Machine Learning and AI via Brain simulations

Andrew Ng

This talk

The idea of "deep learning." Using brain simulations, hope to:
- Make learning algorithms much better and easier to use.
- Make revolutionary advances in machine learning and AI.

Ideas are not only mine; vision shared with many researchers:

E.g., Samy Bengio, Yoshua Bengio, Tom Dean, Nando de Freitas, Jeff Hawkins, Geoff Hinton, Yann LeCun, Honglak Lee, Tommy Poggio, Dawn Song, Josh Tenenbaum, Kai Yu, Jason Weston, ….

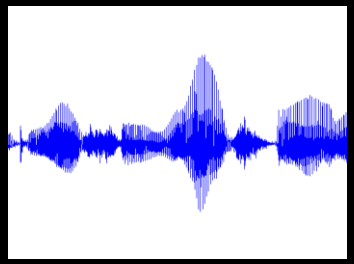
I believe this is our best shot at progress towards real AI.

What do we want computers to do with our data?

Images/video

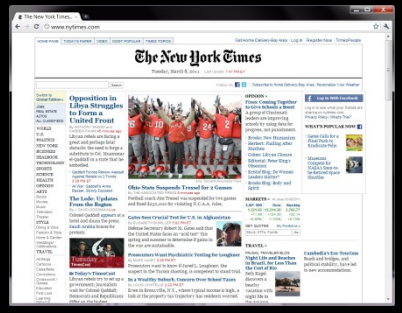
Label: “Motorcycle”
Recognize location, activities, …

Audio



Speech recognition
Music classification
Speaker identification
…

Text



Spam classification
Web search
Machine translation
…

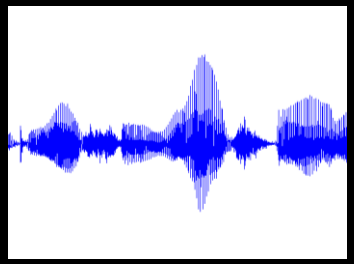
Computer vision is hard!

What do we want computers to do with our data?

Images/video

Label: “Motorcycle”
Recognize location, activities, …

Audio



Speech recognition
Music classification
Speaker identification
…

Text

Spam classification
Web search
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”

Machine learning for image classification





Motorcycles



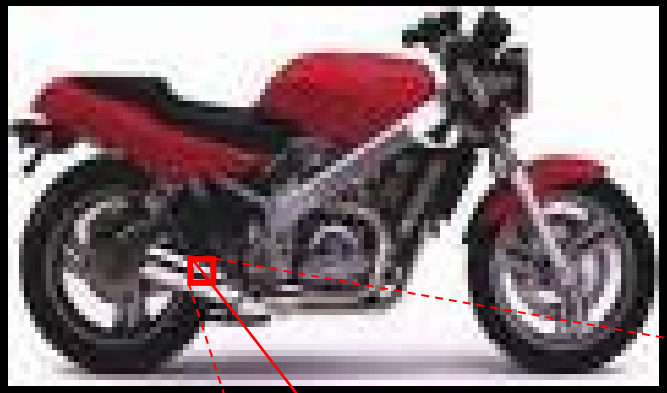



Not a motorcycle

Testing:
What is this?

Why is this hard?

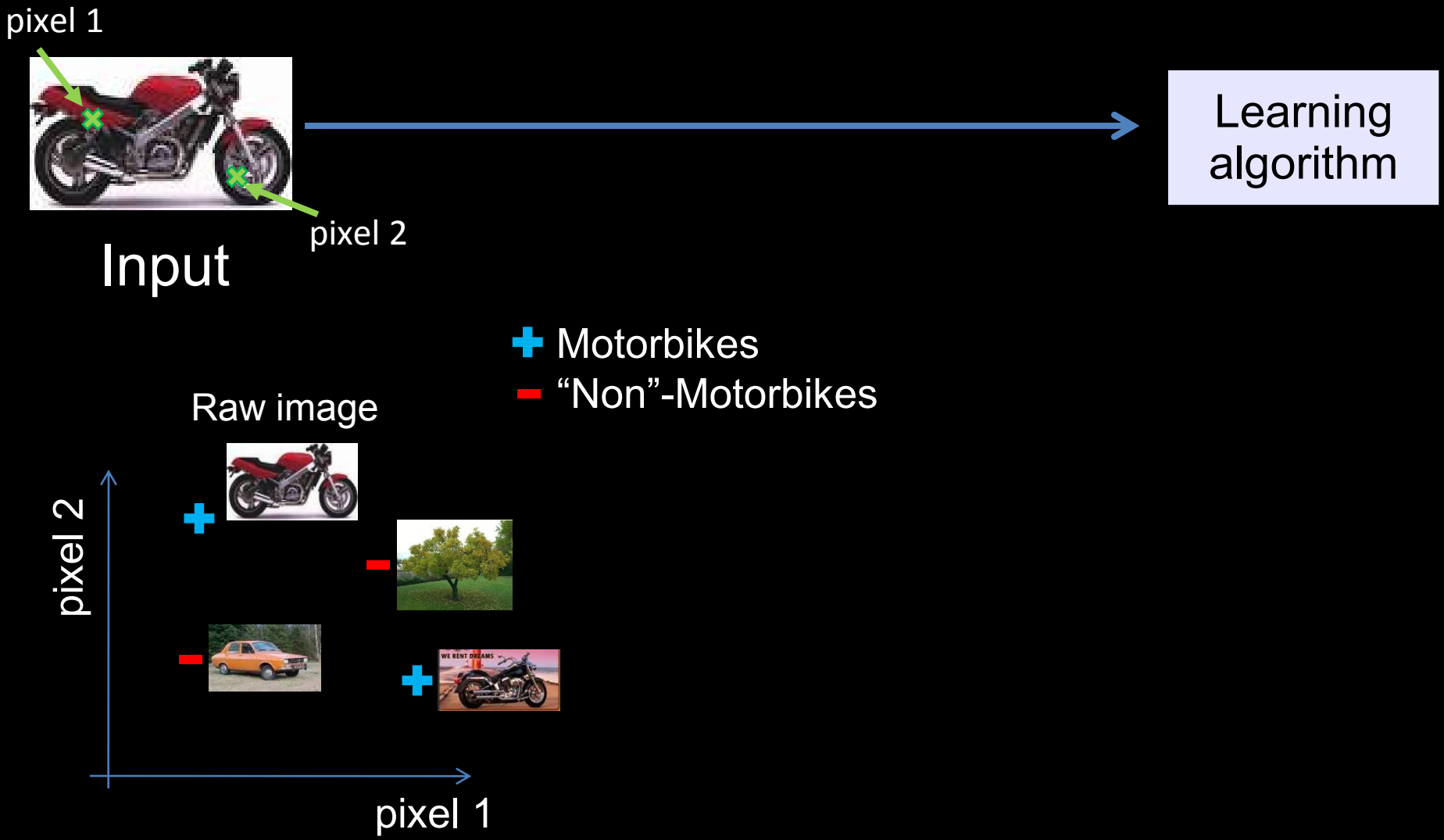
You see this:



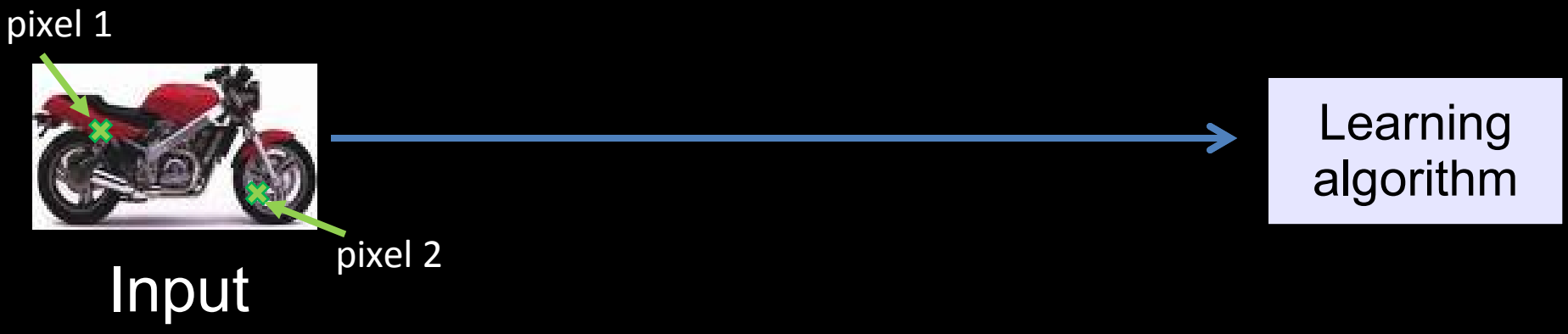
But the camera sees this:

