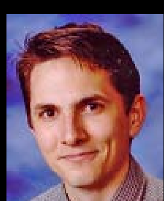
# Machine Learning and AI via Brain simulations


## Andrew Ng

Stanford University



Thanks to:



Adam Coates

Quoc Le

Honglak Lee

Andrew Saxe

Andrew Maas

Chris Manning

Jiquan Ngiam

Richard Socher

Will Zou

Google: Kai Chen, Greg Corrado, Jeff Dean, Matthieu Devin, Andrea Frome, Rajat Monga, Marc'Aurelio Ranzato, Paul Tucker, Kay Le

# This talk: Deep Learning

Using brain simulations:

- Make learning algorithms much better and easier to use.
- Make revolutionary advances in machine learning and AI.

Vision shared with many researchers:

E.g., Samy Bengio, Yoshua Bengio, Tom Dean, Jeff Dean, Nando de Freitas, Jeff Hawkins, Geoff Hinton, Quoc Le, Yann LeCun, Honglak Lee, Tommy Poggio, Marc'Aurelio Ranzato, Ruslan Salakhutdinov, Josh Tenenbaum, Kai Yu, Jason Weston, ….

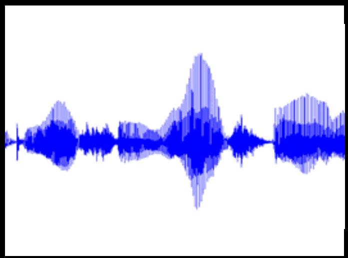
I believe this is our best shot at progress towards real AI.

# What do we want computers to do with our data?

Images/video

Label: "Motorcycle"
Suggest tags
Image search
…

Audio

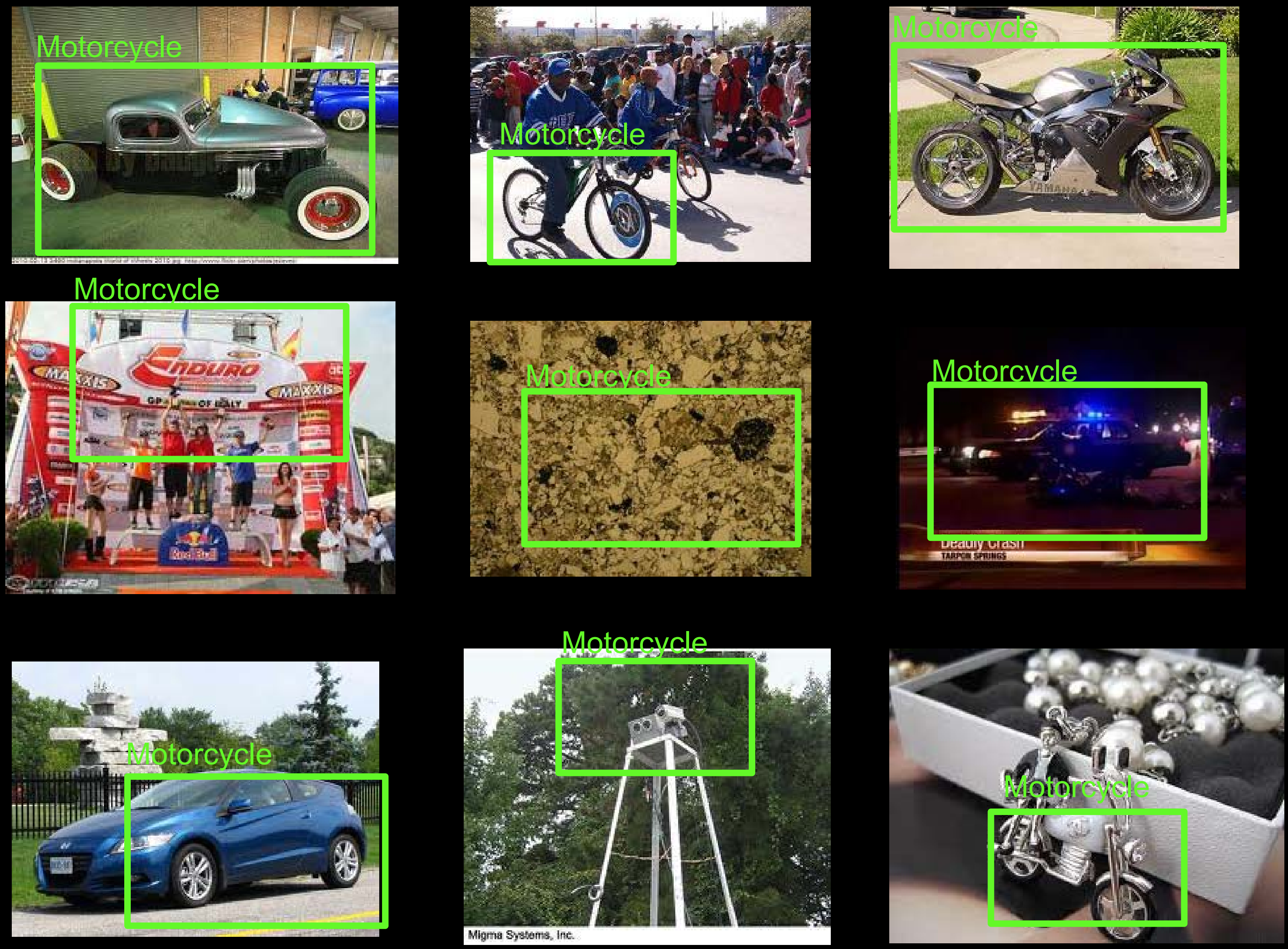


Speech recognition
Music classification
Speaker identification
…

Text


Web search
Anti-spam
Machine translation
…

# Computer vision is hard!



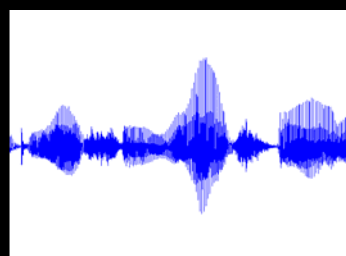

Andrew Ng

# What do we want computers to do with our data?

Images/video

Label: “Motorcycle”
Suggest tags
Image search
…

Audio



Speech recognition
Speaker identification
Music classification
…

Text

The New York Times

Web search
Anti-spam
Machine translation
…

Machine learning performs well on many of these problems, but is a lot of work. What is it about machine learning that makes it so hard to use?

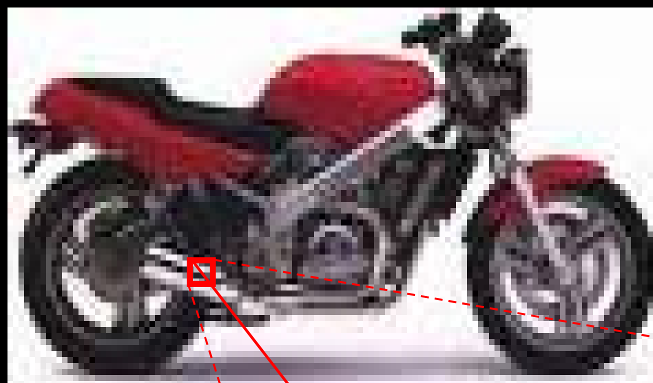
# Machine learning for image classification

→ "Motorcycle"

This talk: Develop ideas using images and audio.
Ideas apply to other problems (e.g., text) too.

# Why is this hard?

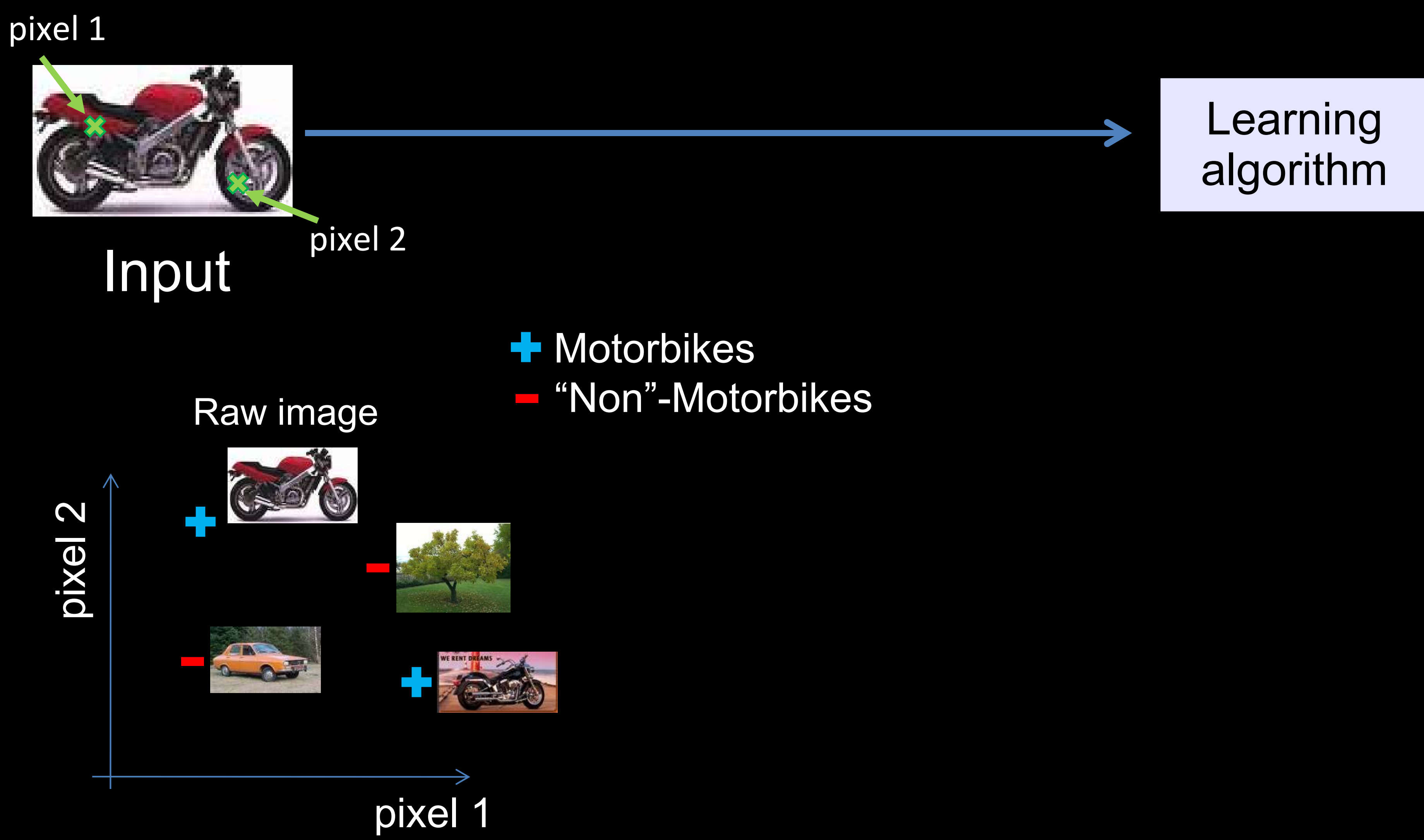
You see this:



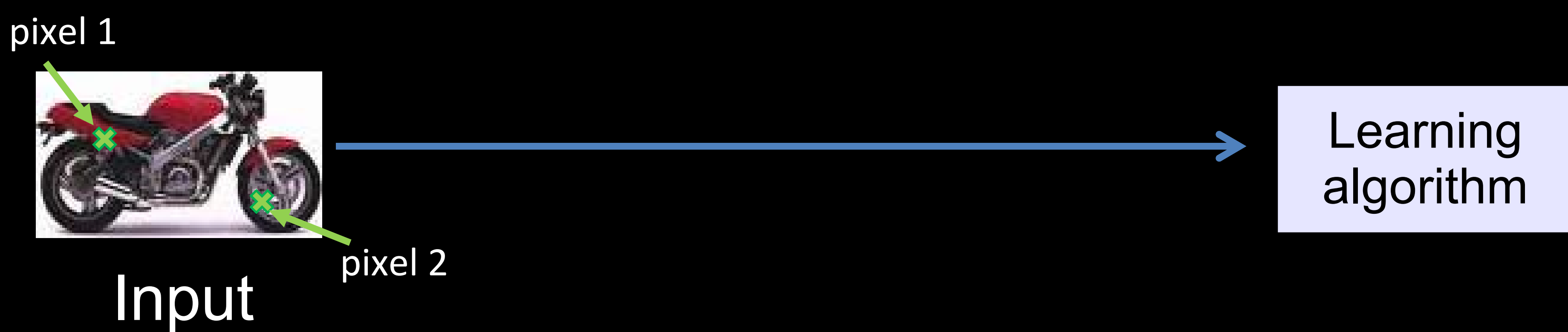
But the camera sees this:

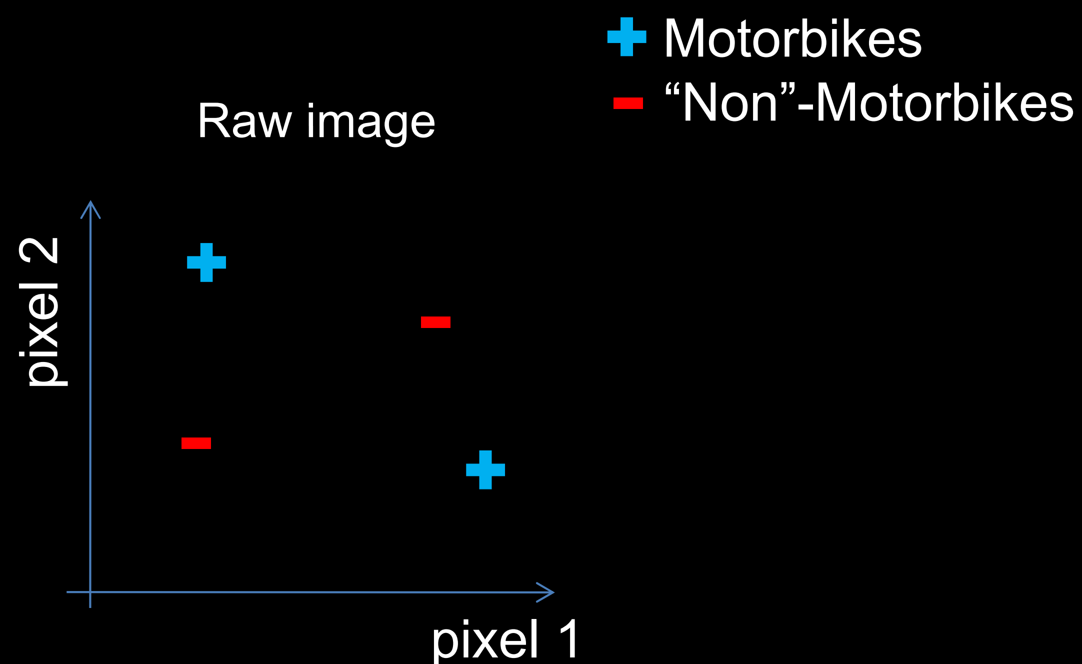
| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 194 | 210 | 201 | 212 | 199 | 213 | 215 | 195 | 178 | 158 | 182 | 209 |
| 180 | 189 | 190 | 221 | 209 | 205 | 191 | 167 | 147 | 115 | 129 | 163 |
| 114 | 126 | 140 | 188 | 176 | 165 | 152 | 140 | 170 | 106 | 78 | 88 |
| 87 | 103 | 115 | 154 | 143 | 142 | 149 | 153 | 173 | 101 | 57 | 57 |
| 102 | 112 | 106 | 131 | 122 | 138 | 152 | 147 | 128 | 84 | 58 | 66 |
| 94 | 95 | 79 | 104 | 105 | 124 | 129 | 113 | 107 | 87 | 69 | 67 |
| 68 | 71 | 69 | 98 | 89 | 92 | 98 | 95 | 89 | 88 | 76 | 67 |
| 41 | 56 | 68 | 99 | 63 | 45 | 60 | 82 | 58 | 76 | 75 | 65 |
| 20 | 43 | 69 | 75 | 56 | 41 | 51 | 73 | 55 | 70 | 63 | 44 |
| 50 | 50 | 57 | 69 | 75 | 75 | 73 | 74 | 53 | 68 | 59 | 37 |
| 72 | 59 | 53 | 66 | 84 | 92 | 84 | 74 | 57 | 72 | 63 | 42 |
| 67 | 61 | 58 | 65 | 75 | 78 | 76 | 73 | 59 | 75 | 69 | 50 |

# Machine learning and feature representations

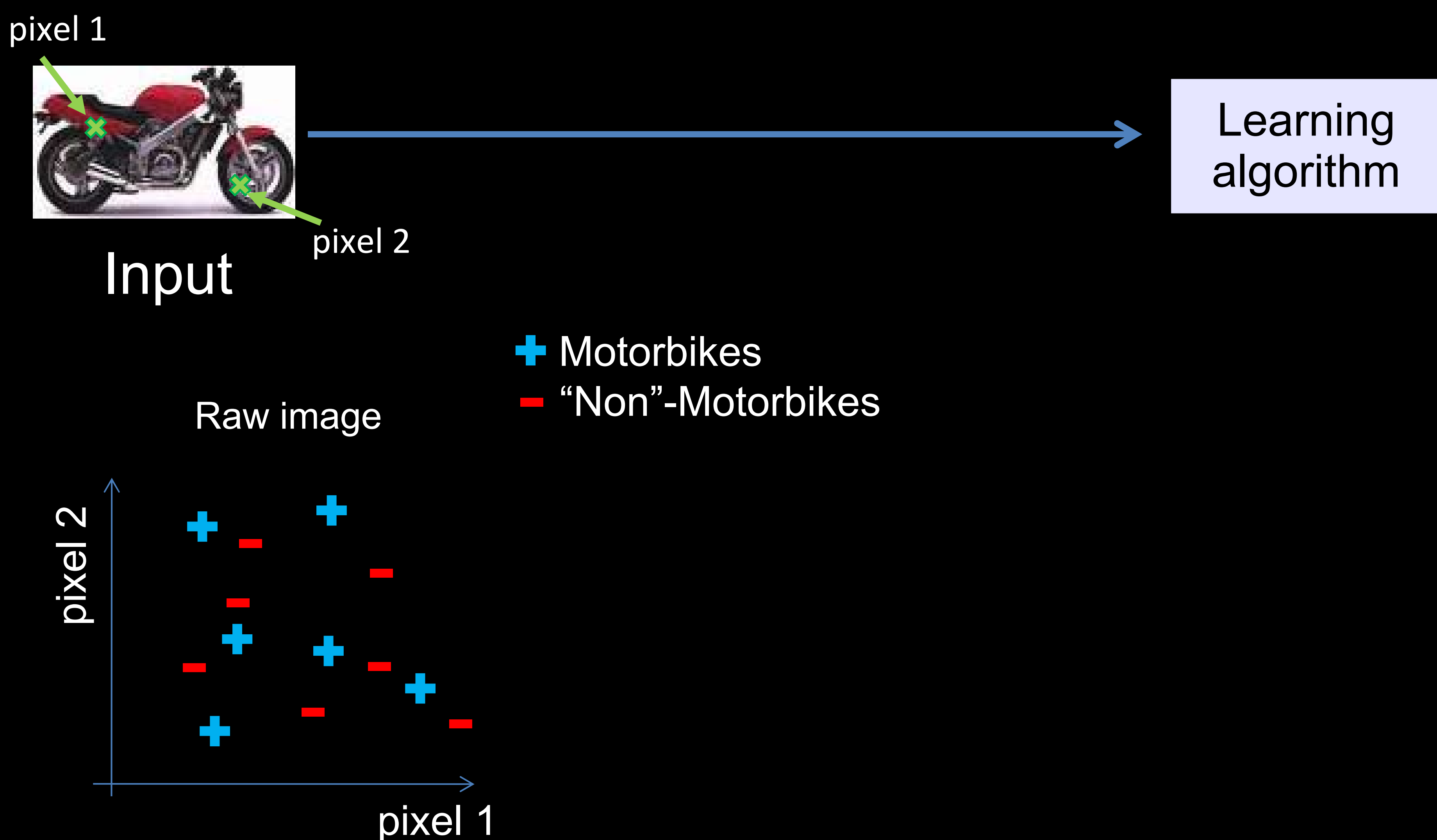


# Machine learning and feature representations



# Machine learning and feature representations



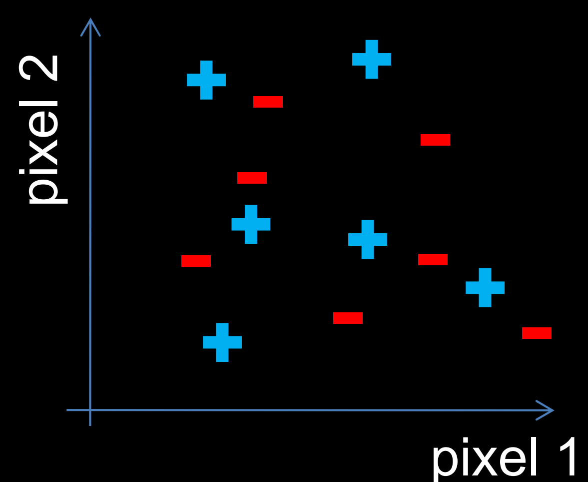
# What we want


Input → Feature representation → Learning algorithm

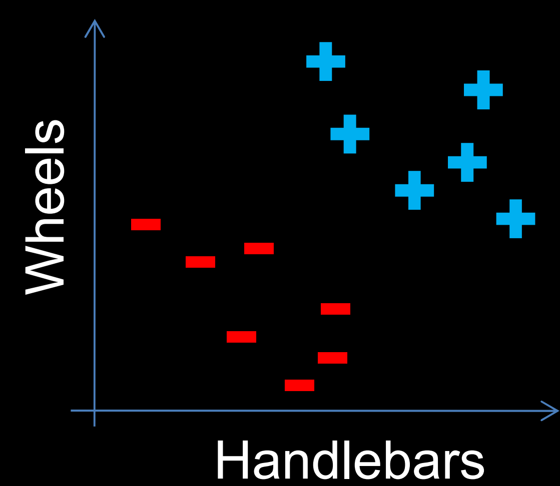
E.g., Does it have Handlebars? Wheels?

+ Motorbikes
– “Non”-Motorbikes

Raw image




Features

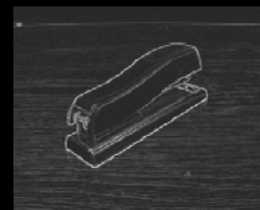



Andrew Ng

# How is computer perception done?

Images/video

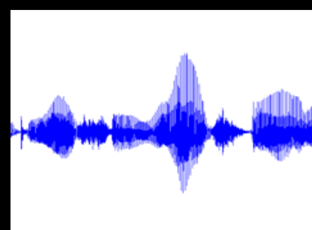
Image



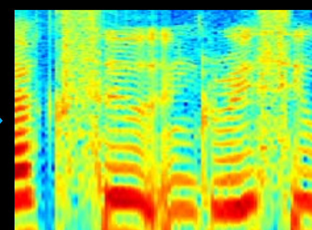
Vision features

Detection

Audio



Audio

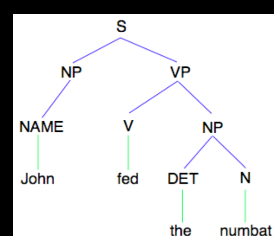


Audio features

Speaker ID

Text

Text

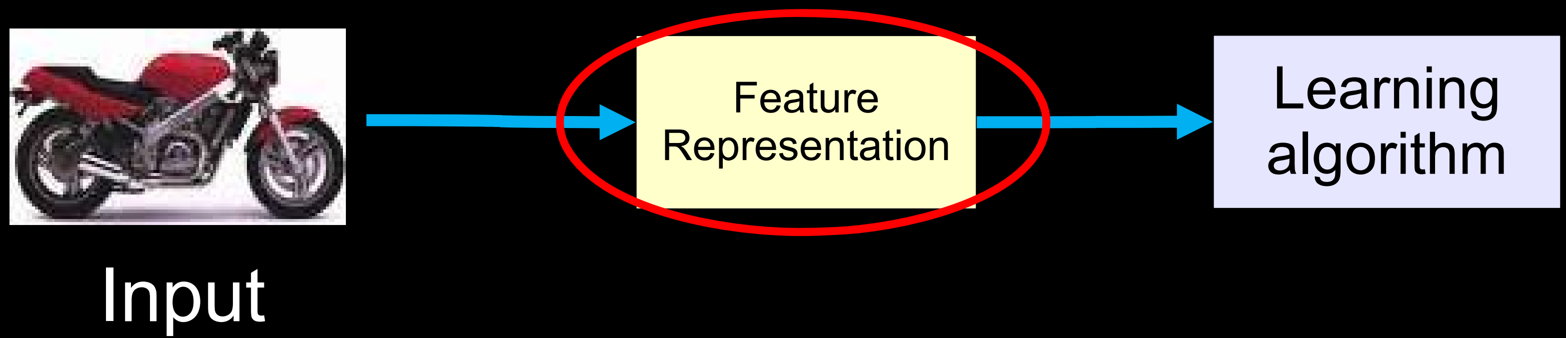



Text features

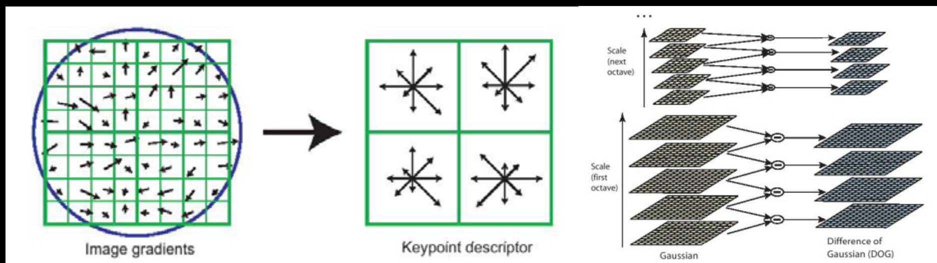
Text classification,
Machine translation,
Information retrieval,
....

Andrew Ng

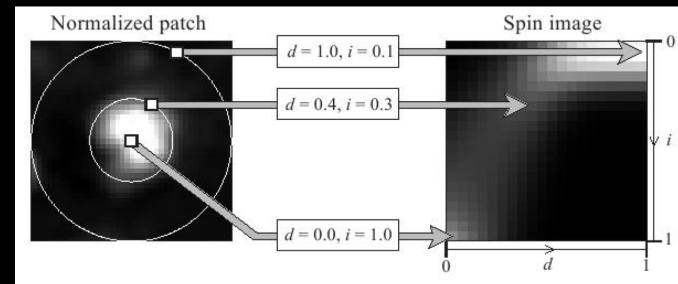
# Feature representations



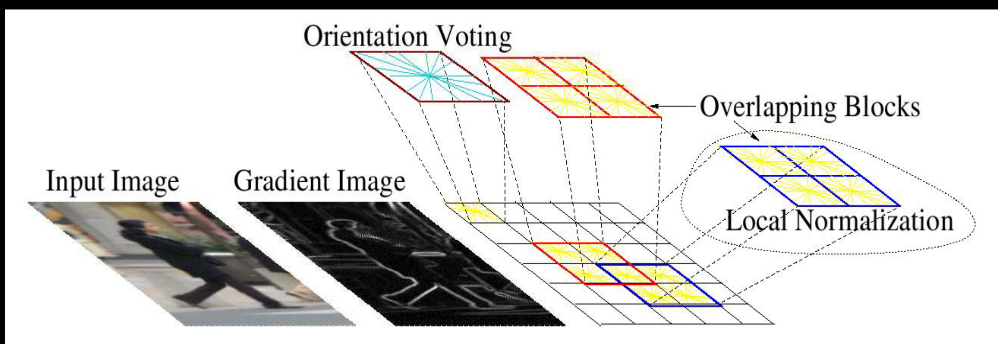
# Computer vision features



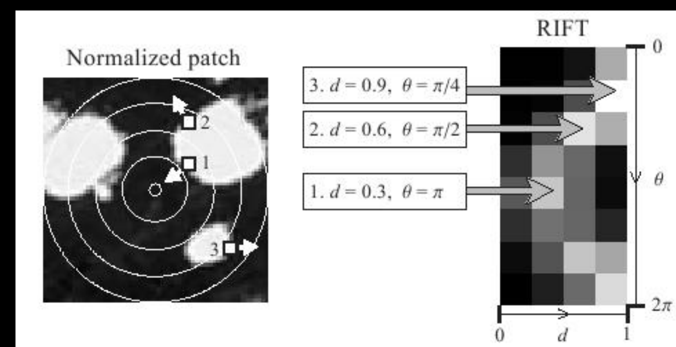

SIFT



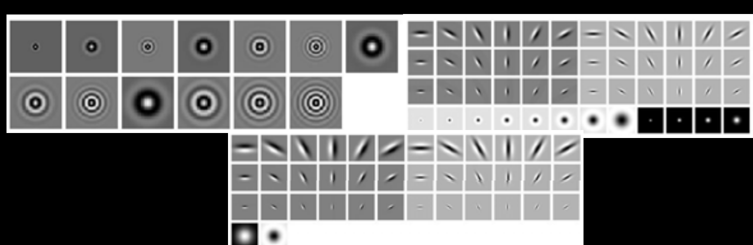

Spin image



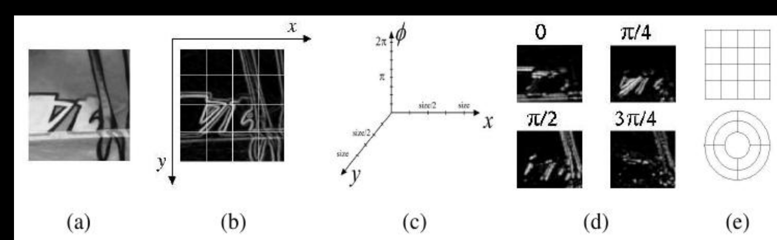

HoG




RIFT

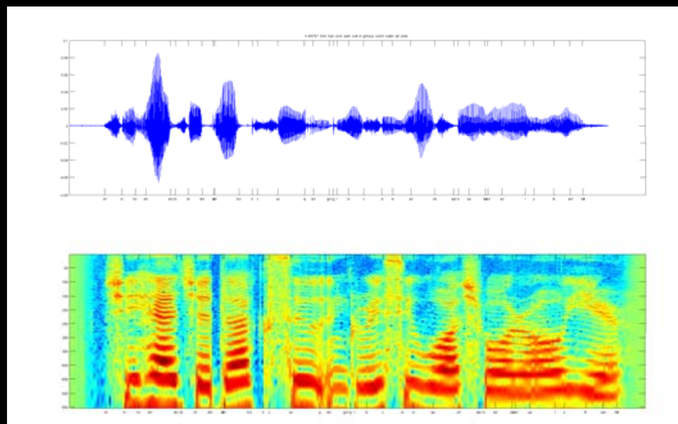


Textons

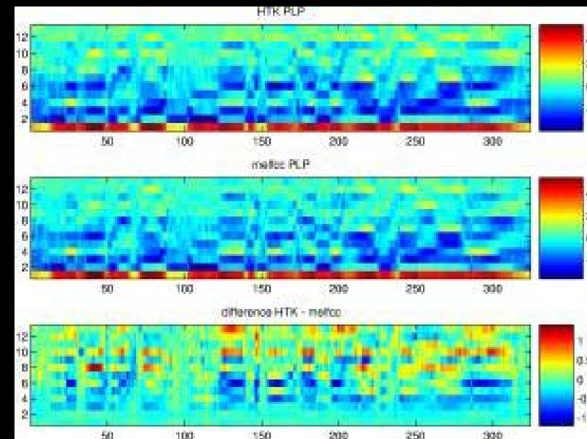



GLOH

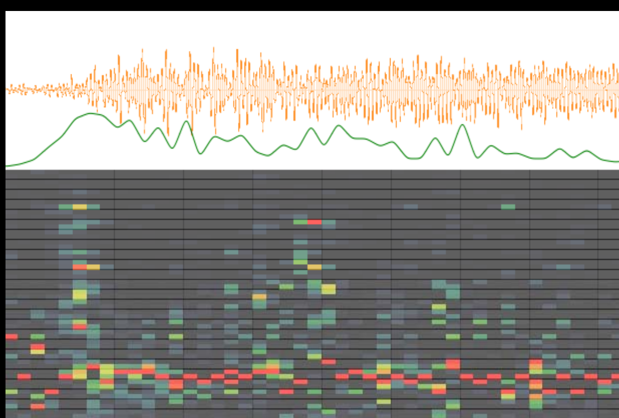
# Audio features



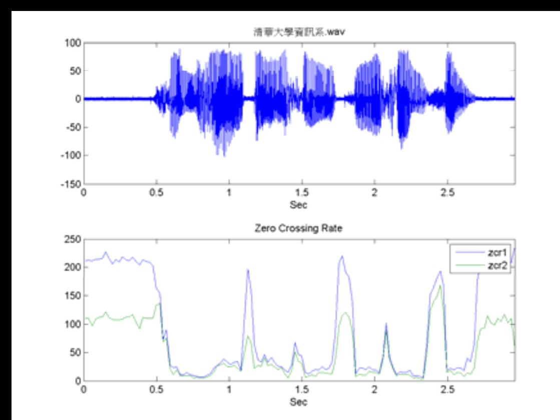
Spectrogram



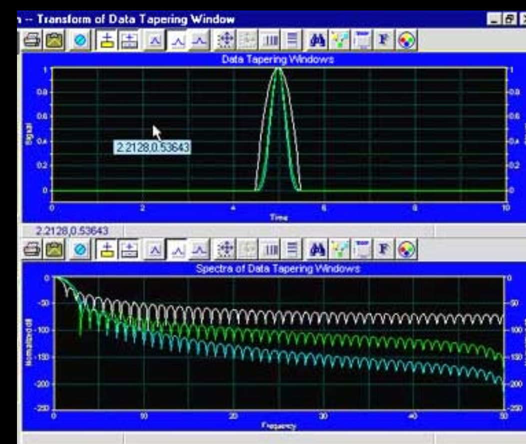

MFCC



Flux



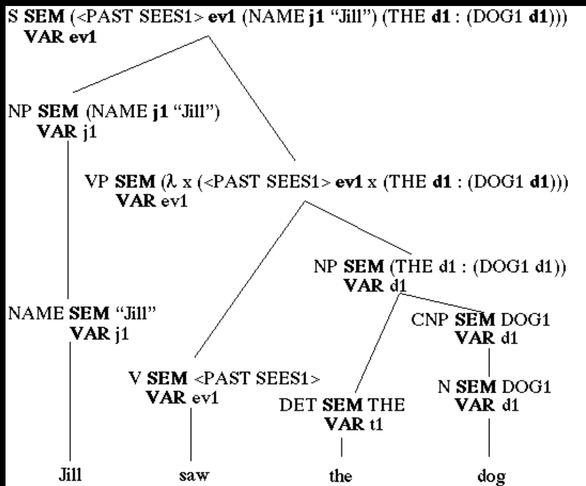

ZCR



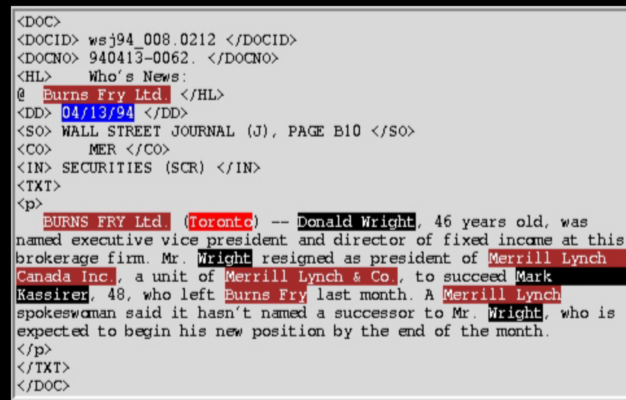

Rolloff

Andrew Ng

# NLP features

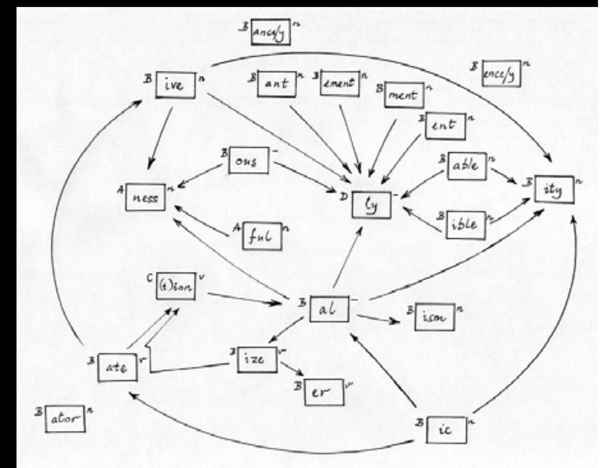



Pars



<DOC>
<DOCID> wsj94_008.0212 </DOCID>
<DOCNO> 940413-0062. </DOCNO>
<HL> Who's News:
@ Burns Fry Ltd. </HL>
<DD> 04/13/94 </DD>
<SO> WALL STREET JOURNAL (J), PAGE B10 </SO>
<CO> MER </CO>
<IN> SECURITIES (SCR) </IN>
<TXT>
<p>
BURNS FRY Ltd. (Toronto) -- Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm. Mr. Wright resigned as president of Merrill Lynch Canada Inc., a unit of Merrill Lynch & Co., to succeed Mark Kassirer, 48, who left Burns Fry last month. A Merrill Lynch spokeswoman said it hasn't named a successor to Mr. Wright, who is expected to begin his new position by the end of the month.
</p>
</TXT>
</DOC>

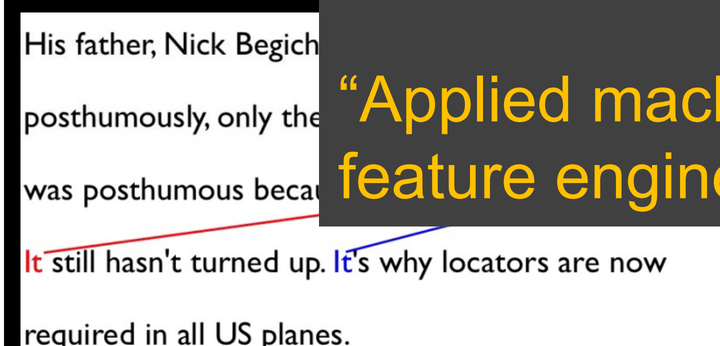
Named entity recognition



Stemming

Coming up with features is difficult, time-consuming, requires expert knowledge.

"Applied machine learning" is basically feature engineering.

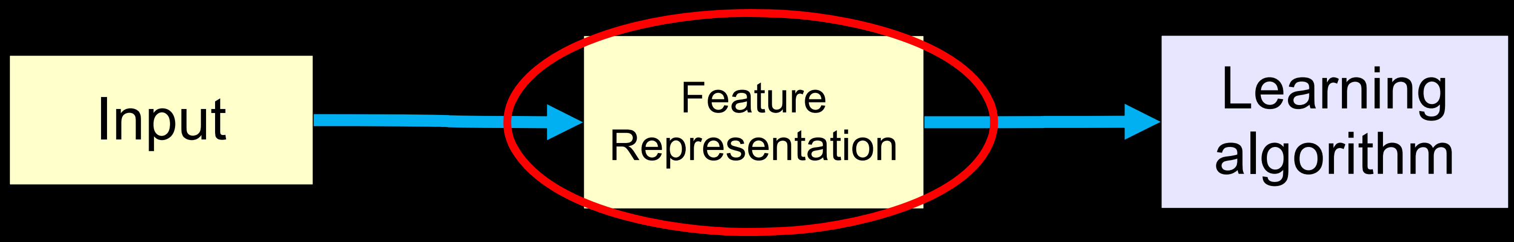



Anaphora

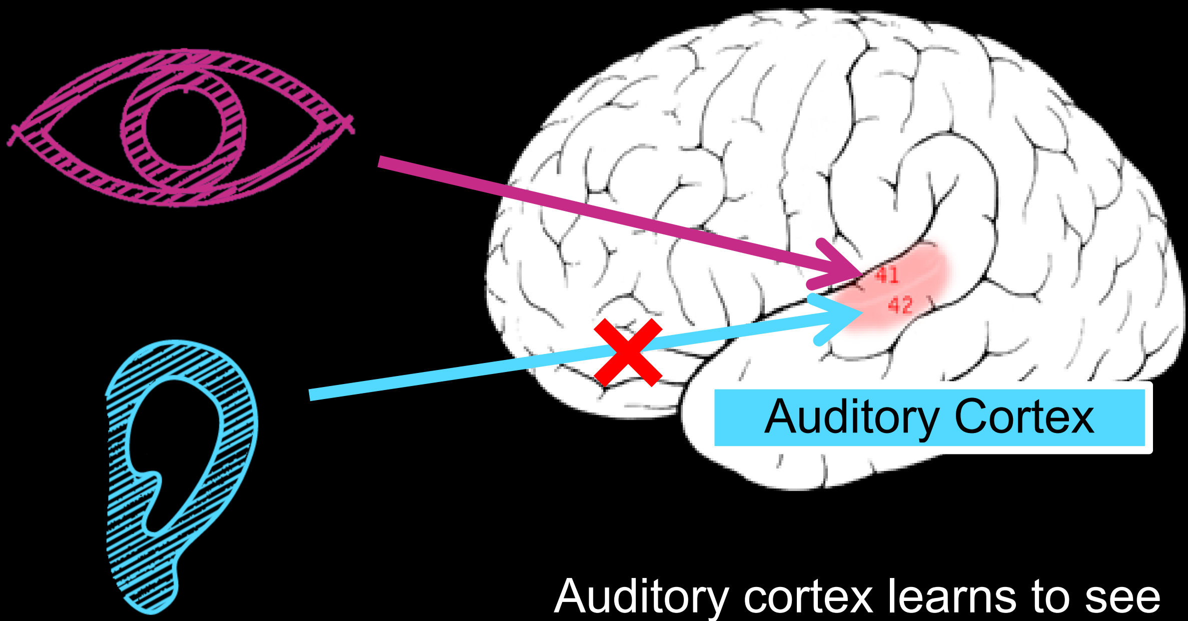

Part of speech


Ontologies (WordNet)

# Feature representations



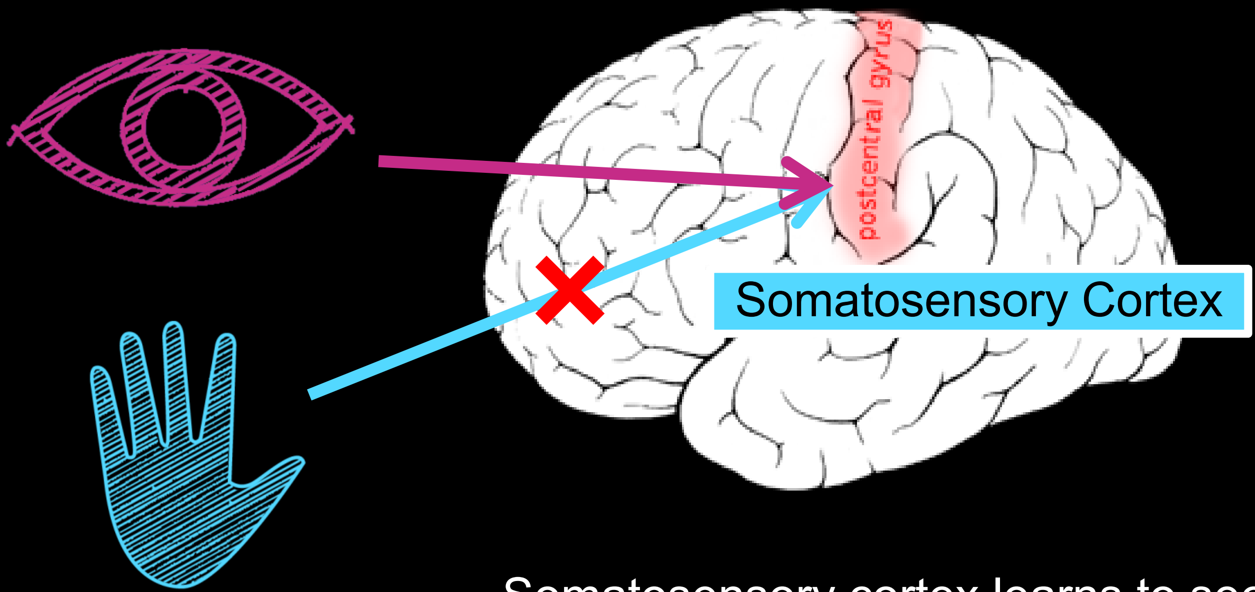
# The "one learning algorithm" hypothesis




Auditory cortex learns to see

[Roe et al., 1992]

# The "one learning algorithm" hypothesis




Somatosensory cortex learns to see

[Metin & Frost, 1989]

# Sensor representations in the brain

Seeing with your tongue

Human echolocation (sonar)

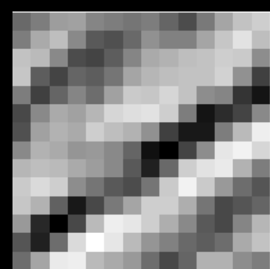

Haptic belt: Direction sense


Implanting a 3rd eye

[BrainPort; Welsh & Blasch, 1997; Nagel et al., 2005; Constantine-Paton & Law, 2009]

# Feature learning problem

- Given a 14x14 image patch x, can represent it using 196 real numbers.



