Deep Learning Fundamentals and Emerging Trends

Deep Learning Successes

Images

Year	Error rate
2011	26.2%
2012	15.3%
2015	4.8%
Human	5.1%

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

Speech

Year	Error rate
2004	15%
2011	12%
2015	8%
Human	4%

Word error rate on the Switchboard task (IBM)

SyncWords
Captions Automated

“Deep Learning” – Deep Buzzword

- Used to be called “neural networks” in the 90s, before they were over-hyped and rejected by a research community who hates hype.
- Old buzzwords get replaced by new buzzwords.
  Now called, “deep learning,” “computational networks,” etc.
- (Although, some “deep” algorithms are extensions of non-NN models.)

What's different now?

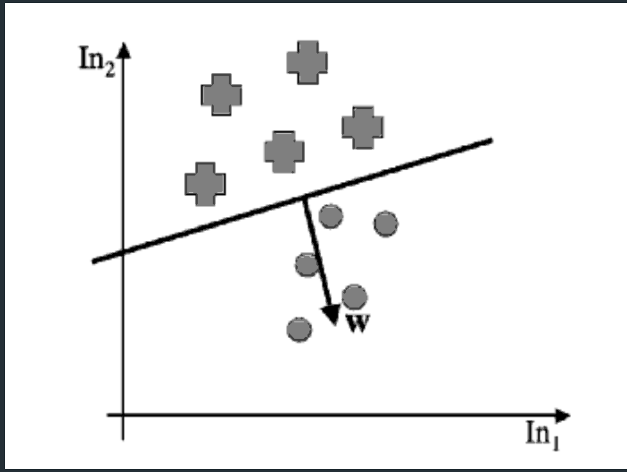
- Moore's law
  - 10,000-fold more computing power (!)
- More data
- More experience with NN algorithms
- New ideas, extensions, tricks

What is Deep Learning?

- Multiple “shallow” models stacked on top of each other.
- Internal representations develop at the boundaries of the models.
- At each step, the shallow model transforms its input into a different representation.

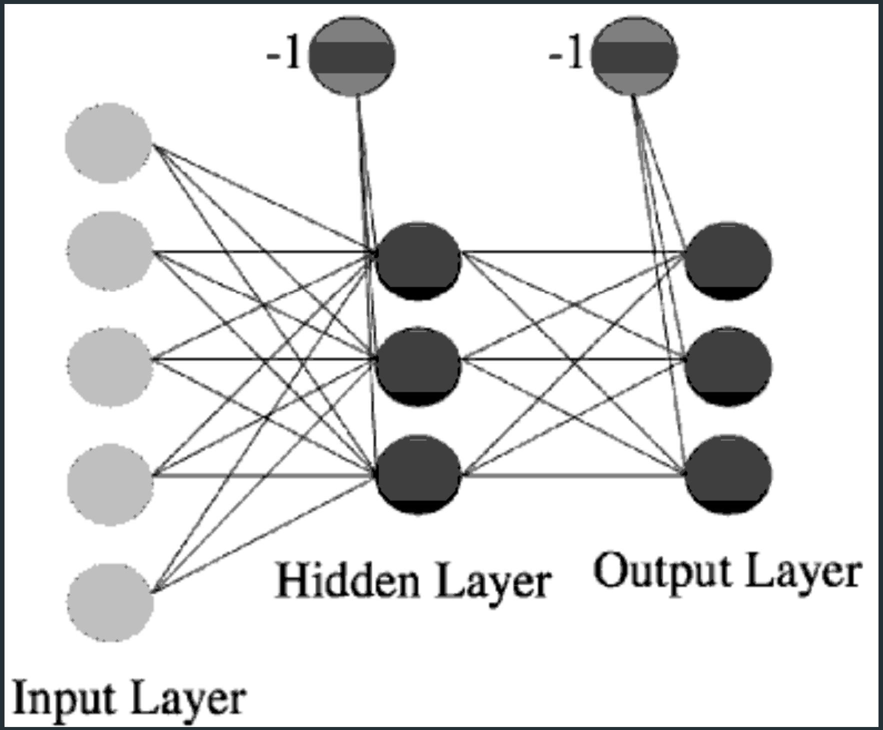
Shallow Learning

- Works very well, given appropriate inputs
- Logistic regression
  - Uses linear combination of inputs:
    f() = a + bx + cy + dz + …
    (the dot product)
  - Requires linearly separable input features.



Learning

- How does each parameter affect the output?
- Trial and error?
- Genetic algorithm?
- Calculus!



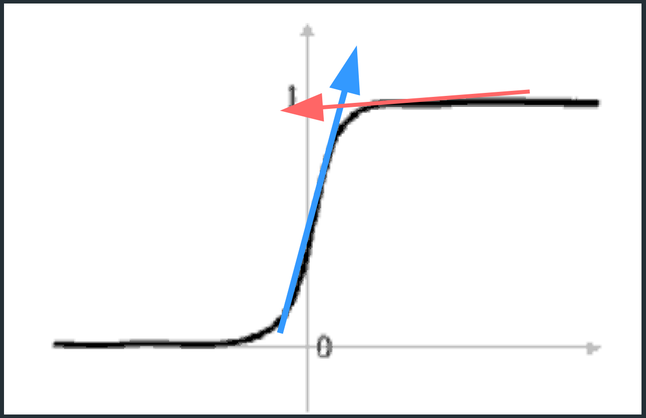
SyncWords
Captions Automated

Gradient Descent

- Modify parameters in the direction of the derivative (“gradient”)
- Any function which is differentiable can be used in a computational network.

