Computer Vision: a Timeline

anguelos.nicolaou@uni-graz.at, nicolas.renet@uni-graz.at

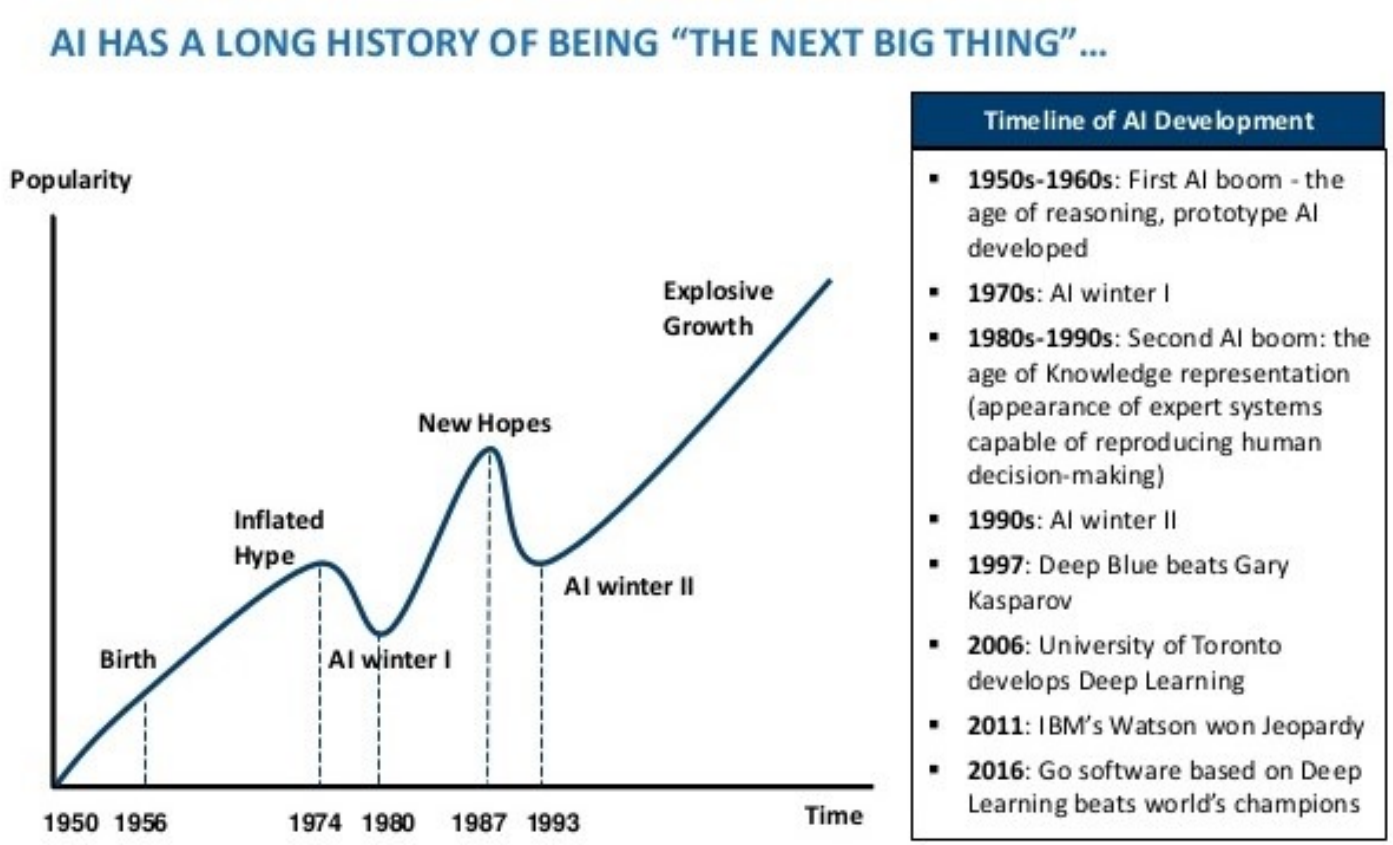
February 2023

1. Simulating vision: the highs and lows of AI

2. Convolutional networks
   - Early developments
   - Efficient ConvNets
   - Around/with the ConvNets: extracting explicit features (2000s)
   - ConvNets: quiet, but significant progress

3. Deep learning revolution

The Next Big Thing?



(Credit: Lim, `https://www.actuaries.digital/2018/09/05/history-of-ai-winters/`)

The AI Winters

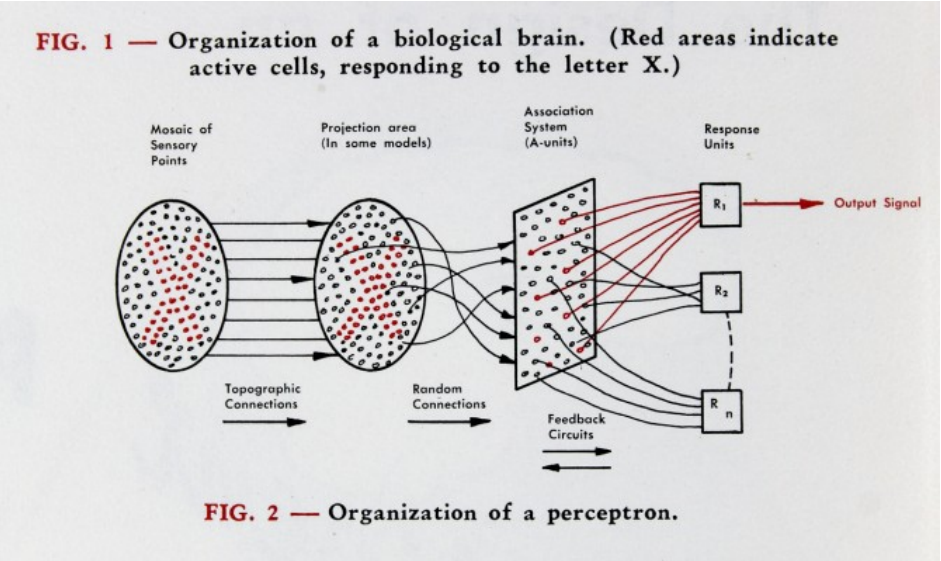
Early 70s	The first connected system (Perceptron) disappoints; after that, interests shift to symbolical, knowledge-based approach to AI
1987-1993	Demise of the LISP market and expert systems.

"Winters" are overstated!

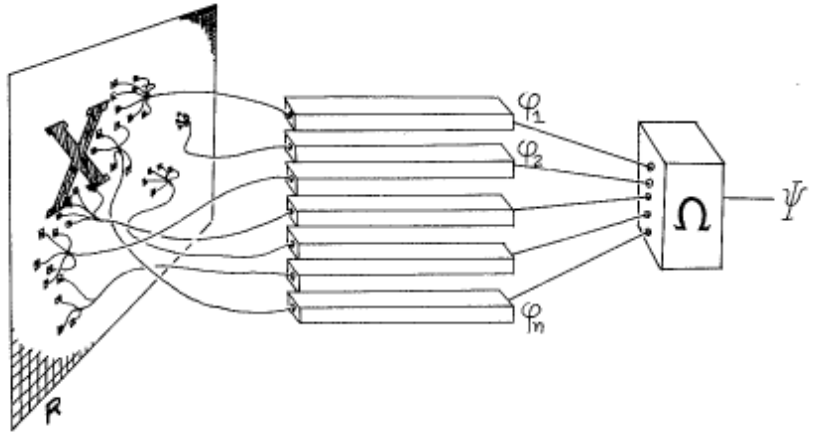
- Some systems are good enough for their specific task, allowing them to survive (in big corporations, mostly)
- New directions are explored quietly, or earlier developments are object of renewed interest $\rightarrow$ neural networks in the late 90s