255
98
93
87
89
91
48
…

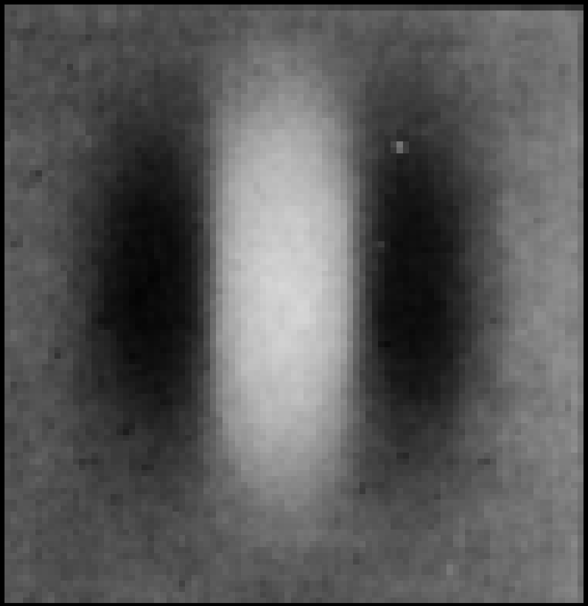
- Problem: Can we find a learn a better feature vector to represent this?

# First stage of visual processing: V1

V1 is the first stage of visual processing in the brain.

Neurons in V1 typically modeled as edge detectors:

Neuron #1 of visual cortex
(model)



Neuron #2 of visual cortex
(model)

# Learning sensor representations

Sparse coding (Olshausen & Field,1996)

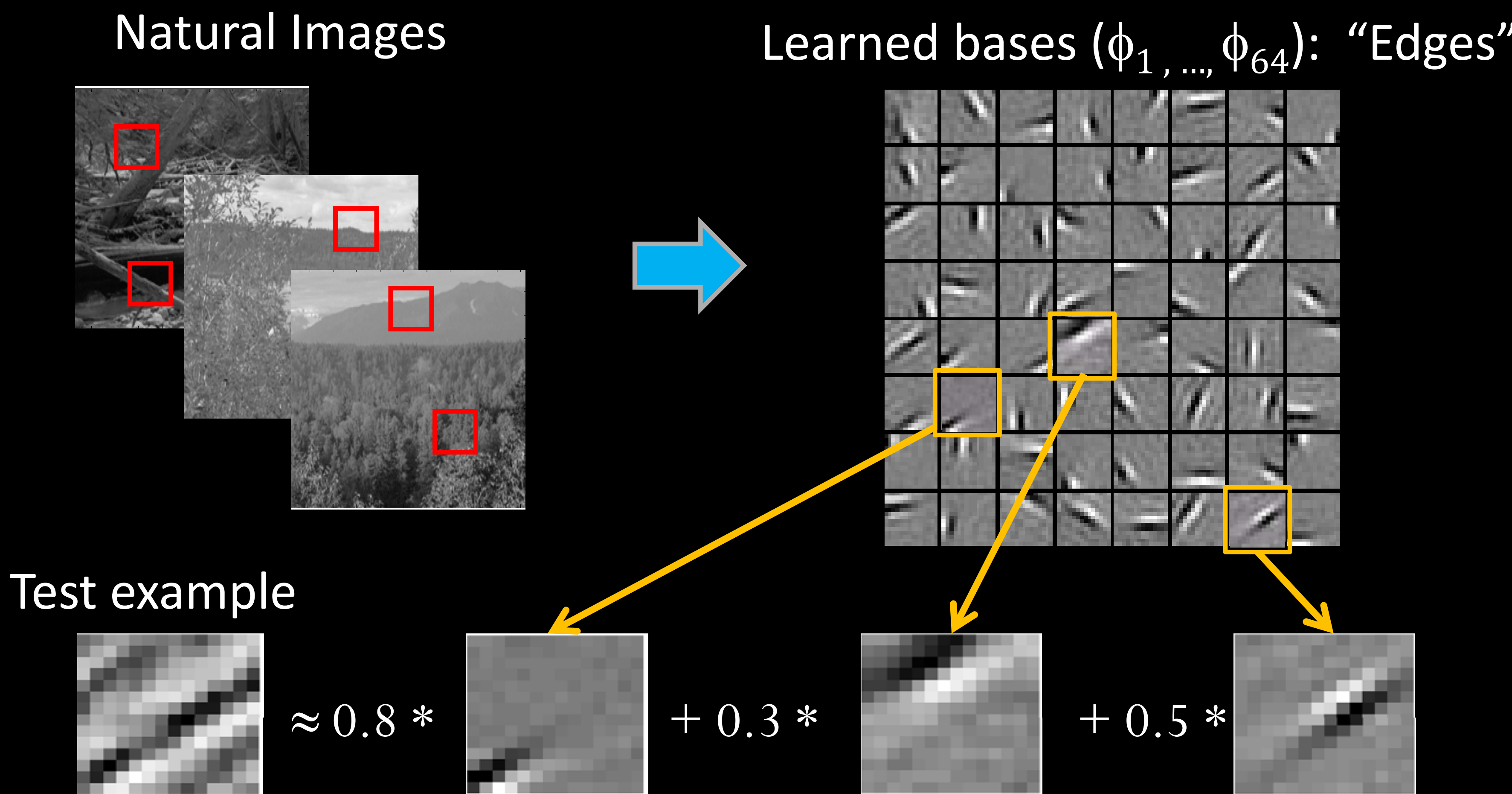
Input: Images $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ (each in $R^{n \times n}$)

Learn: Dictionary of bases $\phi_1, \phi_2, \ldots, \phi_k$ (also $R^{n \times n}$), so that each input x can be approximately decomposed as:

$$x \approx \sum_{j=1}^{k} a_j \phi_j$$

s.t. $a_j$'s are mostly zero ("sparse")

# Sparse coding illustration



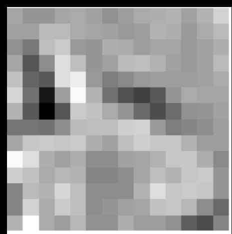
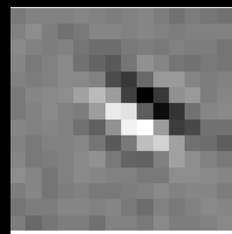
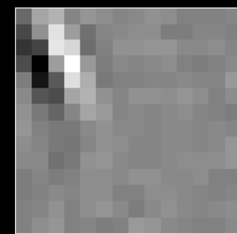
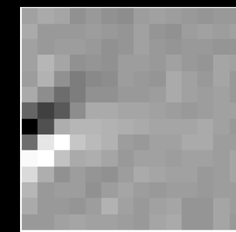

$$x \approx 0.8 * \phi_{36} + 0.3 * \phi_{42} + 0.5 * \phi_{63}$$

$[a_1, \ldots, a_{64}]$ = [0, 0, ..., 0, **0.8**, 0, ..., 0, **0.3**, 0, ..., 0, **0.5**, 0]

(feature representation)

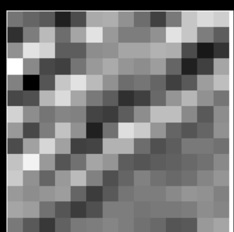
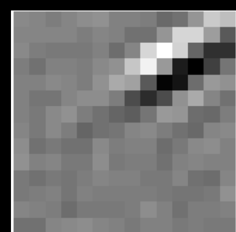
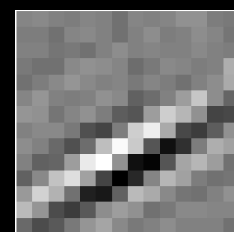
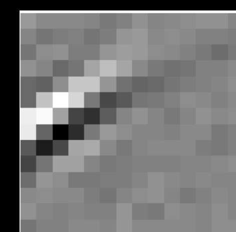
More succinct, higher-level, representation.

# More examples

 ≈ 0.6 *  + 0.8 *  + 0.4 * 

$\phi_{15}$ $\phi_{28}$ $\phi_{37}$

**Represent as: [$a_{15}$=0.6, $a_{28}$=0.8, $a_{37}$ = 0.4].**

 ≈ 1.3 *  + 0.9 *  + 0.3 * 

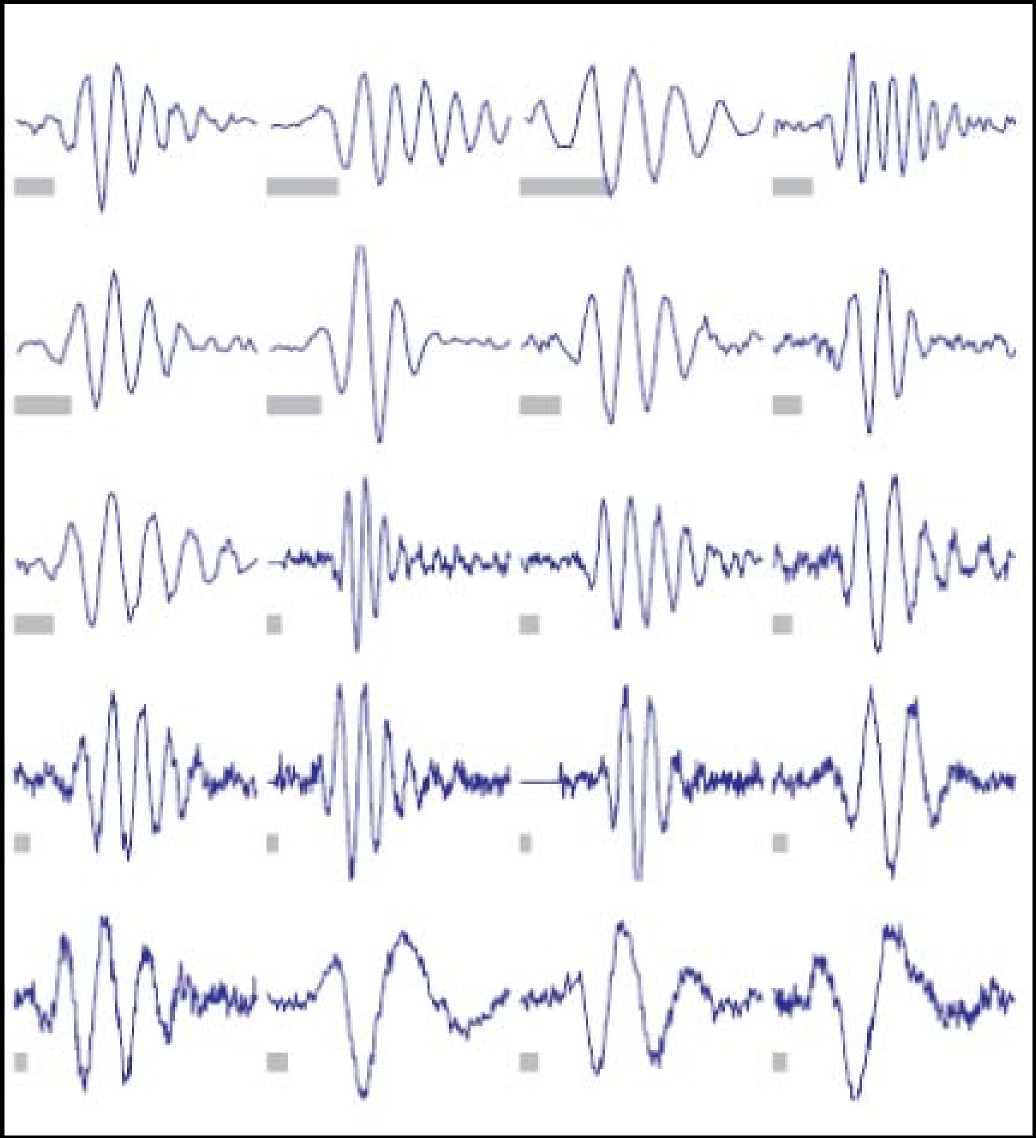
$\phi_{5}$ $\phi_{18}$ $\phi_{29}$

**Represent as: [$a_{5}$=1.3, $a_{18}$=0.9, $a_{29}$ = 0.3].**

- Method "invents" edge detection.
- ]Gives a more succinct, higher-level representation than the raw pixels.
- Quantitatively similar to primary visual cortex (area V1) in brain.

# Sparse coding applied to audio

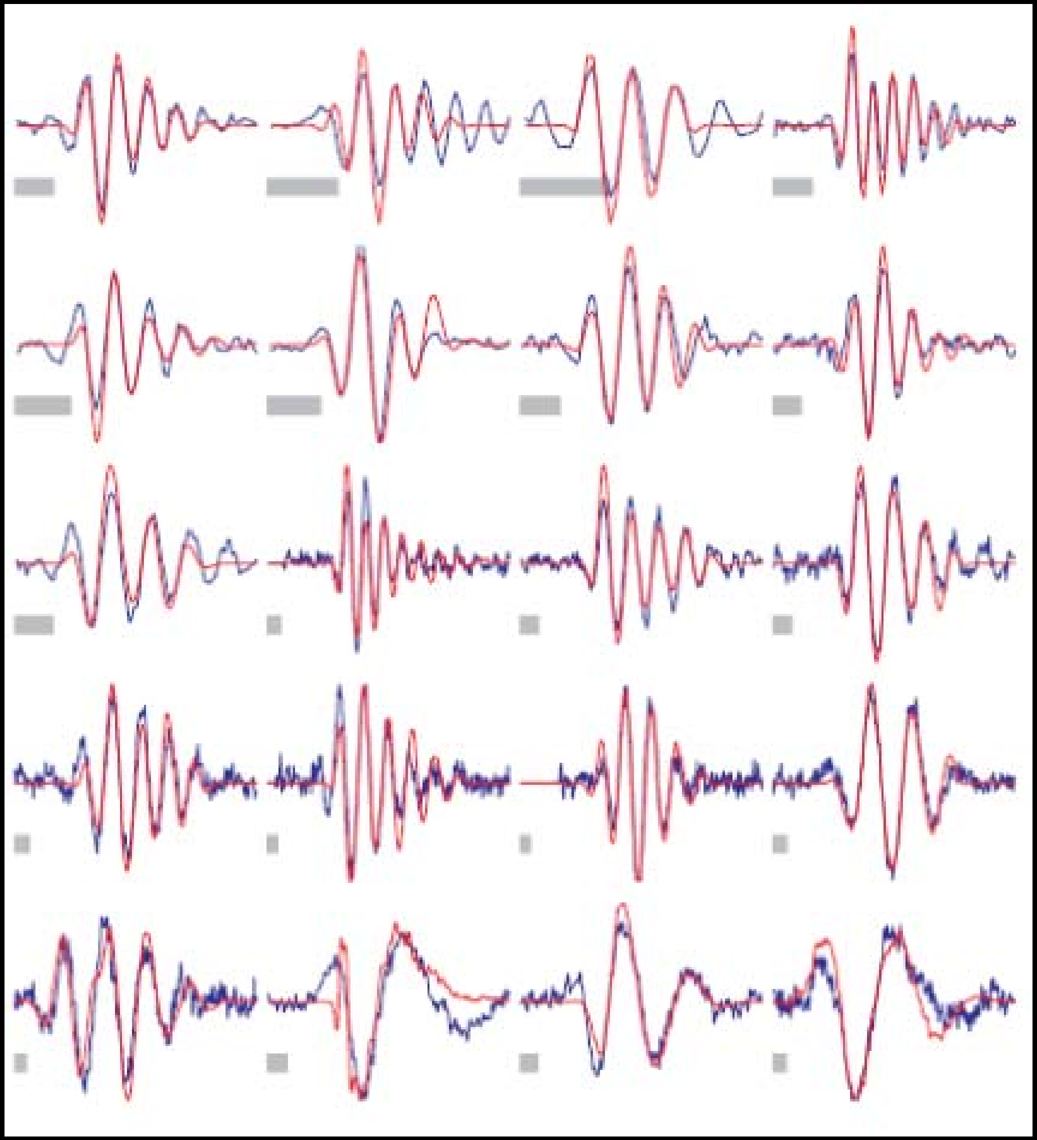
Image shows 20 basis functions learned from unlabeled audio.



[Evan Smith & Mike Lewicki, 2006]

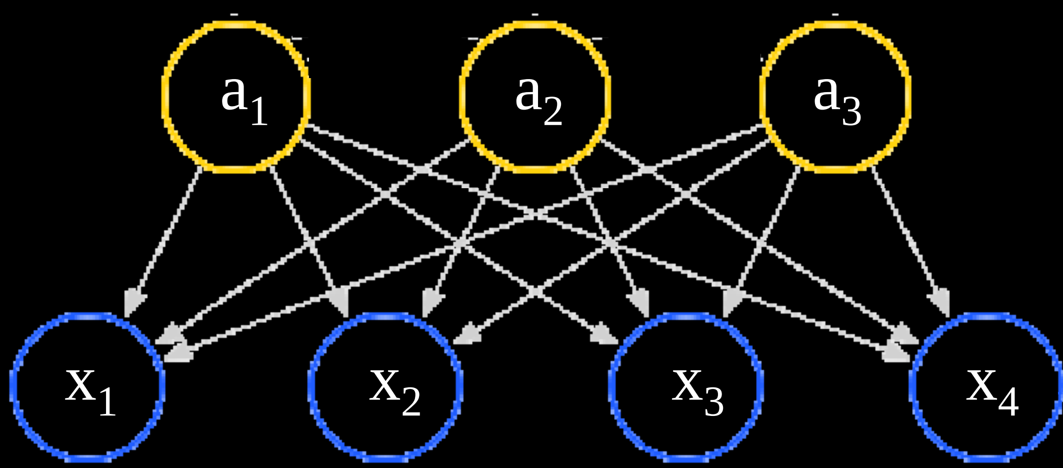
# Sparse coding applied to audio

Image shows 20 basis functions learned from unlabeled audio.



[Evan Smith & Mike Lewicki, 2006]

# Learning feature hierarchies




Higher layer
(Combinations of edges;
cf V2)

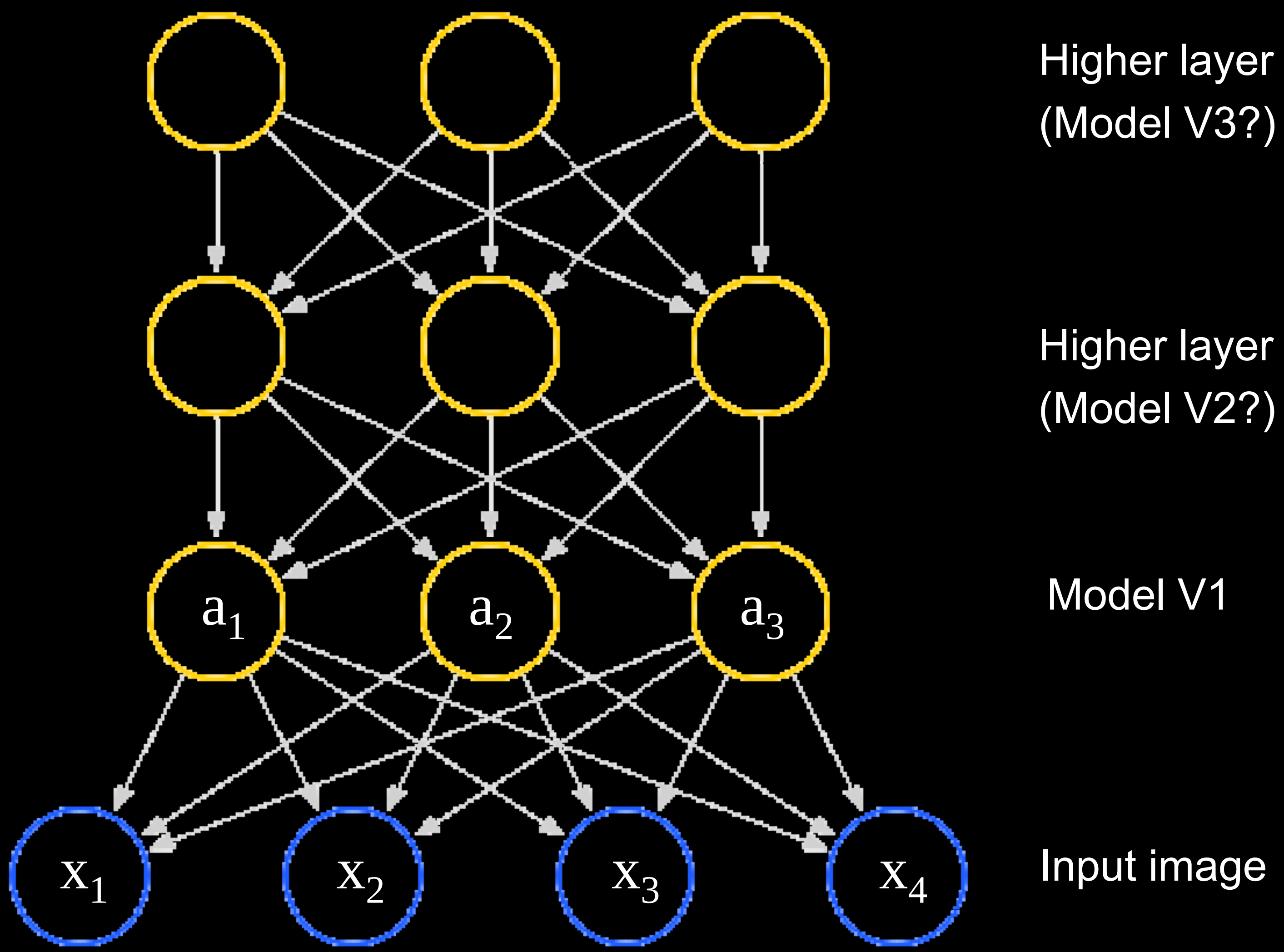
"Sparse coding"
(edges; cf. V1)

Input image (pixels)

[Technical details: Sparse autoencoder or sparse version of Hinton's DBN.]

[Lee, Ranganath & Ng, 2007]

# Learning feature hierarchies

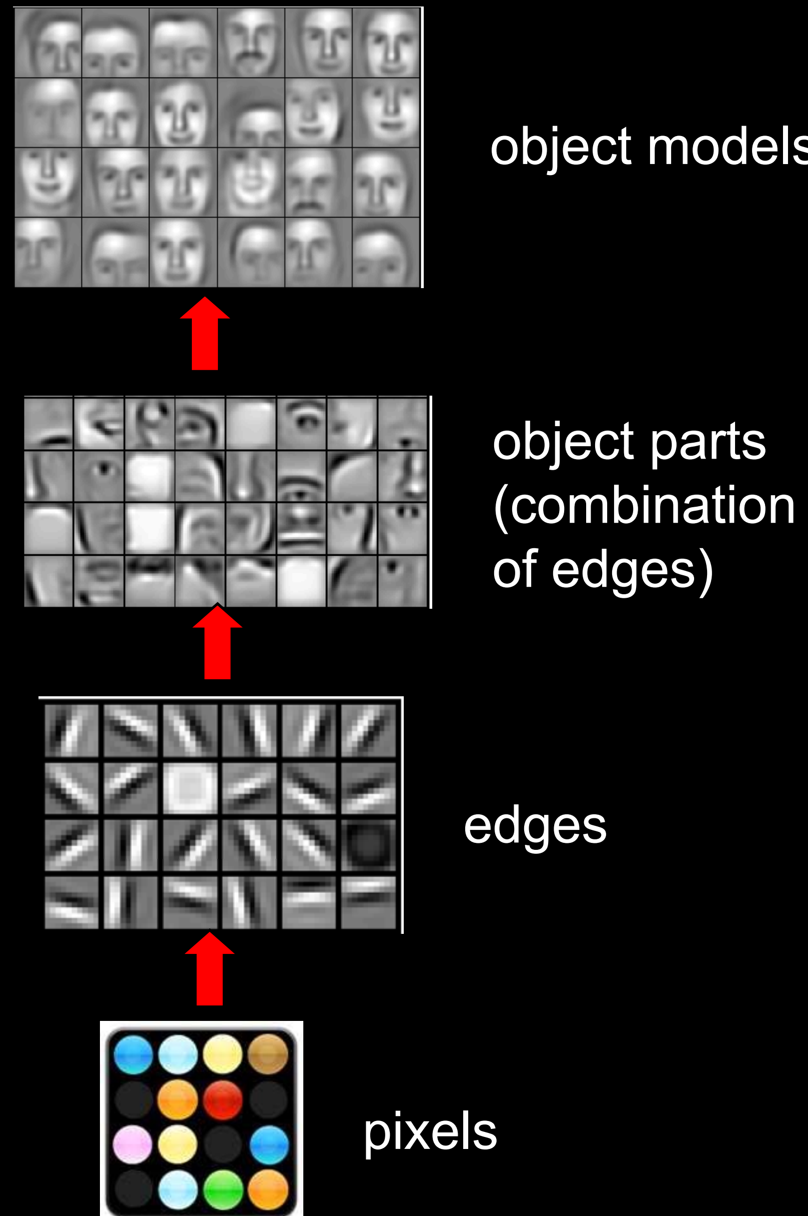



[Technical details: Sparse autoencoder or sparse version of Hinton's DBN.]

[Lee, Ranganath & Ng, 2007]

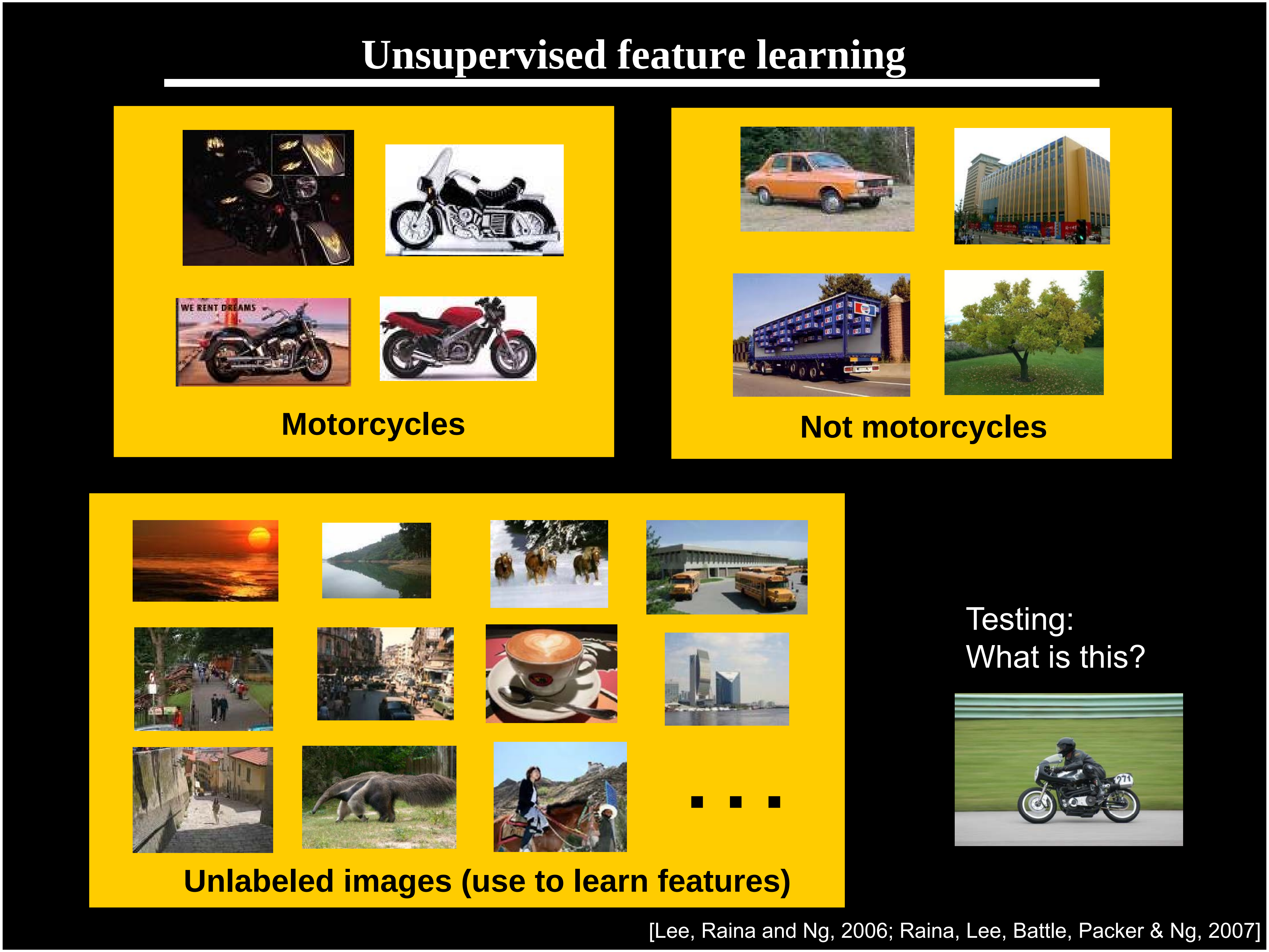
# Hierarchical Sparse coding (Sparse DBN): Trained on face images

Training set: Aligned images of faces.




[Honglak Lee]

# Machine learning applications


Unsupervised feature learning
WE RENT DREAMS
Motorcycles
Not motorcycles
Unlabeled images (use to learn features)
Testing:
What is this?
[Lee, Raina and Ng, 2006; Raina, Lee, Battle, Packer & Ng, 2007]

# Video Activity recognition (Hollywood 2 benchmark)

| Method | Accuracy |
|---|---|
| Hessian + ESURF [Williems et al 2008] | 38% |
| Harris3D + HOG/HOF [Laptev et al 2003, 2004] | 45% |
| Cuboids + HOG/HOF [Dollar et al 2005, Laptev 2004] | 46% |
| Hessian + HOG/HOF [Laptev 2004, Williems et al 2008] | 46% |
| Dense + HOG / HOF [Laptev 2004] | 47% |
| Cuboids + HOG3D [Klaser 2008, Dollar et al 2005] | 46% |
| **Unsupervised feature learning (our method)** | **52%** |

Unsupervised feature learning significantly improves on the previous state-of-the-art.

[Le, Zhou & Ng, 2011]

## Audio

| TIMIT Phone classification | Accuracy |
|---|---|
| Prior art (Clarkson et al.,1999) | 79.6% |
| Stanford Feature learning | **80.3%** |

| TIMIT Speaker identification | Accuracy |
|---|---|
| Prior art (Reynolds, 1995) | 99.7% |
| Stanford Feature learning | **100.0%** |

## Images

| CIFAR Object classification | Accuracy |
|---|---|
| Prior art (Ciresan et al., 2011) | 80.5% |
| Stanford Feature learning | **82.0%** |

| NORB Object classification | Accuracy |
|---|---|
| Prior art (Scherer et al., 2010) | 94.4% |
| Stanford Feature learning | **95.0%** |

## Video

| Hollywood2 Classification | Accuracy |
|---|---|
| Prior art (Laptev et al., 2004) | 48% |
| Stanford Feature learning | **53%** |

| YouTube | Accuracy |
|---|---|
| Prior art (Liu et al., 2009) | 71.2% |
| Stanford Feature learning | **75.8%** |

| KTH | Accuracy |
|---|---|
| Prior art (Wang et al., 2010) | 92.1% |
| Stanford Feature learning | **93.9%** |

| UCF | Accuracy |
|---|---|
| Prior art (Wang et al., 2010) | 85.6% |
| Stanford Feature learning | **86.5%** |

## Text/NLP

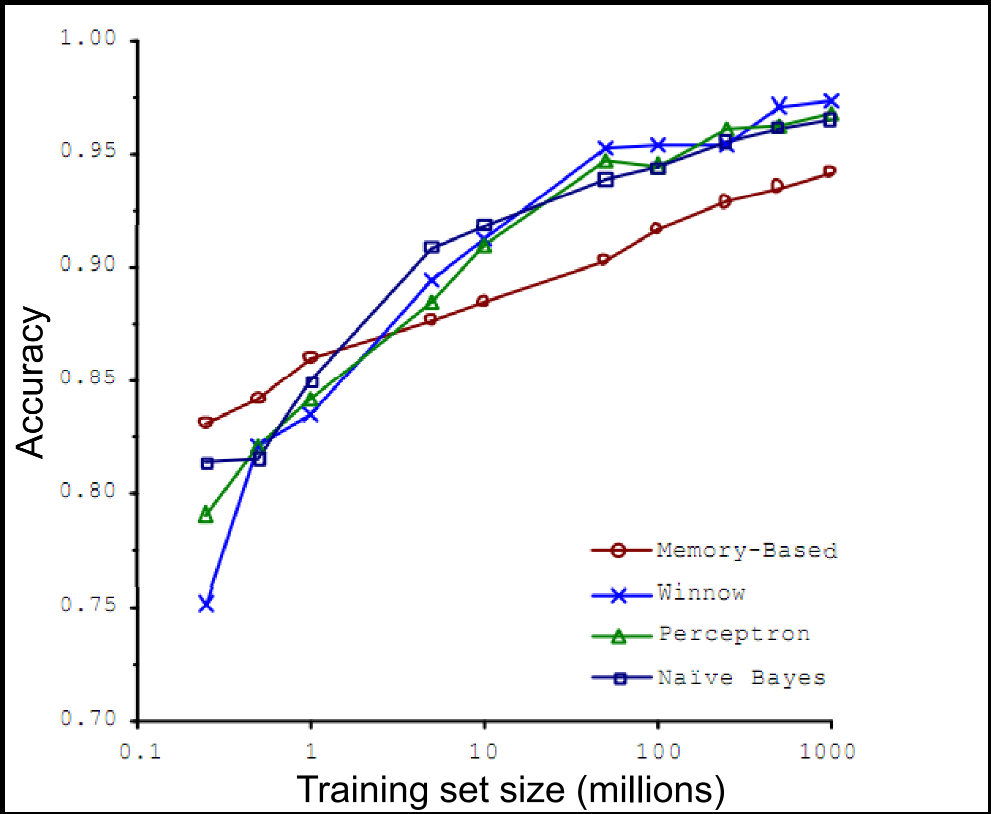
| Paraphrase detection | Accuracy |
|---|---|
| Prior art (Das & Smith, 2009) | 76.1% |
| Stanford Feature learning | **76.4%** |

| Sentiment (MR/MPQA data) | Accuracy |
|---|---|
| Prior art (Nakagawa et al., 2010) | 77.3% |
| Stanford Feature learning | **77.7%** |

How do you build a high accuracy learning system?

# Supervised Learning: Labeled data

- Choices of learning algorithm:
  - Memory based
  - Winnow
  - Perceptron
  - Naïve Bayes
  - SVM
  - ….
- What matters the most?

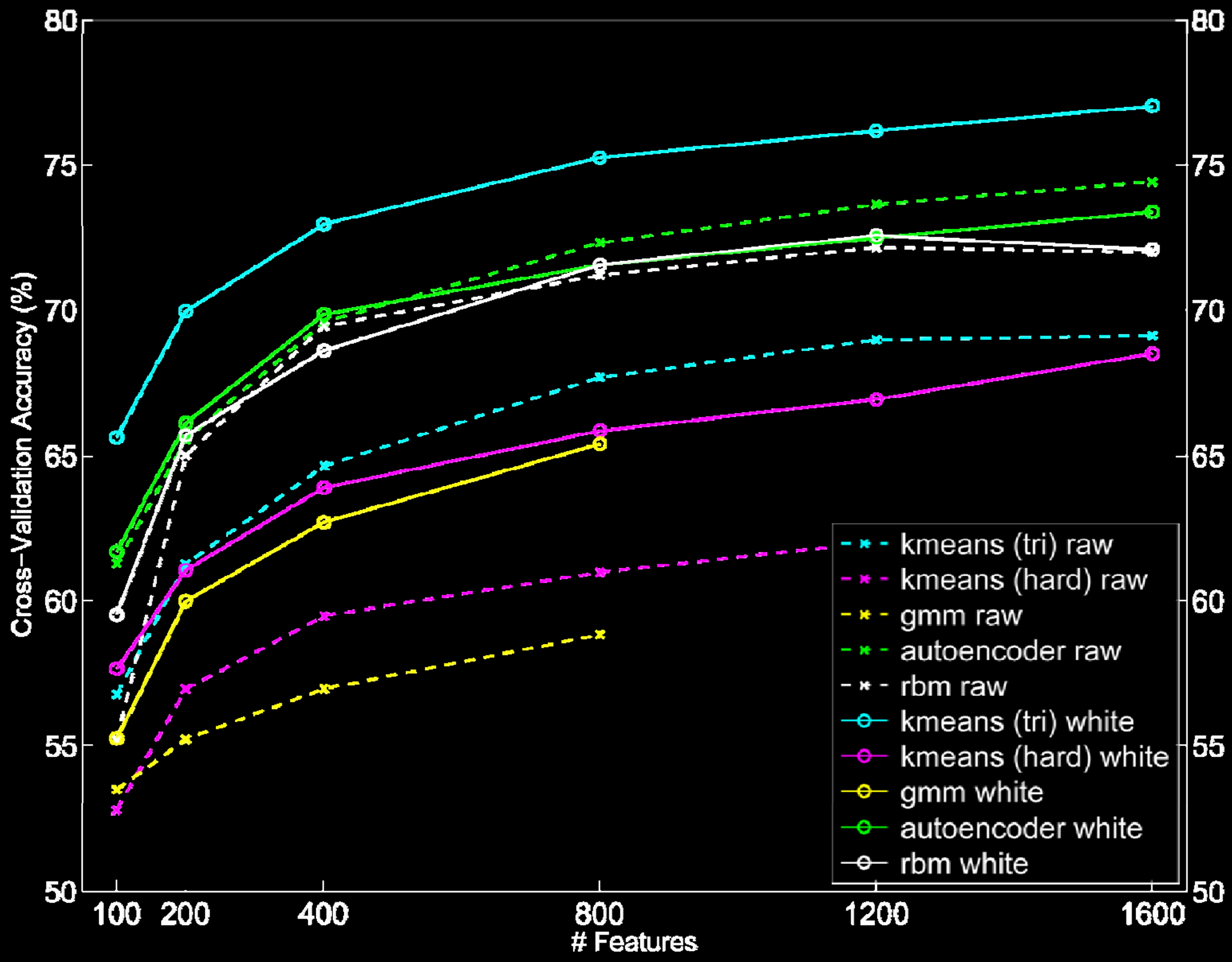



[Banko & Brill, 2001]

"It's not who has the best algorithm that wins. It's who has the most data."

