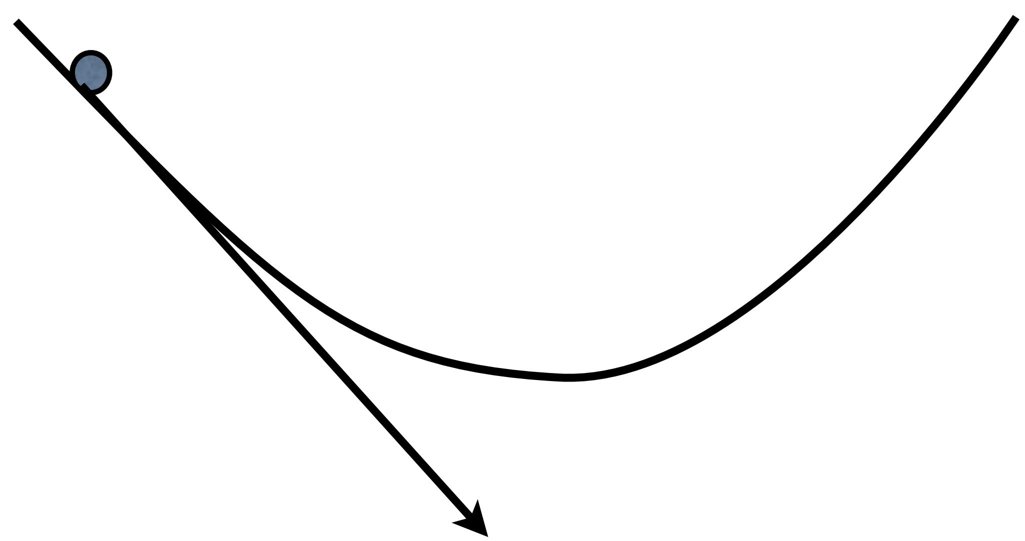
- **while** not done
  - pick a random training case **(x, y)**
  - run neural network on input **x**
  - modify connections to make prediction closer to **y**

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

What can neural nets compute?

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

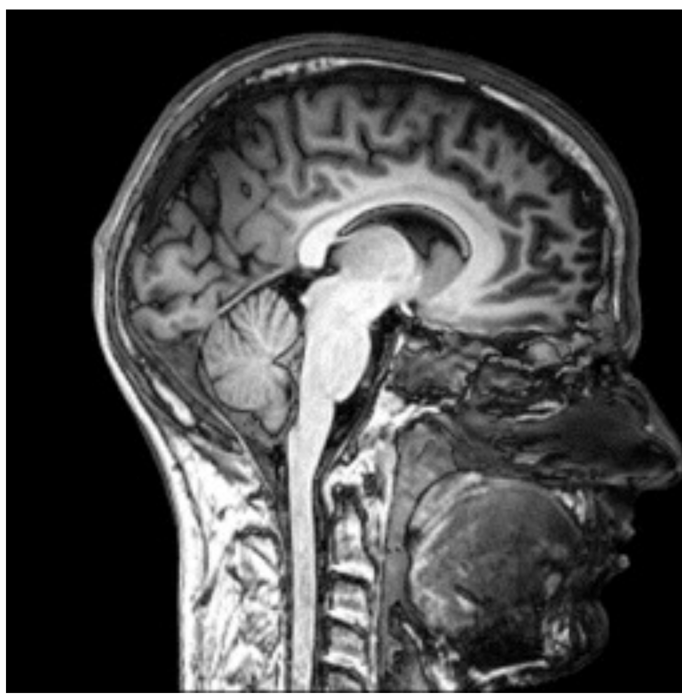
  - Recognize speech ("hear")

  - Recognize emotion

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

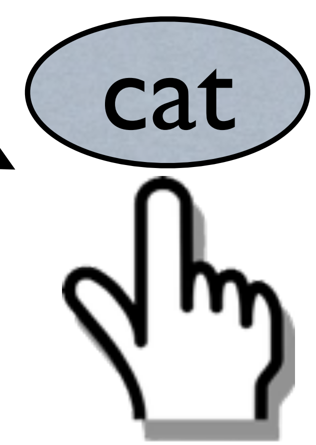
Why do neural networks work?

0.1 sec: neurons fire only 10 times!



see image

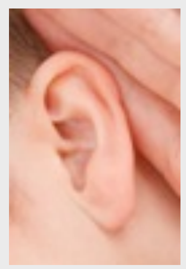
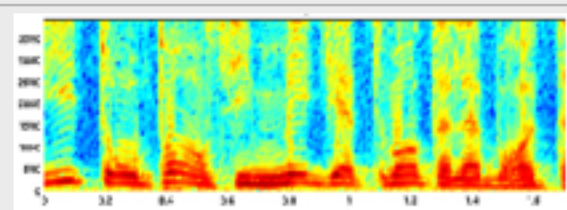
click if cat



Why do neural networks work?

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

Functions Artificial Neural Nets Can Learn

Input	Output
Pixels: 	"ear"
Audio: 	"sh ang hai res taur aun ts"
<query, doc1, doc2>	P(doc1 preferred over doc2)
"Hello, how are you?"	"Bonjour, comment allez-vous?"

Research Objective: Minimizing Time to Results

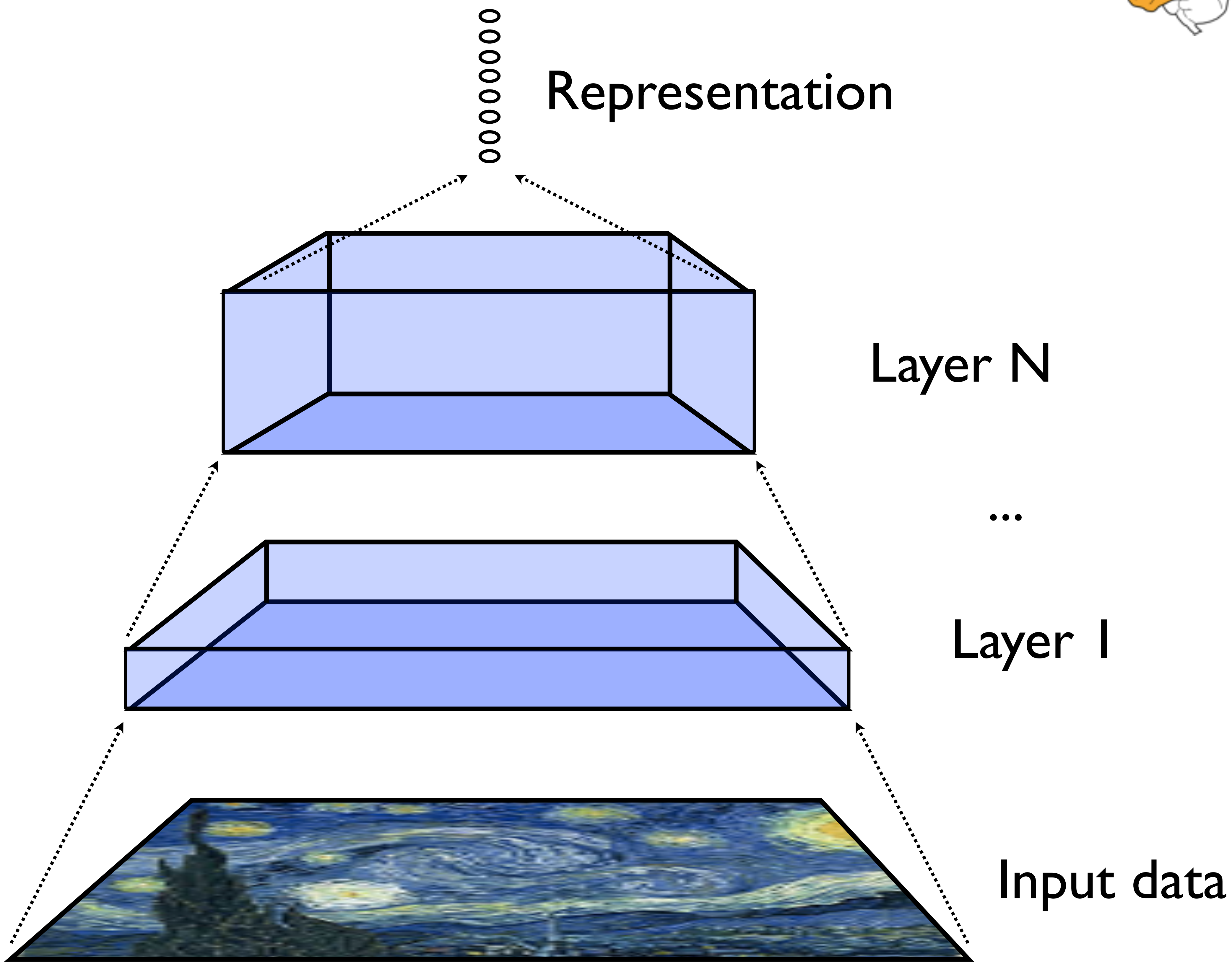
- We want results of experiments quickly
- "Patience threshold": No one wants to wait more than a few days or a week for a result
  - Significantly affects scale of problems that can be tackled
  - We sometimes optimize for experiment turnaround time, rather than absolute minimal system resources for performing the experiment

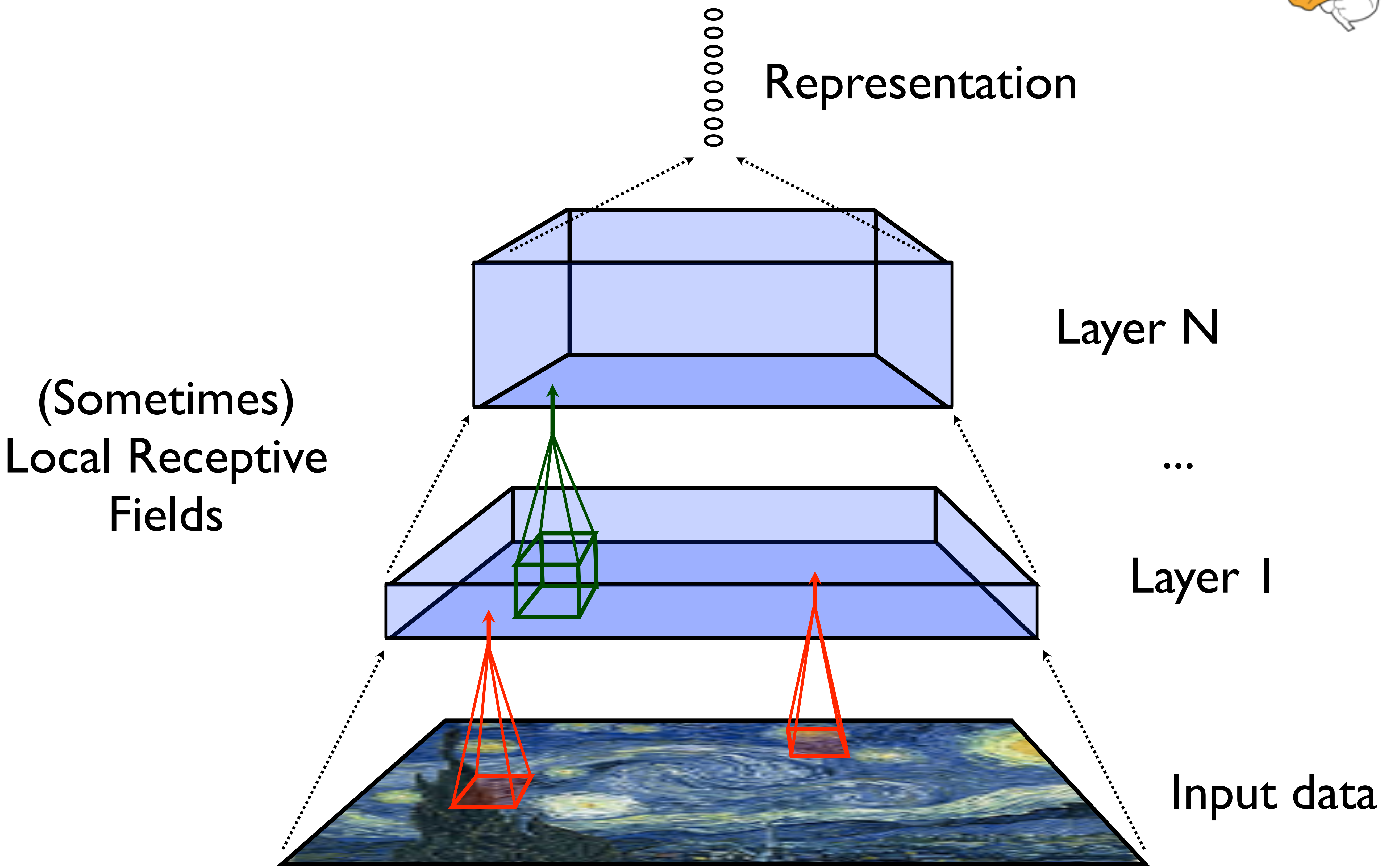
Train in a day what takes a single GPU card 6 weeks

How Can We Train Big Nets Quickly?