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

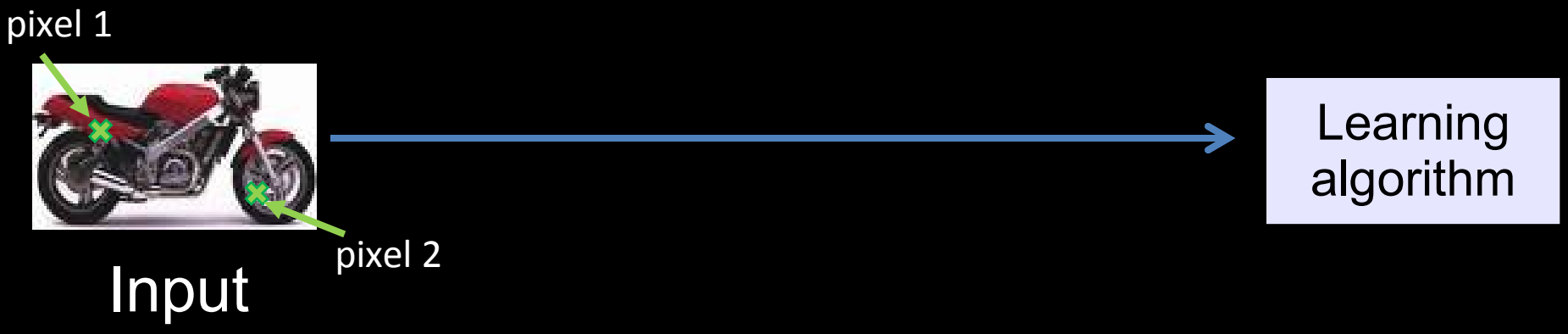
Machine learning and feature representations



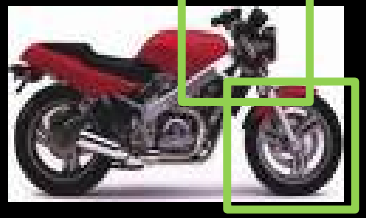
Machine learning and feature representations



Machine learning and feature representations



What we want

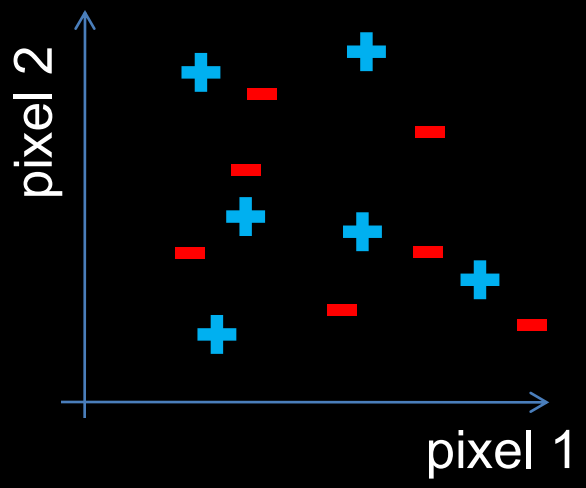


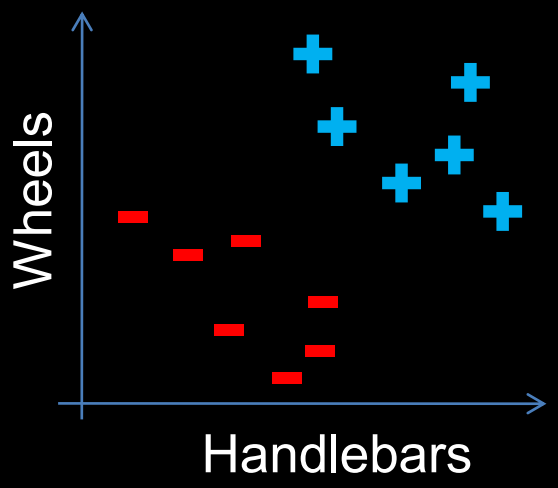
Input

Feature representation

Learning algorithm

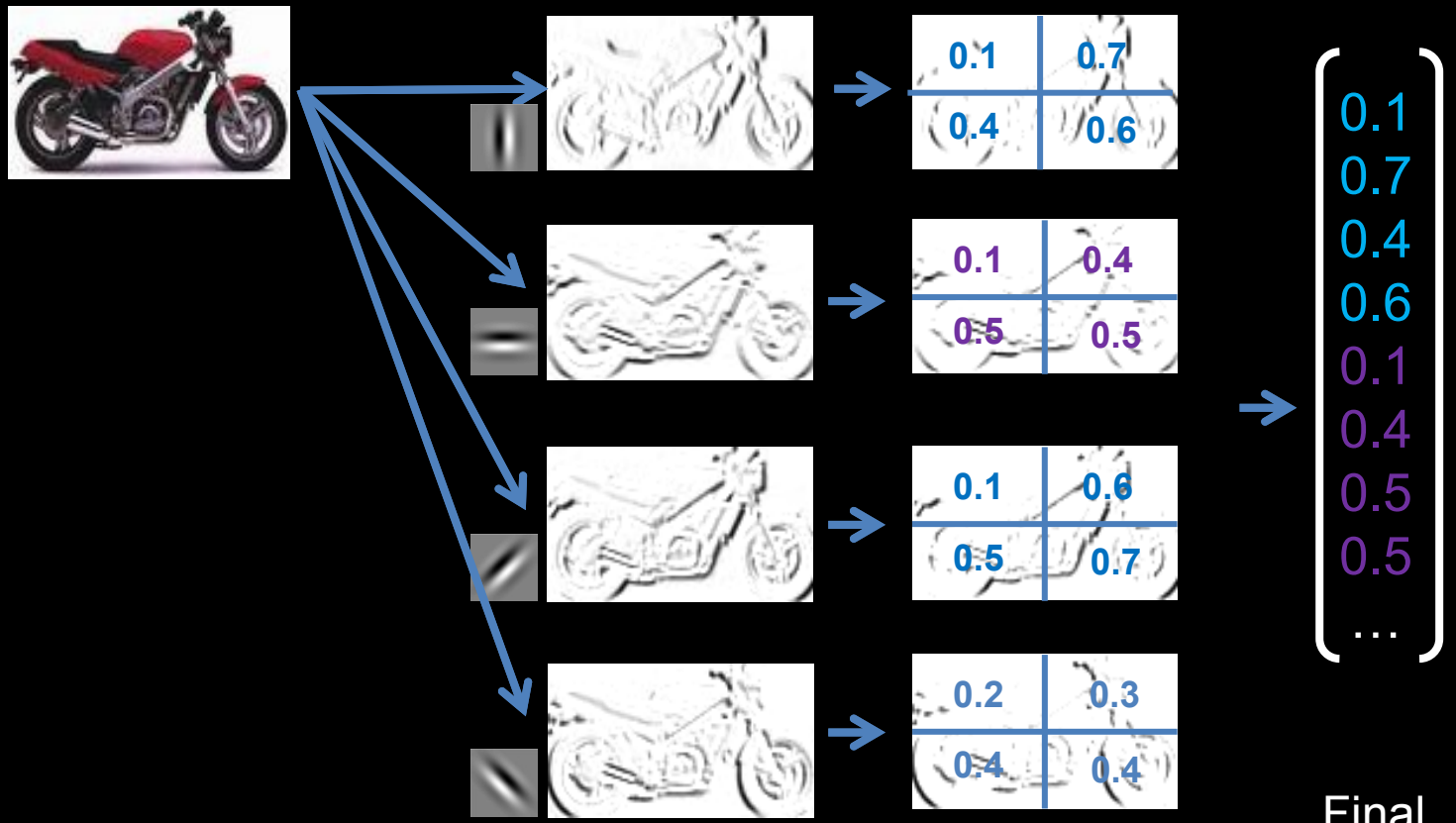
E.g., Does it have Handlebars? Wheels?





Computing features in computer vision

But… we don't have a handlebars detector. So, researchers try to hand-design features to capture various statistical properties of the image.



How is computer perception done?

