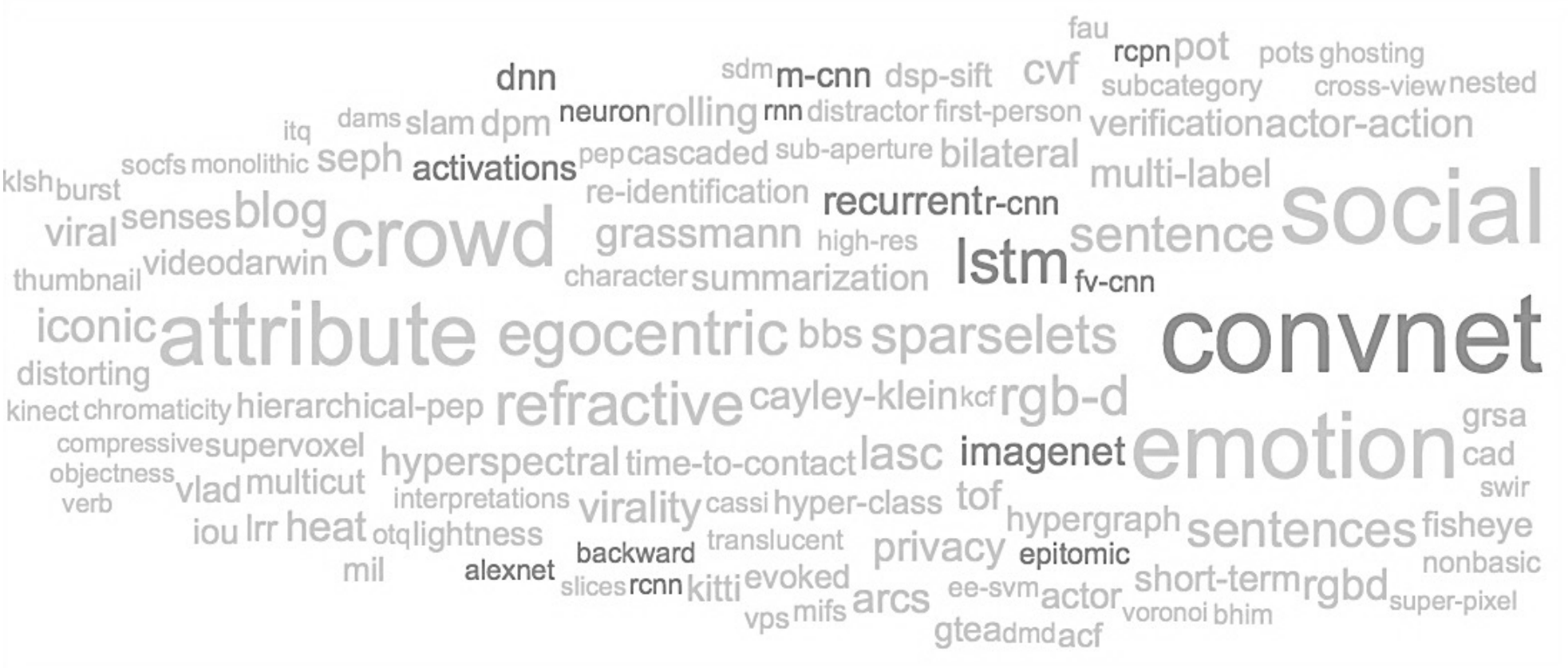
# Introduction to deep learning

Natalia Neverova

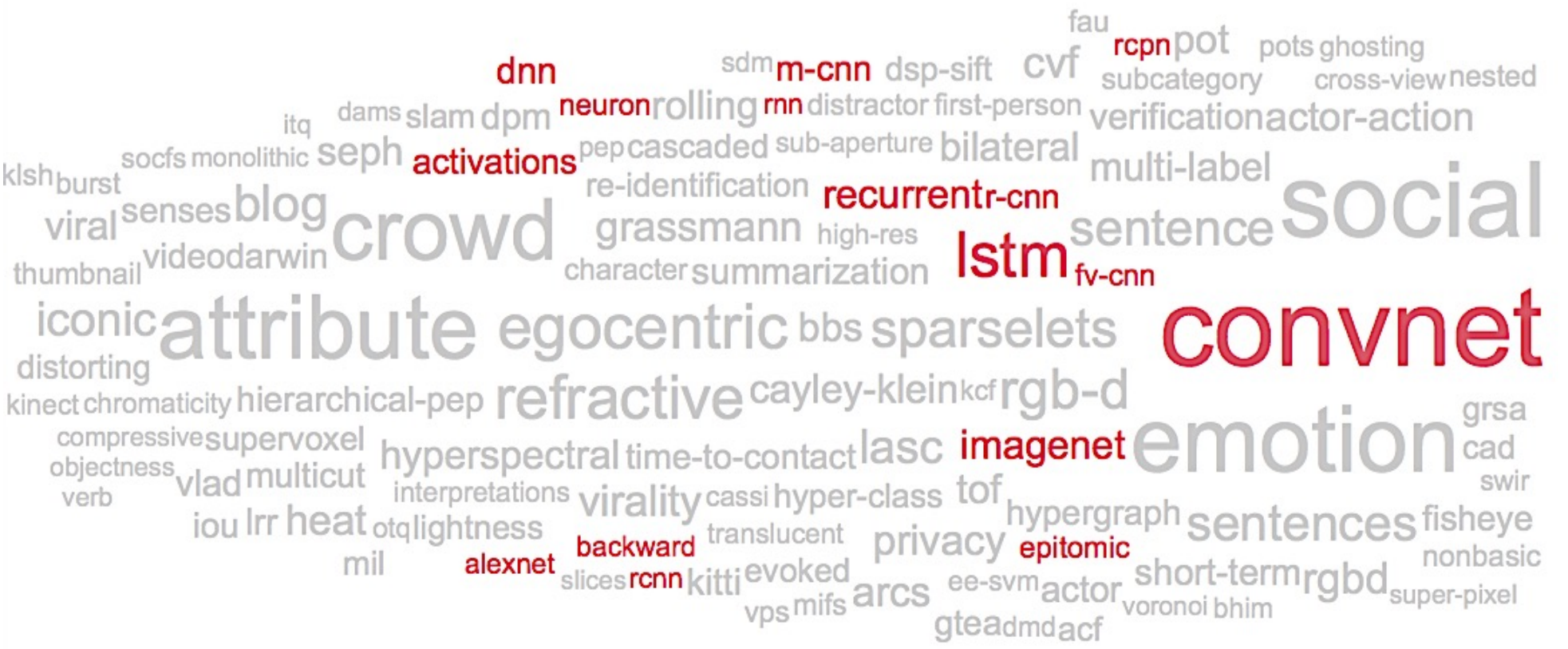
INSA-Lyon, LIRIS CNRS


UNIVERSITY of GUELPH
CHANGING LIVES IMPROVING LIFE
Google
INSA
INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON
LIRIS

# Computer vision 2015:
# what changed since ten years ago?



# Computer vision 2015:
# what changed since ten years ago?

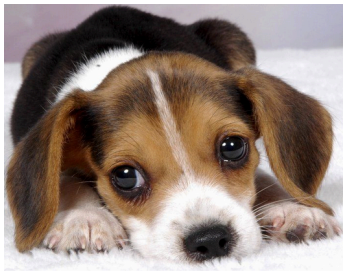


# What is the idea?

## An example: classical supervised learning setting

$x$ ⟶ $y$

input output



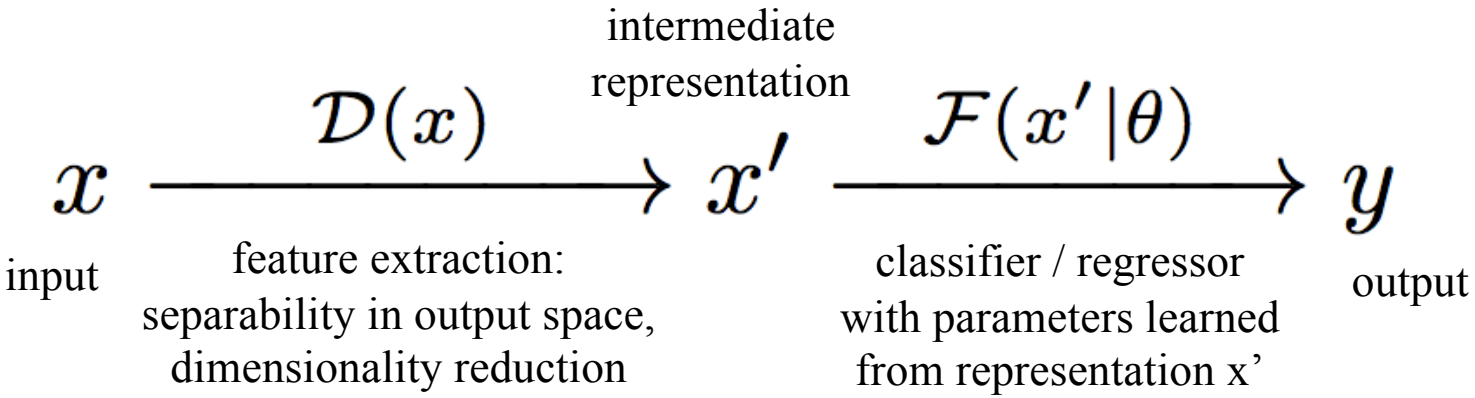
object recognition

{dog, cat, pokemon}

super resolution

# What is the idea?

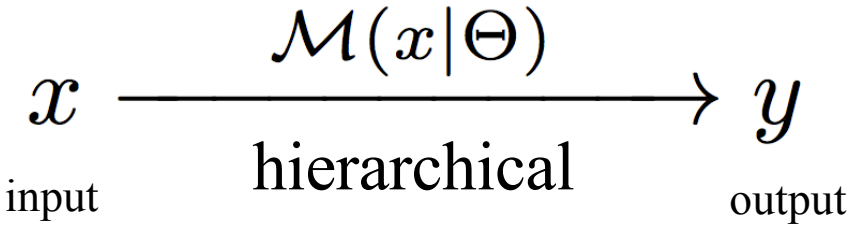
## An example: classical supervised learning setting



where $\mathcal{D}(x)$ ...

# What is the idea?

deep *learning* of data representations




joint learning
of representations with
increased levels of
abstraction
+ classification or
regression

# What is the idea?

biologically inspired model

huge amount of training samples

general and suitable for any input

supervised, unsupervised and reinforcement learning

# What has been achieved?

loosely biologically inspired models

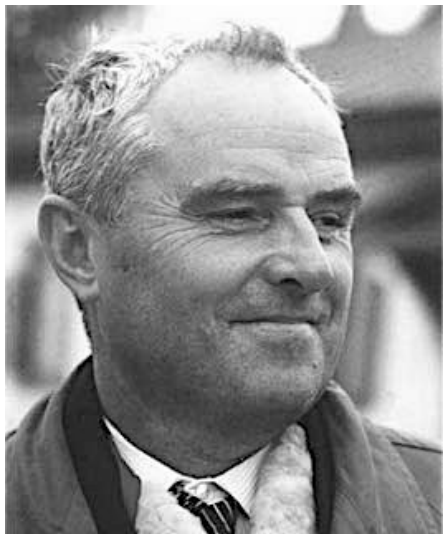
huge amount of training samples when available

general and suitable for many kinds of inputs after adaptation

supervised learning in a product

unsupervised and reinforcement learning – work in progress

# A bit of history



early 1960s

Alexey Ivakhnenko

first works on deep neural networks

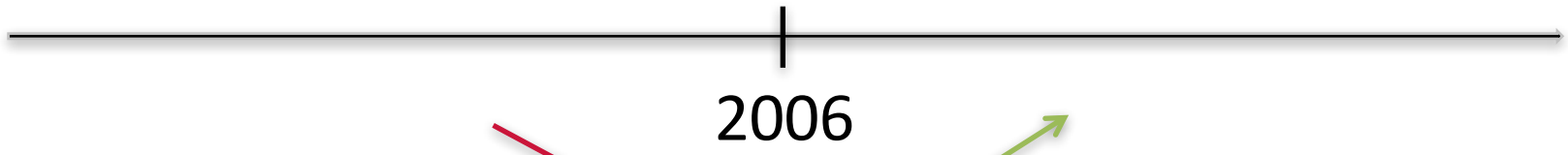
1986

Geoffrey Hinton

backpropagation algorithm in its current form

# A bit of history

computer vision
and speech
communities do not use
neural nets anymore



2006

hard to train
no computational resources
small datasets

efficient training
GPU / multi-core processors
large datasets

# A bit of history

2011

Microsoft

breakthrough
in speech recognition

2012

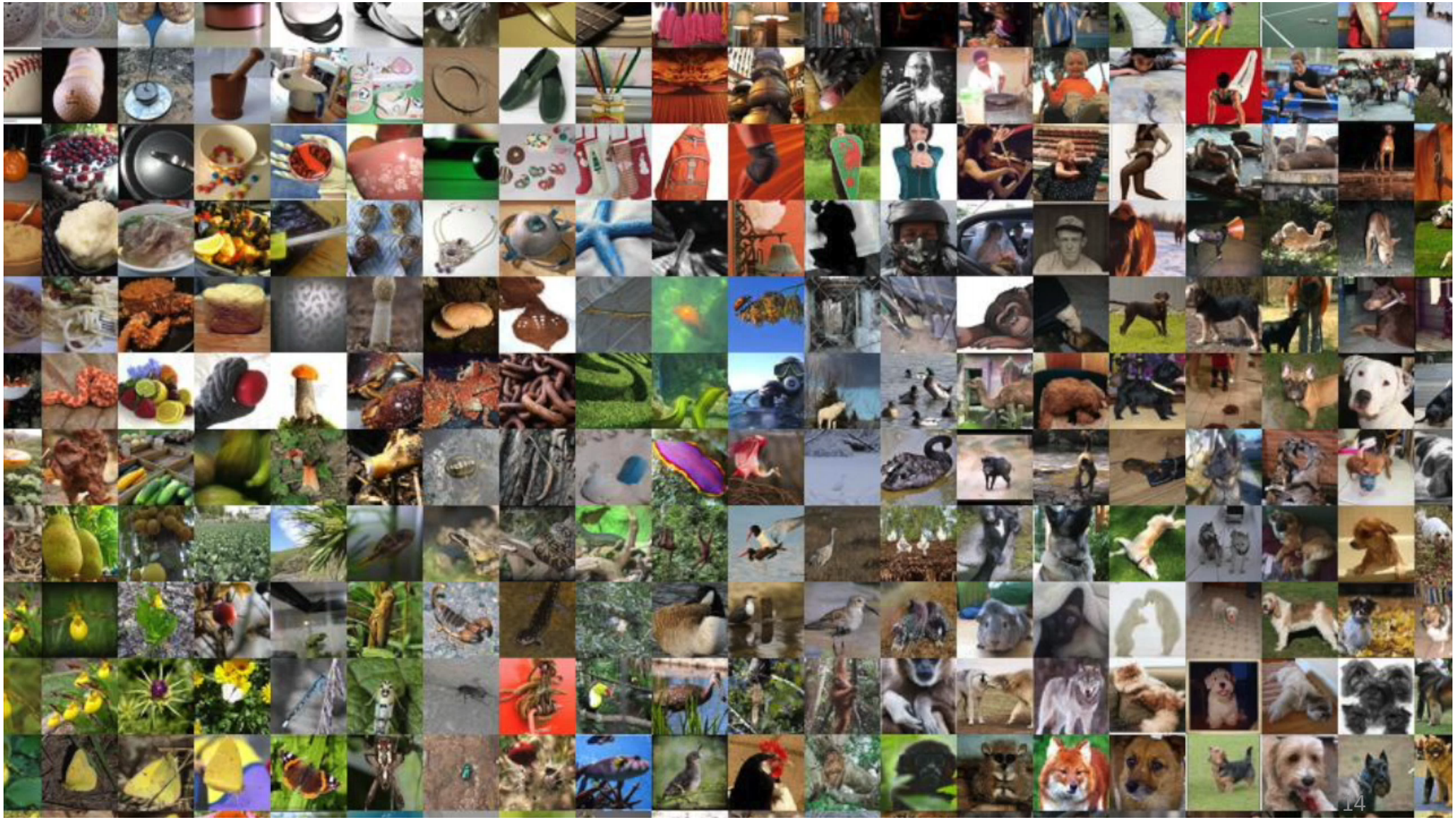
Geoffrey Hinton

breakthrough
in computer vision

IMAGENET

IMAGENET

WordNet hierarchy, 21841 classes
Images: 14,197,122
Objects annotated: 1,034,908
Crowd source annotations



IMAGENET

# Object recognition

| Method | Team | Year | Error |
|---|---|---|---|
| Hand crafted | University of Tokyo | 2012 | 0.2617 |
| AlexNet | University of Toronto | 2012 | 0.1531 |
| Multiple neural nets | Clarify | 2013 | 0.1120 |
| **GoogLeNet** | **Google** | **2014** | **0.0666** |

# Object detection

| Method | Team | Year | MAP |
|---|---|---|---|
| Hand crafted | UvA-Euvision | 2013 | 0.2258 |
| **GoogLeNet** | **Google** | **2014** | **0.4393** |

# Human performance?

Andrej Karpathy's experiment (PhD @ Stanford)

deep learning 6.7%, human 5.1% error

| | Deep learning correct | Deep learning wrong |
|---|---|---|
| Human correct | 1352/1500  | 72/1500<br>• Objects very small or thin<br>• Abstract representations<br>• Image filters |
| Human wrong | 46/1500<br>• Fine-grained recognition<br>• Class unawareness<br>• Insufficient training data | 30/1500<br>• Multiple objects<br>• Incorrect annotations |

# A bit of history

2011  A. Ng founded Google Brain project

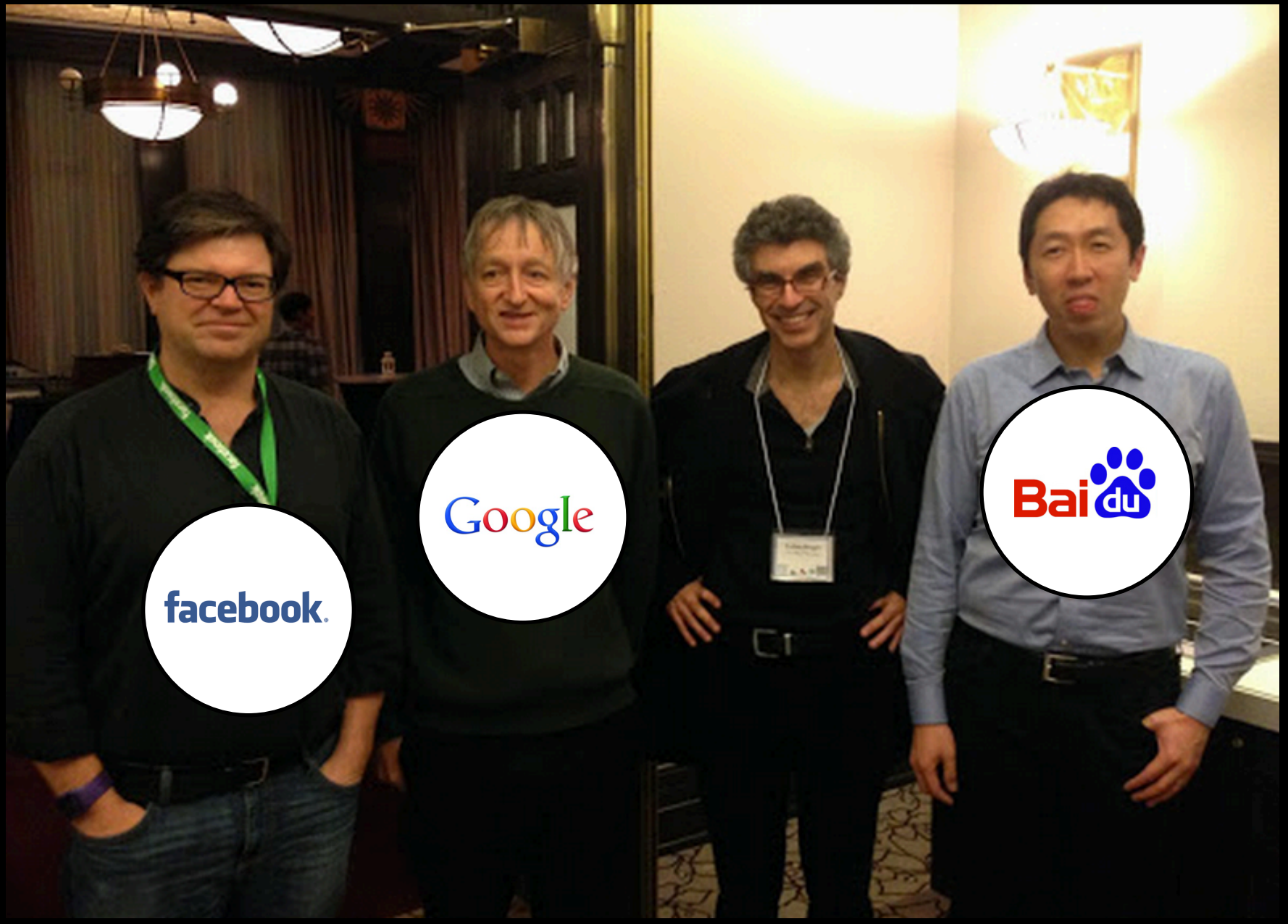
2013  DNNResearch is acquired by Google
Hinton join Google Brain project

