Unsupervised Feature Learning and Deep Learning

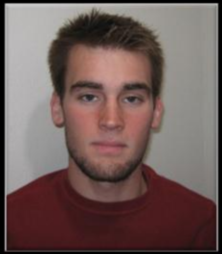
Andrew Ng

Thanks to:

Adam Coates

Quoc Le

Honglak Lee



Andrew Maas

Chris Manning

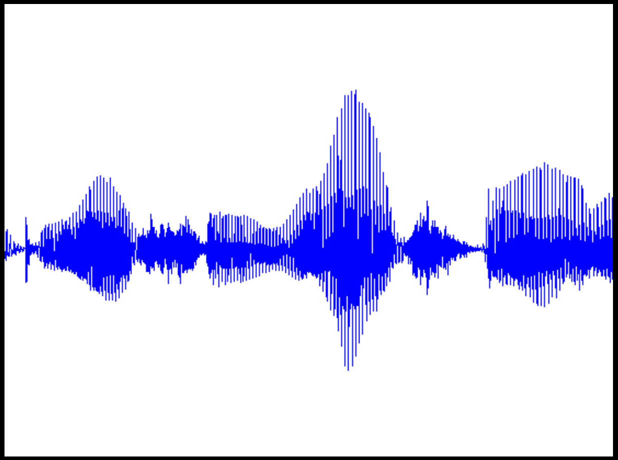
Jiquan Ngiam

Andrew Saxe

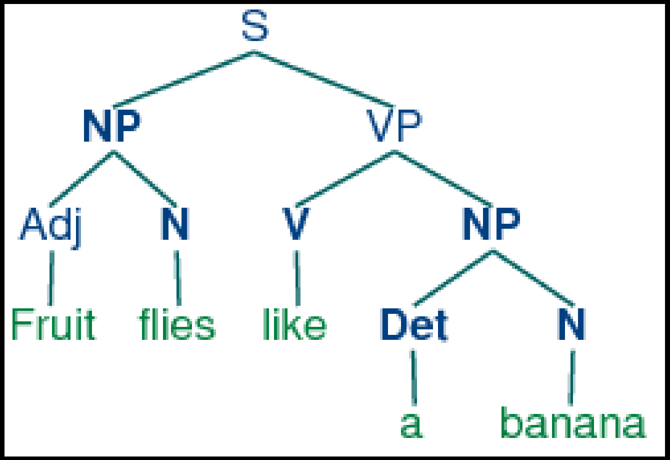
Richard Socher

Develop ideas using…

Computer vision



Audio



Text

Feature representations

Input

Learning algorithm

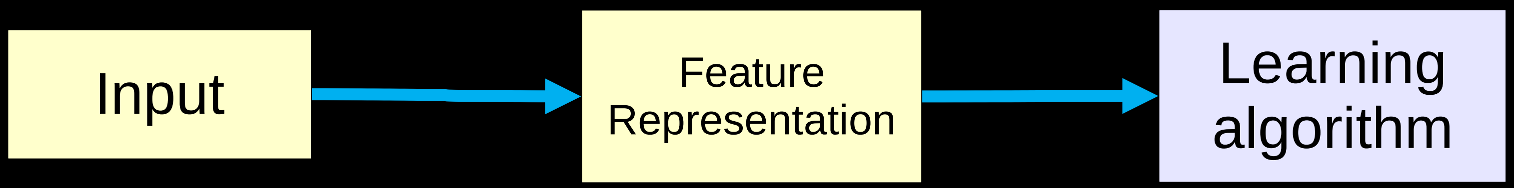
Feature representations

Input

Feature Representation → Learning algorithm

E.g., SIFT, HoG, etc.

Feature representations



How is computer perception done?

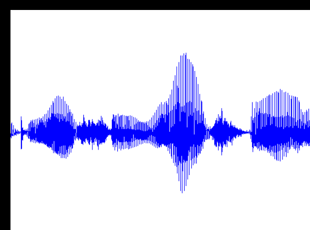
Object detection

Image

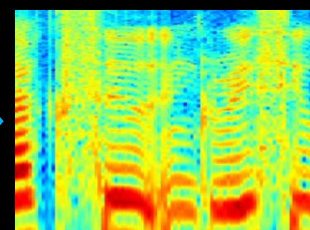
Vision features

Detection

Audio classification



Audio

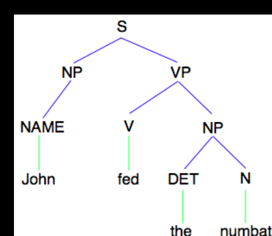


Audio features

Speaker ID

NLP

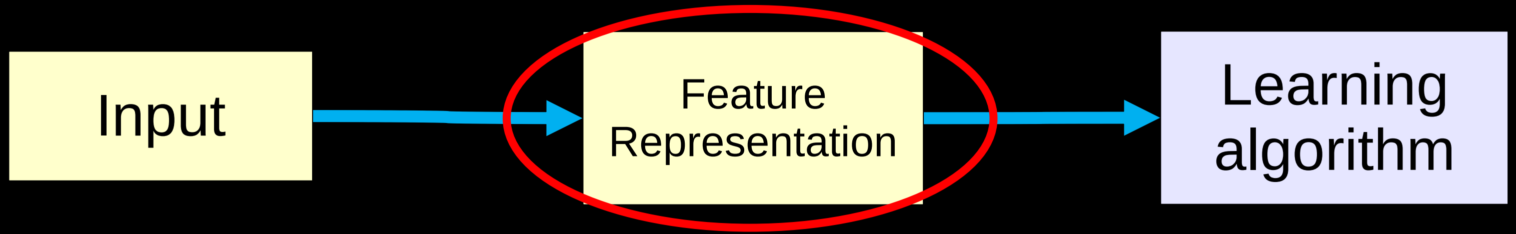
Text



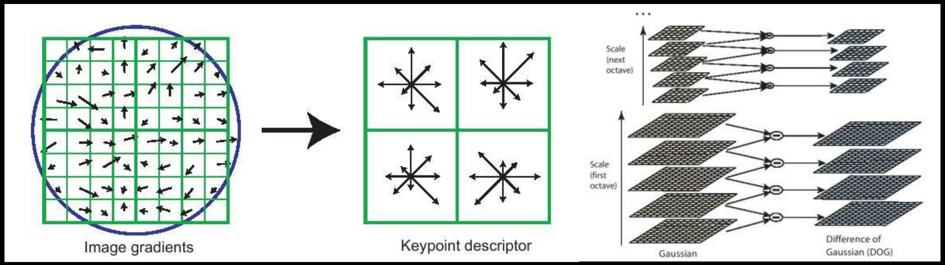
Text features

Text classification,
Machine translation,
Information retrieval,
etc.