Images/video

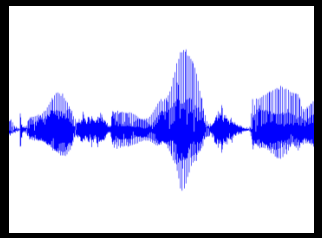
Image

Vision features

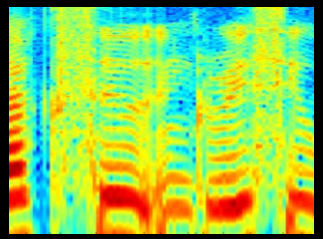
Recognition

"Motorcycle"

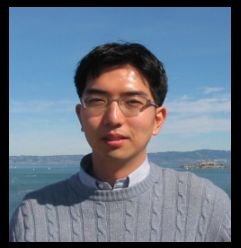
Audio



Audio

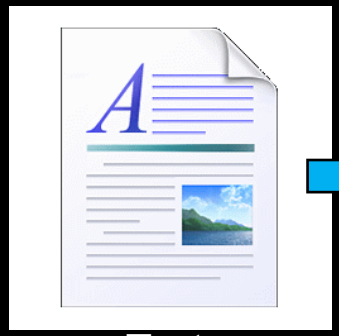


Audio features

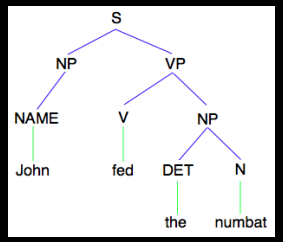


Speaker ID

Text



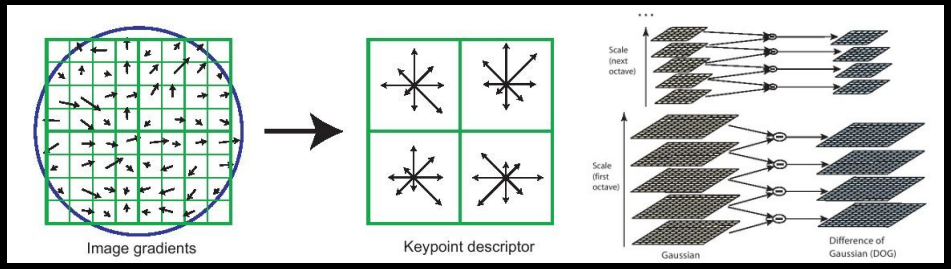
Text



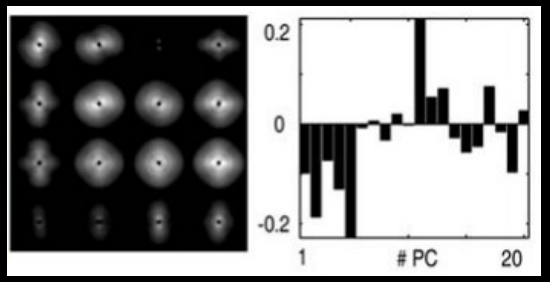
Text features

Information retrieval,
Machine translation,
Text classification,
etc.