2013  Y. LeCun directs Facebook's new AI lab

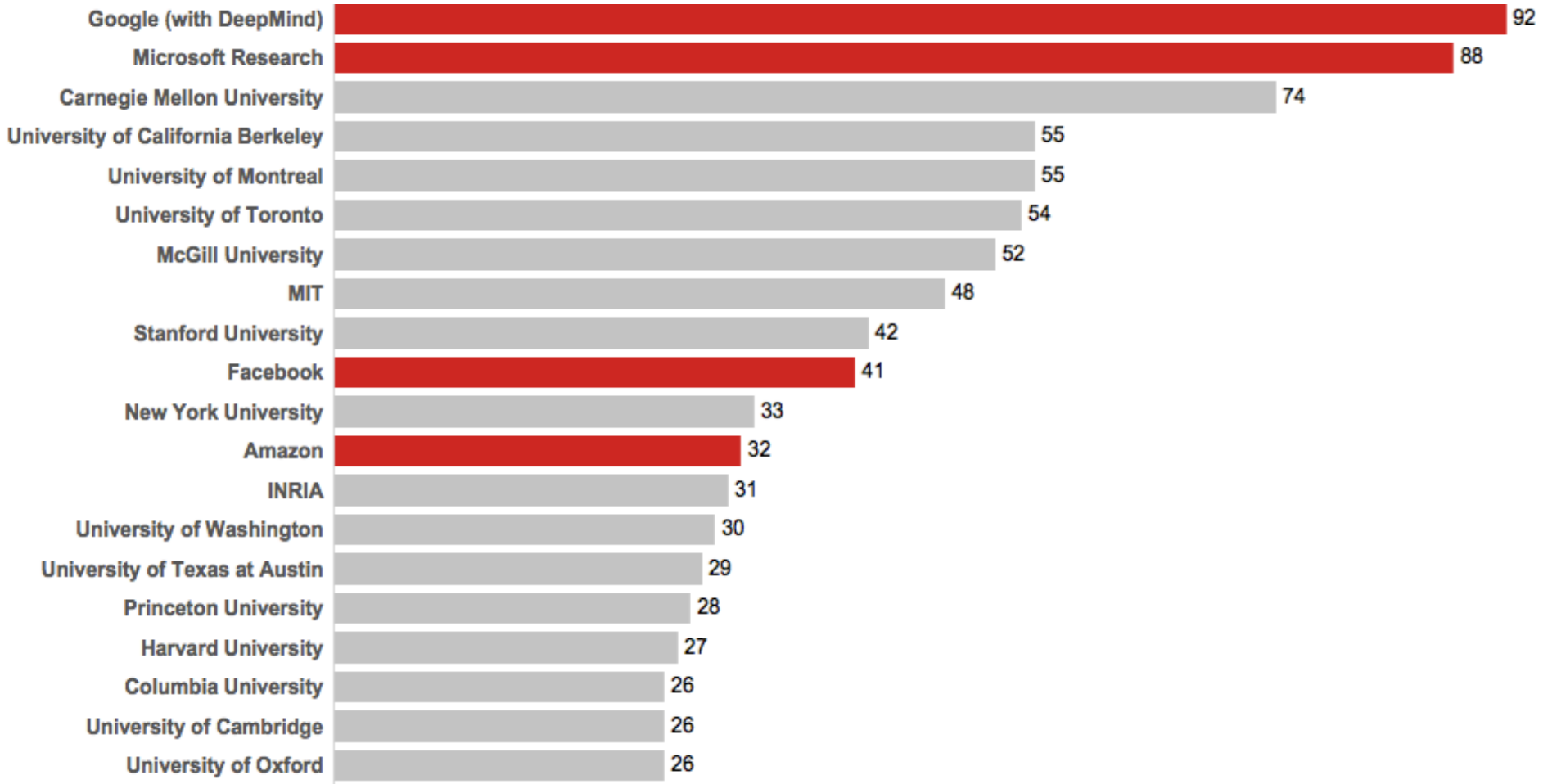
2014 

Is acquired by Google for 400M$

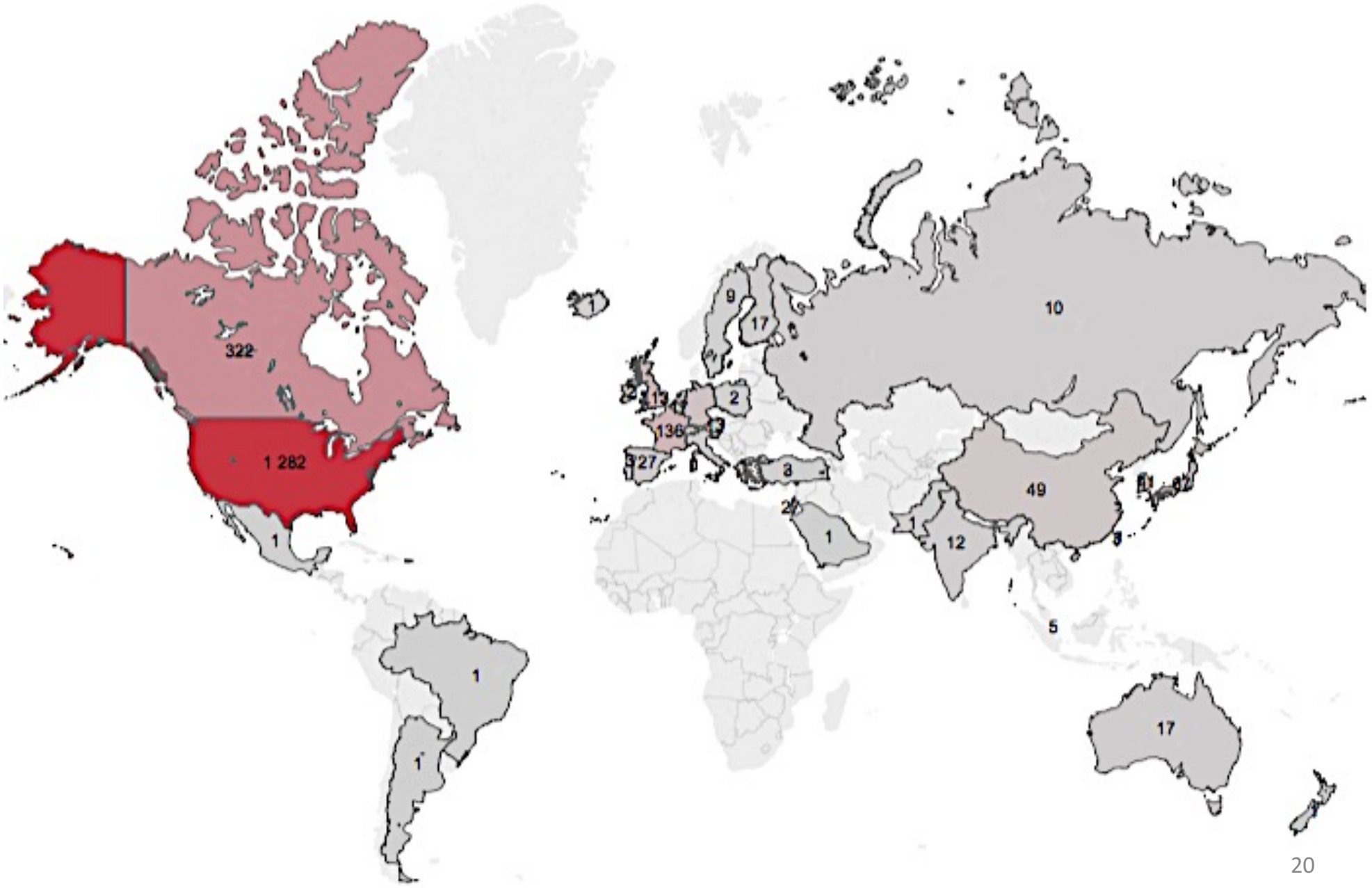
2014  A. Ng directs Baidu's new AI lab


facebook
Google
Baidu

# NIPS 2014: academia vs industry



# NIPS 2014 attendance



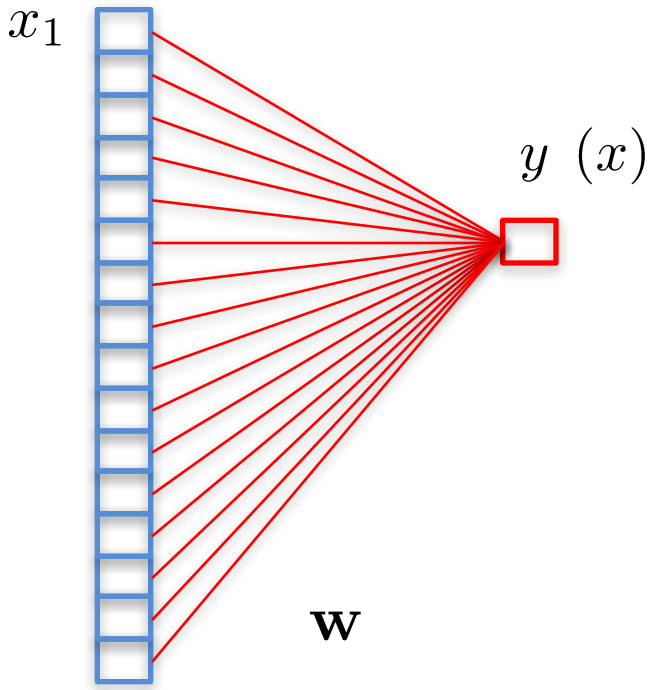
# NIPS 2014 attendance

# NIPS 2014 attendance

# Technology overview

# Perceptron

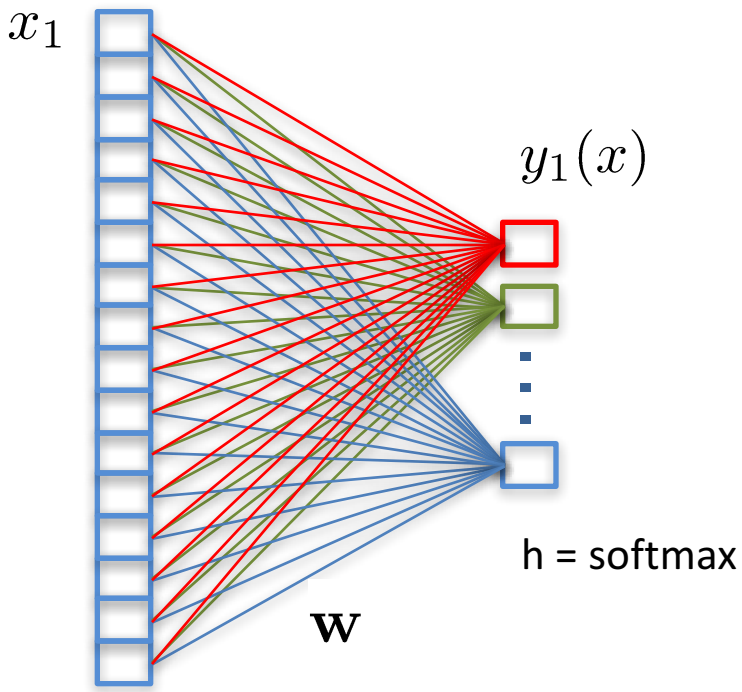
$$y_i(\mathbf{x}, \mathbf{w}) = \sum_{i=0}^{D} w_i x_i$$




input layer, i output layer

# Logistic Regression

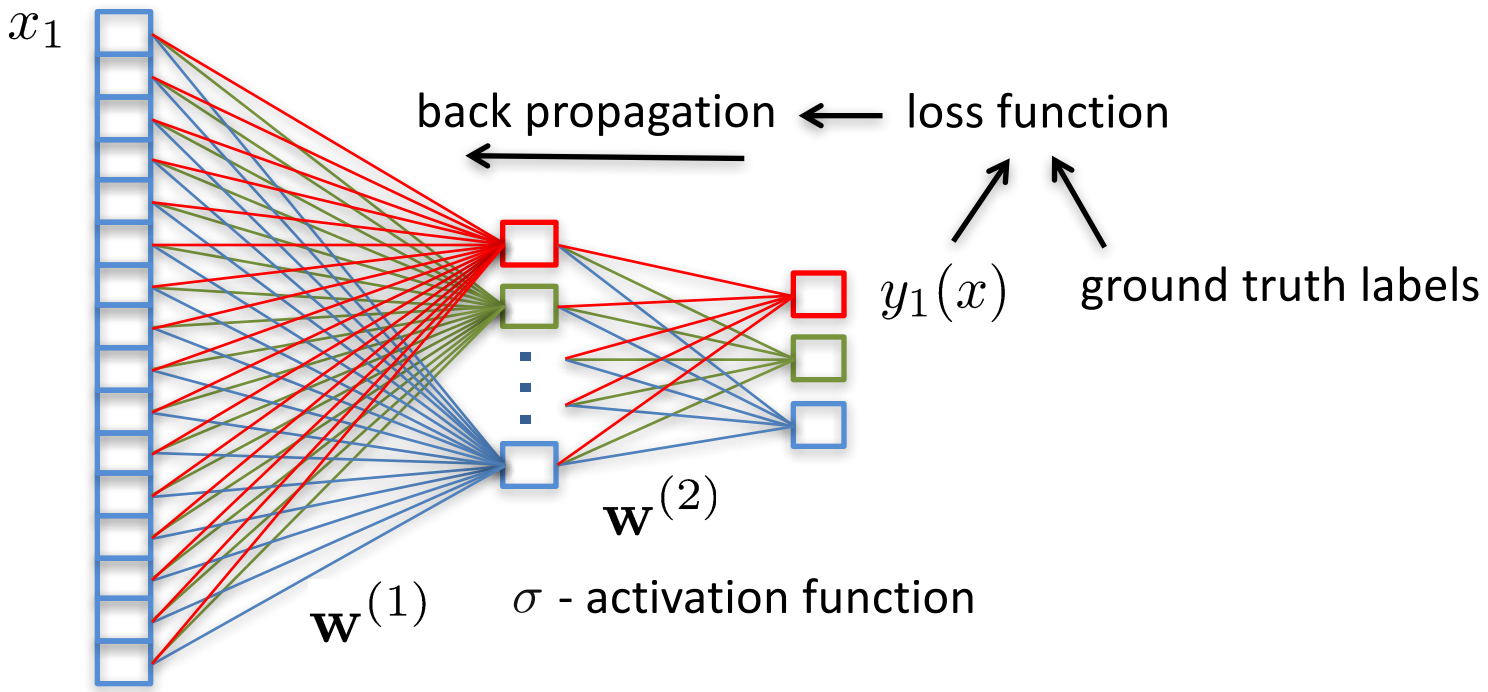
$$y_j(\mathbf{x}, \mathbf{w}) = h\left(\sum_{i=0}^{D} w_{ji}\, x_i\right)$$



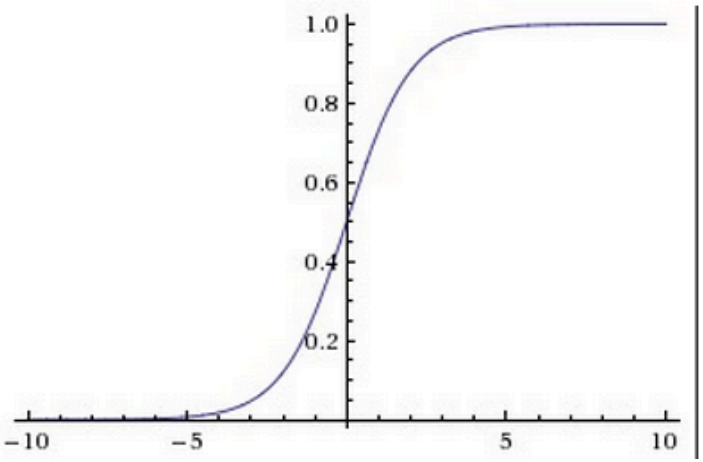

input layer, i output layer, j

# Multi-Layer Perceptron (MLP)

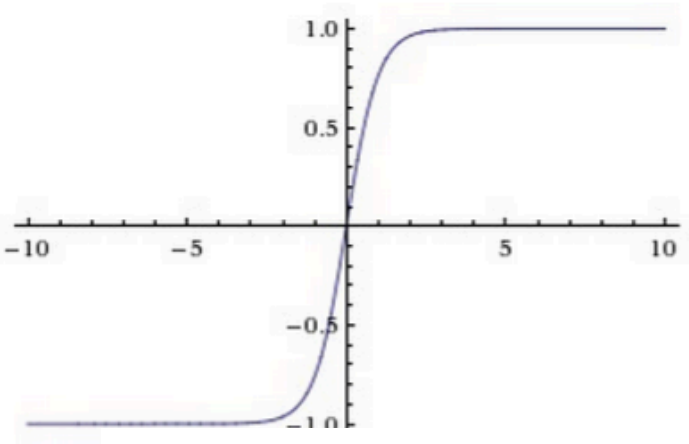
$$y_k(\mathbf{x}, \mathbf{w}) = h\left(\sum_{j=0}^{M} w_{kj}^{(2)}\, \sigma\left(\sum_{i=0}^{D} w_{ji}^{(1)} x_i\right)\right)$$



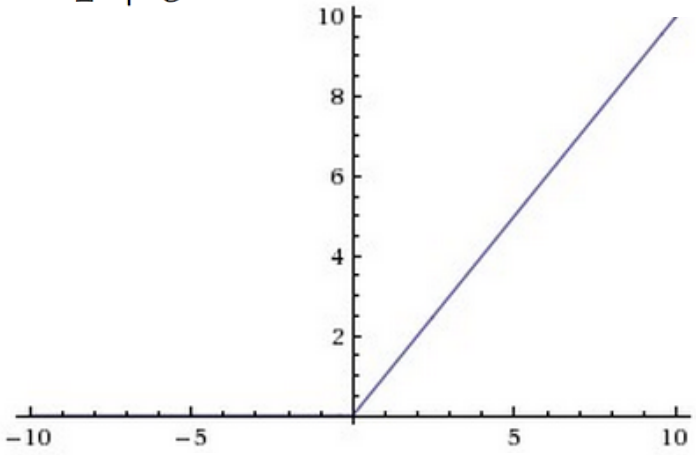
# Activation functions




$$\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$$



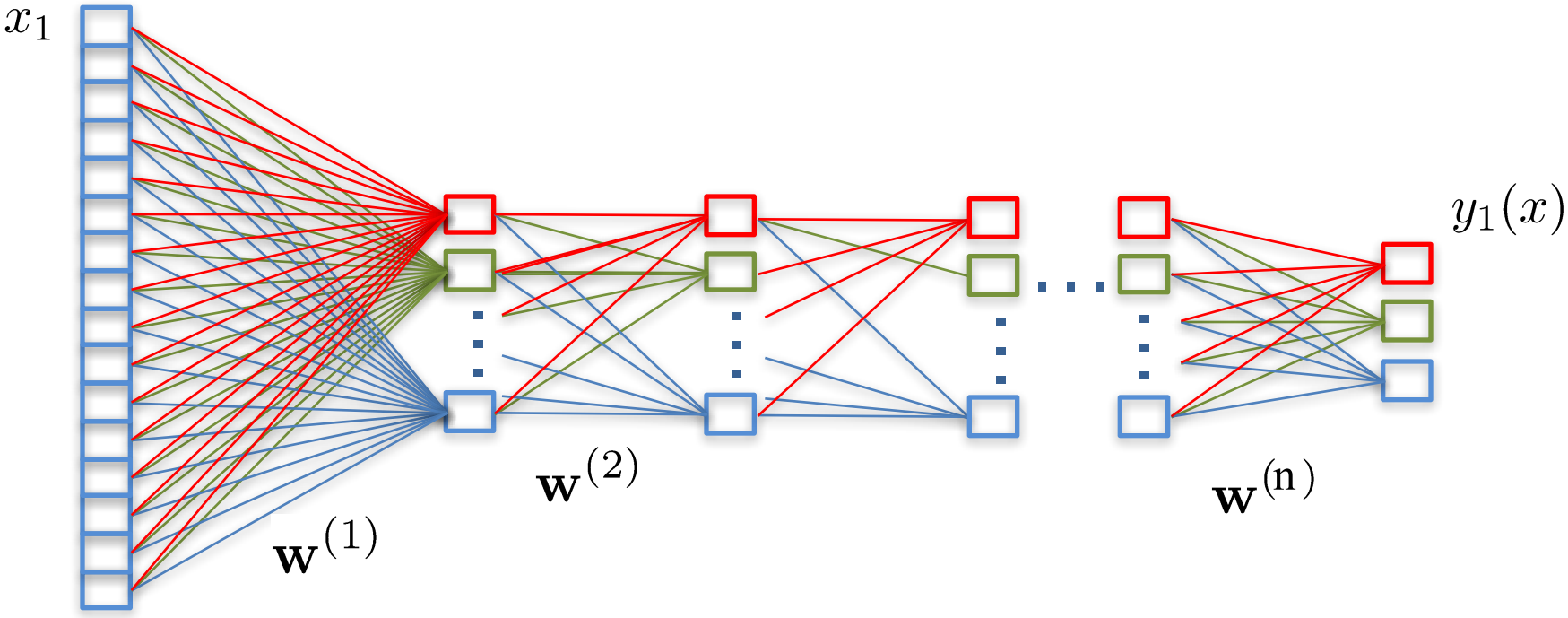

$$\tanh(x) = \frac{1-e^{-2x}}{1+e^{-2x}}$$




$$\text{ReLU}(x) = \max(0, x)$$

# Deep Neural Network (DNN)

$$y_n(\mathbf{x}, \mathbf{w}) = h\left( \ldots \sigma\left( \sum_{j=0}^{M} w_{kj}^{(2)} \sigma\left( \sum_{i=0}^{D} w_{ji}^{(1)} x_i \right) \right) \right)$$




input layer, i hidden layer, j hidden layer, k output layer, n

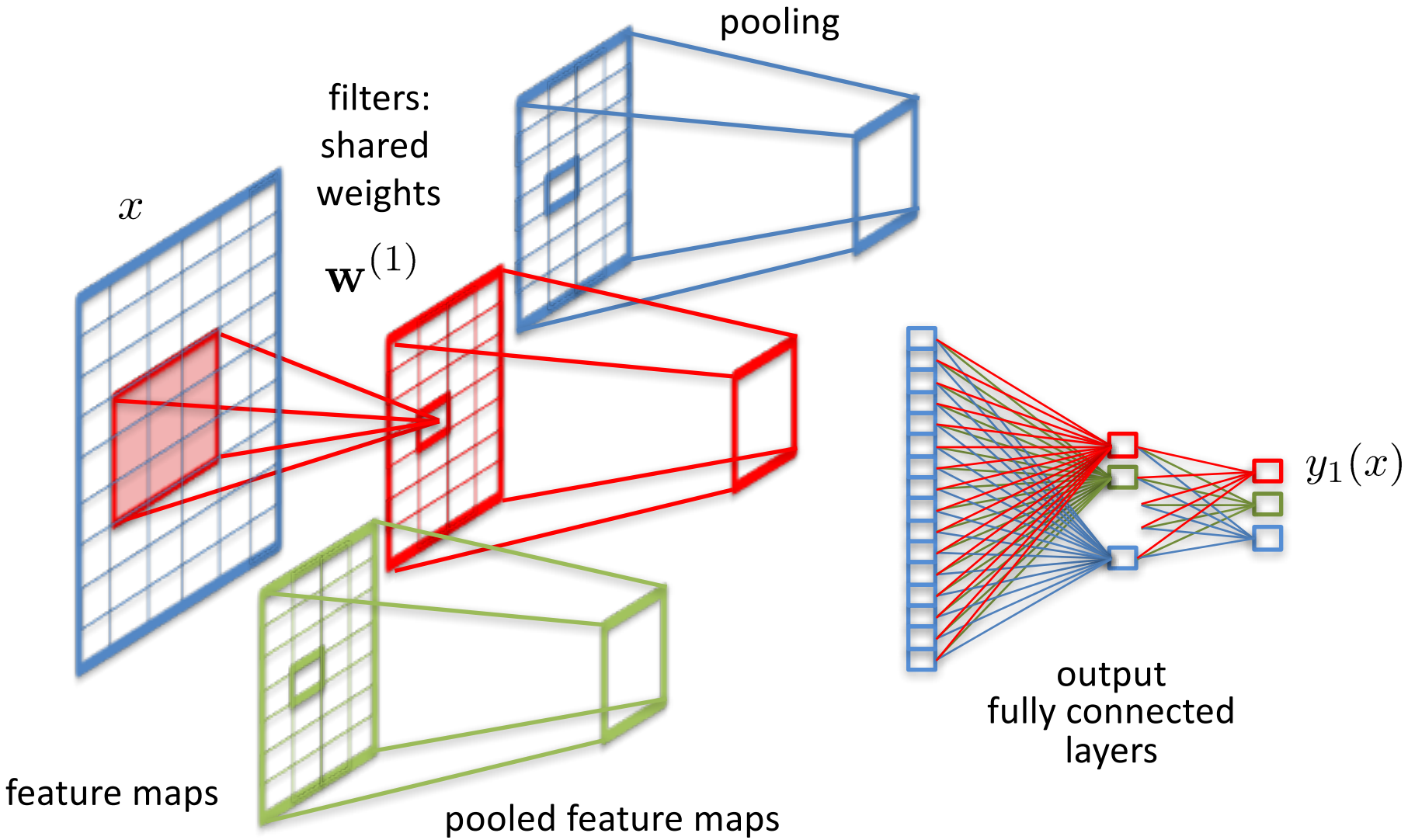
# What is wrong with fully connected deep neural networks?