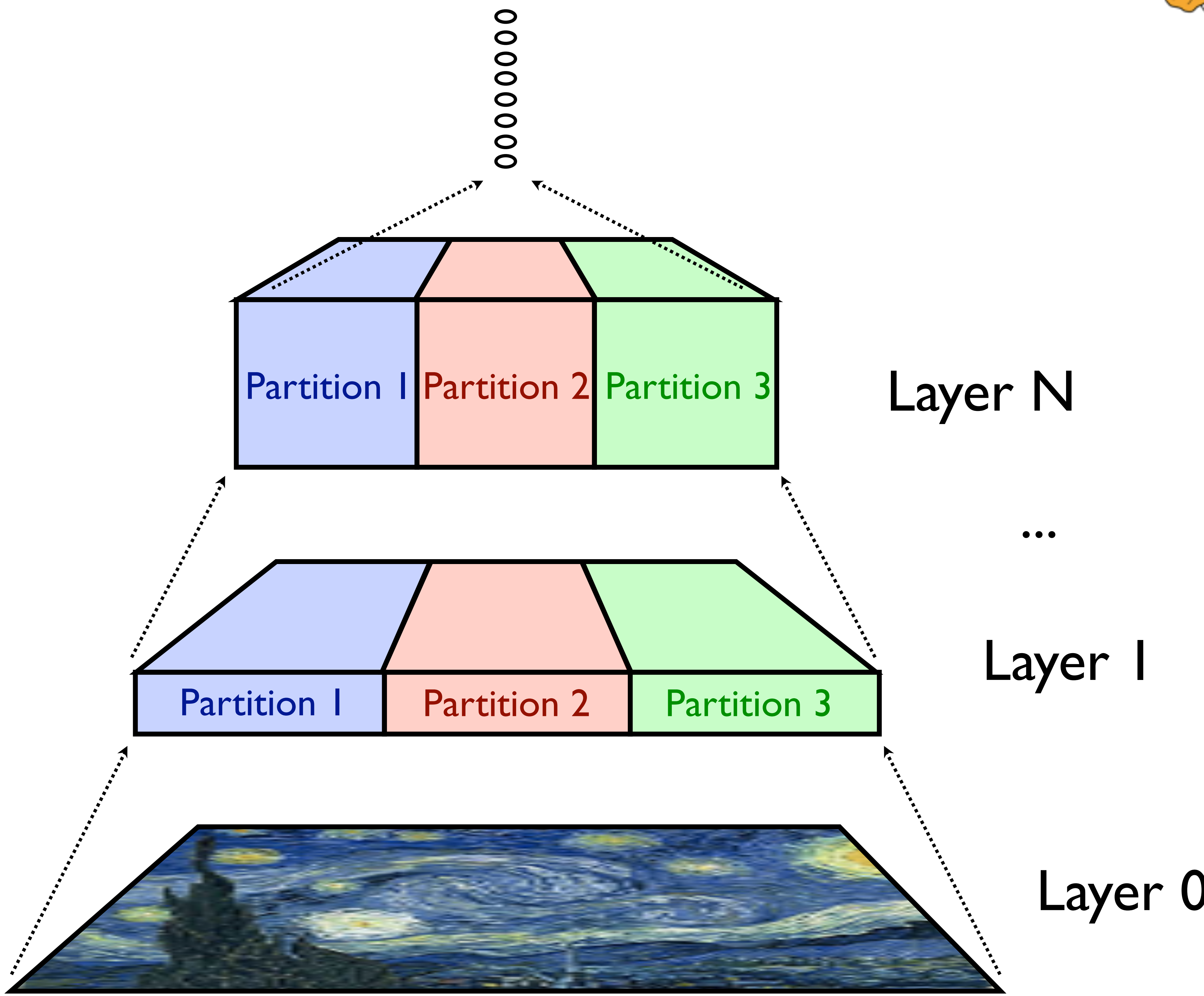
- Exploit many kinds of parallelism

- Model parallelism
- Data parallelism

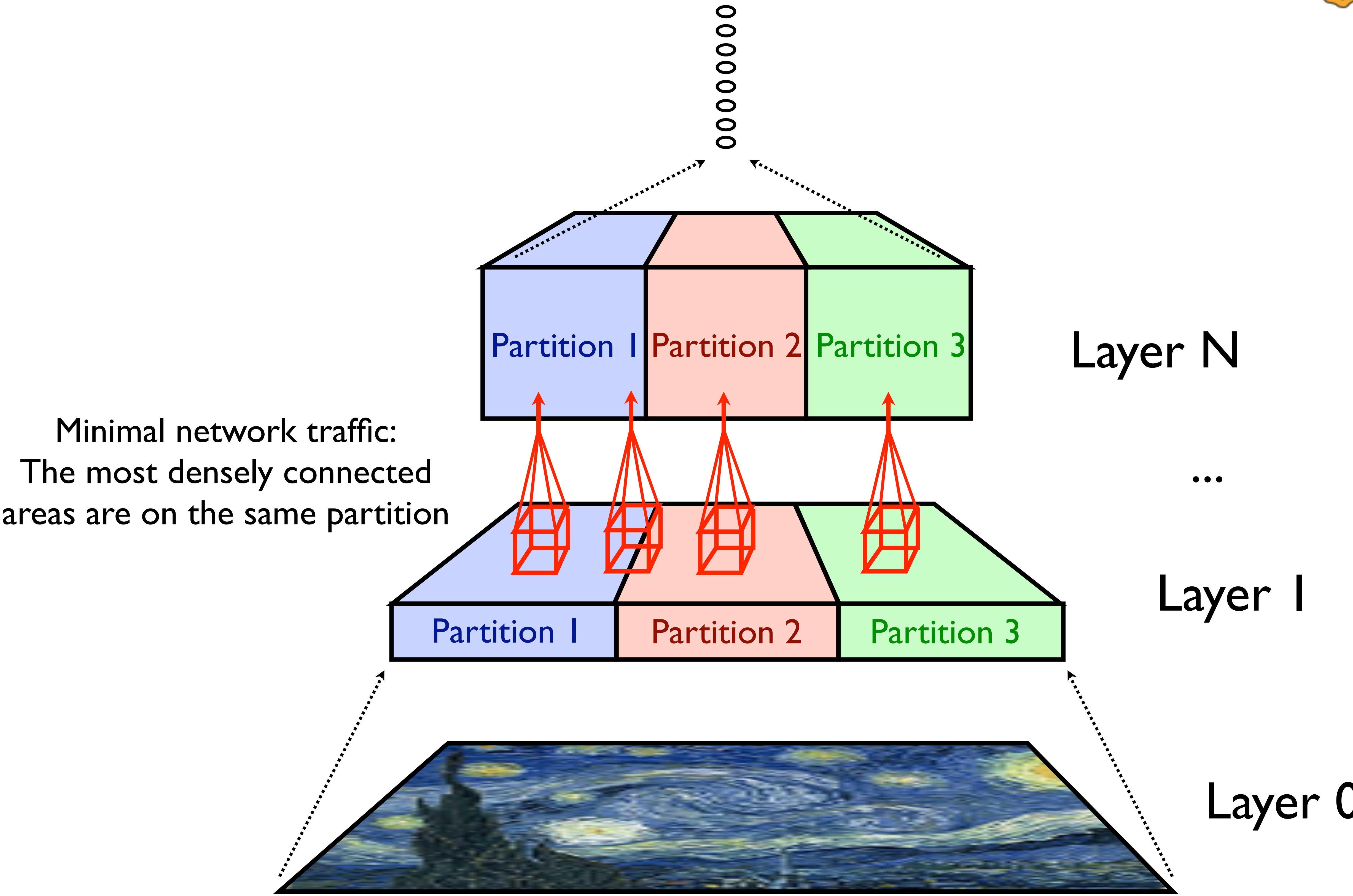

Representation
Layer N
...
Layer 1
Input data


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

Model Parallelism: Partition model across machines

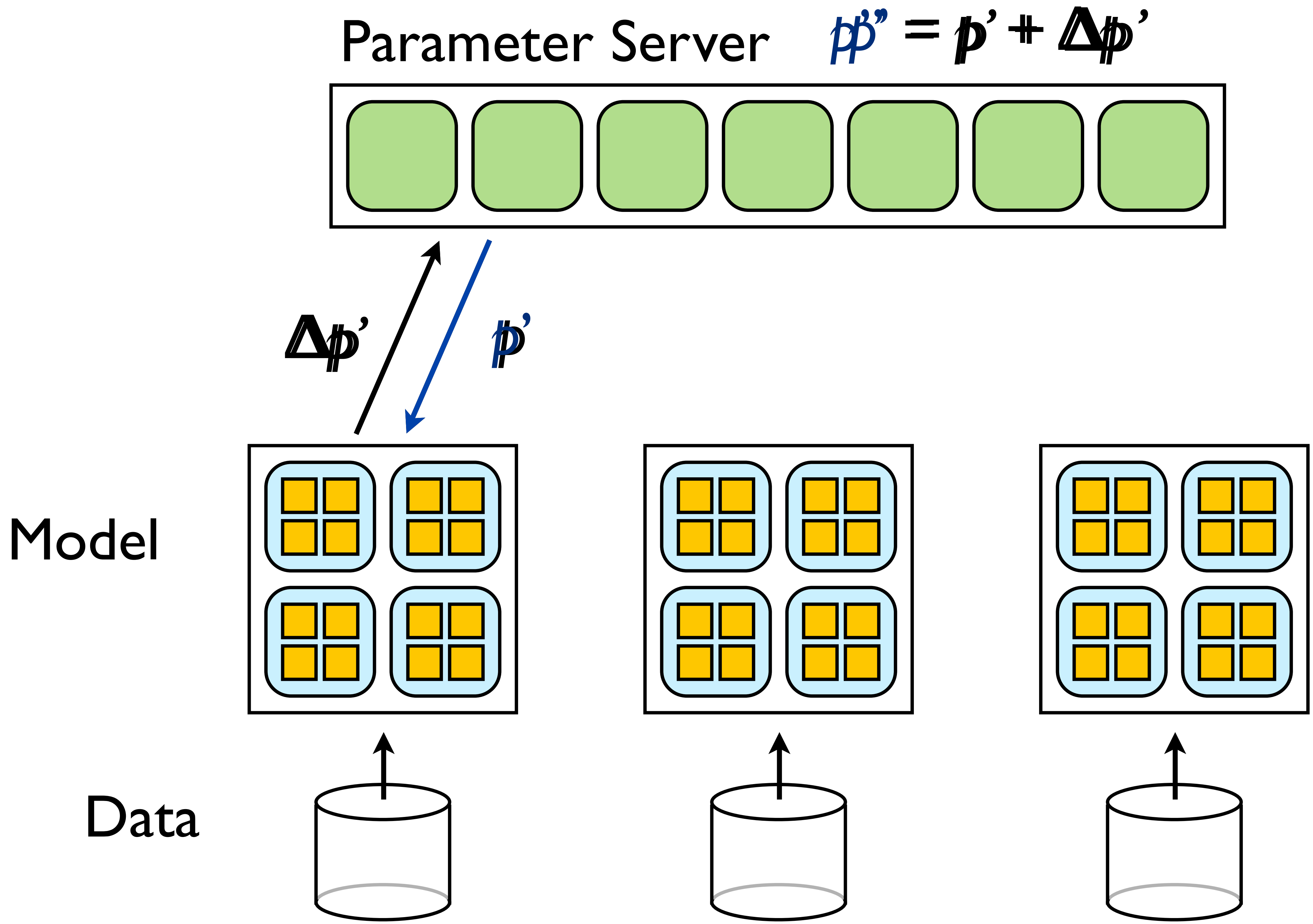


Model Parallelism: Partition model across machines



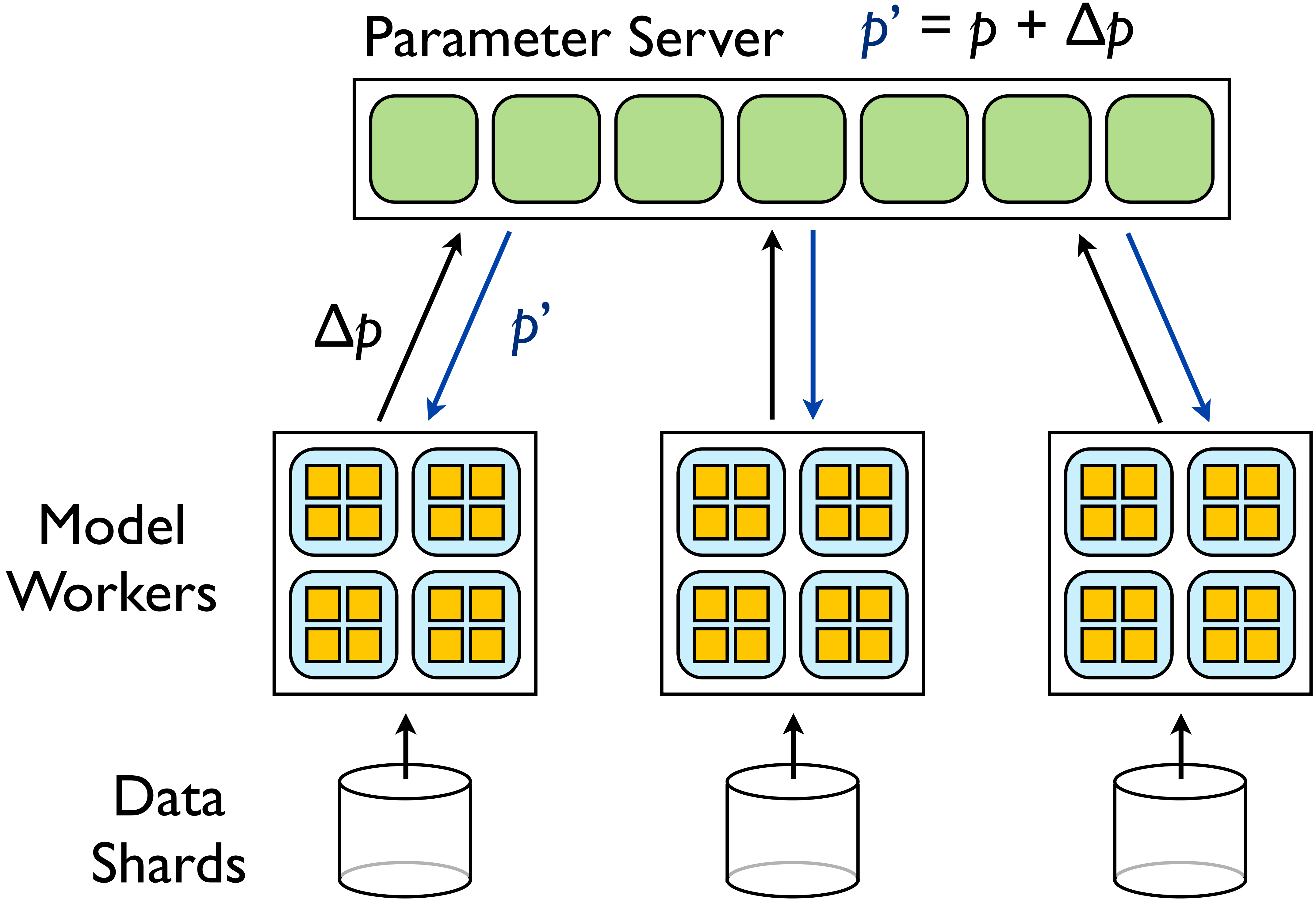
One replica of our biggest model: 144 machines, ~2300 cores

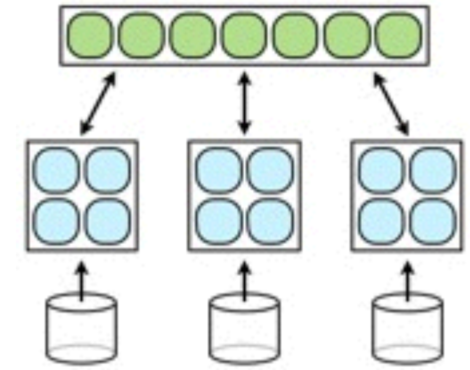
Data Parallelism:
Asynchronous Distributed Stochastic Gradient Descent



Data Parallelism:

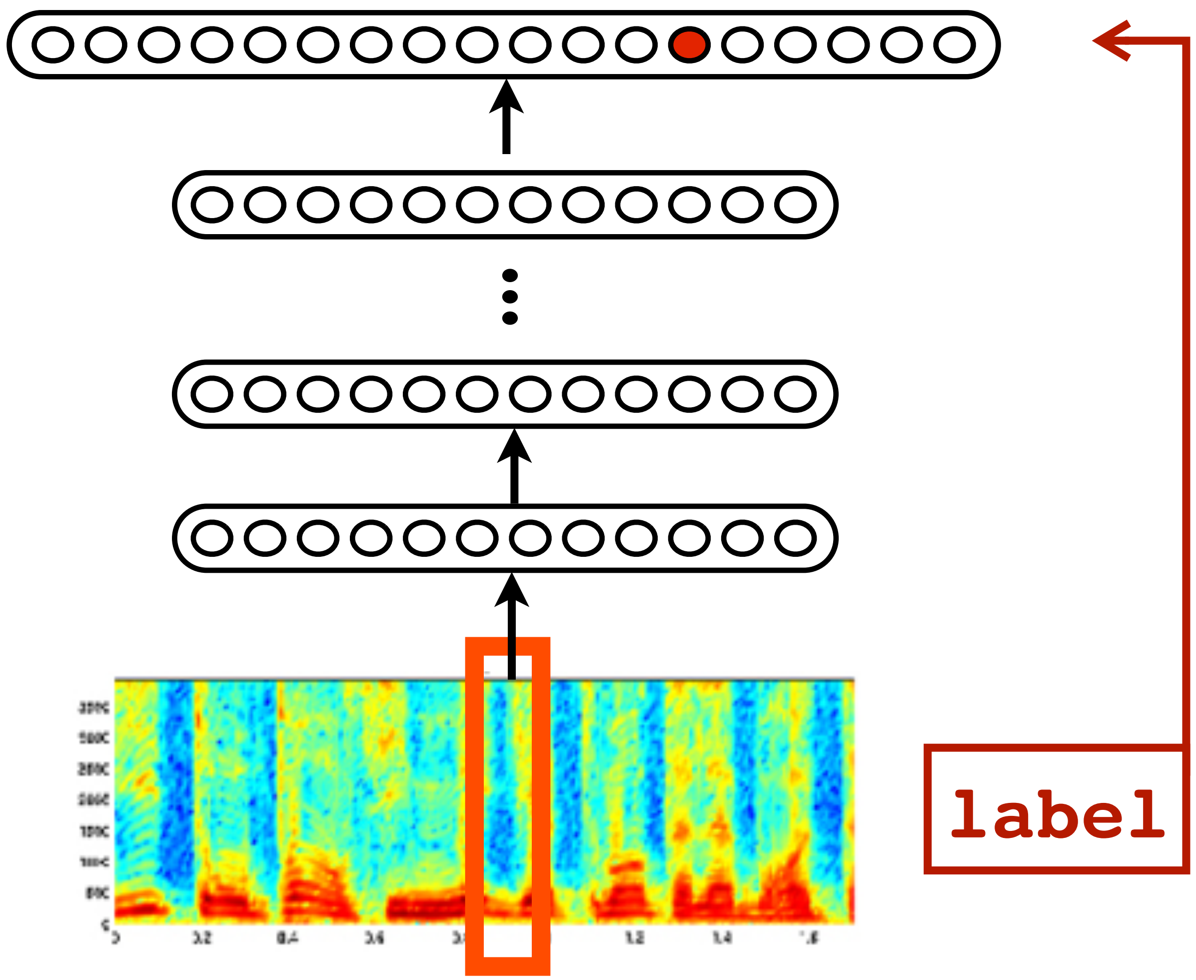
Asynchronous Distributed Stochastic Gradient Descent