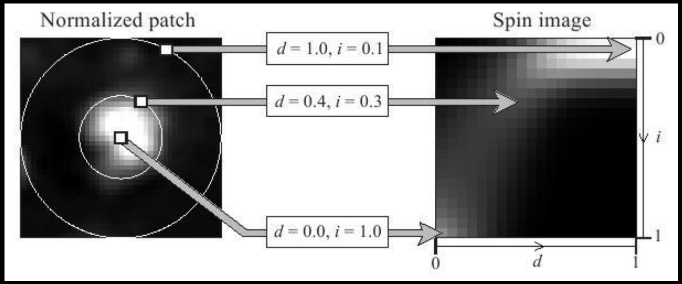
Feature representations



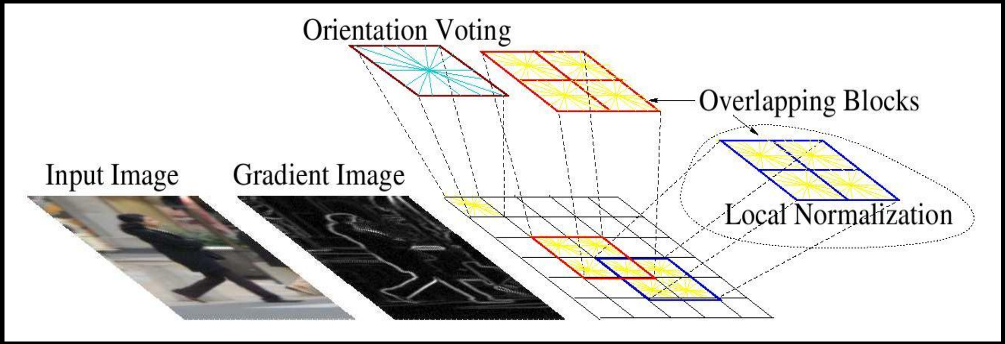
Computer vision features



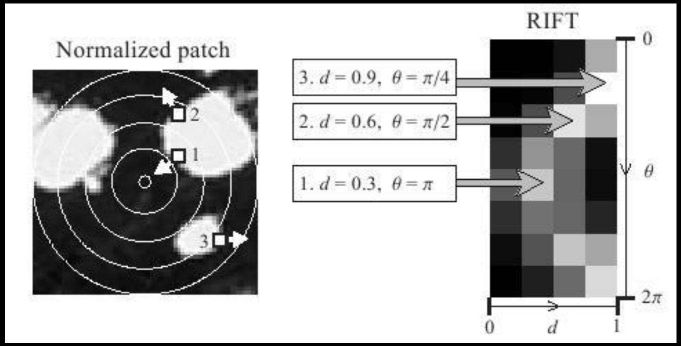
SIFT



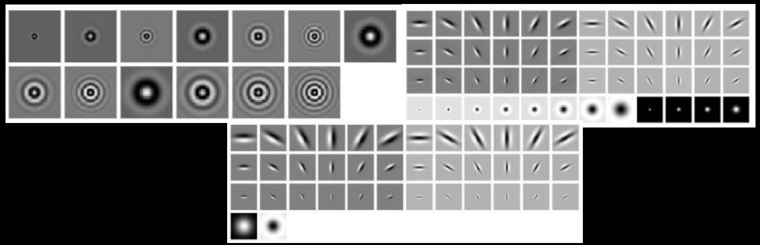
Spin image



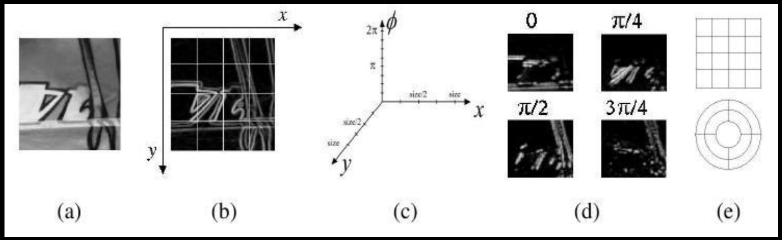
HoG



RIFT



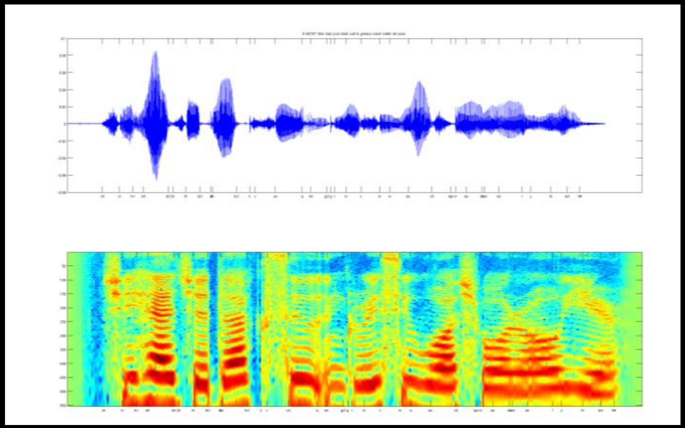
Textons



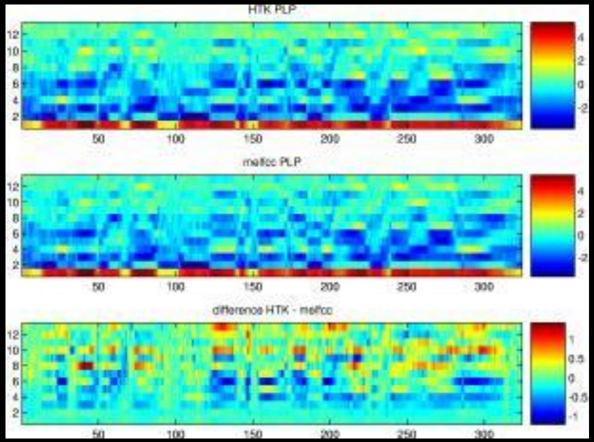
GLOH

Andrew Ng

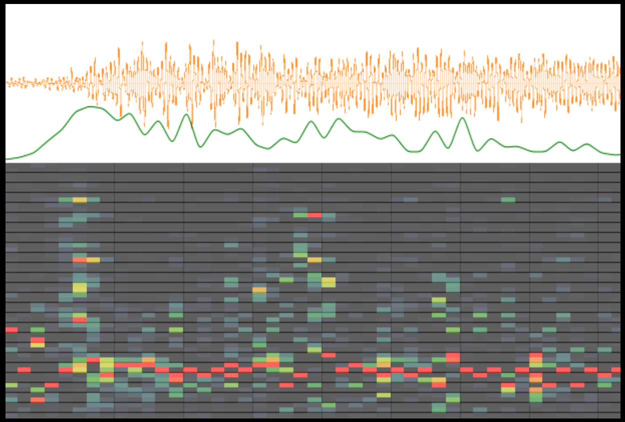
Audio features



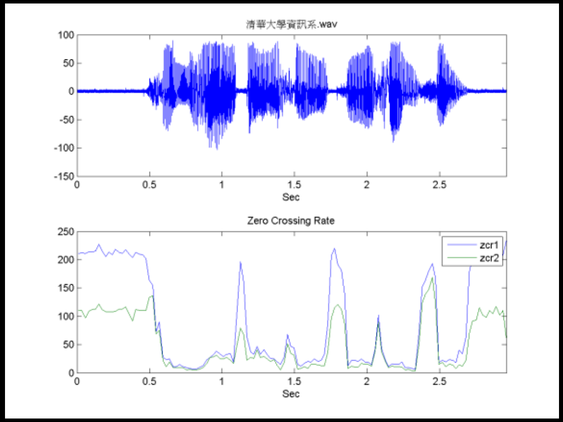
Spectrogram



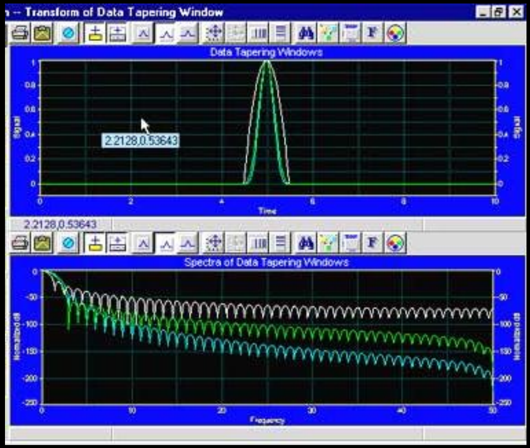
MFCC



Flux

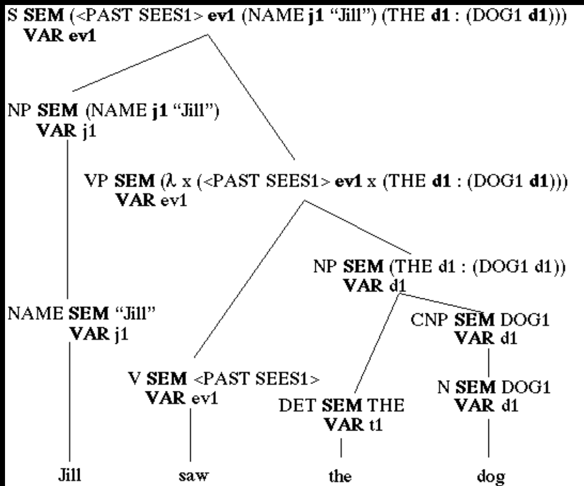


ZCR

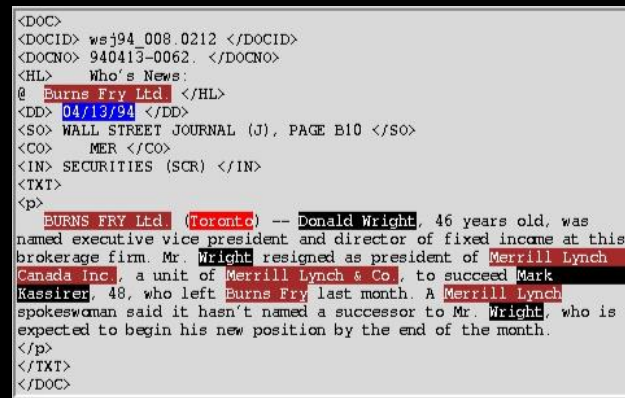


Rolloff

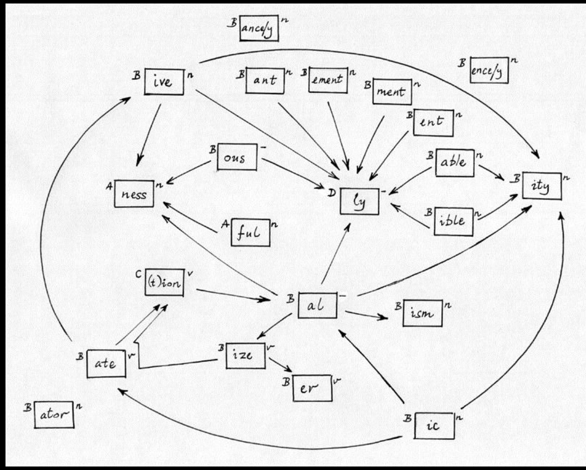
NLP features



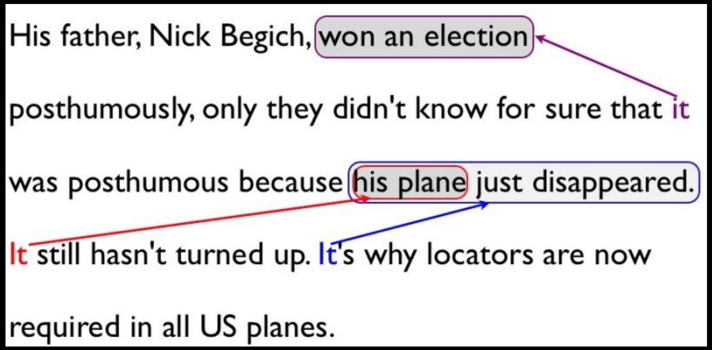
Parser features



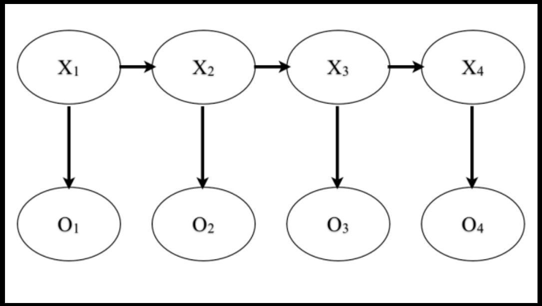
NER/SRL



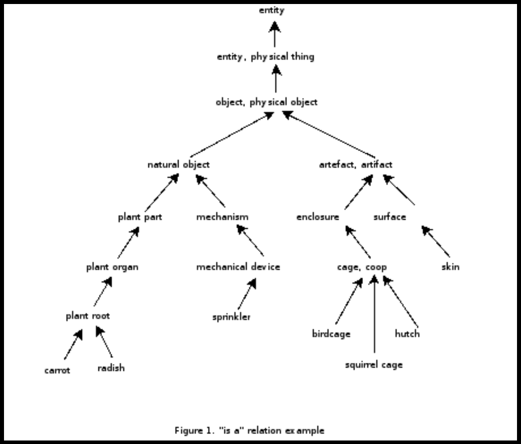
Stemming



Anaphora

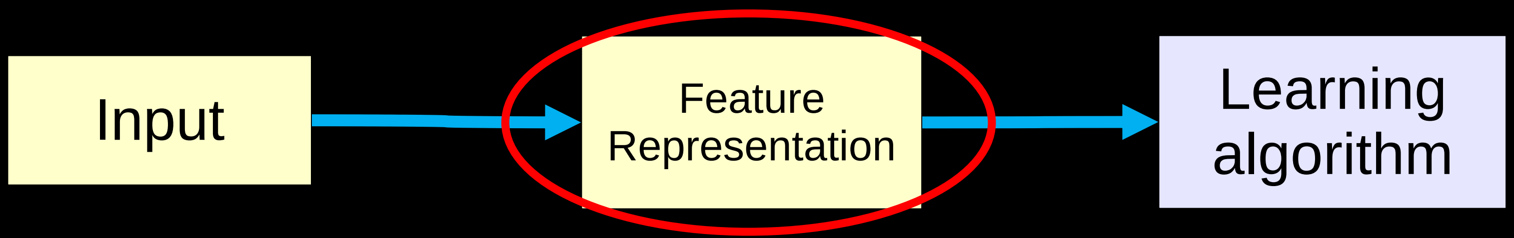


POS tagging

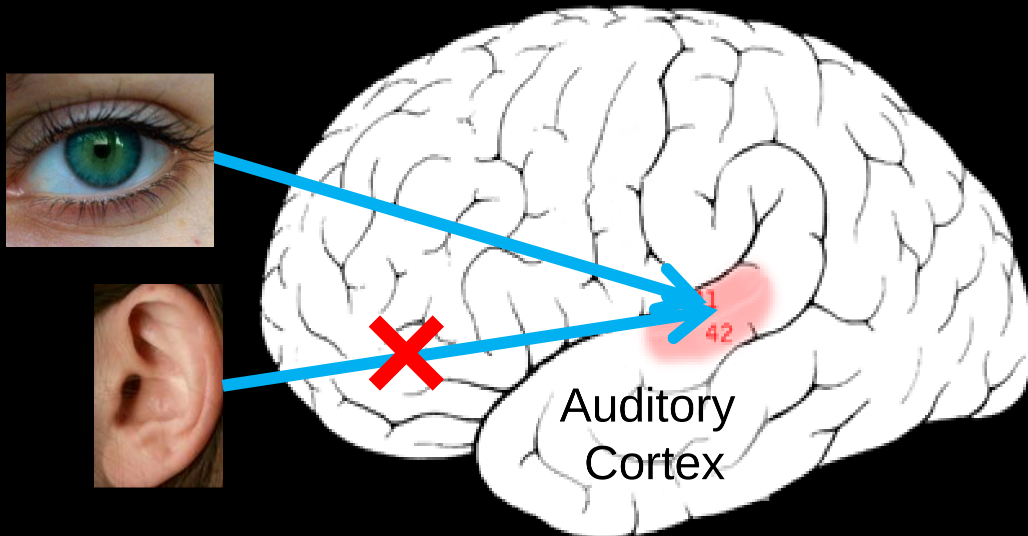


WordNet features

Feature representations



Sensor representation in the brain



Auditory cortex learns to see.

(Same rewiring process also works for touch/ somatosensory cortex.)

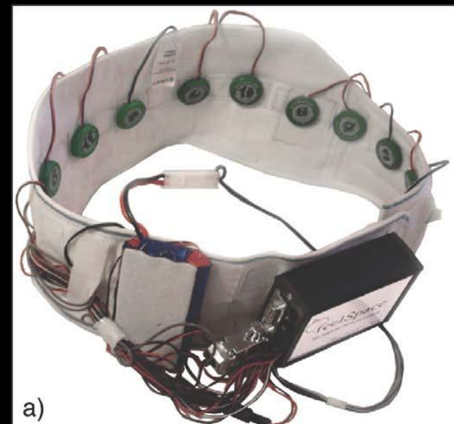
Seeing with your tongue

Human echolocation (sonar)

[Roe et al., 1992; BrainPort; Welsh & Blasch, 1997]

Other sensory remapping examples

Haptic compass belt. North facing motor vibrates. Gives you a “direction” sense.



Implanting a 3rd eye.

[Nagel et al., 2005 and Wired Magazine; Constantine-Paton & Law, 2009]

On two approaches to computer perception

The adult visual (or audio) system is incredibly complicated.

We can try to directly implement what the adult visual (or audio) system is doing. (E.g., implement features that capture different types of invariance, 2d and 3d context, relations between object parts, …).

Or, if there is a more general computational principal/algorithm that underlies most of perception, can we instead try to discover and implement that?

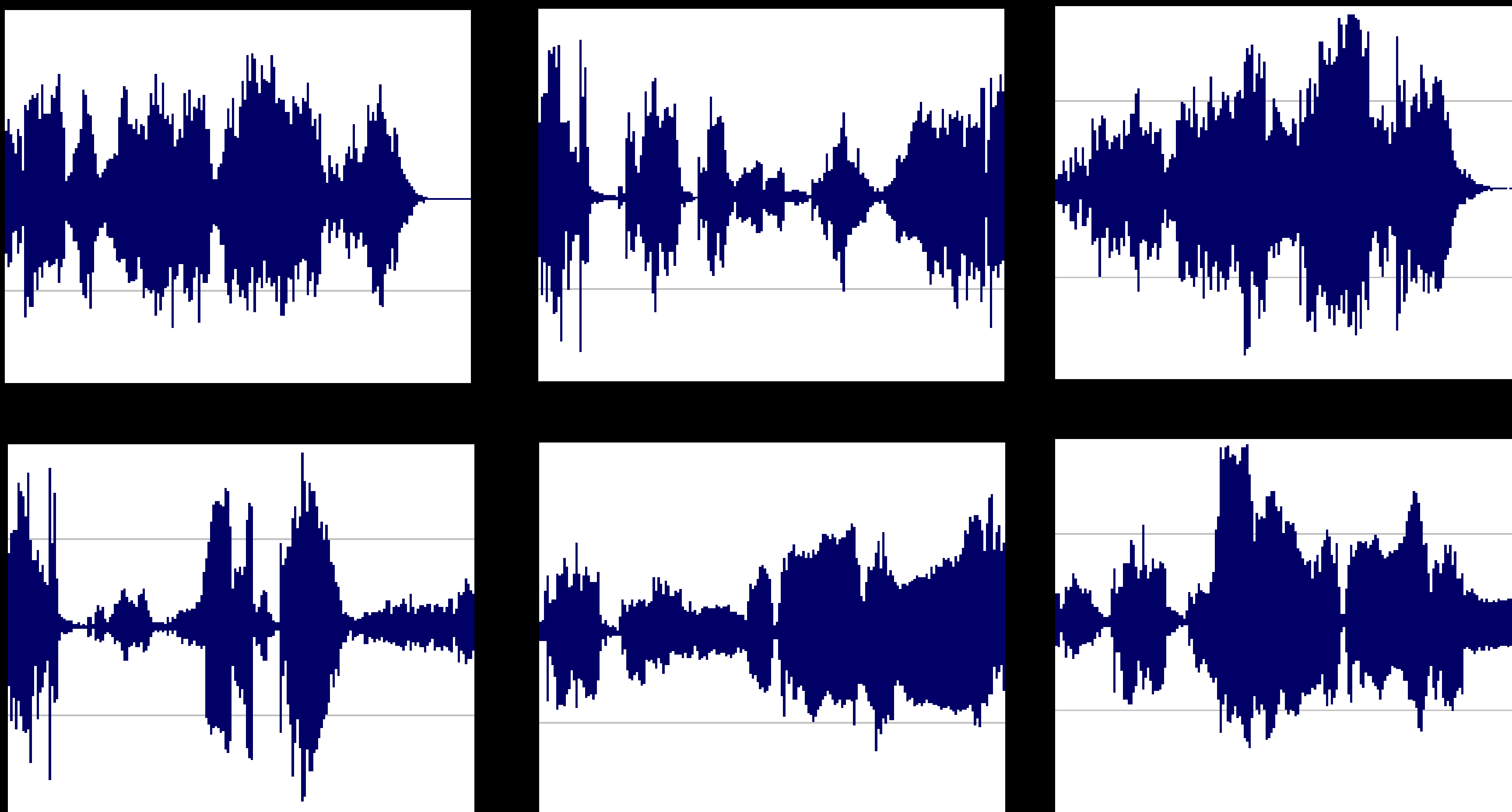
Learning input representations






Find a better way to represent images than pixels.

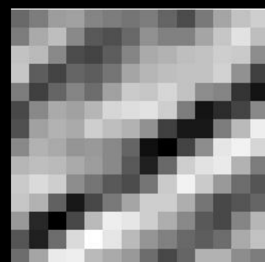
Learning input representations



Find a better way to represent audio.

Feature learning problem

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



$$\begin{pmatrix} 255 \\ 98 \\ 93 \\ 87 \\ 89 \\ 91 \\ 48 \\ \ldots \end{pmatrix}$$

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

Supervised Learning: Recognizing motorcycles

Motorcycles

Not motorcycles

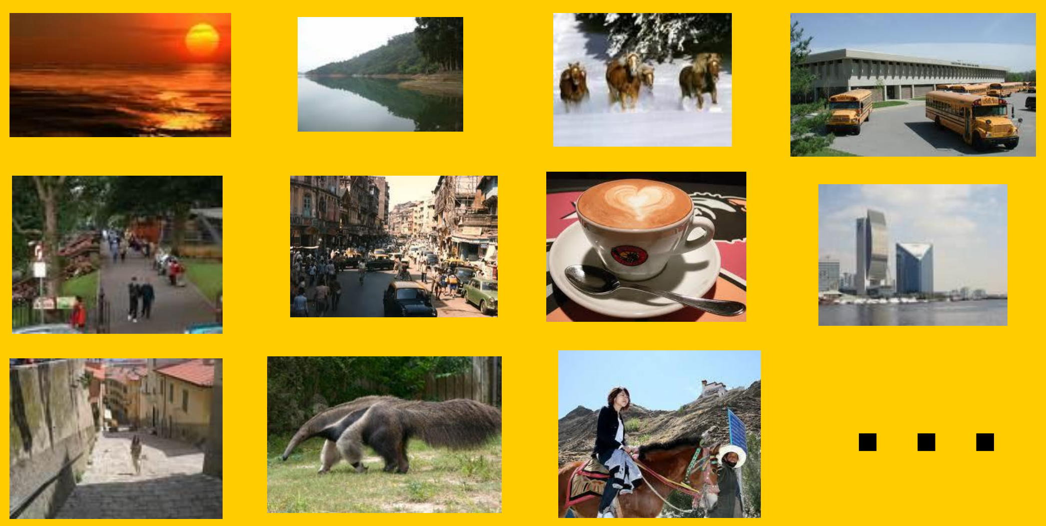
Testing:
What is this?

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

Self-taught learning (Feature learning problem)

Motorcycles

Not motorcycles



Unlabeled images

Testing:
What is this?

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

Feature Learning via Sparse Coding

Sparse coding (Olshausen & Field,1996). Originally developed to explain early visual processing in the brain (edge detection).

