Computer Vision & Deep Learning

Helmut Grabner
helmut.grabner@zhaw.ch

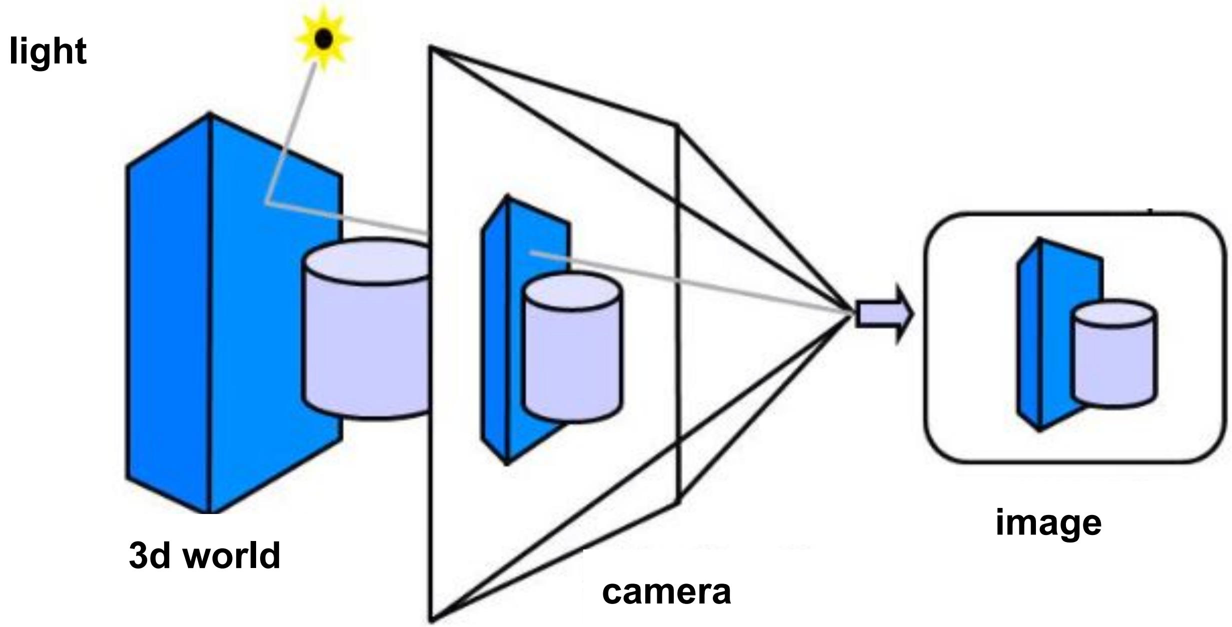
Gesellschaft für Mathematik an Fachhochschulen
2021/12/18

Zurich University
of Applied Sciences

zh
aw
School of
Engineering

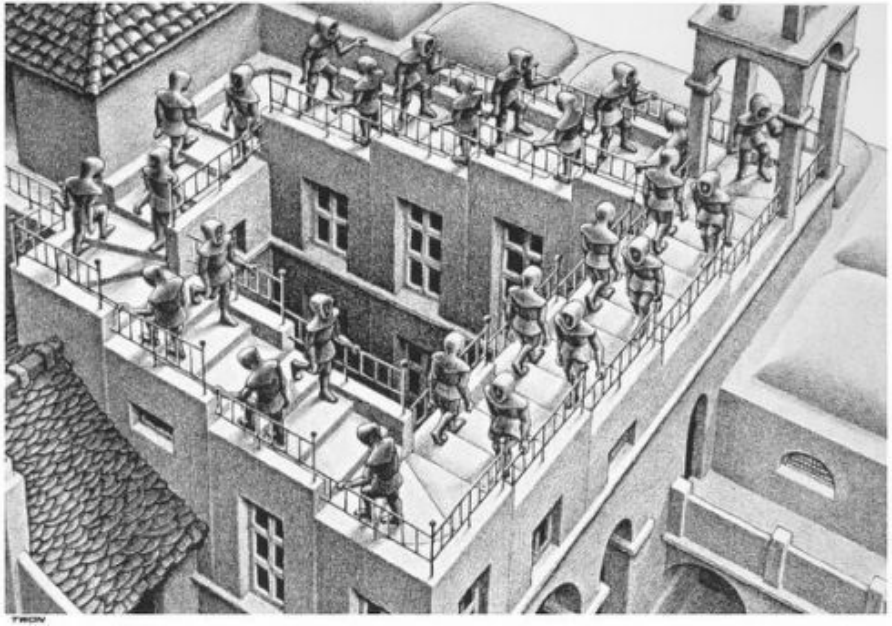

ARRIBA
ARRIBA
Spanish English
UP
UP

Computer Vision



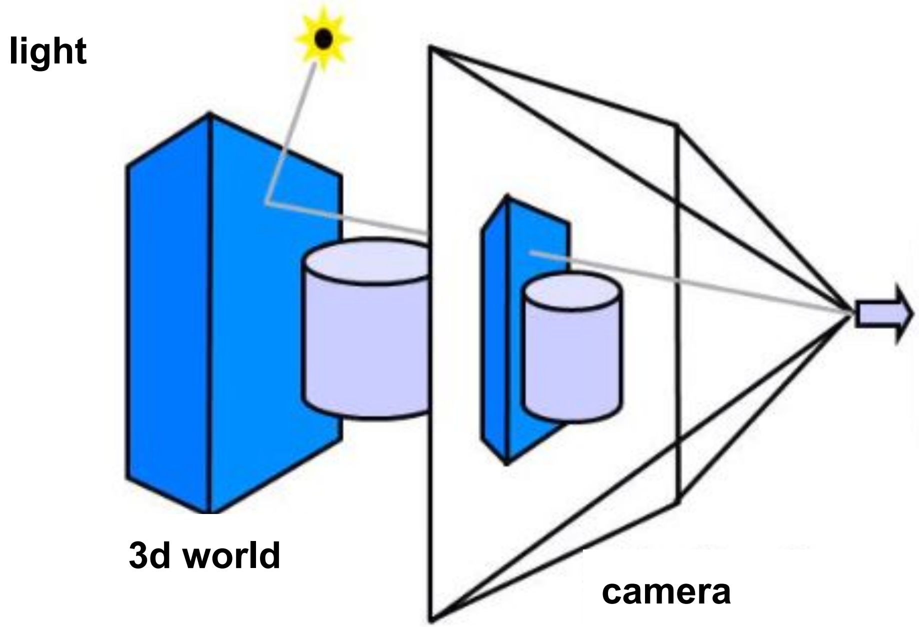
3d - 2d Information Loss

M. C. Escher



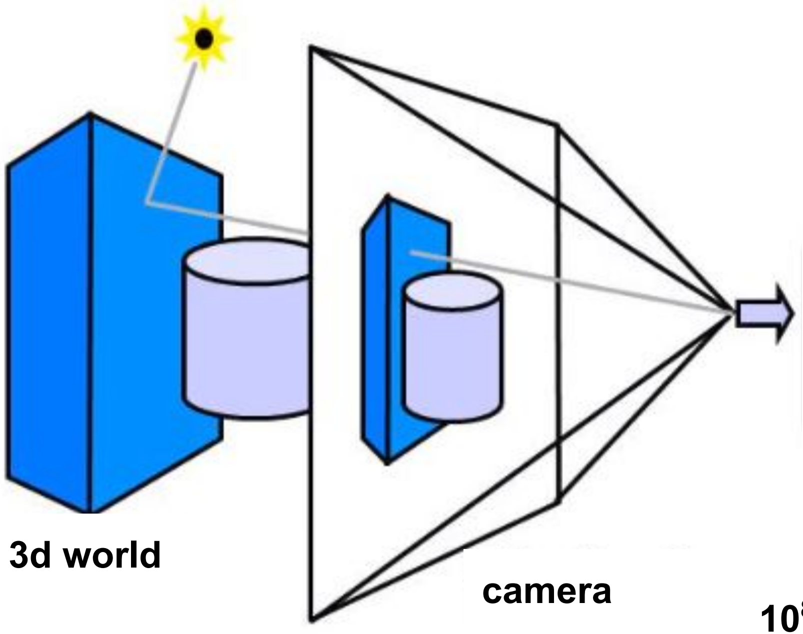
Technorama

Computer Vision -- Why is it hard?



157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

Computer Vision -- Why is it hard?



157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168										
205	174										
190	216										
190	224										
190	214										
187	196										
183	202										

10^{80} atoms in the universe
10^{90} possible images (8 bit 6 x 6px)

Example: Image Tagging

Image

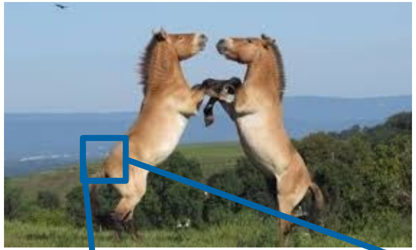
Programming

```
def classify(image):
    #some magic
    return classLabel
```

Horse: yes/ no

Example: Image Tagging

Image



Programming

```
def classify(image):
    #some magic
    return classLabel
```

Horse: yes/ no

157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

Pixels

Magic: No obvious way to hard-code the Algorithm! (unlike e.g., sorting numbers)

Challenges

Viewpoint, illumination, scale, deformation, occlusion (self-occlusion), motion,...