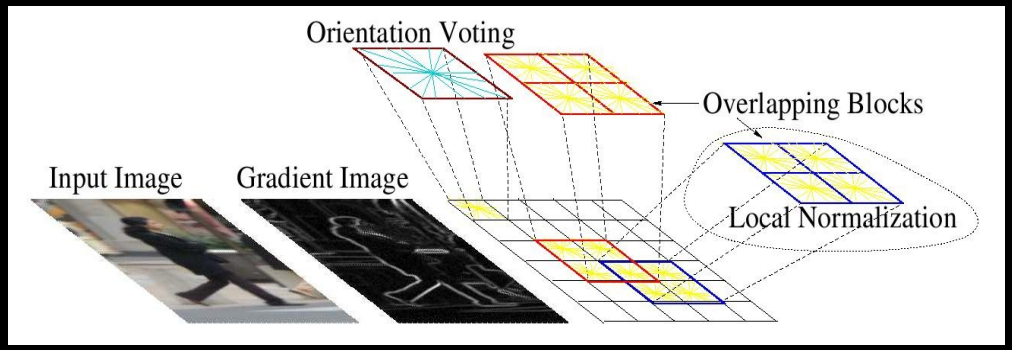
Computer vision features



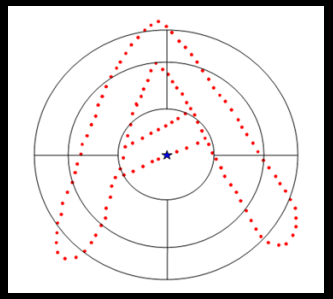
SIFT



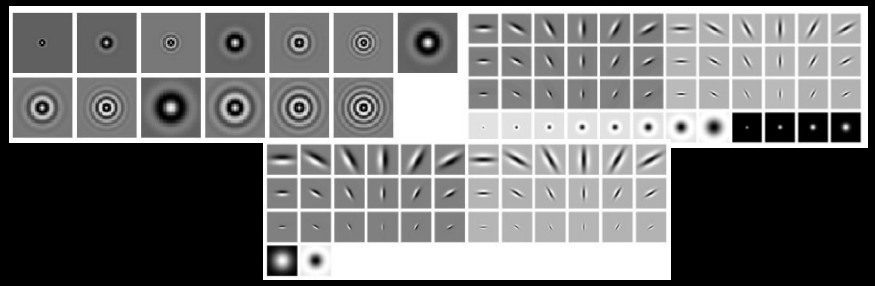
GIST



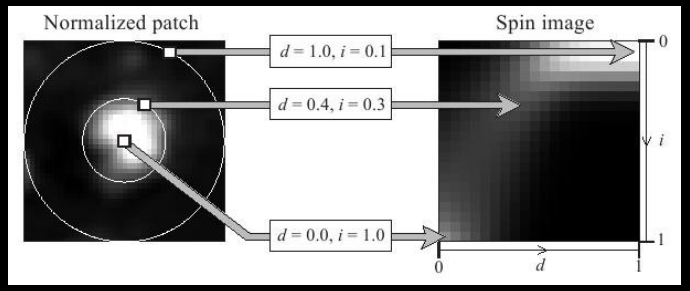
HoG



Shape context

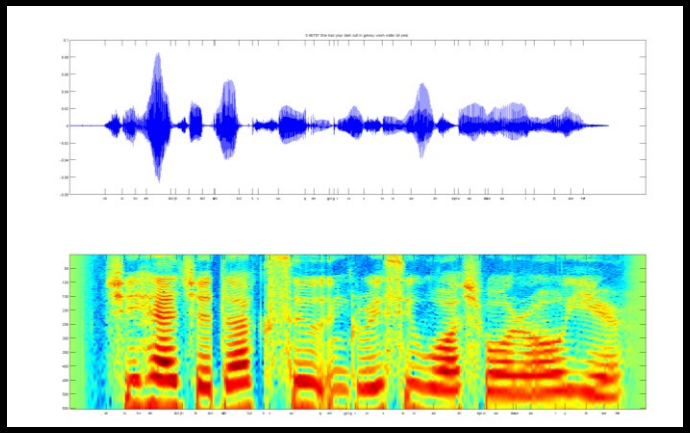


Textons

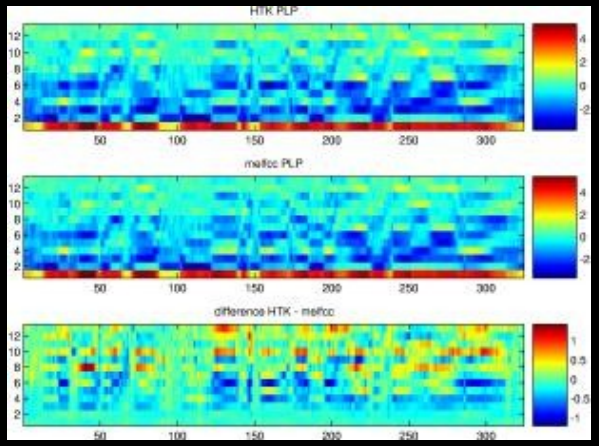


Spin image

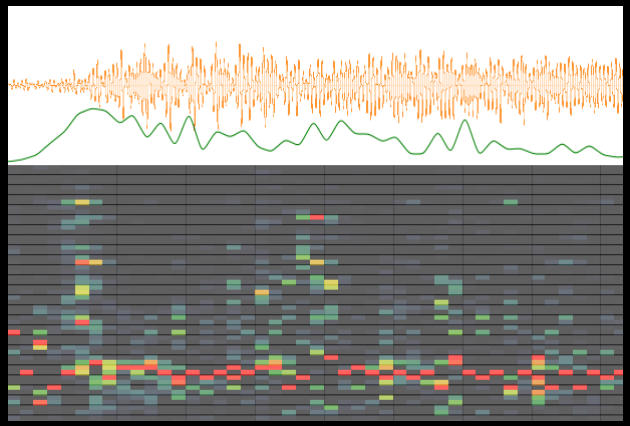
Audio features



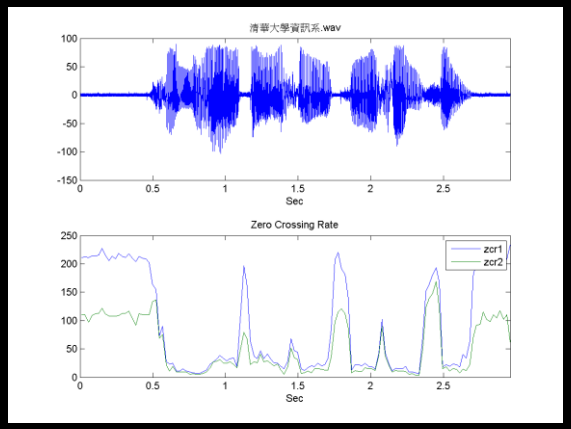
Spectrogram



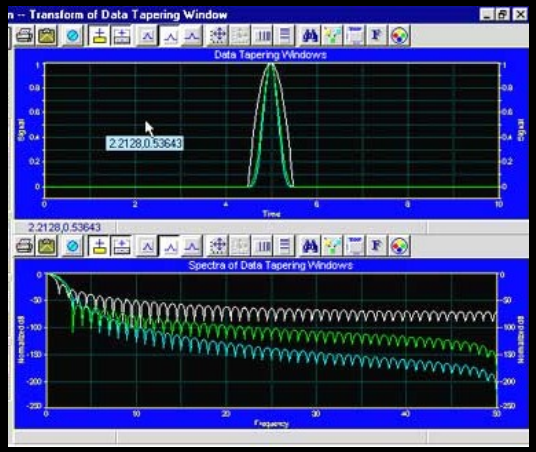
MFCC



Flux

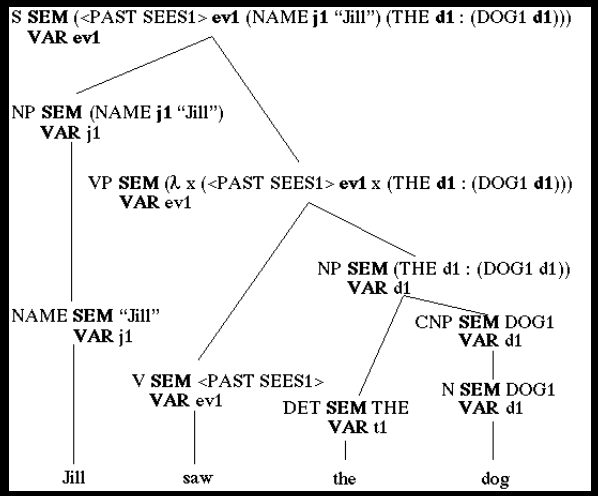


ZCR

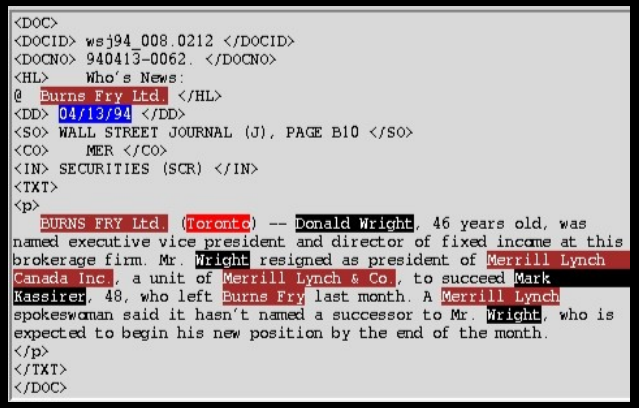


Rolloff

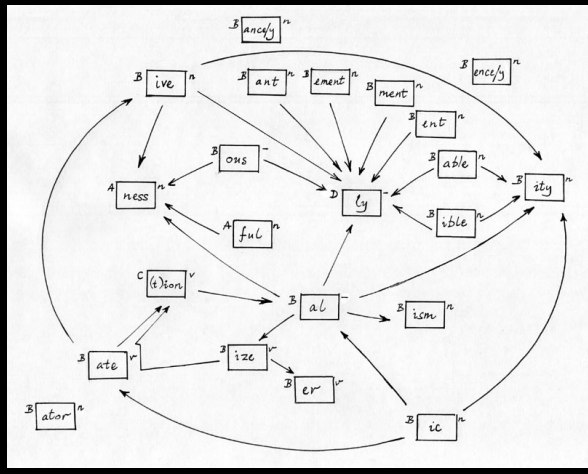
NLP features



Pars



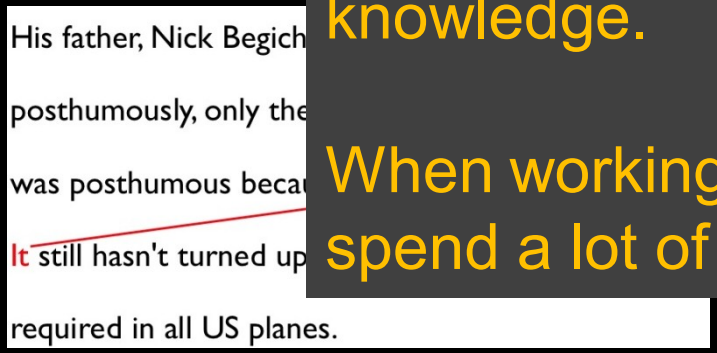
Named entity recognition



Stemming

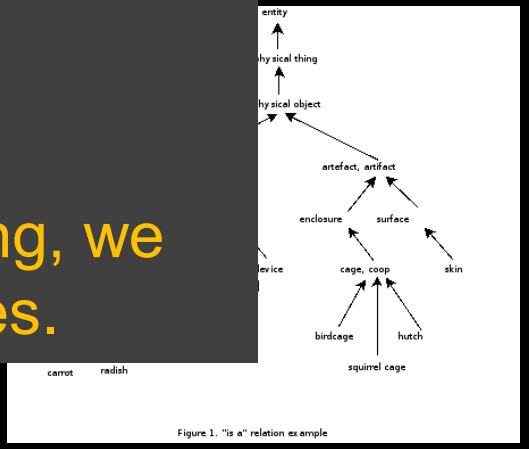
Coming up with features is difficult, time-consuming, requires expert domain knowledge.

When working applications of learning, we spend a lot of time tuning the features.



Anaphora

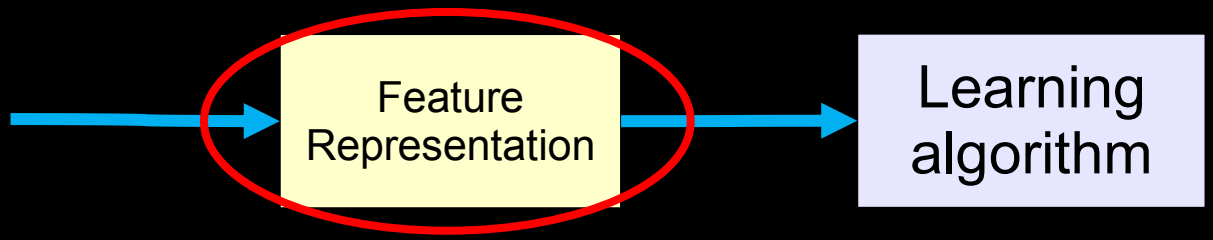
Part of speech



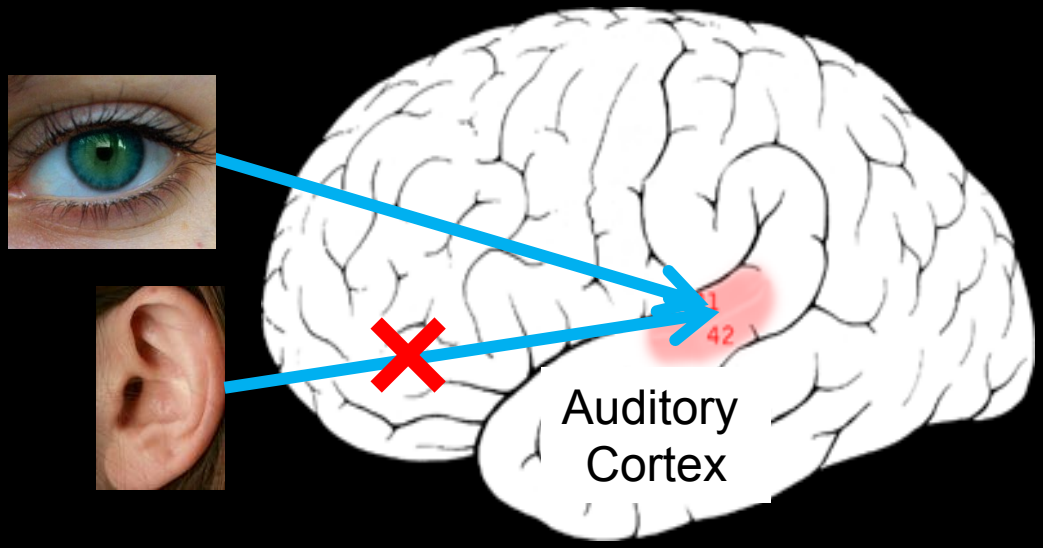
Ontologies (WordNet)