# Unsupervised Learning

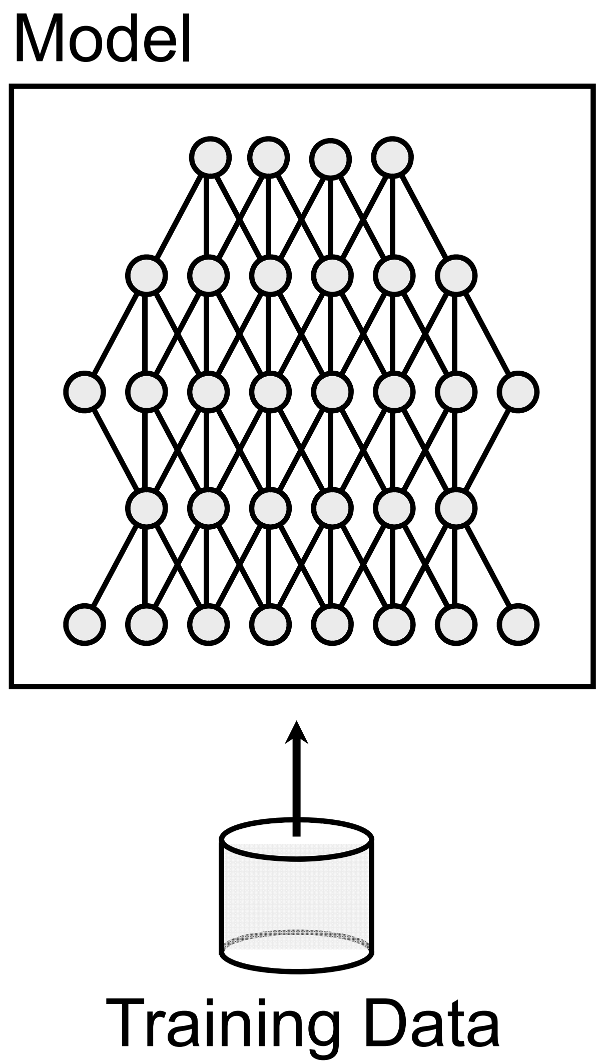
Large numbers of features is critical. The specific learning algorithm is important, but ones that can scale to many features also have a big advantage.




[Adam Coates]

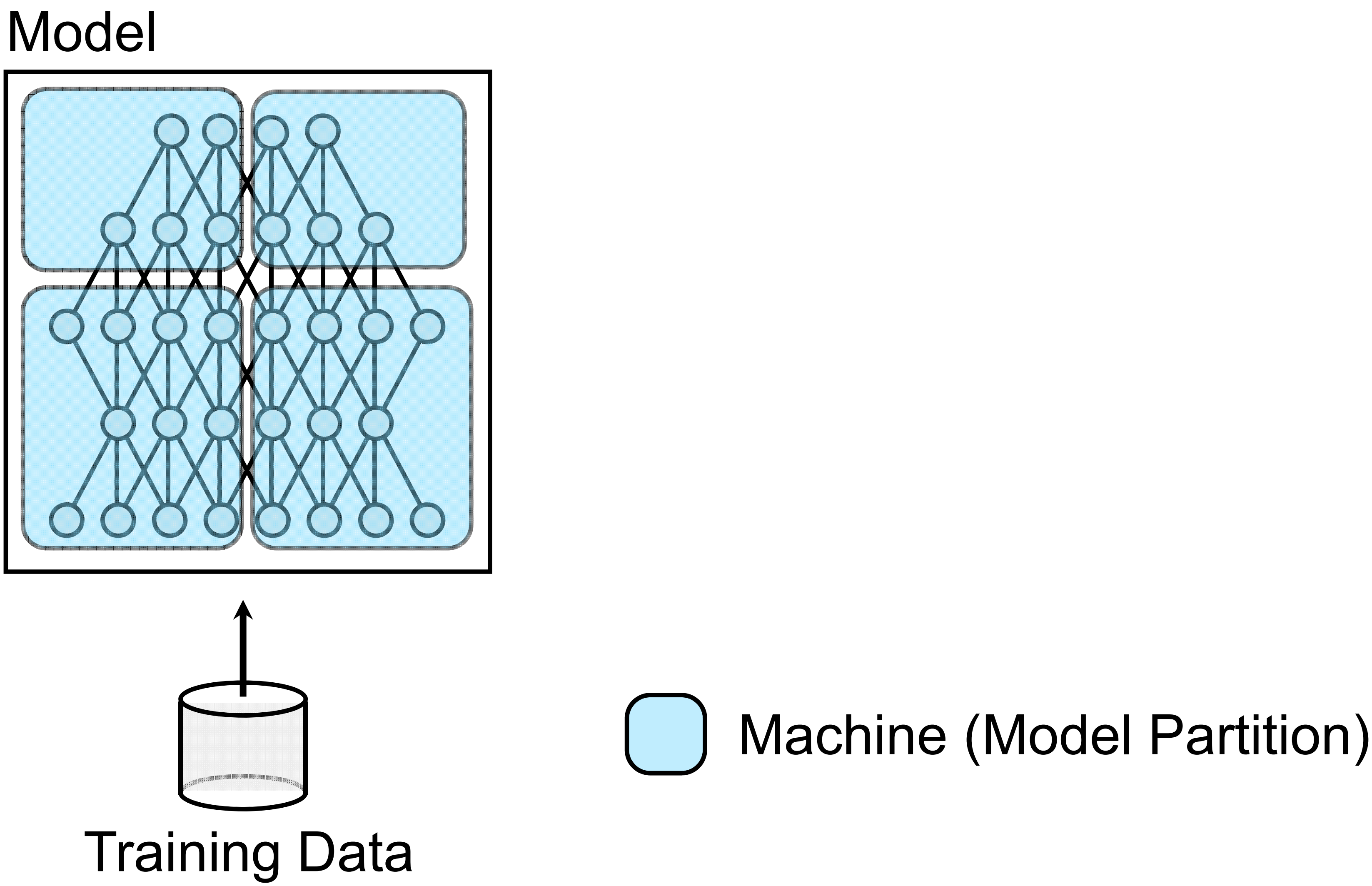
# Learning from Labeled data

# Google-scale Parallel learning



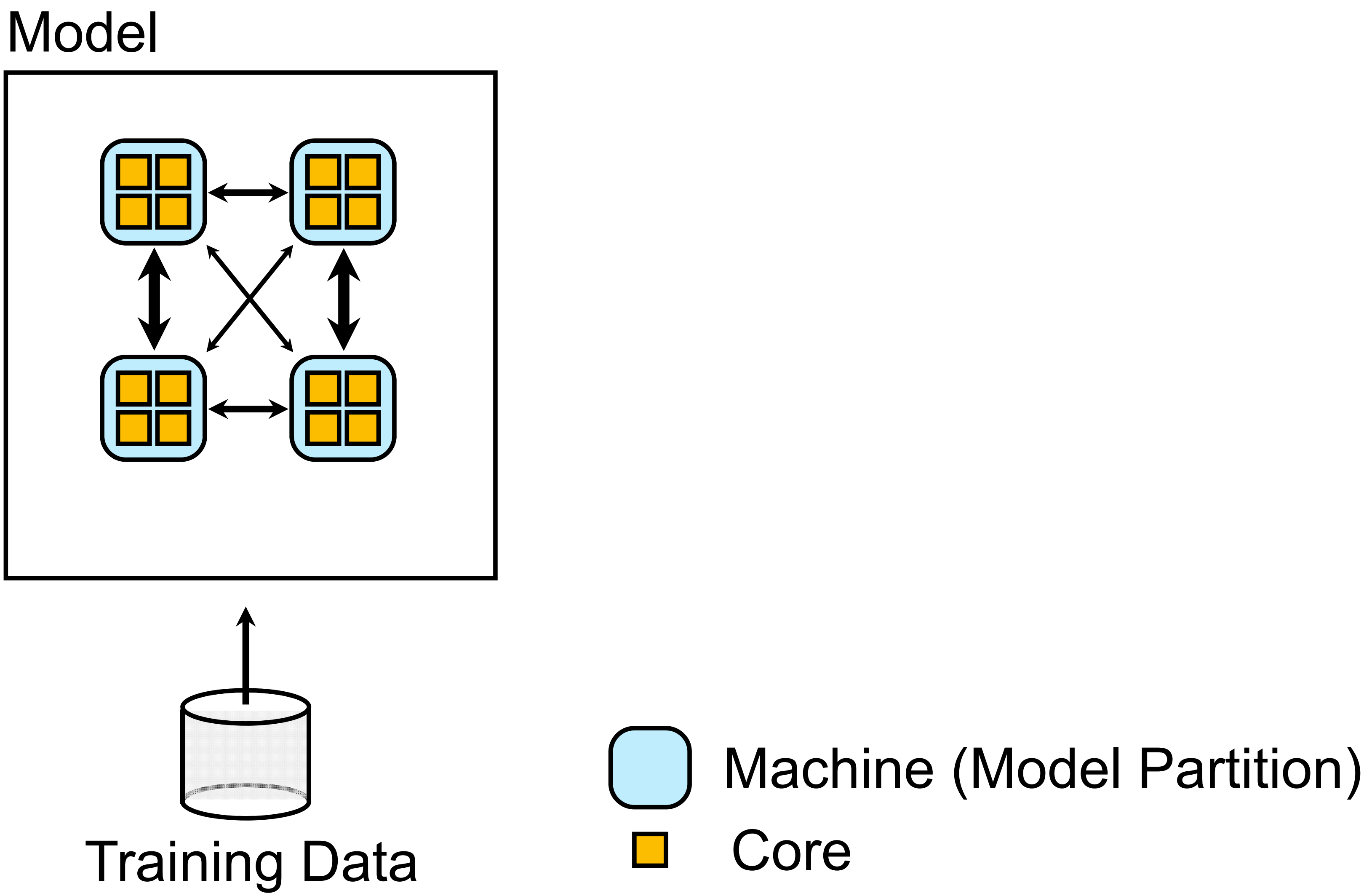

[Adam Coates]

# Google-scale Parallel learning



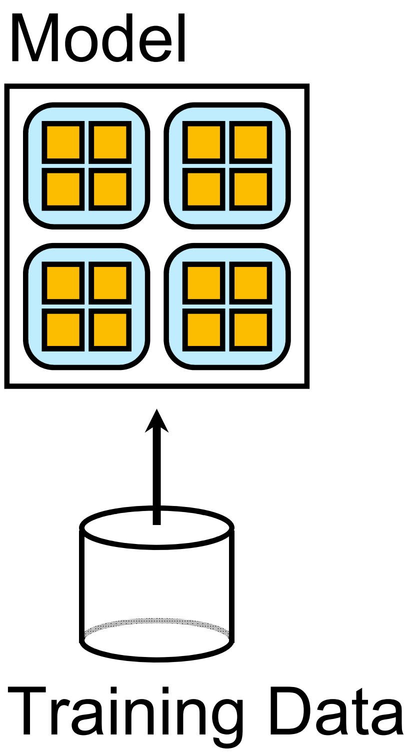

[Adam Coates]

# Google-scale Parallel learning




[Adam Coates]

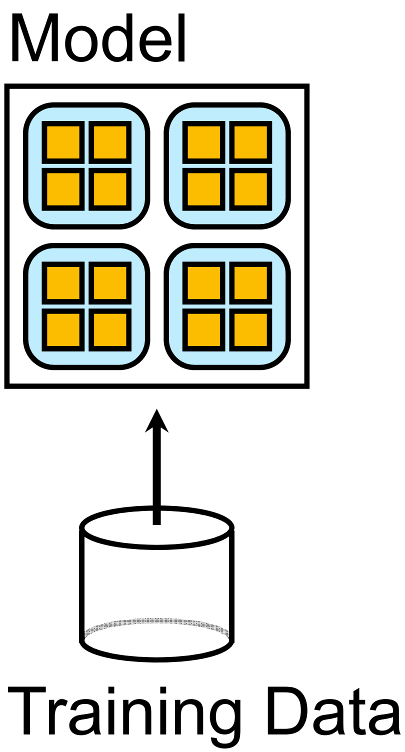
# Google-scale Parallel learning




- Unsupervised or Supervised Objective
- Minibatch Stochastic Gradient Descent (SGD)
- Model parameters sharded by partition
- 10s, 100s, or 1000s of cores per model

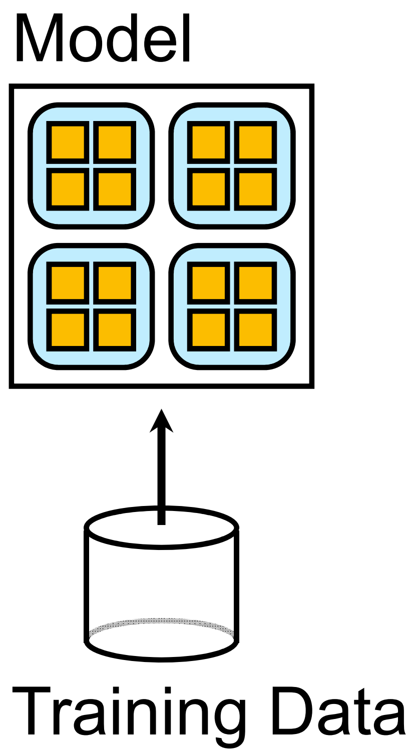
[Adam Coates]

# Basic DistBelief Model Training




- Unsupervised or Supervised Objective
- Minibatch Stochastic Gradient Descent (SGD)
- Model parameters sharded by partition
- 10s, 100s, or 1000s of cores per model

# Basic DistBelief Model Training




Parallelize across ~100 machines (~1600 cores).

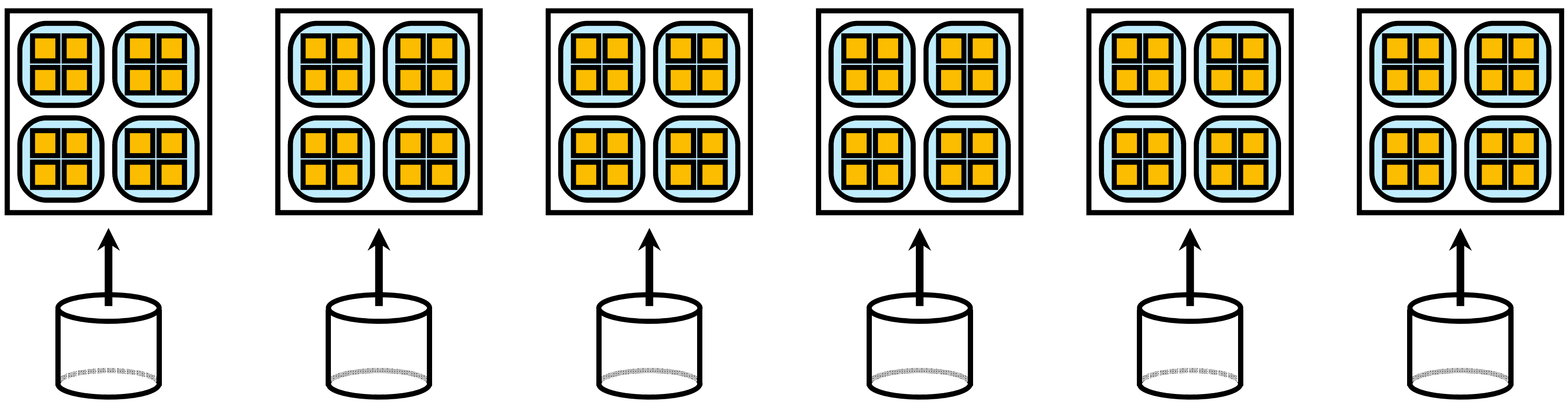
But training is still slow with large data sets.

Add another dimension of parallelism, and have multiple model instances in parallel.

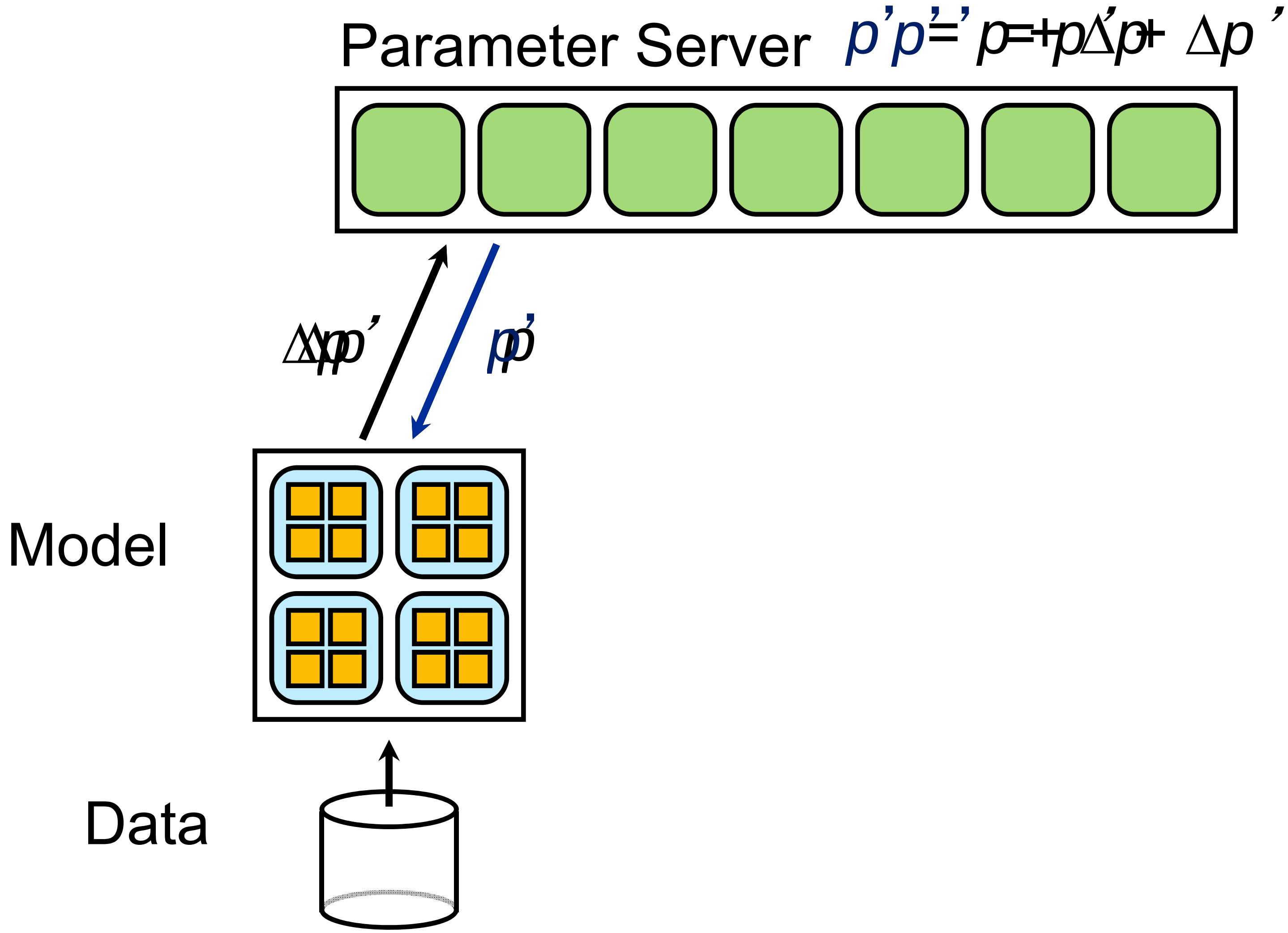
# Two Approaches to Multi-Model Training

(1) Downpour: Asynchronous Distributed SGD

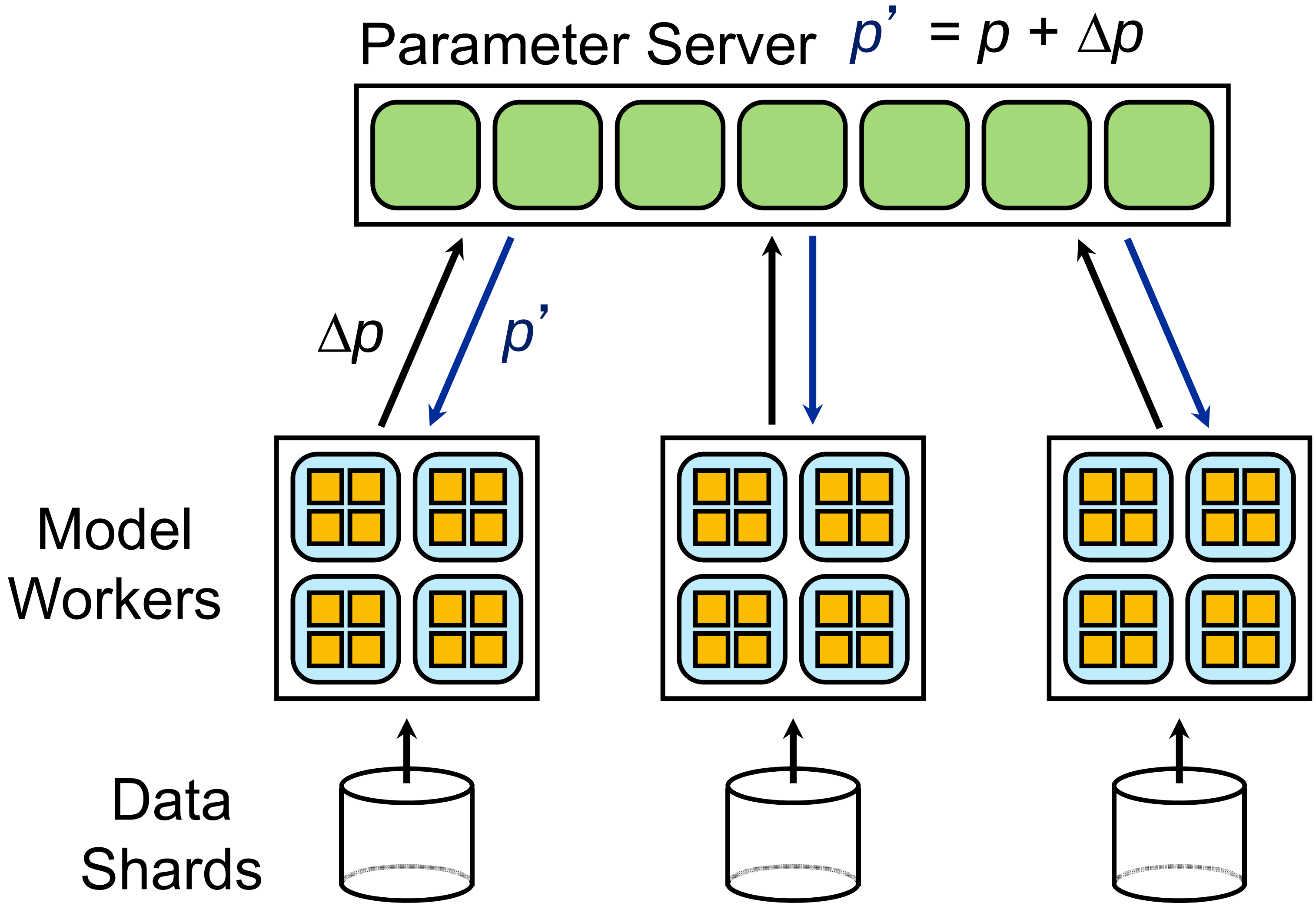
(2) Sandblaster: Distributed L-BFGS



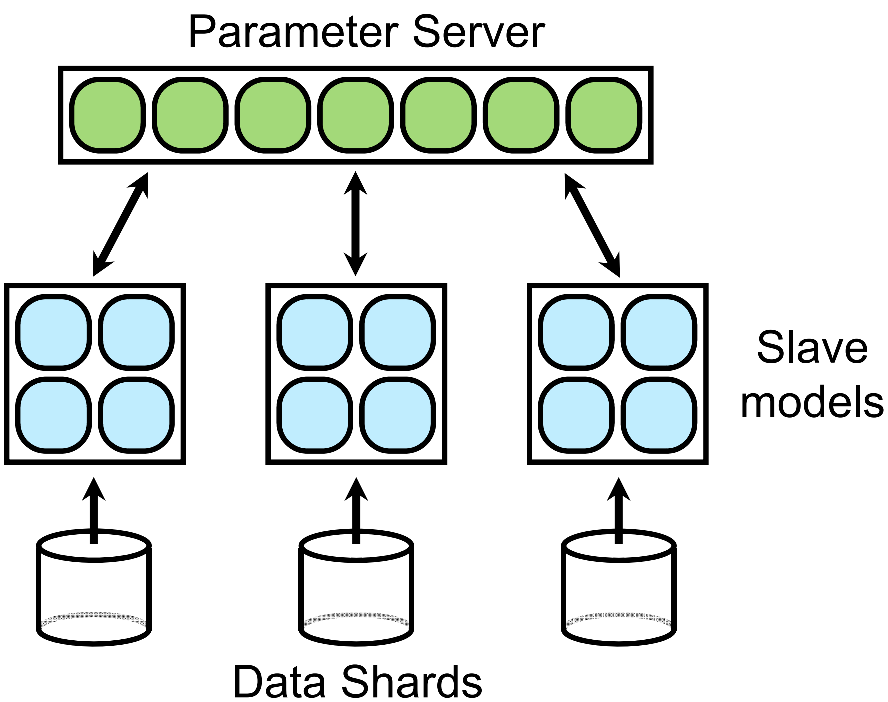
# Asynchronous Distributed Stochastic Gradient Descent



# Asynchronous Distributed Stochastic Gradient Descent



# Asynchronous Distributed Stochastic Gradient Descent




From an engineering standpoint, superior to a single model with the same number of total machines:

- Better robustness to individual slow machines
- Makes forward progress even during evictions/restarts

# L-BFGS: a Big Batch Alternative to SGD.

Async-SGD

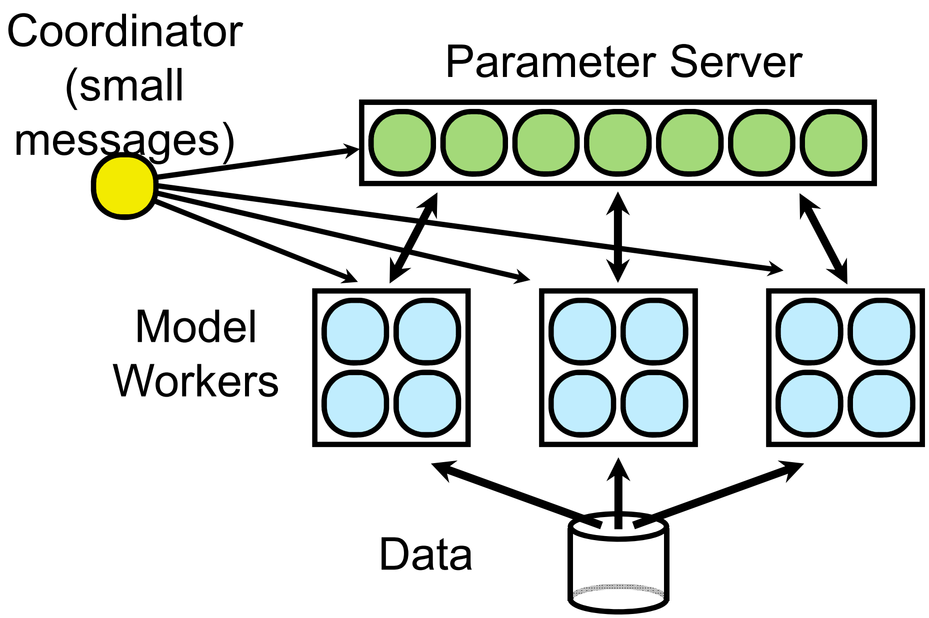
- first derivatives only
- many small steps
- mini-batched data
  (10s of examples)
- tiny compute and data requirements per step
- theory is dicey
- at most 10s or 100s of model replicas

L-BFGS

- first and second derivatives
- larger, smarter steps
- **mega**-batched data
  (millions of examples)
- huge compute and data requirements per step
- strong theoretical grounding
- 1000s of model replicas

# L-BFGS: a Big Batch Alternative to SGD.

Leverages the same parameter server implementation as Async-SGD, but uses it to shard computation within a mega-batch.




Some current numbers:

- 20,000 cores in a single cluster
- up to 1 billion data items / mega-batch (in ~1 hour)

More network friendly at large scales than Async-SGD.

The possibility of running on multiple data centers...

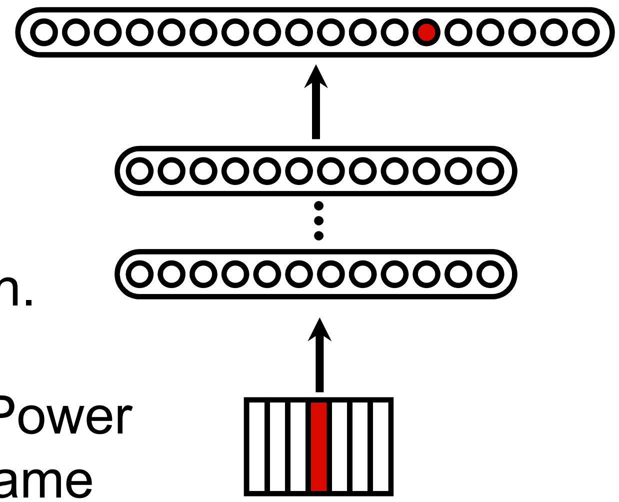
# Acoustic Modeling for Speech Recognition

8000-label Softmax

One or more hidden layers

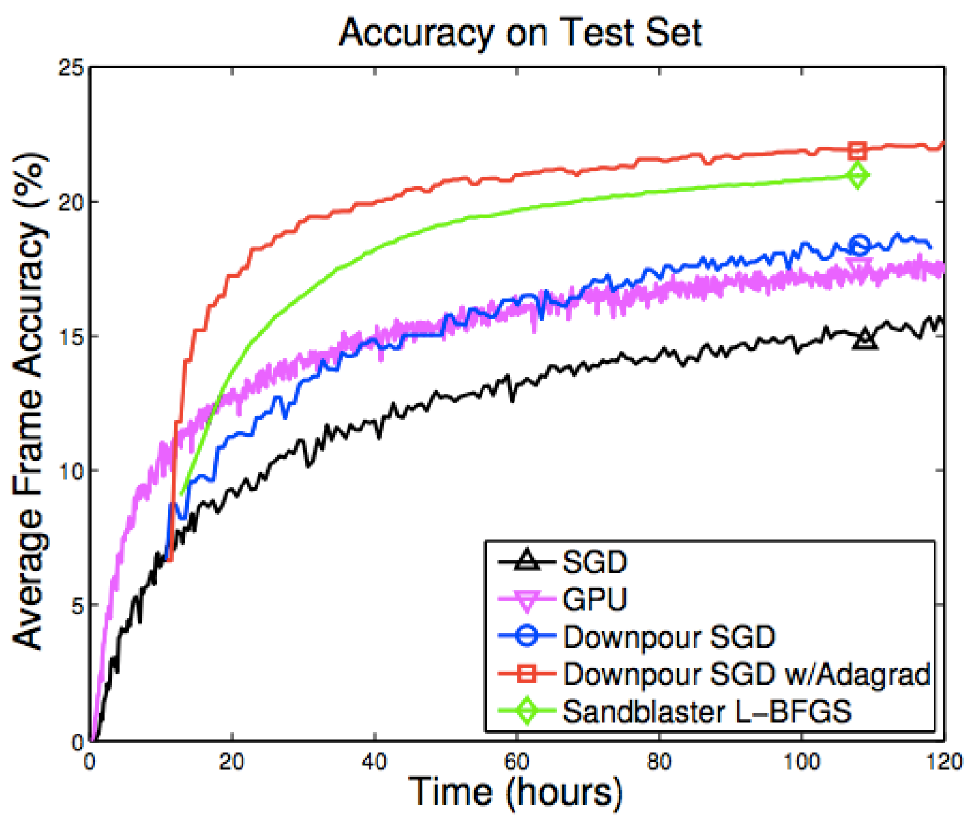
of a few thousand nodes each.

11 Frames of 40-value Log Energy Power
Spectra and the label for central frame



# Acoustic Modeling for Speech Recognition

Async SGD and L-BFGS can both speed up model training.




To reach the same model quality DistBelief reached in 4 days took 55 days using a GPU....

DistBelief can support much larger models than a GPU (useful for unsupervised learning).

# Speech recognition on Android

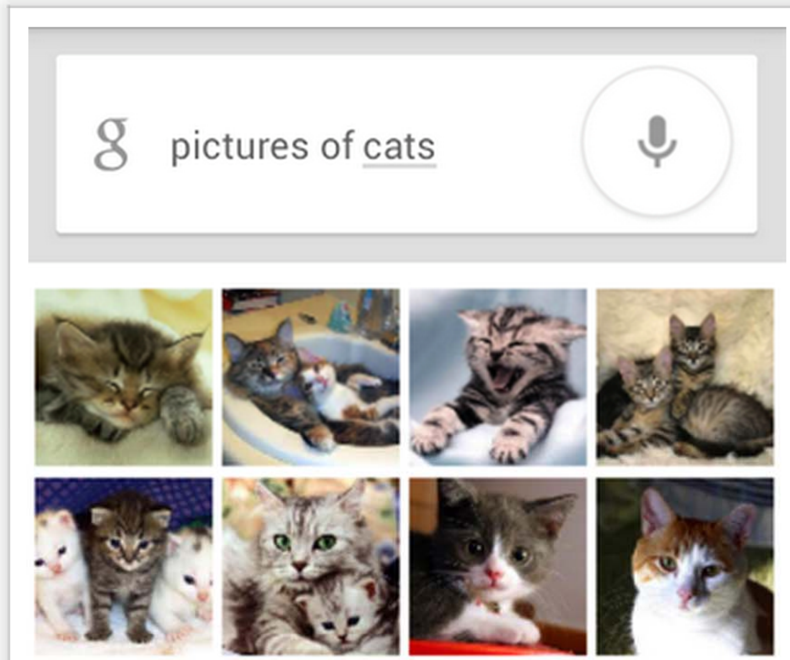
AUG
6

## Speech Recognition and Deep Learning

Posted by Vincent Vanhoucke, Research Scientist, Speech Team

The New York Times recently published an article about Google's large scale deep learning project, which learns to discover patterns in large datasets, including... cats on YouTube!

What's the point of building a gigantic cat detector you might ask? When you combine large amounts of data, large-scale distributed computing and powerful machine learning algorithms, you can apply the technology to address a large variety of practical problems.




With the launch of the latest Android platform release, Jelly Bean, we've taken a significant step towards making that technology useful: when you speak to your Android phone, chances are, you are talking to a neural network trained to recognize your speech.

Using neural networks for speech recognition is nothing new: the first proofs of concept were developed in the late

# Application to Google Streetview

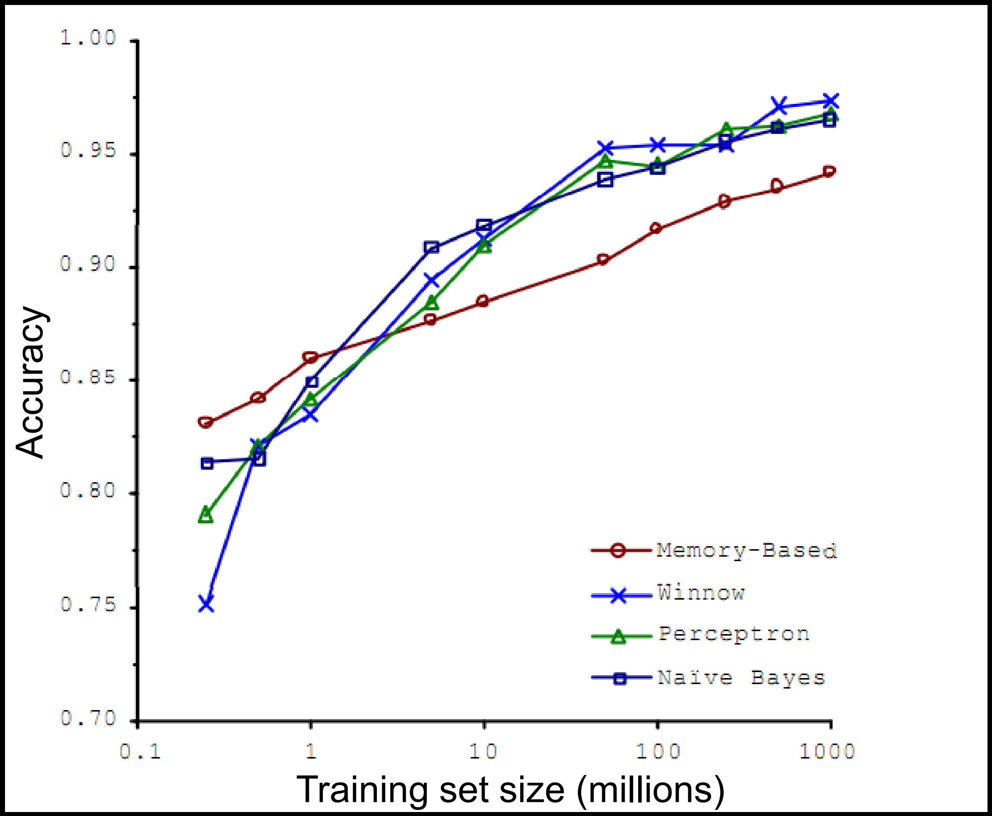
[with Yuval Netzer, Julian Ibarz]

# Learning from Unlabeled data

# Supervised Learning

- Choices of learning algorithm:
  - Memory based
  - Winnow
  - Perceptron
  - Naïve Bayes
  - SVM
  - ….

- What matters the most?

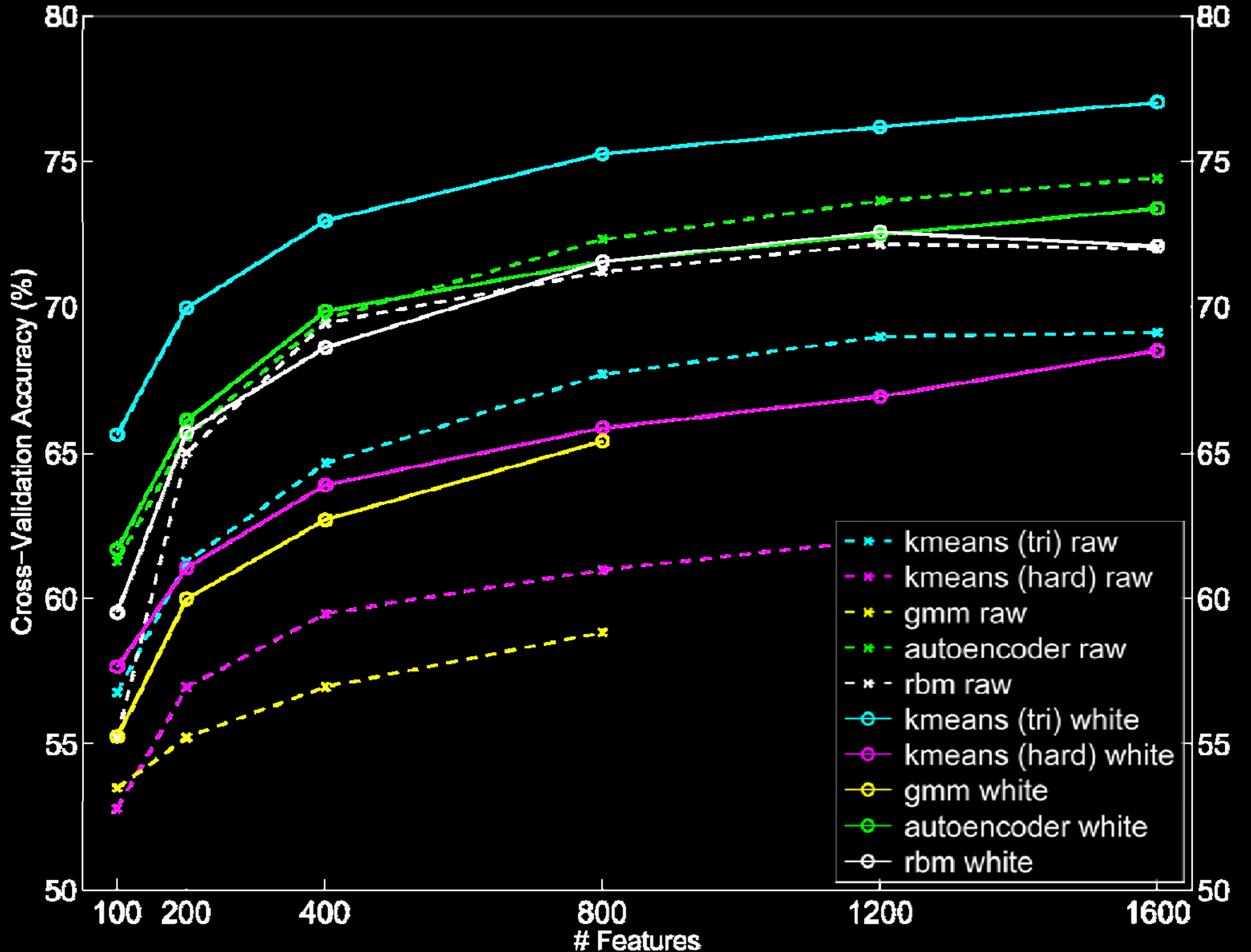



[Banko & Brill, 2001]

"It's not who has the best algorithm that wins.
It's who has the most data."

# Unsupervised Learning

Large numbers of features is critical. The specific learning algorithm is important, but ones that can scale to many features also have a big advantage.




[Adam Coates]

50 thousand 32x32 images

10 million parameters

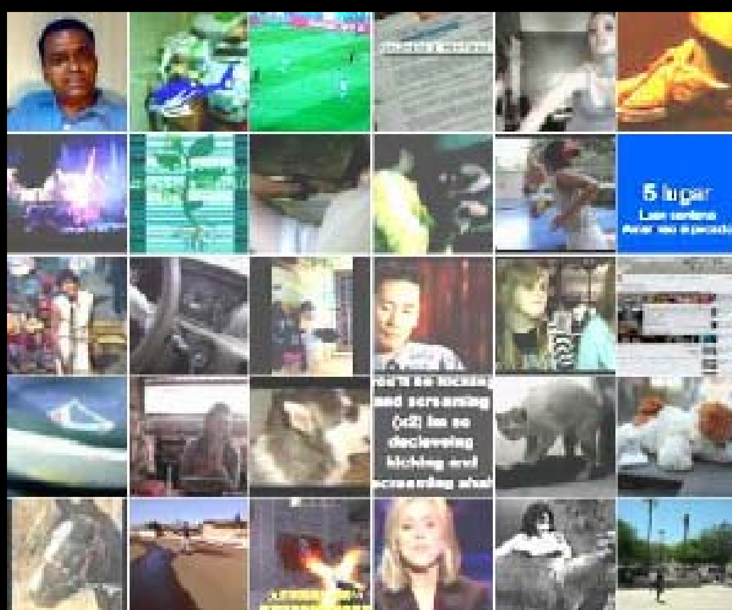
10 million 200x200 images

1 billion parameters

# Training procedure

What features can we learn if we train a massive model on a massive amount of data. Can we learn a "grandmother cell"?

