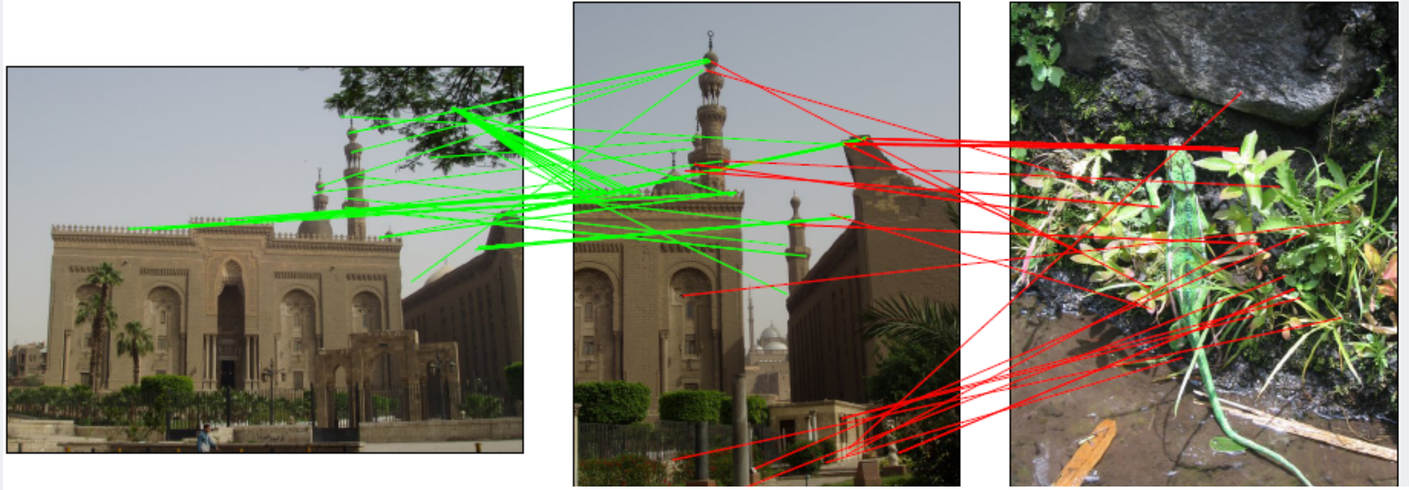
Similarity search for weakly supervised Machine Learning

Matthijs Douze

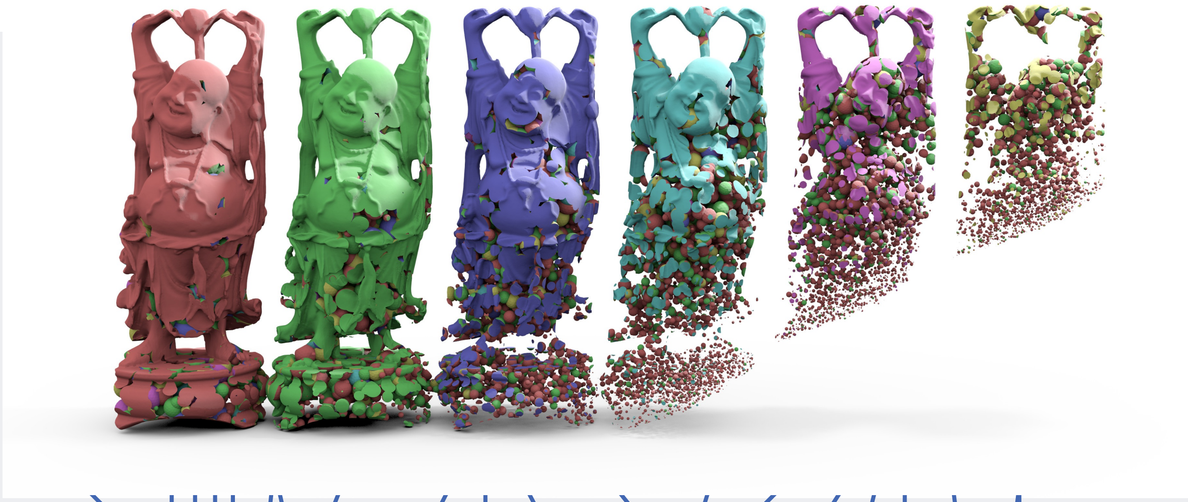
Facebook AI research, Paris

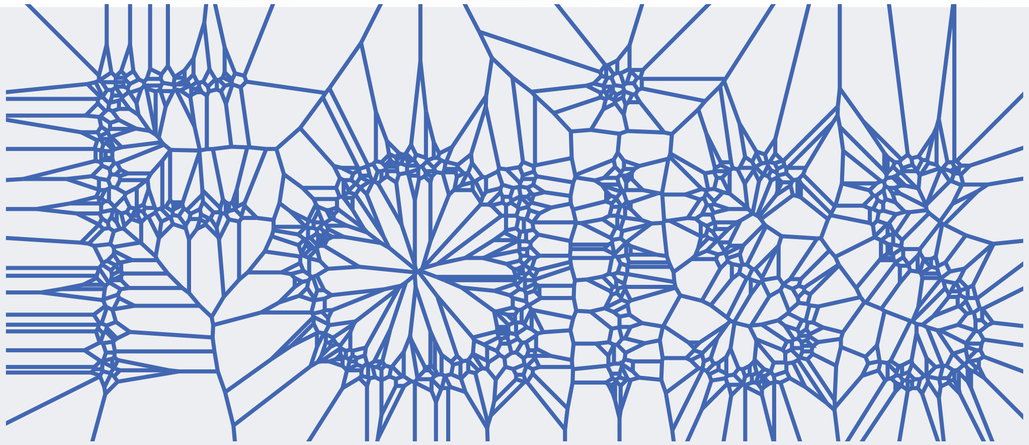
Toulouse symposium on deep learning, Oct 18, 2018

About me



- Eng + PhD ENSEEIHT, 2004
  - Computer vision
- 10 years at INRIA
  - image/video matching
  - large-scale 3D reconstruction
- Facebook since 2015
  - similarity search
  - knn-graphs (clustering, low-shot learning)
  - unsupervised learning
  - video alignment
  - neural net memorization





Facebook AI Research

facebook

10 YEAR ROADMAP

3 YRS

5 YRS

10 YRS

ECOSYSTEMS

- Facebook
- Instagram

PRODUCTS

- Workplace
- Messenger
- Search
- WhatsApp
- Groups
- Video

TECHNOLOGIES

- Connectivity
  - Telco Infra
  - Express Wi-Fi
  - Satellites
  - Terrestrial Solutions
  - Lasers
  - Free Basics
  - Drones
- AI
  - Vision
  - Language
  - Reasoning
  - Planning
- VR / AR
  - BCI
  - Mobile VR
  - Rift
  - Touch
  - Social VR
  - Standalone
  - AR Tech

Facebook AI Research

•Created in 2013 by Yann Le Cun
•Fundamental, open and collaborative research
•**160+ people:** 50/40/10 scientists/engineers/ students

Locations:

- Menlo Park (HQ of Facebook) [2014]
- NYC [2014]
- Paris [2015] + London [2018]
- Montreal [2017]
- Pittsburgh [2018]
- Seattle [2018]

Values:

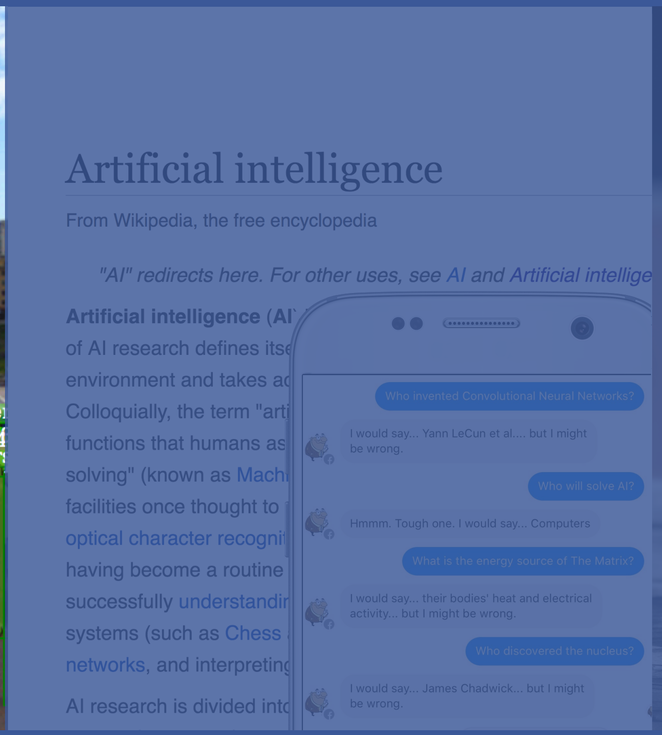
- Openness: publish and open-source
- Freedom: researchers have complete control on their agenda
- Collaboration: with internal and external partners
- Excellence: focus on most impactful projects (~~publish or perish~~)
- Scale: operate at large scale (computation, data)