Applications

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

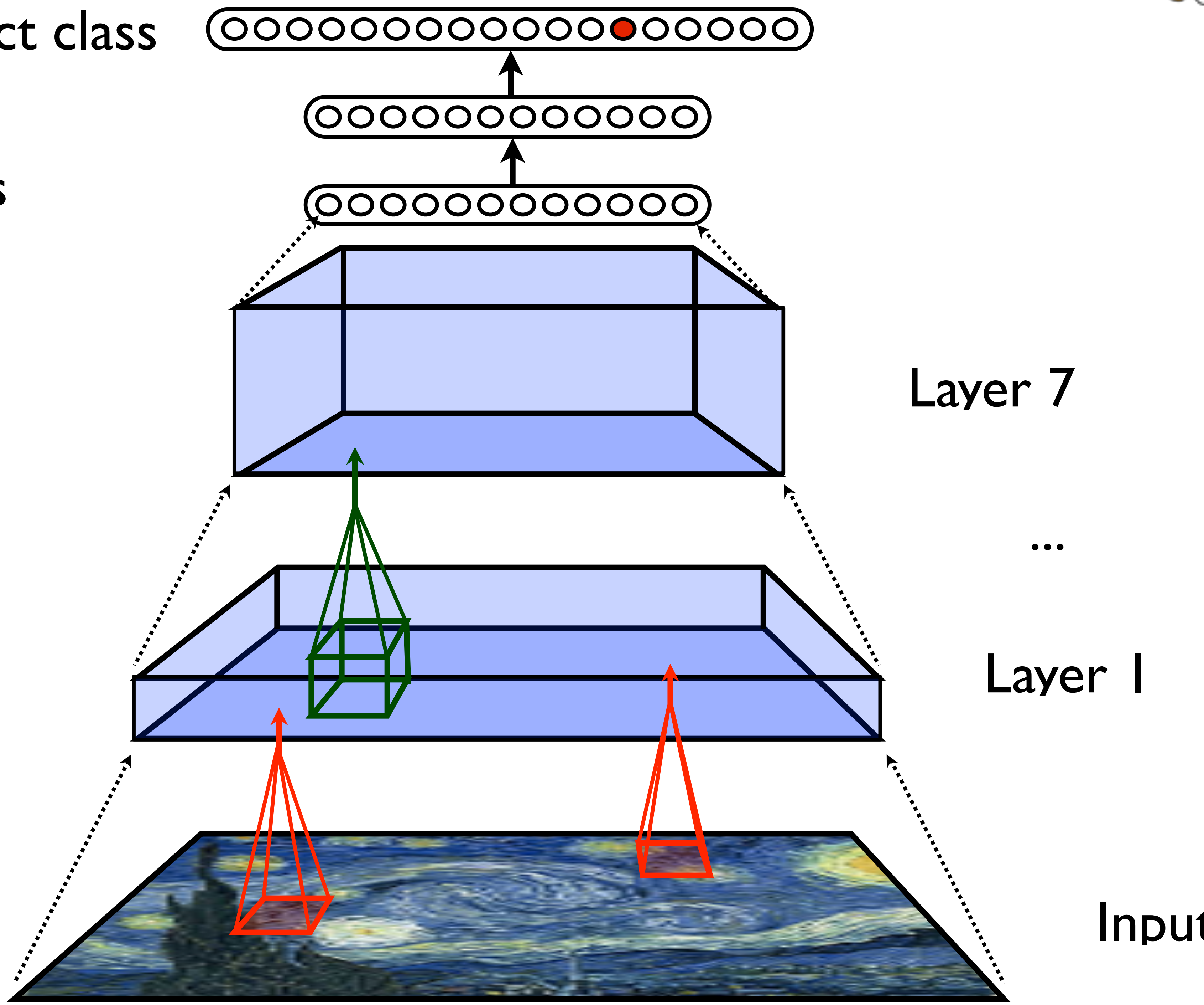
2012-era Convolutional Model for Object Recognition

Softmax to predict object class

Fully-connected layers

Convolutional layers
(same weights used at all spatial locations in layer)

Convolutional networks developed by Yann LeCun (NYU)



Basic architecture developed by Krizhevsky, Sutskever & Hinton (all now at Google).

Won 2012 ImageNet challenge with 16.4% top-5 error rate

2014-era Model for Object Recognition

Module with 6 separate convolutional layers

24 layers deep!

Developed by team of Google Researchers:

Won 2014 ImageNet challenge with **6.66%** top-5 error rate

Good Fine-grained Classification

"hibiscus"

"dahlia"

Good Generalization

Both recognized as a "meal"

Google

Sensible Errors

"snake"

"dog"

Google

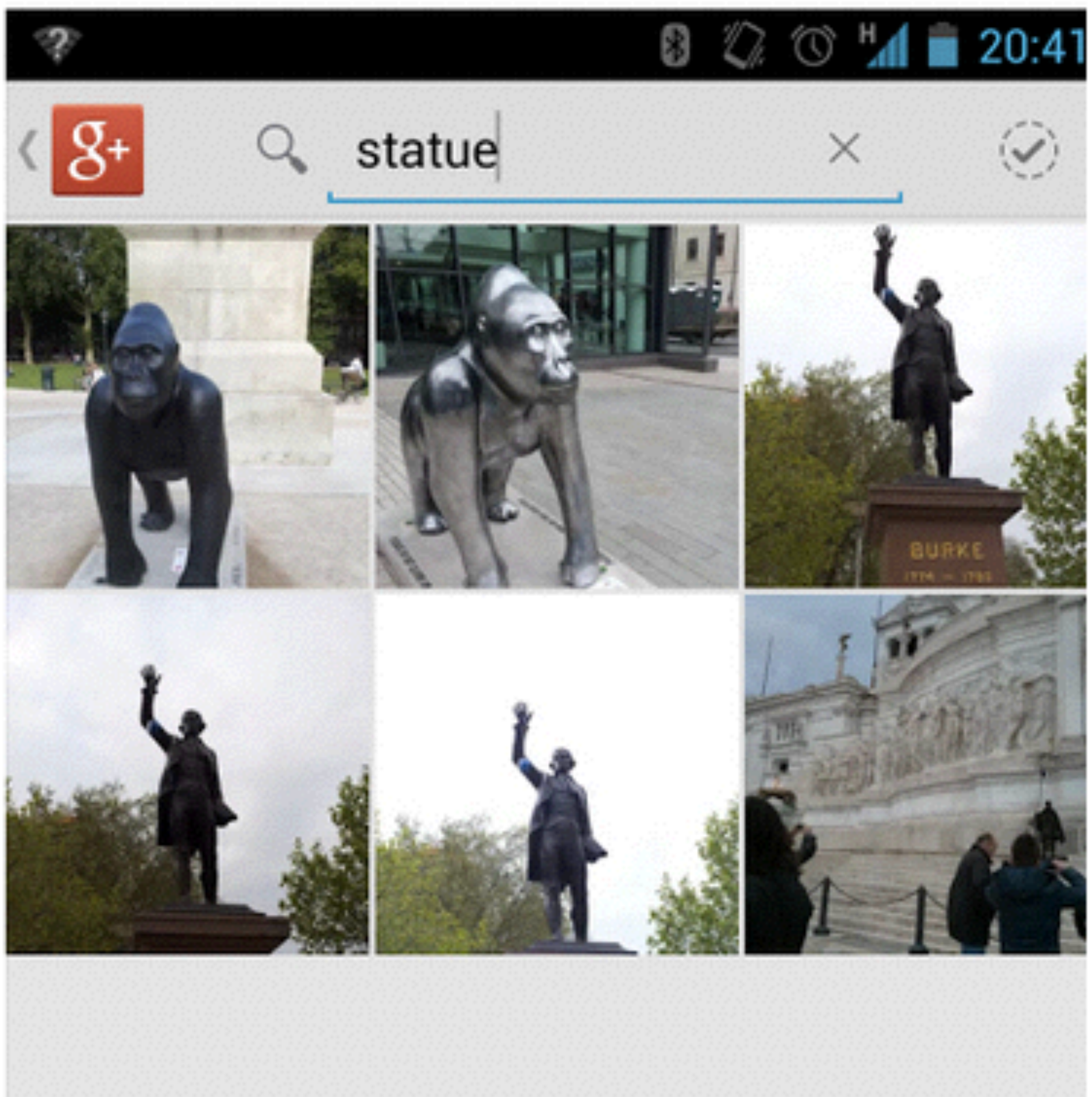
Works in practice

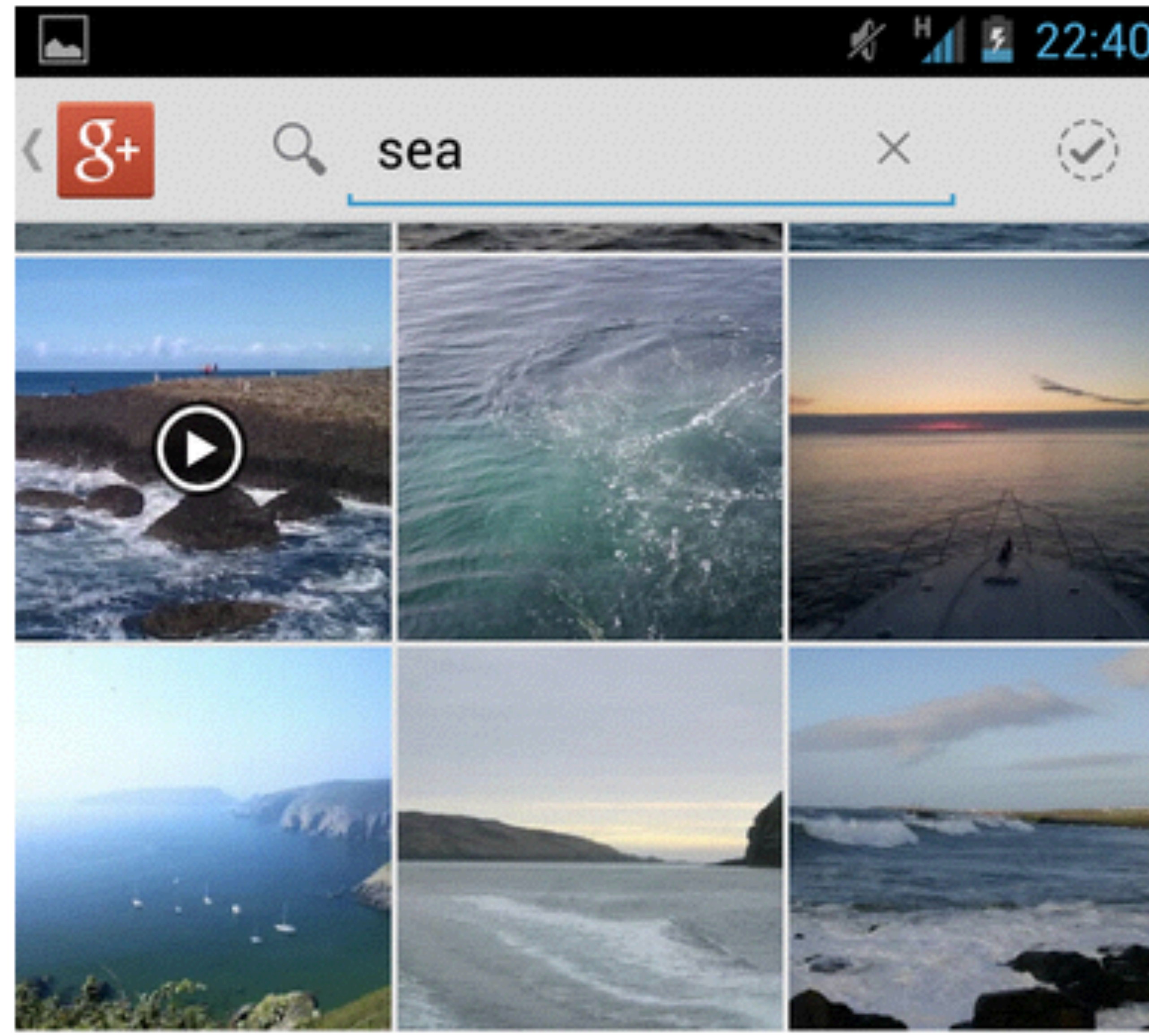
for real users.

Wow.

The new Google plus photo search is a bit insane.

I didn't tag those... :)