Challenges

… intra - and inter class variations

Sheepdog or mop

Chihuahua or muffin

Machine Learning to Rescue!

Data, Examples

“ML - Magic”

Prediction

$\{(x_1, y_2), (x_2, y_2), \ldots (x_n, y_n)\}$

$H: x \rightarrow y$

$y = H(x)$

person

no person

Machine Learning in a Nutshell

Machine Learning = Learning from Examples

To do it well, you need a lot of data.

Computer Vision

Features/ Representation

Size, Color,
Shape, Texture…

Classification

What are good features?

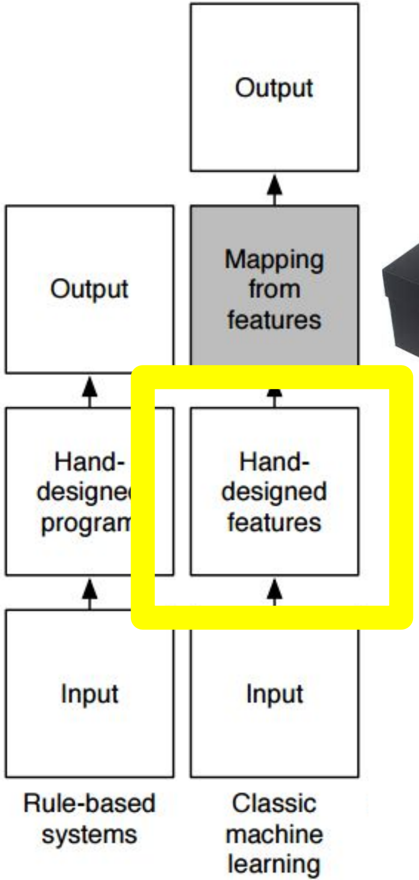
Who to get the representation?

Good Old Fashion (GOF) Machine Learning

Good & compact Features are more important than the model choice!

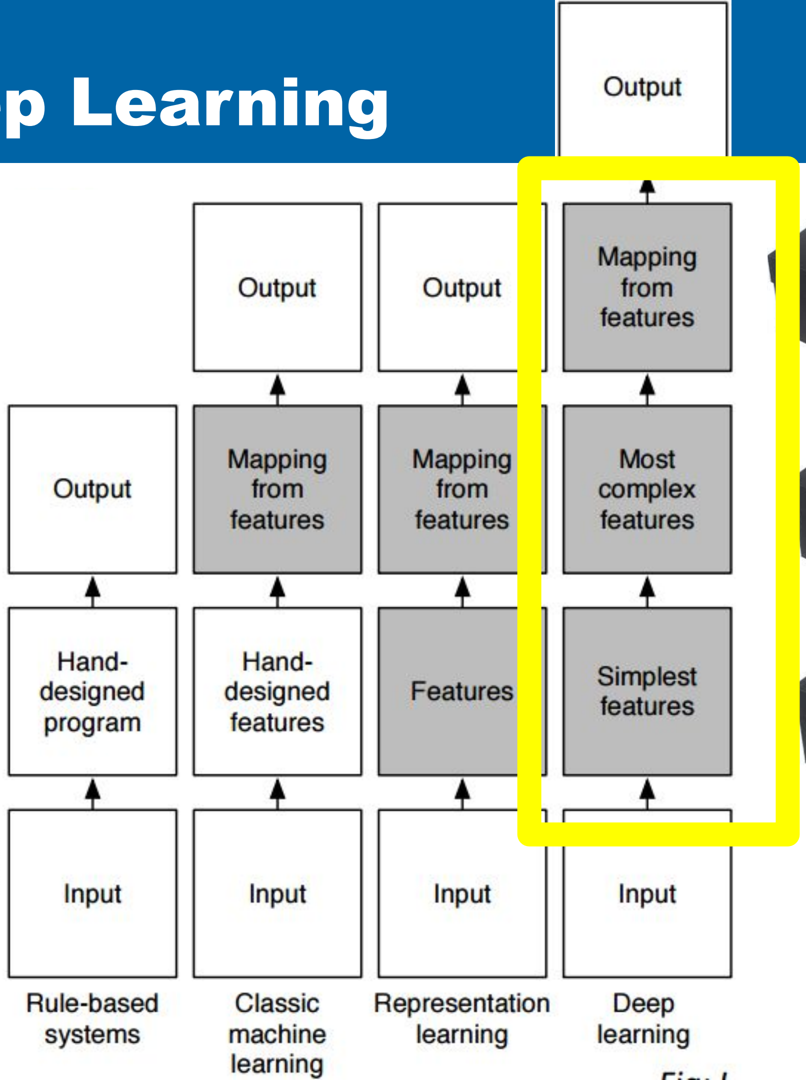
What prior knowledge can easily expressed in certain features and models?

$$t = \sqrt{\frac{2h}{g}}$$



Adapted from Y. Bengio, Deep Learning: Theoretical Motivations, [Lecture](), 2015

GOF ML vs. Deep Learning



$$t = \sqrt{\frac{2h}{g}}$$

Adapted from Y. Bengio, Deep Learning: Theoretical Motivations, Lecture, 2015

Building blocks of a (Deep) Neural Network

Neurons: Linear Combination + Activation Function

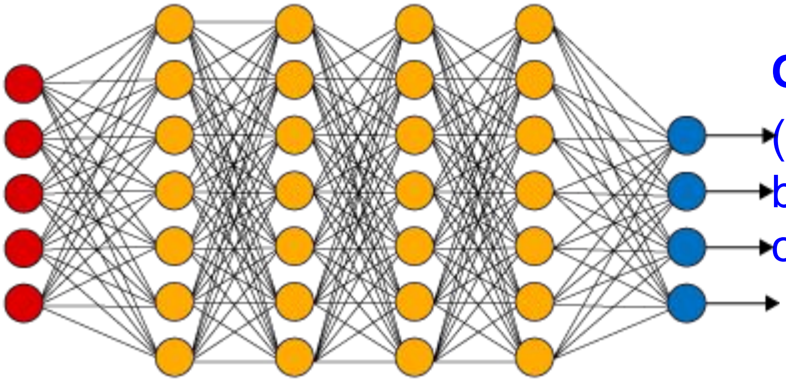
Feature sharing, e.g., Convolutions + Pooling (-> CNNs)

Layers: Set of neurons (Feed forward neural network)

Input

(basically raw)

Weights (to be learnt)

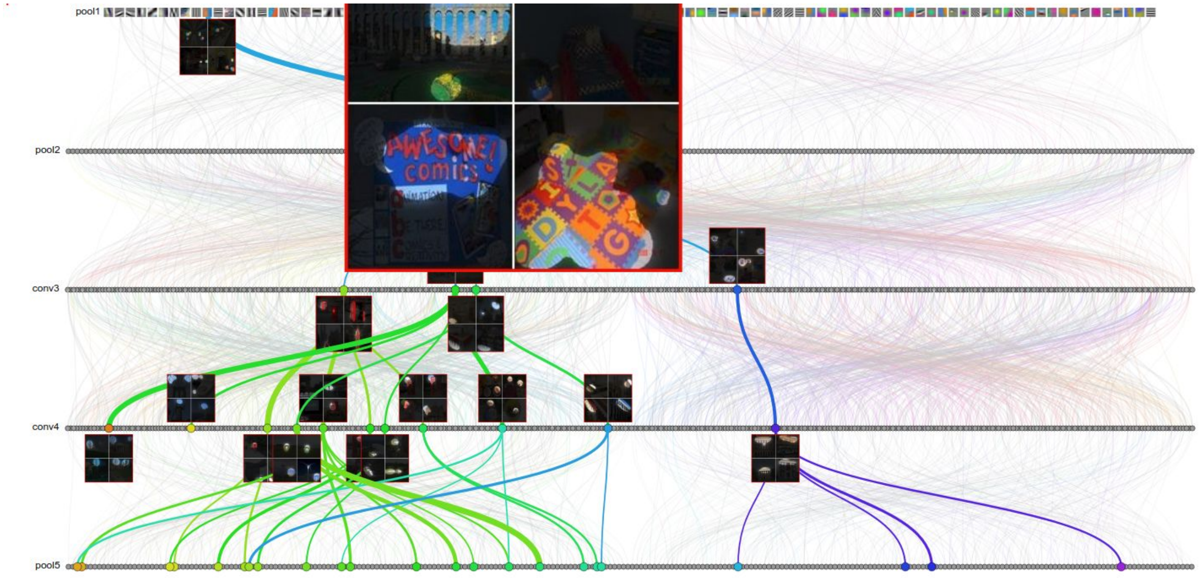


Output

(probability of input belonging to certain class)

CNN Learnt Representation

Antonio Torralba's: [drawNet]



Deep Learning in Computer Vision

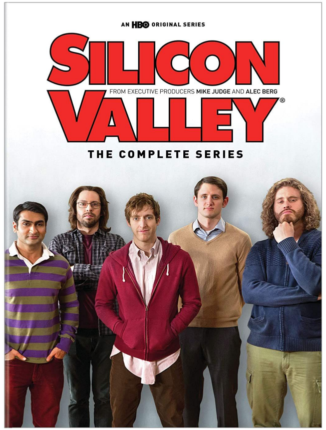
Yann LeCun -- first CNNs in ‘88, “ignored” for a long time!