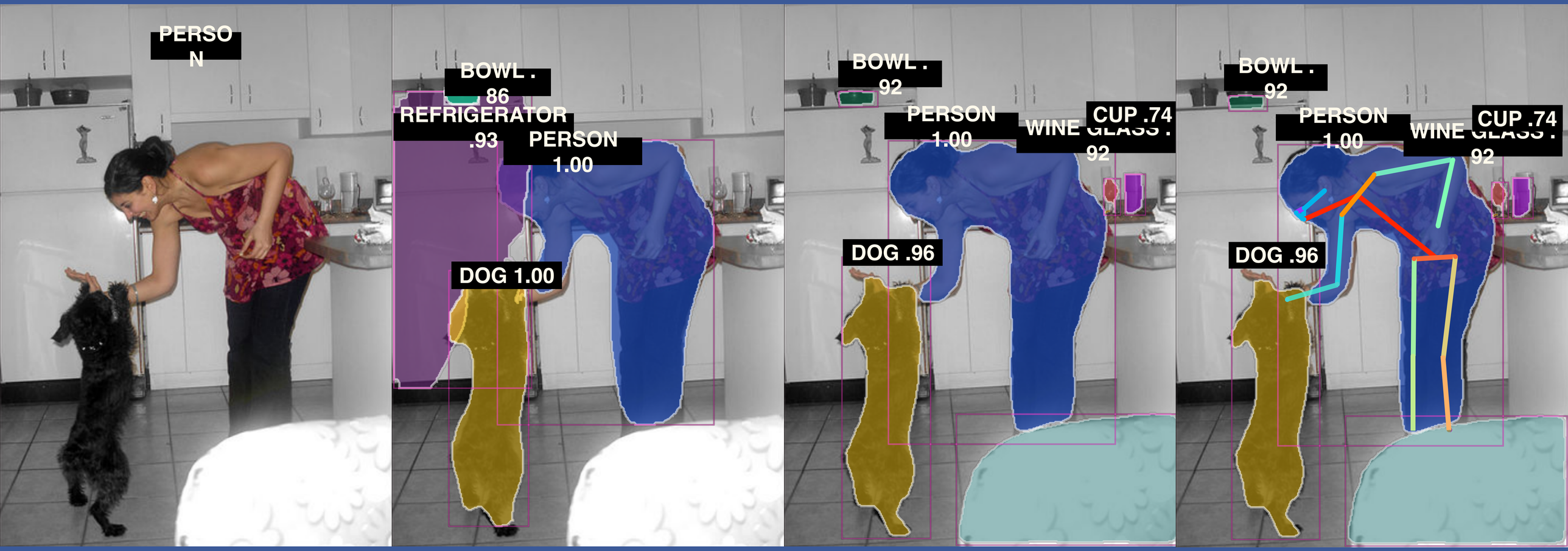
PERCEPTION

UNDERSTANDING & LEARNING

PREDICTION

PLANNING



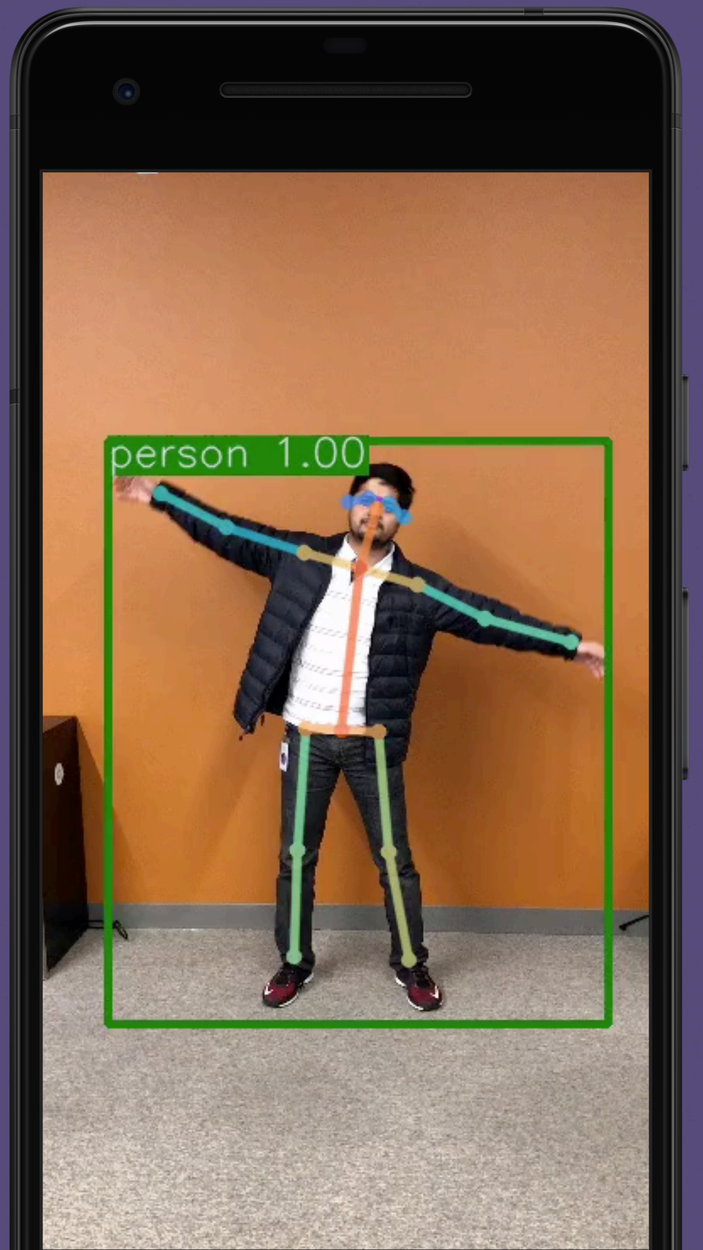


Mask-RCNN
He, Gkioxari, Dollar & Girshick [2018]

Live Pose Detection

BEFORE | 2017

AFTER | 2018



PERCEPTION

UNDERSTANDING & LEARNING

PREDICTION

PLANNING

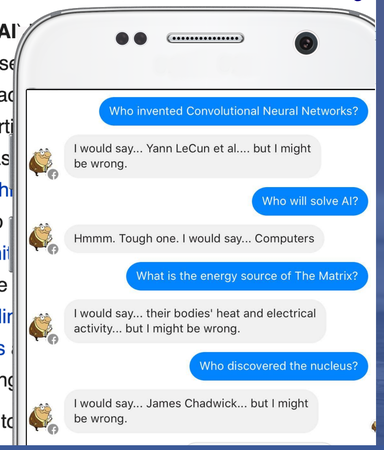
Artificial intelligence

From Wikipedia, the free encyclopedia

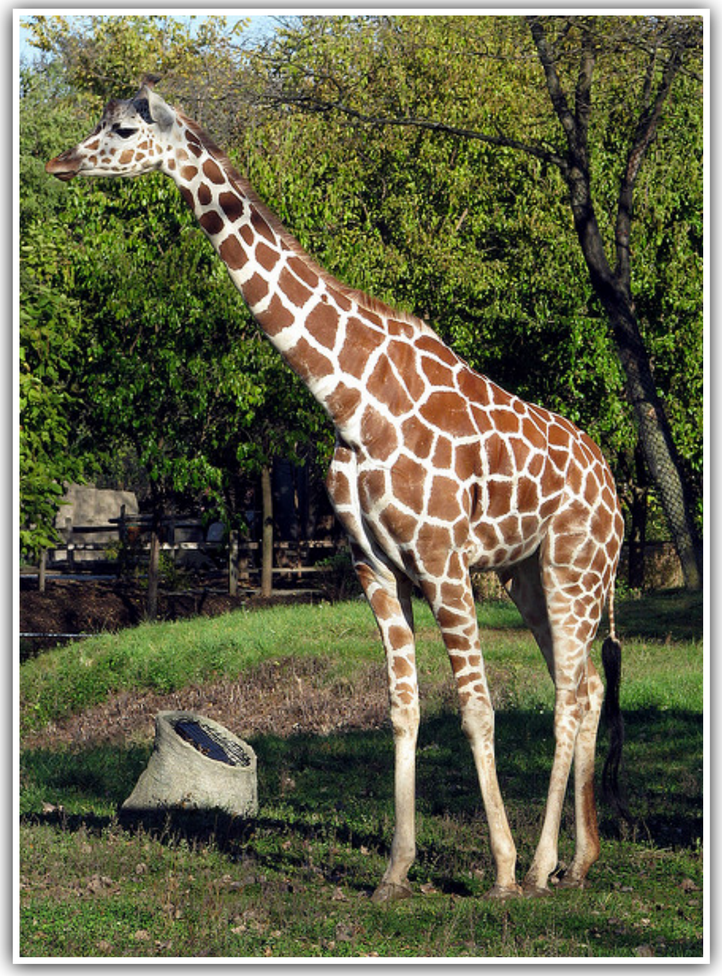
"AI" redirects here. For other uses, see AI and Artificial intellige

Artificial intelligence (AI
of AI research defines itse
environment and takes ac
Colloquially, the term "arti
functions that humans as
solving" (known as Machi
facilities once thought to
optical character recogni
having become a routine
successfully understandi
systems (such as Chess
networks, and interpreting

AI research is divided into



A man riding a wave on a surfboard in the water.



A giraffe standing on the grass next to a tree.

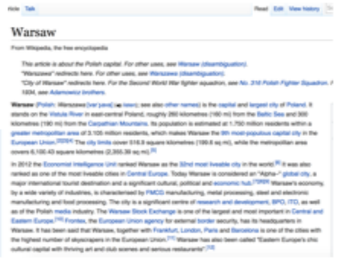
An airplane is parked on the tarmac at an airport.

A man riding a motorcycle on a beach.

READING WIKIPEDIA

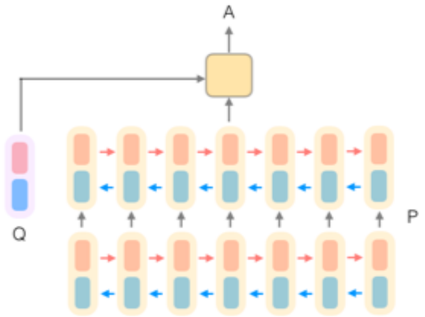
Dr. QA

Q: How many of Warsaw's inhabitants spoke Polish in 1933?