Feature representations

Input



Sensor representation in the brain



Auditory cortex learns to see.

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

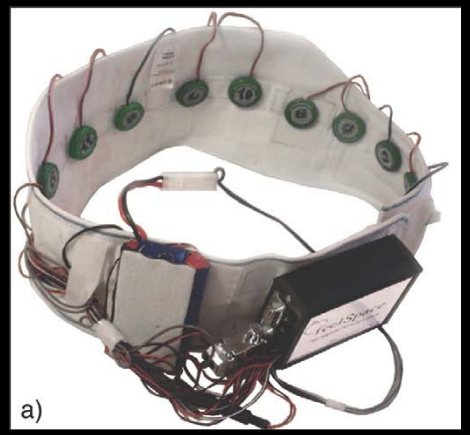
Seeing with your tongue

Human echolocation (sonar)

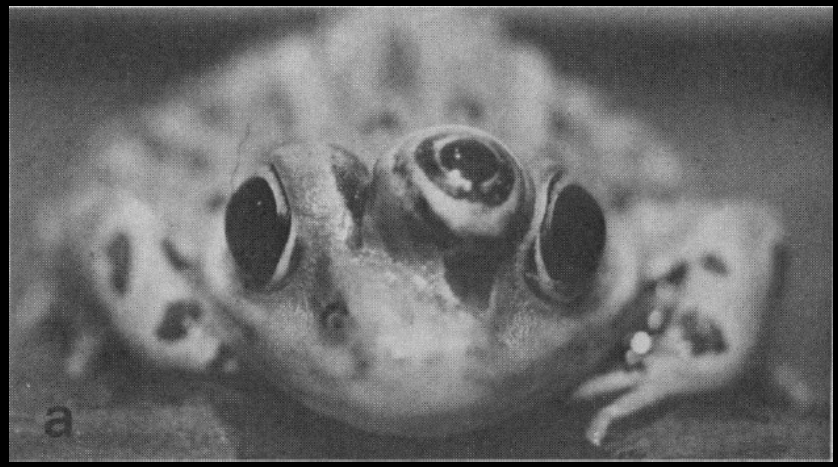
[Roe et al., 1992; BrainPort; Welsh & Blasch, 1997]

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]

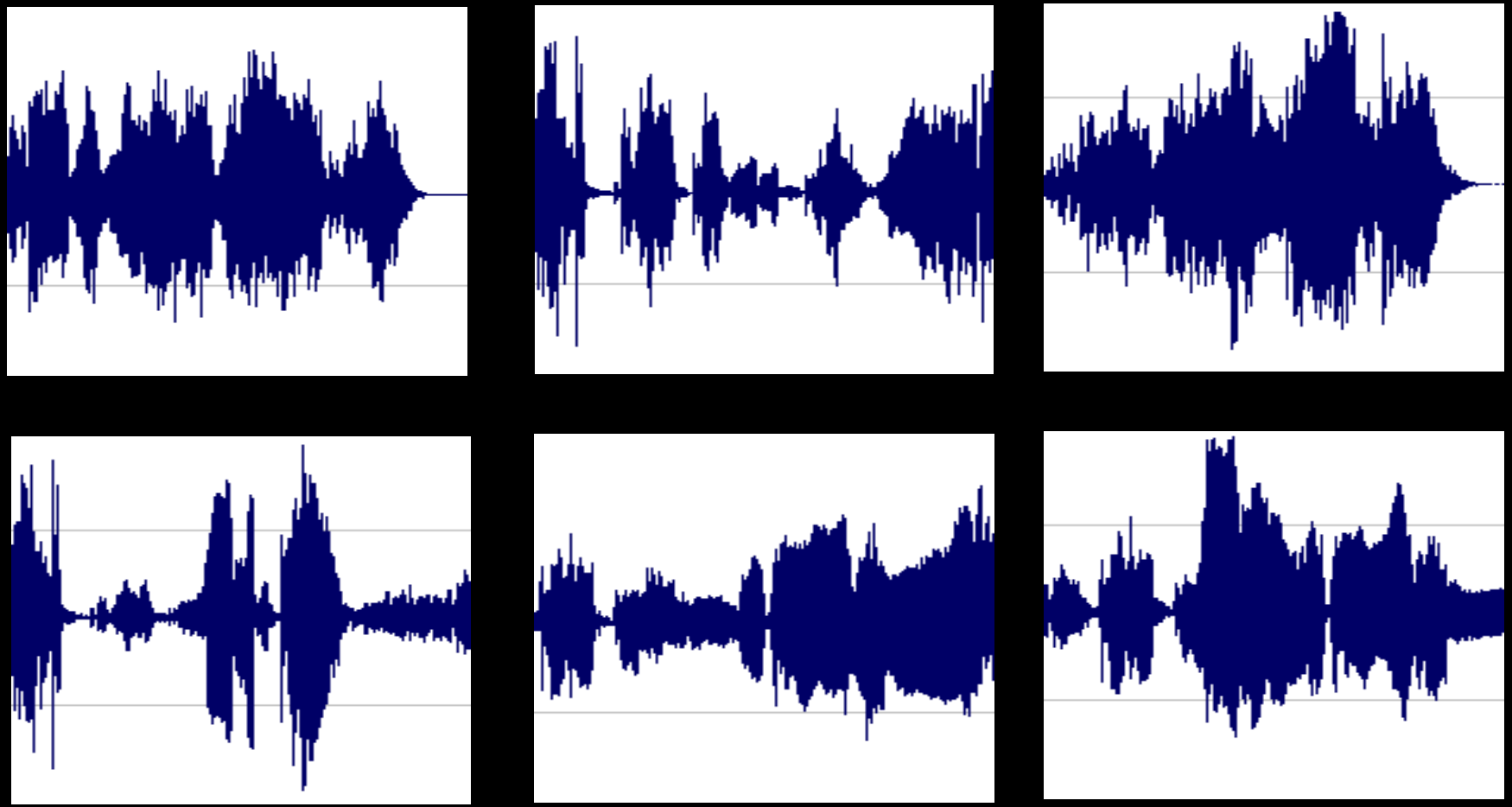
Learning input representations






Rather than hand-engineering features, can we automatically *learn* a better way to represent images than pixels.

Learning input representations



Automatically learn a feature representation for audio.

Feature learning problem

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

$$\begin{bmatrix} 255 \\ 98 \\ 93 \\ 87 \\ 89 \\ 91 \\ 48 \\ \dots \end{bmatrix}$$

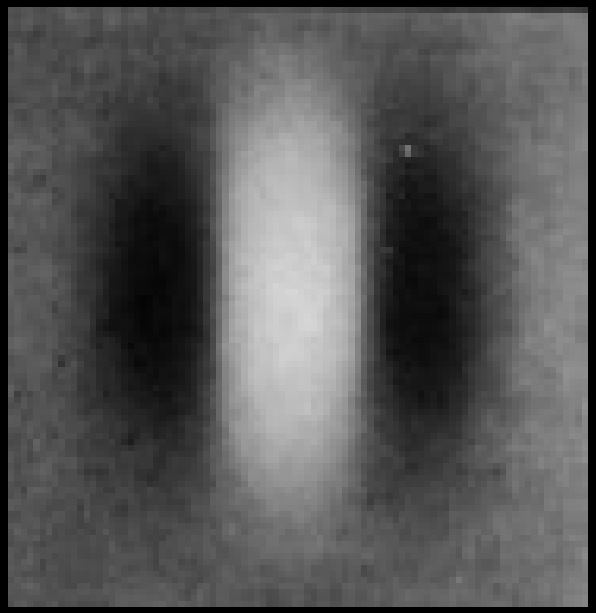
- Problem: Can we find a learn a better feature vector to represent this?