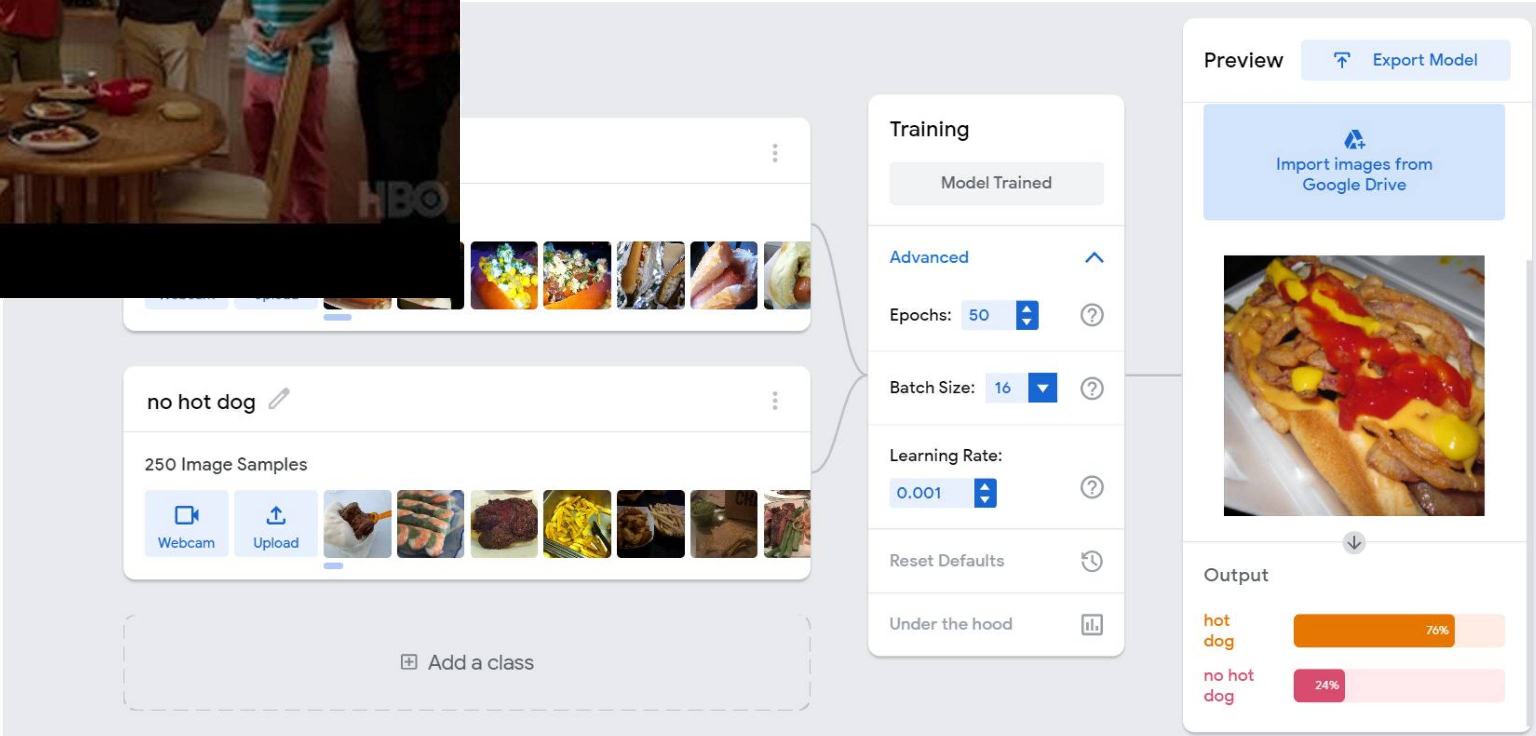

4627
4 6 2 7

“Hello World”: 12 lines of code

```
import tensorflow as tf
mnist = tf.keras.datasets.mnist

(x_train, y_train),(x_test, y_test) = mnist.load_data()
x_train, x_test = x_train / 255.0, x_test / 255.0

model = tf.keras.models.Sequential([
  tf.keras.layers.Flatten(input_shape=(28, 28)),
  tf.keras.layers.Dense(512, activation=tf.nn.relu),
  tf.keras.layers.Dropout(0.2),
  tf.keras.layers.Dense(10, activation=tf.nn.softmax)
])
model.compile(optimizer='adam', loss='sparse_categorical_crossentropy', metrics=['accuracy'])

model.fit(x_train, y_train, epochs=5)
model.evaluate(x_test, y_test)
```

Hot Dog or Not Hot Dog





Large Scale Visual Recognition Challenge

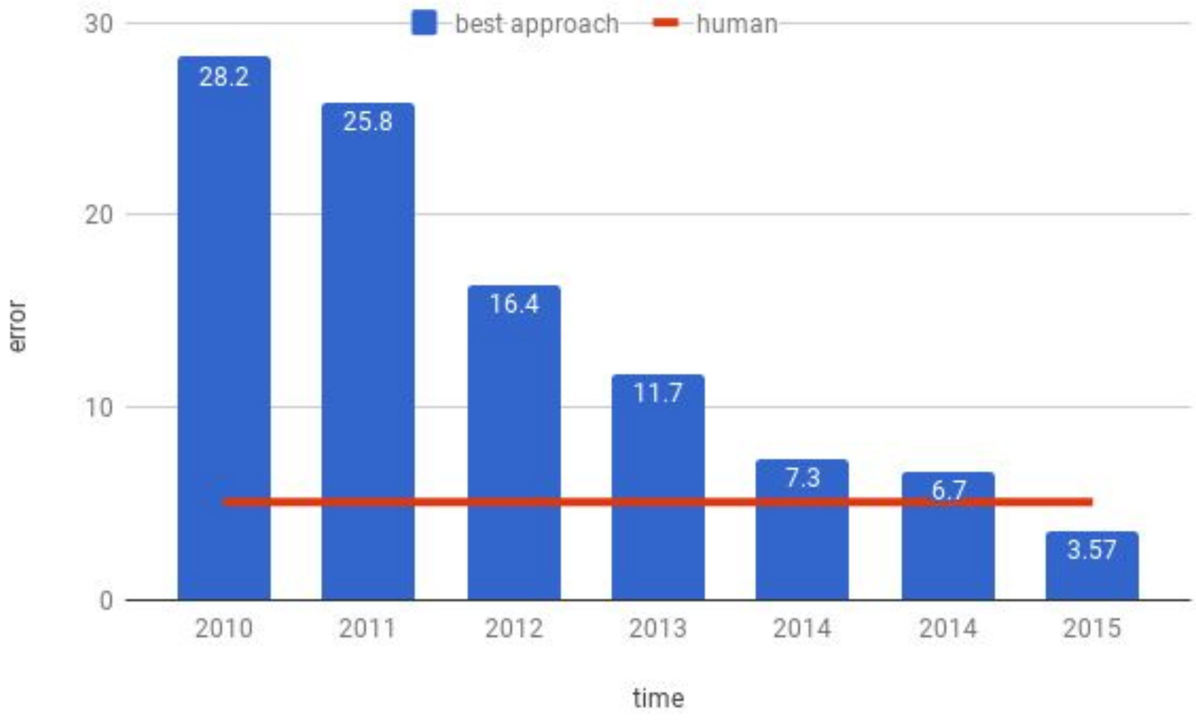
IMAGENET

Benchmark datasets drive progress

> 3 Million labelled images

> 1000 categories

Performance on the ImageNet Challenge



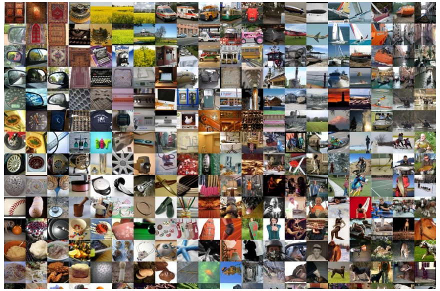
Interests in Deep learning (CVPR'15 WS)

CNN is not Invented Overnight

Time	Task	Computational Power (# of Transistors)	Data (# pixels)
1998 LeCun et al.		 10^6	10^7 NIST