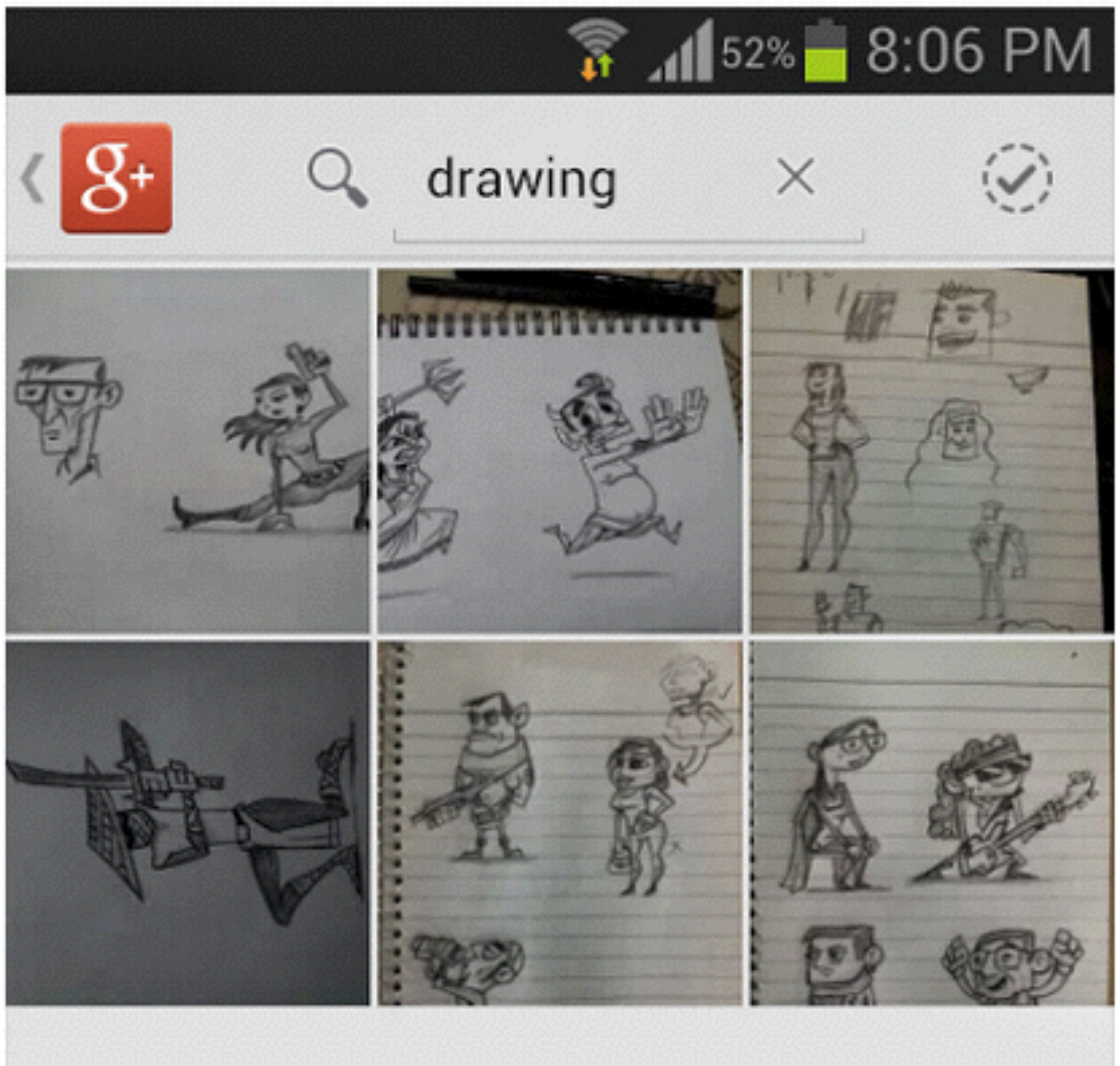


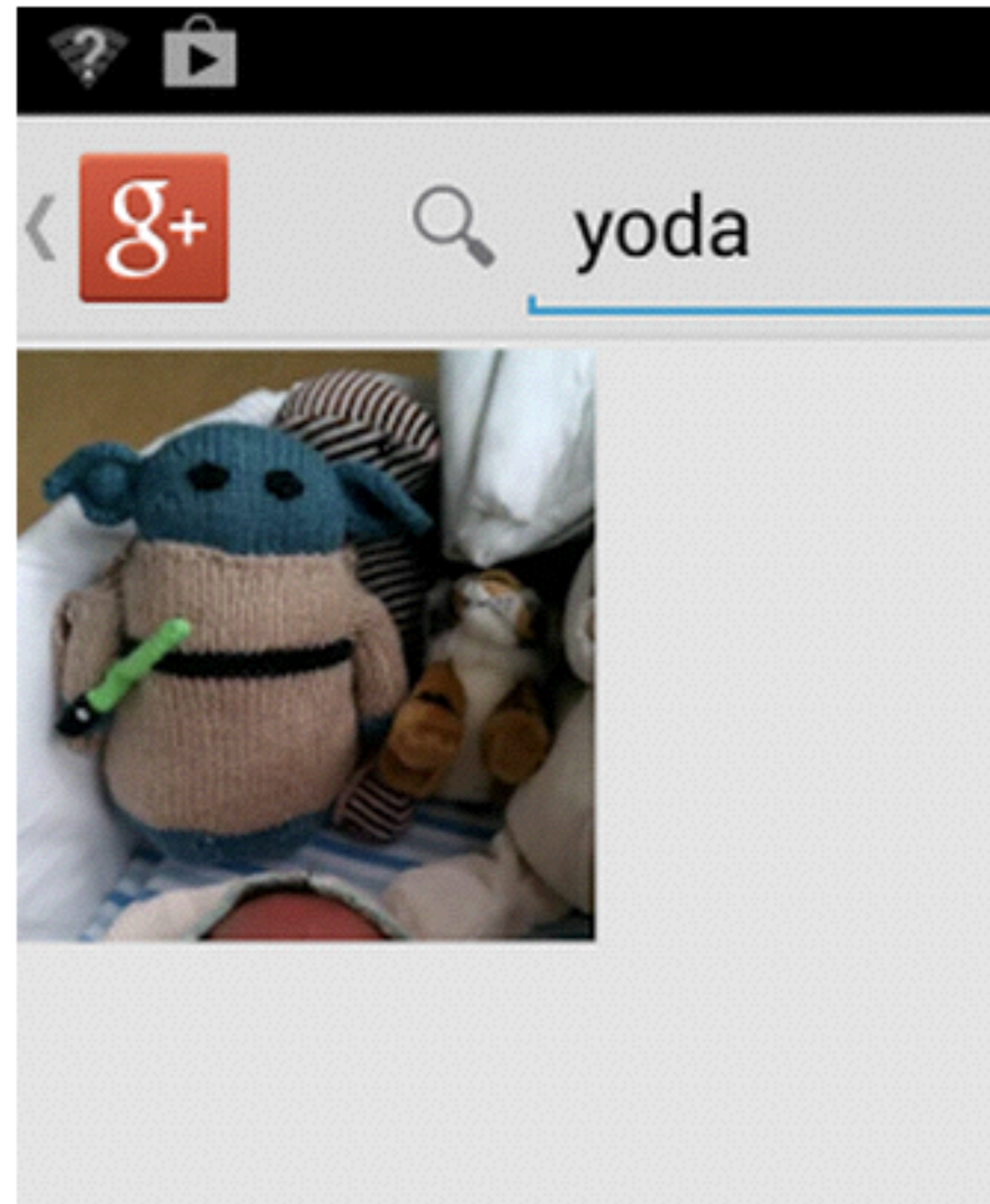


Works in practice

for real users.

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

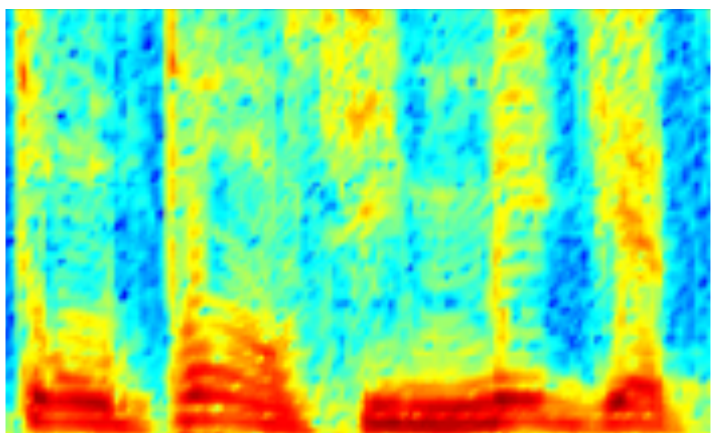


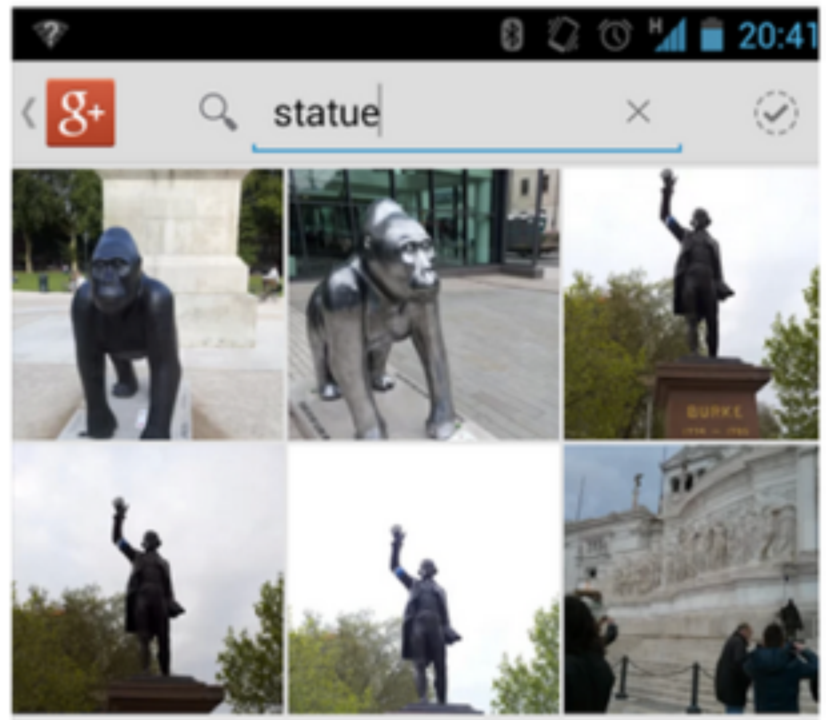



ASIAWIDE TRAVEL
Tel (02) 9745 3355 1ST Floor, 240 BURWOOD RD
aria's Bakery Inn 超羣餅屋
Maria's Bakery Inn 超羣餅屋


Corner
Cubbyhouse
THUMP
www.thumphq.com
CIANO MOTOR ENGINEERS
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
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.

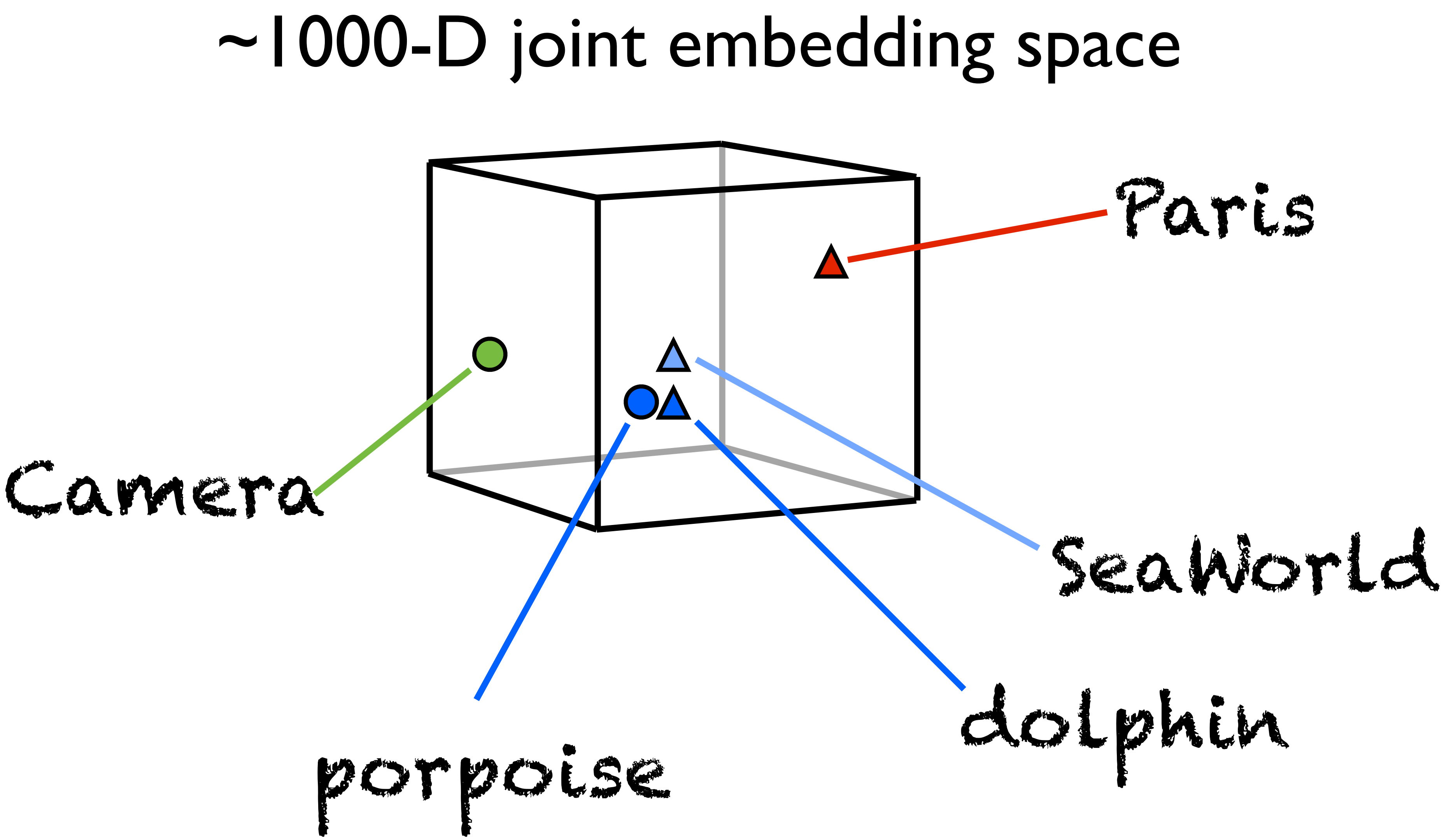




What about domains with sparse input data?

How can DNNs possibly deal with sparse data?

Answer: Embeddings



How Can We Learn the Embeddings?

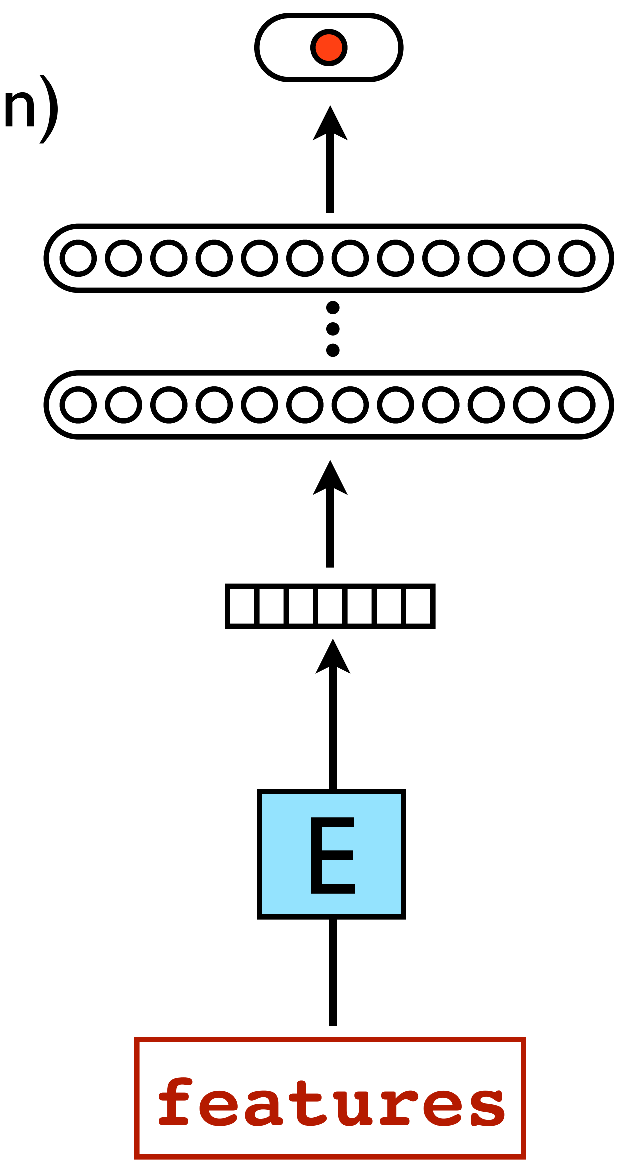
Prediction
(classification or regression)