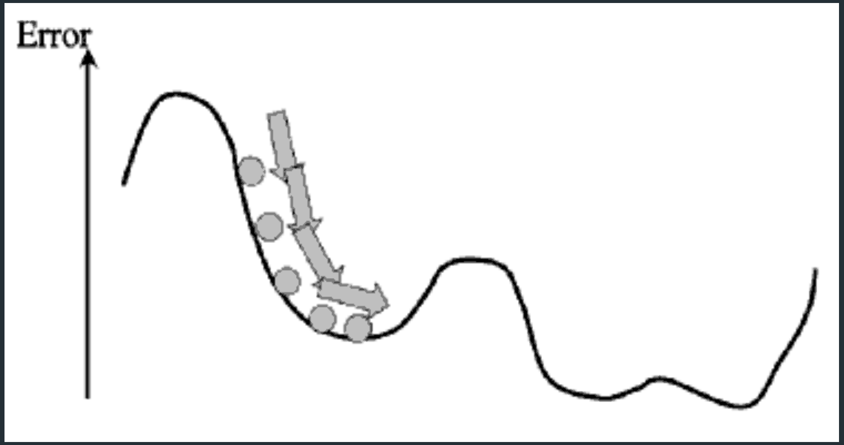
Gradient Descent

- Follow the derivative!
- But how far?
- Crazy, stupid idea:
  - Bigger derivative, bigger change in weight.



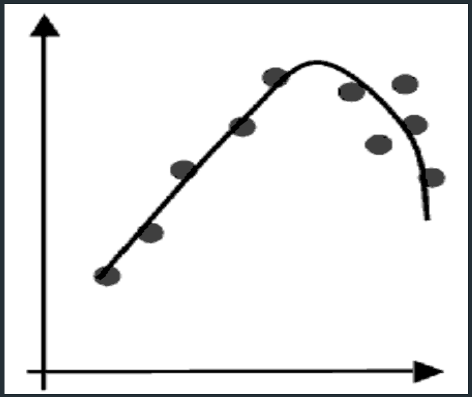
Stochastic GD

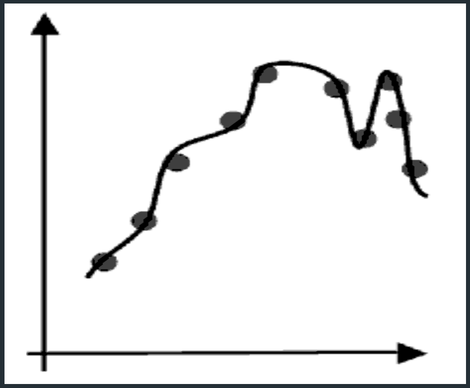
- We meander like a drunk. Unafraid.
- Mini-batches give bad gradients.
- Large learning rate shoots past optimum.
- Drop-out temporarily breaks some neurons.



Overfitting, Data, and Tricks

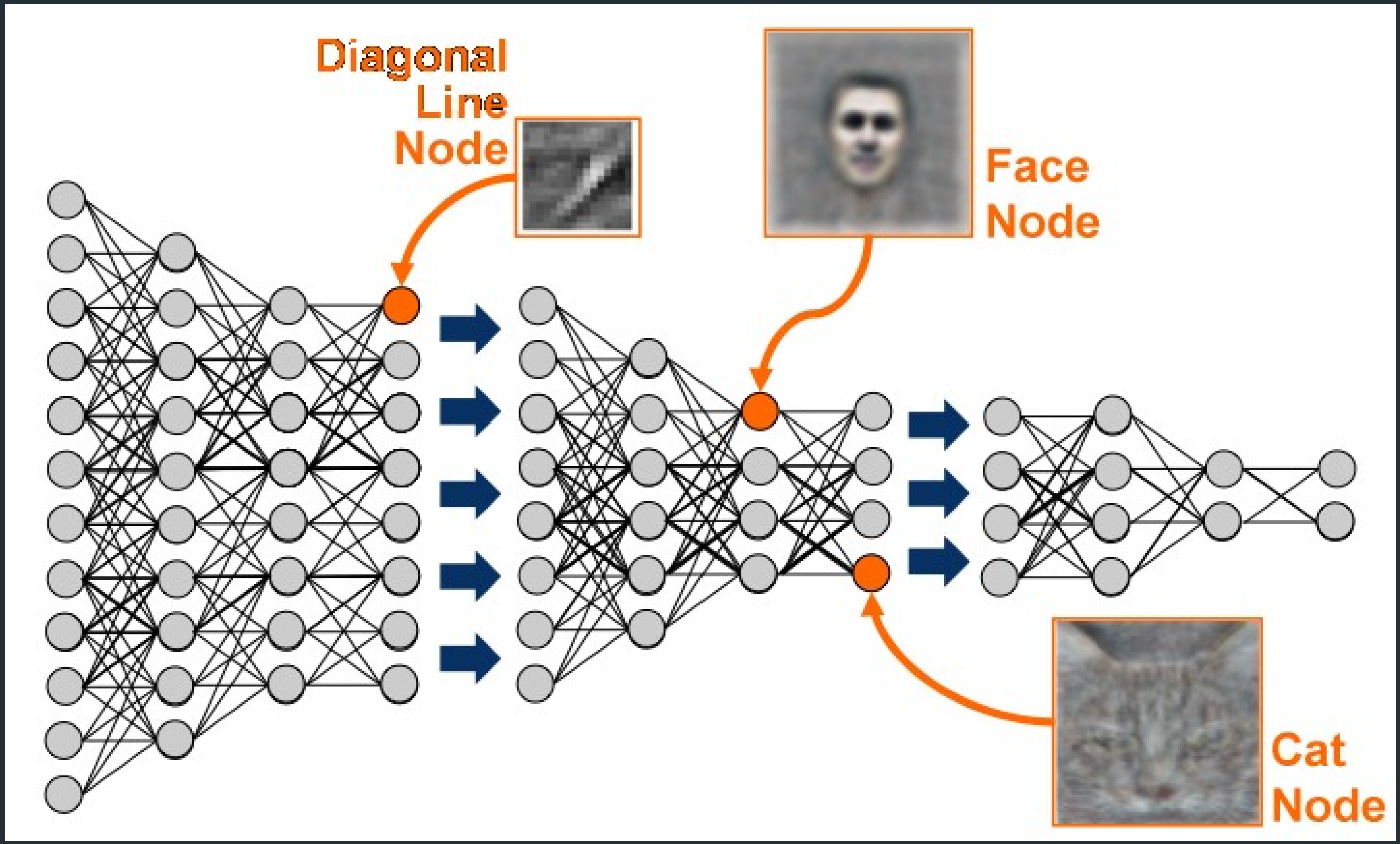
- Don't be too smart.
  Ignorance is creativity.
- More data always helps.
  - Fake it (augmentation).
- Reduce number of parameters.
- Be more stochastic.
- Multi-task learning.