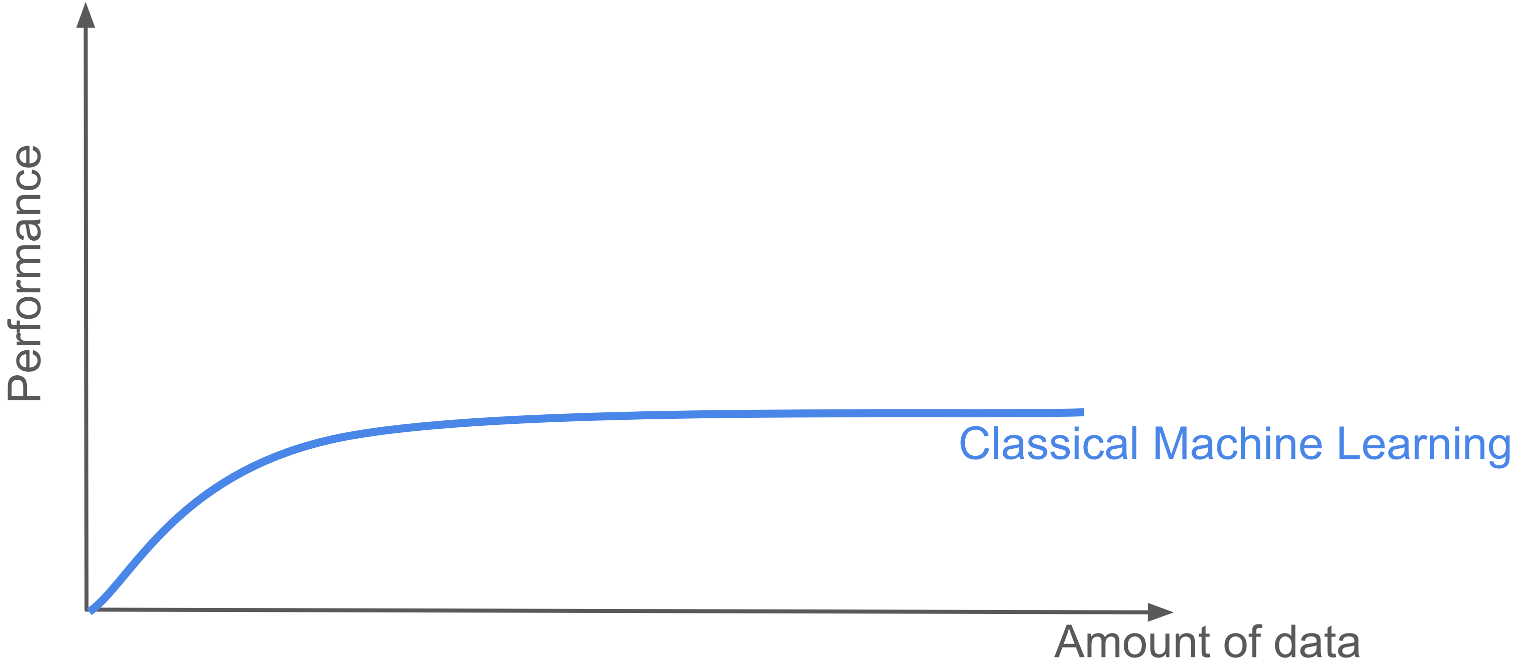
Inspired/ Adapted from: CNN in Practice, 2016, Stanford, [Lecture]

CNN is not Invented Overnight

Time	Task	Computational Power (# of Transistors)	Data (# pixels)
1998 LeCun et al.		 10^6	10^7 NIST
2012 Hinton et al.		 10^9	10^{14} IMAGENET

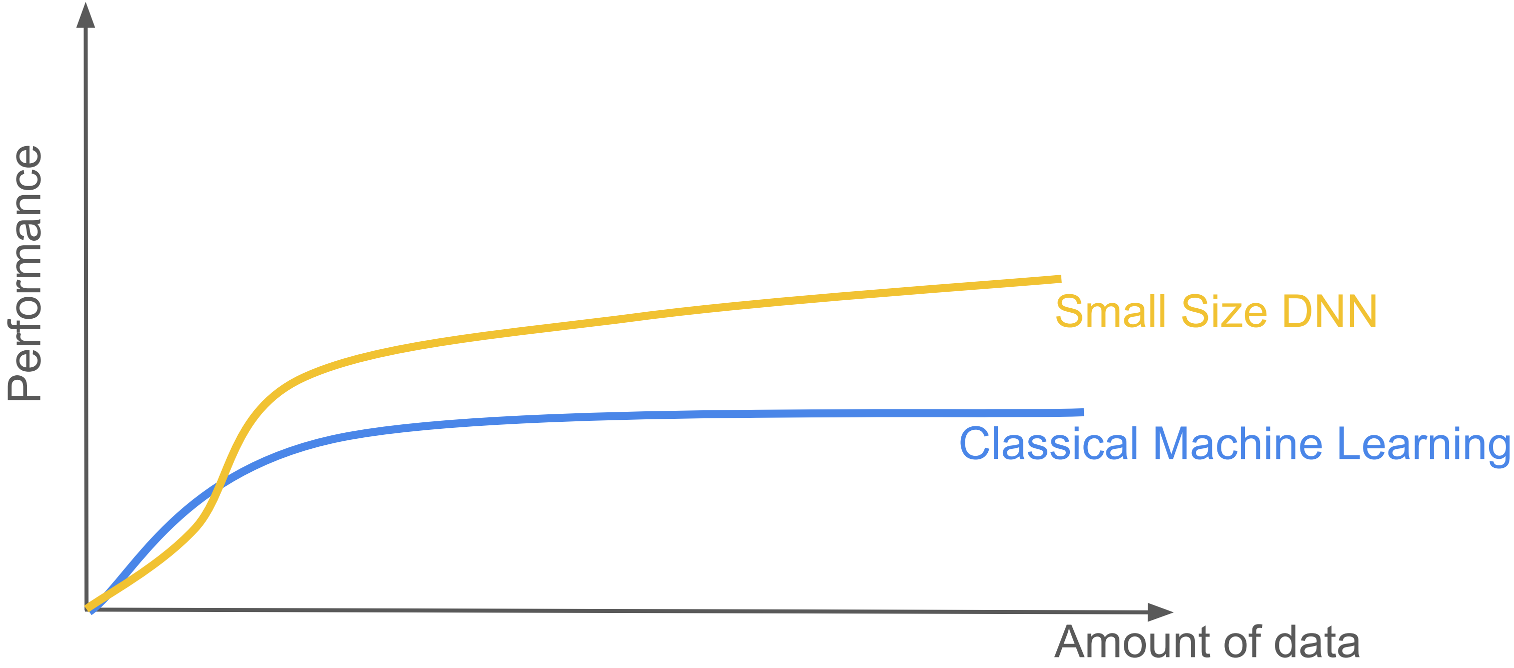
Inspired/ Adapted from: CNN in Practice, 2016, Stanford, [Lecture]

Big Data meets Deep Learning



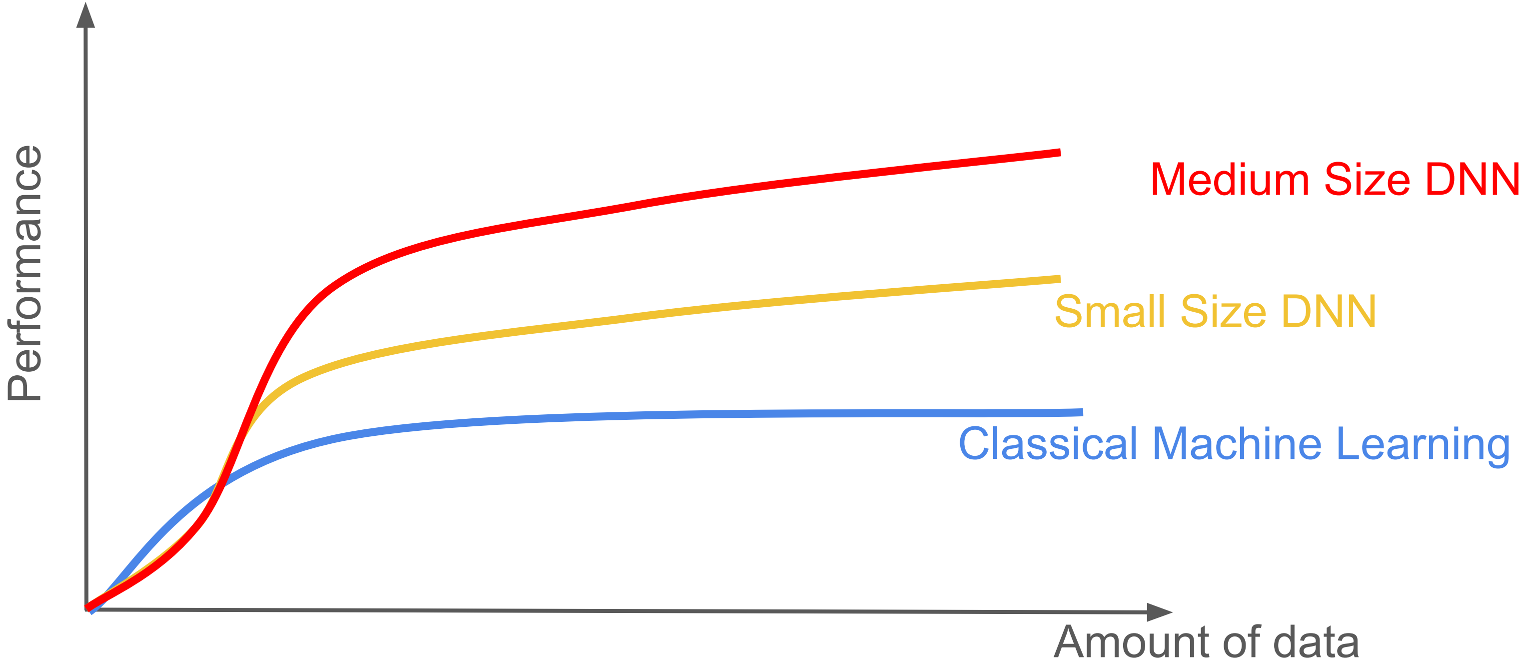
A. Ng., The State of AI, YouTube
N. Tishby et al., *Opening the Black Box of Deep Neural Networks via Information*, pdf, YouTube