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

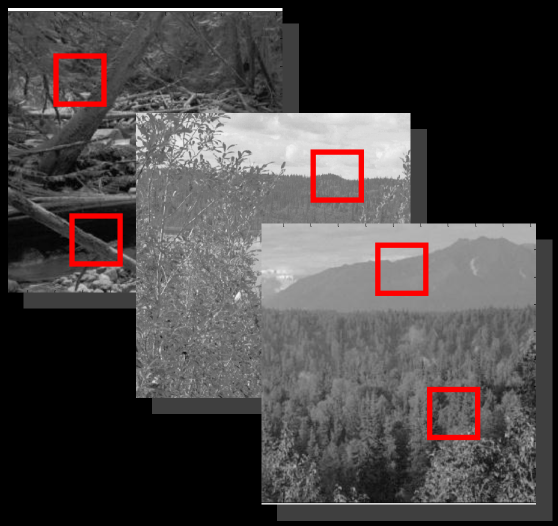
Learn: Dictionary of bases $\phi_1, \phi_2, \ldots, \phi_k$ (also $R^{n \times n}$), so that each input x can be approximately decomposed as:

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

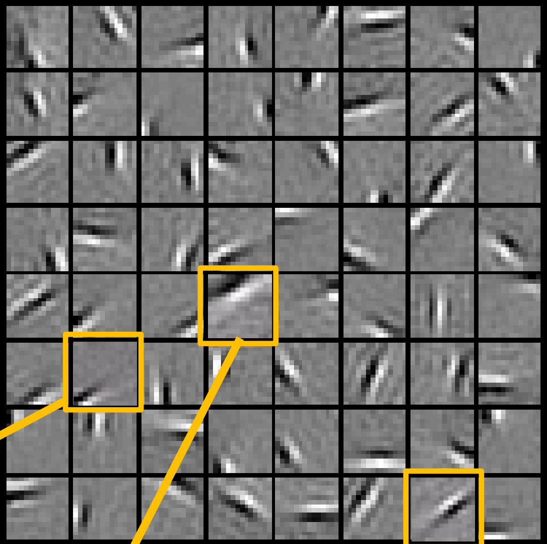
s.t. a_j's are mostly zero ("sparse")

Sparse coding illustration

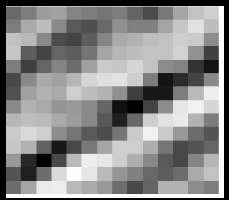
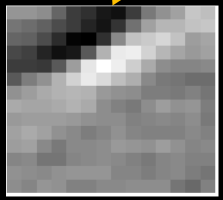
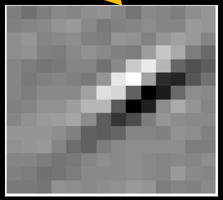
Natural Images



Learned bases ($\phi_1, \ldots, \phi_{64}$): "Edges"



Test example

 $\approx 0.8 *$  $+ 0.3 *$  $+ 0.5 *$ 

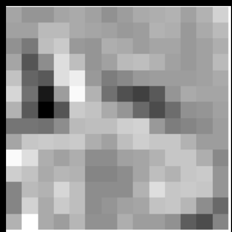
$$x \approx 0.8 * \phi_{36} + 0.3 * \phi_{42} + 0.5 * \phi_{63}$$

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

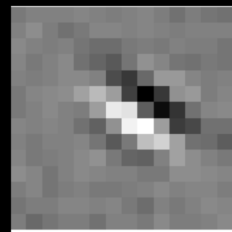
(feature representation)

Compact & easily interpretable

More examples

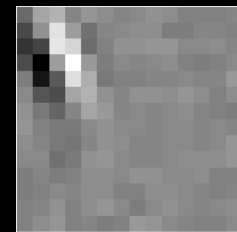


$\approx 0.6 *$



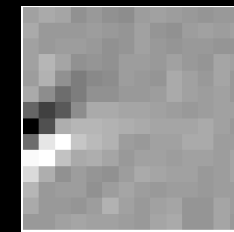
ϕ_{15}

$+ 0.8 *$



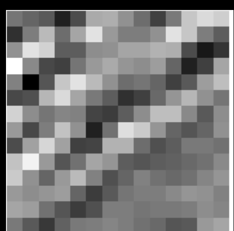
ϕ_{28}

$+ 0.4 *$

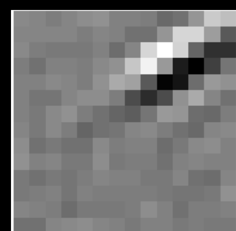


ϕ_{37}

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

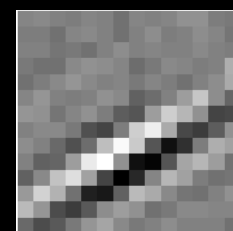


$\approx 1.3 *$



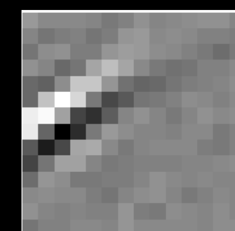
ϕ_{5}

$+ 0.9 *$



ϕ_{18}

$+ 0.3 *$



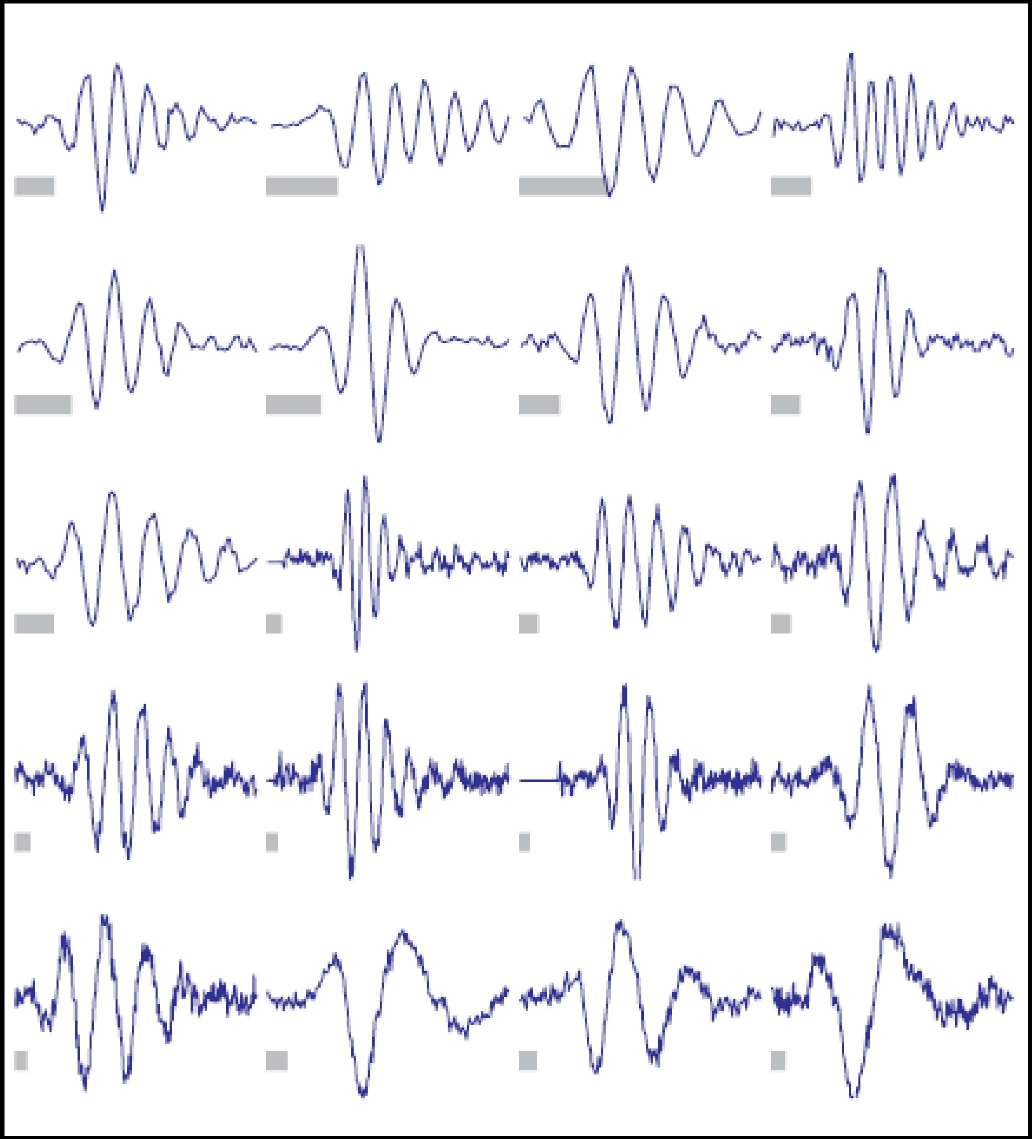
ϕ_{29}

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

- Method “invents” edge detection.
- Automatically learns to represent an image in terms of the edges that appear in it. 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

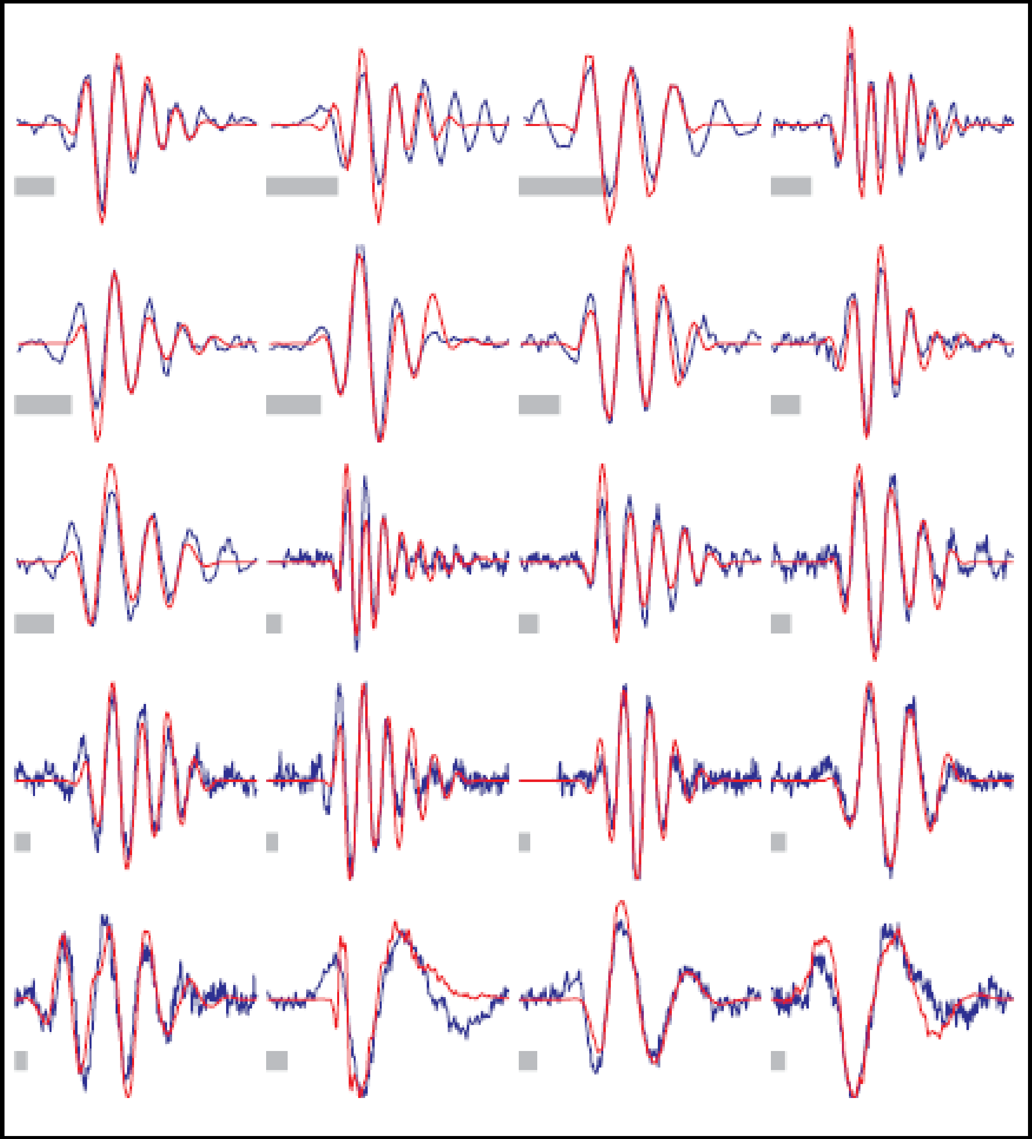
Image shows 20 basis functions learned from unlabeled audio.



[Evan Smith & Mike Lewicki, 2006]

Sparse coding applied to audio

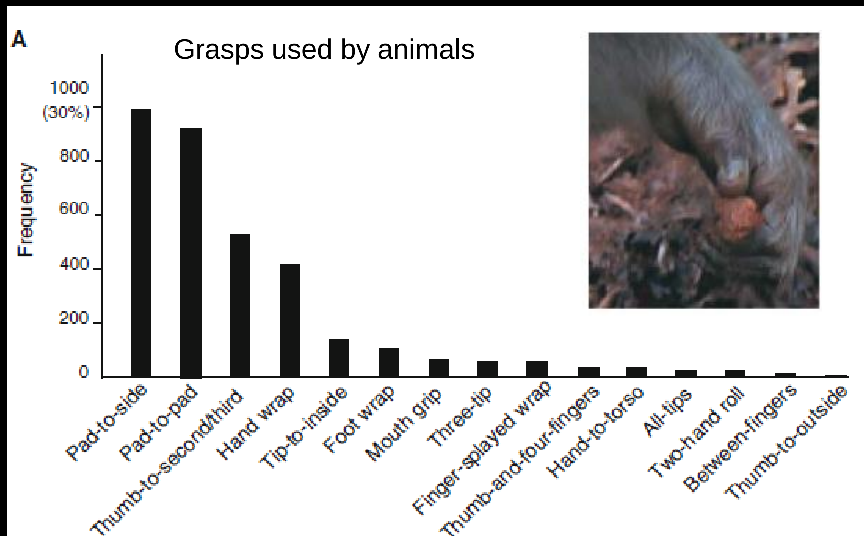
Image shows 20 basis functions learned from unlabeled audio.



[Evan Smith & Mike Lewicki, 2006]

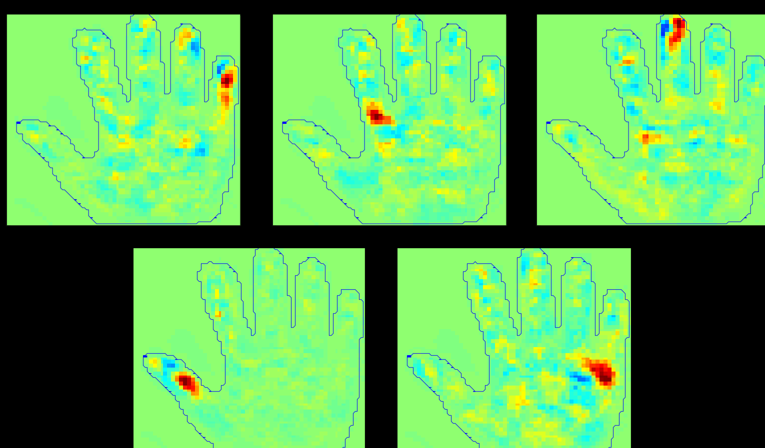
Sparse coding applied to touch data

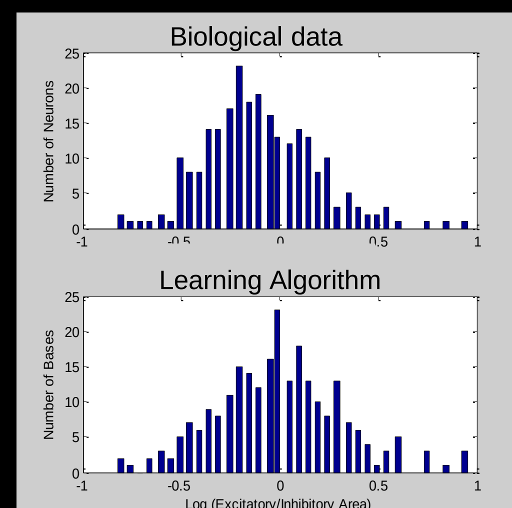
Collect touch data using a glove, following distribution of grasps used by animals in the wild.



[Macfarlane & Graziano, 2009]

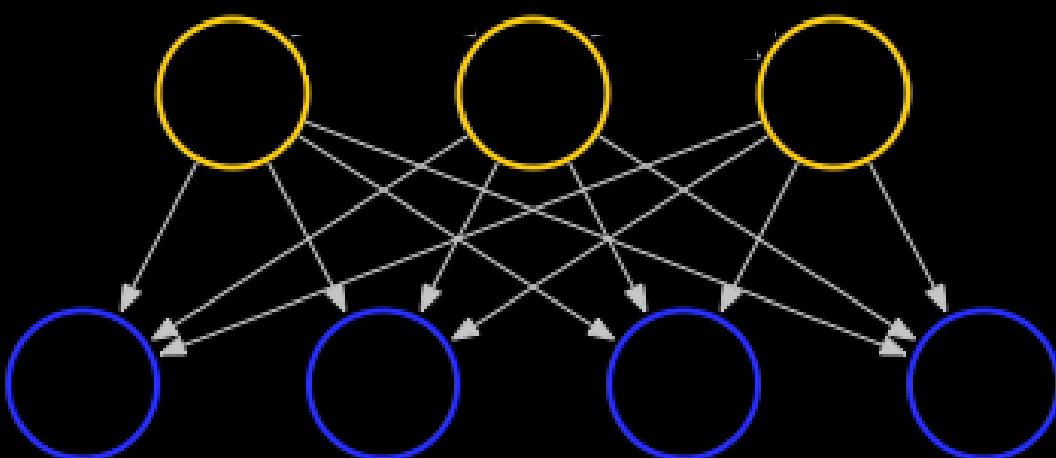
Example learned representations





[Saxe, Bhand, Mudur, Suresh & Ng, 2011]

Learning feature hierarchies



Higher layer
(Combinations of edges;
cf V2)