Perceptron (Rosenblatt, 1958)

Single-layer neural simulation



(a) Modelling the biological brain (Rosenblatt, 1958)



(b) Parallel computing of local patterns

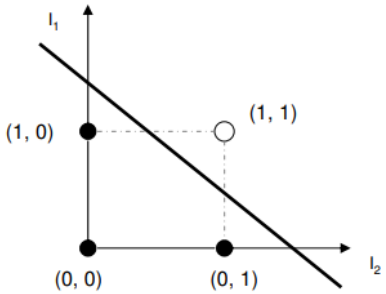
(Credit: Minsky & Papert, 1969)

The Perceptron cannot solve XOR

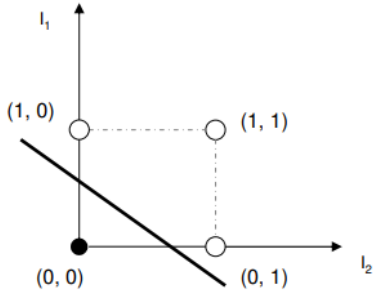
Minsky's critique (The Perceptron, 1969, 1976)

- The Perceptron cannot solve the XOR problem.

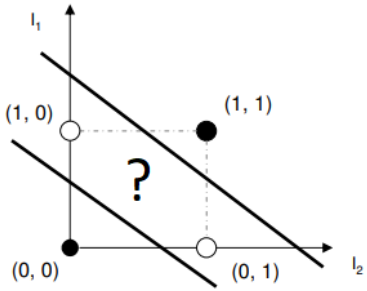
AND		
I_1	I_2	out
0	0	0
0	1	0
1	0	0
1	1	1



OR		
I_1	I_2	out
0	0	0
0	1	1
1	0	1
1	1	1



XOR		
I_1	I_2	out
0	0	0
0	1	1
1	0	1
1	1	0



(Kevin Swingler + Lucas Araujo)

- A call for action, not a death sentence!
- Connectionist paradigm fades nonetheless.
- Illustrates the two flavors of CV research: engineering feats vs. conceptual proof (why does it work?)

AlexNet (2012)

AlexNet (U. Toronto):

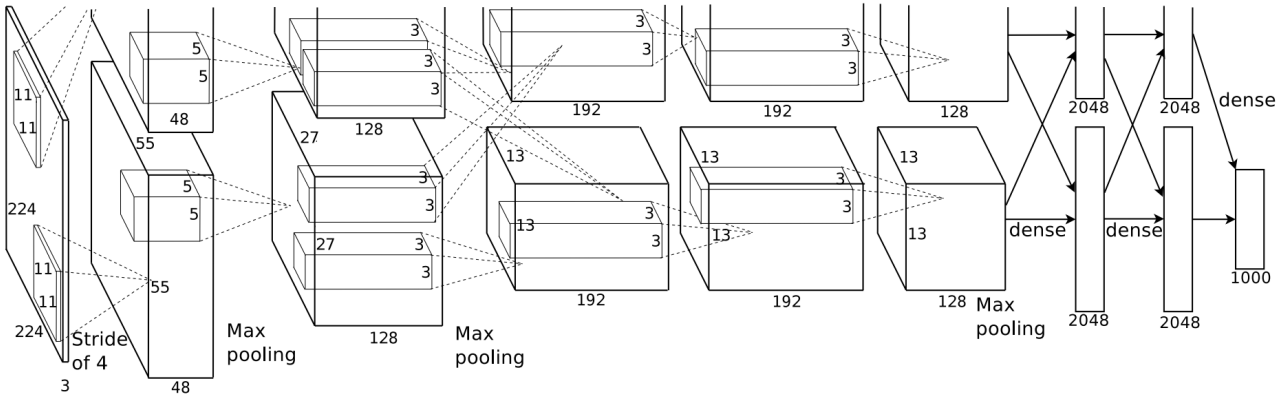
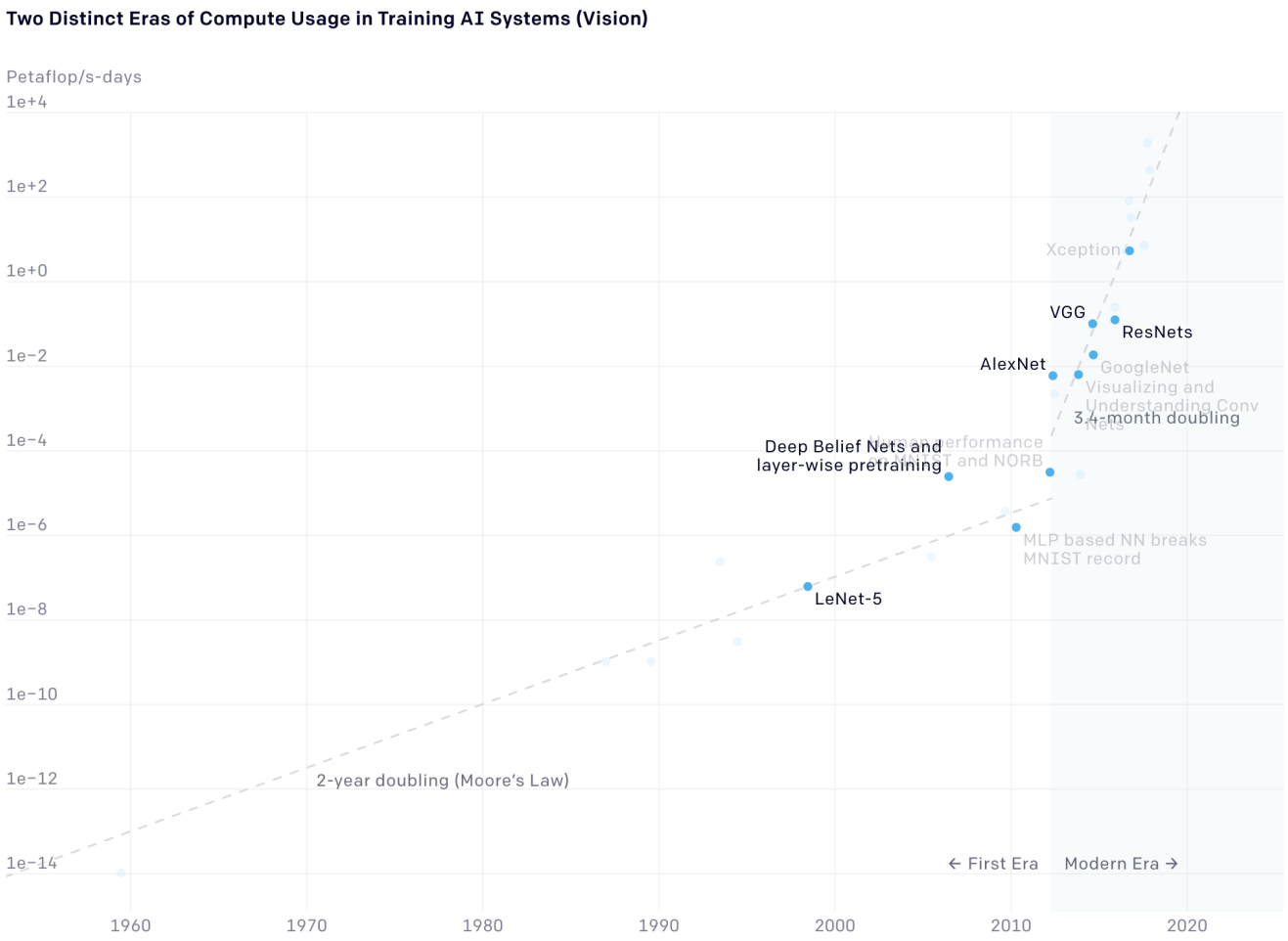


Figure: (Krizhevksy et al., 2012)

- Wins the "ImageNet Large Scale Visual Recognition Challenge" for a wide margin (1.2 millions 3-channel images, to be classified in 1000 categories).
- Most of what has happened in the last 10 years builds upon this milestone.