SyncWords
Captions Automated

Deep Neural Network



SyncWords
Captions Automated

Common Tasks

--

Building Blocks

SyncWords

Captions Automated

Regression

- Approximate some function.
- Objective function: mean square error

Classification

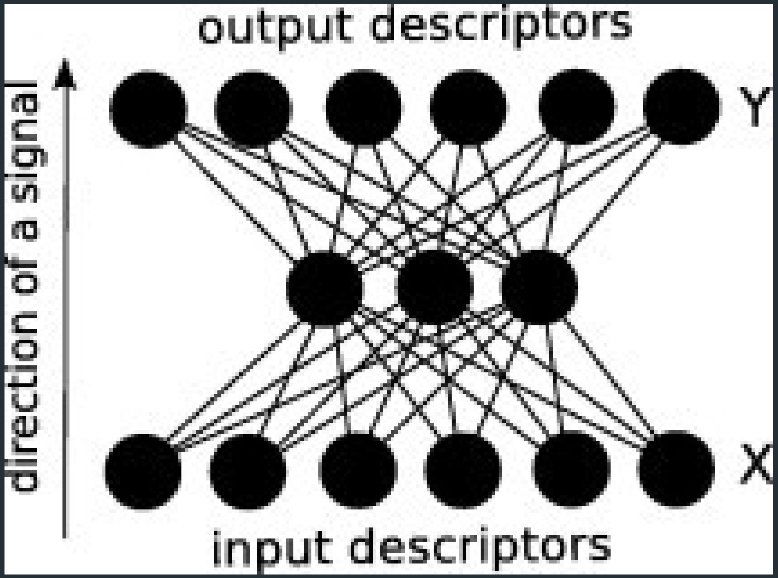
- Categorical prediction
- We get the posterior probability

$$P(\, C \mid X \,)$$

- Sound/image classification (cats, phonemes)
- Language modeling (predict next word)

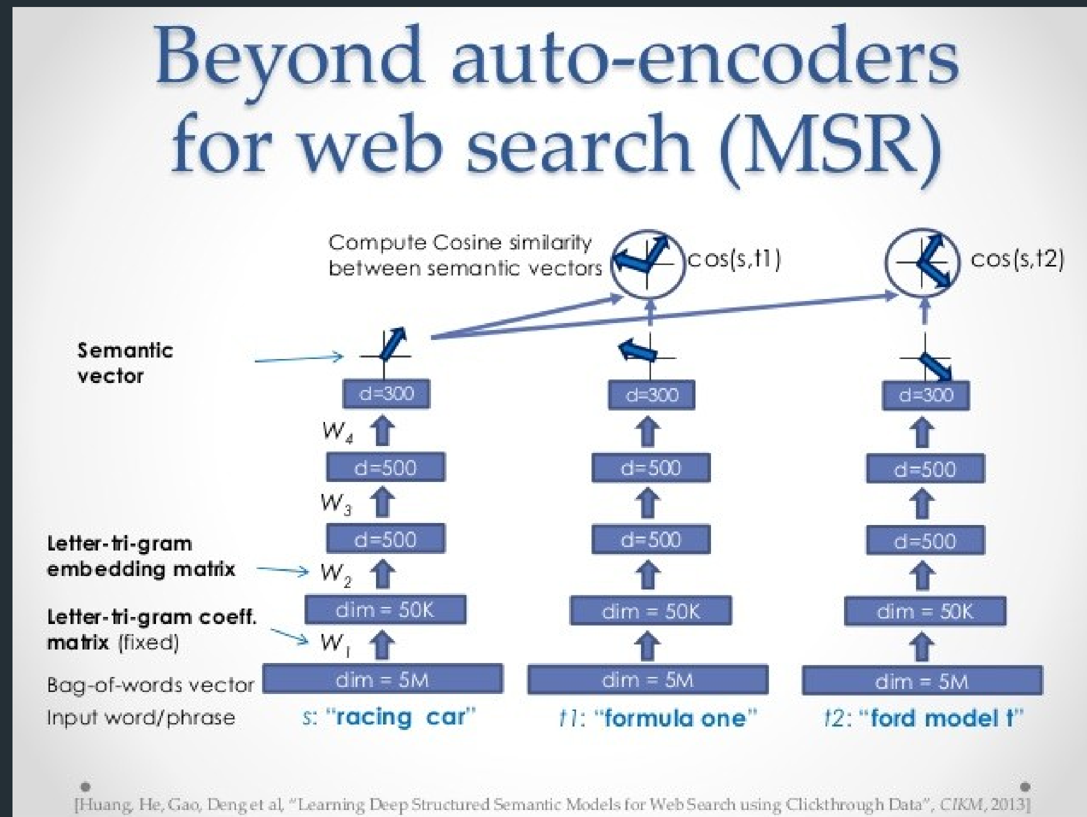
“Compression”

- Auto-encoder
- Creates a compressed internal representation
- (Unsupervised)



Embedding

- DSSM/ Siamese net
- Unsupervised
- Constructs a semantic space where points have meaning.