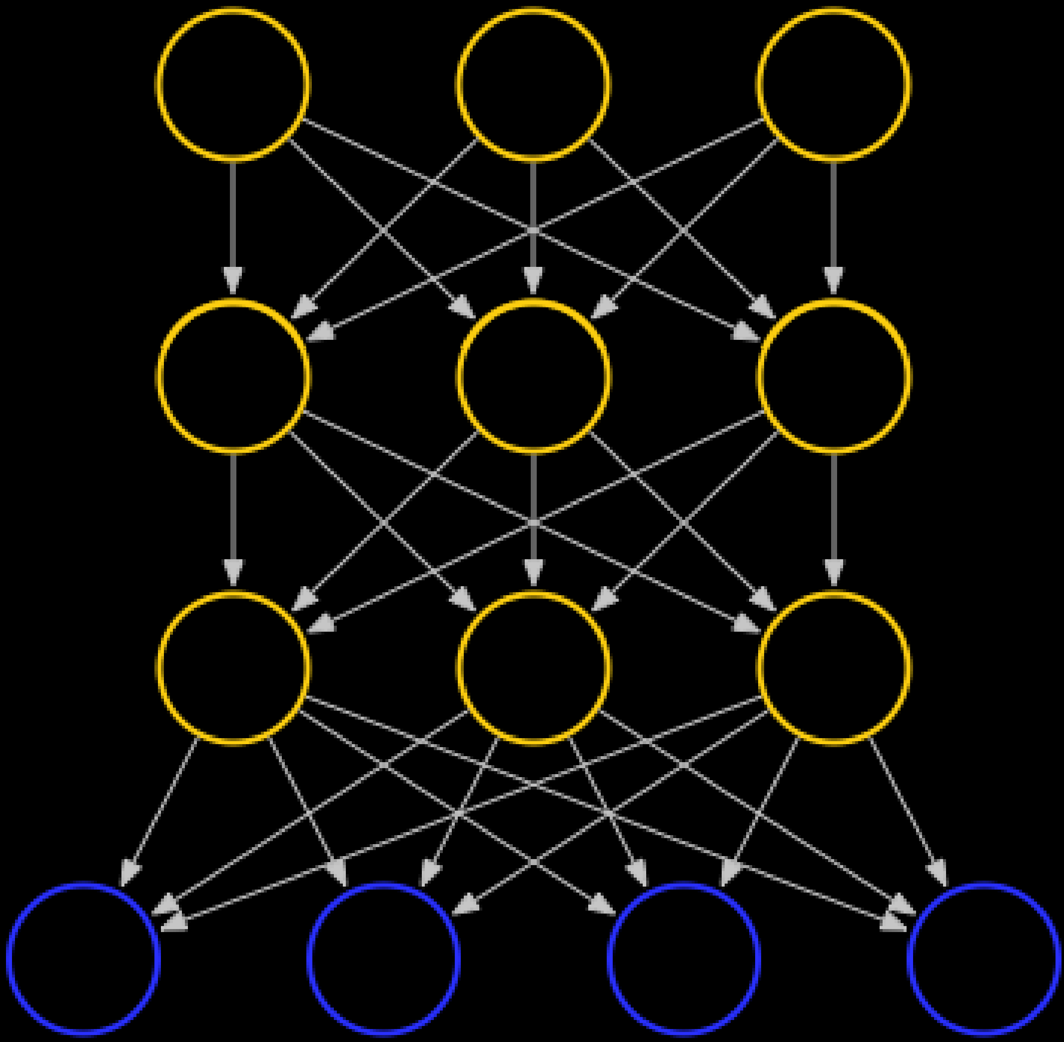
“Sparse coding”
(edges; cf. V1)

Input image (pixels)

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

[Lee, Ranganath & Ng, 2007]

Learning feature hierarchies



Higher layer
(Model V3?)

Higher layer
(Model V2?)

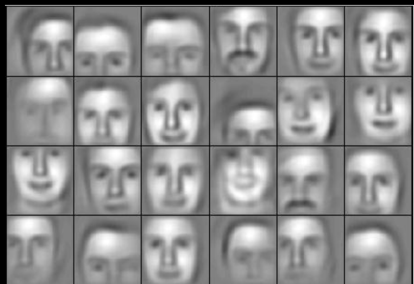
Model V1

Input image

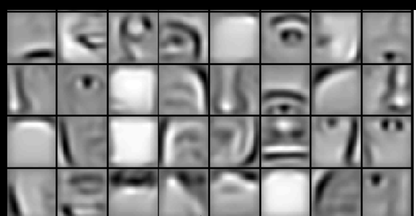
[Technical details: Sparse autoencoder or sparse version of Hinton's DBN.]

[Lee, Ranganath & Ng, 2007]

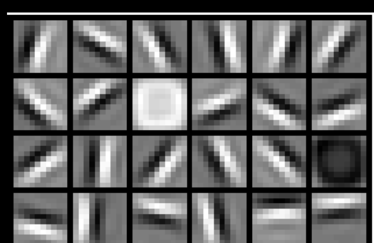
Sparse DBN: Training on face images



object models



object parts
(combination
of edges)



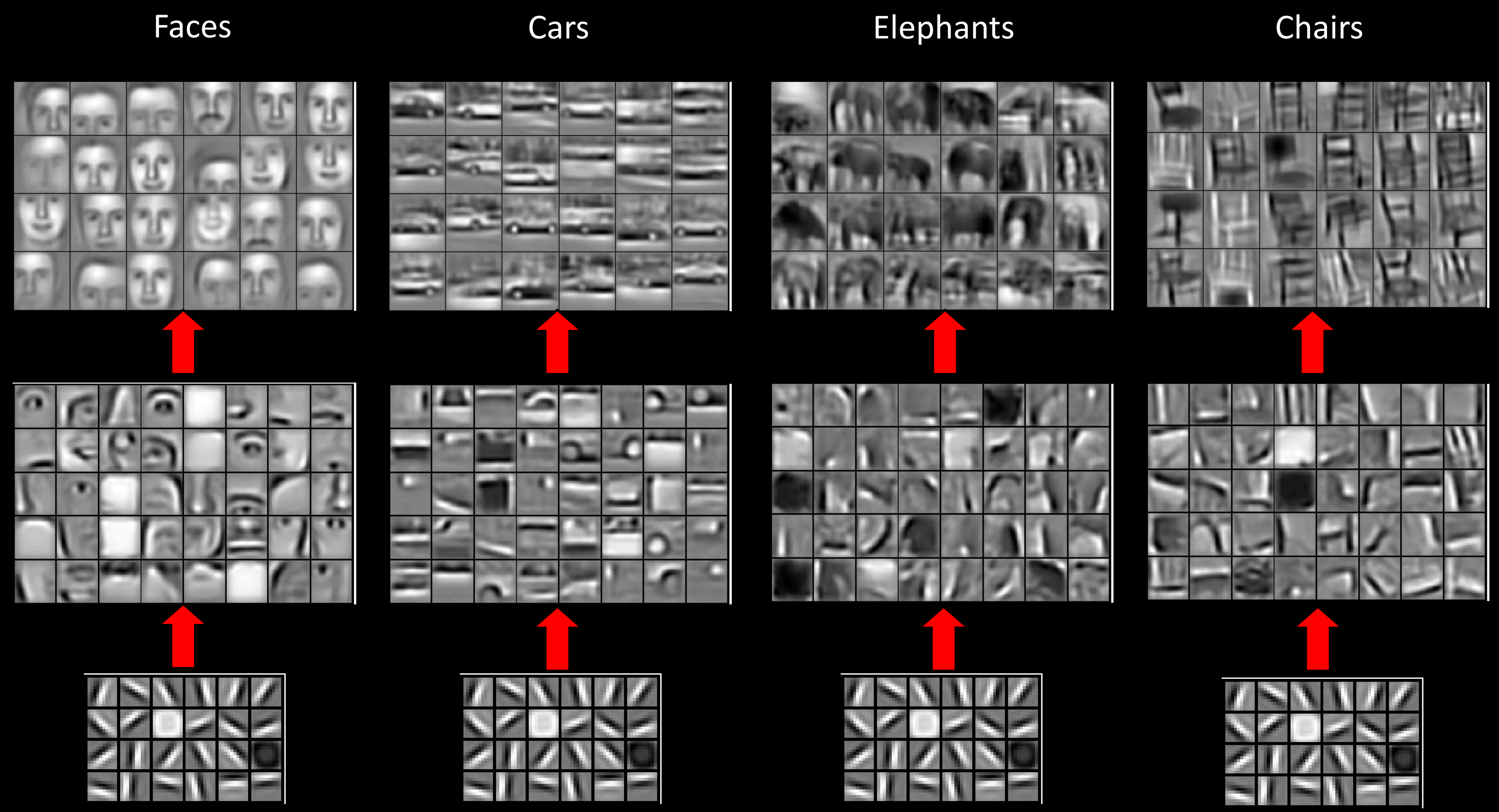
edges

pixels

[Lee, Grosse, Ranganath & Ng, 2009]

Sparse DBN

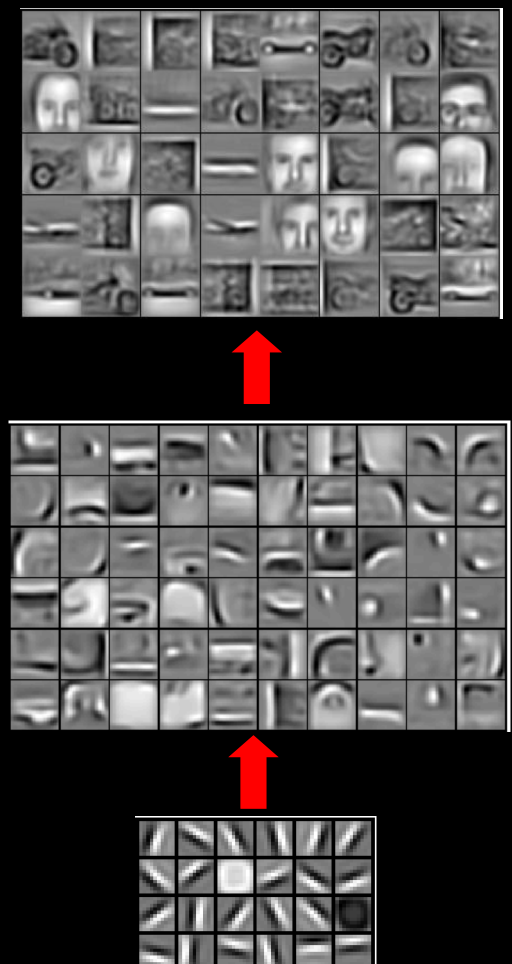
Features learned from different object classes.



[Lee, Grosse, Ranganath & Ng, 2009]

Training on multiple objects

Features learned by training on 4 classes (cars, faces, motorbikes, airplanes).



Object specific features

Features shared across object classes

Edges

[Lee, Grosse, Ranganath & Ng, 2009]

Machine learning applications

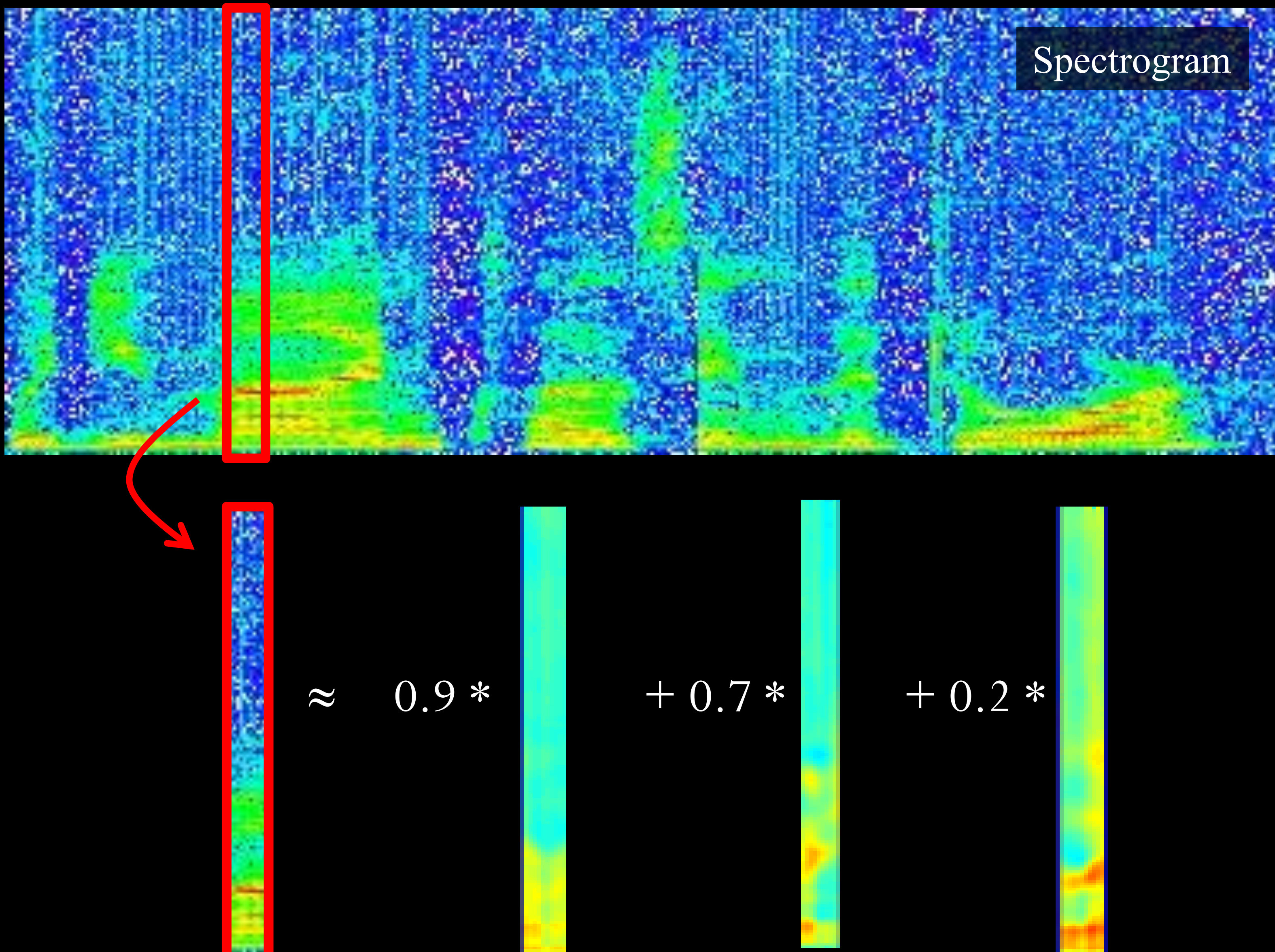
Activity recognition (Hollywood 2 benchmark)

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]

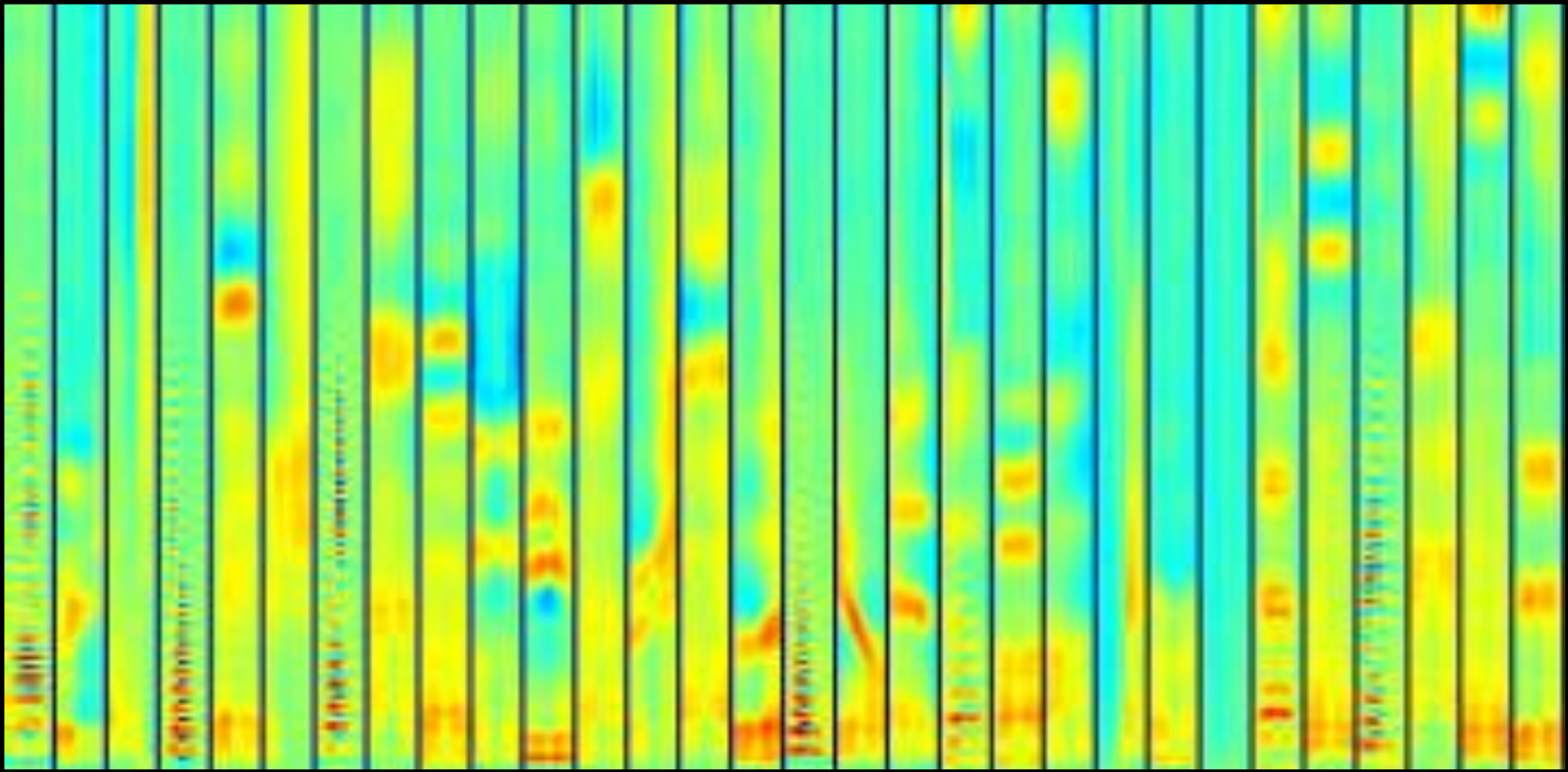
Sparse coding on audio



$$x \approx 0.9 * \phi_{36} + 0.7 * \phi_{42} + 0.2 * \phi_{63}$$

[Lee, Pham and Ng, 2009]