Is the next AI winter around the corner?

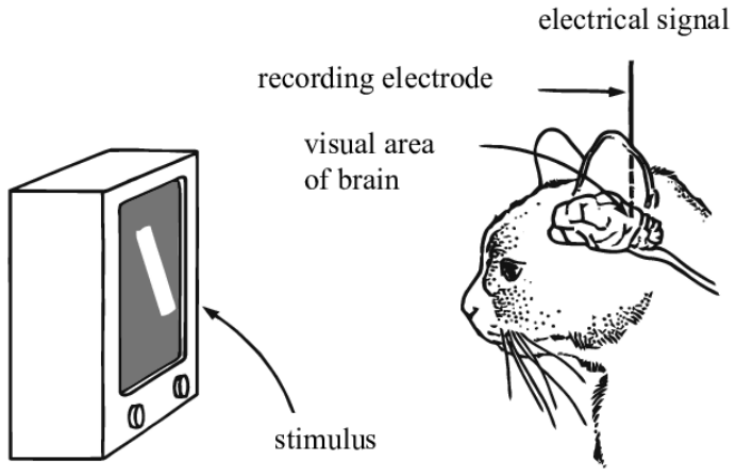
Computing for AI



(Credit: OpenAI, https://openai.com/blog/ai-and-compute/)

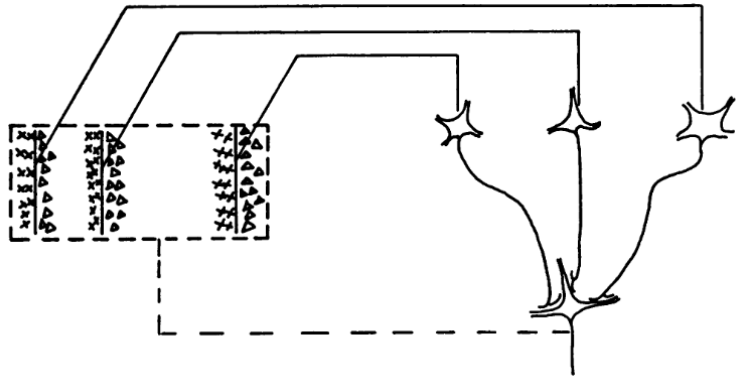
The cat's eye: the biological model

D. Hubel & T. Wiesel, *Journal of Physiology*, 1962



Hypothesis of a hierarchy of neuronal connections:

- simple cells capture elementary features
- complex, higher-level cells synthesize composite patterns



(Credit: Hubel, 1962)

Biology vs. Computer vision

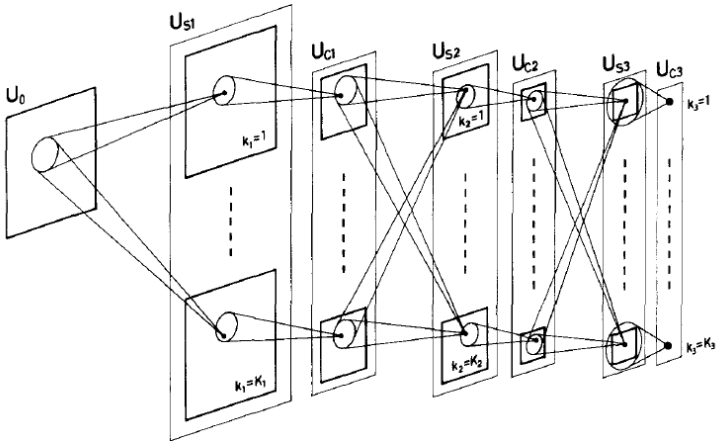
Hubel's model has been considerably revised in the following decades.

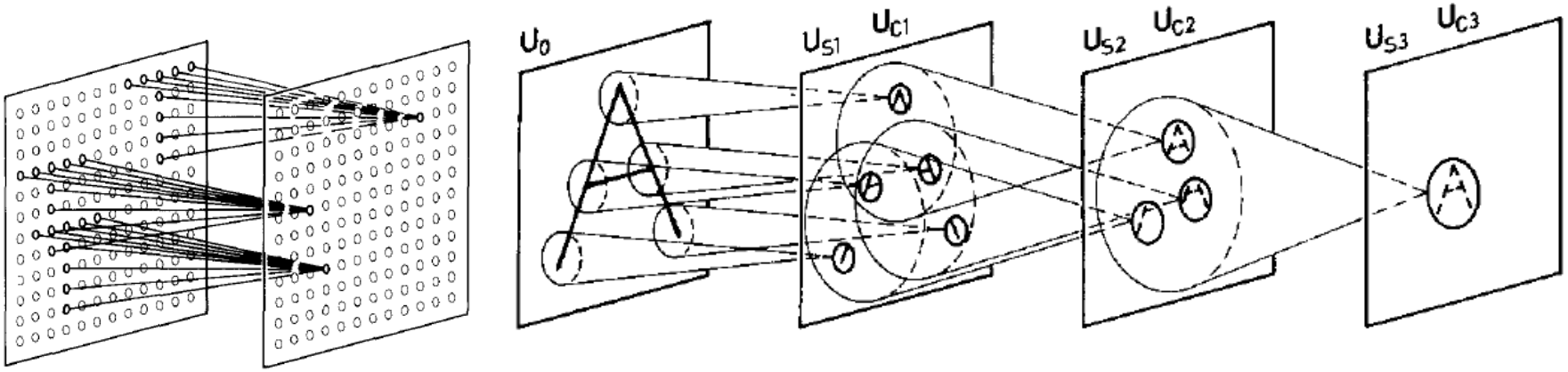
- widely influential model: a template for most connected systems in the following 50 years
- from there on, computer vision research by and large follows a track of its own, without looking back at recent developments in neurology

Neo-Cognitron (Fukushima, 1975-1985)

A 2-tier system, refined over 15 years:

- layer-wise, unsupervised learning for the filter bank
- separately-trained supervised linear classifier for the output layer



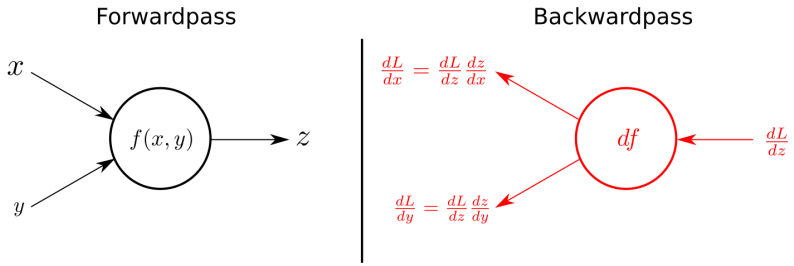


(Credit: Fukushima, 1980)

LeNet (U. of Toronto, 1986-1989)

Yann LeCun and Yoshua Bengio come up with an efficient implementation of ConvNets, which dispenses with the 2-tier architecture.

Geoffrey Hinton (U. of Toronto) had shown how to apply back-propagation on multilayer NNs:



Error propagated to prior neurons with dynamic algorithm, in each layer of the network $\rightarrow$ computational cost decreases by an order of magnitude.

LeNet-5 (1998)

LeNet (LeCun, Bottou, Bengio) trains a 6-layer convolutional neural network on gray-scale digits:

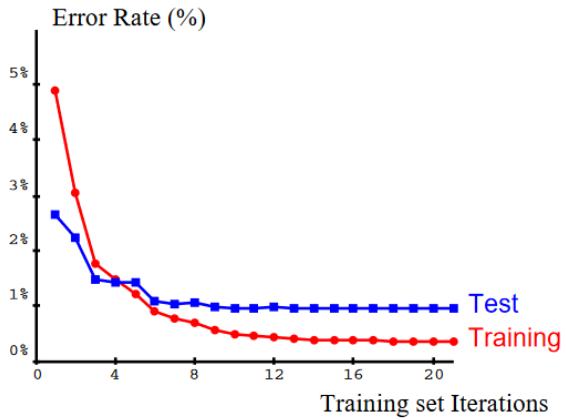


Fig. 5. Training and test error of LeNet-5 as a function of the number of passes through the 60,000 pattern training set (without distortions). The average training error is measured on-the-fly as training proceeds. This explains why the training error appears to be larger than the test error. Convergence is attained after 10 to 12 passes through the training set.

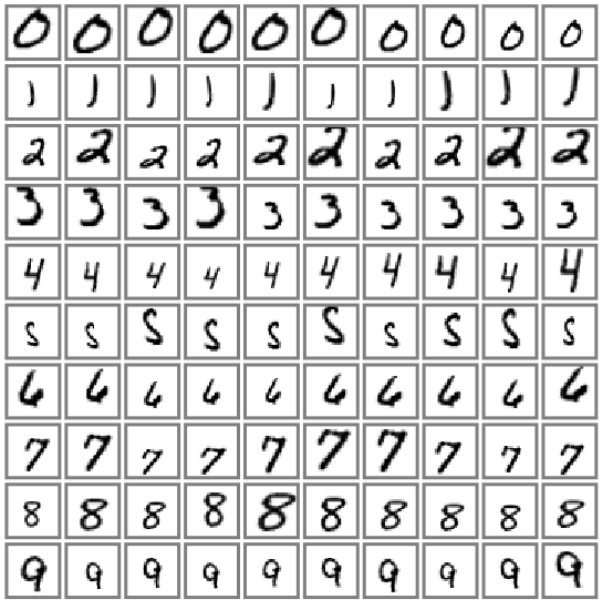
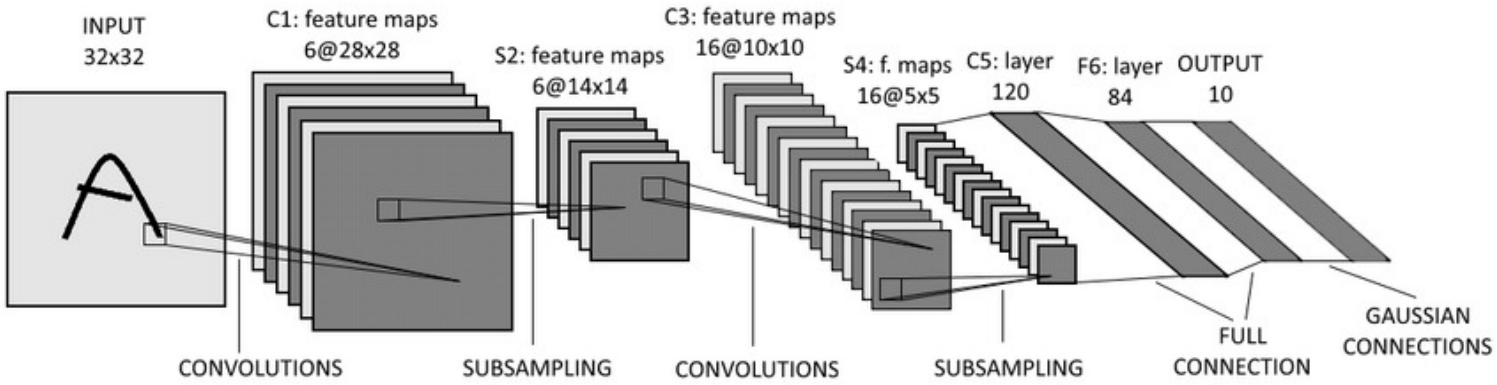


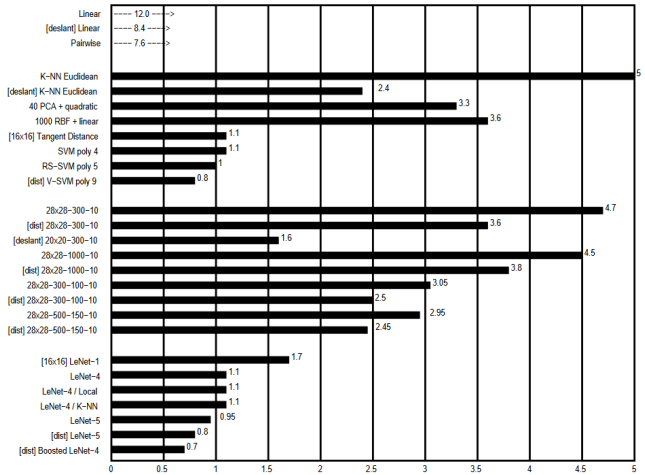
Fig. 7. Examples of distortions of ten training patterns.

- MNIST test data = NIST database of handwritten characters + augmented with distorsions/translations
- By late 90's, 10% of the US checks read by ConvNets (AT&T, 1995).
- >50 000 citations since

LeNet-5 (1998)



- convolution = extraction of local features across all regions of the input (weights/parameters are shared)
- subsampling = reduce the resolution
- fully-connected layer = flattening and classification



LeNet's posterity

Supervised ConvNets prove to be versatile, lending themselves to a wide range of specialized applications.
However, their performance does not place them in a category of their own. Although LeNet is certainly influential, it still leaves some room for approaches based on feature-extraction:

- model is versatile enough to be combined with feature-extraction approach
- LeNet does not start an explosive growth in NN for computer vision: training is costly (lack the computing power, but also the data)

A longtime challenger: SVM

Support Vector Machines (Vapnick et al., 1995)

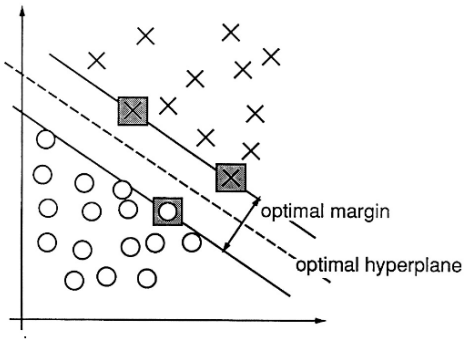


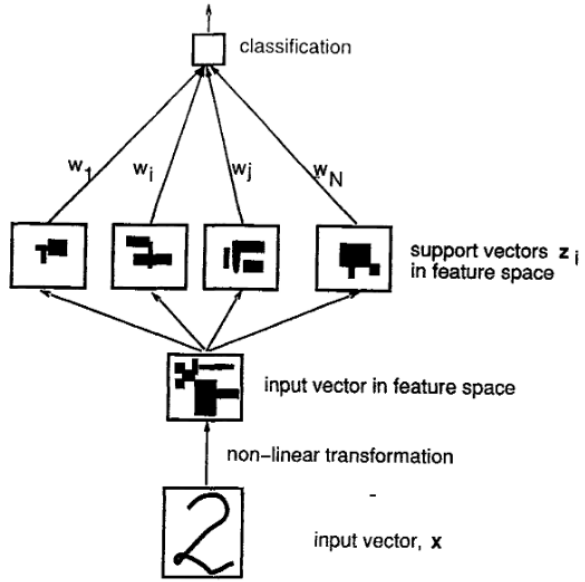
Figure 2. An example of a separable problem in a 2 dimensional space. The support vectors, marked with grey squares, define the margin of largest separation between the two classes.

(Credit: Vapnick et al., 1995)

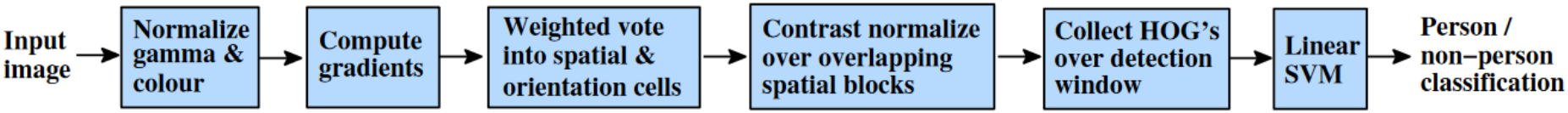
- State of the art for binary classification problems during the 2000s
- Compute an hyperplane in the vector space

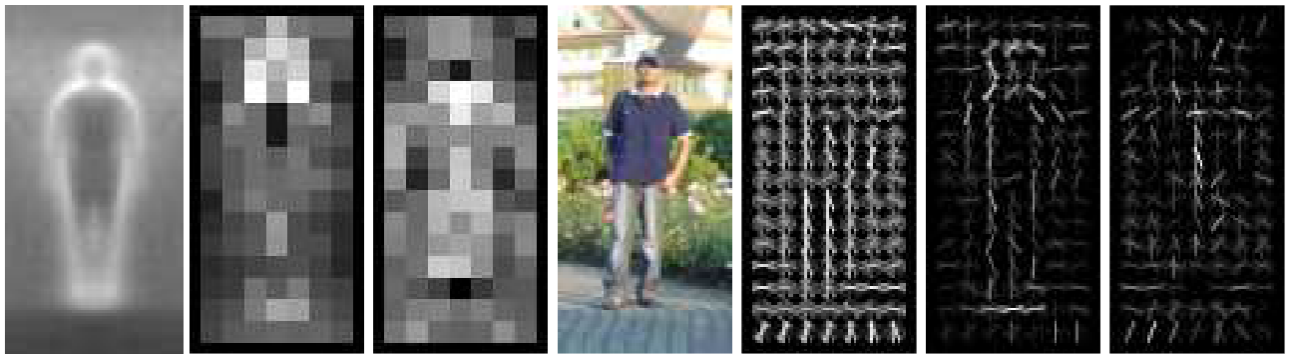
Efficient solution for non-linearly separable data:

- data points are projected into a higher-dimensional space (kernel)
- ... where a separating plane can be found through linear optimization.



Histograms of Gradient Descent (Dalal & Triggs, 2005)



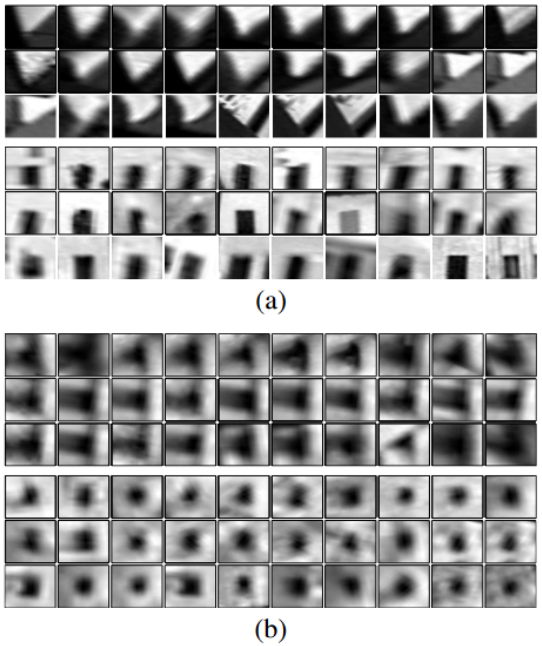


(a) (b) (c) (d) (e) (f) (g)

Video Google, 2003

Text retrieval approach to object matching in videos (Sivic & Zisserman, 2003)

1. viewpoint invariant regions
2. clusters of "visual" words
3. visual indexing
4. scene matching using visual words

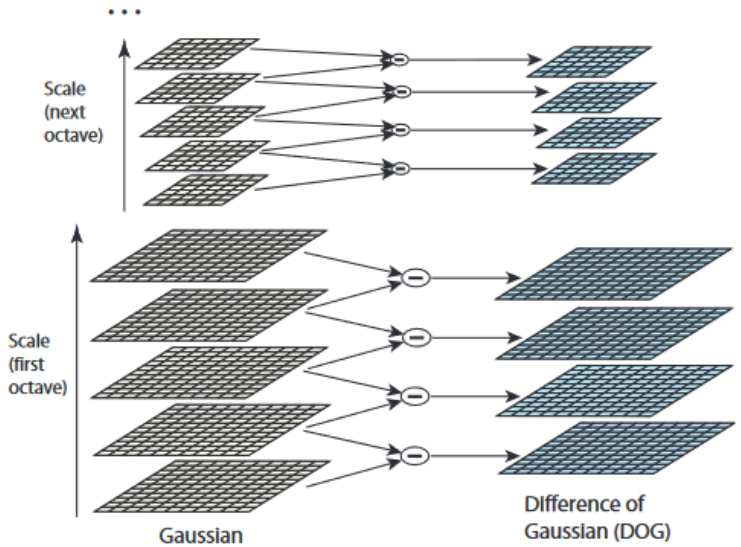


Signal processing approach (SIFT)

Scale-Invariant Feature Extraction (G. Lowe, 2004)

(Credit: G. Lowe)

1. convolution at different scale (Gaussian kernel="blurring")
2. difference of convoluted images (DoG)
3. feature matching and indexing
4. cluster identification (Hough transform)

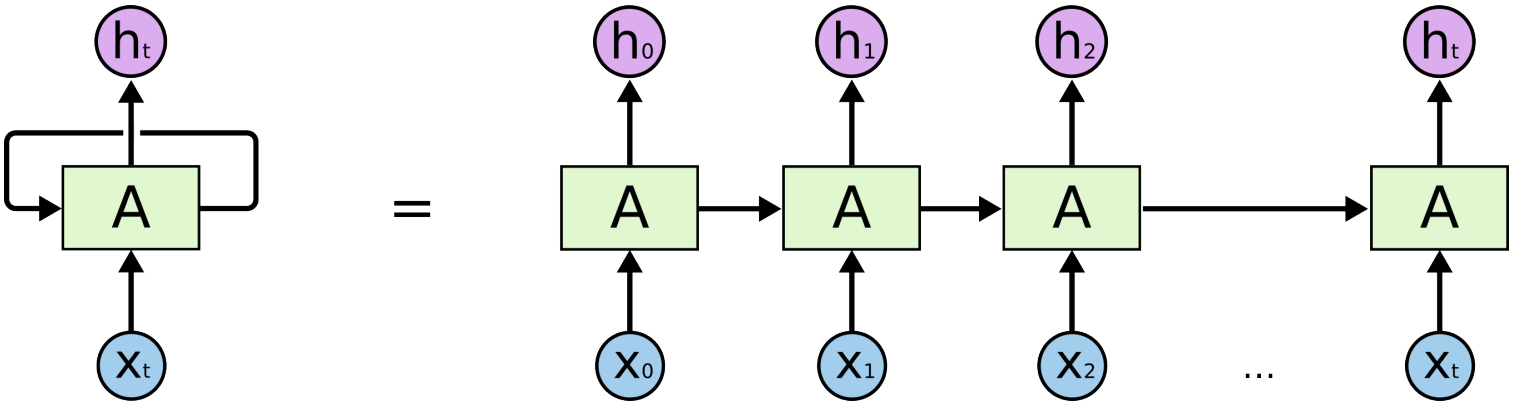


Innovation vs. Revolution

- ConvNets excelled in useful, but relatively well-defined applications (OCR, NLP) where labeled training samples existed and were large enough
- AI was not front page material...
- ... but already everywhere (Google detects license plates and faces in StreeView images; video surveillance systems in airports...)
- Significant advances are made in the early 2000s, paving the way for the recent boom

Recurrent Neural Networks (RNN)

Idea (1986): rolling the many layers of a CNN into a single, recurrent layer that calls itself (recurrence)



(Credit: https://colah.github.io/posts/2015-08-Understanding-LSTMs/)

- size of the model independent from input size
- suited to infinite sequences (speech)
- can store historical information

Issue: long-term memory (vanishing gradient)

LSTM in OCR

ICDAR 2009: Handwriting recognition competition

- University of Technology of Munich (TUM) wins convincingly with an LSTM-based architecture (RNN + "memory unit" - see Schmidhuber, 1997)
- runner-up (UT Valencia) combines probabilistic approach (Hidden Markov Model) with multilayer perceptron.

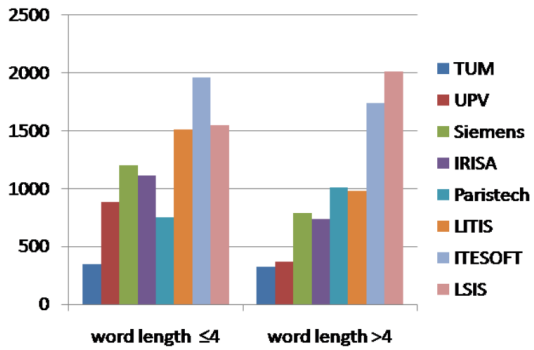


Figure 3. Number of errors vs. word length.

(Credit: Grosicki & El Abed, 2009)

Take-away:

- Neither classifier relies on character segmentation → holistic
- Establishes dominance of LSTM in OCR community (over HMM-based approaches)

Conditions for AI renaissance (1): data

Storage cost plummets, allowing for larger **training data sets**

- MNIST (1998) 60,000 training images and 10,000 testing images (grayscale).
- 80M Tiny Images (2006) 32x32 images labeled with nouns (from Wordnet)
- Pascal VOC challenges (2005-2012) 11530 annotated images
- ImageNet 14M annotated 3-channel images

Conditions for AI renaissance (2): computing power

Computing **power**: fast and programmable general-purpose GPUs (NVIDIA CUDA capable of 1 trillion operations/seconds)

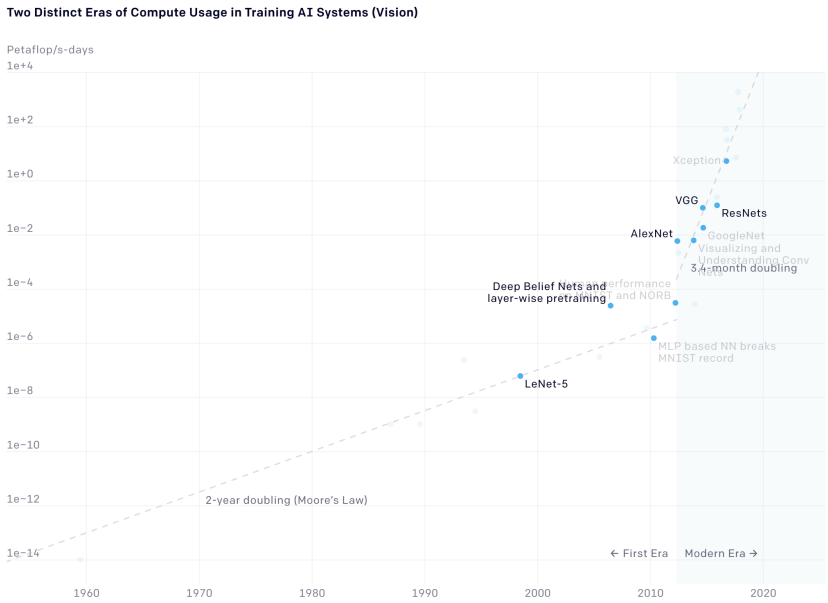


Figure: Computing use by NNs over time (Credit: OpenAI, https://openai.com/blog/ai-and-compute/)

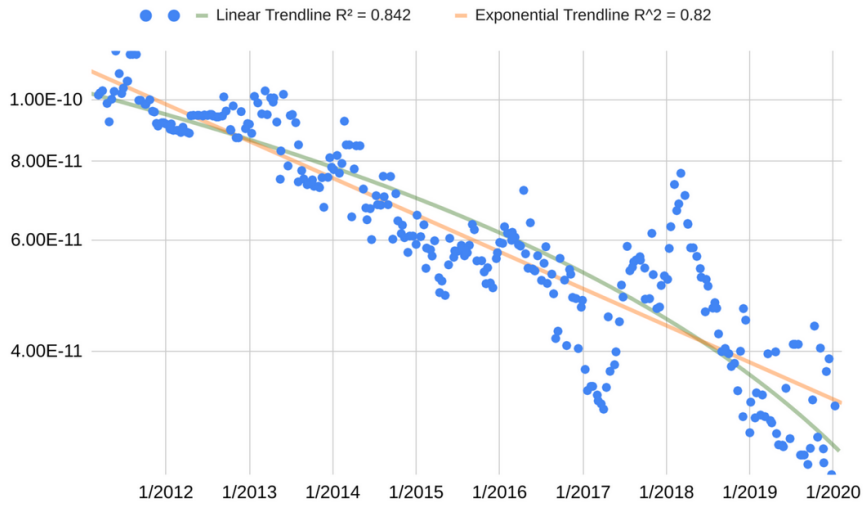


Figure: Price/performance (FLOPS/s)

(Credit: https://aiimpacts.org, 2019)

This prepare the stage for...

AlexNet (2012)

A jump in size, both in the training data and model:

	Parameters	Connections
LeNet5	∼50M	∼ 220 000
AlexNet	∼60M	∼600M

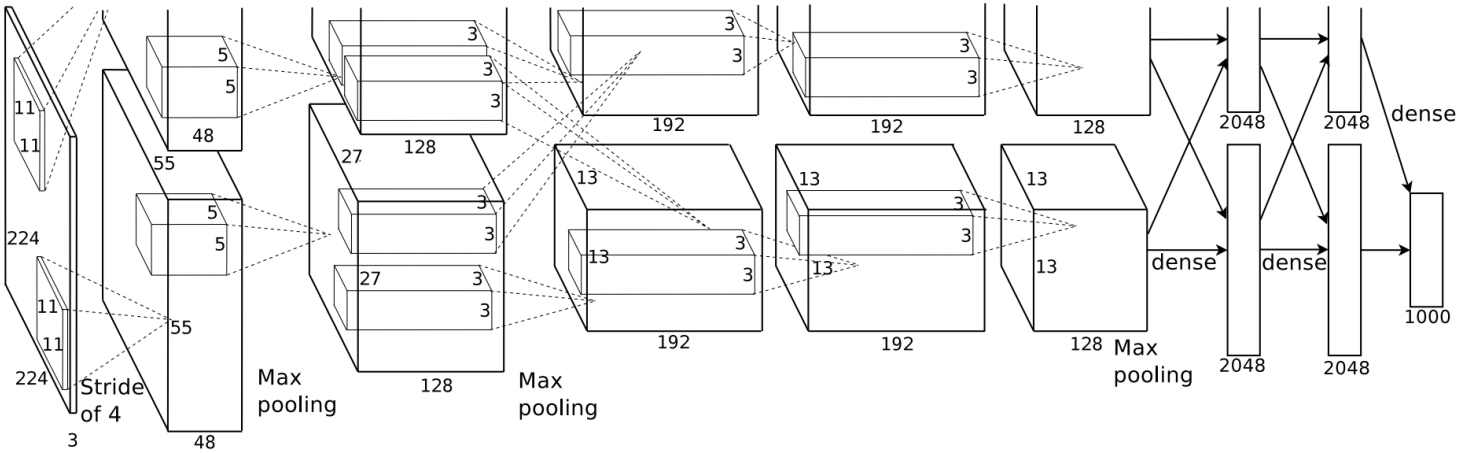
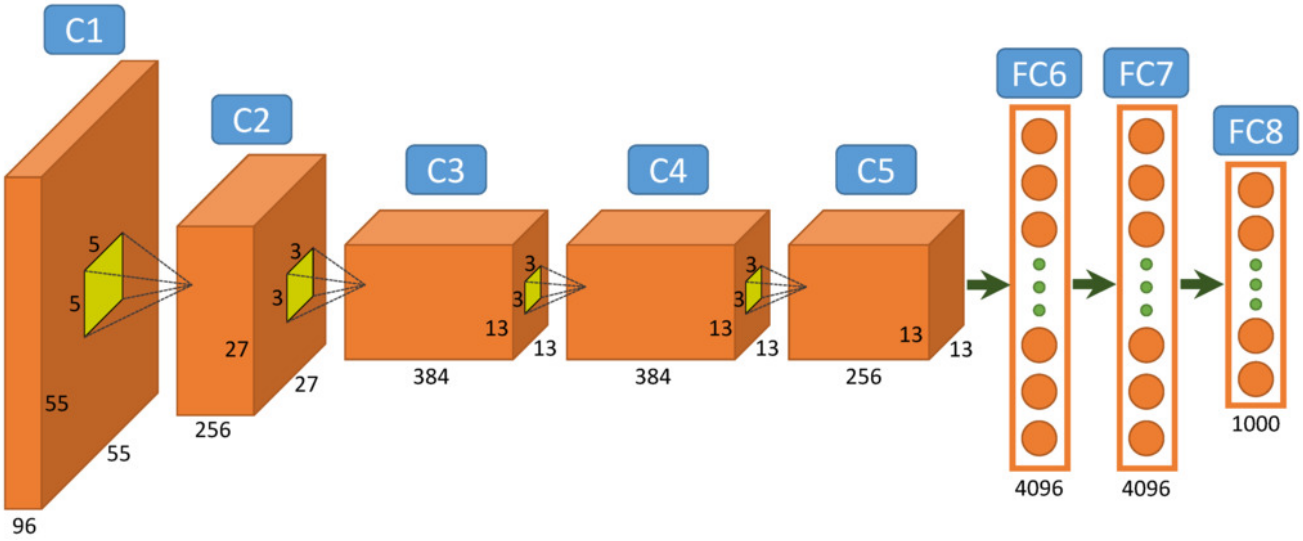


Figure: LeNet-5 vs. AlexNet (LeCun et al. 1998, Krizhevsky et al. 2012)

AlexNet (2012)



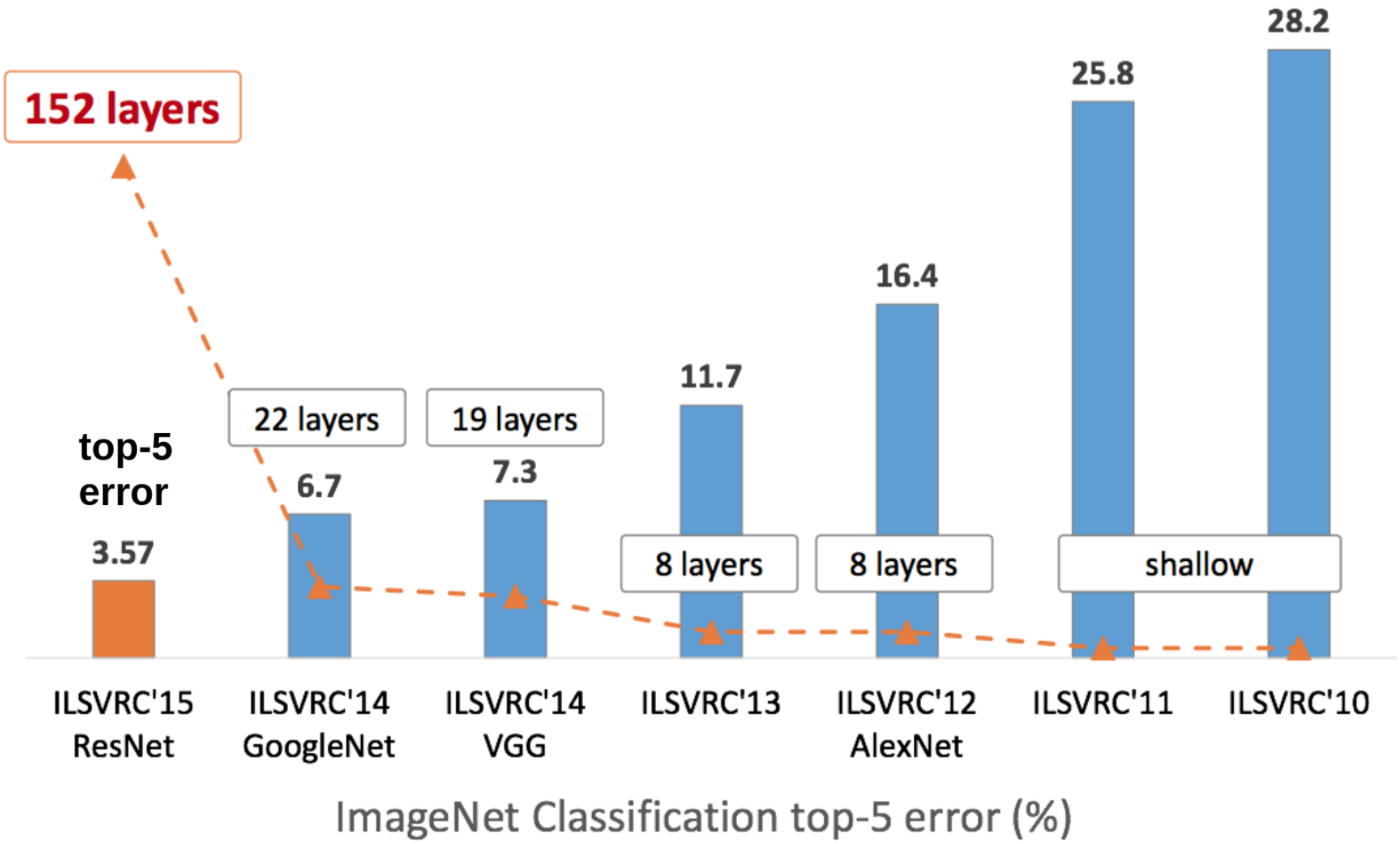
(Pablo Ceres, UW-Madison)

Innovations:

- new activation functions (ReLU) allows for large gain in training time
- normalization layer = reduce overfitting
- dropout layer (probabilistic zeroing of a neuron's input)
- training data augmented by artificial translations and variations of the existing examples

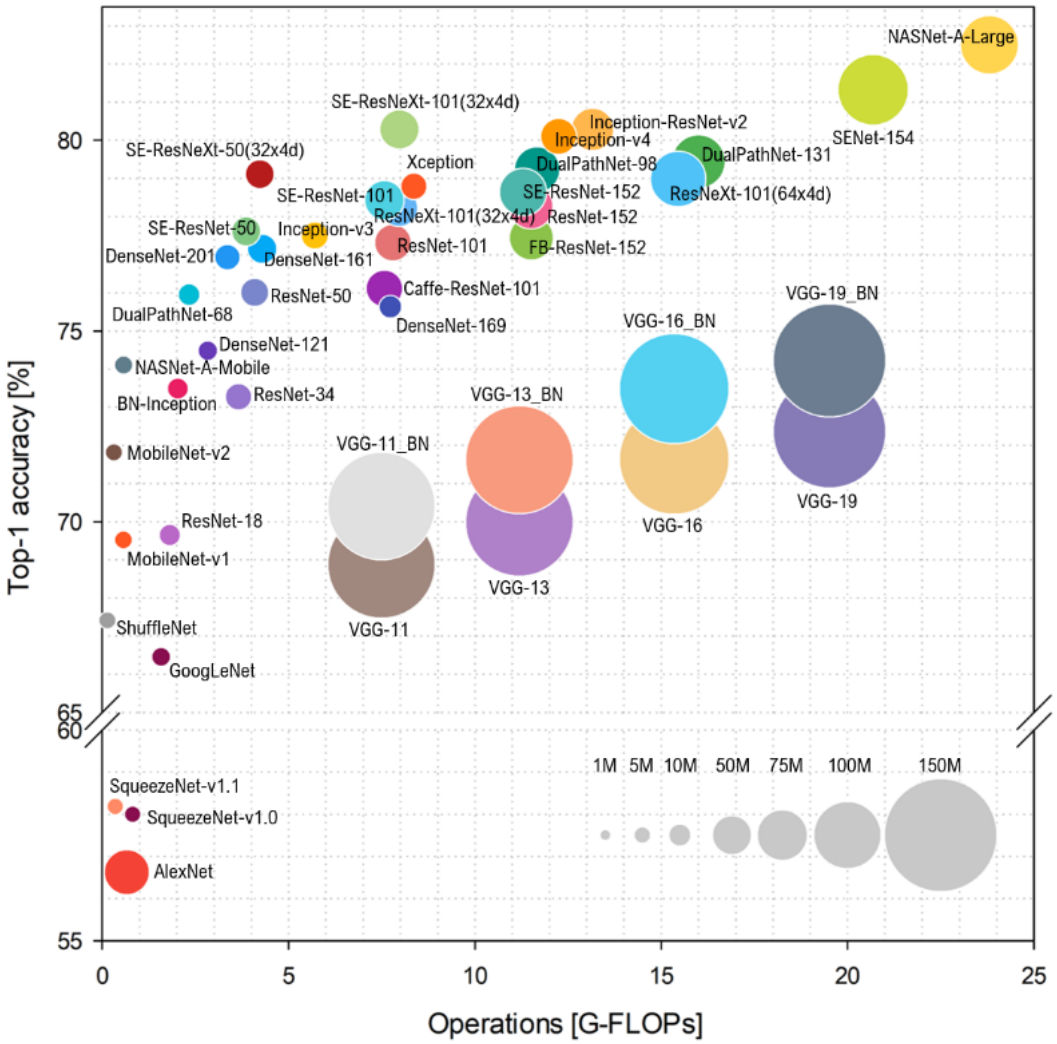
The past decade: ever deeper

An arm race, with shattered records:



(Credit: K. He)

Leaner, better networks

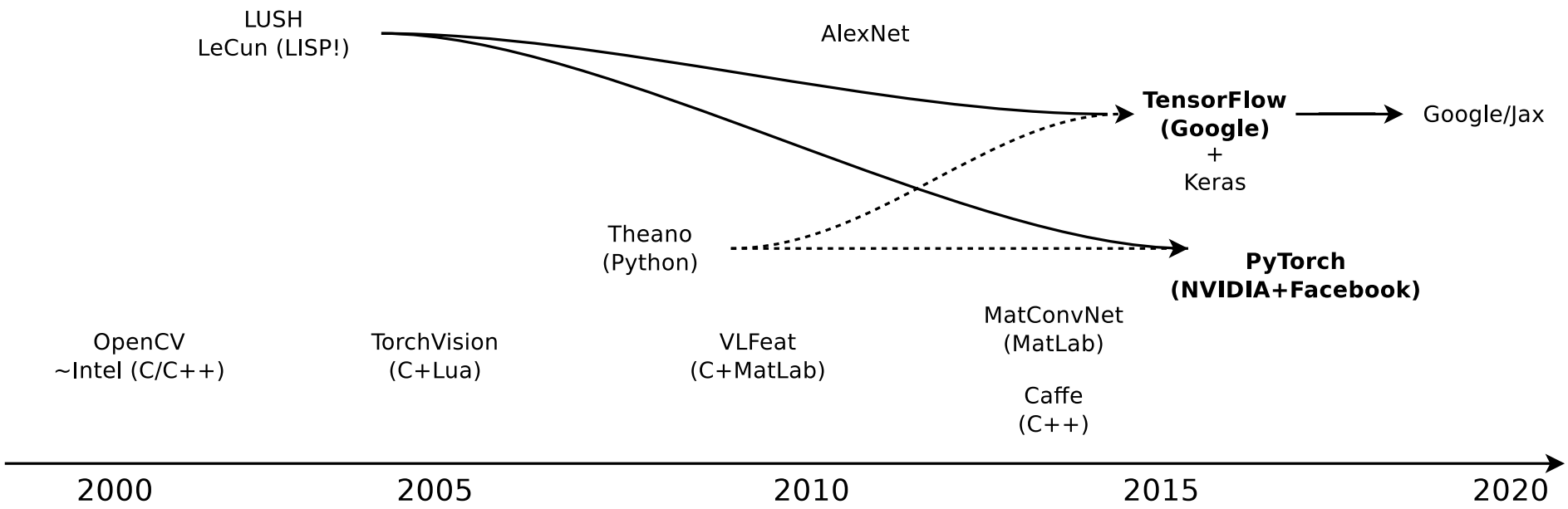


(Credit: Bianco et al., 2018)

Deeper is better. How do we achieve it?

- 1×1 convolution filters (GoogleNet, 2014)
- skipping connections (ResNet, DenseNet)

Deep Learning Frameworks



Trouble in paradise

- Uses? Medical imagery vs. facial recognition
- Bias baked in the training data? 80 Million Tiny Images (2006) as a case study $\rightarrow$ retired in 2020
- Everyone is an artist? Generative Adversarial Networks (GAN - Goodfellow, 2014) vs. protected contents