Too many parameters:
1000] - 2 000 000 parameters

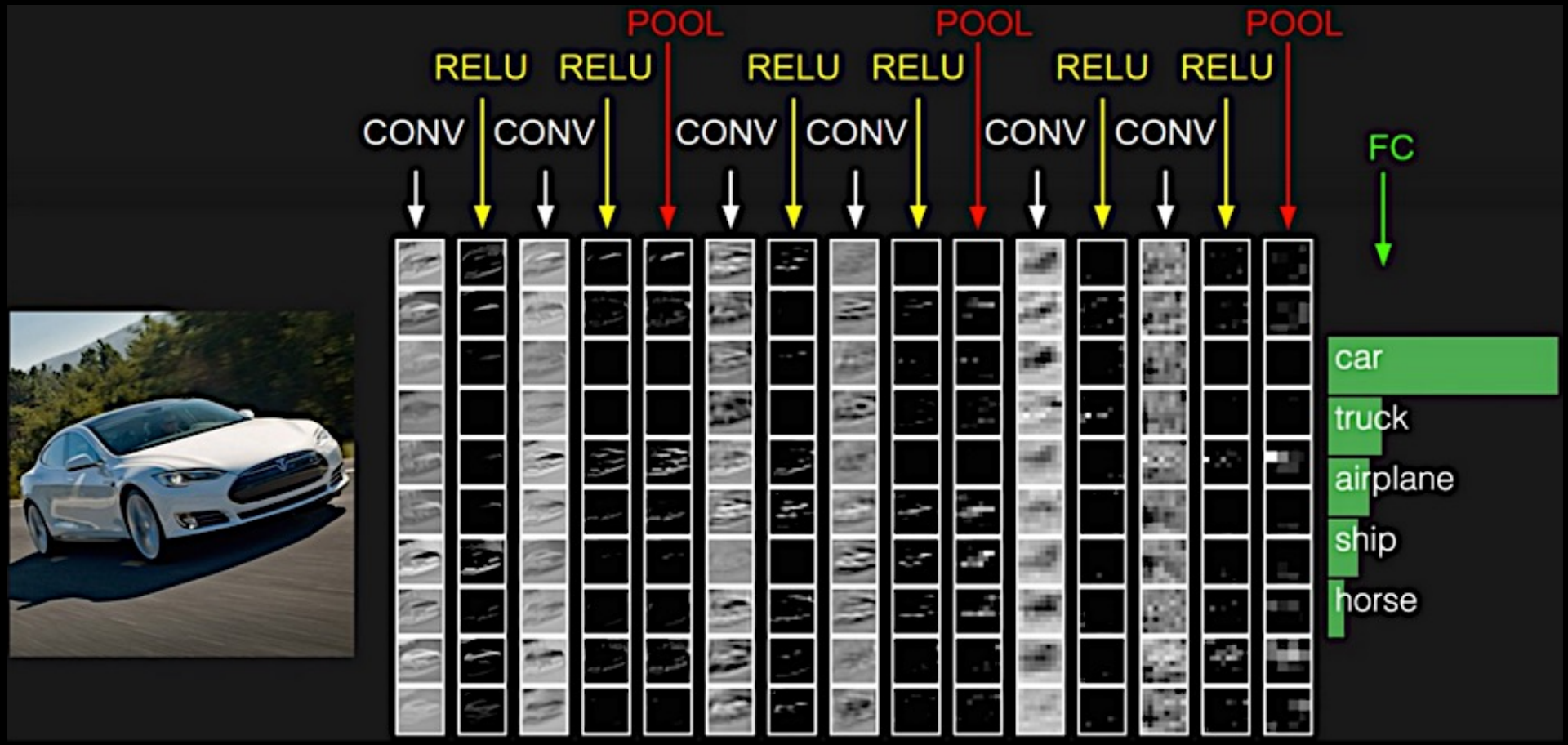
MLP[1000-1000-
require too much
training data

Wasteful and do not generalize:
no spatial invariance for images

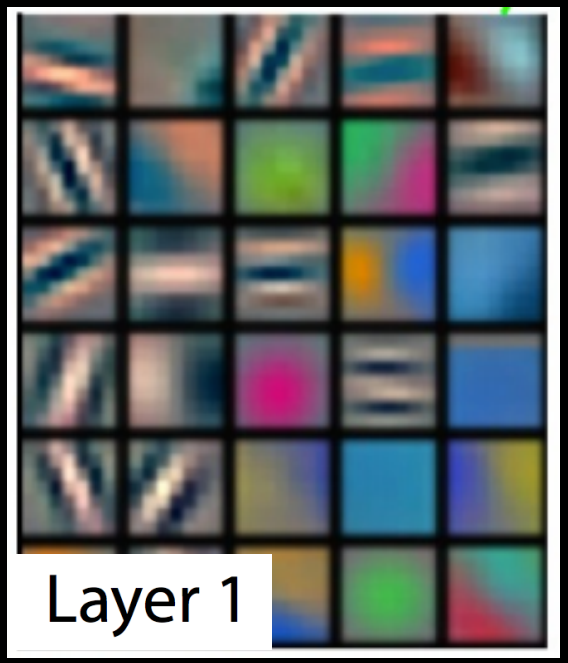
# Convolutional Neural Network (CNN, Convnet)



# What do we learn?

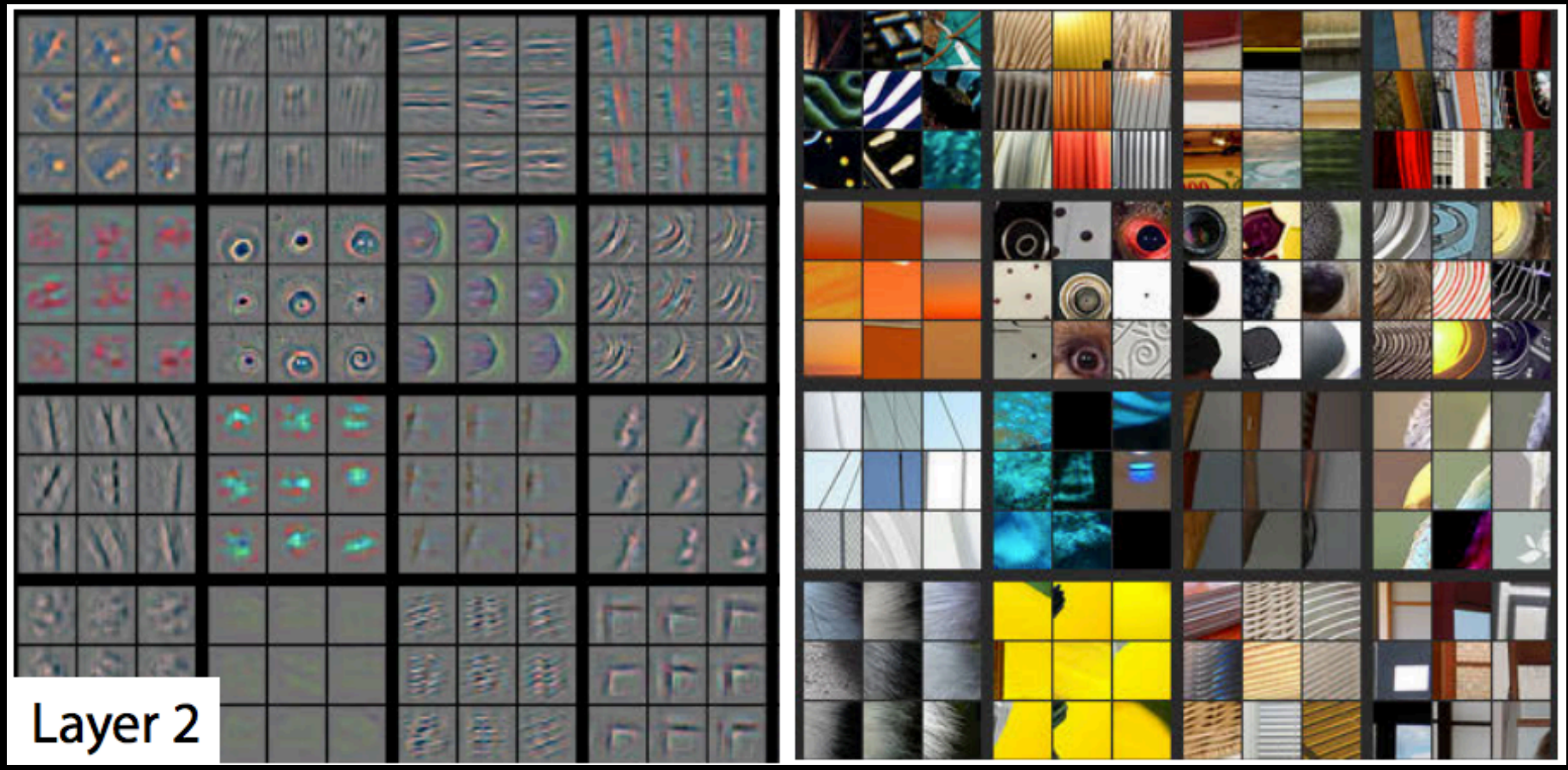


# What do we learn?



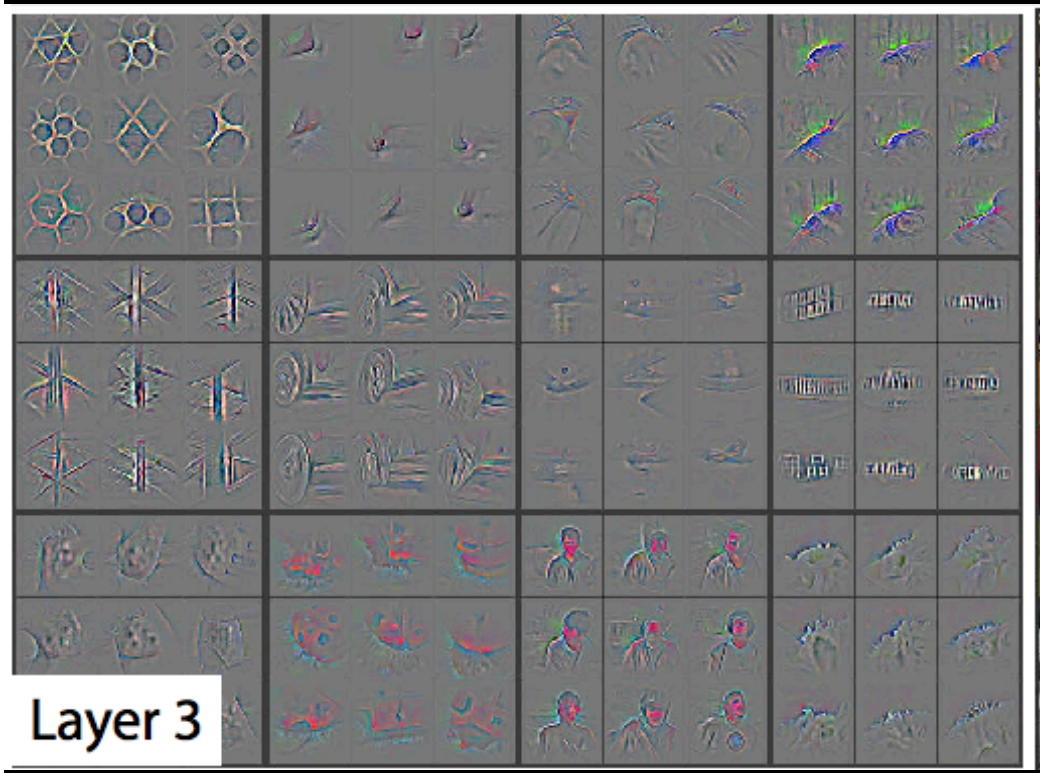

[Zeiler, Fergus, 2013]

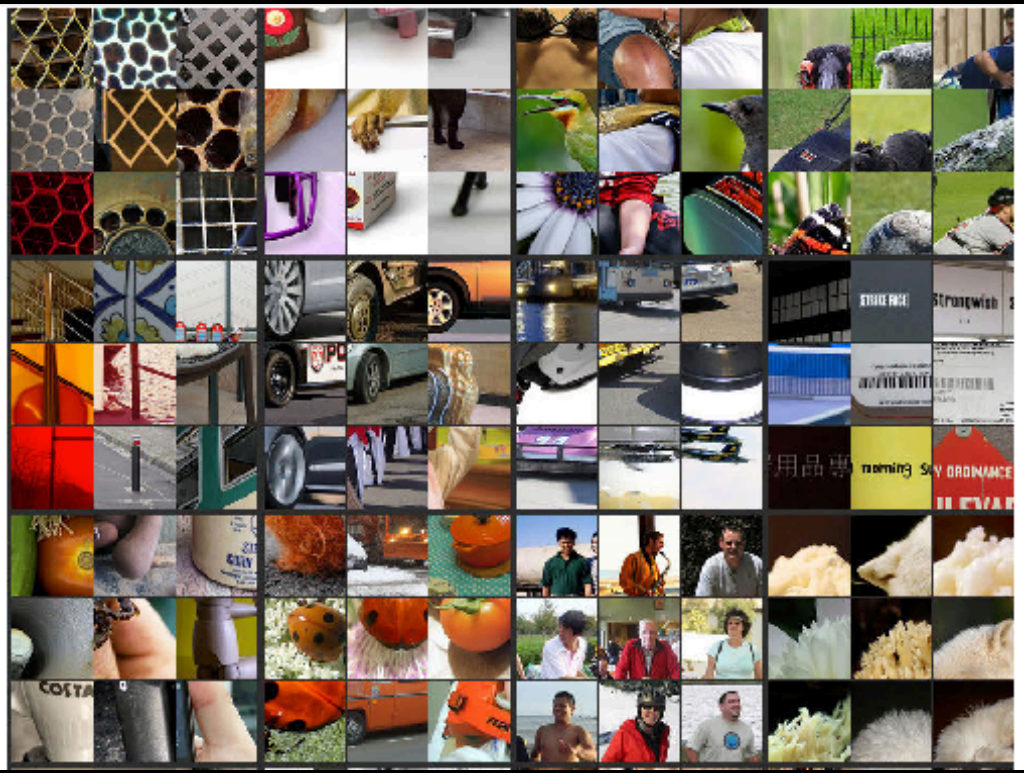
# What do we learn?



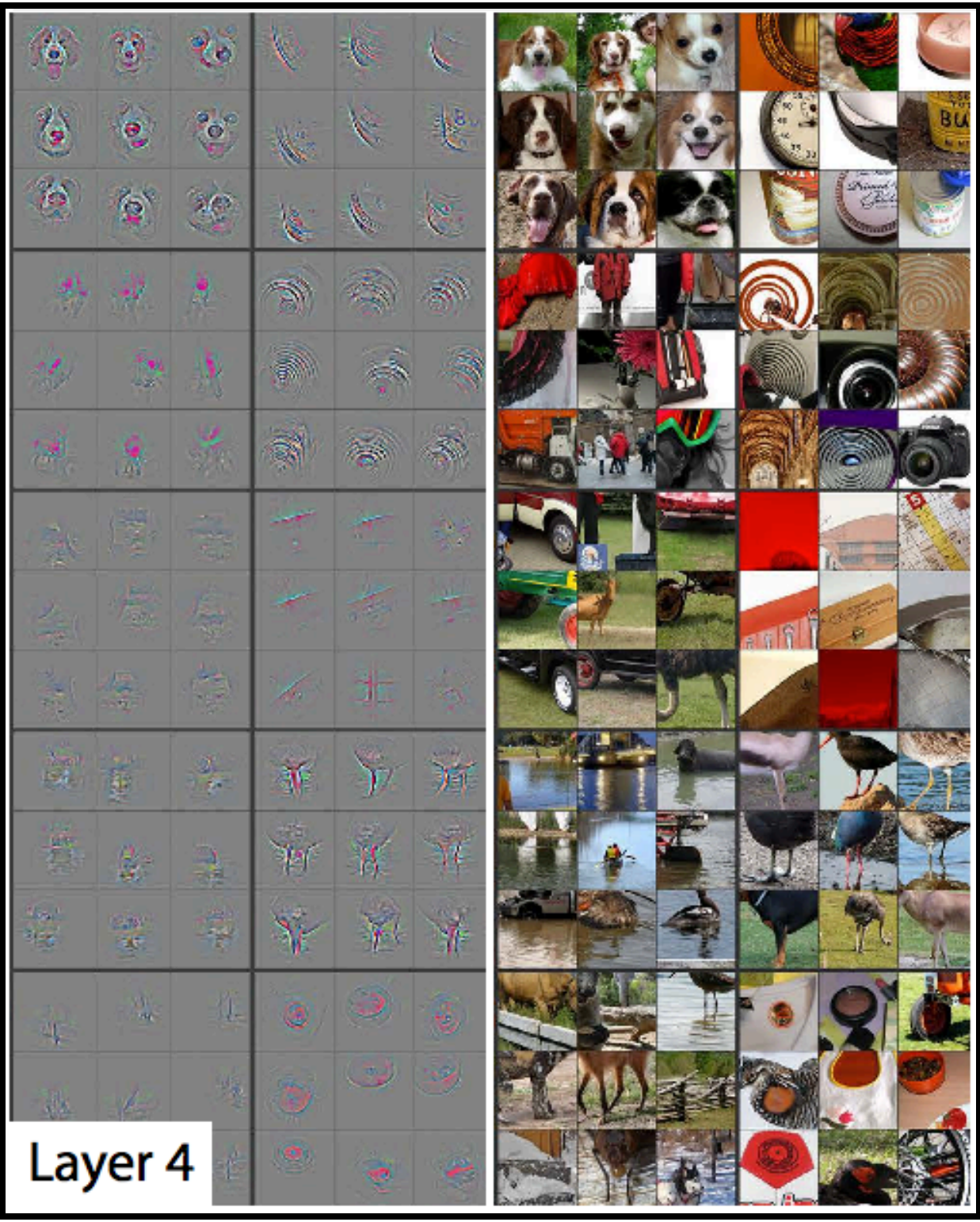

[Zeiler, Fergus, 2013]

# What do we learn?

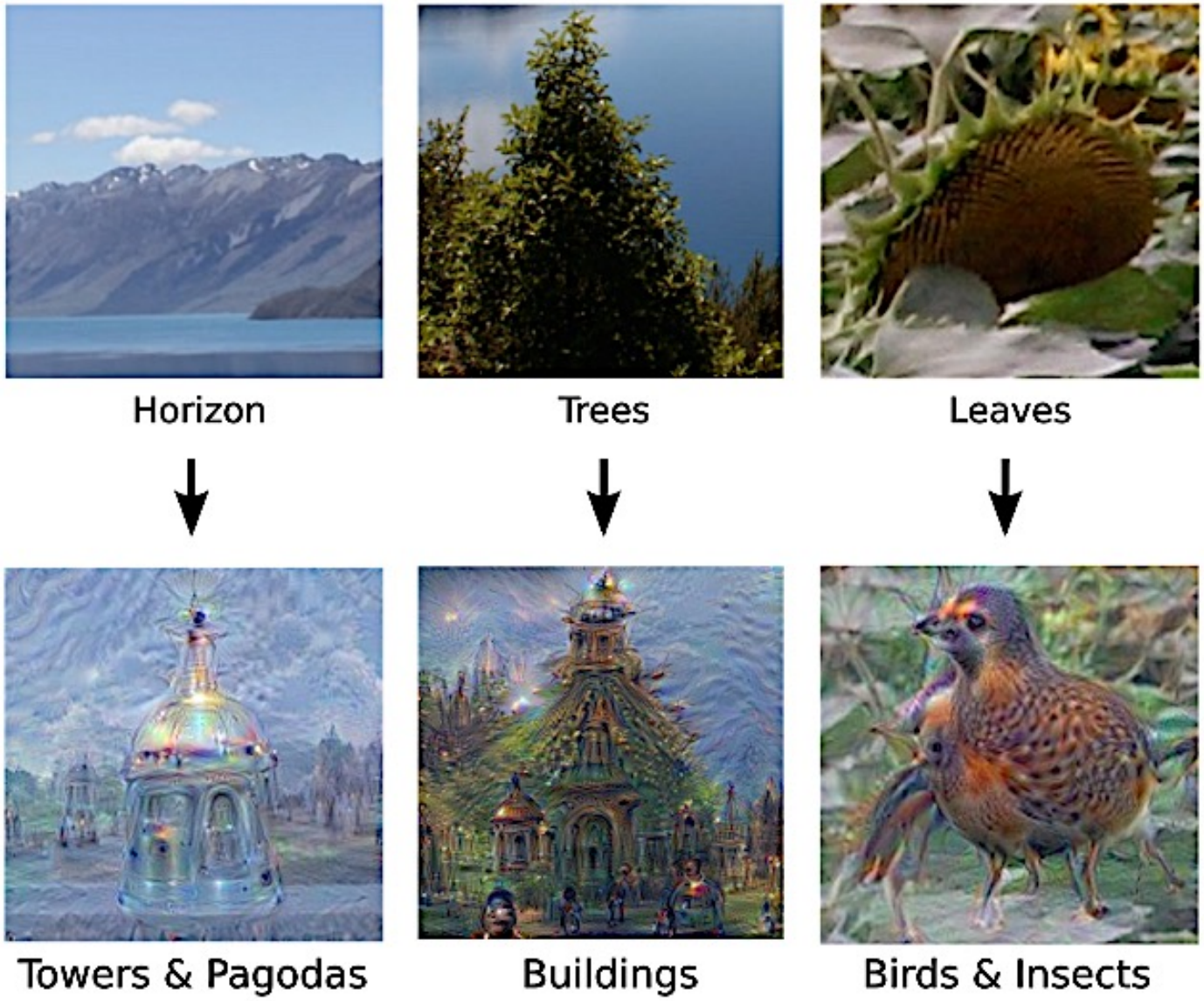


[Zeiler, Fergus, 2013]

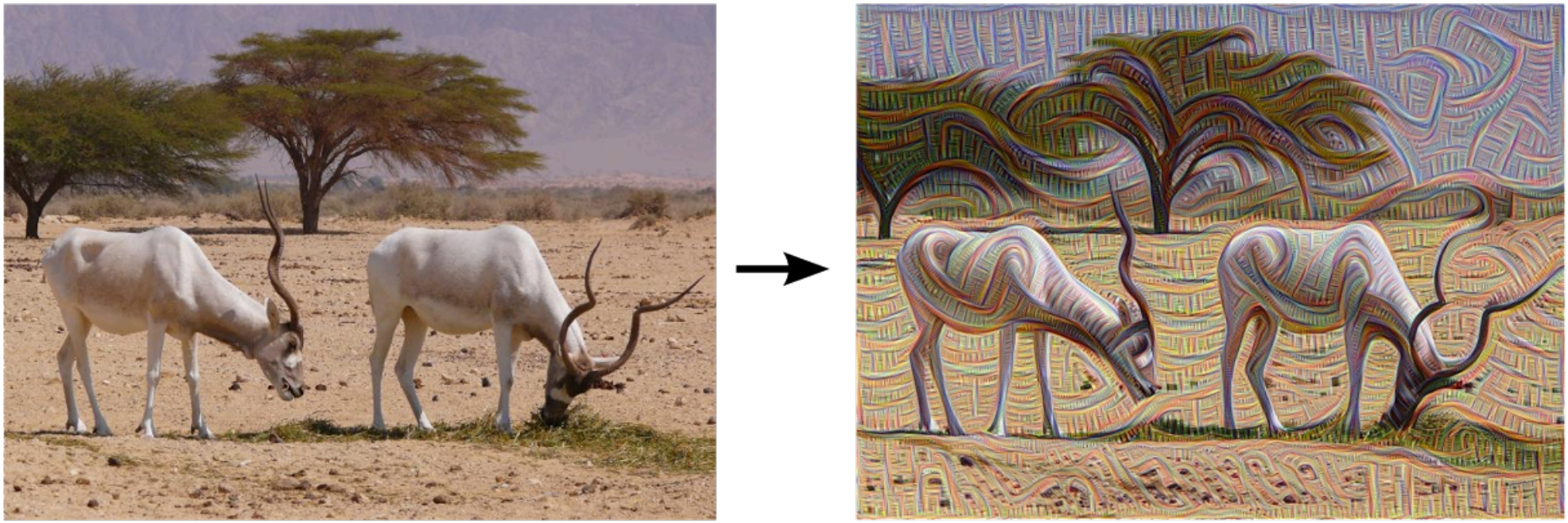

Layer 4

# “Deep dreaming”



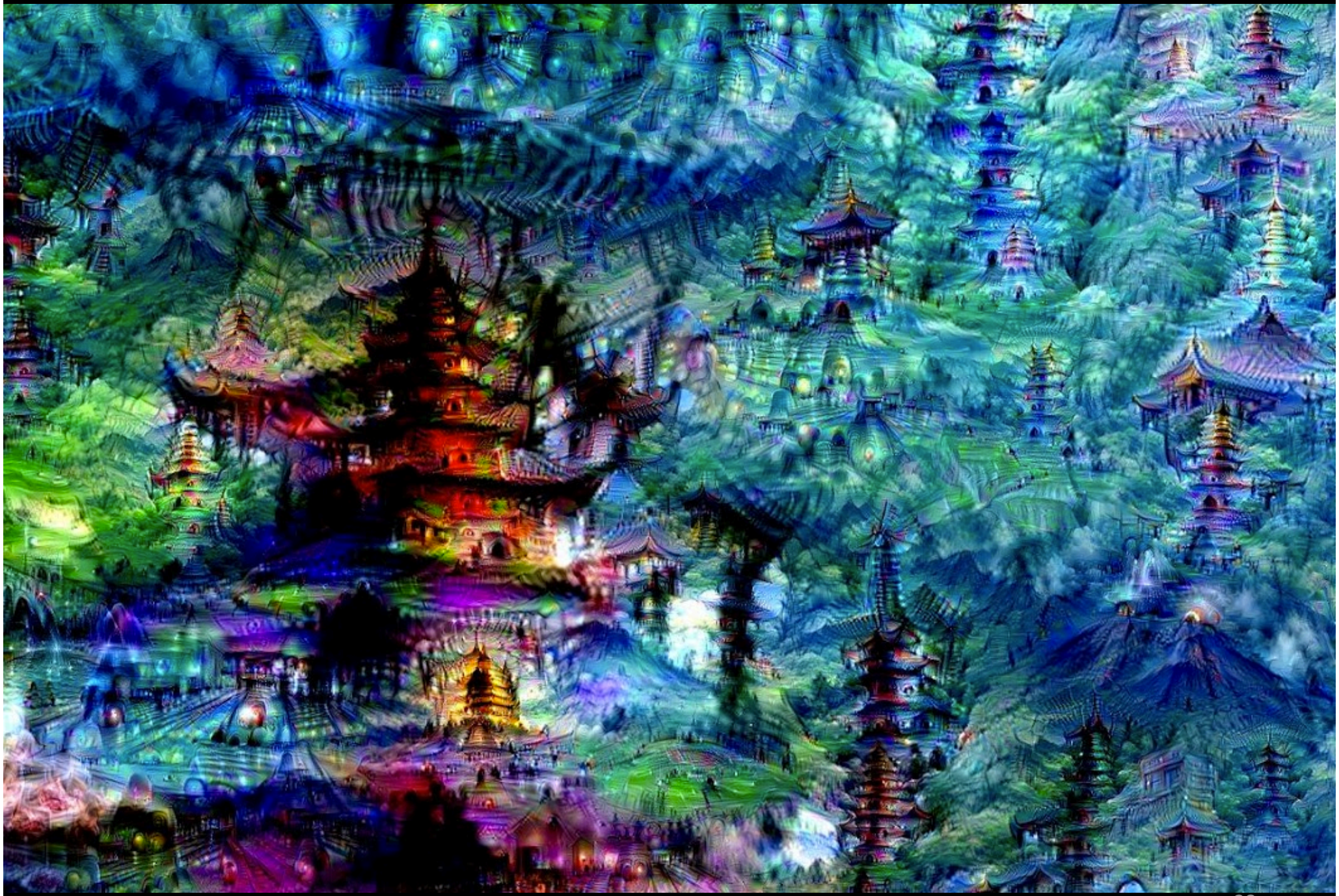

[http://googleresearch.blogspot.fr/]