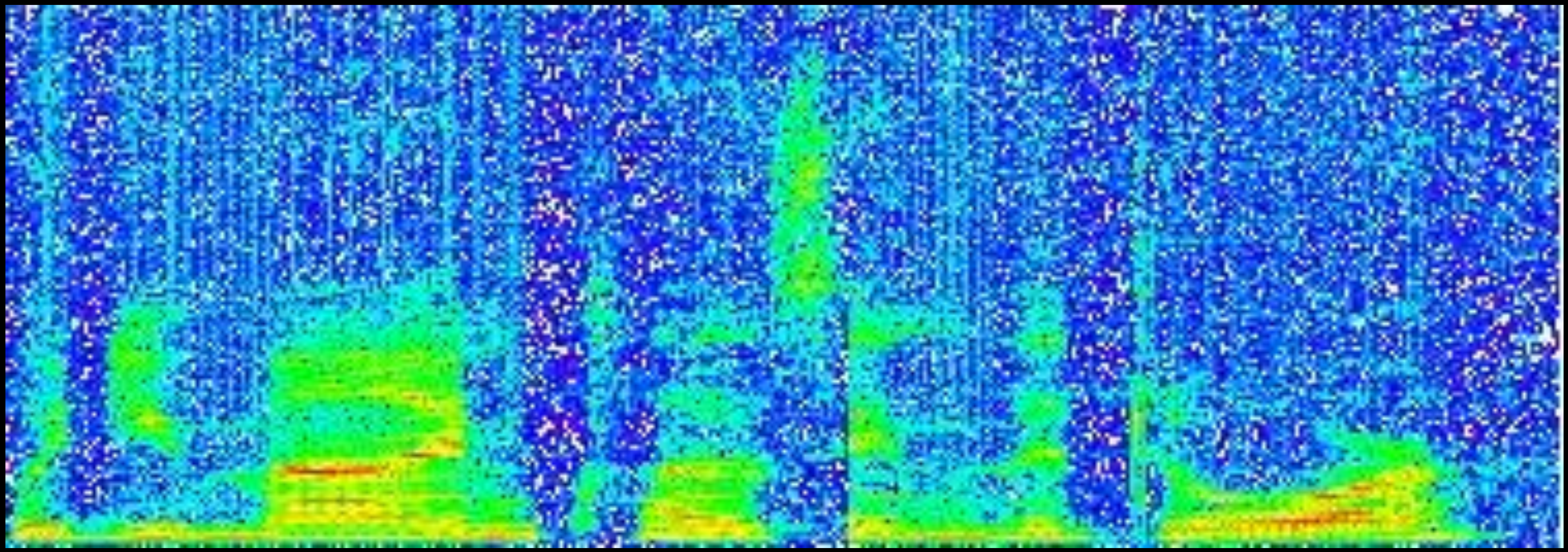
Dictionary of bases ϕ_i learned for speech



Many bases seem to correspond to phonemes.

[Lee, Pham and Ng, 2009]

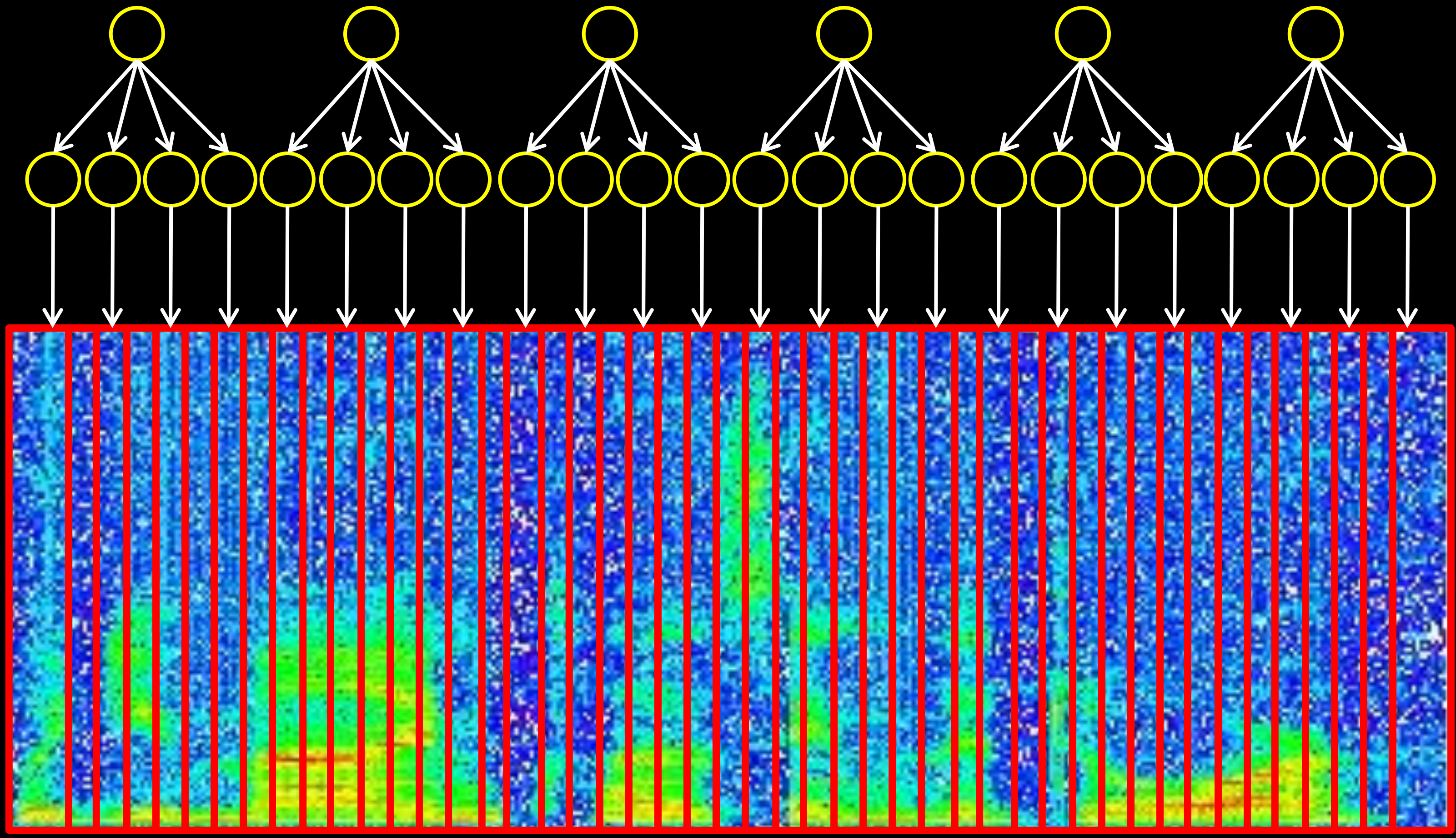
Sparse DBN for audio



Spectrogram

[Lee, Pham and Ng, 2009]

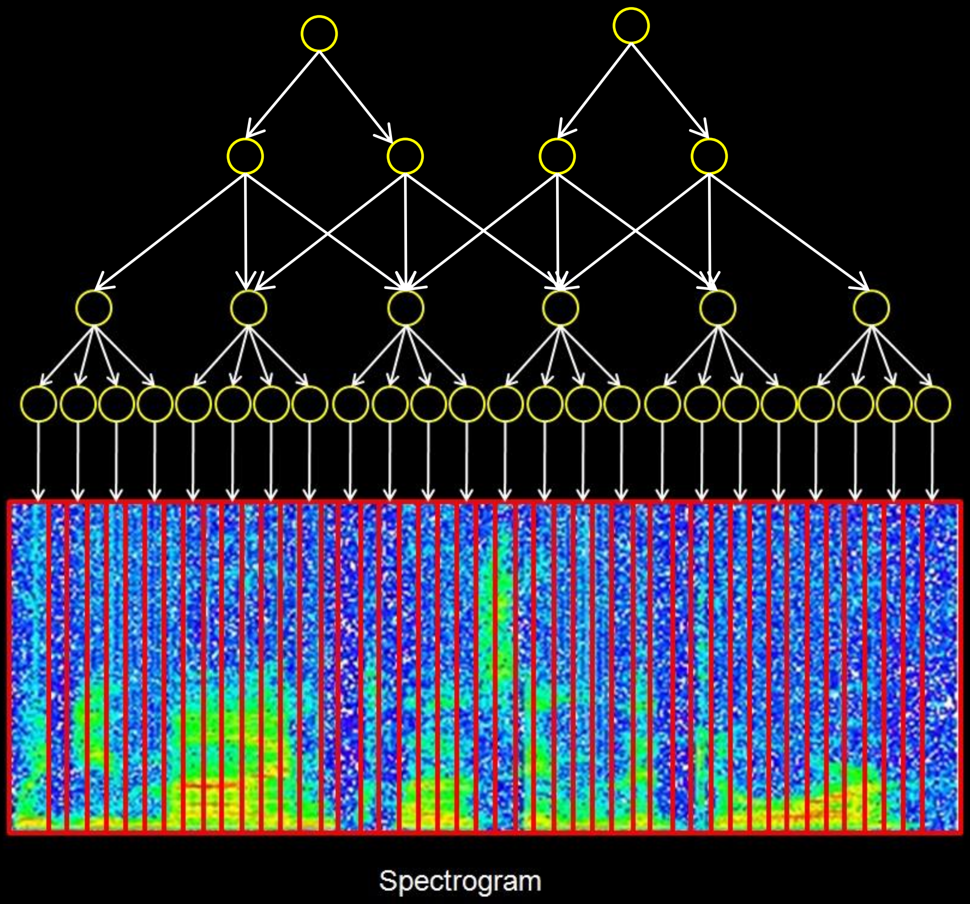
Sparse DBN for audio



Spectrogram

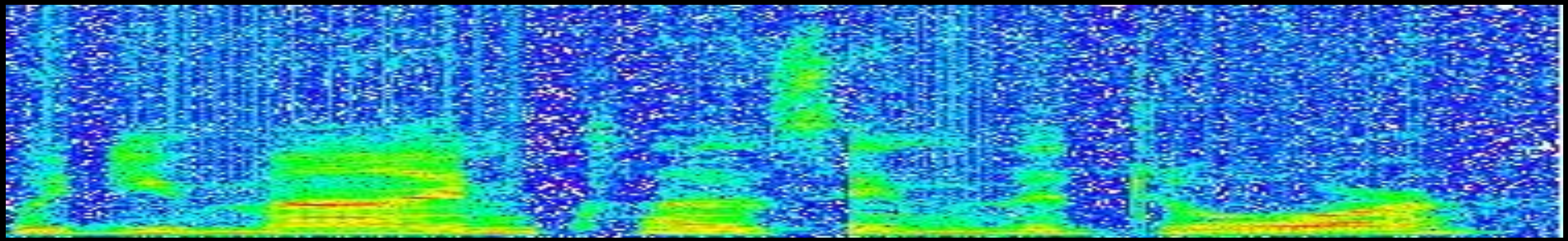
[Lee, Pham and Ng, 2009]

Sparse DBN for audio



[Lee, Pham and Ng, 2009]

Phoneme Classification (TIMIT benchmark)



Method	Accuracy
Clarkson and Moreno (1999)	77.6%
Gunawardana et al. (2005)	78.3%
Sung et al. (2007)	78.5%
Petrov et al. (2007)	78.6%
Sha and Saul (2006)	78.9%
Yu et al. (2006)	79.2%
Unsupervised feature learning (our method)	**80.3%**

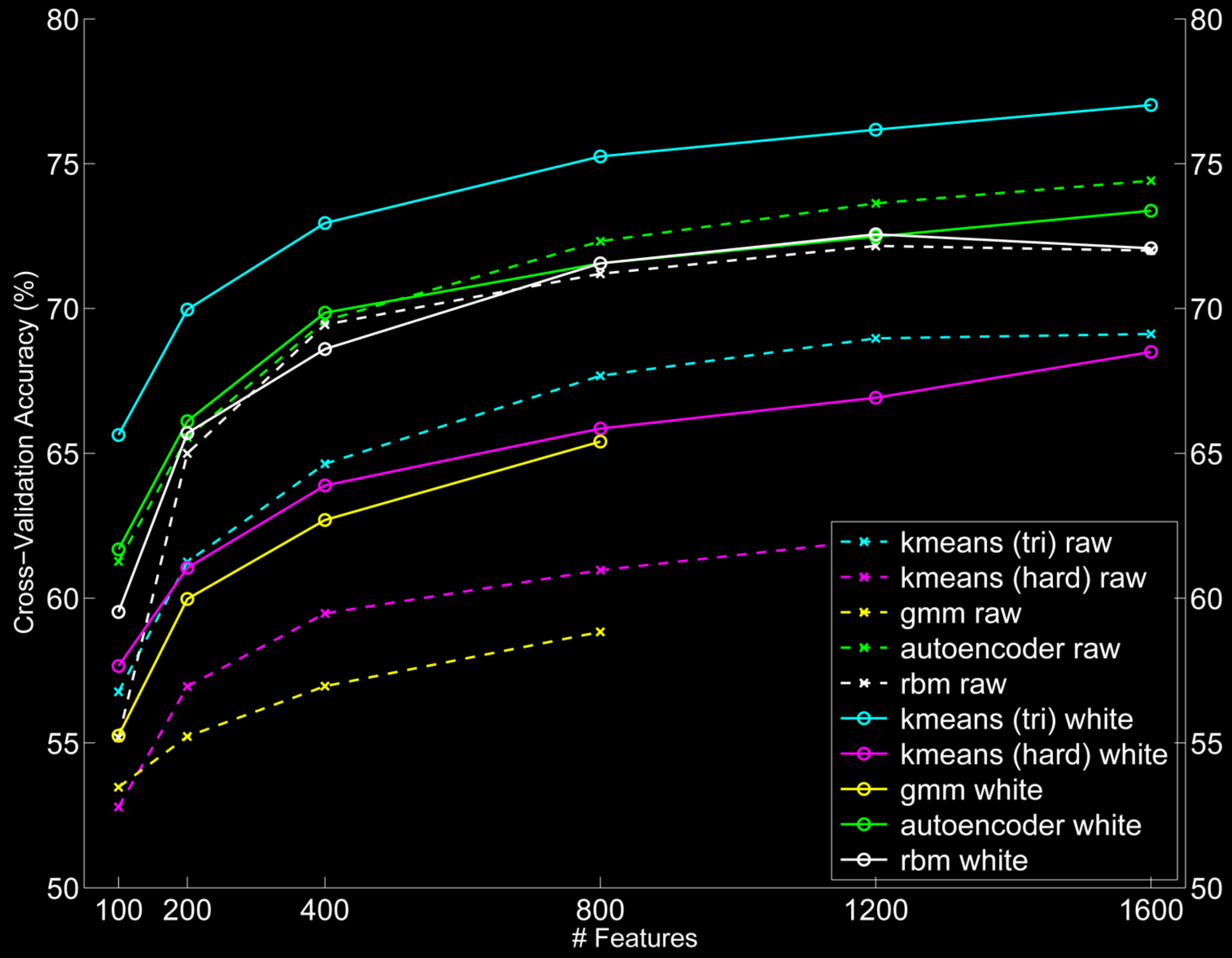
Unsupervised feature learning significantly improves on the previous state-of-the-art.

[Lee, Pham and Ng, 2009]

Technical challenge: Scaling up

Scaling and classification accuracy (CIFAR-10)

Large numbers of features is critical. Algorithms that can scale to many features have a big advantage.