Big Data meets Deep Learning



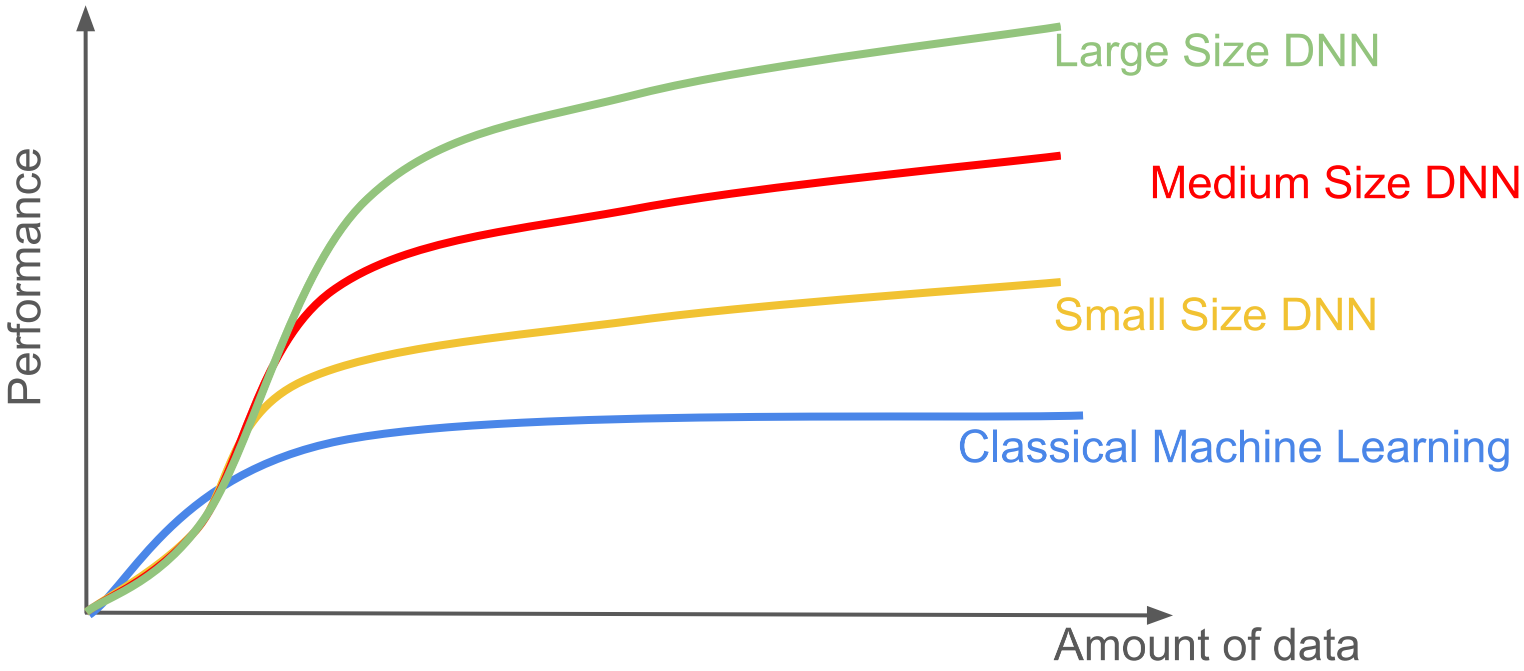
A. Ng., The State of AI, YouTube
N. Tishby et al., *Opening the Black Box of Deep Neural Networks via Information*, pdf, YouTube

Big Data meets Deep Learning



A. Ng., The State of AI, YouTube
N. Tishby et al., *Opening the Black Box of Deep Neural Networks via Information*, pdf, YouTube

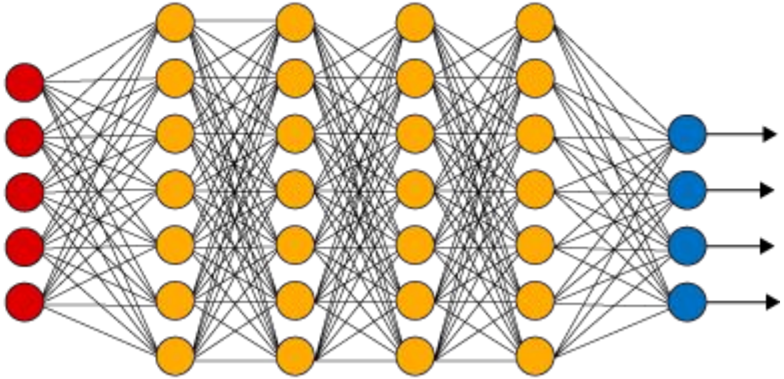
Big Data meets Deep Learning



A. Ng., The State of AI, YouTube
N. Tishby et al., *Opening the Black Box of Deep Neural Networks via Information*, pdf, YouTube

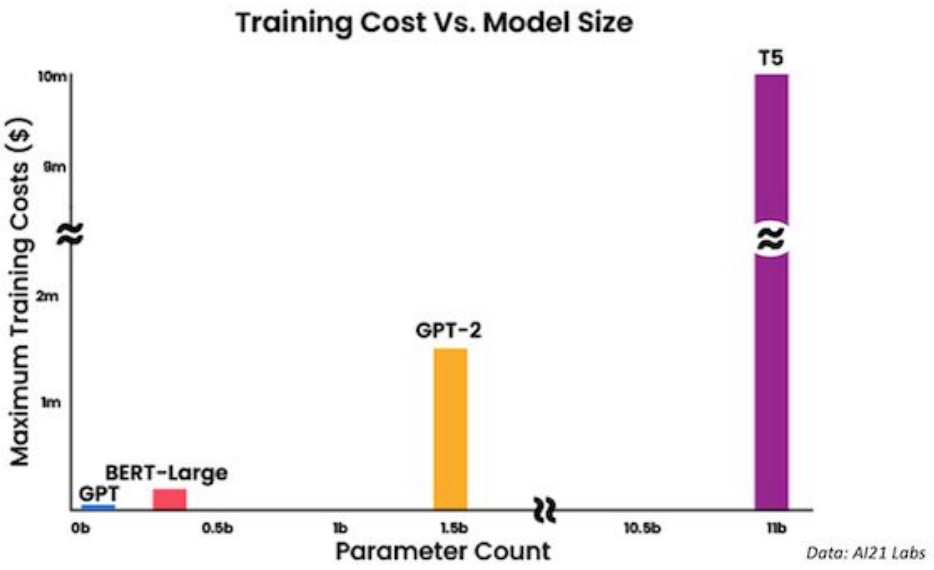
~1 billion = 1,000,000,000,000 parameters

~ 100 - 1,000 parameters



11 billion Parameters!

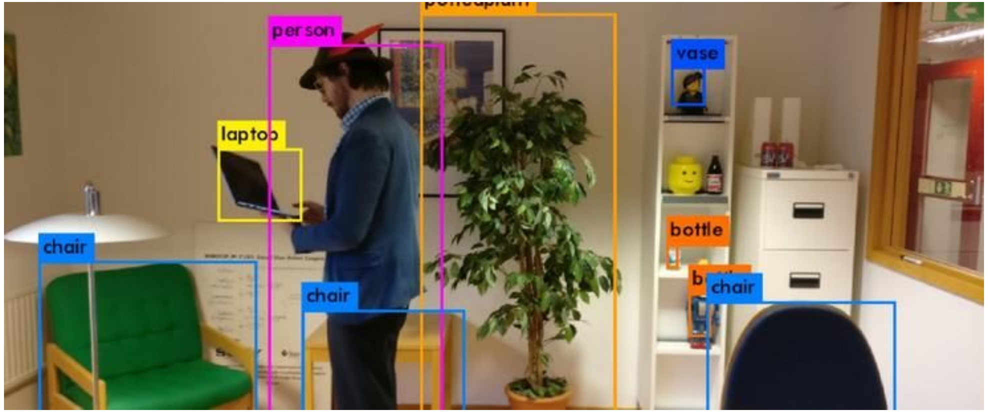
Behind the news: In 2020, researchers estimated the cost of training a model ‹ 1.5 billion parameters (the size of OpenAI's GPT-2) on the Wikipedia and Book corpora at $1.6 million. They gauged the cost to train Google's Text-to-Text Transformer (T5), which encompasses 11 billion parameters, at $10 million. Since then, Google has proposed Switch Transformer, which scales the parameter count to 1 trillion no word yet on the training cost.



Beyond Image Classification

Facebooks's Detectron2 (2019/10)

[Detectron2: A PyTorch-based modular object detection library]()



Applications

Pixels

Machine Learning

Horse: yes/ no

Audio Seg.