# “Deep dreaming”



First-layer activations

[http://googleresearch.blogspot.fr/]

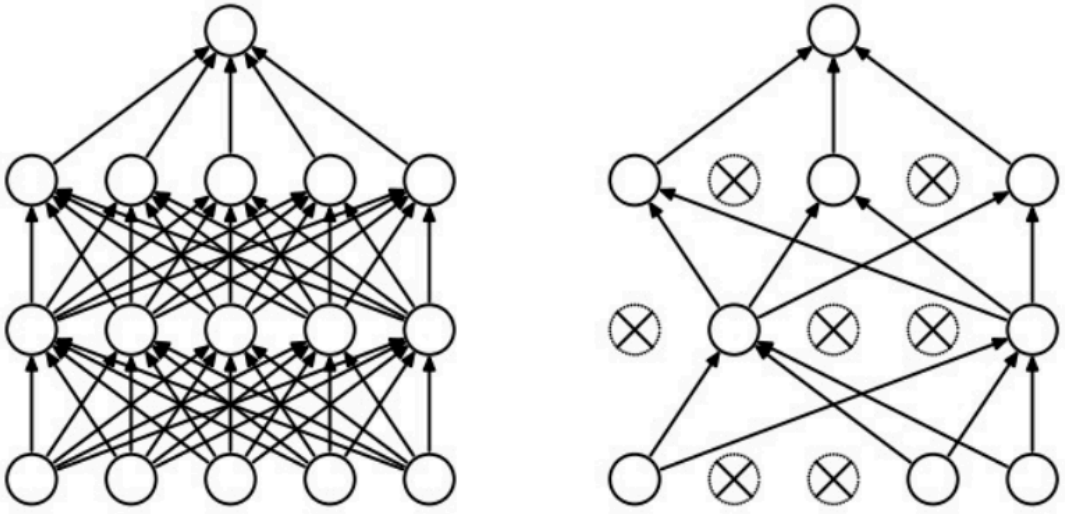


# Practical issues with babysitting neural networks

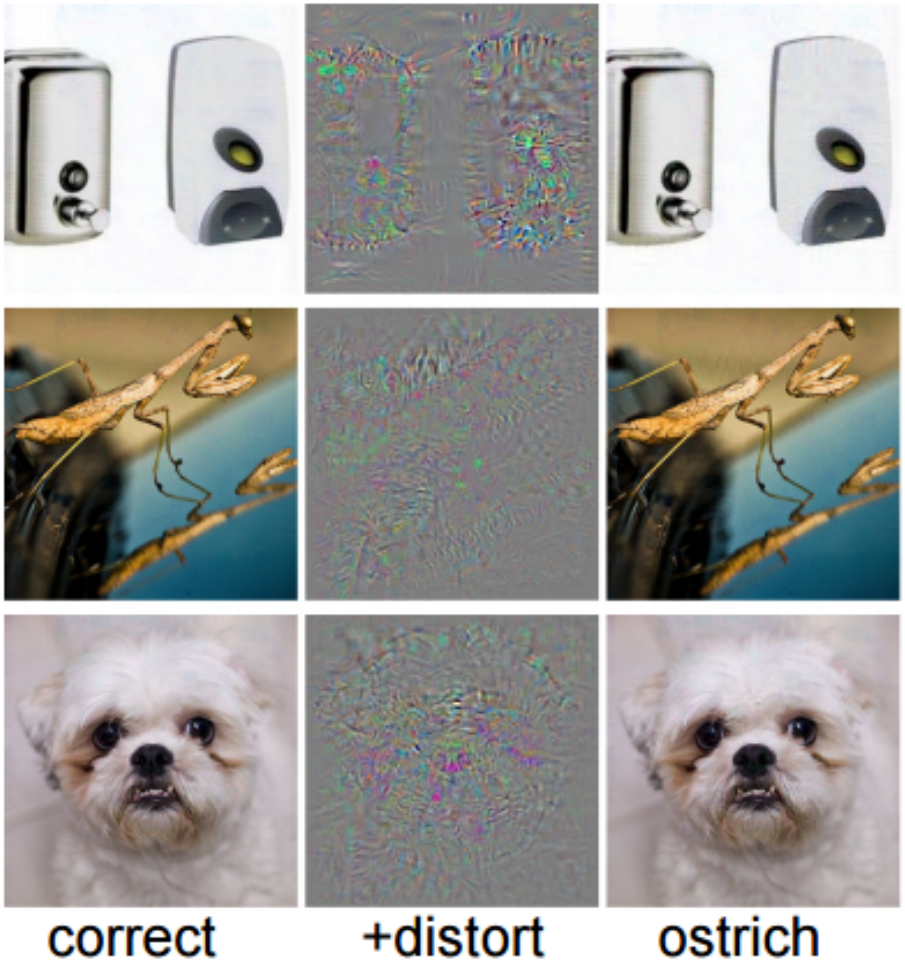
- Data augmentation
- Data normalization
- Architecture optimization
- Loss function
- Weight initialization
- Learning rate and its decay schedule
- Overfitting: regularization, early stopping
- Momentum

# Network regularization: dropout

- During training, for each sample, 50% of the units disabled
- Punishes co-adaptation of units
- Large performance gains



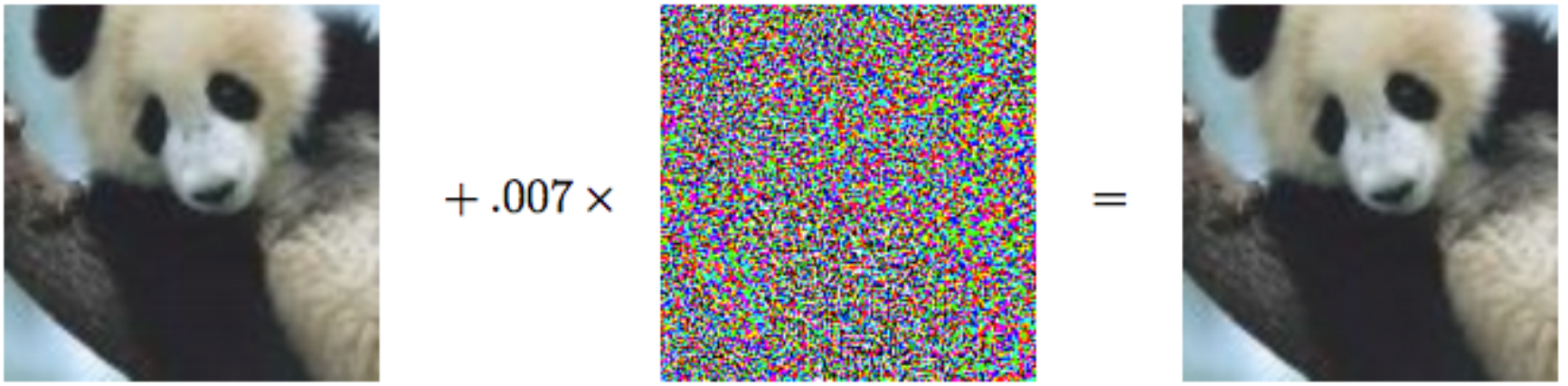
# Intriguing properties of ConvNets


[Szegedy et al, 2013]

# Training with adversarial samples

$$\boldsymbol{w}^\top \tilde{\boldsymbol{x}} = \boldsymbol{w}^\top \boldsymbol{x} + \boldsymbol{w}^\top \boldsymbol{\eta}$$




$\boldsymbol{x}$

"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

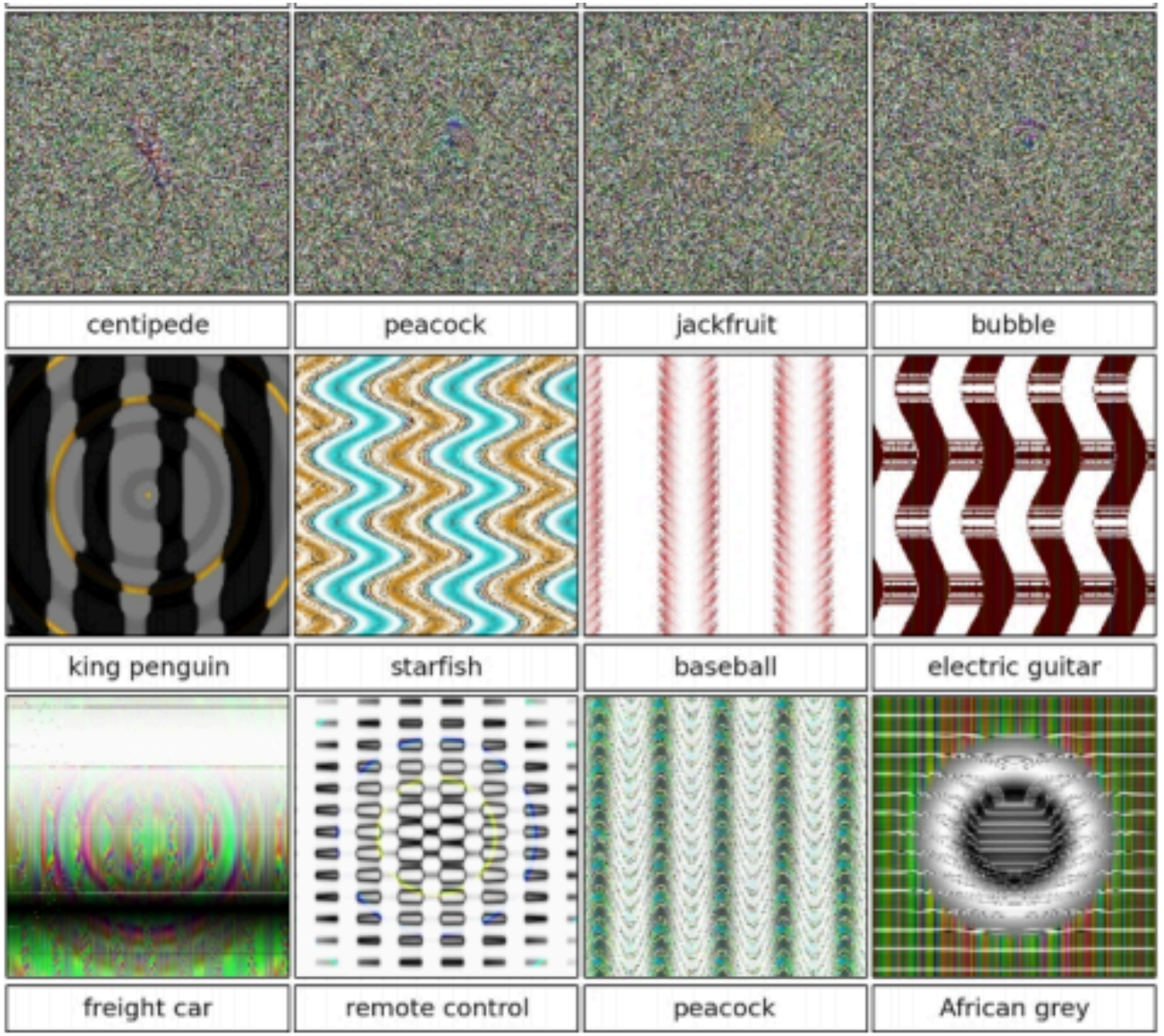
"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

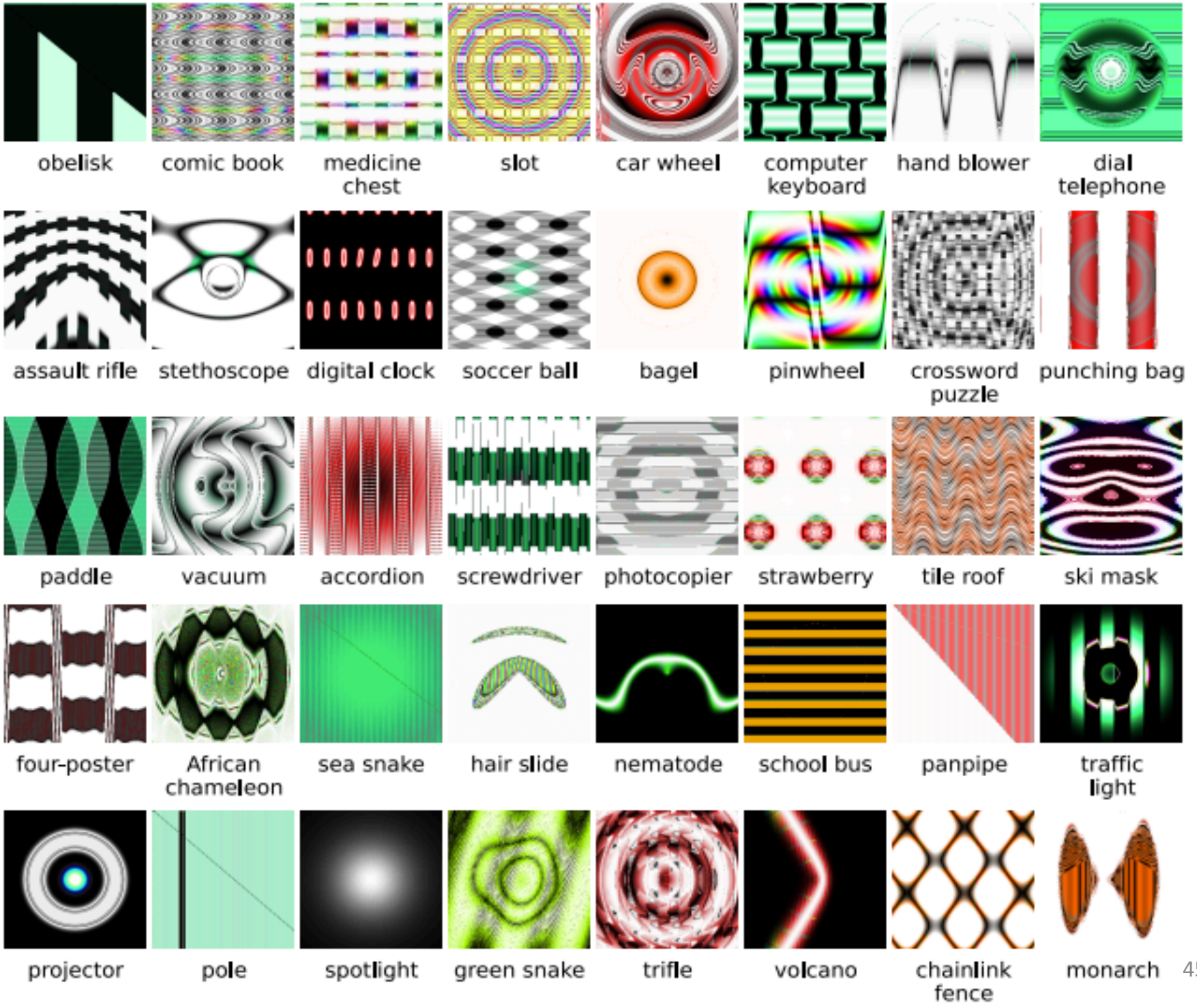
"gibbon"
99.3 % confidence

[Goodfellow et al, 2014]

# Intriguing properties of ConvNets

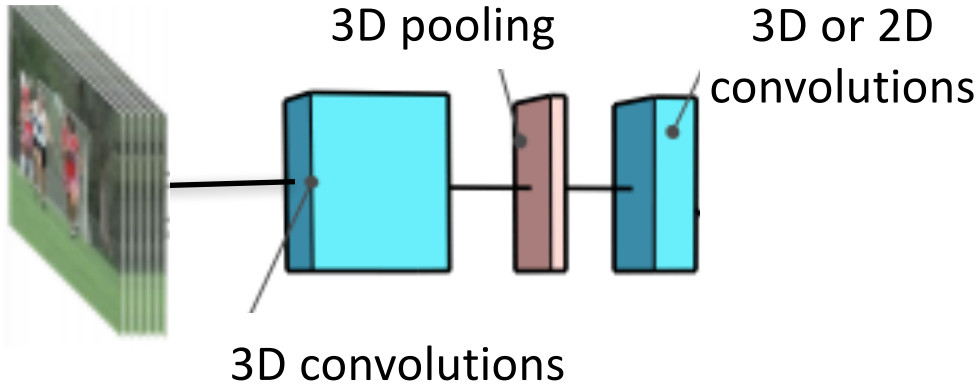



[Nguen et al, CVPR 2015]


obelisk
comic book
medicine chest
slot
car wheel
computer keyboard
hand blower
dial telephone
assault rifle
stethoscope
digital clock
soccer ball
bagel
pinwheel
crossword puzzle
punching bag
paddle
vacuum
accordion
screwdriver
photocopier
strawberry
tile roof
ski mask
four-poster
African chameleon
sea snake
hair slide
nematode
school bus
panpipe
traffic light
projector
pole
spotlight
green snake
trifle
volcano
chainlink fence
monarch

# What about sequential data?

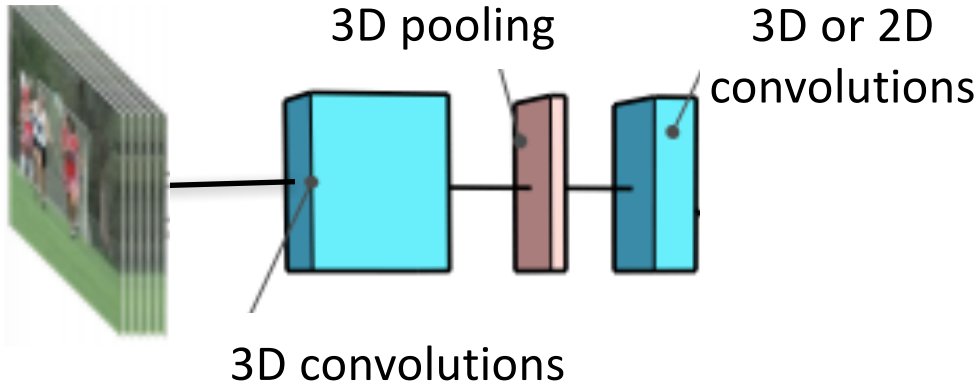
Video: spatio-temporal blocks (fixed length, simple patterns)



# What about sequential data?

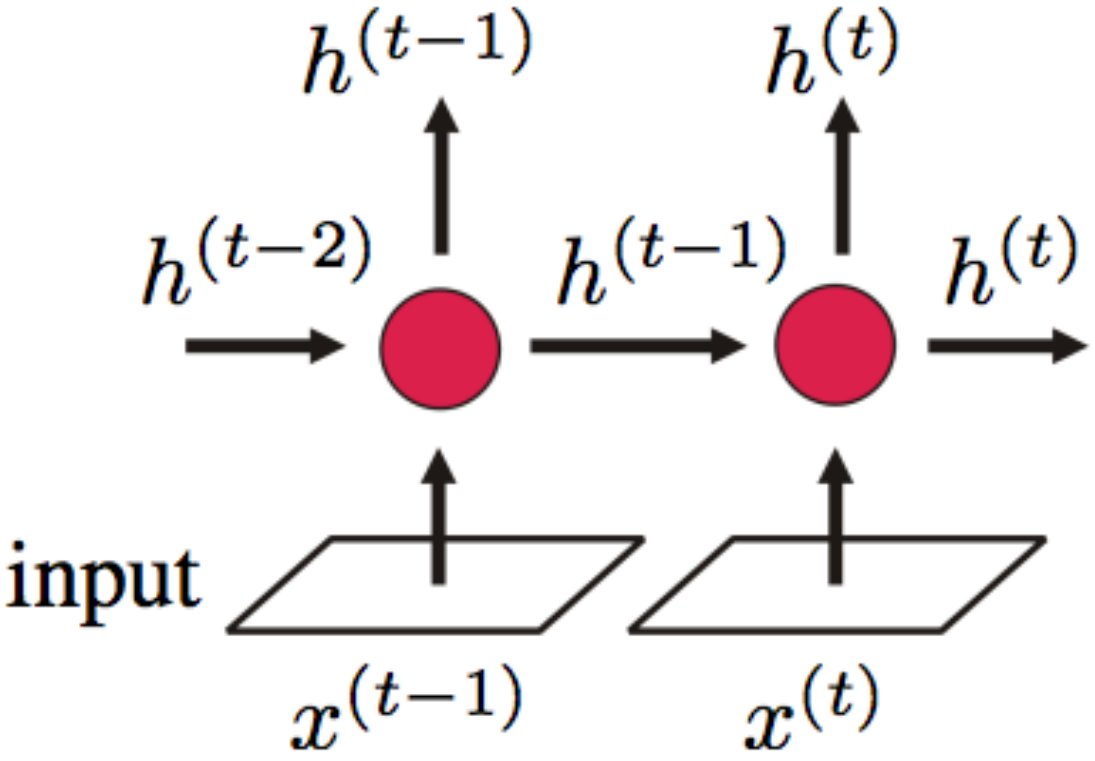
# What about sequential data?