- Train on 10 million images (YouTube)
- 1000 machines (16,000 cores) for 1 week.
- Test on novel images

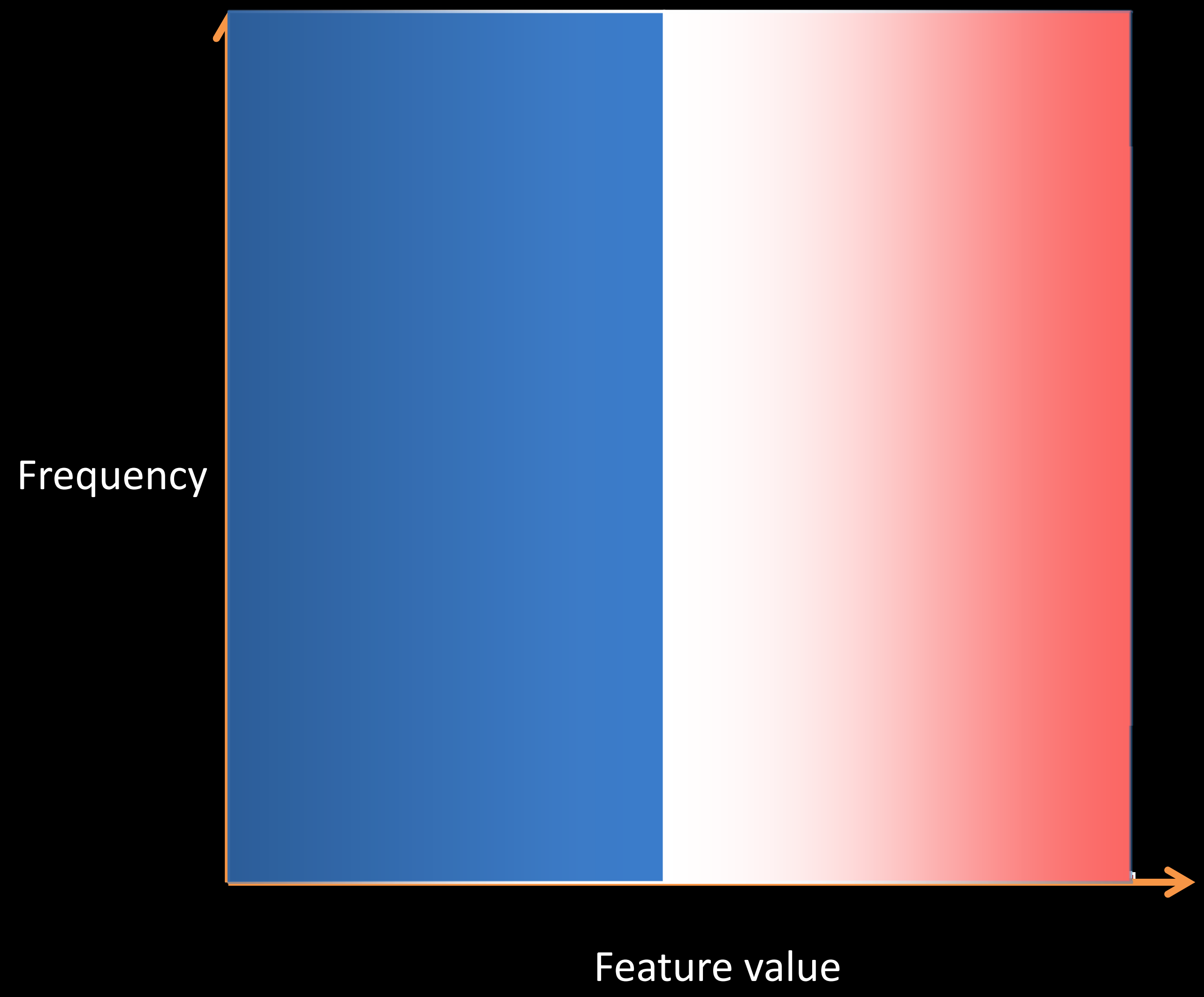


# The face neuron

Top stimuli from the test set

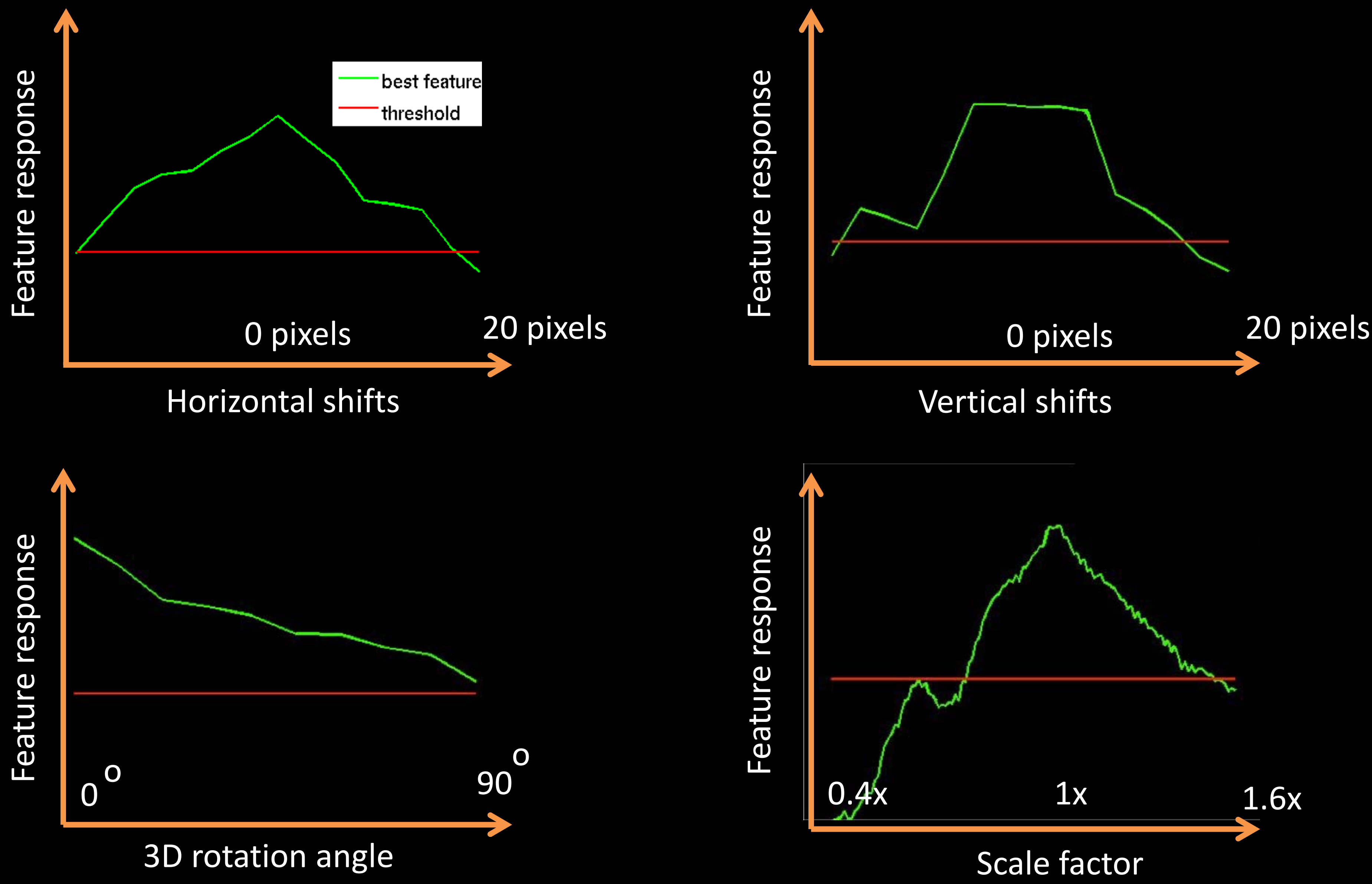
Optimal stimulus
by numerical optimization

Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012




Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

# Invariance properties



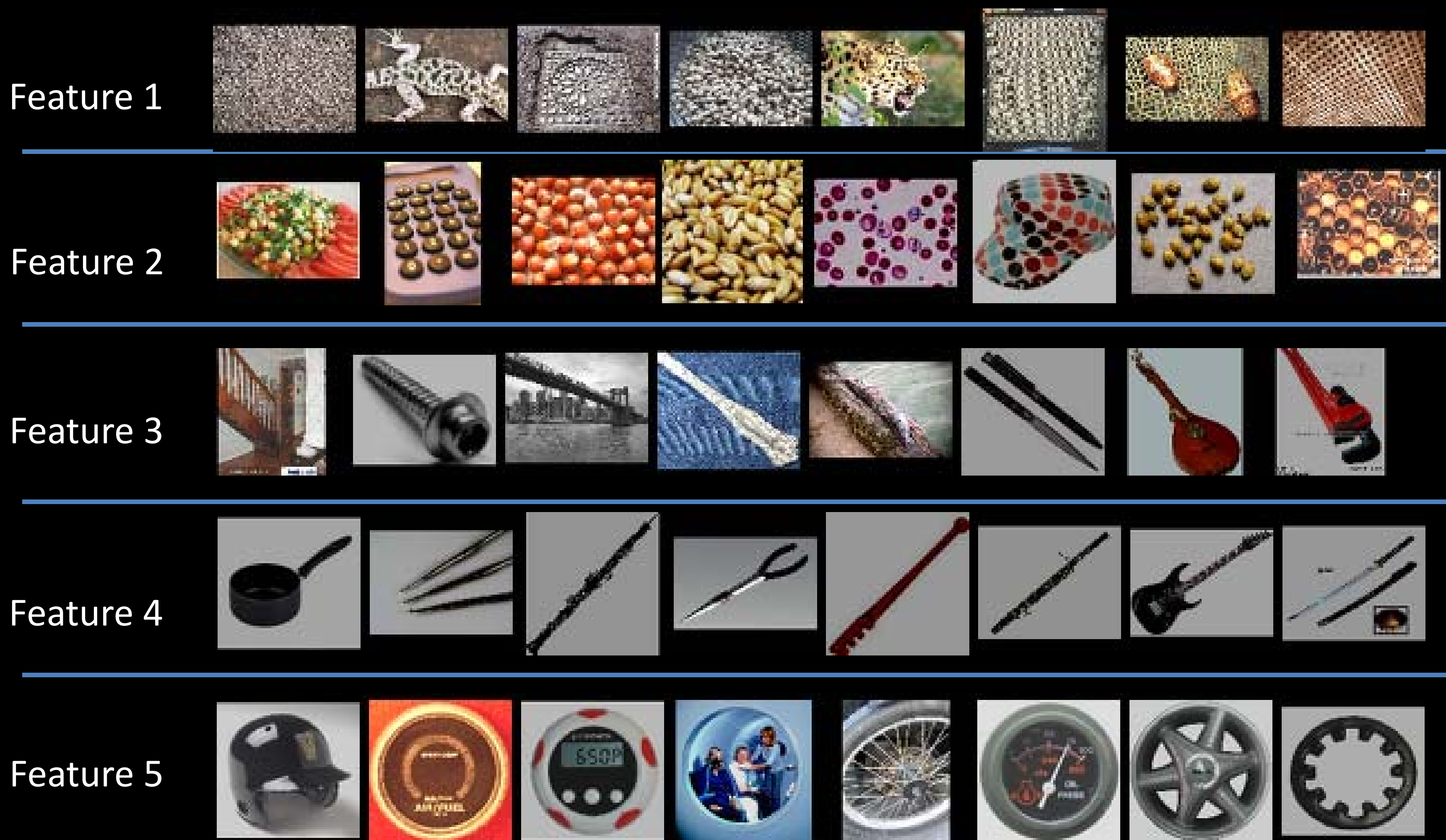

Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

# Cat neuron

Top Stimuli from the test set

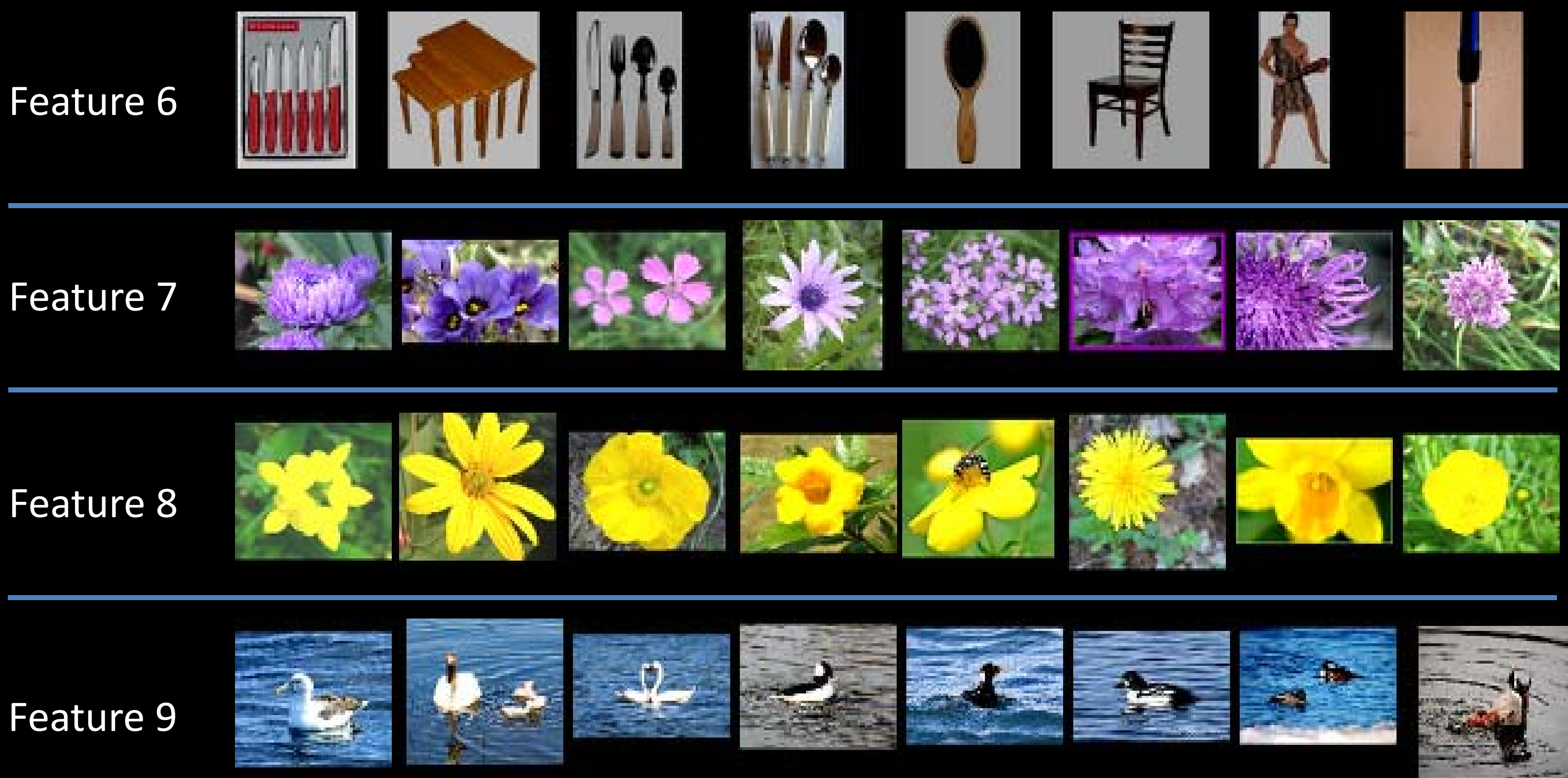
Average of top stimuli from test set

# Best stimuli



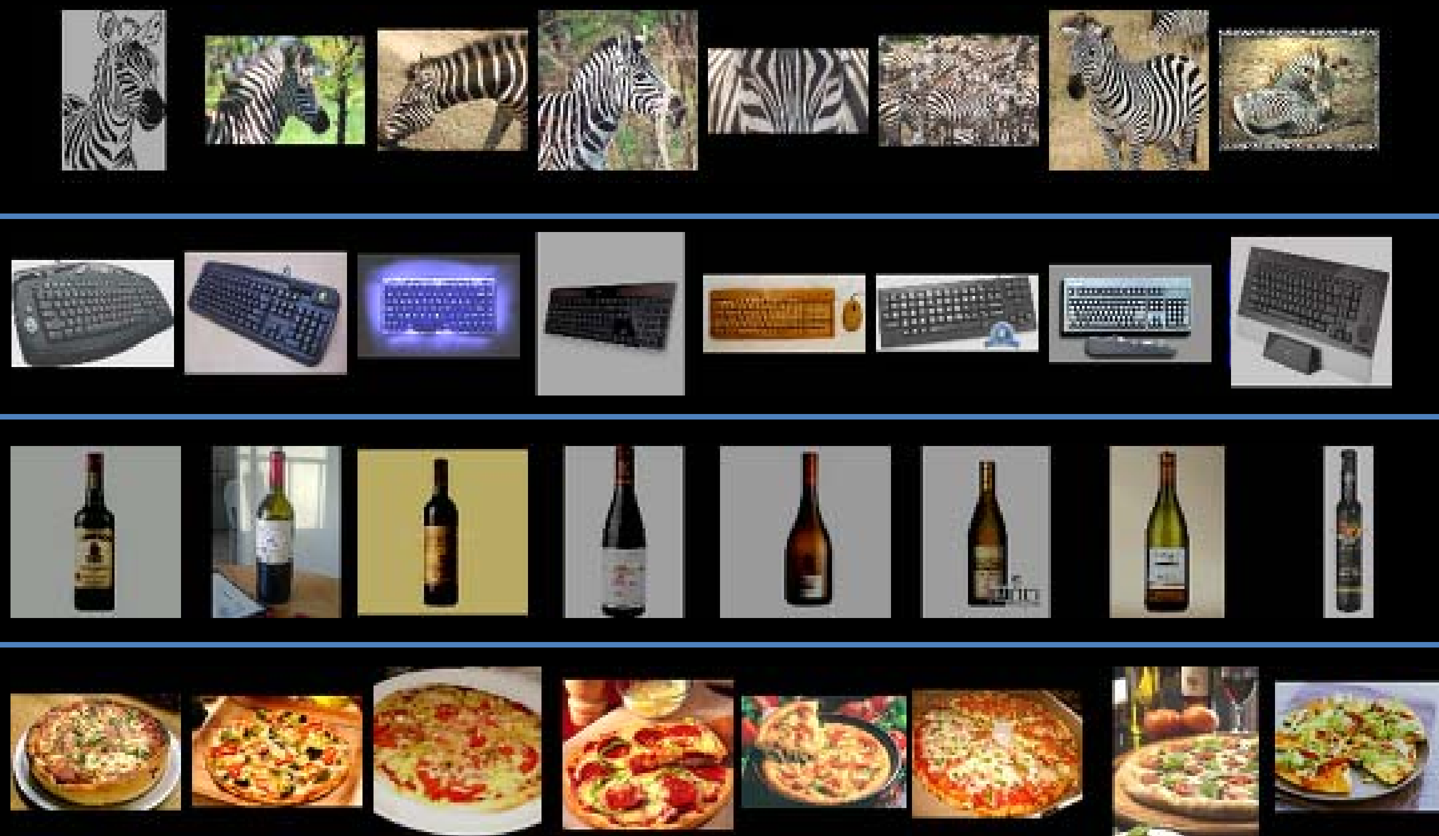

Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

# Best stimuli




Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

# Best stimuli



Feature 10

Feature 11

Feature 12

Feature 13

Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

# ImageNet classification

22,000 categories

14,000,000 images

Hand-engineered features (SIFT, HOG, LBP),
Spatial pyramid,  SparseCoding/Compression

Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

# ImageNet classification: 22,000 classes

…
smoothhound, smoothhound shark, Mustelus mustelus
American smooth dogfish, Mustelus canis
Florida smoothhound, Mustelus norrisi
whitetip shark, reef whitetip shark, Triaenodon obseus
Atlantic spiny dogfish, Squalus acanthias
Pacific spiny dogfish, Squalus suckleyi
hammerhead, hammerhead shark
smooth hammerhead, Sphyrna zygaena
smalleye hammerhead, Sphyrna tudes
shovelhead, bonnethead, bonnet shark, Sphyrna tiburo
angel shark, angelfish, Squatina squatina, monkfish
electric ray, crampfish, numbfish, torpedo
smalltooth sawfish, Pristis pectinatus
guitarfish
roughtail stingray, Dasyatis centroura
butterfly ray
eagle ray
spotted eagle ray, spotted ray, Aetobatus narinari
cownose ray, cow-nosed ray, Rhinoptera bonasus
manta, manta ray, devilfish
Atlantic manta, Manta birostris
devil ray, Mobula hypostoma
grey skate, gray skate, Raja batis
little skate, Raja erinacea
…

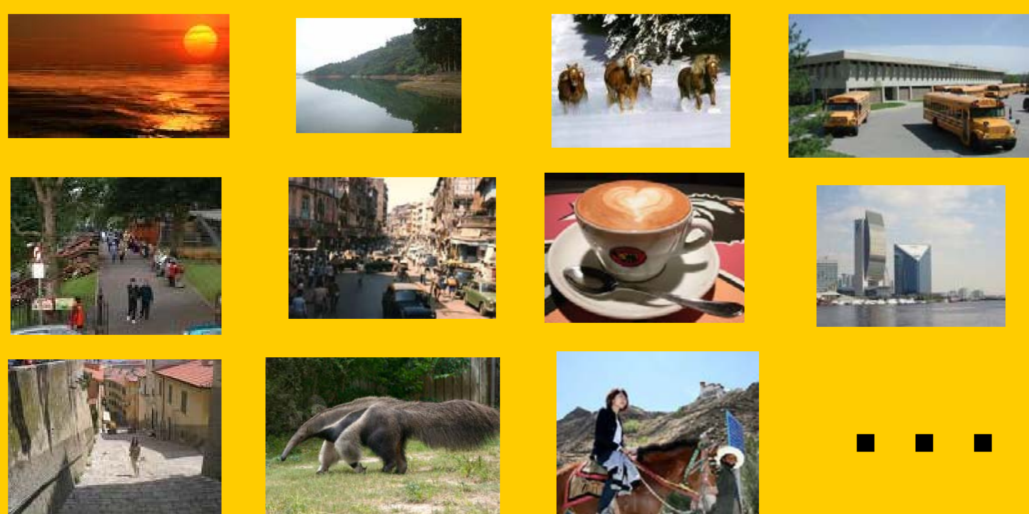
Stingray

Mantaray

# Unsupervised feature learning (Self-taught learning)


Motorcycles

Not motorcycles



…

Unlabeled images

Testing:
What is this?

[Lee, Raina and Ng, 2006; Raina, Lee, Battle, Packer & Ng, 2007]

0.005%

Random guess

9.5%

State-of-the-art
(Weston, Bengio ‘11)

?

Feature learning
From raw pixels

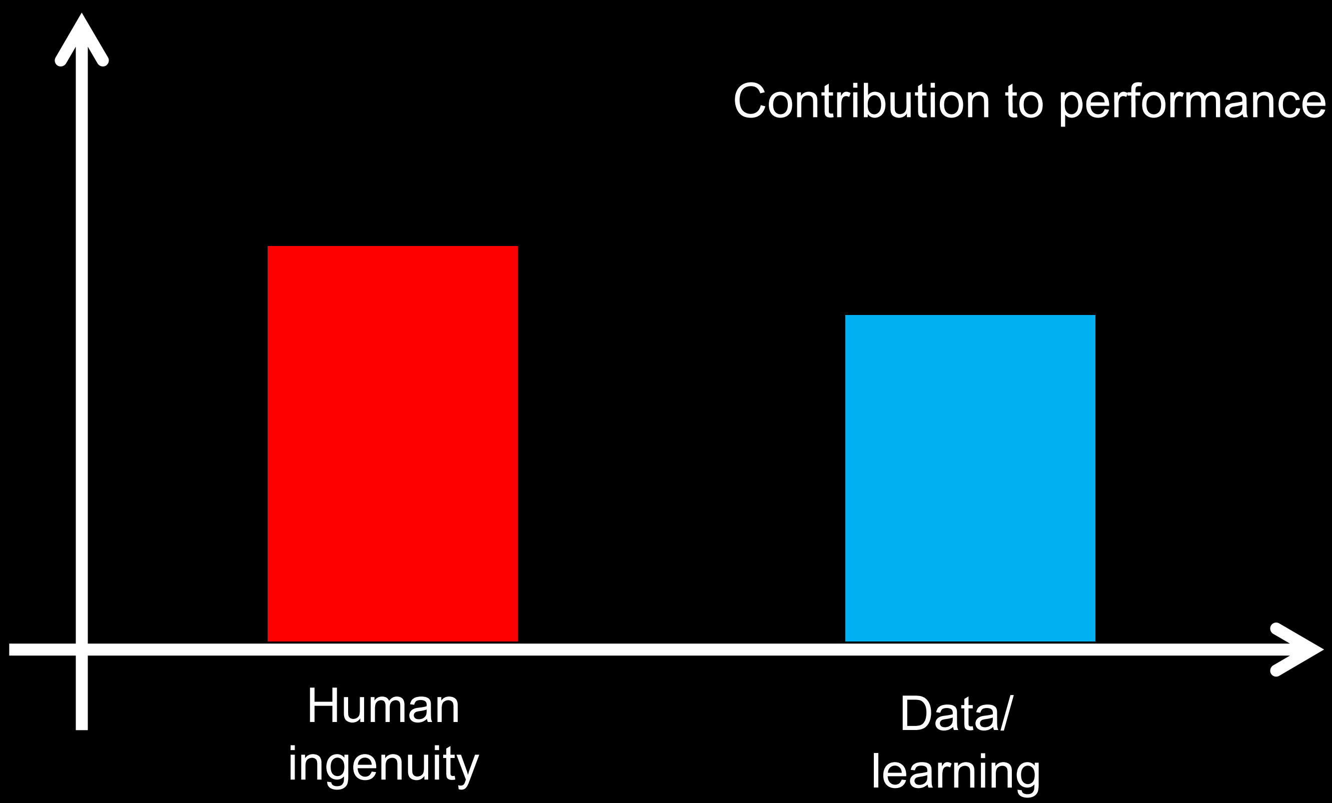
Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

| 0.005% | 9.5% | 21.3% |
| --- | --- | --- |
| Random guess | State-of-the-art (Weston, Bengio '11) | Feature learning From raw pixels |

Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

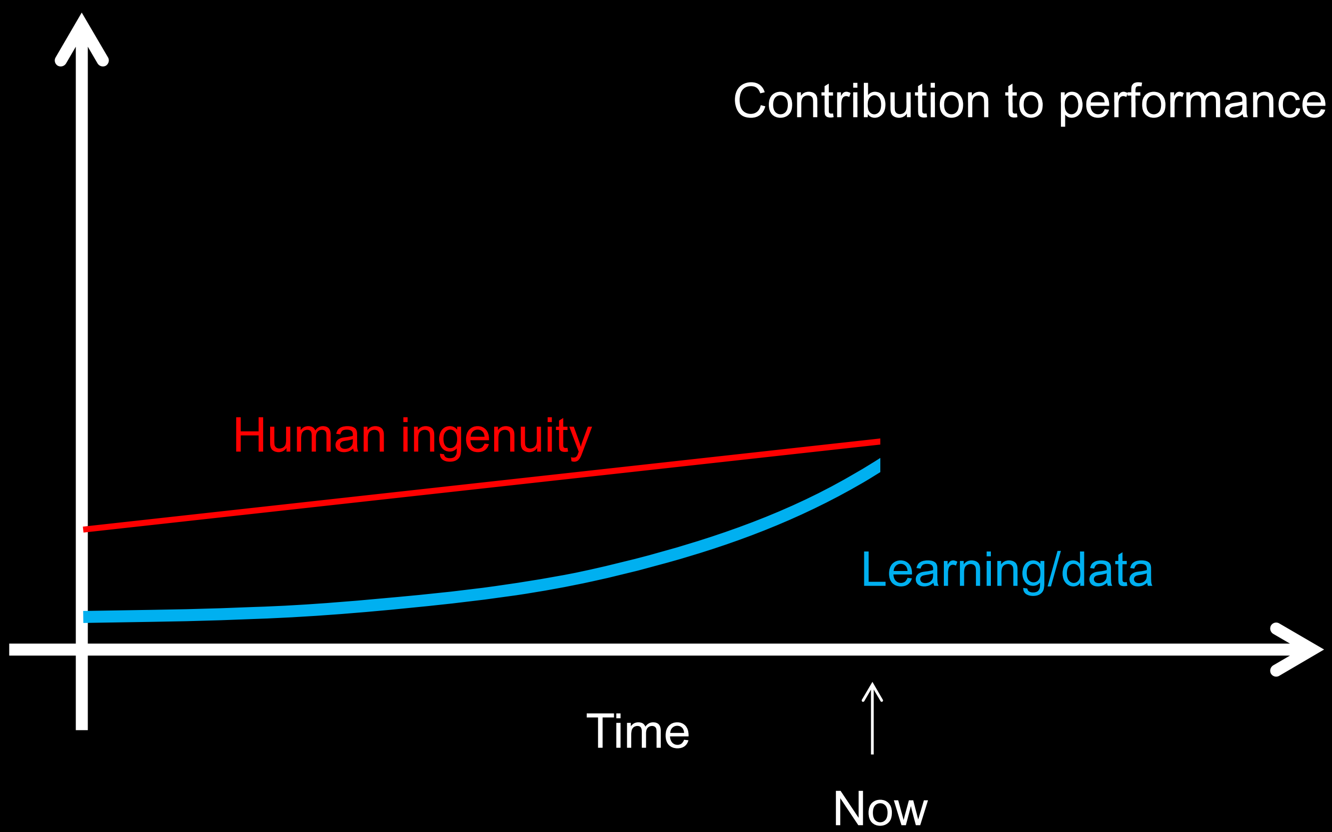
# Discussion: Engineering vs. Data

# Discussion: Engineering vs. Data




Andrew Ng

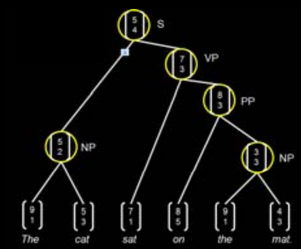
# Discussion: Engineering vs. Data

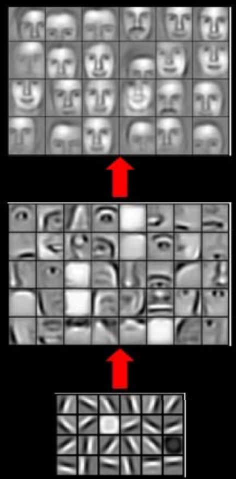



Andrew Ng

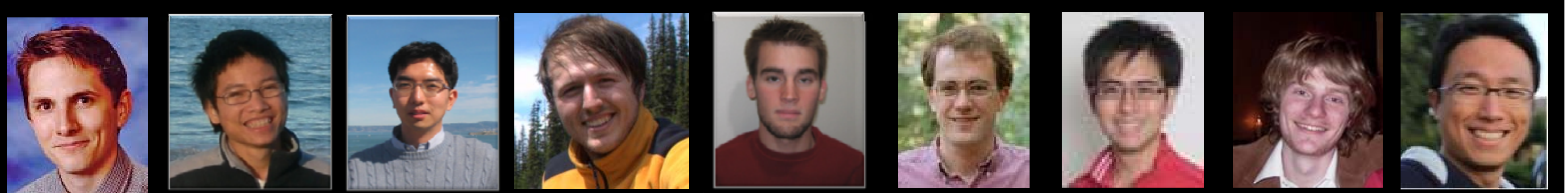
# Deep Learning

- Deep Learning: Lets learn our features.
- Discover the fundamental computational principles that underlie perception.
- Scaling up has been key to achieving good performance.
- Recursive representations for language.
- Online tutorial:
  http://deeplearning.stanford.edu/wiki



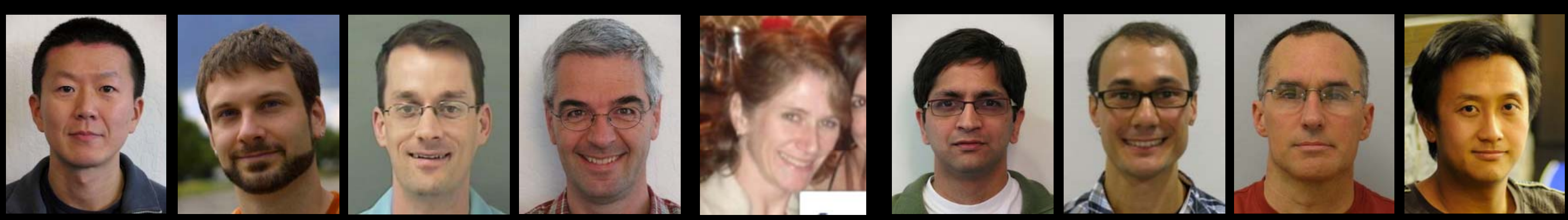



Stanford



Adam Coates | Quoc Le | Honglak Lee | Andrew Saxe | Andrew Maas | Chris Manning | Jiquan Ngiam | Richard Socher | Will Zou

Google



Kai Chen | Greg Corrado | Jeff Dean | Matthieu Devin | Andrea Frome | Rajat Monga | Marc'Aurelio Ranzato | Paul Tucker | Kay Le

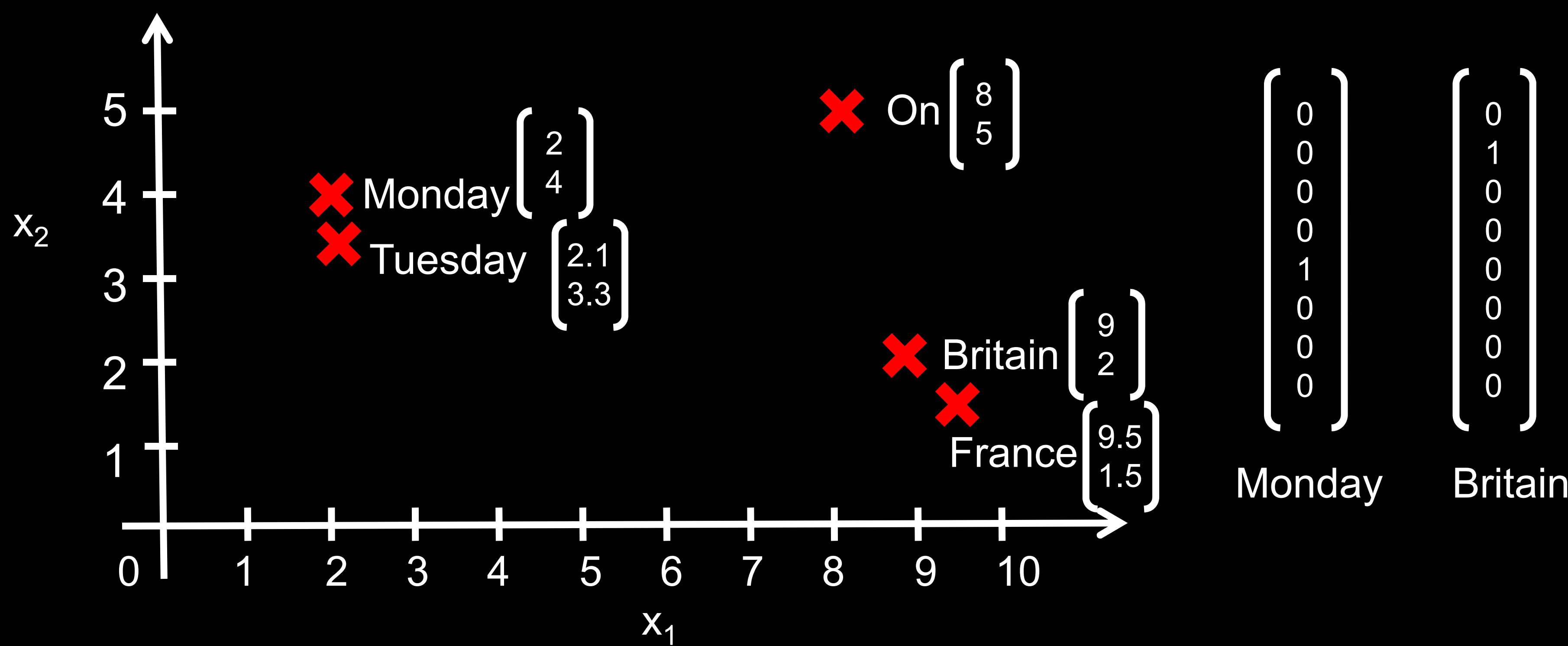
# Language: Learning Recursive Representations

# Feature representations of words

For each word, compute an n-dimensional feature vector for it.

[Distributional representations, or Bengio et al., 2003, Collobert & Weston, 2008.]

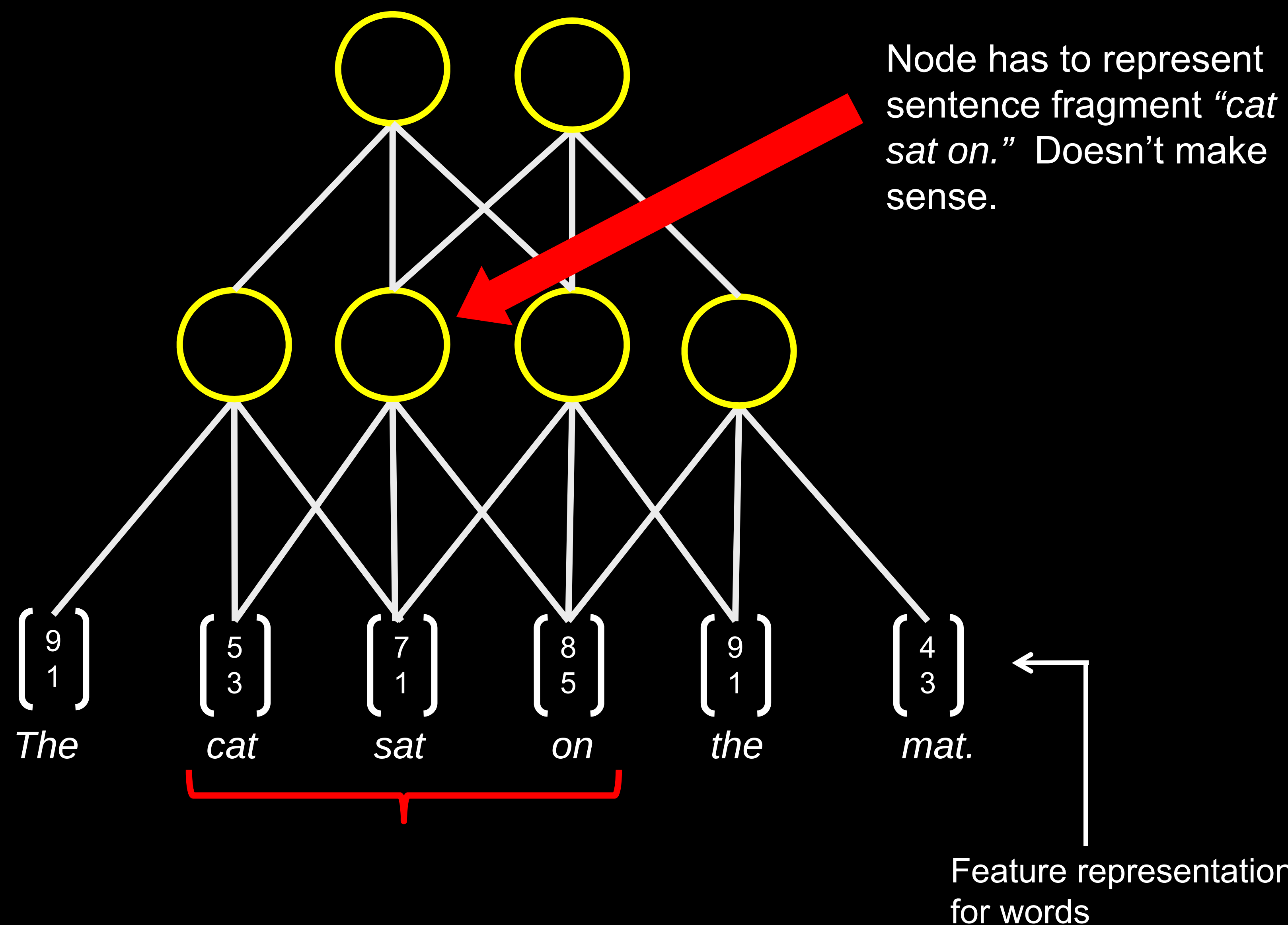
2-d embedding example below, but in practice use ~100-d embeddings.