“See you tomorrow”

Text

“Hello, my name is Bob”

“Hallo, mein Name ist Bob”

Machine Learning Tasks

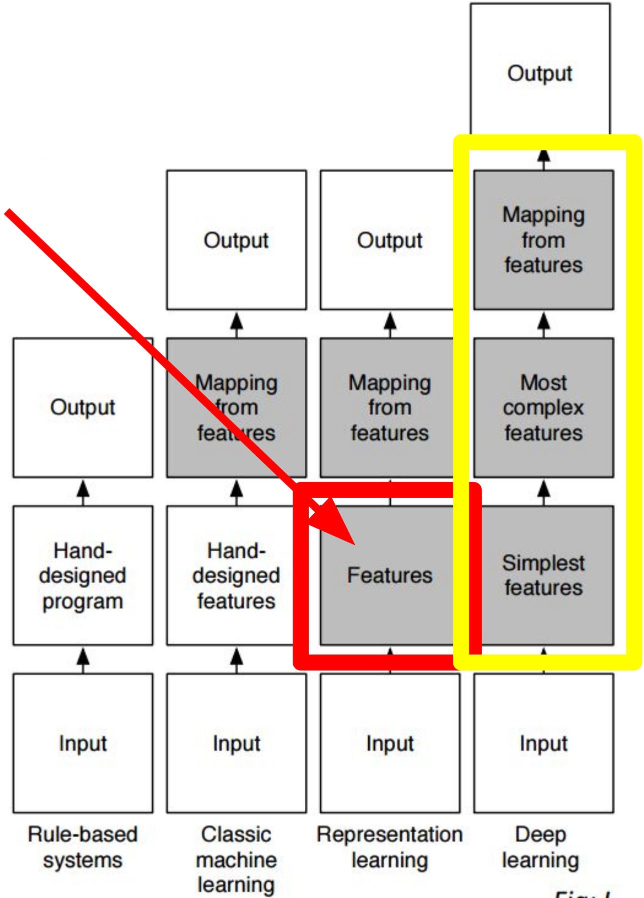
Supervised Learning:	Labeled Data: features -> outputs data + labels
Unsupervised Learning:	Unlabeled Data: Explore the Structure data
Semi-Supervised Learning:	Labeled and Unlabeled Data data + some labels
Self Supervised Learning:	Predict part of the Input data + data
Reinforcement Learning:	Learning policies experience + rewards (feedback)

Machine Learning Tasks

Supervised Learning:	Labeled Data: features -> outputs data + labels
Unsupervised Learning:	Unlabeled Data: Explore the Structure data
Semi-Supervised Learning:	Labeled and Unlabeled Data data + some labels
Self Supervised Learning:	**Predict part of the Input** **data + data**
Reinforcement Learning:	Learning policies experience + rewards (feedback)

(Deep) Representation Learning

“Classical” Representation Learning: Unsupervised



Output = (part of the) Input or automatically generated & task related targets

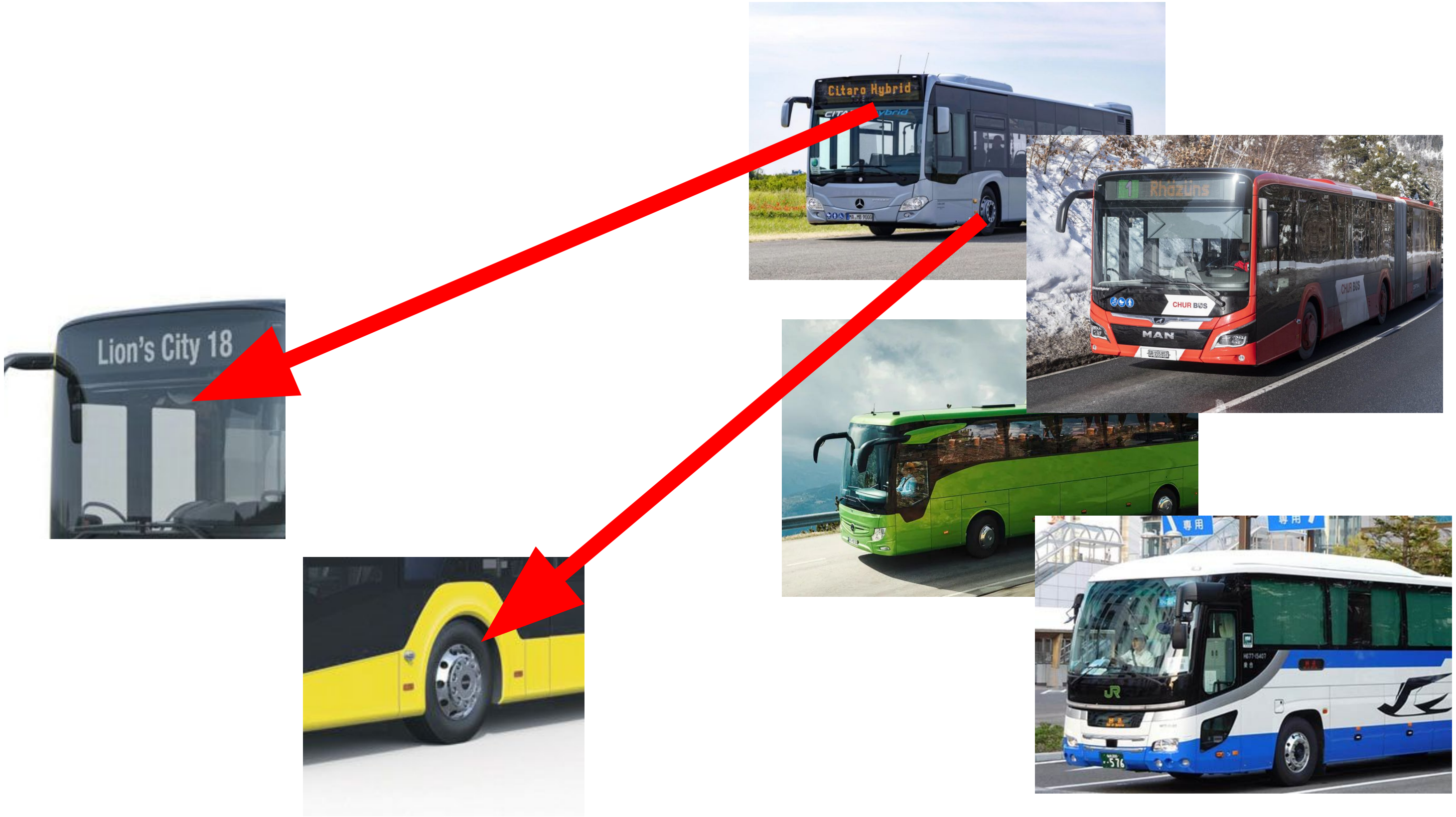
Supervised Rep. Learning

Self Supervised Learning

?	?	?
?		?
?	?	?

Self Supervised Learning

Self Supervised Learning



Learning Representations

Context as supervision

- Predict part of the Image
- Predict location of image patches

Invariance as supervision

- Rotate the image & predict this rotation

180 deg

"Typical" properties

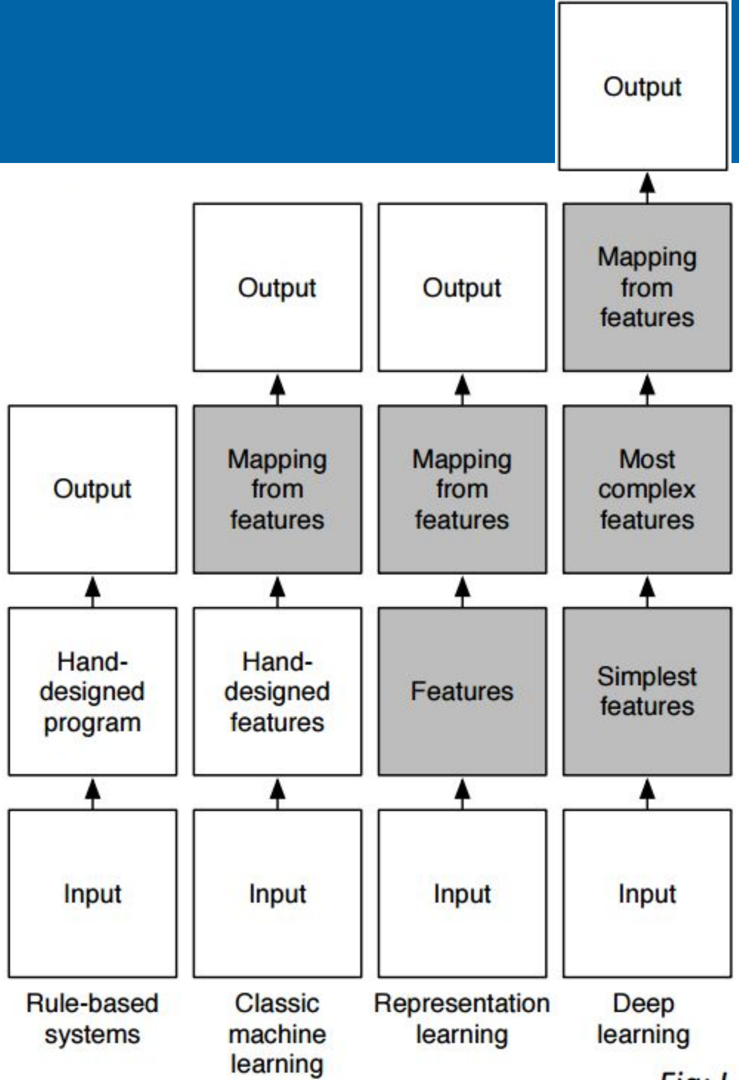
- grey scale -> color

Word2Vec: Word Embedding

Machine learning (ML) is the study of computer algorithms that improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine [redacted] algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so.