How does the brain process images?

Looking to the brain for inspiration:

The first visual processing step in the brain (primary visual cortex, area V1) in looks for “edges” in images.

Neuron #1 of visual cortex (model)



Neuron #2 of visual cortex (model)

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 $R^{n \times n}$)

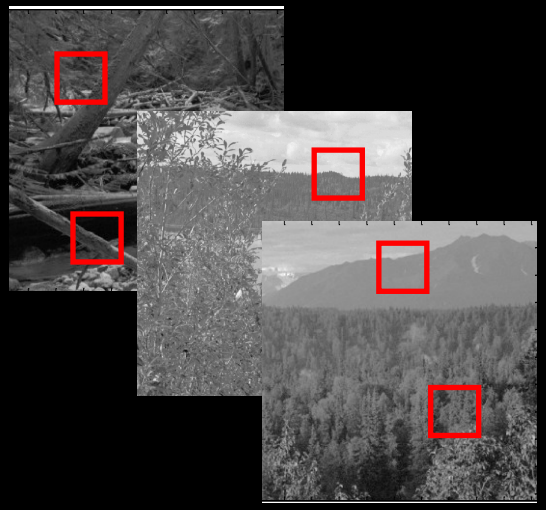
Learn: Dictionary of bases $\phi_1, \phi_2, \ldots, \phi_k$ (also $R^{n \times n}$), so that each input x can be approximately decomposed as:

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

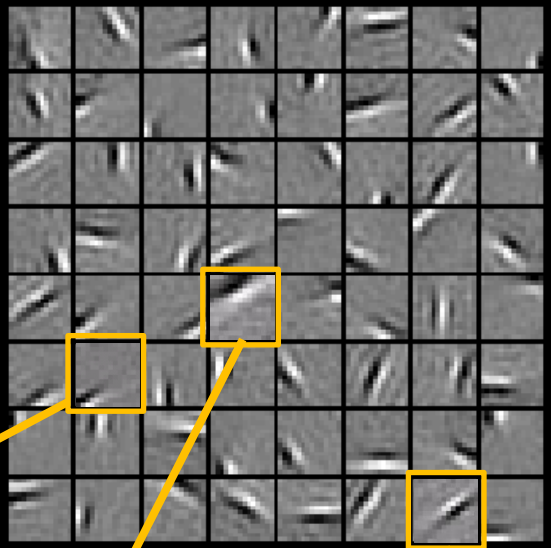
s.t. a_j's are mostly zero ("sparse")

Sparse coding illustration

Natural Images



Learned bases (ϕ_1 , …, ϕ_{64}): "Edges"



Test example

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

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

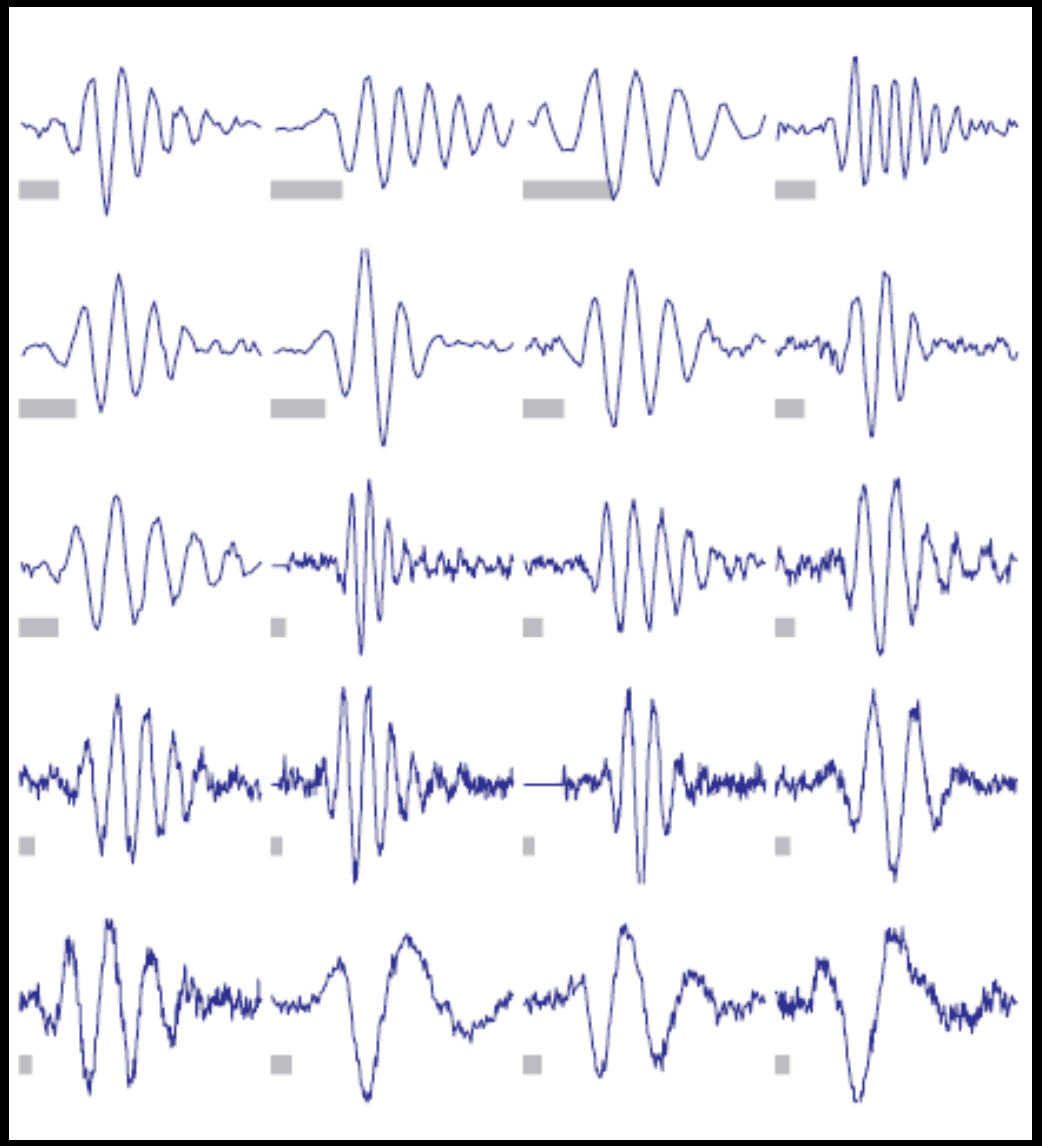
$[a_1, \ldots, a_{64}] = [0, 0, \ldots, 0, \mathbf{0.8}, 0, \ldots, 0, \mathbf{0.3}, 0, \ldots, 0, \mathbf{0.5}, 0]$

(feature representation)