Embedding

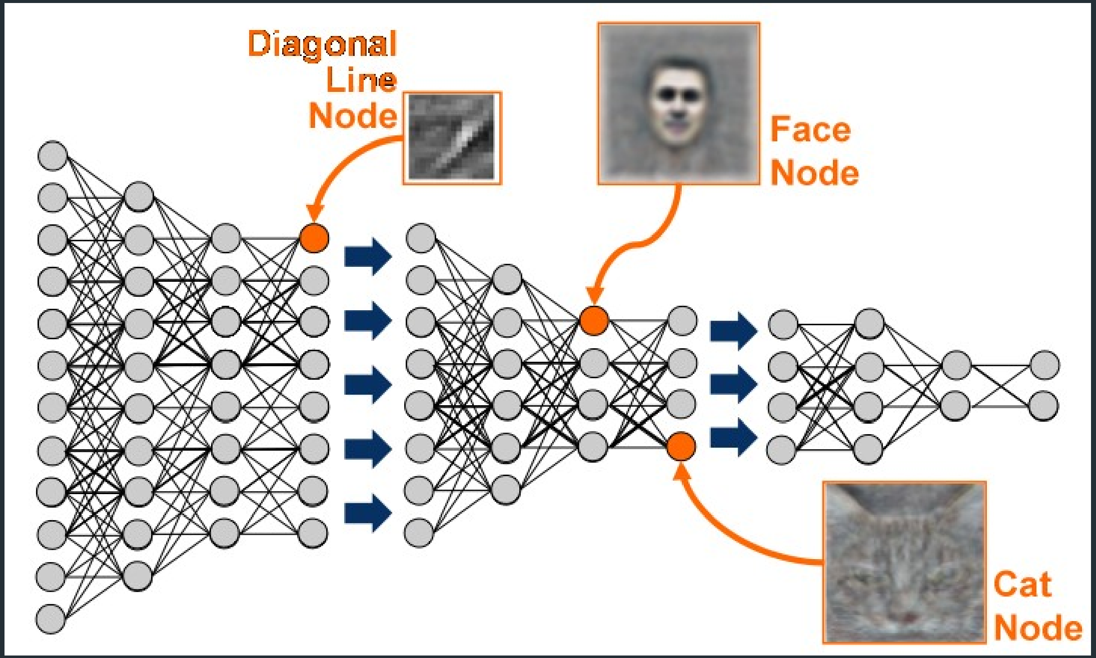
- Search/comparison
- Transformation
  - Images to captions
- Using the embedding as features to train a different (often, simpler) model.
- DSSM allows using a lightly-supervised dataset.

Architectures

SyncWords
Captions Automated

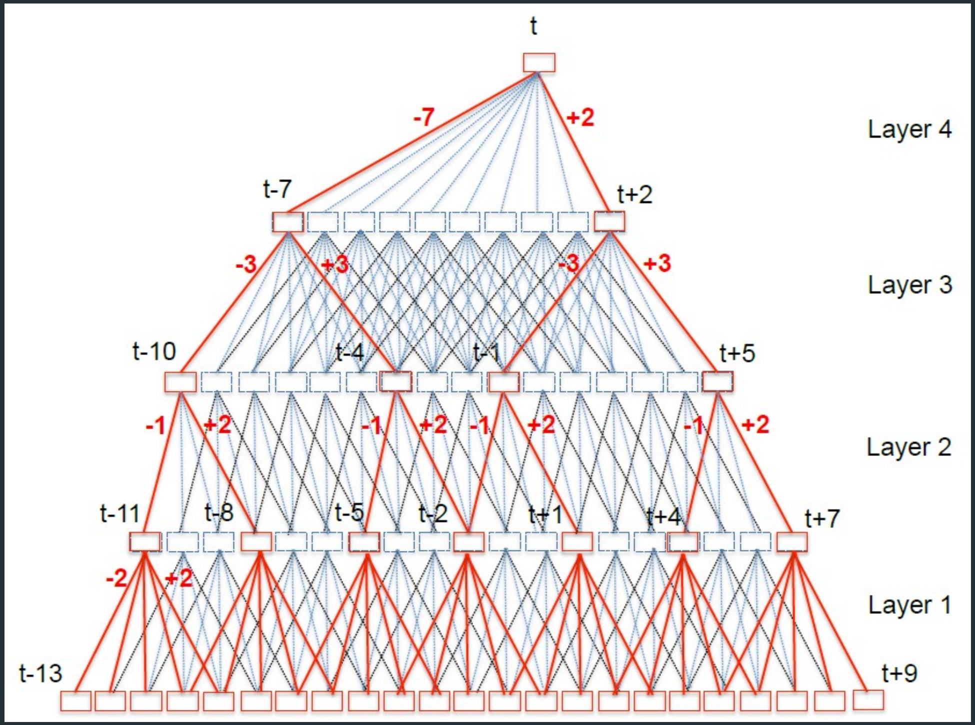
Feed-forward NN (DNN)

- **<u>Simple, one-shot processing</u>**
- **Work well**



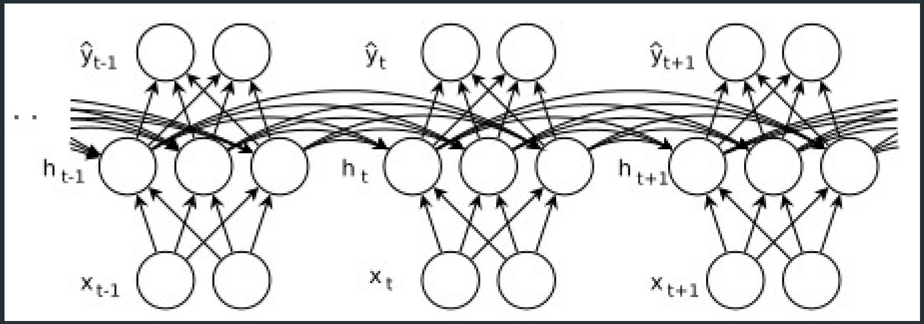
Convolutional NN (CNN)

- **<u>Re-use of blueprints.</u>**
  - Better generalization
  - Less over-fitting
  - Each parameter trains on many more examples