*On Monday, Britain ….*

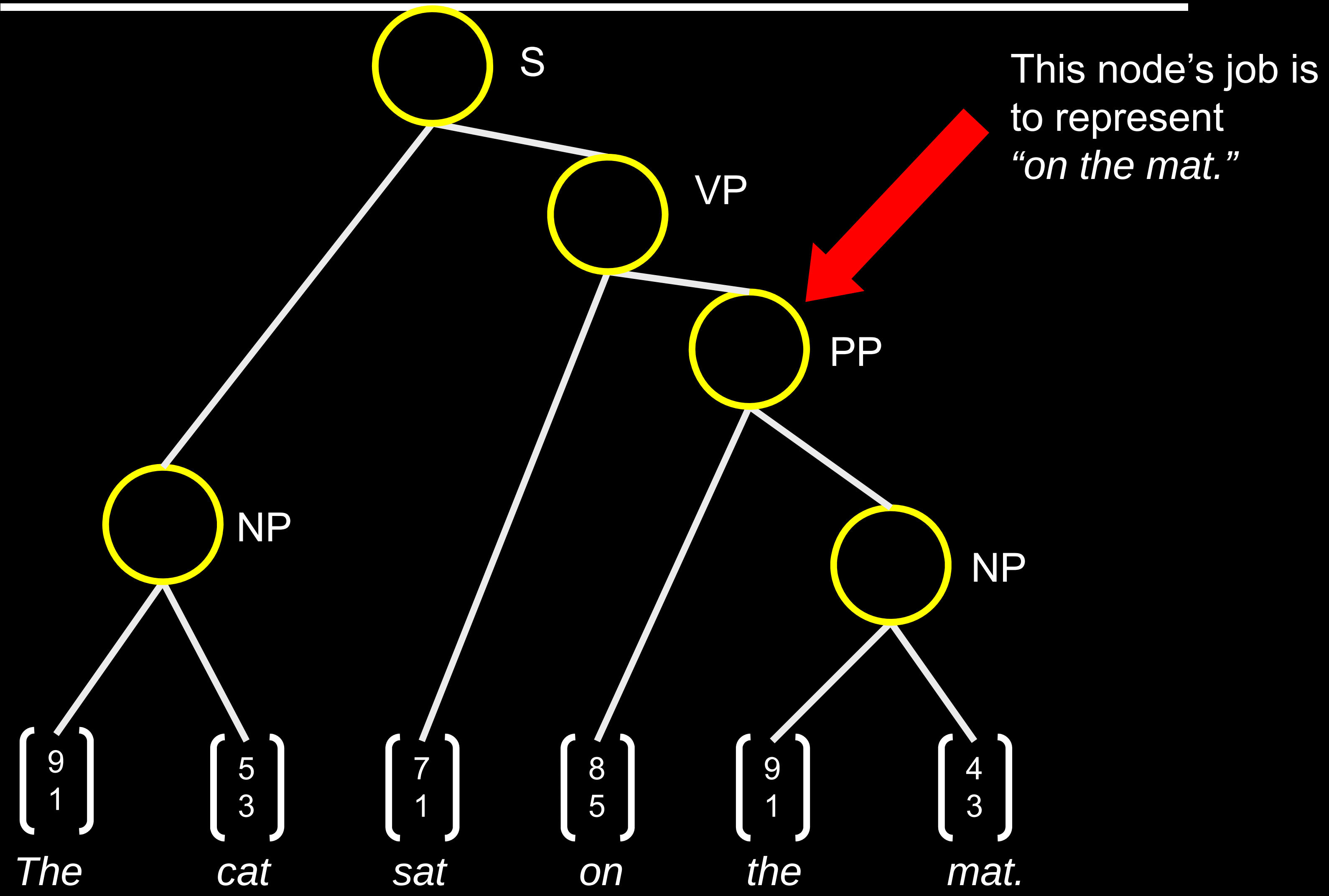
Representation: $\begin{bmatrix} 8 \\ 5 \end{bmatrix}$ $\begin{bmatrix} 2 \\ 4 \end{bmatrix}$ $\begin{bmatrix} 9 \\ 2 \end{bmatrix}$

# “Generic” hierarchy on text doesn’t make sense



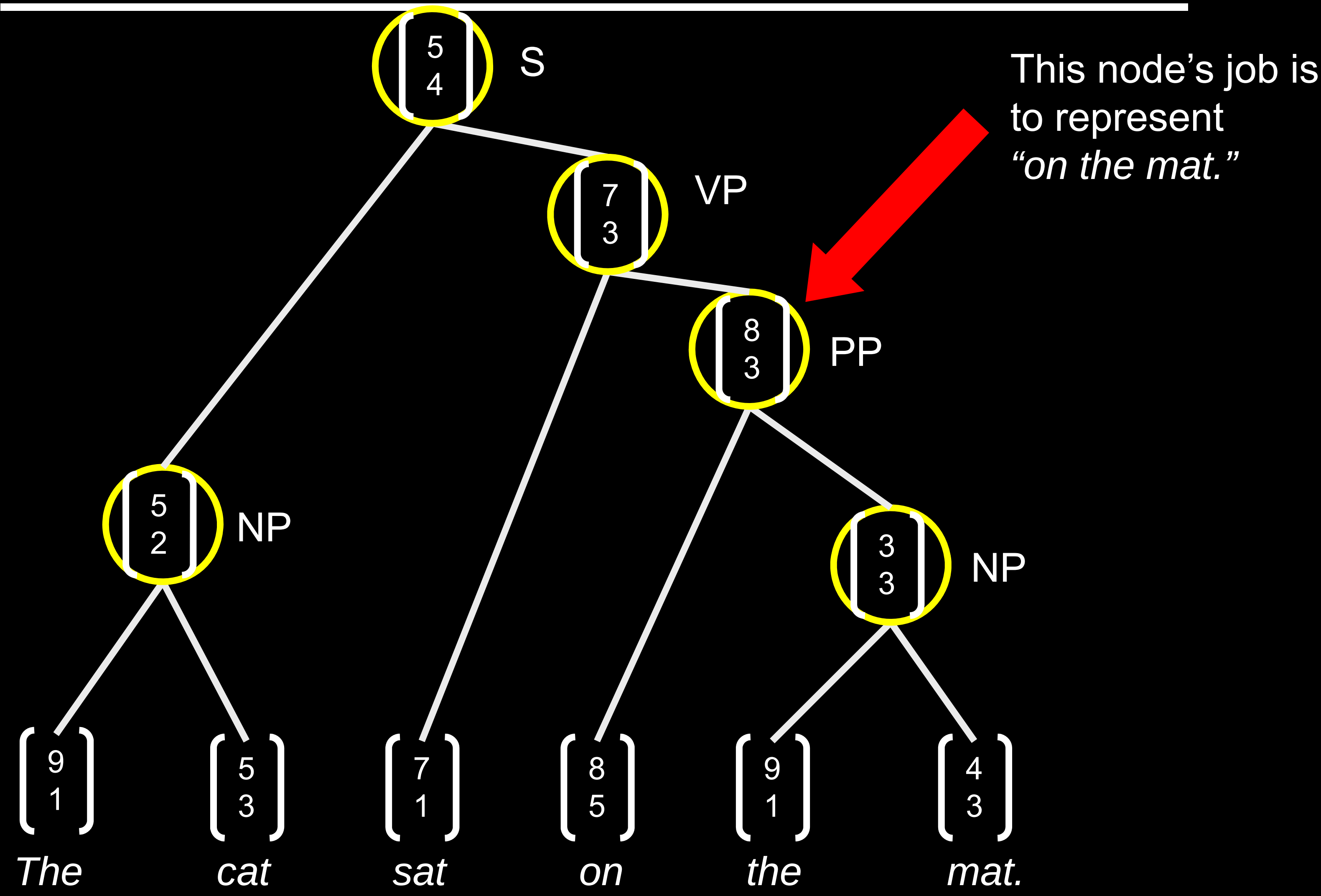

Andrew Ng

# What we want (illustration)

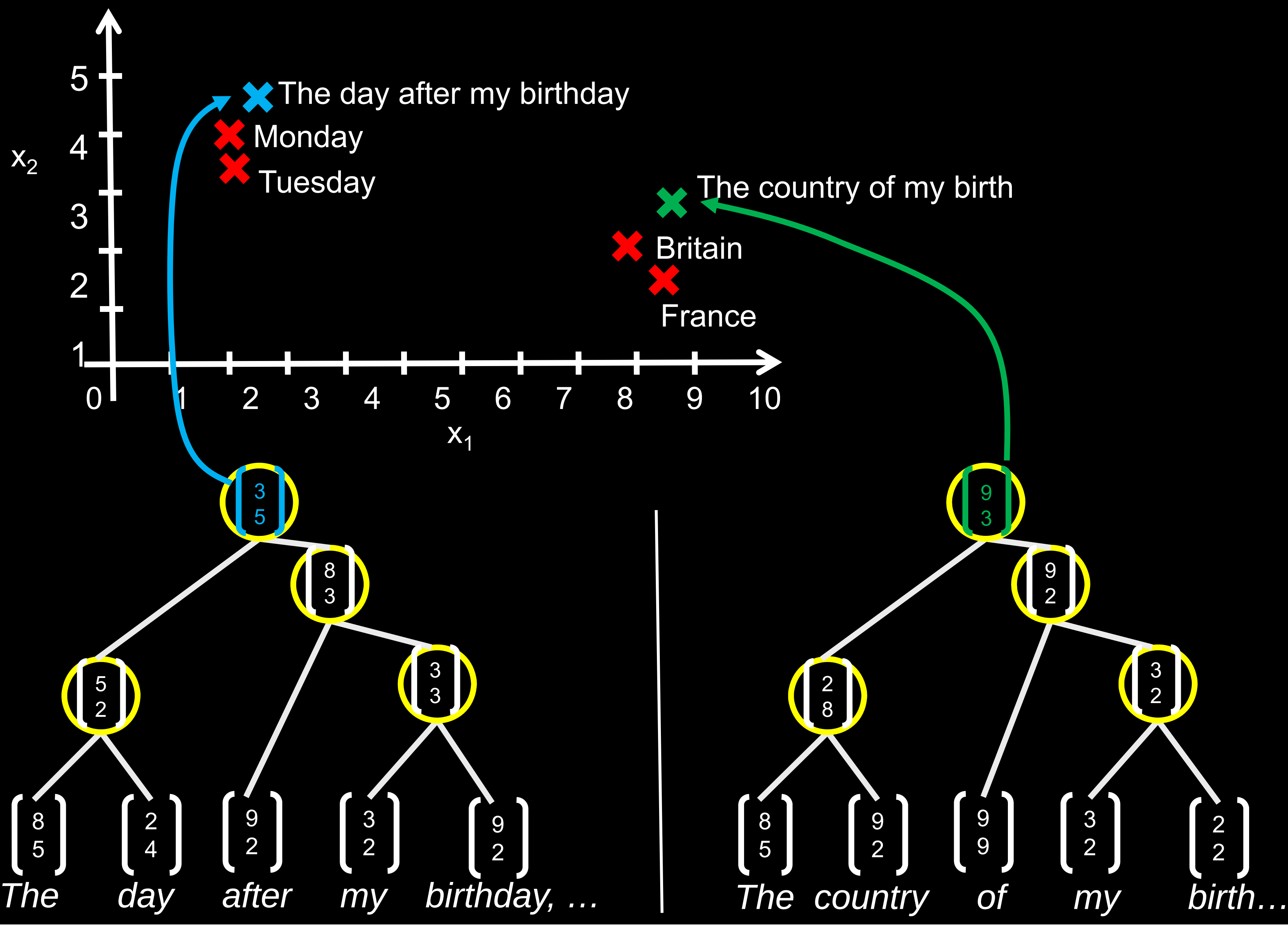



Andrew Ng

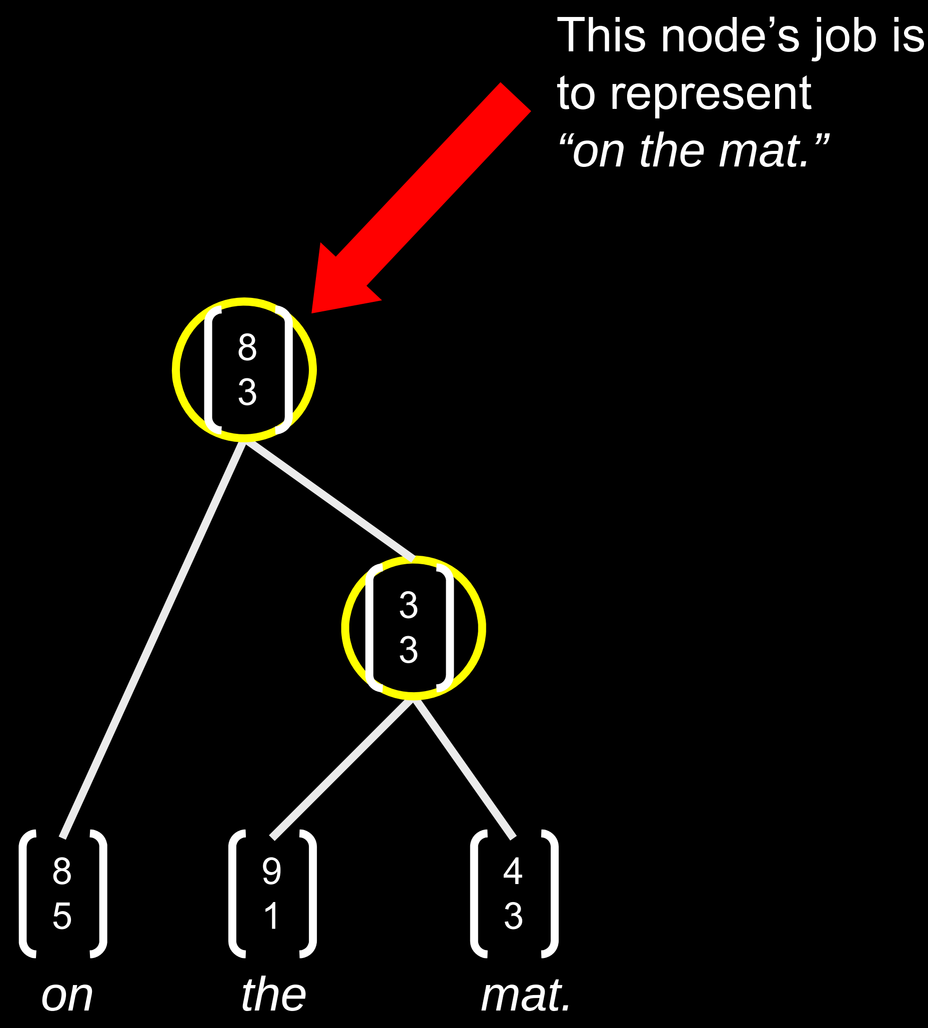
# What we want (illustration)



# What we want (illustration)

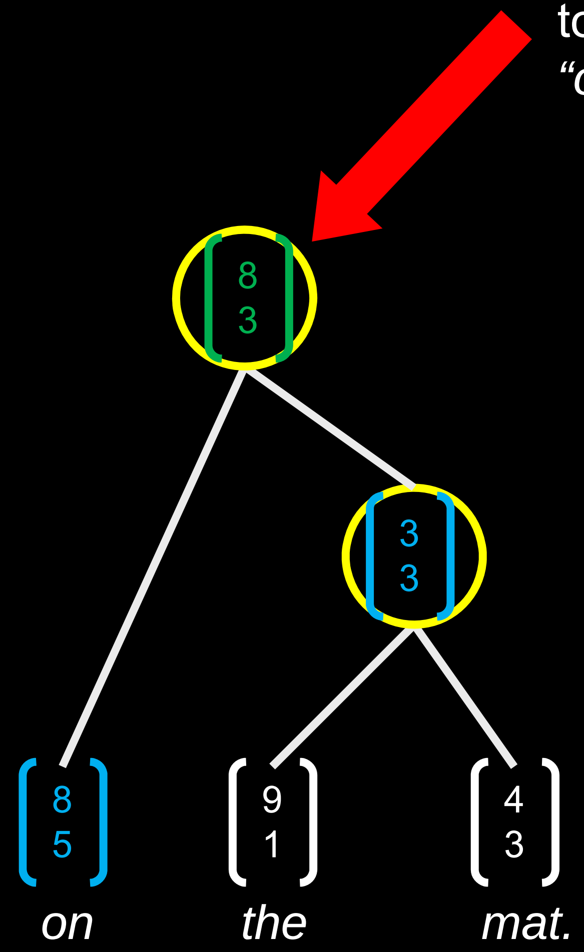


# Learning recursive representations



# Learning recursive representations

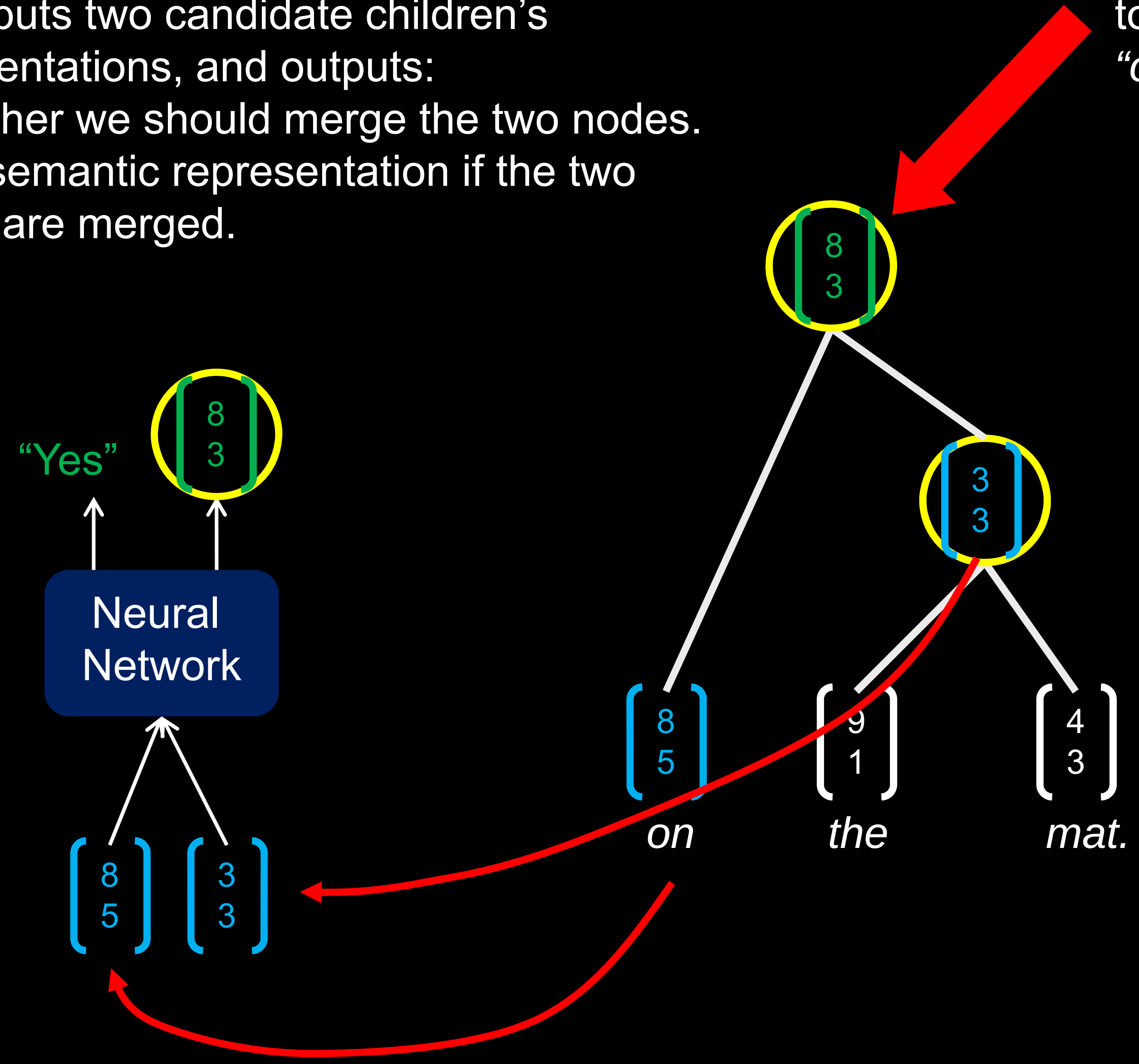
This node's job is to represent *"on the mat."*



# Learning recursive representations

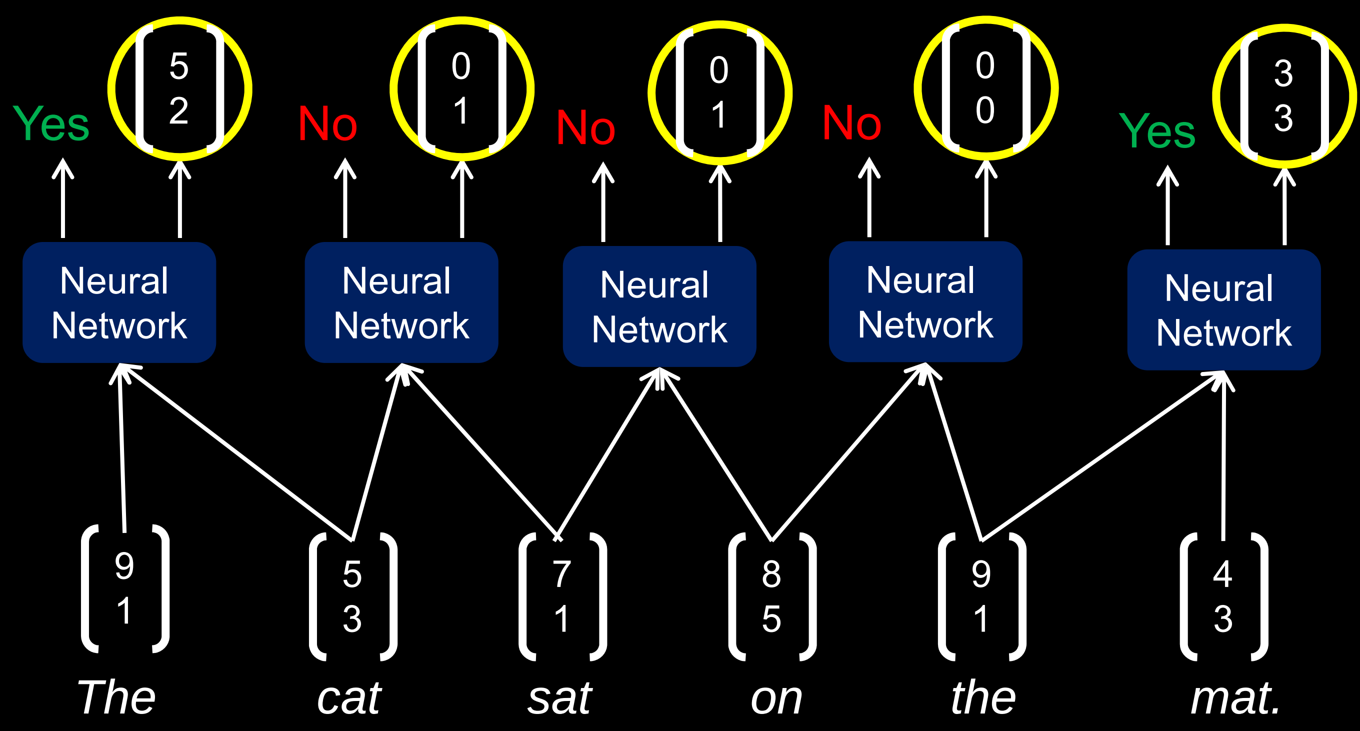
Basic computational unit: Neural Network that inputs two candidate children's representations, and outputs:
- Whether we should merge the two nodes.
- The semantic representation if the two nodes are merged.

This node's job is to represent *"on the mat."*

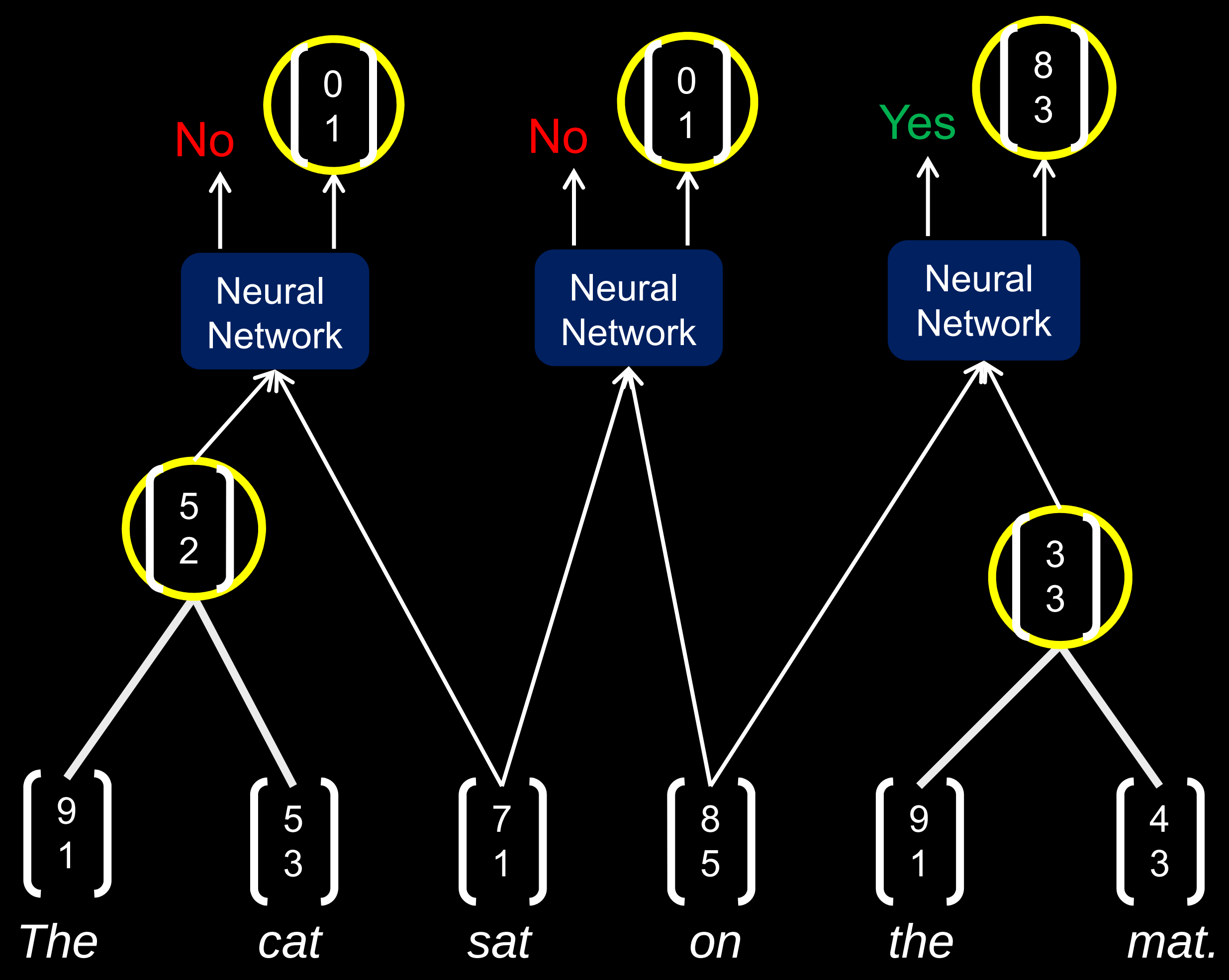



Andrew Ng

# Parsing a sentence

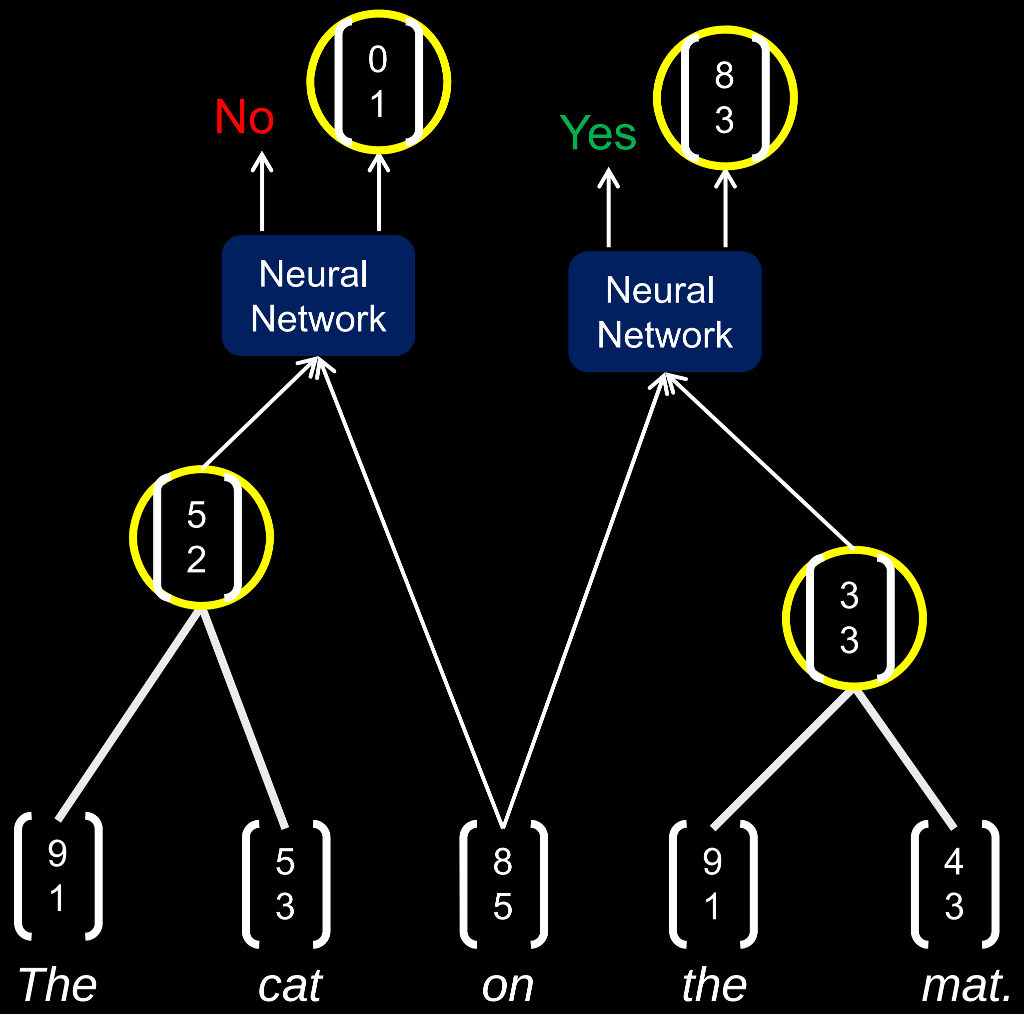


# Parsing a sentence



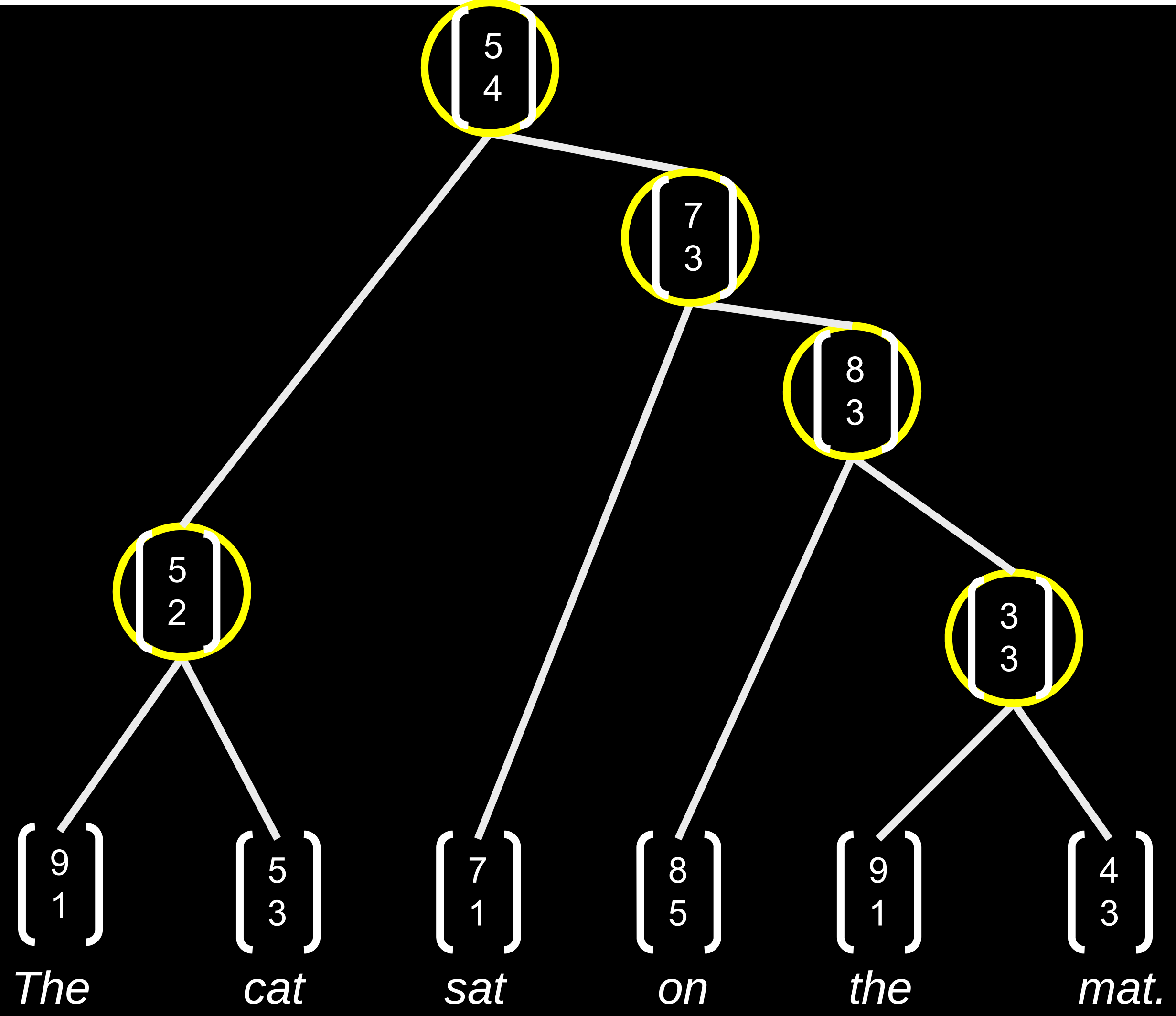

Andrew Ng

# Parsing a sentence




[Socher, Manning & Ng]

# Parsing a sentence



# Finding Similar Sentences

- Each sentence has a feature vector representation.
- Pick a sentence ("center sentence") and list nearest neighbor sentences.
- Often either semantically or syntactically similar. (Digits all mapped to 2.)

| Similarities | Center Sentence | Nearest Neighbor Sentences (most similar feature vector) |
|---|---|---|
| Bad News | Both took further hits yesterday | 1. We 're in for a lot of turbulence ...<br>2. BSN currently has 2.2 million common shares outstanding<br>3. This is panic buying<br>4. We have a couple or three tough weeks coming |
| Something said | I had calls all night long from the States, he said | 1. Our intent is to promote the best alternative, he says<br>2. We have sufficient cash flow to handle that, he said<br>3. Currently, average pay for machinists is 22.22 an hour, Boeing said<br>4. Profit from trading for its own account dropped, the securities firm said |
| Gains and good news | Fujisawa gained 22 to 2,222 | 1. Mochida advanced 22 to 2,222<br>2. Commerzbank gained 2 to 222.2<br>3. Paris loved her at first sight<br>4. Profits improved across Hess's businesses |
| Unknown words which are cities | Columbia , S.C | 1. Greenville , Miss<br>2. UNK , Md<br>3. UNK , Miss<br>4. UNK , Calif |

# Finding Similar Sentences

| Similarities | Center Sentence | Nearest Neighbor Sentences (most similar feature vector) |
|---|---|---|
| Declining to comment = not disclosing | Hess declined to comment | 1. PaineWebber declined to comment<br>2. Phoenix declined to comment<br>3. Campeau declined to comment<br>4. Coastal wouldn't disclose the terms |
| Large changes in sales or revenue | Sales grew almost 2 % to 222.2 million from 222.2 million | 1. Sales surged 22 % to 222.22 billion yen from 222.22 billion<br>2. Revenue fell 2 % to 2.22 billion from 2.22 billion<br>3. Sales rose more than 2 % to 22.2 million from 22.2 million<br>4. Volume was 222.2 million shares , more than triple recent levels |
| Negation of different types | There's nothing unusual about business groups pushing for more government spending | 1. We don't think at this point anything needs to be said<br>2. It therefore makes no sense for each market to adopt different circuit breakers<br>3. You can't say the same with black and white<br>4. I don't think anyone left the place UNK UNK |
| People in bad situations | We were lucky | 1. It was chaotic<br>2. We were wrong<br>3. People had died<br>4. They still are |

# Application: Paraphrase Detection

- Task: Decide whether or not two sentences are paraphrases of each other. (MSR Paraphrase Corpus)

| Method | F1 |
|---|---|
| Baseline | 79.9 |
| Rus et al., (2008) | 80.5 |
| Mihalcea et al., (2006) | 81.3 |
| Islam et al. (2007) | 81.3 |
| Qiu et al. (2006) | 81.6 |
| Fernando & Stevenson (2008) (WordNet based features) | 82.4 |
| Das et al. (2009) | 82.7 |
| Wan et al (2006) (many features: POS, parsing, BLEU, etc.) | 83.0 |
| **Stanford Feature Learning** | **83.4** |

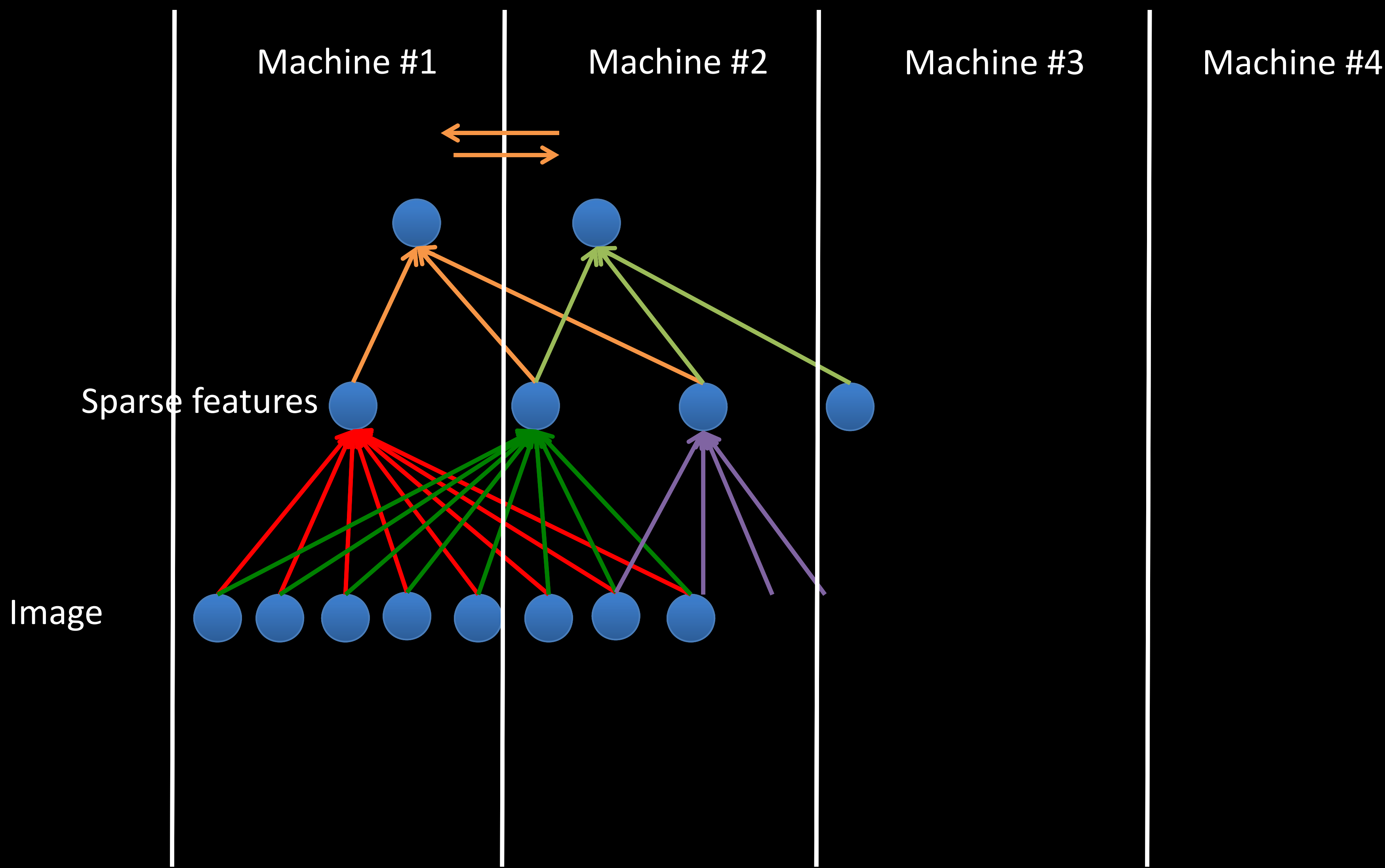
END END END

[redacted]

# Scaling up: Discovering object classes

[Quoc V. Le, Marc'Aurelio Ranzato, Rajat  Monga, Greg Corrado, Matthieu Devin, Kai Chen, Jeff Dean]

# Local Receptive Field networks

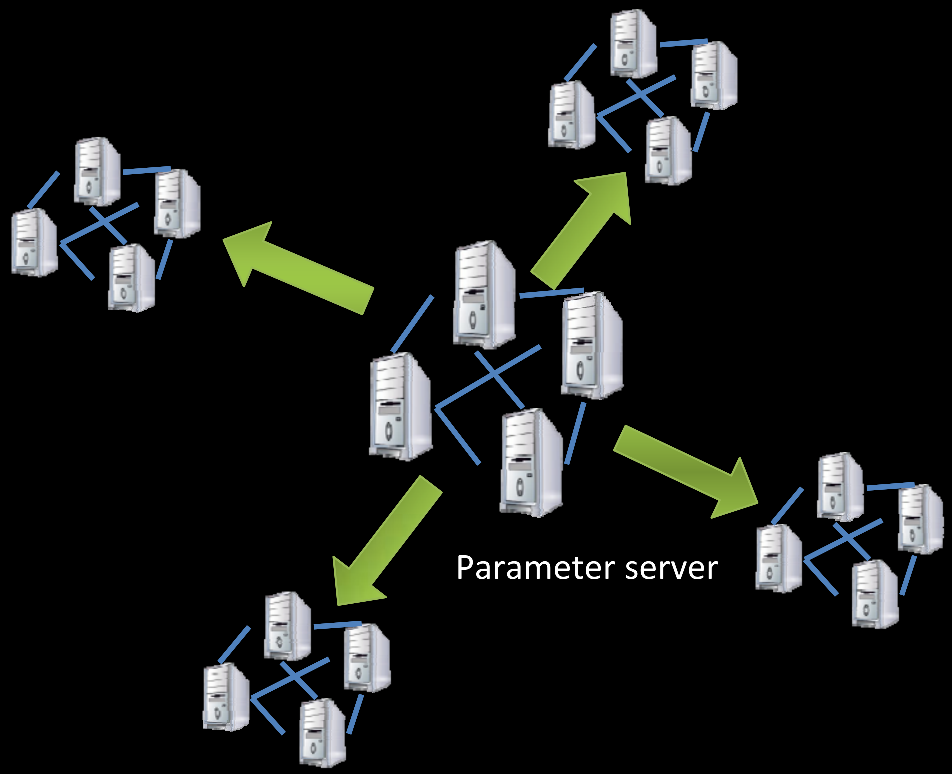



Le, et al., *Tiled Convolutional Neural Networks*. NIPS 2010

# Asynchronous Parallel SGD

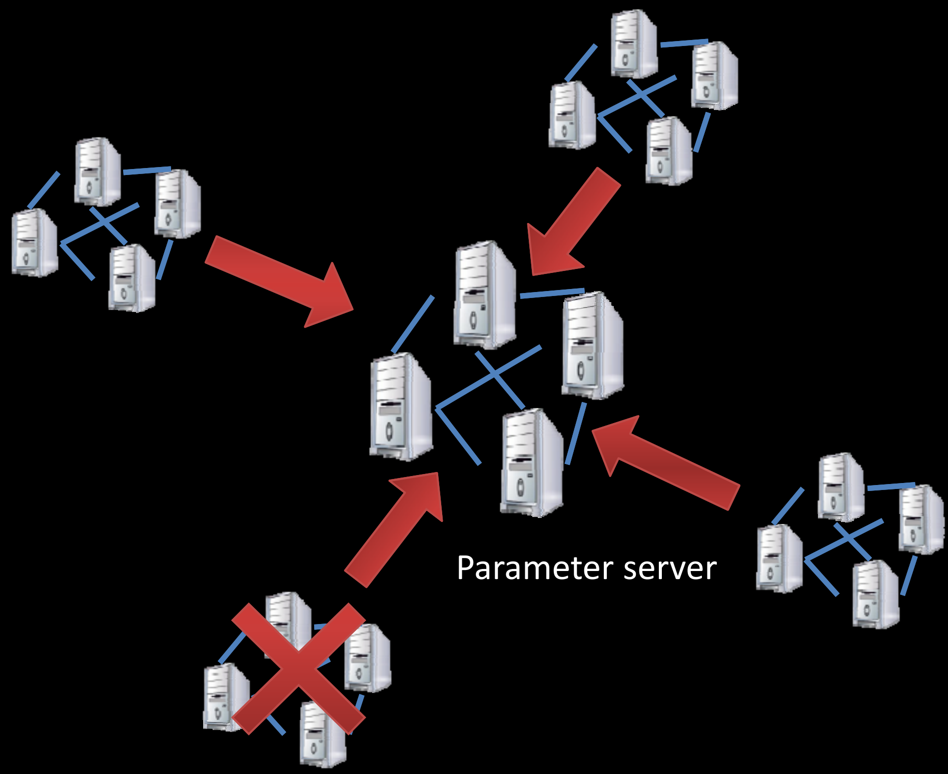
Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

# Asynchronous Parallel SGD




Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

# Asynchronous Parallel SGD

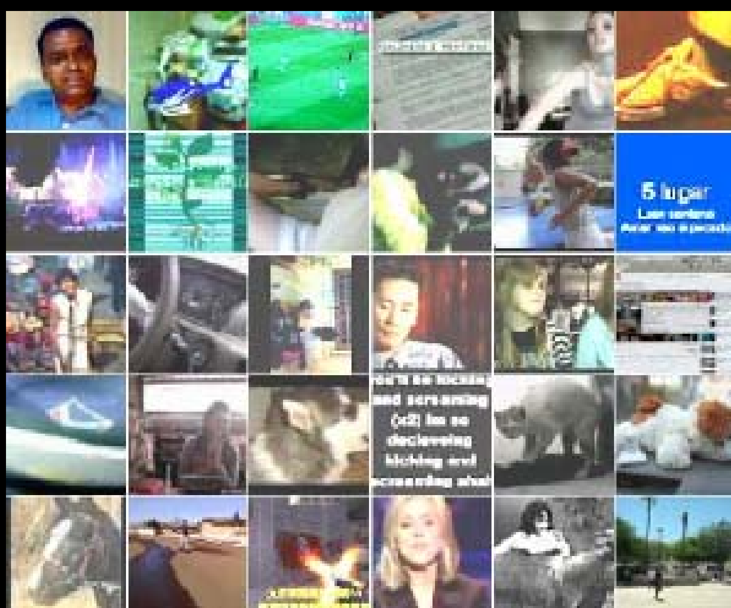



Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

# Training procedure

What features can we learn if we train a massive model on a massive amount of data. Can we learn a “grandmother cell”?