Video: spatio-temporal blocks (fixed length, simple patterns)



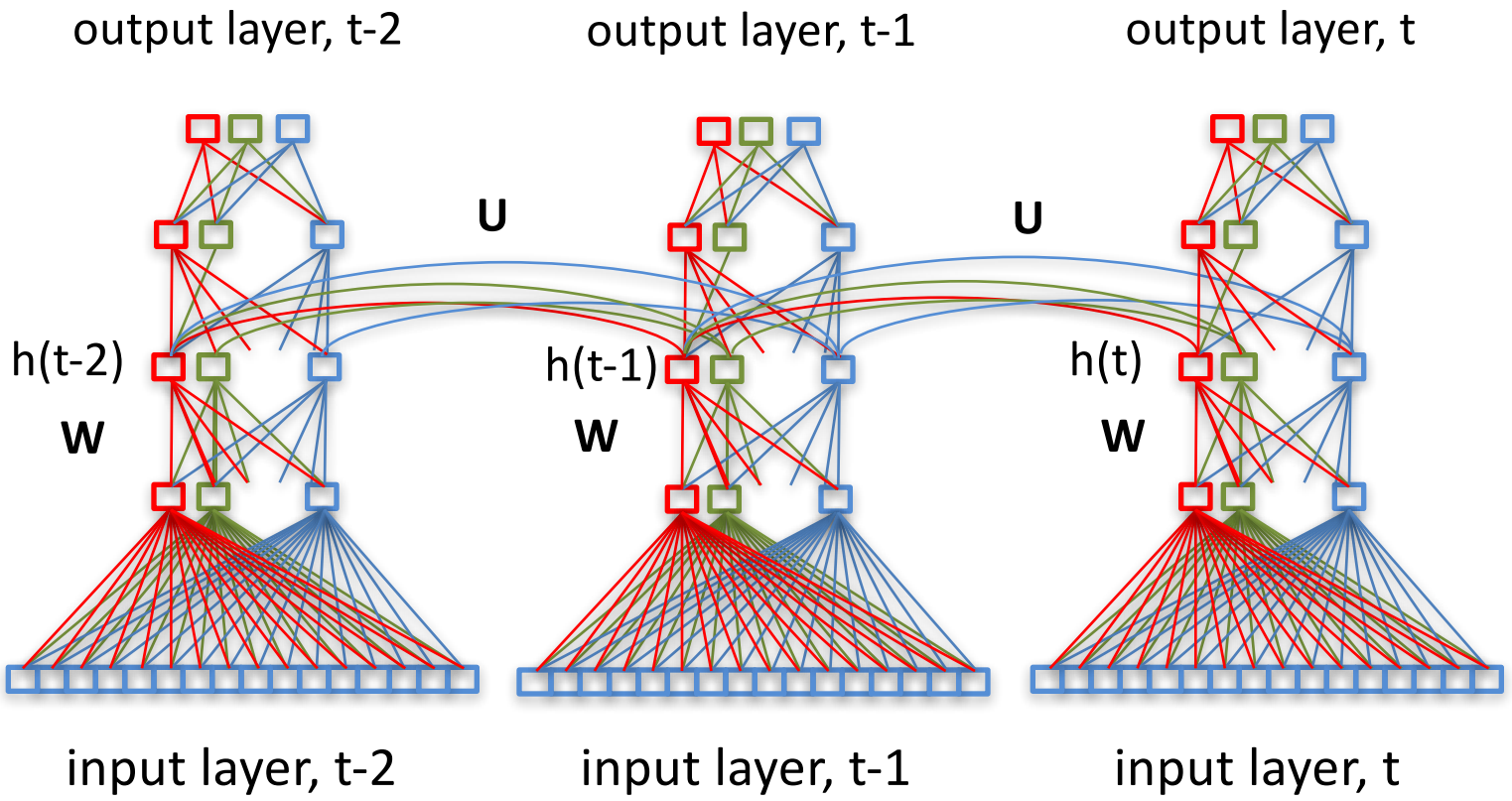

Alternatively: temporal modeling
of short- and long-term dependencies

# Recurrent Neural Network (RNN)

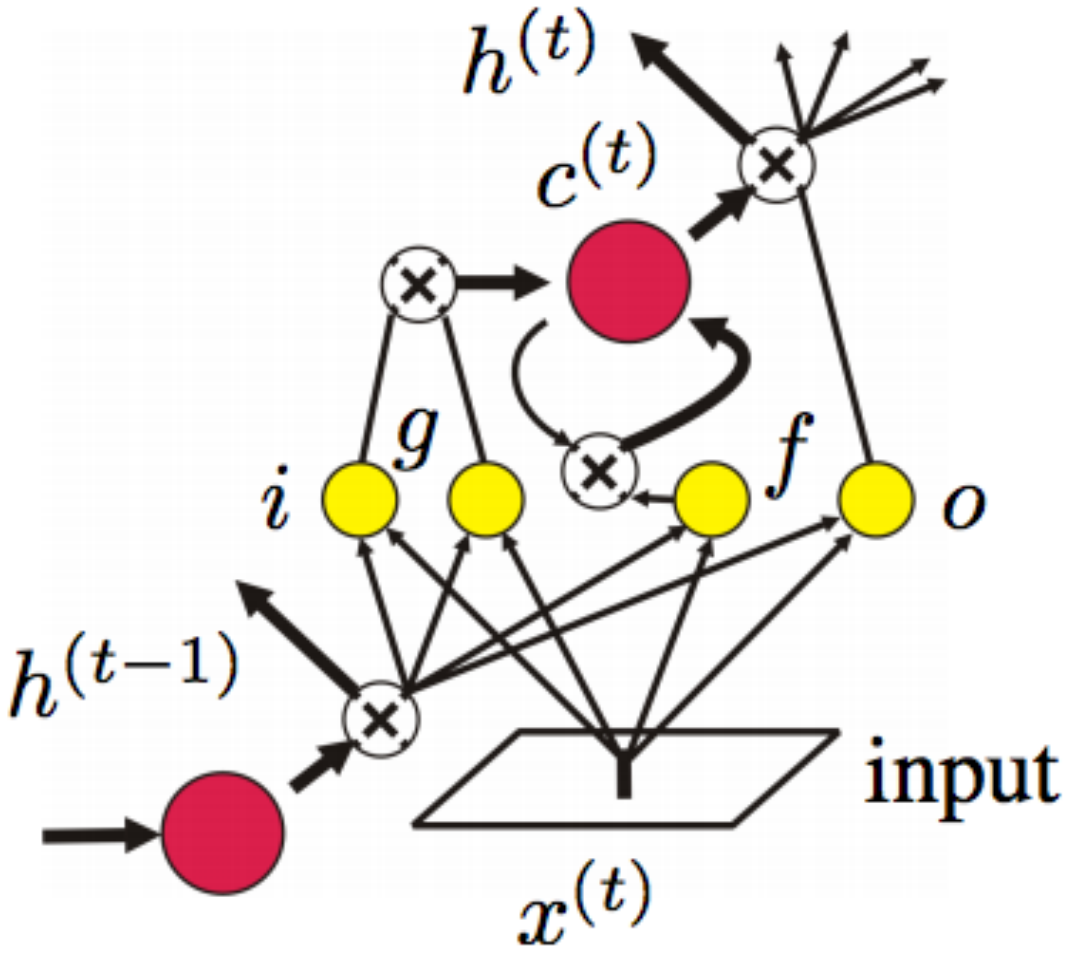


# Recurrent Neural Network (RNN)

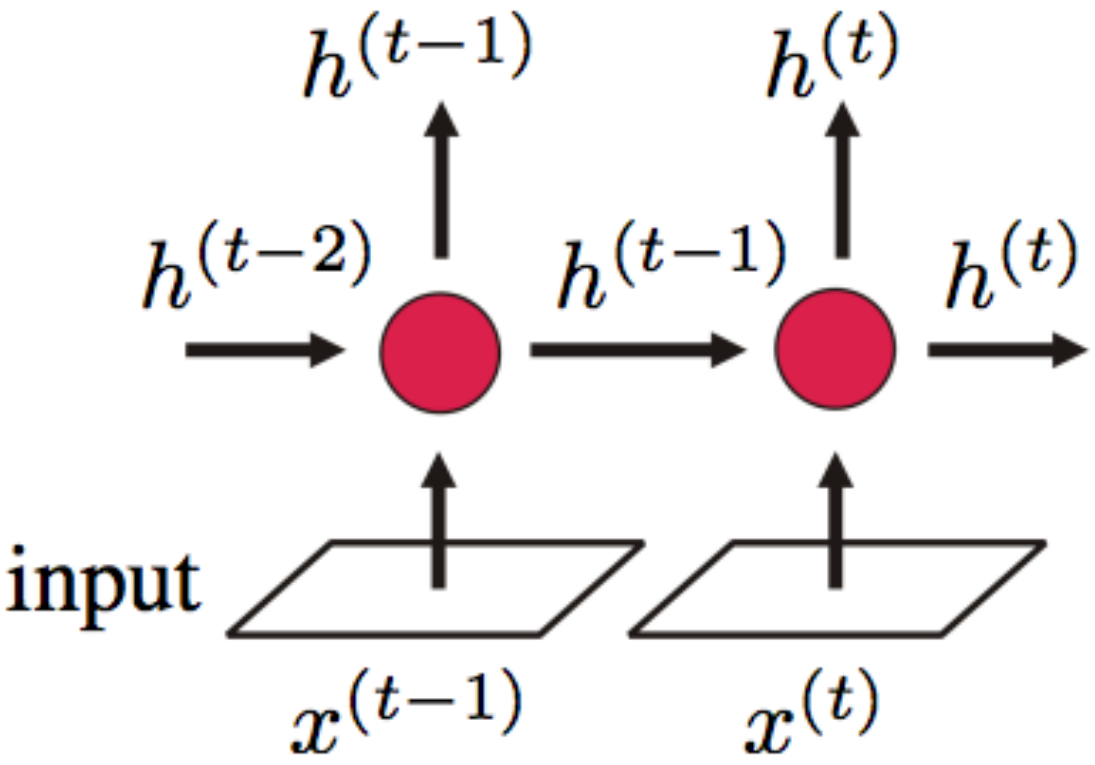
$$h^{(t)} = \psi(Wx^{(t)} + Uh^{(t-1)})$$



# Long Short Term Memory (LSTM)



# Generative temporal models



# Leo Tolstoy's "War and Peace"

Iteration 100:

```
tyntd-iafhatawiaoihrdemot  lytdws  e ,tfti, astai f ogoh eoase rrranbyne 'nhthnee e
plia tklrgd t o idoe ns,smtt   h ne etie h,hregtrs nigtike,aoaenns lng
```

Iteration 300:

```
"Tmont thithey" fomesscerliund
Keushey. Thom here
sheulke, anmerenith ol sivh I lalterthend Bleipile shuwy fil on aseterlome
coaniogennc Phe lism thond hon at. MeiDimorotion in ther thize."
```

Iteration 500:

```
we counter. He stutn co des. His stanted out one ofler that concossions and was
to gearang reay Jotrets and with fre colt otf paitt thin wall. Which das stimn
```

[http://karpathy.github.io/2015/05/21/rnn-effectiveness/]

# Leo Tolstoy's "War and Peace"

Iteration 700:

```
Aftair fall unsuch that the hall for Prince Velzonski's that me of
her hearly, and behs to so arwage fiving were to it beloge, pavu say falling misfort
how, and Gogition is so overelical and ofter.
```

Iteration 1200:

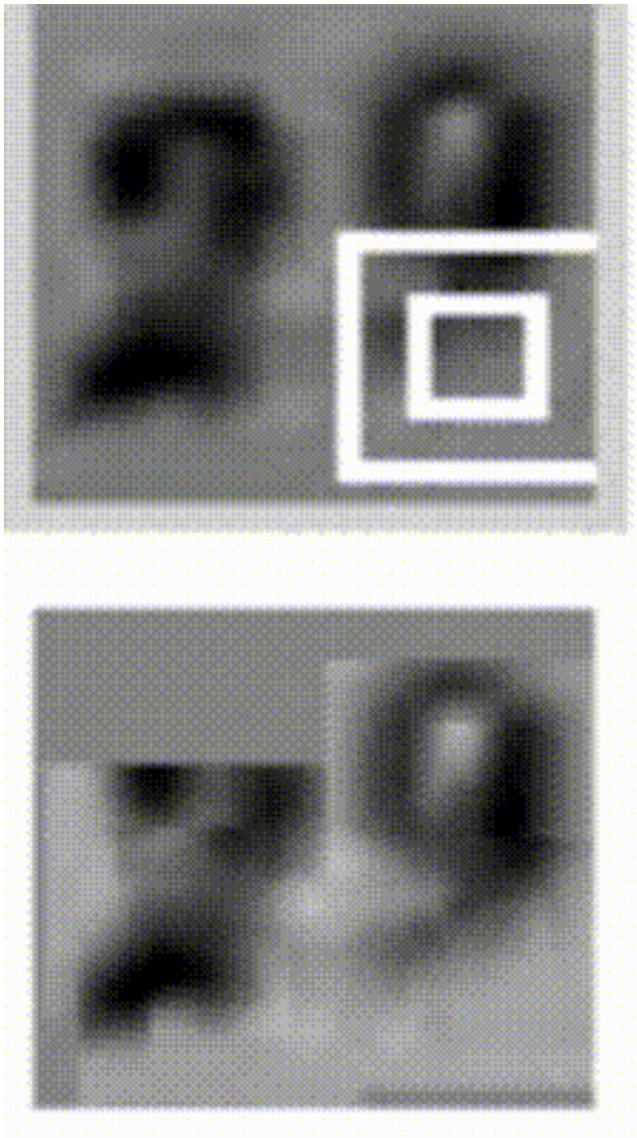
```
"Kite vouch!" he repeated by her
door. "But I would be done and quarts, feeling, then, son is people...."
```

Iteration 2000:

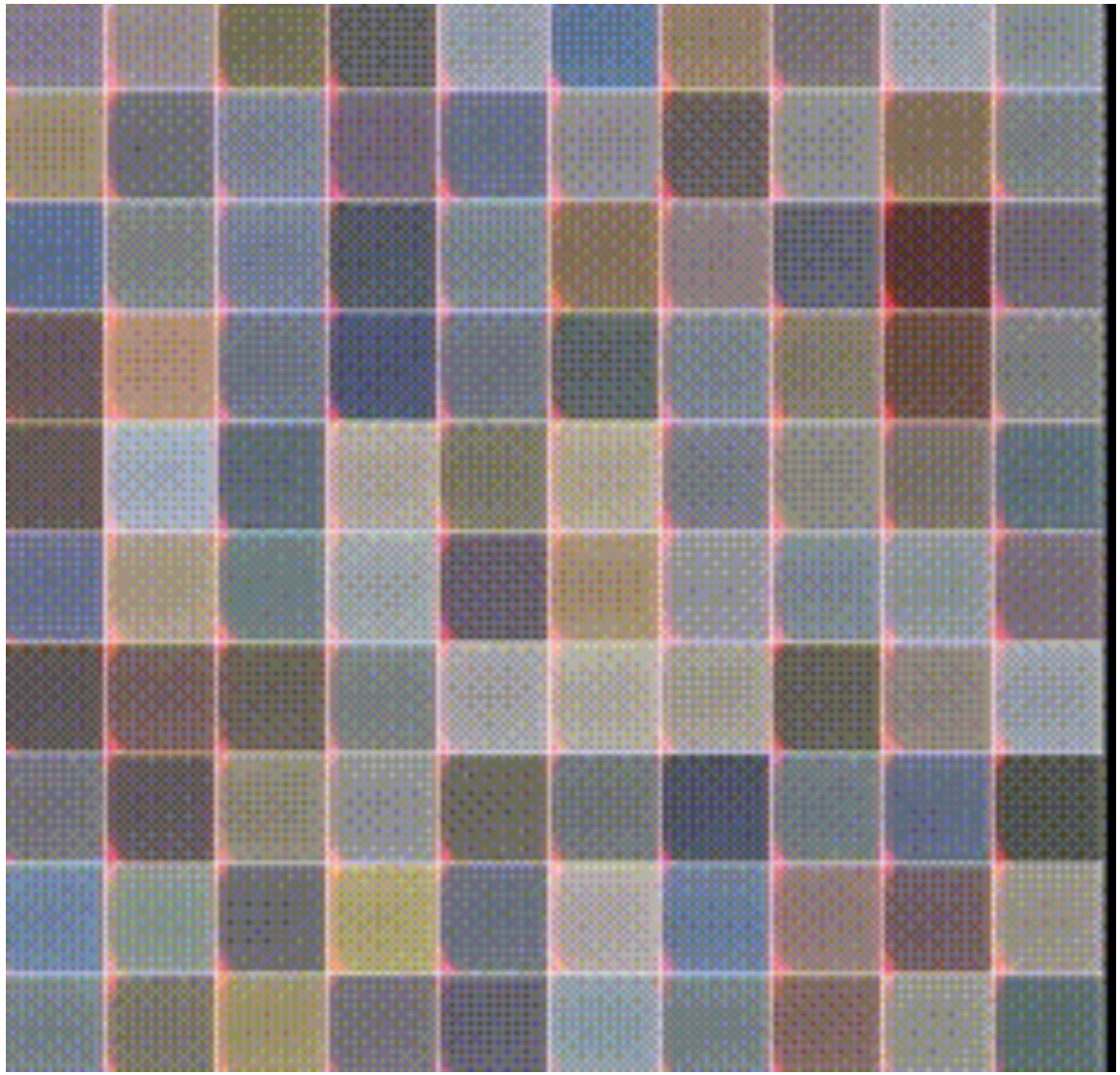
```
"Why do what that day," replied Natasha, and wishing to himself the fact the
princess, Princess Mary was easier, fed in had oftened him.
Pierre aking his soul came to the packs and drove up his father-in-law women.
```

[http://karpathy.github.io/2015/05/21/rnn-effectiveness/]

# Not only for sequential data…



[Ba et el, 2015]



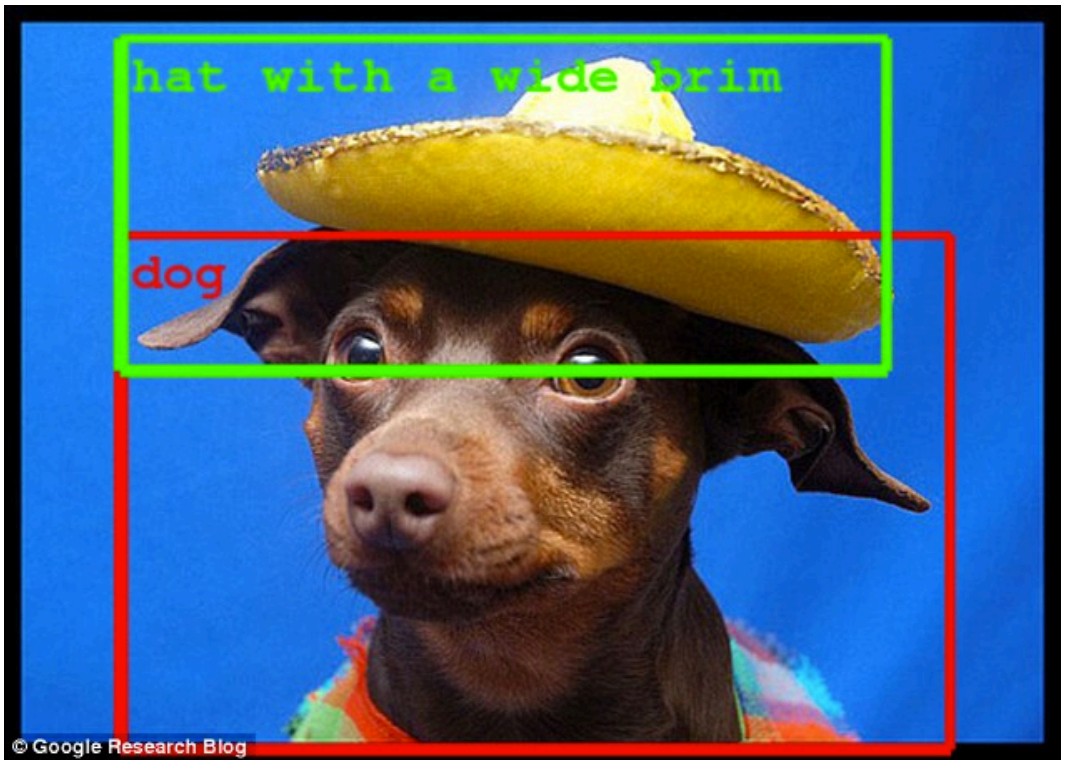
[Gregor et el, 2015]

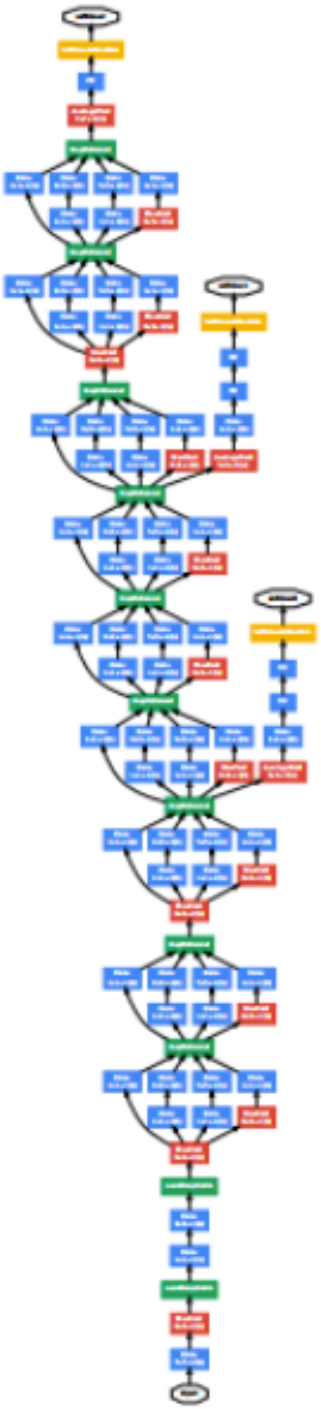
# Applications

Google

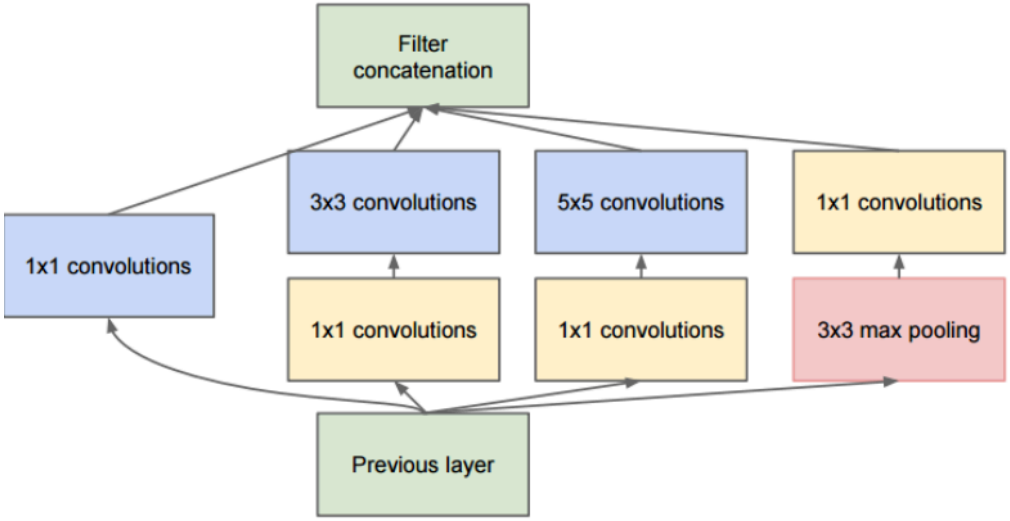
# Object recognition and localization

Deep learning based visual search engine announced in 2013
Internal dataset with 100 000 000 labelled images and 18 000 classes





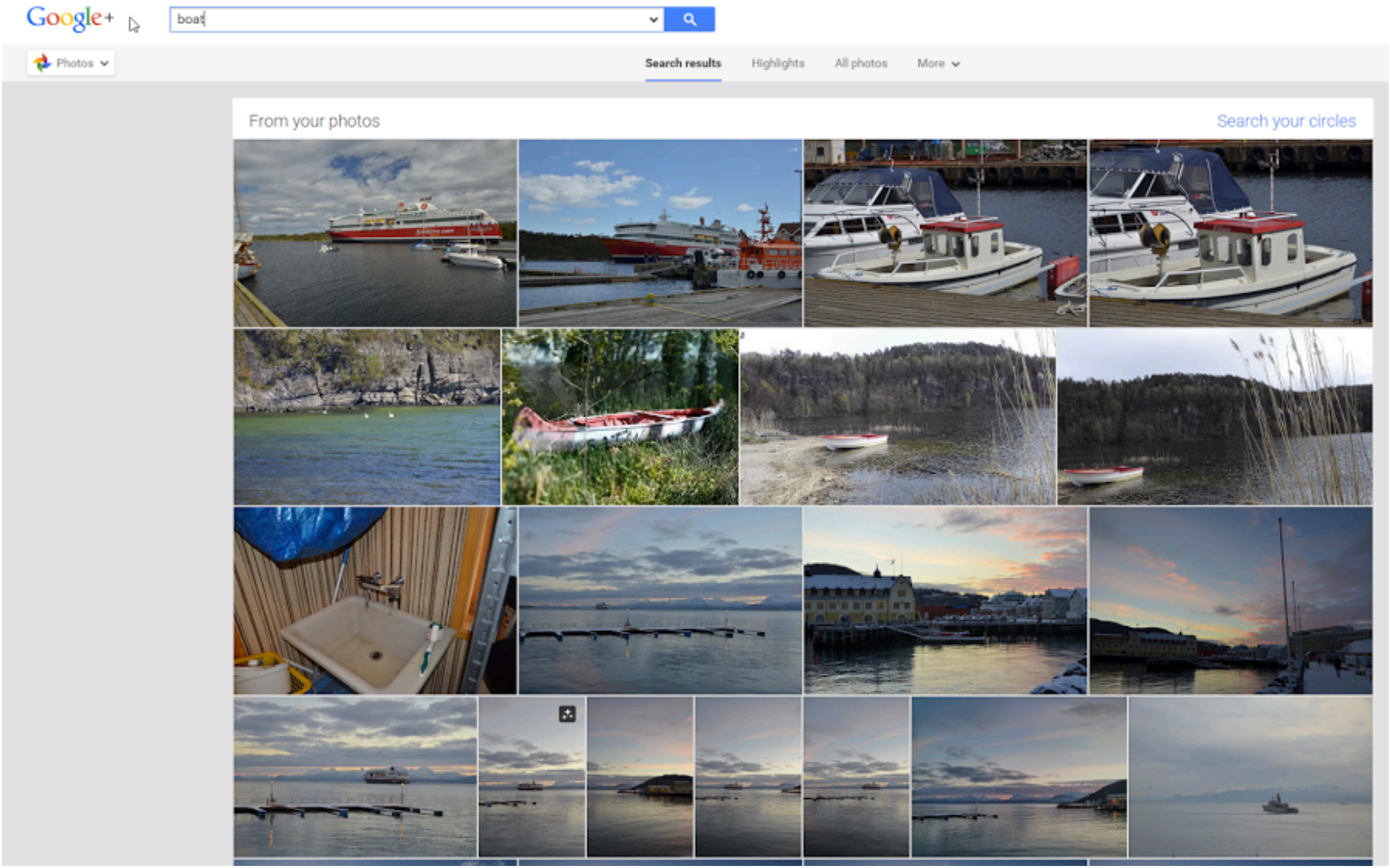
# Inception model (GoogLeNet)