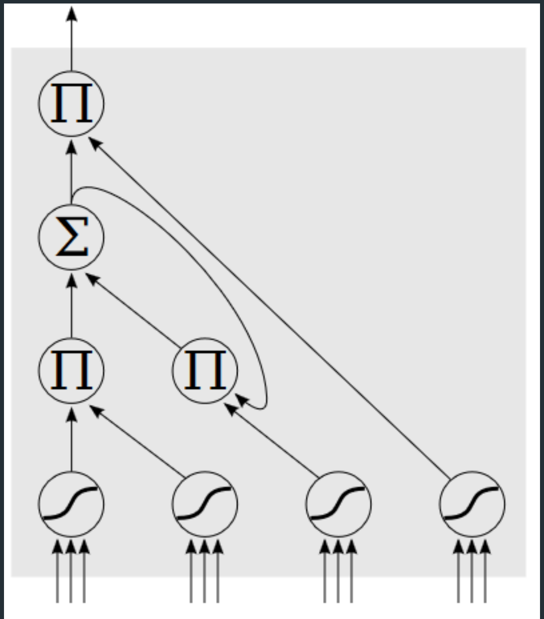
Recurrent NN (RNN)

- **<u>Short-term memory</u>**
- Also re-uses parameters



LSTM RNN

- **<u>Medium-term memory</u>**
- Fixes the partial derivatives
- Makes changes to parameters more stable.



Activation Functions (Neuron Types)

- Sigmoid, Tanh
- Rectifier
- Max-Out
- Leaky, Programmable Rectifier Linear
- LSTM

Artificial Neuroscience

Individual Neurons are Detectors

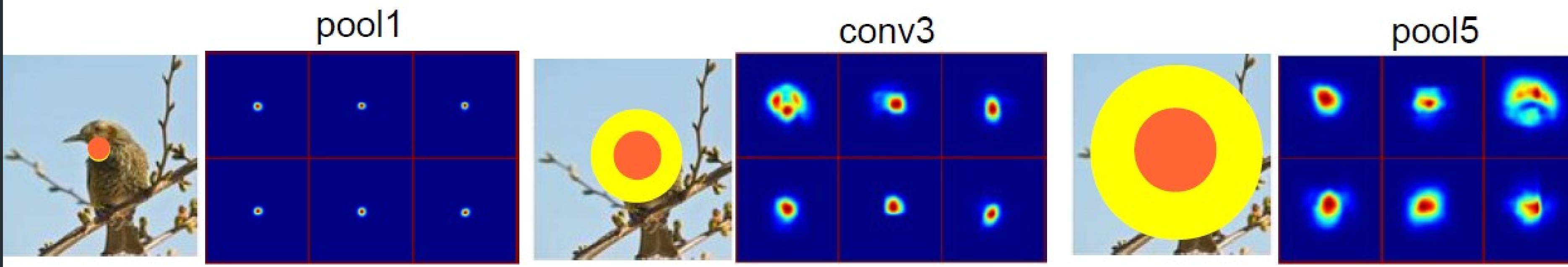
- **Object Detectors Emerge in Deep Scene CNNs**
  Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba
  CVPR 2015

- Monitor which neurons activate, and trace back through the convolutional layers which region of the image contributed to its activation.

