DataScience®
foundation
MAKE BETTER DECISIONS

# What you are too afraid to ask about Artificial Intelligence Part II

Author, Francesco Corea

A Data Science Foundation White Paper

March 2018

--------------------------------------------------

www.datascience.foundation

***Data Science Foundation***
Data Science Foundation, Atlantic Business Centre, Atlantic Street, Altrincham, WA14 5NQ
Tel: 0161 926 3641   Email: admin@datascience.foundation  Web: www.datascience.foundation
Registered in England and Wales 4th June 2015, Registered Number 9624670

# What you are too afraid to ask about Artificial Intelligence (Part II): Neuroscience and AI hardware

*This is the second part on recent developments in AI*

## ***Neuroscience and AI***

Along with the advancements in pure machine learning research, we have done many steps ahead toward a greater comprehension of the brain mechanisms. Although much has still to be understood, we have nowadays a slightly better overview of the brain processes, and this might help to foster the development of an AGI.

It seems clear that try **to fully mimic the human brain is not a feasible approach**, and is not even the correct one. However, drawing inspiration from how the brain works is a completely different story, and the study of neuroscience could both stimulate the creation of new algorithms and architectures, as well as validate the use of current machine learning research toward a formation of an AGI.

More in detail, according to Numenta's researchers AI should be inspired to the human neocortex. Although a common theoretical cortical framework has not been fully accepted by the scientific community, according to **Numenta a cortical theory should be able to explain:**

i. how layers of neurons can learn sequences;
ii. the properties of SDRs;
iii. unsupervised learning mechanism with streaming temporal data flows;
iv. layer to layer connectivity;
v. how brain regions model the world and create behaviors; and finally,
vi. the hierarchy between different regions

These can be seen then as the six principles any biological or artificial intelligence should possess to be defined as such. Intuitively, it sounds a reasonable model, because the neocortex learns from sensory data, and thus it creates a sensory-motor model of the world. Unfortunately, we do not fully comprehend how the neocortex works yet, and this demands a machine intelligence be created flexible as much as robust at the same time.

In a more recent work, Hawkins and Ahmad (2016) turned their attention on a neuroscientific problem who is though crucial to the development of an AGI. They tried to explain **how neurons integrate inputs from thousands of synapses,** and their consequent large-scale network behavior. Since it is not clear why neurons have active dendrites, almost every ANNs created so far do not use artificial dendrites at all, and this would suggest that something is probably missing in our artificial structures.

Their theory explains how networks of neurons work together, assumed all the many thousands of synapses presented in our brain. Given those *excitatory neurons,* they proposed a model for sequence memory that is a universal characteristic of the neocortical tissue, and that if correct would have a drastic impact on the way we design and implement artificial minds.

Rocki (2016) also highlighted few aspects specifically relevant for building a biologically inspired AI — specifically, **the necessary components for creating a general-purpose learning algorithm.** It is commonly assumed that humans do not learn in a supervised way, but they learn (unsupervised) to interpret the input from the environment, and they filter out as much data as possible without losing relevant information (Schmidhuber, 2015).

Somehow, the human brain applies a sort of Pareto's rule (or a *Minimum Description Length* rule otherwise) to information it gathers through sensory representations, and keeps and stores only the information that can explain the most of what is happening. According to Rocki, unsupervised learning regularizes and compresses information making our brain a data compactor (Bengio et al., 2012; Hinton and Sejnowski, 1999).

In addition to being unsupervised, Rocki hypotheses that the architecture of a general-learning algorithm has to be**compositional; sparse and distributed; objectiveless;** and **scalable.** Human brain learns sequentially, starting from simpler patterns and breaking up more complex problems in terms of those simpler bricks it already understood — and this type of hierarchy and compositional learning is indeed well captured by deep learning.

As already pointed out by Ahmad and Hawkins (2015), **sparse distributed representations are essential,** and they are much more noisy-resistant than their dense counterparts. However, there are much more peculiarities that make SDRs preferable: there are no region-specific algorithms in the brain, but the cortical columns act as independent feature detectors. Each column becomes active in response to a certain stimulus, and at the same time, it laterally inhibits other adjacent columns, forming thus sparse activity patterns. Since they are sparse, it is easier to reverse engineer a certain external signal and extract information from it (Candès et al., 2006). The property of being distributed helps instead in understanding the causes of patterns variations. SDRs also facilitates the process described above of filtering out useless information. **They represent minimum entropy-codes**( Barlow et al., 1989) that provide a generalized learning mechanism with simpler temporal dependencies.

The reason why the learning process should not have a clear stated objective is slightly controversial, but Rocki — and Stanley and Lehman (2015) before him — support this argument as the only way to achieve and form transferrable concepts. Moreover, Rocki states scalability as fundamental for a general-learning architecture. **The brain is inherently a parallel machine,** and every region has both computational and storing tasks (and this is why GPUs are much more efficient than CPUs in deep learning). This would suggest an AI to have a hierarchical structure that separates local learning (parallel) from higher-order connections (synapses updates), as well as a memory that can itself compute, in order to reduce the energy cost of data transfers.

Rocki eventually concludes with some further functional rather than structural ingredients for the formation of an AI, namely: **compression; prediction; understanding; sensorimotor; spatiotemporal invariance; context update; and pattern completion.**

We discussed the importance of compression and sensorimotor before, and **we can think of AGI as a general purpose compressor that forms stable representations of abstract concepts —** although this point is controversial according to the no free lunch theorem (Wolpert and Macready, 1997) that indirectly states that this algorithm cannot exist. We can also see prediction as of a weak form of spatiotemporal coherence of the world, and then we can argue learning to predict to be equivalent to understanding. Finally, we need to incorporate a continuous loop of bottom-up predictions and top-down contextualization to our learning process, and this contextual spatiotemporal concept would also allow for a disambiguation in the case of multiple (contrasting) predictions.