Deep neural network

Floating-point vectors

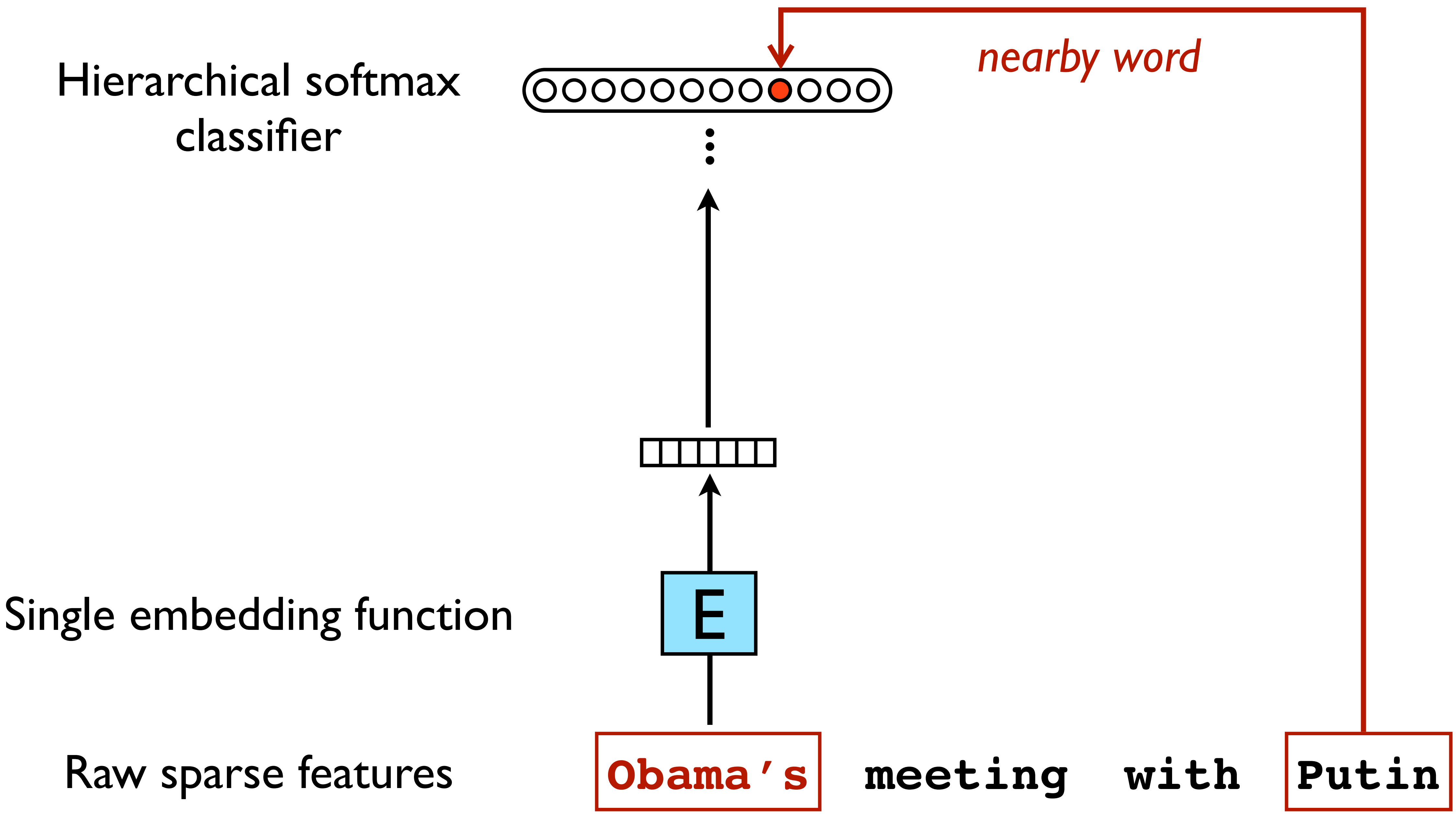
Embedding function

Raw sparse inputs



How Can We Learn the Embeddings?

Skipgram Text Model

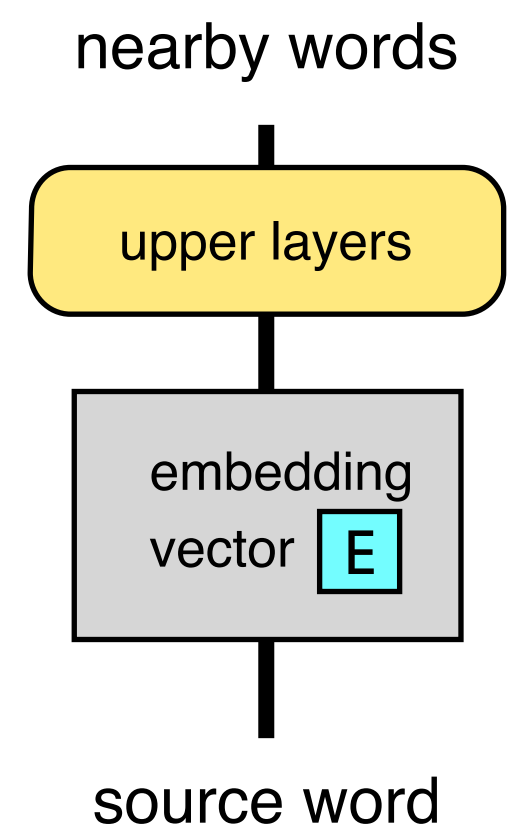


Mikolov, Chen, Corrado and Dean. *Efficient Estimation of Word Representations in Vector Space*, http://arxiv.org/abs/1301.3781.

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



Google

* 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).

$$E(\textit{hotter}) - E(\textit{hot}) \approx E(\textit{bigger}) - E(\textit{big})$$

$$E(\textit{Rome}) - E(\textit{Italy}) \approx E(\textit{Berlin}) - E(\textit{Germany})$$

Solving Analogies

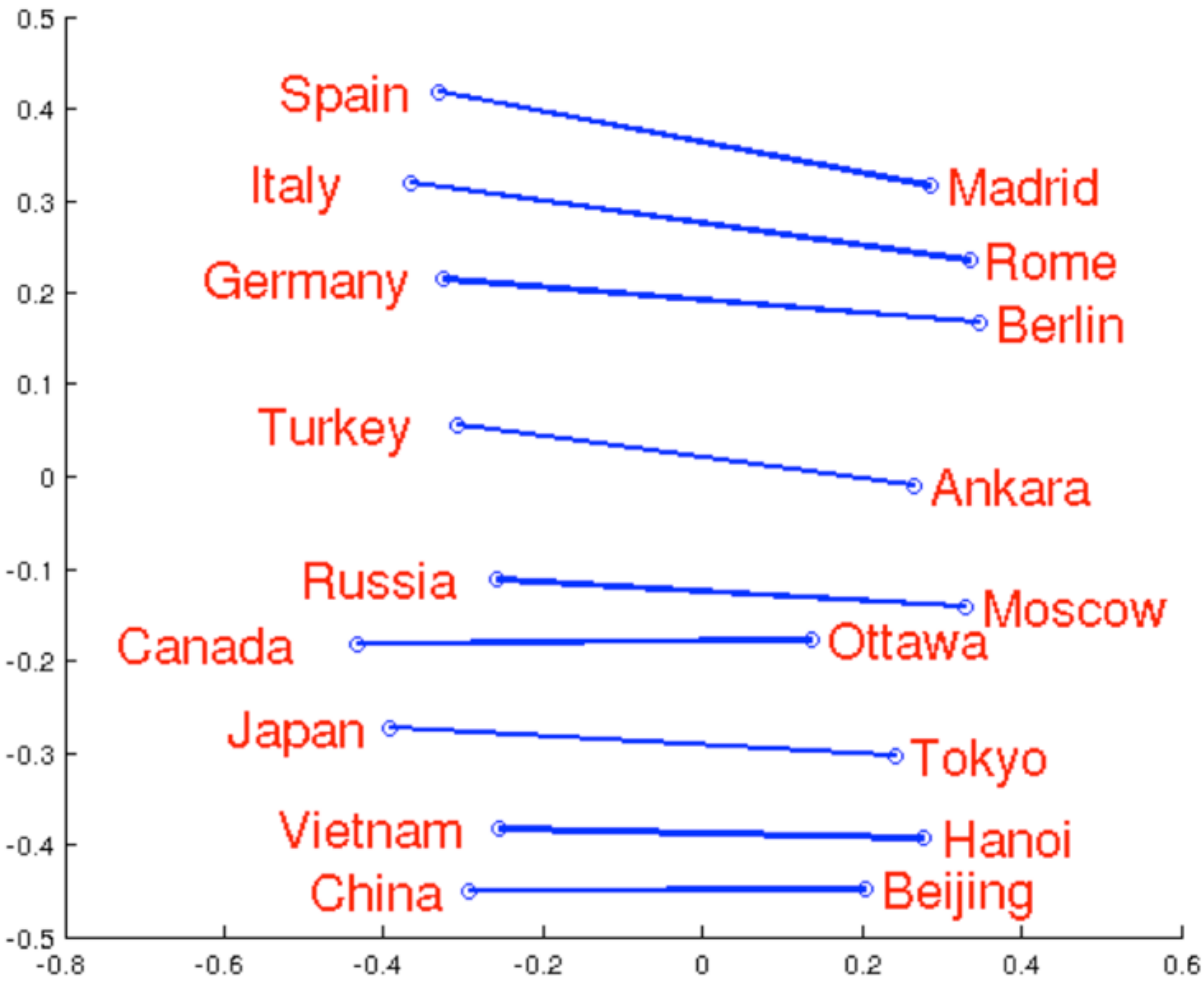
- Embedding vectors trained for the language modeling task have very interesting properties (especially the skip-gram model).

$$E(\textit{hotter}) - E(\textit{hot}) + E(\textit{big}) \approx E(\textit{bigger})$$

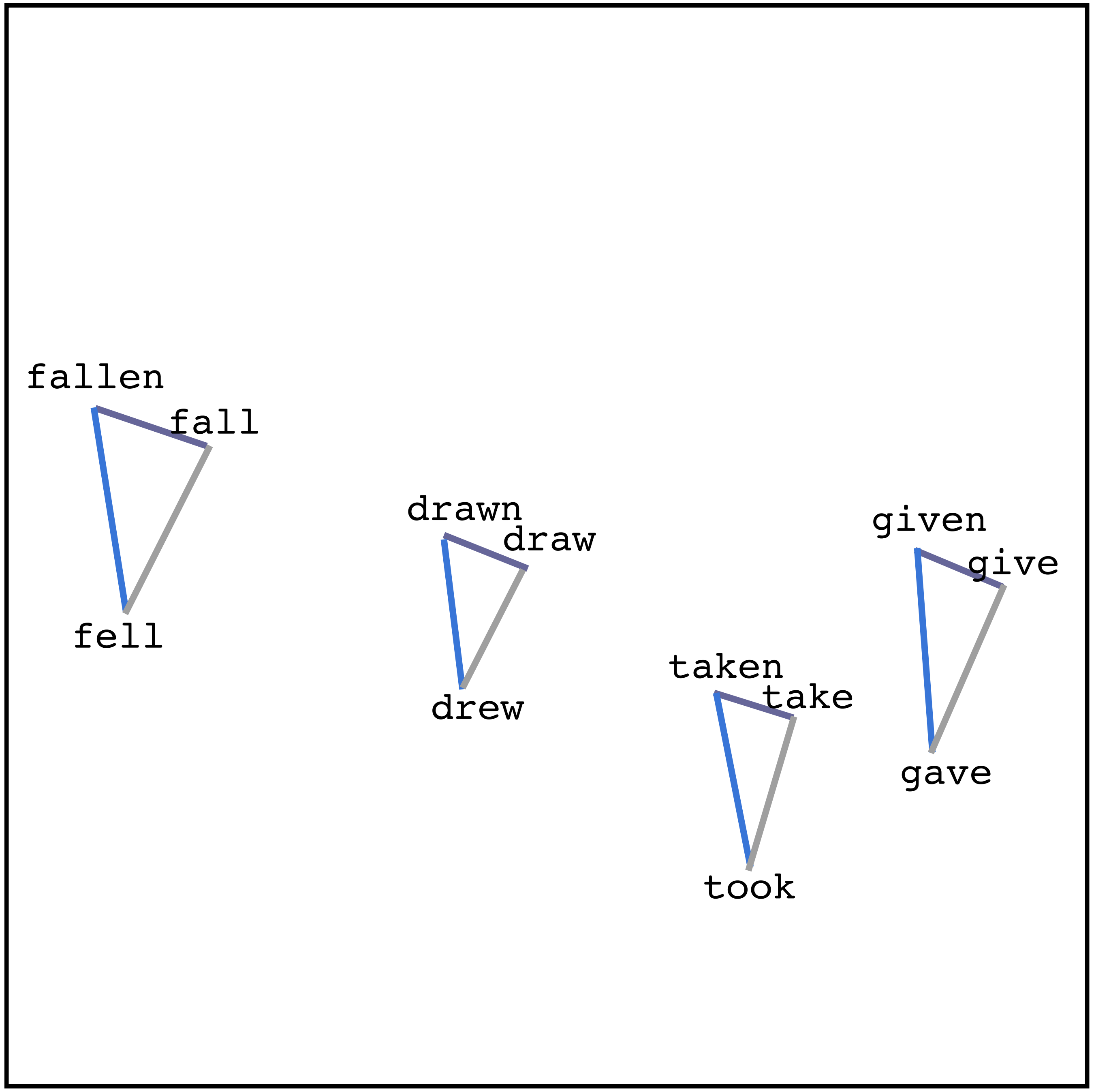
$$E(\textit{Rome}) - E(\textit{Italy}) + E(\textit{Germany}) \approx E(\textit{Berlin})$$

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



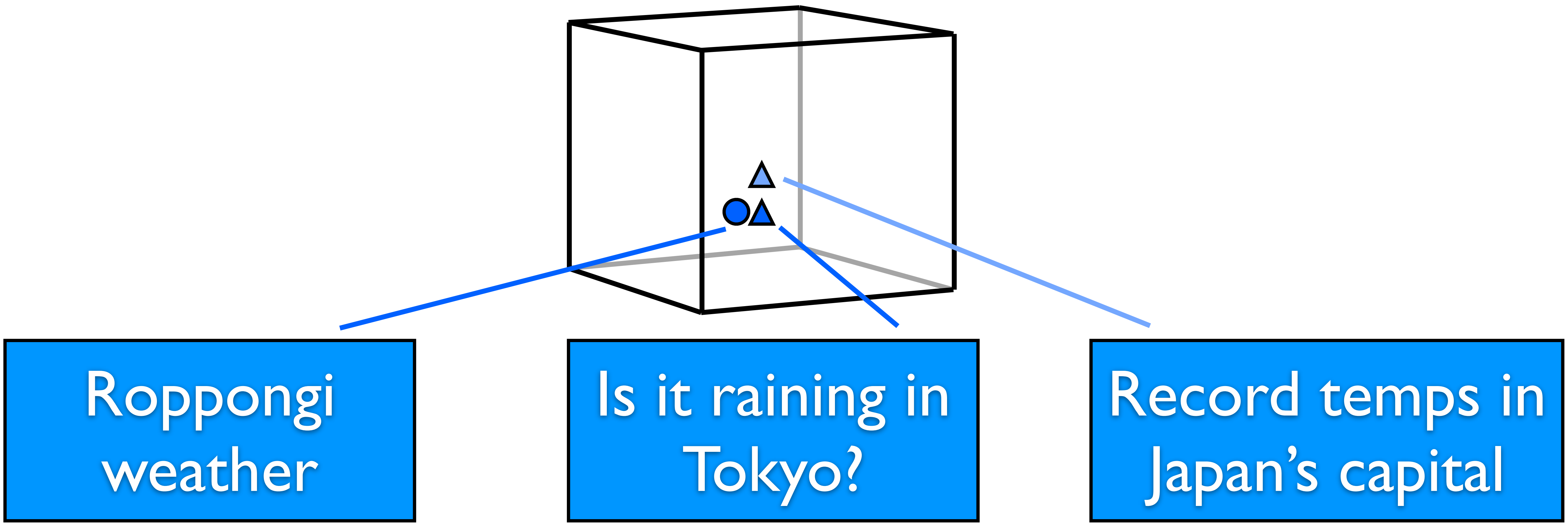
Embeddings are Powerful



Google

Embeddings seem useful.
What about longer pieces of text?