[Coates, Lee and Ng, 2010]

Approaches to scaling up

- Efficient sparse coding algorithms. (Lee et al., NIPS 2006)
- Parallel implementations via Map-Reduce (Chu et al., NIPS 2006)
- GPUs for deep learning. (Raina et al., ICML 2008)
- Tiled Convolutional Networks (Le et al., NIPS 2010)
  - The scaling advantage of convolutional networks, but without hard-coding translation invariance.
- Efficient optimization algorithms (Le et al., ICML 2011)
- Simple, fast feature decoders (Coates et al., AISTATS 2011)

State-of-the-art Unsupervised feature learning

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 (Krizhevsky, 2010)	78.9%
Stanford Feature learning	**81.5%**

NORB Object classification	Accuracy
Prior art (Ranzato et al., 2009)	94.4%
Stanford Feature learning	**97.3%**

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%**

Multimodal (audio/video)

AVLetters Lip reading	Accuracy
Prior art (Zhao et al., 2009)	58.9%
Stanford Feature learning	**65.8%**

Other unsupervised feature learning records:
Pedestrian detection (Yann LeCun)
Different phone recognition task (Geoff Hinton)
PASCAL VOC object classification (Kai Yu)

Kai Yu's PASCAL VOC (Object recognition) result (2009)

Class	Feature Learning	Best of Other Teams	Difference
Aeroplane	88.1	86.6	1.5
Bicycle	68.6	63.9	**4.7**
Bird	68.1	66.7	1.4
Boat	72.9	67.3	**5.6**
Bottle	44.2	43.7	0.5
Bus	79.5	74.1	**5.4**
Car	72.5	64.7	**7.8**
Cat	70.8	64.2	**6.6**
Chair	59.5	57.4	2.1
Cow	53.6	46.2	**7.4**
Diningtable	57.5	54.7	2.8
Dog	59.3	53.5	**5.8**
Horse	73.1	68.1	**5.0**
Motorbike	72.3	70.6	1.7
Person	85.3	85.2	0.1
Pottedplant	36.6	39.1	-2.5
Sheep	56.9	48.2	**8.7**
Sofa	57.9	50.0	**7.9**
Train	86.0	83.4	2.6
Tvmonitor	68.0	68.6	-0.6

- Sparse coding to learn features.
- Unsupervised feature learning beat all the other approaches by a significant margin.

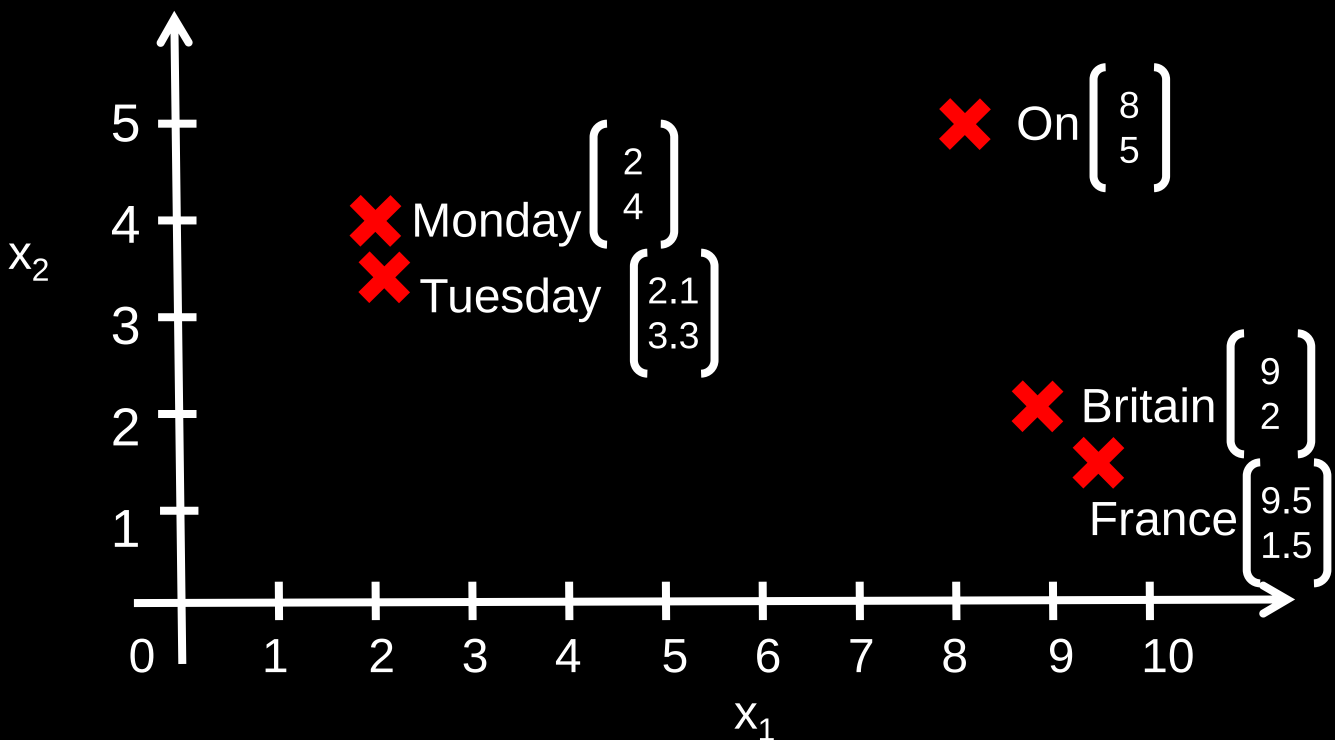
[Courtesy of Kai Yu]

Learning Recursive Representations

Feature representations of words

Imagine taking each word, and embedding it in an n-dimensional space. (cf. distributional representations, or Bengio et al., 2003; Collobert & Weston, 2008).

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



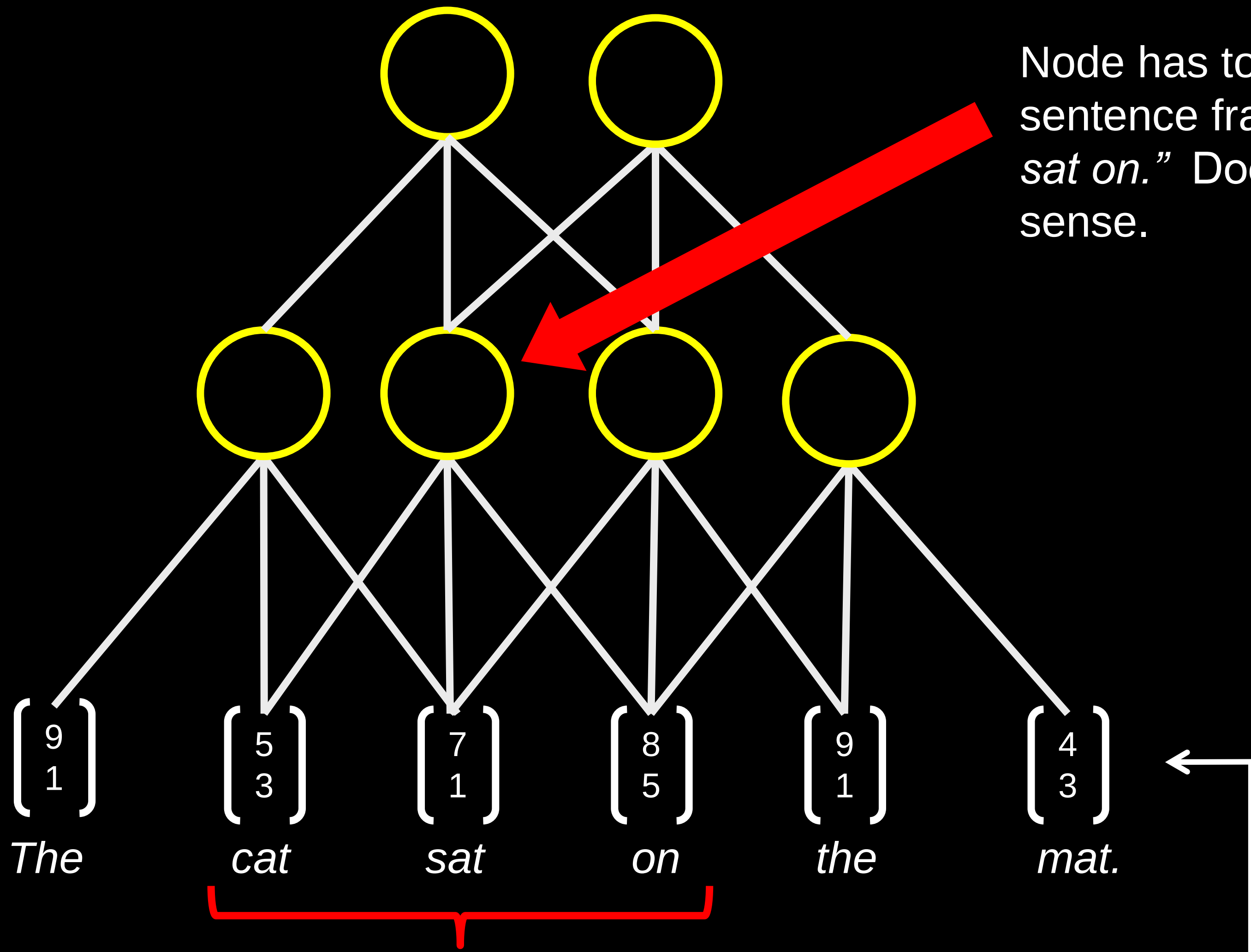
$$\begin{bmatrix}0\\0\\0\\0\\1\\0\\0\\0\end{bmatrix} \qquad \begin{bmatrix}0\\1\\0\\0\\0\\0\\0\\0\end{bmatrix}$$

Monday Britain

On *Monday,* *Britain ….*

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

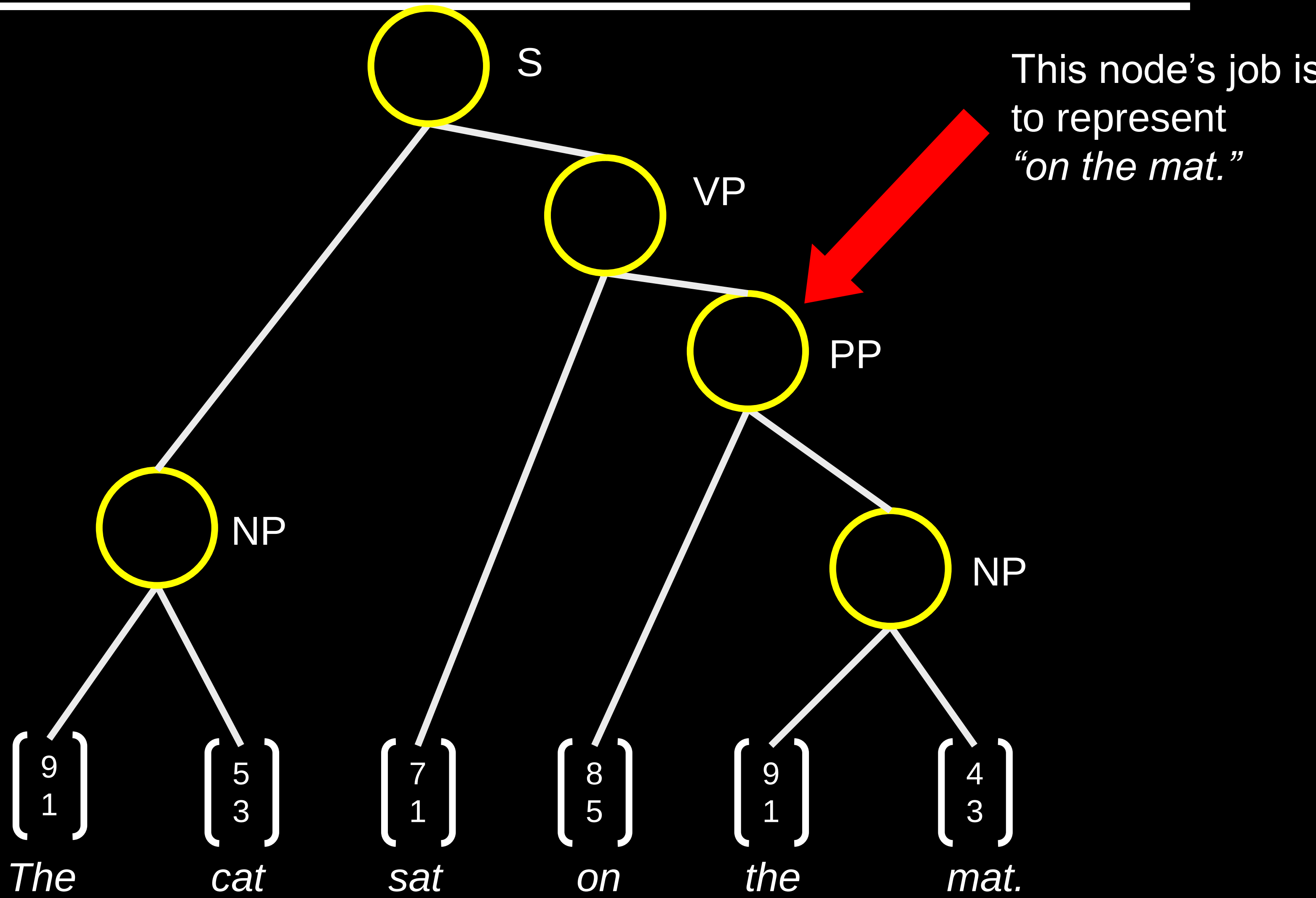
“Generic” hierarchy on text doesn’t make sense



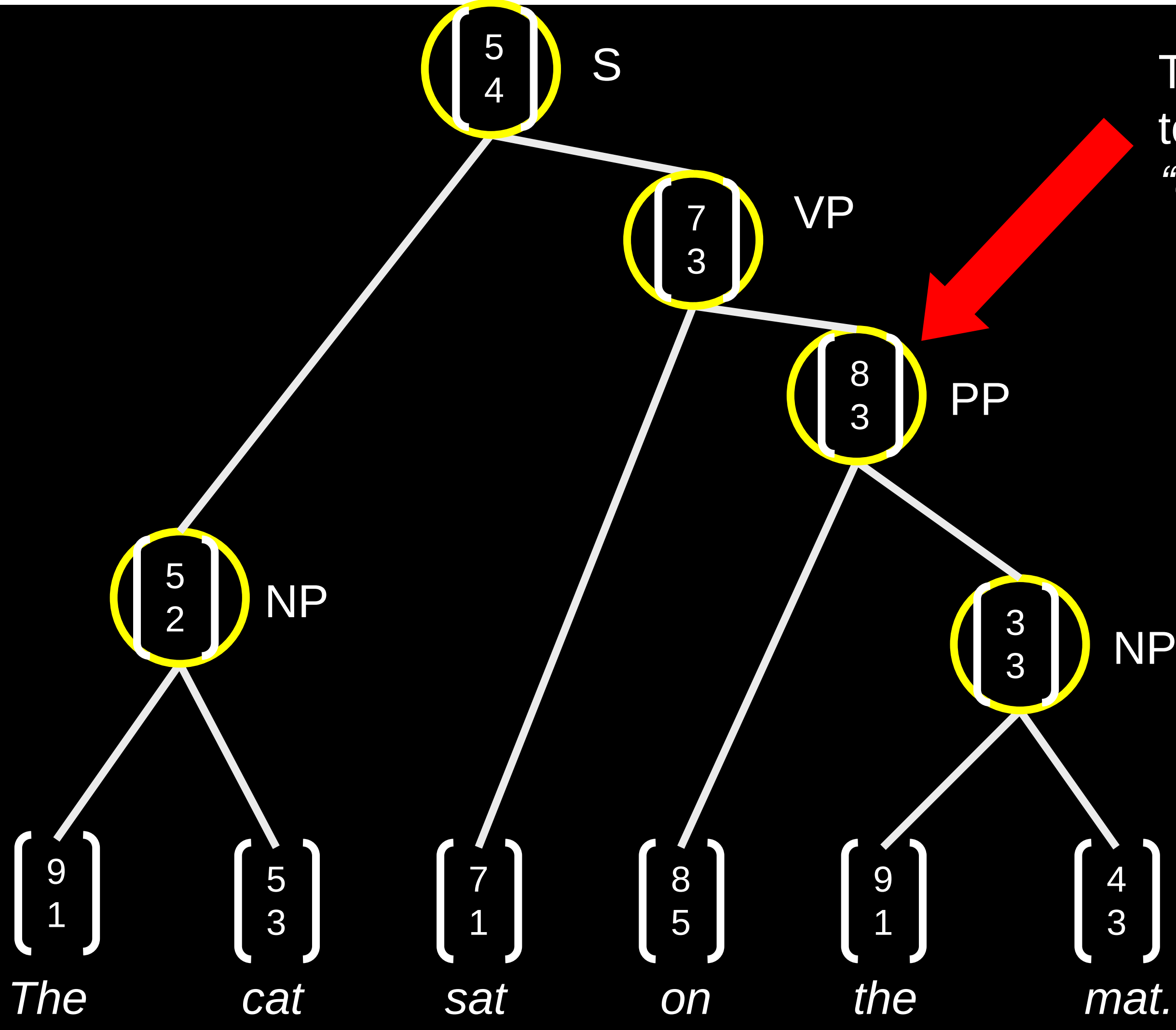
Node has to represent sentence fragment *“cat sat on.”* Doesn’t make sense.