Document Reader

833,500



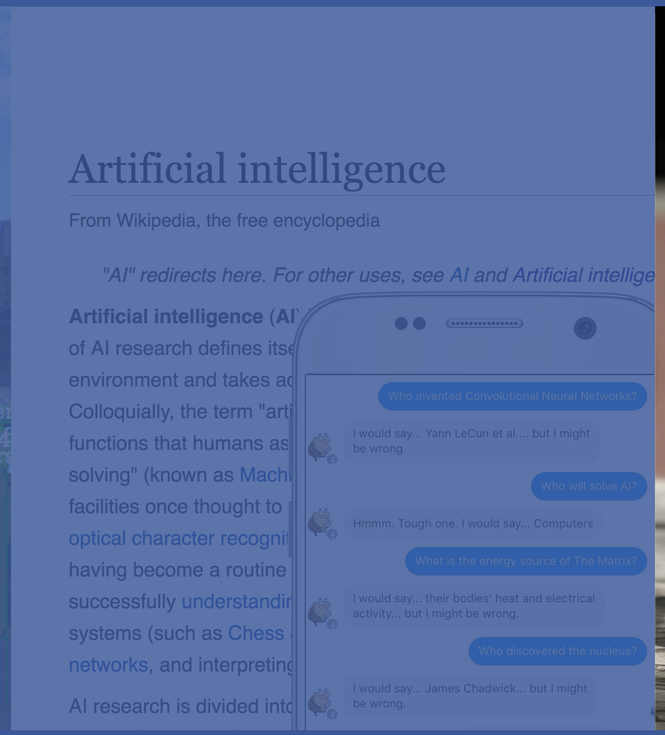
Reading Wikipedia to answer open-domain questions
Chen, Fisch, Weston & Bordes [2017]

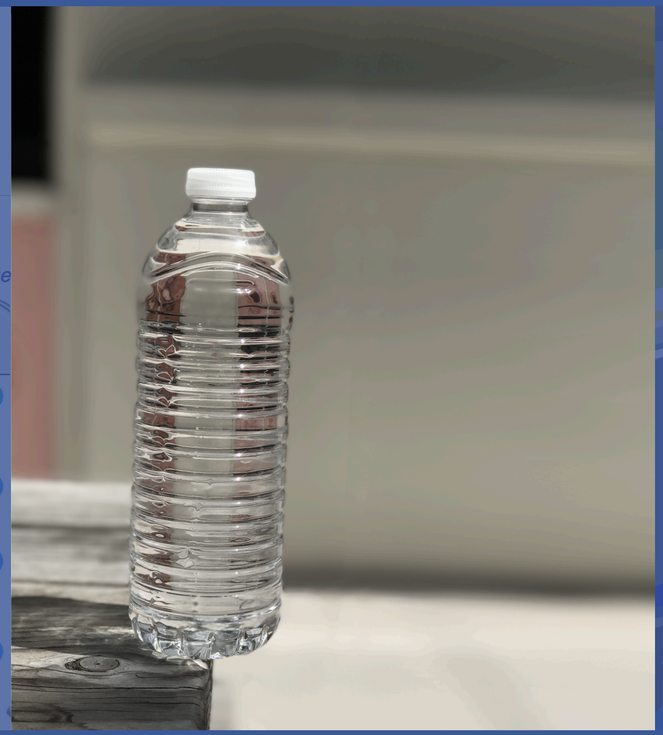
PERCEPTION

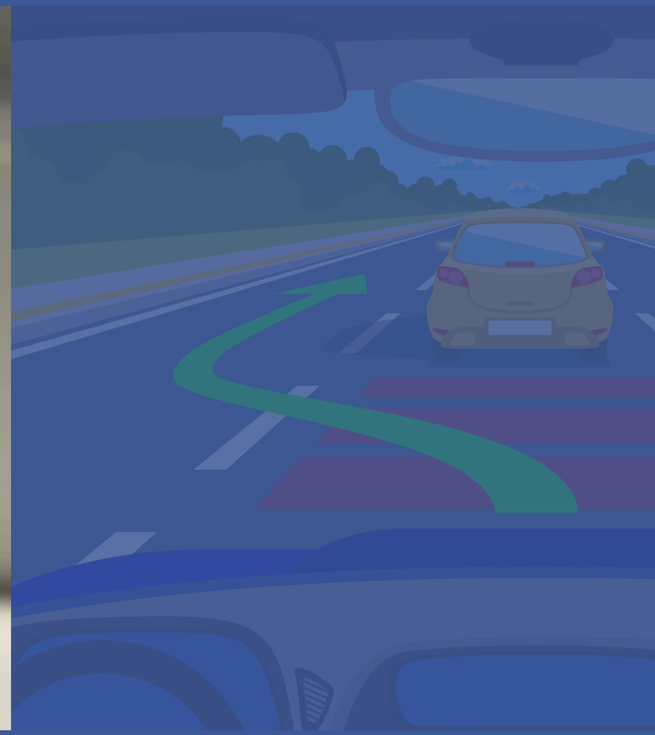
UNDERSTANDING & LEARNING

PREDICTION

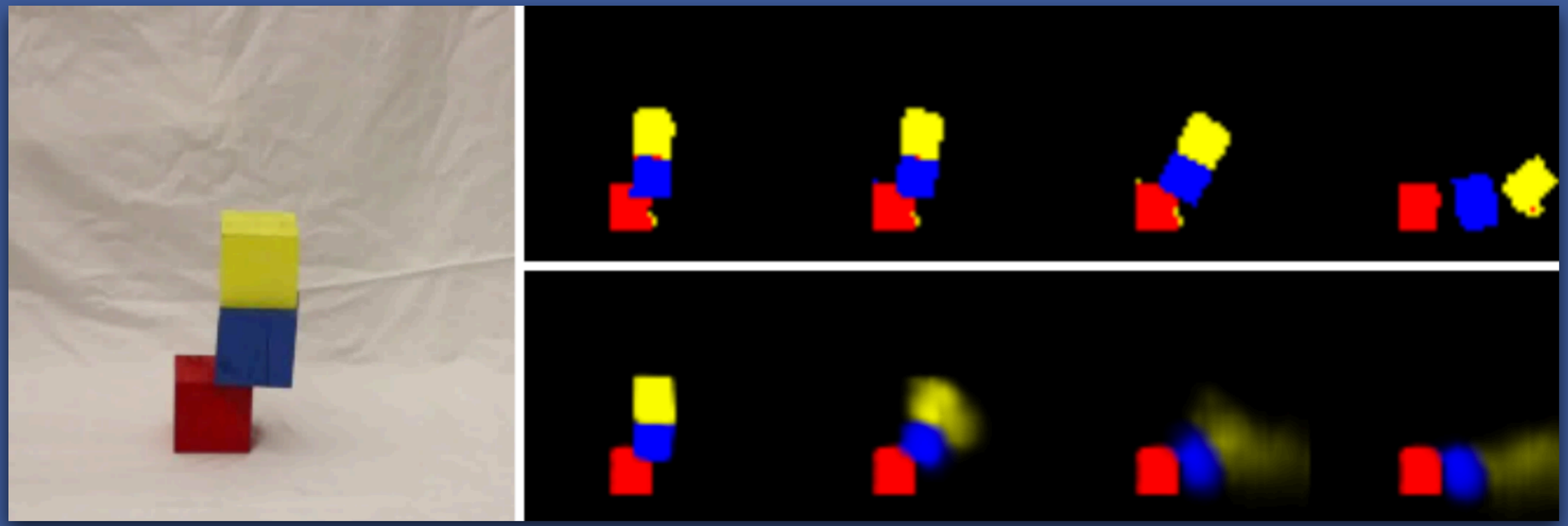
PLANNING







GETTING AN INTUITION OF PHYSICS



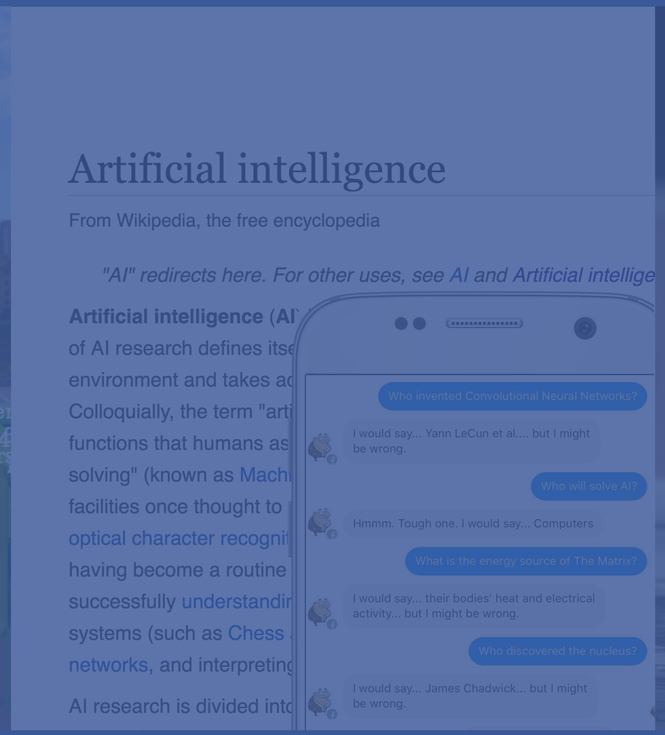
PREDICTING FUTURE FRAMES

PREDICTING FUTURE FRAMES

DesIGN: Design Inspiration from Generative Networks.
Sbai, Couprie, Elhoseiny, Bordes & LeCun [2018]