22 layers with parameters
27 layers including pooling
about 100 building blocks

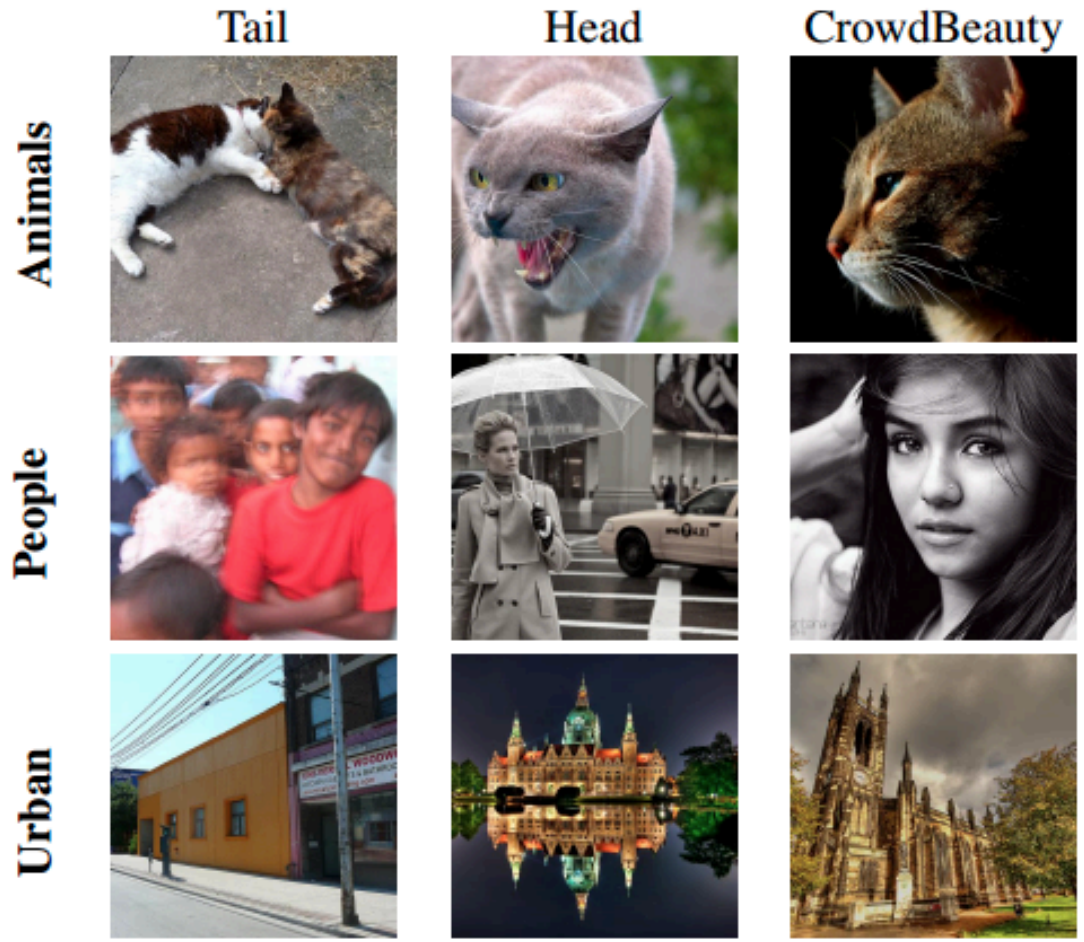
[Szegedy et el, CVPR 2015]

# photos.google.com



flickr™ YAHOO!®

# Computational aesthetics and video creativity




The Hague
ZH, Netherlands
Partly cloudy
68°
61°
68
F

Facebook AI Research

# DeepFace:
# Closing the gap to human-level performance in face verification

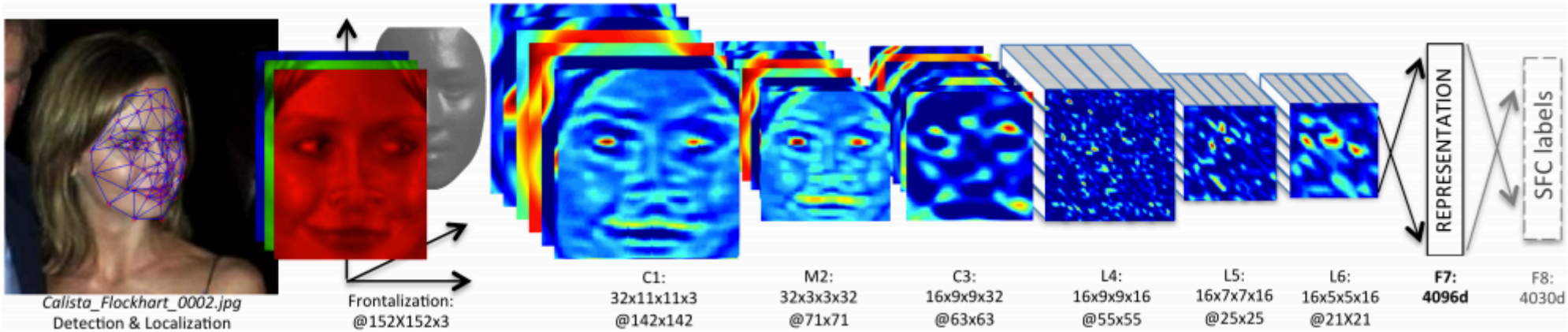
Nicole Kidman Nicole Kidman

Jacquline Obradors Julie Taymor

[Taigman et al, CVPR 2014]

Facebook AI Research

# DeepFace:
# Closing the gap to human-level performance in face verification




97.35% accuracy on the Labeled Faces in the Wild dataset

(27% improvement over the state of the art)

[Taigman et al, CVPR 2014]

Microsoft Research

# Age estimation


how-old.net

Microsoft Research

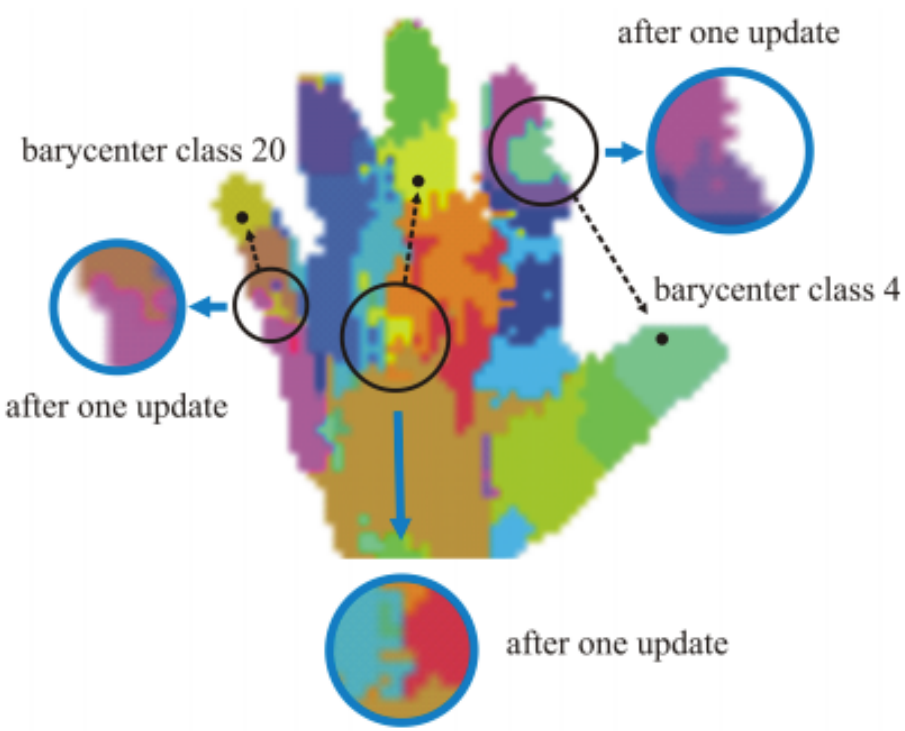
# Semantic segmentation

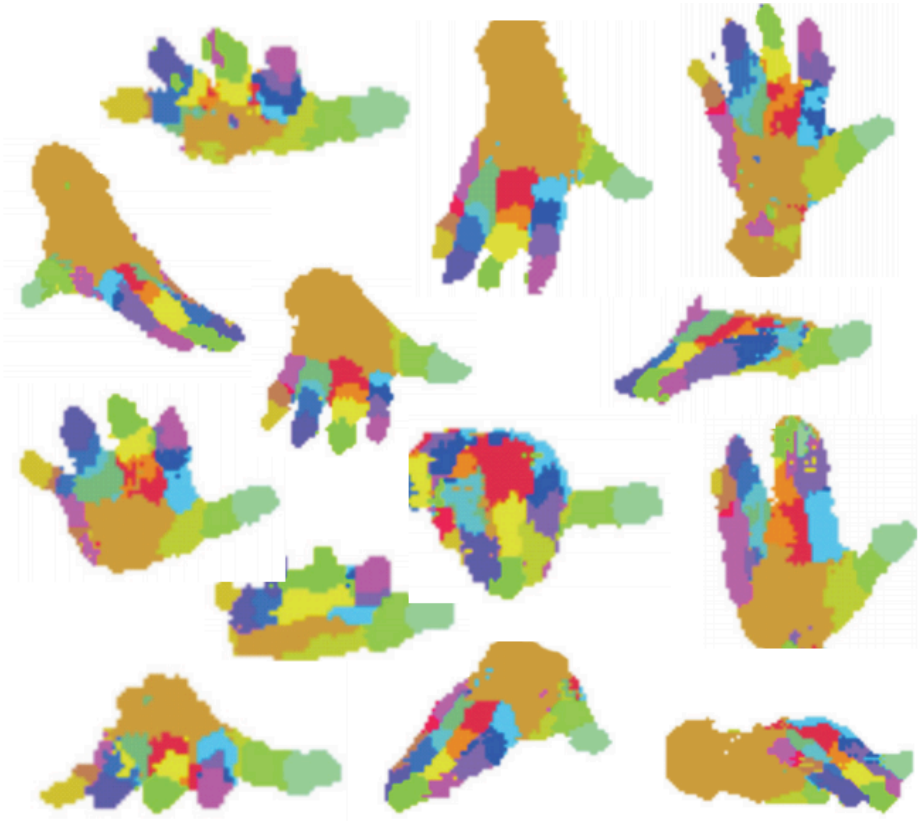

[Dai et al, CVPR 2015]

# Automatic navigation: scene labeling

[Byeon et al, CVPR 2015]

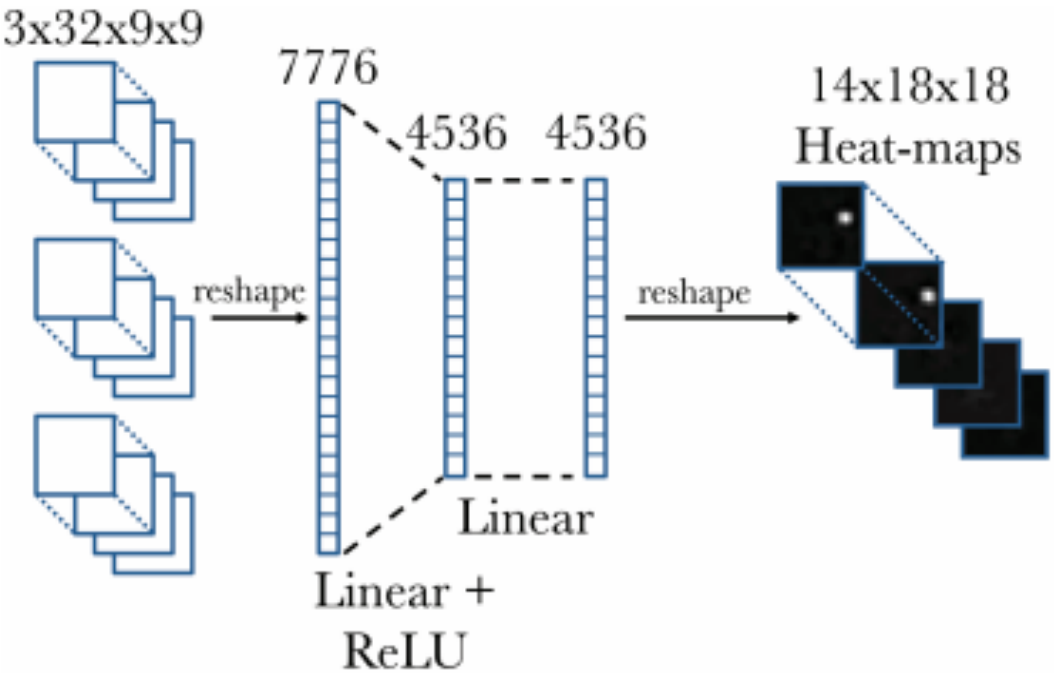
# Human-computer interaction: hand segmentation and pose estimation



[Neverova et al, ACCV 2014]

# Human-computer interaction: hand segmentation and pose estimation

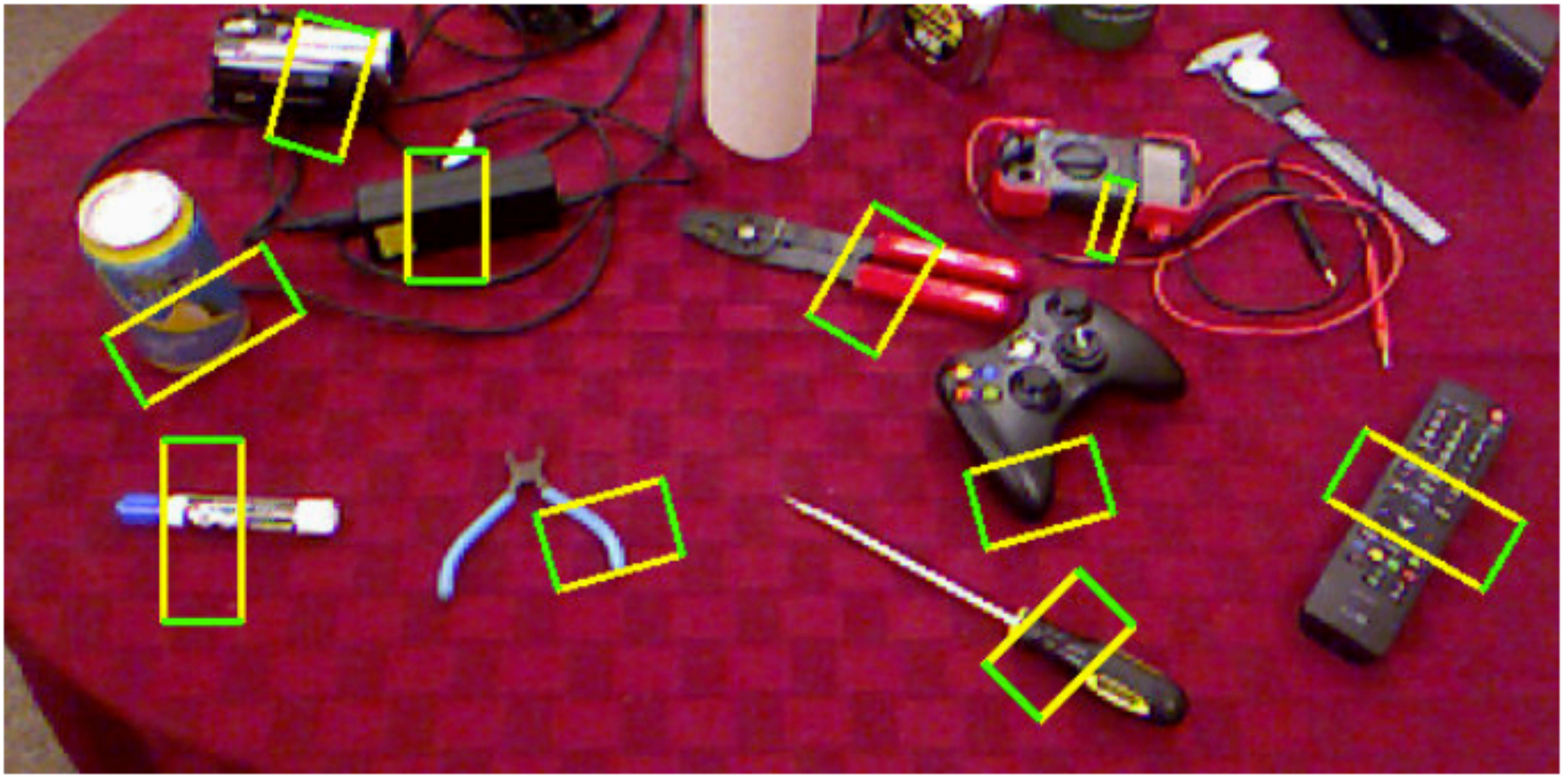



[Tompson et al, SIGGRAPH 2014]

# Human pose estimation

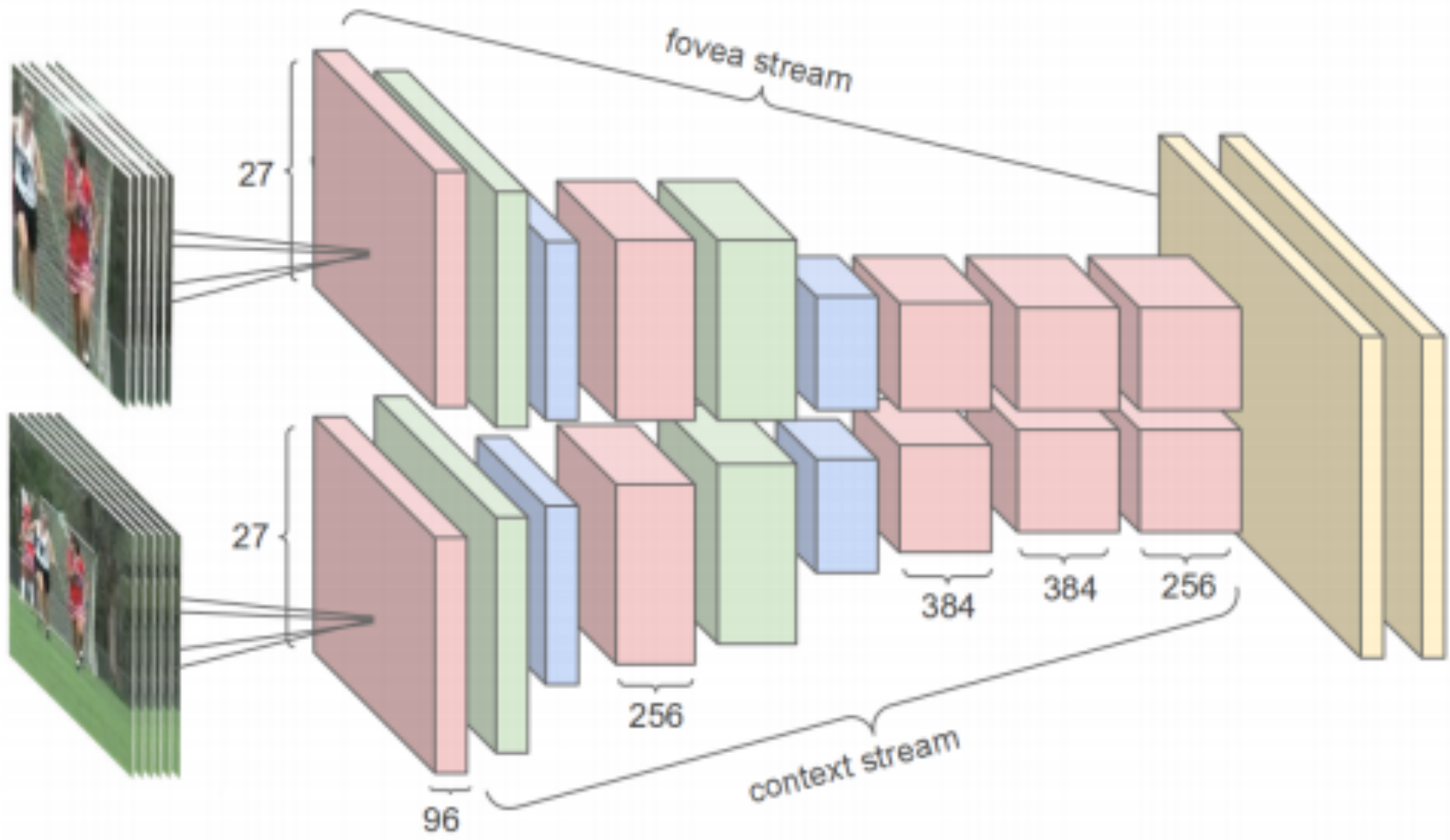
[Tompson et al, CVPR 2015]

# Deep learning for robotics: navigation, human-robot interaction, detecting grasps



[Lenz et al, IJRR 2014]

# Video classification




[Karpathy et al, CVPR 2014]

