UNIVERSITÉ Grenoble Alpes

cnrs

IUF

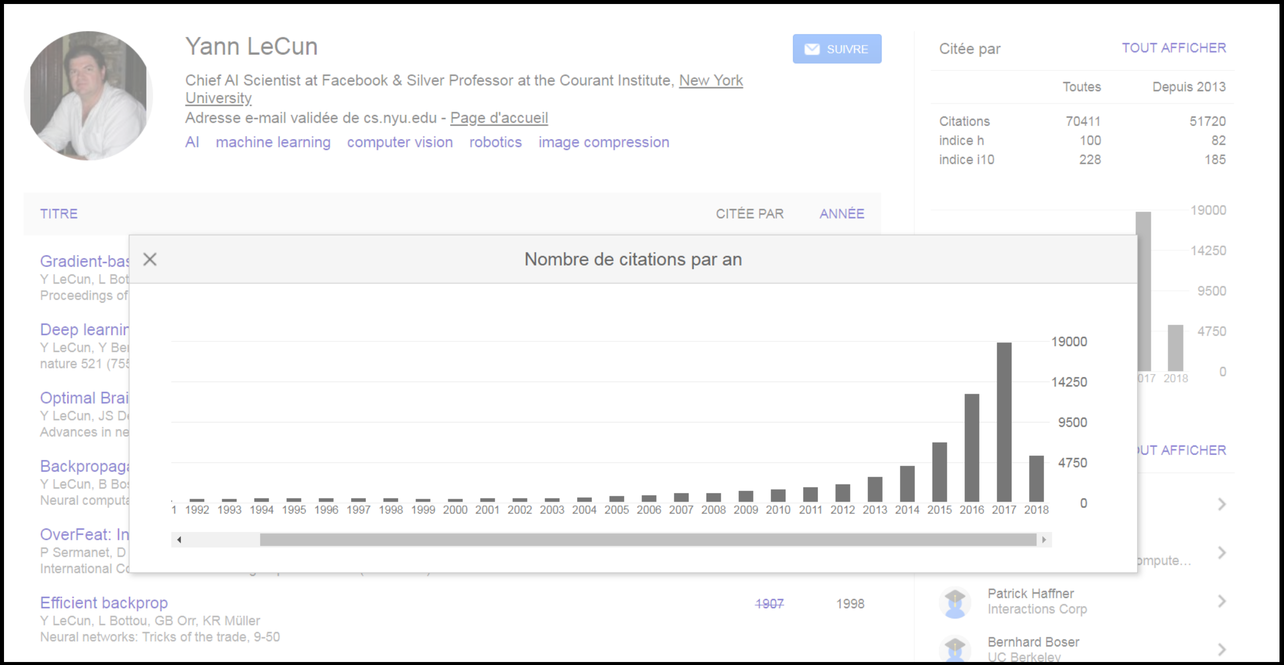
Are Deep Neural Networks Bio-Inspired and To What Extend Could They Reach a Basic Level of Self-Awareness?

Martial Mermillod

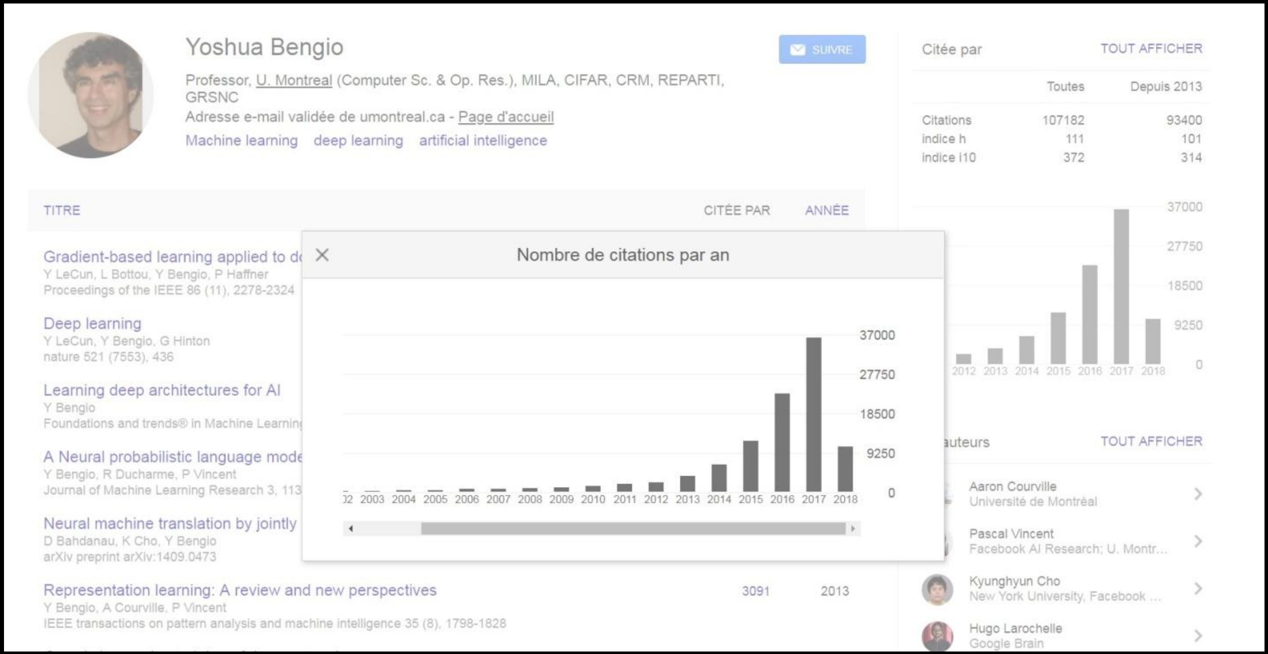
LPNC
Laboratoire de Psychologie et NeuroCognition

The buzz since december 2012 !

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).



Yann LeCun
SUIVRE
Chief AI Scientist at Facebook & Silver Professor at the Courant Institute, New York University
Adresse e-mail validée de cs.nyu.edu - Page d'accueil
AI machine learning computer vision robotics image compression
Citée par TOUT AFFICHER
Toutes Depuis 2013
Citations 70411 51720
indice h 100 82
indice i10 228 185
TITRE CITÉE PAR ANNÉE
Nombre de citations par an
19000
14250
9500
4750
0
1 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018
Efficient backprop 1907 1998
Y LeCun, L Bottou, GB Orr, KR Müller
Neural networks: Tricks of the trade, 9-50
Patrick Haffner
Interactions Corp
Bernhard Boser



Yoshua Bengio
SUIVRE
Professor, U. Montreal (Computer Sc. & Op. Res.), MILA, CIFAR, CRM, REPARTI, GRSNC
Adresse e-mail validée de umontreal.ca - Page d'accueil
Machine learning deep learning artificial intelligence
Citée par TOUT AFFICHER
Toutes Depuis 2013
Citations 107182 93400
indice h 111 101
indice i10 372 314
TITRE CITÉE PAR ANNÉE
Gradient-based learning applied to d
Y LeCun, L Bottou, Y Bengio, P Haffner
Proceedings of the IEEE 86 (11), 2278-2324
Deep learning
Y LeCun, Y Bengio, G Hinton
nature 521 (7553), 436
Learning deep architectures for AI
Y Bengio
A Neural probabilistic language mode
Y Bengio, R Ducharme, P Vincent
Neural machine translation by jointly
D Bahdanau, K Cho, Y Bengio
arXiv preprint arXiv:1409.0473
Nombre de citations par an
37000
27750
18500
9250
0
2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018
Representation learning: A review and new perspectives 3091 2013
Y Bengio, A Courville, P Vincent
IEEE transactions on pattern analysis and machine intelligence 35 (8), 1798-1828
TOUT AFFICHER
Aaron Courville
Université de Montréal
Pascal Vincent
Kyunghyun Cho
Hugo Larochelle
Google Brain

Why on December 2012 ?

3 factors:

- Deep Neural Networks ready for a while

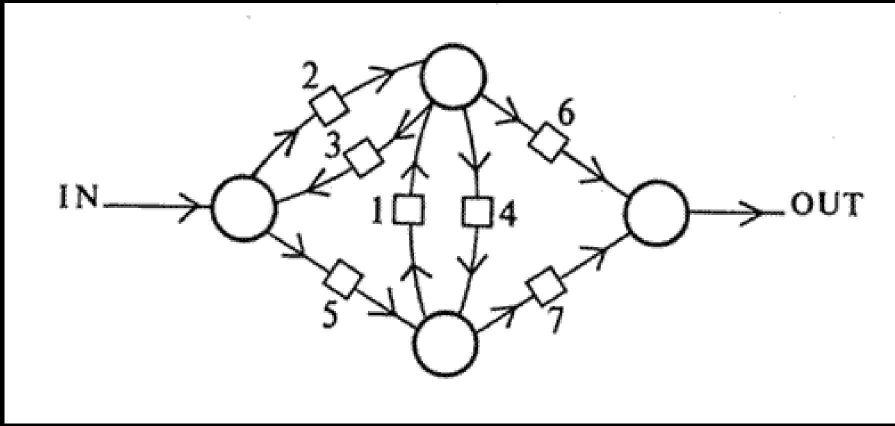
- GAFA -> BIG DATA

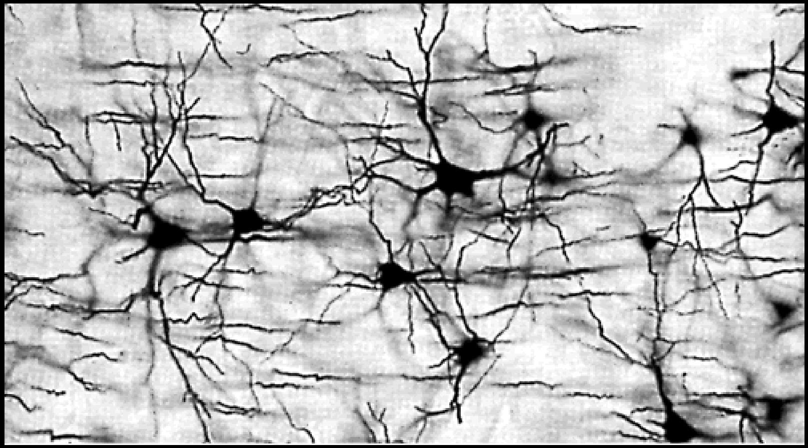
- GPU -> convolution/pooling

To what extend Deep Learning is inspired from a human brain?

The Mainstream: Turing-Von Neumann
And…
The parallel History !

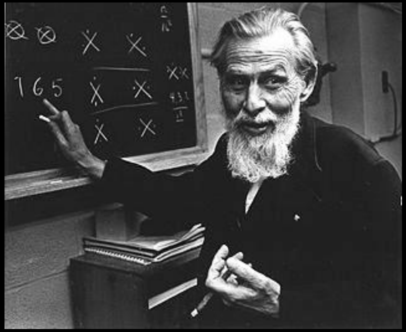
Turing's unorganized machines (1948)

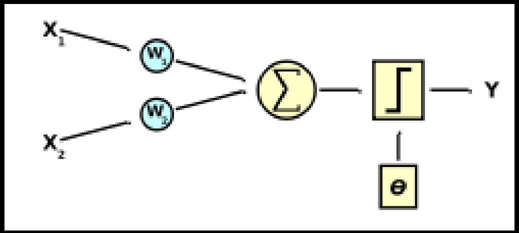


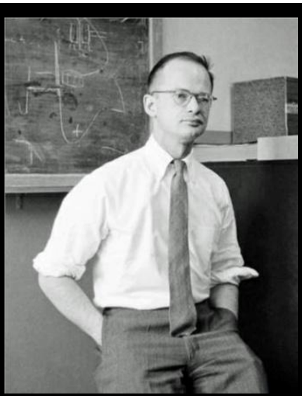


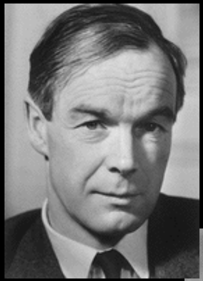
The fundamental component of human mind:
From neurons to psyché.

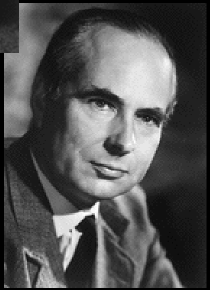
McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133.

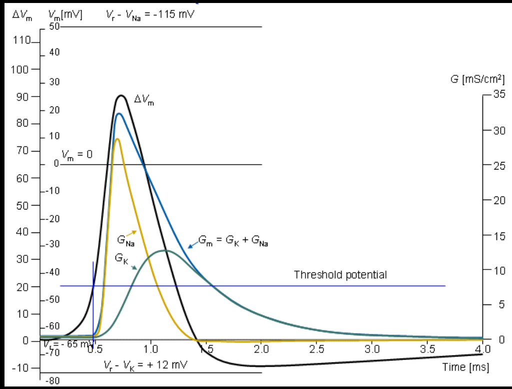












$$i_m = i_{mI} + c_m \frac{\partial V_m}{\partial t} = \frac{1}{r_i + r_o} \frac{\partial^2 V_m}{\partial x^2}$$

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, *117*(4), 500-544

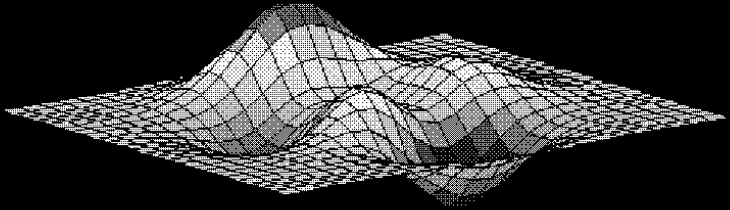
Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. Wiley, New York

$$w_{ij} = \frac{1}{p} \sum_{k=1}^{p} x_i^k x_j^k,$$

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Widrow, B., & Hoff, M. E. (1960). *Adaptive switching circuits* (No. TR-1553-1). STANFORD UNIV CA STANFORD ELECTRONICS LABS.

$$w_{i,j}(t + 1) = w_{i,j}(t) + n * (t_j – o_j) * x_i$$



The rising of Multi-Layer Perceptron (MLP)

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation* (No. ICS-8506). California Univ San Diego La Jolla Inst for Cognitive Science.

D. E. Rumelhart, G. E. Hinton and R. J. Williams, “Parallel Distributed Processing Explorations in the Microstructure of Cognition, Vol. 1 & 2,” MIT Press, Cambridge, 1986.

Geoffrey Hinton
(1947-20XX) : Psychology & Computer Science, University of Toronto. Parallel Distributed Processing Group.

James McClelland
(1948-20XX) : Psychology & Cognitive Science, Stanford University. Parallel Distributed Processing Group.

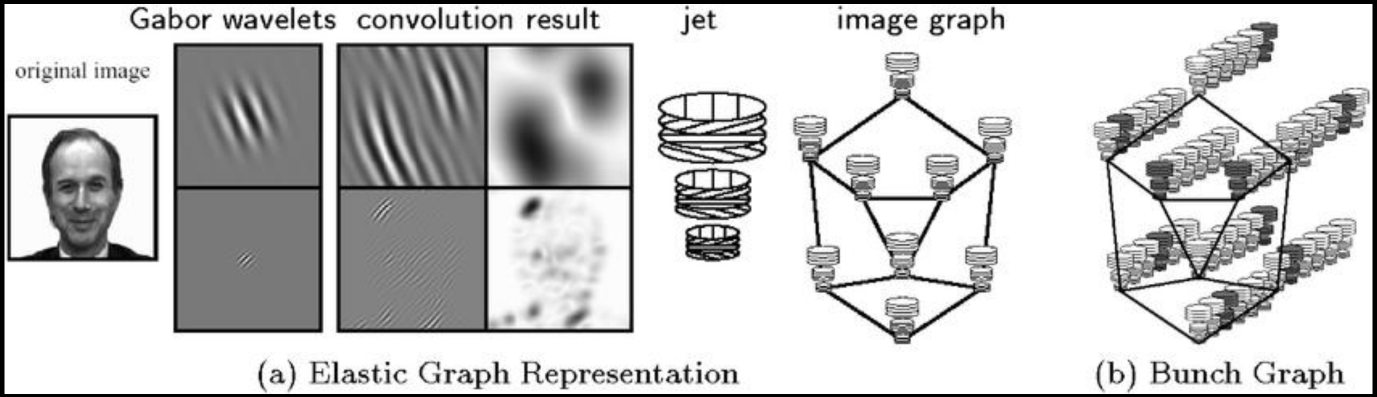
David Everett Rumelhart
(1942-2011) : Psychology, UCSD and Stanford University. Parallel Distributed Processing Group.

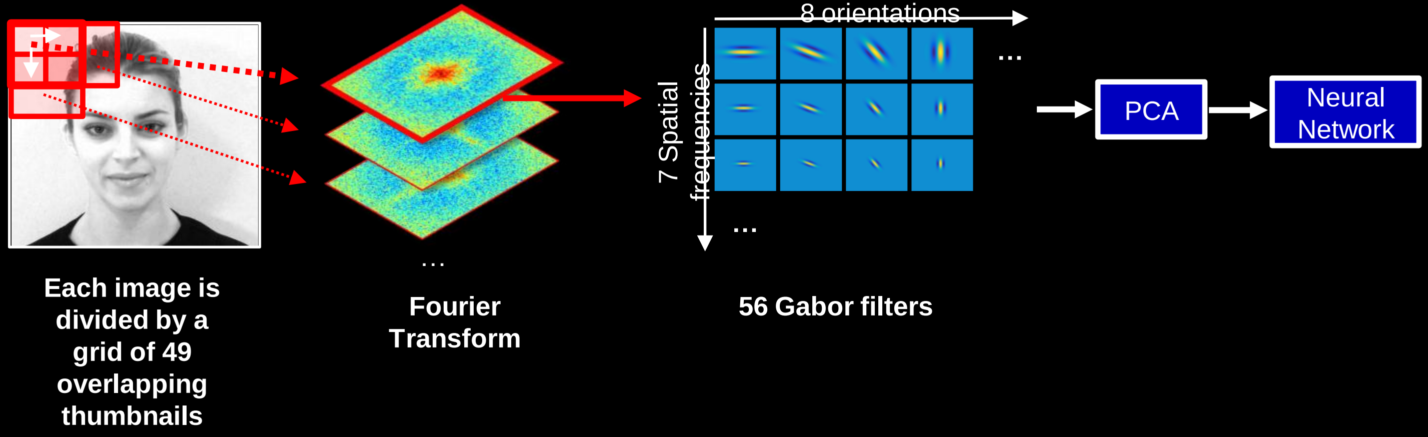
Increasing engineering with bio-inspired neural networks

Wiskott, L., Krüger, N., Kuiger, N., & Von Der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE Transactions on pattern analysis and machine intelligence*, *19*(7), 775-779.

Mermillod, M., Bonin, P., Mondillon, L., Alleysson, D., & Vermeulen, N. (2010). Coarse scales are sufficient for efficient categorization of emotional facial expressions: Evidence from neural computation. *Neurocomputing*, *73*(13-15), 2522-2531.

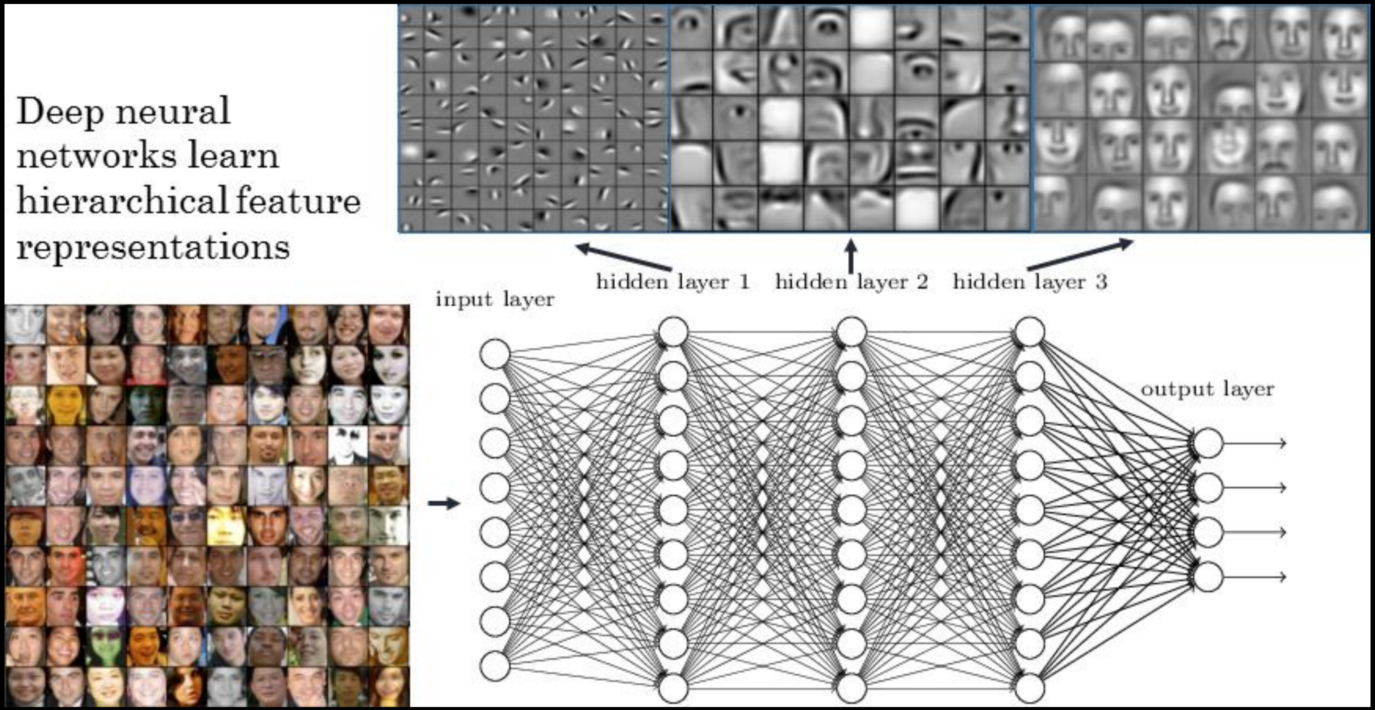
Mermillod, M., Guyader, N., & Chauvin, A. (2005). The coarse-to-fine hypothesis revisited: Evidence from neuro-computational modeling. *Brain and Cognition*, *57*(2), 151-157.





From MLP to DNN

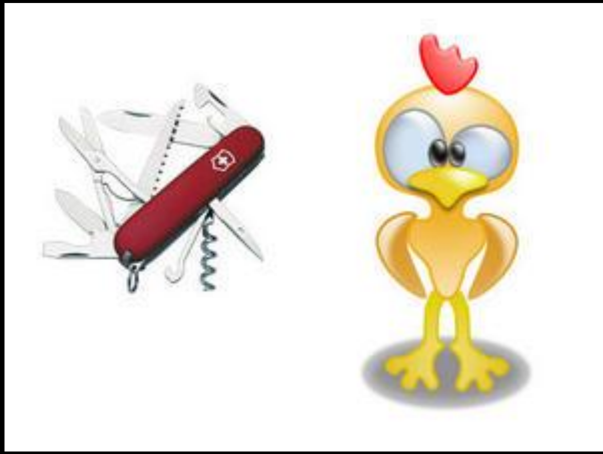
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. N*ature*, *521*(7553), 436.



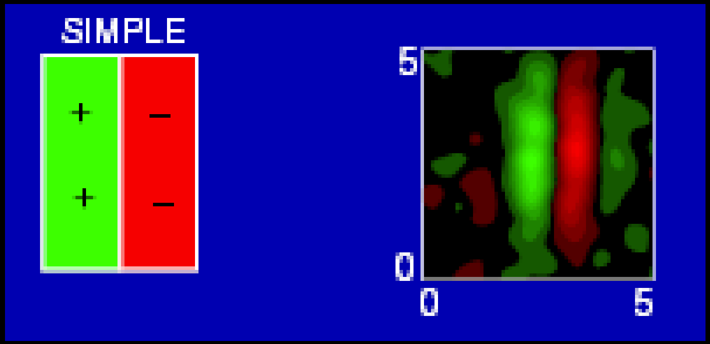
Warning

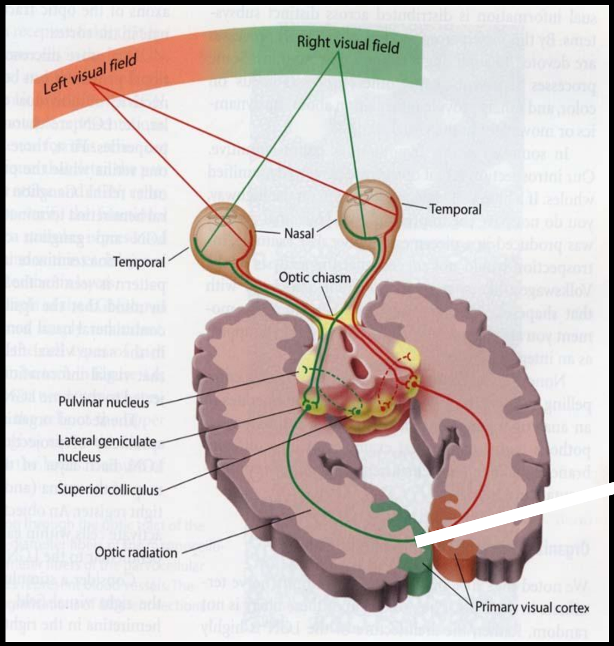
Convolution / pooling is not a magic formula !!!

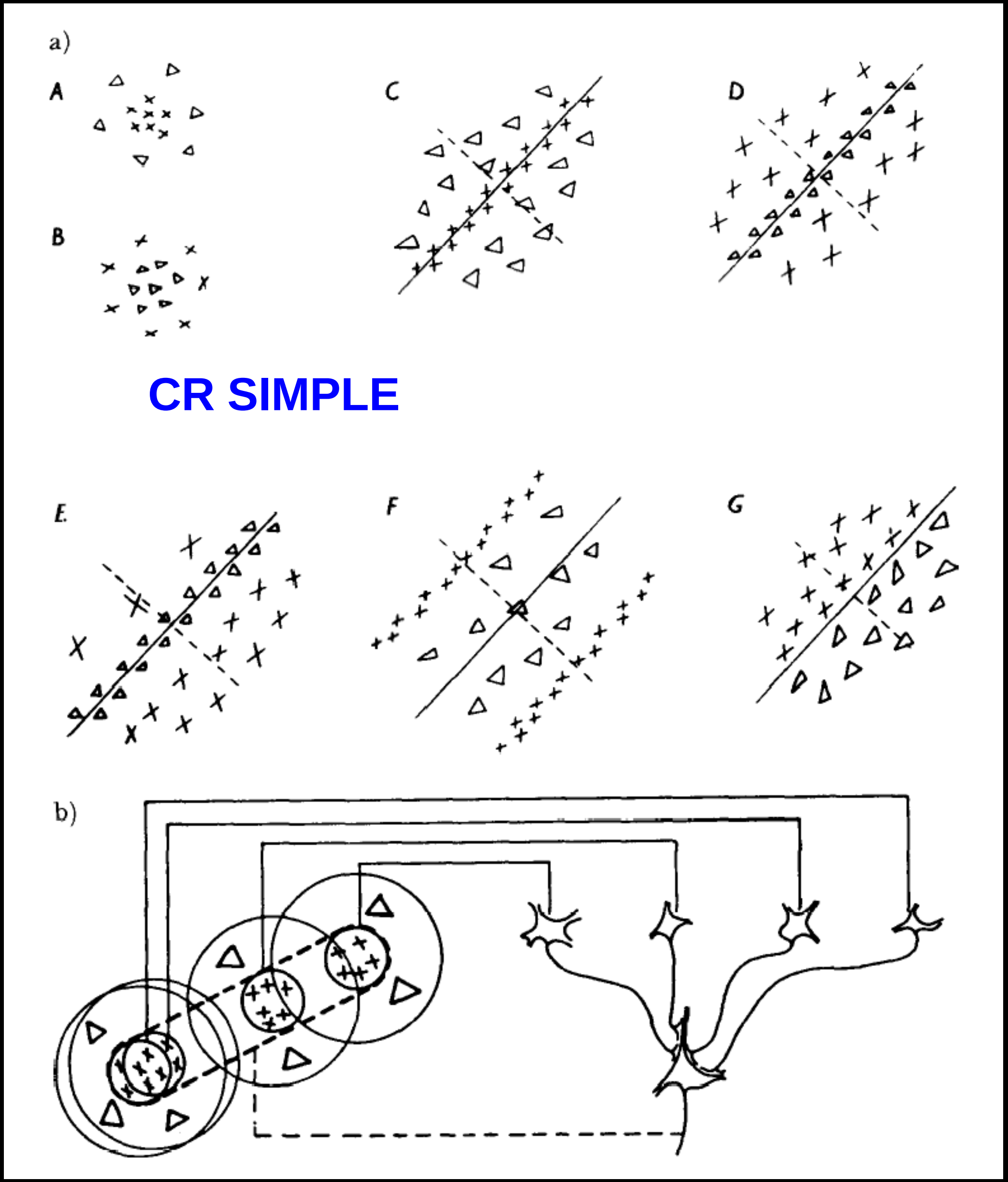
“FOR A MEANINGFUL ARTIFICIAL INTELLIGENCE” (Cédric Villani, 2018)



Is DNN bio-inspired?







Gabor filters in artificial and biological neural networks

Jones, J. P., & Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. Journal of neurophysiology, 58(6), 1233-1258.

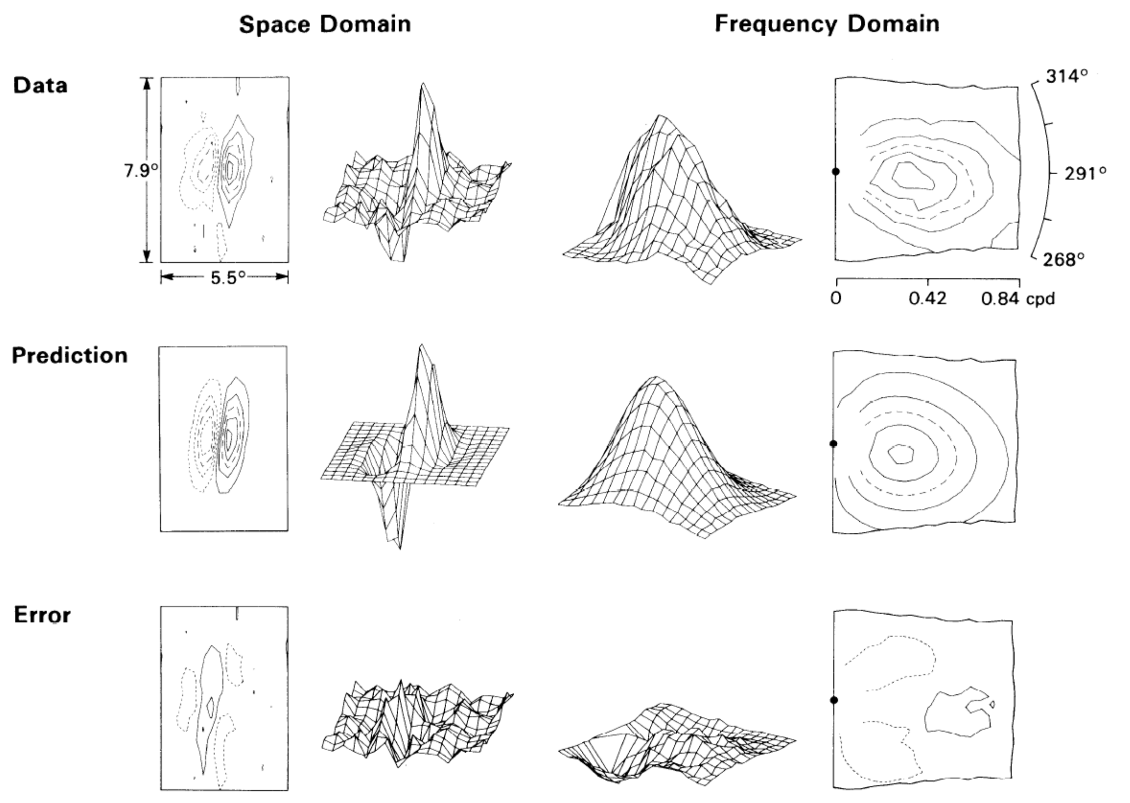
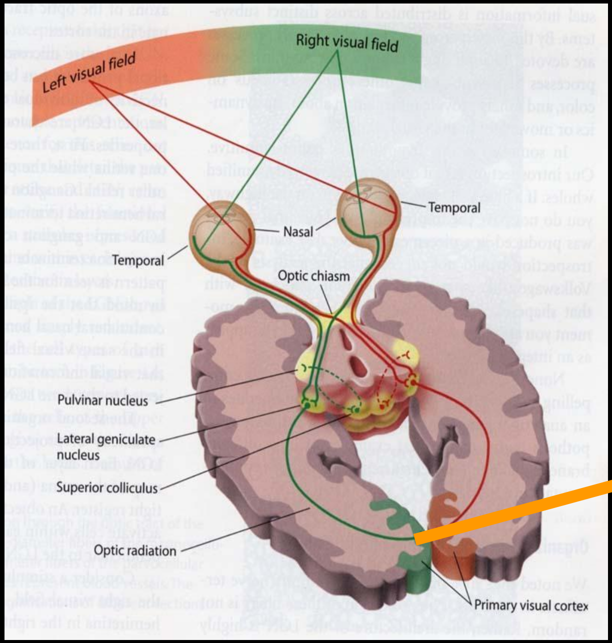
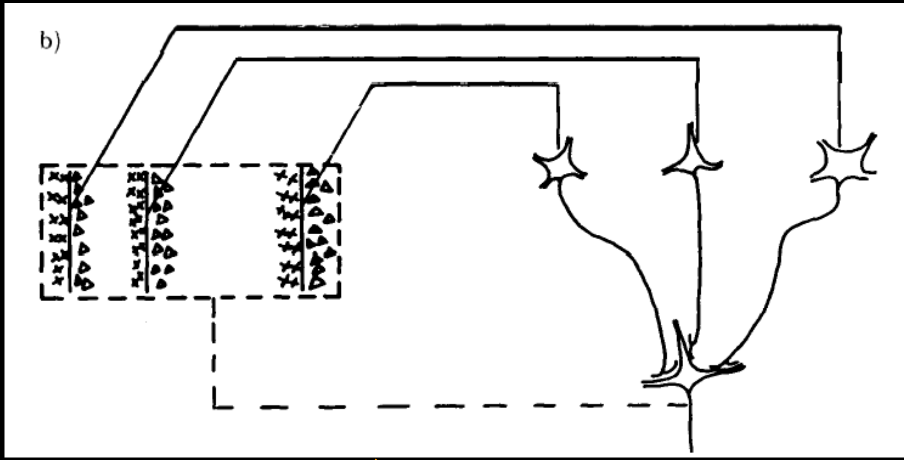


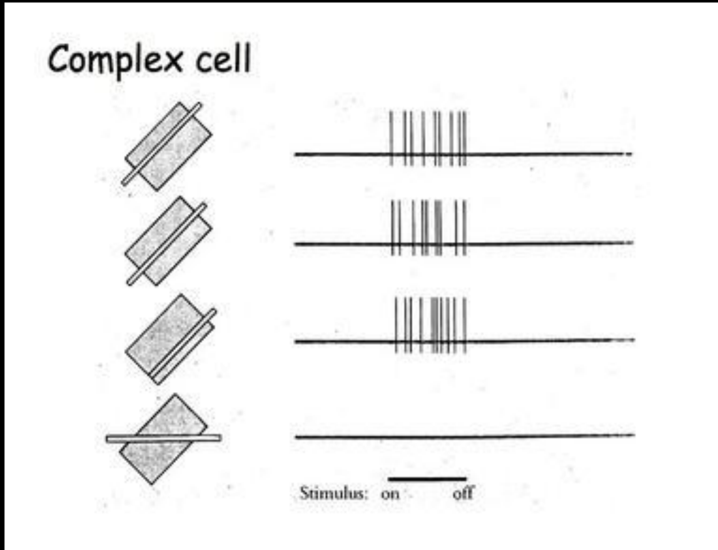
FIG. 14. In many cases, simple cell two-dimensional (2D) spatial and spectral response profiles are Fourier transform pairs, indistinguishable from 2D Gabor filters. *Top row*: the observed response profiles of a single cell. *Middle row*: the 2D spatial response profile predicted by the frequency domain data, and the 2D spectral response profile predicted by the space domain data, using the fitted 2D Gabor filters as intermediate steps in both cases. *Bottom row*: the error.

Champs récepteur/pooling (Hubel & Wiesel, 1968)

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, *195*(1), 215-243.



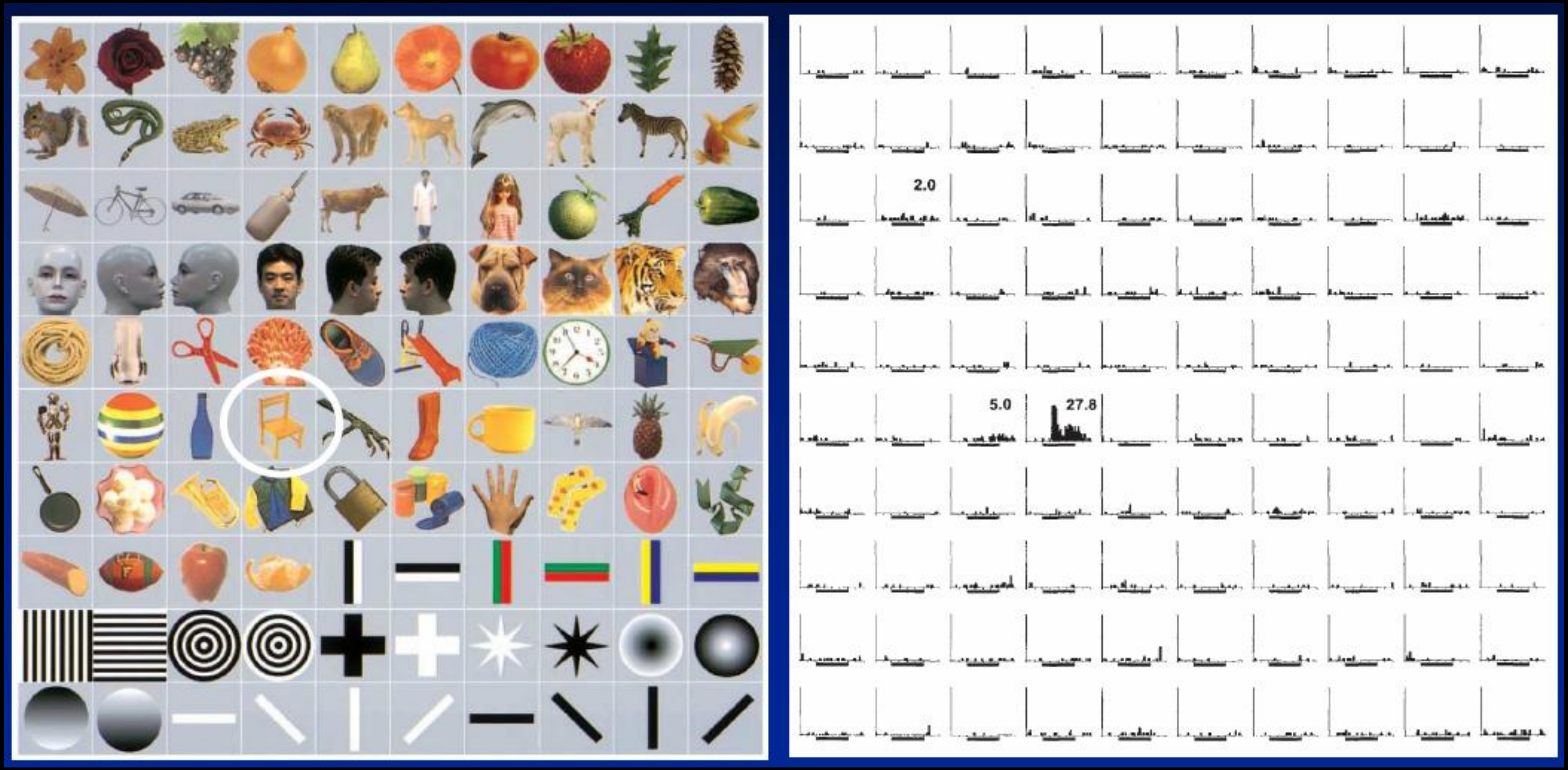




Pooling process in primates (e.g. Keiji Tanaka, Rufin Vogels, etc.)

Tamura, H., & Tanaka, K. (2001). Visual response properties of cells in the ventral and dorsal parts of the macaque inferotemporal cortex. *Cerebral Cortex*,*11*(5), 384-399.

Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. Part 2: single-cell study. *European Journal of Neuroscience*, *11*(4), 1239-1255.



Category-specific neurons in humans

Kreiman, G., Koch, C., & Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature neuroscience*, *3*(9), 946.

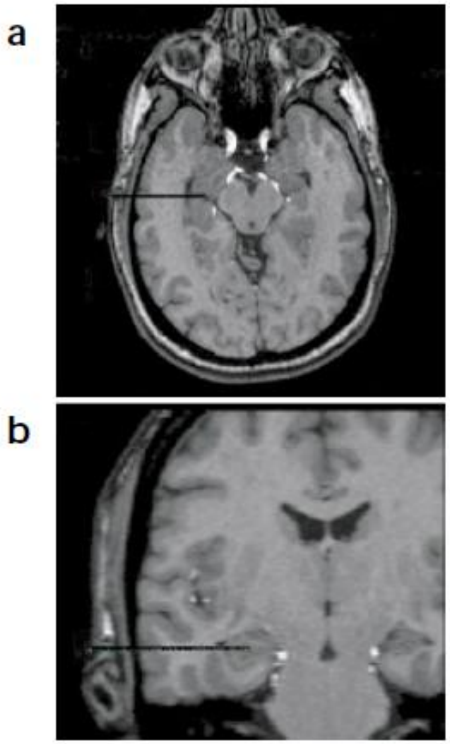
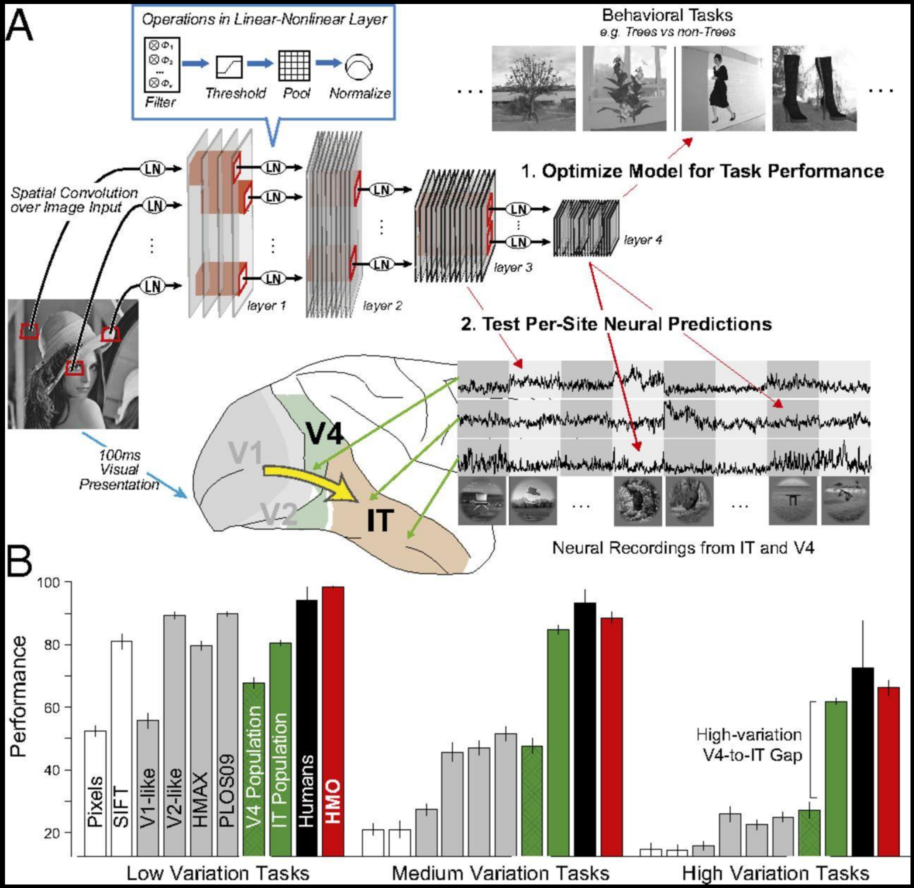


Fig. 1. Electrode placement. The trajectory of an electrode placed in the hippocampus is depicted in axial (**a**) and coronal (detail, **b**) structural MR images (1.5 Tesla scanner). Post-operative CT and MRI were used to confirm the location of the electrode. The CT was co-registered with MRI structural information for anatomic verification. The distal end of the electrode included platinum-iridium microwires from which single neurons were recorded. The microwires extended about 4 mm from the tip, lying on a cone with an opening angle of less than 45 degrees.

Fig. 2. Sample of stimuli presented in each category. Figures (mostly color) were drawn from a group of nine categories that included faces denoting emotional expressions by unknown actors[21], household objects, spatial layouts (including house exteriors, interiors and natural scenes), animals, cars, drawings of famous people or cartoon characters, photographs of famous people, food items and abstract patterns. Stimuli were presented for 1000 ms. Subjects had to indicate by pressing a button whether the image was a human face or not.

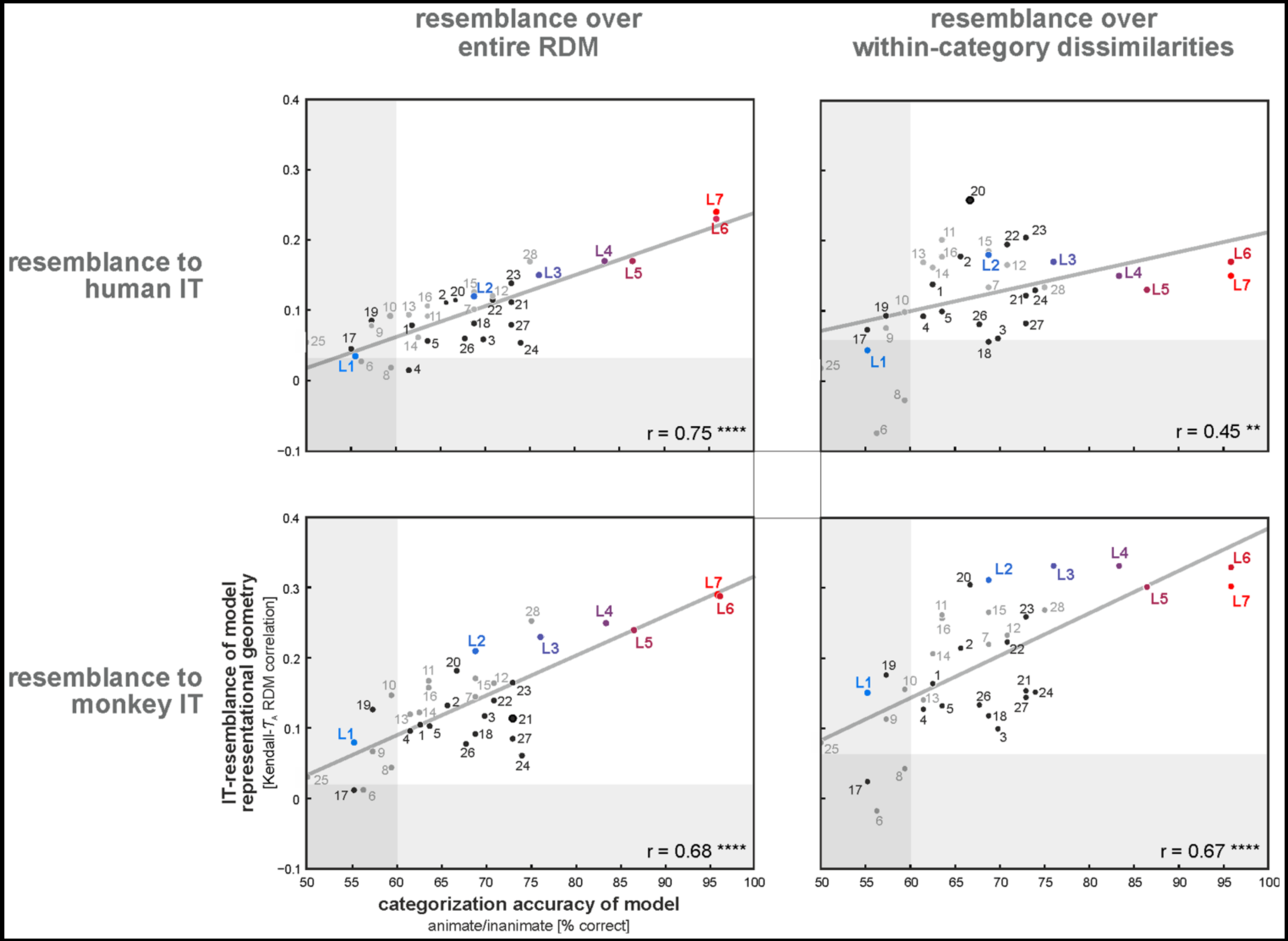
Ok, DNN are brain-inspired, but can we assess that they reliably simulate the ventral stream?
iEEG Data with primates

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619-8624.



fMRI (humans) and iEEG (primates) evidence.

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS computational biology*, *10*(11), e1003915.



iEEG Data with humans

Kuzovkin, I., Vicente, R., Petton, M., Lachaux, J. P., Baciu, M., Kahane, P., ... & Aru, J. (*in press*). Activations of Deep Convolutional Neural Network are Aligned with Gamma Band Activity of Human Visual Cortex. *Nature Communication Biology*.

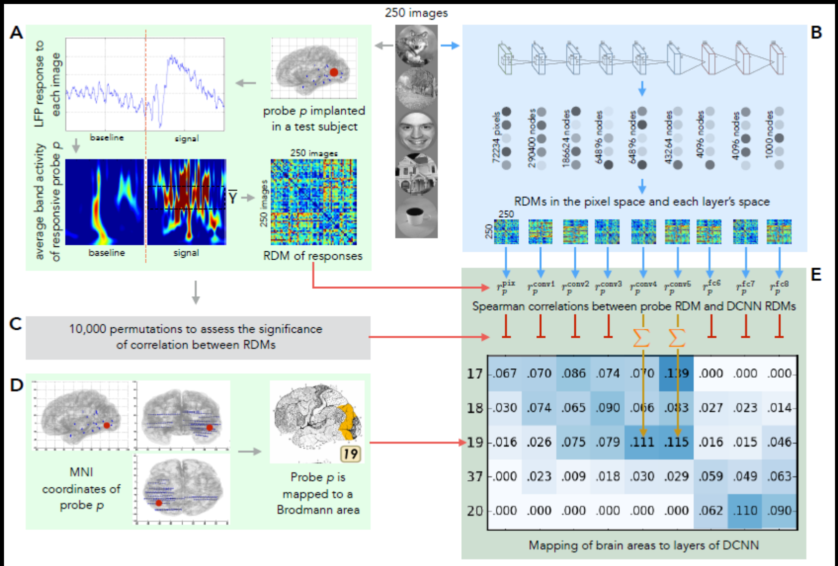


Figure 1 Overview of the analysis pipeline. 250 natural images are presented to human subjects (panel A) and to an artificial vision system (panel B). The activities elicited in these two systems are compared in order to map regions of human visual cortex to layers of deep convolutional neural networks (DCNNs). **A:** LFP response of each of 11293 electrodes to each of the images is converted into the frequency domain. Activity evoked by each image is compared to the activity evoked by every other image and results of this comparison are presented as a representational dissimilarity matrix (RDM). **B:** Each of the images is shown to a pre-trained DCNN and activations of each of the layers are extracted. Each layer's activations form a representation space, in which stimuli (images) can be compared to each other. Results of this comparison are summarized as a RDM for each DCNN layer. **C:** Subject's intracranial responses to stimuli are randomly reshuffled and the analysis depicted in panel A is repeated 10000 times to obtain 10000 random RDMs for each electrode. **D:** Each electrode's MNI coordinates are used to map the electrode to a Brodmann area. The figure also gives an example of electrode implantation locations in one of the subjects (blue circles are the electrodes). **E:** Spearman's rank correlation is computed between the true (non-permuted) RDM of neural responses and RDMs of each layer of DCNN. Also 10000 scores are computed with the random RDM for each electrode-layer pair to assess the significance of the true correlation score. If the score obtained with the true RDM is significant (the value of $p < 0.001$ is estimated by selecting a threshold such that none of the probes would pass it on the permuted data), then the score is added to the mapping matrix. The procedure is repeated for each electrode and the correlation scores are summed and normalized by the number of electrodes per Brodmann area. The resulting mapping matrix shows the alignment between the consecutive areas of the ventral stream and layers of DCNN.

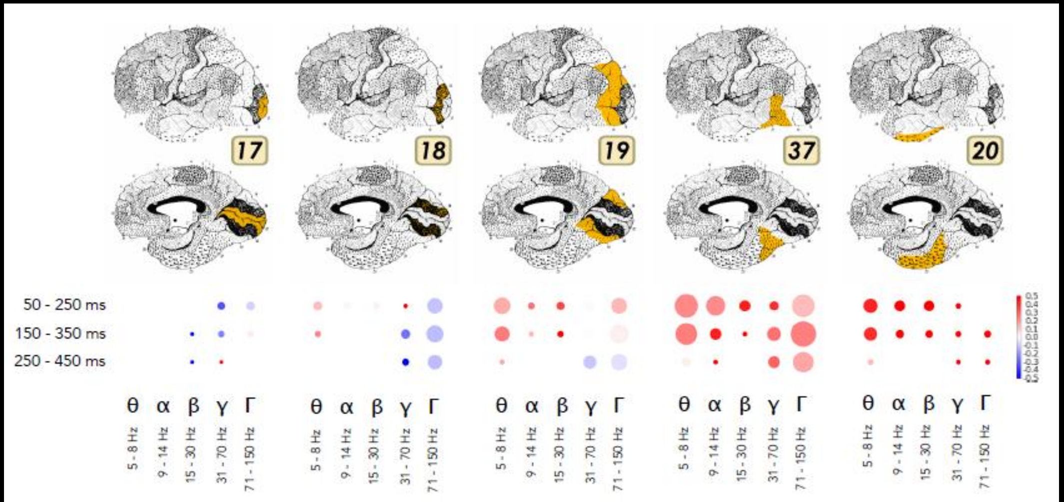
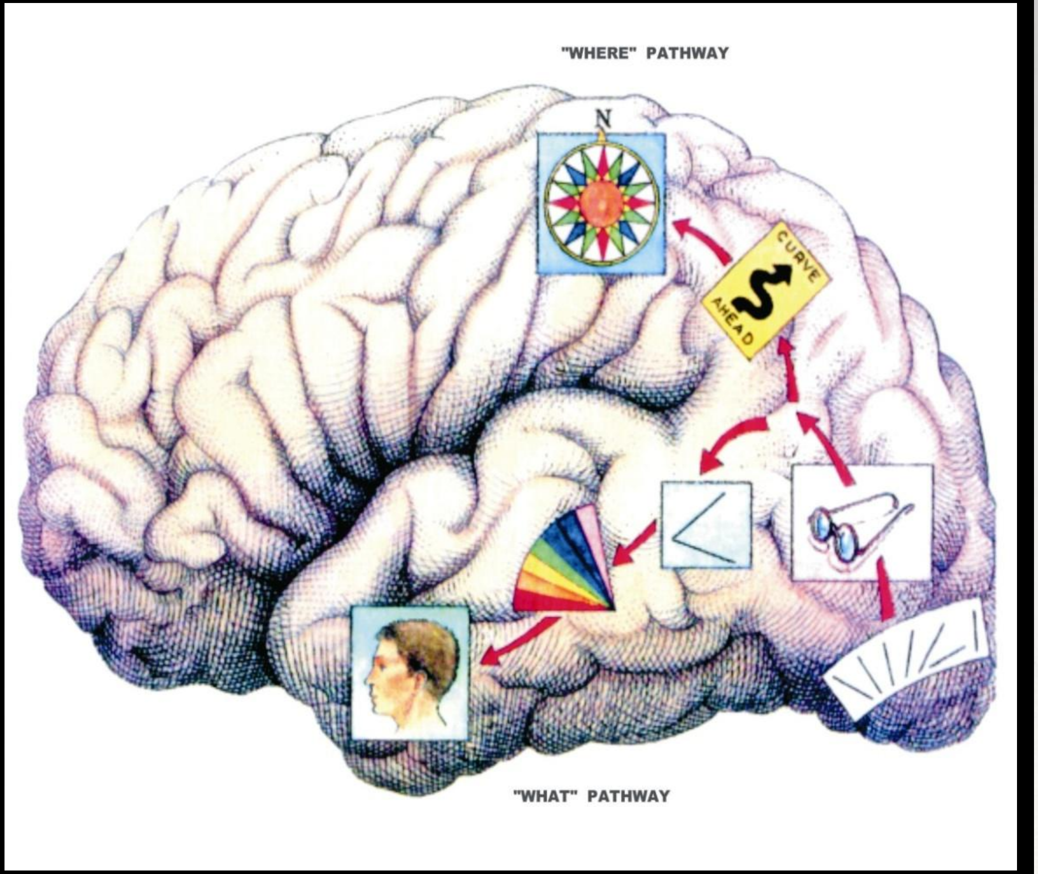
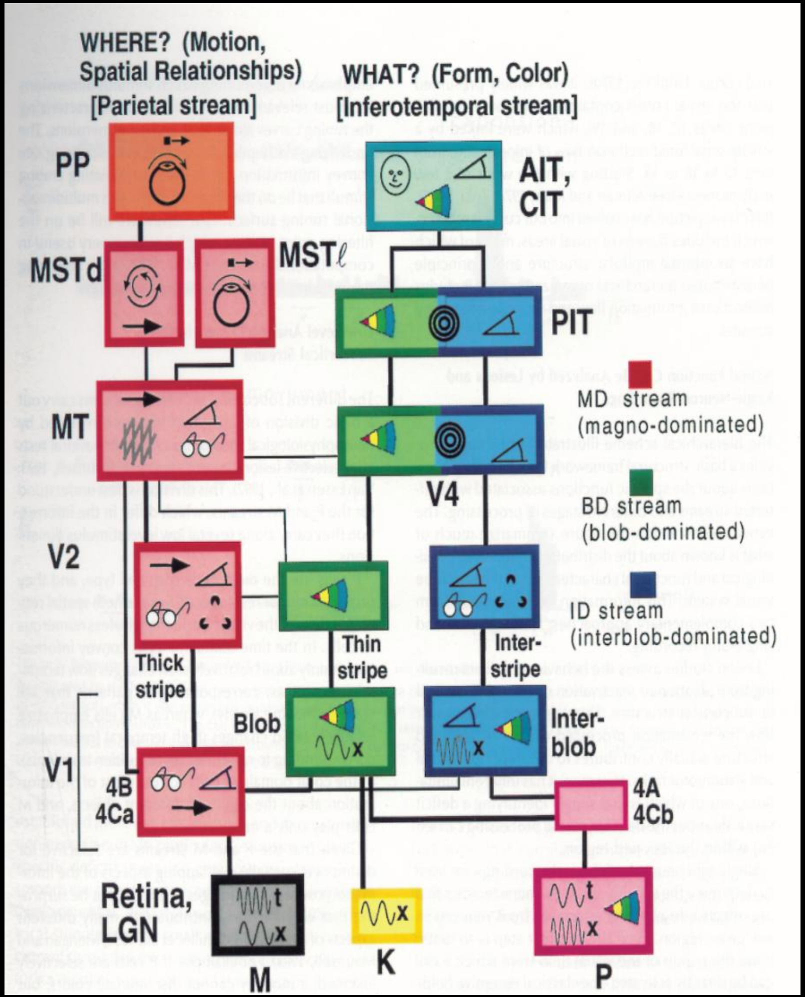


Figure 5 Area-specific analysis of volume of neural activity and complexity of visual features represented by that activity. Size of the marker shows the sum of correlation coefficients between the area and DCNN for each particular band and time window. Color codes the ratio of complex visual features to simple visual features, i.e. the comparison between the activity that correlates with the higher layers (`conv5`, `fc6`, `fc7`) of DCNN to the lower layers (`conv1`, `conv2`, `conv3`). Intensive red means that the activity was correlating more with the activity of higher layers of DCNN, while the intensive blue indicates the dominance of correlation with the lower areas. If the color is close to white then the activations of both lower and higher layers of DCNN were correlating with the brain responses in approximately equal proportion.

Ok, but is interdisciplinarity still required for future AI?





Example of autonomous vehicules.

AI will continue to kill people.

Gestalt process and top-down expectations required !

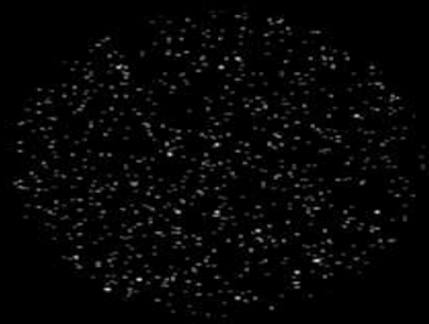
Koffka, K. (1922). Perception: an introduction to the Gestalt-Theorie*.* *Psychological Bulletin, 19*(10), 531.

Importance for movement detection

Newsome, W. T., & Pare, E. B. (1988). A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *Journal of Neuroscience*, *8*(6), 2201-2211.

100% de cohérence

30% de cohérence



5% de cohérence

Importance action understanding and planification!

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, *119*(2), 593-609.

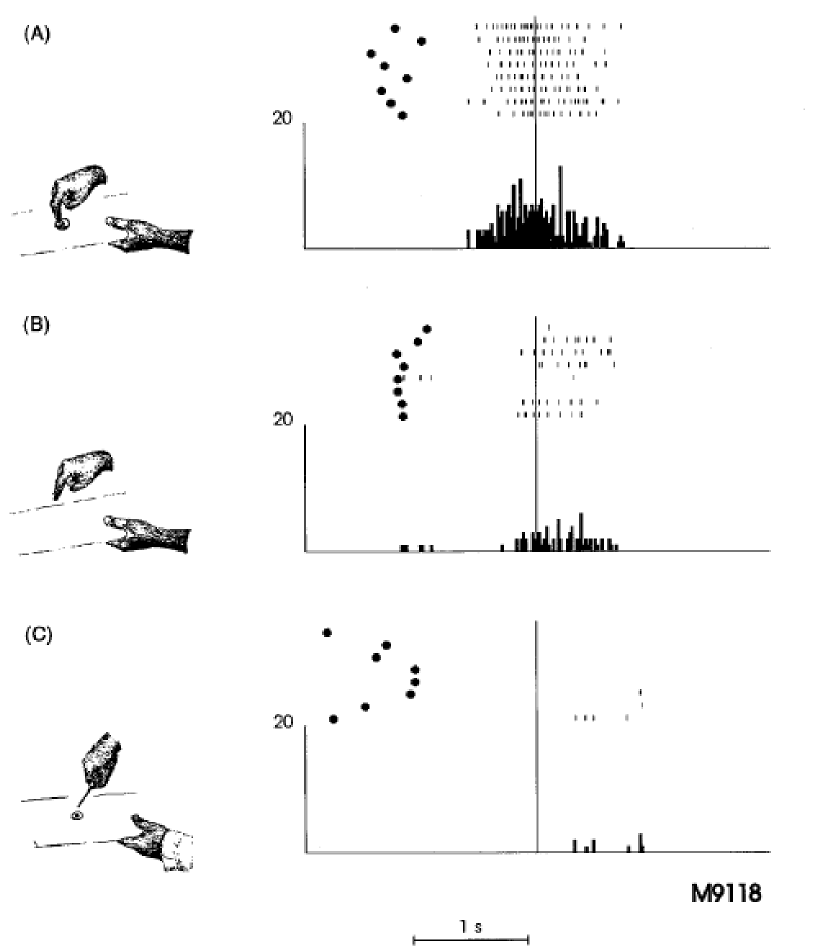
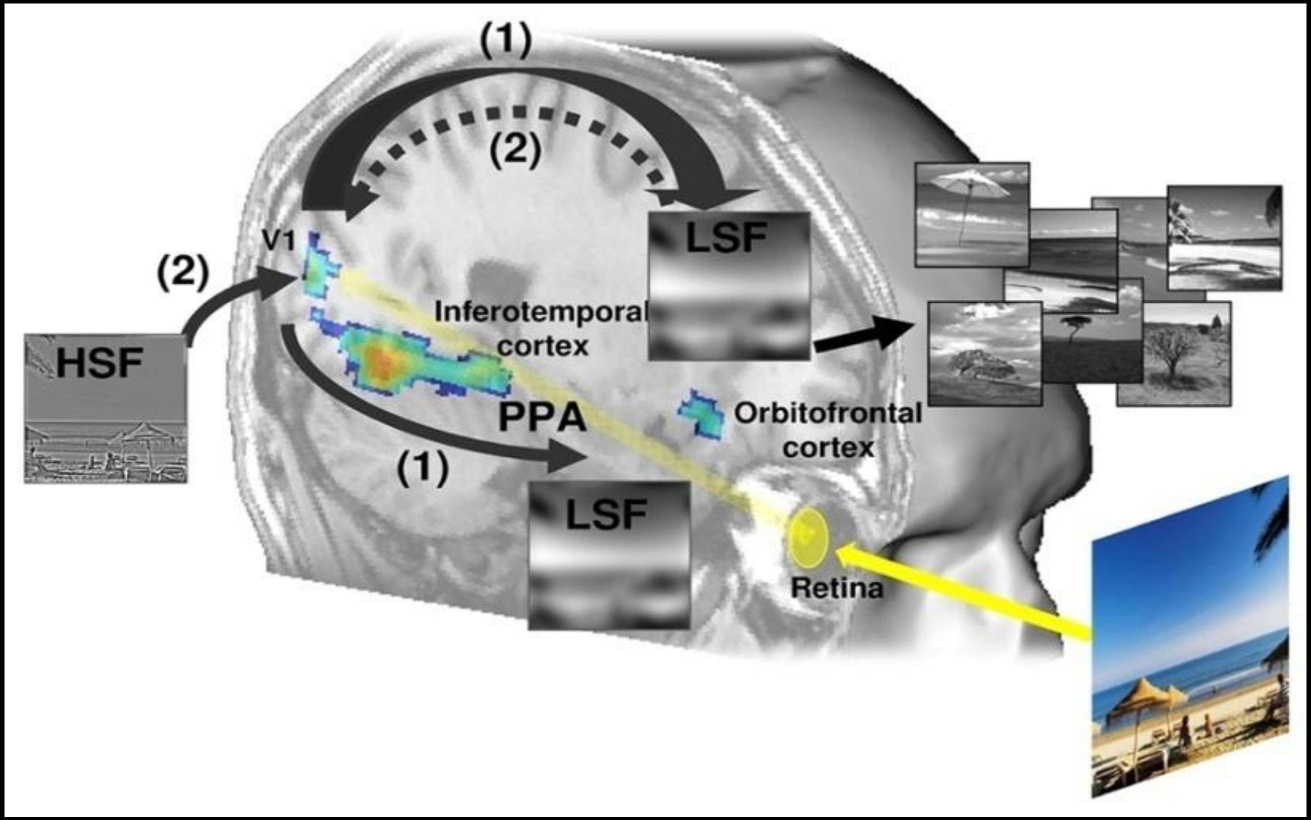


Fig. 4 Visual responses of a manipulating mirror neuron. **(A)** The experimenter retrieved a piece of food placed in a well in a tray, using his index finger. This was the only action that activated the neuron. **(B)** The same action was mimed without food. **(C)** the food was retrieved using a tool. Conventions as in Fig. 1.

Importance for anticipation

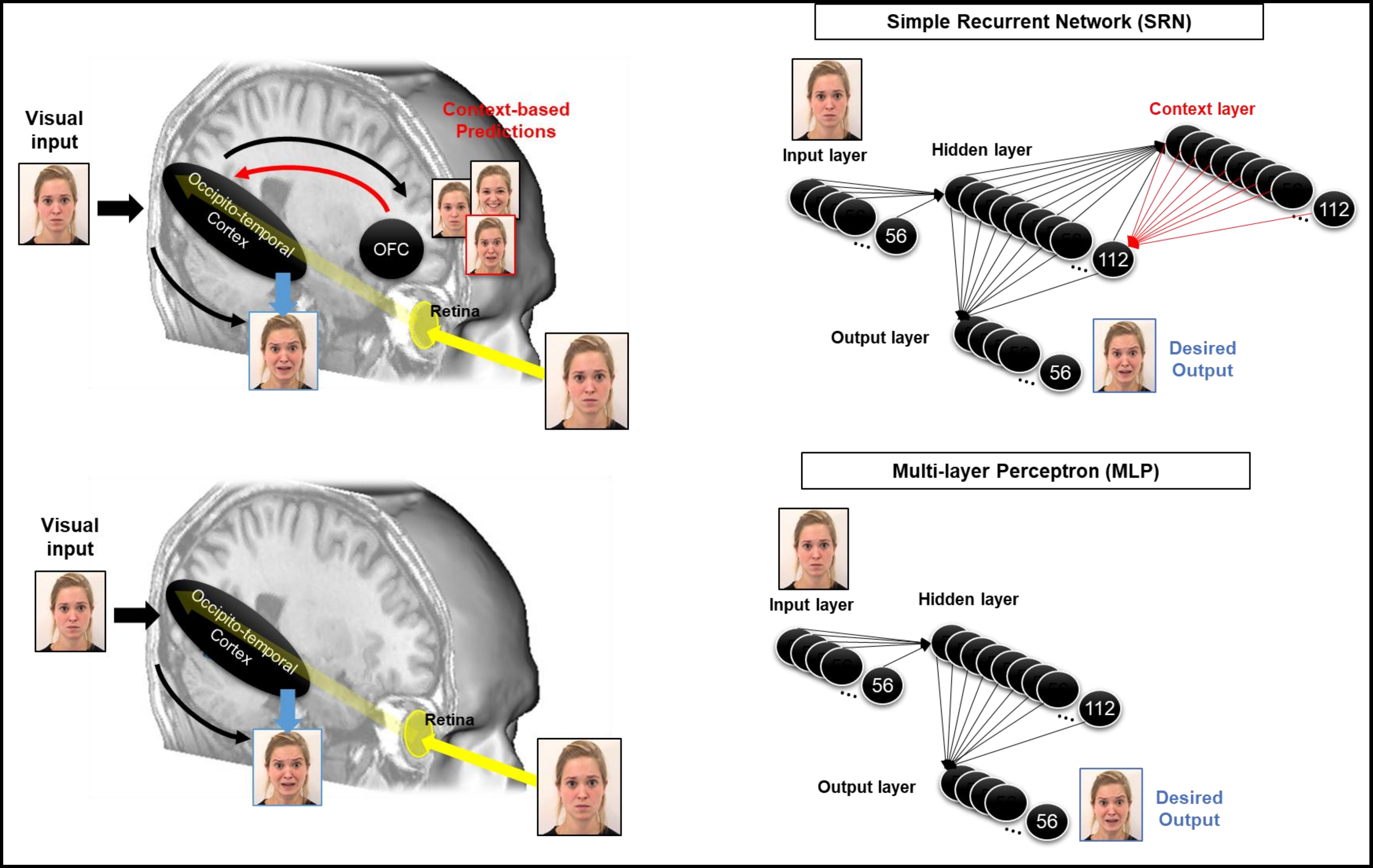
Kauffmann, L., Ramanoël, S., & Peyrin, C. (2014). The neural bases of spatial frequency processing during scene perception. *Frontiers in integrative neuroscience*, *8*, 37.

Beffara, B., Wicker, B., Vermeulen, N., Ouellet, M., Bret, A., Molina, M. J. F., & Mermillod, M. (2015). Reduction of interference effect by low spatial frequency information priming in an emotional Stroop task. *Journal of vision*, *15*(6), 16-16.



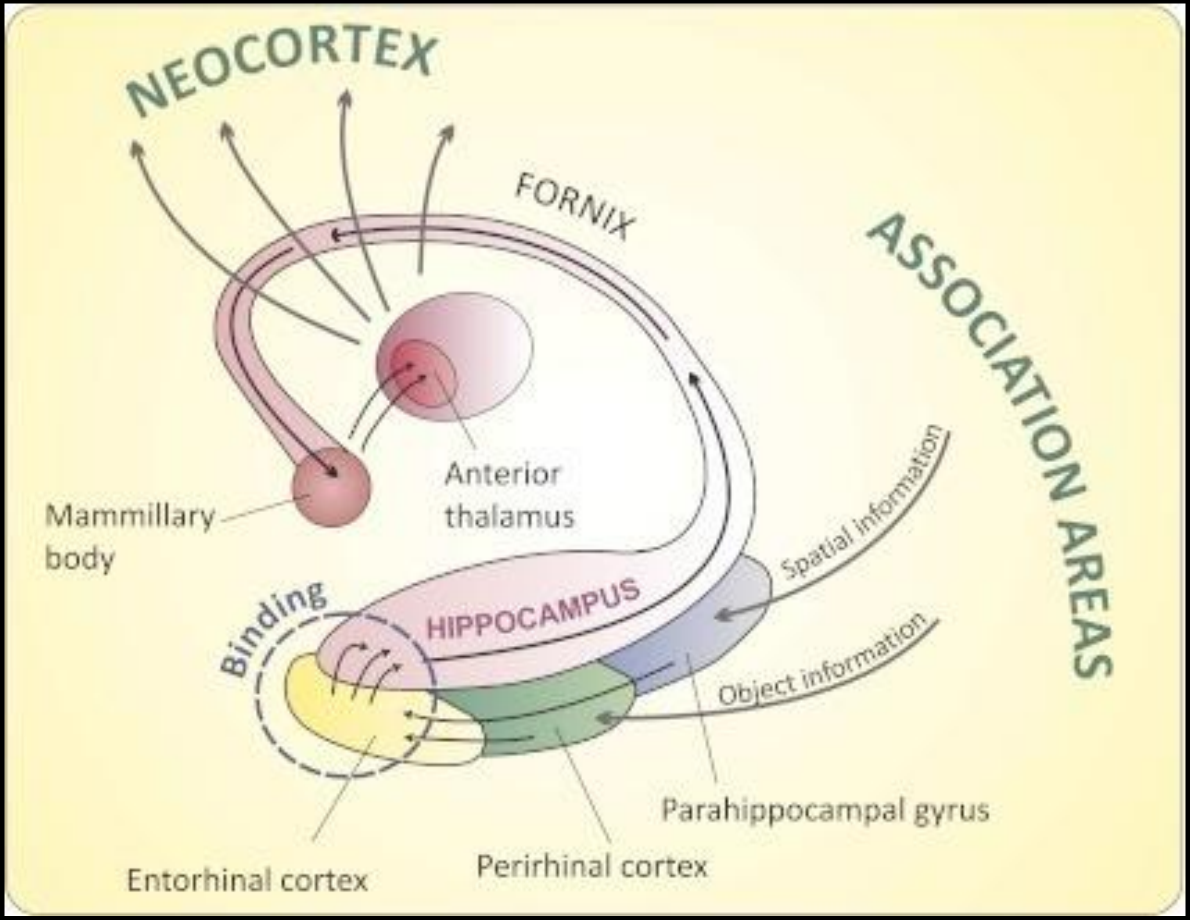
Bio-inspired Predictive Brain

Mermillod et al. (under review). The Importance of Recurrent Top-Down Synaptic Connections for the Anticipation of Dynamic Emotional Expressions. *Neural Networks.*



Importance of binding

Hafting, T., Fyhn, M., Molden, S., Moser, M. B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801.
Fyhn, M., Molden, S., Witter, M. P., Moser, E. I., & Moser, M. B. (2004). Spatial representation in the entorhinal cortex. *Science*, *305*(5688), 1258-1264.
O'keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford: Clarendon Press.



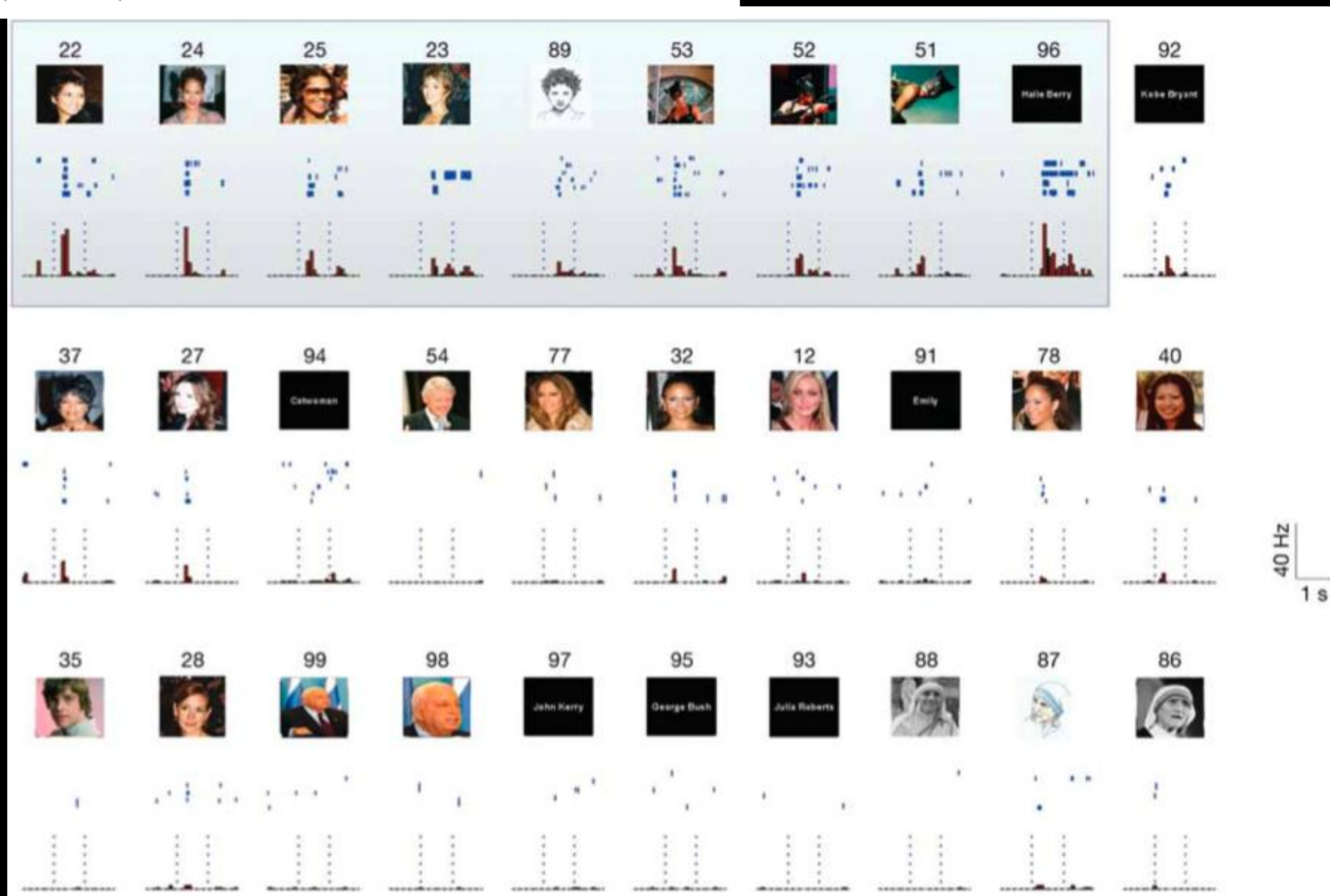
nature

Vol 435|23 June 2005|doi:10.1038/nature03687

LETTERS

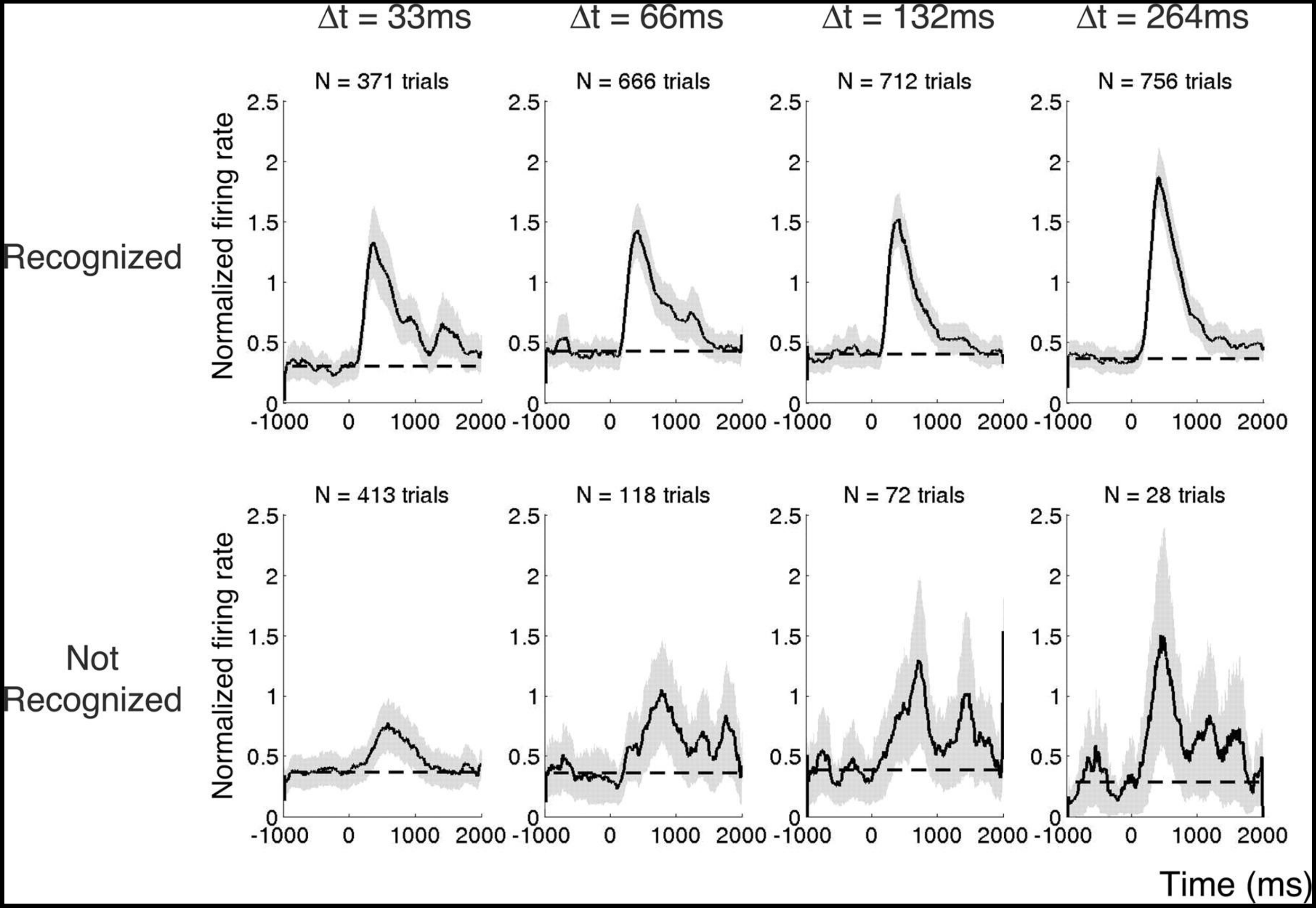
Invariant visual representation by single neurons in the human brain

R. Quian Quiroga[1,2]†, L. Reddy[1], G. Kreiman[3], C. Koch[1] & I. Fried[2,4]



Correlated with exogenous consciousness !

Quiroga, R. Q., Mukamel, R., Isham, E. A., Malach, R., & Fried, I. (2008). Human single-neuron responses at the threshold of conscious recognition. *Proceedings of the National Academy of Sciences*, *105*(9), 3599-3604.



Toward large-scale neural networks on chip

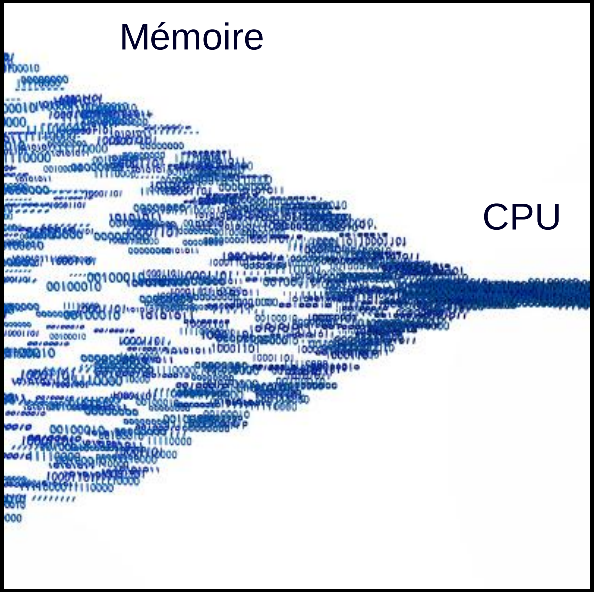
… And possibly self-consciousness?

The Human Brain Project

Perspective: Beyond Turing-Von Neumann machine

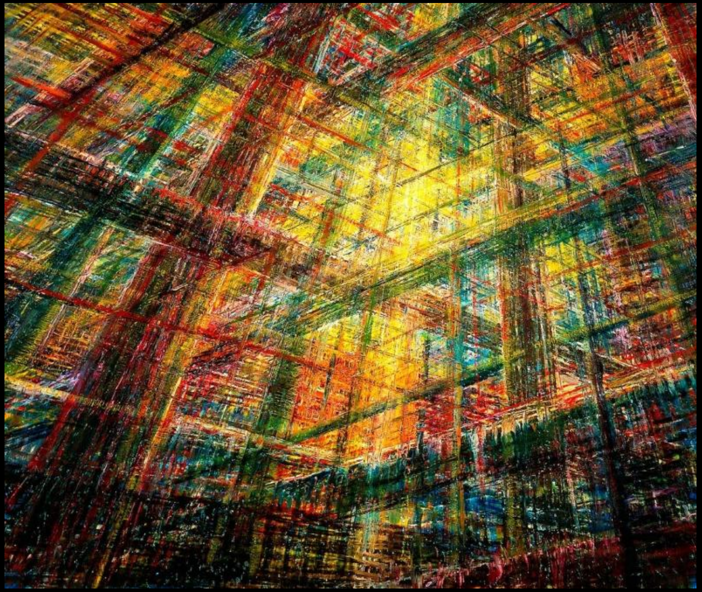
Turing-Von Neumann Machine

- CPU ≠ Memory
- CPU serial processes



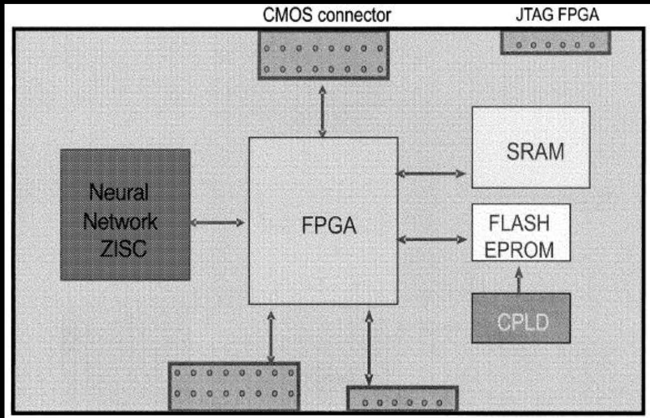
Neural Networks

- CPU = Memory
- Parallel and distributed processes

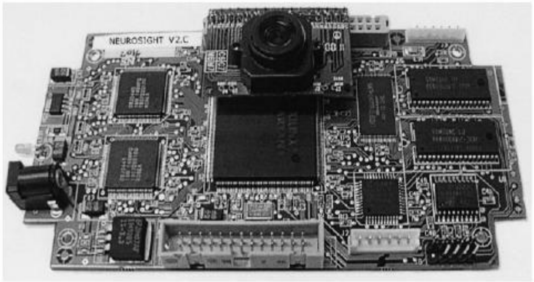


Neural Networks on CMOS (Complementary Metal Oxide Semiconductor).

Yang, F., & Paindavoine, M. (2003). Implementation of an RBF neural network on embedded systems: real-time face tracking and identity verification. *IEEE Transactions on Neural Networks*, *14*(5), 1162-1175.



(a)



(b)

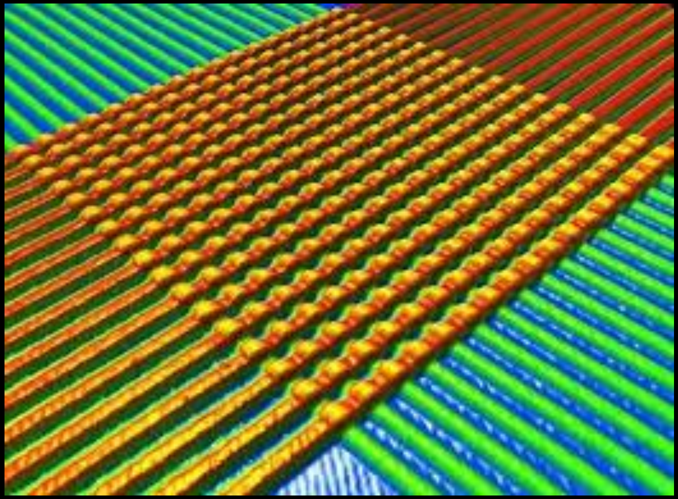
Neural Network on MEMRISTOR.

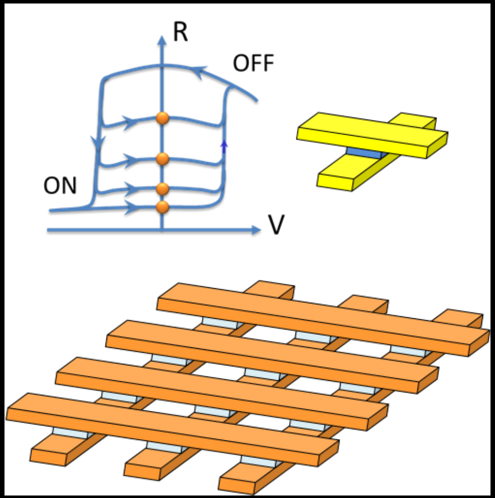
Chua, L. O. & Kang, S. M. Memristive devices and systems. Proc. IEEE 64, 209–223 (1976)
Chua, L. O., & Yang, L. (1988). Cellular neural networks: Applications. *IEEE Transactions on circuits and systems*, *35*(10), 1273-1290.
Chanthbouala, A., Garcia, V., Cherifi, R. O., Bouzehouane, K., Fusil, S., Moya, X., ... & Bibes, M. (2012). A ferroelectric memristor. *Nature materials*, *11*(10), 860.

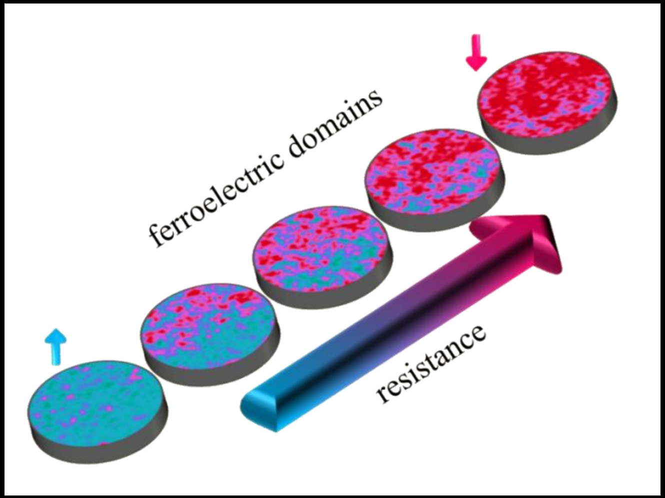
Memristor matrices…

… With massive parallel & distibuted synapse connectivity…

… And brain-inspired learning capacities.







Once artificial intelligence will go beyond the constraints of the Turing-Von Neumann Machine, surpassing human cognitive capacities shall be fast.

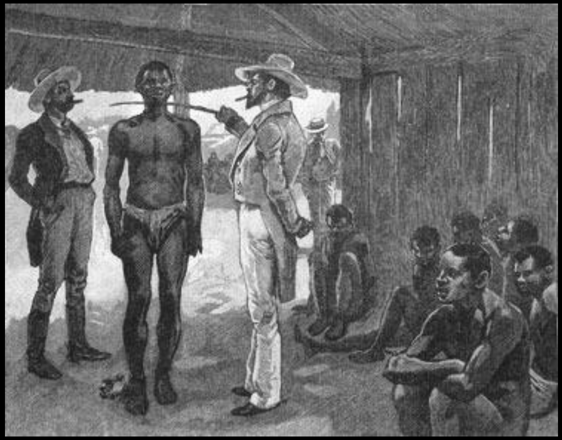
SKYNET

NEURAL NET-BASED ARTIFICIAL INTELLIGENCE

•CYBERDYNE SYSTEMS CORPORATION•

Note for the futur: Do not replicate an entire human brain !!!

Example of dehumanization process

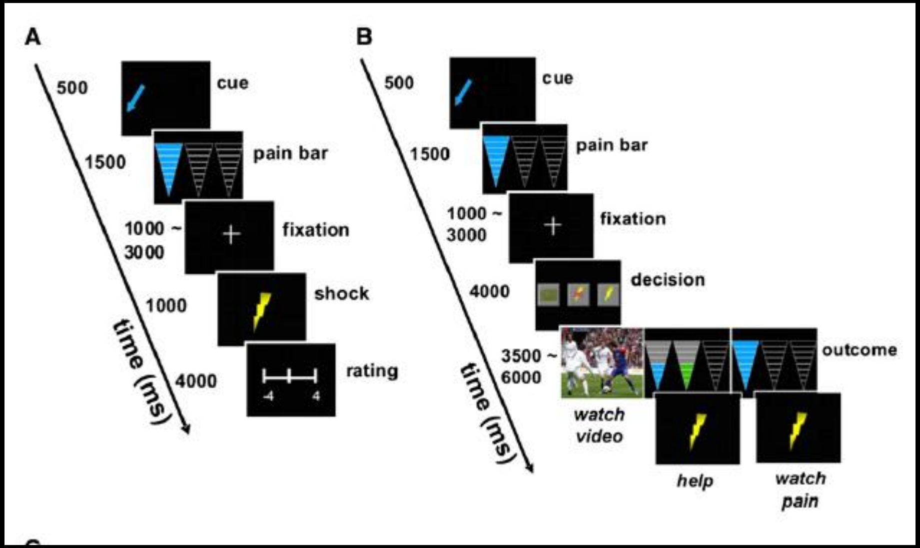



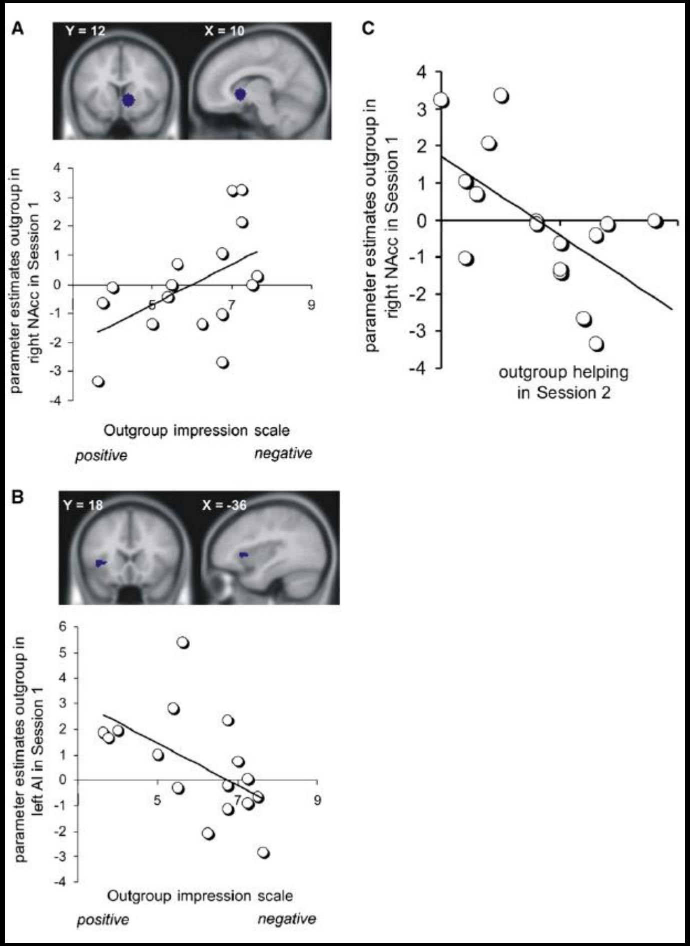

Minimal Group Paradigm.

Tajfel, H., Billig, M., Bundy, R. P. & Flament, C. (1971). Social categorization and intergroup behaviour. European Journal of Social Psychology, 2, 149-178.

The neural basis of dehumanization

Hein, G., Silani, G., Preuschoff, K., Batson, C. D., & Singer, T. (2010). Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. Neuron.



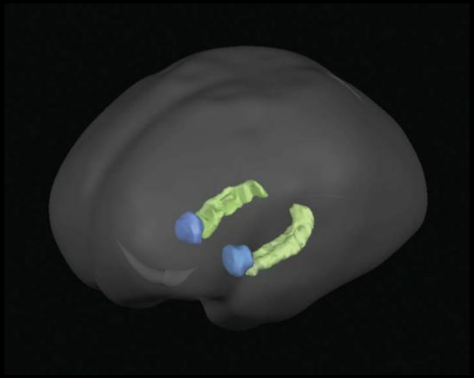


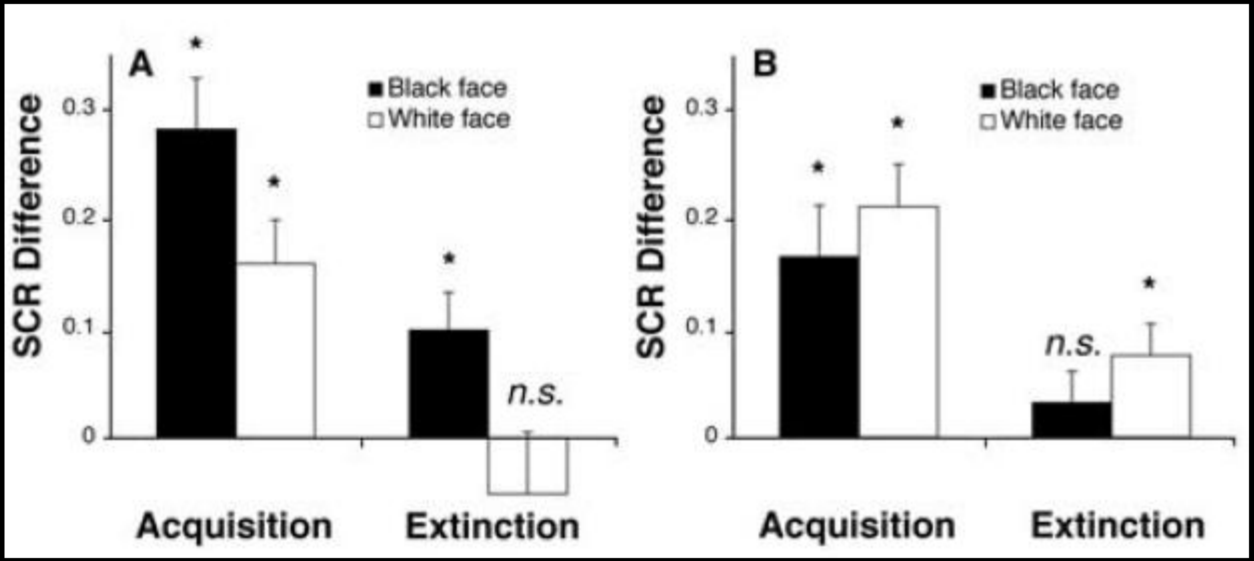
Example of the “Black Sheep Effect”

Marques, J. M., Yzerbyt, V. Y., & Leyens, J. P. (1988). The “black sheep effect”: Extremity of judgments towards ingroup members as a function of group identification. *European Journal of Social Psychology, 18*(1), 1-16.

Example of stereotypes and discrimination

Olsson, Ebert, Banji & Phelps (2005). The role of social groups in the persistence of learned fear. Science.
Olsson & Phelps (2007). Social learning of fear. Nature neuroscience.
Phelps (2006). Emotion and cognition: insights from studies of the human amygdala. Annual Review of Psychology.





Example of emotions

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, *17*(2), 124.

Asimov :

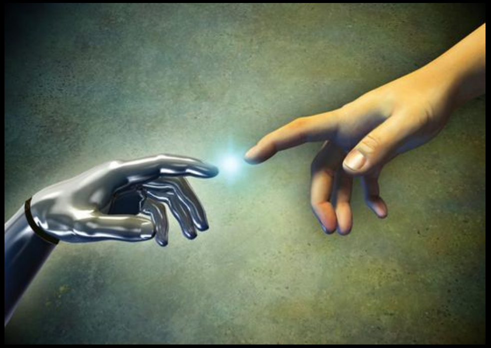
Same (mis)conception of artificial intelligence than Minsky & Papert or McCarthy.

-Première Loi : « Un robot ne peut porter atteinte à un être humain ni, restant passif, laisser cet être humain exposé au danger. » ;

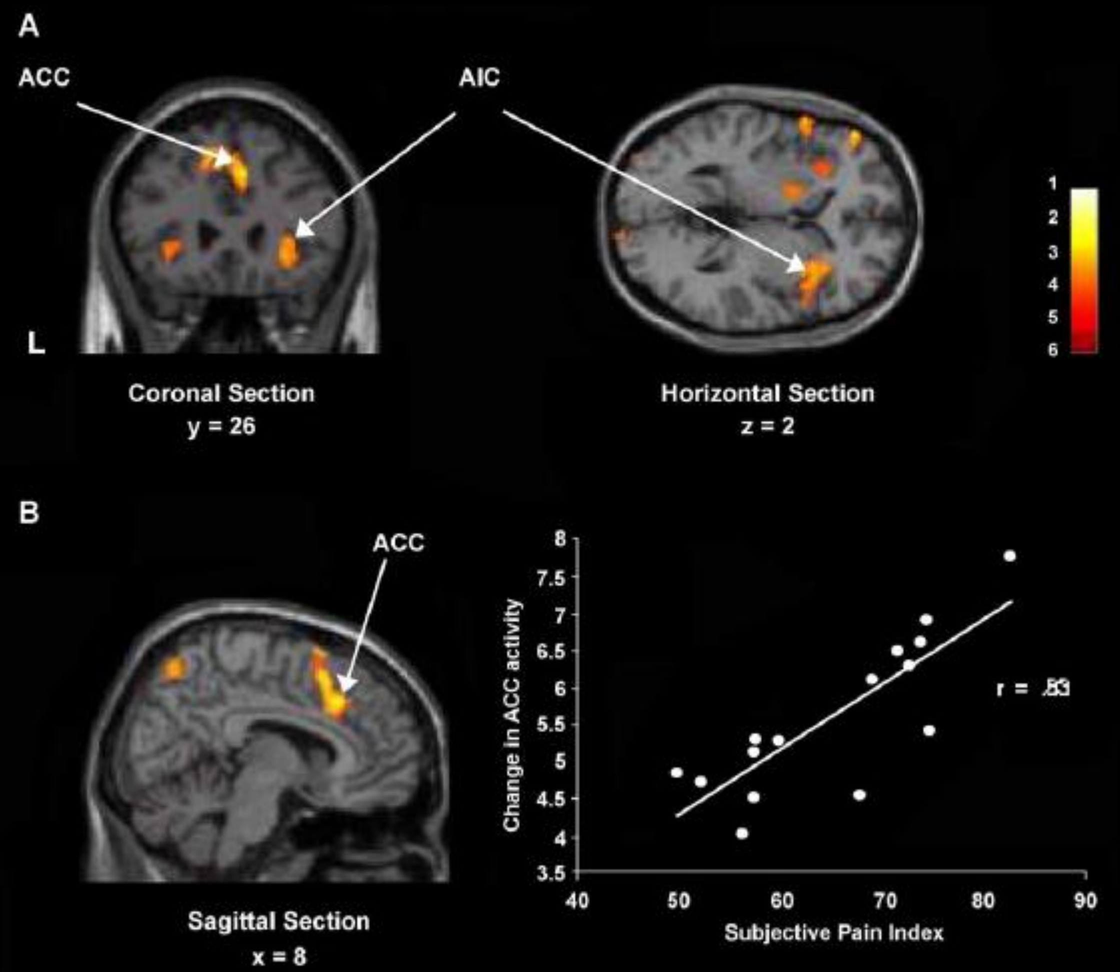
-Deuxième Loi : « Un robot doit obéir aux ordres donnés par les êtres humains, sauf si de tels ordres sont en contradiction avec la Première Loi. » ;

-Troisième Loi : « Un robot doit protéger son existence dans la mesure où cette protection n'entre pas en contradiction avec la Première ou la Deuxième Loi. »

The importance of understanding and replicating the neural substrate of empathy



Jackson, P. L., Meltzoff, A. N., & Decety, J. (2005). How do we perceive the pain of others? A window into the neural processes involved in empathy. *Neuroimage*, *24*(3), 771-779.



The pros and cons of artificial (versus biological) neural networks

- Slow versus metamorphic evolution

- Open to fast auto-evolution

- Hardware and energy does not require to eat other biological systems

- Immortality

Thank you for your attention