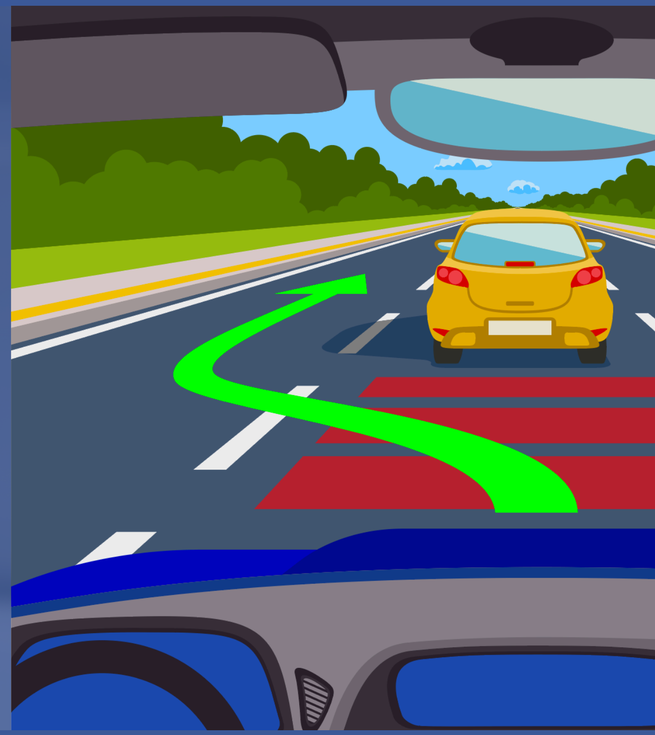
PERCEPTION

UNDERSTANDING & LEARNING



PREDICTION

PLANNING



BEFORE TRAINING

AFTER TRAINING

Common Sense

Humanoid robot

UNSUPERVISED LEARNING

Efficient Learning

Long-term planning

"Good enough" translation

"Autonomous" driving

Narrow Chatbots

Visual-impaired assistance

Go (game)

SUPERVISED LEARNING

Speech Recognition

Handwriting Recognition

Picture segmentation

Chess

Common Sense

Humanoid robot

UNSUPERVISED LEARNING

Efficient Learning

Long-term planning

"Good enough" translation

"Autonomous" driving

Narrow Chatbots

Visual-impaired assistance

Go (game)

SUPERVISED LEARNING

Speech Recognition

Handwriting Recognition

Picture segmentation

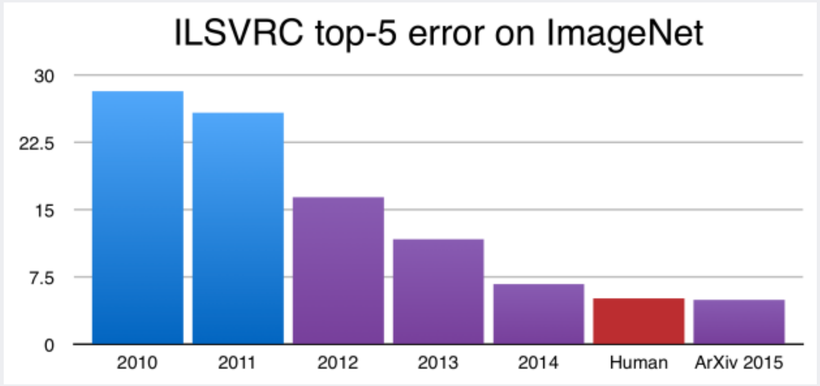
Chess

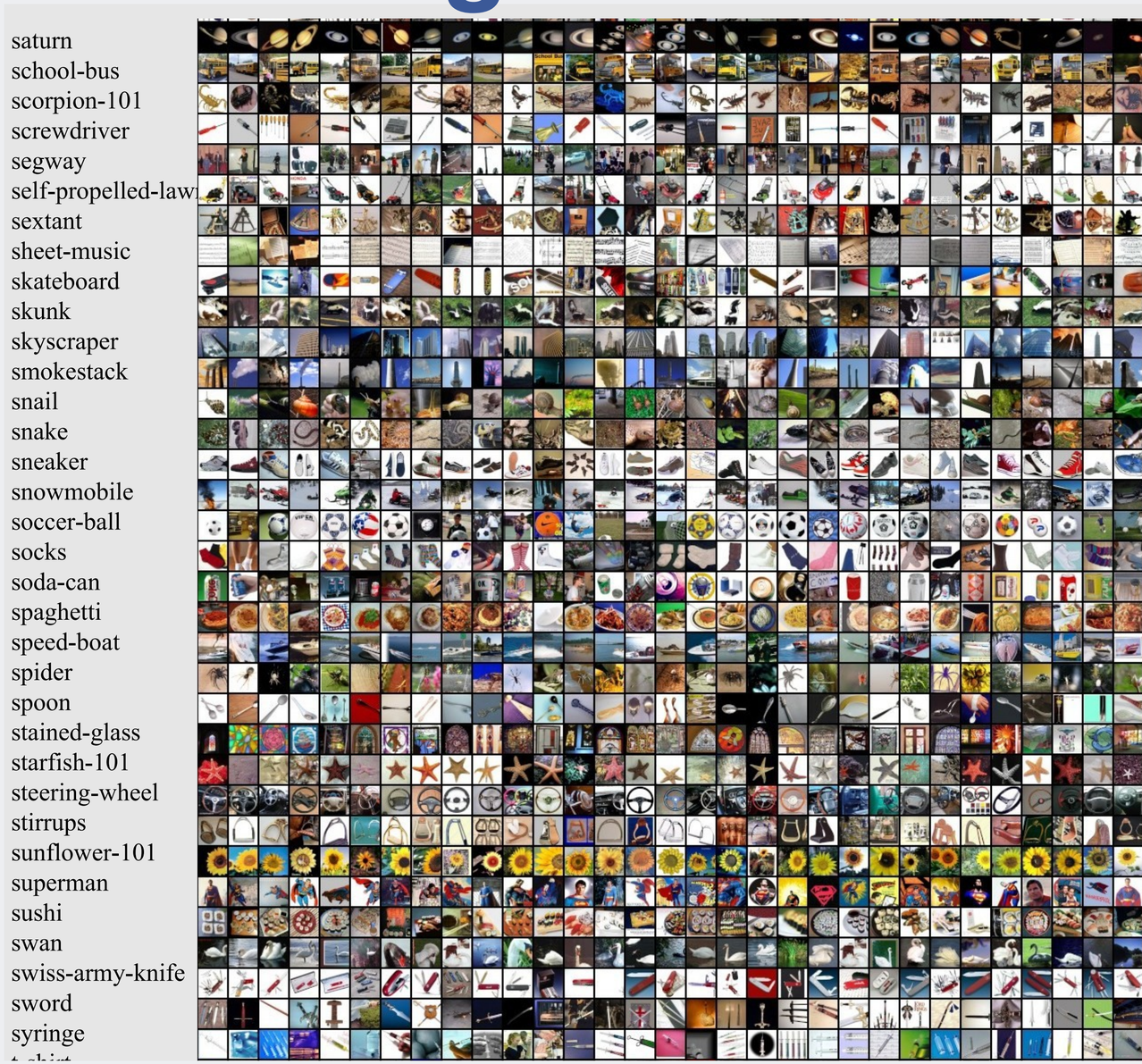
Supervised/unsupervised learning

[ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases, Pierre Stock, Moustapha Cisse, ArXiv 2017]

Supervised learning: the Imagenet case

- training set: a large and balanced set of positive examples, labelled unambiguously
- excellent testbed for the improvement of image classification algorithms
- leap forward in 2010-2015
  - see this morning's presentation
- now: more or less solved





Supervised learning

Behind the scenes

- Getting many images is not a problem

- Manual annotation: an endeavor
  - millions of images
  - several opinions
  - quality control...

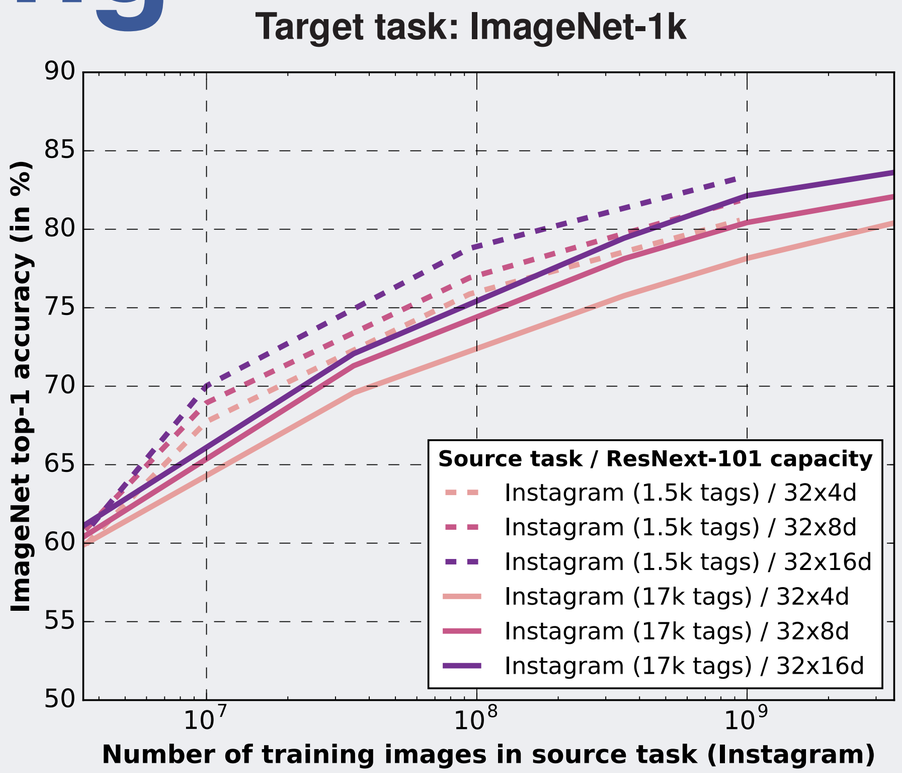
- AI powered by... Human intelligence

- How can we reduce the level of supervision?