Summary

$$t = \sqrt{\frac{2h}{g}}$$

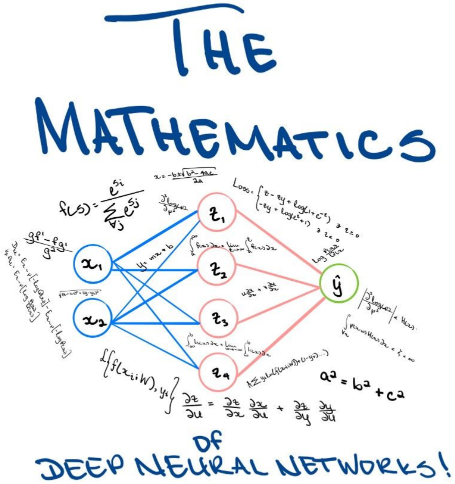


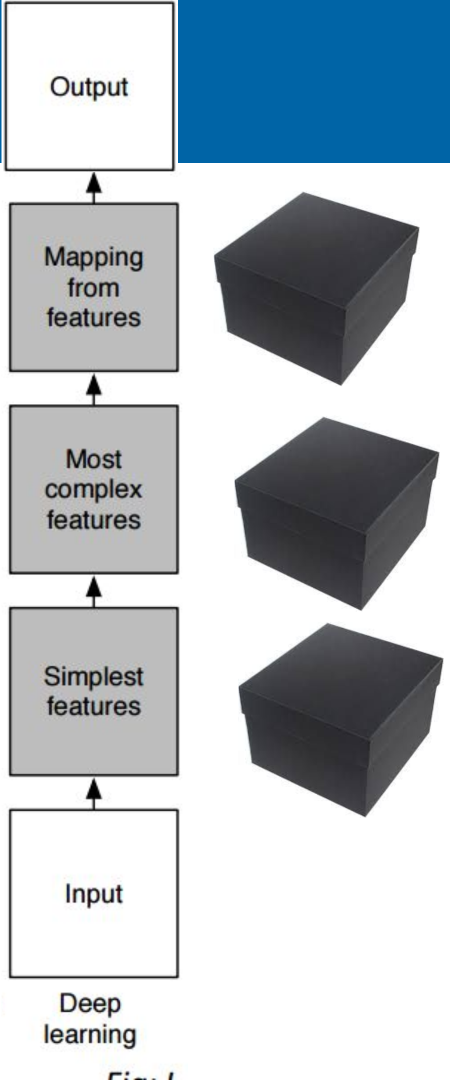
Adapted from Y. Bengio, Deep Learning: Theoretical Motivations, [Lecture](), 2015

Next…

… how to train the network?

Tom Weinmann



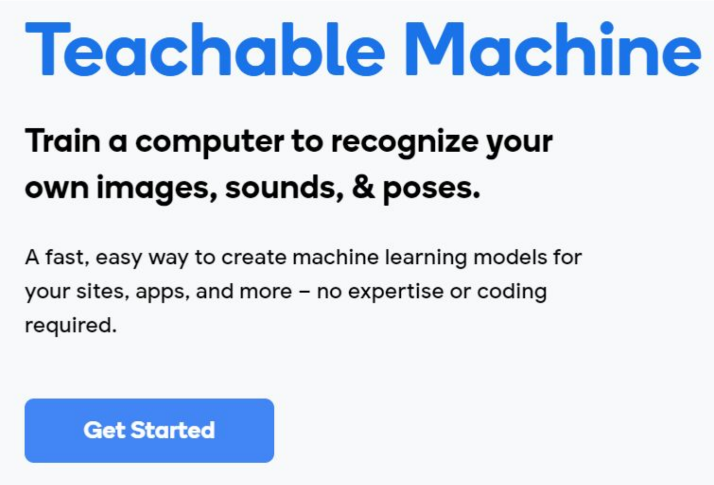


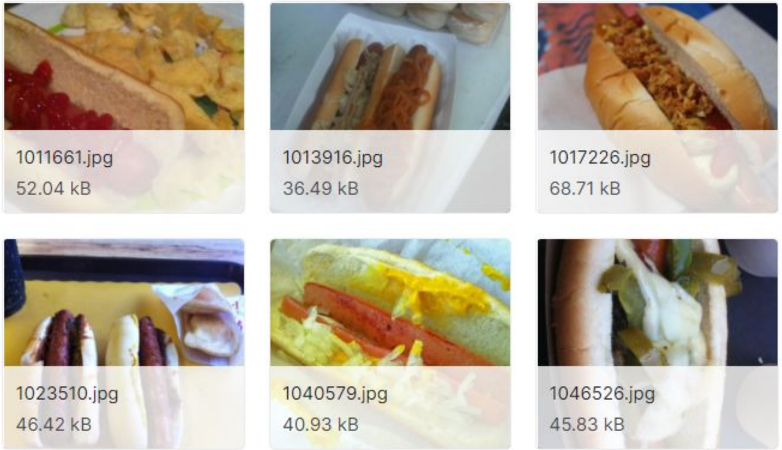

HBO

Your turn!

https://teachablemachine.withgoogle.com/

Data: https://www.kaggle.com/dansbecker/hot-dog-not-hot-dog





Research Interests

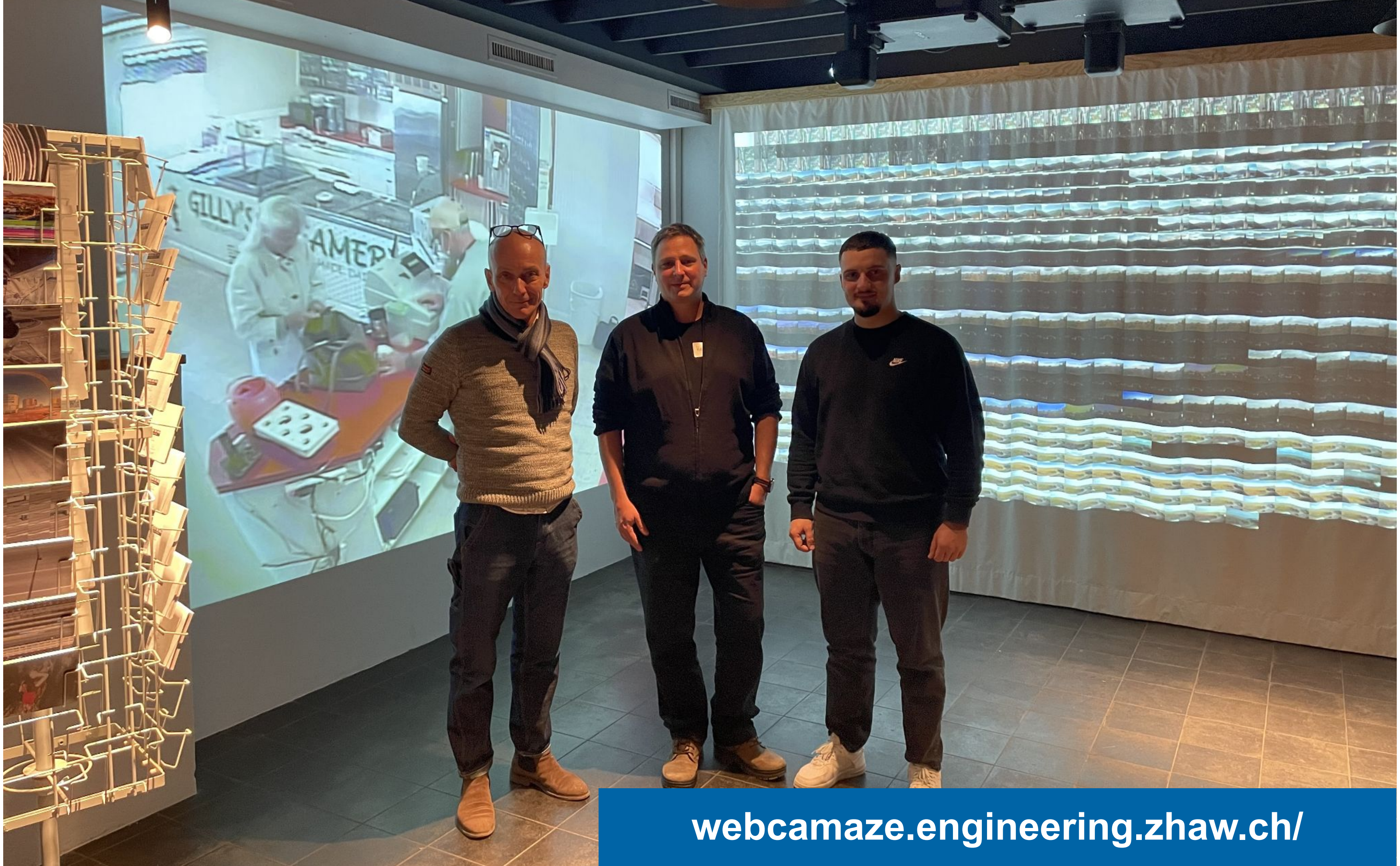
Why does the man smile?

DISCLAIMER: There are no sexist intention at all!