- Train on 10 million images (YouTube)
- 1000 machines (16,000 cores) for 1 week.
- 1.15 billion parameters
- Test on novel images



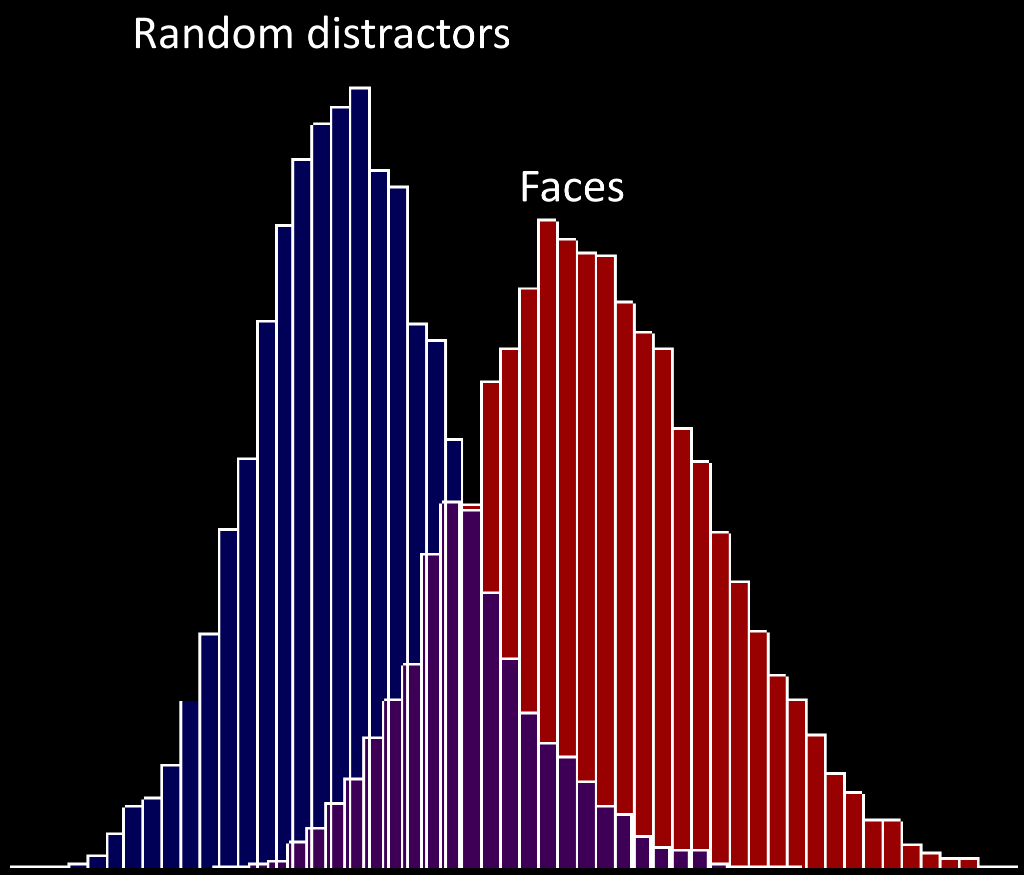
Training set (YouTube)

Test set (FITW + ImageNet)

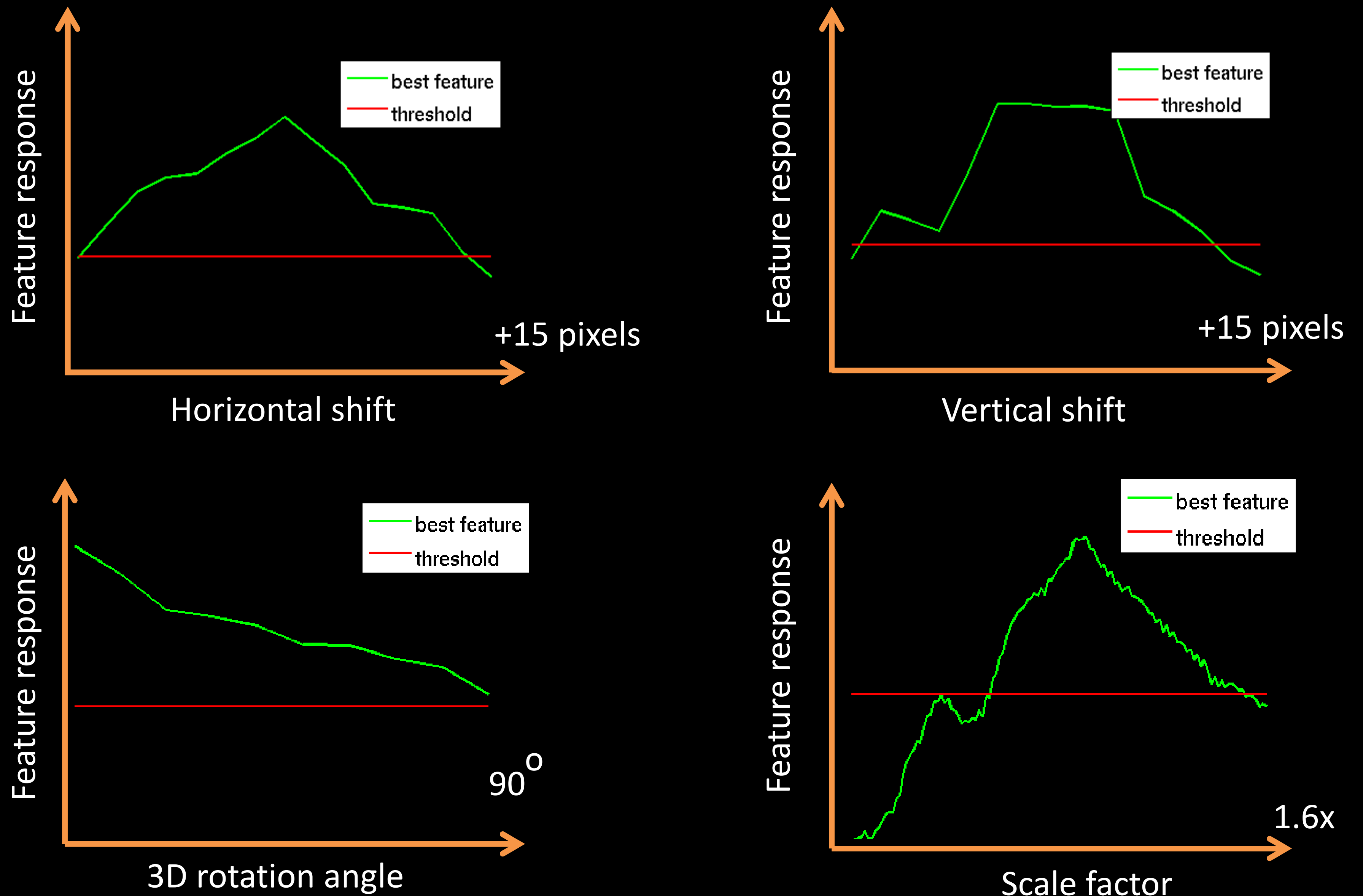
# Face neuron

Top Stimuli from the test set

Optimal stimulus by numerical optimization


Random distractors
Faces

# Invariance properties




Andrew Ng

# Cat neuron

Top Stimuli from the test set

Average of top stimuli from test set

Andrew Ng

# ImageNet classification

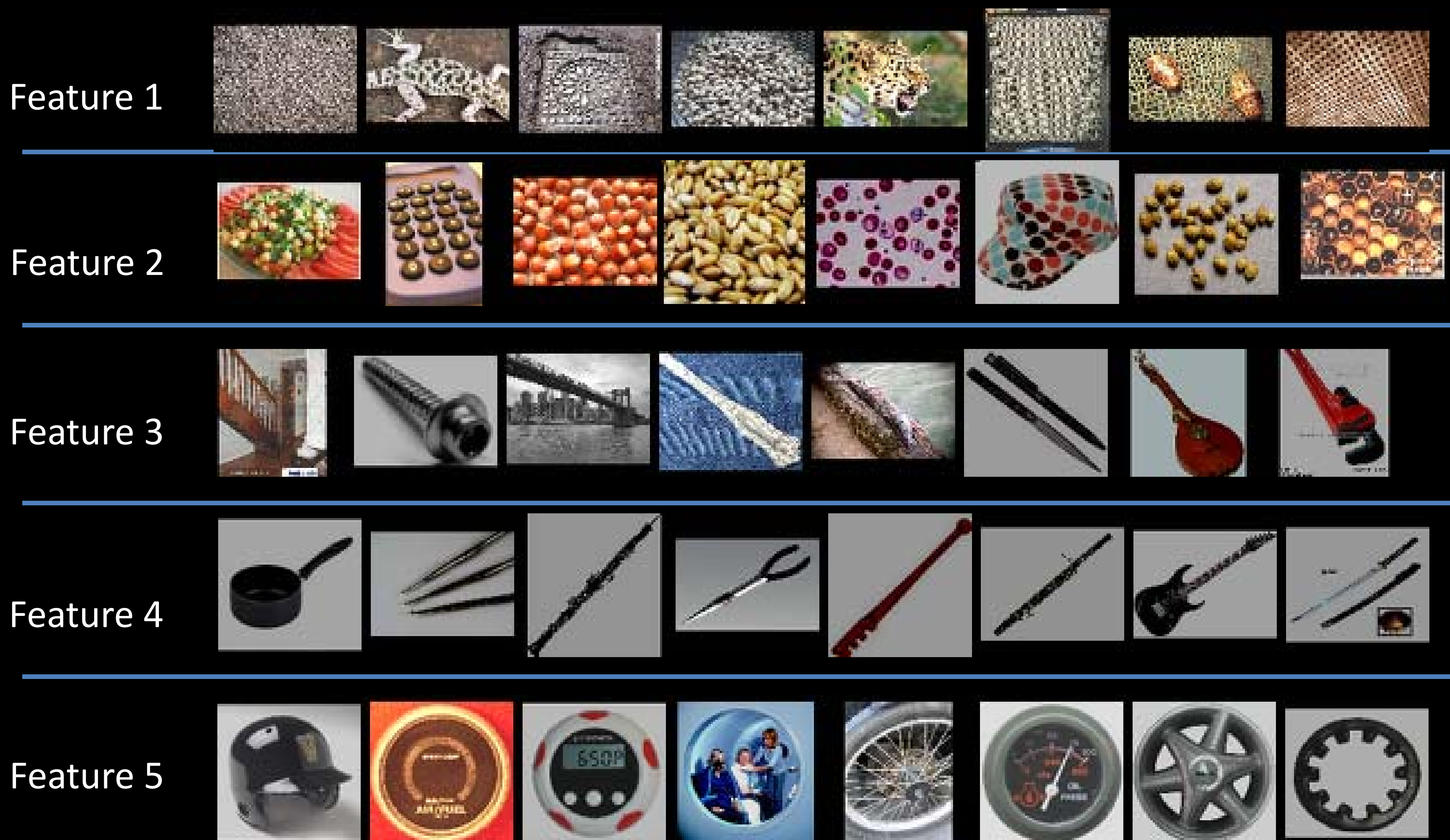
20,000 categories

16,000,000 images

Others: Hand-engineered features (SIFT, HOG, LBP), Spatial pyramid,  SparseCoding/Compression

Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

# Best stimuli




Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

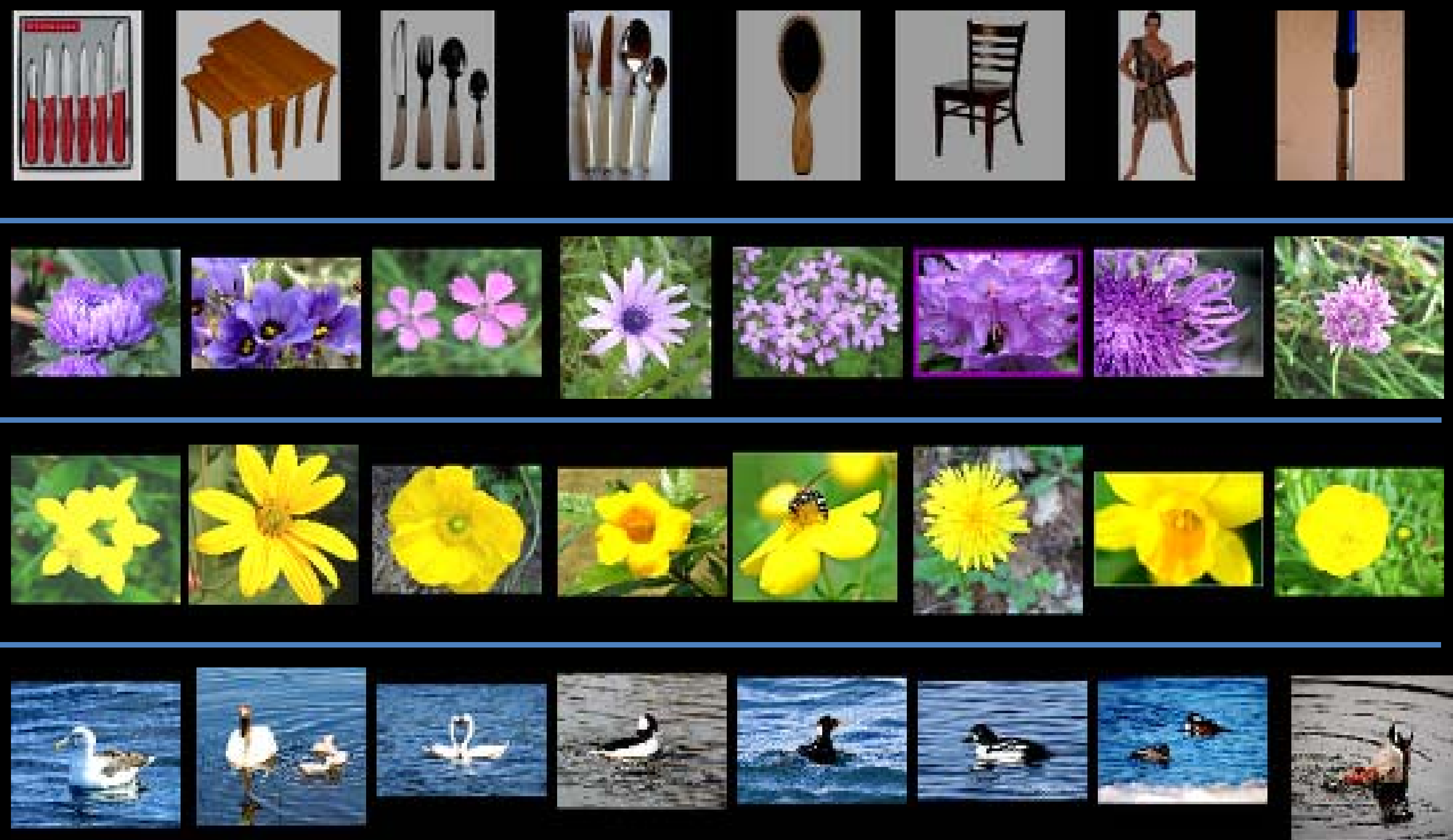
# Best stimuli

Feature 6

Feature 7

Feature 8

Feature 9



Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

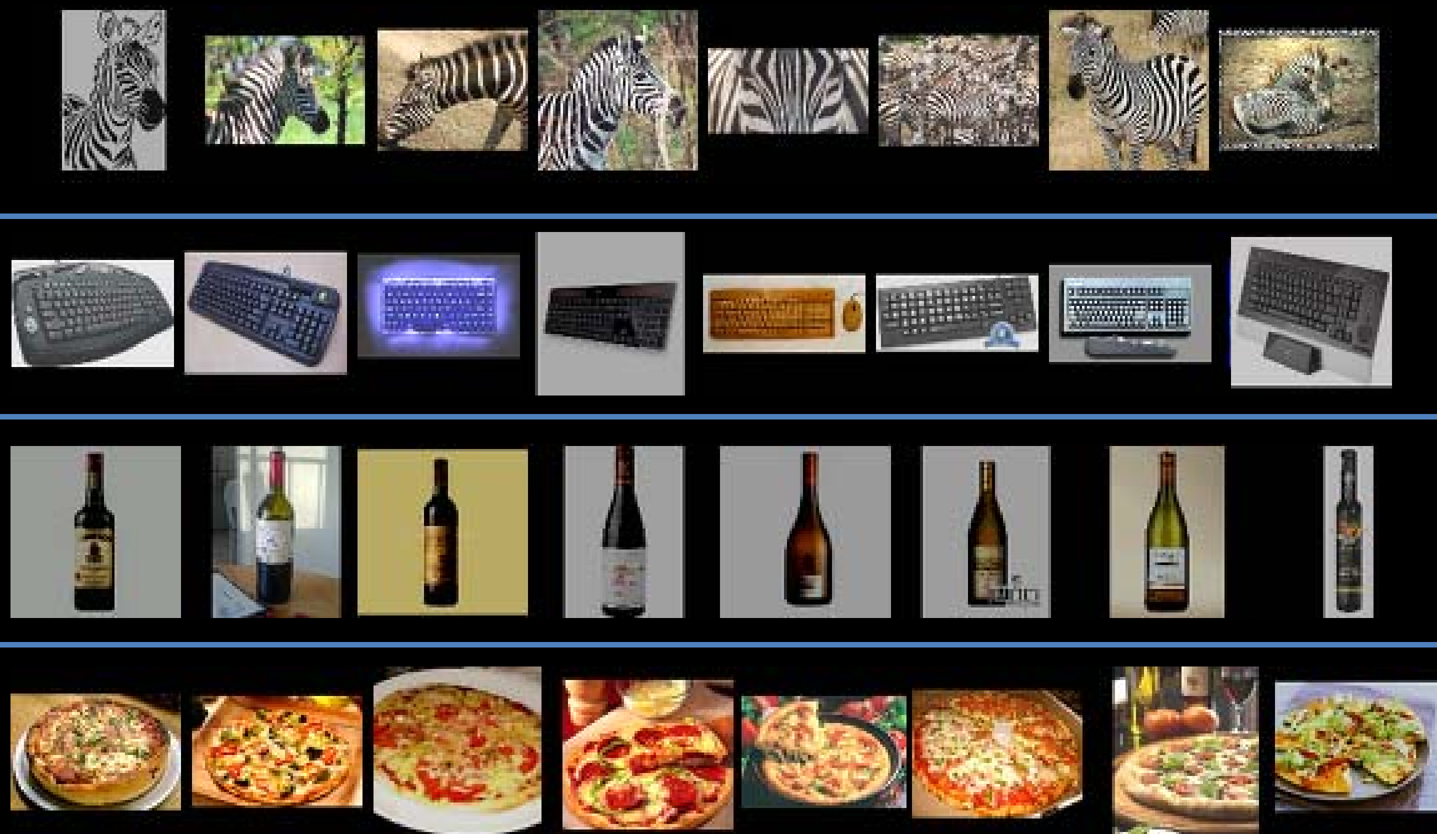
# Best stimuli

Feature 10

Feature 11

Feature 12

Feature 13



Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

# 20,000 is a lot of categories…

…
smoothhound, smoothhound shark, Mustelus mustelus
American smooth dogfish, Mustelus canis
Florida smoothhound, Mustelus norrisi
whitetip shark, reef whitetip shark, Triaenodon obseus
Atlantic spiny dogfish, Squalus acanthias
Pacific spiny dogfish, Squalus suckleyi
hammerhead, hammerhead shark
smooth hammerhead, Sphyrna zygaena
smalleye hammerhead, Sphyrna tudes
shovelhead, bonnethead, bonnet shark, Sphyrna tiburo
angel shark, angelfish, Squatina squatina, monkfish
electric ray, crampfish, numbfish, torpedo
smalltooth sawfish, Pristis pectinatus
guitarfish
roughtail stingray, Dasyatis centroura
butterfly ray
eagle ray
spotted eagle ray, spotted ray, Aetobatus narinari
cownose ray, cow-nosed ray, Rhinoptera bonasus
manta, manta ray, devilfish
Atlantic manta, Manta birostris
devil ray, Mobula hypostoma
grey skate, gray skate, Raja batis
little skate, Raja erinacea
…

Stingray

Mantaray

| 0.005% | 9.5% | ? |
| --- | --- | --- |
| Random guess | State-of-the-art (Weston, Bengio ‘11) | Feature learning From raw pixels |

Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

0.005%
Random guess

9.5%
State-of-the-art
(Weston, Bengio '11)

15.8%
Feature learning
From raw pixels

ImageNet 2009 (10k categories): Best published result: 17%
(Sanchez & Perronnin '11 ),
Our method: 20%

Using only 1000 categories, our method > 50%

Le, et al., *Building high-level features using large-scale unsupervised learning*. ICML 2012

# Speech recognition on Android

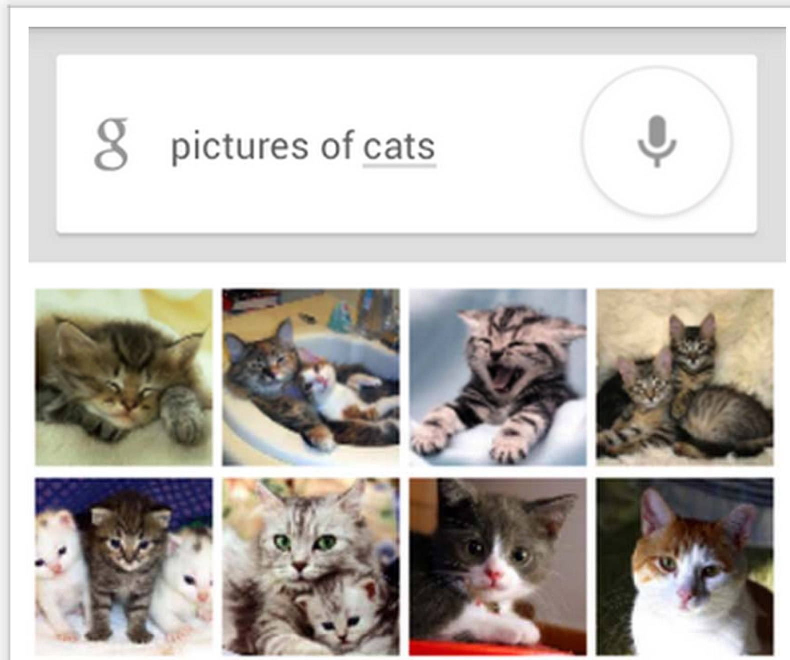
AUG
6

## Speech Recognition and Deep Learning

Posted by Vincent Vanhoucke, Research Scientist, Speech Team

The New York Times recently published an article about Google's large scale deep learning project, which learns to discover patterns in large datasets, including... cats on YouTube!

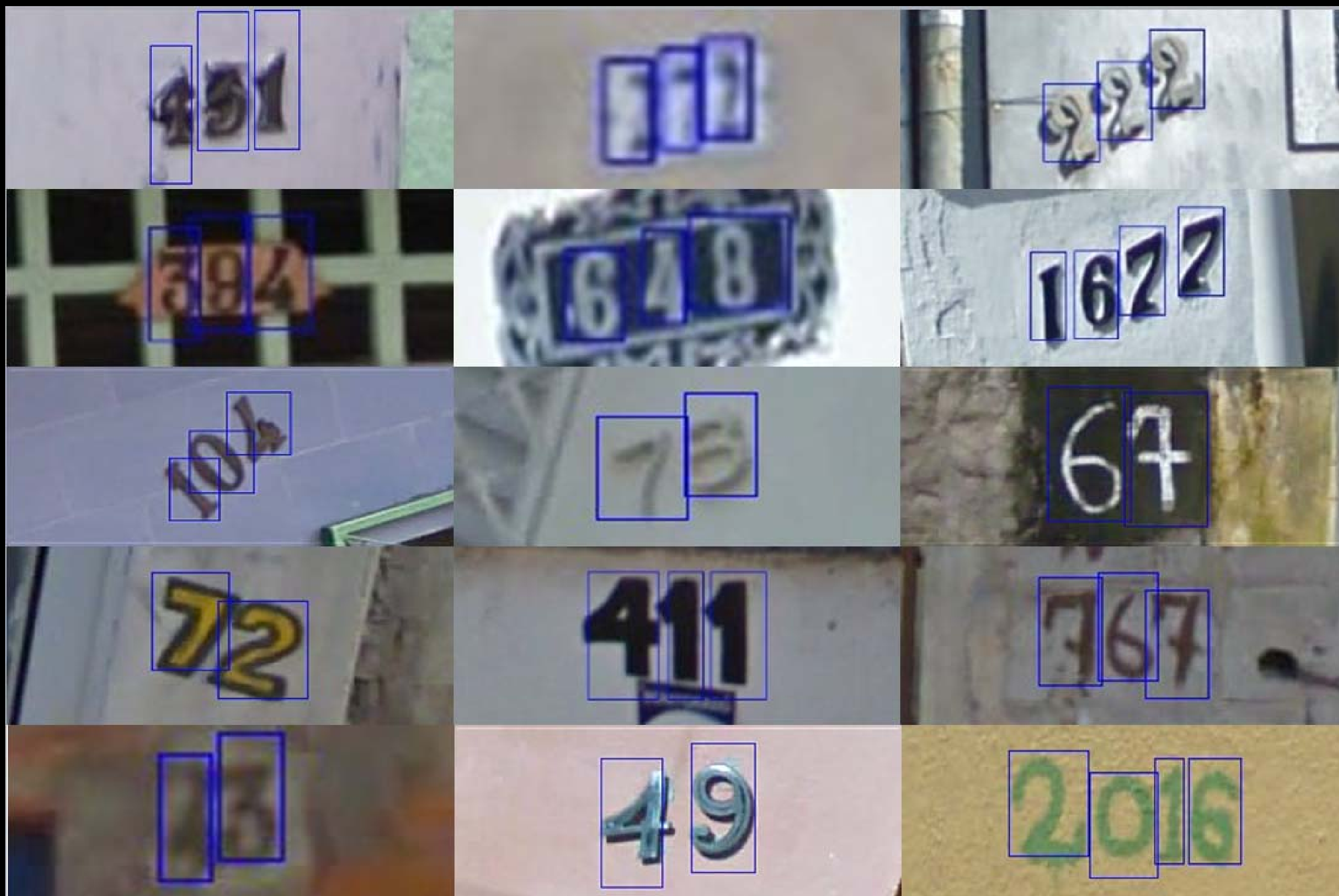
What's the point of building a gigantic cat detector you might ask? When you combine large amounts of data, large-scale distributed computing and powerful machine learning algorithms, you can apply the technology to address a large variety of practical problems.




With the launch of the latest Android platform release, Jelly Bean, we've taken a significant step towards making that technology useful: when you speak to your Android phone, chances are, you are talking to a neural network trained to recognize your speech.

Using neural networks for speech recognition is nothing new: the first proofs of concept were developed in the late

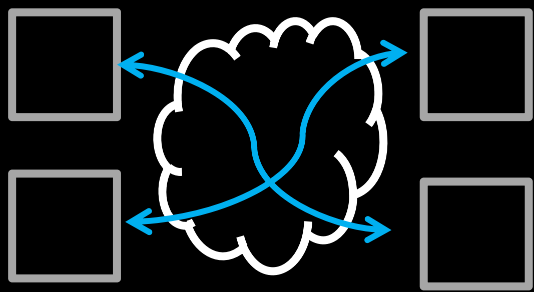
# Application to Google Streetview



[with Yuval Netzer, Julian Ibarz]

# Scaling up with HPC

"Cloud" infrastructure



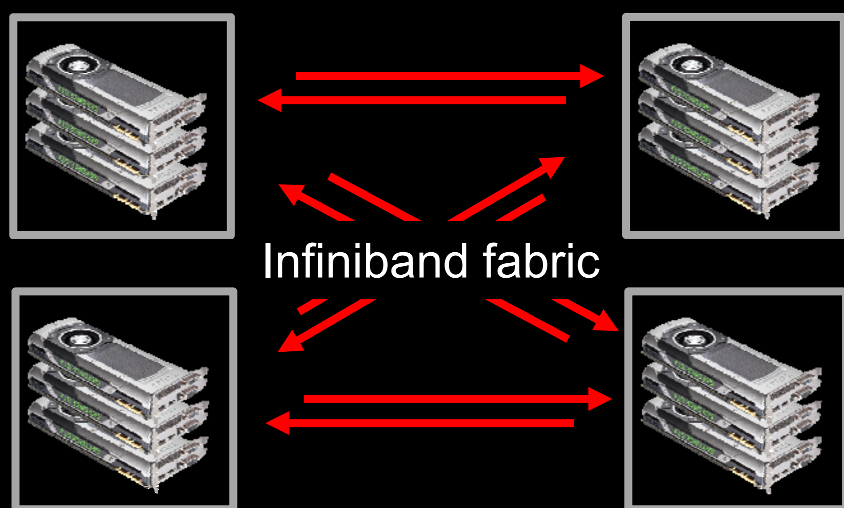
Many inexpensive nodes.

**Comm. bottlenecks, node failures.**

GPUs with CUDA

1 very fast node.

**Limited memory; hard to scale out.**

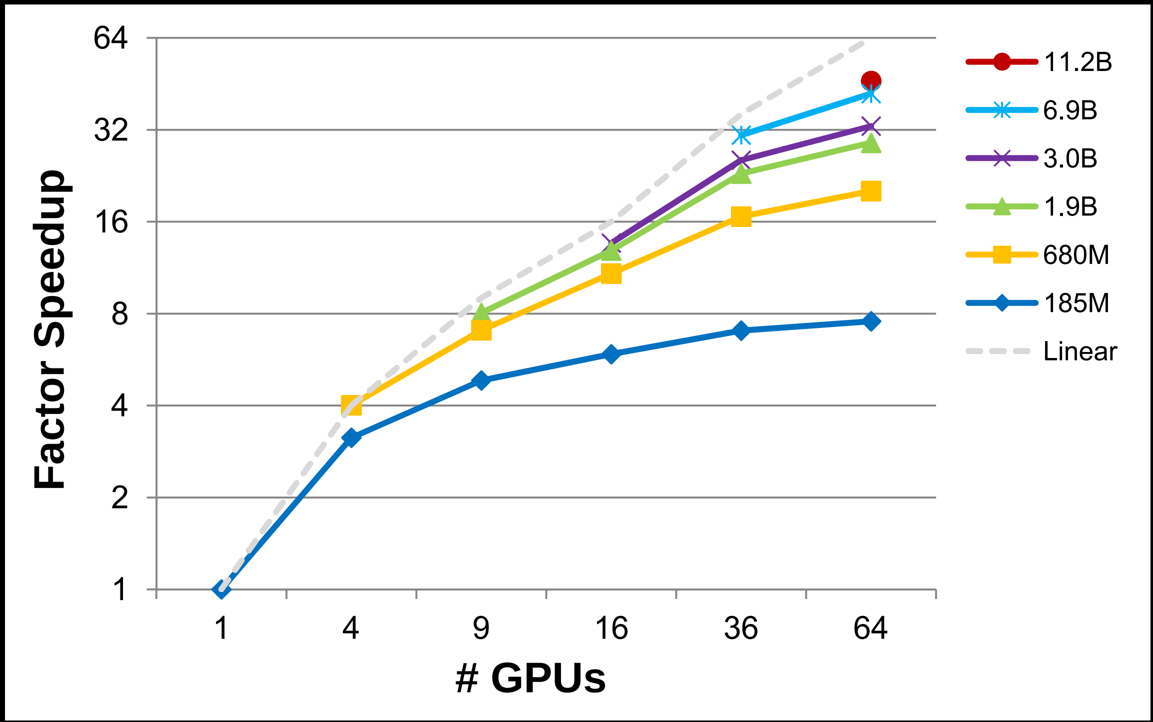



HPC cluster: GPUs with Infiniband

**Difficult to program---lots of MPI and CUDA code.**

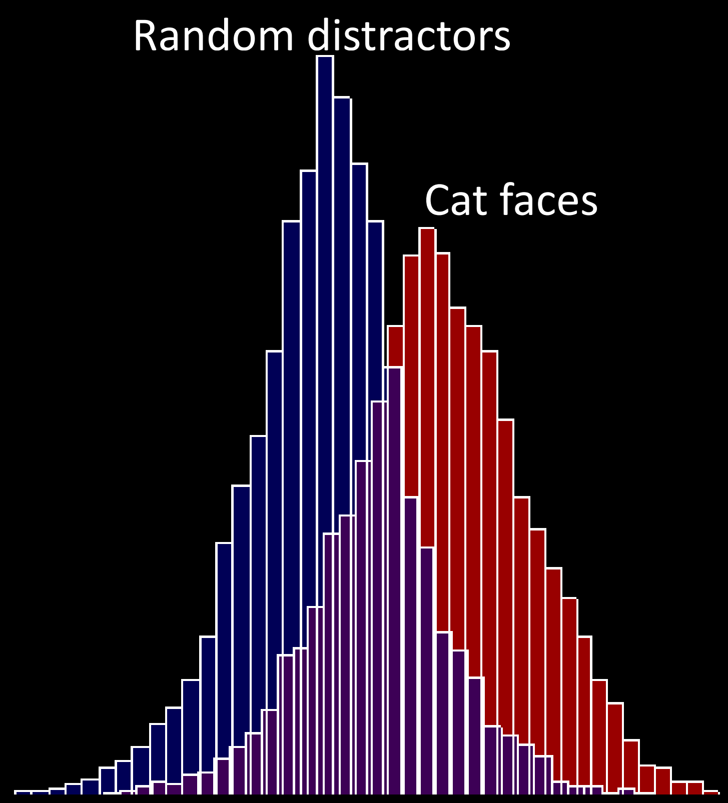
# Stanford GPU cluster

- **Current system**
  - 64 GPUs in 16 machines.
  - Tightly optimized CUDA for UFL/DL operations.
  - 47x faster than single-GPU implementation.




  - Train 11.2 billion parameter, 9 layer neural network in < 4 days.

# Cat face neuron



# Control experiments

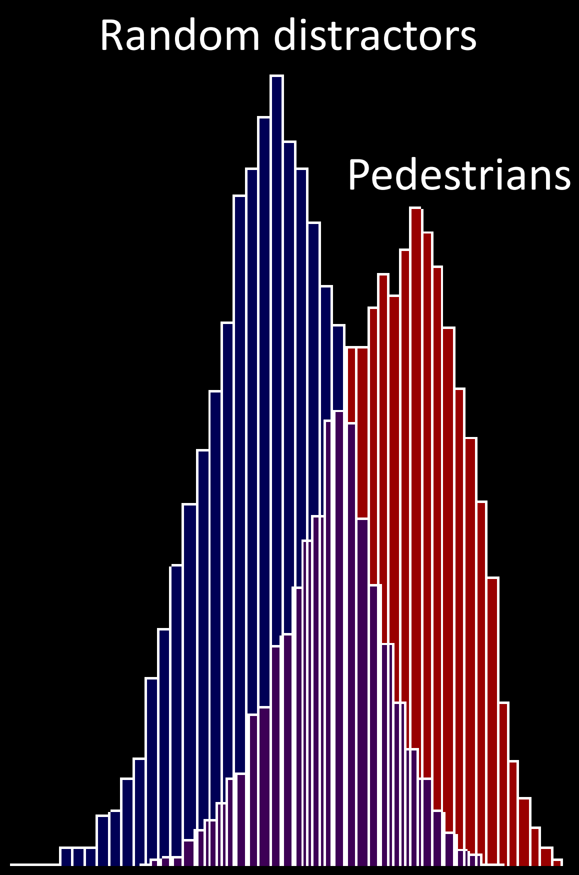
| Concept | Random guess | Random weights | Best linear filter | Best first layer neuron | Best neuron | Best neuron without contrast normalization |
|---|---|---|---|---|---|---|
| Faces | 64.8% | 67.0% | 74.0% | 71.0% | **81.7%** | 78.5% |
| Upright human bodies | 64.8% | 66.5% | 68.1% | 67.2% | **76.8%** | 71.8% |
| Cats | 64.8% | 66.0% | 67.8% | 67.1% | **74.6%** | 69.3% |
| Cars | 64.8% | 65.3% | 65.2% | 65.3% | 65.7% | 65.3% |

# Visualization

Top Stimuli from the test set

Optimal stimulus by numerical optimization

# Pedestrian neuron



# Conclusion

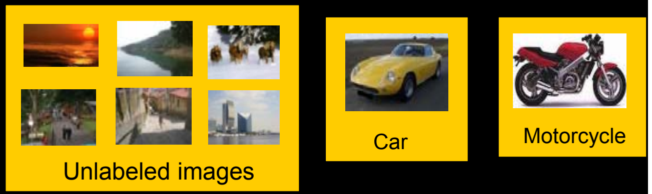
# Unsupervised Feature Learning Summary

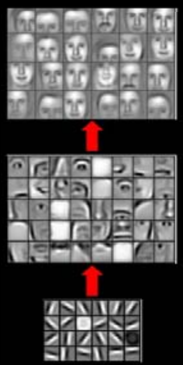
• Deep Learning and Self-Taught learning: Lets learn rather than manually design our features.