Less supervised learning

[Exploring the Limits of Weakly Supervised Pretraining, Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin

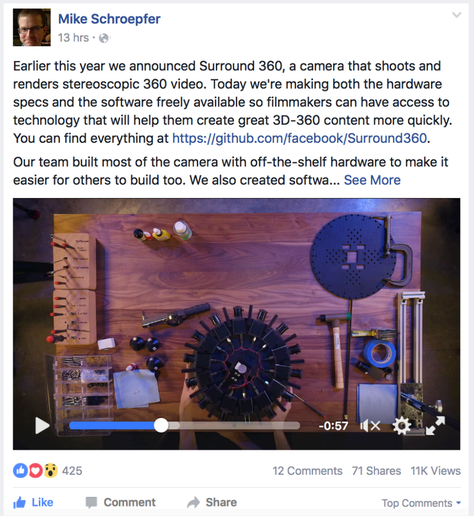
- Noisy supervision: learn from hashtags -- scale up to 3B images
- Weak supervision: object detector with only image-level supervision
- Low-shot learning: train from few images per class
- Zero-shot learning: learn from metadata, eg. a description rather than examples



[Low-Shot Visual Recognition by Shrinking and Hallucinating Features. B Hariharan, RB Girshick - ICCV,

[An embarrassingly simple approach to zero-shot learning, Bernardino Romera-Paredes, Philip H. S. Torr, ICML'15]

Embeddings



Post embedding

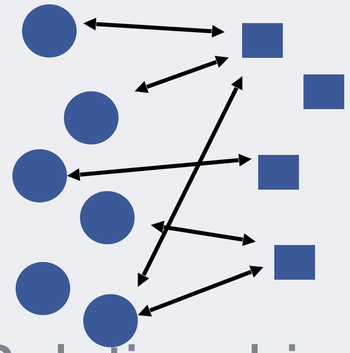
Text embedding (word2vec, **fastText**)

Face embedding

Video embedding

$$x \in \mathbb{R}^d$$

typical: d=100-1000 (dense)



Relationship embedding

User embedding

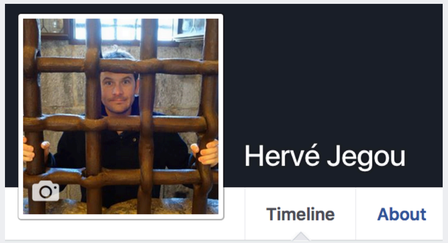
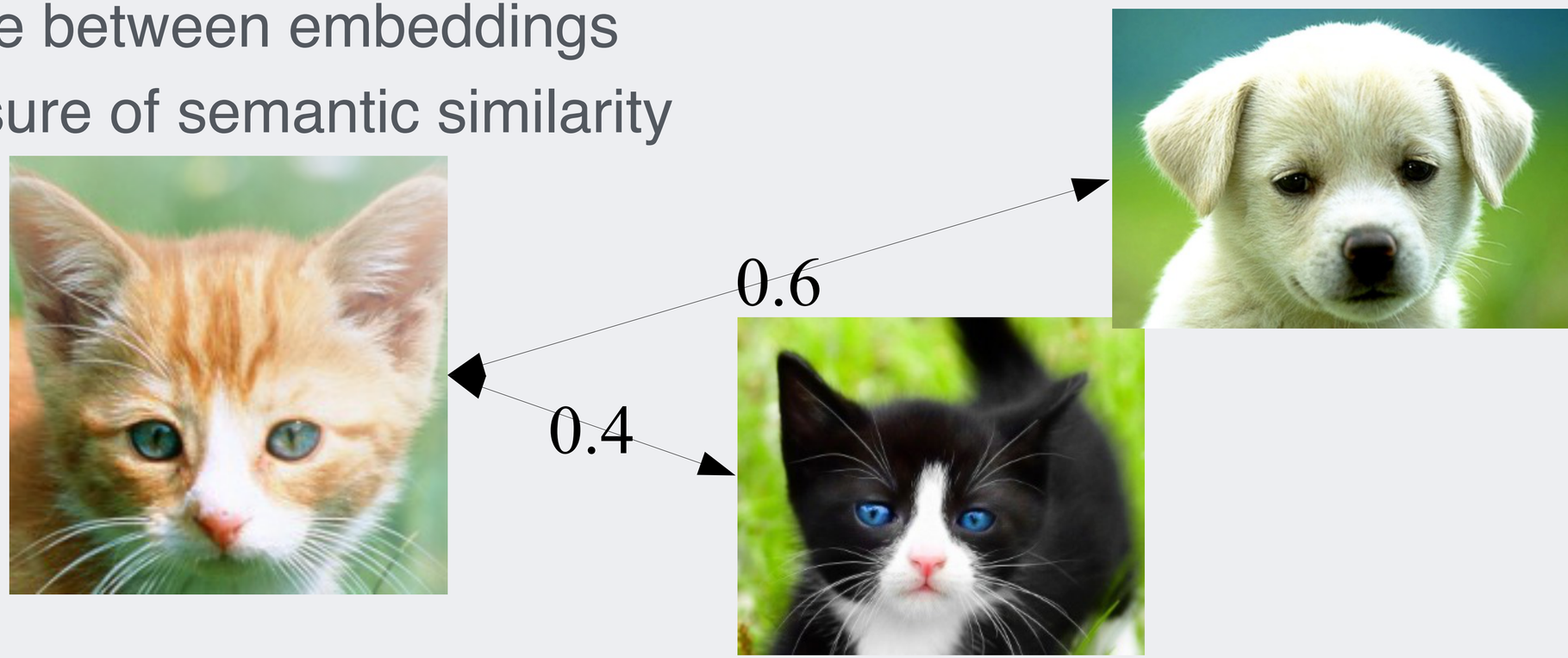


Image Embedding (CNN layer)

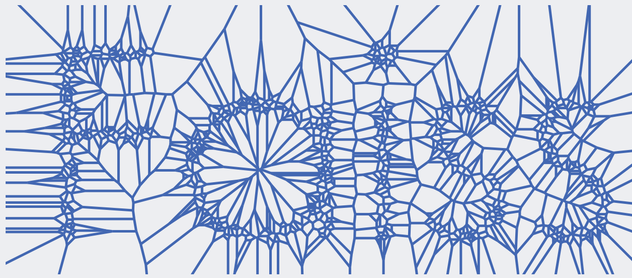
Comparing embeddings

- Using the distance between embeddings
- should be a measure of semantic similarity



- the k-nearest neighbor classifier
- in the following: 2 works on reducing supervision using similarity search

Indexing embeddings



Build index for a collection:

$y_1, y_2, \ldots, y_n \in \mathbb{R}^d$

$\begin{bmatrix}\vdots\\\vdots\\\vdots\end{bmatrix}\ \begin{bmatrix}\vdots\\\vdots\\\vdots\end{bmatrix}\quad \begin{bmatrix}\vdots\\\vdots\\\vdots\end{bmatrix}$

Indexing

Media Description (e.g, CNN)

Index in RAM

Query:

$x \in \mathbb{R}^d$

$\begin{bmatrix}\vdots\\\vdots\\\vdots\end{bmatrix}$

Result: $k - \text{argmin}_{i=1..n} \|x - y_i\|^2$

Criteria: compact, fast, accurate

Low-shot learning

[Low-shot learning with large-scale diffusion,
Douze, Szlam, Hariharan, Jégou, CVPR'18]

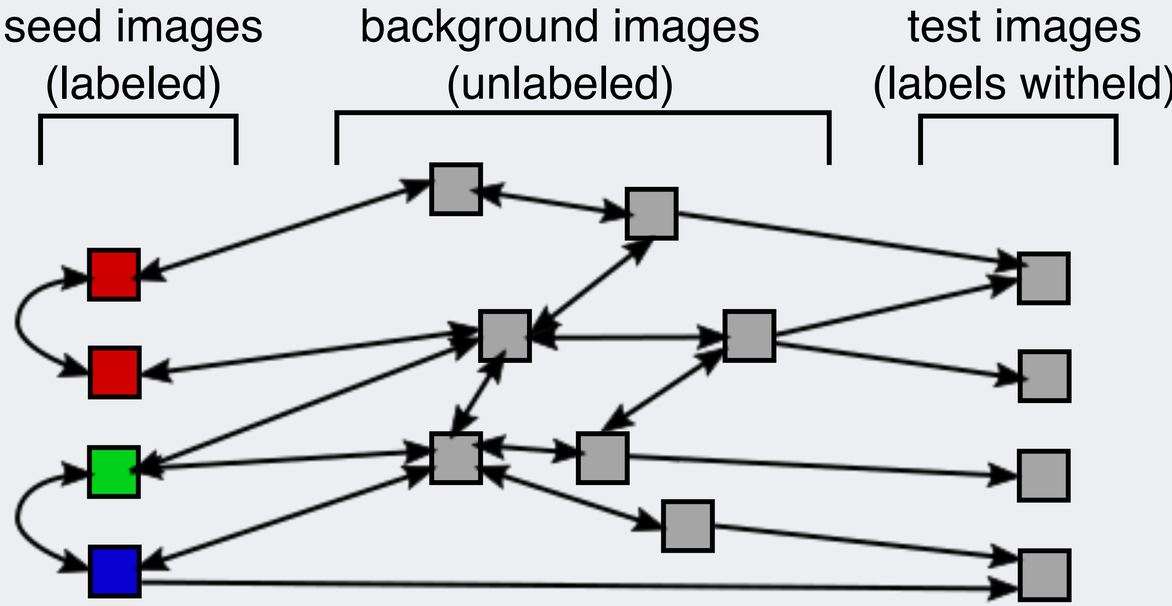
Problem setup

- Assume we have only 1,2, ..10 training images per class
  - Too little data to effectively train a CNN

- Typical approach: transfer learning
  - train an embedding on other classes
  - SVM or logistic regression classifier on top of the embeddings

Diffusion on a knn-graph

- We have:
  - a large set of background images
  - semantic embeddings for the images
- knn-graph
  - nodes = images
  - edges = links to the k nearest images