Compact & easily interpretable

Sparse coding applied to audio

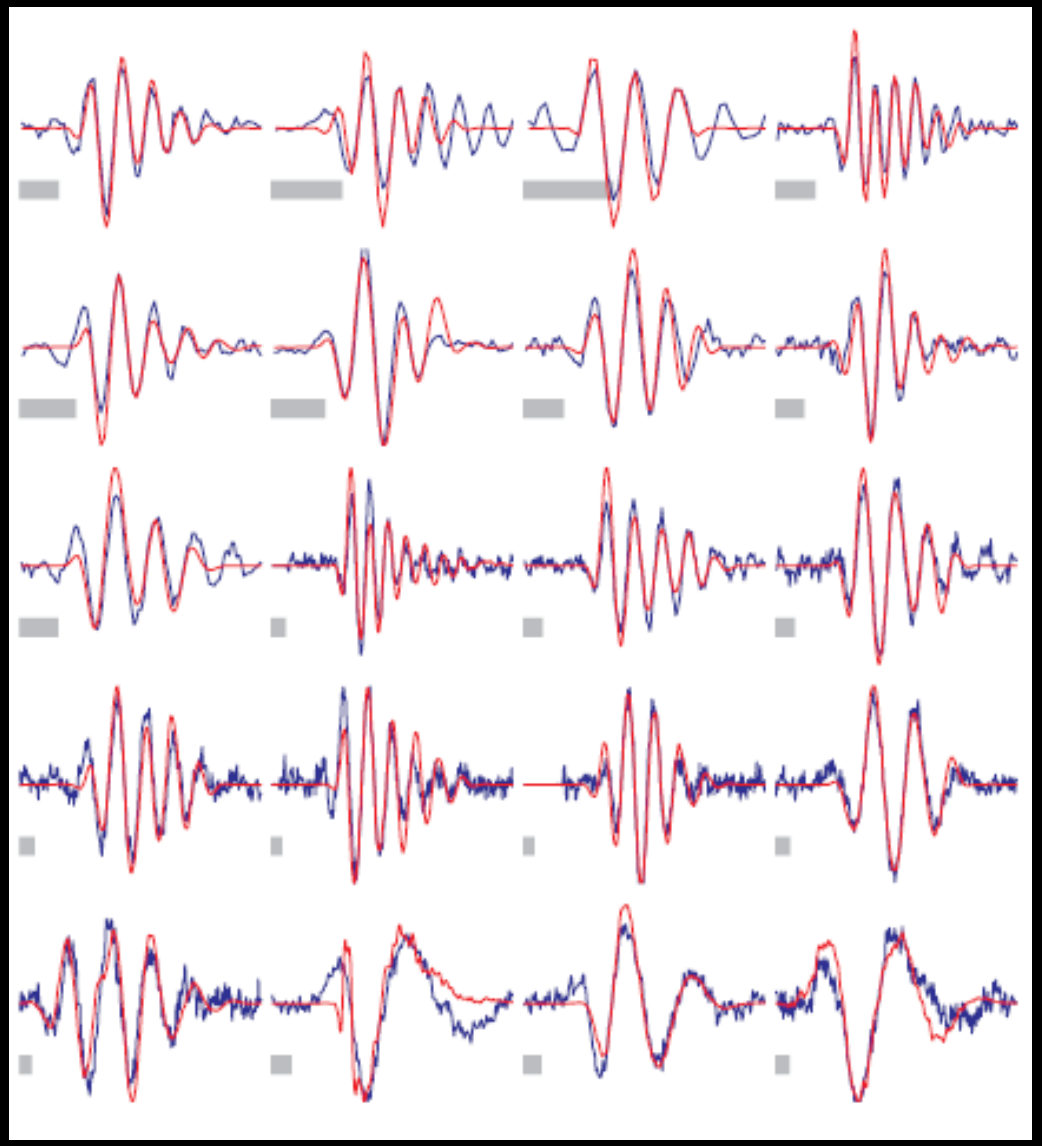
Image shows 20 basis functions learned from unlabeled audio.



[Evan Smith]

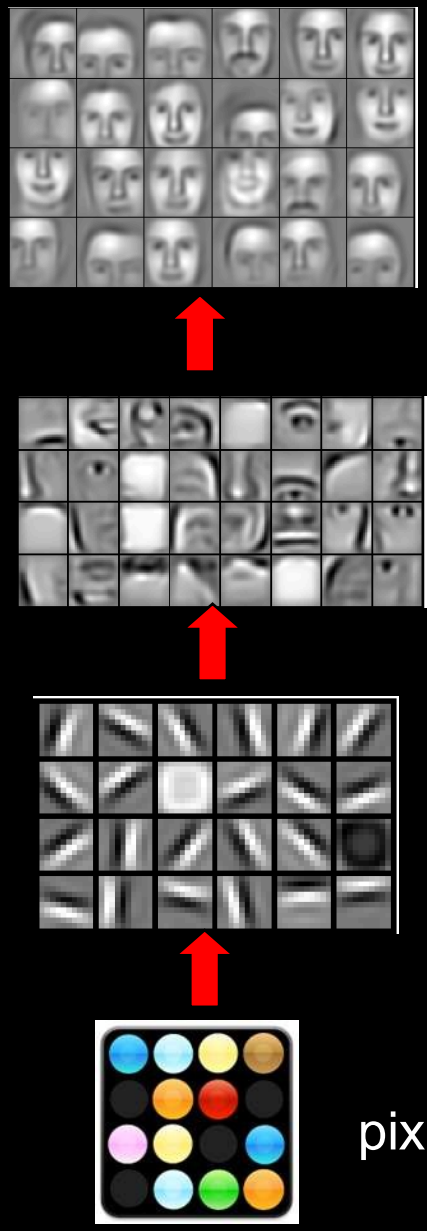
Sparse coding applied to audio

Image shows 20 basis functions learned from unlabeled audio.



[Evan Smith]

Recursive sparse coding: Training on face images



object models

object parts (combination of edges)

edges

pixels

Can we go further?

By recursively applying sparse coding algorithms, get higher-level features.

[Technical details: Sparse autoencoder (Bengio) or Sparse DBN (Hinton).]

Machine learning applications

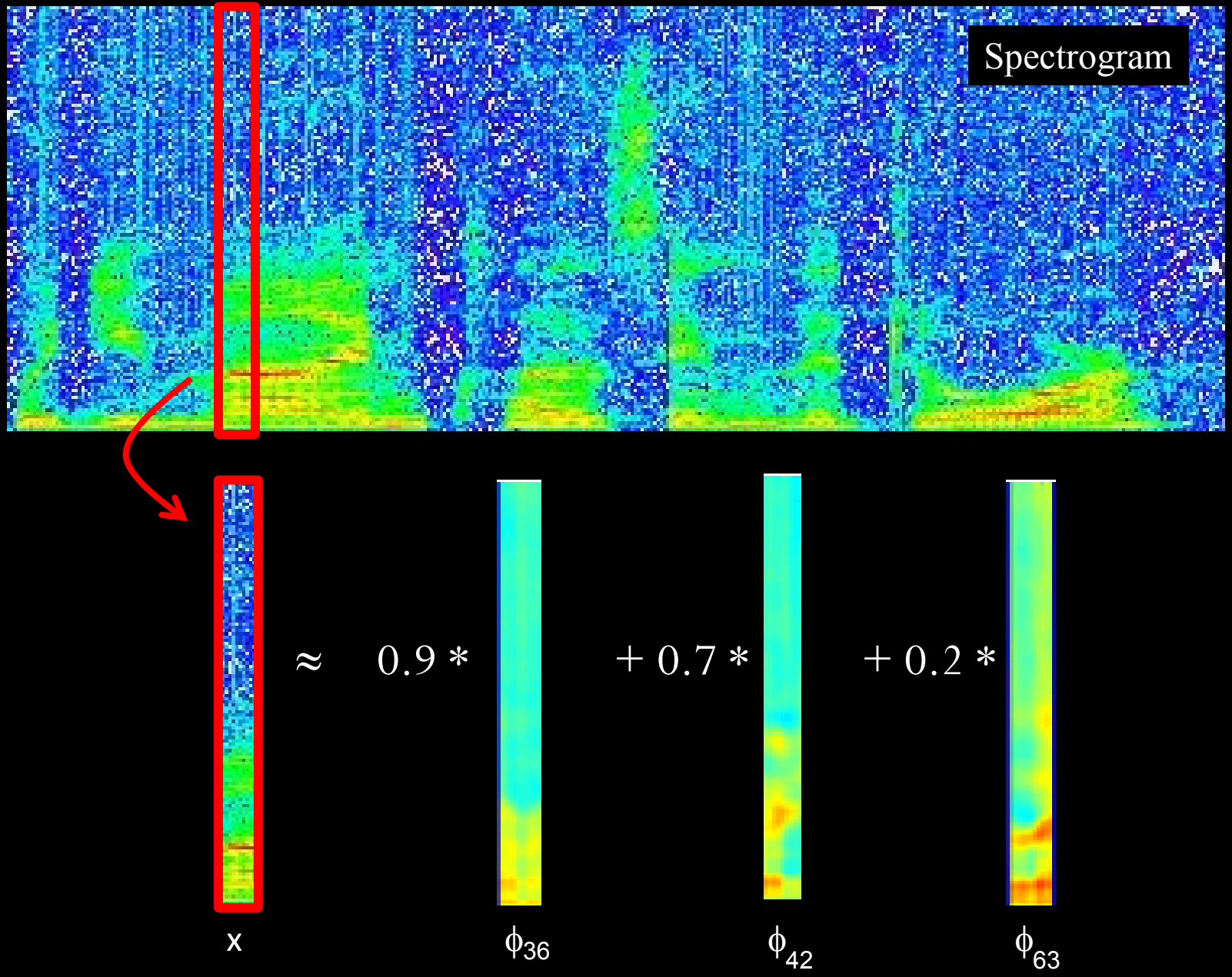
Activity recognition (Hollywood 2 benchmark)