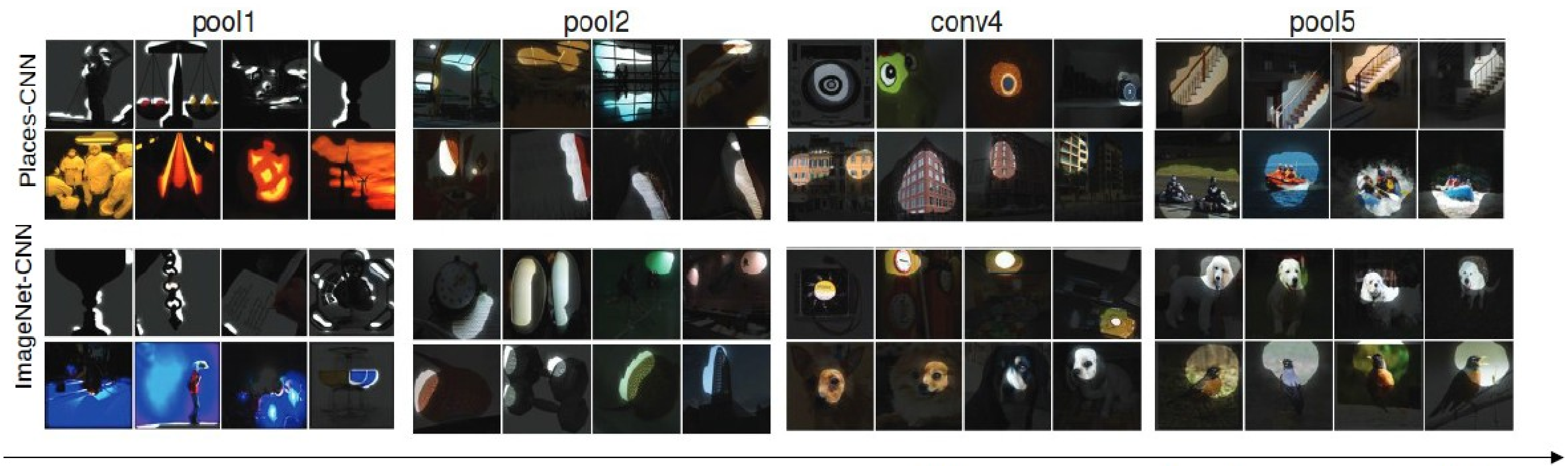
Estimating the Receptive Fields

Estimated receptive fields

Actual size of RF is much smaller than the theoretic size



Segmentation using the RF of Units

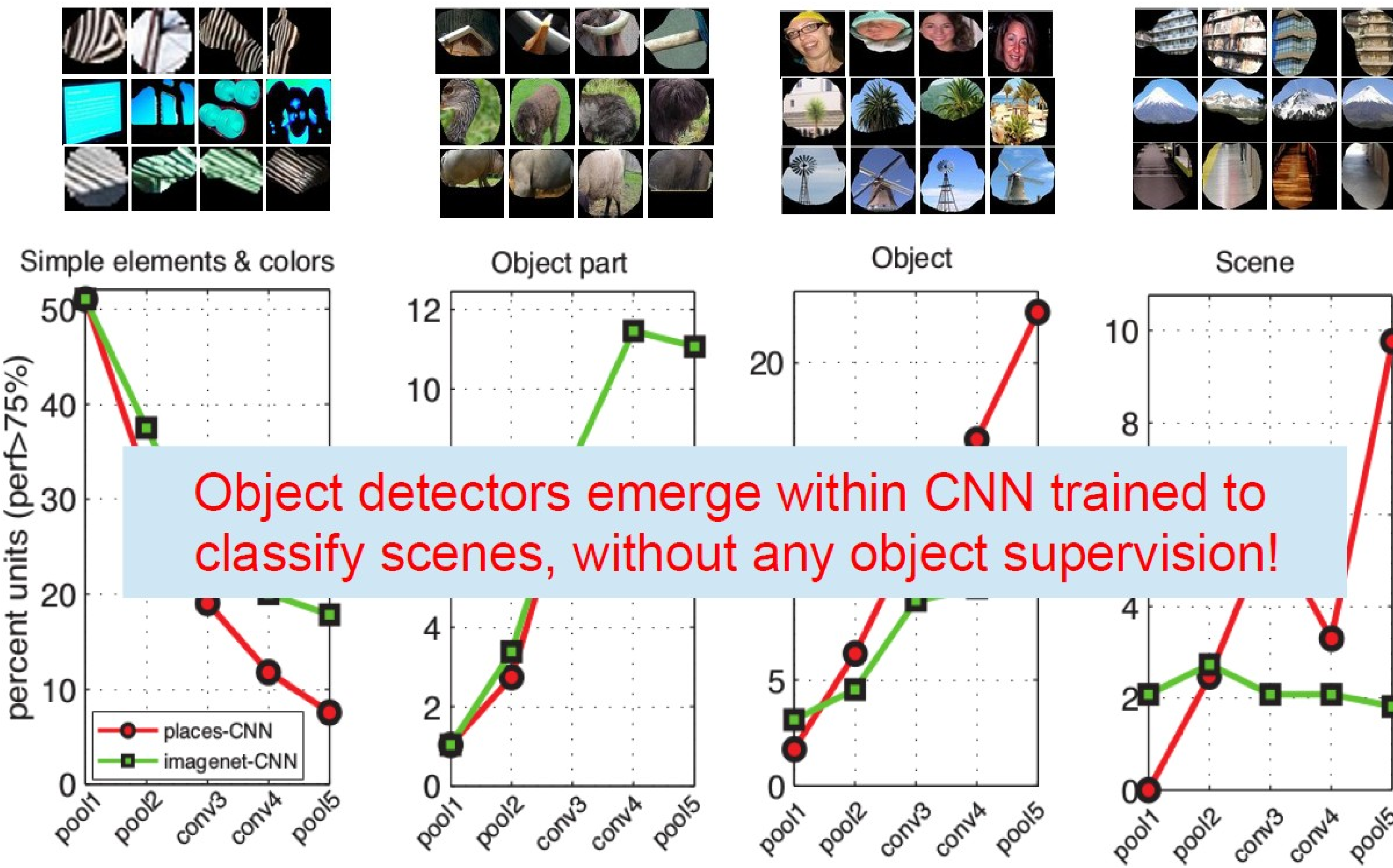


More semantically meaningful

Annotating the Semantics of Units

Pool5, unit 13; Label: Lamps; Type: object; Precision: 84%

Distribution of Semantic Types at Each Layer

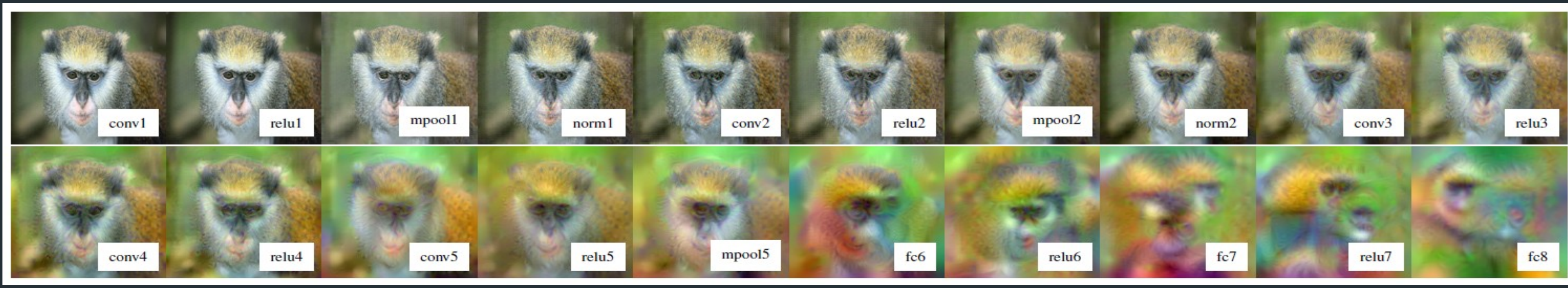


Object detectors emerge within CNN trained to classify scenes, without any object supervision!