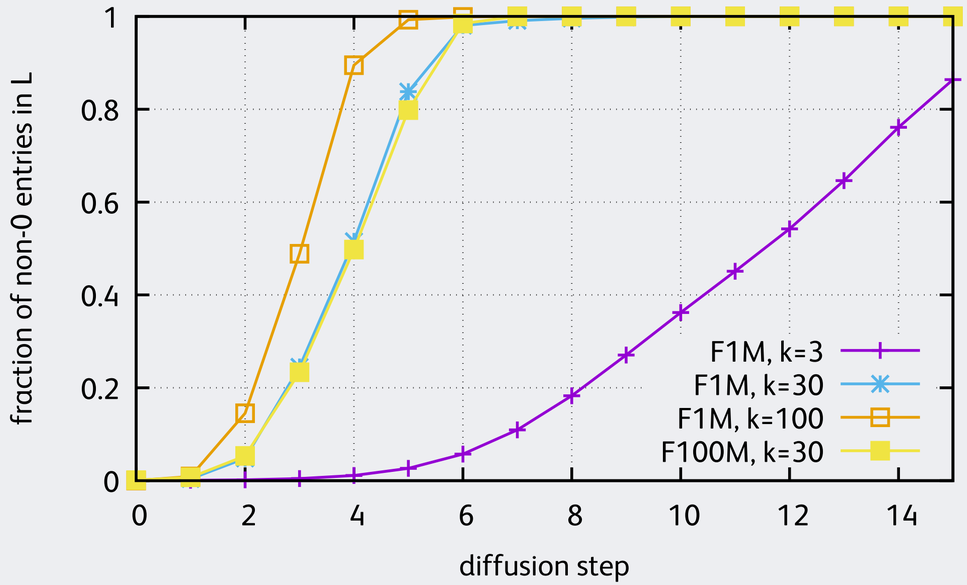
Diffusion on a knn-graph

Matrix view

$$L_0 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad L_{i+1} = \begin{bmatrix} \mathbf{W}_{\mathrm{LL}} & \mathbf{W}_{\mathrm{LB}} \\ \mathbf{W}_{\mathrm{BL}} & \mathbf{W}_{\mathrm{BB}} \end{bmatrix} \times L_i$$

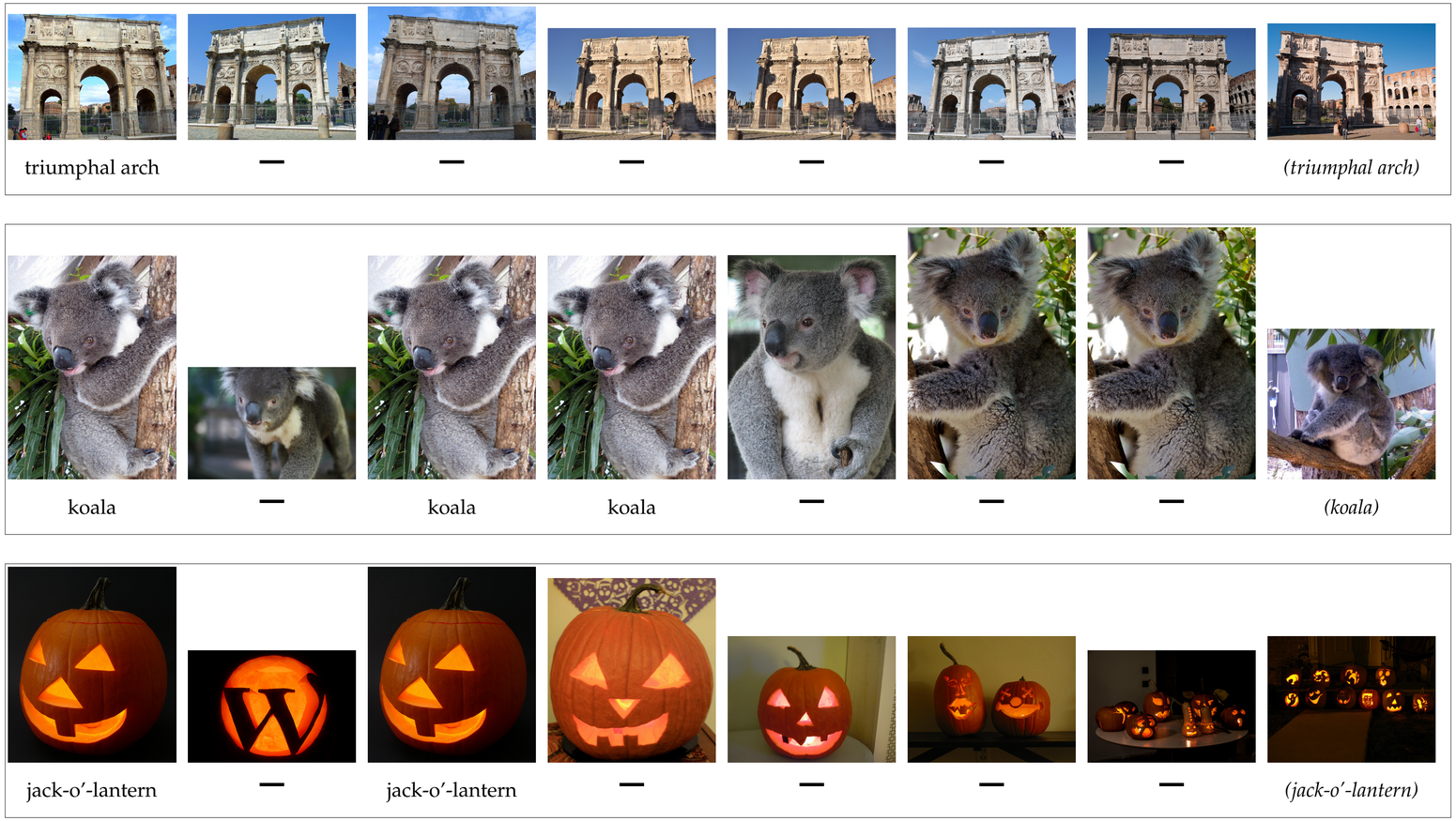
- matrix symmetrized and row-normalized
- no normalizations necessary during diffusion
- L1-normalize the L vector at the end of iterations

- early stopping (validated parameter)
- Can be performed on all classes at once
  - sparse matrix - dense martrix multiplications

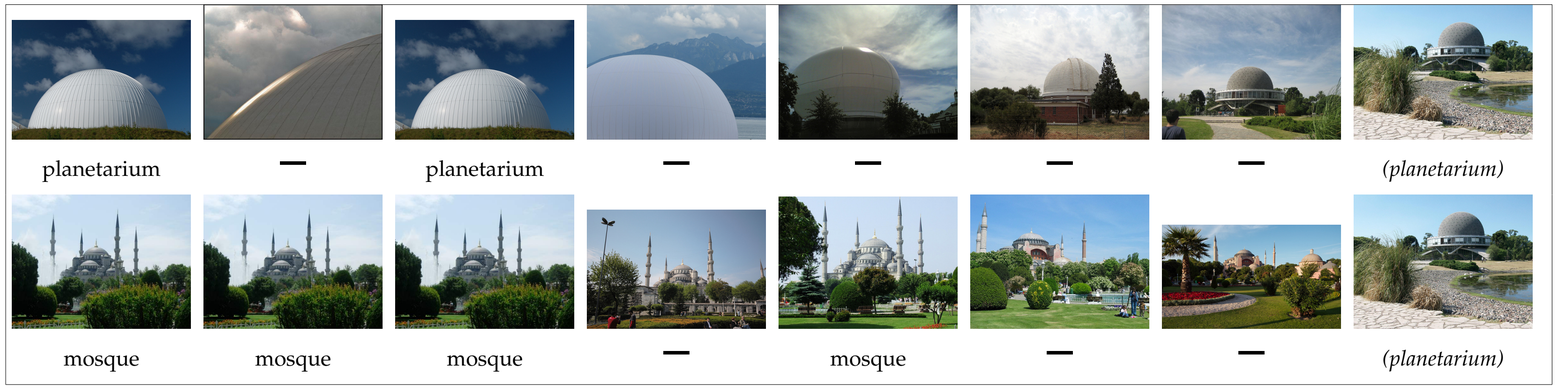


Visualize

- Visualization of the strongest path
- starting from the target images
- follow strongest edges



Failure cases – correct path first, and path produced by the method:



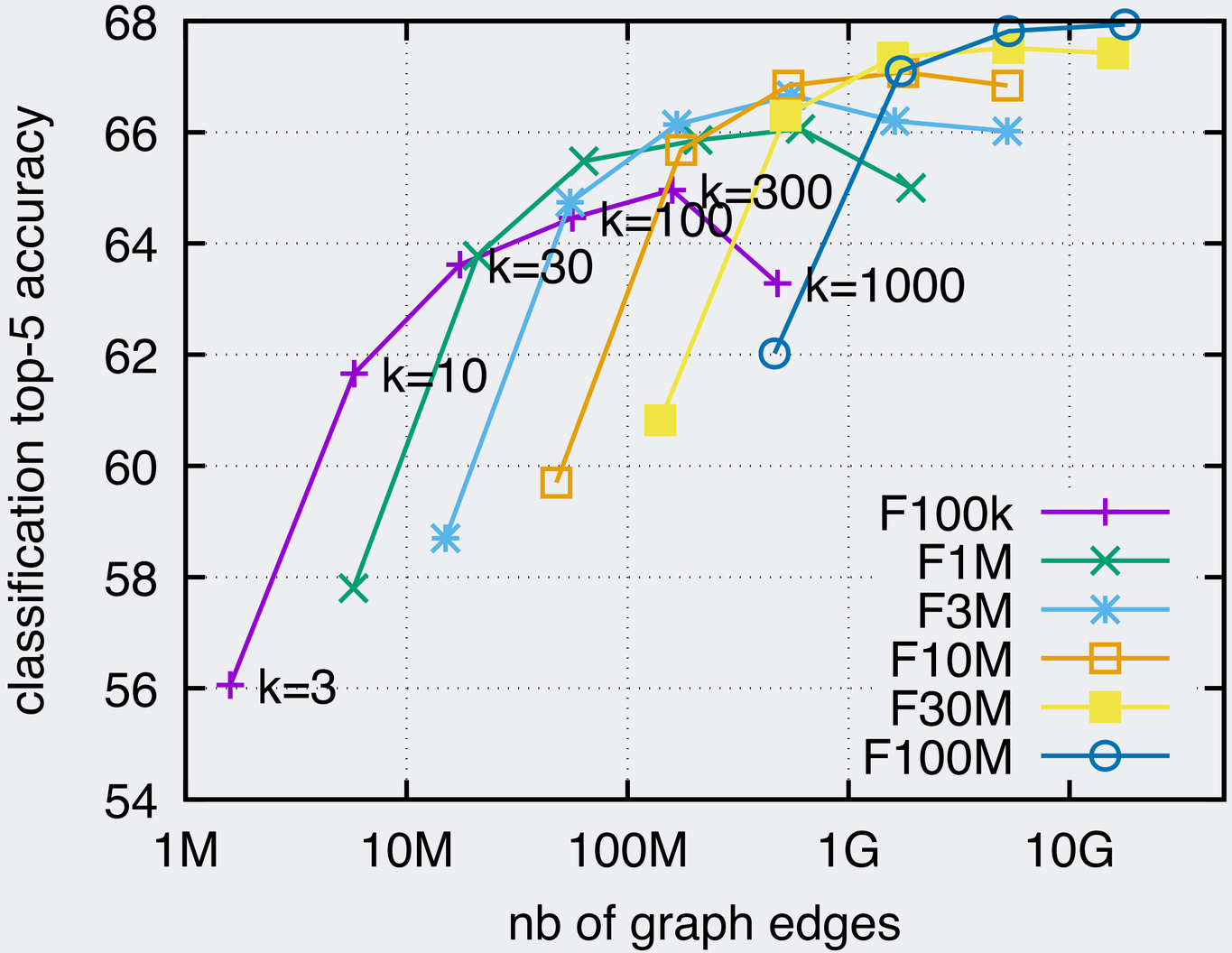
Results

- classification performance on a subset of ImageNet classes, with few (n=1..20) training images

	out-of-domain diffusion				in-domain	logistic	combined		[16]
n	none	F1M	F10M	F100M	Imagenet	regression	+F10M	+ F100M	
1	57.1	60.0	61.4	62.3	68.0	57.3	62.0	62.6	60.6
2	62.5	65.5	66.8	67.8	73.2	66.0	68.7	69.2	68.9
5	68.4	70.6	71.9	73.1	77.8	76.4	76.9	77.4	77.3
10	72.7	74.2	75.3	76.2	80.1	80.9	81.3	81.5	80.6
20	76.0	77.0	77.5	78.6	81.4	83.7	83.9	84.1	82.5

Results

- runtime depends on number of edges
- graph completion is slow
- useful in practice?



background	none	F1M	F10M	F100M
optimal iteration	2	3	4	5
timing: graph completion	2m57s	8m36s	40m41s	4h08m
timing: diffusion	4.4s	19s	3m44s	54m

Deep clustering

[Deep Clustering for Unsupervised Learning of Visual Features, Caron, Bojanowski, Joulin, Douze, ECCV'18]

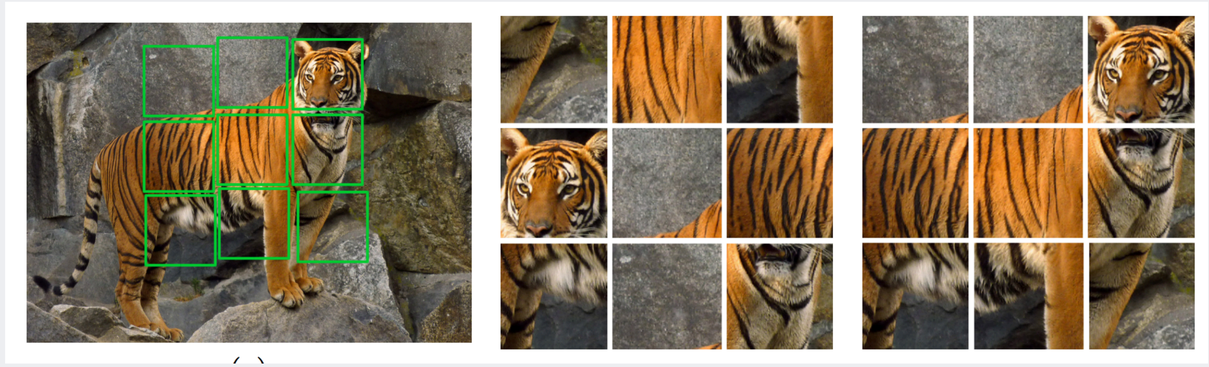
Unsupervised feature learning

Experimental context

- Given many unlabeled images, find an embedding

- How can we check whether the embedding is good?
  - fine-tune the convnet on another dataset that has fewer images (Pascal VOC)
  - see how it performs

- Related work:
  - invariance to data aguentation
  - pretext tasks: counting, patch layout prediction

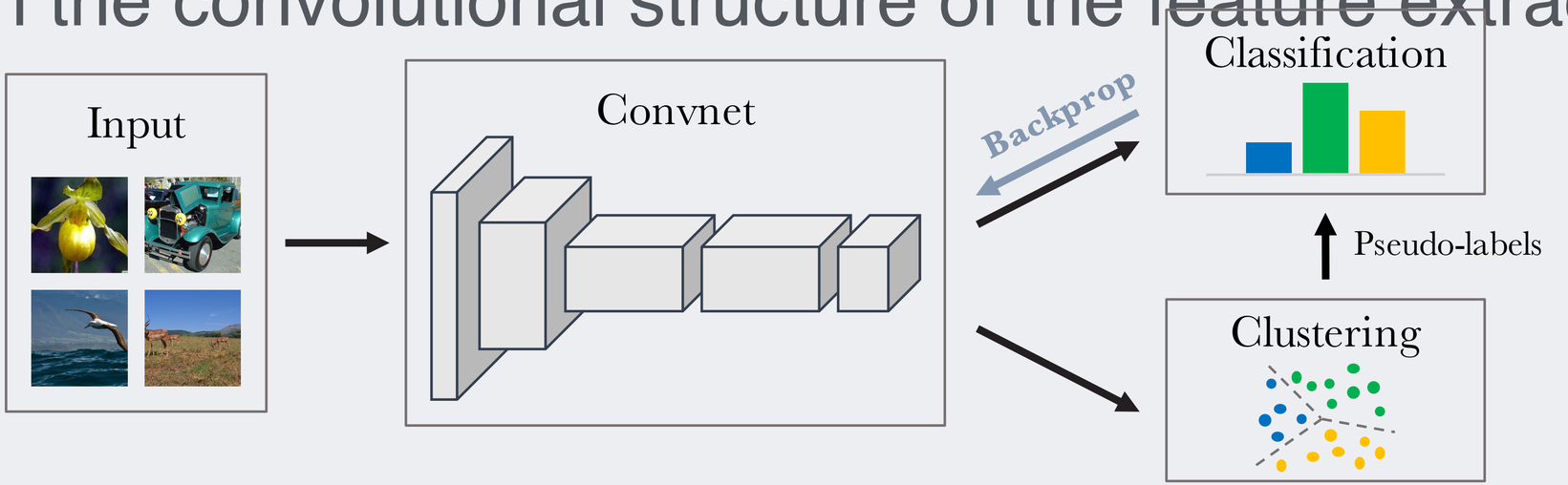


[Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles, Noroozi & Favaro, ECCV'16]

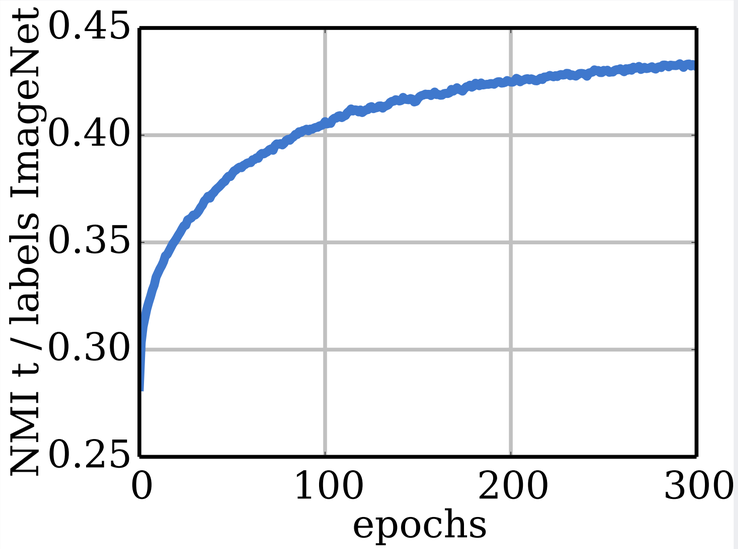
Our approach

Keep it simple!