Facebook AI Research

# Action recognition from video

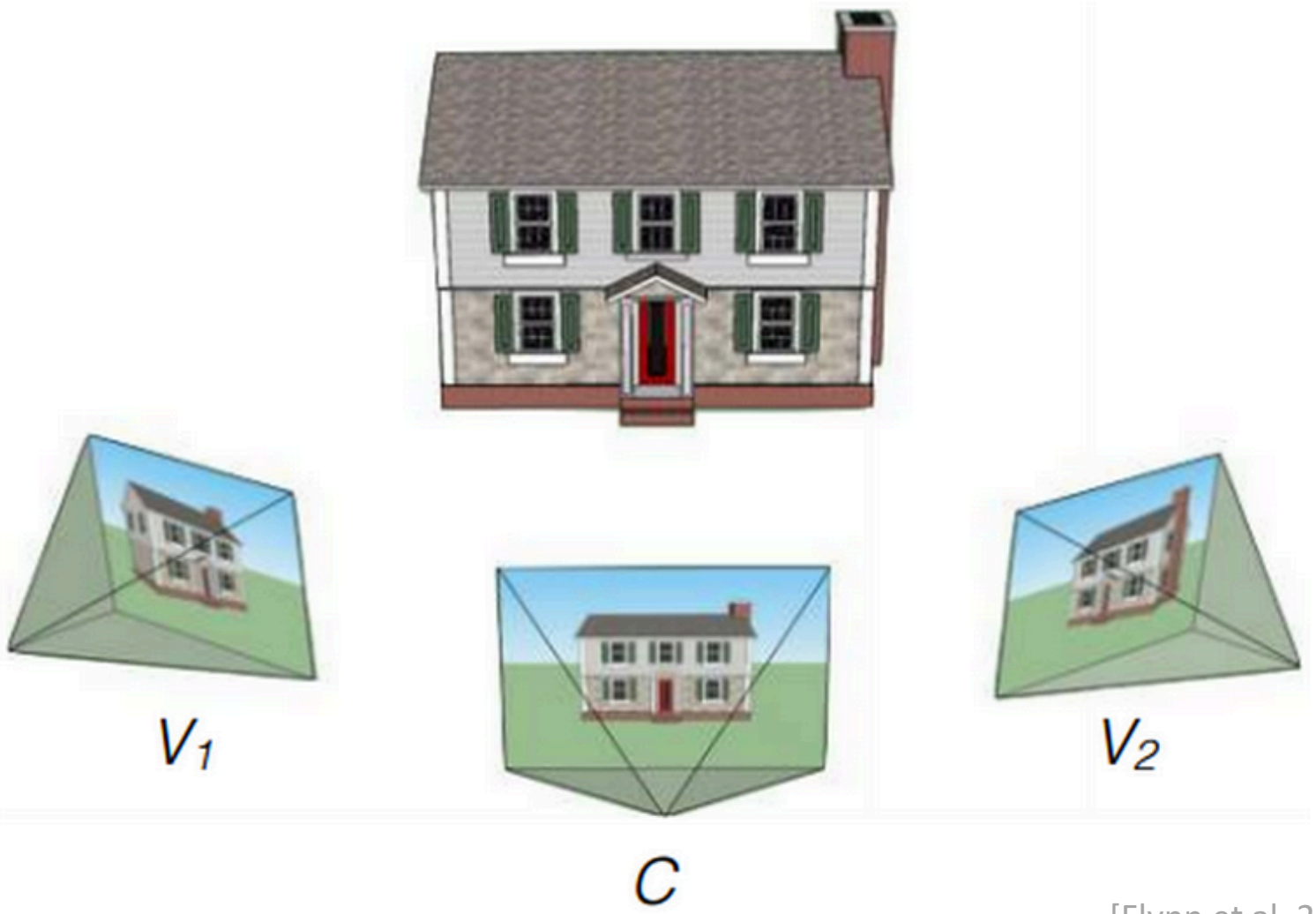
Facebook AI Research

# Natural image generation

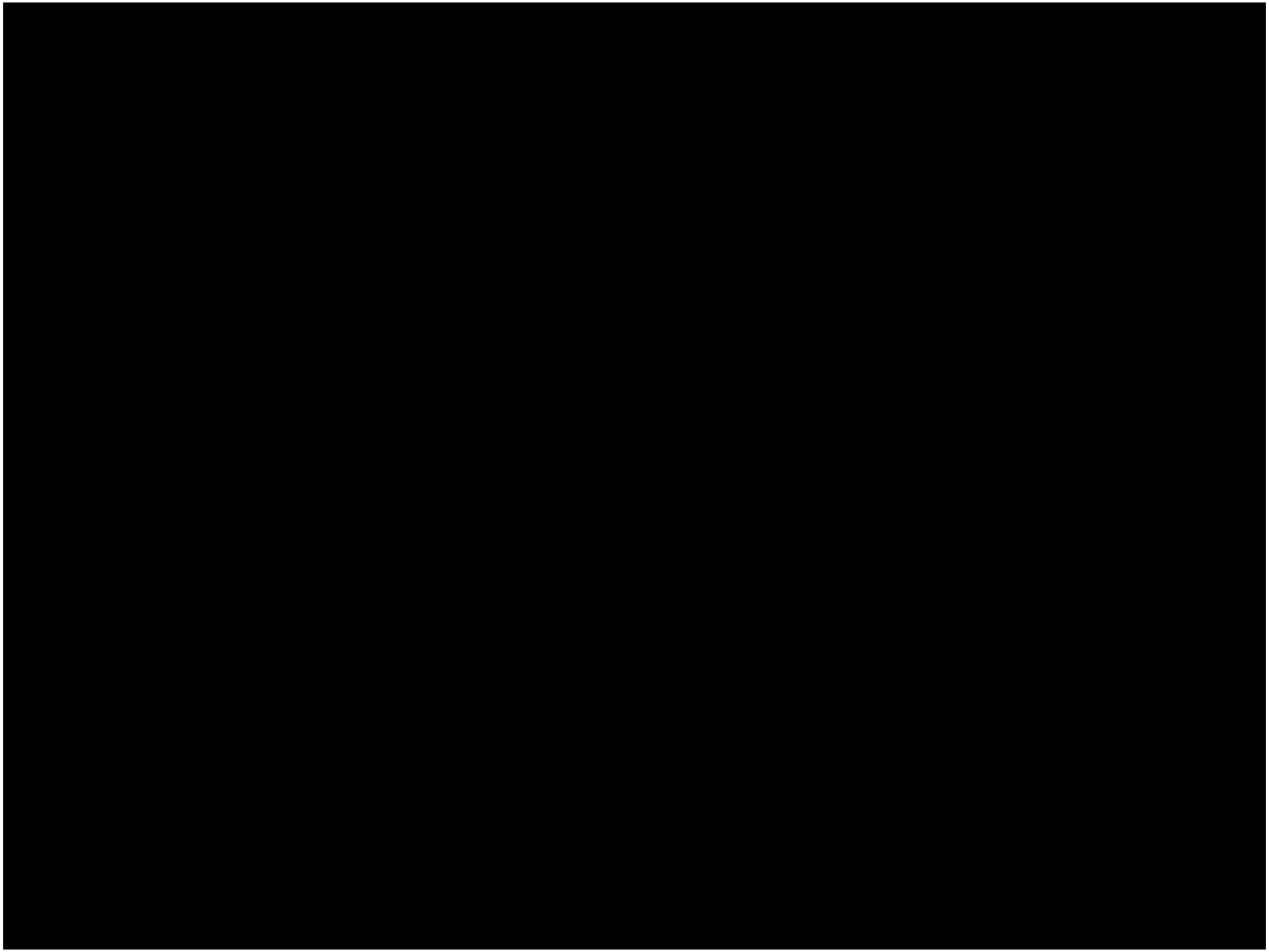
[Denton et al, 2015]

Google

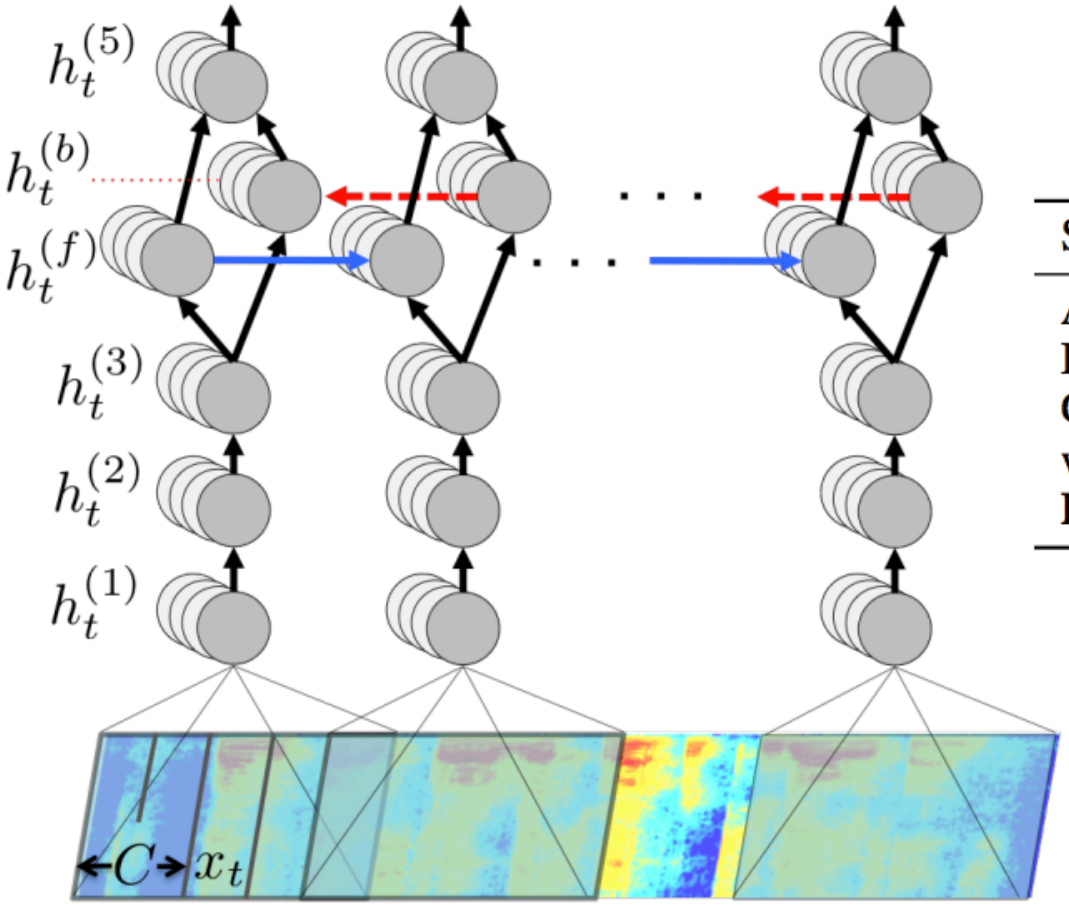
# DeepStereo: View synthesis




[Flynn et al, 2015]



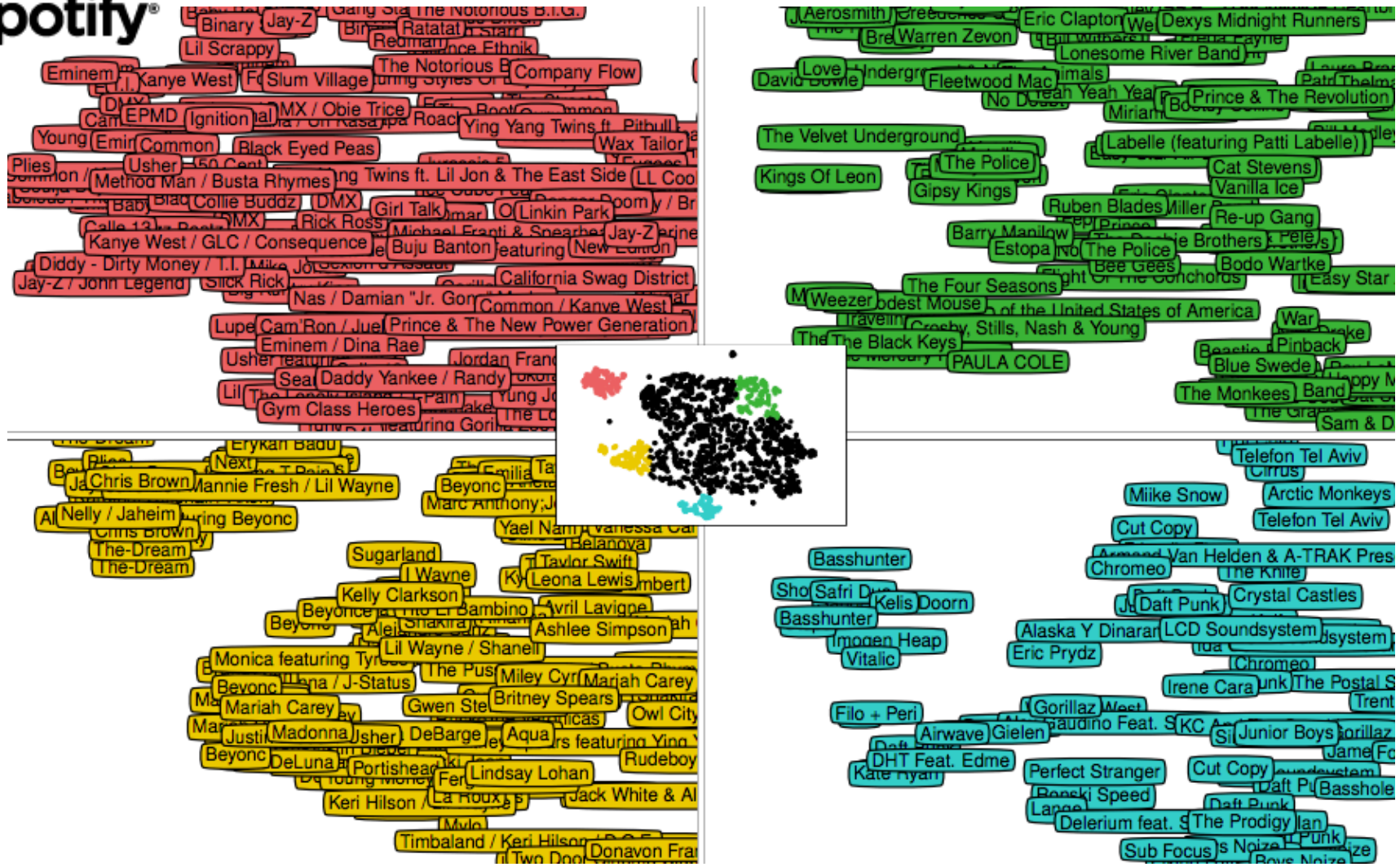
# Deep Speech




| System | Combined (176) |
|---|---|
| Apple Dictation | 26.73 |
| Bing Speech | 22.05 |
| Google API | 16.72 |
| wit.ai | 19.41 |
| **Deep Speech** | **11.85** |

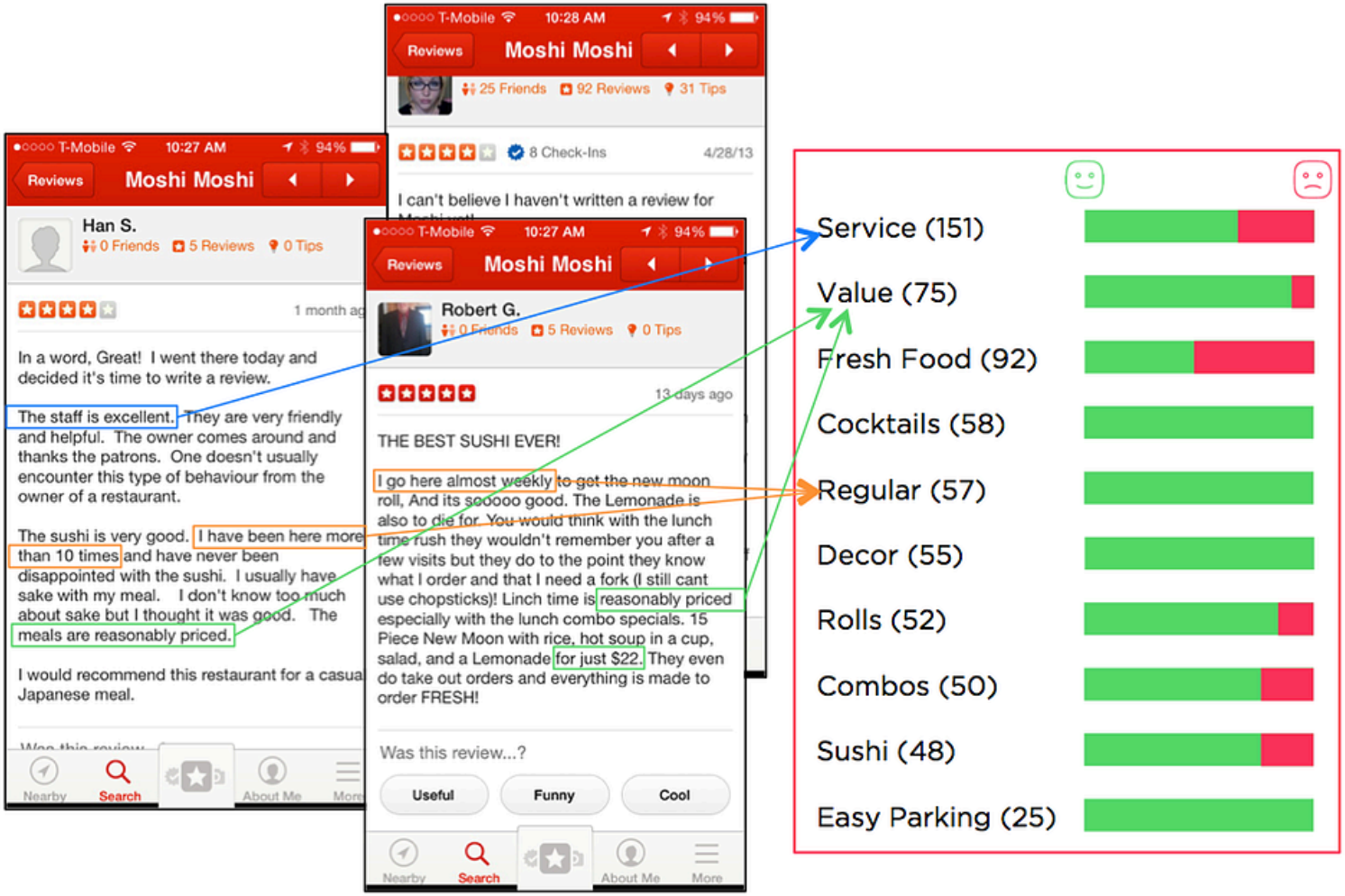
[Hannun et al, 2014]

# Musical intelligence




[Oord, Dieleman, Schrauwen, NIPS 2013]

# Text understanding



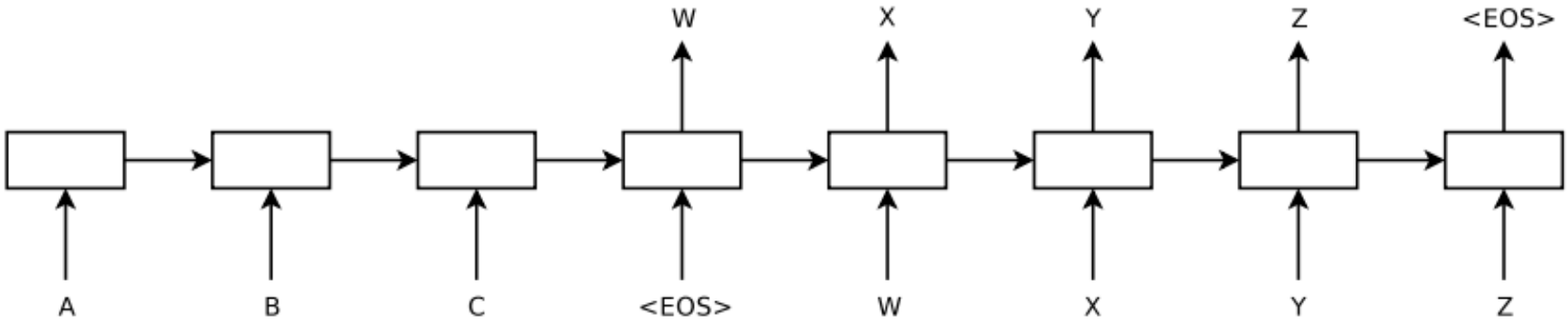
Facebook AI Research

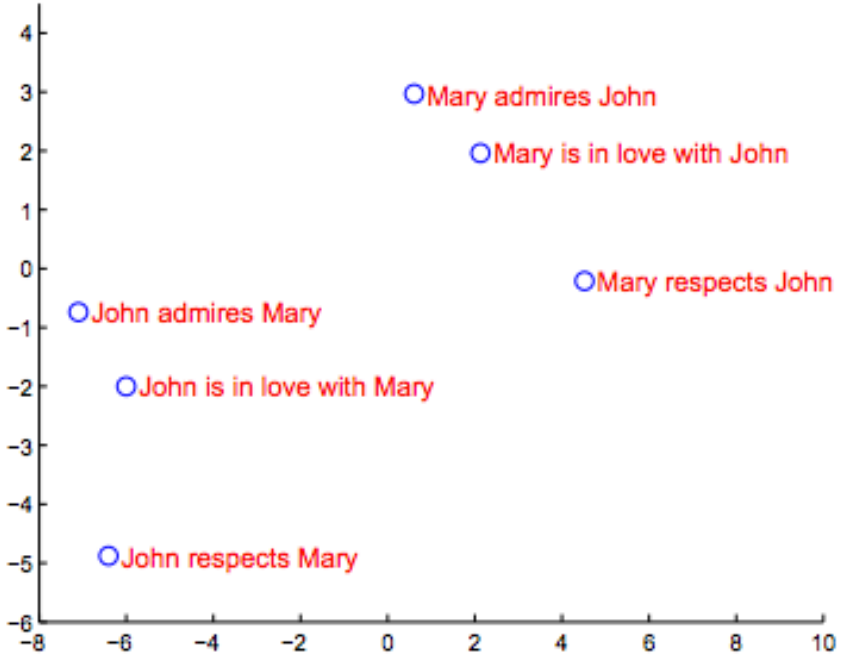
# Memory networks: question answering

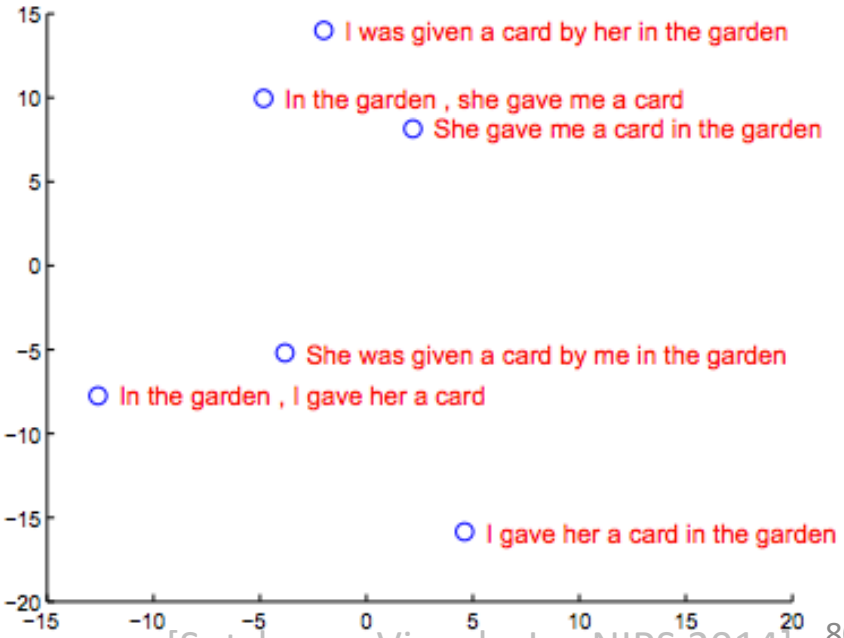
Joe went to the kitchen. Fred went to the kitchen. Joe picked up the milk.
Joe travelled to the office. Joe left the milk. Joe went to the bathroom.
Where is the milk now? A: office
Where is Joe? A: bathroom
Where was Joe before the office? A: kitchen

[Weston et al, ICLR 2015]