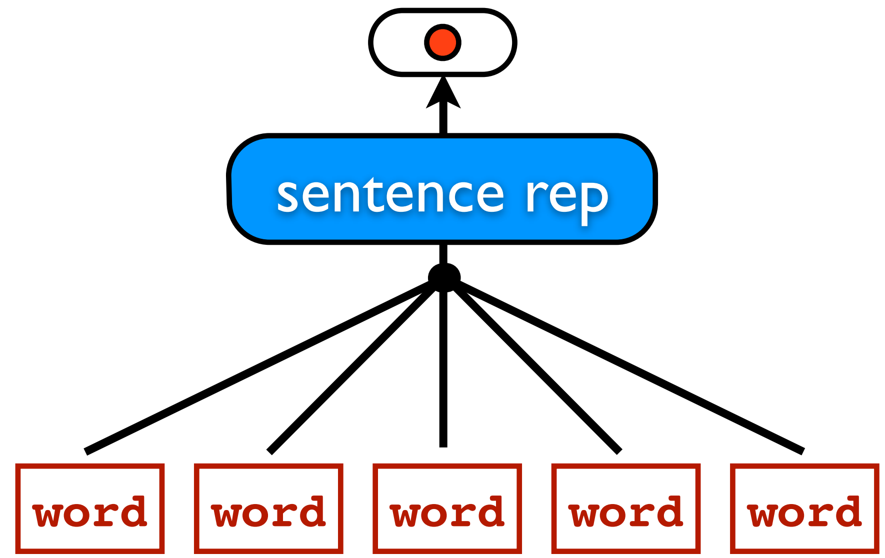
Can We Embed Longer Pieces of Text?



- Query similarity / Query-Document scoring
- Machine translation
- Question answering
- Natural language *understanding*?

Bag of Words:

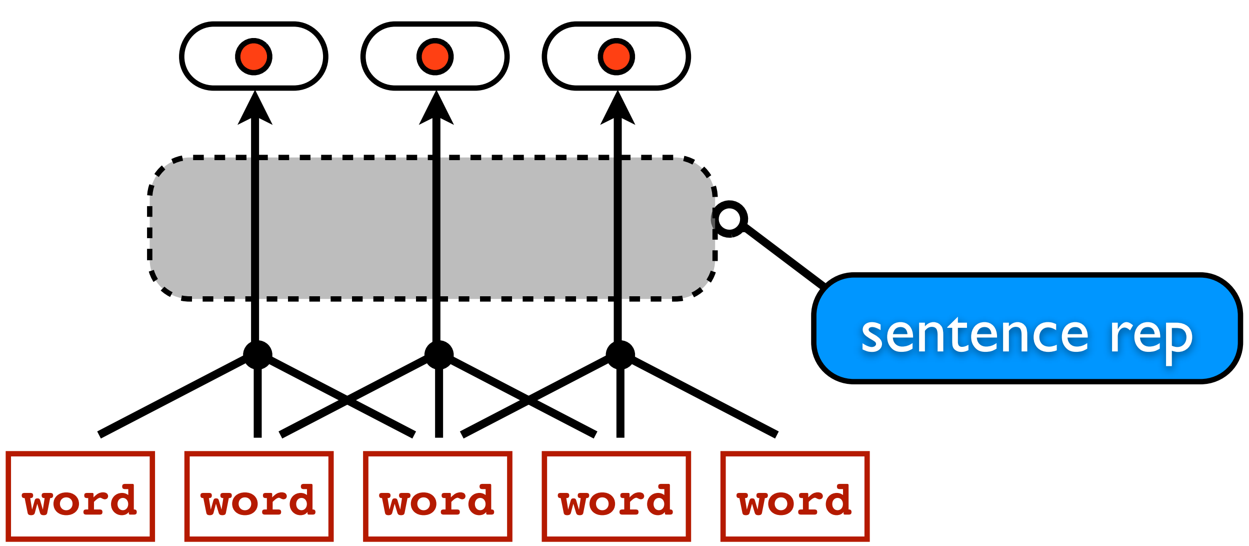
Avg of embeddings



Topic Model:

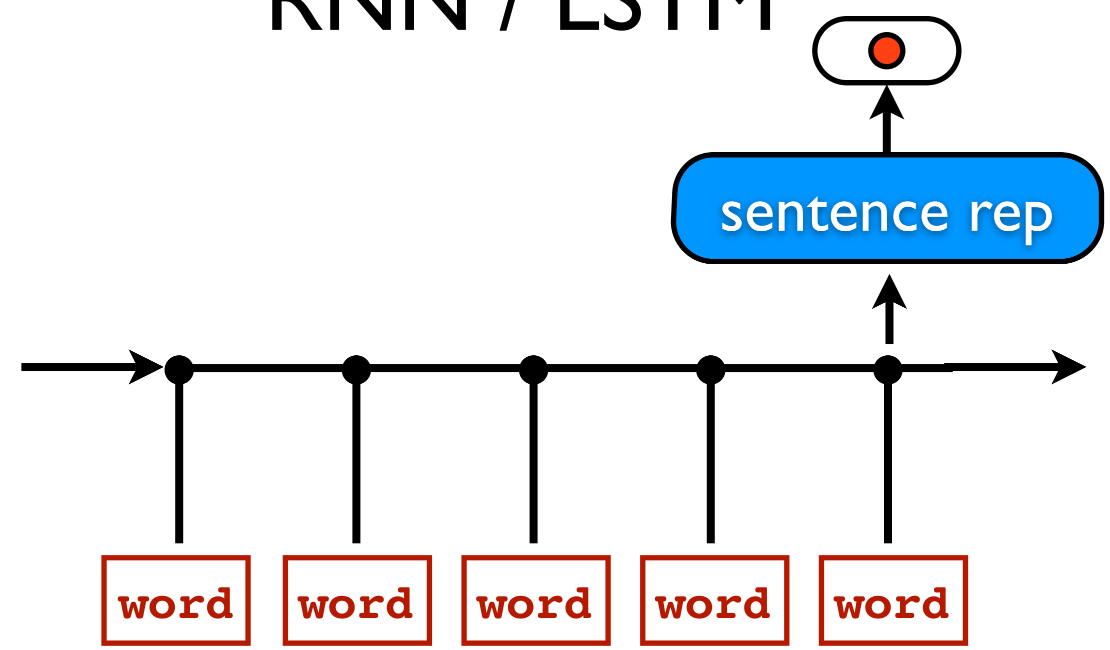
Paragraph vectors

$\vec{\P}$



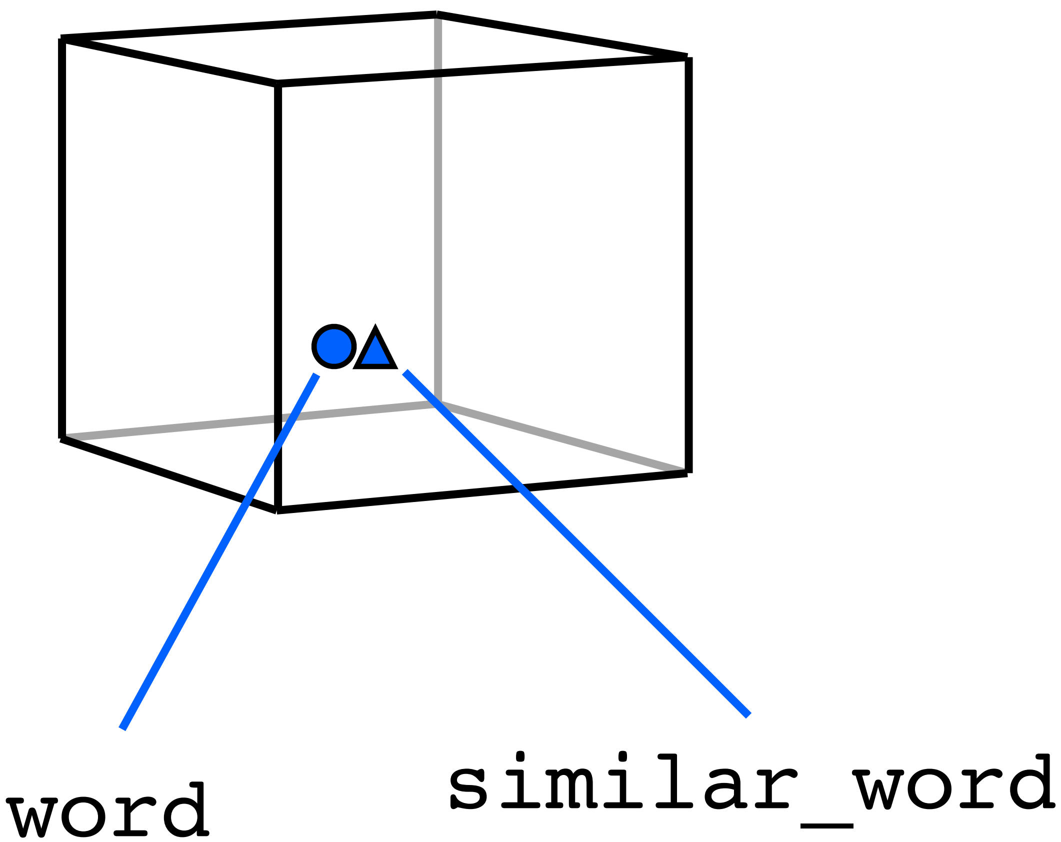
Sequential:

RNN / LSTM

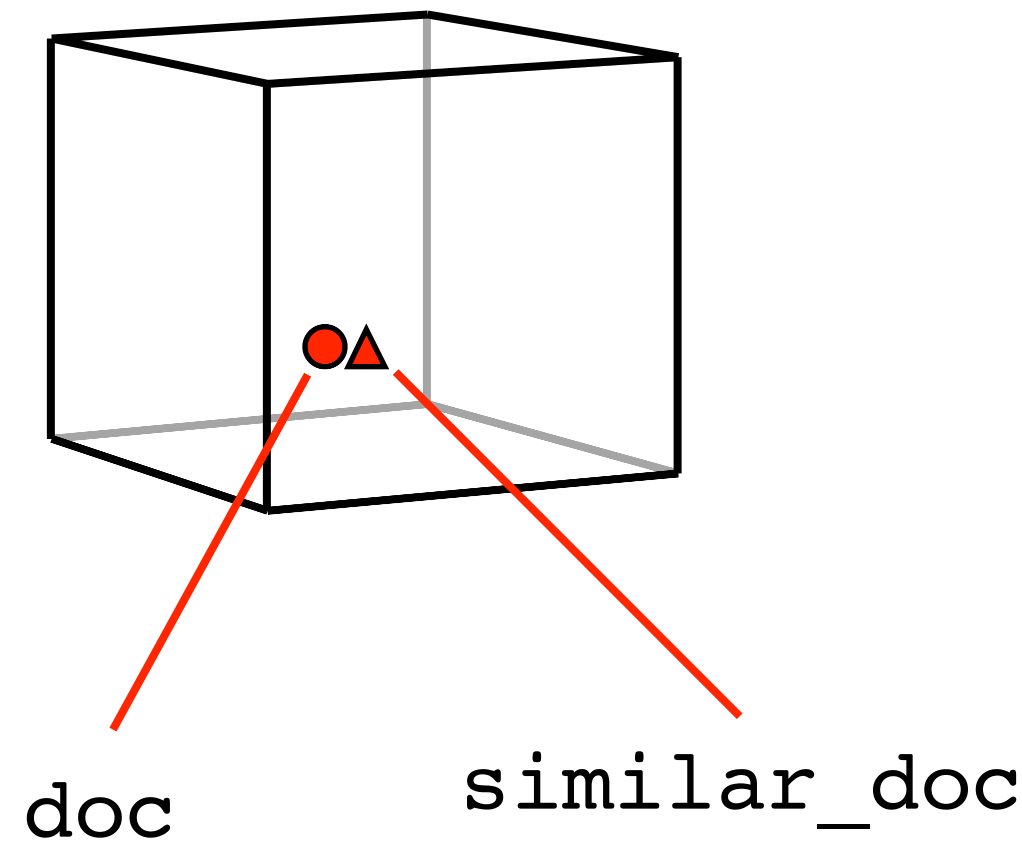


Paragraph Vectors:
Embeddings for long chunks of text.