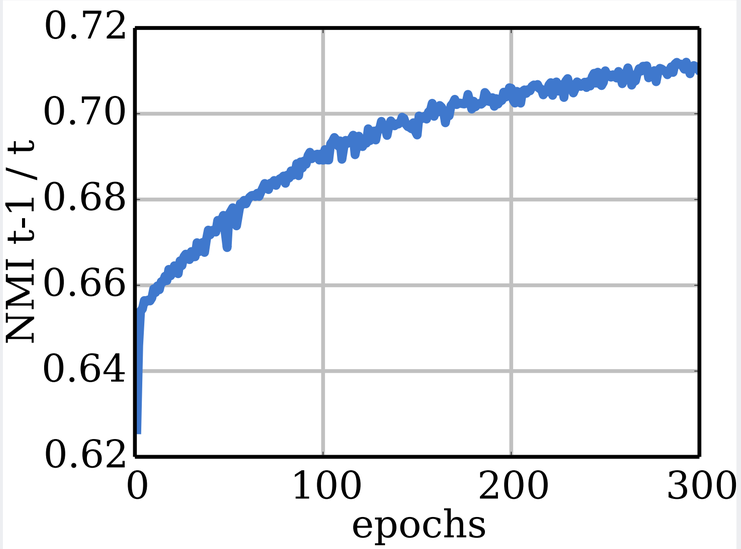
- random initialization of the convnet
- Iterate
  - k-means clustering of descriptors
  - train convnet to predict the clusters
- relies on the convolutional structure of the feature extraction



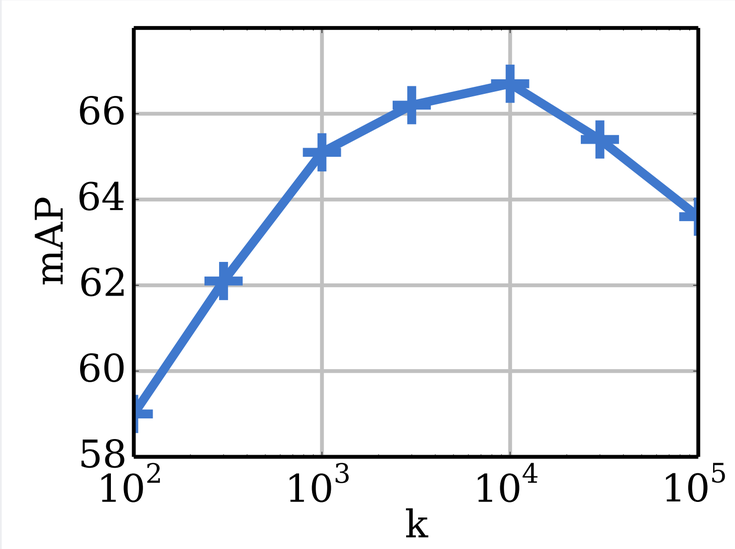
Parameters



(a) Clustering quality



(b) Cluster reassignment

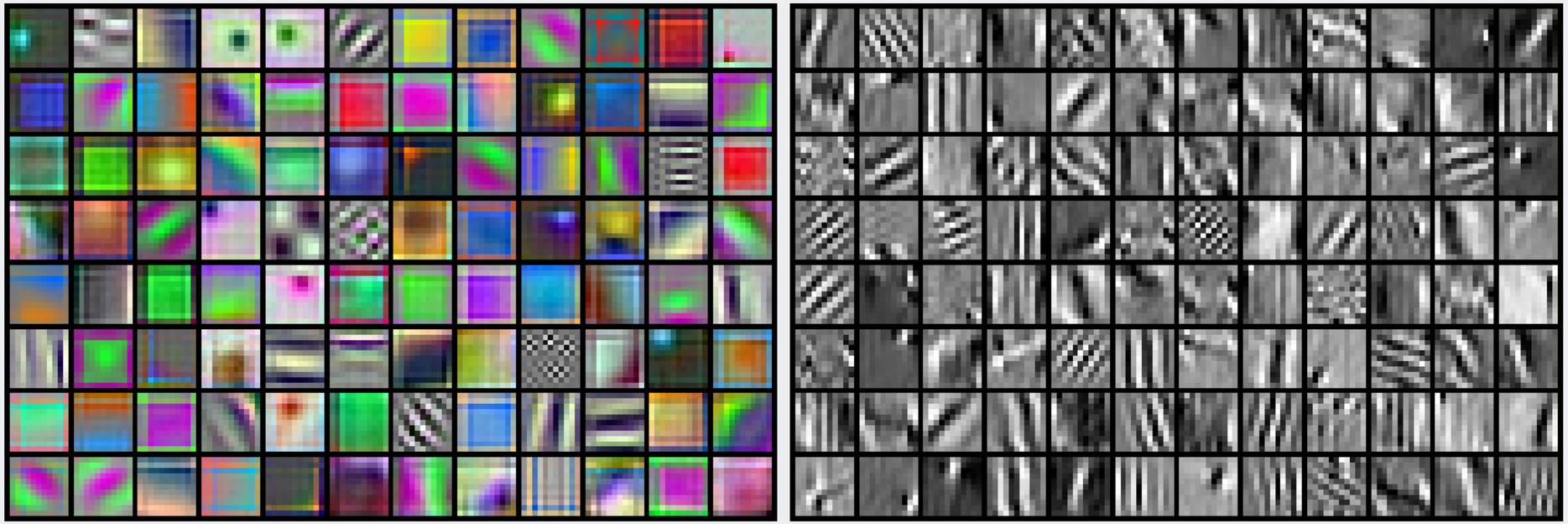


(c) Influence of k

- convergence in ~300 epochs
- Sobel filter
- data augmentation

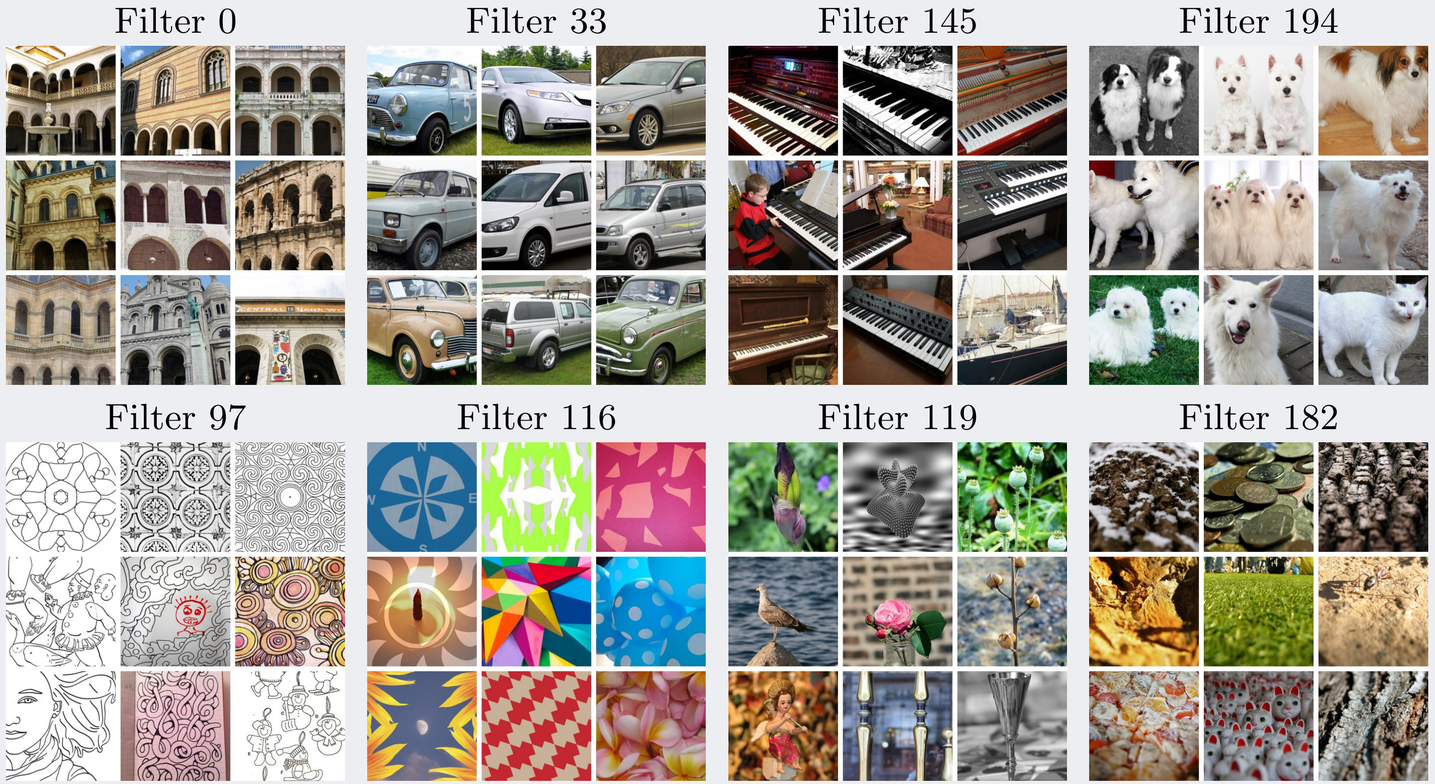
Visualization

- Direct visualization of the first convolutional filters



Visualization

- strongest images for some filters (last convolutional layer)



Results

- Image classification performance when using features at a certain level of the network
- variants:
  - use images from Imagenet vs. Flickr
  - use power iteration clustering vs. k-means

Method		Classification		Detection		Segmentation	
		FC6-8	ALL	FC6-8	ALL	FC6-8	ALL
ImageNet labels		78.9	79.9	–	56.8	–	48.0
Random-rgb		33.2	57.0	22.2	44.5	15.2	30.1
Random-sobel		29.0	61.9	18.9	47.9	13.0	32.0
Pathak *et al.* [38]		34.6	56.5	–	44.5	–	29.7
Donahue *et al.* [20]*		52.3	60.1	–	46.9	–	35.2
Pathak *et al.* [27]		–	61.0	–	52.2	–	–
Owens *et al.* [44]*		52.3	61.3	–	–	–	–
Wang and Gupta [29]*		55.6	63.1	32.8†	47.2	26.0†	35.4†
Doersch *et al.* [25]*		55.1	65.3	–	51.1	–	–
Bojanowski and Joulin [19]*		56.7	65.3	33.7†	49.4	26.7†	37.1†
Zhang *et al.* [28]*		61.5	65.9	43.4†	46.9	35.8†	35.6
Zhang *et al.* [43]*		63.0	67.1	–	46.7	–	36.0
Noroozi and Favaro [26]		–	67.6	–	53.2	–	37.6
Noroozi *et al.* [45]		–	67.7	–	51.4	–	36.6
DeepCluster		**72.0**	**73.7**	**51.4**	**55.4**	**43.2**	**45.1**
DeepCluster	YFCC100M	67.3	69.3	45.6	53.0	39.2	42.2

Conclusion

- We have to go to less supervised learning

- Similarity search:
  - non-parametric method for media matching
  - efficient / scalable

- A fundamental tool for unsupervised learning
  - not only for vision: see [Word translation without parallel data, Conneau, Lample, Ranzato, Denoyer, Jégou, Arxiv'18]

- All works are open-sourced
  - check it out!

Source: sed ut unde omnis

Source: sed ut unde omnis