# Machine translation

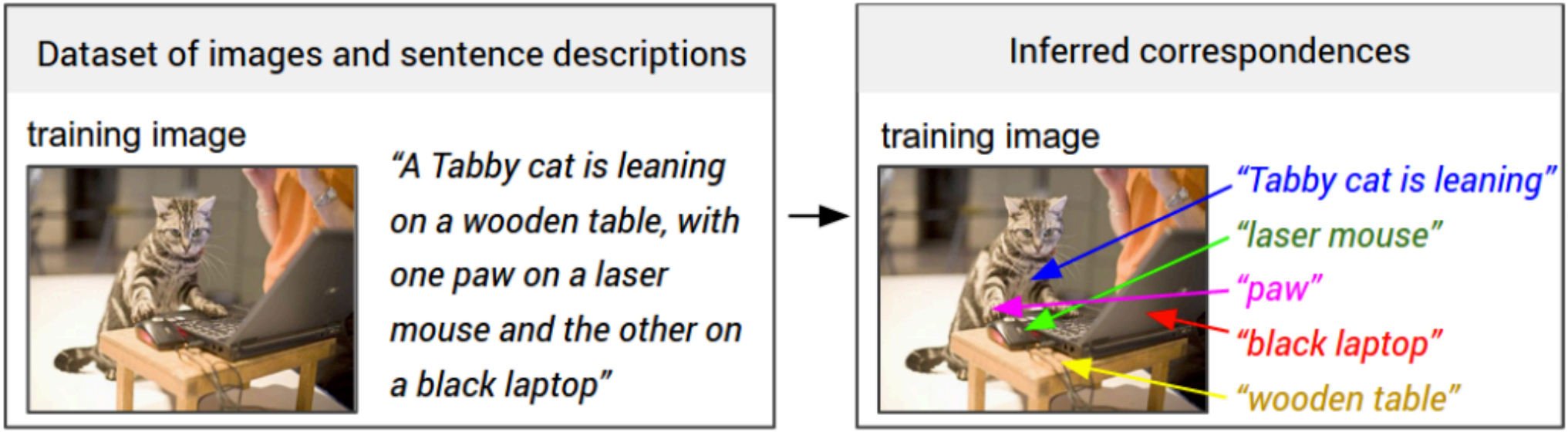


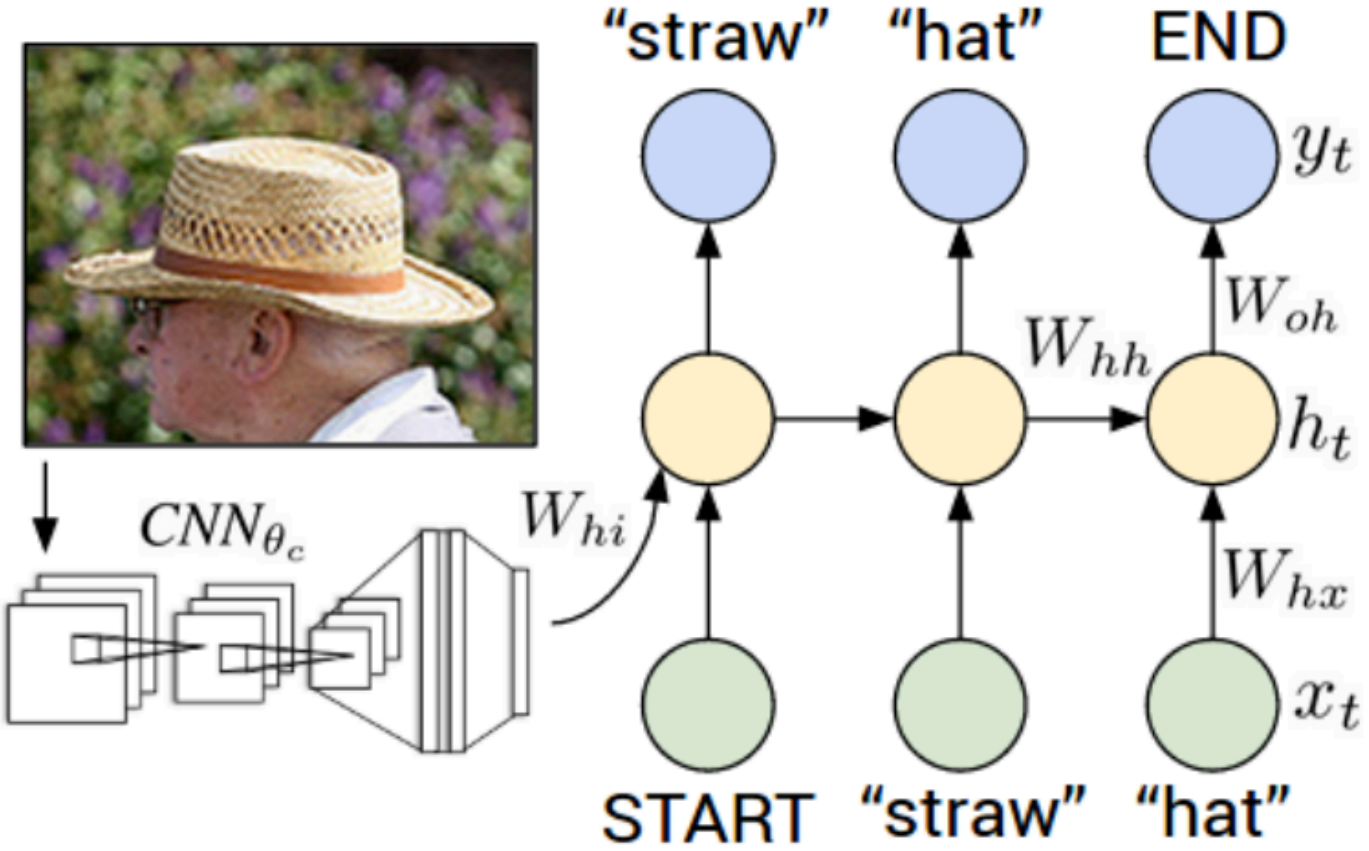

[Sutskever, Vinyals, Le, NIPS 2014]

# Multimodal deep learning: image captioning




[Karpathy et al, CVPR 2015]

# Image captioning




[Karpathy et al, CVPR 2015]

# Image captioning


"man in black shirt is playing guitar."

"girl in pink dress is jumping in air."

"black cat is sitting on top of suitcase."

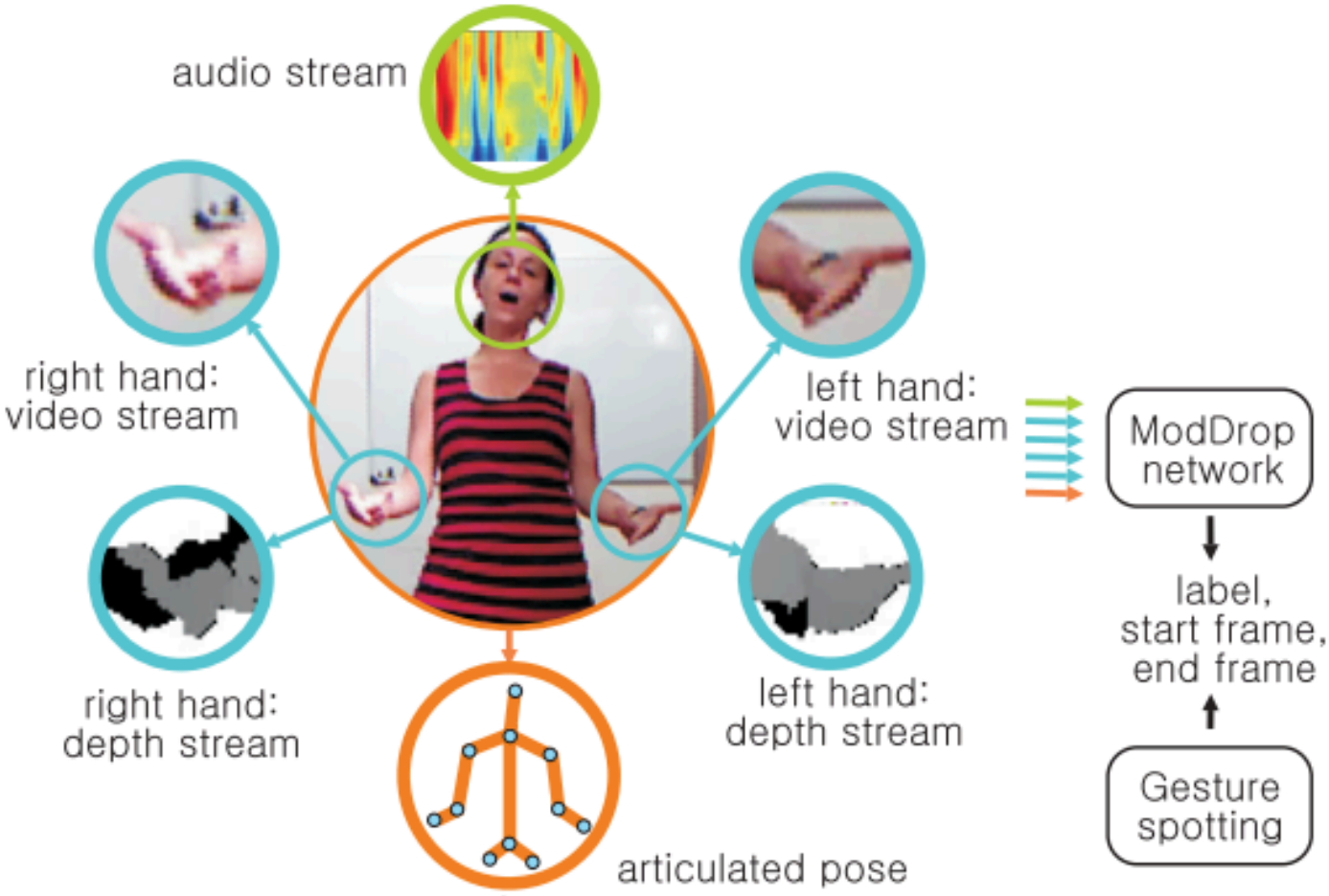
"a cat is sitting on a couch with a remote control."


"a woman holding a teddy bear in front of a mirror."

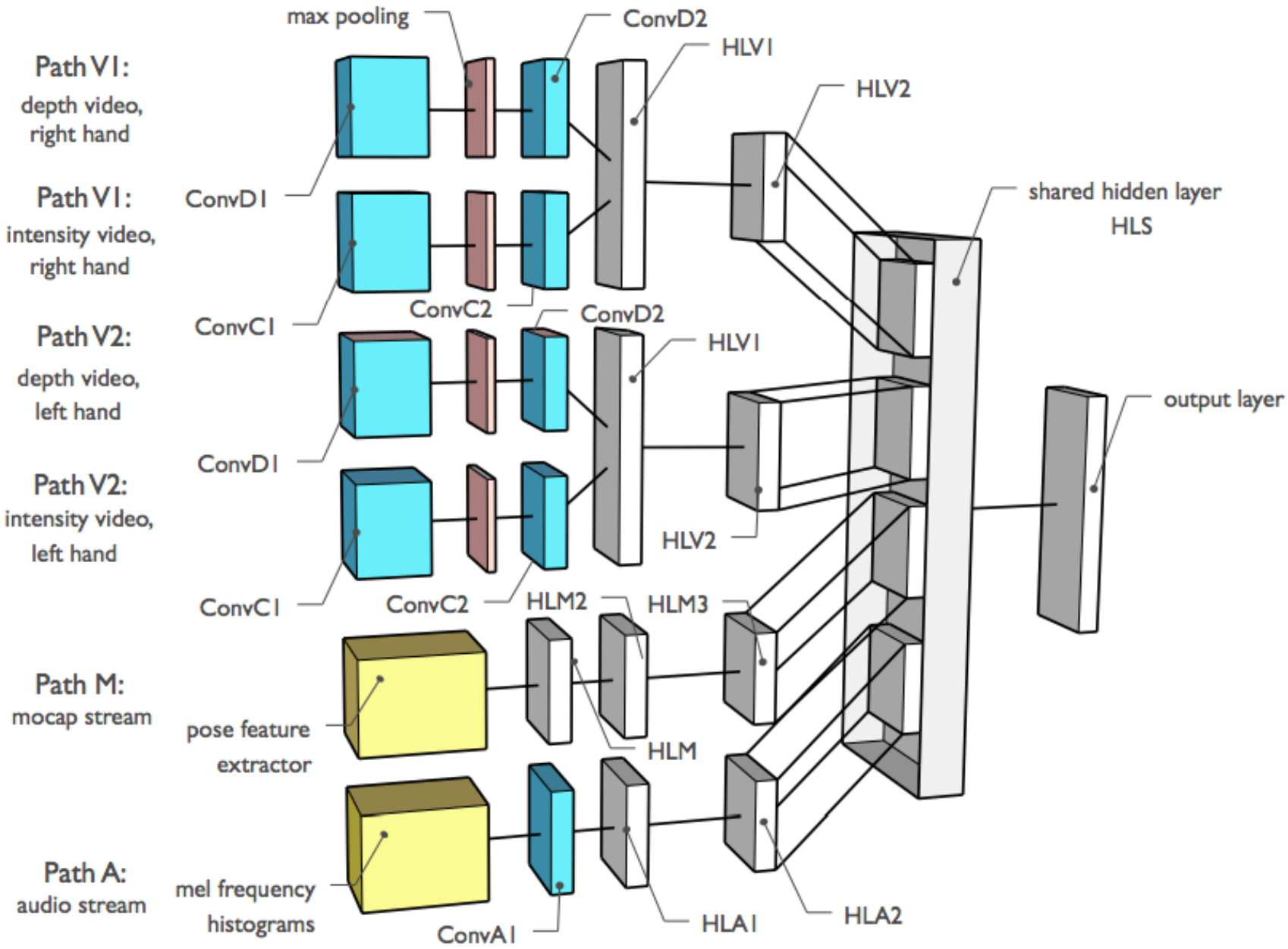
[Karpathy et al, 2015]


Joint modeling of text and images based on their co-occurrence

# Multi-modal deep learning: gesture recognition



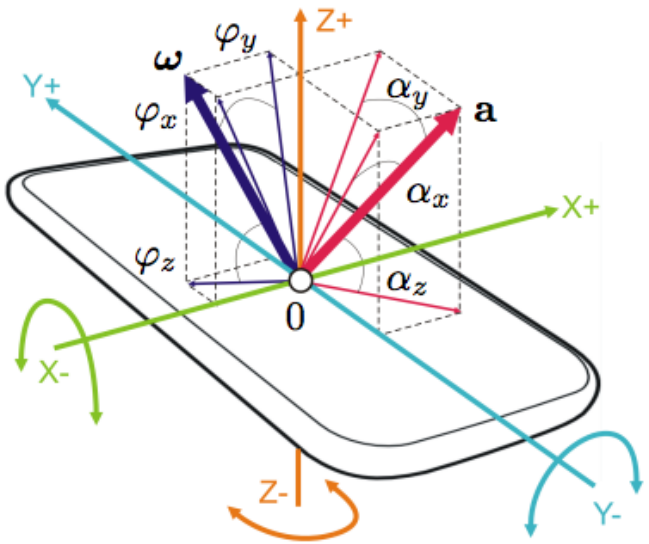

[Neverova et al, PAMI 2015]

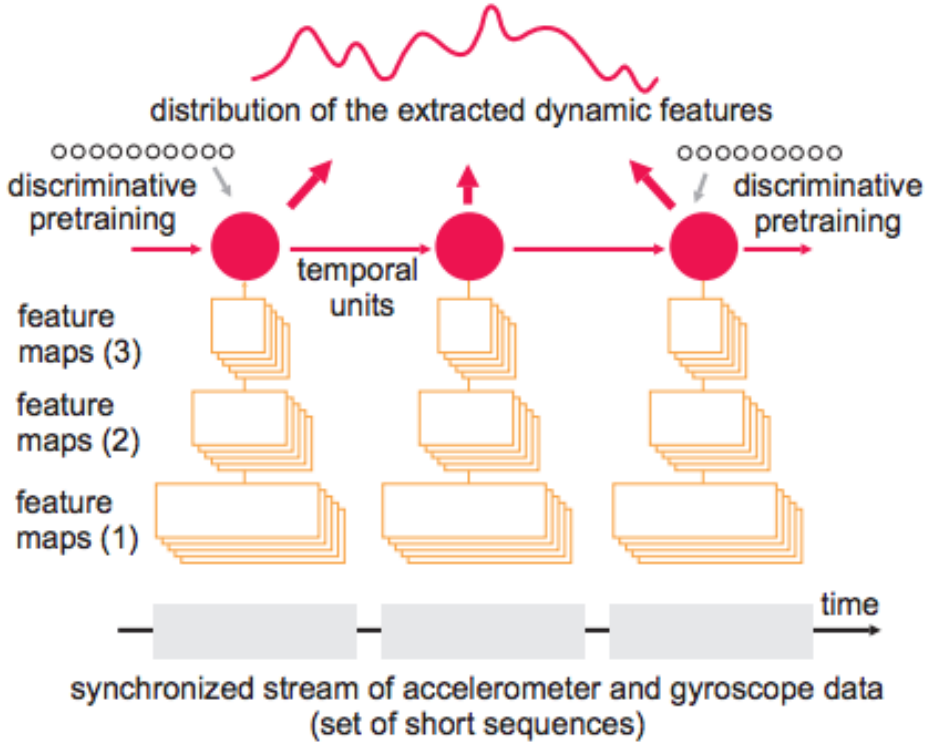



[Neverova et al, PAMI 2015]

Google

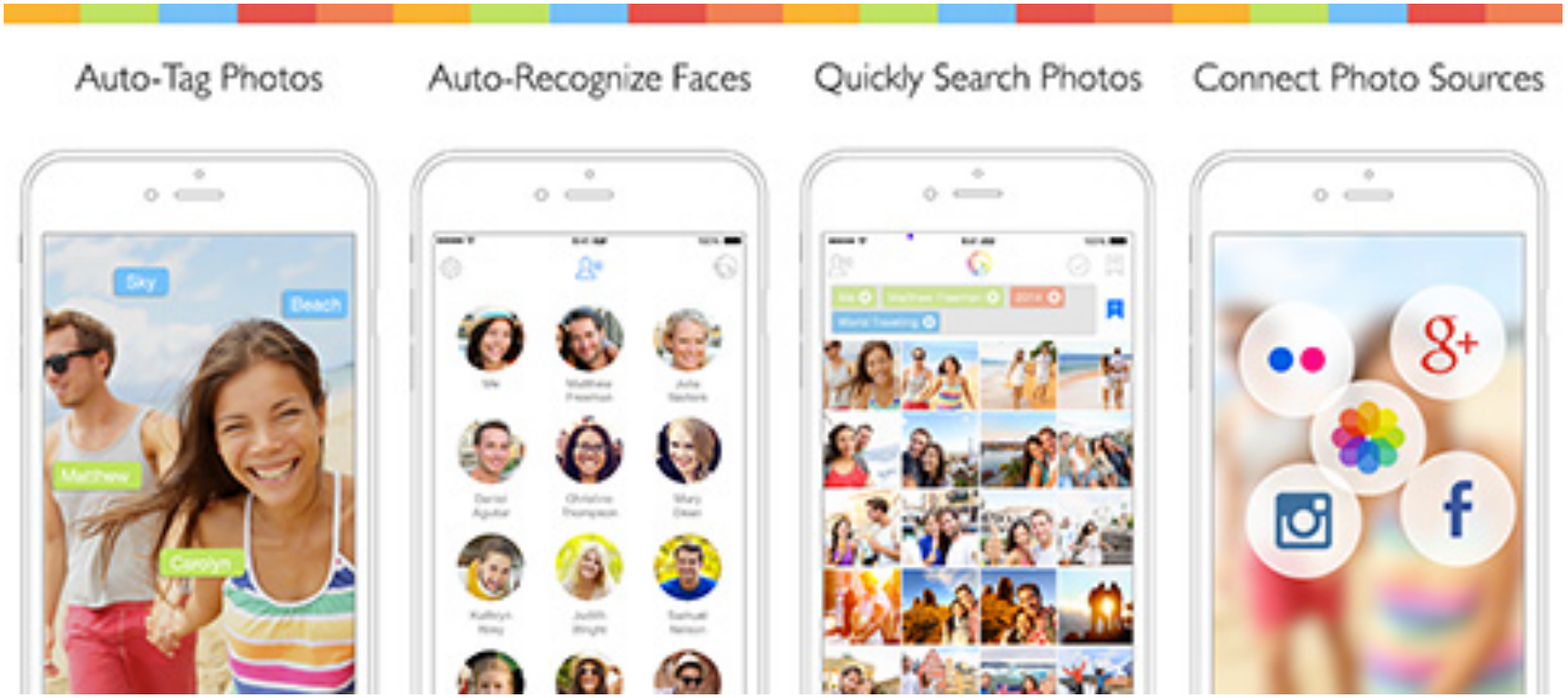
# Multi-modal deep learning: mobile authentication




[Neverova et al, forthcoming publication]


# Deep learning services

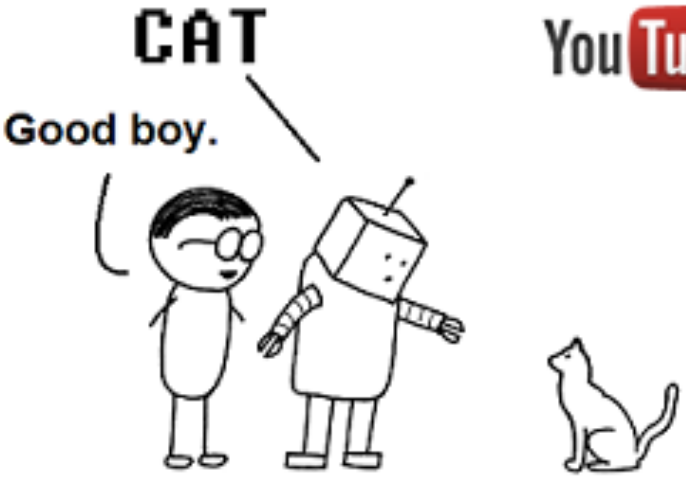


clarifai

coffee croissant beverage morning

breakfast food

# Unsupervised deep learning?

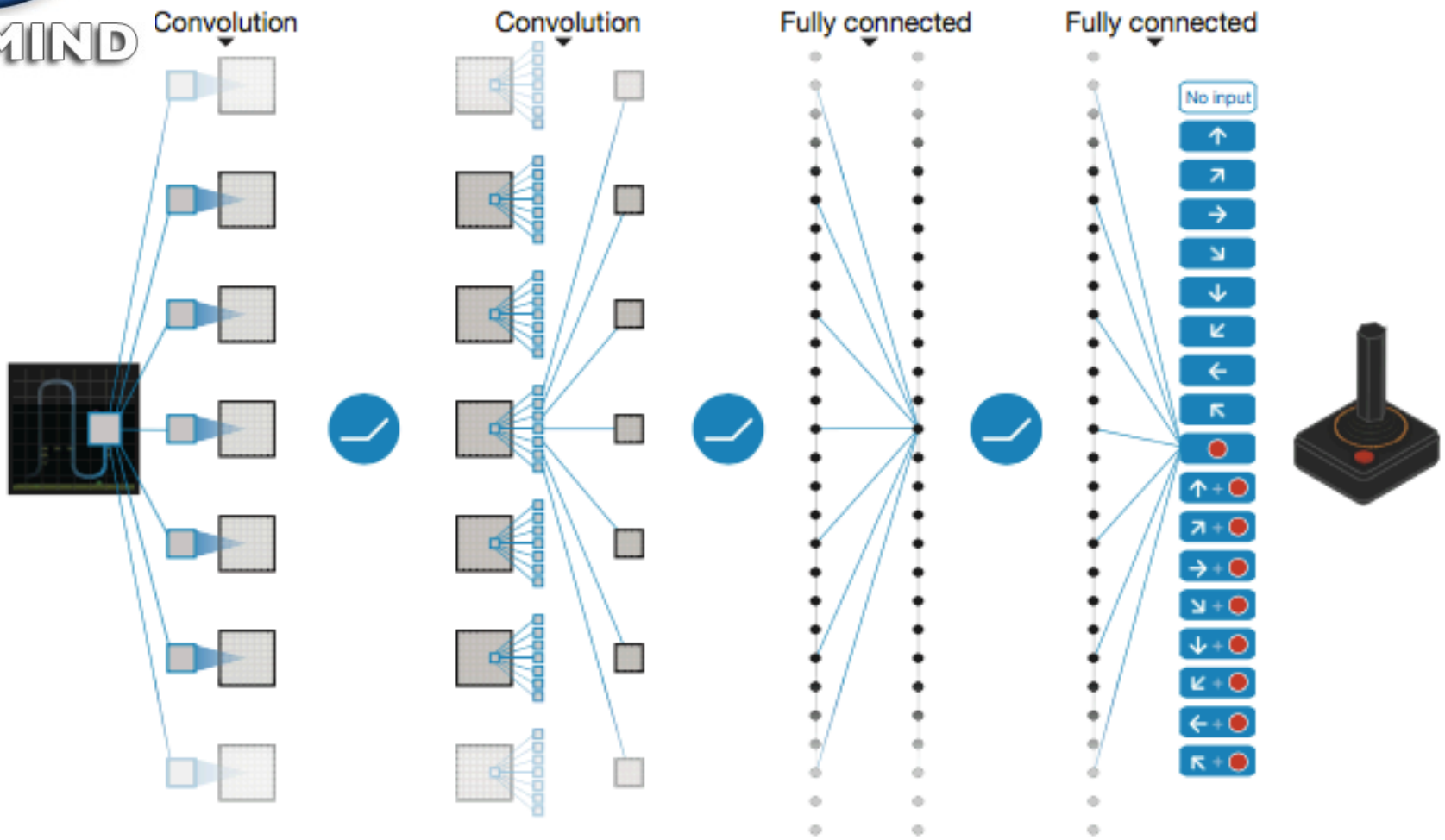






[Le et al, ICML 2012]

# Reinforcement deep learning?




[Mnih et al, Nature 2015]

# Google DeepMind's Deep Q-learning

# All roads lead to deep learning?

- Deep learning methods are general and work well with completely different kinds of data (text, sparse data).
- Depends on large data sets, which are not available in many domains.
- Hungry for computational resources.
- Requires certain expertise to train.

# Future?

- Deep learning: big data + computational resources + efficient training + model complexity
- Deep learning chips, cloud services, specific and general software libraries
- Deep supervised learning in action, unsupervised and reinforcement learning coming soon
- Applications: “understanding” text and visual context, question answering, more efficient temporal models, generating content, new data types

Natalia Neverova, natalia.neverova@liris.cnrs.fr