Word vectors

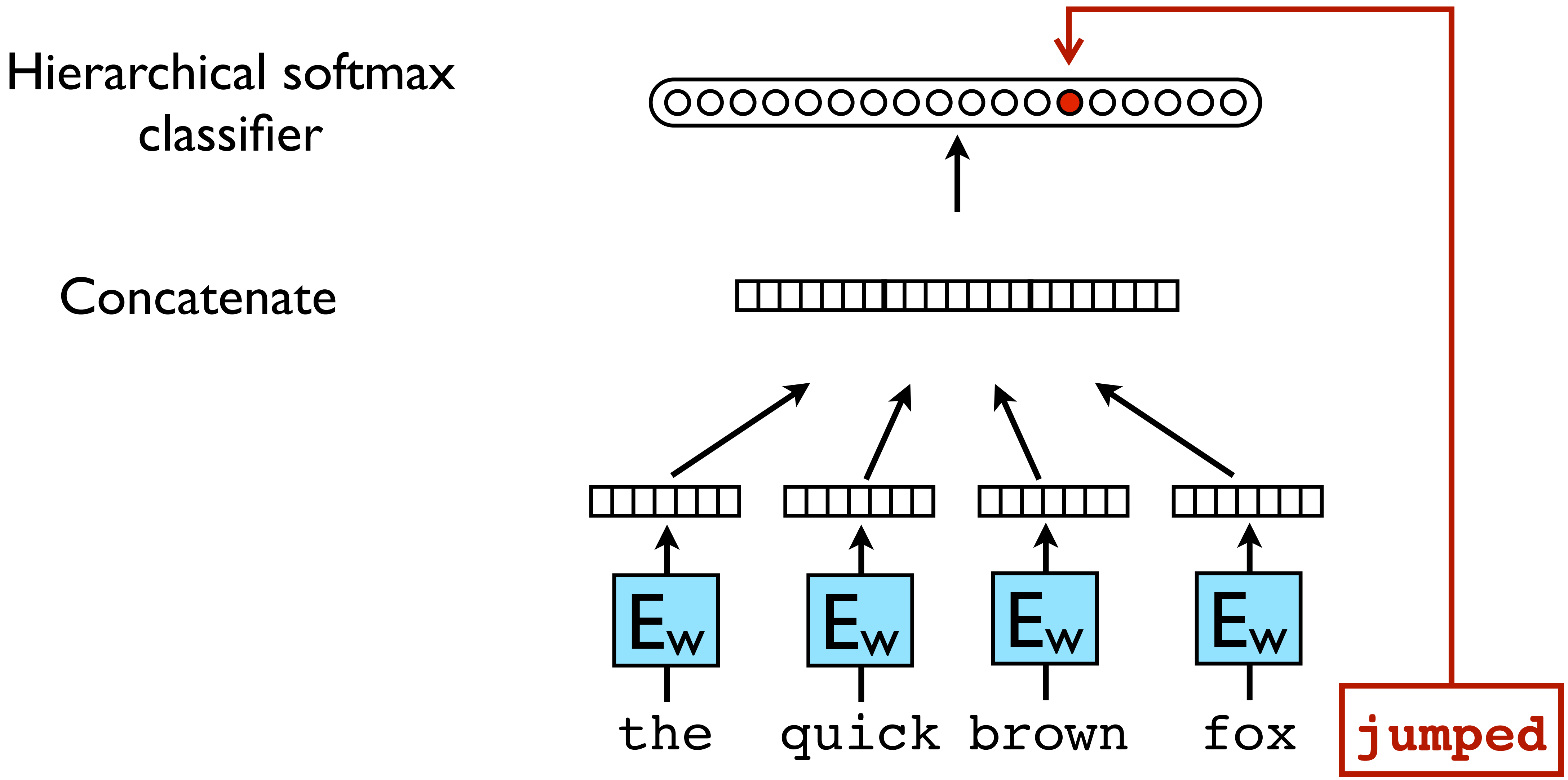


Paragraph Vectors

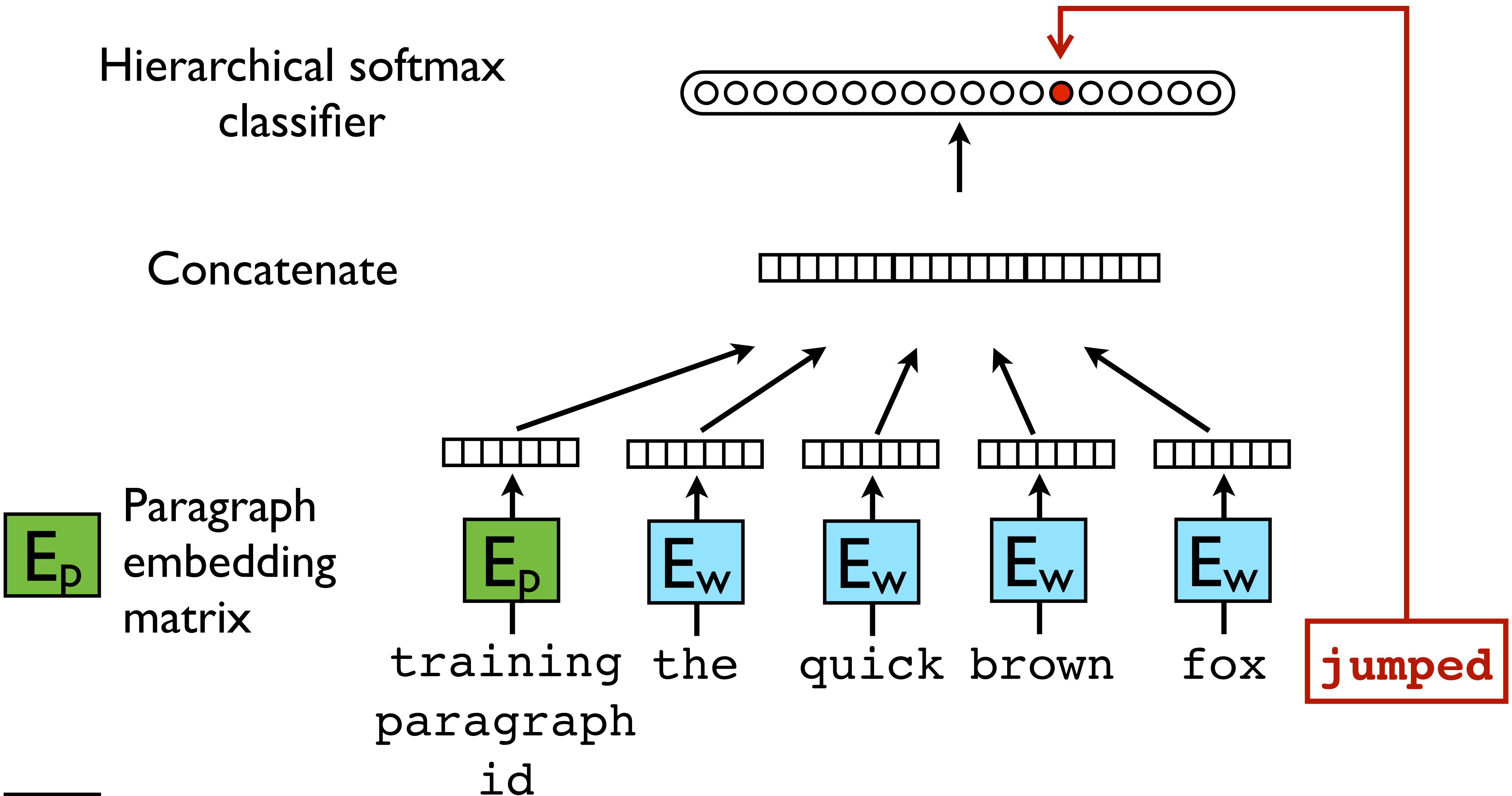


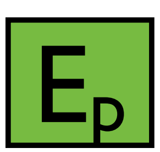
Google

Simple Language Model



Paragraph Vector Model





is a matrix of dimension ||# *training paragraphs*|| x *d*

A paragraph vector captures the complementary, non-local information that is best able to predict the next word

At inference time, for a new paragraph, hold rest of model fixed and run gradient descent on words in paragraph to obtain representation for the paragraph

Details in *Distributed Representations of Sentences and* Documents, by Quoc Le and Tomas Mikolov, ICML 2014, http://arxiv.org/abs/1405.4053

Text Classification

Sentiment analysis on IMDB reviews

50,000 training; 50,000 test

Example 1: *I had no idea of the facts this film presents. As I remember this situation I accepted the information presented then in the media: a confused happening around a dubious personality: Mr. Chavez. The film is a revelation of many realities, I wonder if something of this caliber has ever been made. I supposed the protagonist was Mr.Chavez but everyone coming up on picture

was important and at the end the reality of that entelechy: the people, was overwhelming. Thank you Kim Bartley and Donnacha O´Briain.

*

Example 2: *This movie should have NEVER been made. From the poorly done animation, to the beyond bad acting. I am not sure at what point the people behind this movie said "Ok, looks good! Lets do it!" I was in awe of how truly horrid this movie was. At one point, which very may well have been the WORST point, a computer generated Saber Tooth of gold falls from the roof stabbing the idiot creator of the cats in the mouth...uh, ooookkkk. The villain of the movie was a paralyzed sabretooth that was killed within minutes of its first appearance. The other two manages to kill a handful of people prior to being burned and gunned down. Then, there is a random one awaiting victims in the jungle...which scares me for one sole reason. Will there be a Part Two? God, for the sake of humans everywhere I hope not.

This movie was pure garbage. From the power point esquire credits to the slide show ending.*

Results for IMDB Sentiment Classification (long paragraphs)

Method	Error rate
Bag of words	12.2%
Bag of words + idf	11.8%
LDA	32.6%
LSA	16.1%
Average word vectors	18%
Bag of words + word vectors	11.7%
Bag of words + word vectors + more tweaks	11.1%
Bag of words + bigrams + Naive Bayes SVM	9%
Paragraph vectors	**7.5%**

Important side note:
"Paragraph vectors" can be computed for things that are not paragraphs. In particular:

sentences
whole documents
users
products
movies
audio waveforms
...

Paragraph Vectors:

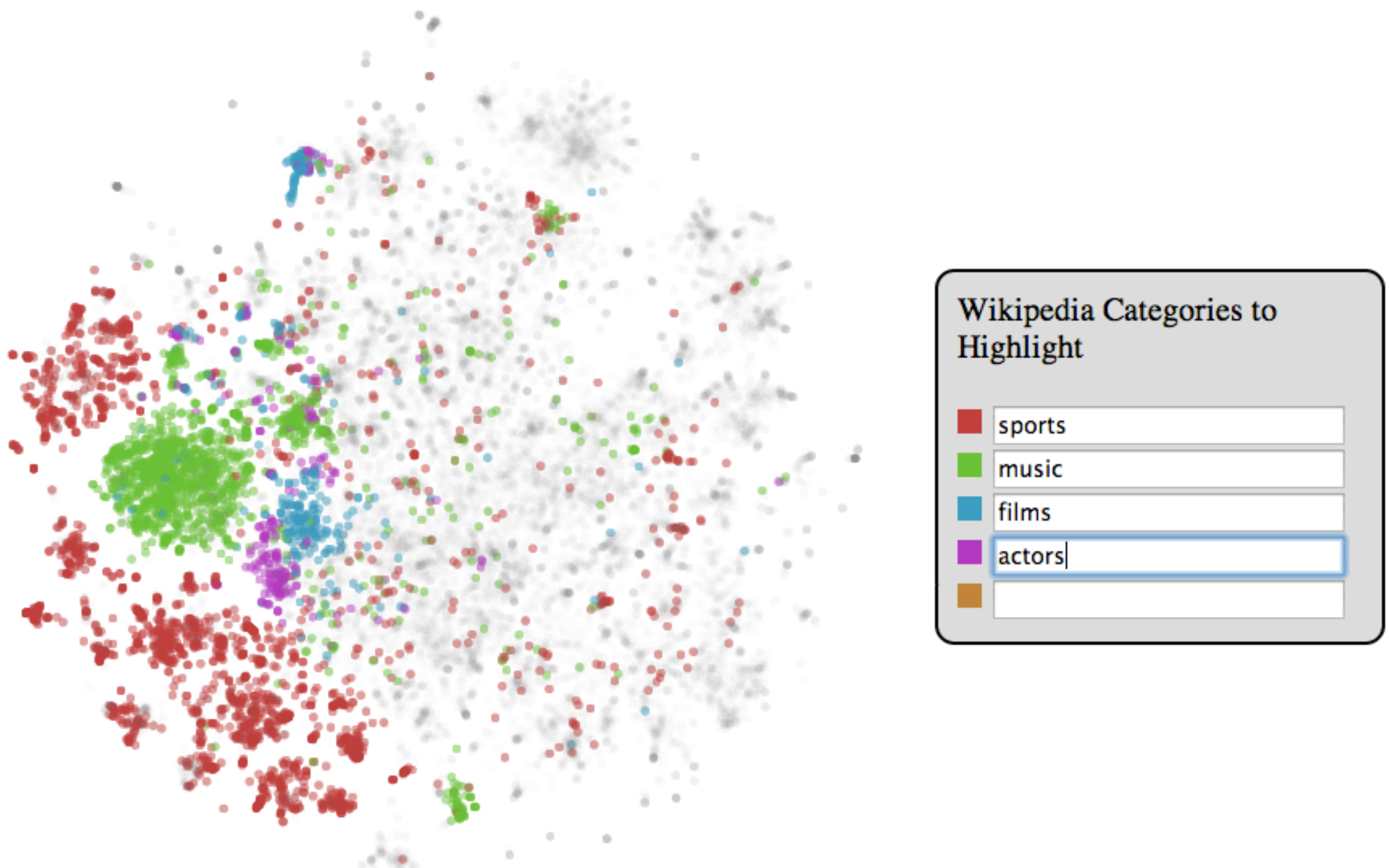
Train on Wikipedia articles

Nearest neighbor articles to article for "`Machine Learning`"

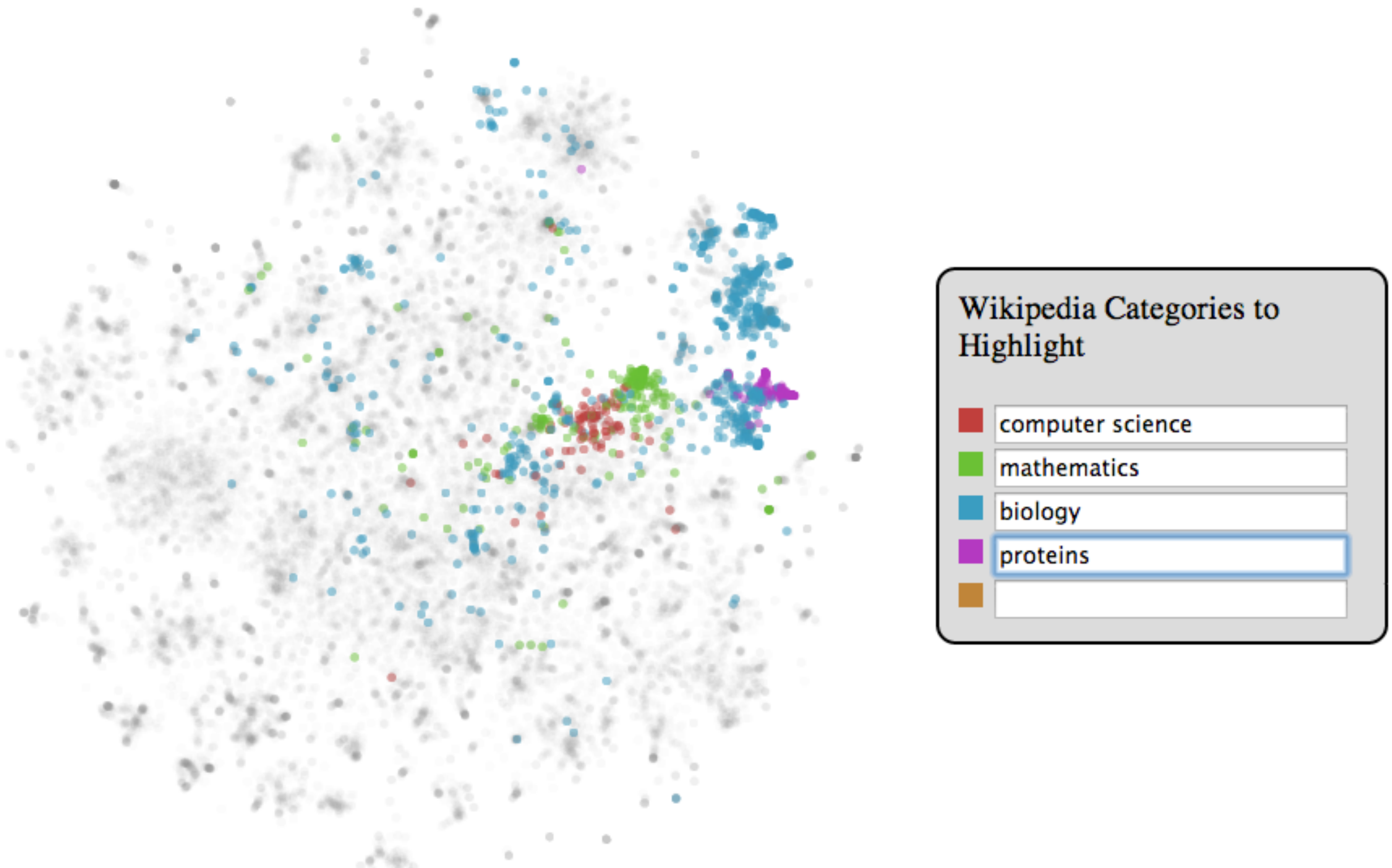
LDA	Paragraph Vectors
Artificial neural network	Artificial neural network
Predictive analytics	Types of artificial neural networks
Structured prediction	Unsupervised learning
Mathematical geophysics	Feature learning
Supervised learning	Predictive analytics
Constrained conditional model	Pattern recognition
Sensitivity analysis	Statistical classification
SXML	Structured prediction
Feature scaling	Training set
Boosting (machine learning)	Meta learning (computer science)
Prior probability	Kernel method
Curse of dimensionality	Supervised learning
Scientific evidence	Generalization error
Online machine learning	Overfitting
N-gram	Multi-task learning
Cluster analysis	Generative model
Dimensionality reduction	Computational learning theory
Functional decomposition	Inductive bias
Bayesian network	Semi-supervised learning