> 10,000 webcams -- an image every 15 minutes -- ~1,000,000 images per day

Machine Learning

Interestingness -- *the power of attracting or holding one's attention.*

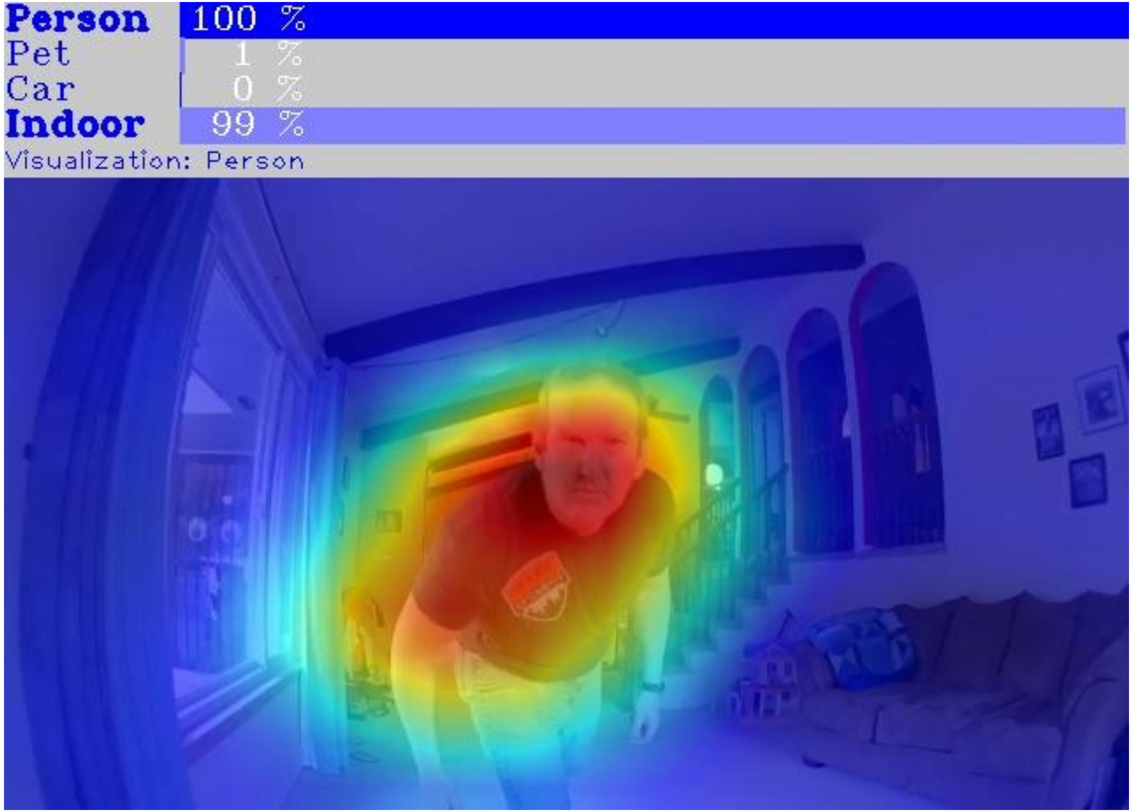


webcamaze.engineering.zhaw.ch/

Additional Slides

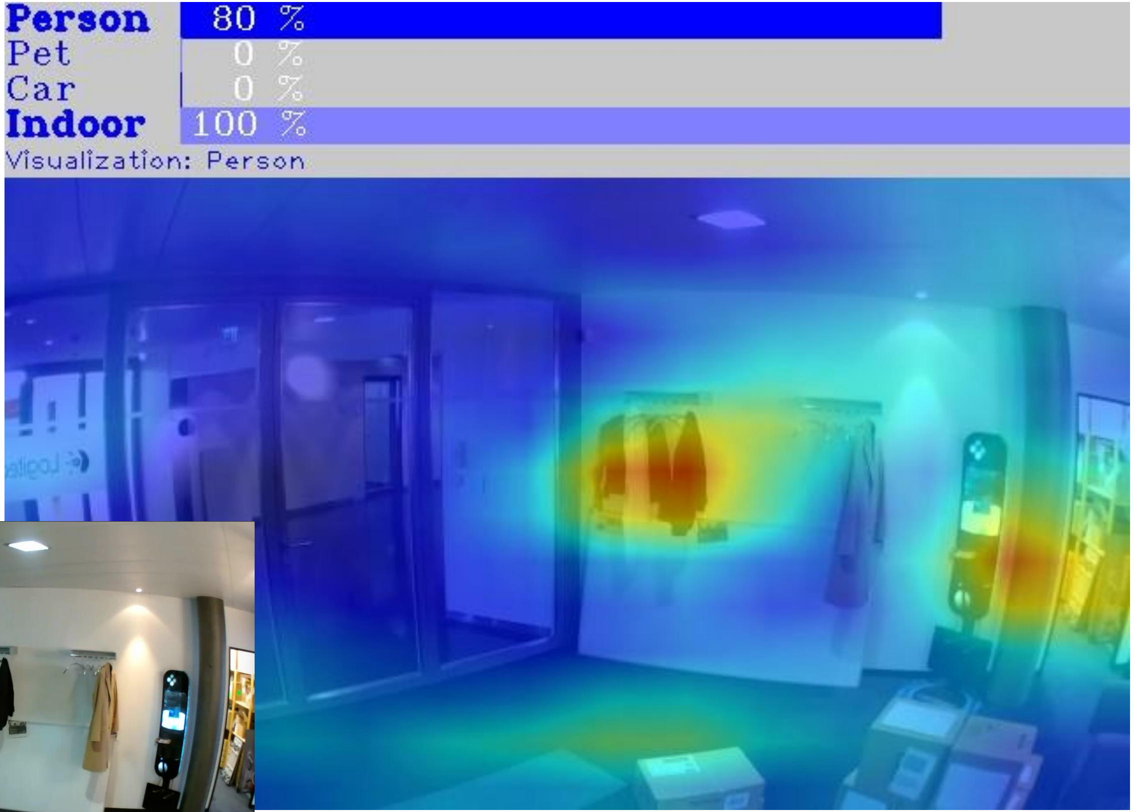
Know when it fails: Debugging Tools

Know when it fails: Debugging Tools



Zhou et al, **Learning Deep Features for Discriminative Localization**, CVPR'16, web

Know when it fails: False Positives



Know when it fails: False Positives

Context

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Recognition in Terra Incognita Sara Beery, Grant Van Horn, and Pietro Perona, link

Errors: Limitations of the Training

Fooling a Classifier

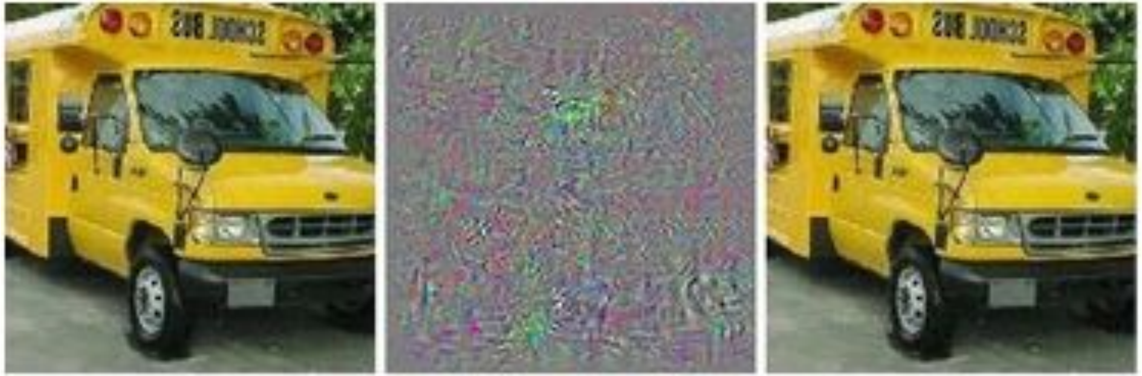
Bus

Szegedy et al., **Intriguing properties of neural networks**

Errors: Limitations of the Training

Fooling a Classifier

Bus | Noise | Bus+Noise



Szegedy et al., **Intriguing properties of neural networks**

Errors: Limitations of the Training

Fooling a Classifier

Bus | Noise | Bus+Noise | Ostrich

Szegedy et al., **Intriguing properties of neural networks**

Brief History

1943 McCullough and Pitts, linear threshold units can compute logical functions

1949 Hebb, learning rule with some physiological plausibility

1950 Turing, Computing and Intelligence

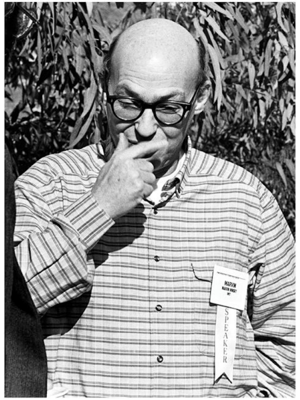
1958 **Rosenblatt**, Perceptron (single layer NN)

1960 Widrow and Hoff, 1st practical system

1968 2001: A space odyssey

1969 **Minsky** and Papert, Casting doubt on viability of Neural Networks

1st winter



Brief History

1st winter

1980er **Hinton** (among others), Backpropagation

1990er Fukushima, **LeCun**, CNN

1995 Vapnik and Cortes, SVM

2nd winter (in favor for other ML techniques)

2000er more compute, more data, deeper networks

2010 **Fei-Fei Li** et al, launch ImageNet challenge

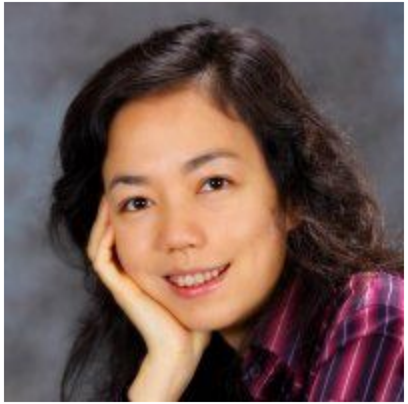
2012 Alex & **Hinton**, ImageNet challenge goes DL

Hubel and Wiesel: Visual Perception

https://knowingneurons.com/2014/10/29/hubel-and-wiesel-the-neural-basis-of-visual-perception/

Resources

Fei-Fei Li at Stanford



History of CV

Stanford, CS231n: Convolutional Neural Networks for Visual Recognition

How we teach computers to understand pictures

TED Talk