JÜRGEN SCHMIDHUBER, THE SWISS AI LAB IDSIA (USI & SUPSI)

JÜRGEN SCHMIDHUBER
YOU_AGAIN SHMIDHOOBUH

DEEP LEARNING IS A HALF CENTURY OLD
(RECENT "TABLOID SCIENCE" STORIES CLAIM IT IS A RECENT THING)

888 references, 88 pages: http://www.idsia.ch/~juergen/deep-learning-overview.html

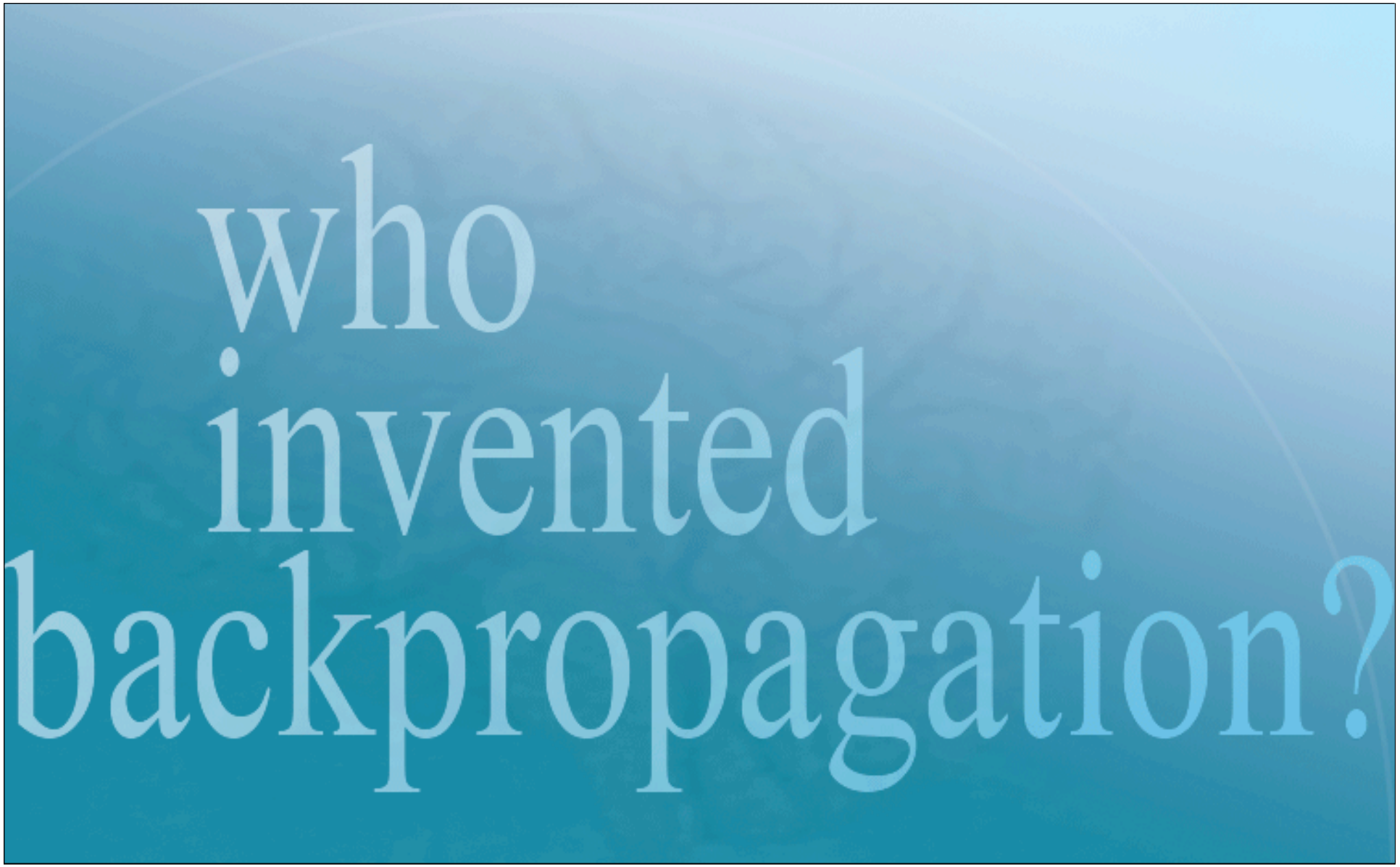
FIRST DEEP LEARNING: IVAKHNENKO, 1965 -

Deep multilayer perceptrons with polynomial activation functions

incremental layer-wise training by regression analysis - learn numbers of layers and units per layer - prune superfluous units

numerous applications

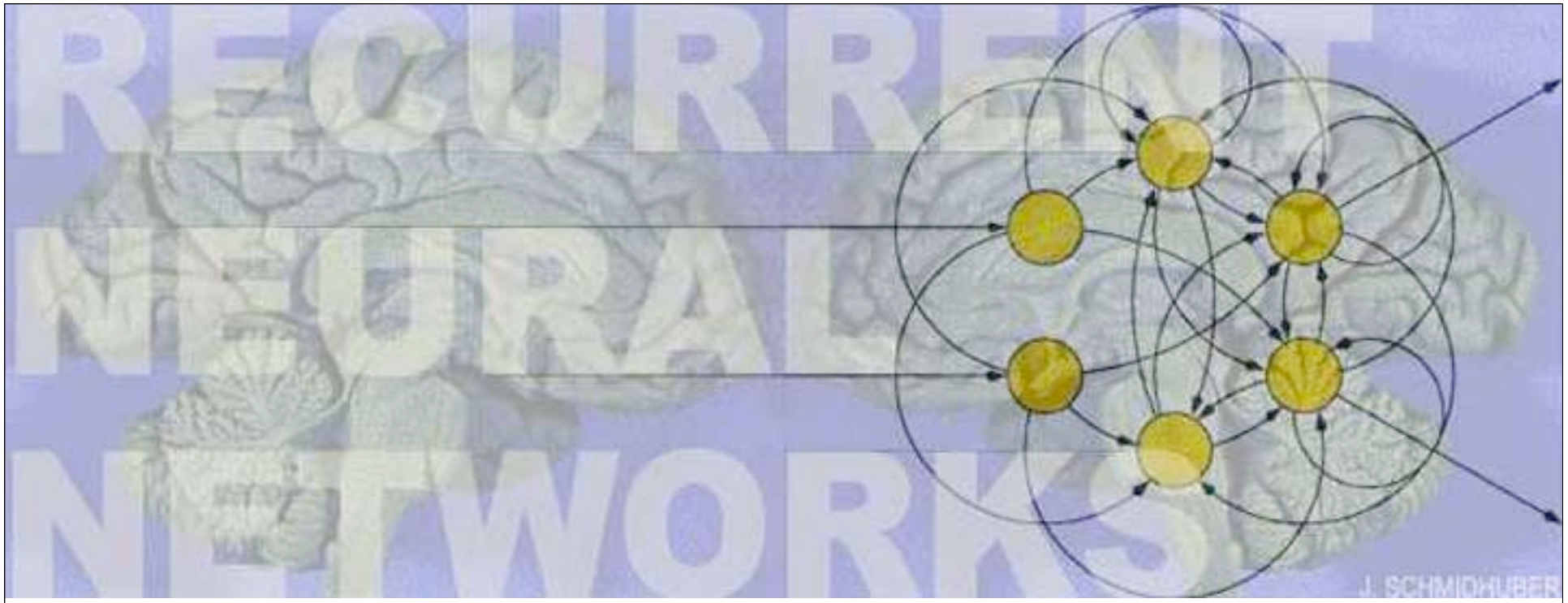
8 layers already back in 1971



http://www.idsia.ch/~juergen/who-invented-backpropagation.html

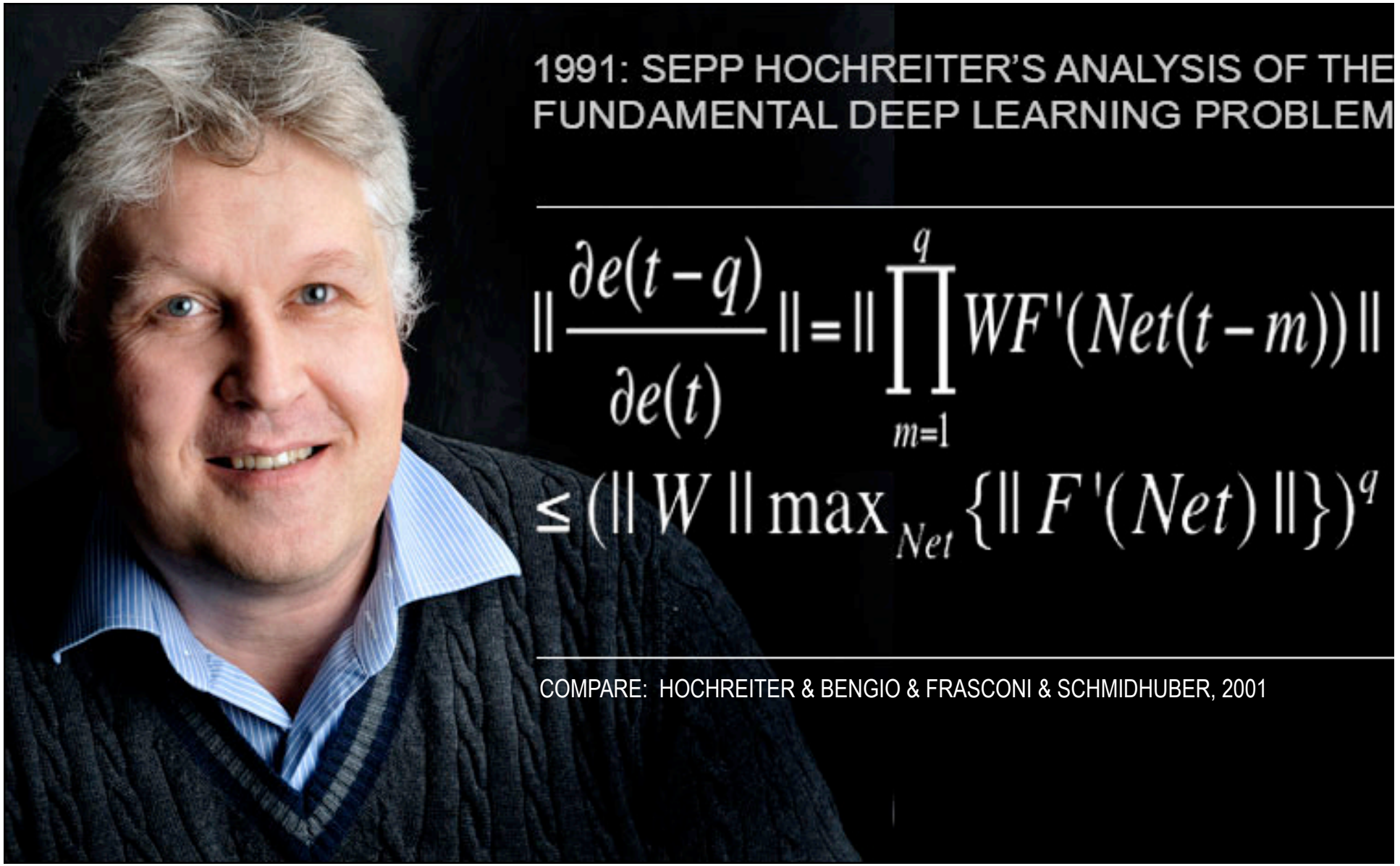
SUPERVISED BACKPROPAGATION (BP)

BP in Euler-LaGrange Calculus + Dynamic Programming: Bryson 1961, Kelley 60, Dreyfus 62, Pontryagin et al 61, Bryson & Ho 69. BP in sparse, discrete, NN-like nets: Linnainmaa 1970, Ostrovskii et al 71, Dreyfus 73, Werbos 74, Automatic Differentiation: Speelpenning 80. BP for NNs: Werbos 1981 (LeCun 85, Parker 85), Rumelhart et al 86; RNNs: e.g., Werbos 88, Williams 89; Robinson & Fallside 87...

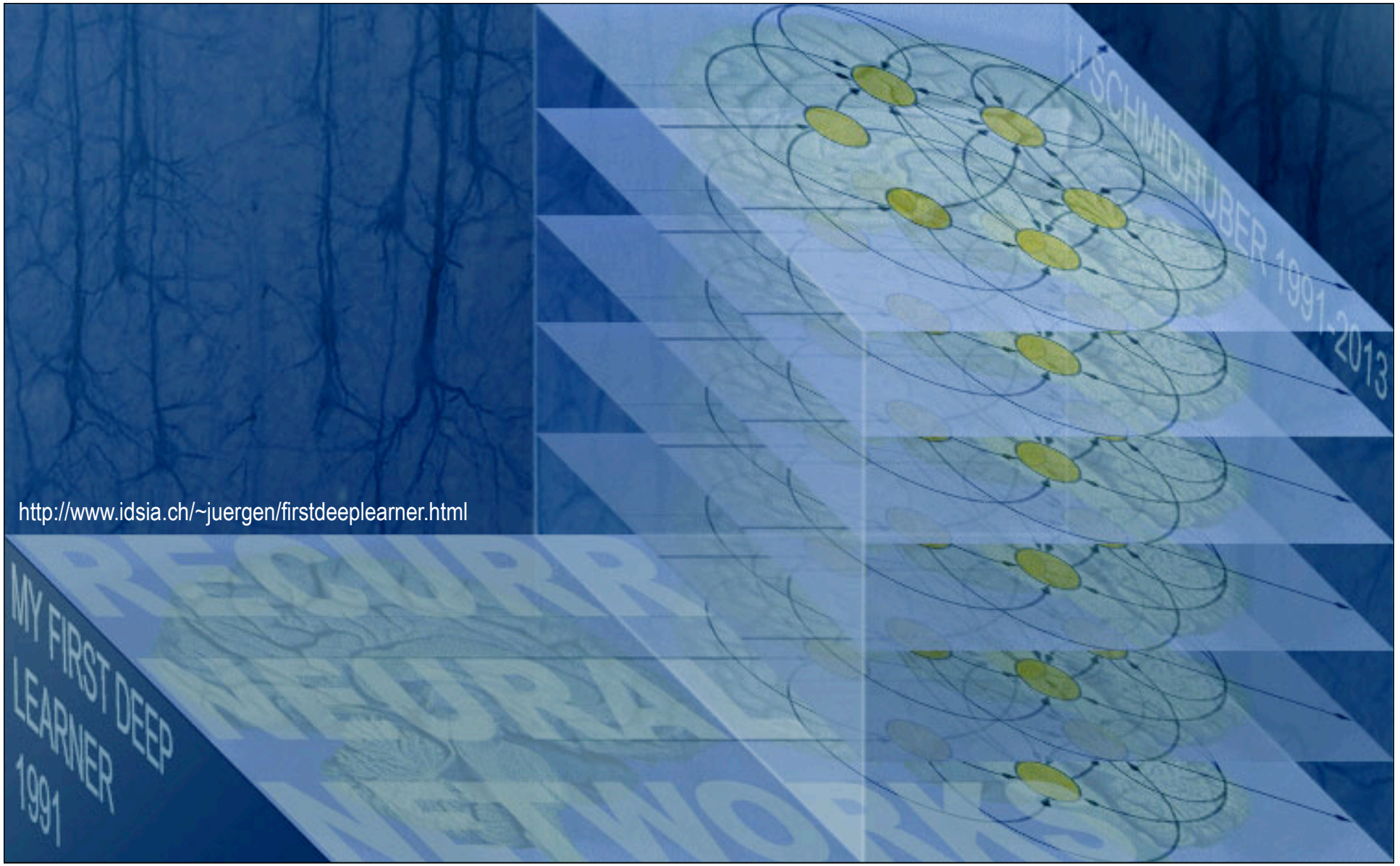


THE DEEPEST NNs:
RNNs ARE GENERAL COMPUTERS
LEARN PROGRAM = WEIGHT MATRIX

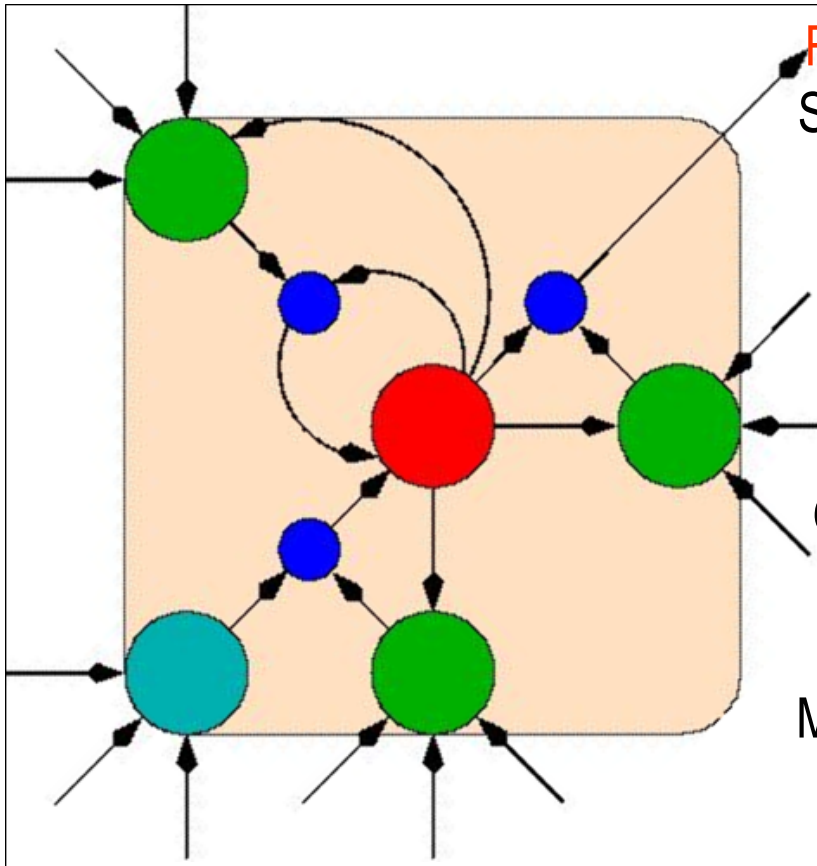
http://www.idsia.ch/~juergen/rnn.html



http://www.idsia.ch/~juergen/fundamentaldeeplearningproblem.html



Schmidhuber 1991: Unsupervised pretraining for Hierarchical Temporal Memory: stack of RNN ➔ history compression ➔speed up supervised learning. Compare feedforward NN case: AutoEncoder stacks (Ballard 1987) and Deep Belief NNs (Hinton et al 2006)



RED:LINEAR UNIT SELF-WEIGHT 1.0: TRANSPORTS ERROR ACROSS THOUSANDS OF TIME STEPS
GREEN: GATES OPEN / PROTECT ACCESS
BLUE: MULTIPLICATIONS

RNN: LONG SHORT-TERM MEMORY

LSTM: NO VANISHING GRADIENTS

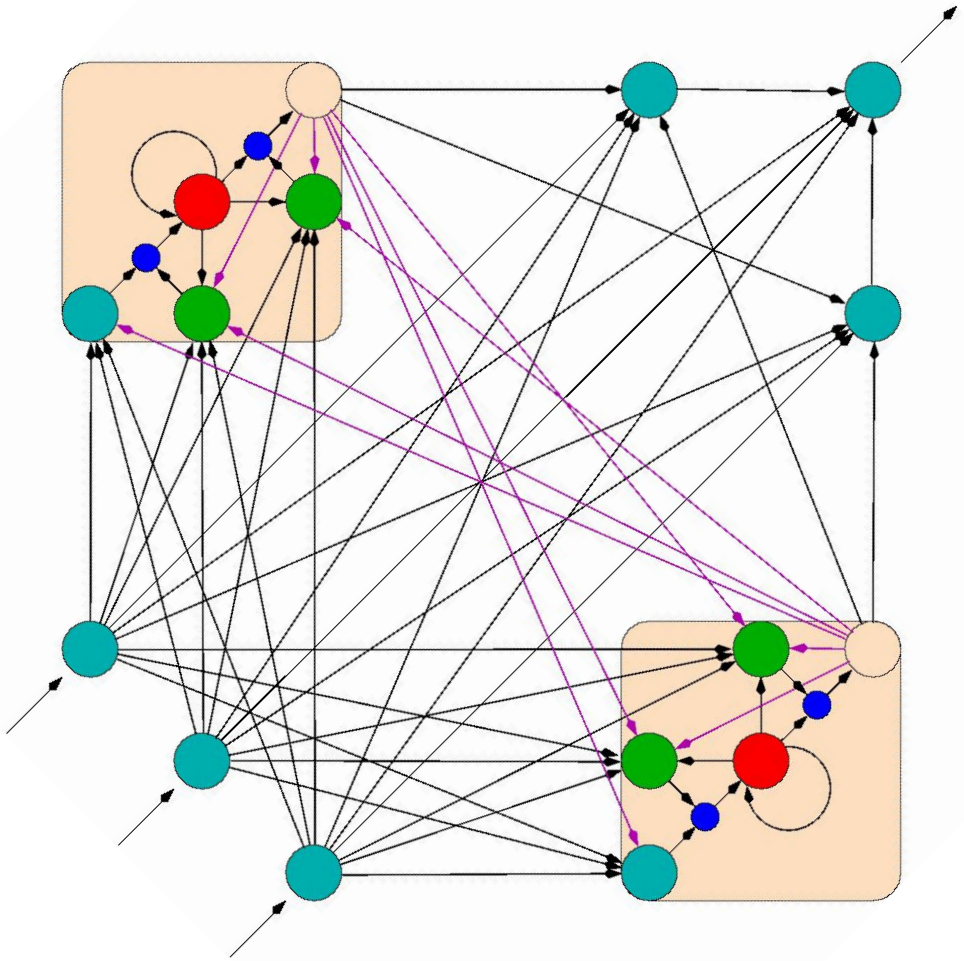
GRADIENT-BASED LSTM (1995, Neural Comp. 1997) LEARNS MANY PREVIOUSLY UNLEARNABLE DEEP LEARNING TASKS: GRAMMARS, BLUES COMPOSITION, R-LEARNING ROBOTS, METALEARNING, SPEECH RECOGNITION (>HMMS), PROTEINS, HANDWRITING. NO BIAS TOWARDS RECENT OR ANCIENT EVENTS!

http://www.idsia.ch/~juergen/rnn.html

Today's LSTM RNNs shaped by:

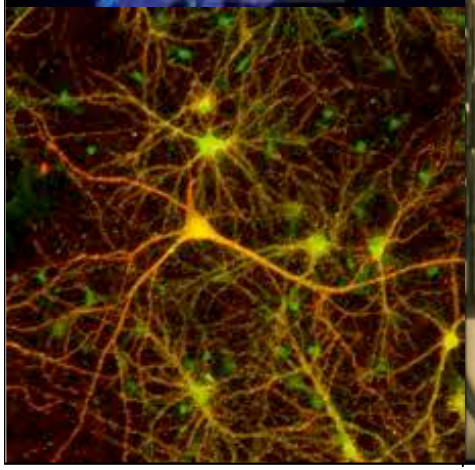
Ex-PhD students (TUM & IDSIA):
Sepp Hochreiter (PhD 1999)
Felix Gers (PhD 2001)
Alex Graves (PhD 2008)
Daan Wierstra (PhD 2010)
Others

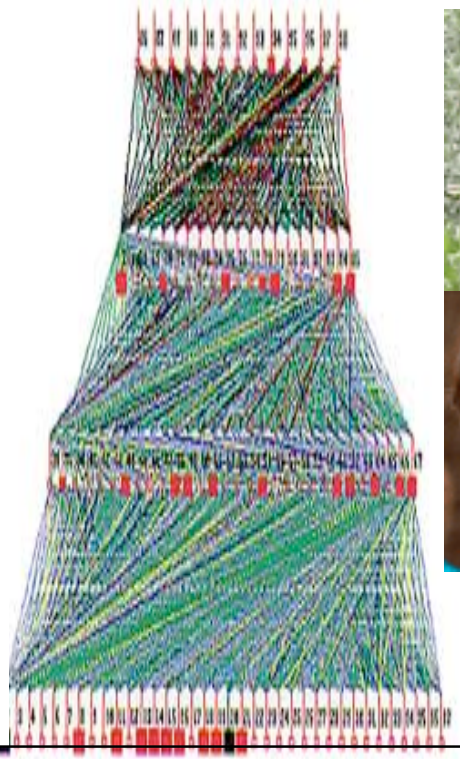
Postdocs at IDSIA (2000s):
Fred Cummins
Santiago Fernandez
Faustino Gomez
Others



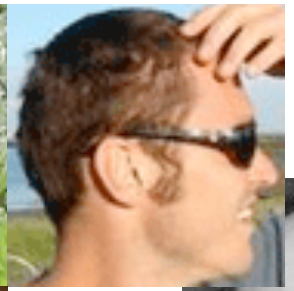
FELIX

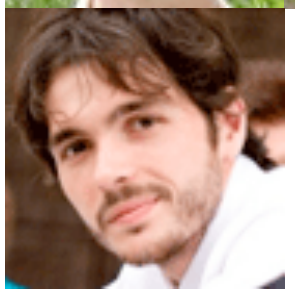
ALEX



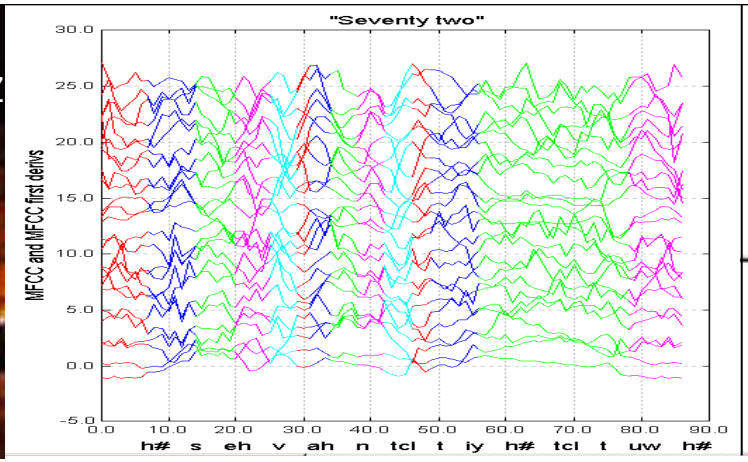


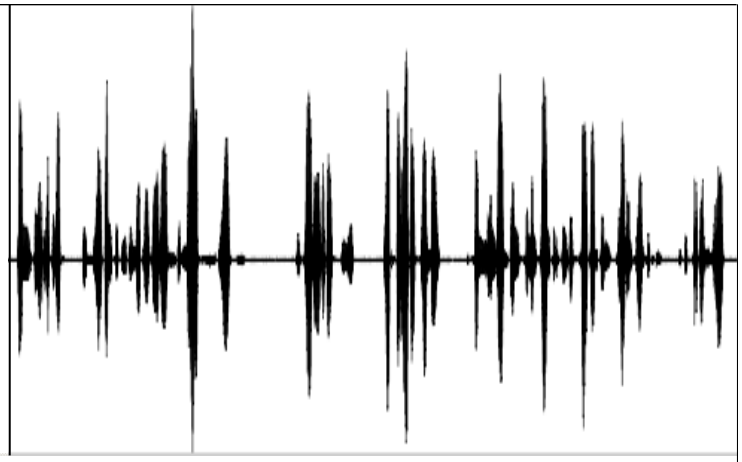
DEEP FEEDFORWARD NNs:
DAN CIRESAN,
UELI MEIER,
JONATHAN MASCI,
ALESSANDRO GIUSTI,
OTHERS

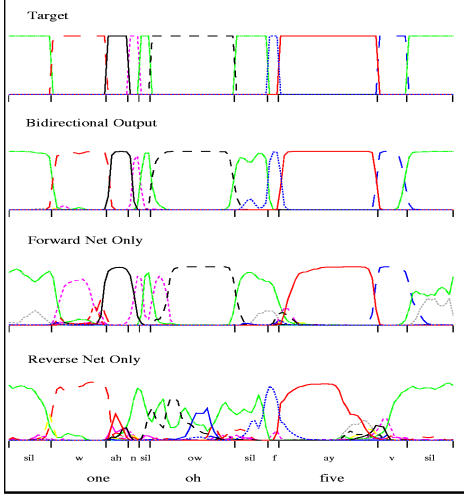


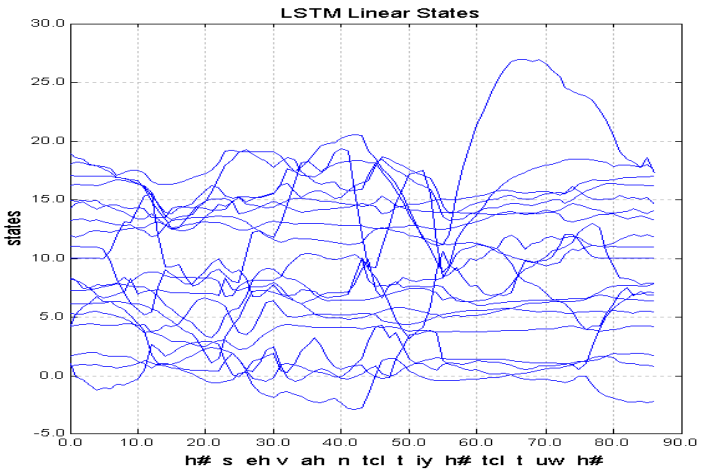


VERY DEEP LEARNING SINCE 1991
FUNDED BY DFG & SNF

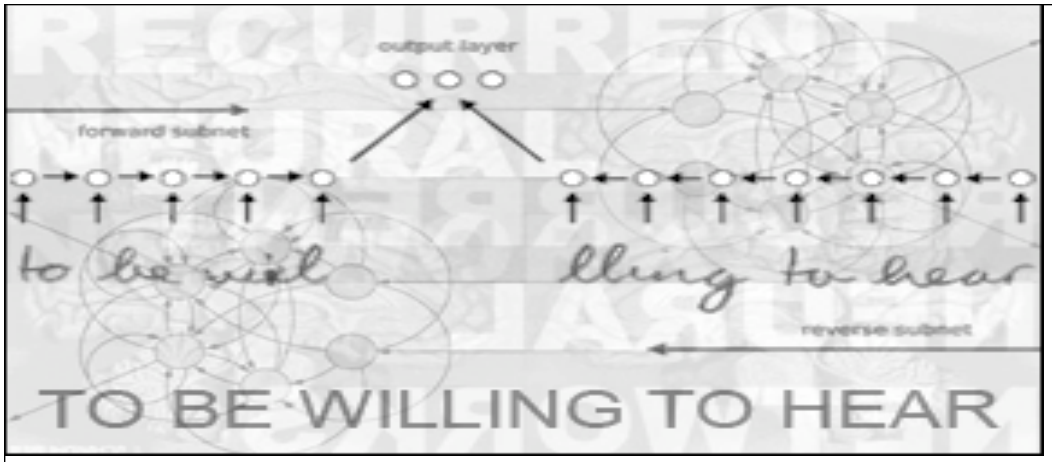






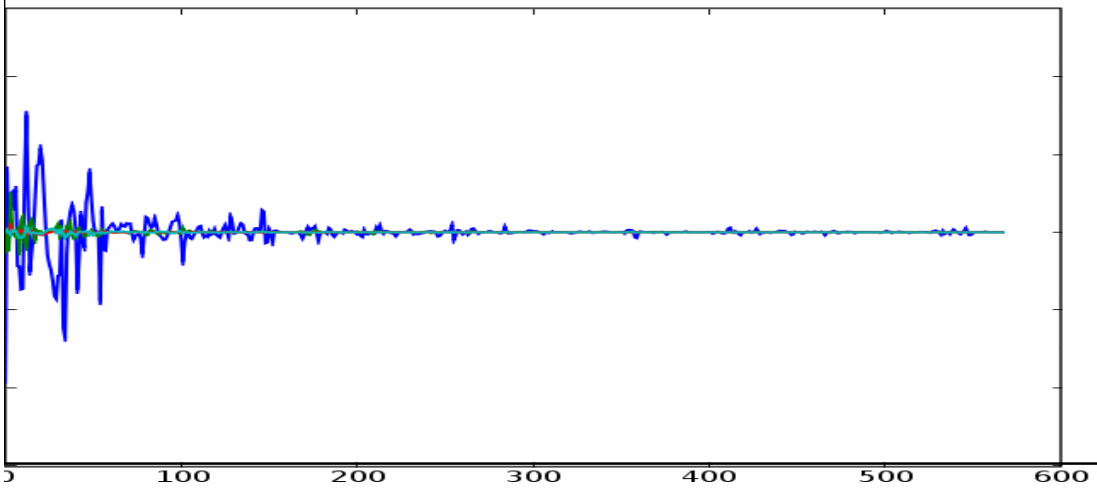


LSTM-RNN FOR SPEECH: 2003 COMPARABLE TO HMMS, IJCAI 2007: LSTM STACK GETS BEST RESULTS ON KEYWORD SPOTTING IN A LARGE CORPUS (VS HMMS)



SUPERVISED LONG SHORT-TERM MEMORY (LSTM) RNN WON 3 ICDAR 2009 CONNECTED HANDWRITING COMPETITIONS

http://www.idsia.ch/~juergen/handwriting.html



NO PRE-SEGMENTED DATA; RNN MAXIMISES PROBABILITY OF TRAINING SET LABEL SEQUENCES

$$O^{ML}(S) = -\sum_{(\mathbf{x},\mathbf{z})\in S} \ln(p(\mathbf{z}|\mathbf{x}))$$

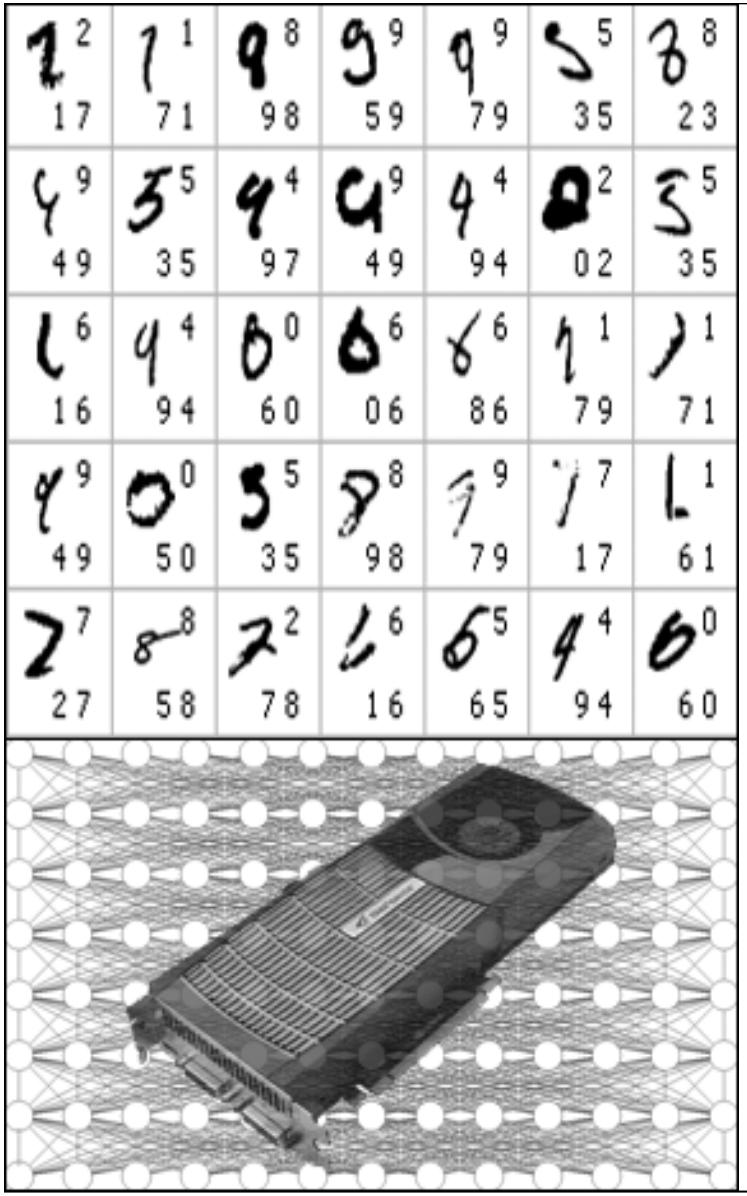
GRAVES, FERNANDEZ, GOMEZ, SCHMIDHUBER ICML 2006 - GRAVES, SCHMIDHUBER, NIPS 2010

GRAVES et al, ICASSP 2013: BEST RESULTS ON TIMIT SPEECH

2014 benchmark records with LSTM RNNs, often at major IT companies:

1. Large vocabulary speech recognition (Sak et al., Google, Interspeech 2014)
2. English to French translation (Sutskever et al., Google, NIPS 2014)
3. Text-to-speech synthesis (Fan et al., Microsoft, Interspeech 2014)
4. Prosody contour prediction (Fernandez et al., IBM, Interspeech 2014)
5. Language identification (Gonzalez-Dominguez et al., Google, Intersp. 2014)
6. Medium vocabulary speech recognition (Geiger et al., Interspeech 2014)
7. Audio onset detection (Marchi et al., ICASSP 2014)
8. Social signal classification (Brueckner & Schulter, ICASSP 2014)
9. Arabic handwriting recognition (Bluche et al., DAS 2014)
10. Image caption generation (Vinyals et al., Google, 2014)
11. Video to textual description (Donahue et al., 2014)

http://www.idsia.ch/~juergen/rnn.html

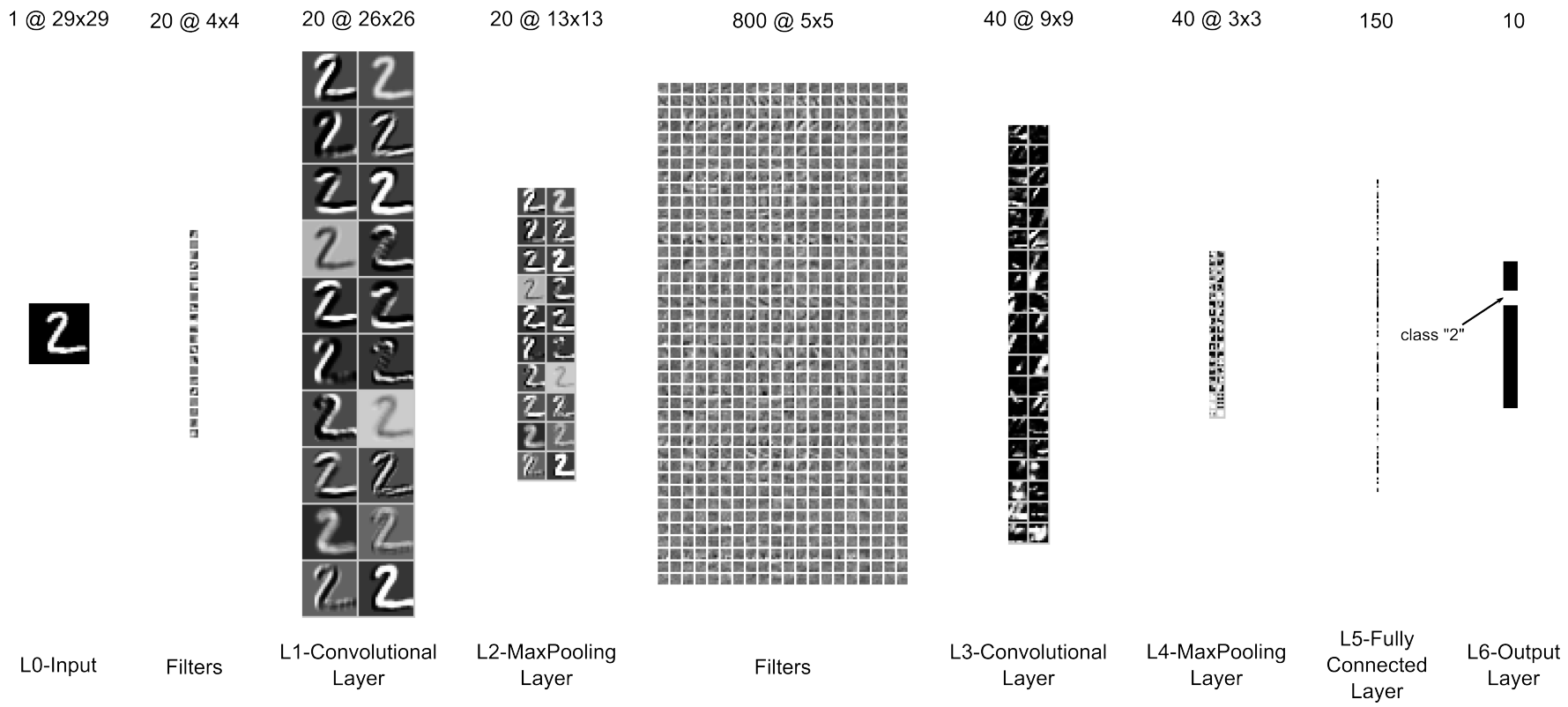


MNIST: 60000 DIGITS FOR TRAINING, 10000 FOR TESTING,
7 LAYER MLP; >12M WEIGHTS; TRAIN 200 DAYS ON CPU = 5 ON GPU; >10^{15} WEIGHT UPDATES, 5B/s,
2010: NEW WORLD RECORD 0.35%

TWO OLD IDEAS: BACKPROP (3-5 DECADES OLD), TRAINING PATTERN DEFORMATIONS (BAIRD 1990, 2 DECADES OLD)

DEEP MAX-POOLING CNN ON GPU (IJCAI 2011)

ONE OUTPUT NEURON PER CLASS NORMALIZED BY SOFTMAX



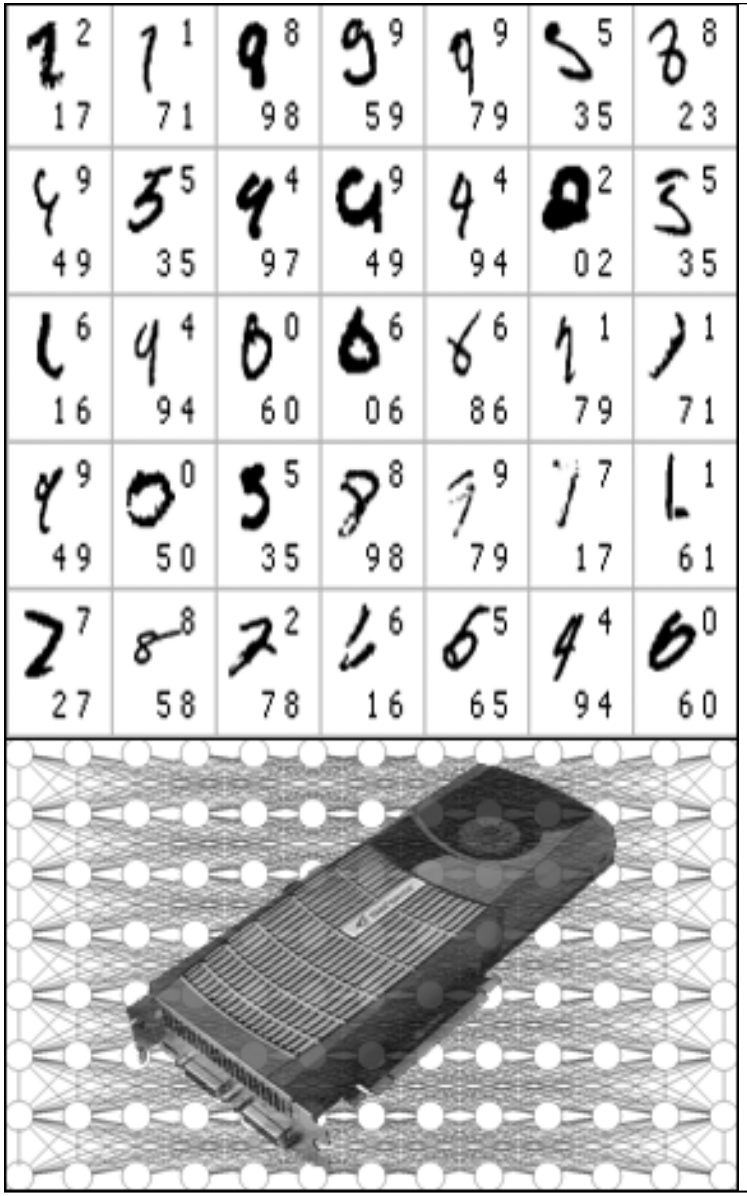
e.g., http://www.idsia.ch/~juergen/deeplearning.html

MPCNNs: 1979 1989
1992 2007 2010 2011

Alternating weight-replicating convolutional layers and subsampling layers: Fukushima 1979. BP for CNNs: LeCun et al 1989. Max-pooling (MP): Weng 1992 (like HMAX, Riesenhuber & Poggio 1999). BP for MPCNNs: Ranzato et al 2007, Scherer et al 2010 - GPUs - Ciresan et al (Swiss AI Lab IDSIA): winning feedforward contests since 2011

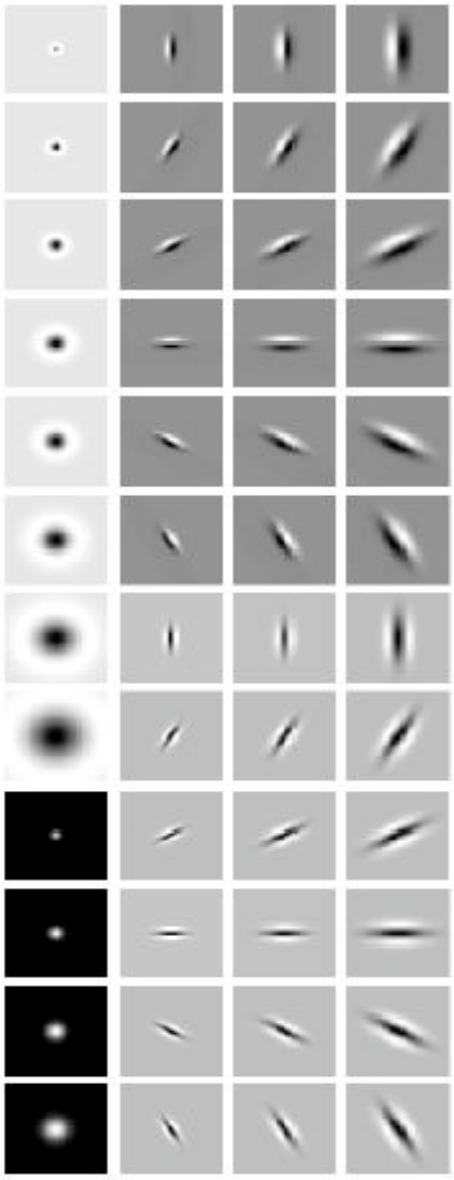
ICDAR 2011 OFFLINE CHINESE HANDWRITING RECOGNITION CONTEST (4000 CLASSES):
1ST & 2ND RANK
OCT 2013: AGAIN BEST RESULTS, NEAR-HUMAN PERFORMANCE

http://www.idsia.ch/~juergen/handwriting.html

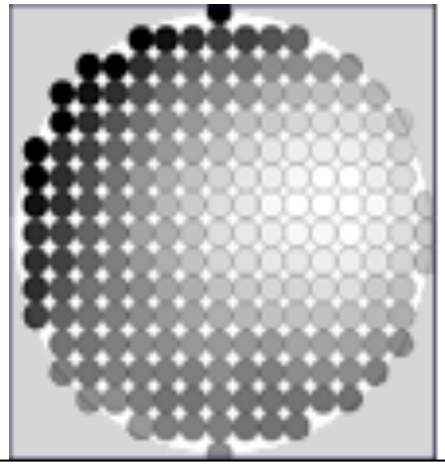


DEEP SPARSE CNN
+ MAX-POOLING
+ MLP ON TOP: 1
YEAR ON CPU = 1
WEEK ON GPU; $5*10^9$
WEIGHT UPDATES/s
2011-2012: FIRST
HUMAN-COMPETITIVE
MNIST RESULT: 0.2%
(after almost a decade of
roughly 0.4%)

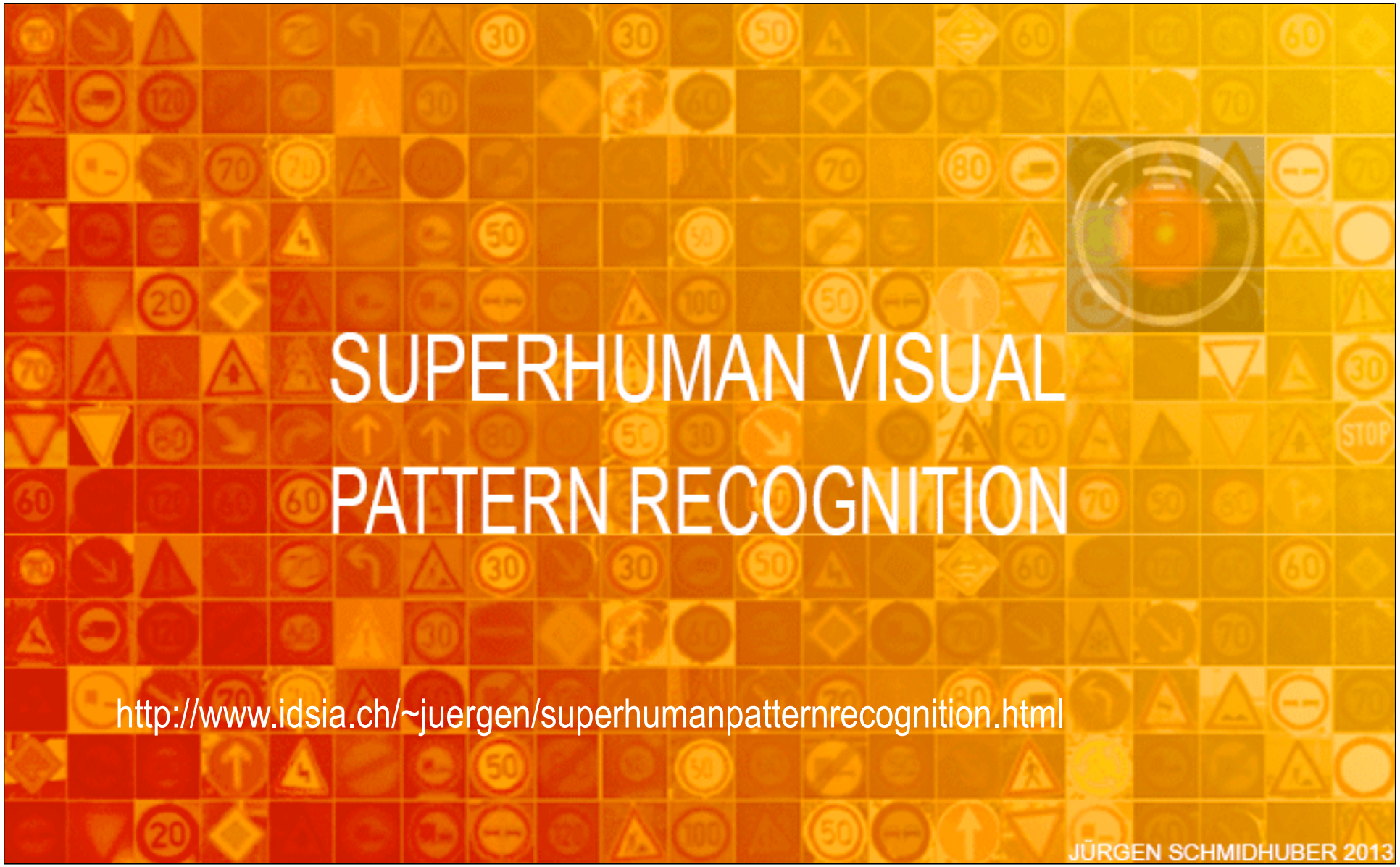
CIRESAN, MEIER, MASCI, GAMBARDELLA, SCHMIDHUBER
(IJCAI 2011) - CIRESAN, MEIER, SCHMIDHUBER (CVPR 2012)



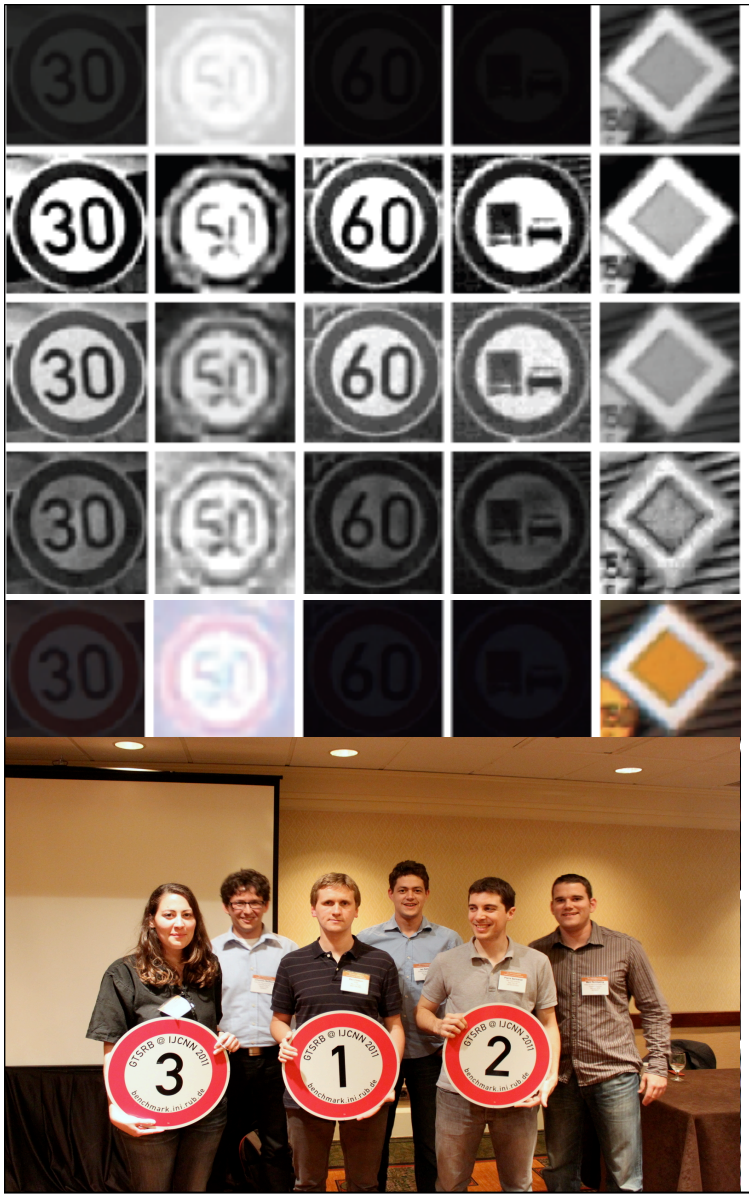
PERIPHERY: FEATURE DETECTORS VERY SIMILAR TO THOSE OF BIOLOGICAL BRAINS OR THOSE FOUND BY OUR UNSUPERVISED METHODS (1992-)



IT'S ALL ABOUT COMPRESSION



Traffic Sign Contest, Silicon Valley, 2011: twice better than humans
three times better than the closest artificial competitor
six times better than the best non-neural method



IJCNN 2011 ON-SITE TRAFFIC SIGN RECOGNITION COMPETITION OF 2 AUGUST 2011:

1ST (0.56% ERROR)

2ND HUMANS (1.16%)

3RD (1.69%)

4TH (3.86%)

Ciresan, Meier, Masci, Schmidhuber, Neural Networks, 2012

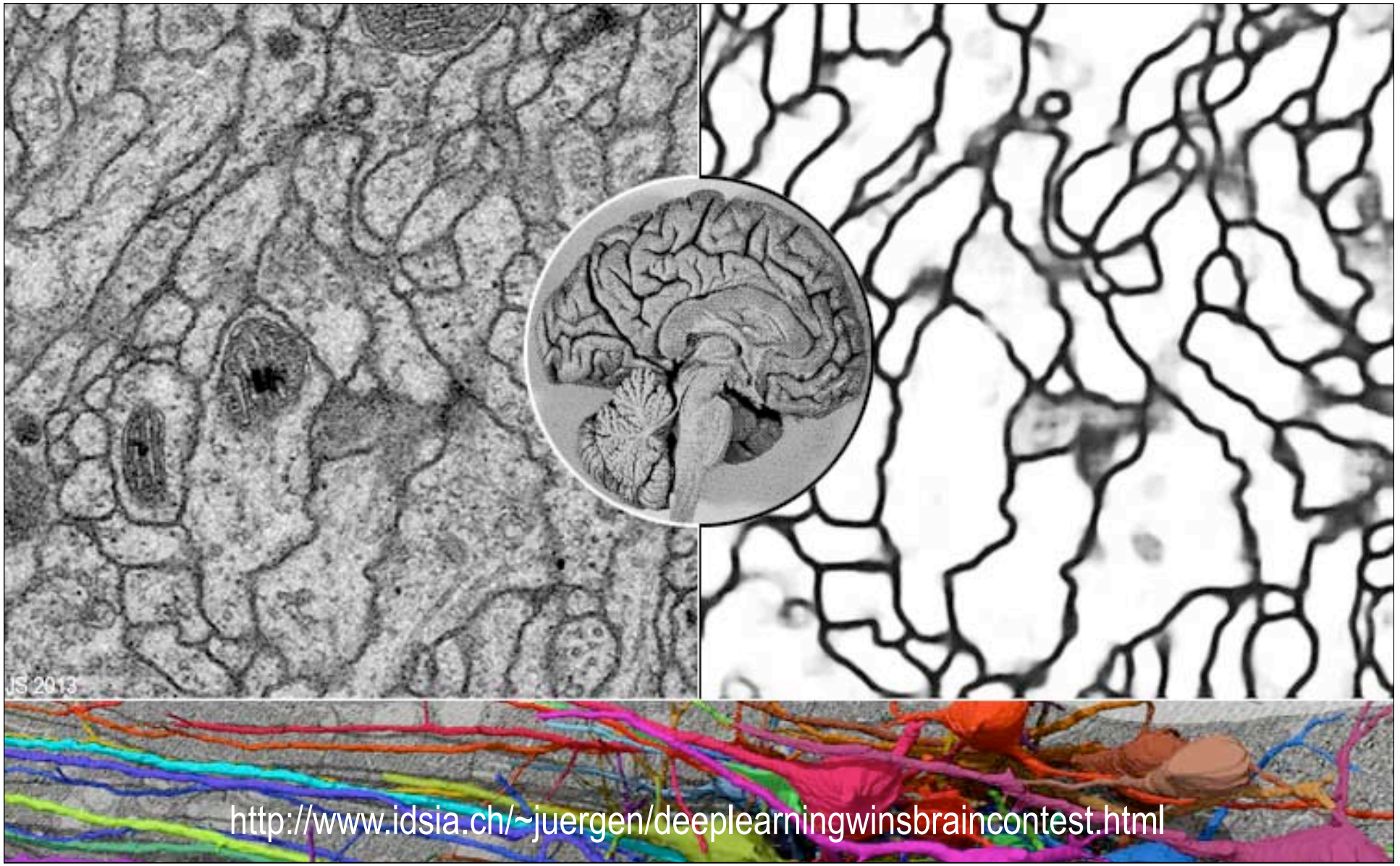
ERNST DICKMANNS
THE ROBOT CAR PIONEER
UNIBW MÜNCHEN, 1980s

ROBOT CARS

1995: MUNICH TO DENMARK AND BACK ON PUBLIC AUTOBAHNS, UP TO 180 KM/H, NO GPS, PASSING OTHER CARS →

2014: 20 YEAR ANNIVERSARY OF SELF-DRIVING CARS IN HIGHWAY TRAFFIC (DICKMANNS, 1994)

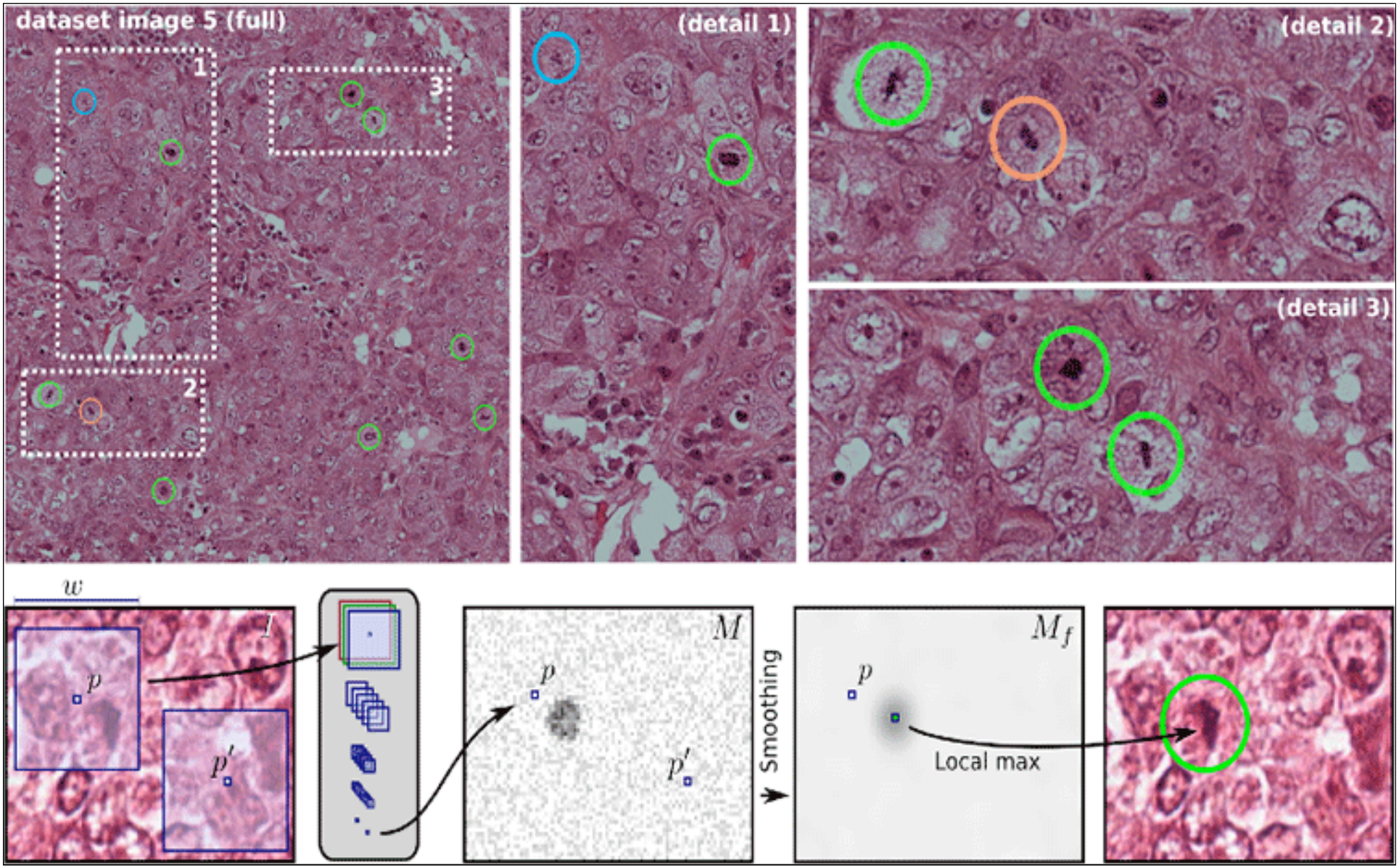
http://www.idsia.ch/~juergen/robotcars.html



Deep Learning Wins ISBI 2012 Brain Image Segmentation Contest

First feedforward Deep Learner to win an image segmentation competition

(but compare deep recurrent LSTM 2009: segmentation & classification)



Deep Learning Wins ICPR 2012 Contest on Mitosis Detection

First pure Deep Learner to win a contest on object detection (in large images)

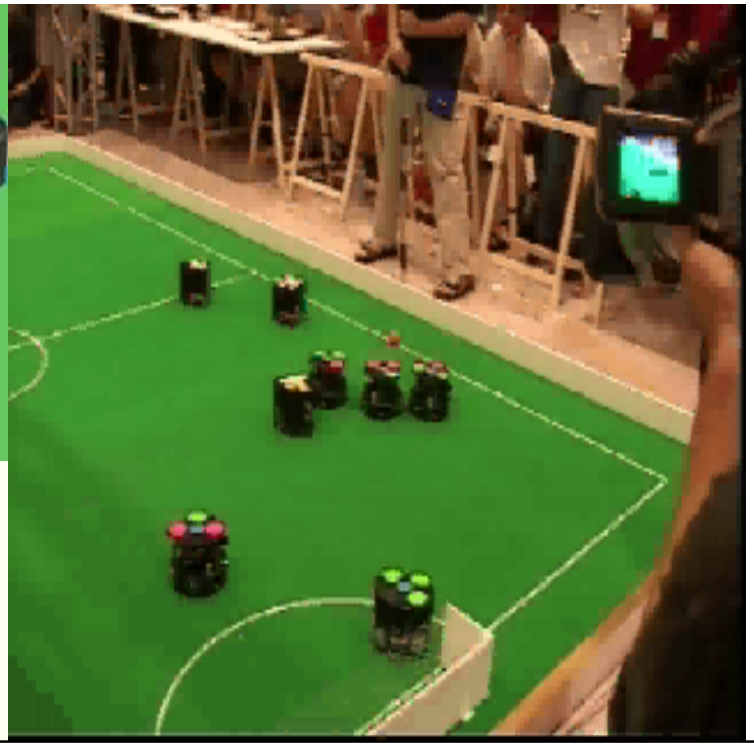
Very fast MPCNN scans: Masci, Giusti, Ciresan, Gambardella, Schmidhuber, ICIP 2013

http://www.idsia.ch/~juergen/deeplearningwinsMICCAIgrandchallenge.html

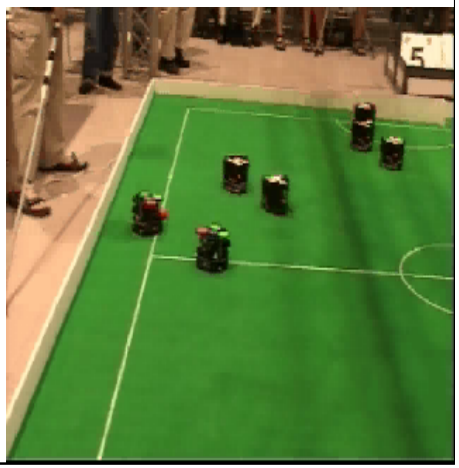
Thanks to Dan Ciresan & Alessandro Giusti

ROBOCUP WORLD CHAMPION 2004
FASTEST LEAGUE, UP TO 5M/S

LOOKAHEAD EXPECTATION & PLANNING WITH NEURAL NETWORKS (J. SCHMIDHUBER, IEEE INNS 1990): SUCCESSFULLY USED FOR ROBOCUP BY ALEXANDER GLOYE-FÖRSTER (NOW IDSIA). MOVIE:

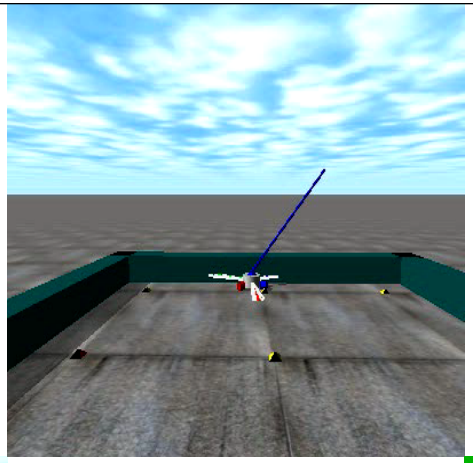


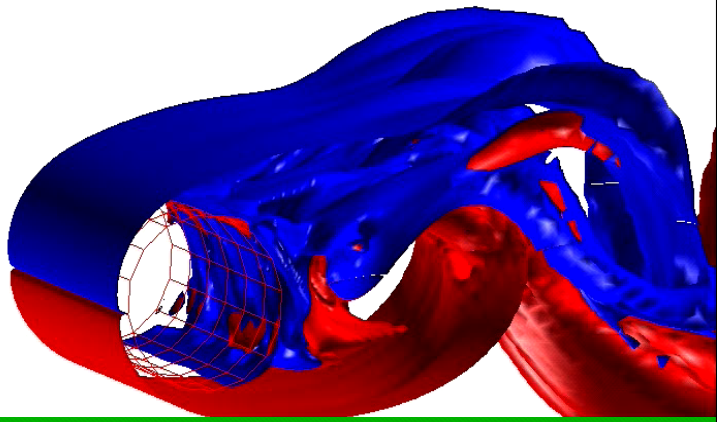
ALEX GLOYE-FÖRSTER, IDSIA, LEADER OF THE FU-FIGHTER ROBOCUP TEAM OF FU BERLIN, WORLD CHAMPION 2004

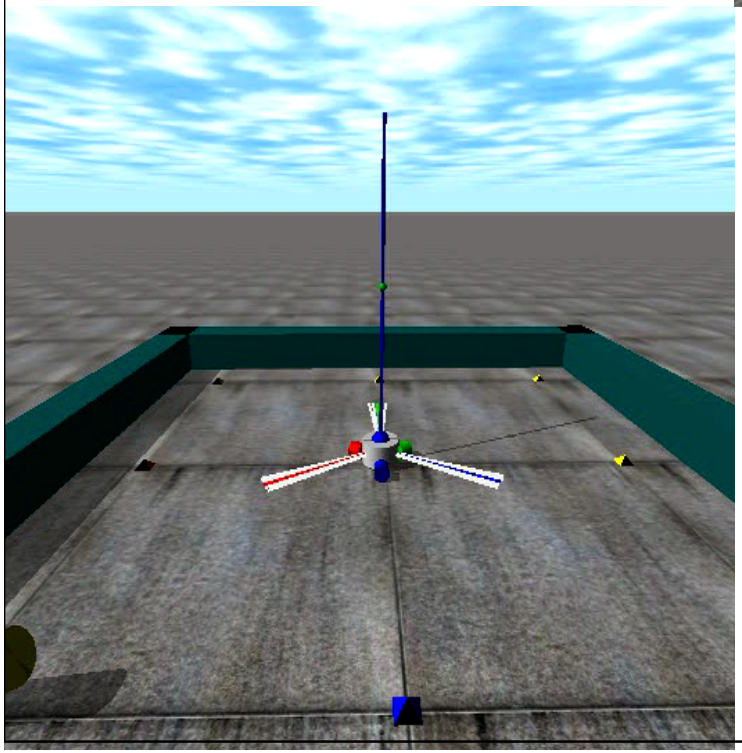


http://www.idsia.ch/~juergen/learningrobots.html

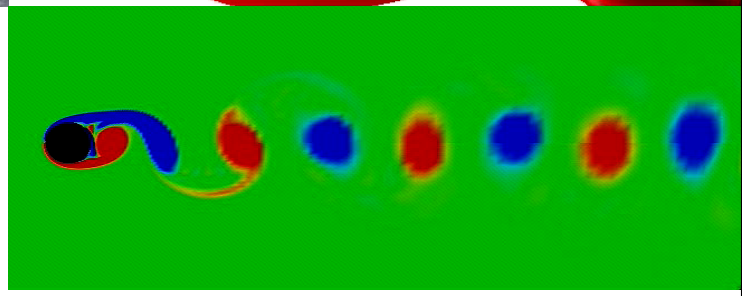
DIRECT POLICY SEARCH BY RNN EVOLUTION ETC



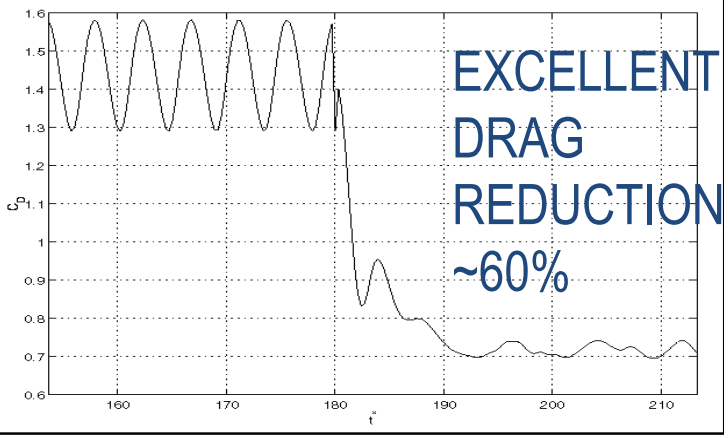




E.G., ROBOT LEARNS TO BALANCE 1 OR 2 POLES (3D JOINT)

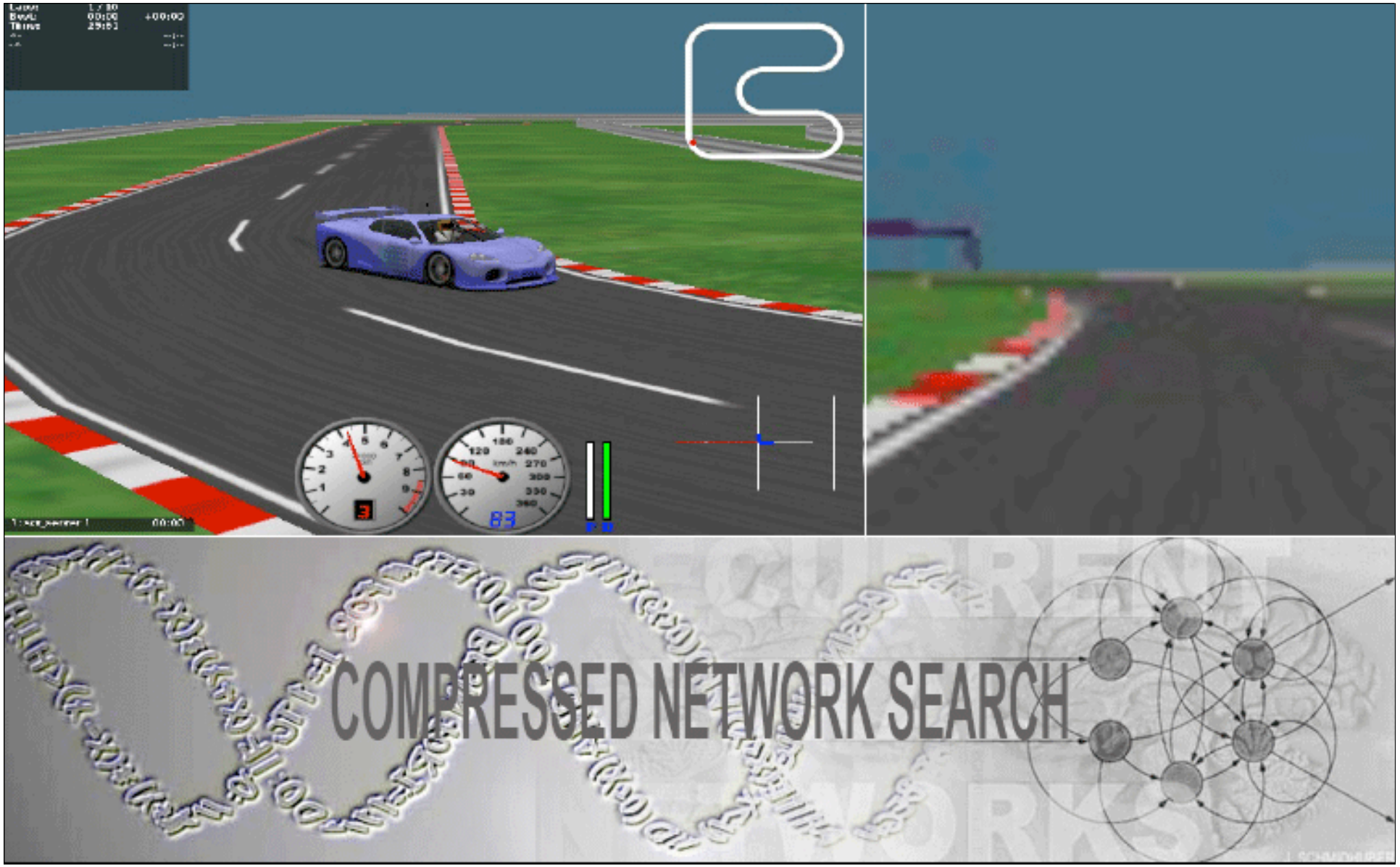


WITH FAUSTINO GOMEZ (IDSIA), MICHELE MILANO (ETHZ) & PETROS KOUMOUTSAKOS



EXCELLENT DRAG REDUCTION ~60%

http://www.idsia.ch/~juergen/evolution.html



Finds Complex Neural Controllers with a Million Weights – RAW VIDEO INPUT!

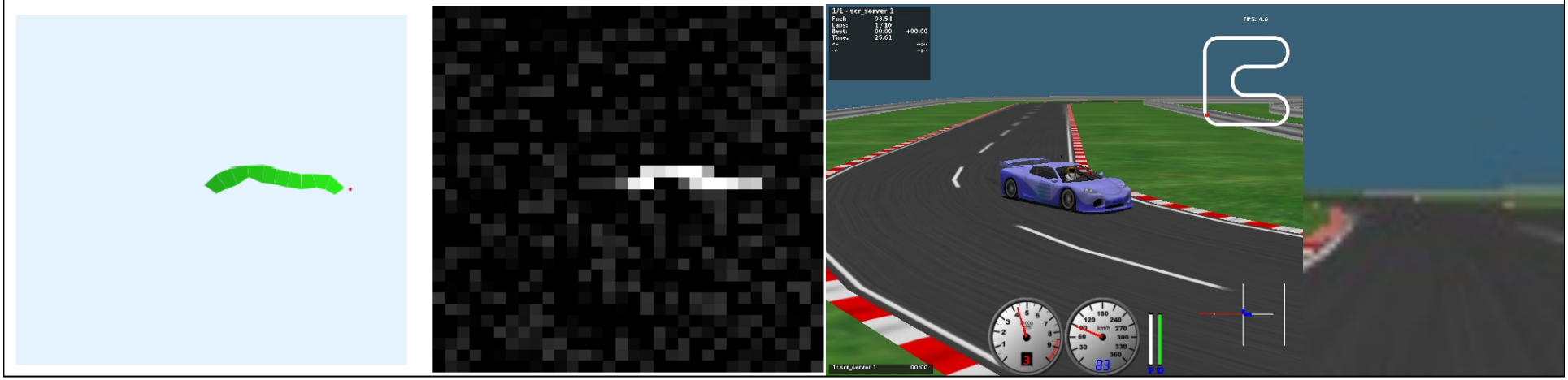
Faustino Gomez, Jan Koutnik, Giuseppe Cuccu, J. Schmidhuber, GECCO 2013

http://www.idsia.ch/~juergen/compressednetworksearch.html

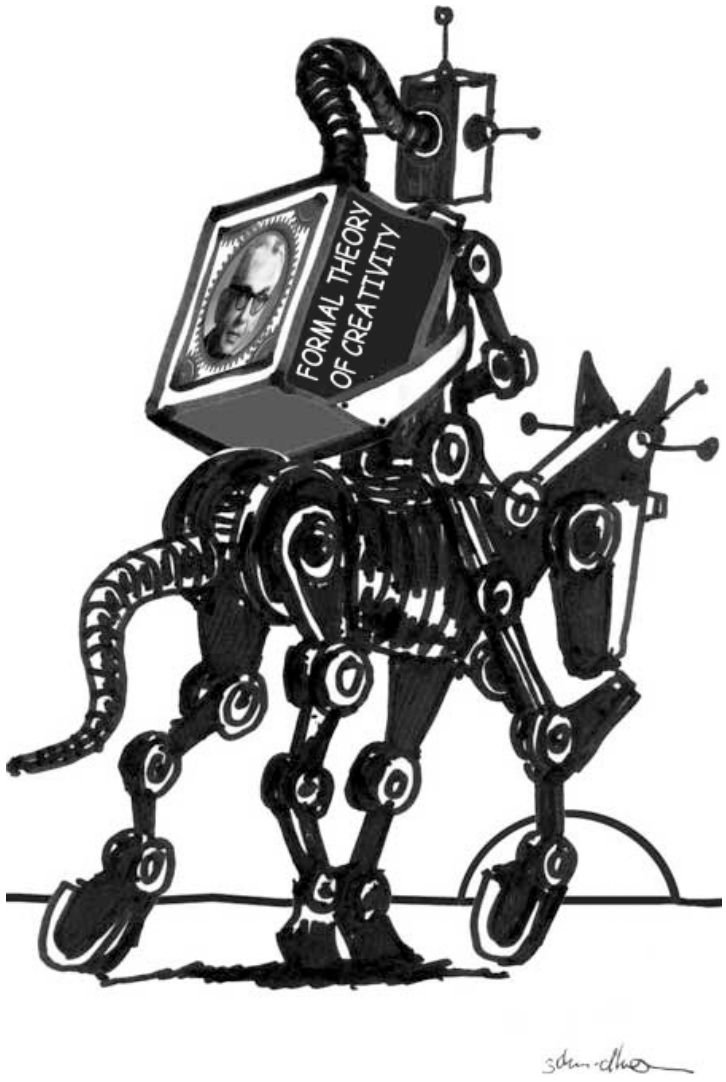
OCTOPUS-ARM CONTROL: 82 IN, 32 OUT, 3'680 W, ONLY 20 DCT COEFFICIENTS, COMPRESSION 1:184

OCTOPUS-ARM WITH LOW-LEVEL VISION, 32X32 IN, 32 OUT, **33'824** W, **160** DCT, COMPRESSION **1:211**

TORCS DRIVING, LOW-LEVEL VISION, 64X64 IN, 3 OUT, **1'115'139** W, **200** DCT, COMPRESSION **1:5575**



http://www.idsia.ch/~juergen/compressednetworksearch.html



Maximize Future Fun(Data X,O(t))~
∂CompResources(X,O(t))/∂t

FORMAL THEORY OF FUN
& NOVELTY & SURPRISE & ATTENTION & CREATIVITY & CURIOSITY & ART & SCIENCE & HUMOR

http://www.idsia.ch/~juergen/creativity.html

E.g., Conn. Sci. 18(2):173-187, 2006; IEEE TAMD, 2(3):230-247, 2010