Google

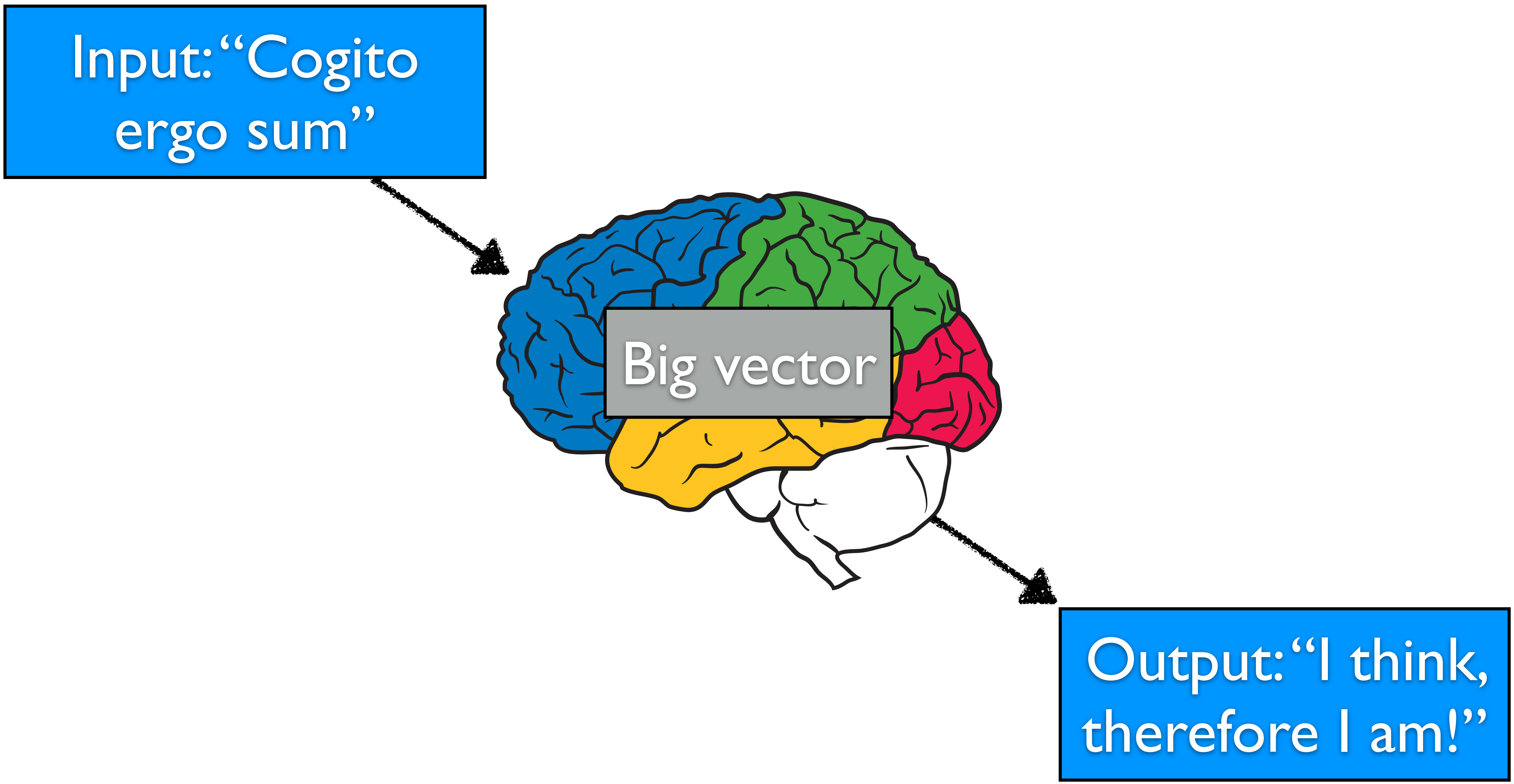
Wikipedia Article Paragraph Vectors visualized via t-SNE



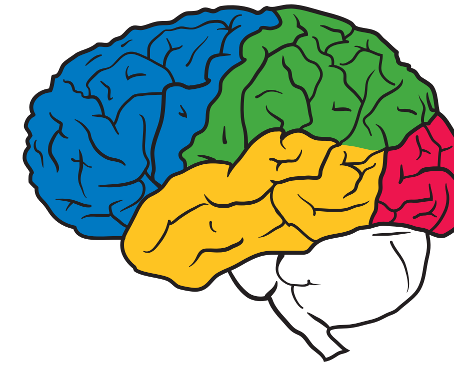
Wikipedia Article Paragraph Vectors visualized via t-SNE



Example of LSTM-based representation: Machine Translation

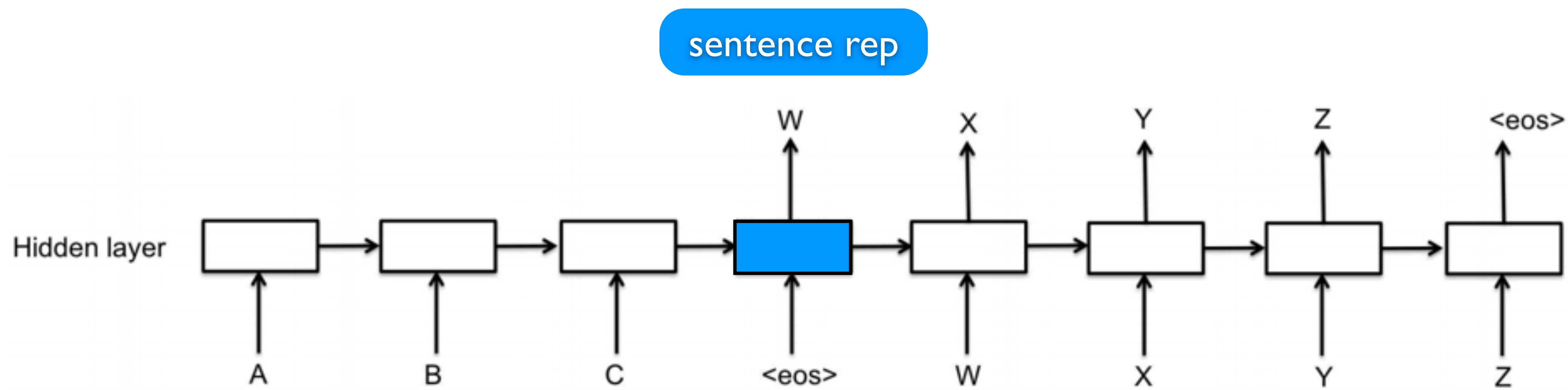


LSTM for End to End Translation



Source Language: A B C

Target Language: W X Y Z



See: *Sequence to Sequence Learning with Neural Networks*, Ilya Sutskever, Oriol Vinyals, and Quoc Le. http://arxiv.org/abs/1409.3215. To appear in NIPS, 2014.

Google

Example Translation

- Google Translate:

As Reuters noted for the first time in July, the seating configuration is exactly what fuels the battle between the latest devices.

- Neural LSTM model:

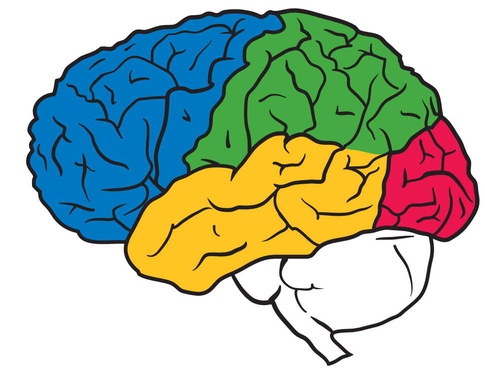
As Reuters reported for the first time in July, the configuration of seats is exactly what drives the battle between the latest aircraft.

- Human translation:

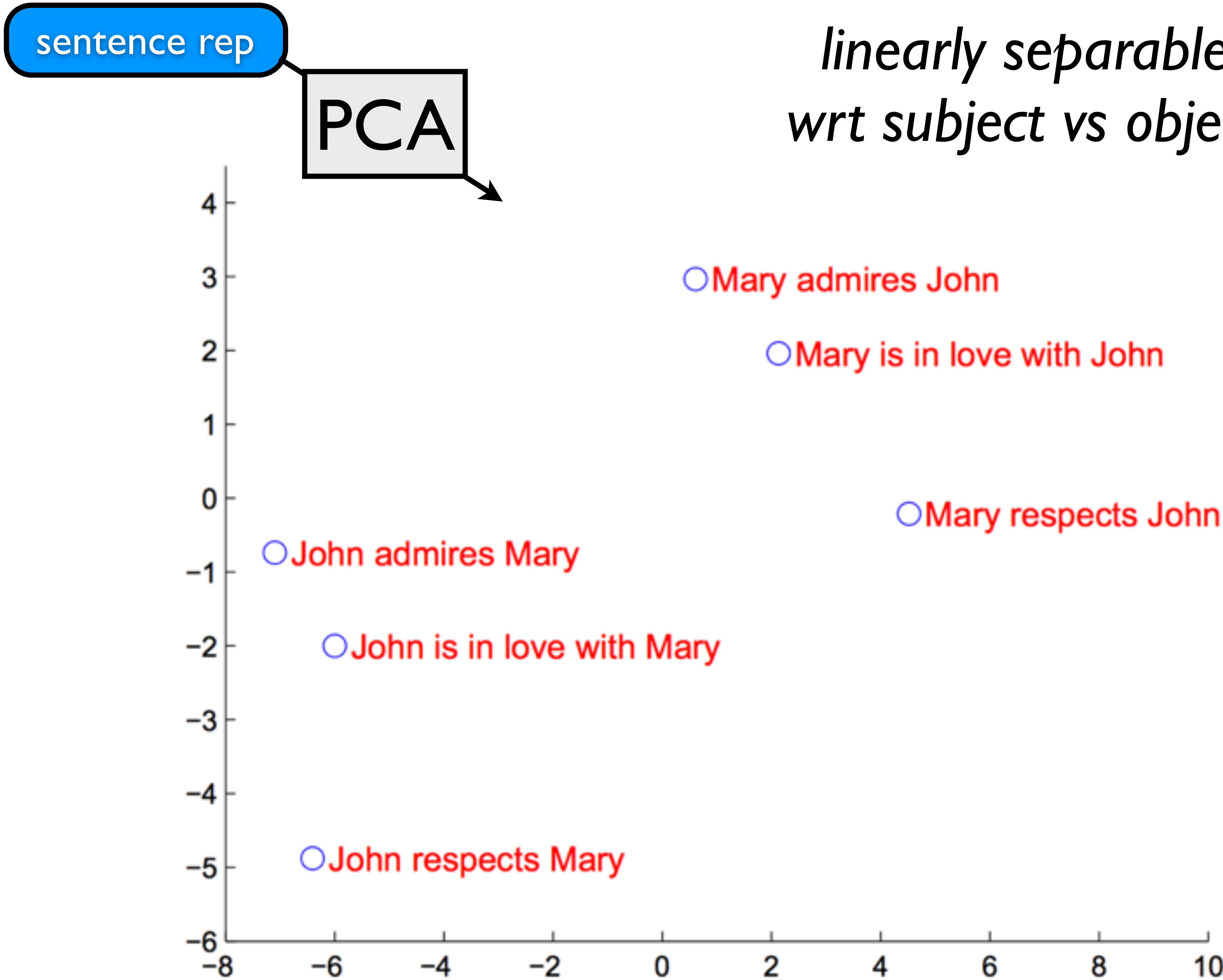
As Reuters first reported in July, seat layout is exactly what drives the battle between the latest jets.

Google

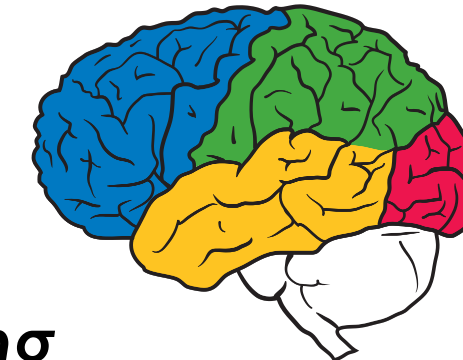
LSTM for End to End Translation



linearly separable
wrt subject vs object



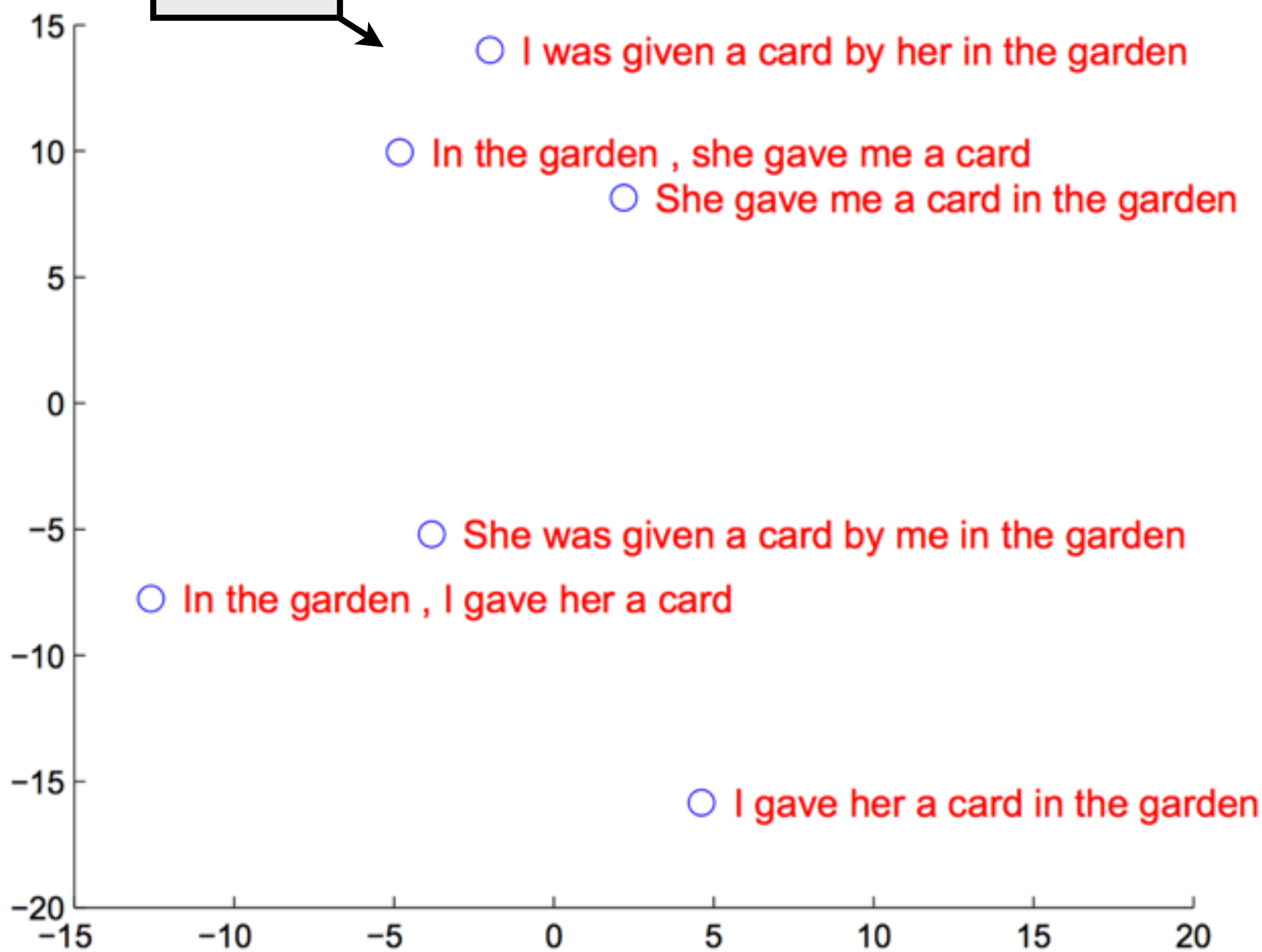
LSTM for End to End Translation



sentence rep

PCA

mostly invariant to paraphasing



Combining modalities
e.g. vision and language

Generating Image Captions from Pixels

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

Model sample 1: A close up of a child holding a stuffed animal.

Model sample 2: A baby is asleep next to a teddy bear.

Google

Work in progress by Oriol Vinyals *et al.*

Generating Image Captions from Pixels

Human: Three different types of pizza on top of a stove.

Model sample 1: Two pizzas sitting on top of a stove top oven.

Model sample 2: A pizza sitting on top of a pan on top of a stove.

Google

Generating Image Captions from Pixels

Human: A green monster kite soaring in a sunny sky.

Model: A man flying through the air while riding a skateboard.

Google

Generating Image Captions from Pixels

Human: A tennis player getting ready to serve the ball.

Model: A man holding a tennis racquet on a tennis court.

Google

Conclusions

- Deep neural networks are very effective for wide range of tasks
  - By using 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
  - Effective in many domains: speech, vision, language modeling, user prediction, language understanding, translation, advertising, …

An important tool in building intelligent systems.

Google

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, Corrado, et al. , *Large Scale Distributed Deep Networks,* NIPS 2012.
- Mikolov, Chen, Corrado and Dean. *Efficient Estimation of Word Representations in Vector Space,* http://arxiv.org/abs/1301.3781.
- *Distributed Representations of Sentences and* Documents, by Quoc Le and Tomas Mikolov, ICML 2014, http://arxiv.org/abs/1405.4053
- Vanhoucke, Devin and Heigold. *Deep Neural Networks for Acoustic Modeling*, ICASSP 2013.
- *Sequence to Sequence Learning with Neural Networks*, Ilya Sutskever, Oriol Vinyals, and Quoc Le. http://arxiv.org/abs/1409.3215. To appear in NIPS, 2014.
- http://research.google.com/papers
- http://research.google.com/people/jeff

Google