Feature representation for words

What we want (illustration)

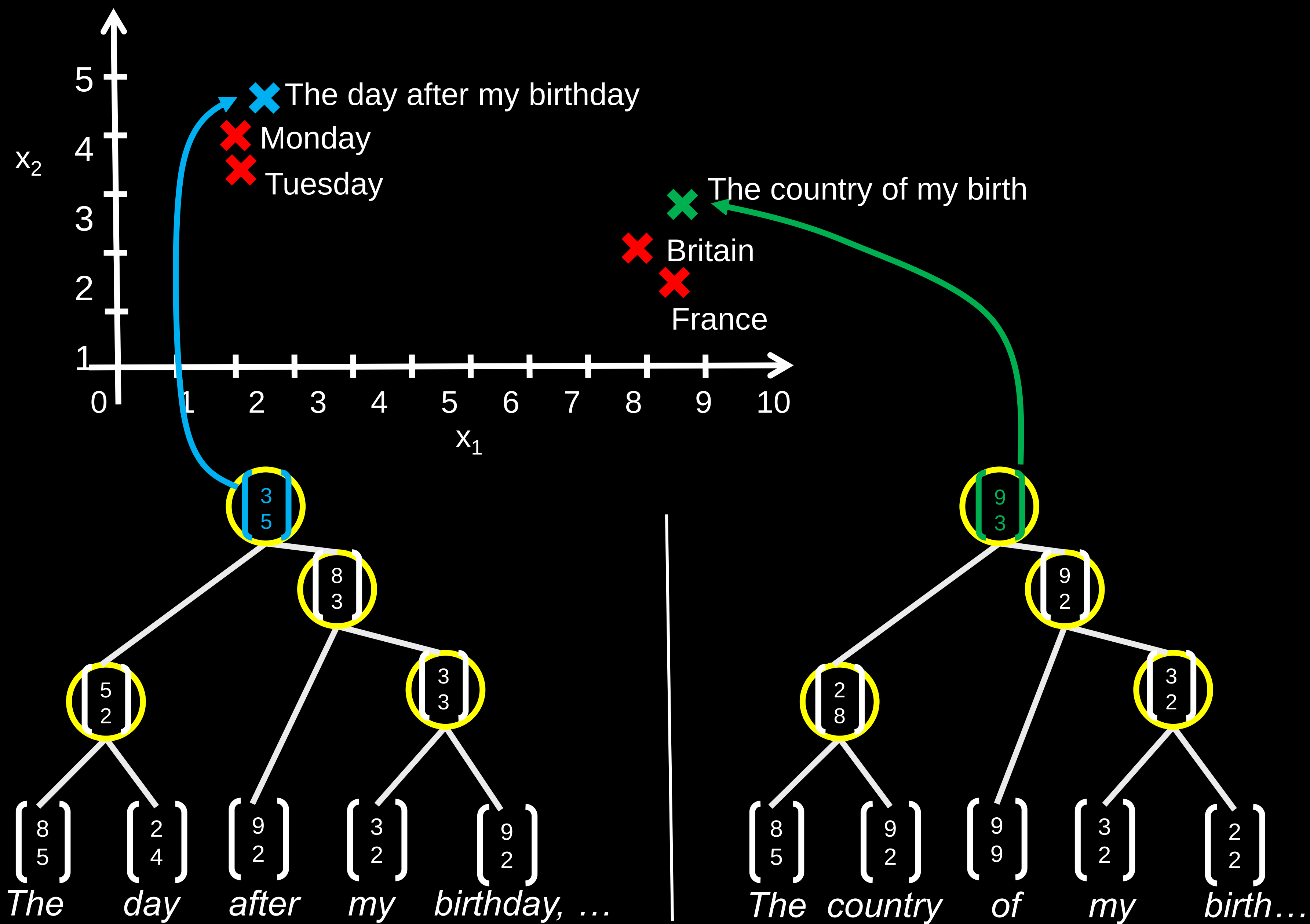


What we want (illustration)



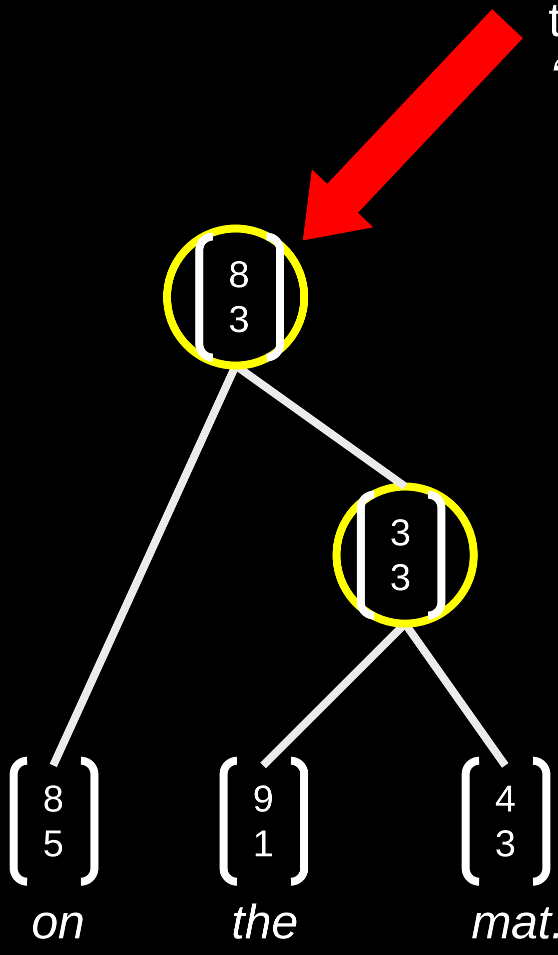
This node's job is to represent *"on the mat."*

What we want (illustration)



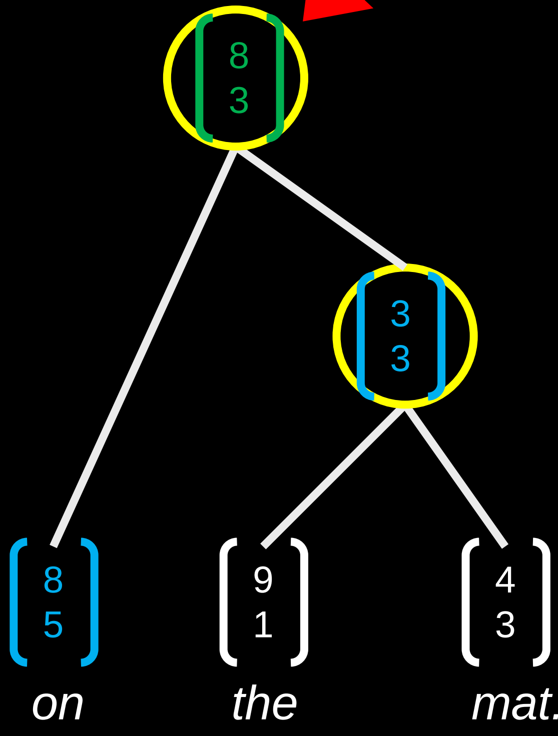
Learning recursive representations

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



Learning recursive representations

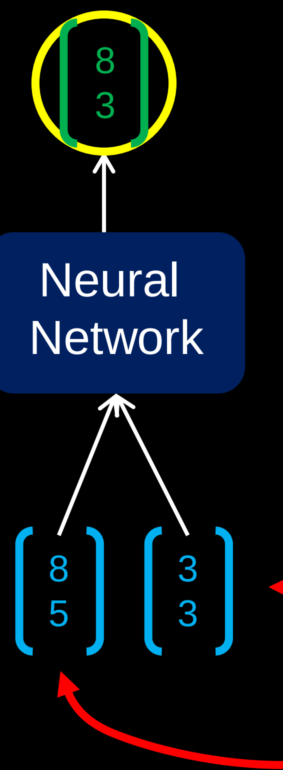
This node's job is to represent *"on the mat."*

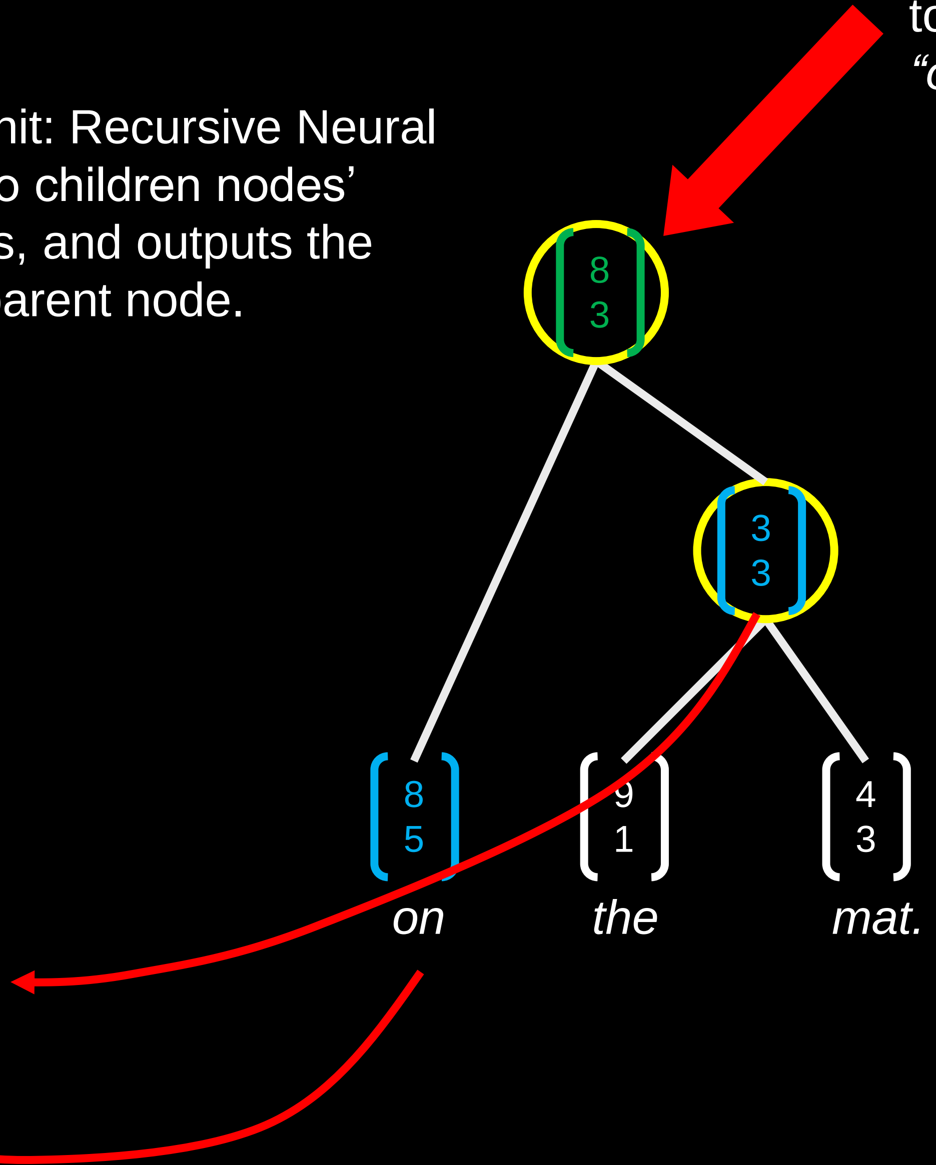


Learning recursive representations

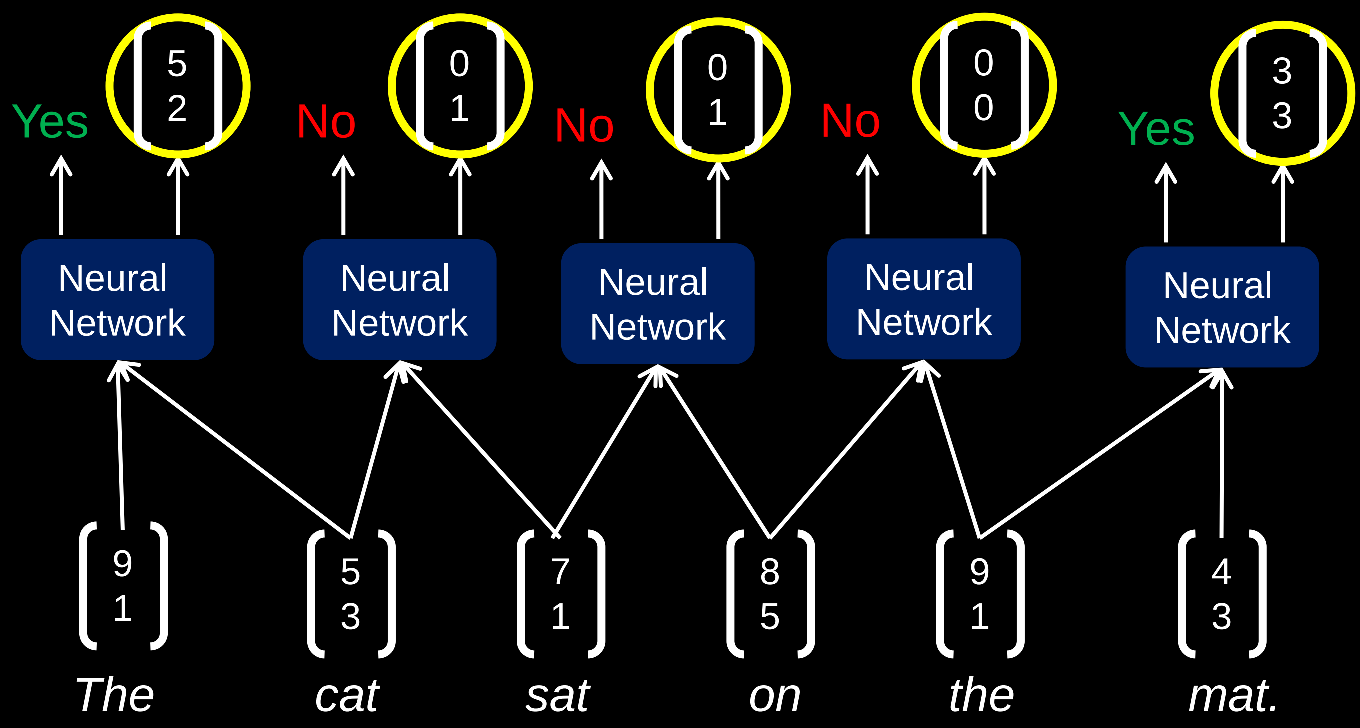
This node's job is to represent *"on the mat."*

Basic computational unit: Recursive Neural Network that inputs two children nodes' feature representations, and outputs the representation of the parent node.