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%**

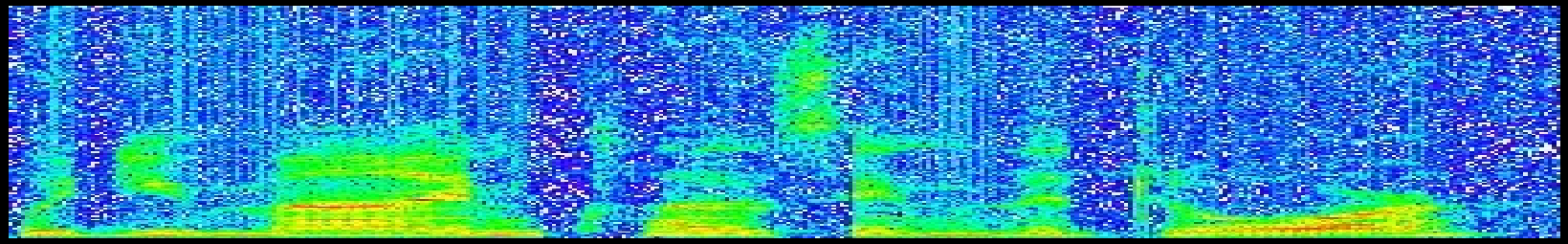
Unsupervised feature learning significantly improves on the previous state-of-the-art.

[Le et al., 2011]

Sparse coding on audio



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.

[Honglak Lee]

State-of-the-art unsupervised feature learning

Audio

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

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

Images

CIFAR Object classification	Accuracy
Prior art (Yu and Zhang, 2010)	74.5%
Feature learning	**80.1%**

NORB Object classification	Accuracy
Prior art (Ranzato et al., 2009)	94.4%
Feature learning	**97.0%**

Video

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

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

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

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

Multimodal (audio/video)

AVLetters Lip reading	Accuracy
Prior art (Zhao et al., 2009)	58.9%
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]

Weaknesses & Criticisms

Weaknesses & Criticisms

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

  A: There was a similar linguists vs. machine learning/IR debate in NLP ~20 years ago….

- Deep 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/audio/text features are also not really human-understandable (e.g, concatenations/combinations of different features). There're also techniques to bound the parts we don't understand.

Technical challenge: Scaling up

Current models

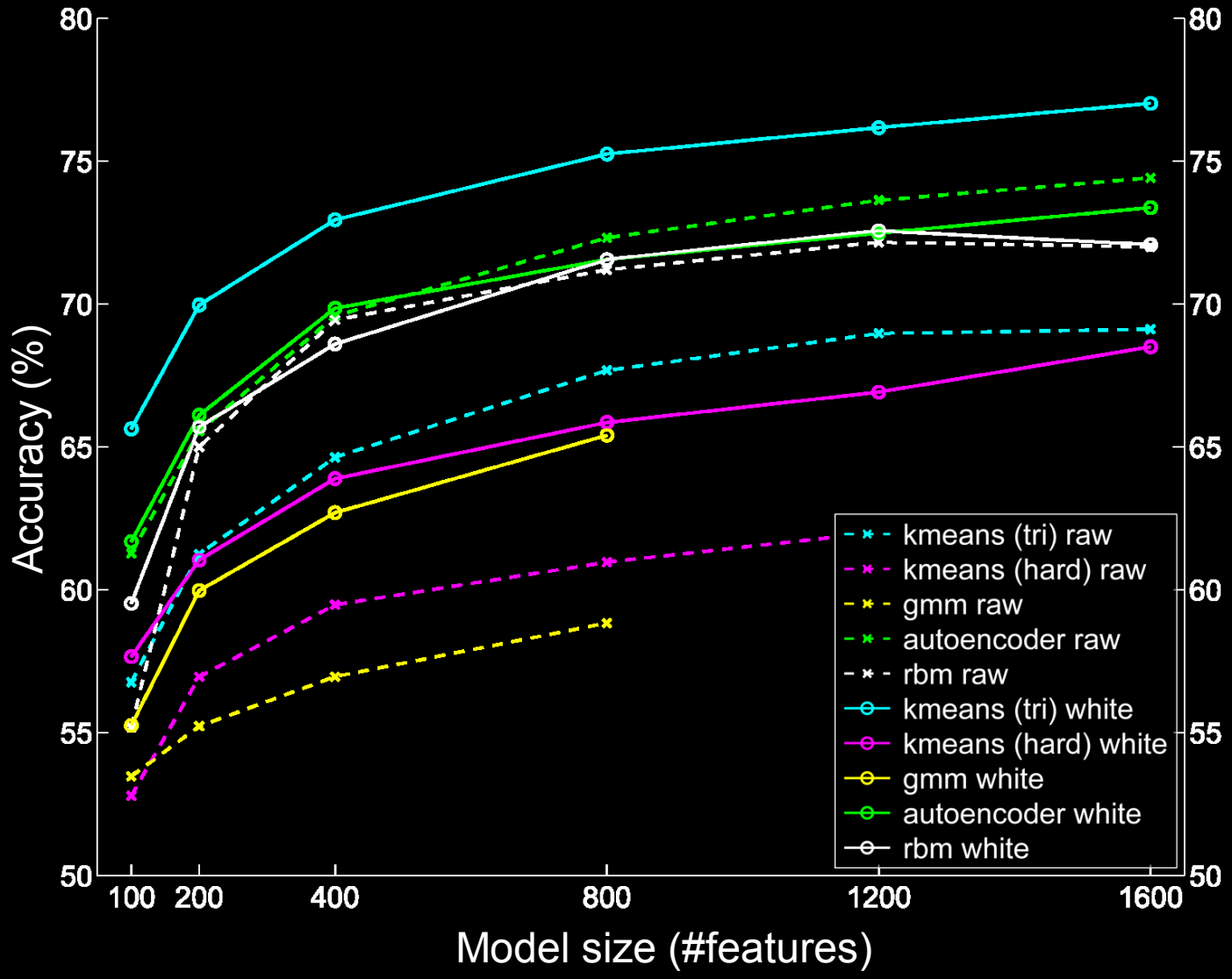
Current models simulate 10^3 - 10^6 neurons.

The brain has 10^{11} neurons.

Larger models:
- Simulate cortical neuronal properties more accurately.
- Work much better on machine learning tasks.

Machine learning application (CIFAR-10)

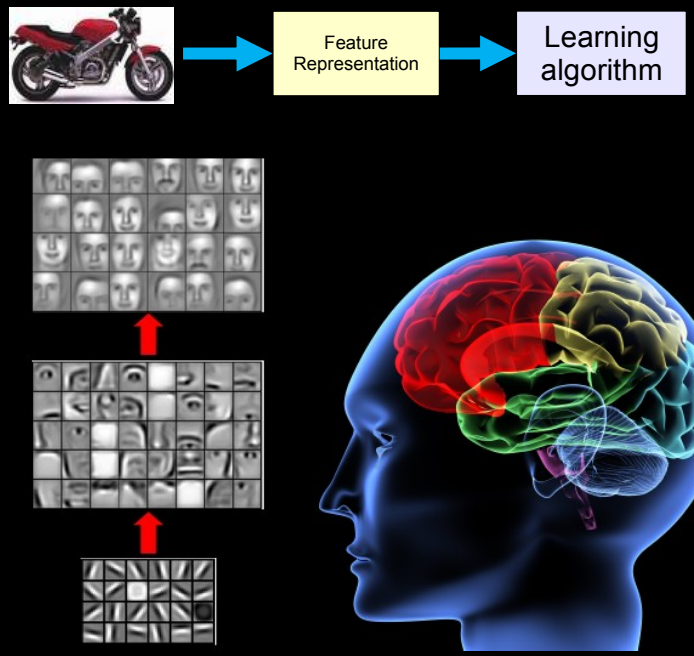
Accuracy accuracy vs. model size (#features).



Summary

Machine Learning & AI Summary

- Lets learn rather than manually design our features.
- Sparse coding and Deep Learning is best method currently for many tasks.
- Discover the fundamental computational principles that underlie perception, make progress on AI.
- Key challenge: Scalability.
- If interested in (i) collaborating on these topics, or (ii) learning more about deep learning, contact me (email ang@cs.stanford.edu). Online tutorials available to Computer forum members.



Thanks to:

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