Inverting CNNs

Understanding Deep Image Representations by Inverting Them

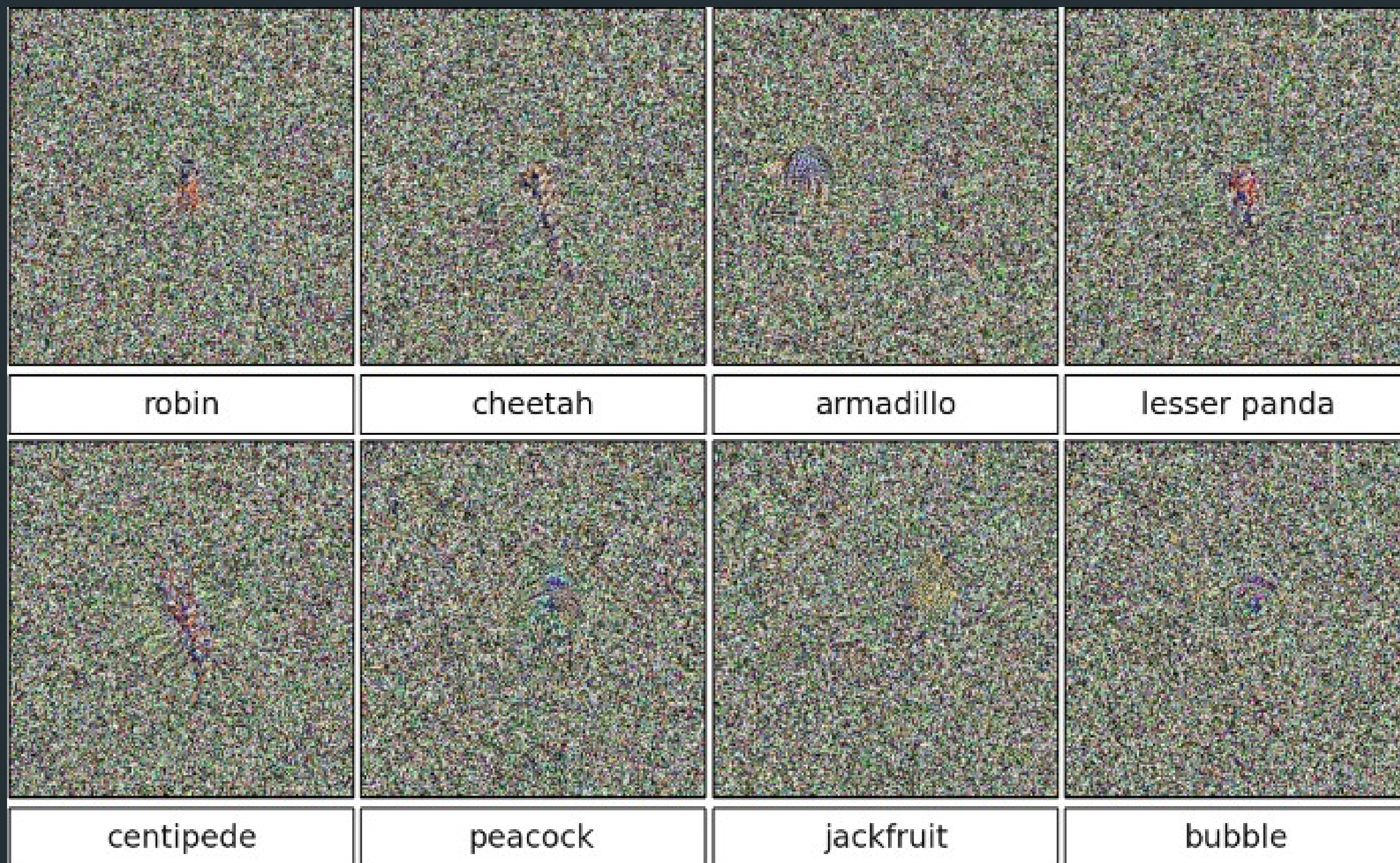
Aravindh Mahendran, Andrea Vedaldi

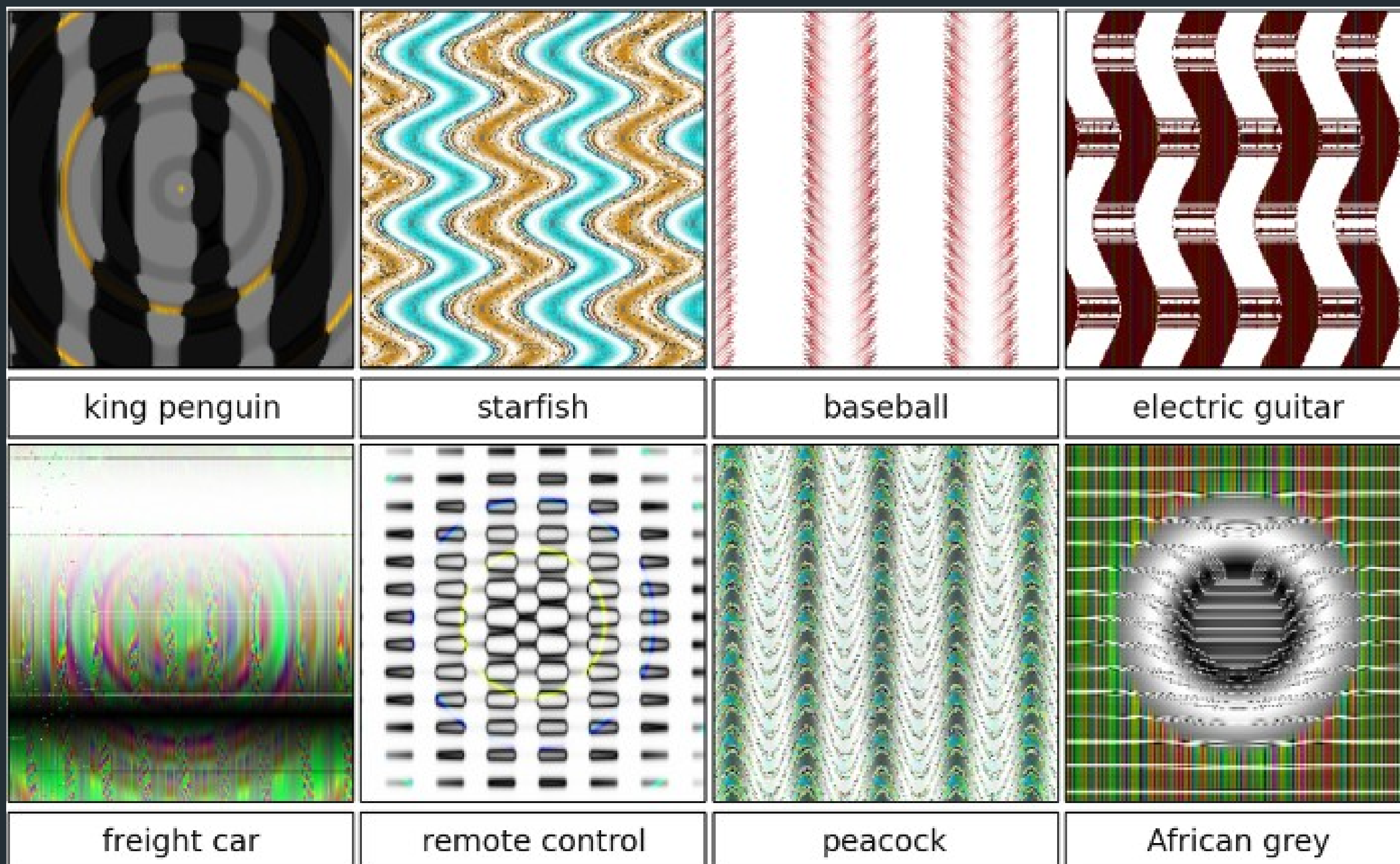

conv1
relu1
mpool1
norm1
conv2
relu2
mpool2
norm2
conv3
relu3
conv4
relu4
conv5
relu5
mpool5
fc6
relu6
fc7
relu7
fc8

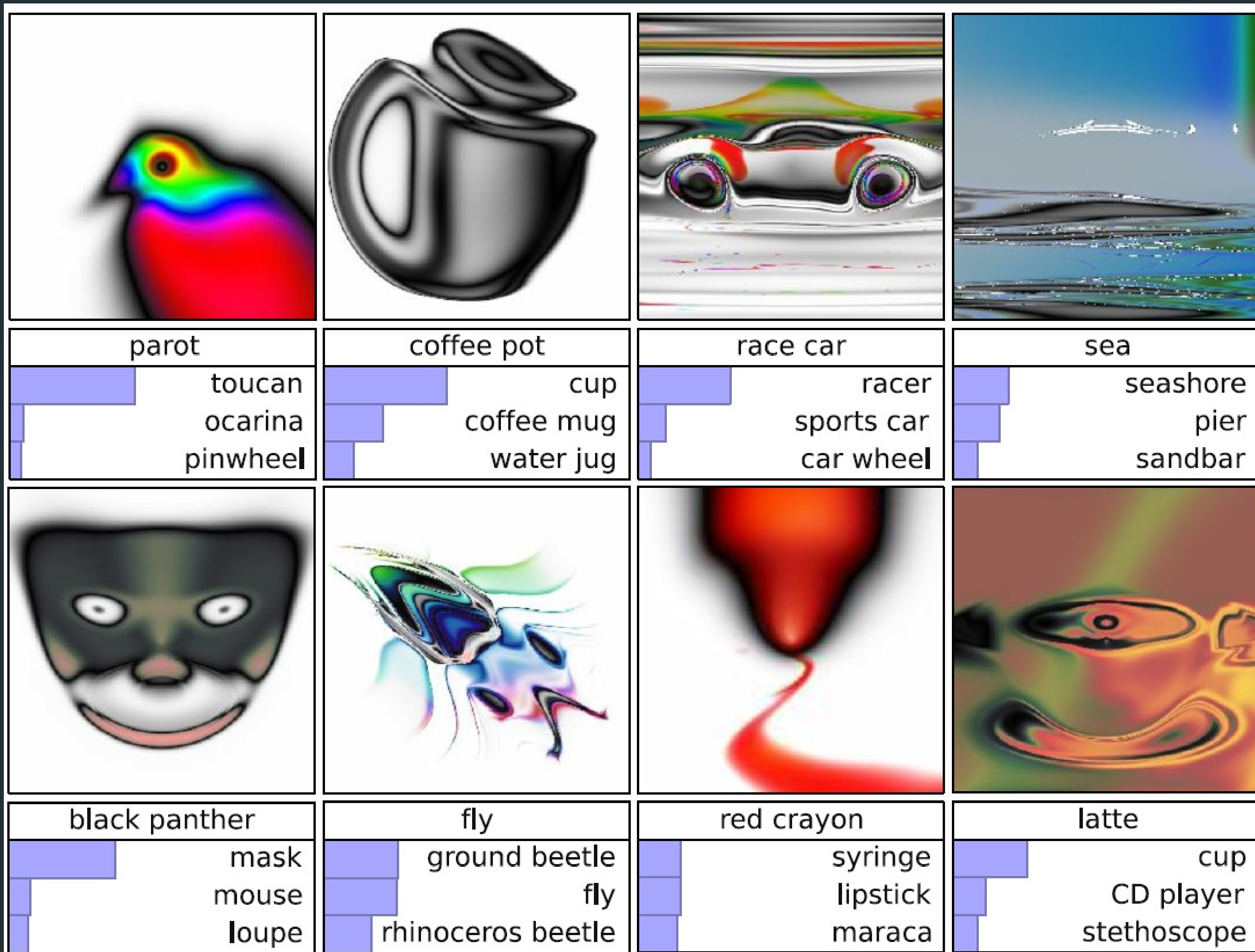
Snow Crash

- **Deep Neural Networks are Easily Fooled: High Confidence** Predictions for Unrecognizable Images
  Anh Nguyen, Jason Yosinski, Jeff Clune

Snow Crash


robin
cheetah
armadillo
lesser panda
centipede
peacock
jackfruit
bubble


king penguin
starfish
baseball
electric guitar
freight car
remote control
peacock
African grey


parot
toucan
ocarina
pinwheel
coffee pot
cup
coffee mug
water jug
race car
racer
sports car
car wheel
sea
seashore
pier
sandbar
black panther
mask
mouse
loupe
fly
ground beetle
fly
rhinoceros beetle
red crayon
syringe
lipstick
maraca
latte
cup
CD player
stethoscope