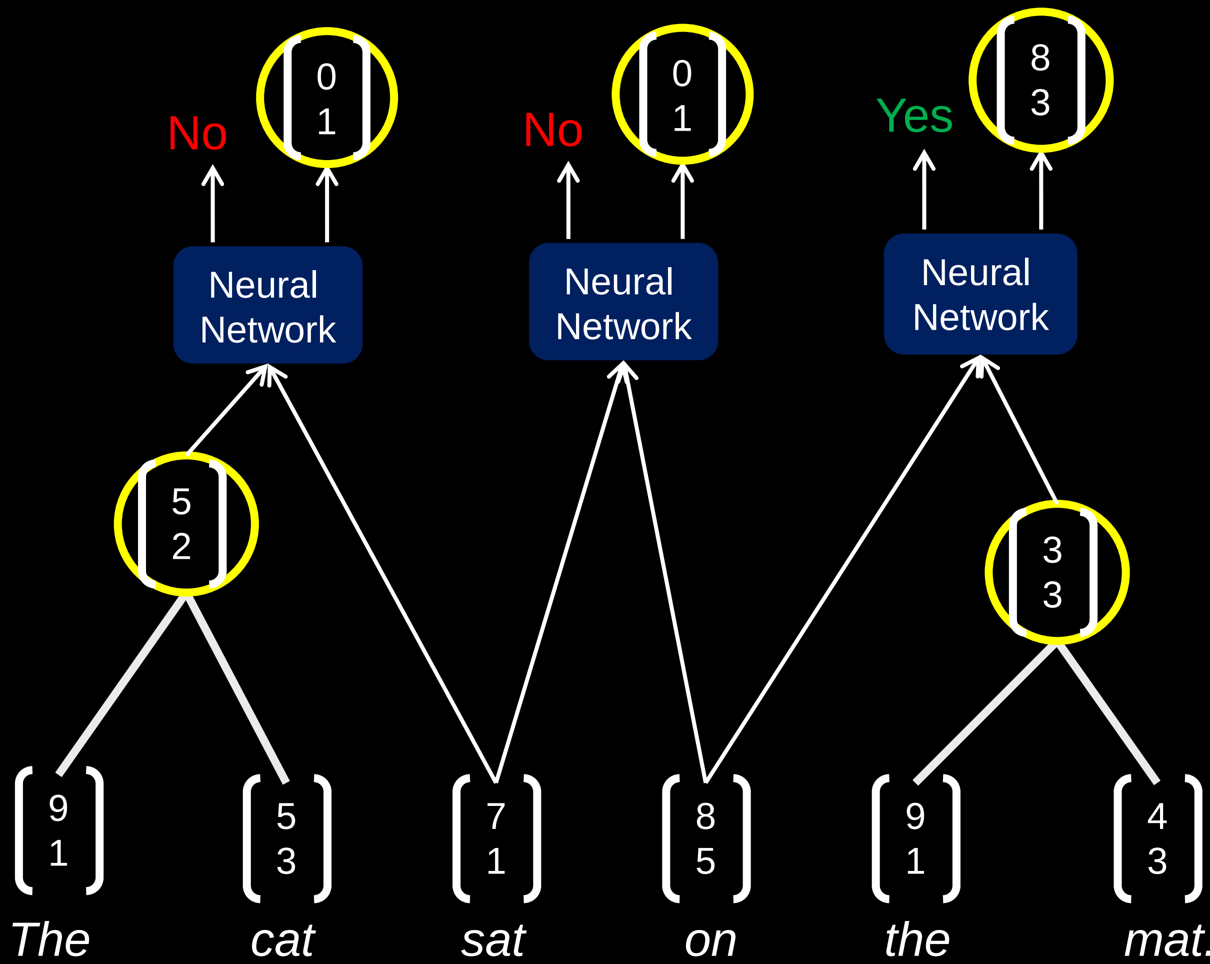


Parsing a sentence



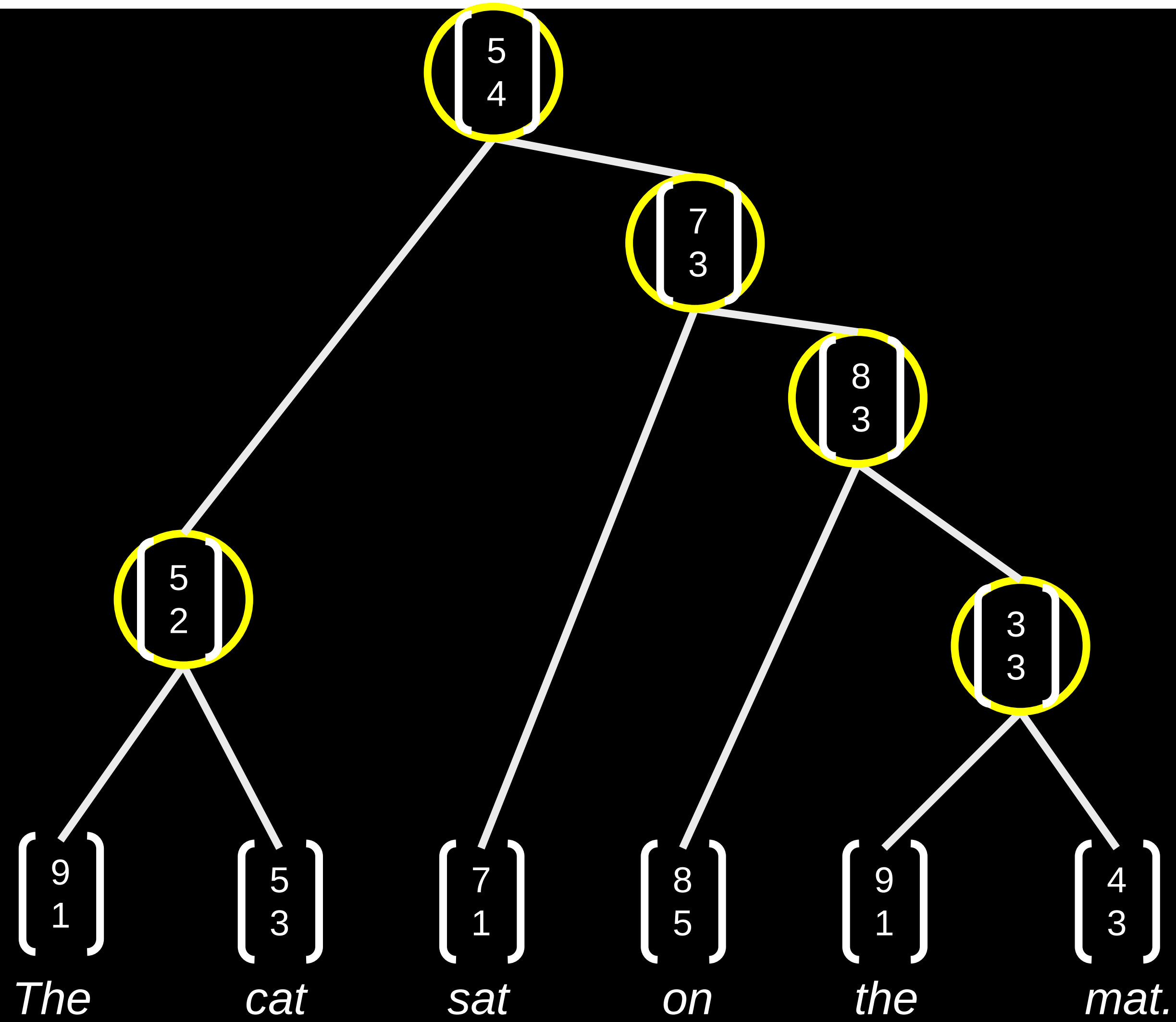
[Socher, Manning & Ng]

Parsing a sentence



[Socher, Manning & Ng]

Parsing a sentence



[Socher, Manning & Ng]

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 ... 2. BSN currently has 2.2 million common shares outstanding 3. This is panic buying 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 2. We have sufficient cash flow to handle that, he said 3. Currently, average pay for machinists is 22.22 an hour, Boeing said 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 2. Commerzbank gained 2 to 222.2 3. Paris loved her at first sight 4. Profits improved across Hess's businesses
Unknown words which are cities	Columbia , S.C	1. Greenville , Miss 2. UNK , Md

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 2. Phoenix declined to comment 3. Campeau declined to comment 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 2. Revenue fell 2 % to 2.22 billion from 2.22 billion 3. Sales rose more than 2 % to 22.2 million from 22.2 million 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 2. It therefore makes no sense for each market to adopt different circuit breakers 3. You can't say the same with black and white 4. I don't think anyone left the place UNK UNK
People in bad situations	We were lucky	1. It was chaotic 2. We were wrong 3. People had died

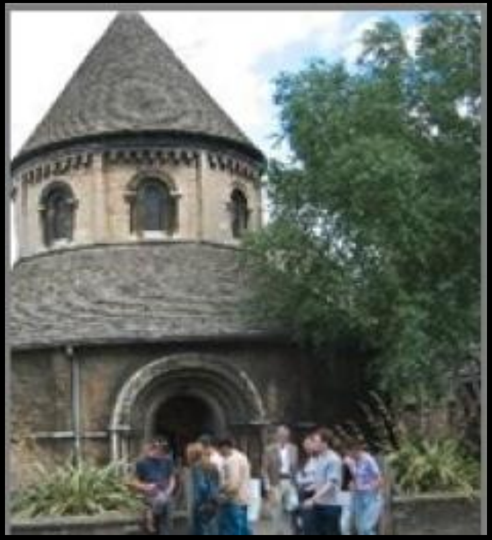
Experiments

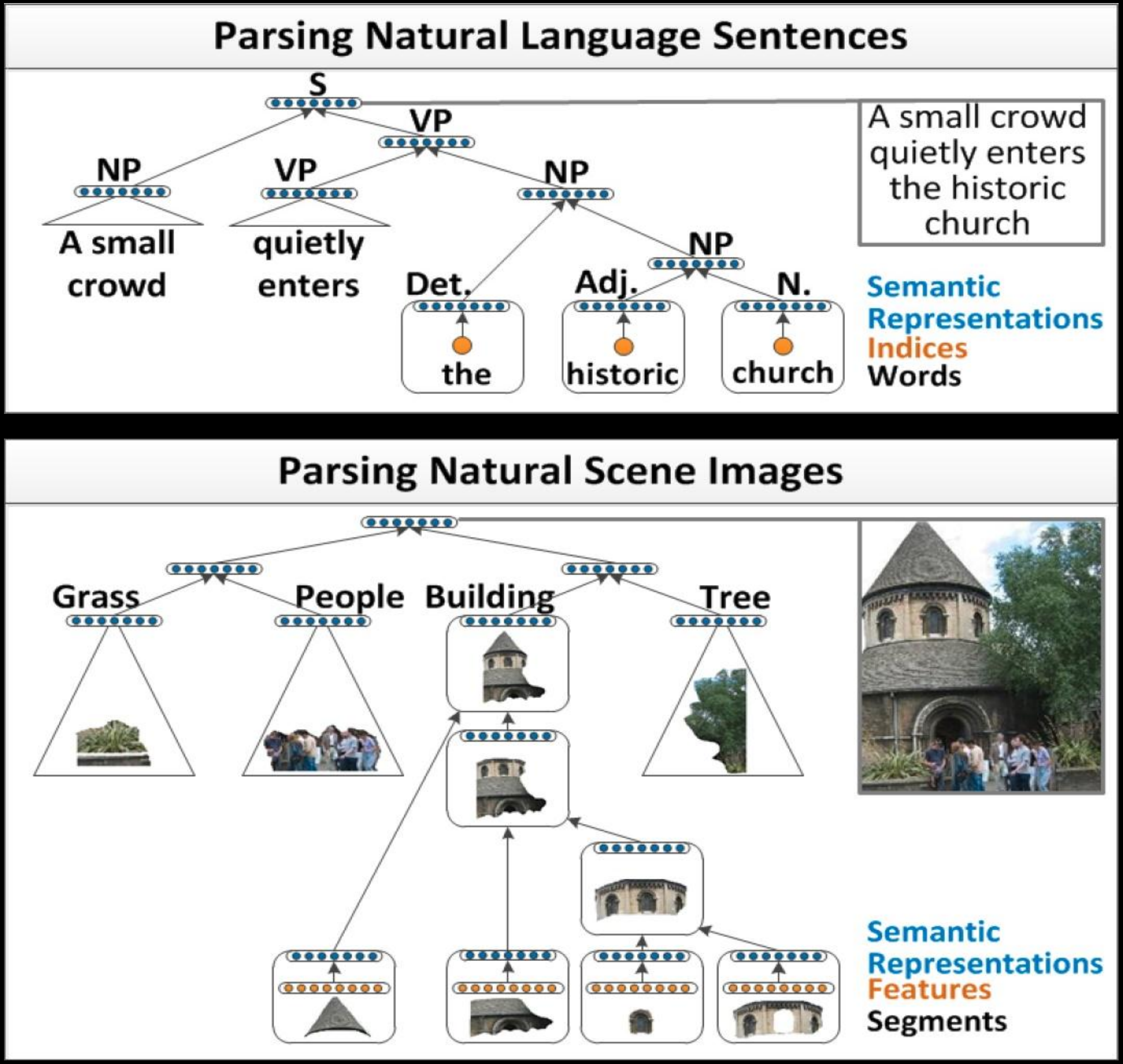
- No linguistic features. Train only using the structure and words of WSJ training trees, and word embeddings from (Collobert & Weston, 2008).
- Parser evaluation dataset: Wall Street Journal (standard splits for training and development testing).

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

Parsing sentences and parsing images

A small crowd quietly enters the historic church.





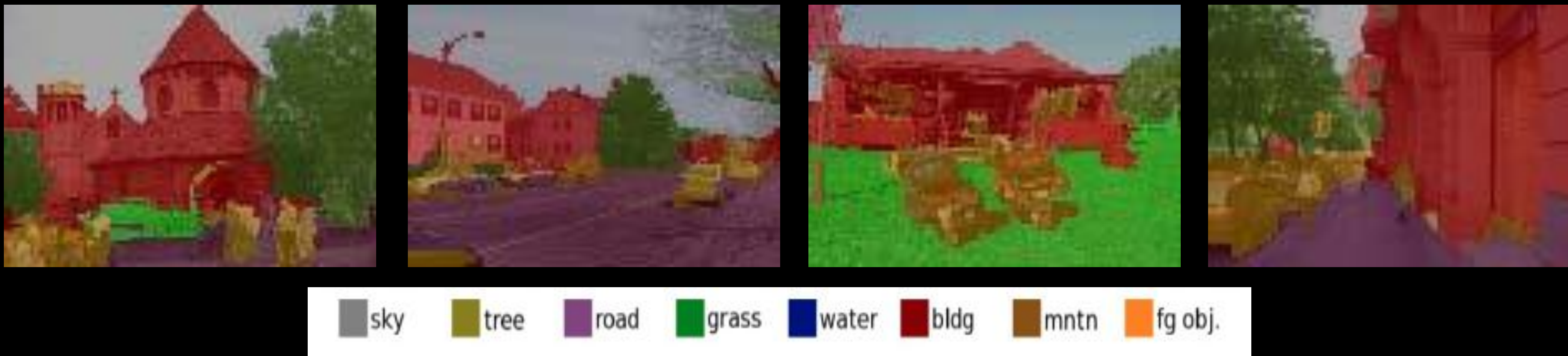
Each node in the hierarchy has a “feature vector” representation.

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
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
Feature learning (out method)	**76.7**

Weaknesses & Criticisms

Weaknesses & Criticisms

- You're learning everything. It's better to encode prior knowledge about structure of images (or audio, or text).

  A: Wasn't there a similar machine learning vs. linguists debate in NLP ~20 years ago….

- Unsupervised feature learning cannot currently do X, where X is:

  ~~Go beyond Gabor (1 layer) features.~~
  ~~Work on temporal data (video).~~
  ~~Learn hierarchical representations (compositional semantics).~~
  ~~Get state-of-the-art in activity recognition.~~
  ~~Get state-of-the-art on image classification.~~
  Get state-of-the-art on object detection.
  Learn variable-size representations.

  A: Many of these were true, but not anymore (were not fundamental weaknesses). There's still work to be done though!

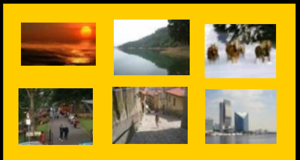
- We don't understand the learned features.

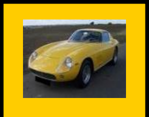
  A: True. Though many vision features are also not really human-understandable (e.g, concatenations/combinations of different features).

Conclusion

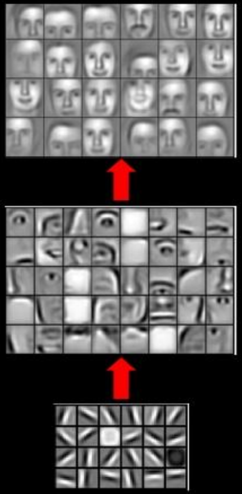
Unsupervised feature learning summary

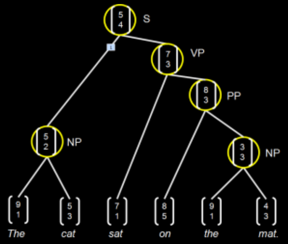
- Unsupervised feature learning.
- Lets learn rather than manually design our features.
- Discover the fundamental computational principles that underlie perception?
- Sparse coding and deep versions very successful on vision and audio tasks. Other variants for learning recursive representations.
- Online tutorial for applying algorithms: http://ufldl.stanford.edu/wiki, or email me.


Natural scenes


Car


Motorcycle



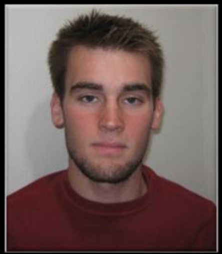


Thanks to:


Adam Coates


Quoc Le


Honglak Lee


Andrew Maas


Chris Manning


Jiquan Ngiam


Andrew Saxe


Richard Socher