***Hardware and AI***

As we explained before, the recent surge of AI and its rapidly becoming a dominant discipline are partially due to the exponential degree of technological progress we faced over the last few years. What it is interesting to point out though is that AI is deeply influencing and shaping the course of technology as well.

First of all, the Graphics Processing Units (GPUs) have been adapted from traditional graphical user interface applications to alternative parallel computing operations. NVIDIA is leading this flow and is pioneering the market with the CUDA platform and the recent introduction of Telsa P100 platform (the first GPU designed for hyperscale data center applications). On top of P100, they also created the first full server appliance platform (named DGX-1), which will bring deep learning to an entirely new level. Very recently, they also released the Titan X, which is the biggest GPU ever built (3,584 CUDA cores).

In general, the most impressive developments we observed are related to chips, especially Neuromorphic Processing Units (NPUs) ideated to emulate the human brain. Specific AIchips have been created by major incumbents: IBM has released in 2016 the TrueNorth chip, which it is claimed to work very similarly to a mammalian brain. The chip is made of 5.4 billion transistors, and it is able to simulate up 1 million neurons and 256 million neural connections. It is equipped with 4,000 cores that have 256 inputs lines (the axons) and as much output lines (neurons), which send signals only when electrical charges achieve a determined threshold.

This structure is quite similar to the Neurogrid developed by Stanford,although the academic counterpart is made of 16 different chips instead of the single one proposed by the software colossus.

Google, on the other hand, announced the introduction design of an application-specific integrated circuit (ASIC) thought and tuned specifically for neural networks — the so-called Tensor Processing Unit (TPU). The TPU optimizes the performance per watt specifically for machine learning problems, and it both powers RankBrain (i.e., Google Search) and DeepMind (i.e., AlphaGO).

Intel is working on similar chips as well, i.e., the Xeon Phi chip series, and the latest release has been named Knights Landing (KNL). KNL has up to 72 cores, and instead of being a GPU, it can be a primary CPU that reduces the need to offload machine learning to co-processors.

Even Qualcomm has invested enormous resources in the Snapdragon 820, and eventually into the deep learning SDK Snapdragon Neural Processing Engine and their Zeroth Machine Intelligence Platform.

The cost for all those chips is huge (on the order of billions for R&D, and hundred thousand dollars as selling cost), and they are not viable for retail consumers yet but only thought for enterprise applications. The main exception to this major trend is the mass-scale commercial AI chip called Eyeriss, released earlier in 2016 by a group of researchers at MIT. This chip — made of 168 processing engines — has been built on a smartphone’s power budget and thus is particularly energy-friendly, but it presents anyway computational limitations.

Even though this is a cost-intensive game, several startups and smaller companies are considerably contributing to the space: Numenta open-source NuPIC, a platform for intelligent computing, to analyze streaming data. Knowm, Inc. has brought memristor chips to the market, which is a device that can change its internal resistance based on electrical signals fed into it (and used as a nonvolatile memory). KnuEdge (and its subsidiaries KnuPath) created LambaFabric, which runs on a completely innovative architecture different not only from traditional GPUs but also from TPUs. Nervana Systems released an ASIC with a new high-capacity and high-speed memory technology called High Bandwidth Memory. Horizon Robotics is another company actively working in the space, as well as krtkl, which has produced a new low-cost dual-core ARM processor (FPGA, Wi-Fi, Bluetooth) named Snickerdoodle.

A final note has to be made in favor of Movidius, which introduced a completely new concept, i.e., an all-in-one USB for deep learning. Codenamed Fathom Neural Compute Stick, it contains a chip called Myriad 2, which has been thought in partnership with Google specifically to tackle down any advanced image recognition issue (but it has been used also to power drones and robots of a perse kind).

**References**

Ahmad, S., Hawkins, J. (2015). “Properties of sparse distributed representations and their application to hierarchical temporal memory”. arXiv preprint arXiv:1503.07469

Barlow, H.B., Kaushal, T.P., Mitchison, G.J. (1989). “Finding minimum entropy codes”. Neural Computation, 1(3): 412–423.

Bengio, Y., Courville, A.C., Vincent, P. (2012). “Unsupervised feature learning and deep learning: A review and new perspectives”. CoRR abs/1206.5538.

Candès, E.J., Romberg, J.K., Tao, T. (2006). “Stable signal recovery from incomplete and inaccurate measurements”. Comm. Pure Appl. Math, 59 (8): 1207–1223.

Hawkins, J., and Ahmad, S. (2016). “Why Neurons Have Thousands of Synapses, A Theory of Sequence Memory in Neocortex”. Frontiers in Neural Circuits, 10.

Hinton, G., Sejnowski, T. (1999). Unsupervised Learning: Foundations of Neural Computation. MIT Press.

Rocki, K. (2016). “Towards Machine Intelligence”. CoRR abs/1603.08262: 1–15.

Schmidhuber, J. (2015). “Deep learning in neural networks: An overview”. Neural Networks, 61: 85–117.

Stanley, K.O., Lehman, J. (2015). Why Greatness Cannot Be Planned — The Myth of the Objective. Springer International Publishing.

Wolpert, D.H., Macready, W.G. (1997). “No free lunch theorems for optimization”. Transactions on Evolutionary Computation, 1(1): 67–82.

This is an excerpt from my book “Artificial Intelligence and Exponential Technologies: business models evolution and new investment opportunities”, edited by Springer (2017).

DataScience®
foundation
MAKE BETTER DECISIONS

## About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

## Contact Data Science Foundation

Email:admin@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation