JÜRGEN SCHMIDHUBER

DEEP LEARNING

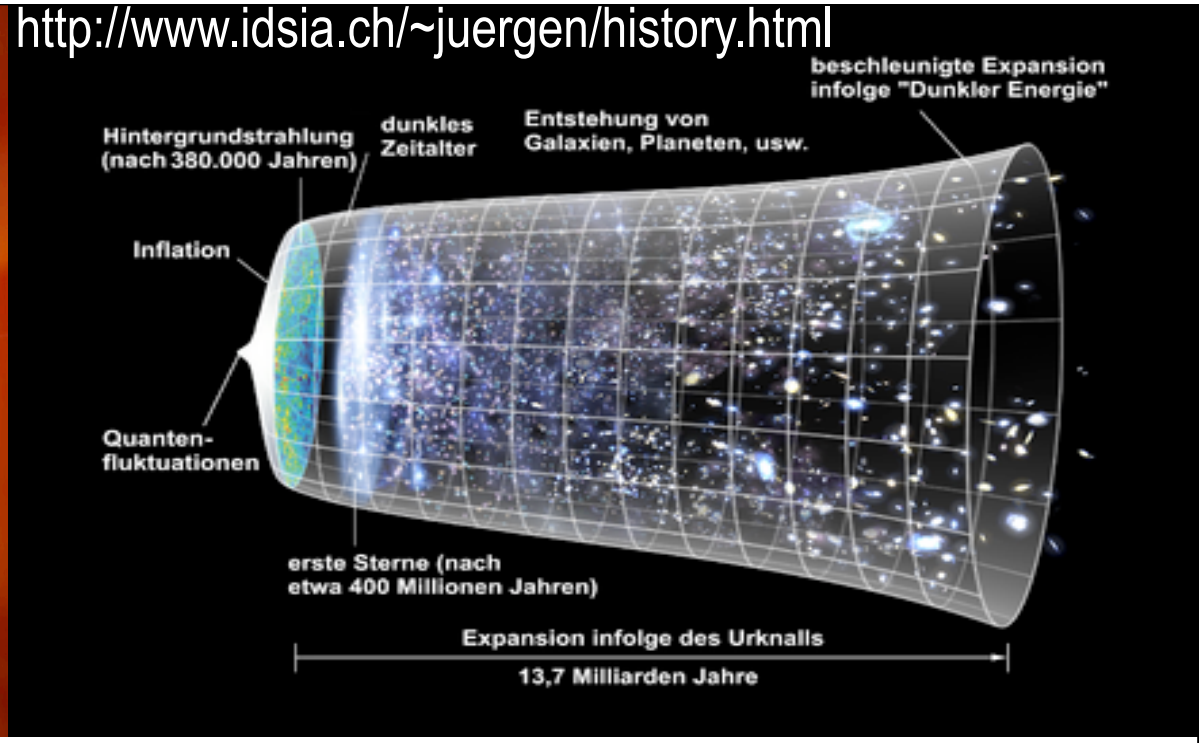
THE SWISS AI LAB
IDSIA - USI & SUPSI

NNAISENSE

Jürgen Schmidhuber

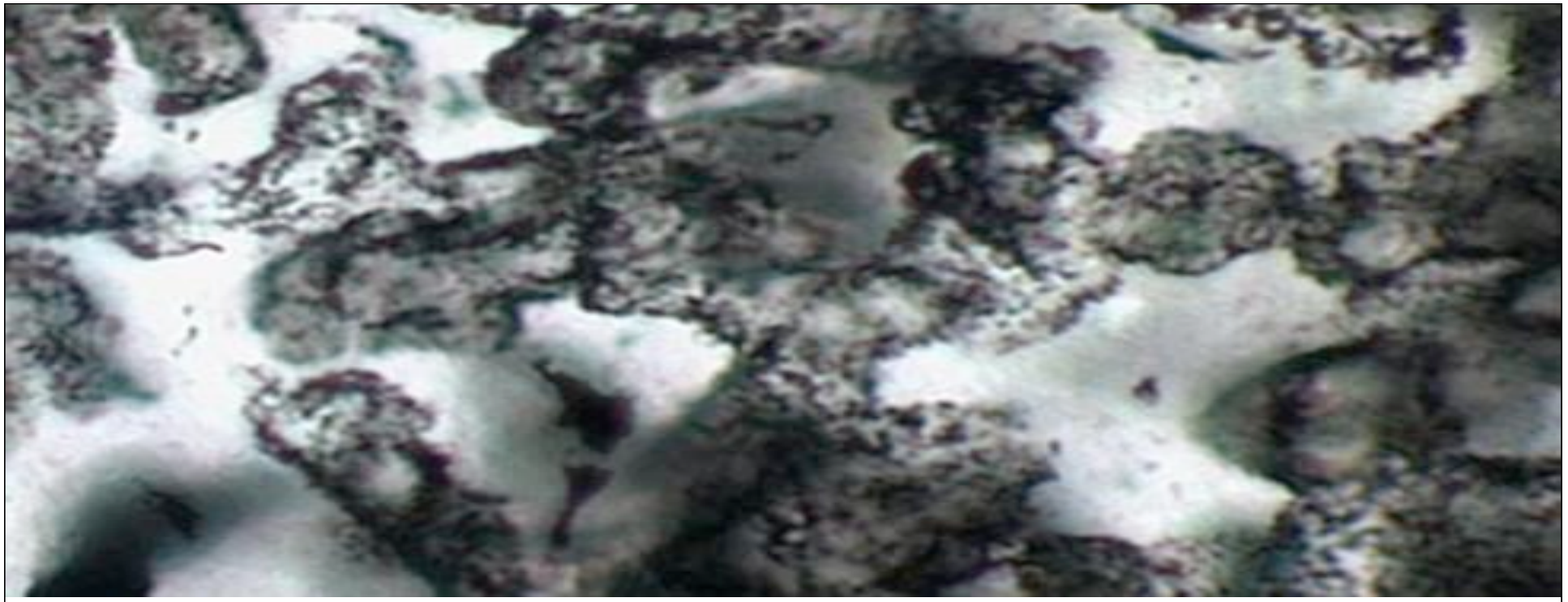
You_again Shmidhoobuh

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



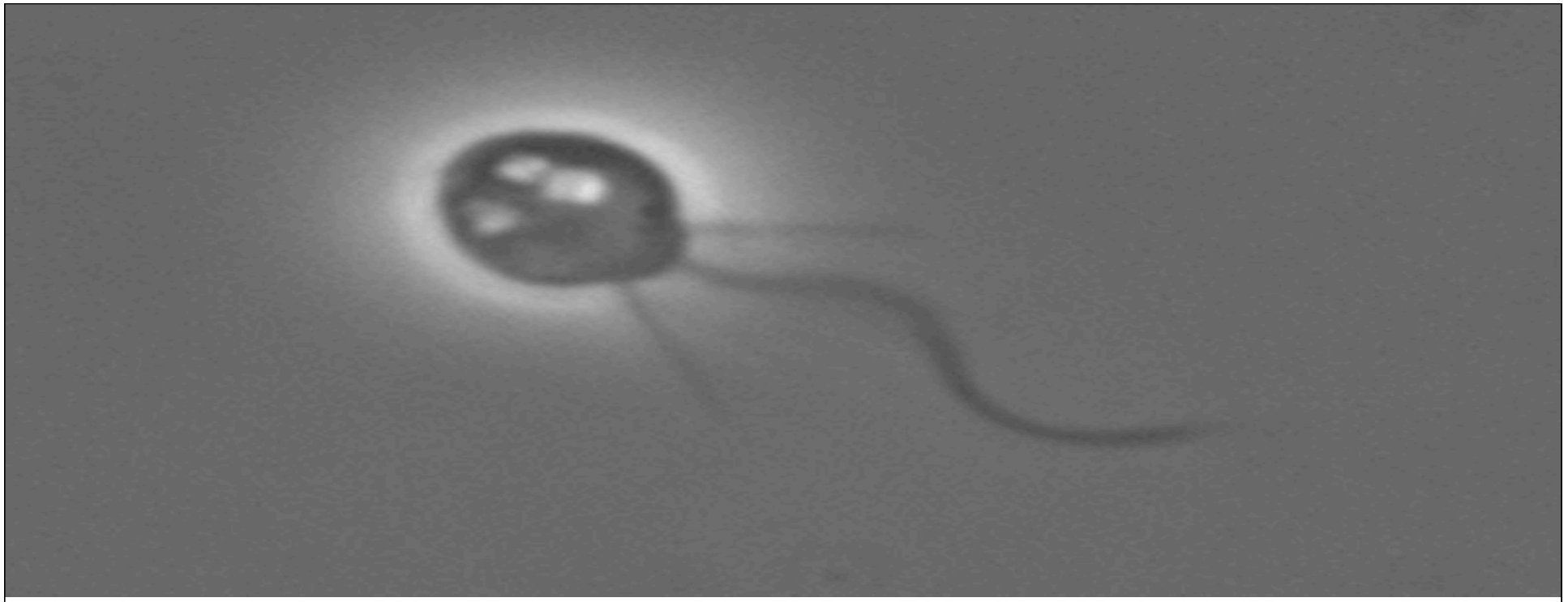
Ultimate Trend
Will history converge
around 2050 = Ω?

Pattern starts at
Ω - 13.8 B years:
Big Bang



Ω - ¼ of this time

Ω - 3.5 B years: Life



Ω - ¼ of this time

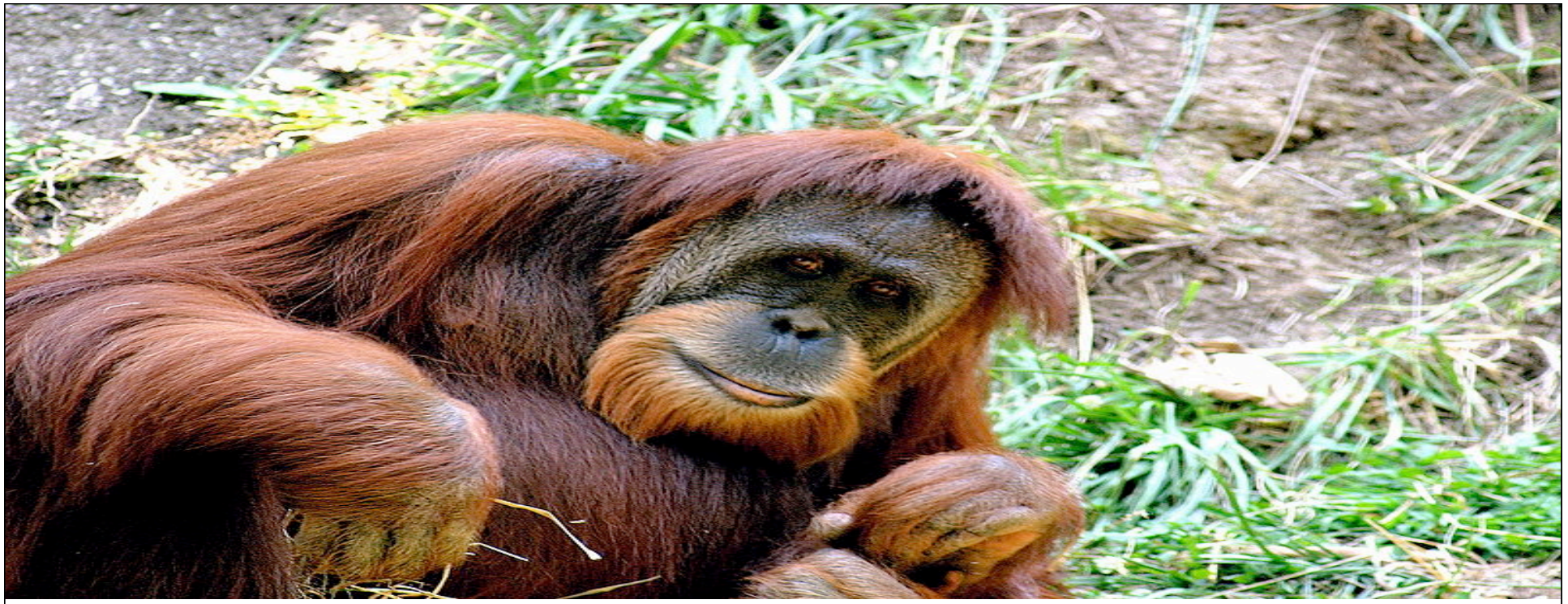
Ω - 0.9 B years:
Animal-like life

Ω - ¼ of this time

Ω - 220 M years: Mammals

Ω - ¼ of this time

Ω - 55 M years: Primates



Ω - ¼ of this time

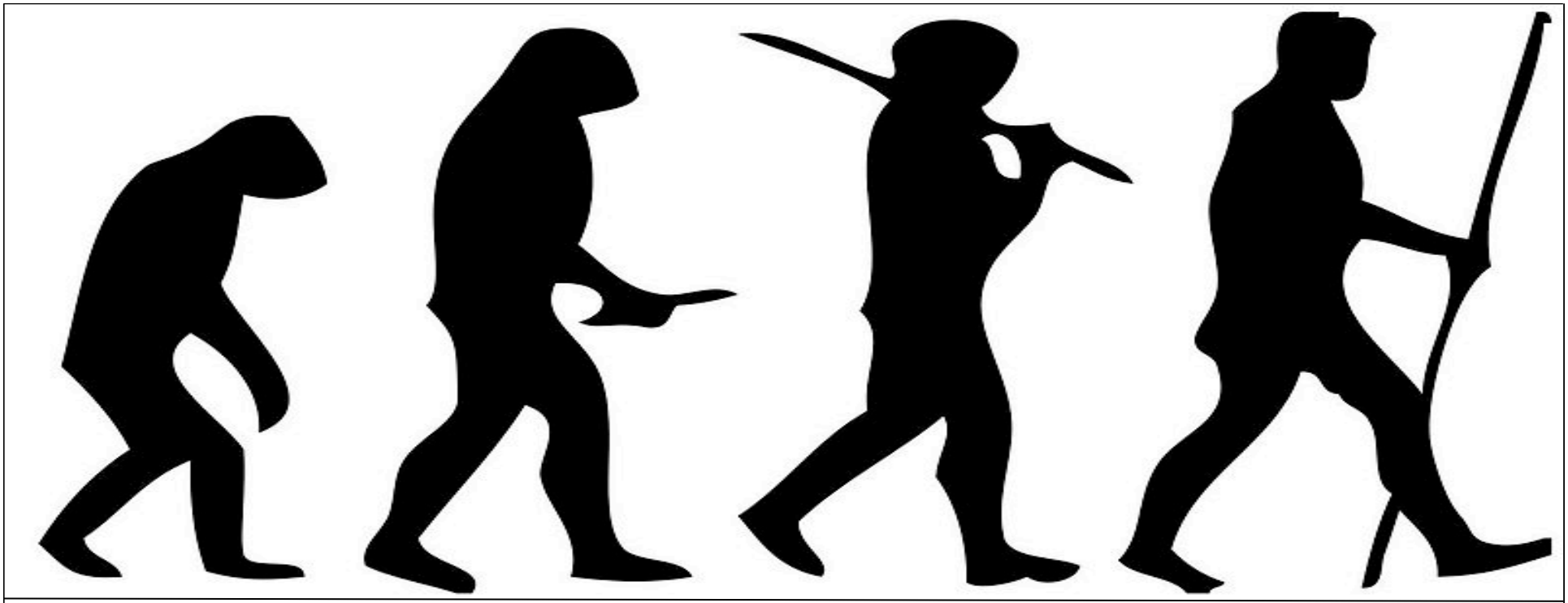
Ω - 13 M years: Hominids

Ω - ¼ of this time

Ω - 3.5 M years: Stone tools

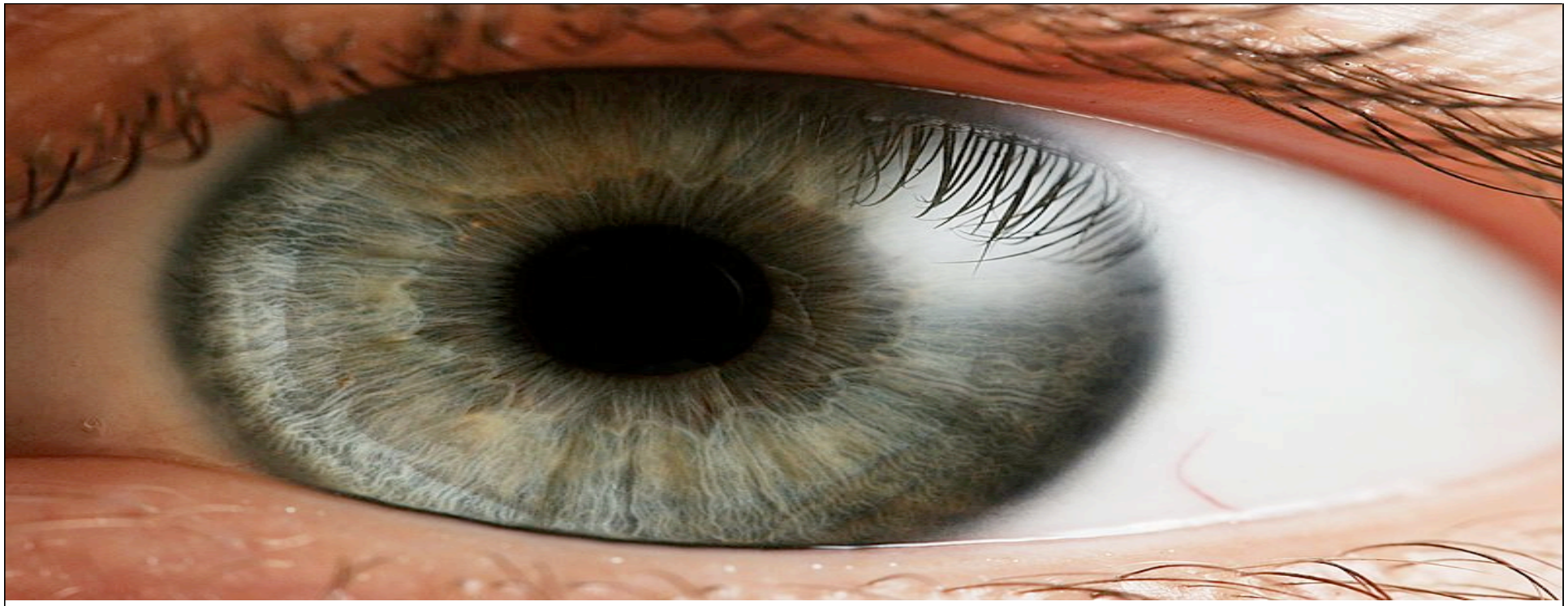
Ω - ¼ of this time

Ω - 850,000 years:
Controlled fire



Ω - ¼ of this time

Ω - 210,000 years: Anatomically modern man



Ω - ¼ of this time

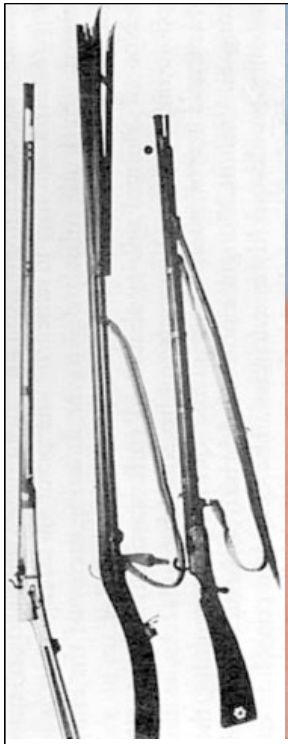
Ω - 50,000 years: Behaviorallly modern man

Ω - ¼ of this time

Ω - 13,000 years:
Neolithic revolution

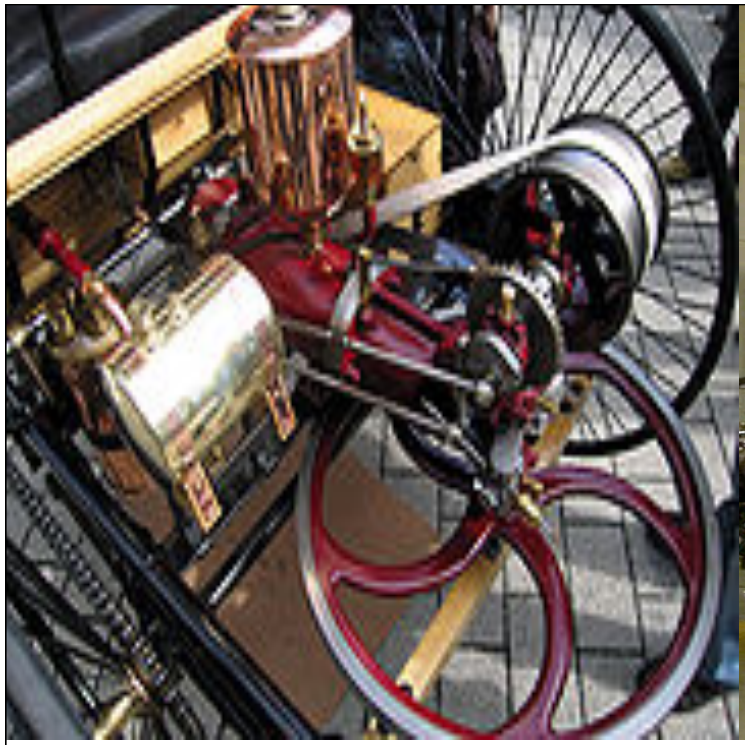
Ω - ¼ of this time

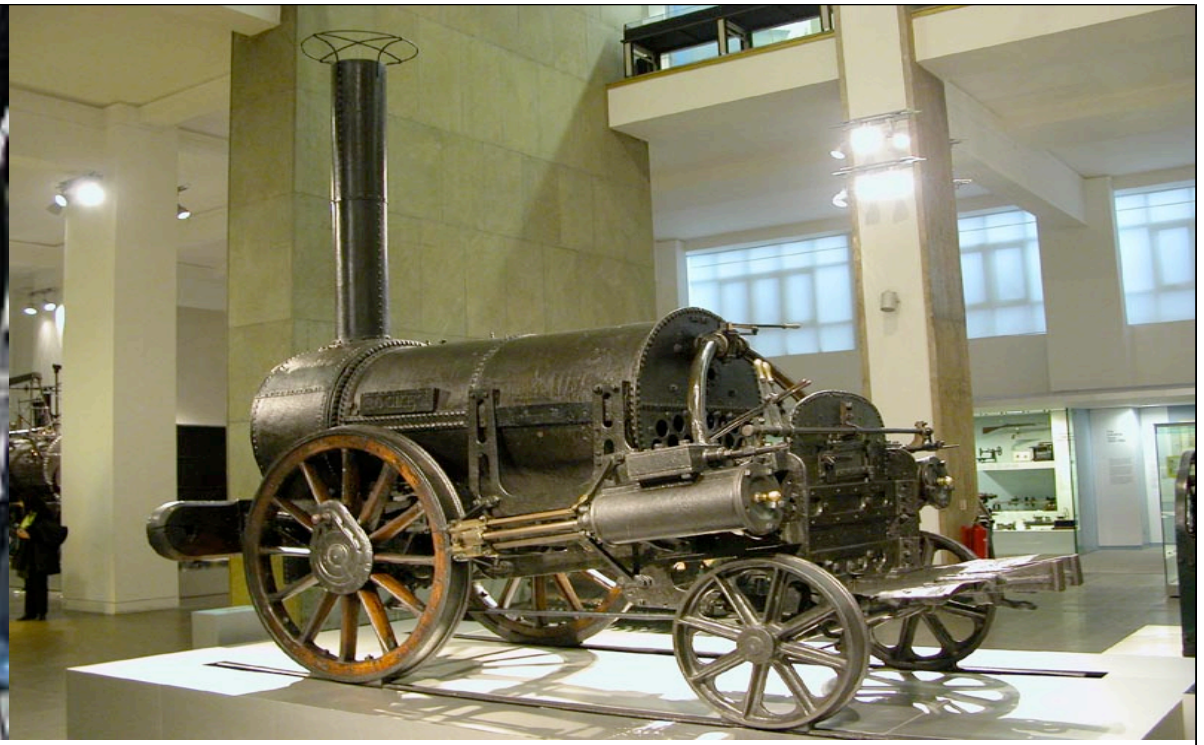
Ω - 3,300 years: Iron age



Ω - ¼ of this time

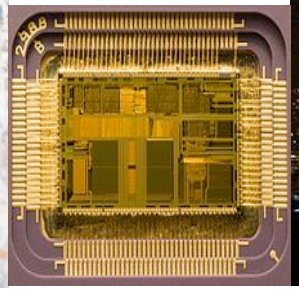
Ω - 800 years:
Guns & rockets





Ω - ¼ of this time

Ω - 200 years:
Industrial revolution



Ω - ¼ of this time

Ω - 50 years (now):
Information revolution

Ω - 12 years

Small computers
with 1 brain power

Ω - 3 years

?

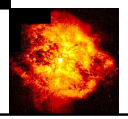
Ω - 9 months

??

Ω - 10 weeks

????

????????



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

JÜRGEN SCHMIDHUBER

DEEP LEARNING

RNNAISSANCE

THE SWISS AI LAB
IDSIA - USI & SUPSI

NNAISENSE

deep learning overview

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

Deep Learning is a half century old although recent "tabloid science" stories claim it is a recent thing

Critique (also at Google+) of paper by self-proclaimed “deep learning conspiracy” (LeCun & Bengio & Hinton) who cite each other but not the pioneers of the field: http://www.idsia.ch/~juergen/deep-learning-conspiracy.html

Father of Deep Learning

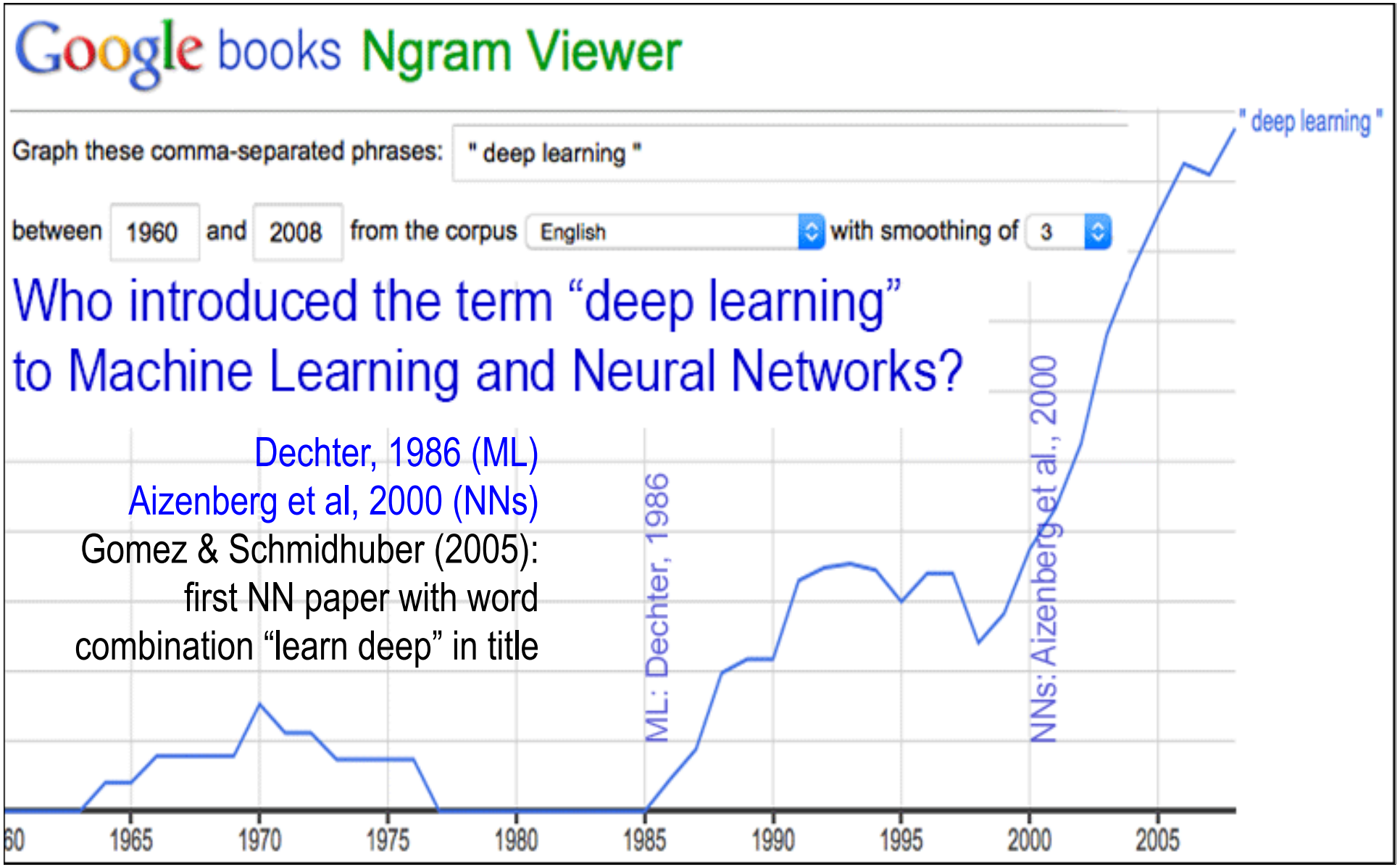
Ivakhnenko et al, since 1965

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

8 layers already back in 1971

still used in the 2000s

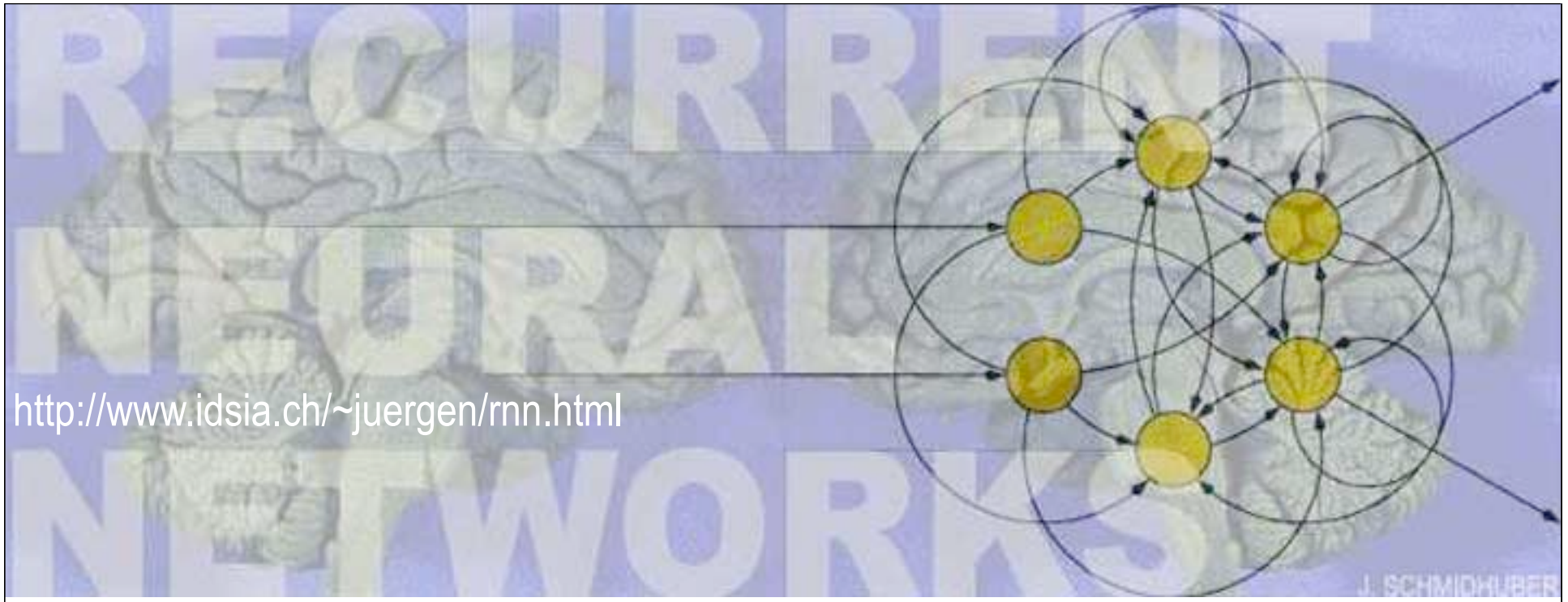

Google books Ngram Viewer
Graph these comma-separated phrases: " deep learning "
between 1960 and 2008 from the corpus English with smoothing of 3
Who introduced the term "deep learning" to Machine Learning and Neural Networks?
Dechter, 1986 (ML)
Aizenberg et al, 2000 (NNs)
Gomez & Schmidhuber (2005): first NN paper with word combination "learn deep" in title
" deep learning "
ML: Dechter, 1986
NNs: Aizenberg et al., 2000
60
1965
1970
1975
1980
1985
1990
1995
2000
2005

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

who invented backpropagation?

Supervised Backpropagation (BP)

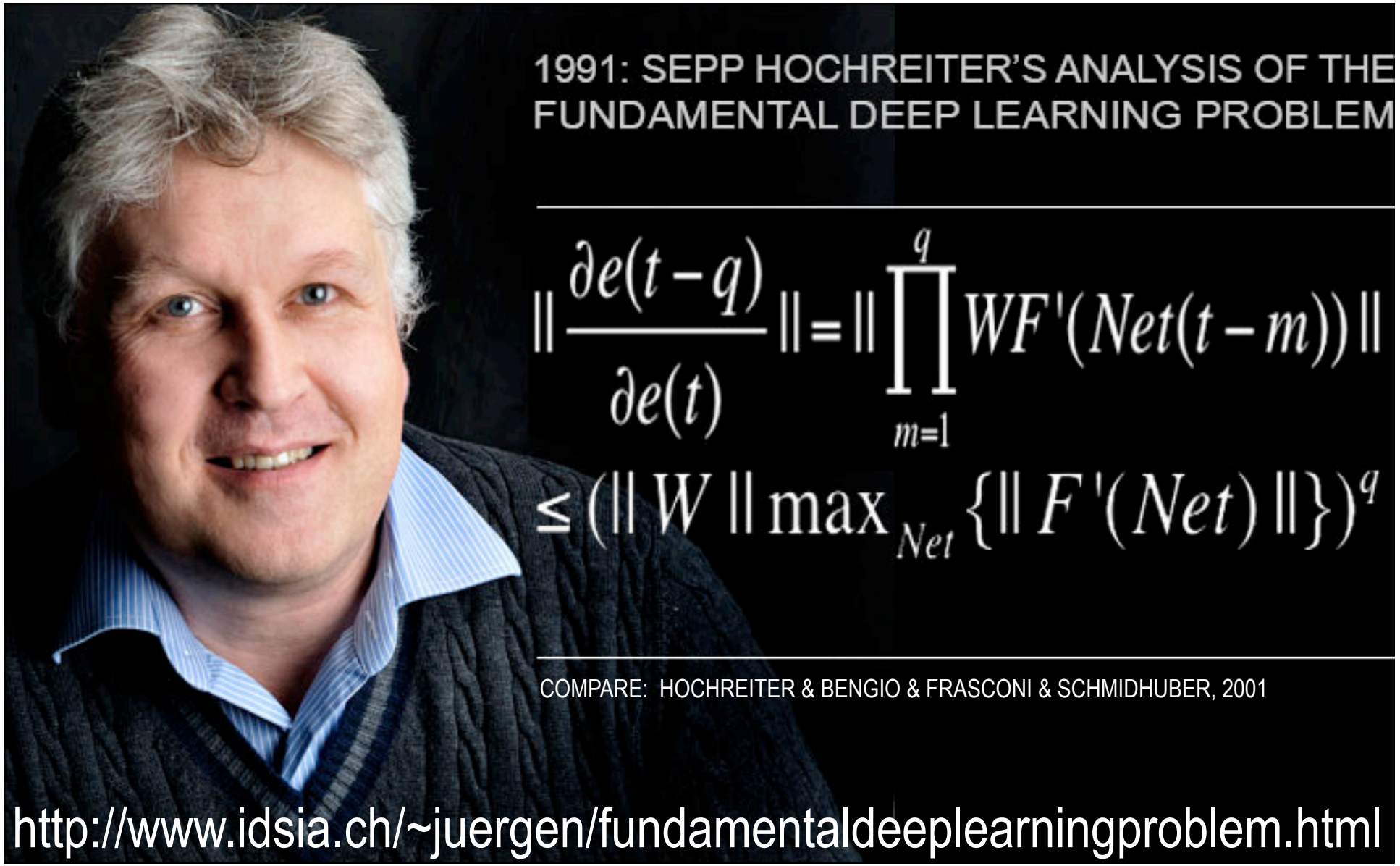
Continuous BP in Euler-LaGrange Calculus + Dynamic Programming: Bryson 1961, Kelley 1960. BP through chain rule only: Dreyfus 1962. `Modern BP' in sparse, discrete, NN-like nets: Linnainmaa 1970. Weight changes: Dreyfus 1973. Automatic differentiation: Speelpenning 1980. BP applied to NNs: Werbos 1982. Experiments & internal representations: Rumelhart et al 86. RNNs: e.g., Williams, Werbos, Robinson, 1980s...



The deepest NNs:

RNNs are general computers

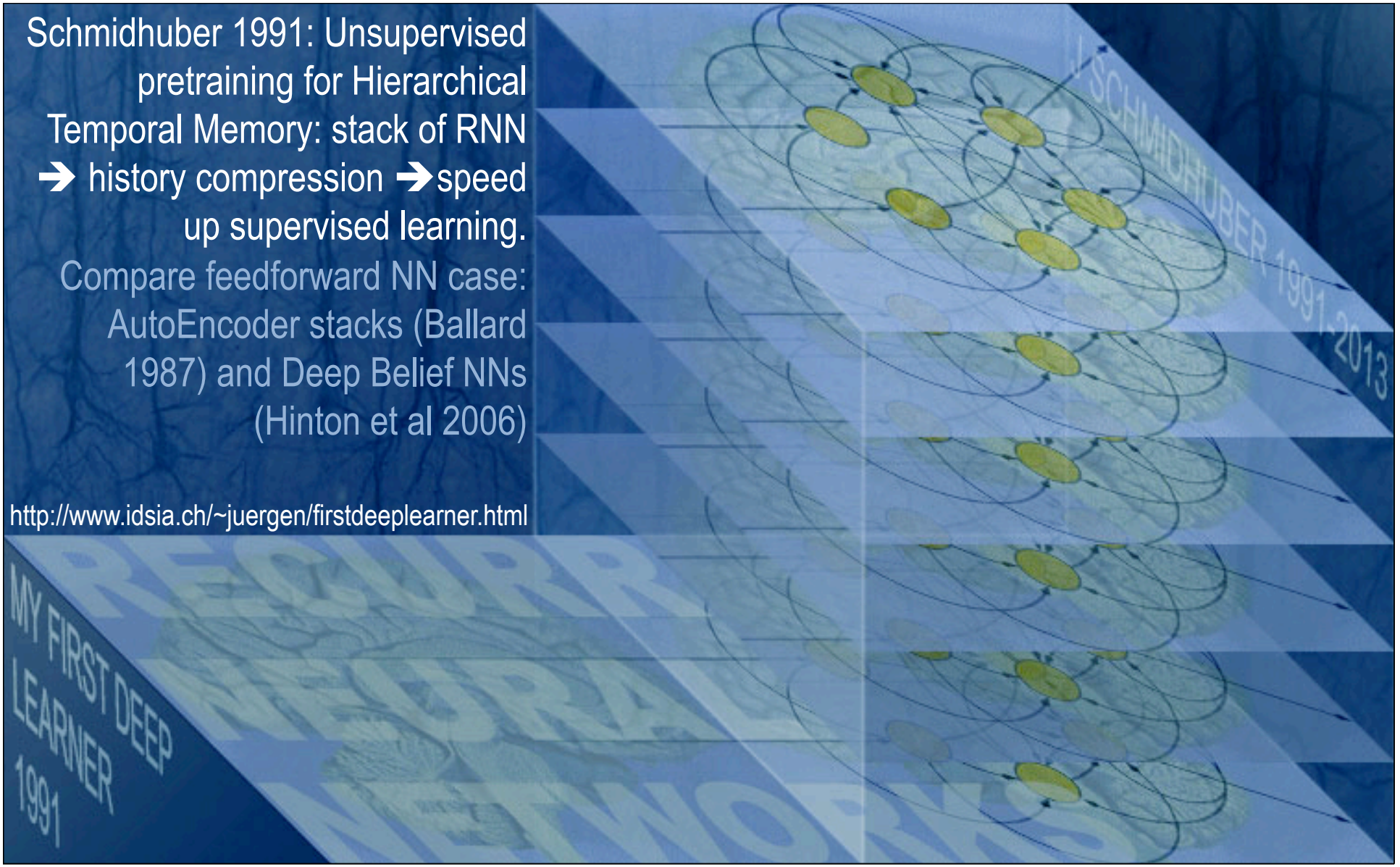
Learn program = weight matrix


1991: SEPP HOCHREITER'S ANALYSIS OF THE FUNDAMENTAL DEEP LEARNING PROBLEM
$$\| \frac{\partial e(t-q)}{\partial e(t)} \| = \| \prod_{m=1}^{q} WF'(Net(t-m)) \|$$
$$\leq (\| W \| \max_{Net} \{ \| F'(Net) \| \})^q$$
COMPARE: HOCHREITER & BENGIO & FRASCONI & SCHMIDHUBER, 2001
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)

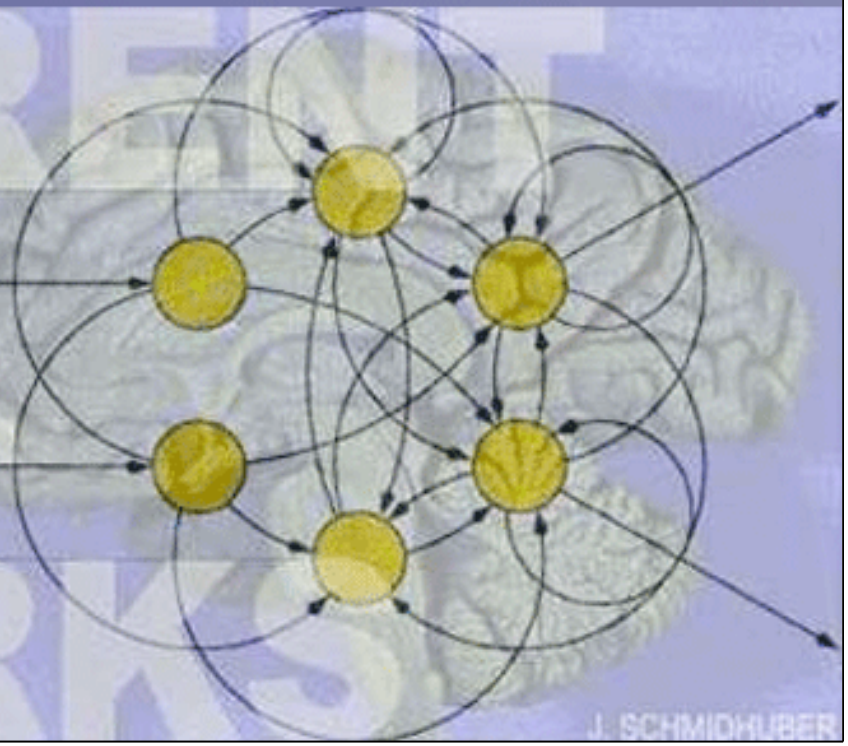
http://www.idsia.ch/~juergen/firstdeeplearner.html

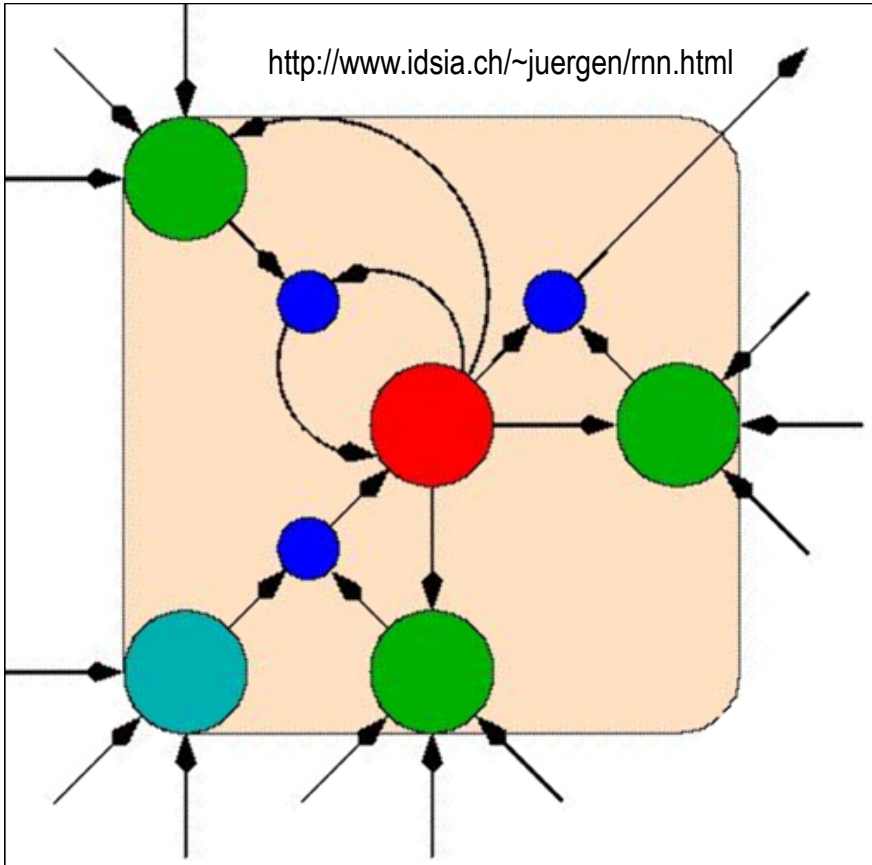


LONG SHORT-TERM MEMORY

1997-2007. Since 2015 on your phone! Google, Microsoft, IBM, others, all use LSTM now

With Hochreiter, Gers, Graves, Fernandez, Gomez, Bayer…





Red: linear unit: self-weight 1.0: transports error across 1000s of time steps. Green: gates open / protect access. Blue: multiplications

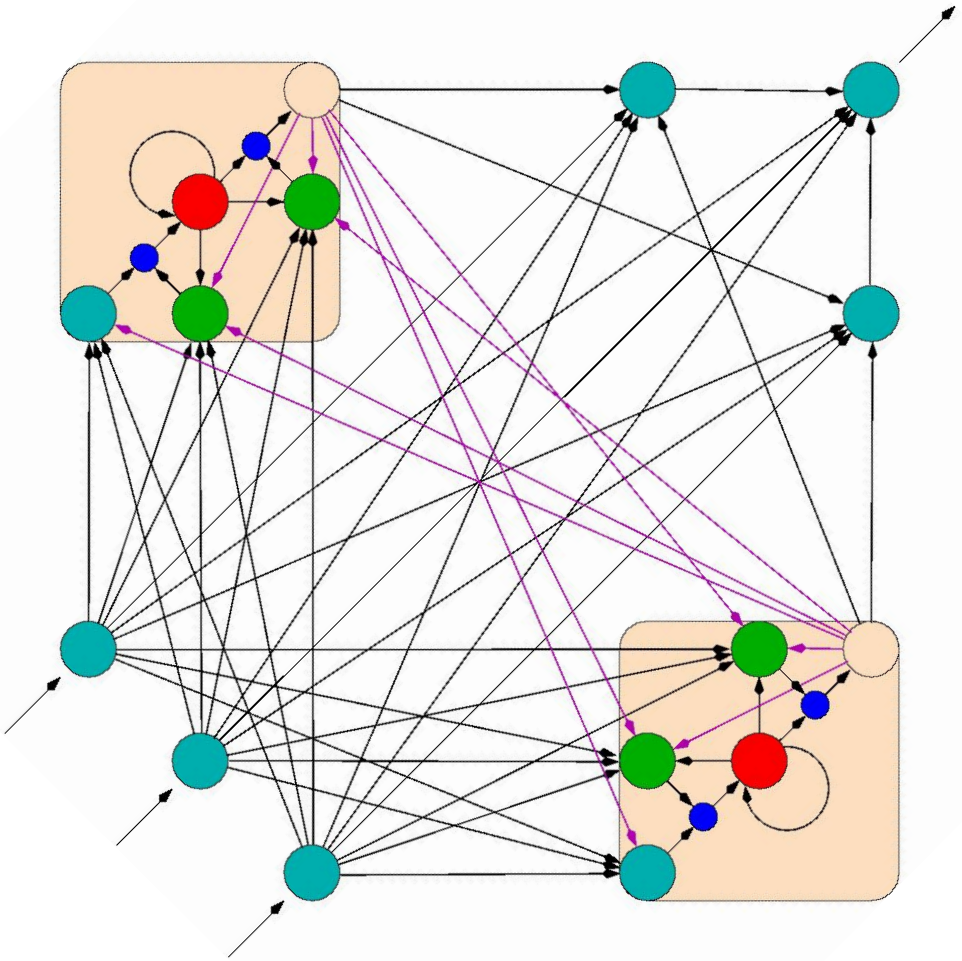
Long Short-Term Memory LSTM: no vanishing gradients

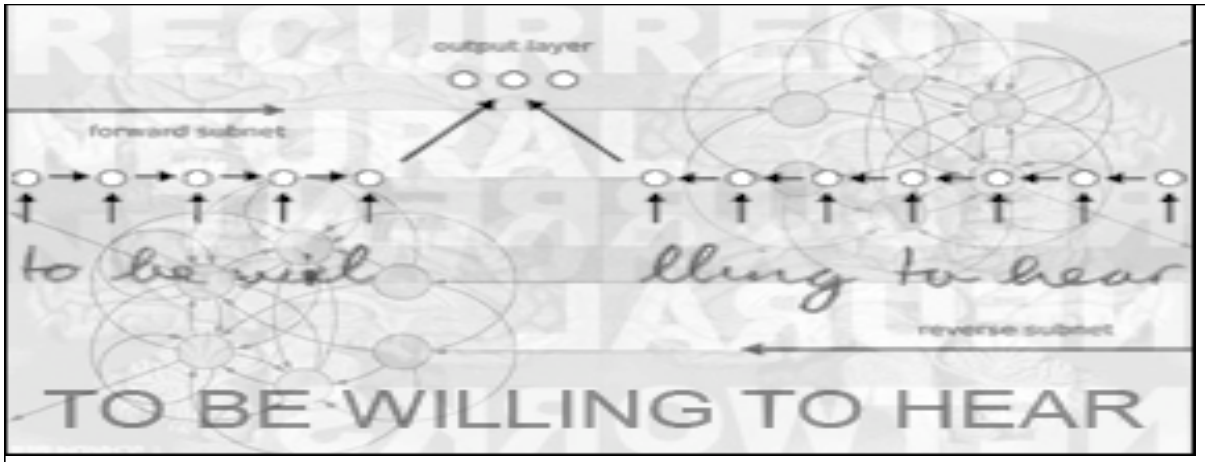
Gradient-based LSTM variants (1997, 1999, 2001, 2005, 2007,…) learn many previously unlearnable Deep Learning tasks: context-sensitive grammars, music composition, R-Learning robots, metalearning, speech recognition (vs HMMs/GMMs), protein prediction, connected handwriting, machine translation…. No bias towards recent or ancient events!

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)
Justin Bayer (2009), others

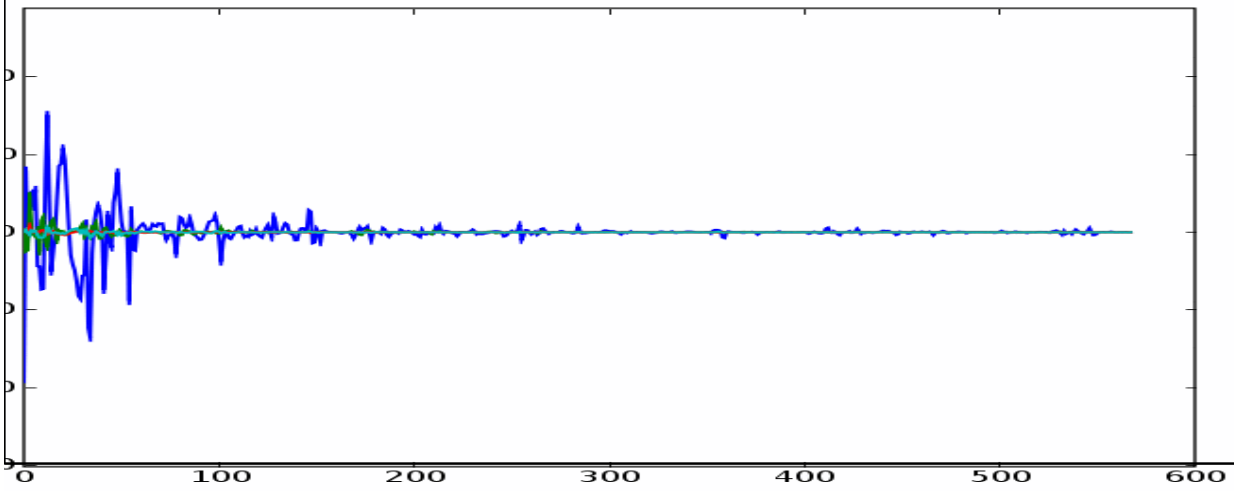
Postdocs at IDSIA (2000s):
Fred Cummins
Santiago Fernandez
Faustino Gomez
Others



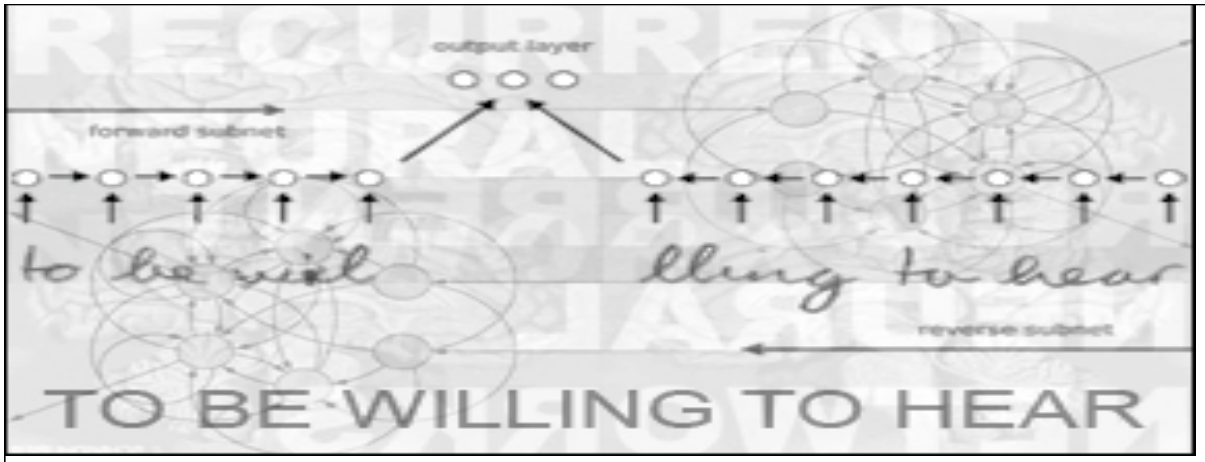


Connectionist Temporal Classification (CTC):
Graves, Fernandez, Gomez, Schmidhuber
ICML 2006

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

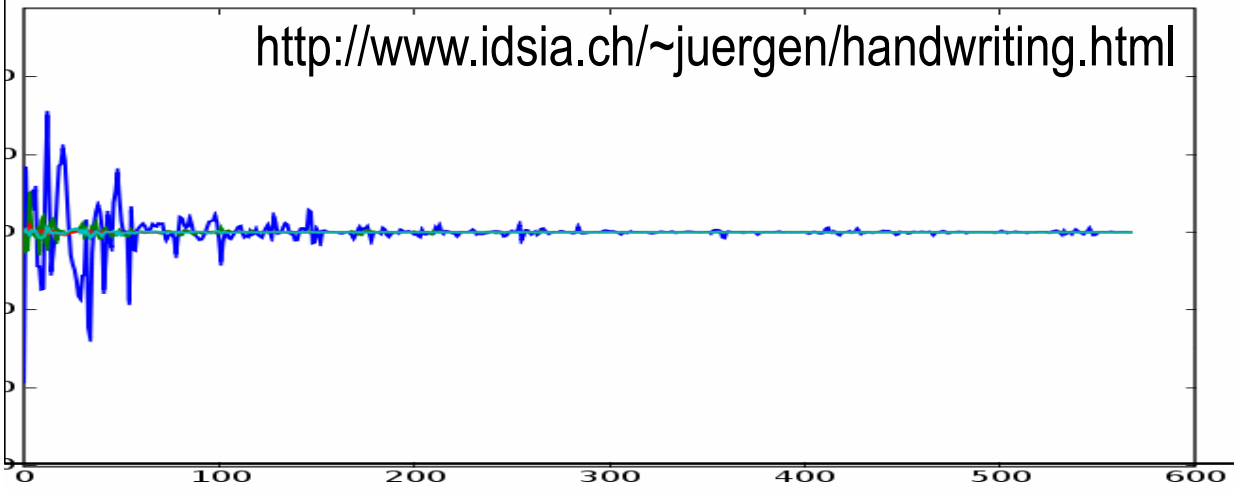


No pre-segmented data; RNN maximises probability of training set label sequences

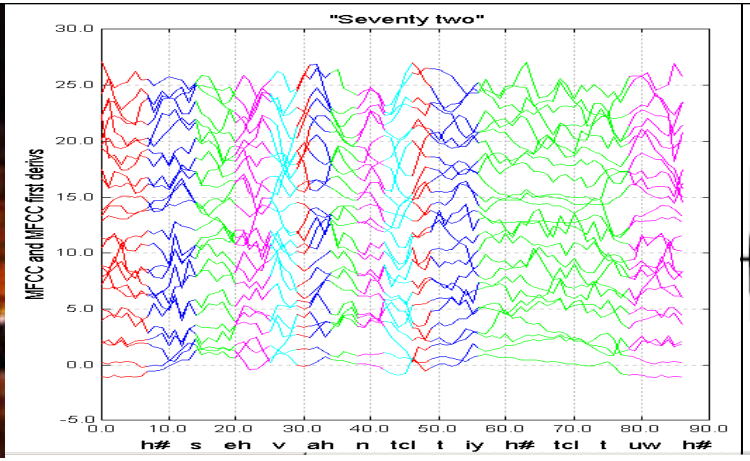


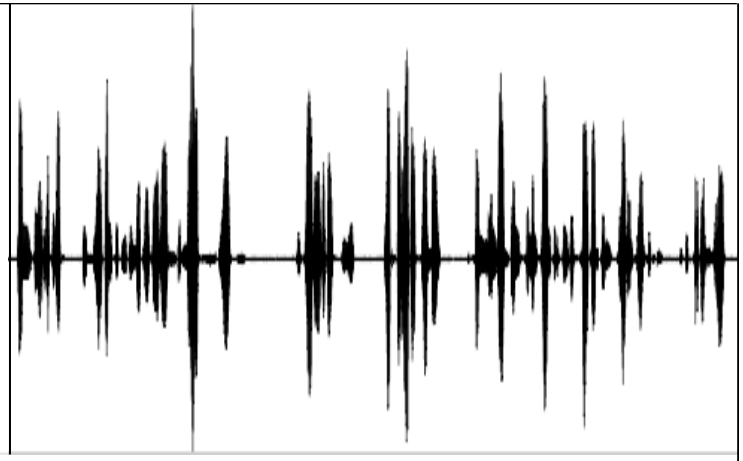
LSTM: First RNN to win contests:
3 ICDAR 2009 connected handwriting competitions

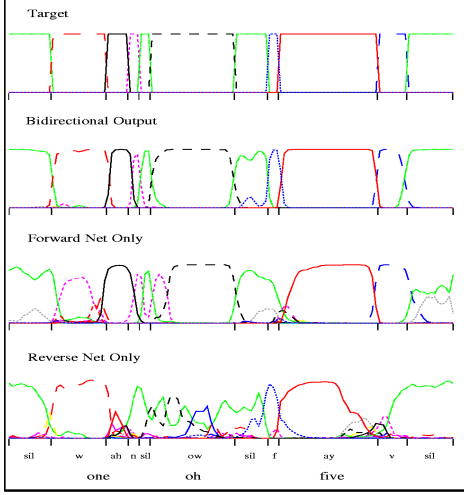
http://www.idsia.ch/~juergen/handwriting.html

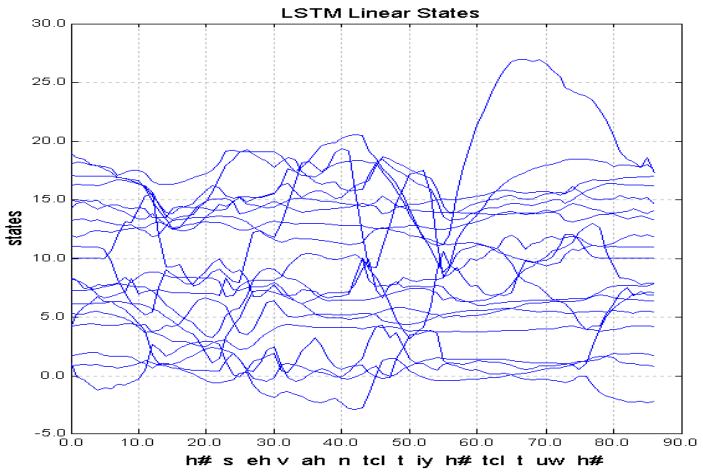


E.g., Graves & Schmidhuber NIPS 2010

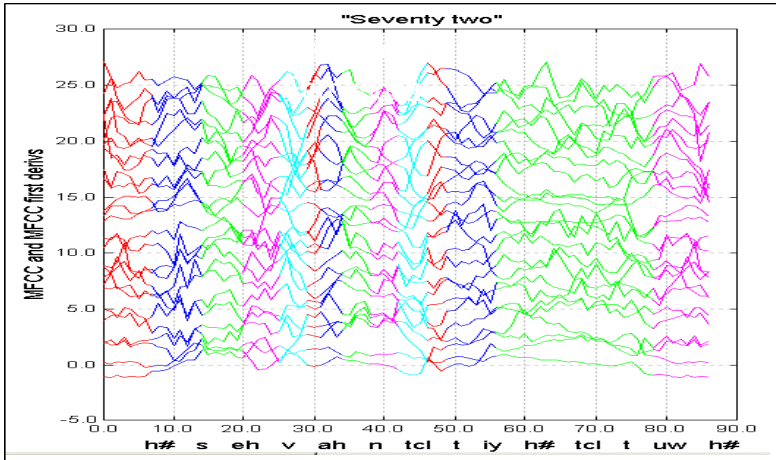


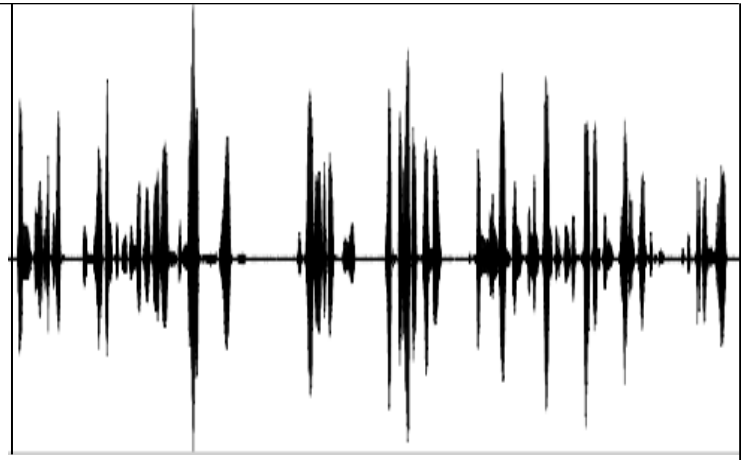






LSTM for speech: 2003 as good as HMMs, 2007: LSTM stack gets best results on keyword spotting in a large corpus (vs HMMs). Today: best large vocabulary speech recognition …





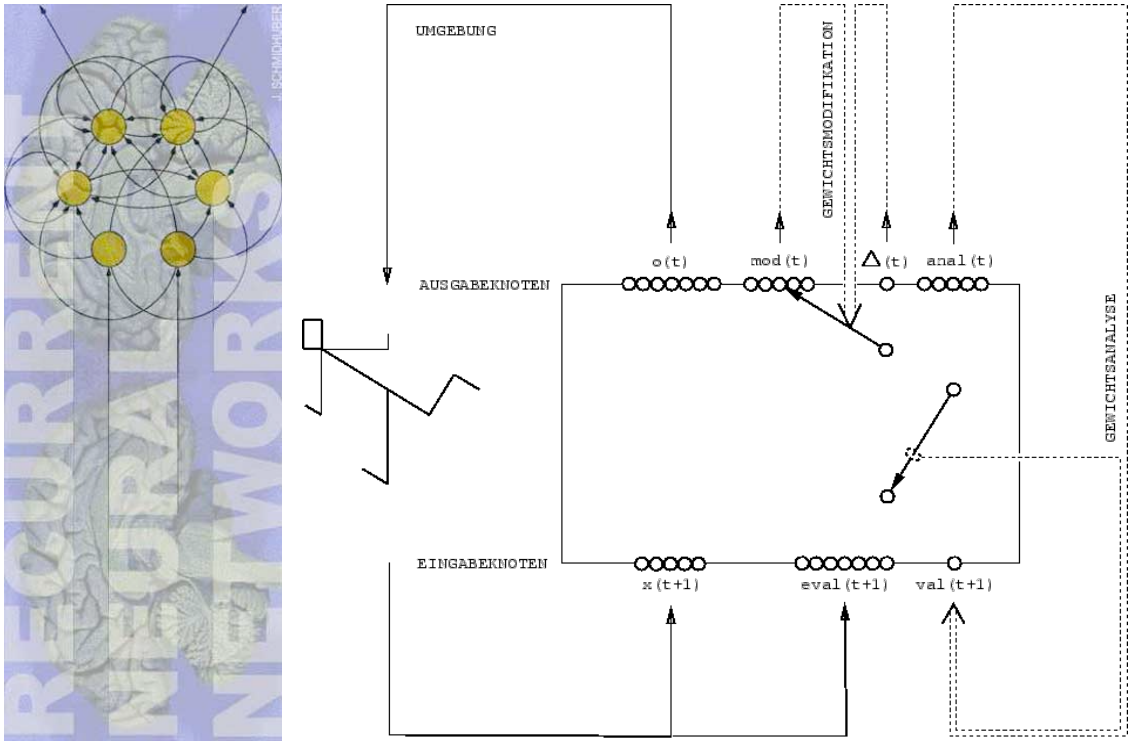
Since 2003/2007: Speech Recognition Revolution through RNNs, e.g., Graves et al, ICASSP 2013: best results on TIMIT through LSTM. Google 2014: best large vocabulary speech rec.

BAIDU's DeepSpeech uses our CTC-based RNNs for end-to-end speech recognition without any HMMs / GMMs (Hannun et al., Baidu, 2014); broke Switchboard benchmark record

A dozen of the many 2014/2015 benchmark records with LSTM RNNs, often at major IT companies:

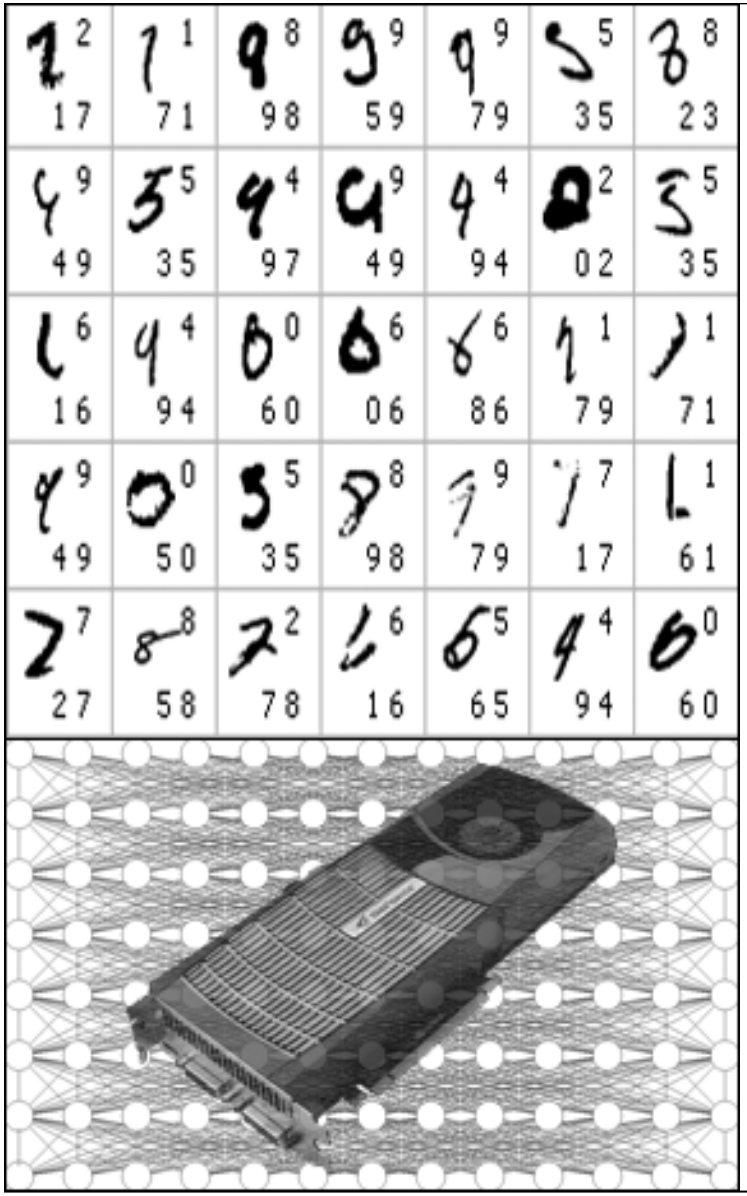
1. Large vocabulary speech recognition (Sak et al., Google, Interspeech 2014)
2. English to French translation (Sutskever et al., Google, NIPS 2014)
3. Text-to-speech synthesis (Fan et al., Microsoft, Interspeech 2014)
4. Prosody contour prediction (Fernandez et al., IBM, Interspeech 2014)
5. Google Voice improved by 49% (Sak et al, 2015, now for >1 billion users)
6. Syntactic parsing for NLP (Vinyals et al., Google, 2014)
7. Photo-real talking heads (Soong and Wang, Microsoft, ICASSP 2015)
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. Keyword spotting (Chen et al., Google, ICASSP 2015)
12. Video to textual description (Donahue et al., 2014; Li Yao et al., 2015)

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



1993: Gradient-based meta-RNNs that can learn to run their own weight change algorithm:
J. Schmidhuber.
A self-referential weight matrix.
ICANN 1993

This was before LSTM. In 2001, however, Hochreiter taught a meta-LSTM to learn a learning algorithm for quadratic functions that was faster than backprop



MNIST: 60,000 digits for training, 10,000 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% (Ciresan et al.) Since then: decline of unsupervised pre-training for FNNs, like in the 1990s for RNNs

Two old ideas: backprop (3-5 decades old), training pattern deformations (Baird, 1990, 2 decades old)

Unsupervised ➔ Supervised

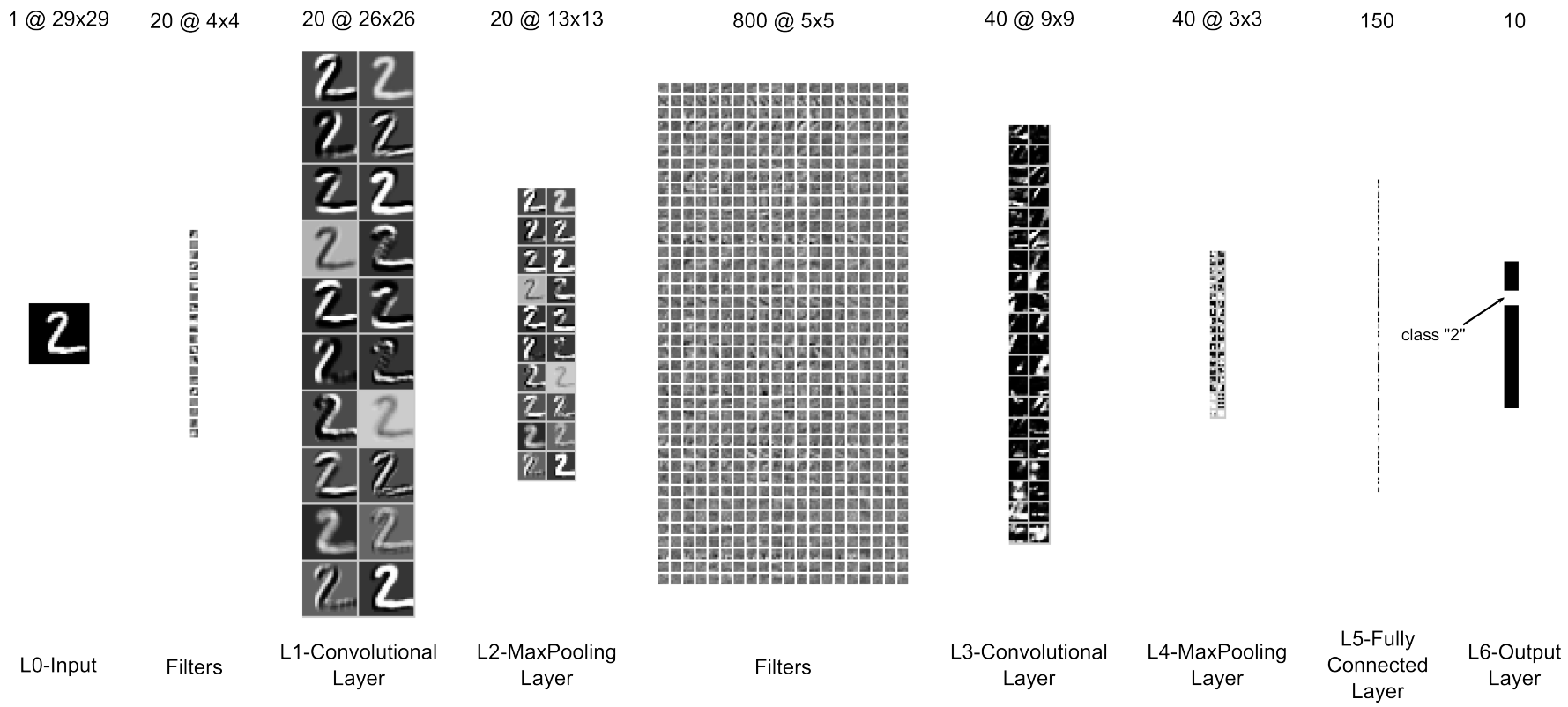
1990s: Trend from unsupervised to supervised RNNs

2000s: Trend from unsupervised to supervised FNNs

Both trends driven by our team

Our Deep GPU-Based Max-Pooling CNNs (IJCAI 2011)

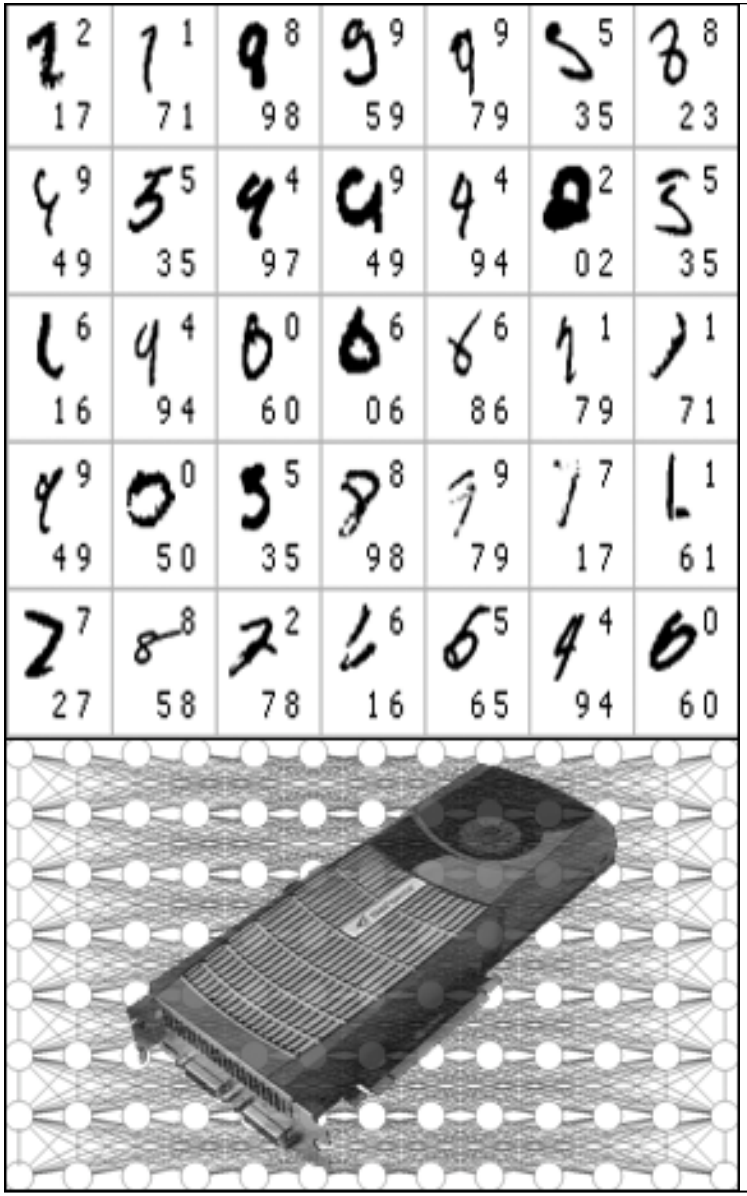
e.g., http://www.idsia.ch/~juergen/deeplearning.html



Alternating convolutional and subsampling layers (CNNs): Fukushima 1979. Backprop for CNNs: LeCun et al 1989. Max-pooling (MP): Weng 1992. Backprop for MPCNNs: Ranzato et al 2007, Scherer et al 2010, GPU-MPCNNs - Ciresan et al (Swiss AI Lab IDSIA, 2011)

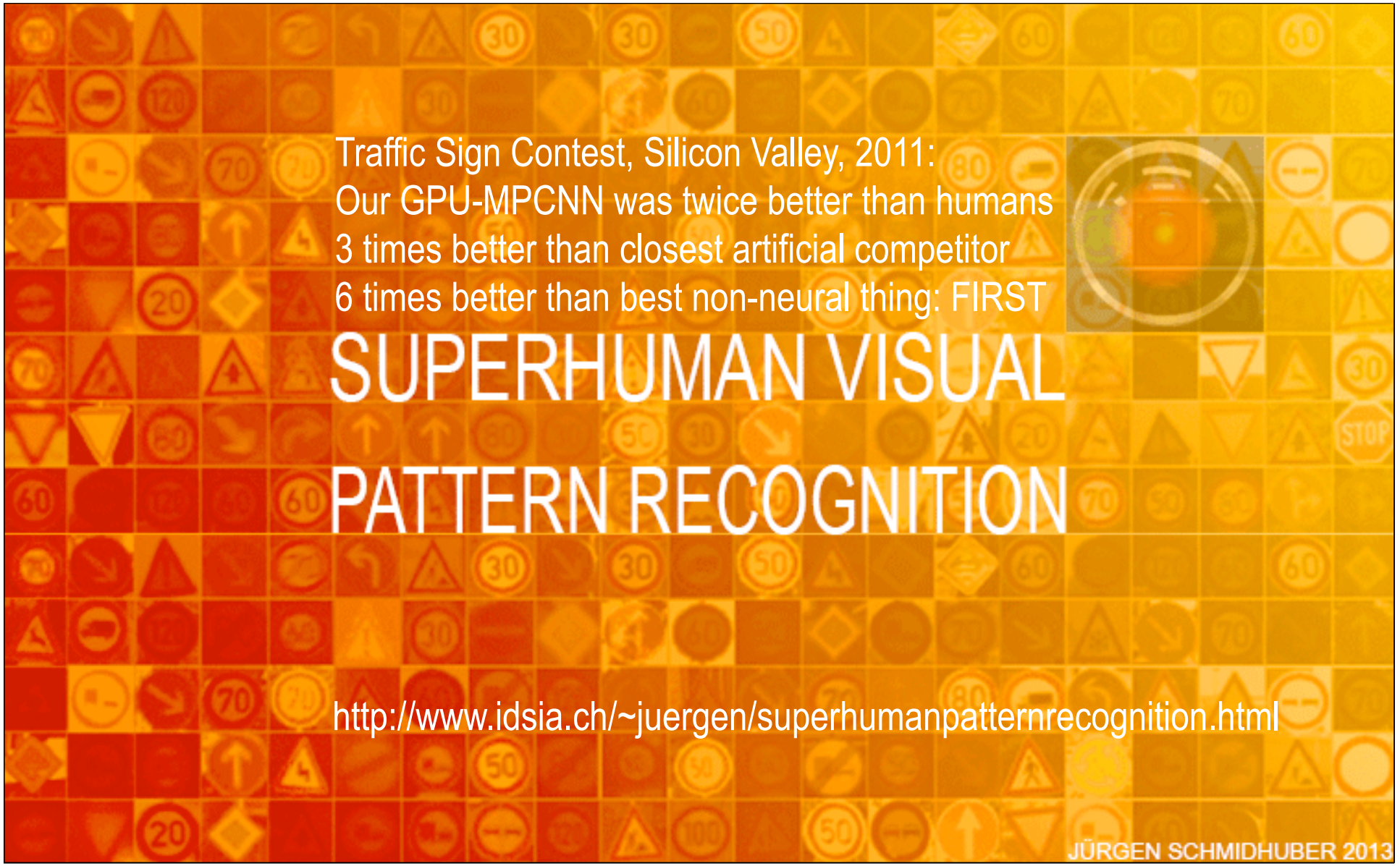
ICDAR 2011 offline Chinese handwriting recognition contest (4000 classes):
1st & 2nd rank
Oct 2013: again best results, first near-human performance
(Ciresan & Schmidhuber, 2013)

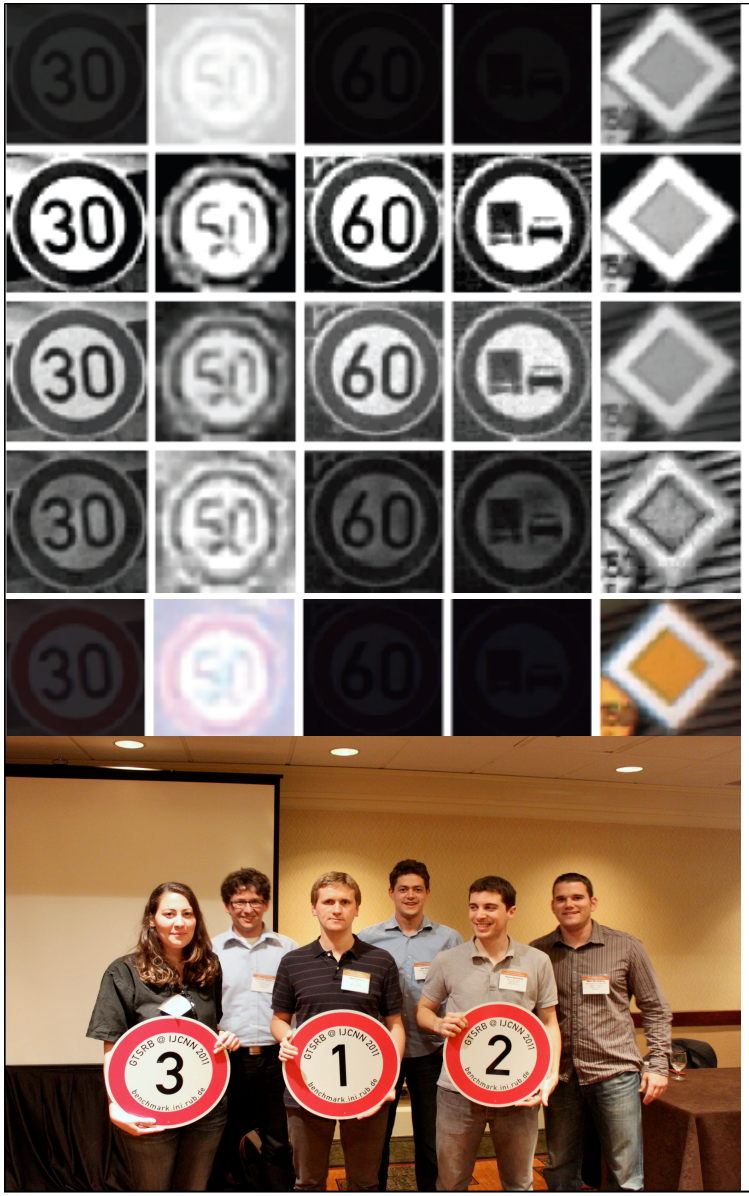
http://www.idsia.ch/~juergen/handwriting.html



Ensembles of deep sparse CNNs + Max-Pooling + MLP on top: 1 year on CPU = 1 week on GPU. 2011-2012: first human-competitive MNIST result: 0.2% (after almost a decade of ~0.4%).

Ciresan, Meier, Masci, Gambardella, Schmidhuber, IJCAI 2011, IJCNN 2011, CVPR 2012


Traffic Sign Contest, Silicon Valley, 2011:
Our GPU-MPCNN was twice better than humans
3 times better than closest artificial competitor
6 times better than best non-neural thing: FIRST
SUPERHUMAN VISUAL
PATTERN RECOGNITION
http://www.idsia.ch/~juergen/superhumanpatternrecognition.html
JÜRGEN SCHMIDHUBER 2013



IJCNN 2011 traffic sign recognition competition, Silicon Valley, 2011:
1ST (0.56% ERROR)
2ND HUMANS (1.16%)
3RD (1.69%)
4TH (3.86%)
Ciresan, Meier, Masci, Schmidhuber, IJCNN 2011, Neural Networks, 2012

Very similar GPU-MPCNNs later used for ImageNet (Krizhevsky & Hinton 2012, Zeiler & Fergus 2013, …)

Robot Cars

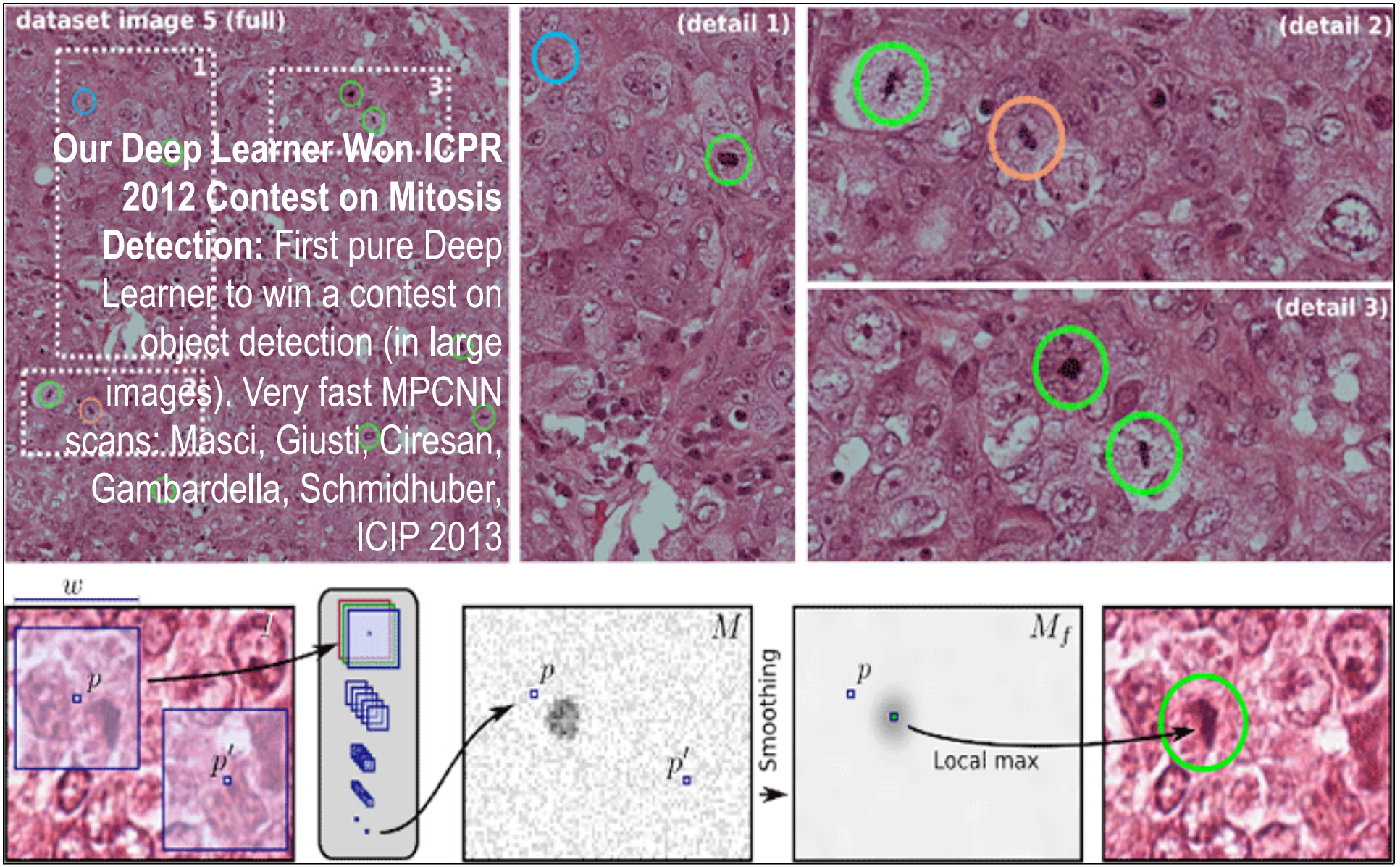
http://www.idsia.ch/~juergen/robotcars.html

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

Ernst Dickmanns, *the* robot car pioneer, Munich, 80s


Our Deep Learner
Won ISBI 2012 Brain
Image Segmentation
Contest:
First feedforward
Deep Learner to win
an image
segmentation
competition
(but compare deep
recurrent LSTM
2009: segmentation
& classification)
JS 2013
http://www.idsia.ch/~juergen/deeplearningwinsbraincontest.html


dataset image 5 (full)
1
3
(detail 1)
(detail 2)
(detail 3)
Our Deep Learner Won 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
w
I
p
p′
M
Smoothing
M_f
Local max

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

Thanks to Dan Ciresan & Alessandro Giusti

Some of Our Deep Learning "Firsts"

- First recurrent NN to win contests (2009)
- First NN to win connected handwriting contests (2009)
- First outperformance of humans in a computer vision contest (2011)
- First deep NN to win Chinese handwriting contest (2011)
- European handwriting (MNIST): old error record almost halved (2011)
- First deep NN to win image segmentation contest (2012)
- First deep NN to win object detection contest (2012)
- First deep NN to win medical imaging contest (2012)
- First RNN controller that reinforcement learns from raw video (2013)
- …

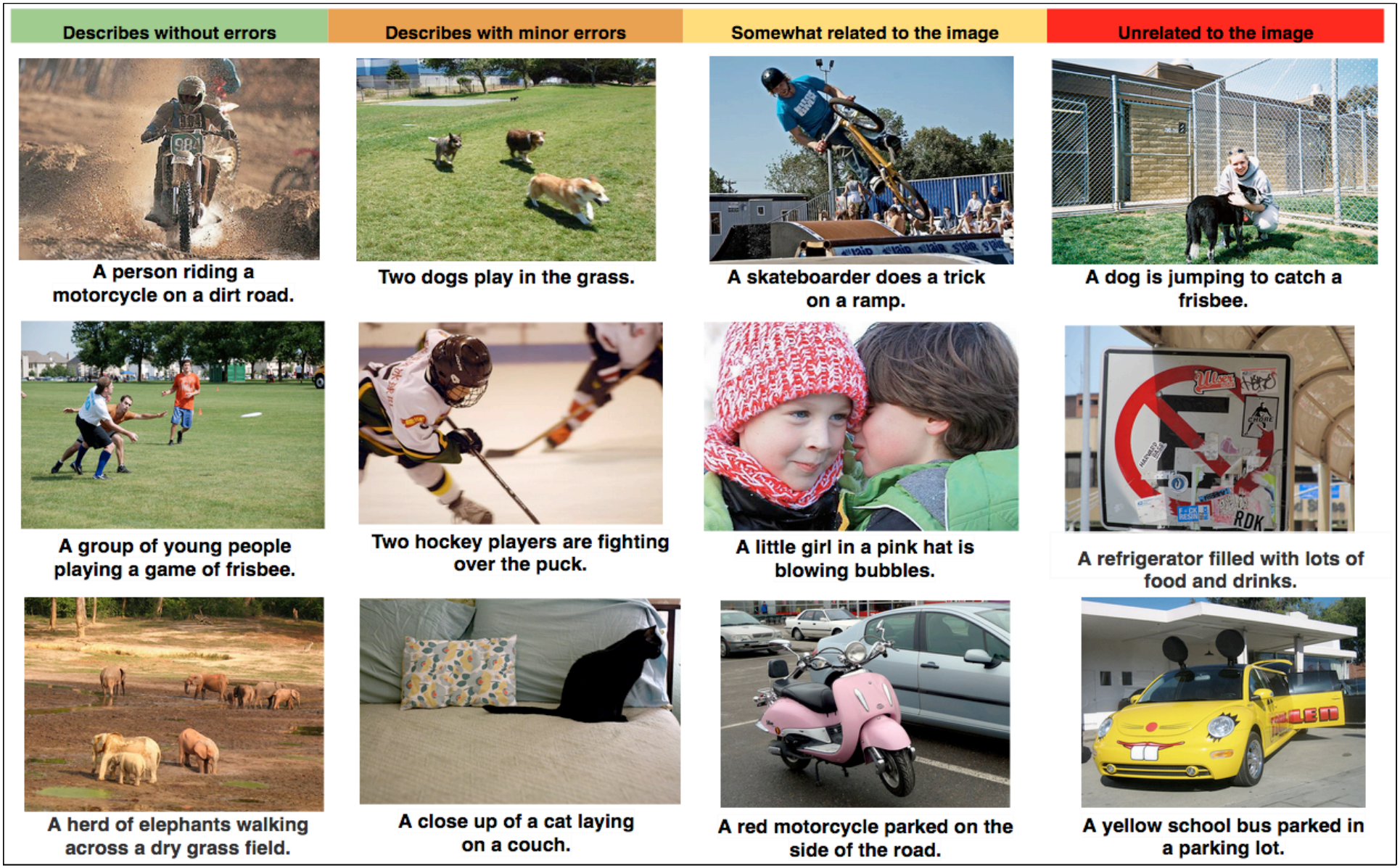
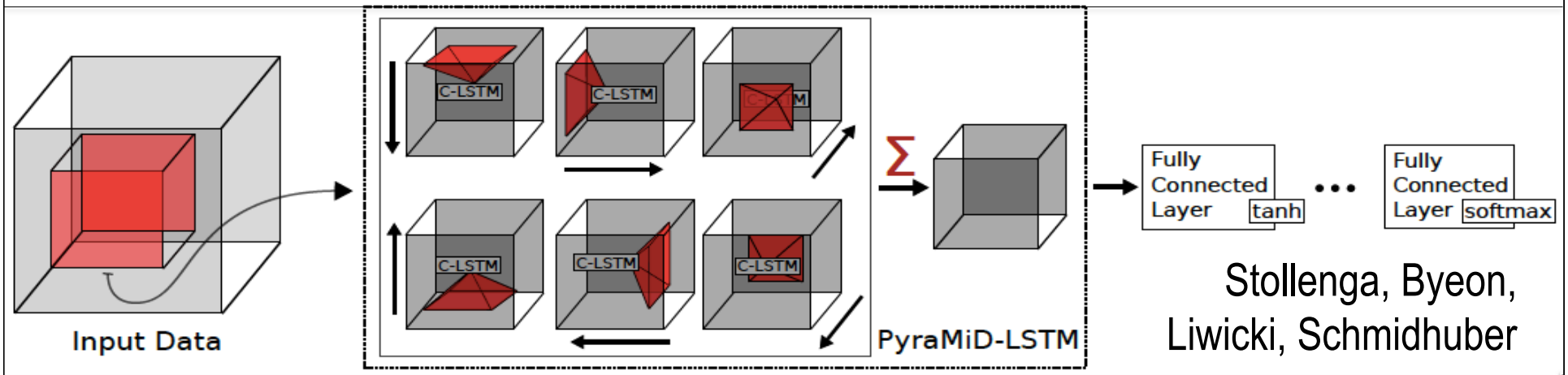
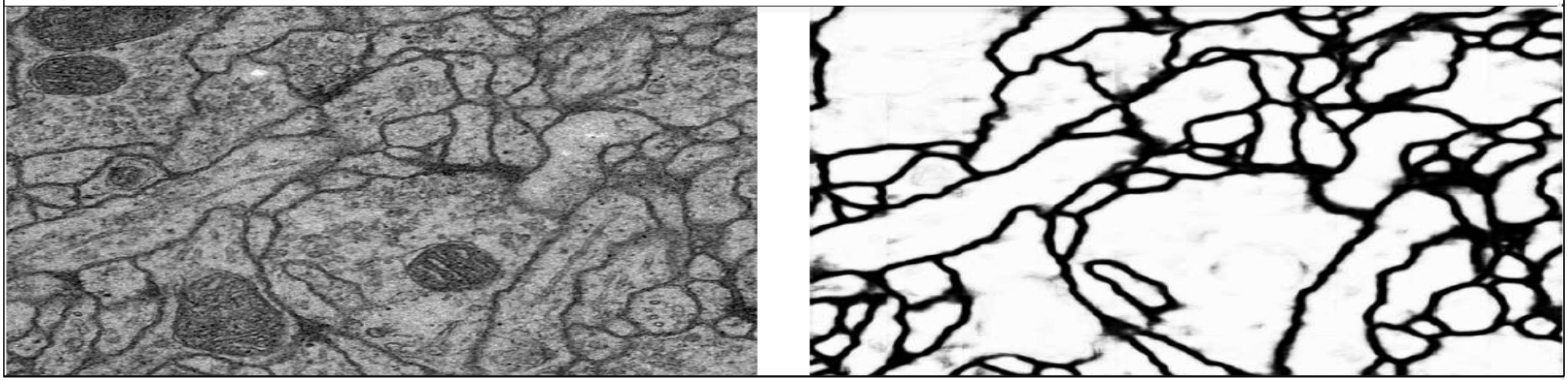


Image caption generation with LSTM RNNs translating internal representations of CNNs
(Vinyals, Toshev, Bengio, Erhan, Google, 2014)

Best Segmentation with PyramMiD-LSTM (NIPS 2015)



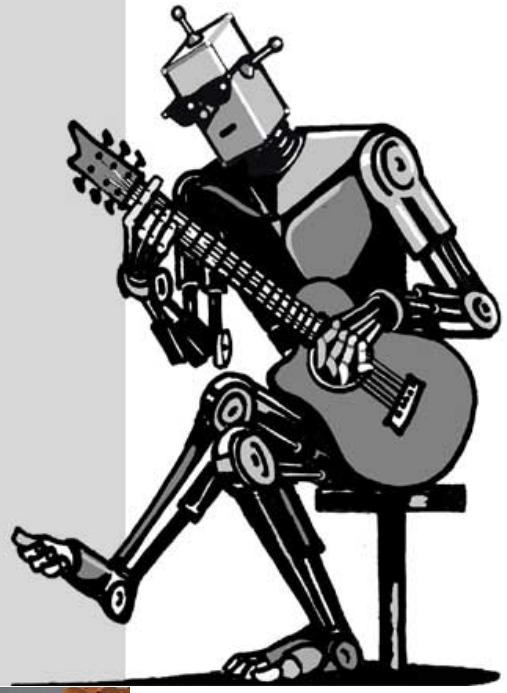
Stollenga, Byeon, Liwicki, Schmidhuber



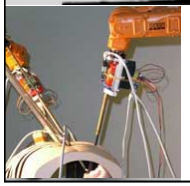
TU MUNICH

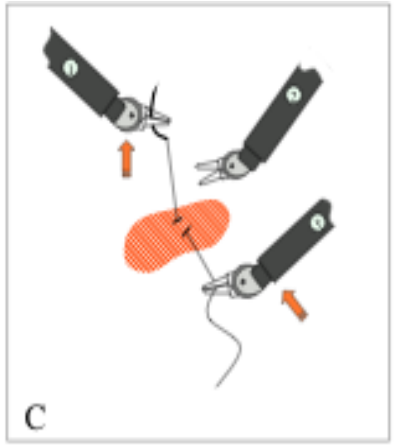
LSTM learns knot-tying tasklets:
Mayr Gomez Wierstra Nagy Knoll
Schmidhuber, IROS'06

J SCHMIDHUBER 2006

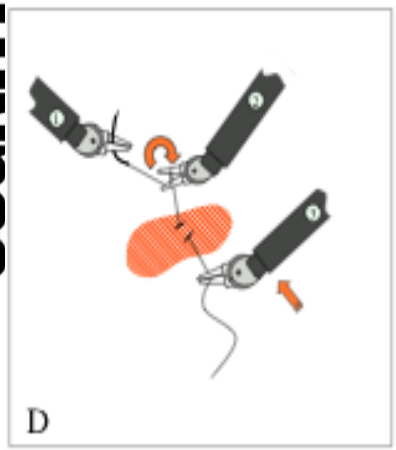


COGNITIVE ROBOTICS

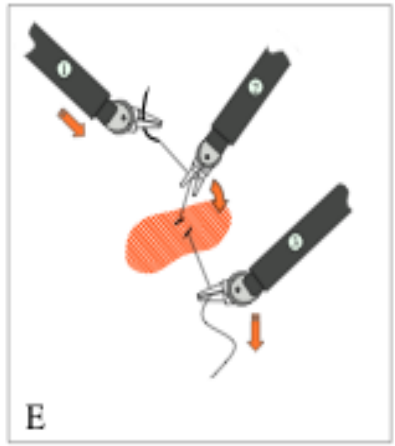




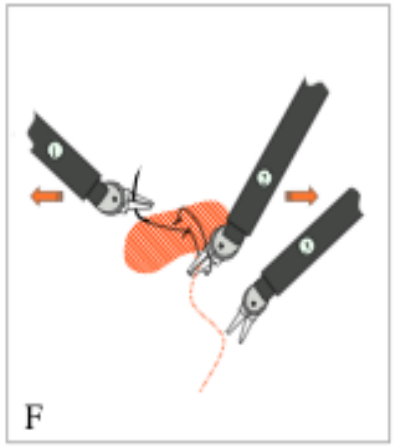
C



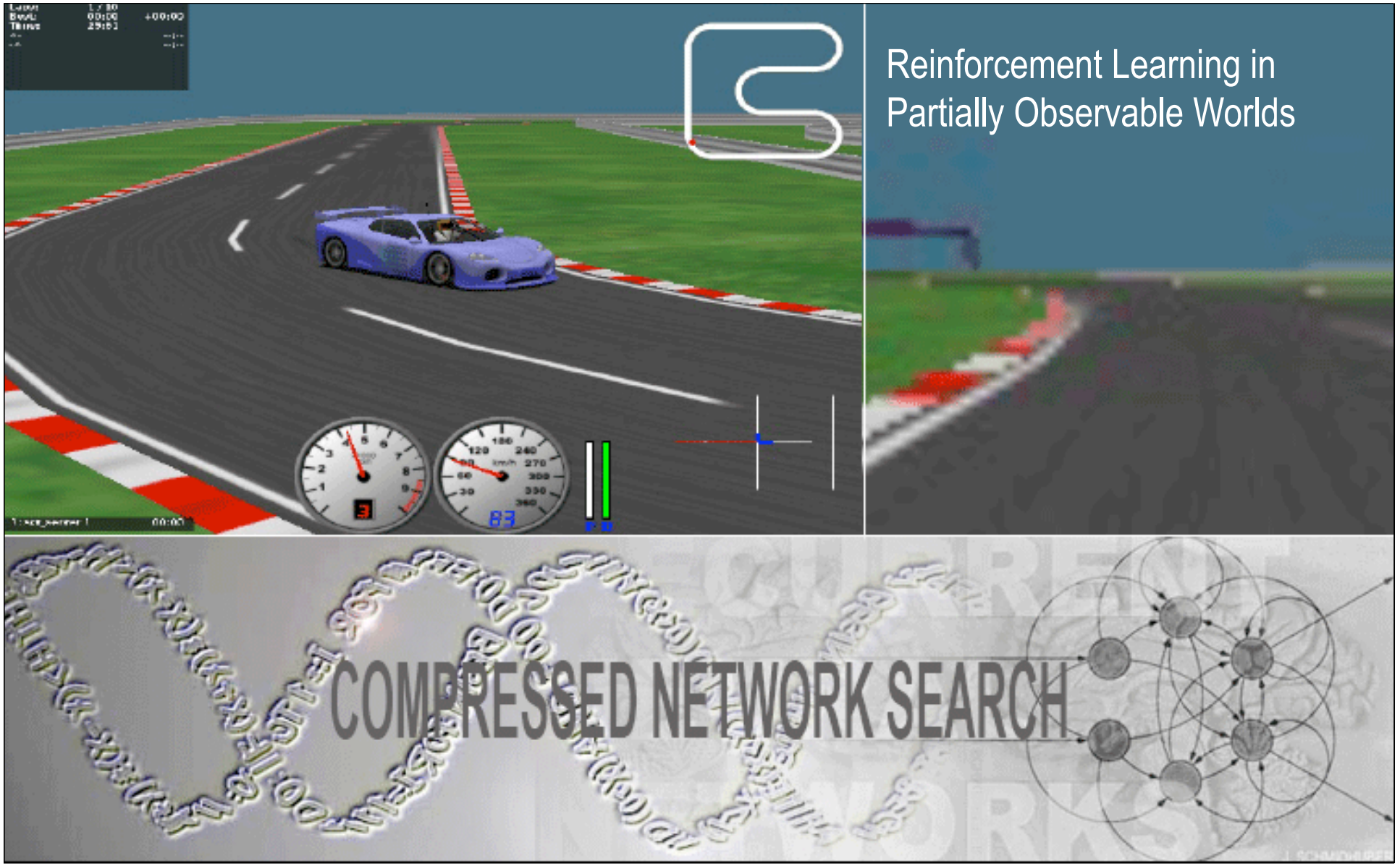
D



E



F



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

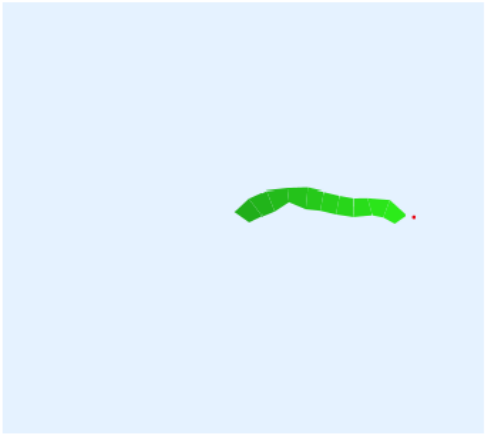
Faustino Gomez, Jan Koutnik, Giuseppe Cuccu, J. Schmidhuber, GECCO 2013

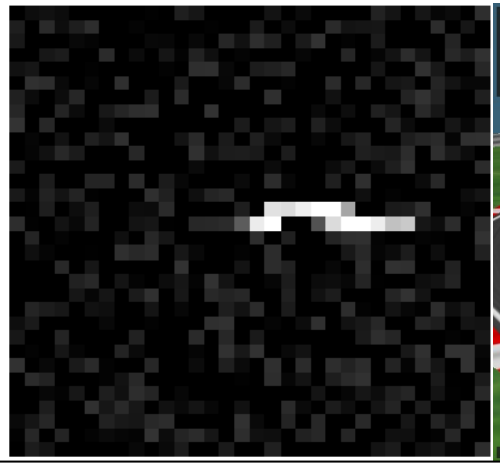

http://www.idsia.ch/~juergen/compressednetworksearch.html

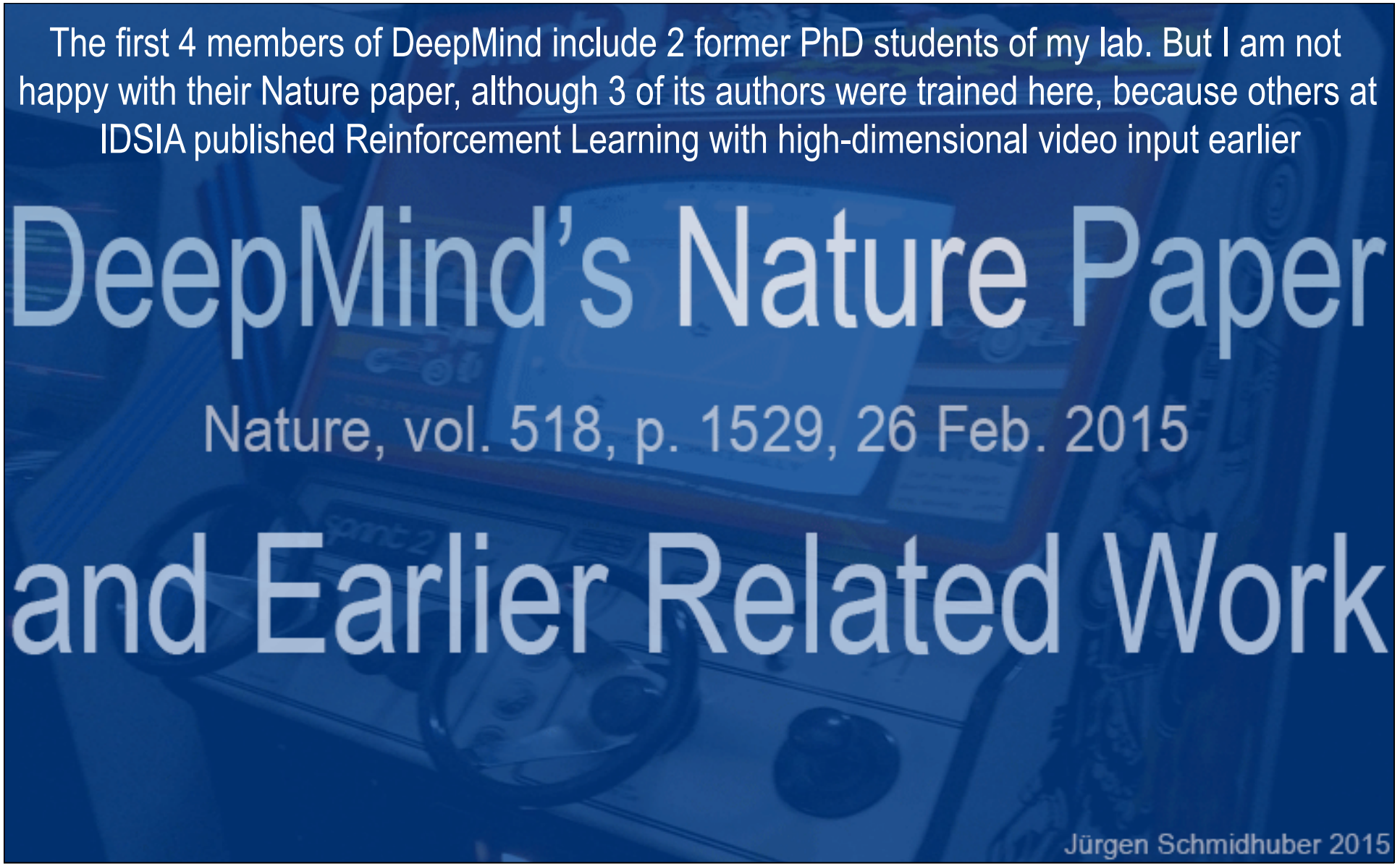
Octopus-arm control: 82 in, 32 out, 3'680 weights, only 20 DCT coefficients, compression 1:184

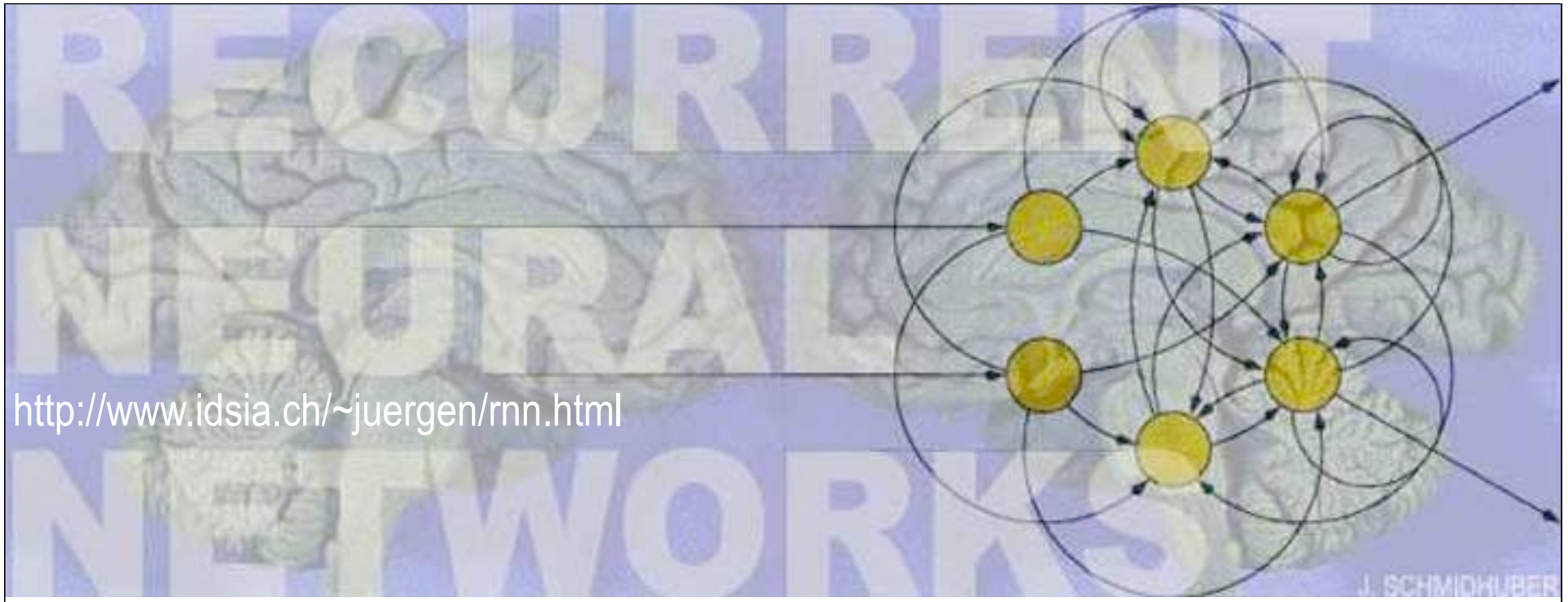
Octopus-arm with low-level vision, 32x32 in, 32 out, **33'824** weights, **160** DCT, compression **1:211**

TORCS driving video game, low-level vision, 64x64 in, 3 out, **1'115'139** weights, **200** DCT, compression **1:5575**






The first 4 members of DeepMind include 2 former PhD students of my lab. But I am not happy with their Nature paper, although 3 of its authors were trained here, because others at IDSIA published Reinforcement Learning with high-dimensional video input earlier
DeepMind's Nature Paper
Nature, vol. 518, p. 1529, 26 Feb. 2015
and Earlier Related Work
Jürgen Schmidhuber 2015



No new NN winter, because physics dictates that future hardware will be 3D-RNN-like: many processors connected by many short and few long wires

IJCNN 1990, NIPS 1991: Reinforcement Learning with Recurrent Controller & Recurrent World Model

A bit like AIXI, but with feasible local search

IJNS 1991: R-Learning of Visual Attention on 1,000,000 times slower computers

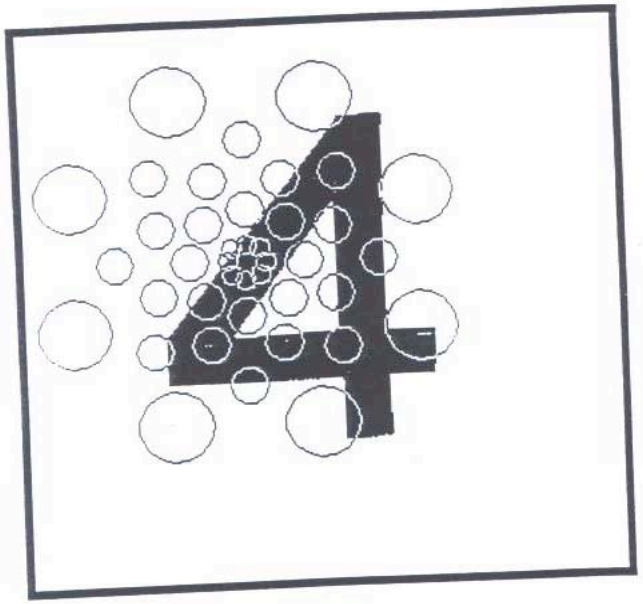


Fig. 1. A typical visual scene. The diameters of the receptive fields of the retina's input units are indicated by circles.

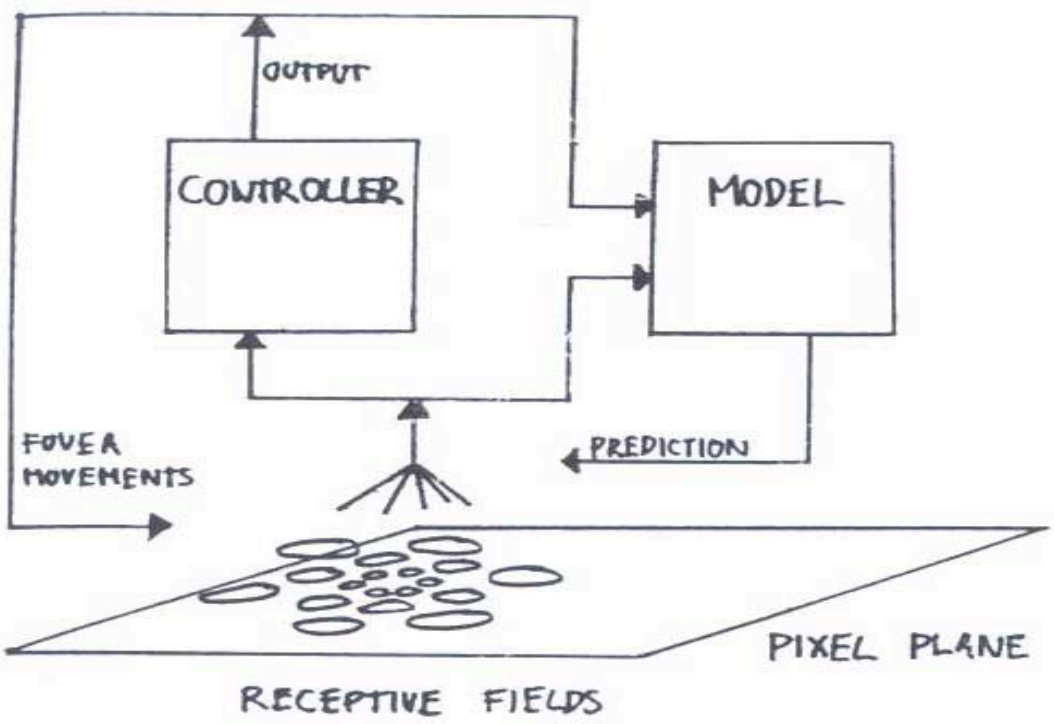
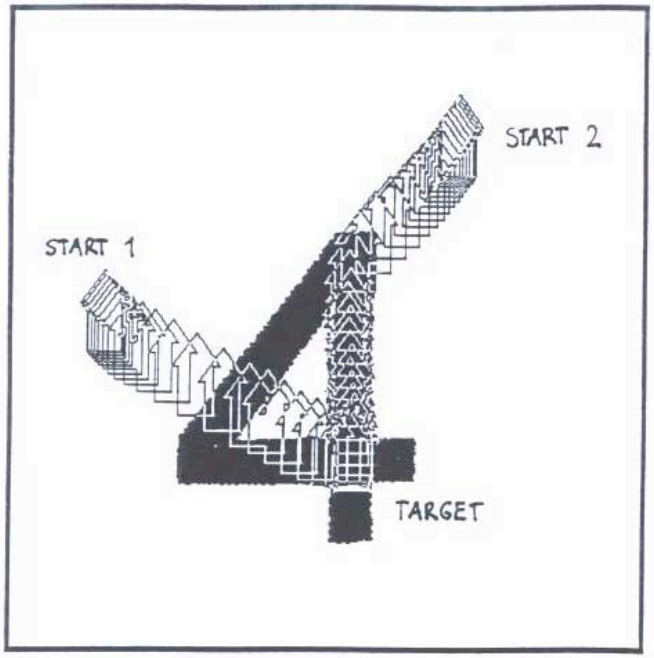


Fig. 2. An artificial fovea provides inputs for a control network which is able to move the fovea around. A model network is trained to predict the next input from the current input and the current controller action.

1991: current goal=extra fixed input
2015: all of this is coming back!



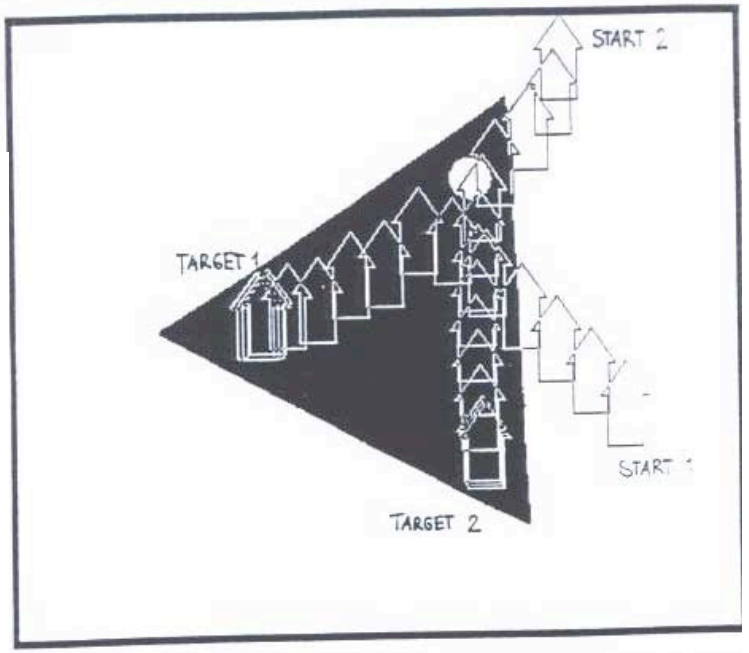
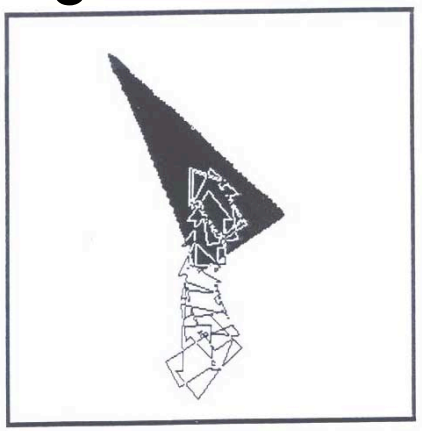
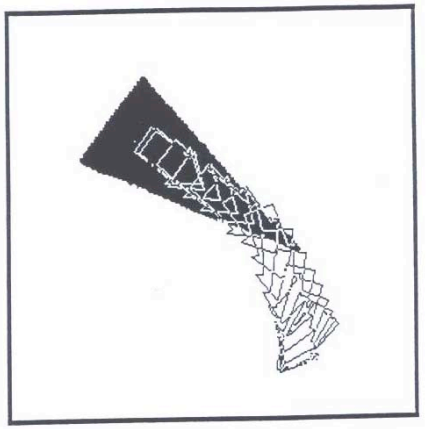


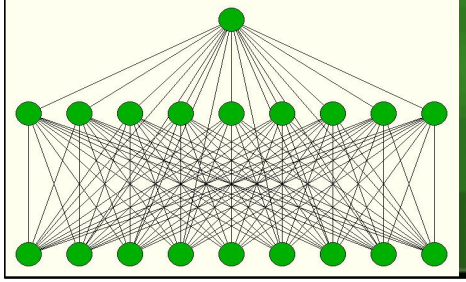
Fig. 5. One controller for various targets specified by ar additional constant input: Examples of fovea trajectorie leading from various start positions to different targets. Th first target is near the left corner of the triangle. The secon target is near the lower corner.

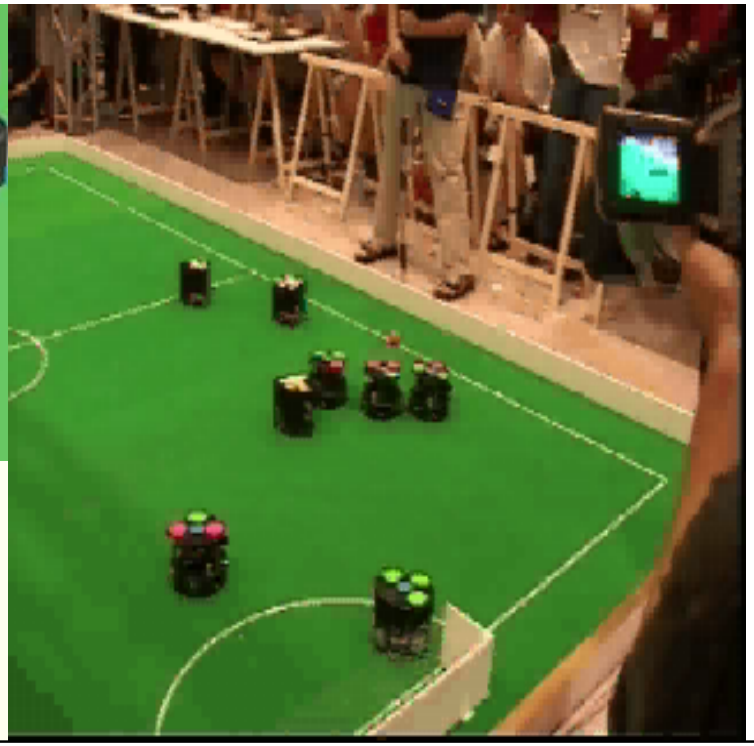




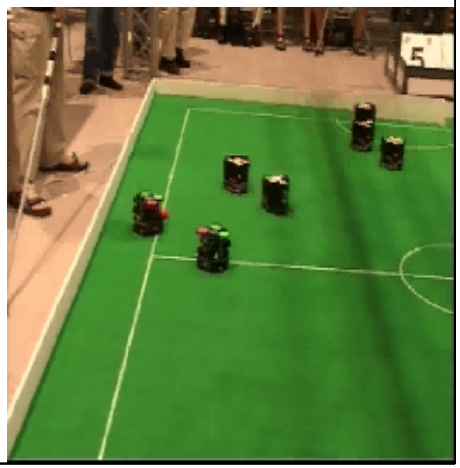
RoboCup World Champion 2004, Fastest League, 5m/s

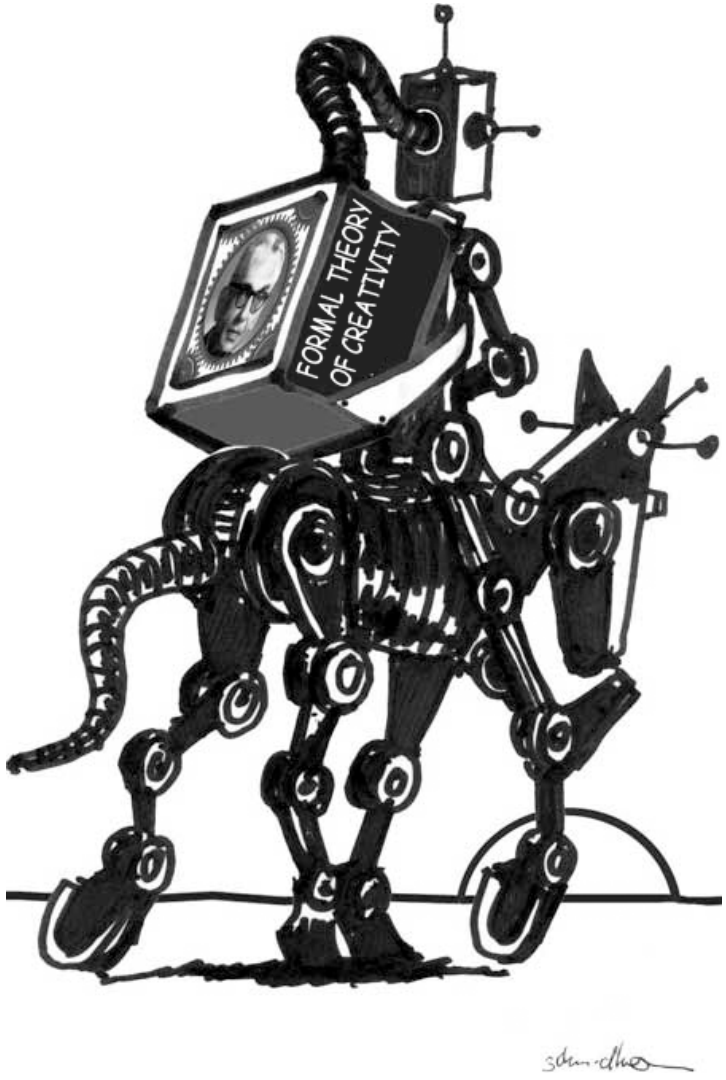
Lookahead expectation & planning with neural networks (Schmidhuber, IEEE INNS 1990): successfully used for RoboCup by Alexander Gloye-Förster (went to IDSIA)
http://www.idsia.ch/~juergen/learningrobots.html





Alex @ IDSIA, led FU Berlin's RoboCup World Champion Team 2004





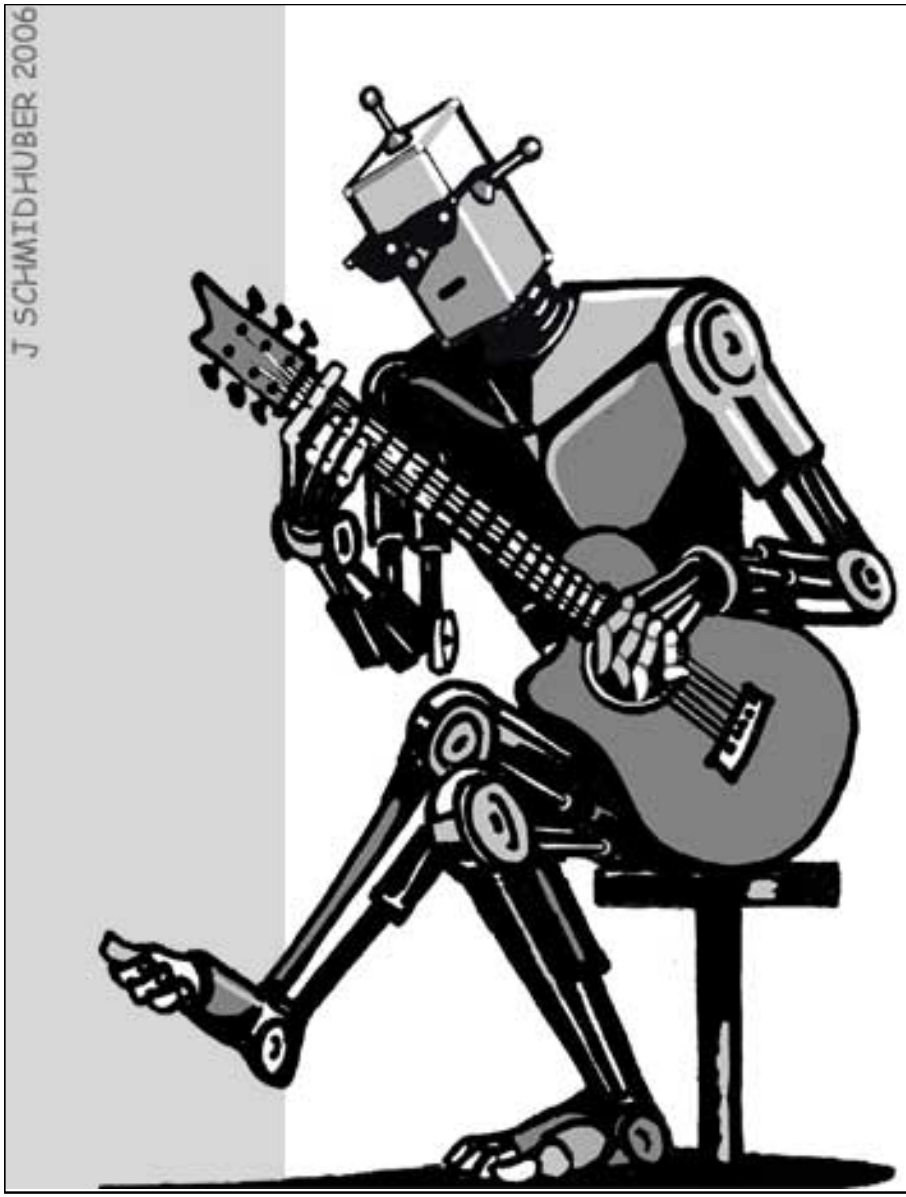
Maximize Future Fun(Data X,O(t))~
∂CompResources(X,O(t))/∂t

Formal theory of fun & novelty & surprise & attention & creativity & curiosity & art & science & humor

E.g., Connection Science 18(2):173-187, 2006
IEEE Transactions AMD 2(3):230-247, 2010
http://www.idsia.ch/~juergen/creativity.html

PowerPlay not only solves but also continually invents problems at the borderline between what's known and unknown - training an increasingly general problem solver by continually searching for the simplest still unsolvable problem

POWERPLAY



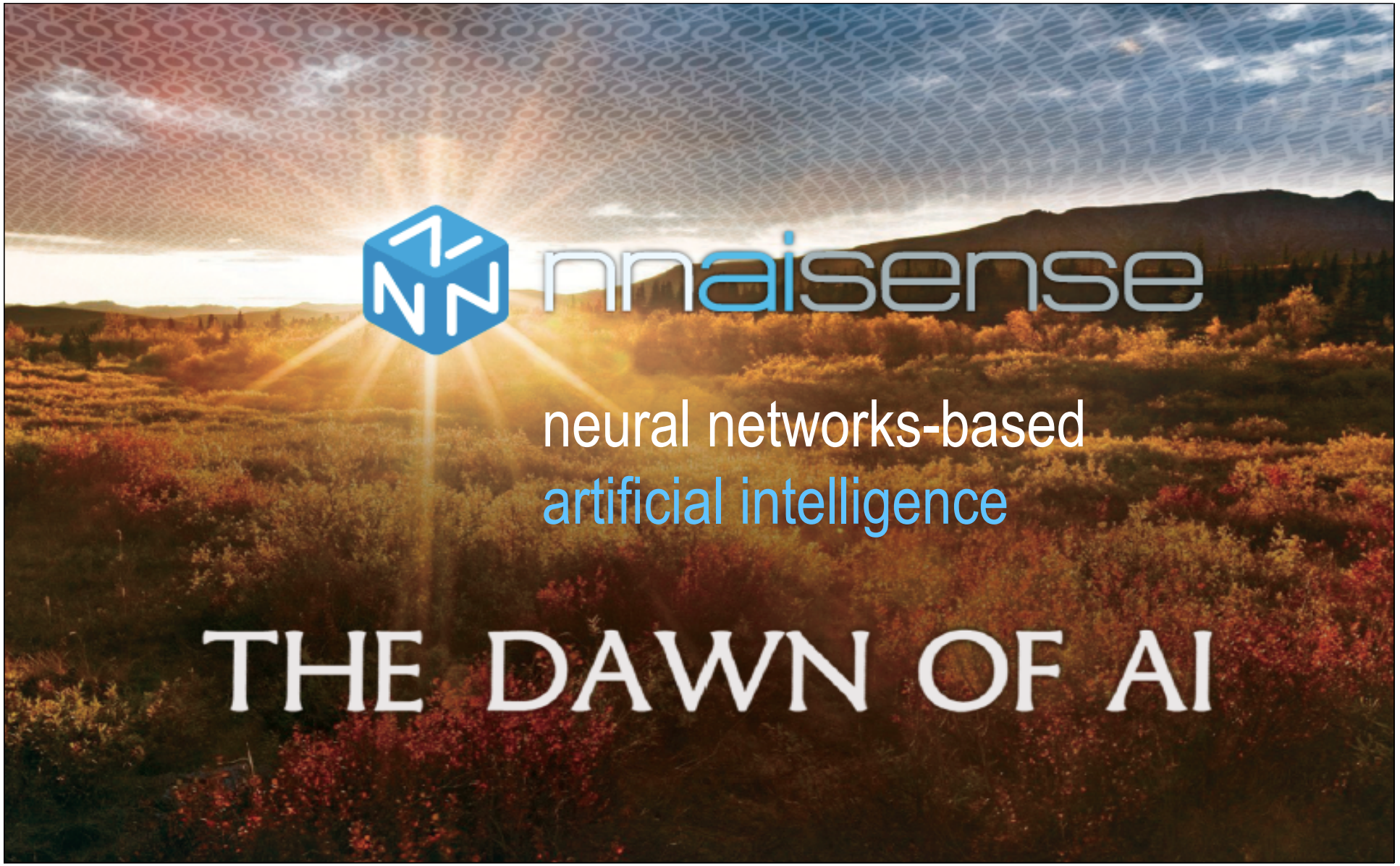
True Artificial Intelligence Will Change Everything

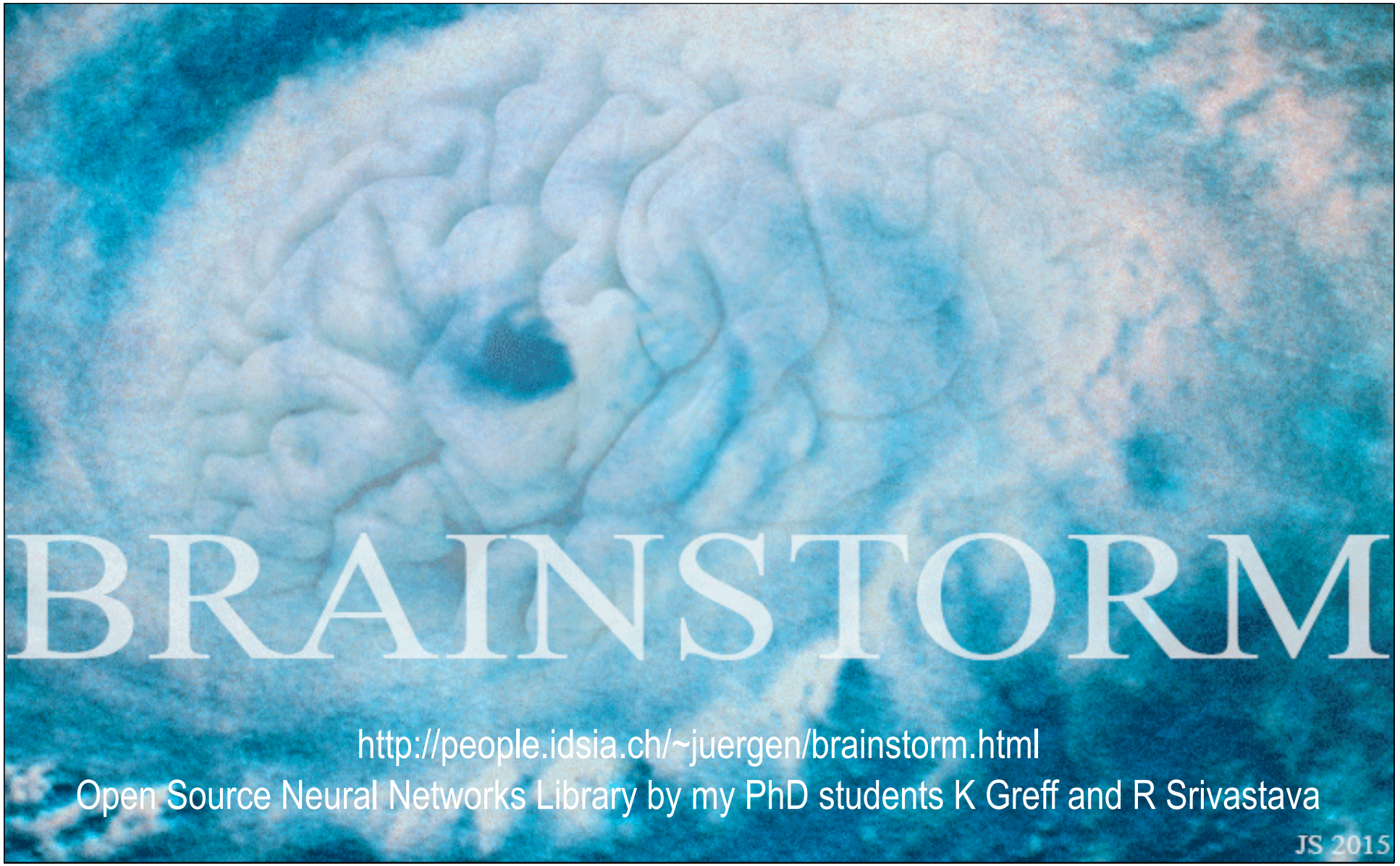
Jürgen Schmidhuber
The Swiss AI Lab IDSIA
Univ. Lugano & SUPSI
http://www.idsia.ch/~juergen

NNAISENSE

Next: build small animal-like AI that learns to think and plan hierarchically like a crow or a capuchin monkey

Evolution needed billions of years for this, then only a few more millions for humans


nnaisense
neural networks-based
artificial intelligence
THE DAWN OF AI


BRAINSTORM
http://people.idsia.ch/~juergen/brainstorm.html
Open Source Neural Networks Library by my PhD students K Greff and R Srivastava
JS 2015