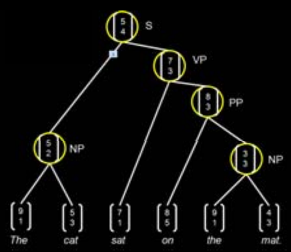
• Discover the fundamental computational principles that underlie perception?

• Sparse coding and deep versions very successful on vision and audio tasks. Other variants for learning recursive representations.

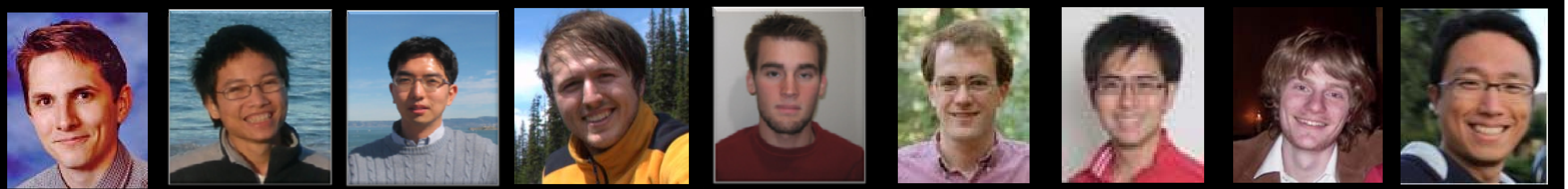
• To get this to work for yourself, see online tutorial:
http://deeplearning.stanford.edu/wiki or go/brain






Stanford



Adam Coates | Quoc Le | Honglak Lee | Andrew Saxe | Andrew Maas | Chris Manning | Jiquan Ngiam | Richard Socher | Will Zou

Google

Kai Chen | Greg Corrado | Jeff Dean | Matthieu Devin | Andrea Frome | Rajat Monga | Marc'Aurelio Ranzato | Paul Tucker | Kay Le

Andrew Ng

# Advanced Topics

Andrew Ng
Stanford University & Google

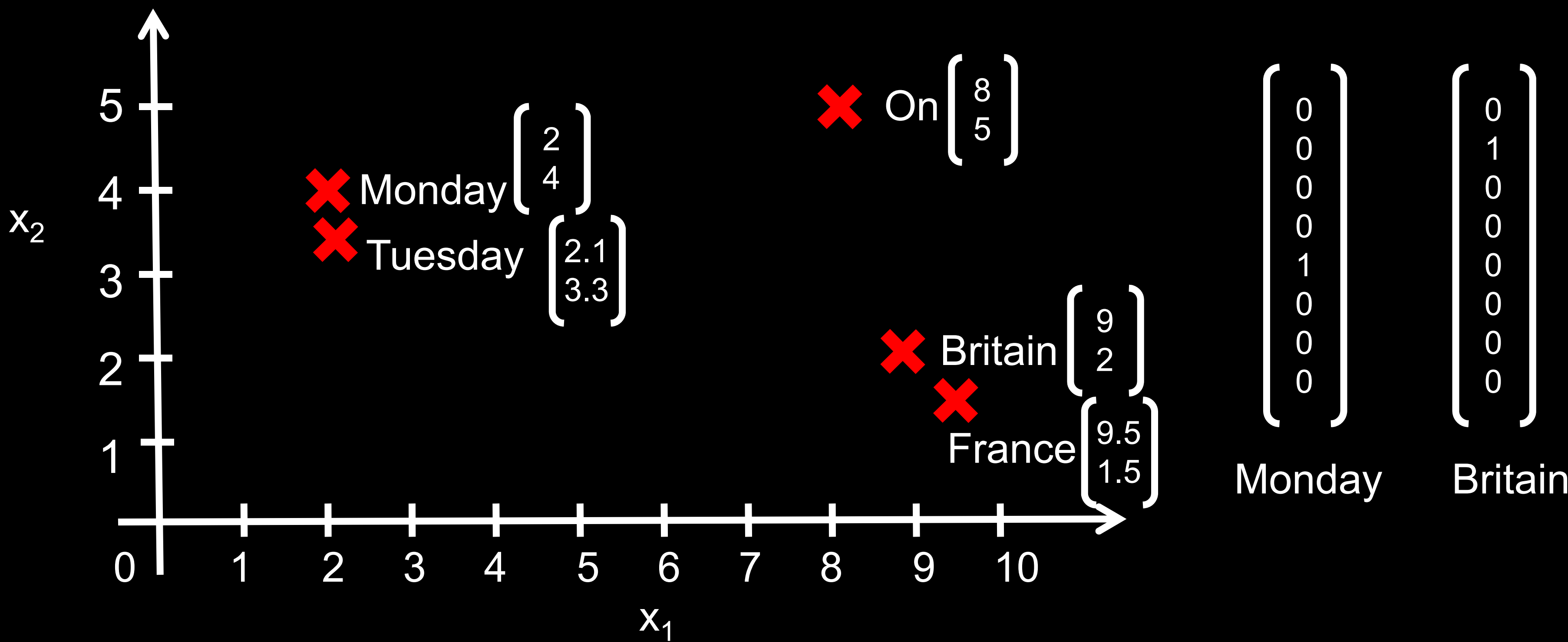
# Language:
# Learning Recursive Representations

# Feature representations of words

Imagine taking each word, and computing an n-dimensional feature vector for it.

[Distributional representations, or Bengio et al., 2003, Collobert & Weston, 2008.]

2-d embedding example below, but in practice use ~100-d embeddings.

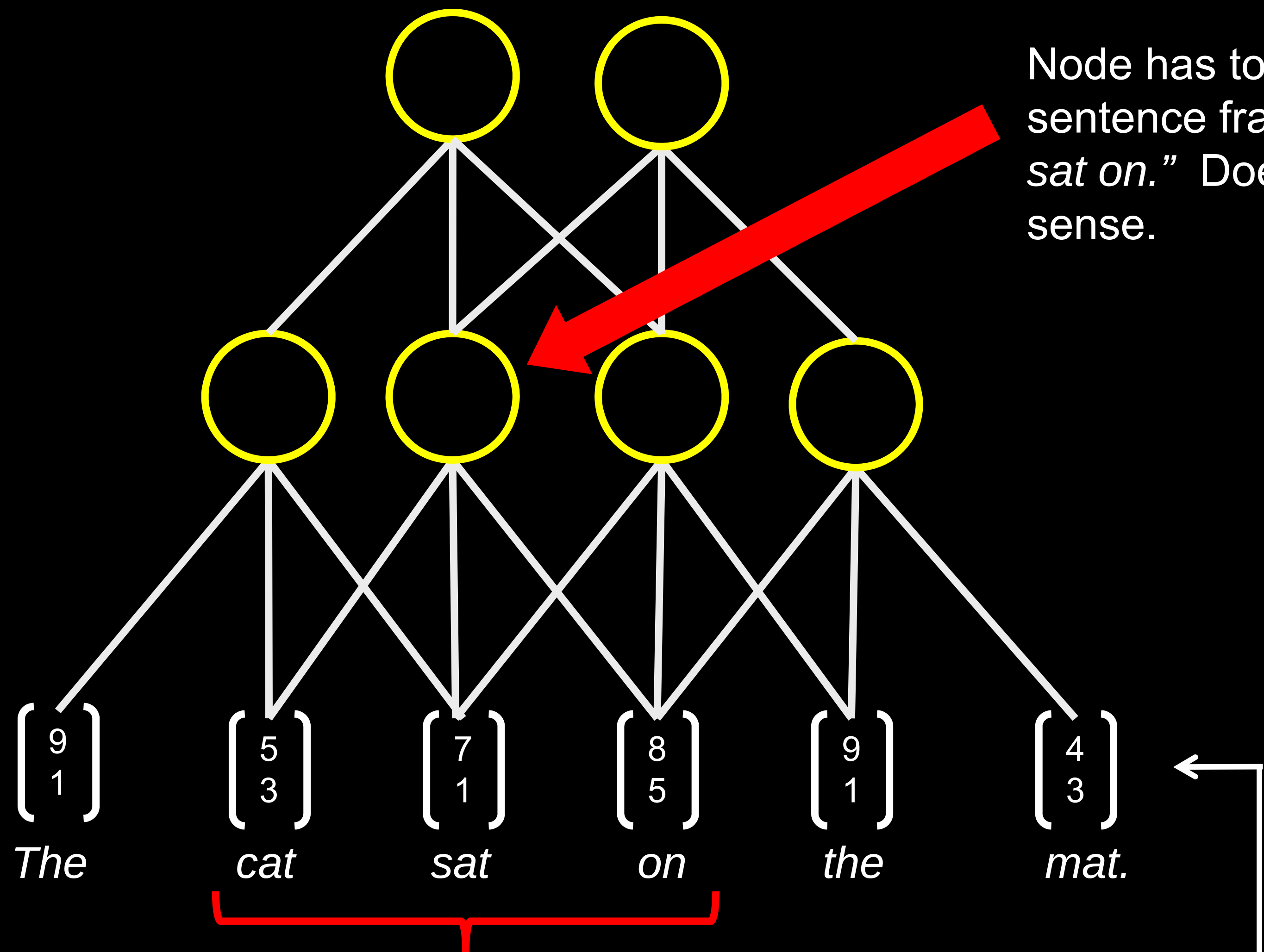



*On Monday, Britain ….*

Representation: $\begin{bmatrix} 8 \\ 5 \end{bmatrix}$ $\begin{bmatrix} 2 \\ 4 \end{bmatrix}$ $\begin{bmatrix} 9 \\ 2 \end{bmatrix}$

Andrew Ng

# "Generic" hierarchy on text doesn't make sense

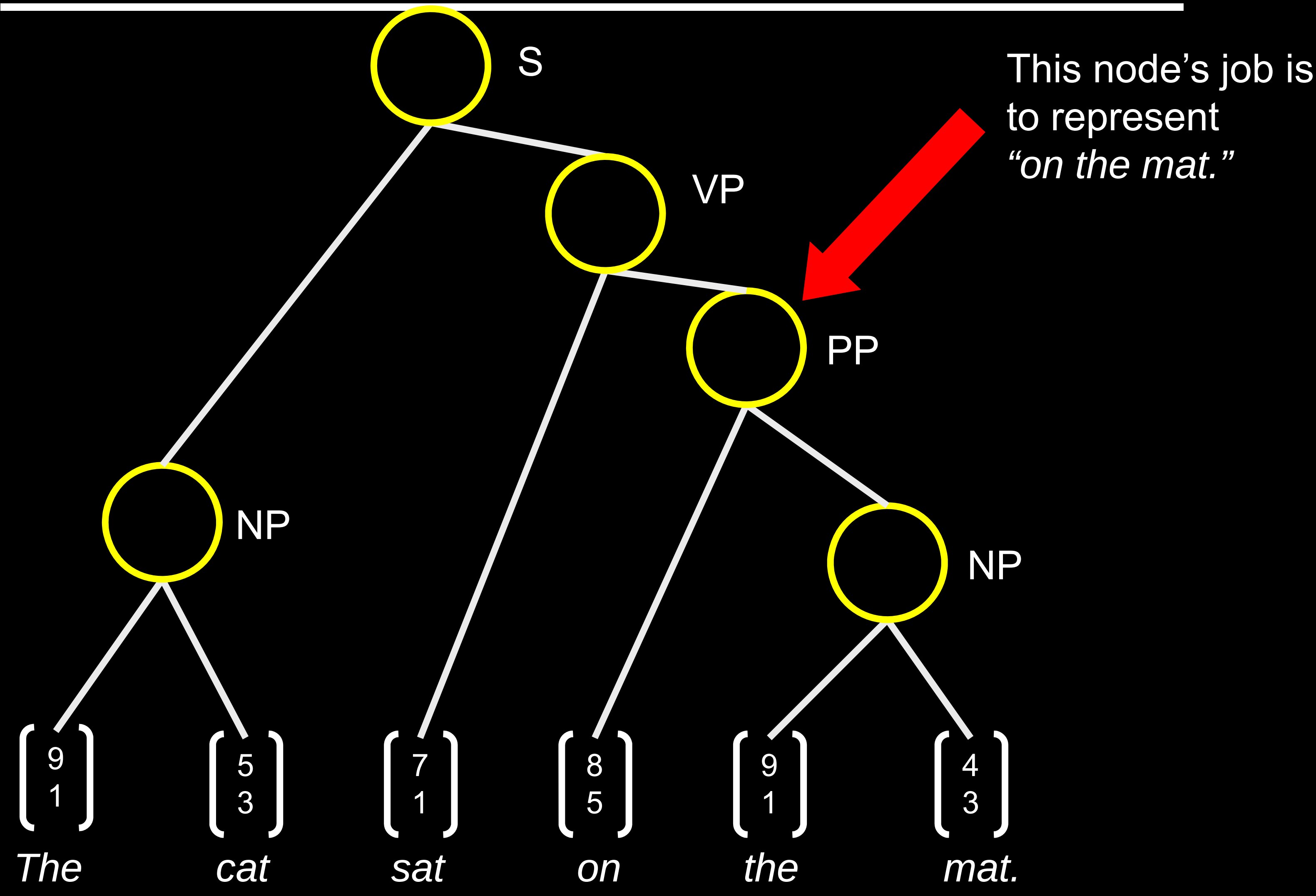



Node has to represent sentence fragment *"cat sat on."* Doesn't make sense.

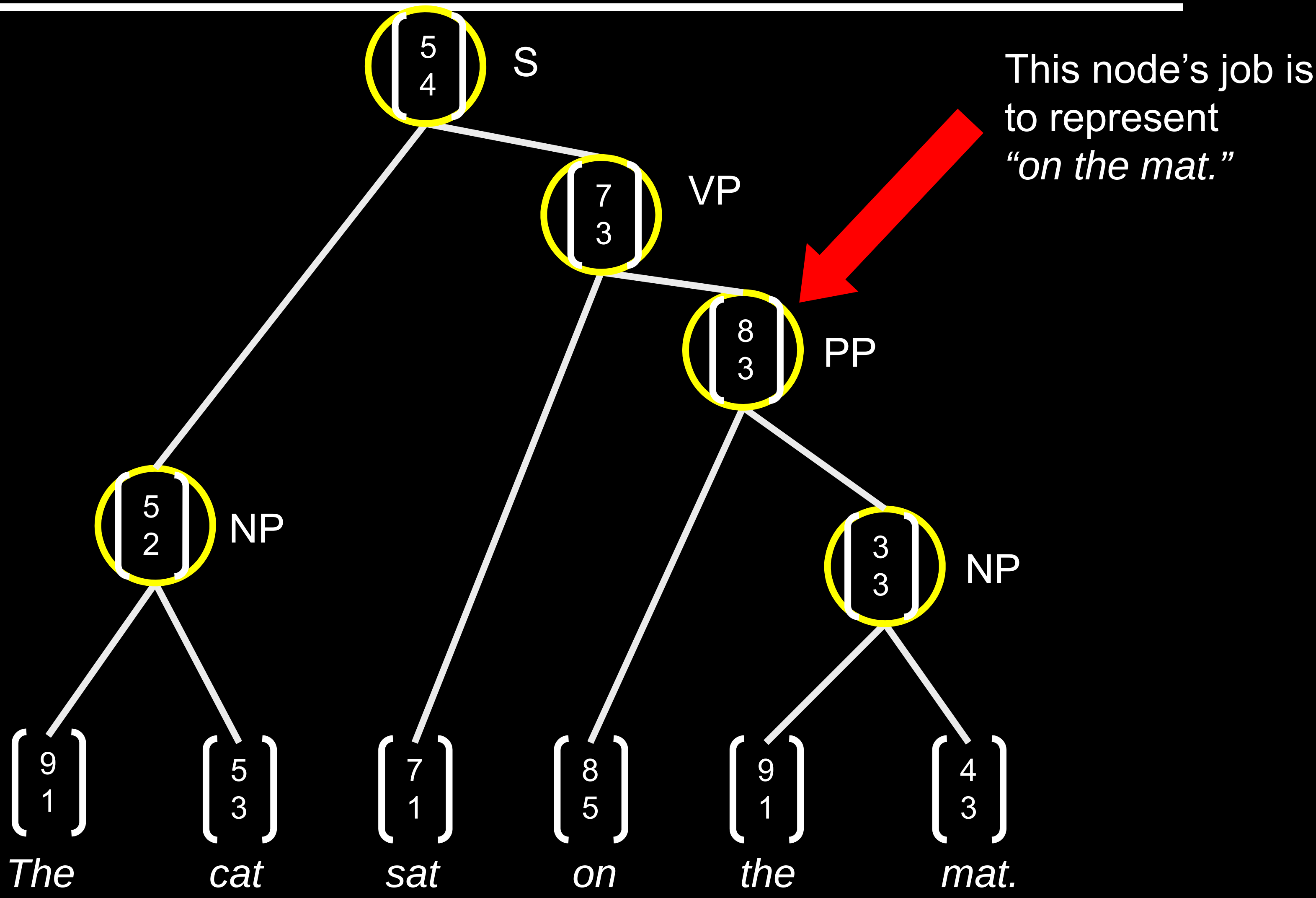
Feature representation for words

Andrew Ng

# What we want (illustration)

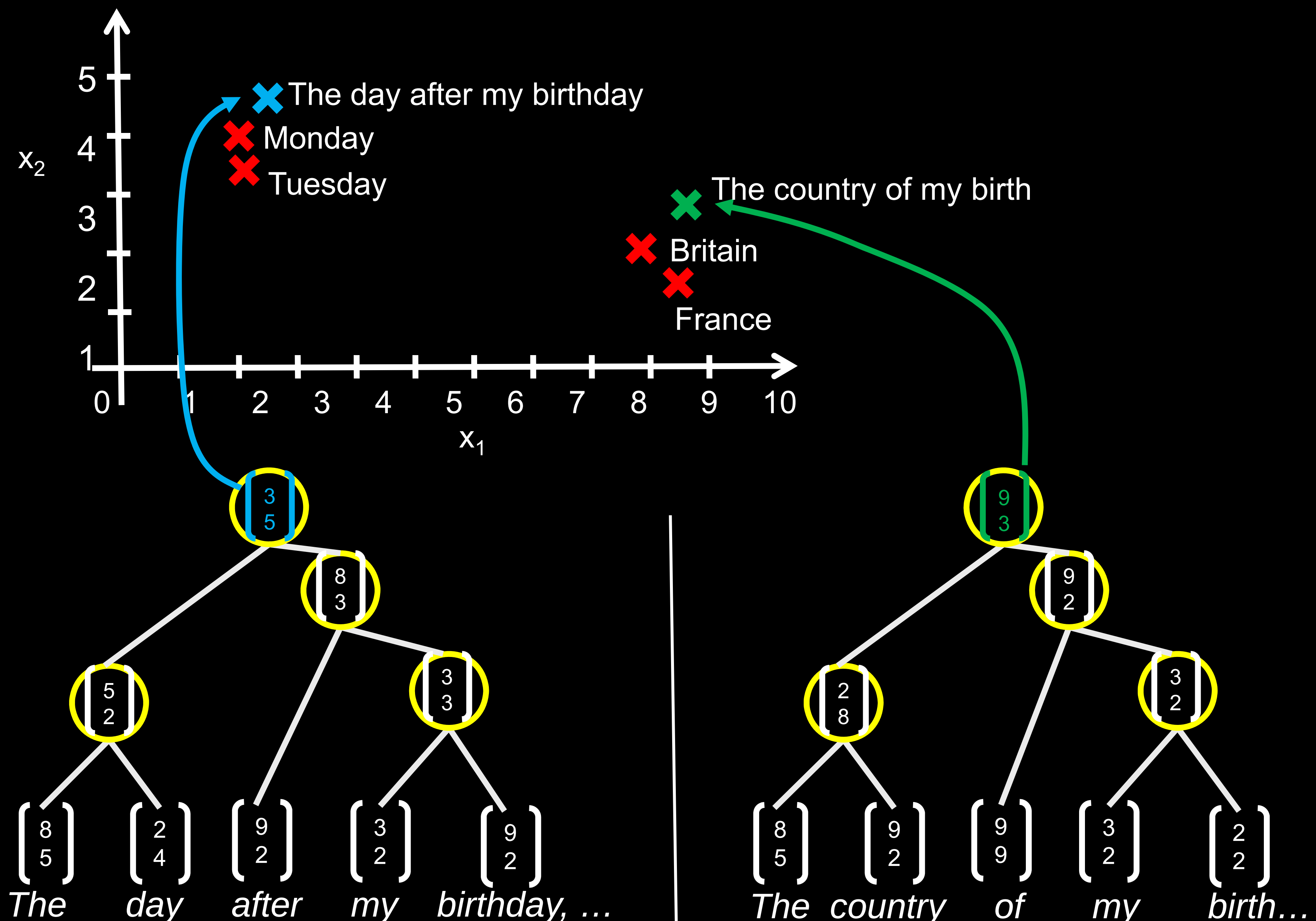


# What we want (illustration)

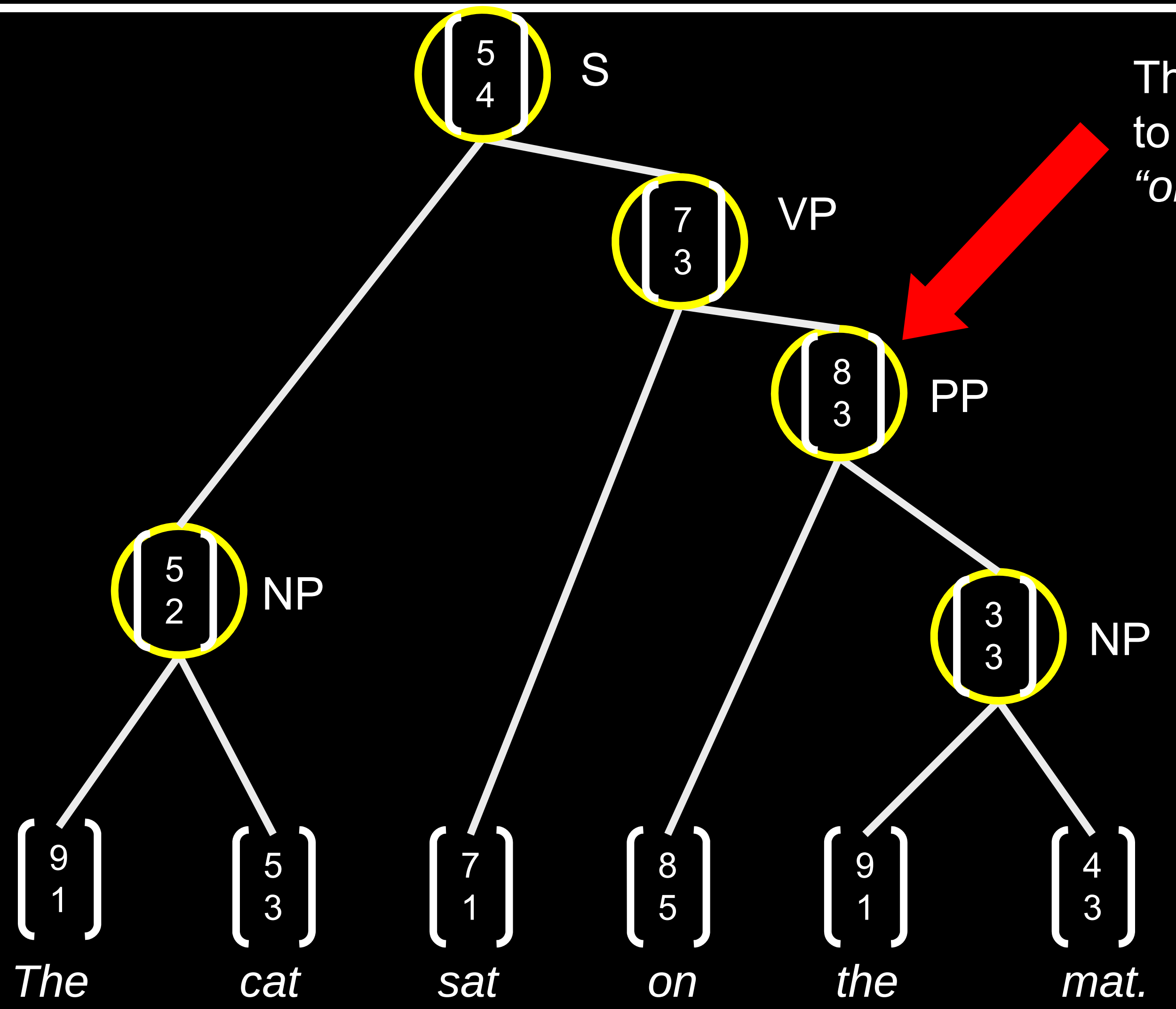



Andrew Ng

# What we want (illustration)



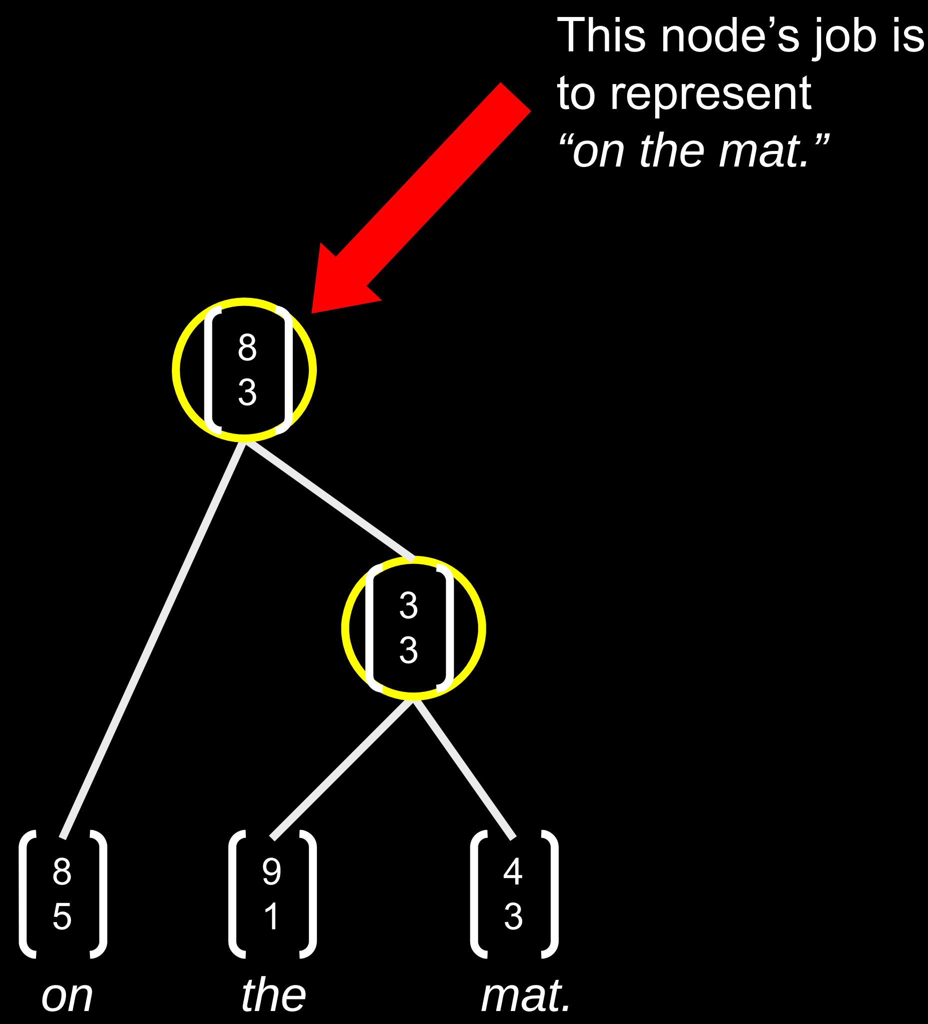
# Learning recursive representations



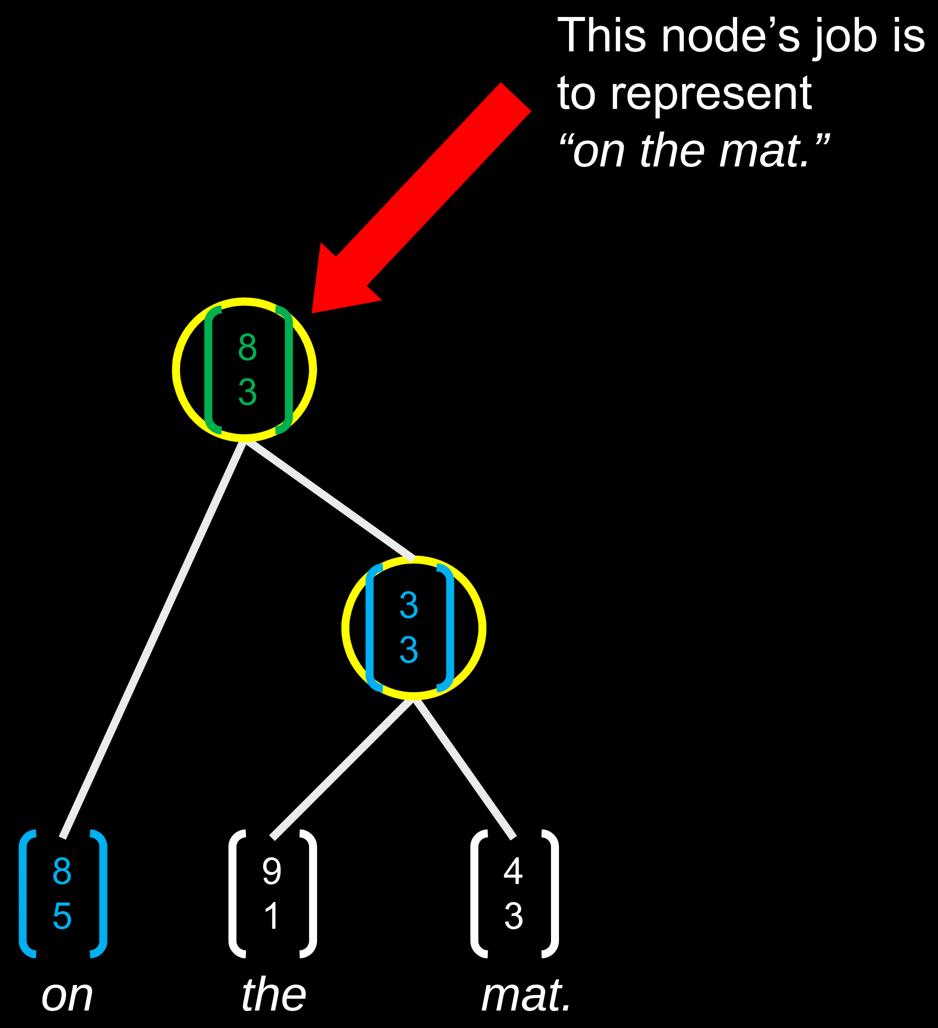

This node's job is to represent *"on the mat."*

Andrew Ng

# Learning recursive representations



# Learning recursive representations



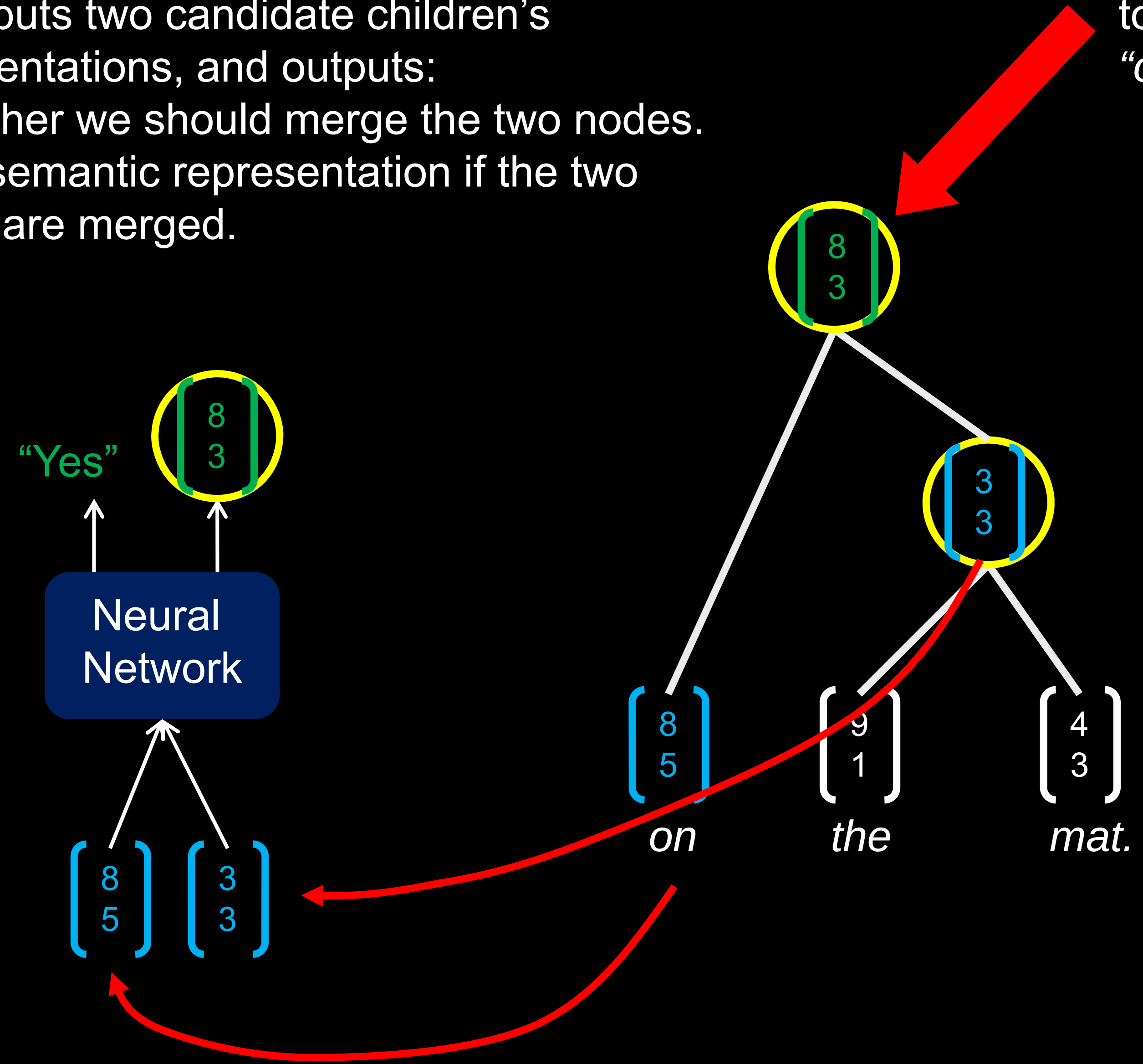

Andrew Ng

# Learning recursive representations

Basic computational unit: Neural Network that inputs two candidate children's representations, and outputs:

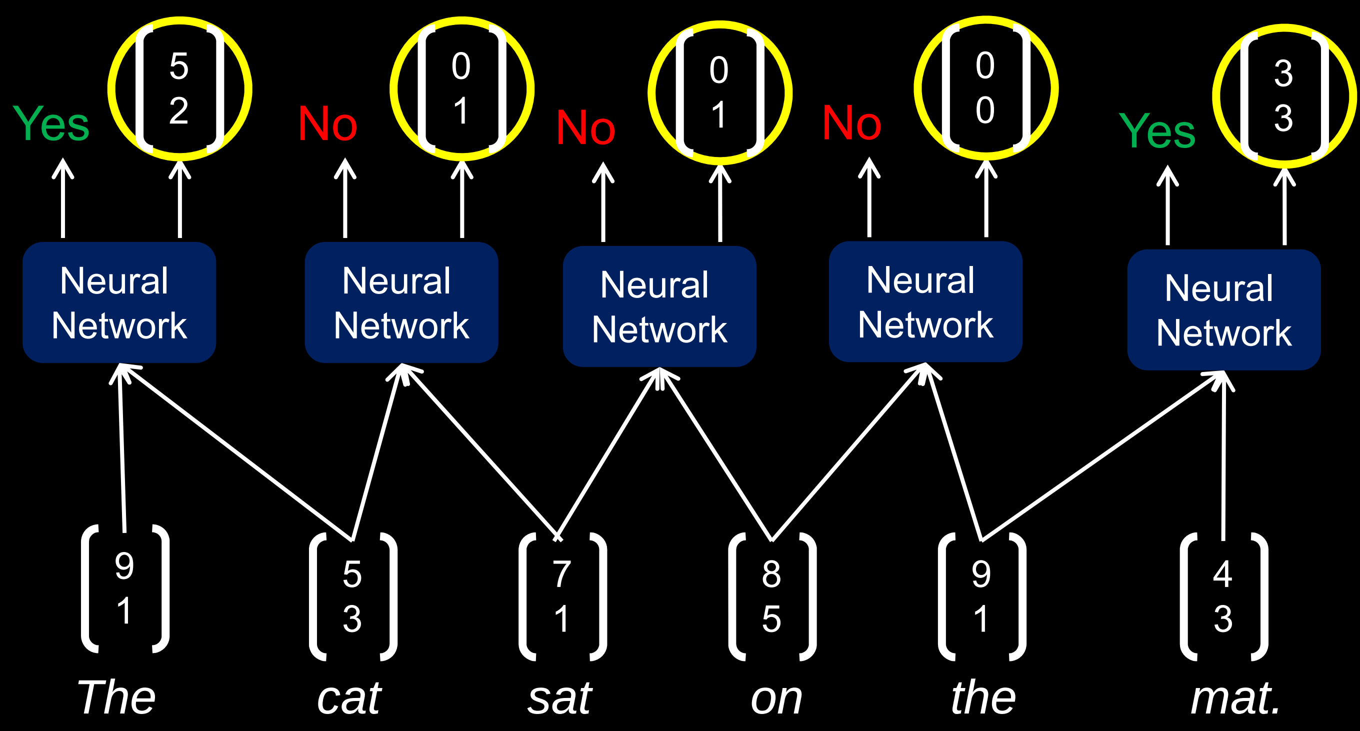
- Whether we should merge the two nodes.
- The semantic representation if the two nodes are merged.

This node's job is to represent *"on the mat."*

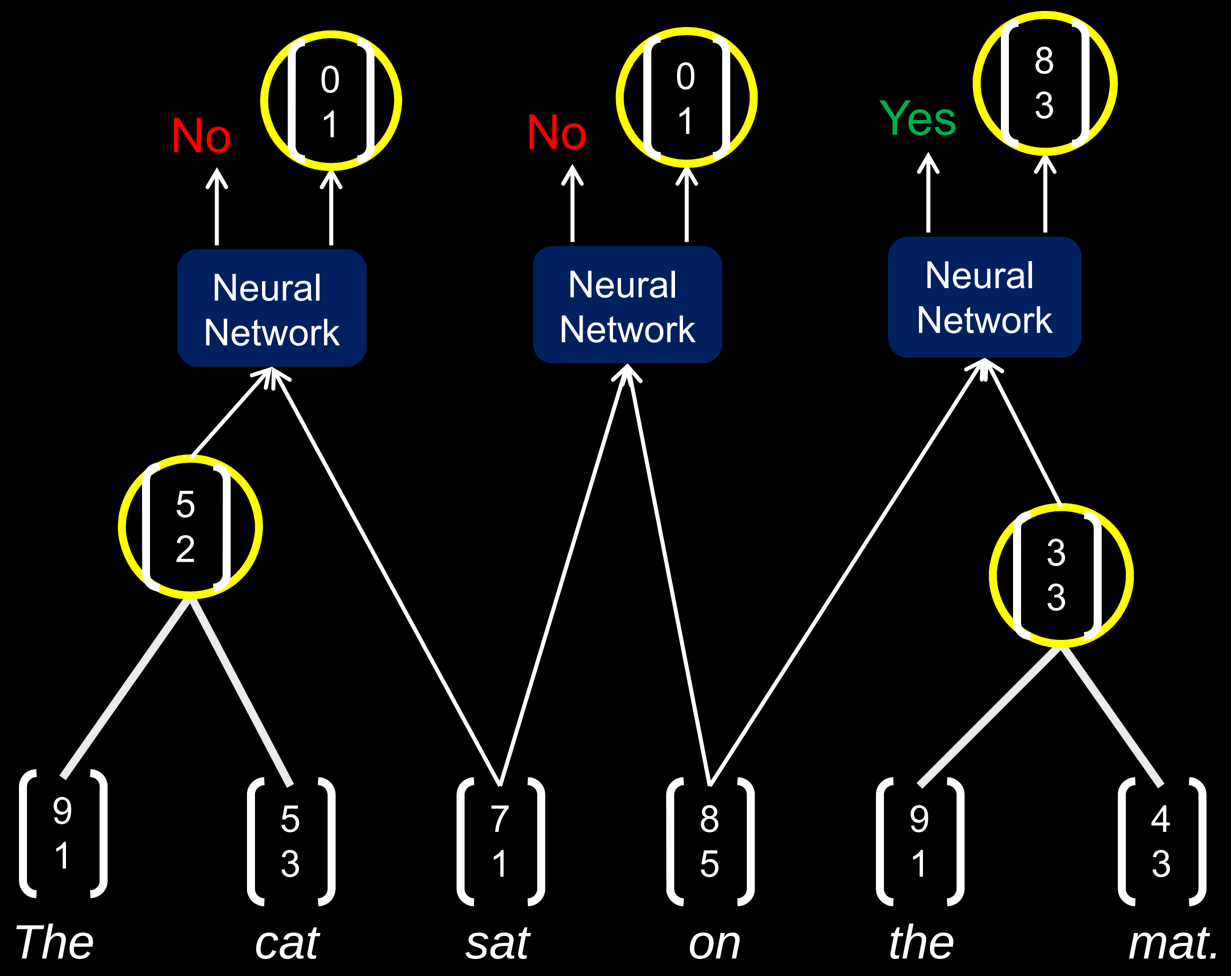



Andrew Ng

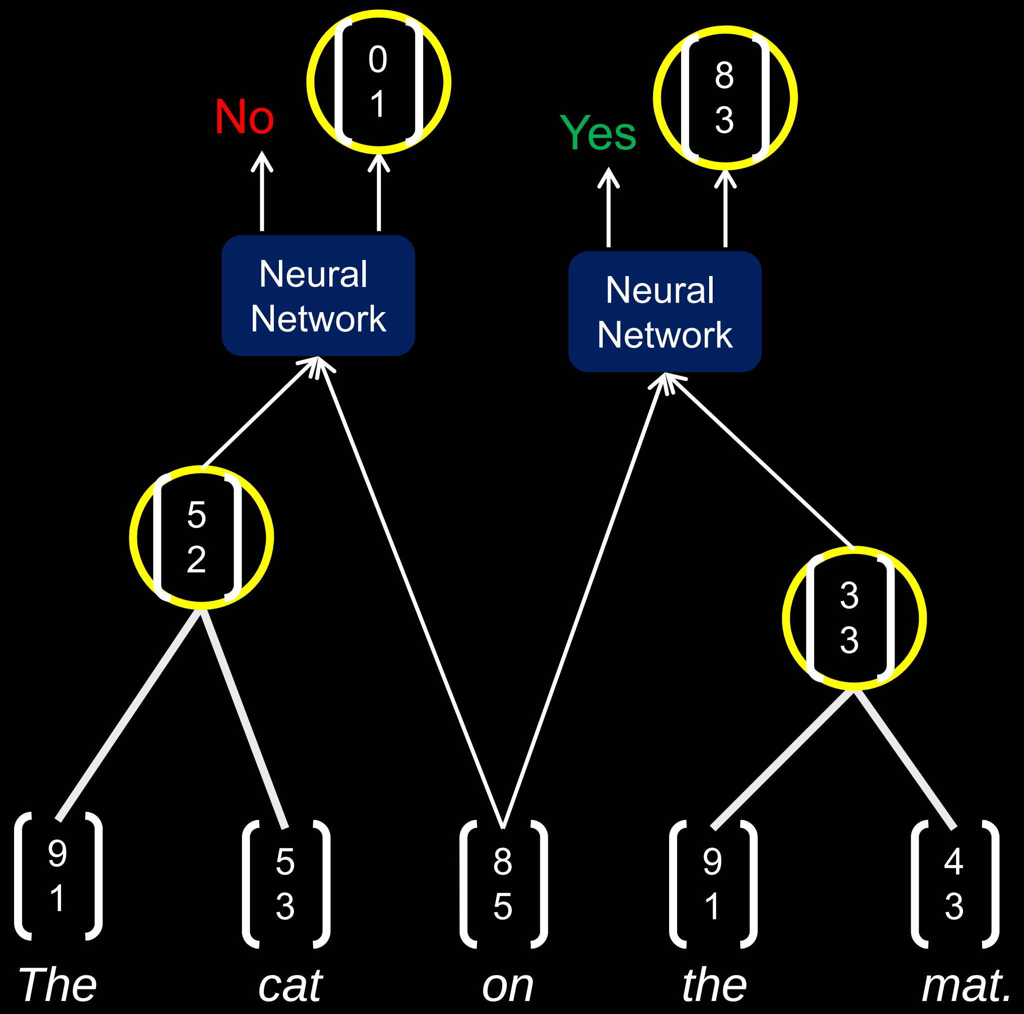
# Parsing a sentence



# Parsing a sentence

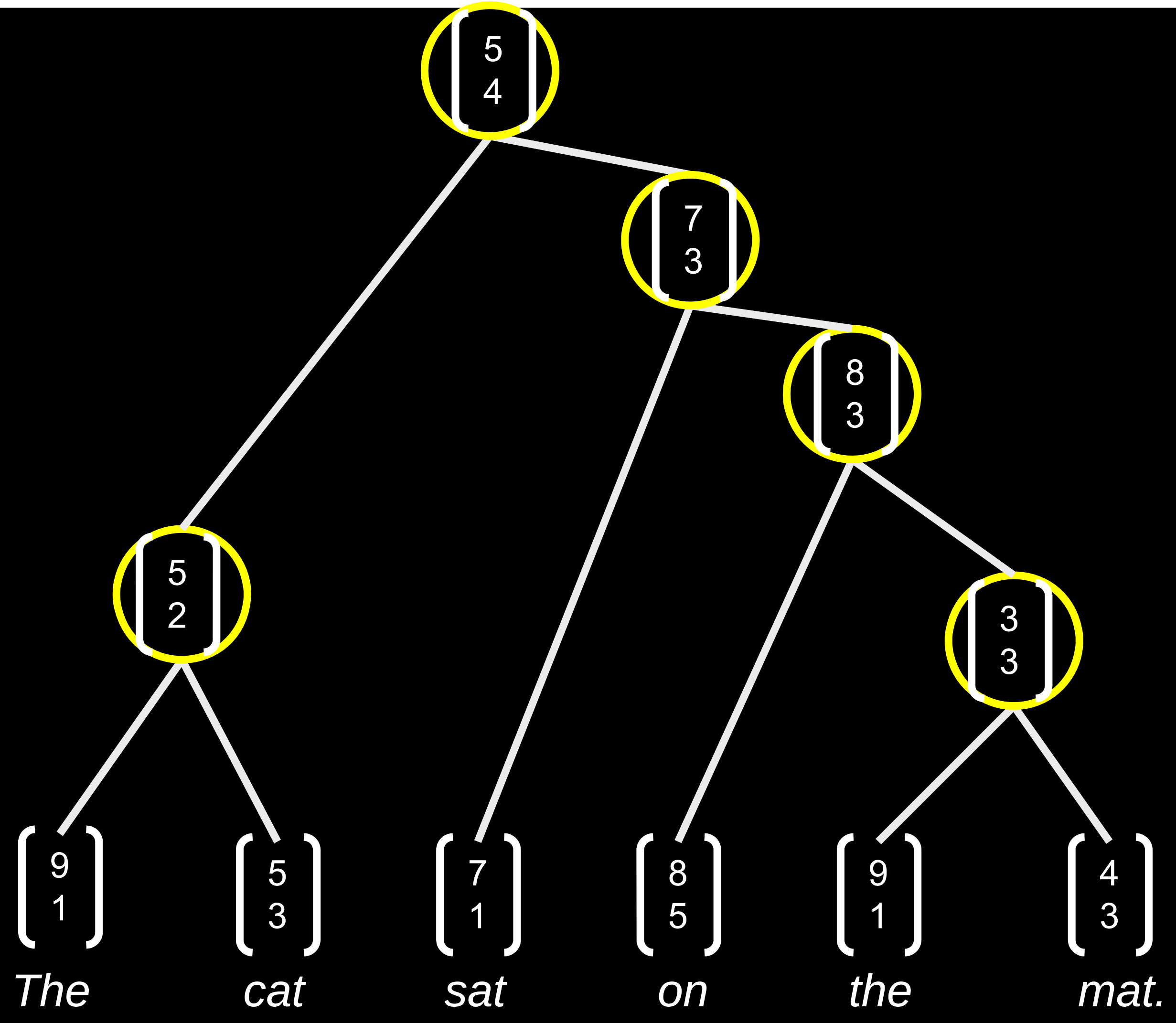


# Parsing a sentence




[Socher, Manning & Ng]

# Parsing a sentence



# Finding Similar Sentences

- Each sentence has a feature vector representation.
- Pick a sentence ("center sentence") and list nearest neighbor sentences.
- Often either semantically or syntactically similar. (Digits all mapped to 2.)

| Similarities | Center Sentence | Nearest Neighbor Sentences (most similar feature vector) |
|---|---|---|
| Bad News | Both took further hits yesterday | 1. We 're in for a lot of turbulence ...<br>2. BSN currently has 2.2 million common shares outstanding<br>3. This is panic buying<br>4. We have a couple or three tough weeks coming |
| Something said | I had calls all night long from the States, he said | 1. Our intent is to promote the best alternative, he says<br>2. We have sufficient cash flow to handle that, he said<br>3. Currently, average pay for machinists is 22.22 an hour, Boeing said<br>4. Profit from trading for its own account dropped, the securities firm said |
| Gains and good news | Fujisawa gained 22 to 2,222 | 1. Mochida advanced 22 to 2,222<br>2. Commerzbank gained 2 to 222.2<br>3. Paris loved her at first sight<br>4. Profits improved across Hess's businesses |
| Unknown words which are cities | Columbia , S.C | 1. Greenville , Miss<br>2. UNK , Md<br>3. UNK , Miss<br>4. UNK , Calif |

# Finding Similar Sentences

| Similarities | Center Sentence | Nearest Neighbor Sentences in Embedding Space |
|---|---|---|
| Bad News | Both took further hits yesterday | 1. We 're in for a lot of turbulence ...<br>2. BSN currently has 2.2 million common shares outstanding<br>3. This is panic buying<br>4. We have a couple or three tough weeks coming |
| Something said | I had calls all night long from the States, he said | 1. Our intent is to promote the best alternative, he says<br>2. We have sufficient cash flow to handle that, he said<br>3. Currently, average pay for machinists is 22.22 an hour, Boeing said<br>4. Profit from trading for its own account dropped, the securities firm said |
| Gains and good news | Fujisawa gained 22 to 2,222 | 1. Mochida advanced 22 to 2,222<br>2. Commerzbank gained 2 to 222.2<br>3. Paris loved her at first sight<br>4. Profits improved across Hess's businesses |
| Unknown words which are cities | Columbia , S.C | 1. Greenville , Miss<br>2. UNK , Md<br>3. UNK , Miss<br>4. UNK , Calif |

# Finding Similar Sentences

| Similarities | Center Sentence | Nearest Neighbor Sentences (most similar feature vector) |
|---|---|---|
| Declining to comment = not disclosing | Hess declined to comment | 1. PaineWebber declined to comment<br>2. Phoenix declined to comment<br>3. Campeau declined to comment<br>4. Coastal wouldn't disclose the terms |
| Large changes in sales or revenue | Sales grew almost 2 % to 222.2 million from 222.2 million | 1. Sales surged 22 % to 222.22 billion yen from 222.22 billion<br>2. Revenue fell 2 % to 2.22 billion from 2.22 billion<br>3. Sales rose more than 2 % to 22.2 million from 22.2 million<br>4. Volume was 222.2 million shares , more than triple recent levels |
| Negation of different types | There's nothing unusual about business groups pushing for more government spending | 1. We don't think at this point anything needs to be said<br>2. It therefore makes no sense for each market to adopt different circuit breakers<br>3. You can't say the same with black and white<br>4. I don't think anyone left the place UNK UNK |
| People in bad situations | We were lucky | 1. It was chaotic<br>2. We were wrong<br>3. People had died<br>4. They still are |

# Experiments

- No linguistic features. Train only using the structure and words of WSJ training trees, and word embeddings from (Collobert & Weston, 2008).
- Parser evaluation dataset: Wall Street Journal (standard splits for training and development testing).

| Method | Unlabeled F1 |
|---|---|
| Greedy Recursive Neural Network (RNN) | 76.55 |
| Greedy, context-sensitive RNN | 83.36 |
| Greedy, context-sensitive RNN + category classifier | 87.05 |
| Left Corner PCFG, (Manning and Carpenter, '97) | 90.64 |
| CKY, context-sensitive, RNN + category classifier (our work) | 92.06 |
| Current Stanford Parser, (Klein and Manning, '03) | 93.98 |

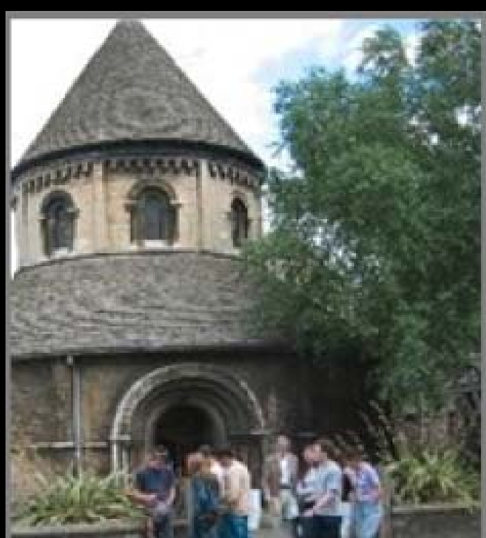
# Application: Paraphrase Detection

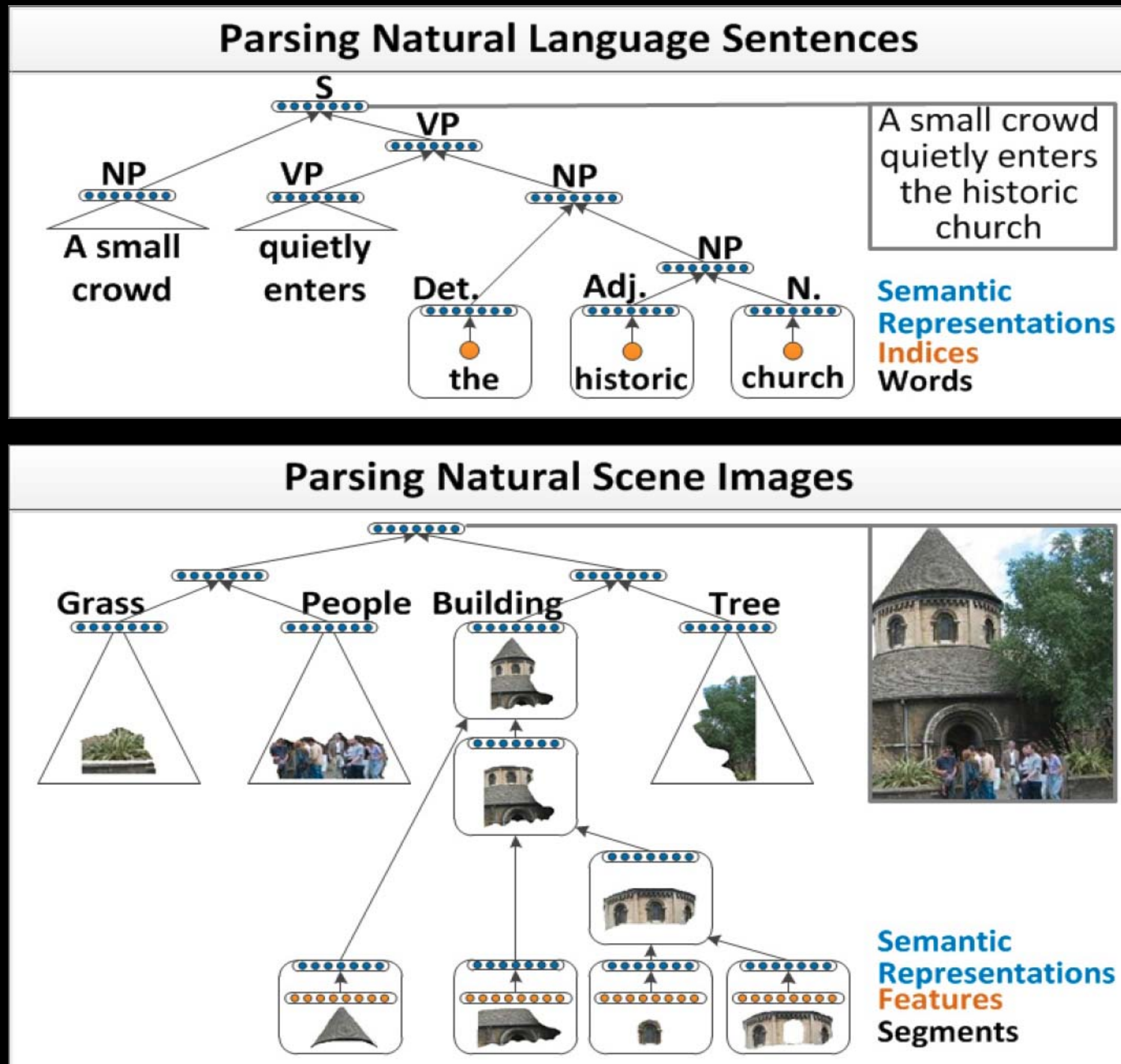
- Task: Decide whether or not two sentences are paraphrases of each other. (MSR Paraphrase Corpus)

| Method | F1 |
|---|---|
| Baseline | 79.9 |
| Rus et al., (2008) | 80.5 |
| Mihalcea et al., (2006) | 81.3 |
| Islam et al. (2007) | 81.3 |
| Qiu et al. (2006) | 81.6 |
| Fernando & Stevenson (2008) (WordNet based features) | 82.4 |
| Das et al. (2009) | 82.7 |
| Wan et al (2006) (many features: POS, parsing, BLEU, etc.) | 83.0 |
| **Stanford Feature Learning** | **83.4** |

# Parsing sentences and parsing images

A small crowd quietly enters the historic church.

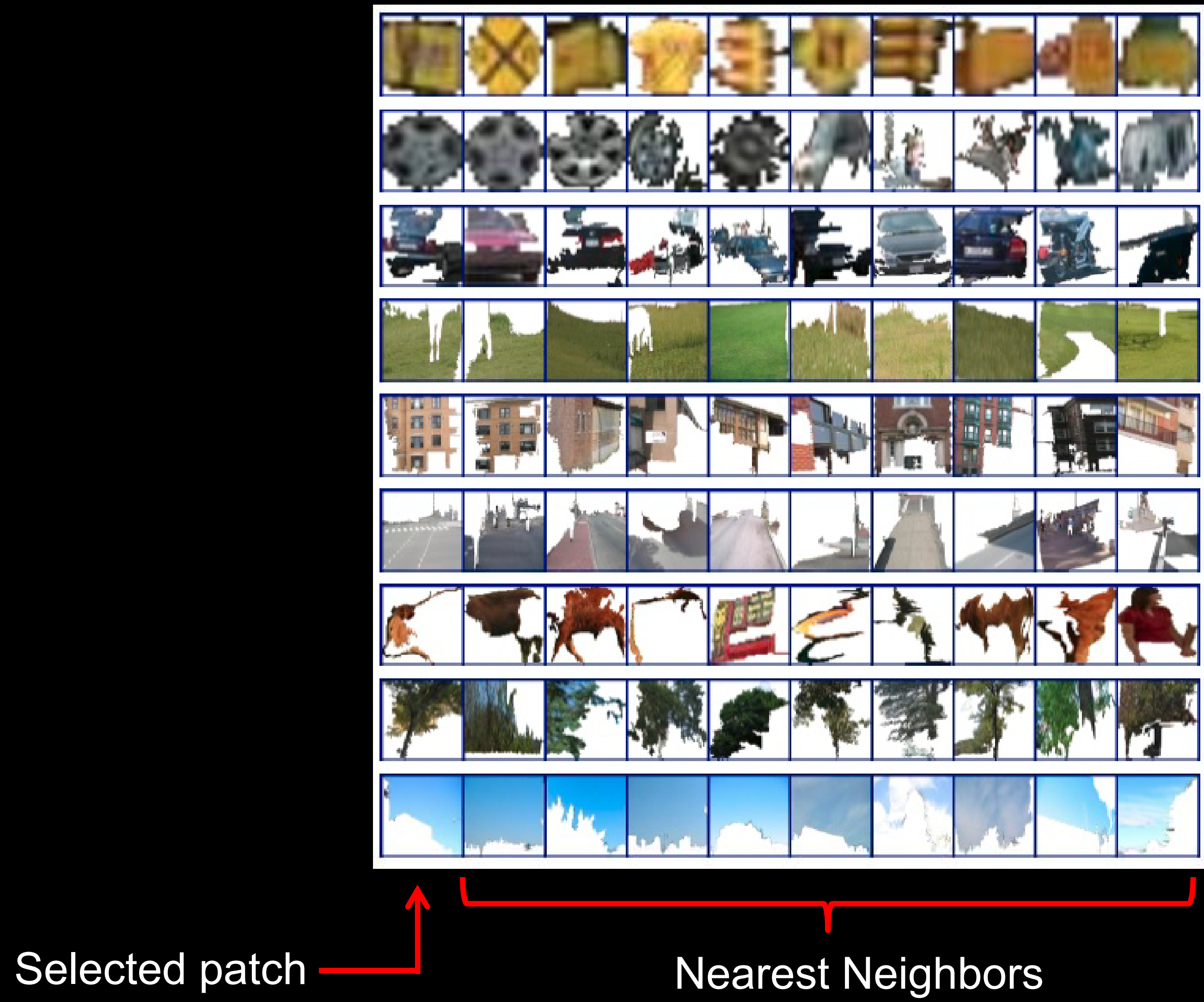





Each node in the hierarchy has a "feature vector" representation.

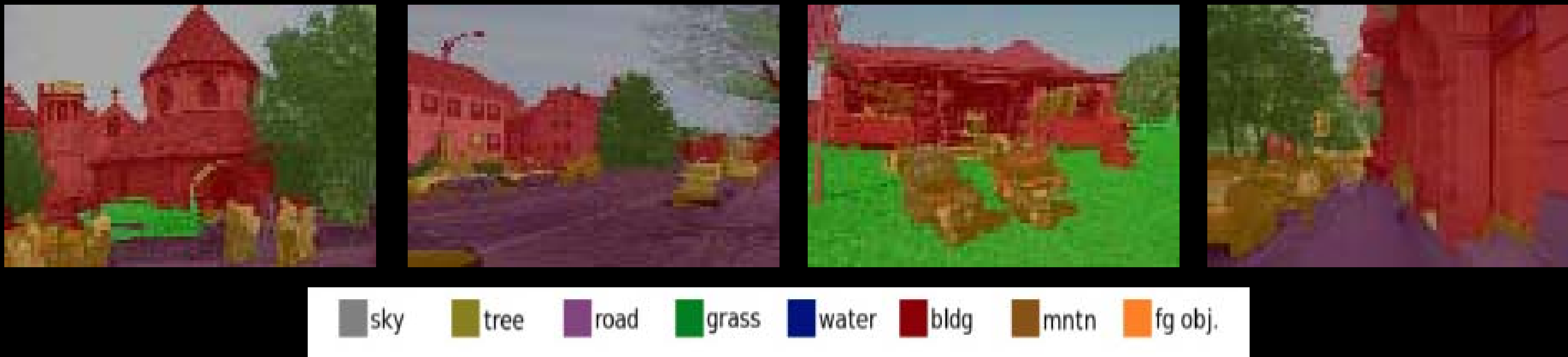
# Nearest neighbor examples for image patches

- Each node (e.g., set of merged superpixels) in the hierarchy has a feature vector.
- Select a node ("center patch") and list nearest neighbor nodes.
- I.e., what image patches/superpixels get mapped to similar features?




Andrew Ng

# Multi-class segmentation (Stanford background dataset)




| Method | Accuracy |
|---|---|
| Pixel CRF (Gould et al., ICCV 2009) | 74.3 |
| Classifier on superpixel features | 75.9 |
| Region-based energy (Gould et al., ICCV 2009) | 76.4 |
| Local labelling (Tighe & Lazebnik, ECCV 2010) | 76.9 |
| Superpixel MRF (Tighe & Lazebnik, ECCV 2010) | 77.5 |
| Simultaneous MRF (Tighe & Lazebnik, ECCV 2010) | 77.5 |
| **Stanford Feature learning (our method)** | **78.1** |

# Multi-class Segmentation MSRC dataset: 21 Classes

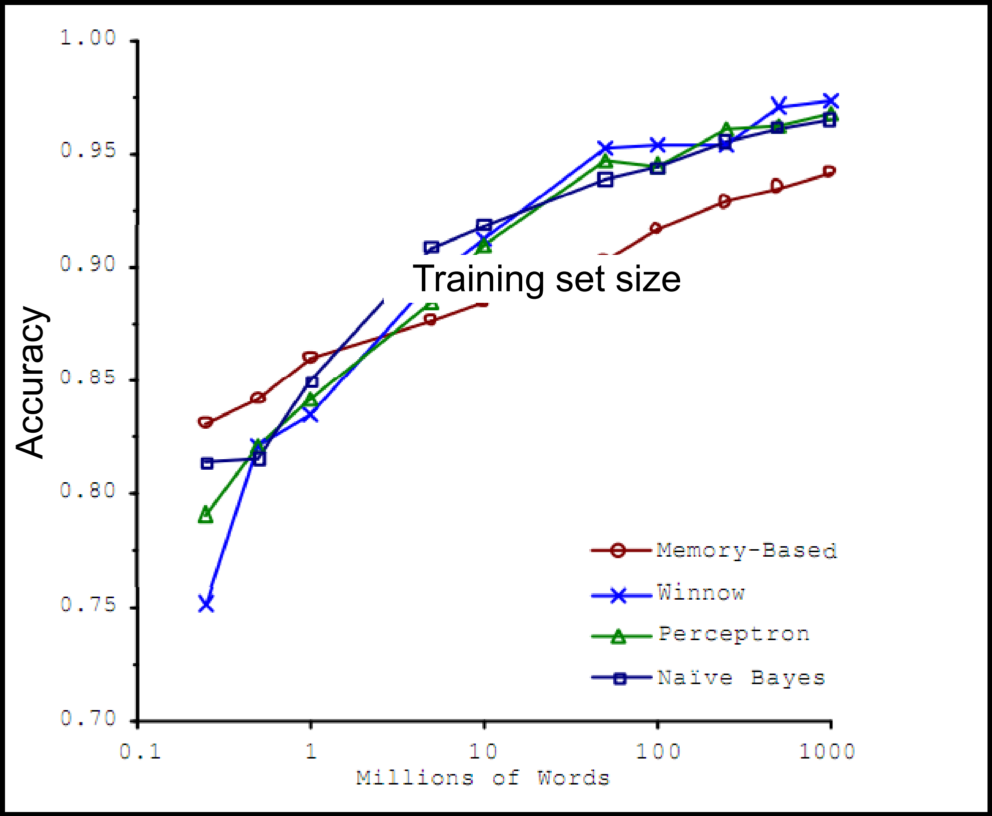

| Methods | Accuracy |
|---|---|
| TextonBoost (Shotton et al., ECCV 2006) | 72.2 |
| Framework over mean-shift patches (Yang et al., CVPR 2007) | 75.1 |
| Pixel CRF (Gould et al., ICCV 2009) | 75.3 |
| Region-based energy (Gould et al., IJCV 2008) | 76.5 |
| **Stanford Feature learning (out method)** | **76.7** |

# Analysis of feature learning algorithms

Andrew Coates    Honglak Lee

# Supervised Learning

- Choices of learning algorithm:
  - Memory based
  - Winnow
  - Perceptron
  - Naïve Bayes
  - SVM
  - ….
- What matters the most?

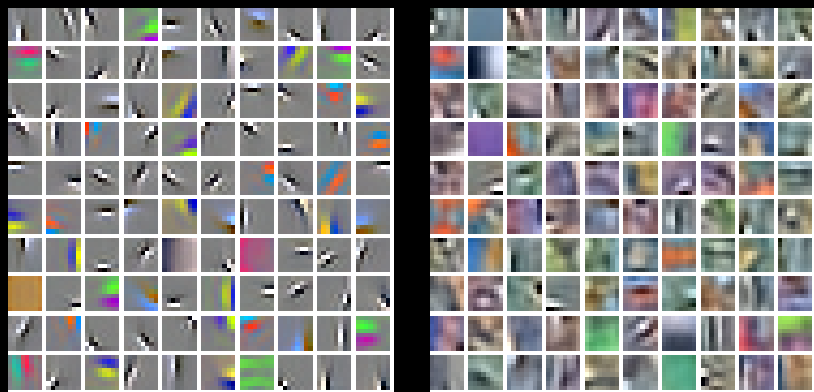



[Banko & Brill, 2001]

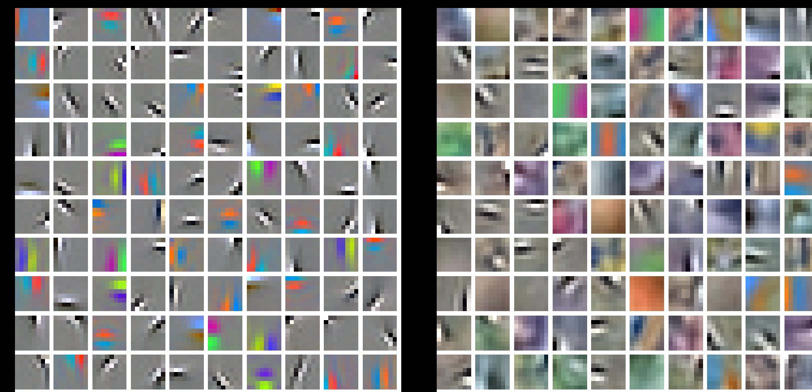
"It's not who has the best algorithm that wins.
It's who has the most data."

# Receptive fields learned by several algorithms

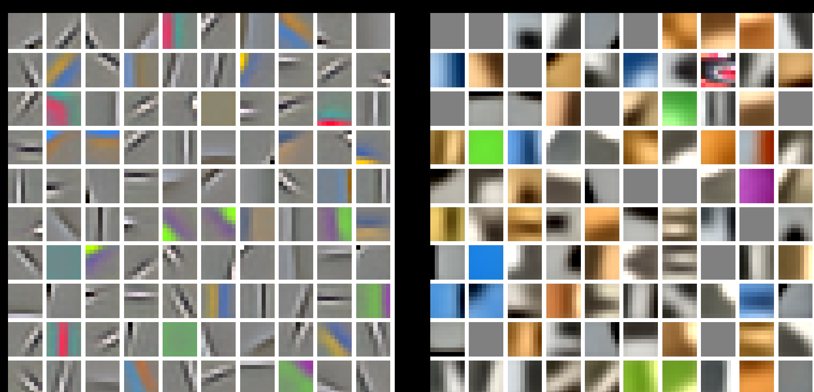
The primary goal of unsupervised feature learning: To discover Gabor functions.



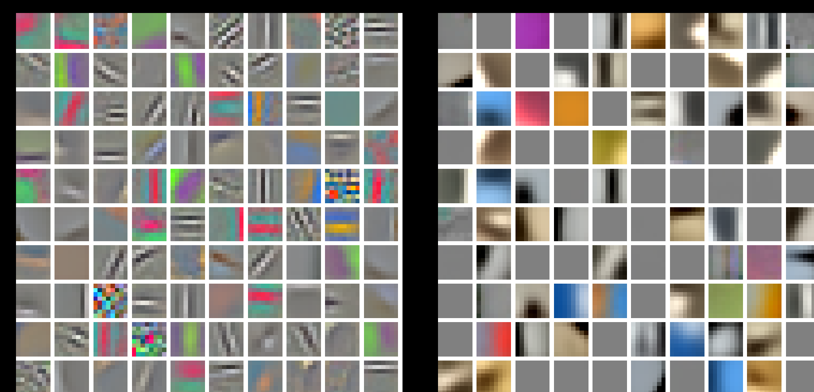
Sparse auto-encoder (with and without whitening)



Sparse RBM (with and without whitening)



K-means (with and without whitening)



Gaussian mixture model (with and without whitening)

# Analysis of single-layer networks

- Many components in feature learning system:
  - Pre-processing steps (e.g., whitening)
  - Network architecture (depth, number of features)
  - Unsupervised training algorithm
  - Inference / feature extraction
  - Pooling strategies

- Which matters most?
  - Much emphasis on new models + new algorithms. Is this the right focus?
  - Many algorithms hindered by large number of parameters to tune.

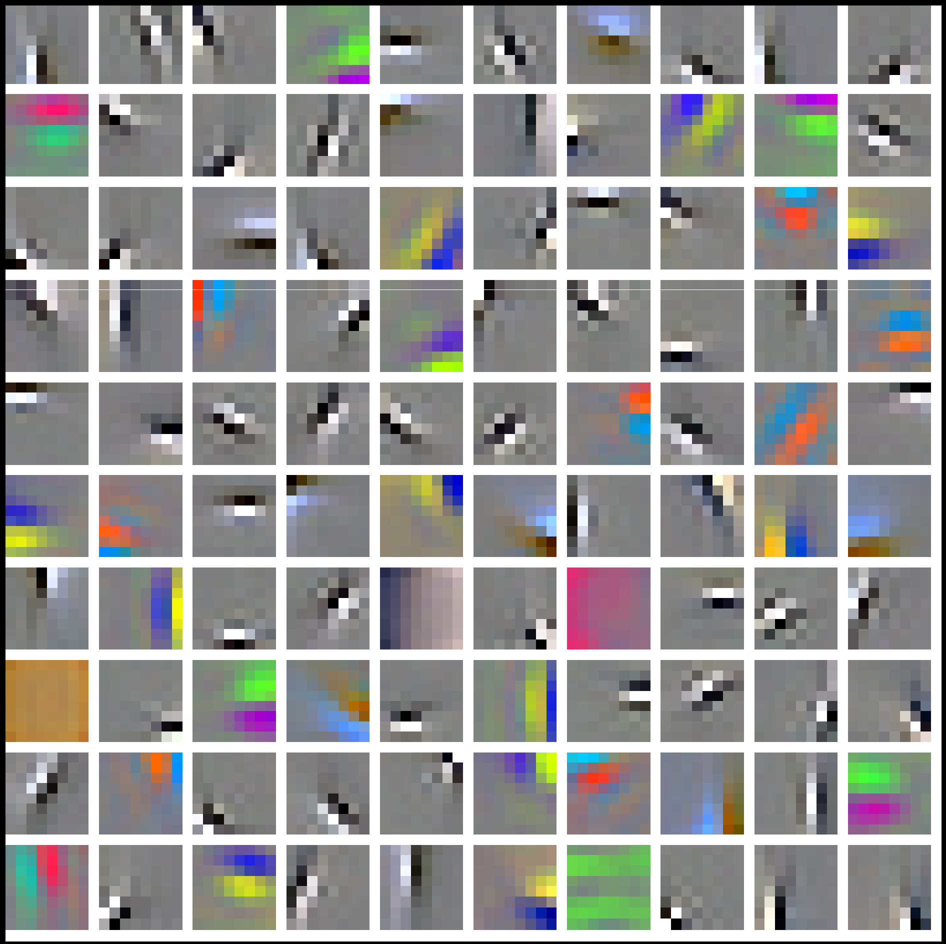
  - Simple algorithm + carefully chosen architecture = *state-of-the-art*.
  - Unsupervised learning algorithm may not be most important part.

# Unsupervised Feature Learning

- Many choices in feature learning algorithms;
  - Sparse coding, RBM, autoencoder, etc.
  - Pre-processing steps (whitening)
  - Number of features learned
  - Various hyperparameters.
- What matters the most?

# Unsupervised feature learning

Most algorithms learn Gabor-like edge detectors.

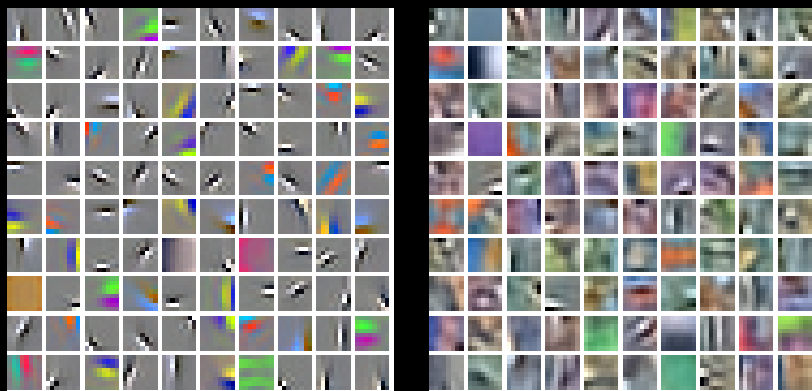


Sparse auto-encoder

# Unsupervised feature learning

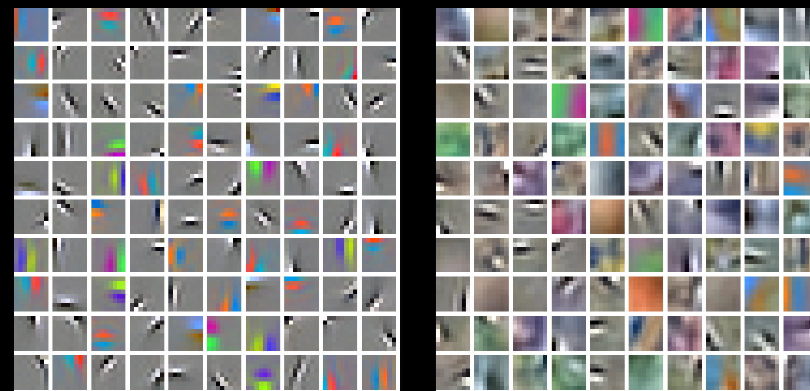
Weights learned with and without whitening.

with whitening | without whitening



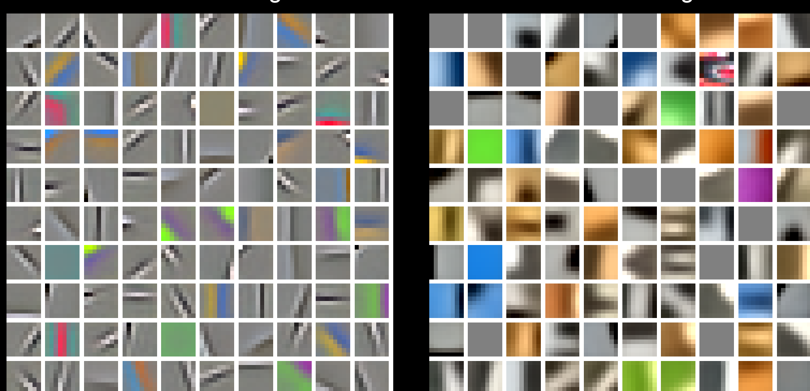
Sparse auto-encoder

with whitening | without whitening



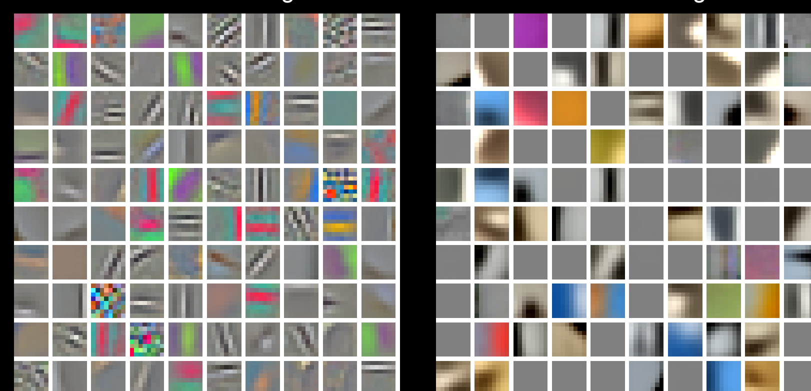
Sparse RBM

with whitening | without whitening



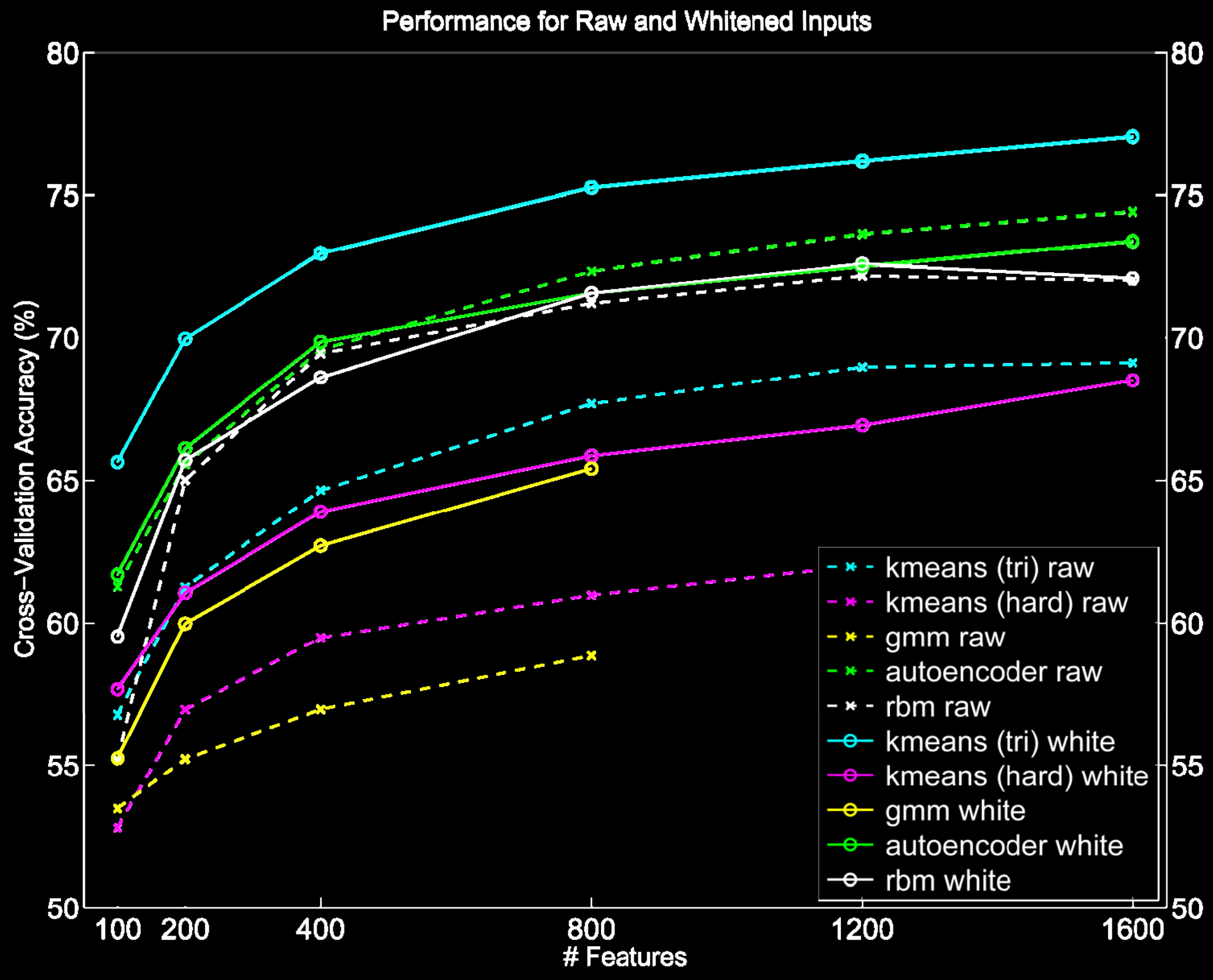
K-means

with whitening | without whitening



Gaussian mixture model

# Scaling and classification accuracy (CIFAR-10)




Andrew Ng